Cathy O'Neil
Weapons of Math Destruction
New York: Crown
2016

INTRODUCTION

When I was a little girl, I used to gaze at the traffic out the car window and study the numbers on license plates. I would reduce each one to its basic elements—the prime numbers that made it up. $45 = 3 \times 3 \times 5$. That's called factoring, and it was my favorite investigative pastime. As a budding math nerd, I was especially intrigued by the primes.

My love for math eventually became a passion. I went to math camp when I was fourteen and came home clutching a Rubik's Cube to my chest. Math provided a neat refuge from the messiness of the real world. It marched forward, its field of knowledge expanding relentlessly, proof by proof. And I could add to it. I majored in math in college and went on to get my PhD. My thesis was on algebraic number theory, a field with roots in all that

factoring I did as a child. Eventually, I became a tenure-track professor at Barnard, which had a combined math department with Columbia University.

And then I made a big change. I quit my job and went to work as a quant for D. E. Shaw, a leading hedge fund. In leaving academia for finance, I carried mathematics from abstract theory into practice. The operations we performed on numbers translated into trillions of dollars sloshing from one account to another. At first I was excited and amazed by working in this new laboratory, the global economy. But in the autumn of 2008, after I'd been there for a bit more than a year, it came crashing down.

The crash made it all too clear that mathematics, once my refuge, was not only deeply entangled in the world's problems but also fueling many of them. The housing crisis, the collapse of major financial institutions, the rise of unemployment—all had been aided and abetted by mathematicians wielding magic formulas. What's more, thanks to the extraordinary powers that I loved so much, math was able to combine with technology to multiply the chaos and misfortune, adding efficiency and scale to systems that I now recognized as flawed.

If we had been clear-headed, we all would have taken a step back at this point to figure out how math had been misused and how we could prevent a similar catastrophe in the future. But instead, in the wake of the crisis, new mathematical techniques were hotter than ever, and expanding into still more domains. They churned 24/7 through petabytes of information, much of it scraped from social media or e-commerce websites. And increasingly they focused not on the movements of global financial markets but on human beings, on us. Mathematicians and statisticians were studying our desires, movements, and spending power. They were predicting our trustworthiness and calculating our potential as students, workers, lovers, criminals.

This was the Big Data economy, and it promised spectacular gains. A computer program could speed through thousands of résumés or loan applications in a second or two and sort them into neat lists, with the most promising candidates on top. This not only saved time but also was marketed as fair and objective. After all, it didn't involve prejudiced humans digging through reams of paper, just machines processing cold numbers. By 2010 or so, mathematics was asserting itself as never before in human affairs, and the public largely welcomed it.

Yet I saw trouble. The math-powered applications powering the data economy were based on choices made by fallible human beings. Some of these choices were no doubt made with the best intentions. Nevertheless, many of these models encoded human prejudice, misunderstanding, and bias into the software systems that increasingly managed our lives. Like gods, these mathematical models were opaque, their workings invisible to all but the highest priests in their domain: mathematicians and computer scientists. Their verdicts, even when wrong or harmful, were beyond dispute or appeal. And they tended to punish the poor and the oppressed in our society, while making the rich richer.

I came up with a name for these harmful kinds of models: Weapons of Math Destruction, or WMDs for short. I'll walk you through an example, pointing out its destructive characteristics along the way.

As often happens, this case started with a laudable goal. In 2007, Washington, D.C.'s new mayor, Adrian Fenty, was determined to turn around the city's underperforming schools. He had his work cut out for him: at the time, barely one out of every two high school students was surviving to graduation after ninth grade, and only 8 percent of eighth graders were performing at grade level in math. Fenty hired an education reformer named Michelle Rhee to fill a powerful new post, chancellor of Washington's schools.

The going theory was that the students weren't learning enough because their teachers weren't doing a good job. So in 2009, Rhee implemented a plan to weed out the low-performing teachers. This is the trend in troubled school districts around the country, and from a systems engineering perspective the thinking makes perfect sense: Evaluate the teachers. Get rid of the worst ones, and place the best ones where they can do the most good. In the language of data scientists, this "optimizes" the school system, presumably ensuring better results for the kids. Except for "bad" teachers, who could argue with that? Rhee developed a teacher assessment tool called IMPACT, and at the end of the 2009–10 school year the district fired all the teachers whose scores put them in the bottom 2 percent. At the end of the following year, another 5 percent, or 206 teachers, were booted out.

Sarah Wysocki, a fifth-grade teacher, didn't seem to have any reason to worry. She had been at MacFarland Middle School for only two years but was already getting excellent reviews from her principal and her students' parents. One evaluation praised her attentiveness to the children; another called her "one of the best teachers I've ever come into contact with."

Yet at the end of the 2010–11 school year, Wysocki received a miserable score on her IMPACT evaluation. Her problem was a new scoring system known as value-added modeling, which purported to measure her effectiveness in teaching math and language skills. That score, generated by an algorithm, represented half of her overall evaluation, and it outweighed the positive reviews from school administrators and the community. This left the district with no choice but to fire her, along with 205 other teachers who had IMPACT scores below the minimal threshold.

