4

DIGITAL TRUTH SERUM

Everybody lies.

People lie about how many drinks they had on the way home. They lie about how often they go to the gym, how much those new shoes cost, whether they read that book. They call in sick when they're not. They say they'll be in touch when they won't. They say it's not about you when it is. They say they love you when they don't. They say they're happy while in the dumps. They say they like women when they really like men.

People lie to friends. They lie to bosses. They lie to kids. They lie to parents. They lie to doctors. They lie to husbands. They lie to wives. They lie to themselves.

And they damn sure lie to surveys.

Here's my brief survey for you:

Have you ever cheated on an exam? _____

Have you ever fantasized about killing someone? _____

Were you tempted to lie? Many people underreport embarrassing behaviors and thoughts on surveys. They want to look good, even though most surveys are anonymous. This is called social desirability bias.

An important paper in 1950 provided powerful evidence of how surveys can fall victim to such bias. Researchers collected data, from official sources, on the residents of Denver: what percentage of them voted, gave to charity, and owned a library card. They then surveyed the residents to see if the percentages would match. The results were, at the time, shocking. What the residents reported to the surveys was very different from the data the researchers had gathered. Even though nobody gave their names, people, in large numbers, exaggerated their voter registration status, voting behavior, and charitable giving.

	REPORTED ON SURVEY	OFFICIAL COUNT
Registered to vote	83%	69%
Voted in last presidential election	73%	61%
Voted in last mayoral election	63%	36%
Have a library card	20%	13%
Gave to a recent Community Chest charitable drive	67%	33%

Has anything changed in sixty-five years? In the age of the internet, not owning a library card is no longer embarrassing. But, while what's embarrassing or desirable may have changed, people's tendency to deceive pollsters remains strong.

A recent survey asked University of Maryland graduates

various questions about their college experience. The answers were compared to official records. People consistently gave wrong information, in ways that made them look good. Fewer than 2 percent reported that they graduated with lower than a 2.5 GPA. (In reality, about 11 percent did.) And 44 percent said they had donated to the university in the past year. (In reality, about 28 percent did.)

And it is certainly possible that lying played a role in the failure of the polls to predict Donald Trump's 2016 victory. Polls, on average, underestimated his support by about 2 percentage points. Some people may have been embarrassed to say they were planning to support him. Some may have claimed they were undecided when they were really going Trump's way all along.

Why do people misinform anonymous surveys? I asked Roger Tourangeau, a research professor emeritus at the University of Michigan and perhaps the world's foremost expert on social desirability bias. Our weakness for "white lies" is an important part of the problem, he explained. "About one-third of the time, people lie in real life," he suggests. "The habits carry over to surveys."

Then there's that odd habit we sometimes have of lying to ourselves. "There is an unwillingness to admit to yourself that, say, you were a screw-up as a student," says Tourangeau.

Lying to oneself may explain why so many people say they are above average. How big is this problem? More than 40 percent of one company's engineers said they are in the top 5 percent. More than 90 percent of college professors say they do above-average work. One-quarter of high school seniors think they are in the top 1 percent in their ability to get along with

other people. If you are deluding yourself, you can't be honest in a survey.

Another factor that plays into our lying to surveys is our strong desire to make a good impression on the stranger conducting the interview, if there is someone conducting the interview, that is. As Tourangeau puts it, "A person who looks like your favorite aunt walks in. . . . Do you want to tell your favorite aunt you used marijuana last month?" * Do you want to admit that you didn't give money to your good old alma mater?

For this reason, the more impersonal the conditions, the more honest people will be. For eliciting truthful answers, internet surveys are better than phone surveys, which are better than in-person surveys. People will admit more if they are alone than if others are in the room with them.

However, on sensitive topics, every survey method will elicit substantial misreporting. Tourangeau here used a word that is often thrown around by economists: "incentive." People have no incentive to tell surveys the truth.

