pay less attention to WMDs like value-added models. Or better yet, jettison them entirely.

At around the same time, New York governor Andrew Cuomo's education task force called for a four-year moratorium on the use of exams to evaluate teachers. This change, while welcome, does not signal a clear rejection of the teacher evaluation WMDs, much less a recognition that they're unfair. The push, in fact, came from the parents, who complained that the testing regime was wearing out their kids and taking too much time in the school year. A boycott movement had kept 20 percent of third through eighth graders out of the tests in the spring of 2015, and it was growing. In bowing to the parents, the Cuomo administration delivered a blow to value-added modeling. After all, without a full complement of student tests, the state would lack the data to populate it.

Tim Clifford was cheered by this news but still wary. "The opt-out movement forced Cuomo's hand," he wrote in an e-mail. "He feared losing the support of wealthier voters in top school districts, who were the very people who most staunchly supported him. To get ahead of the issue, he's placed this moratorium on using test scores." Clifford fears that the tests will be back.

Maybe so. And, given that value-added modeling has become a proven tool against teachers' unions, I don't expect it to disappear anytime soon. It's well entrenched, with forty states and the District of Columbia using or developing one form of it or another. That's all the more reason to spread the word about these and other WMDs. Once people recognize them and understand their statistical flaws, they'll demand evaluations that are fairer for both students and teachers. However, if the goal of the testing is to find someone to blame, and to intimidate workers, then, as we've seen, a WMD that spews out meaningless scores gets an A-plus.

8

COLLATERAL DAMAGE

Landing Credit

Local bankers used to stand tall in a town. They controlled the money. If you wanted a new car or a mortgage, you'd put on your Sunday best and pay a visit. And as a member of your community, this banker would probably know the following details about your life. He'd know about your churchgoing habits, or lack of them. He'd know all the stories about your older brother's run-ins with the law. He'd know what your boss (and his golfing buddy) said about you as a worker. Naturally, he'd know your race and ethnic group, and he'd also glance at the numbers on your application form.

The first four factors often worked their way, consciously or not, Now, you might think that large numbers would bring the scores into focus. After all, New York City, with its 1.1 million public school students, should provide a big enough data set to create meaningful predictions. If eighty thousand eighth graders take the test, wouldn't it be feasible to establish reliable averages for struggling, middling, and thriving schools?

Yes. And if Tim Clifford were teaching a large sampling of students, say ten thousand, then it might be reasonable to measure that cohort against the previous year's average and draw some conclusions from it. Large numbers balance out the exceptions and outliers. Trends, theoretically, would come into focus. But it's almost impossible for a class of twenty-five or thirty students to match up with the larger population. So if a class has certain types of students, they will tend to rise faster than the average. Others will rise more slowly. Clifford was given virtually no information about the opaque WMD that gave him such wildly divergent scores, but he assumed this variation in his classes had something to do with it. The year he scored poorly, Clifford said, "I taught many special education students as well as many top performers. And I think serving either the neediest or the top students—or both—creates problems. Needy students' scores are hard to move because they have learning problems, and top students' scores are hard to move because they have already scored high so there's little room for improvement."

The following year, he had a different mix of students, with more of them falling between the extremes. And the results made it look as though Clifford had progressed from being a failing teacher to being a spectacular one. Such results were all too common. An analysis by a blogger and educator named Gary Rubinstein found that of teachers who taught the same subject in consecutive years, one in four registered a 40-point difference. That suggests that the evaluation data is practically random. It

wasn't the teachers' performance that was bouncing all over the place. It was the scoring generated by a bogus WMD.

While its scores are meaningless, the impact of value-added modeling is pervasive and nefarious. "I've seen some great teachers convince themselves that they were mediocre at best based on those scores," Clifford said. "It moved them away from the great lessons they used to teach, toward increasing test prep. To a young teacher, a poor value-added score is punishing, and a good one may lead to a false sense of accomplishment that has not been earned."

As in the case of so many WMDs, the existence of value-added modeling stems from good intentions. The Obama administration realized early on that school districts punished under the 2001 No Child Left Behind reforms, which mandated high-stakes standardized testing, tended to be poor and disadvantaged. So it offered waivers to districts that could demonstrate the effectiveness of their teachers, ensuring that these schools would not be punished even if their students were lagging.*

The use of value-added models stems in large part from this regulatory change. But in late 2015 the teacher testing craze took what may be an even more dramatic turn. First, Congress and the White House agreed to revoke No Child Left Behind and replace it with a law that gives states more latitude to develop their own approaches for turning around underperforming school districts. It also gives them a broader range of criteria to consider, including student and teacher engagement, access to advanced coursework, school climate, and safety. In other words, education officials can attempt to study what's happening at each individual school—and

* No Child Left Behind sanctions include offering students in failing schools the option of attending another, more successful school. In dire cases, the law calls for a failing school to be closed and replaced by a charter school.

"You'd think I'd have been elated, but I wasn't," he said. "I knew that my low score was bogus, so I could hardly rejoice at getting a high score using the same flawed formula. The 90 percent difference in scores only made me realize how ridiculous the entire value-added model is when it comes to education."

Bogus is the word for it. In fact, misinterpreted statistics run through the history of teacher evaluation. The problem started with a momentous statistical boo-boo in the analysis of the original *Nation at Risk* report. It turned out that the very researchers who were decrying a national catastrophe were basing their judgment on a fundamental error, something an undergrad should have caught. In fact, if they wanted to serve up an example of America's educational shortcomings, their own misreading of statistics could serve as exhibit A.

Seven years after *A Nation at Risk* was published with such fanfare, researchers at Sandia National Laboratories took a second look at the data gathered for the report. These people were no amateurs when it came to statistics—they build and maintain nuclear weapons—and they quickly found the error. Yes, it was true that SAT scores had gone down on average. However, the number of students taking the test had ballooned over the course of those seventeen years. Universities were opening their doors to more poor students and minorities. Opportunities were expanding. This signaled social success. But naturally, this influx of newcomers dragged down the average scores. However, when statisticians broke down the population into income groups, scores for every single group were rising, from the poor to the rich.

In statistics, this phenomenon is known as Simpson's Paradox: when a whole body of data displays one trend, yet when broken into subgroups, the opposite trend comes into view for each of those subgroups. The damning conclusion in the *Nation at Risk* report, the one that spurred the entire teacher evaluation

movement, was drawn from a grievous misinterpretation of the data.

Tim Clifford's diverging scores are the result of yet another case of botched statistics, this one all too common. The teacher scores derived from the tests measured *nothing*. This may sound like hyperbole. After all, kids took tests, and those scores contributed to Clifford's. That much is true. But Clifford's scores, both his humiliating 6 and his chest-thumping 96, were based almost entirely on approximations that were so weak they were essentially random.

