# 2

# Data as Surveillance

Governments and corporations gather, store, and analyze the tremendous amount of data we chuff out as we move through our digitized lives. Often this is without our knowledge, and typically without our consent. Based on this data, they draw conclusions about us that we might disagree with or object to, and that can impact our lives in profound ways. We may not like to admit it, but we are under mass surveillance.

Much of what we know about the NSA's surveillance comes from Edward Snowden, although people both before and after him also leaked agency secrets. As an NSA contractor, Snowden collected tens of thousands of documents describing many of the NSA's surveillance activities. In 2013, he fled to Hong Kong and gave them to select reporters. For a while I worked with Glenn Greenwald and the *Guardian* newspaper, helping analyze some of the more technical documents.

The first news story to break that was based on the Snowden documents described how the NSA collects the cell phone call records of every American. One government defense, and a sound bite repeated ever since, is that the data collected is "only metadata." The intended point was that the NSA wasn't collecting the words we spoke during our phone conversations, only the phone numbers of the two parties, and the date, time, and duration of the call. This seemed to mollify many people, but it shouldn't have. Collecting metadata on people means putting them under surveillance.

An easy thought experiment demonstrates this. Imagine that you hired a private detective to eavesdrop on someone. The detective would plant bugs in that person's home, office, and car. He would eavesdrop on that person's phone and computer. And you would get a report detailing that person's conversations.

Now imagine that you asked the detective to put that person under surveillance. You would get a different but nevertheless comprehensive report: where he went, what he did, who he spoke with and for how long, who he wrote to, what he read, and what he purchased. That's metadata.

Eavesdropping gets you the conversations; surveillance gets you everything else.

Telephone metadata alone reveals a lot about us. The timing, length, and frequency of our conversations reveal our relationships with others: our intimate friends, business associates, and everyone in-between. Phone metadata reveals what and who we're interested in and what's important to us, no matter how private. It provides a window into our personalities. It yields a detailed summary of what's happening to us at any point in time.

A Stanford University experiment examined the phone metadata of about 500 volunteers over several months. The personal nature of what the researchers could deduce from the metadata surprised even them, and the report is worth quoting:

- Participant A communicated with multiple local neurology groups, a specialty pharmacy, a rare-condition management service, and a hotline for a pharmaceutical used solely to treat relapsing multiple sclerosis.
- Participant B spoke at length with cardiologists at a major medical center, talked briefly with a medical laboratory, received calls from a pharmacy, and placed short calls to a home reporting hotline for a medical device used to monitor cardiac arrhythmias.
- Participant C made a number of calls to a firearms store that specializes in the AR semiautomatic rifle platform, and also spoke at length with customer service for a firearm manufacturer that produces an AR line.

- In a span of three weeks, Participant D contacted a home improvement store, locksmiths, a hydroponics dealer, and a head shop.
- Participant E had a long early morning call with her sister. Two days later, she placed a series of calls to the local Planned Parenthood location. She placed brief additional calls two weeks later, and made a final call a month after.

That's a multiple sclerosis sufferer, a heart attack victim, a semiautomatic weapons owner, a home marijuana grower, and someone who had an abortion, all from a single stream of metadata.

Web search data is another source of intimate information that can be used for surveillance. (You can argue whether this is data or metadata. The NSA claims it's metadata because your search terms are embedded in the URLs.) We don't lie to our search engine. We're more intimate with it than with our friends, lovers, or family members. We always tell it exactly what we're thinking about, in words as clear as possible. Google knows what kind of porn each of us searches for, which old lovers we still think about, our shames, our concerns, and our secrets. If Google decided to, it could figure out which of us is worried about our mental health, thinking about tax evasion, or planning to protest a particular government policy. I used to say that Google knows more about what I'm thinking of than my wife does. But that doesn't go far enough. Google knows more about what I'm thinking of *than I do*, because Google remembers all of it perfectly and forever.

I did a quick experiment with Google's autocomplete feature. This is the feature that offers to finish typing your search queries in real time, based on what other people have typed. When I typed "should I tell my w," Google suggested "should i tell my wife i had an affair" and "should i tell my work about dui" as the most popular completions. Google knows who clicked on those completions, and everything else they ever searched on.

Google's CEO Eric Schmidt admitted as much in 2010: "We know where you are. We know where you've been. We can more or less know what you're thinking about."

If you have a Gmail account, you can check for yourself. You can look at your search history for any time you were logged in. It goes back for as long as you've had the account, probably for years. Do it; you'll be surprised. It's more intimate than if you'd sent Google your diary. And even though Google lets you modify your ad preferences, you have no rights to delete anything you don't want there.

There are other sources of intimate data and metadata. Records of your purchasing habits reveal a lot about who you are. Your tweets tell the world what time you wake up in the morning, and what time you go to bed each night. Your buddy lists and address books reveal your political affiliation and sexual orientation. Your e-mail headers reveal who is central to your professional, social, and romantic life.

