Andreas Weigend | Social Data Revolution | Fall 2015
School of Information | University of California at Berkeley | INFO 290A-03

Introduction.
The Social Data Revolution

#

"The only way of discovering the limits of the possible is to venture a little way past them into the impossible."[1]

Arthur C. Clarke

#

At 6:45 AM, the alarm on my mobile phone wakes me up. Before I even get out of bed, I scan through the email, text messages, and Facebook notifications that arrived overnight. I'm eager to get a jumpstart on the day, so I carry my phone into the bathroom. Those handful of steps are registered by my mobile phone's GPS, which logs the shift a few meters east in my phone's longitude and latitude. As I pour myself a cup of coffee and take it out to my deck, the phone's accelerometer tracks how quickly I pace around my apartment and its barometer registers when I'm going up or down the stairs. Because I have Google apps installed on my phone, Google has a record of all these movements.

After breakfast, it's time for me to make my way to Stanford. The electricity company has installed a "smart" meter, which registers the decrease in my electricity consumption as I turn off the lights before heading out for the day. When I go down to the garage and open the door, the meter detects the usage signature specific to the door. Thus, as I pull my car out onto the street, my electricity provider has enough data to know I'm no longer at home. When my phone switches cellular signal towers, so does my mobile phone carrier.

On the road, a camera installed on a street corner takes a photo of my license plate in case I speed through a red light. Thankfully, I'm on my best behavior today so I won't be greeted with a ticket in the mail a few days in the future. But as I go on my way, my license plate is photographed again and again. Some of those cameras belong to the local government, but some belong to private companies that are analyzing the data to identify driving patterns—a product they might sell to police on the level of individual license plates and to developers on the level of aggregated cars.When I get to Stanford, I use the EasyPark app on my phone to pay the parking fee. The money is automatically debited from my bank account, and the university parking team is notified that I'm paid up, so both the school and my bank can see that I'm on campus starting at 9:03 AM. When my phone stops moving at a car's pace, Google logs where I'm parked so that I can look it up later if I can't remember. It's also time to check my Metromile insurance app, which has been recording data about my drive from the car's on-board diagnostic system since I turned the key in the ignition. I can see in an instant that my fuel efficiency was lower today—only nineteen miles per gallon—and that I spent $2.05 on gas to get to work. I'll get my bill for Metromile's pay-as-you-go car insurance at the end of the month.

After class, I'm planning to meet up with a new friend back in San Francisco. We "virtually" met each other when we both commented on a post by a mutual friend on Facebook, and liked each other's take on the issue being debated. It turned out that we had more than thirty Facebook friends in common, yet somehow didn't know each other.

I pull out my mobile phone and put my new friend's address in Google Maps to get suggestions about the best route to his place. Google provides me with an estimated arrival

time by observing how fast traffic is moving. It predicts that I'll get to his place at 7:12 PM, and as usual the prediction is correct within a couple minutes. As it happens, my friend lives above a store that sells tobacco products as well as various paraphernalia used for smoking marijuana. The geolocation tracker on my smartphone doesn't differentiate between the apartment and the store, however, so as far as my carrier and Google are concerned, I've ended my day with a visit to the head shop—a fact vividly communicated to me by the ads Google shows when I log on for a final visit to the web before going to bed.

Welcome to the world of personalization, brought to you by your social data.

Every day, more than a billion people create and share social data like the ones that trailed behind me in the course of the typical day I have described to you. I define *social data* as information about you, such as your movements, behavior, and interests, as well as information about your relationships with other people, places, products, even ideologies.[2] Some of this data are shared knowingly and willingly, as when you are logged into Google Maps and type in your destination, others less so, often without much of thought, part and parcel of the convenience of using the internet and mobile devices. In some cases, it is clear that sharing data is a necessary condition for receiving services: Google can't show you the best route to take if you don't tell it where you are and where you want to go. But you might also happily contribute information to a social network, as when you "like" a friend's Facebook post because you want to reach out to him in some small way.

Social data can be highly accurate, pinpointing your location to within less than a meter—but lots of social data are sketchy, in the sense of being incomplete. For example, unless I log into an app that displays my smart meter's readings (for instance, in order to be sure that I really did turn off all the lights in the house as I make my way to the airport), the electricity company only knows when I am not at home, nothing more than that. It's a rough data point that may or may not be of much use to me. Similarly, as I was visiting my new friend in San Francisco, and while my X-Y coordinates were conveyed with precision, the inferences made about my activities that evening were utterly wrong. That's even sketchier, in the sense that the data seems quite exact but it's very much dependent on interpretation. Sketchy data have a tendency to be incomplete, error-prone, and, sometimes, polluted with fraud.[3]

All together—passive and active, necessary and voluntary, precise and sketchy—the amount of social data is now growing exponentially over very short periods of time. Today, the time it takes for social data to double in quantity is eighteen months. In five years, the amount of social data will have increased by about a factor of 10, or an order of magnitude, and after ten years, it will increase by about a factor of 100. In other words, the amount of data we created over the course of the entire year 2000 is now created over the course of a day. At our current growth rate, in 2020 we'll create that amount of data in less than an hour.

Data scientists are increasingly being asked to act like detectives and artists, painting iteratively clearer sketches of human behavior as they transform raw data into a usable product. Meanwhile, as data are collected and coalesced from disparate sources, we are being forced to navigate the thorny issues of data ownership and privacy.

#

My Stasi file

Over the past year, a debate had been raging over the use and possible abuse of the data that we share with companies.[4] News of a Facebook study into how moods spread from person to person caused an uproar among users who felt the site had "manipulated" their feelings.[5] After months of protest, activists won a court victory against the NSA and its dragnet data surveillance of phone records.[6] Yet, despite the vocal backlash against data collection, very

few of us are getting rid of our mobile phones, email addresses, and social media accounts. Life is just more convenient with these technologies.

I am very aware of the risks of sharing personal information. Indeed, I'm more than aware of how a powerful institution might use information against you. In fact, I find it sobering and scary.

In 1949, my father, a young man of twenty-three, took a job as a teacher in East Germany. When he arrived in his new town, he needed to find someone to share a room with. At the train station, he met a man who was also looking for a place to live. He thought he had been lucky. But a few days after they moved into their place, the roommate disappeared, without a trace. My father was baffled. As my father was making breakfast one morning not long afterwards, there came a knock at the door. When he answered, he was greeted by several men, who informed him that he had won an award for teaching. It was quite a special award, and had to be presented to him in person, and they were there to escort him to the hall where he would be honored. My father was skeptical: it seemed odd that the men were so dour, and that they were all wearing identical trench coats. But he had no choice; he was immediately ushered into the waiting car. To his utter surprise, he discovered that the car doors could not be opened from the inside. He had been arrested by the Soviet occupying forces.

Because he spoke English, my father was charged with being an American spy. He was thrown into solitary confinement in a prison run by the Soviet authorities, where he languished for six years. None of his family or friends knew where he was. He had simply disappeared from the face of the earth.

A decade after the collapse of East Germany, I requested to see what information the Ministry for State Security, also known as the Stasi, had collected about my father. I wasn't the only one curious about what the Stasi had on my family; nearly 3 million people have asked to see their files or those of relatives since the fall of the Berlin Wall.[7] Unfortunately, when the letter came from the commission in charge of sharing the Stasi's files, it seemed that everything about my dad had been destroyed.

However, tucked into the envelope with the letter I discovered a photocopy of another Stasi file: my own. I was amazed. There was a Stasi file on me? I was just a kid, studying physics, observing interactions between subatomic particles. Still, the security agents had started gathering information about me in 1979, when I was a teenager, and had last updated the file in 1987, the year after I had moved to the United States. All that was left of my file was the cover sheet; I'd never know what information the Stasi had collected, why they had collected it, or what, if anything, it had been used for.

[insert art: my Stasi file]

Back in the days of the Stasi, it was tough to get information about "citizens of interest." First, the data had to be gathered by following people, taking photographs of them, intercepting their mail, interviewing their friends, installing microphones in their homes. Then the information had to be analyzed, all by hand. There was so much to scrutinize that, at the time of the collapse of East Germany, one out of every 200 of the nation's citizens worked full-time for the secret police. But the Stasi needed even more resources than that to collect information. At least 174,000 adults, or 1 percent of the working population, were recruited as regular informers.[8] Data collectors today have it easy in comparison.

A new balance

You might think that the shock of discovering my Stasi file would have converted me into a zealot for privacy, here to educate you on how to minimize the digital traces you leave behind

for the NSA, Google, Facebook, and others. But I believe we receive real, tangible value from sharing data about ourselves and our lives—and we could receive even more. It depends on whether we can find ways to ensure that the data companies' interests are aligned with our own.

During my tenure as Chief Scientist of Amazon, I helped develop the company's data strategy and customer-centric culture, including the move from editor-written to consumer-written product reviews and the creation of predictive models for recommendations. In my experiments with data at Amazon, I saw the power of letting people communicate openly about products and of letting them find the things that they wanted to buy rather than the things companies wanted to sell to them. The tools we created for the site fundamentally changed how people make choices about purchases and became the gold standard in e-commerce.

Since leaving Amazon, I have taught courses on "The Social Data Revolution" for scores of students, from undergraduates to PhDs, at Stanford and at the University of California at Berkeley, and I continue to run experiments through the Social Data Lab, a network of scientists and thought leaders that I founded in 2011. Over the past decade, in my work with corporations ranging from Alibaba and AT&T to Walmart and United Healthcare, and at major airlines, financial services firms, and even dating websites, I have advocated to share the power of data with customers and users—regular people like you and me.

It's important to understand that "social data" isn't merely some fancy buzzword for social media. Many social media platforms have been designed either as a method for broadcasting, or with the capacity for it. In the case of Twitter, communication is almost always moving in one direction, from a celebrity or authority to the masses. Social data is far more democratic. You may disseminate information about yourself, your company, your accomplishments, and your opinions through Twitter or Facebook, but the vast array of digital traces you create and share is much deeper and broader than that. Your searches on Google, your purchases on Amazon, your calls on Skype, the minute-by-minute location of your mobile phone—all these and much more come together to produce a unique portrait of you as an individual.

And the sharing of social data doesn't stop there. You constantly share data about the strength of your relationships with family, friends, and colleagues through your communication patterns; you collaborate with friends and strangers alike in creating data, for instance, by reviewing a product or tagging a photo on Instagram. You authenticate your identity to rent a holiday home on the Airbnb website using your Facebook profile in addition to a government-issued ID. The use of social data is being embedded in homes with smart meters, in cars with navigational systems, and in workplaces with team-based productivity software. It is beginning to feature in our doctor's offices, classrooms, and government offices. As mobile phones get loaded up with more sensors and apps, and a variety of devices are adopted to track your behavior at home, in the mall, and on the job, you'll have less and less ability to exert complete control over the social data that describes your daily routine—as well as your deepest wishes.

But more important than the tracking is the decision-making power that can be afforded through the analysis of social data. Your digital traces can be examined and distilled to uncover preferences, reveal trends, and make predictions. No single person can wade through all of the data available today in an effort to make what we used to call an "informed" decision about some aspect of life. But who will have access to the tools that are necessary for taming data in service to our problems and needs? Will the preferences, trends, and predictions buried in data be available to only to a few powerful organizations, or will they be

available for anyone to use? What price will we have to pay to secure the dividends of our social data?

As we come to grips with the value of social data, I believe we must focus not just on access but also on actions. We face some decisions many times each day, others once in a lifetime. Yet, that doesn't mean the social data we create today has a short shelf life. The way we behave today may influence the choices we have fifty years from now. Few people have the ability to observe everything they do and analyze how their behavior might affect them, in the short or long term, but social data technologies will be able to provide that service. The one thing these technologies assuredly cannot do is decide what sort of future we want—as individuals or society. Preferences, trends, and predictions will allow us to better identify the possibilities and probabilities, but the final choice is ours.

As with most new technologies, it's not the "machine" that changes everything. The revolution comes as people learn to use the machine and adjust their expectations and social norms in response to it.

We are now poised at a hinge moment, when the relationship between the people who create data and the organizations that create data products and services is being defined. It's time to take a stand and truly understand how data are and might be used, so that we can realize the personal benefits and monitor the adverse consequences. Only then can we assess if our interests are aligned with the data companies.

Data of the people and by the people can be *for* the people—if we rise to the challenge. I invite you to join the revolution.

#

Notes

---

1. Clarke, Arthur C., "Hazards of Prophecy: The Failure of Imagination, in *Profiles of the Future: An Enquiry into the Limits of the Possible* (New York: Harper & Row, 1962), p. 21.
2. I have been teaching the course entitled "The Social Data Revolution" at Stanford University since 2008 and at the University of California at Berkeley since 2011, but I've been developing the concept of "social data" for longer than that. In the earliest stages, social data was merely data that people socialized, including reviews on Amazon and posts on social media platforms.
3. For those interested in a fuller discussion of sketchy data, I'd recommend the video of the "Sketchy Data" panel at which I spoke at the UC-Berkeley School of Information's 2013 DataEdge conference, http://www.catchtalk.tv/events/dataedge/videos/sketchy-data-panel-discussion-dataedge-2013.
4. **POSSIBLY ADD NOTE HERE [pointer to notable examples of abuse of data shared with companies]**
5. The word "manipulated" appeared in media including *Forbes* magazine, which ran a story under the headline "Facebook Manipulated User News Feeds to Create Emotional Contagion"—not understanding that the study was observing emotional contagion, not *creating* it. I'll spend more time discussing social network experiments in chapter 3.
6. The case before the Second Circuit Court of Appeals was *ACLU v. Clapper*. Crocker, Andrew, "EFF Case Analysis: Appeals Court Rules NSA Phone Records Dragnet Is Illegal," Electronic Frontier Foundation, May 9, 2015, https://www.eff.org/deeplinks/2015/05/eff-case-analysis-appeals-court-rules-nsa-phone-records-dragnet-illegal.

7. Between 1991 and 2011, 2.75 million people requested to see their Stasi file, and approximately 60,000 people put in a new request each year. Pidd, Helen, "Germans Piece Together Millions of Lives Spied on by Stasi," *Guardian,* March 13, 2011, http://www.theguardian.com/world/2011/mar/13/east-germany-stasi-files-zirndorf.
8. Regular Stasi informers were called *inoffizielle Mitarbeiter,* or "unofficial employees." Koehler, John O., *Stasi: The Untold Story of the East German Secret Police* (Boulder, CO: Westview Press, 1999), p. 8. **[TO DOUBLE CHECK THESE NUMBERS]**

Chapter 1.
Becoming Data Literate:

#

"In the 18th century a person able to read aloud familiar passages from the Bible or a catechism would be counted as literate; today someone who could read no more than that would be classified as functionally illiterate—unable to read materials considered essential for economic survival."[1]

George Miller

#

"Data for the people" is not some empty slogan. Every day we are presented with data products and services in form of rankings and recommendations based on social data. The traditional "Mad Men" of marketing have been replaced by data scientists running algorithms on the multitudinous digital traces that more than a billion people trail behind them every day. The balance of power is shifting between sellers and buyers, bankers and borrowers, employers and employees, doctors and patients, and teachers and students. We're living in the midst of an exponentially expanding dataset, but what's more important is the change in mindset taking hold. To be full participants in the social data revolution, we must shed the role of the passive "consumer," who takes in whatever is placed before them, and embrace a new mindset, that of a co-creator of the social data that serves as the raw material for a range of increasingly personalized products and services. This is how data of the people and by the people becomes data for the people.

The demand to ensure data for the people also couldn't be more important. Data is the most important raw material of the twenty-first century: data is the new oil.[2] Although this analogy is bandied about quite a bit, it's illuminating in several ways. For more than a century, our economy and society have been largely shaped by the discovery of oil and the development of techniques for extracting, storing, and refining it to create products that everybody on the planet uses. Today, the capacity to transform raw data into products and services is transforming our lives in ways that will rival the machine age.

Crude oil cannot be used in its raw form. It has to be refined into gasoline, plastics, and many other chemical products. Refined oil has also fueled the machines of the industrial age, and played a role in the manufacture of most of the modern economy's physical products. Similarly, raw data is pretty useless on its own. The value of data is created by refineries that aggregate, analyze, compare, filter, and distribute new data products and services. Instead of powering the apparatus of an industrial revolution, refined data powers the apparatus of a social data revolution.

Happily, data is very different to oil in fundamental ways. The amount of oil in the world is finite, and the less of this resource that remains, the more the cost of exploiting it goes up. In contrast, the amount of data created each year is increasing exponentially, while the cost of the technology required for communication and processing is decreasing at a similar same pace. At the end of 2013, one in five people on the planet owned a smartphone,[3] including 56 percent of all adults in the United States.[4] The average American spends about two hours each day on a mobile phone.[5] It's estimated we touch our phones between 200 and 300 times a day—for most of us, that's more often than we touch our partners in a month.[6] And each time we do so, we create data. That's in large part why the doubling rate for data is eighteen months. In contrast to oil, we will never run out of data.

Our use of energy is also constrained by the fact that oil is scarce and material; our use of data has to take into account that data are now abundant and digital. Only one entity at a time can have the

right to use a particular stock of crude oil or a product refined from it, while many can simultaneously access the same pool of data and create many different products from it. We've established laws and social norms based on the idea that data is in short supply. For instance, in the absence of data, we created insurance—a way to protect ourselves against the possibility of terrible events occurring in our lives. Because it was impossible to know a specific person's chances of being burgled or contracting diabetes, insurers grouped people together, pooling the risk. As more and more data are produced, we may soon have the ability to make such individual risk assessments. We can pretend we don't have the data or we can acknowledge that we do have it and think about how this should change the way we go about our lives. What sort of world do we want to create with this fantastic new resource?

