Remarks as Delivered by Counselor John Podesta
The White House OSTP, UC Berkeley School of Information,
Berkeley Center for Law and Technology Workshop:
"Big Data: Values and Governance"
April 1, 2014

Good afternoon. I want to begin by thanking Deirdre for that kind introduction, and for all of her hard work putting together today's workshop, and for her work and friendship over the years. For decades, Berkeley has been at the forefront of public debates about freedom of speech, freedom of association, and civil rights. So I think it's very appropriate that our third workshop on big data and privacy is happening here. The discussions we've heard today built on and deepened some themes that have come up throughout this review of big data and privacy, particularly at the previous workshops at MIT and NYU. So I also want to thank all of today's panelists for their thoughtfulness and their expertise. And of course I must thank you all for joining us here today.

The technologies and applications we've heard about today and in our previous workshops may be new, but the legal and policy questions these technologies raise are quite old. Big data requires us to ask ourselves: how do we embrace new technologies and the progress they bring to our society while at the same time protecting our fundamental freedoms and values like privacy, fairness and self-determination?

It's been referenced that I was one of the staff authors of the Electronic Communications Privacy Act. I was working for then a young Senator, Pat Leahy, in 1984. And it might be worth just spending a second talking about how we began the journey to write the Electronic Communications Privacy Act. We actually sent a letter to the Attorney General asking whether the existing wiretap statute, which was passed in 1968—most of you who are lawyers know it as Title 3 of the Omnibus Crime Control and Safe Streets Act—applied to electronic communications. It was a simple, straightforward question: did the wiretap act apply to email and other forms of electronic communications? Months later, we heard back from his deputy, who told us that the answer to our question "was neither clear nor obvious." So that reply didn't inspire confidence. So for the next two years, Senator Leahy, along with his House colleagues, Don Edwards, from this area, and Bob Kastenmeier, worked to update the law for modern times. The result, as was noted, was the Electronic Communications Privacy Act, which—with some updates—and I will note with a few more needed—still governs Internet communications today.

It wasn't that we were just moronic back in 1986. We were faced with a series of precedents about how third parties handled data that were dodgy in terms of trying to provide protection for electronic communication. So this was indeed a negotiated compromise, but I think this really was perhaps a unique law in that we took a problem before anyone knew its full dimensions and really dealt with it in a forward-looking manner. And obviously today we are engaged in the Administration to try and think about how we correct inefficiencies, particularly the one mentioned here today about the difference between 180 days and 181 days of storage, and we're engaging again with my old boss on this topic.

A few years after that, I hosted the Washington, D.C., organizing meeting for a new group dedicated to a new civil liberties cause. I knew Lee would be here, so I somehow dug out my membership card—member number 33 of the Electronic Frontier Foundation. Lee Tien and the

EFF are still fighting to find balance between civil rights, personal freedom, and technological advancement, and, of course, I've become "the man."

But we all know rapid change is the one constant in technology. And in the information age, those changes come at an incredible rate. In 1998, when Google was getting its start, less than five percent of the world was online. Today, three-quarters of people in developed countries and four in 10 people worldwide have Internet access.

As Americans, our willingness to innovate, to experiment, to build new things and try new ideas has built the largest and strongest economy in the world. But from a centralized postal service all the way through to SnapChat, each technological advance has prompted serious debate about its implications for other values that we hold dear, like equality, privacy, and personal freedom.

We're about two-thirds of the way through our study of big data and privacy. Since President Obama asked me to lead this review back in January, working with the Secretary of Commerce, the Secretary of Energy, the NEC and OSTP, we have met with privacy advocates and technology companies; with academic researchers and advertisers; with law enforcement agencies and with civil rights groups. We are receiving public input through a request for information – that comment process was due to close yesterday, but we are extending it to accept submissions through the end of this week. We also have asked for feedback from the public through a survey on the White House website—you can find that at whitehouse.gov/bigdata, and I encourage you to fill it out and pass it on to those in your networks.