One way to think about it is that data is content, and metadata is context. Metadata can be much more revealing than data, especially when collected in the aggregate. When you have one person under surveillance, the contents of conversations, text messages, and e-mails can be more important than the metadata. But when you have an entire population under surveillance, the metadata is far more meaningful, important, and useful.

As former NSA general counsel Stewart Baker said, "Metadata absolutely tells you everything about somebody's life. If you have enough metadata you don't really need content." In 2014, former NSA and CIA director Michael Hayden remarked, "We kill people based on metadata."

The truth is, though, that the difference is largely illusory. It's all data about us.

## CHEAPER SURVEILLANCE

Historically, surveillance was difficult and expensive. We did it only when it was important: when the police needed to tail a suspect, or a business required a detailed purchasing history for billing purposes. There were exceptions, and they were extreme and expensive. The exceptionally paranoid East German government had 102,000 Stasi surveilling a population of 17 million: that's one spy for every 166 citizens, or one for every 66 if you include civilian informants.

Corporate surveillance has grown from collecting as little data as necessary to collecting as much as possible. Corporations always collected information on their customers, but in the past they didn't collect very much of it and held it only as long as necessary. Credit card compa-

nies collected only the information about their customers' transactions that they needed for billing. Stores hardly ever collected information about their customers, and mail-order companies only collected names and addresses, and maybe some purchasing history so they knew when to remove someone from their mailing list. Even Google, back in the beginning, collected far less information about its users than it does today. When surveillance information was expensive to collect and store, corporations made do with as little as possible.

The cost of computing technology has declined rapidly in recent decades. This has been a profoundly good thing. It has become cheaper and easier for people to communicate, to publish their thoughts, to access information, and so on. But that same decline in price has also brought down the price of surveillance. As computer technologies improved, corporations were able to collect more information on everyone they did business with. As the cost of data storage became cheaper, they were able to save more data and for a longer time. As big data analysis tools became more powerful, it became profitable to save more information. This led to the surveillance-based business models I'll talk about in Chapter 4.

Government surveillance has gone from collecting data on as few people as necessary to collecting it on as many as possible. When surveillance was manual and expensive, it could only be justified in extreme cases. The warrant process limited police surveillance, and resource constraints and the risk of discovery limited national intelligence surveillance. Specific individuals were targeted for surveillance, and maximal information was collected on them alone. There were also strict minimization rules about not collecting information on other people. If the FBI was listening in on a mobster's phone, for example, the listener was supposed to hang up and stop recording if the mobster's wife or children got on the line.

As technology improved and prices dropped, governments broadened their surveillance. The NSA could surveil large groups—the Soviet government, the Chinese diplomatic corps, leftist political organizations and activists—not just individuals. Roving wiretaps meant that the FBI could eavesdrop on people regardless of the device they used to communicate with. Eventually, US agencies could spy on entire populations and save the data for years. This dovetailed with a changing threat, and they continued espionage against specific governments, while expanding mass surveillance of broad populations to look for potentially dangerous individuals. I'll talk about this in Chapter 5.

The result is that corporate and government surveillance interests have converged. Both now want to know everything about everyone. The motivations are different, but the methodologies are the same. That is the primary reason for the strong public-private security partnership that I'll talk about in Chapter 6.

To see what I mean about the cost of surveillance technology, just look how cheaply ordinary consumers can obtain sophisticated spy gadgets. On a recent flight, I was flipping through an issue of *SkyMall*, a catalog that airlines stick in the pocket of every domestic airplane seat. It offered an $80 pen with a hidden camera and microphone, so I could secretly record any meeting I might want evidence about later. I can buy a camera hidden in a clock radio for $100, or one disguised to look like a motion sensor alarm unit on a wall. I can set either one to record continuously or only when it detects motion. Another device allows me to see all the data on someone else's smartphone—either iPhone or Android—assuming I can get my hands on it. "Read text messages even after they've been deleted. See photos, contacts, call histories, calendar appointments and websites visited. Even tap into the phone's GPS data to find out where it's been." Only $120.

From other retailers I can buy a keyboard logger, or keylogger, to learn what someone else types on her computer—assuming I have physical access to it—for under $50. I can buy call intercept software to listen in on someone else's cell phone calls for $100. Or I can buy a remote-controlled drone helicopter with an onboard camera and use it to spy on my neighbors for under $1,000.

These are the consumer items, and some of them are illegal in some jurisdictions. Professional surveillance devices are also getting cheaper and better. For the police, the declining costs change everything. Following someone covertly, either on foot or by car, costs around $175,000 per month—primarily for the salary of the agents doing the following. But if the police can place a tracker in the suspect's car, or use a fake cell tower device to fool the suspect's cell phone into giving up its location information, the cost drops to about $70,000 per month, because it only requires

people are steered away from traffic jams so they don't add to them. But are we aware of how much data we're giving away?

