Managing the Muddled Mass of Big Data

By Susan Freiwald
University of San Francisco School of Law

At the same time that "Big Data" promises previously unobtainable insights, its use places significant pressure on two significant methods of legal regulation to protect privacy. Specifically, because Big Data merges data from different sources, it renders ineffective legal regulation targeted to the method of data collection and exercised at the time of collection. In addition, Big Data makes obsolete regulation that relies on identifying a particular data holder or on keeping data of one type segregated from data of another type.

Managing the muddled mass of Big Data requires law makers to shift focus from how the data got there to how the data is being used and protected, if at all. It requires consideration of the value versus risk of having large databases, which in turn depends on the quality and security of their data, and the dangers from disclosure. Whenever Big Data projects involve risk to privacy and civil liberties, trustworthy experts should assess the value of the analytics they use in a transparent manner, and those results should be regularly assessed.

What is New About Big Data?

Prior to the era of Big Data, databases[1] held discrete sets of data, whose collection we could regulate, which were stored by an identifiable and stable source. In the private context, companies that sold goods and services recorded information electronically about their customers, as did health care providers,

[1] Databases are not new. I worked full-time as a database programmer more than 25 years ago.

banks, and credit card companies. Even online companies kept information about our web browsing, our searching, and our "likes" in their own proprietary databases. Law enforcement agents gathered information about a particular target, using a particular technique, such as an electronic pen register or a request for stored emails, and stored those records in a database.[2]

Big Data projects merge data from multiple places, which is how they get to be "Big". In the government context, the perceived need to find potential terrorists in our midst has led to the merger of data from multiple sources in fusion centers[3] and to the FBI joining forces with the NSA to gather up huge quantities of information from multiple sources. Big Data projects in the private sector involve data brokers pulling data from multiple sources to create behavioral profiles to yield the most effective targeted marketing.[4] While Big Data projects need good analytical tools based on sound logic, they work best, at least in theory, when they have the richest and deepest data to mine.

The depth of the data in Big Data comes from its myriad sources. To visualize, think of a Big Data database that has more information about a particular person (or entry) as adding to its length, in the sense that it spans a longer period (i.e., 5 years of John Doe's email records rather than 6 months). Adding entries for more people (e.g., adding in the emails of John Doe's wife and kids) increases its width. But Big Data has greater depth as well, in the sense that it can also analyze John Doe's web browsing data and his tweets. Because Big Data information

---

[2] *See, e.g.,* United States v. Forrester, 512 F.3d 500, 511 (9th Cir. 2008) (finding real-time collection of IP addresses by law enforcement agents to be unprotected by the Fourth Amendment); United States v. Warshak, 631 F.3d 266, 288 (6th Cir. 2010) (holding that acquisition of thousands of stored email without a warrant is unconstitutional). In both of these cases, law enforcement surely stored the electronic records they acquired in a database so they could search them for evidence.

[3] *See* Danielle Keats Citron and Frank Pasquale, *Network Accountability for the Domestic Intelligence Apparatus,* 62 HASTINGS L. J. 1441 (2011) (describing and critiquing fusion centers).

[4] *See* Neil M. Richards, *The Dangers of Surveillance*, 126 HARV. L. REV. 1934 (2013).

comes from multiple sources, the entity who analyzes it is quite likely not the one who gathered it.[5]

Regulation Based on Collection

In each of the commercial, law enforcement, and national security contexts, we have traditionally regulated at the point of data collection. Any data that has become untethered from its collector and the method by which it was collected becomes beyond the reach of those laws.[6]

Sectoral privacy laws place limits on what data may be collected, requiring that some personally identifiable data, in some contexts, be gathered only after data subjects give some kind of consent. The Children's Online Privacy Protection Act (COPPA),[7] which regulates the acquisition of information for marketing purposes about those under 13, provides perhaps the most rigorous regime, but regulations in the health care, financial, and cable context provide other examples.[8] Terms of service in the online context also permit, in varying degrees, those who contract with online companies to limit the extent to which those entities may collect and store information.