* Another reason for lying is simply to mess with surveys. This is a huge problem for any research regarding teenagers, fundamentally complicating our ability to understand this age group. Researchers originally found a correlation between a teenager's being adopted and a variety of negative behaviors, such as using drugs, drinking alcohol, and skipping school. In subsequent research, they found this correlation was entirely explained by the 19 percent of self-reported adopted teenagers who weren't actually adopted. Follow-up research has found that a meaningful percent of teenagers tell surveys they are more than seven feet tall, weigh more than four hundred pounds, or have three children. One survey found 99 percent of students who reported having an artificial limb to academic researchers were kidding.

How, therefore, can we learn what our fellow humans are really thinking and doing?

In some instances, there are official data sources we can reference to get the truth. Even if people lie about their charitable donations, for example, we can get real numbers about giving in an area from the charities themselves. But when we are trying to learn about behaviors that are not tabulated in official records or we are trying to learn what people are thinking—their true beliefs, feelings, and desires—there is no other source of information except what people may deign to tell surveys. Until now, that is.

This is the second power of Big Data: certain online sources get people to admit things they would not admit anywhere else. They serve as a digital truth serum. Think of Google searches. Remember the conditions that make people more honest. Online? Check. Alone? Check. No person administering a survey? Check.

And there's another huge advantage that Google searches have in getting people to tell the truth: incentives. If you enjoy racist jokes, you have zero incentive to share that un-PC fact with a survey. You do, however, have an incentive to search for the best new racist jokes online. If you think you may be suffering from depression, you don't have an incentive to admit this to a survey. You do have an incentive to ask Google for symptoms and potential treatments.

Even if you are lying to yourself, Google may nevertheless know the truth. A couple of days before the election, you and some of your neighbors may legitimately think you will drive to a polling place and cast ballots. But, if you and they haven't searched for any information on how to vote or where to vote,

data scientists like me can figure out that turnout in your area will actually be low. Similarly, maybe you haven't admitted to yourself that you may suffer from depression, even as you're Googling about crying jags and difficulty getting out of bed. You would show up, however, in an area's depression-related searches that I analyzed earlier in this book.

Think of your own experience using Google. I am guessing you have upon occasion typed things into that search box that reveal a behavior or thought that you would hesitate to admit in polite company. In fact, the evidence is overwhelming that a large majority of Americans are telling Google some very personal things. Americans, for instance, search for "porn" more than they search for "weather." This is difficult, by the way, to reconcile with the survey data since only about 25 percent of men and 8 percent of women admit they watch pornography.

You may have also noticed a certain honesty in Google searches when looking at the way this search engine automatically tries to complete your queries. Its suggestions are based on the most common searches that other people have made. So auto-complete clues us in to what people are Googling. In fact, auto-complete can be a bit misleading. Google won't suggest certain words it deems inappropriate, such as "cock," "fuck," and "porn." This means auto-complete tells us that people's Google thoughts are less racy than they actually are. Even so, some sensitive stuff often still comes up.

If you type "Why is . . ." the first two Google auto-completes currently are "Why is the sky blue?" and "Why is there a leap day?" suggesting these are the two most common ways to complete this search. The third: "Why is my poop green?" And Google auto-complete can get disturbing. Today,

if you type in "Is it normal to want to . . ." the first suggestion is "kill." If you type in "Is it normal to want to kill . . ." the first suggestion is "my family."

Need more evidence that Google searches can give a different picture of the world than the one we usually see? Consider searches related to regrets around the decision to have or not to have children. Before deciding, some people fear they might make the wrong choice. And, almost always, the question is whether they will regret *not having* kids. People are seven times more likely to ask Google whether they will regret not having children than whether they will regret having children.

After making their decision—either to reproduce (or adopt) or not—people sometimes confess to Google that they rue their choice. This may come as something of a shock but post-decision, the numbers are reversed. Adults with children are 3.6 times more likely to tell Google they regret their decision than are adults without children.

One caveat that should be kept in mind throughout this chapter: Google can display a bias toward unseemly thoughts, thoughts people feel they can't discuss with anyone else. Nonetheless, if we are trying to uncover hidden thoughts, Google's ability to ferret them out can be useful. And the large disparity between regrets on having versus not having kids seems to be telling us that the unseemly thought in this case is a significant one.