New technologies can be empowering, but only when we have the tools to utilize them. Before Gutenberg invented the printing press, books were in scarce supply and sending written communication was expensive. The majority of the population gained no benefit from spending long hours learning how to read. When George Miller, a professor of psychology at Princeton, wrote in 1988 about modern standards of literacy, he was worried too many students were leaving school without the level of advanced reading, mathematical, and scientific literacy necessary to get a job in an economy dominated by the "knowledge industry."[7] Today, there is another, just as pressing need for data literacy, skills like being able to recognize how data is being refined, learning what parameters can and cannot be changed, interpreting errors and probabilities, and understanding the possible consequences of sharing our social data.

Such literacy is necessary for a world in which most decisions will be based on the analysis of data by machines. There are huge benefits to be gained from social data if you're data literate, and there are significant risks if you're not.

The data-refining process

One of the first great data refineries, Amazon, is in the retail sector. This isn't surprising. To succeed as a retailer, you have to know which products to stock for your likely customers, which entails keeping track of data about your inventory and your customers' buying behavior.

Two hundred years ago, the only data a shopkeeper had to work with was the inventory on the shelf and the money in the till at the end of the day. This data was recorded in a paper ledger with a fountain pen. For similar products being sold at similar prices, the customer chose what to buy based on the credibility of the product's promises, the attractiveness of the product's packaging, and the say-so of his neighbors, family, and friends. A little over one hundred years ago, a few companies—most notably, Montgomery Ward and the Sears & Roebuck Company—delighted customers in small towns across America by publishing mail-order catalogs with more than ten thousand products listed in them. These innovative companies knew which items a particular customer ordered and where he wanted them shipped, typically the local general store-cum-post office, and they could see which products sold better in specific regions and which had mass appeal.[8] Fifty years ago, the retail landscape changed again. Mail-order companies could more easily group American customers by using the newly introduced Zip code system for addresses.[9] Over the next couple decades, companies tried to glean demographic information about these clusters, sometimes turning to specialist data processors like Acxiom, which provided no services to individuals in exchange for their data. That was as personalized as data got, pre-web.[10]

The retail industry's reliance on inventory management and customer relationship marketing systems meant it was natural to consider how to take advantage of data at scale when it became technologically feasible. From the advent of e-commerce in the 1990s, retailers had the capacity to track every click and purchase, to capture every abandoned shopping cart. Not all took advantage of this technological capacity, but Amazon did. Amazon is sometimes called the "Everything Store," but

more profoundly, it was the first "Save Everything Store" in the arena of data. Because of the new scale of digital inventory, it's not possible to browse page by page through the entire catalog of the Everything Store. Amazon can't show you every product it has on its warehouse shelves—or, more to the point, every product it could get from a supplier on your behalf within a few days' time. If you don't tell Amazon what you're looking for, there's no way the company can tell to you the products it can deliver that might be a match. You have to give data to gain access to the refinery's essential service: an algorithm to rank items that fit your search. You no longer have the option of keeping your interests to yourself until the moment you reach the checkout.

Saving all the shopping data about customers is not the stuff of revolutions, in and of itself. What set Amazon apart was its commitment to refining data in ways that help customers decide what to buy based on their own interests and preferences.[11] Many companies—Amazon included—were tempted by the ability to know a customer inside and out. But a healthy balance must be struck when it comes to personalization. If a retailer opts to offer everyone basically the same offers, customers may leave the store, since the store has learned so little about their preferences. If a retailer takes the opposite approach and personalizes every product recommendation, it risks scaring off customers when they realize how much the store knows about them. Many retailers wised up and mixed personalized recommendations based on customer searches with unrelated offers. A woman who had abruptly started buying vitamins and unscented hand lotions would still see ads for baby strollers, to use *New York Times* reporter Charles Duhigg's famous example, but to make the ads feel less intrusive, they'd also see some promotions for products unrelated to pregnancy.[12] The mix of generic and targeted marketing would, it was believed, deflect the suspicions of the expecting mother.

In the early days of e-commerce, most companies applied methods invented by Acxiom and other companies for targeting customers in a data-scarce world. They sliced and diced their customers and their shopping data and happily assigned them to a variety of well-established labels: "Suburban Soccer Mom," "Blue Blood Estates," "Shotguns and Pickups," and several dozen more—some even worse in their stereotyping.[13] When I arrived at Amazon, there was no question we could move beyond such simplistic thinking, given that we knew about every interaction a customer had on the site. My team and I identified five hundred personal attributes that might be associated with a specific user. For instance, did the distance between the shipping address and the nearest bookstore or mall make a difference to how often the customer shopped at Amazon or how much he spent? Did a customer's choice of MasterCard over an American Express card predict anything about her future buying patterns? Was a customer who shopped in two or more categories worth more in sales each year to Amazon than someone who only ever bought books? Does a customer order different things during the day versus in the evening? Our goal was to analyze the data in order to make essential business decisions, such as whether to spend a dollar on marketing or a dollar on price reductions.

No matter how we sliced and diced the data, we discovered that the best predictor of a product's likelihood of getting bought was the product's relationship to other products. The similarity between items could be computed in various ways, from a comparison of product specification to an analysis of the overlap in the words that appear in product descriptions. However, the most important similarity for making purchasing recommendations was how often two products shared a "shopping cart." If there was a pattern of customers buying two items together, or buying one item and then later buying the other, those products were mapped as being complementary in customers' eyes. If there was a pattern of customers clicking on two items in the same shopping session, but only buying one of them, those products were mapped as substitutes—once one was bought, there was a very low probability that the customer would be interested in the other. In this way, Amazon's recommendations were built on aggregated clicks and purchasing data, not aggregated customer profiling. The recommendation system doesn't note who bought the two products, or how similar a customer might be to you, based on where you live, what you do for a living, or how you spend your

free time. It's the relationship between items that is most relevant. The effectiveness of item-by-item filtering meant it was better to spend a dollar on data analysis than a dollar on marketing. So instead of buying lists of "Suburban Soccer Moms" and "Pickups and Shotguns" from marketing outfits, Amazon built a platform through which third parties could sell products on the site, even offering space in its warehouse for these companies' products. Amazon could then analyze the relationships between the products sold by these other companies, and add this into the mix of its rankings and recommendations.

Amazon's other big innovation was giving customers the ability to publish product reviews on the website. The experiment turned traditional marketing on its head once again. Customers were eager to share their experiences with other customers, and companies were no longer able to control the release of information about their products. Customers considered other customers to be more trustworthy than the manufacturer's description, which was nothing more than advertising. More surprising, the aggregated opinions of thousands of strangers held far more sway than an expert editor's endorsement. If lots of people gave an item low marks, it didn't matter if a member of staff loved it. Allowing customers to review products also provided far more coverage of everything for sale in the Everything Store, and it gave customers a chance to scan a range of opinions, not simply one person's. As a result, Amazon got rid of its editorial staff and dedicated its resources to developing algorithms for customer reviews so that those considered most useful to other customers were featured at the top of a product's page. A dollar spent on customer-review algorithms created a better shopping experience for customers than a dollar spent on staff-generated reviews.

Amazon built a data refinery and looked for ways to encourage others to create and share data that could be used to improve its recommendations and increase its sales. As a data refinery, Amazon has changed how a billion people shop. Today, more than 50 percent of retail purchases in the United States start with a search at Amazon, regardless of where the customer ends up buying the product.[14]

We've moved far beyond the days where we disclose our interests and intentions through searches and clicks alone. And just as we don't need to understand every intricacy of an internal combustion engine to drive a car, we don't need to understand every intricacy of Amazon's recommendation algorithms to find a product that matches our interests and needs. It's more important that we understand the basic mechanisms for how the machine works and establish rules for safely operating it. As we create and share data from more sources and sensors, we can either stand by and let others decide the terms of use—scrolling through twenty-plus pages after which we blithely click the accept button—or we can choose to help establish new norms of interaction. We can imagine the social data refineries are mysterious "black boxes," or we can become data literate and demand meaningful ways to influence the refineries' inputs and outputs to our benefit.[15]

Learning how the data refineries work is a crucial first step to becoming data literate, because this explains why "we the users" won't be able to exert power the old-fashioned way: by attaching a value to our data and putting it up for sale.

#

What's your data worth?

We already rely on social data in making many everyday decisions, as when we decide which product to buy on Amazon. With the exponential growth in social data, we will increasingly rely and depend on data refineries to help us make the biggest decisions in life—including who we pick as a romantic partner, where and how we work, what food we eat and medications we take, how and what we study, even which politicians we vote for.

Given the amount of social data being created, we have to rely on refineries to help us sort through it. Meaning only emerges by comparing our data to that of others. In many cases, the amount of social data available to the refineries means that we can now hope to get answers to many questions

that we'd never before expected could be answered—and sometimes to questions we'd never thought to ask.

Algorithms find patterns that humans without computers cannot see. The value we reap from sharing data comes in the form of better decision-making when we're negotiating deals, buying products and services, applying for a loan, looking for a job, obtaining healthcare and education for yourself and your family, and improving your community's safety and public services.

So far, the debate over the use of people's data has focused on the data miners, that is, how, when, and why companies and governments collect the deluge of digital information that we create day in and day out in an attempt to extract valuable nuggets about us. Some argue that too much data is being collected, and that the best option for individuals is to share less about themselves—or to demand payment for the data they create and share. I think we should demand something far more valuable than a little financial hand-out in return for our raw data: we should be asking for a seat at the controls of the refineries.

First, let's consider the difference in value between raw and refined data. If I enter "Andreas Weigend" into the Google search box, Google reports that there are "about 122,000" pages with the words "Andreas" and "Weigend" on them. There is no way for anyone to sift through all those pages manually—at the rate of one page every five second, a phenomenally fast click-and-review rate—it would take a full week for you to look at them. I know I couldn't be bothered to devote that much time to most topics, and even if I was especially interested in this topic, I can't imagine spending a week on the task without taking a break or two. So that leaves us reliant on the order in which Google ranks the pages for us. Google could list the most recent mention first. That might be ideal if I'm interested in the most recent news about myself, but not if I'm looking for the video of a class I taught a few years ago. Another option would be to count the number of times my name appears on a page and list the pages in that order, with the most mentions being the most relevant. That might be somewhat helpful if I'm sorting articles and want to find the one in which I'm quoted the most. But think about how this ranking would help if instead of looking up my name, I were searching for a "cheap iPad"—about 356,000 results. The click-bait specialists would load up pages with popular search terms (as many of them still do) and I'd be stuck wading through result after result trying to find a link to a page that was actually useful.

To improve its search results, Google uses multiple data sources to assess the usefulness of pages. In the early days of the site, Google's engineers ranked the relevance of pages based on the number of other pages that pointed to them. These incoming links served as a shortcut for detecting which pages were trusted by others. As people learned about the importance of incoming links in determining a page's rank in search results, the field of "search engine optimization"—including disreputable link farms—was born. Google's algorithms had to adapt, learning how to detect which incoming links were created by honestly interested individuals and which were created on behalf of a page's owner. Now, Google has nearly two decades of information on which pages people choose to visit in response to a search query—and how long they stay on them before navigating back to the list on their web browser. Pages that are less relevant are downgraded in Google's page rankings when many visitors click on the page but, after a cursory glance, swiftly defect in search of something better—called a "short" click.

We all use Google so frequently to search for information on the internet that "to google" is a verb, but you may not be fully aware of the amount of data being created and shared. For instance, how many photos do you think get posted on Facebook each day? Resist the temptation to search for the answer on the internet. Instead, think in orders of magnitude, meaning, in factors of ten. Assume that the number of people on Facebook is 1 billion, since it's certainly more than 100 million and less than 10 billion.[16] Assume that each person on Facebook posts 1 photo per day, whatever the exact number might be, it is probably between 1 photo every ten days and 10 photos a day. This gives us an

order of magnitude of 1 billion for the number of photos posted on Facebook per day. Thinking in terms of orders of magnitude is part of being data literate.

The vast amount of data being created means your stream of raw data isn't particularly valuable to the data refineries. While your personal data might hold a lot of sentimental value for you, the refineries cannot extract much decision-making value from most of it. That adorable photo of your dog that you posted on Facebook is probably of interest to no more than 0.0001 percent of the site's users (i.e., 1,000 people). It is only by aggregating and analyzing data from millions of other people that meaningful correlations and patterns can be found that are of use to a billion people. Subtract one person's data and the refineries develop the same conclusions from everything that's left. The individual misses out but the refineries miss essentially nothing.

[insert art: The deluge of data being actively created online every 60 seconds.]

This is partly—though not entirely—why I think that it's wrongheaded to ask to be paid for your data. With so much data being created, each bit is worth very little—and less and less each day. Perhaps you'll garner the king's ransom of eight dollars a month, the amount the start-up Datacoup has been willing to pay for access to a person's social media accounts, including Facebook. For the most part, Datacoup and similar data brokerage services are willing to pay that handsome rate—for the time being—because you are sharing not just your own data, but the data of everyone in your social network. Because few people are willing to sign up, Datacoup has to offer a lot to attract subscribers. As the number of subscribers increases, the compensation will plummet. Soon enough, the going market rate will be a fraction of a fraction of a cent, if that.

But data brokers aren't the only ones arguing that you should be compensated for the use of your data. Microsoft Research philosopher Jaron Lanier has become a cheerleader for data compensation in the form of "micro-payments," a stance he has presented with great passion since the publication of his book *Who Owns the Future?* in 2013. Lanier seems to understand that raw data is worth very little, yet he still thinks we should charge a micro price whenever content we have posted on a website is read or used in some way.[17] One of his pet examples is the language translation service available from Google. Why should Google get extra advertising revenue, he asks, when all the people who help to improve the company's algorithms by suggesting and correcting translations receive nothing? With each suggestion and correction, Google's model for translating text does improve, even when a new contribution duplicates the work of earlier contributors. The model learns from these duplicates to put more weight on that suggestion, because multiple parties have made it, which means it has more credence. There's not one "best" answer, so to speak, and every contribution to the service is worth marginally less and less. And of course, this sets aside the sunk costs of populating the service with translations for people to respond to—which Google handled by importing the French/English translations of the official Canadian parliamentary record Hansard and, later, United Nations documents published in multiple languages.[18] In addition, Lanier's contributors do receive something for their efforts. There's a high probability that they, too, benefit by using Google's service tool for translating texts. They are paid not in money but in refined data products and services.

Then consider some of the data created on Facebook. If you post a photo of your dog, you clearly created that data. But what if you post a photo of a group of friends at a birthday party. You took the photo and posted it, but the commercial value of the post for Facebook is based on the traffic that it inspires and the refined data about relationships and interests that are embedded in people's interactions with it. Should you get 100 percent of the payments attributed to the sharing of that data? Or should you split it with everyone tagged in the photo? How about with everyone who adds a comment, like, or tag on it, which means the photo becomes part of their activity for friends to see? What about the clicks you create as you visit the results served up to a Google search or the posts that appear in your Facebook News Feed? These data are far more numerous—and far more helpful to

improving these refineries' services and ensuring their revenues. Lanier doesn't talk about this data—presumably because he doesn't think of it as "creative" content worth being paid for. But these digital traces are the bulk of raw data being collected and refined for profit. The services and products that we depend on, day in and day out, are built from these marginal contributions.

If refineries were forced to make a reckoning of the value of all your clicks and searches, all your likes and tags, relative to everyone else who touches that data and adds it, you can bet they'd start asking users to pay for access to search results, recommendations, and rankings. Developing algorithms costs money, and doing this analysis would require developing an algorithm specifically for the purpose of assigning attribution and value to every bit of data—including how the value of data changes over time.

It's not merely the cost of solving the problem of attribution that makes Lanier's proposal for micropayments a non-starter. Instead it is the fact that most people aren't willing to switch off their digital lives. You can decide you no longer want to provide free data to the data refineries, but then you'll have to forego the free products and services they provide. Facebook's profits for all of 2014 totaled $2.9 billion—less than three bucks per user on average. Is having a free communication platform worth more than three dollars a year to you? If so, you're already getting paid for your data.

Data refineries do not reduce you to a bunch of data to be bought and sold—at least, not necessarily. If I want you to take one lesson away from this book, it's that social help *you*—not merely some megacorporation developing a targeted advertising campaign—improve your decisions. I believe that, as much you are the data you create, you are the decisions you make. The value of data is in the decisions you can make from them.

Exploitation versus exploration

The creation of refined data products involves a trade-off between exploitation and exploration. Hearing those words, you might be imagining a dark and seedy street corner, but instead I want to transport you to the blaring neon lights of the Vegas Strip and a bank of slot machines. In the field of machine learning—where computer software responds dynamically to incoming data—the "one-armed bandit" is a sort of king, an exemplar of the dilemma of whether you're better off exploring new options or whether you stick with the best option you have seen so far.[19] Say you just walked into a casino and heard someone seemingly make a fortune at a particular slot machine. What would you do? Would you spend the rest of your evening at that machine, exploiting your observation that it has paid out more than other machines since you arrived, or would you explore other machines, looking for data that might identify a potentially better chance of a jackpot? Ideally, the computer theorists say, you'd spend some time observing the slot machines and try to detect a pattern. Of course, casinos set up the game so that gamblers lose on average (they're in the business of making money, after all). And, of course, you have limited information about each machine, since none of us want to spend every hour of every day in a slot-machine hall. The challenge is deciding if the data you have about collected about each machine is meaningful. Have you observed a machine enough to ascertain its probability of a payoff compared to the others in the hall? You could explore machine after machine in the hope that you'll happen to sit down in front of one just as it hits its scheduled moment for a big payoff. What a disappointment it would be to hear the jangle of hundreds of coins come from the machine that you'd just left.