This didn't seem to be a witch hunt or a settling of scores. Indeed, there's a logic to the school district's approach. Administrators, after all, could be friends with terrible teachers. They could admire their style or their apparent dedication. Bad teachers can *seem* good. So Washington, like many other school systems, would minimize this human bias and pay more attention to scores based on hard results: achievement scores in math and reading. The numbers would speak clearly, district officials promised. They would be more fair.

Wysocki, of course, felt the numbers were horribly unfair, and she wanted to know where they came from. "I don't think anyone understood them," she later told me. How could a good teacher get such dismal scores? What was the value-added model measuring?

Well, she learned, it was complicated. The district had hired a consultancy, Princeton-based Mathematica Policy Research, to come up with the evaluation system. Mathematica's challenge was to measure the educational progress of the students in the district and then to calculate how much of their advance or decline could be attributed to their teachers. This wasn't easy, of course. The researchers knew that many variables, from students' socioeconomic backgrounds to the effects of learning disabilities, could affect student outcomes. The algorithms had to make allowances for such differences, which was one reason they were so complex.

Indeed, attempting to reduce human behavior, performance, and potential to algorithms is no easy job. To understand what Mathematica was up against, picture a ten-year-old girl living in a poor neighborhood in southeastern Washington, D.C. At the end of one school year, she takes her fifth-grade standardized test. Then life goes on. She may have family issues or money problems. Maybe she's moving from one house to another or worried about an older brother who's in trouble with the law. Maybe she's unhappy about her weight or frightened by a bully at school. In

any case, the following year she takes another standardized test, this one designed for sixth graders.

If you compare the results of the tests, the scores should stay stable, or hopefully, jump up. But if her results sink, it's easy to calculate the gap between her performance and that of the successful students.

But how much of that gap is due to her teacher? It's hard to know, and Mathematica's models have only a few numbers to compare. At Big Data companies like Google, by contrast, researchers run constant tests and monitor thousands of variables. They can change the font on a single advertisement from blue to red, serve each version to ten million people, and keep track of which one gets more clicks. They use this feedback to hone their algorithms and fine-tune their operation. While I have plenty of issues with Google, which we'll get to, this type of testing is an effective use of statistics.

Attempting to calculate the impact that one person may have on another over the course of a school year is much more complex. "There are so many factors that go into learning and teaching that it would be very difficult to measure them all," Wysocki says. What's more, attempting to score a teacher's effectiveness by analyzing the test results of only twenty-five or thirty students is statistically unsound, even laughable. The numbers are far too small given all the things that could go wrong. Indeed, if we were to analyze teachers with the statistical rigor of a search engine, we'd have to test them on thousands or even millions of randomly selected students. Statisticians count on large numbers to balance out exceptions and anomalies. (And WMDs, as we'll see, often punish individuals who happen to *be* the exception.)

Equally important, statistical systems require feedback—something to tell them when they're off track. Statisticians use errors to train their models and make them smarter. If Amazon.com,

through a faulty correlation, started recommending lawn care books to teenage girls, the clicks would plummet, and the algorithm would be tweaked until it got it right. Without feedback, however, a statistical engine can continue spinning out faulty and damaging analysis while never learning from its mistakes.

Many of the WMDs I'll be discussing in this book, including the Washington school district's value-added model, behave like that. They define their own reality and use it to justify their results. This type of model is self-perpetuating, highly destructive—and very common.

When Mathematica's scoring system tags Sarah Wysocki and 205 other teachers as failures, the district fires them. But how does it ever learn if it was right? It doesn't. The system itself has determined that they were failures, and that is how they are viewed. Two hundred and six "bad" teachers are gone. That fact alone appears to demonstrate how effective the value-added model is. It is cleansing the district of underperforming teachers. Instead of searching for the truth, the score comes to embody it.

This is one example of a WMD feedback loop. We'll see many of them throughout this book. Employers, for example, are increasingly using credit scores to evaluate potential hires. Those who pay their bills promptly, the thinking goes, are more likely to show up to work on time and follow the rules. In fact, there are plenty of responsible people and good workers who suffer misfortune and see their credit scores fall. But the belief that bad credit correlates with bad job performance leaves those with low scores less likely to find work. Joblessness pushes them toward poverty, which further worsens their scores, making it even harder for them to land a job. It's a downward spiral. And employers never learn how many good employees they've missed out on by focusing on credit scores. In WMDs, many poisonous assumptions are camouflaged by math and go largely untested and unquestioned.

This underscores another common feature of WMDs. They tend to punish the poor. This is, in part, because they are engineered to evaluate large numbers of people. They specialize in bulk, and they're cheap. That's part of their appeal. The wealthy, by contrast, often benefit from personal input. A white-shoe law firm or an exclusive prep school will lean far more on recommendations and face-to-face interviews than will a fast-food chain or a cash-strapped urban school district. The privileged, we'll see time and again, are processed more by people, the masses by machines.

Wysocki's inability to find someone who could explain her appalling score, too, is telling. Verdicts from WMDs land like dictates from the algorithmic gods. The model itself is a black box, its contents a fiercely guarded corporate secret. This allows consultants like Mathematica to charge more, but it serves another purpose as well: if the people being evaluated are kept in the dark, the thinking goes, they'll be less likely to attempt to game the system. Instead, they'll simply have to work hard, follow the rules, and pray that the model registers and appreciates their efforts. But if the details are hidden, it's also harder to question the score or to protest against it.