Let's pause for a moment to consider what it even means to make a search such as "I regret having children." Google presents itself as a source from which we can seek information directly, on topics like the weather, who won last night's game, or when the Statue of Liberty was erected. But sometimes we

type our uncensored thoughts into Google, without much hope that it will be able to help us. In this case, the search window serves as a kind of confessional.

There are thousands of searches every year, for example, for "I hate cold weather," "People are annoying," and "I am sad." Of course, those thousands of Google searches for "I am sad" represent only a tiny of fraction of the hundreds of millions of people who feel sad in a given year. Searches expressing thoughts, rather than looking for information, my research has found, are only made by a small sample of everyone for whom that thought comes to mind. Similarly, my research suggests that the seven thousand searches by Americans every year for "I regret having children" represent a small sample of those who have had that thought.

Kids are obviously a huge joy for many, probably most, people. And, despite my mom's fear that "you and your stupid data analysis" are going to limit her number of grandchildren, this research has not changed my desire to have kids. But that unseemly regret is interesting—and another aspect of humanity that we tend not to see in the traditional datasets. Our culture is constantly flooding us with images of wonderful, happy families. Most people would never consider having children as something they might regret. But some do. They may admit this to no one—except Google.

THE TRUTH ABOUT SEX

How many American men are gay? This is a legendary question in sexuality research. Yet it has been among the toughest questions for social scientists to answer. Psychologists no longer believe Alfred Kinsey's famous estimate—based on surveys that oversampled prisoners and prostitutes—that 10 percent of American men are gay. Representative surveys now tell us about 2 to 3 percent are. But sexual preference has long been among the subjects upon which people have tended to lie. I think I can use Big Data to give a better answer to this question than we have ever had.

First, more on that survey data. Surveys tell us there are far more gay men in tolerant states than intolerant states. For example, according to a Gallup survey, the proportion of the population that is gay is almost twice as high in Rhode Island, the state with the highest support for gay marriage, than Mississippi, the state with the lowest support for gay marriage.

There are two likely explanations for this. First, gay men born in intolerant states may move to tolerant states. Second, gay men in intolerant states may not divulge that they are gay; they are even more likely to lie.

Some insight into explanation number one—gay mobility—can be gleaned from another Big Data source: Facebook, which allows users to list what gender they are interested in. About 2.5 percent of male Facebook users who list a gender of interest say they are interested in men; that corresponds roughly with what the surveys indicate. And Facebook too shows big differences in the gay population in states with high versus low tolerance: Facebook has the gay population more than twice as high in Rhode Island as in Mississippi.

Facebook also can provide information on how people move around. I was able to code the hometown of a sample of openly gay Facebook users. This allowed me to directly estimate how

improved student learning.
es together into a composite
ch measure adds something

rning that many Big Data
the holes that I showed up
er. Remember, he was the
sed lessons learned from a
f American Pharoah.
files and math with me,
veapon: Patty Murray.
intelligence and elite
vr. She also left New York
e than humans," Murray
itional in her approaches
horse agents, personally
k, checking for scars and
s.

er as they pick the final
ray sniffs out problems
data, despite being the
ver collected on horses,

the revelations of Big
ust throw data at any
te the need for all the
the millennia to un-
h other.

8

MO DATA, MO PROBLEMS? WHAT WE SHOULDN'T DO

Sometimes, the power of Big Data is so impressive it's scary. It raises ethical questions.

THE DANGER OF EMPOWERED CORPORATIONS

Recently, three economists—Oded Netzer and Alain Lemaire, both of Columbia, and Michal Herzenstein of the University of Delaware—looked for ways to predict the likelihood of whether a borrower would pay back a loan. The scholars utilized data from Prosper, a peer-to-peer lending site. Potential borrowers write a brief description of why they need a loan and why they are likely to make good on it, and potential lenders decide

whether to provide them the money. Overall, about 13 percent of borrowers defaulted on their loan.

It turns out the language that potential borrowers use is a strong predictor of their probability of paying back. And it is an important indicator even if you control for other relevant information lenders were able to obtain about those potential borrowers, including credit ratings and income.