The problem was that the administrators lost track of accuracy in their quest to be fair. They understood that it wasn't right for teachers in rich schools to get too much credit when the sons and daughters of doctors and lawyers marched off toward elite universities. Nor should teachers in poor districts be held to the same standards of achievement. We cannot expect them to perform miracles.

So instead of measuring teachers on an absolute scale, they tried to adjust for social inequalities in the model. Instead of comparing Tim Clifford's students to others in different neighborhoods, they would compare them with forecast models of *themselves*. The students each had a predicted score. If they surpassed this prediction, the teacher got the credit. If they came up short, the teacher got the blame. If that sounds primitive to you, believe me, it is.

Statistically speaking, in these attempts to free the tests from class and color, the administrators moved from a primary to a secondary model. Instead of basing scores on direct measurement of the students, they based them on the so-called error term—the gap between results and expectations. Mathematically, this is a much sketchier proposition. Since the expectations themselves are derived from statistics, these amount to guesses on top of guesses. The result is a model with loads of random results, what statisticians call "noise."

For a few decades, it may have seemed that industrial workers and service workers were the only ones who could be modeled and optimized, while those who trafficked in ideas, from lawyers to chemical engineers, could steer clear of WMDs, at least at work. Cataphora was an early warning that this will not be the case. Indeed, throughout the tech industry, many companies are busy trying to optimize their white-collar workers by looking at the patterns of their communications. The tech giants, including Google, Facebook, Amazon, IBM, and many others, are hot on this trail.

For now, at least, this diversity is welcome. It holds out the hope, at least, that workers rejected by one model might be appreciated by another. But eventually, an industry standard will emerge, and then we'll all be in trouble.

• • •

In 1983, the Reagan administration issued a lurid alarm about the state of America's schools. In a report called *A Nation at Risk*, a presidential panel warned that a "rising tide of mediocrity" in the schools threatened "our very future as a Nation and a people." The report added that if "an unfriendly foreign power" had attempted to impose these bad schools on us, "we might well have viewed it as an act of war."

The most noteworthy signal of failure was what appeared to be plummeting scores on the SATs. Between 1963 and 1980, verbal scores had fallen by 50 points, and math scores were down 40 points. Our ability to compete in a global economy hinged on our skills, and they seemed to be worsening.

Who was to blame for this sorry state of affairs? The report left no doubt about that. Teachers. The *Nation at Risk* report called for action, which meant testing the students—and using the results to zero in on the underperforming teachers. As we saw in

the Introduction, this practice can cost teachers their jobs. Sarah Wysocki, the teacher in Washington who was fired after her class posted surprisingly low scores, was the victim of such a test. My point in telling that story was to show a WMD in action, how it can be arbitrary, unfair, and deaf to appeals.

But along with being educators and caretakers of children, teachers are obviously workers, and here I want to delve a bit deeper into the models that score their performance, because they might spread to other parts of the workforce. Consider the case of Tim Clifford. He's a middle school English teacher in New York City, with twenty-six years of experience. A few years ago, Clifford learned that he had bombed on a teacher evaluation, a so-called value-added model, similar to the one that led to Sarah Wysocki's firing. Clifford's score was an abysmal 6 out of 100.

He was devastated. "I didn't see how it was possible that I could have worked so hard and gotten such poor results," he later told me. "To be honest, when I first learned my low score, I felt ashamed and didn't tell anyone for a day or so. However, I learned that there were actually two other teachers who scored below me in my school. That emboldened me to share my results, because I wanted those teachers to know it wasn't only them."

If Clifford hadn't had tenure, he could have been dismissed that year, he said. "Even with tenure," he said, "scoring low in consecutive years is bound to put a target on a teacher's back to some degree." What's more, when tenured teachers register low scores, it emboldens school reformers, who make the case that job security protects incompetent educators. Clifford approached the following year with trepidation.

The value-added model had given him a failing grade but no advice on how to improve it. So Clifford went on teaching the way he always had and hoped for the best. The following year, his score was a 96.

were more productive than others. They hung a so-called sociometric badge around each employee's neck. The electronics in these badges tracked the employees' location and also measured, every sixteen milliseconds, their tone of voice and gestures. It recorded when people were looking at each other and how much each person talked, listened, and interrupted. Four teams of call center employees—eighty people in total—wore these badges for six weeks.

These employees' jobs were highly regimented. Talking was discouraged because workers were supposed to spend as many of their minutes as possible on the phone, solving customers' problems. Coffee breaks were scheduled one by one.

The researchers found, to their surprise, that the fastest and most efficient call center team was also the most social. These employees pooh-poohed the rules and gabbed much more than the others. And when all of the employees were encouraged to socialize more, call center productivity soared.

But data studies that track employees' behavior can also be used to cull a workforce. As the 2008 recession ripped through the economy, HR officials in the tech sector started to look at those Cataphora charts with a new purpose. They saw that some workers were represented as big dark circles, while others were smaller and dimmer. If they had to lay off workers, and most companies did, it made sense to start with the small and dim ones on the chart.

Were those workers really expendable? Again we come to digital phrenology. If a system designates a worker as a low idea generator or weak connector, that verdict becomes its own truth. That's her score.

Perhaps someone can come in with countervailing evidence. The worker with the dim circle might generate fabulous ideas but not share them on the network. Or perhaps she proffers priceless advice over lunch or breaks up the tension in the office with a joke. Maybe everybody likes her. That has great value in the workplace. But computing systems have trouble finding digital proxies for these kinds of soft skills. The relevant data simply isn't collected, and anyway it's hard to put a value on them. They're usually easier to leave out of a model.

So the system identifies apparent losers. And a good number of them lost their jobs during the recession. That alone is unjust. But what's worse is that systems like Cataphora's receive minimal feedback data. Someone identified as a loser, and subsequently fired, may have found another job and generated a fistful of patents. That data usually isn't collected. The system has no inkling that it got one person, or even a thousand people, entirely wrong.

That's a problem, because scientists need this error feedback—in this case the presence of false negatives—to delve into forensic analysis and figure out what went wrong, what was misread, what data was ignored. It's how systems learn and get smarter. Yet as we've seen, loads of WMDs, from recidivism models to teacher scores, blithely generate their own reality. Managers assume that the scores are true enough to be useful, and the algorithm makes tough decisions easy. They can fire employees and cut costs and blame their decisions on an objective number, whether it's accurate or not.

Cataphora remained small, and its worker evaluation model was a sideline—much more of its work was in identifying patterns of fraud or insider trading within companies. The company went out of business in 2012, and its software was sold to a start-up, Chenope. But systems like Cataphora's have the potential to become true WMDs. They can misinterpret people, and punish them, without any proof that their scores correlate to the quality of their work.

