Big Data in Private Sector and Public Sector Surveillance

Recent years have seen an explosion in the popularity of big data. This popularity is attributable to a variety of reasons, including the easier collection of data points by computers and the affordability of massive storage devices and computer processing power. Big data has become a trendy catchphrase for the idea that large datasets can often be used to learn interesting relationships that are not obvious at first glance, or that might not be evident in smaller datasets. Unfortunately, public policy has not kept pace with the rising popularity of big data, resulting in dangers to consumer privacy when big data is used by the private sector, and to constitutional norms and rights when used by the government.

Not all uses of big data implicate dangers to privacy or rights, such as datasets that are not about people or what they do. Even when the datasets concern people, such as the analysis of a dataset on health information to find previously unknown links between diseases or the analysis of a traffic dataset to aid in urban planning, analysis aimed at generating insights about large populations may pose relatively less risk than analysis aimed at classifying, sorting, or focusing on particular individuals or groups. Of course, there is always the risk that a data breach or dissemination of the underlying dataset could expose individuals' personal information (even if the desired analysis does not). If the type of analysis that will be done is known ahead of time, however, then it may be possible to mitigate this privacy risk through techniques that scrub or anonymize the data in such a way that only data relevant to the desired statistics remain.

The larger threat to privacy comes when big data is used to individually target people in a certain group found within a dataset. As an example, consider Target's development (using data from its baby registry) of an algorithm that analyzes someone's purchases in order to determine if they are pregnant, and the subsequent use of that algorithm to individually target people not in the registry with baby-related advertising.[1] By running algorithms on its customer dataset, which included looking for, among other variables, customers purchasing unscented wipes and magnesium supplements, Target's use of big data to identify pregnant customers raises questions, like: Is it ethical to analyze its baby registry for a purpose that its customers probably did not know about? To develop an algorithm for identifying potentially pregnant customers knowing

1 Duhigg, Charles. "*How Companies Learn Your Secrets.*" New York Times, February 16, 2012.

that they probably do not want to be identified? To advertise to them? To potentially offer its pregnancy assessment service to employers, insurers, or others? In the biomedical field, ethical guidelines and practices like the Belmont Report and the Federal Policy for the Protection of Human Subjects seek to protect individuals' interest in respect and autonomy.[2] Collection and privacy standards must be recognized when it comes to big data as the information collected (and insight gleaned) can reflect some of the most intimate details of a person's life.

This concern is even greater with respect to the increasing use of big data for government surveillance, such as the government's use of Section 215 of the Patriot Act to collect all Americans' calling records and Section 702 of the Foreign Intelligence Surveillance Amendments Act to indiscriminately collect users' phone calls and emails. There are also highly secret uses of surveillance authorities like classified National Security Policy Directives (NSPDs) and Executive Order 12333 ("EO 12333"). Such "authorities" are collecting data in a similar manner to Section 215 and Section 702. They are used to create massive datasets containing information concerning US and non-US persons. This data includes personal identifiers, sensitive information, personal communications, and potentially other data that may be commingled with data collected under the Foreign Intelligence Surveillance Act (FISA). Such use of big data is difficult to reconcile with the idea of privacy not only because much of the data is collected in secret without the predication required by the Fourth Amendment, but also because the analytics seek to identify particular individuals. Such uses raise the greatest public policy concern and deserve more government and public attention.

Executive Summary

The Electronic Frontier Foundation appreciates the opportunity to submit comments for the Office of Science and Technology Policy's Big Data RFI, OSTP-2014-0003-0001. Our comments focus on certain parts of the first four questions, and we specify what we address here, along with very brief summary answers:

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

2 See http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.htm

by big data analytics?

Big data has serious public policy implications for privacy and fairness. We do not dwell on the potential benefits here, except to warn that good public policy should: (a) not overestimate potential benefits while discounting potential harms; (b) recognize that potential benefits and harms must be assessed within a realistic understanding of economic and political incentives; and (c) recognize that time matters. Because (a) is obvious, we elaborate on (b) and (c).

Realism about incentives is another way of saying that just because something can be done doesn't mean that it will be. We do not expect companies to use big data altruistically; we expect them to use big data to compete, profit, and grow. If big data gives an insurance company an incentive to eliminate more costly policyholders it is reasonable to assume that it will do so. We may wish for government to act in the public interest, but visions of the public interest vary greatly, and government agencies with missions like law enforcement or intelligence present special tensions, especially for constitutional values like due process, transparency, and democratic accountability.

