ACM US Public Policy Council

RESPONSE TO REQUEST FOR INFORMATION
Big Data Review
79 FR 12251
DOCUMENT NUMBER 2014-04660
OFFICE OF SCIENCE AND TECHNOLOGY POLICY

RESPONSE FILED BY:
U.S. PUBLIC POLICY COUNCIL OF THE ASSOCIATION FOR COMPUTING MACHINERY

On behalf of the U.S. Public Policy Council (USACM) of the Association for Computing Machinery (ACM) we are submitting the following comments in response to the Request for Information (RFI) by the Office of Science and Technology Policy (OSTP) on the comprehensive review on big data announced in January 2014.

With over 100,000 members, the Association for Computing Machinery (ACM) is the world's oldest and largest educational and scientific computing society. The ACM U.S. Public Policy Council (USACM) serves as the focal point for ACM's interaction with U.S. government organizations, the computing community, and the U.S. public in all matters of U.S. public policy related to information technology. Our comments are informed by the research experience of our membership. Should you have any questions or need additional information, please contact our Public Policy Office at 212-626-0541 or at acmpo@hq.acm.org.

We welcome the review of issues connected to the intersection of big data and privacy. The concept of big data is still emerging, but it is not too early to review how changes in the ability to collect, analyze and use large amounts of information provide new challenges and opportunities. While the definition of big data used for this RFI focuses on datasets so "large, diverse and/or complex, that conventional technologies cannot adequately capture, store and analyze them" the questions in this RFI (and our responses) can be applicable to large datasets currently captured by conventional technologies. The ability to effectively analyze collected data typically follows the ability to capture and store such data. As capabilities change, it will be important to systematically revisit datasets as our analytical abilities advance, both concurrently and after data collection and storage.

Answers to specific questions in the RFI

(1) What are the public policy implications of the collection, storage, analysis, and use of big data? For example, do the current U.S. policy framework and privacy proposals for protecting consumer privacy and government use of data adequately address issues raised by big data analytics?

The rise of big data highlights tensions that have existed in U.S. efforts to protect consumer privacy and limit government use of data. While the Fair Information Practice Principles

(FIPPs) (which are part of the USACM recommendations on privacy practices[1]) continue to have utility, trends in big data have made certain practices less effective. It has become significantly easier to extract personally identifiable information from nominally de-identified data as more data becomes available. In recent years academic researchers have shown that many data sets thought to be "de-identified" or "anonymized" can be re-identified when the data are correlated with other information that is publicly available.

Restricting data collection to specified purposes runs counter to the efforts of the government and companies to repurpose already collected data. The multitude of purposes for collected data have added to the complexity of privacy policies, documents that consumers find harder to understand. It also complicates efforts to give notice and obtain consent for collected data and associated uses. If big data trends continue to focus on multiple uses of collected data, new tools and processes for protecting and securing that data are needed to augment the protections currently available. Policies on protecting personally identifiable information, handling data breaches, and ensuring data quality become more important – not less – in an era of big data.

That obtaining notice and consent is increasingly difficult does not imply that the unrestricted repurposing of unchecked data is the way forward.

First, the more use that is made of collected data, the more important it is that the data be accurate. If collected data is going to be used for multiple purposes, making sure that data is accurate has additional benefits for the consumer, the data collector and other parties that may reuse that data. An important tool for this is to provide consumers the ability to check that collected data is indeed accurate, and to correct or dispute inaccuracies.

Tracking the flow of data will also become more important as it will likely change hands more frequently, and the ability to correct that data will need to keep up. Making sure the data remains accurate, secure and uncorrupted is important, and some encouragement may be needed for effective controls to be established and/or augmented for big data.

Such efforts would benefit from a sustained effort to develop approaches, technology and standards for the systematic tracking of data provenance and metadata. Agencies could help by funding specific research in this area and by sponsoring forums to discuss and ultimately adopt FIPS for provenance and metadata.

Separately, one could include the ability to remove data and/or opt out of (or consent to) specific uses. The right to do so may depend on who has supplied it; information a consumer supplies, or which is collected from that consumer's online behaviors, may deserve more control than, for example, reports of criminal convictions by that individual. Accuracy, in contrast, including deletion of convictions (where judicially appropriate) is more critical than corrections about consumer information. Some practices that have

[1] http://usacm.acm.org/privsec/category.cfm?cat=7&Privacy%20and%20Security

proven infeasible under current practices may be possible through the cloud, with appropriate policies.

