Editors: Michael Lesk, lesk@acm.org | Jeffrey MacKie-Mason, jmm@umich.edu

Big Data, Big Brother, Big Money

Michael Lesk | Rutgers University

Historically, most of our fears about surveillance and privacy have involved the fear of a controlling government. This is the warning from dystopian fiction such as George Orwell's *1984* or Franz Kafka's *The Trial*, and it is what has happened in Communist states such as East Germany and is now happening in the US. We think of it as purely governmental activity, designed to control what people say and think. What's new today is that surveillance has been outsourced, because most data is in private hands.

Our Former Fears

George Orwell's and Jeremy Bentham's fictional visions of the watched society are well known:[1,2]

> With the development of television, and the technical advance which made it possible to receive and transmit simultaneously on the same instrument, private life came to an end. Every citizen, or at least every citizen important enough to be worth watching, could be kept for twenty-four hours a day under the eyes of the police and in the sound of official propaganda, with all other channels of communication closed. The possibility of enforcing not only complete obedience to the will of the State, but complete uniformity of opinion on all subjects, now existed for the first time.

> The more constantly the persons to be inspected are under the eyes of the persons who should inspect them, the more perfectly will the purpose X of the establishment have been attained. Ideal perfection, if that were the object, would require that each person should actually be in that predicament, during every instant of time. This being impossible, the next thing to be wished for is, that, at every instant, seeing reason to believe as much, and not being able to satisfy himself to the contrary, he should conceive himself to be so.

Bentham's model of continuous observation was implemented in the design of the former British Museum Reading Room (1857–1997). The room was circular, with the reading desks laid out along radii. In the center was a raised platform, where one man could stand and see all the readers. When the Library of Congress opened its own reading room in 1897, it copied the idea of a round room with a dome, including large amounts of glass for light. However, it laid out the reading desks in concentric circles, either not understanding what the British Museum design was about or feeling little risk of theft or vandalism.

While I wrote this article, news of US government snooping came to light, including collection of at least telephone call

1540-7993/13/$31.00 © 2013 IEEE Copublished by the IEEE Computer and Reliability Societies

metadata—called and calling numbers and times—and email involving non-US participants. This snooping, however, was via data collected by others. The large databases about individuals today start with private corporations, not the government. A new business model, exemplified by Facebook, collects as much information as possible about you in order to sell targeted advertising. Some of this is old; from the early days of advertising, products specifically aimed at men, women, or children would appear with broadcast programs or magazine articles aimed at that specific group. What's new today is the amount of data that can be kept and searched, the sense of being constantly observed, the precision of the surveillance, and the merging of data from multiple sources. Searches, emails, geographic position, and other observations combine to deliver ads for specific products in your particular location. So, when the government decided to look for terrorists, it went to the corporations that had the data.

The extent of private data collection far outstrips what's known about public data collection. For example, Chicago has approximately 10,000 video surveillance cameras, of which about 10 percent belong to the police. After the Boston bombings in April 2013, the police found the best images of the perpetrators on private cameras, not public ones. Record-keeping of Web searches, online purchases, and emails is even more pervasive. Other aspects of surveillance technology continue to move into the private sector; a few decades ago, there was virtually no cryptography except by the government, and now we have vast amounts of encoded Web traffic.

We now know that the US government has been piggybacking its surveillance on commercial services. It's been collecting metadata about everybody's telephone calls for more than a decade, while telephone companies have been using the same information commercially. Mobile telephone carriers exploit location data from your calls and sell it to others; see, for example, Verizon's "Precision Market Insights" service.[3] Alerting based on the contents of non-US email seems similar to ad targeting based on Gmail and Twitter content.

We now know that the US government has been piggybacking its surveillance on commercial services.

Datacenters

Can we compare government and private data collection? Little is said openly, but we can look at the size of datacenters. Many organizations don't report details of their facilities, but some scraps of information are available. Amazon is building a US$600 million cloud for the Central Intelligence Agency, but that's not large today. Bloomberg is building a $710 million server, and Microsoft is completing a $1 billion center. The National Security Agency's (NSA) new datacenter in Utah will cost $2 billion, and Facebook is building a single center that will cost $1.5 billion and also has several others. Google has eight datacenters, with more coming; the typical cost for one is $600 million.

