March 31, 2014

To: The White House Office of Science and Technology Policy (OSTP)
Re: Big Data Study; Request for Information

From: Micah Altman[1], Director of Research, MIT Libraries;
Non Resident Senior Fellow, Brookings Institution
David O'Brien, Project Manager, Berkman Center for Internet & Society, Harvard U.
Salil Vadhan, Vicky Joseph Professor of Computer Science and Applied Mathematics,
Director, Center for Research on Computation & Society, Harvard U.
Alexandra Wood, Fellow, Berkman Center for Internet & Society, Harvard U.

on behalf of *Privacy Tools for Sharing Research Data Project*, Harvard U.

We appreciate the opportunity to contribute to the White House's request for information (RFI) on big data and privacy. These comments address selected privacy risks and mitigation methods relevant to this review. Our perspective is informed by substantial advances in privacy science that have been made in the computer science literature and by recent research conducted by the members of the Privacy Tools for Sharing Research Data project at Harvard University.[2]

As a general matter, we are proponents of transparency and open access to data. We and our colleagues have previously published on the benefits to researchers and the public of wider data access.[3] However, we are also acutely aware of the challenges related to confidentiality that arise when collecting, analyzing, and sharing data pertaining to individuals.[4] In our research, we have studied various approaches to sharing sensitive data, and we strongly believe that a sophisticated approach to data disclosure is needed in order to ensure both privacy and utility. As numerous reports by the National Research Council have shown, naïve treatment of information confidentiality and security has become a major stumbling block to efficient access to and use of research data [National Research Council 2005, 2007, 2009, 2010].

These reports and related research have established a general framework for privacy analysis.

[1] The authors take full responsibility for these comments. However, the authors wish to thank the other members of the project for their comments and insights and to acknowledge that these comments build upon joint work conducted with all members of the Privacy Tools project; and to thank Yves Alexandre de Montjoye for review and insightful comments and suggested edits.

[2] The Privacy Tools for Sharing Research Data project is a National Science Foundation funded collaboration at Harvard University involving the Center for Research on Computation and Society, the Institute for Quantitative Social Science, the Berkman Center for Internet & Society, and the Data Privacy Lab. More information about the project can be found at http://privacytools.seas.harvard.edu.

[3] A representative list of publications related to the Privacy Tools for Sharing Research Data project is available at: http://privacytools.seas.harvard.edu/publications. *See, e.g.*, King (1994); Altman & McDonald (2010; 2014); Altman, et al. (2010).

[4] See, e.g., Sweeney (2013); Sweeney et al. (2013); Ziv (2013).

Any analysis of information privacy should address

- the scope of information covered,
- the sensitivity of that information (its potential to cause individual, group, or social harm),
- the risk that sensitive information will be disclosed (by re-identification or other means),
- the availability of control and accountability mechanisms (including review, auditing, and enforcement), and
- the suitability of existing data sharing models,

as applied across the *entire lifecyle of information use*, from collection through dissemination and reuse.

On the whole, a modern approach to privacy requires a realization that the risks of informational harm are generally not a simple function of the presence or absence of specific fields, attributes, or keywords in the released set of data; redaction, pseudonymization, coarsening and hashing, are often neither an adequate nor appropriate practice, nor is releasing less information necessary helpful; and that a thoughtful analysis with expert consultation is necessary in order to evaluate the sensitivity of the data collected, to quantify the associated re-identification risks, and to design useful and safe release mechanisms.

1. Scope of information

The general scope of information covered by this RFI is extraordinarily broad. Both the government and third parties have the potential to collect extensive (sometimes exhaustive), fine grained, continuous, and identifiable records of a person's location, movement history, associations and interactions with others, behavior, speech, communications, physical and medical conditions, commercial transactions, etc. Such "big data" has the ability to be used in a wide variety of ways, both positive and negative. Examples of potential applications include improving government and organizational transparency and accountability, advancing research and scientific knowledge, enabling businesses to better serve their customers, allowing systematic commercial and non-commercial manipulation, fostering pervasive discrimination, and surveilling public and private spheres.

