Big Data RFI Index		
#	Entity	Date
1	**Annie Shebanow**	3/21/2014
2	**Jackamo**	3/21/2014
3	**Peter Muhlberger**	3/24/2014
4	**Consumer Action**	3/25/2014
5	**ARM and AMD**	3/26/2014
6	**USPIRG and Center for Digital Democracy**	3/27/2014
7	**Information Technology Industry Council**	3/27/2014
8	**Consumer Federation of America**	3/28/2014
9	**Leadership Conference on Civil and Human Rights et al.**	3/28/2014
10	**MAG-Net**	3/28/2014
11	**Abraham Wagner**	3/28/2014
12	**Mary Culnan**	3/29/2014
13	**Georgetown University**	3/30/2014
14	**Intrical**	3/30/2014
15	**Access**	3/31/2014
16	**ACLU**	3/31/2014
17	**Advertising Self-Regulatory Council, Council of Better Business Bureaus**	3/31/2014
18	**Association for Computing Machinery**	3/31/2014
19	**Association of National Advertisers**	3/31/2014
20	**BSA | The Software Alliance**	3/31/2014
21	**CDT**	3/31/2014
22	**Center for Data Innovation**	3/31/2014
23	**Center for Digital Democracy**	3/31/2014
24	**Center for National Security Studies**	3/31/2014
25	**Cloud Security Alliance**	3/31/2014
26	**Common Sense Media**	3/31/2014
27	**Computer and Communications Industry Association**	3/31/2014
28	**Computing Community Consortium**	3/31/2014
29	**Consumer Watchdog**	3/31/2014
30	**Dell**	3/31/2014
31	**Direct Marketing Association**	3/31/2014
32	**Durrell Kapan**	3/31/2014
33	**Electronic Transactions Association**	3/31/2014
34	**Federation of American Societies for Experimental Biology**	3/31/2014
35	**Financial Services Roundtable**	3/31/2014
36	**Food Marketing Groups**	3/31/2014
37	**Future of Privacy Forum**	3/31/2014
38	**IMS Health**	3/31/2014

#	Entity	Date
39	**Interactive Advertising Bureau**	3/31/2014
40	**IT Law Group**	3/31/2014
41	**James Cooper**	3/31/2014
42	**Jason Kint**	3/31/2014
43	**Jonathan Sander, STEALTHbits**	3/31/2014
44	**Marketing Research Association**	3/31/2014
45	**McKenna Long & Aldridge**	3/31/2014
46	**Microsoft**	3/31/2014
47	**MITRE Corporation**	3/31/2014
48	**Mozilla**	3/31/2014
49	**NYU Center for Urban Science & Progress**	3/31/2014
50	**Pacific Northwest National Laboratory**	3/31/2014
51	**Privacy Coalition**	3/31/2014
52	**Reed Elsevier**	3/31/2014
53	**Frank Pasquale**	3/31/2014
54	**Sidley Austin**	3/31/2014
55	**Software & Information Industry Association**	3/31/2014
56	**TechAmerica**	3/31/2014
57	**TechFreedom**	3/31/2014
58	**Technology Policy Institute**	3/31/2014
59	**The Internet Association**	3/31/2014
60	**U.S. Chamber of Commerce**	3/31/2014
61	**US Leadership for the Revision of the 1967 Space Treaty**	3/31/2014
62	**VIPR Systems**	3/31/2014
63	**World Privacy Forum**	3/31/2014
64	**Constellation Research**	4/2/2014
65	**Fred Cate, Peter Cullen, and Viktor Mayer-Schönberger**	4/4/2014
66	**Healthcare Leadership Council**	4/4/2014
67	**Brennan Center for Justice**	4/4/2014
68	**Making Change at Walmart**	4/4/2014
69	**Online Trust Alliance**	4/4/2014
70	**MIT**	4/4/2014
71	**Coalition for Privacy and Free Trade**	4/4/2014
72	**EFF**	4/4/2014
73	**Privacy Coalition-Updated**	4/4/2014
74	**EPIC**	4/5/2014
75	**Kaliya Identity Woman 1**	4/5/2014
76	**Kaliya Identity Woman 2**	4/6/2014

#	Entity	Date
77	**Tyrone Grandison**	4/6/2014
78	**Open Technology Institute, New America Foundation**	4/8/2014

Annie Shebanow
[redacted]

My notes on Big Data RFI:

Privacy notices are difficult to draw and deliver for the data use. Policies are needed to safeguard data and information. There should not be differences between the government and the private sectors for policy frameworks or regulations on handling big data. Although we all know there are differences between government and private sectors' data use, the privacy policies should be general enough to apply to both entities.

For any government or private organization:

- Big data policies needed for securely handling, storing, and protection of structure, unstructured, and semi-structures data
- Privacy policy controls for big data must be part of organizations' operations, and those privacy policy controls should be public information and listed on the organizations' Web sites. These policies could be companies' own policies and the regulatory environments policies. Transparency is the key to safeguarding data privacy. These should be in bullet list format (not the large documents written by attorneys in attorneys' language, which no one reads them as some companies provide during signup process.)
- Different privacy policies needed for the type of data use
- Differentiate big data analytics policies for collecting individual's data directly (supplied by individuals), indirectly from third parties (posts to social media, videos, sensor data), or public records
- Global consortiums for big data analytics policies are needed to draft, modify standards and policies, and accepts membership (U.N. of Big Data Analytics Policies)
- Countries' memberships protects their citizens' data, lack of membership may create economic downfall. This will incentivize best practices for big data analytics policies.
- Special privacy policies for those countries that are not members of big data privacy consortiums and when data is originated from those countries
- Fair information practices privacy policies needed for information sharing
- Big data privacy announcement policies needed for data and information sharing
- Random auditing policy for ensuring privacy policies' of organizations (It could be similar of SOX-404 process)
- Big data analytics policies whistleblower measures needed
- Policies needed to encourage big data analytics privacy policies education in our educational systems for the understanding and protection of citizens' fundamental rights and providing information on existing data privacy policies and standards

[redacted]

From: Jackamo <[redacted]
Sent: Friday, March 21, 2014 8:23 PM
To: bigdata@ostp.gov
Subject: Big Data & Big Brother

Dear Big Brother,
I know you love me and want to take care of me, that's why you want "Big Data" to love me too. But I don't love you Big Brother or your friend Big Data. As a matter of fact Big Brother, you can take a flying bit-of-my-ass and Big Data can choke on snot and slobber all over itself-just don't bother my right to be left alone.

I hope this doesn't seem impolite or rude Big Brother, but it seems to me you need to find something more important to do; like finding a cure for clap or the common cold and stop wasting the taxpayers money on this frivolous paranoid set of fantasies you've been having ad infinitum about monsters from the ID, etc. Maybe you need a new therapist Big Brother? Maybe, you can discuss with him or her why you're so anal retentive and obsessed about collecting other people's business. At any rate if, all else fails Big Brother: just relax, pour yourself a favorite drink, then lay down and take it easy, and all those bad thoughts will go way-at least for a while.

Love,
__JPC__

[redacted]

From: Muhlberger, Peter <pmuhlber@nsf.gov>
Sent: Monday, March 24, 2014 2:55 PM
To: bigdata@ostp.gov
Subject: [Big Data RFI]

To contextualize my comments, a little about me: Till Oct. 2013, I was the National Science Foundation's program director for cyber social sciences. This position involved me in questions of cybersecurity and, thereby, data privacy. I am now in an unrelated position, but remain interested in issues of big data privacy. My views here are my own and not the views of the National Science Foundation. My background includes political science, public policy, political psychology, and data intensive social science.

I am still learning about the policy context of big data privacy. Consequently, what I say here is an impression at this point. Nevertheless, it does seem that there are some important points to be made. The following clarifies what I perceive to be limitations of the current policy framework regarding big data and privacy issues, and I sketch a more complete framework. I will address specific questions with respect to this RFI (request for information) after elaborating this framework. I comment on four of the five questions below, identifying them by question number.

It is my impression that policy makers are acutely aware of the potential social and economic value of big data, but are less knowledgeable regarding the long-term implications of big data analytics for individuals and society ('privacy') or the related ethical implications. For instance, Leon Panetta, in the recent MIT / White-House meeting on big data privacy discussed detailed examples of the benefits of big data and, seemingly as an afterthought, mentioned that we should seek to achieve these benefits while preserving our ethical mores with respect to privacy. He did not mention what these mores are, how they bear on big data, or how they might be preserved. The White House data privacy framework identifies privacy as a means to achieving the trust needed to fully exploit big data, but does not elaborate on privacy as an ethical ends or in terms of potential social implications.

The discussion, then, seems framed in terms of how to extract the benefits of big data while providing some protection for a less than fully elaborated notion of privacy. That notion of privacy seems to largely consist in avoiding alarm by the public with respect to the spread of individually-identifiable data. Much attention is also focused on technical solutions such as k-anonymity or differential privacy that allow sharing of data--effectively, expanded use of big data--while making a bow to privacy by preventing identification of individuals. This was much of the focus of the MIT / White-House meeting.

A notion of privacy focused on the spread of individually identifiable data cannot adequately capture the underlying issues behind privacy concerns. Privacy is only a subset of a set of privacy-related concerns raised by big data that pose substantial social and ethical risks. Technological solutions to insuring individuals are not identified in big data are quite inadequate to address these risks. What is needed is an expanded notion of risk and a robust regulatory regime, based on strong rights, that continually incorporates input from the public.

First, it is not evident that technological solutions can even protect the simple notion of privacy: that of anonymizing individual data. Syntactic models of anonymity, such as k-anonymity and l-diversity, have been shown to be seriously flawed because combining an anonymized dataset with other publicly-available datasets

allows re-identification of large numbers of people in many circumstances. Big data is precisely about integrating a diversity of datasets, but this then opens many paths to identifying individuals. Differential privacy circumvents this problem by not sharing anonymized datasets. Data is given to a curator who only answers questions based on the data and the answers contain statistical noise to prevent leakage of individual information. This works, however, only if the number of related questions is constrained. That is, queries must be put on a budget. Overall, this arrangement is far too onerous for organizations to willingly adopt it for internal purposes. It means giving up control over a crucial resource, making data difficult to analyze, and preventing resale to others, as well as constraining uses that are best done with non-anonymized data, such as advertising.

Those organizations and persons most likely to use big data for the public good--university researchers, non-partisan non-profits, and government agencies--are those most likely to be potentially hamstrung by technical solutions to preserving individual anonymity in data. Such organizations and persons are the least able to protect themselves from onerous legal requirements with respect to data privacy.

Second, the social and ethical risks of big data are appreciably broader than a narrow notion of privacy as individual anonymity. Some aspects of this are evident from considering the functions that privacy is meant to serve. An older academic literature on the social psychological functions of privacy, now largely ignored in the more technology-focused privacy literature, discusses privacy as a crucial element in a person's construction of their identity or sense of self. Privacy is about limiting information about oneself so as to construct boundaries that create the conditions for and also help define personal identities.

The need for privacy comes about, in substantial part, from the recognition that the construction of personal identity is vulnerable to social influence. Other people could influence, perhaps even control, the construction of personal identity if they have access to sufficient private information. This information could allow them to utilize means of social control--social shaming, reputational attacks, and interpersonal manipulation--to alter the construction of personal identity and otherwise impose harms, such as limiting future options.

Personal identity construction is not something merely confined to young people, particularly in a society in which people are moving at an increasing pace into new jobs, new roles more generally, and new geographic locations. Research in psychology indicates that few adults have fully integrated identities, which means that they accept alternate and often conflicting roles and identities. Numerous experimental studies have found that framing issues in different ways, ways that often evoke different identities and roles, elicits sharply different attitudes and decisions in individuals.

In other words, many people have roles and identities that are sufficiently poorly integrated that knowledgeable actors can readily manipulate them—cause them to act or think in a way they would not if they knew as much as the manipulator. It is well known that social identities powerfully affect political attitudes and behavior. Such identities have apparently been used quite successfully to manipulate the public in the political sphere. Various lines of psychological and sociological research show that identities and roles guide much of day-to-day behavior as well — as people make their way through the day by enacting such roles as 'parent,' 'employee,' 'consumer,' 'friend,' 'husband,' 'citizen,' and many others. Because these roles and identities are less than perfectly defined and integrated in most individuals, people are open to manipulation. An organization that could subtly shift what a person understands by, for example, 'employee,' 'consumer,' 'citizen,' or 'friend' or can affect when these roles are activated or the tradeoffs between them could powerfully affect behavior.

Big data increasingly puts large organizations in the position of being sufficiently knowledgeable about individuals or groups of individuals that they can effectively impose social controls, particularly manipulation, on individuals. The social science applications of big data and computationally intensive modeling and analysis techniques are expanding rapidly. Today, advertisers build profiles of individuals. In the not too distant future, social scientists may well be able to take the sum total of what people have said and done online and done within view of increasingly omnipresent sensor networks to construct detailed models of people's beliefs and understandings as well as psychological, cognitive, and behavioral propensities.

With such information, organizations will seek to determine what to say and do to people to shift these people to a new belief system, personality, or identity configuration that will prove beneficial to the organization manipulating that person. 'Doing to' a person could be something as subtle as providing targeted discounts for certain activities that shift behavior and self-definitions over the long term. While for the most part organizations have thus far shied away from other means of social control such as shaming and reputational attacks, such means could come to be utilized in the absence of a clear and well-enforced regulatory regime against them or given expanded possibilities to shame or attack people anonymously.

The prospects for capturing key aspects of people from social media and other data and being able to use this information to greatly influence them beyond existing levels of influence are, of course, speculative. Nevertheless, large companies have put down multi-billion dollar wagers that big data about people can help them affect their behavior. We also know that people who know us intimately can 'push our buttons,' that most people are readily manipulated by psychological experiments, that people seem to be manipulated at a substantial scale by political campaigns, and that a person's daily activity and interaction contexts greatly influence behavior. Big data is creating a world in which large organizations will likely have sufficient access to background information about people to potentially know as much about them as intimates, if not more. This is all the more possible to the extent that people increasingly use social media and related contexts to interact with their friends and to develop new identities and roles. These behaviors are permanently captured and then shared between organizations. What remains is for natural language, text mining, and other techniques to be further developed to make powerful inferences from this vast quantity of information.

To the extent that good modeling of individual propensities will be possible in the not too distant future, oligopolistic organizations could use such information to more thoroughly influence people by presenting them with certain frames and belief structures. This is already done in relatively poorly targeted ways today in political campaigns. Organizations could also tamper with people's marginal costs of activity—structuring their environments and behavioral affordances to influence people's development. For example, social media companies could alter the likelihoods of who people are likely to encounter online, to insure they have the 'right' friends to influence them in a particular direction. Prices, advertising, and the availability of information (e.g., search engine top results) could be modified to insure people encounter the 'right' information. Search companies already personalize search results based on past behavior, and companies already seek to influence the prices and information to which individuals are exposed.

The full set of such developments in technologies of manipulation remains speculative, but such developments do seem to be in the realm of possibility. Consequently, the question is whether to risk potentially substantial social and ethical harm by allowing, as we currently are, the building of a vast infrastructure by large organizations to accumulate and analyze big data about people and to influence people's environments in various ways.

A technology of manipulation would, of course, be most effective if attuned to each individual. Nevertheless, even without information about specific individuals, a greatly improved technology of manipulation would still be possible with big data and emerging computationally-intensive methods. Simply knowing, in vast quantity and detail, the behavior and verbal expressions of highly specific types of individuals should prove very useful in determining how to influence individuals of these types. Thus, even if individuals were successfully anonymized in big data, by technical or legal means, it would remain possible for organizations with access to such data to use it to determine how to manipulate rather specific types of people, which these organizations could identify without big data. Thus, the privacy of individuals could be maintained even while individuals were being harmed. Because of this, the ethical implications of big data should be construed more broadly than the issue of individual privacy.

The solutions to the above concerns are apt to be complex and multi-faceted, particularly with respect to the tradeoff between the potential social and private value of big data and its potential harms. Nevertheless, any adequate solutions are apt to embrace a number of key ethical principles that form a set of 'big data rights': the right to know, the right to anonymity, the right to be forgotten, and the right to be heard. Aspects of these principles also appear in the Consumer Privacy Bill of Rights (CPBR). There are, however, a number of ways in which CPBR would need to be fortified to be up to the task of fully addressing the potential social and ethical harms of big data. First, the rights must be treated as strong ethical rights. A strong right is not one that should be subject to definition and voluntary self-restrictions by those entities with the most to gain in violating those rights—organizations with large stakes in big data. The public and academia should be most involved in defining these rights and monitoring their protection in a changing environment. Second, people have a deep right to know—one that extends to the inferences and profiles drawn about them and to exactly how these inferences and profiles are being used to affect them.

People have a deep right to know what organizations know about them and how this knowledge is being used with respect to them and what would be different in their environment or information exposure were they not known to these organizations. For example, people should have access to a service that would show them search results that were withheld from them based on prior information held by a search firm or commercial firm. They should also have access to a service that presents an accessible summary of what organizations know about them, including inferences, and access to organizations' detailed information about them as well as information on how the inferences organizations make about them are being used.

Once they know how their information is used, people need a practical mechanism to help them articulate their concerns and have these concerns impressed upon policy makers. That is, it is insufficient for people to only be able to express private alarm about what they learn concerning the uses of big data; they need public fora that allow them to articulate their concerns and relay collective concerns to policy makers. This is a meaningful 'right to be heard.' Policy makers should welcome continual input from the public with respect to something that could drastically and not necessarily positively shape society.

People should have a default right to anonymity, balanced against the social benefits of the data. Even when people are in public, they have the expectation that they are not being tracked and analyzed--which is stalking. That is, people do not expect, even in public places, that someone is recording their every move and making inferences about those moves that add up to a detailed picture that may well invade their privacy. Similarly, people do not expect their online activities to be carefully tracked across websites, their utterances on social media sites to be forever recorded, analyzed, and data or inferences shared with other organizations. Until they obtain consent from affected individuals, organizations

whose purposes are not entirely focused on the public good should not be allowed to track and analyze individuals via big data methods. Perhaps such organizations could persuade the public to make their data available by having a third option of temporary tracking so the organizations could show people the benefits of their maintaining information about them.

Such data and the consent to collect it should, however, be erasable at request as well. This is the right to be forgotten. People seeking a new start, as many do, were once able to move to a new town and begin anew, without others having preconceptions about them. Identity development and change more generally are also facilitated by a world in which others do not have perfect records of a person's past. A right to be forgotten is therefore important. CPBR is thus far ambiguous with respect to a right to be forgotten. Corporations are, for example, not held responsible for data they do not control. This seems to allow that a company will collect and share, with third parties, very substantial data about an individual, which data is then out of the reach of a right to be forgotten because the company no longer controls the third party data. A serious right to be forgotten would require every piece of data collected by a company to receive a unique identifier that would allow, with the right data systems, all instances of this data to be erased, even if held by third parties. Moreover, not only should the data be erasable, but all inferences built based on this data should be undone. These inferences pose the greatest dangers to privacy.

Having presented a framework for understanding big data and privacy-related issues, the specific questions from this RIF can be better addressed:

Question 1: The current U.S. policy framework is insufficient to fully address the ethical and social implications of big data. Too much emphasis is currently placed on technical solutions to insuring the privacy of individual data, which is not the central issue with respect to the potentially adverse implications of big data. Organizations such as information sector firms will likely not be required to use the more effective technical solutions internally. The existing framework does not capture the most important aspects of the functions of privacy--namely to protect people who are vulnerable to manipulation of their identities, roles, and beliefs. With such concerns in focus, the chief problem presented by big data shifts to the many possibilities big data affords to large organizations for improving their methods of influence. While the prospects for a vastly improved technology of manipulation based on big data remain speculative, these prospects are not implausible, companies are making big wagers that such technologies will prove effective, and a risk-averse approach should be taken to minimize potential harms of such advances. In particular, it is important to take seriously the notion that people have strong rights with respect to privacy that supersede the economic benefit of exploiting new technologies. Taking such rights seriously would require a far more stringent regulatory scheme than is being currently considered.

Question 2: Given the potential for large-scale abuse of big data technologies, there should be measures in place to insure that any organization that does not operate exclusively in terms of the public interest, understood in a non-partisan way, should be subject to constraints with respect to their big data capabilities, particularly as outlined in the 'big data rights' above. Such rights implicitly call for strong regulation of these organizations. In the case of organizations that pursue the public good, rights should be balanced against potential for contributions to the public welfare. Commercial organizations cannot be assumed to be exclusively or even primarily concerned with the public good. Universities and government agencies, on the other hand, are tasked with pursuing the common good and, ideally, act under careful scrutiny and transparency rules, unlike commercial organizations.

Question 3: This question is part of the problem. Technological solutions to insuring the privacy of individual data are not a sufficient answer. Instead,

technology is more likely to be part of the problem. Of concern is the development of technologies that allow deep knowledge of people's propensities from unstructured data about them, such as their movement patterns and social media postings. Also of concern are technologies that allow knowledge about people to be used to shape their opportunities and environment.

Question 4: See answer to Question 2. Though government agencies are tasked with pursuing the common good, agencies have rather different cultures with respect to what constitutes this good and how it might be pursued. Care must be taken to insure that agencies do not undermine a general notion of the public good by avidly pursuing their own narrower conception. A robust public voice would be helpful in insuring that agencies and other organizations operate in the public interest.

consumer action
Education and advocacy since 1971

www.consumer-action.org

PO Box 70037
Washington, DC 20024
202-544-3088

221 Main St, Suite 480
San Francisco, CA 94105
415-777-9648

523 W. Sixth St., Suite 1105
Los Angeles, CA 90014
213-624-4631

Consumer Action comments to the Office of Science and Technology Policy regarding Government "Big Data" Request for Information

Submitted 3/31/14 via bigdata@ostp.gov

Big Data Study
Office of Science and Technology Policy
Eisenhower Executive Building
1650 Pennsylvania Ave. NW
Washington, DC 20502

Dear Deputy Wong:

We are pleased to submit comments to you on the Government's "Big Data" study and appreciate the White House's commitment to providing forums for public input.

As a 42-year-old national nonprofit, Consumer Action works to financially empower disadvantaged communities, including communities of color, immigrants, seniors and those struggling with debt and poverty. We have a national network of over 8,000 community-based organizations to which we provide financial education programs, staff training and expert advocacy on a variety of consumer issues such as credit, housing, insurance and privacy.

For us, the discussion on Big Data has a different lens: the very real and increasingly frequent harms associated with the loss of privacy and its disproportionate impact on underserved communities. Ad networks and marketing companies are including more nuanced and opaque data targeting mechanisms in their collection technology, setting their sights on identifying as much about a user as possible, including sensitive information such as specific demographics, personal characteristics and health conditions.

Because Hispanic and African American populations represent the fastest growing users of mobile technology while also being the most vulnerable to quantifiable privacy harms like identity theft and fraud, this raises serious privacy questions for public policy. For example, researcher Ashkan Soltani has unearthed evidence that points to dynamic pricing logarithms being built in such a way that those determined to be poor or from a minority group *pay more* than those who are profiled as more wealthy. This requires closer investigation by academics and policymakers. The use of location as a data point for marketers makes this even more pressing. The use of GPS data to pinpoint mobile device movement around a retail environment is almost always done without the consumer's knowledge or permission. We believe that location is an especially sensitive

category of information that should be protected under law.

Data collection and sharing is ubiquitous, invisible, intrusive and lawless. Consumers are mostly unaware of the collection and sharing being done and this is not an accident. Industry players have made it extremely difficult for even the most astute consumers to learn about and attempt to use some privacy controls, even though such controls hardly exist in any meaningful way.

Current regulation and policy mostly relies on the Federal Trade Commission (FTC), which must use its limited resources to attack a giant and moving target, and on self-regulation, a thoroughly debunked approach to protecting consumers. Public policy should move away from this inadequate framework toward one that includes a basic consumer privacy bill of rights, such as the one proposed by the White House in 2012, and more robust tools and additional rulemaking authority for the FTC. The primary goal of a baseline bill would effectively be to peel back the layers on data collection in order to effectively legislate current and emerging practices, allowing states to continue innovating on specific privacy areas if they so choose. We also would support the inclusion of a special protection clause for disadvantaged communities in a baseline privacy bill. Ideally, this bill would also include some accountability for companies who violate the law, such as civil penalties and monetary remedies for consumers who are victims of harm.

Another area of concern is the forthcoming data stew of facial recognition technology combined with retail analytics. When this high-powered system targets disadvantaged communities, it could be disastrous. It's not hard to imagine a scenario in the very near future that includes someone receiving a "diagnosis" born from drugstore analytics, inferences made using information such as past medical history, gait, ethnicity and perceived lifestyle habits (a face with more wrinkles might indicate a smoker). This could not only create erroneous profiles that are impossible to correct, it could have the effect of cutting out doctors (the middleman) and leaving consumers vulnerable to the medical "advice" of giant pharmaceutical and healthcare companies who obviously have a financial, rather than ethical stake, in customer well-being.

Thank you for the opportunity to comment on this very important study. We look forward to working with you to make sure consumers are given transparency, control and accountability when it comes to their increasingly public personal data.

Sincerely,

Michelle De Mooy
Senior Associate, National Priorities
CONSUMER ACTION
michelle.demooy@consumer-action.org
301-244-5081

OSTP Consultation on Big Data: Joint ARM AMD Submission

Introduction

ARM and AMD welcome this opportunity to contribute to the debate through OSTP. ARM designs microprocessors, used in over 90% of mobile phones and many other products, including servers. AMD is a global leader in CPU and GPU design.

An important point to keep in mind from the start is that Big Data starts with Little Data.

Big Data is the agglomeration of many little pieces of data. This is already happening: data about what online sites consumers visit, what purchases they make, what their health apps record, is capable of being aggregated. This will increase as the Internet of Things (IoT) revolution gathers pace. IoT holds out the prospect that many objects will have embedded sensors and will be able to record information about their performance or their environment and send that information somewhere else, probably including to a Big Data storage centre in the Cloud. Popular estimates predict that billions of devices will be connected in this way .

We believe that, properly used, (big) data can help drive economic growth. But there will be concerns about the use and misuse of data. It is important that the debate recognises this while at the same time acknowledging that there is a risk that mishandling the data issue risks stifling innovation.

Our Approach

ARM and AMD have drawn up some principles to guide the data debate. An essential component is security.

Our hope is that these principles will form the starting point for an industry led conversation about a framework designed to give consumers confidence in how their data is handled.

Specifically, we believe that the correct approach to the discussion should consider the following **key points**:

1. ***Effective IT security features and technology solutions are the first and final defense against malicious intrusions and cyber-attacks.*** Robust, open standards-based approaches to IT security that promote interoperability are the most efficient means to ensure broad coverage and widespread adoption.

2. ***No single technology, industry standards body, public policy, or government agency can fully address the rapidly evolving security threats in cyberspace.*** The entire IT ecosystem, industry standards bodies, government regulators and enforcement agencies, and individual consumers and businesses must all assume responsibility for securing cyberspace and protecting data and privacy.

3. ***A common set of policy principles and a framework for data-handling is needed for policy discussions.*** Industry bodies, government agencies, technology consumers, and other stakeholders need a shared understanding of key policy issues and approaches to data security and privacy so that policy discussions can be effectively and productively pursued.

Six Principles for the IOT Data Discussion

ARM and AMD propose the following policy principles as a starting point to guide policy discussions for securing and maintaining privacy for IOT data.

• **Consumers should own their own data**

Work by the World Economic Forum (WEF) suggests online data falls into three categories: (i) data volunteered in the context of a contract; (ii) anonymised data (e.g. "How many cars are in a traffic queue based on mobile phone signals?"); and (iii) between these two, data which is observed about someone without their knowledge, whether directly or through the transfer of their data to a third party.

Over time, we could aim for categories (i) and (ii) to expand, thus reducing category (iii) about which there is most concern. This will require consumers to be more aware of the fact that they should be able to determine what is done with their data: in short that they own it.

• **Data can drive economic growth, and provide a multitude of societal and individual benefits**

Data can have significant economic benefits and help to drive wide economic growth. It can also help improve delivery of services in health, environmental management, smart cities etc.

• **Not all data is equally sensitive**

Consumers may be content for their data to go to certain recipients but not others: you might be happy for your health data to go to your doctor but not to you insurance company. Second, provided the chain of custody is secure, consumers may be happy to share their data with a number of recipients who could use it to offer new services, provided it does not fall into the hands of people who might misuse it (for identity theft, or to track movements etc.). Anonymised data is likely to be less sensitive than identifiable data, particularly where it is clear that anonymised data is used for public benefits.

• **Consumers must have confidence in how their data is used, stored, and transported**

More needs to be done to reassure consumers about the security arrangements for (i) protecting their data against hacking and (ii) ensuring their data is not wrongfully transferred

to an unauthorised recipient. An important aspect of this is informing consumers of the benefit they gain from agreeing to their data being used in various ways.

• **Technology is a significant part of the solution**

ARM is working on several areas to improve the security of data, and to enhance consumer confidence in having control over their data.

• **A data-handling framework that categorizes different types of data and associated management strategies is required to unlock the potential of IOT**

This needs to bring together specific proposals on how to put these principles into practice. Its aim should be to reassure consumers while at the same liberating data to drive innovation.

An Initial Idea for a Data-Handling Framework

Different types of data should be managed differently. By establishing specific categories of data and associated responsibilities and mechanisms to manage each category of data, a framework can be established that provides an efficient means of addressing data security and protection.

For example, the following types of data should be managed differently:

(i) Highly sensitive data – health, financial, individual communications, trade secrets, etc;
(ii) Volunteered data in context of a transaction or enabled via consent (i.e. Opt-In);
(iii) Observed data about one's interests, activities, movements, etc. that is collected with one's consent (i.e. Cookies);
(iv) Observed data about one's interests, activities, movements, etc. that is collected without one's consent (i.e. web trackers); and
(v) Anonymized or de-identified data (i.e. anonymous surveys).

Highly sensitive data must be fully protected with very high assurance. Data that is volunteered, depending on the instrument and context, may be less sensitive and require a lesser degree of protection and assurance. Similarly, data that is anonymized or de-identified, assuming assurance that it has been sufficiently anonymized or de-identified, is much less sensitive than data associated with a specific individual. In short, establishing such a framework can help target the specific level of security and privacy protection that is appropriate for different types of data.

Clearly, there is a great deal of work that needs to be done to address the security and privacy issues raised by the Internet of Things era. But just as clearly, the IOT era is already providing tremendous benefits, with the promise of truly transformational change and benefits going forward for individuals and society as a whole.

Your Specific Questions

The Consultation invited us to address some specific questions. Many of our answers can be found in the section above on Our Approach. Additional points are indicated below:

(1) What are the public policy implications of the collection, storage, analysis, and use of big data? For example, do the current U.S. policy framework and privacy proposals for protecting consumer privacy and government use of data adequately address issues raised by big data analytics?

- We need a framework (see above).
- It may helpful to distinguish between certain key uses of data.
- There may be two broad categories:
- (i) consumer as target: this is where the aim of assembling and analysing data about a consumer is to offer them additional products or services. In some cases these offers will be directly linked to an action a consumer has taken on line (researching specific products or entering into a specific contract). In some cases it may be the result of drawing inferences (eg about a consumer's life style based on various sources of information). There is a particularly sensitive area where data is used to make sensitive 'predictions' or inferences about consumers health etc (other than general life style deductions).
- (ii) consumer as topic: this is where the primary aim of analysing data is to generate a conversation about the consumer eg where, without your permission, your health data is sent to your insurer, or your financial data goes to your mortgage company or employer etc. Consumers are likely to be much more sensitive in general about (ii) and about sensitive 'predictions' or inferences .

(2) What types of uses of big data could measurably improve outcomes or productivity with further government action, funding, or research? What types of uses of big data raise the most public policy concerns? Are there specific sectors or types of uses that should receive more government and/or public attention?

- See comments on (1) above.
- Once consumers have confidence in how their data is handled, it can be liberated to drive economic growth. This will generally be focused around providing additional marketing opportunities.
- But Big Data may have important other uses too eg in helping to manage traffic follows more precisely, in saving energy (by helping consumers mange their consumption more effectively) in preserving infrastructure (if we could monitor water flows through pipes we could reduce the incidence of leaks), in researching health outcomes.
- Consumers will be much less sensitive about some of the application in the previous bullet, because data will probably be deidentified or anonymised.

(3) What technological trends or key technologies will affect the collection, storage, analysis and use of big data? Are there particularly promising technologies or new practices for safeguarding privacy while enabling effective uses of big data?

- See comments in ‘Our Approach’ above.
- The security of data transmission (as well as handling) will be a key factor in maintaining consumer confidence.
- We are working together with other partners to advance the ARM® TrustZone® technology. TrustZone® allows consumers and businesses to secure their data and perform secure transactions, such as banking transactions, with a much greater level of trust and protection than current technologies.
- We're making good progress in providing secure technologies, but we also realize that rapidly evolving security threats online cannot be addressed by any one company, standards body or government. The job is too big for anyone to do alone and protecting data and privacy is a shared responsibility. As cybersecurity challenges continue to evolve, we can't point to others to solve the problem. We need collaborative industry development and public-private partnerships to more fully secure cyberspace.

(4) How should the policy frameworks or regulations for handling big data differ between the government and the private sector? Please be specific as to the type of entity and type of use (e.g., law enforcement, government services, commercial, academic research, etc.).Show citation box

(5) What issues are raised by the use of big data across jurisdictions, such as the adequacy of current international laws, regulations, or norms?

- The Big Data market is a global business. We need to ensure that approaches are harmonised between various jurisdictions.

March 2014

ARM, AMD

[redacted]

From: Jeffrey Chester <jeff@democraticmedia.org>
Sent: Thursday, March 27, 2014 10:38 AM
To: bigdata@ostp.gov
Cc: Wong, Nicole; Edelman, R. David
Subject: Big Data RFI (from USPIRG/CDD)
Attachments: USPIRGEFandCDDBigDataReportMar14_1.3web.pdf

Dear Nicole and David:

USPIRG Education Fund and the Center for Digital Democracy released this report today in order to submit it for the "Big Data" review. The report analyzes how Big Data practices are used in today's consumer financial marketplace and calls for new safeguards. Among the issues examined in the report, "Big Data Means Big Opportunities and Big Challenges: Promoting Financial Inclusion and Consumer Protection in the 'Big Data' Financial Era," are the following:

- the plight of "underbanked and unbanked consumers," who face special challenges in the new financial marketplace;

- the impact of data collection and targeted advertising on all Americans, most of whom have no idea that their personal data shape the offers they receive and the prices they pay online;

- the use of murky online "lead generation" practices, especially by payday lenders and for-profit trade schools, to target veterans and others for high-priced financial and educational products; and

- the need for new regulatory oversight to protect consumers from potentially discriminatory and deceptive practices online.

Let me know if the report exceeds your limit; if so, we can send in an excerpt. Ed Mierzwinski and I and colleagues from the financial reform community are also happy to discuss these issues with you as well.

Regards,

Jeff

Jeffrey Chester
Center for Digital Democracy
1621 Connecticut Ave, NW, Suite 550
Washington, DC 20009
www.democraticmedia.org
www.digitalads.org
202-986-2220

Big Data Means Big Opportunities and Big Challenges

Promoting Financial Inclusion and Consumer Protection in the "Big Data" Financial Era

U.S. PIRG Education Fund

CDD center for digital democracy

Big Data Means Big Opportunities and Big Challenges

Promoting Financial Inclusion and Consumer Protection in the "Big Data" Financial Era

Jeff Chester, Center for Digital Democracy

Edmund Mierzwinski, U.S. PIRG Education Fund

March 2014

Acknowledgements

Thanks to Gary Larson of Center for Digital Democracy for editing and citation review. Thanks to Julia Christensen of U.S. PIRG Education Fund for background research.

This research was funded by the Ford Foundation and the Annie E. Casey Foundation. We thank them for their support but acknowledge that the findings and conclusions presented in this report are those of the authors alone, and do not necessarily reflect the opinions of the Foundations.

About the U.S. PIRG Education Fund: With public debate around important issues often dominated by special interests pursuing their own narrow agendas, **U.S. PIRG Education Fund** offers an independent voice that works on behalf of the public interest. U.S. PIRG Education Fund, a 501(c)(3) organization, works to protect consumers and promote good government. We investigate problems, craft solutions, educate the public and offer Americans meaningful opportunities for civic participation. For more information, please visit our website at www.uspirgedfund.org.

About the Center for Digital Democracy, also a 501(c)(3) organization: The Center for Digital Democracy is at the forefront of research, public education, and advocacy on protecting consumers in the digital age. It has helped foster widespread debate, educating a spectrum of stakeholders, and creating a legacy of government and self-regulatory safeguards across a variety of Internet and digital media platforms. CDD's public education programs are focused on informing consumers, policy makers, and the press about contemporary digital marketing issues, including their impact on public health, children and youth, and financial services. For more information, please visit our website at www.democraticmedia.org.

Design by Alec Meltzer, meltzerdesign.net.

Table of Contents

Introduction

Dramatic changes are transforming the U.S. financial marketplace. Far-reaching capabilities of "Big-Data" processing that gather, analyze, predict, and make instantaneous decisions about an individual; technological innovation spurring new and competitive financial products; the rapid adoption of the mobile phone as the principal online device; and advances in e-commerce and marketing that change the way we shop and buy, are creating a new landscape that holds both potential promise and risks for economically vulnerable Americans.[1] Using advances in data analytics specifically to promote economic inclusion and fairness during this period of transformation in the U.S. economy should be a proactive strategy embraced by all stakeholders. While not a panacea to address growing financial inequality, a wise investment in strategies that harvest the potential of the new digital financial system may better enable struggling Americans to maneuver a difficult economic future.

It is also possible, however, that the emergence of a powerful data-driven "Banking 3.0," (as it is sometimes called), and the shift to a digital and mobile services financial system, could provide further obstacles to the consumers most at risk today—imposing new forms of unaffordable loans, discriminatory pricing, escalating fees for services, and unfair marketing practices.

The consequences of the financial meltdown, a low-wage-centered economy, historic barriers to equitable access to education and opportunity, and the too-long-ignored one-in-six Americans who live below the poverty line, have placed tens of millions at the margins—or out of reach—of the financial system.[2] Families and children are at particular risk: More than a third of single mothers and their children now live below the poverty line;

"But beyond the grim statistics and the heart-breaking daily struggles that so many face, there is also the glimmer of a potential opportunity. Underbanked and unbanked Americans are not totally cut off from the changes that are reshaping financial services. In fact, they can help to influence its direction and growth."

since 2000, poverty has grown 54 percent in our suburbs alone.[3]

But beyond the grim statistics and the heart-breaking daily struggles that so many face, there is also the glimmer of a potential opportunity. Underbanked and unbanked Americans are not totally cut off from the changes that are reshaping financial services. In fact, they can help to influence its direction and growth. Digital services—especially mobile phones—play an extraordinary role in how the new financial services marketplace delivers products and services for banking, credit, shopping, and more. Many lower income and other consumers already rely on mobile phones to access the Internet (illustrating the realities of the "digital divide" that makes wireline-based broadband connections out of reach).[4] Nearly 90 percent of underbanked consumers have mobile phones; more than half have "smart phones." Nearly half of the underbanked already use their mobile phones for banking services.[5] As the Federal Reserve recently explained, "smartphone ownership growth for underserved consumers is higher than other consumer groups because of the low cost and PC-like functionality of today's modern mobile handsets. As such, many of the underserved are migrating directly from cash–based payments to mobile (prepaid) accounts." The policy issues raised by the lack of affordable and equitable access to residential broadband continue to require debate and intervention strategies. However, it is increasingly recognized that mobile phones are quickly becoming the leading online device, overtaking the role that personal computers have traditionally played.[6]

As the marketplace looks to define "best" mobile marketing practices, as regulators seek to establish rules and guidelines, and as the financial and other digital industry sectors seek to build new markets (such as "hyper-local" targeted online advertising), ensuring that economically vulnerable Americans benefit should be a strategic goal during this transition period.

This report is intended to provide an overview of key issues for policymakers, thought leaders, and others concerned about financial inclusion and economic opportunity. The authors hope that its recommendations for further research, engagement, and outreach provide guidelines for next steps.

"But beyond the grim statistics and the heart-breaking daily struggles that so many face, there is also the glimmer of a potential opportunity. Underbanked and unbanked Americans are not totally cut off from the changes that are reshaping financial services. In fact, they can help to influence its direction and growth."

I. The "Connected Consumer" and the Underbanked Consumer Now Live in a Big Data World

1. The Data-Dependent World We Live In

Today, we live in an increasingly data-dependent world. A historic transformation of society is taking place, as data processing and the digital media further converge, ultimately blurring the divisions that now exist between the physical and online worlds. Powerful computers, communications networks, sophisticated data-processing techniques, and the use of Internet-connected devices are the foundation of the economy and contemporary society. Each day, billions of transactions from millions of consumers are instantly collected, analyzed, and processed for subsequent marketing.[7]

Financial services companies are repositioning their operations and relationships with consumers to take advantage of these changes. Banks, credit, and retail companies have invested to ensure they have the ability to gather, analyze, and make actionable—instantly—information about our offline and online behaviors—especially those related to our finances and spending. They understand that they must be able to engage interactively in real time with consumers using "omni-channel" communications—reaching current and potential customers through mobile phones, social media, in stores and through digital TVs. The flow of personal and other data coming from these devices is being combined with other information—about our neighborhoods, race, ethnicity, buying habits, social relationships, and more to create detailed profiles and predictions about us and our communities. Increasingly, we are being placed under a powerful "Big Data" lens, through which, without meaningful transparency or control, decisions about our financial futures are being decided.

As an official of FICO (the company best known for its consumer credit scoring services) recently explained, "Companies will decide how to converse with their customers based on a deep and timely analysis of each customer's context, behavior and history… ." We are on the cusp of a major shift in how enterprises formulate and manage their interactions with customers."[8]

Financial services companies are leading the adoption of Big-Data techniques, covering such

activities as sales and marketing, risk management, and new product development. U.S. banks were predicted to spend $41.5 billion on technology in 2013, with JPMorgan Chase, Bank of America, Citigroup, and Wells Fargo said to spend "$7 billion to $10 billion annually" alone.[9] In 2012, the financial services sector spent nearly $4.75 billion for digital marketing to take advantage of how mobile phones and the Internet have changed how consumers interact with banks, loan, and credit card companies.[10] An executive at Fidelity Investments candidly noted that "we've seen a proliferation of data that gives marketers the ability to target consumers more precisely … . [W]e're at a new golden age of marketing [and] have more tools at our disposal than we've ever had before."[11]

TAKEAWAY:

The financial system is at a critical transition period, as it repositions itself to take advantage of the changes made possible by advances in data processing and the growing consumer use of online media, especially mobile phones. Advocates and others concerned about poverty in the U.S. should take advantage of this unique window of opportunity to ensure that the interests of both the poor and those with low incomes are meaningfully addressed.

2. The Role of Unbanked and Underbanked Consumers in the Digital Marketplace

The consequences of the financial meltdown, a low-wage-centered economy, historic barriers to equitable access to education and opportunity, and the too-long-ignored one-in-six Americans who live below the poverty line, have placed tens of millions at the margins—or out of reach—of the financial system.[12] Nearly 30 percent of U.S. households are considered "either unbanked or underbanked" and conduct "some or all of their financial transactions outside of the mainstream banking system," according to a 2012 FDIC report. Seventeen million adults live in unbanked households (8.2 percent of all U.S. households), while 51 million adults—20 percent of U.S. homes—are considered underbanked. Minorities (except Asian), unemployed, younger, and low-income households are groups with the "highest unbanked and underbanked rates," according to FDIC's survey.[13] Sadly, more than 48 million children under 18 today live in low-income or poor families across the ethnic/racial and geographic spectrum, with higher percentages for children from African-American, Hispanic, Native American and other (non-white or -Asian) households.

Underbanked and unbanked consumers rely on increasingly popular prepaid debit cards, as well as products from alternative financial service (AFS) providers (non-banks), for services such as check cashing, payday loans, rent-to-own, and refund-anticipation loans. Today, debit cards and the AFS marketplace are rapidly adopting mobile and online-connected products and services. As the Center for Financial Services Innovation (CFSI) and others have identified, providing services for economically at-risk consumers is an important way for banks and other credit providers to generate new business, delivering significant profit and growth. According to CFSI, underbanked Americans spent $78 billion in 2011 in fees and interest for financial services. "This revenue was generated from an overall market volume of $682 billion in principal loaned, funds transacted, deposits held, and other financial services rendered," it explained.[14] The revenue potential of the underbanked is just one of the emerging markets now eyed by the financial services sector—which also knows that the rewards of serving multicultural consumers (especially Hispanics) and the growing mobile commerce marketplace will be critical for their near-term success.

Seeking to expand into these valuable sectors, as well as serve current customers and build new businesses, established and venture-backed start-up companies are competing to offer new services for credit and loans, banking, and payment. These include debit cards, lending services, mobile payment products, and other online-consumer-friendly tools enabling greater personal control over one's finances. Companies not traditionally identified with the financial services industry—such as Walmart, Google, Verizon, and scores of others—now provide a growing range of financial services and products.

TAKEAWAY:

New opportunities to serve underbanked consumers better may be possible by taking advantage of recent entrants in the financial services market. Telephone, computer, and other tech-based companies may have different policy priorities and market goals than the traditional market participants.

3. We Face a Mobile Moment of Opportunity for Reform

The impact of the mobile phone as an online device is at the core of the fundamental changes that are restructuring how we communicate and engage with both the marketplace and society at large. As we will discuss, the mobile device will be the key to accessing financial services and much more—to pay for goods and services and participate in everyday life. Financial companies are currently making a "massive investment" in their

ability to use mobile devices to reach and serve consumers. U.S. financial services companies comprise "the second highest spender in paid online and mobile media" marketing.[15] In one 2013 survey of the financial industry, "mobile banking" topped the list of the most important products to promote (followed by auto and mortgage loans). More than two-thirds of respondents said that "online advertising and social media" were going to be the "most important" media channels to use, with three-quarters of retail financial institutions involved with at least one social media service (and with "nearly one out of every four financial institutions is on Facebook"). Almost 60 percent said that merging offline data for online targeting was going to be very important as well.

The focus on mobile and online banking is partly being driven by demographic trends, especially younger consumers who are more comfortable conducting their financial activities online, without the need to walk into a bank. More than half of all bank transactions are now made online (although we still heavily use ATM's and other "physical" banking facilities).[16] The growth of mobile financial services also intersects with other developments. In the aftermath of the economic crisis, "households have adopted new financial and decision criteria to determine their lifestyles and credit behavior," explained Equifax.[17] Financial institutions are keenly aware that there is a growing number of dissatisfied customers who can easily switch to other companies, especially those offering digital access. Companies such as Walmart vie to serve what they call the "unhappily banked," consumers not content with the services offered by their current institution, and who seek alternatives that provide them with better financial control.[18]. There is also growing interest by consumers in having accessible, easy-to-use, and informative means of controlling their finances more closely. At-risk consumers can now access services that until recently would have been solely the preserve of the elite. For example, they can now use mobile phones to pay bills electronically, deposit a check without a visit to the bank (using the smartphone camera), receive a balance warning by text message, and add ("reload" or "top off") money to debit cards at thousands of convenient locations.[19] The availability of mobile payment technologies provided by PayPal, Square, and others enables new and community businesses to collect funds from payment cards without having to make a major investment.[20]

A wave of change in how we bank and buy will be a hallmark of what's called mobile payments—a swipe of a card (or merely carrying your mobile phone "wallet") will pay for services and enable banking almost anywhere and anytime. All consumers will require some form of mobile-enabled payment card or device to maneuver through this new landscape in the very near future. The combination of all these developments—mobile payments, new competition and services, digital media behaviors—is understood as a unique moment for the industry. "We've never seen a market opportunity of this size and magnitude … ," explained one venture investor. "There is a massive restructuring taking place."[21]

TAKEAWAY:

As credit is extended to more Americans, their use of new forms of mobile and online banking will also likely grow. We believe that economically vulnerable Americans form the basis of a very important—and potentially influential—constituency that should play a proactive role ensuring that this new marketplace evolves fairly and with the goal of promoting asset building. Through a variety of coalition outreach efforts focused on industry and policy leaders, a set of best practices and rules should emerge that directly benefits low-and moderate income Americans.

4. It is an Opportunity, however, with Looming Risks

However, while the new financial marketplace has the potential to provide greater economic opportunity, it also poses unprecedented challenges and risks for economically vulnerable Americans. It is possible that the shift to the data- and digitally driven financial marketplace could make life harder. For example, the growing use of highly sophisticated and powerful online techniques that enable the "micro-targeting" of vulnerable consumers to apply for high-interest payday loans is just one example. So is the expansion of so-called "e-scores"—a form of invisible (to the consumer) online ratings—that can help determine our credit worthiness, "lifetime value," or even the prices we pay. These e-scores can be used to blacklist or engage in discriminatory practices against individuals or even groups of consumers.

New economic stress may also be placed on economically vulnerable individuals and families, if financial marketers use their data-driven capabilities to focus primarily on more affluent or financially rewarding consumers. How much information about one's financial behavior, race/ethnicity, or health concerns can or should be used in making decisions on credit, now that advances in data processing enable a person's actions and behavior to be tracked and analyzed online and off? In a world where there is the ability to reach and engage the desired individual with growing precision and cost-effectiveness, what are the economic consequences for those citizens and consumers who do not offer the desired financial returns? Another looming problem is that consumers will be continually (and creatively) urged to spend, as alluring real-time offers—increasingly tailored to their interests and behavior because of data profiling—appear on their mobile phones and on social media sites. Marketing, sales, and payment will all seamlessly converge on the mobile device, creating endless opportunities for marketers to convince us to buy products when we are most vulnerable. The forces shaping personalized, data-driven commerce could undermine the financial security of all Americans, but especially those living paycheck to paycheck and on tight family budgets.

In addition, while the rise of "alternative" data is helping make more credit available to the underserved, there are important questions that should be raised about the reliability, fairness, and propriety of using, for example, social media, utility bills, and other records as sources of information.

Typical of what is emerging in hyper-local consumer targeting is the work of one company that is now mapping "data from multiple sources onto a grid of tiles that cover every square foot of the U.S. Each tile is 100 meters by 100 meters, and we inject third-party demographic information about the tile, as well as data on what's physically located there— … retailers and so forth. Then, we connect that data with where a mobile device is in real-time, or where it has recently been … ."[22] The growing capabilities to analyze and make non-transparent or accountable decisions about people and their "micro-neighborhoods" could usher in new forms of digital redlining for all kinds of services.[23]

New forms of discrimination may emerge as financial (and other) marketers deploy geospatial mapping software, tied to demographic, financial, and other databases to closely identify and classify the behaviors of individuals residing in a distinct neighborhood or "micro-community." Communities across the nation are subject to intense scrutiny as "location-centric data science" closely maps and assesses who we are and what we do in very narrowly defined geographic areas. Geo-mapping can identify locational differences, classifying neighborhoods by "social grade," "buying habits," and "blue vs. white collar."[24] Marketers are able, for example, to "learn" about "where users go and how often; when they go and how long they stay there; perhaps as importantly, where they *don't* go."[25] Such analysis can be used to make decisions about investing in some communities and bypassing others, or to take advantage of consumer vulnerabilities that may be harmful (such as the reliance on fast food by a neighborhood's families.[26]

TAKEAWAY:

While there has been understandable enthusiasm and support to help promote new financial products and services to low-income Americans, there are critical concerns that must be addressed. Technology and the use of data can play both a positive and negative role in our society. It is essential to identify and address the obstacles now present or emerging from the marketplace that pose a risk to already economically vulnerable consumers. The growing role of geo-spatial community analysis needs to be reviewed to prevent possible discriminatory practices.

5. Multicultural Communities are at the Epicenter of the Digital Marketplace

Over the last several years, a robust system to identify and target Hispanics and African Americans online has emerged.[27] Marketers are especially aware that Hispanics, African Americans, and Asian Americans are early and enthusiastic adopters of mobile phone use, social media, and online video services.[28] The development of data-driven profiles that are built around taking advantage of consumers' race or ethnicity—whether to sell insurance, a credit card, or fast food—can be used to steer individuals toward making good or poor decisions. The use of racial and ethnic data for financial marketing, including identifying an individual's "language," "assimilation," and "Hispanic country of origin," which can be combined with income, religion, use of a debit card and more, raises important questions about ensuring that historic discriminatory practices are not tolerated in the new financial marketplace.[29]

More than 63 percent of African Americans, 60 percent of Hispanics, and 62 percent of Asian Americans were predicted to own smartphones in 2013, outpacing the white population (at 54 percent). Hispanics and African Americans spend more time on the "mobile Web" and also with apps. Three-quarters of African-American and 68 percent of Hispanic cell phone users "go online" from their phone. More than half of all Americans with incomes below $30,000 a year and who have mobile phones rely on them to connect to the Internet (as do 60 percent of those earning between $30,000 and $50,000 a year).[30] These groups are the subject of intense research and analysis on their buying and financial behaviors by marketers that also incorporate cultural analyses into their targeting plans. One study of Hispanics and their use of mobile phones explained, for example, that "For Hispanic users the Web is more organically integrated into their lives. It's on the go, right now, access to their friends, family and information."[31]

America's multicultural communities include a much higher percentage of youth and young adults than the general population.[32] Younger multicultural consumers are highly prized by marketers because they are considered key "influencers" helping to define trends for the wider culture.[33] Consequently, these "digital natives" are often the focus of intensive campaigns designed to get them involved with brands and products.[34] Food and beverage, retail, and gaming advertisers target multicultural teens (and their younger siblings) using an array of highly sophisticated digital tracking techniques, especially on social media (such as Facebook) and on mobile phones. Youth of color are the focus, for example, of more junk food ads than other groups.[35] To take better advantage of how these and other young people are online today, fast food companies are quickly building new wireless payment systems that will enable a teen to order and pay for a meal instantly using a mobile app.[36] By purposely tapping into the developmental and emotional vulnerabilities of young people

to foster spontaneous decision-making about buying products, marketers place both their physical well-being (such as from obesity and diabetes) and their families' financial resources at greater risk.

The commercial digital media culture plays a powerful role in helping shape the identity and behaviors of youth from all backgrounds. These young people are enthusiastic participants in and creators of the digital culture, helping develop this marketplace (such as through the growing practice of watching TV and being online simultaneously).[37] However, they are being deliberately socialized through a range of digital marketing practices to embrace brands, engage in impulse buying, and care less about protecting their privacy.

TAKEAWAY:

As a key target for digital marketing, multicultural youth will be especially at risk. Not only will digital marketing have an impact on the health and well-being of children and teens, it will likely cause new strains on the emotional relations and budgets of their families. Concerning privacy, although children have some online marketing safeguards, teens are largely unprotected. Adolescents are subject to a growing onslaught of marketing on social media and on mobile devices. Encouraging a broad range of stakeholders to address unfair digital marketing practices targeting multicultural youth should be a priority.

6. The Data-Driven Financial Marketplace Focuses on Today's "Connected Consumer"

The explosion of consumer and transaction data, along with our computer device-driven lifestyle, and the growing capabilities of marketers to analyze and use those data effectively, have combined to become the driving force behind the emergence of "Big Data" in our lives.[38] Financial services companies are investing in an array of data management and other customer relationship management platforms to take advantage of the "pools of data that used to be unreachable."[39] Reflecting the volume, velocity, and variety of data associated with the current Big Data era, "2.8 trillion gigabytes [of information] were created, replicated or consumed in 2012."[40]

Financial services companies now engage in "Precision Marketing" using fine-tuned "customer segmentation" techniques; incorporate "time to event" and other "predictive" models designed to "optimize" the marketing process; and create "micro segments"—data that can all be used in the consumer credit review process. They use "text mining" and "semantic" analytics software to "discover patterns and hidden value" in what consumers say online (such as with social media).[41] Companies are working to connect all their information that was previously "silo-ed," so their various business divisions can access "customer-level data" and use that information to make decisions on credit, collections, fraud, and marketing. Prof. Robert Stine, who teaches statistics at the Wharton School, University of Pennsylvania, and who researches credit scoring, observed that "we're seeing a new leap in the kind of [accessible] data and the technology that is available to manipulate that data."[42]

For example, Capital One "continually seeks to refine its methods for segmenting credit card customers and for tailoring projects to individual risk files … ." The company "conducts more than 65,000 tests each year, experimenting with combinations of market segments and new products." It uses "transaction histories" that indicate a "customer's approximate annual income, spend-

There Is A Range of Tools At Marketers' Disposal

Financial marketers can mix and match a range of tools that identify, both online and offline, financially vulnerable consumers, such as the following:

- Nielsen's P$yche product uses "financial behaviors" that help make up its 58 "actionable" segments, which are "based on age, family structure, income and assets." One can identify consumers labeled "Payday Prospects," "ethnically mixed," and those who "often find themselves living paycheck to paycheck." Or a company could decide to market to, or ignore, "the most financially challenged segment. Those classified as "Bottom Line Blues" "have low educations and insecure jobs, surviving on cash instead of bank or insurance products."[55]
- Datalogix sells our "known financial behavior" so a consumer can be targeted online, including on Facebook. The data it uses cover 110 million U.S. households and are "verified by at least two 3rd party sources," collected from "credit header sources permissible for marketing use," "estimations from modeled credit data," and public records (deeds and the census). Consumers can be targeted based on their use of "credit cards, credit status, net worth, investments, household income" and use of "financial services" for banking and insurance. Datalogix identifies consumers' estimated "credit worthiness," listing whether they are "poor, fair, good, very good or excellent" prospects.[56]
- Data "co-ops" or exchanges buy and sell an array of consumer information that can be bought and sold. Through BlueKai's[57] Exchange, for example, companies can purchase information to identify consumers based on "estimated household income," "employment status," whether they are a "homeowner or renter," and if they have "propensities" for "personal finance" products."[58]
- Leading data company Alliant's information contains "detailed online and offline purchase transactions and payment histories on over 135 million consumers" that are "updated monthly." For online targeting, Alliant sells information on the "Financially Challenged" ("Payment Score: Bottom 50%, Bottom 20%, Bottom 5%, etc.), "Credit Card Rejects," "Credit Challenged," and "Risky Consumers."[59]
- In September, Acxiom introduced a new product so that "for the first time in history … marketers are able to fully leverage all kinds of data—first-party, transactional, digital, social, mobile and other audience information."[60]
- Mobile phones can be targeted with "true precision" by financial marketers using Acxiom, consumer, and purchase data, according to AdHaven Bullseye.[61]

ing habits, online usage patterns, and transaction types, along with how he or she typically makes payments … ."[43]

Harnessing its new data capabilities, one unnamed bank

created a system able to monitor the financial "lives" of customers—tracking deposits, withdrawals and other transactions. When an activity two standard deviations or more from the norm was detected, the system immediately alerted and briefed the customer service function. So, a real-time offer tailored to that change could be made at once. Using immediacy and familiarity—hallmarks of a social interaction—the bank created a 60% action rate, and a 38% conversion rate, by offering relevant contact at that moment of greatest opportunity. The same bank also explored location—by combining the when of customer interactions with the where … .[44]

This ability to analyze, predict, and immediately act on a consumer's data illustrates the changes helping transform the financial services industry. Companies in this sector now enjoy

"unprecedented levels of insight" to use in their consumer decision-making. As TSYS, a leading payment processor noted, mobile "technologies have greatly enhanced this data collection by giving organizations valuable information about individuals' transactions, preferences and online interactions." The harvesting of transaction data, according to TSYS, "will provide a more complete picture of cardholder behavior and, in turn, identify which cardholders are most profitable."[45]

Such comments reflect an intense interest by many in the industry to bypass individuals who have been identified as having a low or less profitable "lifetime value," and focus their best offers and services on the well-to-do.

TAKEAWAY:

In this era in which marketers know so much about individuals, and can reach them literally anytime and anywhere, critical questions about equity should be raised. Will the most attractive offers and opportunities for financial gain be offered only to the fortunate? Will already economically at-risk consumers be identified for their "value" to generate high fees and rates of interest, creating a new cycle that will create additional obstacles to their survival?

7. The Growth of the Consumer-Financial Data Complex

Financial information on a consumer has become a highly sought—and now daily sold—commodity. The amount of financial data that can be readily obtained today on an individual is staggering.[46] There is a literal explosion of firms, including data brokers, retailers, credit companies, and many more that vie to buy and sell information on a consumer in order to assess more accurately how to treat current and prospective customers. Credit bureaus and other data companies have established online products—overflowing with financial information—to complement their traditional offline services. Acxiom, Experian, and Equifax, for example, now have well-defined—and growing—digital divisions.[47]

In a 2012 presentation to its "Financial Services and Insurance" clients, Merkle described the need for companies to have a "single repository" that provides a "consolidated view of the consumer across all touchpoints." That includes capturing and understanding individuals' offline and online media use (including mobile, social, and print), "life events," demographics, and what "Life-Time Value" segment they are in. The ease of merging offline and online data (so-called data onboarding) has presented marketers with new opportunities to evaluate consumers.[48] "Not having the ability to link the digital address with their customer history means you could be missing revenue opportunities," explains Acxiom. Its "Single Customer View & Value" captures all the ways a consumer can interact with digital and in-store marketing, including the use of "bank data."[49] Real-time data collection practices bring together a diverse set of information, including what's called "first-party" and "third-party" information (data on an individual's transactions combined with demographic and online sources, respectively).

Consumers are largely unaware of the extent of data collected today on their activities, and how this is done. Marketing-automation software enables thousands of companies to have the capability to capture a consumer's "digital body language." They can collect data at each interaction, including "website visits, downloads, social networks, and searches," for example. They can know when a consumer has opened up an email and how a person has interacted on a site or series of websites. Companies can increasingly "monitor social conversations to … gain insight, dislikes, and perceptions and [then] drive [consumers] directly back into your social campaigns, social properties and communities."[50] Facebook and other social media are key platforms to promote products, including financial services. On Facebook, campaigns can be tailored to "regions, cities, zip codes, languages, brands, and products to gain complete control of customer targeting … [and] [d]rive traffic to spe-

Will Scores Be Used To Discriminate?

The role of scores as a potential discriminatory tool that can harm the interests of Americans seeking better financial opportunities is reflected in how some of these products are now being used, as the following examples suggest:

- Data and scoring company Alliant explains that its "ProfitSelect accesses the current transaction histories of over 130 million consumers" and allows companies to "cultivate the good, weed out the bad … ," "know who the slow or non-payers are, in advance …" and "identify the best customers early on and focus your best offers on them."[72]
- Scores from Netmining, which use "vast pools of data in real-time," measure the "value each individual is." Consumers are given "true-interest" scores, which dynamically change based on their actions.[73]
- Consumers are awarded "customer quality scores" using "predictive behavioral models [that] evaluate thousands of data features." These scores are used to "'tell apart' high quality visitors from the rest … . [E]ach of your visitors will see a unique section of products."[74]
- "P3 scores" that reflect "Personal, Purchase and Propensity" information on consumers, based on their spending and behaviors, are integrated into "300 million unique cookies" a month) used for online targeting. (Cookies are a form of online profiling software.)[75]
- Location, online behavior, and scoring are merged in Alliant's "real-time offer decisioning" "GeoPerformance" scores. "People tend to live near people like them. So if you know the area, you can predict the performance of the people who live there," Alliant explains. Data used for the score cover the behavioral waterfront: household income, recency of purchase, product preferences, detailed payment and transaction histories.[76]
- Mobile phone users are scored as well. Data-targeting company Dstillery says it can "score and rank the universe of mobile user events … through our observation of billions of user actions over time."[77]

cific offers or exclusive social deals," explains one leading data-oriented marketing company.[51] Some of the largest databrokers in the world are now part of Facebook's data targeting apparatus.[52]

These developments reflect the continuous monitoring and assessment of individual consumers, forms of what the *New York Times* has referred to as "commercial surveillance."[53] As a spokesperson for global advertising giant (and increasingly data-driven) WPP recently explained, "We're all moving to some point in the future where we can all monitor exposure at an individual or household level and that will all get fed into a data management platform." WPP coined the term "adaptive marketing" to describe new capabilities of using data for continuous targeting, using a consumer's "data exhaust" to help inform the next marketing cycle.[54]

TAKEAWAY:

Data collection practices that impact all consumers are growing, helping to fundamentally restructure how we buy and pay for products. Individual consumers can now be tracked by marketers and specifically targeted with personalized offers and even prices. This process could foster additional and unnecessary spending that may harm already fragile budgets.

Using such data products, underbanked consumers can be identified and targeted online for

payday loans, prepaid debit cards, money transfer, and similar services. For example, one data targeting company that focuses on the underbanked relies on a "proprietary formula of 127 predictors" based on the analysis of "1000s of raw data points per individual." This information comes from databases that contain "profile insights on almost every U.S. household and adult consumer," including "financial data, geo-location, purchase history, household data, [and] life stages." The same company also ranks U.S. neighborhoods on their "financial health"—the ability of their residents "to satisfy their existing financial obligations" (broken down using 9-digit Zip codes).[62]

The expansion of the consumer data financial profiling system, and its use in real-time decision-making on a wide range of products and services, require scrutiny on behalf of the underbanked and unbanked. Financial products and services promoted to the underbanked and unbanked should be reviewed for the role that data collection plays in their operations.

8. How Will Invisible Predictions Affect Our Financial Future?

Today, companies use decision-management systems to build a "sophisticated predictive model for every data mining function under the sun." One of the outcomes of this process is the growing array of so-called "e-scores."[63] These scores rate individual consumers based on a number of variables connected to their financial status and behaviors. Such identifiers can signal what companies believe consumers' "lifetime value" (LTV) to be, their "propensity" for purchasing goods, and how they should be treated in terms of offers and customer service.[64] The scoring function is incorporated in "decision management and prediction" software used by banks and others, capable of rating millions of customers in "minutes."[65]

Scores are more than just a more precise segmentation strategy. They can serve as a digital "scarlet letter" to convey a potentially negative assessment of individuals that can affect the services they are offered, whether or not they become a magnet for high-interest payday loans, or are given second-class customer treatment (made to wait longer on the phone for assistance, for example, as those with "better" e-scores are given priority). New forms of both overt and subtle discrimination, hidden from view, may be one of the consequences that e-scores have on economically vulnerable consumers.

In addition to scoring used to influence or determine our financial status, so-called "propensity" scores are sold to help marketers keenly understand consumers' potential interest in specific financial products or that they are likely associated with some negative event. Using data on our behavior, spending patterns, assets, what we have purchased previously, the media we prefer and more, propensity scores "provide rich insight" into how a consumer is likely to "respond, convert and remain loyal … ."[66] They can be used to identify customers who "are likely to spend more," or don't require (or need) significant "discounting." Banking, credit cards, insurance, retail and other markets buy these "propensity" scoring products. (Acxiom alone has "[t]housands of prebuilt, propensity model scores … available.")[67]

The expansion of data collection enables propensity and predictive modeling on a consumer to incorporate information connected to our "favorite ATMs close to work or home, favorite gas stations along a daily commute, preferred supermarkets and preferred online stores for shopping … [and even] our favorite cash withdrawal amounts."[68] It can also include, among the "decision model variables," data related to our "lifetime value," what we purchase, our "clickstream" activity, and how we have interacted with a company previously.[69]

For now, we can only surmise what the impact of secretive, data-driven scoring may be on financial opportunity. If a financial marketer identifiers consumers as having a "propensity" to buy more than they can afford, or in continual need of paycheck-advance loans, will this trigger a flood of payday loan ads on their mobile phones, as well as enticing digital discount coupons that promote over-spending?

Financial marketers are also interested in identifying individuals based on their "influencer"

potential on social media—whether what they say and do online can sway their friends and others to like or purchase. "Influencer scores" are being used based on the analysis of an individual's postings and relationships on Facebook and other social media.[70] Increasingly, the social media scoring models are also being used for financial decisions as well, by lending firms such as Lendup, a recipient of startup funding from Google Ventures, and Moven. While the U.S. Federal Trade Commission and Consumer Financial Protection Bureau regulate the sharing of such information by credit bureaus, its use by a firm to evaluate its own customers and potential customers is not regulated.[71]

For example, credit scoring company FICO has developed a "predictive scorecard" that analyzes the relationship of "social influencers."[78] Epsilon, another major provider of data and which works closely with Facebook, gathers "public" social media information as part of its consumer services, including "tweets, posts, comments, likes, shares, and recommendations," as well as "users' IDs, names, ages, genders, hometown locations, languages and numbers of social connections (e.g. friends or followers)."

Epsilon says that it "does not associate social media data with any other information stored in our databases." But it also says, on its financial services page, that "Understanding customers and when they are 'in market' for financial services and insurance is critical for today's marketer," raising questions about how such social data may ultimately be used.[79]

TAKEAWAY:

Questions should be raised about the overall role and use of scoring, especially for economically vulnerable Americans. In addition to possible discriminatory behavior, it will be essential to know what such scores mean in terms of additional unfair services offered to hard-pressed consumers.[80] One glaring problem is that unlike traditional credit bureau reports, which by law must give consumers free access once a year (as well as numerous other rights and protections), e-scores are unregulated practices. The public needs access to the profiles used to generate these e-scores, along with a public debate on their role in today's financial system including a review of what other consumer protections should apply to their use.

9. New Variables Used for Credit Scoring: Ubiquitous, Around-the-clock, Year-round Surveillance Tracking Systems

The emergence of these products raises other questions that reflect the directions of today's financial marketplace. Companies may wish to surveil underbanked consumers continually (in order to identify when they may be considered for approval). For example, Experian urges its clients to consider "radical processing." To capture these "new entrants" and "credit seekers" better, it recommends implementing a "continuous prospect monitoring process, using propensity scores, triggers and attributes … ." This involves using data on consumers about their use of credit and their "behavior patterns" for as long as the previous two years. This analysis can help financial companies "confidentially identify prospects within the near-prime segments who are trending upward and … make an offer to these receptive consumers."[81]

There are also emerging credit models that rely on a wider variety of consumer data for their decision-making. For example, Zest Finance says its mission is to make sure that people "being left out" of the credit system, even if they may have "bad credit," are considered for loans. To make such loans, Zest "analyzes thousands of potential credit variables—everything from financial information to technology usage—to better assess factors like the potential for fraud, the risk of default, and the viability of a long-term customer relationship … ."[82]

The influx of nontraditional data may help secure credit for consumers who do not have a traditional credit history, but questions on the scope and propriety of the data used must be raised. For example,

- AvantCredit promises to provide "immediacy" through the use of "machine learning" to determine lending risk, rather than relying solely on a credit score. Its platform "analyzes dozens of data sources while the customer is filling out an application, using an algorithm to find a customer's 'true' credit worthiness."[83]
- Credit Optics "supplemental score introduces a new dimension to the assessment of credit worthiness: Stability." The score "gauges risk by examining the velocity of account openings along with changes in the consumer's phone numbers, addresses and additional identifiers—all in real time."[84]
- Moven's CredScore product uses "a combination of financial wellness, social media metrics, transactional insight, and feedback loops to provide customers with the ability to understand their day-to-day financial behavior." Consumers are given their score "in real-time … . CRED is used as a transparent 'relationship' score—so we share your score in real-time" to understand "how that affects your monthly fees, other processing charges, interest rates on savings, availability to credit facilities."[85]

TAKEAWAY:
The availability of new sources of data used for credit review needs to be analyzed for their fairness, effectiveness, and relationship to loan products.

10. The Role of Online Lead Generation

E-scores play a role in another growing practice that impacts vulnerable consumers especially—online lead generation. "Lead gen," as it is called, is the practice of collecting and selling information about an individual as a "lead" who may be seeking a loan, credit card, or a product that requires a significant expenditure. Online lead generation was used as a technique to identify potential customers for subprime loans during the period that led to the financial meltdown.[86] Today it is a nearly $1.7 billion business in the U.S.[87] Websites that may offer loans or other financial products and services—even those that provide online calculators for mortgages, for example—are often only "lead generation" sites. They are designed to capture a

Search Engine Marketing

Financial services advertisers are also heavily involved with search engine marketing to win over potential customers and, by encouraging them to visit a website or fill out an information request, to collect data that can be used for lead generation.[90] Quicken Loans, which relies on the Internet for its marketing, uses nearly 47,000 keywords and spends anywhere from $120,000 to $198,000 *per day.* Lending Tree spends approximately $73,000 daily for its 41,000 search terms; Bankrate.com relies on more than 126,000 terms, spending anywhere from $32,000 to $48,000 daily. Major brick-and-mortar financial companies focus on search marketing as well. Bank of America, for example, uses some 63,000 search terms, spending $11,000 to $122,000 a day.[91] Google has paid close attention to how African Americans and Hispanics use search services to make buying decisions.[92]

person's information, such as street address and financial background, as well as gathering online data so individuals can later be targeted when they are on the Internet. Once a person's information is collected, it can be sold through what is known as a reverse auction. A person's data are auctioned off to the highest bidder—a lender, for example, who will pay the most for what's called a "hot" lead, because they have identified someone "in-market" for a loan. It's unlikely that consumers will get a good deal with reasonable rates once they become part of the lead generation system. Marketers employ an array of stealth tactics to collect information from consumers, including the use of the latest "Big Data" technologies. For example, consumers can be encouraged to fill out an online request for more information—or offered an online calculator to determine the cost of a loan. Consumers' activities can also be surreptitiously gathered online, as their actions are observed on numerous websites. Their data—whether personal information such as name and address (on a form), or cookies placed on their Web browser (in the case of the calculator or other online tracking tool)—are analyzed and scored to create what the industry calls "quality" or "hot" leads. Companies analyze the information to determine the identity and value of that prospect, which is then sold—often in real time—in a well developed lead marketplace.[88] Companies such as Lending Tree, Quicken, and Bankrate are leading sellers of such online leads, which are sold to companies or brokers seeking to sell payday, mortgage, and private school loans and similar products.[89]

A very "sophisticated network of high quality payday lead generation websites" thrives online today, including Spanish-language sites, that helps to sell these often unaffordable loan products.[93] Lead generation companies now use the latest state-of-the-art data-driven technologies to discover individuals who will be responsive to payday loan offers—including when consumers are using their mobile phones or on Facebook.[94] Companies offering loans or leads can feed data into superfast computers that identify individuals as prospects based on their behaviors, actions, and other variables. These consumers can be served an ad in milliseconds and can be followed wherever they go online.[95] Consumers are also unlikely to be aware that online lead generators conduct "testing" to help ensure that their websites trigger the responses they seek from largely unsuspecting consumers.[96]

Leads for loans and other financial products often come from online companies that most consumers believe are informational sites, but they make their revenues from collecting data and selling leads. Among online lead-generation company Datalot's clients are Bankrate.com, Efinancial, HomeAdvisor, and eHealth, Datalot operates its own lead-exchange system called "lead.io," which enables lead generation customers "to acquire and process consumers across multiple channels at scale … [and] provides real-time insight into lead value and traffic quality … ." In another illustration of how technology fused with data practices enables individuals to be assessed, Datalot says it "statistically determines a customer's value, and delivers only the most actionable, high-value prospects to the sales force from the sea of widely varying consumer interest … [using] proprietary tech-

nology to isolate and deliver actionable customer prospects.[97] It provides "one-stop" targeting for users of Facebook, Google, Yahoo, Twitter, and Microsoft's Bing, regardless of whether the consumer uses a desktop or mobile phone. Facebook, which opened its own ad exchange in 2012, is also working with online lead generation companies and payday-style lenders[98]

TAKEAWAY:

The role of the online lead generation industry and its impact on the underbanked and unbanked are not well understood. The industry's use of sophisticated data-driven online techniques enables payday lenders to reach vulnerable consumers regardless of location. Public education and safeguards are required to address these practices, which appear to facilitate the marketing of high-cost, under-regulated, non-transparent loan products, rather than to promote financial inclusion and opportunity.[99]

11. Alternative Credit Data Scoring

Lenders use credit-risk scores to determine the likelihood that a potential customer "will repay their various credit obligations."[100] FICO, a leading provider of these scores, notes that three-quarters of all mortgage loans are "underwritten" with its scores (and nearly "10 billion FICO scores" are used yearly). However, 64 million Americans "have little or no traditional credit history," according to Experian, creating a class of consumers considered not eligible for credit or forced to accept non-prime loan terms.[101] CFSI notes that "millions of Americans continue to go without access to affordable, high-quality credit products, in part, because they lack a long credit history or do not have a credit history at all. This quandary could be at least partially resolved by the use of alternative data."[102] Policy organizations including CFSI, the Political and Economic Research Council (PERC), and the Corporation for Enterprise Development (CFED) have all supported alternative data collection, including the use of so-called full-file utility reporting. [103]

Conversely, while underbanked and other at-risk consumers may benefit from these products, the data they rely on and how that information is subsequently used needs to be reviewed. The use of some non-traditional data has raised concerns from some financial consumer advocates, including the authoritative National Consumer Law Center (NCLC), which has questioned their reliability as meaningful predictors of credit worthiness and the appropriateness of their use. For example, NCLC has pointed out that for the lowest income at-risk consumers to qualify for winter energy relief programs such as LIHEAP in many states, they must first be delinquent in payments. NCLC has also pointed out that utility reporting is extremely inconsistent across sectors and across states.[104]

Recognizing that serving so-called "thin-file" consumers can be a good business, some companies have developed products that use non-traditional data to evaluate consumer applications. Equifax offers insight into how credit companies see the underbanked as critically new important markets, explaining that "Retail banking is undergoing change … making it harder to identify and capture potentially profitable households. … [T]raditional consumer risk assessment tools are limiting many financial institutions' success at the point of sale. They rely heavily on negative information that can be dated and unreliable. As a result, customers that could present significant revenue opportunity are walking out your door and customers that may ultimately charge-off are being approved."[105]

Examples of the data used by these new credit-scoring products (which can also contain a more traditional analysis) include the following:

- Equifax's "Insight Score for Retail Banking," which "leverages its data on 25 million so-called 'unscorable' customers with no traditional credit history," uses mortgage and loan repayments; income; wealth and assets; demo-

graphics; utility, pay-TV and telecommunications bill payments.[106]

- VantageScore (created by credit-reporting companies Equifax, Experian, and TransUnion), one of the leading providers of such scoring services and created by the Big Three credit bureaus to compete with FICO, adds information on a consumer's "rental, utility and cell" payments.[107]
- FICO's "Expansion Score" uses "aggregated data" (see section below on aggregated products) from such sources as "cell and landline telephone utility information, membership club records, [and] judgments." Made for the "credit-underserved market," the score is aimed at "recent immigrants, young adults, recently widowed or divorced and mature cash-spenders."[108]
- CoreLogic's products help to "identify previously hidden risks and new lending opportunities." It explains that "property, landlord/tenant credit and public record data elements represent unique insight into borrower debt and assets." Its so-called "FCRA-compliant" data used for the score include renter lease applications, collections, court records (failure to pay, judgments for rent, eviction writs or warrants), property transactions (liens, property tax amount), alternative credit information (online and storefront cash-advance lending, installment lending, rent-to-purchase inquiries), and borrower-specific public records (judgments for money—child support, deficiency judgment, tax liens, bankruptcies).[109]

Today, then, there is a robust debate among how such data, such as utility bills, should be used in applications to help predict creditworthiness.[110] Looming as a key issue as well is the growing availability of social media and new sources of other financially related digital data.

TAKEAWAY:

The range of information that can be collected and analyzed for consumer credit decision-making will continue to grow. While alternative credit scoring can be a boon for the underbanked, there need to be standards and safeguards to ensure that any new data are used appropriately.

12. Prescreening, Scoring, and the Fair Credit Reporting Act

Banks and other lenders have access to scoring products that enable them to identify, analyze, and then target a prospective consumer more precisely. To the extent that such modeling data are used to "prescreen" an individual as part of the "firm offer of credit" process, the Fair Credit Reporting Act (FCRA) regulates these practices and provides consumers with strict consumer protections, including the explicit right to opt-out of having their credit files used for prescreen marketing.

However, marketers claim that much of the financial and other data they use to make decisions on whom to target fall outside of the FCRA rules—since such data are tied only to advertising to promote interest in a brand or product, not, purportedly, to credit decisions. While the Fair Credit Reporting Act restricts the use of financial (including mode of living) data in credit reports to credit or insurance marketing purposes only (not general target marketing), the firms claim that they are not using such data to make financial offers, only to build audiences. They also claim that the files developed are not on individual consumers, but on clusters of consumers. Not subject to FCRA regulation, they assert, are scores and other products that identify consumers on an aggregate basis—which for them means information narrowed to a small cluster of households at the ZIP+4 level.[111]

However, given the capabilities of the contemporary data-driven consumer landscape, an ar-

ray of detailed information can be used to create a consumer profile and then deliver a "micro-targeted" ad or marketing message designed to initiate a process leading to a transaction (such as the sale of a financial product). As ads for credit cards and loan products are delivered directly to consumers on their computers and mobile phones, and are based on data that have analyzed a consumer's behavior, history, and financial transactions, should these practices not be considered a prescreened offer under the FCRA? What criteria are used to perform the prequalification assessment, and do they or should they trigger the FCRA?

On the one hand, data companies are fairly candid about the capabilities of the marketplace to identify a consumer for a specific product. Experian, for example, recommends to clients that they use an "online acquisitions strategy" involving the "prequalifying [of] consumers online to manage risks of prospects being evaluated for underwriting." Credit marketers "who choose to expand into the online channels have integrated tools that assess a prospect's risk prior to the application process," it explains.[112] Financial services companies like Equifax claim that such "aggregated" scoring products—a "micro-neighborhooded form of the FICO Score to enhance marketing applications" that the company claims is not linked to a specific individual—are permissible under the FCRA.[113]

But an examination of the composition and intent of aggregated products raises questions about the role of these scores, and the need for new safeguards. For example, these unregulated products provide financial companies information on "capacity to pay, financial stress, financial activity," as well as whether a household in a "micro-neighborhood" will "file for bankruptcy," "be looking to purchase a new automobile and looking for financing," or be a credit borrower that it has somehow determined is "likely to become a liability in the near future."[114] Among the 390 metrics available in the CreditStyles Pro product offered by Equifax is its "3.0 Neighborhood Risk Score," which identifies whether there's a "likelihood a household in a particular ZIP+4 will file for bankruptcy."[115]

There are also "aggregated" FICO scores for "marketing applications" that are used to predict "the likelihood that an existing account or potential credit customer will become a serious credit risk within 24 months." The score "identifies and projects the full range of credit risks" for a wide range of financial products, including auto loans,

Customers are leaving pieces of information at numerous touch points that are like bread crumbs and marketers are struggling to make sense of them," the CEO of one onboarding company explained."

bankcards, paying rent or mortgage and more. These data profiles are also considered outside the scope of the FCRA.[116] The explanation of how aggregated scores can be used online also suggests that their intended use is significantly to focus on—if not single out—an individual consumer. As explained by IXI (a division of Equifax), it can now "differentiate visitors in real-time" to "reach more visitors with the desired standard profile and propensities for products and services." It can "serve the right offer with the right message and creative [the ad or marketing content] based on visitors' likely financial position and purchase tendencies."[117]

One can digitally target consumers by "wealth, income, spending and ability to pay, by financial and economic behaviors," as well as behaviors related to products such as mobile phones, cable TV subscriptions, retail shopping, and "travel, leisure and entertainment."[118] IXI provides the following products that it says fall outside of the FCRA:

- Income360 Digital: "a powerful estimate of your prospects' and customers' total household income";
- DS$ Digital: "an estimate of a household's spending after accounting for fixed expenses of life—housing, utilities, public transportation";
- Ability to Pay Digital: "ranks online consumers based on their expected ability to pay their financial obligations";
- Financial Cohorts Digital: "data involving consumer assets, income, spending, and likely availability of credit." Companies can provide "premium offers to visitors likely to have significant financial potential and save lower value offers for others."[119]

Data targeting that enables the incorporation of a consumer's offline and online information—what's called "onboarding"—are also claimed to be "FCRA compliant" (an Orwellian construction meaning these uses of information are, in fact, outside FCRA regulation). "Customers are leaving pieces of information at numerous touch points that are like bread crumbs and marketers are struggling to make sense of them," the CEO of one onboarding company explained. The company takes these "bread crumbs" and merges them with "IP addresses, email addresses and zip codes," which are then "matched with more than 500 data points from approximately 250 million US consumers … ."[120] Yet this practice enables the more precise identification of individuals by connecting their online and offline identity information.

TAKEAWAY:

Credit bureaus and financial services companies are compiling expanding datasets on consumers to make decisions about their financial prospects. Yet the firms claim much of this information isn't personal to an individual consumer nor used for a transactional decision and therefore falls outside current federal FCRA rules. A review of the role of aggregated scoring and other services purported to be "FCRA compliant" is urgently needed to ensure that economically vulnerable consumers are not being unfairly treated by these practics.

II. The Underbanked in the Emerging Mobile Financial Marketplace

1. Prepaid Cards, Mobile Payments, and Digital Wallets

The financial marketplace is in a period of accelerated change, as prepaid cards are tied to online services (such as mobile apps); as the mobile payment becomes a primary way to interact for banking and shopping; and as the mobile phone "morphs" into a credit/debit card (or personal banker and shopping assistant on the go). As the Federal Reserve Bank of Boston stated last May, the "rapid growth in use of smartphones and mobile apps," the role of "non-banks" (including PayPal and Google) offering financial products, and the "convergence of online, mobile and POS (point of sale) channels" is helping drive the growth of the mobile payments marketplace.[121] The underbanked and unbanked will be directly affected by these developments. On the one hand, if companies and government develop safeguards and fair services, they can be new opportunities to conserve resources and spend wisely. But they can also place new pressures on vulnerable consumers and families who will be deluged to spend more, as the data profiling enabled by these services create a steady stream of sophisticated and personalized pitches.

Today, prepaid debit cards, which enable a consumer to "load" money to pay for expenses, are a key financial instrument for underbanked and unbanked consumers. Prepaid cards are a growing part of the financial services industry, enabling consumers to gain access to forms of electronic payment without needing to qualify for credit. The 2012 FDIC Underbanked report says that one in ten households uses this service. "For those with little or no net worth, prepaid remains their primary and often only choice," explains a presentation by Acxiom. Half of prepaid consumers "are unlikely to deal with a traditional bank or credit union," and "are much more likely to be Hispanic or African American," it notes.[122]

Debit cards enable greater control over spending, as one knows exactly how much one has on the card. Economically vulnerable consumers prefer their ability to control spending through prepaid cards.[123] As Consumers Union notes, today a variety of prepaid cards are "mainstream financial products," regularly used by "[m]illions of Americans" with "their wages, government benefits payments, tax refunds and other income regularly loaded." In 2012, General Purpose Reloadable (GPR) and other prepaid cards "were used in

1.3 billion transactions totaling $77 billion," and were expected to grow to $167 billion in 2014."[124] There is a growing number of prepaid card providers, from traditional institutions to recent entrants, such as Walmart and PayPal.[125] However, as the National Consumer Law Center has noted, GPR cards have weaker consumer protections than payroll or government benefit cards or debit cards linked to bank accounts.[126]

The greatest concern of most consumer advocates is that prepaid cards have served as a safe harbor for vulnerable consumers from both high-cost payday loans and bank account overdraft fees. However, cards are beginning to emerge with payday loan and overdraft features that impose new fees, thus eroding the potential to be a positive tool. As prepaid cards connect to the Internet, enabling real-time marketing of additional services, other new fees may be imposed as well.[127] If prepaid cards are to serve as a cost-effective and fair foundation for underbanked and unbanked consumers, they need to maintain low fees and transparent practices during this transition period.

TAKEAWAY:

The prepaid card market has become a cornerstone that provides financial services to the underbanked. As additional features are added to prepaid cards, new fees and features could be imposed that lead to unanticipated expenses for vulnerable consumers. In particular, overdraft fees or payday loan-type features could override the benefits of prepaid cards for vulnerable populations.

2. Protecting Vulnerable Consumers in the Smartphone-Connected Prepaid Card Market

The integration of prepaid cards with mobile phones and apps enables underbanked consumers to access features that can provide more effective control over their resources—such as record-keeping tools, remote check deposit, and opportunities for discounts and rewards.[128] For example, when one is "linked to a PayPal account, and uses its prepaid card, it unlocks services such as online payback rewards, an optional savings account, immediate account alerts, and online budgeting tools.[129] But increased online access also opens up a digital Pandora's box, filled with data collection, new opportunities for user financial profiling, and continuous "calls to action" to spend money.

Financial and other marketers have conducted numerous research studies to determine how best to use mobile phones to sell products. One striking finding is that most individuals view their mobile device as a part of themselves—not some distant digital tool. Google recently commissioned anthropological research on consumer attitudes towards their phones, and found that most consumers would give up chocolate or beer before surrendering their mobile phone.[130]

New ways to get further in debt, however, may also be one of the hallmarks of this new marketplace. Through the collection of underbanked consumers' financial information, including what they spend their resources on and where, Internet-connected prepaid card operators can make hard-to-resist offers that appear precisely during the decision-making process. Real-time credit, with payday loan-type terms, will increasingly become part of the underbanked consumer's financial landscape. The merging of prepaid cards with online access will enable consumers to apply for credit as they consider (through online ads or the use of digital discount coupons in stores) buying products or services, even with critical expenses such

as medical bills.[131] During the most recent holiday season, for example, PayPal offered instant credit access by integrating its app with its "BillMeLater" service. As a PayPal official explained, "For the first time, in the app that launched today, credit is built directly into the app. You can apply for a line of credit from the mobile app and it is not like a three-day process, we will basically give you a decision and a line of credit while you are still in that app in a matter of a minutes."[132]

The mobile phone's role in promoting and processing credit also raises concerns. The mobile device's small screen, and its configuration to serve as a very effective digital "salesperson," restricts how much information can be delivered to a consumer. Currently, there are no rules in the "wild west" world of what's called m-commerce (mobile commerce). Will already vulnerable consumers shopping for their children during the holidays have the time or inclination to see how much interest will be charged, or the terms of service concerning data collection, by reviewing the digital fine print displayed on their mobile phone's small screen? Moreover, as data profiling drives such personalized credit offers in real time (butressed with the increased dimension of locational information), and as advertisers deliberately use messaging that triggers emotions rather than reason, will a consumer be capable of making the wisest decision?

TAKEAWAY:

The introduction of new credit and loan services that are integrated into mobile app-connected prepaid cards requires safeguards, including a set of best practices. Real-time loan offers, especially those tied to a consumer's data profile, raises new financial risks for the underbanked. Principles of transparency, disclosure, and fairness should be applied to this new feature, as well as to the role of mobile devices providing financial-related services.

3. Mobile Payments and Mobile Wallet Markets are Growing Rapidly

The ability to pay for products through a swipe or use of a card or mobile device at numerous Point of Sale (POS) locations, and to engage in financial services nearly anywhere, is at the core of the mobile payment system. POS opportunities will abound, at the grocery shelf, in retail aisles, and for public services. Consumer interactions with financial services companies and their networks will become a routine daily feature.

At the moment there are competing technologies and companies all working to build out the mobile payment environment (with policymakers involved as well). Regardless of what standard or standards prevail, there will be an accelerated transition to POS and mobile payments that helps reconfigure consumer financial services as we now know them. Standalone prepaid and bankcards will eventually merge with smart phones—enabling that device to become the much-discussed mobile wallet. Leading credit card, phone, online marketing, and other companies are all working on mobile wallet initiatives. For example, the Google Wallet enables customers "to store their debit and credit cards" onto its platform and use it to pay for transactions at POS terminals (which use its preferred technology, called near field communications, or NFC).[133] The ISIS mobile wallet consortium, formed by AT&T, Verizon and T-Mobile, is developing its own technological standards for the mobile wallet.[134]

Companies are enthusiastically offering various mobile wallets, as they position themselves both to lead in this new area and also reap the financial benefits gained by offering an integrated set of services. MasterCard's MasterPass, for example, is designed to "unify" all of a user's transactions and provide a "a consistent experience whether the purchase is made at the cash register with a phone or credit card, online, or through a browser on the smartphone."[135] PreCash introduced a mobile wallet aimed at consumers without a bank account or credit card "that enables instant remote check deposits and bill pay from a smartphone." Bills can be paid to "utility, wireless, cable, Internet, auto loan" and other creditors. PreCash's

device can also be used to "top off" the amount of credit on the phones of their families or other contacts abroad.[136]

Various technological and standards groups are working to perfect the mobile payments and mobile wallets system. The Merchant Customer Exchange (MCX), created by Walmart, Target, 7-Eleven, and Best Buy, will enable thousands of stores to accept mobile payments in the U.S.[137]

The convenience of mobile payments and using one's phone (or yet unimagined device) will first be used as a supplement to but will eventually surpass the use of traditional credit and debit cards or online payment systems. But questions abound about what the costs will be. Will a series of daily transactions result in additional fees? Will consumer data be shared by the various partners working to develop these new services, creating new ways to profile and target vulnerable consumers?

TAKEAWAY:
The mobile phone will eventually become a leading—if not dominant—way we pay bills. This marketplace requires analysis and the establishment of safeguards on behalf of the underbanked and unbanked.

4. Mobile Apps and Wallets Pose Privacy Threats that Could Lead To Adverse Behavioral Targeting

While the growth of mobile payment services provides new opportunities and the capabilities to control one's finances, it also opens up the potential for further collection and use of consumer data. As legal scholars wrote, "Mobile payment technologies offer the ability to collect more information than before, and share it with different participants in transactions, providing an attractive service enhancement to both merchants and payment providers … ." They noted that more merchants will be able to "collect personally identifiable" information, and share it with financial services companies (including the companies that help process these transactions).[138]

These potential privacy risks are at odds with the often-voiced industry claim that users will have *greater* control over their information in this new era. The image of "empowered consumers" is often invoked by industry, despite research indicating that few consumers really understand how the data collection process and its various applications actually work.[139] Whether the general public, especially economically vulnerable consumers, will have any privacy or other new consumer rights is an open question at the moment. The mobile payment landscape should be encouraged to reduce fees, provide greater services, and ensure better protection of the information of all consumers.

TAKEAWAY:
Mobile phones pose a major privacy concern because these devices can collect both transactional (what you do and where you go online) and locational information. Consumer advocates need to be encouraged to focus on ways to protect the privacy of economically vulnerable consumers.

5. Loyalty Programs and Rewards Help Firms Collect Information

Leading financial companies are developing loyalty programs that take advantage of both the real-time accessibility of the individual consumer and also the information that can be gathered and monetized.[140] Underbanked consumers may find such programs especially attractive, believing that through earning "points" and other loyalty rewards they can build up resources for needed purchases. But these services can pose risks to their financial well-being. Some 400 financial institutions, including Bank of America, Regions, and PNC, work with loyalty program provider Cardlytics, including with debit, credit, and pre-paid cards. The company mines "detailed purchase data" to identify "an array of buying behaviors for millions of consumers," explaining that they "know what consumers buy—based on actual transactions." This information is made "actionable for marketers" and placed "in the online and mobile banking statements" of customers.[141] "When consumers log into their digital bank statements, they see advertising for products and services, chosen for them based on their recent purchases. They click to accept the offer, visit the store or website, and then use their debit or credit card to receive cash back from their bank."[142] While these rewards are attractive to consumers, few are likely to be aware of the data that are collected and how they may be used to make targeted financial offers.

TAKEAWAY:
Loyalty programs are being embedded into new banking services that take advantage of consumer data to make ongoing personalized offers. This raises both privacy concerns and the specter of another possible way the underbanked may be unfairly singled out to accept new forms of expensive loans and spend more of their limited resources.

III. Big Data and the Shopping Experience

1. SoLoMo and Other "Shopper Science" Technologies

The role that Big Data and new approaches to financial services play in the lives of vulnerable consumers is more than just access to credit and banking or the tools used to pay for products. It is reshaping the daily buying and shopping experience, changing over time what we may pay for appliances, clothing, and even groceries.[143] Much of the innovation in mobile payments, use of smart phones, and online marketing is being spurred by the economic rewards expected as technology fundamentally changes how we shop and pay for our purchases. Retail and grocery chains, online advertising powerhouses like Google, credit card, and phone companies are all actively participating in this transition. The field known as "shopper sciences" is working quickly to bring to local stores the ability to use data and mobile phones to drive sales. Intensively researched to advance the goal of a seamless and continual shopping "experience," the industry has developed a number of paradigms to describe the process—including "Path-to-Purchase, "Zero Moment of Truth," and "**SoLoMo** (combining social, location and mobile data and strategies)."[144] Through consumers' data profiles, which include online, in-store purchase, and financial data, "hyper-local" and personalized mobile targeted ads and e-discount coupons will be sent at the most effective time of day. These communications will be virally promoted by brands' social media messages on products, and payment or loyalty card smart "apps" that know when consumers are near or inside a store will urge them to buy.[145] Advances in data collection and analysis have enabled retailers to link in-store sales with targeted digital marketing, including on mobile phones.[146]

The proliferation of mobile phones will enable distinct marketing—and instant payment—pitches to be sent to multipleindividuals in a family, including children—all to reinforce a message. As a Mondelez (Kraft) executive recently observed, "The mobile phone is the one device that you have with you every second of the day … . Mobile is disrupting consumers' path-to-purchase as well as in-store experience, from the aisle to the register."[147]

TAKEAWAY:

Shopper science is significantly changing how we make buying decisions and interact with stores and services. It is part of the changes connected to mobile payments and online marketing that are ushering in the new financial landscape. A review of how changes in grocery store and other retail shopping from these technologies will affect economically vulnerable consumers is required.

2. You Don't Decide Anymore, They Decide for You

The same predictive modeling and segmentation information used to score consumers for financial services is being applied in the retail sector. When combined with the ability to reach a consumer in real time and at any location, the results can be that stores create "marketing offers so precise, so targeted, that customers think they were developed just for them," according to a KXEN-sponsored report. As one data technologist explained, "With behavioral profiling, companies can determine how much a consumer will pay for a product, and deliver coupons selectively so that each customer's discount reflects what they are willing to pay."[148] The key difference is that in the past customers decided whether or not to look for, collect, and use a coupon, while in the new model companies will determine who gets which coupons. More than 92 million Americans used a mobile coupon in 2012; e-coupons are expected to largely replace paper ones in the near future.[149]

Offers will also be based on our geography—where we live, the streets we cross, and places we visit. Geo-fences and other location-aware technologies are closely mapping and analyzing the individual and collective resources of ever-more discrete communities.[150] Hyper-local technologies can help companies analyze "existing customers" and also identify "people displaying similar behaviors and preferences."[151] This includes how we interact online and offline as well.[152] Illustrating again the cross-industry uses of data analytic technologies, credit-scoring company FICO is examining the use of locational information, explaining that the ability to access GPS data "provides a wealth of hidden predictive information about your customers' activity." Marketers can determine "the interaction between the path taken by the customer" and various community locations, providing a "powerful mechanism to influence a person's" behavior.[153] We do not yet know—but need to—what geo-related designations or inferences are being attributed both to consumers who struggle economically and to their physical environment.[154]

The "one-to-one marketing" model of delivering the "right ad to the right person at the right time," which is at the core of today's advertising-driven e-commerce system, will also begin to influence the prices a consumer may pay. Big Data technologies have helped create "analytic offer managers," which use "sophisticated time-to-event (TTE) scorecard models" based on the observed buying behavior for specified time frames," and which processes thousands of decision variables … ." The result, explains FICO, is a way to "execute targeted offers on a massive scale, in the context of real-time interactions."[155] The potential for new forms of price discrimination exists in this new digital marketing environment, as tools are made available that identify the "right or wrong price at the right time" for a single consumer.

How the data-driven shopping process ultimately influences consumers is still an open question. On the one hand, the Internet mobile phone allows consumers to check and compare prices more easily—what's now called "showrooming." Armed with more information or competitive offers, there's a good chance that a reasonable buying decision will be made.[156] But there is also a very real risk that an individual's ability to have the time and ability to make reasonable consumer decisions will be influenced—if not overwhelmed—by the powerful combination of marketing forces at work. Financially strapped and sensitive consumers could be harmed by these developments, if they are unfairly targeted for products they may not require or at prices they cannot afford or are higher than the prices offered others. There are consequences beyond busting the family budget as well, including to their health, as quick-service restaurants, food and beverage marketers, and even drug companies embrace the new digital model for marketing.[157]

TAKEAWAY:

New forms of community redlining and other discriminatory practices may emerge as marketers take advantage of their ability to "micro-target" individuals and their communities. An examination is required on the growing capabilities and interest of marketers to use personalized pricing for consumers, creating possible new forms of discrimination.

3. Financial Marketing on Social Media

A new frontier for data-driven financial marketing is on social media, especially Facebook and Twitter. Already, Visa, MasterCard, American Express, Chase, and Citibank are among the top-30 advertisers on Facebook.[158] Facebook and other social media sites provide new opportunities for financial services companies to engage in data mining, targeting, and influencing consumers and their networks of friends.[159] The social media structure is a complex, evolving, and purposely opaque system. But because companies such as Facebook require individuals to provide personal information, the amount of data that can be gathered and made actionable is significant.

Facebook itself is candid about its interest in working with banks, credit card companies, and others. As one trade article on a recent ABA presentation by Facebook's head of global marketing for financial services explained, "You don't have to tell Facebook what financial products this pool of people has or doesn't have—they don't care. All Facebook needs to know is that you've identified a type of consumer you'd like to focus on. Facebook uses your list to find users in its system attached to the email addresses and phone numbers you've supplied. Facebook can then build a profile of other users who match the 'digital accountholder' segment you've defined." Facebook says it does this with "astonishing precision."[160]

Financial services companies (like most others) are investing in what are called social commerce solutions.[161] Through these services, which help orchestrate complex social media marketing campaigns, companies can engage in "rich data capture," as Merkle describes it, on individuals and their friends. Other financial marketers are using Facebook for "new leads for their loan and refinance offers," involving "category, behavioral and email" targeting.[162]

TAKEAWAY:

Facebook and other social media are quickly becoming the new "public square," and will grow in importance as places of influence and where marketing and sales occur. These services are also successfully migrating to mobile devices as well. There are opportunities, however, to propose a set of best practices for the emerging social media industry and financial services, especially related to payday loans, lead generation and other products that impact underbanked and unbanked consumers.

4. Food, Beverage, Retail, or Bank Account: All Can Play

Financial, retail, food and beverage, and others are also using the same advanced data targeting structure to track and reach individuals online. As with online leads, financial firms, food companies, and other marketers can "buy" the right to deliver a very targeted message to an individual consumer. Through "ad exchanges" an individual with the desired online profile or record (such as financial behavior) is sold to the highest bidder in milliseconds. That message is then delivered to the individual's home or work computer, mobile phone, and—very soon—even their TV set.[163] Real-time bidding, using ad exchanges and other forms of "programmatic" buying, was predicted to generate $3.34 billion in 2013, comprising a fifth of online ad buying. It is expected to grow to $8.69 billion by 2017.[164]

The "Brave New World" of advertising is now being run by what are called "Math Men" (and women), not just copywriters. Although still largely out of public view, marketing is becoming more embedded into our everyday lives. It will be further integrated into all of our experiences, packaged as appealing entertainment, free services, and even "branded content" disguised as news. But the goal of such advertising will be to observe silently what we and our friends do—and use that information for what will be the lifelong profiling of individuals for commercial purposes.[165]

The intertwined forces of data collection and digital device adoption enable a "360-degree" targeting environment—"anytime and anywhere"—according to the industry refrain. Although there will be a focus on serving the well-to-do and middle class, it is likely everyone will be a target (since "influence" and positive word of mouth are desirable outcomes in addition to buying products in the new social commerce-oriented environment). In other words, economically vulnerable consumers, especially families and youth, will not likely find respite from the increasingly personalized and pervasive pitches that will continually reveal themselves when we shop, transfer funds, send money, or check our balances.

TAKEAWAY:

Technological advances that collect, analyze, and make actionable consumer data are now at the core of contemporary marketing. The public is largely unaware of these changes and there are few safeguards in this new marketplace. Economically vulnerable consumers, and especially youth, will be continually urged to spend their limited resources. Conversely, there are opportunities to use the same tools to urge consumers to budget, save and build assets.

IV: Where Do We Go from Here? Recommendations for Next Steps

The Need for Public Education, Transparency, and Advocacy

The next several years are a critical transition period to ensure that unbanked and underbanked Americans specifically benefit from the developments addressed in this report. We believe that the new financial marketplace can operate in a fair and equitable manner, helping to generate opportunities to promote economic security for individuals and communities. But to accomplish this, we should develop a proactive agenda that specifically identifies how the shift to digital and mobile financial services should be used to protect and serve the interests of America's economically vulnerable consumers. We should further examine the role that data analysis plays in the underbanked and unbanked financial marketplaces. Work should be done that identifies as early as possible best practices, potential harms, and areas requiring industry codes of conduct, public accountability, and regulation. A chief goal would be to nurture the positive potential of the digital financial system and reduce its negative consequences to individuals, families, and children as much as possible. To help create this agenda, an initiative should be created that reaches out to NGOs and coalitions already working on financial reform and economic inclusion and justice. Working with existing partnerships and forming new alliances as required, there should be outreach to other consumer, health, education and parent groups, industry, philanthropy, academia, and government.

A. Public Education

Few consumers—nor even many NGOs—understand the dimensions of the contemporary data-driven digital marketplace. As research shows, consumers are largely unaware of how their information is collected and used, as well as protected by regulatory safeguards.[166]

While there is extraordinary consumer acceptance of the role of mobile phones and digital marketing in their lives, despite concerns about privacy, much is not well publicly known about how data collection practices for the financial services industry actually works.

Consumers need a better understanding of the emerging landscape of data-driven financial services—how products such as mobile payments

and apps operate, and what the rewards and risks ultimately are. Clearly written and publicly accessible materials should be available that provide both the "big picture" of "Big Data" transformation as well as information on specific technologies, services, and issues (including privacy). Such bilingual information should be delivered using a range of media, from print and online (including social media) to mobile communications and perhaps even gaming. Materials should be conceived and then distributed through alliances with religious groups, economic justice advocates, civil rights organizations, and consumer and financial reform advocates.

B. Best Practices

This is a tumultuous period for the financial services sector. Traditional consumer expectations and relationships are changing, as competition further erodes the overwhelming dominance of the major financial brands. Companies are also stepping up the pace of innovation, introducing new products and services (especially for mobile). However, the forces of consolidation are at play as well, likely leaving fewer companies in the long term that rely on increasingly standardized practices for marketing, data practices, etc.[167] A "window of opportunity" is now open to help set standards for products serving the underbanked that help—not impair—their ability to conserve and grow their assets. An objective assessment needs to be conducted on the data-gathering and analytical products used by leading financial companies, to identify where disclosure, transparency, consumer control, and regulatory safeguards are warranted (such as with the use of social media data for financial decision making and whether new forms of digitally-based redlining are emerging).

Consumer groups, advocates for the underbanked, privacy organizations, and other experts have critical roles to play in this area. For each of the major categories of products and services focused on or potentially useful for asset building—such as prepaid cards, mobile payments, alternative scoring, data collection, and the role of social media—we propose the creation of small working groups. These individuals and organizations would be tasked with engaging in an intensive overview, including interviewing industry representatives, scholars, and government officials, and developing a set of "best practices" and potential self-regulatory (or policy-based) guidelines. For example, it is important now to identify and address the likelihood and impact of new or escalating fees imposed for services that are supposed to benefit the underbanked economically.

Through dialogue with industry and allies (such as CFSI and the Asset Funders Network), as well as with the FTC, CFPB, and others, and also through public outreach via the media, this effort would help set the parameters for what is preferred—and what is unacceptable—for products and services offered to underbanked Americans.[168]

C. Coalition Building and Cross-Fertilization of Ideas

The mobile phone is a key instrument for financial inclusion and communication, and will deliver many of the important services (through apps, mobile wallets, wireless payments, for example) The mobile industry is primarily dominated by a handful of companies, including the telephone industry, Apple, and Google. They are helping set the standards for the industry and also distributing much of its content.

A new coalition of consumer, civil rights, and anti-poverty organizations should be formed—or developed from existing alliances—that is specifically focused on the role that mobile devices play serving the underbanked. This group would work to support best practices, extend mobile Internet access to more low-income Americans, and engage with industry stakeholders. It would also work with other groups focused on issues at the FCC, CFPB, and elsewhere. For example, the FCC, which has regulatory oversight of the telephone network, should be encouraged to examine the impact of new forms of credit that will be placed on a consumer's telephone bill ("Direct carrier billing").

D. Industry Standards-building and Government Oversight

Rules for the new financial marketplace are principally being developed by industry. A number of forums or consortiums are known to be working on best practices, technical standards, and other issues that will affect the underserved. For example, in addition to the Merchant Customer Exchange (MCX) initiative discussed earlier, there is also the Mobey Forum North America, a "bank-led industry association driving the evolution of a sustainable and prosperous mobile financial services (MFS) ecosystem."[169] Its members include American Express, Bank of America, Capital One, CIBC, MasterCard, TD Bank Group, US Bank, and Visa, which gather together to deliver sustainable, long-term mobile services to the mass market.[170] There is also the new SmartCard Alliance that is working on near field communications and wireless payments.[171] The Federal Reserve Bank of Boston also has a "Mobile Payments Industry Workgroup."[172]

Industry trade associations working on issues related to the underbanked, including new entrants such as the Online Lenders Alliance, should also be tracked.[173] The concerns of underbanked consumers should be represented in these forums on an ongoing basis, or at least closely followed to ensure they play the most positive role.

E. Policy

While public education and industry engagement are essential, there should also be federal and state safeguards against unfair practices as well as policies that encourage asset-building, budgeting, financial inclusion and opportunity. Working with Americans for Financial Reform and others, the CFPB and FTC should be encouraged to review the products and services that comprise the contemporary underbanked financial marketplace. Rules governing data collection, profiling, and targeting of vulnerable consumers should be implemented. Groups should propose agency review of the FCRA to ensure it addresses these current practices. New safeguards to protect consumer privacy should also be recommended, including addressing the needs of youth. Groups should also develop a strategy to encourage the Federal Communications Commission (FCC) to use its authority to promote affordable access to mobile devices and develop consumer protection standards.

F. Conclusion

As we have explained, the U.S. is now in the initial phase of a significant transition period, as technology, financial products and services, and consumer behavior and expectations undergo significant change. We have no doubt that these changes will create on their own both new opportunities and pitfalls for economically challenged Americans. But we believe they have the most to gain—and lose—with the digitally connected and data-aware marketplace. Either they will have access to an array of products that operate fairly, that help them save and make the best decisions possible, or they will confront a marketplace that takes advantage of them and makes their lives harder. We know that not doing anything will bring us no closer to promoting economic justice and equity. But by taking advantage of the fundamental change that is upon us, we can help shape this shift to promote the interests of those who have a critical stake in the outcome.

APPENDIX: Walmart Positions Itself for the New Financial, E-commerce, and Shopping Marketplace

Walmart is keenly aware of the changes discussed in this report, and how they affect their customers, many of whom are in the underbanked category. The retail giant is significantly investing in financial products and services, big data analytics, mobile phone, e-commerce, and other digital applications. In 2012, it introduced its low-priced Bluebird debit card in partnership with American Express, proving a wide range of banking and credit services to its customers.[174] Knowing a majority of Walmart shoppers (55 percent) already come into the store with a smartphone, it introduced a mobile app. Its app users are some of the retailer's best customers—visiting the store twice as often as the average shopper, and also spending some 40 percent more in the process. Walmart's competitors—including Dollar General and Target—also now offer mobile apps.[175]

Walmart's smartphone "app leverages geolocation to detect when a consumer is nearby to a store and automatically prompts the user to flip the app into store mode—which lets consumers view maps of stores and find products within aisles."[176] There is a "shopping list feature that lets consumers scan in-store bar codes or add products to their shopping lists."[177] Local stores can promote specials or offers to their mobile app users. The app's "Scan & Go" feature allows customers to "clip coupons by tapping their smartphones and having the savings automatically applied." In some stores, the app can scan the products on shelves and counters, enabling a faster checkout. A recent update to the Walmart app enables users to register their phone number during checkout in order to receive automatic electronic receipts for future in-store purchases.[178]

Beyond convenience, Walmart's moves into mobile are also designed to help deliver revenues. Its app-delivered electronic coupons can be used to increase the "basket size from existing customers," according to Alex Campbell, chief innovation officer of mobile marketer Vibes ("basket" referring to total checkout expenditures). Campbell believes that the ability to access consumers through mobile phones "will allow Walmart to do some really dynamic pricing based on individual customers and buying behavior."[179]

Walmart's mobile app is an example of how the

company has entered the Big Data and e-commerce arms race, adding millions of data points daily for customers who no longer need to be within the confines of a Walmart in order to be touched by the retail giant. It has created a data-oriented research and product development facility in Silicon Valley—Walmart Labs. A recent job posting at the Labs makes clear the direction in which the company is headed: "We are building smart data systems that ingest, model and analyze the massive flow of data from online, social, mobile and offline commerce/user activity to set key business attributes for millions of products in real time."[180]

Illustrating the same omnichannel data collection, analysis, and targeting orientation used by the financial sector, Walmart analyzes a broad range of consumer information. As a Walmart Labs blog post explained, "The targeted team … ingests just about every clickable action on Walmart.com: what individuals buy online and in-stores, trends on Twitter, local weather deviations, and other local external events … . We capture these events and intelligently tease out meaningful patterns … . Our big data tools help us personalize the shopping experience and our psychological analysis helps us to dissect even deeper meaning behind the patterns in the data. We apply behavioral economics to find clarity behind both the rational and irrational behavior shoppers experience." Through its "targeting team comprised of PhD's in computer science, statistics, signal processing and behavioral psychologists," Walmart "has developed methodologies" that can identify "what customers might be seeking" as well as other products they may have a strong appeal for them (through an analysis of their "historical buying patterns," for example).[181]

Among the products developed at the Labs was its own Walmart search engine that is "sensitive to how their customers interact on social networks, so the products that show up are much more likely to be relevant and they claim something like a 10-15% higher conversion rate using their own, in-house search." For its e-commerce online site (with $10 billion in sales), Walmart enables its customers to "shop online and pay in-cash."[182]

The transformation of how we shop was one reason why Walmart bought Inkiru last June. That company developed a "real-time analytics solution" that "allows merchants to analyze and predict customer behaviors, improve active customer marketing, optimize for personalized customer engagements, and detect fraud during a live transaction, all prior to the transaction completion."[183]

Through its combination of powerful data analytics, mobile applications, personalized online and in-store targeting, and financial services, Walmart is playing a key role accelerating how the Big Data transformation will affect its underbanked household customer base.

End Notes

1 See, for example, the showcasing of new financial products at Finovate, "Payment Innovators Impress At FinovateFall 2013," Pymnts.com, 13 Sept. 2013, (viewed 10 Feb. 2014).

2 Sasha Abramsky, *The American Way of Poverty* (New York: Nation Books, 2013).

3 Abramsky, *The American Way of Poverty*.

4 Retail Payment Risks Forum, "The U.S. Regulatory Landscape for Mobile Payments," Federal Reserve Bank of Atlanta, ; Bill Siwicki, "10% of U.S. Consumers Access the Web Only by Smartphone," Internet Retailer, 5 Sept. 2013, . See also Kathryn Zickuhr and Aaron Smith, "Home Broadband 2013," Pew Internet and American Life Project, 26 Aug. 2013, ; Mark Hugo Lopez, Ana Gonzalez-Barrera, and Eileen Patten, "Closing the Digital Divide: Latinos and Technology Adoption," Pew Research Hispanic Trends Project, 7 Mar. 2013, (all viewed 10 Feb. 2014).

5 Board of Governors of the Federal Reserve System, "Federal Reserve Survey Provides Information on Mobile Financial Services," 27 Mar. 2013, (viewed 10 Feb. 2014).

6 Tom Standage, "Live and Unplugged: In 2013 the Internet Will Become a Mostly Mobile Medium. Who Will Be the Winners and Losers?" *The Economist*, 21 Nov. 2012, ; Google, "Mobile," Think Insights, (both viewed 10 Feb. 2014).

7 See, for example, a case study on online data provider eXelate, which explains that the company "processes 60 billion transactions a month for over 200 publishers and marketers … . eXelate is in the business of supporting its advertising customers by reducing Big Data to actionable smart data. It ingests huge amounts of click and other data from websites and other sources, builds models describing what that data means, and uses those models to populate databases that optimize bidding for advertising space in real time." David Floyer, "Real-time Big Data: eXelate Case Study," Wikibon, 17 Oct. 2013, (viewed 10 Feb. 2014).

8 FicoWorld, "The Era of Intimate Customer Decisioning Price Optimization Comes of Age—Dynamic and Real Time," personal copy on file with author Jeff Chester. Experian describes it this way: "The emergence of today's new hyper connected and 'always-on' consumer means brands must deliver coordinated and consistent customer experiences across all marketing channels." "Experian Launches Intuitive Cross-channel Marketing Platform in Southeast Asia," 8 Oct. 2013, http://finance.yahoo.com/news/experian-launches-intuitive-cross-channel-020000629.html (viewed 10 Feb. 2014).

9 Michael Hickins, "Banks Using Big Data to Discover 'New Silk Roads,'" CIO Journal, 6 Feb. 2013, http://blogs.wsj.com/cio/2013/02/06/banks-using-big-data-to-discover-new-silk-roads/ (viewed 10 Feb. 2014).

10 Interactive Advertising Bureau, "Internet Ad Revenues Again Hit Record-Breaking Double-Digit Annual Growth, Reaching Nearly $37 Billion, a 15% Increase Over 2011's Landmark Numbers," 16 Apr. 2013, http://www.iab.net/about_the_iab/recent_press_releases/press_release_archive/press_release/pr-041613/. The IAB defines this category as including "commercial banks, Styles agencies, personal credit institutions, consumer finance companies, loan companies, business credit institutions, and credit card agencies. Also includes companies engaged in the underwriting, purchase, sale, or brokerage of securities and other financial contracts." eMarketer, "Customer Transaction Info is Leading Big Data Element,"11 Oct. 2013, http://www.emarketer.com/Article/Customer-Transaction-Info-Leading-Big-Data-Element/1010291 (both viewed 10 Feb. 2014).

11 "An Interview with Jim Speros," eMarketer, 27 June 2013, http://www.emarketer.com/corporate/clients/fidelity (viewed 10 Feb. 2014).

12 Abramsky, *The American Way of Poverty*.

13 Federal Deposit Insurance Corporation, "FDIC Releases National Survey of Unbanked and Underbanked," 12 Sept. 2012, http://www.fdic.gov/news/news/press/2012/pr12105.html (viewed 10 Feb. 2014).

14 The report also identified products that "witnessed very high growth rates between 2009 and 2010" and that "internet-based payday lending" grew by 35 percent. Center for Financial Services Innovation, "Underbanked Consumer Financial Services Market Estimated at $78 Billion," 1 Nov. 2012, http://www.cfsinnovation.com/content/underbanked-consumer-financial-services-market-estimated-78-billion#sthash.yrLGx7Vj.dpuf (viewed 10 Feb. 2014).

15 eMarketer, "The Financial Services Industry Steadily Grows Digital Ad Spend," 2 July 2013, http://www.emarketer.com/Article/Financial-Services-Industry-Steadily-Grows-Digital-Ad-Spend/1010016 (viewed 10 Feb. 2014).

16 Robin Sidel, "Banks Make Smartphone Connection," *Wall Street Journal*, 12 Feb. 2013, http://online.wsj.com/news/articles/SB10001424127887323511804578298192585478794 (viewed 10 Feb. 2014); Victoria Petrock, "The US Financial Services Industry 2013: Digital Ad Spending Forecast And Key Trends," eMarketer, June 2013, personal copy on file with author Jeff Chester.

17 "CreditStyles Pro10 Feb. 2014

18 Walmart, for example, has a goal of serving the "unhappily banked" through its "Bluebird" prepaid card (co-branded by American Express). "AmEx and Wal-Mart Pursue the 'Unhappily Banked' with Their New Bluebird Prepaid Card," Digital Transactions, 8 Oct. 2012, http://www.digitaltransactions.net/news/story/3713. The growing popularity of these digital services with consumers, including "Direct" banks that operate online, is affecting the operation of local branches. Twenty-two hundred bank branches closed in 2012, a reflection of cost cutting and also, perhaps, the shift by consumers to online. There are significant reductions in the cost of serving consumers: an in-person transaction can cost $4.25, versus online at 19 cents each and only 10 cents when using a mobile device. Robin Sidel, "After Years of Growth, Banks Are Pruning Their Branches," *Wall Street Journal*, 31 Mar. 2013, http://online.wsj.com/news/articles/SB10001424127887323699704578326894146325274 (both viewed 10 Feb. 2014).

19 For example, Green Dot provides personalized prepaid cards connected to the Visa and MasterCard networks. Fees can range from free ATM withdrawals at Walmart and other locations to charges of nearly $5 to add money to the card. Green Dot, "About Green Dot and the Green Dot Card," https://www.greendot.com/greendot/help?from=landingpage#whohttps://www.greendot.com/greendot/help?from=landingpage. So-called "peer" lending services enable individuals to "invest" and make loans to borrowers. Such services can impose high fees. See LendingClub, "Rates & Fees," https://www.lendingclub.com/public/borrower-rates-and-fees.action; Prosper, "Personal Loan Rates and Fees," http://www.prosper.com/loans/rates-and-fees/; "What You Need to Know About Prepaid Cards," *Consumer Reports*," July 2013, http://www.consumerreports.org/cro/2013/07/prepaid-cards-fees/index.htm (all viewed 10 Feb. 2014).

20 A number of companies offer money-management tools, so consumers have more information on their spending. Moven wants consumers to "leverage the power of the smartphone as the primary payment device," and claims that its goal is to "provide financial wellness" for its customers. It monitors a consumer's "transactions made by Moven-branded debit cards in real-time, learning user spending trends and patterns," and reports back to the consumer. Jim Marcus, "Moven: From Mobile Banking to Mobile Money," Next Bank, 18 Feb. 2018, http://www.nextbank.org/mobile/moven-from-mobile-banking-to-mobile-money-2/. See also eMarketer, "Digital Banking Trends," Aug. 2013, personal copy on file with author Jeff Chester. Simple is another prepaid debit card that incorporates tools and information so consumers can have greater control of their finances. It now has 40,000 customers and processes more than $1 billion a year. Joshua Reich, "One Year with Our Customers," Simple Blog, 15 July 2013, https://www.simple.com/blog/one-year-with-simple/. See also Simple, "FAQ," https://www.simple.com/faq/. The Houston-based payments processor PreCash recently launched a mobile- and Web-based service that lets U.S. consumers pay bills without needing a bank account or debit card. PreCash, "Consumer Solutions," http://www.precash.com/consumer_financial_solutions.html. The Evolve Money platform offers the ability to pay bills from around 10,000 billers using an Android app or the Evolve Money website, which is optimized for mobile devices. Billers that accept these payments include companies such as Pacific Gas & Electric, Florida Power & Light, State Farm, Liberty Mutual, Time Warner Cable, and numerous others. Evolve Money, http://www.evolvemoney.com/; Flip, http://www.myflipmoney.com/; "PreCash Introduces the First Practical Mobile Wallet for Underbanked Consumers," 12 Sept. 2012, http://www.precash.com/flip_release_20120911.html. "For instant check deposits, PreCash plans to charge $1 plus 1% of the deposit for a payroll or government check plus $1 plus 3% of the amount for personal checks." Daniel Wolfe, "PreCash to Debut 'Flip,' a Mobile Wallet for the Underbanked," PaymentsSource, 10 Sept. 2012, http://www.paymentssource.com/news/precash-to-debut-flip-a-mobile-wallet-for-the-underbanked-3011817-1.html (all viewed 10 Feb. 2014).

21 Rebecca Grant, "AvantCredit Secures $20M to Make Financial Loans More Accessible with Big Data," Venture Beat, 14 Aug. 2013, http://venturebeat.com/2013/08/14/avantcredit-secures-20m-to-make-financial-loans-more-accessible-with-big-data/ (viewed 10 Feb. 2014).

22 EMarketer, "How to Use Location Data to Target Unique Mobile Audiences," 13 Sept. 2013, personal copy on file with author Jeff Chester.

23 Examples of forms of discrimination, including by race, have begun to emerge. See especially the work of Prof. Latanya Sweeney, "Discrimination in Online Ad Delivery," ACMQueue, 1 Mar. 2013, http://queue.acm.org/detail.cfm?id=2460278. For an example of a company deciding not to serve certain consumers because their profiling information suggests that they wouldn't generate sufficient profits as a customer, see the Orbitz example in Al Urbanski, "Making Data Big and Smart with Data Intelligence," Direct Marketing News, 1 Nov. 2012, http://www.dmnews.com/making-data-big-and-smart-with-data-intelligence/article/265534/ (both viewed 10 Feb. 2014).

24 "The [Zions] bank uses the maps to find areas to target, as well as to figure out which branches serve diverse markets. In some less affluent neighborhoods, for example, Zions provides a check-cashing product that was developed based on Geoscape data. 'We will offer our traditional banking products—checking, savings and traditional loans,' [Juancarlos] Judd, [senior vice president of Zions Bank] notes. 'But where there's a higher need for check cashing products, we offer that.' The software has helped the bank realize that texting, alerts and smartphone use is high within the Hispanic community; they're less apt to use online banking." Penny Crossman, "Zions Bank Combs Big Data for Customer Preference Clues," American Banker, 1 Oct. 2013, http://www.americanbanker.com/issues/178_190/zions-bank-combs-big-data-for-customer-preference-clues-1062531-1.html; NBCUniversal Integrated Media, "Personal Grid," The Curve, vol. 2, 2012, http://thecurvereport.com/category/trends/personal-grid-1/ (both viewed 10 Feb. 2014).

25 Mark Prior, "'Where You've Been' is Where It's At: Rethinking Geo-targeting," The Wall, 4 Oct. 2013, http://wallblog.co.uk/2013/10/04/where-youve-been-is-where-its-at-rethinking-geo-targeting/#ixzz2hKLElJqh. See also GFK, "Geomarketing: The Answers to Your Location Questions," http://www.gfk.com/solutions/geo-marketing/Pages/default.aspx; S2 Customer Insight, "Location Location Location," http://www.s2customerinsight.com/customer-engagement/#locationlocationlocation (all viewed 10 Feb. 2014).

26 S2 Customer Insight, "Our Work—Case Studies: Coca-Cola Enterprises," http://www.s2customerinsight.com/casestudies/fmcg-coke/ (viewed 10 Feb. 2014).

27 Luminar, "Luminar Insight APP: A Deeper Dive Into Understanding the Hispanic Consumer," http://luminarinsights.com/solutions/luminar-insight-app/. See also Briabe, "Advertising," http://www.briabemobile.com/Services/Advertising/; Briabe, "Case Studies: Hispanic Mobile Banking," http://www.briabemobile.com/Case-Studies/Case-Studies/Hispanic-Mobile-Banking (all viewed 10 Feb. 2014).

28 eMarketer, "Among Hispanics, Who's Leading Digital Adoption Trends?" 25 Mar. 2013, http://www.emarketer.com/Article/Among-Hispanics-Whos-Leading-Digital-Adoption-Trends/1009755. For an overview of the issue, see Center for Digital Democracy, "Digital Target Marketing to African Americans, Hispanics and Asian Americans: A New Report," 18 Feb. 2013, http://www.democraticmedia.org/digital-target-marketing-african-americans-hispanics-and-asian-americans-new-report (both viewed 10 Feb. 2014).

29 See, for example, Epsilon Targeting, "High Performaing Marketing Data for Financial Services," http://www.epsilon.com/pdf/financial_datacard_092311.pdf (viewed 10 Feb. 2014).

30 Mike Hudson, "US Hispanics and Autos: The Next Generation of Growth," eMarketer, Oct. 2013, personal copy on file with author Jeff Chester. Another recent study said that 87 percent of Hispanics use smartphones and that 60 percent use tablets. eMarketer, "Smartphone, Tablet Uptake Still Climbing in the US," 14 Oct. 2013, http://www.emarketer.com/Article/Smartphone-Tablet-Uptake-Still-Climbing-US/1010297; eMarketer, "Hispanics Lead as Web Users Who are Truly Mobile-First," 8 Oct. 2013, http://www.emarketer.com/Article/Hispanics-Lead-Web-Users-Who-Truly-Mobile-First/1010280. The September 2013 Pew Internet & American Life survey noted that for "non-whites, among those who use their phone to go online, six in ten Hispanics and 43% of African Americans are cell-mostly users, compared with 27% of whites." Around 45 percent of those with a high school diploma or less use their phone for access. Maeve Duggan, Aaron Smith, "Cell Internet Use 2013," Pew Internet & American Life Project, 16 Sept. 2013, http://pewinternet.org/Reports/2013/Cell-Internet.aspx (all viewed 10 Feb. 2014).

31 Center for Digital Democracy, "Targeting the Digital Latino: FTC Needs to Address how Digital Marketers are Tracking Hispanic Consumers, inc. Youth," 28 May 2013, http://www.centerfordigitaldemocracy.org/targeting-digital-latino-ftc-needs-address-how-digital-marketers-are-tracking-hispanic-consumers-inc (viewed 10 Feb. 2014).

32 David Burgos, "Marketing to the New Majority: Strategies for a Diverse World," WPP, http://www.wpp.com/wpp/marketing/consumerinsights/marketing-to-the-new-majority/ (viewed 10 Feb. 2014).

33 Scarborough, "Millennial Influencers Favor Multicultural Media," di@log, 3 Oct. 2013, http://dialog.scarborough.com/index.php/millennial-influencers-favor-multicultural-media/ (viewed 10 Feb. 2014).

34 See, for example, Viacom, "Viacom Insights," http://north.viacom.com/company/research/ (viewed 10 Feb. 2014).

35 Karen Kramer, Liz Schwarte, Mariah Lafleur, and Jerome Williams, "Targeted Marketing of Junk Food to Ethnic Minority Youth: Fighting Back with Legal Advocacy and Community Engagement," ChangeLab Solutions, 2012, http://changelabsolutions.org/sites/default/files/TargetedMarketingJunkFood_FINAL_20120912.pdf (viewed 10 Feb. 2014).

36 Alan J. Liddle, "What They Didn't Tell You about McDonald's' Mobile Payments Trials," Nation's Restaurant News, 23 Sept. 2013, http://nrn.com/blog/what-they-didnt-tell-you-about-mcdonalds-mobile-payments-trials (viewed 10 Feb. 2014).

37 For an overview of these developments, especially relating to food marketing to youth, see Center for Digital Democracy and Berkeley Media Studies Group, "Digital Ads: Exposing How Marketers Target Youth: Publications," http://www.digitalads.org/how-youre-targeted/publications (viewed 10 Feb. 2014).

38 "The 'Social' Credit Score: Separating the Data from the Noise," Knowledge@Wharton, 5 June 2013, http://knowledge.wharton.upenn.edu/article/the-social-credit-score-separating-the-data-from-the-noise/ (viewed 10 Feb. 2014).

39 Big Data "appliance" platforms using Hadoop, MapReduce, etc., are used by the financial and many other sectors. See, for example, Teradata, "Financial Services," http://www.teradata.com/industry-expertise/financial-services/; Teradata Aster, "Deeper Insights for Financial Services," http://www.asterdata.com/solutions/financial.php; Michael Hickins, "Banks Using Big Data to Discover 'New Silk Roads,'" CIO Journal, 6 Feb. 2013, http://blogs.wsj.com/cio/2013/02/06/banks-using-big-data-to-discover-new-silk-roads/; Merkle, "Marketing Imperatives for Banking and Finance," 29 Apr. 2013, http://www.merkleinc.com/sales-collateral/marketing-imperatives-banking-and-finance; Merkle, "Connected Recognition," http://www.merkleinc.com/what-we-do/database-marketing-services/connected-recognition (all viewed 10 Feb. 2014).

40 "The 'Social' Credit Score: Separating the Data from the Noise."

41 Scores are used for the decision process, including the FICO Revenue Score, FICO Credit Capacity Index, and FICO Bankruptcy Score. Fair Isaac Corporation, "2013 Annual Report," personal copy on file with author Jeff Chester.

42 "The 'Social' Credit Score: Separating the Data from the Noise."

43 TSYS discusses a model to drive "cardholder response" with "calls to actions" for increased spending. TSYS, "How Card Issuers Can Leverage Big Data to Improve Cardholder Retention Efforts," http://www.tsys.com/BigDataWhitePaper/index.cfm (viewed 10 Feb. 2014).

44 The same bank also "analyzed use of 'companion devices' to discover which TV shows customers are watching when they interact with mobile banking." Andy Hirst, "Innovating With Big Data—Leading The SoLoMo Cycle," Business 2 Community, 17 Aug. 2013, http://www.business2community.com/big-data/innovating-with-big-data-leading-the-solomo-cycle-0585217#FrUfaXM2geIsueEU.99 (viewed 10 Feb. 2014).

45 TSYS, "How Card Issuers Can Leverage Big Data to Improve Cardholder Retention Efforts"; "The 'Social' Credit Score: Separating the Data from the Noise." The transformation of the financial marketplace is also occurring, of course, on a global level. MasterCard, which recently established an India-based Advanced Analytics Center of Excellence, says that the "big data analytics market is rapidly growing as companies seek real-time insight that allows them to better connect with their consumers." Pete Rizzo, "MasterCard's Secret to Big Data Monetization Might Surprise You," Pymnts.com, 30 Aug. 2013, http://www.pymnts.com/briefing-room/issuers/playmakers/2013/MasterCard-s-Secret-to-Big-Data-Monetization-Might-Surprise-You/ (viewed 10 Feb. 2014).

46 See, for example, Merkle, "Merkle Acquires Brilig, a Leading Data Exchange for Online Advertising," 10 Sept. 2012, http://www.merkleinc.com/news-and-events/press-releases/2012/merkle-acquires-brilig-leading-data-exchange-online-advertising; "Datalogix Acquires Connection Engine," Business Wire, 5 Sept. 2012, http://www.businesswire.com/news/home/20120905006666/en/Datalogix-Acquires-Connection-Engine (both viewed 10 Feb. 2014).

47 Acxiom, http://www.acxiomdigital.com/; Equifax IXI Services, "Powering Digital Marketing with Financial Insights," http://www.ixicorp.com/ixi-digital/; Experian, "Hitwise—Digital Marketing Intelligence," http://www.experian.com/hitwise/. TransUnion is engaged in digital marketing as well. "TransUnion Transforms Digital Marketing with Autonomy, an HP Company," 27 Sept. 2012, http://www8.hp.com/us/en/hp-news/press-release.html?id=1300935#.UlrM8CSE5bJ (all viewed 10 Feb. 2014).

48 LiveRamp, "70+ Integrations At Your Fingertips," http://liveramp.com/partners/ (viewed 10 Feb. 2014).

49 Acxiom explains that it takes bank data and combines them with information it and data broker partners provide about a consumer's "behaviors," "email opens," social media, "search," and "offline" activity. Detailed information regarding an individual can be scored and segmented using "Big Data" techniques, says Acxiom (including knowing that an individual is a "female with small children, searched on site for travel rewards" and also was "served … a card ad." All this online and offline data can be used for "targeting and personalization" involving more data brokers and online targeting companies. Acxiom's "Abilitec Digital" product "links traditional name and address information to your digital contact information," and "recognizes your customers across the digital divide." Acxiom Fact Sheet, personal copy on file with author Jeff Chester.

50 Oracle, "Oracle Social Marketing Cloud Service," http://www.oracle.com/us/products/social-marketing-cloud-service-1842850.pdf (viewed 10 Feb. 2014).

51 Oracle Eloqua, "Data Sheets: Lead Scoring and Profiling," http://www.eloqua.com/resources/data-sheets.html; Oracle, "Oracle Social Marketing Cloud Service: Overview," http://www.oracle.com/us/solutions/social/social-marketing-cloud-service/overview/index.html; Oracle Eloqua, "Lead Scoring," http://www.eloqua.com/resources/best-practices/lead-scoring.html (all viewed 10 Feb. 2014).

52 "Acxiom and Facebook Improve Online Advertising Experience with Partner Categories," MarketWatch, 10 Apr. 2013, http://www.marketwatch.com/story/acxiom-and-facebook-improve-online-advertising-experience-with-partner-categories-2013-04-10 (viewed 10 Feb. 2014).

53 Natasha Singer, "Ways to Make Your Online Tracks Harder to Follow," Bits, 19 June 2013, http://bits.blogs.nytimes.com/2013/06/19/ways-to-make-your-online-tracks-harder-to-follow-2/ (viewed 10 Feb. 2014).

54 David Kaplan, "Mindshare Data Chief Ivins Wants To Tear Down Walls Around First Party Analytics," Ad Exchanger, 16 Sept. 2013, http://www.adexchanger.com/data-exchanges/mindshare-data-chief-ivins-wants-to-tear-down-walls-around-first-party-analytics/#more-82083. See also WPP's data targeting subsidiary, Xaxis, "We Are Xaxis," http://www.xaxis.com/about; and "The Data Alliance and FourthWall Media Announce Data Partnership Agreement," Business Wire, 2 Oct. 2013, http://www.wpp.com/wpp/press/2013/oct/02/the-data-alliance-and-fourthwall-media-announce-data-partnership-agreement/ (all viewed 10 Feb. 2014).

55 Nielsen, "2013 P$YCLE Segmentation System: 43 Payday Prospects," http://www.claritas.com/MyBestSegments/Default.jsp?ID=37&id1=2000&id2=43; Nielsen, "2013 P$YCLE Segmentation System: 58 Bottom-Line Blues," http://www.claritas.com/MyBestSegments/Default.jsp?ID=37&id1=2000&id2=58. Among the other examples reflecting vulnerable economic constraints is "Social Insecurity: The most downscale of the mature segments, Social Insecurity is filled with ethnically diverse widows and widowers who rely on Social Security and Medicare/Medicaid for survival. With downscale incomes and low income-producing assets, these elderly singles barely register for owning stocks, mutual funds, and real estate investments. Nor can they muster the funds to buy insurance products other than some medical and whole life policies acquired earlier in their working lives. Financially strapped, most Social Insecurity residents lead quiet lifestyles in their older city apartments: there's little money for travel, nightlife, or dining out. Instead, this segment is the top-ranked audience for daytime television, particularly game shows, Spanish-language shows, and soaps." "Getting By Blues" is another category of at-risk consumers. "An ethnically diverse segment of 45- to 64-year-olds, members of this segment typically rent older apartments in urban and second-city neighborhoods. With low incomes and few assets—about two-thirds are unemployed—these consumers rank near the bottom for most banking and insurance products. But they do exhibit above-average rates for owning renter's and whole life insurance … ." Nielsen, "2013 P$YCLE Segmentation System: 53 Social Insecurity," http://www.claritas.com/MyBestSegments/Default.jsp?ID=37&id1=2000&id2=53. See product information on Nielsen Prizm Digital, including mobile devices, http://www.claritas.com/MyBestSegments/Default.jsp?ID=70 . As Nielson explains "previously anonymous website visitors now become identifiable Audiences for brand Advertising." Its "Prizm Digital Segments predict the likelihood of any household and purchase behaviors … ." The data include "what a consumer is likely to purchase," including financial services, whether he or she can "afford a specific product," and the "websites that a consumer is likely to visit." This targeting can also be done on mobile devices as well: "previous anonymous mobile web and app visitors now become identifiable … ." Nielsen, "Reach Your Best Customers and Prospects Online with Nielsen Prizm Digital," 2013,http://www.slideshare.net/vivastream/nielsen-prizm-digital-2013 (all viewed 10 Feb. 2014).

56 Datalogix, "DLX Finance," http://www.datalogix.com/wp-content/uploads/2012/06/DLX_Finance_one-pager_0313.pdf. Datalogix says that it converts a customer's data file "into an anonymous online audience" that is a "1:1" match. Datalogix, "DLX OnRamp," http://www.datalogix.com/audiences/online/onramp/ In an illustration of the range of data available today for financial services companies, Datalogix says it can build a "custom" list of targets. This data can include " thousands of data points: specific SKUs of retail purchases, brand level CPG purchases, and granular demographic and finance data. This fine-point data can be identified and combined into an infinite number of custom segments … ." Datalogix says its data comes from the U.S. Census, summarized credit sources and public records. It has partnerships with many data sources. Datalogix, "Build Your Own Audience" http://www.datalogix.com/audiences/online/syndicated-segments/; Datalogix, "DLX Finance," http://www.datalogix.com/industries/finance/ (all viewed 10 Feb. 2014).

57 "Oracle Buys BlueKai," News release, 24 February 2014, available at http://www.oracle.com/us/corporate/press/2150812

58 BlueKai, "Audience Data Marketplace," http://www.bluekai.com/audience-data-marketplace.php (viewed 10 Feb. 2014).

59 Alliant, "Alliant Info Center," http://www.alliantdata.com/news-resources/alliant-info-center-2/. Credit bureau Experian has invested in its ability to develop products for online. The seventh-largest Facebook advertiser in 2012, its "Alchemy" division targets Facebook users by incorporating precise information. It uses data to help marketers target consumers online, including those on Facebook and using mobile phones. Its "Audience IQ" service is "a digital advertising platform that allows marketers to utilize consumer data within display advertising for more relevant messaging." Clients can bring their own database or "prospect list" and Experian will help identify a "Direct match" in order to "tailor offers for cross-sell, up-sell or prospecting." It will help develop custom targeting to "match the highest-value segments with offers." Experian, "Targeted Display Advertising," http://www.experian.com/marketing-services/online-display-advertising.html; Experian, "Audience Targeting for Financial Services," http://www.experian.com/marketing-services/audience-targeting-for-financial-services-industry.html; Experian, "What We Do," http://www.experian.com/social-marketing/what-we-do.html; Jim Edwards, "Meet The 30 Biggest Advertisers On Facebook," Business Insider, 24 Sept. 2012, http://www.businessinsider.com/the-30-biggest-advertisers-on-facebook-2012-9?op=1; Mark Walsh, "Experian Unit Unveils Facebook-Based Research Tool," Online Media Daily, 5 Apr. 2011, http://www.mediapost.com/publications/article/147980/experian-unit-unveils-facebook-based-research-tool.html#axzz2gDaFwAVe (all viewed 10 Feb. 2014).

60 AOS is comprised of three key layers that enable one-to-one marketing at scale:
- The Data Layer: allowing marketers to ingest and unify virtually any type of data—structured, unstructured, first-party CRM, third-party data source, and Acxiom's consumer data—making it accessible in one place via a click.
- The Audience Operations Layer: enabling that data to be cleansed, matched, and contextualized to provide actionable insights about real people that can then be activated across channels and media buys.
- The Applications Layer: activating a growing roster of trusted development partners to create AOS-approved apps customized to meet any marketing need.
- "Acxiom Releases Audience Operating System," Destination CRM.com, 30 Sept. 2013, http://www.destinationcrm.com/Articles/CRM-News/Daily-News/Acxiom-Releases-Audience-Operating-System-92294.aspx. Epsilon has also expanded its data collection for online targeting. See also David Kaplan, "With GetOnboard, LiveRamp Blurs the Lines Between CRM and Advertising," Ad Exchanger, 29 Aug. 2013, http://www.adexchanger.com/data-exchanges/with-getonboard-liveramp-blurs-the-lines-between-crm-and-advertising/ (both viewed 10 Feb. 2014).

61 4INFO, "Adhaven Bullseye," http://www.4info.com/solutions/adhaven-bullseyehttp://www.catalinamarketing.com/news-events/press-releases/details.php?id=328 (both viewed 10 Feb. 2014).

62 TruSignal, "Discover Your Formula," http://www.tru-signal.com/how-it-works/discover-your-formula; TruSignal, "Audiences," http://www.tru-signal.com/audiences/truaudience-custom-segments (both viewed 10 Feb. 2014).

63 The ability to use sophisticated testing to further micro-segment and analyze consumers is another issue that requires consumer and regulatory scrutiny. FICO notes that "businesses can further compress test-and learn cycles—by using experimental design … with which large numbers of decision strategies are tested simultaneously on smaller population subsets." FICO Labs, "Getting Maximum Value Out of Analytics," FICO Labs Blog, 18 Sept. 2013, http://ficolabsblog.fico.com/2013/09/getting-the-value-out-of-analytics-.html (viewed 10 Feb. 2014).

64 "Hundreds of propensity models (one for each action) can be deployed: … these models allow each potential action to be considered based on its probability of acceptance." Scores can also be used by financial services companies, retailers, and others to gauge "the likelihood that a particular customer is a retention risk" and help determine what offers or incentives a financial institution may present to a particular customer. James Taylor, "Managing the Next Best Activity Decision," Decision Management Solutions, 2012, personal copy on file with author Jeff Chester. For an example of how real-time analytics are used for financial decision making, see Penny Crossman, "Bank of the West's CIO Is on a Quest for Real-Time Analytics," American Banker, 30 Sept. 2013, http://www.americanbanker.com/issues/178_189/bank-of-the-wests-cio-is-on-a-quest-for-real-time-analytics-1062492-1.html (viewed 10 Feb. 2014).

65 KXEN, "Scorer," http://www.kxen.com/Products/Scorer. KXEN says its "patented InfiniteInsight® Modeler automates the building of sophisticated predictive models for every data mining function under the sun … ." KXEN, "Modeler," http://www.kxen.com/Products/Modeler (both viewed 10 Feb. 2014).

66 Acxiom explains that Audience Propensities "incorporate consumer behavior, 3rd party transactional, response and other types of data to model purchase propensities, brand affinities, in-marketing timing and shopping channel preference. Advanced analytical algorithms are applied, creating a model score that rates the probability of a specified action and/or affinity. Model scores predict the likelihood of consumers to respond to particular messages and offers or determine the likelihood of a customer to spend a certain amount over the course of their relationship with a brand." Among the propensities addressed are spending, assets, attitude, and behavior. Acxiom, "Acxiom Audience Propensities," https://marketing.acxiom.com/AcxiomAudiencePropensities.html?CMP=701C0000000UvQ5&ls=Other&status=Downloaded (registration required).

67 Acxiom, "Audience Propensities," http://acxiom.com/resources/audience-propensities/ (viewed 10 Feb. 2014).

68 These are called by FICO "Behavior Sorted Lists." These analytical capabilities—"real-time mapping of actual customer behavior"—are also used to identify commonalities ("archetypes") in other consumers. Such analysis now enables financial marketers and others to go beyond identifying the behavior of a single user, and to create "collaborative profiles" using data derived from others that "maps actual customers to multiple archetypes." FICO, "Extracting Value from Unstructured Data," FICO Insights, Oct. 2013, http://www.fico.com/en/Communities/Pages/Insights.aspx (registration required)..

69 FICO, "Which Retail Analytics Do You Need?" Insights, n. 49, Aug. 2012, http://www.ngrsummit.com/media/whitepapers/2012/Fico.pdf (viewed 10 Feb. 2014).

70 KXEN explains that its InfiniteInsight Social product "automatically segments your social networks into communities of individuals with similar interests to give you a whole new level of customer insight." KXEN, "Find Your Influencers," http://www.kxen.com/Products/Social+Network+Analysis. See also KXEN, "Infinite Insight to Your Clients," http://www.kxen.com/Industries/Financial+Services; KXEN, "Deploy Your Scores," http://www.kxen.com/Products/Scorer. See also Acxiom's new partnership with social media marketing company Aditive: Laurie Sullivan, "Social Matures To Support Ad Targeting," Media Post, 2 Oct. 2013, http://www.mediapost.com/publications/article/210409/social-matures-to-support-ad-targeting.html?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+data-and-targeting-insider+%28MediaPost+|+Data+and+Targeting+Insider%29#axzz2iHXorJBs (all viewed 10 Feb. 2014).

71 Stephanie Armour, "Borrowers Hit Social-Media Hurdles,"8 January 2014, The Wall Street Journal, available at http://online.wsj.com/news/articles/SB10001424052702304773104579266423512930050

72 Companies use such scoring to better determine risk, but its role in potentially blacklisting consumers requires further investigation. Alliant, "ProfitSelect," http://www.alliantdata.com/wp-content/themes/alliant/pdf/Alliant_ProfitSelect_10-05-10.pdf (viewed 10 Feb. 2014).

73 Netmining, "Our Technology: How the Scoring Engine Works," http://www.netmining.com/marketing-technology/; Netmining, "Our Clients: Across Industries, Across The World," http://www.netmining.com/clients/; Netmining, "Meet Our Partners," http://www.netmining.com/company/partners/ (all viewed 10 Feb. 2014).

74 TellApart, "Data Science for Driving Sales," http://www.tellapart.com/about/#platform; Tellpart, "Marketing Solutions for Omnichannel Commerce," http://www.tellapart.com/about/#solutions (both viewed 10 Feb. 2014).

75 Adroit Digital, "Data Co-Op," http://www.adroitdigital.com/solutions/co-op/. There is also a "Buyer Score" used to rank a company's email subscribers based on their ability to buy. "Connection Engine Launches Buyer Score to Provide Instant Predictions of Email Subscriber Value Segment Customers by ROI Potential and Identify High Value Customers," PR Newswire, 17 Feb. 2012, http://www.prnewswire.com/news-releases/connection-engine-launches-buyer-score-to-provide-instant-predictions-of-email-subscriber-value-segment-customers-by-roi-potential-and-identify-high-value-customers-139560728.html (both viewed 10 Feb. 2014).

76 Alliant, "Alliant ProfitSelect," http://www.alliantdata.com/solutions/lead-management/alliant-profitselect/; Alliant, "Profit Select." The GeoPerformance scores are "available for online targeting." Alliant, "GeoPerformance Scores," http://www.alliantdata.com/wp-content/themes/alliant/pdf/Alliant_GeoPerformance_Scores_10-05-10.pdf; Alliant, "Alliant GeoPerformance Scores," http://www.alliantdata.com/solutions/customer-file-enhancements/geographic-targeting/ (all viewed 10 Feb. 2014).

77 Dstillery, "Neighborhood Social Targeting," http://www.everyscreenmedia.com/products/everyscreen-data-products/neighborhood-social-targeting/ (viewed 10 Feb. 2014). Mobile payment processer TSYS explains that companies can use Facebook "likes" or Foursquare check-ins "to fine-tune customer targeting … [eventually looking to] harvest data from … social media, call centre notes, online chats, geo-location logs and email … ." Rob Hudson, "How Card Issuers Can Leverage Big Data to Improve Cardholder Retention Efforts," TSYS, 2013 (registration required).

78 "Social network platforms have opened up new opportunities for identifying social influencers," FICO explains. "There are algorithms such as graph theory to measure node centrality and node relevance in a network. They can be applied on social networks with context specific influence metrics to identify the nodes (individuals) that are most central or relevant … . With the advent of social media, analytic scientists have gotten a powerful playground to understand peer influence and identify social influencers. This is made easy by the recent innovations in Big Data, which allow processing and analyzing of extremely large volumes of structured and unstructured data generated by social media platforms." Shafi Rahman, "In with the In Crowd: Targeting Social Influencers," FICO Labs Blog, 26 Sept. 2013, http://ficolabsblog.fico.com/2013/09/being-in-with-the-in-crowd-targeting-social-influencers.html?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+FicoLabsBlog+%28FICO+Labs+Blog%29. See also Neill Crossley, "FICO Lessons in Developing, Applying Decision Modelling Methods," FICO Labs, 20 Dec. 2013 Neill Crossley, FICO Labs, Dec 20, 2013, http://www.kdnuggets.com/2013/12/fico-lessons-developing-applying-decision-modelling-methods.html. The role of social media as a form of consumer scoring is an issue that requires further analysis. According to *Wired*, Rachel Botsman, author of *What's Mine is Yours: The Rise of Collaborative Consumption*, "painted a vision of a kind of utopia where our actions, and the way in which we lead our lives and treat other people, become more important to the economy than our credit rating. 'Personal reputation is being transformed and it's going to become a currency and cornerstone of our society in the next decade,' said Botsman. With peer-to-peer marketplaces taking over every pocket of industry, so too, will reputation capital become the driving force behind it. It sounds good for those struggling to get a mortgage, a loan or even a job, that our actions speak louder than the boxes we tick on a bank manager's form." Liat Clark, "Forget Online Porn, a Teen's Biggest Problem is Reputation Management," *Wired*, 22 July 2013, http://www.wired.co.uk/news/archive/2013-07/22/kids-privacy (all viewed 10 Feb. 2014).

79 Jeanette Fitzgerald, senior vice president and general counsel, Epsilon Data Management, letter in response to Congressional inquiry, 14 Aug. 2012, http://www.epsilon.com/pdf/Epsilon_Congressional_Response_8_14_2012.pdf; Epsilon, "Financial Services & Insurance," http://www.epsilon.com/solutions/industry-solutions/financial-insurance#sthash.ijI4xkFf.dpuf. Facebook marketers can use Epsilon data, for example, to "target people who currently have an auto loan." AdEspresso, "A Quick Guide to Each of Facebook's Partner Categories," AdEspresso Insight, 9 May 2013, http://blog.adespresso.com/facebook-partner-categories-guide/ (viewed 10 Feb. 2014).

80 KXEN, "Managing the Next Best Activity Decision with Predictive Analytics," http://www.kxen.com/News+and+Events/Webinars/Next+Best+Activity+%28James+Taylor%29 (viewed 10 Feb. 2014). According to one industry report discussing the use of scores by banks and other financial services companies, generating profits from consumers requires their adoption of four to six or more products. One question that needs to be answered concerns the role of e-scores in helping identify a set of financial products targeted to at-risk consumers that do not advance their economic interests.

81 Given the capability to monitor and analyze a consumer's actions more closely, Experian explains that "the combination of scores and data attributes (both point-in-time and trended) allow for micro-targeting within segments of the near-prime population." Experian, "Thin File Redefined," http://www.experian.com/consumer-information/thin-file.html; "Experian Announces its Extended View Score," 13 June 2012, http://press.experian.com/United-States/Press-Release/experian-announces-its-extended-view-score.aspx (both viewed 10 Feb. 2014). Experian also urges using what are known as "hot lists."

82 *Wired*, 15 Aug. 2013, http://www.wired.co.uk/news/archive/2013-07/01/zestfinance-douglas-merrill (all viewed 10 Feb. 2014).

83 It can approve and fund a loan of up to $10,000 in near real-time … ." In Virginia, AvantCredit charges up to 95 percent APR. AvantCredit, https://www.avantcredit.comhttps://www.avantcredit.com/rates_terms.html10 Feb. 2014

84 This score is used with the company's credit report that includes "Information compiled from payday loans, installment loans, non-prime credit cards and other sources of non-traditional credit information … Detailed transaction based payment history … Data intelligence, such as bankruptcy data and loan information or status … Real-time reporting for a complete picture of each consumer." The company's identification-verification technologies can analyze "disparities between IP addresses, time zones and geolocation … ." DataX, "DataX Unveils New Advanced Solutions for Mitigating Risk and Lead Fraud," 22 Apr. 2013, http://www.dataxltd.com/2013/04/22/datax-unveils-new-advanced-solutions-for-mitigating-risk-and-lead-fraud/; DataX, "DataX Credit Report," http://www.dataxltd.com/consumer-credit-reports/credit-reporting/; DataX, "Credit Optics," http://www.dataxltd.com/consumer-performance-reports/credit-optics/ (all viewed 10 Feb. 2014).

85 Jim Marous, "Moven: From Mobile Banking to Mobile Money," Next Bank, 18 Feb. 2013, http://www.nextbank.org/mobile/moven-from-mobile-banking-to-mobile-money-2/; "Is The World Ready For Social Media Credit Scores?" The Financial Brand, 14 Aug. 2012, http://thefinancialbrand.com/24733/social-media-credit-score/; Movenbank, "You've got CRED," YouTube, 9 Aug. 2012, http://www.youtube.com/watch?v=vQ30_k6zalI (all viewed 10 Feb. 2014). eMarketer, "Digital Banking Trends: With Consumer Preferences in Flux, Is Omnichannel the Answer?" Aug. 2013, personal copy on file with author Jeff Chester.

86 Richard Waters, "Internet Groups Brace for Subprime Fallout," FT.com, 27 Aug. 2007, http://www.ft.com/intl/cms/s/0/e2a8cf92-54c6-11dc-890c-0000779fd2ac.html?siteedition=intl#axzz2fjzTJIeT (subscription required). See also Larry Dignan, "Mortgage Upheaval Could Ding Online Advertising," ZDNet, 16 Aug. 2007, http://www.zdnet.com/blog/btl/mortgage-upheaval-could-ding-online-advertising/5966 (viewed 10 Feb. 2014).

87 The Interactive Advertising Bureau, a trade group that releases an annual report on Internet revenues, describes lead generation as "Fees paid by advertisers to online companies that refer qualified potential customers (e.g., auto dealers which pay a fee in exchange for receiving a qualified purchase inquiry online) or provide consumer information (demographic, contact, behavioral) where the consumer opts in to being contacted by a marketer (email, postal, telephone, fax). These processes are priced on a performance basis (e.g., cost-per-action, -lead or -inquiry), and can include user applications (e.g., for a credit card), surveys, contests (e.g., sweepstakes) or registrations." Interactive Advertising Bureau, "IAB Internet Advertising Revenue Report: 2013 First Six Months' Results," Oct. 2013, http://www.iab.net/media/file/IABInternetAdvertisingRevenueReportHY2013FINALdoc.pdf (viewed 10 Feb. 2014).

88 eBureau, "Consumer Lead Quality Management," http://www.ebureau.com/lead-quality-management (viewed 10 Feb. 2014).

89 "DoublePositive, Leadscon, & Performance-Based Online Marketing—A Unified Field/Funnel Theory," DevsBuild.It, 5 July 2012, http://devsbuild.it/resources/type/article/doublepositive-leadscon-performance-based-online-marketing-unified (viewed 10 Feb. 2014).

90 According to The Financial Brand website, it costs $9.34 per click for the term "checking account." "Mortgage rates" deliver a $5.18 cpc (cost per click); "debit cards" deliver $2.55; and "personal loans" a $3.14 cpc. The rates companies pay to acquire keywords connected to a specific geographic market also drive up the costs. So "auto loan Chicago" can cost $9.11 while a search for "auto loan Atlanta" is $7.61. "Google AdWords Costs For Banks And Credit Unions," The Financial Brand, 30 Apr. 2013, http://thefinancialbrand.com/29376/google-adwords-costs-for-banks-credit-unions/ (viewed 10 Feb. 2014).

91 "2013 State of Bank & Credit Union Marketing," The Financial Brand, 12 Feb. 2013, http://thefinancialbrand.com/27511/2013-state-of-bank-credit-union-marketing/ (viewed 10 Feb. 2014); "Google AdWords Costs For Banks And Credit Unions." Illustrating how much marketing is done online, the financial services industries delivered 433 billion online ad impressions in 2012.

92 Google, "Four Truths About US Hispanic Consumers," Think Insights, Oct. 2010, http://www.google.com/think/research-studies/four-truths-about-us-hispanic-consumers.html; Google, "Five Truths of the Digital African American Consumer," Think Insights, June 2011, http://www.google.com/think/research-studies/five-truths-of-the-digital-african-american-consumer.html. Google also focuses on the digital marketing of financial services. Google, "Financial Services," Think Insights, http://www.google.com/think/industries/financial-services.html (all viewed 10 Feb. 2014).

93 "LeadFlash Sees Huge Success with Call Center and Hot Transfer Product," 8 Feb. 2013, http://www.ereleases.com/pr/leadflash-sees-huge-success-call-center-hot-transfer-product-99932. See the Spanish-language version of this payday lender, for example: LendUp, https://www.lendup.com/ (both viewed 10 Feb. 2014).

94 MarketView, "Advantage," http://mktview.com/solutions/advantage/; LeadFlash, "About LeadFlash," https://www.leadflash.com/aboutus.aspx (both viewed 10 Feb. 2014).

95 LendUp, which lends small amounts to consumer, advertises on Facebook and uses the real-time services of ad exchange company AppNexus. AppNexus's data partners collect a diverse range of financial information on consumers.

96 For example, Optimo performs the following types of tests on landing pages to insure that affiliates get best possible conversion rates:
- A/B (split) testing
- Multivariate testing (Full Factorial, Fractional Factorial, Adaptive Multivariate Testing Methods)
- Multi-page testing
- Usability testing
- Template variation testing
- Total-experience testing

"Optimo has built in experiments that use analysis of variance, Chi-squared test, correlation and factor analysis, mean square weighted deviation, linear regression and time series analysis, maximum likelihood estimation, Stochastic calculus, Black–Scholes model to automatically come up with highest revenue generating funnels. In addition to data mining and demographic analysis we hire behavioral psychologists to help us improve our funnels by understanding consumer behavior." LeadsMarket, "Our Technology," http://www.6ivi9.com/optimo.aspx?AspxAutoDetectCookieSupport=1. See also "Personal Loan Affiliate Program LeadsMarket.com Releases New Compliance Package for Publishers," PRWeb, 14 May 2013, http://www.prweb.com/releases/2013/5/prweb10720963.htm. Companies can also use eye-tracking technology to make sure consumers seldom stray from the "call-to-action" or some other lure used to gather the desired data. Tobii, "Advertising Research and Eye Tracking," http://www.tobii.com/eye-tracking-research/global/research/advertising-research/ (all viewed 10 Feb. 2014).

97 Datalot, "Business Analytics Coordinator at Datalot in Brooklyn, NY," http://www.jobscore.com/jobs/datalot/business-analytics-coordinator/axzi2Yd1Or462HiGakhP3Q&ref=rss Founder Collective, "Datalot," http://foundercollective.com/companies-Datalot; Vantage Media "Our Approach," http://www.vantagemedia.com/approach.php (all viewed 11 Feb. 2014).

98 As one lead generator using Facebook explained, it "can take offline data, such as email addresses or phone numbers, from an advertiser's CRM or other sources, and find those users on Facebook … targeting to users we historically were unable to identify online … ." Gina Maranto, "Programmatic Buying: Ensuring Your Ads Are Viewed," DoublePositive, 6 Aug. 2013, http://www.doublepositive.com/index.php?cID=651http://www.doublepositive.com/your-needs/online-advertisers/; Double Positive, "Our Solutions," http://www.doublepositive.com/our-solutions/ http://www.doublepositive.com/whos-working-us/clients/ DoublePositive, "Our Blog," http://doublepositiveblogs.com/blog/page/3/. Mainstream financial institutions also use leads and scores to identify and target consumers, and to identify how to treat current and potential customers. For example, U.S. Bank's program enhances the leads it captures by incorporating customer profile information that is then "prioritized via a scoring model. The bank says criteria for its scoring models involve "the quality and temperature of the lead" as well as the "customer value to the bank." Adobe says that all the data it compiles and that are used for clients like U.S. Bank reflect required privacy practices. See Adobe's U.S. Bank, http://apps.enterprise.adobe.com/go/701a0000000lBHCAA2, and SunTrust Bank, http://apps.enterprise.adobe.com/go/701a0000000ltEhAAI, case studies. See also Adobe, "Financial Services," http://www.adobe.com/solutions/financial-services.html (all viewed 11 Feb. 2014).

99 Interactive Advertising Bureau, "Lead Generation Committee," http://www.iab.net/member_center/committees/working_groups/lead_generation_committee; Online Lenders Alliance, http://www.onlinelendersalliance.org/; LeadsCon, http://leadscon.com/ (all viewed 11 Feb. 2014).

100 Frederic Huynh, "Who Gets A FICO Score?" FICO Banking Analytics Blog, "29 Aug. 2013, http://bankinganalyticsblog.fico.com/2013/08/who-gets-a-fico-score.html/ (viewed 11 Feb. 2014).

101 Kristine Snyder, "How Experian is Helping Customers with Little to No Credit History," 14 June 2012, http://www.experian.com/blogs/news/2012/06/14/extended-view-score/. See also Experian, "Credit Service for the Underserved, Unbanked and Underbanked Populations," http://www.experian.com/consumer-information/unbanked.html?intcmp=CIS_sptmod_learn_underserved_080912 (both viewed 11 Feb. 2014); Experian, "Thin File Redefined."

102 Rachel Schneider and Rob Levy, "Alternative Data Can Help Eliminate Credit's Catch-22," American Banker, 7 Nov. 2012, http://www.americanbanker.com/bankthink/alternative-data-can-help-eliminate-credits-catch-1054167-1.html (viewed 11 Feb. 2014).

103 PERC, "Alternative Data is the Achievable Solution: Alternative Data Can be Used to End Credit Invisibility and Drive Financial Inclusion," http://www.perc.net/approach/#drive; Katherine Lucas McKay, "Policy Update: Full File Credit Reporting," CFED, 13 July 2011, http://cfed.org/blog/inclusiveeconomy/policy_update_full-file_credit_reporting/ (both viewed 11 Feb. 2014).

104 National Consumer Law Center, "Full Utility Credit Reporting: Risks to Low-Income Consumers," July 2012, http://www.nclc.org/images/pdf/energy_utility_telecom/consumer_protection_and_regulatory_issues/ib_risks_of_full_utility_credit_reporting_july2012.pdf (viewed 11 Feb. 2014).

105 Equifax, "Insight Score for Retail Banking," http://www.equifaxddasolutions.com/insight.html; Philip Ryan, "Opportunity with the Underbanked for FIs," Bank Innovation, 12 Oct. 2012, http://bankinnovation.net/2012/10/opportunity-with-the-underbanked-for-fis/ (both viewed 11 Feb. 2014).

106 Equifax, "Insight Score for Retail Banking."

107 In 2012 Experian launched its own risk-scoring product for the unbanked, called "Extended View." Experian explained that its "Fair Credit Reporting Act-compliant credit score" provides lenders "with a more robust underwriting arsenal … able to fund financial products for this underpenetrated market." Extended View is promoted to "banks, credit unions, auto lenders, telecommunications companies and utility providers. Of note is Experian's analysis that shows that "[h]ome values of the typically unscoreable mirror the median U.S. home value," which is $154,000. Median home value at the "super prime/prime" credit tier is $147,300, while at "near prime" it is $110,700. Barrett Burns, "Expand & Grow: How to Reach and Educate More Members Using a New Breed of Credit Scores," VantageScore, 2013, http://www.vantagescore.com/images/resources/NAFCU%20Conference%202013.pdf (viewed 11 Feb. 2014).

108 FICO, "FICO Expansion Score," http://www.fico.com/en/wp-content/secure_upload/FICO_Expansion_Score_1709PS.pdf By using alternative data such as utility payment history and property/asset data, FICO says that its models have shown that "60-75% of traditionally unscorable consumers can be assigned a statistically meaningful credit score … ." FICO is competing in this "thin file" market. Its FICO 8 Score, which "boosts predictive strength by more than double," addresses consumers with these "thin files," including "nonprime" borrowers. It is also using nontraditional data to "assess the credit risk of consumers with little or no credit history at the three major credit reporting agencies." They should also "supplement traditional credit data with alternative data that meet regulatory requirements," explains FICO. FICO, "To Score or Not to Score?" Insights, n. 70, Sept. 2013, http://www.fico.com/en/wp-content/secure_upload/70_Insights_To_Score_or_Not_To_Score_3009WP.pdf (both viewed 18 Feb. 2014).

109 CoreLogic Credco, "CreditIQ," http://www.corelogic.com/product-media/asset_upload_file116_23745.pdf (viewed 17 Feb. 2014).

110 FICO says that to serve "unscoreable consumers" companies need to "augment the limited traditional credit information with alternative data that adds predictive value." FICO, "To Score or Not to Score?"

111 "IXI Services enables its clients to differentiate and target consumer households and target markets based on proprietary measures of wealth, income, spending capacity, credit, share-of-wallet, and share-of-market." Equifax IXI Services, "About Us," http://www.ixicorp.com/about/company-overview/ (viewed 11 Feb. 2014).

112 Experian says that this reduces the volume of prospects who go through the full application process, resulting in "better approval rates and ROIs" for consumers found online." Experian, "Expanding the Marketable Universe," 2011, http://www.experian.com/assets/consumer-information/white-papers/universe-expansion-white-paper.pdf (viewed 11 Feb. 2014).

113 Equifax. "Aggregated FICO Scores," http://www.ixicorp.com/products-and-services/customer-targeting-and-scoring/aggregated-fico-scores/ (viewed 11 Feb. 2014).

114 "CreditStyles Pro (viewed 11 Feb. 2014).

115 "CreditStyles Pro" 360i, "Equifax Names 360i Its Lead Agency," 360i Blog, 22 May 2013, http://blog.360i.com/360i-news/equifax-names-360i-its-lead-agency. Equifax recently hired a digital marketing company to help it "develop and execute a data-driven acquisition strategy for its Personal Solutions unit, helping develop a DMP." This will include bringing together the company's online and offline data. Equifax, "Personal Solutions," http://www.equifax.com/home/en_us; IdentityProtection.com, http://www.identityprotection.com/home (all viewed 11 Feb. 2014).

116 "CreditStyles Pro." "Predictive triggers" are another set of aggregated services that are used to target "households within a micro-neighborhood," Equifax explains. There are predictive triggers for "automotive finance, home financing and refinancing, bank cards, retail purchases, student loans, and personal loans." These products enable the targeting of consumers who have a "specific profiled need," even if they have not yet applied for credit or another financial product. "CreditStyles Pro

117 IXI says that financial marketers can "Tailor Online Messaging: Differentiate site visitors on the fly, and develop creative and messaging that will resonate with your target audience based on a better understanding of their financial profiles." Equifax IXI Services, "Financial Cohorts," http://www.ixicorp.com/products-and-services/customer-segmentation/financial-cohorts/. IXI also recently introduced a digital "audience intelligence tool to better evaluate who is actually viewing and responding to their ads." It allows marketers to "evaluate campaigns in real time and make instant adjustments. Equifax IXI Services, "AudienceIntel: Real-Time Online Audience Intelligence," http://www.ixicorp.com/ixi-digital/solutions-for-advertisers-and-agencies/audienceintel/ (both viewed 11 Feb. 2014).

118 Equifax IXI Services, "IXI Digital Targeting Options," http://www.ixicorp.com/ixi-digital/ixi-digital-targeting-options/ (viewed 11 Feb. 2014).

119 Equifax IXI Services, "Financial Cohorts." One "leading bank" is quoted on Equifax's IXI site reporting that by using Financial Cohorts "profile information," it was able to "narrow its target audience" for an offer by 90 percent. While a great revenue success for the bank, consumers not deemed qualified were simply eliminated from learning about the product. Equifax IXI Services, "Financial Cohorts." See also Equifax IXI Services, "IXI Digital Targeting Options"; Equifax IXI Services, "AudienceIntel."

120 Semcasting explains: "Only aggregated and averaged values for publically available demographics, and only Zip+4 level Smart Zone locations are used to make an onboarding match. There is no ability to reverse engineer onboarding to an individual or household level, and no cookies, tracking or any other form of tagging is used." Semcasting, "Onboarding," Semcasting Marketing Appliance, http://www.marketingappliance.com/#!onboarding/c176x (viewed 11 Feb. 2014).

121 The "mobile payments [sector] is a multibillion-dollar opportunity that could easily expand to a multitrillion-dollar opportunity as smartphone penetration increases," explained eMarketer. Bryan Yeager, "Mobile Payments: An Updated Forecast, Early Successes and Visions for the Future," eMarketer, July 2013, personal copy on file with author Jeff Chester.

122 Driving the growth of prepaid cards is the underserved community (globally, not just in the U.S.), according to a recent report commissioned by MasterCard. "… [U]nbanked and underbanked consumers represent a significant opportunity to financial institutions, as they are most likely to access the Internet using a mobile phone. Customers who use prepaid cards want full control of their money—anytime, anywhere. FIS allows financial institutions to serve the needs of all of their prepaid cardholders with an intuitive, easy-to-use mobile prepaid solution." FIS, "Mobile Prepaid," http://www.fisglobal.com/products-mobilefinancialservices#prepaid (viewed 11 Feb. 2014).

123 Forms of prepaid cards are proliferating, with both Social Security and SSI available as "Direct Express Cards" in addition to direct deposit.

124 The projected annual rate of growth for prepaid cards is 22 percent through 2017. MasterCard's report identifies that "governments around the world are increasingly driving financial inclusion," with the prepaid card viewed as an important way to accomplish that task. MasterCard, "2012 Global Prepaid Sizing Study," July 2012, https://www.partnersinprepaid.com/pdf/a-look-at-the-potential-for-global-prepaid-growth-by-2017.pdf?maincategory=TOPIC Ssubcategory=RESEARCH (viewed 11 Feb. 2014).