In terms of storage, again, people are reluctant to give sizes, but one estimate suggested that the NSA was collecting 14 Pbytes per year. Facebook already has 100 Pbytes, Microsoft 300 Pbytes, and Amazon 900 Pbytes. Perhaps the most concrete comparison available is power consumption. The NSA's center will use 65 MW. Facebook's new Oregon center is planned for 78 MW, the new Apple datacenter in North Carolina uses 100 MW, and Google uses 220 MW, despite efforts to be as efficient as possible in running machines. A single Microsoft center in Chicago uses almost 200 MW to run 224,000 servers; Google is estimated to have more than 900,000 computers. Processor counts are harder to pin down than megawatts, because some companies count servers and others count cores. But there's still no question that multiple commercial systems are now bigger than the ones spy agencies run.

Commerce in Data

Huge amounts of personal data are sold constantly. The three main credit recording services with large revenues are Equifax ($1.8 billion in revenues), Experian ($4.4 billion) and Transunion ($1.2 billion); the latter two, particularly Experian, also have other business lines. The largest mailing list vendor, Acxiom, has revenues of $1.13 billion, and the total direct-mail industry has revenues of $11 billion. According to *The New York Times*, Acxiom has records on approximately 500 million people with 1,500 data points on each of them. It has 23,500 servers in a 250,000 sq. ft. datacenter handling 12 Pbytes—and remember, this is only one of its datacenters.[4] So, a company you've probably never heard of is handling data on the same scale as the NSA. Equifax is even bigger; it has 26 Pbytes and boasts that it has more data than the US Federal Bureau of Investigation. To quote its CIO, Dave Webb, "We know more about you than you would care for us to know."[5]

How much do companies pay for your information? Just your name is of limited value; mailing list companies traditionally sell names for 5 to 10 cents each. More precise

data can be very valuable. A tool called Swipe (http://turbulence.org/Works/swipe) estimates the value of different pieces of information. Its website's front page suggests that address plus date of birth plus phone number, Social Security number, and driver's license are worth $13.75 to marketers. Two reporters estimated in 1999 that music data is worth up to $40 apiece to marketers because it can help them spot true fanatics who will buy anything and everything related to their favorite artists.[6]

The US government has acquired commercial data before for other purposes. For example, the Internal Revenue Service can use credit card data to select taxpayers for audit.[7] This hasn't produced anything like the same outcry that arose from national security uses of personal information. And yet, in both cases, generalized scanning is used fairly broadly, with warrants required to follow up with more invasive searches.

Bad Data vs. Good Data

When the media do get outraged about private data, often they highlight mistakes made using erroneous data. According to the Federal Trade Commission, 20 percent of credit reports contain bad information. Correcting these errors is a tedious and drawn-out process (more than one litigant claims to have spent six years trying to get an error removed from personal files). Other bad data problems involve identity theft. The US Department of Justice estimates the cost of identity theft in the US at $13 billion per year. Large commercial data services point out that they constantly use their data for fraud detection and thus avert immense financial losses for both themselves and their customers.

Erroneous data propagates itself into incorrect deductions. Sandy Pentland of the Massachusetts Institute of Technology suggests that 70 to 80 percent of machine learning results are wrong.[8] This doesn't even have to involve mistakes in data; it can simply be a result of searching too hard to find something. When I worked at Bell Labs, a cautionary example was that there was a significant correlation between the number of phone calls made in Washington, DC, and the level of the Potomac River. This might sound to you like nonsense until I tell you that when they're both high, it's raining. So the correlation is real, but don't jump to causation: you can't raise the river level during a drought by asking people to make more phone calls.

Private data collection is subject to few constraints in the US. What you think about big data in private hands depends partly on how it's used.

However, the problems that arise from incorrect data are perhaps less serious than the issues that arise from accurate data, which is being used in controlling and invasive ways. Is it a step forward or back if people are charged higher insurance rates because something about them makes them a riskier prospect? We're used to auto insurance companies charging people more if they live in the wrong place. However, Congress has forbidden health insurance companies to charge more to people who have genetic conditions suggesting susceptibility to disease. What about the way airlines use travel data to offer many different prices to passengers on the same flight? Is that a benefit because flights that might not otherwise cover their costs can operate thanks to differential pricing, or is it in some way cheating the people who are offered the high fares?