2. Information sensitivity

Generally, information policy should treat information as sensitive when that information, if linked to a person, even partially or probabilistically, possibly in conjunction with other information, is likely to cause significant harm to an individual, vulnerable group, or society. There is a broad range of informational harms that are recognized by regulation and by researchers and practitioners in the behavioral, medical, and social science fields [Bankurt & Ander 2006; Lee 1993]. Harms may occur through directly as a result of the reaction by the subject or others to the information, or indirectly as a result of inferences made from that information.[5] Types of potential harms to individuals associated with information include loss of insurability, loss of employability, market discrimination[6], criminal liability, psychological harm, loss of reputation, emotional harm, and loss of dignity (dignitary harm). Broader harms to groups and society include social harms to a vulnerable group (e.g. stereotyping), price discrimination against vulnerable groups, market failures (e.g. through through enabling manipulation, or eliminating uncertainties on which insurance markets are predicated), and the broad social harms arising from surveillance (such as the chilling of speech and action, potential for political discrimination, or for blackmail and other abuses [see Richards 2013; and see Solove 2006 for a history of existing restriction on surveillance])

3. Re-identification risks and private information leakage

A traditional approach to the sharing of privacy-sensitive data is to "anonymize" the data by removing "personally identifying information" such as name, address, social security number, etc. However, it is now well-understood that stripping identifiers provides very weak privacy protections, and it is often easy to "re-identify" individuals in supposedly anonymized datasets. For example, in the late 1990's, Latanya Sweeney showed how to identify the record of William Weld (governor of Massachusetts at the time) in an anonymized medical claims dataset by comparing sex, ZIP code, and date of birth with publicly available voter registration rolls; these three seemingly innocuous traits uniquely identify well over 50% of the US population [Sweeney 1997; Sweeney 2000]. In general, it takes very little information to uniquely identify an individual, and there have been numerous other examples where this phenomenon has been exploited for re-identification (e.g. the re-identification of Netflix Challenge data in [Narayanan & Shmatikov 2008]).

The size and richness of big data datasets makes them notably extremely hard to anonymize [National Research Council 2007]. For example, no method currently exists that allows detailed location data to be anonymized and then safely published. Yves-Alexandre de Montjoye showed that individual mobility traces in a large-scale dataset of 1.5M people are uniquely re-identifiable using 4 spatio-temporal points [de Montjoye, et al. 2013].

Direct re-identification can be avoided by only providing access to aggregate statistics, but even such systems, if not carefully designed, can leak substantial amounts of personal information. It was shown that a large number of aggregate genomic statistics could be used to determine, with

[5] As an example of a potential harm that is indirect and inferential but nevertheless substantial, Kosinski, et al. (2013) demonstrate that Facebook "likes" can be used to "automatically and accurately predict a range of highly sensitive personal attributes including: sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender. "

[6] For an analysis of the new forms of market discrimination enabled by big data see [Carlo 2013].

high statistical confidence, whether an individual was part of the population studied [Homer et al. 2008] and this led NIH to eliminate public access to these statistics [Felch 2008].
As another example, the Israel Central Bureau of Statistics provided a public internet mechanism for people to make "aggregate" statistical queries about the results of an anonymized survey, but Kobbi Nissim and Eran Tromer were able to extract records of more than one thousand individuals by querying the system and, furthermore, demonstrated that it is possible to link these records to identifiable people [Ziv 2013].

A strong new mathematical framework known as *differential privacy* can provide provable guarantees that individual-specific information will not leak (in an appropriately designed system) while still allowing for rich statistical analysis of a dataset [Dwork 2011]. There are a variety of research efforts around the country aiming to bring the mathematical work on differential privacy to practice in different contexts, and overcome various challenges that arise when doing so. See the talks by Cynthia Dwork and Salil Vadhan at the White House-MIT Big Data Privacy workshop on March 3, 2014 for more information.

4. Review, use, reporting, and information accountability

Because of the inherent identifiability and sensitivity of many types of big data, individuals would clearly benefit from regulation that provides increased transparency and accountability for the collection and subsequent use of such data. The current regulatory framework largely lacks mechanisms that would provide accountability for harm arising from misuse of disclosed data.