125 Walmart MoneyCard, "About Our Products," https://www.walmartmoneycard.com/walmart/about-our-products#cardtop; PayPal, "PayPal Prepaid MasterCard," https://www.paypal-prepaid.com/. Green Dot, MasterCard, Visa, and others also offer prepaid cards. Green Dot, "Reloadable Prepaid Cards, https://www.greendot.com/greendot; Visa, "Visa Prepaid Card," http://usa.visa.com/personal/personal-cards/prepaid-cards/index.jsp (all viewed 11 Feb. 2014).

126 The U.S. PIRG Education Fund has recommended best practices for debit cards used to distribute financial aid on campus, or debit cards linked to student IDs.

127 Consumer Federation of America and the National Consumer Law Center, "Advocates Urge CFPB to Ban Overdraft Fees and Payday Loans on Prepaid Cards," 25 July 2012, http://www.consumerfed.org/news/562 (viewed 11 Feb. 2014).

128 Sarah Perez, "American Express Serve Goes After the 'Under-Banked' with Prepaid Cards You Load with Cash in Stores," TechCrunch, 8 Oct. 2013, http://techcrunch.com/2013/10/08/american-express-serve-goes-after-the-under-banked-with-prepaid-cards-you-load-with-cash-in-stores/ (viewed 11 Feb. 2014).

129 With more than 50 million PayPal users in the U.S., consumers can use their PayPal account to pay for products at "millions of merchant locations," including via its mobile phone app and prepaid card. Discover, "PayPal and Discover to Bring PayPal to Millions of In-Store Locations," 22 Aug. 2012, http://www.investorrelations.discoverfinancial.com/phoenix.zhtml?c=204177&p=irol-newsArticle_print&ID=1727640&highlight=; PayPal, "Apps," https://www.paypal.com/webapps/mpp/mobile-apps; PayPal, "PayPal Prepaid MasterCard," https://www.paypal.com/webapps/mpp/paypal-prepaid-mastercard; David Heun, "PayPal's New Prepaid Card Upsells to the Underbanked," American Banker, 14 Feb. 2012, http://www.americanbanker.com/issues/177_31/paypal-prepaid-unbanked-underbanked-1046670-1.html. In September 2013, PayPal unveiled its latest new product—called Beacon. When PayPal users enter a store, and if they have downloaded and turned on a mobile app, they are "checked-in." The "merchant can then serve up any number of convenient commerce offers tied to a hands-free, swipe-less payment experience." Offers can be "targeted to where consumers happen to be in the store—$2.00 off mascara in the beauty aisle, 50 percent off flip flops in the fun-in-the-sun section, for instance." Karen Webster, "PayPal Puts Another Nail In NFC With Beacon Launch," Pymnts, 10 Sept. 2013, http://www.pymnts.com/briefing-room/mobile/playmakers/2013/PayPal-Puts-Another-Nail-In-NFC-With-Beacon-Launch/ (all viewed 11 Feb. 2014).

130 Jesse Haines, "Mobile and ... Anthropology?" Google Mobile Adds Blog, 2 Oct. 2012, http://googlemobileads.blogspot.com/2012/10/mobile-andanthropology.html. As digital market researcher Millward Brown wrote in its 2012 report ,"Mobile devices are indispensable and increasingly central to our lives." Millward Brown, "AdReaction 2012: Mobile Presents Marketers Globally with an Unprecedented Opportunity to Engage with Consumers," Changing Channels 70|20|10, http://www.millwardbrown.com/Sites/Changing_Channels/AdReaction.aspx (all viewed 11 Feb. 2014).

131 David Penn, "Corduro and LendUp Partner to Provide Loans for Medical Expenses," Finovate, 1 July 2013, http://finovate.com/2013/07/corduro-and-lendup-partner-to-provide-loans-for-medical-expenses.html (viewed 11 Feb. 2014).

132 PayPal, "Paying with Your Phone has Never Been So Easy—Or Looked So Good," https://www.paypal-forward.com/mobile/paying-with-your-phone-has-never-been-so-easy-or-looked-so-good/; PayPal, "Special Financing Options," https://www.billmelater.com/cm/paypal/landers/13ppbmlACQappfin.html; Chantal Tode, "PayPal Courts Shoppers with Offers, Instant Credit in Revamped App," Mobile Commerce Daily, 6 Sept. 2013, http://www.mobilecommercedaily.com/paypal-courts-shoppers-with-offers-instant-credit-in-revamped-app; "PayPal," iTunes Preview, https://itunes.apple.com/us/app/paypal/id283646709?mt=8. The CFPB recently launched an investigation into PayPal's BillMeLater product. Carter Dougherty, "EBay Probed by Regulator Over Loans Pioneered by Payday Lenders," *Businessweek*, 22 Oct. 2013, http://www.businessweek.com/news/2013-10-22/ebay-probed-by-regulator-over-loans-pioneered-by-payday-lenders (all viewed 11 Feb. 2014).

133 Sarah Perez, "Google Wallet for Gmail Invites Start Rolling Out to More Users," TechCrunch, 25 July 2013, http://techcrunch.com/2013/07/25/google-wallet-for-gmail-invites-start-rolling-out-to-more-users/ (viewed 11 Feb. 2014).

134 "A new group of payment terminal providers will add Isis' specifications for payments and loyalty programs, enabling the telecom-driven mobile wallet to reach approximately 90% of the addressable market for point of sale hardware in the U.S. ID Tech, OnTrack Innovations Global, PAX Technology, Uniform Industrial Corp. and XAC Automation Corp. have agreed to integrate Isis' SmartTap—a proprietary mobile commerce software specification that leverages NFC to enable users to pay, present loyalty cards and redeem offers as part of the same transaction." John Adams, "As Nationwide Launch Looms, Isis Broadens Terminal Market Reach," PaymentsSource, 4 Sept. 2013, http://www.paymentssource.com/news/as-nationwide-launch-looms-isis-broadens-terminal-market-reach-3015326-1.html. See also MasterCard, "MasterCard PayPass," http://www.mastercard.us/paypass.html (both viewed 11 Feb. 2014).

135 Roger Cheng, "How MasterCard Plans to Transform Mobile Purchases," C|Net, 24 Feb. 2013, http://reviews.cnet.com/8301-13970_7-57570783-78/how-mastercard-plans-to-transform-mobile-purchases/ (viewed 11 Feb. 2014).

136 PreCash, "Consumer Solutions," http://www.precash.com/consumer_financial_solutions.html; Evolve Money, http://www.evolvemoney.com/; Flip, http://www.myflipmoney.com/; PreCash, "PreCash Introduces the First Practical Mobile Wallet for Underbanked Consumers," 12 Sept. 2012, http://www.precash.com/flip_release_20120911.html. "For instant check deposits, PreCash plans to charge $1 plus 1% if the deposit for a payroll or government check plus $1 plus 3% of the amount for personal checks." Daniel Wolfe, "PreCash to Debut 'Flip,' a Mobile Wallet for the Underbanked," PaymentsSource, 10 Sept. 2012, http://www.paymentssource.com/news/precash-to-debut-flip-a-mobile-wallet-for-the-underbanked-3011817-1.html (all viewed 11 Feb. 2014)

137 And, as a fact sheet notes, "creating better shopping and paying experiences for customers and merchants alike." See "MCX Taps FIS to Power Its New Mobile Commerce Payments Network," Business Wire, 11 July 2013, http://www.businesswire.com/news/home/20130711005313/en/MCX-Taps-FIS-Power-Mobile-Commerce-Payments; "FIS Mobile Wallet," http://bcove.me/3o4cqjak (both viewed 11 Feb. 2014).

138 Chris Jay Hoofnagle, Jennifer M. Urban, and Su Li, "Mobile Payments: Consumer Benefits & New Privacy Concerns," Social Science Research Network, 24 Apr. 2012, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2045580 (viewed 11 Feb. 2014).

139 Chandan Sharma, "Banking on Technology: Embracing a New Era of Transparency," 2012, http://www.verizonenterprise.com/resources/articles/banking-on-technology_en_xg.pdf (viewed 11 Feb. 2014).

140 Financial institutions are also using neuromarketing—the influencing of an individual at a subconscious and emotional level—as part of their outreach campaigns See, for example, "Results of Neurological Testing of Advertising Effectiveness," done for "one of the world's top financial services companies," Neurofocus, personal copy on file with author Jeff Chester.

141 Cardlytics, "Cardlytics Releases 2011 Transaction-Driven Marketing Results at NRF Big Show," 16 Jan. 2012, http://cardlytics.com/press/cardlytics-releases-2011-transaction-driven-marketingtm-results-at-nrf-big-show/; Cardlytics, "Cardlytics Wins Judges' Choice Award in Fifth Annual Paybefore Awards," 8 Mar. 2011, http://cardlytics.com/press/cardlytics-wins-judges-choice-award-in-fifth-annual-paybefore-awards/; Cardlytics, "Advertisers FAQ," http://cardlytics.com/advertisers-2/advertisers-faq/ (all viewed 11 Feb. 2014).

142 Cardlytics explains that "When consumers log into their digital bank statements, they see advertising for products and services, chosen for them based on their recent purchases. They click to accept the offer, visit the store or website, and then use their debit or credit card to receive cash back from their bank. No coupons or codes, no registrations, nothing to slow down the checkout process—easy." Cardlytics, "Advertisers FAQ"; Cardlytics, "How It Works," http://cardlytics.com/advertisers-2/how-it-works/; Cardlytics, "Financial Institutions," http://cardlytics.com/financial-institutions-2/.http://www.firstdata.com/summit/insights.html. One leading academic technologist explained that "In the extreme, coupons will be available for all purchases, and smart shopping software on our phones or browsers will automatically search, aggregate, manage, and redeem these coupons, showing coupon-adjusted prices when browsing for products ... . Coupons will probably also merge with 'rewards,' 'points,' discounts, and various other incentives." Arvind Narayanan, "Personalized Coupons as a Vehicle for Perfect Price Discrimination," 33 Bits of Entropy, 25 June 2013, http://33bits.org/2013/06/25/personalized-coupons-price-discrimination/ (all viewed 11 Feb. 2014).

143 Nestor Bailly, "The Sophistication of Shopper Tracking," iQ, 20 Mar. 2013, http://iq.intel.com/iq/29562682/the-sophistication-of-shopper-tracking (viewed 11 Feb. 2014).

144 Path to Purchase Institute, http://p2pi.org/; Google, "Zero Moment of Truth (ZMOT)," http://www.thinkwithgoogle.com/collections/zero-moment-truth.html; SOLOMO Technology, http://solomotechnology.com/ (all viewed 11 Feb. 2014).

145 Sparkfly, "Solution: The Sparkfly Platform The Sparkfly Platform," http://www.sparkfly.com/solution (viewed 11 Feb. 2014).

146 Nielsen Catalina Solutions, for example, "utilizes the buying history of millions of U.S. households to target mobile advertising to a brand's most valuable consumers by relying on shopper-purchase data matched to anonymized households … ." "Catalina Launches Personalized Mobile Advertising for CPG Brands," 11 Mar. 2013, http://www.catalinamarketing.com/news-events/press-releases/details.php?id=328 (viewed 11 Feb. 2014).

147 Jennifer Marlo, "How Startups are Shaping the Future of Mobile," iMedia Connection, 18 Oct. 2013, http://www.imediaconnection.com/content/35053.asp#sHJYAAo1kfzkCLM2.99 (viewed 11 Feb. 2014).

148 Narayanan, "Personalized Coupons as a Vehicle for Perfect Price Discrimination."

149 Mark Huffman, "Are Digital Coupons the New Way to Shop for Discounts?" Consumer Affairs, 30 Sept. 2013, http://www.consumeraffairs.com/news/are-digital-coupons-the-new-way-to-shop-for-discounts-093013.html. Coupons are appearing on phones as you enter a store. See, for example, Mark Walsh, "MLB Tests iBeacon App Feature To Send Mobile Coupons," Online Media Daily, 26 Sept. 2013, http://www.mediapost.com/publications/article/210096/mlb-tests-ibeacon-app-feature-to-send-mobile-coupo.html#axzz2iAQPfc9r (both viewed 11 Feb. 2014). Customers who get regular bills or statements can get cross-promotional offers delivered that use their data to make them targeted offers.

150 For example, AcquireWeb says that its has linked "over 120 million existing cookies to geographic 'micro-zones' (groups of zip9s). This allows us to leverage traditional offline consumer data to target online display advertising campaigns. … Advertising at the ZIP+4 or neighborhood level to +120 million unique consumer devices tagged with actionable cookies … ." AcquireWeb, "Prospect Display Targets via Cookie Targeting," http://www.acquireweb.com/prospect-data-services/prospect-display-targets-via-cookie-targeting/. It combines what it calls "geo-based target information with consumer behavior … use registration data, IP location data, census data, your customer data." It also engages in an increasingly used practice that identifies online and offline information on a consumer, called "data append." "Reverse Append takes advantage of the fact that for many online marketers, the email address is the only identifier within their online database. To find postal addresses, Reverse Append matches your email-only file with the AcquireWeb database of over 750 million records that include email addresses as well as name and postal address. Where an email address matches, Reverse Append will return a file containing the name and deliverable postal address of the individual at the given email address." AcquireWeb, "Reverse Append," http://www.acquireweb.com/customer-data-services/reverse-append/ (both viewed 11 Feb. 2014).

151 Placed, "Placed Targeting," http://www.placed.com/targeting (viewed 11 Feb. 2014).

152 Advances in geo-mapping and the rapid adoption of mobile device use are also creating a new "fourth dimension"—our personal "grid" where we can be influenced for products and services as we "simultaneously live in adjacent online and offline realities." As one marketer describes it, "this fourth dimension acts as a custom zip code … . [The] personal grid uses data to hone in on individual preference." NBCUniversal Integrated Media, "Personal Grid," The Curve Report, vol. 2, 2012, http://thecurvereport.com/category/trends/personal-grid-1/ (registration required).

153 FICO gives this as an example: "For instance, if there is a Pizza Hut on the way from a person's office to home, she is more likely to redeem a Pizza Hut offer than a person who does not pass by Pizza Hut." Shafi Rahman and Amit Sowani, "Location-Based Marketing Using GPS Data," FICO Labs Blog, 5 Feb. 2013, http://ficolabsblog.fico.com/2013/02/location-based-marketing-using-gps-data.html (viewed 11 Feb. 2014).

154 Marketers are taking advantage of "digital tagging" and so-called "check-ins" so they can be a part of our online and physical environments. They predict that "in the future, digital storefronts … will seamlessly sync up with … physical environments and … more naturally intersect" our lives. NBCUniversal Integrated Media, "Personal Grid."

155 FICO, "FICO Analytic Offer Manager," July 2012, personal copy on file with author Jeff Chester. See also FICO Labs, "Choosing the Right Analytics: Uplift Models," FICO Labs Blog, 13 Sept. 2013, http://ficolabsblog.fico.com/2013/09/choosing-the-right-analytics-uplift-models.html; FICO, "Which Retail Analytics Do You Need?" The same technological foundation for financial marketing, with many of the same data companies involved, is also helping provide the tools for retailers and others. Chains and others, for example, can take advantage of Merkle's "Digital Data Integration Engine" to merge their lists with online data. Merkle, "Solutions," http://www.merkleinc.com/industry-solutions/retail-consumer-goods/solutions. Merkle explains that "Utilizing segmentation and model scores to inform targeting strategy has thus escaped the confines of traditional offline marketing and can now be leveraged digitally." Through such practices, companies know more "about *who* we are serving advertising to," but can "also *predict who among digital prospects is more likely*11 Feb. 2014

156 Alex Campbell, "Top 5 Emerging Trends in Mobile Commerce," The Future of Commerce, 1 July 2013, http://www.the-future-of-commerce.com/2013/07/01/emerging-mobile-trends/; Kyle Stock, "Maybe Showrooming Isn't Killing Retailers After All," Businessweek, 12 Sept. 2013, http://www.businessweek.com/articles/2013-09-12/maybe-showrooming-isnt-killing-retailers-after-all (both viewed 11 Feb. 2014).

157 For example, quick service restaurants such as McDonald's are investing in mobile payments and location marketing. Drug store chains have developed mobile apps to help sell products in-store. See, for example, Liddle, "What They Didn't Tell You about McDonald's' Mobile Payments Trials"; Chantal Tode, "Walgreens Tests Closed-loop Mobile Coupons Delivered via PRIMP App," Mobile Commerce Daily, 29 July 2013, http://www.mobilecommercedaily.com/walgreens-tests-closed-loop-mobile-coupons-delivered-via-primp-app (viewed 11 Feb. 2014). QSR's are predicted to spend $616 million dollars for local online advertising by 2017, up from $434 million in 2012. eMarketer, "US Quick-Service/Fast Food Restaurant Local Online Ad Spending, by Format, 2012 & 2017," 2013, personal copy on file with author Jeff Chester.

158 Edwards, "Meet The 30 Biggest Advertisers On Facebook"; eMarketer, "The U.S. Financial Services Industry 2013," June 2013, personal copy on file with author Jeff Chester.

159 Recently the Navy Federal Credit Union sold $200 million in banking services using Facebook. Running an integrated "membership appreciation" campaign on Facebook that featured a contest, the credit union promoted auto loans and the sales of certificates of deposit. It generated 60,000 new members and sold $90 million in CDs and 5,400 auto loans worth $96 million. It went from 520,000 to 829,000 "Likes" as a result. "Navy FCU Sells $200 Million in Banking Services on Facebook," The Financial Brand, 17 Sept. 2013, http://thefinancialbrand.com/33631/using-social-media-to-sell-banking-products/; Navy Federal Credit Union, "Success Stories: Celebrating Members to Attract New Ones," Facebook, https://www.facebook.com/facebookforbusiness/success/navy-federal-credit-union (both viewed 11 Feb. 2014).

160 "'Facebook ads are 90% accurate with our native targeting products—using geo, demo, interest, smartphone, etc., as variables. We can layer this targeting with the bank's data to gain even more efficiency,' Hiltz explains. 'The "match rates" between the bank data tables and Facebook audience tables are contingent upon the quality of the bank's dataset … . We can also work with trusted third-party data providers. We have existing relationships with Acxiom, Epsilon, and DataLogix, and are signing up even more data providers going forward.'" "Facebook Advertising and the Future of Financial Marketing," The Financial Brand, 2 Oct. 2013, http://thefinancialbrand.com/33967/facebook-advertising-in-banking/. For Facebook apps that facilitate the transfer of money, see "Your Facebook Account Can Be Friends With Your Bank Account," The Financial Brand, 30 Sept. 2013, http://thefinancialbrand.com/33918/icici-bank-facebook-banking-app/. See also "Two Ways Financial Marketers Can Fuel Sales in Social Channels," The Financial Brand, 25 Sept. 2013, http://thefinancialbrand.com/33740/generating-social-media-revenue-roi/ (all viewed 11 Feb. 2014).

161 The Discover Card is a client of Social Amp, now owned by Merkle. " Merkle Acquires Social Amp, Leading Facebook Open-Graph Developer," Feb. 2012, http://www.merkleinc.com/news-and-events/press-releases/2012/merkle-acquires-social-amp-leading-facebook-open-graph-developer; Social Amp, http://www.socialamp.com/ (both viewed 11 Feb. 2014). As Merkle explains, "by using Facebook Open Graph within applications on and off Facebook, marketers can gain access to profile and behavioral information that can be included in customer event streams.". Jennifer Veesenmeyer, Peter VanDre, Ron Park, and Andy Fisher, It Only Looks Like Magic: The Power of Big Data and Customer-Centric Digital Analytics (Merkle, 2013), p. 147, http://www.merkleinc.com/what-we-do/database-marketing-services/analytics/it-only-looks-magic (purchase required).

162 In one example for MasterCard, a company used customers' email addresses, phone numbers, and Facebook data to target them. Ampush, "Case Study: Financial Services," http://ampush.com/wp-content/uploads/2013/07/Financial-Services-Client-1.pdf?__hstc=137463663.29a942d23f8e93dea122cb62b5d06c1a.1382465115028.1382465115028.1382500680412.2&__hssc=137463663.1.1382500680412&__hsfp=2950800491 (viewed 11 Feb. 2014).

163 FourthWall Media, "MassiveData," http://www.fourthwallmedia.tv/MassiveData/ (viewed 11 Feb. 2014).

164 Marc Trudeau and Stephen Drees, "Next-Gen Card Marketing: Meeting the Expectations of New Consumers … And Your Existing Ones," Acxiom, 2013, http://www.paymentssource.com/media/pdfs/CFE13-Acxiom.pdf.The process is described this way: "You don't use programmatic just to buy an ad on PCWorld, for example. It has to do with whomever just arrived at that site, the data and profile they fit in, their IP address, plus what their cookie says about them, and, for example, if they've also been to Cisco.com, Intel.com and IBM.com." Christopher Hosford, "Programmatic Ad Buying Gains Momentum," B2B, 16 Sept. 2013, http://adage.com/article/btob/programmatic-ad-buying-gains-momentum/290285/ (both viewed 11 Feb. 2014).

165 Federal Trade Commission, "FTC Native Advertising Workshop on December 4, 2013 Will Explore the Blurring of Digital Ads With Digital Content," 16 Sept. 2013, http://www.ftc.gov/opa/2013/09/nativeads.shtm. For an overview of ad exchanges, see, generally, Ad Exchanger, http://www.adexchanger.com/ (both viewed 11 Feb. 2014).

166 See, generally, "Berkeley Consumer Privacy Survey," http://www.law.berkeley.edu/privacysurvey.htm (viewed 11 Feb. 2014). There is also misinformation about the role of traditional financial products, such as credit scoring. According to one industry and consumer group survey, 40 percent of consumers are unaware that "a credit score is used in decisions about credit availability and pricing of credit cards." Burns, "Expand & Grow: How to Reach and Educate More Members Using a New Breed of Credit Scores."

167 eMarketer, "Consolidation of Mobile Payments Landscape Will Drive Uptake," 25 July 2013, http://www.emarketer.com/Article/Consolidation-of-Mobile-Payments-Landscape-Will-Drive-Uptake/1010074 (viewed 11 Feb. 2014).

168 Groups would be encouraged to engage with the media when important news or development occurs, such as with the pending new consumer study conducted by NORC at the University of Chicago funded by the Federal Reserve that will address underbanked issues. NORC at the University of Chicago, "2013 Survey of Consumer Finances," http://scf.norc.org/ (viewed 11 Feb. 2014).

169 MCX, "Merchant Customer Exchange," http://www.mcx.com/ (viewed 11 Feb. 2014).

170 Mobey Forum, "Mobey Forum Launches North American Chapter to Advance Mobile Financial Services Ecosystem," 16 July 2012, http://www.mobeyforum.org/mobey-forum-launches-north-american-chapter-to-advance-mobile-financial-services-ecosystem/(viewed 11 Feb. 2014).

171 Smart Card Alliance, "Mobile and NFC Council," http://www.smartcardalliance.org/pages/activities-councils-mobile-and-nfc-council (viewed 11 Feb. 2014), which includes a downloadable presentation entitled "ThePowerPoint Standards for the NFC Ecosystem."

172 Federal Reserve Bank of Boston, "Mobile Payments Industry Workgroup," http://www.bostonfed.org/bankinfo/payment-strategies/mpiw/. In an August 2013 presentation for state legislators, the agency noted that among the challenges facing the industry were "ownership of customer data, lack of regulatory direction, fragmentation with diverse nonbank businesses, lack of interoperability and standards, and multiple stakeholders." Marianne Crowe, "U.S. Mobile Payments Landscape," NCSL Legislative Summit 2013, 13 Aug. 2013, http://www.ncsl.org/documents/standcomm/sccomfc/MarianneCrowe_PowerPoint.pdf (both viewed 11 Feb. 2014).

173 Online Lenders Alliance, http://www.onlinelendersalliance.org/ (viewed 11 Feb. 2014).

174 Martinne Geller and David Henry, "Wal-Mart, Amex Take on Banks with Low-priced Debit Card," Reuters, 8 Oct. 2012, http://www.reuters.com/article/2012/10/08/us-walmart-amex-idUSBRE8970H520121008 (viewed 11 Feb. 2014).

175 "Dollar General," Google Play, 10 July 2013, https://play.google.com/store/apps/details?id=com.dollargeneral.android&hl=en; Target, "Mobile," http://www.target.com/spot/mobile/landing (both viewed 11 Feb. 2014).

176 Lauren Johnson, "Walmart App Users Spend 40pc More than Average Shopper," Mobile Commerce Daily, 26 Sept. 2013, http://www.mobilecommercedaily.com/Walmart-app-users-spend-40-percent-more-than-average-shopper (viewed 11 Feb. 2014).

177 Johnson, "Walmart App Users Spend 40pc More than Average Shopper."

178 Lauren Johnson, "Target Tightens Focus on Mobile as In-store Shopping Tool," Mobile Commerce Daily, 30 Aug. 2013, http://www.mobilecommercedaily.com/target-enhances-in-store-mobile-experience-with-weekly-ads-beta-shopping-list (viewed 11 Feb. 2014).

179 "Walmart continues to ramp up its mobile in-store Scan & Go program by giving users the ability to clip coupons by tapping their smartphones and having the savings automatically applied when they check out. Scan & Go, a feature on the Walmart application, enables users to scan merchandise in certain stores and pay at a self-checkout counter." Chantal Tode, "Walmart Boosts Scan & Go Self-checkout with Mobile Coupons," Mobile Commerce Daily, 2 Aug. 2013, http://www.mobilecommercedaily.com/Walmart-boosts-scan-go-self-checkout-with-mobile-coupons (viewed 11 Feb. 2014).

180 "Principal Software Eng WalmartLabs," Walmart eCommerce, http://jobs.walmart.com/brisbane/ecommerce/jobid3859437-personalization-systems-engineer-jobs (viewed 11 Feb. 2014).

181 Patrick Harrington, "Targeting @WalmartLabs," The Official @WalmartLabs Blog, 26 Nov. 2012, http://walmartlabs.blogspot.com/2012/11/targeting-walmartlabs.html. See also Arun Prasath, "The @WalmartLabs Social Media Analytics Project," The Official @WalmartLabs Blog, 11 Jan. 2013, http://walmartlabs.blogspot.com/2013/01/the-walmartlabs-social-media-analytics.html; Nandita Chakravarti "User Research: Walmart and Amex Product—Bluebird," Behance, http://www.behance.net/gallery/User-Research-Walmart-and-Amex-product-Bluebird/7408619 (all viewed 11 Feb. 2014).

182 Kelly Liyakasa, "The @WalmartLabs Way: Why The Online Pure-Play Needs Brick and Mortar (And Vice Versa)," Ad Exchanger, 4 Sept. 2013, http://www.adexchanger.com/ecommerce-2/the-Walmartlabs-way-why-the-online-pure-play-needs-brick-and-mortar-and-vice-versa/#more-81643 (viewed 11 Feb. 2014).

183 James Taylor, "Using Predictive Analytics and Decision Management in Retail," Nov. 2012, personal copy on file with author Jeff Chester. Dylan Tweney, "Walmart Scoops up Inkiru to Bolster its 'Big Data' Capabilities Online," Venture Beat, 10 June 2013, http://venturebeat.com/2013/06/10/walmart-scoops-up-inkiru-to-bolster-its-big-data-capabilities-online/ (viewed 11 Feb. 2014).

Information Technology Industry Council

Submission to the White House Office of Science and Technology Policy

Response to the Big Data Request for Information

Comments of the Information Technology Industry Council

March 27, 2014

I. Introduction

The Information Technology Industry Council (ITI) appreciates this opportunity to provide comments to the Office of Science and Technology Policy (OSTP). ITI is a U.S.-based global trade association representing more than 50 of the world's most dynamic and innovative companies in the information and communications technology (ICT) sector.

Following months of revelations relating to the nation's surveillance programs, on January 17, 2014, President Obama delivered remarks outlining his plans for surveillance reform. Both prior to January 17, and since the President's remarks, ITI provided written comments to the administration, the President's Review Group on Intelligence and Communications Technologies (Review Group), and the Privacy and Civil Liberties Oversight Board (PCLOB).[1] In these comments, ITI outlined the significant negative economic impact that the revelations are having on the technology sector due to the erosion of public trust, as well as the potential long-term implications for innovation and Internet governance on the global economy. ITI recommended

[1] *See* letter to White House Chief of Staff and White House Counsel (August 20, 2013), *available at* http://www.itic.org/public-policy/TechLetteronWaystoProtectCivilLibertiesinGov%E2%80%99tDataCollection.pdf; comments to the Review Group (October 3, 2013) a*vailable at* http://www.itic.org/upload/ITISIIALettertoReviewGroup.pdf; and comments to PCLOB (October 24, 2013), *available at* http://www.itic.org/dotAsset/6/9/697dec86-0d33-448c-9798-960e172a6dc5.pdf.

specific steps that the U.S. government could take to restore public trust, including greater transparency and oversight. In recent testimony before the House Judiciary Committee, and PCLOB, ITI's President and CEO stressed the critical need for reforms.[2] ITI continues to urge the administration and Congress to work together to implement the necessary reforms to restore trust in the innovative products and services that ITI member companies provide, and to maintain the open and borderless Internet that benefits so many individuals, companies, and countries around the world.

In his January 17, 2014 remarks, President Obama also announced that he had appointed John Podesta to conduct a comprehensive review of big data. Recognizing that this review of big data is distinct from the issues in connection with surveillance reform, ITI's comments today relate specifically to this big data review. ITI appreciates the opportunity to provide input on this important topic.

II. Big data

Both the U.S. government and the private sector recognize the potential capabilities of large-scale data analytics. In 2012, the administration announced the National Big Data Research and Development Initiative to improve the "ability to extract knowledge and insights from large and complex collections of digital data."[3] That initiative is committed to helping "accelerate the pace of discovery in science and engineering, strengthen our national security, and transform teaching and learning."

---

[2] *See* testimony of Dean C. Garfield before the House Judiciary Committee (February 4, 2014) *available at* http://www.itic.org/media/news-releases/testimony-of-iti-president-dean-c-garfield-before-the-house-judiciary-committee-regarding-fisa-reform, and testimony of Dean C. Garfield before the PCLOB (March 19, 2014) *available at* http://www.itic.org/media/news-releases/testimony-of-iti-president-and-ceo-dean-garfield-before-the-privacy-and-civil-liberties-oversight-board.

[3] *See* Press Release, White House Office of Science and Technology, Obama Administration Unveils "Big Data" Initiative: Announces $200 Million in New R&D Investments (March 29, 2012) *available at* http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf.

Across the U.S. government, agencies are examining how to derive maximum benefit from data. Last month, the National Oceanic and Atmospheric Administration (NOAA) issued a request for information outlining its interest to "unleash the power of its data" and seeking assistance from the private sector to make NOAA's data available in a "rapid, scalable manner to the public."[4] We encourage the administration to continue its work on maximizing the greatest benefits from big data.

Innovation across sectors, including ICT, depends on the value derived from large-scale data analytics. Specific societal benefits of big data were highlighted in the March 3, 2014 conference at the Massachusetts Institute of Technology (MIT) convened as part of OSTP's big data review. Panelists discussed how big data is enabling tremendous benefits in areas such as medical care, transportation, and education. As acknowledged by John Podesta in his keynote address:

> *The value that can be generated by the use of big data is not hypothetical. The availability of large data sets, and the computing power to derive value from them, is creating new business models, enabling innovations to improve efficiency and performance in a variety of public and private sector settings, and making possible valuable data driven insights that are measurably improving outcomes in areas from education to healthcare.*[5]

[4] *See* Press Release, NOAA, NOAA announces RFI to unleash power of 'big data' Agency calls upon American companies to help solve 'big data' problem *available at* http://www.noaanews.noaa.gov/stories2014/20140224_bigdata.html.

[5] *See* Remarks as Delivered by Counselor John Podesta The White House/MIT "Big Data" Privacy Workshop (March 3, 2014) *available at* http://www.whitehouse.gov/sites/default/files/docs/030414_remarks_john_podesta_big_data.pdf.

While the benefits of big data are considerable, the question has been posed—by OSTP, as well as others—whether existing policy frameworks for protecting consumer privacy sufficiently address the privacy issues that could be implicated by big data. At the same time that this issue is examined, it is fundamentally important that we more fully consider and understand the significant and transformational benefits that big data can have on individuals: improvements in health care through application of personalized medicine; improved living conditions though enhanced urban planning and development; environmental advancements through enhanced sustainable consumption and more efficient use of energy; and countless other beneficial applications as well as innovative products and services.

As the White House conducts its examination of big data, there must be sufficient study of the beneficial applications as well as the potential risks. We encourage OSTP to fully examine the range of benefits and capabilities associated with big data. Without understanding the benefits, it is impossible to understand the possible opportunity cost of risk mitigation strategies. Only with a complete picture can policy approaches to big data be optimized. Because of the capabilities of large-scale data analytics, responsibility requires being mindful of how data is being used, but also what the implications are for not using it.

III. Policy Frameworks

As OSTP examines how existing privacy frameworks can address the issues raised in connection with big data, we agree that such analysis should begin with the privacy framework approach outlined in the 2012 White House report, which includes a "Consumer Privacy Bill of Rights" that incorporates elements of the Fair Information Practice Principles.[6]

[6] White House, Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy (Feb. 2012), ("White House Report") *available at*

As recognized in the White House Report, the "existing consumer data privacy framework in the United States is flexible and effectively addresses some consumer data privacy challenges in the digital age."[7] This framework, which includes sector-specific laws enforced by a number of different U.S. agencies, Federal Trade Commission (FTC) enforcement under Section 5 of the FTC Act,[8] industry best practices, and self-regulatory initiatives, has shown a level of adaptability to technological innovation. We urge OSTP to identify the strengths of the existing privacy framework as it examines big data policy considerations and risks to privacy.

Below, ITI identifies a number of areas that OSTP can consider as it develops a road map of the policy issues to be considered in connection with big data.

A. Risk minimizing technologies and policies

As discussed at the MIT Workshop, big data encompasses predicated data analysis (uncovering results from what is known to be available in the data), to non-predicated analysis, where big data can reveal new insights or patterns not known to exist within the data prior to the analysis. To the extent certain protections as outlined in the Consumer Privacy Bill of Rights are challenged in these contexts, technological tools and policies may be useful in minimizing privacy risks.

---

http://www.whitehouse.gov/sites/default/files/privacy-final.pdf. The Fair Information Practice Principles (FIPPs), which include the concepts of notice, choice, data minimization, purpose specification, and use limitation, serve as the foundation for both policy and regulatory frameworks for privacy. FIPPs are incorporated into the privacy frameworks developed by multilateral fora, such as the Organization for Economic Cooperation and Development (OECD) and the Asia Pacific Economic Cooperation (APEC) forum. The FIPPs can also be seen in legislative frameworks, such as the European Union's privacy regulatory framework, as well as certain U.S. privacy laws.

[7] White House Report, at 6.

[8] 15 U.S.C §§ 41-58, as amended.

For example, technological tools that can render personal data pseudonymous, anonymous, or de-identified is one way of addressing privacy risks. Further technological research into the development of the robust tools in this area is necessary. When the analysis of data (whether it is the entirety of the data being examined, or a subset thereof) in an anonymous, pseudonymous or de-identified form can be achieved through technological means without adversely affecting the quality of the results derived from the analysis, privacy risks can be minimized. Such analysis would not implicate the privacy concerns that can be present when data is linked to individuals. We further note that the measures that an organization should take to anonymize, pseudonymize, or de-identify data will necessarily depend on the intended use of the data, as well as the available methodology and technology for such modifications. As recognized by the Federal Trade Commission, "the nature of the data at issue and the purposes for which it will be used are also relevant" in determining what would constitute an organization achieving a "reasonable level of justified confidence that the data cannot reasonably be used to infer information about, or otherwise be linked to, a particular consumer, computer, or other device."[9] We further note that an organization's practices with respect to the data will inform the measures necessary to achieve that "reasonable level of justified confidence."

B. Risk-based approach to big data use and analysis

The development of a risk-based analysis approach to determine whether a proposed use or analysis of data is appropriate should be considered as a methodology to minimize the privacy impact on individuals, particularly where certain

[9] See FTC, Protecting Consumer Privacy in an Era of Rapid Change, Recommendations for Businesses and Policy-makers, FTC Report ("FTC Privacy Report") (Mar. 2012), p. 21, *available at* http://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-report-protecting-consumer-privacy-era-rapid-change-recommendations/120326privacyreport.pdf.

FIPPs protections are impractical. A risk assessment, that could be based on a common set of factors, might include the type of data, how the data was amassed, the public interest in the use of the data, the benefits of the use, the security measures in place, and the potential harmful impact to individuals resulting from the use. This risk assessment exercise could serve not only as a determination as to whether a particular use or analysis should go forward, but also to implement privacy-protecting safeguards. Through this risk assessment that considers privacy at the outset—essentially, a privacy impact assessment—data analyzers could determine a proper balance of risk minimization and maximum benefit from the data.

C. Accountability

The well-known concept of "accountability" in the privacy realm applies equally in the big data context, and would be useful in connection with both the risk-minimizing technologies and risk-based approach to big data use and analysis discussed earlier. Generally, accountability requires that organizations develop and put into place processes that foster compliance with their privacy-related commitments. Further, it requires organizations to describe how their processes comport with those commitments and be prepared to demonstrate them.

For example, accountability would require organizations to develop processes—and be able to describe such processes—to determine if and when anonymization, pseudonymization, de-identification and other privacy-protecting measures are appropriate. The accountability requirement would similarly apply in the risk assessments that data analyzers would conduct in connection with intended use of data. Organizations would need to develop processes—and be able to describe such processes—related to the conducting of risk assessments based on commonly understood use-based best practices that could be developed in connection with big data.

D. Consumer education

"Big data"—the term of art generally used to represent large-scale data analytics—is not easily understood by consumers. Improved consumer education on how data analytics are being used to provide benefits in a multitude of areas, such as health, transportation, and medicine should be developed. Long disclosures that are not easily understood may not be the most effective way to inform consumers about data practices. Research into the scope of information to be shared with consumers and how that information can be effectively imparted should be a priority, and we encourage the administration to support such research.

IV. International landscape

Big data magnifies the already challenging international environment where barriers to cross-border data flows impede the information that may be available to emerging big data services that can be provide benefits to individuals. Various types of global restrictions to data flows, from localization requirements to overly restrictive data protection requirements, may also interfere with the potential for big data. Data localization requirements limit the availability of data sets for uses based on geographic location and data protection requirements may constrain collection or use of information. While there are legitimate reasons for data protection regulations, they need not be written or implemented in a way that overly constrains collection or use of information for legitimate purposes.

We urge the administration to resist efforts by other jurisdictions to impose data localization restrictions, overly restrictive data protection regimes, and barriers to cross-border data flows. We further encourage the administration to support mechanisms that enable and facilitate cross-border data flows. For example, the U.S.-EU Safe Harbor enables the transfer of data from the U.S. to the EU and the continued

availability of this mechanism is critical across industry sectors—we urge the administration to ensure the continued availability of this mechanism.

Other international efforts to develop interoperable data protection and cross-border data flow regimes should also be supported by the administration. For example, the administration should continue its important work within the Asia Pacific Economic Cooperation forum (APEC) to promote APEC's Cross-Border Privacy Rules (CBPRs).[10] The APEC CBPRs are a set of privacy rules that are consistent with the APEC Privacy Framework, and CBPRs were developed to provide a flexible way for companies to demonstrate their trustworthiness and accountability for personal information. The ultimate goal is for participating companies to be able to transfer data within the APEC region without impediments.

Such cross-border transfer facilitation mechanisms are critical in the big data context in that they allow for the availability of data across jurisdictions and enable large-scale data analytics. The continued development of these transfer mechanisms should be a priority for the administration. A recent collaborative effort involving APEC CBPR's and one of the EU's data transfer mechanisms—Binding Corporate Rules—demonstrates progress in identifying the compatible elements of differing privacy systems.[11] More work needs to be done in this area—and should be supported by the administration—to identify common ground among differing systems in efforts to

[10] *See* APEC, APEC Cross-Border Privacy Rules System, Policies, Guidelines and Procedures, *available at* http://www.apec.org/Groups/Committee-on-Trade-and-Investment/~/media/Files/Groups/ECSG/CBPR/CBPR-PoliciesRulesGuidelines.ashx.

[11] *See* Joint Work between experts from the Article 29 Working Party and from APEC Economies, on A referential for requirements for Binding Corporate Rules submitted to national Data Protection Authorities in the EU and Cross Border Privacy Rules submitted to APEC CBPR Accountability Agents, (March 6, 2014) *available at* http://www.apec.org/~/media/Files/Groups/ECSG/20140307_Referential-BCR-CBPR-reqs.pdf.

promote interoperability.

* * *

ITI appreciates the opportunity to submit these comments to OSTP. If you have any questions about these comments, please contact Yael Weinman, VP, Global Privacy Policy and General Counsel, Information Technology Industry Council, at 202-626-5751, yweinman@itic.org.

Consumer Federation of America
1620 I Street, N.W., Suite 200 * Washington, DC 20006

March 28, 2014

Nicole Wong
Deputy Chief Technology Officer
Office of Science and Technology Policy
Eisenhower Executive Building
1650 Pennsylvania Ave. NW
Washington, DC 20502

Re: Big Data RFI

Dear Deputy Wong:

Consumer Federation of America (CFA), an association of nearly 300 nonprofit consumer organizations across the United States, offers the following brief comments in response to your office's Request for Information concerning "big data."[1]

The ever-increasing amount and scope of data that can be collected about or linked to individuals (which we will refer to as "personal data") has profound implications for consumers and for our society. While the collection, analysis, retention and use of personal data can bring benefits, there are also risks that must be considered and which current U.S. policy does not adequately address. As the data and its applications grow "bigger," so do our concerns in that regard.

The Administration acknowledged some of these concerns in its 2012 report about data privacy.[2] In the preface, President Obama noted that "Never has privacy been more important than today, in the age of the Internet, the World Wide Web and smart phones" and that much of the innovation that these new technologies have fostered "is enabled by novel uses of personal information." While the document focused on the commercial sphere, the President also mentioned the importance of these technologies in enabling individuals to engage in political discourse. Citing the need for consumer trust and the lack of a "clear statement of basic privacy principles that apply to the commercial world, and a sustained commitment of all stakeholders to address consumer data privacy issues as they arise from advances in technologies and business models,"[3] the report called for a Consumer Privacy Bill of Rights, embodying basic privacy principles that can be adapted in the digital era. This was envisioned as a framework not only to spur self-regulatory initiatives but as the basis for legislation that the Administration would put forward to give Americans actual rights. To date, however, no such legislation has emerged.

While we applaud the Administration for launching this new discussion about big data, we are worried that the train is leaving the station. Business models are already proliferating using big data in ways that

[1] 79 FR 12251, March 4, 2014.
[2] *Consumer Data Privacy in a Networked World: a Framework for Protecting Consumer Privacy and Promoting Innovation in the Global Digital Economy*, http://www.whitehouse.gov/sites/default/files/privacy-final.pdf.
[3] Id, see Foreword.

raise concerns about privacy, discrimination, and other issues, and it is clear that existing U.S. law does provide sufficient protection.[4]

Those concerns go beyond consumers and commerce, as big data can be used to categorize individuals for governmental and other purposes. CFA endorses the *Civil Rights Principles for the Era of Big Data*[5] recently issued by a coalition of civil rights organizations.

With continuing advances in technology, it is urgent to address these concerns. Facial recognition, for instance, is rapidly improving and is being offered today for a variety of purposes[6] which could lead to the loss of individual autonomy, manipulative marketing, and other undesirable consequences. The National Telecommunications and Information Administration's multistakeholder process to develop voluntary industry codes of conduct and other self-regulatory efforts, while potentially helpful, are not enough.

What we need is a comprehensive legal framework regarding personal data in the U.S. to guide the development and deployment of big data in a way that comports with our societal values. This would also help to promote U.S trade abroad, which is presently hindered by adequate U.S. privacy laws.[7]

CFA is joining other groups on separate comments in response to this Request for Information. Those comments will make more detailed recommendations in specific areas.

Whatever emerges from this examination of big data, it is crucial for the U.S. Administration to acknowledge the concerns as well as the benefits and to propose concrete steps to address those concerns. This should include a restatement of the need for a comprehensive privacy law and a commitment to introduce such legislation before the end of this year. CFA would welcome the opportunity to play an active role on this and other initiatives that may come out of this exercise.

Sincerely,

Susan Grant, Director of Consumer Protection
Consumer Federation of America

---

[4] See, for instance, *Big Data Means Big Opportunities and Big Challenges*, by the Center for Digital Democracy and U.S. PIRG, http://www.centerfordigitaldemocracy.org/sites/default/files/USPIRGEFandCDDBigDataReportMar14_1.3web.pdfand *Big Data, a Big Disappointment for Scoring Consumer Credit Risk*, by the National Consumer Law Center, http://www.nclc.org/images/pdf/pr-reports/report-big-data.pdf.

[5] See http://www.civilrights.org/press/2014/civil-rights-principles-big-data.html.

[6] See new report by the Center for Digital Democracy at http://www.centerfordigitaldemocracy.org/sites/default/files/NTIAFacialRecognitionCDDFinal032514.pdf.

[7] See critical report on the US/EU Safe Harbor program at http://www.europarl.europa.eu/document/activities/cont/201310/20131008ATT72504/20131008ATT72504EN.pdf and European Commission report at http://ec.europa.eu/justice/data-protection/files/com_2013_847_en.pdf. This program is meant to enable flows of EU consumers' personal information when they engage in cross-border commerce with US companies, in light of the fact that U.S. privacy law has not been deemed "adequate" to meet EU data protection requirements. It has many inherent shortcomings, however, and is not working well. How to deal with cross-border data flows given the wide gaps between U.S. and EU privacy law is also contentious issue in the Transatlantic Trade and Investment Partnership negotiations. See also CFA's July, 2013 presentation to the negotiators at http://www.consumerfed.org/pdfs/TTIP%20presentation%207.10.13.pdf.

The Leadership Conference on Civil and Human Rights

1629 K Street, NW
10th Floor
Washington, DC
20006

202.466.3311 voice
202.466.3435 fax
www.civilrights.org

The Leadership Conference

March 28, 2014

Nicole Wong
Attn: Big Data Study
Office of Science and Technology Policy
Eisenhower Executive Office Building
1650 Pennsylvania Ave. NW
Washington, DC 20502

Dear Ms. Wong,

On behalf of The Leadership Conference on Civil and Human Rights, we appreciate this opportunity to provide comments in response to the Office of Science and Technology Policy's Request for Information (RFI) regarding "Big Data." The Leadership Conference is a coalition charged by its diverse membership of more than 200 national organizations to promote and protect the civil and human rights of all persons in the United States.

The Leadership Conference is pleased to join many other civil rights and media justice organizations in supporting the attached Civil Rights Principles for the Era of Big Data. Released to the public on February 27, 2014, these principles represent the first time that national civil and human rights organizations have spoken publicly about the importance of privacy and big data for communities of color, women, and other historically disadvantaged groups.

Big data is a civil and human rights issue. These new technologies have the potential to improve the lives of all Americans, and at the same time they pose new risks to civil and human rights that may not be addressed by our existing legal and policy frameworks.

Through these principles, we and the other signatory organizations highlight the growing need to protect and strengthen key civil rights protections in the face of technological change. We call for an end to high-tech profiling; urge greater scrutiny of the computerized decisionmaking that shapes opportunities for employment, health, education, and credit; underline the continued importance of constitutional principles of privacy and free association, especially for communities of color; call for greater individual control over personal information; and emphasize the need to protect people, especially disadvantaged groups, from the documented real-world harms that follow from inaccurate data.

Thank you for embarking on this important process. We stand ready to work with you to ensure that the voices of the civil and human rights community are heard in this important, ongoing national conversation. If you have any questions about these comments, please contact Corrine Yu, Leadership Conference Managing Policy Director, at 202-466-5670 or yu@civilrights.org.

Sincerely,

Wade Henderson
President & CEO

Nancy Zirkin
Executive Vice President

Attachment

Civil Rights Principles for the Era of Big Data

February 2014

Technological progress should bring greater safety, economic opportunity, and convenience to everyone. And the collection of new types of data is essential for documenting persistent inequality and discrimination. At the same time, as new technologies allow companies and government to gain greater insight into our lives, it is vitally important that these technologies be designed and used in ways that respect the values of equal opportunity and equal justice. We aim to:

1. **Stop High-Tech Profiling**. New surveillance tools and data gathering techniques that can assemble detailed information about any person or group create a heightened risk of profiling and discrimination. Clear limitations and robust audit mechanisms are necessary to make sure that if these tools are used it is in a responsible and equitable way.
2. **Ensure Fairness in Automated Decisions**. Computerized decisionmaking in areas such as employment, health, education, and lending must be judged by its impact on real people, must operate fairly for all communities, and in particular must protect the interests of those that are disadvantaged or that have historically been the subject of discrimination. Systems that are blind to the preexisting disparities faced by such communities can easily reach decisions that reinforce existing inequities. Independent review and other remedies may be necessary to assure that a system works fairly.
3. **Preserve Constitutional Principles**. Search warrants and other independent oversight of law enforcement are particularly important for communities of color and for religious and ethnic minorities, who often face disproportionate scrutiny. Government databases must not be allowed to undermine core legal protections, including those of privacy and freedom of association.
4. **Enhance Individual Control of Personal Information**. Personal information that is known to a corporation — such as the moment-to-moment record of a person's movements or communications — can easily be used by companies and the government against vulnerable populations, including women, the formerly incarcerated, immigrants, religious minorities, the LGBT community, and young people. Individuals should have meaningful, flexible control over how a corporation gathers data from them, and how it uses and shares that data. Non-public information should not be disclosed to the government without judicial process.
5. **Protect People from Inaccurate Data**. Government and corporate databases must allow everyone — including the urban and rural poor, people with disabilities, seniors, and people who lack access to the Internet — to appropriately ensure the accuracy of personal information that is used to make important decisions about them. This requires disclosure of the underlying data, and the right to correct it when inaccurate.

Signatories:

American Civil Liberties Union
Asian Americans Advancing Justice — AAJC
Center for Media Justice
ColorOfChange
Common Cause
Free Press
The Leadership Conference on Civil and Human Rights
NAACP
National Council of La Raza
National Hispanic Media Coalition
National Urban League
NOW Foundation
New America Foundation's Open Technology Institute
Public Knowledge

Civil Rights and Big Data: Background Material

High-Tech Profiling

- The FBI has recently engaged in a racial and ethnic mapping program that uses crass racial and ethnic stereotypes to map American communities by race and ethnicity for intelligence purposes.
- Police in New York used license plate readers to record all the cars visiting certain mosques, allowing their movements to be tracked later. New technology made this surveillance cheap enough that it could happen without a clear policy mandate.
- Law enforcement can use new social media monitoring tools to investigate nearly anyone at low cost. These systems need audit records and usage rules to ensure they are used fairly.

Automated Decisions

- Financial institutions can now gather detailed information on trivial consumer missteps, such as a one-time overdraft, and use it to bar customers from opening bank accounts.
- A major auto insurer has begun to deny its best rates to those who often drive late at night, such as those working the night shift. The insurer knows each driver's habits from a monitoring device, which drivers must install in order to seek the insurer's lowest rate.

Constitutional Principles

- Information from warrantless NSA surveillance has been used by other federal agencies, including the DEA and the IRS — even though it was gathered outside the rules that normally bind those agencies.
- Databases like the so called “no fly” list are used to bar US citizens and legal residents from flying, without a fair process for reviewing these determinations.
- People who have access to government databases have often used them for improper purposes, including to leak confidential information about public figures and to review without reason the most intimate communications of strangers.

Individual Control of Personal Information

- New financial startups are using social network data and other "digital traces" to microtarget financial products. They claim to act outside the scope of existing consumer protections against unfair lending practices.
- Unscrupulous companies can find vulnerable customers through a new industry of highly targeted marketing lists, such as one list of 4.7 million “Suffering Seniors" who have cancer or Alzheimer’s disease.
- Some advertisers boast that they use web monitoring technologies to send targeted advertisements to people with bipolar disorder, overactive bladder, and anxiety.
- Location-aware social media tools have allowed abusive spouses and partners to learn the whereabouts of their victims in real time.

Risks of Inaccurate Data

- Government employment verification systems such as E-Verify demonstrate a persistently higher error rate for legal immigrants, married women, naturalized citizens, and individuals with multiple surnames (including many Hispanics) than for other legal workers, creating unjustified barriers to employment.
- Background check companies frequently provide inaccurate information on job candidates that stops them from being hired. While under law individuals are supposed to be able to correct these errors, they frequently recur and employers are not required to re-hire victims of misidentification.
- People often lose job opportunities due to criminal history information that is inaccurate, or that has nominally been expunged.

March 28, 2014

Nicole Wong
Deputy Chief Technology Officer
White House Office of Science and Technology Policy

Dear Ms. Wong,

The organizations represented on this letter are members of the Media Action Grassroots Network (MAG-Net). Collectively, our 175 members are working together for media change to end poverty, eliminate racism, and ensure human rights.

As members of MAG-Net, we believe that big data creates significant new risks of racial injustice. In order to ensure a fair and inclusive future for our nation's communities of color, and to enable the potential benefits of these new technologies to be fully realized and broadly shared, it is vitally important that the emerging policy framework for big data explicitly acknowledges and address issues of racial discrimination.

Much of the data that can be used to make important choices — in areas such as crime, lending, housing, education, and health — is deeply infected with racial bias. For example, decades of higher-intensity policing within communities of color have contributed to biased historical statistics on crime, by making it disproportionately likely that crimes committed among people of color will be reflected in the statistics. "Predictive policing" that is derived from these historical numbers may reinforce racial bias, exacerbating the daily reality that people of color will face unwarranted hostility and suspicion from law enforcement.

Institutional racism is similarly entrenched in the commercial marketplace. A long history of residential redlining (compounded in more recent years by discriminatory lending policies that shunted prime-worthy borrowers of color into subprime loans) has created a situation in which consumers of color are less likely than other consumers to possess the conventional hallmarks of financial health, even when they in fact are excellent credit risks. And big data marketing models that single out vulnerable groups such as "struggling seniors" or households that are "barely making it" — groups that are disproportionately comprised of people of color — can easily be used to target predatory products at these groups.

Apart from these entrenched biases, any constellation of thousands of data points will contain many that are close proxies for race. Such systems carry a pronounced risk of disparate impact, and their impacts must be scrutinized. Big data's advocates sometimes claim that in theory, data-driven methods might be used to track or remedy some of these inequalities, but one constant lesson of American history is that markets do not solve for racial justice. It is crucially important that big data systems include auditing and transparency controls to track their disparate impact across racial groups. Regardless of whether the data involved is big or small,

law and policy must protect against systems that differentially and adversely affect communities of color.

We support the Civil Rights Principles for the Era of Big Data, and join the signatories to those Principles — including major media policy, civil and human rights, and technology policy organizations — in urging you to ensure that the protection of civil and human rights for all residents plays an appropriately central role in the emerging law and policy framework for big data.

amalia deloney
Policy Director, Center for Media Justice

"Enclosure: [Civil Rights Principles for the Era of Big Data],"

The undersigned organizations:

1. Media Action Grassroots Network
2. Center for Digital Democracy
3. Alternate ROOTS
4. Media Mobilizing Project
5. Art is Change
6. Urbana Champaign Independent Media Center
7. Working Narratives
8. St. Paul Neighborhood Network
9. Organizing Apprenticeship Project
10. Women, Action & the Media
11. Media Alliance
12. The Greenlining Institute
13. Chicago Media Action
14. Media Literacy Project
15. Line Break Media
16. 1Love Movement
17. The Peoples Press Project

Cybersecurity and Privacy: The Challenge of Big Data[1]

Abraham R. Wagner
Columbia Law School
Columbia University, School of International and Public Affairs

Introduction

Recent history has seen both the rapid evolution of cyberspace, accompanied by an enormous expansion in terms of users and capabilities, as well as unprecedented technological, economic and social revolutions. These new technologies have also led to a virtual explosion in the amounts of data resident in servers and systems across the globe – often referred to as "big data." Along with a host of benefits, the era of big data has also brought with it a set of challenges in terms of security and privacy that increasingly affect the lives of Americans.[2]

Cyberspace has created a new venue for both crime and warfare. At the same time the availability of "big data" has also provided an opportunity for the commercial sector to analyze and utilize the data for non-criminal purposes which may still pose serious security and privacy questions. Increasingly the links between those that store "big data," commercial users and the Government have come under great public scrutiny while the courts are dealing with new cases where constitutional issues of privacy are being decided.[3]

Overall this paradigm shift is not simply one of technology, but embraces radical changes in the economics of information as well as the culture of modern society. This is easily the most significant change in media since the invention of moveable type in the 15th Century. While Americans have been quick to embrace the new technologies and capabilities they offer, public policy and the legal regime

---

[1] Comments in response to the Office of Science and Technology Policy, Government "Big Data" Request for Information, March 4, 2014.
[2] See Abraham Wagner, *Cybersecurity – From Experiment to Infrastructure* (Defense Dossier, 2012) and *Cybersecurity: New Threats and Challenges* (American Foreign Policy Council, 2013).
[3] See, for example, *In the Matter of the Search of Information Associated with [Redacted]@mac.com that is Stored at Premises Controlled by Apple, Inc.* (Magistrate Case No. 14-228 (JMF)). See also Orin Kerr, *Searches and Seizures in a Digital World*, 119 HARV. L. REV. 531 (2005).

are well behind and in need of serious effort.[4] Here the current laws are decades behind the current technologies and the problems that "big data" poses.[5]

It is also the case that Americans themselves see "big data" as well as the security and privacy concerns raised differently than in years past. Greater use of the technologies and increased awareness of potential problems has changed privacy expectations significantly. For their part both state and federal courts have responded to a myriad of cases with a far more encompassing view of the privacy protections afforded under the Fourth Amendment.[6] The current challenge is therefore multi-faceted. As both the government and the private sector continue to collect, analyze and utilize data norms, policies, and statutes are needed which address the privacy and security needs of Americans while promoting the free flow of information in ways that are consistent with these needs.

Evolution of Cyberspace and Big Data

The rapid evolution of cyberspace and the accompanying rise of "big data" has clearly been one of the greatest technological revolutions in recorded history. What began as a Defense Department experiment at the Advanced Research Projects Agency (ARPA – later DARPA) in the late 1960s has transformed almost all aspects of life with new technologies and an explosive growth in e-mail, the web and net-based applications never anticipated.

Security and privacy were not essential elements of the original ARPAnet design. At the outset the ARPAnet was an experiment in optimizing network resources with "switched packet" technology as an alternative to traditional "line switching." E-mail was not even a part of the concept; the web did not yet exist;

---

[4] Policy studies undertaken since the late 1990s have identified serious problems in the infrastructure, but the response by both the government and the commercial sector has proved to be grossly inadequate. See here PDD/NSC-63 *Critical Infrastructure Protection* (1998) and PPD-21 *Presidential Policy Directive - Critical Infrastructure Security and Resilience* (2013). It is striking that these two Presidential directives, coming well over a decade apart, come to the same conclusions with almost nothing have been done in between.

[5] As discussed at greater length below, one good example is the *Electronic Communications and Privacy Act (ECPA)* enacted in 1986 and codified at 18 U.S.C. §§ 2510–2522. The ECPA also added new provisions prohibiting access to stored electronic communications, i.e., the Stored Communications Act, 18 U.S.C. §§ 2701-12.

[6] See Harvey Rishikof and Abraham Wagner, *Cybersecurity and Cyberlaw* (Durham NC: Carolina Academic Press, 2014). See also Michael Warner, "Privacy and Security, Yesterday and Today," in *Cybersecurity and Privacy: Report of the Expert Workshop held for the Defense Advanced Research Projects Agency (DARPA)*(Arlington, VA: Institute for Defense Analyses, 2014).

there were no browsers or net-based content; and, there were no early commercial or national security applications.

Apart from DARPA's developmental work, a wide range of users including the Government, commercial firms, educational institutions and others acquired computers connected to various networks adding data at an exponential rate. With the transition to the Internet, networks were given low-cost global connectivity. For the first time in history, the marginal cost of worldwide communications fell to almost zero, as the "web" made it easier for users with new applications and web-based content growing exponentially.

Few entrants into cyberspace were aware of or cared about the myriad of security vulnerabilities which existed in operating systems, server software, middleware, application layers, router software and elsewhere. For well over a decade, the prevailing notion was that if there were problems, it must be somebody else's job to fix them.

Early Vulnerabilities and Security Efforts: The commercial world was quick to adopt the net, offer a vast range of applications, and generate "big data," but was largely unwilling and uninterested in paying to either secure it or provide privacy. Even banks failed to address the problem until they had been robbed of large sums. Government users were not much better as they quickly embraced cost-effective networked systems but failed to address critical vulnerabilities.

Internet programmers recognized vulnerabilities in operating systems as well as server design. Early attacks generally involved malware which disabled vulnerable computers and exploited data which was not protected, stealing larger amounts of data from servers connected to the net. Microsoft distributed "fixes" and "patches" to deal with some vulnerabilities while third party vendors like Norton sold security software that attempted to deal with a wider range of malware, installed firewalls, and gave users regular updates as new threats were identified.

These early entrants into the field saw the threat from malicious net activity and tried to protect users from malware, removing suspicious code such as viruses, worms and Trojans from infected computers. Other firms offered encryption software, such as PGP, enabling their users to protect sensitive files while a secure version of net protocol (:/https) enabled "secure" transactions over the web. In some ways cyberspace was becoming safer and more secure, but the adversarial threat was advancing at an even greater pace as well.

Growing Threats from Home and Abroad: Growth of e-commerce and "big data" brought new demands for privacy and security, while the proliferation of networked systems national security also required secure networks and applications to high standards. Vulnerabilities continued to be identified while new threats were seen on a daily basis. As the financial sector entered cyberspace, lucrative targets for cybercrime emerged as net-based theft from banks and credit card fraud became a booming business. "Big data" became both and a target and commodity.

While the early threats came largely from youthful hackers and disgruntled system administrators, the past decade has witnessed the evolution of far more serious cyber threats from expert criminals as well as well-trained military units assigned to cyberwarfare missions. Debate continues over the range of potential threats, ranging from denial of service to a type of apocalyptic attack often referred to as a "digital Pearl Harbor" which could involve massive denial of net services, widespread theft of data, or possibly the corruption of data being sent over the net.

Security, Privacy and the Law in the *Jones* Era

The first part of the twenty-first century brought a world of new devices, applications and accompanying "big data." At the same time there have been dramatic changes in user expectations of both privacy and security. In addition, various disclosures as well as major studies about government surveillance programs adopted since the 9/11 terrorist attacks have fueled a broader debate over essential security requirements and competing privacy demands.[7]

It generally is agreed that the legal regime for cyberspace is seriously outdated, and generations behind current technologies. Several key cases are currently before the courts, and proposed legislation is before Congress awaiting action. Major concerns exist as to how new Presidential Directives, laws and court decisions will impact on technology development as well as privacy and national security interests. Certainly the technology path will not stop or be reversed. Increasing amounts of what contribute to big data will continue to accumulate on systems worldwide presenting an ever greater challenge to public policy.

---

[7] See Privacy and Civil Liberties Oversight Board, *Report on the Telephone Records Program Conducted under Section 215 of the USA PATRIOT Act and on the Operations of the Foreign Intelligence Surveillance Court* (January 23, 2014). See also *Liberty and Security in a Changing World: Report and Recommendations of the President's Review Group on Intelligence and Communications Technologies* (12 December 2013). At the same time as these outstanding studies, unlawful disclosures by Edward Snowden first published on June 5, 2013 in the British newspaper *The Guardian* have received widespread media attention and have served to focus additional attention in this critical area.

Increasingly many Americans believe that the Fourth Amendment protects privacy as a right and that freedom and independence may not possible without some semblance of privacy.[8] Earlier Chief Justice Earl Warren predicted the problem that technological innovation has diminished privacy expectations.[9] Justices Douglas, Brandeis and others have also interpreted the Fourth Amendment as providing a fundamental right to privacy that needs to be upheld in order for justice and freedom to prevail through the ages.[10] At the same time, national security requirements have required practices and intelligence operations which in the wake of the 9/11 terrorist attacks have been viewed as critical and more recently have come under increasing attack.[11]

Data privacy suits have increased in number and notoriety in recent years and the issue of "injury in fact" has become an early challenge for privacy plaintiffs to prove. [12] Normally this type of injury is rarely an issue in lawsuits, but is as big an obstacle for data privacy plaintiffs as Mount Kilimanjaro is for hikers.[13] Here the Wiretap Act provides a private right of action against any person who "intentionally intercepts, endeavors to intercept, or procures any other person to intercept or endeavor to intercept, any wire, oral, or electronic communication."[14] Further the Stored Communication Act prohibits providers of electronic communication from "knowingly divulging to any person or entity the contents of a communication."[15]

When Congress passed the Electronic Communications Privacy Act (ECPA) in 1986 it was landmark legislation of its time.[16] That was close to three decades ago, and preceded the start of the Internet by several years. Clearly technology has evolved dramatically in these decades in ways never imagined. [17] Still, by 1986, the

---

[8] *Olmstead*, 277 U.S. at 472-73 (Brandeis, J., dissenting).

[9] *Lopez v. United States*, 373 U.S. 427, 441 (1963)

[10] *Osborn v. United States*, 385 U.S. 323, 343 (1966) (Douglas, J., dissenting); *Olmstead*, 277 U.S. at 472-73 (Brandeis, J., dissenting).

[11] See the opinion of Judge Claire Egan explaining the FISA court's rationale for approving the Section 215 telephone records program, *Amended Memorandum Opinion, In Re Application of the Federal Bureau of Investigation for an Order Requiring the Production of Tangible Things*, No. BR 13-109 (FISA Ct. Aug. 29, 2013).

[12] *In re Google Privacy Litig.*, at *4 (citing *In re iPhone Application Litig.*, No. 5:11–md–02250–LHK, 2013 WL 6212591 (N.D.Cal. Nov. 25, 2013)); *Pirozzi v. Apple Inc.*, 913 F.Supp.2d 840, 847 (N.D.Cal.2012).

[13] *Id.* at *4.

[14] 18 U.S.C. § 2511(1)(a); see id. § 2520.

[15] 18 U.S.C. § 2702(a).

[16] 18 U.S.C. § 2707(a)

[17] Christina Bonnington, Apple Mac at 30: See the Evolution of an Icon, *Wired* (Jan. 25, 2014),; Matt Honan, New Tools Show How Deep Glass will Embed in Our Live, *Wired* (Nov. 19, 2013), and

use of computers and network-related technology had grown significantly and individuals had begun using personal computers to access remote networks and data. [18] When Congress passed the ECPA one goal was to reassure industry that its growth would not be constrained by individuals' fears regarding the privacy of their communications and data maintained on computer servers.[19] Under then-existing Supreme Court precedent, it was far from clear that the Supreme Court would extend Fourth Amendment protection to these new technologies.[20]

Legal scholars have criticized the current law at length. Professor Orin Kerr argues that the lack of a suppression remedy has confused courts on how to remedy an unauthorized interception. [21] Others argue that the all private communication and stored data should be protected equally.[22] Still others have shown that under the current language, the same e-mail is subject to different protection depending on whether it is in transit, stored on a home computer, opened and stored in remote storage, unopened and stored in remote storage for 180 days or less, or unopened and stored in remote storage for more than 180 day, with at least one circuit court going so far as to hold that the lack of protection provided to electronic communication after 180 days in temporary storage provision is unconstitutional because it authorizes less than a probable cause warrant standard to search private communication.[23]

For decades now scholars have debated ways to improve the existing legal regime and its intersection with the Fourth Amendment.[24] One side of this debate

---

Amanda Scherker, Family Banned All Technology Made After 1986, *Huffington Post* (Sept. 3, 2013),

[18] Melissa Medina, The Stored Communications Act: An Old Statute for Modern Times, 63 AM. U. L. REV. 267, 290-91 (2013)

[19] *Id.*

[20] *United States v. Karo*, 468 U.S. 705, 721 (1984); *United States v. White*, 401 U.S. 745, 754 (1971).

[21] Orin S. Kerr, A User's Guide to the Stored Communications Act, and A Legislator's Guide to Amending It, 72 GEO. WASH. L. REV. 1208, 1243 (2004)

[22] See, for example, Robert A. Pikowsky, The Need for Revisions to the Law of Wiretapping and Interception of Email, 10 MICH. TELECOMM. & TECH. L. REV. 1, 49-50 (2003).

[23] See *United States v. Warshak*, 631 F.3d 266, 288 (6th Cir. 2010); Marc Zwillinger, Jacob Sommer. *Warshak Decision: Sixth Circuit's En Banc Reversal in Warshak Sidesteps Constitutionality of Stored Communication Act's Delayed Notification Provision*, BNA PRIVACY & SECURITY LAW REPORT, Vol. 7, No. 31, (Aug. 4, 2008).

[24] See Daniel J. Solove, *Reconstructing Electronic Surveillance Law*, 72 GEO. WASH. L. REV. 1264, 1299-30 (2004); Orin S. Kerr, *The Fourth Amendment and New Technologies: Constitutional Myths and the Case for Caution*, 102 MICH. L. REV. 801, 809-10 (2004); Orin S. Kerr, *The Mosaic Theory of the Fourth Amendment*, 111 MICH. L. REV. 311, 315 (2012); Daniel J. Solove, *Fourth Amendment Codification and Professor Kerr's Misguided Call for Judicial Deference*, 74 FORDHAM L. REV. 747, 749 (2005); Orin S. Kerr, *A User's Guide to the Stored Communications Act, and A Legislator's Guide to Amending It*, 72

proposes a universal search warrant requirement the other argues that Congress is the best suited to enact laws to protect privacy because the Courts are faced with the disadvantage of trying to hit a "moving target," the continuing development of technology, while interpreting a distinct moment in time, the case and controversy before them.[25] Most experts, however, agree that the existing legal regime needs to be modified to improve its application to modern technology and the demands of "big data."[26]

Fourth Amendment Interpretation: Traditionally, the Fourth Amendment right to privacy has been viewed as a property right.[27] Searches of property required a warrant issued by a magistrate supported by probable cause.[28] While Fourth Amendment right to privacy still maintains its foundation in property rights, the Supreme Court has also supplemented property-based privacy rights with a reasonable expectation of privacy outside of any property right.[29]

Current conceptions of privacy are based on the landmark *Katz* case where the Court held that even in a public place, a person may have a reasonable expectation of privacy in his person.[30] Justice Harlan's concurrence in *Katz* has served as the guiding principle for the analysis of whether search violates a reasonable expectation of privacy, establishing two requirements for a reasonable expectation of privacy: (1) a person have exhibited an actual (subjective) expectation of privacy; and (2) the expectation be one that society is prepared to recognize as "reasonable"(objective).[31] Writing for the majority, Justice Stewart, reasoned, "[W]hat a person knowingly exposes to the public, even in his own home or office is not a subject of the 4th Amendment protection."[32] He continued, however, to say, "But what he seeks to

---

GEO. WASH. L. REV. 1208 (2004)

[25] Daniel J. Solove, *Reconstructing Electronic Surveillance Law*, 72 GEO. WASH. L. REV. 1264, 1299-30 (2004). S. Kerr, *Congress, the Courts, and New Technologies: A Response to Professor Solove*, 74 FORDHAM L. REV. 779, 782 (2005).

[26] Robert A. Pikowsky, *The Need for Revisions to the Law of Wiretapping and Interception of Email*, 10 MICH. TELECOMM. & TECH. L. REV. 1, 65-66 (2003); Solove, *Reconstructing Electronic Surveillance Law*, op. cit.

[27] See *United States v. Jones, 132 S. Ct. 945, 954 (2012)*, *Florida v. Jardines,* 133 S. Ct. 1409, 1417-18 (2013).

[28] *Shadwick v. City of Tampa*, 407 U.S. 345, 354 (1972).

[29] See *United States v. Jones, 132 S. Ct. 945, 954 (2012)*, *Florida v. Jardines,* 133 S. Ct. 1409, 1417-18 (2013).

Katz v. United States, 389 U.S. 347, 360 (1967).

[30] *Id.* at 351.

[31] *Id.*

[32] *Id.* at 351 (majority opinion).

preserve as private, even in an area accessible to public, may be constitutionally protected."[33]

For close to half a century now, *Katz* has served as a foundation for determining whether behavior constitutes a violation of the Fourth Amendment right to privacy. Moving beyond communications, the Court has applied these principles in considering whether here is a reasonable expectation of privacy in "open fields" outside of the curtilage of a home.[34] The reasonable expectation test remains the as to whether there is a reasonable expectation of privacy where there is no property right at issue, such as in electronic communications or data storage.

Exposure to the Public: Cases following *Katz* stand for the principle that what one knowingly exposes to the public is not subject to Fourth Amendment protection.[35] Furthermore, as the Court articulated in *California v. Greenwood*, "An expectation of privacy does not give rise to Fourth Amendment constitutional protection unless society is prepared to accept that expectation as objectively reasonable."[36]

Most recently the Court has issued another landmark privacy decision in *United States v. Jones*, a case involving a GPS tracker attached to a drug dealer's vehicle without judicial approval, and then used evidence obtained through tracking him to convict him. The Court held that the reasonable expectation of privacy test supplements the property based expectation of privacy and therefore the placing of a tracker on the Jeep, an effect, constituted an unlawful search.[37]

Concurring in the decision, Justice Sotomayor reasoned that unrestrained power to assemble data that reveals private aspects of identity is susceptible to abuse, warning that that it may be necessary to reconsider the premise that an individual has no reasonable expectation of privacy in information voluntarily disclosed to third parties because it is ill-suited to the digital age.[38] Foretelling the issues of "big data" Justice Sotomayor went on to raise concerns over the comprehensiveness of a record of personal movements and a "wealth of detail about

---

[33] *Id.*
[34] *Oliver v. United States*, 466 U.S. 170, 184 (1984); *United States v. Dunn*, 480 U.S. 294, 305, (1987).
[35] *Katz v. United States*, 389 U.S. 347, 360 (1967).
[36] *California v. Greenwood*, 486 U.S. 35, 39 (1988). In *Greenwood*, held that garbage left at the side of the road is readily accessible to animals, children, scavengers, and other members of the public.
[37] *United States v. Jones, 132 S. Ct. 945, 954 (2012)*.
[38] *Id.* at 954 (Sotomayor, J., concurring); *Smith*, 442 U.S., at 742, 99 S.Ct. 2577; *United States v. Miller*, 425 U.S. 435, 443 (1976).

her familial, political, professional, religious, and sexual associations."[39] To protect the information, however, requires that "Fourth Amendment jurisprudence ceases to treat secrecy as a prerequisite for privacy."[40] As Justice Marshall stated, "privacy is not a discrete commodity, possessed absolutely or not at all."[41]

Subsequently *Jones* has been cited by numerous courts considering a range of privacy issues, including numerous federal appellate courts, the Foreign Intelligence Surveillance Court and the Supreme Court itself.[42]

Meeting the Challenge – Toward a National Policy

It is the unfortunate reality that national policy toward cybersecurity during the 1990s, the Internet's first critical decade was in large part either non-existent, badly managed, poorly funded, and in some cases simply absurd. As the net literally exploded in terms of users and applications, and evolving threats were seen, there was little national consensus as to whose responsibility it was to secure cyberspace and respond to the threats. While the Government and the military became large-scale users, and the "pig at the trough," little was done by to protect this vital resource. As a whole, Government saw this as a responsibility of the commercial service providers while Government programs to deal with it were minimal and inadequate.

What the nation failed to see at that time was the reality of the cyber threat problem, mostly from overseas. As national security, government, and finance became large net users they, adding "big data" to the networked world, they became lucrative targets for both major criminal enterprises, as well as foreign military forces who foresaw the potential for cyberwarfare.[43] At the time America focused largely on defense against hackers and lower level threats, and not looking to the larger evolving threat environment.

While the 9/11 attacks themselves had little to do with cyberwarfare or "big data" they did provide a catalytic shock to the government in terms of looking far more seriously at new threats, particularly in the technology space. Cell phone and

---

[39] Jones, 132 S. Ct. at 955 (Sotomayor, J., concurring); See, e.g., *People v. Weaver*, 12 N.Y.3d 433, 441-42 (2009).

[40] *Jones*, 132 S. Ct. at 957 (Sotomayor, J., concurring)

[41] *Smith*, 442 U.S. at 749.

[42] See *Amended Memorandum Opinion, In Re Application of the Federal Bureau of Investigation for an Order Requiring the Production of Tangible Things*, op. cit., and *Opinion and Order*, No. PR/TT [redacted] (FISA Ct.).

[43] See, for example, *APT1, Exposing One of China's Cyber Espionage Units* (Mandiant, 2013).

Internet use by terrorists and others now became a serious subject of interest. Programs to focus on these technologies which languished in the 1990s received new attention and support. At the same time a number of early cyber-attacks such as Moonlight Maze (from Russia – 1999); Titan Rain (from China – 2004); and others attacking critical systems drove home the reality of increasing threats.

A Strategy for Cyberwarfare: Increasing cyberattacks from foreign groups have raised the specter of cyberwarfare as a realistic area for future conflict. Analysts continue to debate as to how this new type of warfare, which has no geography, differs from the traditional model of kinetic warfare, and what "rules" of warfare apply, and the extent to which the elements of loss of life and destruction of property – the two cornerstones of the kinetic model of warfare – might apply in the cyberwar context.

Cyberspace is Part of a Highly Dynamic World: Cybersecurity has become an essential element of life in the wired world, which is a highly dynamic one where both the technology base and the threats continue to evolve. For some time now this world has moved into an era of "digital everything" with an almost seamless merger of communications, computing, and media of all kinds including "big data" which are largely digital. Coupled with hardware and communications bandwidth that has become increasingly cheap, the marginal costs of communications are free nearly so in many cases, which have caused use of cyberspace to grow by orders-of-magnitude in a few short years.

The enabling technologies and economics have also brought about some major changes in culture. Use of the net, devices, and advent of "big data" have brought about modern cultural artifacts from Internet dating to social awareness streams. Net-based commerce is fast surpassing all other forms, while businesses as well as the government agencies have become almost totally dependent on net based systems.

System architectures are increasingly moving to a cloud concept, while more serious threats from cyber criminals, cyber warriors and cyber terrorists across the globe continue to grow and it is increasingly important that any policy or strategy have effective defense and offensive elements that aid in meeting overall strategic objectives as well as user demands for privacy, security and resilience. Meeting these sometimes competing demands presents an increasing policy challenge.

Building the Technology Base: Implementing a successful national strategy must necessarily start with building the technology base, and in this area largely involves educating people with the skills necessary to meet the emerging challenges. It is also

an area that simply requires the "best and the brightest" to create the type of software and other technologies required. Educating the necessary to meet this challenge requires a new level of commitment to the nation's universities, possibly using the model of the Eisenhower Administration in responding to the Cold War challenges of the "space race."[44]

This model for cyberspace and "big data" makes good sense, and it is reasonably certain that the universities are not going to meet this challenge utilizing only internal resources. In the current economic climate event the major private universities are constrained, while most public universities are under enormous economic pressure. While there is sound logic that shows there are increasing numbers of jobs in cyberspace, the fact does not seem compelling enough to overcome the level of inertia in education today.

Acceleration of Government Programs: Notwithstanding budgetary pressures, it is increasingly clear that the Government cannot continue to be "the pig at the trough" in terms of massive net use; fail to adequately fund effective security programs; and maintain the false expectation that the private sector will recognize the full scope of the problems and remedy them. Efforts to protect the net and "big data" need continued strong and increasing support. Not all of these tasks can be left to the Defense Department and the intelligence agencies. Without exception all other Government agencies have become major users of cyberspace and need to become partners in its ongoing protection.

Partnership with Industry: Aside from limited government funding, one reason national policy on cyberspace failed in the 1990s was a basic misunderstanding of the role industry could and would play in securing the net and protecting "big data." There were unreasonable expectations that industry would recognize the vulnerabilities and fix them. It was believed that it was not essential for the Government to support this in a meaningful way and that user demands, from both the public and private spheres, would drive industry to meet the challenge, a belief that was only partially correct. What was done was largely inadequate, and insufficient to meet the threats that evolved.

---

[44] At that critical point in history the nation undertook a series of coordinated initiatives starting with substantial government investment in science and math education, under the National Defense Education Act (NDEA). The government initiated new technology agencies, such as the Advanced Research Projects Agency (ARPA), the National Science Foundation (NSF) and others.

Policy now requires a more realistic approach to industry involvement on several levels. It is essential to recognize that industry built cyberspace and created "big data" – and they will fix it, irrespective of who pays. By and large the Government can only write checks – not computer code. Even in the most sensitive areas the actual work is out-sourced to commercial firms with few programmers being Government employees.

Here the nation needs to move to a model where the technology companies that dominate cyberspace are made a more integral part of the process. The model what was highly effective in dealing with the communications firms for decades is a useful one that has not been effectively employed where cyberspace is concerned. Certainly some of the traditional telecoms are "within the tent" but many of the most important and critical firms are not. In the final analysis the nation needs to look ahead at what the solution is going to be, and work back from that, making sure that the technology base and the supporting industrial base can meet the very real threats and challenges ahead.

Another key element of this partnership needs to be with the holders of "big data" including the financial sector; the health services industry; as well as the telecoms and internet service providers who hold increasingly large amounts of "big data." One aspect of this partnership needs to be the timely and accurate provision of threat data coming from government sources and vice versa.[45] Another is a far broader national policy and legal regime that recognizes the role that industry servers and clouds have in maintaining "big data" and protecting both the privacy of users and the security of their data.

These are by no means simple issues. They continue to involve a number of complex technical, legal and financial considerations all of which are in a constant state of change. Here the challenge presented by the Office of Science and Technology Policy represents a useful way to engage a wide range of Americans in the process of developing an effective national policy in this most critical area.

[45] The current Defense Industrial Base (DIB) effort is one useful approach, but a far more extensive set of program is needed.

Response to OSTP "Big Data" Request for Information

Question 1: What are the public policy implications of the collection, storage, analysis and use of big data?

Submitted by:

Mary J. Culnan, Ph.D.[1]
Senior Fellow, Future of Privacy Forum
Professor Emeritus, Bentley University
mculnan@bentley.edu

March 28, 2014

Introduction

Big data by definition raises two important issues. The two issues are (1) the fact that big data involves the secondary use or repurposing of data originally collected for another reason, and (2) the need for redress/procedural fairness particularly when big data is used to make automated decisions. These issues in turn pose significant challenges to the current U.S. policy framework and privacy proposals for protecting consumer privacy including the Consumer Privacy Bill of Rights specifically and the Fair Information Practices Principles (FIPPS) more generally. My comments will describe these challenges and suggest steps the White House can take to address them.

The FIPPS were designed for an information ecosystem where a direct relationship exists between an individual who provides information to a single, known organization which retained control over the subsequent use of the information and was responsible for using the information responsibly. This relationship was typically well-understood by consumers. Here, privacy is based on a "notice and choice" paradigm where the organization provides notice about its information practices to the individual, and offers the opportunity to opt out of secondary use when personal information is collected for one purpose and used for other unrelated purposes, thereby violating the context principle. In the commercial world, secondary use typically involves sharing for marketing purposes. For website privacy, organizations provide notice and choice through their online privacy notices.

Online advertising posed one of the first challenges to this paradigm as third party ad networks with whom the consumer had no direct relationship partnered with web content providers and began to collect information about the individual's surfing patterns from content providers' across the web. The ad network typically provided notice and choice on its own website, a practice that provided limited transparency for consumers, many of whom were unaware of the existence of these firms. The growth of the Internet of Things, the mobile ecosystem, the proliferation of sensors and social media, all of which are driving big data, pose even greater challenges to the traditional privacy paradigm.

The Repurposing Challenge of Big Data

The first challenge big data poses to most of the principles in the Consumer Privacy Bill of Rights results from the fact that big data typically involves secondary use. Data collected for one purpose is combined

[1] The views expressed here are my own.

with other sources of data and repurposed after the fact. Often this can involve third parties who are unknown to the individual. It should be noted that big data does not appear to present new challenges to the Security principle in the Consumer Privacy Bill of Rights, while the Accountability principle provides an opportunity to address many of the challenges posed by big data and will be discussed subsequently.

The specific challenges of big data to the Consumer Privacy Bill of Rights from repurposing include the following:

Transparency: With Big Data, data collected for one purpose are often combined with other sources of data after the fact. Because these uses may not be known or anticipated when the information was collected, it is unlikely they are described in the organization's traditional privacy notice with any specificity that would promote understanding by the individual. Providing effective notice later, when these new uses actually occur, is not practical. Further, when big data analytics are used to create PII from non-PII long after collection, this poses additional challenges to providing transparency [2]. The use of third party data also poses transparency challenges as described previously.

Respect for Context: Because big data typically involves secondary use or repurposing, it is likely to violate the context principle. It is impractical to go back after the fact and offer choice for every new use of data if choice is appropriate.

Access and Accuracy: Big data often involves the use of analytics to draw inferences and the data are not in a "usable format" making access and correction infeasible. Further, where an organization has acquired personal information from multiple sources, the organization using the data may not be able to provide access and correction to data when the data are gathered, owned and maintained by a third party, even if the source of the data can be identified[3].

Focused Collection: Big data by definition often conflicts with the collection limitation principle[4].

Using Accountability to Address the Repurposing Challenge of Big Data

Effective governance processes based on Privacy by Design (PBD) can address some of the challenges Big Data poses to the Consumer Privacy Bill of Rights and demonstrate an organization's commitment to accountability. PBD calls for privacy to be imbedded in systems end to end. In particular, conducting a Privacy Impact Assessment (PIA) before undertaking a new big data initiative can address and hopefully avoid privacy issues[5]. Making the results of the PIA available to the public can also help address the

[2] See Kate Crawford & Jason Schultz, "Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms, *Boston College Law Review*, 55: 93-128, 2014. Available at: http://bclawreview.org/files/2014/01/03_crawford_schultz.pdf. Crawford and Schultz cite the familiar Target example where analytics were used to infer a customer was pregnant from her purchases, thereby creating new PII.

[3] This is an issue with data brokers for example.

[4] See for example: Justin Brookman and G.S. Hans, "Why Collection Matters: Surveillance as a De Facto Privacy Harm," September 2013, Available at: http://www.futureofprivacy.org/big-data-privacy-workshop-paper-collection.

[5] For an example of comprehensive guidance on conducting a PIA, see: Department of Homeland Security, *Privacy Impact Assessments: The Privacy Office Official Guidance*, June 2010. Available at: https://www.dhs.gov/xlibrary/assets/privacy/privacy_pia_guidance_june2010.pdf. The DHS guidance

transparency challenge. For big data, some have recommended that a traditional risk-benefit assessment be expanded to include an ethical assessment of whether undertaking something legal is also the right thing to do [6]. As President Obama stated in his January 17, 2014 speech where he announced the big data review process, "can" is not always equal to "should":

> "…the power of new technologies means that there are fewer and fewer technical constraints on what we can do. That places a special obligation on us to ask tough questions about what we should do."

The full range of privacy issues raised by big data are unlikely to be addressed absent organizational commitment to doing the right thing backed up by robust processes[7].

Redress Challenge of Big Data

Big data are often used to make automated decisions ranging from which marketing offers may be of interest to a consumer to determining eligibility for government benefits. The potential for harm obviously varies with the nature of the decision. While the Consumer Privacy Bill of Rights addresses the issue of accuracy of source data, it is silent on the issue providing redress when accurate (or inaccurate) data are used to make an incorrect decision.

Citron cites numerous examples of this issue and calls for "technological due process" to address unfairness resulting from problems such as automation bias (excessive trust in an automated decision), programmers failing to accurately code policy rules, misidentification due to problems with data, the difficulty of diagnosing the cause of errors because of a lack of audit trails or the opacity of the source code, and failure of these systems to provide adequate notice in the event of an adverse decision [8]. Crawford and Schultz make similar arguments for big data given the shortcomings of the current notice and choice paradigm when applied to big data[9]. This problem may be further exacerbated when organizations hire independent third parties to build and operate a system rather than operating it themselves. If the system is based on third party data, the third party may also responsible for providing redress to people with whom it has no relationship in the event of an adverse outcome.[10]

---

[6] See for example, Ryan Calo, "Consumer Subject Review Boards, A Thought Experiment," September 2013, Available at: http://www.futureofprivacy.org/big-data-privacy-workshop-paper-collection

[7] See for example Mary J. Culnan and Cynthia Clark Williams, "How Ethics Can Enhance Organizational Privacy: Lessons from the ChoicePoint and TJX Data Breaches," *MIS Quarterly*, 33(4): 673-687, December 2009, and Mary J. Culnan, "Accountability as the Basis for Regulating Privacy: Can Information Security Regulations Inform Privacy Policy," Privacy Papers for Policy Makers, 2011. Available at: http://www.futureofprivacy.org/wp-content/uploads/2011/07/Accountability%20as%20the%20Basis%20for%20Regulating%20Privacy%20Can%20Information%20Security%20Regulations%20Inform%20Privacy%20Policy.pdf

[8] Danielle Keats Citron, "Technological Due Process," *Washington University Law Review*, 85(6): 1249-1313, 2008. Available at: http://digitalcommons.law.wustl.edu/lawreview/vol85/iss6/

[9] Crawford & Schultz, 2014.

[10] See for example the use of Experian's identify verification system by HHS and SSA to authenticate people creating an account at healthcare.gov or seeking online access their Social Security account respectively. The Experian system is based in part on the individual's credit report. If the individual is rejected by the Experian system, they need to dispute their credit report with Experian.

Using Procedural Fairness to Address the Redress Challenge

Both Citron and Crawford and Schultz make specific recommendations for addressing these challenges including at a minimum, notice about the types of decisions made and the general sources of data on which the decisions are made, and the opportunity to have an adverse decision reviewed. Further, organizations which rely on automated decisions should invest in employee training for the individuals who use these systems about their potential biases and shortcomings.

Recommendations

I do not recommend that the White House pursue privacy legislation for big data at this time. Big data applications, particularly those based on sensors and the Internet of Things, are in their infancy and evolving rapidly, and well-intentioned legislation could have unintended consequences that stifle innovation and the free flow of information without effectively solving the privacy challenges. Further, legislation is unlikely to address the ethical challenges of big data. Finally, even if an effective law could be drafted, it is also difficult to imagine that it could be enacted given the realities of the current political environment. However, big data applications pose significant privacy challenges that should not be ignored because of their potential for harm and the threat they pose to innovation based on big data by undermining consumer confidence[11]. Absent legislation, the White House can lead by example and by executive order by developing and implementing best practices for non-NSA federal agencies [12].

For example, The White House Deputy CTO and the CIO Council could develop a set of best practices that adapt the Consumer Privacy Bill of Rights to big data, and also incorporate clear guidelines for procedural fairness. Subsequently, this group could conduct an inventory of current applications that make automated decisions about individuals, and/or involve the use of information compiled by commercial sources such as information aggregators. While these systems do not necessarily meet the definition of big data in the RFI, they raise many of the same issues [13]. A set of representative applications could be selected to focus on the fairness issues in existing systems[14] . These applications can serve as test cases for the implementation of the best practices where gaps have been identified. The White House could then update the Consumer Privacy Bill of Rights and guidelines for conducting PIA's for big data based on the results of these test cases, and OMB could issue updated guidance to the agencies.

Subsequently, a multistakeholder process could be convened to adapt these best practices to the private sector along the lines of those processes convened by NTIA for transparency of mobile apps and facial recognition. Organizations using big data should be encouraged to implement "reasonable privacy" consistent with the type of data being used, the ways it is being used, and the potential for

---

[11] In the mid-1990's, The White House recognized that privacy concerns posed a similar threat to e-commerce. See*: A Framework for Gobal E-Commerce*, July 1, 1997. Available at: http://clinton4.nara.gov/WH/New/Commerce/read.html

[12] For example, in 2003 OMB issued guidance to agencies on implementing the E-Government Act of 2002 (M-03-22, September 26, 2003) . The Guidance calls for all agencies to conduct PIA's for electronic information systems and collections and make them publicly available.

[13] The RFI defines "big data" as "datasets so large, diverse, and/or complex, that conventional technologies cannot adequately capture, store or analyze them."

[14] See Citron (2008) for examples.

harm resulting from these uses[15]. The results of the White House efforts could provide the basis for an enforceable self-regulatory code of conduct. Organizations signing on to the code would be subject to an enforcement action by the relevant enforcement agency if they failed to comply[16]. While this is not a perfect solution, this proposal represents a practical first step in addressing a significant privacy challenge of big data, and could ultimately provide the basis for legislation.

Thank you for the opportunity to share my thoughts. I wish the White House well with this important endeavor.

[15] See: Culnan (2011) for a discussion of how current requirements for "reasonable security" can serve as the basis for privacy policy.

[16] See: The White House, *Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy*, February 2012.

March 30, 2014

Response to
Office of Science and Technology Policy: Government "Big Data";
Request for Information

By
Georgetown University

Contact: Robert M. Groves, Provost, 202-687-6400, bgroves@georgetown.edu

The RFI does not define "big data;" we will use as a definition: "large, diverse, complex, longitudinal or relational data sets generated from instruments, sensors, internet transactions, email, video, click streams, digitized administrative records, and other digital sources."

We believe that the nation that builds institutions and policies that permit the conjoining of all these data, combined with traditional statistical data sources, will win the future.

(1) What are the public policy implications of the collection, storage, analysis, and use of big data? For example, do the current U.S. policy framework and privacy proposals for protecting consumer privacy and government use of data adequately address issues raised by big data analytics?

Since 1940, the informational infrastructure of the United States has consisted of the Federal Statistical System (e.g., the unemployment rate, Gross Domestic Product), shared scientific data sets produced through Federal grants (e.g., the NIH Health and Retirement Study), summary information from administrative files (e.g., Medicare summary statistics), public opinion polling (e.g., Pew Research Center), private sector subscription statistics (e.g., Nielsen television ratings, Arbitron radio ratings), and customer satisfaction ratings (e.g., automobile manufacturers). We call these "common good" uses of statistical information because they are freely available to all. Collectively, they are a key cornerstone of the democracy, providing tools for an informed citizenry to guide the country.

Many of these informational sources depend on the sample survey as a method

of collection. However, sample surveys are threatened by large increases in costs associated with the use of enumerators, heightened nonresponse, and weakened sampling frames. Thus, there has been a steady trend toward surveys producing a smaller portion of data on the economy and society.

While the future of surveys may be problematic, there has been an exponential rise of "big data" in all sectors of the society much of which is relevant to many challenges or issues confronting society.

The strengths of these new digital resources are:

a. They are timely, offering near real time monitoring of some phenomena
b. They are low cost, designed as auxiliary to ongoing processes
c. They have few geographical restrictions, covering large portions of the world

The weaknesses of these data are:

a. They are lean in variables, sometimes only offering time, place, and some single attribute
b. They are no consistent quality standards for the data
c. They are disproportionately held by business owners who use them to run profit-making enterprises; they are thus not generally made available for common good purposes
d. There is a mismatch in the privacy reality versus the privacy believed involving these data

What are the public policy issues of this new data world?

a. Data useful for common good, societal purposes are not accessible to institutions of the society focusing on those goods (e.g., Federal statistical agencies, academic researchers, nonprofits, NGOs or government agencies) let alone the public at large
b. Data holders and privacy advocates do not possess a safe environment to negotiate common good purposes of the data
c. Data holders and privacy advocates fear legal implications of data sharing
d. Data holders fear being "scooped" by the creation of profitable information products partially based on their data
e. Data holders fear the release of propriety knowledge of the workings of their firms' processes by sharing data
f. The privacy debate has yet to evolve to a state in which the value of big data to address common good issues is recognized
g. Individuals contribute personal information under the misguided perception that the information provided will remain private
h. Hence, building a massive permanently linked data set on persons and businesses is not acceptable, but simultaneously analysis of multiple data sets is necessary

(2) What types of uses of big data could measurably improve outcomes or productivity with further government action, funding, or research? What types of uses of big data raise the most public policy concerns? Are there specific sectors or types of uses that should receive more government and/or public attention?

Individual high-dimensional data sets are often poor information sources relative to the rich multivariate data of sample surveys. This arises often because the dimension on which the data deserve the "high-dimension" label is often merely the number of observations/records, not the number of attributes known about the observations. Multiple big data sets must be conjoined to be of value to most pressing societal questions.

The most difficult observation but the most important is this: the regulatory and legal restrictions preventing the blending together of multiple large data sets is stifling progress for the whole country. The problem is acute both within and outside of the government as even federal executive branch agencies face incredible challenges in sharing data amongst each other. Sharing data across state agencies requires overcoming even more herculean obstacles. The private sector data owners have no incentive to share these data (with full confidentiality protections) to those whose mission it is to inform the society about itself and seek common good actions.

The United States needs a way forward to use the wealth of big data resources for the benefit of the full society.

Specific Sectors with Useful Big Data Resources:

a. Educational record systems should be linked nationwide, should continue to be linked longitudinally, and should be accessible by educational researchers for studies of the attributes of successful programs
b. Facebook, LinkedIn, Twitter, Google+ should be accessible for common-good research purposes
c. Search engine queries (e.g., Bing, Google) should be accessible for research purposes
d. The master address file of the Census Bureau should be available for linkage to big data sets for addition of geospatial and census survey data, with full confidentiality restrictions on personal data
e. Anonymized credit card, health transaction, utility transaction data should be available for research purposes
f. Document repositories without copyright restrictions should be made available for research purposes

(3) What technological trends or key technologies will affect the collection, storage, analysis and use of big data? Are there particularly promising technologies or new practices for safeguarding privacy while enabling effective uses of big data?

Due to the continuing exponential growth of big data, the ability to move, store, and analyze data in a centralized fashion is rapidly becoming infeasible. In addition, the cost and physical constraints associated with transmission, space, power, and cooling is fundamentally driving big data analysis to more decentralized and distributed architectures.

Computation analytics must increasingly be moved to the data, lacking the ability to continue moving data to a single or small set of computation centers. This trend is challenging the manner in which we think of data analysis, driving the industry towards collaborative analytic fabrics versus conventional data processing approaches.

One very promising approach for addressing this trend is to employ the use of knowledge overlay technology that allows data to remain largely at rest at its source. Georgetown is pursuing this approach by creating a distributed knowledge representation framework that involves each source locally managing and curating its data resources, but providing an important data-to-information transformation that enables the salient features of their datasets to be shared with others. This transformation results in a significant reduction in data exchange volume and private information exposure.

The associated techniques leverage the significant recent advances that have been made in graph-based knowledge representation and reasoning. A critical aid to this process is the development and application of new data adapter technology, which holds considerable promise in reducing the enormous "plumbing" burden typically associated with disparate data handling due to the wide diversity in format, design, organization, and storage mechanisms.

For further privacy safeguards, a new blend of "cloaked" pattern processing and "blindfolded" search techniques are emerging that require significantly less direct access to data to achieve the same analytic outcomes. For example, Georgetown's Black Box technique is being deployed to analyze extremely sensitive health record information in order to help track and prevent the spread of HIV/AIDS. This process is accomplished with no human access to any private information. One of the most encouraging results is that such techniques are conducive to formal methods analysis to better ensure implementation proof-of-correctness for both robust privacy and cybersecurity enforcement.

At the hardware level, the emergence of the next generation of exascale super-computing resources is driving an entirely new family of big data-enabled analytic modeling, simulation, and prediction capabilities at regional, national, and

international scale. Increasingly, these new platforms have the ability to represent substantially larger, multidisciplinary problems spanning numerous interdependent research domains due to their extremely large shared memory address spaces and high-performance multithreading architectures. Several powerful variations for specialized graph processing are emerging that will enable performance increases up to six orders of magnitude, awarding extraordinary opportunities for global integrative research.

(4) How should the policy frameworks or regulations for handling big data differ between the government and the private sector? Please be specific as to the type of entity and type of use (e.g., law enforcement, government services, commercial, academic research, etc.).

We are missing an institution in the United States. It would probably need to lie on the intersection of the private sector, the academic sector, and the government sector:

a. The institution would offer a safe environment for common-good uses of high dimensional data. It would be a public-academic-private joint venture.

b. It would be governed partly by privacy protectors, with independent authority to alert the public to any breaches of confidentiality pledges.

c. Data owners from all sectors, private commercial entities local and state governments would make accessible their data resources for specified, publicly-documented uses of data. Data owners would receive tax benefits and liability protections for their contribution to the common good. Additional guidelines and protections would need to be developed for compounded derived information.

d. Multiple data would not be permanently linked but would be simultaneously accessible for statistical operations

e. A governing body would review each proposal for use of the various data sources for common good utility; approved proposals would be publicly available.

f. Both research uses and action uses would be eligible for proposals.

g. It would acknowledge that most big data in the future will not be feasibly movable to another physical location. Thus, it would be an environment creating software exchange tools of delimited ability to extract data from an existing big data source and conjoin them with data from others for the purpose of addressing questions that can be answered only with both data

jointly.

h. Above all, there would be transparencies to permit all US residents to know what data analyses were being performed with what data sources; the statistical information derived from the analyses would lie in the public domain.

(5) What issues are raised by the use of big data across jurisdictions, such as the adequacy of current international laws, regulations, or norms?

The power of big data rests with the ability to conjoin large, complex, longitudinal, and relational data sets from a wide spectrum of diverse sources. With the 21st century data explosion, the sources of these data sets naturally span multiple jurisdictions, each with often differing, and sometimes conflicting, legal authorities, management policies, and handling procedures. Thus, big data invariably is a multijurisdictional integration challenge. Our experience in understanding such challenges is that it is a *quadratic* problem. That is, the amount of time (or cost) associated with integrating n organizations is proportional to n^2. As a result, progress in effectively conjoining of data at any appreciable scale has been very slow, rarely beyond just a very small number of organizations.

Upon analysis of this quadratic cost nature, four key issues have been identified:

a. Plumbing of data: Across multiple jurisdictions, the richness of today's technology offerings has created a vast number of differing data standards, representation formats, models, schemas, field definitions, storage and retrieval tools, access mechanisms, etc. While standardization can help, rigid enforcement also curbs innovation. Thus, the burden of conjoining data across these jurisdictions can be massive, requiring substantial time to understand, negotiate, and resolve the often mundane, but very lengthy list of difficult data "plumbing" issues. In actual practice, the integration of large numbers of databases from numerous organizations can quickly escalate into lengthy multi-year projects requiring hundreds of millions, perhaps billions of dollar investment in software and IT infrastructure. Large development efforts at this scale have been exceedingly difficult to formulate, specify, and subsequently manage successfully resulting in high project failure rates.

b. Protection of data systems: In this current age, cyber security is now a very serious concern and credible threat. While integrating large numbers of data systems creates an enormous big data analytic opportunity, it also substantially increases the risk of cyber attack. That is, if an adversary is able to compromise just one system component, a closely integrated multijurisdictional solution can expose the entire aggregate enterprise to

attack. Finding and mitigating the weakest link in such a large system of systems has proven quite difficult for the computer security industry.

c. Pattern matching of data: Most contemporary big data solutions employ sophisticated technological variants of "data dumpster diving", searching for correlations in whatever mound of data that has been possible to amass. While frequently achieving intriguing results, the process very much follows a constrained query-response model. That is, the data mound is queried, a response is produced. The process repeats. Unfortunately, at the multijurisdictional big data realm, this approach is ineffective for many important problems. With the dynamic nature of data at such scale, the generated response to a query can frequently become obsolete just within moments of its issue. Thus, the meaning and integrity of comprehensive, multistage analysis conducted within this paradigm is suspect. An alternative approach involves the formulation, specification, and detection of patterns that are instead persisted over time, with an asynchronous reporting and continuous alerting mechanism.

d. Privacy assurance: The recent purported disclosures of the U.S. Governments global surveillance activities have reignited the complex debate regarding storage and access of personal information. This debate has placed the desire for national security in tension with constitutional protections of individual liberties. Public opinion has framed this as a tradeoff, balancing sacrifices in civil liberties for gains in national security. As a result, any serious, credible big data endeavor across jurisdictions invariably must address privacy as a fundamental core principal. As with data plumbing, resolving such issues has historically been a very tedious, slow quadratic cost process ultimately involving small incremental changes to jurisdictional policy, law, and international treaty. The true promise of big data cannot be unleashed without an effective solution to this single critical issue.

In cooperation with the national laboratories, commercial industry, the defense and intelligence industry, and other members of academia and the public policy community, Georgetown University has examined these four fundamental issues in considerable depth. The conclusion that has emerged is that all four of these issues are extremely heavily intertwined. In other words, a solution to any one of these items invariably requires a simultaneous solution to each of the other three. As a result of this complex interdependency, it is not surprising that only modest, piecemeal advances have been made throughout government, commercial, and academic sectors. With high confidence that a uniform solution actually does exist, the resulting approach requires a reformulation of the problem that challenges many of today's leading assumptions associated with "Big Data". The reformulation that Georgetown is pursuing fundamentally shifts emphasis from data and its manipulation, to knowledge and collaboration. Such a paradigm shift can only be enabled through a powerful blending of leading-edge policy experts working in very close concert with leading-edge social and economic

researchers, on one hand, and computational scientists, on the other, to address the full, aggregated spectrum of issues versus a narrowed focus on individual, isolated facets.

Response to Request for Information
Government "Big Data"

Document Number: 2014-04660

March 30, 2014

Submitted to: Office of Science and Technology Policy
Eisenhower Executive Office Building
1650 Pennsylvania Ave. NW, Washington, DC 20052
bigdata@ostp.gov

Submitted by: Intrical LLC
1320 R St. NW #1; Washington, DC 20009
240.888.4406 | info@intrical.us | http://intrical.us

Table of Contents

Introduction

Recently emerging techniques for data processing and management enable an unprecedented capacity to harvest data. This new capacity has great potential, but so far lacks the necessary focus to optimally contribute to the public good or to adequately protect the public from abuses of privacy, civil liberties, and civil rights.

Big Data's biggest impact on privacy policy may be cultural rather than technological.

Big Data's biggest impact on privacy may be cultural rather than technological. Increasingly, data logged by routine processes such as digital communication, digital account management, digital media consumption, etc. is seen as a commodity for acquisition and ingestion for purposes including scientific research, commercial research, consumer targeting, law enforcement, and national security. This culture has produced an arms race[1] with higher volume seen as a success metric[2], which creates a decoupling between the drive to acquire/digest this data and the measurable utilization of this data. This, in turn, leads to an indiscriminate spread of data across organizations. Unfortunately, consumers are left with little control over this data, or appropriate awareness of its implications. Vendor privacy policies are rarely well considered or well understood. Meanwhile, national security initiatives are necessarily opaque to minimize potential for malicious subversion.

Regulations on handling data have been enabled by several legislative statutes in the past two decades, but the laws were narrow in scope, focusing only on the particularly sensitive industries of finance and healthcare. Researchers have also indicated a regulatory overlap with scientific standards on research involving human subjects[3], as the computational capacity of Big Data processing has the potential to reveal as much, if not more, than the heavily regulated practice of human subject interviewing. The Obama administration addressed the general concern of consumer data in February 2012's *Consumer Data Privacy in a Networked World*, but applying these principles to the Big Data culture requires more structure.

In this RFI response, we describe some factors affecting the appropriate safeguarding of Big Data systems, both in policy and system architecture, to provide Government and private-sector organizations engaging in Big Data projects with appropriate levels of flexibility while providing individuals with adequate privacy. In the end, our ideas are far from comprehensive, and much research and experimentation remains needed to develop necessary regulatory resources to implement Big Data practices effectively and safely.

---

[1] Tessella *Can You Win the Big Data Arms Race?* http://tessella.com/download/8398
[2] AdSoft Direct *The More Data, The Better* http://www.adsoftdirect.com/the-more-data-the-better/
[3] Stanford University *Consumer Subject Review Boards: A Though Experiment* https://cyberlaw.stanford.edu/files/publication/files/Calo.pdf

Public Policy Implications

Q1: What are the public policy implications of the collection, storage, analysis, and use of big data? For example, do the current U.S. policy framework and privacy proposals for protecting consumer privacy and government use of data adequately address issues raised by big data analytics?

The complexity of applying privacy guards on Big Data implementations grows with the complexity of the datasets themselves. Datasets are likely of high volume, velocity, and variety, requiring the privacy guards to be applied at large scale, high speed, and with appropriate regard for variable levels of privacy concern.

Metadata Standards

One way to facilitate this is to develop standards allowing data to be affixed with appropriate privacy-guarding metadata early in its processing. This will eliminate, or at least reduce, the computational expense inherent in attempting to apply these guards to enormous pools of complex data later, when data has potentially been expanded, dispersed, and/or repeatedly summarized. Privacy policies should define a controlled vocabulary for this metadata, allowing application in broad sweeps across different sectors, both Government and private. These standards should be maintained transparently in the public record. This may require a high level of regulatory authority for technical prescription for successful implementation.

Defining Guarded Tiers

A tiered approach (Figure 1) could adequately frame policies in this arena in both public and private sectors, with technical safeguards in place to construct some number of "guarded" scenarios. With adequate definition and institution of these guards, public policy can provide incentives for their use by:

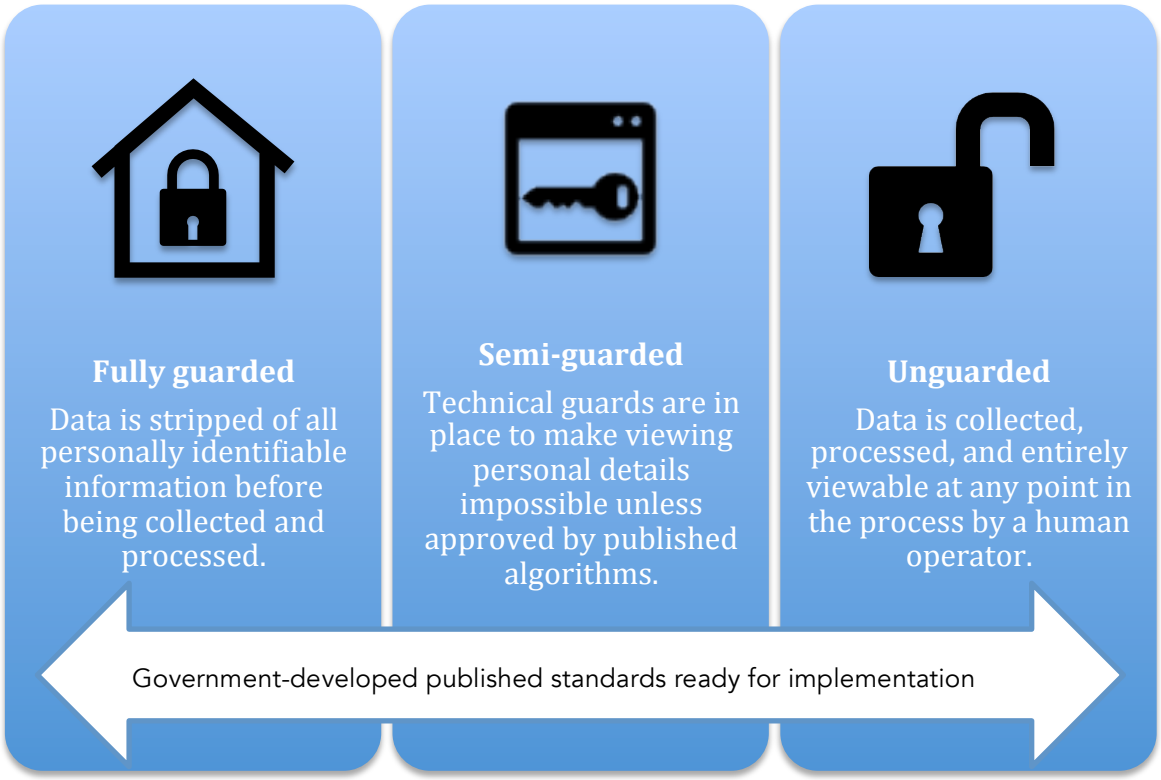
- reducing the difficulty of implementing data safeguards using published standards ready for implementation, and
- incentivizing safeguards while facilitating innovation and competition in industries utilizing data by defining tiers that combine minimal regulatory burden with minimal privacy risks, pushing industries away from privacy intrusion

Such data/system tiers may include:

- **Unguarded:** Data, including personally identifiable information, is collected, processed, and entirely viewable at any point in the process by a human operator. End results may produce specific lists of individuals, implicating them in some category (i.e. individuals to be contacted about certain promotional offers). Ideally, data can only make it into this scenario if described individuals granted deliberate and explicit approval for use of their data.

- **Semi-guarded:** Data, including personally identifiable information, is collected and processed, but technical guards are in place to make viewing personal details impossible unless some algorithmic conclusion is reached (see Suspicion Algorithms below for more explanation). This scenario may be especially useful in the realm of law enforcement and national security, where some governmental interest for public safety may necessitate utilizing data, but competing reverence for privacy of the law-abiding compels a measured approach to handling the data without just cause. Personal information in this case may be tokenized, with tokens relating back to real identities only through a heavily secured index list.
- **Fully guarded:** Data stripped of all personally identifiable information is collected and processed. This data would be used for scientific or commercial inquiry into public habits and disposition, but could not be used to target individuals in any way. Requirements for instituting this kind of data utilization would be minimal, but some regulation would be needed to minimize the possibility that data could be traced back to the individual it describes. There are some concerns with the practicality of truly anonymizing line item data records due to the ability to cross-reference data with publicly available personal identification data; addressing these concerns would require investigation and innovation to maintain the value of the data while potentially blurring the integrity of discrete records.

Figure 1: Potential Tiers for Data/System Safeguarding



The foregoing are only examples of how the privacy of data can be tiered. The principles could be applied to any number of tiers and any level of use prescription. Other factors that may be considered when defining tiers:

- nature of attributes included in data, i.e. biographical data, physical location data, consumption history, medical information
- purpose of data utilization, i.e. commercial targeting, commercial research, scientific research, law enforcement, national security
- provenance of the data, as determined by the nature of the source

Low-Level Implementation Concerns

Most Big Data architecture applies distributed and fault-tolerant software engineering techniques on clustered commodity hardware to achieve faster turnaround on computationally expensive data-processing tasks. This architecture presents special difficulties in applying common security measures. These difficulties include, but are not limited to:

- **Hardware heterogeneity:** Big Data compute clusters are likely to contain many kinds of hardware. The goal of the software paradigm was to enable collections of computers to be thrown into a networked cluster ad-hoc to increase its compute and storage capacity. These cluster nodes are likely to vary considerably in generation, CPU architecture, and any/all performance specifications, making it difficult to apply uniform "hardening" measures to a cluster on the node level.
- **Data redundancy:** To mitigate the high failure rate of variable commodity hardware, data is fragmented into manageable pieces and redundantly distributed across nodes of a cluster. The overhead associated with this data management makes the addition of typical security measures for data at rest, namely encryption/decryption, significantly more time-intensive, inhibiting the gains of applying the distributed techniques in the first place.

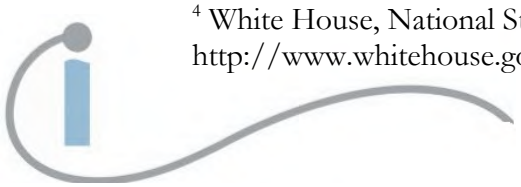
Government Vs. Private Sector

Q4: How should the policy frameworks or regulations for handling big data differ between the government and the private sector? Please be specific as to the type of entity and type of use (e.g., law enforcement, government services, commercial, academic research, etc.).

In contrast to the private sector, Governmental applications of big data, specifically those of law enforcement and national security, differ in that they:

- have the potential to greatly affect national security, as several high-profile lapses in our national security in the past 15 years were the result of inadequate dot-connecting among Government-possessed data records[4]

[4] White House, National Strategy for Information Sharing and Safeguarding http://www.whitehouse.gov/sites/default/files/docs/2012sharingstrategy_1.pdf



- lack elements of perceived explicit consent present in voluntary commercial services
- can result in dire losses if compromised; though both Government and private sectors have an element of opacity involved to maintain a competitive advantage, the Government is in competition with adversaries to protect life and limb

New ideas and public engagement are necessary to address an appropriate mitigation of these competing factors.

Suspicion Algorithms

As data solutions are used for routine law enforcement, a need may arise for a fortified legal framework for how to handle data collected and used for these purposes. If the results of advanced analytics become a precondition for visual inspection of data that would otherwise be hidden for privacy reasons, new legal precedents may need to be set as the autonomously computed equivalent of Reasonable Suspicion to justify that initial visual inspection. Ideally, these suspicion algorithms would be published for review by the public or by other Governmental branches.

Improving Big Data Outcomes

Q2: What types of uses of big data could measurably improve outcomes or productivity with further government action, funding, or research? What types of uses of big data raise the most public policy concerns? Are there specific sectors or types of uses that should receive more government and/or public attention?

Cases where Big Data has made marked contributions in application are scarce. Most stories of Big Data success are anecdotes with little hard evidence that the Big nature of the Data was inherent to the value produced[5][6][7]. For this reason, we believe developing a rigorous method for measuring the success of Big Data programs should be a major near-term goal. Big Data that possess any privacy concerns, but lack appropriate analysis to justify introducing those concerns, should be heavily scrutinized by public policy.

Developing a rigorous method for measuring the success of Big Data programs should be a major near-term goal.

The Government should follow precedent set by the Belmont Report and establish ethical principles and guidelines for public and private

[5] ZDNet *Why big data has (so far) failed medicine* http://www.zdnet.com/why-big-data-has-so-far-failed-medicine-7000021581/
[6] Scientopia *Debunking Two Nate Silver Myths* http://scientopia.org/blogs/goodmath/2012/11/09/debunking-two-nate-silver-myths/
[7] Pacific Standard *Obama's Not-So-Big Data* http://www.psmag.com/navigation/politics-and-law/obamas-big-data-inconclusive-results-political-campaigns-72687/

institutions utilizing Big Data. Guidelines developed in consultation with industry experts would serve not only to to protect consumers, but also to establish best practices and a quality standard by which Big Data solutions are evaluated.

Before implementing Big Data solutions, implementers must first be confident that Big Data can provide a substantial contribution to improve existing or new efforts. Given the potential privacy concerns and the considerable financial costs involved in implementing a Big Data solution, the Government should restrict funding to evidence-based Big Data solutions. To be considered evidence-based, Big Data solutions must undergo pilot tests evaluated against clearly defined success criteria. Only Big Data pilots that show demonstrable positive impacts and undergo a peer-review process should be approved for wide-scale implementation.

It should also be a major policy goal to distill the Big Data solution space in a number of archetypes with demonstrable value, to bring some concrete shape to the field of Big Data applications. A survey of credibly successful Big Data implementations should be conducted to begin enumerating these archetypes. It's difficult to assess, without this type of controlled research, which industries are furthest ahead in the implementation of Big Data initiatives.

Conclusion

The questions posed in the Government "Big Data" RFI are worthy of serious thought and investigation. This response provides our initial thoughts and comments on several areas of concern regarding Big Data and public policy, but it does not serve as a comprehensive assessment of the field. We applaud the White House's recent efforts to engage the community through a series of meetings and events with academic and corporate interests, and look forward to participating in future conversations.

Big Data Study
Office of Science and Technology Policy
Eisenhower Executive Office Building
1650 Pennsylvania Avenue, NW
Washington, D.C. 20502
VIA E-MAIL -- bigdata@ostp.gov

Re: Big Data Study, Document Number 2014-04660

March 31, 2014

Dear Ms. Wong,

Thank you for the opportunity to provide public comment in response to your comprehensive review of "big data" and its implications for privacy, the economy, and public policy. Access (https://www.accessnow.org) is a global organization dedicated to defending and extending the digital rights of users at risk around the world. Access works through its Policy, Technology, and Advocacy teams to achieve this mission. Access provides thought leadership and policy recommendations to the public and private sectors to ensure the internet's continued openness and universality and wields an action-focused global community of nearly half a million users from more than 185 countries. Access also operates a 24/7 digital security helpline that provides real-time direct technical assistance to users around the world.

I. The Challenges of "Big Data"

The growth in large-scale collection, retention, transfer, and analysis of personal data places everyone's privacy at risk. All types of organizations -- consumer-facing companies, third party data brokers, government agencies, and others -- develop comprehensive profiles at times containing identifying information, such as names, addresses, and phone numbers, as well as buying habits, personal interests, ethnic identities, political affiliations, marital status, credit card details, and numerous other data points.[1] Enough information is often collected that even anonymous information can be re-identified easily.[2] In one high-profile case, reporters were able to identify several anonymous users based solely on their AOL search history, which had been publicly released.[3] Information in one user's records provided detailed information on her medical history and love life.

There has been an exponential increase in the amount of data collected and stored by private companies in recent years. Facebook announced in 2012 that its data center had grown 2500x

[1] http://www.newrepublic.com/article/115041/what-big-data-does-and-doesnt-know-about-me
[2] http://www.forbes.com/sites/adamtanner/2013/04/25/harvard-professor-re-identifies-anonymous-volunteers-in-dna-study/
[3] http://www.nytimes.com/2006/08/09/technology/09aol.html?pagewanted=all

since 2008.[4] By 2012, Facebook was collecting about 180 petabytes of data per year. For reference, one petabyte is the equivalent of 20 million 4-drawer filing cabinets filled with text. Retailers, whether focused at online markets or off, also track customers. It is estimated that in one hour Wal-Mart processes about 1 million customer transactions containing 2.5 petabytes of data.

"Free" services offered by companies are often possible because these practices are part of a business model that relies on interpreting high-quality data about their users in order to serve revenue-generating targeted advertising. And over the years, many of these same internet companies have "simplified" their privacy policies by eliminating granular user-controls while increasing the capacity to track each and every online action.[5]

Data collection practices have been connected to specific practices that negatively impact internet users. For example, in 2012, it was discovered that some online travel booking companies, including Orbitz Worldwide Inc., were charging customers using Apple products close to 30% more for flights and hotels than visitors using Windows.[6] Such digital market manipulation leads to economic and privacy harms.[7] A recent breach of Target's systems is estimated to have affected up to one third of all Americans.[8] Ensuring that citizens have adequate knowledge and control over their data would greatly reduce the privacy and other human rights risks associated with big data. Currently, comprehensive standandards apply to medical and financial data, but not other types of sensitive information.

It is not only private entities where data collection has skyrocketed. Recent revelations have shown that US government intelligence agencies have been implementing programs to collect personal information and communications of users around the world at unprecedented levels. Some of these programs are implemented through legal processes, which compel companies to produce user information that the companies have otherwise collected for their own purposes. These collection programs are overseen by the secret FISA Court, which issues orders requiring production while preventing companies from publicly revealing that the collection has occurred.

Under other programs, often authorized under Section 702 of the FISA Amendments Act and Executive Order 12333, the US is tapping fiber optic cables directly (BLARNEY, OAKSTAR,

---

[4] https://www.facebook.com/notes/facebook-engineering/under-the-hood-scheduling-mapreduce-jobs-more-efficiently-with-corona/10151142560538920

[5] http://mattmckeon.com/facebook-privacy/

[6] http://online.wsj.com/news/articles/SB10001424052702304458604577488822667325882?mg=reno64-wsj&url=http%3A%2F%2Fonline.wsj.com%2Farticle%2FSB10001424052702304458604577488822667325882.html

[7] http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2309703

[8] http://www.nytimes.com/2014/01/11/business/target-breach-affected-70-million-customers.html

STORMBREW, FAIRVIEW),[9] breaking into the private links between corporate data centers (e.g., MUSCULAR),[10] or collecting the content of a whole country's phone calls (e.g., MYSTIC/RETRO).[11] Given the preponderance of attacks on the US Government, these mass surveillance places a tremendous amount of users and user data at risk.

II. The Problem of Unauthorized Access

Once collected, bad data security practices have led to the unauthorized access to and use of personal information, compromising users around the world. Data breaches are increasing in frequency. Last year saw the highest total records breached, according to a report by Risk Based Security.[12] In one incident, attackers obtained records with email addresses and passwords from around 152 million Adobe accounts.[13] In another breach, approximately 110 million Target accounts, about a third of the US, were affected by a data breach.[14] While the Adobe and Target breaches are two of the largest known breaches to date, data continues to be compromised with such great frequency that these incidents account for only a small portion of the total data that is known to have been exposed in 2013. Indeed, last year there were 2,164 incidents of data breaches with 822 millions records exposed reported worldwide. Attacks against US entities accounted for nearly half of all breaches globally.[15]

Unauthorized access to user data is not a new problem. For the past 12 years, identity theft has been the biggest source of complaints to the Federal Trade Commission,[16] which underlines that the identity and finances of citizens are consistently at risk due to needless collection practices and insufficient security practices employed by companies online. The economic impact of data breaches, and the accompanying reputational and legal fallout, is undoubtedly huge. Target spent $61 million in breach related costs in the first three months after the breach, which experts estimate may grow to as high as $1 billion.[17] Target's data breach is expected to be so expensive, in part, because it revealed data placing credit at risk.[18] That might be good for credit monitoring agencies, but it can create everyday challenges for victims when they try to get a mortgage, get a credit card, or buy a car. Data breaches are also particularly expensive in the US for the companies who lost or had records stolen. In 2012, companies paid on average $188 per lost or stolen record. That equated to about $5.4 million in loss for each entity with a data

[9] http://www.washingtonpost.com/business/economy/the-nsa-slide-you-havent-seen/2013/07/10/32801426-e8e6-11e2-aa9f-c03a72e2d342_story.html
[10] https://www.accessnow.org/blog/2013/11/01/nsa-hacks-internet-company-data-centers
[11] https://www.accessnow.org/blog/2014/03/20/nsa-bulk-collection-is-out-of-control
[12] https://www.riskbasedsecurity.com/reports/2013-DataBreachQuickView.pdf
[13] http://www.reuters.com/article/2013/11/07/us-adobe-cyberattack-idUSBRE9A61D220131107
[14] http://www.nytimes.com/2014/01/11/business/target-breach-affected-70-million-customers.html
[15] https://www.riskbasedsecurity.com/reports/2013-DataBreachQuickView.pdf
[16] http://www.ftc.gov/news-events/press-releases/2012/02/ftc-releases-top-complaint-categories-2011
[17] http://www.reuters.com/article/2014/02/26/us-target-results-idUSBREA1P0WC20140226
[18] http://www.usnews.com/news/articles/2014/03/26/jackpot-target-data-theft-victims-become-a-credit-agency-goldmine

breach.[19]

Governments also take advantage of insecure data. While the surveillance programs discussed above often operate under a system of compelled production, others skip official channels and, instead, use back doors. One such program is the "Upstream" programs alluded to in slides released in June 2013, and later confirmed by government officials. Upstream collection takes data right off the "backbone" of the internet -- the wires over which information is transmitted from computer to computer. Further revelations have brought to light backbone collection by US and other governments of remotely-activated webcam feeds,[20] e-mail contact lists,[21] and information on internal company networks.[22] It has also been revealed that the government has acted to preserve these collection programs by undermining data security standards.[23]

Unauthorized access or use of information by governments, as well as private actors, fundamentally threatens the internet as we know it. The world's largest internet companies build their business models around user trust in the networks that transmit and entities that store their personal data. Google's public Chief Legal Officer David Drummond, has said, "Our business depends on the trust of our customers." More acutely at risk, U.S.-based cloud computing firms spoke out after losing business following last summer's NSA revelations, and fear losing up to $35 billion in worldwide contracts as European regulators look to tighten restrictions on the cloud. Trust is also eroded when the NSA shares data with government agencies not dealing with foreign intelligence. For example, the NSA has provided evidence to the DEA, which then uses "parallel construction," whereby agents find alternative grounds to justify arrests and skirt legal challenges.[24] Rule of law is threatened when legal limitations fail to protect even the narrow existing privacy protections.

III. The Role of Data Security

As data are transferred from entity to entity, they become increasingly vulnerable, with more points at which unauthorized parties may be able to gain access to those data and use them for unintended purposes. Bad actors may compromise the financial or physical safety of users, and governments could use personal information to target dissidents, stifle speech, or influence

---

19 https://www4.symantec.com/mktginfo/whitepaper/053013_GL_NA_WP_Ponemon-2013-Cost-of-a-Data-Breach-Report_daiNA_cta72382.pdf

20 http://www.foxnews.com/tech/2014/02/27/uk-us-spies-hacked-webcams-millions-yahoo-users/

21 http://www.washingtonpost.com/world/national-security/nsa-collects-millions-of-e-mail-address-books-globally/2013/10/14/8e58b5be-34f9-11e3-80c6-7e6dd8d22d8f_story.html

22 http://www.washingtonpost.com/world/national-security/nsa-infiltrates-links-to-yahoo-google-data-centers-worldwide-snowden-documents-say/2013/10/30/e51d661e-4166-11e3-8b74-d89d714ca4dd_story.html

23 http://www.wired.com/2013/09/nsa-backdoored-and-stole-keys/

24 http://www.washingtonpost.com/blogs/the-switch/wp/2013/08/05/the-nsa-is-giving-your-phone-records-to-the-dea-and-the-dea-is-covering-it-up/

political outcomes.

Access has attempted to move the global conversation on security of big data forward. In March 2014, Access released the Data Security Action Plan.[25] In creating the Data Security Action Plan, Access considered what common-sense practices were needed to mitigate the extreme risk posed by the increasing amounts of data stored online. The Action Plan consists of seven steps that companies should take to protect their users. The seven steps are:

1. Implement strict encryption measures on all network traffic;
2. Executive verifiable practices to effectively store user data stored at rest;
3. Maintain the security of credentials and provide robust authentication safeguards;
4. Promptly address known, exploitable vulnerabilities;
5. Use algorithms that follow security best practices;
6. Enable or support the use of client-to-client encryption; and
7. Provide user education tools on the importance of digital security hygiene.

All entities should support the implementation of these security measures on all relevant data and networks under their control. Widespread adoption would benefit all internet users around the world, and would raise the floor on minimally-acceptable data security practices. If we fail to consider data security in the debate on big data public policy, we are standardizing unacceptable risks for users, companies, and the public at large.

IV. Conclusion

To mitigate the harms of data breach and misuse and to build user trust, the White House should consider what steps are necessary to protect user data. Companies should take proactive steps to protect user data. Specifically, this means adopting privacy-centered approaches to the collection and processing of user data, including: data minimization to limit collection of data where possible; ensuring that data is collected and stored for strictly defined purposes, and not used in a way that is incompatible with those purposes; and applying appropriate security measures to data both in transit and at rest.

Accordingly, Access calls on the government to bolster data protection standards, promote data security, and continue to foster a robust discussion on best practices.

Thank you for the opportunity to provide comment as part of this Big Data Study. For more information, please visit https://www.accessnow.org or contact the authors of this comment, Amie Stepanovich and Drew Mitnick, at amie@accessnow.org and drew@accessnow.org respectively.

[25] More information is available at https://encryptallthethings.net

WASHINGTON
LEGISLATIVE OFFICE

March 31, 2014

Office of Science and Technology Policy
Eisenhower Executive Office Building
1650 Pennsylvania Ave. NW.
Washington, DC 20502
bigdata@ostp.gov

RE: Big Data RFI – ACLU comments on the White House Big Data Initiative

Attention: Big Data Study

The American Civil Liberties Union (ACLU) writes today to describe some concrete, immediate steps the Obama administration can pursue to improve privacy and address the expanded use of personal information and big data.[1]

The Big Data Study Group has an enormous task. According to Senior White House Advisor John Podesta, in a mere 90 days, it will "deliver to the President a report that anticipates future technological trends and frames the key questions that the collection, availability, and use of "big data" raise – both for our government, and the nation as a whole."[2]

We commend the Big Data Study Group for the serious and focused attention it has brought to privacy issues in a short period of time. The first two workshops on the issue have been excellent explorations of some of the cutting edge ethical and legal challenges exposed by the accelerating collection of personal information. But we know that any 90 day review will only be a beginning in addressing big data. We also know that big data does not present wholly – or even mostly – new challenges. In reality these issues have been confronting policymakers since at least the 1970s, when the federal government developed the first version of the Fair Information Practice Principles.

AMERICAN CIVIL LIBERTIES UNION
WASHINGTON
LEGISLATIVE OFFICE
915 15th STREET, NW, 6TH FL
WASHINGTON, DC 20005
T/202.544.1681
F/202.546.0738
WWW.ACLU.ORG

LAURA W. MURPHY
DIRECTOR

NATIONAL OFFICE
125 BROAD STREET, 18TH FL.
NEW YORK, NY 10004-2400
T/212.549.2500

[1] The ACLU is a nationwide, non-partisan organization of more than a half-million members, countless additional activists and supporters, and 53 affiliates nationwide dedicated to enforcing the fundamental rights of the Constitution and laws of the United States. The ACLU's Washington Legislative Office (WLO) conducts legislative and administrative advocacy to advance the organization's goal of protecting privacy rights including use of information by government and the private sector.

[2] John Podesta. "Big Data and the Future of Privacy." *The White House Blog.* Jan. 23, 2014 Available at: http://www.whitehouse.gov/blog/2014/01/23/big-data-and-future-privacy

In fact, we already have solutions for some of the privacy issues that confront us today and there are specific actions the executive branch can take to improve Americans' privacy. With that goal in mind, the bulk of these comments will focus on two main areas. The first area is immediate actions the administration can and should take to improve how the federal government collects and uses personal information. The second area is a few specific subjects where sustained focus and attention could improve privacy knowledge and best practices in the future.

A hallmark of the ACLU is the breadth of our work on privacy and our expertise across a wide range of issues, including commercial data collection and use, law enforcement practices, and national security issues. As a final matter, we have prepared as an appendix to this letter a non-exhaustive review of recent ACLU reports and congressional testimony on privacy issues ranging from collection of phone record data by the NSA, to license plate readers, to immigration databases like E-Verify. We hope these will be a valuable resource for exploring specific subjects in more depth.

I. Subjects for Immediate Action

a. Support for Legislation

There is an almost universal acknowledgement that laws related to privacy are out of date. Technology has changed, but the law has not, creating serious gaps in privacy protections or leaving entire areas almost completely unprotected. Obviously Congress, not the Executive, passes new statutes; however there are three areas where the Administration could contribute to the legislative discussion in a way that would advance privacy.

Endorse the USA Freedom Act. Ongoing revelations regarding the use of big data to gather personal information on the American public by the National Security Agency (NSA) and other members of the intelligence community have highlighted the need for reform. The bipartisan USA Freedom Act reins in bulk collection of American records. It amends Section 215 of the Patriot Act – which is used to collect the phone records of almost every American every day – so that it can no longer be used in such a sweeping fashion. The bill would also require individualized suspicion for national security letters and pen registers, two other Patriot Act tools used to access Americans' records. The bill would make changes to the FISA Amendments Act (FAA) to prevent the government from searching through FAA-collected data for U.S. person data in the absence of an emergency or a court order. Finally, the bill includes the creation of a special advocate before the FISA court and new transparency requirements.

Two independent panels – the Privacy and Civil Liberties Oversight Board (PCLOB) and the President's Review Group on Intelligence and Communications Technologies – have already confirmed that intelligence agencies have interpreted their authority in an overbroad and unconstitutional manner. Further these panels "have not identified a single instance involving a threat to the United States in which the [215 telephone records] program made a concrete

difference in the outcome of a counterterrorism investigation."[3] While the President has just called for legislation to better protect Americans' phone records, our email, internet, financial and other records are just as sensitive and require stricter limitations. Endorsement by the Obama Administration of the USA Freedom Act would strengthen reform efforts and confirm that when big data runs amok it must be reined in.

Update the Electronic Communications Privacy Act (ECPA). Among many other protections, ECPA regulates how the government can access the contents of electronic communications. Unfortunately, it has not been substantially updated since 1986. Under the statute, e-mail, documents stored in the cloud, and other private communications like photos and text messages do not receive the protection of a search warrant approved by a judge (the protection that would apply to physical mail or even electronic communications that are not stored with companies like Google or Yahoo).

There is bipartisan legislation in the Senate and the House, S. 607 and H.R. 1852, which would make a simple fix to the law to assure that regardless of where individuals store their communications, those communications will be safe from unjust government intrusion and only accessible with a search warrant based on probable cause. Some areas must be shield from big data analysis without an appropriate legal predicate.

Seemingly, the only major impediment to passage is an objection by the Securities and Exchange Commission, which would like to use the legislation as an opportunity to expand its investigative authority. Support from the Administration would be a major step toward removing this roadblock.

Release Commercial Privacy Model Language. In February 2012 the Administration released a report outlining the need for a "Consumer Privacy Bill of Rights." The report delineated a strong framework for consumer privacy rights, one based on the fair information practice principles and resulting from a multiyear effort to identify and develop workable practices to make those principles a reality. The President committed, "My Administration will work to advance these principles and work with Congress to put them into law."[4]

Unfortunately, more than two years later, there has been no congressional initiative or filed bill. According to press reports, legislative language has been drafted.[5] This language should be

---

[3] Privacy and Civil Liberties Oversight Board. *Report on the Telephone Records Program Conducted under Section 215 of the USA PATRIOT Act and on the Operation of the FISC*. January 23, 2014. Available at: http://www.pclob.gov/SiteAssets/Pages/default/PCLOB-Report-on-the-Telephone-Records-Program.pdf (pg 11)

[4] The White House, *Consumer Data in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Economy.* February 2012. Available at: http://www.whitehouse.gov/sites/default/files/privacy-final.pdf (presidential introduction)

[5] For more information on this problem please see: Alex Byers. "White House pursues online privacy bill amid NSA efforts." *Politico*. Oct. 7, 2013. Available at: http://www.politico.com/story/2013/10/white-house-online-privacy-bill-nsa-efforts-97897.html

released, perhaps as an appendix to the Study Group's report. This language will help to advance the debate, giving activists and privacy supporters concrete reforms they can point to in the fact of continuing privacy invasions.

b. Administrative Reform

Today we live in a world of records. They are generated by electronic devices and stored in massive databases. It is easy to track each of us using cell phones, automated license plate cameras and a host of other technologies. Too often, the result is that the government stores, accesses, and uses personal information on innocent Americans without a reason. Any meaningful regulation of big data must begin by grappling with this reality by bringing transparency to these practices and then regulating or ending their use.

National security surveillance transparency. The long string of revelations regarding US government surveillance both here in the United States and abroad have highlighted how little the American public knows about these programs and the legal authorities that underpin them. It is critical that the administration to be more forthcoming. While we appreciate the President's transparency on the phone metadata program under section 215 of the Patriot Act, there is still a lack of information about the many other surveillance programs currently underway.

Three critical areas in which the administration can advance transparency are:

- release all remaining undisclosed FISA Court opinions;
- describe operational details, scope and legal underpinnings of existing surveillance programs; and
- address how many Americans have had their personal information swept up in these programs and how the government is using the vast amounts of data it is allegedly collecting.

Location and Telephone Record Information. Currently, records related to law enforcement requests for location information and telephone numbers are almost completely secret. In spite of the fact that tens of thousands of these orders are entered annually, Congress, the courts, nor the public has any clear sense of their scope.[6] These orders often reportedly collect not just information on subjects of an investigation, but also on dozens or hundreds of other, completely blameless, individuals.

The Department of Justice should develop a protocol to avoid indefinite sealing of surveillance orders ("D" orders and pen/trap orders). Specifically, a protocol should provide for:

- Immediate review of all sealed applications and orders under 2703(d) and the Pen/Trap statute, followed by DOJ filing motions in district courts seeking unsealing of all applications and orders that do not relate to a currently ongoing investigation or where unsealing will not result in imminent serious risk of physical harm or death to a person;

[6] For more information on this problem please see: Smith, Stephen W., Gagged, Sealed & Delivered: Reforming ECPA's Secret Docket (May 21, 2012). Harvard Law & Policy Review Vol. 6, 2012 Forthcoming. Available at SSRN: http://ssrn.com/abstract=2071399

- Create a prospective DOJ policy requiring US Attorneys' Offices to either seek unsealing of surveillance applications and orders within a reasonable time after an investigation is no longer active, or include a presumptive expiration date for sealing in the applications. For example a seal could expires 180 days after entry of the court's order, unless the government files a motion before that time certifying that the investigation is still active or that unsealing would cause imminent serious physical harm or death.

Automated License Plate Readers (ALPR). Tens of thousands of commercial and government license plate readers collect billions of records on American's location, often keeping that information for years.[7] This information is extremely sensitive, revealing an individual's location at a specific time and potentially where they worship, spend their nights or engage in First Amendment protected activities. Press reports indicate that several federal agencies, including U.S. Immigration and Customs Enforcement (ICE), the Drug Enforcement Administration (DEA) and the Federal Bureau of Investigation (FBI) are building their own ALPR databases or routinely accessing commercial databases.[8] Yet we have few details on how the federal government regulates access to this information.

The two key questions which need to be answered are:

- to what extent are federal agencies building vast databases of ALPR data, and
- to what extent is the federal government accessing data collected by others. These databases include not just private sector data collection but also databases compiled by state/local actors.

Commercial Databrokers. The Privacy Act does not extend to the federal government's use of commercial databases. As documented by a 2008 GAO Report, the federal government uses such databases frequently for a variety of purposes, such as in support of law enforcement and for background check investigations.[9] These databases often contain incorrect information, but individuals currently have none of the protections such as access, notice, correction, and purpose limitations, which are fundamental to the Privacy Act and fair information practices.

Federal agencies should examine and disclose their commercial data access, notice, correction and use policies and perform the same privacy impact assessments (PIAs) on the use of personal information in commercial databases that are already required on agencies' own databases. These PIAs would create basic transparency by requiring agencies to describe what information

---

[7] For more information see the ACLU's recent report on license plate readers: ACLU. *You Are Being Tracked*. July 2012. Available at: https://www.aclu.org/alpr

[8] Dan Froomkin, "Reports of the Death of a National License-Plate Tracking Database Have Been Greatly Exaggerated." *The Intercept*. March 17, 2014. Available at: https://firstlook.org/theintercept/2014/03/17/1756license-plate-tracking-database/

[9] U.S. Government Accountability Office. (2008, March). Government Use of Data from Information Resellers Could Include Better Protections. (Publication No. GAO-08-543T). Available at: http://www.gao.gov/products/GAO-08-543T

is collected, the purpose of the collection, with whom information will be shared, and how it will be secured.

Surveillance Drones. The federal government increasingly uses unmanned surveillance drones domestically. These small, inexpensive tools have the potential to dramatically increase aerial surveillance and are subject to few legal restrictions. Customs and Border Patrol flies a fleet up to a hundred miles away from both the northern and southern borders. It has also admitted to lending these drones to other federal, state and local law enforcement agencies. According to media reports, such practices have increased eightfold since 2010.[10] Former FBI director Robert Mueller disclosed that the agency uses drones but has yet to develop privacy protocols for that use.[11]

Except in exigent circumstances, agency drone use for criminal investigations should only be conducted pursuant to a particular investigation and after judicial approval. In non-criminal circumstances (such as patrolling the border) drones should not be 'loaned out' or used beyond their stated purpose by the federal agency authorized to use them.

Each of these proposals shares a common idea: that any program that collects data regarding the activities of a substantial number of people for a law enforcement or intelligence purpose, without any individualized suspicion, must be disclosed. When large swaths of people are subject to such collection, fundamental principles of democracy require disclosure so that there can be a public debate about the privacy tradeoffs. That principle should be applied broadly across the federal government.

II. Future Areas of Investigation

As the Study Group focuses on the future impact of big data, it should pay specific attention to two areas – the reality that data collection may exacerbate existing inequality and discrimination and effective research techniques that can be developed to protect privacy while allowing research to flourish.

a. Impact of Big Data on Exacerbating Inequality and Discrimination

The ACLU, along with 13 other civil rights, privacy and media justice organizations, is a signatory to five civil rights principles for the era of big data. These principles recognize the importance of data collection for documenting persistent inequality and discrimination but also seek to build an intellectual framework for assessing how surveillance and data use is being

[10] Craig Whitlock and Craig Timberg, "Border-patrol drones being borrowed by other agencies more often than previously known," *Washington Post*. Jan. 14, 2014. Available at: http://www.washingtonpost.com/world/national-security/border-patrol-drones-being-borrowed-by-other-agencies-more-often-than-previously-known/2014/01/14/5f987af0-7d49-11e3-9556-4a4bf7bcbd84_story.html.

[11] Jake Miller, "FBI Director Acknowledges Domestic Drone Use." *CBS News.* June 19, 2013. Available at: http://www.cbsnews.com/news/fbi-director-acknowledges-domestic-drone-use/

woven into the fabric of ordinary life, sometimes with harmful effects. The specific principles are:

1. Stop High-Tech Profiling.
2. Ensure Fairness in Automated Decisions.
3. Preserve Constitutional Principles.
4. Enhance Individual Control of Personal Information.
5. Protect People from Inaccurate Data.[12]

Data and surveillance are already part of ordinary life, especially in poor or disadvantaged communities that have long faced excessive government scrutiny. For example, a recent news story described an invasive, new police tactic employed by the Chicago Police Department.[13] Using software created by an engineer at the Illinois Institute of Technology, the city developed a "'heat list' — an index of the roughly 400 people in the city of Chicago supposedly most likely to be involved in violent crime." The criteria for placement on the list are secret, but reportedly go beyond indicators like criminal conviction, and raise real questions about racial bias in the selection process.

The results of placement can be very invasive. At least one person reported that a Chicago police commander showed up at his door to let him know the police would be watching him. He hadn't committed a crime or even recently interacted with police. This type of automated profiling is both a privacy problem and a civil rights problem. Use of personal information to make secret determinations is a violation of privacy. When there is significant potential for racial discrimination and police abuse, that's a civil rights problem.

The Chicago list is just the tip of an iceberg of dangerous ways that big data is being used. A Senate Commerce Committee report recently described marketers' use of lists based on racial and other characteristics to identify "the most and least desirable consumers."[14] The government E-Verify database, which many employers check to determine immigration status, has

[12] For a full description of the principles please see: "Civil and Human Rights Orgs Speak Out for the First Time on Privacy and Big Data Policy," Feb. 27, 2014. Available at: http://www.civilrights.org/press/2014/civil-human-rights-orgs-speak-out-on-big-data-privacy.html

[13] Matt Stroud. "The minority report: Chicago's new police computer predicts crimes, but is it racist?" *The Verge.* Feb. 19, 2014. Available at: http://www.theverge.com/2014/2/19/5419854/the-minority-report-this-computer-predicts-crime-but-is-it-racist

[14] U.S. Senate Committee on Commerce, Science, and Transportation, *A Review of the Data Broker Industry: Collection, Use, and Sale of Consumer Data for Marketing Purposes.* December 2013. Available at: http://www.commerce.senate.gov/public/?a=Files.Serve&File_id=0d2b3642-6221-4888-a631-08f2f255b577 (pg 25)

a persistent bias that causes legal immigrants to be wrongly identified as ineligible to work.[15] Police too frequently spy on innocent people who pray at mosques.[16]

All of these examples point to a growing need to consider how privacy and civil rights intersect. As one memorable article recently noted, often the best way to predict the future of surveillance is to ask poor communities what they are enduring right now.[17] The study group should consider these examples and the many others which accompany the principles as it assesses its future focus. While technology and computer analytics sometimes appear to be neutral, in fact they too frequently mirror persistent, existing inequality.

b. Research on Gaining From Big Data without Harming Privacy

In addition to the potential for inequality and privacy invasions, the use of big data may also have very real potential value – for improving medicine, combating climate change and addressing a host of societal ills. It is imperative that privacy not be pitted against those values. Data collection and use should not be a zero sum game where the increased value of data means a decrease in privacy. One of the ways to prevent that outcome is by supporting research which allows scientists to make use of data in noninvasive ways.

Computer scientists are developing new ways to share and analyze large datasets while strongly protecting privacy. One basic risk when removing personal information in order to share a sensitive dataset, is that the data might later be combined with outside information to reveal information on individuals. For example, the research of Dr. Latanay Sweeney has demonstrated that de-identified medical records can be combined with publicly available datasets to re-identify particular individuals and their medical conditions. But groundbreaking advances in differential privacy offer new tools to statistically measure and reduce that risk. The Census Bureau has already adopted differential privacy techniques, using them in its OnTheMap project to publish geographical information about where workers live and work without revealing anyone's specific employment.[18]

Other fundamental breakthroughs in cryptographic research are making it possible to reap the benefits of cloud computing while protecting sensitive information. Cloud computing services currently require access to their users' sensitive data, in order to analyze, search and present the

---

[15] ACLU. *The 10 Big Problems with E-Verify*. May 2013. Available at: https://www.aclu.org/10-big-problems-e-verify

[16] Noa Yachot. "With No Evidence of Wrongdoing, NYPD Treats Entire Mosques as Terrorsits Groups." *The ACLU Blog of Rights*. Aug. 28, 2013 Available at: https://www.aclu.org/blog/national-security-religion-belief-technology-and-liberty-criminal-law-reform/no-evidence

[17] Virginia Eubanks. "Want to Predict the Future of Surveillance? Ask Poor Communities" *The American Prospect*. Jan 15, 2014. Available at: http://prospect.org/article/want-predict-future-surveillance-ask-poor-communities

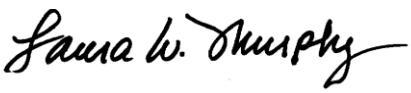
[18] Erica Klarrech. "Privacy by the Numbers: A New Approach to Safeguarding Privacy" *Quanta Magazine.* Dec 10, 2012. Available at: https://www.simonsfoundation.org/quanta/20121210-privacy-by-the-numbers-a-new-approach-to-safeguarding-data/

information. But a new family of approaches called homomorphic encryption may allow cloud providers to offer useful services together with strong privacy — searching and analyzing users' data without decrypting that data. This approach could leave users in control of their own information in a new and technologically robust way, and might also have beneficial legal consequences (because the user's key for decrypting her data may never have to be shared with any third party).

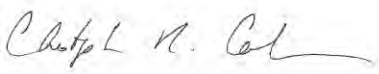
Strong research funding from the federal government will be critical in order to develop these potentially transformative, privacy-strengthening technologies to make them ready for widespread use. Closer collaborations between engineers and the policy community — such as the interactions fostered by this very review — will likewise remain vitally important as this research continues to develop.

We have attempted to offer manageable actions the executive branch can pursue in the short and long term to increase legal protections and transparency, reduce spying on innocent individuals, and address fruitful avenues for future research. The ACLU urges the Study Group to pursue each of these recommendations. For additional questions please contact Chris Calabrese at (202) 715-0839 or ccalabrese@aclu.org.

Sincerely,



Laura W. Murphy
Director, Washington Legislative Office



Christopher Calabrese
Legislative Counsel

Appendix A

- Section 702 of the FISA Amendments Act: Public Hearing Before the Privacy and Civil Liberties Oversight Board: (March 2014) (statement of Jameel Jaffer, Deputy Legal Director of the ACLU) Available at: https://www.aclu.org/sites/default/files/assets/pclob_fisa_sect_702_hearing_-_jameel_jaffer_testimony_-_3-19-14.pdf

- ACLU, U.S. Government Watchlisting: Unfair Process and Devastating Consequences (March 2014). Available at: https://www.aclu.org/sites/default/files/assets/watchlist_briefing_paper_v3.pdf

- Chris Conley, ACLU of Northern California, Metadata: Piecing Together a Privacy Solution (February 2014). Available at: http://www.datascienceassn.org/sites/default/files/Metadata%20Report%202014-%20Piecing%20Together%20a%20Privacy%20Solution.pdf

- The Future of Unmanned Aviation in the U.S. Economy: Safety and Privacy Considerations: Hearing Before the Senate Commerce Committee (January 2014). (statement of Chris Calabrese, Legislative Counsel of the ACLU) Available at: https://www.aclu.org/sites/default/files/assets/domestic_drones_statement_senate_commerce_committee.pdf

- Nicole Ozer and Matt Cagle, ACLU of Northern California, Losing the Spotlight: A Study of California's Shine the Light Law (November 2013) Available at: https://www.aclunc.org/sites/default/files/Losing%20the%20Spotlight%20-%20A%20Study%20of%20California%27s%20Shine%20the%20Light%20Law%20final.pdf

- Jay Stanley, ACLU, Police Body-Mounted Cameras: With Right Policies in Place, A Win For All (October 2013) Available at: https://www.aclu.org/technology-and-liberty/police-body-mounted-cameras-right-policies-place-win-all

- Jennie Pasquarella, ACLU of Southern California, Muslims Need Not Apply (August 2013). Available at: http://www.aclusocal.org/CARRP/

- Catherine Crump, ACLU, You Are Being Tracked: How License Plate Readers Are Being Used to Record Americans' Movements (July 2013). Available at: https://www.aclu.org/technology-and-liberty/you-are-being-tracked-how-license-plate-readers-are-being-used-record

- Strengthening Privacy Rights and National Security: Oversight of FISA Surveillance Programs: Hearing Before the Senate Judiciary Committee (July 2013) (statement of Jameel Jaffer, Deputy Legal Director of the ACLU and Laura Murphy, Director of the ACLU's Washington Legislative Office) Available at: https://www.aclu.org/files/assets/testimony.sjc_.073113.final_.pdf

- Jay Stanley, Senior Policy Analyst, ACLU and Chris Calabrese, Legislative Counsel, ACLU, Prove Yourself to Work: The 10 Big Problems with E-Verify (May 2013) Available at: https://www.aclu.org/files/assets/everify_white_paper.pdf

- State of Federal Privacy and Data Security Law: Lagging Behind the Times?: Hearing Before the Senate Committee on Homeland Security Subcommittee on Oversight (July 2012) (statement of Chris Calabrese, Legislative Counsel of the ACLU) Available at: https://www.privacysos.org/sites/all/files/Calabrese.pdf

March 31, 2014

Nicole Wong
Office of Science and Technology Policy
Attn: Big Data Study
Eisenhower Executive Office Building
1650 Pennsylvania Ave. NW
Washington, DC 20502

Sent via email to bigdata@ostp.gov

Re: Notice of Request for Information, "Big Data RFI," FR Doc. 2014-04660

Dear Ms. Wong,

The Advertising Self-Regulatory Council (ASRC), administered by the Council of Better Business Bureaus (CBBB), is pleased to submit these comments to highlight the importance of independent, public and enforceable self-regulation in spurring innovative uses of "big data" and maximizing the opportunities and free flow of information in and by the private sector while minimizing the risks to consumer privacy. Our comments explain how independent self-regulation can help in building trust between businesses and consumers in the data-driven world of today and tomorrow. Specifically, these comments highlight the work of the Interest-Based Advertising Accountability Program, which enforces the Self-Regulatory Principles for Online Behavioral Advertising, the Self-Regulatory Multi-Site Data Principles and the Mobile Guidance (collectively, the Principles).

We hope that this overview will be helpful in demonstrating (questions one, three and four) how self-regulation can help to address some of the policy implications of the uses of big data in the private sector. With respect to question two, government support could help to ensure that self-regulatory initiatives are used in fast-moving high technology sectors where regulation might not be flexible enough to respond to changes in practices and could stifle innovation. With respect to question five, as explained more fully below, self-regulation can help to create more consistent private sector norms.

About the Council of Better Business Bureaus

- The CBBB is the national arm of the 100+ member Better Business Bureau System

- The BBB system handles 1 million consumer complaints annually; rates businesses based on factors including responsiveness to consumer complaints and government actions provides consumer alerts; and accredits businesses based on uniform standards.

- BBB complaints comprise 21% of data in Consumer Sentinel database

About The Advertising Self-Regulatory Council (ASRC, formerly NARC)

- **In 1971**, the Association of National Advertisers (ANA), the American Association of Advertising Agencies (AAAA), and the American Advertising Federation (AAF) formed an alliance with the CBBB to create an independent self-regulatory body – the National Advertising Review Council (NARC).

- **ASRC** establishes the policies and procedures for advertising industry self-regulation, with programs tailored to provide industry best practices including:

    - National Advertising Division and National Advertising Review Board – 1971: Responded to consumers' concerns about truth and accuracy in advertising.

    - Children's Advertising Review Unit – 1974: Chartered to assure that advertisers would take special care in addressing advertising messages to a vulnerable audience. First Safe Harbor Program approved under COPPA

    - Electronic Retailing Self-Regulation Program – 2004: Developed at the request of ERA to meet special needs of fast-paced direct-response industry.

    - NAD & Council for Responsible Nutrition – 2007: Created in cooperation with CRN to expand NAD's review of dietary-supplement advertising and rein in outrageous claims.

About the Accountability Program

- The Accountability Program was developed at the request of the Digital Advertising Alliance (DAA), an alliance of the major advertising trade associations, as the independent third-party accountability agent for Principles.

- The Accountability Program is administered by the Council of Better Business Bureaus (CBBB), under the policy guidance of the Advertising Self-Regulatory Council (ASRC),

- The Accountability Program enforces industry-wide enforcement of the Principles to ensure that all companies provide consumers with transparency and choice with respect to their collection and use of data for online interest-based advertising and that these data are not used for eligibility determinations such as access to insurance, medical treatment, employment or promotion.

- The Accountability Program's mission is to build consumer trust in the Internet by enforcing the industry best practices set forth in the Principles on all platforms. Without consumer trust, the potential positive uses of the Internet for information exchanges and e-commerce cannot be effectively realized.

How the Accountability Program Enforces the Principles

- The Accountability Program monitors the OBA and MSD practices of all entities collecting or using multisite data for compliance with the Principles.

- The Accountability Program is assisted in this effort by it third-party contractor which uses a technology monitoring platform to generate reports on companies whose business models include OBA or MSD and through the research of its own staff to uncover and investigate potential issues of non-compliance.

- The Accountability Program brings informal and formal inquiries to resolve issues of suspected non-compliance.

- The Accountability Programs works confidentially and constructively with companies that have received its inquiries to resolve those compliance issues.

- At the end of a formal inquiry, the Accountability Program issues a public decision and an accompanying press release to ensure the transparency of its adjudications and inform industry and the public of the outcome of its inquiry.

- The company that was the subject of the inquiry includes a statement in the decision that publicly states its agreement to follow the Accountability Program's recommendations.

- If a subject of an inquiry declines to participate in the process or refuses to implement the Accountability Program's recommendations, the Accountability Program will refer the company to the appropriate federal or state government agency.

Overview of Monitoring and Investigation

Monitoring of Transparency and Consumer Control Compliance:

- The Accountability Program staff monitors the provision of real-time, enhanced notice of OBA and consumer choice by simulating browsing sessions and capturing session traffic through a debugging proxy server application (web debugger). This allows the staff to identify entities collecting data from a browser or serving interest-based (including retargeted) ads which is used to determine if the appropriate notice and choice are being provided.

Opt-Out Data:

- These data concern the functionality and expiration date of opt-out cookies dropped on a device when a covered entity's opt-out mechanism, accessed through their website, is used. The Accountability Program investigates broken opt-out links and opt-out cookies that expire in advance of the industry standard of five years.

Tracking after Opt-Out Data:

- Accountability staff monitor an entity's cookies (browser and LSO/flash) dropped on a device after previously acquiring the same entity's functioning opt-out cookies. The

Accountability Program uses these data to investigate potential cases of OBA tracking after a successfully processed opt-out request.

Privacy Policy Changes:

- Principle V, Material Changes to Privacy Policies, requires a company to obtain consent from affected individuals if the company's OBA data collection and use practices undergo a material change. As noted in the Principles, "a material change would be a more expansive collection or use of data than previously disclosed to the user. The Accountability Program looks for substantive changes to a covered entity's privacy policy that may result in a material expansion of OBA data collection and/or usage.

OBA Disclosure Compliance Monitoring:

- For publishers or first-party website operators, the Accountability Program is manually inspecting websites and looking for a footer notice that links to a disclosure of third party data collection and use for OBA on their website, as required by the Principles.

- For all covered entities, the Accountability Program is manually inspecting websites and is looking for a statement of adherence to the Principles in their Privacy Policy or OBA disclosure as well as information on opt-out mechanisms, as required by the Principles.

Accountability Program Advice and Consultation

- The Accountability Program continues to provide the DAA with independent, expert advice as new DAA Principles and Programs are created.
- Self-regulation is an iterative process, and the Principles have expanded since their inception to address issues of regulatory and consumer concern.

Why This Self-Regulatory Program is Successful

- **Substantive Principles:** The Principles directly and meaningfully address online privacy by providing consumers with real-time notice about interest-based advertising and enabling them to choose whether to participate.

- **Industry-Supported:** The Principles reflect consensus by the advertising industry as to appropriate conduct and practices and therefore are broadly supported.

- **Independently Enforced:** The Accountability Program's enforcement operates completely independent of the DAA, which has no advance knowledge or input into its monitoring or compliance processes.

- **Companies Held Public Accountable:** The outcomes of all formal inquiries are made public, giving policy-makers and the public confidence that industry is being held accountable for compliance with the Principles.

Highlights of Accountability Program Decisions and Administrative Dispositions

- Thirty-two decisions and administrative dispositions have been issued to date with 100 percent compliance (in addition to numerous informal resolutions of compliance issues).

- Decisions enforce the Principles against all members of the advertising ecosystem, including major brands, advertising agencies, web publishers and ad networks to ensure consumers have transparency and choice:

  > *The OBA Principles were developed to address consumers' concerns about the practices of companies collecting and using data for OBA. Consumers are not the only beneficiaries of the OBA Principles. The entire advertising ecosystem—including advertisers and ad agencies—benefit from vigilant, industry-wide compliance. By instructing their advertising agencies and ad network partners to adopt practices that comply with the OBA Principles, advertisers signal to their customers that they are committed to strong privacy practices.*
  >
  > *The Accountability Program's mission is to build trust between consumers and businesses by ensuring that all in the advertising industry comply with the OBA Principles. When an advertiser embraces self-regulation as Kia has done in this instance, it initiates a virtuous cycle with all members of the advertising ecosystem that support its ad campaigns*. Kia Motors America Decision

- Decisions interpret and enforce the Principles and adapt them to cover evolving practices and technologies such as device identification:

  > *As technologies continue to evolve and raise new compliance issues, the Accountability Program will respond to ensure that the OBA Principles are preserved and can extend to meet these novel situations. Companies' commitment to applying the OBA Principles to their new technologies will ensure that the OBA Principles continue to evolve along with technological advances. Technological innovation provides new challenges, but also can lead to innovative solutions that benefit consumers.* Blue Cava Decision

- Company statements in Accountability Program decisions demonstrate industry support of self-regulation and the Accountability Program's enforcement work:

Initiative supports the OBA Principles, and looks forward to helping advance their adoption throughout the advertising ecosystem.

Facebook is a strong proponent of the principles of transparency, consumer control, and accountability that are central to the Self-Regulatory Program for Online Behavioral Advertising (OBA Program) …. We appreciate the collaborative discussions that we have had with the Accountability Program about how to enhance transparency in FBX. We are pleased to confirm that in a further effort to promote transparency we will soon enable partners to use the AdChoices logo to indicate when FBX advertisements are interest-based. Specific Media takes privacy and self-regulation very seriously, and makes a concerted effort to safeguard consumer notice and choice regarding anonymous data collected for ad measurement and targeting purposes. Specific Media ….fully supports the Online Behavioral Advertising (OBA) Program, and the review process of the Better Business Bureau (BBB).

- Decisions, administrative dispositions/closures, and press releases available at:
  - http://www.asrcreviews.org/accountability-program-decisions
  - http://www.asrcreviews.org/2013/10/accountability-program-administrative-dispositions/
  - http://www.asrcreviews.org/category/ap/ap-press-releases/

Accountability Program Compliance Guidance

While trillions of the Ad Choices option icon are delivered on interest-based ads every month, the Accountability Program's monitoring of website compliance has revealed that a few of the other requirements of the OBA Principles are less widely understood and implemented by some businesses. The Accountability Program found that a significant number of otherwise compliant businesses were not meeting their obligation to give consumers a notice on every webpage where third parties collect information for interest-based ads. When clicked, this notice must take consumers to the exact place on the website where they can learn about the website's data collection practices and find an easy-to-use way to opt out from interest-based ads.

To remedy this compliance problem, the Accountability Program on October 14, 2013, released its first industry-wide Compliance Warning, clarifying the website enhanced notice requirement and notifying businesses that strict enforcement will begin on January 1, 2014. Compliance with this aspect of the OBA Principles is particularly important given the current sensitivities about undisclosed collection of online consumer data.

Accountability Program Compliance Guidance Page:
http://www.asrcreviews.org/2013/10/compliance-warnings-and-guidance/

International:

There are now DAA-related programs in Canada and the European Union, with discussions underway in Australia. Such self-regulatory efforts promise to help create greater interoperability in interest-based advertising, furthering the global economy.

The Accountability Program is providing consultation services to Advertising Standards Canada (ASC), which is the Accountability mechanism for the Digital Advertising Alliance of Canada's (DAAC) Online Behavioral Advertising Program. ASC is the national not-for-profit advertising self-regulatory body founded by the advertising industry in 1957. Since that time, ASC has demonstrated that advertising self-regulation best serves the interests of the industry and the public. This principle has guided its work and activities on behalf of its members, the public and the industry for over 50 years.

The DAAC has just released the Canadian OBA Principles, with enforcement beginning after an education program and a commercially reasonable time for companies to implement the technologies necessary to achieve compliance. The DAAC's OBA Principles are modeled on the DAA OBA Principles, but are adapted to be compatible with the requirement of Canadian privacy law and self-regulatory practices in Canada.

- The Accountability Program is advising ASC on interpreting the requirements of the Principles
- The Accountability Program is training ASC staff on technical monitoring techniques to ensure companies' compliance
- The Accountability Program will advise ASC on investigations when compliance begins
- The Accountability Program and ASC will be providing webinars to educate businesses and consumers about the DAAC and its OBA Principles and Program
- The Accountability Program also provides advice and training on OBA self-regulation to the Self-Regulatory Organizations (SROs) in the European Union which will enforce the European version of the DAA Principles in the 28 member states of the EU, under the auspices of the European Advertising Self-Regulatory Alliance, which is the umbrella organization for self-regulation in the European Union.

COPPA Compliance

- The Accountability Program is advising CARU on enforcement issues relating to the revised COPPA Regulation's definition of PI, as it relates to the use of persistent identifiers
- The Accountability Program is providing training for CARU staff on technical and other issues relating to enforcement of the requirements for notice and parental choice concerning collection and use of persistent identifiers and the contours of the exceptions to that requirement

- The Accountability Program is working with CARU to identify technical issues relating to a companies' compliance responsibilities and ambiguities in the application of the new definition

- The Accountability Program has partnered with CARU on a webinar for CARU supporters, featuring Evidon that explained the technical issues involved with compliance

Complaint Handling

The Accountability Program's experience with complaints indicates that enforcement is more comprehensive when complaints are supplemented by independent monitoring and investigation. Consequently, while the Accountability Program has received and responded to almost ten thousand complaints since its inception, routing them to the appropriate BBB complaint-handling source, the majority of its investigatory work is generated by its technology monitoring platform (TP) and the research of its skilled staff who verify leads provided by the TP monitoring data and independently investigate potential issues of non-compliance.
Table 1 below analyzes complaints received from late April 2011 through mid-October 2013. In that time frame, the program received a total of 7,444 consumer complaints. Of these complaints, 1,324, or 17.8 percent, address issues generally related to online advertising. An even smaller number of complaints, 283, or 3.8 percent of all complaints, described a practice that appeared to be facially relevant to online behavioral advertising (relevant complaints). Ultimately, only 148 of the complaints, or 2 percent, contained enough information to enable the Accountability Program to undertake investigations (actionable complaints).

Table 1

Complaints Received April 2011 - December 2013	**Total: 7,656**	
Complaints Related to Online Ads	1,431	18.7%
Complaints Unrelated to Online Ads	6,225	81.3%
Complaints Concerning OBA	305	4%
Complaints Not Concerning OBA	7,351	96%
Actionable Complaints	176	2.3%
Complaints Not Actionable	7,480	97.7%

Accountability Program Praised by Regulators

- **FTC Bureau of Consumer Protection Director Jessica Rich:** "We have long supported BBB's self-regulatory initiatives as an important complement to the FTC's law enforcement, policy, and educational initiatives. Over the years, the FTC has

emphasized that when implemented in tandem, self-regulation and government oversight provide valuable efficiencies and benefits.

In fact, well-constructed industry programs with certain hallmarks – (1) clear requirements, (2) widespread industry participation, (3) active monitoring, (4) effective enforcement, (5) procedures to resolve conflicts, (6) transparent and independent processes, and (7) responsiveness to changing markets and consumers – offer some clear advantages over government regulation alone.
They can be more prompt, flexible, and responsive than when we only enforce through statutes and regulations. They also can be better tailored to reach to particular categories of marketing or particular categories of businesses.
In a nutshell, strong self-regulatory programs provide important guidance to industry, alleviate some of the FTC's burden in monitoring for law violations, and develop workable standards that we all can draw on in future policy and enforcement efforts. And, of course, the FTC is here to serve as a regulatory back-stop when self-regulation fails to bring about compliance."

- **Former FTC Chairman Leibowitz:** "The online advertising industry has been working to develop an icon that consumers could click to opt out of receiving targeted ads. Today, although it is still a work in progress, the ad industry has obtained buy-in from companies that deliver 90 percent of online behavioral advertisements; and, with the Better Business Bureau, it has established a mechanism with teeth to address non-compliance, backed up with FTC enforcement."

- **FTC Commissioner Maureen Ohlhausen:** "Self-regulation works best when it is backed up by serious compliance efforts and tough enforcement. Through the Better Business Bureau and the Direct Marketing Association's efforts, the DAA is making enforceability a reality. I am especially pleased that DAA has not only adopted broadly applicable Principles that apply across sectors and to a wide variety of companies but also, building on the success of the advertising industry's approach to self-regulation, has provided for strong, objective oversight by the Council of Better Business Bureaus (CBBB) and the Direct Marketing Association. These programs provide for prompt follow up on complaints and, in the case of the CBBB, active monitoring of all members of the digital advertising system. Indeed anyone who has met with the CBBB Accountability Program's Director, Genie Barton, understands the seriousness with which she approaches this process. Already the CBBB's Accountability program has publicly reported on nineteen cases and, in all of these cases, the companies have agreed to voluntarily implement the program's recommendations.

  This approach can build public confidence in self-regulation by providing a public, transparent record of real efforts to ensure compliance more quickly and with fewer burdens than is sometimes the case with government enforcement. Though time consuming and arduous, regular compliance work improves the overall health of the online advertising industry by ensuring that companies live up to the promises they make. It also helps nip minor issues in the bud, correcting them before they become

serious problems for consumers. This frees up Commission resources to focus on truly bad actors. In a time of limited government resources, this approach is both efficient and sensible."

How the Accountability Program is Funded

The Accountability Program is administered by the Council of Better Business Bureaus. Accountability Program Employees are CBBB employees. The CBBB is reimbursed for the costs of administering the Accountability Program by the DAA.

RESPONSE TO REQUEST FOR INFORMATION
Big Data Review
79 FR 12251
DOCUMENT NUMBER 2014-04660
OFFICE OF SCIENCE AND TECHNOLOGY POLICY

RESPONSE FILED BY:
U.S. PUBLIC POLICY COUNCIL OF THE ASSOCIATION FOR COMPUTING MACHINERY

On behalf of the U.S. Public Policy Council (USACM) of the Association for Computing Machinery (ACM) we are submitting the following comments in response to the Request for Information (RFI) by the Office of Science and Technology Policy (OSTP) on the comprehensive review on big data announced in January 2014.

With over 100,000 members, the Association for Computing Machinery (ACM) is the world's oldest and largest educational and scientific computing society. The ACM U.S. Public Policy Council (USACM) serves as the focal point for ACM's interaction with U.S. government organizations, the computing community, and the U.S. public in all matters of U.S. public policy related to information technology. Our comments are informed by the research experience of our membership. Should you have any questions or need additional information, please contact our Public Policy Office at 212-626-0541 or at acmpo@hq.acm.org.

