Consumer Privacy Bill of Rights and Big Data:
Response to White House Office of Science and Technology Policy Request for Information

April 4, 2014 - submitted to bigdata@ostp.gov

Daniel J. Weitzner, MIT Computer Science and Artificial Intelligence Lab
Hal Abelson, MIT Department of Electrical Engineering and Computer Science
Cynthia Dwork, Microsoft Research
Cameron Kerry, MIT Media Lab
Daniela Rus, MIT Computer Science and Artificial Intelligence Lab
Sandy Pentland, MIT Media Lab
Salil Vadhan, Harvard University

I. Introduction and Overview

In response to the White House Office of Science and Technology Policy Request for Information on Big Data Privacy we offer these comments based on presentations and discussions at the White House-MIT Workshop "Big Data Privacy Workshop: Advancing the State of the Art in Technology and Practice" and subsequent workshops co-sponsored with Data & Society and NYU Information Law Institute and the UC Berkeley iSchool.

1. Big data analytics offers significant new opportunities for advances in scientific research in many fields. Presentations offered at the MIT workshop showed unique benefits for improved healthcare quality, advances in the understanding of diseases through genomics research, potential to improve educational effectiveness, and more efficient, safe transportation systems.

2. There are real privacy risks raised by ubiquitous collection of personal data and use of big data analytic techniques. Key risks include:

- *Re-identification attacks*
- *Inaccurate data or models*
- *Unfair use of sensitive inferences*
- *Chilling effects on individual behavior*
- *Excess government power over citizens*
- *Large-scale data breach*

3. The White House Consumer Privacy Bill of Rights offers policy approaches to address each of these risks. Some of the principles such as transparency, respect for context, security, access, and accountability will play especially important roles in big data issues. Transparency should be augmented beyond just visibility into policies, to also enable individuals and regulators to see how personal data actually flows and is used. The respect for context principle should be implemented with particular attention to developing and enforcing limits on how personal data is used, especially in circumstances where collection limits and ex ante consent are difficult to achieve.

4. Technical contributions from computer science can assess and in some cases control the privacy impact of data usage in a rigorous, quantitative manner. But as technology will not replace the need for laws and social norms protecting privacy, basic and applied research must be conducted in a cross-disciplinary context so that technical designs will meet social policy needs.

II. Benefits and Risks Specific to Big Data

A variety of presentations at the White House-MIT workshop[1] illustrate significant new knowledge to be gleaned from analysis of large data sets. The large scale analytic results are possible because database technology innovation has made it easier and cheaper to collect, store, and analyze data. As Mike Stonebraker [slides] explained, computer science continues to develop new techniques for collecting and storing ever-expanding volumes of data, and analyzing these with increasingly fine-grained detail. While we can identify core big data analytic technology platforms and associated privacy enhancing technologies, it would be a mistake to conclude that all uses of big data technology implicate the same privacy values or, indeed, that a single set of privacy-enhancing technologies apply to all big data applications.

Computer science faces the challenge of storing and analyzing large amounts of data in part because so much more data is now being collected. Much of the digital infrastructure becoming ubiquitous around the world -- from mobile phone networks to transportation, finance and healthcare systems -- is trending toward recording nearly every action human beings take. John Guttag [video] showed how he was able to learn about the spread of hospital-acquired infection, but needed access to large amounts of patient data, plus personal information about un-infected patients as well as personal details about the doctors, nurses and other staff who were potential disease vectors, even if their interaction with the infected patients were marginal. Manolis Kellis [slides], a computer scientist working with large scale genomic data, explained the need to have large samples of the population to detect very small scale phenomena that, though widely dispersed in the data, are nonetheless critical to understanding the way the human genome functions. Sam Madden [slides] spoke about his Car-Tel experiment using mobile phones in cars that has enabled insights about how to make transportation networks more efficient and even

[1] The complete agenda, video of each presentation and slides and workshop summary report can be found at the workshop website. http://web.mit.edu/bigdata-priv/

reduce risky driving behavior by teenagers. And Anant Agarwal [video] showed how education researchers can use personal data from online courseware systems to learn about the most effective pedagogical techniques for different types of students.

