

# Problems with Probability

## *How overconfident math geeks nearly destroyed the global financial system*

Statistics cannot be any smarter than the people who use them. And in some cases, they can make smart people do dumb things. One of the most irresponsible uses of statistics in recent memory involved the mechanism for gauging risk on Wall Street prior to the 2008 financial crisis. At that time, firms throughout the financial industry used a common barometer of risk, the Value at Risk model, or VaR. In theory, VaR combined the elegance of an indicator (collapsing lots of information into a single number) with the power of probability (attaching an expected gain or loss to each of the firm's assets or trading positions). The model assumed that there is a range of possible outcomes for every one of the firm's investments. For example, if the firm owns General Electric stock, the value of those shares can go up or down. When the VaR is being calculated for some short period of time, say, one week, the most likely outcome is that the shares will have roughly the same value at the end of that stretch as they had at the beginning. There is a smaller chance that the shares may rise or fall by 10 percent. And an even smaller chance that they may rise or fall 25 percent, and so on.

On the basis of past data for market movements, the firm's quantitative experts (often called "quants" in the industry and "rich nerds" everywhere else) could assign a dollar figure, say \$13 million, that represented the maximum that the firm could lose on that position over the

time period being examined, with 99 percent probability. In other words, 99 times out of 100 the firm would not lose more than \$13 million on a particular trading position; 1 time out of 100, it would.

Remember that last part, because it will soon become important.

Prior to the financial crisis of 2008, firms trusted the VaR model to quantify their overall risk. If a single trader had 923 different open positions (investments that could move up or down in value), each of those investments could be evaluated as described above for the General Electric stock; from there, the trader's total portfolio risk could be calculated. The formula even took into account the correlations among different positions. For example, if two investments had expected returns that were negatively-correlated, a loss in one would likely have been offset by a gain in the other, making the two investments together less risky than either one separately. Overall, the head of the trading desk would know that bond trader Bob Smith has a 24-hour VaR (the value at risk over the next 24 hours) of \$19 million, again with 99 percent probability. The most that Bob Smith could lose over the next 24 hours would be \$19 million, 99 times out of 100.

Then, even better, the aggregate risk for the firm could be calculated at any point in time by taking the same basic process one step further. The underlying mathematical mechanics are obviously fabulously complicated, as firms had a dizzying array of investments in different currencies, with different amounts of leverage (the amount of money that was borrowed to make the investment), trading in markets with different degrees of liquidity, and so on. Despite all that, the firm's managers ostensibly had a precise measure of the magnitude of the risk that the firm had taken on at any moment in time. As former *New York Times* business writer Joe Nocera has explained, "VaR's great appeal, and its great selling point to people who do not happen to be quants, is that it expresses risk as a single number, a dollar figure, no less."<sup>1</sup> At J. P. Morgan, where the VaR model was developed and refined, the daily VaR calculation was known as the "4:15 report" because it would be on the desks of top executives every afternoon at 4:15, just after the American financial markets had closed for the day.

Presumably this was a good thing, as more information is gener-

ally better, particularly when it comes to risk. After all, probability is a powerful tool. Isn't this just the same kind of calculation that the Schlitz executives did before spending a lot of money on blind taste tests at halftime of the Super Bowl?

Not necessarily. VaR has been called "potentially catastrophic," "a fraud," and many other things not fit for a family book about statistics like this one. In particular, the model has been blamed for the onset and severity of the financial crisis. The primary critique of VaR is that the underlying risks associated with financial markets are not as predictable as a coin flip or even a blind taste test between two beers. The false precision embedded in the models created a false sense of security. The VaR was like a faulty speedometer, which is arguably worse than no speedometer at all. If you place too much faith in the broken speedometer, you will be oblivious to other signs that your speed is unsafe. In contrast, if there is no speedometer at all, you have no choice but to look around for clues as to how fast you are really going.

