CARL T. BERGSTROM
& JEVIN D. WEST

Calling Bullshit

The Art of Scepticism in a Data-Driven World

ALLEN LANE
an imprint of

ALLEN LANE

UK | USA | Canada | Ireland | Australia
India | New Zealand | South Africa

Allen Lane is part of the Penguin Random House group of companies
whose addresses can be found at global.penguinrandomhouse.com.

Penguin
Random House
UK

First published in the United States by Random House, an imprint
and division of Penguin Random House LLC, New York, 2020
First published in Great Britain by Allen Lane 2020
001

Book design by Barbara M. Bachman
Printed and bound in Great Britain by Clays Ltd, Elcograf S.p.A.

A CIP catalogue record for this book is available from the British Library

HARDBACK ISBN: 978–0–241–32723–4
TRADE PAPERBACK ISBN: 978–0–241–43810–7

www.greenpenguin.co.uk

MIX
Penguin Random House is committed to a

To our wives, Holly and Heather,
for calling us on our bullshit when we need it—
but especially for not, when we don't.

CHAPTER 4

Causality

IF WE COULD GO BACK IN TIME AND PROVIDE ONE PIECE OF ADVICE to our fifteen-year-old selves, it would be this: *Feeling insecure, clueless, unconfident, naïve? Fake it. That's all that anyone else is doing*. Expressing self-confidence and self-esteem go a long way in shaping how others view you, particularly at that age. Indeed, faking social confidence is an act so self-fulfilling that we scarcely consider it bullshit. The kids with abundant self-confidence seemed happy and popular. They had the largest number of friends. They started dating earlier. High school seemed easier for them. The rest of us admired, envied, and occasionally hated them for it.

A recent study titled "Never Been Kissed" appears to illustrate how effective this kind of positive thinking can be. Surveying seven hundred college students, the authors of the study identified the personality traits that go hand in hand with never having kissed a romantic partner before starting college.

The research report is charming in the way it assumes zero prior knowledge of the human experience. We are told that "kissing is generally a positively valanced behavior." We learn that "the first kiss is often considered a very positive experience." We are informed that "physical intimacy is important in romantic relationships, and kissing is a common aspect of that physical intimacy." Best of all we are told, with a phrase only an epidemiologist could turn, that kissing has "an average age of onset of about 15.5 [years]."

So what factors influence whether or not someone has been kissed

tors of having had a first kiss prior to college. What makes people popular on the high school dating scene isn't good looks, intellectual ability, or good taste in music—it's self-confidence.

It's a nice story, but even though the study found an association between self-esteem and kissing, it is not so obvious which way that association goes. It's possible that self-esteem leads to kissing. But it's also possible that kissing leads to self-esteem. Or maybe kissing neither causes nor is caused by self-esteem. Maybe both are caused by having great hair.

This objection introduces us to a pervasive source of bullshit. People take evidence about the association between two things, and try to sell you a story about how one *causes* the other. Circumcision is associated with autism. Constipation is associated with Parkinson's disease. The marriage rate is associated with the suicide rate. But this doesn't mean that circumcision causes autism, nor that constipation causes Parkinson's, nor that marriage causes suicide. It is human nature to infer that when two things are associated, one causes the other. After all, we have evolved to find patterns in the world. Doing so helps us avoid danger, obtain food, deal with social interactions, and so much more. But often we are too quick to leap to conclusions about what causes what. In this chapter, we will show you how to think rigorously about associations, correlations, and causes—and how to spot bullshit claims that confuse one for the other.

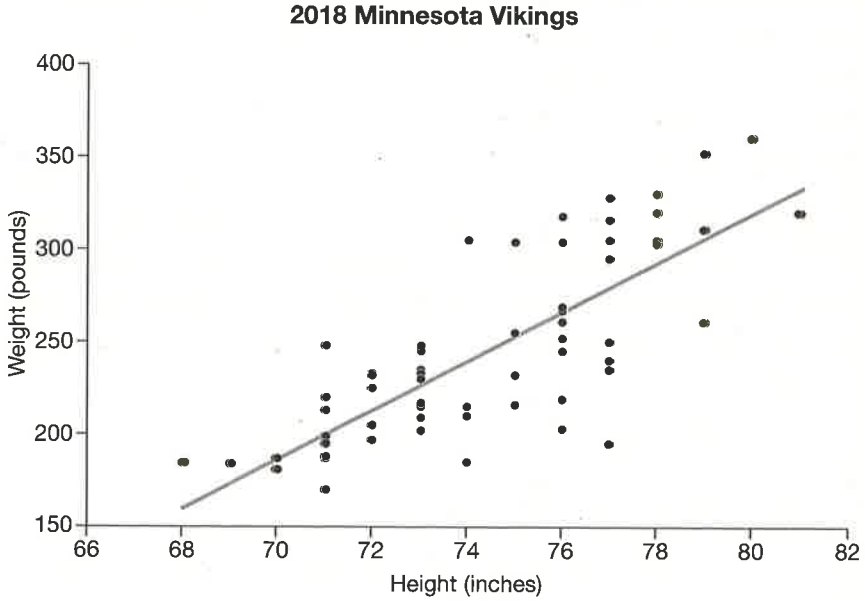
RED SKY AT NIGHT, SAILOR'S DELIGHT

"Red sky in the morning, sailors take warning. Red sky at night, sailor's delight." The rhyme reflects a pattern that people have known for over two thousand years. If you know what the sky looks like now, it tells you something about what the weather will be like later.

In wintertime in Seattle, an overcast sky usually means that it is relatively warm outside, because warm, wet air is sweeping overland from the ocean. When the sky is clear, it is usually colder outside because cold, dry air is blowing in from inland deserts. We don't need to step outside to know whether we need gloves and a hat; it's enough to simply look out the window. The cloud cover is *associated* with the overall temperature. We say that two measurements are associated

about the state of the other. Similarly, people's heights and weights are associated. If I tell you my friend is six feet four inches tall, you can safely guess that he will weigh more than many of my other acquaintances. If I tell you another friend is five feet one inch tall, you can guess that she is probably lighter than average.

In common language, we sometimes refer to associations as correlations. Someone might say, "I heard that your personality is correlated with your astrological sign. Aries are bold, whereas Taurus seek security." (This would be bullshit, but never mind that.) When scientists and statisticians talk about a correlation, however, they are usually talking about a linear correlation.* Linear correlations are so central to the way that scientists think about the world that we want to take a minute to explain how they work.



