CENTER FOR DATA INNOVATION

August 5, 2014

Mr. John Morris
National Telecommunications and Information Administration
U.S. Department of Commerce
1401 Constitution Avenue NW
Room 4725
Attn: Privacy RFC 2014
Washington, DC 20230

RE: Big Data and Consumer Privacy in the Internet Economy, Docket No. 140514424–4424–01

Dear Mr. Morris,

On behalf of the Center for Data Innovation (www.datainnovation.org), I am pleased to submit these comments in response to the National Telecommunications & Information Administration's (NTIA) request for comments on big data and consumer privacy in the Internet economy.[1] The Center for Data Innovation is a non-profit, non-partisan, Washington, D.C.-based think tank focusing on the impact of the increased use of information on the economy and society. In addition, the Center formulates and promotes pragmatic public policies designed to enable data-driven innovation in the public and private sectors, create new economic opportunities, and improve quality of life. The Center is affiliated with the Information Technology and Innovation Foundation (ITIF).

The White House report "Big Data: Seizing Opportunities, Preserving Values" and the President's Council of Advisors on Science and Technology's (PCAST) report "Big Data and Privacy: A Technological Perspective" correctly recognize the enormous potential economic and social benefits of data as well as the fundamental difficulties involved in regulating data collection.[2] However, the White House report focuses disproportionately on potential future harms to consumers and suggests a number of unnecessary restrictions on data collection that would keep companies and government agencies from maximizing the benefits of big data analytics. The Consumer Privacy Bill of Rights, in its current form, is one such set of restrictions, and it is incompatible with attempts to promote more

[1] "Request for Comments on Big Data and Consumer Privacy in the Internet Economy," National Telecommunications & Information Administration, June 4, 2014, http://www.ntia.doc.gov/federal-register-notice/2014/request-comments-big-data-and-consumer-privacy-internet-economy.
[2] John Podesta et al., "Big Data: Seizing Opportunities, Preserving Values," Executive Office of the President, May 2014, http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf.

data-driven innovation.[3] Instead of attempting to impose broad regulations limiting data collection and reuse, policymakers should attempt to narrowly restrict concrete examples of harmful uses of data while otherwise encouraging data reuse. Establishing a responsible use framework not only encourages data-driven innovation, but it also protects individuals even if their data becomes public due to a data breach.

> Any policies intended to preserve privacy use should respect the fact that useful analysis depends on quality data as well as the need to explore and reuse data. In some cases, this may involve creating policies to discourage individuals from opting out of having their data collected (e.g. for certain public health data sets) and allowing organizations to use data for purposes not originally specified. When possible, policymakers should address potential misuses of data by drawing from the large existing body of law protecting individuals from discrimination and other abuses. In addition, government agencies, including the NTIA, should work together to develop a roadmap for researching and developing privacy-enhancing technologies, including best practices around de-identification.

Should any of the specific elements of the Consumer Privacy Bill of Rights be clarified or modified to accommodate the benefits of big data? Should any of those elements be clarified or modified to address the risks posed by big data?

The ability to innovate with data depends on at least three factors: data quality, data exploration, and data reuse. First, the data set must be of sufficient quality to permit analysis. This depends both on the accuracy of individual records and the representativeness of the full data set. Second, it is essential to be able to explore a new data set freely to determine how each variable is represented, calculate the statistical characteristics of the variables, and consider what questions further analysis might be able to answer. This requires sufficient freedom to probe the data set and test hunches that may not contribute to the final analysis. Finally, it is essential to be able to reuse data when new questions arise that the data set could answer or when the addition of new data makes further analysis possible. This requires the data set to be retained and applications beyond the original use permitted.

Three major elements of the Consumer Privacy Bill of Rights, individual control, respect for context, and focused collection, fail to take these factors into account.

[3] "Consumer Data Privacy in a Networked World," The White House, February 2012, http://www.whitehouse.gov/sites/default/files/privacy-final.pdf.

