3108 Fifth Avenue
Suite B
San Diego, CA 92103

Comments of the World Privacy Forum
To: Office of Science and Technology Policy
Re: Big Data Request for Information

Via email to bigdata@ostp.gov

Big Data Study,
Office of Science and Technology Policy,
Eisenhower Executive Office Building,
1650 Pennsylvania Ave. NW.,
Washington, DC 20502

March 31, 2014

[In response to Question 2, FR Doc. 2014-04660, March 4, 2014. <https://www.federalregister.gov/articles/2014/03/04/2014-04660/government-big-data-request-for-information>.]

Thank you for the opportunity to respond to the Request for Information regarding the Big Data Study. The World Privacy Forum is a non-profit, non-partisan public interest research group. We focus on in-depth research on privacy matters in several key areas, including large datasets. More information about our work is available at www.worldprivacyforum.org.

Privacy must be re-imagined for a digital era. We are in the midst of a time in which complex data flows involving large data sets are not just occasional, but commonplace. Big data has many uses, including positive ones. In analyzing the privacy impact of big data, we see a range of issues, but we continue to believe that Fair Information Practices should be the bedrock of any policy for large datasets. Among the many possibilities that big data presents, three issues in particular stand out to us as important focal points:

- The quality of the predictive capacity of the data
- The appropriateness of the uses of the data sets

- Handling problems arising from analysis, including in vulnerable populations

An underlying issue for all big data discussions is the identifiability level of the data sets, and if de-identified, the probabilities for re-identification of the data sets.

Data Quality

Good predictions require good data quality. Current datasets may be easy to collect, but inaccuracies may be significant and variable across datasets and can be made worse if data linkages are of poor quality. In looking at aggregated data, inaccuracy may be less of an issue. But if large datasets are used to make decisions around individuals, particularly identifiable individuals, then errors stemming from either underlying factors or analytic model error rate can be problematic and deserve policy attention. Privacy principles that call for data destruction (or de-identification) and tying data uses to original purposes remain important. Large datasets cannot be exempt from data quality principles. Ultimately, high data quality is good for all parties involved.

Appropriateness of data uses

In a groundbreaking series of articles, the Associated Press used EPA data[1] to map the air quality risk scores for every neighborhood in the U.S. The AP mapped existing EPA toxicity risk scores to socio-economic and racial factors for each neighborhood from the 2000 Census to determine who was breathing the dirtiest air in America. The headlines across the country read, in some variation, that minorities suffer most from industrial pollution.[2]

The results established important understandings about neighborhoods and toxicity, and the resulting snapshot of where and how factory pollution affected neighborhoods and people was deservedly much-discussed. These results are examples of an informative and beneficial use of what today would be called large datasets or "big data."

It is helpful that the EPA has a set of meaningful best practice guidelines for analyzing its data in the EPA Risk Characterization Handbook. It discusses EPA's use of risk characterizations in some detail. The EPA analysis is valuable here:

---

[1] See <http://www.epa.gov/risk/health-risk.htm>. The EPA data in this instance help screen for polluted areas in the U.S. that may need additional study and vetting for potential human health problems.

[2] David Pace, *More Blacks Live With Pollution*, Associated Press (Dec. 13, 2005), <http://onlinenews.ap.org/work/pollution/wrap.py?story=./linked_story/part1.html>. See also http://www.nbcnews.com/id/10452037/ns/us_news-environment/t/minorities-suffer-most-industrial-pollution/>. The EPA uses toxic chemical air releases reported by factories to calculate risk for each square kilometer of the United States. The scores allow comparing risks from long-term exposure to factory pollution from one area to another. The scores are based on: the amount of toxic pollution released by each factory; the path the pollution takes through the air; the level of danger to humans posed by each different chemical released; and the number and ages of males and females living in the exposure paths.

> "Risk characterizations should clearly highlight both the confidence and the uncertainty associated with the risk assessment. For example, numerical risk estimates should always be accompanied by descriptive information carefully selected to ensure an objective and balanced characterization of risk in risk assessment reports and regulatory documents."[3]

The EPA also created excellent documentation on how the analysis of its own data is to be used.[4] The documentation is for its own researchers, but its quality suggests broader applications are appropriate.

It stated, in part:

> "The methods used for the analysis (including all models used, all data upon which the assessment is based, and all assumptions that have a significant impact upon the results) are to be documented and easily located in the report. This documentation is to include a discussion of the degree to which the data used are representative of the population under study. Also, this documentation is to include the names of the models and software used to generate the analysis. Sufficient information is to be provided to allow the results of the analysis to be independently reproduced."[5]

These recommendations should also apply to large data sets applicable to other areas impacting consumers. Usage guidelines like EPA's, plus guidelines which discuss identifiability of consumers, create important fairness benchmarks for many of the uses and applications of large datasets. These benchmarks would go toward improving privacy protections for other big data activities.

Handling problems arising from analysis -- vulnerable populations

When problems are uncovered using big data analysis, careful application of the information is necessary. For example, policies that would mandate identifying and protecting victims of abuse, or other crimes, could have an unfortunate reverse effect. No one wants to create a readily accessible list of identifiable or semi-identifiable victims of abuse, while at the same time, the promise of a proper analysis to pinpoint aid distribution and assistance in a timely way to those who need it most would be welcome. The tension here is real, and we have to acknowledge it and resolve it in a balanced way.