* Linear correlations require variables with numerical values such as height and weight, whereas associations can occur between categorical values such as "favorite color" and "favorite ice cream flavor," or between numerical variables. Correlations are associations, but not all associations are correlations. Moreover, values can be highly predictable without being linearly correlated. For example, consider pairs of numbers $\{x, \sin(x)\}$. If we know x, we can predict exactly what $\sin(x)$ will be, but the correlation coefficient—a measure of linear correlation—between these numbers is zero across a full-cycle sine wave. There is no linear correlation between x and $\sin(x)$ because a best-fit line

The easiest way to understand linear correlations is to imagine a scatter plot relating two kinds of measurements, such as the heights and weights of football players. We call each type of measurement a variable. Loosely speaking, two variables are linearly correlated if we can draw a slanted line that gets close to most of the points.

In the plot on page 52, each dot corresponds to a single player on the 2018 Minnesota Vikings football team. The horizontal position of a dot indicates the player's height, and the vertical position indicates the player's weight. For the Vikings, there is a linear correlation between players' heights and weights. The points lie roughly along the trend line superimposed on the points. Of course, the players' heights and weights don't lie right on the line. Quarterbacks and kickers, for example, tend to be lighter than you would expect given their height, whereas running backs and linemen tend to be heavier.

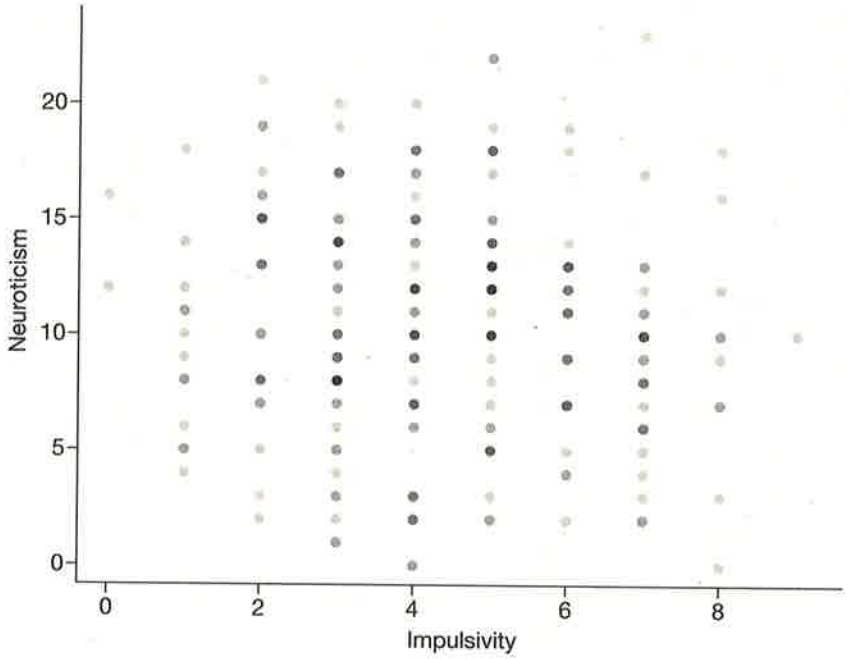
The strength of a linear correlation is measured as a *correlation coefficient,* which is a number between 1 and −1. A correlation of 1 means that the two measurements form a perfect line on a scatter plot, such that when one increases, the other increases as well. For example, distance in meters and distance in kilometers have a correlation coefficient of 1, because the former is just one thousand times the latter. A correlation of −1 means that two measurements form another kind of perfect line on a scatter plot, such that when one increases, the other decreases. For example, the time elapsed and the time remaining in a hockey game add up to sixty minutes. As one increases, the other decreases by the same amount. These two quantities have a correlation of −1.

A correlation coefficient of 0 means that a best-fit line through the points doesn't tell you anything.* In other words, one measurement tells you nothing about the other.† For example, psychologists sometimes use the Eysenck Personality Inventory questionnaire as a way to summarize aspects of personality known as impulsivity, sociability, and neuroticism. Across individuals, impulsivity and neuroticism are essen-

* In the rare case where the data points form either a vertical or horizontal line, the correlation coefficient is undefined. In these cases, knowing one measurement also tells you nothing about the other measurement.

† Strictly speaking, this is true only if you are restricted to using a linear model to predict one

tially uncorrelated, with a correlation coefficient of −0.07. In other words, knowing something about a person's impulsivity tells you very little (if anything) about his neuroticism and vice versa. The plot below illustrates neuroticism and impulsivity scores for a number of people. Darker points indicate multiple individuals with the same score.

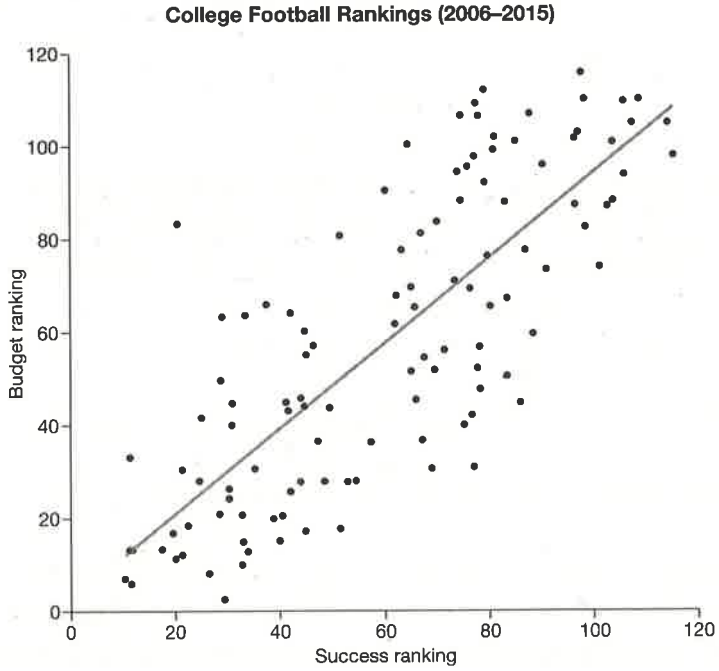


Most correlation coefficients lie somewhere in between 0 and 1, or they are smaller than 0 but bigger than −1. In either case, knowing one value tells us something, but not everything, about what the other value is likely to be.

To continue with sports examples, knowing the amount that a sports team spends tells you something about their likely win-loss record. Everyone knows that huge payrolls help the New York Yankees and FC Barcelona remain perpetual contenders in their respective leagues.

It is more surprising that this pattern holds even in US college sports, where the athletes are purportedly unpaid. If you look at the ranking by budget of college football programs and the ranking by competitive success, there is a strong relationship. On the following

The correlation coefficient between budget ranking and success ranking is 0.78. The powerhouse programs such as Alabama, Michigan, etc., are highly ranked, but they also spend the most money. Of course, the correlation is not perfect; outliers like Boise State have produced more wins than expected given their small budgets. It's not clear which way causality goes: Does money breed success, or does success generate more revenue from television, licensing, and donations? Most likely it goes both ways.