For years, Washington teachers complained about the arbitrary scores and clamored for details on what went into them. It's an algorithm, they were told. It's very complex. This discouraged many from pressing further. Many people, unfortunately, are intimidated by math. But a math teacher named Sarah Bax continued to push the district administrator, a former colleague named Jason Kamras, for details. After a back-and-forth that extended for months, Kamras told her to wait for an upcoming technical report. Bax responded: "How do you justify evaluating people by a measure for which you are unable to provide explanation?" But that's the nature of WMDs. The analysis is outsourced to

coders and statisticians. And as a rule, they let the machines do the talking.

Even so, Sarah Wysocki was well aware that her students' standardized test scores counted heavily in the formula. And here she had some suspicions. Before starting what would be her final year at MacFarland Middle School, she had been pleased to see that her incoming fifth graders had scored surprisingly well on their year-end tests. At Barnard Elementary School, where many of Sarah's students came from, 29 percent of the students were ranked at an "advanced reading level." This was five times the average in the school district.

Yet when classes started she saw that many of her students struggled to read even simple sentences. Much later, investigations by the *Washington Post* and *USA Today* revealed a high level of erasures on the standardized tests at forty-one schools in the district, including Barnard. A high rate of corrected answers points to a greater likelihood of cheating. In some of the schools, as many as 70 percent of the classrooms were suspected.

What does this have to do with WMDs? A couple of things. First, teacher evaluation algorithms are a powerful tool for behavioral modification. That's their purpose, and in the Washington schools they featured both a stick and a carrot. Teachers knew that if their students stumbled on the test their own jobs were at risk. This gave teachers a strong motivation to ensure their students passed, especially as the Great Recession battered the labor market. At the same time, if their students outperformed their peers, teachers and administrators could receive bonuses of up to $8,000. If you add those powerful incentives to the evidence in the case—the high number of erasures and the abnormally high test scores—there were grounds for suspicion that fourth-grade teachers, bowing either to fear or to greed, had corrected their students' exams.

It is conceivable, then, that Sarah Wysocki's fifth-grade students started the school year with artificially inflated scores. If so, their results the following year would make it appear that they'd lost ground in fifth grade—and that their teacher was an underperformer. Wysocki was convinced that this was what had happened to her. That explanation would fit with the observations from parents, colleagues, and her principal that she was indeed a good teacher. It would clear up the confusion. Sarah Wysocki had a strong case to make.

But you cannot appeal to a WMD. That's part of their fearsome power. They do not listen. Nor do they bend. They're deaf not only to charm, threats, and cajoling but also to logic—even when there is good reason to question the data that feeds their conclusions. Yes, if it becomes clear that automated systems are screwing up on an embarrassing and systematic basis, programmers will go back in and tweak the algorithms. But for the most part, the programs deliver unflinching verdicts, and the human beings employing them can only shrug, as if to say, "Hey, what can you do?"

And that is precisely the response Sarah Wysocki finally got from the school district. Jason Kamras later told the *Washington Post* that the erasures were "suggestive" and that the numbers might have been wrong in her fifth-grade class. But the evidence was not conclusive. He said she had been treated fairly.

Do you see the paradox? An algorithm processes a slew of statistics and comes up with a probability that a certain person *might* be a bad hire, a risky borrower, a terrorist, or a miserable teacher. That probability is distilled into a score, which can turn someone's life upside down. And yet when the person fights back, "suggestive" countervailing evidence simply won't cut it. The case must be ironclad. The human victims of WMDs, we'll see time and again, are held to a far higher standard of evidence than the algorithms themselves.

After the shock of her firing, Sarah Wysocki was out of a job for only a few days. She had plenty of people, including her principal, to vouch for her as a teacher, and she promptly landed a position at a school in an affluent district in northern Virginia. So thanks to a highly questionable model, a poor school lost a good teacher, and a rich school, which didn't fire people on the basis of their students' scores, gained one.

• • •

Following the housing crash, I woke up to the proliferation of WMDs in banking and to the danger they posed to our economy. In early 2011 I quit my job at the hedge fund. Later, after rebranding myself as a data scientist, I joined an e-commerce start-up. From that vantage point, I could see that legions of other WMDs were churning away in every conceivable industry, many of them exacerbating inequality and punishing the poor. They were at the heart of the raging data economy.

To spread the word about WMDs, I launched a blog, MathBabe. My goal was to mobilize fellow mathematicians against the use of sloppy statistics and biased models that created their own toxic feedback loops. Data specialists, in particular, were drawn to the blog, and they alerted me to the spread of WMDs in new domains. But in mid-2011, when Occupy Wall Street sprang to life in Lower Manhattan, I saw that we had work to do among the broader public. Thousands had gathered to demand economic justice and accountability. And yet when I heard interviews with the Occupiers, they often seemed ignorant of basic issues related to finance. They clearly hadn't been reading my blog. (I should add, though, that you don't need to understand all the details of a system to know that it has failed.)

I could either criticize them or join them, I realized, so I joined them. Soon I was facilitating weekly meetings of the Alternative

Banking Group at Columbia University, where we discussed financial reform. Through this process, I came to see that my two ventures outside academia, one in finance, the other in data science, had provided me with fabulous access to the technology and culture powering WMDs.