This 90-day review is fundamentally a scoping exercise, not a policy development process. After today, I am making one exception. Listening to Ken Bamberger, we are changing our policy to ensure that Google Glasses being used in the White House Gym haven't been hacked. Really, we hope to raise the relevant questions we must ask now and in the future. We have been casting a wide net as we ask how big data technologies have begun to transform government, commerce, and society, and as we examine the potential implications for our social and civic values.

The President's Council of Advisors for Science and Technology is working in parallel on a report that takes a deep dive into the technologies themselves, from machine learning and algorithms to sensor technology and new models for data storage.

Both our study and PCAST's work also aim to look ahead, as much as that's possible, and to anticipate what today's trends might mean for our lives, our economy, and our world in decades to come.

Without getting too far over my skis about what will be in a report that we haven't written yet, I want to share some observations about what we've learned during this process.

But first, let me say a few words about the ongoing effort inside the Administration to reform how we collect information to keep our country safe. I should be clear, specific intelligence programs are not a part of this study on big data, but they *are* central to the debate in which we find ourselves. Last week, the President announced a plan to significantly change the way the Section 215 telephone metadata program works.

As you know, this program has been the subject of a great deal of controversy in recent months, as we as a nation debated our signals intelligence practices. Under the President's proposal, the government will not collect or hold metadata in bulk; instead, those records will remain with telephone companies, and will be available to intelligence agencies only under court order. We're working with Congress to put this plan into law, and are continuing a wide range of efforts to improve transparency and oversight of our intelligence programs.

The President asked us to conduct this study because he recognized that, even as the Administration takes steps to review and reform intelligence practices, the intelligence community is only part of the picture when it comes to big data. These technologies touch every sector of the economy, from government to industry to research, and will only continue to grow in importance.

I want to say up front that we recognize big data technologies can create considerable economic value and are spurring tremendous innovation. From the very beginning, the Obama Administration has worked to make more public data available through our Open Data Initiative. Entrepreneurs and developers have used these data to create countless applications and build successful businesses.

We've also known for a long time that the information age requires us to carefully consider how we protect personal data. That's why in 2012, the President released a consumer privacy blueprint and the Consumer Privacy Bill of Rights, which sets out principles that aim to protect user privacy without restricting the innovative potential of the Internet.

So in the midst of what some are calling a "big data revolution," we are taking this opportunity to consider the landscape and to really interrogate whether our existing policies are prepared for what's on the horizon technologically.

Throughout this process, we've asked three questions: First, what is new and different about big data technology and techniques, compared to traditional data analysis? Second, what questions should we be asking about big data's relationship to privacy and other rights and values? That's the conversation we've been having here today. And, finally, do our existing policies and in particular the Consumer Privacy Bill of Rights provide an adequate policy framework for the era of big data, or are there gaps where new policy or new research needs to be developed?

Of course, we first had to decide on a definition of big data—something that easily could have taken up the entire 90 days of our review. So for the purposes of this study, we've been thinking of big data as data sets that are so large in *volume*, so diverse in *variety*, and changing with such *velocity* that traditional modes of data analysis are insufficient.

I think there are a few technological trends that bear being drawn out.

First, the declining cost of collection, storage, and processing of data, combined with new sources of data like sensors, cameras, and other observational technologies, particularly geospatial technologies, means that we live in a world of near-ubiquitous data collection. And as more companies produce web-enabled appliances, wearable technology, and advanced sensors to monitor everything from health indicators to energy use to my running speed, the "Internet of Things" will add huge amounts of new data to what is commonly collected today. Data analysis is also being conducted at a speed that is increasingly approaching real-time, with a growing potential to have an immediate effect on our environment or decisions being made about our lives.

Second, one of the most powerful things about big data analysis is what is sometimes called data fusion. By combining multiple sources and types of data, big data can lead to some remarkable insights. But it can also lead to the so-called “mosaic effect,” whereby personal information can be derived or inferred from data sets that do not include personal identifiers within.

Third, the information revealed by big data analysis isn’t necessarily perfect. Identifying a pattern doesn’t establish whether that pattern is significant. Correlation still doesn’t necessarily equal causation. So as these technologies become more prevalent, we will need to be deliberate in developing strong, scientific standards—and strong ethical standards—for judging the results of big data analysis and predictive algorithms, as was raised on the first panel today.