CONTEMPLATING CAUSATION

Ask a philosopher what causation is, and you open an enormous can of worms. When a perfectly struck cue ball knocks the eight ball into the corner pocket, why do we say that the cue ball *causes* the eight ball to travel across the table and drop? The dirty secret is that although we all have an everyday sense of what it means for one thing to cause another, and despite endless debate in the fields of physics and metaphysics alike, there is little agreement on what causation *is*. Fortunately, we don't need to know in order to use the notion of causation. In

poses. We want to know how to cause things. We want to know why things went wrong in the past, so that we can make them go right in the future.

But it is rarely straightforward to figure out what effects an action will have. A large fraction of the time all we have to work with is information about correlations. Scientists have a number of techniques for measuring correlations and drawing inferences about causality from these correlations. But doing so is a tricky and sometimes contentious business, and these techniques are not always used as carefully as they ought to be. Moreover, when we read about recent studies in medicine or policy or any other area, these subtleties are often lost. It is a truism that *correlation does not imply causation*. Do not leap carelessly from data showing the former to assumptions about the latter.*

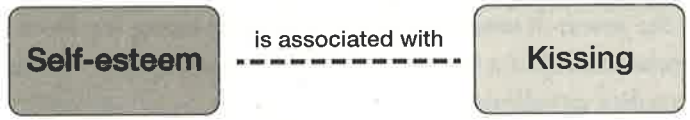
This is difficult to avoid, because people use data to tell stories. The stories that draw us in show a clear connection between cause and effect. Unfortunately, one of the most frequent misuses of data, particularly in the popular press, is to suggest a cause-and-effect relationship based on correlation alone. This is classic bullshit, in the vein of our earlier definition, because often the reporters and editors responsible for such stories don't care what you end up believing. When they tell you that drinking wine prevents heart disease, they are not trying to lead you into alcoholism or away from behaviors that promote cardiac health. At best, they are trying to tell a good story. At worst, they are trying to compel you to buy a magazine or click on a link.

One team of researchers recently attempted to figure out how common this type of misrepresentation is in news stories and on social media. They identified the fifty research studies shared most often on Facebook and Twitter about how factors such as diet, pollution, exercise, and medical treatment were correlated with health or illness. Because it is very difficult to demonstrate causality in a medical study, only fifteen of the fifty studies did a decent job of demonstrating cause-and-effect relationships. Of these, only two met the highest standards for doing so. The rest identified only correlations. That's okay; identifying correlations can generate important hypotheses,

* This principle holds for associations of any sort, not just for linear correlations. Though not as catchy a phrase, it's worth remembering that *association does not imply causation* either. That said, it is

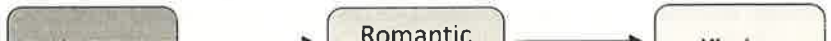
among other things. The problem is how the results were described. In a third of the studies, the medical journal articles themselves suggested causation in the absence of adequate evidence. Matters got worse in the popular press. Nearly half of news articles describing the studies made unwarranted claims about causation. When reading articles about medical trials or any other studies that purport to demonstrate causality, you can't count on the story being presented correctly. You need to be able to see through the bullshit.

Let's return to the never-been-kissed study that started off the chapter. The study found a strong association between positive self-esteem and having been kissed. To illustrate this association, we would draw a diagram such as the following:



The dashed line indicates an association. If we are willing to accept the story that acting confidently leads to social and romantic success, this association would be causal. Having self-esteem would cause being kissed. We can indicate a causal relationship by replacing the dashed line with an arrow from cause to effect:

A causal arrow like this doesn't have to represent absolute certainty. Positive self-esteem doesn't have to *ensure* that one gets kissed. We just mean that the higher one's self-esteem, the more likely one is to engage in kissing. And while an abundance of self-esteem may lead some people to walk up and kiss strangers, that's a bit too much self-esteem for our taste. We might refine our diagram to include an intermediate step, for example:



Alternatively, you might think that kissing is the cause, rather than the effect. Perhaps the wonder of a first kiss works miracles for one's self-esteem. It did for ours. In that case, the direction of causality would be reversed. In our diagram, we simply turn the causal arrow around.

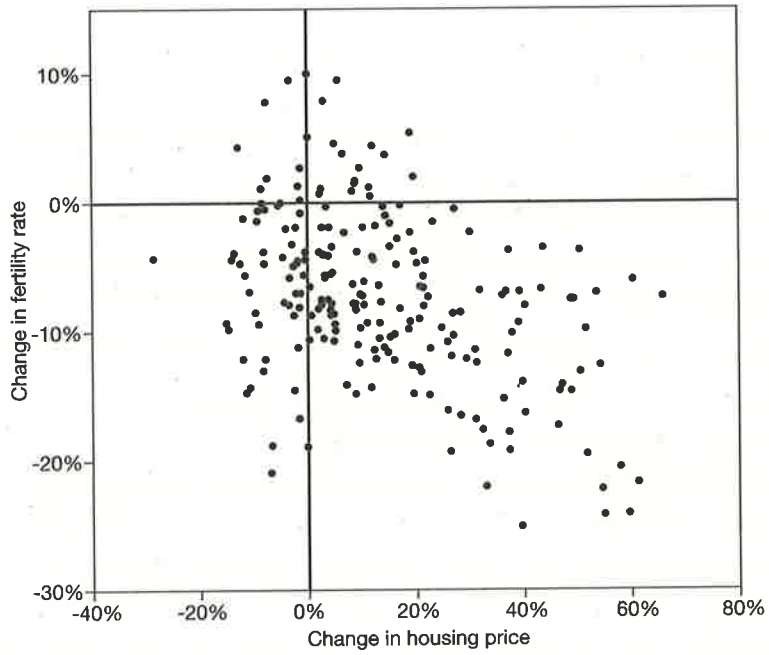
Of course, it's probably a bit more complicated than that. Perhaps it's not the kissing itself that leads to positive self-esteem for adolescents—it's simply being engaged in a romantic relationship. And (as the research study takes pains to note) being involved in a romantic relationship is a strong predictor of kissing. So we might diagram causality as follows:

Causality can even flow in multiple directions and form a feedback loop. Having positive self-esteem may increase an adolescent's likelihood of being engaged in a romantic relationship, and being in such a relationship may in turn increase self-esteem. We would diagram that as follows, with the feedback loop illustrated by the dual arrows at left:

Now that we understand correlations and associations and know how to diagram them, we can look at some of the ways that correla-

CORRELATION DOESN'T SELL NEWSPAPERS

In the summer of 2018, the real estate website Zillow reported a negative correlation between changes in housing prices and changes in birth rates. Cities in which housing prices had increased the most from 2010 to 2016 exhibited greater declines in the fertility rate for women aged 25 to 29. The trend is illustrated below.