"Big data" in government hands is already used for activities like the "no fly" list and border searches. Although law enforcement needs to maintain confidentiality of its procedures, there are at least some oversight rules (warrants, Congressional supervision, and the like). Private data collection is subject to few constraints in the US.

What you think about big data in private hands depends partly on how it's used. For example, researchers at Kaiser Permanente found that children born to mothers who used antidepressant drugs during pregnancy have double the risk of autism-related illness. If this leads to a way to prevent autism, that's good. If it means that medical insurers will start refusing coverage to families in which someone uses antidepressants, depriving them of medical care, that's something most people would consider bad. There are real risks to public health if people avoid seeing doctors for fear that knowledge of medical problems will be used against them.

Intellectually, online search services are now getting quite good at showing you only opinions with which you will agree. Eli Pariser calls this the "filter bubble," a world in which, at the extreme, we'd never see anything new, just things we already believe.[9] Does that increase divisiveness and partisanship in society, or is it just saving you reading time?

The correlations used to judge people's interests have limits, of course. Perhaps the most famous example is exemplified by an article in *The Wall Street Journal* entitled, "If TiVo Thinks You Are Gay, Here's How to Set It Straight,"

describing a man who recorded the movie *The Sex Monster* because he was the writer-director and star.[10] TiVo started suggesting so many gay-themed movies that he tried recording the Playboy Channel to change its opinion of him, resulting in so many soft-core porn movies that his wife got upset.

Mistakes shouldn't obscure the risks from even correctly used information. It's not just that we don't like being watched; it's that some degree of enjoyment in life comes from freedom, which is reduced by imagining constant observation, as Bentham proposed for criminals. It might be frightening to think of the government watching to see if you're breaking the law, or it might be simply annoying to find that your mailbox and browser screen are full of ads resulting from what you thought was an idle inquiry.

Does It Work Both Ways?

If businesses monitor us, can we monitor them? The Web has made comparison shopping much easier. Whereas the automotive industry traditionally had obscure pricing, it's now easy to find automobile prices online. This has reduced the margins on car purchases by 25 percent, and the range of car prices has shrunk 35 percent. It's estimated that the public is saving $1 billion per year. This can be win-win; in an earlier article, I wrote about fishermen in India who used cellphones to find the ports where their catches would yield the highest price, which was win-win for both the fishermen and their customers because fewer fish rotted before they could be sold.[11]

As you might expect, businesses don't like consumers monitoring them. As I write this article, Bloomberg is in disrepute for having snooped on what its banking customers, such as JPMorgan and Goldman Sachs, were doing on their computer terminals—and Bloomberg was looking at message content, not just addressees. Early in the Web's history, Andersen Consulting (now Accenture) introduced a bot to gather price data, and many vendors immediately blocked it. Retailers regularly complain about applications like the Amazon Price Check app that lets you compare prices in conventional stores with Amazon's prices. Former Senator Olympia Snowe (R, ME) denounced the Amazon app as "an attack on Main Street businesses that employ workers in our communities."[12]

In his new book, *Who Owns the Future?*, Jaron Lanier suggested that companies should pay us for the information they gather from searching and browsing.[13] He imagines, referring to Ted Nelson's decades-old proposal for "transclusion," that for each bit of information you provide, Facebook or Google could send a little money back to you when the service based on that information monetizes it. Alternatively, there could be a government claim on publicly produced knowledge, which Lanier calls a "spy tax"; the information collected from everybody should produce some revenue for everybody. His book calls for an attempt to redress the balance of information, rather than as it is today when the companies know a lot about you and you know little about them. By contrast, a variety of "sunshine laws" do make considerable government data available, including, for example, the salaries of every state employee in many states.