Since big data in the context of this RFI is concerned with unconventional technologies, we recommend that new policies, procedures and best practices be as technology-neutral as possible. Technologies and methods will change more quickly than policies in the public and private sectors will be able to adapt.

(2) What types of uses of big data could measurably improve outcomes or productivity with further government action, funding, or research? What types of uses of big data raise the most public policy concerns? Are there specific sectors or types of uses that should receive more government and/or public attention?

Lessons learned in large archival efforts to date (e.g., libraries, music) may be applicable to big data. Ensuring the quality of collected data on par with archived data can make it easier to migrate data between users, allowing for additional uses of collected data. As recognized in the May 2013 executive order and formalized in policy directive M-13-13, standardization of data formats (preferably in open data formats that are machine readable) must be encouraged as well. This will reduce expenses for public and private parties interested in big data, and make data more accessible and readable for any interested party.

Areas of public policy concern related to big data include intellectual property connected to large datasets. The ability to conduct effective research on large datasets could be hampered by limits of access to the data, or restrictions on the ability of research resulting from these datasets to be properly reviewed and replicated. Policy makers should consider incentives to require researchers to make data available for public benefit (e.g., medical research); in some cases this might be required as a quid pro quo for using data about individuals without direct personal benefit.

A useful distinction to make for big data and public policy is between data generated that is somehow related to human activity and data generated that is not. When humans are the subjects of large datasets, a higher level of privacy and security controls will be needed to protect the sensitive data collected. This is particularly true of online behaviors that are increasingly tracked, correlated and shared between corporate entities.

(3) 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?

The challenges of data de-identification (including the increasing ease of re-identification) will be important to big data. Being able to de-identify data is a common practice intended to protect the privacy of subjects of collected data. It is important to remember that de-identification is not a binary state, but a spectrum. It is most appropriately viewed as a single privacy risk control rather than a technique that renders privacy concerns irrelevant. As such, its use does not in and of itself constitute a convincing argument for completely removing such data from a risk management regime or from applicable regulatory oversight. Anonymization and de-anonymization are active areas of research, changes in technology and methods will make current de-identification standards obsolete in ways that cannot be predicted.

Techniques are needed for a consumer to express their preferences, without needing to confront the complexity of all possible circumstances of use. Informed consent is sometimes taken to mean both control such that one's intent is realized, and also control (or at least understanding) over all details. These can conflict, and it may be best to separate them.

Computer security setup may suggest a way forward. Users can configure Microsoft Windows security *approximately* right, by answering a handful of questions about their attitudes and tradeoffs; as systems gain additional controls, experts provide rules to generate new settings. Compared with multipage legalese or settings on dozens of complex features, such tradeoff responses give a less precise but more understandable picture of system behavior.

Data quality issues will matter as well. Besides the reasons mentioned in our response to question 2, preserving the quality of data helps make it more useful to the entities looking to reuse it. That data is accurate and reliable will help make analysis easier, and depending on the datasets and analyses involved, could address the challenges of false positives and false negatives. The utility of metadata is linked to its sensitivity, and both require attention to ensure the quality of the underlying data.

Inexpensive means of providing durable homes for research data should be encouraged. The availability of cloud resources for purchase and lease provides an opportunity to improve data durability. When a research project terminates, its hardware and hence its data traditionally became unavailable. One needed mechanism is to fund archiving, a process to identify data to be archived (if raw data is too large or too sensitive), and connection to a new technical system to do the archiving. With a storage cloud, the data can stay in place, so the last barrier is substantially reduced. (Cloud providers should be required to ensure that data is protected if the provider goes out of business).

Finally, new technologies allowing individuals greater control of data collected about them, especially in online contexts, is an important area of current computing research. New technologies are being explored ranging from "data vaults" to renumeration systems for allowing various kinds of data sharing (returning value to the consumer when the data is used). These technologies are being coupled with other technologies to allow users to permit their data to be shared at various levels of aggregation, while controlling the direct sharing of personal information. These aggregation and control technologies hold promise for safeguarding privacy in the big data era, and the government should encourage their creation and usage.