We welcome the review of issues connected to the intersection of big data and privacy. The concept of big data is still emerging, but it is not too early to review how changes in the ability to collect, analyze and use large amounts of information provide new challenges and opportunities. While the definition of big data used for this RFI focuses on datasets so "large, diverse and/or complex, that conventional technologies cannot adequately capture, store and analyze them" the questions in this RFI (and our responses) can be applicable to large datasets currently captured by conventional technologies. The ability to effectively analyze collected data typically follows the ability to capture and store such data. As capabilities change, it will be important to systematically revisit datasets as our analytical abilities advance, both concurrently and after data collection and storage.

Answers to specific questions in the RFI

(1) What are the public policy implications of the collection, storage, analysis, and use of big data? For example, do the current U.S. policy framework and privacy proposals for protecting consumer privacy and government use of data adequately address issues raised by big data analytics?

The rise of big data highlights tensions that have existed in U.S. efforts to protect consumer privacy and limit government use of data. While the Fair Information Practice Principles

(FIPPs) (which are part of the USACM recommendations on privacy practices[1]) continue to have utility, trends in big data have made certain practices less effective. It has become significantly easier to extract personally identifiable information from nominally de-identified data as more data becomes available. In recent years academic researchers have shown that many data sets thought to be "de-identified" or "anonymized" can be re-identified when the data are correlated with other information that is publicly available.

Restricting data collection to specified purposes runs counter to the efforts of the government and companies to repurpose already collected data. The multitude of purposes for collected data have added to the complexity of privacy policies, documents that consumers find harder to understand. It also complicates efforts to give notice and obtain consent for collected data and associated uses. If big data trends continue to focus on multiple uses of collected data, new tools and processes for protecting and securing that data are needed to augment the protections currently available. Policies on protecting personally identifiable information, handling data breaches, and ensuring data quality become more important – not less – in an era of big data.

That obtaining notice and consent is increasingly difficult does not imply that the unrestricted repurposing of unchecked data is the way forward.

First, the more use that is made of collected data, the more important it is that the data be accurate. If collected data is going to be used for multiple purposes, making sure that data is accurate has additional benefits for the consumer, the data collector and other parties that may reuse that data. An important tool for this is to provide consumers the ability to check that collected data is indeed accurate, and to correct or dispute inaccuracies.

Tracking the flow of data will also become more important as it will likely change hands more frequently, and the ability to correct that data will need to keep up. Making sure the data remains accurate, secure and uncorrupted is important, and some encouragement may be needed for effective controls to be established and/or augmented for big data.

Such efforts would benefit from a sustained effort to develop approaches, technology and standards for the systematic tracking of data provenance and metadata. Agencies could help by funding specific research in this area and by sponsoring forums to discuss and ultimately adopt FIPS for provenance and metadata.

Separately, one could include the ability to remove data and/or opt out of (or consent to) specific uses. The right to do so may depend on who has supplied it; information a consumer supplies, or which is collected from that consumer's online behaviors, may deserve more control than, for example, reports of criminal convictions by that individual. Accuracy, in contrast, including deletion of convictions (where judicially appropriate) is more critical than corrections about consumer information. Some practices that have

[1] http://usacm.acm.org/privsec/category.cfm?cat=7&Privacy%20and%20Security

proven infeasible under current practices may be possible through the cloud, with appropriate policies.

Since big data in the context of this RFI is concerned with unconventional technologies, we recommend that new policies, procedures and best practices be as technology-neutral as possible. Technologies and methods will change more quickly than policies in the public and private sectors will be able to adapt.

(2) What types of uses of big data could measurably improve outcomes or productivity with further government action, funding, or research? What types of uses of big data raise the most public policy concerns? Are there specific sectors or types of uses that should receive more government and/or public attention?

Lessons learned in large archival efforts to date (e.g., libraries, music) may be applicable to big data. Ensuring the quality of collected data on par with archived data can make it easier to migrate data between users, allowing for additional uses of collected data. As recognized in the May 2013 executive order and formalized in policy directive M-13-13, standardization of data formats (preferably in open data formats that are machine readable) must be encouraged as well. This will reduce expenses for public and private parties interested in big data, and make data more accessible and readable for any interested party.

Areas of public policy concern related to big data include intellectual property connected to large datasets. The ability to conduct effective research on large datasets could be hampered by limits of access to the data, or restrictions on the ability of research resulting from these datasets to be properly reviewed and replicated. Policy makers should consider incentives to require researchers to make data available for public benefit (e.g., medical research); in some cases this might be required as a quid pro quo for using data about individuals without direct personal benefit.

A useful distinction to make for big data and public policy is between data generated that is somehow related to human activity and data generated that is not. When humans are the subjects of large datasets, a higher level of privacy and security controls will be needed to protect the sensitive data collected. This is particularly true of online behaviors that are increasingly tracked, correlated and shared between corporate entities.

(3) What technological trends or key technologies will affect the collection, storage, analysis and use of big data? Are there particularly promising technologies or new practices for safeguarding privacy while enabling effective uses of big data?

The challenges of data de-identification (including the increasing ease of re-identification) will be important to big data. Being able to de-identify data is a common practice intended to protect the privacy of subjects of collected data. It is important to remember that de-identification is not a binary state, but a spectrum. It is most appropriately viewed as a single privacy risk control rather than a technique that renders privacy concerns irrelevant. As such, its use does not in and of itself constitute a convincing argument for completely removing such data from a risk management regime or from applicable regulatory

oversight. Anonymization and de-anonymization are active areas of research, changes in technology and methods will make current de-identification standards obsolete in ways that cannot be predicted.

Techniques are needed for a consumer to express their preferences, without needing to confront the complexity of all possible circumstances of use. Informed consent is sometimes taken to mean both control such that one's intent is realized, and also control (or at least understanding) over all details. These can conflict, and it may be best to separate them.

Computer security setup may suggest a way forward. Users can configure Microsoft Windows security *approximately* right, by answering a handful of questions about their attitudes and tradeoffs; as systems gain additional controls, experts provide rules to generate new settings. Compared with multipage legalese or settings on dozens of complex features, such tradeoff responses give a less precise but more understandable picture of system behavior.

Data quality issues will matter as well. Besides the reasons mentioned in our response to question 2, preserving the quality of data helps make it more useful to the entities looking to reuse it. That data is accurate and reliable will help make analysis easier, and depending on the datasets and analyses involved, could address the challenges of false positives and false negatives. The utility of metadata is linked to its sensitivity, and both require attention to ensure the quality of the underlying data.

Inexpensive means of providing durable homes for research data should be encouraged. The availability of cloud resources for purchase and lease provides an opportunity to improve data durability. When a research project terminates, its hardware and hence its data traditionally became unavailable. One needed mechanism is to fund archiving, a process to identify data to be archived (if raw data is too large or too sensitive), and connection to a new technical system to do the archiving. With a storage cloud, the data can stay in place, so the last barrier is substantially reduced. (Cloud providers should be required to ensure that data is protected if the provider goes out of business).

Finally, new technologies allowing individuals greater control of data collected about them, especially in online contexts, is an important area of current computing research. New technologies are being explored ranging from "data vaults" to renumeration systems for allowing various kinds of data sharing (returning value to the consumer when the data is used). These technologies are being coupled with other technologies to allow users to permit their data to be shared at various levels of aggregation, while controlling the direct sharing of personal information. These aggregation and control technologies hold promise for safeguarding privacy in the big data era, and the government should encourage their creation and usage.

Insights | Collaboration | Advocacy

Leading the Marketing Community

March 31, 2014

Big Data Study
Office of Science and Technology Policy
Eisenhower Executive Office Building
1650 Pennsylvania Avenue, NW
Washington, DC 20502

SUBJECT: Request for Information on "Big Data"

Pursuant to the March 4, 2014 *Federal Register* Notice by the Science and Technology Policy Office, the Association of National Advertisers (ANA) appreciates the opportunity to provide these comments on the issue of "big data" involving consumers, businesses and our entire economy. These comments are addressed primarily to Question (1) in the Request for Information *"What are the public policy implications of the collection, storage, analysis, and use of big data? For example, do the current U.S. policy framework and privacy proposals for protecting consumer privacy and government use of data adequately address issues raised by big data analytics?"* We also include comments on other big data matters of interest and concern.

The ANA (Association of National Advertisers) provides leadership that advances marketing excellence and shapes the future of the industry. Founded in 1910, ANA's membership includes more than 575 companies with 10,000 brands that collectively spend over $250 billion annually in marketing and advertising. The ANA pursues "collaborative mastery" that advances the interests of marketers and promotes and protects the well-being of the marketing community. For more information, visit www.ana.net.

In General

ANA believes a number of points deserve emphasis at the outset:

- We agree with the *Federal Register's* Request for Comments definition of "big data:" "Big data' refers to datasets so large, diverse and/or complex, that conventional technologies cannot adequately capture, store or analyze them."

- This definition, however, focuses solely on the quantity and/or complexity of the data involved and the technological capacity to handle, store and analyze it. The definition, therefore, provides no guidance as to the policy implications of any dataset regardless of size.

- The amount of data in and of itself is not determinative of potential concern. Rather, the focus should be on the sensitivity and potential vulnerability to harm of any dataset.

- Various principles (e.g. transparency, accountability, consumer control) apply in both the big and smaller data contexts.

- While big and small data share these characteristics and concerns, they are not identical.

- All information is not created equal; some (health, financial) information is far more sensitive than other data and requires different treatment.

- Data security is, of course, a consistent fundamental interest that substantially impacts all data whether big or small.

- In the commercial arena, the use of data for advertising purposes is a driving force in the U.S. economy. It helps generate employment, sales and economic activity throughout various industrial and other sectors and in every geographical region in our nation.

- Increasingly, advertisers utilize data to provide greater and more relevant information to consumers. This information enables consumers to make informed choices in the marketplace, which helps to enhance economic efficiency, innovation and competition.

- For the purpose of effectively placing advertising, it is often not necessary to have or utilize personally identifiable data.

- Because of the critical role advertising plays in regard to our economy, it is essential that governmental decisions about commercial data collection and use be made carefully, correctly and judiciously.

- It is difficult to see how broad or comprehensive new privacy laws or regulations at the present time could keep pace with the revolutionary and extraordinarily rapid transformation of the Internet and other new media technologies.

- Advertisers have established various major self-regulatory mechanisms to address privacy and other issues related to data collection and use. Especially in a period of accelerating evolving technology, new data-related privacy laws and regulations should be initiated only where it is conclusively demonstrated that existing laws, regulations and industry efforts are clearly inadequate.

Data Provides Tremendous Value to the U.S. Economy

The appropriate collection and use of data provides increasing benefits for our nation's economy, the business sector and consumers. Every industry in America relies on data-driven marketing. One of our industry partners, the Direct Marketing Association (DMA), recently released a study that found that the data-driven marketing economy added $156 billion in revenue to the U.S. economy and fueled more than 675,000 jobs in 2012. The study also found that 70 percent of the value of the data-driven marketing economy depends on the ability of firms to exchange data across the marketplace. The DMA study is available at: http://www.the-dma.org

Advertising is a driving force in the U.S. economy, serving as a generator of job creation and sales. According to a 2013 landmark study conducted by IHS Global Insight, Inc., a highly regarded consulting firm, advertising is a remarkably powerful economic force. Nationally, it generated over $5.8 trillion in economic activity in 2012, or approximately 20 percent of total U.S. economic activity. Sales of products and services stimulated by advertising supported 19.8 million jobs, or 15 percent of the total jobs in the country. The study was based on an economic model developed by Dr. Lawrence R. Klein, recipient of the 1980 Nobel Prize in Economics. More information about the study is available at: www.ana.net/content/show/id/29212

Another Nobel Laureate in Economics, the late Dr. George Stigler, noted that advertising is a critical force in fostering economic efficiency and competition throughout the U.S. economy. In addition, the economic health of most of our country's media, including the online marketplace, rests primarily on the strong financial foundation provided by advertising.

As advertising fuels the Internet engine, advertisers have a great economic interest in the appropriate collection and use of data and are actively engaged in satisfying the privacy concerns of consumers in both the offline and online world. Advertising increasingly is a data-driven industry. It is extremely important, however, that marketers are able to obtain accurate anonymized consumer data, or Internet users will be besieged by unwanted and irrelevant advertisements, often labeled spam and in those circumstances, the Internet would be inundated with a profusion of unwanted and unnecessary traffic.

Not All Data is Created Equal

The United States has historically taken a sectoral approach to privacy regulation, adopting carefully defined rules to apply to specific categories of information. As a result, there are more than ten separate federal regulatory privacy regimes, including: the Children's Online Privacy Protection Act (COPPA), the Cable Communications Policy Act, the Telephone Consumer Protection Act, the Video Privacy Protection Act, the Gramm-Leach-Bliley Act, the Fair Credit Reporting Act, and the Health Insurance Portability and Accountability Act.

This sectoral approach is based on the fact that not all information is created equal. For example:

- Personally identifiable information clearly raises potentially far more serious privacy implications for consumers than non-personally identifiable information.

- Greater sensitivity concerning certain kinds of information, such as medical or financial information, is required. Consumers rightfully expect that this type of information should be more protected because the potential harm from disclosure is greater than for other kinds of information, such as the color of a shirt ordered online by a consumer.

- Information from and about children must also be treated differently. Marketers agree that children deserve greater privacy protection. ANA and others in the business community worked closely with the Congress and the Federal Trade Commission (FTC) to develop COPPA, which provides parents substantial control over this type of data collection.

A majority of the data collected for marketing purposes is anonymous or anonymized and is not sensitive information. Any consideration of government restrictions on the collection and use of this data must include a serious analysis and delineation of any harm that might occur if this information were inappropriately released. Would there be real, rather than theoretical, harms, and what steps could be taken to address them?

We cannot assume that the mere collection and use of data is harmful or will encroach on individual privacy. Major steps have been taken by industry to address specific privacy concerns. The Digital Advertising Alliance (DAA), formed by ANA and four other industry groups, for example, has set forth seven guiding principles designed to address consumer concerns about the use of information, while preserving the innovative and robust advertising that supports the vast array of free online content and the ability to deliver relevant advertising to consumers. These principles apply in both the big and small data contexts, and address the following:

- Education: educate individuals and businesses about the collection and use of data.

- Transparency: provide clear and easily accessible disclosures to consumers about data collection and use practices.

- Consumer control: ensure the ability for consumers to choose whether data is collected and used for specified purposes, including the obtaining of consumer consent in certain circumstances.

- Data security: provide appropriate security for, and limited retention of, data collected and used.

- Material changes: obtain consumer consent before a material change is made in certain collection and use policies, where such change would result in more collection or use of data.

- Sensitive data: recognize that data obtained from children merits heightened protection, and require parental consent for data collection under 13.

- Accountability: develop programs to advance these principles, including programs to monitor and report instances of non-compliance with these principles.

These principles are available at: http://www.aboutads.info

In addition, in the area of Web viewing data, the DAA has set forth a clear framework governing the collection of online Multi-Site Data that also provides consumer choice for the collection of such data. These Principles clearly and explicitly prohibit the collection or use of Multi-Site Data for the purpose of any adverse determination concerning employment, credit, health treatment or insurance eligibility. Additionally, the Multi-Site Data Principles provide specific protections for sensitive data concerning children, health and financial data. These and other industry efforts are directed to very specific potential harms. The principles are available at: http://www.aboutads.info/msdprinciples

The private sector has made substantial progress over the past several years to enhance the level of privacy protection for consumers. At the urging of ANA and other industry groups, almost every major commercial website has adopted and posted privacy policies to tell consumers how information is collected and used. Companies are innovating in the area of privacy and offering consumers new privacy features and tools such as sophisticated preference managers, persistent opt-outs, universal choice mechanisms and shortened data-retention policies. These developments demonstrate that companies are pro-active as well as responsive to consumers, and focusing on privacy as a way to distinguish themselves in the marketplace.

Even in the sensitive area of health information, there are examples of the effective and appropriate use of large scale databases to achieve very positive results while protecting the privacy of consumers. Dr. Michael Nguyen, the Acting Director of the Division of Epidemiology in the Food and Drug Administration's Center for Biologics Evaluation and Research, recently wrote a blog entry describing the use of extensive medical databases to evaluate the safety and effectiveness of prescription medications: "FDA scientists have partnered with the Harvard Pilgrim Healthcare Institute to create such a surveillance system, called Sentinel. Within Sentinel, FDA has supported the development of software that analyzes information from health insurance and health record databases to search for evidence that certain products are linked to specific adverse effects. Although these data are protected behind tight firewalls and remain under the control of the original health insurance plans that created them, the software makes it possible to analyze the information without disclosing identifying information in order to strictly maintain patient privacy." http://blogs.fda.gov/fdavoice/index.php/2014/03/sentinel-harnessing-the-power-of-databases-to-evaluate-medical-products/

ANA does not believe there is a present need for broad new federal privacy legislation. Government regulation does not have the flexibility to adapt effectively to the exceptionally rapid changing technologies and new privacy issues that the Internet and other new media have generated. Rather, we believe that consumers can be best protected through a

combination of existing privacy laws and regulations, privacy enhancing technology, effective and muscular self-regulation and the ultimate backstop of the powers of the FTC to stop false, deceptive or unfair acts or practices.

The Internet of Things

It is clear that the discussion of data collection and use has expanded with the growth in data. For example, the ability of everyday devices, from cars to smart home appliances, to communicate with each other and with people is becoming more prevalent and raises new privacy and security concerns. This development is often referred to as "The Internet of Things."

The FTC held a public workshop last November to explore the consumer privacy and security issues posed by the growing connectivity of devices. In opening remarks, FTC Chairwoman Edith Ramirez stated: "As I see it, the expansion of the Internet of Things presents three main challenges to consumer privacy: first, it facilitates the collection of vastly greater amounts of consumer data; second, it opens that data to uses that may be unexpected by consumers; and third, it puts the security of that data at greater risk."

Chairwoman Ramirez concluded: "With big data comes big responsibility." ANA agrees with this comment, and notes that the advertising community has been actively addressing various issues related to the collection of vast amounts of data.

We believe that in the context of advertising that the DAA model provides the right template for action that can be modified and expanded to meet the privacy concerns of new media and new technologies.

Data Security is Essential

Big or small, data must be secure. Entities that collect and use data must make every effort to ensure that mechanisms are in place to guarantee the integrity of the data.

Larger amounts of data may make the challenge of securing such data more difficult. Larger data fields are available for trolling by unscrupulous entities, and those collecting and storing data will likely face greater challenges with large volumes of data than in the small data context.

The public and government have legitimate privacy concerns related to database hacking that could result in health, financial or other sensitive information ending up in the wrong hands. These are serious issues and there are several bills pending in the Congress and other Congressional activities related to data security issues. For example, on March 26th, the Senate Commerce Committee held the latest of numerous hearings on recent data breaches.

We believe Congress should pass data security legislation which establishes one uniform standard and preempts the 47 different state laws on data security and breach notification.

The Need for Accurate Data

Business in general and advertisers in particular are faced with a new and very dangerous challenge -- web traffic fraud – which does not discriminate based on the amount of data involved. The Internet Advertising Bureau (IAB) recently estimated that approximately 36% of all Internet activity is fraudulent, resulting from viruses and mechanisms that direct computers to particular sites. This "bot traffic" costs advertisers significant amounts, since advertisers pay for the placement of ads that are loaded in response to users visiting Web pages. Forty-six percent of online ads are served to websites and charged to marketers even though consumers never saw the ads. Additionally, advertising inadvertently supports rogue websites that deliberately pirate movies, music and other intellectual property.

The advertising community is responding to these and other measurement challenges through an initiative known as Making Measurement Make Sense (3MS). This initiative, launched in 2011 by the ANA, the IAB and the 4A's, is an industrywide effort to develop effective and accurate advertising metrics that will enhance evaluation of digital media and facilitate cross-platform measurement. More information about 3MS is available at: www.ana.net/content/show/id/d3ms

Data, whether it is "big" or "small," if it is inaccurate or corrupted, is of substantially diminished value both for industry and consumers. This is a growing challenge for the advertising community and for the Internet-based economy as a whole.

Commercial Privacy Issues Must Not be Conflated With Government Privacy Issues

Recent disclosures about surveillance of citizens by the National Security Agency (NSA) as well as several high-profile data breaches from major retailers and other companies have combined to substantially increase the focus of policymakers, the business community and consumers on data security and privacy issues.

In his January 17th speech on NSA reforms, President Obama stated: "The challenges to our privacy do not come from the government alone. Corporations of all shapes and sizes track what you buy, store and analyze our data and use it for commercial purposes. That's how those targeted ads pop up on your computer and your smart phone periodically."

Commercial privacy issues must not be allowed to be conflated with government surveillance and potential reforms at the NSA. These issues must not be confused with interest-based advertising or online behavioral advertising (OBA). The privacy issues in these two areas are very distinct and deserve careful separate consideration. In addition, government access to private sources of data must be considered carefully. Whoever holds collected data must have in place security mechanisms to ensure the safeguarding of data, and methods for

government obtaining such data must be clearly enumerated and subject to close supervision.

Interest-based advertising, which under the auspices of the DAA self-regulatory program primarily utilizes anonymous or anonymized data, is a critical tool to reach the right consumer at the right time with the right message and often at the right price. It provides enormous economic efficiency to the Internet marketplace; provides tremendous competitive benefits; and provides consumers with relevant marketing information rather than spam. Data fuels the engine of this online marketplace.

To address the privacy concerns that some have about interest-based advertising, the marketing community has built one of the most rapidly-growing and successful self-regulatory programs in history – the DAA. The DAA features an icon alerting consumers to the fact that they have been served an ad based on OBA. From this icon, consumers can access information about interest-based ads and learn how to exercise choice about how to opt-out of those interest-based targeted ads if they choose to do so.

Since its launch in 2010, the DAA has rapidly brought enhanced notice and choice to consumers:

- The AdChoices Icon is now served globally trillions of times each month
- 30 million unique visitors have visited our two program sites, www.aboutads.info and www.youradchoices.com
- 3 million unique users have exercised an opt-out choice on our Consumer Choice Page

These are compelling numbers which show that consumers are coming to rely on the DAA program for meaningful choice.

On February 23, 2012, at a White House event announcing President Obama's framework for privacy in the 21st century, the Chairman of the FTC, the Secretary of Commerce and White House officials publicly praised and endorsed the DAA's initiative.

Enforcement of the program is administered by the Council of Better Business Bureaus (CBBB) and the Direct Marketing Association. We believe that strong industry self-regulation is a superior alternative to restrictive new laws and regulations. One of the most important benefits of self-regulation is the flexibility to adapt to changing technologies, consumer behaviors and attitudes. In mid-2013, the DAA principles were extended to cover interest-based ads delivered across mobile applications and the mobile Web. The DAA guidance also addresses location-based data and personal directory data use.

The DAA's self-regulatory principles of choice and transparency for the collection and use of web-viewing data have been adopted by 31other countries, through the Digital Advertising Alliance of Canada and the European Interactive Digital Advertising Alliance.

Two public opinion surveys have found that Internet users recognize the value of online advertising and our industry's self-regulatory program. The first survey, commissioned by DAA and conducted by Zogby Analytics in early 2013, measured attitudes regarding online advertising with a specific focus on interest-based ads. Nearly 70 percent of respondents said they would like at least some ads tailored directly to their interests, compared to only 16 percent who preferred to see only generic ads. More than 90 percent of those surveyed said that free content was important to the overall value of the Internet. More than 75 percent said they prefer content like news, blogs and entertainment sites to remain free and supported by advertising, compared to fewer than 10 percent who said they would rather pay for ad-free content. The full results of the survey are available at: www.aboutads.info/resource/image/Poll/Zogby_DAA_Poll.pdf

In the second poll, also conducted by Zogby Analytics last November, more than half of those surveyed (51.3 percent) said they would be more likely to click on an online ad that included an icon – like the AdChoices icon – that allowed them to opt out of ad-related information collection. Additionally, more than 73 percent of users polled said they would feel more comfortable with interest-based ads if they knew they had access to the protections that the DAA currently provides, such as the ability to opt out, limitations on data collection and third-party enforcement. The full results of the survey are available at: www.aboutads.info/ZogbyDAAOct13PollResults.pdf

It is clear, then, that consumers desire and benefit from interest-based advertising. It is also apparent that there are significant concerns and very serious issues related to governmental surveillance and potential NSA reforms. Any policies developed to respond to these issues should be tailored to address vulnerabilities and potential harms in the governmental sphere and avoid jeopardizing the many commercial benefits of interest-based advertising.

Conclusion

Ultimately, we believe that consumer privacy concerns must be balanced with consumers' desire for more innovative products and services. Because advertising is a major economic engine in the United States and throughout the global economy, great care must be taken not to stifle the many benefits provided by effective advertising. Industry self-regulation, coupled with consumer education, is the best way to strike this balance and to enable innovative technologies to continue to bring new and exciting opportunities to consumers. These objectives apply in both the big and small data environments.

All those who collect and use data must continue to work to ensure the security and proper use of that information. Advertisers will continue to focus on issues of harm that can flow from big data by working cooperatively with other industry participants, governmental policymakers and consumer-related organizations on greater data security, encryption and anonymization. Where relevant, lessons learned in the small data context can be applied to big data. Where necessary, for example, related to enhanced data collection involved with the Internet of Things, protocols may need to be developed to ensure that data remains reliable and secure.

Data and interest-based advertising are fundamental to the efficient use of the Internet, mobile and other emerging technologies. The DAA's self-regulatory program is the best end-to-end program to maximize consumer choice and provide consumer benefits in regard to online behavioral advertising served through any technology.

Any new laws or regulations should only be adopted to fill in where it has been shown clearly that existing requirements or mechanisms are not adequate to protect consumers from harm and that the new law or regulation will ameliorate that harm. Policies should not assume that data, big and small, is the same, but policymakers also should not ignore similarities between big and small data collection and use. Further, policymakers must recognize the rapidly evolving nature of technology and innovation and refrain from constraints that will impede additional benefits and opportunities for users and commercial entities alike.

Thank you for your consideration of our views.

Sincerely,

Daniel L. Jaffe
Group Executive Vice President, Government Relations

March 31, 2014

Ms. Nicole Wong
Deputy U.S. Chief Technology Officer
Attn: Big Data Study, Office of Science and Technology Policy
Eisenhower Executive Office Building
1650 Pennsylvania Ave. NW
Washington, DC 20502

Re: Big Data RFI

Dear Ms. Wong,

Unleashing the power of big data is a key priority for both government and industry, and it is in the interests of the US economy to establish a policy framework that promotes innovation and protects the interests of all stakeholders associated with or affected by the use of the data.

BSA | The Software Alliance ("BSA") appreciates the opportunity to present its views in response to the Request for Information (RFI) on the Obama Administration's comprehensive review of the ways in which "big data" will affect how Americans live and work. Our comments focus on the benefits of big data as well as the implications of collecting, analyzing and using such data for privacy, the economy, and public policy.

BSA is the leading global advocate for the software industry. It is an association of nearly 100 world-class companies that invest billions of dollars annually to create innovative software solutions that make enterprises more productive, competitive, and secure.[1] Through international government relations and educational activities, BSA helps build trust and confidence in digital networks and in the new technologies that drive the global economy.

[1] *BSA | The Software Alliance (www.bsa.org) is the leading advocate for the global software industry before governments and in the international marketplace. Its members are among the world's most innovative companies, creating software solutions that spark the economy and improve modern life. With headquarters in Washington, DC, and operations in more than 60 countries around the world, BSA pioneers compliance programs that promote legal software use and advocates for public policies that foster technology innovation and drive growth in the digital economy.*

BSA's members include: Adobe, Apple, ANSYS, Autodesk, AVG, Bentley Systems, CA Technologies, CNC/Mastercam, Dell, IBM, Intel, Intuit, McAfee, Microsoft, Minitab, Oracle, PTC, Rockwell Automation, Rosetta Stone, Siemens PLM, Symantec, Tekla, and The MathWorks.

20 F Street, NW, Suite 800
Washington, DC 20001

P 202-872-5500
W bsa.org

The role of information is evolving. BSA's response to this RFI reflects our member companies' evolving experience in developing the products and tools that are used to improve the reliability, security and trustworthiness of big data collection, storage and analytics.

Governments, companies and consumers across the world rely on data-driven products and services to derive actionable insights from the ever-expanding pool of digital information. These insights and information help governments better serve their citizens at lower cost; they help businesses grow and expand in unprecedented ways; and, they help consumers better their lives – from improving their personal health and fitness to increasing the efficiency of their commercial interactions with new and established enterprises.

As a threshold matter, BSA would draw attention to the RFI's attempt to define "big data" as "datasets so large, diverse, and/or complex, that conventional technologies cannot adequately capture, store, or analyze them." While we recognize that such a definition allows the Administration to focus on issues and concerns that stand in the way of tremendous potential problems, such a definition provides too narrow a perspective for the dramatic societal and economic gains that an examination of the true scope of big data affords.

We believe that a key feature of big data is that through the application of complex analytics we can identify patterns that can be converted into reliable predictions. These predictive outputs can be used to advance science, improve health, and save energy as well as for a variety of commercial undertakings. Such benefits already are all around us, and to suggest that big data lies beyond our current technological capabilities would be to deny the economic and personal value already being derived from big data analysis.

BSA does appreciate the Obama Administration's focus in the RFI: the opportunity to advance the policy discussion and the big data solutions for the most difficult cases. By examining the impediments here, BSA hopes the Administration can continue to promote interests and advances across the board.

Big data already is having major positive impacts across all aspects of the economy. Developments such as predictive analytics have helped traditional manufacturing companies save millions of dollars in testing and other costs. At the same time, the wealth of new data streams has helped spawn entirely new business lines and revolutionized existing industry models. By way of potential impact, one estimate finds that increased big data analytics could increase annual GDP in retailing and manufacturing by up to $325 million even as it produces cost savings of as much as $285 billion in health care and government services.[2]

In the research realm, big data has the potential to unlock tremendous societal advances ranging from medicine to energy efficiency and the environment. In the health care field, for example, scientists' ability to analyze the immense amounts of data that are increasingly available will help explain disease processes and could identify new treatments and cures.

[2] Game Changers: Five Opportunities for US Growth and Renewal, McKinsey & Co. (July 2013), available at: http://www.mckinsey.com/insights/americas/us_game_changers.

Such data can help identify illnesses – even before symptoms develop. It will help with early cancer detection and lead to much more effective treatments.[3]

It is important to note that this process is playing out against the backdrop of an international conversation about data privacy and international surveillance. All efforts should be made to avoid conflating government and commercial interests in data. Ensuring trust in the full spectrum of the big data environment is key to its success, and it requires that we fully separate the topics in order to avoid missing the promise of "big data."

BSA Responses to the RFI's Questions to the Public

(1) What are the public policy implications of the collection, storage, analysis, and use of big data? For example, do the current U.S. policy framework and privacy proposals for protecting consumer privacy and government use of data adequately address issues raised by big data analytics?

BSA appreciates the attention that the Obama Administration has given to data-related policy in the United States, particularly the work on the 2010 Commerce Department green paper,[4] the 2012 Federal Trade Commission white paper,[5] and the Administration's 2012 Privacy Bill of Rights.[6] In a sign of how quickly the big data economy is developing, none of these recent reports fully assesses the policy implications of the collection, storage, analysis and use of big data. The current White House effort can help close those gaps.

The Administration should ensure that any policy proposals related to big data carefully account for the varying types of data that will fall under any new policy umbrella. The rising tide of information that big data innovations is based on comes in a range of types and from a multitude of sources, including public data, sensor data, transaction information and demographic information. The type and source of this data matters because different types of data implicate a range of different policy concerns.

In recent years, much of the debate about privacy has focused on the ways data is collected and associated notice and consent rules. We believe this is a necessary element but does not fully fit the realities. In most instances, concerns arise when data is used in ways that threaten or cause harm to the individual. Thus, we believe privacy policies should specifically take into account the risk of harm that the exposure or misuse of the relevant data represents. Accordingly, policies should be tailored to account for such risks in context-specific circumstances. The most sensitive data must

[3] The Impact of Big Data on Medical Research, Sanford-Burnham Science Blog, Sanford | Burnham Medical Research Institute (June 25, 2013), available at: http://beaker.sanfordburnham.org/2013/06/the-impact-of-big-data-on-medical-research/#sthash.bUhWRLKu.dpuf.
[4] Commercial Data Privacy and Innovation in the Internet Economy: A Dynamic Policy Framework (December 2010), available at: http://www.ntia.doc.gov/files/ntia/publications/iptf_privacy_greenpaper_12162010.pdf.
[5] Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Businesses and Policymakers (March 2012), available at: http://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-report-protecting-consumer-privacy-era-rapid-change-recommendations/120326privacyreport.pdf.
[6] Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy, (February 2012), available at: http://www.commerce.gov/sites/default/files/documents/2012/february/privacy-final.pdf

be accorded the highest protections. For example, individualized health information in the care of a patient's personal physician cannot be treated in the same manner as weather information that is the product of federally funded meteorological research.

Even with within the narrow categories of information, the sensitivity of data can vary – ranging in the health care space, for example, from information gathered in long-term public health assessments to that in an oncologist's diagnostic files. Taking a risk-based approach that accounts for such variations is particularly important in the evolving big data environment, where the traditional focus on notice and consent fails to account for the ever-increasing size and scope of data sets.

In addition, as the Administration considers updating policy frameworks it should also take into account the steps that can be taken to protect data – and the underlying individuals that data represents. Already BSA member companies build in privacy protections to their systems from the point of inception. This practice of "privacy by design" ensures that companies thoroughly incorporate privacy protections into their products and services.

Such efforts begin with the development and use of adequate privacy and security standards to establish a responsible benchmark for business practices. By putting the appropriate emphasis on security, for example, companies can reduce the risk that data is improperly accessed or disseminated. In addition, by using anonymization, de-identification, and encryption tools a company can further minimize the impact of any breach.

(2) What types of uses of big data could measurably improve outcomes or productivity with further government action, funding, or research? What types of uses of big data raise the most public policy concerns? Are there specific sectors or types of uses that should receive more government and/or public attention?

The Administration already has taken positive steps on big data by increasing access to the government's existing data sets and encouraging additional focus on areas of great societal impact. The recently announced initiative to share climate data with the goal of helping the public better understand risks on coastal flooding serves as one example. BSA would encourage continued similar efforts aimed to produce the greatest potential societal gains, including around such areas as health, public safety, and education.

Government should not focus its limited resources on areas where consumers can more directly shape the emerging big data environment. For example, systems built on user-based preferences for personalization and self-selected preferences in the commercial space should be allowed to mature and evolve to fit consumer needs. Finally, the lowest priority should be given to business analytics where no individuals are exposed to risk or harm.

(3) What technological trends or key technologies will affect the collection, storage, analysis and use of big data? Are there particularly promising technologies or new practices for safeguarding privacy while enabling effective uses of big data?

Big data technologies are enabling and spurring a range of products and services to benefit consumers and governments. Data-rich mapping services already have turned our phones into navigation devices that can be leveraged for restaurant reviews or

guidance for homebuyers. Crime statistics are being used by law enforcement agencies for "predictive policing" that helps reduce crime even before it happens. Facial recognition technologies are being developed that could help improve security and public safety and help tailor customer services.

Big data developments will help create new products and services in innumerable sectors. The benefits of many of these advances are obvious. And while such developments also pose potential privacy and security challenges, the same technology that delivers these advances can help provide solutions to address such concerns. By anonymizing or de-identifying the underlying data, technology can diminish or even extinguish many threats. New encryption technologies enable this, as do methods of data analysis that limit the exposure of the data to interception.

BSA supports the recent efforts of the National Institute of Standards and Technology (NIST) to convene stakeholders to discuss the potential for privacy engineering to contribute to the development of effective and repeatable privacy protections.[7] In doing so NIST and industry can work together to develop a basis for the development of technical standards and best practices for the protection of individual privacy and civil liberties.

(5) What issues are raised by the use of big data across jurisdictions, such as the adequacy of current international laws, regulations, or norms?

Like the Internet, data does not "respect" political borders. Nor should we restrict the innovations enabled by data analysis. At the same time, we must respect that privacy norms vary in different markets, and policymakers and industry should work together to bridge the differences.

As the United States develops policies with respect to big data, we strongly recommend that any such policies should include considerations of the impact of the policy on cross-border data flows and big data uses. The full rewards of big data will best be achieved by enabling the broadest possible economies of scale and ensuring that the benefits of big data accrue to the widest possible global audience.

Recognizing that there is a fine line between protecting privacy and creating unnecessary burdens to the flow of information, the 21 countries of the Asia-Pacific Economic Cooperation (APEC) established the Data Privacy Pathfinder in 2007.[8] The goal of the Pathfinder is to "achieve accountable cross-border flow of personal information" by developing and implementing a system of cross-border privacy rules (CBPRs).

Those rules are consistent with the APEC Privacy Framework, which was developed in 2004 and aims to: 1) improve information sharing among government agencies and regulators; 2) facilitate the safe transfer of information between economies; 3) establish a common set of privacy principles; 4) encourage the use of electronic data as a means to enhance and expand business; and 5) provide technical assistance to those economies that have yet to address privacy from a regulatory or policy perspective.

---

[7] NIST Privacy Engineering Workshop (March 6, 2014). More information available at: http://www.nist.gov/itl/csd/privacy-engineering-workshop.cfm

[8] Asia-Pacific Economic Cooperation (APEC), APEC Data Privacy Pathfinder, at http://www.apec.org/About-Us/About-APEC/Fact-Sheets/APEC-Privacy-Framework.aspx.

This year, a joint effort between the US, APEC and the European Union created a tool to build even more-expansive bridges. A coalition of officials from the Federal Trade Commission, APEC and the European Union (EU) created the "Referential" a tool that maps the APEC CBPR's to the EU's Binding Corporate Rules.[9] This new document is intended to serve as a practical reference guide for companies that seek certification under both the APEC and EU systems. Both of these tools can serve as models for the creation of policy frameworks for enabling continued big data innovation.

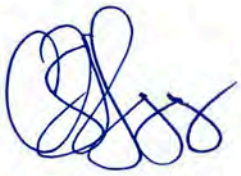
Looking more broadly, BSA believes that trade rules should be developed to enable and ensure the free flow of digital trade generally and the growth of big data services specifically. This entails covering innovative services in trade agreements, keeping borders open to the free flow of data and preventing mandates on where servers or other computing infrastructure must be located.

As the United States examines its policies on big data it should redouble its existing efforts on policy areas that impinge on cross-border data flows, and the government should look for new ways to improve international alignment on privacy and other data-related policy areas.

Conclusion

BSA appreciates this opportunity to comment on the Administration's Big Data RFI. We would be delighted to discuss these comments with the Office of Science and Technology Policy or to answer any questions about them.

Sincerely,



Christopher Hopfensperger
Director, Policy

[9] Asia-Pacific Economic Cooperation (APEC), "Joint work between experts from the Article 29 Working Party and from APEC Economies, on a referential for requirements for Binding Corporate Rules submitted to national Data Protection Authorities in the EU and Cross Border Privacy Rules submitted to APEC CBPR Accountability Agents," at http://www.apec.org/~/media/Files/Groups/ECSG/20140307_Referential-BCR-CBPR-reqs.pdf.

KEEPING THE INTERNET
OPEN • INNOVATIVE • FREE

www.cdt.org

CENTER FOR DEMOCRACY
& TECHNOLOGY

1634 Eye Street, NW
Suite 1100
Washington, DC 20006

March 31, 2014

Office of Science and Technology Policy
Eisenhower Executive Office Building
1650 Pennsylvania Ave. NW
Washington, DC, 20502

Re: Big Data Study

The Center for Democracy & Technology (CDT) is pleased to submit these comments in response to the Office of Science and Technology's Request For Information (RFI) on the implications of big data.

Our comments focus on the continued value of the Fair Information Practice Principles (FIPPs) as the best available framework for addressing the privacy implications of big data practices; the possibility of technical measures, such as de-identification, to safeguard privacy; and the need for immediate reform of current laws, including the Electronic Communications Privacy Act (ECPA). We respond to each of the questions posed in the RFI in turn below.

> *(1) What are the public policy implications of the collection, storage, analysis, and use of big data? For example, do the current U.S. policy framework and privacy proposals for protecting consumer privacy and government use of data adequately address issues raised by big data analytics?*

In our view, big data involves the collection of vast amounts of data from a growing number and variety of sources, combined with powerful analytic techniques that promise to extract useful insights applicable to a range of business and social problems. While the U.S. has a concept of privacy, expressed in the Fair Information Practice Principles (FIPPs) and the Administration's Consumer Privacy Bill of Rights, that can be used to address and mitigate the privacy implications of big data, that concept has not been comprehensively implemented in U.S. law. To the contrary, even before the emergence of big data, it was widely recognized that U.S. law fails to provide adequate privacy protection in the face of the digital revolution.[1] The advent of big data should add urgency to the goal of updating U.S. laws, both for

[1] Testimony of Ari Schwartz, Center for Democracy & Technology, Senate Committee on Commerce, Science, and Transportation, Subcommittee on Interstate Commerce, Trade, and Tourism, "Reauthorization of the Federal Trade Commission" (Sept. 12, 2007), *available at* https://www.cdt.org/files/pdfs/20070912schwartz-testimony.pdf.

businesses and government, to establish comprehensive baseline privacy legislation and stronger standards controlling governmental access.[2]

Data is being collected about individuals in a growing number of ways – when they browse the Internet and use online services,[3] through their electrical energy smart meters, through mobile applications installed on smartphones,[4] through systematic monitoring of their Internet usage by their ISPs,[5] and by tracking their movements in a variety of ways,[6] among other methods. The Internet of Things will vastly magnify the potential for data collection.[7]

The use of this data to compile profiles and to make decisions about individuals raises fundamental issues of fairness. In theory, big data analytics could be used to classify individuals based on race, ethnicity, gender, national origin, age, sexual orientation, or other suspect classes.[8] Big data analytics could also be used in many ways to widen existing power disparities between companies and consumers, by more accurately determining that the precise price that any individual may be willing to pay for a specific commodity or service.

Even before analytic techniques are applied to make decisions about people, the *collection* of data implicates privacy interests.[9] By collecting vast sets of data, companies and governments open themselves up to risk of data breach, unintended exposure, and internal misuse. As entities amass larger databases of information that may be linked to individuals, those databases become tempting

---

[2] Our comments focus on big data applications that involve data about individuals or that draw inferences about individuals. We recognize that there are many big data applications (such as assessing the environmental conditions in a field of corn or the functioning of a jet airplane engine) that do not involve personally identifiable or re-identifiable data.

[3] Testimony of Justin Brookman, Center for Democracy & Technology, Senate Committee on Commerce, Science, and Transportation, "A Status Update on the Development of Voluntary Do-Not-Track Standards" (Apr. 24, 2013), *available at* https://www.cdt.org/files/pdfs/Brookman-DNT-Testimony.pdf.

[4] G.S. Hans, *Lookout's Open Source Privacy Policy Could Change the Game on Mobile App Transparency* (Mar. 27, 2014), *available at* https://www.cdt.org/blogs/gs-hans/2703lookouts-open-source-privacy-policy-could-change-game-mobile-app-transparency.

[5] G.S. Hans, *Should Your ISP Monitor What You Do With Your Internet Service?* (Aug. 13, 2013), *available at* https://www.cdt.org/blogs/gs-hans/1308should-your-isp-monitor-what-you-do-your-internet-service.

[6] Comments of Center for Democracy & Technology to Federal Trade Commission, February 2014 Workshop on Mobile Device Tracking (Mar. 19, 2014), *available at* https://www.cdt.org/files/pdfs/cdt-mobile-device-tracking-comments.pdf.

[7] Comments of Center for Democracy & Technology to Federal Trade Commission, November 2013 Workshop on "Internet of Things" (Jan. 10, 2014), *available at* https://www.cdt.org/files/pdfs/iot-comments-cdt-2014.pdf

[8] *See, e.g.,* Press Release, The Leadership Conference, "Civil Rights Principles for the Era of Big Data", http://www.civilrights.org/press/2014/civil-rights-principles-big-data.html.

[9] Justin Brookman & G.S. Hans, *Why Collection Matters: Surveillance as a De Facto Privacy Harm*, FUTURE OF PRIVACY F., *available at* http://www.futureofprivacy.org/wp-content/uploads/Brookman-Why-Collection-Matters.pdf

targets for malicious third parties seeking to gain unauthorized access. Depending on what information is contained within those databases, and how that information is protected (through de-identification, encryption, or other methods), the effects of a data breach could be catastrophic. As recent high-profile data breaches have demonstrated, sensitive personal and financial data can, in the event of a breach, become accessible to unauthorized third parties and can result in real-world consumer harm, such as identity theft.[10] Therefore, businesses and governments that collect data about individuals should limit their collection practices and only collect data necessary for specific uses.

The current American legal regime does not adequately protect consumer privacy. At present, there are a patchwork of laws, including the FTC Act, the Children's Online Privacy Protection Act, and the Video Privacy Protection Act, that provide varying degrees of privacy protection, but no comprehensive privacy legislation. CDT has long called for Congress to pass baseline privacy legislation,[11] and we have supported the proposals put forward by Congress,[12] the White House,[13] and the Federal Trade Commission.[14] The advent of big data should not be a distraction from this unfinished business; to the contrary, the increased surveillance, analytical and data retention technologies that make Big Data possible should spur the adoption of comprehensive baseline federal privacy legislation.

The FIPPs as a Framework to Protect Privacy in the Era of Big Data

CDT believes that the Fair Information Practice Principles (FIPPs) provide a robust framework to promote the protection of individual privacy interests in the era of big data. For the past thirty years, the dominant concept of information privacy has been expressed in the FIPPs. The Obama Administration adopted

---

[10] G.S. Hans, *Target and Neiman Marcus Testify on Data Breach – But What Reforms Will Result?* (Feb. 7, 2014), *available at* https://www.cdt.org/blogs/gs-hans/0702target-and-neiman-marcus-testify-data-breach-–-what-reforms-will-result.

[11] Testimony of Justin Brookman, Center for Democracy & Technology, House of Representatives Committee on Energy and Commerce, Subcommittee on Commerce, Manufacturing, and Trade, "Balancing Privacy and Innovation: Does the President's Proposal Tip the Scales?" (Mar. 29, 2012), *available at* https://www.cdt.org/files/pdfs/Justin-Brookman-privacy-testimony.pdf

[12] Press Statement, Center for Democracy & Technology, CDT Statement on Release of Draft Consumer Privacy Bill: the Best Practices Act (Jul. 19, 2010), *available at* https://www.cdt.org/pr_statement/cdt-statement-release-draft-consumer-privacy-bill-best-practices-act.

[13] WHITE HOUSE, CONSUMER DATA PRIVACY IN A NETWORKED WORLD: A FRAMEWORK FOR PROTECTING PRIVACY AND PROMOTING INNOVATION IN THE GLOBAL DIGITAL ECONOMY (2012), available at http://www.whitehouse.gov/sites/default/files/privacy-final.pdf.

[14] Statement of Edith Ramirez, Federal Trade Commission, Senate Committee on the Judiciary, "Privacy in the Digital Age: Preventing Data Breaches and Combating Cybercrime" (Feb. 4, 2014), *available at* http://www.ftc.gov/system/files/documents/public_statements/prepared-statement-federal-trade-commission-privacy-digital-age-preventing-data-breaches-combating/140204datasecuritycybercrime.pdf.

the FIPPs as the basis for its Consumer Privacy Bill of Rights in February 2012.[15] Many have argued that big data fundamentally challenges the FIPPs framework. In CDT's view, it is *not* inevitable that big data will overwhelm traditional concepts of privacy. Many of the issues now being cited in connection with big data are actually longstanding concerns (for example, the limitations of notice and consent). Many of the solutions being put forth by academics and others draw upon or echo elements of the traditional FIPPs framework.

For example, many of Paul Schwartz's recommendations sound very similar to core FIPPs.[16] Schwartz recommends, for example, that a company using big data analytics "should develop reasonable mitigation processes and reasonable remedies as appropriate when analytics lead to decisions that harm individuals," which sounds like the redress element of the individual participation FIPP and the accountability FIPP. Likewise, echoing the data quality FIPP, he recommends that a company "should engage in decision-making based on analytic output that is reasonably accurate." At another point, Schwartz recommends that, based on ongoing review and revision of their analytics practices, companies "should only use information that is predictive," which restates the data quality principle to emphasize the reliability of outcomes.

While initially dismissing the traditional privacy framework, Christopher Kuner and co-authors also end up endorsing the FIPPs that focus on outcomes.[17] Given big data's role in decision-making about individuals, they state, "issues such as the accessibility, accuracy and reliability of data may matter as much or maybe more than privacy" (by "privacy," the authors seem to mean collection and use limitations). Of course, accessibility, accuracy and reliability have always been key FIPPs. Also, it is noteworthy that Kuner et al., after severely criticizing the consent model, conclude that there remains a proper role for individual consent, further illustrating how compelling the FIPPs framework is.

---

[15] *See* WHITE HOUSE, *supra* note 13, at 1. The FIPPs were first articulated in both the U.S. and in Europe in the early 1970's and quite rapidly became the focus of privacy policy development on both sides of the Atlantic and, after their adoption by the OECD, globally. *See* Robert Gellman, *Fair Information Practices: A Basic History*, http://bobgellman.com/rg-docs/rg-FIPShistory.pdf. The Department of Homeland Security adopted a version of the FIPPs as the foundation for privacy policy and implementation at DHS in 2008. Hugo Teufel III, Department of Homeland Security, *Privacy Policy Guidance Memorandum* (Dec. 29, 2008) *available at* http://www.dhs.gov/xlibrary/assets/privacy/privacy_policyguide_2008-01.pdf. Here, we use both the language of the Administration's FIPPs, which focused on the consumer context, and the DHS FIPPs, which focused on government practices. The congruence between the Administration's formulation and DHS's shows the consistency in the understanding of information privacy in the U.S.

[16] See Paul Schwartz, *Data Protection Law and the Ethical Use of Analytics*, Privacy & Security Law (Jan. 10, 2011), and Paul Schwartz, *Property, Privacy, and Personal Data*, 117 HARV. L. REV. 2055, 2096 (2004).

[17] Christopher Kuner, Fred H. Cate, Christopher Millard, & Dan Jerker B. Svantesson, *The Challenge of "Big Data" for Data Protection*, 2 INT'L DATA PRIVACY L. 49 (2012).

Omer Tene and Jules Polonetsky note that the FIPPs have a certain adaptability that allows adjustments in emphasis among the various principles.[18] They base their solution on "re-craft[ing] transparency obligations and access rights to make them more useful in practice." They note that "[t]raditional transparency and individual access mechanisms have proven ineffective." However, rather than proposing to replace transparency and access, they call for more effective implementation of these FIPPs, which they argue will both better protect individuals and unleash the power of big data:

> "If organizations provide individuals with access to their data in usable format, creative powers will be unleashed to provide users with applications and features building on their data for new innovative uses. In addition, transparency with respect to the logic underlying organizations' data processing will deter unethical, sensitive data use and allay concerns about inaccurate inferences."

It appears that the FIPPs framework is more durable than many have recently assumed. Even in calling for new approaches to privacy in response to big data, key experts have reverted to concepts that are part of the FIPPs. There is a certain power in this correlation between traditional data protection concepts and new ideas about privacy and big data. Among other things, the correlation offers a response to the paralysis of privacy policy that big data seemed to portend. It is not necessary to have a full rethink of privacy.

Indeed, in our view, rather than tolling the death knell of privacy, the big data phenomenon could be leveraged to spur development of the workable and effective privacy framework that has long been lacking. We believe that any comprehensive federal privacy legislation should use the FIPPs as an organizing framework, and that, in the interim, companies should use the FIPPs as a self-regulatory tool to protect their customer's privacy interests. Below, we discuss the FIPPs in turn and highlight their relevance to big data.

Purpose Specification and Use Limitation (Respect for Context)

It is often said that big data techniques involve the use of data in unanticipated ways. Nevertheless, purpose specification and use limitation are two closely related principles that remain vital to protecting individual privacy. With respect to consumers, the Administration's Privacy Bill of Rights well describes the two principles when it says, "Consumers have a right to expect that companies will collect, use, and disclose personal data in ways that are consistent with the context in which consumers provide the data." Even in the era of big data, purpose specification should be a crucial first step in any system design, requiring entities to detail on what grounds they will collect data and the uses that they plan for it. The use limitation principle requires entities to follow through on the delineated uses and refrain from using the collected data for undisclosed purposes.

[18] Omer Tene and Jules Polonetsky, *Big Data for All: Privacy and User Control in the Age of Analytics*, 11 NW. J. TECH. & INTELL. PROP. 239 (2013).

Limitations on the collection of data are vitally important in a world in which it is becoming less and less expensive to collect increasing amounts of data from a variety of devices. Individual privacy interests are implicated at the point of collection, because of the variety of risks that databases are subject to. When any entity collects data about individuals, that data can be subject to internal misuse, changes in company practices, or data breaches.[19] Some have argued that relying on use limitations would be sufficient to protect privacy, but the threats to privacy arise long before an entity actually uses the data. Use limitations, while important, cannot protect against all possible threat models. As a result, purpose specification, which provides both a basis for and limits on the collection of information, is a vital element to protecting individual privacy interests. It is directly linked to other principles, including minimization (focused collection) and transparency. Companies engaged in big data analytics should be sure to detail the purposes for which they collect information in order to demonstrate their commitment to protecting consumers and their privacy interests. Use limitations are also important. Companies must confine their uses of data to the purposes disclosed to consumers. If the company plans to share data collected with a third party, that sharing should be disclosed to consumers in advance, as should the third party's uses (e.g. analytics).

Especially in the big data context, entities collecting personal information could very well develop new uses of data in future years that are loosely (if at all) related to the uses that the data was originally collected for. If that happens, entities must at the very least provide transparency about those new uses before they begin. Entities holding data should consider whether the new uses can be performed with de-identified data. They should carefully weigh the potential adverse consequences that may befall individuals from the use of such data and design their programs to avoid such consequences or ensure that they are reliable and justified. However, if data custodians conclude that the new uses can only be performed with identifiable data, and are not contextually related to the purposes for which the data was originally collected, they must seek new consent for those new uses. User expectations – and the potential for user *surprise* – are important indicia for whether a new purpose is contextually related to an older one.

Transparency

It is almost certain that the public has very low awareness of the uses currently made of data – much less potential future uses. Therefore, the onus is on the *companies, governmental entities, and others collecting and using data about individuals* to disclose what uses they are making and plan to make, and, when they come up with new uses, to disclose them. Both the private sector and the government will have to educate the public on what big data actually means and why entities are employing it. By being transparent about their collection, use, and retention practices of data, companies, government agencies and other entities will both create better public awareness of their practices and increase public trust.

[19] *See* Brookman & Hans, *supra* note 9.

The limitations of notice have long been recognized. Whether it is corporate privacy policies or Privacy Act System of Records Notices, individuals are unable to sift through the massive volume of verbiage to determine which is relevant. But at the very least, companies and government agencies must make information about all their practices available to the public in some form – whether in a privacy policy, in terms of service, in the statutes and guidelines defining governmental collection authorities, or in other forms of detailed disclosure. The ability for the public to access information on corporate and government practices is vitally important, both for educational purposes and to hold companies and government officials accountable when their public statements fail to correspond with their actual practices. The FTC should continue to undertake investigations and bring enforcement actions against companies that have not sufficiently described their data privacy practices.[20]

Individual Participation (Individual Control)

Related to the transparency principle, the individual participation principle urges companies to give individuals control over what personal data is collected from them and how it is used. The most obvious way that companies can do this is by allowing users to make decisions regarding what data gets collected, and what uses a company can make with that data. Especially where consumers purchased the devices that enable big data analytics, they should be in control over what data those devices collect and transmit to companies. Therefore, companies should solicit the participation of consumers when seeking to access the data that devices can provide.[21] Some data collection and retention can reasonably be done only on an opt-out basis – that is, unless the consumer affirmatively objects. Some – such as data collected and used only for reasonable and focused security purposes – should not be subject to individual control at all. However, certain sensitive categories of data should only be collected and retained with a consumer's informed permission. Information about medical conditions – or information about what users do inside their own homes – are examples of intensely personal information that should only be done on an opt-in basis.

The development of effective notice and consent regimes will play a vital role in enabling responsible big data regimes, as pervasive collection may allow businesses to create highly granular and comprehensive records for individual customers. Because more and more companies have the capacity to monitor users, these controls will in many cases need to be universal. For example, the Do Not Track mechanism has been proposed as an easy and effective way for consumers to express their choice to stop cross-site tracking in the online

[20] G.S. Hans, *Goldenshores Case Demonstrates Flaws in Current Mobile Privacy Practices* (Dec. 23, 2013), *available at* https://www.cdt.org/blogs/gs-hans/2312goldenshores-case-demonstrates-flaws-current-mobile-privacy-practices.

[21] A recent case involving LG TVs that broadcast viewer usage practices to the manufacturer highlights the need for empowering users to make the final say over how their devices behave. *See* Justin Brookman, *Eroding Trust: How New Smart TV Lacks Privacy by Design and Transparency* (Dec. 3, 2013), *available at* https://www.cdt.org/commentary/eroding-trust-how-new-smart-tv-lacks-privacy-design-and-transparency.

context. The online advertising industry should be encouraged to honor users' Do Not Track requests, and other industries should explore similar universal choice mechanisms to allow consumers to more effectively regulate the dissemination of their personal information.[22]

Effective consumer notification will be necessary. Customers may not even be aware what a business can collect from a computer, a mobile device, or wearable devices. Without adequate notice and consent provisions, customers who don't approve of what a particular business does won't be able to "vote with their feet" and choose another business with different practices. Companies should begin developing effective consent models now, rather than deploying them after they finalize their big data collection practices.

Security

The recent spate of high-profile data breaches emphasizes the need for strong security programs for all entities that collect data about individuals.[23] As the big data trend results in the increasing collection of data, businesses and governmental entities must create strong security programs – and monitor and update those programs – in order to protect data. Companies should be held accountable for failing to safeguard the data they maintain and should notify consumers of breaches as they occur in full compliance with current law. Although the FTC's ability to seek enforcement actions against companies for poor data security practices is currently being litigated, CDT thinks that the FTC currently has authority under Section 5 of the FTC Act to regulate data security, and we encourage the FTC to continue to bring enforcement actions against companies that have substandard data security programs.[24]

Data Minimization (Focused Collection)

Data minimization is closely related to data security. Collecting data without a clear (and disclosed) purpose in mind, or the failure to purge old data in accordance with reasonable minimization procedures, should be factors in evaluating whether an entity's data security practices were reasonable. As part of their security programs, companies, government agencies and other entities should implement specific retention periods for data, rather than retaining that information indefinitely. If entities implement minimization procedures and delete unnecessary, outdated, or irrelevant entries, fewer records will be accessible to unauthorized parties if and when a data breach occurs. By removing identifying

[22] For example, FTC Commissioner Julie Brill has launched an initiative, "Reclaim Your Name", that would educate users and empower them to assert control over their personal data. *See* Julie Brill, Op-Ed, *Demanding Transparency from Data Brokers*, WASH. POST (Aug. 15, 2013), *available at* http://www.washingtonpost.com/opinions/demanding-transparency-from-data-brokers/2013/08/15/00609680-0382-11e3-9259-e2aafe5a5f84_story.html.

[23] See Hans, *supra* note 10.

[24] G.S. Hans, *Data Security and Your Next Hotel Stay: How the FTC Encourages Strong Security Practices* (May 21, 2013), *available at* https://www.cdt.org/blogs/gs-hans/2105data-security-and-your-next-hotel-stay-how-ftc-encourages-strong-security-practice.

information and deleting data after it is no longer needed, companies will both protect their customers' security and promote consumer trust.

If a company retains data or shares it with a third party, it should consider anonymizing or pseudonymizing the data it provides in order to protect individual privacy. In its 2012 report on consumer privacy, the FTC set out the following standard to ensure that data is properly anonymized so that it cannot be "reasonably linked" to a particular consumer, computer, or device: "data is not 'reasonably linkable' to the extent that a company: (1) takes reasonable measures to ensure that the data is de-identified; (2) publicly commits not to try to re-identify the data; and (3) contractually prohibits downstream recipients from trying to re-identify the data." [25] CDT believes that this is an appropriate and viable standard for companies to implement to de-identify consumer data. By removing identifying information before sharing data, companies can take an affirmative step to protecting consumers even after the data is out of their direct control by reducing the likelihood that someone else can use the data for undisclosed purposes

Data Quality (Access and Accuracy)

Entities collecting and using data about individuals should also ensure that the data they use and retain is accurate, relevant, and complete. Because of the sensitive nature of data collected for big data purposes, it is vitally important for entities to ensure that their records are accurate. If a promotional offer was delivered to the wrong consumer or if records were not kept suitably secure, customers could become disturbed, inconvenienced, or vulnerable to inappropriate uses.[26]

Accountability and Auditing

In order to ensure that data collection and use practices are followed and security programs are properly implemented, entities must create internal oversight mechanisms and must be subject to external accountability. This will ensure that the practices that are nominally adopted are effectively followed, and will encourage public trust.

(2) What types of uses of big data could measurably improve outcomes or productivity with further government action, funding, or research? What types of uses of big data raise the most public policy concerns? Are there specific sectors or types of uses that should receive more government and/or public attention?

[25] FEDERAL TRADE COMM'N, *Protecting Consumer Privacy in an Era of Rapid Change* (February 2012), *available at* http://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-report-protecting-consumer-privacy-era-rapid-change-recommendations/120326privacyreport.pdf.

[26] Charles Duhigg, *How Companies Learn Your Secrets*, N.Y. TIMES MAGAZINE at MM 30 (Feb. 19, 2012), *available at* http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html.

As discussed above, uses of big data that classify individuals based on suspect classes and treat those individuals differently from the general public would raise the most public policy concerns. Because big data analytics are conducted without public knowledge or disclosure, it is difficult to identify when such classifications are being made. Therefore, the openness principle described above will be particularly important to determine when businesses are making such classifications so that consumers can make a more informed choice about what data they provide to businesses.

Government uses of big data also raise public policy concerns. While the Privacy Act regulates the government's ability to create and use databases, its provisions include multiple exceptions for agencies that engage in law enforcement and foreign intelligence. As a result, individuals whose records are collected and analyzed by the government may not be aware that such analysis is taking place. Increased transparency concerning government use of personal data in big data processes will help limit this type of use.

The use of algorithms to make determinations regarding how individuals should be classified, targeted, or marketed to may raise specific policy concerns. If companies make assumptions about what an individual wants and targets those individuals, those assumptions may be reinforced rather than challenged as inaccurate or outdated. For example, if a store targets individuals with advertisements or coupons for foods with high sugar, fat, or salt content, the store may undermine the customer efforts to adhere to a diet plan. Ryan Calo has described this issue as "digital market manipulation," arguing that such practices deserve attention from regulators.[27]

> *(3) What technological trends or key technologies will affect the collection, storage, analysis and use of big data? Are there particularly promising technologies or new practices for safeguarding privacy while enabling effective uses of big data?*

The trend towards increased collection of data from mobile devices, networked appliances, cars, wearable devices, and online services allows for more extensive big data analytics. Advances in computing power and storage capacity allow companies and government to retain more data for longer periods at reduced costs, and analyze massive datasets to an unprecedented degree.

The possibility of collecting all data on a persistent basis has given rise to dystopic fears of a world in which someone is always watching you, even in private spaces. The United States Constitution specifically identifies realms in which individuals have heightened privacy interests, and big data collection capabilities imperil those spaces. There have been many stories, from FTC enforcement actions to inappropriately targeted advertisements, that highlight how individual privacy has been compromised by systems that were designed to

[27] Ryan Calo, *Digital Market Manipulation*, forthcoming GEO. WASH. L. REV. (2014), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2309703.

constantly collect data without foresight. From laptop cameras spying on individuals during intimate moments,[28] to advertisements identifying a teenager as pregnant before she told her parents,[29] consumers have been threatened by invasive practices that could have been avoided had companies used more forethought in designing systems and releasing products.

Safeguarding privacy, therefore, is of increased importance given that there are more ways than ever before for privacy interests to be implicated by big data practices. Allowing individuals to have as much control as possible over their devices will be paramount in protecting those interests. When a consumer purchases a device that has the capacity to collect data, that consumer should have the ability to the extent possible to control that collection. Some practices that allow for big data analytics – such as mobile device tracking – rely upon data collection that individuals may not be aware is even happening, and may have few ways to prevent. For example, mobile device tracking typically uses a device's broadcast of a Media Access Control (MAC) address to track that device over time and make determinations based on the behavior of the device (and by implication, the device's owner). However, due to the system architecture of most devices, many individuals are not aware that their devices are broadcasting MAC addresses, and are not easily able to prevent it from happening without disabling Wifi and Bluetooth functionality.[30] When consumers purchase devices, they should be the ultimate arbiter of when data collection occurs, how it occurs, and with what frequency. Empowering users to have more control over their devices and the types of data that each device collects will allow individuals to proactively protect their privacy.

Companies should also limit the amount of data they collect. There has been a worrisome trend to focus on use limitations, rather than limitations on collection. However, for the reasons discussed above, there are multiple ways in which the collection of data implicates individual privacy interests. Companies should therefore make deliberate decisions on what types of data are collected and under which circumstances, rather than enabling all possible types of collection and relying upon use limitations to manage their databases. Use limitations are important, but must be coupled with collection limitations in order to create a FIPPs-compliant data management regime.

Some authors have proposed specific technical solutions, such as an increasing reliance on differential privacy as proposed by Cynthia Dwork.[31] Companies have also architected some big data analytic systems to be privacy protective; for example, Facebook created a double-blind system when partnering with

---

[28] G.S. Hans, *Laptop Spying Case Indicates More Aggressive FTC Stance on Privacy* (Oct. 9, 2012), *available at* https://www.cdt.org/blogs/gs-hans/0910laptop-spying-case-indicates-more-aggressive-ftc-stance-privacy.

[29] *See* Duhigg, *supra* note 26.

[30] *See* Comments of Center for Democracy & Technology, *supra* note 6.

[31] Cynthia Dwork, *Differential Privacy: A Survey of Results* (2010), *available at* http://www.cs.ucdavis.edu/~franklin/ecs289/2010/dwork_2008.pdf.

Datalogix in order to ensure that neither company could create a detailed profile based on online and offline data.[32] We at CDT hope that other companies will take steps to develop such technical solutions.

(4) How should the policy frameworks or regulations for handling big data differ between the government and the private sector? Please be specific as to the type of entity and type of use (e.g., law enforcement, government services, commercial, academic research, etc.).

One major difference between policy frameworks for the government and the private sector is that the government is subject to the Constitution. Unfortunately, the case law and statutes have not kept pace with technological developments, and personal data held by third parties is inadequately protected against government access. A major challenge that needs to be addressed, and one where we urge the Administration to take a stronger position, is to ensure that the principles of the Fourth Amendment are extended to cover government access to digital data held by third parties.

Immediate reform is needed with respect to the content that individuals are storing with third party companies more than ever before. This includes emails, photos, address books and documents stored in the cloud.

Under an outdated law, this digital content is not adequately protected from government access. The Electronic Communications Privacy Act (ECPA) says that government agencies do not need a warrant— authorized by a judge and based on probable cause—to demand that third party service providers turn over the contents of their customers' emails and documents. A federal appeals court, in a decision we endorse, has held that ECPA is unconstitutional in this regard. Bi-partisan legislation is pending in both Houses of Congress to address this problem. The Administration should support enactment of S. 607 and H.R. 1852, with no carve-outs or exceptions for civil agencies.

The problem of government access also extends to metadata about communications. Cell site location data is one particularly revealing type of data and is automatically generated by mobile phones used by 91% of the U.S. population.[33] The government argues that it does not need a warrant to access consumers' mobile location data held by communications service providers. Of all the kinds of transactional data, location tracking information is one that clearly should be subject to the warrant protection. Again, bipartisan legislation is pending in both Houses of Congress to require government agencies to obtain a warrant before compelling service providers to disclose location tracking information, and the Administration should support that legislation.

---

[32] Jennifer Martinez, *Facebook's New Ad Partnership Stokes Privacy Concerns*, THE HILL (Sept. 26, 2012), *available at* http://thehill.com/blogs/hillicon-valley/technology/251287-facebooks-new-ad-tracking-partnership-stokes-privacy-concerns-.

[33] Lee Rainie, *Cell Phone Ownership Hits 91% of Adults*, FactTank: Pew Research Center (June 6, 2013), *available at* http://www.pewresearch.org/fact-tank/2013/06/06/cell-phone-ownership-hits-91-of-adults/.

At the root of concerns about government access in the era of big data is the so-called third party doctrine. If the Administration wants to do anything about big data and privacy it at least needs to acknowledge that the scope of the third party doctrine needs to be curtailed. Adopted long before digital technology had become essential to daily life and long before the outlines of the big data phenomenon were apparent, the third party doctrine says that individuals lose all constitutional privacy interest in data voluntarily disclosed to a third party. This doctrine is the basis of arguments that there is no constitutional privacy interest in documents stored in the cloud, in cell phone tracking information, or in records collected by the private sector about our daily activities, ranging from health to finances to travel to entertainment choices. It is the basis of the NSA telephony metadata program, the revelation of which helped prompt this review.

The third party doctrine is especially ill-suited to the era of big data, for it says that all of the big data collected by commercial entities about individuals is unprotected by the Constitution. Until the third party doctrine is addressed, government access issues will be left to a patchwork of statutes, many of which currently allow government access to highly sensitive data under a very weak standard.

(5) What issues are raised by the use of big data across jurisdictions, such as the adequacy of current international laws, regulations, or norms?

Because of disparate international laws and regulations, multinational companies – whether they use big data analytics or not – need to comply with often contradictory regulations. The lack of a comprehensive U.S. privacy framework has made this particularly difficult. The European Union is debating a Data Protection Regulation (DPR) that may impose even further limits on the abilities of companies to conduct big data analytics.[34]

The lack of a baseline consumer privacy law in the U.S. makes it harder for U.S. companies and officials to argue credibly against overbroad or unworkable privacy regulations. The EU DPR was proposed in part to force American companies to institute stronger privacy protections, specifically in response to the fact that current American law does not place baseline requirements upon companies.[35] In our view, continuing U.S. inaction on consumer privacy contributes to proposals in Europe that would be unduly restrictive of the open Internet. Proposed European data localization requirements, for example, would not only negatively affect American businesses by increasing operating costs for companies or leading European users to migrate from American services to

[34] Justin Brookman, *Progress Made Revising EU Data Privacy Laws* (Nov. 12, 2013), *available at* https://www.cdt.org/blogs/justin-brookman/1211progress-made-revising-eu-data-privacy-laws.

[35] Kevin J. O'Brien, *Firms Brace for New European Data Privacy Law*, N.Y. TIMES (May 13, 2013), *available at* http://www.nytimes.com/2013/05/14/technology/firms-brace-for-new-european-data-privacy-law.html.

European analogues.[36] Data localization would also contribute to fragmentation of the open Internet. However, so long as the U.S. lacks a privacy law, countries may be attracted by such extreme measures, and countries outside Europe will likely defer to the European framework for consumer privacy protection when developing their own regulations. If Congress did pass baseline privacy legislation, it would signal to the world that data can be protected without the need for localization.

Even in the absence of baseline federal privacy legislation, businesses should adhere to practices that allow for user control, not only to promote individual privacy but also to increase the likelihood of complying with disparate international law. Technological changes and advancements should not fundamentally change human rights protections or the need for private locations in which individuals can exist unobserved. The rise in networked devices that can silently collect data in the home and mobile devices that can track the owner's location may imperil those private spaces, and companies should consider privacy – which has been considered an international human right – when designing their products and systems.

Conclusion

We thank OSTP for soliciting comments and for its workshop series on big data and its technical, social, and regulatory implications. Faced with the privacy and security risks inherent in big data practices, we believe the FIPPs are as relevant as ever and provide an exemplary framework for promoting individual privacy protections by both business and government. A FIPPs-based framework could address the key challenges of big data:

- provide protections for privacy while still enabling analytics to solve pressing business and social challenges;
- apply consistently across sectors yet still be flexible enough to respond to the particular risks to privacy posed by different applications;
- include mechanisms to hold accountable entities collecting and analyzing data; and
- provide incentives for the adoption of privacy-enhancing technical architectures/models for collecting and sharing data.

[36] Alison Smale, *Merkel Backs Plan to Keep European Data in Europe*, N.Y. TIMES (Feb. 16, 2014), at A6, *available at* http://www.nytimes.com/2014/02/17/world/europe/merkel-backs-plan-to-keep-european-data-in-europe.html.

Sincerely,

/s/
Nuala O'Connor
President & CEO

/s/
Jim Dempsey
Vice President for Public Policy

/s/
Justin Brookman
Director, Consumer Privacy Project

/s/
Joseph Lorenzo Hall
Chief Technologist

/s/
G.S. Hans
Ron Plesser Fellow

/s/
Runa A. Sandvik
Staff Technologist

March 31, 2014

Attn: Big Data Study
Office of Science and Technology Policy
Eisenhower Executive Office Building
1650 Pennsylvania Avenue NW
Washington, DC 20502

Dear Ms. Wong,

On behalf of the Center for Data Innovation (www.datainnovation.org), I am pleased to submit these comments in response to the Office of Science and Technology Policy's (OSTP) request for public comment on the public policy implications of "big data."[1]

The Center for Data Innovation is a non-profit, non-partisan, Washington, D.C.-based think tank focusing on the impact of the increased use of information on the economy and society. The Center formulates and promotes pragmatic public policies designed to enable data-driven innovation in the public and private sectors, create new economic opportunities, and improve quality of life. The Center is affiliated with the Information Technology and Innovation Foundation (ITIF).

In this request for public comment, OSTP seeks responses to the following five questions:

1. What are the public policy implications of the collection, storage, analysis, and use of big data? For example, do the current U.S. policy framework and privacy proposals for protecting consumer privacy and government use of data adequately address issues raised by big data analytics?

2. What types of uses of big data could measurably improve outcomes or productivity with further government action, funding, or research? What types of uses of big data raise the most public policy concerns? Are there specific sectors or types of uses that should receive more government and/or public attention?

3. What technological trends or key technologies will affect the collection, storage, analysis and use of big data? Are there particularly promising technologies or new practices for safeguarding privacy while enabling effective uses of big data?

[1] Federal Register. "Government 'Big Data'; Request for Information." March 4, 2014. https://www.federalregister.gov/articles/2014/03/04/2014-04660/government-big-data-request-for-information. Accessed March 25, 2014.

4. How should the policy frameworks or regulations for handling big data differ between the government and the private sector? Please be specific as to the type of entity and type of use (e.g., law enforcement, government services, commercial, academic research, etc.).

5. What issues are raised by the use of big data across jurisdictions, such as the adequacy of current international laws, regulations, or norms?

Each question is addressed in turn below.

QUESTION 1

What are the public policy implications of the collection, storage, analysis, and use of big data? For example, do the current U.S. policy framework and privacy proposals for protecting consumer privacy and government use of data adequately address issues raised by big data analytics?

RESPONSE

By helping individuals and organizations make better decisions, data has the potential to spur economic growth and improve quality of life in a broad array of fields. The private sector is currently using vast quantities of data for a variety of purposes including optimizing energy efficiency in buildings, reducing mechanical failures in equipment, and improving crop yields on farms.[2] The public sector also has many opportunities to use data to address major social issues such as improving health care, fighting crime, building more sustainable communities, and creating more efficient transportation systems.[3]

While the potential benefits of data are clear, achieving those benefits is far from certain. The opportunities from data will not be realized unless policymakers create the necessary conditions for data-driven innovation to flourish. Unfortunately, most of the policy debate in Washington has been on how to minimize potential harms from data, especially around privacy, rather than on how to enable more and better uses of data. This needs to change if for no other reason than a significant proportion of the benefits from data will come from scientific or business applications that do not involve the use of personal information. To that end, the U.S.

[2] Castro, Daniel and Travis Korte. The Center for Data Innovation. "Data Innovation 101." November 3, 2013. http://www.datainnovation.org/2013/11/data-innovation-101/. Accessed March 28, 2014.

[3] Ibid.

government should create a comprehensive strategy on how to maximize the benefits of data that address issues associated with human capital, technology, and data itself.

First, the federal government can help maximize the benefits from data by ensuring that the workforce can support the growing demand for data scientists and related professionals. Only one-third of 4.4 million available data-related jobs will be filled in 2015, according to 2012 projections from research firm Gartner.[4] The government can accelerate the growth of a data-literate workforce by offering computer science and statistics education in secondary schools, developing open online courses on data-related subjects, and engaging with data science communities around the country with competitions and civic hacking events.

Second, the federal government should promote continued innovation of data-related technologies by supporting data-related research and development (R&D) initiatives. Funding agencies should continue to invest in tools and methods for large-scale data storage, analysis, and visualization. In particular, there is a major need to develop software that links data analysis with distributed systems architectures, as well as producing data analysis software targeted to end users, who may not have extensive training in computer science and statistics.[5]

Other funding can be used to improve technologies that support privacy, such as by funding research on privacy-preserving data mining and new techniques to de-identify data. For example, various techniques can be used to add noise to sensitive datasets so that individual information cannot be extracted.[6] One example of this approach can be found in synthetic data, which strives to preserve the usefulness of sensitive datasets by emulating their underlying statistical characteristics while simultaneously masking individual information.[7] Many agencies are working on similar data privacy problems, but their research efforts are not coordinated. To ensure that federal research dollars are directed to the most pressing privacy challenges and

[4] Gartner. "Gartner Says Big Data Creates Big Jobs: 4.4 Million IT Jobs Globally to Support Big Data By 2015." October 22, 2012. http://www.gartner.com/newsroom/id/2207915. Accessed March 28, 2014.

[5] National Research Council. *Frontiers in Massive Data Analysis*. Washington, DC: The National Academies Press, 2013.

[6] National Institute for Standards and Technology. National Strategy for Trusted Identities in Cyberspace. http://www.nist.gov/nstic/. Accessed March 28, 2014.

[7] Centers for Medicare and Medicaid Services. "Medicare Claims Synthetic Public Use File (SynPUFs)." July 10, 2013. http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/SynPUFs/. Accessed March 28, 2014.

that agencies collaborate on funding initiatives, the federal government should develop a strategic roadmap for federally-funded privacy R&D.[8]

Third, the data economy depends on the ability to share and reuse data, and policymakers should support policies that encourage the free flow of data between different stakeholders. Server localization requirements and consumer privacy laws should not be used to "lock up" data. For example, a patient's medical data is sensitive and should be protected from illicit uses. However, uniting patient data from different hospital systems can enable medical research that would have been impossible with only one source's data. Forward-thinking policy is needed to streamline the process to ensure that researchers can gain access to comprehensive records while still protecting individual privacy. For example, this may be achieved through a combination of data sharing agreements, incentives for patients to participate, and strong consumer protection regulations to ensure that patient data cannot be used for harmful purposes.

Fourth, government should offer incentives for the private sector to release and share data. For example, many U.S. colleges and universities collect data on student satisfaction through the National Survey of Student Engagement, but few schools release this data.[9] Making the information public and accessible could increase overall student satisfaction by helping students make better informed decisions about what college to attend. In light of the national benefits, the federal government should require the release of this data as a condition of schools receiving federal aid.

Finally, government leaders should be careful not demonize big data. By vilifying large-scale data analysis, government officials run the risk of unfairly stigmatizing the technology's beneficial applications and later being unable to use them to achieve social and economic goals. Efforts to block or delay such applications on principle will likely have negative consequences in the future. For example, better data and analysis has helped the National Oceanic and Atmospheric Administration nearly double tornado warning lead times, undoubtedly saving lives in the process.[10] In addition, government leaders should be careful not to conflate

---

[8] Castro, Daniel. The Information Technology and Innovation Foundation. "The Need for an R&D Roadmap for Privacy." August, 2012. http://www2.itif.org/2012-privacy-roadmap.pdf. Accessed March 28, 2013.

[9] Atkinson, Robert D. The Chronicle of Higher Education. "Student-Survey Results: Too Useful to Keep Private." November 15, 2009. http://chronicle.com/article/Student-Survey-Results-Too/49128/. Accessed March 31, 2014.

[10] Korte, Travis. The Center for Data Innovation. "Data Science Is Not PRISM: In Defense of Analytics." June 20, 2013. http://www.datainnovation.org/2013/06/data-science-is-not-prism-in-defense-of-analytics/. Accessed March 28, 2014.

government surveillance with the voluntary collection and use of data by the private sector. While some of the means may be the same, they undoubtedly serve different ends.

QUESTION 2

What types of uses of big data could measurably improve outcomes or productivity with further government action, funding, or research? What types of uses of big data raise the most public policy concerns? Are there specific sectors or types of uses that should receive more government and/or public attention?

RESPONSE

Fundamentally, data analysis helps people and organizations make better decisions. In the private sector, these decisions may take the form of a company buying from one vendor instead of another, a farmer planting at a particular place and time, or a person at home choosing to bring an umbrella on an outing. Key decisions in the government that can be aided with data analysis include determining which programs to cut, which companies to audit, and which business processes to implement.

Government has an important role to play in encouraging big data use in fields including health care, education, road safety, weather prediction, financial reporting, mapping and macroeconomic forecasting.

- Health Care: Big data could account for $300 billion in annual savings to the U.S. health system, according to a 2013 estimate from McKinsey & Company.[11] It could also help medical researchers develop new treatments and help doctors make better decisions.[12] However, in order to reap these benefits, government must help resolve the data quality and availability issues that have inhibited large-scale health data analysis in the past. The federal government should continue to make it attractive for states to adopt electronic health record systems, which are crucial for improving data quality and

[11] Kayyali, Basel et al. McKinsey & Company. "The Big-Data Revolution in U.S. Health Care: Accelerating Value and Innovation." April, 2013. http://www.mckinsey.com/insights/health_systems_and_services/the_big-data_revolution_in_us_health_care. Accessed March 28, 2014.

[12] Harris, Derrick. GigaOm. "Better Medicine, Brought to You by Big Data." July 15, 2012. http://gigaom.com/2012/07/15/better-medicine-brought-to-you-by-big-data/. Accessed March 28, 2014. Henschen, Doug. InformationWeek. "IBM's Watson Could Be Healthcare Game Changer." February 11, 2013. http://www.informationweek.com/software/information-management/ibms-watson-could-be-healthcare-game-changer/d/d-id/1108608. Accessed March 28, 2014.

lowering long-term administrative costs. It should also support states developing health information exchanges to ensure that doctors can make the best possible diagnosis and hospitals can avoid costly repeat visits.

- Education: Education data harbors great potential for beneficial applications, including adaptive tutoring systems that change according to individual students' learning styles, and predictive analytics systems to identify students at risk of dropping out. As is the case in health care, government can encourage data use in the education sector by helping make data available; one approach could be encouraging states to adopt centralized student information databases, from which administrators could conduct analytics and other trusted authorities could develop new educational technologies.[13] The United States Department of Education could also encourage greater education data use by offering funding to data-driven educational applications through the Race to the Top program.[14]

- Road Safety: Driving, especially in heavy traffic, is a highly complex activity in which human error often has deadly consequences. As such, it presents a prime opportunity for data-driven automation to make people safer. Vehicle-to-vehicle and vehicle-to-infrastructure communication systems, which allow vehicles to share data on physical variables such as position, speed, and acceleration, may be able to prevent up to 80 percent of road accidents not involving drunk drivers or mechanical failure.[15] The Department of Transportation should continue to move forward on requiring vehicle-to-vehicle systems in new cars and other light vehicles and supporting the deployment of intelligent transportation systems.

- Weather Prediction: Satellite weather prediction data is widely used in the public and private sector alike, fueling navigation services, agricultural software, disaster response, and other applications. However, one of the United States' major satellite weather data collection programs is likely to lose half its capacity in 2016 due to delays in

[13] Castro, Daniel and Travis Korte. The Center for Data Innovation. "Parents and Educators Should Embrace, Not Fear, Student Data." December 3, 2013. http://www.datainnovation.org/2013/12/parents-and-educators-should-embrace-not-fear-student-data/. Accessed March 28, 2014.

[14] Ibid.

[15] Dockterman, Eliana. Time. "Government Wants Cars To Talk To Each Other." February 4, 2014. http://time.com/4291/government-wants-cars-to-talk-to-each-other/. Accessed March 28, 2014.

constructing a satellite to replace the one currently in orbit.[16] In order to maximize private sector value and ensure that life-saving disaster predictions are as precise as possible in the long term, government must not let such gaps occur in the future. The Weather Forecasting Improvement Act of 2013, which would help ensure that the government could purchase weather data from the private sector in the future, offers one potential solution.[17]

- Financial Data Reporting: Agencies must report their financial data to federal authorities for evaluation, but this data is frequently not standardized and unavailable. Ensuring that digital filing data is published in a structured data format could make it easier for the government to identify fraud, waste, and abuse; it could also support large-scale data analysis, which is costly to conduct using traditional filings. The Digital Accountability and Transparency Act of 2013 would standardize how this data is published and help ensure agencies are performing as well as possible.[18]

- Mapping: Many government agencies create maps using geographic information systems (GIS). For example, cities bolster public safety with crime maps, states consult land use data for planning, and the U.S. National Park Service conducts research and conservation efforts with digital mapping software. These maps have considerable value in the private sector, underlying geolocation apps, news services and social and economic research. The company that produces the software behind much of the agencies' map data has recently made it possible for agencies to easily release this data openly, and the White House should encourage agencies to do so as broadly as possible pursuant to the 2013 Open Data Executive Order.[19]

---

[16] Korte, Travis. FedScoop. "US Weather Data Threat Exposes Poor Data Stewardship." November 22, 2013. http://fedscoop.com/guest-column-us-weather-data-threat-exposes-poor-data-stewardship/. Accessed March 28, 2014.

[17] H.R. 2413, "Weather Forecasting Improvement Act of 2013." 113th Congress (2013-2014). http://beta.congress.gov/bill/113th-congress/house-bill/2413. Accessed March 28, 2014.

[18] S. 994, "Digital Accountability and Transparency Act of 2013." 113th Congress (2013-2014). http://beta.congress.gov/bill/113th-congress/senate-bill/994. Accessed March 28, 2014.

[19] Howard, Alexander. ReadWrite. "Government Agencies Will be Able to Make More Geospatial Data Available to Developers and the Public." February 10, 2014. http://readwrite.com/2014/02/10/esri-enable-thousands-government-agencies-open-gis-data-public. Accessed March 28, 2014. White House. "Executive Order: Making Open and Machine Readable the New Default for Government Information" (Washington, D.C., 2013). http://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government. Accessed March 28, 2014.

- Macroeconomic Forecasting: The national statistical agencies have traditionally collected macroeconomic data through surveys, which are costly and only available after a delay of days or weeks. Private sector data sources have the potential to supplement and in some cases eventually replace these surveys with near-real time data. For example, the Bureau of Economic Analysis is experimenting with using anonymized data from financial software firm Intuit to improve its estimates of employment and sales trends.[20]

QUESTION 3

What technological trends or key technologies will affect the collection, storage, analysis and use of big data? Are there particularly promising technologies or new practices for safeguarding privacy while enabling effective uses of big data?

RESPONSE

Emerging technologies and trends will affect all stages of the data lifecycle, from data collection on through storage, processing, analysis, and use.

- Data Collection: Data is collected from multiple sources. One important trend is the increasing variety of devices that collect data. The "Internet of Things" (IoT) refers to the concept that the Internet can now function as platform for devices to communicate electronically with the world around them. From "smart" thermostats that automatically adjust to users habits, to bridges equipped with wireless sensors that detect structural changes and predict catastrophic failures, IoT technologies will become increasingly pervasive in public and private sector applications over the next decade. Enabling innovators to harness the potential of data flowing from one device to another for economic and social good will be a key responsibility of lawmakers as the technologies mature.

- Data Storage: As the cost of digital storage continues to fall, organizations are able to store and use more data. One important technology underpinning the declining cost for storage is cloud computing. Cloud computing refers to the practice of "renting" remotely-located IT services, including processing capabilities, information storage, and software applications, on an as-needed basis. While cloud-based data storage can often be more cost effective than on-premise storage, adoption has been delayed in some government

[20] Korte, Travis. The Center for Data Innovation. "What Big Data Can Do for National Statistics." March 27, 2014. http://www.datainnovation.org/2014/03/what-big-data-can-do-for-national-statistics/. Accessed March 31, 2014.

applications due to the misconception that the cloud is not secure.[21] In fact, for many government applications, transferring data held in aging, ad-hoc databases to market-tested cloud-based systems can be a substantial security upgrade. The federal government should lead the way in adopting cloud storage to increase confidence among state and local data managers.

- Data Processing: In addition to its importance in data storage, cloud computing will have a key role in the future of data processing. The cloud allows organizations to conduct "big data"-scale computing with a minimum of infrastructure. In cases where the data is so large or complex that it requires cutting-edge hardware to process it, organizations can also save maintenance and expertise costs by using the cloud. For example, researchers at University of Southern California used computer simulations to efficiently detect which organic compounds are most suited for next-generation photovoltaic cells. By using the cloud computing platform offered by Amazon Web Services the research team spent only $33,000 instead of the $68 million it would have cost them to build the equivalent computing infrastructure [22]

  Moreover, many important applications from public safety to finance require the ability to make decisions based on "streams" of data in real time. Stream processing tools, which allow data scientists to analyze and rapidly act upon large, continuous flows of data, are now maturing.[23] Stream processing tools tailored for different kinds of data, such as Twitter data, or different applications, such as business analytics, are widely available. This has placed real-time processing capacity within reach of an increasing number of organizations, and cloud-based services are poised to reduce the overhead for small businesses and government agencies to implement real-time solutions even further.

---

[21] Castro, Daniel and Travis Korte. The Center for Data Innovation. "Parents and Educators Should Embrace, Not Fear, Student Data." December 3, 2014. http://www.datainnovation.org/2013/12/parents-and-educators-should-embrace-not-fear-student-data/. Accessed March 28, 2014.

[22] Darrow, Barb. GigaOM. "Cycle Computing once again showcases Amazon's high-performance computing potential." November 12, 2013. http://gigaom.com/2013/11/12/cycle-computing-takes-aws-to-a-whole-other-level-in-hpc/. Accessed March 28, 2014.

[23] Lorica, Ben. O'Reilly Strata. "Expanding Options for Mining Streaming Data." December 15, 2013. http://strata.oreilly.com/2013/12/expanding-options-for-mining-streaming-data.html. Accessed March 28, 2014.

- Data Analysis: One emerging technology that promises to make a major impact on data analysis is deep learning, a set of methods for identifying patterns in complex data that were loosely inspired by mechanisms of information propagation observed in the human brain. As deep learning is often more effective for modeling very complex data than traditional algorithms, the "big data" environment has broadened its impact greatly.[24] Deep learning methods are being applied with increasing success to problems from reading handwriting to determining whether two photographs show the same individual. These methods hold considerable promise for automating costly and repetitive tasks in private sector fields from manufacturing to translation. They will also play a key role in the future of law enforcement, helping auditors detect fraud, transit police recognize faces, and federal investigators determine when someone is lying.

- Data Use: Data is created not only for humans to act upon, but also for machines. In particular, the devices that make up the Internet of Things will be a major user of data in coming years. For example, a smart thermometer in the home will make decisions about how much to heat or cool a room based on data it receives about energy prices, whether anyone is home, and its user's preferences. From factory automation to autonomous (and semi-autonomous) vehicles, computers will be some of the greatest users of data.[25]

QUESTION 4

How should the policy frameworks or regulations for handling big data differ between the government and the private sector? Please be specific as to the type of entity and type of use (e.g., law enforcement, government services, commercial, academic research, etc.).

RESPONSE

Data-driven innovation can only occur if laws encourage use and reuse of data. An environment in which data is collected once and used many times is far more efficient than one in which data must be collected anew for every ad hoc application. This principle applies to the public and private sector alike. In general, government can encourage use and reuse of data in both sectors by allowing entities to collect, buy, and sell data, while carefully restricting harmful uses.

---

[24] Hof, Robert. MIT Technology Review. "Deep Learning." April 23, 2013. http://www.technologyreview.com/featuredstory/513696/deep-learning/. Accessed March 28, 2014.

[25] Korte, Travis. The Center for Data Innovation. "Data Scientists Should Be the New Factory Workers." September 27, 2014. http://www.datainnovation.org/2013/09/data-scientists-should-be-the-new-factory-workers/. Accessed March 28, 2014.

For example, federal agencies should be allowed, and even encouraged, to buy data from the private sector if this allows them to achieve their missions more effectively. There are of course some exceptions. Some government data, such as sensitive tax information or Census responses, may be available for government use only; likewise, some private sector data may not be available for government use, especially without judicial oversight.

Government policies, mandates, and incentives can help make more data available or available in a more useful format. Government agencies, at all levels of government, should adopt an ethic of "open by default," in which timely and accurate data is released in open and machine-readable formats for free and without restrictions. Some private sector organizations are voluntarily participating in data-sharing initiatives. For example, pharmaceutical companies have come together to share data for research purposes.[26] When voluntary sharing does not occur and there is a strong public interest, public policy can be used to mandate or incentive private sector companies to share data. For example, the Securities and Exchange Commission can require publicly traded companies to disclose certain financial and non-financial data using open standards.[27] Similarly, agencies from the Environmental Protection Agency to the Department of Health and Human Services can require organizations in the private sector to collect and share data that may have high public value.

Some fear that without government intervention there is nothing to limit the amount of data that the private sector will collect about individuals. These fears are misguided. In the private sector, the collection of data by companies is held in check by market forces. Users must voluntarily share their information, and companies that lose the trust of their users will not be able to collect data from them in the future. These same market forces do not apply to the government. In government, data collection is only limited by the decisions of government agencies and the political process that holds government officials accountable for their decisions. And while government agencies, especially those in the intelligence community and law enforcement, have a bias in favor of collecting more information, this must be tempered by competing interests, including the civil liberties of individuals and the economic impact of such decisions. Ultimately decisions by government agencies to collect data should be guided by policy that balances a wide variety of competing public interests.

---

[26] Silverman, Ed. Wall Street Journal. "On Data Sharing, Pharmaceutical Companies Finally Open Up." March 18, 2014. http://blogs.wsj.com/corporate-intelligence/2014/03/18/access-to-pharma-trial-data-how-open-is-open/. Accessed March 28, 2014.

[27] Korte, Travis. "XBRL: How to Save a Good Idea from a Bad Implementation." Center for Data Innovation. Accessed March 29, 2014. http://www.datainnovation.org/2013/11/xbrl-how-to-save-a-good-idea-from-a-bad-implementation/. Accessed March 31, 2014.

CENTER FOR DATA INNOVATION

QUESTION 5

What issues are raised by the use of big data across jurisdictions, such as the adequacy of current international laws, regulations, or norms?

RESPONSE

Large-scale data analysis is capable of driving global-scale innovation, but data exchange between jurisdictions is a major obstacle to many beneficial uses of data. For example, medical researchers can only develop personalized treatments for certain rare medical conditions if they can draw data from an array of international sources.[28]

One major impediment to international data access is data residency laws and other laws restricting information flow.[29] These laws prohibit certain data from being stored or accessed outside a given nation's borders. For example, Danish and Norwegian Data Protection Authorities prevent the use of cloud services when servers are not located domestically, and the Canadian provinces of British Columbia and Nova Scotia have instituted laws mandating that personal information in the custody of a public body can only be stored and accessed within the country.[30] Data residency laws prevent firms from offering cross-border services that might otherwise be beneficial for government agencies. U.S. trade negotiators and diplomats should push back against these digital barriers to trade.

In some post-Soviet and East Asian states, state secrets laws prevent many relatively mundane government data sources from being released. In China, this includes company financial statements, soil pollution data, and information on executions.[31] In Russia, as well as in other countries in Eastern Europe that have adopted Russian legal language, authorities have broad

---

[28] Marcus, Amy Dockser. Wall Street Journal. "Families Push for New Ways to Research Rare Diseases." February 18, 2014. http://online.wsj.com/news/articles/SB10001424127887324432004578306364187833702. Accessed March 28, 2014.

[29] Hughes, Krista. Reuters. "Data Privacy Shapes Up as a Next-Generation Trade Barrier." March 27, 2014. http://www.chicagotribune.com/business/sns-rt-us-usa-trade-tech-analysis-20140327,0,6825057.story. Accessed March 28. 2014.

[30] Ezell, Stephen et al. The Information Technology and Innovation Foundation. "Localization Barriers to Trade: Threat to the Global Innovation Economy." September 2013. http://www2.itif.org/2013-localization-barriers-to-trade.pdf. Accessed March 28, 2014.

[31] Rovnick, Naomi. Quartz. "Pollution Levels, Accounting Records, and Other Things China Classifies as 'State Secrets.' February 28, 2013. http://qz.com/57668/pollution-levels-accounting-records-and-other-things-china-classifies-as-state-secrets/. Accessed March 28, 2014.

discretion over what types of data can be classified as state secrets, which can include economic and environmental information.[32] These laws pose a major obstacle to data-driven innovation in these countries, far beyond the cultural and administrative obstacles found in the West. Although the primary benefit from access to this data may be foreign users, ultimately the United States may also benefit from access to the data. The U.S. should continue to support efforts, such as the G8 Open Data Charter, which encourage greater international standards for sharing government data. The U.S. should also engage its trading partners in developing a "Geneva Convention on the Status of Data" that establishes international legal standards for government access to data.[33]

Another impediment to data-driven innovation is the lack of international data standards in certain technological areas. An effort to standardize Internet of Things data formats was recently launched by a consortium of U.S. companies with the help of several U.S. government agencies, and the relevant U.S. federal government agencies should continue to encourage data standardization efforts in other areas, including mobile health and humanitarian aid reporting.[34]

CONCLUSION

The federal government can play a major role in maximizing the potential benefits of big data, but it must above all encourage use and reuse of data. This means allowing data to be collected and retained for serendipitous future applications that were not foreseen at the time of collection, while restricting harmful applications. When it comes to the future of data the biggest

---

[32] Article 19. "Under Lock and Key: Freedom of Information and the Media in Armenia, Azerbaijan and Georgia." April 2005. http://www.freedominfo.org/wp-content/uploads/documents/FOI_and_Media_in_the_South_Caucasus_English.pdf. Accessed March 28, 2014. Pavlov, Ivan. Institute for Freedom of Information Development. "Access to Information: State Secrets and Human Rights." December, 2006. http://www.freedominfo.org/documents/Russia_Access_Information_Pavlov.pdf. Accessed March 28, 2014.

[33] Castro, Daniel. The Information Technology and Innovation Foundation. "The False Promise of Data Nationalism." December 2013. http://www2.itif.org/2013-false-promise-data-nationalism.pdf. Accessed March 31, 2014.

[34] Hardy, Quentin. The New York Times. "Consortium Wants Standards for 'Internet of Things.'" March 27, 2014. http://bits.blogs.nytimes.com/2014/03/27/consortium-wants-standards-for-internet-of-things/. Accessed March 28, 2014. Payne, Jonathan. mHealth Alliance. "The State of Standards and Interoperability for mHealth." March 2013. http://mhealthalliance.org/images/content/state_of_standards_report_2013.pdf. Accessed March 28, 2014. Rahman, Zara. School of Data. "World Humanitarian Data and Trends, 2013- report review." February 24, 2014. http://schoolofdata.org/2014/02/24/world-humanitarian-data-and-trends-2013-report-review/. Accessed March 28, 2014.

risk is not how it will be used, but whether it will be used. Ultimately, the United States needs data-literate policymakers to champion opportunities to leverage data if we are to achieve its full potential.

Sincerely,

Daniel Castro

Director, Center for Data Innovation
1101 K Street NW, Suite 610
Washington, DC 20005

dcastro@datainnovation.org

31 March 2014

John Podesta, Senior Counselor to the President
Nicole Wong, Deputy Chief Technology Officer, OSTP
Big Data Study
Office of Science and Technology Policy
Eisenhower Executive Office Building
1650 Pennsylvania Ave., NW
Washington, DC, 20502

Dear John and Nicole:

The Center for Digital Democracy (CDD) respectfully submits these comments for the Administration's review. While today's "Big Data"-driven landscape may appear to be a recent development, it is really the consequence of historical trends in data processing, the growth of digital platforms, and the evolution of the commercial online marketing business model. The overall dimensions of today's pervasive data collection complex were set in the middle 1990's. The principle commercial Internet business paradigm (i.e., the collection and analysis of individuals' information so they can be tracked and targeted) has long been based on what is called "one-to-one" marketing. Over the last two decades, such "1:1" marketing has expanded from tracking an individual on a single website to the vast data collection apparatus that operates in real-time today.[1] We are closely monitored and our behaviors analyzed across devices and applications (mobile, PC, gaming); our social media actions—as well as those of our "friends"—are scrutinized; and real-time location and geographic behaviors are gathered. Few Americans know that they are connected to a digital dossier—a so-called profile—that is filled to the brim through the contributions of dozens of increasingly allied data brokers. Or that their information is used to make assessments or predictions about them—such as their "Lifetime Value" for companies engaged in financial services. Nor are they informed that their profiles are increasingly auctioned off in milliseconds to the highest bidder, so they can be targeted wherever they are—including when using their mobile devices.

[1] See, for example, Jeff Chester, *Digital Destiny: New Media and the Future of Democracy* (New York: The New Press, 2008), chapter 7; and Jeff Chester, "Cookie Wars: How New Data Profiling and Targeting Techniques Threaten Citizens and Consumers in the 'Big Data' Era," in Serge Gutwirth, et al, eds., *European Data Protection: In Good Health?* (New York: Springer, 2012): 53-77.

The inability to implement basic privacy rules in the United States to address Internet data collection practices has resulted in the ubiquitous commercial surveillance landscape that today threatens the privacy of Americans—as well as those in the European Union and other countries where U.S. companies collect and transport their information.[2] The absence of a clear legislative proposal from the Administration has contributed to the growing threat to privacy that Americans confront. The online data industry sees no credible challenge from the White House that would encourage it to stem the ever-increasing tide of personalized collection. It has now been more than two years since the White House announced a "Privacy Bill of Rights," explaining that the public "cannot wait" for much-needed consumer protection safeguards. CDD urges the Administration to release with its forthcoming Big Data report a specific legislative proposal to implement these privacy rights. Americans deserve to know where the Administration stands. Does it believe that individuals have the right to control how their data are collected and used? Will the Federal Trade Commission and other agencies responsible for privacy be empowered to engage in necessary rulemakings that protect the public?[3]

---

[2] A key exception to the failure to protect online privacy from the federal government was the enactment of the Children's Online Privacy Protection Act (COPPA) in 1998. The author of this comment, along with Professor Kathryn C. Montgomery of American University, played a key role in its passage. As COPPA demonstrates, a well-crafted policy that places individuals in control over their data collection can address changes in the data collection marketplace—putting to rest the purposefully disingenuous argument that the Internet is so dynamic that regulation is impossible. See Kathryn C. Montgomery, *Generation Digital: Politics, Commerce, and Childhood in the Age of the Internet* (Cambridge, MA: 2007); and Federal Trade Commission, "FTC Strengthens Kids' Privacy, Gives Parents Greater Control Over Their Information By Amending Children's Online Privacy Protection Rule,"19 Dec. 2012, http://www.ftc.gov/news-events/press-releases/2012/12/ftc-strengthens-kids-privacy-gives-parents-greater-control-over. The failure to enact both a fundamental policy law in the U.S. and an effective set of policies for Safe Harbor has placed data flows between the U.S. and EU as risk. See, for example, Jennifer Baker, "Safe Harbor: Reding Warns US that Progress is Needed Before Summer," viEUws, 26 Mar. 2014, http://www.vieuws.eu/ict/safe-harbor-reding-warns-us-that-progress-is-needed-before-summer / (both viewed 30 Mar. 2014).

[3] The White House, "We Can't Wait: Obama Administration Unveils Blueprint for a 'Privacy Bill of Rights' to Protect Consumers Online," 23 Feb. 2012, http://www.whitehouse.gov/the-press-office/2012/02/23/we-can-t-wait-obama-administration-unveils-blueprint-privacy-bill-rights. The Administration's approach to a "multistakeholder" process to develop industry "codes of conduct," part of its "We Can't Wait" announcement, is an inadequate approach to ensuring consumer protection and privacy. An industry-dominated convening, in which there is also no meaningful disclosure of actual data collection practices by leading companies and trade associations, has demonstrated that it cannot develop the appropriate safeguards. For a review of the first NTIA multi-stakeholder proceeding and its failings, see Center for Digital Democracy, "New Report Exposes Flaws in NTIA 'Multistakeholder' Effort to Establish Privacy Safeguards: White House Must Act to Fulfill its Vision for a 'Privacy Bill of Rights'/See Cross-platform Tracking/Users as $ 'whales,'" 29 Aug. 2013, http://www.centerfordigitaldemocracy.org/new-report-exposes-flaws-ntia-%E2%80%9Cmultistakeholder%E2%80%9D-effort-establish-privacy-safeguards-white-house-mus (both viewed 30 Mar. 2014).

As we expressed in a meeting the White House had with several privacy NGOs several weeks ago, CDD believes the Big Data report must address the *realties* of today's commercial data gathering and analysis landscape. While we acknowledge the many positive uses of Big Data, and its potential, the Administration should not gloss over the threats as well.[4] We fear that missing for the most part in the White House's review will be a fact-based assessment of actual commercial data practices conducted by Google, Facebook, Yahoo, data brokers, and many others. Such a review would reveal an out-of-control commercial data collection apparatus, with no restraints, and which is leading to a commercial surveillance complex that should be antithetical in a democratic society. The report should show the consequences of such information gathering on Americans, where the data can be immediately made "actionable." It should address the consequences when predictive analysis and other "insight" identification applications trigger real-time and future decisions about the products and services we are offered, the content we may receive, and even the online "experiences" with which we interact. The report should make clear how its Consumer Privacy Bill of Rights Principles should be interpreted when data collected from Americans are used to unfairly target them—and their families—for products and services that can be harmful to their well-being (such as the delivery of high-interest payday loans, promotion of questionable medical treatments, and the targeting of junk food ads to children, which contributes to the nation's obesity epidemic).[5]

CDD respectfully urges the White House to address in its report the following issues as raising privacy and consumer protection concerns, and requiring action by policymakers. For brevity, we will only briefly describe each issue and provide a few examples.[6]

- *The Growth of Ubiquitous Cross-Platform and Across-Application Tracking of Individuals Online:* Today, consumers are increasingly tracked across the devices they use, such as PC, mobile, gaming, and soon even TV. Companies are also expanding beyond cookies to create single identifiers on a person. The same person confronts a data collection environment that also captures their behavior and actions on social media, mobile apps, websites and more. As one online marketing publication recently explained about the implications of device identification, "Never before has digital tracking become so personal and never before has the argument for

[4] We note, however, that much of the commentary on the potential of Big Data is often akin to public relations. Recent critical analysis of the failure of Google's Flu Trends data to provide meaningful results is one example. Declan Butler, "When Google Got Flu Wrong," *Nature*, 13 Feb. 2013, http://www.nature.com/news/when-google-got-flu-wrong-1.12413 (viewed 30 Mar. 2014).

[5] In addition to CDD, there are other experts specializing in addressing the impact of today's commercial data apparatus on the public. See, for example, Joseph Turow, *The Daily You: How the New Advertising Industry Is Defining Your Identity and Your Worth* (New Haven, CT: Yale University Press, 2012).

[6] However, CDD follows this field very closely and is happy to provide further analysis and documentation of the problems we describe.

consumer privacy controls been so compelling."[7] Such all-encompassing data-gathering practices are at odds with long-standing Fair Information Privacy Practices requiring data minimization, as well as the Administration's "Focused Collection" rights principle.[8]

- *The Emergence of Big-Data-derived Comprehensive Data Profiles on Individuals (Data Management Platforms)*: Evolving from the one-to-one marketing paradigm described earlier, companies are focused on collecting and making actionable as much of a person's information as possible. The data broker company Merkle has called this process "Connected Recognition"; others describe similar "360 degree," "Master" profiles, and "multi-channel" approaches. Few consumers know that all their information is increasingly stored in one central repository such as a data management platform (DMP)—that can include their financial data, social media usage, location, demographics, ethnicity, and much more. These data profiles can be connected to "Experience Managers" and other applications used by marketers and others to make determinations—and take action—concerning the products and services to offer an individual. The continuous gathering and use of a person's information should be challenged as posing a fundamental threat to privacy in America today. The unfettered growth of commercial data profiles is at odds with many of the Administration's

[7] Gavin Dunaway, "ID Is Key: Unlocking Mobile Tracking & Cross-Device Measurement, Part I," AdMonsters, 2 Aug. 2013, http://www.admonsters.com/blog/id-key-unlocking-mobile-tracking-cross-device-measurement-part-I; Gavin Dunaway, "ID Is Key: Unlocking Mobile Tracking & Cross-Device Measurement, Part 2," AdMonsters, 3 Aug. 2013, http://www.admonsters.com/blog/id-key-unlocking-mobile-tracking-cross-device-measurement-part-2 (both viewed 5 Feb. 2014).

[8] See, for example, Google, "The New Multi-Screen World Study," Think Insights, Aug. 2012, http://www.thinkwithgoogle.com/research-studies/the-new-multi-screen-world-study.html. See also Google, "The Customer Journey to Online Purchase," Think Insights, http://www.thinkwithgoogle.com/tools/customer-journey-to-online-purchase.html (both viewed 5 Feb. 2014); Steve Schuler, "How to Reach Consumers Across Devices with Sequential Messages," Yahoo Advertising, 16 Sept. 2013, http://hispanicad.com/agency/digital/how-reach-consumers-across-devices-sequential-messages; "Marketing for Cross-screen Sequencing," ANA Magazine, Issue 8, 2012, p. 7, http://www.ana-thoughtleadership.net/ana-thoughtleadership/2013_issue_8#pg7 (both viewed 5 Feb. 2014); Nielsen, "Unleashing Cross-Platform: The Tip of the Spear," 13 Nov. 2013, http://www.nielsen.com/us/en/newswire/2013/unleashing-cross-platform-the-tip-of-the-spear.html. See also Becky Chappell, "Making Mobile Work Across the Advertising Industry," DoubleClick Advertiser Blog, 1 Nov. 2013, http://doubleclickadvertisers.blogspot.com/2013/11/making-mobile-work-across-advertising.html. "What The Google AdID Means For Ad Tech," Ad Exchanger, 19 Sept. 2013, http://www.adexchanger.com/data-driven-thinking/what-the-google-adid-means-for-ad-tech/. For examples from Facebook, examine the relationship between what a consumer does outside of the social network who can then be targeted while there. See, for example, "Datalogix Announces Facebook Partner Categories for CPG, Retail and Automotive Brands," Datalogix, Apr. 2013, http://www.datalogix.com/2013/04/datalogix-announces-facebook-partner-categories-for-cpg-retail-and-automotive-brands/ (all viewed 30 Mar. 2014).

Privacy Principles, including Transparency, Individual Control, Focused Collection, and Accountability.[9]

- *The Digital Data Collection Apparatus, Including the Use of Multiple Data Sources and the Real-time Buying and Selling of American Internet Users:* Americans do not know—nor can they effectively control—the myriad of data that is collected and used on them. Data brokers and other similar providers increasingly work together to pool and sell their information on a single person or to create a highly refined segment. This information is made available for sale, and is combined with the data resources on individuals that e-commerce and Internet marketing companies routinely capture. Our data profiles feed a far-ranging automated system of online auctions on individual Americans—a dehumanizing process in which we are sold to the highest bidder through ad exchanges, regardless of whether we are online or using a mobile device (and soon even in front of a TV). Without our awareness or consent, our information (including behaviors, ethnicity, race, financial interests, etc.) is treated as a mere commodity for sale. Both the role of data broker alliances and the use of super-fast computers to track and sell Americans raise concerns related to a number of the Administration's Privacy Rights, including Transparency, Individual Control, Respect for Context, Focused Collection, and Accountability.[10]

---

[9] Security is also implicated, of course, as the recent explosion of major data breaches illustrates. Nor is there any meaningful access or methods to ensure accuracy. For background on the growth and use of these comprehensive Big Data dossiers on Americans, see, for example: Merkle, "Connected Recognition: Enabling a 360° View of the Customer," http://www.merkleinc.com/what-we-do/database-marketing-services/connected-recognition; Acxiom, "Macy's Focuses on the Customer; Builds Comprehensive View Across Touch Points," http://acxiom.com/macys-builds-comprehensive-view-customer/; Susan Bidel, "Boosting First-Party Data Effectiveness With DMPs," Forrester, 10 Jan. 2014, http://www.forrester.com/Boosting+FirstParty+Data+Effectiveness+With+DMPs/fulltext/-/E-RES108301; Acxiom, "Multichannel Marketing Solutions," http://www.acxiomdigital.com/services/multichannel-marketing.asp; eMarketer, "To Handle Big Data, Advertisers Turn to DMPs," 24 May 2013, http://www.emarketer.com/Article/Handle-Big-Data-Advertisers-Turn-DMPs/1009919; "Adobe Digital Marketing Summit Hits on Key Theme of Marketing Reinvention," 25 Mar. 2012, http://www.businesswire.com/news/home/20140324006499/en/Adobe-Digital-Marketing-Summit-Hits-Key-Theme#.UzgzI8dGJnY. See also Adobe's new "Master Marketing Profile" application that is integrated into its marketing cloud service. Adobe, "Profile Management," http://www.adobe.com/solutions/digital-marketing/profile-management.html (all viewed 30 Mar. 2014).

[10] Kevin Weil, "Driving Mobile Advertising Forward: Welcoming MoPub to the Flock," Twitter Blog, 9 Sept. 2013, https://blog.twitter.com/2013/driving-mobile-advertising-forward-welcoming-mopub-to-the-flock; Nexage, "Nexage Reaches Major Milestone with More than 50 Percent of Spend Now through RTB," 17 July 2013, http://www.nexage.com/resources/press-releases/nexage-more-than-50-percent-of-spend-now-through-real-time-bidding; Kelly Liyakasa, "Ads Across Amazon: O&O Sites Vary in RTB and Data Readiness," Ad Exchanger, 5 Dec. 2013, http://www.adexchanger.com/ecommerce-2/sizing-amazons-scope-the-owned-and-operated-opportunity/; DoubleClick, "Solutions for Publishers: Sell Across All Screens," http://www.google.com/doubleclick/publishers/solutions/mobile.html (all viewed 4 Feb. 2014);

- *The Growth of Commercial Digital Surveillance at the Community, Hyper-local Level:* Geo-location data are increasingly being gathered and made actionable, with capabilities permitting real-time responses to events and individual behaviors. The rapid adoption of mobile devices and mobile data-driven marketing practices enables companies both to collect data and to reach individual consumers at any time. Mobile analytics identify, track, measure, and help make actionable a wide array of user behaviors (including those that can be used to identify the spending behavior of individuals).

  Intrusive data collection practices increasingly monitor and assess the individuals, businesses, and institutions residing in a discrete "micro-neighborhood." For example, online data marketers have further sub-divided the country into so-called "tiles," which are used to identify unique characteristics of a community. The use of these "tiles" raises serious privacy and consumer concerns. As one hyper-local targeting company explained, "What we do is map data from multiple sources onto a grid of tiles that cover every square foot of the US. Each tile is 100 meters by 100 meters, and we inject third-party demographic information about that area into the tile, as well as data on what's physically located there—points of interest like parks and airports, tourist attractions, retailers, stadiums, and so forth. Then, we connect that data with where a mobile device is in real time, or where it has recently been, to build unique audience segments for brands to target."[11] The information derived from these "tiles" can help generate a "score" on an individual—or a neighborhood—that becomes part of a data profile and a decision point about their value (or lack thereof). Americans should not have to trade away their privacy—let alone make themselves or their neighbors vulnerable to ongoing and invisible scrutiny—in order to discover a gas station or drug store using an online map. The growth of commercial surveillance at the local level raises the specter of new forms of discrimination or unfair practices (and the continuation of historic practices related to redlining, for example). As

---

Natasha Singer, "Your Online Attention, Bought in an Instant," *New York Times*, 17 Nov. 2012, http://www.nytimes.com/2012/11/18/technology/your-online-attention-bought-in-an-instant-by-advertisers.html?pagewanted=all&_r=0. Programmatic buying is being applied to TV, which will eventually have similar data tracking and targeting capabilities now seen online. See, for example, AudienceXpress, "The Platform," http://www.audiencexpress.com/the-platform/. For an example of data partnerships and data coops, see Turn, "The Turn Partner Ecosystem," http://www.turn.com/en-gb/data-partners; Brilig, "Coop Members," http://www.brilig.com/coop-members.php (all viewed 30 Mar. 2014).

[10] Jeremy Lizt, "Data Onboarding System Overview," LiveBlog, 3 Oct. 2012, http://blog.liveramp.com/2012/10/03/data-onboarding-system-overview/; Fahim Zaman, "New LiveRamp Connect Allows Brands to Integrate Data into 70+ Marketing Platforms," LiveBlog, 21 Jan. 2014, http://blog.liveramp.com/2014/01/21/new-liveramp-connect-allows-brands-to-integrate-data-into-70-marketing-platforms/; Datalogix, "DLX OnRamp," http://www.datalogix.com/dlx-onramp; Michel Benjamin, "1st, 2nd, 3rd Party Data: What Does it All Mean?" Lotame, http://lotame.com/1st-2nd-3rd-party-data-what-does-it-all-mean (all viewed 30 Mar. 2014).

[11] Placed, "Placed Targeting," https://www.placed.com/targeting (viewed 5 Feb. 2014).

mobile payment data becomes increasingly merged with geo-location and other data, communities will confront an even more formidable system that can either help or harm their future. Mobile and hyper-local data practices raise each of the Administration's Privacy Rights principles.[12]

- *The Delivery of Financial, Health, and Other Products Linked to Sensitive Data and Uses that Raise Consumer Protection Concerns*: If all the data collected on an individual today were merely being stored, that alone would be a major privacy concern. But such data are analyzed and used to make decisions about us—including for the targeting of products and services tied to our livelihoods, well being, and families. Financial services companies, for example, are using data analytics and generating insights from our information to determine the products and services we are offered in the marketplace. The growing role of so-called "e-scores" can determine—invisibly and without accountability—our "lifetime value" and credit worthiness. CDD filed in this proceeding late last week, along with the U.S. PIRG Education Fund, a new report on Big Data and financial products and services that explores this issue in depth. CDD and a coalition of consumer, public health, and child advocacy groups are also filing today with OSTP a call for new safeguards related to Big Data and the obesity crisis linked to food and beverage marketing to youth online. The White House report should call for the strongest set of Consumer Privacy Bill of Rights safeguards covering sensitive information and their applications, especially when connected to a product or service that involves finance, health, race/ethnicity, young people and seniors.[13]

---

[12] Jesse Haines and Abigail Posner, "The Meaning of Mobile," Think with Google, Oct. 2012, http://ssl.gstatic.com/think/docs/the-meaning-of-mobile_research-studies.pdf; Verizon, "Unprecedented Insights, in Your Neighborhood," http://business.verizonwireless.com/content/dam/b2b/precision/Precision_Phoenix_Market_Infographic.pdf; "Measurement for Mobile App Ads," Facebook Developers, https://developers.facebook.com/docs/ads-for-apps/measurement/; eMarketer, "How to Use Location Data to Target Unique Mobile Audiences," personal copy; Stephen Milton and Duncan McCall, "Apparatus and Method for Profiling Users," United States Patent 8,489,596, 16 July 2013, http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PALL&p=1&u=%2Fnetahtml%2FPTO%2Fsrchnum.htm&r=1&f=G&l=50&s1=8489596.PN.&OS=PN/8489596&RS=PN/8489596 (all viewed 4 Feb. 2014).

[13] Center for Digital Democracy, "Report Examines Both the Promise and the Potential Dangers of the New Financial Marketplace: Leading Reform Groups Call for New Regulations to Protect Consumers from Unfair and Discriminatory 'Big Data' Practices," 27 Mar. 2014, http://www.democraticmedia.org/report-examines-both-promise-and-potential-dangers-new-financial-marketplace-leading-reform-groups-c; Center for Digital Democracy, "Protecting Consumer Privacy and Welfare in the Era of 'E-Scores,' Real-time Big-Data 'Lead-Generation' Practices and other Scoring/Profile Applications [USPIRG/CDD FTC Filing]," 18 Mar. 2014, http://www.democraticmedia.org/protecting-consumer-privacy-and-welfare-era-%E2%80%9Ce-scores%E2%80%9D-real-time-big-data-%E2%80%9Clead-generation%E2%80%9D-practice; Federal Trade Commission, "#516: Request for Comments and Announcement of FTC Workshop on Spring Privacy Series," http://www.ftc.gov/policy/public-comments/initiative-516 (all viewed 30 Mar. 2014).

- *The Failure of Industry Self-regulation and the Limits of the Multi-stakeholder Process*: Finally, CDD urges the White House to look closely at the role and realities of so-called privacy self-regulation. As this comment illustrates, the growth of cross-platform and data broker-enhanced collected information on individuals continues—but without any corresponding ways to protect privacy. While online marketers, for example, declare that they are engaged in "anonymous" data gathering, even a cursory examination of their practices reveals disturbing expansion in the use of our personal data. Data companies, including Google and Facebook, publicly declare that they care about privacy—but actually engage in activities that constantly undermine our ability to make meaningful personal decisions about how and whether our information can be used. Scholarly research has demonstrated the inadequacies of the nearly invisible "icon" that is the most visible component of the marketing industry's self-regulatory program. News reports have revealed industry opposition to modest calls that would create a "Do-Not-Track" system. Industry lobbyists fight any regulatory proposal that would empower citizen and consumer choices for privacy. Self-regulation is designed to give the *appearance* of protecting privacy without actually doing anything to stem the powerful data collection tide. Technological solutions offer no magic digital bullet, either, although they can play a role. When the default is collection and use, which is how the online medium has been purposefully structured, it's not practical for consumers to try to "turn off" the data machine. There have to be regulatory rules that limit the collection of data and empower individuals to make their own privacy decisions.

The Department of Commerce's "multi-stakeholder" process, convened by the NTIA, isn't capable of developing a meaningful solution, either. It has failed to address how the contemporary data collection apparatus *actually works*, which is essential if one is to identify a framework that can better protect Americans. Beyond its unwillingness to focus on the integrated data environment that consumers confront, the NTIA hasn't demonstrated an interest ensuring a robust analysis of the two issues it has tried to address so far. Industry lobbyists also vastly outnumber consumer and privacy groups, and the process is rife with conflicts of interest. Leading companies refuse to discuss their actual practices or plans—even when connected to the issue being discussed. The belief by some in the Administration that the Internet is too dynamic to regulate is misplaced. So too is the focus on permitting so-called stakeholders to help set the rules of the data collection road. Such a process is akin to allowing lobbyists to draft legislation for their own industries. An examination of the role of stakeholders in addressing how to protect privacy will reveal their inability to rise above their own corporate imperatives in order to support pro-consumer policies. Rather than focusing on corporate "codes of conduct," the Administration should make it clear how it would like Congress to interpret its Privacy Bill of Rights.[14]

---

[14] Brad Stenger, "CHI 2012 Conference Q&A With Lorrie Cranor and Pedro Leon," *New York Times*, 17 July 2012, http://open.blogs.nytimes.com/2012/07/17/chi-2012-conference-qa-with-lorrie-cranor-and-pedro-leon/?_php=true&_type=blogs&_r=0; Kate Kaye, "Study: Consumers

Americans should not have to trade away their privacy or consumer protections in order to participate in the commercial arena—or in the marketplace of ideas in the Internet era. The Big Data report from the Obama Administration will be one of its key legacies—and historians and others will look back to see how willing the President and his advisors were to be candid with the American public and urge for the privacy safeguards and practices they deserve. Americans should no longer have to wait.

Respectfully submitted,

Jeff Chester
Executive Director
Center for Digital Democracy

Don't Know What AdChoices Privacy Icon Is," *Ad Age*, 29 Jan. 2014, http://adage.com/article/privacy-and-regulation/study-consumers-adchoices-privacy-icon/291374/; http://www.nytimes.com/2012/10/14/technology/do-not-track-movement-is-drawing-advertisers-fire.html; Natasha Singer, "Do Not Track? Advertisers Say 'Don't Tread on Us,'" *New York Times*, 13 Oct. 2012, http://www.centerfordigitaldemocracy.org/new-report-exposes-flaws-ntia-%E2%80%9Cmultistakeholder%E2%80%9D-effort-establish-privacy-safeguards-white-house-mus (all viewed 30 Mar. 2014).

Center for National Security Studies

Protecting civil liberties and human rights

Director
Kate Martin

Comments to the Office of Science and Technology Policy regarding Government "Big Data" Request for Information

from the Center for National Security Studies Washington, D.C.

Submitted 3/31/14 via bigdata@ostp.gov

The following comments are submitted in response to the numbered questions set forth in the RFI.

(1) The policy and legal frameworks currently governing intelligence and law enforcement access to and use of information about individuals in the United States are not adequate to address issues raised by "big data." Historically, these frameworks have been based in large part on the premise that the government should only collect information on individuals in the United States for law enforcement or intelligence purposes when it has some kind of specific factual predicate about an individual or event. They were not intended to address the possibilities of "Big Data." While many of the statutory changes enacted since 9/11 have weakened or even eliminated such predicates, there has been inadequate public debate and understanding of the implications of such changes, especially in the context of "Big Data." At the same time, the intelligence agencies have apparently devoted substantial resources to the secret development of tools for collection, analysis and use of "Big Data" about individuals.

(2) Uses of "Big Data" by law enforcement and especially intelligence create unique concerns about risks to individual rights as well as democratic governance because such uses are shrouded in much greater secrecy than is the case in the commercial sector or in other governmental agencies. At the same time, the availability of "Big Data" poses separate risks to necessary intelligence capabilities for generating useful and accurate information. As the President's Review Group pointed out, "Big Data"cannot answer all important intelligence questions; accordingly, it is important to ensure that the intelligence community understands the limitations as well as potential benefits of "Big Data."

We urge a greater public effort to identify the ways in which the benefits and risks of law enforcement and intelligence use of "Big Data" differ from issues raised by other sectors' use of "Big Data". We need to develop a common understanding of the ways in which issues regarding "Big Data" that arise in commercial or other contexts both are and are not applicable to law enforcement and intelligence use.

Identifying and understanding the issues, benefits, and risks specific to law enforcement/national security uses of Big Data is an essential prerequisite for designing processes to determine useful and permissible uses as well as safeguards against misuse.

Specific or unique considerations include for example:

-- the government's intent to use such information/analysis to target individuals or groups for law enforcement or intelligence activities;
-- the difficulty or even impossibility of obtaining consent from the subject;
-- the effect of secrecy on a risk-management analysis of such uses and on any after-the-fact analysis of the effectiveness of such uses;
-- the effect of specialized, secret knowledge within the intelligence community on the relevant analyses; and
-- that the risks from "Big Data" are likely to include risks beyond privacy risks, which are not present in many other contexts, e.g., risks posed to democratic governance and public trust in the government as identified by the President's Review Group's report.

Development of protocols/procedures for evaluating and approving "Big Data" uses by law enforcement or intelligence.

Given that law enforcement and intelligence by definition operate in greater secrecy, are important for national security, and at the same time raise significant issues concerning protection of individual rights and democratic governance, it would be useful to outline a set of questions and procedures that should be followed by such agencies when deciding whether and how to use specific sets of "Big Data."

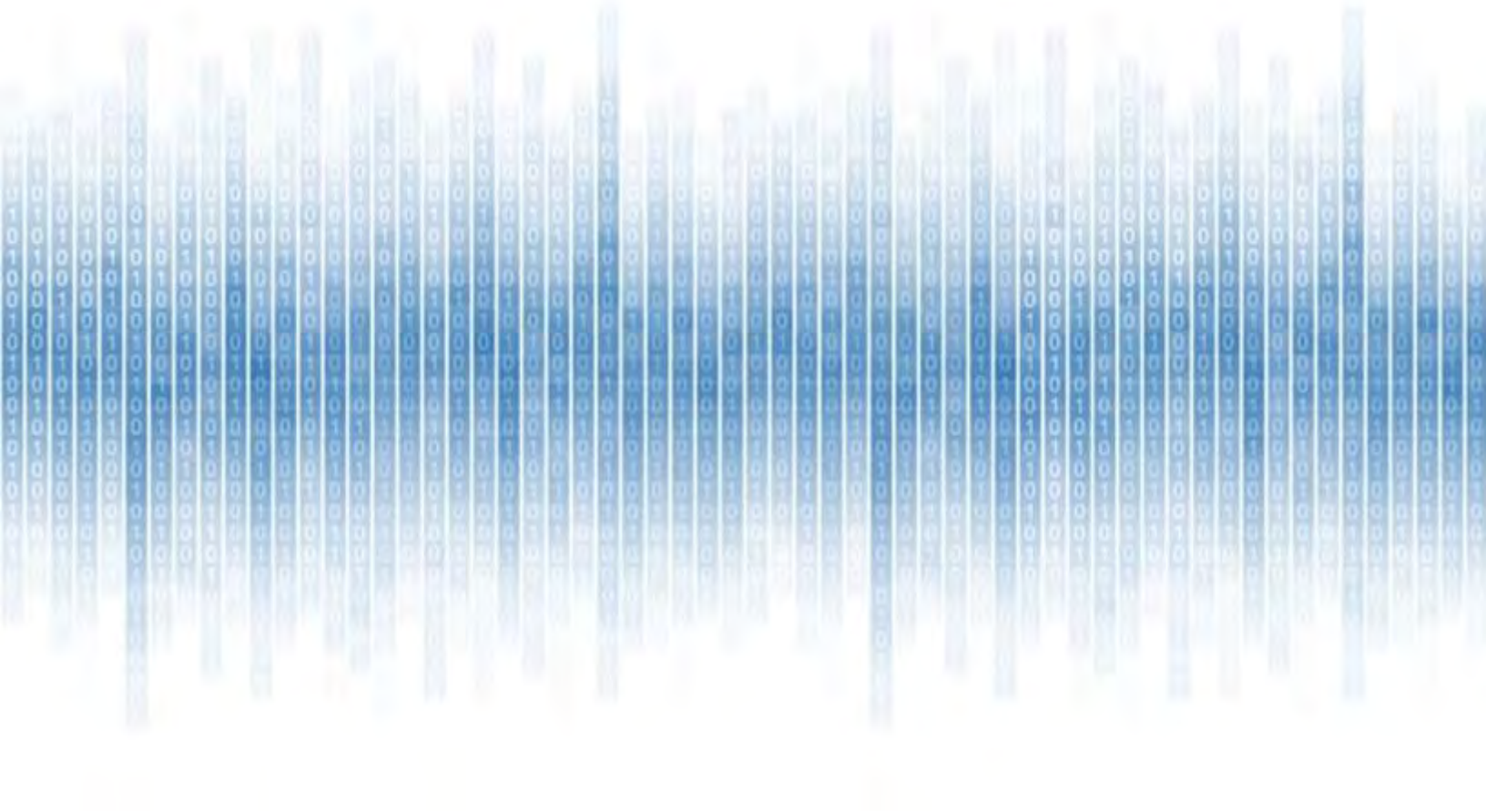
Those questions should include for example:

- What question would be answered with the use of "Big Data," what intelligence would it provide?
- Is the availability of Big Data, especially communications "Big Data," likely to skew intelligence analysis, for example by encouraging a focus on issues which can be addressed by "Big Data" without considering whether more important intelligence priorities are being overlooked?
- Are there alternative means of gathering and analyzing useful intelligence, which would gather less data on individuals or use more targeted collection and analysis methods?
- How can "Big Data" contribute to effective intelligence without reinforcing stereotypes or bias, which prejudice communities and individuals while also diminishing effective U.S. intelligence capabilities.
- In particular, the intelligence community has adopted the metaphor that it needs to have the haystack to find the needles relevant to terrorism. What does this mean for the use of "Big Data"? How can the validity of this claim be measured? What are the costs/benefits of this approach? (See NSF study on counter-terrorism and data-mining.) What kinds of legal/technological safeguards are available to improve the

effectiveness of such an approach and to limit the risks?

- Are there ways in which "Big Data" could provide additional privacy protections and reduce the amount of needed intelligence collection. For example, the intelligence community has for several years claimed that it is difficult to determine which phone calls or emails are from/to a person in the United States, rather than a foreign citizen overseas. This factual claim underlies the government's claim that full Fourth Amendment protections need not be accorded individuals in the United States before seizing the contents of their communications for foreign intelligence purposes, and current law is partly based on that claim. Could the "Big Data" already collected by the government help identify which communications are in fact from/to a person in the United States? If so, that capability should be publicly disclosed so that it can be factored into the legal constitutional analysis concerning government collection of individual communications.

cloud security alliance℠



Big Data Working Group

Comment on Big Data and the Future of Privacy

March 2014

(1) What are the public policy implications of the collection, storage, analysis, and use of big data? For example, do the current U.S. policy framework and privacy proposals for protecting consumer privacy and government use of data adequately address issues raised by big data analytics?

Public policy implications have a bearing on Access, Ownership, Privacy, Liability, and Transparency.

Privacy protection has become an elusive goal in the big data era as researchers have shown that "linkability threats" can re-identity individuals. Due to the highly personal nature of data of individuals, the policy framework should lead to best practices to store and transmit the data. Existing practices focus on keeping data encrypted at rest and in transit with an infrastructure to ensure proper authorization and authentication of entities to get access to the data. With the advent of big data era, analytics tools require access to raw data for generating information of high value for both individuals as well as third party organizations. In practice, such data is shared after sufficient removal of unique identifiers by the processes of anonymization and aggregation. This process which has led to very many instances of re-identification based on big data linkability needs to be strengthened. The policy framework needs to address systematizing privacy preserving data disclosure mitigating linkability threats in the big data era. Specifically, the policy framework might have to lead to enforcement that all linkable data be encrypted. Furthermore, the policy framework needs to address the concerns on the geo location where the data is stored. As well as enforcement of transparency: individuals have the right to know which party has which access to their data, how the (raw) data is used and how it is protected.

(2) What types of uses of big data could measurably improve outcomes or productivity with further government action, funding, or research?

Following are examples of some of the uses of big data to improve outcomes:

1. Jobs data matching geo location data and education data will lead to better employment outcomes.
2. Sharing cyber threat intelligence among multiple businesses will lead to thwarting potential cyber threats to national infrastructure. There is a need for more funding in developing big data analytics techniques in cyber security.
3. Big data analytics on encrypted data to thwart linkability threats.
4. Improvements in health care, leading to more personalized medicine and treatment. This should also result in a more cost effective health care system.
5. Smarter city and transport infrastructure, leading to a most cost effective and greener environment, lesser and faster commute.

What types of uses of big data raise the most public policy concerns?

The following are the most public policy concerns:

1. Correlation of disparate data such as healthcare, financial, demographic and location data.

2. Tracking consumer behavior and sharing them with 3rd party without proper authorization for targeting and other purposes.

3. Big data storage in the cloud across multiple geo boundaries

4. Lack of transparency: who has access to which data, which data is collected and for what reason.

Are there specific sectors or types of uses that should receive more government and/or public attention?

Healthcare Education, Financial wellness, Employment, Mobility, and Information Access are some of the specific sectors that should receive more government/public attention.

(3) What technological trends or key technologies will affect the collection, storage, analysis and use of big data?

Predictive analytics, real time analytics, complex event processing, stream computing, high performance computing, cloud computing, deep machine learning algorithms, open source technologies, visualization and mobile apps usage are some of the technological trends that will affect the collection, storage, analysis and use of big data.

Are there particularly promising technologies or new practices for safeguarding privacy while enabling effective uses of big data?

(Somewhat) Homomorphic Encryption and Differential Privacy are some of the promising technologies for safeguarding privacy while enabling effective uses of big data. It should be noted that although technologies play an important role in the safeguarding of privacy, the approach should also include, amongst others, legal and administrative aspects.

The Cloud Security Alliance (CSA) Big Data Working Group (BDWG) has come up with 100 best practices to enhance the security and privacy of big data:
https://docs.google.com/document/d/1FqeHlA53sliNS3sd3ECy2hwyJu0UJDZT71zUs-02nX4/edit#

The top 10 best practices are listed below:

1. Authorize access to files by predefined security policy

2. Protect data by data encryption while at rest

3. Implement Policy Based Encryption System (PBES)

4. Use antivirus and malware protection systems at endpoints

5. Use big data analytics to detect anomalous connections to cluster

6. Implement privacy preserving analytics

7. Consider use of partial homomorphic encryption schemes

8. Implement fine grained access controls
9. Provide timely access to audit information
10. Provide infrastructure authentication mechanisms

(4) How should the policy frameworks or regulations for handling big data differ between the government and the private sector?

The policy frameworks and regulations for handling big data differ between the government and private entities in the context of the ability of the government to adjudicate -- for example, in policies governing demographics data in law enforcement and government investment of tax dollars.

(5) What issues are raised by the use of big data across jurisdictions, such as the adequacy of current international laws, regulations, or norms?

The question of where the data is stored, where the data is processed and where the data analytics results are distributed influence the cross boundary jurisdictions pertaining to privacy policies and regulations.

March 31, 2014

Nicole Wong, Deputy Chief Technology Officer
Office of Science and Technology Policy
Eisenhower Executive Office Building
1650 Pennsylvania Avenue, NW
Washington, DC 20502

ATTN: **Big Data Study – Comments of Common Sense Media**

Dear Ms. Wong,

Common Sense Media, a nonpartisan, nonprofit organization dedicated to helping kids, families, and educators thrive in a world of media and technology, respectfully submits these comments in response to the Office of Science and Technology Policy's Request for Information in the comprehensive review of issues at the intersection of "big data" and privacy. As the White House frames the key questions that the collection, analysis and use of "big data" raise for our government and nation, **we urge special consideration for the privacy interests of children and teens in both the consumer and education sectors. Today's youth are prolific users of the Internet and uniquely at risk when it comes to "big data."**

I. Introduction and Overview

Today's kids are the first generation to live their entire lives online, creating a vast digital footprint that can track them throughout their lives. The explosive growth of digital devices, smart phones, and social media is literally transforming the lives of these "digital natives" at home, at school, and in between. Even the youngest of our children are migrating online, using smart phones and tablets, downloading apps and content. These new high-tech tools bring a wonderful potential for our children to learn, communicate, and create – as well as the potential to amass a huge collection of personally identifiable information about young people that can be tracked, mined, and exploited by unintended audiences with surprising consequences.

The clear trends for young people to share more information online, to use more mobile technology, and for more technology to be integrated in schools, underscore the need for special protections for children, teens, and students. When it comes to big data and young people, we must ensure that their sensitive personal information is safeguarded, so they can enjoy innovative and engaging content and applications without surrendering their privacy.

Common Sense Media calls for measures to bolster existing protections for children and to empower teens with better choice and control over their online information. Specifically, all young users should be able to use an "Eraser Button" to delete information they post online. Teens should be able to "opt in" to online collection of their personal and geolocation information, and to "opt in" before their personal information, location, and online activity is shared with third parties for uses such as profiling or behavioral marketing.

Common Sense Media also calls for special protections for students, to help ensure that: students' sensitive personal information is used only for educational purposes; students' sensitive personal information and online activity is not used to target advertising to students or families; and appropriate data security, retention, and destruction policies are adopted for students' data.

We want to create a trusted environment where kids can use technology at home, at school, and on the go; where they can harness the power of the Internet for learning, entertainment, communication and collaboration; and where they can find their voice, share with their friends, and explore the world without fear of commercial exploitation or other unintended consequences.

II. Technological Trends Involving Children, Teens, and Students

A. Today's children and teens are heavy tech users, increasingly sharing personal information online.

Our children and teens are growing up in a digital world, surrounded by online and mobile technology. They have unprecedented access to digital products, create and consume enormous amounts of content, and can connect with people and information around the world. While the Internet presents tremendous opportunities for entertainment, innovation, and learning, this digital interaction also raises concerns about kids' and teens' online privacy and the creation of a huge digital footprint.

1. Young people are prolific online users.

Almost a quarter of children begin their digital lives before they are even born and, by the age of two, 92 percent of American children have an online presence.[1] By age five, about half of children go online daily. And, by age eight, more than two-thirds of children use the Internet on any given weekday.[2]

Teens are especially avid daily users of digital technology and social media. Virtually all teenagers (95 percent) use the Internet.[3] Moreover, our kids live in a culture of sharing. Social media has become part of everyday life for many teens and a key communication tool. Our recent survey of teens found that 90 percent of 13- to 17-year-olds have used some form of social media, and 75 percent of teenagers have a profile on a social networking site.[4]

---

[1] Business Wire Press Release, *Digital Birth: Welcome to the Online World – AVG Study Finds a Quarter of Children Have Online Births Before Their Actual Birth Dates* (Oct. 6, 2010), http://www.businesswire.com/news/home/20101006006722/en/Digital-Birth-Online-World

[2] The Joan Ganz Cooney Center at Sesame Workshop, *Always Connected: The New Digital Media Habits of Young Children*, at 16 (Mar. 10, 2011), http://www.joanganzcooneycenter.org/wp-content/uploads/2011/03/jgcc_alwaysconnected.pdf.

[3] Pew Research Center & Berkman Center for Internet and Society, *Teens and Technology 2013*, at 3 (Mar. 13, 2013), http://www.pewinternet.org/~/media//Files/Reports/2013/PIP_TeensandTechnology2013.pdf.

[4] Common Sense Media, *Social Media, Social Life: How Teens View Their Digital Lives* at 9 (Summer 2012), http://www.commonsensemedia.org/sites/default/files/research/socialmediasociallife-final-061812.pdf

Notably, teens are increasingly sharing personal information on social media sites:

- 92 percent posted their real name to the profile they use most often;
- 91 percent posted a photo of themselves;
- 84 percent post their interests, such as movies, music or books they like;
- 82 percent post their birth date; and
- 71 percent post their school name and the city or town where they live.[5]

2. Young people are uniquely at risk online.

As children and teens surf the Internet, a host of cookies and other technologies tracks their movements and builds a profile of their online activities. Significantly, these youth have been tracked and targeted more than adults in online spheres. A 2010 *Wall Street Journal* investigation found that the top 50 websites popular with U.S. children and teens installed 30 percent more tracking technology on personal computers than top websites aimed at adults.[6]

When it comes to social media, although teens frequently share their personal details, some teens may not understand whether and how the information they share is being used by the social media site and by third parties. For instance, recent focus groups suggest that teen Facebook users did not believe that the company would give anyone else access to the information they share.[7]

Other research shows that even when Facebook users think they are using privacy controls to limit the information they are sharing "publicly," they are increasingly revealing information "privately" and fail to appreciate that such "private" disclosures are available to Facebook, third-party apps, and Facebook advertisers. [8]

Moreover, whatever users text or post can be searched, copied, shared, and analyzed by vast social networks of friends and followers – and by vast commercial networks of advertisers, analytics services, and data brokers. This information can be used in unintended ways. According to a 2013 Kaplan survey, for example, 30 percent of college admissions officers discovered something on social media that negatively impacted the applicant's chance of getting into the school.[9] In addition, some lending companies are mining social media to determine creditworthiness and make lending decisions.[10] Conceivably, a social media post about a youthful indiscretion could come back to haunt a young borrower.

---

[5] Pew Research Center & Berkman Center for Internet & Society, *Teens, Social Media, and Privacy*, at 3 (May 21, 2013), http://www.pewinternet.org/files/2013/05/PIP_TeensSocialMediaandPrivacy_PDF.pdf.
[6] Steve Stecklow, *On the Web, Children Face Intensive Tracking*, WALL ST. J. (Sept. 17, 2010).
[7] *Teens, Social Media, and Privacy*, supra note 5, at 10.
[8] Stutzman, Gross & Acquisti, *Silent Listeners: The Evolution of Privacy and Disclosure on Facebook*, J. PRIVACY & CONFIDENTIALITY, Vol. 4: Iss. 2, Article 2 (2013), http://repository.cmu.edu/jpc/vol4/iss2/2.
[9] Press Release, *Kaplan Test Prep Survey: More College Admissions Officers Checking Applicants' Digital Trails, But Most Students Unconcerned* (Oct. 31, 2013), http://press.kaptest.com/press-releases/kaplan-test-prep-survey-more-college-admissions-officers-checking-applicants-digital-trails-but-most-students-unconcerned
[10] Stephanie Armour, *Borrowers Hit Social-Media Hurdles*, WALL ST. J. (Jan. 8, 2014).

B. The growth in mobile technology is especially significant for kids and teens.

The clear trend in recent years has been the huge shift to mobile technology. Not necessarily as evident is how quickly children are starting to use mobile devices at very young ages, how heavily teens have come to rely on mobile devices for Internet access, and how pervasively kids' apps have tracked young users.

1. Young children and teens are early adopters of mobile devices.

Common Sense Media's recent Research Report, *Zero to Eight, Children's Media Use in 2013*, found that almost twice as many young children are using mobile media now as there were just two years ago, and the average amount of time children spend using mobile devices has tripled. Seventy-two percent of children 0 to 8 have used a mobile device for some type of media activity, such as playing games, watching videos, or using apps. In fact, 38 percent of toddlers *under age two* have used a mobile device in the past year.[11]

Teens especially are increasingly connected whenever and wherever they go – more so than adults:

- About three in four (74 percent) teens ages 12-17 say they access the Internet on cell phones, tablets, and other mobile devices at least occasionally;
- One in four teens are "cell-mostly" Internet users — far more than the 15 percent of adults who are cell-mostly;
- Among teen smartphone owners, half are cell-mostly Internet users.[12]

2. Mobile technology presents special risks, particularly for young users.

The explosive growth of mobile technology poses particular challenges because these devices can be used almost anytime, anywhere, providing a ready platform for users to share a nearly constant stream of personal information. The proliferation of apps can access a tremendous amount of personal data and usage information from mobile devices – including their precise geolocation. This sensitive personal information can be shared with advertisers, analytics companies, data brokers and other third parties, often without the user's knowledge or express consent. Even with privacy policies and on-screen permissions, it's not always easy for users to determine what information an app will access, how it will be used, or with whom it will be shared.

Research has shown that mobile apps for kids have been rampant leakers of personal information. A December 2012 Federal Trade Commission Report found that many of the kids' apps shared sensitive information with third parties -- without disclosing that fact to parents. Nearly 60 percent (235) of the apps reviewed transmitted device ID to the developer or, more commonly, an advertising network, analytics company, or other third party. Three percent

[11] Common Sense Media, *Zero to Eight: Children's Media Use in America 2013*, at 11 (Oct. 28, 2013), https://www.commonsensemedia.org/file/zero-to-eight-2013pdf-0/download.

[12] Pew Research Center & Berkman Center for Internet and Society, *Teens and Technology 2013* (Mar. 13, 2013), at http://www.pewinternet.org/~/media//Files/Reports/2013/PIP_TeensandTechnology2013.pdf

(12) of the apps also transmitted the user's geolocation and one percent (3) transmitted the device's phone number. This is concerning because in every instance where an app transmitted geolocation or phone number, it also transmitted the user's device ID, so that third parties could potentially connect the location and phone information with any data previously collected through other apps running on the same device. These apps have been downloaded hundreds of thousands of times.[13]

Similarly, a June 2013 *Wall Street Journal* examination of 40 popular and free child-friendly apps on Google's Android and Apple's iOS systems found that nearly half transmitted to other companies a device ID number, a primary tool for tracking users from app to app, and 70 percent transmitted information about how the app was used.[14]

Although some tracking of young children may be curbed by recent amendments to the COPPA regulations, noted below, the opportunity for tracking, profiling, use and abuse of young people's sensitive information is vast.

C. Increasing use of technology in schools is spawning a proliferation of digital student data.

Our nation's schools are increasingly integrating computers, laptops and tablets in the classroom, and relying on cloud computing services for a variety of academic and administrative functions. This technology, used wisely, has the vast potential to enhance and personalize student learning and to improve school efficiency. To fulfill this potential, we must ensure that students' sensitive personal information is safeguarded.

Through online platforms, mobile applications, digital courseware, virtual forums for interacting with other students and teachers, and cloud computing services, schools and education technology providers collect massive amounts of sensitive informatino about students. This student data includes: school work and academic performance data, online searches, contact information, health information, behavior and disciplinary records, eligibility for free or reduced-price meals – even cafeteria selections and whether or not students ride the bus to school. This information is at risk.

Some online services collect and analyze personal details about students without clear limits on use of the student data for educational purposes.[15] Other online services have failed to adequately secure and encrypt students' personal information from potential misuse.[16] In fact, a recent study by Fordham Law School's Center on Law and Information Policy found that the majority of school district cloud service agreements have serious deficiencies in the protection of student information, "generally do not provide for data security and even allow vendors with

---

[13] FTC, *Mobile Apps for Kids: Disclosures Still Not Making the Grade*, at 10-11 (Dec. 2012).
[14] Jeremy Singer-Vine and Anton Troianovski, *How Kid Apps Are Data Magnets*, WALL ST. J. (June 27, 2013).
[15] *See, e.g.*, Benjamin Herold, *Google Under Fire for Data-Mining Student Email Messages*, Education Week (Mar. 13, 2014), http://www.edweek.org/ew/articles/2014/03/13/26google.h33.html?cmp=ENL-EU-NEWS2.
[16] Natasha Singer, *Data Security Is A Classroom Worry, Too*, New York Times (June 22, 2013).

alarming frequency to retain student information in perpetuity."[17]

Concerns about the privacy and security of this sensitive student information, if not addressed up front, can quash innovation in the education sector. Students shouldn't have to surrender their right to privacy and security at the schoolhouse door. We need clear rules of the road to ensure that schoolchildren's information is not exploited for commercial purposes and stays out of the wrong hands.

III. The Administration Should Ensure That Personal Data of Children, Teens, and Students Is Protected.

As the Administration considers the implications of "big data," we urge that data collected from and about children, teens, and students be given special consideration and heightened protection. All data should not be considered equal. As the Administration recognized in its 2012 Consumer Privacy Bill of Rights, personal data obtained from children and teenagers may require greater protections than data for adults.[18] Indeed, personal information about children and teens is a sensitive data category, like financial and health information, that warrants special protection. Likewise, the Federal Trade Commission's 2012 Privacy Report recognized the general consensus that children's information is a sensitive data category (like financial and health information, and precise geolocation data), and that companies should obtain affirmative express consent from consumers before collecting such data. The FTC further recognized that information about teens is sensitive, warranting consideration of additional protections, as well. [19]

With this in mind, we encourage the Administration to support the following measures:

A. Online protections for minors and the "Eraser Button."

The Children's Online Privacy Protection Act ("COPPA"), enacted in 1998, requires sites, services and apps that are directed to children under 13 or have actual knowledge that the user is under 13 to obtain parental consent before collecting personal information. The FTC recently updated its COPPA Rule to expand the type of personal information that requires parental consent to include geolocation, photos, videos, audio, and persistent identifiers, to account for the explosive growth of social media and mobile technologies.[20]

---

[17] Press Release, *Fordham Law National Study Finds Public School Use of Cloud Computing Services Causes Data Privacy Problems* (Dec. 13, 2013), http://law.fordham.edu/32158.htm; Natasha Singer, *Schools Use Web Tools, and Data is Seen at Risk*, New York Times (Dec. 12, 2013).

[18] *Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy*, at 15 (Feb. 2012).

[19] FTC, *Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Businesses and Policymakers*, at 59-60 (Mar. 2012).

[20] *See* News Release, FTC Strengthens Kids' Privacy, Gives Parents Greater Control Over Their Information By Amending Children's Online Privacy Protection Rule (Dec. 19, 2012), http://www.ftc.gov/news-events/press-releases/2012/12/ftc-strengthens-kids-privacy-gives-parents-greater-control-over.

To enhance control of online information for both children and teens, Common Sense Media has supported requiring Internet companies to provide an "Eraser Button" that would permit minors to remove content or information that they personally posted on websites, online services, and online or mobile apps. Too often, young people post information they later regret but can't delete from the online and mobile world. Children and teens often self-reveal before they self-reflect and may post sensitive personal information about themselves -- and about others -- without realizing such sharing may affect potential college or job opportunities – or even increase the risk of identity theft. All of us -- especially kids -- should be able to delete what we post.

According to a recent Pew poll, pruning and revising social media profile content is an important part of teens' online identity management: 59 percent have deleted or edited something that they posted in the past. Moreover, 19 percent of teens have posted updates, comments, photos, or videos that they later regretted sharing.[21]

In September, 2013, California enacted SB 568, *Privacy Rights for California Minors*. This new "Eraser Button" law requires websites and apps to permit users under 18 to remove content they posted on Internet and social media sites.[22] Several additional states are now considering similar measures. The FTC also has supported the "eraser button" concept, noting that it is consistent with the principles of data access and deletion.[23]

Many companies already provide deletion functions for their users. Although not a panacea for information that has already been re-posted or shared by third parties, an eraser button would provide needed control for online profiles.

B. Additional online protections for teens and an "opt in" standard for collection and use of their sensitive information.

While many current online protections are geared to young children, teenagers also should be accorded special protections online. Teens are heavy users of the Internet and mobile technology, tech-savvy in some ways, yet still requiring training wheels in others. As the FTC has explained, more privacy-protective default settings for teens "can function as an effective 'speed bump' for this audience and, at the same time, provide an opportunity to better educate teens about the consequences of sharing personal information."[24]

Too many online and mobile companies launch services and access users' information automatically, sometimes giving them the opportunity to opt out afterwards. This can mean that a user's personal information is collected and used before the user understands how the service works. Especially when teens are targeted and sensitive information is involved, they should be

---

[21] *Teens, Social Media, and Privacy*, supra note 5, at 9.

[22] Press Release, *Governor Signs Steinberg Bill Protecting Minors' Privacy on the Internet* (Sept. 23, 2013), http://sd06.senate.ca.gov/news/2013-09-23-governor-signs-steinberg-bill-protecting-minors-privacy-internet

[23] FTC, *Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Businesses and Policymakers*, at 70 (March 2012).

[24] Id. at 60.

given the opportunity to opt in.

We recognize that as teens mature it may not always be effective or appropriate to seek parental consent for their online conduct. That said, teens deserve enhanced online protection that would provide them with more transparency, notice, choice and control in the collection and use of their personal information.

Accordingly, Common Sense Media supports measures to require:

- "Opt In" for Collection of Personal Information: Online companies should obtain the teen's express affirmative consent before collecting personal or location information;
- "Opt In" for Behavioral Marketing: Online companies should obtain the teen's express affirmative consent before sharing personal information, location, or online activity for third-party profiling or behavioral marketing; and
- Fair information practice principles: Online companies should establish policies for transparency, collection, use, disclosure, retention, and security of teens' personal information.

These protections are included in the bipartisan, bicameral Do Not Track Kids Act of 2013, which would provide important baseline protections for young teens.[25] The bill would amend COPPA to bolster protection for children under 13 and create new online and mobile privacy protections for teens who use websites/services directed to teens aged 13 to 15 (or that *know* they are collecting personal info from teens 13 to 15). The bill also includes a "Digital Marketing Bill of Rights for Teens" that establishes fair information practice principles and would require online companies to explain the types of personal information collected, how that information is used and disclosed, and the policies for collection of personal information.

As the Administration works with Congress to pass legislation that codifies the Consumer Privacy Bill of Rights into law, we respectfully urge the inclusion of similar measures to provided enhanced protection for teens.

The public overwhelmingly supports such proposals to rein in corporate surveillance, according to a survey recently conducted by Anzalone Liszt Grove Research. Online tracking of children and teens is especially concerning, with 89 percent of Americans stating that it is extremely or very important to keep personal information about our kids private from corporate tracking. In addition, the vast majority (78 percent) support requiring companies to get permission from young teens aged 13 to 15 before collecting any personal information about them or sending them targeted ads based on their online activities.[26]

[25] S. 1700 and H.R. 3481, introduced in November, 2013, by Senator Edward J. Markey and Rep. Joe Barton.

[26] Memo from Anzalone Liszt Grover Research, Americans Concerned about Privacy from Corporate and Government Surveillance (Mar. 31, 2014), http://media.wix.com/ugd/c4876a_e8f4ee3b207344d9aac9a3403118ca9c.pdf.

C. Special privacy and data security protections for students.

It has long been recognized that student data deserves protection. New cloud computing technologies and the proliferation of student data in digital form heighten concerns that existing regulatory frameworks are inadequate. The Family Educational Rights and Privacy Act of 1974 (FERPA),[27] designed in the era of primarily paper records, gives parents the right to access their children's education records and generally requires written permission from the parent before these records are released to third parties. FERPA, however, is extremely complex and there are many significant gaps in coverage. Other statutes, such as the Protection of Pupil Rights Amendment (PPRA) and COPPA, also fail to cover large swaths of sensitive student data.

To pave the way for new personalized digital learning tools, online assessments, and other interactive technologies that help foster and enhance the learning process, the privacy and security of students' personal data must be addressed.

As Secretary of Education Arne Duncan stated, "Put plainly, student data must be secure, and treated as precious, no matter where it's stored. It is not a commodity."[28]

Common Sense Media has proposed three basic principles that attempt to balance the tremendous opportunity provided by education technology with the need to foster a trusted learning environment, where students' personal information is protected:

1. Students' personal information should be used solely for educational purposes;
2. Students' personal information or online activity should not be used to target advertising to students or families; and
3. Schools and education technology providers should adopt appropriate data security, retention, and destruction policies.

There is overwhelming public support for the implementation of such policies and regulations to protect students' private information. A national survey conducted earlier this year on behalf of Common Sense Media found that 90 percent of adults – whether parents or not – are concerned about how non-educational interests are able to access and use students' personal information. Eighty-six percent of Americans agree that protecting children's safety and personal information should be the No. 1 priority, while only 11 percent believe the argument that regulations would be overly burdensome and stifle innovation. [29]

---

[27] 20 U.S.C. § 1232g; 34 C.F.R. Part 99. The Protection of Pupil Rights Amendment (PPRA) also provides parents with the opportunity to opt-out of certain surveys and activities in schools when personal information is collected from the student for marketing purposes. PPRA, however, permits school districts to use personal information that they collect from students to develop, evaluate or provide educational products or services for students or schools. 20 U.S.C. § 1232h; 34 CFR Part 98.

[28] U.S. Dept. of Education, *Technology in Education: Privacy and Progress - Remarks of U.S. Secretary of Education Arne Duncan at the Common Sense Media School Privacy Zone Conference* (Feb. 24, 2014), https://www.ed.gov/news/speeches/technology-education-privacy-and-progress

[29] Press Release, *National Poll Commissioned by Common Sense Media Reveals Deep Concern for How Students' Personal Information Is Collected, Used, and Shared: Americans Overwhelmingly Support*

IV. Conclusion

The extraordinary technological changes and the development of new social, mobile, and educational platforms in recent years have created new and immersive environments for young people, resulting in a proliferation of sensitive digital data about them. We applaud the Administration for reviewing the implications of big data – and urge special attention to children, teens, and students in this review process and the development of public policy.

We look forward to working with the Administration as it continues this study, so we can fashion appropriate measures to safeguard the privacy and security of our children's personal data, and help ensure that the Internet remains a robust platform for education, innovation, and economic growth.

Respectfully submitted,

James P. Steyer
CEO and Founder
Common Sense Media

Reforms to Protect Students, Including Increased Transparency, Tighter Security Standards, and More Restrictions on Companies and Cloud Services (Jan. 22, 2014), at http://www.commonsensemedia.org/about-us/news/press-releases/national-poll-commissioned-by-common-sense-media-reveals-deep-concern.

Before the
White House Office of Science and Technology Policy
Washington, D.C.

In the matter of	)
Government Big Data	)
	)
	)

COMMENTS OF
THE COMPUTER AND COMMUNICATIONS INDUSTRY
ASSOCIATION

1 Introduction

In response to the Request for Information released by the Office of Science and Technology Policy (OSTP) on March 4, 2014,[1] the Computer & Communications Industry Association (CCIA) offers the following comments.

CCIA is an international, nonprofit association representing a broad cross section of computer, communications and Internet industry firms. CCIA remains dedicated, as it has for over 40 years, to promoting innovation and preserving full, fair and open competition throughout our industry. Our members employ more than 600,000 workers and generate annual revenues in excess of $200 billion.[2]

These comments will address four issues: 1) That big data is simply a tool and is not uniform in its implications. Its implications depend on who it is using it and for what; 2) Commercial use of data creates value and avoids harm when users are protected through transparency and control; 3) Many government uses of large data sets are vital for the operation of our country, and the Privacy Act protects the subjects of the data, and; 4) That big data analysis for surveillance purposes is highly risky because of the relative

[1] Government "Big Data"; Request for Information, 79 Fed. Reg. 12251 (Mar. 4, 2014).

[2] A complete list of CCIA members is available at http://www.ccianet.org/members/.

power of the state in law enforcement and national security contexts, and that therefore stringent restrictions and oversight is necessary.

2 Big data is simply a tool

The term "big data" has become a shorthand for an enormous variety of technologies and techniques for the gathering, manipulation, and analysis of large data sets. It is important in a policy examination of big data to realize that the term does not encompass a monolith of practices. Indeed, the subject pertains to a diverse set of tools, and in reality there is still considerable disagreement over what precisely the term refers to.[3]

As OSTP proceeds in examining the implications of big data, it will be important to keep this fact in mind. Questions or solutions that appear to apply to the entire waterfront of big data should be examined closely to make sure that they are not actually applicable to only a subset of "big data", or applicable to different areas in different ways.

In particular, government uses of big data are different from commercial ones for a variety of reasons. Solutions that protect data subjects while promoting innovation and product solving in one situation may be inapplicable or even harmful in another. To the extent that OSTP makes recommendations, they will be more helpful if this fact is kept in mind.

Even within the category of commercial uses of big data, there will be variation in what kinds of data is collected, whether the data is individually marked or aggregated, and the uses that the data is put to. In many cases, some of these factors will inherently protect users from real harm while in others further interaction will be necessary. In any case, OSTP should be conscious of these distinctions if attempting to make broad observations.

When approaching such a broad and complex subject, it is wise to approach the problem with an awareness of and sensitivity to the unknown. OSTP has already demonstrated considerable sensitivity to the as-yet undis-

[3] *See, e.g., The Big Data Conundrum: How To Define It*, MIT Technology Review, Oct. 3, 2013, http://www.technologyreview.com/view/519851/the-big-data-conundrum-how-to-define-it/.

covered technological innovations in the big data area by calling upon the public to advise them through this proceeding and other meetings. Applying that same awareness to any resulting recommendations will ensure that they are relevant and proportional to the issues they are addressing.

3 Commercial applications of big data bring value to users and offer proper controls

The commercial uses of "big data" are as numerous as there are business plans in the country. The basic analysis of the data produced by a company can be considered big data and can contribute to significant consumer benefits, increasing sales, greater efficiencies, streamlining supply chains, and an innumerable host of other applications. In the years to come more and more companies will be exploring the data they generate as a routine part of doing business. Many of these uses will have no privacy implications whatsoever.

There are a number of reasons why privacy may not be implicated. In situations where a company is simply analyzing data about their own operations, such as supply chain information, there is simply no personal information about consumers involved and thus no privacy impact. In other circumstances, a company may be working with data about individuals who are customers of the company, in order to improve services, and where the data never leaves the company itself. In this case the privacy implications are minimal. Aggregated or deidentified data may also pose considerably less of a privacy concern, as data on particular people is not available. This is even more true if the company commits to not attempting to reidentify the data in the course of their work with it.

There are also many cases where privacy interests are implicated. If data collected and analyzed is about identified individuals, and there is a potential for concrete harm to those individuals, then privacy concerns are at their apex. In those cases there are a few ways in which people can be protected from harm. First, there are some areas where the concrete harm can be so great that we have enacted laws to control the use of the data, and

our current legal framework is fully capable of providing remedies for these harms. For example, in situations where particular adverse decisions may be made against a person from some data, the Fair Credit Reporting Act controls how the data may be used and what rights the person has to see and correct information about them.[4] Secondly, where no law applies to the data collection and analysis, it is important that companies be transparent about their actions and give users control over how the data is used, including the opportunity to correct inaccurate data and the possibility of opting-out if they desire. In these circumstances there is always the backstop of the Federal Trade Commission, using its authority to make sure that companies follow the public pledges they make about privacy.

Commercial big data products are important because they make possible a huge variety of features and products that benefit consumers, and even entire business plans. Mobile mapping applications that warn about traffic do so by aggregating huge amounts of location data from phones traveling on the roads. Apple's voice recognition software only works from analyzing a large corpus of voice recordings and refines itself all the time based on the queries it receives from users. Self-driving cars, like those being developed by Google and other companies, rely on having detailed information on the roads that the car will be driving along. While not a perfect crystal ball, large-scale data analysis can also create insights that lead to the next great American company, not just in Silicon Valley, but around the country.

Two years ago the White House proposed a Consumer Privacy Bill of Rights, a framework that was intended to capture common privacy principles in a "comprehensive" way. When it released the Bill of Rights, the White House appropriately praised the strength of the U.S. privacy regime:

> The consumer data privacy framework in the United States is, in fact, strong. This framework rests on fundamental privacy values, flexible and adaptable common law protections and consumer protection statutes, Federal Trade Commission (FTC) enforcement, and policy development that involves a broad array

[4]Fair Credit Reporting Act, 15 U.S.C. §1681 et seq.

> of stakeholders. This framework has encouraged not only social and economic innovations based on the Internet but also vibrant discussions of how to protect privacy in a networked society involving civil society, industry, academia, and the government.[5]

The 2012 report aimed to address concerns raised by those who favored a single approach applying across all sectors by articulating "a clear statement of basic privacy principles" and catalyzing "a sustained commitment of all stakeholders to address consumer data privacy issues as they arise from advances in technologies and business models."

The privacy world has changed considerably since the Administration's 2012 Report, during which time the U.S. has "experimented with new methods of privacy policymaking in the form of NTIAs multistakeholder process. And in corporate America, a culture of privacy awareness has blossomed into robust privacy programs and a thriving market for privacy professionals."[6] In contrast, enforcement of "comprehensive" privacy regulation has been more stagnant and limited, and debate continues about how to update European privacy laws for the twenty-first century.

One final feature of big data for OSTP to keep in mind with regard to commercial uses is the way it can quickly change the landscape of a marketplace, including its own. The inexpensive availability of data analysis, extending to nearly all actors in a marketplace, can cause rapid shifts in business methods and products. In the face of such a quickly moving set of circumstances, regulators must be careful when trying to address issues. Any government-imposed solution risks being outdated before it is even implemented. Multistakeholder processes that involve consumer groups, industry, and government working together are better equipped to tackle privacy issues arising from big data in flexible but privacy protecting ways. Legislation does not easily so adapt to the ever-evolving nature of commercial

[5] The White House, Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy i (2012).

[6] *See* Kenneth A. Bamberger & Deirdre K. Mulligan, *Privacy in Europe: Initial Data on Governance Choices and Corporate Practices*, 81 Geo. Wash. L. Rev. 1529, 1564-65 (2013).

technologies in the same way.

4 Government is beginning to explore big data

There are a number of areas in which government is also beginning to see how big data can make the business of running government more effective and efficient, as well as opening up services that previously would have been impossible. While the techniques of analysis that come with the big data moniker are new to government, the government has always been a major collector of data, at large scale, and of highly personal nature. Of course government must therefore focus on the privacy of people whose data is being analyzed, and the Privacy Act provides the rules for how that happens.

The ways in which the government is using big data are as varied as the government agencies themselves. The Library of Congress is cataloging every public tweet ever published for future analysis or even just posterity. The Center for Disease Control employs statistics and big data to track flu outbreaks.[7] The various statistical agencies, such as the Bureau of Justice Statistics and the Bureau of Labor Statistics, have begun releasing data in machine readable formats, allowing them to be combined with other data sets for greater depth of analysis.[8] All of these efforts show how the power of big data is influencing how government operates.

The privacy of data subjects when the government collects and analyzes large amounts of data is no less important just because the government is the one doing the collecting. Indeed, in some ways, it is more important, as residents of a country have little opportunity to simply select a different government to live under, while users of online services can usually easily pick and choose a service that has privacy options they prefer. Competition for users in the online world provides a form of "market discipline" on companies and how they use data. Because governments are not subject to the same changes in market dynamics, they are not as responsive to privacy

[7] Note that CDC's flu tracker is separate from Google's and studies have shown that the most accurate representation of real-world flu comes from combining the two approaches. *See* Steve Lohr, *Google Flu Trends: The Limits of Big Data*, N.Y. Times, Mar. 28, 2014.

[8] *See generally*, Bureau of Justice Statistics, http://www.bjs.gov.

problems and their feedback mechanisms break easily. For these reasons, protections on government use and analysis are important.

The law that controls how the government treats the data it collects and analyses is the Privacy Act of 1974.[9] The Privacy Act was one of the first attempts to deal with the idea of large computer databases, operated by the government, collecting information about citizens. The act grew out of a process at the Department of Health, Education, and Welfare that codified what are now known as Fair Information Practice Principles (FIPPs). Those principles were adapted into the Privacy Act and today control how the government treats data.

The Privacy Act, however, has not been substantially modified since its passage in 1974, nor has the guidance issued by the Office of Management and Budget in 1975. The intervening 40 years, on the other hand, have seen an explosion of new technology and new applications of that technology to data analysis. For example, the Privacy Act's central definition, a "system of records," is hard to accurately apply in a modern age where many interlocking databases may be maintained, sometimes by different agencies.[10] Similarly, the "routine use" exception to the Privacy Act is poorly defined and in practice operates as a huge loophole for government agencies to share personal information, and could be a focal point for reform.

There have been occasional efforts to update the law, but none have succeeded. OSTP's exploration of the implications of big data should focus on whether the Privacy Act is adequate to address the federal government's data usage today. If the White House recommends upgrades to the Privacy Act, it could catalyze what has been a stalled conversation about the direction of the law in the 21st Century.

[9] Privacy Act, 5 U.S.C. §522a et. seq.

[10] *See, e.g., Veterans Data Breach Highlights Inadequate Privacy Protections*, Center for Democracy and Technology, May 31, 2006, available at https://www.cdt.org/policy/veterans-data-breach-highlights-inadequate-privacy-protections.

5 The use of big data for surveillance is inherently problematic

While normal government use of big data can be troublesome in some contexts (making the Privacy Act an important bulwark against abuse), big data for surveillance purposes, either in a criminal context or for national security, brings in a host of problems. One essential solution available today is for Congress to pass and the President to sign a bill that updates the Electronic Communications Privacy Act of 1986 (ECPA) to ensure a warrant for content standard.[11] In the national security context, there are a series of much-needed reforms to how the government treats the vast amounts of information it collects in the name of security. In particular, the recent decisions with regard to bulk collection of metadata are a good start, but there are still problems to be addressed.

Data collected under surveillance regimes are different from all the other forms of data. Out of all of these categories, the potential harms that can come from abuse are greatest from government surveillance. That is why we have historically had such strong controls on the government's ability to gather information for the purposes of criminal investigation. Those controls were enshrined in the Fourth Amendment to our Constitution and have been, for most of the history of the Republic, the primary protection against abuse by the government.

Today that protection is no longer adequate. More and more of our day to day lives are now spent online. Communication, work, play, and creative pursuits all have moved to the Internet in some form or another. The law, sadly, has not kept up. Beginning with the Third Party Doctrine created by the Supreme Court in Smith v. Maryland in 1979, a distinction has been made between the lives we live offline and the lives we live online.[12] Paper mail is protected while email is not. Files in a file cabinet are protected while files in cloud storage are not. For most Americans today, however, these distinctions are meaningless.

[11] Electronic Communications Privacy Act, 18 U.S.C. §2510 et. seq.
[12] Smith v. Maryland, 422 U.S. 735 (1979).

Fortunately, the solution already exists. A clean bipartisan and bicameral bill exists that would fix the warrant for content problem in a narrow and targeted way. The Senate version, S. 607 (written by Senator Leahy, the original author of ECPA), has already passed the Senate Judiciary Committee on a voice vote, and the House version, H.R. 1852, now has almost 200 co-sponsors from both sides of the aisle.[13] The White House could help make this easy fix a reality, improving Americans' privacy, giving certainty to companies doing business online, and showing the international community that responsible control over government surveillance is a priority of the U.S. government.