[insert art: The one-armed bandit problem]

The one-armed bandit problem may seem to have little to do with the output of data refineries, but the balance between exploitation and exploration is a key issue in how recommendations and rankings are presented to users and how users pick among them. When search engines like Google present a list of websites in response to a query, they don't only present you with pages that are the highest probability match; they seed the results with options that allow you to

discover pages with a range of relevance to your search term. Occasionally, it's very clear that you're searching for information about a particular thing, for instance, if you type *Panthera onca* into the query box. But if you ask for pages about "jaguar," it's highly unlikely that you'll only see websites about the cat (or, for that matter, about the car or the old Mac operating system) on your first page of results.[20] The search engine's algorithms then create clusters of "jaguar" meanings based on the words on the page, the links between pages, and people's click decisions, in order to provide enough range of exploration among the clusters to ensure you find what you're looking for, no matter which cluster interests you.

Similarly, an offshoot of the one-armed bandit is called the "optimal stopping" theory or "fussy suitor" problem, which was first described by Martin Gardner in his "Mathematical Games" column for *Scientific American.* Gardner's version involved slips of paper with numbers written on them, anything from "small fractions of 1 to a number the size of a 'googol' (1 followed by a hundred 0s)."[21] You mix up the slips of paper and turn them over, one by one, until you come to a slip that you think might be the largest number in the stack. Over time, the slips in the thought experiment were transformed into suitors going out on dates. You go on a date with a person and have to decide: do you keep dating around or do you stop because *she's the one* (of those you've met so far)? You're facing a real-world, high-stakes choice between exploration and exploitation.

Obviously, users of dating websites are constantly negotiating the fussy suitor problem. Early dating sites were coded to let users specify their preferences for people based on weight or height or distance ranges and ranked their dating prospects accordingly. You decide to click on a photo of a possible dating prospect, who we'll call Sam. The site doesn't know what it is that inspired you to click on Sam's picture. Is it the fa ct that Sam was the first person on your list? Is it because Sam has dark hair or wears glasses? Is it because there's an ocean view in the background, and you're interested in someone who lives by the beach, or in someone who likes to take beach vacations?

#

Turning data into decisions

"Data! Data! Data!" he cried impatiently. "I can't make bricks without clay."[22]

Sherlock Holmes in "The Adventure of the Copper Beeches"

#

When I was working as a post-doctoral fellow at Xerox PARC—Xerox's research and development center in Palo Alto—in the early 1990s, we used a supercomputer to analyze road traffic patterns. Our goal was to be able to predict when the flow of traffic would shift from smooth to stop-and-go. Being physicists, we studied traffic like a fluid, trying to identify the conditions that caused a "laminar-turbulent" transition—the shift from a smooth to a choppy flow. By today's standards, we didn't have much data, so we had to make a lot of assumptions and built those assumptions into of our traffic simulation models.

Today, that complicated prediction problem has been solved through "simple," real-time observation. One company in the field, a Microsoft spin-off called Inrix, analyzes the geolocation data of more than 100 million individuals each day to identify when cars are moving—and more important, when they're not—to understand trends in the movements of people and products.[23] For its analyses, Inrix collects data from mobile carriers on when those 100 million phones switch cellular connection between phone service towers. You may be sharing your location data without even knowing it. Inrix sells this refined data to Garmin, MapQuest, Ford, BMW, and other companies that want to provide mapping and route-planning services to drivers. Inrix also consults with government officials on urban planning issues, including where to build new bridges, install new traffic lights, and situate new public hospitals and other social services.

Here's another example of refined data influencing real-world decisions. A couple years ago, I enrolled in a service called Google Now.[24] It scanned my gmail for information in my e-tickets and

updated me on the status of my flights, often before the airline did. Simple enough. But the sophisticated data analysis of Google Now still managed to surprise me. One dawn, as I was packing up my bags in Freiburg, I was informed by the app that I needed to leave for the airport right away. According to my schedule, my flight wasn't supposed to leave for hours. I was incredulous. Airlines don't typically shift a scheduled flight's departure time forward. Nothing made sense. Still, trusting Google Now more than my calendar, I decided to get a move on; maybe Google had identified a huge traffic jam on the way to the airport. When I arrived, I realized I had the wrong time in my calendar, but Google had ignored that data and instead reminded me about the departure time based on the e-ticket in my gmail.[25]

Inrix's traffic updates and my Google Now story demonstrate how data from many sources and aggregated by a refinery can be far more empowering for decision-making than any one individual's raw data all alone. Anticipatory systems based on social data will advise us—and potentially nudge us—on a trajectory of decisions about our relationships, our jobs, our health, and many other things. They also highlight the importance of understanding the role of interpretation in data refining. The data refineries provide three levels of output: description, prediction, and prescription.

Descriptive statistics summarizes data, calling out salient features, for example, by grouping similar data into clusters. Such descriptions of data can provide a context for decision-making by setting a benchmark against which to you can measure your particular circumstances. If you want to know the current location of traffic jams in Manhattan, you can look at how quickly cars are moving on the streets by tracking the geolocation of mobile phones and identify bottlenecks. But even this relatively simple exercise involves some interpretation. You might see huge amounts of data indicating stationary cars near the MetLife Building. But is that because it's near Grand Central Terminal and several busy taxi stands, where you've got a number of taxi drivers waiting for passengers, plus a number of passengers joining them, so quite a few mobile phones over-reporting "stalled" traffic there? If you want to know if your store is doing well this holiday season, you can tally up your sales, but you need to compare them against something else. If you compare them against your sales at the same time last year, this would not take into account changes in the local economy. Instead you can compare your sales to similar stores in the area.

However, for descriptive statistics to be useful, you have to detect what is signal and what is noise. "Signal" and "noise" are statistics jargon for data that is relevant—a signal—and data that are random and thus irrelevant—noise. Social data are complicated because what is signal and what is noise may vary from user to user and from context to context. When a friend on Facebook tags you in a photo in which you don't appear, is that signal or noise? It depends. If they've tagged you in the photo accidentally, mis-clicking on your name instead of Andrew's right below you in the list, that's noise—the social data equivalent of static interfering with your radio's reception. If they've tagged you in the photo because they are certain it will be of interest to you, or because you were at an event but didn't make it into the snapshot's frame, there's some relevant information being conveyed. It might be annoying, but it is not noise.

The second way to think about data refining is predictive analytics, which involves taking data and generalizing to future cases, including likely behavior and events. For instance, city planners have used Inrix archives of traffic data collected at one-minute intervals to assess the impact of an event—whether it's a highway accident, a construction project, or a big concert—so that better contingency plans can be formulated for the future. Hedge funds have used Inrix data on the amount of traffic to shopping malls and big-box stores to decide whether to buy and sell stock in retailers long before the release of quarterly sales figures. Analysis of geolocation data collected on Black Friday 2012 correctly forecast a major bump in sales for the entire holiday season.

A data scientist is more likely to build an effective predictive model if he can accurately describe the data going into it. When I was at Amazon, we looked at the time lag between when a person viewed an item and when she bought it. Some of the data points were obviously erroneous because the time difference was negative. Because of Amazon's design, it was impossible for someone to first view a product after she'd purchased it. We didn't know why this data was wrong, but we threw it out. We were left with a bunch of data that indicated that quite a few customers waited eight hours before buying an item. How strange. It was only when we realized that some of the computers at Amazon were set on U.S. Pacific time and others were set on Greenwich Mean Time that we realized the eight-hour lag was an artifact of different international time zones being applied to different clicks. No meaningful analysis could be made until we'd adequately understood the data we had in hand.

All datasets have mistakes; there's no getting around that. In the days of "small data," statisticians got to know each and every data point they collected, weeding out any and all errors so that these wouldn't perturb their analysis. Every single data point was incredibly precious, elevated to a station of great importance, worthy of being acted upon. Because there was a small amount of data, it was possible to check everything. And it's a good thing we did, because often we were making decisions for a community or a state based on data collected from a handful or two of people. An input error of one order of magnitude—a decimal point being off by one digit—in a day's unemployment claims for a state would have a huge impact on the monthly unemployment figures, which in turn would dramatically affect the government's economic policies.

It is reasonable to assume that the rate of errors in data does not depend on the amount of data collected. If you now have access to 100 times more data compared to last year, you'll be working with 100 times more incorrect data points. Yet, for some reason, we still rely on those old laws and social norms that assume that every data point that isn't a "mistake" is significant. We've seen a single Tweet kill a person's career, even though that Tweet was an outlier among the person's Twitter activity. Such judgments are akin to sending a person caught possessing a small amount of an illegal drug to jail on her first offense, rather than waiting to see if he's a "persistent offender." We no longer have to rely on single data points when making decisions. In fact, we shouldn't, because we are bound to surface some spurious correlations because of the amount of data at our disposal. If 1,000 variables can be identified, those variables require the analysis of 1 million pairwise correlations. With the help of refineries, we can access a rich data history. Patterns—and not just single data points or one-to-one correlations—are far more significant.

Indeed, given the amount of data today, we no longer have the ability to hunt down individual errors; instead, we must create machine-based models that take mistakes into account. First, the amount of data means that individual outliers have much less influence when looking at data in aggregate to find trends, set benchmarks, and develop other refined data products and services. Second, because people are constantly creating new data in response to refined data, the algorithms can learn to identify what might be an input error. Google asks if you meant to search for "Andreas Weigend" when you type "Andreas Weigand" by mistake because a percentage of previous users retyped their queries after seeing the results. We've become accustomed to letting the refineries point out mistakes like this and accepting them as part of living with lots of data.

For search engines and other data refineries, analyzing user feedback is key. I don't mean that you're asked to sit down and fill out a customer survey or attend a focus group. Fostering an ongoing "dialogue" with users is the only way to improve the products and services of a data refinery. Your dialogue with a data refinery is expressed through your interaction with the options the refinery presents to you. Each choice you make feeds more data into the refinery's algorithms, which adjust the ranking of the options, and the options themselves, in response. In some cases, as with Google,

you learn to change what words you search for to get results closer to the ones you were expecting. This process of back and forth on the part of users is called "query refinement."

Because many data refineries are in the business of predicting your likely purchasing decisions, you have to be on the look-out for the possibility that the rankings and recommendations have been created in ways that aren't aligned with your interests. One of the earliest big data initiatives was the Sabre Global Distribution System for flight reservations. Sabre was originally launched in 1960 as a project of American Airlines, which invested a huge amount of resources into developing an automated reservation-making system. In 1976, Sabre was installed in travel agent offices and other airlines' flights were added into the mix.[26] By looking at the pattern of flights that were getting booked, American realized that travel agents were most likely to choose the first flight on the screen, and that few if any flights appearing after the first page got selected.[27] American tinkered with the algorithm that ranked the options to give preference to its own flights. The customers didn't know that the top options presented to them might be biased. And the travel agents' interests were in getting a commission, so it made little difference to them if the flight that was booked was more expensive than an option that was buried down in the results. However, two rival airlines, New York Air and Continental, discovered that their flights were getting buried in the results, even when they added new routes and discounted fares—two of the variables that should have helped to raise their flights in the system's rankings.[28] A congressional investigation ensued.[29] The bias finally stopped when the government prohibited it in 1984.[30]

Such manipulation is much harder to pull off when the users of a data refinery's products and services aren't middlemen but the end customers, where there is much more attention to whether the prediction matches the customer's preferences. I worked with Agoda, a Bangkok-based hotel reservation site, on developing a recommendation system. At first, it might seem that the best option for the company's bottom line would be to list hotel options based on the amount of profit that the site makes. If a hotel is willing to give a bigger commission to a travel site, why not list that hotel's rooms as the first choice? In the short term, the customer got a room and Agoda got a quick buck. However, the site was better served by ranking a hotel based on the probability that it matched the customer's preferences. Some customers who got results ranked by Agoda's profit margin on the room went ahead and booked it, but might later regret making the booking. Other customers looked at the top options and assumed Agoda didn't have inventory of the sort of hotel they liked and decided to book a room through a competitor. Agoda was more profitable in the long run when its interests were aligned with its customers'.[31]

The final level of working with data is prescriptive analytics, taking the data you have and determining how to change conditions to reach a desired outcome. A classic example is the data analysis involved in NASA's Moon landing.[32] To get Neil Armstrong and the American flag on the Moon, the engineers at NASA had to collect and analyze a continuous stream of data about the landing module's position in space. The engineers needed to do more than summarize that data (description), and they needed to do more than forecast when and where the module would hit the lunar surface (prediction); they needed to identify which action to take next in response to every new data point about the module's changing situation in order to improve their chances of actually getting a man on the Moon. After each firing of one of the module's jet thrusters, they'd receive data about the exact effect the force had on the module's course. They'd then predict how and when and for how long the thruster needed to be fired again to reach their goal.

Good data scientists use all the three levels of analysis in their repertoire. Description provides opportunities for recognizing "natural experiments," situations in which one variable has been changed without design and effects can be observed, such as when a bug gets incorporated in the roll-out of software. The developers for Amazon's French website somehow forgot to add the shipping cost to customers' shopping carts. Unexpectedly, we were able to describe the effect of free

shipping by comparing the day's purchases to a day before the mistake had been spotted. Of course, prediction is at the heart of the scientific method: a scientist creates a model that makes a prediction, carries out an experiment, and measures whether the outcome corresponds to the prediction. If it doesn't, the scientist changes the model and begin the process again. What most interests me in the realm of social data are experiments that involve prescription, allowing users to see how setting parameters changes a decision. A refinery can use data to identify a traffic jam happening right now and alert drivers, providing a longer predicted travel time for the slow route and suggesting alternates that appear to be moving faster. If most drivers select the same alternate route, a new traffic jam might build up there. A refinery could suggest a variety of options and inform drivers what percentage of other drivers in the area have already selected a particular route, so that they can decide if they might be better off taking a different option. Or the refinery could use that data to try to optimize traffic flow by anticipating where a traffic jam will likely occur in the next few minutes unless the timing of traffic lights—and then change them.

#

Give to get

In the field of data science, a common maxim is to let computers do what computers are good at and let humans do what humans are good at. Humans are pretty good at finding patterns—sometimes too good, so that we sometimes see patterns whether or not they exist. But more important, humans are better than computers at formulating questions, and human judgment is crucial in evaluating the trade-offs intrinsic to decision-making. For this reason, we must insist that humans retain the ability to weigh our options and choose our own actions. This isn't a power we should merrily hand over to the algorithms, one less thing to worry about as we go about our days. Basic data literacy, including the ability to estimate orders of magnitude and the effect of errors, allows us to understand the trade-offs of taking one action over another in light of the machines' analysis. I do not want us to be *driven* by data, I want us to be *empowered* by it.

The laws in place that protect individuals in many countries from discrimination in the workplace or health care may not exist tomorrow—and in some countries, they do not exist even today. Imagine that you opt to share that you're worried about having high cholesterol with a health website or app in order to get advice about diet and exercise regimens. Could your worries be exercised against you in some way? What if the laws made it permissible to charge higher rates for medical care if you didn't change your behavior, continuing to eat fried foods and slouching on the couch even after you'd been presented with a menu of health risks and recommended actions? Or what if a potential employer used a service that crawls the web for information about you, and then, based on what he learned, decides that your lifestyle isn't a good match for a job and he won't consider your application?

In order to assess the rewards and risks of sharing data, we need to be able to hold data refineries to standards of transparency and agency:

- *Transparency*: Is the refinery observing your behavior and your relationships from the "dark" side of a one-way mirror, like a police detective waiting for you to slip up and talk about a crime with your accomplice? Or does the refinery give you a view onto its data detection work, providing data that you can check to see if the outputs seem possible and probable, based on your inputs?
- *Agency*: Are you able to freely act on the refinery's outputs? How much raw data about yourself do you have to share, and how valuable is the refined data you get in return for it? How easy is it for you to identify the refinery's "default" settings? Can you play with some of the parameters and create different scenarios that result in different recommendations?

In the simplest terms, the two-way window of transparency requires that both sides can see the other side. This provides a more symmetrical balance of power between institutions and individuals. It's better for you as a customer and as a citizen if information does not only flow one way.

Transparency for the user is an essential ingredient in the design of the shopping experience at Amazon. When you are about to buy an item, should Amazon remind you that you already bought it and potentially lose a sale because you forgot? If you try to buy a book that you've already bought from Amazon, the site asks you, "Are you sure? You bought this item already, on December 17, 2007." Amazon points you to your purchase history. That's because Amazon's data refinery has been designed with the goal of helping customers make decisions that minimize regret. Compare that to the approach taken by airlines' frequent flyer programs. Should a frequent flyer program remind you that your miles are about to expire and suggest ways to apply them to a purchase, or should it quietly let them expire, reducing the liability on the program's books? Even though it was quite simple to do, for years many programs didn't email a reminder.

Now let's consider the far-too-typical experience of calling your favorite friendly customer service center. At the start of the call, you'll inevitably hear the warning: "This call may be recorded for quality assurance purposes." You're given no choice: you must accept the company's conditions if you want to talk to a representative. But *why* is that recording only accessible to the business? What, really, does "quality assurance purposes" mean if only one side of the conversation is assured of having access to the record of what was agreed? Shouldn't you, the paying customer, also have access to the recording? That would be data symmetry.