By around 2005, with the VaR dropping on desks at 4:15 every weekday, Wall Street was driving pretty darn fast. Unfortunately, there were two huge problems with the risk profiles encapsulated by the VaR models. First, the underlying probabilities on which the models were built were based on past market movements; however, in financial markets (unlike beer tasting), the future does not necessarily look like the past. There was no intellectual justification for assuming that the market movements from 1980 to 2005 were the best predictor of market movements after 2005. In some ways, this failure of imagination resembles the military's periodic mistaken assumption that the next war will look like the last one. In the 1990s and early 2000s, commercial banks were using lending models for home mortgages that assigned zero probability to large declines in housing prices.<sup>2</sup> Housing prices had never before fallen as far and as fast as they did beginning in 2007. But that's what happened. Former Federal Reserve chairman Alan Greenspan explained to a congressional committee after the fact, "The whole intellectual edifice, however, collapsed in the summer of [2007] because the data input into the risk management models generally covered only the past two decades, a period of euphoria. Had instead the models been fitted more appropriately to historic periods

of stress, capital requirements would have been much higher and the financial world would be in far better shape, in my judgment.”<sup>3</sup>

Second, even if the underlying data could accurately predict future risk, the 99 percent assurance offered by the VaR model was dangerously useless, *because it's the 1 percent that is going to really mess you up*. Hedge fund manager David Einhorn explained, “This is like an air bag that works all the time, except when you have a car accident.” If a firm has a Value at Risk of \$500 million, that can be interpreted to mean that the firm has a 99 percent chance of losing no more than \$500 million over the time period specified. Well, hello, that also means that the firm has a 1 percent chance of losing more than \$500 million—much, much more under some circumstances. In fact, the models had nothing to say about how bad that 1 percent scenario might turn out to be. Very little attention was devoted to the “tail risk,” the small risk (named for the tail of the distribution) of some catastrophic outcome. (If you drive home from a bar with a blood alcohol level of .15, there is probably less than a 1 percent chance that you will crash and die; that does not make it a sensible thing to do.) Many firms compounded this error by making unrealistic assumptions about their preparedness for rare events. Former treasury secretary Hank Paulson has explained that many firms assumed they could raise cash in a pinch by selling assets.<sup>4</sup> But during a crisis, every other firm needs cash, too, so all are trying to sell the same kinds of assets. It's the risk management equivalent of saying, “I don't need to stock up on water because if there is a natural disaster, I'll just go to the supermarket and buy some.” Of course, after an asteroid hits your town, fifty thousand other people are also trying to buy water; by the time you get to the supermarket, the windows are broken and the shelves are empty.

The fact that you've never contemplated that your town might be flattened by a massive asteroid was exactly the problem with VaR. Here is *New York Times* columnist Joe Nocera again, summarizing thoughts of Nicholas Taleb, author of *The Black Swan: The Impact of the Highly Improbable* and a scathing critic of VaR: “The greatest risks are never the ones you can see and measure, but the ones you can't see and therefore can never measure. The ones that seem so far outside the boundary of normal probability that you can't imagine they could happen in your

lifetime—even though, of course, they do happen, more often than you care to realize.”

In some ways, the VaR debacle is the opposite of the Schlitiz example in Chapter 5. Schlitiz was operating with a known probability distribution. Whatever data the company had on the likelihood of blind taste testers' choosing Schlitiz was a good estimate of how similar testers would behave live at halftime. Schlitiz even managed its downside by performing the whole test on men who said they liked the other beers better. Even if no more than twenty-five Michelob drinkers chose Schlitiz (an almost impossibly low outcome), Schlitiz could still claim that one in four beer drinkers ought to consider switching. Perhaps most important, this was all just beer, not the global financial system. The Wall Street quants made three fundamental errors. First, they confused precision with accuracy. The VaR models were just like my golf range finder when it was set to meters instead of yards: exact and wrong. The false precision led Wall Street executives to believe that they had risk on a leash when in fact they did not. Second, the estimates of the underlying probabilities were wrong. As Alan Greenspan pointed out in testimony quoted earlier in the chapter, the relatively tranquil and prosperous decades before 2005 should not have been used to create probability distributions for what might happen in the markets in the ensuing decades. This is the equivalent of walking into a casino and thinking that you will win at roulette 62 percent of the time because that's what happened last time you went gambling. It would be a long, expensive evening. Third, firms neglected their “tail risk.” The VaR models predicted what would happen 99 times out of 100. That's the way probability works (as the second half of the book will emphasize repeatedly). Unlikely things happen. In fact, over a long enough period of time, they are not even that unlikely. People get hit by lightning all the time. My mother has had three holes in one.

The statistical hubris at commercial banks and on Wall Street ultimately contributed to the most severe global financial contraction since the Great Depression. The crisis that began in 2008 destroyed trillions of dollars in wealth in the United States, drove unemployment over 10 percent, created waves of home foreclosures and business failures, and saddled governments around the world with huge debts as they struggled

to contain the economic damage. This is a sadly ironic outcome, given that sophisticated tools like VaR were designed to mitigate risk.

