CALLING BULLSHIT

The Art of Skepticism in a Data-Driven World

CARL T. BERGSTROM
& JEVIN D. WEST

RANDOM HOUSE | NEW YORK

Published in the United States by Random House,
an imprint and division of Penguin Random House LLC, New York.

Random House and the House colophon are
registered trademarks of Penguin Random House LLC.

LIBRARY OF CONGRESS CATALOGING-IN-PUBLICATION DATA

Names: Bergstrom, Carl T., author. | West, Jevin D. (Jevin Darwin), author.

Title: Calling bullshit : the art of skepticism in a data-driven world / Carl T. Bergstrom and Jevin D. West.

Description: First edition. | New York : Random House, 2020. | Includes bibliographical references.

Identifiers: LCCN 2020003592 (print) | LCCN 2020003593 (ebook) | ISBN 9780525509189 (hardcover acid-free paper) | ISBN 9780593229767 (international edition acid-free paper) | ISBN 9780525509196 (ebook)

Subjects: LCSH: Skepticism.

Classification: LCC B837 .B47 2020 (print) | LCC B837 (ebook) | DDC 149/.73—dc23

LC record available at https://lccn.loc.gov/2020003592

LC ebook record available at https://lccn.loc.gov/2020003593

Printed in the United States of America on acid-free paper

randomhousebooks.com

987654321

First Edition

Book design by Barbara M. Bachman

Contents

PREFACE xi

Chapter 1: BULLSHIT EVERYWHERE 3

Chapter 2: MEDIUM, MESSAGE, AND MISINFORMATION 18

Chapter 3: THE NATURE OF BULLSHIT 38

Chapter 4: CAUSALITY 50

Chapter 5: NUMBERS AND NONSENSE 77

Chapter 6: SELECTION BIAS 104

Chapter 7: DATA VISUALIZATION 134

Chapter 8: CALLING BULLSHIT ON BIG DATA 180

Chapter 9: THE SUSCEPTIBILITY OF SCIENCE 206

Chapter 10: SPOTTING BULLSHIT 242

Chapter 11: REFUTING BULLSHIT 264

ACKNOWLEDGMENTS 287

BIBLIOGRAPHY 289

INDEX 309

PREFACE

THE WORLD IS AWASH WITH BULLSHIT, AND WE'RE DROWNING IN IT.

Politicians are unconstrained by facts. Science is conducted by press release. Silicon Valley startups elevate bullshit to high art. Colleges and universities reward bullshit over analytic thought. The majority of administrative activity seems to be little more than a sophisticated exercise in the combinatorial reassembly of bullshit. Advertisers wink conspiratorially and invite us to join them in seeing through all the bullshit. We wink back—but in doing so drop our guard and fall for the second-order bullshit they are shoveling at us. Bullshit pollutes our world by misleading people about specific issues, and it undermines our ability to trust information in general. However modest, this book is our attempt to fight back.

The philosopher Harry Frankfurt recognized that the ubiquity of bullshit is a defining characteristic of our time. His classic treatise *On Bullshit* begins as follows:

> One of the most salient features of our culture is that there is so much bullshit. Everyone knows this. Each of us contributes his share. But we tend to take the situation for granted [yet] we have no clear understanding of what bullshit is, why there is so much of it, or what functions it serves. And we lack a conscientiously developed appreciation of what it means to us. In other words, we have no theory.

To eradicate bullshit, it is helpful to know precisely what it is. And here things get tricky.

First of all, "bullshit" is both a noun and a verb. Not only can I get tired of listening to your bullshit (*n.*), I can turn around and bullshit (*v.*) you myself. This is straightforward enough. To bullshit is, at a first approximation, the act of producing bullshit.

But what does the noun "bullshit" refer to anyway? As with many attempts to match philosophical concepts to everyday language, it would be a fool's errand to attempt a definition that incorporates and excludes everything it should. Instead, we will start with a few examples and then try to describe some things that would qualify as bullshit.