Listed below are ten phrases the researchers found that are commonly used when applying for a loan. Five of them positively correlate with paying back the loan. Five of them negatively correlate with paying back the loan. In other words, five tend to be used by people you can trust, five by people you cannot. See if you can guess which are which.

God	lower interest rate	after-tax
promise	will pay	hospital
debt-free	graduate	
minimum payment	thank you	

You might think—or at least hope—that a polite, openly religious person who gives his word would be among the most likely to pay back a loan. But in fact this is not the case. This type of person, the data shows, is less likely than average to make good on their debt.

Here are the phrases grouped by the likelihood of paying back.

TERMS USED IN LOAN APPLICATIONS BY PEOPLE MOST LIKELY TO PAY BACK

debt-free	after-tax	graduate
lower interest rate	minimum payment	

TERMS USED IN LOAN APPLICATIONS BY PEOPLE MOST LIKELY TO DEFAULT

God	will pay	hospital
promise	thank you	

Before we discuss the ethical implications of this study, let's think through, with the help of the study's authors, what it reveals about people. What should we make of the words in the different categories?

First, let's consider the language that suggests someone is more likely to make their loan payments. Phrases such as "lower interest rate" or "after-tax" indicate a certain level of financial sophistication on the borrower's part, so it's perhaps not surprising they correlate with someone more likely to pay their loan back. In addition, if he or she talks about positive achievements such as being a college "graduate" and being "debt-free," he or she is also likely to pay their loans.

Now let's consider language that suggests someone is unlikely to pay their loans. Generally, if someone tells you he will pay you back, he will not pay you back. The more assertive the promise, the more likely he will break it. If someone writes "I promise I will pay back, so help me God," he is among the least likely to pay you back. Appealing to your mercy—explaining that he needs the money because he has a relative in the "hospital"—also means he is unlikely to pay you back. In fact, mentioning any family member—a husband, wife, son, daughter, mother, or father—is a sign someone will not be paying back. Another word that indicates default is "explain," meaning if people are trying to explain why they are going to be able to pay back a loan, they likely won't.

The authors did not have a theory for why thanking people is evidence of likely default.

In sum, according to these researchers, giving a detailed plan of how he can make his payments and mentioning commitments he has kept in the past are evidence someone will pay back a loan. Making promises and appealing to your mercy is a clear sign someone will go into default. Regardless of the reasons—or what it tells us about human nature that making promises is a sure sign someone will, in actuality, not do something—the scholars found the test was an extremely valuable piece of information in predicting default. Someone who mentions God was 2.2 times more likely to default. This was among the single highest indicators that someone would not pay back.

But the authors also believe their study raises ethical questions. While this was just an academic study, some companies do report that they utilize online data in approving loans. Is this acceptable? Do we want to live in a world in which companies use the words we write to predict whether we will pay back a loan? It is, at a minimum, creepy—and, quite possibly, scary.

A consumer looking for a loan in the near future might have to worry about not merely her financial history but also her online activity. And she may be judged on factors that seem absurd—whether she uses the phrase "Thank you" or invokes "God," for example. Further, what about a woman who legitimately needs to help her sister in a hospital and will most certainly pay back her loan afterward? It seems awful to punish her because, on average, people claiming to need help for medical bills have often been proven to be lying. A world functioning this way starts to look awfully dystopian.

This is the ethical question: Do corporations have the right to judge our fitness for their services based on abstract but statistically predictive criteria not directly related to those services?

Leaving behind the world of finance, let's look at the larger implications on, for example, hiring practices. Employers are increasingly scouring social media when considering job candidates. That may not raise ethical questions if they're looking for evidence of bad-mouthing previous employers or revealing previous employers' secrets. There may even be some justification for refusing to hire someone whose Facebook or Instagram posts suggest excessive alcohol use. But what if they find a seemingly harmless indicator that correlates with something they care about?