First, the principle of individual control says that consumers should have a right to exercise control over what personal data companies collect from them and how they use it. While many companies provide consumers the ability to do this, it does not always make sense for companies to spend the time and resources to develop tools for more granular control. Instead, some companies may simply prefer that consumers opt not to use a given product or service (and thus decide not to have data collected about them). In other cases, it may be desirable to not allow individuals to opt out of data collection, such as contagious disease reporting. When consumers opt out of data collection, the data becomes less representative in potentially unpredictable ways, reducing the value of the data set and making it more difficult to derive accurate conclusions. Incomplete data sets are of less value than complete ones.

Second, the principle of focused collection, which says that consumers have a right to reasonable limits on the personal data that companies collect and retain, would make data exploration less useful. It is not always possible to know ahead of time what insights a new data set might contain and exploratory analysis can offer some clues. Data exploration can help generate hypotheses and reveal the importance of variables that might not have been obvious before exploration. In this way, the principle of focused collection gets the established process of exploratory data analysis backwards: analysts do not specify a particular insight they want to glean and then collect data that might yield it, they collect data first and then determine what insights might come from it.

Third, the principle of respect for context would make data reuse impractical or impossible. Organizations often use a single data set to answer questions about a variety of variables, and it is inefficient to require them to collect or obtain the same data multiple times for different analytical inquiries. As the open data movement has shown, there is enormous social and economic value in allowing data sets to be reused for multiple purposes.[4] Moreover, it is unreasonable to expect analysts to anticipate all potential future uses of a data set, since the addition of new data in the future could enable previously impossible analysis.

Should a responsible use framework, as articulated in Chapter 5 of the Big Data Report, be used to address some of the challenges posed by big data? If so, how might that framework be embraced within the Consumer Privacy Bill of Rights? Should it be? In what contexts would such a framework be most effective? Are there limits to the efficacy or appropriateness of a responsible use framework

[4] See for example, "The Economic Impact of Open Data," Center for Data Innovation, April 8, 2014, http://www.datainnovation.org/2014/04/the-economic-impact-of-open-data/ and "The Social Impact of Open Data," Center for Data Innovation, July 23, 2014, http://www.datainnovation.org/2014/07/the-social-impact-of-open-data/.

in some contexts? What added protections do usage limitations or rules against misuse provide to users?

Policymakers should embrace a responsible use framework. This would serve the dual purposes of helping hold companies accountable for harmful uses of data and leaving them free to pursue beneficial applications. A responsible use framework would put fears of unspecific future harms to rest while allowing companies to innovate and develop useful data applications far beyond what regulators can anticipate.

Industry groups, such as the Digital Advertising Alliance, have begun to develop self-regulatory guidelines specifying how data should be used in advertising, and these types of efforts can serve as a model for future initiatives to establish self-regulatory guidelines in other industries.[5]

However, the responsible use framework is not compatible with the Consumer Privacy Bill of Rights as it is currently formulated, as the former allows data collection and reuse while the latter discourages those practices. The Consumer Privacy Bill of Rights principle of focused collection would restrict collection without even pointing to specific harms, undermining companies that collect data for beneficial purposes. The principle of respect for context would impede beneficial reuses of data, also without an eye toward any harm in particular. To add responsible use provisions to a set of principles already limiting collection and restricting reuse would be overly burdensome, as many companies doing beneficial work would be unable to use data in the first place.

Is there existing research or other sources that quantify or otherwise substantiate the privacy risks, and/or frequency of such risks, associated with big data? Do existing resources quantify or substantiate the privacy risks, and/or frequency of such risks, that arise in non-big data (''small data'') contexts? How might future research best quantify or substantiate these privacy risks?