All of this research requires access to large amounts of data. In some cases that data will have been collected for a purpose defined at the time of collection.   But, in a big data environment looking for unanticipated relationships, it is not always possible to foresee at the time of collection all the questions that may subsequently be asked of the data. In some cases, it may be possible to apply new privacy-enhancing techniques such as differential privacy (see Section IV) to enable research with little or no privacy exposure, and thereby reduce the need for consent. However, many workshop participants expect that in some significant set of circumstances these techniques will not be applicable without real loss of useful research results. Researchers at the workshop generally expressed the view that obtaining consent after collection for specific new analytic tasks would likely be impractical in many instances and reduce the ability to gain useful knowledge from data. As will be explained below (section III. B), the Consumer Privacy Bill of Rights anticipates the need to protect privacy and guard against unwanted mission creep in situations in which individualized notice and consent is not feasible.

Acknowledging the social and economic benefits of large scale analytics, we must also recognize substantial privacy risks that have the potential to lead to real harms. Leading risks include the following:

1. *Re-identification attacks*: Data may be disclosed with certain identifiable information removed, but is still susceptible to being re-identified by correlating the weakly de-identified data with other publicly or privately held data.[2]   Re-identification is not the only problem that can occur with release of "de-identified" data. Even without matching an individual to a specific record, but simply being able to conclude that the individual corresponds to one of a small number of records, or one of a possibly large set of records that agree on a particular attribute (HIV positive status, for example), can be harmful. These risk grows as the availability of data increases and the ability to correlate improves.
2. *Inaccurate data or models*: More data is by no means always a guarantee of more accurate results. Data sets may contain either inaccurate data about individuals and/or employ models that are incorrect at least as to particular individuals. This risk increases as larger datasets are to generate increasingly complex models which may be applied to decisions about individuals without rigorous validation. When decisions are made based on either inaccurate data or incorrect or imprecise models, individuals can suffer harm by being denied services or otherwise treated incorrectly.
3. *Unfair use of sensitive inferences:* There are numerous examples of big data analytics able to infer sensitive facts (creditworthiness, sexual orientation, spousal relationships,

[2] Narayanan, Arvind, and Vitaly Shmatikov. "Myths and fallacies of personally identifiable information." *Communications of the ACM* 53.6 (2010): 24-26. Sweeney L. Matching Known Patients to Health Records in Washington State Data. Harvard University. Data Privacy Lab. 1089-1. June 2013.

likely future location, or other information individuals may choose not to disclose) from correlation with other public or less sensitive personal information. Even if these sensitive inferences are accurate, it may be unfair to use these inferred facts about individuals to make certain kinds of decisions.

4. *Chilling effects on individual behavior*: Individuals may alter their behavior and expression in response to a sense of being watched. Awareness of increasingly ubiquitous data collection capable of deriving detailed conclusions about individuals through big data analytic techniques may have the effect of limiting individual participation in a variety of activities, including in constitutionally protected areas such as politics and religion.
5. *Excess government power over citizens:* The power of big data analytics can also be deployed to expand the amount of knowledge governments have about their citizens. Even though liberal democracies may impose limitations on their own collection, use, and retention of large data sets, their use of such analytics may facilitate more pervasive use by authoritarian governments
6. *Large-scale data breach*: Data breach is a continuous risk. As the scale of data collection, the flow of this data to different parties with access to this data, and the ability to derive detailed information all increase, the risk of harm from breach also grows. With big data, data thieves have a richer set of targets and tools available.

III. Guidance from the Consumer Privacy Bill of Rights

For each of the big data privacy risks identified here, the substantive principles in the Consumer Privacy Bill of Rights offer guidance to develop concrete responses to those risks in a manner that provides clarity for individuals and flexibility for innovative big data analytic applications. Given the rapid evolution of big data analytic applications, the unique procedural aspects of the Consumer Privacy Bill of Rights also offers a means by which principle-based privacy approaches to new applications can be developed rapidly as enforceable codes of conduct and then enforced under the FTC's existing statutory authority.

