WEAPONS OF MATH DESTRUCTION

HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY

CATHY O'NEIL

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 x 3 x 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. Admin istrators, 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 bi ases 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.

5

CIVILIAN CASUALTIES

Justice in the Age of Big Data

The small city of Reading, Pennsylvania, has had a tough go of it in the postindustrial era. Nestled in the green hills fifty miles west of Philadelphia, Reading grew rich on railroads, steel, coal, and textiles. But in recent decades, with all of those industries in steep decline, the city has languished. By 2011, it had the highest poverty rate in the country, at 41.3 percent. (The following year, it was surpassed, if barely, by Detroit.) As the recession pummeled Reading's economy following the 2008 market crash, tax revenues fell, which led to a cut of forty-five officers in the police department—despite persistent crime.

Reading police chief William Heim had to figure out how to get the same or better policing out of a smaller force. So in 2013 he invested in crime prediction software made by PredPol, a Big Data start-up based in Santa Cruz, California. The program processed historical crime data and calculated, hour by hour, where crimes were most likely to occur. The Reading policemen could view the program's conclusions as a series of squares, each one just the size of two football fields. If they spent more time patrolling these squares, there was a good chance they would discourage crime. And sure enough, a year later, Chief Heim announced that burglaries were down by 23 percent.

Predictive programs like PredPol are all the rage in budget-strapped police departments across the country. Departments from Atlanta to Los Angeles are deploying cops in the shifting squares and reporting falling crime rates. New York City uses a similar program, called CompStat. And Philadelphia police are using a local product called HunchLab that includes risk terrain analysis, which incorporates certain features, such as ATMs or convenience stores, that might attract crimes. Like those in the rest of the Big Data industry, the developers of crime prediction software are hurrying to incorporate any information that can boost the accuracy of their models.

If you think about it, hot-spot predictors are similar to the shifting defensive models in baseball that we discussed earlier. Those systems look at the history of each player's hits and then position fielders where the ball is most likely to travel. Crime prediction software carries out similar analysis, positioning cops where crimes appear most likely to occur. Both types of models optimize resources. But a number of the crime prediction models are more sophisticated, because they predict progressions that could lead to waves of crime. PredPol, for example, is based on seismic software: it looks at a crime in one area, incorporates it into historical patterns, and predicts when and where it might occur next. (One simple correlation it has found: if burglars hit your next-door neighbor's house, batten down the hatches.)

Predictive crime models like PredPol have their virtues. Unlike the crime-stoppers in Steven Spielberg's dystopian movie *Minority Report* (and some ominous real-life initiatives, which we'll get to shortly), the cops don't track down people before they commit crimes. Jeffrey Brantingham, the UCLA anthropology professor who founded PredPol, stressed to me that the model is blind to race and ethnicity. And unlike other programs, including the recidivism risk models we discussed, which are used for sentencing guidelines, PredPol doesn't focus on the individual. Instead, it targets geography. The key inputs are the type and location of each crime and when it occurred. That seems fair enough. And if cops spend more time in the high-risk zones, foiling burglars and car thieves, there’s good reason to believe that the community benefits.

But most crimes aren’t as serious as burglary and grand theft auto, and that is where serious problems emerge. When police set up their PredPol system, they have a choice. They can focus exclusively on so-called Part 1 crimes. These are the violent crimes, including homicide, arson, and assault, which are usually reported to them. But they can also broaden the focus by including Part 2 crimes, including vagrancy, aggressive panhandling, and selling and consuming small quantities of drugs. Many of these “nuisance” crimes would go unrecorded if a cop weren’t there to see them.

These nuisance crimes are endemic to many impoverished neighborhoods. In some places police call them antisocial behavior, or ASB. Unfortunately, including them in the model threatens to skew the analysis. Once the nuisance data flows into a predictive model, more police are drawn into those neighborhoods, where they’re more likely to arrest more people. After all, even if their objective is to stop burglaries, murders, and rape, they’re bound to have slow periods. It’s the nature of patrolling. And if a patrolling cop sees a couple of kids who look no older than sixteen guzzling from a bottle in a brown bag, he stops them. These types of low-level crimes populate their models with more and more dots, and the models send the cops back to the same neighborhood.

