Big Data Privacy Workshop

Advancing the State of the Art in Technology and Practice

co-hosted by

The White House Office of Science & Technology Policy

&

Massachusetts Institute of Technology

MIT Big Data Initiative at CSAIL

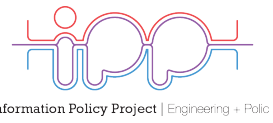
MIT Information Policy Project at CSAIL

Computer Science and Artificial Intelligence Laboratory (CSAIL)

March 3, 2014

WORKSHOP SUMMARY REPORT

Cambridge, Massachusetts



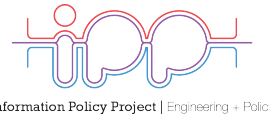
Acknowledgements

We gratefully acknowledge the many contributors to this workshop. This includes all the speakers, panelists, and participants for their thoughtful input. Thank you to all the MIT staff that worked behind the scenes to make the workshop a success. A special thank you to Barbara Mack, Pingry Hill Enterprises, for all the support in drafting and editing this summary report.

This workshop was supported by a generous grant from The Alfred P. Sloan Foundation.

ALFRED P. SLOAN FOUNDATION

Support for the workshop was also provided by the MIT Big Data Initiative at CSAIL, the MIT Information Policy Project at CSAIL, and the Computer Science and Artificial Intelligence Lab (CSAIL).



bigdata@CSAIL
MIT BIG DATA INITIATIVE

Table of Contents

Executive Summary

The Big Data Privacy Workshop, co-hosted by the White House Office of Science & Technology Policy and MIT on March 3, 2014, focused on technology and public policy questions in a big data environment. This MIT workshop was the first in a series of three events held on university campuses in the spring of 2014 as part of the White House administration's 90-day review of big data and privacy.

Today computing is ubiquitous and the widespread ability to collect, integrate, and analyze data, creates important opportunities for increasing knowledge, but also poses new challenges to notions of privacy. Opening remarks at the workshop from L. Rafael Reif, President of MIT, articulated key questions: How do we capitalize on big data's potential for good, while maintaining essential privacy protections? How do we design future technologies, policies, and practices to achieve the right balance for society?

John Podesta, White House Counselor leading the 90-day big data review, cited several prominent initiatives of the Obama Administration centered on open data and making data available to the public, to industry, and the research community. He underscored the importance of maintaining a strong privacy protection regime, citing the principles that serve as the foundation for the Privacy Act of 1974 and the development of the Obama Administration's Consumer Privacy Bill of Rights released in 2012. Penny Pritzker, Secretary of the U.S. Department of Commerce, illustrated how the American economy is grounded in a commitment to the free flow of information and emphasized the role of innovation and the promise of big data in potentially unlocking trillions of dollars in economic value to the global economy. She highlighted the critical role trust plays between consumers, businesses, and governments in achieving these potential gains.

Presentations and discussions at the workshop yielded many technical observations and a set of questions to be addressed.

First, 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 to assure that systems handling personal data are functioning consistent with desired public policies and institutional rules. In some cases, these controls will 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. Developing the science base that enables people to be in control of their data and ensure that it is used accountably is a key challenge. We saw from research presented at the workshop that there are a variety of solutions are available to meet these challenges.

Second, large-scale analytic techniques are creating socially important knowledge from the use of personal data, including advances in health care, genomics research, transportation, and education. Speakers addressed the use of personal data in each of these applications and the tradeoffs involved in terms of protecting privacy. This was made especially clear in the case of medical data and the use of patient medical records, where the benefits include preventing serious illness and saving lives. One example may be seen in the case of Health Care-Associated infections (HAIs), where in the United States 1 in every 20 inpatients develops an infection related to hospital care. Using patient medical records and hospital data, researchers were able to develop a model that successfully predicts patients who are at higher risk of acquiring these infections and as a result, can help prevent the spread of these potentially deadly infections. Another example is in the case of genomics research, where it is necessary to have access to very large data sets on patients, including genomic and clinical data, in order to yield insights on medical disorders and treatments. In some circumstances, researchers need unmediated access to data that includes personal information. In other cases, it may be possible to control access to data in a manner that sufficiently obscures personal identity, while

enabling researchers to extract valuable knowledge. In either case, there are computational techniques that can help to protect individuals from harmful uses of that data.

Throughout the workshop, different big data privacy risks were discussed, including:

- new opportunities for discrimination, including unanticipated or unintentional ways in which data may be used against individuals;
- a lack of transparency and control in the context of multiple data sharing relationships;
- difficulties in obtaining informed consent from users, when there are hundreds of data-collection sources, including many that are machine-generated sources of data;
- re-identification attacks, i.e. cases when anonymized or "de-identified" data sets are made available, but taken in combination with other available data sets it is possible to then re-identify individuals from the anonymized data set;
- indiscriminate collection and use of metadata;
- challenges for organizations operating under different regulatory policies in different countries; and
- increasing opportunities for large-scale data breaches.

The workshop reviewed state of the art in different privacy protecting technologies. Speakers agreed that classical methods of data encryption are no longer sufficient and that we must focus efforts on developing new methods that allow for performing computations on encrypted data, including techniques such as functional encryption and homomorphic encryption. In addition to encryption methods, the speakers discussed the evolution of differential privacy, a technique designed to reveal accurate statistics about a group of respondents, while preserving the privacy of individuals in the group. Progress on developing "accountable systems" was also discussed; this is an approach that entails expressing laws in a computable policy language and provides proof that a

transaction is compliant or not with a given set of policies. Other methods covered included the Secure Function Evaluation (SFE) technique, or Secure Multi-Party Computation, where a function can be run over disparate private data sets and will return an exact output result while keeping the original data sets hidden. All of these technologies are at different stages of development; some are ready for deployment and others require more research to advance towards practical application.

Technology will not replace the need for laws and social norms protecting privacy. We should expect that systems will be built to perform according to privacy rules and will make enforcement of those rules easier; this will be necessary as the scale of data usage increases beyond the point that manual compliance and audits can be effective. However, throughout the workshop a variety of implicit and explicit definitions of privacy were used. Some definitions implied that privacy is synonymous with secrecy and complete confidentiality – 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 sometimes understood as a question of whether personal data is used in a manner that harms the individual. Therefore, in order to realize the goal of designing systems that do a better job of respecting privacy, we must have greater clarity on rules for specific uses of big data applications.

Finally, progress toward enabling beneficial uses of personal information in a manner that protects individual privacy requires answering the following questions:

1. How can institutions (public and private) evolve best practices for handling personal information and taking maximum advantage of privacy enhancing technologies? From the scenario discussion at the end of the workshop, we learned that there are enormous legal and ethical complexities that arise with the use of large-scale analytics on personal data. In some cases, the law or social custom provides guidance, but in many cases it does not.

2. How can government and the private sector work together to test privacy enhancing technologies in practical contexts at large scale?
3. How can government and industry work together to advance the technical state of the art in multidisciplinary privacy research, including developments in cryptographic algorithms that allow private computation on data and differential privacy, as well as accountable systems?

We learned in this workshop that there are a variety of developments from computer science that can help those who work with big data techniques to manage the use of personal data according to rules, as well as to offer scalable solutions for those who assess, manage, and regulate privacy in these contexts.

Daniel J. Weitzner
Principal Research Scientist
MIT Computer Science and Artificial Intelligence Lab
Director, MIT Information Policy Project

Elizabeth J. Bruce
Executive Director
MIT Big Data Initiative at CSAIL

Introductory Remarks by L. Rafael Reif, President of MIT

Adapted Transcript of Talk

Good morning to everyone. At MIT I've learned that there is a clear, direct correlation between snow on the day of a big event, and the importance of the topic. Unfortunately, this correlation hit DC today, and all the flights, I understand, to and from DC have been cancelled. Nevertheless, judging by our turnout this morning, it is clear that the question of big data and privacy is not only very interesting, and important, but it is so to a very broad range of people. This gathering includes leaders from industry, from universities, from government. We have here among us experts on policy, technology, and politics, and our speakers will explore big data privacy issues in terms of healthcare and education, civil liberties, national security, and more, and we need this diversity of perspectives, because the subjects we explore today affect the whole spectrum of society.

We are together this morning because President Obama has focused his administration on finding the best path forward for the nation on the complex and urgent questions around big data. This workshop is the first of three university-based events that the White House is co-sponsoring to bring the most important issues to the surface. Our assignment is to look at privacy. How do we capitalize on big data's potential for good, while maintaining essential privacy protections? How do we design future technologies, policies, and practices to get that balance right for our society? We are here because MIT brings substantial strength in this field, strength magnified today with the insight and experience of our brilliant colleagues from other sectors, and institutions. With all this talent here today, we have been asked to help define the terms of the national conversation, to help set each direction, and to raise its ambitions.

Central to this conversation are the tensions around big data and privacy, the tremendous opportunities, and the profound and pervasive risks. At this gathering of experts on a day that promises superb speakers on these topics, I will be brief, but I do want to offer one example on how these tensions affect our work at MIT, not only our work on big data, but our work in other domains, where big data is a fact of life from biomedicine and healthcare to energy, to digital learning. In these fields and many others, figuring out how to use big data for the most benefit with the least risk of harm presents a fascinating, and deeply important challenge. For instance,

MIT is helping to push the frontiers of digital learning, both for a global audience of millions and for our own students at MIT.

Later this morning we will hear from Anant Agarwal, President of edX, the open source global learning platform that MIT and Harvard launched two years ago. I will just touch on the issues for the digital courses created by MIT faculty on the edX platform, which we call MITx. For us, as educators, today's digital learning technologies offer an incredible new opportunity to learn about learning. To give you a sense of scale and scope, MIT has about 129,000 living alumni, but in less than two years MITx has attracted more than 760,000 unique registered learners from more 190 countries, and they have already generated more than 700 million records of student interactions with the edX platform.

We want to study the huge quantities of data about how MITx students interact with digital courses. We want to make sure what really works. We want to use what we learn to improve the way we teach and to advance the science of teaching overall. There is so much to learn that we also want others to be able to study our data. We intend for MITx data to be constructed and curated as a public trust, all worthy goals, I believe, but at the same time we value privacy and so does the federal government; MITx student data is governed by the Family Education Rights and Privacy Act, or FERPA. FERPA is the same law that says that students over 18 have the right to keep their academic records private, even from the parents paying their tuition. That means that any efforts to use MITx data run into significant practical challenges, as well as serious ethical constraints.

For instance, legally speaking in a massive open online course, who counts as a student with protections under FERPA: Those who register, but never view course content? Those who view about half of the course content? Those who explore the course deeply, but don't take the final exam, or only those who actually earn a certificate? Are all these sub-population students in the FERPA sense? And, to allow for research, can their data be suitable de-identified? MITx classes include a public forum component. To date, MITx users have posted more than 423,000 forum entries. Their postings often include a great deal of personally identifying information. Correlating forum data with institutional releases of identified data could produce serious breaches of personal information. How do we set the boundaries and balance the competing interest?

If you believe in the potential of digital learning, you have to care about the larger question: How can we harness this flood of data to generate positive change without destroying the very idea of privacy? A pile of questions hovers over our

work in field after field. Fortunately, many people in this room are better qualified to answer these questions than I am, including our keynote speaker.

So, it is my pleasure to introduce someone who has been instrumental in shaping U.S. policy on digital privacy since the 1990s, and the person President Obama has trusted to lead the national conversation on big data today. From 1998 until 2001, John Podesta served President Bill Clinton as Chief of Staff with responsibility for all White House policy development, daily operations, Congressional relations, and staff activities. He coordinated the work of Cabinet agencies around federal budget and tax policy, he served in the President's Cabinet, and as a principal on the National Security Council. Previously he has held positions of influence in Washington on a wide variety of subjects, including agricultural patterns, regulatory reform, telecommunications, security, terrorism, government information, and privacy.

A lawyer by training, Mr. Podesta is a Visiting Professor of Law at the Georgetown University Law Center. In 2003, he launched the Center for American Progress, a leading progressive think-tank that focuses on 21st century challenges, such as energy, national security, economic growth and opportunity, immigration, education, and healthcare. In 2008, President Obama chose John Podesta to lead his transition to office. After that assignment, Mr. Podesta returned to the Center for American Progress, as President and CEO. Then last December, President Obama asked him to rejoin the White House team.

In his current role as Counselor to the President, Mr. Podesta is driving a variety of key initiatives forward, from workforce education and the President's Climate Action Plan to this new work around big data. We are very fortunate,and very grateful to be able to count on his wisdom today. Unfortunately, due to the storm that hit DC, Mr. Podesta will be joining us by phone. After his talk, he will welcome a few questions. Please join me in welcoming the voice from DC of the White House Counselor, John Podesta.

Keynote: John Podesta, White House Counselor

Remarks as Delivered by Counselor John Podesta The White House/MIT "Big Data" Privacy Workshop March 3, 2014

Good morning. I'm sorry to be talking to you remotely. In my world, big data squared off against big snow, and big snow won. Secretary Pritzker is travelling from NYC and I hope she'll have better luck and be with you at lunch. And the Administration is well represented by our Deputy CTO Nicole Wong and the head of NTIA Larry Strickling.

I want to start by thanking President Reif, for joining the White House in this important exploration of the technologies driving the big data revolution. And I want to thank Danny Weitzner and Elizabeth Bruce in particular, not only for putting together this outstanding event, but for their ongoing contributions to the research in this area.

This workshop is the first in a series of events that the White House will be co-hosting with academic institutions across the country. So, it is a fitting time for me to provide some background about the 90-day White House study of big data and privacy, the process that we're currently undertaking, and what we hope to accomplish in the next several weeks.

As many of you will recall, on January 17, the President spoke to the American people about how to keep us safe from terrorism in a changing world, and at the same time continue to uphold America's commitment to liberty and privacy that our values and our Constitution require. In that speech, he asked me to lead a comprehensive review of big data and privacy, recognizing that national security is not the only space where changes in technology are altering the landscape of how data is collected and used, and challenging traditional conceptions of privacy. So, one purpose of this study is to get a more holistic view of the state of the technology and the benefits and challenges that it brings. This Administration remains committed to an open, interoperable, secure and reliable Internet – the fundamentals that have enabled innovation to flourish, drive markets, and improve lives. We also recognize that ensuring the continued strength of the Internet requires applying our timeless privacy values to these new technologies, as we have throughout our history, with each new mode of communication from the mail to the telephone to the social network.

We are undergoing 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. The immense volume, diversity, velocity, and potential value of data will have profound implications for privacy, the economy, and public policy. The White House working group will consider all of those issues, and specifically how the present and future state of these technologies might motivate us to re-visit our policies across a range of sectors.

There is a lot of buzz these days about "Big Data" – a lot of marketing-speak and pitch materials for VC funding. For purposes of the White House study, when we talk about "big data" we're referring to 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 requires 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.

There is no question that there is more data than ever before, and no sign that the trajectory is slowing its upward pace. In 2012, there were an estimated 2.4 billion global Internet users. The amount of global digital information created and shared – from documents to photos to videos to tweets – grew 9x in five years to nearly 2 zettabytes in 2011 (a zettabyte is one trillion gigabytes). On Facebook, there are some 350 million photos uploaded and shared every day. On YouTube, 100 hours of video is uploaded every minute. And we are only in the very nascent stage of the "Internet of Things," where our appliances will communicate with each other and sensors may be nearly ubiquitous.

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. For example, The Cancer Genome Atlas, an NIH-funded program, is using large genomic data sets to map the genetic changes in more than 20 cancer types. Their researchers have discovered that breast and ovarian cancers have genomic similarities that may have implications for treating these diseases.