This type of software signals the rise of WMDs in a new realm.

efficiency and profitability, not for justice or the good of the "team." This is, of course, the nature of capitalism. For companies, revenue is like oxygen. It keeps them alive. From their perspective, it would be profoundly stupid, even unnatural, to turn away from potential savings. That's why society needs countervailing forces, such as vigorous press coverage that highlights the abuses of efficiency and shames companies into doing the right thing. And when they come up short, as Starbucks did, it must expose them again and again. It also needs regulators to keep them in line, strong unions to organize workers and amplify their needs and complaints, and politicians willing to pass laws to restrain corporations' worst excesses. Following the *New York Times* report in 2014, Democrats in Congress promptly drew up bills to rein in scheduling software. But facing a Republican majority fiercely opposed to government regulations, the chances that their bill would become law were nil. The legislation died.

• • •

In 2008, just as the great recession was approaching, a San Francisco company called Cataphora marketed a software system that rated tech workers on a number of metrics, including their generation of ideas. This was no easy task. Software programs, after all, are hard-pressed to distinguish between an idea and a simple string of words. If you think about it, the difference is often just a matter of context. Yesterday's ideas—that the earth is round, or even that people might like to share photos in social networks—are today's facts. We humans each have a sense for when an idea becomes an established fact and know when it has been debunked or discarded (though we often disagree). However, that distinction flummoxes even the most sophisticated AI. So Cataphora's system needed to look to humans themselves for guidance.

Cataphora's software burrowed into corporate e-mail and messaging in its hunt for ideas. Its guiding hypothesis was that the best ideas would tend to spread more widely through the network. If people cut and pasted certain groups of words and shared them, those words were likely ideas, and the software could quantify them.

But there were complications. Ideas were not the only groups of words that were widely shared on social networks. Jokes, for example, were wildly viral and equally befuddling to software systems. Gossip also traveled like a rocket. However, jokes and gossip followed certain patterns, so it was possible to teach the program to filter out at least some of them. With time, the system identified the groups of words most likely to represent ideas. It tracked them through the network, counting the number of times they were copied, measuring their distribution, and identifying their source.

Very soon, the roles of the employees appeared to come into focus. Some people were idea generators, the system concluded. On its chart of employees, Cataphora marked idea generators with circles, which were bigger and darker if they produced lots of ideas. Other people were connectors. Like neurons in a distributed network, they transmitted information. The most effective connectors made snippets of words go viral. The system painted those people in dark colors as well.

Now, whether or not this system effectively measured the flow of ideas, the concept itself was not nefarious. It can make sense to use this type of analysis to identify what people know and to match them with their most promising colleagues and collaborators. IBM and Microsoft use in-house programs to do just this. It's very similar to a dating algorithm (and often, no doubt, has similarly spotty results). Big Data has also been used to study the productivity of call center workers.

A few years ago, MIT researchers analyzed the behavior of call center employees for Bank of America to find out why some teams the globe. Companies in many geographies can establish just-in-time supply chains in a snap. These models likewise constitute the mathematical underpinnings of companies like Amazon, Federal Express, and UPS.

Scheduling software can be seen as an extension of the just-in-time economy. But instead of lawn mower blades or cell phone screens showing up right on cue, it's people, usually people who badly need money. And because they need money so desperately, the companies can bend their lives to the dictates of a mathematical model.

I should add that companies take steps not to make people's lives *too* miserable. They all know to the penny how much it costs to replace a frazzled worker who finally quits. Those numbers are in the data, too. And they have other models, as we discussed in the last chapter, to reduce churn, which drains profits and efficiency.

The trouble, from the employees' perspective, is an oversupply of low-wage labor. People are hungry for work, which is why so many of them cling to jobs that pay barely eight dollars per hour. This oversupply, along with the scarcity of effective unions, leaves workers with practically no bargaining power. This means the big retailers and restaurants can twist the workers' lives to ever-more-absurd schedules without suffering from excessive churn. They make more money while their workers' lives grow hellish. And because these optimization programs are everywhere, the workers know all too well that changing jobs isn't likely to improve their lot. Taken together, these dynamics provide corporations with something close to a captive workforce.

I'm sure it comes as no surprise that I consider scheduling software one of the more appalling WMDs. It's massive, as we've discussed, and it takes advantage of people who are already struggling to make ends meet. What's more, it is entirely opaque. Workers often don't have a clue about when they'll be called to work. They are summoned by an arbitrary program.

Scheduling software also creates a poisonous feedback loop. Consider Jannette Navarro. Her haphazard scheduling made it impossible for her to return to school, which dampened her employment prospects and kept her in the oversupplied pool of low-wage workers. The long and irregular hours also make it hard for workers to organize or to protest for better conditions. Instead, they face heightened anxiety and sleep deprivation, which causes dramatic mood swings and is responsible for an estimated 13 percent of highway deaths. Worse yet, since the software is designed to save companies money, it often limits workers' hours to fewer than thirty per week, so that they are not eligible for company health insurance. And with their chaotic schedules, most find it impossible to make time for a second job. It's almost as if the software were designed expressly to punish low-wage workers and to keep them down.

The software also condemns a large percentage of our children to grow up without routines. They experience their mother bleary eyed at breakfast, or hurrying out the door without dinner, or arguing with *her* mother about who can take care of them on Sunday morning. This chaotic life affects children deeply. According to a study by the Economic Policy Institute, an advocacy group, "Young children and adolescents of parents working unpredictable schedules or outside standard daytime working hours are more likely to have inferior cognition and behavioral outcomes." The parents might blame themselves for having a child who acts out or fails in school, but in many cases the real culprit is the poverty that leads workers to take jobs with haphazard schedules—and the scheduling models that squeeze struggling families even harder.

The root of the trouble, as with so many other WMDs, is the modelers' choice of objectives. The model is optimized for with notice of a week or less—often just a day or two, which can leave them scrambling to arrange transportation or child care.

Within weeks of the article's publication, the major corporations it mentioned announced that they would adjust their scheduling practices. Embarrassed by the story, the employers promised to add a single constraint to their model. They would eliminate clopenings and learn to live with slightly less robust optimization. Starbucks, whose brand hinges more than most on fair treatment of workers, went further, saying that the company would adjust the software to reduce the scheduling nightmares for its 130,000 baristas. All work hours would be posted at least one week in advance.