Consent has been, and will be a major tool for regulating information privacy. The mechanisms that are being used for consent are changing, and there is a movement in some areas towards more portable and broader consent for certain uses of information, such as research uses. [for a survey see Vayena et ala 2013]

Consent to data collection, while it should be required, is not generally sufficient protection. Privacy policies are dauntingly complex for individuals to understand, and many of the summaries provided by data collectors are inaccurate [Cranor 2012]. Restrictions on disclosure or transfer to third party and on the on types of permitted uses should be standardized and communicated.

Furthermore, data collectors will generally have to be better informed of potential and actual data use, and, in the face of ubiquitous data collection practices, consumers find it difficult to effectively withhold consent because the playing field is uneven. Thus, transparency, and accountability for misuse are all essential to achieving an optimal balance of social benefit and individual privacy protection [see e.g., Weitzner, et al. 2008]. Accountability mechanisms should enable individuals to find out where data describing them has been distributed and used, set forth penalties for misuse, and provide harmed individuals with an individual right of action to obtain their data, correct it, and take legal action for misuse.

5. Improved data sharing models

As discussed in Section 3, it is very difficult to obtain substantial privacy protections by traditional “anonymization” methods (unless one strips a dataset of almost all useful information). Simple anonymization methods such as removing PII or masking data through aggregation and perturbation individual points is generally insufficient when it comes to big data, short of rendering the data useless (see section 3).

Below we provide a categorization of data sharing models that can be used individually or combined to provide stronger privacy protections for subjects.

- *Aggregation methods protect the privacy of individual through the release of aggregated data or statistics:*
    - *Contingency tables* are tables giving the frequencies of co-occurring attributes. For example, a 3-dimensional contingency table based on Census data for Norfolk County, Massachusetts might have an entry listing how many people in the population are female, under the age of 40, and rent their home.
    - *Data visualizations* are graphical depictions of a dataset’s features and/or statistical properties. Data visualizations are especially useful for comprehending huge amounts of data, perceiving emergent properties, identifying anomalies, understanding features at different scales, and generating hypotheses [for an overview, see Ware 2012; for an example of privacy protecting visualization, see McSherry 2009].
- *Synthetic data* are “fake” data generated from a statistical model that has been developed using the original data set. Methods for generating synthetic data were first developed for filling in missing entries and are now considered attractive for protecting privacy (as a synthetic dataset does not directly refer to any “real” person) [Rubin 1993; Fienberg 1994; Abowd & Vilhuber 2008]. They are however of limited use as only the properties that have been specifically modeled are present in the synthetic dataset.
- *Multiparty computations* are electronic protocols that enable two or more parties to carry out a computation that involves both of their datasets (for example, finding out how many records are shared between the two) in such a way that no party needs to explicitly hand their dataset to any of the others. Techniques from cryptography can ensure that no party learns anything beyond the result of the computation (see the talks by Shafi Goldwasser and Vinod Vaikunathan at the White House-MIT Big Data Privacy workshop on March 3, 2014 for more information.)
- *Interactive mechanisms* are systems that enable users to submit queries about a dataset

and receive corresponding results. The dataset is stored securely and the user is never given direct access to it, but such systems can potentially allow for very sophisticated queries. For example, the Census Bureau's online Advanced Query System allows users to create their own customized contingency tables [U.S. Census Bureau 2004].

- *Personal data stores* are an approach to information handling in which individuals effectively exercise fine-grained control over where information about them is stored and how it is accessed. Thus individuals can choose to share specific personal information at specific times with specific parties. [see de Montjoye 2013 for an open implementation; and Kirkham, et al 2013 for a review] Personal data stores not only provide increased control but, as user-controlled interactive systems, are a potential foundation for developing richer accountability mechanisms, online aggregation method, and advanced security mechanisms. (See section 4)

All of the models outlined above – privacy-aware methods for contingency tables, synthetic data, data visualizations, multiparty computations, personal data stores, and interactive mechanisms – have been successfully used in practice to enable data use while enhancing privacy. As discussed in Section 3, the fact that these systems do not provide direct access to raw data does not automatically ensure privacy, and naively designed systems can leak sensitive personal information. However, each of these models can be implemented with privacy protections and, when made privacy-aware in an appropriate way, can provide strong protection. Indeed, many of these forms of data sharing have even been shown to be compatible with the strong guarantees of differential privacy (cf. Section 3).