With the exponential advance of these capabilities, we must make sure that our modes of protecting privacy – whether technological, regulatory, or social – also keep pace. Now, it's certainly true that data analytics is an old science, dating to the late 1800s. In this study, we want to explore whether there is something truly new in the vast collection of data and lightning-speed analytics that are made possible by new technologies, computational strategies and cratering storage costs. My hope is that this inquiry will anticipate future technological trends to help us frame the key questions arising from the collection, availability, and use of big data — both for our government and the nation as a whole – and develop a workplan to address them.

Today's conference – appropriately set at MIT, which has been the cradle for so many game-changing technologies, is part of this 90-day endeavor and is designed to provide a firm grounding in the current state of technologies and their likely trajectories.

The Administration's Big Data initiatives

It is important to note that the Administration is not starting from scratch when it comes to big data or privacy.

Since the earliest days of this Administration, the Federal Government has taken unprecedented steps to make government data more available to citizens, companies, and innovators. Through the Data.gov platform, which launched in 2009, users have been able to access thousands of government datasets about a wide range of topics. The Open Data Initiative and Executive Order that the President signed last year commits federal agencies to unlocking even more valuable data from the vaults of government in health, energy, education, public safety, finance, and global development.

The natural outgrowth of this commitment to making large data sets available for public innovation is a broad commitment to the technologies that can harness these assets. In 2012, the Administration announced a $200 million commitment by six agencies to invest in big data projects. And just last fall, we showcased 28 public-private partnerships harnessing big data to enhance national priorities, including economic growth and job creation, education and health, energy and sustainability, public safety and national security, and global development. Indeed, one of those projects was launched from here as part of MIT's Big Data Initiative. We are pleased to be collaborating with CSAIL's Big Data Privacy Working Group and we look forward to hearing from some of the researchers engaged in that project later today.

The United States can also be proud of its long history as a leader in information privacy, starting with the pioneering of the Fair Information Practice Principles in the 1970s. Those principles -- known as the "FIPPs" -- are the underpinnings of the Privacy Act of 1974, which articulates the rights of citizens and the obligations of government to protect personal information. The same principles have also become the globally-recognized foundation for privacy protection, adopted by the OECD and providing the framework for privacy regimes around the world. President Obama, from early in his first term, has been working to advance protections for individual privacy in this new age of information technology.

Indeed, the Administration announced a groundbreaking privacy document in 2012, with the release of its consumer privacy blueprint, including the Consumer Privacy Bill of Rights. The blueprint refined the FIPPs to be more focused on consumers in terms they could understand in their own lives. It also re-framed the FIPPs to better accommodate the incredibly innovative online environment in which we all now

live. While the document does not specifically use the term "big data," the blueprint recognized that significant data was being collected about individuals online and that some data would be sensitive. It also assumed that this data could deliver significant value, if properly used, to individual consumers.

What we will be exploring in this study is whether the Consumer Privacy Bill of Rights fully addresses the changes that today we refer to as the "big data revolution" – recognizing that we may only be at the beginning of that revolution. What the President wants to explore, in part, is whether our existing privacy framework can accommodate these changes, or if there are new avenues for policy that we need to consider.

Have we fully considered the myriad ways in which this data revolution might create social value and have we fully contemplated the risks that it might pose to our conceptions of individual privacy, personal freedom, and government responsibility of data?

As we move from predicated analysis of data – that is, using data to find something we already know that we're looking for, to non-predicated, or pattern-based searches – using data to find patterns that reveal new insights -- I think we need to be conscious of the implications for individuals.

How should we think about individuals' sense of their identity when data reveals things they didn't even know about themselves? In this study, we want to explore the capabilities of big data analytics, but also the social and policy implications of those capabilities.

Our work is still in its early stages, but already we're learning important things about the current state of technology and its potential. For example, we recently met with some leaders in higher education to discuss the use of academic performance data to improve learning outcomes. There is some terrific research happening in this area and it worth talking about in a bit more detail.

The Pittsburgh Science of Learning Center, an NSF-funded center that joins the disciplines of cognitive learning and computer science, hosts the "DataShop", the world's preeminent central repository for data on the interactions between students and educational software and a suite of tools to analyze that data. In collaboration with private sector partners, they have made large-scale data sets available to develop learning models aimed at improving math, science, and language curriculums for K-12 students. In one study, the researchers tested a new algebra curriculum for middle school to high school students that utilized education technologies for instruction and performance measurement.

As some of you may know, mathematics proficiency rates of students in the United States – while on the rise – are still far below what they should be and lag behind students in the top-scoring countries. While there is still more work to be done, the

early results of these large-scale studies show significant gains – 8 percentile points more than usual, an amount that is nearly double how much students learn from a typical algebra course.

Importantly, what this type of research underscores is that the use of educational technologies is improving the scope, scale, and granularity of data we have about how kids learn. With that data, and applying cognitive and data science to the problem, we are better able to understand how to help our children move up the performance curve. We are gaining insights that were not previously possible, or maybe were dismissed for lack of concrete data, because of the new capability to capture student performance in detail and at scale in diverse, real-world school contexts.

Of course, there are also privacy implications to be considered when gathering and using this data. While the educators working with the students obviously knew how individual kids were doing on their tests, the researchers who developed those data-driven education tools only had de-identified data and deliberately decided not to collect the demographic data on the students.

Now, given what we already know about effective education policy, demographic data might have been useful both in developing effective curriculum and in addressing the needs of the individual student. But the researchers decided that collection of such information raised privacy and ethical concerns, and that they could make progress without that data.

We can see similar real-world benefits and similar privacy questions raised in a range of areas: from tracking electricity usage in a home to significantly bring down energy costs, to collecting individual location data in order to reduce traffic congestion. I believe we'll be hearing about some of these innovative uses today, including uses in education, genomics, and transportation.

This is the power of big data analytics that could unleash real human potential and so a goal of this study is to look at where the federal government can play a role in supporting this type of work, while continuing to protect personal privacy and other values.

So that is the context of this inquiry.

The Study

Now, let me just take a few moments to explain a bit more about the review, its scope, and what you can expect over the next 90 days.

In his speech, the President asked me to lead a comprehensive review of the way that "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. I will be joined in this effort by Secretary of Commerce Penny Pritzker (who will be your lunchtime keynote later today), Secretary of Energy Ernie Moniz, the President's Science Advisor John Holdren, the President's Economic Advisor Jeff Zients, and other senior government officials.

This is going to be a collaborative effort with four channels of engagement.

First, the President's Council of Advisors on Science and Technology (PCAST) is conducting a parallel study to explore in-depth the technological dimensions of the intersection of big data and privacy. Their report will feed into this broader effort and ensure a substantive grounding in the technologies at issue.

Second, our working group is consulting with a wide range of stakeholders. We have already met with privacy and civil liberties advocates, business leaders, policymakers, international partners, academics, and several government agencies on the significance of and future for these technologies. In the next several weeks, we look forward to hearing from a broad range of private sector companies, particularly those who collect and use data to develop products and deliver services, whether by targeted advertising, improved medical treatment, financial services, and more.

We also will engage international audiences, including international regulators and officials, to help answer the President's charge that we consider "whether we can forge international norms on how to manage this data and how we can continue to promote the free flow of information in ways that are consistent with both privacy and security."

Third, this workshop kicks off a series of events that we are co-hosting around the country to convene stakeholders to discuss these very issues and questions. The next event will be on March 17, co-hosted with the Data & Society Research Institute and NYU, and will focus on the social, cultural and ethical implications of big data. Then, on April 1, we will co-host an event with the School of Information and Berkeley Center for Law & Technology at UC Berkeley, which will focus on the legal and policy issues raised by big data.

Finally, and perhaps most importantly, we want to engage the public. This is not a discussion that should be confined to Washington or academia. This is an issue of such importance, an array of technologies already so pervasive, that it requires public participation in the conversation about how we realize the great benefits of big data while protecting individual privacy and other values. To this end, this week I will be posting a video to the White House website that describes this inquiry and asks the public, "What technologies are most transformative in your life?" and "Which technologies give you pause?" We have also just initiated a process to receive written comments addressing these questions in even more depth. You can find both channels for providing your input on the White House website and we

welcome your comments and ideas. These discussions will help to inform our study.

This study is fundamentally a scoping exercise. We are trying to get a full view of the landscape – the technologies at play, the uses by the government, industry, and academia. Whether we want to examine the Administration's consumer privacy blueprint—including the Consumer Privacy Bill of Rights—and how its principles can be applied in this new landscape. That may prompt us to look harder at some of our existing policies, at our research agenda, or at specific sectors where great gains could be made by the use of big data.

When we complete our work, 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.

While we don't expect to answer all these questions, or produce a comprehensive new policy in 90 days, we expect this work to serve as the foundation for a robust and forward-looking plan of action.

This is a fascinating and complex area, so let me close by throwing out a few questions that we have been thinking about:

- What is genuinely "new" about big data and what, if any, policies should be revisited because of those changes?
- What business models do you think are most dependent, today, on big data? How will that change in, say, 5 years? 15?
- What types of uses of big data could measurably improve social or economic outcomes or productivity with further government action, funding, or research?
- Can we build additional privacy protection into the architecture of big data analytics and should the government and the private sector be investing more in research toward that end.
- For individuals, what do you think will be the most significant effects of these emerging pattern-based data mining techniques?

Thank you for your time this morning and your engagement in this national conversation. I am sorry I am not with you in person, but I'll be watching the feed. If we can manage the technology, I think a have a few more minutes to take some questions.

See Appendix A – Q&A with audience

State of the Art of Privacy Protection -- Cynthia Dwork, Microsoft Research

Modern cryptography offers both a language for talking about privacy and its loss and a methodology for approaching the preservation of privacy. In pre-modern cryptography, a cryptosystem would be proposed, then it would be broken and someone would try to fix it, and then the fix would be broken again. This resulted in a cycle of proposals, breakages, and new proposals. A different approach, heralding the modern era of cryptography, was introduced in the early 1980's in the works of Goldwasser, Micali, and Rivest.[1] In this paradigm, a definition would be proposed, in part to answer the questions, "What do we want from a cryptosystem? Specifically, what does it mean to break the system? And what is the goal of a hypothetical adversary?" The researcher would then develop an algorithm and prove that it satisfied the definition. If the cryptosystem was broken later, it meant that the definition needed to be strengthened; the process would begin again, but with a stronger definition. The focus on increasingly strong definitions and the iteration of algorithms marching in tandem led to the development of stronger cryptosystems and converted the cycle of propose-break-propose to a path of progress in the field.

The "privacy dream" is that a database filled with useful but sensitive information is managed by a well-intentioned, computationally powerful curator. The curator is a computer system that will sit between the raw data and those users who want to access it. In the dream, the curator will "sanitize" the database, removing all privacy risks, producing a data set that may be accessed by data analysts for all research purposes. The question revolves around how the sanitization takes place and how secure the output is. Through the work of Dinur and Nissim,[2] it is clear that there are limits to the level of privacy protection that may be provided if statistical utility is to be maintained. This fundamental law of information recovery says that overly accurate answers to too many questions is blatantly non-private. One popular approach to protecting privacy entails the anonymization or "de-identification" of data, which includes the removal of Personally Identifying Information (PII). However, in some cases, identity can be reconstructed fairly easily. The following examples help to define what kind of privacy might be offered by de-identification. One well-known research project conducted by Sweeney linked public health data to voter registration data and re-identified the anonymized record of William Weld, the Governor of Massachusetts at that time. Sweeney estimated that about 87% of

[1] Goldwasser and Micali ACM STOC 1982, Goldwasser, Micali and, Rivest, IEEE FOCS 1984.

[2] Dinur and Nissim ACM PODS 2003. See also, for example, Dwork, McSherry, and Talwar STOC 2007; Dwork and Yekhanin CRYPTO 200; Kasiviswanathan, Rudelson, Smith, and Ullman STOC 2010; De, Theory of Cryptography Conference (TCC) 2012; Muthukrishnan and Nikolov STOC 2012; and Kasiviswanathan, Rudelson, and Smith, Symposium on Discrete Algorithms (SODA) 2013.

the population could be identified by using the sex, date of birth, and zip code of individuals. However, other fields in the database record might also constitute identifying information, for example, the combination of zip code and dates of previous admissions to a hospital, or previous admissions and family history. What constitutes identifying information depends on what is known by the party trying to do the identifying.

The power of "side information" for enabling re-identification was exploited by Narayanan and Shmatikov, who studied Netflix data and found that it was possible to identify many individuals in a set of approximately 480,000 people merely by the titles of three movies that they had seen and the approximate dates of those rentals.

Various approaches to these problems have significant utility, but they also have flaws. Sweeney and Samarati proposed k-anonymity, which does not account for the richness of side information and also allows for the possibility that sensitive information about an individual will be leaked, even without matching a specific row to a given individual. Sometimes even one's presence in the dataset can be sensitive information, for example, if a study includes a dataset of patients that are HIV positive.[3] L-diversity has also been proposed, but it does not protect against re-identification when there is a series of releases of an evolving database.[4]

Further concerns involve the issue of leakage in general; every time a publicly observable action takes place, some information is leaked, as was demonstrated through the billing system for targeted advertisements on Facebook.[5] Calandrino et al[6] showed that one person's preferences could influence another person's experiences; with knowledge of someone's blog and the use of an evolving "similar items" list, an adversary could infer purchases that were not discussed publicly on the blog. Finally Homer et al[7] combined the single nucleotide polymorphism (SNP) statistics from a genome-wide association study with a target's actual DNA and statistics about the general public; they were able to generate information about the membership in the study. As a result of this work, the publication of aggregate SNP statistics is no longer permitted in NIH-funded studies.

The history and current evolution of cryptography indicate that complexity of the type that is present in the big data environment requires a mathematically rigorous theory of privacy and its loss. It is difficult to discuss tradeoffs between privacy and statistical utility without a measure that can capture cumulative harm over multiple uses. Furthermore, other fields such as economics, ethics, and public policy cannot be brought to bear without a "currency," a measure of privacy by which to gauge risks and weigh benefits. This gives rise to the development of differential privacy

[3] Samarati and Sweeney, 1998.

[4] Xiao and Tao, 2007; LiliVenkatasubmraminan, 2007.

[5] Korlova, 2012.

[6] Calandrino et al, 2011.

[7] Homer et al, 2008.

and its parameter, usually called “epsilon,” which serves to measure the loss of privacy[8]. Differential privacy ensures that the probability of any outcome, good or bad, is essentially unchanged by an individual's choice to join, or refrain from joining a dataset. The parameter epsilon specifies precisely what is meant by "essentially" unchanged and allows the curator to make guarantees about the *ratio* of risks of a bad event when the individual opts in and opts out, even when the actual risks are completely unknowable (which is virtually always the case).

Differentially private algorithms permit the curator to track the potential for privacy loss and to make adjustments. For example, one useful technique is to add carefully calibrated random noise to the outcome of a computation in order to hide the presence or absence of any individual. If the goal is very little privacy loss, then more uncertainty (random noise with large variance) should be introduced; if greater privacy loss is acceptable, then less random noise will be required. Finally, the adversary’s background knowledge is irrelevant in this situation; if an algorithm is differentially private, then it is differentially private regardless of what an adversary might know. In consequence, differentially private algorithms are immune to re-identification attacks. One of the biggest challenges lies in the gray areas: in some cases, it is clear that a large loss in differential privacy -- a mathematical notion of loss -- would lead to a real human notion of privacy loss. In other cases, it is not clear that an actual breach would occur. One step forward would be to publish the epsilons and penalize the parties involved in situations when the epsilon is infinity.