A. Overview of Consumer Privacy Bill of Rights Principles.

1. Individual Control

"Consumers have a right to exercise control over what personal data companies collect from them and how they use it."[3]

The principle of Individual Control in the Consumer Privacy Bill of Rights shifts the focus away from the longstanding principle of notice-and-choice to more dynamic and flexible mechanisms. Notice-and-choice is one important mechanism of privacy protection, but the Commerce Department Green Paper[4] process found that routine checking of boxes puts too much weight

[3] The White House, Consumer Privacy Bill of Rights (February 2012). [hereinafter CPBR]

[4] United States Department of Commerce, Internet Policy Task Force, Commercial Data Privacy and Innovation in the Internet Economy: A Dynamic Policy Framework (December 2010)

on the unmanageable burden of reading privacy policies and does not differentiate among situations that present material privacy risk and those that do not. Whether data is used in a commercial context, or for basic medical or scientific research, may also be relevant to what kind of individual control is warranted. The Consumer Privacy Bill of Rights therefore calls for contextual mechanisms to exercise choice at the time of collection "appropriate for the scale, scope, and sensitivity of the data in question," and also for additional mechanisms to address the use of personal data after collection.

This principle reflects the Big Data environment in two ways. First, it recognizes that the increasing velocity and variety of data collection make notice-and-choice ineffective; consumers are asked for consent too frequently and on devices such as mobile phones that are not suited to deliberate informed consent. Second, it recognizes that the velocity of data includes increased sharing with third parties with whom consumers do not have a direct relationship. Moving away from a one-size-fits-all notice and choice regime in which consumers often face a binary choice (either to give up data control or not to use a service) will strengthen fair exchange of value between consumers and companies by allowing consumers greater choices of how much to share in exchange for a given level of features and benefits.

There are certainly contexts in which individual control will play a minor role as compared to other principles such as Respect for Context and Focused Collection. The expanded use of sensors and other developing forms of automated data collection will make notice-and-choice and other mechanisms of control impossible or infeasible in an increasing number of circumstances. The principles of Consumer Privacy Bill of Rights are intended to apply in interactive and dynamic ways appropriate to the technologies they address; the expansion of Big Data will put a premium on such application.

2. Transparency

"Consumers have a right to easily understandable and accessible information about privacy and security practices."

The Transparency Principle requires companies to disclose when and why they collect individuals' personal data, so that consumers can guard against misuse of their personal data. Beyond just individual awareness, transparency has a vital function for the evolution of privacy norms themselves. In the modern history of information privacy, transparency has enabled consumer advocates, policy makers, enforcement agencies, the press and the interested public to engage in dialogue and criticism about how commercial privacy practices are evolving. It is only with awareness of actual privacy practices that society can have a meaningful dialogue about which practices are acceptable and which fall outside legal and/or social norms.

Meaningful transparency in big data systems will require going beyond just disclosure of policies as to personal data. Enabling citizens, governments and advocates to address big data privacy challenges requires a more active transparency - the ability to be aware of and track the actual flow and use of personal information. Big Data is different from the regular use of personal data in that consumers are not only affected by the primary collection of data, but also by the subsequent aggregation of that data to inform algorithms that govern companies' decision-making and affect individual users. Therefore users need additional tools to help them follow complex data flows and understand what picture of them this data enables, beyond just the general disclosures in privacy policies. For example, disclosing to an individual that a health insurance company knows her address does not inform her of the likelihood that this information is being combined with multiple other databases to create "neighborhood profiles" that in turn could affect pricing for individual customers within those neighborhoods.

The requirement that companies detail how they will use gathered data bears more weight in the Big Data context than requirements that companies simply disclose what personal data they collect. A transparency framework that updates consumers when companies come up with new uses for aggregated personal data will increase user trust and help ensure that accountability takes place at the rate of business growth, rather than at the rate of governmental enforcement.

3. Respect for Context

"Consumers have a right to expect that companies will collect, use, and disclose personal data in ways that are consistent with the context in which consumers provide the data."