This creates a pernicious feedback loop. The policing itself spawns new data, which justifies more policing. And our prisons fill up with hundreds of thousands of people found guilty of victimless crimes. Most of them come from impoverished neighborhoods, and most are black or Hispanic. So even if a model is color blind, the result of it is anything but. In our largely segregated cities, geography is a highly effective proxy for race.

If the purpose of the models is to prevent serious crimes, you might ask why nuisance crimes are tracked at all. The answer is that the link between antisocial behavior and crime has been an article of faith since 1982, when a criminologist named George Kelling teamed up with a public policy expert, James Q. Wilson, to write a seminal article in the *Atlantic Monthly* on so-called broken-windows policing. The idea was that low-level crimes and misdemeanors created an atmosphere of disorder in a neighborhood. This scared law-abiding citizens away. The dark and empty streets they left behind were breeding grounds for serious crime. The antidote was for society to resist the spread of disorder. This included fixing broken windows, cleaning up graffiti-covered subway cars, and taking steps to discourage nuisance crimes.

This thinking led in the 1990s to zero-tolerance campaigns, most famously in New York City. Cops would arrest kids for jumping the subway turnstiles. They'd apprehend people caught sharing a single joint and rumble them around the city in a paddy wagon for hours before eventually booking them. Some credited these energetic campaigns for dramatic falls in violent crimes. Others disagreed. The authors of the bestselling book *Freakonomics* went so far as to correlate the drop in crime to the legalization of abortion in the 1970s. And plenty of other theories also surfaced, ranging from the falling rates of crack cocaine addiction to the booming 1990s economy. In any case, the zero-tolerance movement gained broad support, and the criminal justice system sent millions of mostly young minority men to prison, many of them for minor offenses.

But zero tolerance actually had very little to do with Kelling and Wilson's "broken-windows" thesis. Their case study focused on what appeared to be a successful policing initiative in Newark, New Jersey. Cops who walked the beat there, according to the program, were supposed to be *highly* tolerant. Their job was to adjust to the neighborhood's own standards of order and to help uphold them. Standards varied from one part of the city to another. In one neighborhood, it might mean that drunks had to keep their bottles in bags and avoid major streets but that side streets were okay. Addicts could sit on stoops but not lie down. The idea was only to make sure the standards didn't fall. The cops, in this scheme, were helping a neighborhood maintain its own order but not imposing their own.

You might think I'm straying a bit from PredPol, mathematics, and WMDs. But each policing approach, from broken windows to zero tolerance, represents a model. Just like my meal planning or the U.S. News Top College ranking, each crime-fighting model calls for certain input data, followed by a series of responses, and each is calibrated to achieve an objective. It's important to look at policing this way, because these mathematical models now dominate law enforcement. And some of them are WMDs.

That said, we can understand why police departments would choose to include nuisance data. Raised on the orthodoxy of zero tolerance, many have little more reason to doubt the link between small crimes and big ones than the correlation between smoke and fire. When police in the British city of Kent tried out PredPol, in 2013, they incorporated nuisance crime data into their model. It seemed to work. They found that the PredPol squares were ten times as efficient as random patrolling and twice as precise as analysis delivered by police intelligence. And what type of crimes did the model best predict? Nuisance crimes. This makes all the sense in the world. A drunk will pee on the same wall, day in and day out, and a junkie will stretch out on the same park bench, while a car thief or a burglar will move about, working hard to anticipate the movements of police.

Even as police chiefs stress the battle against violent crime, it would take remarkable restraint not to let loads of nuisance data flow into their predictive models. More data, it's easy to believe, is better data. While a model focusing only on violent crimes might produce a sparse constellation on the screen, the inclusion of nuisance data would create a fuller and more vivid portrait of lawlessness in the city.

And in most jurisdictions, sadly, such a crime map would track poverty. The high number of arrests in those areas would do nothing but confirm the broadly shared thesis of society's middle and upper classes: that poor people are responsible for their own shortcomings and commit most of a city's crimes.

But what if police looked for different kinds of crimes? That may sound counterintuitive, because most of us, including the police, view crime as a pyramid. At the top is homicide. It's followed by rape and assault, which are more common, and then shoplifting, petty fraud, and even parking violations, which happen all the time. Prioritizing the crimes at the top of the pyramid makes sense. Minimizing violent crime, most would agree, is and should be a central part of a police force's mission.