The current literature on privacy draws on a range of frameworks and techniques from algorithms, cryptography, and statistics to convex geometry, complexity theory, machine learning, programming languages, verification databases, and economics. The research is promising and there is much more to be done.

Key questions from the audience were as follows:

1) Can you give an example of a practical use of differential privacy that is in use today, either in industry or academia?

Cynthia Dwork: One example is the on-the-map website from the Census Bureau that has information about where people live and work, so it is useful for studying commuting patterns. One of the significant problems with putting a system forward has been the question of what epsilon should be; if there is no guidance on epsilon, how would differential privacy work? There are several programming platforms, including PINQ for privacy integrated inquiries from Frank McSherry at UT Austin that facilitate doing things in a differentially private way.[9]

---

[8] Dwork, McSherry, Nissim, and Smith, TCC 2006; see also Dwork, ICALP 2006, Dwork and Naor, Journal of Privacy and Confidentiality, 2010.

[9] http://research.microsoft.com/en-us/projects/pinq/

2) Could you talk about the scalability of your data curation algorithm? It seems to be n-squared in a number of queries, so how will this scale?

CD: If everything were only n-squared, we would be thrilled. Some computations are extremely efficient, so the basic idea of having appropriately scaled noise is essentially no less efficient than the non-private version. However, sometimes there is a cost; this is an active area of research.

3) How should we handle data for things that are mandatory, such as elementary school standardized tests, or the MCAS? If we are maintaining this data, do we have different levels of expectations for protecting it - for example, would there be different epsilons for transactions that were less voluntary?

CD: Differential privacy is applicable in the statistical analysis of very large datasets. If you are trying to understand something about a particular individual, it does not apply; the point is to hide the presence or absence of an individual within the data set as a whole.

4) How should we view the choice of being in a given data set or not. For example, if I suspect that I may have a genetic risk factor for Alzheimer's disease and I do not want my employer to know about that, then I could decline to get my genome sequenced. In this case, the information is very safe. If I do get my genome sequenced and it resides in a database, then there is some risk that it could be disclosed?

CD: The question revolves around what data you think is made accessible. What we have in mind is that the raw data would not be accessible to anyone and all access to the data would pass through a differentially private mechanism. If this guarantee were in place, then you would not be at risk.

Panel 1: Big Data Opportunities and Challenges

"State of the Art of Big Data Technology" -- Michael Stonebraker, MIT CSAIL

Big data can be defined by volume: too much data, velocity: data that is coming in too quickly, or variety: data that is coming from too many places in too many formats. In cases where the volume of data is an issue, there are ample and scalable hardware solutions. Large, mature commercial vendors offer data warehouses for structured data and there are a few dozen production databases available in the petascale range. Hadoop is also used in this environment for semi-structured data and there are a number of petascale Hadoop installations as well. One challenging aspect of volume exists in the realms of predictive modeling, non-predicate data mining, and data clustering, which involve very complicated analytics. The construction of efficient and scalable data management systems is not as well understood in these contexts. However, a significant amount of research is underway and solutions will evolve in server side technology to meet the needs.

In cases where the velocity of the data is an issue, the problem is often due to legacy systems. For example, on Wall Street, financial professionals are trying to cope with the tremendous increases in trading volume on the exchanges and substantial IT investments are being made in the infrastructure of investment banks and exchanges. In a sensor network context, velocity can be handled to some degree by aggregation. Work is also being done to integrate query languages and the integration of storage with on-the-wire processing. As with the volume issue, the solutions to the problem are visible and the technology will eventually be capable of handling the velocity of data that results from the "Internet of Things."

The greatest challenge lies in variety; the technology for managing data in data warehouses may scale to several dozen sources, but there are no seamless solutions for integrating thousands of datasets into a single coherent system. Novartis is one example of a firm that has thousands of independently constructed data sources and is seeking a solution for their integration. This is an active area of research, with a great deal of innovation taking place in start-ups and established firms seeking to address customer needs in data integration and management.

Database security is well-defined and has been part of the SQL standard for many years. Encryption is also possible in the DBMS environment and can be entrusted to the database system or the client, depending on the nature of the data and the need for encryption. The greatest concern with security stems from human factors; insiders are often the culprits and unwitting employees exacerbate the situation by failing to secure their desktops properly. One practical policy suggestion for security and privacy is that the database system writes a command log that covers

everything that happened on the system and then adds an auditing system that searches the log for unusual activities. This will not catch all intrusions or thefts the first time that they occur, but it will flag suspicious activities and may prevent further losses. Thorough auditing of information flows in big data systems will also help in this regard.

"Clinical Data: Opportunities and Obstacles" -- John Guttag, MIT CSAIL

The quantity and availability of medical data is growing at a rapid rate, through better instrumentation, the adoption of Electronic Medical Records (EMRs), and forces that support the aggregation of data across institutions. Unfortunately, very little of this data is being used for useful clinical research. Health care providers do not want to expose potentially embarrassing data and patients are perceived as fearing a loss of personal privacy. There is a trade-off between individual privacy concerns in a health care setting and the potential for learning things that can greatly benefit patients.

Five percent of patients admitted to U.S. hospitals acquire an infection peripheral to the reason for their admission; this figure is higher in some other countries. Health care-associated infections are among the top ten contributors to death. Infection with *Clostridium difficile* is one example, with more than 200,000 cases per year in the U.S. alone. There is an opportunity to use modern EMRs to develop accurate models for predicting which patients are most likely to acquire this infection. These models will not only allow health care facilities to understand in advance which patients are most likely to acquire the infection; the models may also be used proactively to reduce the incidence of these infections, thereby lowering the costs to the medical system and avoiding unnecessary harm to their patients.

The data used in the analysis includes information about medications, procedures, locations within hospitals, staff who have come in contact with patients, lab results, patient history, admission details, dates, and demographics. From a privacy perspective, this is a very substantial amount of sensitive data that is related to real treatment and further, it is not de-identified. Once the data was assembled, the technical challenges were revealed; ironically they were not related to big data; the kind of data gathered in this situation was actually too small. If research starts with a set of 3 million admissions and examines an infection that only affects 1% of the population, once the researchers have eliminated institutional differences, time-related differences, and other factors, there may be barely enough data left to draw useful conclusions. Nevertheless, by using a variety of machine learning techniques, the researchers produced a good model that one hospital has integrated into their online system.[10] The case can be made that medical data is special, not only because privacy is important, but also because progress in healthcare is too urgent to wait

[10] "A Study in Transfer Learning: Leveraging Data from Multiple Hospitals to make Hospital-Specific Predictions," J. Wiens, J. Guttag, and E. Horvitz, JAMIA 2014.

for all of the privacy issues to be resolved completely. The risks of the inability to conduct research on EMR datasets and compare results across groups include the financial and human costs of avoidable pain, suffering, and death. The solution may be to focus on auditing mechanisms and to develop systems of enforcement and punishment for those who misuse the data in these special research environments.

"Importance of Access to Large Populations" -- Manolis Kellis, MIT CSAIL

On any given project, the number of datasets that are used will make a tremendous difference in genomics research. With just a few datasets, a pattern may not be visible at all and there may be a large number of hypotheses, but as the number increases, there will eventually be an inflection point where a pattern is revealed. In order to reach such an inflection point, researchers must be able to overcome the limitations on data sharing that are prompted by privacy concerns. The goal is to learn more about the mechanistic basis of human disease and the challenge is that the effects of individual variance in genetics are very small, so extremely large cohorts are needed to discover them.

One example would be a variant that increases a predisposition to age-related macular degeneration. The link from genetic studies takes the researchers directly from the variant to the disease, but does not provide guidance on the mechanism. The researcher needs to understand the specific classes of regulatory elements where the variant is acting, the target genes of these regions, and the intermediary effects that are influenced by the environment and perhaps by the disease itself. This information will provide a mechanistic basis for developing therapeutics.

To piece all of the information together effectively, researchers need to develop thorough knowledge of the genome function. The NIH has assembled a large number of studies that attempt to understand each functional element in the human genome systematically. This has enabled a completely renewed perspective on the nature of human disease, including the role of variants that fall into non-coding regions and a view on what tissues are relevant for many diseases.

A second revelation is that the significant effect variants that have been the subject of much genomics work in the past may diminish in scope now, because complex disease works through a combination of small effect variants. Such variants are ubiquitous and researchers need to work with very large cohorts in order to discover them. However elusive they may be at first, these weaker effect variants coalesce into specific pathways and regions that can be targeted later on.

The challenges in collecting the data arise, in part, from limitations in the consent forms used by hospitals; these forms often do not grant researchers access to the data generated in these environments. Technologies that are designed for privacy protection may help to overcome some of these limitations. The ability to study

large cohorts can lead to new biological insights, including, for example, the discovery that certain conditions such as schizophrenia are actually heritable medical disorders. Therefore, enabling collaboration, consortia, and the sharing of datasets will provide benefits to the research community and society as a whole.

"Cars, Phones, and Sensors: Mobile Big Data" -- Sam Madden, MIT CSAIL

Mobile big data arises from smartphones, vehicles, watches, and other sensor-equipped devices. This type of data is a part of the "Internet of Things," where all of the important objects in our lives are enhanced with sensors and interconnected with networks. Cell phones are the driving force of growth in mobile big data, but sensors in homes and other locations will play an important role as well. As an example, in 2011, there were 5 billion cell phones deployed globally; 1 billion of these phones were Smartphone devices, with broadband Internet connections and the ability to sense certain things about the environment. Examples of sensors include GPS, proximity and motion sensors, accelerometers, and gyroscopes, as well as networking technologies like BlueTooth and Wi-Fi that can be used to measure proximity and location.

At MIT, researchers have been engaged in a project known as "CarTel" for about five years. This project deployed sensors on cars in order to measure certain aspects of the transportation environment. Sample applications include traffic sensing, which measures the delays as users travel, making maps of where potholes are, and also evaluating the behavior of drivers to see if they are executing risky maneuvers in braking, accelerating, or turning. This information was bundled into a "Commute Portal" that allowed users to study their own driving behaviors, observe road conditions that could affect their driving patterns, and share information with friends in a social network.

One of the key "big data" applications in this space is to move from personal analytics to larger societal benefits. In this context, raw data is generated from a combination of location and sensor data; this data is passed through signal processing to produce information that may be of interest to the individual and then aggregated across individuals. It can become the basis for more broadly useful applications, examples include scoring roads to prioritize road repairs and modeling of driver risk and safety for insurance rating purposes.

Sensors can also play an important role on medical monitoring and outpatient care, potentially reducing the amount of time that people spend in hospitals, saving costs and reducing risks of things like hospital-acquired infections. Similarly, sensor-based fitness applications like Runkeeper, Strava, and Fitbit can collect raw data about someone's activities and produce a set of personal performance metrics. If this information is aggregated, it can be used to develop societal level metrics that offer insights into the wellness of a population.

There are tradeoffs between privacy and public good with the use of smartphones and sensors. For example, studies show that young male drivers will dramatically reduce their risky behavior if they know that they are being monitored.[11] On the other hand, this kind of monitoring can be seen as invasive. Managing this tradeoff between privacy and societal benefit is a challenge we as a society must address.

"Big Data Opportunities for Improving Online Education" -- Anant Agarwal, edX

Massive Open Online Courses (MOOCs) yield tremendous amounts of data that may be analyzed to improve learning platforms and processes dramatically. edX is a MOOC provider launched by Harvard and MIT, which now hosts 2 million learners from 196 countries, with a total of 4 million course enrollments. edX offers 150 courses, spanning every discipline, from math, sciences, business, and medicine, to law, humanities, arts, and music. edX has partnered with a number of universities, including Tsinghua in China and IIT Bombay in India, where the courses are deployed on campus, with approximately 10,000 students enrolled on that basis.

In terms of the volume of data that is generated on a MOOC system, the first offering on the edX platform at MIT was a circuits course, with an enrollment of 155,000 students. This inaugural effort produced 230 million clickstream records, each with about 1 kilobyte of information per record, so nearly a quarter of a terabyte of data for the first course alone. Over the range of all course offerings, the data has become quite vast and is useful for the analysis of learning behaviors. For example, students will watch videos and interact with exercises. The data reflects how many attempts they make at each problem in the exercises. Educators can gauge which problems are easy and which are more difficult for a given group of students.

There are also opportunities to study how students interact with each other in peer learning discussion forums, where they work to solve problems together and may also discuss questions with professors. These peer learning efforts frequently cross national boundaries and offer unique opportunities for long-term educational collaboration. Additional areas of study include the role of homework in the learning process. Initially, researchers find that about 70% of the students will watch the course lectures first and then complete the homework. However, in the final weeks of the course, the percentage has flipped and about two thirds of the

[11] McGehee, D.V., Raby, M., Carney, C., Lee, J.D., Reyes, M.L. (2007). Extending parental mentoring using an event-triggered video intervention in rural teen drivers. Journal of Safety Research, 38, 215-227. See, http://garage.a2om.com/documents/driveiq/research/Extending%20parental%20mentoring%20using%20an%20event-triggered%20video%20intervention%20in%20rural%20teen%20drivers.pdf and also http://www.theguardian.com/money/2011/mar/20/coop-telematics-lower-car-insurance-young-drivers?INTCMP=SRCH&guni=Article:in%20body%20link

students will attempt to complete the problem sets first, referring to the videos and lectures as needed. This is a meaningful statistical result that has implications for the way in which courses are taught. Another point involves student engagement with the videos – the median viewing time is about 6 minutes and yet lectures are about an hour long. Finally, a study from Harvard and MIT showed that taking the time to complete homework assignments has a strong positive correlation with overall performance in a course.

Big data in the MOOCs context supplies researchers with statistically significant results that can influence learning methodologies in real time. On the question of privacy, edX shares identified data with its university partners and de-identified data for all of the universities with all of the partners. The scope of the sharing may be broadened further in the future, as the de-identified form of data is established more concretely.

Keynote: The Honorable Penny Pritzker, Secretary of the US Department of Commerce

Good afternoon. I want to thank President Reif and everyone at MIT. Dr. Reif co-chairs the Administration's Advanced Manufacturing Partnership, housed at the Commerce Department.

I also want to thank John Podesta for his leadership on this important issue. I am pleased to be here at the first of three workshops to start a national conversation among business leaders, academic experts, and civil society advocates. As we all know, the American economy has always been grounded in a commitment to the free flow of data and information. We believe this is good for business and good for society as a whole. You can trace this notion all the way back to the writing of our Constitution, which called for a decennial Census.

The 1890 Census comes to mind. A statistician named Herman Hollerith invented a machine that was able to read the holes punched in paper cards used in the 1890 Census. His machine shortened the time it took to complete the Census, not just by months, but by years. It saved the government $5 million dollars – that is roughly $125 million in today's dollars. Notably, his company would later be folded into an iconic American company, IBM.

Fast forward. Today, data and data analytics are a powerful new fuel of the American economy. Here in Massachusetts, many companies have been built on data, from large established companies that provide the backbone of the digital economy – like cloud-computing leader Akamai to EMC, which offers massive data storage for its customers to companies like EnerNoc, born in Boston, which is using data and software to make our buildings and schools more energy efficient to Big Belly Solar that uses data to manage its smart, Internet-enabled trash receptacles, allowing garbage trucks to pick up containers only when they are full and saving on fuel costs here in Boston and in other cities around the United States.

Simply put, each one of us here today benefits from the power of data and information, whether we are looking for nearby restaurants on our smartphones with Yelp, or getting a cab with Uber, or the example John used about the Cancer Genome Atlas that could help treat diseases in better ways. And yes, we are only scratching the surface. A recent McKinsey study examined the economic potential of open data in seven areas (education, transportation, consumer products, electricity, oil and gas, health care, and consumer finance). McKinsey's analysis showed that open data in these sectors could help unlock $3 trillion dollars in additional value to the global economy.