A year later, however, Starbucks was failing to meet these targets, or even to eliminate the clopenings, according to a follow-up report in the *Times*. The trouble was that minimal staffing was baked into the culture. In many companies, managers' pay is contingent upon the efficiency of their staff as measured by revenue per employee hour. Scheduling software helps them boost these numbers and their own compensation. Even when executives tell managers to loosen up, they often resist. It goes against everything they've been taught. What's more, at Starbucks, if a manager exceeds his or her "labor budget," a district manager is alerted, said one employee. And that could lead to a write-up. It's usually easier just to change someone's schedule, even if it means violating the corporate pledge to provide one week's notice.

In the end, the business models of publicly traded companies like Starbucks are built to feed the bottom line. That's reflected in their corporate cultures and their incentives, and, increasingly, in their operational software. (And if that software allows for tweaks, as Starbucks does, the ones that are made are likely to be ones that boost profits.)

Much of the scheduling technology has its roots in a powerful

discipline of applied mathematics called "operations research," or OR. For centuries, mathematicians used the rudiments of OR to help farmers plan crop plantings and help civil engineers map highways to move people and goods efficiently. But the discipline didn't really take off until World War II, when the US and British military enlisted teams of mathematicians to optimize their use of resources. The Allies kept track of various forms of an "exchange ratio," which compared Allied resources spent versus enemy resources destroyed. During Operation Starvation, which took place between March and August 1945, the Twenty-first Bomber Command was tasked with destroying Japanese merchant ships in order to prevent food and other goods from arriving safely on Japanese shores. OR teams worked to minimize the number of mine-laying aircraft for each Japanese merchant ship that was sunk. They managed an "exchange ratio" of over 40 to 1—only 15 aircraft were lost in sinking 606 Japanese ships. This was considered highly efficient, and was due, in part, to the work of the OR team.

Following World War II, major companies (as well as the Pentagon) poured enormous resources into OR. The science of logistics radically transformed the way we produce goods and bring them to market.

In the 1960s, Japanese auto companies made another major leap, devising a manufacturing system called Just in Time. The idea was that instead of storing mountains of steering wheels or transmission blocks and retrieving them from vast warehouses, the assembly plant would order parts as they were needed rather than paying for them to sit idle. Toyota and Honda established complex chains of suppliers, each of them constantly bringing in parts on call. It was as if the industry were a single organism, with its own homeostatic control systems.

Just in Time was highly efficient, and it quickly spread across hour on Friday. It throws their lives into chaos and wreaks havoc on child care plans. Meals are catch as catch can, as is sleep.

These irregular schedules are a product of the data economy. In the last chapter, we saw how WMDs sift through job candidates, blackballing some and ignoring many more. We saw how the software often encodes poisonous prejudices, learning from past records just how to be unfair. Here we continue the journey on to the job, where efficiency-focused WMDs treat workers as cogs in a machine. Clopening is just one product of this trend, which is likely to grow as surveillance extends into the workplace, providing more grist for the data economy.

For decades, before companies were swimming in data, scheduling was anything but a science. Imagine a family-owned hardware store whose clerks work from 9 to 5, six days a week. One year, the daughter goes to college. And when she comes back for the summer she sees the business with fresh eyes. She notices that practically no one comes to the store on Tuesday mornings. The clerk web-surfs on her phone, uninterrupted. That's a revenue drain. Meanwhile, on Saturdays, muttering customers wait in long lines.

These observations provide valuable data, and she helps her parents model the business to it. They start by closing the store on Tuesday mornings, and they hire a part-timer to help with the Saturday crush. These changes add a bit of intelligence to the dumb and inflexible status quo.

With Big Data, that college freshman is replaced by legions of PhDs with powerful computers in tow. Businesses can now analyze customer traffic to calculate exactly how many employees they will need each hour of the day. The goal, of course, is to spend as little money as possible, which means keeping staffing at the bare minimum while making sure that reinforcements are on hand for the busy times.

You might think that these patterns would repeat week after week, and that companies could simply make adjustments to their fixed schedules, just like the owners of our hypothetical hardware store. But new software scheduling programs offer far more sophisticated options. They process new streams of ever-changing data, from the weather to pedestrian patterns. A rainy afternoon, for example, will likely drive people from the park into cafés. So they'll need more staffing, at least for an hour or two. High school football on Friday night might mean more foot traffic on Main Street, but only before and after the game, not during it. Twitter volume suggests that 26 percent more shoppers will rush out to tomorrow's Black Friday sales than did last year. Conditions change, hour by hour, and the workforce must be deployed to match the fluctuating demand. Otherwise the company is wasting money.

The money saved, naturally, comes straight from employees' pockets. Under the inefficient status quo, workers had not only predictable hours but also a certain amount of downtime. You could argue that they benefited from inefficiency: some were able to read on the job, even study. Now, with software choreographing the work, every minute should be busy. And these minutes will come whenever the program demands it, even if it means clopening from Friday to Saturday.

In 2014, the *New York Times* ran a story about a harried single mother named Jannette Navarro, who was trying to work her way through college as a barista at Starbucks while caring for her four-year-old. The ever-changing schedule, including the occasional clopening, made her life almost impossible and put regular day care beyond reach. She had to put school on hold. The only thing she could schedule was work. And her story was typical. According to US government data, two-thirds of food service workers and more than half of retail workers find out about scheduling changes in twenty-seven regions of the brain. Usually, the conclusion of the phrenologist jibed with the observations he made. If a patient was morbidly anxious or suffering from alcoholism, the skull probe would usually find bumps and dips that correlated with that observation—which, in turn, bolstered faith in the science of phrenology.

Phrenology was a model that relied on pseudoscientific nonsense to make authoritative pronouncements, and for decades it went untested. Big Data can fall into the same trap. Models like the ones that red-lighted Kyle Behm and blackballed foreign medical students at St. George's can lock people out, even when the "science" inside them is little more than a bundle of untested assumptions.

7

SWEATING BULLETS

On the Job

Workers at major corporations in America recently came up with a new verb: *clopening*. That's when an employee works late one night to close the store or café and then returns a few hours later, before dawn, to open it. Having the same employee closing and opening, or clopening, often makes logistical sense for a company. But it leads to sleep-deprived workers and crazy schedules.

Wildly irregular schedules are becoming increasingly common, and they especially affect low-wage workers at companies like Starbucks, McDonald's, and Walmart. A lack of notice compounds the problem. Many employees find out only a day or two in advance that they'll have to work a Wednesday-night shift or handle rush that requires a far broader sweep of data and a more ambitious model.