To say that time matters is to highlight not only the way that law and policy often trail technological change, but also path dependence. One might think of this as the macro-version of "privacy by design." It will be much harder to protect privacy when business models or government programs become entrenched in their use of big data.

Fundamentally, the privacy and fairness implications of big data are closely tied to concerns about social, economic and political power. Both the public and private sector have strong interests in collecting and using data about individuals, whether for business or for social control. Since 9/11 it has become more obvious how much data collected by one sector flows to the other; the telephony metadata program using Section 215 of the Patriot Act is merely one especially obvious example. Power and inequality is, of course, an enduring issue for our nation, but big data seems especially troubling because large, powerful entities are more likely to have access to extremely large and complex datasets, and the sophisticated computing power needed to analyze them.

(2) 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?

Data about rocks is different from data about people or what people do. Analysis of the latter always raises potential ethical issues, which is why much research in the medical arena has traditionally been guided by human subjects protocols and legal-ethical constructs like the Common Rule.[3] We attempt to distinguish between big data uses that focus on large populations and those that focus on small groups or individuals, believing that the latter, roughly speaking, is of greater concern. We also believe that big data programs that operate secretly or with little visibility and transparency demand more attention. Accordingly, we believe that the use of big data in the national security and law enforcement realm deserves the greatest scrutiny. Private data collection remains in scope, of course, because many intelligence programs rely heavily on data collected by the private sector.

(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?

We do not address this question below, and merely note two points. First, improved sensing or data collection technology obviously exacerbates the big data problem. More of what people do is capable of being captured, often without their awareness or their ability to avoid collection. Second, EFF is exploring the potential use of differential privacy techniques by large California utilities that collect granular energy usage data generated by smart meters. Federal smart grid incentives, we believe, illustrate a basic big data policy problem: smart meter deployment was stimulated with little consideration of the privacy risks or of how privacy might be designed into the smart grid. Indeed, much of EFF's role in California's utility regulation progress has been to help create a privacy framework for energy usage data in the face of strong government, commercial, and academic demand for this highly revealing data. EFF and its technical experts faced considerable political resistance even as we struggled to educate policymakers and stakeholders about the re-identification risks associated with granular data.

3 See http://www.hhs.gov/ohrp/humansubjects/commonrule

(4) How should the policy frameworks or regulations for handling big data differ between the government and the private sector? Please be specific as to the type of entity and type of use (eg, law enforcement, government services, commercial, academic research, etc.).

First, government use of big data is inherently subject to constitutional constraints, while private sector use of big data is typically subject only to statutory constraints, with two significant caveats. In California, for example, private actors are subject to the state constitutional privacy right. And even under the federal constitution, private actors can in some circumstances violate individual rights under the state action doctrine. Of particular importance are the predication and particularity values of the Fourth Amendment, the due process values of the Fifth Amendment, the reasoned elaboration values of Article III courts and the democratic accountability of the Constitution itself.

Second, the policy framework for law enforcement and intelligence uses of big data is distinguishable from most other contexts by its lack of transparency. Obviously, law enforcement and intelligence agencies typically collect data in secret and without the consent of the people being surveilled. Secrecy also interferes with public knowledge about these surveillance practices and technologies. Particularly in the intelligence realm, the system of classified information and the state secrets privilege distorts normal processes of democratic accountability essential to legitimate constitutional government. And because these law enforcement and intelligence agencies often rely on data collected by the private sector, these distortions also directly affect individuals' trust relationships with business. Any big data policy framework must publicly address government secrecy and over-classification.

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

That NSA surveillance is conducted both domestically and globally has never been a secret, but recent revelations have made the international nature of signals and communications intelligence impossible to ignore. NSA surveillance is obviously not simply a US problem. The relationship between NSA and GCHQ is clearly intimate, and many other governments in some

way partner with the United States. Genuine accountability as to the intelligence community's use of big data will, at the very least, require accountability as to these data flows and partnership arrangements. After all, if NSA and GCHQ share data about each other's citizens, it would make no sense to control NSA but not GCHQ. More generally, it appears that the various national intelligence agencies have developed their own norms of bulk collection in direct conflict with non-intelligence norms of predicated and particularized seizure or collection of communications. The question may not so much be the crossing of a national border but rather intelligence agency exceptionalism. When the United States attempts to justify secret and unaccountable mass surveillance programs with no credible evidence of utility, it can hardly criticize other nations for doing so.