It is clear that we would want to make some of the above forms of sharing an option for members of the public and for researchers when they would offer both better privacy and better utility than traditional anonymization approaches. At the same time, in many cases, having only a single data-sharing model will not suffice for all uses, and thus a "tiered access" framework can be valuable, and is strictly necessary where one chooses to enable all possible analyses of the information [National Research Council 2005, 2007, 2009, 2010]. For example, aggregate statistics in the form of a contingency table might be provided to the public, an interactive query system to a wide community of researchers, and raw data to a small number of analysts who pass a careful screening process.

6. Application of the general analysis framework to RFI questions

The framework and issues discussed above apply generally to the analysis of data privacy and should be considered in the report resulting from this review. Furthermore, most of the general issues raised by the RFI questions would be addressed by conducting analysis within this framework. However, for additional clarity, we review each of the questions posed in the RFI below.

- **What are the public policy implications of the collection, storage, analysis, and use of big data?**

As discussed in Section 1, the scope of big data collection is extraordinarily broad, the risks of re-identification (in a broad sense) are high, and the information associated with individuals is highly sensitive. Current regulatory frameworks are neither sufficient to minimize risk of harm, nor provide adequate accountability for those harms caused.

- **What types of uses of big data provides the most benefits and/or risks?**

Big data and computational social science are rapidly changing the study of humans, human behavior, and human institutions. Initially, the most visible aspect of this change was the advancement of statistical and data analytic methods. It is now clear that, taken as a whole, the evidence base of social science is shifting [Altman & Rogerson 2008; Lazer, et al. 2009; King 2011]. The potential benefits for research are large. At the same time, as described in section 2, the risks of misuse from this data, especially when comprehensive and well integrated (e.g. through the use of location data, as described in section 3), are equally substantial.

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

As described in section 5, there is a spectrum of emerging privacy-aware statistical and cryptographic tools that show promise in safeguarding privacy while enabling effective use. No mechanism is risk-free, and tiered access, along with accountability mechanisms (as described in section 4), will also be necessary. Law and regulation governing big data should be (re)written so as to allow for (and indeed incentivize) the use of promising new privacy-aware technologies, rather than limiting us to the use of traditional and ineffective approaches such as naive anonymization.

- **How should the policy frameworks or regulations for handling big data differ between the government and the private sector?**

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

The hundreds of laws and regulations comprising the current U.S. sectoral approach to privacy vary with respect to the scope of coverage, the definitions of personally identifiable information, the measures of sensitivity, the technical requirements, and the accountability provisions. Moreover, new regulations continue to be proposed that contain *ad hoc* privacy definitions and rules, further complicating this landscape [for a recent example, see Altman, O'Brien, & Wood

2014].

It has been widely observed that this sectoral approach to regulating information privacy results in many gaps and inconsistencies [Ohm 2010]. This approach also increases the cost and complexity of compliance. Moreover, there is clearly a substantial gap between the legal definitions of privacy, and related mathematical, normative, and empirical conceptions thereof [see, respectively, Vadhan, et al. 2010, Nissenbaum 2009, Cranor, et al. 2006]. This approach also is increasingly at odds with international laws that treat privacy as a basic human right [European Union Agency for Fundamental Rights 2014].

7. A Modern Approach to Sharing Information

Addressing privacy risks requires a sophisticated approach, and the privacy protections currently used for big data do not take advantage of advances in data privacy research or the nuances these provide in dealing with different kinds of data and closely matching sensitivity to risk. Like treatment of other risks to subjects, treatment of privacy risks should be based on a scientifically informed analysis that includes the likelihood of such risks being realized, the extent and type of the harms that would result from realization of those risks, the efficacy of computational or statistical methods to mitigate re-identification and risks and monitor access, and the availability of legal remedies to those harmed [Vadhan, et al. 2010].

