ELECTRONIC FRONTIER FOUNDATION
Protecting Rights and Promoting Freedom on the Electronic Frontier

Submitted via privacyrfc2014@ntia.doc.gov

August 5, 2014

John Morris, Associate Administrator and Director of Internet Policy
National Telecommunications and Information Administration
U.S. Department of Commerce
1401 Constitution Avenue, NW, Room 4725
Attn: Privacy RFC2014
Washington, DC 20230

RE: *Request for Public Comment on Big Data and Consumer Privacy in the Internet Economy*

Dear Mr. Morris:

Thank you very much for the opportunity to provide comments to NTIA on the topic of big data, prompted by the White House's Big Data Report and the PCAST Report. In these brief comments, the Electronic Frontier Foundation wishes to focus on one main point: that policymakers should be careful and skeptical about claims made for the value of big data, because over-hyping its benefits will likely harm individuals' privacy.

The Administration is now looking at "how the Consumer Privacy Bill of Rights could support the innovations of big data while at the same time responding to its risks," including a "responsible use framework."[1] In EFF's view, it is difficult at this time to craft such a framework given the problems surrounding de-identified data.[2]
More importantly, this approach assumes too much. That big data is no panacea should be well known by now.[3] But it is disturbing to see how much the PCAST Report, and by extension the NTIA RFC, seems to embrace the big data vision. This is especially concerning given recent criticism of big data analysis by computer scientists[4], economists[5], statisticians[6], and others in the field[7].

---

[1] Executive Office of the President, *Big Data: Seizing Opportunities, Preserving Values* 61 (May 2014).
[2] Narayan, Arvind and Felten, Edward, *No Silver Bullet: De-Identification Still Doesn't Work* (July 9, 2014).
[3] *See, e.g.*, Fung, Kaiser, http://blogs.hbr.org/2014/03/google-flu-trends-failure-shows-good-data-big-data/; Lazer, David and Kennedy, Ryan and King, Gary and Vespignani, Alessandro, *Google Flu Trends Still Appears Sick: An Evaluation of the 2013-2014 Flu Season* (March 13, 2014). Available at SSRN: http://ssrn.com/abstract=2408560.
[4] Lazer, David, et al. *The Parable of Google Flu: Traps in Big Data Analysis*. Science 14 March 2014: 343 (6176), 1203-1205.
[5] Harford, Tim. *Big data: are we making a big mistake?* The Financial Times. (March 28, 2014)

815 Eddy Street • San Francisco, CA 94109 USA
voice +1 415 436 9333 fax +1 415 436 9993 web www.eff.org email information@eff.org

Yet the NTIA request for public comment seems to assume the benefits of big data.

Question 5, for instance, asks: "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?"

But there is not one question about whether existing research substantiates the supposed social benefits of big data. Given the public concern over NSA bulk collection and data mining, we know that NTIA did not intend to suggest that big data's privacy risks are more speculative or less substantiated than its potential benefits. We nevertheless feel compelled to criticize the apparent assumption of the validity of big data analysis.

Thus, while we do not argue that big data analysis cannot sometimes be accurate and effective, it is important to note that this is only the case if the data collection and analysis are done carefully and purposefully. For big data analysis to be valid, one must follow rigorous statistical practices. Simply "collecting it all" and then trying to extract useful information from the data by finding correlations is likely to lead to incorrect (and, depending on the particular application, harmful or even dangerous) results. To see why, we summarize below three of the major technical problems with big data analysis, which we feel must be addressed before any trade-offs with privacy can be explored.[8]

Problem 1: Sampling Bias

The first major technical problem with big data is the assumption that if one has a large enough data set, then the data will automatically be statistically representative of the underlying population. This claim that "that 'N = All', and therefore that sampling bias does not matter, is simply not true in most cases that count."[9] On the contrary, big data sets "are so messy, it can be hard to figure out what biases lurk inside them – and because they are so large, some analysts seem to have decided the sampling problem isn't worth worrying about. It is."[10]