For our part at the Commerce Department, we are pushing to make more and more federal data available. I know the power of Commerce Department data first-hand. I used Census Bureau information to launch my first business over 25 years ago. My team needed to know the right places to build senior living centers and the Census Bureau was critical to our decision-making. Just a few months ago, I announced that unleashing more data for the public good would be top priority for the Department, because we know that data helps businesses make better decisions and data can launch new companies (and even new industries) -- creating good jobs.

One week ago today, we delivered on that promise by taking the first step to partner with industry to unlock more climate data collected by our National Oceanic and Atmospheric Administration – data that already powers a billion-dollar industry including the Weather Channel and weather apps. At almost every business roundtable, I hear from CEOs who emphasize the importance of using data to grow, compete, and innovate. In fact, faculty here at MIT's Sloan School of Management have conducted research showing that firms which adopt data-driven decision-making have productivity gains 5 to 6 percent higher than alternatives.

Overall, what is clear to me are two things: first, the free flow of data and information is good for society as long as it is respected and used properly. Second, the economic winners in the global economy will be businesses, governments, and other institutions that harness the potential of data. And, yet all of this potential hinges on one thing: trust.

The Administration has used a number of tools to protect privacy, protect our networks, protect the free flow of information, and ensure trust. We convened 300 stakeholders to develop voluntary codes of conduct for privacy disclosures affecting consumers who download apps. We worked with the nation's critical infrastructure providers to create a framework of standards and best practices in cybersecurity. We collaborate with European leaders to ensure a viable and effective framework for meeting privacy requirements in different markets. We work with law enforcement officials go after cybercriminals and deny them safe havens. We partner with other governments to ensure the free flow of information so essential to the democratic process. We work to prevent and stop data theft and piracy. And – yes – we protect our country from terrorism and acts of aggression that can be carried out by exploiting data.

The fact is we have made progress. But the pace of technological change is fast and we must continually evaluate the state of trust essential to reaping the benefits of data. Let me be very clear: The Administration and the Commerce Department are using all of the diplomatic and commercial levers at our disposal to ensure trust and to show that the confidence placed in our companies remains rock solid. As a business person myself, I understand the hesitation that some businesses might have about engaging in efforts like the ones I just listed – or in this 90-day review, for that matter. But what I am finding is that more and more American CEOs understand that our country's commitment to free flows of data and information is

part of our competitive advantage in the 21st century.

We must protect that advantage. How? By working together to uncover the best technologies and practices that lead to enhanced trust in our relationships with stakeholders, including our customers. In short, I believe that we must establish principles and policies that encourage and protect trust among all stakeholders. Each one of us has a role and responsibility in building a fabric of trust surrounding data and privacy issues. Governments must cooperate in a number of areas.

This includes law enforcement – so that we can prevent and address data breaches. It includes cybersecurity and data protection so essential to trade and commerce. It includes supporting the multistakeholder process that sets the standards and norms of cyberspace. And it includes working to ensure an open, fair and free Internet. In addition, governments must also cooperate with each other in how we address and harmonize our information technology regimes. Businesses – like many of you in this room -- also play a crucial role in ensuring trust in open and dynamic economies such as ours. You promote trust when you reach out and explain to your customers in very simple and straightforward terms how you plan to use their data. In addition, your capacity to build trust is also affected by the many other decisions you make every day: the data management systems you choose, the investments you make in risk management, even the training standards you adopt for your employees.

Finally, consumers and citizens themselves must actively help build a fabric of trust – in the way they make choices to share or protect information, and the way they behave online to help ensure free and open cyberspace. Overall, I believe that trust is absolutely necessary for any data-driven business to succeed. Without it, no business can survive. Put another way, all of the data in the world is worthless unless consumers trust the companies they buy from, unless citizens trust their governments, and unless institutions of all kinds trust each other to play by the rules.

Like John Podesta, I believe that we must continue this "Big Data" review by asking questions – (and I know that you have already started to do just that this morning.) A few questions on my mind are: What are the principles of trust that businesses and governments need to adopt in the age of big data? How can we ensure that new technologies protect consumer data, while also supporting innovative uses of data to benefit those same consumers? How can we be more accountable and transparent in how our institutions address privacy issues? What can be done to encourage consumers to better understand what data they are sharing, with whom, and for what purposes – in order to give them clear and consistent control? What can we all – government, businesses and consumers, do to address some of the more egregious and unanticipated consequences of big data collection and analytics? And is there consensus that some action is needed? Yes, we want to know what the government should be doing, but we also want to know what you think is the role of business, and what are the responsibilities of the consumer?

Exploring these questions is what this workshop today is all about. And the input from industry throughout this 90-day review is absolutely crucial. We need your help to set the agenda. As members of the Big Data Working Group, our commitment is to engage with you – to listen to your concerns and to push for answers to the questions raised. With the help of everyone here today, I believe that we can continue to successfully apply our American values to the technologies and circumstances of our time. We can achieve success in unleashing the full potential of data for the benefit of society, just like Herman Hollerith did with his new machine during the 1890 Census.

Thank you again to John Podesta and thanks to all of you for stepping up to engage in this important dialogue.

Panel 2: Privacy Enhancing Technologies

"The Promise of Cryptography: From a General Theory to Applications" -- Shafi Goldwasser, MIT CSAIL

In the big data environment, there is great potential for advances in health care, education, economic growth, and law enforcement. There are also enormous risks regarding the loss of control over private data. The risks include sacrificing data integrity, losing your anonymity, being profiled, and relinquishing your competitive edge when everyone has access to your information. The technical question is how to reap the benefits without incurring the risks (or at least minimizing the risks) that are implied by an overwhelming loss of control over your private data.

The classical methods for addressing these risks are not sufficient. Different entities collect and protect data in an uncoordinated manner and cross-referencing of information held by these entities makes matters worse. The traditional solution of data encryption is useful for hiding information, but makes data processing impossible. Anonymizing the data, which is the solution often used in practice, seems to fail in data rich environments, where it may become easy to reconstitute identity through the aggregation of information.

Starting in the 1980s, techniques have been developed that allow for performing specific and targeted computations on data, while keeping it secret. In a sense, they allow one to compute the benefits "in the dark," without seeing the data. The Secure Function Evaluation (SFE) technique developed since the mid 1980's entails a mathematical formulation, where a function can be run over disparate private data sets and will return an exact output result, so that only the output is visible to the parties; the inputs and data sets remain hidden. Hypothetical examples for SFE include analysis of medical data to see if a certain gene is prevalent in a population without seeing all of the medical data, evaluation of surveillance photographs to see if a suspect appears at a certain location without seeing the photographs in their entirety, and assessments of financial stability to see if some banks will become insolvent under certain conditions, without providing a complete explanation of their assets and investment strategies.[12]

The theory of SFE is well worked out. For practice, there is a highly-regarded survey by Lindell[13] on optimized implementations for a simple class of functions that address common queries; the implementations demonstrate very good practical performance. However, SFE still requires interaction and is not robust to an insider that might leak

[12] Lo et al, 2012.

[13] Lindell, 2013.

information. To this end, one may use Fully Homomorphic Encryption (FHE)[14] and Functional Encryption (FE),[15] or partition the data across several data centers, some of which can be relied on not to communicate with the others.

To conclude, SFE, FE, and FHE allow one to evaluate functions on data without seeing it. However, the value of these functions in themselves may reveal too much information about the data, especially if repeated functions are evaluated on the same data, or the function values are aggregations over data sets that were chosen, in part, maliciously. It is important to determine which classes of functions are safe to compute by SFE first. Another important goal is to develop a combination of privacy and secure computation in a two-stage process. In the first stage, there is a decision that the function or algorithm should be computed and in the second stage cryptographic techniques such as SFE, FHE, and FE are applied to perform the computation securely. Moving from data to programs, there are techniques for protecting privacy in browsing, searching, social interactions, and general usage through Program Obfuscation Methods.

There is much more research to be done on the processing of encrypted data and differential privacy protection in the context of both computer programs and web-based activities.

---

[14] Gentry, 2009.

[15] See, for example, Amit Sahai and Brent Waters. Fuzzy identity-based encryption. In Eurocrypt, pages 457–473, 2005.
Sergey Gorbunov, Vinod Vaikuntanathan, and Hoeteck Wee. Attribute-based encryption for circuits in STOC13.
Goldwasser, Kalai, Popa, Vaikuntanathan, Zeldowich, "Reusable Garbled Circuits and Succinct Functional Encryption, in STOC13
Goldwasser, Goyal, Jain, Sahai, "Multi-Input Functional Encryption," to be presented in Eurocrypt 2014.

"Using Cryptography in Databases and Web Applications" -- Nickolai Zeldovich, MIT CSAIL

One of the largest risks in data privacy is the potential for disclosures of vast amounts of private information due to flaws in servers, or deliberate attacks from adversaries. If the server is simply providing a means of storage, then encryption works quite well. However, if computations will be done on the data, then it will be difficult to work with data that is encrypted through traditional methods. Another approach to data protection is to construct walls around the data; there are ways to enforce security policies in operating systems and hardware and there can be firewalls within networks. In this case, the systems and software can become quite complex. Due to the challenges inherent in writing software correctly, attackers are frequently able to find and exploit programming mistakes, thereby compromising system security. A final threat to data protection lies in the employees and vendors who have access to the systems. Even if these individuals are trustworthy, it is possible that someone else can subvert an account and log in as an admin or authorized user and gain access to confidential data that way.

Since system compromises are inevitable, it is important to build systems that can protect data in spite of the breaches. One goal is to develop and refine ways to process encrypted data. A common approach entails storing encrypted data on a server and when processing is required, the client will send a key to the server so that it can decrypt the data and perform the processing tasks. The obvious problem is that if an adversary has compromised the server, then they will be able to take the key and use it for their own purposes. A more secure way would be to allow for computations over encrypted data without decrypting it; then encrypted results could be sent back to the client, who would decrypt them with a key on their end.

A system developed at MIT CSAIL that allows researchers to run database queries over encrypted data is called CryptDB.[16] One scenario would be a database that is running on a cloud computing platform and an application that is issuing database queries to analyze some of the data in the cloud. Assume that the application is trustworthy. CryptDB will interpose a proxy between the database and the application and the proxy will rewrite the queries in a certain way so that they can be run over the encrypted database. The database will provide encrypted results back to the proxy, which holds a master key and will decrypt the results, sending the final answer back to the application. There are numerous forms of encryption schemes to be assessed in this environment. Randomized encryption schemes allow for the movement of data back and forth. Homomorphic encryption permits addition. Other schemes allow for key word searches or equality-like operations over encrypted data. There are ways to join multiple datasets together and

[16] http://people.csail.mit.edu/nickolai/papers/popa-cryptdb-cacm.pdf

correlate them and there is also a scheme called order-preserving encryption that allows for sorting, order comparison, and related computations. Encryption schemes can be selected for efficiency and a significant amount of computation can be performed on the server.

These schemes have different security properties; the top layers have strong semantic security guarantees, while the bottom layers may tend to reveal issues like repeats among data items, or even the order of data items in a series. To adjust the encryption level of the server appropriately, it is possible to construct an "onion" of encryption, where the starting point is every value in the database in plain text form. These values are encrypted with increasingly stronger encryption schemes, perhaps starting with order-preserving encryption, then moving to deterministic encryption, and finishing with randomized encryption. In the end, the initial values are semantically secure and do not reveal anything about the data. As increased functionality is needed for processing, it is possible to strip away the layers to reach the level needed to perform a certain set of operations on the underlying data. The team has built a system that works quite well and supports a range of real data base applications from websites to transactional processing systems and data analytics. The server does not see the real data and does not have a key to decrypt it. The performance overheads are modest, around 20-30%, and a number of companies have adopted the design, including Google, which is running a service called Encrypted Big Query Client.[17]

A second approach is based on a design where there are many users and there may be a more publicly open system, like Gmail or Facebook. In this situation, each user has a distinct key that protects their data within the database and ensures that other users do not have access to it. The team has developed a system on this model called Mylar, which actually performs encryption in web browsers.[18] Each user has a key in their browser that could be derived from their password, for example, and when they enter information into the system, it will be encrypted automatically before it is sent onto the web application server and the database server. Even if these servers become compromised, the user's data will remain encrypted and confidential. When the user logs in again and obtains the data from the server, the browser will automatically decrypt and display it to that person. This system is being deployed with Newton Wellesley Hospital, in support of an application for endometriosis patients.[19]

Research challenges in this area include developing ways to perform a broader set of computations on encrypted data and computing on data encrypted with many different keys, in cases where data from many users may be involved. It will also be useful to develop ways to delegate limited functions over encrypted data, so that third parties can analyze it. In addition, there are practical systems issues and it is

[17] http://code.google.com/p/encrypted-bigquery-client/

[18] http://css.csail.mit.edu/mylar/

[19] http://people.csail.mit.edu/nickolai/papers/popa-mylar.pdf

important to develop mechanisms that provide improved security for end user devices and can audit systems for unintended data disclosures after the fact.

"Computing on Encrypted Data" -- Vinod Vaikuntanathan, MIT CSAIL

Two aspects of big data appear to be in conflict; big data provides an opportunity for a tremendous amount of computation and analytics, with possible benefits for society, but the use of this data is closely intertwined with privacy concerns. The central question is, "What are the best ways to bring functionality and privacy together in a harmonious manner?" New cryptographic tools provide some answers, particularly those that allow for computation on encrypted data; two examples are Homomorphic Encryption and Functional Encryption.

Data on servers must be protected from two different threats; outsiders, namely hackers, who may access millions of records in a single attack, and insider threats which could target data in a number of ways. Encryption is a way to minimize the "attack surface" of the data. Traditional encryption essentially entails putting data in a locked box and providing a key to an authorized party. However, the data cannot be used in this form.

One new approach to data protection is Homomorphic Encryption, a special system that will take encrypted data and perform computations on it, producing a new encryption that contains the results. The original data and intermediate steps are completely hidden and there are several ways to manage the decryption of the results. This method required new mathematics and techniques that were developed expressly for this purpose. As research in this area has advanced over the past few years, there are now fast implementations in software and downloadable libraries that can be used for the encryption. Hardware techniques involving GPUs (Graphics Processing Units) are able to speed the computations up even more.

An augmented approach is called Functional Encryption, which operates as a certification service; it certifies specific people in combination with functions or programs that they are allowed to use. As before, a researcher would run the homomorphic computation and receive the encrypted data results. The certificate stamp would let him decrypt those results. Some questions remain, concerning what constitutes a safe computation. An authority might certify the functions, but how can it check to see if the computations are valid and that this particular researcher is authorized to perform them? It will be necessary to determine what kinds of computations are legal and privacy-preserving from both computer science and public policy perspectives. In addition, it is quite challenging to ensure that the computations are fast, efficient, and practical to perform. Further work on hardware, software, and mathematics are required to put the concepts into practice on a broad scale.

"Current Developments in Differential Privacy" -- Salil Vadhan, Harvard University

Differential privacy is an alternative approach to protecting sensitive information about individuals by mediating access to the dataset through an algorithm and an interface that serves as a trusted curator. When a researcher wants to analyze the dataset, he will submit a query to the curator, which will in turn inject some randomization into the answers. The goal is to protect the privacy of the individuals in the dataset while producing accurate answers to the queries. Differential privacy is distinguished from traditional approaches, such as anonymization, in the use of random noise to obscure the effect of each individual's data within the dataset.