Probability offers a powerful and useful set of tools—many of which can be employed correctly to understand the world or incorrectly to wreak havoc on it. In sticking with the “statistics as a powerful weapon” metaphor that I’ve used throughout the book, I will paraphrase the gun rights lobby: Probability doesn’t make mistakes; people using probability make mistakes. The balance of this chapter will catalog some of the most common probability-related errors, misunderstandings, and ethical dilemmas.

*Assuming events are independent when they are not.* The probability of flipping heads with a fair coin is  $\frac{1}{2}$ . The probability of flipping two heads in a row is  $(\frac{1}{2})^2$ , or  $\frac{1}{4}$ , since the likelihood of two independent events’ both happening is the product of their individual probabilities. Now that you are armed with this powerful knowledge, let’s assume that you have been promoted to head of risk management at a major airline. Your assistant informs you that the probability of a jet engine’s failing for any reason during a transatlantic flight is 1 in 100,000. Given the number of transatlantic flights, this is not an acceptable risk. Fortunately each jet making such a trip has at least two engines. Your assistant has calculated that the risk of both engines’ shutting down over the Atlantic must be  $(1/100,000)^2$ , or 1 in 10 billion, which is a reasonable safety risk. This would be a good time to tell your assistant to use up his vacation days before he is fired. The two engine failures are not independent events. If a plane flies through a flock of geese while taking off, both engines are likely to be compromised in a similar way. The same would be true of many other factors that affect the performance of a jet engine, from weather to improper maintenance. If one engine fails, the probability that the second engine fails is going to be significantly higher than 1 in 100,000.

Does this seem obvious? It was not obvious throughout the 1990s as British prosecutors committed a grave miscarriage of justice because of an improper use of probability. As with the hypothetical jet engine example, the statistical mistake was in assuming that several events were

independent (as in flipping a coin) rather than dependent (when a certain outcome makes a similar outcome more likely in the future). This mistake was real, however, and innocent people were sent to jail as a result.

The mistake arose in the context of sudden infant death syndrome (SIDS), a phenomenon in which a perfectly healthy infant dies in his or her crib. (The Brits refer to SIDS as a “cot death.”) SIDS was a medical mystery that attracted more attention as infant deaths from other causes became less common.\* Because these infant deaths were so mysterious and poorly understood, they bred suspicion. Sometimes that suspicion was warranted. SIDS was used on occasion to cover up parental negligence or abuse; a postmortem exam cannot necessarily distinguish natural deaths from those in which foul play is involved. British prosecutors and courts became convinced that one way to separate foul play from natural deaths would be to focus on families in which there were multiple cot deaths. Sir Roy Meadow, a prominent British pediatrician, was a frequent expert witness on this point. As the British news magazine the *Economist* explains, “What became known as Meadow’s Law—the idea that one infant death is a tragedy, two are suspicious and three are murder—is based on the notion that if an event is rare, two or more instances of it in the same family are so improbable that they are unlikely to be the result of chance.”<sup>5</sup> Sir Meadow explained to juries that the chance that a family could have two infants die suddenly of natural causes was an extraordinary 1 in 73 million. He explained the calculation: Since the incidence of a cot death is rare, 1 in 8,500, the chance of having two cot deaths in the same family would be  $(1/8,500)^2$  which is roughly 1 in 73 million. This reeks of foul play. That’s what juries decided, sending many parents to prison on the basis of this testimony on the statistics of cot deaths (often without any corroborating medical evidence of abuse or neglect). In some cases, infants were taken away from their parents at birth because of the unexplained death of a sibling.

\* SIDS is still a medical mystery, though many of the risk factors have been identified. For example, infant deaths can be reduced sharply by putting babies to sleep on their backs.

The *Economist* explained how a misunderstanding of statistical independence became a flaw in the Meadow testimony:

There is an obvious flaw in this reasoning, as the Royal Statistical Society, protective of its derided subject, has pointed out. The probability calculation works fine, so long as it is certain that cot deaths are entirely random and not linked by some unknown factor. But with something as mysterious as cot deaths, it is quite possible that there is a link—something genetic, for instance, which would make a family that had suffered one cot death more, not less, likely to suffer another. And since those women were convicted, scientists have been suggesting that there may be just such a link.