There is a striking and seductively simple story that one could tell here: Having children is expensive. By some estimates, the financial cost of raising a child to age eighteen is comparable to the median cost of a house. Many people report waiting to start a family until they have enough money. Perhaps couples are forced to choose between buying a house and having a child. But this is only one of many possible explanations. The Zillow report makes that clear and discusses some of the other possibilities:

As a further caveat, the correlation observed here is by no means proof that home value growth *causes* fertility declines.

clustering into certain counties of people with careers that pay well enough for expensive homes but make it difficult to have children before 30; this could cause both trends observed in the chart above. There are many other confounding factors that could explain this relationship as well, such as the possibility that cultural values or the cost of child care varies across counties with some correlation to home values.

So far, no bullshit. This is the right way to report the study's findings. The Zillow article describes a correlation, and then uses this correlation to generate hypotheses about causation but does not leap to unwarranted conclusions about causality. Given that the study looks only at women aged 25 to 29, we might suspect that women with characteristics that make them likely to delay starting a family are also prone to moving to cities with high housing costs. After all, 25 to 29 is a demographic in which the frequency of childbirth differs considerably across socioeconomic strata and geographic regions. And when looking only at mothers in this age range, it is impossible to tell whether women are reducing the number of children they have, or are delaying the births of those children.

Unfortunately, this kind of distinction is often lost in the popular press. Shortly after the Zillow report was released, *MarketWatch* published a story about the Zillow findings. The first line of the story indicates a causal relationship: "Forget about a baby boom—rising home prices appear to be causing many would-be parents to think twice before expanding their family." Even the headline suggests causality: "Another Adverse Effect of High Home Prices: Fewer Babies." While this headline doesn't use the word "cause," it does use the word "effect"—another way of suggesting causal relationships. Correlation doesn't imply causation—but apparently it doesn't sell newspapers either.

If we have evidence of correlation but not causation, we shouldn't be making prescriptive claims. NPR reporter Scott Horsley posted a tweet announcing that "Washington Post poll finds NPR listeners are among the least likely to fall for politicians' false claims." Fair enough. But this poll demonstrated only correlation, not causation. Yet Horsley's tweet also recommended, "Inoculate yourself against B.S. Listen

sible that listening to NPR inoculates people against believing bullshit. If so, Horsley's advice would be merited. But it's also possible that being skeptical of bullshit predisposes people to listen to NPR. In that case, listening to NPR will not have the protective effect that Horsley supposes. NPR listeners were quick to call bullshit on Horsley's error—but this reinforces evidence of the correlation; it still doesn't prove causation.

The NPR example is merely silly, but matters get more serious when people make prescriptive claims based on correlational data in medical journalism. A 2016 study published in the prestigious *Journal of the American Medical Association* reported that people who exercise less have increased rates of thirteen different cancers. This study does not tell us anything about causality. Perhaps exercising reduces cancer rates, or perhaps people who do not exercise have other characteristics that increase their cancer risk. While the researchers tried to control for obvious characteristics such as smoking or obesity, this does not mean that any remaining differences are causal. The press ignored this subtlety and suggested a causal connection anyway. "Exercise Can Lower Risk of Some Cancers by 20%," proclaimed *Time* magazine in their headline about the study. "Exercising Drives Down Risk for 13 Cancers, Research Shows," announced the *Los Angeles Times*. "Exercise Cuts Cancer Risk, Huge Study Finds," declared *U.S. News & World Report*.

What people really want to read about, especially where health news is concerned, is not just the fact of the matter—they want to know what they ought to be doing. It's a small step from the causal claim that exercise cuts cancer risk, to a recommendation such as "exercise thirty minutes a day to prevent cancer." Much of the prescriptive advice that we read in the popular press is based on associations with no underlying evidence of causality.

Original scientific articles can make this mistake as well. Nutritionists have debated the merits of whole milk versus reduced-fat milk in preventing obesity, and typically favor reduced-fat milk. However, a recent study of children in San Francisco revealed that children who consumed more milk fat were less likely to be severely obese. The authors of the study correctly cautioned that this is a correlation and does not demonstrate a causal relationship.

But the title of the article suggests otherwise: "Full Fat Milk Con-

phasis added). This is causal language. Evidence of correlation is being miscast as evidence of causation. Worse yet, the authors make a prescriptive suggestion: "These results call into question recommendations that promote consumption of lower fat milk." No! There's no evidence here that consuming milk fat causes a reduction in obesity, and no reason to question milk-drinking recommendations from previous studies. Whenever you see a prescriptive claim, ask yourself whether there is causal evidence to back it up.

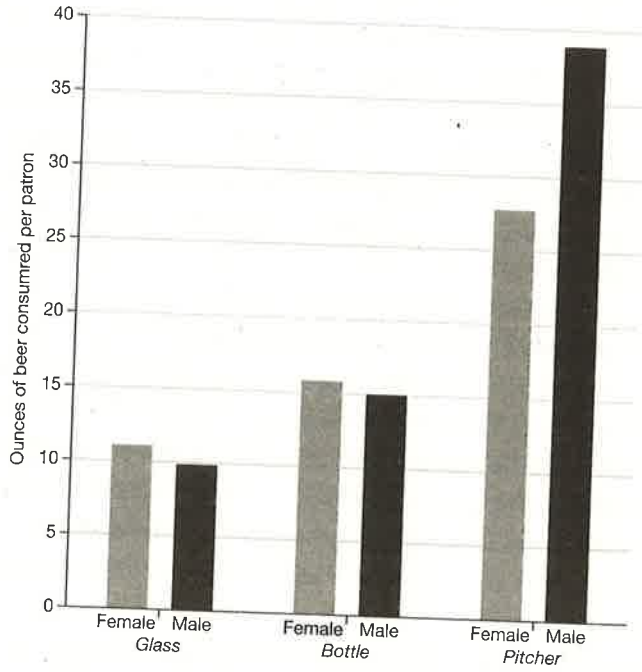
Moving on, what if someone argued that smoking doesn't cause cancer—but rather that cancer causes smoking? Crazy as it sounds, this is precisely what Ronald A. Fisher, one of the greatest statisticians of all time, tried to argue. Fisher noted that chronic inflammation of the lungs is associated with cancerous or precancerous states. Perhaps, he conjectured, this inflammation creates a discomfort that can be soothed by the act of smoking. If so, people in the process of developing cancer might take to smoking as a way to alleviate their symptoms. Those not developing cancer would be less likely to take up the habit. Would it be a stretch, then, to say that cancer causes smoking? Fisher turned out to be wrong, of course, but he was making a point about the challenges of inferring causality—and probably justifying his beloved pipe-smoking habit at the same time. Fisher's suggestion about cancer and smoking never got much traction, but the tobacco industry found other ways to seed doubts about whether or not smoking caused disease. Their efforts delayed antismoking legislation for decades.