Most people think they're pretty good at spotting bullshit. That may be true when bullshit comes in the form of rhetoric or fancy language, what we call *old-school bullshit*. For example:

- Our collective mission is to functionalize bilateral solutions for leveraging underutilized human resource portfolio opportunities. (In other words, we are a temp agency.)
- We exist as transmissions. To embark on the myth is to become one with it. (We might call this new-age old-school bullshit.)
- Just as our forefathers before us, we look to the unending horizons of our great nation with minds fixed and hearts aflame to rekindle the dampened sparks of our collective destiny. (Spare us. How are you going bring jobs back to the district?)

Old-school bullshit doesn't seem to be going away, but it may be overshadowed by the rise of what we call *new-school bullshit*. New-school bullshit uses the language of math and science and statistics to create the impression of rigor and accuracy. Dubious claims are given a veneer of legitimacy by glossing them with numbers, figures, statistics, and data graphics. New-school bullshit might look something like this:

- Adjusted for currency exchange rates, our top-performing global fund beat the market in seven of the past nine years.

(How exactly were returns adjusted? How many of the company's funds failed to beat the market and by how much? For that matter,

was there a single fund that beat the market in seven of nine years, or was it a different fund that beat the market in each of those seven years?)

- While short of statistical significance ($p = 0.13$), our results underscore the clinically important effect size (relative odds of survival at five years = 1.3) of our targeted oncotherapy and challenge the current therapeutic paradigm.

(What does it mean for a result to be clinically important if it is not statistically significant? Is five-year survival a relevant measure for this particular cancer, or are most patients deceased within three years? Why should we imagine that any of this "challenges the current therapeutic paradigm"?)

- The team's convolutional neural net algorithm extracts the underlying control logic from a multiplex network composed of the human metabolome, transcriptome, and proteome.

(What is a multiplex network? Why are connections among these various -omes meaningful, and how are they measured? What does the author mean by "control logic"? How do we know there *is* an underlying control logic linking these systems, and if there is, how do we know that this approach can actually capture it?)

- Our systematic screening revealed that 34 percent of behaviorally challenged second graders admit to having sniffed Magic Markers at least once in the past year.

(Why does it matter? And if it does, is marker-sniffing a cause or a consequence of being "behaviorally challenged"? What fraction of nonchallenged second graders admit to sniffing Magic Markers? Perhaps that fraction is even higher!)

New-school bullshit can be particularly effective because many of us don't feel qualified to challenge information that is presented in quantitative form. That is exactly what new-school bullshitters are

counting on. To fight back, one must learn when and how to question such statements.

WE HAVE DEVOTED OUR careers to teaching students how to think logically and quantitatively about data. This book emerged from a course we teach at the University of Washington, also titled "Calling Bullshit." We hope it will show you that you do not need to be a professional statistician or econometrician or data scientist to think critically about quantitative arguments, nor do you need extensive data sets and weeks of effort to see through bullshit. It is often sufficient to apply basic logical reasoning to a problem and, where needed, augment that with information readily discovered via search engine.

We have civic motives for wanting to help people spot and refute bullshit. It is not a matter of left- or right-wing ideology; people on both sides of the aisle have proven themselves skilled at creating and spreading misinformation. Rather (at the risk of grandiosity), we believe that adequate bullshit detection is essential for the survival of liberal democracy. Democracy has always relied on a critically thinking electorate, but never has this been more important than in the current age of fake news and international interference in the electoral process via propaganda disseminated over social media. In a December 2016 *New York Times* op-ed, Mark Galeotti summarized the best defense against this form of information warfare:

> Instead of trying to combat each leak directly, the United States government should teach the public to tell when they are being manipulated. Via schools and nongovernmental organizations and public service campaigns, Americans should be taught the basic skills necessary to be savvy media consumers, from how to fact-check news articles to how pictures can lie.

As academics with decades of experience teaching data science, statistics, and related subjects at a public university, we know how to teach this sort of thinking. We believe that it can be done without taking sides, politically. You may not agree with us about the optimal size of the federal government, about what constitutes an acceptable degree of government involvement in our private lives, or how the

country should conduct itself on the world stage—but we're good with that. We simply want to help people of all political perspectives resist bullshit, because we feel that a democracy is healthiest when voters can see through the bullshit coming from all sides.