But how about crimes far removed from the boxes on the PredPol maps, the ones carried out by the rich? In the 2000s, the kings of finance threw themselves a lavish party. They lied, they bet billions against their own customers, they committed fraud and paid off rating agencies. Enormous crimes were committed there, and the result devastated the global economy for the best part of five years. Millions of people lost their homes, jobs, and health care.

We have every reason to believe that more such crimes are occurring in finance right now. If we've learned anything, it's that the driving goal of the finance world is to make a huge profit, the bigger the better, and that anything resembling self-regulation is worthless. Thanks largely to the industry's wealth and powerful lobbies, finance is underpoliced.

Just imagine if police enforced their zero-tolerance strategy in finance. They would arrest people for even the slightest infraction, whether it was chiseling investors on 401ks, providing misleading guidance, or committing petty frauds. Perhaps SWAT teams would descend on Greenwich, Connecticut. They'd go undercover in the taverns around Chicago's Mercantile Exchange.

Not likely, of course. The cops don't have the expertise for that kind of work. Everything about their jobs, from their training to their bullet-proof vests, is adapted to the mean streets. Clamping down on white-collar crime would require people with different tools and skills. The small and underfunded teams who handle that work, from the FBI to investigators at the Securities and Exchange Commission, have learned through the decades that bankers are virtually invulnerable. They spend heavily on our politicians, which always helps, and are also viewed as crucial to our economy. That protects them. If their banks go south, our economy could go with them. (The poor have no such argument.) So except for a couple of criminal outliers, such as Ponzi-scheme master Bernard Madoff, financiers don't get arrested. As a group, they made it through the 2008 market crash practically unscathed. What could ever burn them now?

My point is that police make choices about where they direct their attention. Today they focus almost exclusively on the poor. That's their heritage, and their mission, as they understand it. And now data scientists are stitching this status quo of the social order into models, like PredPol, that hold ever-greater sway over our lives.

The result is that while PredPol delivers a perfectly useful and even high-minded software tool, it is also a do-it-yourself WMD. In this sense, PredPol, even with the best of intentions, empowers police departments to zero in on the poor, stopping more of them, arresting a portion of those, and sending a subgroup to prison. And the police chiefs, in many cases, if not most, think that they're taking the only sensible route to combating crime. That's where it is, they say, pointing to the highlighted ghetto on the map. And now they have cutting-edge technology (powered by Big Data) reinforcing their position there, while adding precision and "science" to the process.

The result is that we criminalize poverty, believing all the while that our tools are not only scientific but fair.

▪ ▪ ▪

One weekend in the spring of 2011, I attended a data "hackathon" in New York City. The goal of such events is to bring together hackers, nerds, mathematicians, and software geeks and to mobilize this brainpower to shine light on the digital systems that wield so much power in our lives. I was paired up with the New York Civil Liberties Union, and our job was to break out the data on one of the NYPD's major anticrime policies, so-called stop, question, and frisk. Known simply as stop and frisk to most people, the practice had drastically increased in the data-driven age of CompStat.

The police regarded stop and frisk as a filtering device for crime. The idea is simple. Police officers stop people who look suspicious to them. It could be the way they're walking or dressed, or their tattoos. The police talk to them and size them up, often while they're spread-eagled against a wall or the hood of a car. They ask for their ID, and they frisk them. Stop enough people, the thinking goes, and you'll no doubt stop loads of petty crimes, and perhaps some big ones. The policy, implemented by Mayor Michael Bloomberg's administration, had loads of public support. Over the previous decade, the number of stops had risen by 600 percent, to nearly seven hundred thousand incidents. The great majority of those stopped were innocent. For them, these encounters were highly unpleasant, even infuriating. Yet many in the public associated the program with the sharp decline of crime in the city. New York, many felt, was safer. And statistics indicated as much. Homicides, which had reached 2,245 in 1990, were down to 515 (and would drop below 400 by 2014).

Everyone knew that an outsized proportion of the people the police stopped were young, dark-skinned men. But how many did they stop? And how often did these encounters lead to arrests or stop crimes? While this information was technically public, much of it was stored in a database that was hard to access. The software didn’t work on our computers or flow into Excel spreadsheets. Our job at the hackathon was to break open that program and free the data so that we could all analyze the nature and effectiveness of the stop-and-frisk program.