Other mistaken assumptions about causality have had a serious impact in debates about drugs and public health. In the 1980s, American university administrators and policy makers were concerned about the prevalence of binge drinking on university campuses. Psychologists, epidemiologists, public health experts, and others searched for ways to stem this epidemic of intemperance.

And why not? There are worse places to do fieldwork. In an influential 1986 paper titled "Naturalistic Observations of Beer Drinking among College Students," psychologist Scott Geller and colleagues looked at factors associated with greater consumption of beer at college pubs. What are "naturalistic observations"? They are the observations you make of a subject, in this case the college students, in their natural habitat, in this case the pub. We are amused by this detail f

main as inconspicuous as possible by sitting at tables and *behaving as normal patrons*" (emphasis added). Does this mean drinking beer themselves? One must take pains to blend in, after all.

The researchers observed the number of beers that each student consumed and recorded whether each was purchased by the glass, the bottle, or the pitcher. They observed a strong correlation between the vessel in which beer was served and the amount consumed.



Students who drank beer from pitchers drank roughly two to four times as much beer as those who drank their beer by the glass or by the bottle. The original study was careful not to claim a causal relationship.* But the claim evolved as reports of the study filtered through the popular press and into the broader discussion about alcohol abuse on college campuses. "People drink more *when* beer is consumed in pitchers" was taken to mean "People drink more *because* beer is con-

* Geller and colleagues write: "It would be instructive to determine how much of the relationship between container type and drinking behavior

sumed in pitchers." Based on this, people started making prescriptive claims: "We should ban pitchers so that students will drink less."

Perhaps you already see the problem with this inference. Students aren't necessarily drinking more beer because they ordered a pitcher. They are probably ordering pitchers because they intend to drink more beer. When we two authors go to a bar and want one beer each, we each order a glass. When we want two beers each, we order a pitcher and split it. And we are the kind of fellows who follow through on their intentions: When we intend to drink more beer, we usually do.

Geller's study doesn't necessarily show us that people drink more when served from pitchers. Instead, he may have discovered that people who want to drink lots of beer order more beer than people who want to drink a little bit of beer. Unfortunately, that doesn't make for very exciting headlines, so one can see why the newspapers tried to spin it in a different direction.

The two cases that we just treated are relatively clear-cut, at least with the advantage of hindsight. Smoking causes cancer; intending to drink more beer is associated both with ordering more beer and with drinking more beer. But in many cases we do not know which way causality flows. Studies have found an association between poor sleep and the buildup of beta-amyloid plaques that cause Alzheimer's disease. One hypothesis is that sleep provides a sort of downtime during which the brain can clean up these plaques. If so, a lack of sleep may be a cause of Alzheimer's. But from the available data, it is also possible that causality goes in the opposite direction. A buildup of beta-amyloid plaques may interfere with sleep, in which case it would be that Alzheimer's (or pre-Alzheimer's) causes poor sleep. As yet we simply don't know.

There are many ways to imply causality. Some are overt: "Smoking causes cancer," or "Red wine prevents heart disease." Some make prescriptions: "To avoid cancer, exercise three times a week." But others are less obvious. We can even imply causality with subtle grammatical shifts. We might express a correlation with a plain factual statement in the indicative mood: "If she *is* a Canadian, she *is* more likely to be bilingual." We express causation using a counterfactual statement in the subjunctive mood: "If she *were* a Canadian, she *would be* more likely to be bilingual." The former statement simply suggests an association between

bilinguality. The former statement suggests that people are selected at random from a large group and their attributes compared: "If [the person we happen to pick] is a Canadian . . ." The second suggests that we pick someone and then change some of that person's characteristics: "If [the person we picked] were [turned into a] Canadian."

It is subtle, but the subjunctive mood sneaks in a suggestion of causality. "Where minimum wage is higher, poverty is lower" is not the same claim as "If minimum wage were to increase, poverty would decrease." The first reports a trend across cities: Those with higher minimum wage have lower poverty rates. The second suggests how one might go about reducing poverty in a particular city.

Data graphics can also suggest causality in subtle ways. Think back to the scatter plot of housing price changes versus fertility changes. In many graphs of this sort, the horizontal axis is used to illustrate the variable that causes—or at least influences—the variable shown on the vertical axis. In the Zillow graph, housing prices are shown on the horizontal axis and fertility rates are shown on the vertical axis. Without a word of causal language, this graph conveys a subtle suggestion that housing prices determine fertility rate. A graph like this can trick readers into presuming a causal relationship. When you see scatter plots and related forms of data visualization, ask yourself (and maybe the person who created the graph): Is the structure of the graph suggesting a causal relationship that isn't there?

DELAYED GRATIFICATION AND COMMON CAUSE

One of the hallmark discoveries of social psychology is the role that delayed gratification plays in a successful life. At the heart of delayed gratification theory is an experiment known as the marshmallow test. A four-year-old is presented with alternative rewards: one marshmallow or two marshmallows. He is told that he can have a single marshmallow anytime—but if he can wait for a while, he can have two marshmallows. The experimenter then leaves the room and measures the amount of time until the child says screw it, and takes the single marshmallow. (After fifteen minutes of open-ended waiting, a child who has not yet given up receives the two-marshmallow reward. But seriously—do you remember how long fifteen minutes seemed at that

A number of studies have shown that children who can wait longer at age four have higher SAT scores in high school, and are rated by their parents as better-adjusted during adolescence. The authors of the original studies were careful to explain that their results demonstrated correlation: Delayed gratification is predictive of later academic success and emotional well-being. They did not demonstrate causation: The ability to delay gratification does not necessarily *cause* later success and well-being.* But as these results filtered through the popular press, the line between correlation and causation became blurred. The results of the marshmallow test and other related studies were reported as evidence that ability to delay gratification causes success later in life.

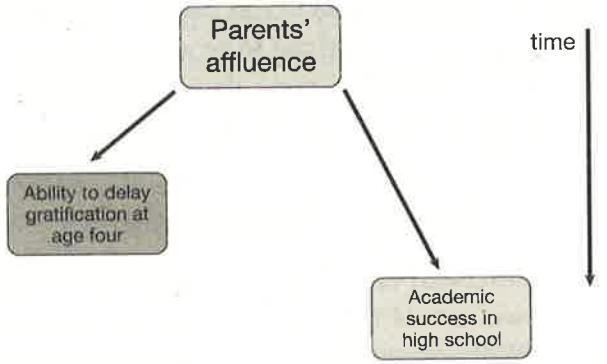
These assumptions about causation are often used as grounds to make a prescription: Improve your future by learning to delay gratification. In response to the marshmallow test, pop-psych and pop-business outlets promote training methods. *Lifehacker* exhorted us to "Build the Skill of Delayed Gratification." "If you can manage to turn delaying gratification into a regular habit," read the copy beneath a stock photo of marshmallows in *Fast Company,* "you may be able to take your own performance from just mediocre to top-notch." In an article titled "40-Year-Old Stanford Study Reveals the 1 Quality Your Children Need to Succeed In Life," *Inc.* magazine explains how to cultivate this ability in our children:

> In other words, actively establish a system for delayed gratification in your young one's brain by promising small rewards for any work done and then delivering on it. If you keep doing this, their brain will automatically gravitate toward doing hard work first; it's classical conditioning at work.