Those mechanisms are of limited use for those entities that operate outside of the specific parameters of the statutory definitions or outside of the contracts

---

[5] WhileTwitter stores tweets for some period of time, many other public and private are engaging in social network scraping, where they collect and store publicly available information, so a compiler of tweets may not be Twitter.

[6] This is in contrast to the European approach, which regulates data processing generally. *See* LOTHAR DETERMANN, DETERMANN'S FIELD GUIDE TO INTERNATIONAL DATA LAW COMPLIANCE xiii (2012) ("European data protection laws are first and foremost intended to restrict and reduce the automated processing of personal data – even if such data is publicly available.").

[7] 15 U.S.C. §§ 6501-6506 (as amended).

[8] Sectoral privacy laws regulate information gathered in the contexts of health care, (the Health Insurance Portability and Accountability Act Regulations or HIPAA); banking (the Gramm-Leach Bliley Act of 1999), cable (the Cable Communications Policy Act), videotape rentals (Video Privacy Protection Act) and others.

that terms of service arguably create. No sectoral law yet covers data brokers, for example, so their collection practices face no statutory regulation. And those who are covered by either statutory or contractual limits generally find ways to transfer information to third parties who are free of those limits. Once data ends up in the hands of Big Data processors, then, it has become free of legal constraints based on collection.

Privacy protections over data privacy in the law enforcement context reside in controls over how law enforcement may conduct surveillance. The Electronic Communications Privacy Act (ECPA) imposes procedural safeguards before agents may use of electronic devices to gather up information (email intercepts or modern pen registers) or compel the disclosure of electronic and related communications information from service providers.[9] But ECPA places no limits on buying data in bulk from commercial vendors, or amassing it in fusion centers, both of which enable the use of Big Data analysis for preventative law enforcement.

The recent revelations about Section 215 of the USA PATRIOT Act illustrate the executive branch's use of a terrorism-prevention rationale to avoid regulations geared towards collection. Even though the statute requires that information be gathered only when it is "relevant" to "protect against international terrorism or clandestine intelligence activities"[10] the executive branch has admitted to collecting all telephony metadata (non-content information) for calls within the United States over a period of years; apparently it does not query the database without some suspicion of wrongdoing.[11] By avoiding the statutory collection

[9] *See* 18 U.S.C. §§ 2510–2522 (2002) (regulating the interception of electronic communications); at 18 U.S.C. §§ 3121–27 (2010) (regulating the use of pen registers); 18 U.S.C. §§ 2701–11 (2010) (regulating the acquisition of stored communications and records).

[10] 50 U.S.C. § 1861(b)(2)(A) (2006).

[11] [best cite tbd].

limit, the executive is apparently subjecting itself to its own undisclosed and discretionary limits on its data access. The danger to civil liberties is obvious; through its almost certainly unconstitutional practices, the executive has amassed a gigantic database filled with all of our personal communication information.

Regulation Based on Identification

Another important method to protect privacy that Big Data renders ineffective is that dependent on the identification of a stable data collector. When someone becomes the target of inappropriate or unlawful data collection, she needs to be able to identify the data holder to have that holder purge the improperly collected data. That may be impossible to do with Big Data.

In the commercial context, COPPA requires that website operators accede to demands by parents to purge their databases of information about their children.[12] From the recently decided *Maryland v. King* case, we know that, under the state statute whose constitutionality the Supreme Court upheld, authorities destroy the DNA information of any arrestee who is subsequently found not guilty.[13] The minimization provisions of the Foreign Intelligence Surveillance Act (FISA) purport to get rid of improperly intercepted communications of U.S. persons as soon as it is determined that they are not relevant to foreign intelligence. For all of these mechanisms to work effectively, however, the data holder has to be stable and identifiable, and the data has to remain in place with that entity.

After data has been copied and sold to other entities, having it purged by the original collector does no good. When fusion centers merge data from private and

[12] 15 U.S.C. § 6502(b)(1)(B)(ii) (requiring the parent to "refuse to permit the operator's further use or maintenance in retrievable form, or future online collection, of personal information from that child").