The definition of differential privacy considers any two datasets, D and D', where D contains a given individual's data and the data is removed from D' (or is replaced by some unrelated data). We require that, at the end of the process, it is not possible to distinguish D and D'. The indistinguishability is measured through the use of a parameter, epsilon, which measures how close the distributions of results produced under D and D' are; the smaller the epsilon, the greater the level of privacy protection. Differential privacy is compatible with the vast size of big data. As a simple way to express the relationship, given N number of individual data points in a set, if N is a large number, then each data point will have a smaller proportional effect on the computation and it will be easier to hide this effect by injecting randomness overall.

A second point about differential privacy is that is offers a strong guarantee of privacy for all databases, regardless of the type of background information an adversary might have about the original data. This implies that differential privacy is scalable; once the code is written for the interface, then it will be possible to tune the parameter epsilon to the level of privacy desired and there is no need to introduce a privacy expert for each analysis or release of statistical information. There is a substantial amount of work on designing differentially private algorithms for analytical tasks on various types of data sets.[20] One powerful option is synthetic

[20] See histograms [Dwork, McSherry, Nissim, Smith 2006]; contingency tables [Barak, Chaudhuri, Dwork, Kale, McSherry, Talwar 2007, Gupta, Hardt, Roth, Ullman 2011, Thaler, Ullman, Vadhan 2012, Dwork, Nikolov, Talwar 2014]; machine learning [Blum, Dwork, McSherry, Nissim 2005, Kasiviswanathan, Lee, Nissim, Raskhodnikova 2008]; regression & statistical estimation [Chaudhuri, Monteleoni, Sarwate 2011, Smith, "Asymptotically Optimal and Private Statistical Estimation,"2011, Kifer, Smith, Thakurta 2011, Smith, Thakurta 2012, Jain, Thakurta 2013]; clustering [Nissim, Raskhodnikova, Smith 2007]; social network analysis [Hay, Li, Miklau, Jensen 2009, Gupta, Roth, Ullman 2011, Kasiviswanathan, Raskodnikova, Smith, Yaroslavstev 2011, Blocki, Blum, Datta, Sheffet 2013]; approximation algorithms [Gupta, Ligett, McSherry, Roth, Talwar 2010]; singular value decomposition [Hardt, Roth 2012, Hardt, Roth 2013, Kapralov, Talwar 2013, Dwork, Talwar, Thakurta, Zhang 2014]; streaming algorithms [Dwork, Naor, Rothblum, Yekhanin 2010, Dwork, Naor, Pitassi, Rothblum 2010, Mir, Muthukrishnan, Nikolov, Wright 2011]; mechanism design [McSherry, Talwar 2007, Nissim, Smorodinsky, Tennenholz 2010, David Xiao 2011, Nissim, Orlandi,

data generation, where the individual data points have been constructed in a way that retains many of the statistical properties of the original data set, but the synthetic data points do not correspond in a one to one way with any of the original individual data points.[21] The computational complexity is significant, but practical implementations have been successful.[22] Another possibility relates to tasks that are common in statistical inference and machine learning, where the convergence rates are essentially the same as for non-private algorithms as N grows.[23] There have also been practical differentially private algorithms developed and implemented for many specific inference and learning problems of interest, such as logistic regression, support vector machines, and empirical risk minimization.[24]

There are some challenges that arise when bringing differential privacy to practice. One is getting good performance on datasets with a small or moderate number N of observations, but this should be less of a problem as we move more towards big data. It is also is important to model and manage privacy loss over time, especially across different analysts and databases. A further challenge relates to the culture of data analysis, where analysts are accustomed to having access to raw data and will now be working with an interface that deliberately inserts noise into the results. A tiered access model could be constructed to handle this issue, using differential privacy to allow wider access to data than would be possible otherwise, and also providing access to raw data through a process with strict terms of use and a security protocol. This kind of model has been used in the census for many years. Since differential privacy is focused on protecting individual information, it is useful when "global" or population-level computations are of interest; it is not appropriate when the intended use is extracting information about specific individuals. Finally, the current privacy law and regulatory structures tend to frame the issue of privacy in terms of anonymization, including de-identifying data and removing data fields. Since differential privacy is an alternative mechanism for privacy protection, the law and policy framework might have to change to encompass the innovations in this area. A multidisciplinary project at Harvard University is developing tools to use in the Dataverse Network, which is open source software used for sharing, analyzing, citing, and archiving research data in virtual repositories.[25]

Other efforts to bring differential privacy into practice include:

Smorodinsky 2012, Chen, Chong, Kash, Moran, Vadhan 2012, Huang, Kannan 2012, Kearns, Pai, Roth, Ullman 2012]

[20]See Simons Institute Workshop on Big Data & Differential Privacy 12/2013

[21] See, for example results by Blum, Liggett, and Roth 2008 and Hardt, Rothblum 2010.

[22] See Dwork, Naor, Rheingold, Rothblum, Vadhan 2009 and Ullman, Vadhan 2011, Jonathan Ullman, "Answer n^{2+o(1)} queries with differential privacy is hard," 2013; see also Hardt, Ligett, McSherry 2012 and Gaboardi, Gallego Arias, Hsu, Roth, Wu 2014.

[23] See Kasiviswanathan, Lee, Nissim, Raskhodnikova 2008 and Smith 2011.

[24] See optimizations and practical implementations for logistic regression, ERM, LASSO, SVMs in Rubinstein, Bartlett, Huang, Taft 2009, Chaudhuri, Monteleoni, Sarwate 2011, Smith, Thajurta 2013, and Jain,and Thakurta 2014.

[25] See http://privacytools.seas.harvard.edu, which is supported by an NSF Secure and Trustworthy Cyberspace "Frontier" grant and seed funding from Google.

- CMU-Cornell-Penn State - "Integrating Statistical and Computational Approaches to Privacy." See http://onthemap.ces.census.gov/
- UCSD - "Integrating Data for Analysis, Anonymization, and Sharing" (iDash)
- UT Austin - "Airavat: Security & Privacy for MapReduce"
- UPenn - "Putting Differential Privacy to Work"
- Stanford-Berkeley-Microsoft - "Towards Practicing Privacy"
- Duke-NISSS - "Triangle Census Research Network"

"Accountable Systems" -- Daniel Weitzner, MIT CSAIL

In assessing the relationship between law and computer science with regard to privacy, it is important to connect the policy expectations clearly with the system designs, so there are fewer gaps between policy aspirations and functional implementations. One promising area is the development of accountable systems,[26] an approach that draws on standards of practice in the world of financial accounting and applies similar principles in the world of information systems.

A central figure in the history of computer science and privacy is Alan Westin, who laid the foundation for the modern understanding of privacy with his book *Privacy and Freedom* in 1967. Noting that a concrete definition was elusive, Westin stated, "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." Jerry Salzer and Mike Schroeder echoed this definition in an article that appeared in the Communications of the Association Computing Machinery in 1974.[27] In the systems context, a tractable definition for privacy was critical, because it informed design requirements. While the focus on privacy as related to communication of information from one place to another has been useful for many years, it is no longer sufficient for addressing privacy needs in big data environments.

Since both communications networks and data repositories have grown dramatically, it is important to develop a more sophisticated understanding of what privacy is and how it should be protected, recognizing that there are complicated balances between individual rights and benefits, groups rights and benefits, and the risks inherent in data sharing environments. Some uses of personal information are recognized as being good, or at least acceptable, and others are deemed inappropriate, or even threatening. One area that requires much deeper thought encompasses usage restrictions; at the present time it is very difficult to assess how information is being used, and, in some, cases, even to know who is using the information. As an example, in the summer of

[26] 2nd International Workshop on Accountability: Science, Technology and Policy see http://dig.csail.mit.edu/2014/AccountableSystems2014/

[27] Weitzner, Daniel J., et al. "Information accountability." Communications of the ACM 51.6 (2008): 82-87.

2013, the chief judge of the Foreign Intelligence Surveillance Court asserted, "…the court lacks the tools to independently verify how often the government's surveillance breaks the court's rules that aim to protect Americans' privacy."[28] This statement could apply to both administrative and technical tools and underscores the complexity of privacy protection in the modern age.

There are at least three definitions of privacy that address the issues at different levels. First there is a notion that privacy is a simple binary choice; people either have access to an individual's data or they don't. At a second level, privacy can be viewed in the context of data sets and whether an individual's data is exposed against the backdrop of the set as a whole. Finally, there is an activity-based view, where privacy is seen in relation to access to data and permission for analyses that may be performed on it. This framework moves beyond addressing the basic question of whether data is exposed or not and leads to insights on how to develop accountable systems for information access and data protection. The goal of these systems is to determine how information is being used and to pinpoint inappropriate usage.[29]

The research on accountable systems entails expressing laws in a computable policy language and logging certain types of transactions into the system. The system analyzes the application of the law as expressed in the policy language and provides a proof that a transaction is or is not compliant. Each policy is represented as rules and patterns in a policy file and the definitions and classifications are represented in an ontology file. One challenge in this work is to ensure that the policy language is sufficiently expressive to handle legal rules, with some degree of nuance. If the language becomes too expressive, then issues with computational efficiency may arise.

It is important to provide for the evaluation of usage post-collection and analysis. Further, it is useful to generate clear explanations on compliance, as they will help organizations and individuals come to the right decisions. A related challenge involves threat models where there may be malicious insiders present. Finally, it is essential to support incompleteness and inconsistency within the analytical framework. In some cases, it will not be possible to obtain a simple answer to a legal question, so the system may be able to provide helpful guidance, but will not be able to make the final determination itself; further human intervention and interpretation will be necessary.

This work is in progress at MIT and work on developing policy languages and reasoners is underway at other places, including CMU. Additional research includes inferring purposes and other properties through statistical methods,[30] creating a formal definition of accountable systems to assess the reliability of systems,[31] and conducting

[28] Statement by Judge Reggie B. Walton, in *The Washington Post*, August 15, 2013.
[29] "Information Accountability," Weitzner, D.J., Abelson, H., Berners-Lee, T., et al. Communications of the ACM (June 2008), 82-87.
[30] Tschantz, Michael Carl, Anupam Datta, and Jeannette M. Wing. "Formalizing and enforcing purpose restrictions in privacy policies." *Security and Privacy (SP), 2012 IEEE Symposium* in IEEE, 2012.
[31] Feigenbaum, Joan, Aaron D. Jaggard, and Rebecca N. Wright. "Towards a formal model of accountability." *Proceedings of the 2011 workshop on new security paradigms*. ACM, 2011.

research on operating system and hardware-level architectures, where some of the accountability and enforcement could be grounded at the silicon level.[32] One goal in the work on accountable systems is to support and enhance the level of public trust. As with the model of a general ledger feeding into a financial balance sheet, personal information transactions could feed into a personal information balance sheet. In this model, developing clear and accountable compliance systems will help to address some of the challenges facing privacy in the big data context.

Panel 3: Roundtable Discussion of Large-Scale Analytics Case Study

Carol Rose, Executive Director, American Civil Liberties Union Foundation of Massachusetts
John DeLong, Director of Compliance, National Security Agency
Mark Gorenberg, Founding Managing Member, Zetta Venture Partners
David Hoffman, Director of Security Policy and Global Privacy Officer, Intel
Karen Kornbluh, Executive Vice President, Nielsen
Andy Palmer, Founder, KOA Lab
Mona Vernon, Senior Director of Emerging Technologies, Thomson Reuters
Latanya Sweeney, Professor, Harvard University
Vinod Vaikuntanathan, Assistant Professor, MIT CSAIL

Panel Moderator: Daniel Weitzner, Principal Research Scientist, MIT CSAIL

This panel discussed big data privacy concerns based on a hypothetical case study centered at MIT, where there is an increasing interest in the "quantified campus," and creating a more data-driven environment on campus.[33] The premise of this case study is that in 2015, MIT decides to embrace the potential of data-powered, analytics-driven systems in all aspects of campus life, from education to health care to community sustainability.

The panel discussion explored privacy issues as the decision unfolds through five phases: 1) MIT departments require students use the edX MOOC platform for courses, and researchers begin to look at correlations and patterns in the student data; 2) MIT provides everyone on campus with a self-tracking device; 3) MIT provides the local public transportation system real time student course schedules so they can better meet demand, and government agencies request access to student data so that they can monitor for students at risk of being recruited into terrorist activities; 4) data from 50-100 top universities is aggregated by a new single commercial venture, creating opportunities for new research and services; 5) MIT approaches École Normale Supérieure(ENS) in Paris about creating an educational alliance (*Note: All of these scenarios are hypothetical and bear no relationship to*

[32] Shrobe, Howard, Thomas Knight, and Andre de Hon, "TIARA: Trust Management, Intrusion-tolerance, Accountability, and Reconstitution Architecture." (2007).
[33] See http://web.mit.edu/bigdata-priv/CaseStudy-WH-MITBigDataPrivacyWorkshop.pdf

actual MIT or EdX plans.)

Discussions focused on the major privacy concerns in each situation and how MIT can fulfill the twin goals of making data available for research purposes and taking maximum advantage of large scale analytics to improve campus life and educational quality, all while ensuring appropriate privacy protection for the students and the community. Included below is a summary of each of the panelists concluding remarks following the scenario discussion. See Appendix B for a transcript of the full roundtable discussion.

Panelists concluding remarks:

VINOD VAIKUNTANATHAN (MIT): My point is short, we have, in the computer science, privacy, and cryptography communities, many tools that enable privacy. Some that are more expensive than the others and some have little cost at all. When you make policy about big data and privacy issues, you have to be aware of what technologies are available, so that policy can be informed by technology. Some of these technologies are surprising; there are things that you wouldn't even imagine you can do. Better policies will be made by being aware of the remarkable technology that's out there.

LATANYA SWEENEY (Harvard): We have to enable these technologies to grow and give us an easier, better way of sharing data. In order to do this, we have to be more transparent in showing actual harms. This means being willing to let experiments, vulnerability assessments, and transparent data sharing happen, so that we can make the case for what's really the right value proposition.

MONA VERNON (Thomson Reuters): Operating and innovating on a global footprint requires us to think in that lens.

ANDY PALMER (KOA Lab): The long pole in the tent here is the understanding of use, holding people accountable, and determining whether there's harm or not.

KAREN KORNBLUH (Nielson): Somebody said earlier that through math we can find a solution to both privacy and innovation; this is a great way to think about how we can have all these things together. We have to think about these things on a global scale and ensure that we develop interoperable solutions.

DAVID HOFFMAN (Intel): Paraphrasing Reed Hunt, distrust is the cancer that may kill the digital economy. This means that we need an evolution, not a revolution; the FIPPs are enduring, we should look to them. But we need to create comprehensive U.S. privacy legislation that could become a model for the rest of the world, encouraging economic growth, while still protecting individuals.

MARK GORENBERG (Zetta Venture Partners): This is the greatest opportunity for economic growth and societal good since computers were taught to be

programmed. Given that, the collection of this data will not slow down. It's up to our technologists to figure out solutions to move this forward in balance with regard to privacy and security systems and it's up to our policy people to figure out ideas as they did for digital rights management that helped the music industry evolve in this new world; that'll be our challenge.

JOHN DE LONG (NSA): Big data, big rules, big compliance. In addition to applying the technology solutions in the big data context, we need to increasingly think about the rules themselves. This will help to lash up policy, the intent of policy, and the actual words in policy with the actions that are occurring every day because of the way the technology and the policy works.