We are not setting up a platform from which we can call bullshit on things that we don't like. For that reason, the examples in this book are seldom the most egregious ones we know, let alone the instances that anger us the most. Rather, our examples are chosen to serve a pedagogical purpose, drawing out particular pitfalls and highlighting appropriate strategies for responding. We hope you'll read, think, and go call bullshit yourself.

OVER A CENTURY AGO, the philosopher John Alexander Smith addressed the entering class at Oxford as follows:

> Nothing that you will learn in the course of your studies will be of the slightest possible use to you [thereafter], save only this, that if you work hard and intelligently you should be able to detect when a man is talking rot, and that, in my view, is the main, if not the sole, purpose of education.

For all of its successes, we feel that higher education in STEM disciplines—science, technology, engineering, and medicine—has dropped the ball in this regard. We generally do a good job teaching mechanics: students learn how to manipulate matrices, transfect cells, run genomic scans, and implement machine learning algorithms. But this focus on facts and skills comes at the expense of training and practice in the art of critical thinking. In the humanities and the social sciences, students are taught to smash conflicting ideas up against one another and grapple with discordant arguments. In STEM fields, students seldom are given paradoxes that they need to resolve, conflicting forms of evidence that they must reconcile, or fallacious claims that they need to critique. As a result, college graduates tend to be well prepared to challenge verbal arguments and identify logical fallacies but surprisingly acquiescent in the face of quantitative claims. Of course, the same holds true for secondary education. If STEM education incorporated the interrogatory teaching practices already com-

mon in the humanities, schools could shape a generation of students prepared to call bullshit on statistical statements and artificial intelligence analyses just as comfortably as current students can on political, ethical, artistic, and philosophical claims.

For a number of reasons, we draw heavily on examples from research in science and medicine in the chapters that follow. We love science and this is where our expertise lies. Science relies on the kinds of quantitative arguments we address in this book. Of all human institutions, science seems as though it ought to be free from bullshit—but it isn't. We believe that public understanding of science is critical for an informed electorate, and we want to identify the many obstacles that interfere with that understanding.

But we want to stress that nothing we say undermines science as a successful, institutionalized means to understand the physical world. For all our complaints, for all the biases we identify, for all the problems and all the bullshit that creeps in, at the end of the day *science works*. With science on our side, we fly in airplanes, talk on videophones, transplant organs, eradicate infectious diseases, and apprehend phenomena ranging from the early moments after the Big Bang to the molecular basis of life.

New forms of information technology have changed the way we communicate in both science and society. As access has improved, information overload has worsened. We hope this book will help you face the onslaught and separate fact from fiction.

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 looks around nervously, then gets out of his chair and walks over to

the fellow administering the test. "Before I put anything in writing," he says in a low whisper, "what do you want it to be?"

Numbers are ideal vehicles for promulgating bullshit. They feel objective, but are easily manipulated to tell whatever story one desires. Words are clearly constructs of human minds, but numbers? Numbers seem to come directly from Nature herself. We know words are subjective. We know they are used to bend and blur the truth. Words suggest intuition, feeling, and expressivity. But not numbers. Numbers suggest precision and imply a scientific approach. Numbers appear to have an existence separate from the humans reporting them.

People are so convinced of the primacy of numbers that skeptics claim they "just want to see the data," or demand to be shown "the raw numbers," or insist that we "let the measurements speak for themselves." We are told that "the data never lie." But this perspective can be dangerous. Even if a figure or measurement is correct, it can still be used to bullshit, as we'll demonstrate later in the chapter. For numbers to be transparent, they must be placed in an appropriate context. Numbers must presented in a way that allows for fair comparisons.