[13] Maryland v. King, 133 S .Ct. 1958, 1967 (2013).

public sources into one master database, they do not indicate that to the original subject so that person can bring claims based on inappropriate use. Maryland may purge its own DNA database, but if the defendant's DNA has already been transferred to a central repository, it is unlikely to be purged after the defendant's acquittal. And of the many revelations that have come to light about the FISA minimization procedures, one indicates that the inadvertently collected communications of U.S. persons may be forwarded to the FBI for any law enforcement purpose.[14]

Regulation Based on Segregation

The merger of information in the foreign intelligence and law enforcement context illustrates another method of privacy protection that Big Data renders ineffective. Historically, the law has distinguished between data held by private entities from data held government entities. It has also treated surveillance for law enforcement purposes under an entirely different set of rules than surveillance for foreign intelligence gathering. Big Data has merged all data together.

Traditionally, we have been more concerned about private data in the hands of the government than we have been about private data in private hands. That is why the Privacy Act[15] regulates government data collection only and does not address private collection. It is also why ECPA permits electronic communications services providers (those who provide email, cell phone services, etc.) to voluntarily divulge records of those services to any non-government entity but not to governmental entities.[16] Once private intermediaries acquire such

[14] [FAA 702 minimization procedures recently released. Best cite tbd]
[15] 5 U.S.C. § 552a (2006).

[16] 18 U.S.C. §2702 (a)(3) (2006).

records, however, they are free to sell or give them to the government, which is undoubtedly how fusion center databases become populated with information.

In the past we erected virtual walls between the workings of domestic law enforcement and foreign intelligence agents. The former operated under much stricter standards, because citizens have constitutional rights that foreigners lack, and because protecting the nation's security carries more weight than ordinary crime fighting. Recent disclosures indicate that the FBI and the NSA have been working closely together to gather up the giant metadata database described above. The NSA apparently uses metadata databases (of both telephony and internet data) to hone its foreign intelligence queries. These actions mandate reform because it seems clear that the executive is operating under the weaker foreign intelligence standards to engage in work that further ordinary law enforcement goals. Big Data may be the focus of such reform.

Handling the Muddy Mass

With recognition of the problem the first step towards solving it, the next step does not involve reinventing the wheel. Academics[17] and expert commissions[18] have studied data mining at some length and come to several conclusions about how to minimize harm. Those insights themselves need to be mined as we supplement our ineffective legal approaches with ones that are effective for Big Data.

---

[17] *See e.g.*, Fred. H. Cate, *Government Data Mining, the Need for a Legal Framework,* 43 Harv. C.R.-C.L. L. Rev. 435 (2008); K.A. Taipale, *Data Mining and Data Security: Connecting the Dots to Make Sense of Data*, 5 COLUM. SCI. & TECH. L. REV. 2 (2005).

[18] *See, e.g.,* The Task Force on National Security in the Information Age, Markle Found., Creating a Trusted Network for Homeland Security (2003); Task Force on National Security in the Information Age, Markle Found., Mobilizing Information to Prevent Terrorism (2006); Tech. and Privacy Advisory Comm., U.S. Dep't of Def., Safeguarding Privacy in the Fight Against Terrorism (2004).

Those who have studied the issue agree on several key principles. Importantly, we must not be intimidated by the technically sophisticated nature of Big Data analysis. Even if we have to engage independent experts to do it, we should subject our data queries to oversight for effectiveness, and make sure we do not attribute unwarranted legitimacy to the results of Big Data queries.[19] Big Data programs must be much more transparent than they now are, so that the efficacy and fairness of their use can be monitored.

In addition, we must better appreciate that the mere accumulation of data in one place creates a risk both from insiders who abuse their access and outsiders who gain access. Because of those risks, data security, immutable audit trails, and meaningful accountability are also crucial features of effective Big Data regulations.

Conclusion

Big Data's depth represents its value and its challenge. By pulling data from a variety of sources into a single source, Big Data promises new answers to questions we may never have thought to ask. But it also fundamentally challenges regulations based on collection, identification, and segregation. Instead, we need to focus on transparency, expert review, efficacy, security and audit to reap the benefits of Big Data while minimizing the costs.

[19] Some computer experts have questioned the very premise of searching large databases for terrorist-planning patterns because we lack enough terrorist events to know what a plan looks like.