The Center for Data Innovation undertook a review of all of the harms from big data identified in the previously mentioned White House report on big data.[6] We found that even though many commentators had expressed concern about big data, the report failed to identify almost any concrete examples of how big data is actually causing consumers economic, physical, or social harm. In fact, after reviewing all 37 concerns identified in the report, the Center found that all but two of

[5] "Application of Self-Regulatory Principles to the Mobile Environment," Digital Advertising Alliance, July 2013, http://ww.aboutads.info/DAA_Mobile_Guidance.pdf.
[6] Podesta et al.

them were purely speculative, i.e., the authors cited no evidence that the concerns mentioned were occurring today, and many were vague and ill-defined.[7]

This is a crucial point. If the White House had identified a broad series of tangible examples of how big data was presently harming consumers, then it would be legitimately justified in calling for policymakers to adopt comprehensive consumer privacy rules. But since it did not, this raises the question of whether there is even a compelling need for policy intervention at this stage. After all, many theoretical concerns may never be realized if factors, such as market forces, cultural norms, and new technologies, intervene. Thus policymakers should be extremely cautious about trying to regulate on the basis of purely speculative concerns which might not even come to pass, especially when doing so might curtail substantial economic and social benefits, many of which are already occurring today.

Of the two tangible harms the report cited, the first was as follows (pp. 7-8):

> *Unfortunately, "perfect personalization" also leaves room for subtle and not-so-subtle forms of discrimination in pricing, services, and opportunities. For example, one study found web searches involving black-identifying names (e.g., "Jermaine") were more likely to display ads with the word "arrest" in them than searches with white-identifying names (e.g., "Geoffrey"). Outcomes like these, by serving up different kinds of information to different groups, have the potential to cause real harm to individuals.*

Clearly, it is harmful for advertisements to reinforce negative stereotypes about marginalized groups. But given that advertisers frequently come under criticism for accusations of racism, sexism, and ageism, this concern may have less to do with big data than it does with the conduct of advertisers. To address this issue, the advertising industry, as well as individual advertisers and ad platforms, could adopt their own self-regulatory guidelines and company policies banning the negative portrayal of a particular race or ethnicity.

The second example concerns retailers offering different consumers different prices for the same goods (pp. 46-47):

> *Recently, some offline retailers were found to be using an algorithm that generated different discounts for the same product to people based on where they believed the customer was located. While it may be that the price differences were driven by the lack of competition in*

[7] Daniel Castro and Travis Korte, "A Catalog of Every 'Harm' in the White House Big Data Report," The Center for Data Innovation, July 15, 2014, http://www.datainnovation.org/2014/07/a-catalog-of-every-harm-in-the-white-house-big-data-report/.

> *certain neighborhoods, in practice, people in higher-income areas received higher discounts than people in lower-income areas.*

Ironically, one of the solutions to the above problem is actually more data. As the original *Wall Street Journal* article noted, one likely reason for the lower prices in higher-income areas was greater competition in well-off suburban neighborhoods. If low-income shoppers had more access to data about the prices paid by others, they could make better informed decisions about where to buy and thereby create a more competitive market.

In short, the White House report identified only two concrete consumer harms from big data, neither of which would justify new privacy laws. Rather than preemptively trying to curtail the use of data, policymakers would be better off narrowly focusing on identifying specific consumer concerns and then constructing targeted remedies to address those particular problems.

The PCAST Report states that in some cases ''it is practically impossible'' with any high degree of assurance for data holders to identify and delete ''all the data about an individual'' particularly in light of the distributed and redundant nature of data storage. Do such challenges pose privacy risks? How significant are the privacy risks, and how might such challenges be addressed? Are there particular policy or technical solutions that would be useful to consider? Would concepts of ''reasonableness'' be useful in addressing data deletion?

The PCAST report is correct. It can be difficult to completely delete individual records from messy data that has been pieced together from multiple heterogeneous sources. In addition, organizations may back up data in secondary archives that survive even after records are removed from primary databases.

In part because this condition is inevitable among organizations with complex data systems, deleting data is not the best way to forestall any future harms that might arise through data misuse. A better strategy is placing limits on uses that have been deemed harmful, so that even if some data remains after attempts to delete it, individuals will still be protected.