Let's start by thinking about where people find their numbers. Some numbers are obtained directly, through an exact tally or an immediate measurement. There are 50 states in the United States of America. There are 25 prime numbers less than 100. The Empire State Building has 102 floors. Baseball legend Tony Gwynn had 3,141 hits in 9,288 at-bats, for a lifetime Major League batting average of .338. In principle, exact counts should be fairly straightforward; there is a definitive answer and usually a definitive procedure for counting or measuring that we can use to find it. Not that this process is always trivial—one can miscount, mismeasure, or make mistakes about what one is counting. Take the number of planets in the solar system. From the time that Neptune was recognized as a planet in 1846 through to 1930 when Pluto was discovered, we believed there were eight planets in our solar system. Once Pluto was found, we said there were nine planets—until the unfortunate orb was demoted to a "dwarf planet" in 2006 and the tally within our solar system returned to eight.

More often, however, exact counts or exhaustive measurements are impossible. We cannot individually count every star in the measurable universe to arrive at the current estimate of about a trillion trillion.

Likewise we rely on estimates when we consider quantities such as the average height of adults by country. Men from the Netherlands are supposedly the tallest in the world, at an average height of 183 cm (six feet), but this was not determined by measuring every Dutch man and taking an average across the entire population. Rather, researchers took random samples of men from the country, measured the members of that sample, and extrapolated from there.

If one measured only a half dozen men and took their average height, it would be easy to get a misleading estimate simply by chance. Perhaps you sampled a few unusually tall guys. This is known as sampling error. Fortunately, with large samples things tend to average out, and sampling error will have a minimal effect on the outcome.

There can also be problems with measurement procedures. For example, researchers might ask subjects to report their own heights, but men commonly exaggerate their heights—and shorter men exaggerate more than taller men.

Other sources of error, such as bias in the way a sample is selected, are more pernicious. Suppose you decide to estimate people's heights by going to the local basketball court and measuring the players. Basketball players are probably taller than average, so your sample will not be representative of the population as a whole, and as a result your estimate of average height will be too high. Most mistakes of this sort aren't so obvious. We devote the rest of the chapter to considering the subtle ways in which a sample can turn out to be uncharacteristic of the population.

In these examples, we are observing a population with a range of values—a range of heights, for instance—and then summarizing that information with a single number that we call a *summary statistic*. For example, when describing the tall Dutch, we reported a mean height. Summary statistics can be a nice way to condense information, but if you choose an inappropriate summary statistic you can easily mislead your audience. Politicians use this trick when they propose a tax cut that will save the richest 1 percent of the population hundreds of thousands of dollars but offer no tax reduction whatsoever to the rest of us. Taking a mean of the tax cuts, they report that their tax plan will save families an average $4,000 per year. Maybe so, but the average family—if by that we mean one in the middle of the income range—will save nothing. The majority of us would be better off

knowing the tax cut for a family with the *median* income. The median is the "middle" income: Half of US families earn more and half earn less. In this particular case, the median family would get no tax cut at all, because the tax cut benefits only the top 1 percent.

Sometimes one cannot directly observe the quantity one is trying to measure. Recently Carl blew through a speed trap on a straight flat highway in the Utah desert that was inexplicably encumbered with a 50 mph speed limit. He pulled over to the side of the road with the familiar red and blue lights glaring in the rearview mirror. "Do you know how fast you were going?" the state patrolman asked.

"No, I'm afraid not, Officer."

"Eighty-three miles per hour."

Eighty-three: a hard number with the potential to create serious problems. But where did this number come from? Some traffic cameras calculate speed by measuring distance you travel over a known interval of time, but that's not how the state patrol does it. The patrolman measured something different—the Doppler shift in radio waves emitted from his portable radar gun when they reflected off of Carl's speeding vehicle. The software built into his radar gun used a mathematical model grounded in wave mechanics to infer the car's speed from its measurements. Because the trooper was not directly measuring Carl's speed, his radar gun needed to be regularly calibrated. A standard method for getting out of a speeding ticket is to challenge the officer to produce timely calibration records. None of that in Carl's case, though. He knew he had been speeding and was grateful to face only a stiff fine for his haste.

Radar guns rely on highly regular principles of physics, but models used to infer other quantities can be more complicated and involve more guesswork. The International Whaling Commission publishes population estimates for several whale species. When they state that there are 2,300 blue whales left in the waters of the Southern Hemisphere, they have not arrived at this number by going through and counting each animal. Nor have they even sampled a patch of ocean exhaustively. Whales don't hold still, and most of the time you can't see them from the surface. So researchers need indirect ways to estimate population sizes. For example, they use sightings of unique individuals as identified by markings on the flukes and tail. To whatever

degree these procedures are inaccurate, their estimate of population size may be off.