Whenever I hear that my call might be recorded, I announce to the customer service rep that I'll also be recording the call for quality assurance purposes. Most of the time, the rep plays along. Occasionally, however, the rep hangs up abruptly. In those cases, I don't even know who I was speaking to. Of course, you could record the call yourself without asking for the rep's permission—which, I should note, is against the law in some places. Then, if you don't get what you were promised, you could throw the audio file online in the hopes that it goes viral. That might get the company's swift and diligent attention—as it did when one Comcast customer tried to cancel services but was rebuffed again and again, only succeeding after his mp3 file started trending on Twitter.[33] But it shouldn't be that customers have to break the law in order to acquire this symmetry of access to a two-way communication, and the power that comes with being able to publish it. You need to be pushing for more information to be public, not less. This will help to ensure that transparency becomes the new default.

But transparency isn't enough; you also need *agency*, the power you have to make free choices based on the outputs of the data refineries. This means not simply being able to click freely on a recommendation, but being able to ask the data refineries to provide information to you *on your terms*.

Many marketers talk about targeting, segmentation, and conversion. I don't know about you, but I don't want to be targeted, segmented, sliced, or diced. I don't want to be converted to anything that isn't of my own choosing. These aren't expressions of agency. Sometimes, when they're feeling more sociable toward their customers, marketers will use the word "engagement" to signal that they want to have a two-way conversation (or pretend that's what they want). Too often, though, the engagement that occurs is a battle for your attention as a likely customer.

On a fundamental level, customer agency involves giving people the ability to create data that is useful to them. Amazon wholeheartedly embraced uncensored customer reviews. It didn't matter to the company if the reviews were good or bad, five stars or one, written out of a desire to gain approval from others or to achieve a lifelong dream of becoming a book critic. What mattered was that they were relevant to other customers when they were deciding what to purchase. Reviews provided a

window into whether a customer regretted a purchase even though she did not return the item for a refund. That data helped customers figure out for themselves if a recommended product was the best choice for them. Amazon gave customers more agency when it came to evaluating goods.

Social data creators have a stake in the new data refineries, but they can't extract value from information that is buried, raw, and unrefined. To gain dividends from social data, you must give information away. Period. The value you reap from sharing data comes in the form of better decision-making ability, when negotiating deals, buying products and services, getting a loan, finding a job, obtaining healthcare and education for your family, and improving your community's safety and public services. The price you pay in sharing data should be offset by what you receive. There are risks involved in sharing data, but there are also ways to communicate the level of that risk to a refinery's users. Transparency about what refineries are learning and doing—including what they're providing directly to you as a benefit—is essential. So, too, is the ability to have some control over a refinery's products and services. Otherwise, how could you possibly judge what you give against what you get?

If data refineries are held to standards of transparency and agency, it will lead to what I call "sign flips"—reversals in the traditional relationships between individuals and institutions, from negative to positive. Amazon's decision to shift from letting companies write most of the content about products to letting customers do so is a sign flip, and the social data revolution will provide many similar opportunities. As individuals gain more tools to help them make better decisions for themselves, old-fashioned marketing and manipulation are becoming less effective. Gone is the day when a company could tell a powerless customer what to buy. Soon, you will get to tell the company what to make for you. In some places, you already can.

Sign flips are an important element in how physicists see the world. They are often associated with phase transitions, where a change in an external condition results in an abrupt alteration in the properties of matter—water changing from a liquid into a gas when it is heated to the boiling point. Our newfound capacity to process and transmit data are acting on society like increasing amounts of heat applied to the physical world. Under certain conditions—when refineries increase transparency and agency for users—the flip that takes place will benefit the individual over the institution, that is, it will benefit *you*.

Rather than trying to extort payment for data, you need to be demanding more sophisticated ways to gain control over how you share, when you share, what your data can be used for, and what you get as a result. The data refineries that are most successful encourage people to contribute raw data that improves the refined data products offered to you. But you can ask for more—to quantify transparency and agency, so that you can judge how much your decision-making is improved by using a particular refinery. We spend a lot of time as a society debating if limitations should be put on how organizations can use people's raw data, and not nearly enough time on what tools could be put into place to provide you with more useful and pertinent services and products.

Before we take a look at the personal controls that I argue must be built into data refineries for our benefit, I want to consider three sources of data—your clicks, your connections, and your context—and how to understand them through the lens of data literacy. As we'll see, these raw data streams challenge many of society's existing norms, including strongly held, often emotional issues. How do we establish a personal identity? To what extent is privacy an illusion? What does it mean to be a friend? How do you decide who to trust, when, and for what? How much are we influenced by our environment, and how much do we influence our environment? It may surprise you to learn that your Google search history, your Facebook profile, and your mobile phone appear to hold the answer to these questions.

#

Notes

1. Miller, George A., "The Challenge of Universal Literacy," *Science,* vol. 241 (September 9, 1988), p. 1293.
2. I have been using this metaphor for many years in my teaching, and described data refineries at two talks in 2011, one at the United Nations and the other at the O'Reilly Strata Summit, but I am not alone in making the comparison. Among those who refer to data as the new oil are Clive Humby, who helped create the British supermarket Tesco's Clubcard, one of the earliest loyalty cards to track all of the items in your grocery cart. My talk at the Strata Summit was covered by Miller, Rich, "In the Pipeline: A Tidal Wave of Data," *Data Center Knowledge,* September 26, 2011. My talk to the United Nations was part of the Secretary-General's Global Pulse initiative for Big Data innovation; video is available at www.youtube.com/watch?v=lbmsDH8RJA4. See also my Q&A with *BloombergBusiness*: Brustein, Joshua, "Consultant Andreas Weigend on Big Data Refineries," *BloombergBusiness,* March 6, 2014, http://www.bloomberg.com/bw/articles/2014-03-06/consultant-andreas-weigend-on-big-data-refineries.
3. The number of smartphones on the planet hit the 1 billion-mark way back in the fall of 2012. Heggestuen, John, "One in Every 5 People in the World Own a Smartphone, One in Every 17 Own a Table," *Business Insider,* December 15, 2013, http://www.businessinsider.com/smartphone-and-tablet-penetration-2013-10; Rushton, Katherine, "Number of Smartphones Tops One Billion," *Telegraph,* October 17, 2012, http://www.telegraph.co.uk/finance/9616011/Number-of-smartphones-tops-one-billion.html
4. Rogowsky, Mark, "More than Half of Us Have Smartphones, Giving Apple and Google Much to Smile About," *Forbes,* June 6, 2013, http://www.forbes.com/sites/markrogowsky/2013/06/06/more-than-half-of-us-have-smartphones-giving-apple-and-google-much-to-smile-about. **[UPDATE THIS LATER]**
5. Lunden, Ingrid, "80% of All Online Adults Now Own a Smartphone, Less Than 10% Use Wearables," *TechCrunch,* January 12, 2015, http://techcrunch.com/2015/01/12/80-of-all-online-adults-now-own-a-smartphone-less-than-10-use-wearables.
6. Tecmark, "Smartphone Usage Statistics 2014: UK Survey of Smartphone Users," October 8, 2014, http://www.tecmark.co.uk/smartphone-usage-data-uk-2014. If this number sounds high to you, consider an earlier "feasibility" study on a sample of one conducted by former Nokia executive and author Tomi T. Ahonen at his blog, *Communities Dominate Brands*. Ahonen, Tomi T., "An Attempt to Validate the 150x Per Day Number Based on 'Typical User,'" *Communities Dominate Brands,* January 22, 2013, http://communities-dominate.blogs.com/brands/2013/01/an-attempt-to-validate-the-150x-per-day-number-based-on-typical-user.html. Ahonen, Tomi T., "Remember That 150x Per Day Statistic?," *Communities Dominate Brands,* October 8, 2014, http://communities-dominate.blogs.com/brands/2014/10/remember-that-150x-per-day-statistic-now-there-is-a-survey-of-how-many-times-we-look-at-our-phones.html.
7. Miller, George A., "The Challenge of Universal Literacy," *Science,* vol. 241 (September 9, 1988), p. 1293.
8. As mass-market retailers, it's not surprising that mail-order catalog companies like Montgomery Ward and Sears & Roebuck evolved showroom-style department stores after the Second World War, when the car became a central fixture of American life. Kaufman, Leslie, with Claudia H. Deutsch, "Montgomery Ward to Close Its Doors," *New York Times,* December 29, 2000, http://www.nytimes.com/2000/12/29/business/montgomery-ward-to-close-its-doors.html.
9. The wide-scale promotion and adoption of credit cards starting in the mid-1960s simultaneously provided a method for efficiently collecting transaction data by individual customer account.

10. Founded in 1969, by the early 2000s Acxiom was generating revenues of more than $1 billion each year by selling consumer data to companies for use in advertising and marketing campaigns. Behar, Richard, "Never Heard of Acxiom? Chances Are It's Heard of You," *Fortune*, February 23, 2004, http://archive.fortune.com/magazines/fortune/fortune_archive/2004/02/23/362182/index.htm.
11. The title of *Bloomberg Businessweek* reporter Brad Stone's history of Amazon has become common shorthand for Jeff Bezos' entrepreneurial idea, which Stone encapsulates as a "grandiose vision of a single store that sells everything." Stone, Brad, *The Everything Store* (New York, Little Brown, 2013), p. 13. However, I and other Amazon insiders think of Amazon first and foremost as a data company—and indeed, Jeff came up with the idea while he was at D.E. Shaw & Company, the hedge fund that reinvented investing by focusing on big data analysis.
12. Duhigg, Charles, "How Companies Learn Your Secrets," *New York Times Magazine*, February 16, 2012, http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html.
13. "Suburban Soccer Mom" is a classification in the retailer Best Buy's database of 75 million customer households, one of the more aggressive efforts by a big-box chain to use data for personalizing offers. Kotler, Philip and Kevin Lane Keller, *Marketing Management 14* (Upper Saddle River, NJ: Prentice Hall, 2012), p. 71. Zmuda, Natalie, "Best Buy Touts Data Project as Key to Turnaround," *Advertising Age*, February 27, 2014, http://adage.com/article/datadriven-marketing/buy-touts-data-project-key-turnaround/291897. "Blue Blood Estates" and "Shotguns and Pickups" are two categories in the Potential Rating Index by Zip Markets (PRIZM) program developed by the marketing firm Claritas. Claritas is now a unit of the Nielsen Company—the people who provide point-of-purchase sales data to manufacturers and who measure TV audiences using set meters. Kotler, Philip and Kevin Lane Keller, *Marketing Management 14* (Upper Saddle River, NJ: Prentice Hall, 2012), p. 215.
14. **TO ADD CITE [Amazon 50% of purchasing starts]** *I cannot find a source for this statistic (50 percent of retail purchases start with an Amazon search) that you mentioned to me. Do you have one?*
15. Legal scholar Frank Pasquale has popularized the idea of "black box" algorithms, and led the argument railing against them, most recently in his book *The Black Box Society: The Secret Algorithms That Control Money and Information* (Cambridge, Mass.: Harvard University Press, 2015). As I discuss in more detail in this chapter and chapter 5, algorithms are mostly "secret" in the sense that they are difficult for most of us—even many scientists—to decipher. But it's also worth noting that the cost of building a data refinery from the ground up is not insubstantial, so the algorithms are also "secret" to the extent that they are the intellectual property of commercial enterprises.
16. Sedghi, Ami, "Facebook: 10 Years of Social Networking, in Numbers," *Guardian*, February 4, 2014, http://www.theguardian.com/news/datablog/2014/feb/04/facebook-in-numbers-statistics.
17. Lanier, Jaron, *Who Owns the Future?* (New York: Simon & Schuster, 2013), pp. 273–4.
18. Hansard has been used by several machine-based translation systems, including ones developed by IBM. Carter, Dave, and Diana Inkpen, "Searching for Poor Quality Machine Translated Text: Learning the Difference between Human Writing and Machine Translation" in Leila Kosseim and Diana Inkpen (eds.), *Advances in Artificial Intelligence: 25th Canadian Conference on Artificial Intelligence,* Toronto, May 28–30, 2012 (Berlin: Springer-Verlag, 2012), pp. 52–53.

19. Gittins, J.C., "Bandit Processes and Dynamic Allocation Indices," *Journal of the Royal Statistical Society B (Methodological)* 41 (2: 1979): pp. 148–77, http://www.jstor.org/stable/2985029.
20. **[Add credit to Jan Pedersen at Google]**
21. Gardner tackled the so-called secretary problem in the February/March 1960 issue of *Scientific American*, and later expanded the column for a collection of his writing. In the collection, the problem is presented under the heading "The Game of Googol." Gardner, Martin, *Martin Gardner's New Mathematical Diversions* (New York: Simon & Schuster, 1966), p. 35.
22. This is one of Holmes' most frequently quoted lines of dialogue. Although Holmes suggests he has come up with seven possible explanations for the odd offer of employment made to a governess, the reader does not get to hear him; instead, the full facts (data) of the case are revealed and the matter settled in short order. Doyle, Sir Arthur Conan, "The Adventure of the Copper Beeches," *Strand Magazine* (June 1892).
23. Inrix is just one of many commercial enterprises that analyze mobile phone location data to study traffic; data from Garmin and other dedicated GPS devices for route-planning and traffic reporting are also used to analyze trends, though of course the user base is more limited. In addition, there have been multiple academic studies on how call logs from mobile phones can be used to extrapolate commuting routes. See, for instance, Wang, Pu, et al., "Understanding Road Usage Patterns in Urban Areas," *Scientific Reports* 2 (1001: 2012), doi:10.1038/srep01001.
24. Samsung has a similar app, called Terrain Home, and others will no doubt be developed.
25. I shared a version of this story in an interview with *MIT Technology Review*. Unfortunately, some of the facts got garbled; the reporter couldn't understand how Google Now could alert me to the fact that I needed to get to the airport two hours *earlier* than scheduled. Regalado, Antonio, "The Data Made Me Do It," *MIT Technology Review*, May 3, 2013, http://www.technologyreview.com/news/514346/the-data-made-me-do-it.
26. Snyder, Brett, "Sabre Makes the Wrong Choice by Removing American Airlines," CBS News Moneywatch, January 7, 2011, http://www.cbsnews.com/news/sabre-makes-the-wrong-choice-by-removing-american-airlines.
27. American Airlines, "November Line of Sale Analysis," memo to R.E. Murray from S.D. Nason, December 3, 1981.
28. Tefft, Sheila, "Reservation Systems' Bias a Sore Spot for Smaller Airlines," *Chicago Tribune*, February 11, 1983, http://archives.chicagotribune.com/1983/02/11/page/87/article/new-technology.
29. Whiteley, David, *An Introduction to Information Systems* (New York: Palgrave Macmillan, 2013), p. 109.
30. After the Airline Deregulation Act of 1978 and the congressional debates over computer reservations systems in the early 1980s, the Civil Aeronautics Board adopted anti-bias regulations. Pearlstein, Debra J., Robert E. Iloch, et al., ed., *Antitrust Law Developments*, vol. 1 (Chicago: American Bar Association, 2002), p. 1428.
31. Agoda was acquired by Priceline in 2007. **TO ADD CITE HERE [based on ASW consulting with Agoda?]**
32. In engineering, prescriptive analytics is called "control theory."
33. In case you missed it, you can hear the excruciating recording of the call on SoundCloud: Block, Ryan, "Comcastic Service Disconnection (Recording Starts 10 Mins into Call)," SoundCloud, July 14, 2014, https://soundcloud.com/ryan-block-10/comcastic-service. Comcast issued a public apology for the customer service representative's actions on July 15, 2014, and Block's services were cancelled. However, Block argued on Twitter that the rep

was only trying to do the job that the company expected of him: to keep customers from leaving.

Chapter 2.
Character and Characteristics:
Are we the data we create?

#

"One does what one is; one becomes what one does."
Robert Musil

#

I'm a physicist by training, as are many of the people experimenting with social data today. That's not really surprising when you think about it: the digital traces we leave behind as we browse the web and use our mobile phones are very much like the trails and counts captured by a particle detector. In fact, working in experimental particle physics was the perfect training ground for conducting experiments in e-commerce.

In high-energy physics, you can't really observe the particle in and of itself. All you can do is observe the interactions the particle has with the matter in the detector you built. Physicists infer the properties of the particle by recording those interactions. At the CERN particle physics lab in Switzerland, I worked on data from a bubble chamber experiment, which involved measuring the trajectory and radius of the microscopic bubbles formed as a particle entered the chamber and interacted with a near-boiling liquid.[1] The path and size of those bubbles were used to calculate the particle's electrical charge and mass.[2] This principle of indirect observation applies to every experiment in particle physics: no one will ever see a Higgs boson, but most physicists are sure the particle exists given the indirect traces that have been observed.[3]

[insert art—bubble chamber trails]

People, like particles, have properties that can only be ascertained by watching how they interact and value other people and things, a bit like the mathematician hero of Robert Musil's *The Man Without Qualities* who depends on the world around him to define his very character.[4] The traces we leave reveal a great deal about us. In addition, new social data platforms, where we actively create and share data about ourselves, give us an opportunity to express the self as never before. You are able to "curate" the self, not just by changing the cut of your clothes or the style of your hair, but by choosing the information you want friends, strangers, companies, the whole wide world to see. This is a much richer definition of identity than the old-fashioned name, rank, and serial number.

As we rush to share unprecedented amounts of data, it feels natural to want to establish rules for safeguarding our character and our characteristics. For many, the obvious solution is to develop stronger privacy protections, but why? Privacy, as we understand it today, is a pretty new concept. It hasn't stood the test of time, or the test of evolving technology. Social data force us to develop new norms rather than fall back on rules that no longer do the job.