Having to delete data can prevent organizations from developing beneficial applications in the future. Data reuse, which refers to analysis that leverages data beyond the purpose for which it was originally collected, is a crucial technique in sectors from medical research to particle physics.[8] Data

[8] James Cimino, "Collect Once, Use Many: Enabling the Reuse of Clinical Data through Controlled Terminologies," AHIMA, February 2007, http://library.ahima.org/xpedio/groups/public/documents/ahima/bok1_033473.hcsp?dDocName=bok1_033473.

reuse is also a hallmark of longitudinal research, which is only possible when organizations retain data for years or decades. For example, the Framingham Heart Study, a long-term cardiovascular study that has been active since 1948, has recently been applied to studies of obesity and divorce.[9]

Because it is often hard to know at the time of collection what insights might eventually be drawn from a given data set, particularly those that might become possible with additional data in the future, deleting data can mean having to acquire it anew for future applications. This includes even the most unobjectionable applications, such as using old data to refine website designs or recommendation algorithms, or making individuals' own data available to them in perpetuity. Just as in the open source software community, new data applications often build on prior work. Forcing organizations to collect the same data repeatedly is inefficient and unnecessary, particularly given that complete data deletion is not always technically feasible. Placing limits on harmful uses of data would protect individuals from misuse while enabling the reuse at the core of so many beneficial applications.

How significant are the privacy risks posed by unindexed data backups and other ''latent information about individuals?'' Do standard methods exist for determining whether data is sufficiently obfuscated and/or unavailable as to be irretrievable as a practical matter?

As discussed above, it is inevitable that data will sometimes remain even after deletion efforts. Similarly, data breaches can unfortunately expose private information. De-identification standards and guidelines can help reduce the risk of misuse, and the federal government should support efforts to develop these best practices. The National Institute of Standards and Technology (NIST) has previously worked to develop a wide range of cryptographic standards which has helped standardize and improve information security in both government and the private sector. It should make similar efforts to develop best practices for de-identification techniques.

The government should also encourage a risk-based approach to de-identification, encouraging organizations sharing sensitive information to have a higher threshold for de-identification than those working with non-sensitive information, as well as promoting the use of non-disclosure

Maggie McKee, "Does a new particle lurk in data from sleeping LHC?" NewScientist, June 11, 2014, http://www.newscientist.com/article/dn25710-does-a-new-particle-lurk-in-data-from-sleeping-lhc.html#.U7xlZvldXng.

[9] Rich Morin, "Is divorce contagious?" Pew Research Fact Tank, October 21, 2013, http://www.pewresearch.org/fact-tank/2013/10/21/is-divorce-contagious/.

Gina Kolata, "Study Says Obesity Can Be Contagious," The New York Times, July 25, 2007, http://www.nytimes.com/2007/07/25/health/25cnd-fat.html.

agreements, audits, and other safeguards to reduce risk.[10] One way to encourage greater adoption of de-identification techniques is through data breach notification laws. Federal and state data breach laws should exempt organizations from notification requirements if the data that is stolen or accidentally released had previously been de-identified. This would encourage organizations to invest greater resources into developing these techniques, resulting in better privacy and security for consumers.

As the PCAST Report explains, ''it is increasingly easy to defeat [de- identification of personal data] by the very techniques that are being developed for many legitimate applications of big data.'' However, de-identification may remain useful as an added safeguard in some contexts, particularly when employed in combination with policy safeguards. How significant are the privacy risks posed by re-identification of de- identified data? How can de- identification be used to mitigate privacy risks in light of the analytical capabilities of big data? Can particular policy safeguards bolster the effectiveness of de-identification? Does the relative efficacy of de-identification depend on whether it is applied to public or private data sets? Can differential privacy mitigate risks in some cases? What steps could the government or private sector take to expand the capabilities and practical application of these techniques?

The risk of re-identification of individuals from properly de-identified data is much lower than is often reported in the popular press.[11] In particular, low-dimensional data, which includes few variables, can often be de-identified in a manner that mathematically precludes re-identification.[12] In the case of high-dimensional data with many variables, where the possibility of re-identification is higher, strict access controls could help reduce the risk of re-identification.