A modern approach to privacy protection recognizes the following three principles:

- *The risks of informational harm are generally not a simple function of the presence or absence of specific fields, attributes, or keywords in the released set of data.* Instead, much of the potential for harm stems from what one can learn or infer about individuals from the data release as a whole or when linked with available information.

- *Redaction, pseudonymization, coarsening and hashing, are often neither an adequate nor appropriate practice, and releasing less information is not always a better approach to privacy.* As noted above, simple redaction of information that has been identified as sensitive is often not a guarantee of privacy protection.

- *Thoughtful analysis with expert consultation is necessary in order to evaluate the sensitivity of the data collected, to quantify the associated re-identification risks, and to design useful and safe release mechanisms.* Naïve use of any data sharing model, including those we describe above, is unlikely to provide adequate protection.

8. References

Abowd, John M., and Lars Vilhuber, "How Protective Are Synthetic Data?," *Privacy in*

Statistical Databases (2008): 239-49.

Altman, Micah, David O'Brien, and Alexandra Wood, Re: Proposed Rule: Improve Tracking of Workplace Injuries and Illnesses (2014), available at http://www.regulations.gov/#!documentDetail;D=OSHA-2013-0023-1207.

Altman, Micah, and Michael McDonald, "Public Participation GIS: The Case of Redistricting," *Proceedings of the 47th Annual Hawaii International Conference on System Sciences* (2014).

Altman, Micah, and Kenneth Rogerson, "Open Research Questions on Information and Technology in Global and Domestic Politics – Beyond 'E-,'" *PS Political Science and Politics*, 41.4 (2008), 1-8.

Altman, Micah, et al., *Principles for Transparency and Public Participation in Redistricting* (Washington: Brookings, 2010).

Altman, Micah, and Michael McDonald, "The Promise and Perils of Computers in Redistricting," *Duke Journal of Constitutional Law & Public Policy* 5 (2010).

Bankert, Elizabeth A., and Robert J. Andur, *Institutional Review Board: Management and Function* (Boston: Jones and Bartlett, 2006).

Ryan, Carlo M. "Digital Market Manipulation." University of Washington School of Law Research Paper 2013-27 (2013).

Cranor, Lorrie Faith, "Necessary but Not Sufficient: Standardized Mechanisms for Privacy Notice and Choice," *Journal on Telecommunications & High Technology Law* 10 (2012): 273.

de Montjoye, Yves-Alexandre, et al., "Unique in the Crowd: The Privacy Bounds of Human Mobility." *Nature Scientific Reports* 3 (2013).

de Montjoye, Yves-Alexandre, et al. "On the Trusted Use of Large-Scale Personal Data." IEEE Data Eng. Bull. 35.4 (2012): 5-8. and http://openpds.media.mit.edu/

Dwork, Cynthia, "A Firm Foundation for Private Data Analysis," *Communications of the ACM* (2011): 1, 86-95.

European Union Agency for Fundamental Rights (FRA), *Handbook of European Data Protection Law* (Luxembourg: Publications Office of the European Union, 2014).

Felch, J., "DNA databases blocked from the public," *Los Angeles Times,* page A31 (29 August 2008).

Fienberg, Stephen E., "Conflicts Between the Needs for Access to Statistical Information and Demands for Confidentiality," *Journal of Official Statistics* 10 (1994): 115-32.

Fung, Benjamin C.M., et al., *Introduction to Privacy-preserving Data Publishing: Concepts and Techniques* (Boca Raton: CRC Press, 2010).

Gonzalez, Marta C., Cesar A. Hidalgo, and Albert-Laszlo Barabasi, "Understanding Individual Human Mobility Patterns," *Nature* 453.7196 (2008): 779-782.

Homer N, Szelinger S, Redman M, Duggan D, Tembe W, et al., “Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays,” *PLoS Genetics* 4(8): e1000167. doi:10.1371/journal.pgen.1000167 (2008).