Big Data Facilitates Private Sector Surveillance

The collection and analysis of big data, which was a niche field within computer science just two decades ago, has exploded into a $100 billion industry.[4] Big data is now used in sectors as diverse as energy, medicine, advertising, and telecommunications. Because of the explosive growth of this field, companies ranging from startups in Silicon Valley to established multi-national corporations are adopting the mantra of "collect it all," in the belief that running a variety of analytics on big data will increase the value of their products or the companies themselves.

In many cases companies outsource the use of big data to intermediary entities known as data brokers, which collect, analyze, and sell consumer information that can include highly personal details like marital status, religion, political affiliation, tax status, and others. A website may have an agreement with a data broker to better identify who their customers are so they can place more effective ads—often in exchange for their customers' browsing habits and demographic information. Data brokers receive and aggregate consumer data from a variety of sources: transactional data from retailers and stores, loyalty cards, direct responses and surveys, social media and website interactions, public records, and more.[5] They then aggregate this information across sources and use it to create highly detailed profiles about individuals—one

4 *"Data, data everywhere."* The Economist, Feb. 25, 2010. https://web.archive.org/web/20131207192955/http://www.economist.com/node/15557443. Last accessed March 28, 2014.

5 See Dixon, Pam. "*What Information Do Data Brokers Have on Consumers?*" World Privacy Forum, December 18, 2013. Last accessed March 30, 2014.

particular data broker is said to have 1,500 data points on over 700 million individuals.[6] It's been revealed that these highly detailed profiles include names like "Ethnic Second-City Strugglers," "Rural and Barely Making It," and "Credit Crunched: City Families," as well as sensitive lists such as police officers and their home addresses; lists of rape victims; genetic disease sufferers; and Hispanic payday loan responders.[7] The vast majority of information data brokers use to create these lists is data which consumers unintentionally expose in large part because they simply do not know how or when they are being tracked, or what information is being collected. As a result the information is almost perfectly asymmetric: brokers know a great deal about consumers, but most consumers have no idea these parties actually even exist.

This asymmetry is related to the first harm consumers are exposed to as a result of private-sector big data usage, namely the significant power imbalance between consumers and the companies wielding the data and analysis tools. For example, if a company uses big data analysis to inform its hiring decisions (say by analyzing a database on the web browsing habits of potential employees acquired from a data broker), would a rejected prospective employee learn why she was not offered a job, be able to see the data that led to the decision or the algorithm that processed the data, or dispute the correctness of either?[8] In general, the fact that people may be treated differently based on data and algorithms that they know little about and have no recourse for correcting creates elementary fairness and transparency problems.[9]

A related problem results from the fact that even if consumers are aware of what data they are providing about themselves and who they are providing it to, they frequently believe wrongly that the law or a company's privacy policies block certain uses of that data or its dissemination. As explained by Chris Hoofnagle and Jennifer King in their study of Californians' perceptions of online privacy:

---

6 See Brill, Julie. "*Demanding transparency from data brokers*." The Washington Post, August 15, 2013. http://www.washingtonpost.com/opinions/demanding-transparency-from-data-brokers/2013/08/15/00609680-0382-11e3-9259-e2aafe5a5f84_story.html. Last accessed March 30, 2014.

7 See Dixon, Pam. "*What Information Do Data Brokers Have on Consumers?*" World Privacy Forum, December 18, 2013. Last accessed March 30, 2014.

8 One could argue that it would be in a company's best interests to use data that is as accurate as possible. However, a company's ultimate goal is to be as profitable as possible, and big data analysis is only carried out to further that goal. No rational company would acquire better quality data when the cost of doing so would be greater than the estimated returns. This exposes the fundamental mismatch in incentives between companies (whose big data will only be as accurate as profitability dictates) and individuals (who primarily care about whether the data about they themselves is accurate). Even a competitive market might not be able to completely resolve this issue, since making sure all the data is accurate 100% of the time will likely require human-intensive, and therefore costly, dispute/redress processes.