Researchers at Cambridge University and Microsoft gave fifty-eight thousand U.S. Facebook users a variety of tests about their personality and intelligence. They found that Facebook likes are frequently correlated with IQ, extraversion, and conscientiousness. For example, people who like Mozart, thunderstorms, and curly fries on Facebook tend to have higher IQs. People who like Harley-Davidson motorcycles, the country music group Lady Antebellum, or the page "I Love Being a Mom" tend to have lower IQs. Some of these correlations may be due to the curse of dimensionality. If you test enough things, some will randomly correlate. But some interests may legitimately correlate with IQ.

Nonetheless, it would seem unfair if a smart person who happens to like Harleys couldn't get a job commensurate with his skills because he was, without realizing it, signaling low intelligence.

In fairness, this is not an entirely new problem. People have long been judged by factors not directly related to job performance—the firmness of their handshakes, the neatness of their dress. But a danger of the data revolution is that, as more of our life is quantified, these proxy judgments can get more esoteric yet more intrusive. Better prediction can lead to subtler and more nefarious discrimination.

Better data can also lead to another form of discrimination, what economists call price discrimination. Businesses are often trying to figure out what price they should charge for goods or services. Ideally they want to charge customers the maximum they are willing to pay. This way, they will extract the maximum possible profit.

Most businesses usually end up picking one price that everyone pays. But sometimes they are aware that the members of a certain group will, on average, pay more. This is why movie theaters charge more to middle-aged customers—at the height of their earning power—than to students or senior citizens and why airlines often charge more to last-minute purchasers. They price discriminate.

Big Data may allow businesses to get substantially better at learning what customers are willing to pay—and thus gouging certain groups of people. Optimal Decisions Group was a pioneer in using data science to predict how much consumers are willing to pay for insurance. How did they do it? They used a methodology that we have previously discussed in this book. They found prior customers most similar to those currently looking to buy insurance—and saw how high a premium they were willing to take on. In other words, they ran a doppelganger search. A doppelganger search is entertaining if it helps

us predict whether a baseball player will return to his former greatness. A doppelganger search is great if it helps us cure someone's disease. But if a doppelganger search helps a corporation extract every last penny from you? That's not so cool. My spendthrift brother would have a right to complain if he got charged more online than tightwad me.

Gambling is one area in which the ability to zoom in on customers is potentially dangerous. Big casinos are using something like a doppelganger search to better understand their consumers. Their goal? To extract the maximum possible profit—to make sure more of your money goes into their coffers.

Here's how it works. Every gambler, casinos believe, has a "pain point." This is the amount of losses that will sufficiently frighten her so that she leaves your casino for an extended period of time. Suppose, for example, that Helen's "pain point" is $3,000. This means if she loses $3,000, you've lost a customer, perhaps for weeks or months. If Helen loses $2,999, she won't be happy. Who, after all, likes to lose money? But she won't be so demoralized that she won't come back tomorrow night.

Imagine for a moment that you are managing a casino. And imagine that Helen has shown up to play the slot machines. What is the optimal outcome? Clearly, you want Helen to get as close as possible to her "pain point" without crossing it. You want her to lose $2,999, enough that you make big profits but not so much that she won't come back to play again soon.

How can you do this? Well, there are ways to get Helen to stop playing once she has lost a certain amount. You can offer her free meals, for example. Make the offer enticing enough, and she will leave the slots for the food.

But there's one big challenge with this approach. How do you know Helen's "pain point"? The problem is, people have different "pain points." For Helen, it's $3,000. For John, it might be $2,000. For Ben, it might be $26,000. If you convince Helen to stop gambling when she lost $2,000, you left profits on the table. If you wait too long—after she has lost $3,000—you have lost her for a while. Further, Helen might not want to tell you her pain point. She may not even know what it is herself.

So what do you do? If you have made it this far in the book, you can probably guess the answer. You utilize data science. You learn everything you can about a number of your customers—their age, gender, zip code, and gambling behavior. And, from that gambling behavior—their winnings, losings, comings, and goings—you estimate their "pain point."

You gather all the information you know about Helen and find gamblers who are similar to her—her doppelgangers, more or less. Then you figure out how much pain they can withstand. It's probably the same amount as Helen. Indeed, this is what the casino Harrah's does, utilizing a Big Data warehouse firm, Terabyte, to assist them.