What we found, to no great surprise, was that an overwhelming majority of these encounters—about 85 percent—involved young African American or Latino men. In certain neighborhoods, many of them were stopped repeatedly. Only 0.1 percent, or one of one thousand stopped, was linked in any way to a violent crime. Yet this filter captured many others for lesser crimes, from drug possession to underage drinking, that might have otherwise gone undiscovered. Some of the targets, as you might expect, got angry, and a good number of those found themselves charged with resisting arrest.

The NYCLU sued the Bloomberg administration, charging that the stop-and-frisk policy was racist. It was an example of uneven policing, one that pushed more minorities into the criminal justice system and into prison. Black men, they argued, were six times more likely to be incarcerated than white men and twenty-one times more likely to be killed by police, at least according to the available data (which is famously underreported).

Stop and frisk isn’t exactly a WMD, because it relies on human judgment and is not formalized into an algorithm. But it is built upon a simple and destructive calculation. If police stop one thousand people in certain neighborhoods, they’ll uncover, on average, one significant suspect and lots of smaller ones. This isn’t so different from the long-shot calculations used by predatory advertisers or spammers. Even when the hit ratio is miniscule, if you give yourself enough chances you’ll reach your target. And that helps to explain why the program grew so dramatically under Bloomberg’s watch. If stopping six times as many people led to six times the number of arrests, the inconvenience and harassment suffered by thousands upon thousands of innocent people was justified. Weren’t *they* interested in stopping crime?

Aspects of stop and frisk were similar to WMDs, though. For example, it had a nasty feedback loop. It ensnared thousands of black and Latino men, many of them for committing the petty crimes and misdemeanors that go on in college frats, unpunished, every Saturday night. But while the great majority of university students were free to sleep off their excesses, the victims of stop and frisk were booked, and some of them dispatched to the hell that is Rikers Island. What's more, each arrest created new data, further justifying the policy.

As stop and frisk grew, the venerable legal concept of probable cause was rendered virtually meaningless, because police were hunting not only people who might have already committed a crime but also those who might commit one in the future. Sometimes, no doubt, they accomplished this goal. By arresting a young man whose suspicious bulge turned out to be an unregistered gun, they might be saving the neighborhood from a murder or armed robbery, or even a series of them. Or maybe not. Whatever the case, there was a logic to stop and frisk, and many found it persuasive.

But was the policy constitutional? In August of 2013, federal judge Shira A. Scheindlin ruled that it was not. She said officers routinely "stopped blacks and Hispanics who would not have been stopped if they were white." Stop and frisk, she wrote, ran afoul of the Fourth Amendment, which protects against unreasonable searches and seizures by the government, and it also failed to provide the equal protection guaranteed by the Fourteenth Amendment. She called for broad reforms to the practice, including increased use of body cameras on patrolling policemen. This would help establish probable cause—or the lack of it—and remove some of the opacity from the stop-and-frisk model. But it would do nothing to address the issue of uneven policing.

While looking at WMDs, we're often faced with a choice between fairness and efficacy. Our legal traditions lean strongly toward fairness. The Constitution, for example, presumes innocence and is engineered to value it. From a modeler's perspective, the presumption of innocence is a constraint, and the result is that some guilty people go free, especially those who can afford good lawyers. Even those found guilty have the right to appeal their verdict, which chews up time and resources. So the system sacrifices enormous efficiencies for the promise of fairness. The

Constitution’s implicit judgment is that freeing someone who may well have committed a crime, for lack of evidence, poses less of a danger to our society than jailing or executing an innocent person.

WMDs, by contrast, tend to favor efficiency. By their very nature, they feed on data that can be measured and counted. But fairness is squishy and hard to quantify. It is a concept. And computers, for all of their advances in language and logic, still struggle mightily with concepts. They “understand” beauty only as a word associated with the Grand Canyon, ocean sunsets, and grooming tips in *Vogue* magazine. They try in vain to measure “friendship” by counting likes and connections on Facebook. And the concept of fairness utterly escapes them. Programmers don’t know how to code for it, and few of their bosses ask them to.