CAROL ROSE (ACLU): Technology alone will not solve privacy problems; the law needs to keep pace. Metadata is personal information data; it is just like content data and we need to have privacy protections and government transparency, so that we don't continue to have the lack of trust that has been caused by the recent revelations, the secrecy, and cover-up that has arisen in conjunction with the NSA spying. I would encourage everybody to take a look at the open letter from researchers in cryptography and information security that was released on January 24, 2014 that sets out what researchers, business, and civil liberties advocates believe are important objectives if we're going to address these problems.

DANIEL WEITZNER: If you think about all of the challenges that surfaced in the case study of what a little place like MIT ought to do when faced with an interesting collection of data. The challenges facing the U.S. Administration, other governments, and all of us who are interested in seeing progress, centers on understanding how we engender a sense of progress. We're not going to have a light bulb moment, where we answer the privacy questions by passing a single law or inventing a single brilliant algorithm. We'll need to step through these issues guided by important principles like transparency and accountability. This will help to create a path for progress.

References

John Podesta's talk:
- data.gov platform - https://www.data.gov/
- Open Government Initiative - http://www.whitehouse.gov/open
- Fair Information Practices Principles - http://www.nist.gov/nstic/NSTIC-FIPPs.pdf
- Privacy Act of 1974 - http://www.justice.gov/opcl/privstat.htm
- Consumer Privacy Blueprint - http://www.commerce.gov/os/ogc/developments/administration-releases-blueprint-consumer-privacy-global-digital-economy
- Consumer Privacy Bill of Rights - http://www.whitehouse.gov/sites/default/files/privacy-final.pdf

Daniel Weitzner's talk:
- Weitzner, Abelson, Berners-Lee, Feigenbaum, Hendler, Sussman, Information Accountability, *Communications of the ACM, Jun. 2008, 82-87.*
- Weitzner, Needles in Haystacks: Creating Information Balance Sheets for Personal Data, Remarks before the United States Privacy and Civil Liberties Oversight Board Workshop Regarding Surveillance Programs Operated Pursuant to Section 215 of the USA Patriot Act and Section 702 of Foreign Intelligence Surveillance Act, July 9, 2013
- C. Hanson, L. Kagal, D. Weitzner, Integrated Policy Explanations via Dependency Tracking (IEEE Policy 2008)
- Pato et all, Aintno: Demonstration of Information Accountability on the Web, Third IEEE International Conference on Information Privacy, Security, Risk and Trust (PASSAT 2011)
- Khandelwal, J. Bao, L. Kagal, I Jacobi, L. Ding, J. Hendler, Analyzing the AIR Language: A Semantic Web (Production) Rule Language 2010
- Waterman and Wang, Prototyping Fusion Center Information Sharing; Implementing Policy Reasoning Over Cross-Jurisdictional Data Transactions Occurring in a Decentralized Environment, IEEE Conference on Homeland Security Technologies (IEEE HST 2010)
- Senevirante, Augmenting the Web with Accountability, World Wide Web Conference 2012 PhD Symposium, April 2012
- Second International Workshop on Accountable Systems: Science, Technology and Policy. http://dig.csail.mit.edu/2014/AccountableSystems2014/
- For more: http://dig.csail.mit.edu/

Work supported by National Science Foundation grant CNS-0831442 CT-M: Theory and Practice of Accountable Systems, IARPA Policy Assurance for Private Information Retrieval grant FA8750-07-2-0031, and the Department of Homeland Security Accountable Information Systems grant N66001-12-C-0082.

Appendix A – Q&A Session with John Podesta, White House Counselor

Adapted from transcript

DANIEL WEITZNER: John, thanks very much. While the audience is digesting your very comprehensive talk, allow me to ask one question: what do you think are the most important questions we can be working on here to contribute to the policy debate?

JOHN PODESTA: I tried to kick out a few of the things worth thinking about towards the end, but I think that particularly our mind is given the ubiquitousness of collection now, the "Internet of Things," the ability to collect geolocation data, the ability to mine those data patterns, these profound changes in the way we collect and store data, and then the analytics that work on top of them, what does that mean for the existing policy framework? Danny, you worked on the policy framework that was released in 2012 when you were working here in the White House. So, those, I think, are the big issues in our mind: What are the technological trends? Do they require us to kind of update or rethink our policy proposals, and then finally, what is the real change in going from, and this is as much of a question for our government as it is in the private sector, predicated searches, searches where you at least know, even if the person's name is not known to you, what you're going after or looking for a search involving an individual, to ones that are non-predicated, where you're looking for a pattern. Does that change the way we need to be thinking about the government operates with regard to its handling of data with regard to academic institutions, and of course, with regard to the private sector?

Q: I work here at MIT in the Industrial Liaison Program. How is this initiative over the next 90 days going to inform the NSA on how to improve its business practices?

JOHN PODESTA: Well, I think that's a good question, and I think that as I mentioned at the outset of my remarks, the President did a six-month review of the issues that arose from the NSA and from the leaks that were the result of Edward Snowden's activities. In that review that he summed up in his speech on January 17th, he ordered a certain number of initiatives that will go forward. This is one of them, but he also ordered the government to think about and to change its practices with regard to certain surveillance practices, particularly with regard to the collection of metadata, the so-called Section 215 program. The review of that is ongoing, led by the Justice Department, and the intelligence community, and some options to change the way the use of that program to try to maintain and capture the ability to get the upside of being able to find suspects involved in terrorism, as we saw in the Boston Marathon bombing case, can't take place without the government holding that metadata. So, that review is underway in a somewhat separate track and hopefully our larger perspective, looking at this beyond just the intelligence community and its authorities into the broader picture of the technology review, particularly in consultation with PCAST may help inform intelligence policy going forward, but

really I think that these are parallel tracks and we are proceeding as the President has ordered.

Q: I'm the National Chair for the Restore the Fourth Coalition and I am interested in your thoughts not simply related to the NSA and how it deals with surveillance data, but also with the movement from a predicate-based model of analytics to a non-predicated-based model of analytics, as it relates to the broader mission of law enforcement. If the 4th Amendment inherently involves having a predicate and having probable cause before the government investigates you and makes the decision as to whether to detain you or not, then isn't there a fundamental hostility between big data and the Constitution? I'd welcome your comments on that.

JOHN PODESTA: Well, that's an excellent question, and I highlighted it at the end to suggest that that's something that we really want to engage both with 4th Amendment experts, with the law enforcement community, and with the public at large to think through. It's a challenging question. Back in the 1980s, I was the Chief Counsel to Senator Leahy when he drafted the Electronic Communications Privacy Act. There is considerable attention on the Electronic Communications Privacy Act today. Legislation to update it is pending before the House and the Senate. The administration is engaged in developing its own policy with respect to reacting to that legislation that's been put forward by the Congress. I think there's no question that the issue that you pose will be presented in that context and I have to say that I don't have a full answer to your question, because that's really what, I think, this study is trying to accomplish, which is to take on board the effects of new technology, and the questions about whether there is something new and different with respect to the jurisprudence around the 4th Amendment that has developed over the years, and that challenges, particularly some of the decisions that have been rendered by the Supreme Court with respect to third party holding of data, etc. So, we're looking at that and hopefully in 90 days we'll have something more specific to say about what our conclusions are with respect to the issue, but I think you're raising a very, very important point.

Q: So, my question, perhaps is more appropriate for the other two workshops, but I'm very interested in the question of the balance of the roles between government and big business. Right now we see the large companies – Google, Amazon, you name it – who have enormous amounts of both technology and capability, and for whom the collection, management, and use of all this data is central to their business. On the other hand, we see the government stepping up as perhaps the first customer in much of that work. The balance in the United States is very different from, perhaps, in some of the other countries, certainly European countries, where the role of government and regulation has much more of an effect, perhaps, on the behaviors of companies. So, how do we find our way through all of this space that is not just about technology and capability, not just about the legal system, even, but also integrating the question of how do our companies continue to be able to do the business that they do under the conditions of privacy and identity management?

JOHN PODESTA: Another great question. I met with the German Foreign Minister last Friday to discuss these topics and obviously the Europeans, and particularly the Germans, have some different views on this topic. I think to date our privacy regime has been effective with respect to commercial data – at least it's pointed the way. I mentioned the FIPPs earlier in terms of trying to provide citizens with rights, and to ensure that people can understand; they have notice of what the data practices of large companies have been here. We have, I think, at least as strong enforcement mechanisms for individuals as the Europeans have, but they do have somewhat different conceptualization of the set of rules. I would say that you argued in favor of balance, and I think that's appropriate, because there's no question, and I tried to lay out a few examples in education and healthcare, where the capacity of big data has an enormous contribution to make to positive social outcomes. So, what we need to do is to try to figure out whether that's by technological means and that's some of the work that's going on at MIT, which looks to try to build better privacy protection into the data structures themselves, or whether it's by clear and better data practices by the private sector. Generally the way the administration has approached that is through a multi-stakeholder process, where codes of conduct can be developed that are strong and give the public a right and a sense that the data that's being collected about them will be used appropriately, or whether it's by further regulation, I think that we need to get the balance right, so we can get the upside with respect to the capabilities, the innovation, the practical and positive effects that this can have on people's lives, but in doing that to protect their personal freedom and their privacy. So, that's what we're searching for is that right balance, and I think this day is set up to explore those questions in-depth.

Q: Robert Ellis Smith with Privacy Journal, John. It's good to hear from you. And you may have answered this, but is there anything to be learned from Europe and Canada on this whole question? I'm not referring just to the question of regulation, but innovative uses of technologies to solve some of these problems?

JOHN PODESTA: Well, we're going to sit down with a broader range of our international partners in the weeks ahead, and we, of course, remain open to listening to their views. But I think the one thing we should all want to avoid is the balkanization of the Internet. I think that would be a mistake with respect to both the capacity for innovation, but also for the free flow of information and ideas.

So I think what we don't want to do is set up different technological regimes that really balkanize and create high walls for movement of data across national boundaries. We'll be looking to hear from them as to their current thoughts, but I remain convinced that we can find a way in which we can accommodate each other's perspective. Right now, the U.S. and EU are in consultation about updating the safe harbor provisions that are a feature of the way the EU is treated U.S. companies doing business in the EU. And I think those discussion have been positive. So, we just need to find a way that is appropriate for the thinking of both cultures.

My bottom line is, we want to ensure that we keep an open and free flowing Internet. So, that's an assumption, presumption, or a bias that I have going in. But given that, we'll listen to our European colleagues and colleagues from the rest of the world for their thoughts on that.

Appendix B

Panel 3: Roundtable Discussion of Large-Scale Analytics Case Study

Adapted Transcript

Carol Rose, Executive Director, American Civil Liberties Union Foundation of Massachusetts
John DeLong, Director of Compliance, National Security Agency
Mark Gorenberg, Founding Managing Member, Zetta Venture Partners
David Hoffman, Director of Security Policy and Global Privacy Officer, Intel
Karen Kornbluh, Executive Vice President, Nielsen
Andy Palmer, Founder, KOA Lab
Mona Vernon, Senior Director of Emerging Technologies, Thomson Reuters
Latanya Sweeney, Professor, Harvard University
Vinod Vaikuntanathan, Assistant Professor, MIT CSAIL

Panel Moderator: Daniel Weitzner, Principal Research Scientist, MIT CSAIL

DANIEL WEITZNER: In this last panel we will spend a little over an hour discussing a case study that presents some novel uses of personal information and novel data integration techniques.[34] We have slides that summarize the case study as we go along. We chose MIT and the educational establishment for this exercise. This case study centers on privacy issues that are likely to arise as universities move more and more towards an online learning environment and as people want to ask and answer questions about what's going on in that environment. At MIT and on campuses all around the world, there's an increasing interest in what might be called the "quantified campus", a more data-driven campus environment. For researchers, these are great laboratories for experimenting with new uses of information in all kinds of ways. However, as these experimentations are taking place, we are also actively working on questions about what the right privacy approaches to the use of this personal data should be.

I should say just as a disclaimer, I didn't check with Anant Agarwal or Ike Chuang at MIT, so this is a purely hypothetical version of MIT that we're discussing, not MIT itself. As you heard from Anant Agarwal, there's certainly enormous excitement around the world about moving to online educational platforms. The current instantiation of that involves MOOCs, which tend to be outwardly focusing, that is to say, they support education for students who are not on campus. My own experience in teaching here at MIT is that the local educational experience is moving online as well. It's not perfectly synchronized with the edX platform, but it may happen over time, mostly because there is enormous learning potential to connect

[34] The full text of the case study is available here: http://web.mit.edu/bigdata-priv/CaseStudy-WH-MITBigDataPrivacyWorkshop.pdf

the rest of the world to campuses, so that students on campus can learn from students and others off campus.

The premise of this case study is that there is a convergence between the MOOC platform and other online educational tools that support the local learning environment.

In this case study, our presumption is that four academic departments have decreed that students have to use some iteration of the future edX platform for all of their courses. What that means is that anyone who can see the data on that MOOC platform can learn quite a bit about the details of students' educational activities. Anant commented on the scale of the click stream data. In this scenario, we want to look at what the academic departments think they can learn from having all this data. We also want to address some concerns on the part of students, who may wonder about whether people who have access to this data can learn who the students are working with, where they're studying, when they work, and when they sleep. In this scenario, there is not yet a policy in place about how the access to this data is logged. So there's an open environment in which there will be certain access to the data. In keeping with the quantified campus environment, additional information from campus will be combined with the MOOC data. For example, campus health services data will be commingled, as will student food purchases through the MIT dining plan. This relates to an interest in the correlations between health status in general and educational achievement.

MIT has a firm commitment, both in this case study and in actuality, to the belief that this data should be held in the public trust; it should be available to researchers to evaluate various aspects of the educational experience. So consistent with the open nature of the edX platform, MIT's desire is to make this data as open as possible for researchers on the campus and around the world.

So, for the panelists, based on this part of the scenario, could you reflect on what we learned in the last panel? What do you think MIT's practical options are today and what might be some of the aspirational hopes for the future on how to fulfill these twin goals of making this data available for research purposes, but also creating an environment where there is a sense of privacy protection still at work here?

VINOD VAIKUNTANATHAN: One of the problems in working with data and worrying about privacy is the unexpected ways in which the data can be used once it's released into the world. You post something on Facebook and you are not really thinking about where this data might end up. So at this moment, depending on how concerned people are about privacy, a natural and somewhat immediate solution is to give access to all of this data via an interface, which can regulate the ways in which this data is accessed. As you heard before in the differential privacy talks, there are two main approaches to releasing aggregate data. One is to say that you have an entity called a curator, which receives questions about what aggregates people want to know, checks to see if this is acceptable or not, may add some noise,

and then will release this information. The other approach is to create what's called a synthetic database. So you take all of this data, aggregate it, and release it in a form that may not look like the data itself, but exists in a way where you can ask a large class of questions and obtain answers. Clearly it does take more infrastructure to maintain a curator and ensure that privacy policies are met, but this is one approach, from the very beginning.

One unexpected thing that can happen concerns the issue of composition. You ask one question about the data, then you ask a second question, and a third question; each of these sounds innocuous by itself, but put together, these can be constructed as the type of linkage attacks that Latanya discovered and that differential privacy tries to model. These are surprising attacks that one should be really careful about.

So that's the privacy answer to this question. What differential privacy and associated technologies do is to try to answer the question of what computations are legal, what are safe computations that can be performed on this data? Then there's the cryptography part, which handles the question of once you decide on the safe computations, how can you do these computations safely and how can you store encrypted data in a server, for example. And still be able to do all these computations? That's the crypto angle of things. These are complex questions that differential privacy is well-positioned to answer, but they also involve policy questions.

DANIEL WEITZNER: So do we find that this entails tradeoffs; are we trading off some amount of knowledge about what's actually going on in these environments in exchange for the kind of differential privacy commitments you want MIT to make here?