#

A brief history of the brief history of privacy

For most of history, humans have lived our lives in public. We shared our living spaces not just with close intimates, but with extended family, who gathered with us around an open hearth and knew our comings and goings. "Neighborly" behavior was maintained through close observation. Those who broke society's rules were subjected to merciless gossip; when people *really* broke the rules, they were ostracized, or worse.

Chimneys might very well be the first privacy-enabling technology.[5] In the seventeenth

century, the chimney became a common feature in European houses, allowing more families to divide their homes into private rooms with walls and doors that shielded individuals from the prying eyes of their relatives. Around the same time, a cluster of agricultural innovations changed how people could earn a living.[6] By the middle of the eighteenth century, food production was growing faster than the population—and the population was growing fast. Many people moved to the cities, where the first factories of the Industrial Revolution were going up.

City living was—and remains—anonymous living, and urban dwellers locked their doors to prevent strangers from entering their abodes. Their newfound privacy wasn't entirely pleasant; the design of fireplaces, which were usually quite deep to permit pots to be placed over the fire, were incredibly inefficient. Smoke built up in small rooms—this was the price of privacy. Conditions finally improved in the 1740s, when, who else, Benjamin Franklin suggested a design, the "Pennsylvanian Fire Place," that did a better job of heating a room while also forcing the fire's smoke up the chimney.[7] Finally, ordinary people could shut their doors without fear of asphyxiation. The home became the *sanctum sanctorum*, a place where any person could expect some semblance of privacy and safety. With the Fourth Amendment to the U.S. Constitution, Franklin and his fellow Founding Fathers went further, enshrining a right to security against illegal search and seizure of the person in the protection of his dwelling space.

While home life was becoming more insulated and private, political life was not. In our earliest experiments in democracy, voting was a decidedly social activity. The point, after all, was to encourage freedom of expression among regular citizens. As recounted by Harvard history professor Jill Lepore, for the first century of the United States, men voted in public, raising their hands or lining up on one side of a room (just as they still do in the Iowa presidential caucuses). "Casting a 'secret ballot'," Lepore says, was deemed "cowardly, underhanded, and despicable," an undermining of the openness and direct communication many thinkers believed was an essential ingredient in democratic governance.[8] For example, beginning in the 1850s,[9] the English philosopher John Stuart Mill argued a secret vote was vulnerable to "selfish" interests, and that "not secrecy, but publicity, should be the rule."[10] A gentleman should vote with public rather than private interests in mind, and what better way to ensure this than to have his vote out in the open and accountable, in other words, transparent.[11]

In those days, when only white property owners were enfranchised, the best technology available—paper ballots—was deemed to be elitist. Paper required voters to be literate, and not all men with property were. Eventually, however, paper won out as a more stable medium for recording votes than flailing hands or moving bodies. At first, voters were required to bring their own ballot papers to the polls. To make voting easier, they could mark the ballot in advance, or have somebody else do so for them. Those who could afford to print up ballots for others' use were practically invited to manipulate the choices available. Sometimes this led to flagrantly self-serving campaign tactics, as when a political party listed only its own candidates on a ballot paper supplied to loyal party members as well as random passersby. The paper ballot was not adopted for its privacy-enabling attributes but for its permanence. A permanent record could be re-counted.[12]

The first secret, government-printed ballot was used in an election in the Australian town of Victoria in 1856. It took a generation for the system to be adopted in Britain; cities and states in America only began the switch in the late 1880s. The new approach was a limited success. The percentage of the American electorate who turn up at the polls has never again reached the 80-percent norm of the mid to late nineteenth century,[13] partly because there is little social cost to not voting.

Around the time the secret ballot was gaining popularity, a couple of lawyers in Boston were making the case for a new "right to privacy." In what is considered to be the first use of the phrase in 1890, former law firm partners Samuel Warren and Louis Brandeis railed in the *Harvard Law Review* against the increasing intrusions into people's private lives. The offenders? "Recent inventions and business methods"—including photographs and circulation-hungry newspapers trading in gossip.[14] As

with many inventions, this "right to privacy" was devised to solve a personal problem: Warren and his family had recently been the victim of unflattering and unwanted sketches in the society columns.[15] They clearly didn't live at a time when a billion photos a day were being posted on Facebook.

Alas, the good lawyers, so eager to save their wives and daughters from social embarrassment, conflated the desire to control depictions of yourself with the right to manage what other people say about experiences they've shared with you. In a robust democracy, no one is compelled to express their thoughts and feelings to others, not even someone suspected of illicit behavior. If you share your secret with someone, there is always going to be a possibility that other people will hear it. (This is also true when it comes to data such as your geolocation, collected through the movements of your mobile phone; you have the freedom to switch off the service.) Laws can't stop such indiscretions, but social norms might—in those cases when the "public" gains more benefit by keeping things quiet. In engineering, it's often said that the purpose of communication is to transmit information. But as Facebook's founder, Mark Zuckerberg, intuited, the purpose of information is to give people an excuse to communicate.

A century ago, when Brandeis was confirmed to a seat on the U.S. Supreme Court, the right to privacy was among his most passionate causes. Brandeis began to tie the right to privacy to Americans' strong beliefs in personal liberty. Take *Meyer v Nebraska*, a case fought over whether the state of Nebraska could make it illegal for teachers to give instruction in the German language. This was just after the First World War, and anti-German sentiment was running high. The Supreme Court's majority asserted that a person had the right "to contract, to engage in any of the common occupations of life, to acquire useful knowledge, to marry, establish a home and bring up children, to worship God according to the dictates of his own conscience, and generally to enjoy those privileges long recognized at common law as essential to the orderly pursuit of happiness by free men," no matter where he made his home.[16] An attack on the right to privacy was an attack on freedom, according to the law.

Our personal choices seemed secure, locked away from peeping eyes and judgmental tongues—or so we fooled ourselves into believing. Aberrations like the McCarthy hearings were "outliers," a snooping into personal politics that was permitted because Communism was a great menace to free society. But our default assumption of privacy would dramatically flip with the development of tools for discovering information and communicating on the internet.

#

From walls to windows

In 1996, when Larry Page and Sergey Brin attacked the problem of web search by looking at the link structure between web pages, they had to rely on public data. Every web page crawled by Google was public. Someone had written the page and posted it on the internet so other folks could read it, and someone had linked to it.

Developing Google's algorithms and building its network of servers required a lot of money, and Larry and Sergey decided to pay for that by selling advertising space based on a user's search. Advertisers "purchased" keywords, phrases, and categories they believed matched the interests of their likely customers. The payoff was immediate. Advertisers who used Google's personalized ad option had four times as many people clicking through to their sites versus the average, including compared to ads placed on pages with content related to the product.[17] People's search data were a valuable commodity because they provided an observable trace of the things that attracted attention.

In April 2004, Google obtained another source of data about user attention when it launched Gmail, which conducts keyword analysis of a user's emails to determine the ads that appear on the service's web interface. Until this moment, most people thought of an email as equivalent to a letter—something sealed in an envelope, intended for the eyes of the addressed recipient only. Privacy advocates argued that Gmail users would be "giving away" their most personal communications to

Google if they signed up. Today it is the most widely used web-based email service in the world, with more than half a billion accounts active every month.[18] Users are comfortable having their correspondence scanned by Google's computers in exchange for free email. This trade-off is acceptable to most of the public.

Google's aspirations as a company go well beyond search. The prototype for Google Glass, released in March 2013, incorporated sensors capable of observing and recording the wearer's surrounding environment from her point-of-view. Critics raised concerns that Glass would be used to publicly share conversations without a person's consent. But it's not as though Glass is the only piece of technology that can be used for such a purpose. A handy audio-recorder or mini video camera can easily do the same job. Or you could pull out your mobile phone, which has those sensors and several more. Instead of having to pull out a device to record a moment, wearables like Glass are out and about with us, all the time.

A few months before Gmail's launch, a little website called Facemash went live at Harvard. In what is now a legendary story, then Harvard undergrad Mark Zuckerberg wrote software to "scrape" headshots from the online directories of nine residential houses so students could vote on which of two random photos was "hotter."[19] The site was wildly popular with Mark's classmates—and wildly controversial. Mark was in a lot of hot water with the university, which said he had violated copyright and individual privacy rights when he published the photos without permission. Once again, photos were being used by a new communication format as a way to satisfy the very human desire to gossip. Judge Brandeis would be horrified, but within a decade Facebook would be the default communication system for a large portion of the world's population.[20] As of the beginning of 2015, more than 20 percent of the world's population were on Facebook, and 1.25 billion people accessed it each month via their mobile phone.[21] Mark was pushing the boundaries of social norms, and a lot of people were eager to enter the uncharted territory of digital identity with him.

Of course, as Facebook grew, the company realized that it, like Google, would need to start selling advertising to generate revenue to support its services. The content of people's posts created even greater potential for targeting ads than an email message. People were identifying their relationship status, education level, political persuasion, and religious beliefs; creating lists of their favorite movies, TV shows, books, and music; reporting their travels; and sharing opinions about a host of brands and ad campaigns. They were uploading photos of themselves, their kids, their beloved dogs and cats. All of this was intended for a "public" of family and friends).[22] I was at Facebook on the day the site began running ads, and reviewed the feedback coming in from users.[23] It was eye-opening. For the most part, people weren't complaining that there were ads, or that the ads were based on their personal information; they were complaining that the ads had not done a good enough job taking into account the information they had shared. A typical example: "My profile page clearly says I'm a man and I'm interested in men. Why am I getting ads for a site where I can 'meet fifty-plus women'?" Users were asking to see ads for stuff they might actually want. A few years later, Facebook allowed users not simply to hide unwanted ads and "sponsored stories," but to explain why they didn't want to see the ad in the future so Facebook could improve its algorithms.[24]

The year 2016 marks the thirteenth birthday of Facebook. While it's unlikely that more than a few if any of the first batch of Harvard student users were posting baby photos to their Facebook profiles, very soon we're going to witness a generation of people who have had their entire childhood shared on Facebook by their parents and grandparents long before the kids could officially open their own accounts. In the past, a person graduated from high school with a handful of identity documents: a birth certificate, an immunization record, a transcript of grades, and a diploma. Most would have a driver's license. Some might have a reference from an employer or religious authority, maybe a passport. Every tween now comes preloaded with social data, created by parents, grandparents, aunts, uncles, older siblings, and family friends. You can find sonograms from before a child's birth,

commentary on difficulties maintaining a toddler's discipline, prayers about ill health, and details of physical appearance, skills, and hobbies. Why does Facebook still require its users to be at least thirteen years old in order to use the service? It makes more sense to set up a Facebook account for every baby at birth, since at least this would give people some ability to curate the data shared about them when they're old enough to decide for themselves.

We've evolved from the open hearth and its assumption of a public existence, with little experience or expectation of privacy, to enshrining a "right" to personal and political privacy behind the walls of our bedrooms and voting booths. As the internet became entwined in the fabric of social existence, we were more than happy to "go public" with our lives in exchange for free and immediate contact with family, friends, even strangers. The building of the idea of privacy and its dismantlement all happened in the span of a couple centuries—a blip in human history.

village gossip
no privacy

chimneys and urban migration (17c–1796)
social anonymity and the invention of privacy

U.S. Fourth Amendment (1792) and the adoption of the secret ballot (1856–1896)
privacy gets political

"The Right to Privacy" (1890)
privacy enshrined in law

Facebook, Gmail, and beyond (2004 and on)
privacy is an illusion—and we like *to share*

For the past hundred years we've cherished privacy, but the time has come to recognize that privacy is nothing more than an illusion. We *want* tools for managing attention, belonging, and communication. Judge Brandeis came up with a nice idea, but it was an idea of his time, when data was scarce, communities were localized, and communicating with others was costly. It was easy back then to stop someone from publishing a photo of you that you didn't like. Not so today. More important, the romantic conception of privacy that many people are fighting to defend assumes anonymity is the default setting of democracy, and thus it's essential to freedom. But history shows us otherwise, and it's better to write rules for the realities of the present and the possibilities of the future than to hope the rules of the past will continue to protect us. How can a right to privacy address developments like machine learning or image recognition software other than by shutting down the flow of raw data into the refineries altogether? To put data to work for the people, we need transparency and agency, more choices about how data are used to describe us, more windows with a view to the future and many fewer walls.

Rather than expending energy on delineating what's public and what's private, and then building walls to keep the data in (or keep it out), let's focus on the ability to truly be ourselves, including at the level of our passive digital fingerprints. Doing so will allow us to take full advantage of the data refineries and come to terms with the potential downsides of sharing.

#

On the internet, everybody knows you're a dog

When it comes to social data, it's no longer about whether you have privacy or not. Not anymore. There's a classic *New Yorker* cartoon by Peter Steiner with the punchline, "On the Internet, nobody

knows you're a dog."[25] Things have changed a lot since 1993, when the cartoon made its debut.[26] Today, a better adage would be, "On the Internet, everybody knows you're a dog. Sports a blue collar. Interested in cats. And your owners are on vacation." That's because you've shared this information with social data refineries in order to communicate with your friends and get recommendations about how best to navigate the web. The price involved seeing an ad for Puppy Chow.

People felt secure on the internet because they assumed they were mostly anonymous. But even before Facebook, our data exposed our identity. Shortly after the *New Yorker* published Steiner's cartoon, computer scientist Latanya Sweeney decided to find out exactly how anonymous an "anonymous" database of health data was.[27] The Commonwealth of Massachusetts decided it was in the public interest to share information about state employee hospital visits with researchers. The government officials weren't dumb; they knew it was inappropriate to share this data with people's names attached, so they removed identifiers, including each person's name, address, and Social Security number. In order for the data to be useful for improving health policy, they kept a few bits of relevant data: sex, birth date, and Zip code. By comparing those three bits of data to a second database—the voter registration rolls for the city of Cambridge, which was publicly available for a twenty-dollar fee—Sweeney was able to pinpoint the record of the state's governor, and "in a theatrical flourish, Dr. Sweeney sent the Governor's health records (which included diagnoses and prescriptions) to his office."[28]

Sweeney estimated that 87 percent of the American public could be identified if you knew a person's sex, birth date, and Zip code.[29] Later research put the figure closer to 63 percent—still a staggeringly high number given this could be done without access to more unique characteristics like the ones people share every day on Facebook and other social data sites.[30] Simple mathematics shows why it takes so few data points to pinpoint a person's identity. With about 42,000 active Zip codes in the United States, and a total population of around 300 million people, on average a Zip code has about 7,000 residents, approximately half of whom are male and half female.[31] If you assume an even distribution of births across the days of a calendar year, only about ten men or women in a Zip code would share a birthday; factor in the birth year, and it's plausible that you can narrow it down to one or two individuals.[32]

Now, consider the social data available to a typical refinery. The idea that a person couldn't be identified by digital traces was shattered when two big refineries shared "anonymized" social data with researchers. First, AOL released three months' worth of anonymized search logs of 658,000 users for academic study. The data was inadvertently posted so that anyone could download it, however, and two *New York Times* journalists tracked down several individuals based on search history alone.[33] They could do this quite easily because people like to search for themselves and their relatives, or for directions from their home address. A couple months later, video-renting site Netflix announced a contest to increase the accuracy of predicting a person's future rating of a movie. The researchers needed access to data in order to build their new models, so Netflix provided "100 million movie ratings, along with the date of the rating" from 480,000 customers.[34] The customers' names weren't included, but two researchers at the University of Texas at Austin identified people in the data set by comparing the "anonymized" data to reviews publicly posted on IMDb.com, the Internet Movie Database. Since those reviews were already public knowledge, what difference did it make? The users didn't post reviews for every movie rented from Netflix, and some of the "private" movie picks were quite revealing—or so argued the plaintiff in *Doe v. Netflix,* who feared her identity as a lesbian had been outed to the 50,000 researchers who had access to the Netflix Prize database. (The lawyer who filed Jane Doe's suit had previously fought to close down a Facebook feature that automatically posted a user's Blockbuster video rentals for friends to see; going forward you had to "opt in" to such sharing.)

The stream of live queries coming into Google and other search engines reveal a great deal

about humanity and about you. If you're like the majority of people, the most frequent address you enter into Google Maps is your home address. Where you live, where you want to go, what you need to buy, who you are curious about, and what you are worried about: these are among the most intimate details of our lives. The terms we search for reflect our societal preoccupations. Google encapsulates these concerns in the refined data of Google Trends. Most people think about celebrity news stories trending online, but Google Trends also reveals the rise in interest in "cyberbullying" and "transgender" over the past couple years. On the decline are searches for "privacy" and "transsexual."[35]

Now imagine being able to see search terms for an individual IP address. On a visit to an internet search engine start-up in the 1990s, I watched a string of queries go by. One of them caught my attention: someone had just searched for "how to commit suicide."[36] What do you do? Do you track down the user's identity from his internet service provider and alert a suicide hotline? Would that be an invasion of privacy? Do you first try to develop some way to analyze the query against the user's history, hoping to interpret the person's motivation and put a more specific probability on the action "predicted" by the query? Perhaps he is a novelist researching a character, and harbors no intent to harm himself; only additional data on the person's searches would allow you to discover if there is a correlation between the query and the user's future actions. But then you see the person's next search is for "Golden Gate Bridge"—where more than 1,600 people have committed suicide.[37] So you'd have to dig deeper into the person's digital history of behavior. Or do you step back from the monitor, from the individual, and merely collect the data along with all the other search terms, focusing instead on delivering the most useful search results and ignoring that a person's life might be at risk? There's no easy answer here.