Ill-conceived mathematical models now micromanage the economy, from advertising to prisons. These WMDs have many of the same characteristics as the value-added model that derailed Sarah Wysocki's career in Washington's public schools. They're opaque, unquestioned, and unaccountable, and they operate at a scale to sort, target, or "optimize" millions of people. By confusing their findings with on-the-ground reality, most of them create pernicious WMD feedback loops.

But there's one important distinction between a school district's value-added model and, say, a WMD that scouts out prospects for extortionate payday loans. They have different payoffs. For the school district, the payoff is a kind of political currency, a sense that problems are being fixed. But for businesses it's just the standard currency: money. For many of the businesses running these rogue algorithms, the money pouring in seems to prove that their models are working. Look at it through their eyes and it makes sense. When they're building statistical systems to find customers or manipulate desperate borrowers, growing revenue appears to show that they're on the right track. The software is doing its job. The trouble is that profits end up serving as a stand-in, or proxy, for truth. We'll see this dangerous confusion crop up again and again.

This happens because data scientists all too often lose sight of the folks on the receiving end of the transaction. They certainly understand that a data-crunching program is bound to misinterpret people a certain percentage of the time, putting them in the wrong groups and denying them a job or a chance at their dream

house. But as a rule, the people running the WMDs don't dwell on those errors. Their feedback is money, which is also their incentive. Their systems are engineered to gobble up more data and fine-tune their analytics so that more money will pour in. Investors, of course, feast on these returns and shower WMD companies with more money.

And the victims? Well, an internal data scientist might say, no statistical system can be *perfect*. Those folks are collateral damage. And often, like Sarah Wysocki, they are deemed unworthy and expendable. Forget about them for a minute, they might say, and focus on all the people who get helpful suggestions from recommendation engines or who find music they love on Pandora, the ideal job on LinkedIn, or perhaps the love of their life on Match.com. Think of the astounding scale, and ignore the imperfections.

Big Data has plenty of evangelists, but I'm not one of them. This book will focus sharply in the other direction, on the damage inflicted by WMDs and the injustice they perpetuate. We will explore harmful examples that affect people at critical life moments: going to college, borrowing money, getting sentenced to prison, or finding and holding a job. All of these life domains are increasingly controlled by secret models wielding arbitrary punishments.

Welcome to the dark side of Big Data.

1

BOMB PARTS

What Is a Model?

It was a hot August afternoon in 1946. Lou Boudreau, the player-manager of the Cleveland Indians, was having a miserable day. In the first game of a doubleheader, Ted Williams had almost single-handedly annihilated his team. Williams, perhaps the game's greatest hitter at the time, had smashed three home runs and driven home eight. The Indians ended up losing 11 to 10.

Boudreau had to take action. So when Williams came up for the first time in the second game, players on the Indians' side started moving around. Boudreau, the shortstop, jogged over to where the second baseman would usually stand, and the second baseman backed into short right field. The third baseman moved

to his left, into the shortstop's hole. It was clear that Boudreau, perhaps out of desperation, was shifting the entire orientation of his defense in an attempt to turn Ted Williams's hits into outs.

In other words, he was thinking like a data scientist. He had analyzed crude data, most of it observational: Ted Williams *usually* hit the ball to right field. Then he adjusted. And it worked. Fielders caught more of Williams's blistering line drives than before (though they could do nothing about the home runs sailing over their heads).

If you go to a major league baseball game today, you'll see that defenses now treat nearly every player like Ted Williams. While Boudreau merely observed where Williams usually hit the ball, managers now know precisely where every player has hit every ball over the last week, over the last month, throughout his career, against left-handers, when he has two strikes, and so on. Using this historical data, they analyze their current situation and calculate the positioning that is associated with the highest probability of success. And that sometimes involves moving players far across the field.

Shifting defenses is only one piece of a much larger question: What steps can baseball teams take to maximize the probability that they'll win? In their hunt for answers, baseball statisticians have scrutinized every variable they can quantify and attached it to a value. How much more is a double worth than a single? When, if ever, is it worth it to bunt a runner from first to second base?

The answers to all of these questions are blended and combined into mathematical models of their sport. These are parallel universes of the baseball world, each a complex tapestry of probabilities. They include every measurable relationship among every one of the sport's components, from walks to home runs to the players themselves. The purpose of the model is to run different

scenarios at every juncture, looking for the optimal combinations. If the Yankees bring in a right-handed pitcher to face Angels slugger Mike Trout, as compared to leaving in the current pitcher, how much more likely are they to get him out? And how will that affect their overall odds of winning?

Baseball is an ideal home for predictive mathematical modeling. As Michael Lewis wrote in his 2003 bestseller, *Moneyball*, the sport has attracted data nerds throughout its history. In decades past, fans would pore over the stats on the back of baseball cards, analyzing Carl Yastrzemski's home run patterns or comparing Roger Clemens's and Dwight Gooden's strikeout totals. But starting in the 1980s, serious statisticians started to investigate what these figures, along with an avalanche of new ones, really meant: how they translated into wins, and how executives could maximize success with a minimum of dollars.

"Moneyball" is now shorthand for any statistical approach in domains long ruled by the gut. But baseball represents a healthy case study—and it serves as a useful contrast to the toxic models, or WMDs, that are popping up in so many areas of our lives. Baseball models are fair, in part, because they're transparent. Everyone has access to the stats and can understand more or less how they're interpreted. Yes, one team's model might give more value to home run hitters, while another might discount them a bit, because sluggers tend to strike out a lot. But in either case, the numbers of home runs and strikeouts are there for everyone to see.