VINOD VAIKUNTANATHAN: I think the technology in differential privacy is exciting, but the major concerns that I see are in the computational complexity. First, if you want to support more and more complex computations, you have to spend computational effort. The question is, "How large is that?" Second, there may be a tradeoff in the accuracy of answers that you release; if you want privacy, you must give something up and that turns out to be accuracy in the aggregate data. This brings us to the question of the "epsilon," which was covered in Cynthia Dwork's talk.

ANDY PALMER: I do not understand this; the obfuscation of the data itself in the interest of protecting people bothers me a lot. Aren't we just increasing the opportunity for bad actors to take bad actions? If you're really technical, if you're a hacker, if you have the wrong things at heart and now there is this second way to go in and influence or change the data, isn't this just another point of failure, rather than a way to provide more transparency?

VINOD VAIKUNTANATHAN: So your point is, "Where is this data stored and how do I regulate access to it?"

ANDY PALMER: So you suggested the presence of a curator and then there is probably a person who is responsible for making sure that all of these encryptions work the way that they're supposed to work and that they are statistically significant and aligned with the intent, or the sort of provisional data. Isn't every one of those people a potentially bad actor who could do something wrong? These may also be situations where there are technical problems in organizations that are already under-resourced, that already don't have enough compute power, storage, and people.

DANIEL WEITZNER: So, David Hoffman, you've left your job as the Intel Chief Privacy Officer because you took care of all their privacy problems, you're now the Chief Privacy Officer at MIT in our study. How would you try to guide MIT in thinking through this challenge?

DAVID HOFFMAN: If I could rephrase the question, is it that, "As we are trying to secure the data better, are we driving up the risk, or are we driving down the risk?" I think to a large degree that's a technical question, so if we're talking about using cryptography in new ways, people can talk about how strong and well-tested the algorithms are and how long they have been subjected to peer review. Those are all really important questions. What I would be wary of would be to say that, "If we're going to collect more data, we may have increased the risk." I don't think that that's necessarily true and it runs counter to the concepts of innovation and creativity that should be driving this conversation. We should have a deeper discussion about the particular security safeguards that we want to put in place and how confident we feel about them. We may be improving the situation for privacy here.

MARK GORENBERG: I would agree with David that it's not about collection. There's a great article that came out in Foreign Affairs this month by Craig Monday, who's also on PCAS and was the Chief Research Officer at Microsoft, called, "Privacy Pragmatism." It promotes the idea of focusing on use and not on collection. The problem of trying to hold back collection is that horse is already out of the barn; given all of the data that has already been collected, consents are just almost impossible, because there are hundreds of sources. On top of that, there is machine-generated data and how are you going to obtain consent for that? Passive data is being collected all the time, how are you going to get consents about that? So the focus should be much more about how it's used when you're talking about who uses it, and when it's used.

This morning, Mike Stonebraker talked about audit systems and Danny discussed accountable systems. Then the question becomes, "How does that propagate?" One interesting nuance is that in the industry context, the question is associated with the fact that they have granular data and come out with results that could help their customers, but they are not just giving away all of the data. I sat on the board of a company called Omniture that was the leader in the web analytics. Anant talks about quarter of a billion of click streams for the course on the edX platform; by

2009 when Omniture was acquired by Adobe, they were doing over a trillion click streams a quarter. That was leading to 20 to 30% better results for marketers. You keep getting better results as you have more data to mine. So you're not going to be able to hold that back.

DAVID HOFFMAN: I agree with that; I don't think that we're moving away from collection. There still has to be an analysis of whether and when the collection is going to be inappropriate. We need to recognize that a situation is evolving, where a much greater percentage of the data that relates to us as individuals is not coming directly from us in a capacity where we could provide the kind of protections we want for individuals by looking primarily at collection and not looking at use, accountability, security safeguards, and other mechanisms.

MARK GORENBERG: The response to that is really transparency. If you say transparently that the data is going to be made available, a lot of people in the community will be up in arms about it. So if you're transparent about what data you're collecting, if you're transparent about who gets access to that data. There is a professor at NYU now, whose opinion is that part of the solution is to make sure the individuals get access to all of the detailed data, because then at least they know what's out there about them, even if they're not going to be able to solve the collection problem.

ANDY PALMER: There is another pragmatic problem when you build systems, which is that you will work on collection, and maybe encryption, and then at some point way down the road, you get to the question of how you are going to use that data. By the time you get to that question, you've usually run out of money, right? So the questions of use are almost always radically underfunded. They're almost radically underdeveloped, because all of the resources get consumed with all the work that's done around the collection and the aggregation of this data. This should be turned on its head; we need to focus on use first.

DANIEL WEITZNER: Here is another hypothetical. The FBI and the Secret Service are investigating a cyber-attack at a bank that happens to be based in Boston, but has global reach. I don't actually have one in mind. The investigators end up with big chunks of code that they think are responsible in some way in the systems for a whole set of insider attacks. They're completely stumped and had to figure out how it got there and who's responsible for it. They've learned that this bank employs a lot of MIT students and they decide, having exhausted all other options, that they're going to try to match code signatures of the malicious code with code developed by Course 6 students at MIT. So MIT is served with a subpoena that says they want all the code submitted for all of the problem sets in Course 6 over the last five years, because that's a reasonable amount of time to look. It turns out that they identify a few students through this code signature matching technique and they identify more individuals by linking some social network data they have obtained from the discussion forums with public social networks. So they catch a group of people and convict them, they are now in jail. Carol, did MIT do the right thing?

CAROL ROSE: Well, the first thing I would tell them to do would be to read Professor Abelson's report about data.

DANIEL WEITZNER: And what would they learn when they read that? Who should read it and what would they learn?

CAROL ROSE: All of us should read it. What's really important here touches on both use and transparency and centers on the fact that this kind of knowledge gives you a lot of power and who the "you" is influences who has the power, whether it's MIT or the FBI. There's a huge amount of power that is associated with that and yet our laws still have a distinction between metadata and content data. What we've learned here today is that that distinction is rapidly going away; big data can tell you a lot about yourself, more than you might voluntarily provide or even know. So metadata needs to have much more privacy protection than they were given when we last updated the law in 1986. This is one thing that I hope the White House will take to heart: we need to give the same privacy protections to metadata that we do to content data now. The second issue is the problem of secrecy that exists, especially when third parties hold the data. The FBI or district attorneys here in Massachusetts can go and use either national security letters or the state equivalent of a state administrative subpoena to obtain data about you and never let you know.

DANIEL WEITZNER: So you're saying the students should have been notified?

CAROL ROSE: The students should have been notified and the students should not have been required to buy into the database in order to get education in exchange. That was a coercive move by MIT in this hypothetical.

DANIEL WEITZNER: Well it wasn't coercive in that most of these students knew that MIT was going to do this when they signed up. The case study notes that the fact that the education has migrated to this platform and has been subject to rigorous analytics means that we've tuned our courses and our problem sets so that everyone is doing better. This was an educational benefit.

CAROL ROSE: So if I want privacy, I have to go to Harvard?

DANIEL WEITZNER: In this situation, the FBI didn't know where they were going to look except that they thought there was a reason to suspect MIT students. So they needed the data. Should MIT have said, "No, you can't have all that data."?

CAROL ROSE: MIT could've said, "We need to tell the students that you're getting the data, so that they know that this is coming and they can correct it if it's wrong."

DANIEL WEITZNER: So it would've been okay to give the data, provided students had notice? Could the students object?

CAROL ROSE: I think students should have the standing in court.

DANIEL WEITZNER: If you were representing a student, what would be the basis of the student's objection?

CAROL ROSE: I would go into cautious subpoena, but the bottom-line is that right now under the current law, MIT would own the data, not the student. That's the problem.

DANIEL WEITZNER: So John, take your NSA hat off for the moment. How do you think about this from a compliance perspective?

JOHN DELONG: Andy's comments on use and resources capture a lot of it. If I were to distill one important thought from the workshop, it is that big data requires big rules and big compliance and you have to think about all three of these things in lockstep. You can't start with one and add the other two, or start with the middle and add the other ones. We tend to focus on the data side, I think that's natural. However, big rules and big compliance are really important and you have to think about them up front. One of the slides earlier showed that having no policy is like the kryptonite of compliance. You can't do compliance without a policy, so it's very hard to create systems. The FBI and the NSA both operate under attorney general-approved procedures. So we know in advance, for example, if we're going to bring in data differently between the FBI and the NSA, so we have to build a compliance design in for that from the beginning.

Referring back to Cynthia Dwork's talk, she had a stone tablet on her slides. This is a great way to explain big rules and big compliance; you have the law and the policy, which is like a version of software code. It has if-then statements, and go tos, and global variables; you have the actual technical code, the software which again is an expression of the law and policy. Then you have the internal policies and all three of those things have to be in lockstep. It's like the Rosetta Stone - the same expression, same meaning in three different languages. We need to think about applying big data techniques to the rules themselves. Danny Weitzner's slides picked up on that – looking at a kind of intermediate policy language, which is tied to the data and helps keep those three things together.

DANIEL WEITZNER: Faced with a request like this for a lot of data on somewhat speculative grounds, it certainly seems plausible that no one having anything to do with the cyber-attack would have ever been near an MIT course. But faced with that kind of request, assuming MIT was not persuaded by Professor Abelson's report, or by Carol's interpretation of the report, is there a middle ground, more privacy-preserving response that doesn't entail handing everything over to the FBI?

VINOD VAIKUNTANATHAN: Yes, in fact there is. This is something that Shafi Goldwasser talked about: the notion of secure computation that has been around since the 1980s. The idea is that I have this big piece of data and you want to know

something about it. So do I ship the entire data to you? This seems bad for several reasons. You could use that as the authority to learn what you are looking for and you could potentially discover other things about the individual. A better way to do this from a cryptographic point of view, is to run a crypto algorithm and, at the end, the authority will get what he is looking for, which is this one bit saying, "Is this code malicious? Was this code generated by these people, or not?" That's all you need to learn and that's all you should learn.

DANIEL WEITZNER: Interesting points. Any questions on the case study as far as we've gone, Eric?

ERIC LANDER (Audience): I'm sort of troubled by the analysis if I understand it now, which is that we've got all this massive source code and maybe there's a guilty person there and maybe there's not. There's somebody within the database of MIT students who writes code that looks similar to code written by the guilty person. It sounds as though you're concerned about the extra uses with regard to all of the innocent parties, but you're not particularly concerned about the use with regard to the guilty party. Maybe we should get access not just to MIT's Course 6 data, but also to the private diaries on all of the laptops; I'm bothered by the idea that law enforcement can have blanket access to all of this information.

ANDY PALMER: I really think the conversation has to shift more to use, because if we look for fancy ways to encode this data, irrespective, assuming away the use and the intention of the people that are accessing the data, I think we're making a radically unfair assumption. All that matters in guiding us in what we collect and how we store it is, "Who's going to use it and what are they going to use it for?" This helps to identify bad actors and also speaks to the allocation of resources. You can't collect all the data in the world, as much as some of us would like to try and do that. So how do you determine which data is important to collect? Use should drive the resource prioritization decisions as well.

MONA VERNON: So one of the things that I keep thinking about is what John Podesta said this morning, which is that we're moving from predicate search to pattern-based data mining, looking for things we don't know. In that framework I'm not sure how we can figure out the use first. Isn't the whole point of big data to identify information that we're not already aware of - we're looking for new patterns. So that makes defining the potential uses quite difficult, doesn't it?

CAROL ROSE: That's pretty dangerous in a law enforcement context, because our whole system has been built upon the principle that if someone commits a crime, then you hold them responsible. You don't decide if someone has the potential to commit a crime. You can draw conclusions and analogies that may be very incorrect and we have found that time and time again that's precisely what happens in the law enforcement context. There was a gentleman on an airplane, for example, who was looking at maps of airplanes. He was detained and held for many hours. It turns out he builds model airplanes. Part of the problem is the data, but it matters who

interprets it and how they interpret it. This goes back to the questions of use and transparency. There needs to be a probable cause standard for metadata, the same way there is for content data.

MARK GORENBERG: The other part of this question is retention and there are two sides to that discussion. The positive side of keeping lots of data is that sometimes benefits emerge from looking at the data that were not apparent several years ago. In one example, Kaiser Permanente went back and looked at certain types of data and was able to correlate autism with drugs that women had taken during pregnancy. They would never thought about that at the time when they were collecting the data. So that's a positive reason why you'd want to keep data for many years. But by the same token you can make the argument that data gathered five years ago may not be relevant for this hypothetical situation at MIT. Therefore, should MIT consider a retention policy where they don't keep the detailed data for more than a year; in this case, you could only look at current data to make these choices.

DANIEL WEITZNER: Here is another hypothetical: the FBI controversy has passed and the students were ultimately exonerated on a technicality. Since MIT has been so successful in creating this highly instrumented environment, a company first called FitBit, later called Fitbyte in the big data sense, is interested in participating in this environment. They would like to explore what will happen when they combine all of their data with other things that are happening on campus. They give every person on campus, students, faculty members, and staff a free device and are even willing to replace devices that are broken or lost. MIT is enthusiastic about learning more and the MIT Health Plan even says that if you wear your FitBit, share your data with the health plan, and are willing to receive periodic healthy lifestyle reminders based on your activity, then you will get a 10% discount on your health insurance rates.

Everyone goes along with this; instructors start to use this data to send reminders to students who seem to be sluggish physically and underperforming in their classes, suggesting either additional exercise, or additional study on the theory that one will lead to the other. The MIT medical staff are figuring out ways to get their patients to have healthier lives. So, Latanya, what should MIT be thinking about in putting together all of the data with this very fine grained activity information, which, by the way, also has location information. What are some of the risks that MIT ought to be thinking about in agreeing to this partnership?

LATANYA SWEENEY: I have to provide a disclaimer in that whatever I say does not represent the FTC or any of its commissioners. On the question here, there is a lot that could be learned, in terms of the promises of the future of privacy enhancing technologies as articulated through the lens of differential privacy and through the lens of homomorphic functions, in terms of some of the problems. So when the data are isolated in one repository, then there might be a future promise of solving some of the computational issues. We also have to realize that FERPA already controls

some of the MOOCs data, if not all of it and that HIPAA might apply to some of the medical data at a first glance. What's interesting is, it actually would not apply to this FitByte data.

One of the things that happens when you make data more transparent is that you'll get a good sense of where to look for harms. Where are the problem areas? Then you can really understand whether the remedies that you're seeking will be helpful against the harms and how they will correct for them? This graph from the datamap.org is basically a plot from FOIA requests and breach notices, all publicly available data of data sharing. If you click on the node, it shows particular companies and so forth. What's really interesting is that everything that you see in bold is not covered by HIPAA. This is all medical data, it's all personal information about your healthcare and only about half of the edges that we're able to document are covered by HIPAA. The other data are not covered by HIPAA and are often released in ways that are far more identifiable.

So, independent of what MIT thinks about it's decisions with respect to the health data, there are many other ways that I can get access to the same data. So even if MIT says, "We're going to only give the data to X sources," one of the things we see here is who else X is giving the data to - note these long chains. The fastest growing area on that chart are companies whose only goal is analytics; they take data that you would not think of as being connected, like health data and cable viewing data and they link them together. This is suggested in your scenario, except that in the business context, it represents new opportunities as they make data products that build ever-increasing profiles. As a first question we would want to determine, "Can we make data sharing transparent? Can there be a mechanism by which we can know all of the places that MIT is providing this data?" This could be done with the requirement that everybody who receives the data will also let MIT know who they give the data to. Then an individual can begin to answer whether or not they've been harmed.