In 2004, the British government announced that it would review 258 trials in which parents had been convicted of murdering their infant children.

**Not understanding when events ARE independent.** A different kind of mistake occurs when events that *are* independent are not treated as such. If you find yourself in a casino (a place, statistically speaking, that you should not go to), you will see people looking longingly at the dice or cards and declaring that they are “due.” If the roulette ball has landed on black five times in a row, then clearly now it must turn up red. No, no, no! The probability of the ball’s landing on a red number remains unchanged: 16/38. The belief otherwise is sometimes called “the gambler’s fallacy.” In fact, if you flip a fair coin 1,000,000 times and get 1,000,000 heads in a row, the probability of getting tails on the next flip is still  $\frac{1}{2}$ . The very definition of statistical independence between two events is that the outcome of one has no effect on the outcome of the other. Even if you don’t find the statistics persuasive, you might ask yourself about the physics: How can flipping a series of tails in a row make it more likely that the coin will turn up heads on the next flip?

Even in sports, the notion of streaks may be illusory. One of the most famous and interesting probability-related academic papers refutes the common notion that basketball players periodically develop a streak of good shooting during a game, or “a hot hand.” Certainly most sports

fans would tell you that a player who makes a shot is more likely to hit the next shot than a player who has just missed. Not according to research by Thomas Gilovich, Robert Vallone, and Amos Tversky, who tested the hot hand in three different ways.<sup>6</sup> First, they analyzed shooting data for the Philadelphia 76ers home games during the 1980–81 season. (At the time, similar data were not available for other teams in the NBA.) They found “no evidence for a positive correlation between the outcomes of successive shots.” Second, they did the same thing for free throw data for the Boston Celtics, which produced the same result. And last, they did a controlled experiment with members of the Cornell men’s and women’s basketball teams. The players hit an average of 48 percent of their field goals after hitting their last shot and 47 percent after missing. For fourteen of twenty-six players, the correlation between making one shot and then making the next was negative. Only one player showed a significant positive correlation between one shot and the next.

That’s not what most basketball fans will tell you. For example, 91 percent of basketball fans surveyed at Stanford and Cornell by the authors of the paper agreed with the statement that a player has a better chance of making his next shot after making his last two or three shots than he does after missing his last two or three shots. The significance of the “hot hand” paper lies in the difference between the perception and the empirical reality. The authors note, “People’s intuitive conceptions of randomness depart systematically from the laws of chance.” We see patterns where none may really exist.

Like cancer clusters.

**Clusters happen.** You’ve probably read the story in the newspaper, or perhaps seen the news exposé: Some statistically unlikely number of people in a particular area have contracted a rare form of cancer. It must be the water, or the local power plant, or the cell phone tower. Of course, any one of those things might really be causing adverse health outcomes. (Later chapters will explore how statistics can identify such causal relationships.) But this cluster of cases may also be the product of pure chance, even when the number of cases appears highly improbable. Yes, the probability that five people in the same school or church or workplace

will contract the same rare form of leukemia may be one in a million, *but there are millions of schools and churches and workplaces*. It's not highly improbable that five people might get the same rare form of leukemia in one of those places. We just aren't thinking about all the schools and churches and workplaces where this hasn't happened. To use a different variation on the same basic example, the chance of winning the lotto may be 1 in 20 million, but none of us is surprised when *someone* wins, because millions of tickets have been sold. (Despite my general aversion to lotteries, I do admire the Illinois slogan: "Someone's gonna Lotto, might as well be you.")

Here is an exercise that I do with my students to make the same basic point. The larger the class, the better it works. I ask everyone in the class to take out a coin and stand up. We all flip the coin; anyone who flips heads must sit down. Assuming we start with 100 students, roughly 50 will sit down after the first flip. Then we do it again, after which 25 or so are still standing. And so on. More often than not, there will be a student standing at the end who has flipped five or six tails in a row. At that point, I ask the student questions like "How did you do it?" and "What are the best training exercises for flipping so many tails in a row?" or "Is there a special diet that helped you pull off this impressive accomplishment?" These questions elicit laughter because the class has just watched the whole process unfold; they know that the student who flipped six tails in a row has no special coin-flipping talent. He or she just happened to be the one who ended up with a lot of tails. When we see an anomalous event like that out of context, however, we assume that something besides randomness must be responsible.