Baseball also has statistical rigor. Its gurus have an immense data set at hand, almost all of it directly related to the performance of players in the game. Moreover, their data is highly relevant to the outcomes they are trying to predict. This may sound obvious, but as we'll see throughout this book, the folks building WMDs routinely lack data for the behaviors they're most interested in. So they substitute stand-in data, or proxies. They draw statistical

correlations between a person's zip code or language patterns and her potential to pay back a loan or handle a job. These correlations are discriminatory, and some of them are illegal. Baseball models, for the most part, don't use proxies because they use pertinent inputs like balls, strikes, and hits.

Most crucially, that data is constantly pouring in, with new statistics from an average of twelve or thirteen games arriving daily from April to October. Statisticians can compare the results of these games to the predictions of their models, and they can see where they were wrong. Maybe they predicted that a left-handed reliever would give up lots of hits to right-handed batters—and yet he mowed them down. If so, the stats team has to tweak their model and also carry out research on why they got it wrong. Did the pitcher's new screwball affect his statistics? Does he pitch better at night? Whatever they learn, they can feed back into the model, refining it. That's how trustworthy models operate. They maintain a constant back-and-forth with whatever in the world they're trying to understand or predict. Conditions change, and so must the model.

Now, you may look at the baseball model, with its thousands of changing variables, and wonder how we could even be comparing it to the model used to evaluate teachers in Washington, D.C., schools. In one of them, an entire sport is modeled in fastidious detail and updated continuously. The other, while cloaked in mystery, appears to lean heavily on a handful of test results from one year to the next. Is that really a model?

The answer is yes. A model, after all, is nothing more than an abstract representation of some process, be it a baseball game, an oil company's supply chain, a foreign government's actions, or a movie theater's attendance. Whether it's running in a computer program or in our head, the model takes what we know and uses it to predict responses in various situations. All of us carry thousands

of models in our heads. They tell us what to expect, and they guide our decisions.

Here's an informal model I use every day. As a mother of three, I cook the meals at home—my husband, bless his heart, cannot remember to put salt in pasta water. Each night when I begin to cook a family meal, I internally and intuitively model everyone's appetite. I know that one of my sons loves chicken (but hates hamburgers), while another will eat only the pasta (with extra grated parmesan cheese). But I also have to take into account that people's appetites vary from day to day, so a change can catch my model by surprise. There's some unavoidable uncertainty involved.

The input to my internal cooking model is the information I have about my family, the ingredients I have on hand or I know are available, and my own energy, time, and ambition. The output is how and what I decide to cook. I evaluate the success of a meal by how satisfied my family seems at the end of it, how much they've eaten, and how healthy the food was. Seeing how well it is received and how much of it is enjoyed allows me to update my model for the next time I cook. The updates and adjustments make it what statisticians call a "dynamic model."

Over the years I've gotten pretty good at making meals for my family, I'm proud to say. But what if my husband and I go away for a week, and I want to explain my system to my mom so she can fill in for me? Or what if my friend who has kids wants to know my methods? That's when I'd start to formalize my model, making it much more systematic and, in some sense, mathematical. And if I were feeling ambitious, I might put it into a computer program.

Ideally, the program would include all of the available food options, their nutritional value and cost, and a complete database of my family's tastes: each individual's preferences and aversions. It would be hard, though, to sit down and summon all that

information off the top of my head. I've got loads of memories of people grabbing seconds of asparagus or avoiding the string beans. But they're all mixed up and hard to formalize in a comprehensive list.

The better solution would be to train the model over time, entering data every day on what I'd bought and cooked and noting the responses of each family member. I would also include parameters, or constraints. I might limit the fruits and vegetables to what's in season and dole out a certain amount of Pop-Tarts, but only enough to forestall an open rebellion. I also would add a number of rules. This one likes meat, this one likes bread and pasta, this one drinks lots of milk and insists on spreading Nutella on everything in sight.

If I made this work a major priority, over many months I might come up with a very good model. I would have turned the food management I keep in my head, my informal internal model, into a formal external one. In creating my model, I'd be extending my power and influence in the world. I'd be building an automated me that others can implement, even when I'm not around.

There would always be mistakes, however, because models are, by their very nature, simplifications. No model can include all of the real world's complexity or the nuance of human communication. Inevitably, some important information gets left out. I might have neglected to inform my model that junk-food rules are relaxed on birthdays, or that raw carrots are more popular than the cooked variety.

To create a model, then, we make choices about what's important enough to include, simplifying the world into a toy version that can be easily understood and from which we can infer important facts and actions. We expect it to handle only one job and accept that it will occasionally act like a clueless machine, one with enormous blind spots.

Sometimes these blind spots don't matter. When we ask Google Maps for directions, it models the world as a series of roads, tunnels, and bridges. It ignores the buildings, because they aren't relevant to the task. When avionics software guides an airplane, it models the wind, the speed of the plane, and the landing strip below, but not the streets, tunnels, buildings, and people.