9 Dwork and Mulligan, "It's Not Privacy, and It's Not Fair," 66 STAN. L. REV. ONLINE 35 (2013).

> Californians who shop online believe that privacy policies prohibit third-party information sharing. A majority of Californians believes that privacy policies create the right to require a website to delete personal information upon request, a general right to sue for damages, a right to be informed of security breaches, a right to assistance if identity theft occurs, and a right to access and correct data.[10]

Additionally, users may not know to what extent data is shared with unknown third-parties: an online project called "theDataMap" reflects this data-sharing landscape.[11]

But even a good understanding of the legal and policy protections for data is insufficient to protect a consumer from harm, due in large part to the next danger: loss of privacy due to individualized analysis and tracking by private-sector use of big data. By "connecting the dots" between different, disparate datasets, or even by analyzing data from the same dataset that on its face does not seem to have any connection, companies can infer characteristics about people that they might not otherwise wish to be made public, or at least not wish to share with certain third-parties (for example, the well-known Target pregnancy example). Very few consumers realize the power of statistical analysis and other big data algorithms. Even if consumers are aware of what specific data they are sharing, they may not understand what inferences could be made based on that data.

The risk of abuse of the underlying datasets remains. As the recent hack on Target's credit card systems demonstrates, even large, well-financed companies can suffer from massive data breaches that put consumers' data in the hands of a malicious third-party.[12] This danger is especially grave when companies collect and save all data possible, regardless of its current value, with the idea that a profitable use might later emerge. Unfortunately, the collection of data into more concentrated repositories creates a tempting target for malicious agents. Additionally, EFF has long been concerned that private-sector mass data accumulation strongly facilitates government data accumulation, given the many ways that companies can be induced or compelled to provide data to the government.

Finally, even if the above dangers are avoided, we emphasize that many "common sense" approaches to preserving privacy and anonymity in big data do not actually accomplish their goals. Malicious actors could use a variety of sophisticated statistical and information-theoretic

---

10 Hoofnagle, Chris Jay and King, Jennifer, "*What Californians Understand about Privacy Online*." (September 3, 2008). Available at SSRN: http://ssrn.com/abstract=1262130 or http://dx.doi.org/10.2139/ssrn.1262130

11 See http://thedatamap.org/

12 Elgin, Ben; Lawrence, Dune; Matlack, Carol; Riley, Michael. *"Missed Alarms and 40 Million Stolen Credit Card Numbers: How Target Blew It."* Bloomberg BusinessWeek, March 13, 2014. https://web.archive.org/web/20140313132757/http://www.businessweek.com/articles/2014-03-13/target-missed-alarms-in-epic-hack-of-credit-card-data. Last accessed March 29, 2014.

algorithms to extract identifiable data from what appears to be an anonymized dataset.[13] This is especially true if the malicious agent has access to individual datasets that might not pose a privacy risk on their own, but when combined together can be used to infer private information.

Suggested Approaches

Unfortunately, current US policy frameworks and privacy proposals for protecting consumer privacy when it comes to the use of big data by the private sector are woefully inadequate. In order to remedy this and counteract the dangers described above, we propose a range of suggestions.

The private sector could adopt and adapt the White House's Consumer Privacy Bill of Rights to the collection and usage of big data.[14] Part of adapting to big data lies in companies' being more mindful of the knowledge asymmetry problems discussed above. In particular, companies should adopt clear policies that enable consumers to understand their data collection, use, and dissemination practices, including what specific personal data companies have about consumers and the algorithms (including scoring protocols) used to make decisions about consumers.[15] Because the results of big data analytics are often hard to predict,[16] consumers should also be able to see what a company's analytics are inferring about them. In essence, "consumers have a right to exercise control over what personal data companies collect from them and how they use

---

13 Anderson, Nate. *"'Anonymized' data really isn't—and here's why not."* ArsTechnica, Sep. 8, 2009. https://web.archive.org/web/20140123133104/http://arstechnica.com/tech-policy/2009/09/your-secrets-live-online-in-databases-of-ruin/. Last accessed Mar. 29, 2014.

14 *"Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation."* The White House, February 2012. http://www.whitehouse.gov/sites/default/files/privacy-final.pdf. Last accessed March 31, 2014.