There are many ways for error to creep into facts and figures that seem entirely straightforward. Quantities can be miscounted. Small samples can fail to accurately reflect the properties of the whole population. Procedures used to infer quantities from other information can be faulty. And then, of course, numbers can be total bullshit, fabricated out of whole cloth in an effort to confer credibility on an otherwise flimsy argument. We need to keep all of these things in mind when we look at quantitative claims. They say the data never lie—but we need to remember that the data often mislead.

DISTILLING NUMBERS

Though unaged whiskey has become trendy of late,* freshly distilled whiskey can often be harsh and laden with undesirable by-products from the distillation process. A couple of years in a freshly charred oak barrel (for bourbon) or a longer period in a previously used barrel (for scotch) brings about a remarkable transformation. Flavors from the wood penetrate the liquor and some of the undesirable chemicals in the liquor are extracted through the wood.

This alchemy does not transpire for free. As the liquor ages in the barrel, a portion seeps out and evaporates into the air. A barrel that begins full will hold only a fraction of its initial volume by the time the aging process is complete. The portion of spirits lost to evaporation is known as the "angels' share." Romantic imagery aside, the angels' share represents a substantial cost in the production of bourbon and scotch.

How can we best describe this cost? We could start with the total loss: Approximately 440,000 barrels of whiskey are lost to evaporation in Scotland each year. Most people don't know how big a whiskey barrel is (about 66 gallons), so we might do better to say that in Scotland, about 29 million gallons are lost every year to the angels.

* At the risk of cynicism, we conjecture that this has everything to do with the proliferation of new microdistilleries that don't want to wait three or more years before bringing in revenue, and little to do with the flavor or other characteristics of unaged whiskey. If so, much of the advertising copy about the marvels of unaged whiskey is—you guessed it—bullshit.

We usually encounter whiskey in 750 ml bottles rather than gallons, so perhaps we would want to report this as a loss of 150 million bottles a year.

Aggregate totals are hard to grasp unless one knows the total amount of scotch being produced. We could break these numbers down and describe the amount of liquid lost by a single distillery during the process of barrel aging. Running at full capacity, the large Speyside distillery Macallan loses about 220,000 LPA—liters of pure alcohol—per year. (Notice yet another type of measurement; a distillery's capacity is often reported by tallying only the alcohol produced, not the total volume including water.) By contrast, the smaller Islay distillery Ardbeg loses about 26,000 LPA per year.

Because distilleries vary widely in size, perhaps we should report loss per barrel or, better yet, as a percentage of the starting volume. During the aging process for the legendary Pappy Van Winkle twenty-three-year-old bourbon, 58 percent of its initial volume is lost to evaporation. But instead of describing the loss as a percentage of the starting volume, I could describe it as a percentage of the final volume. For this bourbon, 1.38 liters are lost to evaporation for every liter bottled, so we could report the loss as 138 percent of the final volume. It is the exact same figure as the 58 percent of starting volume described above, but this way of presenting the data makes the loss appear larger.

Of course, different whiskies are aged for different amounts of time. Maybe instead of describing the total loss, it would make more sense to describe the annual loss. Scotch whiskies lose about 2 percent in volume per year of aging, or roughly 0.005 percent per day. Bourbons typically are aged under higher temperatures than scotch, and thus experience higher rates of evaporation; some may lose upward of 10 percent per year. Moreover, the rate of loss is not constant. The aforementioned Pappy Van Winkle loses about 10 percent of its volume over its first year in the barrel, but this drops to about 3 percent per year later in the aging process.

There are other decisions to make as well. For example, alcohol and water leave the barrel at different rates; we could report changes in alcohol volume, water volume, or the total. And then there is the issue of units: Metric or imperial? Liters or milliliters? Gallons or ounces?