A pioneer in this field is Gild, a San Francisco–based start-up. Extending far beyond a prospect's alma mater or résumé, Gild sorts through millions of job sites, analyzing what it calls each person's "social data." The company develops profiles of job candidates for its customers, mostly tech companies, keeping them up to date as the candidates add new skills. Gild claims that it can even predict when a star employee is likely to change jobs and can alert its customer companies when it's the right time to make an offer. But Gild's model attempts to quantify and also *qualify* each worker's "social capital." How integral is this person to the community of fellow programmers? Do they share and contribute code? Say a Brazilian coder—Pedro, let's call him—lives in São Paulo and spends every evening from dinner to one in the morning in communion with fellow coders the world over, solving cloud-computing problems or brainstorming gaming algorithms on sites like GitHub or Stack Overflow. The model could attempt to gauge Pedro's passion (which probably gets a high score) and his level of engagement with others. It would also evaluate the skill and social importance of his contacts. Those with larger followings would count for more. If his principal online contact happened to be Google's Sergey Brin, or Palmer Luckey, founder of the virtual reality maker Oculus VR, Pedro's social score would no doubt shoot through the roof.

But models like Gild's rarely receive such explicit signals from the data. So they cast a wider net, in search of correlations to workplace stardom wherever they can find them. And with more than six million coders in their database, the company can find all kinds of patterns. Vivienne Ming, Gild's chief scientist, said in an interview with *Atlantic Monthly* that Gild had found a bevy of talent frequenting a certain Japanese manga site. If Pedro spends

time at that comic-book site, of course, it doesn't predict superstardom. But it does nudge up his score.

That makes sense for Pedro. But certain workers might be doing something else offline, which even the most sophisticated algorithm couldn't infer—at least not today. They might be taking care of children, for example, or perhaps attending a book group. The fact that prospects don't spend six hours discussing manga every evening shouldn't be counted against them. And if, like most of techdom, that manga site is dominated by males and has a sexual tone, a good number of the women in the industry will probably avoid it.

Despites these issues, Gild is just one player. It doesn't have the clout of a global giant and is not positioned to set a single industry standard. Compared to some of the horrors we've seen—the predatory ads burying families in debt and the personality tests excluding people from opportunities—Gild is tame. Its category of predictive model has more to do with rewarding people than punishing them. No doubt the analysis is uneven: some potential stars are undoubtedly overlooked. But I don't think the talent miners yet rise to the level of a WMD.

Still, it's important to note that these hiring and "onboarding" models are ever-evolving. The world of data continues to expand, with each of us producing ever-growing streams of updates about our lives. All of this data will feed our potential employers, giving them insights into us.

Will those insights be tested, or simply used to justify the status quo and reinforce prejudices? When I consider the sloppy and self-serving ways that companies use data, I'm often reminded of phrenology, a pseudoscience that was briefly the rage in the nineteenth century. Phrenologists would run their fingers over the patient's skull, probing for bumps and indentations. Each one, they thought, was linked to personality traits that existed reject those candidates but instead to provide them with help—whether English classes or onsite day care—to pull them through.

This is a point I'll be returning to in future chapters: we've seen time and again that mathematical models can sift through data to locate people who are likely to face great challenges, whether from crime, poverty, or education. It's up to society whether to use that intelligence to reject and punish them—or to reach out to them with the resources they need. We can use the scale and efficiency that make WMDs so pernicious in order to help people. It all depends on the objective we choose.

...

So far in this chapter, we've been looking at models that filter out job candidates. For most companies, those WMDs are designed to cut administrative costs and to reduce the risk of bad hires (or ones that might require more training). The objective of the filters, in short, is to save money.

HR departments, of course, are also eager to save money through the hiring choices they make. One of the biggest expenses for a company is workforce turnover, commonly called churn. Replacing a worker earning $50,000 a year costs a company about $10,000, or 20 percent of that worker's yearly pay, according to the Center for American Progress. Replacing a high-level employee can cost multiples of that—as much as two years of salary.

Naturally, many hiring models attempt to calculate the likelihood that each job candidate will stick around. Evolv, Inc., now a part of Cornerstone OnDemand, helped Xerox scout out prospects for its calling center, which employs more than forty thousand people. The churn model took into account some of the metrics you might expect, including the average time people stuck around on previous jobs. But they also found some intriguing correlations. People the system classified as "creative types"

tended to stay longer at the job, while those who scored high on "inquisitiveness" were more likely to set their questioning minds toward other opportunities.

But the most problematic correlation had to do with geography. Job applicants who lived farther from the job were more likely to churn. This makes sense: long commutes are a pain. But Xerox managers noticed another correlation. Many of the people suffering those long commutes were coming from poor neighborhoods. So Xerox, to its credit, removed that highly correlated churn data from its model. The company sacrificed a bit of efficiency for fairness.

While churn analysis focuses on the candidates most likely to fail, the more strategically vital job for HR departments is to locate future stars, the people whose intelligence, inventiveness, and drive can change the course of an entire enterprise. In the higher echelons of the economy, companies are on the hunt for employees who think creatively and work well in teams. So the modelers' challenge is to pinpoint, in the vast world of Big Data, the bits of information that correlate with originality and social skills.

Résumés alone certainly don't cut it. Most of the items listed there—the prestigious university, the awards, even the skills—are crude proxies for high-quality work. While there's no doubt some correlation between tech prowess and a degree from a top school, it's far from perfect. Plenty of software talent comes from elsewhere—consider the high school hackers. What's more, résumés are full of puffery and sometimes even lies. With a quick search through LinkedIn or Facebook, a system can look further afield, identifying some of a candidate's friends and colleagues. But it's still hard to turn that data into a prediction that a certain engineer might be a perfect fit for a twelve-member consultancy in Palo Alto or Fort Worth. Finding the person to fill a role like up with its own automated assessment program in the late '70s, just as Apple was releasing its first personal computer, represented a bold experiment.

It turned out, however, to be an utter failure. St. George was not only precocious in its use of mathematical modeling, it seemed, but also an unwitting pioneer in WMDs.

As with so many WMDs, the problem began at the get-go, when the administrators established the model's twin objectives. The first was to boost efficiency, letting the machine handle much of the grunt work. It would automatically cull down the two thousand applications to five hundred, at which point humans would take over with a lengthy interviewing process. The second objective was fairness. The computer would remain unswayed by administrators' moods or prejudices, or by urgent entreaties from lords or cabinet ministers. In this first automatic screening, each applicant would be judged by the same criteria.

And what would those criteria be? That looked like the easy part. St. George's already had voluminous records of screenings from the previous years. The job was to teach the computerized system how to replicate the same procedures that human beings had been following. As I'm sure you can guess, these inputs were the problem. The computer learned from the humans how to discriminate, and it carried out this work with breathtaking efficiency.

In fairness to the administrators at St. George's, not all of the discrimination in the training data was overtly racist. A good number of the applications with foreign names, or from foreign addresses, came from people who clearly had not mastered the English language. Instead of considering the possibility that great doctors could learn English, which is obvious today, the tendency was simply to reject them. (After all, the school had to discard three-quarters of the applications, and that seemed like an easy place to start.)