---

[6] Fung, Kaiser. *Two unsolved problems of Big Data studies: confirmation and controls*. http://junkcharts.typepad.com/numbersruleyourworld/2014/07/two-unsolved-problems-of-big-data-studies-confirmation-and-controls.html

[7] Taleb, Nassim. *Beware the Big Errors of 'Big Data.'* Wired Magazine (February 8, 2013)

[8] There are of course other problems with much of big data, such as the general lack of replication of results in the literature (due in large part to the proprietary nature of most big data sets collected by companies). For brevity's sake, however, we will focus in these comments on the major three technical problems.

[9] Harford, *supra* fn. 5.

[10] *Id*.

Correcting for sampling bias is especially important given the digital divide. By assuming that data generated by people's interactions with devices, apps, and websites are representative of the population as a whole, policy-makers risk unintentionally redlining large parts of the population.[11] Simply put, "with every big data set, we need to ask which people are excluded. Which places are less visible? What happens if you live in the shadow of big data sets?"[12]

The only way to correct for this sort of error is by employing carefully thought out sampling techniques, or by using statistically validated weighting techniques post-sampling. Both of these methods, however, require understanding and knowledge about the underlying data sets and the populations they represent. Simply taking a data set and throwing some statistical or machine learning algorithms at it and assuming "the numbers will speak for themselves" is not only insufficient—it can lead to fundamentally flawed results. If those results are used in ways that have a dramatic effect on people's lives (e.g. determining allocation of government resources, deciding the terms of a mortgage, influencing health insurance premiums, etc.) then the result may be significant, completely avoidable harm to innocent people.

Problem 2: Correlation is Not Causation (And Sometimes, Correlation is Not Correlation)

Even if one tackles the sampling problem, a fundamental problem with big data is that

"although big data is very good at detecting correlations…it never tells us which correlations are meaningful. A big data analysis might reveal, for instance, that from 2006 to 2011 the United States murder rate was well correlated with the market share of Internet Explorer: Both went down sharply. But it's hard to imagine there is any causal relationship between the two."[13]

In other words, while big data may allow one to discover new correlations in the underlying data, it doesn't follow that those correlations are meaningful. Unfortunately, this holds true even with correlations that are less obviously disconnected than the murder-Internet-Explorer correlation described above. As a result, one should always be suspicious about taking any action based strictly on correlation that was not explicitly being tested for, no matter how convincing it may seem.
Even more problematic, however, is the fact that "big data may mean more information, but it also means more false information."[14] This contributes to what is known as the

---

[11] Schradie, Jen. *Big Data Not Big Enough? How the Digital Divide Leaves People Out*. http://www.pbs.org/mediashift/2013/07/big-data-not-big-enough-how-digital-divide-leaves-people-out/ (July 31, 2013).
[12] Crawford, Kate. *The Hidden Biases in Big Data*. http://blogs.hbr.org/2013/04/the-hidden-biases-in-big-data/ (April 1, 2013).
[13] Marcus, Gary, and Davis, Ernest. *Eight (No, Nine!) Problems With Big Data*. The New York Times (April 6, 2014).
[14] Taleb, *supra* fn. 7.

"multiple-comparisons" problem: if you have a large enough data set, and you do enough comparisons between different variables in the data set, some comparisons that are in fact flukes will appear to be statistically significant. As Tim Harford explains:

"There are various ways to deal with this but the problem is more serious in large data sets, because there are vastly more possible comparisons than there are data points to compare. Without careful analysis, the ratio of genuine patterns to spurious patterns – of signal to noise – quickly tends to zero."[15]

In other words, the number of false correlations almost always grows faster than the number of true correlations. This problem demonstrates the falsehood of one of the fundamental maxims frequently espoused by big data proponents: that by collecting enough data, one can tease out interesting new meanings and correlations. In fact the opposite is true—by gathering enough data, one is guaranteed to find spurious correlations and come to invalid conclusions. What is "interesting" is not necessarily true or valid.