The principle of Respect for Context builds on the recognition that expecting users to read notices and make choices for every single individual collection and use of personal data element is unsustainable. Requiring users to make such a large volume of choices is not fair to the individual. In the Big Data environment of unstructured data and unforeseen correlations, rigid insistence on notice and choice would also render impractical many of the applications that can result in important new scientific discoveries, more efficient public infrastructure, and innovative new commercial services in a Big Data environment . The Respect for Context principle therefore recognizes that consent can be inferred in some circumstances and that privacy protection will depend on ensuring that the *uses* of personal information are faithful to the original context in which the individual provided the data (whether actively such as in a transaction, or passively such as via sensing devices). Giving users the right to expect that the context of collection be respected in further uses will protect individuals from unwanted surprise, and at the same time allow the development of valuable new big data applications.

4. Security

"Consumers have a right to secure and responsible handling of personal data."

Collecting, storing and using personal data comes with inherent risks, including the possibilities of privacy loss and data theft, modification or destruction. The security principle recognises this, calling on companies to assess these risks and maintain reasonable safeguards, because without them data-driven economic growth will be limited by a lack of trust. In addition, the risks of data re-identification need to be investigated and mitigated as much as possible. When multiple anonymous, heterogeneous databases are combined, prediction algorithms can be used to re-identify individuals. While the security principle does not explicitly recognise de-anonymization risks, they deserve special mention because they are extremely difficult to assess in practice, especially given the unpredictability of additional sources of information that may be employed in the attempt to de-anonymize. Currently, these risks are assessed by simply attempting to de-anonymize the data: clearly, new innovations are needed to address this tough problem.

5. Access and Accuracy

"Consumers have a right to access and correct personal data in usable formats, in a manner that is appropriate to the sensitivity of the data and the risk of adverse consequences to consumers if the data is inaccurate."

The Access and Accuracy principle requires consumers to have the ability to access and correct personal data in usable format. It addresses Big Data by recognizing that the opportunity to access and correct personal data is especially important where" [a]n increasingly diverse array of entities uses personal data to make decisions that affect consumers …." With such access and ability to correct, errors or inaccurate data can be more easily discovered and detected due to possible contradictions with other datasets. More accurate inferences can be made about consumers with a less likelihood of misinterpretation of information itself. In turn, these inferences could also be used in discovering anomalous pattern within the different datasets and hence be useful in understanding and detecting possible additional errors.

6. Focused Collection

"Consumers have a right to reasonable limits on the personal data that companies collect and retain."

Large-scale undirected collection of information that is seemingly unrelated to the main use of an application exposes consumers to unnecessary risk, as even collection of seemingly innocuous data may allow unwanted intrusions or inferences about sensitive details of a person's life.

This CPBR principle is a conscious move away from the data minimization principle. It recognizes that Big Data applications often make use of wide-ranging data sets related to the user that do not have obvious initial value. Hence, the principle does not demand absolute minimization. Declining costs for pervasive sensing devices such as health trackers and mobile phones has led to increasingly broad collection of data, while the cost per bit of storing such information has dropped. These combined factors incentivize companies to collect and store such information indefinitely, regardless of that data's usefulness, just in case future exploitation of such data might prove useful. The Focused Collection principle allows for collection of large data sets, but as a check against unrestricted and possibly unreasonable collection practices, it calls for thoughtful decisions about what data to collect or retain and new innovation, both in the policy and technical arena, to enable companies to target their collection.

7. Accountability

> *“Consumers have a right to have personal data handled by companies with appropriate measures in place to assure they adhere to the Consumer Privacy Bill of Rights*.”

The Accountability principle recognizes the necessity for organizations using personal information to have strong training, internal policies, and audit mechanisms to assure compliance with legal requirements and, more broadly, wise and responsible use of the data they hold. The MIT workshop revealed a number of big data scenarios for scientific research in which the benefits of large scale analytics can only be achieved by allowing wide-ranging internal analysis of the data, with controls on the ultimate use of the data. Therefore, it is especially important that organizations employing these techniques have clear internal policies to prevent against misuse of data, that employees within the organization have a clear sense of what these policies mean, and that audit mechanisms are in place to help guard against either intentional or unintentional misuse.