Similarly, any e-commerce transactions reveal attributes about you, and occasionally about others. For Amazon to deliver your order to you, it has to have your credit card information, including your name and shipping address. It's in your interest to share your correct address; otherwise, you won't get your package. Your shopping history, however, may be a confused mix of purchases for yourself and for others. Amazon gives you the option to mark an item as a gift or to ignore it when making product recommendations. (You can do this on your "My Amazon" page, https://www.amazon.com/gp/yourstore/iyr). Using this data, the personalization algorithm learns to treat an item you say you bought for someone else differently from your other purchases. If you buy a shirt for a woman as a present, the purchase may seem to say very little about her. Yet, you are sharing data about her physical build when you select the shirt's size. If you buy the shirt in the week or two before Mother's Day, and the recipient has the same family name as you, Amazon's algorithm can infer your relationship. The site might send you an email a year later recommending great Mother's Day gifts.

The "Your Amazon" page provides some degree of transparency and agency for users. You can see your raw data inputs—your purchase history—and control which of these are refined into personalized recommendations. You can also add items that you've purchased elsewhere, whether recently or decades ago. In 2014, Facebook adopted a similar approach, giving you access to your Activity Log, a list of friend requests and status, likes, stories and photos in which a user is tagged, event RSVPs, and more. You can delete individual data points from that history, should you wish. (Wouldn't it be nice for a thirteen-year-old to be able to edit her parents' posts about her?) Your digital identity on Facebook is used to generate personalized ads, so deleting bits of your Facebook history also allows you to influence which ads you see.

In addition, Facebook activity reflects users' personality attributes quite accurately, as David Stillwell of the Psychometrics Centre at Cambridge University has found in his research. Stillwell recruited thousands of Facebook users to take a test that assessed the strength of their "Big Five" personality traits—openness, conscientiousness, extraversion, agreeableness, and neuroticism—and

then asked a separate group of subjects to review the individuals' Facebook profiles and give their own assessment of their personality. The two assessments matched surprisingly well. People tend to present an accurate portrait of themselves on Facebook—they are themselves, even when they are curating a profile on social media.[38] If a group of human strangers can assess your personality from your Facebook Timeline, you can be sure an algorithm can, as well. Deleting a like or two or twenty from your history of activity is unlikely to hide your overall pattern of behavior. Revealing weak conscientiousness on Facebook is the price you pay to keep your friends up to date on the events in your life.

Data about your attributes may coalesce without your active involvement. The huge numbers of photos posted online are a case in point. Not every photo of you is within your control, let alone your copyright. If you attend an event and someone snaps a photo as you're passing by, it's only a matter of time before your face is identifiable. Facebook's artificial intelligence research group, established in 2015 by Yann LeCun, can identify if two photos show the same face, nearly matching the performance of humans (who accomplish this feat with 97.53 percent accuracy).[39] The software, called DeepFace,[40] doesn't yet put a name to a face, but if a human tags one photo with a name, the algorithm can assume another photo with the same face is likely of the same person. Other software is being developed to analyze the background and context of a photo, distinguishing whether you're standing in a crowded bar or on an isolated mesa. If you tend to be photographed in one situation more than another, an algorithm might classify you as a social butterfly or a lonesome adventurer.

As Microsoft Research scientist Cynthia Dwork and others have shown, the very existence of data and databases opens everybody up to information disclosures. The point of maintaining a database is to get answers from queries, and a series of queries can be framed in such a way that only one person in the database provides a positive answer to all of them. Cynthia often demonstrates this with the example of asking what percentage of people in a medical database of Microsoft employees carry the sickle cell trait, then asking how many employees who are not female, curly-haired distinguished scientists have it. The difference between the two answers tells you if Cynthia—the only female, curly-haired distinguished scientist at Microsoft—has the trait.[41] "Differential privacy" involves adding noise to the system to avoid having people be distinguishable like this. As a database receives more queries, the added noise is less capable of protecting a person's identity. An individual is more identifiable the more a database gets used.

The main reason a person shares data with a refinery is to obtain personalized recommendations and other outputs that help steer decision-making. A database like the one described by Cynthia Dwork is relatively specific and constrained about the sort of data it collects, that is, "small" data. In comparison, the traces being collected by today's "big" data refineries are mind-boggling. To get useful outputs from refinery, you have to provide accurate inputs, such as your true interests and preferences. If you are not willing to share this data with a refinery, you can expect nothing better than the recommendations for the "average" person in the population, meaning, you're going to get whatever is most popular or relevant to Joe Public. If you supply incorrect data, you risk getting outputs that are totally useless to you. The price of gaining a bit more privacy is paid in less utility.

#

What's in a pseudonym?

Our distinct digital traces make anonymity practically impossible; worse, anonymity requires us to give up the benefits of having an ongoing identity and history. Still, it wasn't until Facebook that real names were a common sight on social data platforms. Pseudonyms were the norm. This was partly an issue of logistics. Some names are so common it was impossible to let everyone adopt a username that was the same as their real name when there was no other means for differentiating the users; some sites didn't provide usernames with enough characters to accommodate longer names. At the same

time, there were also people who didn't feel comfortable revealing their real name in certain situations because they feared identity theft, stalking, or repercussions at work or in their community for voicing unpopular opinions. In any case, you could maintain a different username for every website you visited, if you so wanted. As a result, the first two decades of the web were marked by an unprecedented ability to fragment identity. And in the process of adopting multiple pseudonyms, we redefined the function of identity in many of our interactions with other people.

Traditionally, a person's identity has consisted of simple data, like your name, date of birth, physical appearance, nationality, and place of residence—basic demographic information used to "authenticate" that you are who you say you are. Authenticating identity is important to enforcing many rules and norms of human interaction. Your age or nationality grants you certain privileges and responsibilities in society, such as the ability to vote or drink alcohol in public spaces, or the duty to pay taxes or serve in the military. For centuries, we've used identity passes to prove we're allowed to enter a territory or can pay for something with money safeguarded in a distant bank vault. We accept that we have to hand over a government-issued ID or number, type in a password, or answer a series of questions about our frequent flyer numbers or childhood pets in order to do quite a lot in life.

Many of the digital traces you leave are produced through your physical interaction with devices, and quite a few of these interactions are distinct enough to prove who you are, too. As people spend more time accessing the internet through mobile devices, social data refineries are investing a great deal of time and money to stitch together a person's identity across devices. One way to do this is to develop specialist apps with username log-ins, but there are more subtle clues, like the fonts you have installed and which ones you tend to use when you compose emails. Uri Rivner, the co-founder of Israel-based company called Biocatch, has described it as the potential to develop "a way to authenticate your mind by observing what you do and how you do it."[42] Biocatch and its competitors aren't interested in what you're searching for but how you're doing it. Do you thump your touchscreen vigorously or gently pat it? How much tremor is in your hand when you hold your mobile phone? How quickly do you drag your mouse? Do you prefer to open new tabs or navigate back and forth? What typographical errors do you regularly make, and which ones do you typically catch and correct? There are also highly predictable patterns in the order in which you tend to browse web pages each day, or the typical route you take, but you probably wouldn't want your bank to deny a transaction for which you broke out of your daily routine.

There are other components to identity, including your occupation, educational degree, and other credentials. We use these categories of identity-building to assign a reputation to people and predict how they will behave in the future. An established reputation attached to a name provides a shortcut for decision-making. For example, we often rely on the reputation of a corporate brand rather than investigating every ingredient in a food product or reading reviews of laundry detergents. If you're interviewing an Ivy League grad for a job, you might assume the candidate has a record of academic accomplishment without a glance at her college transcript. The reputation of a university provides a shortcut in deciding which candidates to interview.

Some markers of a good reputation can be faked, however. If you're in a hospital and a person wearing a white coat and a stethoscope around his neck asks you to undress, you're probably going to assume that the person is a legitimate doctor. Yet, people have been known to adopt a false identity, for one reason or another. In January 2015, a seventeen-year-old was taken into custody by police after spending a month at a Florida medical center posing as a doctor—a white coat and stethoscope having got him past the hospital's security guards. Likewise, the UK Competition and Markets Authority discovered some businesses had hired people to write "fake negative reviews to undermine rivals, for malicious reasons, or for personal gains," while others paid for "fake reviews to boost their ratings… compared to rivals." There were also been reports of customers were in allegedly "using the threat of a poor review to 'blackmail' businesses" into discounting services.[43]

Historically, pseudonyms have been used as a means for exercising freedom of expression. When "Publius" published the first of *The Federalist* papers in October 1787, "he"—that is, Alexander Hamilton, James Madison, and John Jay—was batting off scathing criticisms of the newly released draft of U.S. Constitution. Few of the combatants in the debate revealed their identities.[44] George Eliot, born Mary Ann Evans, adopted her pen name to avoid the stereotypes commonly attached to nineteenth-century woman writers, who she said—in an anonymous essay, no less—wrote "silly novels" marked by "the frothy, the prosy, the pious, or the pedantic."[45] She wanted her characters and her words to be taken seriously, which, she believed, would be impossible if readers pre-judged her writing by the name on the book's cover.

Sometimes, the motivation for adopting a pseudonym is less about freedom of expression than about making a break from past history. In 1947, a man called Hans Fallada (his "real name" was Rudolf Ditzen) penned *Every Man Dies Alone*, the fictionalized account of a German husband and wife who begin a quiet campaign of resistance against the Nazis. Fallada had been commissioned by a Soviet cultural attaché to review Gestapo files and weave a great anti-fascist tale out of them.[46] Yet, Fallada did not use a pseudonym because he was worried about linking his real identity to the politics of his books. It appears he wanted to divorce his writing career and his literary reputation from a failed suicide and subsequent stays in a sanatorium.[47]

These three famous pseudonyms share a characteristic: the owners of them wanted them to be persistent and gain a reputation. Publius always wrote in support of ratifying the Constitution. Eliot and Fallada used their names for all of their published works. These authors wanted to have their creative output tied to a single identity.

In the early days of the internet, adopting several pseudonyms seemed like a great option. Unfortunately, there was a problem: online, it's easy to create a new pseudonym if your original pseudonym proves disreputable to the community at large. Because it's easy to shed a pseudonym, newcomers are distrusted. How can you be sure this new username does not correspond to a person who was booted off the site a week earlier? A site might insist on having an email address registered with every pseudonym, but email accounts are easy to create, too. Some social data platforms responded by creating complicated registration forms, which make it slightly more costly to set up a new account, but such obstacles don't prevent a dedicated fraudster from hiring a legion of people or bots to fill them out. The "social cost of cheap pseudonyms," in the coinage of economist Eric Friedman and information scientist Paul Resnick, can't be eradicated this way.[48]

Could the cost of pseudonyms be increased enough to make them as useful as a real name? It depends on the situation. When you need to establish trust from the very first interaction, it makes sense to adopt a "real name" policy; doing so allows you to import a history of past behavior—say, with your bank or your credit card company. In contrast, adopting a pseudonym requires you to build up a reputation over time, starting from zero. The longer a pseudonym has been active, the more it is trusted, because other users can review the pseudonym's history of behavior.

When I was at Amazon, we looked at whether customer reviews were more valuable to other users if they were posted under a pseudonym or a real name.[49] We knew logging in to an Amazon account and adopting a pseudonym of some sort reduced the chances of "unuseful" reviews. We'd also seen that customers gave more weight to non-anonymous reviews. And whenever an Amazon customer changes pseudonyms, all of his reviews, past and present, are updated to display the new name so that each person's reviewing history remains intact; the person's identity is persistent even if the pseudonym isn't. Amazon could have insisted on a real-name policy for reviewers, since every one of Amazon's customers has a real name—as confirmed by the account's credit card. Would people who saw reviews from real names be more likely to buy a product, even if the reviews were no more glowing than those from reviews written under a persistent pseudonym? In the end, the most important factor was whether Amazon could verify the reviewer had actually bought the product.

People put more trust in opinions that are signed—but in this case, with data authenticating the reviewer's use of the product, rather than a total stranger's real name. In June 2015, Amazon increased the weight on verified-purchase reviews, making those reviews more visible and adjusting the "average" star ratings for a product accordingly.[50] (Amazon has also sued several companies for allegedly paying customers to write "five-star" reviews.[51])

Anonymity has other drawbacks. Consider the subtle differences between filling out a paper "comment card" at a place you frequent and answering an online survey. Even though a "comment card" is ostensibly anonymous, many people don't fill them in, and not just because they're lazy. People know they can be identified from a number of attributes—their handwriting, their word choice, the topics that they raise. They may fear repercussions for sharing negative comments. An anonymous comment is also, by definition, a "one-time game." There's no dialogue between the two sides, no chance to clarify a person's meaning or intentions, and no incentive to cooperate. This allows the recipients of the feedback to dismiss it—as noise, an outlier, something specific to the moment and not indicative of the need to make a change. Anonymous feedback can be shrugged off as malicious or self-serving.

Online discussion boards like Reddit have to address these failures of anonymity through machine learning. Reddit usernames can be utilized for every visit to the community or for a single post, never to be heard from again. Every pseudonym is given free rein to express itself all it wants, and users are encouraged to try on various personas when making comments on wide-ranging topics. At no time in the process are users asked to attach an email address or real name to the account; the site's founders had no desire to hold people accountable in this way. Accountability develops in other ways. Soon enough, other users chime in, ratifying a person's comments or arguing against them, either making their own comments or upvoting or downvoting someone else's. If a comment is downvoted four times, it is obscured, pushed to the bottom of the rankings of useful comments and marked as "deleted/removed." Instead of placing the opinion in a vacuum, a dialogue develops between users, who decide which comments are worth their time and which ones aren't. They quickly see if other people think there's truth to a claim. This is especially important in cases where a single piece of data might receive outsized weight, treated as though it's indicative of a pattern rather an outlier.

What matters most to Reddit is that those discussions which get featured on its "hot," "rising," and "controversial" lists are genuinely interesting to lots of different people, not lots of different pseudonyms. Being ranked among the top twenty-five discussions on one of those lists often leads to widespread attention across the internet. Instead of spending lots of time and money asking human moderators to enforce rules and regulations, Reddit has relied on machine learning to decrease the effect of "vote fraud," where users adopt different personas not to express opinions but to upvote their postings and downvote others'. When multiple pseudonyms are active and in synch, originating from similar IP addresses or displaying similar writing styles, the pseudonyms are "ringed" together as confederates. Downvotes received from confederates are given less weight, and occasionally ignored, when deciding whether a comment is obscured from a conversation or a topic gets elevated to Reddit's "hot" page.

#

Honest signals

Millions of people turn to websites and apps to meet people for casual encounters, dating, courtship, and long-term relationships. The problem is how to find the right person, that certain someone who matches your desires, wants what you have to offer and offers what you want, and—the hardest part—returns your interest. How can you judge the character of a stranger from a dating profile?

When it comes to dating, some people, on some topics, are truthful, some of the time. The distribution of truthfulness varies from one person and one situation to another. Sometimes people

"lie" or "cheat" because they want to gain some sort of benefit at another's expense, sometimes because they simply don't know what they really want. What people say is one signal. What they *do* is another. The signals revealed through a person's actual behavior are what social scientists call "honest signals."

Designing a dating app's user interface and recommendation algorithms is particularly challenging, since users might consider certain attributes dear to their heart when their site usage tells a very different story about who they find attractive. Christian Rudder, one of the co-founders of OKCupid, has shown that users may not fully realize or want to admit the strength of racial and ethnic preferences.[52] There's little incentive to be perfectly honest when answering such questions on your dating profile if you fear your answers might raise some eyebrows. But a simple count of clicks and contact messages quickly reveals a user's preference for a specific race or ethnicity.

When you're exploring dating options, would you prefer to base your decisions on the honest signals detected by software, or wait to see if you can deduce for yourself what's true and what's not? To some extent, it might depend on what people are lying about. While working as a consultant with one dating site in the mid 2000s, I discovered there were far more people reporting their age to be twenty-nine than thirty. This could not be true. Eventually, I traced the abnormal distribution back to the design of the site, which only allowed people to search by age ranges in five-year increments. You could look for someone who was "25–29" or "30–34." More people on the site were interested in users in the "25–29" bracket, so if you had just turned thirty, you had an incentive to shave a year off your age. Why not provide more fungible age searches that better reflect people's actual preferences? If a person is twenty-eight years old, she might like to meet people who are "26–33." And perhaps she'd like to be able to see who else might pop up in the results if instead she slid the upper bound to "34" or "35" or "36." In the real world, most people don't dismiss a potential date for being a few months too old for them. Redesigning the query interface gave users more decision-making power in filtering who they might want to meet.

This got me thinking more about how and when and why people lie on dating profiles. Did the fake twenty-nine-year-olds lie about their age when they first set up their profile or did they change their birth date after interacting with the app and realizing they were "too old" to show up in users' searches? One fix might involve letting users see the history of edits that others have made to their dating profile. Some edits would be considered acceptable and understandable to most users. After a few dates, you might want to revise the interests you listed because you felt you were overselling your rock-climbing prowess or discounting how much you enjoy going to concerts. Similarly, you might revise the description of who and what you are looking for. Other changes would be frowned upon by most users—such as frequently toggling between different relationship statuses. Surfacing a profile's edit history means that changes are far more likely to affect the person changing his profile, and far less likely to affect other users.