DANIEL WEITZNER: So David, in your role as MIT's Chief Privacy Officer, we don't really have a roadmap for sharing the data and there are no HIPAA requirements for a lot of it. However, there are privacy issues, so how do you think MIT ought to work through this question of how to define the rules? Pretty much everyone here has agreed that rules need to be clear and uses need to be transparent. But if we don't know what the rules are, where are we?

DAVID HOFFMAN: We've been talking about how the technology has to be a part of the solution, but not all of the solution. We could think about that as the layers in the stack; some pieces of this are legislative and we need to understand what laws cover this and where the gaps are. Let me describe that as the hardware in the stack. Then we acknowledge the importance of policy, meaning the internal policies that the organization will be applying with regard to responsibilities and commitments extending beyond what is going to be legislatively required. We could think of that as the firmware. On top of that might be the technology, which we'll describe as the

software in the stack that will provide us with further protections. Finally, we have the implementation of accountability measures, which will be put in place to make real on all of those commitments.

Then we can consider what can be done about the harms that may be created here with the different pieces of that stack. This panel keeps coming back to transparency. As the CPO, I would have to say. "Look, I've tried to do my job by going out and talking to some of these students. I'm very hesitant to think that we will be able to communicate all of this to a student in a way that they can fully understand and will say, "Oh good, it's all transparent to me now, this is the piece I consent to."" Obtaining consent and providing notice to individuals plays an incredibly important role. It's going to be an enduring formation practice principle, but we need other things to supplement that, including an understanding of use, use restrictions, accountability, and increased ability for individual participation.

LATANYA SWEENEY: Just to be clear, when I say the word "transparency" here, absolutely none of the data map documentation is about the transparency to an individual. It was the fact that we were able to issue requests to certain groups and that breach notices required certain information. We were able to mine through the thousands of breach notices and so forth. If those two things did not exist, we would not even be able to give you that much documentation about where personal data goes. One significant concern in healthcare, where we think about what's the new big data, is that one of those big silos is health information exchanges. However, it's impossible to make the case for the evolution of differential privacy and homomorphic functions without being able to show why they're making a difference. The water flows to the lowest level and the lowest level for business is typically the thing that is the easiest thing to do. Differential privacy might raise the bar for HIPAA standards, but it will never be used. The only way to make that bar come up and to begin to make the case for these technologies is through a kind of transparency, so that people can show actual harm, so they can do these types of computations, so that you can make the case for why the risk is worth it.

DANIEL WEITZNER: In this case study, MIT's intention is to be a leader in best practices on the use of this information. Let's hear some comments on this from an entrepreneurial perspective. Part of the reason that we're excited about doing this kind of project is due to all of the innovative new services that could develop; FitByte is participating because they want to stimulate an ecosystem of new applications and services around their data. How can we square the circle between the complexity of Latanya's graph and all of the legal and technical uncertainties on one hand, while on the other hand, a bunch of scrappy entrepreneurs are going to take this stuff and run with it?

MONA VERNON: This is something that we think about a lot, because in our company the trust principle is what defines us. Thomson Reuters is based on this idea that trust really matters. So we try to be really precise on what we can do and we're seeing this rise of desire from customization and innovation on all this usage

data, but at the same time we're trying to be respectful of our clients' privacy needs. It's a really interesting time and there are things that you can do now in big data with easier tools, so it's becoming easier to innovate. The question is how to balance between driving the innovation and at the same time being concerned about the privacy aspects.

ANDY PALMER: Looking at Latanya's chart, if you were an entrepreneur is you would focus on the needs of that patient and the design principles for an interface for that person to control whether or not someone has used their information to do them harm. That is the right way to think about this and this should improve over time, if you look at what the Sage Bionetworks folks have done around genomics as an example; it's evolving, you can see it happening. Even 23 and Me is a great reference, when you look at the simplicity of that interface, regardless of whether the medical information is that they're providing you is useful or not. We're processing how people are going to consume very complex information in very simple ways. There's a serious lowest common denominator here and my biggest concern about these kinds of things is that as those designs happen and as those systems get built, there are many opportunities for bad actors to insert themselves into these systems and to do the wrong things. The government has to play an important role in making sure that if all of this data exists and all of these log files are out there and are accessible, there is a way to protect it. Even if you try and control it, there will always be some hacker that's smarter than you who is going to be able to crack your encryption. If we assume that's the case, do we have the legal infrastructure and a mechanism in place to go back and hold them accountable for any harm? If we don't have mechanisms to hold people accountable, it is a serious concern that the bad actors are going to control these systems.

CAROL ROSE: I agree with that and so much of what's going on right now has been done in secret. If it weren't for Edward Snowden, we wouldn't be even having this conversation right, because everything's being done in secret. Not once was any of the information about what the data the government was collecting on people and how they were using it, brought into discussion publicly. In fact, when groups like the ACLU tried to challenge it on behalf of journalists, lawyers, and human rights workers, they were told "You don't have standing to come in court, because you can't prove that you were being spied on." So there is no mechanism for people to get redressed, because there's no mechanism to even know when the government is obtaining data that might be used to do you harm and that's a huge problem.

MARK GORENBERG: One of the real questions here is that the students are looking for a trusted entity that is watching over them and they assume it's MIT. In reality, MIT has formed this deal with FitByte to make them sort of the owner, or at least the licenser of all the data. MIT should take care to figure out what the negotiation is and what makes sense. For example, should MIT be keeping GPS data, or should it follow a scheme similar to Snapchat, for example, where the data basically disappears into the ether - it's only good at that second, but not further?

ANDY PALMER: I lived through the e-commerce revolution in the early 90s and I feel like we're going through this again. Many of us were in a similar environment then, where people were worried about sharing their credit card information online, absolutely freaked out. Were those fears realized in a way? Well, maybe, but now all of this infrastructure exists now and there are good actors and bad actors. The reality is that we're getting comfortable with sharing information online and I think it will go the same way with our medical information.

DANIEL WEITZNER: Karen can you look at this for us from a more global perspective? The view on the panel is somewhere between "We have a bunch of technical tools that will kind of enforce privacy protection to whatever definition we decide" and "Well, we're going to innovate our way out of this problem – we're going to take that complex web turn it into something simple and users are going to feel in charge." The e-commerce analogy is apt, because it has worked quite well by many measures in the United States. However, other countries look at what we've done in the online personal information environment and think we're totally crazy. So, how do you think this would play out if we were having this discussion in a different culture?

KAREN KORNBLUH: I almost have the converse reaction, just speaking as somebody who has been living abroad and talking in this international economic standard setting organization, where the U.S. is the envy of the world because of our innovation and economic growth. In this workshop, we've talked about the benefits of big data and we've emphasized research, but we've also talked about economic growth as a benefit in innovation. In other countries, people complain about the U.S., but everybody would like to have a Silicon Valley and that's universally felt. Earlier, somebody said that industry will go to the least required solution; I don't think it's necessarily true. In a workshop like this, we're sharing not only industry best practices, but also best theoretical practices. If you could take that even further, I think there's a role for the government in terms of bringing down costs for the best theoretical options and letting people share that information.

When you and I were working on the OECD Internet Policy Making Principles, we emphasized that it was for the Internet economy, because we thought we could get people to an interoperable solution if we talked to them about innovation. So in the big data and privacy conversation, some people may look at Europeans and others who have different norms and say, "Well, we all have to arrive at the same norms." I think what we really need to do is to the take the problem to another dimension. How do we find some kind of interoperable work around? We've had the safe harbor with EU, we've had the OECD policy principles, the FIPPS and so we could assess, how would all those work in this world of big data and how do we make sure that big data is a platform for innovation just the way the Internet has been?

DANIEL WEITZNER: Here is another hypothetical in our case study. A problem has emerged where, all of a sudden MIT and Harvard students start going back and forth between campuses at an alarming rate. This has had a heavy impact on the subway

line, the T, here and the MBTA board has decided to go to a congestion pricing system. It puts students at an enormous disadvantage, so the MBTA then decides to give students a break if they have a bona fide need to travel that is established by proving that you need to get from one place to another, based on your class schedule. The MBTA wants access to everyone's class schedule. That helps the students because the train fares are reasonable again. However, it doesn't solve the congestion problem, so MBTA decides it needs access to everyone's schedule, in order to tune the train schedules precisely to the needs of the commuters and the students. Are we back in the same data and privacy situation again?

CAROL ROSE: Let's say that there's a bad guy who wants to get certain MIT students. If I know when they're traveling, then that's the kind of information I would try to get. It could also create perverse incentives for the individual. Let's say I'm a FitBit student and I want to get my count up because I'm feeling competitive and I know my professor is looking at it, so I'm going to start taking methamphetamine. Then I would like to get some substance abuse counseling, but if they know where and when I'm traveling, then they're going to figure it out. You can play out some of the scenarios and they're really bad, so we need to think very carefully about location data; this is tremendously revealing of who we are; who we are is where we are. If you turn location information over to the MBTA, then you have the possibility, not only for bad guys to get the data, but also to create a perverse incentive system.

LATANYA SWEENEY: I was going to pull together some of the threads that were there very quickly. One is the example of breaches, which illustrates the point that I made earlier about water going to the lowest spot. Most of these breaches involve data that is not encrypted, which is a simple solution that's been around a long time. So if data's at rest, you could encrypt it and you're done, but we don't do that and we end up with breaches. At least 99% of them are of that nature. The second thing concerns the relationship between innovation and our privacy regime; they go hand in hand. Computer scientists are about tomorrow's technology and the technology that we don't have today; so we want that innovation to happen in a responsible way. There are many times I have argued that Zuckerberg could not have produced Facebook in Cambridge, England, but he could in Cambridge, Massachusetts and that's specifically because of the privacy regime we have here being sort of patch work - anything kind of goes until it doesn't and then we might patch it again. That's worked out extremely well for our innovation companies. We've gone through an era where there is a lot of discussion about open data; you should just live your life in an open way; with transparency this is okay, then let's have open data sharing. You collected the data, you put it in the vault, and now nobody can know where it goes. And yet I can be harmed and that seems unfair. So we have the regime we have and we want innovation, we just want to make sure that we know where the harms exist.

MONA VERNON : For global companies like ours to innovate when we're positioned in Cambridge, Mass and Cambridge, England, it is actually very difficult, because we have to figure out which privacy policies should we worry about, Germany, France,

U.S.? One thing that would be a great help would be to have a U.S. policy that takes into account the global nature of innovation and helps us to operate in that place.

DANIEL WEITZNER: So the last stop in our case study is an alliance that's proposed between MIT and École Normale Supérieure, the great institution of French learning that has produced many of the great thinkers in France. However, there is this fundamental human right to privacy in the European Union Charter of Fundamental Rights. How would you advise MIT to persuade the École Normale that this alliance can work and that their values can be respected, along with U.S. practices?

KAREN KORNBLUH: This gets at the point that the Internet is global and the potential advantages of this big data are global. Some of the biggest advantages are found in the innovation that you would find at places like MIT and the École Normale, but there are plenty of others as well. We have to find a way to respect each other's differences. I remember a conversation we had at the OECD where my French colleague said, "Well, I didn't mean transparency in the Anglo-Saxon sense of the word." So we're not going to change each other, but I do think we have found workarounds and interoperable means. Safe harbor is a great example and we have to find a way to preserve that through the OECD. I would urge the École Normale to use this as a test case to see what some of the benefits of collaboration and experimentation could be.

DAVID HOFFMAN: If we dig underneath a bit, we might look at some of the opinions that are coming out of the Article 29 working party. The collection under Article 29 of the 1995 Data Protection Directive has participants from each of the individual member states coming together to express opinions on data protection and implementation of the directive. If you look at some of their opinions, particularly their recent opinion on purpose limitation, and some of the enforcement actions that are coming out of the FTC, we can actually see increasing convergence on some of the nuances when it comes down to protecting individuals. There's been a great deal of excellent analysis done, although it's not clear if that detailed analysis is getting back to influence the "hardware," meaning the actual legislation that's governing the model that is not working completely well in Europe. Of course, many people, including the administration, have said the model isn't working completely well in the United States either. A lot of excellent work was done on the Consumer Privacy Bill of Rights; respect for context, transparency was there. There is a real opportunity to dust that off and start another real conversation. It called for comprehensive privacy legislation, so we could talk about the role that it could play in a big data environment.

KAREN KORNBLUH: Some of the diplomacy that you did, that Cam did, that so many others did was truly important and I think we lost some ground with the Snowden revelations. I remember when we had a conversation with a bunch of privacy folks in Europe early in the Obama Administration and they said, "You've got to convince us that you care about privacy. We understand that you have a completely different framework; we have a tendency to stereotype the U.S., so explain to us that our

stereotypes are wrong." The messaging about the reality that we do have strong privacy rules is another important aspect to this, besides any changes to the firmware.

MONA VERNON: In the U.S. there are so many global companies that actually startups are starting to think of their market as global. What we're looking for is making that ability to innovate around big data easier, while respecting privacy regulations. The fact is that patchwork of laws and legislation makes it really difficult. It might create permanent chief privacy officer jobs, but that's not sufficient. Anything that could help the conversation on how to operate in a global framework would be fantastic.

SHAFI GOLDWASSER (Audience): So I was wondering about your Fitbyte example, where you say that MIT is considering -- one option was to make it mandatory and the other option was to give them a discount on their medical insurance. Since we are thinking of this as an economic benefit, we should note that some people care more about privacy and some people care less about privacy. If you could create an economic incentive to reveal your data in some sophisticated way, then those who care less, could go ahead with it. If I were a research institute, I might have a lot of data, and I would be willing to share with somebody else who has less data, if I get more benefit. I wonder if one could put some sort of economic model on it.

LATANYA SWEENEY: This idea of economic value of data is a very powerful one, because it changes the framework. So if MIT makes me a financial offer, then what did they tell me was going to happen to my data, what was disclosed exactly? How do I take them to court, if there was a misrepresentation, or I was ultimately harmed by this transaction?

CAROL ROSE: You need to be very careful not to create a dual regime where people who have wealth are able to have privacy and people who are poor are unable to have privacy. We already have this kind of inequality in terms of where we have things like hot spots and predictive policing; we target poor communities. So I think while there may be some merit to developing an economic model, we need to be cautious and ensure that due process protections would attach to an economic regime like that.

LATANYA SWEENEY: Privacy is already the domain of those who can afford it; you can see this in the use of loyalty cards, people who get welfare assistance, or have Medicaid - their data is far more readily available than people who don't.

TOM EASLEY (Audience): I'm from the Journal of the American Medical Association. Andy you said that you thought within 20 years people would be freely sharing their medical records the way they share credit cards now? There was also talk this morning about the value that will be created by demonstrating why patients should want to share that kind of data. How do we get from here to there and what are some of the land mines in the near term?

ANDY PALMER: We're at the very beginning of a radical consumerization of health information. One example is blood pressure. We all know that if you had better awareness of your blood pressure on a day to day basis and that awareness was shared with your closest family members and the people that care about you, it would create an incentive to manage your blood pressure better. This has been proven over and over again. The information that you would generate by using a blood pressure cuff on a daily basis would be your information; it wouldn't be the hospital's information, or a doctor's information. Now if the doctor used that information, the hospital might require them to bring it in and make it a part of your electronic medical record. However, those benefits are so significant and we're so close to having these kinds of capabilities in the hands of consumers in a relatively affordable way. The power of that information and the impact it can have on improving human health is going to be the real driver there. The idea of people owning their own health information and the information that matters is not necessarily focused on all the information that lives in the EMR over at MGH, but rather the information that you might generate from sensors that you wear on your body every single day; it's a very powerful dynamic.