A model's blind spots reflect the judgments and priorities of its creators. While the choices in Google Maps and avionics software appear cut and dried, others are far more problematic. The value-added model in Washington, D.C., schools, to return to that example, evaluates teachers largely on the basis of students' test scores, while ignoring how much the teachers engage the students, work on specific skills, deal with classroom management, or help students with personal and family problems. It's overly simple, sacrificing accuracy and insight for efficiency. Yet from the administrators' perspective it provides an effective tool to ferret out hundreds of apparently underperforming teachers, even at the risk of misreading some of them.

Here we see that models, despite their reputation for impartiality, reflect goals and ideology. When I removed the possibility of eating Pop-Tarts at every meal, I was imposing my ideology on the meals model. It's something we do without a second thought. Our own values and desires influence our choices, from the data we choose to collect to the questions we ask. Models are opinions embedded in mathematics.

Whether or not a model works is also a matter of opinion. After all, a key component of every model, whether formal or informal, is its definition of success. This is an important point that we'll return to as we explore the dark world of WMDs. In each case, we must ask not only who designed the model but also what that person or company is trying to accomplish. If the North Korean government built a model for my family's meals, for example, it

might be optimized to keep us above the threshold of starvation at the lowest cost, based on the food stock available. Preferences would count for little or nothing. By contrast, if my kids were creating the model, success might feature ice cream at every meal. My own model attempts to blend a bit of the North Koreans' resource management with the happiness of my kids, along with my own priorities of health, convenience, diversity of experience, and sustainability. As a result, it's much more complex. But it still reflects my own personal reality. And a model built for today will work a bit worse tomorrow. It will grow stale if it's not constantly updated. Prices change, as do people's preferences. A model built for a six-year-old won't work for a teenager.

This is true of internal models as well. You can often see troubles when grandparents visit a grandchild they haven't seen for a while. On their previous visit, they gathered data on what the child knows, what makes her laugh, and what TV show she likes and (unconsciously) created a model for relating to this particular four-year-old. Upon meeting her a year later, they can suffer a few awkward hours because their models are out of date. Thomas the Tank Engine, it turns out, is no longer cool. It takes some time to gather new data about the child and adjust their models.

This is not to say that good models cannot be primitive. Some very effective ones hinge on a single variable. The most common model for detecting fires in a home or office weighs only one strongly correlated variable, the presence of smoke. That's usually enough. But modelers run into problems—or subject *us* to problems—when they focus models as simple as a smoke alarm on their fellow humans.

Racism, at the individual level, can be seen as a predictive model whirring away in billions of human minds around the world. It is built from faulty, incomplete, or generalized data. Whether it comes from experience or hearsay, the data indicates

that certain types of people have behaved badly. That generates a binary prediction that all people of that race will behave that same way.

Needless to say, racists don't spend a lot of time hunting down reliable data to train their twisted models. And once their model morphs into a belief, it becomes hardwired. It generates poisonous assumptions, yet rarely tests them, settling instead for data that seems to confirm and fortify them. Consequently, racism is the most slovenly of predictive models. It is powered by haphazard data gathering and spurious correlations, reinforced by institutional inequities, and polluted by confirmation bias. In this way, oddly enough, racism operates like many of the WMDs I'll be describing in this book.

• • •

In 1997, a convicted murderer, an African American man named Duane Buck, stood before a jury in Harris County, Texas. Buck had killed two people, and the jury had to decide whether he would be sentenced to death or to life in prison with the chance of parole. The prosecutor pushed for the death penalty, arguing that if Buck were let free he might kill again.

Buck's defense attorney brought forth an expert witness, a psychologist named Walter Quijano, who didn't help his client's case one bit. Quijano, who had studied recidivism rates in the Texas prison system, made a reference to Buck's race, and during cross-examination the prosecutor jumped on it.

"You have determined that the . . . the race factor, black, increases the future dangerousness for various complicated reasons. Is that correct?" the prosecutor asked.

"Yes," Quijano answered. The prosecutor stressed that testimony in her summation, and the jury sentenced Buck to death.

Three years later, Texas attorney general John Cornyn found

that the psychologist had given similar race-based testimony in six other capital cases, most of them while he worked for the prosecution. Cornyn, who would be elected in 2002 to the US Senate, ordered new race-blind hearings for the seven inmates. In a press release, he declared: "It is inappropriate to allow race to be considered as a factor in our criminal justice system. . . . The people of Texas want and deserve a system that affords the same fairness to everyone."

Six of the prisoners got new hearings but were again sentenced to death. Quijano's prejudicial testimony, the court ruled, had not been decisive. Buck never got a new hearing, perhaps because it was his own witness who had brought up race. He is still on death row.

Regardless of whether the issue of race comes up explicitly at trial, it has long been a major factor in sentencing. A University of Maryland study showed that in Harris County, which includes Houston, prosecutors were three times more likely to seek the death penalty for African Americans, and four times more likely for Hispanics, than for whites convicted of the same charges. That pattern isn't unique to Texas. According to the American Civil Liberties Union, sentences imposed on black men in the federal system are nearly 20 percent longer than those for whites convicted of similar crimes. And though they make up only 13 percent of the population, blacks fill up 40 percent of America's prison cells.