Now, while the human beings at St. George's had long tossed out applications littered with grammatical mistakes and misspellings, the computer—illiterate itself—could hardly follow suit. But it could correlate the rejected applications of the past with birthplaces and, to a lesser degree, surnames. So people from certain places, like Africa, Pakistan, and immigrant neighborhoods of the United Kingdom, received lower overall scores and were not invited to interviews. An outsized proportion of these people were nonwhite. The human beings had also rejected female applicants, with the all-too-common justification that their careers would likely be interrupted by the duties of motherhood. The machine, naturally, did the same.

In 1988, the British government's Commission for Racial Equality found the medical school guilty of racial and gender discrimination in its admissions policy. As many as sixty of the two thousand applicants every year, according to the commission, may have been refused an interview purely because of their race, ethnicity, or gender.

The solution for the statisticians at St. George's—and for those in other industries—would be to build a digital version of a blind audition eliminating proxies such as geography, gender, race, or name to focus only on data relevant to medical education. The key is to analyze the skills each candidate brings to the school, not to judge him or her by comparison with people who seem similar. What's more, a bit of creative thinking at St. George's could have addressed the challenges facing women and foreigners. The *British Medical Journal* report accompanying the commission's judgment said as much. If language and child care issues posed problems for otherwise solid candidates, the solution was not to automatic systems to winnow down piles of résumés. In fact, some 72 percent of résumés are never seen by human eyes. Computer programs flip through them, pulling out the skills and experiences that the employer is looking for. Then they score each résumé as a match for the job opening. It's up to the people in the human resources department to decide where the cutoff is, but the more candidates they can eliminate with this first screening, the fewer human-hours they'll have to spend processing the top matches.

So job applicants must craft their résumés with that automatic reader in mind. It's important, for example, to sprinkle the résumé liberally with words the specific job opening is looking for. This could include positions (sales manager, chief financial officer, software architect), languages (Mandarin, Java), or honors (summa cum laude, Eagle Scout).

Those with the latest information learn what machines appreciate and what tangles them up. Images, for example, are useless. Most résumé scanners don't yet process them. And fancy fonts do nothing but confuse the machines, says Mona Abdel-Halim. She's the cofounder of Resunate.com, a job application tool. The safe ones, she says, are plain vanilla fonts, like Ariel and Courier. And forget about symbols such as arrows. They only confuse things, preventing the automatic systems from correctly parsing the information.

The result of these programs, much as with college admissions, is that those with the money and resources to prepare their résumés come out on top. Those who don't take these steps may never know that they're sending their résumés into a black hole. It's one more example in which the wealthy and informed get the edge and the poor are more likely to lose out.

To be fair, the résumé business has always had one sort of bias or another. In previous generations, those in the know were careful to organize the résumé items clearly and consistently, type them

on a quality computer, like an IBM Selectric, and print them on paper with a high rag content. Such résumés were more likely to make it past human screeners. More times than not, handwritten résumés, or ones with smudges from mimeograph machines, ended up in the circular file. So in this sense, the unequal paths to opportunity are nothing new. They have simply returned in a new incarnation, this time to guide society's winners past electronic gatekeepers.

The unequal treatment at the hands of these gatekeepers extends far beyond résumés. Our livelihoods increasingly depend on our ability to make our case to machines. The clearest example of this is Google. For businesses, whether it's a bed-and-breakfast or an auto repair shop, success hinges on showing up on the first page of search results. Now individuals face similar challenges, whether trying to get a foot in the door of a company, to climb the ranks—or even to survive waves of layoffs. The key is to learn what the machines are looking for. But here too, in a digital universe touted to be fair, scientific, and democratic, the insiders find a way to gain a crucial edge.

. . .

In the 1970s, the admissions office at St. George's Hospital Medical School, in the South London district of Tooting, saw an opportunity. They received more than twelve applications for each of their 150 openings each year. Combing through all those applications was a lot of work, requiring multiple screeners. And since each of those screeners had different ideas and predilections, the process was somewhat capricious. Would it be possible to program a computer to sort through the applications and reduce the field to a more manageable number?

Big organizations, like the Pentagon and IBM, were already using computers for such work. But for a medical school to come

I consider personality tests in hiring departments to be WMDs. They check all the boxes. First, they are in widespread use and have enormous impact. The Kronos exam, with all of its flaws, is scaled across much of the hiring economy. Under the previous status quo, employers no doubt had biases. But those biases varied from company to company, which might have cracked open a door somewhere for people like Kyle Behm. That's increasingly untrue. And Kyle was, in some sense, lucky. Job candidates, especially those applying for minimum-wage work, get rejected all the time and rarely find out why. It was just chance that Kyle's friend happened to hear about the reason for his rejection and told him about it. Even then, the case against the big Kronos users would likely have gone nowhere if Kyle's father hadn't been a lawyer, one with enough time and money to mount a broad legal challenge. This is rarely the case for low-level job applicants.*

Finally, consider the feedback loop that the Kronos personality test engenders. Red-lighting people with certain mental health issues prevents them from having a normal job and leading a normal life, further isolating them. This is exactly what the Americans with Disabilities Act is supposed to prevent.

▪ ▪ ▪

The majority of job applicants, thankfully, are not blackballed by automatic systems. But they still face the challenge of moving their application to the top of the pile and landing an interview. This has long been a problem for racial and ethnic minorities, as well as women.

In 2001 and 2002, before the expansion of automatic résumé

* Yes, it's true that many college-bound students labor for a summer or two in minimum-wage jobs. But if they have a miserable experience there, or are misjudged by an arbitrary WMD, it only reinforces the message that they should apply themselves at school and leave such hellish jobs behind.

readers, researchers from the University of Chicago and MIT sent out five thousand phony résumés for job openings advertised in the *Boston Globe* and the *Chicago Tribune*. The jobs ranged from clerical work to customer service and sales. Each of the résumés was modeled for race. Half featured typically white names like Emily Walsh and Brendan Baker, while the others with similar qualifications carried names like Lakisha Washington and Jamaal Jones, which would sound African American. The researchers found that the white names got 50 percent more callbacks than the black ones. But a secondary finding was perhaps even more striking. The white applicants with strong résumés got much more attention than whites with weaker ones; when it came to white applicants, it seemed, the hiring managers were paying attention. But among blacks, the stronger résumés barely made a difference. The hiring market, clearly, was still poisoned by prejudice.

The ideal way to circumvent such prejudice is to consider applicants blindly. Orchestras, which had long been dominated by men, famously started in the 1970s to hold auditions with the musician hidden behind a sheet. Connections and reputations suddenly counted for nothing. Nor did the musician's race or alma mater. The music from behind the sheet spoke for itself. Since then, the percentage of women playing in major orchestras has leapt by a factor of five—though they still make up only a quarter of the musicians.