Imagine a scenario where users have the ability to see not just edits but also communication history. A common problem on straight dating apps is that women are often inundated with hundreds of messages while some men get no messages at all. To stimulate more symmetrical communication, dating apps have tried charging users to contact other users. If you have to pay per approach, you are more likely to narrow down your options through closer study of user profiles, and contact the people you've ascertained have a greater probability of replying. Another approach would be to limit the number of messages a person can send over a certain period of time. But because the inventory on dating apps changes daily, with users cycling in and out of circulation, this is a recipe for frustration. What if Miss Match activates her profile the day after you've used up your monthly quota, and by the time you can send messages again she's disappeared? You don't know if the algorithm has buried her further down the search results because you didn't get in touch earlier, or if she's started dating someone else. Why not use the power of transparency to reveal the honest signals of user behavior?

For example, each profile could indicate how many messages the person sends and receives each month, as well as the average response rate and response time. This would put you in a better position to decide whom to contact.

Dashboards of interest statistics are already being used on some dating apps. The gay men's chat and dating app Jack'd provides data on a user's reply rate to incoming messages and descriptive statistics about the people he's actually shown interest in (not just what he said interested him on his profile). This transparency allows users to fully explore not just their options but their chances. If a guy you want to approach only replies to 18 percent of messages, then you might decide to spend your time contacting someone else—especially if 44 percent of the users who catch his eye describe themselves as having "big muscles" and you are anything but that. Interestingly, Jack'd's data on a user's taste aren't based on incoming or outgoing messages, but on the user's "Favorites" list and the "Match Finder" tool, which allows users to express interest in someone and only alerts the other party if both people say they're interested.

Interpreting the motivation behind a click or contact is complicated, however. When working with Match.com, I came across a user who had blocked a large number of black women, all of whom described themselves as "curvy." The obvious hypothesis was that he prefers thin, non-black women, right? Wrong! When we looked at his filter settings and his clicks, it became clear that the opposite was true: He was singularly interested in curvy black women. He was blocking those women he'd already tried to contact with no luck. These are the fun problems you solve as a data detective. For a data scientists, coming up with good stories and telling them well is important..

To tell a story with data, you have to find a way to get inside the mind of the user. When I was working with another dating site, Singapore-based Fridae, we analyzed the remarks users had jotted down for themselves about other members. These annotations were visible only to the user who made them. We removed the usernames and looked through thousands of these private notes. They ranged from "sent me 5 messages; need to write back," and "met him—not my type" to "graduated with honors in chemistry" and "looks much older than 29." The notes accomplished two important goals. First, they helped Fridae's users remember the people they'd messaged or met with no luck so that they could avoid the effort or embarassment of contacting these individuals again. Second, the notes allowed Fridae's data scientists to uncover attributes that were important to users but not yet captured in Fridae's profile fields. Users were interested in very different things at 4 o'clock on a Friday afternoon compared to 4 o'clock on a Sunday morning. Who we are and what we want changes with the time of day—and the time of night.

Our recent experiences also shape how we think about personal identity, including our answers to big questions. OKCupid is known for the hundreds of user-generated questions that are deployed to determine relationship compatibility. A match-finding game poses a series of questions to be answered in ten seconds, ranging from "Do you believe in God?" to "Do you have a problem with racial jokes?" Some questions, like these, tend to have relatively static answers, but how about "Would you enjoy going to an all-night dance party?" or "Do you prefer to be rough or gentle in bed?" Would your answer be unaffected by what you'd done the previous night, or by the memory—happy or not—of your most recent relationship? OKCupid allows users to edit their answers. It also allows anyone on the site to view other users' answers. The order in which the questions and answers are shown to others could make or break your chance of making contact. Do you want to see where you disagree with a potential date, or where you agree? How about letting users apply weights to questions, indicating when a certain answer is a deal-breaker or indicates a sure thing?

Dating sites are increasingly giving users the option of revealing a real name and identity, such as a link to a Facebook page or Twitter account, as a means for showing their intention to engage in honest communication and behavior. Sebastiaan Boer, a data scientist at the mobile dating app Skout, wrote an algorithm to filter out inappropriate messages.[53] It was informally called the "creep

filter." What was inappropriate? Whatever was identified by the clicks and contact patterns of the app's users. If someone was blocked by lots of users, he was a creep—probabilistically speaking. But if someone sent repeated, unreciprocated messages to a specific user, he may be creepy—to that person. Over time, the algorithm learned what content tended to be found in messages that got a user blocked or went unanswered. Typically, negativity creeped in to the messages. Words such as "nasty" or "ugly" were one hallmark, but defining inappropriate was more nuanced. One person's turn-off was someone else's turn-on. Only after a pattern of blocking emerged did the creep filter stop a message from being delivered. In addition, excessive messaging to a specific user without ever getting a response would get throttled. The creep filter maintains a positive environment for the majority of users while giving the creep the chance to explore whether he can find someone who likes him.

I started this chapter by mentioning how my experience as a physicist have helped me design, run, and analyze experiments with social data. Many social data experiments involve watching how changes in the design of a refinery influence people's behavior. If you let a dating app user see that a person he is interested in rarely responds to messages, will he spend time carefully crafting a message for the object of his attraction, or will he skip the effort and look for someone more likely to respond? What's more likely to encourage a creep to stop sending messages, an administrator telling him to stop or the feedback of getting zero responses? How does experimenting with identity attributes change a user's response rates? If a person experiments with identity, when does it cross the line and turn him into nothing more than fraudster in the eyes of other users? Greater transparency about users' behavior allows people to decide for themselves whether the character presented in a profile truly matches up with the characteristics of their ideal soulmate.

#

Calling for accountability

"When it comes to privacy and accountability, people always demand the former for themselves and the latter for everyone else."

David Brin

#

As people spend more time accessing the internet through mobile devices, another pseudonym is playing a significant part in the formation of identity: your phone number. When phones were first installed in homes, an operator would call and ask whether you wanted to accept the call from the other party. With the development of rotary pulses, people could call each other directly. The caller's identity was no longer announced in advance. You had to answer the call to find out who was on the other end of the line. Yet, as long as the price of making a phone call remained pretty high, families got very few unwanted phone calls. When prices started falling, telemarketing became financially viable. Then, in 1990, around the same time as the web was invented, a tone-dialing system was introduced, making Caller ID possible.

Initially, there was some resistance to the idea that your phone number—and possibly your name, if your number was listed—would be automatically transmitted to the person you were calling. Now, this has flipped, and people are unlikely to accept a call when the caller's identity has not been shared with them. "Unknown" numbers are sent to voicemail. In order to get people to answer your call, you have to let them know it is you calling. You have to make your number public in some way. You might feel more secure hiding your identity, and you might feel more secure if the other person is communicating their identity, but you can't have it both ways. Phone communication works better when it is symmetrical—that is, when both sides of the call know each other's identity.

Alex Algard, the founder of the online phone number directory White Pages, believes it's possible to force greater transparency on phone communications for the benefit of all users. By taking advantage of its massive database of phone numbers, White Pages provides identification for incoming calls, regardless of the caller's settings or the contact list on your mobile phone, a

particularly useful service in these days of increasing phone "spam." If the caller hasn't got a published phone number, White Pages can often still identify the caller by category, such as "telemarketer," based on mining online sources and analyzing the pattern of calls from the number to White Pages subscribers. It can also attach other components of identity, such as a Facebook photo, to the incoming call, to remind you of the face that goes with the name, for instance. Identity is richer than ten digits; it's your history of interaction with the person you're calling right now, yesterday, and further in the past. As a society, we have to decide if both sides of the conversation have a right to know the identity of the person at the other end of the line. If the answer is yes, then we have to decide what you are allowed to do with that data.

This is an especially tricky issue because a persistent identity is a necessary but not a sufficient condition for generating trust. Knowing who someone is merely provides a means by which you can call an individual to account if they misbehave. I know a couple people who have posted screenshots on Facebook of what they felt was inappropriate behavior on a dating app. In the first case, the other party wouldn't take "I'm not interested" as an answer; in the second case, the other party was insulting. Both recipients could have hit the "block" button and left it at that. But both of them decided to share the bad behavior with their friends.

What are the expectations of privacy when you communicate with someone today? The recipients of the unwanted dating messages could argue that publishing this "private" message served the best interests of their community of friends. The screenshots served as a warning to other individuals who may be on the app about the person's misbehavior. Indeed, neither culprit's photo nor username was deleted from the screenshots, so there was nowhere to hide among the recipients' Facebook friends. Similarly, if your boss sends you an unfair email rant, you can easily forward it to your friends or post it online. The law might say you were in the wrong for doing so, because the email was a "confidential" communication, only intended for distribution within the company. But sharing the email also has a public benefit—it lets potential employees know a bit more about the company's working conditions.

To some extent, how we react to a person sharing a private communication depends on how trustworthy we think she is. A screenshot can be faked very easily. On a discussion board like Reddit where users are essentially anonymous, you have little way of authenticating the poster's identity, let alone the messages themselves. On Facebook, the poster is usually a person we know (or who knows someone we know). And since accounts are infrequently hacked, the person's decision to post the screenshot is reined in by the downside of having friends know she might share private messages with others. Still, this doesn't mean we should assume the screenshot is real. It might have been fabricated with the intent to discredit someone.

Once the screenshot has been posted, it was just like any other bit of data: it could be shared freely by anyone who came across it. What if one of those Facebook friends was outraged or amused by the message and took a screenshot to share with her friends? Or if a friend of that friend decided to Tweet it? At some point, someone—or some algorithm—will probably recognize the person's face, and a real name will get attached to the "bad" behavior. By then, it will no longer have the cover of providing some benefit to a group of friends, or the context of the original poster's identity. But it may be discoverable when searching for the person's name.

What protection might a person get for their character online in the future? One option was suggested by the European Court of Justice in May 2014, when it ruled in favor of a person's "right to be forgotten." A Spanish man was tired of prospective employers and landlords turning up an article about how he'd lost his home because of unpaid taxes, especially since he'd gone on to pay off his debts.[54] He didn't want to purge the record of the foreclosure, or to remove the news article from archives. He simply wanted the page to stop showing up when people searched for his name on Google. The court decided that people should have the right to have links "delisted" from search

results when they felt they were being harmed by them. On the first day after the EU ruling went into effect, more than 12,000 requests were submitted to Google; in the first year, more than 275,000 flooded in.

Google has publicized some of the requests to remove links that were submitted to the site after the Spanish man's victory. An Italian woman asked that an article about the murder of her husband more than a decade earlier be removed from searches for her name. A Latvian activist who was injured during a protest asked that an article about the protest be removed from searches for the person's name. A German teacher "convicted for a minor crime" more than a decade earlier asked to have an article about the conviction removed from searches, too. In each of these cases, Google decided that the individual's right to be forgotten outweighed the "public interest in the content."[55]

These requests appear admirable, but should it be up to Google's algorithms—and, presumably, its lawyers—to decide what is in the public interest? Google notifies media of articles to which links been removed, but not who has put in the request, or what personal harm has been alleged. Grappling with the question of public interest, the *Guardian* newspaper decided to buck the law and inform readers of the first six stories from the paper removed from some search results.[56] Other media outlets followed suit, with the BBC publishing a list of 182 delinked articles.[57] By informing media about delisted articles, Google provides fodder for further publicity. To ensure more agency for people, I believe the individuals should choose

One particular story demonstrates the folly of trying to erase information from the internet. Just days after the directive was adopted, private-equity investor Greg Lindae asked that a 1998 *Wall Street Journal* article about a Tantra workshop in which he had participated be removed from Google results served up to European IP addresses. The order—and the too-sexy-to-resist Tantra session—garnered a new round of coverage in the U.S. edition of the newspaper, whose editors decided to track down the person who had put in the request because they felt the article continued to serve "the public interest." Ironically, the extra publicity means there are now more high-ranking search results for "Tantra" that mention Lindae by name. Lindae has decided to take the coverage in stride: "If it adds a little more context… that is not a problem. That is actually better."[58]

Back in 1890, when those two legal eagles, Samuel Warren and Louis Brandeis, made their case for a "right to privacy," they were particularly interested in an individual's fundamental ownership of a *personality*, a concept they borrowed from the philosophers of my homeland, Germany.[59] Who wants to have an embarrassing snapshot published for everyone to see, without you having any say in the matter? The idea was that the law should require people to treat each other humanely. The "right to privacy" enshrined in the law was intended to preserve dignity, not foster liberty.[60] It was commonly believed at the time that unchecked liberty would lead to a tyranny of the masses. Liberty was a bad thing.

An insightful paper by two law professors, Paul Schwartz at Berkeley and Karl-Nikolaus Peifer at the University of Cologne, looks at how notions of privacy and personality protect us (or fail to do so) in the courts.[61] They describe a "kiss-and-tell" memoir by a bestselling American author recounting her struggle with vaginal pain and its effect on her physical and psychological health, including her relationship with her former boyfriend. The author never named the boyfriend and changed details of his life, but the man said his friends and business associates all knew about their relationship, and the depiction of their sex life had caused him to suffer "severe personal humiliation" and "considerable damage to his reputation."[62] The judge agreed that the man was identifiable, and that the ex had been painted in a pretty bad light, but asserted that the public benefit of the memoir was greater than the harm to him. Only a small circle of people could identify him based on publicly available attributes, and this was most important. The other book, published in Germany, was an autobiographical novel featuring thinly veiled versions of the author's ex-girlfriend and her mother. Though the novel contained "a traditional disclaimer that all characters in it were invented," the

German judge found that any person who knew the girlfriend or mother could see the characters were based on them.[63] However, the judge said, only the girlfriend's rights had been harmed, since the girlfriend's sex life was ostensibly private, taking place behind closed doors, while the mother's interfering involved other people, and thus was publicly known already. The right to personality protected the girlfriend from the indignity of having her sex life sold for public titillation. The novel was pulped.

These two court decisions seem almost quaint when you think about my friends posting screenshots from a dating app on Facebook. Let's assume the screenshot is real, that the offending party doesn't try to pretend the words aren't his. How would a judge consider the right to privacy or the right to personality there? Is a chat on a dating app presumed to be private? Would declaring a motivation of warning off friends protect the poster from getting in trouble? Would it matter if the screenshot obscured the name and photo of the person who'd sent the unwanted messages? What if someone—either the offended party or a friend—tagged the photo with a person who had a similar name and physical appearance, all but saying it was the same person and notifying the person's friends?

I also mention these decisions because of the critical role of weighing public benefit against private harm plays in them. The exponentially increasing amount of social data clearly presents unprecedented opportunities. How much personal harm is required before we feel the harm outweighs the benefit to the aggregated masses, and how do we measure it? Would you sacrifice being able to see data about a potential date's changes to his profile because such reputational information could be inadvertently or maliciously shared with his friends or coworkers? With more data refineries being designed for supporting decisions in areas ranging from employment to healthcare to education, we need to develop more sophisticated tools for evaluating these trade-offs.

As science fiction author and technology writer David Brin notes, it seems everybody wants privacy for themselves and accountability for the people they interact with. You can't have it both ways. And because privacy is an illusion, we all need to get used to being more accountable. A good start is to be more accountable to our friends.

#

Notes

---

1. I worked on bubble-chamber experiments at the CERN (the European Organization for Nuclear Research) near Geneva while an undergraduate student at the University of Karlsruhe, and later, as a graduate student at Stanford, I designed, ran, and analyzed experimental results from the Mark II particle detector at the SLAC National Accelerator Laboratory in Menlo Park, California.
2. Thanks to Gregor Hochmuth, founder and data engineer/strategist at DADA, who first suggested this analogy in a workshop I hosted on the social data revolution in 2010. At that time he was a product manager at Google working on interest-based ads and the Chrome web browser. **[CONFIRM]**
3. The Higgs boson was not detected in a bubble chamber but at two particle accelerators at CERN's Large Hadron Collider.
4. The German title of Musil's book, *Der Mann ohne Eigenschaften,* translates literally as "The Man Who Has No Properties." I suspect the English publisher was worried that readers might imagine that the protagonist hadn't managed to buy any real estate.
5. Credit for introducing me to the idea of the chimney as a privacy-enabling technology goes to the remarkable John Taysom, founder of BlinkBox Music and Reuters Venture Capital. The ability to maintain some privacy while also gaining access to the personalized services of

the data refineries has been a key project for John, who holds two patents for a "method of anonymising an interaction between devices." More on that idea in chapter 5.
6. This was particularly true in Britain, where land was shifted from the commons to private ownership as part of the enclosure movement. Enclosure was among the most pertinent changes in society, since it led many people to seek their fortune in the city. But other agricultural innovations starting around 1750 were also instrumental in creating conditions that intensified urban migration. These included new and imported techniques such as crop rotation to reduce the span of time when productive soil was left fallow; selective breeding of livestock to increase meat yields; use of more efficient metal plows, a technology copied from the Chinese; and adoption of more efficient land draining, a technology imported by the Dutch. On the "communication" side, an extensive network of canals was constructed so that crops could be transported to centralized markets more quickly than by horse and cart. This was a perfect storm of technologies, fundamentally reshaping societal norms and laws, as proposed in Overton, Mark, *Agricultural Revolution in England: The Transformation of the Agrarian Economy 1500–1850* (Cambridge: Cambridge University Press, 1996).
7. Benjamin Franklin's Pennsylvanian Fire Place was actually a stove, not a bricked part of the house, but his chimney was unquestionably revolutionary. However, Franklin isn't the only person deserving some credit for inventing a "smoke-free" chimney. Sir Benjamin Thompson, later Count Rumford, spent considerable time perfecting the design in London, publishing his approach in 1796. The Rumford fireplace was the standard for much of the nineteenth century. The details of chimney engineering are based on the work of Orville R. Butler, a historian at the American Institute of Physics. Butler, Orville R., "Smoke Gets in Your Eye: The Development of the House Chimney," http://www.ultimatehistoryproject.com/chimneys.html.
8. This history of the secret ballot is greatly indebted to Lepore's fascinating account in *The New Yorker*. Lepore, Jill, "Rock, Paper, Scissors: How We Used to Vote," *The New Yorker,* October 13, 2008, http://www.newyorker.com/reporting/2008/10/13/081013fa_fact_lepore?currentPage=all.