Finally, there are some promising technological means to better protect privacy in a big data world. Encrypting data, perturbing data so it no longer represents real individuals, or giving users more say over how their data are used through personal profiles or controls are among the technological solutions. But none of them are perfect, and technology alone cannot protect privacy absent strong social norms and a responsive policy and legal framework.

I want to share a few examples of what all of that means in the real world.

The United States spends more on health care per capita, with worse outcomes, than any other developed country. As a consequence, health care costs are a major driver of our federal deficit. It’s urgently important that we find ways to bring down costs while also improving health outcomes. Big data is helping show us how.

One study synthesized millions of data samples from monitors in a neonatal intensive care unit to try and pinpoint which babies were likely to contract potentially fatal infections. By analyzing all of the data—not just what doctors noted on their rounds—the project was able to identify factors, like increases in temperature and heart rate, that serve as early warning signs that an infection may be taking root. These early signs of infection are not something even an experienced and attentive doctor would catch through traditional practices.

In another instance, by crunching huge amounts of data from electronic medical records, researchers determined characteristics that increase or decrease the likelihood that a patient will be readmitted to the hospital after receiving treatment. This makes it possible for doctors to proactively schedule follow-up care for the patients who are most likely to need it, and reduce costly readmissions in the process.

These are just two examples of how big data techniques are leading to astonishing advances in patient care. But data mining can also pose privacy concerns, especially in a health care context. Our medical history is among our most sensitive personal information. Latanya Sweeney, a Harvard researcher who is currently the Chief Technology Officer at the FTC, has repeatedly demonstrated that it is possible to use publicly available data to identify individual patients in supposedly de-identified medical data sets.

In other contexts, however, there are big data sets that pose no, or very remote, risks at all to personal privacy. One example is public data on climate, weather, and environmental factors. Collected by huge networks of sensors and a fleet of NOAA and NASA satellites, these data have fueled our understanding of how our climate is changing. These data get plugged into the complex computer models that predict sea-level rise, ocean acidification, storm surge risk, and other climate impacts.

The economic and political import of these data is extraordinary. But by way of example, without NOAA data, we wouldn't have the Weather Channel. Without GPS signals made available by Department of Defense satellites, our smartphones wouldn't be able to tell us how to get from here to the Golden Gate Bridge.

At the same time, even civic-minded big data applications can have troubling implications not just for personal privacy, but for equality and nondiscrimination as well— an issue we are thinking a lot about in the course of our study. For instance, in Boston, the city released an app called Street Bump that used smartphone sensors to detect potholes and report them to the department of public works.

But what happened after Street Bump was first rolled out in Boston should give us pause. Because poor people and the elderly were less likely to carry smartphones, let alone download the app, the app wound up systematically directing city services to wealthier neighborhoods. Now, to its credit, the city of Boston figured this out and tweaked the app to account for underreporting, so that everyone would have equal access to city services. But the lesson here is that we need to pay careful attention to what unexpected outcomes the use of big data might lead to, and how to remedy any unintended discrimination or inequality that may result.

As I said earlier, this review is a scoping exercise, aimed at identifying trends and defining the questions that will inform future research and policy development. But a few big issues are coming into focus, and were reinforced throughout the day, and I'd like to touch on some of those before I conclude.

First, the notice-and-consent framework that has governed data collection for decades is coming under stress in the big data context. In public spaces, people are often unaware of the degree to which sensors, cameras, and other data collection tools are recording information about them. They may not fully understand how that information gets used and shared.

And if people aren't reading terms of use agreements on their computers before they click "OK," it seems unlikely that they will parse pages of legalese that may accompany their smart watch, Nike Fuel band (it's time we let FitBit off the hook) or other wearable device.

So there is an active and ongoing debate about how we can shore up the notice and consent framework that has been the bedrock of consumer privacy for four decades, or whether we need to create other ways to ensure people's personal information, or information collected about them, is being responsibly used in a big data world.