Applying state-of-the art approaches to institutional compliance, the most important thing is execution - achieving actual compliance with rules or establishing why noncompliance has occurred. To that end, it's important to distinguish objective performance measurement or diagnosis from normative assignment of responsibility (or - even more so - blame, liability, or culpability). The latter invite a backward-looking defensive posture that gets in the way of honest appraisal of the facts and forward-looking improvements. That appraisal is more effective if individual responsibility is secondary.

B. Addressing Risks

Consider how each of the Big Data privacy risks identified above can be addressed under the framework of the Consumer Privacy Bill of Rights.

<u>1. Re-identification risk:</u> The risk that personal data can leak from big data research platforms is real. Principles including Transparency, Security, Focused Collection, and Accountability will all be important to manage this risk. T*ransparency* will enable regulators, enforcement authorities, and interested members of the public such as advocates and academics to know what kind of data is being released and in what form. Assessing whether the users’ rights to have data held *securely* should include an assessment of who is able to access the data and therefore whether the re-identification risk can be minimized by binding those individuals to legal commitments to avoid re-identification. The right to have only *focused collection* of user data will also reduce re-identification risk by limiting gratuitous collection of data. And finally, an organization with strong institutional *accountability* procedures in place should handle data carefully and only release it publicly after evaluating the risk of re-identification. If the organization fails to consider this risk, then appropriate parties can be held accountable for resulting harm.

2. Data and model inaccuracy: The Consumer Privacy Bill of Rights can reduce the risk that decisions are made about an individual based on inaccurate information or an incorrect model. The principles of Transparency, Respect for Context, and Access and Accuracy are all useful to ensure fairness in big data decisionmaking. Since the Fair Credit Reporting Act (FCRA) was enacted, individuals have had basic *transparency* rights enabling them to know that personal information about them is being used for important decisions, as well as the right to *access* and correct personal data to ensure that it is *accurate*. Such transparency is critical to make sure that individuals know their data is being used therefore be able to assure its accuracy or decide to exclude themselves from uses they object to.

Similarly, the Access and Accuracy affords a mechanism to assure that data and the inference drawn from are accurate.[5] While most consumers will not be able to identify errors in models, transparency on inferences drawn by a model may shine light on algorithmic errors,

The CPBR Transparency principle also requires that companies explain how they will use data and this should be understood to include relevant information about the decisionmaking models and algorithms. There is work to be done to define how much of the decisionmaking metrics should be exposed, as some of that information will be proprietary. Enough context about the decision metrics should be made available to enable consumer protection enforcement agencies and other stakeholders to assess whether the decisions models are fair.

These principles are reinforced by the Respect for Context principle.   When data is used out of context, the CPBR provides that if “companies decide to use or disclose personal data for purposes that are inconsistent with the context in which the data was disclosed, they must provide heightened measures of Transparency and Individual Choice.” This will help individuals to flag uses of information that are likely to create risk and increase the likelihood that both personal data and models derived from personal data are accurate.

3. Unfair use of sensitive inferences: Even if inferences are accurate, it may be unfair as a matter of ethics or public policy to use such information for certain purposes. For example, behavioral profiling techniques used for marketing purposes can provide advertisers the ability to reach audiences defined by age, ethnicity, race, gender and other sensitive categories. The recent statement of privacy principles from leading civil rights organizations (“Civil Rights Principles for the Era of Big Data”) offers useful guidance on this point.The Respect for Context principle was specifically designed to prevent misuse of such profiles for more sensitive, harmful discriminatory purposes. As the CPBR explains:

[5] “An increasingly diverse array of entities uses personal data to make decisions that affect consumers in ways ranging from the ads they see online to their candidacy for employment. Outside of sectors covered by specific Federal privacy laws, such as the Health Insurance Portability and Accountability Act (HIPAA) and the Fair Credit Reporting Act, consumers do not currently have the right to access and correct this data.” CPBR p20

> The Administration also encourages companies engaged in online advertising to refrain from collecting, using, or disclosing personal data that may be used to make decisions regarding employment, credit, and insurance eligibility or similar matters that may have significant adverse consequences to consumers…. Such practices also may be at odds with the norm of responsible data stewardship that the Respect for Context principle encourages."[6]