**The prosecutor's fallacy.** Suppose you hear testimony in court to the following effect: (1) a DNA sample found at the scene of a crime matches a sample taken from the defendant; and (2) there is only one chance in a million that the sample recovered at the scene of the crime would match anyone's besides the defendant. (For the sake of this example, you can assume that the prosecution's probabilities are correct.) On the basis of that evidence, would you vote to convict?

I sure hope not.

The prosecutor's fallacy occurs when the context surrounding statistical evidence is neglected. Here are two scenarios, each of which could explain the DNA evidence being used to prosecute the defendant.

**Defendant 1:** This defendant, a spurned lover of the victim, was arrested three blocks from the crime scene carrying the murder weapon. After he was arrested, the court compelled him to offer a DNA sample, which matched a sample taken from a hair found at the scene of the crime.

**Defendant 2:** This defendant was convicted of a similar crime in a different state several years ago. As a result of that conviction, his DNA was included in a national DNA database of over a million violent felons. The DNA sample taken from the hair found at the scene of the crime was run through that database and matched to this individual, who has no known association with the victim.

As noted above, in both cases the prosecutor can rightfully say that the DNA sample taken from the crime scene matches the defendant's and that there is only one in a million chance that it would match with anyone else's. But in the case of Defendant 2, there is a darn good chance that he could be that random someone else, the one in a million guy whose DNA just happens to be similar to the real killer's by chance. *Because the chances of finding a coincidental one in a million match are relatively high if you run the sample through a database with samples from a million people.*

**Reversion to the mean (or regression to the mean).** Perhaps you've heard of the *Sports Illustrated* jinx, whereby individual athletes or teams featured on the cover of *Sports Illustrated* subsequently see their performance fall off. One explanation is that being on the cover of the magazine has some adverse effect on subsequent performance. The more statistically sound explanation is that teams and athletes appear on its cover after some anomalously good stretch (such as a twenty-game winning streak) and that their subsequent performance merely reverts back to what is normal, or the mean. This is the phenomenon known as reversion to the mean. Probability tells us that any outlier—an observation that is particularly far from the mean in one direction or the other—is likely to be followed by outcomes that are more consistent with the long-term average.

Reversion to the mean can explain why the Chicago Cubs always

seem to pay huge salaries for free agents who subsequently disappoint fans like me. Players are able to negotiate huge salaries with the Cubs after an exceptional season or two. Putting on a Cubs uniform does not necessarily make these players worse (though I would not necessarily rule that out); rather, the Cubs pay big bucks for these superstars at the end of some exceptional stretch—an outlier year or two—after which their performance for the Cubs reverts to something closer to normal.

The same phenomenon can explain why students who do much better than they normally do on some kind of test will, on average, do slightly worse on a retest, and students who have done worse than usual will tend to do slightly better when retested. One way to think about this mean reversion is that performance—both mental and physical—consists of some underlying talent-related effort plus an element of luck, good or bad. (Statisticians would call this random error.) In any case, those individuals who perform far above the mean for some stretch are likely to have had luck on their side; those who perform far below the mean are likely to have had bad luck. (In the case of an exam, think about students guessing right or wrong; in the case of a baseball player, think about a hit that can either go foul or land one foot fair for a triple.) When a spell of very good luck or very bad luck ends—as it inevitably will—the resulting performance will be closer to the mean.

Imagine that I am trying to assemble a superstar coin-flipping team (under the erroneous impression that talent matters when it comes to coin flipping). After I observe a student flipping six tails in a row, I offer him a ten-year, \$50 million contract. Needless to say, I'm going to be disappointed when this student flips only 50 percent tails over those ten years.

At first glance, reversion to the mean may appear to be at odds with the “gambler’s fallacy.” After the student throws six tails in a row, is he “due” to throw heads or not? The probability that he throws heads on the next flip is the same as it always is:  $\frac{1}{2}$ . The fact that he has thrown lots of tails in a row does not make heads more likely on the next flip. Each flip is an independent event. However, we can expect the results of the ensuing flips to be consistent with what probability predicts, which is half heads and half tails, rather than what it has been in the past, which is all tails. It’s a virtual certainty that someone who has flipped all tails will

begin throwing more heads in the ensuing 10, 20, or 100 flips. And the more flips, the more closely the outcome will resemble the 50-50 mean outcome that the law of large numbers predicts. (Or, alternatively, we should start looking for evidence of fraud.)