The trouble is that few professions can engineer such an evenhanded tryout for job applicants. Musicians behind the sheet can actually perform the job they're applying for, whether it's a Dvorak cello concerto or bossa nova on guitar. In other professions, employers have to hunt through résumés, looking for qualities that might predict success.

As you might expect, human resources departments rely on take care of" *or* "Sometimes, I need a push to get started on my work."

The *Wall Street Journal* asked an industrial psychologist, Tomas Chamorro-Premuzic, to analyze thorny questions like these. The first item, Chamorro-Premuzic said, captured "individual differences in neuroticism and conscientiousness"; the second, "low ambition and drive." So the prospective worker is pleading guilty to being either high-strung or lazy.

A Kroger question was far simpler: Which adjective best describes you at work, unique or orderly?

Answering "unique," said Chamorro-Premuzic, captures "high self concept, openness and narcissism," while "orderly" expresses conscientiousness and self control.

Note that there's no option to answer "all of the above." Prospective workers must pick one option, without a clue as to how the program will interpret it. And some of the analysis will draw unflattering conclusions. If you go to a kindergarten class in much of the country, for example, you'll often hear teachers emphasize to the children that they're unique. It's an attempt to boost their self-esteem and, of course, it's true. Yet twelve years later, when that student chooses "unique" on a personality test while applying for a minimum-wage job, the program might read the answer as a red flag: Who wants a workforce peopled with narcissists?

Defenders of the tests note that they feature lots of questions and that no single answer can disqualify an applicant. Certain patterns of answers, however, can and do disqualify them. And we do not know what those patterns are. We're not told what the tests are looking for. The process is entirely opaque.

What's worse, after the model is calibrated by technical experts, it receives precious little feedback. Again, sports provide a good contrast here. Most professional basketball teams employ data geeks, who run models that analyze players by a series of metrics, including foot speed, vertical leap, free-throw percentage, and a host of other variables. When the draft comes, the Los Angeles Lakers might pass on a hotshot point guard from Duke because his assist statistics are low. Point guards have to be good passers. Yet in the following season they're dismayed to see that the rejected player goes on to win Rookie of the Year for the Utah Jazz and leads the league in assists. In such a case, the Lakers can return to their model to see what they got wrong. Maybe his college team was relying on him to score, which punished his assist numbers. Or perhaps he learned something important about passing in Utah. Whatever the case, they can work to improve their model.

Now imagine that Kyle Behm, after getting red-lighted at Kroger, goes on to land a job at McDonald's. He turns into a stellar employee. He's managing the kitchen within four months and the entire franchise a year later. Will anyone at Kroger go back to the personality test and investigate how they could have gotten it so wrong?

Not a chance, I'd say. The difference is this: Basketball teams are managing individuals, each one potentially worth millions of dollars. Their analytics engines are crucial to their competitive advantage, and they are hungry for data. Without constant feedback, their systems grow outdated and dumb. The companies hiring minimum-wage workers, by contrast, are managing herds. They slash expenses by replacing human resources professionals with machines, and those machines filter large populations into more manageable groups. Unless something goes haywire in the workforce—an outbreak of kleptomania, say, or plummeting productivity—the company has little reason to tweak the filtering model. It's doing its job—even if it misses out on potential stars.

The company may be satisfied with the status quo, but the victims of its automatic systems suffer. And as you might expect, As Kronos grew, it developed a broad range of software tools for workforce management, including a software program, Workforce Ready HR, that promised to eliminate "the guesswork" in hiring, according to its web page: "We can help you screen, hire, and onboard candidates most likely to be productive—the best-fit employees who will perform better and stay on the job longer."

Kronos is part of a burgeoning industry. The hiring business is automating, and many of the new programs include personality tests like the one Kyle Behm took. It is now a $500 million annual business and is growing by 10 to 15 percent a year, according to Hogan Assessment Systems Inc., a testing company. Such tests now are used on 60 to 70 percent of prospective workers in the United States, up from 30 to 40 percent about five years ago, estimates Josh Bersin of the consulting firm Deloitte.

Naturally, these hiring programs can't incorporate information about how the candidate would actually perform at the company. That's in the future, and therefore unknown. So like many other Big Data programs, they settle for proxies. And as we've seen, proxies are bound to be inexact and often unfair. In fact, the Supreme Court ruled in a 1971 case, *Griggs v. Duke Power Company*, that intelligence tests for hiring were discriminatory and therefore illegal. One would think that case might have triggered some soul-searching. But instead the industry simply opted for replacements, including personality tests like one that red-flagged Kyle Behm.

Even putting aside the issues of fairness and legality, research suggests that personality tests are poor predictors of job performance. Frank Schmidt, a business professor at the University of Iowa, analyzed a century of workplace productivity data to measure the predictive value of various selection processes. Personality tests ranked low on the scale—they were only one-third as predictive as cognitive exams, and also far below reference checks. This

is particularly galling because certain personality tests, research shows, can actually help employees gain insight into themselves. They can also be used for team building and for enhancing communication. After all, they create a situation in which people think explicitly about how to work together. That intention alone might end up creating a better working environment. In other words, if we define the goal as a happier worker, personality tests might end up being a useful tool.

But instead they're being used as a filter to weed out applicants. "The primary purpose of the test," said Roland Behm, "is not to find the best employee. It's to exclude as many people as possible as cheaply as possible."

You might think that personality tests would be easy to game. If you go online to take a Five Factor Personality Test, it looks like a cinch. One question asks: "Have frequent mood swings?" It would probably be smart to answer "very inaccurate." Another asks: "Get mad easily?" Again, check no. Not too many companies want to hire hotheads.

In fact, companies can get in trouble for screening out applicants on the basis of such questions. Regulators in Rhode Island found that CVS Pharmacy was illegally screening out applicants with mental illnesses when a personality test required respondents to agree or disagree to such statements as "People do a lot of things that make you angry" and "There's no use having close friends; they always let you down." More intricate questions, which are harder to game, are more likely to keep the companies out of trouble. Consequently, many of the tests used today force applicants to make difficult choices, likely leaving them with a sinking feeling of "Damned if I do, damned if I don't."

McDonald's, for example, asked prospective workers to choose which of the following best described them:

"It is difficult to be cheerful when there are many problems to by the personality test he'd taken when he applied for the job. The test was part of an employee selection program developed by Kronos, a workforce management company based outside of Boston. When Kyle told his father, Roland, an attorney, what had happened, his father asked him what kind of questions had appeared on the test. Kyle said that they were very much like the "Five Factor Model" test, which he'd been given at the hospital. That test grades people for extraversion, agreeableness, conscientiousness, neuroticism, and openness to ideas.