For the first time in history, governments and corporations have the ability to conduct mass surveillance on entire populations. They can do it with our Internet use, our communications, our financial transactions, our movements . . . everything. Even the East Germans couldn't follow everybody all of the time. Now it's easy.

## HIDDEN SURVEILLANCE

If you're reading this book on a Kindle, Amazon knows. Amazon knows when you started reading and how fast you read. The company knows if you're reading straight through, or if you read just a few pages every day. It knows if you skip ahead to the end, go back and reread a section, or linger on a page—or if you give up and don't finish the book. If you highlight any passages, Amazon knows about that, too. There's no light that flashes, no box that pops up, to warn you that your Kindle is sending Amazon data about your reading habits. It just happens, quietly and constantly.

We tolerate a level of electronic surveillance online that we would never allow in the physical world, because it's not obvious or advertised. It's one thing for a clerk to ask to see an ID card, or a tollbooth camera to photograph a license plate, or an ATM to ask for a card and a PIN. All of these actions generate surveillance records—the first case may require the clerk to copy or otherwise capture the data on the ID card—but at least they're overt. We know they're happening.

Most electronic surveillance doesn't happen that way. It's covert. We read newspapers online, not realizing that the articles we read are recorded. We browse online stores, not realizing that both the things we buy and the things we look at and decide not to buy are being monitored. We use electronic payment systems, not thinking about how they're keeping a record of our purchases. We carry our cell phones with us, not understanding that they're constantly tracking our location.

Buzzfeed is an entertainment website that collects an enormous amount of information about its users. Much of the data comes from traditional Internet tracking, but Buzzfeed also has a lot of fun quizzes, some of which ask very personal questions. One of them—"How Privileged Are You?"—asks about financial details, job stability, recreational activities, and mental health. Over two million people have taken that quiz, not realizing that Buzzfeed saves data from its quizzes. Similarly, medical information sites like WebMD collect data on what pages users search for and read.

Lest you think it's only your web browsing, e-mails, phone calls, chats, and other electronic communications that are monitored, old-fashioned paper mail is tracked as well. Through a program called Isolation Control and Tracking, the US Postal Service photographs the exterior, front and back, of *every piece of mail sent in the US*. That's about 160 billion pieces annually. This data is available to law enforcement, and certainly other government agencies as well.

Off the Internet, many surveillance technologies are getting smaller and less obtrusive. In some cities, video cameras capture our images hundreds of times a day. Some are obvious, but we don't see a CCTV camera embedded in a ceiling light or ATM, or a gigapixel camera a block away. Drones are getting smaller and harder to see; they're now the size of insects and soon the size of dust.

Add identification software to any of these image collection systems, and you have an automatic omnipresent surveillance system. Face recognition is the easiest way to identify people on camera, and the technology is getting better every year. In 2014, face recognition algorithms started outperforming people. There are other image identification technologies in development: iris scanners that work at a distance, gait recognition systems, and so on.

There's more hidden surveillance going on in the streets. Those contactless RFID chip cards in your wallet can be used to track people. Many retail stores are surreptitiously tracking people by the MAC addresses and Bluetooth IDs—which are basically identification numbers—broadcast by their smartphones. The goal is to record which aisles they walk down, which products they stop to look at, and so on. People can be tracked at public events by means of both these approaches.

In 2014, a senior executive from the Ford Motor Company told an audience at the Consumer Electronics Show, "We know everyone who breaks the law, we know when you're doing it. We have GPS in your car, so we know what you're doing." This came as a shock and surprise, since no

one knew Ford had its car owners under constant surveillance. The company quickly retracted the remarks, but the comments left a lot of wiggle room for Ford to collect data on its car owners. We know from a Government Accountability Office report that both automobile companies and navigational aid companies collect a lot of location data from their users.

Radar in the terahertz range can detect concealed weapons on people, and objects through eight inches of concrete wall. Cameras can "listen" to phone conversations by focusing on nearby objects like potato chip bags and measuring their vibrations. The NSA, and presumably others, can turn your cell phone's microphone on remotely, and listen to what's going on around it.

There are body odor recognition systems under development, too. On the Internet, one company is working on identifying people by their typing style. There's research into identifying people by their writing style. Both corporations and governments are harvesting tens of millions of voiceprints—yet another way to identify you in real time.

This is the future. Store clerks will know your name, address, and income level as soon as you walk through the door. Billboards will know who you are, and record how you respond to them. Grocery store shelves will know what you usually buy, and exactly how to entice you to buy more of it. Your car will know who is in it, who is driving, and what traffic laws that driver is following or ignoring. Even now, it feels a lot like science fiction.