As a curious side note, researchers have also documented a *Businessweek* phenomenon. When CEOs receive high-profile awards, including being named one of *Businessweek*’s “Best Managers,” their companies subsequently underperform over the next three years as measured by both accounting profits and stock price. However, unlike the *Sports Illustrated* effect, this effect appears to be more than reversion to the mean. According to Ulrike Malmendier and Geoffrey Tate, economists at the University of California at Berkeley and UCLA, respectively, when CEOs achieve “superstar” status, they get distracted by their new prominence.<sup>7</sup> They write their memoirs. They are invited to sit on outside boards. They begin searching for trophy spouses. (The authors propose only the first two explanations, but I find the last one plausible as well.) Malmendier and Tate write, “Our results suggest that media-induced superstar culture leads to behavioral distortions beyond mere mean reversion.” In other words, when a CEO appears on the cover of *Businessweek*, sell the stock.

**Statistical discrimination.** When is it okay to act on the basis of what probability tells us is likely to happen, and when is it not okay? In 2003, Anna Diamantopoulou, the European commissioner for employment and social affairs, proposed a directive declaring that insurance companies may not charge different rates to men and women, because it violates the European Union’s principle of equal treatment.<sup>8</sup> To insurers, however, gender-based premiums aren’t discrimination; they’re just statistics. Men typically pay more for auto insurance because they crash more. Women pay more for annuities (a financial product that pays a fixed monthly or yearly sum until death) because they live longer. Obviously many women crash more than many men, and many men live longer than many women. But, as explained in the last chapter, insurance companies don’t care about that. They care only about what happens on average, because if they get that right, the firm will make money. The interesting thing

about the European Commission policy banning gender-based insurance premiums, which is being implemented in 2012, is that the authorities are not pretending that gender is unrelated to the risks being insured; they are simply declaring that disparate rates based on sex are unacceptable.\*

At first, that feels like an annoying nod to political correctness. Upon reflection, I'm not so sure. Remember all that impressive stuff about preventing crimes before they happen? Probability can lead us to some intriguing but distressing places in this regard. How should we react when our probability-based models tell us that methamphetamine smugglers from Mexico are most likely to be Hispanic men aged between eighteen and thirty and driving red pickup trucks between 9:00 p.m. and midnight when we also know that the vast majority of Hispanic men who fit that profile are not smuggling methamphetamine? Yep, I used the profiling word, because that's the less glamorous description of the predictive analytics that I described so glowingly in the last chapter, or at least one potential aspect of it.

Probability tells us what is more likely and what is less likely. Yes, that is just basic statistics—the tools described over the last few chapters. But it is also statistics with social implications. If we want to catch violent criminals and terrorists and drug smugglers and other individuals with the potential to do enormous harm, then we ought to use every tool at our disposal. Probability can be one of those tools. It would be naive to think that gender, age, race, ethnicity, religion, and country of origin collectively tell us nothing about anything related to law enforcement.

But what we can or should do with that kind of information (assuming it has some predictive value) is a philosophical and legal question, not a statistical one. We're getting more and more information every day about more and more things. Is it okay to discriminate if the data tell us that we'll be right far more often than we're wrong? (This is the origin of the term "statistical discrimination," or "rational discrimination.") The

\* The policy change was ultimately precipitated by a 2011 ruling by the Court of Justice of the European Union that different premiums for men and women constitute sex discrimination.

same kind of analysis that can be used to determine that people who buy birdseed are less likely to default on their credit cards (yes, that's really true) can be applied everywhere else in life. How much of that is acceptable? If we can build a model that identifies drug smugglers correctly 80 out of 100 times, what happens to the poor souls in the 20 percent—*because our model is going to harass them over and over and over again.*

The broader point here is that our ability to analyze data has grown far more sophisticated than our thinking about what we ought to do with the results. You can agree or disagree with the European Commission decision to ban gender-based insurance premiums, but I promise you it will not be the last tricky decision of that sort. We like to think of numbers as "cold, hard facts." If we do the calculations right, then we must have *the right answer*. The more interesting and dangerous reality is that we can sometimes do the calculations correctly and end up blundering in a dangerous direction. We can blow up the financial system or harass a twenty-two-year-old white guy standing on a particular street corner at a particular time of day, because, according to our statistical model, he is almost certainly there to buy drugs. For all the elegance and precision of probability, there is no substitute for thinking about what calculations we are doing and why we are doing them.