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 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 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 cloopenings 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 cloopenings, 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 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 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 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 price less 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. 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.

“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 damming 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.”

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 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.
* 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.