Next, it's also clear that what constitutes "responsible use" will be different in different contexts. Companies and researchers working with sensitive health or financial information have different responsibilities than marketers looking to connect you with online advertisements that match what they predict you might want to buy.

In particular, as the President said in January, the United States government must hold itself to a higher standard when it comes to using data responsibly. He was speaking of how we handle data in the intelligence context, but protecting sensitive data is a challenge that many federal agencies have taken seriously for decades, including the Census Bureau, the IRS, and HHS.

Defining responsible use standards for federal agencies is essential for public trust, and will need to be balanced against the economic benefits of opening more government data.

With big data, the government can lead by example in other ways as well—for instance, by making it easier for people to access data that the government holds about them. The Department of Veterans Affairs Blue Button tool helps veterans access their medical records, manage their health care, and correct any errors in their files, all in one easily accessible place. The My Data Initiative at the Department of Education seeks to give every student access to his or her academic data so they can make more informed decisions about college choices and study needs. Efforts like these encourage greater transparency in how data are collected and used, and give people more control over their personal information.

The flow of big data is, of course, global in scale, and recognizing that, the President charged us "to consider how we can forge international norms on how to manage this data." So to inform the process, we've met with a range of foreign counterparts, including privacy regulators from across Europe and the Americas. That's why it's important that Rainer Stentzel made the trek to be here with us today. There is tremendous interest in how big data challenges existing regulatory and ethical paradigms for privacy. There are no simple answers but it's clear that there is a whole lot riding on getting the answers at least in the right zone, including the future of a single open, interoperable, secure and reliable global Internet, and the free flow of information. Big data raises real policy questions for governments around the world. But I would note that even the European regulators we met with noted that the conversation here in the U.S. is as deep and more technologically informed than anywhere.

It's clear from the conversations amongst all the panels today, the policy implications are serious. One of the issues that has emerged powerfully today is the need to have a serious conversation about civil rights and discrimination in the big data context. Big data analysis of information voluntarily shared on social networks has showed how easy it can be to infer information about race, ethnicity, religion, gender, age, and sexual orientation, among other personal details. We have a strong legal framework in this country forbidding discrimination based on these criteria in a variety of contexts. But it's easy to imagine how big data technology, if used to cross legal lines we have been careful to set, could end up reinforcing existing inequities in housing, credit, employment, health and education. This is a place where Cynthia Dwork's question on the last panel about how to think about the relationship between utility and fairness comes into highest relief.

All this leads to a few important, technical conclusions about the world of big data in which we increasingly live. First, that the costs and physical footprint of recording data about this world is shrinking to zero, so our expectations of actions being recorded will only increase.

Second, the streams of information that might fall into the category of "surveillance" are rapidly increasing in number and diversity — and being put to some very interesting uses. Behind all the powerful functionality in, for instance, our smartphones or cars, are sensors for temperature, audio, acceleration, light, and heat that are more sensitive than anything available to consumers three decades ago.

Third, as the opportunities for big data analytics increase, so will the demand for high-performance computing capable of supporting it. The architecture for high performance computing to do computational simulation of complex systems is merging with big data analytic architectures. That will have big implications from energy to manufacturing to research and development.

And finally, it is clear that the new paradigm in big data will be defined by the move from 'search' to 'prediction,' or from a predicated search to a non-predicated search, raising a whole number of practical and ethical questions for consumers, companies, and particularly for governments.

For those of you looking for what we think the most important social and ethical question these trends raise: stay tuned. The report will come out in a matter of weeks.

So with those insights in mind, let me end on this note. Melvin Kranzberg's six laws of technology begin by stating: Technology is neither good nor bad. Nor is it neutral. Like Kranzberg, I believe technological progress demands that we engage with the social, ethical, legal and political questions that arise.

So I want to thank you all again for your time and for your thoughtful engagement during today's workshop and throughout our review of big data and privacy. I'm confident that this is a conversation that will continue inside the Administration, but more importantly in every corner of America and around the world. After today, we'll be returning to Washington to draft and review our report, and we look forward to releasing our findings in the very near term.

Thank you for your attention – I think we have some time for questions.