In the national security arena, there is still a lot that the White House should do to bring the National Security Agency (NSA)'s practices into line with principles that will protect people around the world and encourage trust in the online marketplace. CCIA supports the Reform Government Surveillance principles promulgated by many of the large tech firms.[14] These principles deal with the concept of big data directly and should serve as guideposts for the administration as it considers how it will move forward with its national security work.

The first principle, "Limiting Governments' Authority to Collect Users' Information," calls for governments to target surveillance at specific known users and only for lawful purposes. By limiting collection to this category, this principle seeks to take national security surveillance out of the category of big data entirely. Indiscriminate gathering of information about the public, in order to sift it for possible indications of wrongdoing, is not compatible with this idea.

With regard to particularity of government data collection, the President's recent announcement of restrictions on the bulk collection of telephone metadata is very welcome, however there are two remaining issues that the government should address. First, the announcement applies to telephone metadata only at the moment. While it appears as if the NSA

[13] Electronic Communications Privacy Act Amendments Act, S. 607, 113th Cong. (2013); Email Privacy Act, H.R. 1852, 113th Cong. (2013).

[14] *See* Reform Government Surveillance, https://www.reformgovernmentsurveillance.com/.

is not currently gathering metadata from online transactions,[15] under the plan proposed by the President it could resume at any time. The new rule should apply to the bulk collection of any metadata, no matter where it lives. Secondly, the new rule still maintains the troubling problem of permitting the NSA to chain "hops" of people who communicate together to reach non-targets. This is problematic particularly when one of the hops is a phone number that a large number of people call, such as a pizza delivery number or a government service such as a Department of Motor Vehicles.

The second and third principles cover the equally important question of what happens surrounding the collection of data. Governments should place their collection apparatuses under proper oversight and must hold accountable the groups doing the collection.

The fifth principle, "Avoiding Conflicts Among Governments," also pertains to big data in surveillance. Mutual Legal Assistance Treaties (MLATs) are structures by which the investigating authorities in one country can obtain information from companies located in another country. The government of the requesting country makes the demand of the government where the company is located and the second government then makes the demand of the company. This process, when it works, gets investigators the information they need, while at the same time forcing them to follow proper channels (rather than simply attempting to intimidate any local staff of the target company into turning over information on risk of imprisonment).

The MLAT process, however, almost never works as it should. The process is convoluted and inefficient. Requests can take up to 18 months to process in some cases.[16] Fortunately, the administration has signalled its intention to fix it. The President in January announced his intention to reform the MLAT process and the Department of Justice has asked Congress

[15]Glenn Greenwald & Spencer Ackerman, *NSA collected US email records in bulk for two years under Obama*, The Guardian, June 27, 2013, available at http://www.theguardian.com/world/2013/jun/27/nsa-data-mining-authorised-obama.

[16]The President's Review Group on Intelligence and Communications Technologies reported an average of 10 months backlog and some cases that go back substantially further than that. *See*, *e.g.*, The President's Review Group on Intelligence and Communications Technologies, Liberty and Security in a Changing World (2013) at 227.

for a larger budget to help hire more lawyers and staff to deal with the MLAT backlog.[17]

While those are good first steps, there are still other things that the U.S. government could be doing to improve the process. As in most other areas of surveillance, transparency about the MLAT process should be a priority. Only through knowing how the current process is working can we effectively attempt to fix it. Transparency will also show where the most requests come from and which countries are using the process, and help therefore highlight where the most effective reforms might take place.

6 Conclusion

The White House, in exploring the issues that are brought about by big data, should keep in mind that uses of big data may take many forms and have many different implications. Attempting to address all of them in the same way will be ineffective. In particular, OSTP should be conscious of the different categories outlined in these comments and think about how those categories pose different issues. Finally, because of the immense harm that can arise, OSTP should be most critical of the use of big data for surveillance purposes.

CCIA thanks OSTP for the opportunity to comment on this important matter, and would welcome the opportunity to answer any questions or make any clarifications that are requested.

[17] See Press Release, U.S. Department of Justice, Attorney General Holder Announces President Obama's Budget Proposes $173 Million for Criminal Justice Reform (Mar. 4, 2014), available at http://www.justice.gov/opa/pr/2014/March/14-ag-224.html.

OSTP Response – Big Data

Dr. Gregory Hager
Professor and Chair, Department of Computer Science, Johns Hopkins University
Vice Chair, Computing Community Consortium

March 29, 2014

(2) What types of uses of big data could measurably improve outcomes or productivity with further government action, funding, or research? What types of uses of big data raise the most public policy concerns? Are there specific sectors or types of uses that should receive more government and/or public attention?

It is clear that many sectors of the national and international economy are already acquiring and exploiting "big data" for commercial gain. Google, Facebook, Twitter and others are creating "the information economy," that is built on new platforms for information search and social interaction. The enormous commercial leverage to be gained by these enterprises will continue to drive new innovations in these sectors, while raising challenging issues regarding the ethics and regulation of the acquisition and use of personal data.

We assert that big data will play an equal or larger role in the advance of scientific discovery and related technological innovations. The evidence for this is already apparent. *Physics* has evolved from the traditional model of theory and experimentation to the use of *data-intensive computing* to identify physical phenomena, to discover and classify galaxies, and to analyze the results of large-scale simulation. *Biology* has been revolutionized by the enormous advances in individual genome sequence assembly and analysis which, in just a decade, has moved from a theoretical possibility to a practical reality. We are seeing the early signs of similar revolutions in medicine, chemistry, neuroscience, cell biology, economics, sociology, materials science … practically every branch of pure and applied science is being revolutionized by data. Why? Because computing and big data lets us couple human inspiration with computational "perspiration" – creating possibilities for discovery that would not otherwise be possible.

The challenge is thus to match the rapid growth in data acquisition and analytics in the private sector with comparable advances in data-intensive computing research. These advances will, in turn, support discoveries and innovation that will advance our long-term economic competitiveness and support innovations for the public good. Our specific recommendations are:

1) Strongly support the creation of platforms, tools, and best practices for the acquisition and sharing of data in support of basic and applied science. This must involve a multi-pronged strategy: support for basic and applied research in computing and data science to develop platforms and tools, together with the creation of best practices and incentives to deploy and use common platforms and tools into the broader research community.

2) Strongly support research in computer science, statistics, and mathematics, with a particular eye toward encouraging partnerships in the development of new data

analytic methods that will have long-term impact across multiple domains. Couple this with support for the creation of broadly-based computer science and data science education programs that will build capacity and community and the application of big data across the entire spectrum of big data applications.

3) Strongly support research in privacy, data provenance, and data anonymization so that scientific research that requires the use of information on individuals or populations of individuals can be carried out with maximal effectiveness and minimal risk of inadvertent disclosure of information.

4) Ensure regulations involving the acquisition and use of data are guided by ideas and knowledge so that they are enforceable and reasonable. In particular, technology evolves at a rapid pace – regulations that are implicitly or explicitly predicated on a particular technological frame of reference will be in danger of becoming irrelevant (thus losing their ability to provide enforcement) or counter to prevailing trends (thus *impeding* progress).

By taking these actions, the government will ensure that our nation is well-positioned to support data-enabled advances in science, in healthcare, in education, and in many other national priorities. It will also ensure our continued international leadership in science and technological innovation.

(3) What technological trends or key technologies will affect the collection, storage, analysis and use of big data? Are there particularly promising technologies or new practices for safeguarding privacy while enabling effective uses of big data?

Many of the key technologies affecting the acquisition and use of big data are already in place. Nearly all communications are computer-mediated and digital, thus making that data easily stored and accessed. The growth of cloud-based storage and computing means the computing infrastructure necessary to "mine" this data is growing and becoming cheaper. In short, data is now a commodity, and computing, and storage are all widely available utilities. Economic incentives continue to drive innovation and growth of both.

A *new trend* is the coupling of computing and information with devices – sometime referred to as the "internet of things." Early signs of this trend are already in place -- for example mobile phones, cars, televisions, phones, surgical systems, manufacturing systems, and agricultural equipment, are rapidly becoming equipped with sensors and automation and are partially or completely computer-controlled. This opens up entire new opportunities for the acquisition and exploitation of data on what people do, and how they do it. It also raises new concerns about privacy, safety, and appropriate use of data.

Protecting privacy within this framework is challenging and has no easy technical solution. Recent ideas, such as differential privacy[1] demonstrate that there are fundamental tradeoffs between the ability to draw statistical inferences from data, and the "leakage" of

---

[1] Dwork, Cynthia. "Differential privacy." *Automata, languages and programming*. Springer Berlin Heidelberg, 2006. 1-12.
Chaudhuri, Kamalika, and Daniel Hsu. "Sample Complexity Bounds for Differentially Private Learning." *Journal of Machine Learning Research-Proceedings Track* 19 (2011): 155-186.

information (i.e. what can be learned about an individual or group) from the data. There can be no free lunch. There is a growing threat of "predictive privacy" [2] whereby private information can be predicted from information not normally considered sensitive – e.g. predicting a customer is sick or pregnant based on past purchases and other demographic information.

We recommend the following actions related to the use of data in contexts where privacy is at risk due to inadvertent or intentional release of data:

1) Support research and development of methods that support the ability to trace an inference based on data back to the sources of data it relied on. Big data inference often involves the combination of multiple sources of data, leading to a conclusion that is then acted upon in some way – for example online behavior combined with locality data may be combined to place an ad. At one level, ensuring that data carries information on provenance ensures that the quality of an inference (vis-à-vis the data it relies on) can be computed. Conversely, it implies that it is possible for a user or a regulatory agency to determine the data sources were obtained appropriately for the designated purpose.

2) Support the development of frameworks for placing time, geographic, or sector-based limitations on data: Data should be constructed in a way that limits sharing as part of the data construct itself rather than through controls on an overall data set. For example, the European Data Protection Directive strictly limits the time that data on an individual can be held[3], and dictates sanctions in case those limitations are not met. As already noted above, having an intrinsic means of tracking the acquisition and use of data is an essential building block in enforcement.

3) Support the broader discussion of ways that technical means can be combined with social and economic means to achieve privacy. For example, many of the recent compromises of private information reported in the news are not a failure of technical approaches to security – they occurred due to lack of implementation and oversight. By making data sources explicit and placing limitations on sharing and use it becomes possible to concretely and automatically detect and render data "users" accountable for lapses and abuses of data placed in their care.

In summary, privacy fundamentally involves the ability of an individual and/or group to control sharing of data and, by extension, the uses of the data. Therefore, to ensure privacy, it must be possible for an individual or group to appeal to uniform standards that define and guard against inappropriate or unpermitted use of data. We caution that any such regulations should dictate ends, but not means, as those means (i.e. methods or technologies) are likely to change at the same speed as data and technology itself.

---

[2] Crawford, Kate, and Jason Schultz. "Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms." *Boston College Law Review* 55.1 (2014): 93.

[3] Heisenberg, Dorothee. *Negotiating privacy: The European Union, the United States, and personal data protection*. BoulderColorado: Lynne Rienner Publishers, 2005.

John Podesta, Senior Counselor to the President
Nicole Wong, Deputy Chief Technology Officer, OSTP
Big Data Study
Office of Science and Technology Policy
Eisenhower Executive Office Building
1650 Pennsylvania Ave., NW
Washington, DC, 20502

March 31, 2014

Via Email

Dear Mr. Podesta and Ms. Wong,

Consumer Watchdog, a nonpartisan, nonprofit public interest group, thanks you for the opportunity to offer comments on the White House "Big Data" policy review. On one hand we are pleased that you are delving into the issue, but on the other hand we are frustrated that these issues have been on the table without satisfactory resolution since the 1970s when the government developed the Fair Information Practice Principles. Indeed, two years ago President Obama called for the implementation of the Consumer Privacy Bill of Rights and called for baseline privacy legislation. Sadly, it has yet to be introduced in Congress.

The guiding principles for governing "Big Data" are straightforward: People must be able to know what information is gathered about them, how long it is kept and for what the information will be used. They should, in fact, have control over whether their data is even collected in the first place. People should be able to correct errors in data files about them and request the deletion of data not required to complete a business transaction they initiated. Large data sets used for research purposes should be aggregate data that has been de-identified.

Sadly, little of this is true currently. In the murky world of data brokers there is virtually no transparency. People don't know what digital dossiers have been assembled about them, what the data is used for or what decisions are being made about them without their knowledge.

We call on the Administration to introduce baseline privacy legislation and to implement the Consumer Privacy Bill of Rights. You must protect a person's right to control whether data about him or her is collected and how it is used. In other comments you will receive, we have joined a group of public interest and consumer groups in outlining six requirements your final report must address: Transparency, oversight, accountability, robust privacy techniques, meaningful evaluation and control. Thank you for your consideration.

Sincerely,

John M. Simpson
Privacy Project Director

2701 Ocean Park Blvd., Suite 112
Santa Monica, CA 90405
Tel: 310-392-0522 • Fax: 310-392-8874

EXPOSE. CONFRONT. CHANGE.
www.ConsumerWatchdog.org

413 E. Capitol St., SE, First Floor
Washington, D.C. 20003
Tel: 202-629-3064 • Fax: 202-629-3066

Dell Inc.
1225 I Street, NW
Suite 300
Washington, DC 20005
tel +1 202 408 3355
fax +1 202 408 7664
www.dell.com

Submission to the White House Office of Science and Technology Policy

Response to the Request for Information on Big Data

March 31, 2014

INTRODUCTION

Dell appreciates the opportunity to provide comments to the Office of Science and Technology Policy on Big Data. Over the past decade, extraordinary improvements in computing technology, global connectivity and analytical capabilities have enabled a powerful new phase in the IT revolution. The emergence of affordable storage, ubiquitous computing, data aggregation technologies and advanced analytics has opened the door to the data economy.

Our comments below focus on a few areas of technology necessary to bring about the benefits of the data economy and data innovation. In answering questions two (2) and three (3) of the OSTP request for information, we stress the importance of implementing strong security solutions to mitigate potential risks in the use of large data sets.

First, we mention two examples of the benefits the data economy can bring to society, highlighting the broader point that we should give due weight to the benefits in measuring against potential risks.

Critical services, such as disaster and emergency preparedness and response, can be enhanced by using data. Predictive modeling can better pinpoint the potential occurrence of events and mitigate the damage they inflict. For example, Dell's Digital Command Center at the American Red Cross headquarters has improved the sourcing and identification of trends in disaster-affected areas, significantly improving the ability of the Red Cross to anticipate and respond to the public's needs and enabling them to quickly connect people with resources during a disaster.

Medicine is critical to improving the quality of life. The ability to analyze huge genetic data sets is leading to new and more effective individualized treatments. In 2011, Dell began a partnership to apply technology to improve treatment of children diagnosed with neuroblastoma, one of the deadliest forms of pediatric cancer, deploying high-performance computing to allow our partner, T-Gen, to map the genomes of children in the trial more quickly – a process that once took 10 days now takes six hours. And through cloud technology, doctors can receive the information and share learnings with each other in real time, allowing treatments to be quickly individualized for each child.

Below we turn to practical technical areas that can increase the security of large data sets. Safeguarding privacy starts with security.

BIG DATA AND TECHNOLOGY TRENDS

(2) What types of uses of big data could measurably improve outcomes or productivity with further government action, funding, or research? What types of uses of big data raise the most public policy concerns? Are there specific sectors or types of uses that should receive more government and/or public attention?

In this response, we focus on practical technical areas that can increase the security of large data sets providing protection of data from unauthorized access through proper identity governance. A good security program starts with information about the data, information about how the data will be used and information about the user accessing the data. This information is then utilized to create controls that appropriately manage a user's access to data. This access is monitored, generating audit reports and regular attestation as to the need for the data access. Identity governance is the term associated with this type of model, which goes beyond simple identity and access management and includes data classification, controls and audit. Access to large data sets must similarly be appropriately controlled and monitored. Breaches of data security often result from an inappropriate identity governance model. Without appropriate identity controls, data innovation will not reach its full potential due to public concern with inappropriate data usage and security breaches.

Further, the area of automated and fine-grained data classification is in its infancy and could benefit greatly from additional government funded research. Granular encryption key management and encryption research can also help control and manage access to large data sets, including work in emerging encryption algorithms (e.g., homomorphic encryption). Continued funding and support for programs like the National Strategy for Trusted Identities in Cyberspace allow us to move away from classic user IDs and passwords to next generation identity services that enable true identity governance.

(3) What technological trends or key technologies will affect the collection, storage, analysis and use of big data? Are there particularly promising technologies or new practices for safeguarding privacy while enabling effective uses of big data?

A number of key technology trends are helping to improve the storage capacity, the speed, the level of correlation and the ability to run more sophisticated analytics. These are important and useful advancements, but the advanced technology must sit on top of a secure environment that allows access only by appropriate people with appropriate rights and intents. This only occurs if we move beyond simple identity and access management systems to a strong identity governance model that enables appropriate access to the data, is continuously audited, and requires attestation for access. And with granular data classification, users are limited to only the data that they are specifically authorized to use, and not to the entire data set. Advanced technology is being developed to measure the risk of allowing a user access to data at the time of authentication: technology that utilizes advanced biometrics to prove the validity of the user, technology that continuously authenticates the user, technology that proactively looks for anomalies in access patterns. Technology is also being developed to connect solutions at the

endpoint, network, server and storage layers to reduce risk and safeguard access. Safeguarding privacy starts with security, and requires this level of identity governance and access control.

CONCLUSION

As noted by the Information Technology Industry Council, the White House examination of Big Data should include sufficient study of the beneficial applications as well as the potential risks. Without understanding the benefits, it is impossible to understand the possible opportunity cost of risk mitigation strategies.

One area where the U.S. government will play a key role is in enabling cross-border data flows. The most beneficial analytics projects will be based on the largest data collections, containing data from many regions and countries. Data innovation magnifies the already challenging international environment where barriers to cross-border data flows impede the quality of the services that can be provided to individuals. As a result, the Administration should do all it can to support the U.S. – European Union Safe Harbor Framework, the Asia Pacific Economic Cooperation forum to promote APEC's Cross-Border Privacy Rules, the EU's Binding Corporate Rules regime and other similar mechanisms that facilitate cross-border transfers of personal data.

Finally, Dell is committed to protecting the personal data of our employees, customers and suppliers. Dell has robust privacy policies and online privacy statements describing how Dell uses and protects personal data. These policies are brought to life by Dell's Privacy Office, which includes our Chief Privacy Officer and lawyers around the globe who advise the company on local data privacy legal requirements.

Again, we look forward to being stakeholders and engaging with the White House, as well as other governments, as they ensure that the benefits of data innovation continue to be unleashed, that the right technologies are available and that proper policy frameworks are in place to do so. We are happy to discuss these comments further and explain how identity governance, access monitoring and data classification play a role in the security/privacy equation around data innovation.

* * *

If you have any questions about these comments, please contact Rebecca Karnak, Senior Manager, Global Policy, at 202-714-5668, Rebecca_Karnak@Dell.com.

Advancing and Protecting
Responsible Data-Driven Marketing

March 31, 2014

Submitted Via Email: bigdata@ostp.gov

Nicole Wong
Office of Science and Technology Policy
Eisenhower Executive Office Building
1650 Pennsylvania Avenue
Washington, DC 20504

RE: Office of Science and Technology Policy Request for Information on "Big Data"

Dear Ms. Wong:

On behalf of the Direct Marketing Association ("DMA"), we appreciate the opportunity to provide comments in response to the Office of Science and Technology Policy's ("OSTP") request for information ("RFI") on "big data" published on March 4, 2014.[1] The DMA is the world's largest trade association dedicated to advancing and protecting responsible data-driven marketing in the United States and globally.[2] Founded in 1917, DMA represents thousands of companies and nonprofit organizations that use and support responsible data-driven marketing practices and techniques. DMA provides data-driven marketers the voice to shape policy and public opinion, the connections to grow members' businesses, and the tools to ensure full compliance with responsible and best practices as well as professional development.

Marketers have engaged in the responsible collection and use of data for marketing purposes for more than 100 years. The recent advent of large data sets or "big data" has enhanced, but not changed, the basic role that marketing plays in the U.S. economy. Namely, the analysis of big data by marketers has made it easier to connect consumers with the products and services they desire. According to a recent study, the resulting Data-Driven Marketing Economy ("DDME) added $156 billion in revenue to the U.S. economy and fueled more than 675,000 jobs in 2012 alone.[3]

The remarkable growth of the DDME is due in part to the robust framework of sectoral laws and self-regulatory protections that already apply to the use of big data for marketing purposes. This framework combines specific legal restrictions, which focus on the potential misuse of data, with enforceable industry self-regulation that responds to a rapidly changing business landscape. It is this flexible framework of protections governing the responsible use of data for marketing purposes that has helped drive innovation and fuel the U.S. economy.

---

[1] "Big Data" Request for Information, 79 Fed. Reg. 42, 12251-52 (Mar. 4, 2014).
[2] www.thedma.org
[3] Deighton and Johnson, *The Value of Data: Consequences for Insight, Innovation & Efficiency in the U.S. Economy* (2013), *available at* http://ddminstitute.thedma.org/#valueofdata (hereinafter "*The Value of Data*").

Given that a robust and successful framework of protections already govern the use of "big data" for marketing purposes, the DMA encourages OSTP to continue the U.S. tradition of focusing on discernible, concrete harms to consumers when considering any additions to an already well-functioning and data-driven global economy.

1. **What are the public policy implications of the collection, storage, analysis, and use of big data? For example, do current U.S. policy framework and privacy proposals for protecting consumer privacy and government use of data adequately address issues raised by big data analytics?**

The DMA maintains that the current regulatory framework of sectoral laws designed to address concrete harms associated with the misuse of data, complemented by self-regulatory codes of conduct backed by enforcement mechanisms, appropriately fosters market innovation and addresses consumer privacy considerations. In the 1990s, the government considered comprehensively regulating the Internet and related connectivity through formal legislation, but ultimately maintained the long standing approach we have today in the United States toward privacy regulation – a sectoral framework that addresses particular areas of concern, such as children's online privacy or specific sectors perceived as handling sensitive information (*e.g.*, certain healthcare and financial services). These sectoral laws are supplemented by industry self-regulatory principles to effectively promote the responsible online and offline collection and use of data for marketing purposes.

"Big Data" has not created new concerns regarding the private sector's responsible use of marketing data for marketing purposes. Regardless of quantity, data is data – it may be used for good or for harm. The scope of data has not made existing protections less valuable or less effective, and the focus of policy frameworks should continue to be on uses of data shown to be harmful to consumers, rather than on restricting the responsible use of data for marketing purposes.

a. Sectoral Laws Address Concrete Harms. The current framework provides protections for consumers in particular areas where the nature of the data, if misused or misappropriated, could cause discernible harm to consumers. The United States has wisely taken a harm-based approach to these protections, identifying areas where consumers may be harmed and regulating those areas. For example, the Health Information Portability and Accountability Act ("HIPAA") protects the use of patient health data, the Fair Credit Reporting Act ("FCRA") protects against the use of consumer data for eligibility purposes, the Children's Online Privacy Protection Act ("COPPA") protects children's privacy on the Internet, and the Gramm–Leach–Bliley Act ("GLB") protects consumers' financial privacy. This harm-based approach to regulation has allowed the private sector to use data responsibly to improve consumer interactions with businesses with clear limits on certain uses, while at the same time enabling the delivery of more relevant marketing. The existing sectoral framework has thus proven to be a successful means of advancing innovation while also providing consumers with transparency and control over their data choices.

b. Self-Regulation Effectively Regulates Marketing Practices. For decades, the private sector has developed and enforced robust self-regulatory codes of conduct to complement

the sectoral legal framework. Unlike legislation, which is static and runs the risk of codifying practices that may become out-of-date even before a bill turns into law, industry self-regulation is nimble by its very nature and thus better suited to provide protections in cutting-edge areas such as the information economy.

The DMA itself promulgates and enforces its *Guidelines for Ethical Business Practice* ("*DMA Guidelines*"), which set forth guidance for how marketers may responsibly collect and use data for marketing purposes.[4] The *DMA Guidelines* require choice and transparency regarding responsible collection and use of marketing data. Disclosures about marketing data collection and use should be provided as appropriate, but in a manner that fits the situation. In addition, the *DMA Guidelines* require marketing data to be used solely for marketing purposes, and not for decisions about credit worthiness, employment, or other eligibility purposes.

The *DMA Guidelines* are updated regularly by DMA's Ethics Policy Committee to account for changes in the way consumers and marketers create and engage with data. For more than four decades, DMA has ensured that data is used responsibly through the robust enforcement of the *DMA Guidelines*. The DMA reviews complaints from several sources, such as consumers, member companies, non-members, and consumer protection agencies. Complaints are referred to the DMA's Ethics Operating Committee and are reviewed for potential violations of the *DMA Guidelines*. Penalties may be assessed, and some violations may be referred to the Federal Trade Commission or other appropriate law enforcement agencies as appropriate.

In addition, the Digital Advertising Alliance ("DAA"), an organization that the DMA helped spearhead, has also released its *Self-Regulatory Principles for Online Behavioral Advertising*, *Self-Regulatory Principles for Multi-Site Data*, and *Application of Self-Regulatory Principles to the Mobile Environment* (collectively, "Self-Regulatory Principles") to provide consumers with effective transparency and choice regarding the collection and use of web viewing and data gathered from mobile devices including precise location data and personal directory data.[5] These programs effectively regulate marketing data practices, delivering enhanced transparency and control to consumers. "Big data" has not changed these fundamental principles of transparency and control.

Self-regulation has worked for more than forty years to ensure responsible use of marketing data for marketing purposes, while enabling the growth of a strong data-driven economy.

c. The Current Framework Fosters Economic Prosperity. Within this current framework, the responsible use of data for marketing purposes has helped drive innovation and fuel the U.S. economy. A recent study, commissioned by DMA's Data-Driven Marketing Institute ("DDMI") and conducted independently by Professors John Deighton of Harvard Business School and Peter Johnson of Columbia University, entitled, *The Value of Data:*

---

[4] Direct Marketing Association, *Guidelines for Ethical Business Practice* (Jan. 2014), *available at* http://thedma.org/wp-content/uploads/DMA_Guidelines_January_2014.pdf.

[5] Digital Advertising Alliance, The DAA Self-Regulatory Principles, *available at* http://www.aboutads.info/principles/.

Consequences for Insight, Innovation & Efficiency in the U.S. Economy ("*Value of Data*"), quantifies this fact.[6] The *Value of Data* study found that the DDME added $156 billion in revenue to the U.S. economy and fueled more than 675,000 jobs in 2012 alone. The study also found that an additional 1,038,000 jobs owe part of their existence to these DDME jobs.[7] The study estimated that 70% of the value of the DDME – $110 billion in revenue and 475,000 jobs nationwide – depends on the ability of firms to share data across the DDME.

The current privacy framework remains effective today because it is sufficiently flexible to ensure that data is used responsibly while accommodating the use of "big data" analytics to benefit society as a whole.

2. **What types of uses of big data could measurably improve outcomes or productivity with further government action, funding, or research? What types of uses of big data raise the most public policy concerns? Are there specific policy concerns? Are these specific sectors or types of uses that should receive more government and/or public attention?**

The use of "big data" can measurably improve outcomes and productivity across the economy. DMA members constantly seek new ways to responsibly harness "big data" for the benefit of consumers and society. The DMA would welcome government investment in and support of research into how data may continue to improve outcomes and productivity in the private sector, while avoiding one-size-fits all rules in this developing area of the economy. Additionally, when appropriate, data maintained by the government should be made available to the private sector for use in creating technological advances, new products and services, and yet unknown benefits to society.

a. Responsible Data Use Helps Consumers & Society. For nearly a century now, DMA members have been collecting and analyzing consumer data responsibly to help companies reach the "right" audiences for their products and services. Such techniques have expanded to the online environment, helping businesses provide customized offerings to an even broader consumer audience. This growth has occurred in an environment that continues to focus on expanding transparency and limiting harms to consumers.

Innovative applications of marketing analytics to "big data" are providing benefits to society every day. Marketing data analytics make it more likely that an offer will be valuable to the consumer who receives it. For example, a retailer might look at what a customer has purchased at a particular store, through its website, from its mobile site and otherwise, and then analyze those purchases in comparison to others who have bought those items. Using analytics, the retailer will predict whether a customer is more likely to want a coupon for jewelry or for kitchen appliances and use the same data to identify and improve the channels where consumers are more likely to purchase and engage with the retailer's products.

---

[6] Deighton and Johnson, *The Value of Data: Consequences for Insight, Innovation & Efficiency in the U.S. Economy* (2013), *available at* http://ddminstitute.thedma.org/#valueofdata (hereinafter "*The Value of Data*").
[7] *The Value of Data* at 74.

Marketing data analytics also offer broader social benefits, outside the realm of commercial marketing. For example, charities and other nonprofit organizations use "big data" analytics to keep fundraising costs down by focusing on the most likely donors and tailoring their approaches to engage those in greatest need. A nonprofit might create a demographic profile of major donors and then search for new donors that fit a similar profile. Such analytics help lower fundraising costs, freeing up donated funds to work on achieving the organization's societal mission.

These responsible uses of "big data" for marketing purposes do not create new public policy concerns because there is no new or current risk of harm presented by such activity.

b. Self-Regulation Maintains Consumer Trust. DMA members understand that customer trust is the bedrock of their business. For more than four decades, DMA and its members have ensured the responsible use of marketing data by developing a rigorous regime of self-regulation, the *DMA Guidelines*. The *DMA Guidelines* keep pace with the rapidly changing economy and strive to maintain the consumer trust that has led to the success of the current framework. For example, the January 2014 update to the *DMA Guidelines* includes provisions that require DMA members to establish data security programs in order to protect their marketing databases. This addition to the *DMA Guidelines* reflects the fact that DMA members are constantly iterating on new ways to protect data from unwarranted exposure.

c. Public Policy Focus Should Be on Data Security and Government Use of Data.

Given that a robust and effective set of sectoral laws and self-regulatory protections already govern the use of big data for data-driven marketing, public policy discussions led by OSTP should focus on making it easier for companies to keep data secure, as well as considering when government access to commercial data is appropriate.

While the *DMA Guidelines* provide significant data security and breach notification requirements to which companies must adhere, the DMA has supported the creation of a national data breach notification standard to aid companies in the vital activity of protecting customer data for nearly a decade. A single national data breach notification law would ease the significant burden of complying with the current patchwork of state laws, lowering the cost of compliance and increasing market efficiency.

Improving the legal framework that controls government access to consumer data is another effective way to mitigate the harm that such access can present to consumers. Thus, DMA also supports reforming the Electronic Communications Privacy Act ("ECPA") to require law enforcement to obtain a warrant before gaining access to the content of digital consumer data. As currently written, ECPA provides a loophole through which law enforcement can gain access to consumer data without a warrant if it is stored on a third party server for more than 180 days.

These legal reforms would help assure consumers that their data is safe when they share it with marketers, from both unwarranted exposure to the government or bad actors. Adding these new pieces to the existing framework, while allowing the private sector to continue to

responsibly use data within the current self-regulatory framework, would be an adequate way to address new policy concerns presented by "big data."

3. **What technological trends or key technologies will affect the collection, storage, analysis and use of big data? Are there particularly promising technologies or new practices for safeguarding privacy while enabling effective uses of big data?**

To help facilitate innovation in the data-driven economy, the DMA encourages OSTP to refrain from focusing on or endorsing any specific technology that may be used in the collection, storage, and analysis of data. The Administration should not position itself to choose winners or losers of specific technologies. Instead, the marketplace is better equipped to identify specific technologies that may be useful in advancing the DDME. Data analytics is an exciting new space that engages skilled professionals across disciplines who are constantly innovating. "Big data" combined with today's computing power enable DMA members to tackle problems in new and often unexpected ways. The government sanctioning or otherwise approving of specific technologies impacting data could prematurely stunt this data-driven growth and unnecessarily hinder the economy. References to technology by OSTP, if any, should therefore be presented in a neutral fashion.

As mentioned previously, enforceable industry self-regulation offers one such technologically-neutral tool that we encourage OSTP to promote as a means of responsibly and safely collecting and using data.

4. **How should the policy frameworks or regulations for handling big data differ between the government and the private sector? Please be specific as to the type of entity and type of use (e.g., law enforcement, government services, commercial, academic research, etc.).**

The policy frameworks that oversee government and private sector data practices should continue to be distinct from one another. While government collection, access, and use of data raises important issues for consideration – including concerns over government surveillance – marketing data is collected and used solely for marketing purposes and thus does not raise similar concerns.

The current policy framework has placed constitutional and legal restrictions on the ability of the government to access and use data because such actions present a fundamental threat of harm to individuals through government overreach or abuse. Conversely, marketing data is used to facilitate commerce and to drive the economy, not for law enforcement, investigatory, or other governmental purposes. Thus, the responsible collection and use of marketing data for marketing purposes by the private sector presents no similar harm to individuals and in most areas has been allowed to appropriately regulate itself through robust and enforceable self-regulatory programs. To that end, marketers have long had a strong incentive to maintain the trust and confidence of consumers, and the DMA has helped to further that commitment through self-regulation.

Policy frameworks should continue to respect the important and real differences between the use of data by the government and the private sector, and allow the private sector to continue forward with the existing framework that has both protected consumers and grown the economy for decades.

5. What issues are raised by the use of big data across jurisdictions, such as the adequacy of current international laws, regulations, or norms?

The flow of data across borders contributes to an efficient global economy. To help promote global interoperability and economic growth, OSTP should encourage policies that reduce barriers to trans-border data flows. Currently, both industry self-regulatory efforts and government safe harbor frameworks play an important role in facilitating the responsible use of data across jurisdictions. These means of promoting data flows should be allowed to continue.

a. Self-Regulation Aids Global Data Flows. One of the benefits of current self-regulation for marketing data practices is that it can reach across borders. The *DMA Guidelines* apply to both DMA members and non-members, whether they are based in the United States or abroad. The same principles apply to the use of data regardless of where that data is, and self-regulation and specific harm-based regulations present the best approach for addressing these issues.

b. Government & Private Sector Partnerships Support Data Flows. Another set of programs that currently facilitate international flows of data are the U.S.-EU and U.S.-Swiss Safe Harbor programs. These programs regulate the flow of data from Europe into the United States by helping to ensure that personal information is adequately protected. Companies that are members of the DMA may select the DMA as their safe harbor dispute resolution mechanism.[8] As of mid-2013, the DMA Safe Harbor Program was serving 62 participating member companies. These programs offer an efficient and enforceable protection for the international flow of data into the United States from Europe.

For the DDME to continue to thrive, such self-regulatory efforts and safe harbor programs should be promoted as vetted and acceptable means of promoting global interoperability in a responsible manner.

* * *

DMA thanks you for the opportunity to submit these comments, and we look forward to working with OSTP on these important matters. Please do not hesitate to contact me with any questions at (202) 861-2420.

[8] *See* Direct Marketing Association, The DMA Safe Harbor Program: A Guide for Businesses, *available at* http:www.dmaresponsibility.org/safeharbor/.

Sincerely,

Peggy Hudson
Senior Vice President, Government Affairs
Direct Marketing Association

CC: Stu Ingis, Venable LLP

[redacted]

From: Durrell Kapan [redacted]
Sent: Monday, March 31, 2014 5:11 PM
To: bigdata@ostp.gov
Subject: Big Data RFI

Attn: Big Data Study, Office of Science and Technology Policy Eisenhower Executive Office Building
1650 Pennsylvania Ave. NW.
Washington, DC 20502.

From: Durrell D. Kapan, Ph.D., Mill Valley, CA 94941

(1) What are the public policy implications of the collection, storage, analysis, and use of big data? For example, do the current U.S. policy framework and privacy proposals for protecting consumer privacy and government use of data adequately address issues raised by big data analytics?

The most important policy challenge facing our country RE: big data are how efforts to improve privacy policy simultaneously helps individuals maintain control over their data (an hence, avoiding negative pitfalls of loss of control of private, individually identifiable information) but hurts opportunities to capitalize on big data for the benefit of these same individuals and groups to which they belong. This tradeoff is inherent in all phases of collection, storage and analysis of big data.

(2) What types of uses of big data could measurably improve outcomes or productivity with further government action, funding, or research?

Big data improvements are potentially ground-breaking at three levels:

1) For individuals, big data, if 'owned, controlled and released' by said individuals can have a major positive impact on personal health and well-being. For example, if health care was integrated with behavioral data now capable of being automatically generated with a new era of smart devices, then patients and doctors could more easily 'be on the same page'. Needless to say extensive research needs to occur to implement a way that individuals can control health care records as well as selectively reveal their behavioral data to their individual providers. These data could have implications to populations and larger groups (see below).

2) This same data has potentially huge impacts on population health by first improving the 'snapshots' we take with current clinical trials and other research methods into longitudinal phenotype data bases. Imagine clinical trails where enrollee's provided comprehensive genomic and environmental (phenotype) data to researchers in exchange for pin-pointed health advice based not only on hypothesis driven research but also on big data driven correlational evidence derived from like patients in like environments. Suddenly aggregate data could be 'personalized' and medicine could move beyond the statistically 'robust' but biologically very weak pronouncements of 'the effect of factor X on group Y is an Z percent increase in the odds of …' to strong inference and a heads up to health care consumers: "folks with similar genotypes at the following loci [ ] who have a history of diets high in X & Y that don't exercise with more than Z frequency form a group that is 100x more likely to die from heart disease than those with the similar genotypes, but have modified their diet and exercise to eliminate or reduce X & Y and increase exercise above amount Z'.
We need research to begin to facilitate this type of personalized medicine as well as research into the tools that allow the safe capture, merging and sharing of such BIG data across disparate data types and entities.

3) Across all populations, public good is lost when we are not able to seamlessly and safely provide input based on our data to benefit society in general without fear of privacy problems and potential financial or safety concerns. For

instance, correlated patterns of decreased movement in smart-phones in urban areas (relative to normal commute patterns) could really help authorities identify emerging or resurging infectious diseases. However, no-one wants the government to monitor them at this level. Research should pinpoint methods to analyze these data in a locally distributed way without capturing or saving any personally identifiable information.

What types of uses of big data raise the most public policy concerns?

1) The type of data doesn't raise the concern, it is how it is handled. Health records are mishandled in almost all settings, either inefficiently captured or captured on paper then lost. Clearly once we integrate behavioral data then how and why this is captured becomes important, but the handling of this data is most critical. Does the user keep a 'private key' to the identifiable bits of his or her data trail? How else do we upload, store, protect and eventually purge said data?

Are there specific sectors or types of uses that should receive more government and/or public attention?
I think geolocated data and private genomic data have the greatest potential for abuse by the government / law enforcement and the insurance industry. I think that automated protocols that decrease data precision for particular elected uses need to be instituted so that generators of big data derive the most benefit and avoid downfalls.

(3) What technological trends or key technologies will affect the collection, storage, analysis and use of big data?

I think that individuals could keep their own 'big data' in certain categories, and that they should be aware of it and where it goes. Clearly a GPS track of individual movement aggregated over a lifetime might be hard for individuals to keep, but we wouldn't want this persistent data to be available forever either.

(3 continued) Are there particularly promising technologies or new practices for safeguarding privacy while enabling effective uses of big data?

As a non expert I imagine strong encryption and private-public key methods to open encrypted files to be critical. Somehow if the private key could include the precision of the big data in it, then one could offer low precision data 'for sale' to companies without revealing potentially damaging high-precision data. This should be investigated further. One idea is that you should be able to 'pull' your big data from a preauthorized use if you decide the entity using your data is not being a good steward of the data. Pull requests should include the private key holder's authentication and should allow seamless blanking of the public portion of the record(s) upon request. This will require serious re-tooling of data handling and communication protocols to include record level encryption and control.

(4) How should the policy frameworks or regulations for handling big data differ between the government and the private sector?

I think they should NOT. I think that a gold standard should exist that protects the individual at the cost of so called "programs for your own good" that reek of big-brother. With private-public key encryption + designed levels of precision and consumer control of their data streams (such as clear 'opt-in' policies across private / public data aggregators) then consumers should be able to protect themselves while still deriving a benefit.

Please be specific as to the type of entity and type of use (e.g., law enforcement, government services, commercial, academic research, etc.).

(5) What issues are raised by the use of big data across jurisdictions, such as the adequacy of current international laws, regulations, or norms?

Unregulated e-commerce is potentially the biggest area of concern here. Buying patterns and other big data provide a huge advantage to companies. If some offshore entities utilize big-data outside of the legal norms of US companies they

could simultaneously derive an unfair advantage and put US consumers at risk. Since I know little of international laws in this area I can only say this is an area for concern.

ELECTRONIC TRANSACTIONS ASSOCIATION

1101 16th Street NW
Suite 402
Washington, DC 20036

www.electran.org
T 800.695.5509
T 202.828.2635
F 202.828.2639

March 31, 2014

Nicole Wong
Office of Science and Technology Policy
Attn: Big Data Study
Eisenhower Executive Office Building
1650 Pennsylvania Ave. NW
Washington, DC 20502

Sent via email to bigdata@ostp.gov

Re: Notice of Request for Information, "Big Data RFI," FR Doc. 2014-04660

Dear Ms. Wong:

The Electronic Transaction Association ("ETA") provides comments in response to the White House Office of Science and Technology Policy's ("OSTP") request for information regarding "big data" published on March 4, 2014.[1]

The Electronic Transactions Association ("ETA") is an international trade association representing more than 500 companies that offer electronic transaction processing products and services. The purpose of ETA is to influence, monitor and help shape the merchant acquiring industry by providing leadership through education, advocacy and the exchange of information. ETA's membership spans the breadth of the payments industry, from financial institutions to transaction processors to independent sales organizations (ISOs) to equipment suppliers.

The members of the Electronic Transaction Association leverage data to provide a wide range of products and services designed to enhance and secure electronic transfers. Our members rely on data to help reduce fraud and to authenticate transactions between businesses and consumers in order to make such transactions seamless and secure. To keep pace with and address changes in technology and cyber threats that may arise with the use of data, ETA has established the Technology Committee to respond to these issues. We provide this comment for your consideration.

(1) What are the public policy implications of the collection, storage, analysis, and use of big data? For example, do the current U.S. policy framework and privacy proposals for protecting consumer privacy and government use of data adequately address issues raised by big data analytics?

The collection, storage, analysis, and use of data drive the U.S. and global economies today. Data facilitates and promotes commercial activity between consumers and businesses, not only through the Internet but also in shops and storefronts around the world as well. Electronic payment methods like credit, debit, prepaid, gift and benefits cards, as well as automatic withdrawals, provide speed and efficiency for connecting

[1] "Big Data" Request for Information, 79 Fed. Reg. 42, 12251-52 (Mar. 4, 2014).

buyers and sellers around the world. Today, someone with a card issued by a bank in Fargo, North Dakota, can visit a shop in Laos and make a purchase with that credit card. In a matter of seconds the merchant can transmit information for authentication and processing, with consumers receiving their goods immediately and merchants receiving payment within a few days. Millions of such transactions are made possible every day because of the free flow of data. Access to and use of data opens new markets, lowers the barriers to market entrance for businesses, and increases competition.

Because of the success and efficiency of the industry model, consumers have come to expect that they can make payments and purchase goods electronically by transmitting data through payment networks at point of sale terminals and through online portals. Today, 70 percent of consumer spending is done electronically, a share that is expected to increase even more with time as the mobile payment industry continues to innovate and expands its options for consumers.[2]

The strong embrace by consumers and businesses alike provides strong testimony to the benefits of the electronic transactions model and signals the importance encouraging more innovation and development in this space. Over the past few decades, U.S. policy has encouraged innovative—yet responsible—uses of data in the payments industry and the industry has flourished. Going forward, policymakers should continue to build on this successes, specifically by promoting:

1. the wide availability of good quality data, including the free availability of publicly-funded data;
2. the free flow of data across local, state , and international boundaries, to foster a seamless digital marketplace; and
3. investment in infrastructure that supports the flow and access to data.

ETA believes that a one-size-fits all approach to the regulation of data would not work in today's modern economy and would impose significant costs on consumers and businesses alike. The current U.S. framework is comprised of a combination of regulation and flexible, enforceable standards. In the electronic transactions industry, certain data like financial information is governed by federal law, including the Gramm-Leach-Bliley Act, while other data and its uses are governed by robust self-regulatory programs including the Payment Card Industry Data Security Standard ("PCI-DSS"), which sets forth requirements designed to ensure companies that process, store or transmit credit card information maintain a secure environment for such data. Through a calibrated mixture of federal law and industry codes, policymakers have recognized that for certain data and uses, prescriptive federal laws could burden commerce and innovation. In these areas, industry has been encouraged to self-regulate to keep pace with the fast moving technology and emerging channels like mobile.

[2] Why are Electronic Payments Important (2013), available at http://www.electran.org/wp-content/uploads/ETAPaymentsInfographic.pdf.

(2) What types of uses of big data could measurably improve outcomes or productivity with further government action, funding, or research? What types of uses of big data raise the most public policy concerns? Are there specific sectors or types of uses that should receive more government and/or public attention?

The growth of the payments industry is accelerating, as is the use of data to support this growth. As noted, consumers and businesses alike benefit from the growth in the industry by providing greater security and efficiency for electronic transactions and data. The rapid acceleration of industry growth favors market entry for small startups developing novel processes as well as investment in research and development by existing companies, creating jobs and opportunity for the U.S. economy. Government policy should be designed to encourage the further growth of the payments industry to achieve these important benefits.

3) What technological trends or key technologies will affect the collection, storage, analysis and use of big data? Are there particularly promising technologies or new practices for safeguarding privacy while enabling effective uses of big data?

To keep pace with consumer demand for better, faster and more ubiquitous payment options, the payments industry has been innovating rapidly in recent years. For example, companies have developed tools that give merchant portfolio owners and managers insight into the performance not only of their own portfolios, but of the portfolios of others in the market as well, using sophisticated techniques that provide greater context, sample size, micro-segmenting ability and other critical and informative features. These Big Data breakthroughs drive value in many ways and are reflective of the future of the industry. They are made possible by policies that promote access to and use of data to advance the field of payments processing.

In the coming years, the payment card industry will migrate to EMV technology that will provide enhanced security to consumers while continuing to foster a payment system that facilitates seamless and timely electronic transactions. EMV and other proposed measures, such as encryption and tokenization, are being considered as methods for safeguarding consumer privacy while enabling effective uses of data. ETA encourages a robust examination of these measures and their impacts on rates of consumer fraud, purchase processing times, and other related issues.

(4) How should the policy frameworks or regulations for handling big data differ between the government and the private sector? Please be specific as to the type of entity and type of use (e.g., law enforcement, government services, commercial, academic research, etc.).

Government access to private citizen and corporate data raise issues that are distinct from the private sector. ETA supports efforts to clarify by law the appropriate balance between citizen privacy and government access to private data. With respect to a policy framework for private sector handling of data, ETA believes the current U.S. framework is the appropriate approach. The combination of federal and state law focused on concrete harms supplemented by industry codes of conduct has proven to be an effective means for regulating data – big or small. The policy framework should support the collection and use of data as well as the tools used to gather and analyze big data sets. This approach will foster research and innovation that will deliver new benefits to society and consumers.

While ETA believes that the market forces are best source to shape practices, there is one area in need of government involvement- data breach notification. ETA supports a uniform national data breach notification law to provide consumers with consistent, predictable expectations about how and when they will be notified of a breach, Currently there are 46 different state laws governing the details of notification in the event of data breach.

(5) What issues are raised by the use of big data across jurisdictions, such as the adequacy of current international laws, regulations, or norms?

ETA recommends that OSTP consider ways to foster the use of data by lowering barriers to cross-board transfers. Consumers live and shop globally and should be provided a secure experience regardless of where in the world they elect to purchase goods or services. The U.S. should encourage other nations to improve data infrastructure to facilitate data flow and promote international commerce.

Restrictions to the free flow of data across local, state, and international boundaries should be removed. ETA believes that arbitrary distinctions that impose limitations on access to and use of data impede the conduct of efficient and secure transactions. The U.S. should promote the removal of such obstacles to open new markets to the world economy and to make U.S. goods available for purchase to consumers in foreign lands. Efforts should be undertaken to encourage more interoperability between nations and make commerce more interconnected and seamless for consumers. The U.S. Government could foster these efforts by making the data it maintains more accessible and useable by the private sector. Access to open data sets will improve governance, deliver benefits to society, and drive innovation.

* * *

ETA thanks you for the opportunity to submit these comments. Please do not hesitate to contact me with any questions at 202 677-7403.

Scott Talbott
Senior Vice President for Government Affairs
Electronic Transaction Association

FASEB
Federation of American Societies for Experimental Biology

Representing Over 115,000 Researchers

301.634.7000
www.faseb.org

9650 Rockville Pike
Bethesda, MD 20814

March 31, 2014

Nicole Wong, JD
Office of Science and Technology Policy
Eisenhower Executive Office Building
1650 Pennsylvania Ave. NW
Washington, DC 20502

RE: Request For Information: Government "Big Data"

Comments submitted electronically to: bigdata@ostp.gov

Dear Ms. Wong:

The Federation of American Societies for Experimental Biology (FASEB) is composed of 26 scientific societies, collectively representing more than 115,000 researchers. FASEB recognizes the increasingly important role "big data" plays in research and throughout society. We thank the Office of Science and Technology Policy (OSTP) for this opportunity to provide comments on government "big data" and resulting privacy issues. Our comments are drawn from previous FASEB statements that address different aspects of government "big data" as related to biomedical and life science research. We have appended the relevant statements to this response and call your attention to several points that are most pertinent to the questions listed in the OSTP request for information (RFI).

Responses to questions 1-3 and 5 from the OSTP RFI:

(1) Current U.S. policy frameworks and privacy proposals are insufficient to ensure the privacy of human research subjects in perpetuity. In comments on the proposed National Institutes of Health (NIH) Genomic Data Sharing Policy (see attached), FASEB stated that "de-identification cannot be guaranteed for certain types of data, including whole genomic sequences." FASEB, therefore, recommended the consideration of alternative models to protect human research subjects, such as shifting from a privacy-protection paradigm to "one that provides research subjects with substantive legal protections against the misuse of or inappropriate access to their data."

(2) In several statements, FASEB has noted the potential value of "big data" analysis and data sharing. These possible benefits include improved health, quality of life, and clinical care, as well as the development of new and transformative technologies.

(3) There is a critical need for more tools and systems to promote high quality metadata collection. The development of these is essential to the creation of datasets and their ultimate utility (see attached comments on the proposed NIH Data Catalog).

However, the possibility of re-identification of individuals increases as more data from both research and non-research sources become available, scientific knowledge increases, and analytical tools improve.

(5) FASEB is currently developing comments to submit to the Institute of Medicine (IOM) request for public feedback on the responsible sharing of data from clinical trials that address variation across international regulations. We encourage OSTP to review the ongoing IOM study and discussion framework document. We are also willing to share our comments to IOM with OSTP once they are available.

FASEB appreciates your consideration of our comments and looks forward to working with OSTP on these issues. Please let us know if we can be of further assistance.

Sincerely,

Margaret K. Offermann, MD, PhD
FASEB President

Attachments
FASEB Comments on the draft NIH Policy for Genomic Data Sharing
FASEB's Response to the Request for Information on the Development of an NIH Data Catalog

FASEB
Federation of American Societies for Experimental Biology

Representing Over 110,000 Researchers

301.634.7000
www.faseb.org

9650 Rockville Pike
Bethesda, MD 20814

November 6, 2013

Genomic Data Sharing Policy Team
Office of Science Policy, National Institutes of Health
6705 Rockledge Drive
Suite 750
Bethesda, MD 20892

Dear NIH Genomic Data Sharing Policy Team:

The Federation of American Societies for Experimental Biology (FASEB) appreciates the opportunity to comment on the National Institutes of Health's (NIH) draft Genomic Data Sharing Policy (NOT-OD-13-119). FASEB is comprised of 27 scientific societies, collectively representing over 110,000 biological and biomedical researchers. FASEB recognizes the importance of promoting the sharing of data resulting from genomic research studies and commends NIH for developing a draft policy to encourage such behavior among NIH-funded investigators.

FASEB agrees with the general principles underlying the draft Policy, as the exchange of research findings serves to increase the efficiency of scientific research and accelerate discoveries that could improve human health. We are, however, concerned about potential unintended impacts; specifically, large increases in administrative burdens for investigators and institutions engaged in genomics research and a decrease in participation of human subjects in clinical research with genomics components due to concerns regarding misuse of shared data. Therefore, FASEB strongly recommends that NIH revise the draft Policy to address the concerns detailed below prior to its implementation.

Increased Administrative Burdens for Investigators and Institutions

While FASEB appreciates the role data sharing plays in advancing scientific discovery, we are concerned that the Policy may introduce unintended administrative burdens at every stage of the research process, from development of a data sharing plan and timeline in the grant proposal to managing the submission of data to databases, including de-identification and development of coding schemes and maintaining this information after the conclusion of the grant award. Below, we highlight some key areas that NIH should address to improve the clarity of the draft Policy.

Lack of Clarity Regarding Types of Research Covered by Policy

FASEB's greatest concern is the lack of clarity regarding the types of research that would be covered by this Policy. Appendix A contains only four examples of the types of research to be covered, which vary greatly in terms of sample size, species of research subject, and overall detail of the volume of sequence data that would trigger coverage by the Policy. The table describing expectations for data submission and data release is similarly vague. While the intent may have been to produce a policy with flexibility to adapt to the rapidly changing field of genomics research, this lack of clarity will likely result in confusion and increase administrative burden as investigators and institutions struggle to determine what constitutes compliance. In response, some institutions may treat all research utilizing genetic methods as the "large-scale" genomic research addressed by this Policy, extending administrative burdens far beyond NIH's intention.

To reduce confusion, FASEB strongly recommends that NIH supplement the final Policy with a workflow diagram or chart to help investigators and institutions navigate and implement the Policy as intended. For example, the supplemental grant application instructions for the PHS 398 and SF 424 forms include scenarios to help guide researchers in determining whether the proposed research does not include human subjects research, includes non-exempt human subjects research, or includes exempt human subjects research. The supplemental instructions also link to the Department of Health and Human Services Office of Human Research Protections where additional information regarding each category can be obtained. A similar set of detailed scenarios for the draft Genomic Data Sharing Policy could serve to enhance implementation and compliance. This would also decrease the risk of over-regulation by institutions, a phenomenon that FASEB's recent survey of federally funded researchers found to be a major source of unnecessary administrative burden.

Option of "Just-In-Time" Data Sharing Plans for Funded Research Proposals
While FASEB agrees that investigators should consider plans for data sharing in the early stages of their research project, we are concerned that the draft Policy requires fairly detailed data sharing plans at the research proposal stage. In fiscal year 2012, the success rate for NIH grant applications was 17.6 percent; NIH Director Francis Collins stated that the success rate for the current year will be 15 percent. The time devoted to the development of a detailed data sharing plan for a proposal with a high likelihood of not being funded is wasteful, and thus we urge NIH to consider adopting a "just-in-time" process that allows investigators to submit basic data sharing plans at the time of proposal submission. Institutional certification and associated documentation should only be required for those proposals recommended for funding. This is commensurate with current policy regarding development and submission of detailed project budgets and has also proven to be successful with human subjects and animal research protocols.

Lack of Clarity Regarding Activities that Constitute Data Sharing
Throughout the draft Policy, "data sharing" is described as deposition of sequence data into "any widely used data repository, whether NIH-funded or not." Per these guidelines and NIH's current policy regarding open-access to publications resulting from federal support, it is unclear whether publication of research results in scientific journals would be accepted as compliance with this draft Policy. To alleviate confusion and potential non-compliance, FASEB urges NIH to clarify the role of publication activity in compliance with the Genomic Data Sharing Policy.

Additional Barriers to Human Subject Research and Research Participation
Genetics and genomics investigators have been leaders within the scientific community in their recognition of the ethical, social, and legal implications (ELSI) associated with their research. FASEB has identified three aspects of the draft Genomic Data Sharing Policy that may be problematic in light of current ELSI research.

Lack of Guidance Regarding Data for which De-identification Cannot Be Guaranteed in Perpetuity
The draft Policy does not address the fact that de-identification **cannot** be guaranteed for certain types of data, including whole genomic sequences. As more data become publically available in both biomedical and non-biomedical databases (such as those for personal ancestry research), and as genotype-phenotype correlations become more predictive, the possibility increases that a genomic sequence could allow for re-identification of research participants. Therefore, NIH should consider alternative models to protect human research subjects. Shifting from a paradigm centered on an institution's responsibility to ensure

data privacy to one that provides research subjects with substantive legal protections against the misuse of or inappropriate access to their data may be a more effective way to minimize risks to research participants.

Exemptions to Data Sharing for Individual Research Participants
The draft Policy has striking implications for the recruitment of patients for clinical trials. Can individuals who, at the time of consent, restrict use of their genomic data to only the original research team still be allowed to enroll in the study? In certain situations, allowing enrollment of these individuals may be critical to the integrity of the research project. The following are a few possible challenges an investigator may face: (1) a higher frequency of non-consent to data sharing among some populations and groups, which would reduce the representativeness of the study population; (2) for rare diseases, loss of only a few individuals from a study, which could be detrimental to achieving a sufficient sample size; and (3) increases in time and costs of study recruitment greatly beyond what their source of funding can provide.

Exemptions to Data Sharing for Select Research Projects
Finally, FASEB is concerned that the draft Policy does not specifically allow investigators to seek exemptions for an entire study when the research is associated with "at risk" or vulnerable populations, including children or specific racial, ethnic, or tribal groups. Some subjects may be more distrustful due to historical abuses and may wish to limit use of their data to the original research team. Also, specific restrictions regarding data sharing may be requested during community consultations, and it is important that researchers are able to respect these requests. Limited technological and biomedical literacy within some populations could also pose a barrier to ensuring informed consent for genetic data sharing. FASEB is concerned this Policy could have an unintended chilling effect on research designed to address health issues or ameliorate health disparities found among these populations.

In conclusion, FASEB commends NIH for its leadership in the development of policies to guide the rapidly changing field of genomics research and appreciates the opportunity to provide comments on the draft Genomic Data Sharing Policy. While we recognize the important role data sharing plays in furthering scientific discovery, if the draft Policy is implemented as currently written, we are concerned it will cause a large increase in administrative burdens for investigators and institutions engaged in genomics research and a decrease in participation of human subjects in clinical research with genomics components. Therefore, we urge NIH to make significant revisions to the areas of the draft Genomic Data Sharing Policy noted above and provide the public with an opportunity to comment on the revised version prior to its finalization and implementation.

Sincerely,

Margaret K. Offermann, MD, PhD
FASEB President

FASEB
Federation of American Societies for Experimental Biology

301.634.7000
www.faseb.org

9650 Rockville Pike
Bethesda, MD 20814

Representing over 110,000 researchers

The American Physiological Society

American Society for Biochemistry and Molecular Biology

American Society for Pharmacology and Experimental Therapeutics

American Society for Investigative Pathology

American Society for Nutrition

The American Association of Immunologists

American Association of Anatomists

The Protein Society

Society for Developmental Biology

American Peptide Society

Association of Biomolecular Resource Facilities

The American Society for Bone and Mineral Research

American Society for Clinical Investigation

Society for the Study of Reproduction

The Teratology Society

The Endocrine Society

The American Society of Human Genetics

Environmental Mutagenesis and Genomics Society

International Society for Computational Biology

American College of Sports Medicine

Biomedical Engineering Society

Genetics Society of America

American Federation for Medical Research

The Histochemical Society

Society for Pediatric Research

Society for Glycobiology

Association for Molecular Pathology

July 12, 2013

Jennie Larkin, PhD
National Heart Lung and Blood Institute
National Institutes of Health
6701 Rockledge Drive
Rockledge II, Room 8200
Bethesda, MD 20892-7940

RE: Request for Information (RFI): Input on Development of a NIH Data Catalog, Notice Number: NOT-HG-13-011

Comments submitted electronically to: data-catalog@mail.nih.gov

Dear Dr. Larkin:

The Federation of American Societies for Experimental Biology (FASEB) is composed of 27 scientific societies collectively representing over 110,000 biomedical researchers. FASEB recognizes the importance of facilitating data sharing across the biomedical research enterprise and expresses its support for the proposed National Institutes of Health (NIH) data catalog to document publically available datasets and house related metadata. FASEB thanks NIH for this opportunity to provide comments on the development of this resource.

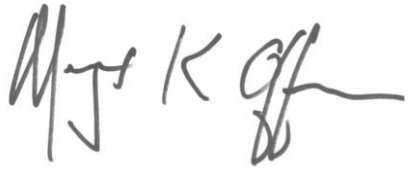
The FASEB Subcommittee on Research IT Issues convened a special Task Force to guide the Federation's comments on this topic. A list of Subcommittee and Task Force members and their professional affiliations is included as an appendix. FASEB's comments on the specific points raised in the NIH request for information (RFI) are also attached. The Task Force identified several key barriers for NIH to address during the development and maintenance of the catalog. One critical barrier that we would like to highlight is the **lack of incentive for researchers to submit catalog entries** as the majority of the effort is to be borne by the initial investigator for the benefit of secondary users. Other barriers identified by our Subcommittee and Task Force include a need for continuous standardization and curation of catalog entries and metadata, the lack of tools and systems to promote high quality metadata collection and catalog entries, and a variety of currently recognized barriers to data sharing in general.

For this proposed resource to be considered a long-term success, the data catalog must progressively expand in scope and be viewed a valuable tool by researchers in all life science fields and specialties. To achieve this goal, FASEB recommends that NIH focus its initial cataloguing efforts on data types that are already commonly deposited in databanks or data repositories and for which metadata reporting has become fairly standardized. Learning from this initial process, NIH can then support standardization

efforts for other data types and eventually incorporate these into the catalog. Finally, FASEB strongly recommends that NIH not limit catalog entries to research supported by NIH, and instead accept biomedical and life science catalog entries regardless of the original funding source. To communicate this to the research community, we encourage NIH to adopt a more generic name for the catalog.

Creation of data catalogs is generally recognized to be highly useful and easy to conceive, but difficult to implement effectively and cost efficiently. Therefore, it is critical that NIH carefully develop the data catalog to maximize the value of this investment, and FASEB encourages NIH to continue working with data catalog designers, stakeholders, and potential users throughout this process. FASEB appreciates your consideration of our comments and looks forward to working with NIH on these issues. Please let us know if we can be of further assistance.

Sincerely,



Margaret K. Offermann, MD, PhD
FASEB President

FASEB's comments on specific points raised in the RFI:

The critical barriers, opportunities, or incentives to making data more easily discoverable and citable, and the possible impact of a Data Catalog.

FASEB supports the creation of an NIH data catalog to document publically available datasets and house relevant metadata for these datasets. Database entries should be both human and machine readable. FASEB recommends inclusion of the following minimum metadata elements for all catalog entries:

- Unique identifier (e.g., accession number)
- Information content (e.g., species name)
- Information format (e.g., XML)
- Location of the dataset (e.g., name of databank and a link)
- Authors/creators of the dataset (including any standardized identification system NIH may adopt in the future, such as ORCID)
- Any associated publications (e.g., PMID)

The following five major barriers that could diminish value of the NIH data catalog:

1. There is a **lack of incentive for researchers to submit catalog entries** as the majority of the effort is to be borne by the initial investigator for the benefit of secondary users. Therefore, efforts by NIH to encourage depositions and increase the benefits for the original investigator are essential.
    - Simplification and standardization of data citation practices so as to enhance the perceived value of data sharing by the biomedical research community. To support improved data citation practices, FASEB makes the following recommendations: (1) all entries in the catalog should have a unique identifier (accession number); (2) NIH should work with the National Library of Medicine to develop standardized citation systems for different types of data; and (3) catalog functionality should permit users to download citation information into common bibliographic platforms.
    - NIH should also address the valuation of data sharing within the research community and ensure that these activities are given appropriate consideration when reviewing grant applications and progress reports.
    - Finally, FASEB recommends that NIH explore ways the catalog could be designed so as to promote research collaborations rather than mere reuse of data.

2. **Metadata standards will need to be frequently reviewed and harmonized**. NIH should also develop ontologies for metadata elements to enhance catalog query functionality. Attendant standards should be based on data type as opposed to field or topic of research (i.e., metadata standards for DNA sequences in general, rather than microbiome genomic metadata or human DNA sequence metadata). Standards should also be continuously harmonized and updated for each data type and between data types that are commonly used in combination by researchers, as was done for NMR and crystallography metadata.

3. FASEB recognizes that some early stage research methodologies, including high throughput approaches, and disciplines acquiring new data types, may not be ready for the development of meaningful metadata standards. Therefore, FASEB recommends that for early inclusion in the catalog NIH **focus on data types for which there already exists a demonstrated interest in shared and accessed databanks, and private and communal data repositories have been developed, so as to take advantage of metadata reporting that may have become fairly standardized.** This will help clarify and guide the continued support for future development of standards for other data types.

4. FASEB emphasizes **the need to ensure the high quality of the reported metadata**.
   - Allow the research community to provide comments or annotations to catalog entries on an ongoing basis, which would function as a dynamic review and consultation system. Notably, this has been found to be particularly valuable in the marketplace (e.g., Amazon.com).
   - FASEB further recommends that NIH ensure the development of tools to simplify and automate collection and reporting of metadata, such as a basic functionality that permits users to copy and edit metadata from existing entries, as well as import it from commonly used databases and data acquisition systems. These tools would help reduce the administrative burden of productive data sharing. NIH should explore ways to encourage quality reporting of metadata. For example, the introduction of a "plausibility tool" to scan depositions into the Protein Data Bank discouraged the accumulation of "sloppy" data.
   - Finally, it is essential that NIH develop a system to monitor data integrity and update entries as datasets are moved or lost.

5. FASEB notes that **many barriers to data sharing remain.** These include: (1) privacy and informed consent; (2) intellectual property protection and the Patent and Trademark Law Amendments Act of 1980; (3) biosecurity concerns; and (4) agency and funding mechanism-specific data sharing rules that require documentation of all secondary data users. Failure by NIH and the federal government to address these barriers will greatly limit the datasets available for inclusion in the NIH data catalog.

Possible Data Catalog linkage to existing data repositories to ensure data within the repository are findable and how to ensure that such linkages remain up to date and accurate.

It is critical that catalog users are accurately directed to the original datasets and that links to datasets are maintained. Whenever possible, the catalog should provide a direct link to the dataset of interest, and in all other cases, provide a link to the database, repository, or website within which the data are stored. The catalog should provide reliable verification that the links are "live" and the data described are still available at that location. BitTorrent may serve as a useful model of how to dynamically track and update dataset locations.

If your research field has no existing repositories to store data, comments can include how a Data Catalog might usefully link out to the data and where such data might be located.

The NIH data catalog should allow investigators to create entries that link to data housed outside of a standard data repository system. NIH should also encourage databases and repositories housing new data types to join, even if they do not yet have fully developed metadata standards, so as to foster their development.

The useful level of granularity for a Data Catalog entry. For instance, a Data Catalog entry may correspond to all the data in a publication, only a particular data type within a given study, or individual dataset from a single experiment.

FASEB strongly recommends that NIH structure catalog entries according to data type, so that each entry contains a single discrete set or independent unit of data, rather than bundling by publication. Storing multiple types of data in a single entry would be particularly confusing in the case of publications that utilize a wide variety of experimental methods. Such a bundling approach would also capture many types of datasets that the catalog is not initially prepared to handle. A data type-based approach would also facilitate discoverability within the catalog across disciplines, and would support an effective query function.

Any potential requirements for Data Catalog registration of data by NIH-funded or supported investigators.

FASEB encourages NIH to focus on data types in cases where there already exists a demonstrated interest in shared and accessed databanks, and private and communal data repositories have been developed, so as to take advantage of metadata reporting that may have become fairly standardized. This will help clarify and guide the evolution of metadata standards for other data types which may be in the early stages of development. Given the need for standardization, and the need to ensure that participation requires the minimal administrative effort possible and that its value is perceived by the scientific community, FASEB discourages NIH from making any broad requirements at this time. FASEB also recommends that any future requirements be applied only to new awards due to the difficulties of retrofitting already collected data and metadata for a catalog entry. NIH can encourage investigators to include catalog submission as part of a data management plan and encourage journals to request a accession number for publication when applicable, as is frequently done for the GenBank accession numbers of sequence data.

Whether a Data Catalog entry benefits from a scientific abstract that describes the data, including its potential uses and the rationale for its creation.

FASEB believes the most valuable information elements the catalog could provide are standardized, data type-specific metadata collection fields, which can be more easily queried than an abstract and have greater utility for machine-readable outputs. Noting the many aforementioned barriers to catalog participation, FASEB discourages NIH from making catalog submission an excessively time consuming process. FASEB also cautions NIH that submitted abstracts may be viewed by journal publishers as too similar to text within the corresponding journal article and, thus, may be considered a violation of copyright. While NIH should allow investigators the option to provide a detailed description of the data, providing such text should not be required.

The feasibility of the development of a Data Catalog to potentially support future uses.

To be of great value to the research community, the NIH data catalog will require a substantial and sustained investment of time and effort and should be seen as an evolving process. NIH should look to historical examples, such as the development of the Protein Data Bank, to better understand the long-term commitment and continued development required. NIH should also be mindful of past efforts and the lessons learned from other database and catalog efforts including caBIG and the fPET scan database.

Appendix: Members of the FASEB Task Force

Robert J. Robbins, PhD (FASEB Research IT Subcommittee Chair, FASEB Science Policy Committee)
Visiting Scientist, RCN4GSC Project
Center for Research in Biological Systems
University of California at San Diego

Daniel J. Bernard, PhD (FASEB Science Policy Committee)
Associate Professor of the Department of Pharmacology and Therapeutics
McGill University

James M. Musser, MD, PhD (FASEB Board of Directors, FASEB Research IT Subcommittee)
Executive Vice President and Co-Director
Fondren Foundation Distinguished Endowed Chair
Director of the Center for Molecular and Translational Human Infectious Diseases Research
Vice Chairman of the Department of Pathology
The Methodist Hospital Research Institute

Harel Weinstein, DSc (FASEB Science Policy Committee, FASEB Research IT Subcommittee)
Maxwell M. Upson Professor of Physiology and Biophysics
Chairman of the Department of Physiology and Biophysics
Director of the Institute for Computational Biomedicine
Weill Cornell Medical College of Cornell University

William York, PhD (FASEB Research IT Subcommittee)
Professor of Biochemistry and Molecular Biology
Adjunct Professor of Computer Science
Complex Carbohydrate Research Center
University of Georgia

FINANCIAL SERVICES ROUNDTABLE

March 31, 2014

Attn: Big Data Study
Office of Science and Technology Policy
Eisenhower Executive Office Building
1650 Pennsylvania Avenue NW
Washington, D.C. 20502

Re: "Big Data" – Request for Information

Ladies and Gentlemen:

The Financial Services Roundtable ("FSR")[1] is pleased to respond to the government's request for information concerning the collection, analysis and use of "big data" published in the Federal Register on March 4, 2014 (the "RFI") by the Office of Science and Technology Policy (the "Office").

Background and Overview

On January 17, 2014, President Obama called for a comprehensive review of how "big data," defined in the RFI as "datasets so large, diverse, and/or complex, that conventional technologies cannot adequately capture, store, or analyze them," will affect the everyday lives of Americans. The Office issued the RFI to facilitate that review and requested voluntary responses from both the public and private sector. The RFI poses five questions aimed at gathering responses on the implications of collecting, analyzing and using big data for privacy, the economy and public policy, with a focus on how

[1] As *advocates for a strong financial future*™, FSR represents 100 integrated financial services companies providing banking, insurance, and investment products and services to the American consumer. Member companies participate through the Chief Executive Officer and other senior executives nominated by the CEO. FSR member companies provide fuel for America's economic engine, accounting directly for $98.4 trillion in managed assets, $1.1 trillion in revenue, and 2.4 million jobs.

technological advances and broadening uses of such data can be maximized while minimizing the risks to privacy.

FSR and its members are strongly committed to protecting the privacy of Americans. We share the Office's view that big data can be used to "spur innovation and maximize the opportunities and free flow of this information," but that consumers must be provided with meaningful protections to ensure the privacy and security of data about them, including personal information. Our response to the RFI addresses this balance of interests, first, by providing an overview of the many ways in which financial institutions currently use certain data about their consumers to provide financial services (*i.e.*, from enhancing fraud prevention to complying with anti-money laundering regulations); and second, by summarizing the primary federal statutes and regulations and industry guidelines already in place governing how financial institutions collect, use, share and secure information about consumers.

This response follows on the heels of the March 27, 2013 meeting at the White House between representatives from the financial services industry and Administration officials. At that meeting, BITS (the technology policy division of FSR) and other financial services executives emphasized to Administration officials the importance of data analytics for the purposes of fraud reduction and cybersecurity, and discussed other direct and indirect benefits to consumers. There is no question that increased access to "big data" not only will combat fraud and improve security, but also will provide new insights and opportunities to improve financial products and customer relationships.

We welcome the Office's efforts to undertake a review of "big data." We note, however, that the concept of "big data" is an evolving one, and therefore, any questions, policies or frameworks that may be developed to address it should be formulated in ways that do not unnecessarily stymie its possible beneficial effects on society, individuals and the economy. Big data and enhanced data analytics, in general, can be used to strengthen national security, drive effective marketing, improve health care, enable a cleaner environment, and build safer cities. To the extent there are concerns about big data – whether it is the creepiness factor or that it may lead to profiling or discrimination – the financial services industry is vigilant about these concerns and operates not only in strict compliance with existing privacy and data security laws and regulations, but also works with BITS and other industry organizations to continually develop best practices for the industry.

We appreciate this opportunity to share our industry's experiences and expertise with the Office and look forward to being part of the government's continuing dialogue about big data in the future.

Overview of Uses of Consumer Data

In general, financial institutions collect, analyze and use data about consumers to provide better, more secure financial products to them. The data that is reviewed is not necessarily "big data," as defined in the RFI, but as big data becomes easier to access and manage, it undoubtedly will be used for the same purposes. An overview of some of the key ways in which consumer data is used today is provided below.

To Improve Access to Financial Products

Consumers today require quick access to banks, credit, and other financial services. In order to make rapid, reliable, and appropriate decisions about credit, insurance, and other consumer loans, financial institutions need to have ready access to a range of information about consumers. This information provides two downstream effects: *first*, it reduces the cost of financial services, and *second*, it increases the availability of those services. Banks are able to reduce costs by pooling consumer loans (securitization), practical only when accurate consumer information is available. Credit is provided based on historical consumer data including credit (FICO) scores, and is already highly regulated by the Fair Credit Reporting Act.

As more consumer data becomes available in the future (*e.g.*, in the form of "big data"), banks may be able to better gauge the creditworthiness of consumers, including those who have not yet established credit, by reviewing a broader array of relevant data and not relying solely on FICO scores. The data also may be used to create new financial products personalized to the consumer. In short, by using enhanced analytics, financial institutions will be able to better define and service their customers.

Enhancing Fraud Prevention and Customer Service

The ability of financial institutions to use big data to detect and prevent fraudulent activity saves billions of dollars each year for consumers and for financial institutions. In 2010, 73% of banks reported losses from check fraud, totaling around $893 million, but *attempted* check fraud amounted to around $11 billion.[2] Banks are estimated to have prevented around $13 billion in fraudulent transactions that would have affected consumers in 2012, in no small part because they have been able to use consumer data to spot these transactions early on.[3]

---

[2] Association for Financial Professionals, *2013 AFP Payments Fraud and Control Survey, available at* http://www.afponline.org/fraud/.

[3] American Bankers Association, *Banks Stop $13 Billion in Fraud Attempts in 2012, available at* http://www.aba.com/Press/Pages/121213DepositAccountFraud.aspx.

Financial institutions generally bear the burden of fraudulent transactions: they refund consumers and retailers affected by the fraud. To stem these losses and protect their consumers, they rely heavily on access to consumer transaction histories which allow them to detect and prevent fraudulent activity. By sharing consumer data with affiliates, they also are able to deter broader fraudulent activity across affiliate accounts.

Access to consumer data also allows financial institutions to provide better, more responsive customer service, including across affiliates. This can include not only helping customers when they have problems with their accounts, but also offering targeted or bundled services to customers with particular needs. *Compliance*

Financial institutions are subject to anti-money laundering regulations and other laws that require mandatory reporting of suspicious transactions. In particular, banks are required to notify the government of high-value currency transactions and similar suspicious activity. Access to consumer data can efficiently limit the occurrence of false positives when a bank checks suspicious names against a sanctions list. In addition, by responsibly monitoring customer activity over time, banks also can improve the accuracy of their reporting to the government.

Marketing

Financial institutions also use consumer data to identify the needs of their customers and ensure more relevant advertisements are reaching those customers. Targeted marketing can reduce unwanted or duplicative advertising, and engage consumers more efficiently. Consumers have the ability to control whether to receive such advertising by opting out of receiving emails, phone calls and direct mail solicitations.

Technological Trends in the Collection and Use of Big Data (Question 3)

Financial institutions collect consumer data directly from the consumer, from affiliates and from non-affiliates – with notice to the consumer – through a variety of "traditional" methods, including through the institution's website, at branches or other physical locations, and by phone. Due to technological advances, the types of information they are able to collect and the means by which they can collect it have expanded in recent years, as detailed below. The collected data, in turn, is used to provide better financial products and to improve customer relationships.

Mobile Applications and Social Media

Today virtually every major financial services institution offers mobile applications (*e.g.*, a mobile banking application), which offer convenience and accessibility to users. Mobile applications present a new opportunity to improve communication between customers and financial institutions, permitting more "real time"

interactions like balance notifications, potential fraudulent activity alerts, and other up-to-the-minute information. They also offer consumers a portable means of accessing their financial data. Data collected from mobile applications can include personal information, financial information and location data. Mobile privacy has received significant attention in recent years. The Federal Trade Commission (the "FTC") and California's Attorney General issued mobile privacy guidelines in 2013 to address the unique privacy concerns raised by mobile applications, including the collection of location data. Those guidelines serve as guide posts for the financial services and other industries.

Financial institutions also are increasingly engaging with consumers through social media platforms for marketing purposes, but social media is not a primary source for consumer information.

Location Data and Biometrics

The kinds of personal information available to financial institutions have expanded in recent years. A primary example is consumer location data, which is used to provide customer services (*e.g.*, to identify the location of nearby ATMs through a mobile banking app) and to detect possible fraud (*e.g.*, to verify transactions based on the location of the consumer).

Fingerprint recognition technology is also being used by banks in countries like Brazil to secure transactions and protect customers against fraud. However, further research and consideration of the associated privacy and security risks will be required before biometrics are adopted by the U.S. financial services industry in any meaningful way.

Online Behavioral Advertising

For marketing purposes, financial institutions today engage in some level of online behavioral advertising ("OBA"). OBA basically is advertising targeted to consumers based on their prior actions online. In the financial services context, OBA primarily takes the form of "retargeting" advertisements: consumers are shown ads for products or services they previously viewed online. Retargeting provides consumers with more relevant and useful advertising based on expressed needs, and can decrease the amount of unwanted and unnecessary advertising consumers see or receive.

Many financial institutions are members of the Digital Advertising Alliance (DAA)'s self-regulating program, which requires enhanced transparency and optimizes consumer choice with respect to OBA. The program allows consumers to opt out of their data being used for OBA by clicking on the "ad choices" icon, a universal symbol found near advertisements or on Internet pages where data is collected for OBA purposes.

Existing Privacy Laws Governing the Financial Services Sector (Questions 1, 3, and 5)

As noted above, banks and other financial institutions necessarily collect, analyze and use a significant amount of consumer information in the ordinary course of business. For that reason, in addition to privacy regulations applicable to all industries (*e.g.*, Section 5 of the FTC Act, which prohibits "unfair or deceptive acts or practices in or affecting commerce", and similar state laws), the financial services sector has long been subject to a set of specific federal and state laws that regulate how personal information may be collected, used, shared and secured by financial institutions.

Importantly, the laws are in place to protect the consumer and seek to accomplish this primarily through transparency and notice. Under the existing legal framework, financial institutions have affirmative disclosure obligations to ensure that consumers are aware of the types of information that are being collected and how that data may be used or shared by financial institutions. Consumers are also provided with meaningful choice as to how that data may be used or shared by affiliated or unaffiliated entities (*e.g.*, through opt-out notices). Financial institutions also provide customers with the option to limit email, telephone and direct mail solicitations.

The federal laws are reinforced by various U.S. state law requirements as well as industry best practices. Nearly all states have enacted laws that regulate the collection and use of consumer credit and financial data as well as laws requiring data breach notification. And some states, like California, afford even greater privacy protections to the financial information of consumers. Through its partnership with organizations like BITS, the financial services industry also has developed and implemented data security best practices. Together, these laws and standards establish a comprehensive framework for maintaining the highest standards of protection and privacy for consumer data.

- *The Gramm-Leach-Bliley Financial Modernization Act of 1999 ("GLBA")*[4]

  The GLBA is the primary law governing the privacy of consumer financial information. First, financial institutions covered by the GLBA are required to adopt privacy policies and make their information-sharing practices transparent to customers in annual privacy notices. The privacy policy must plainly inform consumers and customers of what information is collected, identify with whom the information will be shared, and describe how that information will be protected. Second, the GLBA generally prohibits financial institutions from sharing nonpublic and personally identifiable financial information with unaffiliated third parties, unless the customer receives notice and opportunity to opt-out. Lastly, the GLBA requires financial institutions to develop, implement

[4] 15 U.S.C. § 6801 *et seq.*

and maintain a "comprehensive information security program" designed to safeguard customer data.

- *The Fair Credit Reporting Act of 1970 ("FCRA")*[5]

  The FCRA regulates the practices of consumer reporting agencies that compile consumer information used by companies, including financial institutions, to make credit, employment, or insurance decisions affecting consumers. The FCRA also regulates the users of that consumer report information. Financial institutions may only use consumer report information for the purposes specified in the statute. Depending on the proposed use of the information, certain disclosures are required either before obtaining this information, in connection with using the information to take adverse action, or both. Consumers may opt out of the sharing of certain information between affiliates. And in the marketing context, there are rules about pre-screened offers for credit or insurance, restrictions on the sharing of information between affiliates for marketing purposes, and mechanisms for consumer choice.

- *The Fair and Accurate Credit Transactions Act of 2003 ("FACTA")*[6]

  FACTA, which substantially amended the FCRA, enhanced consumer protections by requiring federal agencies to adopt affiliate marketing, disposal, and identity theft red flag rules. The affiliate marketing provisions of FACTA generally prohibit companies from using consumer information received by an affiliate to make marketing solicitations, unless the consumer is provided with clear and conspicuous notice and the opportunity to opt out. Importantly, the rules apply to information that is otherwise excluded from the scope of "consumer report" information under the FCRA. The Disposal Rule protects against unauthorized access or use of consumer information and obligates companies to securely dispose of information in consumer reports. Financial institutions must incorporate disposal practices into the information security program required by the GLBA Safeguards Rule. Finally, under the Identity Theft Red Flag Rule, financial institutions and creditors that hold any consumer account for which there is a reasonably foreseeable risk of identity theft must implement programs designed to detect, prevent, and mitigate these risks.

[5] 15 U.S.C. § 1681 *et seq*.

[6] Pub. L. No. 108-159, 117 Stat. 1952 (Dec. 4, 2003).

- *The California Financial Information Privacy Act ("SB1")*[7]

  California state privacy laws are widely considered the most comprehensive and stringent of the state financial privacy laws. SB1 imposes obligations on financial institutions operating in its jurisdiction that are stricter than those provided for under federal law. Namely, SB1 defines identifiable information more broadly than federal law, requires opt-in as opposed to opt-out consent under certain circumstances and contains stricter limitations on the sharing of covered information with affiliates. For example, affirmative opt-in consent is required under California law before financial institutions may share covered information with nonaffiliated third parties. An opt-out opportunity must also be provided to consumers before financial institutions share covered information with affiliates in different lines of business.

- *BITS Cybersecurity and Fraud Reduction Best Practices*

  As the technology policy division of FSR, BITS addresses issues at the intersection of financial services, technology and public policy, where industry cooperation serves the public good, such as cybersecurity, critical infrastructure protection, fraud prevention, and the safety of financial services and its consumers. BITS, which was formed in 1996, works with subject matter experts from within its 100 member companies in each of the areas noted to develop best practices related to safe and sound computing, the protection of consumer information and protection of its members and their consumers from cyber attacks and fraud schemes. (See more at: http://www.bits.org)

- *Federal Financial Institutions Examination Council ("FFIEC") Guidance*

  The Federal Financial Institutions Examination Council, or FFIEC, is a government organization that works to promote uniform supervision of financial institutions. The FFIEC has issued a number of data security guidance documents, including standards for authentication that recommend the use of multi-factor identification or other means of identifying consumers (including biometric templates) to increase security and prevent unauthorized access.[8] The FFIEC guidance statements represent evolving best practices and are another helpful mechanism for ensuring the application of uniform, sufficient controls for safeguarding consumer data in a rapidly changing landscape.

---

[7] Cal. Fin. Code §§ 4050-5060.

[8] FFIEC, "Security Controls Implementation: Authentication," *available at* http://ithandbook.ffiec.gov/it-booklets/information-security/security-controls-implementation/access-control-/authentication.aspx.

- *The Financial Services Information Sharing & Analysis Center ("FS-ISAC") Data Security Standards*[9]

  Another key component critical to safeguarding sensitive consumer information held by financial institutions is collaboration and information sharing among industry members and between industry and the government. To that end, FS-ISAC was formed in 1999 to facilitate partnership between the public and private sectors working to defend the nation's critical infrastructures from cyber threats. There are thousands of member institutions primarily consisting of large financial services firms. The FS-ISAC model allows members to share threat, vulnerability, and incident information anonymously to protect the sector as a whole. It also developed best practices for mitigating system risks, as well as the development and testing of crisis management procedures.

Conclusion

Access to big data – whether it is personal information collected from the consumer or information about their transaction histories collected from third parties – is crucial for the provision of financial services and the security of consumers. Perhaps more than any sector, the financial services industry has had to balance these important interests against the risks of minimizing consumer privacy. We believe that the existing legal framework governing the financial services sector, including data best practices adopted by the industry, accomplish just that through various mandatory notice obligations and security standards. We would be happy to provide the Office with any additional information as it proceeds with its work of framing the main questions and policy concerns surrounding big data.

Thank you for the opportunity to respond to the RFI. If you have any questions, please feel free to contact me at (202) 589-2424.

Respectfully submitted,

Richard Foster
Vice President & Senior Counsel for
Regulatory and Legal Affairs
Financial Services Roundtable

[9] *See* "Industry Best Practices," *available at* https://www.fsisac.com/news/industry_best_practices.

March 31st, 2014

John Podesta, Counselor to the President
Nicole Wong, Deputy Chief Technology Officer, OSTP
Big Data Study, Office of Science and Technology Policy
Eisenhower Executive Office Building
1650 Pennsylvania Ave. NW.
Washington, DC 20502

Dear Mr. Podesta and Ms. Wong:

We, the undersigned, are organizations concerned about the proliferation of marketing of unhealthful foods and beverages that targets children and adolescents. We are dedicated to eliminating harmful food marketing—particularly marketing aimed at those who are most vulnerable to obesity and other nutrition-related diseases—by actively identifying, investigating, and advocating changes to marketing practices that undermine health.

As organizations concerned about the youth obesity epidemic in the United States, we urge the OSTP, in its forthcoming "Big Data" report to the President, to recommend safeguards to ensure young people are protected from data-driven marketing practices that could adversely affect their health and potentially undermine the progress the country has achieved so far in addressing obesity.

As you know, children and adolescents are growing up today in an era where they are always connected online—especially with mobile devices. Contemporary marketing relies on a set of "Big Data" practices that can target youth across media platforms and devices, creating digital profiles on their behaviors, location, interests, race, or ethnicity. Food and beverage companies are at the forefront in the use of an array of these powerful practices, which target young people on social networks, online videos, mobile phones, and games. We are concerned that without proper safeguards to address this new era of marketing to youth, the initial promising progress the country has made addressing the obesity crisis could be undermined.

We know the Administration, in its 2012 Consumer Privacy Bill of Rights blueprint, acknowledged that children and adolescents should receive special protections against unfair data collection practices. The Federal Trade Commission, in its own 2012 privacy report, also noted the need to ensure adolescents receive appropriate protection. As public health, child advocacy, research, and consumer organizations, we commend the Administration for this important inquiry into the impact of the use of personal information by the commercial sector. We respectfully urge the Administration to incorporate the following principles designed to protect the welfare of young people:

Recognize children's and teen's data as sensitive information requiring special safeguards - The Children's Online Privacy Protection Act and the proposals to treat teen information as sensitive data should be reflected in the White House report.

Ensure data collection minimization– Food and beverage companies should be encouraged to collect and use the least amount of data if they market to children and teens. They should make all their data collection practices transparent, permitting parents and other stakeholders to assess the scope and potential impact of these techniques.

No "Big-Data" oriented marketing of nutrient-poor foods and beverages - Companies should not use current data practices to promote unhealthy foods and beverages.

Safeguards on predictive analytics used to profile and target youth – Big Data has offered companies incredible insight into the preferences and intent of consumers. The use of these powerful analytic tools on children and youth is unfair and takes advantage of many of their developmental vulnerabilities. Food and beverage companies should confirm that they will not use predictive analytics to profile and target youth.

Sincerely,

African American Collaborative Obesity Research Network
American Academy of Child and Adolescent Psychiatry
American Academy of Sports Dietitians and Nutritionists
Berkeley Media Studies Group
Campaign for a Commercial-Free Childhood
CANFIT
Center for Digital Democracy
Center for Global Policy Solutions
Center for Science in the Public Interest
ChangeLab Solutions
Children Now
Common Sense Media
Consumer Federation of America
Consumers Union
Eat Drink Politics
Food Sleuth, LLC
Interfaith Center on Corporate Responsibility
Momsrising
National Consumers League
Northwest Coalition for Responsible Investment
Partnership for Prevention

The Praxis Project
Prevention Institute
Public Citizen
Public Health Advocacy Institute
Salud America!
Shape Up America!
Yale Rudd Center for Food Policy and Obesity

March 31, 2014

Via e-mail: bigdata@ostp.gov

Nicole Wong, Esq.
Big Data Study
Office of Science and Technology Policy
Eisenhower Executive Office Building
1650 Pennsylvania Avenue NW
Washington, DC 20502

Re: Public Comments, Big Data RFI

Dear Ms. Wong:

The Future of Privacy Forum (FPF) is a think tank seeking to advance responsible data practices and is supported by leaders in business, academia, and consumer advocacy.[1] FPF submits these Comments in response the White House Office of Science and Technology Policy (OSTP) Request for Information (RFI) dated March 4, 2014. In the RFI, the OSTP seeks public comment on how best to ensure innovation and maximize the opportunities and free flow of big data while minimizing any risks to privacy.[2]

Unlocking the value of data and instituting responsible data practices go hand-in-hand, and both have been an important focus of FPF's work since our founding in 2008. FPF recognizes the enormous potential benefits to consumers and to society from sophisticated data analytics,[3] yet FPF also understands that taking advantage of big data may require evolving how we implement traditional privacy principles. Through our work on inter-connected devices and applications and the emerging Internet of Things, FPF has acquired experience with the technologies involved in data collection and use. FPF appreciates this opportunity to provide Comments and share its insights into how best to promote the benefits of big data while minimizing any resulting privacy risks or harms.

Responding to the President's call to review how big data is changing our society, OSTP's Big Data Review has been a helpful exercise in soliciting thought leadership from academics, researchers, and industry. There is much than can be done to promote innovation in a way that advances privacy, and we are pleased to provide our recommendations. Specifically, we recommend that the OSTP Big Data Review report:

---

[1] The views herein do not necessarily reflect those of the Advisory Board or supporters of the Future of Privacy Forum.
[2] White House Office of Science and Technology Policy, Government "Big Data": Request for Information, 79 Fed. Reg. 12,251 (Mar. 4, 2014).
[3] Omer Tene & Jules Polonetsky, *Big Data for All: Privacy and User Control in the Age of Analytics*, 11 NW. J. TECH. & INTELL. PROP. 239, 243-51 (2013).

1) **Embrace a flexible application of Fair Information Practice Principles (FIPPs).** Traditional FIPPs have guided privacy policy nationally and around the globe for more than 40 years, and the White House Consumer Privacy Bill of Rights is the most recent effort to carry these principles forward into a world of big data. FPF supports the continued reliance on the FIPPs and believes they remain flexible enough to address many of the challenges posed by big data when applied in a practical, use-based manner. Our Comments recommend a nuanced approach to their applicability that accounts for modern day technical realities.

2) **Promote the benefits of big data in society.** Researchers, academics, and industry have demonstrated how big data can be useful in driving economic growth, advancing public safety and health, and improving our schools. Yet, privacy advocates and the public appear skeptical of these benefits in the face of certain outlier uses. More work is needed to understand the ways big data is already improving society and making businesses more efficient and innovative. This report should highlight the importance of big data's benefits and identify additional opportunities to promote positive uses of big data.

3) **Support efforts to advance practical de-identification, including policy and technological solutions.** While the Federal Trade Commission (FTC) has acknowledged that data that is effectively de-identified poses no significant privacy risk, there remains considerable debate over what effective de-identification requires. FPF believes that technical anonymization measures are only one component of effective de-identification. Instead, a broader understanding that takes into account how administrative and legal safeguards, as well as whether data is public or non-public, should inform conversations about effective de-identification procedures.

4) **Encourage additional work to frame context and promote enhanced transparency.** The context in which data is collected and used is an important part of understanding individuals' expectations, and context is a key principle in both the Consumer Privacy Bill of Rights and the FTC Privacy Framework. Respect for context is an increasingly important privacy principle, yet more work by academics, industry, and policymakers is needed about how to properly frame and define this principle. The Department of Commerce-led Internet Policy Task Force (IPTF) should continue its work convening stakeholders and hold programs that could help frame context in an age of big data. At the same time, another important tool that can be used to promote public trust in big data is enhanced transparency efforts. In particular, FPF has called for more transparency surrounding high-level decisional criteria that organizations may use to make decisions about individuals.

5) **Encourage efforts to promote accountability by organizations working with big data.** Data privacy frameworks increasingly rely on organizational accountability to ensure responsible data stewardship. In the context of big data, FPF supports the further development of the concept of internal review boards that could help companies weigh the benefits and risks of data uses. In conjunction with the evolving role of the privacy professional, accountability measures can be put in place to ensure big data projects take privacy considerations into account.

6) **Promote government leadership on big data through its own procedures and practices.** The federal government is one of the largest producers and users of data, and, as

a result, the government may inform industry practice and help demonstrate the value of data through its own uses of big data across and among agencies. The Federal Chief Information Officer (CIO) Council is particularly well-positioned to ensure the federal government can maximize the potential of big data with an eye toward privacy protection.

7) **Promote global efforts to facilitate interoperability.** Recent privacy developments in the Asia Pacific and the European Union have given new life to constructive collaboration on the cross jurisdictional issues presented by big data. FPF urges government to actively promote and maintain existing frameworks to facilitate interoperability, including the US-EU Safe Harbor and the Asia Pacific Economic Cooperation's (APEC) Cross Border Privacy Rules (CBPR) System**.**

These broad next steps are suggested as a helpful beginning to the work that needs to be done. In the remainder of this submission, we respond to the questions posed in the RFI.

(1) What are the public policy implications of the collection, storage, analysis, and use of big data?

Big data may be one of the biggest public policy challenges of our time.[4] The debate surrounding big data will ask policy makers to pit compelling interests such as national security, public health and safety, and sustainable development against risks to personal autonomy from high-tech profiling and discrimination, increasingly-automated decision making, inaccuracies and opacity in data analysis, and strains in traditional legal protections.[5] However, the traditional Fair Information Practice Principles (FIPPs) remain flexible enough to address many of these concerns when applied in a practical, use-based manner. What is needed is additional research on the benefits of big data and on how to advance practical de-identification and other measures to protect privacy.

I. Big Data and the Fair Information Practice Principles

There is considerable dispute today over how best to properly calibrate the FIPPs to protect privacy and encourage innovative uses of data. On one hand, some increasingly suggest that foundational privacy practices such as a notice and choice and purpose limitation are either impractical or less relevant due to big data and other emerging technologies.[6] While privacy advocates and regulators recognize limitations with our notice and choice framework, they worry that big data may provide an excuse to override individual rights in order to facilitate intrusive marketing or ubiquitous surveillance. FPF would caution against disposing of sound principles that have guided privacy policy for more than forty years. Our Comments advocate for a nuanced approach, based upon a practical application of the FIPPs that accounts for modern day technical realities around collection and use of personal data.

[4] Jules Polonetsky, Omer Tene & Christopher Wolf, *How To Solve the President's Big Data Challenge*, IAPP PRIVACY PERSPECTIVES (Jan. 31, 2014), https://www.privacyassociation.org/privacy_perspectives/post/how_to_solve_the_presidents_big_data_challenge.
[5] Civil Rights Principles for the Era of Big Data, http://www.civilrights.org/press/2014/civil-rights-principles-big-data.html (last visited March 15, 2013).
[6] For example, the growing network of smart, connected devices known as the "Internet of Things" is commonly understood to rely upon the capture, sharing, and use of data, including data about who we are and what we do at any given moment. *See, e.g.,* Bill Wasik, *Welcome to the Programmable World*, WIRED (May 14, 2013), http://www.wired.com/gadgetlab/2013/05/internet-of-things/.

In their various formulations, the FIPPs establish core principles guiding the collection, use, and disclosure of data.[7] Some of the most important FIPPs are: (1) Notice – individuals should be provided with timely notice of how their data will be collected, used, and disclosed; (2) Choice – individuals should be given choices about whether and how their data will be used; (3) Purpose Specification – the purposes for which personal data are collected should be specified prior to or at the time of collection; and (4) Use Limitation – personal data should only be used for those purposes specified prior to or at the time of collection; and (5) Data Minimization – organizations should seek to limit the amount of personal data they collect and that might be retained.[8] These principles are each challenged by big data in different ways.

The White House Consumer Privacy Bill of Rights has recognized this challenge. Based on the FIPPs, the general principles put forward by the Administration's privacy framework explicitly afford companies discretion in how they implement them. This flexibility was designed both to promote innovation and to "encourage effective privacy protections by allowing companies, informed by input from consumers and other stakeholders, to address the privacy issues that are likely to be most important to their customers and users, rather than requiring companies to adhere to a single, rigid set of requirements."[9]

A. Notice and Choice: The Need for Flexibility

Notice is often considered the most "fundamental" principle of privacy protection.[10] Yet there is wide acknowledgement that a privacy framework based on notice and choice has significant limitations.[11] The vast majority of consumers do not read privacy policies,[12] and further, studies have shown that consumers make privacy decisions not based on policies but rather on the context in which they are presented by a use of their data.[13] In the age of big data, the implementation of notice and choice through detailed privacy policies may only result in the publication of even more unread policies. Furthermore, notice and choice presents particular problems for connected devices or other "smart" technologies that will not be equipped with interactive screens or other easily accessible user interfaces. Information collected in "public" spaces and used for data analytics may also prove problematic.

---

[7] The FIPPs generally have been thought of as establishing high-level guidelines for promoting privacy. They do not establish specific rules prescribing how organizations must protect privacy in all contexts, but rather they provide principles that can inform the implementation of specific codes of practice. Initially proposed in a 1973 advisory committee report for the Department of Housing, Education, and Welfare, the FIPPs emerged due to concern about the increased use of personal data in record-keeping systems. Subsequently, the FIPPs have become the basis of global privacy law and remain relevant in a world of big data.

[8] *See* Christopher Wolf & Jules Polonetsky, *An Updated Privacy Paradigm for the "Internet of Things"* (2013), http://www.futureofprivacy.org/wp-content/uploads/Wolf-and-Polonetsky-An-Updated-Privacy-Paradigm-for-the-"Internet-of-Things"-11-19-2013.pdf.

[9] WHITE HOUSE, CONSUMER DATA PRIVACY IN A NETWORKED WORLD: A FRAMEWORK FOR PROTECTING PRIVACY AND PROMOTING INNOVATION IN THE GLOBAL DIGITAL ECONOMY 2 (2012), http://www.whitehouse.gov/sites/default/files/privacy-final.pdf [hereinafter WHITE HOUSE BLUEPRINT].

[10] FEDERAL TRADE COMMISSION, PRIVACY ONLINE: A REPORT TO CONGRESS 7 (1998).

[11] Fred Cate, *Looking Beyond Notice and Choice*, PRIVACY & SECURITY LAW REPORT (Mar. 29, 2010), http://www.hunton.com/files/Publication/f69663d7-4348-4dac-b448-3b6c4687345e/Presentation/PublicationAttachment/dfdad615-e631-49c6-9499-ead6c2ada0c5/Looking_Beyond_Notice_and_Choice_3.10.pdf (citing Former FTC Chairman Jon Liebowitz conceding that the "notice and choice" regime offered by the FIPPs hasn't "worked quite as well as we would like.").

[12] Aleecia M. McDonald & Lorrie Faith Cranor, *The Cost of Reading Privacy Policies*, 4 I/S: J. L. & POL'Y FOR INFO. SOC'Y 543 (2008).

[13] *See, e.g.*, Alessandro Acquisti et al., What Is Privacy Worth? 27-28 (2010) (unpublished manuscript), *available at* http://www.heinz.cmu.edu/~acquisti/papers/acquisti-ISR-worth.pdf.

Although technological solutions may help to facilitate notice and choice options, it will be impractical to premise data collection and use in the world of big data and other emerging technologies based solely on traditional implementations of notice and choice. For that matter, a number of data applications may not require any integration of privacy protections. Consider a smart TV that learns the volume preferences of particular users and adjusts its volume accordingly – if that information is not transmitted out of the TV it should raise few issues. Most machine-to-machine communications of contextually aware devices also make notice and choice unnecessary, especially if any information flows for these devices are contained.

That said, privacy policies still have value. They remain helpful as accountability and enforcement mechanisms: they set the boundaries for data use by businesses beyond those that might be prescribed by law and they create enforceable legal obligations. Disclosure requirements by themselves can force companies to evaluate their privacy practices and instill discipline in how they treat consumer information.[14]

Flexibility will be especially important with respect to the concepts of notice and choice. There remains much uncertainty over what information matters for disclosure, and this problem is only exacerbated by big data. Organizations need to think creatively about how to provide consumers with meaningful insight into commercial data practices, and regulators and policymakers should encourage these efforts.[15]

This challenge, however, is recognized by the Consumer Privacy Bill of Rights, which recommends organizations seek innovative ways to provide consumers with more individual control, and if that remains impractical, organizations should embrace and augment other FIPPs or elements of the Consumer Privacy Bill of Rights in order to adequately protect consumer privacy.[16] The OSTP Big Data Review report should call for renewed efforts by industry to develop new models to inform individuals about the collection and use of their personal data. Techniques to inform consumers of data practices might include symbols, short phrases, colors, diagrams, or any of the tools otherwise available to designers seeking to provide users with an engaging user experience. Engaging consumers about data use should be viewed as an essential feature and a core part of the user experience. In the end, design features that "communicate" information to users may be more helpful than traditional notice models.

B. Purpose Specification & Use Limitation: Context Is Key

One of the exciting challenges presented by big data is that much of the new value from data is being discovered in surprising ways.[17] Consider the innovations pioneered by the United Nations Global Pulse that are enabled by the analysis of mobile phone data. Global Pulse has helped us understand mobility, social interaction and economic activity.[18] By analyzing mobile interactions,

---

[14] *See* Peter P. Swire, *The Surprising Virtues of the New Financial Privacy Law,* 86 MINN. L. REV. 1263, 1314 (2002).

[15] Testimony Before the California State Assembly Joint Committee Hearing on Privacy (Dec. 12, 2103) (statement of Jules Polonetsky, Executive Director, Future of Privacy Forum, at 3), *available at* http://www.futureofprivacy.org/wp-content/uploads/CA-Assembly-Hearing-Privacy-Policies_Does-Disclosure-Transparency-Adequately-Protect-Consumers-Privacy-Final.pdf (citing Susanna Kim Ripken, *The Dangers and Drawbacks of the Disclosure Antidote: Toward a More Substantive Approach to Securities Regulation*, 58 BAYLOR L. REV. 139, 147 (2006) (calling for regulators to "lay aside the gospel of disclosure in favor of more substantive laws that regulate conduct directly")).

[16] WHITE HOUSE BLUEPRINT, *supra* note 9, at 13.

[17] *E.g.*, VIKTOR MAYER-SCHÖNBERGER & KENNETH CUKIER, BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK (2013).

[18] Robert Kirkpatrick, Beyond Targeted Ads: Big Data for a Better World (2012), *available at*

UN researchers were able to examine the post-earthquake population migration caused by the Haiti earthquake. Global Pulse has been able to track the spread of disease and better understand socio-economic activity in a number of countries around the world. However, under our traditional privacy frameworks, some valuable uses of data may be constrained.

Most privacy regimes endorse a principle of use limitation, which is generally implemented by requiring that personal information be used *only* as specified at the time of collection.[19] Most of the innovative secondary uses of information – including breakthroughs in medicine, data security, or energy usage – are impossible to anticipate when notice is first provided, often long before a new benefit is uncovered through data analysis.[20] Companies can neither provide notice for a purpose that is yet to exist, nor can consumers provide informed consent for an unknown.[21]

However, these principles may be implemented by instead limiting the use of information based upon the *context* in which it is collected.[22] Often, context is understood to mean that personal information should be used only in ways that individuals would expect given the context in which information was disclosed and collected. However, there are uses of data that may be outside individual expectations but have high societal value and minimal privacy impact that should be encouraged. More work is needed to define and frame context.

C. Data Minimization: Moving Toward More Accountability Measures

While it has been overshadowed by the principles of notice and choice,[23] data minimization has long been another important traditional privacy practice.[24] Data minimization promotes privacy by limiting the amount of personal information in circulation.[25] Yet it is not clear that minimizing information collection is always a practical approach to privacy in the age of big data.[26] Almost by definition, "big" data requires a significant amount of data to be available in order to discern previously unnoticed patterns and trends. As the Consumer Privacy Bill of Rights notes, "wide-ranging data collection may be essential for some familiar and socially beneficial internet services and applications."[27] These uses, as well as many others yet to be developed, would be stymied if companies were required to limit the amount of data they collect.

---

http://www.slideshare.net/unglobalpulse/strata-14934034.

[19] *See, e.g.*, European Parliament and Council Directive 95/46/EC - on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data, 1995 O.J. (L 281) 31, *available at* http://www.refworld.org/docid/3ddcc1c74.html (last visited Mar. 15, 2014).

[20] SCHÖNBERGER & CUKIER, *supra* note 17, at 153.

[21] *Id.*

[22] Wolf & Polonetsky, *supra* note 8, at 9.

[23] Fred Cate, *The Failure of the Fair Information Practice Principles* 15 (2009), http://www.informationpolicycentre.com/files/Uploads/Documents/Centre/Failure_of_Fair_Information_Practice_Principles.pdf

[24] *See* OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data, ORG. FOR ECON. CO-OPERATION & DEV. (Sept. 23, 1980), http://www.oecd.org/document/18/0,3343,en_2649_34255_1815186_1_1_1_1,00.html. Data minimization involves limiting an organization's collection of personal data to the minimum extent necessary to obtain specified and legitimate goals. The principle further instructs organizations to delete data that is no longer used for the purposes for which it was originally collected, and to implement restrictive policies with respect to the retention of personal data in identifiable form.

[25] With less data to process and analyze, many believe that companies will have less capability to use data in new, privacy-invasive ways – and consumers will be protected from unwarranted access to their information. *See, e.g.*, Justin Brookman & G.S. Hans, *Why Collection Matters* (2013), http://www.futureofprivacy.org/wp-content/uploads/Brookman-Why-Collection-Matters.pdf.

[26] Omer Tene & Jules Polonetsky, *Privacy in the Age of Big Data: A Time for Big Decisions*, 64 STAN. L. REV. ONLINE 63 (2012), http://www.stanfordlawreview.org/online/privacy-paradox/big-data.

[27] WHITE HOUSE BLUEPRINT, *supra* note 9, at 21.

There is still a role for sensible retention policies and efforts to reasonably limit data collection should not be dismissed out-of-hand. However, concerns around data collection and use may be mitigated through additional accountability measures, such as internal controls and internal review boards, which we discuss below. Further, when organizations use adequately de-identified data sets, their use of that data mitigates privacy risk, which demonstrates how further research around de-identification could prove helpful within the context of big data. Thus, a more sophisticated analysis of data minimization should take into account the de-identification and other privacy safeguards that have been implemented.

II. Advancing Practical De-Identification Solutions

Clarifying the scope of information subject to privacy law has become an increasingly important policy question. During this review's "Advancing the State of the Art in Technology and Practice" workshop at MIT, the question of what information is properly de-identified or anonymous emerged throughout the day's discussion.[28] Personally identifiable information (PII) is one of the central concepts in information privacy regulation, but there is no uniform definition of PII.[29] Similarly, there is no standard for what constitutes adequate de-identification of PII.[30]

This is important because resolving the spectrum of PII and non-PII also addresses some of the concerns facing traditional FIPPs. As the FTC acknowledged in its March 2012 report, *Protecting Consumer Privacy in an Era of Rapid Change*, data that has been effectively de-identified does not raise significant privacy concerns.[31] However, laws often turn on whether or not information is PII or not, and this bi-polar approach based on labeling information either "personally identifiable" or not is not appropriate given the messiness of big data.[32] FPF proposes that PII instead be defined based on a risk matrix taking into account the risk, intent, and potential consequences of re-identification, as opposed to a dichotomy between "identifiable" and "non-identifiable" data.

De-identification should be understood as a process that takes into account legal and administrative safeguards, as well as technical measures, in order to protect privacy. Unfortunately, at the moment, much of our discourse around de-identification focuses on the technical possibility of re-identification and the assumption that all data will be made publicly available.[33] While computer scientists have repeatedly shown that anonymized data, either released publicly or poorly de-identified, can be re-identified, organizations and policymakers must recognize that non-public data presents a lessened privacy risk than information released publicly.

---

[28] Big Data Privacy Workshop: Advancing the State of the Art in Technology and Practice, http://web.mit.edu/bigdata-priv/ (last visited Mar. 15, 2014).

[29] Paul M. Schwartz & Daniel J. Solove, *The PII Problem: Privacy and a New Concept of Personally Identifiable Information*, 86 NYU L. REV. 1814 (2011).

[30] Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization,* 57 UCLA L. REV. 1701 (2010) (arguing that "scientists have demonstrated that they can often 'reidentify' or 'deanonymize' individuals hidden in anonymized data with astonishing ease." *But see* Daniel Barth-Jones, *Re-Identification Risks and Myths, Superusers and Super Stories*, CONCURRING OPINIONS (Sept. 6, 2012), http://www.concurringopinions.com/archives/2012/09/re-identification-risks-and-myths-superusers-and-super-stories-part-ii-superusers-and-super-stories.html (citing Ohm's suggestion that public policy should not "inappropriately conflat[e] the rare and anecdotal accomplishments of notorious hackers with the actions of typical users . . .")).

[31] FED. TRADE COMM'N, PROTECTING CONSUMER PRIVACY IN AN ERA OF RAPID CHANGE: RECOMMENDATIONS FOR BUSINESSES AND POLICYMAKERS 20-22 (2012), http://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-report-protecting-consumer-privacy-era-rapid-change-recommendations/120326privacyreport.pdf [hereinafter FTC PRIVACY REPORT].

[32] *See generally* SCHÖNBERGER & CUKIER, *supra* note 17, at 32-49 (suggesting the value of big data may require an approach to data analysis that is "comfortable with disorder and uncertainty.").

[33] Yianni Lagos & Jules Polonetsky, *Public vs. Nonpublic Data: The Benefits of Administrative Control*, 66 STAN. L. REV. ONLINE 103 (2013), http://www.stanfordlawreview.org/online/privacy-and-big-data/public-vs-nonpublic-data.

While any analysis of effective de-identification should consider the legal and administrative controls around data, there remains work to be done to advance technical de-identification measures. The Administration should support this effort through funding and by convening workshops that can help further de-identification research. This work could focus on de-identification techniques that maintain data utility for researchers and industry, and should help frame a de-identification debate that recognizes the value of non-technical safeguards.

III. Framing Context and Meaningful Transparency

A principle of respect for context relies on what individuals expect from their relationship with an organization. Consumers expect that companies will share their personal information with other companies to fulfill orders and that companies will use personal information to engage in first-party marketing.[34] When personal information is used in those ways or in others that individuals would reasonably expect, there is no privacy violation.

But respect for context can become difficult to meet when faced with innovative data practices.[35] Focusing solely on individual expectations not only hampers some benefits that could accrue to those individuals, but it also ignores that company-to-consumer relationships evolve. As the Consumer Privacy Bill of Rights recognizes, respect for context must admit that a relationship between an organization and an individual may change over time in ways not foreseeable at the time of collection, and that "such adaptive uses of personal data may be the source of innovations that benefit consumers."[36] Consider a company that collects personal fitness information from wearable sensors that track sleep, steps taken, pulse or weight. Analysis of such data, collected originally only to report basic details back to users, may yield unanticipated health insights that could be provided individually to users or used in the aggregate to advance medical knowledge. Rigidly and narrowly specifying context could trap knowledge that is available and critical to progress.

The challenge facing organizations and policymakers is that respect for context requires an appreciation for dynamic social and cultural norms.[37] Context includes not only an objective component, but also a number of subjective variables including an individual's level of trust and his perceived value from the use of his information.[38] Public-facing efforts to inform consumers about big data will be essential to provide individuals with more context around data practices. Companies could frame relationships by "setting the tone" for new products or novel uses of information. Even where new uses of data are contextually similar to existing uses, information and education are essential. Amazon serves as a prime example of this approach: its website is able to pursue a high degree of customization without violating consumer expectations, given its clear messaging about customization and its provision of a user interface frames how data is used.

---

[34] *See* WHITE HOUSE BLUEPRINT, *supra* note 9, at 16-17.
[35] Jules Polonetsky & Omer Tene, *It's Not How Much Data You Have, But How You Use It* 5 (2012), http://www.futureofprivacy.org/wp-content/uploads/FPF-White-Paper-Its-Not-How-Much-Data-You-Have-But-How-You-Use-It_FINAL1.pdf.
[36] WHITE HOUSE BLUEPRINT, *supra* note 9, at 16.
[37] Carolyn Nguyen, Director, Microsoft Technology Policy Group, Contextual Privacy, Address at the FTC Internet of Things Workshop (Nov. 19, 2013) (transcript available at http://www.ftc.gov/sites/default/files/documents/public_events/internet-things-privacy-security-connected-world/final_transcript.pdf).
[38] *Id.*

To complement a principle focused on respect for context, organizations must be much more transparent about how they are using data. Many of the concerns around big data applications center on worries about untoward data usage, and enhanced transparency may help alleviate fears that an individual's personal information is somehow being used against them. Transparency can be a tool that can help demystify big data.[39] For example, organizations should disclose the criteria underlying their decision-making processes to the extent possible without compromising their trade secrets or intellectual property rights. While there are practical difficulties in requiring these disclosures, distinctions can be drawn between sensitive, proprietary algorithms and high-level decisional criteria.

Further transparency efforts have been endorsed by both the Consumer Privacy Bill of Rights and the FTC's 2012 privacy report.[40] But transparency cannot simply mean better privacy policies. Instead, policymakers should encourage companies to engage with consumers in a meaningful conversation where both parties' interests and expectations can be aligned.[41] FPF has previously called for the "featurization" of data, transforming data analysis into a consumer-side application by granting individual access to their personal data in intelligible, machine-readable forms. Mechanisms such as personal clouds or data stores will allow individuals to contract with third-parties who would get permission to selectively access certain categories of their data to provide further analysis or value-added services.[42] "Featurization" will allow individuals to declare their own policies, preferences and terms of engagement, and do it in ways that can be automated both for them and for the companies they engage.[43]

IV. Enhanced Accountability through Internal Review Boards

Big data may warrant a shift in focus toward accountability mechanisms that ensure organizations are responsibly managing personal information.[44] Several privacy scholars have suggested that our current privacy framework stresses mere compliance, when emphasizing institutional accountability may be more necessary to promote better data stewardship.[45] While there are many strategies to augment accountability in the age of big data, it will be important for organizations to engage in a practical balancing of privacy considerations and data use.

A formalized review mechanism could help to review and approve innovative data projects.[46] Some have also called for big data "algorithmists" that could evaluate the selection of data sources, the choice of analytical tools, and the interpretation of any predictive results.[47] As organizations

---

[39] Tene & Polonetsky, *supra* note 3, at 270-72.
[40] WHITE HOUSE BLUEPRINT, *supra* note 9, at 16-17; FTC PRIVACY REPORT, *supra* note 31, at 60.
[41] *See, e.g.*, Omer Tene & Jules Polonetsky, *A Theory of Creepy: Technology, Privacy and Shifting Social Norms* , 16 YALE J.L. & TECH. 59 (2013).
[42] Tene & Polonetsky, *supra* note 3, at 263-70.
Jules Polonetsky, *Big Data for All: Privacy and User Control in the Age of Analytics*, 11 NW. J. TECH. & INTELL. PROP. 239, 243-51 (2013).
[43] The rise of privacy and reputation management vendors points to a future where organizations will be able to unlock additional value by collaborating with their users. One interesting first step is the launch of "About the Data" by Acxiom, the nation's largest data broker. "About the Data" is a consumer-facing tool that gives individuals control over certain categories of information (such as personal characteristics, interests, and finances) gathered by Acxiom. The site allows consumers to correct information, suppress any data they see, or opt-out of Acxiom's marketing profile system altogether via an approachable user interface.
[44] *See* Fred H. Cate & Viktor Mayer-Schoenberger, Data Use and Impact Global Workshop (2013), http://cacr.iu.edu/sites/cacr.iu.edu/files/Use_Workshop_Report.pdf.
[45] *Id.* at 5.
[46] Ryan Calo, *Consumer Subject Review Boards: A Thought Experiment*, 66 STAN. L. REV. ONLINE 97 (2013).
[47] SCHÖNBERGER & CUKIER, *supra* note 17, at 180.

increasingly face interesting new proposals for using data, these professionals could operate across the public and private sectors and conduct cost-benefit analyses of data uses.

Industry increasingly faces ethical considerations over how to minimize data risks while maximizing benefits to all parties. Formal review processes may serve as an effective tool to infuse ethical considerations into data analysis. Institutional review boards (IRBs) were the chief regulatory response to decades of questionable ethical decisions in the field of human subject testing; big data internal review boards could similarly serve as a proactive response to concerns regarding data misuse. In many respects, these review boards would be a further expansion of the role of privacy professionals within the industry today. While creating internal review boards would present a unique set of challenges, encouraging companies to create sophisticated structures and personnel to grapple with these issues and provide oversight would be invaluable.

Any successful approach to big data must be guided by a cost-benefit analysis that takes into account exactly how the benefits of big data will be distributed. So far, our procedural frameworks are largely focused on traditional privacy risks and assessing what measures can be taken to mitigate those risks. In 2010, for example, the Department of Commerce's Internet Policy Task Force endorsed the use of privacy impact assessments (PIAs) both to help organizations decide whether it is appropriate to engage in innovative data uses and to identify alternative approaches that could reduce relevant privacy risks.[48] However, human research IRBs also take into account anticipated benefits and even the importance of any knowledge that may result from research.[49]

FPF is exploring a framework to provide a similar accounting of the rewards of big data. While organizations and privacy professionals have developed expertise at evaluating risk, we believe decision makers need better processes to evaluate how to assess, prioritize, and to the extent possible, quantify a project's potential benefits. To that end, we intend to propose a methodology that assesses a project's value based upon several criteria, including culture-specific values and probability of success. FPF is eager to engage in further discussion about how best to develop big data review mechanisms.

(2) What types of uses of big data could measurably improve outcomes or productivity with further government action, funding, or research? What types of uses of big data raise the most public policy concerns? Are there specific sectors or types of uses that should receive more government and/or public attention?

One of the chief challenges facing big data analytics is determining how much of "big data" is new, or whether it is simply another buzzword. There have been detailed efforts to show how the emerging technologies fueled by big data are reshaping society.[50]

I. The Social Ramifications of Predictive Analytics

The fundamental problem posed by big data may be less a question of how it impacts our privacy and more that it upsets our society's sense of fairness. The debate around big data is often couched

[48] U.S. DEP'T OF COMMERCE, INTERNET POL'Y TASK FORCE, COMMERCIAL DATA PRIVACY AND INNOVATION IN THE INTERNET ECONOMY: A DYNAMIC POLICY FRAMEWORK 34-35 (2010), http://www.commerce.gov/sites/default/files/documents/2010/december/iptf-privacy-green-paper.pdf.
[49] 45 CFR § 46.111.
[50] *E.g.*, RICK SMOLAN & JENNIFER ERWITT, THE HUMAN FACE OF BIG DATA (2012).

as something that implicates traditional privacy principles and that the use and inferences drawn from our data invade our privacy, but this obscures the larger public policy challenge. The real concerns presented by big data are increasingly abstract or inchoate risks that may have very little to do with privacy practices per se.

A. Shifting User Expectations: "Creepiness"

Since the revelation that Target was able to predict a teenager's pregnancy before her family was even aware of it, much of the concern surrounding predictive analytics is that it is somehow "creepy." The Target example is illustrative precisely because it did not involve any explicit breach of the FIPPs, and was not explicitly harmful to the consumer in question. Instead, it was a novel use of data that dramatically upset individual expectations. Disrupting one's expectations can lead to unexpected benefits, but it also limits an individual's ability to feel comfortable or in control.[51]

While creepiness is inherently subjective, creepy behaviors are detrimental to the development of any trust-based relationship – whether between friends, consumer and company, or government and citizen.[52] Due to a lack of trust, individuals have been quick to dismiss the benefits of big data as a result of some of its more surprising results. The use of big data is outpacing the evolution of social norms, and addressing creepiness may simply be a matter of organizations exploring how to set the right tone when seeking intimate relationships with individuals via their data.[53] Self-regulatory frameworks are likely to be the best mechanism to address shifting cultural norms, and industry should be encouraged to be proactive in this regard.

B. Potential Concerns Surrounding Civil Liberties

There are other broad concerns about big data that are outside the ambit of traditional privacy law. Recently, critics, including some of the United States' leading civil rights organizations, have argued that big data could be the "civil rights" issue of this generation.[54] In particular, there are worries that big data undermines equal opportunity and equal justice through hidden or new forms of discrimination.[55] Big data could achieve these harms by contributing to currently illegal practices, allowing otherwise unlawful activity to go undetected due to a lack of transparency or access surrounding data analysis.[56] Alternatively, big data may introduce societal biases that may impact protected classes or otherwise vulnerable populations disproportionately or unfairly.[57]

The United States has enacted a series of powerful legislative remedies to combat discrimination in the context of employment, education, housing, and credit worthiness. These laws specifically

---

[51] Francis T. McAndrew & Sara S. Koehnke, (On the Nature of) Creepiness, Poster presented at the annual meeting of the Society for Personality and Social Psychology (SPSP), Jan 18, 2013, *available at* http://www.academia.edu/2465121/Creepiness ("Being 'creeped out' is an evolved adaptive emotional response to ambiguity about the presence of threat that enables us to maintain vigilance during times of uncertainty.").

[52] *See generally id.*

[53] *See* Omer Tene & Jules Polonetsky, *A Theory of Creepy: Technology, Privacy and Shifting Social Norms* , 16 YALE J.L. & TECH. 59 (2013).

[54] Alistair Croll, *Big Data Is Our Generation's Civil Rights Issue, and We Don't Know It*, SOLVE FOR INTERESTING (July 31, 2012), http://solveforinteresting.com/big-data-is-our-generations-civil-rights-issue-and-we-dont-know-it/; Civil Rights Principles for the Age of Big Data, *supra* note 5.

[55] Civil Rights Principles for the Age of Big Data, *supra* note 5.

[56] Pam Dixon, *On Making Consumer Scoring More Fair and Transparent*, IAPP PRIVACY PERSPECTIVES (Mar. 19, 2014), https://www.privacyassociation.org/privacy_perspectives/post/on_making_consumer_scoring_more_fair_and_transparent.

[57] See Kate Crawford, *The Hidden Biases in Big Data*, HBR BLOG NETWORK (Apr. 1, 2013), http://blogs.hbr.org/2013/04/the-hidden-biases-in-big-data/.

protect access to certain opportunities by prohibiting organizations from taking into account certain factors. For example, Title VII of the Civil Rights Act of 1964 prohibits employers from discriminating against applicants and employees on the basis of race, color, religion, sex, and national origin.[58] The Equal Credit Opportunity Act forbids creditors from asking about a candidate's marital status or plans to have children.[59] Our antidiscrimination laws have even had to take new technologies into account: the Genetic Information Nondiscrimination Act of 2008, for example, prohibits employers from using an applicant's or an employee's genetic information as the basis of an employment decision, and it also limits the ability of health insurance organizations to deny coverage based solely on a genetic predisposition to develop a disease.[60]

Antidiscrimination laws are not truly data privacy laws, however. Instead, they work to address classifications and decisions that society has deemed either irrelevant or illegitimate. Big data makes it easier to discover an individual's race, religion, or gender, and it also encourages ever more granular categorization and segmentation of individuals. The challenge is that there are no clear guidelines as to where value-added personalization and segmentation – which may provide positive consumer benefits in some cases – turn into harmful discrimination.[61] We already have legal protections in place that would prohibit the use of big data to engage in certain kinds of specific discrimination. Further academic and expert analysis is necessary in order to understand which of the claimed data practices are already illegal and merely need additional enforcement and which create new uses that warrant further policy analysis.

However, the bigger question is whether big data may impact individuals or classes of individuals in ways that we might deem unfair. Take the example of an Atlanta man who returned from his honeymoon to find his credit limit slashed from $10,800 to $3,800 because he had used his credit card at locations where *others* were likely to have a poor repayment history.[62] Is this an efficient use of data analysis or fundamentally unfair? Are there industry practices or remedies that could avoid problems and concerns? If new data practices are deemed unfair, illegitimate, or have a disparate impact against an already vulnerable group, more work is necessary to understand how best to address these practices. These issues are already receiving priority attention at the FTC, which in 2012 entered into an $800,000 settlement with Spokeo for marketing personal information to employers in violation of the Fair Credit Reporting Act,[63] and more recently, held a workshop on the use of predictive analytics to create consumer scores.[64] FPF and other groups could play a positive role by further exploring these issues.

C. Transparency and Opacity: Filter Bubbles and Surveillance

Another more inchoate concern presented by big data is that it may allow institutions to know more about an individual than that individual even knows. Even if organizations have the best of

[58] 42 U.S.C. § 2000e-2.
[59] 15 U.S.C. § 1691.
[60] Pub.L. 110–233, 122 Stat. 881
[61] Michael Schrage, *Big Data's Dangerous New Era of Discrimination*, HBR BLOG NETWORK (Jan. 29, 2014), http://blogs.hbr.org/2014/01/big-datas-dangerous-new-era-of-discrimination/.
[62] *See* Lori Andrews, *Facebook Is Using You*, N.Y. TIMES (Feb. 4, 2012), http://www.nytimes.com/2012/02/05/opinion/sunday/facebook-is-using-you.html.
[63] Press Release, Fed. Trade Comm'n., Spokeo to Pay $800,000 to Settle FTC Charges Company Allegedly Marketed Information to Employers and Recruiters in Violation of FCRA (June 12, 2012), http://www.ftc.gov/news-events/press-releases/2012/06/spokeo-pay-800000-settle-ftc-charges-company-allegedly-marketed.
[64] Fed. Trade Comm'n., Spring Privacy Series: Alternative Scoring Products (Mar. 19, 2014), http://www.ftc.gov/news-events/events-calendar/2014/03/spring-privacy-series-alternative-scoring-products.

intentions, the knowledge gained from analysis of big data can quickly lead to over-personalization. Individuals are more easily segmented, classified, and potentially placed into "filter bubbles" at the expense of their autonomy.[65] This also produces a more segregated society.

Feelings of being surveilled, which can arise from the continuous collection and use of our information, may also impact how people behavior, causing a chilling effect on civil discourse. For example, pervasive web tracking presents the possibility that people may avoid certain searches or sources of information out of fear that accessing that information would reveal interests, medical conditions, or other characteristics they would prefer be kept hidden.[66] Combined with a lack of transparency about how this information is being used, individuals may feel anxiety over consequential decisions about them being made opaquely, inducing a sense of powerlessness.[67]

The challenge is that some of the new risks associated with big data are not easily mapped to recognizable harms or are difficult to link to accepted privacy risks. To what degree they even present real challenges to society remains an open question. More work is needed by researchers to determine how these abstract concerns should inform any big data privacy analysis.

II. A Positive Role for Government

The White House has previously shown significant commitment to using big data to advance the national good, and it can do more to highlight these benefits and alleviate any concerns. Government can play a pivotal role as a convener, encouraging the benefits of the big data while promoting efforts within industry to advance the framework presented by the Consumer Privacy Bill of Rights. At the same time, it can also provide leadership and guidance on big data through its own procedures and practices.[68] The federal government is one of the biggest producers of big data. More than $200 million was committed in 2012 as part of a National Big Data Research and Development Initiative.[69] Last year, the Administration continued to be active in convening multiple stakeholders to explore big data applications that can improve economic growth, job creation, education, health, energy, sustainability, public safety, advanced manufacturing, science and engineering, and global development.[70]

The federal government, through Open Data efforts and day-to-day interactions with the public, generate massive amounts of data. In 2012, for example, the President launched an initiative entitled Digital Government: Building a 21st Century Platform to Better Serve the American People.[71] The aims of the initiative are to offer the public access to government information and

[65] *E.g.*, ELI PARISER, THE FILTER BUBBLE: HOW THE NEW PERSONALIZED WEB IS CHANGING WHAT WE READ AND HOW WE THINK (2012).
[66] Felix Wu, *Big Data Threats* 2 (2013), http://www.futureofprivacy.org/wp-content/uploads/Wu-Big-Data-Threats.pdf.
[67] *Id.*
[68] Adelaide O'Brien, Iron Mountain, *The Impact of Big Data on Government* (Oct. 2012), http://www.ironmountain.com/Knowledge-Center/Reference-Library/View-by-Document-Type/White-Papers-Briefs/Sponsored/IDC/The-Impact-of-Big-Data-on-Government.aspx.
[69] Press Release, White House Office of Science & Tech., Obama Administration Unveils "Big Data" Initiative: Announces $200 Million in New R&D Investments (Mar. 29, 2012), http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release.pdf.
[70] Big Data Senior Steering Group, The Networking and Information Technology Research and Development Program, http://www.nitrd.gov/nitrdgroups/index.php?title=Big_Data_(BD_SSG)#title (last visited Mar. 15, 2014).
[71] White House, Digital Government: Building a 21st Century Platform to Better Serve the American People, http://www.whitehouse.gov/sites/default/files/omb/egov/digital-government/digital-government.html (last visiting Mar. 15, 2014).

services anytime, anywhere, on any device, and to better leverage the rich wealth of federal data by promoting open data and machine-readable information.

Increasingly, government agencies are also using data in more creative ways. The Consumer Financial Protection Bureau (CFPB) has been especially active in using data analytics tools, arguing that "a 21st-century agency should use 21st-century tools."[72] In order to fulfill its mission and effectively monitor financial practices, the bureau has gathered vast amounts of consumer finance data on information varying from overdraft fees to credit card add-on products, and it is building databases that will integrate consumer credit information with loan and property records.[73] Yet, the CFPB also serves as an example of some the key privacy challenges that may come with innovative data use. The bureau has received criticism over concerns about its transparency and accountability when it comes to using individual financial data to police bank behavior.[74]

While critics are quick to argue that any of big data's benefits are underwhelming when weighed against potential harms, [75] the federal government can demonstrate the benefits of big data in a way that protects and promotes privacy. In particular, it can do this by supporting the development of big data tools and augmented accountability mechanisms across government. For example, the Federal Chief Information Officer (CIO) Council is well-positioned to ensure the federal government can maximize the potential of big data with an eye toward privacy protection.[76] The government can also do more to provide additional definition and framing as to how privacy principles like transparency and accountability can be used to alleviate concerns about big data.

Demonstrating how big data can be used to improve government function and services is only one component of the government's role, however. While the CIO Council can coordinate action within government, IPTF, again, can also bring agencies together to advance public policy that promotes big data.[77] Agencies can do more to support and highlight how big data is being used to benefit consumers in the private sector. Public/private partnerships can advance the necessary work to support innovation and advance privacy. IPTF should continue its work convening stakeholders and advancing further discussion about big data.

(3) What issues are raised by the use of big data across jurisdictions, such as the adequacy of current international laws, regulations, or norms?

As previously mentioned, FPF believes that big data may require a focus on accountability mechanisms to ensure responsible data stewardship across borders. To this end, FPF has advocated strengthening the existing US-EU Safe Harbor agreement and believes its continuation should be a

[72] Rebecca Sausner, *Warren's CFPB Embraces Big Data*, AM. BANKER (Dec. 1, 2010), http://www.americanbanker.com/btn/23_12/warrens-cfpb-embraces-big-data-1029410-1.html.
[73] Karuna Mintaka Kumar, *CFPB Tangles with Bankers over Big Data*, PYMNTS (July 19, 2013), http://www.pymnts.com/uncategorized/2013/cfpb-tangles-with-bankers-over-big-data/.
[74] *E.g.*, Op-Ed, *Consumer Financial Cover-Up*, WALL ST. J. (Mar. 17, 2014), http://online.wsj.com/news/articles/SB10001424052702303795904579431484040822904 ; Carter Dougherty, *Richard Cordray and the CFPB Are Monitoring Your Banking Habits*, BLOOMBERG BUSINESSWEEK (Apr. 25, 2013), http://www.businessweek.com/articles/2013-04-25/richard-cordray-and-the-cfpb-are-monitoring-your-banking-habits.
[75] For a discussion critiquing the benefits of big data, *see* Paul Ohm, *The Underwhelming Benefits of Big Data*, 161 U. PA. L. REV. ONLINE 339 (2013), http://www.pennlawreview.com/online/161-U-Pa-L-Rev-Online-339.pdf.
[76] CIO.gov, https://cio.gov/about/ (last visited Mar. 15, 2014).
[77] Nat'l Telecomm. & Info. Admin., Internet Pol'y Task Force, http://www.ntia.doc.gov/category/internet-policy-task-force (last visited Mar. 26, 2014).

top priority in the context of international data transfers.[78] More recently, the United States, in conjunction with representatives from the 21-nation Asia Pacific Economic Cooperation (APEC) and the EU's Article 29 Working Party announced the endorsement of a common referential.[79] This jointly-endorsed document identifies points of commonality under the APEC Cross Border Privacy Rules (CBPR) System and the EU's system of Binding Corporate Rules (BCRs). Each of these initiatives incorporates accountability mechanisms and clearly demonstrates the concept's utility in the context of international data transfers.

The Administration has previously committed to pursuing international interoperability through the mutual recognition of commercial data privacy frameworks that incorporate both effective enforcement and accountability mechanisms.[80] However, realizing the full potential of interoperability requires sustained senior-level engagement at the Department of Commerce. FPF urges the United States to use all levers of diplomatic, policy and regulatory activities across a range of international venues to achieve interoperability of these frameworks and consensus on their applicability in the era of big data.

Conclusion

Big data presents many benefits and potential risks. A thoughtful, balanced analysis of the value choices now at hand is essential. The Administration's efforts to convene thought leaders have produced many fruitful conversations, and more are needed. At the same time, it will be essential that the Administration provide transparency and a clear plan of action to all stakeholders moving forward.

Big data offers the United States a great opportunity to provide global leadership on promoting innovation – and protecting privacy. It also presents a challenge, but we have the privacy principles and frameworks needed to thoughtfully address that task.

FPF thanks the White House Office of Science and Technology Policy for considering these Comments, and we look forward to further engagement and collaboration on the issue of big data.

Sincerely,

Jules Polonetsky
Director and Co-Chair
Future of Privacy Forum

Christopher Wolf
Founder and Co-Chair
Future of Privacy Forum

Josh Harris
Policy Director
Future of Privacy Forum

Joseph Jerome
Policy Counsel
Future of Privacy Forum

[78] FUTURE OF PRIVACY FORUM, THE US-EU SAFE HARBOR: AN ANALYSIS OF THE FRAMEWORK'S EFFECTIVENESS IN PROTECTING PERSONAL PRIVACY (2013), http://www.futureofprivacy.org/wp-content/uploads/FPF-Safe-Harbor-Report.pdf

[79] Joint work between experts from the Article 29 Working Party and from APEC Economies, on a referential for requirements for Binding Corporate Rules submitted to national Data Protection Authorities in the EU and Cross Border Privacy Rules submitted to APEC CBPR Accountability Agents (Mar. 7, 2014), *available at* http://www.apec.org/~/media/Files/Groups/ECSG/20140307_Referential-BCR-CBPR-reqs.pdf.

[80] WHITE HOUSE BLUEPRINT, *supra* note 9, at 31.

Subject: [Big Data RFI]
To: White House Office of Science and Technology Policy, bigdata@ostp.gov
From: Frank Pasquale, Professor of Law at University of Maryland Francis King Carey School of Law.

His research agenda focuses on challenges posed to information law by rapidly changing technology, particularly in the health care, internet, and finance industries.

RFI Question (1) What are the public policy implications of the collection, storage, analysis, and use of big data?

The collection and analysis of "big data" has considerable implications for demographic-based discrimination, and productive research aims. The question for policymakers is how to discourage the former and encourage the latter.

For example, if you are childless, shop for clothing online, spend a lot on cable TV, and drive a minivan, data brokers are probably going to assume you're heavier than average.[1] We know that drug companies may use that data to recruit research subjects. Marketers could utilize the data to target ads for diet aids, or for types of food that research reveals to be particularly favored by people who are childless, shop for clothing online, spend a lot on cable TV, and drive a minivan.

But the data can be put to darker purposes: for example, to offer credit on worse terms to the obese (stereotype-driven assessment of looks and abilities is used from Silicon Valley to experimental labs).[2] And perhaps someday it will be put to higher purposes: for example, identifying "obesity clusters" that might be linked to overexposure to some contaminant.[3]

A rough ranking of these goals might be:

(1) Curing illness or precursors to illness (identifying the obesity cluster; clinical trial recruitment)

(2) Helping match those offering products to those wanting them (food marketing)

(3) Promoting the classification and de facto punishment of certain groups (identifying a certain class as worse credit risks)

---

[1] Joseph Walker, *Data Mining to Recruit Sick People: Companies Use Information From Data Brokers, Pharmacies, Social Networks*, WALL ST. J. (Dec. 17, 2013, 4:32 PM), http://online.wsj.com/news/articles/SB10001424052702303722104579240140554518458.

[2] Danielle Citron, Bright Ideas: *Deborah Rhode's The Beauty Bias*, CONCURRING OPINIONS (Apr. 19, 2010), www.concurringopinions.com/archives/2010/04/bright-ideas-deborah-rhodes-the-beauty-bias.html; Peter Dizikes, *Study: Attractive Men Fare Best In Gaining Venture Capital*, MITNEWS (Mar. 17, 2014), http://web.mit.edu/newsoffice/2014/study-says-attractive-men-fare-best-in-gaining-venture-capital.html; Frank Pasquale, *Decomposing Pulchritude's Perks*, MADISONIAN.NET (Nov. 9, 2006), http://madisonian.net/2006/11/09/decomposing-pulchritudes-perks/.

[3] Tom Philpott, *Can Antibiotics Make You Fat*?, MOTHER JONES (Jan. 2, 2013, 4:01 AM), www.motherjones.com/environment/2013/12/can-antibiotics-make-you-fat.

Current law and policy does not do enough to recognize how valuable goals like (1) are, and how destructive (3) could become. In fact, to the extent (1) is highly regulated, and (3) is unregulated, law may perversely help channel capital into discriminatory ventures and away from socially productive ones. We need to update anti-discrimination law and policy (such as the Fair Credit Reporting Act (FCRA), codified at 15 U.S.C. § 1681 et seq., and the Equal Credit Opportunity Act (ECOA), codified at 15 U.S.C. § 1691 et seq.) to account for how big data is being used, and new laws and policies need to distinguish between innovation and discrimination.

RFI Question (2) What types of uses of big data could measurably improve outcomes or productivity with further government action, funding, or research? What types of uses of big data raise the most public policy concerns?

We need to update anti-discrimination law and policy. New laws and policies need to distinguish between innovation and discrimination.[4]

Reputation intermediaries outside the health sector are now using data not covered by HIPAA to impute health conditions to individuals.[5] As the former CIO of Google (& CEO of ZestFinance) puts it, "[A]ll data is credit data, we just don't know how to use it yet." A lawyer might respond: "all data is health data," too, and should be subject to HIPAA and HITECH strictures.

If a firm finds out that the obese (or people with minivans) are worse credit risks, and imposes a higher interest rate on them, I question whether that is "innovation" as valuable as, say, finding better ways of curing a disease, growing food, or cooking a meal. It may, instead, merely be a way for industry to arrogate to itself a quasi-juridical role of punishing one group and forcing them to generate more rents for the finance sector.

Currently our laws fail to incentivize socially beneficial innovation. Our innovation (and privacy) law must recognize that a cancer cure is of greater value than a tool that helps companies avoid hiring people who are likely to have cancer. A recent review of Julia Angwin's excellent book "Dragnet Nation" (a muckraking take on privacy) concluded that its "lack of a more radical critique of digital capitalism may say more about the scope of the problem than our paucity of solutions."[6]

One approach would be to tax the profits of big data users and use that money toward endeavors beneficial to the entire pubic, and not just a company's profits. Tarleton Gillespie offers one way of doing so:

---

[4] *See* Scott R. Peppet, *Regulating the Internet of Things: First Steps Toward Managing Discrimination, Privacy, Security & Consent*, 82 TEX. L. REV (forthcoming 2014), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2409074.

[5] Nicolas P. Terry, *Big Data Proxies and Health Privacy Exceptionalism*, HEALTH MATRIX (forthcoming 2014), *available at* www.ftc.gov/system/files/documents/public_comments/2014/01/00007-88437.pdf.

[6] Jacob Silverman, *'Dragnet Nation' Looks At The Hidden Systems That Are Always Looking At You*, L.A. TIMES (Mar. 6, 2014), http://www.latimes.com/books/jacketcopy/la-ca-jc-julia-angwin-20140309,0,3129669.story.

> "The third party data broker who buys data from an e-commerce site I frequent, or scrapes my publicly available hospital discharge record, or grabs up the pings my phone emits as I walk through town [is] building commercial value on my data, but offer me no value to me, my community, or society in exchange. So what I propose is a "pay it back tax" on data brokers. . . .
>
> "If a company collects, aggregates, or scrapes data on people, and does so not as part of a service back to those people . . . then they must grant access to their data and access 10% of their revenue to non-profit, socially progressive uses of that data. This could mean they could partner with a non-profit, provide them funds and access to data, to conduct research. Or, they could make the data and dollars available as a research fund that non-profits and researchers could apply for. Or, as a nuclear option, they could avoid the financial requirement by providing an open API to their data. . . . . I think there could be valuable partnerships: Turnstyle's data might be particularly useful for community organizations concerned about neighborhood flow or access for the disabled; health data could be used by researchers or activists concerned with discrimination in health insurance. There would need to be parameters for how that data was used and protected by the non-profits who received it, and perhaps an open access requirement for any published research or reports." [7]

Gillespie's proposal addresses core problems of our increasingly big data driven (and intermediary driven) economy: law's agnosticism as to the ultimate productive value of what innovators are doing.[8] Yiren Lu recently asked, "Why do . . . smart, quantitatively trained engineers, who could help cure cancer or fix healthcare.gov, want to work for a sexting app?"[9] The answer is pretty obvious: the money. If we develop an elaborate set of laws that channels billions of dollars to at best reallocative (and at worst, flat out discriminatory) endeavors, we shouldn't be surprised when tech talent flocks to them.[10]

If we want entrepreneurs to use big data for higher ends, we have to change the incentives. The question is not: "should the U.S. have an industrial policy for big data?"—we already have a highly dysfunctional one. We should, instead, focus on improving returns for those who contribute to real gains in productivity.

---

[7] Tarleton Gillespie, *A "Pay It Back Tax" On Data Brokers*, CULTURE DIGITALLY (Mar. 18, 2014), http://culturedigitally.org/2014/03/a-pay-it-back-tax-on-data-brokers/.

[8] Frank Pasquale, *A "Content Loss Ratio" for Cable Companies?*, MADISONIAN.NET (Jan. 4, 2010), http://madisonian.net/2010/01/04/a-content-loss-ratio-for-cable-companies/

[9] Yiren Lu, *Silicon Valley's Youth Problem*, N.Y. TIMES (Mar. 12, 2014), www.nytimes.com/2014/03/16/magazine/silicon-valleys-youth-problem.html?_r=0

[10] PAUL KEDROSKY & DANE STANGLER, FINANCIALIZATION AND ITS ENTREPRENEURIAL CONSEQUENCES (Mar. 2011), *available at* www.kauffman.org/~/media/kauffman_org/research%20reports%20and%20covers/2011/03/financialization_report_32311.pdf.

Subject: [Big Data RFI]
To: White House Office of Science and Technology Policy, bigdata@ostp.gov
From: A submission from Frank Pasquale, excerpting key sections from Frank Pasquale & Danielle Citron, *The Scored Society,* 89 WASHINGTON LAW REV. 1 (2014).

Pasquale and Citron are Professors of Law at University of Maryland Francis King Carey School of Law.

RFI Question (2) What types of . . . uses should receive more government and/or public attention?

Our key recommendations are on page 4. To introduce them:

Big Data is increasingly mined to rank and rate individuals. Predictive algorithms assess whether we are good credit risks, desirable employees, reliable tenants, valuable customers—or deadbeats, shirkers, menaces, and "wastes of time." Crucial opportunities are on the line, including the ability to obtain loans, work, housing, and insurance. Predictive algorithms are increasingly rating people in countless aspects of their lives. Consider these examples. Job candidates are ranked by what their online activities say about their creativity and leadership.[1] Software engineers are assessed for their contributions to open source projects, with points awarded when others use their code.[2] Individuals are assessed as likely to vote for a candidate based on their cable-usage patterns.[3] Recently released prisoners are scored on their likelihood of recidivism.[4]

In one area of particular concern for policy makers should be credit scoring. While there are currently regulations of credit, the outdated laws focus on credit history, not the derivation of scores from big data. Evidence suggests that what is supposed to be an objective aggregation and assessment of data—the credit score—is arbitrary and has a disparate impact on women and minorities. Critiques of credit scoring systems come back to the same problem: the secrecy of their workings and growing influence as a reputational metric. Scoring systems cannot be meaningfully checked because their technical building blocks are trade secrets.

We are not completely in the dark though about credit scores' impact. Evidence suggests that credit scoring does indeed have a negative, disparate impact on traditionally disadvantaged groups.[5] Concerns about disparate impact have led many states to regulate the use of credit

---

1. *See* Don Peck, *They're Watching You at Work*, ATLANTIC MONTHLY, Dec. 2013, at 72, 76.

2. *See* E. GABRIELLA COLEMAN, CODING FREEDOM 116–22 (2013) (exploring Debian open source community and assessment of community members' contributions).

3. *See* Alice E. Marwick, *How Your Data Are Being Deeply Mined*, N.Y. REV. BOOKS, Jan. 9, 2014, at 22, 22.

4. Danielle Keats Citron, *Data Mining for Juvenile Offenders*, CONCURRING OPINIONS (Apr. 21, 2010, 3:56 PM), http://www.concurringopinions.com/archives/2010/04/data-mining-for-juvenile-offenders.html.

5. BIRNY BIRNBAUM, INSURERS' USE OF CREDIT SCORING FOR HOMEOWNERS INSURANCE IN OHIO: A REPORT TO THE OHIO CIVIL RIGHTS COMMISSION 2 (2003) ("Based upon all the available information, it is our opinion that insurers' use of insurance credit scoring for underwriting,

scores in insurance underwriting.[6] The National Fair Housing Alliance (NFHA) has criticized credit scores for disadvantaging women and minorities.[7] Insurers' use of credit scores has been challenged in court for their disparate impact on minorities. After years of litigation, Allstate agreed to a multi-million dollar settlement over "deficiencies in Allstate's credit scoring procedure which plaintiffs say resulted in discriminatory action against approximately five million African-American and Hispanic customers."[8] As part of the settlement, Allstate allowed plaintiffs' experts to critique and refine future scoring models.[9]

How are these scores developed? Predictive algorithms mine personal information to make guesses about individuals' likely actions and risks.[10] A person's on- and offline activities are turned into scores that rate them above or below others.[11] Private and public entities rely on predictive algorithmic assessments to make important decisions about individuals.[12]

And there is far more to come. Algorithmic predictions about health risks, based on information that individuals share with mobile apps about their caloric intake, may soon result in higher insurance premiums.[13] Sites soliciting feedback on "bad drivers" may aggregate the information, and could possibly share it with insurance companies who score the risk potential of insured individuals.[14]

The scoring trend is often touted as good news. Advocates applaud the removal of human beings and their flaws from the assessment process. Automated systems are claimed to rate all individuals in the same way, thus averting discrimination. But this account is misleading. Because human beings program predictive algorithms, their biases and values are embedded into the software's instructions, known as the source code and predictive algorithms.[15] Scoring systems mine datasets containing inaccurate and biased information provided by people.[16] There

---

rating, marketing and/or payment plan eligibility very likely has a disparate impact on poor and minority populations in Ohio.").

6. *Credit-Based Insurance Scoring: Separating Facts from Fallacies*, NAMIC POL'Y BRIEFING (Nat'l Ass'n of Mut. Ins. Cos., Indianapolis, Ind.), Feb. 10, at 1, *available at* http://iiky.org/documents/NAMIC_Policy_Briefing_on_Insurance_Scoring_Feb_2010.pdf.

7. *The Future of Housing Finance: The Role of Private Mortgage Insurance: Hearing Before the Subcomm. on Capital Mkts., Ins. & Gov't Sponsored Enters. of the H. Comm. on Fin. Servs.*, 111th Cong. 16 (2010) (statement of Deborah Goldberg, Hurricane Relief Program Director, The National Fair Housing Alliance). The NFHA has expressed concern that "the use of credit scores tends to disadvantage people of color, women, and others whose scores are often lower than those of white borrowers." *Id.* at 57. The NFHA has also expressed "growing concern about how useful credit scores are for predicting loan performance and whether the financial sector is placing too much reliance on credit scores rather than other risk factors such as loan terms." *Id.*

8. Dehoyos v. Allstate, 240 F.R.D. 269, 275 (W.D. Tex. 2007). The parties settled after the Fifth Circuit decided that federal civil rights law was not reverse preempted by the McCarran-Ferguson Act's allocation of insurance regulatory authority to states. *See* Dehoyos v. Allstate Corp., 345 F.3d 290, 299 (5th Cir. 2003). The Equal Credit Opportunity Act (ECOA), which regulates lending practices, does not preempt state laws that are stricter than ECOA.

9. *Dehoyos*, 240 F.R.D. at 276.

10. Frank Pasquale, *Restoring Transparency to Automated Authority*, 9 J. ON TELECOMM. & HIGH TECH. L. 235, 235–36 (2011).

11. Hussein A. Abdou & John Pointon, *Credit Scoring, Statistical Techniques and Evaluation Criteria: A Review of the Literature*, 18 INTELLIGENT SYSTEMS ACCT. FIN. & MGMT. 59, 60–61 (2011).

12. *See* Marwick, *supra* note 3, at 24; *see also* Jack Nicas, *How Airlines Are Mining Personal Data In-Flight*, WALL ST. J., Nov. 8, 2013, at B1.

13. *See* Marwick, *supra* note 3, at 24.

14. *See* Frank Pasquale, *Welcome to the Panopticon*, CONCURRING OPINIONS (Jan. 2, 2007), http://www.concurringopinions.com/archives/2007/01/welcome_to_the_14.html.

15. Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1260–63 (2008).

16. *Id.*; Danielle Keats Citron, *Open Code Governance*, 2008 U. CHI. LEGAL F. 355, 363–68 [hereinafter Citron, *Open Code Governance*].

is nothing unbiased about scoring systems.

Supporters of scoring systems insist that we can trust algorithms to adjust themselves for greater accuracy. In the case of credit scoring, lenders combine the traditional three-digit credit scores with "credit analytics," which track consumers' transactions. Suppose credit-analytics systems predict that efforts to save money correlates with financial distress. Buying generic products instead of branded ones could then result in a hike in interest rates. But, the story goes, if consumers who bought generic brands also purchased items suggesting their financial strength, then all of their purchases would factor into their score, keeping them from being penalized from any particular purchase.

Does everything work out in a wash because information is seen in its totality? We cannot rigorously test this claim because scoring systems are shrouded in secrecy. Although some scores, such as credit, are available to the public, the scorers refuse to reveal the method and logic of their predictive systems.[17] No one can challenge the process of scoring and the results because the algorithms are zealously guarded trade secrets.[18] As this Article explores, the outputs of credit-scoring systems undermine supporters' claims. Credit scores are plagued by arbitrary results. They may also have a disparate impact on historically subordinated groups.

Just as concerns about scoring systems are more acute, their human element is diminishing. Although software engineers initially identify the correlations and inferences programmed into algorithms, Big Data promises to eliminate the human "middleman" at some point in the process.[19] Once data-mining programs have a range of correlations and inferences, they use them to project new forms of learning. The results of prior rounds of data mining can lead to unexpected correlations in click-through activity. If, for instance, predictive algorithms determine not only the types of behavior suggesting loan repayment, but also automate the process of learning which adjustments worked best in the past, the computing process reaches a third level of sophistication: determining *which* metrics for measuring past predictive algorithms were effective, and recommending further iterations for testing.[20] In short, predictive algorithms may evolve to develop an artificial intelligence (AI) that guides their evolution.

RFI Question (1) What are the public policy implications of the collection, storage, analysis, and use of big data? For example, do the current U.S. policy framework and privacy proposals for protecting consumer privacy and government use of data adequately address issues raised by big data analytics?

The current regulatory models failure to regulate credit scoring mechanisms using big

---

17. Tal Zarsky, *Transparent Predictions*, 2013 ILL. L. REV. 1503, 1512.

18. Evan Hendricks, *Credit Reports, Credit Checks, Credit Scores*, A.B.A. GPSOLO, July/Aug. 2011, at 32, 34.

19. Chris Anderson, *The End of Theory: The Data Deluge Makes Scientific Inquiry Obsolete*, WIRED (June 23, 2008), http://www.wired.com/science/discoveries/magazine/16-07/
pb_theory.

20. A pioneer of artificial intelligence described this process in more general terms: "In order for a program to improve itself substantially it would have to have at least a rudimentary understanding of its own problem-solving process and some ability to recognize an improvement when it found one. There is no inherent reason why this should be impossible for a machine." Marvin L. Minsky, *Artificial Intelligence*, SCI. AM., Sept. 1966, at 246, 260.

data is rooted in credit scoring companies' immunity from defamation law.[21] By limiting the possible penalties for reputational injuries, the Fair Credit Reporting Act (FCRA) opened the door to tactics of stalling, obstinacy, and obfuscation by the credit industry.[22] As such, credit scoring happens in an unregulated "block box." The government doesn't know and can't control what is happening, and neither can the consumer monitor or seek redress for wrongs against them.

To correct these errors, our recommendations are as follows:

- Regulatory Oversight Scheme
    1. Transparency to Facilitate Testing
        a. The FTC should be given access to credit-scoring systems and other scoring systems that unfairly harm consumers
    2. Risk Assessment Reports and Recommendations
        a. Once the FTC evaluates credit-scoring systems to detect "arbitrariness-by-algorithm," it should issue a Privacy and Civil Liberties Impact Assessment evaluating a scoring system's negative, disparate impact on protected groups, arbitrary results, mischaracterizations, and privacy harms.
- Individual Protections
    1. Notice Guaranteed by Audit Trails
        a. Individuals are owed notice when governmental systems make adverse decisions about them.
    2. Interactive Modeling
        a. Give consumers the chance to see what happens to their score with different hypothetical alterations of their credit histories

A Brief Introduction

Increasing transparency of credit scoring systems is essential. Regulation accomplishing this would borrow from our due process tradition. A "technological due process" would introduce human values and oversight back into the picture. Scoring systems and the arbitrary and inaccurate outcomes they produce must be subject to expert review.

Regulators should be able to test scoring systems to ensure their fairness and accuracy. Individuals should be granted meaningful opportunities to challenge adverse decisions based on scores miscategorizing them. Without such protections in place, systems could launder biased and arbitrary data into powerfully stigmatizing scores.

Procedural regularity is essential given the importance of predictive algorithms to people's life opportunities—to borrow money, work, travel, obtain housing, get into college, and far more. Scores can become self-fulfilling prophecies, creating the financial distress they claim

---

21. SMITH, *supra* note **Error! Bookmark not defined.**, at 320. Note, though, that the FCRA is riddled with many exceptions, exceptions to exceptions, and interactions with state law.

22. 15 U.S.C. § 1681 et seq.; *see* Schramm-Strosser, *supra* note **Error! Bookmark not defined.**, at 170–71 ("What started out as an improvement over how the common law dealt with credit-reporting issues has evolved into a regulatory scheme that tends to favor the credit reporting industry . . . . One example of the FCRA's overly broad preemptive scope is the prohibition of injunctive relief for consumers who bring common law defamation claims against CRAs.").

merely to indicate.[23] The act of designating someone as a likely credit risk (or bad hire, or reckless driver) raises the cost of future financing (or work, or insurance rates), increasing the likelihood of eventual insolvency or un-employability.[24] When scoring systems have the potential to take a life of their own, contributing to or creating the situation they claim merely to predict, it becomes a *normative* matter, requiring moral justification and rationale.[25]

Scoring systems should be subject to fairness requirements that reflect their centrality in people's lives. Private scoring systems should be as understandable to regulators as to firms' engineers. However well an "invisible hand" coordinates economic activity generally speaking, markets depend on reliable information about the practices of firms that finance, rank, and rate consumers. Brandishing quasi-governmental authority to determine which individuals are worthy of financial backing, private scoring systems need to be held to a higher standard than the average firm.

How should we accomplish accountability? Protections could draw insights from what one of us has called "technological due process"—procedures ensuring that predictive algorithms live up to some standard of review and revision to ensure their fairness and accuracy.[26] Procedural protections should apply not only to the scoring algorithms themselves (a kind of technology-driven rulemaking), but also to individual decisions based on algorithmic predictions (technology-driven adjudication).

This is not to suggest that full due process guarantees are required as a matter of current law. Given the etiolated state of "state action" doctrine in the United States, FICO and credit bureaus are not state actors; however, much of their business's viability depends on the complex web of state supports and rules surrounding housing finance. Nonetheless, the underlying values of due process—transparency, accuracy, accountability, participation, and fairness[27]—should animate the oversight of scoring systems given their profound impact on people's lives. Scholars have built on the "technological due process" model to address private and public decision-making about individuals based on the mining of Big Data.[28]

---

23. *See* Michael Aleo & Pablo Svirsky, *Foreclosure Fallout: The Banking Industry's Attack on Disparate Impact Race Discrimination Claims Under the Fair Housing Act and the Equal Credit Opportunity Act*, 18 B.U. PUB. INT. L.J. 1, 5 (2008) ("Ironically, because these borrowers are more likely to default on their loans, the banks, to compensate for that increased risk, issue these borrowers loans that feature more onerous financial obligations, thus increasing the likelihood of default.").

24. *See id.*

25. This is part of a larger critique of economic thought as a "driver," rather than a "describer," of financial trends. *See generally* DONALD MACKENZIE, AN ENGINE, NOT A CAMERA: HOW FINANCIAL MODELS SHAPE MARKETS (2006) (describing how economic theorists of finance helped create modern derivative markets); Joel Isaac, *Tangled Loops: Theory, History, and the Human Sciences in Modern America*, 6 MOD. INTELL. HIST. 397, 420 (2009) ("[S]cholars are rejecting the traditional notion that economics attempts to create freestanding representations of market processes (which economic sociologists must then insist leaves out power, or cultural context, or the fullness of human agency)."). Some commentators have argued that we need to "recognize economics not as a (misguided) science of capitalism but as its technology, that is, as one of the active ingredients in the production and reproduction of the market order." Marion Fourcade, *Theories of Markets and Theories of Society*, 50 AM. BEHAV. SCI. 1015, 1025 (2007).

26. *See generally* Citron, *supra* note 15.

27. Martin H. Redish & Lawrence C. Marshall, *Adjudicator, Independence, and the Values of Procedural Due Process*, 95 YALE L.J. 455, 478–89 (1986).

28. Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. Rev. (forthcoming 2014) (relying on a "technological due process" model to address Big Data's predictive privacy harms), *available at* http://lsr.nellco.org/nyu_plltwp/429/; Neil M. Richards & Jonathan H. King, *Three Paradoxes of Big Data*, 66 STAN. L. REV. ONLINE 41, 43 (2013) (calling for a "technological due process" solution to governmental and corporate decision-making by Big Data predictions).

We offer a number of strategies in this regard. Federal regulators, notably the Federal Trade Commission (FTC), should be given full access to credit-scoring systems so that they can be reviewed to protect against unfairness. Our other proposals pertain to individual decision-making based on algorithmic scores. Although our recommendations focus on credit scoring systems, they can extend more broadly to other predictive algorithms that have an unfair impact on consumers.

Detailed Analysis of Policy Proposals

A. Regulatory Oversight over Scoring Systems

The first step toward reform will be to clearly distinguish between steps in the scoring process, giving scored individuals different rights at different steps. These steps include:

1) Gathering data about scored individuals;

2) Calculating the gathered data into scores;

3) Disseminating the scores to decisionmakers, such as employers;

4) Employers' and others' use of the scores in decisionmaking.

We believe that the first step, data gathering, should be subject to the same strictures as FCRA—whatever the use of the data—once a firm has gathered data on more than 2,000 individuals.[29] Individuals should have the right to inspect, correct, and dispute inaccurate data, and to know the sources (furnishers) of the data. Ironically, some data brokers now refuse to give out their data sources because of "confidentiality agreements" with sources.[30] That position (hiding behind privacy interests to violate consumer privacy) would not stand for consumer reporting agencies covered by FCRA. It should not stand for data brokers and the like.

Second, at the calculation of data stage, ideally such calculations would be public, and all processes (whether driven by AI or other computing) would be inspectable. In some cases, the trade secrets may merit protection, and only a dedicated, closed review should be available. But in general, we need to switch the default in situations like this away from an assumption of secrecy, and toward the expectation that people deserve to know how they are rated and ranked.

The third stage is more difficult, as it begins to implicate First Amendment issues. Given the Supreme Court's ruling in *Sorrell v. IMS Health Inc.*[31] and other rulings in cases involving the regulation of ranking systems,[32] courts may look askance at rules that limit the dissemination of data or scores.[33] Nevertheless, scored individuals should be notified when scores or data are

---

29. This number is meant to permit small businesses' consumer research to be unregulated; we are open to suggestion as to whether the number should be higher or lower.

30. Casey Johnston, *Data Brokers Won't Even Tell the Government How It Uses, Sells Your Data*, ARS TECHNICA (Dec. 21, 2013, 12:07 PM), http://arstechnica.com/business/2013/12/data-brokers-wont-even-tell-the-government-how-it-uses-sells-your-data/.

31. __U.S.__, 131 S. Ct. 2653 (2011).

32. *See, e.g.*, Pasquale, *Beyond Innovation and Competition*, *supra* note **Error! Bookmark not defined.**, at 117–19 (discussing the successful First Amendment defense of the Avvo lawyer ratings site).

33. *Sorrel*, 131 S. Ct. at 2670–72 (holding that drug companies have a constitutional right to access certain types of data without undue state interference); *see also* NEIL M. RICHARDS, INTELLECTUAL PRIVACY: CIVIL LIBERTIES AND INFORMATION IN A DIGITAL AGE ch. 5 (forthcoming 2014) (exploring why *Sorrell* does not lay down a blanket rule that all data is speech for purposes of the First Amendment and more narrowly rested on concerns about viewpoint discrimination among other reasons). For a critical description of the stakes of *Sorrell*, see David Orentlicher, *Prescription Data Mining and the Protection of Patients' Interests*, 38 J.L. MED. & ETHICS 74, 81 (2010) ("When people develop relationships

communicated to an entity. That notification only *increases* speech; it does not restrict or censor communication. Coerced speech can implicate the First Amendment, but like Professor Neil Richards, we do not understand *Sorrell* to lay down a blanket rule that all data is speech.[34] Transparency requirements are consistent with First Amendment doctrine.

The fourth and final stage is the most controversial. We believe that—given the sensitivity of scoring and their disparate impact on vulnerable populations—scoring systems should be subject to licensing and audit requirements when they enter critical settings like employment, insurance, and health care. Such licensing could be completed by private entities that are themselves licensed by the EEOC, OSHA, or the Department of Labor.[35] This "licensing at one remove" has proven useful in the context of health information technology.[36]

Given scoring's sensitivity, fair, accurate, and replicable use of data is critical. We cannot rely on companies themselves to "self-regulate" toward this end—they are obligated merely to find the most *efficient* mode of processing, and not to vindicate other social values including fairness. Licensing can serve as a way of assuring that public values inform this technology.

Licensing entities could ensure that particularly sensitive data does not make it into scoring. For example, data brokers sell the names of parents whose child was killed in car crash,[37] of rape victims,[38] and of AIDS patients.[39] Licensors could assure that being on such a list does not influence scoring. Public hearings could be held on other, troubling categories to gather input on whether they should be used for decisionmaking. Data brokers pigeonhole individuals on the basis of who-knows-what data and inferences. Before letting such monikers become de facto scarlet letters,[40] we need to have a broader societal conversation on the power wielded by data brokers and, particularly, the level of validity of such classifications.

Many of our proposals would require legislation. We are under no illusions that Congress is presently inclined to promote them. However, as in the case of the massive health IT legislation of 2009 (HITECH), it is important to keep proposals "ready to hand" for those brief moments of opportunity when change can occur.[41]

---

with their physicians and pharmacists, they are entitled to the assurance that information about their medical condition will be used for their benefit and not to place their health at risk or to increase their health care costs."); Frank Pasquale, *Grand Bargains for Big Data*, 72 MD. L. REV. 682, 740 (2013); Andrew Tutt, *Software Speech*, 65 STAN. L. REV. ONLINE 73, 75 (2012).

34. *See* RICHARDS, *supra* note 33, at ch. 5.

35. For a relevant case regarding the potentially discriminatory impact of a scoring system or its use, see *EEOC v. Kronos Inc.*, 620 F.3d 287, 298 n.5 (3d Cir. 2010) ("[Regarding] the low score on the Customer Service Assessment she had completed as part of the application process[, the manager] noted from the Customer Service Assessment that Charging Party potentially might be less inclined to deliver great customer service.").

36. Frank Pasquale, *Private Certifiers and Deputies in American Health Care*, 92 N.C. L. REV. (forthcoming 2014).

37. *See* Kashmir Hill, *OfficeMax Blames Data Broker for 'Daughter Killed in Car Crash' Letter*, FORBES (Jan. 22, 2014, 12:09 PM), http://www.forbes.com/sites/kashmirhill/
2014/01/22/officemax-blames-data-broker-for-daughter-killed-in-car-crash-letter/.

38. Amy Merrick, *A Death in the Database*, NEW YORKER (Jan. 23, 2014), http://www.newyorker.com/online/blogs/currency/2014/01/ashley-seay-officemax-car-crash-death-in-the-database.html.

39. *Id.*

40. Frank A. Pasquale, *Rankings, Reductionism, and Responsibility*, 54 CLEV. ST. L. REV. 115, 122 (2006).

41. This is commonly known as the "garbage can" theory of political change—rather than being rationally planned, most legislative efforts depend on whatever plans are at hand. J. Bendor et al., *Recycling the Garbage Can: An Assessment of the Research Program*, 95 AM. POL. SCI. REV. 95, 169 (2001).

Fortunately, the Federal Trade Commission does have statutory authority to move forward on several parts of the "scored society" agenda. The FTC can oversee credit-scoring systems under its authority to combat "unfair" trade practices under Section 5 of the Federal Trade Commission Act.[42] It can use this authority to develop much more robust oversight over credit scoring, which could then be a model for legislation for other scoring entities (or for state consumer protection authorities and state attorneys general with authority to promote fair information practices).

"Unfair" commercial practices involve conduct that substantially harms consumers, or threatens to substantially harm consumers, which consumers cannot reasonably avoid, and where the harm outweighs the benefits.[43] In 2008, the FTC invoked its unfairness authority against a credit provider for basing credit reductions on an undisclosed behavioral scoring model that penalized consumers for using their credit cards for certain transactions, such as personal counseling.[44]

The FTC's concerns about predictive algorithms have escalated with their increasing use. In March 2014, the FTC is hosting a panel of experts to discuss the private sector's use of algorithmic scores to make decisions about individuals, including individuals' credit risk with certain transactions, likelihood to take medication, and influence over others based on networked activities.[45] The FTC has identified the following topics for discussion:

- How are companies utilizing these predictive scores?
- How accurate are these scores and the underlying data used to create them?
- How can consumers benefit from the availability and use of these scores?
- What are the privacy concerns surrounding the use of predictive scoring?
- What consumer protections should be provided; for example, should consumers have access to these scores and the underlying data used to create them?
- Should some of these scores be considered eligibility determinations that should be scrutinized under the Fair Credit Reporting Act?[46]

FTC Chairwoman Edith Ramirez has voiced her concerns about algorithms that judge individuals "not because of what they've done, or what they will do in the future, but because inferences or correlations drawn by algorithms suggest they may behave in ways that make them poor credit or insurance risks, unsuitable candidates for employment or admission to schools or other institutions, or unlikely to carry out certain functions."[47] In her view, predictive

---

42. *See* Federal Trade Commission Act § 5, 15 U.S.C. § 45 (2012). *See generally A Brief Overview of the Federal Trade Commission's Investigative and Law Enforcement Authority*, FED. TRADE COMM'N, http://www.ftc.gov/about-ftc/what-we-do/enforcement-authority (last updated July 2008).

43. 15 U.S.C. § 45(n) (2012).

44. Stipulated Order for Permanent Injunction and Other Equitable Relief Against Defendant CompuCredit Corp., FTC v. CompuCredit Corp., No. 1:08-CV-1976-BBM-RGV (N.D. Ga. Dec. 19, 2008), *available at* http://www.ftc.gov/sites/default/files/documents/cases/2008/12/081219compucreditstiporder.pdf. For a compelling account of the crucial role that the FTC plays in regulating unfair consumer practices and establishing a common law of privacy, see Daniel J. Solove & Woodrow Hartzog, *The FTC and the New Common Law of Privacy*, 114 COLUM. L. REV. (forthcoming 2014), *available at* http://ssrn.com/abstract=2312913 (last updated Oct. 29, 2013).

45. *See Spring Privacy Series: Alternative Scoring Products*, FED. TRADE COMM'N, http://www.ftc.gov/news-events/events-calendar/2014/03/spring-privacy-series-alternative-scoring-products (last visited Feb. 11, 2014).

46. *Id.*

47. Ramirez, *supra* note **Error! Bookmark not defined.**, at 7.

correlations amount to "arbitrariness-by-algorithm" for mischaracterized consumers.[48]

Indeed, as Chairwoman Ramirez powerfully argues, decisions-by-algorithm require "transparency, meaningful oversight and procedures to remediate decisions that adversely affect individuals who have been wrongly categorized by correlation."[49] Companies must "ensure that by using big data algorithms they are not accidently classifying people based on categories that society has decided—by law or ethics—not to use, such as race, ethnic background, gender, and sexual orientation."[50]

With Chairwoman Ramirez's goals in mind and the FTC's unfairness authority, the FTC should move forward in challenging credit-scoring systems. The next step is figuring out the practicalities of such enforcement. How can the FTC translate these aspirations into reality given that scoring systems are black boxes even to regulators?