Just because it is possible to learn or infer a sensitive characteristic of an individual, that does not imply that it is either legally or ethically permissible to use such an inference (no matter how accurate or inaccurate) for all purposes. However, addressing the use of such characteristics is a matter of social policy broader than privacy policy. Antidiscrimination laws and norms of countries around the world regularly prohibit acting in a discriminatory manner based on information about an individual, even if it is publicly available. Indeed, some of the personal characteristics that entail the highest degree of legal concern include gender and race, attributes of individuals that are readily observable and in most cases public information.

The Transparency and Access Accuracy principles provide mechanisms that can be helpful in identifying where data collected about individuals is used in ways contrary to legal or ethical principles. Despite this, reflexive and poorly justified application of the Fourth Amendment third party doctrine can lead to the "unwarranted" assumption that as soon as personal data is public it can be used for any purpose. The Respect for Context principle stands in opposition to this view and squarely for the proposition that privacy interests in personal information are determined as much by how the data is to be used as is the public or non-public status of the data.

4. Chilling effects on individual behavior: Among the paramount constitutional concerns at the heart of privacy is protection of the freedom of association enshrined in the First Amendment of the US Constitution. That is why President Obama introduced the Consumer Privacy Bill of Rights by recognizing the importance of upholding individual freedom of association:

> Citizens who feel protected from misuse of their personal information feel free to engage in commerce, to participate in the political process, or to seek needed health care. This is why we have laws that protect financial privacy and health privacy, and that protect consumers against unfair and deceptive uses of their information. This is why the Supreme Court has protected anonymous political speech, the same right exercised by the pamphleteers of the early Republic and today's bloggers.[7]

The CPBR can protect against chilling effects in the first instance through the Individual Control principle. Citizens who feel in more control over their personal data will feel more free to engage in activities online and offline. Transparency is critical to help assure individuals that they have

[6] CPBR p 18

[7] President's Preface to the Consumer Privacy Bill of Rights. February 23, 2012.

some understanding of when their personal data is collected and how their data is used. Respecting the context in which personal information is collected and avoiding out-of-context uses will limit the degree to which individuals are surprised but subsequent data usage and increase trust. Finally, knowing that they the right to access and correct personal data will reduce mistrust and fear that data could be inaccurately used against an individual's legitimate interest.

5. Excess government power over citizens: Large scale analytics can expand government investigative power by revealing otherwise hidden knowledge about social relationships, political action, and other details of citizens' First Amendment-protected associative or expressive activity. Traditionally, civil liberties concerns about undue expansion of government surveillance and data gathering power are addressed in the United States by limiting government action with respect to collection of private information and property through the Fourth Amendment protections against unreasonable search and seizure. Today, however, enhancement in the government's law enforcement and national security investigative power are driven by big data analysis of information that often originates with private sector organizations such as network operators and Internet edge services. The primary venue for deciding on the proper scope of government surveillance power should be norms of government action that are beyond the scope of the Consumer Privacy Bill of Rights.

As much of the data in question originated with the private sector, several principles in the CPBR are relevant to managing the risk of that private collection of information contributes to excessive government power. *Individual choice* will help uses avoid having personal data collected if they are engaging in activities that they would want to remain outside government visibility. *Transparency* will help individuals make those choices and *Access and Accuracy* will make sure that personal data that does come into government hands is accurate. *Security* will help prevent surreptitious surveillance.

6. Large-scale data breach: With big data comes risk of bigger data breach. The CPBR *Focused Collection* principle can have some role in limiting harm when data that was never actually needed is subject to a breach. Most importantly, organizations that collect, use and store large repositories of personal data must follow good security practices that "best fit the scale and scope of the personal data that they maintain." Furthermore, the Administration's call for Federal Data Breach notification law is all the more important give the increased security risks from the growing size and scope of big data repositories.