Scott Gnau, general manager of Terabyte, explains, in the excellent book *Super Crunchers*, what casino managers do when they see a regular customer nearing their pain point: "They come out and say, 'I see you're having a rough day. I know you like our steakhouse. Here, I'd like you to take your wife to dinner on us right now.'"

This might seem the height of generosity: a free steak dinner. But really it's self-serving. The casino is just trying to get customers to quit before they lose so much that they'll leave for

an extended period of time. In other words, management is using sophisticated data analysis to try to extract as much money from customers, over the long term, as it can.

We have a right to fear that better and better use of online data will give casinos, insurance companies, lenders, and other corporate entities too much power over us.

On the other hand, Big Data has also been enabling consumers to score some blows against businesses that overcharge them or deliver shoddy products.

One important weapon is sites, such as Yelp, that publish reviews of restaurants and other services. A recent study by economist Michael Luca, of Harvard, has shown the extent to which businesses are at the mercy of Yelp reviews. Comparing those reviews to sales data in the state of Washington, he found that one fewer star on Yelp will make a restaurant's revenues drop 5 to 9 percent.

Consumers are also aided in their struggles with business by comparison shopping sites—like Kayak and Booking.com. As discussed in *Freakonomics*, when an internet site began reporting the prices different companies were charging for term life insurance, these prices fell dramatically. If an insurance company was overcharging, customers would know it and use someone else. The total savings to consumers? One billion dollars per year.

Data on the internet, in other words, can tell businesses which customers to avoid and which they can exploit. It can also tell customers the businesses they should avoid and who is trying to exploit them. Big Data to date has helped both sides in the struggle between consumers and corporations. We have to make sure it remains a fair fight.

THE DANGER OF EMPOWERED GOVERNMENTS

When her ex-boyfriend showed up at a birthday party, Adriana Donato knew he was upset. She knew that he was mad. She knew that he had struggled with depression. As he invited her for a drive, there was one thing Donato, a twenty-year-old zoology student, did not know. She did not know her ex-boyfriend, twenty-two-year-old James Stoneham, had spent the previous three weeks searching for information on how to murder somebody and about murder law, mixed in with the occasional search about Donato.

If she had known this, presumably she would not have gotten in the car. Presumably, she would not have been stabbed to death that evening.

In the movie *Minority Report*, psychics collaborate with police departments to stop crimes before they happen. Should Big Data be made available to police departments to stop crimes before they happen? Should Donato have at least been warned about her ex-boyfriend's foreboding searches? Should the police have interrogated Stoneham?

First, it must be acknowledged that there is growing evidence that Google searches related to criminal activity do correlate with criminal activity. Christine Ma-Kellams, Flora Or, Ji Hyun Baek, and Ichiro Kawachi have shown that Google searches related to suicide correlate strongly with state-level suicide rates. In addition, Evan Soltas and I have shown that weekly Islamophobic searches—such as "I hate Muslims" or "kill Muslims"—correlate with anti-Muslim hate crimes that

week. If more people are making searches saying they want to do something, more people are going to do that thing.

So what should we do with this information? One simple, fairly uncontroversial idea: we can utilize the area-level data to allocate resources. If a city has a huge rise in suicide-related searches, we can up the suicide awareness in this city. The city government or nonprofits might run commercials explaining where people can get help, for example. Similarly, if a city has a huge rise in searches for "kill Muslims," police departments might be wise to change how they patrol the streets. They might dispatch more officers to protect the local mosque, for example.

But one step we should be very reluctant to take: going after individuals before any crime has been committed. This seems, to begin with, an invasion of privacy. There is a large ethical leap from the government having the search data of thousands or hundreds of thousands of people to the police department having the search data of an individual. There is a large ethical leap from protecting a local mosque to ransacking someone's house. There is a large ethical leap from advertising suicide prevention to locking someone up in a mental hospital against his will.

The reason to be extremely cautious using individual-level data, however, goes beyond even ethics. There is a data reason as well. It is a large leap for data science to go from trying to predict the actions of a city to trying to predict the actions of an individual.

