2012

The Challenge of "Big Data" for Data Protection

Fred H. Cate
Indiana University Maurer School of Law, fcate@indiana.edu

Christopher Kuner
Brussels Privacy Hub

Christopher Millard
Cloud Legal Project

Dan Jerker B. Svantesson
Bond University

Follow this and additional works at: http://www.repository.law.indiana.edu/facpub

Part of the Computer Law Commons, Information Security Commons, and the Privacy Law Commons

Recommended Citation
Cate, Fred H.; Kuner, Christopher; Millard, Christopher; and Svantesson, Dan Jerker B., "The Challenge of 'Big Data' for Data Protection" (2012). Articles by Maurer Faculty. 2620.
http://www.repository.law.indiana.edu/facpub/2620
Editorial

The challenge of ‘big data’ for data protection

Christopher Kuner*, Fred H. Cate**, Christopher Millard**, and Dan Jerker B. Svantesson***

Data protection, like almost everything else in our lives, is challenged by the advent of ‘big data’. The Economist reports in its 2012 Outlook that the quantity of global digital data expanded from 130 exabytes in 2005 to 1,227 in 2010, and is predicted to rise to 7,910 exabytes in 2015.¹

An exabyte is a quintillion bytes. If you find that hard to visualize, consider this: someone has calculated that if you loaded an exabyte of data on to DVDs in slimline jewel cases, and then loaded them into Boeing 747 aircraft, it would take 13,513 planes to transport one exabyte of data. Using DVDs to move the data collected globally in 2010 would require a fleet of more than 16 million jumbo jets.

And exabytes are rapidly becoming passé. The volume of stored information in the world is growing so fast that scientists have had to create new terms, including zetta-byte and yottabyte, to describe the flood of data.

The importance of big data is not just a result of its size or how fast it is growing (about 60 per cent a year), but also the reality that the data come from an amazing array of sources. The Internet captures lots of data. Facebook alone has more than 800 million active users, more than half of whom log in every day, where they generate more than 900 million web pages and upload more than 250 million photos every day.

In 2010, a lifetime ago in Internet time, Google sites were used by more than 1 billion unique visitors every month who spent a collective 200 billion minutes on its sites. Google-owned YouTube passed 1 trillion video playbacks in 2011. Email, IM, VOIP calls, and other communications generate tens of trillions of recorded messages every year.

Credit and debit cards, checks, and other financial activities provide a steady stream of billions of financial transactions recorded every month.

And increasingly sensor networks—video surveillance cameras, embedded computers in automobiles, the more than 5 billion cell phones we carry—record locations, movements, and activities. We can now talk meaningfully about ubiquitous data collection, in which almost everything we do results in data being captured and stored by one or more third parties.

It is significant that those data are digital. They can be stored, shared, searched, combined, and duplicated with extraordinary speed and at very little cost. And they are accompanied by metadata—data about when and where and how the underlying information was generated. Some experts estimate that there may be five times more metadata than the information we are aware of creating, and this metadata can be extraordinarily revealing.

We used to define ‘big data’ as being data sets so large that a supercomputer was needed to process them, but another aspect of big data has been that not only has analytical capacity soared, but also become far more inexpensive and widely distributed. It is not just that today’s mobile devices have more computing power than the desktop machines of a decade ago, but also that we can now link data and computers virtually so that huge computational tasks can be undertaken affordably and conveniently.

In fact, we are witnessing the movement of more of that computational power, as well as storage of the tidal wave of data we are generating and collecting, into the ‘cloud’. Cloud computing is all the rage, but despite the overuse and misuse of the term, it is increasingly clear that many of the data and resources we used to believe that we had to possess locally—in computers, handheld devices, entertainment systems, and business record systems—can now be provided with greater security and reliability (and at lower cost) remotely.

When thinking about the importance of ‘big data’, it is critical to remember that access to so much data, from so many different sources, and to the computing

---


* Editor-in-Chief
** Editor
***Managing Editor

© The Author 2012. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com
power necessary to process it, increasingly means we can perceive patterns, engage in discoveries, and discover secrets that were heretofore hidden.

This new-found power has already yielded fabulous successes in fields such as medical research, where drug interactions and the efficacy of treatments can be assessed in ways never before possible and without endangering lives through interventional research. But the same capacity has been shown to make deidentification more difficult, as Google and Netflix discovered when making available anonymized data sets for research.

One significant result of, and contributor to, big data development is how much we rely on data-based systems for critical decisions and applications. It is no exaggeration to say that we are nothing more than a collection of data to most of the institutions—and many of the people—with whom we deal. It is not simply that ‘our biographies are etched in the ones and zeros we leave behind in daily digital transactions,’ as Stanford Law School Professor Kathleen Sullivan has written, it is that it is those collections of zeros and ones are what identify, describe, and increasingly define us to others.

Big data poses enormous challenges for data protection—both by processors and regulators. It simultaneously changes the context and raises the stakes for data protection. Not surprisingly, given the pace of the change, there is little evidence that data protection is keeping up.