Regardless of the quality of de-identification, however, data breaches occur frequently both now and for the foreseeable future. This means that some personally identifiable information will become public regardless of how effectively organizations de-identify data they release intentionally.[13] The best approach to this eventuality is not attempting to individually de-identify all the world's data sets, but rather formulating use restrictions to ensure that consumers are protected from harm even when their data becomes public through data breaches or improper de-identification. Conversely, data

[10] Khaled El Emam, "Risk-Based De-Identification of Health Data," Privacy Interests, May/June 2010, http://www.privacyanalytics.ca/wp-content/uploads/2013/12/riskdeid.pdf.
[11] Ann Cavoukian and Daniel Castro, "Big Data and Innovation, Setting the Record Straight: De-identification Does Work," June 16, 2014, http://www2.itif.org/2014-big-data-deidentification.pdf.
[12] Ibid
[13] Dan Lohrmann, "Are Data Breaches Inevitable?" Government Technology, February 11, 2013, http://www.govtech.com/pcio/Are-Data-Breaches-Inevitable-.html.

collection restrictions cannot prevent these incidents, and offer no consumer protections when they occur.

The Big Data Report concludes that ‘‘big data technologies can cause societal harms beyond damages to privacy, such as discrimination against individuals and groups’’ and warns ‘‘big data could enable new forms of discrimination and predatory practices.’’ The Report states that ‘‘it is the responsibility of government to ensure that transformative technologies are used fairly’’ and urges agencies to determine ‘‘how to protect citizens from new forms of discrimination that may be enabled by big data technologies.’’ Should the Consumer Privacy Bill of Rights address the risk of discriminatory effects resulting from automated decision processes using personal data, and if so, how? How could consumer privacy legislation (either alone or in combination with anti-discrimination laws) make a useful contribution to addressing this concern? Should big data analytics be accompanied by assessments of the potential discriminatory impacts on protected classes?

Numerous existing laws already address discrimination of individuals based on various characteristics, and there is no reason these laws cannot be applied to discrimination of those characteristics as represented in data. Title VII of the Civil Rights Act of 1964, the Equal Pay Act, the Fair Credit Reporting Act, the Age Discrimination Employment Act, the Americans with Disabilities Act, and the Genetic Information Nondiscrimination Act are just a few examples of laws that contain antidiscrimination provisions. These laws are, in essence, use restrictions on information, so they are naturally applicable to discrimination based on data. They also extend naturally to algorithmically generated models; if a particular variable is prohibited from informing a human’s employment decision, it should also be excluded from models that do the same job. To prevent algorithms from using variables that may be proxies for a prohibited variable, such as using an individual’s zip code as a proxy for race, enforcement agencies could ensure that organizations not use models that make predictions or recommendations correlating above a certain threshold with membership in a protected class. To the extent that new types of discrimination are identified that are not covered under existing laws, they can be addressed through use restrictions modeled after the earlier legislative efforts.

It may be reasonable to require organizations to assess potential discriminatory impacts, and such provisions already exist in some areas. For example, the Equal Opportunity Employment Commission's (EEOC) Uniform Guidelines on Employee Selection Procedures state that employers should maintain information on whether their employee selection procedure for a particular job has an adverse impact on any protected class and update this information at least annually for groups that represent a significant portion of the workforce.[14] This can be straightforwardly applied to employee selection procedures that use data. If an employer takes online influence metrics—such as those offered by social media analytics company Klout—into account, and this process has an adverse impact on certain groups because it is correlated with access to the Internet and is not otherwise central to the job requirements, the employer would need to reassess its use of this method.

Moreover, data-driven automated systems can and should be used to reduce discrimination. For example, a properly tuned home loan evaluation algorithm could be made much fairer than a human loan officer; by definition, a successful loan evaluation algorithm should award loans to the most deserving and ignore concerns like race, religion, or sexual orientation that cloud human judgment. Even in areas where automated decisions are the norm today, such as demographic assessment in insurance pricing, data-driven systems could offer improvement; pricing insurance based on the output of a car-bound sensor that measures driving ability is undeniably less discriminatory than pricing it based on the risk associated with a particular demographic profile.

Can emerging privacy enhancing technologies mitigate privacy risks to individuals while preserving the benefits of robust aggregate data sets?