So you might think that computerized risk models fed by data would reduce the role of prejudice in sentencing and contribute to more even-handed treatment. With that hope, courts in twenty-four states have turned to so-called recidivism models. These help judges assess the danger posed by each convict. And by many measures they're an improvement. They keep sentences more consistent and less likely to be swayed by the moods and biases of judges. They also save money by nudging down the length of the average sentence. (It costs an average of $31,000 a year to house an inmate, and double that in expensive states like Connecticut and New York.)

The question, however, is whether we've eliminated human bias or simply camouflaged it with technology. The new recidivism models are complicated and mathematical. But embedded within these models are a host of assumptions, some of them prejudicial. And while Walter Quijano's words were transcribed for the record, which could later be read and challenged in court, the workings of a recidivism model are tucked away in algorithms, intelligible only to a tiny elite.

One of the more popular models, known as LSI–R, or Level of Service Inventory–Revised, includes a lengthy questionnaire for the prisoner to fill out. One of the questions—"How many prior convictions have you had?"—is highly relevant to the risk of recidivism. Others are also clearly related: "What part did others play in the offense? What part did drugs and alcohol play?"

But as the questions continue, delving deeper into the person's life, it's easy to imagine how inmates from a privileged background would answer one way and those from tough inner-city streets another. Ask a criminal who grew up in comfortable suburbs about "the first time you were ever involved with the police," and he might not have a single incident to report other than the one that brought him to prison. Young black males, by contrast, are likely to have been stopped by police dozens of times, even when they've done nothing wrong. A 2013 study by the New York Civil Liberties Union found that while black and Latino males between the ages of fourteen and twenty-four made up only 4.7 percent of the city's population, they accounted for 40.6 percent of the stop-and-frisk checks by police. More than 90 percent of those stopped were innocent. Some of the others might have been drinking underage

or carrying a joint. And unlike most rich kids, they got in trouble for it. So if early "involvement" with the police signals recidivism, poor people and racial minorities look far riskier.

The questions hardly stop there. Prisoners are also asked about whether their friends and relatives have criminal records. Again, ask that question to a convicted criminal raised in a middle-class neighborhood, and the chances are much greater that the answer will be no. The questionnaire does avoid asking about race, which is illegal. But with the wealth of detail each prisoner provides, that single illegal question is almost superfluous.

The LSI–R questionnaire has been given to thousands of inmates since its invention in 1995. Statisticians have used those results to devise a system in which answers highly correlated to recidivism weigh more heavily and count for more points. After answering the questionnaire, convicts are categorized as high, medium, and low risk on the basis of the number of points they accumulate. In some states, such as Rhode Island, these tests are used only to target those with high-risk scores for antirecidivism programs while incarcerated. But in others, including Idaho and Colorado, judges use the scores to guide their sentencing.

This is unjust. The questionnaire includes circumstances of a criminal's birth and upbringing, including his or her family, neighborhood, and friends. These details should not be relevant to a criminal case or to the sentencing. Indeed, if a prosecutor attempted to tar a defendant by mentioning his brother's criminal record or the high crime rate in his neighborhood, a decent defense attorney would roar, "Objection, Your Honor!" And a serious judge would sustain it. This is the basis of our legal system. We are judged by what we do, not by who we are. And although we don't know the exact weights that are attached to these parts of the test, any weight above zero is unreasonable.

Many would point out that statistical systems like the LSI–R

are effective in gauging recidivism risk—or at least more accurate than a judge's random guess. But even if we put aside, ever so briefly, the crucial issue of fairness, we find ourselves descending into a pernicious WMD feedback loop. A person who scores as "high risk" is likely to be unemployed and to come from a neighborhood where many of his friends and family have had run-ins with the law. Thanks in part to the resulting high score on the evaluation, he gets a longer sentence, locking him away for more years in a prison where he's surrounded by fellow criminals—which raises the likelihood that he'll return to prison. He is finally released into the same poor neighborhood, this time with a criminal record, which makes it that much harder to find a job. If he commits another crime, the recidivism model can claim another success. But in fact the model itself contributes to a toxic cycle and helps to sustain it. That's a signature quality of a WMD.

• • •

In this chapter, we've looked at three kinds of models. The baseball models, for the most part, are healthy. They are transparent and continuously updated, with both the assumptions and the conclusions clear for all to see. The models feed on statistics from the game in question, not from proxies. And the people being modeled understand the process and share the model's objective: winning the World Series. (Which isn't to say that many players, come contract time, won't quibble with a model's valuations: "Sure I struck out two hundred times, but look at my *home runs* . . .")

From my vantage point, there's certainly nothing wrong with the second model we discussed, the hypothetical family meal model. If my kids were to question the assumptions that underlie it, whether economic or dietary, I'd be all too happy to provide them. And even though they sometimes grouse when facing

something green, they'd likely admit, if pressed, that they share the goals of convenience, economy, health, and good taste—though they might give them different weights in their own models. (And they'll be free to create them when they start buying their own food.)

I should add that my model is highly unlikely to scale. I don't see Walmart or the US Agriculture Department or any other titan embracing my app and imposing it on hundreds of millions of people, like some of the WMDs we'll be discussing. No, my model is benign, especially since it's unlikely ever to leave my head and be formalized into code.

The recidivism example at the end of the chapter, however, is a different story entirely. It gives off a familiar and noxious odor. So let's do a quick exercise in WMD taxonomy and see where it fits.