We suspect that different applications of large datasets to different populations will warrant slightly different approaches. Again, we are most concerned about privacy-related challenges in the use of big data when the data sets can be re-identified back to

---

[3] U.S. Environmental Protection Agency, Science Policy Council, Risk Characterization Handbook (. December 2000), <http://www.epa.gov/spc/pdfs/rchandbk.pdf >.p. A5.
[4] U.S. Environmental Protection Agency, *Policy for Use of Probabilistic Analysis in Risk Assessment*, (May 15, 19970, <http://www.epa.gov/spc/pdfs/probpol.pdf>.
[5] Id, p. 2.

specific vulnerable consumer groups, or when the data sets are sensitive and are, or can become, personally identifiable to individuals.

Research is needed to understand how vulnerable populations in particular are affected by analysis and predictions based on such data, and what systematic biases could be potentially introduced into algorithms through faulty data and assumptions. In some cases, even loosely aggregate data has proven problematic.

In working to ensure beneficial uses of large datasets in vulnerable or sensitive areas while mitigating potential harm, we share several thoughts.

Of assistance in determining large dataset policy in identifiable datasets is the Common Rule[6] for protection of human subjects of research, and the Belmont Report[7] regarding Ethical Principles and Guidelines for the Protection of Human Subjects of Research. Informational risks in research must be measured against a firm standard, one that is not affected by every change in technology or commercial practice. For example, the HIPAA privacy standard establishes a firm set of Fair Information Practices. While there is considerable flexibility in the application of the HIPAA privacy rules in some contexts, the standards themselves are not subject to change because of external factors. Patients can expect the HIPAA standards to protect their health information in the same way.

The same should be true for human subjects research, which despite the size of a large dataset containing identifiable individuals, is still research and analysis applicable to individuals. The need for a baseline of privacy protection must be a constant for research even though the degree of informational risk can vary from project to project. The need for rules governing collection, use, and disclosure is constant. The need for openness and accountability is a constant. The need to consider individual participation rights (access and correction) is a constant. Thus, whatever the risk involved in a given project, the need for sufficient privacy protections for personally identifiable information is a constant.

Looking at this issue of identifiable data with more specificity would include for example, ensuring that recourse for discovery of accuracy-related problems is built in to the process. We are interested in policies that develop overall good practices in this area. Accuracy and recourse for correction for individuals identified in health care datasets is a foundational area for further inquiry. Some big data activities have been a part of health and other research for a long time, and there is nothing new in some respects. The demands of researchers can overwhelm existing institutions (like institutional review boards) that do not have the necessary privacy or security expertise.

We support the use of large datasets in medical research, but researcher obligations to protect data and to protect vulnerable populations from problems resulting from analysis need to be defined in law. Any disclosure for health research in large datasets should be

[6] < 45 CFR part 46, Subpart A-D. <http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html>.
[7] <http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html>.

limited by law, regulation, and contract as appropriate.[8] HIPAA requirements that protect health information when held by providers and insurers may not apply to researchers.

These research principles need to be applied to other vulnerable populations undergoing large dataset analysis. For example, financially vulnerable populations are another group deserving of more attention. Aggregate credit scores applied to neighborhoods (versus individuals) are an example of how aggregate but specific predictions based on large datasets may be lead to potentially unfair practices based on primarily geographical factors. If the results of the analysis are not managed correctly or exposed to consumers, errors in prediction may never surface, and other usage issues can arise.

Other examples of important vulnerable populations exist, we do not attempt to be comprehensive here. The overall impetus of the policy guidance should be to identify potential risks for specific populations and in sensitive data, and plan for recourse and checks and balances to mitigate harm or abuse and encourage the best possible uses and results.

Suggestions for Research

We would like the outcome of increased big data adoption to be better insight and more innovation, with adequate and robust protection for vulnerable populations. To do this, we need a significant study of large datasets that focuses on understanding how they are affecting vulnerable populations. This is an under-researched area. The questions we do not have adequate answers for yet include:

- How is big data affecting vulnerable populations?
- What risks are associated with big data and vulnerable populations?
- Which are the vulnerable populations most at risk?
- What sources of data are most problematic for vulnerable populations?
- For these sources of data, what safeguards are in place to insure data quality, and allow for discovery and corrections of inaccuracies?
- What populations are most at risk with current practices?
- What ways has big data been used to assist the protection of vulnerable populations?

We look forward to continuing to work in this key area of privacy. We welcome feedback you may have, and we would be happy to provide answers to any questions you may

[8] See, e.g., Robert Gellman, *The Deidentification Dilemma: A Legislative and Contractual Proposal*, 21 Fordham Intellectual Property, Media & Entertainment Law Journal 33 (2010), http://bobgellman.com/rg-docs/RG-Fordham-ID-10.pdf.

have.

Respectfully submitted,

Pam Dixon, Executive Director
World Privacy Forum
www.worldprivacyforum.org