15 One of the more common objections to this recommendation is that the algorithms underlying big data analysis are sometimes too complex for even the practitioners of big data to understand, so it would be pointless to show the algorithm to a lay-person (or even an expert) to offset any privacy or accuracy concerns. This objection is true only to a certain extent: while some big data analysis does involve complicated machine learning algorithms, much of it is fairly straightforward, such as the Google Flu project. (In essence, the advantage of many machine-learning algorithms lies not in their complexity, but in the fact that they automatically "tune" themselves using data that is presented by the designer in order to train them. Once a machine learning algorithm has been trained and is put into use on actual data it is usually fairly straightforward, at least for a computer scientist, to follow the data flow and understand *how* the algorithms works–just not *why* it ended up that way.) However, the use of an algorithm that is too complicated to understand raises the serious question of how such an algorithm can be proven to be accurate and unbiased. Without being able to explain how the algorithm works, how can a company guarantee that the algorithm doesn't have the effect (whether intended or not) of discriminating against certain classes of consumers? This problem serves to reinforce the recommendation that companies be transparent in their use of big data algorithms that can affect consumers' lives; only by being transparent will independent watchdogs be able to test big data algorithms for fairness.

16 Indeed, this ability of big data to tease out non-obvious relationships is one of the reasons for its increasing popularity.

it," as well as "a right to access and correct personal data... in a manner that is appropriate to the sensitivity of the data and the risk of adverse consequences to consumers if the data is inaccurate."[17]

1. Companies can use technical means to minimize the privacy risk to consumers. Since numerous studies have shown that simple de-identified datasets are too easy to re-identify, steps should be taken to the greatest extent possible to reduce the production and collection of identifiable information. For example, a brick-and-mortar retailer might track peoples' smartphones to analyze how consumers move from display to display or department to department, but assign each phone it sees a new, random ID (not based in any way on the phone's unique MAC address or other identifying information) at the beginning of each business day. This way, there is very little risk of privacy-invasion based on operations on the dataset. Of course there is still the risk that third-parties who got access to the dataset itself could use it to infer private things by combining it with other knowledge. An even better approach would be for companies to design products that do not have hard-coded unique ID numbers which are shared or transmitted by the device in the normal course of its operation. For example, MAC addresses could be designed so that they are rarely transmitted, and random identifiers are transmitted instead.[18]
2. Much of the data collected today is not shared on purpose by consumers, but is "found" data: data collected incidental to the use of products and services whose purpose (at least to the consumer) has nothing to do with big data.[19] We believe that if the government or companies do not take action to implement some of the solutions described above, more consumers will begin to use tools and products that "leak" less private data to third-parties. Already tools such as Tor[20] exist to prevent consumers from "leaking" data to their ISPs about the websites they visit; XPrivacy[21] exists to stop apps from gathering

---

17 *"Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation."* The White House, February 2012. http://www.whitehouse.gov/sites/default/files/privacy-final.pdf. Last accessed March 31, 2014.

18 This is how mobile phone systems work: a randomly generated Temporary Mobile Subscriber Identity (TMSI) is used instead of the unique International Mobile Subscriber Identity (IMSI).

19 Harford, Tom. *"Big data: are we making a big mistake?"* Financial Times Magazine, March 28, 2014. http://www.ft.com/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html. Last accessed March 31, 2014.

20 See https://www.torproject.org/.

21 See http://www.xprivacy.eu/.

unnecessary private information from Android smartphones; and browser plug-ins are being developed to allow people to automatically block third-party trackers on the Internet. It is perfectly appropriate for knowledgeable consumers to take technical precautions against surveillance, but it would be best if consumers could have legitimate trust and confidence that their everyday actions, online or offline, were not being routinely and secretly monitored.

The Government's Big Data Problem

Government use of big data raises many of the problems described above, which are exacerbated by the government's greater resources and its greater ability to exercise power over people's lives. Especially alarming are the recent revelations about intelligence activities that show government surveillance leveraging private sector tracking for advertising purposes as well as exploiting private-sector tracking implementations. In one example, it was revealed that the NSA uses advertising cookies to track a user's location and exfiltrate data off a computer.[22]

Even outside of the classified arena, the government has championed big data with little attention to privacy; the administration's "Big Data Research and Development Initiative" touted big data projects in areas with obvious privacy implications, such as military autonomous systems; cybersecurity; the smart grid; and transactional data from web searches, sensors, and cell phone records. No projects are aimed at addressing the privacy implications of big data.[23] It is clearly one-sided to stimulate the production and aggregation of highly revealing data and of sophisticated tools for analyzing that data without also stimulating privacy design awareness and privacy protection tools.

That needs to change. The government must conduct a full assessment of its big data policies, create new rules and oversight for big data, and be transparent about how it uses big data and algorithms.