At first, losing one minimum-wage job because of a questionable test didn't seem like such a big deal. Roland Behm urged his son to apply elsewhere. But Kyle came back each time with the same news. The companies he was applying to were all using the same test, and he wasn't getting offers. Roland later recalled: "Kyle said to me, 'I had an almost perfect SAT and I was at Vanderbilt a few years ago. If I can't get a part-time minimum-wage job, how broken am I?' And I said, 'I don't think you're that broken.'"

But Roland Behm was bewildered. Questions about mental health appeared to be blackballing his son from the job market. He decided to look into it and soon learned that the use of personality tests for hiring was indeed widespread among large corporations. And yet he found very few legal challenges to this practice. As he explained to me, people who apply for a job and are red-lighted rarely learn that they were rejected because of their test results. Even when they do, they're not likely to contact a lawyer.

Behm went on to send notices to seven companies—Finish Line, Home Depot, Kroger, Lowe's, PetSmart, Walgreen Co., and Yum Brands—informing them of his intent to file a class-action suit alleging that the use of the exam during the job application process was unlawful.

The suit, as I write this, is still pending. Arguments are likely to focus on whether the Kronos test can be considered a medical exam, the use of which in hiring is illegal under the Americans with Disabilities Act of 1990. If this turns out to be the case, the court will have to determine whether the hiring companies themselves are responsible for running afoul of the ADA, or if Kronos is.

The question for this book is how automatic systems judge us when we seek jobs and what criteria they evaluate. Already, we've seen WMDs poisoning the college admissions process, both for the rich and for the middle class. Meanwhile, WMDs in criminal justice rope in millions, the great majority of them poor, most of whom never had the chance to attend college at all. Members of each of these groups face radically different challenges. But they have something in common, too. They all ultimately need a job.

Finding work used to be largely a question of whom you knew. In fact, Kyle Behm was following the traditional route when he applied for work at Kroger. His friend had alerted him to the opening and put in a good word. For decades, that was how people got a foot in the door, whether at grocers, the docks, banks, or law firms. Candidates then usually faced an interview, where a manager would try to get a feel for them. All too often this translated into a single basic judgment: Is this person like me (or others I get along with)? The result was a lack of opportunity for job seekers without a friend inside, especially if they came from a different race, ethnic group, or religion. Women also found themselves excluded by this insider game.

Companies like Kronos brought science into corporate human resources in part to make the process fairer. Founded in the 1970s by MIT graduates, Kronos's first product was a new kind of punch clock, one equipped with a microprocessor, which added up employees' hours and reported them automatically. This may sound banal, but it was the beginning of the electronic push (now blazing along at warp speed) to track and optimize a workforce.

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.

6

INELIGIBLE TO SERVE

Getting a Job

A few years ago, a young man named Kyle Behm took a leave from his studies at Vanderbilt University. He was suffering from bipolar disorder and needed time to get treatment. A year and a half later, Kyle was healthy enough to return to his studies at a different school. Around that time, he learned from a friend about a part-time job at Kroger. It was just a minimum-wage job at a supermarket, but it seemed like a sure thing. His friend, who was leaving the job, could vouch for him. For a high-achieving student like Kyle, the application looked like a formality.

But Kyle didn't get called back for an interview. When he inquired, his friend explained to him that he had been "red-lighted" 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-one 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.

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

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

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

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

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

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

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

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

Published in the United States by Broadway Books, an imprint of
the Crown Publishing Group, a division of Penguin Random House LLC, New York.
crownpublishing.com

Broadway Books and its logo, B \ D \ W \ Y, are trademarks of
Penguin Random House LLC.

Originally published in hardcover in the United States by Crown,
an imprint of the Crown Publishing Group, a division of
Penguin Random House LLC, New York, in 2016.

Library of Congress Cataloging-in-Publication Data
Name: O'Neil, Cathy, author.
Title: Weapons of math destruction: how big data increases inequality and threatens democracy / Cathy O'Neil
Description: First edition. | New York: Crown Publishers [2016]
Identifiers: LCCN 2016003900 (print) | LCCN 2016016487 (ebook) | ISBN 9780553418811 (hardcover) | ISBN 9780553418835 (pbk.) | ISBN 9780553418828 (ebook)
Subjects: LCSH: Big data—Social aspects—United States. | Big data—Political aspects—United States. | Social indicators—Mathematical models—Moral and ethical aspects. | Democracy—United States. | United States—Social conditions—21st century.
Classification: LCC QA76.9.B45 064 2016 (print) | LCC QA76.9.B45 (ebook) | DDC 005.7—dc23
LC record available at https://lccn.loc.gov/2016003900

ISBN 978-0-553-41883-5
Ebook ISBN 978-0-553-41882-8
International Edition ISBN 978-0-451-49733-8

PRINTED IN THE UNITED STATES OF AMERICA

Book design by Ralph Fowler
Cover design by Elena Giavaldi

10

First Paperback Edition

THIS BOOK IS DEDICATED TO

ALL THE UNDERDOGS

"O'Neil's vision of a world run by algorithms is laced with dark humor and exasperation—like a modern-day *Dr. Strangelove* or *Catch-22*. It is eye-opening, disturbing, and deeply important."

—Steven Strogatz, Cornell University, author of *The Joy of x*

"The stuff of technophobes' nightmares . . . Unlike some other recent books on data collection, [O'Neil's] is not hysterical; she offers more of a chilly wake-up call as she walks readers through the ways the 'big data' industry has facilitated social ills."

—*Publishers Weekly*

"There's a little Philip K. Dick, a little Orwell, a little Kafka in her portrait of powerful bureaucracies ceding control of the most intimate decisions of our lives to hyperempowered computer models riddled with all of our unresolved, atavistic human biases."

—*Paris Review*

"Combating secrecy with clarity and confusion with understanding, this book can help us change course before it's too late."

—Astra Taylor, author of *The People's Platform*

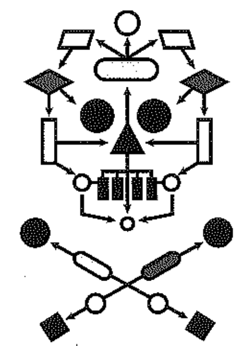
"A fantastic, plainspoken call to arms."

—Cory Doctorow, author of *Little Brother* and coeditor of *Boing Boing*

"[O'Neil's] analysis is superb, her writing is enticing, and her findings are unsettling."

—Danah Boyd, founder of Data & Society and author of *It's Complicated*

WEAPONS OF MATH DESTRUCTION



HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY

CATHY O'NEIL

B\D\W\Y
BROADWAY BOOKS
NEW YORK