Kenthapadi, Krishnaram, Nina Mishra, and Kobbi Nissim, “Denials Leak Information: Simulatable Auditing,” *Journal of Computer and System Sciences* 79.8 (2013): 1322-1340.

Kirkham, Tom, Sandra Winfield, Serge Ravet, and Sampo Kellomaki. "The Personal Data Store Approach to Personal Data Security." Security & Privacy, IEEE 11, no. 5 (2013): 12-19.

King, Gary, “Ensuring the Data Rich Future of the Social Sciences,” *Science* 331.11 (Feb. 2011): 719-721.

King, Gary, "Replication, Replication," *PS: Political Science & Politics* 28.03 (1995): 444-452.

Kosinski, Michal, David Stillwell, and Thore Graepel. "Private traits and attributes are predictable from digital records of human behavior." Proceedings of the National Academy of Sciences 110.15 (2013): 5802-5805.

Lazer, David, et al., "Life in the Network: The Coming Age of Computational Social Science," *Science* 323.5915 (2009): 721.

Lee, Raymond M., *Doing Research on Sensitive Topics* (London: SAGE, 1993).

McSherry, Frank D., "Privacy Integrated Queries: An Extensible Platform for Privacy-preserving Data Analysis," *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data* (2009).

National Research Council, *Proposed Revisions to the Common Rule for the Protection of Human Subjects in the Behavioral and Social Sciences* (Washington: National Academies Press, 2014).

National Research Council, *Conducting Biosocial Surveys: Collecting, Storing, Accessing, and Protecting Biospecimens and Biodata* (Washington: National Academies Press, 2010).

National Research Council, *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health through Research* (Washington: National Academies Press, 2009).

National Research Council, *Putting People on the Map: Protecting Confidentiality with Linked Social-Spatial Data* (Washington: National Academies Press, 2007).

National Research Council, *Expanding Access to Research Data: Reconciling Risks and Opportunities* (Washington: National Academies Press, 2005).

Richards, Neil. "The Dangers of Surveillance." Harvard Law Review (2013).

Rubin, Donald B., “Discussion of Statistical Disclosure Limitation,” *Journal of Official Statistics* 9 (1993), at 461-68.

Solove, Daniel J. A Brief History of Information Privacy Law in PROSKAUER ON PRIVACY, PLI (2006).

Sweeney, Latanya, Akua Abu, and Julia Winn, “Identifying Participants in the Human Genome Project by Name,” Data Privacy Lab, IQSS, Harvard University (2013).

Sweeney, Latanya, “Uniqueness of Simple Demographics in the US Population,” Technical report, Carnegie Mellon University, Data Privacy Lab, Pittsburgh, PA (2000).

Sweeney, Latanya, “Weaving technology and policy together to maintain confidentiality,” *Journal of Law, Medicine and Ethics* 25 (1997).

U.S. Census Bureau, *Census Confidentiality and Privacy: 1790-2002* (2004), available at http://www.census.gov/prod/2003pubs/conmono2.pdf.

Vadhan, Salil, et al., Re: Advance Notice of Proposed Rulemaking: Human Subjects Research Protections (2010), available at http://dataprivacylab.org/projects/irb/Vadhan.pdf.

Vayena, Effy, Anna Mastroianni, and Jeffrey Kahn. "Caught in the web: informed consent for online health research." Science translational medicine 5.173 (2013): 173fs6-173fs6.

Ware, Colin, *Information Visualization: Perception for Design,* 3rd ed. (Boston: Morgan Kaufmann, 2012).

Weitzner, Daniel J., et al., “Information Accountability,” *Communications of the ACM* 51.6 (2008): 82-87.

Zimmerman, Dale L., and Claire Pavlik, “Quantifying the effects of mask metadata disclosure and multiple releases on the confidentiality of geographically masked health data,” *Geographical Analysis* 40.1 (2008): 52.

Ziv, Amitai, “Israel's ‘Anonymous’ Statistics Surveys Aren't So Anonymous,” *Haaretz* (Jan. 7, 2013), http://www.haaretz.com/news/national/israel-s-anonymous-statistics-surveys-aren-t-so-anonymous-1.492256.