1.Transparency to Facilitate Testing

The FTC should be given access to credit-scoring systems and other scoring systems that unfairly harm consumers. Access could be more or less episodic depending on the extent of unfairness exhibited by the scoring system. Biannual audits would make sense for most scoring systems; more frequent monitoring would be necessary for those which had engaged in troubling conduct.[51]

We should be particularly focused on scoring systems which rank and rate individuals who can do little or nothing to protect themselves. The FTC's expert technologists[52] could test scoring systems for bias, arbitrariness, and unfair mischaracterizations. To do so, they would need to view not only the datasets mined by scoring systems[53] but also the source code and programmers' notes describing the variables, correlations, and inferences embedded in the scoring systems' algorithms.[54]

For the review to be meaningful in an era of great technological change, the FTC's technical experts must be able to meaningfully assess systems whose predictions change pursuant to AI logic. They should permitted to test systems to detect patterns and correlations tied to classifications that are already suspect under American law, such as race, nationality, sexual orientation, and gender. Scoring systems should be run through testing suites that run expected and unexpected hypothetical scenarios designed by policy experts.[55] Testing reflects the

---

48. *Id.* at 8.

49. *Id.*

50. *Id.*

51. *See* Helen Nissenbaum, *Accountability in a Computerized Society*, 2 SCI. & ENGINEERING ETHICS 25, 37 (1996) (describing commentators' calls for "simpler design, a modular approach to system building, meaningful quality assurance, independent auditing, built-in redundancy, and excellent documentation").

52. The FTC's Senior Technologist position has been filled by esteemed computer scientists Professor Edward Felten of Princeton University, Professor Steven Bellovin of Columbia University, and now by Professor LaTanya Sweeney of Harvard University.

53. *See, e.g.*, Zarsky, *supra* note 17, at 1520.

54. We thank Ed Felten for suggesting that oversight of automated systems include access to programmers' notes for the purpose of assessing source code. Ed Felten, Comment to Danielle Citron, Technological Due Process Lecture at Princeton University Center on Information Technology Policy Lecture Series (Apr. 30, 2009); *see also Danielle Citron: Technological Due Process*, CTR. FOR INFO. TECH. POL'Y, https://citp.princeton.edu/event/citron/ (last visited Feb. 11, 2014). The question we shall soon address is whether the public generally and affected individuals specifically should *also* have access to the data sets and logic behind predictive algorithms.

55. Citron, *supra* note 15, at 1310.

norm of proper software development, and would help detect both programmers' potential bias and bias emerging from the AI system's evolution.[56]

2.Risk Assessment Reports and Recommendations

Once the FTC evaluates credit-scoring systems to detect "arbitrariness-by-algorithm"—as Chairwoman Ramirez astutely puts it—it should issue a Privacy and Civil Liberties Impact Assessment evaluating a scoring system's negative, disparate impact on protected groups, arbitrary results, mischaracterizations, and privacy harms.[57] In those assessments, the FTC could identify appropriate risk mitigation measures.

An important question is the extent to which the *public* should have access to the data sets and logic of predictive credit-scoring systems. We believe that each data subject should have access to all data pertaining to the data subject. Ideally, the logics of predictive scoring systems should be open to public inspection as well. There is little evidence that the inability to keep such systems secret would diminish innovation. The lenders who rely on such systems want to avoid default—that in itself is enough to incentivize the maintenance and improvement of such systems. There is also not adequate evidence to give credence to "gaming" concerns—i.e., the fear that once the system is public, individuals will find ways to game it. While gaming is a real concern in online contexts, where, for example, a search engine optimizer could concoct link farms to game Google or other ranking algorithms if the signals became public, the signals used in credit evaluation are far costlier to fabricate.[58] Moreover, the real basis of commercial success in "big data" driven industries is likely the quantity of relevant data collected *in the aggregate*—something not necessarily revealed or shared via person-by-person disclosure of data held and scoring algorithms used.

We must also ensure that academics and other experts can comment on such scoring systems. Kenneth Bamberger and Deidre Mulligan argue that Privacy Impact Assessments required by the E-Government Act are unsuccessful in part due to the public's inability to comment on the design of systems whose specifications and source codes remain obscured.[59]

As Tal Zarsky argues, the public could be informed about the datasets that predictive systems mine without generating significant social risks.[60] Zarsky demonstrates that—when it comes to "the collection of data and aggregation of datasets"—it is evident that "providing information regarding the kinds and forms of data and databases used in the analysis . . . generate[s] limited social risks . . . [usually only in the context of] secretive

---

56. Batya Friedman & Helen Nissenbaum, *Bias in Computer Systems*, 14 ACM TRANSACTIONS ON INFO. SYSTEMS 330, 334 (1996).

57. Zarsky, *supra* note 17, at 1529; *see also* Citron, *Open Code Governance*, *supra* note 16, at 370–71 (exploring the untapped potential of federally required Privacy Impact Assessments). For example, the Office of Civil Rights and Civil Liberties of the Department of Homeland Security is required to draft Civil Liberties Impact Assessments in response to new programs and policies impacting minorities. *Civil Rights & Civil Liberties Impact Assessments*, U.S. DEP'T OF HOMELAND SEC., https://www.dhs.gov/civil-rights-civil-liberties-impact-assessments (last visited Feb. 11, 2014).

58. They are, in this sense, more likely to be "honest signals," and we should not expend a great deal of effort to assure their integrity without stronger evidence that they are likely to be compromised. *See, e.g.*, SANDY PENTLAND, HONEST SIGNALS (2010).

59. Kenneth A. Bamberger & Deidre K. Mulligan, *Privacy Decisionmaking in Administrative Agencies*, 75 U. CHI. L. REV. 75, 81–82, 88–89 (2008). Twelve percent of agencies do not have written processes or policies for all listed aspects of Privacy Impact Assessment (PIA) and sixteen percent of systems covered by the PIA requirement did not have a complete or current PIA. *Id.* at 81.

60. Zarsky, *supra* note 17, at 1524 (exploring the practical and normative implications of varying kinds of transparency for governmental predictive systems).

governmental datasets."[61]

The more difficult question concerns whether scoring systems' source code, algorithmic predictions, and modeling should be transparent to affected individuals and ultimately the public at large. Neil Richards and Jonathan King astutely explain that "there are legitimate arguments for some level of big data secrecy," including concerns "connected to highly sensitive intellectual property and national security assets."[62] But these concerns are more than outweighed by the threats to human dignity posed by pervasive, secret, and automated scoring systems. At the very least, individuals should have a meaningful form of notice and a chance to challenge predictive scores that harm their ability to obtain credit, jobs, housing, and other important opportunities.

B. Protections for Individuals

In constructing strategies for technological due process in scoring contexts, it is helpful to consider the sort of notice individuals are owed when governmental systems make adverse decisions about them. Under the Due Process Clause, notice must be "reasonably calculated" to inform individuals of the government's claims against them.[63] The sufficiency of notice depends upon its ability to inform affected individuals about the issues to be decided, the evidence supporting the government's position, and the agency's decisional process.[64] Clear notice decreases the likelihood that agency action will rest upon "incorrect or misleading factual premises or on the misapplication of rules."[65]

Notice problems have plagued agency decision-making systems. Automated systems administering public benefits programs have terminated or reduced people's benefits without any explanation.[66] That is largely because system developers failed to include audit trails that record the facts and law supporting every decision made by the computer.[67] Technological due process insists that automated systems include immutable audit trails to ensure that individuals receive notice of the basis of decisions against them.[68]

1.Notice Guaranteed by Audit Trails

Aggrieved consumers could be guaranteed reasonable notice if scoring systems included audit trails recording the correlations and inferences made algorithmically in the prediction process. With audit trails, individuals would have the means to understand their scores. They could challenge mischaracterizations and erroneous inferences that led to their scores.

Even if scorers successfully press to maintain the confidentiality of their proprietary code and algorithms vis-à-vis the public at large, it is still possible for independent third parties to review it. One possibility is that in any individual adjudication, the technical aspects of the

---

61. *Id.*

62. Richards & King, *supra* note 28, at 43.

63. Dusenbery v. United States, 534 U.S. 161, 168 (2002).

64. JERRY L. MASHAW, DUE PROCESS IN THE ADMINISTRATIVE STATE 176 (1985).

65. Goldberg v. Kelly, 397 U.S. 254, 268 (1970).

66. Citron, *supra* note 15, at 1276–77.

67. *Id.* at 1277 (describing automated public benefits systems that failed to include audit trails and how thus the systems were "unable to generate transaction histories showing the 'decisions with respect to each eligibility criterion for each type of assistance' in individual cases").

68. *Id.* at 1305. Immutable audit trails are essential so that the record-keeping function of audit trails cannot be altered. Citron & Pasquale, *supra* note **Error! Bookmark not defined.**, at 1472.

system could be covered by a protected order requiring their confidentiality. Another possibility is to limit disclosure of the scoring system to trusted neutral experts.[69] Those experts could be entrusted to assess the inferences and correlations contained in the audit trails. They could assess if scores are based on illegitimate characteristics such as race, nationality, or gender or on mischaracterizations. This possibility would both protect scorers' intellectual property and individuals' interests.

2.Interactive Modeling

Another approach would be to give consumers the chance to see what happens to their score with different hypothetical alterations of their credit histories. Imagine an interface where each aspect of a person's credit history is represented on a wiki.[70] To make it more concrete, picture a consumer who is facing a dilemma. She sees on her credit report that she has a bill that is thirty days overdue. She could secure a payday loan to pay the bill, but she'd face a usurious interest rate if she takes that option. She can probably earn enough money working overtime to pay the bill herself in forty days. Software could give her an idea of the relative merits of either course. If her score dropped by 100 points when a bill went unpaid for a total of sixty days, she would be much more likely to opt for the payday loan than if a mere five points were deducted for that term of delinquency.

Just as the authors of the children's series *Choose Your Own Adventure* helped pave the way to the cornucopia of interactive entertainment now offered today,[71] so, too, might creative customer relations demystify credit scoring. Interactive modeling, known as "feedback and control," has been successfully deployed in other technical contexts by a "values in design" movement.[72] It has promoted automated systems that give individuals more of a sense of how future decisions will affect their evaluation. For example, Canada's Immigration Bureau lets individuals enter various scenarios into a preliminary "test" for qualification as a permanent resident.[73] The digital interface allows users to estimate how different decisions will affect their potential to become a Canadian citizen. Learning French or earning a graduate degree can be a great help to those in their thirties; on the other hand, some over sixty years old can do "everything right" and still end up with too few points to apply successfully. The public scorecard does not guarantee anyone admittance, and is revised over time. Nevertheless, it provides a rough outline of what matters to the scoring process, and how much.

Credit bureaus do need some flexibility to assess a rapidly changing financial environment. Any given score may be based on hundreds of shifting variables; a default may be

---

69. *See* Dan L. Burk & Julie E. Cohen, *Fair Use Infrastructure for Rights Management Systems*, 15 HARV. J.L. & TECH. 41, 62 (2001); Pasquale, *Beyond Innovation and Competition*, *supra* note **Error! Bookmark not defined.**, at 162.

70. For general information on wikis, see Daniel Nations, *What is a Wiki?*, ABOUT.COM, http://webtrends.about.com/od/wiki/a/what_is_a_wiki.htm (last visited Feb. 11, 2014).

71. Grady Hendrix, *Choose Your Own Adventure*, SLATE (Feb. 18, 2011, 7:08 AM), http://www.slate.com/id/2282786/.

72. Comments of Deirdre K. Mulligan, Professor, Univ. of Calif. at Berkeley & Nicholas P. Doty in Response to the National Telecommunications & Information Administration's Request for Comments on the Multistakeholder Process To Develop Consumer Data Privacy Codes of Conduct, Docket No. 120214135-2135-01, at 11 (May 18, 2012), *available at* http://www.ntia.doc.gov/files/ntia/mulligan_doty_comments.pdf. *See generally* HELEN NISSENBAUM, PRIVACY IN CONTEXT: TECHNOLOGY, POLICY, AND THE INTEGRITY OF SOCIAL LIFE (2010); PROFILING THE EUROPEAN CITIZEN 67 (Mireille Hildebrandt & Serge Gutwirth eds., 2008).

73. *Determine Your Eligibility — Federal Skilled Workers*, GOV'T OF CANADA, http://www.cic.gc.ca/english/immigrate/skilled/apply-who.asp (last updated June 20, 2013).

much less stigmatizing in a year of mass foreclosures than in flush times. Credit bureaus may not be capable of predicting exactly how any given action will be scored in a week, a month, or a year. Nevertheless, they could easily "run the numbers" in old versions of the scoring software, letting applicants know how a given decision would have affected their scores on, for example, three different dates in the past.

We need innovative ways to regulate the scoring systems used in the finance, insurance, and real estate industries, and perhaps might even consider a "public option" in credit scoring. Even if it were first only tried in an experimental set of loans, it could do a great deal of good. If a public system could do just as well as a private one, it would seriously deflate industry claims that scoring needs to be secretive—a topic explore in more depth in the next section.

Overcoming Objections

Credit bureaus will object that transparency requirements—of any stripe—would undermine the whole reason for credit scores. Individuals could "game the system" if information about scoring algorithms were made public or leaked in violation of protective orders.[74] Scored consumers would have ammunition to cheat, hiding risky behavior and routing around entities' legitimate concerns such as fraud.

We concede that incidental indicators of good credit can become much less powerful predictors if everyone learns about them. If it were to become widely known that, say, the optimal number of credit accounts is four, those desperate for a loan may be most likely to alter their financial status to conform with this norm.

However, we should also ask ourselves, as a society, whether this method of judging and categorizing people—via a secretive, panoptic sort—is appropriate. It has already contributed to one of the greatest financial crises in American history, legitimizing widespread subprime lending by purporting to scientifically rank individuals' creditworthiness with extraordinary precision. Secretive credit scoring can needlessly complicate the social world, lend a patina of objectivity to dangerous investment practices, and encode discriminatory practices in impenetrable algorithms.[75]

The benefits of secrecy are murkier than these costs. Moreover, the secrecy of credit scoring can impede incremental innovation: how can outsiders develop better scoring systems if they have no way of accessing current ones? Secret credit scoring can undermine the public good, since opaque methods of scoring make it difficult for those who feel—and quite possibly are—wronged to press their case.

If scorers can produce evidence about the bad effects of publicity, that might justify

---

74. Odysseas Papadimitriou, *Occupy Wall Street & Credit Score Reform*, WALLETBLOG (Mar. 21, 2012), http://www.walletblog.com/2012/03/credit-score-reform/ ("[T]he Occupiers are off-base in suggesting that we centralize credit scoring and make the underlying formulas public. This would only make it easier for people to game the system, which would make existing credit scores less useful to banks and lead more of them to create their own proprietary scores that consumers would have no way of accessing."). But bureaus may have more "economic" incentives to keep their methods hidden. *See* Eric Pitter, *The Law of Unintended Consequences: The Credit Scoring Implications of the Amended Bankruptcy Code—and How Bankruptcy Lawyers Can Help*, 61 CONSUMER FIN. L. Q. REP. 61, 65 (2007) ("CRAs have refused to disclose their credit scoring formula to anyone, even the Federal Reserve Board. The CRAs' full exclusivity of their credit scoring model protects their niche and their unique role in the credit markets.").

75. Amar Bhide, *The Hidden Costs of Debt Market Liquidity* 17–19 (Ctr. on Capitalism & Soc'y, Columbia Univ., Working Paper No. 79, 2013), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2206996.

keeping the correlations, inferences, and logic of scoring algorithms from the public at large. But that logic would not apply to the FTC or third-party experts who would be bound to keep proprietary information confidential.

Another objection is that our proposal only works when the very existence of scoring systems is public knowledge, as in the case of credit scores. In non-credit contexts, entities are under no legal obligation to disclose scoring systems to the public generally and to impacted individuals specifically. Some scoring systems are not a secret because their business model is the sale of scores to private and public entities. Data brokers, for instance, rank, categorize, and score consumers on non-credit bases so they can avoid the obligations of FCRA.[76]

To be sure, it is impossible to challenge a scoring system that consumers do not even know exists. Secret scores about people's health, employability, habits, and the like may amount to unfair practices even though they fall outside the requirements of FCRA. In that case, the FTC would have authority to require entities to disclose hidden scoring systems.

Of course, scoring systems that remain secret would be difficult for the FTC to identify and interrogate. Lawmakers could insist upon the transparency of scoring systems that impact important life opportunities. California, for instance, has been at the forefront of efforts to improve the transparency of businesses' use of consumer information.[77] The FTC has called upon federal lawmakers to pass legislation giving consumers access to the information that data brokers hold about them.[78] In September 2013, Senate Commerce Committee Chairman Jay Rockefeller announced his committee's investigation of the information collection and sharing practices of top data brokers.[79] We are particularly supportive of such efforts—scoring systems can only be meaningfully assessed if they are known and subject to challenge.

---

76. Pam Dixon, Exec. Dir., World Privacy Forum, Testimony Before Senate Committee on Commerce Science and Transportation: What Information Do Data Brokers Have On Consumers, and How Do They Use It? 3 (Dec. 18, 2013), *available at* http://www.worldprivacyforum.org/wp-content/uploads/2013/12/WPF_PamDixon_CongressionalTestimony_DataBrokers_2013_fs.pdf. For a discussion of the Fair Credit Reporting Act model, see Frank Pasquale, *Reputation Regulation: Disclosure and the Challenge of Clandestinely Commensurating Computing*, *in* THE OFFENSIVE INTERNET 107, 111–2 (Saul Levmore & Martha C. Nussbaum eds., 2010).

77. ACLU OF CAL., LOSING THE SPOTLIGHT: A STUDY OF CALIFORNIA'S SHINE THE LIGHT LAW 13 (2013), *available at* http://www.aclunc.org/R2K.

78. FED. TRADE COMM'N, PROTECTING CONSUMER PRIVACY IN AN ERA OF RAPID CHANGE 14 (2012), *available at* http://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-report-protecting-consumer-privacy-era-rapid-change-recommendations/120326privacyreport.pdf.

79. Tom Risen, *Rockefeller Expands Investigation of Consumer Data Brokers*, U.S. NEWS & WORLD REP. (Sept. 25, 2013), http://www.usnews.com/news/articles/2013/09/25/rockefeller-expands-investigation-on-consumer-data-brokers.

`

March 31, 2014

Office of Science and Technology Policy
Eisenhower Executive Office Building
1650 Pennsylvania Avenue, N.W.
Washington, DC 20502

Attention: Big Data Study

Submitted Electronically by Email to: BigData@OSTP.gov

Dear Sir or Madam:

IMS Health respectfully submits information in response to the Office of Science and Technology Policy's "Government 'Big Data': Request for Information".

We are a leading global information and technology services company providing clients in the healthcare industry with comprehensive solutions to measure and improve a wide range of activities relating to healthcare. We serve key healthcare organizations and decision makers around the world, spanning the breadth of life science companies, including pharmaceutical, biotechnology, consumer health and medical device manufacturers, as well as distributors, providers, payers, government agencies, policymakers, researchers and others. IMS Health is an international expert in health information stewardship — including personal privacy and data protection. We firmly believe that:

- The free flow of information used wisely and responsibly advances economic growth, innovation, AND, in, health care big data offers real value for patients by improving quality, safety, value and outcomes; and,

- Personal privacy must be preserved and protected.

As a company, we are personal privacy and data protection advocates AND leading experts in the collection, standardization, organization, structure, integration and analysis of health information. Since our founding more than 60 years ago, IMS has pioneered practices to de-identify personal sensitive data, while serving a broad array of health stakeholders, including the FDA and other agencies of the United States (US) Department of Health and Human Services. IMS relies on a combination of resources, policies and practices to ensure the leadership and expertise necessary to manage information in a manner that balances vital societal values, including improved health care and personal privacy.

Overview:

IMS applauds and supports the work of the Office of Science and Technology Policy to explore approaches to "continue to promote the free flow of information in ways that are consistent with both privacy and security". We believe the free flow of data, if carried

Kimberly S. Gray, Esq., CIPP/US
Chief Privacy Officer, Global
Work: 484 567 6045
KGray@us.imshealth.com
IMS Health® | One IMS Way | Plymouth Meeting Pa 19462

out with robust and routinely employed personal privacy, security, and data stewardship practices, can lead to important improvements in economic growth, innovation and, more specifically as shown through our work, improved health care delivery. To accomplish the important twin objectives of the free flow of data and personal privacy protection, there must be a laser focus on the creation and development of data stewardship in conjunction with data availability for all entities, public and private, that create and handle data. We support policy on big data that is applied to public, private and non-profit entities that:

- Encourages self-regulation to achieve technology-neutral goals for personal data privacy and security;
- Develops public-private collaboration to identify and disseminate best practices; and,
- Remains relevant and adaptable as technology improves and more data becomes available.

Specific Responses to Questions Posed by OSTP :

Question 2: What types of uses of big data could measurably improve outcomes or productivity with further government action, funding, or research? What types of uses of big data raise the most public policy concerns? Are there specific sectors or types of uses that should receive more government and/or public attention?

Types of Uses that Could Measurably Improve Outcomes or Productivity:

The responsible use of big data (e.g. patient-anonymous health information) has already, over many years, helped improve health outcomes and productivity. Specific examples abound, including the Centers for Disease Control and Prevention's initiative to ensure appropriate antibiotic use and control antimicrobial resistance; the Food and Drug Administration's mini-sentinel project; and health care systems' quality management programs. These specific projects which are flourishing, underpin the recent McKinsey Global Institute report on big data that identified the health care sector as one of the five sectors in which use of big data would be transformative. In fact, the McKinsey Global Institute report estimates that deployment of big data more fully in health care could yield $300 billion of value through reduced health cost growth and improved productivity in the next decade.[1]

The following are just a few examples of the important uses of big data in health care:

- Hospital readmissions can be reduced and outcomes improved through analysis of health care utilization and analysis of real world clinical practice to identify best practices. Provider practice patterns can then be assessed against best practice benchmarks with data driven feedback supporting continuing education and care improvement.
- Population health can be measured, tracked, and compared worldwide using anonymous patient records collected globally. With this information, health care

[1] McKinsey Global Institute. "Big data: The next frontier for innovation, competition, and productivity", May 2011.

Kimberly S. Gray, Esq., CIPP/US
Chief Privacy Officer, Global
Work: 484 567 6045
KGray@us.imshealth.com
IMS Health® | One IMS Way | Plymouth Meeting Pa 19462

programs and systems can understand significant and broad health threats and build community wide interventions to reduce disease burden.

- Consumer on-line shopping tools, cost estimators, health alerts/education and other technology-enabled tools are being developed using health care big data. Today, consumers can assess allergy risks, flu activity, and other health alerts. We are at the beginning of a revolution in the kinds of information and tools that will become available and accessible to patients, allowing each of us to participate in better healthcare choices.
- Biopharmaceutical development and distribution is safer and more productive and efficient with use of a broad array of big data assets. Data allows both the FDA and manufacturers to focus communications about drugs to the appropriate audiences among subsets of healthcare practitioners, both reducing costs and better communicating product information including effectiveness and safety information.

Types of Uses that Raise Public Policy Concerns:

The debate and discussion over 'appropriate use' of truthful information and data inevitably draws lines that exclude important and valued uses. One person's 'marketing' is another person's means to evaluate consumer products. Rather than focus on the type of use or the type entity that may be accessing data, we urge federal policy to focus on whether information or data access:

1. serves consumers by improving the economy, advancing knowledge, creating efficiency, or fueling innovation; and,
2. adequately protects personal or confidential information.

In addition, big data will allow the linking of information about identified patients across different care settings, enhancing the information available to providers to improve care. However, some applications may raise confidentiality, bioethical and other concerns that will need to be addressed in a thoughtful manner (including the ability of the patient to control information about themselves).

Specific Sectors or Types of Uses that Should Receive More Government and/or Public Attention:

Health data is widely viewed as sensitive and extensive federal regulations govern the use and disclosure of health information by Covered Entities (i.e., health care providers, health plans, health clearinghouses) and their Business Associates via the Health Insurance Portability and Accountability Act's (HIPAA) Privacy Rule and Security Rule and the Health Information Technology Act (HITECH).

However, health data is increasingly being collected, shared and stored by entities that are not regulated by HIPAA/HITECH, which may lead to public policy concerns related to the privacy and security of health information, affecting both health care consumers (patients) and those individuals and institutions handling health information.

Kimberly S. Gray, Esq., CIPP/US
Chief Privacy Officer, Global
Work: 484 567 6045
KGray@us.imshealth.com
IMS Health® | One IMS Way | Plymouth Meeting Pa 19462

We urge support for a common privacy and security framework for the handling of health information that would ensure that health data is handled with the same due care no matter the entity possessing it or the circumstances in which it was created.

Question 3: What technological trends or key technologies will affect the collection, storage, analysis, and use of big data? Are there particularly promising technologies or new practices for safeguarding privacy while enabling effective uses of big data?

Longstanding practices, updated and revised as technology evolves, exist that provide important privacy and security protection for the transfer and use of data. Specifically, the removal or encryption of personal identifiers is a technology- enabled practice vital to personal privacy and security. And, when this practice is combined with physical, administrative, and technical safeguards, personal privacy protections are further enhanced. This process is known as anonymization or de-identification.

Multiple methods for anonymization or de-identification exist and an entire discipline of experts is focused on work to develop methods and evaluate risks. The HIPAA Privacy Rule relies upon de-identification as a patient privacy protection and establishes two methods for de-identifying (these methods being the "Safe Harbor" method and the "Expert Determination" method), and defines such resulting data as "de-identified". It will be important to continue developing guidance on how to apply these methods to de-identification in big data, and encourage the use of standards and certifications to increase the level of transparency of organizational practices in both the private and public sectors. IMS Health urges the Administration to encourage adoption and use of best practice methods for rendering patient data non-identifiable or anonymous whenever possible, and to call out HIPAA "de-identification" as an example of practices that provide strong protection and under which no real world re-identification has occurred.[2]

Further, for government data sets, it is imperative that the federal government engages, like the HIPAA de-identification guidance recommends, in updated and continuous risk evaluation to determine the need for additional or new administrative, technical and physical safeguards. Risk varies with each use, analysis and data recipient. "No single universal solution addresses all privacy and identifiability issues. Rather, a combination of technical and policy procedures are often applied to the de-identification task." [3]

IMS Health notes, based on its long history of data stewardship in a big data environment, that maintaining the privacy, security, and integrity of de-identified data is an on-going necessity and requires significant investment in training, security, and technology.

---

[2] Ann Cavoukian, Ph.D. and Khaled El Emam, Ph.D., Dispelling the Myths Surrounding De-identification: Anonymization Remains a Strong Tool for Protecting Privacy, The Information and Privacy Commissioner Office Ontario, Canada, June 2011. Accessed at: http://www.ipc.on.ca/images/Resources/anonymization.pdf

[3] http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html

Kimberly S. Gray, Esq., CIPP/US
Chief Privacy Officer, Global
Work: 484 567 6045
KGray@us.imshealth.com
IMS Health® | One IMS Way | Plymouth Meeting Pa 19462

Question 5: What issues are raised by the use of big data across jurisdictions, such as the adequacy of current international laws, regulations, or norms?

Challenges with the Current System

Current international laws and data protection regulations have contributed towards a solid framework for addressing data stewardship and cross-border transfers in certain jurisdictions. For example, the legalizing measures for transfer of personal information from the European Union (EU) to the US (i.e., Model Contracts, Binding Corporate Rules, Safe Harbor) are relatively well defined and are navigated by many organizations. However, the framework is neither complete nor easy to navigate — nor does it promote the free flow of information.

Even with the relative clarity of the EU Data Protection Directive, the individual EU member countries often have differing standards and requirements that are not aligned with their sister countries. It is not enough to comply with the EU Directive; there is a need to understand and further comply with the EU members' laws. Further, the paperwork required to comply with EU requirements can be overly burdensome.

In the last year, foreign distrust of US data protection as a result of the Snowden/NSA matter adds yet another challenge to cross-border data flows. IMS Health believes that the majority of US companies who rely on the EU Safe Harbor, Model Clauses or BCRs are in compliance with their legal obligations and, in US Trade and diplomacy this compliance should be actively proffered to counter the distrust.
While the data protections expectations are relatively clear in some regimes around the world, there are many countries that don't yet have data protection requirements or whose data protection schemes are still open to interpretation. As new data protection laws are enacted around the world, there is a constant need to understand their requirements and to ensure data handling practices are aligned with the new requirements.

There is some commonality around data protection laws worldwide, as most data protection laws and rules are based on a relatively standard set of fair information practice principles. For example, the definitions of "personal information" and "sensitive personal information" are very similar from country to country. Likewise, there are many countries that treat employee data with special care, and many countries also have comparable requirements concerning transferring personal information from their country's borders into another country. Nonetheless, compliance requirements are dissimilar enough that nothing short of a country-by-country legal analysis is needed in order to fully appreciate the nuances between each country's data protection structure.

Two models that offer a more holistic way of addressing global data protection requirements are the EU's Binding Corporate Rules (BCRs) and the Cross-Border Privacy Rules (CBPRs) of the Asia-Pacific Economic Cooperation (APEC). By offering consistent and adequate levels of protection, both frameworks embrace the dual goals of protection of personal privacy while enhancing the free flow of data. IMS Health appreciates the US support for both APEC's CBPR and EU's BCR programs.

Kimberly S. Gray, Esq., CIPP/US
Chief Privacy Officer, Global
Work: 484 567 6045
KGray@us.imshealth.com
IMS Health® | One IMS Way | Plymouth Meeting Pa 19462

In a global data economy in which the free flow of data is needed to fuel innovation, and the value of data is just being recognized, data protection must be developed alongside the amassing of data without stifling creativity.

There are Bigger Challenges and Bigger Rewards with Big Data

Under normal circumstances, navigating a hodgepodge of data protection laws and rules is challenging. In a big data world, this is magnified.

An emerging trend is to attempt to keep up with the changes in technology that go hand in hand with big data (e.g., mobile apps, cloud data storage, cybersecurity) by considering the creation of new privacy, information security and data protection laws. Unfortunately, technology develops at a lightning-fast pace and will always outpace the law, especially in a global economy with a myriad of global data protection laws. Therefore, finding an interoperable privacy and security framework that will allow for a universal baseline of protection, accountability and self-regulation is a much more logical approach.

A Privacy and Security Framework to Support Big Data: Two Key Tenets

1. ***Where possible, consider using anonymous or de-identified data.***
   Anonymizing or de-identifying personal information is a privacy-enhancing technology that can support the responsible use and transfer of information in a big data global economy.

   Today there are inconsistent views among the various legislators and regulators worldwide as to when data is sufficiently anonymized to adequately reduce the risk of re-identification to a very low level and as to the role that administrative, technical and physical safeguards play in reaching and maintaining this low level.

   In healthcare IMS Health has strongly supported the development and implementation of a common standard for de-identification/anonymization. IMS Health suggests creation of a de-identification/anonymization standard and that the HIPAA de-identification standard be the primary basis for this common standard.

2. ***Promote transparency and accountability.***
   Organizations (public, private, non-profit) should internally develop and implement uniform approaches for global access to and for uses of personal information. It will be important for everyone to contribute to the development of guidance on how to de-identify health information in a big data setting, and encourage the use of standards and certifications to increase the level of transparency of organizational practices in both the private and public sectors. Further, they should conduct assessments of their privacy posture by engaging in privacy impact assessments, codes of conduct or other means which demonstrate accountability and which results can be shared to promote trust.

Kimberly S. Gray, Esq., CIPP/US
Chief Privacy Officer, Global
Work: 484 567 6045
KGray@us.imshealth.com
IMS Health® | One IMS Way | Plymouth Meeting Pa 19462

Changes to the Risk/Benefit Analysis

Perhaps most important is the need to re-consider the risk-benefit analysis that has been bantered about to date. Rather than merely focusing so narrowly on privacy risks, policy debate should also include a real assessment of the additional risks associated with choking off big data. If big data isn't leveraged but is stymied, it will have negative economic impact and a very real impact on healthcare quality and innovation. Big healthcare data has the potential to save countless lives and this must be part of the rational public policy discussion about balance of the dual goals of privacy protection and improving healthcare.

Conclusion:

Personal information must be handled with the utmost care, whether in a traditional or big data setting. Finding the pathway to protecting personal privacy and to sharing data that will improve healthcare and save lives should be everyone's goal.

Respectfully submitted,

Kimberly S. Gray, Esq., CIPP/US
Chief Privacy Officer, Global
IMS Health

March 31, 2014

Nicole Wong
Office of Science and Technology Policy
Attn: Big Data Study
Eisenhower Executive Office Building
1650 Pennsylvania Ave. NW
Washington, DC 20502

Sent via email to bigdata@ostp.gov

Re: Notice of Request for Information, "Big Data RFI," FR Doc. 2014-04660

Dear Ms. Wong,

The Interactive Advertising Bureau ("IAB") provides these comments in response to the White House Office of Science and Technology request for information regarding "big data."

Founded in 1996 and headquartered in New York City, the IAB (www.iab.net) represents over 600 leading companies that actively engage in and support the sale of interactive advertising, including leading search engines and online publishers. Collectively, our members are responsible for selling over 86% of online advertising in the United States. The IAB educates policymakers, consumers, marketers, agencies, media companies and the wider business community about the value of interactive advertising. Working with its member companies, the IAB evaluates and recommends standards and practices and fields critical research on interactive advertising. The IAB is committed to promoting best practices in interactive advertising, and is one of the leading trade associations that released cross-industry self-regulatory privacy principles for the collection of web viewing data.[1] These Self-Regulatory Principles are administered by the Digital Advertising Alliance ("DAA"), and have been widely implemented across the online advertising industry, and are enforceable through longstanding and effective industry self-regulatory enforcement programs.

IAB supports access to and use of data, which fuels innovation, provides tremendous benefits to consumers and our economy, and helps ensure our nation's current competitive position globally. IAB believes current U.S. regulatory approach appropriately addresses concrete harms while promoting the free flow of data. Moreover, IAB has long supported, and continues to support, robust self-regulatory enforcement efforts as a means of promoting accountability within the advertising ecosystem while ensuring the flexibility and adaptability of the industry and its enforcement efforts. To continue to build on the successes of the data-driven economy, IAB urges the U.S. Government to identify ways to provide access to more government data. IAB also asks that the Administration promote the success of the U.S. model in discussions with international partners to avoid the creation of unnecessary barriers to the free flow of data that would harm U.S. competitiveness.

[1] Press Release: *Key Trade Groups Release Comprehensive Privacy Principles for Use and Collection of Behavioral Data in Online Advertising*, July 2, 2009, *available at* http://www.iab.net/about_the_iab/recent_press_releases/press_release_archive/press_release/pr-070209.

Response to RFI Questions

(1) What are the public policy implications of the collection, storage, analysis, and use of big data? For example, do the current U.S. policy framework and privacy proposals for protecting consumer privacy and government use of data adequately address issues raised by big data analytics?

A. Value of Data

As the Office of Science and Technology Policy ("OSTP") stated in the introduction to this Request for Information, we are undergoing a "revolution" with data, which has already demonstrated its transformative power in virtually every industry. The collection, storage, analysis and use of data has fueled economic growth and provided benefits for consumers and businesses alike. Consumers have only begun to benefit from the development of new products and services and job creation in the healthcare, financial services, information technology, transportation, retail, and marketing industries, to name only a few areas of growth driven by the use of data. These industries use data to provide products and services that improve our daily lives, such as through new medical devices, rapid processing times for transactions, mobile shopping applications and countless other ways. In the online advertising context, companies collect data for numerous operational purposes including ad delivery, ad reporting, site rendering, accounting, and network efficiencies and optimization, and site or application customization. These operations are necessary for a seamless cross-channel experience and a functioning digital economy, as well as to support and monetize the applications and services expected by customers in the marketplace today.

Moreover, the data revolution has had a profound impact on the democratization of knowledge across the citizenry through the public release of information previously held by the government or otherwise made inaccessible to the public. This data has reshaped the concept of an informed citizenry, leading to a more efficient marketplace, a more vibrant democracy, and a safer and more secure country. A particularly striking example of this dynamic is the Federation for Internet Alerts, which partners with the U.S. National Center for Missing and Exploited Children and the U.S. National Oceanographic and Atmospheric Agency to deliver life-saving alerts when a child is abducted, or a natural disaster is imminent.[2] This is only one of many applications fueled by data, and businesses have only scratched the surface of innovation and solutions driven by access to large and diverse data sets and the ability to process and apply the data for the benefit of consumers. For the benefit of this effort to better understand the role that data is playing in reshaping our society, we provide a discussion of our industry, the online advertising industry.

B. Online Advertising

Advertising fuels the Internet economic engine. For two decades online advertising has fueled the growth of the Internet by delivering innovative tools and services used by consumers and business to connect and communicate. Revenues from online advertising support and facilitate e-commerce and subsidize the cost of content and services that consumers value, such

[2] http://www.internetalerts.org/

as online newspapers, blogs, social networking sites, mobile applications, email, and phone services. Because of advertising support, consumers can access a wealth of online resources at low or no cost. These advertising-supported resources have transformed our daily lives. The support provided by online advertising is substantial and growing despite the difficult economic times.

Online advertising revenue supports market entry for businesses, new communication channels (*e.g.*, micro-blogging sites and social networks), free or low-cost services and products (*e.g.*, email, photo sharing sites, weather, news, and entertainment media), and enables consumers to compare prices, learn about products, and find out about new local opportunities. The ad-supported model also provides a supplemental revenue source for subscription and other business models online. Digital advertising also drives competition. It particularly empowers small businesses, lowering barriers to entry and enabling them to flourish and compete where costs would otherwise hinder their ability to enter and compete in the market. This leads to a greater diversity of online companies, products, and services, from which consumers gain value.

As a result of this advertising-based model, the Internet has been able to grow and deliver widespread consumer benefit. According to a September 2012 study entitled *Economic Value of the Advertising-Supported Internet Ecosystem* conducted for IAB by Harvard Business School Professor John Deighton, between 2007 and 2011, a period when U.S. civilian employment remained flat, the number of jobs that rely on the U.S. ad-supported internet doubled to 5.1 million. The study found that the ad-supported digital industry directly employs 2 million Americans, and indirectly employs a further 3.1 million in other sectors. Calculating against those figures, the interactive marketing industry contributed $530 billion to the U.S. economy last year, also close to double figures from 2007 that placed it at $300 billion. The study, designed to provide a comprehensive review of the entire Internet economy and answer questions about its size, what comprises it, and the economic and social benefits Americans derive from it, revealed key findings that analyze the economic importance, as well as the social benefits, of the Internet.

The backbone of this thriving Internet economy is data. Data fuels not only the research and development responsible for business to business (B2B) advertising innovations like programmatic buying, ad exchanges, and technological developments for marketplace efficiency; but, enables marketers and publishers to reach the right consumer, with the right advertisement, at the right moment. Business to consumer (B2C) interaction would be unwieldy, if not impossible, without data to help narrow the audience and tailor custom-made content for the consumer.

Consumers have embraced the ad-supported model of the Internet and use it to create value in all areas of life, whether through e-commerce or through free access to valuable content. They are increasingly aware that the data collected about their interactions and behavior on the web and in-application is then used to create an enhanced and tailored experience. Importantly, research demonstrates that consumers are generally not reluctant to participate online due to advertising and marketing practices.

For these reasons, digital advertising is thriving, creating jobs, and expanding access to information and communication channels. Without data, the Internet economy does not exist.

C. Public Policy Implications

As an engine for transforming our economy and producing abundant opportunities for growth and innovation, the public policy implications of data are profound at the national and individual levels of society. The paramount goal must be to nurture the data revolution to achieve its maximum transformative potential for the benefit of consumers, businesses, and the economy writ large. The question is not whether to support or oppose the data-driven economy, but how to lock in and expand its benefits while guarding against potential costs.

Many issues surrounding the collection and use of data are not new. For example, businesses have long collected and analyzed information in order to improve products and services for existing customers and to identify new customers. Consumers have examined the costs and benefits associated with these activities and businesses have reacted accordingly, demonstrating a vibrant and effective marketplace. The same dynamic holds true in the big data environment as the practices evolve to incorporate analytics driven outcomes and research and development in secure data "sandboxes." With regard to online advertising, as with many other areas of the data-driven Internet ecosystem, the oversight process has been calibrated to promote innovation and growth while protecting consumers, as demonstrated by consumers' widespread embrace of the model.

The current model addresses concerns through the application of current law and through robust enforcement of self-regulatory programs. As we go forward in the new data economy, the same principles and governance structure that has led to this achievement should drive the future of the Internet. In short, the approach should not be monolithic, but rather be determined based on the proposed uses of data and weighed alongside the social benefits produced by such uses.

As policymakers confront the technological breakthroughs that are driving the collection, storage, analysis, and use of large data sets they should take care not to impose unnecessary and burdensome regulation. IAB believes the appropriate path forward is clear: industry self-regulation is the preferred approach to addressing policy concerns associated with data collection and use. As demonstrated, and described in more detail below, industry has designed and implemented a program backed by credible enforcement that governs the collection and use of web viewing and mobile data.

(2) (a) What types of uses of big data could measurably improve outcomes or productivity with further government action, funding, or research?

The focus of the government should be to avoid restricting or restraining current data practices in order to continue to promote the economic growth and innovation that data is driving. The government is well-positioned to drive consumer education, and help raise awareness about the data privacy tools available to the consumer in the marketplace today. Several government agencies play an important role in consumer protection and education is fundamental to the success of this mission.

Furthermore, the government can improve outcomes and productivity through the release of government data and the government's support for private sector research using this data. The federal government is one of the nation's largest collectors of data, and the data in its possession

has the potential to lead to fundamental scientific, economic, and technological breakthroughs if made available to the public. The government can and should provide access to data it maintains in an appropriate form and fashion so as to enable the public to conduct research and innovate accordingly.

For example, the government holds a vast trove of health information that, if made available to the public, could be analyzed with private sector analytics tools to identify trends and correlations that may unlock the secrets to new and life-saving medical breakthroughs and treatments. The government's data on transportation, labor, domestic and international trade financial services, housing, and other sectors can lead to similar developments and innovations, provided the public has access to the data to perform the research. This research will generate new valued services and drive technological innovation, propelling the U.S. ahead of our global competitors.

(b) What types of uses of big data raise the most public policy concerns?

The vast collection of data by the federal government creates a unique opportunity for private sector research and development; however, the government's control over this data leaves the broader citizenry vulnerable to the attendant abuses the Constitution was designed to remedy. Diverse government uses of big data have the potential to run afoul of not just long standing Due Process principles, but First Amendment rights, Equal Protection, and even fundamentals of commerce. Strict legislative and judicial oversight should not be limited to revelations about the NSA and metadata collection; rather, cut across all government bodies that collect, house, and use data.

The public policy considerations regarding government big data collection are very distinct from private sector collection and use. Government collection and use is layered with inherent and potential harm to core fundamental rights.

Within the context of the private sector and big data use, significant public policy concerns were identified before the advent of the term Big Data; but are nonetheless equally addressed by law, regulation, and self-regulation. For example, market segmentation is a long established and legitimate business practice that, like many other industry practices, is regulated to prevent behavior that could lead to harmful discrimination. The use of data—big or small—for eligibility determination for credit, health, insurance, and employment is regulated by the Fair Credit Reporting Act and the use of web viewing data for such eligibility decisions is strictly prohibited by the Digital Advertising Alliance Self-regulatory program and backed up by both enforcement by the Council of Better Business Bureaus, and Section 5 of the Federal Trade Commission Act.

(c) Are there specific sectors or types of uses that should receive more government and/or public attention?

Policymakers have wrestled with these complex questions for a long time and have established legal frameworks that reflect the need to protect consumers from concrete harms while promoting innovation and job creation. For example, certain data, if misappropriated, could cause harm to consumers, such as certain health information and financial data. For this

reason, Congress passed the Health Insurance Portability and Accountability Act and the Gramm-Leach-Bliley Act, as well as other laws that address concrete, identifiable harms.

It is important to stress that the collection, storage, analysis, or use of data does not per se present any new considerations or issues that cannot be addressed through the existing frameworks. This is particularly true with respect to digital advertising. The practice of obtaining information about consumers' interests and tailoring the advertising of products and services to appeal to those interests dates back more than a century or longer. The migration of this smart and efficient business practice to the online environment and the advances made by the industry to deliver to consumers' relevant digital advertising across the entire internet ecosystem does not change the fundamental nature of the service—the delivery of relevant advertising—. In fact, as is explained below, to the extent that individual consumers have preferences or wish to exercise choice, the online environment offers such consumers greater control than ever before.

(3) (a) What technological trends or key technologies will affect the collection, storage, analysis and use of big data?

While certain broad trends in the use of data are important to examine, such as the shift to mobile, a study of how access to and use of data is evolving to improve the economy and society should be technology neutral. The Internet provides consumers with the opportunity to access information and content on an unprecedented scale. Increasingly, mobile technology facilitates this access and delivers to consumers what they need and when they need it. Businesses are able to take advantage of new mobile technologies to deliver more relevant content and services to their customers and prospective customers. As the use of mobile technology continues unabated, consumers will seek out, and businesses will create, new products, services, and solutions driven by advanced collection, storage, analysis, and uses of data.

The online advertising ecosystem is constantly evolving to meet consumers' needs and developing new ways to provide relevant offerings to interested consumers. In so doing, the industry supports the Internet's "long tail" of publishers that subsist on the revenue generated by interest-based advertising - which may be as much as 200 percent higher than revenue from non-interest-based advertising. While online publishers of all sizes rely on external advertising exchanges and other third-party advertising technologies, smaller Web sites depend on them for a significantly greater portion of their advertising revenue. According to a recent study, ads for which cookie-related information was available sold for three-to-seven times higher than ads without cookies.[3]

In studying these and other trends, the government should maintain a technology neutral approach to maximize the benefits gained from free market competition.

[3] An Empirical Analysis of the Value of Information Sharing in the Market for Online Content, Beales, J. Howard and Eisenach, Jeffrey A., January 2014; available at: http://www.aboutads.info/resource/fullvalueinfostudy.pdf

(b) Are there particularly promising technologies or new practices for safeguarding privacy while enabling effective uses of big data?

In many ways, businesses are only as successful as the trust that consumers place in them. Indeed, companies are increasingly offering consumers new privacy features and tools such as sophisticated preference managers, persistent opt-outs, universal choice mechanisms, and shortened data retention policies. Progress is underway to expand choice mechanisms to the browser, to limit precise URL data, and to safeguard data leakage. Companies are working to keep best practices on pace with technology advancement. These developments demonstrate that companies are responsive to consumers and that companies are focusing on privacy as a means to distinguish themselves in the marketplace. IAB believes that this impressive competition and innovation should be encouraged.

In particular, IAB supports empowering consumers with control over interest-based advertising. As the technology of online advertising has developed and matured, IAB has worked to promote enhanced transparency and implement a uniform choice mechanism with respect to interest-based advertising, based on a set of technology-neutral principles developed by the nation's leading media, marketing and technology companies, known as the Digital Advertising Alliance ("DAA"). The DAA self-regulatory principles and the implementing and enforcement programs were recognized by the White House in 2012 as a strong privacy protection model.

Since late 2010, IAB has participated in the deployment of the DAA's Advertising Option Icon and the related website that allows consumers to control their participation in online interest-based advertising. IAB has been integrally involved in the development of this easy-to-use choice option that gives consumers the ability to conveniently opt-out of some or all online behavioral ads delivered by companies participating in the self-regulatory program.[4] In 2011, the program expanded beyond the collection of data for interest based advertising purposes to cover all uses of web viewing data collected from a particular computer or device.

Consumers can exercise choice through this tool and are directed to this choice page by clicking through the Advertising Option Icon and notices provided in or near ads or on web pages where data is collected or used for online behavioral advertising purposes. Once arriving at this choice page, consumers can:

- easily learn which participating companies have currently enabled customized ads for their browser;
- see all the participating companies on this site and learn more about their advertising and privacy practices, including whether the data will be transferred to a non-affiliate for interest-based advertising purposes;
- check whether they have already opted out from participating companies;
- opt out of browser-enabled interest-based advertising by some or all participating companies; or
- use the "Choose All Companies" feature to opt out from all currently participating companies in one easy step.

[4] *See* The Program's Consumer Choice Page, available at www.aboutads.info/choices.

This tool empowers consumers to better understand online advertising, express their preferences, and make granular decision about how ads are targeted to their preferences.

Looking at the present and the future of practices that promote consumer choice, we note that in June of 2013, the broad industry coalition that released the *Self-Regulatory Principles for Online Behavioral Advertising* and *Multi-Site Data* ("*Principles*") issued guidance on applying the principles to the mobile environment including precise location data and personal directory data. The expansion of the principles to the mobile environment has been developed in spite of challenging technical issues requiring close coordination from a large set of industry stakeholders. As discussed below, these efforts are part of a broader self-regulatory framework that the advertising and marketing industry has established to ensure that consumers continue to have control and transparency as technology evolves.

(4) How should the policy frameworks or regulations for handling big data differ between the government and the private sector? Please be specific as to the type of entity and type of use (e.g., law enforcement, government services, commercial, academic research, etc.).

Access and the use of data by the government and the private sector require two separate approaches. The government's access to and use of data raises unique constitutional concerns that require significant oversight, including legislative and judicial regardless of whether in the law enforcement or services context, to ensure compliance with obligations and limitations intended to protect against governmental abuses of power. IAB supports efforts to clarify how and when government may access private citizen data and communications.

The private sector does not require such a heavy-handed approach. Private sector use of data is primarily driven to research, develop, market, and monetize varying consumer products and services. Businesses compete with each other to gain customers, and this fundamental difference with government means that all market participants have an incentive to engage in practices that promote trust. Unlike government access and use, consumers have a choice regarding business practices. Coupled with the strong privacy regulations that exist today, self-regulatory programs can adequately meet evolving consumer privacy expectations in the digital marketplace. For this reason, in the online advertising industry, development of voluntary codes of conduct for commercial data practices continues to be the appropriate approach for addressing the interplay of online privacy and online advertising practices. This approach has been successful in addressing consumers' concerns while ensuring that the marketplace is not stifled or restrained by overreaching and rigid regulation. Unlike formal regulations, which can become quickly outdated in the face of evolving technologies, voluntary codes developed through self-regulation provides industry with a nimble way of responding to new challenges presented by the evolving Internet ecosystem.

To this end, as noted above, IAB was centrally involved with the development of the *Principles*, a framework and a platform that addresses a wide range of issues involving the collection and use of web viewing data, including matters of transparency and choice.

A prominent feature of this self-regulatory program requires companies to provide enhanced notice outside of the privacy policy so that consumers could be made aware of the companies they interact with while using the Internet. Companies engaged in interest-based

advertising indicate their adherence to the *Principles* by providing the Advertising Option Icon, described above, to link consumers to disclosures about data collection and use practices associated with interest-based advertising. Backed by technical specifications governing the deployment of the icon, industry has established uniform standards for communicating online data practices creating a consistent user experience. This approach helps ensure more standardized notice and greater transparency for consumers, which fosters consumers' trust and confidence in how information is gathered from them online and how it is used to deliver advertisements based on their interests. It also strikes an appropriate balance by ensuring meaningful information is conveyed to consumers in easy-to-find locations while providing companies flexibility in how they provide this information.

This cross-industry self-regulatory initiative represents an unprecedented, collaborative effort by the entire marketing-media ecosystem. The release of additional principles and guidance shows the ability of the approach to adapt to new technologies and challenges.

While the IAB supports efforts to improve transparency and consumer control, we oppose establishing prescriptive requirements for the form or substance of consumers' notice and control. Given the complexity of today's online environment, companies need flexibility in how they communicate with their customers and must be able to tailor notices for the underlying technology involved and needs of their customers. Companies also require the flexibility to adapt their communications as practices and technologies evolve. Imposing rigid or one-size fits all legal standards that impact all participants across media channels could have unintended consequences for new and emerging channels. Self-regulation strikes a measured balance by ensuring meaningful communication with consumers and providing companies flexibility in how they provide this information.

IAB is concerned that regulation could overly burden the access to and use of data, in particular if the approach is inflexible or one-size-fits-all. The government should not impose additional restrictions on its collection, storage, analysis, or use. The free flow of data provides enormous benefits to consumers and is driving our economy forward in ways we have only started to realize. Choking off access to data in response to theoretical harms or to achieve other unrelated policy interests would deal a severe blow to the economy and deprive consumers of the rapid pace of innovation and improvement to which they have grown accustomed recently through the use of data. IAB encourages OSTP to foster the use of data across industries so that the benefits from data can be enjoyed by all sectors of the economy and society.

Furthermore, following the release of the Consumer Data Privacy Framework in 2012, the National Telecommunications and Information Administration ("NTIA") began conducting multi-stakeholder meetings with the goal of developing voluntary codes of conduct that protect privacy and promote innovation in the digital economy. As with the online advertising industry's self-regulatory initiatives, the Department of Commerce' multi-stakeholder processes are based on the well-founded idea that creating effective policies requires the collection of input from a wide range of stakeholders. IAB applauds the flexible policy development mechanism intended by the NTIA processes; and, we encourage cooperation between industry and government in developing meaningful solutions to policy concerns.

(5) What issues are raised by the use of big data across jurisdictions, such as the adequacy of current international laws, regulations, or norms?

The approach to data governance in the United States has fostered tremendous positive and transformative effects across all sectors of the economy. The hallmark of the U.S. approach to data governance is its limited regulation and its recognition of the power and success of industry self-regulatory programs. This approach has permitted the free flow of data that not only has improved virtually every existing industry across, but also has birthed entirely new industries that collect, store, analyze, and use data on behalf of other businesses. These changes have immeasurably improved the American economy, not least of which through the creation of millions of jobs. Indeed, the history of U.S. data governance is in many respects the history of the rise Silicon Valley and the ascendance of America's global dominance in the areas of data management and analytics.

The U.S. approach to data works because of the seriousness and efficiency with which industry tackles privacy and data security issues on behalf of consumers. Self-regulatory programs, best practices, and codes of conduct are able to adapt and update quickly to address new threats and vulnerabilities in ways that legislators and regulators are unable to do. Moreover, self-regulatory solutions are effective precisely because they are industry-driven. The companies themselves know the scope and nature of the issues best and know the capabilities of their businesses, and it is this knowledge that results in effective policies that promote consumer choice and data security as well as effective self-regulatory enforcement. The U.S. framework has enabled industry to develop all of these benefits for consumers through simple market-based choices and transparency.

Other legal and regulatory data governance regimes have worked to the significant economic detriment of the host country. In Europe, heavy regulation of data has resulted in the continent falling far behind the U.S. in terms of data-driven economic growth. Recent research highlights the importance of interest-based advertising. At a hearing before U.S. House of Representatives Committee on Energy and Commerce, Subcommittee on Commerce, Manufacturing, and Trade, titled "Internet Privacy: The Impact and Burden of EU Regulation," the Subcommittee heard testimony from Professor Catherine Tucker about the effect on advertising performance of the European Union's e-Privacy Directive, which limits the ability of companies to collect and use behavioral data to deliver relevant advertising. Professor Tucker's research study on this question found that the e-Privacy Directive was associated with a 65% drop in advertising performance, measured as the percent of people expressing interest in purchasing an advertised product. The study also found that the adverse effect of such regulation was greatest for websites with content that did not relate obviously to any commercial product, such as general news websites.

The U.S. model has not only produced vastly superior economic results, but also has proven a more effective enforcement mechanism. While European data privacy regulations have been inconsistently enforced, U.S. self-regulatory programs have a proven track record of successful enforcement of data privacy standards against companies that have violated codes of conduct and voluntary standards.

IAB believes that strong independent enforcement is the key to any self-regulatory program. With our member companies, IAB has developed extensive standards for our membership. IAB has also developed overarching privacy principles for interactive advertising, which apply to all IAB members,[5] as well as focused guidance for businesses in areas such as email data management[6] and online lead generation.[7] We have established a Member Code of Conduct, which builds on the DAA's Self-Regulatory Program.[8] All IAB members are required to adhere to this code and compliance is monitored and enforced by the Council of Better Business Bureaus ("CBBB").

The CBBB is a leader in building enforcement programs around difficult advertising policy issues and has successfully partnered with the FTC in the past on issues such as food and beverage advertising and online marketing to children. CBBB has brought dozens of enforcement actions and has a one hundred percent compliance rate. The CBBB utilizes a monitoring technology platform and staff research to review the practices of companies in the advertising ecosystem that are collecting and using information across websites and over time to tailor ads to consumers' interests and for other purposes covered by the *Self-Regulatory Principles*. When the CBBB identifies a compliance issue, it has discretion to initiate an inquiry of the company to demonstrate its compliance with the *Principles*. The CBBB may begin a formal review process in which it provides the company with guidance and recommendations, followed by the release of a public report detailing the nature of the review and its outcome, including information about the company's involvement in the inquiry and its implementation of the recommendations. The CBBB may, in its discretion, refer the matter to the appropriate government agency for further action. In special circumstances, the CBBB may choose to dispose of the inquiry or to administratively close the case.

Industry has diligently worked to build a comprehensive, robust self-regulatory and enforcement framework for online behavioral advertising. This effort has yielded an unprecedented comprehensive self-regulatory framework for interest-based advertising, as well as significant consumer educational resources, and has made tremendous progress toward the goal of delivering consumer-friendly standards and tools for online behavioral advertising across the Internet. In fact, success is being achieved both at home and abroad. IAB is promoting the U.S. Framework for the collection and use of web viewing data in more than 30 countries. The DAA standard and the self-regulatory program is already being accepted and implemented in Canada and Europe, and fruitful dialogue is taking place with the Chinese. Establishing a uniform, global standard that is based on the U.S. principles helps ensure a consistent experience for consumers worldwide, but also helps reduce a company's compliance costs associated with operationalizing privacy standards in multiple jurisdictions.

---

[5] IAB, "Privacy Principles" (adopted 2008), available at http://www.iab.net/guidelines/508676/1464.
[6] IAB, "Email Data Management Best Practices" (2008), available at http://www.iab.net/media/file/email_data_mgt_best_practices0908.pdf.
[7] IAB, "Online Lead Generation: B2C and B2B Best Practices for U.S.-based Advertisers and Publishers" (2008), available at http://www.iab.net/media/file/B2CandB2BBestPracticesFINALv3.pdf.
[8] IAB Website, "IAB Member Code of Conduct," available at http://www.iab.net/public_policy/codeofconduct.

It is critically important that U.S. officials are aware and embrace the success and superiority of the U.S. model, in terms of its economic effects and its ability to enforce data privacy principles that protect consumers from concrete harms while preserving and promoting innovation and job creation. The U.S. Administration should embrace this approach not only for continuing to foster domestic growth, but also for protecting this model in negotiations with international partners. For example, in discussions with the European Union, U.S. officials should reject the creation of arbitrary barriers to data flow across international boundaries to protect America's competitiveness. Moreover, the Administration should exercise caution in issuing statements or reports that call for new legislation or regulation of data practices in the U.S. Doing so undermines our position abroad for commercial data practices. The U.S. should not take steps that would harm one of the key engines of our economy. Instead, the U.S. should work with our international partners to create safeguards that ensure the protection of legitimate international data transfers and prevent the use of local data storage requirements as a precondition to serving customers internationally.

Conclusion

The Internet is a tremendous engine of economic growth. It has become the focus and a symbol of the United States' famed innovation, ingenuity, inventiveness, and entrepreneurial spirit, as well as the venture funding that follows. Simply put: the Internet economy and the interactive advertising industry creates jobs. In recent years, the fuel that has driven this growth has been free and open access to and use of data. IAB and its member companies have seen and felt the powerful effects of data in transforming the online advertising and the marketing ecosystem for the benefit of the economy and for consumers. OSTP's examination of data must start from the proposition that the policies and frameworks that have birthed America's data-driven economic revolution should be embraced, continued, and defended to the maximum extent possible so that consumers in the U.S. and overseas can reap the benefits of our innovation.

IT Law Group

www.itlawgroup.com

Francoise Gilbert
Attorney at Law
555 Bryant Street, #603
Palo Alto, CA 94301
(650) 804 - 1235
fgilbert@itlawgroup.com

March 31, 2014

Via email: bigdata@ostp.gov

Attn: Big Data Study
Office of Science and Technology Policy, Eisenhower
Executive office Building
1650 Pennsylvania Avenue NW
Washington D.C. 20502

Re: Big Data RFI

The IT Law Group (ITLG) is pleased to submit comments in response to the Office of Science and Technology's Request for Information on the ways in which "big data" will affect how Americans live and work, and the implications of collecting, analyzing and using such data for privacy, the economy, and public policy. In the following paragraphs, I have addressed the five Questions to the Public. Please note that the opinions below are mine and not those of my clients, potential clients, or prospective clients or those of the IT Law Group.

(1) What are the public policy implications of the collection, storage, analysis, and use of big data? For example, do the current U.S. policy framework and privacy proposals for protecting consumer privacy and government use of data adequately address issues raised by big data analytics?

In most cases, it could be argued that the use of personal data for certain big data processing purposes is illegal because the affected data subjects never received proper notice or gave proper consent to the use of their personal data, or other data generated by their online activities, for big data processing purposes.

The current FIPPS, and the other principles found in other base documents addressing personal data protection or privacy rights, such as the White House Privacy Bill of Rights, the EU 1995 Data Protection Directive, the OECD Privacy Principles, the APEC Privacy Framework, or most privacy notices publicly available on companies' websites, are based

on the same pattern, which requires some form of "notice and informed consent" for the use of personal data for a specified purpose.

However, in most cases, organizations are processing data without proper notice and/or informed consent, and for purposes that might be drastically different from the original purposes for which they gave notice to the affected individual (assuming any notice was provided), and for purposes that might be also drastically different from the purpose contemplated by the concerned individuals.

For example, the person who places a phone call or send an electronic message to place order for a pizza does not contemplate that the information might be retained, combined with other data, processed, analyzed, and then sold to insurance companies, who might become informed of his consummation of fat and carbs, which, in turn, might cause an increase in the person's insurance premium.

(2) What types of uses of big data could measurably improve outcomes or productivity with further government action, funding, or research? What types of uses of big data raise the most public policy concerns? Are there specific sectors or types of uses that should receive more government and/or public attention?

There are numerous examples of uses of personal data that might produce a productive outcome. These include, for example, those that are related to healthcare, health policy, research (involving or not personal data), such as research for new components with unique capabilities, or other uses where statistical analysis or pattern recognition technologies may allow the identification of recurring patterns in the causes or consequences of certain diseases.

Some uses of big data that might raise more concern – and for which it might be useful to have additional governmental oversight - include those that might result in the creation of individual profiles, and especially those profiles that would reveal "sensitive information" as the term is defined or understood in the various privacy or data protection cultures throughout the world, such as US based privacy culture (e.g. health, financial) or in the EU/OECD data protection culture (e.g. religion, race, political opinion). A combined list of these sensitive topics might include, for instance:

- Personal preferences, personal profiles, patterns that may lead to unwanted contacts (e.g. Interest in certain categories of movies, interest in certain products), especially in connection with advertising)
- Profiles that may lead to make certain important decisions about an individual (e.g. interest rates, employment)
- Health
- Financial
- Behavior that might have a negative financial consequence (e.g., insurance premium or insurance offering, mortgage rates)
- Religious, philosophical preferences

- Race, ethnic origin
- Political association, trade union membership
- Sex life, sexual preferences

(3) What technological trends or key technologies will affect the collection, storage, analysis, and use of big data? Are there particularly promising technologies or new practices for safeguarding privacy while enabling effective uses of big data?

Advertising; marketing; business development.
Risk management.

(4) How should the policy frameworks or regulations for handling big data differ between the government and the private sector? Please be specific as to the type of entity and type of use (e.g., law enforcement, government services, commercial, academic research, etc.).

Law enforcement might have unique needs for access to certain data for national security purposes and these activities might be governed by a different set of laws (either in the US or in other countries).

Except for law enforcement, private and public sectors are frequently closely aligned, so that the distinction between the two categories might be illusory. Staff moves frequently between government and non-government jobs. Activities that are allocated to government agencies may be performed by private entities. Research that starts in universities may be licensed to commercial enterprises. As a result, the boundaries between public and non-public uses or purposes are becoming so blurred that the dichotomy between government and non-government activities might be illusory.

(5) What issues are raised by the use of big data across jurisdictions, such as the adequacy of current international laws, regulations, or norms?

Except for a few areas of the world that have closed political regimes, in practice, countries boundaries are blurring. However, political considerations and legal regimes are different from the day-to-day commercial or practical reality. Countries still want to enforce their own laws for matters that concern or affect their own citizens.

The public would greatly benefit from the adoption and enforcement of a minimum set of common standards throughout the world. However, experience shows that even countries that are part of larger political entities (and thus previously made a specific commitment to act in a cohesive way throughout the group), as is the case, for example for the members of the European Union, European Economic Area or the APEC, have significant trouble agreeing on common standards.

It is likely that it would take a very long time – or a dramatic event – for all or most countries to agree on a global standard that would set the conditions, limits, or principles governing big data processing.

It might be wise to work on both the national and the international aspect concurrently.

Thank you for the opportunity to submit comments. Please do not hesitate to contact me with further questions.

Sincerely,

Francoise Gilbert
Founder/Managing Director; IT Law Group

Comment
Office of Science and Technology Policy
Big Data Study

March 31, 2014

James C. Cooper
George Mason University School of Law
Law & Economics Center

"Big Data," as it has come to be known broadly is the application of analytics to large datasets that come close to covering an entire population, which allows researchers to find relationships that working with small samples would not reveal.[1] By discovering these new relationships, big data stands to be transformative: Google Flu Trends and IBM's Watson are among the standard bearers of big data, but are merely the tip of the iceberg. Big data also has allowed dramatic improvements in fraud detection, and promises to make driverless cars and smart homes a reality. Working with large data sets, moreover, provides researchers with deeper understandings of social and physical phenomena that may improve living in ways we currently cannot apprehend.

At the same time, some are concerned about the risks that big data may pose. The ubiquitous collection of observations from both our online and offline lives that feeds big data has the potential to intrude on privacy. Further, some worry about the security of these immense data caches. Finally, others have expressed concern over the use of big data to sort consumers in increasing granular categories that will determine the offers and information they receive.

This brief comment makes two key points about regulatory approaches to big data. First, any regulation must be guided by an empirically grounded benefit-cost analysis, not unsupported hypotheticals. Second, because the reduction in private information improves the efficiency of markets, the ability of big data to make more granular classifications – either through firm sorting or consumer signaling – should be considered a benefit rather than a harm.

---

[1] *See* VIKTOR MAYER-SCHÖNBERGER & KENNETH CUKIER, BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW

A. Benefit-Cost Analysis

Any regulatory framework that addresses big data must be guided by empirically grounded benefit-cost analysis. The benefits of big data are tangible and can be measured objectively,[2] and regulations that retard big data will deprive consumers of these benefits. Accordingly, proponents of restrictions on big data should have the burden to demonstrate empirically that such policies are necessary to ameliorate actual or likely consumer harm, and that the avoided harm is greater than the foregone benefits.

Broadly, privacy harms can be classified as tangible or intangible. Tangible harms include the extent to which big data is likely to increase the risk of identity fraud or reputational harm from data breaches of sensitive financial or personal information. Such harms can be measured objectively with metrics like fraudulent charges, inconvenience costs, or lost marketplace opportunities due to stigma. Intangible harms include the discomfort associated with ubiquitous observation and the revelation of embarrassing information. Further, some scholars recently have written about the harm associated with receiving information flows only tailored to predicted interests.[3] These harms are suffered internally and therefore are not amendable to objective measurement.

This is not to say that intangible harm should be ignored or could never form the basis for regulatory action. Before relying on intangible harms as a justification for restrictions on big data, however, policy makers should have a firm grasp on their variance and magnitude. The harm associated with unauthorized monitoring of intimate activities or revelation of sensitive health information, for example, is probably significant for most of the population. At the same time, any discomfort associated with the collection and analysis of anonymized data streams for honing predictive algorithms, or receiving tailored information or advertisements, is likely to vary widely among consumers.

When the variance in sensitivity to a particular form of observation and analysis is low, uniform standards will approximate optimality for most of the population. Alternatively, when harm suffered varies widely, a uniform standard –

---

[2] *See id*.; Thomas M. Lenard & Paul H. Rubin, *The Big Data Revolution: Privacy Considerations*, at 4-9 (Dec. 2013), *at* https://www.techpolicyinstitute.org/files/lenard_rubin_thebigdatarevolutionprivacyconsiderations.pdf.

[3] *See*, *e.g.*, ELI PARISER, THE FILTER BUBBLE: HOW THE NEW PERSONALIZED WEB IS CHANGING WHAT WE READ AND HOW WE THINK (2012); Cynthia Dwork & Deirdre K. Mulligan, *It's Not Privacy and Its Not Fair*, 66 STANFORD L. REV. ONLINE 35, 37 (2013).

especially one geared toward to those who are most sensitive – will impose costs on wide swaths of the population. Many willingly would accept less privacy in exchange for lower prices, richer and more customized content, or superior functionality. Consequently, is it crucial that policy makers refrain from imposing a "one size fits all" solution based on "worst-case" hypotheticals that lack any empirical grounding.

B. Classification Harms

A common theme in several works concerning the potential threats of big data is that data driven algorithms increasingly will be used to sort consumers into categories that will determine the types of offers and information they receive.[4] As explained below, the reduction in private information should not be considered harm. Instead, because increased information improves market efficiency through better matching of buyers and sellers, the more granular classifications made possible by big data are likely to be beneficial.

First, businesses currently categorize consumers using the data that is available, and more data typically will allow for more, not less, accurate estimates of parameters like credit risk, health status, or interests. Further, firms have incentives to place consumers into correct categories; companies that systematically offer high interest credit cards to people with good credit or expensive auto insurance to good drivers, will see their profits suffer.

Although more accurate categorization likely means that some consumers receive worse terms, it also means that many consumers will receive better terms. Because ability to pay is negatively correlated with income, moreover, poorer individuals who were previously priced out of a market are likely benefit from differential pricing. What's more, differential pricing does not occur in a vacuum; just as big data may allow firms to charge consumers with relatively inelastic demand higher prices, the same data driven algorithms will allow their competitors to target these consumers with discounts. In this manner, big data driven classification actually can *intensify* competition.[5]

---

[4] *See*, *e.g.*, Dwork & Mulligan, *supra* note 3; Scott Peppet, *Unraveling Privacy: The Personal Prospectus and the Threat of a Full-Disclosure Future*, 105 NORTHWESTERN U. L. REV. 1116 (2012); Ryan Calo, *Digital Market Manipulation*, 82 GEO. WASH. L. REV. (forthcoming 2014).

[5] *See* Kenneth S. Corts, *Third-Degree Price Discrimination in Oligopoly: All-Out Competition and Strategic Commitment*, 29 RAND J. ECON. 306 (1998); James C. Cooper *et al.*, *Does Price Discrimination Intensify Competition? Implications for Antitrust*, 72 ANTITRUST L. J. 327 (2005). This type of competition can be seen regularly at the grocery store, when after purchasing one brand of diapers or cereal, a consumer often will receive a coupon for a competing brand.

Relatedly, some have expressed concern about the flip side of firms sorting consumers based on big data – consumer signaling.[6] In this scenario, the ubiquity of cheap monitoring technology is increasingly allowing consumers to send credible signals to firms that they possess qualities that should lead to better offers (*e.g.*, lower health risks, better driving skills). The upshot is that those who do not agree to be monitored will be presumed to possess less desirable qualities (*e.g.*, poor health or diving skills), and hence will receive worse terms (*e.g.*, higher health or car insurance rates). It is unclear how providing contracting parties with more information about their true types will harm societal welfare. Again, although some parties will receive worse terms, those with above-average qualities will receive better terms. As more information is revealed to the market, matches between firms and consumers will improve, increasing welfare. Finally, limiting classification will create a moral hazard; the inability to receive the benefits from desirable behavior (*e.g.*, healthy eating, safe driving, or good grades), or to be punished for undesirable behavior, will reduce incentives to engage in desirable behavior.

C. Conclusion

A responsible regulatory approach must rest on an empirical showing that government action is necessary to prevent substantial consumer harm. Because markets work more efficiently with greater amounts of information, regulators should refrain from restricting the ability to use big data driven algorithms to make more accurate classifications. Concerns over inequities generated by improved classification are best deal with through transfer payments to those who are negatively impacted, not by the forced concealment of private information, which will reduce the efficiency of the price system. Big data stands to be transformative, and regulators should exercise appropriate caution and humility to avoid unnecessarily depriving consumers of the substantial benefits that big data promises to offer.

[6] *See* Peppet, *supra* note 4.

Jason Kint

Government "Big Data"; Request for Information

Without limiting the foregoing, commenters should consider the following:

(1) What are the public policy implications of the collection, storage, analysis, and use of big data? For example, do the current U.S. policy framework and privacy proposals for protecting consumer privacy and government use of data adequately address issues raised by big data analytics?

As a veteran of the digital media industry my focus is on the use of big data by the private media sector. The current framework ***does not*** protect consumer privacy. It allows for a large and growing number of third parties to collect and target consumer data without any consent, relationship or even awareness by consumers. This framework is the default setting for any U.S. user of the internet and there is no practical option to prevent it from happening. Unrestrained, this third party tracking also negatively impacts the content companies who invest in the media that drives much of the consumption in the first place.

(1) **Data collection is accelerating.** When a consumer browses the web, his or her software application is repeatedly "tagged" by hundreds of third party cookies in the course of a single session. Each of these third party cookies is similar to an identification badge allowing a company to connect individual users to their expressed behaviors and data. ***The number of third party cookies collected continues to increase geometrically*** due to the economic growth of the tracking industry. Today, this includes entirely new categories of intermediaries and more and more actors[1] in each of these categories.

(2) **Consumer choice is a red herring.** The ability for a consumer to express and maintain privacy from third party tracking agents has become increasingly difficult despite the FTC's landmark behavioral advertising framework established in 2007. As recommended by the FTC in 2010, Do Not Track has rightly become a feature, albeit deeply hidden, in all major web

[1] LUMA Partners (2014). Display Media Landscape.
http://www.slideshare.net/fullscreen/tkawaja/luma-display-ad-tech-landscape-2010-1231/1

browsers but it is ignored by almost all sites except for a few industry leaders (e.g. Twitter, Pinterest). Therefore, ***consumers are being presented with a false-choice*** as many industry reporters[2] don't even understand that this feature is being ignored by nearly all web servers.

Additionally, the self-regulatory group, the Digital Advertising Alliance (DAA), has secured billions of impressions to promote its AdChoices icon program intended to allow consumers to opt-out of behavioral advertising. The program is ***flawed by design*** as it remarkably requires the use of persistent third party cookies to opt-out of third party cookies. Most consumers believe the best way to protect their privacy is to delete their browser cookies. In fact, security software is often set to regularly delete browser cookies. In the case of the AdChoices icon program, all opt-out information is lost whenever a consumer deletes his or her cookies. Additionally, only 115 companies/websites are members of the alliance as of March 23, 2014.

(3) **Content is impacted.** A lesser discussed outcome of this mass-sharing of data and loss of privacy is the effect on companies who produce content for the web. While the industry has attempted to argue behavioral advertising is the lifeblood of content companies, it is my experience that in the long-term the opposite is true[3]. The current ecosystem heavily favors companies with the most data and the ability to leverage this data to serve laser-targeted ads wherever a cookie can be matched most cheaply. The digital pie is merely shifting away from the sites and services being consciously consumed to the companies that can *track* the consumption. Due to lack of constraints on the collection of data and targeting by consumers, the ***U.S. internet will continue to grow as a direct response medium*** fueled by conversion tactics. I like to use the analogy of old-fashioned direct mail. If a company could gain access to lists of homeowners' addresses, could print flyers and acquire postage all free-of-charge -- then marketers would flood our mailboxes with ads. They would decidedly reduce purchasing ads on

[2] Yahoo! Finance (2014). How to keep companies from tracking you online — for good. http://finance.yahoo.com/news/keep-data-brokers-from-tracking-you-online-anonymously-202256230.html
[3] Ad Age (2014). Behavioral advertising may not be as crucial as you think. http://adage.com/article/datadriven-marketing/behavioral-advertising-crucial/291858/

television and in newspapers resulting in valuable loss of revenue for content providers. This is our internet today.

Media companies sell to marketers access to a relationship with their consumers. This requires the media company to invest in content and build a trusted relationship between the consumer and the website which can be shared with the marketer. The readily available data ecosystem and limitless lowest-common-denominator inventory on the web has been a significant factor in a race to the bottom in content production, content investment and the type of advertising to support it. The industry has spawn a complex web of intermediaries who much like high-frequency traders move dollars towards third parties arbitraging on supply and demand of inventory which in this case is the consumer's data. I disagree when it is suggested that regulation of third party data will negatively impact the diversity or quality of content on the web. ***Online behavioral advertising is an insignificant (<10%) percentage of revenue for most publishers.*** However, constraints on third party data will restore value in the intellectual property and production of content and relationships with consumers and marketers. I don't believe I'm exaggerating my point when I say third party data regulation is a critical factor in protecting the future investment in quality free content.

(2) What types of uses of big data could measurably improve outcomes or productivity with further government action, funding, or research? What types of uses of big data raise the most public policy concerns? Are there specific sectors or types of uses that should receive more government and/or public attention?

First party big data that is captured through a trusted relationship is incredibly valuable. The FTC made the difference between first party and third party data very clear beginning in their 2007 report. There is little controversy about the workings of first party cookies which are tied only to the sites visited and that bring fabulous features such as a personalized CNN, recommended movies on Netflix or individually-tailored games on ESPN. Additionally, third party data should be able to be captured about users who have granted permission or not opted out of being tracked. The government

should continue to fund research exploring how big data can be used to enhance experiences and everyday life for consumers. The government should support academic research around matters of trust and privacy in digital services. This trust chain includes the content, marketers and the consumers and I believe it's a vital factor in an open society.

(3) What technological trends or key technologies will affect the collection, storage, analysis and use of big data? Are there particularly promising technologies or new practices for safeguarding privacy while enabling effective uses of big data?

It is my opinion that Do Not Track is one of the most important technologies to be implemented in the next year. The standard feature as originally defined as a http header setting is simple and elegant. The industry needs to accept that this option should dictate both the collection and tracking of data by true third parties with caveats in place for fraud prevention and site measurement. I believe the industry also needs to accept a free-market in which any software company can design Do Not Track to be as useful and simple as possible. I also argue it should be able to be marketed as "on by default" in a privacy-focused product solution. This will accelerate adoption of the feature, increase value in trusted relationships and negatively impact bad actors who do not have any trusted relationships in the exchange of advertising on the internet. The industry self-regulatory group, the DAA, should adjust their policies to honor Do Not Track as defined above and as they promised to the White House in February 2012[4]. The FTC should be given powers to investigate any sites who do not honor Do Not Track in true third party requests. Organizations should be funded to audit this on behalf of the FTC.

[4] White House announcement, February 2012.
http://www.whitehouse.gov/the-press-office/2012/02/23/we-can-t-wait-obama-administration-unveils-blueprint-privacy-bill-rights

(4) How should the policy frameworks or regulations for handling big data differ between the government and the private sector? Please be specific as to the type of entity and type of use (e.g., law enforcement, government services, commercial, academic research, etc.).

I believe Do Not Track should have a caveat for law enforcement with probable cause. I'll leave the rest of this to the experts in this area.

Big Data RFI Response
Government "Big Data"; Request for Information
A Notice by the Science and Technology Policy Office on 03/04/2014

Prepared by:
Jonathan Sander
@sanderiam
jonathan.sander@STEALTHbits.com
Strategy & Research Officer at STEALTHbits Technologies, Inc.

(1) What are the public policy implications of the collection, storage, analysis, and use of big data? For example, do the current U.S. policy framework and privacy proposals for protecting consumer privacy and government use of data adequately address issues raised by big data analytics?

Current policy is not sufficient to address big data issues. There are many good proposals, but the speed with which they are taking shape means they are being lapped by the constantly shifting realities of the technology they are meant to shape. NSTIC (National Strategy for Trusted Identities in Cyberspace http://is.gd/JOrjCw) has been an excellent example of this. That digital identity is core to properly addressing big data should be obvious. How can we hope to protect privacy if we cannot identify the proper steward of data? How can we identify data stewards if the people who ought to be identified have no consistent digital identity? The very founding notion of NSTIC, trusted identities, begs the question of if we are prepared to approach empowering people via assigning responsibility. If we do not have identities that can be trusted, then we don't even have one of the basic building blocks that would be required to approach big data as a whole.

That said, the implications that big data has are too large to ignore. In "The Social, Cultural & Ethical Dimensions of Big Data" (http://is.gd/EGe7tD), Tim Hwang raised the notion that data is the basic element in (digital) understanding; and further that understanding can lead to influence. This is the big data formulation of the notion that knowledge leads to rights, and rights lead to power – the well tested idea of Michel Foucault. In the next century, the power of influence will go to those who have understanding culled from big data. This will be influence over elections, economies, social movements and the currency that will drive them all – attention. People create big data in what they do, but they also absorb huge amounts of data in doing so. The data that can win attention will win arguments. The data that gets seen will influence all choices. We see this on the internet today as people are most influenced not by what they read which is correct but rather what they see that holds their attention. And gaining that influence seems to be playing out as a winner takes all game. With nothing short of the ethical functioning of every aspect of human life on the line, big data policy implications cannot be understated.

(2) What types of uses of big data could measurably improve outcomes or productivity with further government action, funding, or research? What types of uses of big data raise the most public policy concerns? Are there specific sectors or types of uses that should receive more government and/or public attention?

The amount of data involved in some big data analysis and the startlingly complex statistical and mathematical methods used to power them give an air of fairness. After all, if it's all real data powered by cold math, what influence could there be hiding in the conclusions? It is when big data is used to portray something as inherently fair, even just, that we need to be the most concerned. Any use of big

data that is meant to make things “more fair” or “evenly balanced” should immediately provoke suspicion and incredulity. Just a small survey of current big data use shows this to be true. Corrine Yo from the Leadership Conference gave excellent examples of how surveillance is unevenly distributed in minority communities, driven by big data analysis of crime. Clay Shirky showed how even a small issue like assigning classes to students can be made to appear fair through statistics applied to big data when there are clearly human fingers tipping the scales. There are going to be human decisions and human prejudices built into every big data system for the foreseeable future. Policy needs to dictate what claims to fairness and justice can be made and outline how enforcement and transparency must be applied in order to be worthy of those claims.

The best way for government to speed the nation to ethical big data will be to fund things that will give us the building blocks of that system. In no particular order a non-exhaustive list of these ethical building blocks will be trusted identity, well defined ownership criteria of data generated by an individual directly and indirectly, simple and universal terms for consent to allow use of data, strong legal frameworks that protect data on behalf of citizens (even and especially from the government itself), and principles to guide the maintenance of this data over time, including addressing issues of human lifecycles (e.g. what happens to data about a person once they are dead?). There are many current proposals that apply here, e.g. NSTIC as mentioned above. All of these efforts could use more funding and focus.

(3) What technological trends or key technologies will affect the collection, storage, analysis and use of big data? Are there particularly promising technologies or new practices for safeguarding privacy while enabling effective uses of big data?

Encryption is often introduced as an effective means to protect rights in the use of data. While encryption will doubtless be part of any complete solution, today’s actual encryption is used too little and when used often presents too small a barrier for the computing power available to the determined. Improvements in encryption, such as quantum approaches, will surely be a boon to enforcement of any complete policy. The continued adoption of multi-factor authentication, now used in many consumer services, will also be an enabler to the type of strong identity controls that will be needed for a cooperative control framework between citizens and the multitude of entities that will want to use data about the citizens. As machines become better at dealing with fuzzy logic and processing natural language, there will be more opportunities to automate interactions between big data analysis and the subjects of that analysis. When machines can decide when they need to ask for permission and know how to both formulate and read responses to those questions in ways that favor the human mode of communication, there will be both more chances for meaningful decisions on the part of citizens and more easily understood records of those choices for later analysis and forensics.

(4) How should the policy frameworks or regulations for handling big data differ between the government and the private sector? Please be specific as to the type of entity and type of use (e.g., law enforcement, government services, commercial, academic research, etc.).

Policies governing interaction of government and private sector is one area where much of what is defined today can be reused. Conversely, where the system is abused today big data will multiply opportunities for abuse. For example, law enforcement data, big or not, should always require checks and balances of the judicial process. However, there is likely room for novel approaches where large sets of anonymized data produced from law enforcement could be made available to the private and educational sectors en masse as long as that larger availability is subject to some judicial check on behalf of the people in place of any individual citizen. Of course, this assumes a clear understanding of things

being “anonymized” – one of many technical concepts that will need to be embedded in the jurisprudence to be applied in these circumstances. There are cracks in the current framework, though. This can allow data normally protected by regulations like HIPPA to seep out via business partners and other clearing house services that are given data for legitimate purposes but not regulated. All instances of data use at any scale must be brought under a clear and consistent policy framework if there is any hope to forge an ethical use of big data.

Big Data RFI
Comments to the White House
From the Marketing Research Association
Submitted to: bigdata@ostp.gov
March 31, 2014

……...

The Marketing Research Association (MRA) applauds the White House for launching a public consultation on the issue of Big Data and privacy and how federal policy can best grapple with the topic. However, we do worry that it is far too broad to be tackled in such a short period of time, especially since the investigation appears to be shoe-horning multiple concepts into one big bucket.

Background

On January 17, 2014, the president addressed concerns about NSA surveillance, but turned the conversation to a much broader range of issues: "Challenges to our privacy do not come from government alone. Corporations of all shapes and sizes track what you buy, store and analyze our data, and use it for commercial purposes."[1] He launched a 90-day review of Big Data and privacy, led by White House Counselor John Podesta, which Podesta said aimed "to deliver to the president a report that anticipates future technological trends and frames the key questions that the collection, availability, and use of Big Data raise – both for our government, and the nation as a whole. It will help identify technological changes to watch, whether those technological changes are addressed by the U.S.'s current policy framework and highlight where further government action, funding, research and consideration may be required."[2] The White House asked for public comment,[3] and even launched a crude online survey to collect public opinion,[4] although we have sincere concerns about the effort's execution and effectiveness.[5]

Whether considered as the interconnected data-sharing nature of everyday consumer devices, from refrigerators and thermostats to mobile phones and medical devices, or the much maligned brokers of consumer-related data, or the ever-advancing realm of large data sets and predictive

[1] http://www.whitehouse.gov/the-press-office/2014/01/17/remarks-president-review-signals-intelligence
[2] http://www.whitehouse.gov/blog/2014/01/23/big-data-and-future-privacy
[3] https://www.federalregister.gov/articles/2014/03/04/2014-04660/government-big-data-request-for-information
[4] http://www.whitehouse.gov/issues/technology/big-data-review
[5] "White House survey on Big Data and privacy is a good idea, poorly executed." March 24, 2014. http://www.marketingresearch.org/news/2014/03/24/white-house-survey-on-big-data-and-privacy-is-a-good-idea-poorly-executed

analytics, Big Data serves as sort of a cypher. The term is widely-referenced but poorly understood, and people in the public policy community see in it a variety of hopes and fears, particularly as they relate to data privacy and data security.

White House Counselor John Podesta outlined on March 3 how he was approaching the term: "data sets that are so large, so diverse, or so complex that the conventional tools that would ordinarily be used to manage data simply don't work. Instead, deriving value from these data sets require a series of more sophisticated techniques, such as Hadoop , NoSQL, MapReduce and machine learning. These techniques enable the discovery of insights from big data sets that were not previously possible."[6]

The White House seeks to look at Big Data from many standpoints:[7] data collection, sharing and use by government entities and private sector companies; government or law enforcement surveillance, as well as public health; what privacy protections exist or should exist, and how to apply them in all these different circumstances; and how the federal government could or should be encouraging or subsidizing Big Data innovation in the public and private sectors.

Each sub-category would be worthy of its own thorough investigation.

However, MRA is generally agnostic on government collection, sharing and use, except as it regards the decennial Census and the American Community Survey,[8] so we will focus our comments on private-sector data collection, sharing and use.

Survey, opinion and marketing research

MRA is a non-profit national membership association representing the survey, opinion and marketing research profession.[9] MRA promotes, advocates for, and protects the integrity of the research profession, and works to improve research participation and quality.

---

[6] http://www.whitehouse.gov/sites/default/files/docs/030414_remarks_john_podesta_big_data.pdf

[7] The White House asked 5 questions to help guide comments in this review: "(1) What are the public policy implications of the collection, storage, analysis, and use of big data? For example, do the current U.S. policy framework and privacy proposals for protecting consumer privacy and government use of data adequately address issues raised by big data analytics? (2) What types of uses of big data could measurably improve outcomes or productivity with further government action, funding, or research? What types of uses of big data raise the most public policy concerns? Are there specific sectors or types of uses that should receive more government and/or public attention? (3) What technological trends or key technologies will affect the collection, storage, analysis and use of big data? Are there particularly promising technologies or new practices for safeguarding privacy while enabling effective uses of big data? (4) How should the policy frameworks or regulations for handling big data differ between the government and the private sector? Please be specific as to the type of entity and type of use (e.g., law enforcement, government services, commercial, academic research, etc.). (5) What issues are raised by the use of big data across jurisdictions, such as the adequacy of current international laws, regulations, or norms?"

[8] "The Census Bureau's American Community Survey (ACS)." http://www.marketingresearch.org/the-census-bureau%E2%80%99s-american-community-survey-acs

[9] The research profession is a multi-billion dollar worldwide industry, comprised of pollsters and government, public opinion, academic and goods and services researchers, whose members range from large multinational corporations and small businesses to academic institutes, non-profit organizations and government agencies.

Survey and opinion research is the scientific process of gathering, measuring and analyzing public opinion and behavior. On behalf of their clients – including the government (the world's largest purchaser), media, political campaigns, and commercial and non-profit entities – researchers design studies and collect and analyze data from small but statistically-balanced samples of the public.[10] Researchers seek to determine the public's opinion and behavior regarding products, services, issues, candidates and other topics. Such information is used to develop new products, improve services, and inform policy.

Analysis of massive data sets is playing a growing role in the research business, both on its own and in conjunction with more traditional research methodologies. While Big Data analysis can and is done for research purposes, analysis in and of itself is not necessarily a research activity.

MRA, in consultation with the broader research profession, has developed a legal definition of bona fide survey, opinion and marketing research, which implicitly includes Big Data analytical research: "the collection and analysis of data regarding opinions, needs, awareness, knowledge, views, experiences and behaviors of a population, through the development and administration of surveys, interviews, focus groups, polls, observation, or other research methodologies, in which no sales, promotional or marketing efforts are involved and through which there is no attempt to influence a participant's attitudes or behavior."

Survey, opinion and marketing research is thus sharply distinguished from commercial activities, like marketing, advertising and sales. In fact, MRA and other research associations prohibit and attempt to combat sales or fundraising under the guise of research (referred to as "sugging" and "frugging"), "push polls,"[11] and any attempts to influence or alter the attitudes or behavior of research participants as a part of the research process.[12] Quite to the contrary, professional research has as its mission the true and accurate assessment of public sentiment in order to help individuals, companies and organizations design products, services and policies that meet the needs of and appeal to the public.

Data's use and purpose matter

Question #2 asks, among other things, "*What types of uses of big data could measurably improve outcomes or productivity with further government action, funding, or research? What types of uses of big data raise the most public policy concerns? Are there specific sectors or types of uses that should receive more government and/or public attention?*"

---

[10] A "sample" is a subset of a population from which data is collected to be used in estimating parameters of the total population.

[11] "'Push polls' - Deceptive Advocacy/Persuasion Under the Guise of Legitimate Polling." http://www.marketingresearch.org/push-polls-deceptive-advocacypersuasion-under-the-guise-of-legitimate-polling

[12] For instance, in the *MRA Code of Marketing Research Standards:* "7. Ensure that respondent information collected during any study will not be used for sales, solicitations, push polling or any other non-research purpose. Commingling research with sales or advocacy undermines the integrity of the research process and deters respondent cooperation. In addition, the possibility of harm from data sharing – such as health insurance companies adjusting an individual's costs based on information disclosed about their health behaviors or financial companies denying someone credit based on their propensity for online shopping – are the focus of growing public debate about Big Data and data brokers. Respondents should be assured that information shared in a study will only be used for research." http://www.marketingresearch.org/code

Question #4 also asks: "*How should the policy frameworks or regulations for handling big data differ between the government and the private sector?*"

Consumer concerns, such as they exist, rightly focus more on data use than its mere existence or collection. Certain types of data may be considered more sensitive than others, but even those are dependent both on an individual's subjective judgment and the context in which those data are used.

MRA generally supports a privacy model based on intended use – different protections and requirements for data privacy, depending on the uses to which that data will be put. Data to be collected, used and transferred strictly for bona fide survey, opinion and marketing research should be held to a different standard than ordinary commercial use, which will differ from purely transactional use. Those uses should all be treated differently than data used for determining a consumer's eligibility for things like health insurance, credit or a mortgage, or data used to prosecute crimes or prevent terrorism. Research purposes, unlike most of the other purposes, involve data collection, sharing and use of information about individuals only to understand broader population segments and demographic groups.

As noted, we appreciate the president's interest in Big Data and privacy. However, we are also extremely concerned that, by lumping NSA spying and private-sector data collection into the same bucket in his January 17 speech and in this review, the president has inadvertently minimized the importance of public concern about the NSA and Edward Snowden's revelations (whether or not they are legitimate) and redirected that concern to a completely unrelated target. We sincerely hope that, through the course of this review, the White House is able to properly separate these spheres of activity, since such data uses are so extremely different.

While Big Data should be approached with common privacy principles, the application of those principles must differ depending on the purpose and use of the data in question, in both the public and privacy sectors. Even in the public sector, those purposes and uses will vary dramatically. The purposes of IRS data collection – normally, for help in collecting taxes owed by individuals – are extremely different from the Census Bureau, which needs to collect individuals' data in order to learn about diverse groups of Americans. Neither sort may be deemed as intrusive by some Americans as data surveillance by law enforcement and espionage agencies, and thus may be held to different standards. Alternatively, Americans may wish to give even greater leeway to such agencies, since their purposes are so important.

Public policy implications of Big Data

Question #1 asks: "*What are the public policy implications of the collection, storage, analysis, and use of big data? For example, do the current U.S. policy framework and privacy proposals for protecting consumer privacy and government use of data adequately address issues raised by big data analytics?*"

Among the concerns raised about the policy implications of Big Data are: (1) notice and consent for data sharing, access and correction; (2) data minimization; (3) deidentification; and (4) eligibility decision-making.

(1) Notice and consent for data sharing, access and correction

Many concerns about massive data sets center on how to apply consumer notice and consent to the workings of massive data sets.

MRA already requires that researchers seek tailor-made approaches to transparency with regard to clients, research participants, and the public at large that are appropriate to different modes and methods of research. Research best practices require disclosure of what data is being collected and used, and for what purpose, that research organizations designate a chief privacy officer to take responsibility for the privacy of respondents and their data,[13] and that participants have the opportunity to opt out.

But should consumers be required to be notified when their data is shared and be able to either opt out of such sharing, or be required to opt into it?

While some survey, opinion and marketing research indicates that consumers, on average, are concerned about their privacy, notifying consumers of every minute thing happening with their data could work counter to the interest of actually keeping consumers well informed, since over-notification and excessively lengthy privacy policies already may be causing consumers to stop paying close attention to their own privacy needs and wants and growing more careless in how they handle their own data.

Moreover, turning to opt in to any great extent for data sharing could severely tilt the competition in the data business towards the largest companies (like Google and Facebook), who have so much internal data that they don't need to share.

The question of whether or not consumers should be given access and control over data regarding them appears to be more complicated.

The demand of access to consumer data may make sense in contexts of eligibility, where such data (particularly if inaccurate) could adversely impact a consumer's credit rating, personal or professional reputation, or in the likelihood of becoming a victim of identity theft or fraud. However, none of these conditions should reasonably be assumed to apply to survey, opinion and marketing research data. Participation in survey, opinion and marketing research is voluntary.

The cost of access and correction could potentially be quite onerous, especially for smaller research companies and organizations, given a potential deluge of frivolous or pointless inquiries. Since the research process is interested in broad groups, not individuals, compiling and tracking individual consumer data, by the individual, would require complex and expensive procedures and infrastructure not currently in use. Moreover, such tracking could lead to a much greater threat of harm from data leakage and empower the kind of consumer tracking that causes privacy concerns.

[13] "Designating a Privacy Officer." http://www.marketingresearch.org/designating-a-privacy-officer

The ability of companies to authenticate the identity of consumers requesting access is another potential problem. That kind of authentication would require collecting and checking even more data, which runs counter to our interest in data minimization and limited data retention. Plus, necessary authentication procedures and processes would add to the cost in money and time on the part of research organizations.

MRA supports the concept of a "sliding scale" for access and correction responsibilities in order to reconcile the vague benefits with the expected costs. We propose that the availability and extent of access should depend on the data actually being susceptible to use for criminal or fraudulent purposes. As pointed out earlier, purpose and use matter.

(2) Data minimization

As a broad principle, data minimization – not collecting or maintaining more data than necessary to fulfill a certain purpose – makes sense. However, within various modes and methods of research, the need to retain data will vary, and should be properly subject to those needs, not an arbitrary decision by fiat of law, or by the decision of a regulatory body unfamiliar with the processes and practices of research. Additionally, a major objective of research is to understand attitudes, behaviors and opinions over time. The collection and analysis of this information often leads to new theories over time, requiring the re-visiting of older data. Prescribed retention periods would thus diminish the long-term value of data collected for research purposes.

MRA would thus be concerned about a law or a regulatory agency setting time constraints without being familiar with the processes and practices of all businesses that would be impacted by their implementation, including the many processes and practices of survey and opinion research.

(3) Deidenification

Anonymizing or de-identifying personal information whenever possible, by aggregating or pseduonymizing it, is a sensible principle, and one that MRA and the research profession support. In fact, carving out a safe harbor from many privacy regulations for data that has been protected in this way makes a lot of sense. We are concerned with the specific definition of how and to what extent to do it.

There is an ongoing debate in the academic and policy arenas on whether or not data can ever be fully de-identified or anonymized. If it cannot, then any piece of data could ultimately be personally identifiable.

Speakers at a conference in Washington, DC in December 2011 clashed extensively over this very question.[14] Several researchers, like Latanya Sweeney, director of the Data Privacy Lab at Harvard University, contended that most any data could be re-identified, based on a pair of her

[14] "Personal Information: The Benefits and Risks of De-Identification." A conference held by the Future of Privacy Forum (FPF) on December 5, 2011. http://www.futureofprivacy.org/2011/10/19/upcoming-fpf-event-one-day-conference-dec-5/

studies. Several other researchers responded that the two biggest re-identification studies were very limited cases and not generalizable.[15]

To illustrate the debate, consider a data point such as date of birth, which could be considered personally identifiable because it splits the population into 25,000 cells and can enable re-identification. If you combine such data with a zip code containing only a handful of people in a certain age range, it may be very easy to re-identify. Professor Peter Swire of Ohio State University made an analogy at the conference to a cop collecting clues. A suspect is male, tall, with red hair. That would not be enough to re-identify, but it would certainly make it easier. It is more a matter of how much legwork, analysis and extra data is available and accurate. That is what weighs against the public being able to re-identify de-identified data, according to Professor Swire.[16]

Khaled El Eman, a researcher at the University of Ottawa, felt that the data re-identification efforts by Sweeney and company were the exceptions that proved the rule. Most attacks fail miserably, he said. According to Eman, the studies that succeeded are too small, too few, too ambiguous, too heterogeneous and with confidence intervals that are way too large. Eman concluded that, "Re-identification is hard." He suggested that there would need to be 40-50 replicable studies to start to change such a conclusion.[17]

It is also important to remember that de-identification carries costs as well as benefits. Daniel Barth-Jones, epidemiology professor at Columbia University, warned at that conference that excessive de-identification of data can yield huge statistical errors and inaccurate research results: the greater the level of de-identification, the less statistically useful the data becomes.[18] Blanket de-identification could grind statistical research and number-crunching to a standstill. Ultimately, is there a point in de-identification to a level where there are significantly easier and cheaper ways of getting the data? Professor Barth-Jones ended his presentation with a warning about trade-offs, that the real harm is not the ephemeral threat to privacy but the real threat of "not catching the next HIV epidemic".[19]

There may be benefit to engaging the government agencies in the broader public debate over de-identification, but they should not be encouraged to arbitrarily end that debate.

(4) Eligibility decision-making

Identifying (and especially, quantifying) the harm arising in data privacy often proves elusive for both privacy activists and regulatory authorities. Aside from legitimate concerns about identity theft, fraud and other criminal abuse, the only potential harm agreed upon to date from Big Data is negative impact on eligibility decisions, such as whether or not consumers should receive health insurance, what kind of credit or mortgage rate they should be offered, or discounts or premiums they should be offered or charged when shopping for products and services.

[15] FPF Conference.
[16] FPF Conference.
[17] FPF Conference.
[18] FPF Conference.
[19] FPF Conference.

A recent Senate hearing on data brokers[20] provides useful illustrations.

Professor Joseph Turow from the Annenberg School for Communication expressed fears at the hearing about the power of advanced computing and statistics "blending your information into complex algorithms" to "better understand you," which he thinks is "turning into an actuarial activity." He touched throughout the discussion on the unfairness of Big Data predictive analytics, from the more well-known algorithmic decision-making for mortgage loans to the dramatically variable pricing of airline tickets. He even suggested that lawmakers should consider "how many data points companies should be allowed to buy about us at a time, and how they can be merged with other data points."

Senator Ed Markey touched on similar concerns at the hearing, decrying the attachment of "propensity scores" to American consumers, without their knowledge and consent, which then become the basis for targeting offers or benefits. High value prospects get good offers, he said, but many others end up getting shut out.

As a recent *Forbes* article described such price discrimination, "In a traditional bazaar a seller might charge a well-dressed buyer twice as much as another based on visual clues or accents. Big data allows for a far more scientific approach to selling at different prices, depending on an individual's willingness to pay."[21]

A more recent workshop hosted by the Federal Trade Commission (FTC) on "Alternative Scoring Products"[22] discussed many similar concerns. Edmund Mierzwinski, consumer program director with the U.S. Public Interest Research Group, pointed out that his organization is not concerned about Big Data, but that they are "concerned about its use and its impact on financial opportunity." However, as pointed out by Stuart Pratt, president and CEO for the Consumer Data Industry Association, the reward and reinforcement of customer loyalty is nothing new, and more than a few participants pointed out that the use or abuse of Big Data for credit and other financial services determinations are already covered by other federal laws, such as the Fair Credit Reporting Act.

MRA happens to think that eligibility decisions are an issue worthy of potential regulatory action regarding Big Data, but crafting the right response has so far proven elusive. Instead of precisely and objectively defining the eligibility decisions that require regulatory oversight and action, the privacy debate has, so far, run a wide and subjective gamut. Meanwhile, some private sector

[20] "What Information Do Data Brokers Have on Consumers, and How Do They Use It?" Senate Commerce Committee. December 18, 2013. http://www.commerce.senate.gov/public/index.cfm?p=Hearings&ContentRecord_id=a5c3a62c-68a6-4735-9d18-916bdbbadf01&ContentType_id=14f995b9-dfa5-407a-9d35-56cc7152a7ed&Group_id=b06c39af-e033-4cba-9221-de668ca1978a

[21] "Different Customers, Different Prices, Thanks To Big Data." by Adam Tammer. Forbes. April 14, 2014. http://www.forbes.com/sites/adamtanner/2014/03/26/different-customers-different-prices-thanks-to-big-data/

[22] FTC workshop. March 19, 2014. http://www.ftc.gov/news-events/events-calendar/2014/03/spring-privacy-series-alternative-scoring-products

companies are innovating on their own to try to bring transparency to some of that decision-making, such as Acxiom's AboutTheData website.[23]

Harmonization of international data privacy protection

Question #5 asks: "*What issues are raised by the use of big data across jurisdictions, such as the adequacy of current international laws, regulations, or norms?*"

The rise of Big Data is one of multiple issues driving the European Union (EU) to revamp and expand the 1998 European Commission's Directive on Data Protection, a regulation looked to as the model for privacy regulation by many countries around the world, whether they like it or not.

The Data Directive prohibits the transfer of "personal data" to non-EU nations that do not meet the European "adequacy" standard for privacy protection. The EU Data Directive places significant restrictions on the collection, use and disclosure of personal data that prove taxing for many researchers. Intentionally or not, the EU wields the Data Directive and its "adequacy" standard as an anti-competitive trade measure, discriminating against U.S. companies in digital trade because they do not deem the U.S. to have "adequate" data privacy protections. The main way a nation can be deemed "adequate" is by passing laws that closely imitate the Data Directive.

Fortunately, for now, in addition to adopting binding corporate rules, U.S. companies can self-certify to the U.S. Department of Commerce that they comply with the seven principles of the U.S.-EU Safe Harbor[24] and at least have some mechanism for data transfer. While it is a self-certification, the FTC enforces compliance with the Safe Harbor under its Section 5 authority to prosecute deceptive practices (not living up to one's public claims). As the EU tries to rewrite their Data Directive, it is essential that we maintain the Safe Harbor – our primary protection for the conduct of digital commerce and research. High level EU officials have made a habit recently of attacking the standing of the U.S.-EU Safe Harbor[25] and the EU Parliament recently voted to essentially get rid of it.

Comprehensive data privacy proposals have been advanced for the last few years by the FTC, the White House, and Members of Congress, all with the goal of better emulating the EU privacy regime so that the U.S. will be deemed "adequate" in its privacy protections by the EU.

MRA supports some form of baseline consumer data privacy law and we feel that is an important issue with which the White House must grapple as part of this Big Data and privacy review. However, the expansive measures envisioned by some parties go far beyond the baseline – with questionable promise of success. "Harmonization" of U.S. law to an EU standard may not make

[23] https://aboutthedata.com/
[24] Notice, Choice, Onward Transfer (to Third Parties), Access, Security, Data Integrity and Enforcement. http://export.gov/safeharbor/eu/index.asp
[25] "EU questions decade-old US data agreement." By Nikolaj Nielsen. EUObserver.com, July 22. http://euobserver.com/justice/120919 and "EU outlines improvements to US data agreement." By Kate Tummarello. *The Hill*, November 27. http://thehill.com/blogs/hillicon-valley/technology/191618-eu-outlines-improvements-to-us-data-agreement

the most sense economically. According to Congressman Lee Terry (R-NE-01), the U.S. has "flexibility" in its privacy regime, allowing for the "emergence of the data economy," which he has identified as "a reason why we are the dominant innovators in this area and Europe is not."[26] Similarly, as outlined by several large technology companies' chief privacy officers at an Internet Association panel discussion in 2013, innovative data businesses generally develop and grow in the US, not in Europe, and our approach to data privacy may be a key factor in our competitive advantage.[27]

More importantly, over the course of many public and private engagements in the past couple of years, Members of the European Parliament and European Commission have indicated that none of the comprehensive proposals offered so far in the U.S. would, if enacted, win the U.S. the coveted "adequacy" designation by the EU. It is possible that nothing short of a complete substitution of EU law for U.S. law would satisfy EU authorities.

Discussions of "harmonization" in trans-Atlantic privacy regulation, particularly given the swelling potential of Big Data, should focus on incorporating the U.S.' strong enforcement mechanisms[28] and self-regulatory entrepreneurship to the EU's more bureaucratic framework of privacy regulation.

What about the president's multistakeholder process for privacy?

As part of this review, the White House strangely seems to have overlooked the president's own multistakeholder privacy process, where the National Telecommunications and Information Administration brings together technology, policy, legal and other experts from innovation companies, trade associations, activist groups, academic institutions and other organizations to develop and agree upon enforceable privacy codes of conduct. While the first such effort, on mobile apps privacy,[29] ended somewhat ambiguously last year, the second effort, on facial recognition privacy, is already well underway.[30] MRA has been an active participant, including presenting a white paper[31] as part of a panel on February 5.

We would like to see the White House give this innovative approach more time to work before declaring where privacy regulation should go.

---

[26] "Rep. Terry: Data is the New Gold." October 29, 2013. http://www.marketingresearch.org/news/2013/10/29/rep-terry-data-is-the-new-gold

[27] "Corporate privacy officers discuss global compliance, trans-Atlantic competition, a comprehensive privacy law, and the US-EU Safe Harbor." March 7, 2013. http://www.marketingresearch.org/news/2013/03/07/corporate-privacy-officers-discuss-global-compliance-trans-atlantic-competition-a-co

[28] "U.S. takes the gold in doling out privacy fines." by Jay Cline. *Computerworld*. February 17, 2014. http://www.computerworld.com/s/article/9246393/Jay_Cline_U.S._takes_the_gold_in_doling_out_privacy_fines?taxonomyId=17

[29] "The NTIA Multistakeholder Process for Mobile Apps Privacy." http://www.marketingresearch.org/the-ntia-multistakeholder-process-for-mobile-apps-privacy

[30] "Facial recognition privacy: Kicking off another NTIA multistakeholder process." February 5, 2014. http://www.marketingresearch.org/news/2014/02/05/facial-recognition-privacy-kicking-off-another-ntia-multistakeholder-process-and-an-

[31] "The Marketing Research Applications of Facial Recognition Technology." MRA white paper. February 6, 2014. http://www.marketingresearch.org/sites/mra/files/pdfs/MRA_facial-recognition-MR-applications_2-6-14.pdf

Conclusion

Survey, opinion and marketing researchers already encounter significant public apathy with respect to research participation. Research "response" rates have been falling for the last couple of decades, driving up the cost of and time involved in achieving the required number and strata of participants to reach viable representative samples for most research studies. That always informs MRA's approach to any new regulatory impediments to research: that the issues identified above could make it harder to reach and involve research participants, increase non-response bias, make it more difficult to share and learn from data, and adversely impact the accuracy of research results.[32]

We've highlighted in these comments some areas of Big Data privacy regulation that may hold the most need and promise for consideration, such as notice and choice for consumer data sharing, access and correction, data minimization, deidentification, eligibility decision-making, and our digital trade relationship with the EU.

MRA looks forward to working with the White House on these and other important privacy issues.

However, the White House's review, as well-intended and necessary as it may be, spawned from a debate not about most of the issues we've discussed in these comments, but about the surveillance of citizens by American intelligence authorities. As observed recently by the *Washington Post*'s Catherine Rampell, "All the information the government collects in secret probably does little to cultivate trust in the collection that occurs more transparently."[33]

Privacy issues in the private sector, while certainly a concern, are already being tackled on multiple fronts. Between robust enforcement by the FTC and other agencies of unfair or deceptive practices, and the president's nudging towards more effective self-regulation through his multistakeholder process, there is plenty going on already. MRA urges the White House to strongly consider where its attention would best be focused and to not dictate direction too strongly to the independent authorities and processes already innovating in ways the White House would want.

---

[32] This wouldn't just impede bona fide survey and opinion research. It would ultimately result in higher costs for research – costs which would be passed on to all Americans, in the form of: higher prices for goods and services; lengthier time before new or better goods and services are brought to the marketplace; delayed introduction of new or better public policies; and a decreased amount of research ordered by companies, who might then bring less well-tested and researched products and services to market, harming consumers in the end because the goods and services did not fulfill consumer expectations or needs.

[33] "'Big data' needs a helping hand in Washington." By Catherine Rampell. *The Washington Post*. March 27, 2014. http://www.washingtonpost.com/opinions/catherine-rampell-big-data-needs-a-helping-hand-in-washington/2014/03/27/11b1f90e-b5bd-11e3-8cb6-284052554d74_story.html

Albany
Atlanta
Brussels
Denver
Los Angeles
Miami
New York

McKenna Long & Aldridge LLP
1900 K Street, NW
Washington, DC 20006
Tel: 202.496.7500
mckennalong.com

Northern Virginia
Orange County
Rancho Santa Fe
San Diego
San Francisco
Seoul
Washington, DC

DANIEL W. CAPRIO, JR.
Direct Phone: 202.496.7348
Direct Fax: 202.496.7756

EMAIL ADDRESS
dcaprio@mckennalong.com

March 31, 2014

Ms. Nicole Wong
Big Data Study
Office of Science and Technology
The White House
Washington, DC 20500

Re: Big Data Study

Dear Ms. Wong:

On behalf of Transatlantic Computing Continuum Policy Alliance,[1] I am pleased to submit these comments in response to the White House request for information on Big Data. We commend the White House for this request for information.

The Need to Foster Innovation

Big Data is a term used to describe the use of data from traditional and digital sources that represents a source for ongoing discovery and analysis. Any analysis of big data needs to consider the Internet of Things (IoT), a broad term that describes the ecosystem of sensors that interact with each other, persons, and services in computer-aware environments supported by analytics. Among others, this includes sensors that will interact with each other, sensors that will interact with the broad ecosystem through local area networks (LANs) as well as sensors that may be in direct contact with the Internet. All actors need to consider the breadth and potential implications of all policy actions on this emerging, yet complex, ecosystem.

[1] The Transatlantic Computing Continuum Policy Alliance consists of AT&T Corporation, General Electric, Intel Corporation and Oracle Corporation.

Big Data and the IoT represents transformative 21st century technology that promises to revolutionize homes, cars, health care, education and industry in general. Big Data and the IoT present both the opportunity and challenge of protecting privacy and security while encouraging innovation and creative new services. Whether we call it the Smarter Planet, the Internet of Everything, or the Industrial Internet - Big Data and the IoT are about innovation and the future of the internet ecosystem itself.

We are still at the very beginning of the promise of Big Data and the IoT. Business models must be allowed flexibility to develop. The potential benefits of Big Data and the IoT are only now emerging. For instance, sensors can and will interact with other objects or people in computer-aware environments to make use of cloud-based services supported by Big Data and powerful analytics. While consumers are already using internet enabled devices to reap the benefits of social media and e-commerce, industry has only begun to explore ways in which connected devices can improve the safety and reliability of complex industrial processes; achieve greater energy and operational efficiencies; support sustainable consumption, create faster more cost effective means of communications; and improve the safety of medical devices and services and drive personalized medicine.

While some IoT devices will interact directly with the consumer and be designed principally for the consumer, others (such as connected airplane engines, wind turbines and locomotives) will operate principally in the industrial space and therefore involve a separate set of considerations on issues that are not tied to personal data. Other applications of the IoT and Big Data may involve the use of limited personal information or that information in the aggregate or will be one to many applications that are ill suited to scenarios of one-to-one consent. If the vast societal and economic benefits of the Big Data and the IoT are to be realized, the White House must embrace a broad vision and confront the opportunities and challenges with evidence-based work toward practical solutions that protect the individual, encourage responsible use of data, and foster robust innovation.

The Need to Focus on Transparency, Use and Security

Some Big Data and IoT applications will challenge traditional notions of how to apply privacy frameworks like the Fair Information Practice Principles (FIPPS) that have been in place since the 1970's. Those established frameworks serve us well, but much has changed since the era of centralized databases, highly structured data and relatively straightforward consumer transactions involving one buyer and one seller. Unlike the client server infrastructures of fifteen years ago, today's internet ecosystem contains an abundance of unstructured data, is highly transactional and thrives on a one-to-many model with many players including cloud services providers, intermediates routing traffic and establishing connectivity and entities providing enhanced security. Similarly, other industries using IoT devices, like the health care industry, now involve a complex network of providers, payor entities, product providers and service agendas, and researchers.

The advent of Big Data and the IoT compels policy makers, industry and civil society to confront traditional notions of how to operationalize notice and choice, security and accountability and consent. While we should not abandon the FIPPS, we do need to adapt, interpret and update them in a way that makes sense for the future environment.

Particularly in the context of the collection and use of Big Data in a complex connected ecosystem, we need to move away from an approach centered on specific notices and consents associated with the collection of data to focus in practical terms on what happens to that data and how it's used, bearing in mind the real world harms and consequences. That does not mean that there is no role for notice and choice, but rather that we must review the context of the implementation, and potential societal benefits from how the information may be used, to determine what controls are needed to protect privacy within the circumscribed use. We need to think through how we manage notice and choice - not to change existing privacy principles, but to provide more guidance about how to apply the existing principles through new innovative approaches for this new Big Data and IoT environment. One of the major areas of promise for Big Data lies in the ability to use correlation analysis to identify new topics of inquiry in endeavors of great societal benefit such as health care, the environment and urban planning. A correlation analysis that identifies such a field of inquiry, by definition cannot be predicated on a solely consent-based model as the query is generated by the correlation. We must explore how to create needed assurances and safeguards to replace consent in such circumstances; enhancing other FIPPs; applying technological, policy or legal controls; and enhancing transparency may be some of those ways.

The need to broaden our view of the elements of the FIPPs is particularly critical as we enter a computing continuum era of technology where many of the devices make it difficult or impossible for an individual to read something that looks like a privacy policy. Data aggregation from sensors and machine-to-machine communications - Big Data and the IoT - and the increased value from data analytics mean individuals will not always know who holds data relating to them. We must adapt notice to individuals to provide real time, context-specific information. These context-specific choices are something engineers, working alongside privacy and security professionals, can help bake into products.

Moreover, we must fully consider mechanisms to demonstrate the responsible use of data in contexts in which consent would be difficult to administer or place undue burden on the consumer. Unfortunately, notice paradigms may not be capable of providing consumers a full understanding how their data will be used in complex environments such as transportation logistics, health care product research or internet security, rendering a consumer’s consent less meaningful; moreover, consent may be impossible to obtain in many other contexts, such as support and development of improved industrial goods and services. Government and industry working together should continue to invest in finding creative and innovative consent mechanisms. These new mechanisms should provide individuals with easier and more automated methods for consenting where appropriate, while also protecting privacy in those contexts where consent is not possible. Finding these new mechanisms is where we need to spend time thinking through use cases and outcomes.

Therefore, it is important to more fully explore what are appropriate and accountable uses of data. A focus on accountability and use shifts the burden from the individual back to the organization that holds the data, as it encourages responsible behavior even for situations where consent cannot be obtained. This shift, in turn, will promote innovation and the development of new business models, while encouraging responsible behavior even for situations where consent cannot be obtained.

The complex data environment of the computing continuum will put an even higher priority on security. As technology allows entities to store more data relating to individuals, the threat of potential exposure of the information will increase. These threats require increased focus on the physical, technical and administrative mechanisms and technology tools used to secure data.

The future of technology shows us both an environment where we can no longer burden the individual with having to make choices about all the various issues concerning the processing of their data and a future of immense opportunity for societal benefits resulting from Big Data and the IoT. In sum, we must increase transparency and safeguard security while we work together to define appropriate uses of data.

Conclusion

Policy experts, technologists, academics and regulators have, on the whole, not succeeded in predicting the emergence and success of future business models. To avoid well-intentioned but unintended consequences, the White House needs to avoid unnecessary constraints on innovation or adoption of proscriptive polices to allow Big Data and IoT markets to develop. Industry stands ready to work together with government and civil society in the spirit of a true multistakeholder process to address these very important issues. The potential societal benefits of Big Data and the IoT are enormous and we need to ensure the policy framework to protect privacy and security supports trust and confidence in Big Data and the IoT going forward.

Very truly yours,

Daniel W. Caprio, Jr.
Senior Strategic Advisor

DWC

Microsoft Corporation
One Microsoft Way
Redmond, WA 98052-6399

Tel 425 882 8080
Fax 425 936 7329
http://www.microsoft.com/

March 31, 2014

Ms. Nicole Wong
Big Data Study
Office of Science and Technology Policy
Eisenhower Executive Office Building
1650 Pennsylvania Ave. NW.
Washington, DC 20502

Re: Government "Big Data" (FR Doc. 2014-04660)

Dear Ms. Wong:

Thank you for the opportunity to submit comments in response to the Office of Science and Technology Policy's Request for Information regarding big data. Microsoft applauds the Administration for initiating a comprehensive review of big data and its public policy implications and for seeking broad input through the RFI and the various public events. We are responding primarily to the first and third questions posed by the RFI.

Big data holds tremendous promise for society. Like many other developments enabled by technology, however, big data raises important public policy questions. We believe that government policy ought to be directed at promoting the learning that can be gained by analysis of big data while ensuring that privacy is not sacrificed in the process. The benefits enabled by analysis of huge datasets will not be possible if data is locked up in government or private silos. Therefore, we believe that government ought to consider carefully how best to ensure that, in general, data is broadly available to enable big data analysis. Some of this data will relate to people. To address that, government should strengthen—but also adapt—privacy law to enable the collection and use of large datasets while preserving privacy. This will no doubt entail tradeoffs on which reasonable people may have widely varying views. In determining how best to address privacy concerns, we believe government should draw upon a full range of tools—not only law, but also technology, articulated best practices and technical standards. We address these points below.

The Promise of Big Data

Advances in technology have led to the digitization of massive amounts of information. We're increasingly surrounded by sensors (in our smartphones, tablets, cars and even common appliances), all

constantly recording data. Vast amounts of data are generated in other ways across the private and public sectors. As the costs of computing and storage have dropped, it has become increasingly possible to collect, retain, aggregate and analyze all this data. Sophisticated analytical techniques enable researchers to detect trends and correlations among disparate phenomena and thereby make important predictions.

The key to unlocking the promise of big data is to enable the collection and broad availability of large volumes of data. Unlike random sampling, which was the foundation of statistical analysis in the 20th century, big data largely relies upon the collection of as much information as possible pulled from a variety of sources to create new, combined datasets. The collection of data may occur before anyone even realizes what insights may later be drawn from it. Using all of the data available related to a particular phenomenon—collected from various datasets over time—facilitates detecting patterns and improves the quality of prediction.

A few examples drawn from work done at Microsoft Research illustrate the point. More than ten years ago Microsoft researchers demonstrated the power of big data in natural language processing through their work to improve a grammar checker for Microsoft Word.[1] They began by pulling words—the data for their work—from a variety of sources, such as news articles, scientific abstracts, government transcripts and literature. They "trained" their grammar checker on increasingly large datasets drawn from these sources, and found that its accuracy greatly improved as the training dataset grew. For example, one of the grammar algorithms was only 75% accurate in predicting grammar problems when trained with a corpus of a half million words, but 95% accurate when trained with a corpus of a billion words.

More recently, Microsoft researchers were able to assist doctors aiming to understand HIV mutation by applying analytical methods first developed for an entirely different purpose—fighting email spam. HIV is hard to address because it is constantly mutating to avoid attack by the human immune system. Email spammers program their spam to constantly mutate too—to avoid email filters. By applying methods initially developed to analyze spam mutations, doctors could better understand how different immune systems respond to the mutations of the HIV virus.

While data about spam was helpful in researching HIV, search queries on Bing proved useful for other medical research—to discover potentially dangerous drug interactions. Microsoft Research worked with researchers at Stanford University on an analysis that identified side effects when a patient takes Paxil, a widely used antidepressant, together with Pravachol, a leading cholesterol-reducing drug.[2] Using Bing search engine logs, the researchers determined that people who searched on the names of both of those drugs had a much higher likelihood of also searching for diabetes-related side effects (such as

[1] Michele Banko and Eric Brill, "Scaling to Very Very Large Corpora for Natural Language Disambiguation," *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 26-33 (2001).
[2] Ryen White, Nicholas Tatonetti, Nigam Shah, et al. "Web-scale Pharmacovigilance: Listening to Signals from the Crowd," *J Am Med Inform Assoc*, doi:10.1136/amiajnl-2012-001482 (2013).

headache or fatigue) than a person who searched only for one of the drugs. This suggested a dangerous interaction between the two drugs that could result in diabetic blood sugar levels.

Government Should Promote the Broad Availability of Big Data

Government can play multiple roles in ensuring that society realizes the benefits of big data while other important values are protected.

Government is obviously an important source of data, about taxes, education, labor, defense, energy consumption, weather, health, communications, transportation, entitlement programs and more. It is a steward of that data as well. Access to data such as this will be important if new insights are to be gained and innovation promoted across disciplines as varied as health, security and economics. Government should establish policies with a view toward making data generally available, but limited as appropriate to address other important societal values. Information that essentially belongs to society as a whole, such as data describing the physical or economic world in which we live, ought to be made generally available in an efficient manner. Where data relates to individuals, government should employ a risk-based approach to assess the benefits of data sharing against harms to privacy, taking into account de-identification approaches and other techniques that may help to reduce privacy risks.

In its role as policymaker, the government should be cognizant of its duty to "promote the Progress of Science and useful Arts" (the constitutional basis for copyright) when considering the application of copyright law to large datasets. Large datasets and the aggregation of smaller pieces of data encourage uses that advance the progress of science and research, and computational uses of such data do not impinge upon the dataset owner's reasonable expectations. Such uses are generally allowed today because U.S. law does not provide inherent copyright protection for databases, and such a use would likely be considered permissible "fair use" in any event. In considering further developments in copyright law, government should strive to maintain approaches such as these that enable third parties to make use of large datasets in ways that are transformative and socially beneficial.

As an emissary to other governments, the U.S. government has a role to play in encouraging data to flow across borders while ensuring that privacy is respected. This will require the U.S. government to work closely with governments around the world to ensure that privacy regimes do not pose obstacles to the cross-border flow of data but rather appropriately protect the rights of people. Similarly, the U.S. government should work with other governments to promote copyright laws that foster access to data for all kinds of uses.

Government Should Strengthen Privacy Regulation and Adapt It for Big Data

Some large datasets will relate to people—their activities, interests, health and the like. Given the likely ubiquity of sensors in the years to come and the increasing digitization of so many human activities, there is a real risk that that data about people could be misused. We believe that the promise of big data will not be realized unless approaches are established to address privacy and civil liberties concerns.

Privacy regulation in the United States is a complex (yet incomplete) patchwork of federal and state rules that apply to particular industry sectors, particular types of data or particular data uses. While these rules are generally based on the Fair Information Practice Principles, there has been a heavy emphasis on the principles of notice and consent at the time of data collection. Today, however, the notice and consent paradigm has begun to show significant signs of strain under the weight of big data. We believe the notice and consent paradigm, and privacy regulation as a whole, should be strengthened, but also adapted to a big data world, in order to address this.

Today's heavy reliance on notice and consent places most of the burden of privacy protection on individuals. This is a problem. People are confronted with lengthy, detailed and often complex privacy statements from nearly every retailer, online service provider and other organization with which they interact. (In providing these long statements, organizations are aiming to comply with current law.) In theory, people would read these statements and then make informed choices on the basis of them. In practice, people often fail to do so, as they would be overwhelmed if they tried.[3] Instead, they quickly discard paper privacy statements and click through online statements to "agree" with the terms of privacy notices without reading the terms, much less trying to understand them. The law on the books is satisfied, but this is weak privacy protection.

There is a second problem: Even as the existing notice and consent paradigm may fail to provide real protection for people, it may serve to preclude beneficial uses of data that would present little privacy risk. Big data analysis often depends upon using datasets, often in combination with others, in ways that were not contemplated when the data was originally collected. If the original privacy notices did not foresee a beneficial use, the data may not be available for big data analysis.

To address all this, we believe that government should look at ways to focus use of notice and consent in those areas where decisions really can be informed and meaningful, and where privacy concerns are significant.

Some data uses are widely expected or understood, provide high potential societal benefit, or create a low risk of harm. For example, when people purchase an item over the Internet, they understand that the retailer will use the mailing address they provide in order to ship the item to them. Bloating privacy notices with detail about such uses distracts from disclosures that are more important.

Other data uses may entail a high risk of privacy harm and little societal benefit. It might be appropriate to generally preclude such uses, rather than allow them as long as "notice" is nominally provided in some multi-page legal document. For example, merely providing "notice" should not enable firms to use big data to discriminate against vulnerable communities in ways that would not be allowed in other circumstances.

[3] For example, it has been estimated that, on average, every American Internet user would have to spend 244 hours every year to read all the privacy statements he or she encounters. See Aleecia M. McDonald and Lorrie Faith Cranor, "The Cost of Reading Privacy Policies," 4 ISJLP 543 (2008).

In between these two cases are a wide range of potential data uses where reasonable people may not expect particular data uses, or may find particular uses objectionable. This is where it is important that people be provided with meaningful notice and an opportunity to consent, or not, to uses of data about them.

This approach could be complemented by adapting privacy regulation to take greater account of a broader range of Fair Information Practice Principles. Where notice and consent are unwarranted or impractical (as for data collected by small sensors), other protections should be called into play. These may include data security; maintenance of the confidentiality, integrity and availability of the data; data minimization through the use of de-identification techniques and mechanisms for transparency (beyond consumer notices) and other means of creating accountability for all data uses.

Privacy regulation could also be strengthened through the adoption of a more consistent and comprehensive approach to privacy law in the United States—one that reflects the broader and balanced approach to the Fair Information Practice Principles described above. Microsoft has long supported adoption of an omnibus, baseline federal privacy law. We believe this is the right approach because the increasingly complex patchwork of state and federal laws has resulted in an overlapping, inconsistent and incomplete approach to protecting privacy. This approach is confusing from the perspective of consumers, and unnecessarily burdensome for organizations.

The sectoral approach to privacy regulation that we have today may be even more problematic in the context of big data. As noted above, the value of big data is often realized when data is combined and analyzed in new ways. Yet such combinations may be precluded by differing privacy regimes applying to data first collected in varying sectors. A baseline federal privacy law could help ensure that all companies in the big data ecosystem are applying a clear set of responsible data practices, while also enabling the societal value of big data to be more easily realized.

Government Should Explore a Variety of Ways to Protect Privacy in a World of Big Data

The challenge of unlocking the value of big data while protecting the privacy of those whose data is included requires a multifaceted solution. The government should look to technology, best practices, law and standards to help address this.

Technology. Data privacy has traditionally been concerned with the collection, use and disclosure of data that identifies individuals. Privacy frameworks have generally divided data into one of two categories: data that does not identify individuals or data that does, with the assumption that most data is in one category or the other. We will likely need to abandon this binary model. In a world of big data, data that used to be considered non-identifiable, such as the colors of cars, their makes and models, and the times of day when they are on the road, might be used to identify people, especially when combined with other information about where people live and work. While perfect anonymization of big data sets may not be mathematically possible, a variety of techniques hold promise to greatly reduce the practical risk of privacy harm. As in other areas of public policy, government should consider the likelihood and severity of potential risks, and weigh them against other societal benefits.

It may be best to treat identifiability of data on a continuum. Some data is highly identifiable to anyone who sees it, and other data is nearly anonymous and requires significant computing power to re-identify, with gradations between the two extremes. Government should promote research on defining and advancing pseudonymization and de-identification techniques, which would help to address privacy concerns while enabling big data analysis. Better definitions of these concepts would help society know what promises can be made at different levels of de-identification. That would have two benefits: it would avoid overpromising what de-identification can achieve while encouraging the use of de-identification for what it does achieve. This research should build on recent work, such as work on *k*-anonymity and on understanding how easily big data can be re-identified using other sources of information.[4]

One helpful technique that should be explored further is called "differential privacy." Differential privacy does not rely on removing information from a database or changing it, but rather limits access to the underlying data and provides results to queries of the data that include random but small levels of inaccuracy, or distortion.[5] If the level of distortion is set correctly, the datasets can be usefully exploited without revealing information that could be used to re-identify individuals.

Government should explore greater use of cryptographic technologies as well. De-identification often involves cryptographically hashing information that directly identifies individuals, such as account numbers. More aggressive uses of cryptography may become practical over time. For example, current research is exploring how to use encrypted data sets to answer questions about the underlying data even though the data itself cannot be recovered, and some of the technologies are promising.[6] Further research may expand the uses of this research into wider application.

Many promising techniques for de-identification, encryption and differential privacy are still only in the research stage and further work to commercialize them should be encouraged. By recognizing the potential for misuse of big data and the risks of various levels of de-identification, policymakers could help society quantify the value that these technologies offer. That, in turn, would likely encourage industry investments in those potentially useful technologies.

Best Practices. Technologies such as those described above should be supplemented by best practices, in government and the private sector, regarding the use of those technologies. Such best practices would help to guard against attacks on the system or other attempts to circumvent privacy protection. For example, encryption is only as good as the practices put in place to safeguard the security of

---

[4] For more information about *k*-anonymity, see, Latanya Sweeney. "*k*-anonymity: a model for protecting privacy," *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems,* 10 (5), 557-570 (2002), and subsequent research.

[5] See, e.g., Cynthia Dwork, "Differential Privacy," *33rd International Colloquium on Automata, Languages and Programming, part II* (2006).

[6] For example, secure multi-party computation as described by Yehuda Lindell and Benny Pinkas, "Privacy Preserving Data Mining," *Journal of Cryptology*, 15(3):177-206 (2002) and Yehuda Lindell and Benny Pinkas, "Secure Multiparty Computation for Privacy-Preserving Data Mining," *Journal of Privacy and Confidentiality*, 1(1):59-98 (2009). Another example is structured encryption being developed by the MetaCrypt project at http://research.microsoft.com/metacrypt.

decryption keys. Pseudonyms can be compromised if they are independently mapped to identifiers, such as through look-up tables.

Best practices should be standardized and accompanied by robust audit mechanisms. Auditable standards for information security already exist and provide a way for organizations to document their policies that can be understood and verified by third parties. Audits should focus on the organizational controls that organizations put in place. For example, access by users of a dataset to other information that can re-identify individuals in the first dataset should be controlled and logged. Technology has a role to play in standardizing and enforcing these policies, including by associating or "tagging" policy requirements directly to datasets.

Law. Some technological solutions should be backed up by legal requirements. This is important because in many cases people may not know the identities of all the organizations that have access to data about them, much less effective notice of the privacy policies of those organizations. Uniform legal rules could help address this kind of risk. For example, if appropriate legal rules were established that generally prohibited the re-identification of de-identified data, de-identified data could be shared with greater confidence that privacy would be preserved. Any such rules should, of course, be carefully crafted to avoid locking in technologies now that may need to be changed quickly as knowledge of big data and its privacy implications progresses.

Standards. One way to provide technological flexibility is to craft law that relies upon industry technical standards. The Federal Information Security Management Act of 2002 (FISMA) is an example of this approach. To help implement FISMA, the National Institute of Standards and Technology (NIST) issued NIST Special Publication (SP) 800-53, which: (1) defines a risk management process; (2) specifies the risks that stakeholders must consider; and (3) provides lists of effective mitigations. The risks in the FISMA context are associated with information security, essentially the risk of loss, corruption or inappropriate disclosure of data. A standard for big data and privacy could adopt a similar approach, where the risks would include the improper re-identification or other illegitimate use of big data datasets (such as for illegal discrimination).

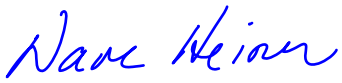
A standard for big data and privacy should seek to continuously improve how risks are addressed and to adapt to new risks, as FISMA and NIST SP 800-53 have done. Such a standard should describe a process of assessing risk, taking measures to reduce risk, assessing the effectiveness of the measures, and then returning to reassess risk. Finally, a successful standard should require any organization that follows it to document how it does so and to submit to third-party validation.

It may make sense for NIST to be responsible for the development of any such standard. NIST has considerable experience serving in this role in other contexts. If NIST were to undertake such an effort, it should canvass a broad set of stakeholders, from the public and private sectors, to identify the risks that any successful big data and privacy standard should address.

Conclusion

Microsoft appreciates the opportunity to comment on this RFI. We hope our comments will prove useful as the Administration continues its study of this important topic.

Yours sincerely,



David A. Heiner
Vice President & Deputy General Counsel, Legal and Corporate Affairs
Microsoft Corporation

[redacted]

From: Blackburn, Duane M. <dblackburn@mitre.org>
Sent: Monday, March 31, 2014 5:05 PM
To: bigdata@ostp.gov
Cc: Cook, Jim; Byrne, Rich; Hughes, Eric R.; Grewe, Barbara A.
Subject: Big Data RFI
Attachments: MITRE response to OSTP Big Data RFI.pdf

Nicole,

MITRE is pleased to submit the attached response to your RFI on *Government "Big Data".* As a not-for-profit operator of Federally-Funded Research and Development Centers (FFRDCs), our sole mission is to partner with the federal government to address issues of critical national importance. Moreover, because we only operate FFRDCs, we emphasize knowledge sharing as an important aspect of our public interest orientation. We apply what we learn from researching and evaluating industry trends and solutions with our experiences in addressing our various sponsors' domains to similar issues faced by other federal agencies. We approached our development of this RFI response from the same mindset: combine the insights gained from our experiences with government and industry, and our research, into a single, succinct, response.

Based on the tone set by the President in launching this effort, and the potentially public nature of RFI responses, we have purposefully approached this not as an opportunity to advocate for solutions, but as a tool to inform dialogue. We would welcome the opportunity to discuss any portion(s) of our response in greater depth with OSTP whenever you feel it would be beneficial for your project. At that time, we could explore more hypothetically a range of possibilities for future, deeper analysis and evaluation.

Thank you for the opportunity to share our insights with the OSTP team. Please do not hesitate to let us know how we could be of further assistance. You may reach out to me at dblackburn@mitre.org or (434) 964-5023; or Jim Cook, our Vice President and Director for the Center for Enterprise Modernization (CEM) at: (703) 983-2684, or via e-mail at: jecook@mitre.org.

Respectfully,

Duane Blackburn

Duane Blackburn
The MITRE Corporation
dblackburn@mitre.org
434-964-5023

http://www.linkedin.com/in/duaneblackburn

MP140200
March 31, 2014

Jim Cook
Eric Hughes
Rich Byrne
Barbara Grewe
Duane Blackburn

MITRE Response to OSTP "Big Data" RFI

For additional information about this response, please contact:

Duane Blackburn
dblackburn@mitre.org
(434) 964-5023

This page intentionally left blank.

Introduction

The MITRE Corporation is a not for profit company that runs Federally-Funded Research and Development Centers (FFRDCs) for the U.S. government. MITRE's FFRDCs serve agencies in a variety of areas that impact the public in direct and indirect ways, such as national security; aviation safety and administration; tax administration; homeland security; healthcare; benefits services (Veterans and Medicaid and Medicare); and other missions. We are pleased to respond to your request for information regarding big data based on our broad perspective gained from serving a variety of government missions, and from the unique perspective of a systems engineering company that combines a strong technical base with an informed awareness of the larger policy and context in which government operations are conducted.

As you are aware, there are many opportunities for both the private and public sector to use big data to have profound positive impacts on the public. Several factors make this possible: (a) the exponential growth of available data as a result of a largely "connected" populace; (b) the rapid expansion of data collection capabilities as more enterprises and the government move to on-line or electronic collection methods; and (c) the rapid emergence of technological capability to analyze big data, coupled with the cross-cutting relevance of this capability to government and industry.

While these factors expose more opportunities, their very nature also creates new challenges. At a macro-level, the relationship between statutes, policy, use, and technological capabilities must be more closely aligned if we are to take advantage of the tremendous potential of big data while not compromising privacy and undermining public trust. Being more clear and transparent regarding how big data will be used, the rationale and objectives for its use, and the public's rights and protections, would help develop public trust. Privacy protection options also need to be built into a system design in early stages rather than waiting until after a proof of concept has been developed, as has often been the case. These concerns are most critical for the federal government given its unique obligation for maintaining the public trust.

Question 1: *What are the public policy implications of the collection, storage, analysis, and use of big data? For example, do the current U.S. policy framework and privacy proposals for protecting consumer privacy and government use of data adequately address issues raised by big data analytics?*

The current U.S. policy framework embodies many important key principles:

- *Transparency* gives the public an understanding of which data is being used, as well as confidence in how and why it is being used.
- *Assurance* ensures that data is protected from theft, tampering or improper use, and that the data accurately represents what is it intended to represent.
- *Service* ensures that data about people and organizations is being used to provide things of benefit to them. This principle applies to the protections and

services the government provides its citizens. It also applies to the products and services industry provides to consumers. It is important to note the distinction, though, that government services are measured on their contribution to the public good, while private sector services are measured on their economic return.

- *Choice* is another principle that applies differently to government and the private sector. In government situations, individuals must comply with the law, so choice is limited to trying to effect change in the government, or facing the consequences of non-compliance. In industry, choice is manifested in contracts and in consumers' right to take their business to a competitor.

Current U.S. policies and privacy proposals represent good but imperfect attempts to apply the principles to a complex range of domains and contexts. For example, "notice" is a problematic approach to provide transparency, because individuals may not be able to understand the implications of things conveyed in a notice. Additionally, aggregated data raises more privacy concerns to the individual than they likely considered when granting authorization to a singular application. "Consent" is an approach to choice, which breaks down when the choice is not an informed one. Transparency and choice are imperfectly applied in situations such as people who are deceased, children, or people who are incapacitated by conditions such as Alzheimer's. Organizations sometimes erroneously deny services based on analysis of data, so rights of individuals need to be protected in these situations to uphold the service principle. Big data calls for rethinking our approaches to these and possibly other principles, rather than, as some are calling for, a devaluation or discarding of the principles themselves.

With big data analytics, it is often possible to identify sensitive attributes of an individual by combining multiple types of data. Policies that attempt to deal with this in specific contexts often drift out of alignment with key principles. For example, the Computer Matching Act rightly attempts to provide transparency and assurance in use of multiple datasets by government programs, but at times significantly reduces the ability of agencies to deliver intended services. In national security, transparency of mission and objectives is vitally important, but transparency of detailed operations would compromise assurance and service.

Implications of U.S. policies and privacy proposals also must be considered in a global context. The policies and practices we develop to govern uses of big data must reinforce and build on, not limit or deter, opportunities for innovation and improve the efficiency and effectiveness in order to drive economic competitiveness. They must also address the fact that data is not bounded by geographic boundary lines and inconsistent policies across those boundary lines at a minimum cause confusion and at worst nullify any protections built into our policies.

Question 2: *What types of uses of big data could measurably improve outcomes or productivity with further government action, funding, or research? What types of uses of big data raise the most public policy concerns? Are there specific sectors or types of uses that should receive more government and/or public attention?*

Big data can significantly improve outcomes across a broad range of missions. The types of uses of big data that raise the most concerns are areas where policies are less well aligned with key principles. In some cases, the benefit of big data accrues disproportionately to businesses or agencies rather than to individuals, resulting in misalignment with the service principle.

There are many potential uses of big data – each with the potential for improving outcomes, so long as related policy changes are aligned and transparent. These do not represent a comprehensive list, nor are we advocating directly for these potential uses – they are served here as thought-provoking examples. Our only hope is that the incredible potential that exists is not swept aside because the policy challenges are difficult. Instead, because the potential is so great, it is in the public interest that these challenges be tackled openly and with persistence.

The following examples reflect both significant potential, but also draw the most public concern:

- Use of big data to improve healthcare quality and patient safety could dramatically improve the health of people and the U.S. economy. Efforts to analyze across the many types of healthcare data (rather than just focusing on collection of data) seem most likely to improve health outcomes and reduce healthcare costs.
- Big data could be used more widely to disambiguate identities for appropriate medical care and accountability, access to entitlements and benefits, and law enforcement. Big data might also be used to correct errors in personal data.
- Big data could be used to significantly identify the potential for, and detect and reduce fraud, waste and abuse in healthcare, financial services, and government programs.
- Big data could be used to make government operations more efficient, effective, and transparent. Implemented correctly, the use of Big Data has the potential to reinforce "government as a servant" rather than "government as big brother".

Some big data analytics efforts raise public policy concerns, especially where the value (benefit/cost) accrues primarily to a business or agency. Institutional Review Board (IRB) processes that explicitly highlight who benefits from human subjects research might provide useful ideas for consideration of big data analytics efforts. Risk-benefit consideration should not replace informed consent (transparency).

There is a growing concern over how to determine whether analysis results have sufficient validation, especially when big data is used. This can be viewed as misalignment with transparency, assurance and service principles, whether in policies or day-to-day governance of an activity.

Question 3: *What technological trends or key technologies will affect the collection, storage, analysis and use of big data? Are there particularly promising technologies or new practices for safeguarding privacy while enabling effective uses of big data?*

One concerning trend is locating infrastructure and platform service providers outside the U.S. or even in international waters (e.g., floating data centers). These arrangements draw into question which laws and policies apply. This trend can negatively impact transparency (e.g.; does the individual know where their data is held and thus which policies apply), assurance (under what regime is their data protected) and choice (can I demand that my data be kept within the U.S.).

Computing on encrypted or masked data is a promising trend for protecting privacy while enabling effective use of big data. De-identification is useful in many contexts but because data can often be re-identified should not be used to exclude datasets from privacy policy and regulatory regimes. These approaches may allow the same level of service to be delivered with greater levels of assurance because fewer personnel would have access to data elements that could cause harm if disclosed.

Some privacy professionals are calling for a "risk-based" approach to privacy as a substitute for one grounded in fundamental principles of individual rights and due process. Key principles should remain foundational, and risk should be assessed according to the service (benefit) provided, balanced by the challenges to assurance and transparency.

Question 4: *How should the policy frameworks or regulations for handling big data differ between the government and the private sector? Please be specific as to the type of entity and type of use (e.g., law enforcement, government services, commercial, academic research, etc.).*

Policy frameworks for government and the private sector should begin with the same key principles. As noted under Question 1, the choice principle has very different application in government as compared to the private sector. Hybrid public-private arrangements should also be considered (e.g., government agencies using private companies to perform eligibility checks). In accordance with the assurance principle, individuals should have means to correct data pertinent to them, whether held by government (e.g., marital status for eligibility) or by private industry (e.g., financial status).

Government use of big data must meet a higher standard in accordance with these principles. Government sets the policy by which government agencies must operate, so any shortcoming can undermine public trust. Policies that are open to interpretation in ways inconsistent with key principles undermine use, and reduce government's effectiveness.

March 31, 2014

Office of Science and Technology Policy
Eisenhower Executive Office Building
1650 Pennsylvania Avenue
Washington, DC 20504

RE: Office of Science and Technology Policy Request for Information: Big Data

To Whom It May Concern:

Mozilla submits these comments in response to the March 4, 2014 Notice of Request for Information on Big Data.

These comments emphasize that we as a multi-stakeholder Internet society are in the early stages of understanding big data, counseling deliberation and patience; note the complexities associated with surveillance and accentuated by big, global data; and suggest ambitious research and development to advance big data opportunities.

We greatly appreciate the Administration's efforts to lead a multi-stakeholder process to explore big data issues through this request for information and the three workshops.

On behalf of Mozilla, we thank you for the opportunity to comment on this request for information. Please do not hesitate to contact us with questions or for additional input.

Respectfully Submitted,

/s/

Alexander Fowler, Global Privacy and Public Policy Lead
M. Chris Riley, Senior Policy Engineer

Mozilla
2 Harrison St
San Francisco, CA 94105

Comments on the Office of Science and Technology Policy Request for Information: Big Data

Prepared by Mozilla and Submitted on May 31, 2014

I. Introduction

Mozilla appreciates the opportunity to provide input to the Office of Science and Technology Policy (OSTP) for its "Big Data" review process. The expansion of massive data gathering, storage, and processing capabilities, together with the reduction in cost, are having a significant impact on society's collective understanding of the proper norms and approaches in this space. This OSTP process can build on top of existing academic, industry, and civil society efforts to catalyze next steps forward on research and multi-stakeholder discussion of these issues.

Mozilla is a global community of people working together since 1998 to build a better Internet. As a non-profit organization, we are dedicated to promoting openness, innovation, and opportunity online. Mozilla and its contributors make technologies for consumers and developers, including the Firefox web browser and Firefox OS phone used by more than half a billion people worldwide. As a core principle, we believe that the Internet, as the most significant social and technological development of our time, is a precious public resource that must be improved and protected.

Privacy and security are important considerations for Mozilla. They are embraced in the products and services we create, and derive from a core belief that consumers should have the ability to maintain control over their entire web experience, including how their information is collected, used and shared with other parties. We strive to ensure privacy and security innovations support consumers in their everyday activities whether they are sharing information, conducting commercial transactions, engaging in social activities, or browsing the web.

At the outset, it is important to establish improved trust as the goal of big data policy.

This process is not about best practices to extract maximum revenue from big data sets, or disrupting industries. Instead, the challenge is to address the policy and normative concerns that arise from big data, and to understand how frameworks for privacy and trust developed for a different world should extend and adapt to this one. Within this challenge, the fundamental risk associated with big data policy is trust in the global Internet and information ecosystems, and whether the world's people and businesses will continue to participate in these ecosystems and realize their benefits for social, economic, and political activity.

These comments will raise three main themes responsive to the questions articulated by OSTP:

1. First, we as a multi-stakeholder community working on, and with, big data are very early in the process of understanding how everything works, including how to apply the norms developed for an earlier data and privacy world, and where both fuzzy and bright lines should be drawn around data handling practices that support innovation and growth, on the one hand, while preserving user control and driving public benefits.
2. Second, we must tackle head-on the heightened sensitivities and trust risks associated with government access to personal data, or we will not have a strong, internally cohesive, collaborative community to tackle these issues.
3. Finally, we should think "big." Big data presents big problems, but there are also big opportunities, and we should embrace them, not disregard ambitious or long-term solutions.

II. Policy Challenges that Arise from or are Amplified by Big Data

This section highlights a few key issues that, while not exhaustive, we think are the most important for the OSTP to grapple with to start:

A. the complex nature of "personal" information,
B. the significance of data portability,
C. growing concerns over balkanization of data systems, and
D. the inherently unique nature of government collection of and access to data.

Across these issues, the basic concepts of the Consumer Privacy Bill of Rights persist, including the value of control and transparency and the importance of context. Trust remains the ultimate normative goal. Yet, these all carry somewhat different meanings and implications in the big data world, and we don't yet have a full understanding of how to apply and implement them.

A. Properly Defining "Personal" Data

Responsive to Question 1: What are the public policy implications of the collection, storage, analysis, and use of big data? For example, do the current U.S. policy framework and privacy proposals for protecting consumer privacy and government use of data adequately address issues raised by big data analytics?

One fundamental challenge to getting data policy right is changing the binary concept of "personal" and "not personal".[1] Big data policy will struggle to be accurate if it rests on a broken categorization of unit data. This is responsive to Question 1, in that big data can amplify and extend the policy problems that arise from a misconceptualization of unit data, and in that big data creates new nuances of personalization arising from combinations of unit data that may not be "personal" in isolation or, importantly, fall under existing regulations.

The "Respect for Context" principle of the Consumer Privacy Bill of Rights touches on this, if interpreted dynamically. In its original formulation, the "context" is explained as the purposes and business processes that generate the data. In the big data world, those same processes also combine the data with other data – about the same individuals and about other individuals – in ways not always made transparent to each subject. This is a new kind of context, relevant to privacy and data policy in the same way other contexts are, but which may necessitate rewording or reformulating of the original principle text.

Mozilla's previous filing with the Federal Trade Commission articulated an Internet user's social graph as an example of aggregations of individual elements of less personal data that in combination become more personal.[2] Extending that example into the big data world, a combination of millions of Internet users' social graphs further changes the analysis. And yet, that combination is precisely what is being done be a number of private sector, intelligence, and law enforcement actors. Certainly, the subject matter of such work must be considered "personal" to some degree.

Ordinary web browsing activity includes a great deal of "personal" data, according to our users. For example, Firefox and other browsers store the list of URLs representing a user's web browsing history for that person's use and reference. An ordinary user views that information as "personal," though such metadata is outside the scope of regulation.

---

[1] Comments of Mozilla, Federal Trade Commission, Protecting Consumer Privacy (Feb. 23, 2011), http://www.ftc.gov/sites/default/files/documents/public_comments/preliminary-ftc-staff-report-protecting-consumer-privacy-era-rapid-change-proposed-framework/00480-58110.pdf

[2] *Id.* at 5.

Over the past decade, very few new forms of PII have been articulated, but the scope of such "personal" information has grown. Additionally, machine-generated data built from ordinary activity, coupled with non-personal data, can uniquely identify a user and dynamically personalize her use of an online service, through pricing, advertising display, and other activities that feel very "personal."

Adapting and expanding the scope of "personal" data is an essential precursor to a proper understanding of privacy in a big data world.

B. Promoting Meaningful, Protected Portability

Responsive to Question 3: What technological trends or key technologies will affect the collection, storage, analysis and use of big data? Are there particularly promising technologies or new practices for safeguarding privacy while enabling effective uses of big data?

One trend is a growing number of silos, vertically and horizontally integrated services, applications, and devices that generate a diverse and broad amount of user data, and share and analyze that data across integrated and partner services. This is responsive to both question 3 and question 1, in that it is a significant technology development with ramifications for big data policy generally.

Integration has created a world where the concept of "big data" is salient even within a single company. Such integration can obviate the financial or technical need for portability of data, and make it easier, or cheaper, to lock data within a single organization. The result is often reduced transparency and user control over personal data and combinations of data with personal impact. It can also create policy problems above and beyond the traditional locus of privacy, transparency, and control issues, such as competition, innovation, and economic growth challenges.

The contrary vision is one in which data sets are portable and not locked within a single company or ecosystem. Portability necessitates a degree of transparency and enables control through choice of platform and environment.

Mozilla is committed to developing, promoting, and promulgating interoperable and standards-driven technologies, and to opposing silos and walled gardens of data and technology. Further OSTP study of data portability and its relation to big data and to data and service silos would greatly help advance the vision of the Consumer Privacy Bill of Rights in the big data world.

C. Resisting Barriers and Balkanization

Responsive to Question 5: What issues are raised by the use of big data across jurisdictions, such as the adequacy of current international laws, regulations, or norms?

One opposite value to portability is integrated, locked ecosystems, as described above – pressures endogenous to the dataset and those processes operating on it. A related, yet highly distinct force is balkanization and division at geographic or regional boundaries – pressures generally exogenous from the data set itself. These forces have been increasing in past months and years as a result of significant economic opportunities for nations through internalizing larger segments of the Internet ecosystem, as well as defensive responses to revelations of expansive surveillance. These issues are responsive to question 5 in that they color global policy design and harmonization for big data, and create proximate challenges to building and using global big data systems.

External drivers that would mandate national borders for data or introduce restrictions to data portability represent a major concern for nascent big data policy and for current systems. As the OSTP process will no doubt reiterate time and time again, big data is at an early stage of understanding. Technical and legal mechanisms that hamstring its growth by imposing artificial and unnecessary barriers serving parochial interests must be avoided and resisted. OSTP can, and should, support further study of the harmful impact of such barriers, and the benefits of safe harbors and other mechanisms to advance open information flow across jurisdictions, and should advocate for open data flow within Administration policy processes and discussions.

D. Restoring Trust by Fixing Surveillance Practices

Responsive to Question 4: How should the policy frameworks or regulations for handling big data differ between the government and the private sector? Please be specific as to the type of entity and type of use (e.g., law enforcement, government services, commercial, academic research, etc.).

Government use of big data carries with it one major contextual difference from private sector use: there is usually no inherent optionality. Although caveats and qualifiers abound, in general, people perceive a choice to use or not use services offered by the private sector, and therefore a fundamental ability to escape those data collection and use mechanisms. People do not have a choice to opt out of being subject to law enforcement or intelligence activities. This doesn't mean private sector use of big data doesn't raise major and legitimate concerns. But it does mean that private sector use is

fundamentally different from government use. It means that some concerns arising from misuse of personal data in all its forms, as well as the harm of not knowing how data is being used, are heightened when it is the government using it. This is responsive to question 4 in that respecting these heightened concerns demands that government use, particularly those uses that are not optional, be viewed differently and separately from private sector use of big data.

Two issues have arisen in the context of government surveillance practices that are salient to this point. First, one of the major objectives behind ECPA reform efforts[3] is addressing the third party doctrine, the notion that data voluntarily ceded to a private sector company loses privacy rights with respect to future sharing of that data with the government. In a world where government use of data and private use of data present different normative balances of interests, this concept is out of date and needs to be changed - particularly where increasingly frequent National Security Letters (NSLs) result in divulgence of the privately held data without any legal pathway to inform the subject. Second, specific government surveillance conducted through interceptions of data center communications[4] represent a direct method for government access to private sector held data, without the knowledge or assistance of the company that collected and transferred the data. These issues render it difficult if not impossible to implement appropriate differences in policy between government and private sector access to, and use of, big data – counter to the policy need to respect heightened concerns associated with government use.

Overall, to the extent that the objective of this OSTP process is to encourage government and private sector collaborative efforts to shape policies and best practices for big data, and that establishing and defending trust in that ecosystem must be a key goal, surveillance and surveillance reform have a proximate, significant impact.

III. Moving Forward

The outcome of this OSTP process is unlikely to be any line drawing as to what is or is not a good practice with regards to big data – and that's proper at this stage, because much more work needs to be done to understand the context. Part of the path forward,

---

[3] *See, e.g.*, Rainey Reitman, "Deep Dive: Updating the Electronic Communications Privacy Act," *Electronic Frontier Foundation* (Dec. 6, 2012), at https://www.eff.org/deeplinks/ 2012/12/deep-dive-updating-electronic-communications-privacy-act.

[4] Barton Gellman and Ashkan Soltani, "NSA infiltrates links to Yahoo, Google data centers worldwide, Snowden documents say," *Washington Post* (Oct. 30, 2013), at http://www.washingtonpost.com/world/national-security/nsa-infiltrates-links-to-yahoo-google-data-centers-worldwide-snowden-documents-say/2013/10/30/e51d661e-4166-11e3-8b74-d89d714ca4dd_story.html.

therefore, is more research on technologies and policies of big data. Another part is continued multi-stakeholder engagement as the industry evolves, to avoid determining policies in a vacuum. OSTP has a facilitating, sponsoring, and convening role to play for both of these directions.

Collective understanding and development of big data policy is at an early stage, as are the technologies, options, and social, political, and market structures around big data. But it's not too early to have positive impact on the future of big data. Research and experimentation will help – on better architectures to advance the principles of transparency and control, on a better understanding of social perspectives around big data, and on legal and policy systems to improve trust. Changes in data policy generally – particularly around the evolving concept of "personal" data, surveillance practices, and growing challenges with portability and barriers – will have an impact wherever little data combines into big.

OSTP and every commenter participating in this proceeding have a mutual opportunity and a collective responsibility to work towards improving public awareness and literacy around big data, its technologies and policies. Meaningful, not merely superficial, public engagement with these issues as they develop would prove hugely helpful to advancing the public interest.

Finally, the big challenges of big data demand big thinking. Even in those dimensions that have yet to bear fruit, such as tagging data to improve transparency and control, it is too early to give up, and more investment may yet produce huge positive returns. Any combination of policy and law that can help make big data more tractable would be worth the efforts involved.

CENTER FOR URBAN
SCIENCE+PROGRESS

1 MetroTech Center, 19FL
Brooklyn, New York 11201
tel: 646.997.0500
fax: 646.997.0560
web: cusp.nyu.edu

Ms. Nicole Wong
Office of Science and Technology Policy
Eisenhower Executive Office Building
1650 Pennsylvania Avenue NW
Washington, DC 20502

March 31, 2014

RE: Government "Big Data"; Request for Information, 79 Fed. Reg. 12251 (March 4, 2014)

Dear Ms. Wong:

On March 4, 2014 the Office of Science and Technology Policy published a Request for Information soliciting responses to five questions that will help inform the review by senior government officials of the ways in which "big data" affect how Americans live and work, and the implications of collecting, analyzing and using such data for privacy, the economy, and public policy. New York University's Center for Urban Science + Progress (CUSP) appreciates this opportunity to provide these comments in order to help consider the implications of "big data."

On April 23, 2012, New York City and New York University formed CUSP as part of the Applied Sciences NYC initiative. CUSP is about the intersection of two simple but compelling themes: cities and data. Cities are where the people are; about half of humanity lives in urban environments today and that number will grow to 80 percent by the middle of this century. The ability to ability to collect, transmit, store and analyze data is rapidly growing, and if properly acquired, integrated, and analyzed. Big data can take governments beyond today's imperfect and often anecdotal understanding of cities to better operations, better planning and better policy.

While the focus in recent years has been on the exploitation of big data for commercial and national security purposes, CUSP believes data science has much to contribute to the public good. Data science can help address major questions about urban infrastructure, the urban environment, and the interactions of people with each other, with institutions, their interactions as organizations, as well as their interactions with the built and natural environments. Yet it is the ever finer temporal and spatial granularity of data about individuals and the increasing power of informatics tools to combine and mine these streams of data that stoke concerns about privacy and data access, particularly when these tools are in the hands of individuals or organizations whose interests are not perceived as being aligned with those of the data subjects. CUSP is dedicated to developing the big data tools that will help cities and the City of New York, in particular, become more productive, livable, equitable and resilient.

CUSP commends the Administration for providing a forum where issues such as the way "big data" will affect the way we live and work, the relationship between government and citizens, and how public and private sectors can spur innovation and maximize the opportunities and free flow of this information while minimizing the risks to privacy can be explored.

(1) What are the public policy implications of the collection, storage, analysis, and use of big data? For example, do the current U.S. policy framework and privacy proposals for protecting consumer privacy and government use of data adequately address issues raised by big data analysis?

Beginning with Fair Information Practice Principles (FIPPs) in 1973,[1] the underpinnings of the current U.S. privacy framework are collection limitation, purpose specification, use limitation, notice, and choice. Embodied in sector-specific privacy laws, such a framework focuses primarily on the collection or disclosure of information, not on its use.[2] In general, organizations are required to specify the purposes for which they will use the data they collect, collect only that data needed to achieve those purposes, and use the data only for those specified purposes. In addition, organizations are required to provide data subjects with privacy notices and disclose non-public personal information to third parties only if data subjects first are provided with the ability to opt out of such sharing.

The past two decades have seen rapid advances in embedded sensors, wireless communication, database technologies, search engines, data mining, machine learning, statistics, distributed computing, visualization, and modeling & simulation. These technologies, which collectively underpin "big data" and allow organizations to acquire, transmit, store, and analyze all manner of data in greater volume, with greater velocity, and of greater variety. Designed with traditional administrative or transactional data in mind, current privacy frameworks are particularly challenged by sensor and imagery data that are generated as byproducts of other primary activities, often collected without asking explicitly and thus without any assumed permission to use beyond uses compatible with the purpose for which the data was collected. Given technological advances, the exploitation of data originally acquired for another purpose can be considered a hallmark of big data.

The problem with the current articulation of these privacy frameworks is that too much emphasis is placed on action by the individual, particularly with respect to notice and choice. In 1970 Alan Westin wrote that "privacy is the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others."[3] Furthermore, for the reasons discussed above, if the organization using the data cannot specify the purpose for which the data is being collected and should not be limited in the data it collects and in the use of the data, then "a law should mandate a minimum floor of safe data-handling practices on every data handler in the U.S."[4] This position is consistent with the accountability principle found in the OECD Guidelines[5] and the right

---

[1] U.S. Department of Health, Education and Welfare, *Records, Computers and Rights of Citizens: Report of the Secretary's Advisory Committee on Automated Personal Data Systems* (1973).

[2] Paul Ohm, *Changing the Rules: General Principles for Data Use and Analysis,* to be published in *Privacy, Big Data, and the Public Good: Frameworks for Engagement* (Cambridge University Press 2014) [hereinafter *Ohm*).

[3] Alan Westin, *Privacy and Freedom*, abstract reprinted in Daniel Solove and Paul Schwartz, *Privacy, Information, and Technology, p*41 (Wolters Kluwer 2d ed. 2009).

[4] Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. Rev. 1701, 1762 (2010).

[5] *Organization for Economic Co-Operation and Development Guidelines on the Protection of Privacy and Transborder Flows of Personal Data* available at http://www.oecd.org/internet/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderdataflowsofpersonaldata.html [hereinafter *OECD Guidelines*].

to accountability found in the Administration's proposed Consumer Privacy Bill of Rights.[6] The OECD Guidelines state in part:

> "A data controller should
> a) Have in place a privacy management programme that
> . . .
> ii. is tailored to the structure, scale, volume and sensitivity of its operations;
> iii. provides for appropriate safeguards based on privacy risk assessment;
> iv. is integrated into its governance structure and establishes internal oversight mechanisms;
> . . .
> vi. is updated in light of ongoing periodic monitoring and periodic assessment . . . ."[7]

This is a risk based approach. Any privacy framework governing big data "should be calibrated to the sensitivity of the information stored in the database, meaning according to the risk of harm to individuals or groups. Certain types of data bases are riskier than others . . . ."[8]

The proposed accountability principle recognizes that "[p]rivacy protection depends on companies being accountable to consumers" and states that: "[c]onsumers have a right to have personal data handled by companies with appropriate measures in place to assure they adhere to the Consumer Privacy Bill of Rights.[9]

The development of accountability principles has led to an articulation of the elements of an accountable organization:

- Organization commitment to accountability and adoption of internal policies with external criteria,
- Mechanisms to put privacy policies into effect, including tools, training and education,
- Systems for internal ongoing oversight and assurance reviews and external verification,
- Transparency and mechanisms for individual participation, and
- Means for remediation and external enforcement.[10]

Accountability based programs are responsive to changing business models and new technologies and are appropriate to their particular context.[11] Thus, an accountability based framework has the flexibility necessary to address the demands of big data while, by establishing robust governance processes, reducing the likelihood of harm to individuals or groups.[12]

Rather than focusing on the data subject, the burden should be shifted to the user of the data. The focus should be on the use of the data and holding users accountable. This approach is reflected in *Data Protection Principles for the 21st Century* (the "Proposed OECD Principles") which proposes revisions to

---

[6] *Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation,* p10 (2012) [hereinafter *Privacy Bill of Rights*].

[7] *OECD Guidelines.*

[8] *Ohm, p8.*

[9] *Privacy Bill of Rights, pp*21, 22.

[10] *Essential Elements of Accountability* available at http://informationaccountability.org/.

[11] Mary Culnan, *Accountability as the Basis for Regulating Privacy: Can Information Security Regulations Inform Privacy Policy? p7*, available at http://www.futureofprivacy.org-/wp-content/uploads/2011/07/Accountability%20as%20.

[12] *Id.*

the 1980 OECD Guidelines.[13] The Proposed OECD Principles remove the concept of limitation from both the collection and use principles, the purpose specification principle has been eliminated, and the emphasis on notice and choice is reduced. Instead, the Proposed OECD Principles focus on the role of data stewards and on holding them liable for reasonably foreseeable harms to individuals and on demonstrating accountability (see Appendix).

The Proposed OECD Principles make accountability the core mechanism of protecting information privacy rather than informational self-determination.[14] The benefits of this approach are that it is technology agnostic and is flexible enough to adjust as innovations develop beyond what currently can be imagined. For laws to remain relevant they must be capable of adapting to change. The Proposed OECD Principles have retained those elements of the FIPPs which are still applicable and have discarded those elements which are no longer pertinent.

(2) What types of uses of big data could measurably improve outcomes or productivity with further government action, funding, or research? What types of uses of big data raise the most public policy concerns? Are there specific sectors or types of uses that should receive more government and/or public attention?

The Open Government Initiative, pursuant to the Executive Order dated May 9, 2013,[15] and the Open Data Policy[16] issued thereunder, require the federal government to make data resources easy to find, accessible and usable.[17] This goal is accomplished by making the default state of federal government structured information machine readable and open formats. [18]

The term "open data" refers to publicly available data structured in a way that enables the data to be fully discoverable and usable by end users.[19] The Open Data Policy sets out seven criteria for government open data:

- *Public*. A "presumption of openness" is required unless there are privacy, confidentiality, security or other valid restrictions on a particular dataset.
- *Accessible*. Convenient, modifiable, and open formats should be used so that open data can be retrieved, downloaded, indexed, and searched. Formats should be machine-readable (i.e., data are reasonably structured to allow automated processing).
- *Described*. Consumers of the data have sufficient information to understand the data.
- *Reusable*. There are no restrictions on the use of the data.
- *Complete*. Data is published in its original forms with the finest possible level of granularity practicable. Derived or aggregate data must reference the original data.
- *Timely*. Data are made available as quickly as possible.

---

[13] Fred Cate, Peter Cullen, Viktor Mayer-Schonberger, *Data Protection Principles for the 21st Century* available at http://www.microsoft.com/en-us/downloads/confirmation.aspx?id=41191 [hereinafter *Proposed OECD Principles*].

[14] Resources are available to assist in demonstrating that an organization is accountable. *See The Privacy Office Guide to Demonstrating Accountability* available at www.nymity.com/demonstratingaccountability.

[15] *Executive Order – Making Open and Machine Readable the New Default for Government Information* (5/19/13) available at http://ww.whitehouse.gov/the-press-office/2013.05/09/executive-order-making-open-and-machine-readable-new-default-government-information [hereinafter *Executive Order*].

[16] *Open Data Policy – Managing Information as an Asset* (5/9/13) available at http://whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf [hereinafter *Policy*].

[17] *Executive Order*, section 1.

[18] *Id.*

[19] *Policy, p* 4.

- *Managed Post-Release*. A point of contact must be designated to assist with issues after the data has been made available to the public.[20]

The Open Government Initiative is primarily prospective in nature. With respect to existing datasets, agencies are encouraged to make them "open" by "prioritizing those that have already been released to the public or otherwise deemed high-value or high-demand through engagement with customers."[21] Agencies are cautioned to "exercise judgment before publicly distributing data residing in an existing system by weighing the value of openness against the cost of making those data public."

Rather than encouraging agencies to make existing datasets "open," agencies should be required to do so unless there is a compelling reason not to. An example of existing datasets that should be open are online archives. Currently, online archives generally are available for an arbitrary number of years. If online archives are electronically available, they should be made available publicly for as far back as they are kept in electronic format. For example, budget data is only publicly available for five years but is electronically available for much longer; all electronically available budget data should be open. Making data easily available can fall victim to understandable causes such as overtaxed staff, a failure of imagination within the agency that anyone would want all the available data, or aesthetic considerations in design of a webpage. As one small example, federal budget data is supposed to be one of the most readily available types of data. One can easily find all relevant Congressional documents for the appropriations process in an electronic format (pdf) on the Library of Congress' Thomas website back to 1998,[22] yet the Department of Energy makes only the most recent 10 years of its Congressional Budget Justifications available on its CFO's website.[23] The Department has the documents in electronic format back to 1977, the year the Department was established and could very easily post them, as the US EPA does back to 1967.[24]

The Open Data Policy requires agencies to describe information using the Common Core Metadata Schema (the "Schema").[25] Metadata is structured information that describes, explains, locates or otherwise makes it easier to retrieve, use or manage an information resource.[26] The Schema establishes a common vocabulary by defining and naming standard metadata fields so that a data consumer has sufficient information to process and understand the described data. Some of the "common core" fields are: title (human-readable name of the asset in plain English and in sufficient detail to facilitate search and discovery), description (human-readable description with sufficient detail to enable a user to quickly understand whether the asset is of interest), and keyword (terms that help technical and non-technical users discover the dataset). While the fields are standardized, the content of them is not.

Different agencies use different terms to refer to the same thing. Not having a common taxonomy complicates data access. The National Archives, NTIA, and/or the Library of Congress should map terms

---

[20] *Id*. p5.
[21] *Id*.
[22] The Library of Congress Thomas, *Status of Appropriations Legislation for Fiscal Year 2014*. Available at http://thomas.loc.gov/home/approp/app14.html (accessed September 28, 2013).
[23] Energy.Gov, Office of the Chief Financial Officer, U.S. Department of Energy, Budget (Justification & Supporting Documents). Available at http://energy.gov/cfo/reports/budget-justification-supporting-documents (accessed September 28, 2013).
[24] U.S. Environmental Protection Agency, *Historical Planning, Budget, and Results Reports*. Available at http://www2.epa.gov/planandbudget/archive#Justification (accessed September 29, 2013). Since EPA was established in 1970, this includes 3 years of budget data from predecessor agencies.
[25] *Policy, p* 6.
[26] *Common Core Metadata Schema* 1 available at http://project-open-dat.github.io/schema/.

that have the same meaning and create a standard taxonomy in order to make data more navigable. This problem is magnified when a user tries to link up local, state and federal government datasets.

The City of New York provides a clear example, but similar cases abound. The primary identifier for New York City data on buildings is the BBL # (Borough Block Lot Number). This number corresponds to a lot in the city (multiple buildings can be on one lot). Each building can also have a BIN (Building Identification Number). Multiple BINs can correspond to a BBL #. However, BINs are not widely used across datasets, forcing researchers to rely on BBL #s as a method of matching up data. As a result, researchers are often forced to use the lot level when building level data could be available. New York City's open data law, Local Law 11, suffers from the same deficiency as the federal Open Data Policy, a single taxonomy is not used. Until terms are mapped to create a common language at all levels of government, open datasets only will have limited accessibility.