The first question: Even if the participant is aware of being modeled, or what the model is used for, is the model opaque, or even invisible? Well, most of the prisoners filling out mandatory questionnaires aren't stupid. They at least have reason to suspect that information they provide will be used against them to control them while in prison and perhaps lock them up for longer. They know the game. But prison officials know it, too. And they keep quiet about the purpose of the LSI–R questionnaire. Otherwise, they know, many prisoners will attempt to game it, providing answers to make them look like model citizens the day they leave the joint. So the prisoners are kept in the dark as much as possible and do not learn their risk scores.

In this, they're hardly alone. Opaque and invisible models are the rule, and clear ones very much the exception. We're modeled as shoppers and couch potatoes, as patients and loan applicants, and very little of this do we see—even in applications we happily sign up for. Even when such models behave themselves, opacity can lead to a feeling of unfairness. If you were told by an usher,

upon entering an open-air concert, that you couldn't sit in the first ten rows of seats, you might find it unreasonable. But if it were explained to you that the first ten rows were being reserved for people in wheelchairs, then it might well make a difference. Transparency matters.

And yet many companies go out of their way to hide the results of their models or even their existence. One common justification is that the algorithm constitutes a "secret sauce" crucial to their business. It's *intellectual property*, and it must be defended, if need be, with legions of lawyers and lobbyists. In the case of web giants like Google, Amazon, and Facebook, these precisely tailored algorithms alone are worth hundreds of billions of dollars. WMDs are, by design, inscrutable black boxes. That makes it extra hard to definitively answer the second question: Does the model work against the subject's interest? In short, is it unfair? Does it damage or destroy lives?

Here, the LSI–R again easily qualifies as a WMD. The people putting it together in the 1990s no doubt saw it as a tool to bring evenhandedness and efficiency to the criminal justice system. It could also help nonthreatening criminals land lighter sentences. This would translate into more years of freedom for them and enormous savings for American taxpayers, who are footing a $70 billion annual prison bill. However, because the questionnaire judges the prisoner by details that would not be admissible in court, it is unfair. While many may benefit from it, it leads to suffering for others.

A key component of this suffering is the pernicious feedback loop. As we've seen, sentencing models that profile a person by his or her circumstances help to create the environment that justifies their assumptions. This destructive loop goes round and round, and in the process the model becomes more and more unfair.

The third question is whether a model has the capacity to grow

exponentially. As a statistician would put it, can it scale? This might sound like the nerdy quibble of a mathematician. But scale is what turns WMDs from local nuisances into tsunami forces, ones that define and delimit our lives. As we'll see, the developing WMDs in human resources, health, and banking, just to name a few, are quickly establishing broad norms that exert upon us something very close to the power of law. If a bank's model of a high-risk borrower, for example, is applied to you, the world will treat you as just that, a deadbeat—even if you're horribly misunderstood. And when that model scales, as the credit model has, it affects your whole life—whether you can get an apartment or a job or a car to get from one to the other.

When it comes to scaling, the potential for recidivism modeling continues to grow. It's already used in the majority of states, and the LSI–R is the most common tool, used in at least twenty-four of them. Beyond LSI–R, prisons host a lively and crowded market for data scientists. The penal system is teeming with data, especially since convicts enjoy even fewer privacy rights than the rest of us. What's more, the system is so miserable, overcrowded, inefficient, expensive, and inhumane that it's crying out for improvements. Who wouldn't want a cheap solution like this?

Penal reform is a rarity in today's polarized political world, an issue on which liberals and conservatives are finding common ground. In early 2015, the conservative Koch brothers, Charles and David, teamed up with a liberal think tank, the Center for American Progress, to push for prison reform and drive down the incarcerated population. But my suspicion is this: their bipartisan effort to reform prisons, along with legions of others, is almost certain to lead to the efficiency and perceived fairness of a data-fed solution. That's the age we live in. Even if other tools supplant LSI–R as its leading WMD, the prison system is likely to be a powerful incubator for WMDs on a grand scale.

So to sum up, these are the three elements of a WMD: Opacity, Scale, and Damage. All of them will be present, to one degree or another, in the examples we'll be covering. Yes, there will be room for quibbles. You could argue, for example, that the recidivism scores are not totally opaque, since they spit out scores that prisoners, in some cases, can see. Yet they're brimming with mystery, since the prisoners cannot see how their answers produce their score. The scoring algorithm is hidden. A couple of the other WMDs might not seem to satisfy the prerequisite for scale. They're not huge, at least not yet. But they represent dangerous species that are primed to grow, perhaps exponentially. So I count them. And finally, you might note that not all of these WMDs are universally damaging. After all, they send some people to Harvard, line others up for cheap loans or good jobs, and reduce jail sentences for certain lucky felons. But the point is not whether some people benefit. It's that so many suffer. These models, powered by algorithms, slam doors in the face of millions of people, often for the flimsiest of reasons, and offer no appeal. They're unfair.

And here's one more thing about algorithms: they can leap from one field to the next, and they often do. Research in epidemiology can hold insights for box office predictions; spam filters are being retooled to identify the AIDS virus. This is true of WMDs as well. So if mathematical models in prisons appear to succeed at their job—which really boils down to efficient management of people—they could spread into the rest of the economy along with the other WMDs, leaving us as collateral damage.

That's my point. This menace is rising. And the world of finance provides a cautionary tale.