Consider, for example, the fascination shown by the EU data protection directive and the proposed EU General Data Protection Regulation, similarly to law in most of the rest of the world, with ‘notice’ and ‘choice’ or ‘consent’ as key tools of data protection. Despite mounting evidence that individuals ignore notices, often do not understand the choices (which often aren’t meaningful in any event), and resist making them unless compelled to do so (in which case they almost always make the choice required to obtain the desired service or product), regulators continue to cling to these concepts. But irrespective of the success of notice and choice to date, how will these tools fare in a world of ubiquitous surveillance, and thousands of data exchanges by and about every individual on the planet every day? In short order the largest database on the planet may be of legally required privacy notices that no one has read.

Even where legislative drafters demonstrate awareness that data processing on a very large scale may raise particular concerns, evidence that the practical risk implications are understood may be lacking. For example, the draft EU General Data Protection Regulation provides for an exception to the general prohibition on transfers of personal data to countries that lack adequate protection where a transfer ‘cannot be qualified as frequent and massive’. While the use of the term ‘massive’ hints at an appreciation of the challenge of big data, no attempt is made to define the concept or even to put it in a relative context.

Big data will also place data protection in a different context. We often talk about data being the ‘currency’ of the information age, but in a world in which data represent individuals in more and more transactions, and provide the basis for decision making, issues such as the accessibility, accuracy, and reliability of data may matter as much or maybe more than privacy.

This seems especially likely to prove true with cloud computing. Just as national security has tended to trump privacy in most anti-terrorism and law enforcement programmes, as critical data and applications are stored remotely, accessibility may become more important than privacy.

Big data also ratchet up the importance of harmonization, or even standardization, in data protection standards. As personal data are universally collected and shared across sectoral and national boundaries, inconsistent data protection laws pose increasing threats to individuals, institutions, and society.

Perhaps the greatest impact of big data is the pressure it brings for new thoughtful, informed, multinational debate about the key principles that should undergird data protection. Most data protection laws continue to rely on the 1980 OECD Guidelines. The Guidelines have weathered the intervening three decades well, but it is important to remember that they were crafted not merely before big data, but also before the World Wide Web, portable laptops, GPS, smartphones, tablet devices, or the myriad other innovations that make big data possible.

Identifying common principles to undergird data protection laws is critical not only to harmonizing those laws, but also to ensure that they serve worthwhile and appropriate ends. Big data highlights the need to focus not only on ‘what’ and ‘how’, but also on ‘why’. For example,

---


3 Article 44(1)(h).
• Are there limits as to what can be done with accurate, lawfully obtained information? Do they apply across the board, even to anti-terrorism and law enforcement activities? Focusing on use takes on new importance as technological applications lessen the role of law in regulating collection and sharing.

• Should there continue to be a ‘Collection Limitation Principle’ in a world of ubiquitous data and an apparently insatiable willingness of the public to supply its data online and off?

• Is privacy implicated if the government or a business looks at, but does not act on, lawfully obtained data, for example, when matching passenger lists with databases of suspected terrorists?

• Is there a proper role for individual consent? We believe the answer is clearly yes, but determining the contours of that role in a world of big data is not proving easy.

• Should there be a right to be forgotten—to have lawfully collected data deleted at the direction of the data subject? What does that mean even in a world of big data? Deleting data is easy, but stopping it from reappearing from another source is much harder. How do we balance the effort to do so with shared national commitments to freedom of expression?

• What is the proper and practical role for government in the face of a deluge of digital data? The sheer volume of personal data suggests that government must establish in law basic rights and obligations, but how are these to be enforced given the imbalance in resources? How do governments or independent regulators exercise meaningful oversight of the more than 60 billion credit and debit card transactions, 90 trillion emails, 1.4 trillion Internet searches, and vast number of other data transfers that occur every year?

• Governments are also among the largest collectors and users of personal data. Indeed, the advent of big data has already shown signs of marking a feeding frenzy among government agencies wanting extraordinary access to information to identify money laundering and terrorist financing transactions, locate suspect offenders, identify and block child pornography or other regulated expression, enforce intellectual property laws, or just alleviate the administrative burden of governments. A recent example of an ambitious approach to online intelligence gathering is an expression of interest by the Strategic Information and Operations Center of the US Federal Bureau of Investigations (‘FBI SIOC’) in developing a ‘social media alert, mapping, and analysis application solution’. Among many other things, FBI SIOC would like to ‘[g]eo-spatially locate bad actors or groups and analyze their movements, vulnerabilities, limitations, and possible adverse actions’. Beyond that, the application might enable FBI SOIC to ‘[p]redict likely developments in the situation or future actions taken by bad actors...’4

• Do we need to rethink the concept or the application of national sovereignty in the context of inherently global data flows? As European Data Protection Supervisor Peter Hustinx noted in 2007: ‘Whereas the jurisdiction of the Community legislator is limited to the territory of the European Union, the external borders become less relevant for data flows. The economy depends more and more on global networks... In general, the physical place of a processing operation is less relevant.’5

*International Data Privacy Law* exists in large part to help facilitate a thoughtful, informed, multinational debate about the principles to undergird society’s responses to these and other data challenges. In the pages that follow in this issue and future issues, our contributors not only address ways of improving data protection in the face of big data, but also remind us of the vital need to work together to identify the principles that should guide our efforts.


---