Let's return to suicide for a moment. Every month, there are about 3.5 million Google searches in the United States re-

lated to suicide, with the majority of them suggesting suicidal ideation—searches such as "suicidal," "commit suicide," and "how to suicide." In other words, every month, there is more than one search related to suicide for every one hundred Americans. This brings to mind a quote from the philosopher Friedrich Nietzsche: "The thought of suicide is a great consolation: by means of it one gets through many a dark night." Google search data shows how true that is, how common the thought of suicide is. However, every month, there are fewer than four thousand suicides in the United States. Suicidal ideation is incredibly common. Suicide is not. So it wouldn't make a lot of sense for cops to be showing up at the door of everyone who has ever made some online noise about wanting to blow their brains out—if for no other reason than that the police wouldn't have time for anything else.

Or consider those incredibly vicious Islamophobic searches. In 2015, there were roughly 12,000 searches in the United States for "kill Muslims." There were 12 murders of Muslims reported as hate crimes. Clearly, the vast majority of people who make this terrifying search do not go through with the corresponding act.

There is some math that explains the difference between predicting the behavior of an individual and predicting the behavior in a city. Here's a simple thought experiment. Suppose there are one million people in a city and one mosque. Suppose, if someone does not search for "kill Muslims," there is only a 1-in-100,000,000 chance that he will attack a mosque. Suppose if someone does search for "kill Muslims," this chance rises sharply, to 1 in 10,000. Suppose Islamophobia has skyrocketed and searches for "kill Muslims" have risen from 100 to 1,000.

In this situation, math shows that the chances of a mosque being attacked has risen about fivefold, from about 2 percent to 10 percent. But the chances of an individual who searched for "kill Muslims" actually attacking a mosque remains only 1 in 10,000.

The proper response in this situation is not to jail all the people who searched for "kill Muslims." Nor is it to visit their houses. There is a tiny chance that any one of these people in particular will commit a crime. The proper response, however, would be to protect that mosque, which now has a 10 percent chance of being attacked.

Clearly, many horrific searches never lead to horrible actions.

That said, it is at least theoretically possible that there are some classes of searches that suggest a reasonably high probability of a horrible follow-through. It is at least theoretically possible, for example, that data scientists could in the future build a model that could have found that Stoneham's searches related to Donato were significant cause for concern.

In 2014, there were about 6,000 searches for the exact phrase "how to kill your girlfriend" and 400 murders of girlfriends. If all of these murderers had made this exact search beforehand, that would mean 1 in 15 people who searched "how to kill your girlfriend" went through with it. Of course, many, probably most, people who murdered their girlfriends did not make this exact search. This would mean the true probability that this particular search led to murder is lower, probably a lot lower.

But if data scientists could build a model that showed that the threat against a particular individual was, say, 1 in 100,

we might want to do something with that information. At the least, the person under threat might have the right to be informed that there is a 1-in-100 chance she will be murdered by a particular person.

Overall, however, we have to be very cautious using search data to predict crimes at an individual level. The data clearly tells us that there are many, many horrifying searches that rarely lead to horrible actions. And there has been, as of yet, no proof that the government can predict a particular horrible action, with high probability, just from examining these searches. So we have to be really cautious about allowing the government to intervene at the individual level based on search data. This is not just for ethical or legal reasons. It's also, at least for now, for data science reasons.

CONCLUSION

HOW MANY PEOPLE FINISH BOOKS?

After signing my book contract, I had a clear vision of how the book should be structured. Near the start, you may recall, I described a scene at my family's Thanksgiving table. My family members debated my sanity and tried to figure out why I, at thirty-three, couldn't seem to find the right girl.

The conclusion to this book, then, practically wrote itself. I would meet and marry the girl. Better still, I would use Big Data to meet the right girl. Perhaps I could weave in tidbits from the courting process throughout. Then the story would all come together in the conclusion, which would describe my wedding day and double as a love letter to my new wife.

Unfortunately, life didn't match my vision. Locking myself in my apartment and avoiding the world while writing a book probably didn't help my romantic life. And I, alas, still need to find a wife. More important, I needed a new conclusion.