But here's the thing. These prescriptions are unwarranted, because we don't actually have strong evidence that the ability to delay gratification causes subsequent success. When a research team went back

* In a 1990 paper by a core group of researchers in the area, Shoda and colleagues, we are cautioned in the discussion that "stability in parent child-rearing practices and in the psychosocial environment in the family and the community may be a common factor underlying both preschool children's delay of gratification behavior and their cognitive and self-regulatory competence in ado-

and attempted to replicate the original marshmallow studies with a larger sample size and additional controls, they found only a fraction of the original effect. Moreover, there was a single factor that seemed responsible both for a child's ability to delay gratification and for success during adolescence: the parents' socioeconomic status.* Children from wealthy families were better able to wait for the second marshmallow. Why? Perhaps they felt a greater sense of overall stability, had greater trust in adults, recalled previous situations in which waiting had proven fruitful, and felt relative indifference—a marshmallow might not be such a special treat for these children. Parental wealth also is a major determinant of an adolescent's educational success. So both the ability to delay gratification and academic success are consequences of parental wealth. Neither effect causes the other. In a case like this, where parental wealth is a common cause of patience and success, we diagram it as follows.



This causal diagram features an arrow to indicate the direction of time. Children are able to delay gratification (or not) at age four long before they are academically successful (or not) in adolescence. Causation flows only forward in time. If A happens before B does, we know that B doesn't cause A. That's useful. In this case, we can immediately rule out the possibility that academic success in high school causes the ability to delay gratification at age four.

But if one is not careful, looking at events chronologically can be

* Statisticians sometimes use the term *confounding* to refer to situations where a common cause in-

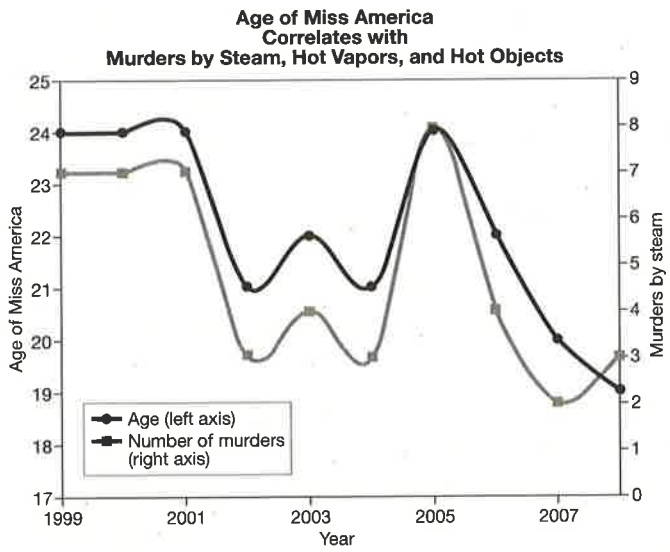
misleading. Just because A happens before B does not mean that A causes B—even when A and B are associated. This mistake is so common and has been around for so long that it has a Latin name: *post hoc ergo propter hoc*. Translated, this means something like "after this, therefore because of it."

It's human to make these mistakes. We are excellent at seeing patterns, and this ability can help us generalize from one experience to another. We might learn that flying black insects don't sting, whereas flying yellow and black insects do. The observations we make now help us anticipate events in the future. We might notice that every time there is a heavy downpour, the river runs higher the next day and should be crossed with caution. We often apply rules of thumb, such as "if two things are associated, the one that happens first causes the one that happens second." Droughts and wildfires are associated; droughts happen first and are a cause of wildfires. But this pattern-seeking ability can also mislead us. If migrating geese arrive in early September every year and coho salmon begin to run later in the month, we might assume that the geese have something to do with calling the fish up the rivers. Of course, the fish don't give a damn about the geese. This is another example of the *post hoc ergo propter hoc* fallacy.

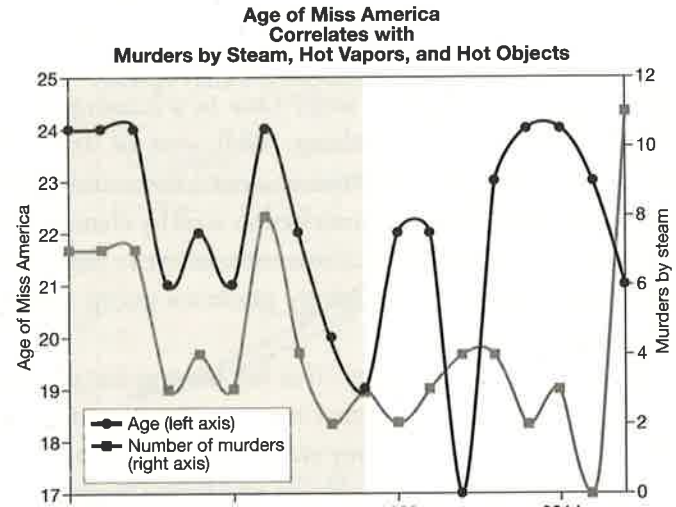
SPURIOUS CORRELATIONS

So far, we've discussed cases where there are meaningful correlations between two events or measurements, but people draw the wrong inferences about causality. Ordering pitchers and drinking more beer are legitimately associated, but it was a mistake to assume that ordering pitchers causes people to drink more beer. Some correlations do not rise even to this standard. They exist by chance, don't tell us anything meaningful about how the world works, and are unlikely to recur when tested with new data. Author Tyler Vigen has collected a delightful set of examples, and has a website where you can uncover these *spurious correlations* yourself. For example, did you know that the age of Miss America is tightly correlated to the number of people murdered with steam, hot vapors, and other hot objects?*

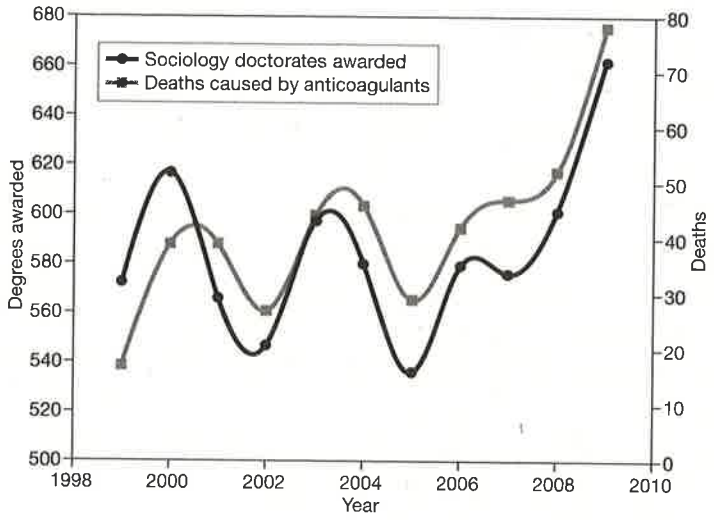
* In order to draw a smooth curve without jagged corners at each point, Vigen uses a technique



There is no way that this correlation captures something meaningful about how the world works. What possible causal connection could there be between these two things? Intuitively, we know this has to be a spurious correlation. It's just random chance that these measurements line up so well. Because it's just chance, we do not expect this trend to hold in the future. Indeed, it doesn't. If we continue the time series across the intervening years since Vigen published this figure, the correlation completely falls apart.