To tell an honest story, it is not enough for numbers to be correct. They need to be placed in an appropriate context so that a reader or listener can properly interpret them. One thing that people often overlook is that presenting the numbers by themselves doesn't mean that the numbers have been separated from any context. The choices one makes about how to represent a numerical value *sets a context* for that value.

So what does it mean to tell an honest story? Numbers should be presented in ways that allow meaningful comparisons.

As one of us (Carl) is writing this chapter, the supply of malted milk ball candies in the box of Hershey's Whoppers on his desk is gradually dwindling—but there is no guilt, because a prominent splash of color on the box announces "25% Less Fat* Than the Average of the Leading Chocolate Candy Brands." The asterisk directs us to small print reading "5 grams of fat per 30 gram serving compared to 7 grams of fat in the average of the leading chocolate candy brands." Here's an example of a number provided without enough context to be meaningful. What brands are chosen as comparators? Is this an apples-to-apples comparison or are we comparing chocolate-covered malt balls to straight chocolate bars? How about sugar? Refined sugar may be a bigger health concern than fat; are Whoppers higher or lower in sugar? Are there other bad ingredients we should be concerned about? And so on, and so forth. The 25 percent figure sounds like an important nutritional metric, but it is really just *meaningless numerosity*.

PERNICIOUS PERCENTAGES

The twelfth chapter of Carl Sagan's 1996 book, *The Demon-Haunted World,* is called "The Fine Art of Baloney Detection." In that chapter, Sagan rips into the advertising world for bombarding us with dazzling but irrelevant facts and figures. Sagan highlights the same problem we address in this chapter: people are easily dazzled by numbers, and advertisers have known for decades how to use numbers to persuade. "You're not supposed to ask," Sagan writes. "Don't think. Buy."

Sagan focuses on the marketing tactics used in drug sales, a problem that expanded in scope a year after Sagan wrote his essay. In 1997, the

United States legalized direct-to-consumer marketing of prescription drugs. But rather than take on that troubling and complex issue, let's instead consider an amusing and largely harmless example.

Arriving in Washington, DC, late one evening, Carl was searching for something to drink before bed. He picked up a packet of instant hot cocoa in the hotel lobby. "99.9% caffeine-free," the packaging boasted. Given that he was already dealing with jet lag, a 99.9 percent caffeine-free drink seemed like a prudent alternative to a cup of coffee. But pause and think about it for a minute. Even though there's a lot of water in a cup of cocoa, caffeine is a remarkably powerful drug. So is a 99.9 percent caffeine-free drink really something you want to drink right before bed?

Let's figure it out. How much caffeine is in a cup of coffee? According to the Center for Science in the Public Interest, there are 415 milligrams of caffeine in a 20-ounce Starbucks coffee. That corresponds to about 21 mg of caffeine per ounce. A fluid ounce of water weighs about 28 grams. Thus, a Starbucks drip coffee is about 0.075 percent caffeine by weight. In other words, strong coffee is also 99.9 percent caffeine free![*]

So while there's nothing inaccurate or dangerous about the 99.9 percent assertion, it's a pointless claim. Most regular coffees could be labeled in the exact same way. Nestlé has provided us with an excellent example of how something can be true and still bullshit of a sort. It is bullshit because it doesn't allow us to make meaningful comparisons the way a claim such as "only 1% as much caffeine as a cup of coffee" would.

A notorious *Breitbart* headline similarly denied readers the opportunity to make meaningful comparisons. This particular bit of scaremongering proclaimed that 2,139 of the DREAM Act recipients—undocumented adults who came to the United States as children—had been convicted or accused of crimes against Ameri-

* While we don't have an exact figure for this brand of cocoa, most cocoas have about 20 mg of caffeine in an 8-ounce cup; i.e., they're about 0.01 percent caffeine by weight. Thus we initially thought that perhaps the 99.9 percent figure referred to the powder, not the finished drink. But Nestlé's website makes it clear that they are referring to the prepared drink, not the powder: "With rich chocolate flavor and only 20 calories per single-serve packet, this cocoa makes a 99.9% caffeine-free, 8 fl oz serving." We also need to be careful about the difference between a fluid ounce, which is a measure of volume, and an ounce, which is a measure of weight. At room temperature at sea level, a fluid ounce of water weighs roughly 1.04 ounces. But we like our coffee hot, and as water approaches its boiling point its weight approaches 1.00 ounces per fluid ounce.