So fairness isn’t calculated into WMDs. And the result is massive, industrial production of *unfairness*. If you think of a WMD as a factory, unfairness is the black stuff belching out of the smoke stacks. It’s an emission, a toxic one.

The question is whether we as a society are willing to sacrifice a bit of efficiency in the interest of fairness. Should we handicap the models, leaving certain data out? It’s possible, for example, that adding gigabytes of data about antisocial behavior might help PredPol predict the mapping coordinates for serious crimes. But this comes at the cost of a nasty feedback loop. So I’d argue that we should discard the data.

It’s a tough case to make, similar in many ways to the battles over wiretapping by the National Security Agency. Advocates of the snooping argue that it’s important for our safety. And those running our vast national security apparatus will keep pushing for more information to fulfill their mission. They’ll continue to encroach on people’s privacy until they get the message that they must find a way to do their job within the bounds of the Constitution. It might be harder, but it’s necessary.

The other issue is equality. Would society be so willing to sacrifice the concept of probable cause if everyone had to endure the harassment and indignities of stop and frisk? Chicago police have their own stop-and-frisk program. In the name of fairness, what if they sent a bunch of patrollers into the city’s exclusive Gold Coast? Maybe they’d arrest joggers for jaywalking from the park across W. North Boulevard or crack down on poodle pooping along Lakeshore Drive. This heightened police presence would probably pick up more drunk drivers and perhaps uncover a few cases of insurance fraud, spousal abuse, or racketeering. Occasionally, just to give everyone a taste of the unvarnished experience, the cops might throw wealthy citizens on the trunks of their cruisers, wrench their arms, and snap on the handcuffs, perhaps while swearing and calling them hateful names.

In time, this focus on the Gold Coast would create data. It would describe an increase in crime there, which would draw even more police into the fray. This would no doubt lead to growing anger and confrontations. I picture a double parker talking back to police, refusing to get out of his Mercedes, and finding himself facing charges for resisting arrest. Yet another Gold Coast crime.

This may sound less than serious. But a crucial part of justice is equality. And that means, among many other things, experiencing criminal justice equally. People who favor policies like stop and frisk should experience it themselves. Justice cannot just be something that one part of society inflicts upon the other.

The noxious effects of uneven policing, whether from stop and frisk or predictive models like PredPol, do not end when the accused are arrested and booked in the criminal justice sys tem. Once there, many of them confront another WMD that I discussed in chapter 1, the recidivism model used for sentencing guidelines. The biased data from uneven policing funnels right into this model. Judges then look to this supposedly scientific analysis, crystallized into a single risk score. And those who take this score seriously have reason to give longer sentences to prisoners who appear to pose a higher risk of committing other crimes.

And why are nonwhite prisoners from poor neighborhoods more likely to commit crimes? According to the data inputs for the recidivism models, it’s because they’re more likely to be jobless, lack a high school diploma, and have had previous run-ins with the law. And their friends have, too.

Another way of looking at the same data, though, is that these prisoners live in poor neighborhoods with terrible schools and scant opportunities. And they’re highly policed. So the chance that an ex-convict returning to that neighborhood will have another brush with the law is no doubt larger than that of a tax fraudster who is released into a leafy suburb. In this system, the poor and nonwhite are punished more for being who they are and living where they live.

What's more, for supposedly scientific systems, the recidivism models are logically flawed. The unquestioned assumption is that locking away "high-risk" prisoners for more time makes society safer. It is true, of course, that prisoners don't commit crimes against society while behind bars. But is it possible that their time in prison has an effect on their behavior once they step out? Is there a chance that years in a brutal environment surrounded by felons might make them more likely, and not less, to commit another crime? Such a finding would undermine the very basis of the recidivism sentencing guidelines. But prison systems, which are awash in data, do not carry out this highly important research. All too often they use data to justify the workings of the system but not to question or improve the system.

Compare this attitude to the one found at Amazon.com. The giant retailer, like the criminal justice system, is highly focused on a form of recidivism. But Amazon's goal is the opposite. It wants people to come back again and again to buy. Its software system targets recidivism and encourages it.

Now, if Amazon operated like the justice system, it would start by scoring shoppers as potential recidivists. Maybe more of them live in certain area codes or have college degrees. In this case, Amazon would market more to these people, perhaps offering them discounts, and if the marketing worked, those with high recidivist scores would come back to shop more. If viewed superficially, the results would appear to corroborate Amazon's scoring system.