The French Constitution of 1795 dictated election by secret ballot. But with at least 16,500 "counter-revolutionaries" executed at the guillotine in the Reign of Terror, the *sans-culottes* may have seen some wisdom in keeping their political preferences to themselves. After the republic broke down the secret ballot became history, and only in 1913 did France re-establish secret voting. Fremont-Barnes, Gregory, *Encyclopedia of the Age of Political Revolutions and New Ideologies, 1760–1815,* vol. 1 (Westport, Conn.: Greenwood Press, 2007). 617.
9. There is great debate among academics about the timing of Mill's shift from being an advocate for the secret vote to being an opponent of it. Mill had taken up the opposing view by at least 1853, according to a review of his surviving letters. Buchstein, Hubertus, "Public Voting and Political Modernization," in John Elster, ed., *Secrecy and Publicity in Votes and Debates* (Cambridge: Cambridge University Press, 2015), pp. 29, 30.
10. Mill, John Stuart, "Thoughts on Parliamentary Reform," in *Dissertations and Discussions: Political, Philosophical, and Historical,* vol. 4 (New York: Henry Holt and Company, 1873), pp. 36–7.
11. Buchstein, Hubertus, "Public Voting and Political Modernization," in John Elster, ed., *Secrecy and Publicity in Votes and Debates* (Cambridge: Cambridge University Press, 2015), p. 31.

Mill was a champion of individual freedom. In his masterwork, *On Liberty*, he set forward the "harm principle": "The only purpose for which power can be rightfully exercised over any member of a civilized community, against his will, is to prevent harm to others."

Mill, John Start, *On Liberty* (Oxford: Oxford University Press, 1859), p. 22. Have some free time? The entire 1859 edition of *On Liberty* is available to read at Google Books. https://books.google.com/books/about/On_Liberty.html?id=3xARAAAAYAAJ **[May move this EN / graf to chapter 11]**

12. Paper ballots still needed to be counted by humans, and often that meant the ballots weren't 100-percent secret. To my amazement, when Thomas Edison applied for a patent in 1869 for a mechanical voting machine that used toggle levers, he found there was no market for it. Politicians wanted to hear from their public—and, in what we might imagine were quite a few cases, influence them. It wasn't until the late 1950s that Edison's lever-voting machine gained some backers, when an American company began to sell a model under the slogan, "Behind the Freedom Curtain." People were finally "free" from anyone knowing their political allegiances unless they chose to share them because a ballot could no longer be traced back to an individual, only to a machine. Stephey, M.J., "A Brief History of Ballots in America," *Time*, November 3, 2008, http://content.time.com/time/politics/article/0,8599,1855857,00.html.

Machine-counted ballots aren't the norm everywhere. For instance, in the United Kingdom, "secret" ballot papers are still voted by hand and can be traced back to the voter who cast them by means of a serial number—an anti-fraud protection that some claim was used by British intelligence services to identify individuals who voted for Communist Party candidates in the 1960s. Beetham, David, and Stuart Weir, *Political Power and Democratic Control in Britain* (London: Routledge, 1999), p. 77

13. Of course, for decades some Americans were barred from voting, even when they showed up on election day, through the widespread use of "poll tests"—literacy tests specifically targeted at those individuals, mostly black men in the South, whom local election officials preferred to send home. Poll tests were a bald-faced rebuke to the Fourteenth Amendment, which affirmed that no person's "life, liberty, or property" could be deprived by a member state of the Union.

14. Warren, Samuel D., and Louis D. Brandeis, "The Right to Privacy," *Harvard Law Review* 4 (5: December 15, 1890), http://groups.csail.mit.edu/mac/classes/6.805/articles/privacy/Privacy_brand_warr2.html.

15. Glancy, Dorothy J., "The Invention of the Right to Privacy," *Arizona Law Review* 21 (1: 1979), pp. 9–10, http://digitalcommons.law.scu.edu/facpubs/317.

16. The decision meant that the teacher was allowed to teach German, but the precedent has been used as the foundation to a "right to privacy" in many other areas of life, from a married couple's personal decision to use birth control (*Planned Parenthood v. Casey*) to a gay couple's personal decision to have consensual sex (*Lawrence v. Texas*). *Meyer v. Nebraska*, 262 U.S. Supreme Court 390 (1923), p. 399, https://supreme.justia.com/cases/federal/us/262/390/case.html.

17. Google, "Google's Targeted Keyword Ad Program Shows Strong Momentum with Advertisers," press release, August 16, 2000, http://googlepress.blogspot.co.uk/2000/08/googles-targeted-keyword-ad-program.html.

18. This is a conservative figure based on published news stories. By 2012, Google was publicly trumpeting in earnings calls with investors that it had 425 million "monthly active users." However, Google has not been transparent when it comes to data about its user base. Ludwig, Sean, "Gmail Finally Blows Past Hotmail to Become the World's Largest Email Service," *VentureBeat*, June 28, 2012, http://venturebeat.com/2012/06/28/gmail-hotmail-yahoo-email-users.

19. This wasn't a novel idea: comparing people's photos was first popular on University of California–Berkeley grads James Hong and Jim Young's "Hot Or Not" rating site, which went online in October 2000.
20. The extent of Facebook use is limited in part by blocks in several countries. For instance, since at least 2009 the People's Republic of China, stating that the site has been used by groups to organize violent protests, has mostly blocked access to Facebook. Rival sites like Renren have gained users as a result. Chen, George, "China to Lift Ban on Facebook—But Only Within Shanghai Free-Trade Zone," *South China Morning Post,* September 24, 2013, http://www.scmp.com/news/china/article/1316598/exclusive-china-lift-ban-facebook-only-within-shanghai-free-trade-zone.
21. "Number of Monthly Active Facebook Users Worldwide as of 1st Quarter 2015 (in Millions)," Statista, http://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide.
22. This number is based on a survey by the Pew Research Center, since Facebook does not release user statistics. Smith, Aaron, "Six New Facts About Facebook," Pew Research Center, February 3, 2014, http://www.pewresearch.org/fact-tank/2014/02/03/6-new-facts-about-facebook.
23. I was visiting a former student who had taken a job as one of the first data scientists at the company. **[TO ADD APPROX. DATE]**
24. McDermott, John, "Facebook Is Cracking Down on Ads Users Hate," *Digiday,* September 12, 2014, http://digiday.com/platforms/facebook-ads-hate.
25. As a side note, I only learned that Peter Steiner was the person to draw those two pups and dream up the punchline because the top search result out of 591,000 told me that he was the cartoon's author.
26. The cartoon has been reproduced more than any other in the magazine's history, according to the *New York Times.* Fleishman, Glenn, "Cartoon Captures Spirit of the Internet," *New York Times,* December 14, 2000, http://www.nytimes.com/2000/12/14/technology/cartoon-captures-spirit-of-the-internet.html.
27. At the time of the study, Sweeney was a Ph.D. candidate. She is now professor of government and technology at Harvard University and director of the Data Privacy Lab at Harvard.
28. Ohm, Paul, "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization," *UCLA Law Review* 1701 (2010), p. 1720.
29. Sweeney, Latanya, *Uniqueness of Simple Demographics in the U.S. Population,* Laboratory for International Data Privacy working paper LIDAP-WP4-2000. Later studies, according to Paul Ohm in "Broken Promises of Privacy," have put the figure at two-thirds of the population. And that's without access to more unique identifiers like the ones we share every day on sites like Facebook.
30. Golle, Philippe, "Revisiting the Uniqueness of Simple Demographics in the U.S. Population," *Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society* (New York: Association for Computing Machinery, 2006), pp. 77–80.
31. I looked up the current number of active Zip codes at the U.S. Post Office FAQ (http://faq.usps.com), since it seemed probable that not all of the 90,000 possible numbers (10000–99999) had been allocated. If they had, the percentage of people who are uniquely identifiable would be higher.
32. One reason the identification rate isn't higher is because the population isn't uniformly distributed across Zip codes, and about 22 percent of Zip codes are assigned to P.O. boxes.
33. One, Thelma Arnold of Lilburn, Georgia, agreed to have her identity revealed in the newspaper. Barbaro, Michael, and Tom Zeller, Jr., "A Face Is Exposed for AOL Searcher No.

4417749," *New York Times,* August 9, 2006, http://www.nytimes.com/2006/08/09/technology/09aol.html.
34. Singel, Ryan, "Netflix Spilled Your *Brokeback Mountain* Secret, Lawsuit Claims," *Wired*, December 17, 2009, http://www.wired.com/2009/12/netflix-privacy-lawsuit.
35. Based on Google Trends, "big data" wasn't on the public's mind until 2011.
36. A meta-analysis conducted by researchers at Oxford University found that young people who self-harm often search online for information about self-harming: "In one of the studies reviewed, well over half (59%) of young people interviewed said they had researched suicide online. Meanwhile, of 15 teenagers who had carried out particularly violent acts of self-harm, 80% said they had gone online to research self-harm beforehand." Daine, Kate, Keith Hawton, Vinod Singaravelu, Anne Stewart, Sue Simkin, and Paul Montgomery, "The Power of the Web: A Systematic Review of Studies of the Influence of the Internet on Self-Harm and Suicide in Young People," *PLoS One,* October 30, 2013, http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0077555.
37. This figure comes from former California Highway Patrol sergeant Kevin Briggs' TED Talk about patrolling the Golden Gate Bridge. Briggs shares heart-breaking and eye-opening perspective on the bridge and advice on speaking with a love one who you think may be contemplating suicide. Briggs, Kevin, "The Bridge between Suicide and Life," TED, March 21, 2014, https://www.ted.com/talks/kevin_briggs_the_bridge_between_suicide_and_life.
38. Bachrach, Yoram, Michal Kosinski, Thore Graepel, Pushmeet Kohli, and David Stillwell, "Personality and Patterns of Facebook Usage," in *Proceedings of the 4th Annual ACM Web Science Conference* (New York: ACM, 2012), pp. 24–32.
39. Simonite, Tom, "Facebook's New AI Research Group Reports a Major Improvement in Face-Processing Software," *MIT Technology Review,* March 17, 2014, http://www.technologyreview.com/news/525586/facebook-creates-software-that-matches-faces-almost-as-well-as-you-do; Taigman, Yaniv, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf, "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," paper presented at the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, Ohio, June 24, 2014, https://research.facebook.com/publications/480567225376225/deepface-closing-the-gap-to-human-level-performance-in-face-verification.
40. DeepFace is a "deep learning" process, which relies on feeding large amounts of data into a so-called neural network, a computational model based on connections of processes similar to the structure of communication between neurons in the brain. We'll investigate network theory as well as DeepFace and other deep learning projects in more detail in the next two chapters.
41. Cynthia often uses this example in discussing differential privacy, for example in the lecture, "I'm in the Database (But Nobody Knows), Dean's Lecture, University of California–Berkeley School of Information, February 4, 2015. Dwork, Cynthia, "Differential Privacy," in *Automata, Languages, and Programming: Lecture Notes in Computer Science* 4052 (2006), pp. 1–12.
42. Leber, Jessica, "Forget Passwords: This Startup Wants to Authenticate Your Mind," *Fast Company Exist*, July 24, 2014, http://www.fastcoexist.com/3033383/forget-passwords-this-startup-wants-to-authenticate-your-mind.
43. Peachey, Kevin, "Online Reviews 'Used as Blackmail,'" *BBC News,* June 19, 2015, http://www.bbc.co.uk/news/business-33184207. [Possibly add this example: "Ashley Booth Griffin, from New York, supposedly posted a positive review for a loan website, but in fact she was killed in a car crash seven years ago. The photo used in the review came from her memorial website."]

44. In 1804, Alexander Hamilton was the first of the men to remove the cloak from the authorship of the papers, sharing an itemized list of each essay's author with his lawyer a few days before his duel with rival Aaron Burr, during which Hamilton was killed. By then, the debate over ratification of the Constitution had been settled.
45. Anonymous, "Silly Novels by Lady Novelists," *Westminster Review* new series 10 (October 1856), p. 442.
46. Wilkes, Geoff, "Afterword," in *Alone in Berlin* [English title of *Jeder stirbt für sich allein* ("Every man dies alone")] (London: Penguin, 2009), pp. 578–9.
47. Fallada killed his companion, but did not succeed in killing himself. Oltermann, Philip, "The Cow, the Shoe, Then You," *London Review of Books* 34 (5: March 8, 2012), p. 27.
48. The seminal paper outlining this concept is Friedman, Eric J., and Paul Resnick, "The Social Cost of Cheap Pseudonyms," *Journal of Economics and Management Strategy* 10 (2), pp. 173–99.
49. There were three levels under discussion at Amazon: pragmatically anonymous reviews, where it is extremely simple to create a new username; traceable pseudonymous reviews, where reviewers can pick any username, but the name has to be coupled to an Amazon account set up with an authenticated credit card; or real-name reviews, where the username is generated by the authenticated credit card but the reviewers have the choice of using initials rather than a first name if they do not want to reveal their gender.
50. Rubin, Ben Fox, "Amazon Looks to Improve Customer-Reviews System with Machine Learning," *CNet,* June 19, 2015, http://www.cnet.com/news/amazon-updates-customer-reviews-with-new-machine-learning-platform.
51. Rubin, Ben Fox, "Amazon Sues Alleged Reviews-for-Pay Sites," *CNet,* April 9, 2015, http://www.cnet.com/news/amazon-sues-alleged-reviews-for-pay-sites. In some cases, it appears a company shipped an empty box or envelope by tracked mail to attain "Verified Purchase" status.
52. Rudder originally shared OKCupid analysis of racial preferences in the likelihood that a person's first contact would be answered on the OKTrends blog; he updated the figures in 2014 around the time he published a book based on site data. Rudder, Christian, "How Your Race Affects the Messages You Get," OKTrends, October 5, 2009, http://blog.okcupid.com/index.php/your-race-affects-whether-people-write-you-back; "Race and Attraction, 2009–2014," OKTrends, September 10, 2014, http://blog.okcupid.com/index.php/race-attraction-2009-2014. For those who date by the numbers, there are far more details in Rudder's *Dataclysm: Who We Are (When We Think No One's Looking)* (New York: Crown, 2014).
53. I served on Skout's board of directors from **[ADD DATES?]**.
54. n.a., "The Man Who Sued Google to Be Forgotten," Reuters, May 30, 2014, http://www.newsweek.com/man-who-sued-google-be-forgotten-252854.
55. These are among twenty-two examples highlighted by Google—all of them cases with obvious cause to approve or deny the request. Among the denials publicized by Google: a British "media professional" who asked for removal of links to four news articles about "embarrassing content he posted to the Internet." Google, "Transparency Report: European Privacy in Search," July 1, 2015, http://www.google.com/transparencyreport/removals/europeprivacy.
56. After some investigation, the *Guardian* reported that one article had been removed not because of the content of the journalism but because one of the *commenters* on the article had come to regret his words. The paper wouldn't allow comments to be deleted or edited—this, the paper said, would alter the context of later commenters' remarks. Ball, James, "EU's Right to Be Forgotten: Guardian Articles Have Been Hidden by Google," *Guardian,* July 2,

2014, http://www.theguardian.com/commentisfree/2014/jul/02/eu-right-to-be-forgotten-guardian-google. All of the articles could still be found via the U.S. version of the Google website.
57. McIntosh, Neil, "List of BBC Web Pages Which Have Been Removed from Google's Search Results," BBC Internet Blog, June 25, 2015, http://www.bbc.co.uk/blogs/internet/entries/1d765aa8-600b-4f32-b110-d02fbf7fd379.
58. Schechner, Sam, "Google Honors 'Right to Forget' Tantric Workshop," *Wall Street Journal*, July 18, 2014, http://www.wsj.com/articles/google-honors-right-to-forget-tantric-workshop-1405717183.
59. Although the revered Immanuel Kant was the leading philosophical voice at the time, it was Friedrich Karl von Savigny (inspired by Kant) who fully developed the idea of "freedom based upon… human autonomy, human will, and human personality" in German law. Eberle, Edward J., "The German Idea of Freedom," *Oregon Review of International Law* 10 (2008), p. 16. **[CK PP]**
60. Bloustein, Edward J., "Privacy as an Aspect of Human Dignity: An Answer to Dean Prosser," *New York University Law Review* 39 (962), p. 1964.
61. Schwartz, Paul M., and Karl-Nikolaus Peifer, "Prosser's *Privacy* and the German Right of Personality: Are Four Privacy Torts Better than One Unitary Concept?," *California Law Review* 98 (2010), pp. 1925–86.
62. Schwartz, Paul M., and Karl-Nikolaus Peifer, "Prosser's *Privacy* and the German Right of Personality: Are Four Privacy Torts Better than One Unitary Concept?," *California Law Review* 98 (2010), pp. 1931. **[CONVERT TO IBID.]**
63. Schwartz, Paul M., and Karl-Nikolaus Peifer, "Prosser's *Privacy* and the German Right of Personality: Are Four Privacy Torts Better than One Unitary Concept?," *California Law Review* 98 (2010), p. 1934. **[CONVERT TO IBID.]**