Vigen finds his spurious correlation examples by collecting a large number of data sets about how things change over time. Then he uses a computer program to compare each trend with every other trend. This is an extreme form of what data scientists call *data dredging*. With a mere one hundred data series, one can compare nearly ten thousand pairs. Some of these pairs are going to show very similar trends—and thus high correlations—just by chance. For example, check out the correlation between the numbers of deaths caused by anticoagulants and the number of sociology degrees awarded in the US:



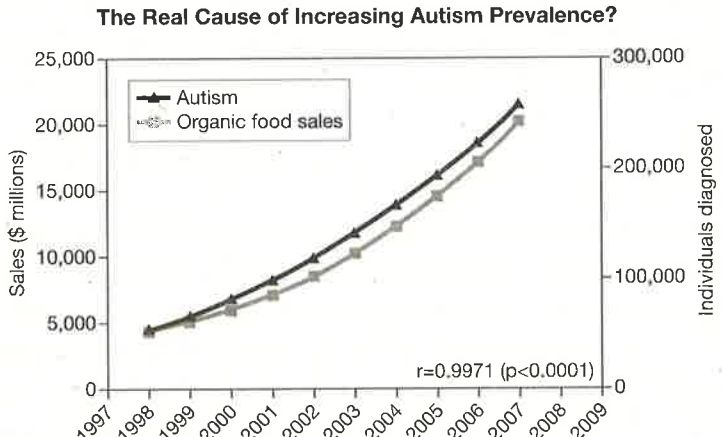
You look at these two trends and think, Wow—what are the chances they would line up that well? One in a hundred? One in a thousand? That must *mean* something. Well, sort of. It means that Vigen had to look through a hundred or even a thousand other comparisons before he found one that matched so well by chance. It doesn't mean that there is any meaningful connection between the two trends. It certainly doesn't mean that sociology grads are going around with rat poison and killing people.

In Vigen's case this is just funny. But his mining for silly correlations parallels a serious problem that can arise in scientific analyses. Particularly in the early exploratory stages, much of science involves a search for patterns in nature. As larger and larger data sets become

for patterns can start to look more and more like Vigen's humorous dragneting expedition.

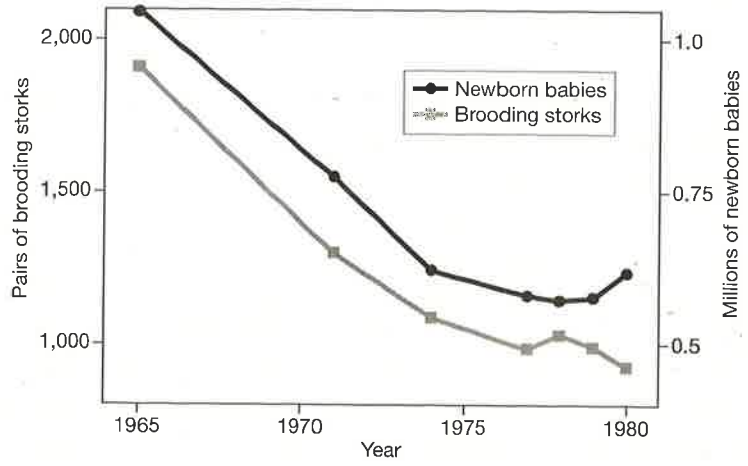
Researchers have collected large survey data sets in which participants are asked dozens of questions about many aspects of their lives, values, personality traits, etc. When digging through these data sets to test hypotheses, researchers need to be careful that they aren't inadvertently doing what Vigen does on purpose: making so many different comparisons that they end up finding similarities that exist by chance rather than as a reflection of any real-world relationship.

One of the easiest ways to get a spurious correlation in trends over time is to look at very simple trends. There are millions of different things we could go out and measure. Many of them are increasing over time: the number of emails in Jevin's in-box, Amazon stock prices, a child's height, the cost of a new car, even the year on the Gregorian calendar. Many others are decreasing over time: the area of arctic sea ice on New Year's Day, cesium-137 levels in Chernobyl, rates of early-onset lung cancer, the cost of storing 1 gigabyte of data. If we compare any two of the increasing quantities, their values will be positively correlated in time. The same thing happens if we compare any two decreasing quantities. (If we compare an increasing quantity with a decreasing quantity, we will get a correlation as well—but it will be a negative correlation.) For the vast majority of pairs, there will be no causal connection at all. In a not-too-subtle jab at the natural health community, one user of the Reddit website posted the following graph.



Obviously there is no reason to assume any sort of causal relationship between organic food sales and autism, but that is the point of the joke. This erroneous attribution of causality is the same mistake that the natural health community makes in ascribing autism to vaccination.

In the late 1980s, a chemist used the same trick to publish a humorous graph in *Nature,* one of the top research journals in the world. The graph, titled "A New Parameter for Sex Education," serves as a cautionary tale against inferring too much from a correlation.



This graph compares two declining trends: pairs of brooding storks in West Germany and the number of newborn human babies. The winking implication is that one should consider a possible causal relationship here. Perhaps the old tall tale is right: Perhaps storks do bring babies after all. If the storks go away, no more babies.

SMOKING DOESN'T KILL?

In our discussions of causality, we have been talking about probabilities, not certainties. We say that drunk driving causes automobile accidents not because every drunk driver crashes or because every crash involves a drunk driver, but rather because driving drunk greatly increases the risk of a crash. There is a key distinction between a *probabilistic cause* (A increases the chance of B in a causal manner), a *sufficient cause* (if A happens, B always happens), and a *necessary cause* (unless A happens, B can't happen).

The distinction between necessary and sufficient causes is sometimes misused, particularly by those interested in denying causal relationships. For example, Mike Pence once made the following argument against government regulation of tobacco:

> Time for a quick reality check. Despite the hysteria from the political class and the media, smoking doesn't kill. In fact, 2 out of every three smokers does not die from a smoking related illness and 9 out of ten smokers do not contract lung cancer.