But unlike the WMDs in criminal justice, Amazon does not settle for such glib correlations. The company runs a data laboratory. And if it wants to find out what drives shopping recidivism, it carries out research. Its data scientists don't just study zip codes and education levels. They also inspect people's experience within the Amazon ecosystem. They might start by looking at the patterns of all the people who shopped once or twice at Amazon and never returned. Did they have trouble at checkout? Did their packages arrive on time? Did a higher percentage of them post a bad review? The questions go on and on, because the future of the company hinges upon a system that learns continually, one that figures out what makes customers tick.

If I had a chance to be a data scientist for the justice system, I would do my best to dig deeply to learn what goes on inside those prisons and what impact those experiences might have on prisoners' behavior. I'd first look into solitary confinement. Hundreds of thousands of prisoners are kept for twenty-three hours a day in these prisons within prisons, most of them no bigger than a horse stall. Researchers have found that time in solitary produces deep feelings of hopelessness and despair. Could that have any impact on recidivism? That's a test I'd love to run, but I'm not sure the data is even collected.

How about rape? In *Unfair: The New Science of Criminal Injustice*, Adam Benforado writes that certain types of prisoners are targeted for rape in prisons. The young and small of stature are especially vulnerable, as are the mentally disabled. Some of these people live for years as sex slaves. It's another important topic for analysis that anyone with the relevant data and expertise could work out, but prison systems have thus far been uninterested in cataloging the long-term effects of this abuse.

A serious scientist would also search for positive signals from the prison experience. What's the impact of more sunlight, more sports, better food, literacy training? Maybe these factors will improve convicts' behavior after they go free. More likely, they'll have varying impact. A serious justice system research program would delve into the effects of each of these elements, how they work together, and which people they're most likely to help. The goal, if data were used constructively, would be to optimize prisons—much the way companies like Amazon optimize websites or supply chains—for the benefit of both the prisoners and society at large.

But prisons have every incentive to avoid this data-driven approach. The PR risks are too great—no city wants to be the subject of a scathing report in the *New York Times*. And, of course, there's big money riding on the overcrowded prison system. Privately run prisons, which house only 10 percent of the incarcerated population, are a $5 billion industry. Like airlines, the private prisons make profits only when running at high capacity. Too much poking and prodding might threaten that income source.

So instead of analyzing prisons and optimizing them, we deal with them as black boxes. Prisoners go in and disappear from our view. Nastiness no doubt occurs, but behind thick walls. What goes on in there? Don’t ask. The current models stubbornly stick to the dubious and unquestioned hypothesis that more prison time for supposedly high-risk prisoners makes us safer. And if studies appear to upend that logic, they can be easily ignored.

And this is precisely what happens. Consider a recidivism study by Michigan economics professor Michael Mueller-Smith. After studying 2.6 million criminal court records in Harris County, Texas, he concluded that the longer inmates in Harris County, Texas, spent locked up, the greater the chance that they would fail to find employment upon release, would require food stamps and other public assistance, and would commit further crimes. But to turn those conclusions into smart policy and better justice, politicians will have to take a stand on behalf of a feared minority that many (if not most) voters would much prefer to ignore. It’s a tough sell.

▪ ▪ ▪

Stop and frisk may seem intrusive and unfair, but in short time it will also be viewed as primitive. That’s because police are bringing back tools and techniques from the global campaign against terrorism and focusing them on local crime fighting. In San Diego, for example, police are not only asking the people they stop for identification, or frisking them. On occasion, they also take photos of them with iPads and send them to a cloud-based facial recognition service, which matches them against a database of criminals and suspects. According to a report in the *New York Times*, San Diego police used this facial recognition program on 20,600 people between 2011 and 2015. They also probed many of them with mouth swabs to harvest DNA.

Advances in facial recognition technology will soon allow for much broader surveillance. Officials in Boston, for example, were considering using security cameras to scan thousands of faces at outdoor concerts. This data would be uploaded to a service that could match each face against a million others per second. In the end, officials decided against it. Concern for privacy, on that occasion, trumped efficiency. But this won't always be the case.