(4) How should the policy frameworks or regulations for handling big data differ between the government and the private sector? Please be specific as to the type of entity and type of use (e.g., law enforcement, government services, commercial, academic research, etc.)

The current system of protection for research involving human subjects is found in the Federal Policy for the Protection of Human Subjects, or the Common Rule, which has been codified in separate regulations by 15 federal departments and agencies based on the United States Department of Health and Human Services codification.[27] The Common Rule outlines the basic provisions of institutional review boards (IRBs) and informed consent. Each institution which is engaged in research covered by the Common Rule and which is conducted or supported by a federal department or agency must provide written assurance that it will comply with the requirements set forth in the Common Rule. In order to approve research covered by the Common Rule, the IRB must determine that the risks to individual research subjects are minimized, risks to subjects are reasonable in relation to anticipated benefits, selection of subjects is equitable, informed consent will be sought and appropriately documented, the research plan makes adequate provisions to protect the privacy of subjects and to maintain the confidentiality of data, and when some or all of the subjects are likely to be vulnerable to coercion or undue influence, additional safeguards have been included.

Private companies that study human behavior are not subject to the same obligations with which academic institutions that conduct research involving people are obligated.

The FTC's March 2012 Report, Protecting Consumer Privacy in an Era of Rapid Change,[28] makes reasonable recommendations that can apply to all commercial entities that collect or use consumer data that can be reasonably linked to a specific consumer, computer or other device. Under such a framework, data would not be reasonably linkable to a particular consumer or device to the extent three protections for that data are implemented: First, reasonable measures must be taken to ensure that the data is de-identified, which means that a company must achieve a reasonable level of justified confidence that the data cannot reasonable be used to infer information about, or otherwise linked to, a particular consumer, computer or other device. Second, a public commitment must be made to maintain and use the data in a de-identified fashion and not to attempt to re-identify the data. Third, if

---

[27] 45 CFR part 46.

[28] *Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Businesses and Policymakers FTC Report* (2012) available at http://www.ftc.gov/reports/protecting-consumer-privacy-era-rapid-change-recommendations-businesses-policymakers.

de-identified data is made available to third parties they should be contractually prohibited from attempting to re-identify the data.[29]

Neither the HIPAA Privacy Rule nor the FTC's March 2012 Report require no risk of re-identification. It should be recognized that the risk of re-identification should be low but not nonexistent. As long as the potential of re-identification has been adequately reduced, and subsequent measures to prevent re-identification are taken, then de-identification should remove the data from applicability of FIPPS, including collection limitation, purpose specification and use limitation.

Likewise, since it frequently will be impracticable to re-notify individuals if the purpose for the collection or use of the data has changed and to provide individual choice, waiver or alteration of the collection limitation, purpose specification and use limitation by an IRB should be permitted. Even if a use can be articulated at the time the data is collected, the research community needs to address the question of under what conditions data collected specifically for a particular research purpose may be reused for another research purpose without compromising the privacy of the individuals from whom the data was collected. Under the current frameworks, personal data could only be used with consent or if the data was de-identified. An IRB should be able to review the purpose of the secondary use and allow it to be used if it is for research or education purposes and disallow it if it is solely for commercial purposes.

Conclusion

CUSP appreciates the opportunity to submit these Comments and hopes that these Comments contribute to the review of the issues discussed above.

Respectfully submitted,

Michael J. Holland
Chief of Staff

[29] *Id.* p21-22.

Appendix
The Proposed OECD Principles

"Principle Applicable to the Collection of Personal Data

1. Collection Principle
   a. Personal data should not be collected:
      i. In violation of restrictions imposed by law;
      ii. Through deception; or
      iii. In ways that are not apparent to or reasonably discernible by and not reasonably anticipated by the individual.
   b. In addition to the requirements of paragraph (1)(a), a governmental entity should not collect data:
      i. Outside the scope of its legal authority; or
      ii. Without a legitimate purpose.

Principles Applicable to the Use of Personal Data

1. Use Principle
   a. The permissibility of uses of personal data should be determined by balancing:
      i. The degree and likelihood of benefits resulting from such uses;
      ii. The degree and likelihood of harm posed by such uses; and
      iii. The measures in place to guard against such harm.
   b. Uses of personal data reasonably likely to result in:
      i. No or minimal harm to an individual should be permitted with the basic protections required by these Principles;
      ii. Significant harm such as physical injury or loss of life should be prohibited; and
      iii. Other harms should be permitted with protections in place appropriate to the risk and degree of harm.
   c. Individual choice should be required as a protection only if meaningful, and if required should be:
      i. Clear;
      ii. Used to provide real choice; and
      iii. Accompanied by relevant, understandable information about the choice and its consequences.
   d. Given the importance of privacy and the flow of personal data, each nation should undertake through a transparent process to determine clearly how harms and benefits are to be evaluated; through uses of personal data which are to be permitted, prohibited, or permitted only with appropriate protections in place; and in what settings or condition individual consent is an appropriate protection. Nations are encouraged to cooperate in making these determinations and to coordinate their legislative measures.
2. Data Quality Principle – Personal data used for a decision affecting individuals should be relevant to the purposes for which they are used and, to the extent necessary for those purposes, should be accurate, complete, and up-to-date.
3. Individual Participation Principle
   a. A data user that uses personal data in any manner affecting the education, employment, physical or mental health, financial position, or legally protected rights of an individual should provide notice to the individual that personal data is being used, and should make readily available to the individual, without charge, a clear and understandable description of:
      i. The types of personal data used;
      ii. The sources of personal data used;
      iii. How those personal data were or will be used; and
      iv. The individual's legal rights under this Principle.

b. An individual should have the right with regard to personal data used in any manner affecting the education, employment, physical or mental health, financial position, or legally protected rights of that individual to:
    i. Obtain access to such personal data relating to the individual within a reasonable time; at a charge, if any, that is not excessive; in a reasonable manner; and in a form that is readily intelligible to the individual;
    ii. Challenge the processing and accuracy of personal data relating to the individual and, if the challenge is successful, to have the data erased, rectified, completed, or amended; and
    iii. Be given reasons if a request made under subparagraphs (i) and (ii) is denied, and to be able to challenge such denial.
c. A data steward should, upon request, correct inaccurate personal data or provide legitimate reasons for its failure to do so.

Principles Applicable to the Collection, Use, or Other Processing of Personal Data

4. Openness Principle – There should be a policy of openness about practices and policies with respect to the processing of personal data.
    a. Means should be readily available for establishing in general terms the existence of personal data processing, the nature of personal data being processed, how such data is protected, and, if applicable, the major purposes for which it is used.
    b. The identity, principal location, and contact information (including email address) of data stewards should be readily accessible.
5. Security Safeguards Principle – Personal data should be protected by reasonable security safeguards against external and internal risks including unauthorized loss, access, destruction, use, modification, or disclosure.
6. Accountability Principle
    a. Anyone who collects, uses, or otherwise processes personal data should be a responsible steward of the data and, to that end, should:
        i. Be accountable for complying with measures that give effect to these Principles;
        ii. Provide appropriate redress to individuals consistent with these Principles;
        iii. Be liable for reasonably foreseeable harm caused by the data steward's failure to give full effect to these Principles;
        iv. Upon reasonable request of a regulator, be able to demonstrate that the data steward has developed and implemented appropriate risk assessments, policies, processes, and procedures designed to follow data processing rules consistent with these Principles.
    b. No one should be held accountable under these Principles for any act or omission concerning data that is not personal data.
7. Enforcement Principle
    a. Nations should have in place adequate regulatory arrangements, competent bodies, and appropriate financial and human resources to ensure that laws enacted pursuant to these Principles are enforced.
    b. Enforcement of data protection laws should achieve effective compliance with these Principles and applicable law, while minimizing the burden on individuals and on lawful information flows."[30]

[30] *Proposed OECD Principles, pp* 14-21.

PNNL-101824

PNNL Response to OSTP Big Data RFI

March 2014

WA Pike
MT Greaves
K Kleese-Van Dam
A Endert
DA Thurman

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062;
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@adonis.osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
email: orders@ntis.gov <http://www.ntis.gov/about/form.aspx>
Online ordering: http://www.ntis.gov

This document was printed on recycled paper.
(8/2010)

PNNL-101824

PNNL Response to OSTP Big Data RFI

WA Pike
MT Greaves
K Kleese-Van Dam
A Endert
DA Thurman

March 2014

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99352

Contents

1.0 Introduction

We are in an age of measurement. The increased availability of data about all aspects of our lives and environment opens up new opportunities for discovery—yet mere *access* to data does not mean that this data can be put to profitable use. Indeed, recent revelations such as the finding that the Google Flu Trends algorithm routinely overpredicted the rate of doctor visits for influenza-like illness compared to traditional biosurveillance remind us that data collection activities must be complemented by equal, if not greater, attention to developing sound models, understanding underlying phenomenology, and creating and applying appropriate analytic methods (Lazer et al. 2014). Techniques that work well with small or static datasets may not necessarily lead to sound conclusions when applied to extremely large, heterogeneous, and dynamic data streams.

While the emergence of massive, linked data raises concerns for privacy by enabling inferences to be drawn about individuals even from purportedly anonymized data, it also opens opportunities for increased protection of privacy and identity, such as new means to verify that a user who presents themselves for a transaction is indeed the individual they purport to be. New architectures can also be established that enable distributed analytics and machine reasoning systems to answer questions about big data without exposing detailed personal records to human eyes.

In this document we discuss six trends and challenges related to the use of big data in science, security, and energy over the next five to ten years. (In this document we often use the term "data-intensive computing" rather than "big data" to indicate that data-driven discovery requires more than the agglomeration of observations; it motivates new approaches to computing that differ significantly from the algorithms and architectures used for traditional analysis, modeling, and simulation.) In particular, we recognize that while the research and development community thinks most heavily about data *volume* as the leading challenge and opportunity in big data (and, secondarily, data *variety* from the profusion of sensors and media platforms), properly dealing with data *velocity* and *veracity* are underexplored issues. These six challenges, and the associated questions in the RFI to which they respond, are:

1. The changing meaning of identity in the cyber and physical worlds; the application and protection of the emerging *superidentity* that represents the association among biometric, cybermetric, biographical, and psychological attributes; and the opportunity superidentity presents for privacy-preserving analytics (responds to Questions 1 and 3).
2. Infrastructure for data-intensive analytics that complements the movement toward open data access with equal attention to the analysis techniques and systems required to make sense of that data (responds to Question 3).
3. Distributed data stores and *analysis at the edge* as a means of performing data-intensive analysis without the need for centralized repositories (responds to Question 3).
4. The increasing need for streaming, real-time analytics that help make sense of evolving phenomena, which we term *analysis in motion* (responds to Question 3).
5. Data stewardship, sharing, and reproducibility as keys to making productive use of open access data repositories and ensuring long-term data veracity (responds to Questions 3 and 5).
6. The increasing public ability to work with big data resources, a trend we term the *democratization of big data analytics* (responds to Question 1).

2.0 Understanding and Protecting Identity in the Cyber and Physical Worlds[1]

As more of our lives are measured—from the traces we leave in our online activities, to the biometric observations captured via increasingly pervasive imaging systems—it becomes possible to draw associations between aspects of identity to produce a fuller picture of an individual's identity and to predict characteristics that are not expressed in direct measurements. Society is at a pivotal stage of development, in which explosive growth in access to cyberspace has helped fuel an "always-connected" society that blurs identities between the cyber world and the natural world. Our exposure to cyber risks will continue to grow as mobile devices become richer sensors of their users and their environments—the activation rate of Android devices is estimated to exceed 1.5 million devices per day (up from 700,000 pe day just a year ago), nearly five times the estimated growth-rate of the human species. The exponential growth in this mobile sensor network means that elements of an individual's cyber-identity may be spread widely across cyberspace—distributed through online social networks, blogs, web fora or other content (e.g., photo sharing websites), online retailers and recommendation systems, and so on. There are also channels where a user may be unaware that they are sharing information (e.g., through metadata associated with content such as a geo-tagged image) or their footprint can be aggregated in order to gain author identification and other intelligence. This unconscious sharing is difficult to mask and may be one of the keys to protecting—or exploiting—the full identity exposure.

Other elements of identity exist or are derivable in the natural world, such as biometric measurements (e.g., fingerprints, vein patterns, ear patterns). These measures may enable connections across the cyber/natural world divide through their association with cyber identities (such as connections to real names that may also appear in email addresses or usernames, or through photos of faces in surveillance video that can be used as query input in web image searches, revealing text content that describes those people). Once any transition across the cyber/physical divide has been made, the amount of additional insight about an identity that can be inferred is significant.

A multinational team, of which the Pacific Northwest National Laboratory (PNNL) is a part, is constructing a new model for contemporary identity, called *Superidentity*. Superidentity is the fusion of physical and digital identities that connects cyber, biological, biographical, and psychological attributes—it is the core identity of an individual to which all of their personas and physical traits belong (Hodges et al. 2013). The more accurate and efficient we can become in recognizing Superidentity, the more confident we can be in correctly identifying individuals, making identities ultimately more secure by incorporating behavioral, biometric, digital, and personality attributes. The sum of these attributes is more difficult to spoof than the single attributes such as social security numbers and credit card numbers that are all that is needed for identity theft today. Superidentity addresses the problems currently experienced with identification of individuals that use measures in isolation—such as fingerprints, DNA, or retinal scans—by combining this information with more novel metrics, such as hand geometry, vein patterns, and online behavioral traces, highlighting the systemic linkages between them.

[1] This section is based on material derived from a collaboration between PNNL and six UK universities (Oxford, Southampton, Bath, Leicester, Kent, and Dundee). More information is available at http://www.southampton.ac.uk/superidentity/.

New notions of identity have the potential to transform law enforcement, intelligence analysis, and commerce, but can also help individuals protect their identities by understanding how inferences can be drawn about them based on the traces they leave in the world. A sound model for identity also allows users of cyberspace to measure the exposure risk they face by revealing certain facets of their identity. For instance, a publicly accessible version of the superidentity model can allow consumers to ask, if I share this information about myself online, what else could someone infer about me?

Just as biometrics describe physical identity traits, "cybermetrics" are the traits that uniquely identify individuals online. Compared to the physical world, however, individuals can have many "identities" online—although they all point back to the same underlying individual, people represent themselves differently in different online settings. The increase in the variety of digital devices and platforms with which we interact can foil identity deception and theft by providing unusually nuanced data through interfaces such as touchscreen mobile phones, greatly enhancing the potential for accurate identification. Superidentity research has shown that phone gestures such as length and weight of a screen swipe can be used to accurately predict personality traits, to determine linked biometrics such as gait, or infer biographical data such as height or gender. Using four simple feature extractions—gesture length, completion time, touch pressure, and gesture thickness—our team has been able to accurately distinguish mobile users by their gender, age range, and handedness. Cognitive psychological cues can also be derived from cybermetrics and are an element of Superidentity. For instance, we have shown that it is possible to predict password complexity from personality traits and thus identify online accounts more susceptible to brute force theft like dictionary attacks.

A mathematical model for Superidentity amounts to a graph connecting *elements* of identity through *transforms* that indicate the associations that can be made between one element of identity and another, along with the associated confidence of those transforms. The extensible model currently contains over one hundred elements of identity, where each element represents a single atomic characteristic of identity, from biographical information (such as name, date of birth, home address), natural-world elements (such as location traces, biometric measures, CCTV imagery), and cyber identities (such as usernames on a given site, email addresses) or indicators associated with access technologies (such as IMEI, IMSI, IP, or MAC addresses). Available biographic, biometric, cybermetric, and psychological observations can then be inserted into the model so that one element can predict, converge with, or refute another. Transformations from one identity element to another allow a small set of seed observations about an identity to grow into a larger set of likely assertions about that identity, through a reachability matrix that leads to a holistic picture about the identity being modeled.

Through the construction of an integrated Superidentity model, it is possible to significantly enhance our capability to make identity attributions and to associate confidence levels with those attributions in a given context; deception or inconsistencies will appear as conflicts within the Superidentity. Our hypothesis is that this makes it substantially more difficult for anyone trying to trick an identification system as one would find it almost impossible to accurately mimic or mask the many physical and virtual cues that make up a Superidentity.

Superidentity makes possible new methods for privacy-preserving analytics. An identity model that reveals pathways between identity elements allows us to measure how well methods for protecting privacy through anonymization actually work; given the inferences that Superidentity allows us to draw, even if certain identifiers are removed from an identity record, can the system still infer the unique underlying identity? (As an example, the Electronic Frontier Foundation's Panopticlick demonstrated

that most individual web users' browser configurations are unique, allowing all of a user's supposedly anonymous browsing to be traced back to them if even one of the sites the user visits contains valid identity information).

Superidentity also enables organizations to answer a critical question in an era of big data: just how much data is enough? For identification and attribution decisions, there reaches a point at which additional data does not help increase the confidence of a determination. The Superidentity model can test how *little* data is needed to accurately make identifications, allowing organizations to actually pare back their data holdings. Likewise, Superidentity can indicate which data elements are critical to identification determinations, and can be used to guide additional data collection activities to fill those gaps.

Finally, Superidentities allow new models of authentication in the cyber world. For instance, instead of usernames and passwords (which are readily compromised and, once compromised, serve as valid authentication tokens until detected or changed), *continuous authentication* can constantly compare a user's activity to elements of a Superidentity, determining the likelihood at any point in time that the user interacting with a system is indeed the owner of the purported identity. In an increasingly mobile world, and one where users have a multitude of devices and sensors to manage, the current model of static usernames and passwords for each of these devices (or the absence of security entirely) is unlikely to provide real protection. A continuous authentication model built on a foundational mathematical representation of identity (and one that, like Superidenity, contains behavioral and even psychological cues, which can be captured continuously and are difficult to spoof) can allow devices, processes, and data to be reliably and securely associated with users without requiring their constant logging in; continuous authentication can make security and privacy more usable, rather than features that many users disable because they interfere with productivity.

3.0 Infrastructure for Data-intensive Analytics

The last decade has seen a growing democratization of data production, access, and analysis worldwide. The number of data producing actors has exploded globally, matched by an increasing interest across science, security, business, and the general public in accessing the information contained in this digital data. This open data movement is the necessary first step in democratizing the analysis process, but it is not sufficient. While we have the raw material (data) and the customer (user), we are lacking the infrastructure to allow users of very different abilities and levels of computer and data literacy to access, extract, and derive sound findings from the data with ease and confidence. We see the start of a nascent trend to complement open data with open analytics—web-accessible tools and techniques that can be composed to help people make sense of data, and to do so in statistically valid ways. Today, we usually rely on custom, one-off solutions to serve data and information to narrowly defined user communities. In this, we fail to engage many other potential consumers of our data, as our services might be too difficult or too simple for their purposes. Furthermore, we may have reached the end of our ability to significantly increase human capacity to provide one-off solutions, as numerous recent calls for more data scientists have demonstrated. In short, we need to rethink how we provide infrastructure support that goes beyond pure storage, network, and computing hardware.

3.1 Composable and Customizable Infrastructure Services

Instead of expert interfaces and one-off solutions, we need to provide composable and customizable services through which service providers and users can easily generate and customize their data analysis environments. We need to recognize that different classes of user will have very different expectations and needs; while some are comfortable with basic services such as optimized data transfer, others only want to combine results from different sources without the need to think about the necessary plumbing. Services for the latter example need to be assembled from the building-block services used in the former example. Furthermore, different classes of users will need different instantiations of the same service, such as to incorporate domain-specific information and tools; therefore, it is important that the services are not only composable, but can be easily customized by end users. The principle behind this approach is not to limit choice through standardization on a set number of analytic services, but to facilitate real choice through standardization of *interfaces* that enable composability and high degrees of customization. In using a smaller set of services that interact with each other, it will also be easier to introduce privacy control measures as a core part of their application programming interfaces (APIs), as described later in this document. This overall approach represents a divergence from current thinking; programs like the U.S.-led Earth System Grid Federation (ESGF), the European Union's European Infrastructure Project (EUDAT), and many others instead focus on the provision of one standard set of services representing lowest common denominator needs.

PNNL has previously developed such composable data analysis environments for scientific user communities, such as Velo (Gorton, 2012; Gorton, 2013; Lansing, 2011; Kleese van Dam, 2014). Velo provides a reusable, domain independent data management and analysis infrastructure for modeling and analysis. Velo integrates and extends open source collaborative and data management technologies to create a scalable core platform that can be tailored for specific science domains. The Velo user interface layer is highly customizable and allows the creation of easy to use, tailored interfaces for different users. Furthermore, its rich client interface allows for the seamless integration of third party tools, providing the user with an integrated simulation and analysis environment where they can reuse familiar tools.

While composable analytic services for science and security communities need continued development, we see a need for this capability to be extended to the consumer level as part of a movement toward "citizen scientists" and increased civic engagement (see Section 7). Although some analyses can be performed on a user's local desktop, as data volumes and the associated complexity of our questions grow, the computational resources required to meaningfully engage with data will exceed consumer capabilities. Thus, there is a need for "open analytics" platforms—web-accessible computing environments that allow users to access data-intensive computing hardware and analysis algorithms and link them to data. So then, who provides this infrastructure—is it a commercial service or a government function? Is there a complement to data.gov called analysis.gov, where users can compose analytic workflows to make discoveries, access shared government, non-profit, or private sector computing resources, and discuss their findings? And how do we abstract the use of these systems so that lay users can ask complex questions of data without needing degrees in computer science to answer them? There is an opportunity for science, technology, engineering, and math education as part of this process, since it is possible to teach principles of sound statistical and visual analysis through guided, interactive exploration of data that is meaningful to a particular user.

3.2 Machine Reasoning as a Privacy Aid

With the rise of analytic algorithms that help humans make sense of complex data also comes an opportunity to think about privacy protection in new ways. One of the motivators behind data anonymization and related privacy enhancing techniques is that consumers often do not want other people or organizations to have access to private information, for fear of misuse or embarrassment. In the case of data like health records, this information could be used to help us (such as identifying early indicators of disease) or to make decisions that adversely affect us (such as insurance rates or eligibility). Are there circumstances in which machine reasoning is a more acceptable alternative to human access to our data? Can machine reasoning algorithms, which can rigorously enforce policy constraints around data sharing and compartmentalization, help ensure that information collected on a user in one domain does not bleed over to analyses in another? Further, how might this require us to consider *ethical boundaries* placed on such machine reasoning processes?

Alternatively, there are times when the increasing trend toward automation needs to be exposed to greater scrutiny. To what extent will different communities—whether the general public, scientific communities, law enforcement, intelligence, and so on—accept and trust findings generated through entirely automated processes? For instance, in advertisement targeting on the web, we may accept an entirely automated process that does not allow users access to the underlying algorithms that resulted in the ad placement. But in a law enforcement context, an indicator of suspicion or probable cause may need to be understandable and defensible by public safety personnel—a black box may not stand up to scrutiny or trust. One new area for research and development is in how to expose the workings of machine learning techniques in a human understandable way. How can we express *why* an algorithm reached a particular finding? How can a user easily re-train that algorithm based on his or her assessment of the validity or veracity of that finding? Active learning techniques are one approach to human-in-the-loop machine learning, but these typically focus on capturing input from users as training data for a classifier (such as many tasks on Amazon's Mechanical Turk). The next trend for active learning is much deeper collaboration between human and machine reasoning components, in which there is a productive discourse between them.

4.0 Analysis at the Edge

Due to a combination of factors in our data-intensive future—the sheer volume of data produced, the increasing desire for real-time data analysis, organizational policies that limit data sharing, and privacy protections—we must rethink how analysis will be conducted for data that will increasingly be highly distributed, both geographically and in terms of ownership. Instead of focusing on collecting and integrating data in a single location for analysis as is the dominant paradigm today, we must create the ability to compose analyses across distributed resources, push elements of analysis to where the appropriate data reside, and subsequently integrate the results of these analysis elements into a collective answer.

Analysis at the edge describes a distributed approach to analytics that allows benefit to be realized from vast data stores while protecting data owners' policy, classification, proprietary, or privacy interests by pushing analytics to the data instead of pulling the data to the analysis. Such an undertaking requires new methods for data discovery and assessment, computational frameworks and algorithms for

distributed analysis, knowledge-based methods for data alignment, and new methods for asynchronous interaction with distributed analytics. More importantly, it requires a fundamentally different way of thinking about how analysis is conducted, a new paradigm for how we talk about the resources that must be available to support analysis, and the policy handshakes required to enable analytic codes to traverse networks and touch shared data.

The emerging analytics at the edge capability will derive from advances in seven areas of technology:

- Resource discovery, assessment, and alignment
  - How can users identify appropriate data and analysis methods/algorithms for a task at hand, composing and executing distributed analysis algorithms? Discoveries increasingly occur at domain interfaces, while data collection efforts remain largely within a single domain; semantic descriptions of data are needed to support greater cross-domain interoperability and application of data. Irrespective of the geographic or organizational distribution of data, the distribution across domains can also only be addressed by finding ways to effectively link the data and service description vocabularies used in those domains.
- Distributed adaptive analysis
  - Analysis description languages are needed to articulate what portions of an algorithm execute against which data in which repository; users will be able to compose descriptions of a required analysis using abstractions from their specific application domain, from which executable analysis codes would be generated that are cognizant of the distributed data, analysis services, policy restrictions, and computational resources.
- Asynchronous visualization methods
  - Data and information visualization provide powerful mechanisms to guide users' understanding of analysis results; while current methods typically require a full result in memory from which to produce visualizations, the ability to operate on incomplete and asynchronously arriving results from different sources will require interactive visualizations of the analysis process and partial results. Visualization can also serve as a useful abstraction to communicate patterns in data without exposing individual, private records to inspection.
- Knowledge frameworks
  - Users will not need to understand the complexities of this distributed environment; questions can be answered without requiring knowledge of where data exist or what specific computing resources are available. Higher level of abstractions of analysis services at multiple levels will be required, as will representation of uncertainty associated with both available data and its applicability to any particular analysis problem. In addition, uncertainty can be introduced by organizations sharing data at different levels of abstraction (e.g., aggregated to various geographic units to protect identities).
- Analytical repository services
  - The fundamental building block for distributed analysis is the availability of analytic services that live on top of data wherever it resides, enabling users to execute algorithms of their choice over that data. Multiple levels of analysis will likely be required, ranging from on-call value added processing against data already housed in a repository to mechanisms for users to craft their own algorithms for execution on multiple repositories. Each of these levels has associated risks and

requirements for an execution sandbox in which the analysis is conducted independent of other activities in the repository and intermediate results are held for subsequent operations and/or retrieval.

- Distributed computing architectures
  - With access to multiple repositories offering analytic services, a user can begin to think about how to interrelate data and compose an analysis consisting of multiple stages executed across different repositories, ideally without consideration for where the data reside; provisioning a computational framework that supports this composition and coordination of analyses provides another set of scientific and technical challenges, including a dynamic search infrastructure to identify relevant data, distributed execution engines for multi-element analysis processes, and intermediate "steering" of analysis based on partial results
- Resource integrity and resiliency
  - Key to the success of this widely distributed, highly interconnected web of data and computational resources is data integrity and resource resiliency. Adequate metadata to describe data and what steps were taken to validate and verify it will be essential to enabling users' trust in the validity of shared data. Likewise, similar information on analysis services offered, benchmarks for performance and reliability, and adequate descriptions to enable effective alignments with questions of interest must be provided. Multi-level security mechanisms providing for various classes of users to have differing levels of access to data and analysis services will also be important, as is the ability to verify code provided by outside users prior to execution. Finally, repositories will need effective "sandbox" environments that enable execution of outside algorithms without compromising security of the data or other computations. Finally, mechanisms that enable users to verify that their algorithms are being executed correctly will be essential to ensuring user trust in the overall environment.

Analysis at the edge can enable privacy-enhanced analytics because it enables questions to be answered of data where the answer does not include personally identifiable information (PII), but the intermediate processing might. It is no longer necessary to import large datasets that contain significant "bycatch" of PII to answer questions that do not actually require that PII. For instance, suppose a public health study requires disease rates by ethnicity per zip code. The "answer" this question needs does not contain PII, only aggregate rates over a geographic area. However, the data required to calculate this rate *does* contain PII. A medical records bureau may have diagnosis codes and unique identifiers for patients (SSNs, or tuples of name, address, date of birth, etc.) The U.S. Census Bureau has name, address, and ethnicity data (but does not currently release data below the census block level for privacy reasons). Currently, we would have to pull each dataset together and join them based on a personal identifier, which brings privacy protection concerns. Anonymizing the records before performing the join damages the accuracy of the join. In an analytics at the edge environment, however, a researcher could set loose an analytic algorithm that attaches to the medical record and census datasets, performs a temporary join, calculates an average disease rate, and returns to the user only this rate information—the intermediate analytic steps containing PII are never cached, and the PII itself never leaves the data owner's control.

5.0 Analysis in Motion

Although we experience the world as a continuous stream of events, our analytic capabilities have typically required us to discretize our study of the world into a "batch-mode" process. That is, data collection may not always be continuous, and even when it is, prevailing analysis tools are designed for forensic analysis of static datasets. To accelerate scientific discovery, enable timely threat discovery, and make critical decisions, it is necessary to identify and interpret phenomena as they emerge and to adapt our analysis and data collection methods as those phenomena evolve. However, the dominant analysis paradigm in many fields remains post hoc evaluation of results, often relying heavily on manual effort for critical knowledge construction activities like hypothesis generation and testing. This strongly human-centered approach does not always offer the scalability required for timely decision making in big data environments that are increasingly marked by continuous data streams, and can result in delaying or preventing necessary decisions and actions. Over the next five years, how must our analytic methods evolve to match the streaming nature of the real world?

Emerging capabilities for Analysis in Motion (AIM) will lead to a new analysis paradigm for continuous, automated synthesis of new knowledge and dynamic control of measurement systems contemporaneously with observed phenomena. By focusing on new semi-automated methods for hypothesis construction from streaming data, and by rebalancing effort between humans and machines, AIM will result in an improved ability to rapidly put new observations in the context of evolving domain knowledge. To achieve this goal, research advances must be made in three areas, the combination of which accelerates a complete analysis cycle that connects data collection, interpretation, and action:

- **streaming data characterization** methods that can identify and tag features of importance in high rate, large volume data streams
- **hypothesis generation and testing** methods based on continuously evolving models that can relate features in data to candidate explanations to assist humans and machines in interpreting results and identifying possible future states
- **human-machine feedback** based on capturing human background knowledge through new interaction paradigms, to evaluate candidate hypotheses rapidly and steer models and data collectors in response to evolving knowledge.

Dealing efficiently with streaming data to identify events, threats, or anomalies is a major challenge in many real-world applications. In particular, a common approach that includes building a static model via machine learning and deploying it for classification of streaming data often quickly results in model failure because of changes in the underlying dynamics of the system that have not been captured or foreseen in generating the model. This has led to a field of incremental machine learning, which is a class of algorithms that were developed to reduce the time, memory, and storage required to train models from fixed massive training sets. These algorithms, however, still remain largely untested for streaming data and are focused primarily on classification and little on anomaly detection on-the-fly. In a streaming world, we are faced with always incomplete data and need algorithms that can adapt to a changing stream, identify hidden variables to help steer future data collection, and quantify the change in an explanatory model over time. At the same time, new machine inference techniques will enable us to continuously shed light on machine-generated hypotheses as new knowledge arrives from streams of data. Deductive reasoning, for instance, is commonly used to enable machines to draw conclusions from data, but has not

been extended to deal with always-on streams. Imagine a future when the data available on data.gov are not static snapshots, but live streams to which users (or more likely, algorithms) can subscribe for real-time awareness and decision making. Given finite storage, we cannot necessarily remember every observation in a stream indefinitely (particularly from high volume sources like imagery). New methods to automatically determine which assertions in a data stream are worth remembering for later use are needed. Techniques for "intentional forgetfulness" where only the relevant observations are preserved will emerge over the next five years as a solution to data volume challenges.

Next, human knowledge needs to be seamlessly incorporated into these streaming reasoning environments. Current approaches to incorporating users' domain expertise into visual and mathematical systems rely on interfaces that require humans to explicitly steer the underlying computation—for instance through sliders, menus, and knobs—that correspond directly to parameters of the specific algorithm underlying a visualization. However, as data rates and analytic complexity increase, users can no longer be expected to manually adjust model parameters. Rather, a new interaction method is needed in which user knowledge is captured continuously and implicitly as users work, with little or no additional burden on them. This knowledge is then fed back into evolving models and hypothesis generation algorithms (this human feedback is also a mechanism for inserting ethical judgment into the automated analysis process). Emerging research in new human-machine feedback methods will advance a user interaction paradigm called semantic interaction wherein interaction with an information space is used as evidence for the underlying cognitive discourse users have with information. In the coming years, there will be a greater trend toward capturing the tacit knowledge associated with user interaction in a mathematically structured form, leading to a more fluid cognitive process of discovery and producing faster and more meaningful insights from big data. If successful, this leads to additional challenges, such as how to evaluate these user-steered models that now reason on behalf of users.

6.0 Data Stewardship, Sharing, and Reproducibility

There is an ever increasing number of data producers in the big data environment, particularly driven by new sensor technology that has led to a proliferation of sensors in our environment, as well as more complex sensors that can generate vast data volumes and rates. Much of the data is currently in private, inaccessible collections (by design or through inability to provide access) and thus privacy is protected by obscurity (small target, difficult to find and reach). However, there is a strong and legitimate drive to make much of the data more easily accessible through open access, offering new capabilities to analyze and correlate data to further scientific progress, commercial innovation, and overall prosperity through better access to information. When considering privacy protection in this evolving open access environment it is necessary to address five considerations:

1. reference models for data stewardship
2. lifetime data guardianship
3. data sharing across institutional and national boundaries
4. data quality
5. reproducibility.

6.1 Reference Models for Data Stewardship

The Open Archival Information System (OAIS) ISO standard offers the only current reference model that describes the functionalities and processes a data facility needs to offer to provide adequate data stewardship. However, the model was originally developed in the late 1990s and only considers standalone archives. It does not consider the role of data stewardship in a distributed environment or the impact of adding analytical services or engagement outside designated user communities. With the democratization of data access and analysis, new reference models are required that elaborate on methods and processes for privacy protection in the context of interactions between data providers (e.g., guidelines for data exchange, data attribution in transfer) and as part of data analysis workflows (offered by the facility itself or more importantly by third parties). Furthermore, the current reference model and indeed most real life implementations assume that data owners will curate the data forever, and there are no provisions to address what will happen to the data should the service fold (are we suddenly left with orphaned data without a guardian protecting its appropriate use?).

6.2 Lifetime Data Guardianship

One of the key challenges of data guardianship in today's distributed big data environments is the loss of control over the data once it leaves the original data provider's site. Today, owners and users of data have no visibility or control over where the data moves or how it is used. In many cases, this may seem unnecessary, in particular if data has no privacy or proprietary restrictions associated with it. However, consider even here the case of a scientist who produces valid results, but along the way the open access data that went into those results is corrupted (deliberately or by accident); once new results are published based on the erroneous data, the reputation of the original scientist could be irreparably tarnished, even though he provided the correct data and had no hand in any later changes to it. Digital Object Identifiers (DOI) offer the ability to publish datasets with a permanent link to the original data. However, the costs associated with issuing DOIs (infrastructure, cost to join the DOI system, and long term maintenance) prevent potential users from utilizing them at a granular level (such as for a single data file or record) or more regularly (such as for different versions of a data object). Furthermore, once a DOI has been issued, the referenced data object needs to be preserved in perpetuity; this not only adds to the cost, but not all data has long-term value. Provenance is another form of capturing the "pedigree" of a particular data product. However, today's provenance systems capture provenance separately from the actual data object and usually in centralized, project-specific services. Most systems capture data provenance only as long as it is in the sphere of the project, and rarely is there import or export of provenance from other systems. Most research is currently centered on provenance vocabularies and less on its exploitation and usage for specific purposes. PNNL has been investigating how provenance can be harvested from different sources (Stephan 2013), how it could be embedded into the data object itself and be evaluated at time of use to highlight unexpected activities (by prior users or during usage).

6.3 Data Sharing across Institutional and National Boundaries

An unsolved problem is the protection of data privacy and confidentiality in a distributed data sharing environment. Traditionally, access control is enforced at the source, but does not address downstream data sharing (i.e., when you use or share my data with others, will you enforce my privacy requirements?). Today we only have one-off agreements to govern the exchange of data across national boundaries, often highly complicated by the different laws in the participating countries. Once the

agreement is signed, there are no automated control mechanisms to monitor and reinforce the potential diverging data policies. Similar hurdles exist between institutions within the U.S. in highly regulated fields. Rather than leading to an increased number of privacy violations, this has resulted in a very conservative approach to data sharing in those domains. As a consequence, critical research findings can be significantly delayed. Next to clearer international regulations in support of data sharing, we need technical solutions that can monitor and enforce privacy protection during data sharing and analysis. There is currently little research in this field, but one promising approach is the development of intelligent data products that include their own privacy rules, access control, provenance, and reporting services. PNNL has developed prototype "active products" that are live documents that contain both the findings of an analysis activity as well links to the underlying data and analysis steps that underlie those findings, a mechanism for alerting users to updates in the data or findings, and policy rules on the use of both. These data products require a lightweight privacy rule ontology standard that can express what can and cannot be done with a specific data object. Furthermore, a microcode would need to be developed that sends status information back to the source on how that data is further shared downstream. For instance, today credit reporting agencies track data access requests and can report to a consumer each time a creditor accesses that consumer's data. Scaling out this approach, imagine a model by which any user can subscribe to be alerted to any "touch point" of data about them or under their control. (Consumers are already familiar with such approaches through the sandboxes that many mobile operating systems provide; users are asked to "opt in" to enabling a new app to access information in another app, such as their contacts. However, new technology is needed to scale out this model across an entire multi-provider data ecosystem). Data-protecting microcode, perhaps resident inside each data object, also needs to be able to protect the data object from obvious misuse such as unregulated transfer outside the country. These developments expand the notion of lifetime data guardianship with a level of active control; datasets can become interactive executables themselves instead of passive files.

6.4 Data Quality

As more processes and people are creating and consuming data, increased attention must be paid to data quality. While there are many approaches to ensuring good data quality at the time of collection, it is important to recognize that these will not always be applied. As a result, it is increasingly necessary to be able to communicate quality and "fitness for purpose" to consumers of data, who may not understand the circumstances of that data's collection. Data quality includes attributes like measurement errors, error reports from a sensor, data provenance reports, and uncertainty measures. There are currently no standards for communicating the analytic suitability of data—given a question, can a particular dataset provide a valid answer? For instance, spatial datasets may contain random jittering as a privacy protection measure, but the displacement is not communicated in the data such that linking this data to others can produce erroneous results.

6.5 Reproducibility

As we are moving beyond raw data and information sharing, more and more people carry out their own complex data analysis and correlation to support scientific research, business, and policy decisions. However, while in the past such analysis often relied solely on local resources and was under the control of the analyst, today's distributed data and analysis environments include much more heterogeneous resources. This is of particular concern when a decision or result is likely to be scrutinized at a later date and reproducibility of the process is required to verify the outcome in question. Traditionally, provenance

has excelled at capturing process details from workflows, and many contemporary analysis systems include standard provenance capture functionalities. However, while provenance captures the steps taken in an analysis, it usually does not include references to the actual data sources or tools used, thus making it difficult if not impossible to recreate the analysis process even a short time after the actual analysis took place. Needed are new provenance approaches that combine a traditional provenance vocabulary with techniques that enable the embedding of direct, executable references to specific tools, data, and resources used (Stephan 2013); in an "internet of things" future, such rich provenance can allow an analytic finding to be validated by "re-executing" it on the same distributed infrastructure that created it. More research is needed to develop compact representation formats for these extended provenance records to make them feasible at extreme scales. Furthermore, it still needs to be determined how these types of records are maintained over time as data sources potentially disappear or change and analysis tools are no longer executable on current operating systems.

7.0 The Democratization of Big Data Analytics

One principal way that big data will impact the way that Americans live and work is also one of the most critical for the health of a democracy. Jefferson wrote that, "wherever the people are well informed they can be trusted with their own government." The big data revolution will certainly accelerate scientific progress and lubricate the wheels of commerce, as we have described above. But it also has the potential to reshape citizen participation in their government, increase citizen engagement and understanding of government processes, and lead to better and more supported public policy outcomes.

The overall big data revolution has been driven several technical and economic factors. First, the most well-known technical factor is the increasing scale and availability of large-scale data resources themselves. Life sciences 'omics data is the most well-known example of these, but the data avalanche can also easily be seen in areas as diverse as transportation, commerce, manufacturing, media, education, and others. Second, coupled with the growing corpus of data resources there has been a concurrent revolution in the general availability of powerful computing infrastructure, as well as a huge increase in the usability of the analytic software required to integrate and make sense of the data. Third, the economics of actually using these data resources, computational infrastructure, and analytic software has recently become dramatically more favorable. Commercial cloud services, open-source software, and a growing ethos of low-cost or free web-accessible data resources means that sophisticated big data analysis can now be performed anywhere with an Internet connection, at exceptionally low cost, and (to a growing extent) without requiring the services of a professional statistician or data scientist. The combination of all of these factors has given rise to the sudden ubiquity of big data analysis in science, open government, and the private sector.

However, the big data revolution has not yet run its course, and PNNL sees another key trend emerging: the coming *democratization of data analysis*. By this, we mean that the driving factors that we have outlined above have reached a tipping point for the ability of ordinary citizens to engage with hitherto-inaccessible data problems. The early evidence for this trend is compelling. In the past three years, we have watched the rise of *data journalism*, where reporters who are often only loosely trained in formal data analysis have been able to use available databases and tools to shed light on previously-opaque topics like campaign finance or differential policing activity. Web-based *social knowledge construction software* like wikis and specialized forums have been commercially successful and are now

commonplace, and they provide both scientists and nonscientists with tools and collaboration support for addressing problems of mutual interest. *Emerging companies* like Ontodia and Kaggle have shown that data analysis talent and desire exists broadly through the population and can be profitably leveraged. And extremely *powerful yet accessible analysis tools* like Wolfram Alpha and Amazon's AWS Big Data Analytics stack are on a clear trendline of becoming cheaper, more powerful, and easier to use. These types of evidence will accumulate because in each case they are driven by strong commercial imperatives.

PNNL believes that this increasing democratization of data analysis is a powerful trend and will substantially increase the engagement of citizens with their government, by increasing the transparency with which government decisions are made, and helping citizens to more fully understand the option space in which government operates. Citizens already have a significant interest in government at all levels, from local and municipal decisions about zoning and education to the larger-scale impact of state and federal budget and regulatory policies. Before big data, though, it was often a time-consuming and lengthy undertaking for a member of the public to gain sufficient data to be able to impact the context in which government decisions get made. These obstacles are becoming much less steep. In fact, the U.S. government has already taken several steps to encourage the democratization of data analysis. The launch of data.gov was a milestone. Data.gov now provides open access to nearly 100,000 government datasets. Coupled with the President's Executive Order last year on "Making Open and Machine Readable the new Default for Government Information," this should result in a substantial acceleration of democratic data analysis. This work of creating the data fuel for the engine of democratic data analysis should be a high priority, as it will enhance the outcomes of government.

Finally, there are several public policy implications of this trend. First, analytics over large datasets can often reveal traditionally private information, and so many of the privacy approaches described earlier in this report will need to be fully explored. Second, data and the associated visualizations can make very powerful arguments, and there remains truth in the cliché that statistical analysis can be manipulated to support almost any conclusion. Given this, it is in the government's interest not only to be the trusted, neutral source of data, but to encourage a diverse marketplace of high-quality tools, data resources, and data analysis expertise. The government can do this by continuing to prioritize efforts to make its own data easily accessible in a standards-based way, and by ensuring that the diverse datasets, analysis tools, and visualization software that have been paid for by public resources are also shared fully with the public. Third, it is in the government's interest to encourage the highest quality analysis over publically-available data. The government can respond by continuing to emphasize science, technology, engineering, and math at all levels of education, and by ensuring that every level of government has a positive responsibility to engage the public on matters of data analysis and data semantics.

8.0 References

Hodges D, S Creese, and M Goldsmith. 2013. "A Model for Identity in the Cyber and Natural Universes." *European Intelligence and Security Informatics Conference* (EISIC), August 12–14 2013.

Gorton I, Sivaramakrishnan C, Black G, White S, Purohit S, Lansing C, Madison M, Schuchardt K, Liu Y. 2012. "Velo: A Knowledge-Management Framework for Modeling and Simulation." *Computing in Science & Engineering* 14(2): 12-23.

Gorton I, Liu Y, Lansing C, Elsethagen T, and Kleese van Dam K. 2013. "Build Less Code Deliver More Science: An Experience Report on Composing Scientific Environments Using Component-Based and Commodity Software Platforms." *Proceedings of the 16th International ACM Sigsoft Symposium on Component-Based Software Engineering.* Vancouver, British Columbia, Canada: ACM. 159-168.

Kleese van Dam K, Lansing C, Elsethagen T, Hathaway J, Guillen Z, Dirks J, Skorski D, Stephan E, Gorrissen W, Gorton I, and Liu Y. 2014. "Nationwide Buildings Energy Research Enabled through an Integrated Data Intensive Scientific Workflow and Advanced Analysis Environment." *Building Simulation*: 1-9. DOI: 10.1007/s12273-014-0171-x.

Lansing C, Liu Y, Yin J, Corrigan A, Guillen Z, Kleese van Dam K, and Gorton I. 2011. "Designing the Cloud-based DOE Systems Biology Knowledgebase." *Proceedings of the 25th IEEE International Parallel & Distributed Processing Symposium*. Shanghai: 1062-1071. DOI: 10.1109/IPDPS.2011.261.

Lazer D, R Kennedy, G King, and A Vespignani. 2014. "The Parable of Google Flu: Traps in Big Data Analysis." *Science* 343(6176):1203–1205. DOI: 10.1126/science.1248506.

Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by **Battelle** *Since 1965*

U.S. DEPARTMENT OF
ENERGY

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99352
1-888-375-PNNL (7665)
www.pnnl.gov

March 31, 2014

Big Data Study
Office of Science and Technology Policy
Eisenhower Executive Building
1650 Pennsylvania Ave. NW
Washington, DC 20502

Dear Deputy Wong:

Our organizations favor the White House review of Big Data and the Future of Privacy. As the President has explained, both the government and the private sector collect vast amounts of personal information. "Big data" supports commercial growth, government programs, and opportunities for innovation. But big data also creates new problems including pervasive surveillance; the collection, use, and retention of vast amounts of personal data; profiling and discrimination; and the very real risk that over time more decision-making about individuals will be automated, opaque, and unaccountable.

That is the current reality and the likely future that the White House report must address. We therefore urge the White House to incorporate these requirements in its final report on Big Data and the Future of Privacy:

TRANSPARENCY: Entities that collect personal information should be transparent about what information they collect, how they collect it, who will have access to it, and how it is intended to be used. Furthermore, the algorithms employed in big data should be made available to the public.

OVERSIGHT: Independent mechanisms should be put in place to assure the integrity of the data and the algorithms that analyze the data. These mechanisms should help ensure the accuracy and the fairness of the decision-making.

ACCOUNTABILITY: Entities that improperly use data or algorithms for profiling or discrimination should be held accountable. Individuals should have clear recourse to remedies to address unfair decisions about them using their data. They should be able to easily access and correct inaccurate information collected about them.

ROBUST PRIVACY TECHNIQUES: Techniques that help obtain the advantages of big data while minimizing privacy risks should be encouraged. But these techniques must be robust, scalable, provable, and practical. And solutions that may be many years into the future provide no practical benefit today.

MEANINGFUL EVALUATION: Entities that use big data should evaluate its usefulness on an ongoing basis and refrain from collecting and retaining data that is not necessary for its intended purpose. We have learned that the massive metadata program created by the NSA has played virtually no role in any significant terrorism investigation. We suspect this is true also for many other "big data" programs.

CONTROL: Individuals should be able to exercise control over the data they create or is associated with them, and decide whether the data should be collected and how it should be used if collected.

We continue to favor the framework set out in the Consumer Privacy Bill of Rights and see that as an effective foundation on which to build other responses to the challenges of Big Data.

Signatories:

Advocacy for Principled Action in Government
American Association of Law Libraries
American Library Association
Association of Research Libraries
Bill of Rights Defense Committee
Center for Digital Democracy
Center for Effective Government
Center for Media Justice
Consumer Action
Consumer Federation of America
Consumer Task Force for Automotive Issues
Consumer Watchdog
Council for Responsible Genetics
Electronic Privacy Information Center (EPIC)
Foolproof Initiative
OpenTheGovernment.org
National Center for Transgender Equality
Patient Privacy Rights
PEN American Center
Privacy Journal
Privacy Rights Clearinghouse
Privacy Times
Public Citizen, Inc.

March 31, 2014

Nicole Wong
Office of Science and Technology Policy
Attn: Big Data Study
Eisenhower Executive Office Building
1650 Pennsylvania Ave. NW
Washington, DC 20502

Submitted via email to bigdata@ostp.gov

Re: Notice of Request for Information, "Big Data RFI," FR Doc. 2014-04660

Dear Ms. Wong:

Reed Elsevier and LexisNexis are pleased to respond to the White House Office of Science and Technology (OSTP) request for information ("RFI") regarding "big data." Reed Elsevier is a world leading provider of professional information solutions. We help scientists make new discoveries, doctors save lives, corporations build commercial relationships, insurance companies assess risk, and government and financial institutions detect fraud. Our LexisNexis Risk Solutions business uses cutting-edge technology, unique data and advanced analytics to help our customers detect and prevent identity theft and fraud and manage risks.

The LexisNexis Risk Solutions High Performance Computing Cluster (HPCC) Systems technology allow our customers to process large amounts of data to make better decisions and get better results. Our HPCC Systems is an open source, big data processing platform that links disparate data sources together on a large scale and at high speed, handling approximately 30 million transactions per hour.

Based on our extensive experience with big data analytics, we provide the following responses to the questions posed in the RFI.

RESPONSES TO REQUEST FOR INFORMATION REGARDING "BIG DATA"

(1) What are the public policy implications of the collection, storage, analysis, and use of big data? For example, do the current U.S. policy framework and privacy proposals for protecting consumer privacy and government use of data adequately address issues raised by big data analytics?

Today, big data and analytics provides value to the economy at large, to individual consumers, and to federal, state, and local governments. Specifically, big data supports academic and legal research, scientific breakthroughs, fraud prevention, and law enforcement applications. Through LexisNexis products and services, big data provides solutions that help private sector and government clients across all industries assess, predict, and manage risk associated with the people and companies with which they do business.

Reed Elsevier believes the U.S. policy framework and existing privacy laws sufficiently protect consumers in the context of big data analytics. The U.S. framework of sectoral laws has identified specific types of personal information, the misuse of which could potentially cause real harms to consumers. It regulates

these types of information only, leading to a legal system of maximum flexibility, where harms are addressed without business being encumbered by excessive legal regulation.

For example, under the Fair Credit Reporting Act ("FCRA"), consumers are entitled to receive notice of the most important adverse actions that may involve the use of personal information, such as denials of employment, credit, insurance, or housing. The FCRA also provides consumers with a copy of the information that the consumer reporting agency at issue maintains about such consumer. The FCRA allows consumers to seek corrections to any data that is inaccurate or incomplete. The protections of the FCRA apply to any personal data used to generate consumer reports, including data used in big data systems with advanced analytics.

The Gramm-Leach-Bliley Act ("GLBA") regulates financial institutions that have access to nonpublic personal information pertaining to consumers. It requires these consumers to receive an annual notice of privacy practices, and provides that consumer with the right to opt-out of certain information sharing practices. GLBA calibrates privacy protections over consumer financial information with providing consumers with transparency and choice. The protections of the GLBA apply to all covered nonpublic personal information, including data used in big data systems.

The Health Insurance Portability Accountability Act ("HIPAA") protects sensitive personal health information created and maintained by health providers and health plans ("protected health information" or "PHI"), severely restricting the use and disclosure of PHI for third party activities, and providing requirements for data security and third party vendors who have access to such health information. The protections of HIPPA apply to all PHI, including data used in big data systems.

The Video Privacy Protection Act ("VPPA"), which regulates disclosure of video rental information—information that is deemed by some consumers as too revealing of personal activities and thus, sensitive for that reason—is another example. As with the other laws listed, the protections of the VPPA apply to all covered personal information, including data used in big data systems.

All of these examples demonstrate that the current sectoral framework works. Particular types of data requiring protection have been identified by policymakers and appropriate legal regimes were developed to protect them. These laws apply to this data regardless of the size of the data pool, because misuse of this data has been identified to lead to harm.

If specific gaps are identified unique to big data analytics, Reed Elsevier supports the development of voluntary, enforceable industry codes of conduct to address such gaps. With the quickly evolving technology associated with big data analytics, this approach is the best way to ensure that whatever standards are eventually adopted are dynamic enough to adapt with the pace of change.

(2) What types of uses of big data could measurably improve outcomes or productivity with further government action, funding, or research? What types of uses of big data raise the most public policy concerns? Are there specific sectors or types of uses that should receive more government and/or public attention?

Many uses of big data have a measurable positive impact on outcomes and productivity. Areas such as record linkage, graph analytics, deep learning and machine learning have been demonstrated as critical to help fight crime, reduce fraud, waste and abuse in the tax and healthcare systems, combat identity theft and fraud, and many other aspects that help society as a whole. For example, LexisNexis is working with the states of Georgia, Louisiana, Indiana, and Connecticut to help combat income tax refund fraud. Georgia deployed the LexisNexis Tax Refund Investigative Solutions in January 2012, stopping more than $30 million of fraudulent returns from 2012 through 2013. Another area where big data analytics holds promise is the health care area. Application of big data technology and advanced analytics to extract knowledge from health data can lead to better patient outcomes and less costly treatments.

(3) What technological trends or key technologies will affect the collection, storage, analysis and use of big data? Are there particularly promising technologies or new practices for safeguarding privacy while enabling effective uses of big data?

Customers such as leading banks, insurance companies, utilities, law enforcement and the Federal government depend on LexisNexis technology and information solutions to help them verify identity, assess risk, predict fraud, comply with legislation and know customers better. To manage, sort, link, and analyze billions of records, LexisNexis developed a data intensive supercomputer. The supercomputer, a high performance computing cluster (HPCC) has been proven with enterprise customers, through LexisNexis products and services, who need to process large volumes of data in mission-critical 24/7 environments.

In 2011, LexisNexis decided to open source the proprietary data intensive supercomputer and launch it into the marketplace as HPCC Systems for Big Data analytics processing. HPCC Systems helps organizations gain competitive advantages by leveraging data to help scale for innovation and growth. The streamlined platform needs fewer resources to operate and eliminates expensive legacy technology.

Big Data processing capabilities are not enough to ensure entity resolution success. Data must be linked quickly and accurately. LexisNexis has a proprietary linking system that turns disparate information into meaningful insights. This technology enables customers to identify, link and organize data associated with a record so that identities and entities can be disambiguated quickly with a high degree of accuracy.

One important trend is the move to the cloud. LexisNexis will soon offer cloud services utilizing HPCC Systems, enabling businesses, researchers, data analysts and developers to easily and cost-effectively process vast amounts of data. Using these services allows customers to focus on mission-critical tasks instead of purchasing, configuring, and maintaining cluster hardware. Cloud services are an exciting technology that allows for processing vast amounts of data for web indexing, data mining, log file analysis, machine learning, financial analysis, bioinformatics research and other applications. They remove the need for hardware and software maintenance by our clients.

There are also a number of research initiatives in this area. Particularly promising technologies to safeguard privacy while enabling certain uses of big data are Differential Privacy, Private Information Retrieval and Homomorphic encryption. While one or more of these technologies may not prove to be commercially viable in the long run, Differential Privacy seems to be the most promising of them, with certain vendors already starting to support it.

(4) How should the policy frameworks or regulations for handling big data differ between the government and the private sector? Please be specific as to the type of entity and type of use (e.g., law enforcement, government services, commercial, academic research, etc.).

We believe that the existing policy framework and privacy laws should apply to big data analytics just as they do in other uses and applications of personal information. Big data is fundamentally no different than any other data applications involving personal information. To the extent that there are regulations regarding access and use of personal information imposed on either the private sector or government users, these regulations should also apply to big data applications.

As is the case with the existing policy framework, there are instances where regulations regarding access and use of personal information differ between the government and the private sector, such as government applications used for law enforcement and national security purposes. Such distinctions should also apply to big data analytics.

We note that any policies developed in this area should recognize the important role that the private sector plays in providing data and analytics to government agencies to help them in accomplishing their critical missions. Any future policy decisions in this area should not restrict the ability of private sector companies to collect and analyze information in support of government agencies.

(5) What issues are raised by the use of big data across jurisdictions, such as the adequacy of current international laws, regulations, or norms?

Reed Elsevier is a long-time supporter of continued efforts by U.S. government entities in the area of increased cooperation among global privacy enforcement authorities. We are especially interested in, and supportive of, facilitation of cross-border data transfers.

At the same time, Reed Elsevier believes that the United States' privacy framework provides the most flexible and nuanced approach to privacy protection in existence today. The explosion in innovation in e-commerce in this country and our status as the world's leader in internet technology and content are clear indications of this. While the legal regime regulating the internet can be improved and strengthened, it is important that we not do anything that will stifle innovation and adversely impact e-commerce. The OSTP should encourage other countries to follow our lead and adopt more flexible privacy frameworks that mimic the U.S. approach. Our framework provides consumer protection in areas where it is needed, while providing enforcement authority against bad actors when required.

We appreciate the opportunity to respond to this RFI and provide our thoughts on the questions raised regarding big data analytics. If you have any questions, please contact Steve Emmert at (202) 857-8254.

Sincerely,

Steven M. Manzo
Vice President
Global Government Affairs Americas

SIDLEY AUSTIN LLP
SIDLEY

SIDLEY AUSTIN LLP
1501 K STREET, N.W.
WASHINGTON, D.C. 20005
(202) 736 8000
(202) 736 8711 FAX

araul@sidley.com
(202) 736 8477

BEIJING HONG KONG SHANGHAI
BOSTON HOUSTON SINGAPORE
BRUSSELS LONDON SYDNEY
CHICAGO LOS ANGELES TOKYO
DALLAS NEW YORK WASHINGTON, D.C.
FRANKFURT PALO ALTO
GENEVA SAN FRANCISCO

FOUNDED 1866

March 31, 2014

By Email: bigdata@ostp.gov

Hon. John Podesta
Counselor to the President
The White House
Office of Science and Technology Policy
Attn: Big Data Study
Eisenhower Executive Office Building
1650 Pennsylvania Ave NW
Washington, DC 20502

Re: Big Data RFI

Dear Mr. Podesta:

The White House's attention to the important privacy issues implicated by the advent of "Big Data" is timely, appropriate, and appreciated. This field presents a novel set of issues that increasingly entails questions that will shape and define the fabric of our society. We thank you for the opportunity to comment and participate in this "scoping exercise," as you have described it, and hope this letter and the enclosed white paper will provide you with a useful perspective and context. We hope our contribution in these comments will help advance understanding about the strength and adaptability of the current U.S. system for regulating and enforcing data privacy. We believe the current U.S. model is actually very well suited to address and adapt to the challenges ahead.

Providing a substantial opportunity for the public to comment on the Big Data review is particularly important given the broad range of data and analytical tools increasingly available across broad sectors of the economy. Given the vast complexity *and* promise of Big Data, it is vital that, in developing "the foundation for a robust and forward-looking plan of action," the White House consider information, advice and input from all stakeholders and assure that any "plan" takes account of all the attendant costs and benefits. The transparency with which the review committee has approached these issues to date is laudable; it is imperative for an issue of this importance that decisions and directions not be determined through an obscure, inscrutable or idiosyncratic process.

The Administration of President Obama has of course directly recognized, and sought to harness, the benefits of Big Data "to help accelerate the pace of discovery in science and engineering, strengthen our national security, and transform teaching and learning."[1] Indeed, the Administration announced its own major "Big Data Research and Development Initiative," on March 29, 2012, "to … improv[e] our ability to extract knowledge and insights from large and complex collections of digital data, … [and] help solve some the Nation's most pressing challenges."[2]

In light of this transformative potential, we urge the Big Data review committee to ensure thata rigorous application of cost-benefit analysis is brought to bear on any recommendations which are put forth. In suggesting this approach, we understand that the committee has been charged with producing a report that frames key questions and identifies areas in which further government action may be required; and that the committee is not charged with drafting or proposing new laws or regulations. Even so, in asking these questions, the committee should consider the costs and benefits of singling out areas of concern – questions as innocuous as "should analytical tools be allowed to be used on data gathered from wearable devices?" can have dramatic impacts on innovation, investment and opportunity for providing new products and services to consumers.

In Executive Order 13,563, "Improving Regulation and Regulatory Review," President Obama embraced and enhanced the cost-benefit analysis principles set forth by President Clinton in Executive Order 12,866 ("Regulatory Planning and Review"). We urge the review committee not merely *assume* that Big Data poses threats to personal privacy; rather, the committee should rigorously identify and characterize any potential harms that could warrant policy attention. Good public policy requires that any privacy risks of Big Data must be analyzed just as objectively as the compliance costs and economic impacts that could result from unduly burdensome regulation. Cass Sunstein, the former OMB Administrator for the Office of Information and Regulatory Affairs, has acknowledged cost-benefit analysis must play a role in privacy policymaking. Indeed, Professor Sunstein has noted that "people are concerned that dignity, privacy, and other values could be trump cards that would allow the agencies to escape the analytic discipline of cost-benefit analysis. And that is a worry that needs to be answered."[3] We hope your review committee will actively address these concerns.

The United States has driven the development of the technology that powers Big Data. Even as the United States continues to lead from a technological perspective, so should we continue to provide thoughtful and measured recommendations for the future of privacy. As you

---

[1] *See* http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal.
[2] *See* http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf.
[3] *See* http://www.cfr.org/economics/regulation-behavior-paternalism/p31076.

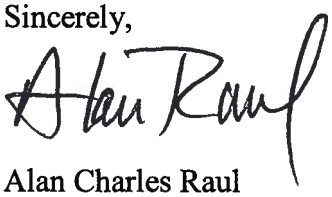
noted in your speech to the White House/MIT Workshop, "the United States can also be proud of its long history as a leader in information privacy." The United States has, and continues to lead on privacy. The European Union has taken a procedurally distinct approach to privacy, prescribing restrictive rules and creating cumbersome compliance regimes. We believe that the United States benefits from the nimble and proportional response of the agencies and enforcement "ecosystem" charged with upholding privacy. Most importantly, the issues raised by "Big Data" should not be treated as monolithic. A fixed, codified approach to regulating data could impose undue compliance costs without commensurate benefits. These issues are explored in more depth in the enclosed white paper, "The Strength of the U.S. Commercial Privacy Regime." (We believe the attached paper responds to RFI questions 1, 4 and 5.)

We further encourage the review group to ask federal and state agencies to identify and quantify how much they spend, and obligate businesses under their jurisdiction to spend, on issues relating to privacy compliance. While "dollars spent" should be by no means considered the defining metric of how well privacy is protected, understanding the allocation of such resources is a valuable data point. The review group should further quantify privacy sections imposed via federal and state penalties, including fines as well as consent orders. In evaluating the latter, it will be valuable to quantify and understand the budgets and resources available and utilized by companies based in the United States for privacy related functions. The White House Big Data review group is in a unique position to request, compile, analyze, and make public such figures – an action that will allow citizens, corporations, as well as governments and regulators around the world to understand how, and how seriously, the United States approaches issues related to the future of privacy. In addition, regulators at home will be assisted by concrete information that details the costs and benefits of various regulatory approaches. We believe the data will confirm that privacy compliance, commitments and enforcement are all highly robust in the U.S. today.

Big Data has a dramatic potential to enhance the American economy and benefit U.S. consumers and citizens. The impact of our ability to gather, analyze, interpret, understand, and use information that is increasingly generated in, on, and around us is only beginning to be understood. The issues presented to the Big Data review committee are vitally important to the future of innovation and economic growth. Assessing the costs and benefits of proposed paths forward is difficult but necessary to protect and promote innovation and opportunities to benefit society.

Thank you for the opportunity to comment and participate in this process.

Sincerely,



Alan Charles Raul

Enclosure: "The Strength of the U.S. Commercial Privacy Regime"

SIDLEY AUSTIN LLP
SIDLEY

MEMORANDUM

By Email: bigdata@ostp.gov

TO: Big Data Study Group

FROM: Alan Charles Raul
Vivek K. Mohan

RE: The Strength of the U.S. Commercial Privacy Regime

DATE: March 31, 2014

I. INTRODUCTION

The United States commercial privacy regime is arguably the oldest, most robust, well developed and effective in the world. While this statement may be debated in certain quarters, particularly in Europe, it is consistent with the facts. The purpose of this paper is to promote a thoughtful understanding of the U.S. privacy regime, and to help begin a discussion that recognizes some of the advantages of the relatively flexible and non-prescriptive nature of the American system, which counts on enforcement and deterrence more than detailed prohibitions and rules.[1] Most important, it is crucial to recognize that the U.S. model of privacy regulation today is in fact comprehensive and capable of identifying and addressing harmful new uses of commercial information.

Legal protection of privacy in civil society has been recognized in the U.S. common law since 1890 when the article "The Right to Privacy" was published in the Harvard Law Review by Professors Samuel D. Warren and Louis D. Brandeis. From that time on, the United States government, the American people, and stakeholders ranging from corporations to academics to nonprofits have worked in concert to evolve, develop, understand and protect privacy. Moreover, from its conception by Warren and Brandeis, the U.S. system for protecting privacy in the commercial realm has been focused on addressing technological innovation. The Harvard professors astutely noted that "[r]ecent inventions and business methods call attention to the next step which must be taken for the protection of the person, and for securing to the individual … the right "to be let alone."" Eighty-four years later, in enacting the Privacy Act of 1974 regulating government databases, Congress found that "the right to privacy is a personal and fundamental right protected by the Constitution of the United States."

[1] This paper should be viewed as a work in progress, rather than final product. As the White House "Big Data" initiative progresses, the authors would be pleased to work with the Administration in advancing and refining the material presented herein.

In practice, privacy and data protection are no more "fundamental rights" in the E.U. than they are in the U.S. On both sides of the Atlantic, consumer privacy is highly valued, and on both sides it is subject to the principle of "proportionality," meaning that the right is not absolute. Privacy and data protection are balanced in the E.U., as in the U.S., in accordance with other fundamental rights and interests that our respective societies need to prosper and flourish – namely, economic growth and efficiency, technological innovation, property and free speech rights and, of course, the shared values of promoting human dignity and personal autonomy.

Conceptions of the right to privacy have come a long way from the "right to be let alone" as articulated by Professors Warren and Brandeis in their seminal article. The dynamic and robust system of privacy governance in the United States marshals the combined focus and enforcement muscle of the U.S. Federal Trade Commission, State Attorneys General, Federal Communications Commission, Consumer Financial Protection Bureau (and other financial and banking regulators), the Department of Health and Human Services, Department of Education, the judicial system, and last – but certainly not least – the highly motivated and aggressive U.S. plaintiffs' bar. Taken together, this enforcement ecosystem has proven to be nimble, flexible, and effective in adapting to rapidly changing technological developments and practices, responding to evolving consumer and citizen expectations, and serving as a meaningful agent of deterrence and accountability. Indeed, the U.S. enforcement and litigation-based approach appears to be particularly well suited to deal with "recent inventions and business methods" – i.e., new technologies and modes of commerce – that pose ever changing opportunities and unpredictable privacy challenges.

In an increasingly interconnected world, privacy and data protection are of course global concerns. The United States is a favorite target of criticism from the European Union, which has taken – from a legislative and organizational perspective – a structurally different approach to privacy oversight. Yet the concepts that underpin such protection – and the substantive protections afforded – are markedly similar upon closer inspection. In fact, the values underpinning the U.S. and E.U. frameworks for privacy and data protection are largely congruent – and share common roots in principles and practices developed in the U.S. starting in the 1970s.

The absence of a comprehensive or "omnibus" commercial privacy law in the U.S. is of no particular substantive significance. The U.S. has specific privacy laws for the types of citizen and consumer data that are most sensitive and at risk: financial, insurance and medical information; information about children and students; telephone, Internet and other electronic communications and records; credit and consumer reports and background investigations, at the federal level, and a further extensive array of specific privacy laws at the state level. Moreover, the U.S. is the unquestioned world leader in mandating information security and data breach notification, without which information privacy is not possible. If one of the sector-specific federal or state laws does not cover a particular category of data or information practice, then the Federal Trade Commission Act, and each state's "Little FTC Act" analogue, comes in to play. Those general consumer protection statutes broadly and comprehensively proscribe (and authorize tough enforcement against) "unfair or deceptive" acts or practices. And, in addition, information privacy is further protected by a network of common law torts, including "invasion" of privacy, public disclosure of private facts, "false light," appropriation or infringement of the

right of publicity or personal likeness, and of course, remedies against general misappropriation or negligence. In short, there are no substantial lacunae in the regulation of commercial data privacy in the U.S.

The advent of Big Data and the insights, advancements, and changes that it promises does not change the fundamental reality that the United States has an effective and comprehensive governance framework for privacy. If an information practice is arguably wrongful and damaging, potential liability in the U.S. is not reasonably avoidable. In other words, if conduct is unfair or deceptive, unreasonably offensive or harmful, or violates any one of numerous sector-specific laws, a remedy is likely available. And there are numerous motivated enforcers to prosecute the cause – to say nothing of the discipline of the commercial marketplace and stock market.

In taking both a general ("unfair or deceptive") and sectoral (HIPAA, GLBA, FCRA, ECPA, COPPA, FERPA, etc.) approach to commercial privacy governance, the United States has empowered government agencies to oversee data privacy where the categories and uses of data are manifestly sensitive, and where the data uses are not obviously delicate, to adapt public policy flexibly in light of new "inventions and business methods" (such as drones, facial technology, Internet of Things, etc.). Comprehensive, monolithic regulation of "data" is not necessarily more effective in addressing new and previously unforeseen privacy issues, and is certainly more likely to inhibit innovation and creativity. As President Obama wrote in the Wall Street Journal on January 18, 2011, outdated, intrusive and unreasonable rules, and rules that restrict consumer choices, can "stifl[e] innovation and have … a chilling effect on growth and jobs."[2]

What defines the U.S. system of privacy and data protection in the commercial sector is responsiveness – responsiveness to what the public wants and expects. For example, the most concrete form of privacy harm – identity theft and data breaches – has generated powerful responses from the FTC, State AGs, class action lawyers, consumers voting with their feet, and the stock market. Moreover, these privacy players shape new data protection policies and practices. Likewise, the White House has already helped further shape public expectations and commercial mores ("best practices") by developing, publishing and promoting its Consumer Privacy Bill of Rights and convening stakeholder groups to elaborate upon and implement these standards. In other words, by taking up these issues as the White House has done, the Administration has sponsored significant thought leadership leading to pro-privacy impacts and practices. Thus, the U.S. system has proven effective at protecting consumer information privacy in a deliberate, measured and proportional manner that does not overreact or deny consumers the benefit of products, services and technological innovation they want.

II. BRINGING COST-BENEFIT ANALYSIS TO PRIVACY GOVERNANCE

In the U.S., regulatory policy, at its best, is designed to be reasonable, fair and balanced in achieving the intended benefits and preventing the specified harms. Regulators and elected

[2] Obama, "Toward a 21st-Century Regulatory System," Wall St. J. (Jan. 18, 2011), http://online.wsj.com/news/articles/SB10001424052748703396604576088272112103698.

officials ask what is efficient and cost-effective; what benefits overall consumer welfare; and what promotes consumer protection while preserving economic growth and innovation, and respecting property rights? In other words, competing objectives often need to be reconciled with each other.

President Obama has directed the federal government to use cost-benefit analysis as part of the regulatory process. In Executive Order 13563, "Improving Regulation and Regulatory Review," President Obama brought the cost-benefit analysis principles set forth by President Clinton in Executive Order 12,866 ("Regulatory Planning and Review") into the 21st century. The application of cost-benefit analysis to the questions posed by the intersection of Big Data and privacy is undoubtedly complex, but nonetheless essential.

Cass Sunstein, the former OMB Administrator for the Office Of Information and Regulatory Affairs, has acknowledged cost-benefit analysis must play a role in privacy policymaking. Professor Sunstein has noted that "people are concerned that dignity, privacy, and other values could be trump cards that would allow the agencies to escape the analytic discipline of cost-benefit analysis. And that is a worry that needs to be answered."[3]

Notably, the application of cost-benefit analysis has led even the Europeans to question their own privacy governance regime. The new E.U. draft Data Protection Regulation was accompanied by a detailed "impact assessment" regarding the projected costs of compliance with the new rules. But that assessment has been soundly criticized. The Office of the Information Commissioner in Britain, as well as the UK Parliament, have criticized the E.U.'s 2012 draft privacy directive as overly prescriptive and has called for the E.U. to adopt a "truly risk-based approach,"[4] and some representatives of the European Commission acknowledge that new privacy regulations must be tailored to protect innovation and allow new technologies (like services provided in the cloud) to flourish.

III. UNITED STATES PRIVACY GOVERNANCE

Every business in the United States is subject to privacy laws and regulations at the federal level, and frequently at the state level. These privacy laws and regulations are actively enforced by federal and state authorities, as well as in private litigation. Some laws and regulations focus on particular industries (healthcare, financial services, telecommunications); some on particular activities (e.g., collecting information about children online or sharing data about consumers' credit); and some on particular types of data. The Federal Trade Commission, the Executive Branch and state attorneys general also issue policy guidance on a number of general and specific privacy topics.

Like many other jurisdictions, the United States does not have a central *de jure* privacy regulator. Instead, a number of authorities – including, principally, the Federal Trade Commission and state consumer protection regulators – exercise broad authority to protect privacy. In this sense, the

[3] *See* http://www.cfr.org/economics/regulation-behavior-paternalism/p31076.
[4] *See* http://ico.org.uk/news/~/media/documents/library/Data_Protection/Research_and_reports/data_protection_reform_latest_views_from_the_ico.ashx.

U.S. has more than 50 *de facto* privacy regulators overseeing companies' information privacy practices. Compliance with the FTC's guidelines and mandates on privacy issues is not necessarily coterminous with the extent of an entity's privacy obligations under federal law – a number of other agencies, bureaus, and commissions are endowed with substantive privacy enforcement authority.

Oversight of privacy is by no means exclusively the province of the federal government – state attorneys general have increasingly established themselves in this space, often drawing from authorities and mandates similar to those of the FTC. The plaintiff's bar increasingly exerts its influence, imposing considerable privacy discipline on the conduct of corporations doing business with consumers.

At the federal level, Congress has passed robust laws protecting consumers' sensitive personal information, including health and financial information, information about children, and credit information. At the state level, nearly all 50 states have data breach notification laws on the books,[5] and many state legislatures—notably California[6]—have passed privacy laws that typically impact businesses operating throughout the United States.[7]

Privacy rights have long been recognized and protected by common law. The legal scholar William Prosser created a taxonomy of four privacy torts in his 1960 article "Privacy" and later codified the same in the *American Law Institute's Restatement (Second) of Torts.* The four actions for which an aggrieved party can bring a civil suit are intrusion upon seclusion or solitude, or into private affairs; public disclosure of embarrassing private facts; publicity which places a person in a false light in the public eye; and appropriation of one's name or likeness. These rights protect not only the potential abuse of information, but govern generally its collection and use.

An exhaustive discussion of all federal, state, and local authorities and enforcement actions area is beyond the scope of this paper.[8] What follows is a brief overview of the framework for commercial privacy governance in the United States.

A. PROTECTING CONSUMERS BY PREVENTING UNFAIRNESS AND DECEPTION: A FEDERAL AND STATE IMPERATIVE

THE FEDERAL TRADE COMMISSION

[5] *See* http://www.ncsl.org/research/telecommunications-and-information-technology/security-breach-notification-laws.aspx.
[6] *See* http://www.ncsl.org/research/telecommunications-and-information-technology/state-laws-related-to-internet-privacy.aspx.
[7] *See e.g.* http://www.ncsl.org/research/telecommunications-and-information-technology/security-breach-notification-laws.aspx and http://www.ncsl.org/research/telecommunications-and-information-technology/state-laws-related-to-internet-privacy.aspx.
[8] See, e.g., a list of federal privacy laws put together by the Center for Democracy and Technology - https://www.cdt.org/privacy/guide/protect/laws.php.

The FTC is the most influential government body that enforces privacy and data protection[9] in the United States.[10] It oversees essentially all business conduct in the country affecting interstate (or international) commerce and individual consumers.[11] Through exercise of powers arising out of Section 5 of the Federal Trade Commission Act, the FTC has taken a leading role in laying out general privacy principles for the modern economy. Section 5 charges the FTC with prohibiting "unfair or deceptive acts or practices in or affecting commerce."[12] The FTC's jurisdiction spans across borders - Congress has expressly confirmed the FTC's authority to provide redress for harm abroad caused by companies within the U.S.[13]

As FTC Commissioner Julie Brill has noted, "the FTC has become the leading privacy enforcement agency in the United States by using with remarkable ingenuity, the tools at its disposal to prosecute an impressive series of enforcement cases."[14] Using this authority, the FTC has brought numerous privacy deception and unfairness cases and enforcement actions, including over 100 spam and spyware cases and over 40 data security cases.[15]

The FTC has sought and received various forms of relief for privacy related "wrongs" or bad acts, including injunctive relief, damages, and the increasingly popular practice of "consent decrees." Such decrees require companies to unequivocally submit to the ongoing oversight of the FTC and implement controls, audits, and other privacy enhancing processes during a period of time that can span decades. These enforcement actions have been characterized as shaping a common law of privacy that guides companies' privacy practices.[16]

"Deception" and "unfairness" effectively cover the water front of possible privacy related actions in the marketplace. "Unfairness" is understood to encompass unexpected information practices, such as inadequate disclosure. The FTC has taken action against companies for "deception" when false promises, such as those relating to security procedures that are purportedly in place, have not been honored or implemented in practice. As part of this new common law of privacy (which has developed quite aggressively in the absence of judicial review), the FTC's enforcement actions include both online and offline consumer privacy practices across a variety of industries, and often target emerging technologies such as the "Internet of Things."

---

[9] This discussion refers generally to "privacy" even though, typically, the subject matter of FTC action concerns "data protection" more so than privacy. This approach follows the usual vernacular in the U.S.

[10] See Daniel J. Solove & Woodrow Hartzog, *The FTC and the New Common Law of Privacy*, 114 Columbia L. Rev. __ (forthcoming 2014) ("It is fair to say that today FTC privacy jurisprudence is the broadest and most influential force on information privacy in the United States—more so than nearly any privacy statute and any common law tort."), available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2312913.

[11] *See* http://export.gov/static/sh_en_FTCLETTERFINAL_Latest_eg_main_018455.pdf.

[12] 15 U.S.C. § 45.

[13] 15 U.S.C. § 45(a)(4).

[14] Commissioner Julie Brill, Privacy, Consumer Protection, and Competition, Loyola University Chicago School of Law (Apr. 27, 2012), available at http://www.ftc.gov/speeches/brill/120427loyolasymposium.pdf.

[15] See Commissioner Maureen K. Ohlhausen, Remarks at the Digital Advertising Alliance Summit (Jun. 5, 2013), available at http://www.ftc.gov/speeches/ohlhausen/130605daasummit.pdf.

[16] *See, e.g.*, Solove, Daniel J. and Hartzog, Woodrow, *The FTC and the New Common Law of Privacy* (August 15, 2013). 114 COLUMBIA LAW REVIEW (2014, forthcoming), *available at* http://ssrn.com/abstract=2312913 or http://dx.doi.org/10.2139/ssrn.2312913 (last accessed Mar. 2, 2014).

THE STATES

State attorneys general retain powers to prohibit unfair and/or deceptive trade practices similar to the FTC arising from powers granted by so-called UDAP ("Unfair or Deceptive Acts and Practices") statutes. Recent privacy events have seen increased cooperation and coordination in enforcement amongst state attorneys general, whereby multiple states will jointly pursue actions against companies that experience data breaches or other privacy allegations. Coordinated actions among state attorneys general often exact greater penalties from companies than would typically be obtained by a single enforcement authority. In the past year and a half, several state attorneys general have formally created units charged with the oversight of privacy, including state such as California, Connecticut, and Maryland.

B. SECTOR SPECIFIC LAWS

For the most sensitive data, the United States takes a sectoral approach to privacy governance in which agencies, such as the Department of Health and Human Services, are authorized and required to oversee privacy related issues related to their substantive mandate. This model ensures that the agencies and regulatory bodies that are most familiar with the substantive issues posed by the collection, analysis, and/or use the specific data in question are provided with the legal tools and oversight authority to leverage their expertise. Instead of being required to pass the baton to a generalist data protection authority, subject-matter expert agencies are able to swiftly, effectively, and often proactively identify and act upon potential concerns.

FEDERAL LAWS

Congress has passed laws protecting personal information in the most sensitive areas of consumer life, including health and financial information, information about children, and credit information. Various federal agencies are tasked with rulemaking, oversight, and enforcement of these legislative directives.

The scope of these laws and the agencies that are tasked with enforcing them is formidable. Laws such as Children's Online Privacy Protection Act of 1998, the Health Insurance Portability and Accountability Act of 1996, Gramm-Leach-Bliley, the Fair Credit Reporting Act, the Electronic Communications Privacy Act, the Communications Act (regarding consumer proprietary network information), and the Telephone Consumer Protection Act of 1991, to name just a few, prescribe specific statutory standards to protect the most sensitive consumer data.

STATE LAWS

In addition to the concurrent authority that state attorneys general share for enforcement of certain federal privacy laws, state legislatures have been especially active on privacy issues that states view worthy of targeted legislation.

In the areas of online privacy and data security alone, state legislatures have passed laws

covering abroad array of privacy related issues,[17] cyberstalking,[18] data disposal,[19] privacy policies, security breach notification,[20] employer access to employee social media accounts,[21] unsolicited commercial communications,[22] and electronic solicitation of children,[23] to name but a few.

California is viewed as a leading legislator in the privacy arena, and its large population and high-tech sector means that the requirements of California law receive particular attention and often have de facto application to businesses operating across the United States.[24] The combined legislative and enforcement authority of federal and state governments ensures that the policy leadership articulated at the federal level – like the White House's 2012 Privacy Report – can be implemented effectively in practice.

C. PRIVACY OUTSIDE THE EXECUTIVE BRANCH

CO-REGULATION/INDUSTRY SELF REGULATION

To address concerns about privacy practices in various industries, industry stakeholders have worked with government, academics, and privacy advocates to build a number of co-regulatory initiatives that adopt domain-specific, robust privacy protections that are enforceable by the FTC under Section 5 and by state attorneys general pursuant to their concurrent and conforming authority. These cooperatively-developed accountability programs establish expected practices for use of consumer data within their sectors, which is then subject to enforcement by both governmental and non-governmental authorities. This approach has had notable success, such as the development of the About Advertising icon by the Digital Advertising Alliance and the opt-out for cookies set forth by the Network Advertising Initiative.[25]

PRIVATE ACTORS – THE PLAINTIFF'S BAR

The plaintiff's bar is highly incentivized to vindicate commercial privacy rights – through consumer class action litigation. The wave of lawsuits that a company faces after being accused in the media of mis-using consumer data, or being victimized by a hacker or suffering a data breach incidents, is well known by large and small across the country. It is perhaps this

---

[17] *See* http://www.ncsl.org/research/telecommunications-and-information-technology/state-laws-related-to-internet-privacy.aspx.
[18] *See* http://www.ncsl.org/research/telecommunications-and-information-technology/cyberstalking-and-cyberharassment-laws.aspx.
[19] *See* http://www.ncsl.org/research/telecommunications-and-information-technology/data-disposal-laws.aspx.
[20] *See* http://www.ncsl.org/research/telecommunications-and-information-technology/security-breach-notification-laws.aspx.
[21] *See* http://www.ncsl.org/research/telecommunications-and-information-technology/employer-access-to-social-media-passwords-2013.aspx.
[22] *See* http://www.ncsl.org/research/telecommunications-and-information-technology/unsolicited-commercial-communication-laws.aspx.
[23] *See* http://www.ncsl.org/research/telecommunications-and-information-technology/electronic-solicitation-or-luring-of-children-sta.aspx.
[24] *See* https://oag.ca.gov/privacy/privacy-laws.
[25] *See* http://www.aboutads.info/; http://www.networkadvertising.org/choices/?partnerId=1//.

fear, and the incentive to invest in privacy protections and compliance that is the most tangible benefit accruing from the involvement of the plaintiff's bar.

IV. PRESCRIBING PRIVACY IN THE EUROPEAN UNION

While it might fairly be said that the United States' approach to privacy governance is complex and multi-polar, Europe's purported centralization of responsibilities for privacy and data protection oversight has yielded no less complex a system.

This EU governance framework is often praised – and distinguished from the U.S. system – on the ground that the right to privacy and data protection is a fundamental right in the Europe, but not the U.S. However, in reality, Europe balances "privacy" as a fundamental right against a litany of other such rights, employing a principle of proportionality.[26] Thus, while Europe as a comprehensive, omnibus data protection law, it does not provide more absolute or effective substantive rights than the U.S. system.

V. THE UNITED STATES: TAKING A RIGHT AND REASONABLE APPROACH

Comparisons between privacy regulation in the United States and European Union sometimes praise the "comprehensive" nature of the E.U.'s data protection regulation while critiquing the simultaneously general and sectoral, and federal and state, approach the U.S. has taken. Proponents of the European approach favor a single document capturing privacy principles that apply across domains, rather than the approach adopted under U.S. law, which entails a broad baseline and imposes tailored privacy restrictions to address sectors where data use raises the risk of specific, concrete harms. To advocate for the European model in the United States is to elide the lessons of our history.

The constitutional system of checks and balances, and federalism, and the institutions that have grown around it, are systematically designed to favor a limited, sectoral approach to privacy governance backed up by powerful general enforcement against unfair, deceptive or unreasonably offensive or damaging conduct.

Given the U.S. track record for protecting personal information privacy, why is the U.S. viewed by some as a data protection outlier – namely, a jurisdiction that does not provide "adequate" privacy protection in the judgment of E.U. regulators? This misimpression derives from a failure to appreciate the extensive, pervasive and consequential nature of the U.S. data protection regime. While Europe and countries whose data protection regimes have been deemed "adequate" by E.U. regulators (including, for example, Andorra, Faroe Islands, Guernsey, and the Isle of Man, amongst others), have established one comprehensive and omnibus framework law for protecting privacy, it is certainly not clear that this approach works better in practice (or in theory) than the U.S. model of protecting privacy and information security through a dense and intersecting array of federal and state statutes, and common law theories.

[26] *See* http://eur-lex.europa.eu/legal-content/en/ALL/;jsessionid=1vn5T4yHsgkr10nnMZ2Hn9zvngpyqYw11X705wGdnSkcnVMJqCQh!1896676610?uri=CELEX:52012DC0009; http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf.

Nonetheless, many in the privacy advocacy community praise the E.U.'s highly prescriptive regime and intensely bureaucratic approach as good, and the U.S. model of more flexible standards disciplined through vigorous enforcement, as bad. They ask why U.S. regulators can't be more like their E.U. counterparts. The reason, of course, is that the flexibility of the U.S. approach pays dividends in innovation and other social benefits – with a powerful back-stop of punishment to deter real wrong-doing.

Though the E.U. oft disparages the U.S. approach, it must be acknowledged that Europe is moving toward the U.S. in a number of ways. The latest E.U. privacy proposal contains a substantially more detailed cost-benefit analysis than any U.S. privacy policymaker has performed to date. The proposal also cuts some red tape and promotes streamlined E.U.-wide regulatory approvals. It also focuses more heavily on what has been a priority in the U.S., namely information security and data breach notification requirements.

The E.U.'s promulgation and reliance on strict and comprehensive legal codes has had mixed results. In fact, Jeffrey Rosen, a highly astute analyst of many important privacy themes, commented in the Stanford Law Review, in February 2012 (64 Stan. L. Rev. Online 88), that there has long been a dichotomy in Europe between strict laws on paper and loose enforcement in practice. Addressing the dramatic impact that the proposed new 2% annual revenue penalty (since upped to 5%!) could have on freedom of expression on the Internet when combined with the new "right to be forgotten," Professor Rosen wrote as follows:

> It's possible, of course, that although the [proposed draft] European regulation defines the right to be forgotten very broadly, it will be applied more narrowly. Europeans have a long tradition of declaring abstract privacy rights in theory that they fail to enforce in practice.

No system of data protection any where in the world that produced more legal settlements, judgments, settlements, consent decrees and, perhaps most importantly, corporate compliance programs that seek to protect and ensure privacy than the United States. Even though every member state of the European Union has a Data Protection Authority, they vary greatly in terms of aggressiveness and resources. Indeed, a recent study found that the very "unpredictability" of FTC's broad mandate proves a stronger incentive to invest in privacy than the European regulators siloed mandate.[27]

This is not to argue, of course, that privacy policy in the U.S. is fully evolved. But as discussed above, the U.S. system possesses the adaptability to respond to new inventions and business methods. Products and services in today's connected information economy often involve broad information collection practices, involving data whose utility may not be apparent when the service is rendered. There would surely be considerable social loss in dramatically curtailing digital data flows if the world went European by setting up reams of new paper privacy protections that prohibit data collection and use by default. Denying society the benefit of

[27] Bamberger, Kenneth A. and Mulligan, Deirdre K., *Privacy on the Books and on the Ground* (November 18, 2011). STANFORD LAW REVIEW, Vol. 63, January 2011; UC Berkeley Public Law Research Paper No. 1568385. *Available at* http://ssrn.com/abstract=1568385.

finding new utility in old data would defeat the very promise of the Big Data revolution. Consumers and workers on both sides of the Atlantic would benefit if the U.S. government would stand up for the American approach to privacy based on relatively flexible regulation, low bureaucracy, reasonable transparency, serious enforcement (and thus deterrence) and the promotion of legal mechanisms that help propagate a culture of compliance among corporations based or doing business in the United States.

VI. LOOKING FORWARD: THE GOALS OF PRIVACY GOVERNANCE

The 2012 White House report "Consumer Data Privacy in a Networked World," and the attendant Consumer Privacy Bill of Rights, recognized the strength of the existing United States commercial privacy framework and recommended improvements to achieve the goals of continued, strong enforcement and international interoperability. As noted in the Foreword to the report:

> The consumer data privacy framework in the United States is, in fact, strong. This framework rests on fundamental privacy values, flexible and adaptable common law protections and consumer protection statutes, Federal Trade Commission (FTC) enforcement, and policy development that involves a broad array of stakeholders. This framework has encouraged not only social and economic innovations based on the Internet but also vibrant discussions of how to protect privacy in a networked society involving civil society, industry, academia, and the government.[28]

The White House report has stirred the U.S. business and NGO communities to action, and much has changed in the intervening two years. The multistakeholder approach has provided tangible benefits and enhanced commitment to accountability that may not have otherwise been achieved so quickly. This process has reemphasized the value of the multi-polar approach taken by the United States. The White House should continue to work to further the goals of international interoperability and ensuring dynamic and proactive governance. To further these goals, we recommend that the White House engage in a benchmarking process to adequately characterize and quantify the costs and benefits of privacy governance, as proposed below.

A. INTERNATIONAL INTEROPERABILITY

Today, while the U.S. and E.U. share a common objective to protect the privacy and personal information of their citizens, perceived and procedural differences in the respective data protection regimes have resulted in counter-productive conflicts and burdensome inconsistencies. This discordance does not primarily reflect material differences in fundamental values or substantive objectives. Indeed, even Commissioner Brill of the FTC, who is seen as one of the United States' most ardent champion of consumer privacy, does not dispute that the United States provides "adequate" data protection within the meaning of the E.U. privacy framework. For example, in remarks to the Council on Foreign Relations in December 2013 Commissioner

[28] *See* http://www.whitehouse.gov/sites/default/files/privacy-final.pdf.

Brill was asked affirmed her belief that the existing privacy governance framework in the United States was "adequate" for purposes of the E.U. Directive.[29]

In general, the U.S. is actively seeking to promote ongoing regulatory alignment and develop common approaches to regulation in ways that will benefit consumers and industry across international borders. For example, on May 1, 2012, President Obama issued Executive Order 13609, "Promoting International Regulatory Cooperation."[30] ("International Regulatory Cooperation Order"). In that International Regulatory Cooperation Order, President Obama specifically directed "the promotion of good regulatory practices internationally, as well as the promotion of U.S. regulatory approaches." The Order seeks to advance "appropriate strategies for engaging in the development of regulatory approaches through international regulatory cooperation, particularly in emerging technology areas." The Order states:

> The regulatory approaches taken by foreign governments may differ from those taken by U.S. regulatory agencies to address similar issues. In some cases, the differences between the regulatory approaches of U.S. agencies and those of their foreign counterparts might not be necessary and might impair the ability of American businesses to export and compete internationally. … [and] international regulatory cooperation can identify approaches that are at least as protective as those that are or would be adopted in the absence of such cooperation. International regulatory cooperation can also reduce, eliminate, or prevent unnecessary differences in regulatory requirements.

This same approach should be applied in the digital arena; the United States should be proud of its strong commercial privacy regime – there is no reason to be defensive abroad.[31]
The role of the United States in international privacy governance must go beyond merely standing up for our system on the international stage. The United States should continue to work towards building a reality wherein international interoperability is feasible for multinational corporations. The very benefits promised by Big Data may be missed entirely should data flows be restricted for the purposes of semantic differences in privacy protections. The European Union is beginning to understand this. As Neelie Kroes, the Vice President of the European Commission and a leader of the European privacy community, noted in a recent speech (aptly titled "Data isn't a four letter word"):

> One thing is clear: the answer to greater security is not just to build walls. Many millennia ago, the Greek people realised that. They realised that you can build walls as high and as strong as you like – it won't make a difference, not without the right awareness, the right risk management, the right security, at every link in the chain. If only the Trojans had realised that too! The same is true in the digital age: keep our data locked up in Europe, engage in an impossible dream of isolation, and we lose an opportunity; without gaining any security.[32]

---

[29] *See* http://www.c-spanvideo.org/event/229156.
[30] 77 Fed. Reg. 26413 (May 4, 2012).
[31] *See* http://op.bna.com/pl.nsf/id/kjon-97urdj/$File/Digital_Trade_Coalition_TTIP_Comments_Final_5-10-13.pdf; http://federal.eregulations.us/rulemaking/document/USTR-2012-0028-0038 (comment from Alan Raul).
[32] *See* http://europa.eu/rapid/press-release_SPEECH-14-229_en.htm.

Both U.S. and E.U. policy makers should do a better job of applying rigorous regulatory impact assessments and coordination when conducting privacy and data protection regulation and enforcement. The fact that many of the harms and risks arising from violations of an individual's information privacy rights and interests are intangible does not mean they can be merely assumed, rather than carefully characterized and weighed. Likewise, just as the benefits of privacy regulation should be clearly considered, so too should the burdens of compliance. To the extent that privacy regulation imposes constraints and costs that are disproportionate to or incommensurate with the resulting benefits, it could cause significantly negative impacts on economic growth, innovation and consumer choice.

B. PRIVACY ENFORCEMENT AND ISSUE SPOTTING

The 2012 White House report rightly called for the FTC to lead the federal government by pursuing enforcement action and to identify developing issues for which codes of conduct may be appropriate. The FTC has responded to this call to action with vigorous enforcement and respected thought leadership.

The FTC noted in recent testimony to Congress that enforcement actions have focused "protecting financially distressed consumers from fraud, stopping harmful uses of technology, protecting consumer privacy and data security, prosecuting false or deceptive health claims, and safeguarding children in the marketplace."[33] The FTC's approach to emerging issues can be informal and inclusive, allowing for productive working relationships that have helped shape the development of products and services in a way that protects consumers while allowing the government to better understand the technology. The use of public meetings and workshops, such as a November 2013 event on the "Internet of Things," to help identify cutting-edge issues raised by technology, is an example of such an approach.[34] The FTC has noted that issues that are likely to capture their "privacy" attention in the years ahead include big data, mobile technologies and connected devices, and protection of sensitive data, particularly health information and information that related to children. Entities known as "data brokers" have captured the attention of the FTC and Senator Rockefeller, and are likely to be targets for future enforcement and oversight. If nothing else, the robust public debate surrounding these issues is indicative of engaged, capable policy makers.

This attentive, forward-looking oversight (and the backstop of enforcement) has influenced the business community. Professors Bamberg and Mulligan found in their study of corporate privacy management that enforcement has been successful in pushing corporate privacy managers to look beyond the letter of the law to develop state-of-the-art privacy practices which

[33] *Id.*

[34] Prepared Statement of the Federal Trade Commission on "The FTC at 100: Where Do We Go From here?" before the United States House of Representatives Committee on Energy and Commerce Subcommittee on Commerce, Manufacturing, and Trade. (Dec. 2013)

anticipate FTC enforcement actions, best practices, and other forms of FTC policy guidance.[35] The study found that corporations find the unpredictability of future enforcement by the FTC (and their state counterparts), paired with the deterrent effect of enforcement actions against peer companies, provide strong motivation to proactively develop privacy policies and practices that exceed industry standards. Companies have responded to regulation and oversight by expanding privacy leadership functions, redoubling compliance and training efforts, and engaging in proactive and ongoing dialogs with federal and state regulators.

VII. PRIVACY BENCHMARKING: A WORTHWHILE GOAL FOR THE WHITE HOUSE

Recent privacy enforcement actions have captured public attention as they ensnare some of our largest companies. Notable examples from recent years include the protracted investigations into Google's Street View collection practices (including the allegedly unauthorized capture of Wi-Fi data streams) and their alleged subsequent use of cookies in circumvention of privacy settings. In the United States, the Street View case resulted in a $7M fine paid to state attorneys general, Google further paid $17M to settle claims brought by 37 state attorneys general as well as a $22.5M fine levied by the FTC relating to the unauthorized placement of cookies. While European regulators have assailed such practices, it is not at all clear that their enforcement sanctions have been commensurate with their aggressive rhetoric.

In order to further evaluate and understand the strength of the United States commercial privacy regime, the White House may consider undertaking a systematic effort to document or review the respective commitments to enforcement, sanctions, regulatory budgets, etc., at the federal, state and E.U. levels. For example, upon (anecdotal) information and belief, it is understood that FTC's data privacy and security budget may be approximately as large as that of the European Commission and of the 28-member data protection authorities combined. As part of such a survey, it would be worthwhile to include a comparison of the privacy and security enforcement budgets of the 50 state attorneys general.

Where the E.U. DPAs and the U.S. FTC and State Attorneys General have sanctioned the same (alleged) privacy-offending conduct, such as the Google's Street View and Safari Cookie Placement cases, it would be interesting to compare the respective penalties across the Atlantic. It would likewise be of interest to rigorously compare the 50-largest privacy and data security enforcement penalties, judicial awards or legal settlements in all U.S. and E.U. jurisdictions. Such an analysis should factor in the cost and impact of consent decrees, an increasingly popular tool used by the FTC for privacy enforcement. To fully understand these costs and contextualize the cost of compliance in the U.S. and E.U., it would be helpful to compare the privacy compliance budgets of the 100 largest U.S.-based companies and their counterparts based in the E.U.

[35] Bamberger, Kenneth A. and Mulligan, Deirdre K., *Privacy on the Books and on the Ground* (November 18, 2011). STANFORD LAW REVIEW, Vol. 63, January 2011; UC Berkeley Public Law Research Paper No. 1568385. *Available at* http://ssrn.com/abstract=1568385.

The authors of this paper are prepared to assist the White House review on developing these benchmark comparisons if that would be useful. We believe that while the results may be surprising, they may go towards demonstrating that the state of privacy enforcement, regulation and compliance in the U.S. is strong.

VIII. CONCLUSION

The acute policy focus on privacy and data protection on both sides of the Atlantic demonstrates the importance of the issue, and the Administration's careful attention to it is encouraging and timely.

Recent years have provided many examples of the U.S.'s layered and complex privacy regime. The U.S.'s multifaceted privacy laws and multi-polar enforcement and regulatory authority, generally provide consumers with equivalent, if not superior, protection compared to the E.U.'s omnibus privacy law. At the same time, the E.U.'s highly prescriptive privacy laws may fail to protect consumers if, as has proven to be the case over the past two years, they are not adequately enforced or cannot evolve and adapt to meet new privacy challenges. Indeed, it is the U.S. system's inherent flexibility and broad range that has demonstrated, since at least 1890, the adaptability to address any risks associated with new inventions and business methods.

There is no doubt that the future of privacy governance can be simpler, better coordinated, more cost-effective and efficient, and less burdensome than today's. Yet calls for increased privacy governance in the age of Big Data should not be mistaken for a need to codify or ossify privacy policy. We urge the White House to appreciate – and not rashly undermine – the effective and responsive privacy system that can continue to provide U.S. consumers and citizens with a desirable balance between costs and benefits, and well serve the overall economy and society.

Software & Information Industry Association
1090 Vermont Ave NW Sixth Floor
Washington, DC 20005-4095

Submitted via email: bigdata@ostp.gov

March 31, 2014

John Podesta
Counselor to the President
Executive Office of the President
The White House
1600 Pennsylvania Ave, NW
Washington, DC 20500

RE: Big Data RFI

Dear Mr. Podesta,

On behalf of the Software & Information Industry Association (SIIA), thank you for the opportunity to comment on your review of the ways in which "big data" will affect how Americans live and work, and the implication of collecting, analyzing and using such data for privacy, the economy and public policy.

SIIA is the principal trade association for the software and digital information industry, representing more than 800 member companies. SIIA represents the industries that publish and distribute digital information, provide software applications and related Web-based services. These industries are among the fastest-growing and most important industries of the U.S. and global economies, and they are critical drivers of data-driven innovation and digital trade.

I. Introduction

Data and analytics have been around for quite some time. What is new is the increasing capacity for enterprises and governments to more effectively gather this data, and to analyze and use it effectively—from a variety of voluminous sources of structured and unstructured data, real-time and static—to innovate and improve the outcomes of everyday life. Entrepreneurs, established businesses, educational institutions and governments have increased abilities to put data to work to change the world for the better, applying these innovative abilities to everything from infrastructure, to financial services, education, healthcare, food production and consumer goods and services.

The term "big data" refers to very large sets of data that outstrip the memory capacity that computers use for processing, which led to the development of well-known processing technologies like MapReduce and its open-source equivalent, Hadoop, as well as newer technologies like the High Performance Computing Cluster (HPCC), another massive parallel-processing computing platform for

Tel: +1.202.289.7442
Fax: +1.202.289.7097
www.siia.net

large-scale data processing and analytics. [1] It also relates to the different kinds of data that are typically used in analysis, unstructured text, video and audio data that are not organized in neat, hierarchical patterns. Finally, big data includes rapidly changing data sets that are a sharp departure from the older static data bases that could be analyzed over a period of days or weeks. The "three Vs" slogan that big data consists of new data analysis techniques put to work on data sets of increased volume, variety and velocity is derived from these underlying realities.[2]

Larger data sets and more affordable analytical techniques increasingly enable greater insights and create greater value for organizations and individuals that were previously possible. One key novelty is that, in addition to finding answers to specific queries, big data analysis allows insights that could not be anticipated empirically or theoretically before the analysis took place. Data analysis is no longer simply hypothesis testing. Instead, the data "speak" and tell the data scientists something they did not know before. This contrasts with historical practices of coming to the data with preset conclusions, as enhanced analytics often enable data to reveal previously unknown insights.

In addition to the broad societal benefits, and those obvious to governments and businesses, it is individuals and consumers that stand to benefit most from tools they have never had access to before, to harness the power of their data to deliver practical benefits. Individuals use data to make better decisions about everything from what they buy to how they plan for the future. These decisions can be minor, such as customized services to an individual, or they can be major such as deciding where to go to college based on school evaluations or predictions of future career earnings.[3]

For these reasons, SIIA agrees with a recent White House Blog post that big data analytics *includes* "a revolution in the way that information about our purchases, our conversations, our social networks, our movements, and even our physical identities are collected, stored, analyzed and used.[4] But it is important to remember that the vast majority of big data is not personal or sensitive data, and the vast majority of new insights generated from big data analysis do not rely on personal information. A realistic policy review of big data should reflect this critical point.

As the Office of Science and Technology Policy (OSTP) noted in 2013, Big Data carries tremendous opportunities, ranging from empowering consumers with the full landscape of information they need to make optimal energy decisions; to enabling civil engineers to monitor and identify at-risk infrastructure; to informing more accurate predictions of natural disasters; and more.[5] SIIA strongly

---

[1] Mayer-Schonberger, Viktor; Cukier, Kenneth Big Data: A Revolution That Will Transform How We Live, Work, and Think Houghton Mifflin Harcourt, 2013, p. 6

[2] "To these, IBM's Michael Schroeck adds Veracity. In other words, a firm's imperative to screen out spam and other data that is not useful for making business decisions." Nielsen, Lars; Burlingame, Noreen, A Simple Introduction to DATA SCIENCE New Street Communications, LLC, 2012, p. 10

[3] Data Innovation 101: An Introduction to the Technologies and Policies Supporting Data-Driven Innovation, Daniel Castro & Travis Korte, November 4, 2013.

[4] Big Data and the Future of Privacy, January 23, 2014.

[5] Unleashing the Power of Big Data, Tom Kalil and Fen Zhao, April 18, 2013

concurs with the goals of the President's Big Data initiative to harness the power of data to advance national goals such as economic growth, education, health, and clean energy; use competitions and challenges; and foster regional innovation.

SIIA produced a white paper in 2013 explaining how this phenomenon, known as data-driven innovation, presents tremendous economic and social value, capable of transforming the way we work, communicate, learn and live our lives.[6] In the paper, we explained the nature of this innovation, how it empowers enterprises and governments to benefit individuals, and we highlight how it is already enabling economic growth.

Technologists, privacy advocates and policy makers can work together to foster the societal, governmental and business opportunities provided by data-driven innovation, while also meeting the challenge of protecting privacy.

SIIA's overarching recommendation for policymakers is to proceed cautiously when considering new data policies, as these are likely to steer the future of data-driven innovation and the scope of what is possible for American innovation for decades to come. Policies that seek to curb the use of data or have this as a foreseeable effect could stifle this nascent technological and economic revolution before it can truly take hold. SIIA therefore urges you to oppose broad policies that will dramatically curb data collection and analysis.

II. Data-Driven Innovation is a Driver of Economic Growth

As we highlighted in our 2013 white paper, a range of previously unimaginable applications of data-driven innovation are already being produced—or will be in the near future. These innovations are making people's lives better and safer and more prosperous, while also improving energy efficiency and saving money. In turn, data-driven innovation has already begun to spur substantial economic and job growth in the U.S. and around the world.

While data-driven innovation is clearly a powerful economic driver for the U.S. and global economies, providing enormous benefit for individuals, businesses and society, it is difficult to quantify the full economic impact because it is taking place across various different sectors of the economy. However, recent research has begun to accomplish this from various different methods.

In research around big data or "data collected and analyzed from every imaginable source," Gartner has concluded that these technologies are becoming an engine of job creation as businesses discover ways to turn data into revenue. By 2015, the firm expects data to lead to the creation of 4.4 million IT jobs globally, of which 1.9 million will be in the U.S. Further, applying an economic multiplier to those jobs, Gartner expects that each "big data" IT job added to the economy will

[6] Data-Driven Innovation, A Guide for Policymakers: Understanding and Enabling the Economic and Social Value of Data, SIIA, April 2013.

create employment for three more people outside the tech industry in the U.S., adding six million jobs to the economy.[7]

Gartner's conclusions closely track European research by the Centre for Economics and Business Research (Cebr). In an independent economic study conducted in 2012, Cebr investigated how organizations in the United Kingdom could harness the economic value of data through the adoption of data analytics. Cebr established a measure of the aggregate economic benefits that could be gained for organizations in the private and public sectors in the UK, terming the economic value of data as "data equity." In identifying six mechanisms, including customer intelligence, supply chain intelligence, performance, quality and risk management and fraud detection, Cebr estimates that data equity was worth £25.1 billion to UK private and public sector businesses in 2011. Further, Cebr notes that increasing adoption of big data analytics technologies will result in bigger gains, and we expect these to reach £40.7 billion on an annual basis by 2017.[8]

Additionally, the U.S. International Trade Commission pointed out in a report produced last year on digital trade, that not only is digital trade strong for the United States, but that further increase in international digital trade is probable, with the United States in the lead, noting that U.S. exports of "digitally enabled services" have exceeded imports in every year from 2007 rough 2011, and the U.S. surplus has widened during this period.[9]

III. Questions

(1) What are the public policy implications of the collection, storage, analysis, and use of big data? For example, do the current U.S. policy framework and privacy proposals for protecting consumer privacy and government use of data adequately address issues raised by big data analytics?

First, policies striving to account for big data need to reinterpret the application of traditional Fair Information Practice Principles (FIPPs) to allow this traditional privacy framework to evolve in response to the new possibilities arising from technological innovation.

FIPPs have provided guidelines for policymakers and data stewards regarding responsible information management practices for many years. However, there is wide recognition that changing technological capabilities and shifting expectations of privacy have challenged the application of these principles.

For instance, the practicability of obtaining true and informed consent is questionable.[10] The Federal Trade Commission has also identified this to be true, that particularly in recent years, the limitation of the notice-and-choice model have become increasingly transparent, and noting that consumers

[7] Thibodeau, Patrick. "Gartner: Big Data to Create 1.9M IT Jobs in U.S. by 2015." InfoWorld. October 22, 2012.
[8] Data Equity: Unlocking the Value of Big Data, Centre for Economics and Business Research Ltd., April, 2012.
[9] Global Trade in the U.S. and Global Economies, Part I, U.S. International Trade Commission, 2013.
[10] Data Protection Principles for the 21st Century: Revising the 1980 OECD Guidelines.

face a substantial burden in reading and understanding privacy policies and exercising the choices offered to them.[11]

Seeking affirmative consent is a barrier to socially beneficial uses of information, not because people object to the collection or use, but because the process of obtaining consent is itself too cumbersome or entirely impractical as data collection continues to increase by devices with small or no screens. Notice and consent will remain critical components in many specific or sensitive circumstances, but it cannot be the sole or even the primary mechanism for privacy protection in the age of big data.

Appropriate data use policies need to effectively balance principles of privacy against societal values such as public health, national security, economic growth, environmental protection, education and more. And this needs to be done in ways that shift the responsibility away from data subjects toward data users, and increases the emphasis on responsible data stewardship and accountability. Taking into consideration the tremendous opportunities posed by data-driven innovation, 21st Century policies need a balanced privacy framework based on risk assessment and data use, keeping in mind that socially acceptable norms of privacy are evolving along with technology.

Beyond notice and choice, opportunities presented by data-driven innovation challenge many interpretations of data minimization, where data purpose specification and use limitation are overly rigid or prescriptive. The notion of data minimization is meant to protect individuals from privacy harms by collecting only the minimum amount of data and then destroying it as soon as possible. However, while the objective is laudable and the approach very practical in certain instances, there is a tension between this method of protecting privacy and the new capabilities presented by data-driven innovation, which thrive on enormous volumes of data and the discovery of novel, unanticipated connections within them.

Data-driven innovation is about maximizing data to identify new meaning and values among a wide range of seemingly unrelated data. In this context, data minimization should not become a rigid construct. Rather it must continue to remain a key element of good data stewardship, which balances risk. For instance, there is no business need to store credit card security codes after a transaction has been processed, and saving such information creates substantial fraud risks. A reinterpreted data minimization principle would dictate that such information not be retained. In the absence of such demonstrated risk from data retention, however, data retention for further analysis should be allowed. The combination of privacy by design techniques and adherence to a set of responsible data principles can create an effective framework for data minimization that balances privacy with innovation and accounting appropriately for risk.

For these reasons, it is useful to think creatively about a new policy regime governing privacy in the "era of big data," one which increases risk assessment and appropriate data uses by entities. In doing so, it is critical to first assess the current U.S. policy framework. Existing laws have in many

[11] Protecting Consumer Privacdy in an Era of Rapid Change: Recommendations for Businesses and Policymakers, U.S. Federal Trade Commission, 2012.

ways continued to function effectively and provide a significant degree of protection, even in light of rapid technological innovation, increased data collection and analytics. Together, the combination of various sectoral privacy laws such as the Health Information Privacy Protection Act (HIPPA), Gramm-Leach-Bliley Act (GLBA), Fair Credit Reporting Act (FCRA), and Section 5 of the FTC Act, provides a framework where data privacy and security is commensurate with the sensitivity of data.

For instance, SIIA recently published a white paper highlighting how the FCRA consumer protection framework is keeping pace with technological innovation to continue protecting consumers, and it provides a good model for privacy policy in the age of data-driven innovation.[12] This approach is somewhat at variance from the standard notice and choice framework of privacy regulation. Instead of simply describing information use (that is, giving notice) and providing consumer choice, the "harm framework" seeks to identify the likely harms that the activities of these companies might cause, and then target any needed regulatory interventions to mitigate or reduce the risks of harm in a way that balances the costs and benefits involved.[13] The continued effectiveness of this model in the age of data-driven innovation was demonstrated in the FTC Spokeo case, where it was determined that the law effectively covers entities that use the most advanced technology, including online data aggregation, social media and mobile apps for a wide range of eligibility contexts as the law was designed several decades ago.[14]

Second, "Big Data" policies need to promote technology neutrality and avoid technology mandates. Technology neutrality has long been a widely recognized guiding principle for technology policies, particularly Internet-based information and communications technologies (ICT). This was first recognized within the U.S. government in 1997, with the Framework for Global Electronic Commerce, a framework that has stood the test of time in establishing broad principles for regulating ICT, that "rules should be technology neutral (i.e., the rules should neither require nor assume a particular technology) and forward looking (i.e., the rules should not hinder the use or development of technologies in the future)."[15] By contrast, Government-mandated technology standards, can freeze the development of new technologies, or disadvantage entire categories of market players.

These long-held principles for resisting technological mandates and maintaining technological neutrality are especially important for a complex IT ecosystem that drives data-driven innovation, an IT environment inherently subject to constant innovation. For example, given the range of devices that lead to the collection and utilization of data, it is impractical and ineffective to create policies based solely on a specific type of device, or an arbitrary characteristic of a device, like whether it is mobile like a smartphone or automobile sensor, or whether it is stationary, such as a computer or a refrigerator. While it might seem practical to target specific devices or platforms, this approach is

[12] How the FCRA Protects the Public, SIIA, December 2013.
[13] J. Howard Beales, III & Timothy J. Muris, Choice or Consequences: Protecting Privacy in Commercial Information 75 U. Chi. L. Rev. 109 2008 pp. 109-120.
[14] Spokeo to Pay $800,000 to Settle FTC Charges Company Allegedly Marketed Information to Employers and Recruiters in Violation of FCRA, U.S. Federal Trade Commission, June 12, 2012.
[15] The Framework for Global Electronic Commerce, 1997.

not likely to be effective today, and it will continue to become less effective over a period of months due to the rapid evolution of IT.

For example, mandating types of encryption or approaches to de-identification might seem like good approaches for enhancing privacy and data security, but such approaches continue to prove incapable of keeping up with technological evolution. The Federal Trade Commission adopted a contractual approach to de-identification rather than a technological mandate.[16] There is almost always a better way to accomplish a given purpose waiting around the corner. Policies must continue to encourage innovation to find faster, better, and less expensive ways to protect privacy and security.

(2) What types of uses of big data could measurably improve outcomes or productivity with further government action, funding, or research? What types of uses of big data raise the most public policy concerns? Are there specific sectors or types of uses that should receive more government and/or public attention?

Big data is critical to improving educational outcomes and productivity, though education is still in the nascent stages of big data compared to other fields. Today, new technology tools and analytical techniques are allowing educational institutions and agencies to enhance their analysis of big data in more cost effective and sophisticated ways to inform teaching and decision making, enhance operations, and ultimately to increase student performance and the productivity of our schools. Among the results are the ability for school systems to both better identify students at risk of failure, and to better identify which interventions would best meet the unique needs of each student. In many cases, education officials are working with school service providers for data management and learning analytics.

It is important for stakeholders to understand educational data governance models. In many cases, personally identifiable information is either not collected, or if it is collected, it is de-identified. Typical program evaluations require student-level data but do not need to know precisely who each student in the program is. In some cases, a software application is licensed and enables schools to build their own data systems. In other cases, educational service providers provide a platform and tools, whereby only the educational entity is able to access and control the personal student information.

To the greatest extent possible, bit data analysis to improve educational programs should protect student privacy by using anonymous or pseudonymous data. However, this is not always possible or desirable. In cases where personal student information is necessary, adequate additional steps must be taken to safeguard student privacy and data security, including steps to protect student information from unauthorized access and to prevent it from being used for non-authorized purposes.

---

[16] Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Businesses and Policymakers, U.S. Federal Trade Commission, 2012.

Policies must continue to balance the need of protecting the privacy of students, while enabling data driven innovation to greatly enhance the teaching and learning for the betterment of the students, institutional supports and society as a whole. For example, it is critical to understand that current federal laws such as the Family Educational Rights and Privacy Act (FERPA) provide such protections and allow the use of personally identifiable student information only for a narrow set of educational purposes, and provide protections in those cases. It is also critical to understand that many access, deletion, breach notification and other policies cannot practically put the decision or requirement on the service provider, as they are simply stewards of the data and not owners of the data.

Following are several ways by which the federal government could further leverage the power of big data to improve education:

- Targeting resources to new big data research models in the areas of evaluation and development, including through public private partnerships.
- Support further education and training needed to increase the number and expertise of professionals to bridge big data and education, as there are now insufficient number of educators and educational researchers skilled in big data.
- Provide guidance to the field to leverage big data while appropriately safeguarding student privacy and data security.

(3) What technological trends or key technologies will affect the collection, storage, analysis and use of big data? Are there particularly promising technologies or new practices for safeguarding privacy while enabling effective uses of big data?

For DDI to reach its full potential, it must be built on a foundation of good data stewardship and trust. Without the appropriate precautions, the collection and usage of some data can pose risks. Therefore, enterprises and governments must not underestimate this risk, and they must think strategically about providing privacy and security protections commensurate to the sensitivity of data, to ensure that it is used, and not abused.

Privacy by Design (PbD) is widely recognized internationally as an effective practice for developing privacy compliant information systems. PbD does not mean privacy by default, but rather it advances the view that the strength of a privacy measure should be commensurate with the sensitivity of the data it protects. It also supports the premise that voluntary privacy protections can be more effective than regulatory frameworks, as entities focus on privacy assurance from the beginning. Customers can hold entities accountable to this, as customer trust is critical for success.

Some of the most important outcomes of data-driven innovation do not rely on personally identifiable information. Indeed, a lot of high-value analytics can be done by simply looking at faceless customer data. In this approach, analysts only know each customer by an arbitrary, non-traceable number. Therefore, even if personal information is collected, it can often be immediately de-identified in a way that does not affect its value or utility for accomplishing important public and social objectives. This allows for robust privacy protection, since the data can be effectively purged of all reference to a specific individual for innovative and societally beneficial purposes.

Therefore, SIIA urges policymakers to encourage de-identification as a way to balance the needs of data-driven innovation and privacy protection, but avoid broad mandates to this end. Further, an additional caution is that if information that is not individually identifiable comes under full remit of privacy laws based on a possibility of it being linked to an individual at some point in time through some conceivable method—no matter how unlikely—this could not only prohibit many beneficial uses and benefits of data-driven innovation, but it could also destroy the incentive to de-identify the data. The Federal Trade Commission avoided this mistake in its contractual approach to de-identification.[17]

(4) How should the policy frameworks or regulations for handling big data differ between the government and the private sector? Please be specific as to the type of entity and type of use (e.g., law enforcement, government services, commercial, academic research).

The recent revelations about the U.S. government data collection and surveillance programs have sparked an examination of the policies and practices of the U.S. intelligence community, and we support this review and discussion to expand the dialogue with governments around the world. We appreciate the opportunities provided for SIIA to engage in this discussion. To that end, SIIA submitted detailed comments to the White House Review Group on Intelligence and Communications Technology and the Privacy and Civil Liberties Oversight Board (PCLOB) in 2013, recommending steps that the U.S. government can take to regain the public trust by incorporating better privacy and civil liberty practices without compromising the government's responsibility to also protect the nation. Additionally, in January SIIA produced a set of Global Principles for Governments Collecting Private Sector Data from Commercial Entities in conjunction with the Information Technology Industry Council (ITI). These principles should serve as a foundation for international discussions about reform government surveillance policies.

The U.S. technology sector is driving transformative innovations that are accelerating the global transition from the industrial to the information age. Concerns about government surveillance and access to personal information jeopardize essential elements of the innovation ecosystem, thus harming the economic position of the United States and our innovative companies.

That said, it is of great concern to us for the review of policies pertaining to government access to data to be conflated with policy debates pertaining to commercial privacy issues. These have long been separate frameworks for considering very different issues, and the notion of joining this debate provides greater risk to splintered global policy frameworks than opportunities to improve privacy.

With respect to government's ability to harness data-driven innovation, SIIA urges governments to adopt policies that leverage cutting-edge technology and data to make governance more efficient and effective, and to reduce government waste. To that end, we support the goals of the President's Big Data Initiative.

---

[17] Protecting Consumer Privacdy in an Era of Rapid Change: Recommendations for Businesses and Policymakers, U.S. Federal Trade Commission, 2012.

Today, governments at all levels are under increasing pressure to reduce the overall cost of operations, while improving productivity and providing better citizen services. Government's acceptance and utilization of new technologies is needed to enhance government's mission. Technologies that leverage data analytics to provide innovative functions and services hold the key for governments to provide improved services and to better understand how well they are fulfilling their missions. On a day-to-day basis, government agencies collect, create, store and manage large volumes of data. Whether the data is from one-on-one interactions with citizens, transactions online, visits to web pages, or interactions on social media, government agencies are creating enormous volumes of structured and unstructured data daily, which makes extracting knowledge a challenge.

Leveraging data analytics effectively can help governments understand program trends and match data across government, helping weed out waste, fraud, and abuse. Innovative data technologies help governments identify problem areas before an improper or erroneous payment occurs, and to track the information after its award for assistance with recovery. Data analytics can also help government analysts identify patterns, highlight trouble spots, and extract useful data from an ongoing data flow. Reducing fraud in government programs, performing real-time analysis of traffic patterns and increasing citizen health and safety are all examples of the way DDI can transform governments at all levels. The net result is more a more efficient and effective government that does more with less.

Therefore, governments should adopt policies that embrace a culture of analytics, where the focus is on knowing, rather than guessing. Specifically, policies should increase the use of data analytics—pulling data from myriad sources—to make strategic decisions, to encourage research and development around data science, and encourage teaching and training for data scientists and professionals, arming them with strong data analytics skills that are already in high demand in both the public and private sectors.

Second, governments should continue to embrace open data policies and public-private partnerships that maximize access to critical public data. The U.S. Federal Government, state and local governments, and governments around the world possess treasure troves of valuable data that have gone largely untapped for many years. More than ever before, citizens want access to government data, and they want it applied in innovative ways to which they are increasingly becoming accustomed.

In response, governments are making more data available with the hope that users will utilize raw data sets to perform analysis, experiments and enhance learning, with the hope that they will in turn develop applications relevant to the mission of government. The US federal government has made nearly 500,000 data sets available to developers on the web through the Data.gov initiative started in 2009, with an overarching goal of a more open and accountable government. To date, nearly 250 citizen developed apps have been created as a result of this initiative – everything from an app that allows you to see where the various superfund sites are to one that measures obesity by

county. Under Data.gov, the government itself has created over 1,200 new apps, including popular apps that track US financials and FDA recalls.

Similarly, state governments are enormous data generation engines. As recently described by the U.S. National Association of State CIOs, (NASCIO), U.S. states and governments are generating data at higher volumes. NASCIO has determined that "the sky is the limit in terms of future data generation based on the growth in mobile applications, sensors, cloud services and the growing public-private partnerships that must be monitored for performance and service levels."[18] However, many governments are still struggling to enact policies that enable a streamlined approach to providing open data, and to enable innovative applications to draw from this data. Even worse, governments are continuing to implement policies that restrict access to public data, often in contrast to public records laws that encourage openness and transparency.

In April 2013, in what was a very disappointing ruling for data-driven innovation, the U.S. Supreme Court unanimously upheld Virginia's citizens-only restriction on public records access.[19] In the case, McBurney v. Young, the Court found that the effect of the "citizen restriction" on commerce was merely "incidental" since the state created the market for the information through a monopoly, it could discriminate against noncitizens who want to access that information.

To the contrary, an effective way to maximize the full potential of data-driven innovation is for governments to embrace open data policies, to use public-private partnerships to provide access to critical public data, and to adopt enterprise architectures that enable sharing. These steps will put public sector data to innovative uses that can reap the economic and societal benefits of data-driven innovation.

(5) What issues are raised by the use of big data across jurisdictions, such as the adequacy of current international laws, regulations, or norms?

As highlighted above, data-driven innovation and digital trade pose tremendous opportunities not only for U.S. businesses, but also for U.S. economic growth and job creation. To that end, uninhibited cross-border, or cross-jurisdictional, data flows is probably the single greatest need for innovative U.S. companies to continue growing around the world.

This is not an issue that is new to the phenomenon of big data. Policies pertaining to data flows have been critically important for many years, particularly given rise by the evolution of IT over the last decade from centralized computing to network dependent systems with distributed assets and distributed management responsibilities. The cross-jurisdictional policy challenges we are facing today largely mirror those that took shape years ago within discussions about "cloud computing."

---

[18] Sweden, Eric. "Is Big Data a Big Deal for State Governments?" NASCIO, Aug. 2012.
[19] LeDuc, David. "Supreme Court Supports State's Right to Restrict Access to Data." Digital Discourse. SIIA, May 7,2013.

Over several years, these global policy discussions have morphed from rubric of "cloud computing," to that of "big data." However, regardless of the label, the fundamental tenets that we espoused in a 2011 white paper on cloud computing very much applies to this review of international policymaking for "big data."[20]

In summary, SIIA supports the following key policy to enable data-driven innovation and digital trade to continuing flourishing in the U.S. and around the world:

1. Promote policies that allow to the greatest extent possible, unrestricted transfer of data across borders,
2. Avoid localization mandates that seek to balkanize the Internet, or any policies that would give preference to data processors using only local facilities or operating locally,
3. Seek interoperable privacy regimes in which countries recognize each other's privacy rules to the greatest extent possible.
4. Encourage rules governing data to travel with the data in order to adequately recognize varying jurisdictional requirements, and ensure data subjects do not lose protection when their data is stored and processed in any remote computing environment,
5. Embrace a global approach to cybersecurity that recognizes the global nature of interconnected systems and provides for data to be protected regardless of where it is located, and that seeks international consensus standards that avoid fragmented, unpredictable national requirements.

In this context, SIIA urges the Administration to continue vigorously engaging with foreign officials to ensure that differing policy regimes are interoperable, enabling cross-border data flows to continue facilitating economic growth and opportunity. The U.S.-EU Safe Harbor Framework, a mechanism established nearly two decades ago, is a critical tool serving to recognize and account for different types of privacy compliance regimes, where the US system is risk-based and the European system is regulatory.

Additionally, SIIA supports the trade negotiating objectives contained in the current version of the proposed Trade Priorities Act on digital trade. It is critical that the trade agreements currently under consideration – the Transpacific Partnership, the Transatlantic Trade and Investment Partnership, the Trade in Services Agreement and the Information Technology Agreement – state that cross-border data flows is the norm, and that exceptions may not be used as means to restrict trade.

Further with respect to international laws and regulations pertaining to data, the recent revelations about U.S. government surveillance continue to threaten the competitiveness of U.S. IT products and services around the world. This is a complex policy issue with broad implications for SIIA members. Not only are customers around the world inherently more concerned about privacy from U.S. government surveillance, policymakers around the world are seeking to restrict the cross-border flow of data and to enact technology localization requirements.

[20] Guide to Cloud Computing for Policymakers, SIIA, July 2011.

As you know, many of these discussions conflate private sector data collection with government surveillance. As we stated above and is particularly true in international policy discussions, conflation of these two very separate issues, governed by separate policy frameworks provides greater risk of splintered global policy frameworks than opportunities to improve privacy.

IV. Conclusion

Again, thank you for the opportunity to comment on this very important topic. We look forward to working with you and other Obama Administration officials as you conduct the initial review ordered by the President on January 17, 2014, and as you continue to explore and develop policies and recommendations pertaining to big data and data-driven innovation. If you have further questions or would like to discuss, please do not hesitate to contact David LeDuc, SIIA's Senior Director for Public policy, at dleduc@siia.net or (202) 789-4443.

Sincerely yours,

Ken Wasch
President

TechAmerica Input to the White House Office of Science and Technology Policy Big Data Request for Information

March 31, 2014

TechAmerica, the leading U.S. technology trade association, respectively submits industry input on the White House OSTP Big Data RFI. The input was submitted by several members of the TechAmerica Big Data Committee. Our Committee members responded to questions two through five from the RFI inquiry.

Question Two

What types of uses of big data could measurably improve outcomes or productivity with further government action, funding, or research? What types of uses of big data raise the most public policy concerns? Are there specific sectors or types of uses that should receive more government and/or public attention?

- Data-driven decisioning - ways to apply data analytics to impact outcome and productively are truly only limited by imagination and budget.
    - Public Welfare / Public Health / Healthcare
        - Discover the paths that lead people to health & social challenges - deeper insight can help government facilitate proactive "prevention" efforts instead of reactive and costly "cure" efforts
        - Uncover hidden negative health & social trends before they become chronic - Discovery analytics can identify trends and predictive analytics can use such trends to help see into the future
        - Increase the effectiveness of fraud, waste and abuse efforts – discovery/predictive analytics can help uncover new patterns of abuse and identify potential situations more rapidly, preventing the need to reclaim funds
        - Outcome – healthier, happier citizens – lower costs to taxpayers
    - Tax & Revenue Authorities
        - Discovery and Predictive analytics are a key capability to help drive tax compliance
        - Increase tax revenue by focusing resources on cases with the best opportunity of success
    - Public Safety / Homeland Security / Defense
        - Drive better preparation and planning – the combination of all available data coupled with discovery and predictive analytics can help ensure a team, department, agency is ready for the engagement – natural disaster, terrorist attack, defensive incident
        - More effective response – have all the data and analytic capabilities available as life or death decisions are made –

know where to send resources and when – predictive trouble areas and proactively respond
        - Mitigate future risks – Use discovery and predictive analytics to monitor for possible threats and make intelligent decisions on ways to alleviate or appropriately respond to the threat
        - Continuous learning and improvement based on constantly feeding new data to the analytics platform
    - The use cases most concerning to the public are the ones that are not fully understood or the cases conducted in secret.
        - Internet monitoring – obvious the concern (in secret)
        - Social media mining – public domain vs. private domain issues (not understood)
        - Healthcare research – is PII used? if so how? Who sees the data/results? (not understood)
- Big data analytics opens the door to sifting through large, unstructured sources of data that are outside the capabilities of traditional database systems. Predictive analytics and fraud and waste analysis are two areas that lend themselves well to the value derived from big data.
- Predictive analytics uses statistical techniques to perform data mining against current and historical information. This analysis provides a framework for forming predictions about future data events or behaviors. Performing predictive analysis against large volumes of data can be applied to a diverse range of subject areas, from climate change forecasting to homeland security examination of global travel information.
- Fraud and waste detection benefits greatly from the strengths of big data. For example, electronic health data is growing exponentially as an information asset. By examining this huge volume of data across geographies and population cohorts, suspicious patterns in behavior, like claim transactions, can be detected and examined. Additionally, one can discover and act upon inefficiencies in business process flows.
- With the growth of data volumes across all elements of the information spectrum comes public concern regarding the use and disclosure of personal information. Healthcare data, as mentioned above, is probably one of the largest concerns when it comes to privacy. Local law enforcement agencies are collecting and storing the data from police license plate scanners, building a database that can potentially track citizen's locations, with very little oversight or regulation. This is an area where lawmakers are just beginning to weigh policies to address privacy concerns.

Question Three

What technological trends or key technologies will affect the collection, storage, analysis and use of big data? Are there particularly promising technologies or new practices for safeguarding privacy while enabling effective uses of big data?

- The perfect storm of innovations such as:
  - Internet of Things
  - Cloud
  - Mobile devices
  - Social media
  - Lower cost of storage
  - Lower cost of compute resources
- Has driven significant/promising advances and opened new opportunities for:
  - Analytics platforms
    - Perpetual Discovery Analytics – iterative, agile, rapid – an on-going process whereby data is analyzed repeatedly from various perspectives to uncover patterns and connections
    - Predictive Analytics – use models and patterns to predict some future action or behavior
    - N-byte scale – ability to scale analytics to petabytes/zettabytes
    - Unified data – analyze across typically silo'd environments – offers detailed data and serves as the underlying data layer for discovery and predictive analytics
- And this group of innovations has exposed gaps in:
  - Data Security
    - Increased frequency of cyber-attacks - internal and external
    - More significant negative impact of recent data breaches
  - Data Privacy
    - Unauthorized access
    - Derived PII
    - Unintended use
- Better addressing data security:
  - Continue the vigilance of defending the perimeter
  - Leverage data and analytics in a discovery fashion, give defenders the ability to see more completely into their environment – to uncover anomalous behaviors and patterns that may be unique to their architecture/ infrastructure
  - Secure the actual data itself, not just the perimeter, systems and applications – encryption, digital signatures, digital rights management
- Better addressing privacy:
  - Integrate/unify data from multiple data sources – counter intuitive because one of the big privacy concerns is the ability to derive PII from multiple sources – however, if all data is already integrated/unified, significantly better controls can be put in place to minimize, if not

eliminate the risk of inadvertently answering a question that should not be answerable (derived information)
    - It is impractical to assume it is possible to limit the questions people may ask of data, however, with a more unified approach to data, responses can be better filtered to remove or mask data that should not be shared with the user – further, responses can be monitored and trigger escalations based on configurable business rules
- The emergence and acceptance of cloud computing is a key element in adopting big data solutions. The data volumes in big data are so large that the only way to process the data is to spread the computing out over multiple computers (nodes). A big data cluster can be composed of hundreds or even thousands of nodes. Although this can be accomplished without the cloud, it is faster and generally cheaper to use a cloud computing provider, like Amazon Web Services. Nodes can be added or subtracted as needed in a short period of time.
- Another key element affecting how big data is used and analyzed is the ability to use off-the-shelf tools to perform instant data analysis. As recently as a couple of years ago, analyzing big data was strictly a batch mode process with delayed results. Today tools, like Cloudera's Impala, help users use traditional business intelligence applications, like Business Objects or Cognos, to analyze information in a big data cluster.
- All of the security practices that apply to traditional databases apply to big data. These practices include:
    - identity management to control access to the information
    - the physical security of the hardware
    - privacy management to ensure that personal information that is stored is encrypted.

Question Four

How should the policy frameworks or regulations for handling big data differ between the government and the private sector? Please be specific as to the type of entity and type of use (e.g., law enforcement, government services, commercial, academic research, etc.).

In the government and private sectors, collected and analyzed personal information should always be stored in a protected and encrypted environment. The key differentiator between different entities will be the audience that consumes the data. Data that is strictly for internal use can be managed differently from data that is exposed

(either in part or in full) for external use. Externally exposed data requires a comprehensive policy to determine what can be shared publicly, and an enforceable framework of access methods and user identity management.

Question Five

What issues are raised by the use of big data across jurisdictions, such as the adequacy of current international laws, regulations, or norms?

A robust data governance policy is key to using big data across jurisdictions. This is a policy issue, rather than a technical one. Data governance policy manages and enforces the data rules that are shaped by the laws and regulations of the various geo-political entities involved with the data. Privacy laws vary greatly in the global domain and it is important that any organization involved in collecting and analyzing big data appoint one or more data stewards to be responsible for managing compliance with these laws in the data environment.

TECH FREEDOM

Comments of

TechFreedom[1]

In the Matter of

Government "Big Data"

Request for Information
A Notice by the Office of Science and Technology Policy

March 31, 2014

[1] TechFreedom is a a non-profit, non-partisan technology policy think tank with 501(c)(3) tax-exempt status. Questions may directed to Berin Szoka, President of TechFreedom, at bszoka@techfreedom.org.

Understanding the benefits and costs of Big Data and even beginning to weigh them against each other is likely not something that can be achieved in the limited 90-day window given to the Office of Science and Technology Policy. Mr. Podesta seemed to acknowledge this in his blog post about this inquiry:

> we expect to deliver to the President a report that anticipates future technological trends and frames the key questions that the collection, availability, and use of 'big data' raise – both for our government, and the nation as a whole. It will help identify technological changes to watch, whether those technological changes are addressed by the U.S.'s current policy framework and highlight where further government action, funding, research and consideration may be required.[2]

Above all, we urge OSTP, the Administration, and those following this inquiry to keep clearly in mind that this report is the beginning of an ongoing process, not the end, that it will frame many more questions than it can possibly answer. Even with this more limited ambition, the Report can offer invaluable guidance to policymakers struggling to understand Big Data and what, if anything, to "do" about it.

Economics Must Guide Any Study of Big Data

On the benefits of Big Data, we urge OSTP to keep in mind two cautions. First, Big Data is merely another trend in an ongoing process of disruptive innovation that has characterized the Digital Revolution. Even before the advent of the Internet, the semiconductor industry saw change accelerate at a pace that was scarcely before conceivable. We now know that this was Moore's Law at work: the doubling of computing power roughly every eighteen months. One industry after another has been disrupted by digital technologies, which allow new companies to emerge out of nowhere with new ways of doing things that can quickly render obsolete not just existing companies but existing ways of doing business – and radically change consumer expectations.[3]

---

[2] http://www.whitehouse.gov/blog/2014/01/23/big-data-and-future-privacy

[3] *See generally* Larry Downes & Paul Nunes, *Big Bang Disruption: Strategy in the Age of Devastating Innovation* (2014).

In hindsight, the benefits to consumers of this topsy-turvy process loom large in many aspects of American life. Yet the process has been painful, not just for incumbent industries and business models, but for those uncomfortable with "What Hath [Technology] Wrought"[4] in our daily lives, from transforming media to unsettling our most deeply held assumptions about privacy, security, child-rearing and a host of other emotionally wrought topics. The only safe generalization that can be made is that, however difficult these changes may seem to us in hindsight, we tend to forget how painful they were at the time, both because it was difficult for even experts to predict the benefits of new technologies and because risk aversion so is deeply rooted in human nature that few at the time would really have believed even the most accurate predictions that it was worth it. Had "the future" been put up for a vote, it probably would have been banned. The point is that any inquiry into future benefits should begin from the assumption that many of the greatest benefits remain unknowable *ex ante* and that any attempt to weigh unknown future benefits against more easily imaginable potential harms will fundamentally bias policymakers against innovation.

Second, cost-benefit analyses generally, and especially in advance of evolving technologies, tend to operate in aggregates because those are more easily measured: How large an economic boost might Big Data make to our society? What are the current costs of, say, identity theft? These predictions can be useful for providing directional indications of future trade-offs, but they should not be mistaken for anything more than that. Life operates, at all levels, not in terms of aggregate, but on the margin: aggregate benefits tell us little about the trade-offs involved in specific practices, and where regulation can be most helpful – or harmful.

These two cautions should lead this inquiry to begin from a stance of humility and a general presumption that we are limited in our ability both to predict and to shape the future in ways that will actually benefit consumers. The task of economics is not to make specific predictions so that policymakers can pull "policy levers" with a clear sense of what the resulting effect of their manipulations will be, but, as Friedrich Hayek famously put it, "to demonstrate to men how little they really know about what they imagine they can design."

Economics *can* play a vital role in this inquiry, however, if assessment of trade-offs on the margins is integrated throughout, even in topics that may seem to have little to do with economics. Nowhere is economics more sorely needed than in the debate over the efficacy

[4] "What hath God wrought," a phrase from the Book of Numbers, was the first message transmitted by telegraph in 1844.

of de-identification, which is in fact a debate over the cost-effectiveness of re-identification. It is not enough to assert that a data set *can* be re-identified. After all, "Three monkeys hitting keys at random on typewriters for an infinite amount of time will almost surely produce Hamlet."[5] The key question is: is a particular data set *likely* to be re-identified based on the potential value of the uses of the data and the costs of re-identification. In other words, how many monkeys and how long *would* it take? And on the other end, how much de-identification is adequate is also as much an economic question as it is a computer science or statistical question.[6]

An economics-informed assessment of these trade-offs should lead us to more carefully weigh the costs and benefits of large data sets and to focus regulation, and the limited enforcement resources of regulators, on areas where regulation can do more good than harm. This is true on most, if not all, of the concerns raised by Big Data, from privacy to data security. Economics can help understand the trade-offs involved in addressing "non-economic" concerns like societal and constitutional values. Even if economists do not have the final word on policy decisions, they have an invaluable role to play as advisors.

Big Data *Is* Speech: This Inquiry Must Address the First Amendment

Since the purpose of this inquiry is, in the end, to shape policymaking, it must also confront another dimension of trade-offs: regulation of the private sector's use of "Big Data" largely means regulation of speech protected by the First Amendment. The Supreme Court made clear in *Sorrell v. IMS Health, Inc.*, 131 S. Ct. 2653 (2011) that data *is* speech:

> This Court has held that the **creation and dissemination of information are speech** within the meaning of the First Amendment. See, e.g. , Bartnicki , supra , at 527 ("[I]f the acts of 'disclosing' and 'publishing' information do not constitute speech, it is hard to imagine what does fall within that category, as distinct from the category of expressive conduct" (some internal quotation marks omitted)); Rubin v. Coors Brewing Co. , 514 U. S. 476, 481 (1995) ("information on beer labels" is speech); Dun & Bradstreet, Inc. v. Greenmoss Builders, Inc. , 472 U. S. 749, 759 (1985) (plurality opinion) (credit report is

[5] David Ives, *Words, Words, Words* (1987).

[6] See generally, Jane Yakowitz, *Tragedy of the Data Commons*, 25 Harv. Jrnl. Law & Tech 1 (2011), http://jolt.law.harvard.edu/articles/pdf/v25/25HarvJLTech1.pdf

> "speech"). Facts, after all, are the beginning point for much of the speech that is most essential to advance human knowledge and to conduct human affairs. There is thus a strong argument that prescriber-identifying information is speech for First Amendment purposes.
>
> The State asks for an exception to the rule that **information is speech**...[7]

It is by no means clear how the Court's jurisprudence on First Amendment protection will evolve. The Court has *always* struggled to apply free speech principles as technology has changed, and Big Data will, in that respect, be much like the telegraph, telephone, film, television, the Internet and video games. Given that OSTP's competence is technical rather than legal, this inquiry, and the future studies it engenders, should focus on the forms of "speech" enabled by Big Data and how it might "advance human knowledge" within its overall inquiry into the benefits of Big Data. This will help policymakers to approach Big Data with greater caution than they have traditionally approached new media.

This does not necessarily mean *less* regulation but does mean *better* and more constitutionally defensible regulation. Even those who think the government should have a lower burden in regulating Big Data than it would in regulating speech more generally should find the general approach of First Amendment analysis a useful heuristic for thinking about how best to deal with Big Data: What, exactly, is the government's interest? How substantial is it? Are the means chosen appropriately or narrowly tailored to address that interest? Are they over-broad? Are there other, less restrictive means available to address the problem? Is the approach either over- or under-inclusive?[8]

These are difficult questions that will either be dealt with carefully by policymakers or, if not, by courts who send legislators back to the drafting board. This inquiry cannot, of course, address all of them, but it must begin the process of integrating an assessment of First Amendment values and doctrines, along with economics, into the study of Big Data and its policy implications.

---

[7] 131 S. Ct. at 2667.

[8] See generally Berin Szoka, The Progress & Freedom Foundation, *Privacy Trade-Offs: How Further Regulation Could Diminish Consumer Choice, Raise Prices, Quash Digital Innovation & Curtail Free Speech*, Comments to the FTC Privacy Roundtables (Dec. 7, 2009), available at http://www.scribd.com/doc/22384078/PFF-Comments-on-FTC-Privacy-Workshop-12-7-09

Government Threats to Privacy

The low-hanging fruit for this inquiry – the areas where policymakers can do the greatest good at the lowest cost in terms of lost innovation, economic benefits or meddling in the still-evolving speech platforms of the Digital Age – is clear: focus on government. Government is not the only source of harm to consumers, but it is the source of the greatest and clearest harms.

Long before Edward Snowden's revelations, TechFreedom and dozens of other non-governmental civil liberties organizations, trade associations and companies joined together in the Digital Due Process Coalition to advance four simple principles for reforming the Electronic Communications Privacy Act of 1986.[9] A clear, broad consensus now exists around the need to ensure that law enforcement agencies cannot access content without a warrant. Indeed, the Sixth Circuit has even ruled that ECPA's failure to require a warrant for content in general violates the Fourth Amendment's protection against unreasonable searches and seizures.[10] Essentially the entire court in *U.S. v. Jones* clearly indicated their discomfort with the failure of our laws to protect Fourth Amendment values as technology has changed.[11] Justice Sotomayor warned that "Awareness that the Government may be watching chills associational and expressive freedoms" and called for the Court "to reconsider the premise that an individual has no reasonable expectation of privacy in information voluntarily disclosed to third parties." Justice Alito and three other Justices explicitly called on Congress to address "concern about new intrusions on privacy" through legislation because Chief Justice Taft's warning that "regulation of wiretapping was a matter better left for Congress has been borne out."

Yet, four years later, while the courts have made great progress, including a scathing magistrate decision scolding the Department of Justice for not meaningfully complying with *Warshak*,[12] Congress has talked about the issue but has done nothing – but at least action

---

[9] http://digitaldueprocess.org/index.cfm?objectid=A77781D0-2551-11DF-8E02000C296BA163

[10] *U.S. v. Warshak*, 631 F.3d 266 (6th Cir. 2011).

[11] *U.S. v. Jones*, 565 U.S. 945 (2012).

[12] In Matter of United States of America for a Search Warrant for a Black Kyocera Corp Model C5170 Cellular Telephone with FCC ID: V65V5170 (D.D.C. March 7, 2014), available at https://ecf.dcd.uscourts.gov/cgi-bin/show_public_doc?2014mj0231-2

finally appears imminent: ECPA Reform legislation now has 193 sponsors in the House.[13] This momentum towards long overdue reform has built slowly but steadily – with no help whatsoever from this Administration.

It takes a special kind of temerity for the President to loftily promise a "Consumer Privacy Bill of Rights"[14] – while doing nothing to protect the *real* Bill of Rights, the Fourth Amendment that is the crown jewel of the civil liberties: the warrant requirement that was among the chief inspirations for the American Revolution.[15]

This Administration's Department of Justice has sought warrants for email content only when ordered to do so by the Sixth Circuit in *Warshak* and even then, did not take the requirement seriously, as the recent magistrate decision makes scathingly clear. Worse, the Administration has actively worked to sabotage ECPA reform by orchestrating opposition to ECPA reform from nominally independent agencies, which appear in fact to be working in conjunction with the Department of Justice. In particular, the fanatic insistence by the Securities and Exchange Commission, now joined by the Federal Trade Commission and other agencies, that administrative agencies should be exempt from the general requirement for a warrant to access content information, has stalled ECPA reform in the Senate.

Meanwhile, the Administration has simply ignored a WhiteHouse.gov petition signed by 110,423 Americans entitled "Reform ECPA: Tell the Government to Get a Warrant."[16] Despite promising to respond "in a timely fashion" to any petition that receives 100,000 signatures within 30 days,[17] the Administration has done nothing[18] – yet it has found plenty of time to respond to a petition by *Star Wars* fans urging the Administration to begin building a Death

---

[13] H.R. 1852: Email Privacy Act, https://www.govtrack.us/congress/bills/113/hr1852

[14] http://www.whitehouse.gov/sites/default/files/privacy-final.pdf

[15] See Testimony of Berin Szoka, TechFreedom, before the House Energy & Commerce Committee Subcommittee on Commerce, Manufacturing, and Trade, hearing on *Balancing Privacy and Innovation: Does the President's Proposal Tip the Scale?*, at 4-5, March 29, 2012, available at http://democrats.energycommerce.house.gov/sites/default/files/documents/Testimony-Szoka-CMT-Balancing-Privacy-and-Innovation-President-Proposal-2012-3-29.pdf

[16] https://petitions.whitehouse.gov/petition/reform-ecpa-tell-government-get-warrant/nq258dxk

[17] https://petitions.whitehouse.gov/how-why/terms-participation

[18] Mark Stanley, Center for Democracy & Technology, *White House Still Silent on Warrantless Email Snooping*, March 31, 2014, https://cdt.org/blogs/mark-stanley/3103white-house-still-silent-warrantless-email-snooping

Star by 2016 with the clever title "This Isn't the Petition Response You're Looking For."[19] We are *not* amused.

This stubborn opposition to sensible, bi-partisan privacy reform is outrageous and shameful, a hypocrisy outweighed only by the Administration's defense of its blanket surveillance of ordinary Americans – a problem so well known that it requires no special description here.

It's time for the Administration to stop dodging responsibility or trying to divert attention from the government-created problems by pointing its finger at the private sector, by demonizing private companies' collection and use of data while the government continues to flaunt the Fourth Amendment.

This inquiry offers the Administration a chance to redeem itself, at least in part. This report should assess the full costs, both in economic terms and in constitutional values, of easy surveillance and access to private data by law enforcement and national security agencies. The report should recommend ECPA reform as outlined by the Digital Due Process Coalition, especially a clear email requirement for access to content and location data that applies to *all* law enforcement agencies, including regulators. OSTP's report should support real and meaningful reforms to national security agencies' collection of, and access to, private communications, both their content and metadata.

Regulating the Private Sector

Getting government's own house in order does not mean ignoring legitimate concerns raised by Big Data, such as how privacy companies may use data they collect and how they secure it against breaches. This inquiry can proceed along both tracks. But rather than get bogged down in abstract debates about the ideal regulatory regime for privacy and data security, an intellectual quagmire in which Washington has been stuck since the FTC first endorsed comprehensive privacy legislation in 2000 (over the vigorous objections of two Commissioners),[20] this inquiry should at least begin with, if not focus on, the legal regime that currently exists for regulating Big Data and other new technologies. That means assessing not merely what the FTC has done about privacy and data security in the past but, more importantly, *how* it has operated.

---

[19] https://petitions.whitehouse.gov/response/isnt-petition-response-youre-looking

[20] http://www.ftc.gov/reports/privacy-online-fair-information-practices-electronic-marketplace-federal-trade-commission

FTC leadership increasingly point to what they call a "common law" of digital consumer protection, meaning the dozens of enforcement actions they have settled across a wide range of cases, from online fraud to data brokers to data security to user interface design. A case-by-case method does indeed have great virtues over *ex ante* regulation for precisely the reasons mentioned above: it is difficult to predict the future, especially the unknowable benefits of new technologies, and attempts to encode today's expectations in law often do more harm than good. As the FTC declared in its 1980 Policy Statement on Unfairness: "[Section 5 of the FTC act] was deliberately framed in general terms since Congress recognized the impossibility of drafting a complete list of unfair trade practices that would not quickly become outdated or leave loopholes for easy evasion."[21]

But even if the FTC has reached the right policy outcome in many, or even most cases, its version of the "common law" is a hollow one, devoid of the very analytical rigor by which the adversarial process of litigation weighs competing theories and advances doctrine.

The FTC regulates privacy, and will regulate Big Data, primarily through its deception and unfairness powers. Yet in over seventeen years of dealing with digital consumer protection cases, the FTC has done little to develop these rich legal concepts beyond their application in the traditional marketing contexts, which the FTC was originally created to police.

This is chiefly because companies so rarely challenge enforcement actions and when the Commission settles an enforcement action, Section 5(b) requires only that (a) the Commission has "reason to believe" a violation of law has occurred and (b) believes that *opening* the enforcement action would be in the public interest. Section 5(b) does not require *any* justification or process for *settling* a case unless the Commission seeks a monetary penalty (e.g., for violations of existing consent decrees). Thus, the settlements cited by the Commission as "guidance" do not even, by their own terms, purport to reach the merits of underlying issues. The Bureau of Economics, which has played a vital role in helping to shape what may far more accurately be called the "common law" of antitrust over the course of decades, has played little apparent role in guiding the FTC's approach to consumer protection. This has led the FTC to prioritize creative theories of harm and issues that might make compelling law review topics over clear consumer harms such as identity theft. While

[21] http://www.ftc.gov/ftc-policy-statement-on-unfairness

identity theft remains far and away the leading source of consumer complaints to the FTC,[22] the FTC has not held a workshop on the topic under this Administration.

The FTC has, commendably, begun to remedy its shortcomings in other areas, most notably by trying to build an in-house technologist capability. But it has resisted changing its overall approach for the simple, understandable reason that law enforcement agencies rarely, if ever, want to make their jobs even slightly more difficult. It is no more realistic to expect the FTC to reform its own processes without significant external pressure than it is to expect the NSA to do so. Once again, what is required is leadership from the Administration and Congress into the FTC's processes.

We believe the FTC's underlying legal standards are fundamentally sound and already provide basis for "comprehensive privacy regulation," including Big Data. But if the FTC is to be trusted with the sweeping, vague power it currently holds over nearly every company in America, it is critical that a serious inquiry begin into *how* the FTC operates. Clearly, the courts have failed to play the role both the FTC and Congress assumed they would when the FTC declared, in an effort to defuse a heated stand-off with an outraged Congress over the FTC's abuse its authority,[23] that:

> The present understanding of the unfairness standard is the result of an evolutionary process. The statute was deliberately framed in general terms since Congress recognized the impossibility of drafting a complete list of unfair trade practices that would not quickly become outdated or leave loopholes for easy evasion. The task of identifying unfair trade practices was therefore assigned to the Commission, **subject to judicial review**, in the expectation that the underlying criteria would evolve and develop over time. As the Supreme Court observed as early as 1931, the ban on unfairness "belongs to that class of phrases which do not admit of precise definition, but the meaning and application of which must be arrived at by what this court

[22] http://www.ftc.gov/news-events/press-releases/2014/02/ftc-announces-top-national-consumer-complaints-2013

[23] Howard Beales, *The FTC's Use of Unfairness Authority: Its Rise, Fall, and Resurrection*, May 30, 2003, http://www.ftc.gov/public-statements/2003/05/ftcs-use-unfairness-authority-its-rise-fall-and-resurrection

elsewhere has called 'the **gradual process of judicial inclusion and exclusion**.'"[24]

Our FTC: Technology & Reform Project, composed of leading FTC experts and veterans, has begun an inquiry into how the FTC operates and how its processes could be improved to draw on many of the benefits of a true common law. [25] Like this inquiry, we see our own project as the beginning of an ongoing dialog. But already it has become clear that a series of relatively small changes could vastly improve how the FTC weighs concerns raised by new technologies, most notably ensuring clearer analysis of the component elements of its unfairness and deception powers, and greater incorporation of economics and First Amendment values in its analysis. By carefully amending Section 5 to create procedural safeguards for how the FTC settles cases and by examining why defendants essentially *always* settle, Congress may be able to help the FTC better execute its mission of advancing consumer welfare by focusing on clear harms to consumers that are not outweighed by greater benefits and that consumers themselves cannot effectively avoid.

OSTP's inquiry offers an invaluable opportunity to refocus the endless, unconstructive "privacy debate" on the concrete "how" of privacy law: FTC process. This, more than any abstract legal theory, will ultimately shape the regulation of Big Data.

---

[24] http://www.ftc.gov/ftc-policy-statement-on-unfairness

[25] *Consumer Protection & Competition Regulation in a High-Tech World: Discussing the Future of the Federal Trade Commission: Report 1.0 FTC: Technology & Reform Project*, (Dec. 2013) http://docs.techfreedom.org/FTC_Tech_Reform_Report.pdf

TECHNOLOGY POLICY INSTITUTE
Studying the Global Information Economy

March 31, 2014

Office of Science and Technology Policy
Eisenhower Executive Office Building
1650 Pennsylvania Ave., NW
Washington, DC 20502
Attn: Big Data Study

Re: Big Data Request for Information

These comments are in response to the Office of Science and Technology's March 4, 2014 Notice of Request for Information. OSTP is requesting input on the Administration's "comprehensive review of the ways in which 'big data' will affect how Americans live and work, and the implications of collecting, analyzing and using such data for privacy, the economy, and public policy."

These comments are largely based on a 2013 Technology Policy Institute paper[1] and address Question 1 of the RFI:

(1) What are the public policy implications of the collection, storage, analysis, and use of big data? For example, do the current U.S. policy framework and privacy proposals for protecting consumer privacy and government use of data adequately address issues raised by big data analytics?

The emergence of big data and the Internet of Things, which generates a growing supply of objects from which data can be collected, has raised the question of whether big data are associated with new privacy harms and a concomitant increase in the need for government action. If so, should policy makers look to the standard solutions involving notice and choice, use specification and limits, and data minimization to solve any privacy problems brought about by big data?

In the study referenced above, we conclude that there is no evidence at present that big data used for commercial and other non-surveillance purposes have caused privacy harms. Moreover, the standard solutions associated with the U.S. policy framework and privacy proposals—Privacy by Design, the Fair Information Practice Principles (FIPPs) and the Organization of Economic Cooperation and Development (OECD) principles—represent a potentially serious barrier to much of the innovation we anticipate from the big data revolution.

---

[1] Thomas M. Lenard and Paul H. Rubin, "The Big Data Revolution: Privacy Considerations," December 2013, available at http://www.techpolicyinstitute.org/files/lenard_rubin_thebigdatarevolutionprivacyconsiderations.pdf.

The Promise of Big Data[2]

Big data's potential comes from "the identification of novel patterns in behavior or activity, and the development of predictive models, that would have been hard or impossible with smaller samples, fewer variables, or more aggregation."[3] Data are now available in real time, at larger scale, with less structure, and on different types of variables than previously.[4]

Because big data analysis involves finding previously unobserved correlations and patterns, it almost necessarily involves uses of data that were not anticipated at the time the data were collected. Examples of the serendipitous uses of data are numerous and include health studies, economic studies, marketing, and the development of new products that help consumers gain access to credit and search for lower prices. Big data are also used to protect against adverse events ranging from credit card fraud to terrorism. Many of the innovations described above use multiple sources of data, which involves transferring data to third parties.

Potential Privacy Threats

Advocates have highlighted a number of potential privacy threats from big data, but as of now there is no evidence that any of these threats has materialized. I discuss them in turn.

Big data increase the risks associated with identity fraud and data breaches.[5]

In theory, big data could increase or decrease identity fraud and data breaches. On the one hand, there are more data at risk. On the other hand, the data themselves are useful in preventing fraud. Moreover, countervailing forces provide strong incentives for data holders (e.g., credit card companies) to protect their data. So, it is useful to examine what the data on identity fraud and data breaches show.

In fact, the proliferation of big data in recent years does not appear to have increased identity fraud and/or data breaches. Since 2005, the overall incidence of identity fraud has been relatively flat and the total dollar amount of fraud has fallen—from an average of $29.1 billion for 2005-2009 to $19.2 billion for 2010-2013. While there has been a slight increase in the number of U.S. data breaches, the trend in records breached since 2005 is relatively constant or even declining slightly. All these measures show more of a decline when considered relative to the growth of the economy and e-commerce.

---

[2] See Lenard and Rubin, pp. 1-10.

[3] Liran Einav and Jonathan Levin, "The Data Revolution and Economic Analysis", Prepared for the NBER Innovation Policy and the Economy Conference, April, 2013, p. 2.

[4] Einav and Levin, pp. 5-6.

[5] See Lenard and Rubin, pp. 10-15.

The use of big data to develop predictive models is harmful to consumers.[6]

The assertion that predictive models harm consumers, if valid, would apply to quantitative analysis used for decision-making throughout the economy. Much of the concern seems to be that the predictive models may not be totally accurate. However, big data improve accuracy.

Use of credentials and test scores, from credit scores to class rankings, is ubiquitous in American life. These decisions are based on "small data"—sometimes, one test score or one data point. Big data can only improve this process. If more data points are used in making decisions, then it is less likely that any single data point will be determinative, and more likely that a correct decision will be reached.

Companies that devote resources to gathering data and undertaking complex analysis do so because it is in their interest to reduce errors. The data and the models have limited value unless they improve accuracy. Thus, big data should lead to fewer consumers being mis-categorized and less arbitrariness in decision-making.

The use of big data in marketing decisions favors the rich.[7]

The argument that data collection favors the rich over the poor is presented without evidence. Likely the concern relates to price discrimination, which involves charging different prices to different consumers for the same product based on their willingness to pay.[8] Online data collection can yield information that can be used to infer a consumer's willingness to pay for a good and in that way facilitates price discrimination.[9]

Price discrimination involves charging prices based on a consumer's willingness to pay, which in general is positively related to a consumer's ability to pay. This implies that a price discriminating firm will, other things the same, charge lower prices to lower-income consumers. Indeed, in the absence of price discrimination, some lower-income consumers would be unable or unwilling to purchase some products at all. So, the use of big data, to the extent it facilitates price discrimination, should usually work to the advantage of lower-income consumers.

Moreover, big data are being used to develop products that specifically benefit lower-income consumers. For example, ZestFinance, using many more variables than traditional credit scoring, helps lenders determine whether or not to offer small, short-term loans to people who are otherwise poor credit risks.[10] This provides a better alternative to people who otherwise

---

[6] See Lenard and Rubin, pp. 15-18.

[7] See Lenard and Rubin, pp. 20-22.

[8] See Tene, ¶ 4,6.

[9] Hal R. Varian, "Differential Pricing and Efficiency", *First Monday* Vol. 1, No. 5, August, 1996, http://www.firstmonday.dk/ojs/index.php/fm/article/view/473/394.

[10] See http://www.zestfinance.com/how-we-do-it.html.

might rely on payday lenders or even loan sharks. LendUp, BillFloat, and ThinkFinance are companies following similar models that can provide better loan options for lower-income consumers, while Kabbage and On Deck Capital provide lending services to very small businesses.

Big data have the potential to provide cost savings to consumers through the analysis and comparison of the prices of goods sold over the internet. Two successful startups, Farecast and Decide.com, use big data to help consumers find the lowest prices.[11] Farecast uses billions of flight-price records to predict the movement of airfares, saving purchasers an average of $50 per ticket. Decide.com predicts price movements for millions of products with potential savings for consumers of around $100 per item.

Firms use big data to manipulate consumers.[12]

Some writers argue that firms use big data to manipulate consumers to behave irrationally and purchase things they don't really want.[13] Drawing a boundary between what is called "manipulation" and the provision of information that helps a consumer make purchasing decisions is difficult. In general, it is not possible to determine whether any given purchase is "rational" or not, because consumers' utility functions are not directly observable.

In a market economy, firms are rewarded for giving consumers what they want. The economist's criterion of performance is how close the economy comes to maximizing "total surplus." Firms want to capture as much of that surplus as possible, and may use data to more precisely target times when they can charge consumers a higher price. However, the transaction will still be beneficial to the consumer, or the transaction would not occur; she may just capture less consumer surplus. Moreover, this is also a way that firms can efficiently price discriminate. Others may get a lower price. Importantly, such price discrimination may be necessary to cover costs of production and for the product to be available at all.

An implicit assumption in discussions of manipulation is that firms lack competition. Even assuming firms can manipulate consumers and thereby earn super-competitive profits, unless there are barriers to entry, other firms will be induced to enter and compete away those profits. This is a check on whatever manipulation might be possible.

---

[11] Mayer-Schönberger and Cukier, "Big Data: a revolution that will transform how we live, work and think", Houghton Mifflin Harcourt, 2013, p. 124.

[12] See Lenard and Rubin, pp. 22-24.

[13] Ryan Calo, "Digital Market Manipulation", Legal Studies Research Paper and George Washington Law Review (forthcoming), 2013, University of Washington School of Law, available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2309703.

Individuals will be forced to reveal data about themselves, thereby eroding privacy.

With advanced information technologies, individuals will increasingly be able to voluntarily make available a range of *verified* (because it is linked directly to the source) information about themselves, including health information, employment records, court records, driving behavior and credit history. For example, your health data may be generated by wearable monitors, your driving behavior by sensors in your car, etc.

The data made available could determine eligibility and the terms for many economically important items, including jobs, insurance and admission to schools. Those with the most favorable data will find it in their interest to reveal it. Others will then be "forced" to reveal their data because failure to do so will reflect negatively on those who do not. Some are concerned that this contains within it the threat of unraveling privacy altogether.[14]

Generally, this phenomenon is considered efficient, because it provides information to the market and helps solve asymmetric information problems. In the absence of information, markets may "unravel" in another way. This is the well-known "lemons" problem.[15]

In the same way that prohibiting producers from hiding defects in their products leads to better products, there are positive incentive effects when individuals are unable to conceal adverse personal information. If individuals were able to conceal their credit histories, we would find more delinquent payments, which would raise the costs of borrowing generally. The fact that automobile insurance rates are lower for young males with a better grade point average is an incentive to study harder, or, at least, for the parents to make sure the student studies harder.

A simple example illustrates the potential costs of restricting this type of information sharing. It is now possible to monitor driving behavior for a variety of purposes. Mapping programs do this in order to direct drivers to the fastest route at any given time. But data can also be used by insurance companies to set rates. A driver can install a device in her car to monitor the time of day she drives and the distance traveled and have the data automatically delivered to her insurance company.[16] Presumably, safe drivers will want to do this so they can get lower rates. Insurance companies might rationally assume that drivers who failed to install such devices were less safe than those who did and charge them a higher rate. This would likely result in at least a partial "unraveling" as more and more drivers installed the monitoring devices.

Prohibiting this practice, as some privacy advocates suggest, would mean there is no payoff to voluntarily providing your monitoring data to the insurance company. This would penalize safe drivers to the benefit of average and less-safe drivers. The prohibition would increase accidents,

---

[14] See Scott Peppet, *Unraveling Privacy: The Personal Prospectus and the Threat of a Full-Disclosure Future,* Northwestern University Law Review, Vol. 105, No. 3, p. 1166.

[15] See George A. Akerlof, "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism", *The Quarterly Journal of Economics,* Vol. 84, No. 3, August, 1970, pp. 488-500.

[16] See for example: http://www.progressive.com/auto/snapshot-how-it-works/.

because even the safest drivers will drive more carefully when they know they are being monitored. There may be a significant increase in safety from drivers further down the spectrum, who would be induced to install the monitor.

Policy Considerations

There is no obvious reason to approach privacy policy issues arising from big data differently than we approach issues involving smaller amounts of data. The same questions are relevant:

- *Is there evidence of a market failure or harm to consumers?* The recent literature on big data does not provide such evidence, at least as far as the legal use of data for commercial purposes is concerned. Discussions of harm are largely speculative and hypothetical. Moreover, the available data do not indicate that there has been an increase in harm to consumers from identity fraud or data breaches.

- *If evidence of market failure or harm is found, is there an available remedy (or remedies) that can reasonably be expected to yield benefits greater than costs and therefore yield net benefits to consumers?* Since the harms are largely hypothetical, so are the benefits.

The threshold question is whether there are harms that can be reduced by the adoption of privacy policies. Otherwise, there are no benefits to privacy regulations. The absence of identified harms implies that privacy policies cannot be expected to yield significant benefits, even in the absence of costs.

The privacy remedies typically discussed are, however, likely to impose costs. A standard solution long promoted by privacy advocates is that data should only be collected for a specific identified purpose. This is reflected in the FIPPs dating back to the 1970s, the OECD Privacy Principles of 1980, the current European Union regulations, and the recommendations of the FTC's 2012 Privacy Report.[17]

Requiring that data only be collected for an identified purpose is particularly ill-suited to the world of big data. Using data in unanticipated ways has been a hallmark of the big data revolution, for commercial, research and even public sector uses. Therefore, the standard solutions that would limit the reuse or sharing of data would be particularly harmful if applied to big data because they are inconsistent with the innovative ways in which data are being used.

This also calls into question the principles of notice and choice, which become almost meaningless when data may be used in unpredictable ways. Even absent questions concerning big data, these principles have become increasingly irrelevant. As Beales and Muris note, "The reality that decisions about information sharing are not worth thinking about for the vast majority

[17] The Federal Trade Commission, *Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Businesses and Policymakers,* March, 2012.

of consumers contradicts the fundamental premise of the notice approach to privacy." They continue, "The FIPs principle of choice fares no better." [18]

Concern about the use of data for predictive scoring and the possibility that algorithms may mis-categorize individuals sometimes leads to recommendations for greater transparency. The notion that consumers should understand who is collecting their data and how they are being used is an appealing one, but it is largely meaningless, especially in the big data era where scores may be based on hundreds of data points and very complex calculations. For example, it is not clear that a person rejected for credit by a complex algorithm would particularly benefit by being shown the equation used. The FICO score, an early example of a calculation based on a complex algorithm, is virtually impossible to explain to even an informed consumer because of interactions and nonlinearities in the way that elements enter into the score.[19]

Giving consumers the ability to correct their information may be more complicated than it might appear, even aside from the administrative complexities. Consumers do have the right to correct information used in deriving their credit scores, but it is made difficult to do so, for good reason. An individual who thinks she has been wrongly categorized clearly has an interest in correcting erroneous information if that information has a negative effect. But she might also have an interest in "correcting" valid information that would adversely affect the decision, or inserting incorrect information that would have a positive effect. Distinguishing between these various "corrections" may be quite difficult.

The purpose of collecting information that affects decisions about individuals—e.g., credit decisions, insurance decisions, or employment decisions—is to ameliorate asymmetric information problems. As Beales and Muris point out, "In our economy, there are vital uses of information sharing [such as credit reporting] that depend on the fact that consumers cannot choose whether to participate."[20]

Moreover, if we make it easier for individuals to access their data then we also make it easier for those bent on fraud to access the same data. If fraudsters have access to large amounts of data about a person, they can more easily defraud that individual. Thus, ease of consumer monitoring is at best a two-edged sword.

Conclusion

Any attempt to limit "harmful" uses of information will limit beneficial uses as well. Although many have suggested "meaningful oversight" as a remedy for what they perceive as harms to

---

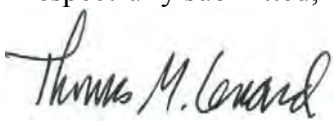
[18] J. Howard Beales and Timothy Muris, *Choice or Consequences: Protecting Consumer Privacy in Commercial Information,* University of Chicago Law Review, 2008, pp. 113-118.

[19] The major inputs to a credit score are well known; however, the calculation of credit scores from credit report data is proprietary and exceedingly complex. See for example FDIC, *Credit Card Activities Manual*, Ch. 8 – Scoring and Modeling, 2007, available at: http://www.fdic.gov/regulations/examinations/credit_card/.

[20] Beales and Muris, p. 115.

consumers, there has been no evidence of any concrete privacy harms. Given this lack of data and analysis, particularly in a new market such as the electronic use of information, it is much more likely that uninformed regulation will stifle innovation rather than provide net benefits. The "familiar solutions"—such as those that would limit the reuse or sharing of data—would seem to be particularly harmful because they are inconsistent with the new ways in which big data are being used.

Respectfully submitted,



Thomas M. Lenard
President and Senior Fellow
Technology Policy Institute

Comments of the Internet Association in Response to the White House Office of Science and Technology Policy's Government "Big Data" Request for Information[1]

The Internet Association[2] welcomes the opportunity to submit the following comment in response to the White House Office of Science and Technology Policy's (OSTP) request for information on "big data" to inform its 90-day review. OSTP seeks comments from interested parties to examine how "big data" will affect Americans' daily lives, the relationship between government and citizens, and how the public and private sectors can spur innovation and maximize the opportunities and free flow of information while minimizing the risks to privacy.[3] OSTP has a unique opportunity as part of this 90-day review to educate the public on the benefits of big data to society and how the government and companies are leveraging big data in a privacy-enhancing way consistent with the robust and effective privacy regime that exists in the United States (U.S.).