As surveillance fades into the background, it becomes easier to ignore. And the more intrusive a surveillance system is, the more likely it is to be hidden. Many of us would refuse a drug test before being hired for an office job, but many companies perform invasive background checks on all potential employees. Likewise, being tracked by hundreds of companies on the Internet—companies you've never interacted with or even heard of—feels much less intrusive than a hundred market researchers following us around taking notes.

In a sense, we're living in a unique time in history; many of our surveillance systems are still visible to us. Identity checks are common, but they still require us to show our ID. Cameras are everywhere, but we can still see them. In the near future, because these systems will be hidden, we may unknowingly acquiesce to even more surveillance.

## AUTOMATIC SURVEILLANCE

A surprising amount of surveillance happens to us automatically, even if we do our best to opt out. It happens because we interact with others, and *they're* being monitored.

Even though I never post or friend anyone on Facebook—I have a professional page, but not a personal account—Facebook tracks me. It maintains a profile of non-Facebook users in its database. It tracks me whenever I visit a page with a Facebook "Like" button. It can probably make good guesses about who my friends are based on tagged photos, and it may well have the profile linked to other information it has purchased from various data brokers. My friends, and those sites with the Like buttons, allow Facebook to surveil me through them.

I try not to use Google search. But Google still collects a lot of information about the websites I visit, because so many of them use Google Analytics to track their visitors. Again, those sites let Google track me through them. I use various blockers in my browser so Google can't track me very well, but it's working on technologies that will circumvent my privacy practices.

I also don't use Gmail. Instead, I use a local ISP and store all of my e-mail on my computer. Even so, Google has about a third of my messages, because many of the people I correspond with use Gmail. It's not just Gmail.com addresses; Google hosts a lot of organizations' e-mail, even though those organizations keep their domain name addresses. There are other examples. Apple has a worldwide database of Wi-Fi passwords, including my home network's, from people backing up their iPhones. Many companies have my contact information because my friends and colleagues back up their address books in the cloud. If my sister publishes her genetic information, then half of mine becomes public as well.

Sometimes data we only intend to share with a few becomes surveillance data for the world. Someone might take a picture of a friend at a party and post it on Facebook so her other friends can see it. Unless she specifies otherwise, that picture is public. It's still hard to find, of course—until it's tagged by an automatic face recognition system and indexed by a search engine. Now that photo can be easily found with an image search.

I am constantly appearing on other people's surveillance cameras. In

cities like London, Chicago, Mexico City, and Beijing, the police forces have installed surveillance cameras all over the place. In other cities, like New York, the cameras are mostly privately owned. We saw the difference in two recent terrorism cases. The London subway bombers were identified by government cameras, and the Boston Marathon bombers by private cameras attached to businesses.

That data is almost certainly digital. Often it's just stored on the camera, on an endless loop that erases old data as it records new data. But increasingly, that surveillance video is available on the Internet and being saved indefinitely—and a lot of it is publicly searchable.

Unless we take steps to prevent it, being captured on camera will get even less avoidable as life recorders become more prevalent. Once enough people regularly record video of what they are seeing, you'll be in enough of their video footage that it'll no longer matter whether or not you're wearing one. It's kind of like herd immunity, but in reverse.

## UBIQUITOUS SURVEILLANCE

Philosopher Jeremy Bentham conceived of his "panopticon" in the late 1700s as a way to build cheaper prisons. His idea was a prison where every inmate could be surveilled at any time, unawares. The inmate would have no choice but to assume that he was always being watched, and would therefore conform. This idea has been used as a metaphor for mass personal data collection, both on the Internet and off.

On the Internet, surveillance is ubiquitous. All of us are being watched, all the time, and that data is being stored forever. This is what an information-age surveillance state looks like, and it's efficient beyond Bentham's wildest dreams.

# 3

# Analyzing Our Data

In 2012, the *New York Times* published a story on how corporations analyze our data for advertising advantages. The article revealed that Target Corporation could determine from a woman's buying patterns that she was pregnant, and would use that information to send the woman ads and coupons for baby-related items. The story included an anecdote about a Minneapolis man who'd complained to a Target store that had sent baby-related coupons to his teenage daughter, only to find out later that Target was right.

The general practice of amassing and saving all kinds of data is called "big data," and the science and engineering of extracting useful information from it is called "data mining." Companies like Target mine data to focus their advertising. Barack Obama mined data extensively in his 2008 and 2012 presidential campaigns for the same purpose. Auto companies mine the data from your car to design better cars; municipalities mine data from roadside sensors to understand driving conditions. Our genetic data is mined for all sorts of medical research. Companies like Facebook and Twitter mine our data for advertising purposes, and have allowed academics to mine their data for social research.

Most of these are secondary uses of the data. That is, they are not the