C. Applicability of the Consumer Privacy Bill of Rights to Big Data privacy challenges

Large scale analytics has long been a factor in privacy policy, beginning with the enactment of the Fair Credit Reporting Act of 1970, the law that was written to regulate the leading big data enterprise of that era, the consumer credit bureaus. The credit reporting agencies took the then-unprecedented and to many, quite alarming step of collected detailed transactional data on the financial life of a significant proportion of the adult population in the United States and then subjecting it to sophisticated analysis for the purpose of developing credit risk scores. So the challenge of addressing large-scale integration of data used for purposes that an have a real impact, positive or negative, on the lives of individuals, is not new. We therefore see no reason to abandon time-tested privacy principles and have shown how the modernized version of these principles in the Consumer Privacy Bill of Rights can apply to big data analytics.

Addressing privacy challenges given the “velocity, volume and variety” that characterizes new big data systems does call for greater reliance on some of the principles in the CPBR than others. We have emphasized transparency, respect for context, security, access and accuracy, and accountable as elements of the CPBR that will bear significant weight in big data privacy protection.

First and foremost, an expanded commitment to *transparency* is necessary to guard against the risk of unfair, inaccurate use of personal data. The variety of personal information in big data systems requires a more active transparency in which individuals, consumer advocates and enforcement agencies can understand precisely how personal data is used, in some cases with resolution down to the level of individual data elements. Recognizing that individual control and consent may not be practical for high velocity collection and use of personal data, such systems will place more reliance on respect for context, assuring the information is only collected where the context makes such collection reasonably apparent, and that the use is consistent with the original context of collection. In context use should be able to proceed without individual consent, but out of context use would require increased transparency and individual control.

Large collections of personal data create increased risk of breach and loss, to security must be given special attention. As important decisions may be made through big data systems, access and accuracy rights are vital to be sure individuals are not treated unfairly. And finally, institutional accountability mechanism are vital to assure that all of the principles in the Consumer Privacy Bill of Rights are adhered to the use of big data systems.

Beyond just the substantive principles of the Consumer Privacy Bill of Rights, the larger policy process that the Administration’s privacy framework puts into place has a dynamic, flexible quality that will be especially important to help American society evolve new privacy norms in response to the challenge of large scale analytics. As explained by Ken Bamberger and Deirdre Mulligan, the evolution of ‘privacy on the ground’[8] has enabled the evolution of privacy rules in a manner that is responsive to public requirements while at the same allowing flexibility for the development of new services and business models. The Consumer Privacy Bill of Rights framework is designed to facilitate the continuous evolution of norms and rules as large-scale analytics drive new business models.

[8] Bamberger, Kenneth, and Deirdre Mulligan. "Privacy on the Books and on the Ground." *Stanford Law Review* 63 (2011).

IV. Research agenda based on technical and policy observations from workshops

Technical contributions from computer science can assess and in some cases control the privacy impact of data usage in a rigorous, quantitative manner. These techniques can help assure that systems handling personal data are functioning consistent with desired public policies and institutional rules. In some cases, these controls can prevent disclosure or misuse of data up front. In other cases, systems can detect misuse of personal data and enable those responsible to be held accountable for violating relevant rules. These technologies are at various stages of development, some ready for broad deployment and others needing more research to enable practical application. Developing the technological base that enables people to be in control of their data and assure it is used accountably is a key challenge. Solutions to this challenge are feasible.

The privacy risks we have identified above (III.B) along with the principles in the Consumer Privacy Bill of Rights provide guidance for shaping a research agenda to expand the established theoretical foundations of privacy enhancing technologies, to integrate them into systems design, and to develop public policy models for big data that address these risks in line with CPBR. A five-prong, cross-disciplinary research agenda is called for:

- Theory: Work in cryptography and theoretical computer science provides the foundation for privacy-sensitive controls on how personal data can be accessed and computed with. Techniques such as fully homomorphic encryption enable computation on data while it remains in encrypted form, thus reducing the risk that the data could be released in raw form, or that partially de-identified data could be re-identified. Secure multiparty computation and functional encryption can allow those who seek to use data to perform specifically-approved computation on personal data but limit access only to execution of those functions. Finally, differential privacy provides a framework and tools for enabling statistical analysis of data while ensuring that personal information does not leak. Continued work on the theoretical foundations of personal information may yet yield new techniques beyond these.