Yes. There exists a variety of technologies that could have wide-reaching implications for protecting consumer privacy, although in many cases further research will be required to capture the technologies' full benefits. For example, research on differential privacy could offer insight into what a given dataset could possibly reveal, without revealing sensitive personal information; better algorithmic and statistical approaches to de-identifying data could preserve the statistical characteristics of some data sets while ensuring that re-identification is very difficult or impossible. Similarly, additional research into computer-readable privacy policies could result in the ability to create policies bound to data so that, for example, data that has been de-identified stays de-identified.[15] Or additional research on chains of trust could help keep track of multiple parties sharing data, such as in cloud-based systems.

In order to coordinate this research and establish a shared vision for its development, the National Telecommunications and Information Administration should work with the National Science Foundation, the National Institute of Standards and Technology, and the Federal Trade Commission,

[14] Uniform Guidelines on Employee Selection Procedure, Section 15-A, EEOC, 1978, http://www.uniformguidelines.com/uniformguidelines.html.

[15] Rakesh Agrawal, et al., "Hippocratic Databases," Proceedings of the 28th VLDB Conference, Hong Kong, China, 2002, http://www.utdallas.edu/~muratk/courses/course_material/hippocratic.pdf.

to develop a roadmap for privacy R&D.[16] This would establish a set of priorities shared among multiple stakeholders and help ensure that the highest-priority research receives funding.

What other approaches to big data could be considered to promote privacy?

In many cases, algorithms and automated systems can promote privacy by reducing human access to information. For example, many common privacy intrusions come from interactions with humans, such as doctor's visits or in-person retail purchases. Because unauthorized information disclosure is such a salient issue in fields that involve sensitive information, such as medicine and law, professionals in these fields have had to establish elaborate norms for preventing it. However, even with these safeguards, people are still wary of telling their doctors embarrassing information. Research indicates that people may disclose medical information more honestly and openly to a computer than to a doctor.[17] Increasing automation in transactions that involve sensitive information would reduce human-to-human contact and thereby reduce both feelings of embarrassment and risk of future disclosure.

What other questions should we be asking about big data and consumer privacy?

One key question policymakers must ask themselves is: what are the net social benefits that personal privacy concerns are impeding? For example more medical data sharing and fewer portability restrictions would save lives.[18] Google CEO Larry Page has estimated that 100,000 lives per year are lost because of fears about data mining.[19] In advocating for strong collection and portability restrictions around medical data, policymakers implicitly make the case that abstract fears about potential future harms outweigh demonstrably lifesaving benefits. Policymakers should ask themselves: is it worth it?

[16] Daniel Castro, "The Need for an R&D Roadmap for Privacy," The Information Technology and Innovation Foundation, August 2012, http://www2.itif.org/2012-privacy-roadmap.pdf.

[17] Tom Jacobs, "I'd Never Admit That to My Doctor. But to a Computer? Sure," Pacific Standard, June 20, 2014, http://www.psmag.com/navigation/health-and-behavior/id-never-admit-doctor-computer-sure-84001/ and Richard Feinberg and Kathy Walton, "The Computers Are Coming: A Study of Human-Computer Social Interaction," Home Economics Research Journal, 1883, 11: 319–326. doi: 10.1177/1077727X8301100401.

[18] Jeremy Farrar, "Sharing NHS Data Saves Lives; EU Obstruction Will Not," The Telegraph, January 14, 2014, http://www.telegraph.co.uk/health/nhs/10569467/Sharing-NHS-data-saves-lives-EU-obstruction-will-not.html.

[19] Alex Hern, "Google: 100,000 Lives a Year Lost Through Fear of Data-Mining," The Guardian, June 26, 2014, http://www.theguardian.com/technology/2014/jun/26/google-healthcare-data-mining-larry-page.

They might ask themselves, “how many preventable deaths would outweigh my fears?” If the answer is that no number is high enough, they should reconsider their logic. Of course government and industry should strive to prevent data from being used in harmful ways, but as long as fear alone is not a harm unto itself, it must be taken into account only in proportion to the benefits it obstructs. Ultimately, policymakers should be paying as much attention to ensuring they are realizing the potential benefits of data-driven innovation as they are to reducing the potential harms associated with data.

Sincerely,

Daniel Castro

Director, Center for Data Innovation
1101 K Street NW, Suite 610
Washington, DC 20005

dcastro@datainnovation.org