As technology advances, we're sure to see a dramatic growth of surveillance. The good news, if you want to call it that, is that once thousands of security cameras in our cities and towns are sending up our images for analysis, police won't have to discriminate as much. And the technology will no doubt be useful for tracking down suspects, as happened in the Boston Marathon bombing. But it means that we'll all be subject to a digital form of stop and frisk, our faces matched against databases of known criminals and terrorists.

The focus then may well shift toward spotting *potential* lawbreakers—not just neighborhoods or squares on a map but individuals. These preemptive campaigns, already well established in the fight against terrorism, are a breeding ground for WMDs.

In 2009, the Chicago Police Department received a $2 million grant from the National Institute of Justice to develop a predictive program for crime. The theory behind Chicago's winning application was that with enough research and data they might be able to demonstrate that the spread of crime, like epidemics, follows certain patterns. It can be predicted and, hopefully, prevented.

The scientific leader of the Chicago initiative was Miles Wernick, the director of the Medical Imaging Research Center at the Illinois Institute of Technology (IIT). Decades earlier, Wernick had helped the US military analyze data to pick out battlefield targets. He had since moved to medical data analysis, including the progression of dementia. But like most data scientists, he didn't see his expertise as tethered to a specific industry. He spotted patterns. And his focus in Chicago would be the patterns of crime, and of criminals.

The early efforts of Wernick's team focused on singling out hot spots for crime, much as PredPol does. But the Chicago team went much further. They developed a list of the approximately four hundred people most likely to commit a violent crime. And it ranked them on the probability that they would be involved in a homicide.

One of the people on the list, a twenty-two-year-old high school dropout named Robert McDaniel, answered his door one summer day in 2013 and found himself facing a police officer. McDaniel later told the *Chicago Tribune* that he had no history of gun violations and had never been charged with a violent crime. Like most of the young men in Austin, his dangerous West Side neighborhood, McDaniel had had brushes with the law, and he knew plenty of people caught up in the criminal justice system. The policewoman, he said, told him that the force had its eye on him and to watch out.

Part of the analysis that led police to McDaniel involved his social network. He knew criminals. And there is no denying that people are statistically more likely than not to behave like the people they spend time with. Facebook, for example, has found that friends who communicate often are far more likely to click on the same advertisement. Birds of a feather, statistically speaking, *do* fly together.

And to be fair to Chicago police, they're not arresting people like Robert McDaniel, at least not yet. The goal of the police in this exercise is to save lives. If the four hundred people who appear most likely to commit violent crimes receive a knock on the door and a warning, maybe some of them will think twice before packing a gun.

But let's consider McDaniel's case in terms of fairness. He hap pened to grow up in a poor and dangerous neighborhood. In this, he was unlucky. He has been surrounded by crime, and many of his acquaintances have gotten caught up in it. And largely because of these circumstances—and not his own actions—he has been deemed dangerous. Now the police have their eye on him. And if he behaves foolishly, as millions of other Americans do on a regular basis, if he buys drugs or gets into a barroom fight or carries an unregistered handgun, the full force of the law will fall down on him, and probably much harder than it would on most of us. After all, he's been warned.

I would argue that the model that led police to Robert McDaniel's door has the wrong objective. Instead of simply trying to eradicate crimes, police should be attempting to build relationships in the neighborhood. This was one of the pillars of the original "broken-windows" study. The cops were on foot, talking to people, trying to help them uphold their own community standards. But that objective, in many cases, has been lost, steamrollered by models that equate arrests with safety.

This isn’t the case everywhere. I recently visited Camden, New Jersey, which was the murder capital of the country in 2011. I found that the police department in Camden, rebuilt and placed under state control in 2012, had a dual mandate: lowering crime and engendering community trust. If building trust is the objective, an arrest may well become a last resort, not the first. This more empathetic approach could lead to warmer relations between the police and the policed, and fewer of the tragedies we’ve seen in recent years —the police killings of young black men and the riots that follow them.

From a mathematical point of view, however, trust is hard to quantify. That’s a challenge for people building models. Sadly, it’s far simpler to keep counting arrests, to build models that assume we’re birds of a feather and treat us as such. Innocent people surrounded by criminals get treated badly, and criminals surrounded by a law-abiding public get a pass. And because of the strong correlation between poverty and reported crime, the poor continue to get caught up in these digital dragnets. The rest of us barely have to think about them.