This is just plain bullshit, and bullshit of a higher grade than usually appears in print. In one sentence Pence says literally "Smoking doesn't kill," and in the very next he says that a third of smokers die of smoking-related illness.* Pence is conflating sufficient cause with probabilistic cause. Smoking is not sufficient to guarantee lung cancer or even smoking-related illness, but it does greatly increase the probability that someone will die of one or the other. A related argument might be that smoking doesn't cause lung cancer because some lung cancer victims—miners, for example—never smoked. This argument conflates necessary cause with probabilistic cause.

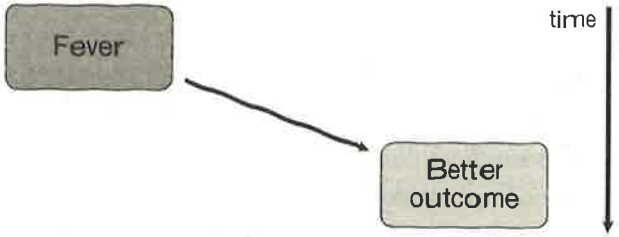
WHEN ALL ELSE FAILS, MANIPULATE

With all these traps and pitfalls, how can we ever be confident that one thing causes another? Scientists struggle with this problem all the time, and often use manipulative experiments to tease apart correlation and causation. Consider the biology of fever. We commonly think of fever as something that disease *does to us,* the way a cold gives us a sore throat, or measles covers the skin with pox. As such, physicians might aim to block or prevent fever, using drugs such as aspirin, Tylenol, or Advil. But fever seems to be different from a sore throat or an outbreak of pox. Multiple lines of evidence suggest that a moder-

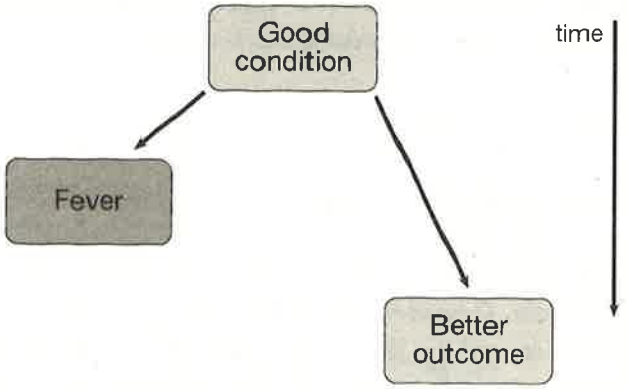
* Pence's claim is a massive underestimate of the fraction killed by smoking-related illness. About two-thirds of smokers die from smoking-related illness, according to a recent large-scale study by Emily Banks and colleagues.

ate fever is one of the body's defenses against infection. For example, people who mount a fever are more likely to survive a bloodstream infection. But this is a correlation, not causation.

Does fever *cause* better outcomes, as diagrammed below?



Or are patients who are in better condition (healthier overall, not malnourished, with less severe infections, or better off in any number of other ways) the ones who are able to mount a fever in the first place? Because these patients are in better condition to start with, we would expect these patients to have better outcomes, irrespective of the effects of fever.



How do we distinguish between these two possibilities? How can we figure out whether fever actually causes better disease outcomes?

Experiments. With fever, there are "natural experiments" taking place all the time, in the sense that absent any experimenter, different patients are treated in different ways. In particular, when visiting the

hospital or the general practitioner, some patients are given drugs to reduce fever, whereas others are not. Overall and across numerous studies, there is a strong trend: Patients who are given fever-reducing (antipyretic) drugs take longer to recover from viral infections.

Does this mean that fever is beneficial? Not necessarily, because fever-reducing drugs are not assigned randomly to patients. The group of patients who are given drugs may have properties different from the group of patients who do not receive drugs. What we are looking at here is a form of *selection bias*. In particular, it may be that people who are in poorer condition are more likely to receive drugs to reduce their fevers. If so, it might look like taking fever-reducing drugs caused negative outcomes. But in fact, it would be that people who were likely to experience worse outcomes were *selected* to receive fever-reducing drugs.

To get around this problem, we might explicitly *randomize* the treatments that patients receive. Then any differences in outcome would be due to the effects of the treatment, rather than differences in the conditions of the patients. While we cannot ethically randomize whether people get treated or not for life-threatening illnesses, we can—with consent from patients—do so for less severe diseases. Using this approach, researchers have found that drugs that block fever tend to slow the rate at which patients recover, and increase the chance that patients spread the illness to others. But we still don't know for certain that temperature is the main cause of these differences. It could be that the drugs themselves, not the changes in temperature that they induce, are responsible. Is it that fever-reducing drugs cause fever to drop, and that the reduction in fever causes worse disease outcomes? Or do fever-reducing drugs have their own negative consequences—independent of their effect on body temperature?

To rule out this possibility, scientists turned to experiments with laboratory animals. They physically cooled the animals. Doing so had the same effect on disease outcomes as did fever-reducing drugs. This suggests that the negative consequences of fever-reducing drugs arise through their effect on body temperature. With this piece in place, we now have a solid chain of evidence supporting the idea that fever is a beneficial defense against disease. Manipulative experiments offer some of the strongest evidence of causality because of the ability to

isolate the purported cause and keep everything else constant. The problem is, such experiments are not always feasible, so we have to rely on other forms of evidence. That's all good and well, but when you do so don't be taken in by an unfounded leap from correlation to causation.

CHAPTER 5

Numbers and Nonsense

OUR WORLD IS THOROUGHLY QUANTIFIED. EVERYTHING IS COUNTED, measured, analyzed, and assessed. Internet companies track us around the Web and use algorithms to predict what we will buy. Smartphones count our steps, measure our calls, and trace our movements throughout the day. "Smart appliances" monitor how we use them and learn more about our daily routines than we might care to realize. Implanted medical devices collect a continuous stream of data from patients and watch for danger signs in real time. During service visits, our cars upload data about their performance and about our driving habits. Arrays of sensors and cameras laid out across our cities monitor everything from traffic to air quality to the identities of passersby.

Instead of collecting data about what people do through costly studies and surveys, companies let people come to them—and then record what those consumers do. Facebook knows whom we know; Google knows what we want to know. Uber knows where we want to go; Amazon knows what we want to buy. Match knows whom we want to marry; Tinder knows whom we want to be swiped by.

Data can help us understand the world based upon hard evidence, but hard numbers are a lot softer than one might think. It's like the old joke: A mathematician, an engineer, and an accountant are applying for a job. They are led to the interview room and given a math quiz. The first problem is a warm-up: What is 2 + 2? The mathematician rolls her eyes, writes the numeral 4, and moves on. The engineer pauses for a moment, then writes "Approximately 4." The accountant