First, there must be review of the government's use of big data in surveillance. Last June, documents revealed previously unknown collections of huge datasets by the National Security Agency (NSA). One such instance was the collection of Americans' calling records using Section

---

22 Gellman, Barton; Peterson, Andrea; Soltani, Ashkan. The Washington Post, December 10, 2013. http://www.washingtonpost.com/blogs/the-switch/wp/2013/12/10/nsa-uses-google-cookies-to-pinpoint-targets-for-hacking/

23 See Fact Sheet: Big Data Across the Federal Government (March 29, 2012).

215 of the Patriot Act. Although Section 215 is supposed to be used by the FBI, the NSA was able to compile a huge database of domestic calling records and run advanced algorithms on the dataset. These algorithms included querying a specific phone number in the dataset as well as using the dataset to create a social graph of certain phone numbers, often called "social chaining." The dataset is also used for other, still classified, techniques. There is a very real threat that Section 215 is not only used for the mass or bulk collection of calling records, but also for bulk collection of financial records, car rental records, and other "business records." Indeed, a recently released FISA Court order strongly suggests that Section 215 of the Patriot Act has been used to obtain mass financial records or purchase records.[24]

We've also seen such mass collection using Section 702 of the Foreign Intelligence Surveillance Amendments Act. Section 702 is used for at least two types of collection; other uses remain classified. The first type, known as PRISM, compels companies like Internet providers to turn over data like voice communications, email, video, chat messages, stored data, file transfers, VoIP calls, and other "digital network information" for information that is to, from, or about a "selector."[25] All of this information is collated into NSA databases, and disseminated to other database located at the FBI, NCTC, and CIA.[26] The second type is for "upstream" collection, under which telecom and Internet providers are required to work with NSA to copy, scan, and filter Internet and phone traffic coming through their physical infrastructure.[27] Both types of Section 702 collection target a foreign entity but acquire communications of persons in the United States.

All three of these types of collection are unconstitutional. In this age of big data it is beyond obvious that large datasets of metadata are highly revealing, and call detail records are especially so. Collection under Section 702 involves communications surveillance that has been

---

24 See http://www.dni.gov/files/documents/0328/104.%20BR%2010-82%20supplemental%20opinion%20-%20Redacted%2020140328.pdf. See also Senator Wyden, *Meet the Press*. March 30, 2014. *http://www.nbcnews.com/meet-the-press/meet-press-transcript-march-30-2014-n67356*.

25 Selectors are not exclusive to email address, phone calls, or other personally identifiable terms. Selectors are also called "targets" or "targeting selector."

26 Memorandum Opinion of October 3, 2011 by the Foreign Intelligence Surveillance Court ("Bates Opinion."). https://www.eff.org/document/october-3-2011-fisc-opinion-holding-nsa-surveillance-unconstitutional. Last accessed February 21, 2014.

27 "*NSA slides explain the PRISM data collection program*." Washington Post, June 6, 2013. http://www.washingtonpost.com/wp-srv/special/politics/prism-collection-documents/. Last accessed February 21, 2014.

within the purview of the Fourth Amendment for decades. Since *Katz v. United States*, the Supreme Court has repeatedly emphasized that the Fourth Amendment "protects people, not places,"[28] and has said that electronic surveillance presents a significant threat of "broad and unsuspected governmental incursions into conversational privacy."[29] These programs exemplify a big data nightmare of secret mass collection of data, secret human and automated big data analysis, and secret use of the results.

The government's current use of big data in these contexts must stop. Traditional investigatory techniques based on Constitutional norms of particularized and individualized suspicion have long acted as bedrock principles of intelligence collection and law enforcement techniques. Unfortunately, the lure of big data technology, shielded from public visibility, has led to a regime of bulk collection of communications, with so little judicial involvement that these programs are hardly distinguishable from the general warrants and writs of assistance that were the *raison d'etre* of the Fourth Amendment.