Living with Our Past

The permanence of personal data has social impacts that last for decades because "the Internet never forgets."[14] Students are routinely warned that they must be very careful about what they post online, but all sorts of information can appear to influence your later life. For example, in 1987, Douglas Ginsburg was nominated for the US Supreme Court but had to withdraw his name over allegations that he had smoked marijuana in the 1960s and 1970s. If young people must keep thinking about anything they do that might be later captured and used against them, the logical consequence will be that people avoid anything risky. Some of our most successful corporate founders—people like Steve Jobs, Larry Ellison, and Bill Gates—didn't have conventional careers but took many risks. Even though David Mamet wrote that the only second chance in life is the chance to make the same mistake again, we would like to believe in second chances and the ability to recover from an error.[15] This has historically been recognized in the culture of startup companies, which institutionalized an old saying that good judgment comes from experience, and experience comes from bad judgment. Eliminating opportunities for those who have made a single misstep in their youth could stifle entrepreneurship, which would be a disaster for our economic future.

If the fear of Big Brother is now a fear of large corporations as well as the government, is there anything we can do about it? Rules about search warrants don't apply to private industry, and when you click on the typical license to use a service, you usually waive any rights you think you have. Sometimes, to preserve privacy, people will pay with cash; as Netflix was growing, I observed that although conventional video stores were closing, the stores selling pornographic DVDs were surviving, presumably because people didn't want to use credit cards.

Remaining below the radar screen is increasingly difficult, although some people try hard. The benefits of systems such as email

and telephony are too great; we need instead to look at regulatory solutions. For criminal law, by the time you read this, there will have been substantial discussion of the use of warrants and the special court for foreign intelligence. For commerce, European countries have laws about collecting data on individuals, for example, requiring that individuals have access to a copy of the data and a chance to correct errors. The European Data Protection Directive might serve as a first example of regulation to preserve at least some economic privacy. ■

References

1. G. Orwell, *1984*, Secker and Warburg, 1949.
2. J. Bentham, *Panopticon, Letters to Lord Pelham*, 1798.
3. J. Lieber, "How Wireless Carriers Are Monetizing Your Movements," *MIT Technology Rev.*, 12 Apr. 2013; www.technologyreview.com/news/513016/how-wireless-carriers-are-monetizing-your-movements.
4. N. Singer, "Mapping, and Sharing, the Consumer Genome," *New York Times*, 16 June 2012; www.nytimes.com/2012/06/17/technology/acxiom-the-quiet-giant-of-consumer-database-marketing.html?pagewanted=all.
5. K.S. Nash, "Equifax Eyes Are Watching You—Big Data Means Big Brother," *CIO*, 15 May 2012; www.cio.com/article/706457/Equifax_Eyes_Are_Watching_You_Big_Data_Means_Big_Brother.
6. J. Sullivan and C. Jones, "How Much Is Your Playlist Worth?," *Wired*, 3 Nov. 1999; www.wired.com/science/discoveries/news/1999/11/32258?currentPage=all.
7. R. Satran, "Next Target of IRS Robo-Audits: Small Business," *US News & World Report*, 9 May 2013; http://news.yahoo.com/next-target-irs-robo-audits-small-business-205502276.html.
8. A. Pentland, "Big Data's Biggest Obstacles," Harvard Business Rev. Blog, 2 Oct. 2012; http://blogs.hbr.org/cs/2012/10/big_datas_biggest_obstacles.html.
9. E. Pariser, *The Filter Bubble*, Penguin, 2011.
10. J. Zaslow, "If TiVo Thinks You Are Gay, Here's How to Set It Straight," *The Wall Street J.*, 26 Nov. 2002; http://online.wsj.com/article/SB103826193687235690 8.html.
11. R. Jensen, "The Digital Provide: Information (Technology), Market Performance, and Welfare in the South Indian Fisheries Sector," *Quarterly J. Economics*, vol. 122, no. 3, 2007, pp. 879–924.
12. J. Easley, "Snowe Rips Amazon for 'Incentivizing Consumers to Spy on Local Shops,'" *The Hill*, 9 Dec. 2011; http://thehill.com/blogs/blog-briefing-room/news/198329-snowe-rips-amazon-for-incentivizing-consumers-to-spy-on-local-shops.
13. J. Lanier, *Who Owns the Future?*, Simon & Schuster, 2013.
14. J. Rosen, "The Web Means the End of Forgetting," *New York Times*, 21 July 2010, p. MM30.
15. D. Mamet, *State and Main* (movie script), Fine Line Features, 2010.

Michael Lesk is a professor at Rutgers University in New Brunswick, New Jersey. Contact him at lesk@acm.org.