  While the theoretical foundations for a number of the above approaches are well-established, comments from a number of presenters and participants at the MIT workshop reveal the need to support research to scalability and models for computational tractability of these techniques.

- *Systems*: Systems research provides a variety of promising avenues to create data architectures that hold personal data while providing for a variety of privacy protections. Encrypted databases such as CryptoDB allows queries directly into encrypted data stores, thereby increasing the security of personal data while allowing access for queries by authorized users. Continued research in this area can make such database architecture more robust, scalable, and offer improved security properties. Accountable systems instrument traditional databases and other structured data repositories,tracking the flow and use of personal information in order to assess compliance with rules governing that data. Using knowledge representation, formal reasoning systems, and linked data architectures, accountable systems can provide a scalable means of helping data users to comply with known rules, and to demonstrate to the public that data is being used only for specified purposes and that there is accountability for violation of privacy rules. Further research in this area will increase the expressivity of the policy languages used to represent rules and explore new reasoning techniques that scale with increased data volume and variety.

- *Human Computer Interaction:* Many of the principles in the Consumer Privacy Bill of Rights (Individual Control, Transparency, Respect for Context, Access and Accuracy, to name a few) seek to give individual users greater awareness and control over their personal information relationships with others. Researchers, consumer advocates and companies are all aware of the ongoing challenge of designing systems to enable users to understand and control their personal data. As the scale and complexity of personal data usage grows, these problems will be all the more challenging. Research into user experience design, machine-assisted assessment of context, and good security mechanisms will be critical to support the efforts of end-user systems designers.

- *Policy*: Technology will not replace the need for laws and social norms protecting privacy or the use of personal information. We should expect that systems are built to perform according to privacy rules and make enforcement of those rules easier, especially as the scale of data usage increases beyond the point that manual compliance and audit can be effective. Policymakers also have much to learn from advances in technology design. Scientifically-informed privacy policy research can help guide policymakers to understand how to take best advantage of privacy-enhancing technologies. While broad privacy principles and risks may be clear, there are still conceptual questions about how to apply these principles to big data systems, challenges in human computer interaction to design privacy enhancing systems and better understanding of how privacy enhancing technologies might actually be used at Web scale.

- Technology/Policy Integration: In each of these research areas, multi-disciplinary collaboration will be critical. Designing and developing algorithms and systems to enhance privacy cannot succeed as a monolithic technical exercise. However, throughout the technical and policy discussions at each of the three workshops, a variety of implicit and explicit definitions of privacy were used. Some implied that privacy is synonymous with secrecy and complete confidentiality – that as soon as personal information is available to anyone else, privacy is lost. Others suggested that privacy is properly understood as the ability to control how personal information is disclosed and/or used. Finally, privacy is understood by some as a question of whether personal data is used in a manner that harms the individual.

One example of a technology policy integration is the Living Lab is being deployed at MIT. The goal of this project is to concretely explore privacy and big data policies and their effects, and to promote greater idea flow within MIT. Software tools such as our openPDS (Personal Data Store) system and Accountable Systems tools gives people the ability to have active transparency and control over where their information goes and what is done with it. Data sharing maintains provenance and permissions are associated with data, and automatic, tamper-proof auditing is supported. This allows enforcement and compliance with information usage rules and helps to minimize the risk of unauthorized information leakage.

Realizing the goal of designing systems that do a better job of respecting privacy requires continued research and dialogue with a wide range of disciplines that come together to define and refine privacy requirements. As those requirements are sure to be context driven, expertise from the wide range of privacy contexts will be required for successful research efforts.

V. Conclusion

Meeting big data privacy challenges requires social and legal consensus on how our fundamental privacy values apply in this new context. There is much scientific, social and commercial value to be realized. As members of the academic community, we are committed not only to expanding our research efforts so that technical tools for privacy protection keep up with the rapid pace of large scale data analysis, and but also to playing our part in the evolution of the legal rules and social norms that are necessary to maintain public trust in this arena.