We urge the Administration to immediately stop misusing Section 215 of the Patriot Act and to support statutory reform to end mass collection of business records. Surveillance agencies should publicly disclose their mass spying techniques and issue Privacy Impact Assessments that set standards and address whether the agency is meeting them. As we've seen, disclosure in a responsible manner allows the public to engage in a vital discussion, and we already know the NSA is conducting such assessments; however, they remain classified.[30]

Concern over government use of big data extends to non-intelligence agencies. For instance, the Department of Homeland Security (DHS) is developing a new system called FALCON integrating over 17 different databases of information ranging from records of individuals who encounter law enforcement to student visa holders.[31] FALCON is an example of aggregating formerly separate databases to create a database with greater potential for privacy intrusions.[32]

---

28 *Katz v. United States* 389 US 347, 351-53 (1967).
29 *Katz v. United States* 389 US 347, 351-53 (1967).
30 Business Records FISA NSA Review June 25, 2009. Pg 44. https://www.eff.org/document/nsa-business-records-fisa-redactedex-ocr.
31 See http://www.dhs.gov/publication/dhsicepia-038-%E2%80%93-falcon-data-analysis-research-trade-transparency-system-falcon-dartts.
32 See page 1-3 of

DHS has also solicited bids to build and maintain a national database of motor vehicle license plate data.[33] Though the initial solicitation was recalled, it's disturbing that this bid was even issued in the first place. This trove of location data would reveal where you've been and when, and could be aggregated to present a detailed picture of your life and whom you associate with. Unsurprisingly, DHS had planned to use the data to target individuals. It wanted to be able to create its own "hot lists" of suspect vehicles from the data. Whether officers would have been required to articulate any individualized suspicion before putting a vehicle on a "hot list" is unclear; it's equally unclear how a vehicle would ever get off such a list.[34]

DHS also proposed sharing its "hot lists" with other agencies and wanted to be able to communicate with other users, "establish Lists submissions, flag license plates, and conduct searches anonymously." Meaningful oversight of the program would be impossible if officers could use the system anonymously. And while EFF has focused on big data privacy issues, both law enforcement and intelligence activity raise significant racial, ethnic and religious discrimination problems, exemplified by ICE's Secure Communities program and the NYPD's stop and frisk policy and surveillance of Muslim communities.[35]

Given the government's appetite for big data surveillance programs, one would expect strong evidence of their value in finding criminals and terrorists. But a 2008 study by the National Research Council concluded that data-mining in order to spot potential terrorists was ineffective, and that there was no clear evidence that behavioral surveillance was useful for counterterrorism operations.[36] In addition, a recent study by the New America Foundation analyzed 255 different terrorism cases and concluded NSA's bulk surveillance programs were of minimal value in supporting investigations.[37] Instead, the study demonstrated the strength of traditional investigative methods for initiating and advancing the investigations. There has also been extensive writing on the huge harms caused by sample bias and sample error in big

---

http://www.dhs.gov/sites/default/files/publications/privacy_pia_ice_falcondartts_january2014_0.pdf.
33 See https://www.eff.org/document/dhs-national-license-plate-reader-database-solicitation
34 See https://www.eff.org/deeplinks/2014/01/los-angeles-cops-should-release-automatic-license-plate-reader-records-eff-aclu
35 See http://fcir.org/2011/10/20/report-secure-communities-encourages-racial-profiling-lack-of-due-process/
36 National Research Council. *Protecting Individual Privacy in the Struggle Against Terrorists: A Framework for Program Assessment*. Washington, DC: The National Academies Press, 2008.
37 See http://newamerica.net/publications/policy/do_nsas_bulk_surveillance_programs_stop_terrorists

datasets.[38]

Suggested Approaches

To its credit, DHS—unlike intelligence agencies—seeks to follow the FIPPs[39] and issues Privacy Impact Assessments. The FIPPs provide a framework for the collection and usage of personal information generally, and can be seen as guiding principles for government and nongovernmental agencies dealing with sensitive personal information in a wide range of circumstances. The principles include:

> Purpose Specification: DHS should specifically articulate the authority that permits the collection of PII and specifically articulate the purpose or purposes for which the PII is intended to be used.
> Data Minimization: DHS should only collect PII that is directly relevant and necessary to accomplish the specified purpose(s) and only retain PII for as long as is necessary to fulfill the specified purpose(s).
> Use Limitation: DHS should use PII solely for the purpose(s) specified in the notice. Sharing PII outside the Department should be for a purpose compatible with the purpose for which the PII was collected.[40]

At a minimum, every agency should announce and use similar principles to guide its data activities, including its use of big data. Privacy Impact Assessments and System of Record Notices (SORNs) are also useful as they provide an overview of the collection, use of collection, and retention periods of collection.