Once again, the only way to counteract this problem is with knowledge and understanding of the underlying problem or question one is trying to answer—one cannot simply collect a couple of petabytes of data in a bucket, shake it up with a sieve of nonlinear regression analysis, and assume that what remains is a genuine gold nugget.

Problem 3: Fundamental Limitations of Machine Learning

Many computer scientists would argue that one way to combat false correlations is to use more advanced algorithms, such as those involved in machine learning.[16] But even machine learning suffers from some fundamental limitations.

First and foremost, "getting machine learning to work well can be more of an art than a science."[17] Unfortunately, the complicated nature of most machine-learning algorithms means that they have a variety of design parameters (variables which can be tuned to improve the algorithm's performance) which must be picked *a priori* by the human analyst, and which can affect the results in unexpected ways. Unfortunately this means that such algorithms can be subject to human error: even though the "machine" is "learning," it is still doing so based on inputs from the person who programmed the algorithm and is subject to that person's biases (subconscious or otherwise). Techniques like cross-validation can help reduce these biases, so it is important to make sure that such techniques are used before accepting the results of any machine-learning algorithm.

---

[15] Harford, *supra* fn. 5.

[16] The line between simple statistics and machine learning is blurry; when we say "machine learning," we mean algorithms that would be more likely to be taught in a computer science class than a statistics class, such as neural networks, support vector machines, Markov models, decision trees, etc.

[17] Bradski, Gary, and Kaehler, Adrian. *Learning OpenCV: Computer Vision with the OpenCV Library*. O'Reilly Media, September 2008.

Second, machine-learning algorithms are just as susceptible to sampling biases as regular statistical techniques, if not more so. The failure of the Google Flu Trends experiment is a prime example of this: machine-learning algorithms are only as good as the data they learn from.[18] If the underlying data changes, then the machine-learning algorithm cannot be expected to continue functioning correctly. (For example, if news coverage of a flu pandemic elsewhere in the world leads to more Americans searching Google using terms related to the flu, then an algorithm that uses search terms to predict U.S. flu rates would have to be re-worked in order to remain accurate.)

Additionally, many machine-learning techniques are fragile: if their input data is perturbed ever so slightly, the results will change significantly.[19] This is an important point to consider when the results of such machine learning algorithms could affect people's lives in dramatic ways: even the most well-trained, cross-validated, neural network can be presented with data that is just slightly perturbed, and it will output a wildly incorrect result.

Finally, machine learning, especially model-free learning[20], is not a valid replacement for more careful statistical analysis (or even machine learning using a model). To quote Gary Marcus:

"If fifty years of research in artificial intelligence has taught us anything, it's that every problem is different, that there are no universally applicable solutions. An algorithm that is good at chess isn't going to be much help parsing sentences, and one that parses sentences isn't going to be much help playing chess. A faster computer will be better than a slower computer at both, but solving problems will often (though not always) require a fair amount of what some researchers call 'domain knowledge'—specific information about particular problems, often gathered painstakingly by experts."[21]

This leads once again to our fundamental message: simply "collecting it all" and assuming that a general machine-learning algorithm will extract useful information from a collection of big data is naïve at best. The only way around this is to have some knowledge of the problem at hand. This once again means that it is important that the data be collected with a specific purpose in mind, so that it can be analyzed using rigorous statistical techniques: not simply re-used because it happened to be sitting

---

[18] Lazer, *supra* fn 4.
[19] Szegedy, Christian, *et al*. *Intriguing properties of neural networks*. http://arxiv.org/abs/1312.6199.
[20] Model-free learning refers to algorithms in which researchers attempt to solve a problem or answer a question using a general framework that doesn't make any assumptions about the structure of the data. This can be contrasted to model-based learning, in which researchers use a mathematical model based on some *a priori* knowledge of the problem at hand, and attempt to learn the values of the variables in that model.
[21] Marcus, Gary. *Steamrolled by Big Data*. The New Yorker (March 29, 2013).

around and a computer scientist decided to throw the latest deep-learning algorithm at it to see what came out.