The current U.S. policy framework is critically important to the continued growth of the Internet ecosystem. It provides a framework by which companies respect and promote the privacy of the people who use their services, while allowing for technological advancements benefitting individual users and society. We encourage the Administration to highlight the

[1] The Internet Association comments electronically submitted at bigdata@ostp.gov.

[2] The Internet Association represents the world's leading Internet companies including: Airbnb, Amazon, AOL, eBay, Expedia, Facebook, Gilt, Google, IAC, LinkedIn, Lyft, Monster Worldwide, Netflix, Practice Fusion, Rackspace, reddit, Salesforce.com, SurveyMonkey, TripAdvisor, Twitter, Uber Technologies, Inc., Yelp, Yahoo!, and Zynga.

[3] White House Office of Science and Technology Request for Information, 79 Fed. Reg. 12251. (Mar. 4, 2014).

1100 H Street NW, Washington, DC 20005
www.InternetAssociation.org

advantages and successes of the existing framework in its 90-day report, just as it did in releasing the Consumer Privacy Bill of Rights in 2012.[4]

Finally, the Internet Association encourages the federal government to devote research and development towards determining solutions to make more government data available, fund research towards emerging technologies as well as support research to determine methods to educate consumers on "big data" practices.

I. The White House Office of Science and Technology Policy should dedicate efforts to further exploration of methods and policies to effectuate government surveillance reform.

The Administration's review of this issue comes at a challenging time for the Internet industry. The ongoing revelations concerning the nature and scope of government surveillance programs create the potential for diminished user trust and confidence in Internet services. It is critical, therefore, that the Administration and Congress focus on policy solutions that are directly responsive to concerns that have surfaced in light of these revelations.

As it explores the policy frameworks under which "big data" issues should be examined, we urge the Administration to be cognizant that government surveillance and commercial privacy are separate and distinct issues. Given that Internet companies aim to provide transparency, choice, and control to consumers, efforts to conflate these issues are

[4] The White House, Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy at i (Feb. 2012), *available at* http://www.whitehouse.gov/sites/default/files/privacy-final.pdf.

1100 H Street NW, Washington, DC 20005
www.InternetAssociation.org

counterproductive, particularly given how little transparency citizens currently have when it comes to government surveillance.

As discussed in greater detail in Section III, our current public policy framework provides an effective and balanced approach in allowing for this innovation while protecting consumers. The success of existing laws and frameworks in guiding the private sector does not justify a need to initiate wholesale changes in how we approach these policies. Rather, we urge the Administration to continue its effort to address concerns that emerged in the aftermath of the NSA revelations. There have been real consequences from the disclosure of current surveillance practices, particularly outside the U.S. and the Administration should act swiftly to prevent declining confidence in U.S.-based Internet companies. Reforming surveillance law will help to rebuild user trust and to maintain the U.S.'s competitive edge in technological innovation. In the near term, the Administration can advance this debate by supporting the following reforms:

- **First, the Administration should endorse legislation pending in Congress that would update the Electronic Communications Privacy Act to** require governmental entities to obtain a warrant before they can compel online companies to disclose the content of users' communications. The Internet Association - along with more than 100 companies, trade associations, and civil society organizations - supports legislation pending in the House (H.R. 1852) and Senate (S. 607) that would update ECPA in this manner. Over 100,000 citizens have petitioned the White House to support this update to ECPA without exceptions that would whittle away at the bright-line, warrant-for-content rule that these bills would create. This update will provide certainty to both service providers and citizens that their content stored online will receive the same Fourth Amendment protections as their offline information.

- **Second, the Administration should build on the emerging consensus that U.S. law should prohibit the bulk collection of communications metadata to support comprehensive surveillance reform.** We understand from recent reports that the Administration intends to end the existing bulk telephony metadata collection program in favor of a legal regime that is narrowly tailored and subject to greater judicial oversight. The general thrust of Congressional legislation is consistent with this approach, and we believe it is critical that the bulk collection of Internet metadata be

1100 H Street NW, Washington, DC 20005
www.InternetAssociation.org

encompassed within Congressional legislation. More remains to be done, and we urge the Administration to continue engaging Internet industry stakeholders on policy prescriptions that would form the basis for comprehensive government surveillance reform.

We encourage the Administration to lead the world in making reforms that ensure government surveillance efforts are clearly restricted by law, proportionate to the risks, transparent and subject to independent oversight. In so doing, the Internet Association recommends the Reform Government Surveillance principles.[5] By addressing these serious gaps in federal law, the government will demonstrate that it takes seriously its responsibility to protect the privacy of millions of Americans.

II. The ability to leverage datasets in a privacy-protective manner has allowed for continued innovation and important advancements within the Internet industry. The government should devote research and development to generate additional innovative solutions and to educate consumers on government and commercial "big data" practices.

The Internet industry uses large datasets to enable activities that promote the public good and facilitate the creation of beneficial products and services, consistent with robust, existing legal frameworks that safeguard against harms arising from the misuse of personal data. For instance, cloud service providers' housing of datasets allows scientists and researchers to address societal changes relating to cancer research and climate change. Additionally, these datasets allows Internet companies to ensure the efficient operation of their platforms. The following examples illustrate these benefits:

- **Cloud services for scalable and reusable analytics.** With the increasing complexity of genomic research, scientists are using Amazon Web Services (AWS) as a platform to store and capture large volumes of scientific data in a cost-effective and timely manner.

[5] *See* Reform Government Survhttps://www.reformgovernmentsurveillance.com/ (last visited Mar. 31, 2014).

1100 H Street NW, Washington, DC 20005
www.InternetAssociation.org

AWS allows researchers to build scalable and reusable analytical tools. For instance, AWS hosts data for the 1000 Genomes Project, an international public-private effort that seeks to build the most detailed map of human genetic variation available. The project has grown to 200 terabytes of genomic data including DNA sequenced from more than 1,700 individuals that researchers can now access on AWS for use in disease research. The samples for the 1000 Genome Project are collected via informed consent and are mostly anonymous. AWS permits anyone with an account to access vast amounts of data to use in research to gain further insights into human health and diseases.

- **Improved communication on efficient and secure platforms.** It is common for a company to randomly distribute information across many servers. For example, to make its infrastructure more efficient, Facebook uses data analysis to intelligently distribute information by mapping data storage based on an understanding of how people communicate with their friends. This analysis protects privacy by relying on aggregated review of communication patterns at scale, and it not only promotes energy efficient infrastructure but also helps people communicate more reliably.

- **Improving users' daily lives.** Google Now helps Google users better manage their lives. With the affirmative consent of users that avail themselves of the service, Google Now uses information in the background to bring users the information they want, when they want it - showing the weather as you start your day, finding the best route to your next event to avoid traffic, telling you a flight is delayed or checking your favorite sports team's score. To do this, Google integrates information from the user with a number of back-end data sets, like maps, flight schedules, calendars, emails, weather, and public transit. Google Now also surfaces AMBER alerts, weather warnings, and other public alerts. Taking these many kinds of data - operational data, process data, statistical data, aggregated data, linguistic data, ethnographic data, and metadata - and linking them together to do something useful is one of the challenges of "big data," and provides significant value to Google users.

- **Spam filtering.** Also based on data analytics, Internet companies are able to enhance their ability to analyze message traffic to prevent spam, while leveraging de-identification technologies that reduce unwanted communications for all users without the need to maintain or share identifiable information about individual's communications.

These beneficial uses of big data in the Internet space are only the beginning. More work must be done in order for our society to gain the full benefits that can be achieved in areas of public health, economic growth, education, and social research from the analysis of large data sets. As a key component of its "big data" review, the Administration should commit to devoting substantial resources towards research and development aimed at unlocking the societal

1100 H Street NW, Washington, DC 20005
www.InternetAssociation.org

benefits of large datasets, in both the private and public sectors. The Administration's current Open Data Initiative, intended to solve complex problems ranging from consumer to environmental issues, is a commendable first step at unleashing government data to fuel scientific discovery and spur innovative growth. For example, the Administration's recent launch of the National Oceanic and Atmospheric Administration (NOAA) open data website[6] will undoubtedly promote both of these goals by making publicly available scores of datasets containing valuable environmental data. Scientists will be able to harness this data to assess environmental risks, and start-ups will utilize it to provide critical products and services to consumers. While this initiative continuously seeks to make positive contributions, more needs to be done. The Administration should increase its efforts to make more government data readily accessible, which could create significant societal and economic benefits, including job creation. And, it should do this in a way that illustrates how large data sets can be used for research in a privacy-preserving manner.

Additionally, the National Science Foundation and other research funding should focus on areas of emerging research such as: (a) how to analyze big data effectively, (b) how to de-identify large data sets (e.g., privacy enhancing technologies), (c) how to build accountable systems, and (d) how to safely release data for research purposes. This research is just beginning, and it is critical to our long-term ability to attain big data benefits. The Administration should prioritize support for these areas of emerging research before seeking to circumscribe cutting

[6] *See* Press Release, NOAA, *NOAA announces RFI to unleash power of 'big data' Agency calls upon American companies to help solve 'big data' problem, available at* http://www.noaanews.noaa.gov/stories2014/20140224_bigdata.html (Feb. 24, 2014).

1100 H Street NW, Washington, DC 20005
www.InternetAssociation.org

edge research in the private sector.

Lastly, although many of the discussions at workshops convened as a part of this "big data" review have focused on data collected by websites, the reality is that "big data" is collected in many contexts beyond websites, and its collection, analysis, and use does not depend on the existence of the Internet. The Internet of Things, mobile devices, wearable computing, and many other sectors and platforms are also involved in collecting data and performing large-scale analytics, and the retail sector has analyzed "big data" since before the commercial Internet existed. One key outcome of the Administration's work on "big data" should be an effort to educate people about the full range of "big data" practices – particularly in sectors that, unlike the Internet, may not be at the top of consumers' minds when thinking about privacy. In this regard, the government should focus research funding on efforts to pioneer methods that organizations, including government agencies, can use to help consumers gain meaningful understanding of how their data is collected and used. For example, the government can support usability studies aimed at discovering the best methods to inform people about the life cycle of their data, or how they can exercise control when organizations share their data with third parties.

III. There is nothing dramatically new that would suggest a wholesale move away from our existing framework for regulating data, particularly given the breadth and effectiveness of federal and state enforcement as compared to regulatory regimes in other jurisdictions.

The U.S.'s flexible, multi-layered privacy regime is capable of responding forcefully to remedy violations of consumers' privacy, while permitting businesses that engage in privacy-protective practices to flourish. Under the current U.S. regime, organizations that engage in harmful information practices, in the big data context or otherwise, are subject to a wide range of

1100 H Street NW, Washington, DC 20005
www.InternetAssociation.org

laws and regulations at both the federal and state levels. At the federal level, privacy laws protect information in the financial, insurance, educational, telecommunications, credit and health sectors, as well as information about children. At the state level, privacy laws cover these areas and more: employee data, spam, event data recorders, phishing and spyware. The U.S. also requires robust information security and data breach notification, which is the front line in preserving and protecting information privacy.

Beyond these sector-specific laws, the Federal Trade Commission Act and equivalent laws at the state level broadly prohibit "unfair or deceptive" acts or practices, and authorize enforcement actions by regulators. A defining feature of the U.S. commercial privacy regime is that it is calibrated to respond to the greatest public concerns. The Federal Trade Commission (FTC) and state attorneys general have acted on consumers' concerns about identity theft and data breaches by taking swift action to punish bad actors in the ecosystem and protect consumers who have been harmed. The FTC and state attorneys general are sensitive to emerging privacy and consumer protection concerns ranging from deceptive health claims to mobile tracking, the "Internet of Things," and "data brokers." Through panels, reports, investigations, consent decrees, and consumer education initiatives, U.S. regulators and law enforcement officials have proven remarkably adept at protecting consumer privacy in a balanced and agile manner that focuses administrative resources on the worst harms while allowing industry to consumers offer innovative products and services.

Nearly two years ago, the White House proposed a Consumer Privacy Bill of Rights, a framework that was intended to capture common privacy principles in a *comprehensive* way. The White House lauded the strength of the U.S. privacy regime when it released the Consumer Privacy Bill of Rights:

1100 H Street NW, Washington, DC 20005
www.InternetAssociation.org

> "the consumer data privacy framework in the United States is, in fact, strong. This framework rests on fundamental privacy values, flexible and adaptable common law protections and consumer protection statutes, Federal Trade Commission (FTC) enforcement, and policy development that involve a broad array of stakeholders. This framework has encouraged not only social and economic innovations based on the Internet but also vibrant discussions of how to protect privacy in a networked society involving civil society, industry, academia, and the government."[7]

Since the release of the Consumer Privacy Bill of Rights, we have seen the continued and sustained success of privacy regulation at the federal and state levels, as well as the first successes arising from multi-stakeholder efforts to build new sectoral privacy improvements that are legally enforceable but would not have been feasible through legislative means.

We urge the White House to recognize in its 90-day report that our current legal framework in the U.S. can robustly address commercial data practices and is a flexible model for the continued growth of the innovation economy. Numerous jurisdictions are considering measures that would restrict the free flow of data, including data localization requirements and restrictive privacy provisions. Such measures would impede economic growth and even extinguish the promise of big data benefits. The Safe Harbor Framework— an important mechanism for U.S. companies that transfer data from Europe to the U.S.— is in danger of being scaled back or even suspended. Trade negotiations, in particular the Transatlantic Trade and Investment Partnership (T-TIP) with the EU have been adversely impacted by both the NSA revelations and a perception that the U.S. privacy regime is not as privacy-protective as the EU model.

The Internet industry appreciates the Administration's commitment to promoting the continued global competitiveness of U.S. businesses, and consistent with that commitment we urge the White House to uphold its responsibility to protect the economic interests of U.S.

[7] The White House, Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy at i (Feb. 2012), *available at* http://www.whitehouse.gov/sites/default/files/privacy-final.pdf.

1100 H Street NW, Washington, DC 20005
www.InternetAssociation.org

industry by accurately characterizing the U.S. commercial privacy regime as fully protective of consumer privacy.

IV. Conclusion

The Internet Association is pleased to provide input in response to OSTP's request for information on "big data." As the current U.S. public framework is balanced and effectively achieves its goal of protecting consumers while allowing for continued innovation, the system is not in need of wholesale changes. We hope that government surveillance reform remains a top priority for the Administration, and resources are devoted to making government data available and exploring emerging areas of research.

Respectfully submitted,

/s/Michael Beckerman
Michael Beckerman
President & CEO
The Internet Association

March 31, 2014

1100 H Street NW, Washington, DC 20005
www.InternetAssociation.org

CHAMBER OF COMMERCE
OF THE
UNITED STATES OF AMERICA

WILLIAM L. KOVACS
SENIOR VICE PRESIDENT
ENVIRONMENT, TECHNOLOGY &
REGULATORY AFFAIRS

1615 H STREET, N.W.
WASHINGTON, D.C. 20062
(202) 463-5457

March 31, 2014

VIA ELECTRONIC FILING

Ms. Nicole Wong
Deputy Chief Technology Officer
Office of Science and Technology Policy
Eisenhower Executive Office Building
1650 Pennsylvania Avenue, NW
Washington, DC 20502

Re: *Big Data Study*

Dear Ms. Wong:

The U.S. Chamber of Commerce ("Chamber")[1] is pleased to submit these comments to the Office of Science and Technology Policy ("OSTP") in response to its Request for Information regarding the "Big Data" study.[2] Given the range of technologies and market participants involved in big data, the Chamber believes that self-regulation and best business practices—that are technology neutral—along with consumer education serve as the preferred framework for protecting consumer privacy and security while enhancing innovation, investment, and competition essential to big data. Additionally, the Chamber calls on the Administration to support legislation updating the Electronic Communications Privacy Act.

1) What are the Public Policy Implications of the Collection, Storage, Analysis, and Use of Big Data?

To help our economy continue to recover, the Chamber believes that big data will be a key component in the creation of jobs and innovation. All sectors of the U.S. economy—including financial services, manufacturing, transportation and many more—collect and use big data to spur sales and job growth, enhance productivity, enable cost savings, improve efficiency, and protect consumers. Big data is used in many beneficial ways in our economy and by our society, including but certainly not limited to: improving healthcare, enabling businesses to offer

[1] The U.S. Chamber of Commerce is the world's largest business federation, representing the interests of more than three million businesses of all sizes, sectors, and regions, as well as state and local chambers and industry associations, and dedicated to promoting, protecting, and defending America's free enterprise system.
[2] *Request for Information Regarding Government "Big Data,"* 79 Fed. Reg. 12251 (Mar. 4, 2014), *available at* http://www.gpo.gov/fdsys/pkg/FR-2014-03-04/pdf/2014-04660.pdf. (*"RFI"*).

enhanced customer service, detecting and preventing fraud as well as authenticating individual identities, and refining the manufacturing of products.

According to market research-firm IDC, the market for big data in 2014 is expected to reach $16.1 billion and grow 6 times faster than the overall IT market.[3] The efficient use of big data allows manufacturers to reduce the cost of product development and assembly by up to 50 percent, and decrease the amount of required working capital by up to 7 percent.[4] The value of big data to U.S. healthcare could be "more than $300 billion in value every year, two-thirds of which would be in the form of reducing national health care expenditures by about 8 percent."[5] Big data is used in fraud detection where identifying "weak and indirect relationships among bad actors" is key because businesses often face fraud perpetrated by numerous people finding and exploiting vulnerabilities across various parts of a company's operation.[6]

Insights derived from the large scale collection and analysis of data will drive economic and societal growth. Massive data sets combined with today's computing power enable problems to be tackled in new and often unexpected ways. By drawing on researchers from across disciplines, this data can be examined and interpreted in new and exciting ways. Data analytics resulting from rapid prototyping and experimentation will drive innovation and progress. Therefore, any federal policies in this area need to be flexible and adaptive to accommodate different uses of data along with rapidly developing technology.

Analyzing large data sets containing data about people can be accomplished ethically and responsibly. Companies work hard to ensure that their products and services are deemed trustworthy and those that fail to meet customers' privacy and security expectations can expect to face swift and decisive marketplace and reputational consequences. Allowing companies to collect, use and experiment with data more freely, and applying more harms based controls on data use would allow for robust research and development in secure protected data environments. De-identifying personal data where feasible can also help to mitigate potential harms.

Federal policy should recognize that differing risks of harm are caused by different types of data collection and usage. For example, there are fewer risks associated with non-personally identifiable data, especially when anonymized or aggregated, than with data that identifies a user. Similarly, encrypted data also results in reduced risk. Additionally, societal benefits should be taken into account and certain uses, such as fraud prevention, may warrant fewer restrictions.

---

[3] Gil Press, "$16.1 Billion Big Data Market: 2014 Predictions From IDC And IIA," *Forbes*, Dec. 12, 2013, *available at* http://www.forbes.com/sites/gilpress/2013/12/12/16-1-billion-big-data-market-2014-predictions-from-idc-and-iia/.

[4] McKinsey Global Institute, *Big Data – The Next Frontier for Innovation, Competition, and Productivity*, at 8, May 2011, *available at* http://www.mckinsey.com/~/media/McKinsey/dotcom/Insights%20and%20pubs/MGI/Research/Technology%20and%20Innovation/Big%20Data/MGI_big_data_full_report.ashx (*McKinsey Report).*

[5] *Id.* at 2.

[6] Quentin Hardy, "IBM's Big Hope for Fraud," *The New York Times*, Mar. 20, 2014, *available at* http://bits.blogs.nytimes.com/2014/03/20/ibms-big-hope-for-fraud/?_php=true&_type=blogs&_r=0.

Thus, the Chamber urges OSTP to recognize that data-driven innovation is vital to the economy and any rigid regulations on data collection, storage, and use will hinder its potential.

2) What Types of Uses of Big Data Could Measurably Improve Outcomes or Productivity with Further Government Action, Funding, or Research?

Use of big data can measurably improve outcomes and productivity across the economy. Arguably, companies are more cognizant of public policy concerns associated with big data than government because of the severe reputational and financial consequences that can result. These uses of big data in the private-sector generally present no or few public policy concerns, while, in other cases, it is too early to know whether any such concerns will arise. Thus, to avoid jeopardizing the tremendous innovation and growth in this area, policymakers should restrain from regulating big data unless there is a market failure. Also, it is worth noting, for example, that where specific concerns have been expressed, the Department of Commerce's National Telecommunications and Information Administration has convened a multistakeholder process to address such issues as facial recognition and mobile application transparency.

To help better protect consumers and solve other government public policy issues, policymakers should increase the number and raise the profile of contests and/or prizes issued to private-sector technologists. The government could apply data analytics to offer better services and information to its citizens. Additionally, policymakers should challenge federal agencies to focus on effectiveness and measurement of government projects and regulations. Agencies should focus on calculating and analyzing workflows, regularly seek data to guide their priorities and management, train their staff on data-driven policymaking, and prioritize data quality controls to ensure best results.

For example, according to a recent survey of federal IT professionals, "63% feel that big data will help track and manage population health more efficiently, 62% view big data as a way to significantly improve patient care within military health and Veterans Affairs (VA) systems, and 60% believe big data will improve how preventive care is delivered."[7] Yet, only "29% of those surveyed have hired trained professionals to manage and analyze big data, or educated senior management on related issues."

3) What Technological Trends or Key Technologies Will Affect the Collection, Storage, Analysis and Use Of Big Data? Are There Particularly Promising Technologies or New Practices for Safeguarding Privacy While Enabling Effective Uses of Big Data?

The increasing importance of data analytics to our society has given rise to data science, a multi-disciplinary field of research. Additionally, data literacy will become important as well to help train people on how to view, communicate, and use data once it is analyzed. The United

[7] Elena Malykhina, "Agencies See Big Data as Cure for Healthcare Ills,"*InformationWeek*, Mar. 26, 2014, *available at* http://www.informationweek.com/government/big-data-analytics/agencies-see-big-data-as-cure-for-healthcare-ills/d/d-id/1127924.

States could face a shortage by 2018 of "140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions."[8]

The Chamber believes that privacy protective technologies and practices will develop more rapidly in the marketplace as consumers look to protect their privacy through increasing interactions in the digital environment. As we see with the rapid increase in encrypted messaging, the market for companies selling easy to use privacy tools will grow organically in response to consumer demands. Cybersecurity professionals are continuously devising new and improved ways to protect data and combat threats from criminals and, sometimes, governments. Data scientists are working on new and improved means of disguising or de-identifying personally identifiable information. These types of technological advancements promoting consumer protection show the marketplace is working and that government intervention is not needed. Policymakers should restrain from regulating big data unless there is a market failure.

4) How Should the Policy Frameworks or Regulations for Handling Big Data Differ Between the Government and the Private Sector? Please Be Specific as to the Type of Entity and Type of Use (e.g., Law Enforcement, Government Services, Commercial, Academic Research, etc.).

The Chamber strongly urges the U.S. government to distinguish privacy issues associated with national security from those related to commercial privacy practices. A distinction must be made between big data that is collected and used by the government—backed by the inherent power of its authority—with no opt-out available and commercial privacy practices, where there are marketplace curbs on bad behavior and, for some business sectors, legal and regulatory requirements to safeguard consumer data.

Additionally, the Chamber urges that the big data study recommend support for legislation updating the Electronic Communications Privacy Act (ECPA) as an action that could immediately improve consumer privacy.[9] Big data is a manifestation of the technology revolution that has changed the way business operates and the way individuals communicate in their daily lives. More and more, Americans in their personal and professional capacities are voluntarily storing information in the cloud (e.g., email, documents, photographs, medical records, business plans, etc.) containing private or proprietary data encompassing the whole range of human existence.

ECPA, which sets standards for government access to private communications, is critically important to businesses, government investigators and ordinary citizens. Though the law was forward-looking when enacted in 1986, technology has advanced dramatically and ECPA has been outpaced. Courts have issued inconsistent interpretations of the law, creating uncertainty for service providers, for law enforcement agencies, and for the hundreds of millions

[8] *McKinsey Report* at 3.
[9] The Chamber is a member of Digital Due Process (http://www.digitaldueprocess.org), a coalition that advocates for ECPA reform.

of Americans who use the Internet in their personal and professional lives. Moreover, the Sixth Circuit Court of Appeals has held that a provision of ECPA allowing the government to obtain a person's email without a warrant is unconstitutional.

Bipartisan legislation—S. 607, the "Electronic Communications Privacy Act Amendments Act of 2013," and H.R. 1852, the "Email Privacy Act"—would update ECPA in one key respect, making it clear that, except in emergencies, or under other existing exceptions, the government must obtain a warrant in order to compel a service provider to disclose the content of emails, texts or other private material stored by the service provider on behalf of its users.

This standard would create a more level playing field for technology. These bills would cure the constitutional defect identified by the Sixth Circuit, allow law enforcement officials to obtain electronic communications in all appropriate cases while protecting Americans' constitutional rights, and provide clarity and certainty to law enforcement agencies at all levels and to American businesses developing innovative new services and competing in a global marketplace.[10]

5) What Issues are Raised by the use of Big Data Across Jurisdictions, such as the Adequacy of Current International Laws, Regulations, or Norms?

Some governments are using concerns over government access to data for law enforcement and national security purposes as a basis to pass misguided rules that either threaten to cut off the international flow of data or require localized servers and storage. Some of these rules are good faith attempts to address public concerns, but several governments appear to be advancing protectionist measures under the guise of national security concerns.

Ultimately, rather than creating jobs, these rules merely serve to raise costs on local businesses, reduce efficiency, and cut off access to customers while depriving consumers of the ability to use the products and services of their choosing. Whether blatantly protectionist or out of genuine concerns, cross-border data transfer restrictions are misguided for a number of reasons, but ultimately only serve to harm domestic economies.

Regardless of intent, data transfer restrictions imposed by national laws that impede the free flow of data also cause significant ramifications globally as well. The Chamber urges the U.S. government to fight against any attempts by other governments to restrict the ability to transfer, store, and process data across jurisdictions.

The European Union and a number of other foreign governments are considering new approaches to data privacy that may curtail the ability to utilize data across jurisdiction. It is essential that the U.S. government continue efforts to preserve the U.S. – EU Safe Harbor and

[10] *See* Multi-Industry Letter Supporting S. 607, the "ECPA Amendments Act" to the Senate Judiciary Committee, Apr. 22, 2013, *available at* https://www.uschamber.com/letter/multi-industry-letter-supporting-s-607-ecpa-amendments-act.

expand efforts to develop approaches to privacy that can bridge the differences from different privacy regimes around the world. The APEC Cross-Border Privacy Rules (CBPR) represent one such example that can be expanded to other regions and governments and both the U.S. – EU TransAtlantic Trade and Investment Partnership (TTIP) and Trade in Services Agreement (TISA) provide additional unique opportunities.

Finally, global progress and democracy depend on the free flow of information and ideas. The upcoming meeting of the U.N.'s International Telecommunications Union (ITU) in October combined with the plans to transition oversight of the Internet Corporate of Assigned Names and Numbers (ICANN) have created a precarious situation for the future of the Internet and the use of data across multiple jurisdictions. It is essential for the U.S. government to continue to guard against any efforts by foreign governments to upend the current, successful multistakeholder governance model under which the Internet has heretofore thrived.[11]

Thank you for the opportunity to provide comments on this important matter.

Sincerely,

William L. Kovacs

[11] See generally http://europa.eu/rapid/press-release_IP-14-142_en.htm.

To the Office of Science and Technology
Government "Big Data" Request for Information-March 31st 2014
Project: US Leadership for the Revision of the 1967 "Space Treaty"

Introduction

The following discussion is submitted on behalf of an innovative civil society initiative which is proposing for a comprehensive, US led revision to the "1967 Treaty on the Peaceful Uses of Outer Space" and the early placement of a specialized Library of Congress unit dedicated to this topic as a primary mechanism for US leadership. An original focus of the program will be for the establishment of a global information (big data) platform for planetary development and sustainability through expanded, treaty related UN mechanisms.

Discussion

Our proposal seeks to address the questions posed and to direct the big data potential into an inclusive and ongoing democratic and participative global governance forum. The leading inquiry as to "Whether the United States can forge international norms on how to manage this data" identifies the need for US leadership in the definition and direction of big data usage for national and global development and security/defense purposes. While American organizations are generally acknowledged as the original pioneers and proponents of big data infrastructure systems, we may typically assume that other countries will presently acquire the same abilities. International collaboration and cooperation within the emerging fields will depend, to a large extent, upon the designation and availability of a suitable and accommodating global resource center, and the provision of secure and verifiable user group environments. Centralized resources of such types may be described and formally brought forward through a specialized unit at the US Library of Congress, and suitably investigated and coordinated through agency, NGO and policy structures.

A) What are the public policy implications of the collection, storage, analysis and use of big data? B) For example, do the current U.S. policy framework and privacy proposals for protecting consumer privacy and government use of data adequately address issues raised by big data analytics?

An effective US policy will be required for the establishment of globalized big data applications that enable both national and international enterprises, while protecting localized interests and the specifics of secure systems. The US has already recognized the value of big data resources and technology, and the early adaptation of collaborative platforms will provide prospects for additional developments, as governments and commercial interests address big data management through international partnerships and consortiums within various fields of influence. In this sense, US policy for "big data" can provide a genuine opportunity for the world community. The proposed US Library of Congress unit and informational system will offer critical resources for this effort to build internationalized commitments, strategies and applications in this field.

Visionary US policy may serve to properly demonstrate the effectiveness and benefits of big data applications and uses, defining standardized and equitable guidelines and parameters, while negotiating

controversy and competing interests. The continuing objective for civil society assurance (at many levels) needs to be well formulated through a US led initiative, finding necessary and compatible responses to the pervasive impact of forthcoming big data technologies at both individual, national and global scales.

A) What types of uses of big data could measurably improve outcomes or productivity with further government action, funding or research? B) What types of uses of big data raise the most public policy concerns? Are there specific sectors or types of uses that should receive more government and/or public attention?

Big data applications will create worthwhile insights and opportunities in areas such as industrial and manufacturing reform, economic development, sustainability, educational and health programs, natural resource development, agricultural productivity, energy and other infrastructure development, to mention a few. Yet related security, privacy and societal concerns for big data applications may be considered to be controversial, although they can be adequately addressed through both technological and legislative stakeholders. Real time access to qualified resources through data based techniques will enable notable efficiencies within essential systems, while permitting flexible adjustments according to local and transient needs. While such abilities in the commercial arena will help inform and facilitate open market systems, governmental priorities for the big data planning system should also be formally directed toward the capable provision of value added public services. In this sense government oversight for big data collation can provide positive objectives for the supporting economic and commercial trends and optimize problem solving prospects, ensuring that civil society interests remain the primary motivating factor for “big data” systems. We suggest that these mutable trends be evaluated utilizing a “risk based approach”, identifying, accessing and analyzing applicable historical information and facilitating forward looking projections.

A) What technological trends will affect the collection storage and analysis of big data? B) Are there particularly promising technologies or new practices for safeguarding privacy while enabling effective uses of big data?

The evolution and expansion of space based and related communications technologies promises to rapidly multiply the ongoing advances in data storage and collection, while “social media” applications continue to bring forward innovated data based systems that are commercially viable. Within the next decade, new space based systems and communications assets will significantly expand the reach and capacity of both governmental and private sector interests within the national and global arenas.

Recent cryptographic technology and ability, as evidenced in the evolution of crypto currencies, the Tor system, and the prospects of internet based governance platforms, indicate the probable development of secure communications and systems, which can be combined with evolving point-to-point communications technology in coming to terms with invasive, intrusive or antagonistic approaches that threaten both personal and national security. Once again, a centralized and reliable informational unit may serve to inform stakeholders not only about security and privacy threats, but also propose and communicate means of dealing with them.

A) How should policy frameworks or regulations for handling big data differ between the government and the private sector? B) Please be specific as to the type of entity and type of use (e.g., law enforcement, government services, commercial, academic research, etc.).

Taking note of how US policy deals with other areas that involve government vs. private sector interests, we observe that government tends to reserve to itself certain critical activities, with a greater reliance on a highly regulated private sector. In the case of Big Data, where gathering and collation activity by private sector interests may inevitably be reviewed or used for governmental activities, we can take note of how the telecommunications industry is currently regulated and how checks and balances are placed on government requests and uses for data. Thus, the complementary differences between the government framework and the private sector framework for “Big Data” must necessarily be reconciled because the government will utilize private sector data resources under standardized guidelines, and the private sector will receive government data based resources according to formal rules and regulations.

In this new field, without the presence of a mandated federal agency, oversights of both government and private sector “big data” application may be legislated haphazardly by each local and regional government. Though predicting the specific directions Big Data will take, and which agencies will be most directly influenced is a complex task since all levels of government and private endeavor will be affected, it is apparent that the evolving technologies of “E-governance” will facilitate dialog, referendum and detailed civil society proposals, as evidenced by early adopters in select communities.

What issues are raised by the use of big data across jurisdictions, such as the adequacy of current international laws, regulations or norms?

In light of the pressing nature of existing (and often unexpected or invasive) applications of this technology, within both private and public interests, and taking into account the unregulated technological innovations that may being used, communities and governments across the world are becoming concerned about the impact these new applications will have on their privacy, freedom, and self-determination. Internationalized treaty mechanisms may be considered as the most appropriate means for the harmonious formulation and definition of the big data prospects, especially those which automatically cross national borders because they lie directly ahead, circling the globe. For this reason we propose that the **“1967 Treaty on the Peaceful Uses of Outer Space”** provides an appropriate venue and offers genuine opportunity for establishing equitable and balanced background conditions within the rapidly evolving technological fronts.

Through the early vision of a “Big Data” platform within the **“1967 Treaty on the Peaceful Uses of Outer Space”** we may come to terms with significant issues related to security, both for national and international interests, obtaining a genuine pathways for topics such as crisis containment and hot spot deployment, arms control and verifications processes, to mention just a few. Other types of immediate impacts can be expected, particularly for the assessment of climate change variables, public health, and epidemiology, along with other arenas that require multiple time-sensitive responses and solutions.

Conclusion

Choices must be made for the way forward, as innovative big data prospects are designed, established, distributed and implemented by private and public interests across the world. It is apparent that US leadership can alleviate many concerns through the key strategies of international agreements, partnerships and participatory ventures, while ensuring secure and verifiable big data structures and equitable development of applications. As a venue for such a task, the **“1967 Treaty on the Peaceful Uses of Outer Space”, should** be considered as an explicit statement of the multiple global identities and interests, permitting it to also become the mechanism for dealing with coming global “Big Data” innovations.

In short, we believe it is essential that America take up a leading responsive and responsible position for the positive evolution of big data applications and their objectives. We propose that policy in this area should be oriented towards the provision of big data platforms within the UN mandates for global development, utilizing legislative purview within international collaborations via the expansion and updating (revision) of the “1967 Treaty on the Peaceful Uses of Outer Space”.

Prepared by Amalie Sinclair, Manuel F. Perez, March 2014

Reviewed by the “Dupont Summit” WorkGroup for the revision of the **“1967 Treaty on the Peaceful Uses of Outer Space”.**

[redacted]

From: Chris Sontag <chris.sontag@viprsystems.com>
Sent: Monday, March 31, 2014 1:11 PM
To: bigdata@ostp.gov
Subject: Big Data RFI

Dear Mr. Podesta,

We are responding to the request for information regarding the issues surrounding the use of Big Data. One of the key questions your panel has asked is "are there particular promising technologies or new practices for safeguarding privacy while enabling effective uses of big data?" My company, VIPR Systems, has such a technology, and it will be deployed for mainstream use in the fourth quarter 2014.

If we were to lay out the characteristics of the "optimal solution" to the problems raised by the use of Big Data there are 3 that stand far above the others:

1. 1. The individual should be in control of the use of his/her personal information and be able to protect it from fraud and abuse.
2. 2. The various commercial enterprises should be able to use data for legitimate purposes, such as analyzing consumer behavior, developing targeted marketing initiatives, etc. yet not be able to abuse that data in ways the individual does not want nor authorize.
3. 3. Government, and particularly law enforcement should be able to retrieve specific personal data that may be necessary without violating privacy protections afforded individuals.

At first glance these 3 points would seem to be completely incompatible, and in the case of #1 and #2, mutually exclusive. However, the patented VIPR System technology accomplishes these three things, and does so in the most desirable way possible – it utilizes existing infrastructure both in terms of data transaction exchange and in terms of regulatory oversight. The technology enables consumers to use a "tool" to control the access and use of personally identifying information (PII) so that any information that is collected is "authorized" for use, and can be subsequently removed if the collectors abuse the initial authorization.

We would be very interested in providing more detail about our technology and the release to the general public later this year. We believe our technology solves the current problems of privacy protection, fraud prevention, and identity management both today and in the future. If you would like to receive more details, please contact me at the address below.

Thank you.

Chris Sontag
VIPR Systems, LLC
[redacted]
chris.sontag@viprsystems.com

3108 Fifth Avenue
Suite B
San Diego, CA 92103

Comments of the World Privacy Forum
To: Office of Science and Technology Policy
Re: Big Data Request for Information

Via email to bigdata@ostp.gov

Big Data Study,
Office of Science and Technology Policy,
Eisenhower Executive Office Building,
1650 Pennsylvania Ave. NW.,
Washington, DC 20502

March 31, 2014

[In response to Question 2, FR Doc. 2014-04660, March 4, 2014. <https://www.federalregister.gov/articles/2014/03/04/2014-04660/government-big-data-request-for-information>.]

Thank you for the opportunity to respond to the Request for Information regarding the Big Data Study. The World Privacy Forum is a non-profit, non-partisan public interest research group. We focus on in-depth research on privacy matters in several key areas, including large datasets. More information about our work is available at www.worldprivacyforum.org.

Privacy must be re-imagined for a digital era. We are in the midst of a time in which complex data flows involving large data sets are not just occasional, but commonplace. Big data has many uses, including positive ones. In analyzing the privacy impact of big data, we see a range of issues, but we continue to believe that Fair Information Practices should be the bedrock of any policy for large datasets. Among the many possibilities that big data presents, three issues in particular stand out to us as important focal points:

- The quality of the predictive capacity of the data
- The appropriateness of the uses of the data sets

- Handling problems arising from analysis, including in vulnerable populations

An underlying issue for all big data discussions is the identifiability level of the data sets, and if de-identified, the probabilities for re-identification of the data sets.

Data Quality

Good predictions require good data quality. Current datasets may be easy to collect, but inaccuracies may be significant and variable across datasets and can be made worse if data linkages are of poor quality. In looking at aggregated data, inaccuracy may be less of an issue. But if large datasets are used to make decisions around individuals, particularly identifiable individuals, then errors stemming from either underlying factors or analytic model error rate can be problematic and deserve policy attention. Privacy principles that call for data destruction (or de-identification) and tying data uses to original purposes remain important. Large datasets cannot be exempt from data quality principles. Ultimately, high data quality is good for all parties involved.

Appropriateness of data uses

In a groundbreaking series of articles, the Associated Press used EPA data[1] to map the air quality risk scores for every neighborhood in the U.S. The AP mapped existing EPA toxicity risk scores to socio-economic and racial factors for each neighborhood from the 2000 Census to determine who was breathing the dirtiest air in America. The headlines across the country read, in some variation, that minorities suffer most from industrial pollution.[2]

The results established important understandings about neighborhoods and toxicity, and the resulting snapshot of where and how factory pollution affected neighborhoods and people was deservedly much-discussed. These results are examples of an informative and beneficial use of what today would be called large datasets or "big data."

It is helpful that the EPA has a set of meaningful best practice guidelines for analyzing its data in the EPA Risk Characterization Handbook. It discusses EPA's use of risk characterizations in some detail. The EPA analysis is valuable here:

---

[1] See <http://www.epa.gov/risk/health-risk.htm>. The EPA data in this instance help screen for polluted areas in the U.S. that may need additional study and vetting for potential human health problems.

[2] David Pace, *More Blacks Live With Pollution*, Associated Press (Dec. 13, 2005), <http://onlinenews.ap.org/work/pollution/wrap.py?story=./linked_story/part1.html>. See also http://www.nbcnews.com/id/10452037/ns/us_news-environment/t/minorities-suffer-most-industrial-pollution/>. The EPA uses toxic chemical air releases reported by factories to calculate risk for each square kilometer of the United States. The scores allow comparing risks from long-term exposure to factory pollution from one area to another. The scores are based on: the amount of toxic pollution released by each factory; the path the pollution takes through the air; the level of danger to humans posed by each different chemical released; and the number and ages of males and females living in the exposure paths.

> "Risk characterizations should clearly highlight both the confidence and the uncertainty associated with the risk assessment. For example, numerical risk estimates should always be accompanied by descriptive information carefully selected to ensure an objective and balanced characterization of risk in risk assessment reports and regulatory documents."[3]

The EPA also created excellent documentation on how the analysis of its own data is to be used.[4] The documentation is for its own researchers, but its quality suggests broader applications are appropriate.

It stated, in part:

> "The methods used for the analysis (including all models used, all data upon which the assessment is based, and all assumptions that have a significant impact upon the results) are to be documented and easily located in the report. This documentation is to include a discussion of the degree to which the data used are representative of the population under study. Also, this documentation is to include the names of the models and software used to generate the analysis. Sufficient information is to be provided to allow the results of the analysis to be independently reproduced."[5]

These recommendations should also apply to large data sets applicable to other areas impacting consumers. Usage guidelines like EPA's, plus guidelines which discuss identifiability of consumers, create important fairness benchmarks for many of the uses and applications of large datasets. These benchmarks would go toward improving privacy protections for other big data activities.

Handling problems arising from analysis -- vulnerable populations

When problems are uncovered using big data analysis, careful application of the information is necessary. For example, policies that would mandate identifying and protecting victims of abuse, or other crimes, could have an unfortunate reverse effect. No one wants to create a readily accessible list of identifiable or semi-identifiable victims of abuse, while at the same time, the promise of a proper analysis to pinpoint aid distribution and assistance in a timely way to those who need it most would be welcome. The tension here is real, and we have to acknowledge it and resolve it in a balanced way.

We suspect that different applications of large datasets to different populations will warrant slightly different approaches. Again, we are most concerned about privacy-related challenges in the use of big data when the data sets can be re-identified back to

---

[3] U.S. Environmental Protection Agency, Science Policy Council, Risk Characterization Handbook (. December 2000), <http://www.epa.gov/spc/pdfs/rchandbk.pdf >.p. A5.
[4] U.S. Environmental Protection Agency, *Policy for Use of Probabilistic Analysis in Risk Assessment*, (May 15, 19970, <http://www.epa.gov/spc/pdfs/probpol.pdf>.
[5] Id, p. 2.

specific vulnerable consumer groups, or when the data sets are sensitive and are, or can become, personally identifiable to individuals.

Research is needed to understand how vulnerable populations in particular are affected by analysis and predictions based on such data, and what systematic biases could be potentially introduced into algorithms through faulty data and assumptions. In some cases, even loosely aggregate data has proven problematic.

In working to ensure beneficial uses of large datasets in vulnerable or sensitive areas while mitigating potential harm, we share several thoughts.

Of assistance in determining large dataset policy in identifiable datasets is the Common Rule[6] for protection of human subjects of research, and the Belmont Report[7] regarding Ethical Principles and Guidelines for the Protection of Human Subjects of Research. Informational risks in research must be measured against a firm standard, one that is not affected by every change in technology or commercial practice. For example, the HIPAA privacy standard establishes a firm set of Fair Information Practices. While there is considerable flexibility in the application of the HIPAA privacy rules in some contexts, the standards themselves are not subject to change because of external factors. Patients can expect the HIPAA standards to protect their health information in the same way.

The same should be true for human subjects research, which despite the size of a large dataset containing identifiable individuals, is still research and analysis applicable to individuals. The need for a baseline of privacy protection must be a constant for research even though the degree of informational risk can vary from project to project. The need for rules governing collection, use, and disclosure is constant. The need for openness and accountability is a constant. The need to consider individual participation rights (access and correction) is a constant. Thus, whatever the risk involved in a given project, the need for sufficient privacy protections for personally identifiable information is a constant.

Looking at this issue of identifiable data with more specificity would include for example, ensuring that recourse for discovery of accuracy-related problems is built in to the process. We are interested in policies that develop overall good practices in this area. Accuracy and recourse for correction for individuals identified in health care datasets is a foundational area for further inquiry. Some big data activities have been a part of health and other research for a long time, and there is nothing new in some respects. The demands of researchers can overwhelm existing institutions (like institutional review boards) that do not have the necessary privacy or security expertise.

We support the use of large datasets in medical research, but researcher obligations to protect data and to protect vulnerable populations from problems resulting from analysis need to be defined in law. Any disclosure for health research in large datasets should be

[6] < 45 CFR part 46, Subpart A-D. <http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html>.
[7] <http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html>.

limited by law, regulation, and contract as appropriate.[8] HIPAA requirements that protect health information when held by providers and insurers may not apply to researchers.

These research principles need to be applied to other vulnerable populations undergoing large dataset analysis. For example, financially vulnerable populations are another group deserving of more attention. Aggregate credit scores applied to neighborhoods (versus individuals) are an example of how aggregate but specific predictions based on large datasets may be lead to potentially unfair practices based on primarily geographical factors. If the results of the analysis are not managed correctly or exposed to consumers, errors in prediction may never surface, and other usage issues can arise.

Other examples of important vulnerable populations exist, we do not attempt to be comprehensive here. The overall impetus of the policy guidance should be to identify potential risks for specific populations and in sensitive data, and plan for recourse and checks and balances to mitigate harm or abuse and encourage the best possible uses and results.

Suggestions for Research

We would like the outcome of increased big data adoption to be better insight and more innovation, with adequate and robust protection for vulnerable populations. To do this, we need a significant study of large datasets that focuses on understanding how they are affecting vulnerable populations. This is an under-researched area. The questions we do not have adequate answers for yet include:

- How is big data affecting vulnerable populations?
- What risks are associated with big data and vulnerable populations?
- Which are the vulnerable populations most at risk?
- What sources of data are most problematic for vulnerable populations?
- For these sources of data, what safeguards are in place to insure data quality, and allow for discovery and corrections of inaccuracies?
- What populations are most at risk with current practices?
- What ways has big data been used to assist the protection of vulnerable populations?

We look forward to continuing to work in this key area of privacy. We welcome feedback you may have, and we would be happy to provide answers to any questions you may

[8] See, e.g., Robert Gellman, *The Deidentification Dilemma: A Legislative and Contractual Proposal*, 21 Fordham Intellectual Property, Media & Entertainment Law Journal 33 (2010), http://bobgellman.com/rg-docs/RG-Fordham-ID-10.pdf.

have.

Respectfully submitted,

Pam Dixon, Executive Director
World Privacy Forum
www.worldprivacyforum.org

[redacted]

From: Steve Wilson <Steve@constellationr.com>
Sent: Wednesday, April 02, 2014 3:08 PM
To: bigdata@ostp.gov
Cc: R "Ray" Wang
Subject: [Big Data RFI] Constellation Research submission
Attachments: Constellation Research BIG DATA PRIVACY White House Submission 1.4a.pdf

Dear Nicole Wong,

Please find attached a submission from Constellation Research on Big Data privacy.

We are delighted with the outreach and engagement initiated by the White House in this vital area, and we have enjoyed participating (albeit remotely) in the recent #BigDataPrivacy seminars at MIT and Berkeley. These were very high quality events indeed; they obviously surfaced important issues, and appear to have harnessed significant energies from industry, academia, the law, the privacy profession and government.

Constellation Research is an independent research and analysis firm, based in the San Francisco Bay area, focused on the impact of disruptive technologies on business and government. We are particularly interested in Big Data and associated technologies such as the Internet of Things, wearable computing, and ubiquitous computing, all of which provide ever more information to fuel the Big Data revolution.

Our research shows that despite a certain cynicism amongst technologists about the law, existing data privacy norms and regulations do in fact accommodate many Big Data features and can serve to control many of Big Data's potential excesses. On the other hand, there are dynamic Big Data processes that challenge the generally static, document-based controls in conventional privacy regulations. Our research therefore points to a new blended approach to Big Data privacy. We have identified the need for a new compact between businesses and individuals; the compact is the subject of ongoing research by Constellation, and an outline is included in our submission attached.

The attached paper addresses RFI question (5): What issues are raised by the use of big data across jurisdictions, such as the adequacy of current international laws, regulations, or norms?

We trust the submission is a useful contribution to you work.

We will continue to watch the OSTP program closely, and would be happy to hold deeper discussions with the Office at any time.

Sincerely,

Steve Wilson.

Steve Wilson | VP & Principal Analyst | Constellation Research, Inc.
[redacted]

Government Submission

Meeting the Privacy Challenges of Big Data

Privacy Regulations can Shock Data Miners, yet Big Data demands a New Privacy Compact

Steve Wilson

Vice President and Principal Analyst

April 2, 2014

constellation RESEARCH

Submission to the White House OSTP

Constellation Research welcomes the opportunity to make this submission to the White House Office of Science and Technology policy (OSTP) in response to the Request for Information seeking public comment on the ways in which big data may impact privacy, the economy, and public policy.

Executive Summary

Big Data – the extraction of knowledge and insights from the vast underground rivers of unstructured data that course unseen through cyberspace – represents one of the biggest challenges to privacy and data protection society has yet seen. Never before has so much personal information been available so freely to so many.

Personally Identifiable Information (PII) is the lifeblood of most digital enterprises today. Many social media business models are fueled by a generally one-sided bargain for PII, and the fairness of this value exchange is currently a hot topic. Data analytics and data mining are able to pull PII almost out of thin air. And so department stores now have the power to predict when female customers have become pregnant, and social network operators can tell who their members are hanging out with, through facial recognition. Collectively, digital businesses may have gone too far in their enthusiasm for Big Data, sacrificing the trust of their users for short term commercial gain.

Big Data promises vast genuine benefits for a great many stakeholders. For example, town planners can extract detailed patterns in the way people and goods move about our urban environments; engineers can better predict wear and tear on infrastructure; retailers can detect how people actually behave in stores without relying on 'likes' and complaints; police can pick out suspicious communications suggestive of criminal preparations; medical professionals and epidemiologists can discover deep patterns in health and lifestyle data to help better manage the entire population's future well-being.

However the vital society-wide Big Data project is threatened by the excesses of a few. Many cavalier online businesses are propelled by a naive assumption that data in the "public domain" is up for grabs. Technocrats may think the law has not kept pace with developments, but they are often caught out by conventional data protection regulations. The extraction of PII from raw data may be interpreted as a *collection* and as such is subject to long standing data protection statutes. On the other hand, orthodox privacy policies and user agreements do not cater for the way PII can be conjured tomorrow from raw data collected today. So privacy compliance efforts need to move beyond the current preoccupation with unwieldy policy documents and simplistic notices about cookies.

Thus the fit between Big Data and data privacy standards is complex and sometimes surprising. While existing laws are not to be underestimated, Constellation calls for a fresh compact with users that engages them in the far-reaching upside of Big Data. Organizations need to work out new privacy controls and user interfaces for opting in and out and in again, as they see fit, while the Big Data value proposition continues to evolve.

We call this new compact Big Privacy.

The Big Business of Big Data

It's not for nothing people call it "data mining". The raw material of Big Data – namely all the ones and zeroes coursing beneath us in the digital environment – is often likened to crude oil, alluding to the enormous riches to be extracted from an undifferentiated matrix.

Consider photo data and the rapid evolution of tools for monetizing it. These tools range from simple metadata embedded in digital photos which record when, where and with what sort of device they were taken, through to increasingly sophisticated pattern recognition and facial recognition algorithms. Image analysis can extract places and product names from photos, and automatically pick out objects. It can identity faces by re-purposing biometric templates that originate from social network users tagging their friends for fun in entirely unrelated images. Image analysis lets social media companies work out what people are doing, when and where, and who they're doing it with, thus revealing personal preferences and relationships, without anyone explicitly "liking" anything or "friending" anyone.

The ability to mine photo data defines a new digital gold rush. Like petroleum engineering, image analysis is very high tech. There is extraordinary R&D going on in face and object recognition. Companies have invested enormously in their own R&D and in acquiring start-ups, often paying over-the-odds for photo companies like Picasa, Instagram and Snapchat[1]: not merely because photos are fun and tagging them is cool, but because the potential for extracting intelligence from images is unbounded.

So more than data mining, Big Data is really about data *refining*.

Business models for monetizing photo data are still embryonic. Some entrepreneurs are beginning to access photo data from online social networks. For example "Facedeals", a proof of concept from advertising invention lab Redpepper, provides automated check-in to retail stores by face recognition; the initial registration process draws on images and other profile information made available by Facebook (with the member's consent) over a public API (see http://redpepperland.com/lab/details/check-in-with-your-face). It is not clear if Facedeals accesses the biometric templates, but nothing in Facebook's privacy and data use policies restrains them from providing or selling the templates. But as we shall see, international privacy regulations do in fact restrict the uses that can be made of the by-products of Big Data, should they be personal. Facebook has been taken to task for stretching social data analytics beyond what members reasonably expected to occur,[2] and we believe more surprises await for digital businesses in retail, healthcare and other industries.

Big Data Cannot Ignore Privacy Law

It's often said that 'technology has outpaced privacy law', yet by and large Constellation suggests that view is overly cynical, and is belied by several notable enforcement cases internationally. Technology has certainly outpaced the intuitions of consumers, who may not grasp how IT works, and are increasingly alarmed at what Big Data can reveal about them, when it is drawn to attention. However OECD data privacy principles set down in 1980 still work well, despite predating the World Wide Web by decades. Enforcement of

privacy laws is gaining momentum everywhere. Outside the US, rights-based privacy law has proven effective against many of today's more worrying business practices. Digital entrepreneurs can feel entitled to make any use they like of data that comes their way, but in truth thirty year old privacy law says otherwise. Information innovators ignore international privacy law at their peril. In this section we will see why, by reviewing the surprising definition of Personal Information, and how good old technology neutral privacy principles are as relevant as ever.

Classical Data Privacy Controls

The OECD's Privacy Principles were handed down in 1980 to deal with the emerging threats of computerization (see http://oecdprivacy.org). For at least a decade before that, the burgeoning databases of governments, police forces and insurance companies were seen as a danger to civil liberties; the cover of Newsweek magazine on July 27, 1970 screamed "IS PRIVACY DEAD?" Since that time, over 100 countries world-wide have legislated data privacy protections based on the OECD principles or the more advanced European Union Privacy Directive.[3]

Despite this strong global trend, American legislators have mostly declined to enact broad-based privacy law, although particular sectors, such as American healthcare, feature some of the strictest data protection rules anywhere the in world. The Fair Information Practice Principles (FIPPs) – which reflect most of the OECD principles though crucially not all – have been adopted only sporadically.

Conventional "rights based" privacy principles are relatively straightforward. They neatly side-step philosophical complexities like the "self" and data ownership, and instead essentially require data custodians to be restrained, careful and transparent in how all PII is handled.

In the context of Big Data, two of the standard international privacy principles stand out:

- **The Collection Limitation Principle** requires that organizations collect only the PII they need for legitimate and transparent purposes. We like to stress that Collection Limitation is about discipline rather than prohibition. Privacy does not stop businesses collecting the PII they truly need; rather it requires that businesses justify what they collect.

- **The Openness Principle** requires that data custodians set out for all to see *what* PII they collect, *why* they need to collect it, *how*, *when* and *where* they collect it, and *who* else the PII may be shared with.

Neither of these principles is strongly reflected in the FIPPs.

Data privacy principles generally apply to all PII, whether it is collected directly (by form or questionnaire for example) or indirectly through data mining. And so we come to the two-fold challenge in Big Data privacy: open ended Big Data can lead to brand new discoveries, which cannot be fully envisaged and outlined at the time raw data was collected. Do we want the potential benefits of Big Data to be inhibited by the finality of the Collection Limitation? If there are shared benefits in the possibility of new PII being

uncovered in raw data, how can the privacy promises of Collection Limitation and Openness be honestly kept? Constellation believes the answer lies in organizations keeping up an open dialog with their users and customers, instead of trying to freeze a privacy understanding at the time raw data is gathered.

Big Data "Oil Spills"

On several occasions, data analytics and other "innovative" information business practices have led to major privacy breaches and non-compliance actions, in ways that have surprised the practitioners. And further surprises are in store for companies that do not grasp the meaning of PII and international privacy law.

Google StreetView Wi-Fi Collection

While they drive around photographing towns and cities, Google's StreetView cars listen out for Wi-Fi hubs and collect the geographical coordinates of any transmitters they find. Google collects Wi-Fi landmarks for its geo-location database, as used by maps and all sorts of services. In 2010 it was found that some StreetView software was also inadvertently collecting Wi-Fi network traffic, some of which contained PII like user names, banking details and even passwords. Privacy Commissioners in Australia, Japan, Korea, the Netherlands and elsewhere found Google was in breach of respective data protection law. Google apologized and destroyed all the Wi-Fi traffic that had been gathered.

The nature of the privacy offence confused some commentators and technologists who argued that Wi-Fi data in the public domain is not private, and could not be private. Therefore some believed Google was within its rights to do whatever it liked with such data. But that reasoning fails to grasp the technicality that Data Protection laws in Europe, Australia and elsewhere do not essentially distinguish "public" from "private". In fact the word "private" doesn't even appear in Australia's "Privacy Act". If data is identifiable, then various privacy rights generally attach to it irrespective of how it is collected.

Facebook Photo Tagging

Photo tagging helps users of photo sharing services organize their albums. Facebook creates biometric templates that mathematically represent facial features, allowing other photos to be identified. When Facebook makes automatic "Tag Suggestions", its facial recognition algorithms run in the background over all photo albums making putative matches; when a photo thus identified is next displayed to a member, the tag suggestion is displayed and the member invited to confirm it. According to the definition of Personally Identifiable Information, when Facebook's software adds a name to a hitherto anonymous photo record, it turns that record into PII; the tagging process therefore *collects* PII.

European privacy regulators found in 2012 that collecting biometric data in this way without consent was a serious privacy breach. They forced Facebook to shut down facial recognition and tag suggestions for all its EU operations and delete all biometric data collected to that time.

So even if data miners synthesize PII using sophisticated data processing algorithms, they are subject to Privacy Principles, such as Openness, Collection and Use Limitation. Crucially, until 2012, Facebook's Privacy Policy and Data Usage Policy had not even mentioned that biometric facial recognition templates were created during tagging, let alone that they were subsequently used to automatically identify people in other photos.

More Privacy Shocks are Likely to Come

Digital businesses are availing themselves of an ever richer array of signals created automatically as we go about our lives online. We should remember that instrumenting their customers is a core aspect of new augmented reality technologies. Worryingly, the Privacy Policies of many technology businesses tend to be silent on what they plan to do with the by-products of Natural Language Processing, face recognition, object recognition and similar "digital bread crumbs". At the same time, informaticians are discovering ever more clever ways to de-anonymise us in cyberspace. In one of the more spectacular recent examples, self-described "DNA hackers" at MIT's Whitehead Institute for Biomedical Research in 2012 worked out how to leverage publicly available genealogical data to decipher the names of anonymous DNA donors in the Thousand Genomes research program.[4]

Do these developments mean "privacy is dead" after all? No. The fact is that anonymity is threatened by information technologies, but anonymity (or secrecy) is not the same thing as privacy. In many parts of the world, undoing anonymity represents an act of PII collection and as such is regulated. If honest Big Data businesses are alert to the broad definition of PII, and to the technology neutrality of data privacy regulations, then they can reduce the chances of more shocks to come.

"Big Privacy" to deal with Big Data

Big Data represents a deep qualitative shift to how business is done in the digital economy rather than just a quantitative change. It demands some new ways of doing privacy, and updating long standing principles, rather than just throwing more privacy controls at the problem. As we've seen, to many technologists' evident surprise, principles-based data protection laws have proven very powerful, to constrain Big Data processes, even though scenarios like facial recognition in social networks could not have been envisaged 30 years ago when the OECD privacy principles were formulated.

And yet orthodox data privacy principles do struggle to cope with the open-endedness of Big Data. Orthodox privacy management entails telling individuals what information is collected about them and when it is collected. But with Big Data, even if a company wants to be completely transparent, it may not be able to say today what PII it's going to find tomorrow. Any promise of Openness with Big Data cannot be made once and forgotten; it needs to be continually revisited.

There is a bargain at the heart of most social media business today, in which personal data is traded for a rich array of free services. Most sophisticated Internet users know 'there is no such thing as a free lunch' and that the things they now take for granted online, like search, maps, cloud mail and blogging, are paid for through the monetization of PII. Now there is nothing intrinsically wrong with business models that extract valuable

PII from anonymous raw data. But the main privacy problem is transparency. Today's service-data bargain is almost always opaque; personal data is harvested unseen; the information businesses are coy even about how valuable PII is.

The fact that there is a real price to pay, one way or another, for things like Online Social Networking has led some privacy advocates to call for an overt user-pays model. They say it would be fairer than making money from users without telling them. That might well be an option for some, but we believe the minimal privacy requirement is more about transparency. Social Network members deserve to be told about the PII trade (including details of what information-based businesses do with all this PII) so they can make up their own minds about it.

A New "Big Privacy" Compact

Constellation calls for a fresh compact between businesses and customers, which acknowledges the central importance of personal data, and honors the right of individuals to participate in how PII about them is exploited. The elements of a new Big Privacy compact should include:

- Respect and Restraint
- Transparency not only about PII but also about information-based business models
- Fair transparent deals for PII
- Innovation in Privacy which honors privacy principles.

Endnotes

[1] Google acquired the cloud photo storage service Picasa in 2004 for an unknown price; Facebook bought photo sharing network Instagram in 2012 for approximately $1B; in late 2013, instant photo messaging service Snapchat turned down an offer from Facebook for approximately $3B.

[2] See "Facebook's challenge to the Collection Limitation Principle", S. Wilson and A. Johnstone, in Encyclopedia of Social Network Analysis and Mining (forthcoming); extract at constellationr.com/content/facebooks-challenge-collection-limitation-principle.

[3] "The Influence of European Data Privacy Standards Outside Europe: Implications for Globalisation of Convention 108", Graham Greenleaf, University of New South Wales, Faculty of Law, International Data Privacy Law, Vol. 2, Issue 2, 2012.

[4] See "Genealogy databases enable naming of anonymous DNA donor" J. Bohannon, Science, 18 January 2013, discussed at lockstep.com.au/blog/2013/02/08/dna-privacy-letter-to-science.

About Steve Wilson

Steve Wilson is Vice President and Principal Analyst at Constellation Research, Inc. He focuses on digital identity, privacy and cyber security across the business research themes of Consumerization of IT and Next-Generation Customer Experience.

Steve has worked in ICT innovation, research, development and analysis for over 25 years, of which the last 19 years has been dedicated to digital identity and privacy. He has been awarded nine patents for identity management innovations and privacy enhancing technologies. Steve has long been involved in security public policy and industry development, including as member of the Australian Law Reform Commission's Developing Technology committee (2007-08), the Federal Privacy Commissioner's PKI Reference Group (2000) and the National E-Authentication Council (1998-2001). He contributed to the American Bar Association *PKI Assessment Guidelines* (1999-2002) and was co-author of the APEC *Electronic Authentication* guidelines (1998-2001). Steve chaired the Certification Forum of Australasia over 1999-2002 and the OASIS PKI Committee from 2007 to 2008. He is a current member of the International Association of Privacy Professionals, the Australian Government *Gatekeeper* PKI Advisory Committee, and the Privacy Coordination Committee of the National Strategy for Trusted Identities in Cyberspace (NSTIC). He has provided advice on identity frameworks to the governments of Australia, Hong Kong, New Zealand, Malaysia, Singapore, Kazakhstan and Macau. Contact Steve at steve@constellationr.com.

About Constellation Research

Constellation Research is a research and advisory firm that helps organizations navigate the challenges of digital disruption through business models transformation and the judicious application of disruptive technologies. This renowned group of experienced analysts, led by R "Ray" Wang, focuses on business-themed research, including Digital Marketing Transformation; Future of Work; Next-Generation Customer Experience; Data to Decisions; Matrix Commerce; Technology Optimization and Innovation; and Consumerization of IT and the New C-Suite.

- Founded and headquartered in the San Francisco Bay Area in 2010.
- Named New Analyst Firm of the Year in 2011.
- Serving over 225 buy-side and sell-side clients around the globe.
- Experienced research team with an average of 21 years of practitioner, management and industry experience.
- Creators of the Constellation Supernova Awards – the industry's first and largest recognition of innovators, pioneers and teams who apply emerging and disruptive technology to drive business value.
- Organizers of the Constellation Connected Enterprise – an innovation summit and best practices knowledge-sharing retreat for business leaders.
- Founders of Constellation Academy, experiential workshops in applying disruptive technology to disruptive business models.

San Francisco | Andalucia | Austin | Belfast | Boston | Chicago | Colorado Springs | Denver | London | Los Angeles | Monta Vista | New York | Pune
Sacramento | San Diego | Santa Monica | Sedona | Sydney | Tokyo | Toronto | Washington D.C.

Governance of Big Data

Fred H. Cate[1]
Peter Cullen[2]
Viktor Mayer-Schönberger[3]

April 4, 2014

Summary

In an age of big data, privacy is more essential than ever, but if we are to protect it effectively, while continuing to enjoy the benefits that big data is already making possible, we need to evolve better, faster, and more scalable governance mechanisms to achieve more effective data protection.

Over the past three years, we have led a multinational research initiative addressing the information privacy and security challenges presented by big data and the technologies that facilitate it. This work has included nine workshops; involving leading regulators, industry executives, public interest advocates, and academic experts from 19 countries on five continents; and three published reports in addition to numerous articles and blog postings.

This research suggests that four elements are essential to protecting privacy while unlocking the potential of big data—namely, a *risk-management* approach to data protection that:

1. Places more responsibility for data stewardship, and liability for reasonably foreseeable harms, on the users of data rather than using notice and consent to shift the burden to individuals. At present, individuals forced to make choices at the time of data collection are saddled both with the burden of making those choices and with the consequences of those choices. The impact on individuals expands exponentially in a world of big data because of dramatic increases in the ubiquity of data collection, the volume and velocity of information flows, and the range of data users (and re-users). A continuing broad reliance on notice and choice under-protects privacy, can seriously interfere with—and raise the cost of—subsequent beneficial (re)uses of data, diminishes the effectiveness of these measures in settings where they could be used appropriately, and runs the risk of ignoring more effective tools—including transparency and redress—for strengthening the role of individuals in data protection.

2. Focuses more on *uses* of big data as opposed to the mere collection or retention of data or the purposes for which data were originally collected. There is often a compelling reason for personal data to be disclosed or collected. Assessing the risk to individuals posed by those data almost always requires knowing the context in which they will be used. Data used in one context or for one purpose or subject to one set of protections may be both beneficial and desirable, where exactly the same data used in a different context or for another purpose or without appropriate protections may be both dangerous and undesirable. A greater focus on risk assessment at the time of use treats data in context, protects individuals in those increasingly

[1] Distinguished Professor and C. Ben Dutton Professor of Law, Indiana University Maurer School of Law; director, Center for Applied Cybersecurity Research, Indiana University.

[2] General Manager, Trustworthy Computing, Microsoft Corporation.

[3] Professor of Internet Governance and Regulation, Oxford Internet Institute, University of Oxford; coauthor (with Kenneth Cukier) of *Big Data: A Revolution That Will Transform How We Live, Work, and Think* (2013).

frequent settings in which data are generated without direct contact with the individual (for example, collected by sensors or inferred from existing data), and recognizes that with the advent of big data personal data may have substantial valuable uses that were wholly unanticipated when they were collected.

3. Is guided by a broad framework of cognizable harms identified through a transparent, inclusive process including regulators, industry, academics, and individuals. The goal of a risk management approach focused on data uses is to reduce or eliminate the harm that personal information can cause to individuals. Accomplishing this, however, requires a clear understanding of what constitutes "harm" or other undesired impacts in the privacy context. A framework for recognized harms is critical to ensuring that individuals are protected and enhancing predictability, accountability, and efficiency. The goal should not be to mandate a one-size-fits-all approach to risk analysis, but rather to provide a useful, practical framework, to ensure that a wide range of interests and constituencies are involved in crafting it, and to highlight measures for reducing risk.

4. Provides meaningful transparency and redress to individuals. Effective transparency and redress protect the rights of individuals, but also serve the vital purposes of enhancing the accuracy and effectiveness of big data tools, and creating disincentives for deploying tools inappropriately. Meaningful transparency and redress, together with effective enforcement, not only provide remedies for current harms, but also help to prevent future ones.

Introduction

The growth of big data presents an array of governance issues many of the most important of which concern privacy and the impact of big data on individuals.[4] Some of these issues are familiar and serve to remind us of the shortcomings of our current governance approaches, especially our heavy reliance on notice and consent to protect privacy. Some of the issues are more novel, for example, the extent to which we may overvalue or inappropriately rely on predictions about individual behavior made with big data. All require effective governance mechanisms that take into account both the value of big data and data-based innovation and the risks that uses of big data pose.

Over the past three years, we have led a multinational research initiative addressing the information privacy and security challenges presented by big data and the data collection and generation technologies that facilitate it. This work has included regional privacy dialogues in Washington, Brussels, Singapore, Sydney, and São Paulo during the summer of 2012, featuring leading regulators, industry executives, public interest advocates, and academic experts; a global privacy summit in September 2012 at Microsoft's headquarters in Redmond, Washington, involving more than 70 privacy and data protection experts from 19 countries on five continents; three multinational

[4] We use "big data" to refer to collections of data that are not only large, but also increasingly complete and granular, and might be contrasted with data analysis that involves incomplete or sample data, or data that are available only in aggregated or more abstract forms. See Viktor Mayer-Schönberger & Kenneth Cukier, *Big Data: A Revolution That Will Transform How We Live, Work, and Think* (2013). A unified or linked group of datasets that list every vehicle owned by every individual in a given city would be a simple example of big data, and might be contrasted with broad inferences of vehicle types and values based on aggregated census data that provide average results based on post code.

workshops on revising the OECD privacy guidelines, data uses and impacts, and data risk management in 2013; and a series of related reports.[5]

This initiative, in addition to our individual research and experience relating to big data, suggests that effective governance of big data requires at least four core elements. There are likely others, but we believe these four are critical to protecting the privacy of individuals while unlocking the potential of big data.

1. A risk management approach that places more responsibility for data stewardship, and liability for reasonably foreseeable harms, on the users of data rather than using notice and consent to shift the burden to individuals

Most data protection laws in the United States and elsewhere effectively place some or all of the responsibility for protecting privacy on individual data subjects through the operation of notice and consent. This is evident in the Organisation for Economic Co-operation and Development (OECD) Guidelines on the Protection of Privacy and Transborder Flows of Personal Data adopted in 1980,[6] which require that personal information be collected, "where appropriate, with the knowledge or consent of the data subject," and which prohibit the reuse of personal information except: "(a) with the consent of the data subject; or (b) by the authority of law."[7] In short, if an individual can be persuaded to consent, then the subsequent use is often permissible under laws adopted in conformance with the OECD Guidelines.

This focus on the role of the individual is especially evident in the United States. In 1998, for example, the U.S. Federal Trade Commission (FTC), after reviewing the "fair information practice codes" of the United States, Canada, and Europe, reported to Congress that "[t]he most fundamental principle is notice. . . . [because] [w]ithout notice, a consumer cannot make an informed decision as to whether and to what extent to disclose personal information."[8] The FTC continued, "[t]he second widely-accepted core principle of fair information practice is consumer choice or consent . . . . [over] how any personal information collected from them may be used."[9]

U.S. statutes and regulations have tended to parallel the FTC's emphasis on notice and choice. The Obama Administration's 2012 Consumer Privacy Bill of Rights includes as its first principle: "Consumers have a right to exercise control over what personal data companies collect from them and how they use it."[10] But while presented as a right, the emphasis on notice and consent in reality often

---

[5] See, e.g., Fred H. Cate & Viktor Mayer-Schönberger, *Data Use and Impact Global Workshop*, Center for Applied Cybersecurity Research (2013); Fred H. Cate, Peter Cullen & Viktor Mayer-Schönberger, *Data Protection Principles for the 21st Century*, Oxford Internet Institute (2013); Fred H. Cate & Viktor Mayer-Schönberger, *Notice and Consent in a World of Big* Data, Microsoft Corporation (2012); Fred H. Cate & Viktor Mayer-Schönberger, "Notice and Consent in a World of Big Data," *International Data Privacy Law*, vol. 3, no. 2 at 67 (2013).

[6] O.E.C.D. Doc. (C 58 final) (Oct. 1, 1980), http://www.oecd.org/document/20/0,3343,en_2649_34255_15589524_1_1_1_1,00.html.

[7] *Id*. at ¶¶ 7, 10.

[8] U.S. Federal Trade Commission, *Privacy Online: A Report to Congress* 7 (1998), http://www.ftc.gov/reports/privacy3/priv-23a.pdf.

[9] *Id.* at 8 (citations omitted).

[10] The White House, *Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation* 47 (2012), http://www.whitehouse.gov/sites/default/files/privacy-final.pdf.

ends up creating a duty on individuals to make choices they are often ill-prepared to make and to accept the consequences of those choices.[11]

While the critique of notice and choice as primary mechanisms for data protection originated with academics—in 1999 Professor Paul Schwartz argued that "social and legal norms about privacy promise too much, namely data control, and deliver too little"[12]—it has more recently been echoed by regulators, industry, and privacy advocates. For example, the FTC noted the dangers of over-reliance on notice and choice in its staff and commission reports (issued in 2010 and 2012, respectively) on the future of privacy protection.[13] In its 2010 staff report, for example, the Commission wrote:

> In recent years, the limitations of the notice-and-choice model have become increasingly apparent….. [C]onsumers face a substantial burden in reading and understanding privacy policies and exercising the limited choices offered to them….. Additionally, the emphasis on notice and choice alone has not sufficiently accounted for other widely recognized fair information practices, such as access, collection limitation, purpose specification, and assuring data quality and integrity.[14]

On December 1, 2009, the EU Article 29 Data Protection Working Party and the Working Party on Police and Justice adopted a "joint contribution" on the future of privacy, in which they argued that "consent is an inappropriate ground for processing" in the "many cases in which consent cannot be given freely, especially when there is a clear unbalance between the data subject and the data controller (for example in the employment context or when personal data must be provided to public authorities)."[15]

---

[11] The growing focus on notice and consent at the time of collection is not limited to the United States. The European Union's Data Protection Directive, for example, is significantly focused on individual choice. Article 7 of the directive provides seven conditions under which personal data may be processed. The first is "the data subject has unambiguously given his consent." *Directive 95/46/EC of the European Parliament and of the Council on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data* (Eur. O.J. 95/L281), Preamble, art. 7(a), http://ec.europa.eu/justice_home/fsj/privacy/docs/95-46-ce/dir1995-46_part1_en.pdf and http://ec.europa.eu/justice_home/fsj/privacy/docs/95-46-ce/dir1995-46_part2_en.pdf. Article 8 restricts the processing of sensitive data, but then provides that the restriction shall not apply where "the data subject has given his explicit consent to the processing of those data." *Id.*, art. 8(2)(a). Article 26 identifies six exceptions to the provision prohibiting the export of personal data to non-European countries lacking "adequate" data protection. The first is that "the data subject has given his consent unambiguously to the proposed transfer." *Id.*, art. 26(1)(a).

The Asia-Pacific Economic Cooperation (APEC) Privacy Framework, adopted in 2004, is similarly focused on notice and choice: "Where appropriate, individuals should be provided with clear, prominent, easily understandable, accessible and affordable mechanisms to exercise choice in relation to the collection, use and disclosure of their personal information." Asia-Pacific Economic Cooperation*, APEC Privacy Framework*, 2004/AMM/014rev1 (2005), at 17.

In Canada, Philippa Lawson and Mary O'Donoghue have written that "[t]he requirement for consent to the collection, use, and disclosure of personal information is a cornerstone of all three common-law regimes." Philippa Lawson & Mary O'Donoghue, "Approaches to Consent in Canadian Data Protection Law," in Ian Kerr, Valerie Steeves & Carole Lucock, eds., *Lessons from the Identity Trail: Anonymity, Privacy and Identity in a Networked Society* 23 (2009), http://www.idtrail.org/content/view/799.

[12] Paul M. Schwartz, "Privacy and Democracy in Cyberspace," 52 *Vanderbilt Law Review* 1607, 1657 (1999).

[13] U.S. Federal Trade Commission, *Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Businesses and Policymakers*, FTC Report (2012), www.ftc.gov/os/2012/03/120326privacyreport.pdf; U.S. Federal Trade Commission, *Protecting Consumer Privacy in an Era of Rapid Change: A Proposed Framework for Businesses and Policymakers*, Preliminary FTC Staff Report (2010), www.ftc.gov/os/2010/12/101201privacyreport.pdf.

[14] Preliminary FTC Staff Report (2010), 19-20.

[15] Article 29 Data Protection Working Party and the Working Party on Police and Justice, *The Future of Privacy: Joint Contribution to the Consultation of the European Commission on the Legal Framework for the Fundamental Right to Protection*

At present, individuals faced with choices at the time of data collection are saddled both with the burden of making those choices and with the consequences of those choices, whether as a result of overly complex notices, limited choices, lack of understanding, or an inability to weigh future risks against present benefits.

The challenge is especially clear in the context of big data. In 1980, there was far less data collection and use and far fewer data collectors and users. Businesses and governments were also using personal data in more straightforward ways, often for a single, well-defined purpose, and not sharing it with numerous third parties and developing complex data sets. Under these circumstances, individuals were more likely to understand the purpose for which their data was being collected and used. And ultimately they could be held accountable for supplying informed consent when given adequate notice.

Today, with the proliferation of new information technologies, applications, and data uses, individual consent is rarely exercised as a meaningful choice. Individuals are overwhelmed with many long, complex online privacy policies just as they are attempting to access a desired resource. Scrolling through numerous policies and clicking "I agree" rarely provides meaningful privacy protection.

Even if individuals wished to read the endless privacy policies they encounter, actually doing so is not practical because of society's growing reliance on personal data. One 2008 study calculated that reading the privacy policies of just the most popular websites would take an individual 244 hours—or more than 30 full working days—each year.[16] In addition, a meaningful assessment of even a single privacy policy may require a sophisticated understanding of how data is used today. This would mean investing even more time to stay current on how data collection and data use are evolving.

In a world of big data, all of these issues are magnified. Individual choice is increasingly impractical and undesirable as the principal way to protect privacy. Dramatic increases in the ubiquity of data collection, the volume and velocity of information flows, and the range of data users (and re-users) place an untenable burden on individuals to understand the issues, make choices, and then engage in oversight and enforcement. A continuing broad reliance on notice and choice at time of collection both under-protects privacy and seriously interferes with—and raises the cost of—subsequent beneficial (re)uses of data.

Moreover, personal information is increasingly used by parties with no direct relationship to the individual, or generated by sensors (or inferred by third parties) over which the individual not merely exercises no control, but has no relationship. Consider that the *New York Times* reported in 2012 that one U.S. company that has no direct interaction with consumers engages in more than 50 *trillion* transactions involving recorded personal data every year.[17] More problematic still is the fact that notice and consent are used to shift responsibility to individuals and away from data users. If a subsequent use of data proves to be harmful or threatening to the individual, data users generally receive legal protection from the fact that the individual consented.

---

of Personal Data (02356/09/EN, WP 168) 17 (Dec. 1, 2009), http://ec.europa.eu/justice_home/fsj/privacy/docs/wpdocs/2009/wp168_en.pdf.

[16] Alecia M. McDonald and Lorrie Faith Cranor, "The Cost of Reading Privacy Policies," *I/S: A Journal of Law and Policy for the Information Society* (2008), http://moritzlaw.osu.edu/students/groups/is/files/2012/02/Cranor_Formatted_Final.pdf.

[17] Natasha Singer, "You for Sale: Mapping, and Sharing, the Consumer Genome," *N.Y. Times*, June 17, 2012, at BU1, http://www.nytimes.com/2012/06/17/technology/acxiom-the-quiet-giant-of-consumer-database-marketing.html?pagewanted=all&_r=0.

If there was any one point on which virtually all of the participants in the 2012 and 2013 workshops that we conducted agreed it was that data users should be accountable for this processing of personal information, and that effective governance of big data requires shifting more responsibility away from individuals and toward data collectors and data users, who should be held accountable for how they manage data rather than whether they obtain individual consent. To be certain, notice and consent may provide meaningful privacy protection in appropriate contexts, but this approach is increasingly ineffective as the primary mechanism for ensuring information privacy, especially so in the case of big data. Over-reliance on notice and consent both diminishes the effectiveness of these measures in settings where they could be used appropriately and often ignores more effective tools—including transparency and redress—for strengthening the role of individuals in data protection.

2. A risk management approach focused more on *uses* of big data as opposed to the mere collection or retention of data or the purposes for which data were originally collected

Another consistent theme of our research has been the importance of focusing more attention on the "use" of personal information rather than on its "collection," given the increasingly pervasive nature of data collection and surveillance, growing governmental demands on industry to collect personal data, and the development of valuable new uses for personal data. Focusing on the use of personal data does not replace responsibilities or regulation relating to data collection, nor should a focus on consent in specific or sensitive circumstances be abandoned. Rather, in many situations, a more practical, as well as sensitive, balancing of valuable data flows and more effective privacy protection is likely to be obtained by focusing more attention on appropriate, accountable use.

As Helen Nissenbaum and others have noted, the extent to which privacy is respected or invaded often depends on the context in which data is used, more than on the data themselves.[18] Data used in one context or for one purpose or subject to one set of protections may be both beneficial and desirable, where exactly the same data used in a different context or for another purpose or without appropriate protections may be both dangerous and undesirable. A greater focus on risk assessment at the time of use treats data in context and would lead to assessments by data users that would recognize that data collected in one context (for example, online marketing) might be insufficiently reliable or otherwise inappropriate to use in a different context (for example, determining eligibility for a job or loan).

Under a more use-based approach, data users would evaluate the appropriateness of an intended use of personal data not by focusing primarily on the terms under which the data were originally collected, but rather on the likely risks to, or impacts on, individuals (of benefits and harms) associated with the proposed use of the data. Such a focus on use is more intuitive because most individuals and institutions already think about uses when evaluating their comfort with proposed data processing activities. "What are you going to do with the data?" "How do you intend to use it?" "What are the benefits and risks of the proposed use?" These are the types of questions that many individuals ask—explicitly or implicitly—when they inquire about data processing activities. They can be answered only in connection with specific uses or categories of uses, and they are precisely the questions that data users would be required to ask—and answer—regarding proposed uses of data.

[18] Helen Nissenbaum, *Privacy in Context* (2010).

A greater focus on use places data protection in the broader context of consumer protection, and recognizes benefits as well as harms to individuals associated with data uses. For example, personal financial data may be used for fraud or other illegal purposes, which implicates laws far beyond data protection. The same information may be used to provide an individual with a product or service that may have great value for the individual, but would nonetheless require the security and accountability required by data protection regulations.

A greater focus on use would, in the words of one data protection official who participated in our workshops, "make privacy relevant again" and enhance privacy protection. It would also help enhance trust by making data protection more relevant and more predictable, reducing the burden on individuals, and creating disincentives for data uses likely to cause harm.

One of the most pronounced changes that will result from the evolution toward a greater focus on use is to diminish the role of the purpose for which data were originally collected. The OECD 1980 Guidelines explicitly provide for a Purpose Specification Principle:

> The purposes for which personal data are collected should be specified not later than at the time of data collection and the subsequent use limited to the fulfillment of those purposes or such others as are not incompatible with those purposes and as are specified on each occasion of change of purpose.

This principle is already problematic today for many reasons, including the fact that, precisely because of it, data processors usually specify exceptionally broad purposes that provide little meaningful limit on their subsequent use of data. In addition, increasingly data are generated in ways that involve no direct contact with the individual (for example, collected by sensors or inferred from existing data) so there is never a purpose specified. Moreover, with the advent of big data and the analytical tools that have accompanied it, personal data may have substantial valuable uses that were wholly unanticipated when the data were collected, yet the data were collected in such a way or are so vast as to make contacting each individual to obtain consent for the new use impractical, as well as potentially undesirable where the beneficial use depends on having a complete data set (as is often the case with medical research).[19]

Some modern data protection systems have dealt with these problems by creating broad exceptions to this principle, interpreting "not incompatible" so broadly as to undermine the principle, or simply ignoring it altogether. Taking maximum advantage of big data will require a more thoughtful approach to purpose specification. The principle will have less relevance in many settings. This will certainly be true when data are observed or inferred without any contact with the individual, but it will also likely be true in many other settings. Instead, it is the analysis of risks associated with an intended use that determines whether, and subject to what protections, a use is appropriate.

The terms under which data were collected would remain relevant when a specific purpose is provided at time of collection and is a meaningful factor in obtaining access to data. This would be especially clear in settings where users had made meaningful choices (for example, specifying a preferred medium for future communications) or where the data processor had agreed to specific limits as a condition of obtaining personal information (for example, an explicit promise not to share the data).

[19] Institute of Medicine, *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health through Research* 210 (2009).

But a greater focus on risk-assessment of specific uses, rather than focusing on consent, is essential to ensure that data users do not evade their commitments, that valuable uses of data are not inappropriately deterred, and that data protection laws are not claimed to reflect a principle that increasingly they do not.

One key reason why a more use-focused approach is necessary to capture the value of big data is that the analysis of big data doesn't always start with a question or hypothesis, but rather may reveal insights that were never anticipated.[20] As a result, data protection based on a notice specifying intended uses of data and consent for collection based on that notice can result in blocking socially valuable (re)uses of data, lead to meaninglessly broad notices, or require exceptions to the terms under which the individual consented. In each of these cases, a valuable use of data is restricted or privacy is compromised or both. If privacy protection is instead based on a risk analysis of a proposed use, then it is possible to achieve an optimum benefit from the use of the data and optimum protection for data fine-tuned for each intended use.

Achieving this optimum outcome is critical because significant unanticipated benefits may result from the sheer size and comprehensiveness of big data. There are many examples where the benefits of having appropriate access to big data for analysis can be life-saving, for example, when analysis of big data reveals harmful drug interactions or ineffective medical treatments.

At the same time, big data can lead to serious privacy incursions if used inappropriately or protected inadequately. Assessing the risk of harms and benefits posed by a specific use in light of the privacy protections in place is the best way to achieve the optimum outcome.

3. A risk management approach guided by a broad framework of cognizable harms identified through a transparent, inclusive process including regulators, industry, academics, and individuals

A critical component of the evolution toward a more use-focused data protection system is the development of a simple, transparent approach to assessing risks to individuals. This is not an easy task, but it is an essential one.

Measuring risks connected with data uses is especially challenging because of the intangible and subjective nature of many perceived harms. Any risk assessment must be both sufficiently broad to take into account the wide range of harms (and benefits), and sufficiently simple, so that it can be applied routinely and consistently. Perhaps most importantly, the assessment should be transparent to facilitate fairness, trust, and future refinement.

The goal of a risk management approach focused on data uses is to reduce or eliminate the harm that personal information can cause to individuals. Accomplishing this, however, requires a clear understanding of what constitutes "harm" or other undesired impact in the privacy context. Surprisingly, despite almost 50 years of experience with data protection regulation, that clear understanding is still lacking both in the scholarly literature and in the law. This is due in part to the focus on notice and consent in data protection, so that harm was considered collecting personal information without

[20] See Mayer-Schönberger & Cukier, supra.

providing proper notice or without obtaining consent, or using data outside of the scope of that consent.

That does not equate with the way most people think about data-related harms, which is more focused on data being used in a way that might cause them injury or embarrassment, rather than the presence or content of privacy notices. So there is a widespread need to think more critically about what constitutes a harm that the risk management framework should seek to minimize or prevent when evaluating data uses.

A framework for recognized harms is critical to ensuring that individuals are protected and enhancing predictability, accountability, and efficiency. Regulators such as the FTC are well placed to help lead a transparent, inclusive process to articulate that framework. The goal should not be to mandate a one-size-fits-all approach to risk analysis, but rather to provide a useful, practical reference point, and to ensure that a wide range of interests and constituencies are involved in crafting it.

There are a wide range of possibilities for what might constitute a harm, but it seems clear that the term must include not only a wide range of tangible injuries (including financial loss, physical threat or injury, unlawful discrimination, identity theft, loss of confidentiality, and other significant economic or social disadvantage), but also intangible harms (such as damage to reputation or goodwill, or excessive intrusion into private life) and potentially broader societal harms (such as contravention of national and multinational human rights instruments). What matters most, though, is that the meaning of harm be defined through a transparent, inclusive process, and with sufficient clarity to help guide the risk analyses of data users.

Risk assessment is not binary and is likely to be influenced by a number of factors within the data user's control. So the goal of the risk assessment isn't simply to indicate whether a proposed data use is likely to be appropriate or not, but also to highlight the steps that the data user can take to make that use more acceptable (for example, by truncating, encrypting, or de-personalizing data). Risk assessment is a process, not an end.

The ultimate goal of risk assessment, after taking into account those measures that the data user can take to reduce risk, is to create presumptions concerning common data uses so that both individuals and users can enjoy the benefits of predictability, consistency, and efficiency in data protection. So, for example, some uses in settings that present little likelihood of negligible harms occurring might be expressly permitted, especially if certain protections such as appropriate security were in place. Conversely, some uses in settings where there was a higher likelihood of more severe harms occurring might be prohibited or restricted without certain protections in place. For other uses that present either little risk of more severe harms or greater risk of less severe harms greater protections or even specific notice and/or consent might be required so that individuals have an opportunity to participate in the decision-making process.

4. Transparency and redress

Big data is increasingly being used to make decisions about individuals, and even to predict their future behavior, with often significant consequences. The one certainty of data analysis is that there will be errors—errors resulting from problems with data matching and linking, erroneous data, incomplete or inadequate algorithms, and misapplication of data-based tools. Moreover, our society often seems obsessed with data (whatever its size) and often acts as if we believe that just because something is

"data-based" it is not merely better, but infallible. As a result, we often deploy new data-based tools without adequate testing, oversight, or redress. (Data mining activities for aviation security are an obvious example, where the federal government denied or delayed boarding for thousands of innocent passengers for years before finally putting in place a redress system.)

Whenever big data is used in ways that affects individuals, there must be effective transparency and redress. This is necessary to protect the rights of individuals, but is also serves the vital purposes of enhancing the accuracy and effectiveness of big data tools, and creating disincentives for deploying tools inappropriately. Meaningful transparency and redress, together with effective enforcement, not only provide remedies for current harms, but also help to prevent future ones.

Moreover, while few individuals demonstrate much interest in inquiring into data processing activities until there is a perceived harm, when they are often more interested in learning how data was used and to what effect. Ensuring that there is meaningful redress will not only create disincentives for risky data processing and help repair the damage that such processing can cause, but it will also provide meaningful rights to individuals at the very time they are most interested in exercising them. It is an essential requirement for responsible use of big data.

Conclusion

In an age of big data, privacy is more essential than ever before, but if we are to protect it effectively, while continuing to enjoy the benefits that big data is already making possible, we need to evolve better, faster, and more scalable mechanisms. Our three-year, multinational research suggests that four elements are essential—namely, a *risk-management* approach to data protection that:

1. Place more responsibility for data stewardship, and liability for reasonably foreseeable harms, on the users of data rather than using notice and consent to shift the burden to individuals;

2. Focuses more on *uses* of big data as opposed to the mere collection or retention of data or the purposes for which data were originally collected;

3. Is guided by a broad framework of cognizable harms identified through a transparent, inclusive process including regulators, industry, academics, and individuals; and

4. Provides meaningful transparency and redress.

There are likely many other measures that also will be useful, but we believe these four are critical to protecting privacy while unlocking the potential of big data.

HEALTHCARE LEADERSHIP COUNCIL

April 4, 2014

Dear Sir or Madam:

The Healthcare Leadership Council (HLC) applauds the White House Office of Science and Technology Policy (OSTP) for taking on the challenge of comprehensively reviewing how data will impact American society. Undoubtedly, how data affects the health and healthcare of people is a critical topic to explore and we are encouraged that the White House is engaging in this effort.

HLC is a coalition of chief executives from all disciplines within American healthcare, and is the exclusive forum for the nation's healthcare leaders to jointly develop policies, plans, and programs to achieve their vision of a 21st century system that makes affordable, high-quality care accessible to all Americans. See attached list of HLC members.

HLC envisions a future in which public and private sector healthcare organizations securely share information in an efficient, effective manner that is accessible and useful for all stakeholders. Improved accessibility and quality of health data can accelerate progress in medical research, improve the quality of care delivery, reduce costs, and will lead to other benefits that we cannot yet imagine. With that in mind, HLC members have developed and endorsed a set of "HLC Principles on Data Policy" that we are pleased to share with OSTP as you examine the ways in which data affects the way we live and work, as well as the relationship between government and citizens. Working together, the public and private sectors can use data to spur innovation and maximize opportunities to share essential health information while minimizing the risks to privacy.

We would be pleased to work with OSTP staff as they engage in this important public policy effort. Please contact Tina Olson Grande, Senior Vice President for Policy, at tgrande@hlc.org if you would like to discuss HLC's work in data policy in further detail.

Sincerely,

Mary R. Grealy

President

Enclosure

HLC MEMBERS

2014

(Alphabetized by Company)

HLC Chairman
Greg Irace
President & CEO
Sanofi US

Mark Bertolini
Chair, President & CEO
Aetna

Todd Ebert
CEO
Amerinet

Steven Collis
President & CEO
AmerisourceBergen

Rolf Hoffmann
SVP, U.S. Commercial Operations
Amgen

Anthony Tersigni, EdD, FACHE
President & CEO
Ascension

Paul Hudson
Executive Vice President, North America
AstraZeneca

Joel Allison
CEO
Baylor Scott & White Health

Marc Grodman, M.D.
Chairman, President & CEO
Bio-Reference Laboratories, Inc.

William Gracey
President & CEO
BlueCross BlueShield of Tennessee

Greg Behar
President & CEO
Boehringer Ingelheim Pharmaceuticals

George Barrett
Chairman & CEO
Cardinal Health

Toby Cosgrove, M.D.
CEO & President
Cleveland Clinic Foundation

Tim Ring
Chairman & CEO
C. R. Bard

Michael A. Mussallem
Chairman & CEO
Edwards Lifesciences

Alex Azar
President, Lilly USA
Eli Lilly and Company

John Finan, Jr.
President & CEO
Franciscan Missionaries of Our Lady Health System, Inc.

Patricia Hemingway Hall
President & CEO
Health Care Service Corporation

Robert Mandel, M.D.
CEO
Health Dialog

Daniel Tassé
Chairman & CEO
Ikaria

Daniel Evans, Jr.
President & CEO
Indiana University Health

Paul Meister
Chairman & CEO
inVentiv Health

Jennifer Taubert
Company Group Chairman, North American Pharmaceuticals
Johnson & Johnson

Brian Ewert, M.D.
President
Marshfield Clinic

John Noseworthy, M.D.
President & CEO
Mayo Clinic

John Hammergren
Chairman & CEO
McKesson Corporation

Chris O'Connell
EVP & President, Restorative Therapies Group
Medtronic

Barry Arbuckle, Ph.D.
President & CEO
MemorialCare Health System

Robert McMahon
President, U.S. Market
Merck

Steven Corwin, M.D.
CEO
NewYork-Presbyterian Hospital

Mark Neaman
President & CEO
NorthShore University HealthSystem

Christi Shaw
EVP and Region Head, North America
Novartis

Jesper Hoiland
President
Novo Nordisk, Inc.

Craig Smith
President & CEO
Owens & Minor

Susan DeVore
President & CEO
Premier healthcare alliance

Chris Wing
President & CEO
SCAN Health Plan

Tim Scannell
Group President, MedSurg & Neurotechnology
Stryker

Paul Uhrig
Acting CEO
Surescripts

Doug Cole
President
Takeda Pharmaceuticals U.S.A.

Douglas Hawthorne, FACHE
CEO
Texas Health Resources

Frank Tarallo
CEO
Theragenics

Curt Nonomaque
President & CEO
VHA Inc.

Gregory Wasson
President & CEO
Walgreens

James Chambers
President & CEO
Weight Watchers International

Jaideep Bajaj
Chairman
ZS Associates

HEALTHCARE LEADERSHIP COUNCIL

HLC Principles on Data Policy

HLC envisions a future in which public and private sector healthcare organizations securely share information in an efficient, effective manner that is accessible and useful for all stakeholders. HLC members have already proven that they can harness data to improve care and value in healthcare. Improved accessibility and quality of health data can accelerate progress in medicines, improve the quality of care delivery, reduce costs, and will lead to other benefits that we cannot yet imagine.

Access to Data

- **As taxpayer-funded entities, it is the responsibility of government health agencies to maximize public benefit from data collected through their operations.** We applaud current work by HHS to reduce the time lag and improve compatibility of data released by the agency, but there is still significant room for improvement. By allowing regular access to data at minimal cost to organizations that are subject to consumer protection laws, organizations throughout the country can develop novel ways to fight disease, improve the quality of care, reduce costs, and accelerate innovation. Increased coordination among federal government agencies to reduce data "silos" and support for crossagency data access by private sector organizations will allow innovative new research that benefits consumers.

- **Timeliness, format, and regulatory flexibility are critical for organizations serving consumers to make the most of data held by the federal government's health programs.** Federal "data use agreement" restrictions keep many healthcare organizations from gaining access to data that would allow them to improve care and reduce costs. These agreements should be revised to allow organizations to get preapproval for real-time access to Centers for Medicare and Medicaid Services (CMS) data for appropriate uses. While many restrictions are important and necessary to protect patient confidentiality, others, such as restrictions on combining data sets, inhibit the true potential of data analysis in healthcare. The current practice of precluding some organizations from purchasing data at all and substantive lag time in the availability of key information slows progress that could benefit everyone.

- **Federal health data should no longer be denied to entities perceived to have a commercial interest.** All entities should be allowed access to federal data to conduct research of interest to federal programs, such as provider and product performance improvement activities. Healthcare organizations are using advanced data analytics to

improve healthcare quality, better manage population health, and address consumer health needs using private-sector patient-level data. Healthcare organizations can do even better with appropriate access to federal program data. At the same time, healthcare organizations have a responsibility to abide by consumer protection laws, such as the Health Insurance Portability and Accountability Act (HIPAA), when handling federal health data. In the era of value-based healthcare and performance-driven reimbursement, all entities arguably have a "commercial interest" in federal program data. These data are important for all healthcare sectors to drive toward value. Commercial entities could be held to the same Data Use Agreement standards as noncommercial entities, including, for example, that the research be relevant to public programs.

Consensus Standards

- **Voluntary, consensus-based standards for observational research must be established that are broadly agreed upon among all healthcare stakeholders and healthcare sectors.** As it becomes technically easier and less costly to use real world heathcare data to establish treatment guidelines and protocols, to make coverage decisions, and to set reimbursement rates, it becomes increasingly important that we work together to ensure that the research is robust. To that end, we need to understand and agree upon the limitations of various data sources and data sets – establish consensus ideas of which data are fit for what purpose. We need consensus on appropriate research methods for nonexperimental observational research, including dataset management. We also need to agree that once research is conducted and findings released, all interested stakeholders should be able to review detailed information about the data set(s) used, how the data were curated, and the research methods employed. The research process should be documented and transparent so that another researcher could replicate a given study.

- **As health data increasingly flow among organizations to improve care, standards for the ownership of health data should be established.** These standards would serve to reduce legal uncertainty and facilitate important information flows.

Secondary Use of Data

- **Efforts to provide consumer transparency of healthcare prices must provide practical, consumer-friendly information that facilitates decisionmaking.** Consumer-accessible data should not include "input prices" but rather prices at the point of service. The price a hospital pays for equipment is not helpful to consumers, but the cost paid by patients for an intervention could be important. In fact, transparency of input prices could cause those prices to regress to the mean over time while still not helping the consumer make informed choices.

- **Any collection and publication of provider price or payment data should be released alongside information on quality in order to drive value in healthcare.** HLC members are continually innovating to drive higher quality and better value in healthcare. There is a significant risk that consumers, when given provider payment information, will make erroneous assumptions about quality based on reimbursement – defeating our efforts to drive toward better value. We urge policymakers to take a thoughtful approach to the release of any cost data to ensure that consumers make a judgment based on value.

- **Electronic Health Records (EHRs)/Electronic Medical Records (EMRs) data should be made available for research and other healthcare innovation.** Despite the fact that the installation of EHRs nationally has been dramatically subsidized by the federal government, it is not yet clear if data collected by the federal government from EHRs as part of the Meaningful Use, Medicare Shared Savings Programs and others will be accessible. Government policy should encourage and foster efforts to use this data to broaden knowledge, improve provider performance, engage patients, and conduct health outcomes research. With appropriate protections for privacy and proprietary information, government policy should support the development of applications that connect various government data sources for approved purposes.

- **Common approaches to risk-adjusting data must be developed to ensure consumer decisionmaking is based on accurate comparisons.** The impact of multiple factors, such as socioeconomic status, on clinical outcomes is well documented. Adjusting for these factors is necessary if data are to be used accurately for comparisons.

<u>Governance and Data Privacy Protections</u>

- **The information protection framework established by the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule should be maintained.** HIPAA established a framework for acceptable uses and disclosures of individually identifiable health information within healthcare delivery and payment systems for the privacy and security of medical information. Confidentiality of patient medical data is of the utmost importance in the delivery of medical care. We must maintain the trust of the patient as we strive to improve healthcare quality. At the same time, providers should have as complete a patient history as is necessary to treat patients. Having access to a complete and timely medical record allows providers to remain confident that they are well informed in the clinical decisionmaking process.

- **A privacy framework should be consistent nationally so that providers, health plans, and researchers working across state lines may exchange patient health data efficiently and effectively to provide treatment, extend coverage, and advance medical knowledge,** whether through a national health information network or another

means of health information exchange. To the extent not already provided under HIPAA, simple, uniform confidentiality rules should apply to all individuals and organizations that create, compile, store, transmit, or use personal health information. Patients' private medical information should have the strictest protection from others outside the medical delivery system and should be supplied only to those necessary for the provision of safe and high-quality care.

- **In order to improve safety and quality, healthcare organizations must have a safe and legal way to match the right patient to his or her own medical record across time and place.** The privacy of individuals in a modern health system must be respected and privacy laws should be vigorously enforced. It is critical that health organizations have a means to gain access to the correct individual patient's medical record in order to provide the right treatment to the right patient at the right time.

- **The timely and accurate flow of deidentified data, with appropriate protections for consumer privacy, is crucial to achieving the true potential of data analytics in healthcare.** Federal privacy policy should abide by HIPAA regulations for the deidentification and/or aggregation of data to allow access to properly deidentified information. This allows researchers, public health officials, and others to assess quality of care, investigate threats to the public's health, respond quickly in emergency situations, and collect information vital to improving healthcare safety and quality.

BRENNAN
CENTER
FOR JUSTICE

Brennan Center for Justice
At New York University School of Law

Washington, D.C. Office
1730 M Street, N.W.
Suite 413
Washington, D.C. 20036
202.249.7190 Fax 202.223.2683

April 4, 2014

Big Data Study
Office of Science and Technology Policy
Eisenhower Executive Office Building
1650 Pennsylvania Ave. NW.
Washington, DC 20502
Via email: bigdata@ostp.gov

Re: Big Data Request for Information

To whom it may concern:

On March 4, 2014, the Office of Science and Technology Policy (OSTP) issued a request for public comment "on the ways in which big data may impact privacy, the economy, and public policy."[1] This request is part of the OSTP's "comprehensive review of how 'big data' will affect how Americans live and work,"[2] which is expected to be released by the end of April. We are pleased that the White House and OSTP are considering these issues, and we welcome the prospect of a comprehensive review of both the benefits and pitfalls of the use of "big data" by government and private entities.

On March 3, John Podesta told a forum at the Massachusetts Institute of Technology that the White House's review of recently revealed government surveillance programs was occurring on a "somewhat separate track," but that the big

[1] Request for Information, 79 FR 12251 (Mar. 4, 2014), *available at* https://www.federalregister.gov/articles/2014/03/04/2014-04660/government-big-data-request-for-information#h-7.
[2] *Deadline Extension: There is Still Time to Join the Conversation on Big Data and Privacy*, WHITE HOUSE OFFICE OF SCIENCE AND TECHNOLOGY POLICY (Mar. 31, 2014, 7:10 PM), http://www.whitehouse.gov/blog/2014/03/31/deadline-extension-there-still-time-join-conversation-big-data-and-privacy (also indicating that comments are now due April 4, 2014).

data review "may help inform intelligence policy going forward."[3] We thus write to focus on the use of "big data," including data mining, for counterterrorism and other national security purposes.

There is substantial consensus among scientists and national security experts that the use of big data and data mining in the counterterrorism arena is ineffective. A drive to amass information on the scale of "big data" may even be counterproductive in the national security context to the extent it overwhelms intelligence agencies with unhelpful information. In addition, there are heightened risks to privacy and civil liberties associated with the accumulation of deep databases of information for counterterrorism and other national security purposes. In light of the potential impact of OSTP's proposals on intelligence policy, it is critical that OSTP's review acknowledge the limited utility of big data analytics in the counterterrorism context. At the very least, OSTP should affirm that principles for the use of big data that may be beneficial in certain contexts may be inappropriate, underinclusive, or overbroad in other contexts, particularly with respect to national security.

These comments address the following issues, which are responsive to Questions 1, 2, and 4 of the Request for Information:

- Public policy implications of the collection, storage, analysis, and use of big data;
- Uses of big data that raise significant public policy concerns; and
- Distinctions between policy frameworks for use of big data by different sectors – specifically, by counterterrorism-focused and other national security agencies.

Pattern-Based Data Mining Has Limited Value in the Counterterrorism Context

One potential use for "big data" is data mining, or "pattern prediction": analyzing a store of data to tease out patterns connected to certain behaviors, and then looking for matching patterns in other datasets to predict other instances in which those behaviors are likely to occur.[4] When it comes to counterterrorism, however, a

[3] Kate Tummarello, *'Big Data' review to focus on private sector*, THE HILL TECH. BLOG (Mar. 3, 2014, 12:47 PM), http://thehill.com/blogs/hillicon-valley/technology/199710-white-house-big-data-review-to-focus-on-companies.

[4] MARY DEROSA, CTR. FOR STRATEGIC AND INT'L STUDIES, DATA MINING AND DATA ANALYSIS FOR COUNTERTERRORISM 4 (2004), *available at* http://csis.org/files/media/csis/pubs/040301_data_mining_report.pdf; K. A. Taipale, *Data Mining and Domestic Security: Connecting the Dots to Make Sense of Data*, 5 COLUM. SCI. & TECH. L. REV. 1, 22-23 (2003), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=546782 ("Data mining

study commissioned by the Department of Defense concluded that "there is no credible approach that has been documented … to accurately anticipate" terrorist threats.[5] Put another way, there is no known way to effectively identify a potential terrorist by pattern analysis.

Credit card companies are probably the best-known and most successful users of the pattern-matching model. Their success in detecting credit card fraud is due to a number of factors that are almost entirely lacking in the counterterrorism context: the massive volume of credit card transactions provides a rich body of data; a relatively high rate of credit card fraud means the model can be tested and refined; regular and identifiable patterns accompany the fraud (such as testing a card at a gas station to ensure that it works and then immediately purchasing more expensive items); and the cost of a false positive — what happens when the system erroneously concludes that a card has been stolen — is relatively minimal: a call to the owner and, at worst, premature closure of a legitimate account.[6]

By contrast, there have been — statistically speaking — a relatively small number of attempted or successful terrorist attacks, which means that there are no reliable "signatures" to use for pattern modeling.[7] Even if the number of attacks were to rise significantly, it is improbable that they would exhibit enough common characteristics to allow for successful modeling. Indeed, government agencies and experts who have engaged in rigorous empirical studies of "radicalization" have

---

generally identifies patterns or relationships among data items or records that were not previously identified (and are not themselves data items) but that are revealed in the data itself. Thus, data mining extracts information that was *previously unknown*.") (internal citations omitted).

[5] JASON, MITRE CORP., RARE EVENTS § 1.5, at 8 (2009), *available at* http://www.fas.org/irp/agency/dod/jason/rare.pdf. A National Academies of Science report echoed this finding, determining that terrorist identification via data mining (or by "any other known methodology") was "neither feasible as an objective nor desirable as a goal of technology development efforts." NAT'L RESEARCH COUNCIL, PROTECTING INDIVIDUAL PRIVACY IN THE STRUGGLE AGAINST TERRORISTS: A FRAMEWORK FOR PROGRAM ASSESSMENT 3-4 (2008), *available at* http://epic.org/misc/nrc_rept_100708.pdf.

[6] *See* Bruce Schneier, *Why Data Mining Won't Stop Terror*, WIRED (Mar. 9, 2006), http://archive.wired.com/politics/security/commentary/securitymatters/2006/03/70357; *see also* Richard Barrington, *2011 Credit Card Facts and Statistics*, INDEXCREDITCARDS (Jan. 10, 2011), http://www.indexcreditcards.com/finance/creditcardstatistics/2011-report-on-credit-card-usage-facts-statistics.html (noting that as of 2010, there were nearly 1.5 billion credit cards in circulation in the United States, and nearly 55 million credit card transactions every day).

[7] *See* JEFF JONAS & JIM HARPER, CATO INST., POLICY ANALYSIS NO. 584: EFFECTIVE COUNTERTERRORISM AND THE LIMITED ROLE OF PREDICTIVE DATA MINING (2006), *available at* http://www.cato.org/publications/policy-analysis/effective-counterterrorism-limited-role-predictive-data-mining ("With a relatively small number of attempts every year and only one or two major terrorist incidents every few years – each one distinct in terms of planning and execution – there are no meaningful patterns that show what behavior indicates planning or preparation for terrorism. Unlike consumers' shopping habits and financial fraud, terrorism does not occur with enough frequency to enable the creation of valid predictive models").

concluded that there is no particular pathway to terrorism or a common terrorist profile.[8]

Moreover, a counterterrorism data-mining program would look not just at a single type of data, such as credit card transactions, but "trillions of connections between people and events": merchandise purchases, travel preparations, emails, phone calls, meetings, business arrangements, and more.[9] It is close to impossible to identify coherent patterns that could be used to predict terrorist activity within this welter of data.

In addition, the adverse consequences of a false positive are vastly more damaging to an individual in the counterterrorism context. As security expert Bruce Schneier has suggested, given the almost overwhelming amount of data available, the most accurate imaginable system would still generate on the order of "1 billion false alarms" — that is, emails, meetings, associations, phone calls, and other items falsely tagged as terrorism-related — "for every real terrorist plot it uncovers."[10] A person falsely suspected of involvement in a terrorist scheme will become the target of long-term scrutiny by law enforcement and intelligence agencies. She may be placed on a watchlist or even a no-fly list, restricting her freedom to travel and ensuring that her movements will be monitored by the government. Her family and friends may become targets as well.

And unlike credit card fraud, a conclusion of possible terrorist involvement is more likely to be influenced by activities that may be protected by the First Amendment, such as email or phone communications, political activism, religious involvement, or connections to certain ethnic groups. In short, there is a reason the Cato Institute has warned that data mining for counterterrorism purposes "would waste taxpayer dollars, needlessly infringe on privacy and civil liberties, and misdirect the valuable time and energy of the men and women in the national security community."[11]

---

[8] FAIZA PATEL, BRENNAN CTR. FOR JUSTICE, RETHINKING RADICALIZATION 8 (2011), *available at* http://www.brennancenter.org/sites/default/files/legacy/RethinkingRadicalization.pdf; MARC SAGEMAN, LEADERLESS JIHAD: TERROR NETWORKS IN THE TWENTY-FIRST CENTURY 72 (2008); Clark McCauley & Sophia Moskalenko, *Mechanisms of Political Radicalization: Pathways Toward Terrorism*, 20 TERRORISM & POLITICAL VIOLENCE 415, 418 (2008); RICHARD ENGLISH, TERRORISM: HOW TO RESPOND 52 (2009).

[9] Schneier, *supra* note 6.

[10] *Id*. The Department of Defense JASON study described this problem as the high risk of "false alarm rates … in the face of massive clutter." JASON, *supra* note 5.

[11] JONAS & HARPER, *supra* note 7, at 1.

Collection of Large Amounts of Data May Be Detrimental to National Security

Second, the large-scale collection of information by national security agencies has been repeatedly associated with failures of intelligence by those agencies. Thus, while there may be many circumstances in which the accumulation of large quantities of data is critical – to facilitate credit card fraud detection, for instance, or to enable government agencies to track pandemics and safeguard public health – the wholesale collection of data by national security agencies is likely to undermine rather than enhance effective analysis.

To be sure, "big data" is qualitatively different from "large data," and not enough is publicly known about these agencies' databases to accurately assess whether they qualify as "big data." Regardless of the technical classification, however, the risks associated with far-reaching information collection will surely be magnified if and when big data is deployed in the national security context.

For instance, experts concluded that an overabundance of data contributed significantly to the failure of the intelligence community to intercept the so-called "underwear bomber" — the suicide bomber who nearly brought down a plane to Detroit on Christmas Day 2009. As an official White House review of the attempted attack observed, a significant amount of critical information was available to the intelligence agencies but was "embedded in a large volume of other data."[12] Similarly, the independent investigation of the FBI's role in the shootings by U.S. Army Major Nidal Hasan at Fort Hood concluded that the "crushing volume" of information was one of the factors that hampered accurate analysis prior to the attack.[13]

Officials across a range of agencies have echoed this assessment. As one veteran CIA agent described it, "The problem is that the system is clogged with information," most of which "isn't of interest."[14] A former official in the Department of Homeland Security branch that handled information coming from fusion centers (state- or regional-based centers that collect, analyze and share threat-related

[12] THE WHITE HOUSE, SUMMARY OF THE WHITE HOUSE REVIEW OF THE DECEMBER 25, 2009 ATTEMPTED TERRORIST ATTACK 3 (n.d.), *available at* http://www.whitehouse.gov/sites/default/files/summary_of_wh_review_12-25-09.pdf.
[13] *Lessons from Fort Hood: Improving Our Ability to Connect the Dots: Hearing Before the Subcomm. on Oversight, Investigations, and Mgmt. of the H. Comm. on Homeland Security*, 112th Cong. 2 (2012) (statement of Douglas E. Winter, Deputy Chair, William H. Webster Comm'n on the Fed. Bureau of Investigation, Counterterrorism Intelligence, and the Events at Fort Hood, Texas on Nov. 5, 2009).
[14] David Ignatius, *A Breakdown in CIA Tradecraft*, WASH. POST, Jan. 6, 2010, *available at* http://articles.washingtonpost.com/2010-01-06/opinions/36805490_1_cia-base-cia-veteran-agency-officers.

information among the federal government, the state, and other partners) characterized the problem as "a lot of data clogging the system with no value."[15] Even former Defense Secretary Robert Gates has acknowledged that a decade after 9/11, one would need to ask, "'Okay, we've built tremendous capability, but do we have more than we need?'"[16]

In addition, centralized storehouses of data are particularly vulnerable to intentional as well as inadvertent security breaches, which can have significant consequences both for operational effectiveness and for the victims of the breaches. In mid-2008, for instance, it was revealed that the director of the secretive Strategic Technical Operations Center at the Marine Corps' Camp Pendleton had been feeding reams of classified federal surveillance files to a local terrorism task force.[17] These information-sharing breaches may become more common as more data is aggregated and available through a single access point. In fact, the federal Government Accountability Office has reported a significant increase in data breaches since 2006, with more than a third of the incidents in 2011 involving personally identifiable information, and the compilation of massive databases is likely to further amplify the risk.[18]

Large-Scale Collection of Information Threatens Civil Liberties and Freedom of Expression

Finally, the collection and retention of personal information on a vast scale poses well-recognized risks to privacy, invites abuse, and chills freedom of expression and dissent. As the Senate's Church Committee recognized over four decades ago, "The massive centralization of … information creates a temptation to use it for improper purposes, threatens to 'chill' the exercise of First Amendment rights, and is inimical to the privacy of citizens."[19] Of course, the nature of "big data"

---

[15] STAFF OF THE SUBCOMM. ON INVESTIGATIONS OF THE S. COMM. ON HOMELAND SEC., 112TH CONG., FEDERAL SUPPORT FOR AND INVOLVEMENT IN STATE AND LOCAL FUSION CENTERS 27 (Comm. Print 2012), *available at* http://www.hsgac.senate.gov/subcommittees/investigations/media/investigative-report-criticizes-counterterrorism-reporting-waste-at-state-and-local-intelligence-fusion-centers.

[16] Dana Priest & William Arkin, *Top Secret America: A Hidden World, Growing Beyond Control*, WASH. POST, July 19, 2010, *available at* http://projects.washingtonpost.com/top-secret-america/articles/a-hidden-world-growing-beyond-control/.

[17] *See* Rick Rogers, *Records Detail Security Failure in Base File Theft*, SAN DIEGO UNION-TRIBUNE, May 22, 2008, *available at* http://www.utsandiego.com/uniontrib/20080522/news_1n22theft.html; *Law Enforcement Records Sought in Stolen Pendleton Surveillance Documents*, ACLU OF SAN DIEGO (July 15, 2008), http://www.aclusandiego.org/law-enforcement-records-sought-in-stolen-pendleton-surveillance-documents-massive-number-of-files-stolen-according-to-press-report/.

[18] U.S. GOV'T ACCOUNTABILITY OFFICE, GAO-12-961T, FEDERAL LAW SHOULD BE UPDATED TO ADDRESS CHANGING TECHNOLOGY LANDSCAPE 13 (2012), *available at* http://www.gao.gov/assets/600/593146.pdf (statement of Gregory Wilshusen).

[19] SELECT COMM. TO STUDY GOV'T OPERATIONS WITH RESPECT TO INTELLIGENCE ACTIVITIES, FINAL REPORT, S. REP. NO. 94-755, bk. III, at 778, *available at* http://www.intelligence.senate.gov/churchcommittee.html.

is that it may not be immediately susceptible to targeted searching or retrieval. Nevertheless, attempts to tackle and fix the big data problem will inevitably result in the easier availability of large stores of information, and the risks of abuse will rise as well.

At a minimum, these significant drawbacks mean that any collection of personal information by intelligence and law enforcement agencies must be justified by a significant benefit. There is ample reason to question whether the gathering of information without any basis to suspect wrongdoing is useful to counterterrorism efforts in any manner.[20] Certainly, as noted above, the indiscriminate collection of information for pattern analysis – the premise of a "big data" approach – is not a useful counterterrorism tool. At the same time, the more data that is collected, the greater the potential for abuse, chilling effect, and privacy intrusion.

These risks are not merely theoretical. For instance, recent disclosures about the National Security Agency have revealed both inadvertent and intentional misuses of the agency's broad surveillance authority[21] – problems that were revealed only after repeated assurances that the agency was operating in strict conformance with applicable legal standards.[22] The agency also succeeded in obtaining approval to search its storehouses of data for information about Americans without a warrant, and recently confirmed that it has conducted such warrantless searches.[23] While this

---

[20] *See, e.g.*, EMILY BERMAN, BRENNAN CTR. FOR JUSTICE, DOMESTIC INTELLIGENCE: NEW POWERS, NEW RISKS (2011), *available at* http://www.brennancenter.org/publication/domestic-intelligence-new-powers-new-risks; PATEL, *supra* note 8; RACHEL LEVINSON-WALDMAN, BRENNAN CTR. FOR JUSTICE, WHAT THE GOVERNMENT DOES WITH AMERICANS' DATA (2013), *available at* http://www.brennancenter.org/sites/default/files/publications/What%20Govt%20Does%20with%20Data%20100813.pdf.

[21] *See, e.g.*, Barton Gellman, *NSA Broke Privacy Rules Thousands of Times per Year, Audit Finds*, WASH. POST, Aug. 15, 2013, *available at* http://articles.washingtonpost.com/2013-08-15/world/41431831_1_washington-post-national-security-agency-documents; Adam Gabbatt and agencies, *NSA Analysts 'Wilfully Violated' Surveillance Systems, Agency Admits*, GUARDIAN, Aug. 24, 2013, *available at* http://www.theguardian.com/world/2013/aug/24/nsa-analysts-abused-surveillance-systems; Chris Strohm, *Lawmakers Probe Willful Abuses of Power by NSA Analysts*, BLOOMBERG, Aug. 24, 2013, *available at* http://www.bloomberg.com/news/2013-08-23/nsa-analysts-intentionally-abused-spying-powers-multiple-times.html; Press Release, Office of Sen. Chuck Grassley, Grassley Presses for Details about Intentional Abuse of NSA Authorities (Aug. 28, 2013), *available at* http://www.grassley.senate.gov/news/Article.cfm?customel_dataPageID_1502=46858.

[22] *See, e.g.*, Dan Farber, *President Obama Outlines Four NSA Reform Initiatives*, CNET (Aug. 9, 2013, 1:13 PM) http://news.cnet.com/8301-13578_3-57597814-38/president-obama-outlines-four-nsa-reform-initiatives/ (quoting President Obama as saying that NSA "programs are operating in a way that prevents abuse"); Edward Moyer, *NSA Admits to Some Deliberate Privacy Violations*, CNET (Aug. 23, 2013, 1:08 PM), http://news.cnet.com/8301-13578_3-57599916-38/nsa-admits-to-some-deliberate-privacy-violations/ (noting that earlier that month, then-NSA Director Keith Alexander said that "no one has willfully or knowingly disobeyed the law or tried to invade your civil liberties or privacy").

[23] *See* Spencer Ackerman and James Ball, *NSA performed warrantless searches on Americans' calls and emails – Clapper*, GUARDIAN, Apr. 1, 2014 *available at* http://www.theguardian.com/world/2014/apr/01/nsa-surveillance-loophole-americans-data; Julian

warrantless access was approved by the FISA Court, it highlights the fact that information collected for one purpose may eventually be used for another, making it particularly critical that close attention be paid to the accumulation of information that may be susceptible to abuse down the line.

Additionally, in the years after 9/11, the Federal Bureau of Investigation improperly gathered, recorded, and retained information about individuals' First Amendment-protected activities, often leading to targets' inclusion in federal databases from which it became almost impossible to escape.[24] Fear of inclusion in such databases – and potential scrutiny based on religion, national origin, or other suspect category – may lead to self-censorship or other chilling effects. Indeed, this phenomenon has been documented in the context of one local surveillance program.[25]

Even innocuous information is vulnerable to abuse, often for petty reasons. For instance, a special agent with the U.S. Commerce Department was indicted for and pled guilty to misusing a federal database to track a former girlfriend and her family. The agent had previously threatened to kill the girlfriend or have her and her family deported, and he accessed the database over 150 times in a one-year period to monitor her movements.[26] Recent reports by the FBI's Office of Professional Responsibility depict FBI employees misusing government databases to look up friends working as exotic dancers and conduct searches on celebrities they "thought were hot."[27] Many additional examples have been documented on the state level.[28]

Hattem, *Lawmakers incensed over NSA 'loophole,'* THE HILL TECH. BLOG (Apr. 1, 2014, 5:04 PM), http://thehill.com/blogs/hillicon-valley/technology/202350-lawmakers-incensed-over-nsa-loophole.

[24] OFFICE OF THE INSPECTOR GEN., U.S. DEP'T OF JUSTICE, A REVIEW OF THE FBI'S INVESTIGATIONS OF CERTAIN DOMESTIC ADVOCACY GROUPS 176, 182-84 (2010), *available at* http://www.justice.gov/oig/special/s1009r.pdf (describing investigations of PETA, Greenpeace, and Catholic Worker, among others).

[25] *See, e.g.*, MUSLIM AM. CIVIL LIBERTIES COAL. ET AL., MAPPING MUSLIMS: NYPD SPYING AND ITS IMPACT ON AMERICAN MUSLIMS 29-32, 40-45 (2013), *available at* http://www.law.cuny.edu/academics/clinics/immigration/clear/Mapping-Muslims.pdf (describing the consequences of the New York Police Department's monitoring of the city's Middle Eastern and South Asian population, including alienating Muslims from their mosques and religious communities, hindering social activism and political debate, and prompting student groups on monitored campuses to ban constitutionally-protected political discussions in group spaces).

[26] *United States v. Robinson*, No. 5:07-cr-00596-JF (N.D. Cal. Aug. 25, 2009); Henry K. Lee, *Ex-Agent Indicted in Misuse of Database*, S.F. GATE (Sept. 19, 2007, 4:00 AM), http://www.sfgate.com/bayarea/article/Ex-agent-indicted-in-misuse-of-database-2522021.php.

[27] Scott Zamost & Kyra Phillips, *FBI Misconduct Reveals Sex, Lies and Videotape*, CNN (Jan. 27, 2011, 10:07 AM), http://articles.cnn.com/2011-01-27/us/siu.fbi.internal.documents_1_fbi-employees-occasional-employee-fbi-s-office?_s=PM:US.

[28] *See, e.g.*, Danielle Bell, *Ottawa Cop Demoted for Database Misuse*, OTTAWA SUN, Sept. 26, 2012, *available at* http://www.ottawasun.com/2012/09/26/ottawa-cop-demoted-for-misuse-of-data-bases (senior staff sergeant accessed police databases 169 times over nearly four years for personal reasons); *Former Montreal Detective Used Police Database to Help Mafia*, TORONTO SUN, Nov. 22, 2012, *available at* http://www.torontosun.com/2012/11/22/former-montreal-detective-used-police-database-

Conclusion

In sum, the accumulation of information on a "big data" scale appears to be both ineffective and counterproductive in the national security context. It also poses particular risks to privacy, civil liberties, and freedom of expression and association. While OSTP may not be explicitly addressing the use of big data in the national security context, the principles it articulates are likely to be influential in the national

---

to-help-mafia# (Montreal police detective used a police database to run license plates and pass information to members of an organized crime syndicate); Christine Hauser, *Sergeant Said to Misuse Terror-Watch Database*, N.Y. TIMES, Nov. 21, 2008, at A31, *available at* http://www.nytimes.com/2008/11/21/nyregion/21sergeant.html; *see also* Sewell Chan, *Police Sergeant Guilty of Misusing Terror Database*, N.Y. TIMES, Jan. 14, 2009, http://cityroom.blogs.nytimes.com/2009/01/14/police-sergeant-pleads-guilty-to-misusing-database/ (reporting that a New York City police sergeant illicitly used a state database to retrieve information from a national terrorist watch list for an acquaintance involved in a child-custody case); Lee, *supra* note 26 (special agent with U.S. Commerce Department indicted in 2007 by federal grand jury for misusing a federal database to track a former girlfriend and her family; agent had previously threatened to kill the girlfriend or have her and her family deported, and he accessed database over 150 times in a one-year period to monitor her movements); Jessica Lussenhop, *Is Anne Marie Rasmusson Too Hot to Have a Driver's License?*, CITY PAGES (Feb. 22, 2012), http://www.citypages.com/2012-02-22/news/is-anne-marie-rasmusson-too-hot-to-have-a-driver-s-license/ (over a hundred officers from eighteen agencies across Minnesota accessed the driving records of a female ex-police officer to look at her picture and glean personal details about her, claiming the practice was common place despite state laws requiring all searches to have an investigative purpose); Tom Lyons, *The Odd Loose Ends in Database Misuse*, SARASOTA HERALD-TRIBUNE, Oct. 11, 2012, at BNV1, *available at* http://www.heraldtribune.com/article/20121011/ARCHIVES/210111025 (secretaries at Florida state attorney's office accessed driver and vehicle information database, limited to official police and prosecutorial use, to perform unauthorized searches for information on candidate for state attorney); Allison Manning, *Cops Criticized for 'Misuse' of Databases*, POLICEONE.COM (Apr. 2, 2012), http://www.policeone.com/police-products/software/Data-Information-Sharing-Software/articles/5360910-Cops-criticized-for-misuse-of-databases/ (officials misusing police databases in Ohio included police officer who looked up a woman's personal information and stopped her car more than a dozen times, police officer who "threw items into the front yard of two people he looked up," and three deputies who looked up the "wife of a man with whom one of the deputies had a dispute"); *Former Montgomery Co. Officer Guilty of Police Database Misuse*, DAILY RECORD (Apr. 27, 2011, 4:46 PM), http://thedailyrecord.com/2011/04/27/former-montgomery-co-officer-guilty-of-police-database-misuse/ (former police officer accessed law enforcement databases to assist her drug-dealing fiancé); Levi Pulkkinen, *IRS Worker Caught Snooping on Ex, Others*, SEATTLEPI.COM (Apr. 23, 2012, 9:44 PM), http://www.seattlepi.com/local/article/IRS-worker-caught-snooping-on-ex-others-3498550.php (IRS technician who had previously looked up her ex-husband's tax return pled guilty to misusing her access to IRS databases to review other people's personal information, including a relative with whom she had had a falling out); Aaron Rupar, *In Minneapolis, Private Information Database Abuse 'Endemic,' Attorney Says*, CITY PAGES (Sept. 26, 2012, 12:27 PM), http://blogs.citypages.com/blotter/2012/09/in_minneapolis_private_information_database_abuse_endemic_attorney_says.php (employees in Minneapolis's department of housing charged with accessing driver's license databases for personal purposes; one of the employees also shared his log-in information with other employees); *Utah Launches Investigation of Leak of Immigrants' Information*, CNN (July 22, 2010, 4:12 PM), http://www.cnn.com/2010/US/07/22/utah.attorney.general/index.html?eref=rss_latest&utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+rss%2Fcnn_latest+%28RSS%3A+Most+Recent%29 (employees of state's Department of Workforce Services generated and circulated a list of 1,300 state residents whom they falsely accused of being illegal immigrants).

security and intelligence arena. Accordingly, we urge OSTP to recognize that while basic principles such as transparency, accountability, oversight, and privacy protections are likely to be relevant regardless of the context, other principles or uses of big data may have unintended consequences if imported from one area to another. We also ask that OSTP recognize the pitfalls associated with the utilization of big data in the national security context, to help ensure that those uses are analyzed with particular attention to the concerns articulated above.

Please do not hesitate to contact us if we can be of further assistance as the review proceeds. Liza Goitein may be reached at elizabeth.goitein@nyu.edu or 202-249-7192, and Rachel Levinson-Waldman may be reached at rachel.levinson.waldman@nyu.edu or 202-249-7193.

Sincerely,

Elizabeth Goitein
Co-Director, Liberty and National Security Program

Rachel Levinson-Waldman
Counsel, Liberty and National Security Program

MAKING CHANGE AT WALMART our community. our future.

VIA ELECTRONIC SUBMISSION

April 4, 2014

Ms. Nicole Wong
Deputy Chief Technology Officer
Attn: Big Data Study
White House Office of Science and Technology Policy
Eisenhower Executive Office Building
1650 Pennsylvania Avenue, NW
Washington, DC 20502

Dear Ms. Wong:

We write today to urge the Office of Science and Technology Policy to consider the impacts of big data collection on low-income communities and people of color. As the Office of Science and Technology Policy rightly recognizes, the implications of collecting, analyzing, and using such data warrant careful review, given their rapidly increasing prevalence. Specifically, we would like to address Question 2, on the types of uses of big data that raise the most public policy concerns.

Communities of color have adopted smartphone technology at faster rates than their white counterparts. Additionally, low-income Americans are more likely to access the internet predominantly through cell phones than more affluent Americans.[1] Mobile technology offers greater opportunities for tracking than ever before. Companies are taking advantage of this technology, largely unfettered, to market to consumers based on this tracking. Take as an example Walmart, which has made clear its massive big data ambitions;[2] has clearly spelled out its intent to market heavily to communities of color;[3] and has thus far provided little transparency to allow advocates and the public to understand how their information is being used, or abused.

As companies like Walmart increasingly use data, both real and predicted, to put people into categories, the risk grows that some groups will fall disproportionately into categories which receive less favorable treatment. Among the more disturbing aspects of this possibility is the complete lack of transparency and accountability. Whereas many forms of discrimination have historically been "testable," discrimination by big data will likely be very difficult to detect and—even when reasonably suspected—to prove. As MIT researcher Kate Crawford said at a recent conference, "It's not that big data is

[1] Duggan, Maeve and Aaron Smith. "Cell Internet Use 2013." Pew Research Internet Project. 16 Sept 2013.
[2] Evan Clark. "Walmart Sets Aggressive Digital Plan." *Women's Wear Daily*. 2 May 2013.
[3] Wentz, Laurel. "Walmart's Tony Rogers: 100% of Growth Is Multicultural." *Ad Age*. 31 Oct 2012.

effectively discriminating—it is, we know that it is. It's that you will never actually know what those discriminations are."[4]

A recent study of Walmart's online data collection by the Center for Media Justice, ColorOfChange, and SumOfUs illustrates what is at stake as online retailers and advertisers gather increasingly more information about consumers: Walmart could have exhaustive consumer data on more than 145 million Americans – more than 60 percent of U.S. adults. The company says it has "petabytes" of data on consumers. At a September 2013 conference, Walmart U.S. CEO Bill Simon told the crowd that the company's "ability to pull data together is unmatched."[5] And yet, there is no mechanism for consumers to fully understand how Walmart uses that data or how Walmart and its partners interpret that information.

The same study found that Walmart shares consumers' online data with more than 50 third parties. This information includes every page and product viewed, unique identifiers for users and their devices, system information such as device type and operating system version, and location information.[6] We believe that companies collecting information like this should be transparent about what is collected and how it is used. Further, consumers should have the ability to have such information deleted or corrected, or to opt out of tracking altogether. This information should be used fairly and accurately.

Tracking and surveillance technologies have evolved rapidly, and despite industry attempts at self-regulation, consumers are not yet guaranteed that this progress will bring greater safety, economic opportunity, and convenience to everyone. Therefore, we urge the Office of Science and Technology Policy to carefully consider the impacts of big data on all communities, and to ensure that technological advances are implemented in ways that respect the values of equal opportunity and equal justice.

Thank you for the opportunity to provide input on this issue.

Sincerely,

Dan Schlademan	Brandon Garrett	Amalia Deloney
Director	Member	Policy Director
Making Change at Walmart	OUR Walmart	Center for Media Justice

[4] Pressman, Aaron. "Big Data Could Create an Era of Big Discrimination." *Yahoo Finance*. 14 Oct 2013.
[5] "Edited Transcript: WMT – Walmart at Goldman Sachs Global Retailing Conference." Thomson Reuters StreetEvents. 11 Sept 2013.
[6] The Center for Media Justice, ColorOfChange, SumOfUs. "Consumers, Big Data, and Online Tracking in the Retail Industry: A Case Study of Walmart." Nov 2013.

Online Trust Alliance

April 4, 2014

Nicole Wong, Esq.
Office of Science and Technology Policy
Eisenhower Executive Office Building
1650 Pennsylvania Avenue NW
Washington, DC 20502

RE: Big Data Study - Notice of Request for Information 79 FR 12251

Dear Ms. Wong:

Thank you for providing the Online Trust Alliance (OTA) the opportunity to submit comments in response to the White House Office of Science and Technology Policy (OSTP)'s Request for Information. As a 501c3 non-profit organization, OTA's mission is to enhance online trust and empower users while promoting innovation and the vitality of the Internet. OTA works to educates businesses, policy makers and stakeholders about best practices and tools that enhance the protection of users' security, privacy and identity. OTA supports collaborative public-private partnerships, benchmark reporting, meaningful self-regulation and data stewardship.

OTA applauds the Administration's leadership and commitment in facilitating this multi-stakeholder discussion. We share the mutual goal of maximizing big data's benefits and the free flow of information while minimizing the privacy and related data security risks. This goal requires government and business entities to become stewards of data, moving from a minimal compliance perspective to one that takes full account of consumer context and societal expectations.

With big data, we have witnessed an unparalleled transformation of how information is collected, analyzed and used. We have also seen the impact of U.S. policy and businesses practices reaching far beyond our physical boarders. Big data brings many benefits and implications to the global economy. There is significant economic potential in fraud detection, security and threat modeling technologies. However, we are simultaneously grappling with the balkanization of the Internet, proposed trade barriers and challenges to the existing US-EU Safe Harbor. Therefore we must establish policies facilitating collection and sharing that have controls and processes preventing abuse.

Big data involves the collection of vast amounts of information from a growing number and variety of sources. These powerful analytic tools include *both* on and off-line data collection. These tools foster useful insights applicable to a range of business, medical and social issues. In the right context, consumers can realize significant benefits in services and offers tailored for their needs and lifestyles. In

the wrong context real harm can occur, ranging from public embarrassment to denial of benefits, manipulated costs of products and lost employment opportunities.[1]

Today's digital reality goes beyond the "reasonable expectation of privacy," compromising the very foundation we have built the Internet upon. If left unchecked, consumer trust will continue to decline, the adoption of anti-tracking technologies will continue to increase and the long-term health of the internet will be undermined.

OTA looks forward to continued dialog on these issues and provides the following responses:

(1) What are the public policy implications of the collection, storage, analysis, and use of big data?

Data collected about individuals and businesses is growing through new and innovative means, spanning both the online and off line world. The "Internet of Things" has brought an era of dependence and connectivity to our mobile devices, homes, automobiles and applications. With the blurring of work and personal boundaries, there are significant policy implications. Controls must be established for both the public and privacy sector. Otherwise, we risk finding ourselves in the dystopian world such as "Minority Report".[2]

Big data collection is occurring in multiple dimensions. Analytic capabilities and data mining tools are transforming what appears to be benign and individual data elements into a comprehensive picture of one's life and lifestyle. For example, medical data, financial information and social relationships are often collected and analyzed to better understand a consumer's personal interests. Yet in the wrong hands and wrong context, this usage could violate an individual's basic right of privacy.[3]

Unintended consequences such as accidental exposure, internal misuse or data loss incidents are real and present dangers and can cause disastrous effects. Unencumbered access to sensitive information could further drive fraud and identity theft. When a determined criminal tries to compromise a company, no matter what level of security protection is in place, both the company and their customers become victims. Recent data loss incidents and rouge employee access demonstrates how sensitive information can become accessible to third parties and result in harm to consumers, both U.S Citizens and consumers abroad.

In order to demonstrate their commitment to proactive privacy ethics and practices, entities engaged in big data analytics should detail the purposes for which they collect and use information. From a policy perspective, user expectations are a good indication of whether purpose of collection relates to its intended use. Similar to credit reports, individuals should have the right to, access, examine and

[1] See, Ryan Calo. *Digital Market Manipulation*. UNIV.OF WASH. SCHOOL OF LAW NO. 2013-27**.** (Aug. 13, 2013)
[2] *Minority Report* is a science fiction film where crime is virtually. See, *Xplenty* Yaniv Mor. *Big Data and Law Enforcement: Was "Minority Report" Right?* WIRED**,** (Mar. 5, 2014). http://www.wired.com/2014/03/big-data-law-enforcement-minority-report-right/
[3] See, Woodrow Hartzog and Evan Selinger. *Big Data in Small Hands*. 66 STAN. L. REV. ONLINE 81, (Sept. 3, 2013).

challenge collected individualized data.[4] Access by the consumer to data collected on that individual is essential to long term consumer trust in the data ecosystem.

Additionally to protect consumer privacy, governments and businesses alike must be transparent and inform individuals about the collection and use of their personal data. These entities must take steps to protect the data from abuse and their infrastructure from compromise. Just like first responders to a fire, entities should have data managers and cyber responders trained, equipped and empowered to deal with an incident. OTA recommends related entities adopt leading security practices, including comprehensive end-to-end encryption as well as developing a data breach readiness and incident response plan.[5]

(2) What types of uses of big data could measurably improve outcomes or productivity with further government action, funding or research? What types of uses of big data raise the most public policy concerns? Are there specific sectors or types of uses that should receive more government and/or public attention?

Society has experienced tremendous benefits in predictive analysis for environmental and medical research. Individual health care professionals may see what may appear to be one or two unrelated patients. Aggregated data, however, can provide insights into potential epidemics and identify trends before they publicly spread. The pharmaceutical industry, aggregating years of research, has expedited the availability of new treatments and medicines and the Center for Disease Control (CDC) has used big data to create flu-tracking systems.[6] The benefits will only increase as these tools and data collection practices become more sophisticated and dynamic. As our dependence on these systems and big data increase, OTA recommends that the government increase funding and research for cyber security measures that will protect these systems vitality and promote innovation.

Recognizing that these tools and specific technologies will evolve, OTA recommends the OSTP should focus the discussion on fundamental consumer concerns of data collection, usage/sharing and obligations. The OSTP should consider how to leverage its unique role as a convening authority for both private and public participants. Learning the lessons from past multistakeholder processes, a working group should be formed with equal numbers of representatives from each constituency in order to promote a collaborative and balanced view and meaningful self-regulatory best practices.

Balancing privacy with security is another fundamental concern when forming public policy. As new technologies such as location trackers and persistent identifiers are integrated into devices and applications, safeguards must be put in place to help prevent or detect fraud and abuse.[7] However it is

[4] One example is Acxiom's "About the Data," a tool that gives individuals control over categories of information collected and allows consumers to correct this information, suppress any data they see, or opt-out of Acxiom's system. https://www.aboutthedata.com/#education

[5] Online Trust Alliance, *The 2014 Data Protection & Breach Readiness Guide (Guide),* Released Jan. 28, 2014. http://otalliance.org/breach.html

[6] Julie Bort, *How the CDC is Using Big Data To Save You From The Flu*. BUSINESS INSIDER**,** (Dec. 13, 2012). http://www.businessinsider.com/the-cdc-is-using-big-data-to-combat-flu-2012-12

[7] Location Privacy Protection Act of 2014. http://www.franken.senate.gov/files/video/140327stalkingapps.mp4 For the full text of the bill: https://www.govtrack.us/congress/bills/113/s2171/text

important that these safeguards do not impede data collection and sharing. Such technologies may be used for marketing and other purposes that can provide enhanced online experience or other consumer benefits. When this data is being collected, the user should be provided a clear understanding of the use of data, including any third party sharing, with an ability to opt-out.

OTA believes anonymous web analytics for research or operational purposes should be exempt from proposed legislation.[8] Web analytics is generally performed by third party service providers for the purpose of providing insight into industry trends. This research helps inform industry investments, website feature developments and facilitates efficient e-commerce and innovation. Such processed data is not used to target any individual or device via on or off line advertising or alter content viewed by the individual based on his/her individualized behavior and activities. Providers of such analytic services that aggregate, weigh, anonymize and otherwise process the collected data are in accordance with industry best practices.

(3) How should the policy frameworks or regulations for handling big data differ between the government and the private sector?

With the limited exception of national security, government entities should follow the same policies as the private sector. The government and private sector struggle to keep up with ever increasing sophisticated attack methods. This challenge is in part due to the lack of automated real-time information sharing. The sharing of threat intelligence information is a key arena for the government and private sector. Major players in both sectors must work diligently to usher in cooperation with the objective of improving protection against threats.

Threat intelligence sharing must move from individual silos of data to being shared cross sector to help identify trends. This shared data will help mitigate threats and aid in the prevention of cybercriminal activities.[9] To better facilitate communication among and between members of the private and public sectors, OTA believes we must universalize incongruent standards, procedures, data formats and abuse reporting.[10]

(4) What issues are raised by the use of big data across jurisdictions, such as the adequacy of current international laws, regulations, or norms?

Multinational companies must comply with often contradictory regulations due to disparate international laws and regulations. This engenders excessive technical investments, legal and consulting fees. OTA supports the White House commitment to pursuing international interoperability though the mutual recognition of commercial data privacy frameworks that incorporate effective enforcement and

[8] Examples include, but are not limited to, data used to optimize site operations, site traffic analysis and frequency capping.

[9] Internet Identity, *Sharing The Wealth, And The Burden, Of Threat Intelligence: Why Security Experts Must Unite Against Cyber Attacks And What's Stopping Them From Collaborating More Effectively.* (Sept. 2013) http://internetidentity.com/wp-content/uploads/2014/02/IID_Infoshare_WhitePaper.pdf

[10] For example, the Research and Education Networking Information Sharing and Analysis Center (REN-ISAC) is a private information sharing community that improves timely local protection against cyber security threats by sharing security-event data, in near-real time, within a trusted federation. http://www.ren-isac.net/ses/

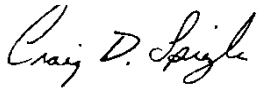
accountability mechanisms.[11] However, the United States lacks a comprehensive privacy framework, making it harder for U.S. companies and officials to argue credibly against overbroad and unworkable privacy regulations.

Data localization regulations that are currently being proposed in the European Union would contribute to the balkanization of the internet.[12] This would also negatively affect American businesses by increasing global operating costs or leading users to migrate towards non-US based services. Synchronization of data regimes will allow businesses to adopt more privacy protective schemes. Clear direction from the OSTP, accompanied with incentives and safe harbors, will support businesses in fulfilling and exceeding privacy requirements.

Conclusion

In summary, OTA looks forward to continued collaboration with all stakeholders in the ecosystem including online brands, technology providers, trade organizations, government agencies and advocacy groups who recognize the benefits and risks of big data. OTA recognizes that business accountability, data stewardship and data protection are all important privacy issues when considering implications of data collection. The goal of this discussion should be to maximize the potential benefits of data collection and the free flow of information while minimizing the privacy and related data security risks. The Online Trust Alliance thanks the White House Office of Science and Technology Policy for considering these comments.

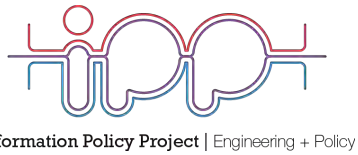
Sincerely,



Craig D. Spiezle
Executive Director and President
Online Trust Alliance
Craigs@otalliance.org
https://otalliance.org
+1 425-455-7400

[11] White House, *Consumer Data Privacy In A Networked World: A Framework For Protecting Privacy And Promoting Innovation In The Global Digital Economy* 31 (2012). http://www.whitehouse.gov/sites/default/files/privacy-final.pdf

[12] http://ec.europa.eu/justice/data-protection/index_en.htm



Consumer Privacy Bill of Rights and Big Data: Response to White House Office of Science and Technology Policy Request for Information

April 4, 2014

Daniel J. Weitzner, MIT Computer Science and Artificial Intelligence Lab
Hal Abelson, MIT Department of Electrical Engineering and Computer Science
Cynthia Dwork, Microsoft Research
Cameron Kerry, MIT Media Lab
Daniela Rus, MIT Computer Science and Artificial Intelligence Lab
Sandy Pentland, MIT Media Lab
Salil Vadhan, Harvard University

I. Introduction and Overview

In response to the White House Office of Science and Technology Policy Request for Information on Big Data Privacy we offer these comments based on presentations and discussions at the White House-MIT Workshop "Big Data Privacy Workshop: Advancing the State of the Art in Technology and Practice" and subsequent workshops co-sponsored with Data & Society and NYU Information Law Institute and the UC Berkeley iSchool.

1. Big data analytics offers significant new opportunities for advances in scientific research in many fields. Presentations offered at the MIT workshop showed unique benefits for improved healthcare quality, advances in the understanding of diseases through genomics research, potential to improve educational effectiveness, and more efficient, safe transportation systems.

2. There are real privacy risks raised by ubiquitous collection of personal data and use of big data analytic techniques. Key risks include:

- *Re-identification attacks*
- *Inaccurate data or models*
- *Unfair use of sensitive inferences*
- *Chilling effects on individual behavior*
- *Excess government power over citizens*
- *Large-scale data breach*

3. The White House Consumer Privacy Bill of Rights offers policy approaches to address each of these risks. Some of the principles such as transparency, respect for context, security, access, and accountability will play especially important roles in big data issues. Transparency should be augmented beyond just visibility into policies, to also enable individuals and regulators to see how personal data actually flows and is used. The respect for context principle should be implemented with particular attention to developing and enforcing limits on how personal data is used, especially in circumstances where collection limits and ex ante consent are difficult to achieve.

4. Technical contributions from computer science can assess and in some cases control the privacy impact of data usage in a rigorous, quantitative manner. But as technology will not replace the need for laws and social norms protecting privacy, basic and applied research must be conducted in a cross-disciplinary context so that technical designs will meet social policy needs.

II. Benefits and Risks Specific to Big Data

A variety of presentations at the White House-MIT workshop[1] illustrate significant new knowledge to be gleaned from analysis of large data sets. The large scale analytic results are possible because database technology innovation has made it easier and cheaper to collect, store, and analyze data. As Mike Stonebraker [slides] explained, computer science continues to develop new techniques for collecting and storing ever-expanding volumes of data, and analyzing these with increasingly fine-grained detail. While we can identify core big data analytic technology platforms and associated privacy enhancing technologies, it would be a mistake to conclude that all uses of big data technology implicate the same privacy values or, indeed, that a single set of privacy-enhancing technologies apply to all big data applications.

Computer science faces the challenge of storing and analyzing large amounts of data in part because so much more data is now being collected. Much of the digital infrastructure becoming ubiquitous around the world -- from mobile phone networks to transportation, finance and healthcare systems -- is trending toward recording nearly every action human beings take. John Guttag [video] showed how he was able to learn about the spread of hospital-acquired infection, but needed access to large amounts of patient data, plus personal information about un-infected patients as well as personal details about the doctors, nurses and other staff who were potential disease vectors, even if their interaction with the infected patients were marginal. Manolis Kellis [slides], a computer scientist working with large scale genomic data, explained the need to have large samples of the population to detect very small scale phenomena that, though widely dispersed in the data, are nonetheless critical to understanding the way the human genome functions. Sam Madden [slides] spoke about his Car-Tel experiment using mobile phones in cars that has enabled insights about how to make transportation networks more efficient and even

[1] The complete agenda, video of each presentation and slides and workshop summary report can be found at the workshop website. http://web.mit.edu/bigdata-priv/

reduce risky driving behavior by teenagers. And Anant Agarwal [video] showed how education researchers can use personal data from online courseware systems to learn about the most effective pedagogical techniques for different types of students.

All of this research requires access to large amounts of data. In some cases that data will have been collected for a purpose defined at the time of collection. But, in a big data environment looking for unanticipated relationships, it is not always possible to foresee at the time of collection all the questions that may subsequently be asked of the data. In some cases, it may be possible to apply new privacy-enhancing techniques such as differential privacy (see Section IV) to enable research with little or no privacy exposure, and thereby reduce the need for consent. However, many workshop participants expect that in some significant set of circumstances these techniques will not be applicable without real loss of useful research results. Researchers at the workshop generally expressed the view that obtaining consent after collection for specific new analytic tasks would likely be impractical in many instances and reduce the ability to gain useful knowledge from data. As will be explained below (section III. B), the Consumer Privacy Bill of Rights anticipates the need to protect privacy and guard against unwanted mission creep in situations in which individualized notice and consent is not feasible.

Acknowledging the social and economic benefits of large scale analytics, we must also recognize substantial privacy risks that have the potential to lead to real harms. Leading risks include the following:

1. *Re-identification attacks*: Data may be disclosed with certain identifiable information removed, but is still susceptible to being re-identified by correlating the weakly de-identified data with other publicly or privately held data.[2] Re-identification is not the only problem that can occur with release of "de-identified" data. Even without matching an individual to a specific record, but simply being able to conclude that the individual corresponds to one of a small number of records, or one of a possibly large set of records that agree on a particular attribute (HIV positive status, for example), can be harmful. These risk grows as the availability of data increases and the ability to correlate improves.
2. *Inaccurate data or models*: More data is by no means always a guarantee of more accurate results. Data sets may contain either inaccurate data about individuals and/or employ models that are incorrect at least as to particular individuals. This risk increases as larger datasets are to generate increasingly complex models which may be applied to decisions about individuals without rigorous validation. When decisions are made based on either inaccurate data or incorrect or imprecise models, individuals can suffer harm by being denied services or otherwise treated incorrectly.
3. *Unfair use of sensitive inferences:* There are numerous examples of big data analytics able to infer sensitive facts (creditworthiness, sexual orientation, spousal relationships,

[2] Narayanan, Arvind, and Vitaly Shmatikov. "Myths and fallacies of personally identifiable information." *Communications of the ACM* 53.6 (2010): 24-26. Sweeney L. Matching Known Patients to Health Records in Washington State Data. Harvard University. Data Privacy Lab. 1089-1. June 2013.

likely future location, or other information individuals may choose not to disclose) from correlation with other public or less sensitive personal information. Even if these sensitive inferences are accurate, it may be unfair to use these inferred facts about individuals to make certain kinds of decisions.

4. *Chilling effects on individual behavior*: Individuals may alter their behavior and expression in response to a sense of being watched. Awareness of increasingly ubiquitous data collection capable of deriving detailed conclusions about individuals through big data analytic techniques may have the effect of limiting individual participation in a variety of activities, including in constitutionally protected areas such as politics and religion.
5. *Excess government power over citizens:* The power of big data analytics can also be deployed to expand the amount of knowledge governments have about their citizens. Even though liberal democracies may impose limitations on their own collection, use, and retention of large data sets, their use of such analytics may facilitate more pervasive use by authoritarian governments
6. *Large-scale data breach*: Data breach is a continuous risk. As the scale of data collection, the flow of this data to different parties with access to this data, and the ability to derive detailed information all increase, the risk of harm from breach also grows. With big data, data thieves have a richer set of targets and tools available.

III. Guidance from the Consumer Privacy Bill of Rights

For each of the big data privacy risks identified here, the substantive principles in the Consumer Privacy Bill of Rights offer guidance to develop concrete responses to those risks in a manner that provides clarity for individuals and flexibility for innovative big data analytic applications. Given the rapid evolution of big data analytic applications, the unique procedural aspects of the Consumer Privacy Bill of Rights also offers a means by which principle-based privacy approaches to new applications can be developed rapidly as enforceable codes of conduct and then enforced under the FTC's existing statutory authority.

A. Overview of Consumer Privacy Bill of Rights Principles.

1. Individual Control

"Consumers have a right to exercise control over what personal data companies collect from them and how they use it."[3]

The principle of Individual Control in the Consumer Privacy Bill of Rights shifts the focus away from the longstanding principle of notice-and-choice to more dynamic and flexible mechanisms. Notice-and-choice is one important mechanism of privacy protection, but the Commerce Department Green Paper[4] process found that routine checking of boxes puts too much weight

[3] The White House, Consumer Privacy Bill of Rights (February 2012). [hereinafter CPBR]
[4] United States Department of Commerce, Internet Policy Task Force, Commercial Data Privacy and Innovation in the Internet Economy: A Dynamic Policy Framework (December 2010)

on the unmanageable burden of reading privacy policies and does not differentiate among situations that present material privacy risk and those that do not. Whether data is used in a commercial context, or for basic medical or scientific research, may also be relevant to what kind of individual control is warranted. The Consumer Privacy Bill of Rights therefore calls for contextual mechanisms to exercise choice at the time of collection "appropriate for the scale, scope, and sensitivity of the data in question," and also for additional mechanisms to address the use of personal data after collection.

This principle reflects the Big Data environment in two ways. First, it recognizes that the increasing velocity and variety of data collection make notice-and-choice ineffective; consumers are asked for consent too frequently and on devices such as mobile phones that are not suited to deliberate informed consent. Second, it recognizes that the velocity of data includes increased sharing with third parties with whom consumers do not have a direct relationship. Moving away from a one-size-fits-all notice and choice regime in which consumers often face a binary choice (either to give up data control or not to use a service) will strengthen fair exchange of value between consumers and companies by allowing consumers greater choices of how much to share in exchange for a given level of features and benefits.

There are certainly contexts in which individual control will play a minor role as compared to other principles such as Respect for Context and Focused Collection. The expanded use of sensors and other developing forms of automated data collection will make notice-and-choice and other mechanisms of control impossible or infeasible in an increasing number of circumstances. The principles of Consumer Privacy Bill of Rights are intended to apply in interactive and dynamic ways appropriate to the technologies they address; the expansion of Big Data will put a premium on such application.

2. Transparency

"Consumers have a right to easily understandable and accessible information about privacy and security practices."

The Transparency Principle requires companies to disclose when and why they collect individuals' personal data, so that consumers can guard against misuse of their personal data. Beyond just individual awareness, transparency has a vital function for the evolution of privacy norms themselves. In the modern history of information privacy, transparency has enabled consumer advocates, policy makers, enforcement agencies, the press and the interested public to engage in dialogue and criticism about how commercial privacy practices are evolving. It is only with awareness of actual privacy practices that society can have a meaningful dialogue about which practices are acceptable and which fall outside legal and/or social norms.

Meaningful transparency in big data systems will require going beyond just disclosure of policies as to personal data. Enabling citizens, governments and advocates to address big data privacy challenges requires a more active transparency - the ability to be aware of and track the actual

flow and use of personal information. Big Data is different from the regular use of personal data in that consumers are not only affected by the primary collection of data, but also by the subsequent aggregation of that data to inform algorithms that govern companies' decision-making and affect individual users. Therefore users need additional tools to help them follow complex data flows and understand what picture of them this data enables, beyond just the general disclosures in privacy policies. For example, disclosing to an individual that a health insurance company knows her address does not inform her of the likelihood that this information is being combined with multiple other databases to create "neighborhood profiles" that in turn could affect pricing for individual customers within those neighborhoods.

The requirement that companies detail how they will use gathered data bears more weight in the Big Data context than requirements that companies simply disclose what personal data they collect. A transparency framework that updates consumers when companies come up with new uses for aggregated personal data will increase user trust and help ensure that accountability takes place at the rate of business growth, rather than at the rate of governmental enforcement.

3. Respect for Context

"Consumers have a right to expect that companies will collect, use, and disclose personal data in ways that are consistent with the context in which consumers provide the data."

The principle of Respect for Context builds on the recognition that expecting users to read notices and make choices for every single individual collection and use of personal data element is unsustainable. Requiring users to make such a large volume of choices is not fair to the individual. In the Big Data environment of unstructured data and unforeseen correlations, rigid insistence on notice and choice would also render impractical many of the applications that can result in important new scientific discoveries, more efficient public infrastructure, and innovative new commercial services in a Big Data environment . The Respect for Context principle therefore recognizes that consent can be inferred in some circumstances and that privacy protection will depend on ensuring that the *uses* of personal information are faithful to the original context in which the individual provided the data (whether actively such as in a transaction, or passively such as via sensing devices). Giving users the right to expect that the context of collection be respected in further uses will protect individuals from unwanted surprise, and at the same time allow the development of valuable new big data applications.

4. Security

"Consumers have a right to secure and responsible handling of personal data."

Collecting, storing and using personal data comes with inherent risks, including the possibilities of privacy loss and data theft, modification or destruction. The security principle recognises this, calling on companies to assess these risks and maintain reasonable safeguards, because without them data-driven economic growth will be limited by a lack of trust. In addition, the risks

of data re-identification need to be investigated and mitigated as much as possible. When multiple anonymous, heterogeneous databases are combined, prediction algorithms can be used to re-identify individuals. While the security principle does not explicitly recognise de-anonymization risks, they deserve special mention because they are extremely difficult to assess in practice, especially given the unpredictability of additional sources of information that may be employed in the attempt to de-anonymize. Currently, these risks are assessed by simply attempting to de-anonymize the data: clearly, new innovations are needed to address this tough problem.

5. Access and Accuracy

"Consumers have a right to access and correct personal data in usable formats, in a manner that is appropriate to the sensitivity of the data and the risk of adverse consequences to consumers if the data is inaccurate."

The Access and Accuracy principle requires consumers to have the ability to access and correct personal data in usable format. It addresses Big Data by recognizing that the opportunity to access and correct personal data is especially important where" [a]n increasingly diverse array of entities uses personal data to make decisions that affect consumers …." With such access and ability to correct, errors or inaccurate data can be more easily discovered and detected due to possible contradictions with other datasets. More accurate inferences can be made about consumers with a less likelihood of misinterpretation of information itself. In turn, these inferences could also be used in discovering anomalous pattern within the different datasets and hence be useful in understanding and detecting possible additional errors.

6. Focused Collection

"Consumers have a right to reasonable limits on the personal data that companies collect and retain."

Large-scale undirected collection of information that is seemingly unrelated to the main use of an application exposes consumers to unnecessary risk, as even collection of seemingly innocuous data may allow unwanted intrusions or inferences about sensitive details of a person's life.

This CPBR principle is a conscious move away from the data minimization principle. It recognizes that Big Data applications often make use of wide-ranging data sets related to the user that do not have obvious initial value. Hence, the principle does not demand absolute minimization. Declining costs for pervasive sensing devices such as health trackers and mobile phones has led to increasingly broad collection of data, while the cost per bit of storing such information has dropped. These combined factors incentivize companies to collect and store such information indefinitely, regardless of that data's usefulness, just in case future exploitation of such data might prove useful. The Focused Collection principle allows for collection of large data sets, but as a check against unrestricted and possibly unreasonable collection practices, it calls for thoughtful decisions about what data to collect or retain and new innovation, both in the

policy and technical arena, to enable companies to target their collection.

7. Accountability

"Consumers have a right to have personal data handled by companies with appropriate measures in place to assure they adhere to the Consumer Privacy Bill of Rights."

The Accountability principle recognizes the necessity for organizations using personal information to have strong training, internal policies, and audit mechanisms to assure compliance with legal requirements and, more broadly, wise and responsible use of the data they hold. The MIT workshop revealed a number of big data scenarios for scientific research in which the benefits of large scale analytics can only be achieved by allowing wide-ranging internal analysis of the data, with controls on the ultimate use of the data. Therefore, it is especially important that organizations employing these techniques have clear internal policies to prevent against misuse of data, that employees within the organization have a clear sense of what these policies mean, and that audit mechanisms are in place to help guard against either intentional or unintentional misuse.

Applying state-of-the art approaches to institutional compliance, the most important thing is execution - achieving actual compliance with rules or establishing why noncompliance has occurred. To that end, it's important to distinguish objective performance measurement or diagnosis from normative assignment of responsibility (or - even more so - blame, liability, or culpability). The latter invite a backward-looking defensive posture that gets in the way of honest appraisal of the facts and forward-looking improvements. That appraisal is more effective if individual responsibility is secondary.

B. Addressing Risks

Consider how each of the Big Data privacy risks identified above can be addressed under the framework of the Consumer Privacy Bill of Rights.

1. Re-identification risk: The risk that personal data can leak from big data research platforms is real. Principles including Transparency, Security, Focused Collection, and Accountability will all be important to manage this risk. T*ransparency* will enable regulators, enforcement authorities, and interested members of the public such as advocates and academics to know what kind of data is being released and in what form. Assessing whether the users' rights to have data held *securely* should include an assessment of who is able to access the data and therefore whether the re-identification risk can be minimized by binding those individuals to legal commitments to avoid re-identification. The right to have only *focused collection* of user data will also reduce re-identification risk by limiting gratuitous collection of data. And finally, an organization with strong institutional *accountability* procedures in place should handle data carefully and only release it publicly after evaluating the risk of re-identification. If the organization fails to consider this risk, then appropriate parties can be held accountable for resulting harm.

2. Data and model inaccuracy: The Consumer Privacy Bill of Rights can reduce the risk that decisions are made about an individual based on inaccurate information or an incorrect model. The principles of Transparency, Respect for Context, and Access and Accuracy are all useful to ensure fairness in big data decisionmaking. Since the Fair Credit Reporting Act (FCRA) was enacted, individuals have had basic *transparency* rights enabling them to know that personal information about them is being used for important decisions, as well as the right to *access* and correct personal data to ensure that it is *accurate*. Such transparency is critical to make sure that individuals know their data is being used therefore be able to assure its accuracy or decide to exclude themselves from uses they object to.

Similarly, the Access and Accuracy affords a mechanism to assure that data and the inference drawn from are accurate.[5] While most consumers will not be able to identify errors in models, transparency on inferences drawn by a model may shine light on algorithmic errors,

The CPBR Transparency principle also requires that companies explain how they will use data and this should be understood to include relevant information about the decisionmaking models and algorithms. There is work to be done to define how much of the decisionmaking metrics should be exposed, as some of that information will be proprietary. Enough context about the decision metrics should be made available to enable consumer protection enforcement agencies and other stakeholders to assess whether the decisions models are fair.

These principles are reinforced by the Respect for Context principle. When data is used out of context, the CPBR provides that if "companies decide to use or disclose personal data for purposes that are inconsistent with the context in which the data was disclosed, they must provide heightened measures of Transparency and Individual Choice." This will help individuals to flag uses of information that are likely to create risk and increase the likelihood that both personal data and models derived from personal data are accurate.

3. Unfair use of sensitive inferences: Even if inferences are accurate, it may be unfair as a matter of ethics or public policy to use such information for certain purposes. For example, behavioral profiling techniques used for marketing purposes can provide advertisers the ability to reach audiences defined by age, ethnicity, race, gender and other sensitive categories. The recent statement of privacy principles from leading civil rights organizations ("Civil Rights Principles for the Era of Big Data") offers useful guidance on this point.The Respect for Context principle was specifically designed to prevent misuse of such profiles for more sensitive, harmful discriminatory purposes. As the CPBR explains:

[5] "An increasingly diverse array of entities uses personal data to make decisions that affect consumers in ways ranging from the ads they see online to their candidacy for employment. Outside of sectors covered by specific Federal privacy laws, such as the Health Insurance Portability and Accountability Act (HIPAA) and the Fair Credit Reporting Act, consumers do not currently have the right to access and correct this data." CPBR p20

> The Administration also encourages companies engaged in online advertising to refrain from collecting, using, or disclosing personal data that may be used to make decisions regarding employment, credit, and insurance eligibility or similar matters that may have significant adverse consequences to consumers…. Such practices also may be at odds with the norm of responsible data stewardship that the Respect for Context principle encourages."[6]

Just because it is possible to learn or infer a sensitive characteristic of an individual, that does not imply that it is either legally or ethically permissible to use such an inference (no matter how accurate or inaccurate) for all purposes. However, addressing the use of such characteristics is a matter of social policy broader than privacy policy. Antidiscrimination laws and norms of countries around the world regularly prohibit acting in a discriminatory manner based on information about an individual, even if it is publicly available. Indeed, some of the personal characteristics that entail the highest degree of legal concern include gender and race, attributes of individuals that are readily observable and in most cases public information.

The Transparency and Access Accuracy principles provide mechanisms that can be helpful in identifying where data collected about individuals is used in ways contrary to legal or ethical principles. Despite this, reflexive and poorly justified application of the Fourth Amendment third party doctrine can lead to the "unwarranted" assumption that as soon as personal data is public it can be used for any purpose. The Respect for Context principle stands in opposition to this view and squarely for the proposition that privacy interests in personal information are determined as much by how the data is to be used as is the public or non-public status of the data.

4. Chilling effects on individual behavior: Among the paramount constitutional concerns at the heart of privacy is protection of the freedom of association enshrined in the First Amendment of the US Constitution. That is why President Obama introduced the Consumer Privacy Bill of Rights by recognizing the importance of upholding individual freedom of association:

> Citizens who feel protected from misuse of their personal information feel free to engage in commerce, to participate in the political process, or to seek needed health care. This is why we have laws that protect financial privacy and health privacy, and that protect consumers against unfair and deceptive uses of their information. This is why the Supreme Court has protected anonymous political speech, the same right exercised by the pamphleteers of the early Republic and today's bloggers.[7]

The CPBR can protect against chilling effects in the first instance through the Individual Control principle. Citizens who feel in more control over their personal data will feel more free to engage in activities online and offline. Transparency is critical to help assure individuals that they have

[6] CPBR p 18
[7] President's Preface to the Consumer Privacy Bill of Rights. February 23, 2012.

some understanding of when their personal data is collected and how their data is used. Respecting the context in which personal information is collected and avoiding out-of-context uses will limit the degree to which individuals are surprised but subsequent data usage and increase trust. Finally, knowing that they the right to access and correct personal data will reduce mistrust and fear that data could be inaccurately used against an individual's legitimate interest.

5. Excess government power over citizens: Large scale analytics can expand government investigative power by revealing otherwise hidden knowledge about social relationships, political action, and other details of citizens' First Amendment-protected associative or expressive activity. Traditionally, civil liberties concerns about undue expansion of government surveillance and data gathering power are addressed in the United States by limiting government action with respect to collection of private information and property through the Fourth Amendment protections against unreasonable search and seizure. Today, however, enhancement in the government's law enforcement and national security investigative power are driven by big data analysis of information that often originates with private sector organizations such as network operators and Internet edge services. The primary venue for deciding on the proper scope of government surveillance power should be norms of government action that are beyond the scope of the Consumer Privacy Bill of Rights.

As much of the data in question originated with the private sector, several principles in the CPBR are relevant to managing the risk of that private collection of information contributes to excessive government power. *Individual choice* will help uses avoid having personal data collected if they are engaging in activities that they would want to remain outside government visibility. *Transparency* will help individuals make those choices and *Access and Accuracy* will make sure that personal data that does come into government hands is accurate. *Security* will help prevent surreptitious surveillance.

6. Large-scale data breach: With big data comes risk of bigger data breach. The CPBR *Focused Collection* principle can have some role in limiting harm when data that was never actually needed is subject to a breach. Most importantly, organizations that collect, use and store large repositories of personal data must follow good security practices that "best fit the scale and scope of the personal data that they maintain." Furthermore, the Administration's call for Federal Data Breach notification law is all the more important give the increased security risks from the growing size and scope of big data repositories.

C. Applicability of the Consumer Privacy Bill of Rights to Big Data privacy challenges

Large scale analytics has long been a factor in privacy policy, beginning with the enactment of the Fair Credit Reporting Act of 1970, the law that was written to regulate the leading big data enterprise of that era, the consumer credit bureaus. The credit reporting agencies took the then-unprecedented and to many, quite alarming step of collected detailed transactional data on

the financial life of a significant proportion of the adult population in the United States and then subjecting it to sophisticated analysis for the purpose of developing credit risk scores. So the challenge of addressing large-scale integration of data used for purposes that an have a real impact, positive or negative, on the lives of individuals, is not new. We therefore see no reason to abandon time-tested privacy principles and have shown how the modernized version of these principles in the Consumer Privacy Bill of Rights can apply to big data analytics.

Addressing privacy challenges given the "velocity, volume and variety" that characterizes new big data systems does call for greater reliance on some of the principles in the CPBR than others. We have emphasized transparency, respect for context, security, access and accuracy, and accountable as elements of the CPBR that will bear significant weight in big data privacy protection.

First and foremost, an expanded commitment to *transparency* is necessary to guard against the risk of unfair, inaccurate use of personal data. The variety of personal information in big data systems requires a more active transparency in which individuals, consumer advocates and enforcement agencies can understand precisely how personal data is used, in some cases with resolution down to the level of individual data elements. Recognizing that individual control and consent may not be practical for high velocity collection and use of personal data, such systems will place more reliance on respect for context, assuring the information is only collected where the context makes such collection reasonably apparent, and that the use is consistent with the original context of collection. In context use should be able to proceed without individual consent, but out of context use would require increased transparency and individual control.

Large collections of personal data create increased risk of breach and loss, to security must be given special attention. As important decisions may be made through big data systems, access and accuracy rights are vital to be sure individuals are not treated unfairly. And finally, institutional accountability mechanism are vital to assure that all of the principles in the Consumer Privacy Bill of Rights are adhered to the use of big data systems.

Beyond just the substantive principles of the Consumer Privacy Bill of Rights, the larger policy process that the Administration's privacy framework puts into place has a dynamic, flexible quality that will be especially important to help American society evolve new privacy norms in response to the challenge of large scale analytics. As explained by Ken Bamberger and Deirdre Mulligan, the evolution of 'privacy on the ground'[8] has enabled the evolution of privacy rules in a manner that is responsive to public requirements while at the same allowing flexibility for the development of new services and business models. The Consumer Privacy Bill of Rights framework is designed to facilitate the continuous evolution of norms and rules as large-scale analytics drive new business models.

[8] Bamberger, Kenneth, and Deirdre Mulligan. "Privacy on the Books and on the Ground." *Stanford Law Review* 63 (2011).

IV. Research agenda based on technical and policy observations from workshops

Technical contributions from computer science can assess and in some cases control the privacy impact of data usage in a rigorous, quantitative manner. These techniques can help assure that systems handling personal data are functioning consistent with desired public policies and institutional rules. In some cases, these controls can prevent disclosure or misuse of data up front. In other cases, systems can detect misuse of personal data and enable those responsible to be held accountable for violating relevant rules. These technologies are at various stages of development, some ready for broad deployment and others needing more research to enable practical application. Developing the technological base that enables people to be in control of their data and assure it is used accountably is a key challenge. Solutions to this challenge are feasible.