More importantly, the government must commit to increasing its transparency around its big datasets. Overclassification is a chronic government problem. Time after time we've seen the witches' brew of ambiguity and secrecy poison democracy and the rule of law. It is widely noted by government officials, academics, and others that the classification system is broken.[41] The

---

38 See http://www.ft.com/intl/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html#axzz2xSL2ejQx
39 See, http://www.dhs.gov/xlibrary/assets/privacy/privacy_policyguide_2008-01.pdf
40 See http://www.dhs.gov/xlibrary/assets/privacy/privacy_policyguide_2008-01.pdf
41 See http://www.brennancenter.org/publication/reducing-overclassification-through-accountability

government—and particularly the intelligence agencies enamored with big data—must inform the public about what information they are collecting about US persons, and even innocent foreigners. Algorithmic transparency is key. Given the amount of data that the government collects, and the incentives of law enforcement and national security agencies to conceal much of what they do (either to collect or to use that data), attention to these problems in the specific context of government secrecy is crucial. No reform of big data for national security can be complete without this added transparency, which must include more affirmative disclosures as opposed to reactive declassifications. Director of National Intelligence General James Clapper recently lamented the fact that NSA should have disclosed the Section 215 Business Records FISA program collecting all Americans' calling records earlier. In particular, he said:

> …had we been transparent about this from the outset right after 9/11—which is the genesis of the 215 program—and said both to the American people and to their elected representatives, we need to cover this gap, we need to make sure this never happens to us again, so here is what we are going to set up, here is how it's going to work, and why we have to do it, and here are the safeguards…We wouldn't have had the problem we had.[42]

General Clapper should take this advice and apply it to any and all intelligence agency programs collecting big data about Americans.

The government must also stop the collection of big datasets that are clearly unconstitutional, like the collection of innocent Americans' calling records, phone calls, and emails. In addition, we suggest:

1. Stopping the big datasets collected by the illegal and unconstitutional use of Section 215 of the Patriot Act and Section 702 of the Foreign Intelligence Surveillance Amendments Act. These uses are conducted via an Executive interpretation of the law and can be stopped by the President. As such, we urge the Administration to simply stop misusing section 215 of the Patriot Act and to support statutory reform that ends mass collection of business records.
2. Calling for a Congressional investigation reviewing unclassified and classified intelligence programs collecting big data on US persons and innocent foreigners. Such a review could include a review of our surveillance programs; foreign policy implications of these programs; efficacy of the various collection techniques; a

---

42 Lake, Eli. *Spy Chief: We Should've Told You We Track Your Calls*, The Daily Best, February 17, 2014. https://web.archive.org/web/20140317070031/http://www.thedailybeast.com/articles/2014/02/17/spy-chief-we-should-ve-told-you-we-track-your-calls.html

review of the classification regime; and a review of the current oversight system, which includes reviewing the current Congressional oversight system.

3. The White House engage in a legislative strategy that not only encompasses surveillance reform, but also addresses the state secrets privilege or the standing to sue in surveillance litigation challenges.
4. Mandate Fair Information Practices Procedures across all government agencies. FIPPs can serve as a control on government's insatiable appetite for data. It will allow for proportional collection, identify the primary purpose of the data, and allow for a notice and disclosure to users. Other avenues include, mandating the release of Privacy Impact Assessments (PIAs) for intelligence agencies. Privacy Impact Assessment and System of Record Notices (SORNs) have provided the public with a notification of the collection and storage practices of users' information. Both can be improved. Sometimes PIAs and SORNs are not updated and fall out-of-date. Both should be reviewed annually to make sure they conform with the current use of the system.

Conclusion

It is a truism that big data is not going away but is only going to become more prevalent. In the future storage will only get cheaper, processing and analytics will only get faster, and both the private-sector and the government will be more incentivized to squeeze every last bit of information out of any data they can acquire. Additionally, the number and types of sources of big data will only increase as our daily lives become more digital. Self-driving cars equipped with cameras and other sensors and consumer products designed to be part of the Internet of things (eg, networked home appliances) will all soon have the capability to collect more data on individuals, which will no doubt be funneled back to company and government databases. Given these new challenges, not to mention the existing problems and dangers we have described, it is extremely important that the government take notice of how big data is being used by the private sector and work to ensure that consumers' privacy is preserved. Even more importantly, the government must take strong steps to end the misuse of big data by law enforcement and intelligence agencies, if for no other reason than to preserve Americans' Fourth Amendment rights. It is for these reasons that EFF strongly recommend that new policy

frameworks and regulations be implemented to fundamentally change how big data is collected, managed, and used.