Not all big data can overcome these technical challenges

Despite these three major technical obstacles, we do not argue that it is impossible for the results of big data analysis to be valid or beneficial to society. With careful forethought and the use of proper statistical techniques, it is possible to use big data to answer difficult questions and come up with wonderful new ways of helping society as a whole. But it is important to note that this is only true for one particular type of big data analysis: analysis that attempts to learn a trend or correlation about a population as a whole (e.g. to identify links between symptoms and a disease, to identify traffic patterns to enable better urban planning, etc.).

This is how much science has worked for the last century: a researcher posits a hypothesis about a phenomenon, determines how to take data to test that hypothesis, collects the data, and then performs statistical analysis to determine whether or not the data bears out the hypothesis. Simply having more data doesn't affect the underlying method—it just means that less prominent effects can be teased out from the data (assuming precautions are taken to avoid spurious correlations, as described above). One still needs to carefully determine *ahead of time* what data to collect and how it will be analyzed. (Of course, Google Flu Trends did not do this, which is why it eventually failed.)

Other uses of big data by their very nature cannot overcome these technical obstacles. Consider the idea of targeting individuals on a massive scale based on information about them collected for a secondary purpose. By using "found" data that was not intended for the specific use it is being put to, sampling biases are inevitable (i.e. Problem 1).

Or consider the claim by proponents of big data that by "collecting it all" and then storing it indefinitely, they can use the data to learn something new at some distant point in the future. Not only will such a "discovery" likely be subject to sampling biases, but any correlations that are discovered in the data (as opposed to being explicitly tested for) are likely to be spurious (i.e. Problem 2).

At the same time, these sorts of uses (individualized targeting, secondary use of data, indefinite data retention, etc.) pose the greatest privacy threats, since they involve using data for purposes for which consent was not originally given and keeping it longer than otherwise necessary.

Thus, before NTIA starts asking about how to reconcile big data with privacy issues, it would first make sense to ask if it's even worth it, given the technical obstacles that prevent big data from being used effectively and accurately in these sorts of privacy-invasive scenarios.

To quote author Edward Tenner: "So far at least, there is no reliable quantitative information on implementation and success of big data projects,…partly because most big data projects are proprietary."[22] In other words, "there are no big data about big data."[23]

Let's not forget about national security uses of big data

To return to Question 5 of NTIA's request for public comments ("Is there existing research or other sources that…substantiate the privacy risks…associated with big data?") we would answer that the most direct harm to privacy comes from uncritical acceptance of the "collect it all" mentality, which may lead to calls to create express legal accommodation for collection of data about individuals without consent or even notice.

Of course, a main reason for current public concern with big data is that the NSA's bulk collection programs perfectly exemplifies this "collect it all" mentality. EFF has previously criticized the Administration for excluding the subject of national security surveillance from its big data workshops. After all, if lack of transparency is problematic for corporate experiments like Google Flu Trends, it is far more problematic given the near-impenetrable secrecy that surrounds national security surveillance.

The approach of the NTIA RFC is thus troubling. For the Administration to build on the big data workshops by further proceeding to legitimize big data methods without insisting on genuine scientific methodology—evidence-based policy—risks not only our privacy as ordinary consumers, but as ordinary citizens. The more we laud big data without critically evaluating its bona fides, the more we unnecessarily legitimize secret bulk collection of personal information.

Sincerely,

Lee Tien, Senior staff attorney
Jeremy Gillula, Staff technologist
Electronic Frontier Foundation

---

[22] Tenner, Edward. *Big Data: Here to Stay, but with Caveats*. The American. (July 30, 2014)
[23] *Id*.