The privacy risks we have identified above (III.B) along with the principles in the Consumer Privacy Bill of Rights provide guidance for shaping a research agenda to expand the established theoretical foundations of privacy enhancing technologies, to integrate them into systems design, and to develop public policy models for big data that address these risks in line with CPBR. A five-prong, cross-disciplinary research agenda is called for:

- Theory: Work in cryptography and theoretical computer science provides the foundation for privacy-sensitive controls on how personal data can be accessed and computed with. Techniques such as fully homomorphic encryption enable computation on data while it remains in encrypted form, thus reducing the risk that the data could be released in raw form, or that partially de-identified data could be re-identified. Secure multiparty computation and functional encryption can allow those who seek to use data to perform specifically-approved computation on personal data but limit access only to execution of those functions. Finally, differential privacy provides a framework and tools for enabling statistical analysis of data while ensuring that personal information does not leak. Continued work on the theoretical foundations of personal information may yet yield new techniques beyond these.

  While the theoretical foundations for a number of the above approaches are well-established, comments from a number of presenters and participants at the MIT workshop reveal the need to support research to scalability and models for computational tractability of these techniques.

- *Systems*: Systems research provides a variety of promising avenues to create data architectures that hold personal data while providing for a variety of privacy protections. Encrypted databases such as CryptoDB allows queries directly into encrypted data stores, thereby increasing the security of personal data while allowing access for queries

by authorized users. Continued research in this area can make such database architecture more robust, scalable, and offer improved security properties. Accountable systems instrument traditional databases and other structured data repositories,tracking the flow and use of personal information in order to assess compliance with rules governing that data. Using knowledge representation, formal reasoning systems, and linked data architectures, accountable systems can provide a scalable means of helping data users to comply with known rules, and to demonstrate to the public that data is being used only for specified purposes and that there is accountability for violation of privacy rules. Further research in this area will increase the expressivity of the policy languages used to represent rules and explore new reasoning techniques that scale with increased data volume and variety.

- *Human Computer Interaction:* Many of the principles in the Consumer Privacy Bill of Rights (Individual Control, Transparency, Respect for Context, Access and Accuracy, to name a few) seek to give individual users greater awareness and control over their personal information relationships with others. Researchers, consumer advocates and companies are all aware of the ongoing challenge of designing systems to enable users to understand and control their personal data. As the scale and complexity of personal data usage grows, these problems will be all the more challenging. Research into user experience design, machine-assisted assessment of context, and good security mechanisms will be critical to support the efforts of end-user systems designers.

- *Policy*: Technology will not replace the need for laws and social norms protecting privacy or the use of personal information. We should expect that systems are built to perform according to privacy rules and make enforcement of those rules easier, especially as the scale of data usage increases beyond the point that manual compliance and audit can be effective. Policymakers also have much to learn from advances in technology design. Scientifically-informed privacy policy research can help guide policymakers to understand how to take best advantage of privacy-enhancing technologies. While broad privacy principles and risks may be clear, there are still conceptual questions about how to apply these principles to big data systems, challenges in human computer interaction to design privacy enhancing systems and better understanding of how privacy enhancing technologies might actually be used at Web scale.

- Technology/Policy Integration: In each of these research areas, multi-disciplinary collaboration will be critical. Designing and developing algorithms and systems to enhance privacy cannot succeed as a monolithic technical exercise. However, throughout the technical and policy discussions at each of the three workshops, a variety of implicit and explicit definitions of privacy were used. Some implied that privacy is synonymous with secrecy and complete confidentiality – that as soon as personal information is available to anyone else, privacy is lost. Others suggested that privacy is properly understood as the ability to control how personal information is disclosed and/or used. Finally, privacy is understood by some as a question of whether personal data is

used in a manner that harms the individual.

One example of a technology policy integration is the Living Lab is being deployed at MIT. The goal of this project is to concretely explore privacy and big data policies and their effects, and to promote greater idea flow within MIT. Software tools such as our openPDS (Personal Data Store) system and Accountable Systems tools gives people the ability to have active transparency and control over where their information goes and what is done with it. Data sharing maintains provenance and permissions are associated with data, and automatic, tamper-proof auditing is supported. This allows enforcement and compliance with information usage rules and helps to minimize the risk of unauthorized information leakage.

Realizing the goal of designing systems that do a better job of respecting privacy requires continued research and dialogue with a wide range of disciplines that come together to define and refine privacy requirements. As those requirements are sure to be context driven, expertise from the wide range of privacy contexts will be required for successful research efforts.

V. Conclusion

Meeting big data privacy challenges requires social and legal consensus on how our fundamental privacy values apply in this new context. There is much scientific, social and commercial value to be realized. As members of the academic community, we are committed not only to expanding our research efforts so that technical tools for privacy protection keep up with the rapid pace of large scale data analysis, and but also to playing our part in the evolution of the legal rules and social norms that are necessary to maintain public trust in this arena.

April 4, 2014

Nicole Wong, Esq.
Big Data Study
Office of Science and Technology Policy
Eisenhower Executive Office Building
1650 Pennsylvania Avenue NW
Washington, DC 20502

Via e-mail: bigdata@ostp.gov

Re: Public Comments, Big Data RFI

Dear Ms. Wong:

The Coalition for Privacy and Free Trade submits these comments in response to the White House Office of Science and Technology Policy's (OSTP) Request for Information dated March 4, 2013. The Coalition is an organization promoting cross-border data flows and privacy through interoperability of differing national legal frameworks. The Coalition's policy goals reflect input from business as well as academics and experts working in the fields of privacy and international trade.

Big Data holds promise for society in many different ways. At the same time, there are privacy challenges related to the collection and use of big data. As the Administration explores ways to address the privacy challenges, we urge a focus on cross-border data flows that will be essential to the full-realization of the benefits of Big Data globally.

There is a danger of overly-restrictive privacy regimes unnecessarily impeding cross-border data flows and, as a result, impeding the realization of the full benefits of Big Data. As the global information technology landscape shifts to account for "Big Data," it is imperative that stable policy and legal mechanisms develop to account for and protect the cross-border mobility of data. These new mechanisms should be built on tested frameworks and ensure interoperability between the privacy regimes of diverse jurisdictions.

Thus, we urge the Administration to focus on ways to ensure cross-border data flows and the interoperability of privacy frameworks as a way to ensure continued international cooperation on realizing the potential of Big Data.

Thank you for your consideration.

Sincerely yours,

Christopher Wolf
On Behalf of the Coalition for Privacy and Free Trade

Big Data in Private Sector and Public Sector Surveillance

Recent years have seen an explosion in the popularity of big data. This popularity is attributable to a variety of reasons, including the easier collection of data points by computers and the affordability of massive storage devices and computer processing power. Big data has become a trendy catchphrase for the idea that large datasets can often be used to learn interesting relationships that are not obvious at first glance, or that might not be evident in smaller datasets. Unfortunately, public policy has not kept pace with the rising popularity of big data, resulting in dangers to consumer privacy when big data is used by the private sector, and to constitutional norms and rights when used by the government.

Not all uses of big data implicate dangers to privacy or rights, such as datasets that are not about people or what they do. Even when the datasets concern people, such as the analysis of a dataset on health information to find previously unknown links between diseases or the analysis of a traffic dataset to aid in urban planning, analysis aimed at generating insights about large populations may pose relatively less risk than analysis aimed at classifying, sorting, or focusing on particular individuals or groups. Of course, there is always the risk that a data breach or dissemination of the underlying dataset could expose individuals' personal information (even if the desired analysis does not). If the type of analysis that will be done is known ahead of time, however, then it may be possible to mitigate this privacy risk through techniques that scrub or anonymize the data in such a way that only data relevant to the desired statistics remain.

The larger threat to privacy comes when big data is used to individually target people in a certain group found within a dataset. As an example, consider Target's development (using data from its baby registry) of an algorithm that analyzes someone's purchases in order to determine if they are pregnant, and the subsequent use of that algorithm to individually target people not in the registry with baby-related advertising.[1] By running algorithms on its customer dataset, which included looking for, among other variables, customers purchasing unscented wipes and magnesium supplements, Target's use of big data to identify pregnant customers raises questions, like: Is it ethical to analyze its baby registry for a purpose that its customers probably did not know about? To develop an algorithm for identifying potentially pregnant customers knowing

[1] Duhigg, Charles. "*How Companies Learn Your Secrets.*" New York Times, February 16, 2012.

that they probably do not want to be identified? To advertise to them? To potentially offer its pregnancy assessment service to employers, insurers, or others? In the biomedical field, ethical guidelines and practices like the Belmont Report and the Federal Policy for the Protection of Human Subjects seek to protect individuals' interest in respect and autonomy.[2] Collection and privacy standards must be recognized when it comes to big data as the information collected (and insight gleaned) can reflect some of the most intimate details of a person's life.

This concern is even greater with respect to the increasing use of big data for government surveillance, such as the government's use of Section 215 of the Patriot Act to collect all Americans' calling records and Section 702 of the Foreign Intelligence Surveillance Amendments Act to indiscriminately collect users' phone calls and emails. There are also highly secret uses of surveillance authorities like classified National Security Policy Directives (NSPDs) and Executive Order 12333 ("EO 12333"). Such "authorities" are collecting data in a similar manner to Section 215 and Section 702. They are used to create massive datasets containing information concerning US and non-US persons. This data includes personal identifiers, sensitive information, personal communications, and potentially other data that may be commingled with data collected under the Foreign Intelligence Surveillance Act (FISA). Such use of big data is difficult to reconcile with the idea of privacy not only because much of the data is collected in secret without the predication required by the Fourth Amendment, but also because the analytics seek to identify particular individuals. Such uses raise the greatest public policy concern and deserve more government and public attention.

Executive Summary

The Electronic Frontier Foundation appreciates the opportunity to submit comments for the Office of Science and Technology Policy's Big Data RFI, OSTP-2014-0003-0001. Our comments focus on certain parts of the first four questions, and we specify what we address here, along with very brief summary answers:

(1) What are the public policy implications of the collection, storage, analysis, and use of big data? For example, do the current US policy framework and privacy proposals for protecting consumer privacy and government use of data adequately address issues raised

[2]See http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html

by big data analytics?

Big data has serious public policy implications for privacy and fairness. We do not dwell on the potential benefits here, except to warn that good public policy should: (a) not overestimate potential benefits while discounting potential harms; (b) recognize that potential benefits and harms must be assessed within a realistic understanding of economic and political incentives; and (c) recognize that time matters. Because (a) is obvious, we elaborate on (b) and (c).

Realism about incentives is another way of saying that just because something can be done doesn't mean that it will be. We do not expect companies to use big data altruistically; we expect them to use big data to compete, profit, and grow. If big data gives an insurance company an incentive to eliminate more costly policyholders it is reasonable to assume that it will do so. We may wish for government to act in the public interest, but visions of the public interest vary greatly, and government agencies with missions like law enforcement or intelligence present special tensions, especially for constitutional values like due process, transparency, and democratic accountability.

To say that time matters is to highlight not only the way that law and policy often trail technological change, but also path dependence. One might think of this as the macro-version of "privacy by design." It will be much harder to protect privacy when business models or government programs become entrenched in their use of big data.

Fundamentally, the privacy and fairness implications of big data are closely tied to concerns about social, economic and political power. Both the public and private sector have strong interests in collecting and using data about individuals, whether for business or for social control. Since 9/11 it has become more obvious how much data collected by one sector flows to the other; the telephony metadata program using Section 215 of the Patriot Act is merely one especially obvious example. Power and inequality is, of course, an enduring issue for our nation, but big data seems especially troubling because large, powerful entities are more likely to have access to extremely large and complex datasets, and the sophisticated computing power needed to analyze them.

(2) What types of uses of big data raise the most public policy concerns? Are there specific sectors or types of uses that should receive more government and/or public attention?

Data about rocks is different from data about people or what people do. Analysis of the latter always raises potential ethical issues, which is why much research in the medical arena has traditionally been guided by human subjects protocols and legal-ethical constructs like the Common Rule.[3] We attempt to distinguish between big data uses that focus on large populations and those that focus on small groups or individuals, believing that the latter, roughly speaking, is of greater concern. We also believe that big data programs that operate secretly or with little visibility and transparency demand more attention. Accordingly, we believe that the use of big data in the national security and law enforcement realm deserves the greatest scrutiny. Private data collection remains in scope, of course, because many intelligence programs rely heavily on data collected by the private sector.

(3) What technological trends or key technologies will affect the collection, storage, analysis and use of big data? Are there particularly promising technologies or new practices for safeguarding privacy while enabling effective uses of big data?

We do not address this question below, and merely note two points. First, improved sensing or data collection technology obviously exacerbates the big data problem. More of what people do is capable of being captured, often without their awareness or their ability to avoid collection. Second, EFF is exploring the potential use of differential privacy techniques by large California utilities that collect granular energy usage data generated by smart meters. Federal smart grid incentives, we believe, illustrate a basic big data policy problem: smart meter deployment was stimulated with little consideration of the privacy risks or of how privacy might be designed into the smart grid. Indeed, much of EFF's role in California's utility regulation progress has been to help create a privacy framework for energy usage data in the face of strong government, commercial, and academic demand for this highly revealing data. EFF and its technical experts faced considerable political resistance even as we struggled to educate policymakers and stakeholders about the re-identification risks associated with granular data.

[3] See http://www.hhs.gov/ohrp/humansubjects/commonrule

(4) How should the policy frameworks or regulations for handling big data differ between the government and the private sector? Please be specific as to the type of entity and type of use (eg, law enforcement, government services, commercial, academic research, etc.).

First, government use of big data is inherently subject to constitutional constraints, while private sector use of big data is typically subject only to statutory constraints, with two significant caveats. In California, for example, private actors are subject to the state constitutional privacy right. And even under the federal constitution, private actors can in some circumstances violate individual rights under the state action doctrine. Of particular importance are the predication and particularity values of the Fourth Amendment, the due process values of the Fifth Amendment, the reasoned elaboration values of Article III courts and the democratic accountability of the Constitution itself.

Second, the policy framework for law enforcement and intelligence uses of big data is distinguishable from most other contexts by its lack of transparency. Obviously, law enforcement and intelligence agencies typically collect data in secret and without the consent of the people being surveilled. Secrecy also interferes with public knowledge about these surveillance practices and technologies. Particularly in the intelligence realm, the system of classified information and the state secrets privilege distorts normal processes of democratic accountability essential to legitimate constitutional government. And because these law enforcement and intelligence agencies often rely on data collected by the private sector, these distortions also directly affect individuals' trust relationships with business. Any big data policy framework must publicly address government secrecy and over-classification.

(5) What issues are raised by the use of big data across jurisdictions, such as the adequacy of current international laws, regulations, or norms?

That NSA surveillance is conducted both domestically and globally has never been a secret, but recent revelations have made the international nature of signals and communications intelligence impossible to ignore. NSA surveillance is obviously not simply a US problem. The relationship between NSA and GCHQ is clearly intimate, and many other governments in some

way partner with the United States. Genuine accountability as to the intelligence community's use of big data will, at the very least, require accountability as to these data flows and partnership arrangements. After all, if NSA and GCHQ share data about each other's citizens, it would make no sense to control NSA but not GCHQ. More generally, it appears that the various national intelligence agencies have developed their own norms of bulk collection in direct conflict with non-intelligence norms of predicated and particularized seizure or collection of communications. The question may not so much be the crossing of a national border but rather intelligence agency exceptionalism. When the United States attempts to justify secret and unaccountable mass surveillance programs with no credible evidence of utility, it can hardly criticize other nations for doing so.

Big Data Facilitates Private Sector Surveillance

The collection and analysis of big data, which was a niche field within computer science just two decades ago, has exploded into a $100 billion industry.[4] Big data is now used in sectors as diverse as energy, medicine, advertising, and telecommunications. Because of the explosive growth of this field, companies ranging from startups in Silicon Valley to established multi-national corporations are adopting the mantra of "collect it all," in the belief that running a variety of analytics on big data will increase the value of their products or the companies themselves.

In many cases companies outsource the use of big data to intermediary entities known as data brokers, which collect, analyze, and sell consumer information that can include highly personal details like marital status, religion, political affiliation, tax status, and others. A website may have an agreement with a data broker to better identify who their customers are so they can place more effective ads—often in exchange for their customers' browsing habits and demographic information. Data brokers receive and aggregate consumer data from a variety of sources: transactional data from retailers and stores, loyalty cards, direct responses and surveys, social media and website interactions, public records, and more.[5] They then aggregate this information across sources and use it to create highly detailed profiles about individuals—one

---

[4] *"Data, data everywhere."* The Economist, Feb. 25, 2010. https://web.archive.org/web/20131207192955/http://www.economist.com/node/15557443. Last accessed March 28, 2014.

[5] See Dixon, Pam. *"What Information Do Data Brokers Have on Consumers?"* World Privacy Forum, December 18, 2013. Last accessed March 30, 2014.

particular data broker is said to have 1,500 data points on over 700 million individuals.[6] It's been revealed that these highly detailed profiles include names like "Ethnic Second-City Strugglers," "Rural and Barely Making It," and "Credit Crunched: City Families," as well as sensitive lists such as police officers and their home addresses; lists of rape victims; genetic disease sufferers; and Hispanic payday loan responders.[7] The vast majority of information data brokers use to create these lists is data which consumers unintentionally expose in large part because they simply do not know how or when they are being tracked, or what information is being collected. As a result the information is almost perfectly asymmetric: brokers know a great deal about consumers, but most consumers have no idea these parties actually even exist.

This asymmetry is related to the first harm consumers are exposed to as a result of private-sector big data usage, namely the significant power imbalance between consumers and the companies wielding the data and analysis tools. For example, if a company uses big data analysis to inform its hiring decisions (say by analyzing a database on the web browsing habits of potential employees acquired from a data broker), would a rejected prospective employee learn why she was not offered a job, be able to see the data that led to the decision or the algorithm that processed the data, or dispute the correctness of either?[8] In general, the fact that people may be treated differently based on data and algorithms that they know little about and have no recourse for correcting creates elementary fairness and transparency problems.[9]

A related problem results from the fact that even if consumers are aware of what data they are providing about themselves and who they are providing it to, they frequently believe wrongly that the law or a company's privacy policies block certain uses of that data or its dissemination. As explained by Chris Hoofnagle and Jennifer King in their study of Californians' perceptions of online privacy:

---

[6] See Brill, Julie. *"Demanding transparency from data brokers."* The Washington Post, August 15, 2013. http://www.washingtonpost.com/opinions/demanding-transparency-from-data-brokers/2013/08/15/00609680-0382-11e3-9259-e2aafe5a5f84_story.html. Last accessed March 30, 2014.

[7] See Dixon, Pam. *"What Information Do Data Brokers Have on Consumers?"* World Privacy Forum, December 18, 2013. Last accessed March 30, 2014.

[8] One could argue that it would be in a company's best interests to use data that is as accurate as possible. However, a company's ultimate goal is to be as profitable as possible, and big data analysis is only carried out to further that goal. No rational company would acquire better quality data when the cost of doing so would be greater than the estimated returns. This exposes the fundamental mismatch in incentives between companies (whose big data will only be as accurate as profitability dictates) and individuals (who primarily care about whether the data about they themselves is accurate). Even a competitive market might not be able to completely resolve this issue, since making sure all the data is accurate 100% of the time will likely require human-intensive, and therefore costly, dispute/redress processes.

[9] Dwork and Mulligan, "It's Not Privacy, and It's Not Fair," 66 STAN. L. REV. ONLINE 35 (2013).

> Californians who shop online believe that privacy policies prohibit third-party information sharing. A majority of Californians believes that privacy policies create the right to require a website to delete personal information upon request, a general right to sue for damages, a right to be informed of security breaches, a right to assistance if identity theft occurs, and a right to access and correct data.[10]

Additionally, users may not know to what extent data is shared with unknown third-parties: an online project called "theDataMap" reflects this data-sharing landscape.[11]

But even a good understanding of the legal and policy protections for data is insufficient to protect a consumer from harm, due in large part to the next danger: loss of privacy due to individualized analysis and tracking by private-sector use of big data. By “connecting the dots” between different, disparate datasets, or even by analyzing data from the same dataset that on its face does not seem to have any connection, companies can infer characteristics about people that they might not otherwise wish to be made public, or at least not wish to share with certain third-parties (for example, the well-known Target pregnancy example). Very few consumers realize the power of statistical analysis and other big data algorithms. Even if consumers are aware of what specific data they are sharing, they may not understand what inferences could be made based on that data.

The risk of abuse of the underlying datasets remains. As the recent hack on Target's credit card systems demonstrates, even large, well-financed companies can suffer from massive data breaches that put consumers' data in the hands of a malicious third-party.[12] This danger is especially grave when companies collect and save all data possible, regardless of its current value, with the idea that a profitable use might later emerge. Unfortunately, the collection of data into more concentrated repositories creates a tempting target for malicious agents. Additionally, EFF has long been concerned that private-sector mass data accumulation strongly facilitates government data accumulation, given the many ways that companies can be induced or compelled to provide data to the government.

Finally, even if the above dangers are avoided, we emphasize that many "common sense" approaches to preserving privacy and anonymity in big data do not actually accomplish their goals. Malicious actors could use a variety of sophisticated statistical and information-theoretic

---

[10] Hoofnagle, Chris Jay and King, Jennifer, "*What Californians Understand about Privacy Online."* (September 3, 2008). Available at SSRN: http://ssrn.com/abstract=1262130 or http://dx.doi.org/10.2139/ssrn.1262130

[11] See http://thedatamap.org/

[12] Elgin, Ben; Lawrence, Dune; Matlack, Carol; Riley, Michael. *“Missed Alarms and 40 Million Stolen Credit Card Numbers: How Target Blew It.”* Bloomberg Businessweek, March 13, 2014. https://web.archive.org/web/20140313132757/http://www.businessweek.com/articles/2014-03-13/target-missed-alarms-in-epic-hack-of-credit-card-data. Last accessed March 29, 2014.

algorithms to extract identifiable data from what appears to be an anonymized dataset.[13] This is especially true if the malicious agent has access to individual datasets that might not pose a privacy risk on their own, but when combined together can be used to infer private information.

Suggested Approaches

Unfortunately, current US policy frameworks and privacy proposals for protecting consumer privacy when it comes to the use of big data by the private sector are woefully inadequate. In order to remedy this and counteract the dangers described above, we propose a range of suggestions.

The private sector could adopt and adapt the White House's Consumer Privacy Bill of Rights to the collection and usage of big data.[14] Part of adapting to big data lies in companies' being more mindful of the knowledge asymmetry problems discussed above. In particular, companies should adopt clear policies that enable consumers to understand their data collection, use, and dissemination practices, including what specific personal data companies have about consumers and the algorithms (including scoring protocols) used to make decisions about consumers.[15] Because the results of big data analytics are often hard to predict,[16] consumers should also be able to see what a company's analytics are inferring about them. In essence, "consumers have a right to exercise control over what personal data companies collect from them and how they use

---

[13] Anderson, Nate. *"'Anonymized' data really isn't—and here's why not."* ArsTechnica, Sep. 8, 2009. https://web.archive.org/web/20140123133104/http://arstechnica.com/tech-policy/2009/09/your-secrets-live-online-in-databases-of-ruin/. Last accessed Mar. 29, 2014.

[14] *"Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation."* The White House, February 2012. http://www.whitehouse.gov/sites/default/files/privacy-final.pdf. Last accessed March 31, 2014.

[15] One of the more common objections to this recommendation is that the algorithms underlying big data analysis are sometimes too complex for even the practitioners of big data to understand, so it would be pointless to show the algorithm to a lay-person (or even an expert) to offset any privacy or accuracy concerns. This objection is true only to a certain extent: while some big data analysis does involve complicated machine learning algorithms, much of it is fairly straightforward, such as the Google Flu project. (In essence, the advantage of many machine-learning algorithms lies not in their complexity, but in the fact that they automatically "tune" themselves using data that is presented by the designer in order to train them. Once a machine learning algorithm has been trained and is put into use on actual data it is usually fairly straightforward, at least for a computer scientist, to follow the data flow and understand *how* the algorithms works–just not *why* it ended up that way.) However, the use of an algorithm that is too complicated to understand raises the serious question of how such an algorithm can be proven to be accurate and unbiased. Without being able to explain how the algorithm works, how can a company guarantee that the algorithm doesn't have the effect (whether intended or not) of discriminating against certain classes of consumers? This problem serves to reinforce the recommendation that companies be transparent in their use of big data algorithms that can affect consumers' lives; only by being transparent will independent watchdogs be able to test big data algorithms for fairness.

16 Indeed, this ability of big data to tease out non-obvious relationships is one of the reasons for its increasing popularity.

it," as well as "a right to access and correct personal data... in a manner that is appropriate to the sensitivity of the data and the risk of adverse consequences to consumers if the data is inaccurate."[17]

1. Companies can use technical means to minimize the privacy risk to consumers. Since numerous studies have shown that simple de-identified datasets are too easy to re-identify, steps should be taken to the greatest extent possible to reduce the production and collection of identifiable information. For example, a brick-and-mortar retailer might track peoples' smartphones to analyze how consumers move from display to display or department to department, but assign each phone it sees a new, random ID (not based in any way on the phone's unique MAC address or other identifying information) at the beginning of each business day. This way, there is very little risk of privacy-invasion based on operations on the dataset. Of course there is still the risk that third-parties who got access to the dataset itself could use it to infer private things by combining it with other knowledge. An even better approach would be for companies to design products that do not have hard-coded unique ID numbers which are shared or transmitted by the device in the normal course of its operation. For example, MAC addresses could be designed so that they are rarely transmitted, and random identifiers are transmitted instead.[18]
2. Much of the data collected today is not shared on purpose by consumers, but is "found" data: data collected incidental to the use of products and services whose purpose (at least to the consumer) has nothing to do with big data.[19] We believe that if the government or companies do not take action to implement some of the solutions described above, more consumers will begin to use tools and products that "leak" less private data to third-parties. Already tools such as Tor[20] exist to prevent consumers from "leaking" data to their ISPs about the websites they visit; XPrivacy[21] exists to stop apps from gathering

---

[17] *"Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation."* The White House, February 2012. http://www.whitehouse.gov/sites/default/files/privacy-final.pdf. Last accessed March 31, 2014.

[18] This is how mobile phone systems work: a randomly generated Temporary Mobile Subscriber Identity (TMSI) is used instead of the unique International Mobile Subscriber Identity (IMSI).

[19] Harford, Tom. *"Big data: are we making a big mistake?"* Financial Times Magazine, March 28, 2014. http://www.ft.com/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html. Last accessed March 31, 2014.

20 See, https://www.torproject.org/.

21 See, http://www.xprivacy.eu/.

unnecessary private information from Android smartphones; and browser plug-ins are being developed to allow people to automatically block third-party trackers on the Internet. It is perfectly appropriate for knowledgeable consumers to take technical precautions against surveillance, but it would be best if consumers could have legitimate trust and confidence that their everyday actions, online or offline, were not being routinely and secretly monitored.

The Government's Big Data Problem

Government use of big data raises many of the problems described above, which are exacerbated by the government's greater resources and its greater ability to exercise power over people's lives. Especially alarming are the recent revelations about intelligence activities that show government surveillance leveraging private sector tracking for advertising purposes as well as exploiting private-sector tracking implementations. In one example, it was revealed that the NSA uses advertising cookies to track a user's location and exfiltrate data off a computer.[22]

Even outside of the classified arena, the government has championed big data with little attention to privacy; the administration's "Big Data Research and Development Initiative" touted big data projects in areas with obvious privacy implications, such as military autonomous systems; cybersecurity; the smart grid; and transactional data from web searches, sensors, and cell phone records. No projects are aimed at addressing the privacy implications of big data.[23] It is clearly one-sided to stimulate the production and aggregation of highly revealing data and of sophisticated tools for analyzing that data without also stimulating privacy design awareness and privacy protection tools.

That needs to change. The government must conduct a full assessment of its big data policies, create new rules and oversight for big data, and be transparent about how it uses big data and algorithms.

First, there must be review of the government's use of big data in surveillance. Last June, documents revealed previously unknown collections of huge datasets by the National Security Agency (NSA). One such instance was the collection of Americans' calling records using Section 215 of the Patriot Act. Although Section 215 is supposed to be used by the FBI, the NSA was

---

http://www.washingtonpost.com/blogs/the-switch/wp/2013/12/10/nsa-uses-google-cookies-to-pinpoint-targets-for-hacking/

23 See Fact Sheet: Big Data Across the Federal Government (March 29, 2012).

able to compile a huge database of domestic calling records and run advanced algorithms on the dataset. These algorithms included querying a specific phone number in the dataset as well as using the dataset to create a social graph of certain phone numbers, often called "social chaining." The dataset is also used for other, still classified, techniques. There is a very real threat that Section 215 is not only used for the mass or bulk collection of calling records, but also for bulk collection of financial records, car rental records, and other "business records." Indeed, a recently released FISA Court order strongly suggests that Section 215 of the Patriot Act has been used to obtain mass financial records or purchase records.[24]

We've also seen such mass collection using Section 702 of the Foreign Intelligence Surveillance Amendments Act. Section 702 is used for at least two types of collection; other uses remain classified. The first type, known as PRISM, compels companies like Internet providers to turn over data like voice communications, email, video, chat messages, stored data, file transfers, VoIP calls, and other "digital network information" for information that is to, from, or about a "selector."[25] All of this information is collated into NSA databases, and disseminated to other database located at the FBI, NCTC, and CIA.[26] The second type is for "upstream" collection, under which telecom and Internet providers are required to work with NSA to copy, scan, and filter Internet and phone traffic coming through their physical infrastructure.[27] Both types of Section 702 collection target a foreign entity but acquire communications of persons in the United States.

All three of these types of collection are unconstitutional. In this age of big data it is beyond obvious that large datasets of metadata are highly revealing, and call detail records are especially so. Collection under Section 702 involves communications surveillance that has been within the purview of the Fourth Amendment for decades. Since *Katz v. United States*, the

---

[24] See http://www.dni.gov/files/documents/0328/104.%20BR%2010-82%20supplemental%20opinion%20-%20Redacted%2020140328.pdf. See also Senator Wyden interview on *Meet the Press. http://www.nbcnews.com/meet-the-press/meet-press-transcript-march-30-2014-n67356*

25 Selectors are not exclusive to email address, phone calls, or other personally identifiable terms. Selectors are also called "targets" or "targeting selector."

[26] Memorandum Opinion of October 3, 2011 by the Foreign Intelligence Surveillance Court ("Bates Opinion."). https://www.eff.org/document/october-3-2011-fisc-opinion-holding-nsa-surveillance-unconstitutional. Last accessed February 21, 2014.

[27] *"NSA slides explain the PRISM data collection program."* Washington Post, June 6, 2013. http://www.washingtonpost.com/wp-srv/special/politics/prism-collection-documents/. Last accessed February 21, 2014.

Supreme Court has repeatedly emphasized that the Fourth Amendment "protects people, not places,"[28] and has said that electronic surveillance presents a significant threat of "broad and unsuspected governmental incursions into conversational privacy."[29] These programs exemplify a big data nightmare of secret mass collection of data, secret human and automated big data analysis, and secret use of the results.

The government's current use of big data in these contexts must stop. Traditional investigatory techniques based on Constitutional norms of particularized and individualized suspicion have long acted as bedrock principles of intelligence collection and law enforcement techniques. Unfortunately, the lure of big data technology, shielded from public visibility, has led to a regime of bulk collection of communications, with so little judicial involvement that these programs are hardly distinguishable from the general warrants and writs of assistance that were the *raison d'etre* of the Fourth Amendment.

We urge the Administration to immediately stop misusing Section 215 of the Patriot Act and to support statutory reform to end mass collection of business records. Surveillance agencies should publicly disclose their mass spying techniques and issue Privacy Impact Assessments that set standards and address whether the agency is meeting them. As we've seen, disclosure in a responsible manner allows the public to engage in a vital discussion, and we already know the NSA is conducting such assessments; however, they remain classified.[30]

Concern over government use of big data extends to non-intelligence agencies. For instance, the Department of Homeland Security (DHS) is developing a new system called FALCON integrating over 17 different databases of information ranging from records of individuals who encounter law enforcement to student visa holders.[31] FALCON is an example of aggregating formerly separate databases to create a database with greater potential for privacy intrusions.[32]

DHS has also solicited bids to build and maintain a national database of motor vehicle

---

[28] *Katz v. United States* 389 US 347, 351-53 (1967).
[29] *Katz v. United States* 389 US 347, 351-53 (1967).
[30] Pg 44. Business Records FISA NSA Review June 25, 2009. https://www.eff.org/document/nsa-business-records-fisa-redactedex-ocr.
[31] http://www.dhs.gov/publication/dhsicepia-038-%E2%80%93-falcon-data-analysis-research-trade-transparency-system-falcon-dartts.
[32] See page 1-3 of http://www.dhs.gov/sites/default/files/publications/privacy_pia_ice_falcondartts_january2014_0.pdf.

license plate data.[33] Though the initial solicitation was recalled, it's disturbing that this bid was even issued in the first place. This trove of location data would reveal where you've been and when, and could be aggregated to present a detailed picture of your life and whom you associate with. Unsurprisingly, DHS had planned to use the data to target individuals. It wanted to be able to create its own "hot lists" of suspect vehicles from the data. Whether officers would have been required to articulate any individualized suspicion before putting a vehicle on a "hot list" is unclear; it's equally unclear how a vehicle would ever get off such a list.[34]

DHS also proposed sharing its "hot lists" with other agencies and wanted to be able to communicate with other users, "establish Lists submissions, flag license plates, and conduct searches anonymously." Meaningful oversight of the program would be impossible if officers could use the system anonymously. And while EFF has focused on big data privacy issues, both law enforcement and intelligence activity raise significant racial, ethnic and religious discrimination problems, exemplified by ICE's Secure Communities program and the NYPD's stop and frisk policy and surveillance of Muslim communities.[35]

Given the government's appetite for big data surveillance programs, one would expect strong evidence of their value in finding criminals and terrorists. But a 2008 study by the National Research Council concluded that data-mining in order to spot potential terrorists was ineffective, and that there was no clear evidence that behavioral surveillance was useful for counterterrorism operations.[36] In addition, a recent study by the New America Foundation analyzed 255 different terrorism cases and concluded NSA's bulk surveillance programs were of minimal value in supporting investigations.[37] Instead, the study demonstrated the strength of traditional investigative methods for initiating and advancing the investigations. There has also been extensive writing on the huge harms caused by sample bias and sample error in big datasets.[38]

---

[33] https://www.eff.org/document/dhs-national-license-plate-reader-database-solicitation
[34] https://www.eff.org/deeplinks/2014/01/los-angeles-cops-should-release-automatic-license-plate-reader-records-eff-aclu
[35] http://fcir.org/2011/10/20/report-secure-communities-encourages-racial-profiling-lack-of-due-process/
[36] National Research Council. *Protecting Individual Privacy in the Struggle Against Terrorists: A Framework for Program Assessment.* Washington, DC: The National Academies Press, 2008.
[37] See http://newamerica.net/publications/policy/do_nsas_bulk_surveillance_programs_stop_terrorists
[38] http://www.ft.com/intl/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html#axzz2xSL2ejQx

Suggested Approaches

To its credit, DHS—unlike intelligence agencies—seeks to follow the FIPPs[39] and issues Privacy Impact Assessments. The FIPPs provide a framework for the collection and usage of personal information generally, and can be seen as guiding principles for government and nongovernmental agencies dealing with sensitive personal information in a wide range of circumstances. The principles include:

> Purpose Specification: DHS should specifically articulate the authority that permits the collection of PII and specifically articulate the purpose or purposes for which the PII is intended to be used.
> Data Minimization: DHS should only collect PII that is directly relevant and necessary to accomplish the specified purpose(s) and only retain PII for as long as is necessary to fulfill the specified purpose(s).
> Use Limitation: DHS should use PII solely for the purpose(s) specified in the notice. Sharing PII outside the Department should be for a purpose compatible with the purpose for which the PII was collected.[40]

At a minimum, every agency should announce and use similar principles to guide its data activities, including its use of big data. Privacy Impact Assessments and System of Record Notices (SORNs) are also useful as they provide an overview of the collection, use of collection, and retention periods of collection.

More importantly, the government must commit to increasing its transparency around its big datasets. Overclassification is a chronic government problem. Time after time we've seen the witches' brew of ambiguity and secrecy poison democracy and the rule of law. It is widely noted by government officials, academics, and others that the classification system is broken.[41] The government—and particularly the intelligence agencies enamored with big data—must inform the public about what information they are collecting about US persons, and even innocent

---

[39] See, http://www.dhs.gov/xlibrary/assets/privacy/privacy_policyguide_2008-01.pdf
[40] See http://www.dhs.gov/xlibrary/assets/privacy/privacy_policyguide_2008-01.pdf
41 See http://www.brennancenter.org/publication/reducing-overclassification-through-accountability

foreigners. Algorithmic transparency is key. Given the amount of data that the government collects, and the incentives of law enforcement and national security agencies to conceal much of what they do (either to collect or to use that data), attention to these problems in the specific context of government secrecy is crucial. No reform of big data for national security can be complete without this added transparency, which must include more affirmative disclosures as opposed to reactive declassifications. Director of National Intelligence General James Clapper recently lamented the fact that NSA should have disclosed the Section 215 Business Records FISA program collecting all Americans' calling records earlier. In particular, he said:

> …had we been transparent about this from the outset right after 9/11—which is the genesis of the 215 program—and said both to the American people and to their elected representatives, we need to cover this gap, we need to make sure this never happens to us again, so here is what we are going to set up, here is how it's going to work, and why we have to do it, and here are the safeguards…We wouldn't have had the problem we had.[42]

General Clapper should take this advice and apply it to any and all intelligence agency programs collecting big data about Americans.

The government must also stop the collection of big datasets that are clearly unconstitutional, like the collection of innocent Americans' calling records, phone calls, and emails. In addition, we suggest:

1. Stopping the big datasets collected by the illegal and unconstitutional use of Section 215 of the Patriot Act and Section 702 of the Foreign Intelligence Surveillance Amendments Act. These uses are conducted via an Executive interpretation of the law and can be stopped by the President. As such, we urge the Administration to simply stop misusing section 215 of the Patriot Act and to support statutory reform that ends mass collection of business records.
2. Calling for a Congressional investigation reviewing unclassified and classified intelligence programs collecting big data on US persons and innocent foreigners. Such a review could include a review of our surveillance programs; foreign policy implications of these programs; efficacy of the various collection techniques; a review of the classification regime; and a review of the current oversight system, which includes reviewing the current Congressional oversight system.

[42] Lake, Eli. *Spy Chief: We Should've Told You We Track Your Calls*, The Daily Best, February 17, 2014. https://web.archive.org/web/20140317070031/http://www.thedailybeast.com/articles/2014/02/17/spy-chief-we-should-ve-told-you-we-track-your-calls.html

3. The White House engage in a legislative strategy that not only encompasses surveillance reform, but also addresses the state secrets privilege or the standing to sue in surveillance litigation challenges.
4. Mandate Fair Information Practices Procedures across all government agencies. FIPPs can serve as a control on government's insatiable appetite for data. It will allow for proportional collection, identify the primary purpose of the data, and allow for a notice and disclosure to users. Other avenues include, mandating the release of Privacy Impact Assessments (PIAs) for intelligence agencies. Privacy Impact Assessment and System of Record Notices (SORNs) have provided the public with a notification of the collection and storage practices of users' information. Both can be improved. Sometimes PIAs and SORNs are not updated and fall out-of-date. Both should be reviewed annually to make sure they conform with the current use of the system.

Conclusion

It is a truism that big data is not going away but is only going to become more prevalent. In the future storage will only get cheaper, processing and analytics will only get faster, and both the private-sector and the government will be more incentivized to squeeze every last bit of information out of any data they can acquire. Additionally, the number and types of sources of big data will only increase as our daily lives become more digital. Self-driving cars equipped with cameras and other sensors and consumer products designed to be part of the Internet of things (eg, networked home appliances) will all soon have the capability to collect more data on individuals, which will no doubt be funneled back to company and government databases. Given these new challenges, not to mention the existing problems and dangers we have described, it is extremely important that the government take notice of how big data is being used by the private sector and work to ensure that consumers' privacy is preserved. Even more importantly, the government must take strong steps to end the misuse of big data by law enforcement and intelligence agencies, if for no other reason than to preserve Americans' Fourth Amendment rights. It is for these reasons that EFF strongly recommend that new policy frameworks and regulations be implemented to fundamentally change how big data is collected, managed, and used.

Privacy Coalition - Updated

April 4, 2014

Big Data Study
Office of Science and Technology Policy
Eisenhower Executive Building
1650 Pennsylvania Ave. NW
Washington, DC 20502

Dear Deputy Wong:

Our organizations favor the White House review of Big Data and the Future of Privacy. As the President has explained, both the government and the private sector collect vast amounts of personal information. "Big data" supports commercial growth, government programs, and opportunities for innovation. But big data also creates new problems including pervasive surveillance; the collection, use, and retention of vast amounts of personal data; profiling and discrimination; and the very real risk that over time more decision-making about individuals will be automated, opaque, and unaccountable.

That is the current reality and the likely future that the White House report must address. We therefore urge the White House to incorporate these requirements in its final report on Big Data and the Future of Privacy:

TRANSPARENCY: Entities that collect personal information should be transparent about what information they collect, how they collect it, who will have access to it, and how it is intended to be used. Furthermore, the algorithms employed in big data should be made available to the public.

OVERSIGHT: Independent mechanisms should be put in place to assure the integrity of the data and the algorithms that analyze the data. These mechanisms should help ensure the accuracy and the fairness of the decision-making.

ACCOUNTABILITY: Entities that improperly use data or algorithms for profiling or discrimination should be held accountable. Individuals should have clear recourse to remedies to address unfair decisions about them using their data. They should be able to easily access and correct inaccurate information collected about them.

ROBUST PRIVACY TECHNIQUES: Techniques that help obtain the advantages of big data while minimizing privacy risks should be encouraged. But these techniques must be robust, scalable, provable, and practical. And solutions that may be many years into the future provide no practical benefit today.

MEANINGFUL EVALUATION: Entities that use big data should evaluate its usefulness on an ongoing basis and refrain from collecting and retaining data that is not necessary for its intended purpose. We have learned that the massive metadata program created by the NSA has played virtually no role in any significant terrorism investigation. We suspect this is true also for many other "big data" programs.

CONTROL: Individuals should be able to exercise control over the data they create or is associated with them, and decide whether the data should be collected and how it should be used if collected.

We continue to favor the framework set out in the Consumer Privacy Bill of Rights and see that as an effective foundation on which to build other responses to the challenges of Big Data.

Signatories:

Advocacy for Principled Action in Government
American Association of Law Libraries
American Library Association
Association of Research Libraries
Bill of Rights Defense Committee
Center for Digital Democracy
Center for Effective Government
Center for Media Justice
Consumer Action
Consumer Federation of America
Consumer Task Force for Automotive Issues
Consumer Watchdog
Council for Responsible Genetics
Doctor Patient Medical Association
Electronic Privacy Information Center (EPIC)
Foolproof Initiative
Government Accountability Project
OpenTheGovernment.org
National Center for Transgender Equality
Patient Privacy Rights
PEN American Center
Privacy Journal
Privacy Rights Clearinghouse
Privacy Times
Public Citizen, Inc.

COMMENTS OF THE ELECTRONIC PRIVACY INFORMATION CENTER

to

THE OFFICE OF SCIENCE AND TECHNOLOGY POLICY

Request for Information: Big Data and the Future of Privacy

April 4, 2014

By notice published on March 4, 2014, the Office of Science and Technology Policy ("OSTP") requests public comment on "big data."[1] Pursuant to OSTP's notice, the Electronic Privacy Information Center ("EPIC") submits these comments to: (1) warn the OSTP about the enormous risk to Americans in the current "Big Data" environment; (2) make clear that the challenges of Big Data are not new; (3) call for the swift enactment of the Consumer Privacy Bill of Rights ("CPBR") and the end of opaque algorithmic profiling; (4) highlight the need for stronger privacy safeguards for "Big Data"; and (5) draw attention to international frameworks that provide strong models for safeguarding privacy.

EPIC is a public interest research center in Washington, DC. EPIC was established in 1994 to focus public attention on emerging civil liberties issues and to protect privacy, the First Amendment, and constitutional values. EPIC has a particular interest in safeguarding personal privacy and preventing harmful data practices. For example, EPIC routinely submits comments to federal agencies, urging them to uphold the Privacy Act and protect individual privacy in mass government databases.[2] EPIC has adamantly opposed government use of "risk-based" algorithmic profiling.[3] EPIC highlighted the problems inherent in profiling programs like the Department of Homeland Security's ("DHS") Secure Flight in previous testimony and comments. In testimony before the National Commission on Terrorist Attacks Upon the United States (more commonly known as "the 9/11 Commission"), EPIC President Marc Rotenberg explained, "there are specific problems with information technologies for monitoring, tracking, and profiling. The techniques are imprecise, they are subject to abuse, and they are invariably applied to purposes other than those originally intended."[4] EPIC is also a leading consumer advocate before the Federal Trade Commission ("FTC"). EPIC has a particular interest in protecting consumer privacy, and has played

---

[1] Government "Big Data," 79 Fed. Reg. 12,251 (Mar. 4, 2014).

[2] *See, e.g.*, EPIC et al., *Comments on the Terrorist Screening Database System of Records, Notice of Privacy Act System of Records and Notice of Proposed rulemaking, Docket Nos. DHS 2011-0060 and DHS 2011-0061* (Aug. 5, 2011), *available at* http://epic.org/privacy/airtravel/Comments_on_DHS-2011-0060_and_0061FINAL.pdf; EPIC, *Comments on Secure Flight, Docket Nos. TSA-2007-28972, 2007-28572* (Sept. 24, 2007), *available at* http://epic.org/privacy/airtravel/sf_092407.pdf; EPIC, *Secure Flights Should Remain Grounded Until Security and Privacy Problems are Resolved, Spotlight on Surveillance Series* (August 2007), *available at* http://epic.org/privacy/surveillance/spotlight/0807/default.html; *Passenger Profiling*, EPIC, http://epic.org/privacy/airtravel/profiling.html (last visited Apr. 3, 2014); *Secure Flight*, EPIC, http://epic.org/privacy/airtravel/secureflight.html (last visited Apr. 3, 2014); *Air Travel Privacy*, EPIC, http://epic.org/privacy/airtravel/ (last visited Apr. 3, 2014).

[3] *See, e.g.*, EPIC et al., *Comments Urging the Department of Homeland Security To (A) Suspend the "Automated Targeting System" As Applied To Individuals, Or In the Alternative, (B) Fully Apply All Privacy Act Safeguards To Any Person Subject To the Automated Targeting System* (Dec. 4, 2006), *available at* http://epic.org/privacy/pdf/ats_comments.pdf; EPIC, *Comments on Automated Targeting System Notice of Privacy Act System of Records and Notice of Proposed Rulemaking, Docket Nos. DHS-2007-0042 and DHS-2007-0043* (Sept. 5, 2007), *available at* http://epic.org/privacy/travel/ats/epic_090507.pdf. *See also*, *Automated Targeting System*, EPIC, https://epic.org/privacy/travel/ats/.

[4] Marc Rotenberg, President, EPIC, *Prepared Testimony and Statement for the Record of a Hearing on Security & Liberty: Protecting Privacy, Preventing Terrorism Before the National Commission on Terrorist Attacks Upon the United States* (Dec. 8, 2003), *available at* http://www.epic.org/privacy/terrorism/911commtest.pdf.

a leading role in developing the authority of the FTC to address emerging privacy issues and to safeguard the privacy rights of consumers.[5]

On January 17, 2014, President Obama announced a plan to take a comprehensive look at the privacy implications of "Big Data."[6] Almost immediately after the White House announcement, EPIC, joined by a coalition of consumer privacy, public interest, scientific, and educational organizations, petitioned OSTP to meaningfully engage the public by accepting public comments on Big Data and the Future of Privacy.[7] The Privacy Coalition urged OSTP to involve the public because it is the public's privacy and future that is at stake when the government and private companies amass big data obtained from the public. The Privacy Coalition encouraged OSTP to consider an array of big data privacy issues, including:

(1)What potential harms arise from big data collection and how are these risks currently addressed?
(2) What are the legal frameworks currently governing big data, and are they adequate?
(3) How could companies and government agencies be more transparent in the use of big data, for example, by publishing algorithms?
(4) What technical measures could promote the benefits of big data while minimizing the privacy risks?
(5) What experience have other countries had trying to address the challenges of big data?
(6) What future trends concerning big data could inform the current debate?

Less than a month after the Coalition filed its petition, the White House announced this public comment opportunity. EPIC appreciates this effort as well as related efforts to encourage public comments on this important policy process.[8]

As discussed below in detail, private organizations and government entities are amassing data with little understanding of the consequences and too few safeguards. In many instances, the organizations gathering the Big Data obtain the benefits, but the individuals bear the consequences.[9] This leads to asymmetries of power and new more subtle means of control. We urge OSTP to incorporate the following observations and recommendations into its final report.

1. The current "Big Data" environment poses enormous risk to Americans

The ongoing collection of personal information in the United States without sufficient privacy safeguards has led to staggering increases in identity theft, security breaches, and financial fraud. Additionally, the use of personal information to make automated decisions and segregate individuals based on secret, imprecise and oftentimes impermissible factors presents clear risks to fairness and due process. Far too many organizations collect detailed personal information and use it with too little regard for the consequences. The current Big Data environment is plagued by data breaches and discriminatory uses of predictive analytics.

---

[5] *See, e.g.*, Letter from EPIC Executive Director Marc Rotenberg to FTC Commissioner Christine Varney, EPIC (Dec. 14, 1995) (urging the FTC to investigate the misuse of personal information by the direct marketing industry), http://epic.org/privacy/internet/ftc/ftc_letter.html; DoubleClick, Inc., *FTC* File No. 071-0170 (2000) (Complaint and Request for Injunction, Request for Investigation and for Other Relief), http://epic.org/privacy/internet/ftc/DCLK_complaint.pdf; Microsoft Corporation, FTC File No. 012 3240 (2002) (Complaint and Request for Injunction, Request for Investigation and for Other Relief), http://epic.org/privacy/consumer/MS_complaint.pdf; Choicepoint, Inc., FTC File No. 052-3069 (2004) (Request for Investigation and for Other Relief) , http://epic.org/privacy/choicepoint/fcraltr12.16.04.html.

[6] John Podesta, *Big Data and the Future of Privacy*, THE WHITE HOUSE BLOG (Jan. 23, 2014, 3:30 PM), http://www.whitehouse.gov/blog/2014/01/23/big-data-and-future-privacy.

[7] EPIC et al., Petition for OSTP to Conduct Public Comment Process on Big Data and the Future of Privacy, Feb. 10, 2014, http://epic.org/privacy/Ltr-to-OSTP-re-Big-Data.pdf.

[8] *Join the Conversation: Big Data, Privacy, and What it Means to You*, THE WHITE HOUSE, http://www.whitehouse.gov/issues/technology/big-data-review (last visited Apr. 3, 2014).

[9] *Big Data and the Future of Privacy*, EPIC, http://epic.org/privacy/big-data/default.html (last visited Apr. 4, 2014).

The use of predictive analytics by the public and private sector undermines our freedom of association. Our online social connections, participation in online debates, and our interests expressed through our online activities can now be used by the government and companies to make determinations about our ability to fly, to obtain a job, a clearance, or a credit card. The use of our associations in predictive analytics to make decisions that have a negative impact on individuals directly inhibits freedom of association. It chills online interaction and participation when those very acts and the associations they reveal could be used to deny an individual a job or flag an individual for additional screening at an airport because of the determination of an opaque algorithm, that may consider a person's race, nationality, or political views.

The ability to predict sensitive data and reveal associations raises the potential for abuse by both the government and the private sector. The information gleaned from predictive analytics could be used in a variety of ways to skirt current legal protections regarding, for example, fairness in housing and employment and First Amendment freedoms of religion and association.[10]

A. Commercial Institutions Collecting Data Have Insufficient Data Security to Protect Americans' Privacy

Over the past year, many disastrous data breaches have occurred. During the busy holiday shopping season, millions of American customers who shopped at Target and Neiman Marcus suffered data breaches. Target suffered a data breach that affected nearly 70 million after its point-of-sale terminals were hacked and compromised because of its own insufficient security standards.[11] This included the account data for roughly 40 million account holders, including their credit and debit card numbers, expiration dates, the three-digit CVV security code, and even PIN data.[12] The customers of Neiman Marcus suffered a very similar data breach in which 1.1 million debit and credit card numbers were compromised.[13]

Last September, a data breach at Adobe exposed the user account information of 38 million users.[14] The breach resulted in the theft of close to 3 million customer credit card numbers.[15] The user account information was similarly exposed in a data breach of LivingSocial that compromised the data of nearly 50 million users.[16] Government agencies routinely lose control of the databases containing detailed personal information they have acquired in the "big data" environment.[17]

---

[10] Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms* 99-101 (Public Law & Legal Theory Research, Working Paper No. 13-64), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2325784.

[11] Target: data breach FAQ, https://corporate.target.com/about/shopping-experience/payment-card-issue-FAQ.

[12] Sarah Perez, *Target's Data Breach Gets Worse: 70 Million Customs Had Info Stolen, Including Names, Emails, and Phones*, TechCrunch, Jan. 10, 2014, http://techcrunch.com/2014/01/10/targets-data-breach-gets-worse-70-million-customers-had-info-stolen-including-names-emails-and-phones/.

[13] Elizabeth A. Harris, Nicole Perlroth & Nathaniel Popper, *Neiman Marcus Data Breach Worse Than First Said*, NYTimes, Jan. 23, 2014, http://www.nytimes.com/2014/01/24/business/neiman-marcus-breach-affected-1-1-million-cards.html.

[14] Brian Krebs, *Adobe Breach Impacted at Least 38 Million Users*, Oct. 29, 2013, Krebs on Security, http://krebsonsecurity.com/2013/10/adobe-breach-impacted-at-least-38-million-users/.

[15] *Id*.

[16] Nicole Perlroth, *LivingSocial Hack Exposes Data for 50 Million Customers*, N.Y. Times, Apr. 26, 2013.

[17] *See, e.g.*, U.S. GOVT' ACCOUNTABILITY OFFICE, GAO-14-487T, INFORMATION SECURITY: FEDERAL AGENCIES NEED TO ENHANCE RESPONSES TO DATA BREACHES (2014), *available at* http://www.gao.gov/assets/670/662227.pdf; William Jackson, *VA Settlement Demonstrates Just How Costly Lax Security Can Be*, GCN, Feb. 2, 2009, http://gcn.com/Articles/2009/02/02/VA-data-breach-suit-settlement.aspx; Majority Staff of H. COMM. ON OVERSIGHT AND GOV'T REFORM, *Information Security Breach at TSA: The Traveler Redress Website* (January 2008), *available at* http://web.archive.org/web/20080131043651/http://oversight.house.gov/documents/20080111092648.pdf; Spencer S. Hsu, *TSA Hard Drive With Employee Data Is Reported Stolen*, WASHINGTON POST (May 5, 2007), http://www.washingtonpost.com/wp-dyn/content/article/2007/05/04/AR2007050402152.html.

In addition to the failure of organizations to adequately safeguard the information they collect, many private companies and government agencies now use opaque and often imprecise techniques that make determinations about individuals that carry real consequences. "Predictive analytics" use algorithms on vast amounts of data to unearth correlations that would otherwise remain hidden.[18] Often, the algorithms leverage seemingly innocuous information to make predictions about sexuality, whether a woman is pregnant, political leanings, and more. One of the more problematic uses of predictive analytics is preemptive predictions that make a specific determination about an individual.

Preemptive predictions limit a person's options by assessing "the likely consequences of allowing or disallowing a person to act in a certain way."[19] Preemptive predictions are made from the perspective "of the state, a corporation, or anyone who wishes to prevent or forestall certain types of action."[20] Examples of preemptive predictions include inclusion on a no-fly list and determinations of credit worthiness. Preemptive predictions are particularly problematic because they are often completely automated decisions made behind a veil of secrecy that lack clear or effective recourse for those individuals who feel they have been wronged by the decision.

The private sector uses big data analytics to make important decisions that affect individuals. A digital lending company has established a loan and credit scoring service that uses big data analytics to assess a person's credit worthiness.[21] The company collects data from social networks, among other sources, to make the automated determination in seconds using a self-learning algorithm.[22]

Even when predictive analytics are not used to make a determination about an individual, they still can be problematic by predicting and, in some instances, revealing sensitive information. The retail chain Target used predictive analytics to predict which female customers were pregnant.[23] This information was given to marketers who revealed the pregnancy of a young woman prior to her telling her parents.[24]

Often, the companies and institutions that are the victims of large-scale data breaches make efforts after-the-fact to improve security and privacy. But this leaves numerous other entities still exposing the personal information of its customers. This problem will only get worse because as John Podesta stated, "There is no question that there is more data than ever before, and no sign that the trajectory is slowing its upward pace."[25]

B. Students are Particularly Vulnerable to Big Data Privacy Risks

Recent large-scale security breaches at educational institutions have compromised student (and faculty) privacy. Last month, a University of Maryland ("UMD") database containing 309,079 student, faculty, staff, and personnel records was breached; the "breached records included name, Social Security number, date of birth, and University identification number" and included records covering a span of 20 years.[26] The university acknowledged that it could have implemented privacy enhancing techniques by purging some of those records "long before the breach."[27] Soon after the UMD breach, Indiana University reported that it had stored names, addresses, and Social

---

[18] VIKTOR MAYER-SCHÖNBERGER & KENNETH CUKIER, BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK 11-12 (Houghton Mifflin Harcourt 2013).
[19] Ian Kerr & Jessica Earle, *Prediction, Preemption, Presumption: How Big Data Threatens Big Picture Privacy* 66 Stan. L. Rev. Online 65, 67 (2013).
[20] *Id.*
[21] Kreditech: Digital Lending, https://www.kreditech.com/loan-and-credit-scoring/.
[22] *Id.*
[23] Charles Duhigg, *How Companies Learn Your Secrets*, N.Y. Times, Feb. 16, 2012, http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html.
[24] *Id.*
[25] Counselor John Podesta, Remarks at the White House/MIT "Big Data" Privacy Workshop (Mar. 3, 2014), *available at* http://www.whitehouse.gov/sites/default/files/docs/030414_remarks_john_podesta_big_data.pdf.
[26] Letter from President Loh, Letter from Brian D. Voss concerning UMD Data Breach, http://www.umd.edu/datasecurity/.
[27] Mark Albert, *UMD Testifies to Congress on Massive Data Breach*, WUSA 9, Mar. 27, 2014, http://www.wusa9.com/story/news/local/2014/03/26/university-of-maryland-congress-data-breach/6942023/.

Security numbers for "approximately 146,000 students and recent graduates" in an "insecure location" for almost a year, thus potentially exposing students to identity theft and other forms of fraud.[28] Johns Hopkins University also recently experienced a breach that compromised the names, contact information, and "student-entered comments" of approximately 850 students that were enrolled over a seven-year span.[29] Hackers posted information stolen from the breach, including employee information, on the internet. In response to the breach, Johns Hopkins is exploring privacy enhancing techniques, such as deleting outdated information.[30] These examples illustrate that Big Data places students at risk because schools are not using adequate security standards to protect student records.

Additionally, the mass collection of student information has led to the creation of student dossiers over which students have little to no control. For example, statewide longitudinal databases collect troves of student information comprised of "preschool, K-12, and postsecondary education as well as workforce data."[31] A 2009 Fordham Law School report analyzing statewide longitudinal databases highlights that (1) "most states collected information in excess of what is needed" for government reporting requirements"; (2) student databases "generally had weak privacy protections"; (3) "many states do not have clear access and use rules regarding the longitudinal database"; (4) most states "fail to have data retention policies"; and (5) "several states . . . outsource the data warehouse without any protections for privacy in the vendor contract."[32] Because statewide longitudinal databases collect so much student information and because that information is not adequately protected, Big Data in student statewide longitudinal databases significantly raises the risks that students will be stigmatized throughout their academic career and in the workforce.

Last year, EPIC testified before the Colorado State Board of Education and discussed the growing privacy risks that students face as private companies routinely collect sensitive student records. EPIC discussed how private companies might access extensive disciplinary records, and even facilitate "principal watch lists."[33]

C. Government Collection of Big Data is Particularly Problematic

The government has also abused Big Data. Documents obtained by EPIC through a Freedom of Information Act request show that the Census Bureau provided the Department of Homeland Security statistical data on people who identified themselves on the 2000 census as being of Arab ancestry.[34] The DHS agent who requested the census data explained that it was needed to determine which languages signage should be posted in at major international airports.[35] However, there was no indication that DHS requested similar information about any other ethnic groups.[36] The ultimate abuse of Census information came during World War II, when the Census Bureau provided statistical information to help the War Department round up more than 120,000 innocent Japanese Americans and confine them to internment camps.

Today, Americans are in more government databases than ever. Government agencies routinely amass PII, but absolve themselves of any legal duties or responsibilities to safeguard individual privacy. For example, the Federal Bureau of Investigation's Data Warehouse System hoards individual information, including:

---

[28] Indiana University Reports Potential Data Exposure, Feb. 25, 2014, http://news.iu.edu/releases/iu/2014/02/data-exposure-disclosure.shtml.
[29] Johns Hopkins Statement: Breach of a University Server, Mar. 7, 2014, http://releases.jhu.edu/2014/03/07/server-breach/.
[30] *Id.*
[31] *Statewide Longitudinal Data Systems*, EDUCATION DEPARTMENT, https://www2.ed.gov/programs/slds/factsheet.html (last visited Apr. 3, 2014).
[32] CHILDREN'S EDUCATIONAL RECORDS AND PRIVACY: A STUDY OF ELEMENTARY AND SECONDARY SCHOOL STATE REPORTING SYSTEMS, EXECUTIVE SUMMARY (Fordham Law Ctr. on Law and Info. Policy, 2009).
[33] Testimony and Statement for the Record, Khaliah Barnes, EPIC Administrative Law Counsel, Study Session Regarding inBloom, Inc., May 16, 2013, *available at* https://epic.org/privacy/student/EPIC-Stmnt-CO-Study-5-13.pdf.
[34] *Freedom of Information Documents on the Census: Department of Homeland Security Obtained Data on Arab Americans From Census Bureau*, EPIC, http://epic.org/privacy/census/foia/ (last visited Apr. 3, 2014).
[35] EPIC FOIA documents: Email exchange between DHS and Census Bureau, http://epic.org/privacy/census/foia/census_emails.pdf.
[36] *Id.*

> biographical information (such as name, alias, race, sex, date of birth, place of birth, social security number, passport number, driver's license, or other unique identifier, addresses, telephone numbers, physical descriptions, and photographs); biometric information (such as fingerprints); financial information (such as bank account number); location; associates and affiliations; employment and business information; visa and immigration information; travel; and criminal and investigative history, and other data that may assist the FBI in fulfilling its national security and law enforcement responsibilities.[37]

Incredibly, the agency has exempted itself from Privacy Act requirements that the FBI maintain only "accurate, relevant, timely and complete" personal records.[38] The FBI has also exempted itself from Privacy Act requirements permitting individuals to access and amend inaccurate records.[39] Other agencies, like the Department of Homeland Security and the National Security Agency, have exempted databases containing detailed, sensitive personal information from well-established Privacy Act safeguards.[40] EPIC has routinely objected to agencies gathering personally identifiable information while eschewing privacy protections, noting:

> It is inconceivable that the drafters of the Privacy Act would have permitted a federal agency to propose a profiling system on U.S. citizens and be granted broad exemptions from Privacy Act obligations. Consistent and broad application of Privacy Act obligations are the best means of ensuring accuracy and reliability of the data used in a system that profoundly affects millions of individuals as they travel throughout the United States on a daily basis.[41]

Like the private sector, the government also uses predictive analytics to the detriment of millions of individuals. For example, the Department of Homeland Security's TSA PreCheck program collects vast amounts of PII including biometric information to perform a "security threat assessment" of "law enforcement, immigration, and intelligence databases, including a fingerprint-based criminal history check conducted through the Federal Bureau of Investigation."[42] The TSA uses automated data processing to determine which individuals will be scrutinized upon traveling throughout the United States.[43] The decisions are completely opaque and lack an effective recourse option. Remarkably, the TSA itself has lost sensitive personal information that it has collected from its employees.[44] The TSA lost a portable drive containing the bank account numbers, Social Security numbers, names and birth dates of more than 100,000 people who worked at the TSA over a three-year period.

It is vitally important to update current privacy laws to minimize collection, secure the information that is collected, and prevent abuses of collected data through the use of predictive analytics.

---

[37] Privacy Act of 1974; System of Records, 77 Fed. Reg. 40,630, 40,631 (July 10, 2012), *available at* http://www.gpo.gov/fdsys/pkg/FR-2012-07-10/pdf/2012-16823.pdf.
[38] 28 C.F.R. §16.96 (v).
[39] *Id*.
[40] *See, e.g.*, EPIC et al., *Comments on the Department of Defense Privacy Program* (Oct. 21, 2013), *available at* https://epic.org/privacy/nsa/Coal-DoD-Priv-Program-Cmts.pdf; *see also supra* note 3, *Comments Urging the Department of Homeland Security To (A) Suspend the "Automated Targeting System"*.
[41] EPIC, *Comments on TSA PreCheck Application Program System of Records Notice and Notice of Proposed Rulemaking and TSA Secure Flight System of Records Notice*, 5 (Oct. 10, 2013), *available at* http://epic.org/apa/comments/TSA-PreCheck-Comments.pdf.
[42] Privacy Act of 1974: Implementation of Exemptions; Department of Homeland Security Transportation Security Administration, DHS/TSA-021, TSA PreCheck Application Program System of Records, 78 Fed. Reg. at 55,657 (proposed Sept. 11, 2013), *available at* http://www.gpo.gov/fdsys/pkg/FR-2013-09-11/pdf/2013-22069.pdf.
[43] Privacy Act of 1974; Department of Homeland Security Transportation Security Administration--DHS/TSA—019 Secure Flight Records System of Records, 78 Fed. Reg. 55,270, 55,271 (proposed Sept. 10, 2013), *available at* http://www.gpo.gov/fdsys/pkg/FR-2013-09-10/pdf/2013-21980.pdf.
[44] Thomas Frank, *TSA Seeks Hard Drive, Personal Data on 100,000*, USA TODAY, May 5, 2007, available at http://usatoday30.usatoday.com/news/washington/2007-05-04-harddrive-tsa_N.htm?csp=1.

2. The Challenges that Big Data Present Are Not New

Many of the problems that Americans are confronting today were anticipated when Congress first addressed the challenges of "Big Data" and automating personal information with the Privacy Act of 1974. The Privacy Act incorporates the Code of Fair Information Practices that the Health, Education, Welfare Advisory Committee on Automated Data Systems issued in 1973.[45] The Code of Fair Information Practices sets out five obligations for all organizations that collect personal data:

1. There must be no personal data record-keeping systems whose very existence is secret.
2. There must be a way for a person to find out what information about the person is in a record and how it is used.
3. There must be a way for a person to prevent information about the person that was obtained for one purpose from being used or made available for other purposes without the person's consent.
4. There must be a way for a person to correct or amend a record of identifiable information about the person.
5. Any organization creating, maintaining, using, or disseminating records of identifiable personal data must assure the reliability of the data for their intended use and must take precautions to prevent misuses of the data.[46]

In passing the Privacy Act of 1974, Congress found that: (1) individual privacy is "directly affected by the collection, maintenance, use, and dissemination of personal information by Federal agencies"; (2) big data in the government sector "greatly magnified the harm to individual privacy"; (3) misuse of government big data can threaten "the opportunities for an individual to secure employment, insurance, and credit, and his right to due process"; (4) privacy is a constitutionally-protected "personal and fundamental right"; and (5) "in order to protect the privacy of individuals identified in information systems maintained by Federal agencies, it is necessary and proper for the Congress to regulate the collection, maintenance, use, and dissemination of information by such agencies."[47]

The findings in the US Privacy Act of 1974 make clear the risks of "big data," long before the term was used.[48] However, the United States has been slow to update its privacy laws. Other countries and regions are moving more effectively to respond to the modern challenge of big data. For example, the European Union Data Protection Directive of 1995 actually anticipated the problem of secretive decisionmaking that would undermine fairness.[49] The right of access, familiar to many in the US, is not limited to simply knowledge about the personal data that is collected but also to how the data is used. According to Article 12 of the Directive:

> *Right of access*
>
> Member States shall guarantee every data subject the right to obtain from the controller:
>
> (a) without constraint at reasonable intervals and without excessive delay or expense: confirmation as to whether or not data relating to him are being processed and information at least as to the purposes of the processing, the categories of data concerned, and the recipients or categories of recipients to whom the data are disclosed; communication to him in an intelligible form of the data undergoing processing and of any available information as to their source; *knowledge of the logic*

[45] *The Code of Fair Information Practices*, EPIC, http://epic.org/privacy/consumer/code_fair_info.html.
[46] U.S. Dep't. of Health, Education and Welfare, Secretary's Advisory Committee on Automated Personal Data Systems, Records, computers, and the Rights of Citizens viii (1973).
[47] Public Law 93-579, 93rd Congress, S.3418, Privacy Act, Section 2 (a) (Dec. 31, 1974).
[48] In the 1960s and 1970s, commentators and policy makers were more likely to say "databanks" or "databases." *See, e.g.*, ARTHUR R. MILLER, THE ASSAULT ON PRIVACY: COMPUTERS, DATA BANKS, AND DOSSIERS (University of Michigan Press 1971); ALAN F. WESTIN, PRIVACY AND FREEDOM (Bodley Head 1970).
[49] S*ee generally EU Data Protection Directive,* EPIC, http://epic.org/privacy/intl/eu_data_protection_directive.html.

> *involved in any automatic processing of data* concerning him at least in the case of the automated decisions referred to in Article 15(1);[50]

As a document prepared for Europeans explains:

> You must also have access to the logic on which automated decisions are based. Decisions, which significantly affect the data subject, such as the decision to grant a loan or issue insurance, might be taken on the sole basis of automated data processing. Therefore, the data controller must adopt suitable safeguards, such as giving the data subject the opportunity to discuss the rationale behind the data collected or to contest decisions based on inaccurate data.[51]

Although the Privacy Act of 1974 anticipated many of the challenges that Big Data present, the current legal frameworks fail to safeguard individual privacy by adequately implementing Fair Information Practices ("FIPs") and adhering to privacy enhancing techniques. Because Big Data has threatened individual privacy for many years, and the risks to Americans increase daily, it is imperative that this Administration confronts Big Data problems expeditiously. Among the changes that are needed, the law should be updated to guarantee algorithmic transparency.

3. Congress Should Swiftly Enact the Consumer Privacy Bill of Rights and the Government Should Immediately Suspend its "Risk Based" Profiling Programs

In 2012, President Obama announced the Consumer Privacy Bill of Rights ("CPBR").[52] It is a critical policy framework that provides a blueprint for protecting privacy in the modern age. Based on FIPs, the CPBR is a framework that grants consumer rights and places obligations on private companies collecting consumer information:

- Individual Control: Consumers have a right to exercise control over what personal data companies collect from them and how they use it.
- Transparency: Consumers have a right to easily understandable and accessible information about privacy and security practices.
- Respect for Context: Consumers have a right to expect that companies will collect, use, and disclose personal data in ways that are consistent with the context in which consumers provide the data.
- Security: Consumers have a right to secure and responsible handling of personal data.
- Access and Accuracy: Consumers have a right to access and correct personal data in usable formats, in a manner that is appropriate to the sensitivity of the data and the risk of adverse consequences to consumers if the data is inaccurate.

---

[50] EU Directive 95/46/EC—The Data Protection Directive, art 15 (1), 1995 (emphasis added), http://www.dataprotection.ie/docs/EU-Directive-95-46-EC--Chapter-2/93.htm. Article 15(1) is expansive and includes "data intended to evaluate certain personal aspects relating to him, such as his performance at work, creditworthiness, reliability, conduct, etc."

[51] EUROPA, Data Protection in the European Union, 9, *available at* http://ec.europa.eu/justice/policies/privacy/docs/guide/guide-ukingdom_en.pdf.

[52] White House, Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Economy, Feb. 23, 2012, http://www.whitehouse.gov/sites/default/files/privacy-final.pdf [hereinafter White House, CPBR]; *see also White House Sets Out Consumer Privacy Bill of Rights*, EPIC, http://epic.org/2012/02/white-house-sets-out-consumer-.html (last visited Apr. 4, 2014).

- Focused Collection: Consumers have a right to reasonable limits on the personal data that companies collect and retain.

- Accountability: Consumers have a right to have personal data handled by companies with appropriate measures in place to assure they adhere to the Consumer Privacy Bill of Rights.[53]

The Consumer Data Privacy Report identifies several high-profile privacy challenges, including online advertising, data brokers, and children's privacy. The report encourages online advertising companies to "refrain from collecting, using, or disclosing personal data that may be used to make decisions regarding employment, credit, and insurance eligibility" and cited a "Do Not Track" mechanism as an example of a beneficial privacy-enhancing technology.[54] The report calls on data brokers to "seek innovative ways to provide consumers with effective Individual Control."[55] Finally, the report notes, "the practices in the Consumer Privacy Bill of Rights may require greater protections for personal data obtained from children and teenagers than for adults."[56]

More than two years have passed since President Obama announced the CPBR. But with no action on the recommendations or the framework, the problems with Big Data have increased. Last month, forty consumer privacy organizations urged the White House to work with Congress to propose legislation enacting the Consumer Privacy Bill of Rights into law. The groups stated:

> Never has the need to update the privacy laws of the United States been more urgent. . . . The Consumer Privacy Bill of Rights is a sensible framework that would help establish fairness and accountability for the collection and use of personal information. . . . the key to progress is the enactment by Congress of this important privacy framework. Only enforceable privacy protections create meaningful safeguards.[57]

The time to act is now. The White House must work with Congress to enact the CPBR and protect the privacy of Americans.

Additionally, to combat the problems with preemptive predictions based on government Big Data, the government must immediately cease "risk based" automated profiling. For example, the Department of Homeland Security ("DHS") uses its Automated Targeting System ("ATS") to assign risks to individuals traveling to, from, and throughout the United States.[58] DHS uses PII to determine whether an individual, based on personal immutable characteristics—not conduct—should undergo investigation, monitoring, and denial of her constitutional right to travel.[59] This is almost certainly constitutionally impermissible.[60] Moreover, because ATS risk assessment compares PII of individuals that have no criminal history against "patterns of suspicious activity," this increases the likelihood that CBP and ATS profile innocent individuals of certain racial, ethnic, or religious groups.[61]

---

[53] *See supra* note 52, White House, CPBR. at 1.
[54] *Id.* at 12.
[55] *Id.* at 13.
[56] *Id.* at 15.
[57] Privacy Coalition Letter on Consumer Privacy Bill of Rights, Feb. 24, 2014, *available at* http://epic.org/privacy/Obama-CPBR.pdf.
[58] Notice of Privacy Act System of Records, 77 Fed. Reg. 30297 (proposed May 22, 2012).
[59] *Sáenz v. Roe*, 526 U.S. 489 (1999).
[60] *See supra* note 3, *Comments Urging the Department of Homeland Security To (A) Suspend the "Automated Targeting System." See also* EPIC, *Comments to CBP regarding Automated Targeting System*, June 21, 2012, *available at* http://epic.org/privacy/travel/ats/EPIC-ATS-Comments-2012.pdf.
[61] Dep't of Homeland Sec., U.S. Customs and Border Protection, Privacy Impact Assessment for the Automated Targeting System, DHS/CBP/PIA-006(b), 19 (June 1, 2012), *available at* http://www.dhs.gov/xlibrary/assets/privacy/privacy_pia_cbp_ats006b.pdf.

EPIC and other privacy and civil liberties organizations have repeatedly called for suspension of automated "risk based" profiling.[62] In conducting its review on Big Data in the government sector, the White House should advocate for the immediate suspension of automated "risk based" profiling.

Big companies, including internet advertisers, should also be far more transparent about their profiling and pricing practices. Every indication points to the increasing use of secretive profiles, filled with sensitive data, to make determinations about American consumers.[63] Those practices should end. Companies should not be allowed to make decisions about individuals without setting out in detail the basis for the decision, including the factors that it considered. And government agencies must be on the lookout for the use of factors, such as race, gender, and nationality that are Constitutionally impermissible.

4. Stronger Big Data Privacy Safeguards Are Needed to Protect Individuals

A. *Current Practices*

It is imperative that stronger safeguards are implemented to protect the privacy and personal information of Americans. Transparency is often foregone in order to avoid accountability for the accuracy of the data or for how the data is used. Various entities access Big Data with little accountability. For example, private information aggregators increasingly sell consumer profiles that are not clearly protected under current legal frameworks.

When legal frameworks are inapplicable to new uses of Big Data, companies collecting Big Data are not held accountable to regulatory bodies. Spokeo, a people-finder service, is one such company. Spokeo sells detailed consumer profiles, including emails, physical addresses, phone numbers, marital status, occupation, family background, and more.[64] Although Spokeo profits from selling consumer profiles like credit reporting agencies do, it makes no warrants regarding accuracy of its profiles.[65]

Professor Anita Ramasastry notes that Spokeo and its ilk compile "robust data sets . . . that creditors want, and at present, it is unclear how actively they stop such companies from using this information." Because Spokeo's data sets are "used for a major life decision, such as whether someone might be hired or not, the person affected has no recourse, or ability to correct the errors." Professor Ramasastry has stated that Spokeo and other information aggregators "need[]to be subject to some regulatory scrutiny. At a minimum, consumers should have the ability to see their data, to correct it if needed, and to understand who might be buying their data for commercial purposes." Transparency is not sufficient and mechanisms for oversight are also needed.

Many state education departments, for example, lack adequate oversight of schools' use of technology and outsourcing of student records. Sheila Kaplan, a student privacy advocate and founder of Education New York, has endorsed state education chief privacy officers as an independent mechanism to "oversee, audit, consult, and report on matters that affect privacy and security of school records that contain personally identifiable information."[66]

---

[62] *See, e.g.*, EPIC et al., Letter to DHS Secretary Janet Napolitano, Re: TSA Racial Profiling Audit, Dec. 1, 2011, http://epic.org/privacy/airtravel/12-01-11-Coalition-Racial-Profiling-Audit-DHS-Letter.pdf.

[63] *See* Pam Dixon & Robert Gellman, World Privacy Forum, *The Scoring of America: How Secret Consumer Scores Threaten Your Privacy and Your Future* (Apr. 2, 2014), *available at* http://www.worldprivacyforum.org/wp content/uploads/2014/04/WPF_Scoring_of_America_April2014_fs.pdf.

[64] Spokeo, www.Spokeo.com (last visited Apr. 4, 2014).

[65] Anita Ramasastry, *The Spokeo Lawsuit and the Perils of the New People Finder Companies*, http://verdict.justia.com/2014/02/11/spokeo-lawsuit-perils-new-people-finder-companies.

[66] MODEL STATE LAW: CHIEF PRIVACY OFFICE FOR EDUCATION ACT § 1 (Sheila Kaplan, Educ. N.Y.), *available at* http://educationnewyork.com/files/CPOforED-2-01.pdf.

B. Support for Privacy Coalition Principles

As a starting point for a policy framework that could address the challenges of Big Data, EPIC supports the Principles set out by the Privacy Coalition and recommends that the White House incorporate these principles in its report:

> **TRANSPARENCY**: Entities that collect personal information should be transparent about what information they collect, how they collect it, who will have access to it, and how it is intended to be used. Furthermore, the algorithms employed in big data should be made available to the public.
>
> **OVERSIGHT**: Independent mechanisms should be put in place to assure the integrity of the data and the algorithms that analyze the data. These mechanisms should help ensure the accuracy and the fairness of the decision-making.
>
> **ACCOUNTABILITY**: Entities that improperly use data or algorithms for profiling or discrimination should be held accountable. Individuals should have clear recourse to remedies to address unfair decisions about them using their data. They should be able to easily access and correct inaccurate information collected about them.
>
> **ROBUST PRIVACY TECHNIQUES**: Techniques that help obtain the advantages of big data while minimizing privacy risks should be encouraged. But these techniques must be robust, scalable, provable, and practical. And solutions that may be many years into the future provide no practical benefit today.
>
> **MEANINGFUL EVALUATION**: Entities that use big data should evaluate its usefulness on an ongoing basis and refrain from collecting and retaining data that is not necessary for its intended purpose. We have learned that the massive metadata program created by the NSA has played virtually no role in any significant terrorism investigation. We suspect this is true also for many other "big data" programs.
>
> **CONTROL**: Individuals should be able to exercise control over the data they create or is associated with them, and decide whether the data should be collected and how it should be used if collected.[67]

These requirements are needed as entities try to take advantage of big data more and more without taking on the responsibility that should come with it. Greater transparency is an important place to start. As a recent Senate Majority report noted about Data Brokers, "Since data brokers generally collect information without the consumers' knowledge, consumers have limited means of knowing how the companies obtain their information, whether it's accurate, and for what purposes they are using it."[68] And public availability of data should not excuse companies or the government from being responsible data stewards. As danah boyd and Kate Crawford note, "The process of evaluating the research ethics cannot be ignored simply because the data is seemingly accessible. Researchers must keep asking themselves – and their colleagues – about the ethics of their data collection, analysis, and publication."[69] This same sentiment should apply to all entities that collect data.

---

[67] Privacy Coalition Letter on Big Data and the Future of Privacy, Mar. 31, 2014, *available at* http://privacycoalition.org/Big.Data.Coalition.Ltr.pdf .

[68] Office of the Oversight and Investigations Majority Staff, "A Review of the Data Broker Industry: Collection, Use, and Sale of Consumer Data for Marketing Purposes," 5 (Dec. 18, 2013), *available at* http://www.commerce.senate.gov/public/?a=Files.Serve&File_id=0d2b3642-6221-4888-a631-08f2f255b577.

[69] danah boyd and Kate Crawford. *Six Provocations for Big Data* at 11. Research paper presented at Oxford Internet Institute's "A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society" (Sep. 21, 2011), http://ssrn.com/abstract=1926431.

C. *Privacy Enhancing Techniques and Other Practices*

There are other recommendations that should be incorporated in the White House report. In the consumer space, University of Washington Law Professor Ryan Calo has suggested the creation of Consumer Subject Review Boards ("CSRBs"). CSRBs would be internal committees within private companies and they would operate under predetermined ethical rules to adequately vet and oversee the big data privacy implications of consumer behavioral research.[70] Among other benefits, CSRBs "could increase regulatory certainty, perhaps forming the basis for an FTC safe harbor if sufficiently robust and transparent" and they "could add a measure of legitimacy to the study of consumers for profit."[71]

Additionally, the public and private sector should implement Privacy Enhancing Technologies ("PETs") that "minimize or eliminate the collection of personally identifiable information."[72] Computer scientists have created various privacy enhancing mechanisms that should be deployed in the Big Data space. Distinguished Scientist at Microsoft Research Cynthia Dwork has espoused "differential privacy" as a "privacy-preserving analysis."[73] Differential privacy "ensures that the removal or addition of a single database item does not (substantially) affect the outcome of any analysis."[74] Although not an "absolute guarantee of privacy," differential privacy "ensures that only a limited amount of additional risk is incurred by participating in the socially beneficial databases."[75] Current FTC Chief Technologist Latanya Sweeney has created various algorithms that maintain confidentiality by "providing the most general version of the data."[76]

Jeff Jonas, Chief Scientist for the IBM Analytics Groups, describes the need to "bake in" privacy protection by, for example, "the ability to anonymize the data at the edge, where it lives in the host system, before you bring it together to share it and combine it with other data."[77]

The techniques are particularly important to address the potential abuses of predictive analytics. Where decisions are being made about individuals using predictive analytics, a process is needed to ensure the fairness of the decision. The "technological due process" as described by Danielle Citron provides a good basis for ensuring fairness in automated decisions.[78] Citron suggests audit trails that provide the information used to make a determination would provide transparency to users and a means to affectively challenge these decisions. The audit trails would also provides a means of oversight and accountability in the use of predictive analytics.

5. International Guidelines Provide Models for Better Privacy Protection

In addition to the Organisation for Economic Cooperation and Development Guidelines, there are other international frameworks that can serve as models to protect privacy in big data. In 2012, the European Commission proposed the "EU General Data Protection Regulation," ("GDPR") which has gained support from numerous U.S. consumer organizations. U.S. groups support the Regulation because it "establishes single, national data protection authorities in each [EU] member state," "adopts several innovative approaches to privacy

[70] *Id*.
[71] *Id*.
[72] Testimony and Statement for the Record of Marc Rotenberg, Executive Director, EPIC, Hearing on Privacy in the Commercial World, Before the Committee on Commerce, Trade, and Consumer Protection (Mar. 1, 2001), http://epic.org/privacy/testimony_0301.html; *See also* Herbert Burkert, *Privacy Enhancing Technologies: Typology, Critique Vision* in PHIL E AGRE AND MARC ROTENBERG, TECHNOLOGY AND PRIVACY: THE NEW LANDSCAPE 125-42 (MIT Press 1998).
[73] Cynthia Dwork, *Differential Privacy: A Survey of Results*, 1, 2008, http://www.cs.ucdavis.edu/~franklin/ecs289/2010/dwork_2008.pdf.
[74] *Id*. at 2.
[75] *Id*. at 2-3.
[76] Latanya Sweeney, *Datafly: a System for Providing Anonymity in Medical Data*, 15, http://dataprivacylab.org/datafly/paper2.pdf.
[77] IBM's Jeff Jonas on Baking Data Privacy into Predictive Analytics, *Data Informed*, Nov. 20, 2013, http://data-informed.com/ibms-jeff-jonas-baking-data-privacy-predictive-analytics/#sthash.hBM0lg1N.dpuf
[78] Danielle Keats Citron, Technological Due Process (2008).

protection, such as privacy by design and privacy by default," and "builds on the right to data deletion."[79] In 2013, the European Parliament Committee approved the regulation. [80]

The Madrid Declaration is an international "commitment to privacy protection" that "reaffirms international instruments for privacy protection, identifies new challenges, and call[s] for concrete actions."[81] Formally endorsed by hundreds of domestic and international civil society groups, privacy experts, and individuals, the Declaration promotes ten propositions concerning data protection.[82] For example, it reaffirms support for Fair Information Practice global implementation, genuine Privacy Enhancing techniques and Privacy Impact Assessments, and "independent data protection authorities."[83] It calls for a moratorium on mass surveillance technology; including body scanners, facial recognition, and RFID tracking, "subject to a full and transparent evaluation by independent authorities and democratic debate."[84]

The Council of Europe Convention 108 is an agreement, signed by the member states of the Council of Europe in 1995, which "protects the individual against abuses which may accompany the collection and processing of personal data and which seeks to regulate at the same time the transfrontier flow of personal data."[85] Convention 108 imposes certain rules regarding the methods by which signatory countries must regulate personal data collection and retention, and also forbids processing of "sensitive" data on a person's race, politics, health, religion, sexual life, criminal record, etc., in the absence of proper legal safeguards. The Convention also "enshrines the individual's right to know that information is stored on him or her and, if necessary, to have it corrected."[86] The Convention still remains the only binding international legal instrument with a worldwide scope of application in the field of data privacy, open to any country, including countries that are not Members of the Council of Europe.[87]

While Convention 108 itself originated in the Council of Europe, the United States' demonstrated interest in the privacy principles protected by the Convention align perfectly. The principles underlying Convention 108 are directly based on the Universal Declaration of Human Rights, adopted by the United Nations in 1948.[88] It was the United States and Eleanor Roosevelt that helped craft the Universal Declaration, and it was the United States that ratified the Council of Europe Convention on Cybercrime and urged its allies to do the same.

Moreover, technical experts and legal scholars have expressed support for the U.S. ratification of Convention 108. On January 28, 2010, twenty-nine members of the EPIC Advisory Board wrote to then Secretary of State Hillary Rodham Clinton to urge that the United States begin the process of ratification of Council of Europe Convention 108.[89] In that letter, the members of the EPIC Advisory Board explained, "Just as communications networks can be used for good and ill, so too can computer technology. It can help sustain aid programs, spur innovation, and encourage economic growth. Or it can track the activities of dissidents, monitor the

---

[79] Letter from U.S. Consumer Organizations on EU General Data Protection Regulation to Jan Philipp Albrecht, Rapporteur, Comm. on Civil Liberties, Justice and Home Affairs, and Lara Comi, Rapporteur, Comm. on Internal Market and Consumer Protection, European Parliament (Sept. 5, 2012), *available at* https://epic.org/privacy/intl/US-Cons-Grps-Support-EU-Priv-Law.pdf.

[80] Press Release, European Parliament, Civil Liberties MEPs Pave the Way for Stronger Data Protection in the EU (Oct. 21, 2013), http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-%2f%2fEP%2f%2fTEXT%2bIM-PRESS%2b20131021IPR22706%2b0%2bDOC%2bXML%2bV0%2f%2fEN&language=EN.

[81] Madrid Privacy Declaration: Global Privacy Standards for a Global World, The Public Voice (Nov. 3, 2009), *available at* http://thepublicvoice.org/madrid-declaration/.

[82] *Id*.

[83] *Id*.

[84] *Id*.

[85] Council of Europe: Treaty Office: Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (last accessed May 9, 2013), *available at* http://conventions.coe.int/Treaty/en/Summaries/Html/108.htm.

[86] *Id*.

[87] *Council of Europe Privacy Convention*, EPIC, http://epic.org/privacy/intl/coeconvention/ (last visited Apr. 4, 2014).

[88] Letter from EPIC Advisory Board to Secretary Clinton, January 28, 2010, *available at* http://epic.org/privacy/intl/EPIC_Clinton_ltr_1-10.pdf.

[89] *Id.*

private lives of citizens, and maintain elaborate systems of identification for laborers and immigrants... the protection of privacy is a fundamental human right. In the 21st century, it may become one of the most critical human rights of all. Civil society organizations from around the world have recently asked that countries which have not yet ratified the Council of Europe Convention 108 and the Protocol of 2001 to do so as expeditiously as possible."[90] The next day, the U.S. Privacy Coalition, comprised of twelve privacy groups, including EPIC, also signed a resolution to the U.S. Senate endorsing Convention 108.[91]

The White House should look to these international models in developing necessary safeguards for the challenge of "Big Data." Technology has outpaced the law, but it is not too late to establish the safeguards that allow for the insights offered by big data, while protecting the fundamental rights of Americans

Conclusion

EPIC appreciates this opportunity to comment and looks forward to continued public engagement on the issue of big data and privacy.

Respectfully submitted,

Marc Rotenberg
EPIC President and Executive Director

Julia Horwitz
EPIC Consumer Protection Counsel

Jeramie Scott
EPIC National Security Counsel

Khaliah Barnes
EPIC Administrative Law Counsel

Electronic Privacy Information Center (EPIC)
1718 Connecticut Avenue, NW, Suite 200
Washington, DC 20009
(202) 483-1140

[90] *Id.*
[91] Privacy Coalition, Resolution, United States Senate, Jan. 29, 2009 *available at* http://privacycoalition.org/resolution-privacy_day.pdf.

Response to the Science and Technology Policy Office Request for Information about Government “Big Data”

April 4th, 2014

By, Kaliya “Identity Woman”

Independent advocate for the rights and dignity of our Digital Selves

Blog: http://www.identitywoman.net

Founder of efemurl, consumer coop web platform.
Site: http://www.efemurl.com

Founder and Executive Director of the **Personal Data Ecosystem Consortium**
Site: http://www.pde.cc

World Economic Forum, **Young Global Leader**, Class of 2012
Contributor to the World Economic Forum **Rethinking Personal Data Project**
Site: http://www.weforum.org/issues/rethinking-personal-data

Co-Founder 2005, Co-Producer, Co-Facilitator of the **Internet Identity Workshop**
Site: http://www.internetidentityworkshop.com

Network Director **Planetwrok** NGO
Site http://www.planetwork.net

Small Business & Entrepreneur Rep on the Management Council of the Identity Ecosystem Steering Group [IDESG] **National Strategy for Trusted Identities in Cyberspace** [NSTIC].
Site: http://www.idecosystem.org

Founder of **She’s Geek**y an Unconference for Women in STEM
Site: http://www.shesgeeky.org

Principle of Unconfernece.net a firm that Designs and Facilitates Unconferneces
Site: http://www.unconference.net

Fellow of the Co-Intelligence Institute
Site: http://www.co-intellignece.org

Member of the National Coalition for Dialogue and Deliberation
Site:http://www.ncdd.org

Introduction & Context

I am a member of the bridge generation a 1.5 generation digital immigrant. Those older then me who didn’t have the web in college are first generation immigrants and those younger then me are digital natives (they grew up with the tools). I was born in 1976, I had a rotary telephone and remember the black and white TV we had with a nob to turn through the 13 channels. I remember the world before the saturation of digital information and communication technologies.

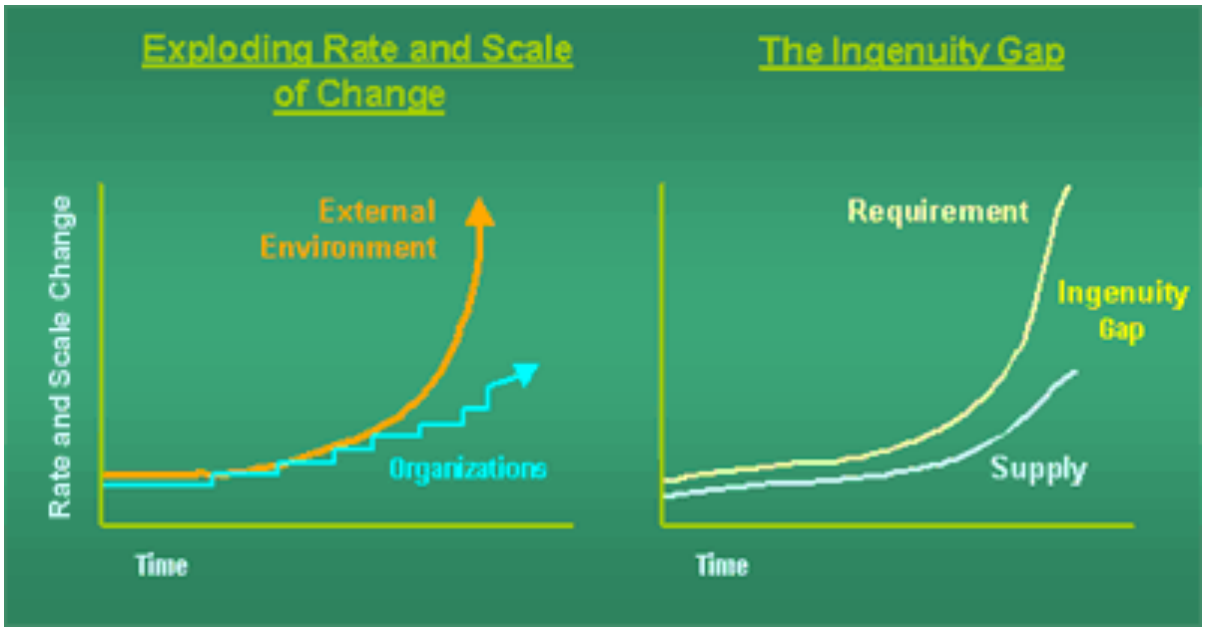
I have lived in the SF Bay Area since 1995 when I began as a freshman at UC Berkeley. While in college I watched the first tech boom come and go. I studied Political Economy and Human Rights and while there learned about and was inspired by the practice of Nonviolence embodied

by the struggles lead by Gandhi and Martin Luther King for large scale justice. Out of college I worked with various organizations focused in this area.

I first learned about digital identity technologies via a community and network - Planetwork. The founders where inspired to host a conference in 2000 with the theme ***Global Ecology and Information Technology*** because they saw ICT as the only thing growing as fast as the crises we face as a species on this planet.

The speakers were amazing the working applications and painted a vision of how technology could play a significant role in helping us as humanity deal with the pending/and ongoing environmental crisis was inspiring. Jim Fournier, one of the founders presented a vision of how ICT could helps re-align with the natural systems of the planet - getting to Meta-Nature (see http://www.metanature.org)

The work of Douglas Engelbart, a computer technology pioneer (he is most well known for inventing the mouse) who's lab at Stanford Research Institute in the 60's focused on Augmenting Human Intelligence and collective capacity for collaboration to solve problems was highlighted given the rapid need to solve problems because of change. This same gap between change and the capacity to deal with it has also been highlighted by contemporary Canadian intellectual Thomas Homer Dixon



Engelbart (right) shows the lagging performance of human organizations in the face of an exploding rate and scale of change. Homer-Dixon (left) sketches a widening gap between the need for and availability of ingenuity. Notice how the curves for ingenuity requirement and supply veer off to the right.

Source: https://web.archive.org/web/20020208050313/http://www.fleabyte.org/eic-6.html#2A12

Douglas Engelbart did not believe that data alone was useful in helping us think / take action together to in short be collectively intelligent. He emphasized that learning how to learn in groups, organizations and as whole societies was also key. Thomas Homer Dixon also emphasizes social, political and economic innovations as necessary to bridge the Ingenuity Gap.

Data is not enough.

It must be used to help us be collectively wise.

We have collected massive amounts of data about the changing climate since the 1970's. More and more and more "big data". Are we actually shifting our actions as a whole country? Do we have the practices we need to collectively discern the meaning of the data we have access to now as a society?

Question 2 asks: Are there specific sectors or types of uses that should receive more government and/or public attention?

Yes - I think it is critical to focus government attention on how to use "Big Data" generated by the society is used to help make meaning about the whole society by "the people" via methods of collective discernment the focusing public attention using democratic methods that lie outside of voting for representatives every few years. These methods have been extensively documented by the National Coalition for Dialogue and Deliberation and Tom Atlee at the Co-Intelligence Institute and his book the Tao of Democracy.

Question 2 asks: What types of uses of big data could measurably improve outcomes or productivity with further government action, funding, or research?

These types of citizen centered discernment can be used to focus government attention and action on critical problems that have seemed politically intractable. Why do we have 1 in 7 Americans so poor they must rely on food stamps. Big Data with public discernment and mandates for government action could be powerful in helping address the challenges we face. Our society is complex system and big data is a key way to get at the results of changes over time and respond to changes. We must develop not just the technological capacity to look at big data but the social tools and collective capacity to make meaning from and take action based on it.

We must also be prepared when collectively looking at issues using "Big Data" to encounter truths that trigger collective shame and vulnerability reactions. It is shameful that so many people live in poverty, it is shameful that so many of those in poverty are descendants of slaves brought here from Africa, the collective wounds and wrongs of the past will surface in meaningful and truthful engagement with "Big Data". Actively acknowledging these and other difficult truths by truly engaging with the data and making real meaning with it will require a new collective emotional maturity. It might be necessary to work with people like Brene Brown (author of Daring Greatly) to consider how to work

Question 1 asks: (1) What are the public policy implications of the collection, storage, analysis, and use of big data?

In the request for information you outline that information about people's purchases, conversations, social networks, movements and physical identities is being collected, stored, analyzed and used.

Lets start out by asking **who** is collecting it and **how**?

Right now the applications, devices, software, hardware and platforms we use; along with the stores we buy from, the credit card and banks we use, sites we read on, music and radio we stream all follow us and track our behaviors. This tracking is not just from our digital interactions but our physical movements everywhere we go.

All of these types of entities actively seek persistent correlateable identifiers from us and then package up our data and sell it on data markets. Only a few pieces of PII are needed to correlate data from different sources together and weave a comprehensive digital dossier.

The data we actively volunteer (knowingly give or post) is put together with data passively generated and collected (geolocation logs from our phones or GPS devices) and can be used to infer much about us including quite sensitive information, for example, our religious beliefs (because we go to particular a house of worship once a week).

The proposed Commercial Privacy Bill of Rights Act of 2011

Defines "Personally identifiable information," as (1) the first and last name of an individual; (2) postal (residential) address; (3) email address; (4) telephone number or mobile number; (5) social security number; (6) credit card number; (7) "unique identifier information that alone can be used to identify a specific individual"; or (8) "biometric data," including fingerprints and retina scans.

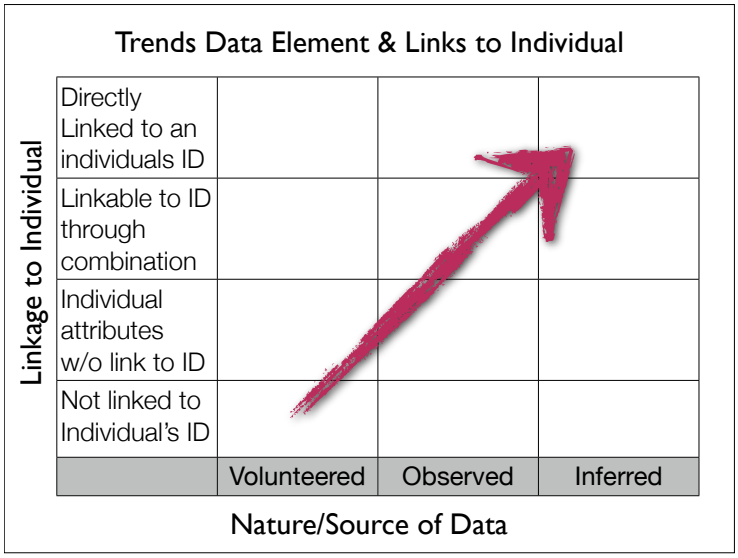
The definition of PII also covers any of the following if stored or used along with (1) through (8) above: (1) date of birth; (2) birth certificate number; (3) place of birth; (4) unique identifier information "that alone cannot be used to identify a specific individual;" (5) "precise geographic location," excluding general geographic information that can be derived from an IP address; (6) information about an individual's use of "voice services, regardless of technology used;" and (7) a catch-all.

The Bill also contains a third category of information which it calls "sensitive" PII, which includes medical/ health information and the "religious affiliation" of an individual.

http://blog.ericgoldman.org/archives/2011/04/a_look_at_the_c.htm

Each of us have extensive digital dossiers about us held by business entities we have no business relationship with or knowledge that they hold information about us. **Why** is this being done? This information is bought and sold and used by commercial entities to target us, to make decisions about whether to do business with us or not and on what terms.

The Re-Thinking Personal Data project at the World Economic Forum came up with the following diagram to articulate the situation. Things get creepier as one moves up and to the right.



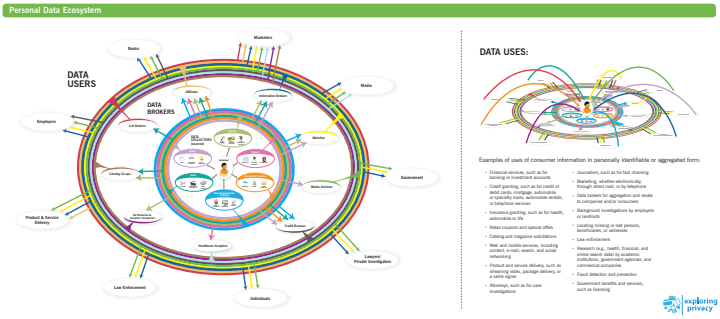
We - the people - who the generate the data are not empowered with the tools to collect their own data.

Since 2010 I have been catalyzing the global community of entrepreneurs seeking to build those tools and develop a Personal Data Ecosystem - at its core are services where individuals collect and manage their own data. Just like we use banks to collect and manage our money, why not have data banks, data vaults, data stores or what is now being called Personal Clouds to store the data we voluntarily and implicitly generate.

Question 1 asks: Do the current U.S. policy framework and privacy proposals for protecting consumer privacy and government use of data adequately address issues raised by big data analytics?

No they do not. This question is broad and the answer is no from several different perspectives.

Daniel Solove wrote in his 2004 book, *The Digital Person* about the nature of the data broker industry. In the 10 years since it has continued to grow - unabated. A map of what happens to our data was submitted to the Federal Trade Commission in 2009 called Exploring Privacy that maps out what happens to our data after we interact with retailers, websites, cable companies, medical offices - basically any aspect of our day-to-day lives were a record of a transaction is entered into a computer or an electronic device is involved. More and more data is being collected with basically no protections on how it can be aggregated to profile us as particular individuals.



http://www.ftc.gov/sites/default/files/documents/public_events/exploring-privacy-roundtable-series/personaldataecosystem.pdf

The current practices on the web involve the placement various forms of device identification, browser configuration, tracking beacons and cookies. These all in effect enable private companies to stalk us as we move around the web. These practices if the patterns they follow were replicated by people or little physical devices placed on us by as we moved about the physical world would be prohibited.

Continuing on with question 1: Question 1 asks: Do the current U.S. policy framework and privacy proposals for protecting consumer privacy and government use of data adequately address issues raised by big data analytics?

One of the big issues raised by big data analytics is the information asymmetry that exists between individuals and communities they are a part of and the corporate and government entities that are the collecting, storing, analyzing and using data.

People are not cognizant of how much information they voluntarily give away and how much meta-data their every day interactions on digital devices and shopping in a retail store generate. One way the government change the current policy framework in one simple way.

It could mandate that individual consumer/citizens have a right to

- a copy of all data (that a company is collecting) while they are using digital devices (whether they own them or not)
- a right to digital copy of all records of transactions they do (right now one gets a paper receipt at a grocery store and they have a digital record of the transaction).

The government has worked on initiatives around various buttons - blue for medical records, green for utility company records, red for educational records and the latest I have heard about is gold for financial records. We should have buttons for "everything" - the data we generate is as much ours as it is the firms who are providing us services. Because these are by their very nature digital objects - that is bits, bits of data. The frame of property ownership is not an appropriate one. We must begin to anchor the conversation in rights and responsibilities. What is my right to the data I generate. What are the responsibilities of the company or services provider I am using (and in the process generating data). We both should have rights to the data and responsibilities. We could use some of the deliberative democracy methods I mentioned earlier to consider what those should be and then move to establish them in law.

The right to our own data will kick start the market for personal clouds and new ethical data markets.

Question 3: What technological trends or key technologies will affect the collection, storage, analysis and use of big data?

Personal Clouds are emerging as a new technology that is designed to put people at the center of their own data lives. Some of tongue and cheek in response to big data called this "small data". It is empowering the individual to collect their own "Big Data" and use tools and services to use analytical data methods on one's self.

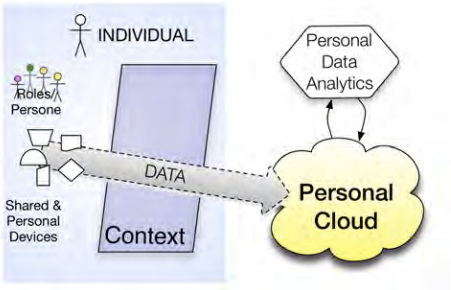
It means that individuals can then see the image of themselves that they are releasing via the activities they do that generate data. The Personal Data Ecosystem that I lead wrote to both the Federal Trade Commission and Federal Communications commission in response to their respective white and green papers on privacy.

* Link to FTC response: https://dl.dropbox.com/u/13895906/FTC-DNTResponse021811.pdf

* Link to FCC response: http://www.ntia.doc.gov/comments/101214614-0614-01/comment.cfm?e=01D6196B-4C69-4C2B-8DA5-6D78D64AF527

Since then we have further articulated the different market models for data that align with consumer/citizen interest.

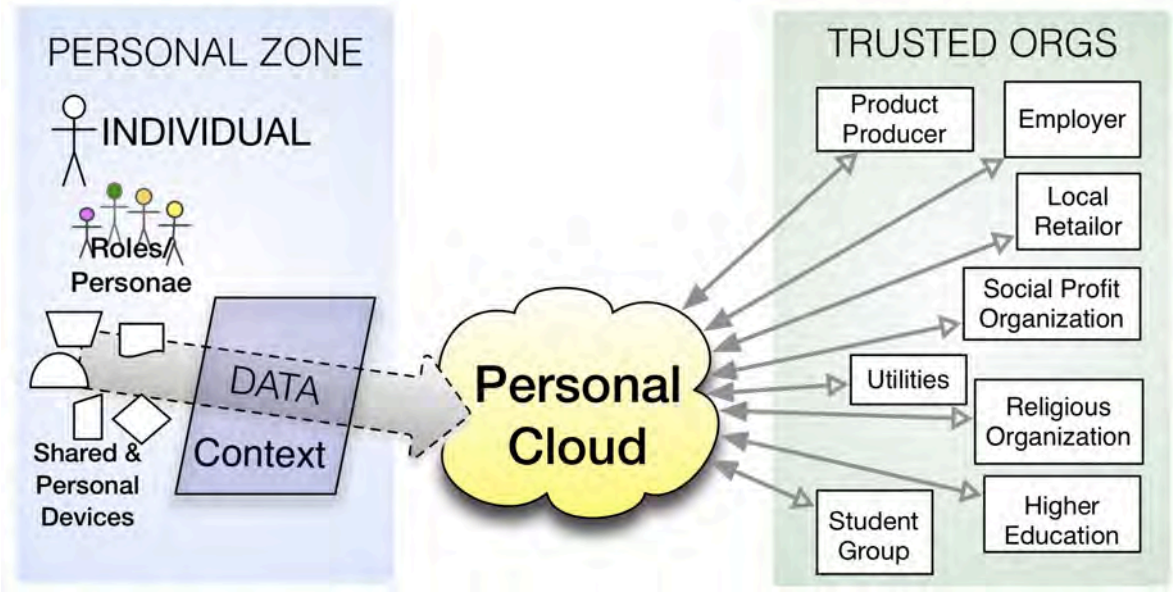
Meaningful, understandable control over the collection and use of personal data by individuals enables companies, organizations, and governments to create value while increasing public trust. Three broad categories emerged: *Vendor Relationship Management, Infomediary Markets,* and *Data Aggregation Markets*. These models place the user at their center, and reflect ways that context will influence the roles a person plays or how they behave.



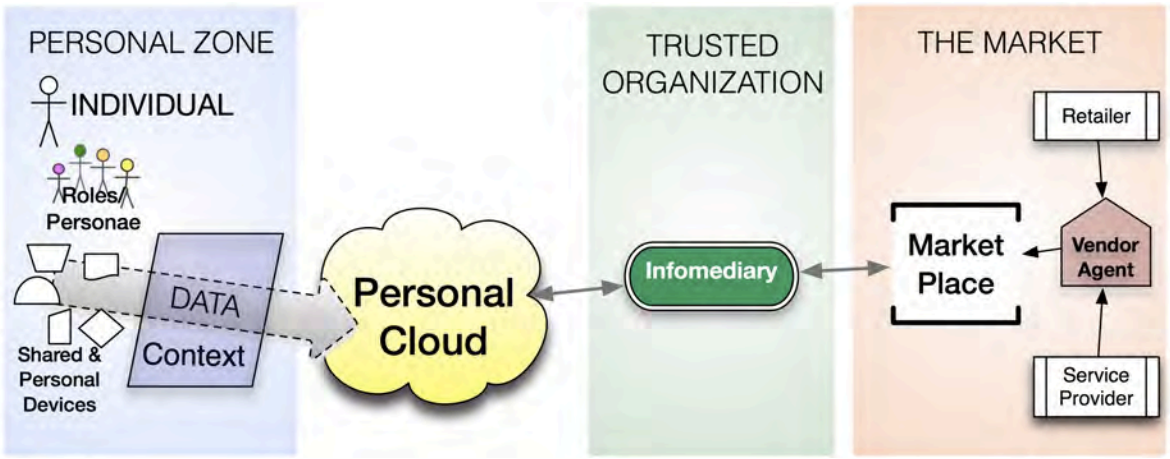
At the heart of each model is a *personal cloud* that helps people to:
•Securely collect and store data they generate in daily life
•Connect to service providers so they get value from their data
•Limit how the data they provide is used by service providers

Vendor Relationship Management

CRM tools are standard in today's businesses where they are used by companies to manage their relationships with customers and potential customers.
Vendor Relationship Management (VRM) tools work on behalf of the individual to connect people to businesses, organizations and public services. Individuals can share more detailed information with companies they like. VRM offers a direct channel people control between themselves and the organizations they engage. Data sharing is governed by terms set by the user. In this model, people have the power to withdraw from relating to the vendor and their choice will be respected.

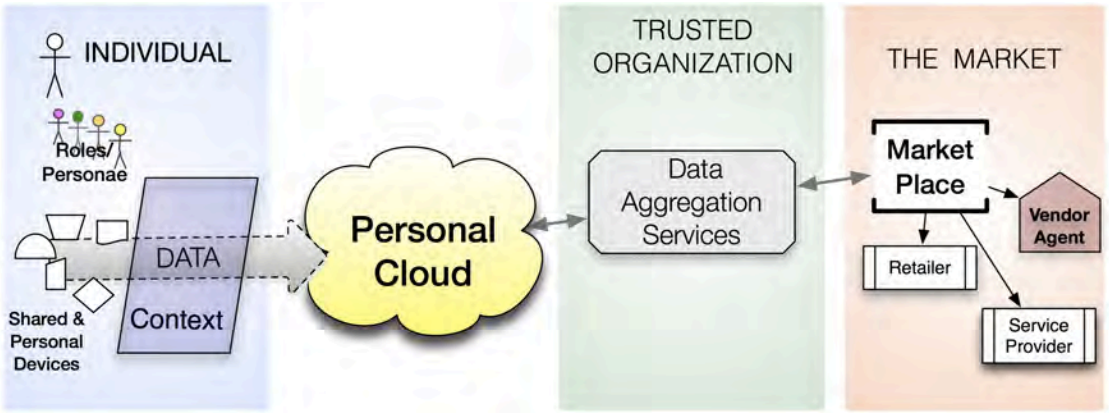


Infomediary Markets



In markets where vendors seek attention from potential customers, *Infomediaries* act as data brokers on the users' behalf. The Infomediaries create permission-based channels, based on accurate personal data that the user provides, but without revealing personally identifying information in the market.
Companies want to find customers for their products and services, governments want to reach citizens with important relevant information, and organizations want to attract new interested members. Today's marketing methods can be ineffective, annoying and intrusive to get at people's personally identifiable information without their awareness or consent.

Data Aggregation Markets

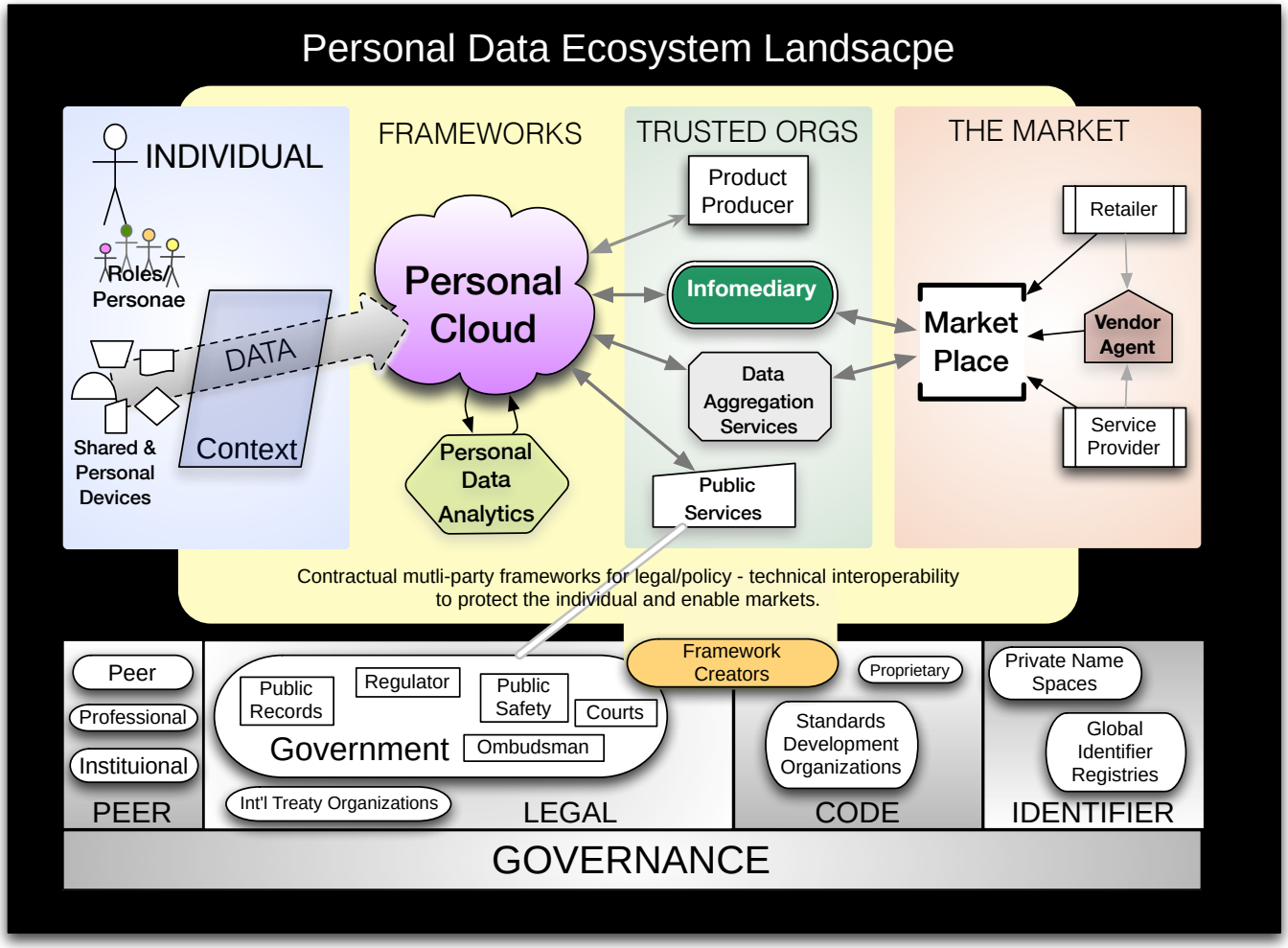


Data Aggregation Services read the details of users' data, but provide companies, governments, and researchers with summary or aggregate values. Corporate research and marketing have many legitimate, ethical uses of "big data." Individuals willingly participate when they believe their privacy will be respected. With that trust, they willingly provide more honest, complete, and detailed personal information to these services. This market model was used for media audience metrics for fifty years by companies like Nielsen and Arbitron.
Big Data is big business and many firms today collect vast troves of information about specific individuals and households, from public records (Axiom), social networks (RapLeaf), or foot traffic via mobile phones (Sense Networks)

Question 3: Are there particularly promising technologies or new practices for safeguarding privacy while enabling effective uses of big data?

A critical aspect of how the system above works is accountability frameworks. Sometimes called Trust Frameworks - these technology/policy sandwiches enable information to flow. If regular citizens are to trust these systems I continue believe that we should not call the frameworks by the name of the feeling we hope that they foster in the overall system. We have to be able to talk about them and ask about the trustability or trustworthiness of them. What is the trustability of a particular trust framework is just so awkward. This is a link to a post that I wrote covering this issue - The Trouble with Trust and the Case for Accountability Frameworks. http://www.identitywoman.net/the-trouble-with-trust-the-case-for-accountability-frameworks

Here is a diagram of the whole ecosystem together - all three market models outlined above, accountability frameworks (and their straddling of legal and code governance) along with other modalities of governance at play in the system. The next page has a description of each box found on this landscape.



Personal Data Landscape Term Definitions

Individual: A person

Devices: Mobile phones, computers, self tracking devices, medical monitoring devices, e-readers.

Context: Where a person is (home, school, work). The Role they are playing (parent, coach, spouse, employee, supervisor, athletic team member). Persona they are presenting (video game player, professional, goofy hobby identity).

Data: The bits generated explicitly such as photos, tweets, status updates.

Frameworks: Contractual mutli-party frameworks connect legal/policy agreements to technical interoperability to protect the individual and enable markets. The Personal Cloud service provider is at the heart of these frameworks, chosen by the end user, and works on their behalf managing their data and its participation in the framework.

Personal Data Analytics: Services that help people gain insight into their own personal data. An example being one's daily health status or a personal annual report.

[Trusted Organizations]

Product Producer: This is an example of a Vendor Relationship Management connection where a consumer who bought a product from a producer manages an open channel with the maker of the product they bought and willingly share information under favorable terms they the user set.

Infomediary: A service trusted to have insight into a person's data and working on their behalf. They have an individual's personally identifiable information (PII) and protect that data and put it to use.

Data Aggregation Services: Services create aggregate data sets from personal data, like music listening habits. Aggregators may compensate people for their data, people may share altruistically, or people may unknowingly share.

Public Services: Governments delivering services to their constituents can enable use of personal data stores for better access and data quality.

[The Market]

Market Place: This is where an Individual's business agents with PII meet Vendor agents without PII.

Retailers: Companies that sell goods to customers.
Service Providers: Companies that provide services to people.

Vendor Agents: Companies that help retailers and service providers find good potential leads. They do not have personally identifiable information.

[Governance]

How systems are regulated take many forms. Governance starts with laws and regulation but also includes cultural practices, business norms, and, in digital systems, how identifiers are allocated and the code that connects them.

LEGAL:

Government: plays many roles in the systems:

Regulator: Governments set baseline rules for how markets work. They provide the court system where contract law is adjudicated.

Public Records: Governments record births, marriages, divorces, deaths along with licensing, and property title registries.

Public Safety: Policing and law enforcement.
Ombudsman: Many states have a data protection commissioner who protects constituents.

International Treaty Organizations: They support the coordination of international treaties and provide a meta-international law that hold governments accountable to each other.

CODE: Computer code and how it runs determines what is possible in computer systems. The phrase "Code is Law" was popularized by Lawrence Lessig.

Standards Development Organizations: Bruce Sterling said "If code is law then standards are like the Senate." Standards bodies agree on how code works regardless of the particular language it is written in or system it is running on. For example, the W3C standardizes the HTML specification for presenting web pages.

IDENTIFIER: Networks run on identifiers for each endpoint. How these are allocated, and the terms and conditions of use in a network, govern the network.

Global Identifier Registries: Examples include the phone system, Domain Names, ISBN numbers, RFID.

Private Name Spaces: examples include Twitter, Skype, Google, Facebook etc.

PEER: This kind of governance is the most powerful in many ways and helps social systems operate.

Peer-to-Peer: People have opinions about each other and also about businesses and services they interact with - like Yelp for small businesses.

Professional: Doctors, lawyers, engineers, geologists, and architects are professions that peer regulate.

Institutional: Institutions figure out what other peer institutions are - such as banks worldwide in SWIFT.

Framework Creators: Organizations that create contractual legal-policy/technology frameworks that govern complex multi-party networks.

Response to the Science and Technology Policy Office Request for Information about Government “Big Data”

April 4th, 2014

By, Kaliya “Identity Woman”

Independent advocate for the rights and dignity of our Digital Selves

Blog: http://www.identitywoman.net

Founder of efemurl, consumer coop web platform.
Site: http://www.efemurl.com

Founder and Executive Director of the **Personal Data Ecosystem Consortium**
Site: http://www.pde.cc

World Economic Forum, **Young Global Leader**, Class of 2012
Contributor to the World Economic Forum **Rethinking Personal Data Project**
Site: http://www.weforum.org/issues/rethinking-personal-data

Co-Founder 2005, Co-Producer, Co-Facilitator of the **Internet Identity Workshop**
Site: http://www.internetidentityworkshop.com

Network Director **Planetwrok** NGO
Site http://www.planetwork.net

Small Business & Entrepreneur Rep on the Management Council of the Identity Ecosystem Steering Group [IDESG] **National Strategy for Trusted Identities in Cyberspace** [NSTIC].
Site: http://www.idecosystem.org

Founder of **She’s Geek**y an Unconference for Women in STEM
Site: http://www.shesgeeky.org

Principle of Unconfernece.net a firm that Designs and Facilitates Unconferneces
Site: http://www.unconference.net

Fellow of the Co-Intelligence Institute
Site: http://www.co-intellignece.org

Member of the National Coalition for Dialogue and Deliberation
Site:http://www.ncdd.org

Dear Ted Wackler,

I am a member of the bridge generation a 1.5 generation digital immigrant. Those older then me who didn’t have the web in college are first generation immigrants and those younger then me are digital natives (they grew up with the tools). I was born in 1976, I had a rotary telephone and remember the black and white TV we had with a nob to turn through the 13 channels. I remember the world before the saturation of digital information and communication technologies.

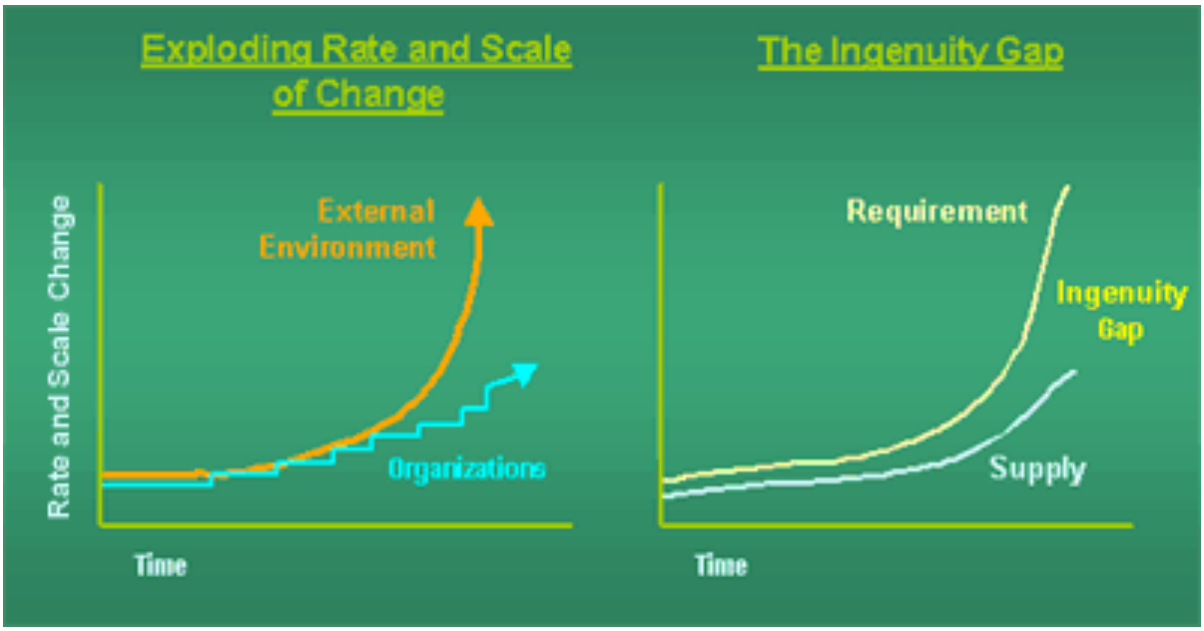
I have lived in the SF Bay Area since 1995 when I began as a freshman at UC Berkeley. While in college I watched the first tech boom come and go. I studied Political Economy and Human

Rights and while there learned about and was inspired by the practice of Nonviolence embodied by the struggles lead by Gandhi and Martin Luther King for large scale justice. Out of college I worked with various organizations focused in this area.

I first learned about digital identity technologies via a community and network - Planetwork (http://www.planetwork.net). The founders where inspired to host a conference in 2000 with the theme *Global Ecology and Information Technology* because they saw ICT as the only thing growing as fast as the crises we face as a species on this planet.

The speakers were amazing and along with working early stage applications painted a vision of how technology could play a significant role in helping us as humanity deal with the pending/and ongoing environmental crisis. Jim Fournier, one of the founders presented a vision of how ICT could helps re-align with the natural systems of the planet - getting to Meta-Nature (see http://www.metanature.org)

The work of Douglas Engelbart, a computer technology pioneer (he is most well known for inventing the mouse) who's lab at Stanford Research Institute in the 60's focused on Augmenting Human Intelligence and collective capacity for collaboration to solve problems problems arising from rapid change. This same gap between change and the capacity to deal with it has also been highlighted by contemporary Canadian intellectual Thomas Homer Dixon



Engelbart (right) shows the lagging performance of human organizations in the face of an exploding rate and scale of change. Homer-Dixon (left) sketches a widening gap between the need for and availability of ingenuity. Notice how the curves for ingenuity requirement and supply veer off to the right.

Source: https://web.archive.org/web/20020208050313/http://www.fleabyte.org/eic-6.html#2A12

Douglas Engelbart did not believe that data alone was useful in helping us think / take action together to in short be collectively intelligent. He emphasized that learning how to learn in groups, organizations and as whole societies was also key. Thomas Homer Dixon also emphasizes social, political and economic innovations as necessary to bridge the Ingenuity Gap.

Data is not enough.

It must be used to help us be collectively wise.

We have collected massive amounts of data about the changing climate since the 1970's. More and more and more "big data". Are we actually shifting our actions as a whole country? Do we have the practices we need to collectively discern the meaning of the data we have access to now as a society?

Question 2 asks: Are there specific sectors or types of uses that should receive more government and/or public attention?

Yes - I think it is critical to focus government attention on how actually make meaning from "Big Data" generated by society. This should involve the whole society, "the people" via methods of collective discernment, the focusing of public attention using democratic methods that lie outside of voting for representatives every few years. These methods have been extensively documented by the National Coalition for Dialogue and Deliberation and Tom Atlee at the Co-Intelligence Institute (http://www.co-intelligence.org) and his book the *Tao of Democracy* (http://www.taoofdemocracy.com). I worked with Tom several years ago to articulate how three of the methods that were most successful at supporting the emergence of a voice of "we the people" could be augmented with Technology.

Ways to Generate an Inclusive, Legitimate, Informed, Coherent and Trustworthy Voice of "We the People"

Type of Citizen Deliberative Council	Picking an Issue	Framing the Issue	Selecting Deliberators	Information and Expertise	Deliberations	Decision-making	Dissemination and Impact	Organizational Support
Citizen Jury	Picked by Convening Authority - • Government Agency • Large NGO • Corporation • University • Wisdom Council • Automatic part of government operations	Organizers usually create a "charge" naming options deliberators must choose among and describing pros, cons and tradeoffs	* Random selection, usually with stratified sampling to reflect demographic profile of the larger community * 12-24 jurors	Oversight committee of diverse partisans and/or respected neutral experts choose briefing materials and expert witnesses	* Normal agenda-based meeting facilitation, often includes values analysis and voting * 4-5 days	* Usually majority or supermajority vote * Generate findings and recommendations	* Results sent to convening authority and media (with varying degrees of publicity) * Wisdom council reports to community meeting + high expectation from participant selection * **Sometimes** -- other dialogues organized before, during and/or after -- officials take action or explain why not -- institutionalized outcomes, e.g., popular vote, legislative action, placement of findings in voters pamphlets…	* Professional (or other high quality) organizers * $20,000 and up
Consensus Conference		Citizen panel frames the issue within their mandate, in liaison with the organizers	* Random from whole country/community database and/or newspaper recruitment; select people who know little about the issue * 12-24 panelists	Similar to citizen jury, but citizens have final say on expert witnesses	* Moderated public hearings followed by facilitated consensus process * 2 briefing weekends, then 3-4 day conference	* Usually consensus, sometimes reporting the nature of any remaining differences * Generate findings and recommendations		* Professional (or other high quality) organizers * $30,000 and up
Wisdom Council	Picks its own issue(s)	Citizen panel frames and reframes the issue as they proceed through dynamic facilitation	* As close to pure random selection as possible, chosen in public ceremony to generate public interest * 12-24 members	Citizens are experts in their own experience, and can choose other experts if they wish.	* Dynamic facilitation of choice-creating process * 2-5 days, culminating in public meeting	* Usually emergent consensus, but sometimes a more crafted agreement * Generate statement		* Can be done by grass-roots citizens from manuals * $2,000 and up
Tao of Extreme Democracy (ideas for)	Some method of surfacing issues online on an ongoing basis through popular paricipation?	National Issues Forums-style issue framing, which provides 3-5 approaches w/ arguments for and against, trade-offs, values, etc., for each - and invites deliberators to move beyond them.	* Random selection (using demographics) from large pool of volunteers who have provided demographic information? * Random selection of and from diverse groups (NGOs, churches, unions, etc.)? * 24-100 or more deliberators	* Info from issue framings and web searches * Experts available from pools of diverse volunteer experts, accessible via all telecommunications media (online, teleconference, etc.)	* Volunteer facilitators following standard guidelines? * Numerous groups of deliberators simultaneously considering the same issue ("parallel processing" a la German "planning cells")?	* Probably supermajority * Mixing and matching members of diverse parallel groups may increase common sense agreements * Could have feedback between deliberators and public before decision made	Grassroots advocacy for recommendations, using blogs and MoveOn-type organizing, etc.	* Since not-for-profit and very experimental, needs major investment in experiments (high ROI of social change when successful!) * Needs grassroots support, especially from techies
Resources and Comments — ***Purpose --*** *To facilitate the emergence of an inclusive, legitimate, informed, coherent and trustworthy voice of We the People*	* *Need to cover popular issues and emerging dangerous ones*	* Universities; graduate students * Existing "issue books" * Wikipedia of issue framings co-created through a citizen journalism movement * *Must be demonstrably inclusive and/or unbiased*	* Database and Selection software * *Needs to be as unbiased and inclusive (wide spectrum diversity) as feasible, to nurture both legitimacy and collective wisdom*	* Wikipedia pattern map for solutions? * *Needs to be as unbiased and inclusive (wide spectrum diversity) as feasible, to nurture both legitimacy and collective wisdom*	* Dialogue Circles * NIF/Kettering * *Needs facilitation to help diverse views evolve towards wise agreement*	• Synanim.com • *The smaller the group, the more agreement they must demonstrate in order to be seen as representing the whole community*	* *Building expectations builds "buzz" afterwards* * *Partisan advocacy tools can be used to advocate for inclusive solutions*	* *Needs to be easily replicable and inexpensive - and hopefully very appealing ("sticky meme")* Tom Atlee (w/Kaliya Hamlin) cii@igc.org co-intelligence.org

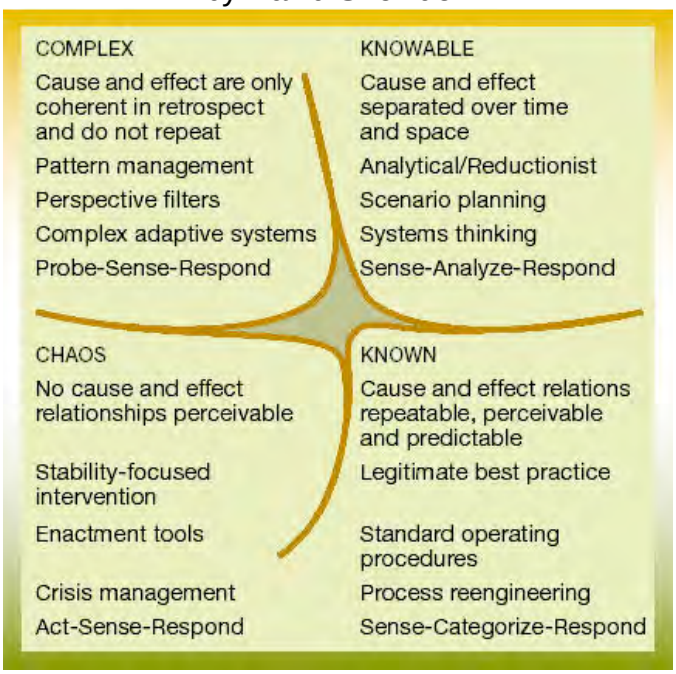
1. In democratic theory, a leader, institution, system or policy is ***legitimate*** to the extent people will voluntarily go along with it without being coerced. Force -- importing *extrinsic* energy into a system -- does not achieve stable outcomes. Intelligence (which collectively involves dialogue) is an alternative to force -- learning the *intrinsic* energies, tendencies and patterns that can be worked *with* (as in permaculture).

2. Things to consider: Imagineering. Wisdom Civilization. Civic Intelligence / CPSR. Anthony Judge. CWPR. URI. "How Not to Make a Decision." Pattern language. Noo. Edmonton Sean. *The future is here -- it's just not well distributed yet!*

Question 2 asks: What types of uses of big data could measurably improve outcomes or productivity with further government action, funding, or research?

These types of citizen centered discernment can be used to focus government attention and action on critical problems that have seemed politically intractable. Why do we have 1 in 7 Americans so poor they must rely on food stamps. Big Data with public discernment and mandates for government action could be powerful in helping address the challenges we face. Our society is a complex adaptive system. Using the Cynefin Framework to understand how we need to interact with such systems we must Probe, Sense and Respond. Big data used to address key issues can be used to measure results and effectiveness - to sense how effective various experiments to address challenges fair. We must develop not just the technological capacity to look at big data but the social tools and collective capacity to make meaning from and take action based on it.

CYNEFIN FRAMEWORK
by David Snowden



We also must be prepared when collectively looking at issues using “Big Data” to encounter truths that trigger collective shame and vulnerability reactions. It is shameful that so many people live in poverty, it is shameful that so many of those in poverty are descendants of slaves brought here from Africa, the collective wounds and wrongs of the past will surface in meaningful and truthful engagement with “Big Data”. Actively acknowledging these and other difficult truths by truly engaging with the data and making real meaning with it will require a new collective emotional maturity. It might be necessary to work with people like Brene Brown (http://www.brenebrown.com), author of *Daring Greatly*, to consider how to do this.

Question 1 asks: (1) What are the public policy implications of the collection, storage, analysis, and use of big data?

In the request for information you outline that information about people’s purchases, conversations, social networks, movements and physical identities is being collected, stored, analyzed and used.

Lets start out by asking **who** is collecting it and **how**?

Right now the applications, devices, software, hardware and platforms we use; along with the stores we buy from, the credit card and banks we use, sites we read on, music and radio we stream all follow us and track our behaviors. This tracking is not just from our digital interactions but our physical movements everywhere we go.

All of these types of entities actively seek persistent correlateable identifiers from us and then package up our data and sell it on data markets. Only a few pieces of PII are needed to correlate data from different sources together and weave a comprehensive digital dossier.

The proposed Commercial Privacy Bill of Rights Act of 2011

Defines "Personally identifiable information," as (1) the first and last name of an individual; (2) postal (residential) address; (3) email address; (4) telephone number or mobile number; (5) social security number; (6) credit card number; (7) "unique identifier information that alone can be used to identify a specific individual"; or (8) "biometric data," including fingerprints and retina scans.

The definition of PII also covers any of the following if stored or used along with (1) through (8) above: (1) date of birth; (2) birth certificate number; (3) place of birth; (4) unique identifier information "that alone cannot be used to identify a specific individual;" (5) "precise geographic location," excluding general geographic information that can be derived from an IP address; (6) information about an individual's use of "voice services, regardless of technology used;" and (7) a catch-all.

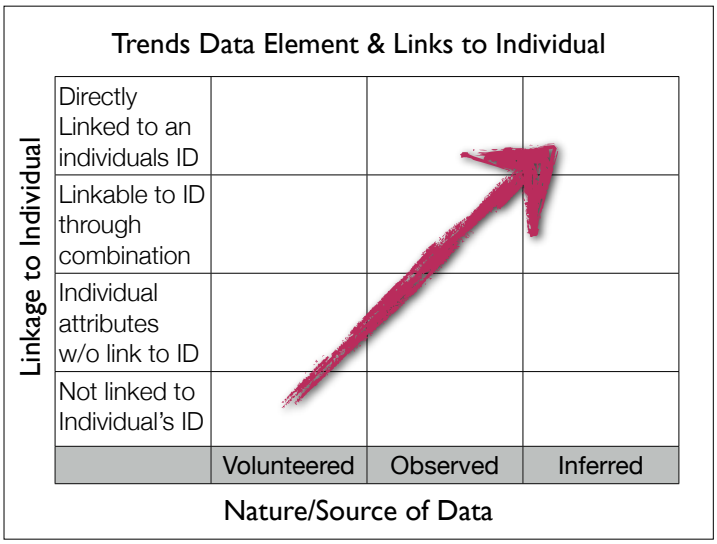
The Bill also contains a third category of information which it calls "sensitive" PII, which includes medical/ health information and the "religious affiliation" of an individual.

http://blog.ericgoldman.org/archives/2011/04/a_look_at_the_c.htm

The data we actively volunteer (knowingly give or post) is put together with data passively generated and collected (geolocation logs from our phones or GPS devices) and can be used to infer much about us including quite sensitive information, for example, our religious beliefs (because we go to particular a house of worship once a week).

Each of us have extensive digital dossiers about us held by business entities we have no business relationship with or knowledge that they hold information about us. **Why** is this being done? This information is bought and sold and used by commercial entities to target us, to make decisions about whether to do business with us or not and on what terms.

The Re-Thinking Personal Data project (http://www.weforum.org/issues/rethinking-personal-data) at the World Economic Forum came up with the following diagram to articulate the situation. Things get creepier as one moves up and to the right.



We - the people - who the generate the data are not empowered with the tools to collect our own data.

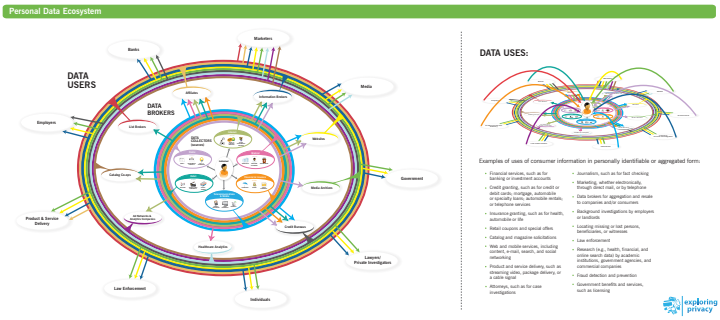
Since 2010 I have been catalyzing the global community of entrepreneurs

seeking to build those tools and develop a Personal Data Ecosystem (http://www.pde.cc)- at its core are services where individuals collect and manage their own data. Just like we use banks to collect and manage our money, why not have data banks, data vaults, data stores or what is now being called Personal Clouds to store the data we voluntarily and implicitly generate.

Question 1 asks: Do the current U.S. policy framework and privacy proposals for protecting consumer privacy and government use of data adequately address issues raised by big data analytics?

No they do not. This question is broad and the answer is no from several different perspectives.

Daniel Solove wrote in his 2004 book, *The Digital Person* about the nature of the data broker industry. In the 10 years since it has continued to grow - unabated. A map of what happens to our data was submitted to the Federal Trade Commission in 2009 called Exploring Privacy that maps out what happens to our data after we interact with retailers, websites, cable companies, medical offices - basically any aspect of our day-to-day lives were a record of a transaction is entered into a computer or an electronic device is involved. More and more data is being collected with basically no protections on how it can be aggregated to profile us as particular individuals.



http://www.ftc.gov/sites/default/files/documents/public_events/exploring-privacy-roundtable-series/personaldataecosystem.pdf

The current practices on the web involve the placement of various forms of device identification, browser configuration, tracking beacons and cookies. These all in effect enable private companies to stalk us as we move around the web. These practices currently happening to us as we move around the digital world if they were replicated by people following us or by companies placing little physical tracking devices on us on us as we moved about the physical world would be prohibited.

One of the big issues raised by big data analytics is the information asymmetry that exists between individuals and communities they are a part of and the corporate and government entities that are collecting, storing, analyzing and using data.

People are not cognizant of how much information they voluntarily give away and how much meta-data their every day interactions on digital devices and shopping in a retail store generate. One simple way the government could change the current policy framework to balance the information asymmetry issue is to mandate that individual consumer/citizens have a right to:

- a copy of all data (that a company is collecting) while they are using digital devices (whether they own them or not)
- a right to a digital copy of all records of transactions they do (right now one gets a paper receipt at a grocery store and they have a digital record of the transaction).

The government has worked on initiatives around various buttons - blue for medical records, green for utility company records, red for educational records and the latest I have heard about is gold for financial records. We should have buttons for "everything" - the data we generate is as much ours as it is the firms who are providing us services. Because these are by their very nature digital objects - that is bits, bits of data. The frame of property ownership is not an appropriate one. We must begin to anchor the conversation in rights and responsibilities. What is my right to the data I generate. What are the responsibilities of the company or services provider I am using (and in the process generating data). We both should have rights to the data and responsibilities. We could use some of the deliberative democracy methods I mentioned earlier to consider what those should be and then move to establish them in law.

We must ask ourselves are we free people, with control over our own data just as we have control over our physical selves. Or are we going to be increasingly enclosed on digital plantations or surfs on digital feudal estates. Some have contrasted this emerging digital feudalism with the potential for a digital enlightenment. (See this slide show by Marc Davis (http://www.slideshare.net/marcedavis/20130509-marc-davis-on-metaphors-and-models-of-personal-data-implications-for-policy-and-technology-at-iiw16)

The right to our own data will kick start the market for personal clouds and new ethical data markets.

Question 3: What technological trends or key technologies will affect the collection, storage, analysis and use of big data?

Personal Clouds are emerging as a new technology that is designed to put people at the center of their own data lives. Some call this "small data". It is empowering the individual to collect their own "Big Data" and use tools and services to use analytical data methods on one's self.

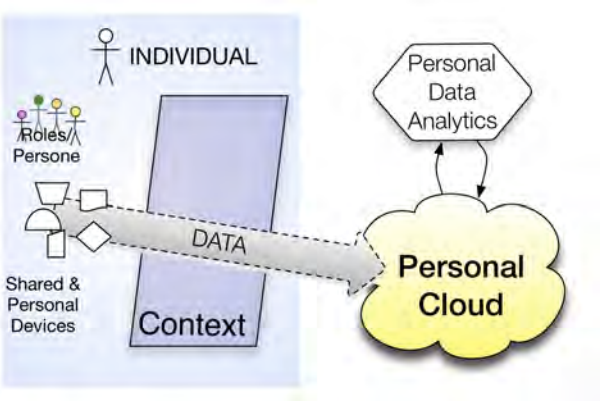
It means that individuals can then see the image of themselves that they are releasing via the activities they do that generate data. The Personal Data Ecosystem that I lead wrote to both the Federal Trade Commission and Federal Communications commission in response to their respective white and green papers on privacy.

* Link to FTC response: https://dl.dropbox.com/u/13895906/FTC-DNTResponse021811.pdf

* Link to FCC response: http://www.ntia.doc.gov/comments/101214614-0614-01/comment.cfm?e=01D6196B-4C69-4C2B-8DA5-6D78D64AF527

Since then we have further articulated the different market models for data that align with consumer/citizen interest.

Meaningful, understandable control over the collection and use of personal data by individuals enables companies, organizations, and governments to create value while increasing public trust. Three broad categories emerged: *Vendor Relationship Management, Infomediary Markets,* and *Data Aggregation Markets*. These models place the user at their center, and reflect ways that context will influence the roles a person plays or how they behave.



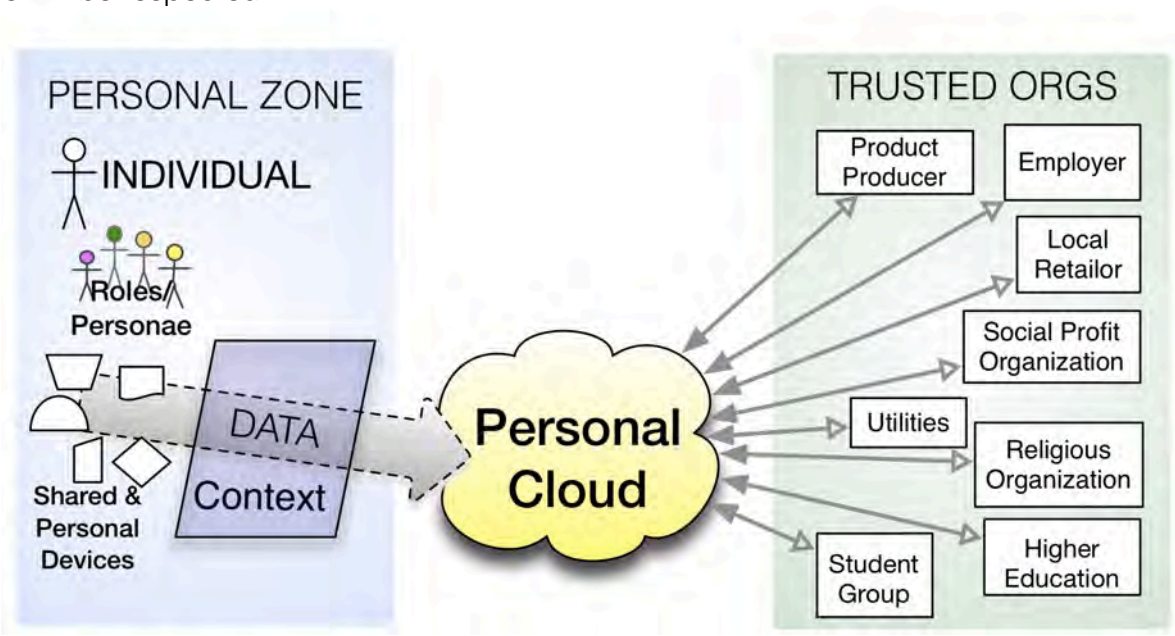
At the heart of each model is a *personal cloud* that helps people to:

- Securely collect and store data they generate in daily life
- Connect to service providers so they get value from their data
- Limit how the data they provide is used by service providers

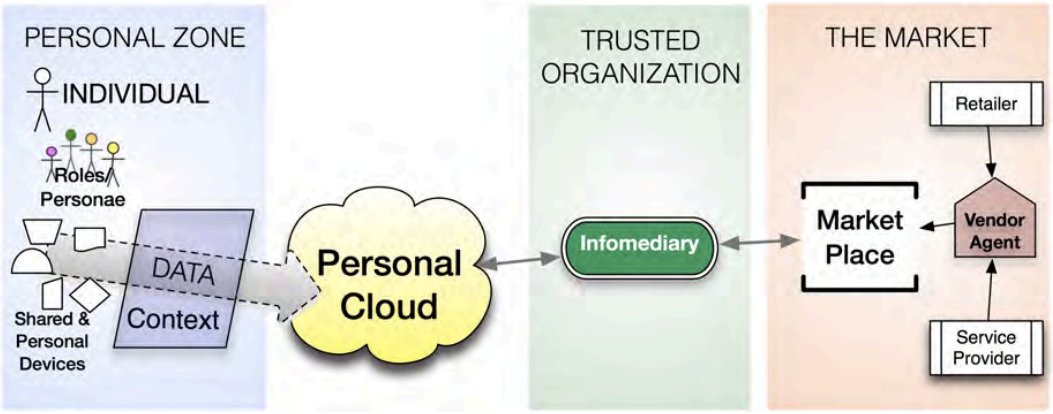
Vendor Relationship Management

CRM tools are standard in today's businesses where they are used by companies to manage their relationships with customers and potential customers.

Vendor Relationship Management (VRM) tools work on behalf of the individual to connect people to businesses, organizations and public services. Individuals can share more detailed information with companies they like. VRM offers a direct channel people control between themselves and the organizations they engage. Data sharing is governed by terms set by the user. In this model, people have the power to withdraw from relating to the vendor and their choice will be respected.

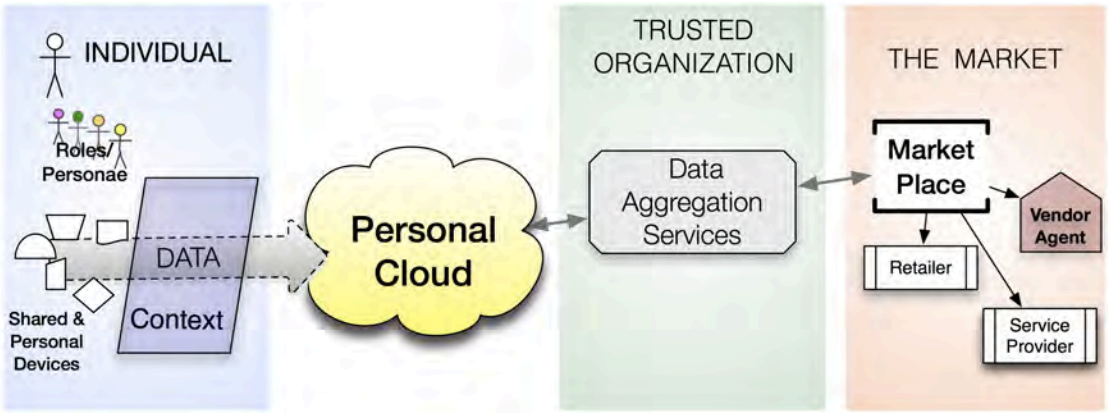


Infomediary Markets



In markets where vendors seek attention from potential customers, Infomediaries act as data brokers on the users' behalf. The Infomediaries create permission-based channels, based on accurate personal data that the user provides, but without revealing personally identifying information in the market.
Companies want to find customers for their products and services, governments want to reach citizens with important relevant information, and organizations want to attract new interested members. Today's marketing methods can be ineffective, annoying and intrusive to get at people's personally identifiable information without their awareness or consent.

Data Aggregation Markets

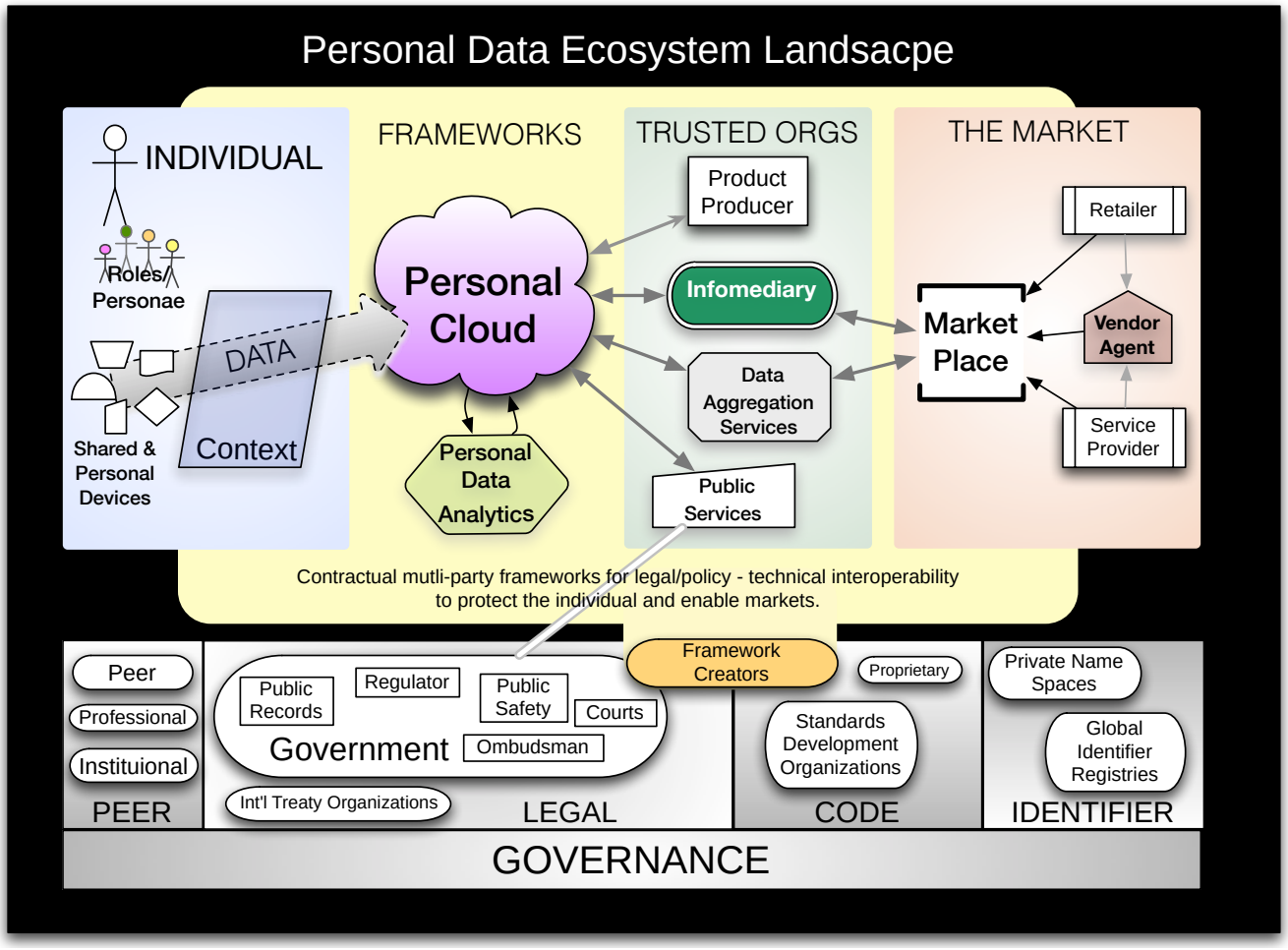


Data Aggregation Services read the details of users' data, but provide companies, governments, and researchers with summary or aggregate values. Corporate research and marketing have many legitimate, ethical uses of "big data." Individuals willingly participate when they believe their privacy will be respected. With that trust, they willingly provide more honest, complete, and detailed personal information to these services. This market model was used for media audience metrics for fifty years by companies like Nielsen and Arbitron.
Big Data is big business and many firms today collect vast troves of information about specific individuals and households, from public records (Axiom), social networks (RapLeaf), or foot traffic via mobile phones (Sense Networks)

Question 3: Are there particularly promising technologies or new practices for safeguarding privacy while enabling effective uses of big data?

A critical aspect of how the system above works is accountability frameworks. Sometimes called Trust Frameworks - these technology/policy sandwiches enable information to flow. If regular citizens are to trust these systems I continue to believe that we should not call the frameworks by the name of the feeling we hope that they foster in the overall system. We have to be able to talk about them and ask about the trustability or trustworthiness of them. It is one thing to have a conversation about the trustability of an accountability framework. Asking "what is the trustability of a particular trust framework" is just to awkward. This is a link to a post that I wrote covering this issue - The Trouble with Trust and the Case for Accountability Frameworks. http://www.identitywoman.net/the-trouble-with-trust-the-case-for-accountability-frameworks

Here is a diagram of the whole ecosystem together - all three market models outlined above, accountability frameworks (and their straddling of legal and code governance) along with other modalities of governance at play in the system. The next page has a description of each box found on this landscape.



Personal Data Landscape Term Definitions

Individual: A person

Devices: Mobile phones, computers, self tracking devices, medical monitoring devices, e-readers.

Context: Where a person is (home, school, work). The Role they are playing (parent, coach, spouse, employee, supervisor, athletic team member). Persona they are presenting (video game player, professional, goofy hobby identity).

Data: The bits generated explicitly such as photos, tweets, status updates.

Frameworks: Contractual mutli-party frameworks connect legal/policy agreements to technical interoperability to protect the individual and enable markets. The Personal Cloud service provider is at the heart of these frameworks, chosen by the end user, and works on their behalf managing their data and its participation in the framework.

Personal Data Analytics: Services that help people gain insight into their own personal data. An example being one's daily health status or a personal annual report.

[Trusted Organizations]

Product Producer: This is an example of a Vendor Relationship Management connection where a consumer who bought a product from a producer manages an open channel with the maker of the product they bought and willingly share information under favorable terms they the user set.

Infomediary: A service trusted to have insight into a person's data and working on their behalf. They have an individual's personally identifiable information (PII) and protect that data and put it to use.

Data Aggregation Services: Services create aggregate data sets from personal data, like music listening habits. Aggregators may compensate people for their data, people may share altruistically, or people may unknowingly share.

Public Services: Governments delivering services to their constituents can enable use of personal data stores for better access and data quality.

[The Market]

Market Place: This is where an Individual's business agents with PII meet Vendor agents without PII.

Retailers: Companies that sell goods to customers.
Service Providers: Companies that provide services to people.

Vendor Agents: Companies that help retailers and service providers find good potential leads. They do not have personally identifiable information.

[Governance]

How systems are regulated take many forms. Governance starts with laws and regulation but also includes cultural practices, business norms, and, in digital systems, how identifiers are allocated and the code that connects them.

LEGAL:

Government: plays many roles in the systems:

Regulator: Governments set baseline rules for how markets work. They provide the court system where contract law is adjudicated.

Public Records: Governments record births, marriages, divorces, deaths along with licensing, and property title registries.

Public Safety: Policing and law enforcement.
Ombudsman: Many states have a data protection commissioner who protects constituents.

International Treaty Organizations: Support the coordination of int'l treaties and provide a meta-int'l law that hold governments accountable to each other.

CODE: Computer code and how it runs determines what is possible in computer systems. The phrase "Code is Law" was popularized by Lawrence Lessig.

Standards Development Organizations: Bruce Sterling said "If code is law then standards are like the Senate." Standards bodies agree on how code works regardless of the particular language it is written in or system it is running on. For example, the W3C standardizes the HTML specification for presenting web pages.

IDENTIFIER: Networks run on identifiers for each endpoint. How these are allocated, and the terms and conditions of use in a network, govern the network.

Global Identifier Registries: Examples include the phone system, Domain Names, ISBN numbers, RFID.

Private Name Spaces: examples include Twitter, Skype, Google, Facebook etc.

PEER: This kind of governance is the most powerful in many ways and helps social systems operate.

Peer-to-Peer: People have opinions about each other and also about businesses and services they interact with - like Yelp for small businesses.

Professional: Doctors, lawyers, engineers, and architects are professions that peer regulate.

Institutional: Institutions figure out what other peer institutions are - such as banks worldwide in SWIFT.

Framework Creators: Organizations that create contractual legal-policy/technology frameworks that govern complex multi-party networks.

(4) How should the policy frameworks or regulations for handling big data differ between the government and the private sector? Please be specific as to the type of entity and type of use (e.g., law enforcement, government services, commercial, academic research, etc.).

Most of the answers that you will get to this question will have 'the answer' from their own point of view. There is not one right answer to these questions. We live in a complex society and complex trade offs must be made. I believe we must tap our collective wisdom and intelligence to figure out where to draw the lines in the grey sand.

We must leverage the deliberative democracy tools I mention above and do so following the Principles of public engagement outlined by the National Coalition for Dialogue and Deliberation (http://www.ncdd.org).

An example that could be followed is how the Canadian Province of British Columbia recently engaged a randomly selected panel of 36 citizens over 2 full weekends to figure out how key policy questions in the implementation of their eID system. I wrote about this recently in the Re:ID magazine (http://www.identitywoman.net/wp-content/uploads/2011/09/reid_spring_14-BC.pdf)

I would like to close with some thoughts about Question 2: What types of uses of big data could measurably improve outcomes or productivity with further government action, funding, or research?

We must look at how data can serve communities as a whole and how they can impact the lives of various groups.

A recent conference hosted by UC Berkeley CITRIS focused on Open Data and Democracy I saw two different projects with a primary focus on communities using Big Data and understanding deeper social issues in a meaningful way. The Center for Community Informatics in Vancouver British Columbia - http://communityinformatics.net/ and the Justice Mapping Center - http://www.justicemapping.org/ and their Justice Atlas - http://www.justiceatlas.org/

Public Engagement Principles by the NCDD

These seven principles reflect the common beliefs and understandings of those working in the fields of public engagement, conflict resolution, and collaboration.In practice, people apply these and additional principles in many different ways.

1. Careful Planning and Preparation
Through adequate and inclusive planning, ensure that the design, organization, and convening of the process serve both a clearly defined purpose and the needs of the participants.

2. Inclusion and Demographic Diversity
Equitably incorporate diverse people, voices, ideas, and information to lay the groundwork for quality outcomes and democratic legitimacy.

3. Collaboration and Shared Purpose
Support and encourage participants, government and community institutions, and others to work together to advance the common good.

4. Openness and Learning
Help all involved listen to each other, explore new ideas unconstrained by predetermined outcomes, learn and apply information in ways that generate new options, and rigorously evaluate public engagement activities for effectiveness.

5. Transparency and Trust
Be clear and open about the process, and provide a public record of the organizers, sponsors, outcomes, and range of views and ideas expressed.

6. Impact and Action
Ensure each participatory effort has real potential to make a difference, and that participants are aware of that potential.

7. Sustained Engagement and Participatory Culture
Promote a culture of participation with programs and institutions that support ongoing quality public engagement.

In contrast at an Open Oakland Camp hosted put on by the Code for America Brigade at Oakland City Hall they had a session focused on a use of big data/ open data for crime spotting. Oakland is a majority minority city with a history of policy brutality and misconduct towards African American men. Except for my partner who is African American the whole audience was white. He wasn't really into helping develop applications who's primary purpose seemed to be to put more brown and black men in prison.

To close on this same theme I must say that issues around Big Data are gendered.

It is mostly men who build, design, big data systems and the algorithms that are used to glean insight from the data. They are the people who choose to fund data startups. They are the ones who lead the large and small companies engaged in data analytics. In my years working in Silicon Valley it has become very clear that the perspective of men and how they move through the world - the assumptions they make about what is and is not an acceptable in the design of systems. Their is also the delicate issue of the preponderance of Aspergers Syndrome / Autism Spectrum Disorder by those working in the technology industry and therefore building these systems. Although all those on the spectrum are unique it is not un-common for various reasons to experience some social blindness - that is a difficulty perceiving and aligning with social norms/boundaries. I have known male programmer friends who are on the spectrum to hoard data, all kinds of it just because it might be useful one day and to not really consider the social implications of doing so. This is an additional reason that broad social engagement is required to figure out where clear lines should be drawn.

I have lead a women's technology unconference since 2007. When women who understand technology are with each other they discuss critical issues that I almost never hear raised in any other forum. It is critical that you seek out and include the voices of women and women technologists when figuring out the answers to the policy questions in the real of big data.

Thank you for soliciting public comment on these critical questions.

If you should like to contact me for further information on any of the ideas I have presented here please don't hesitate to do so.

Kaliya "Identity Woman" Hamlin
[redacted] is the best e-mail address for me.
I can be reached at [redacted] (although I rarely check voice-mail)

Thank you for considering my contribution.

I will be applying to be considered in this third round of Presidential Innovation Fellows.

Regards,

- Kaliya

[redacted]

From: Dr Tyrone W A Grandison <tgrandison@proficiencylabs.com>
Sent: Sunday, April 06, 2014 5:56 PM
To: bigdata@ostp.gov
Subject: Public Comments for RFI

In response to the White House Office of Science and Technology policy (OSTP) group's Request for Information (RFI), I want to say the following as an American citizen (and not as a computer scientist with over fifteen years of experience in the fields of Computer Security, Privacy and Trust Management).

There is no doubt that Big Data collection and processing may have a far-reaching negative impact on the public. As with all technology, when placed in the wrong hands (or used for the wrong purpose or with the wrong intention), the harm or damage to the governement's ecosystem of trust and to the individuals who are wronged will be tangible, real and create (amongst other things) an image/brand problem for the Federal government that may be extremely difficult to recover from.

It is with this knowledge that I request that privacy be the primary focus of the Federal government in this arena, i.e. Big Data. I am not referring to anonymization, de-identification or other controls one may want to put into place to increase accountability . That is a distracting, non-productive and doomed-to-fail approach. For more on this, please read http://www.tyronegrandison.org/1/post/2014/04/lets-have-a-honest-discussion-about-de-identification.html. I am talking about inventing new ways of performing Big Data processing such that the processes are privacy-preserving. I am referring to building Big Data systems that have privacy as a fundamental construct, such that the underlying platform is naturally (or natively) privacy-preserving (and secure).

There is no doubt that Big Data holds great promise. However, to ignore the Great Dangers that are involved would be setting up future generations (and the current administration) to fail.

Regards,
Dr Tyrone Grandison

--
Dr Tyrone W A Grandison
BCS Fellow, RSA Fellow, ACM Distinguished Engineer,
IEEE Senior Member, IBM Master Inventor
[redacted]
Business: http://www.proficiencylabs.com
[redacted]

Open Technology Institute, New America Foundation

April 4, 2014
Office of Science and Technology Policy
Eisenhower Executive Office Building
1650 Pennsylvania Ave. NW
Washington, DC 20502

Re: Big Data RFI—OTI comments on the White House's Big Data Initiative

In response to the Office of Science and Technology Policy's Request for Information on the implications of "big data," the Open Technology Institute (OTI) at New America submits the following short essay, "The Biggest Data of All." This essay was first delivered as a talk at New York University by OTI Policy Director Kevin S. Bankston as part of the workshop on "Social, Cultural & Ethical Dimensions of 'Big Data,'" the second of three workshops co-hosted by the Office of Science and Technology Policy as a part of the White House's big data initiative.

The Biggest Data of All

When considering the future of big data and privacy, we must consider the biggest data of all, the data set that encompasses almost all of the others: the data that transits the Internet.

As our offline activities and records move online—our shopping, our consumption of news and entertainment, our financial and legal and medical records and transactions, and an ever-increasing number of personal and business communications of every kind, even the most sensitive—the depth and breadth of this massive data set continues to expand. As all roads once led to Rome, today, nearly all data streams eventually flow into and through the great river of data that is the Internet.

Therefore, when considering the ethics of big data and privacy, it is necessary to look to the ISPs and governments—including our own—that have access to that river of data, often subject to unclear or insufficient legal restrictions.

What are the duties that these stewards of cyberspace owe to us? Would it be ethical, for example, for an ISP to secretly monitor and record everything you are reading and searching for online, so it could better serve you targeted ads or sell that information to data brokers? What if it gave you notice? Or even a discount on your service in exchange for permission?

And what of the government? Would it be right for our government to secretly install massive automated surveillance stations on top of major Internet exchange points inside the U.S., vacuuming up all of the data under questionable legal authorities, and filtering for suspicious identifiers and patterns? Would it be right for our government to secretly tap into the fiber lines that link the data centers of U.S. companies whose services are used by masses of innocents both here and abroad?

These questions are not hypothetical. And the ultimate report of the big data working group will be incomplete if it does not at least attempt to address some of these questions. There are three answers to these questions, in particular, that I hope we'll eventually see.

The first answer is transparency. Whether talking about ISPs or governments, tapping of the Internet backbone shouldn't occur without the knowledge and consent of us, the customers and the governed. That is why I hope that the Administration will soon finally admit to a fact that's been on the front pages of every major newspaper since December 2005, that's been evidenced by whistleblower documents leaked from inside of AT&T and the government, that's implicit but obvious in FISA court opinions and procedures that are now declassified, and that's even been admitted to by Senate Intelligence Committee Chair Dianne Feinstein.

For us to have a meaningful public debate in the public square, Congress, or the courts, the Administration must declassify the open secret that everyone already knows: the NSA is tapping the Internet backbone inside of the United States.

The second answer is encryption. Much of the big data discussion has focused on the risks to data in storage, and on the anonymization and

encryption tech that might protect that data. But we also must focus on encryption for data in transit—be it encryption that protects data sent between me and you, between me and your web site, between our email provider's servers, or between Google and Yahoo's data centers. When it comes to protecting our digital privacy, code is law, and encryption is one of the strongest laws on the books so long as we use it. Therefore, I urge this working group to conclude, as the President's NSA Review Group did, that the U.S. government should strongly support rather than undermine the widespread use of encryption technology—not only for data at rest but also in transit.

A third and final answer to the problem of privacy when it comes to the biggest data of all is the much-needed re-evaluation of the distinction between communications content and non-content (or addressing or "meta") data about those communications, where content has long been considered the most sensitive, such that non-content has been accorded little legal protection. As Danny Weitzner, the convener of the first in this series of workshops, testified to the Privacy & Civil Liberties Oversight Board: "Metadata at scale is at least as revealing as content." Particularly in the Internet context, metadata can provide an intensely revealing portrait of one's private life, including revealing facts or patterns of behavior that would never be revealed in the content of a communication and in many cases would not even be recognized by the person doing the communicating.

Countless legal and technical experts, including Justice Sotomayor, the oversight board, and the review group have called into question the continued validity of this distinction between content and metadata, and the review group specifically recommended that the government commission a study interrogating that distinction.

I hope that this working group will be the first step in such a study. I hope that this working group will highlight the importance of an encrypted Internet to the future of privacy and security. And I hope that this working group will be a force for greater transparency around how our data is collected and used, whether by our ISPs or by our National Security Agency.

Thank you,

Kevin S. Bankston
Policy Director, Open Technology Institute
New America
1899 L St NW, Suite 400
Washington, DC 20036