cans.* That sounds like a big number, a scary number. But of course, the DREAM Act pertains to a huge number of people—nearly 700,000 held DACA status concurrently and nearly 800,000 were granted DACA status at some point before the program was eliminated. This means that only about 0.3 percent of all DACA recipients—fewer than 1 in 300—have been accused of crimes against Americans. That sounds better, but how does this number compare to similar rates for American citizens? With 0.75 percent of Americans behind bars, US citizens are twice as likely to be presently incarcerated as DACA recipients are to have been accused of a crime. About 8.6 percent of American citizens have been convicted of a felony at some point in their lives, making the DACA population look better still.

Of course, DACA recipients are younger and had generally not been convicted of a crime prior to being awarded DACA status,† so they have had less time to commit crimes than your average American. But it turns out that 30 percent of Americans have been arrested for something other than a traffic violation by age twenty-three. Even assuming that *Breitbart*'s figures were correct, the news outlet presented them without the appropriate information a reader would need to put them into context.

Listing a raw total like this can make a small quantity, relatively speaking, appear to be large. We put this number into context by expressing it as a percentage. Indeed, percentages can be valuable tools for facilitating comparisons. But percentages can also obscure relevant comparisons in a number of ways. For starters, percentages can make large values look small.

In a blog post, Google VP of engineering Ben Gomes acknowledged the problem that their company faces from fake news, disinformation, and other inappropriate content:

> Our algorithms help identify reliable sources from the hundreds of billions of pages in our index. However, it's become very apparent that a small set of queries in our daily traffic

* Note that 2,139 is actually the number of DACA recipients who had lost DACA status at the publication date of the article due to felony or misdemeanor convictions, arrests, or suspected gang affiliation.

† To be eligible for DACA status, one must not have been convicted of a felony, a significant misdemeanor, or three misdemeanors of any kind.

> (around 0.25 percent), have been returning offensive or clearly misleading content, which is not what people are looking for.

Two things are going on here. First, a large and largely irrelevant number is presented as if it helps set the context: "hundreds of billions of pages in our index." Juxtaposed against this vast number is a tiny one: "0.25 percent." But the hundreds-of-billions figure is largely irrelevant; it is the number of pages indexed, not anything about the number of search queries. It doesn't matter whether ten thousand or a hundred billion pages are indexed; if 0.25 percent of Google searches are misleading, you have a one-in-four-hundred chance of getting bullshit in your search results.*

We are not told how many search queries Google handles per day, but estimates place this figure at about 5.5 billion queries a day. So while 0.25 percent sounds small, it corresponds to well over thirteen million queries a day. These two ways of saying the same thing have very different connotations. If we tell you that Google is returning inappropriate search results only one time in four hundred, the system seems pretty sound. But if we tell you that more than thirteen million queries return inappropriate and inaccurate content every day, it sounds as though we are facing a serious crisis in information delivery.

Percentages can be particularly slippery when we use them to compare two quantities. We typically talk about percentage differences: "a 40 percent increase," "22 percent less fat," etc. But what is this a percentage *of*? The lower value? The higher value? This distinction matters. In the month of December 2017, the value of the bitcoin digital currency first surged to $19,211 per unit on the seventeenth of the month, and then plummeted to a low of $12,609 per unit thirteen days later. This is a decrease of $6,602 per unit. But what was the percentage change? Should we say it was 34 percent (because $6,602 is 34.3 percent of $19,221), or is it 52 percent (because $6,602 is 52.4 percent of $12,609)?

* The 0.25 percent figure is probably an underestimate when it comes to misinformation and disinformation in the news. We don't expect a lot of misinformation in response to queries such as "molecular weight of sodium" or "late night pizza delivery near me." Misinformation about health and politics and matters of that sort must occur more than 0.25 percent of the time in order to generate an overall average of 0.25 percent.

One can make a legitimate case for either number. In general, we advocate that a percentage change be reported with respect to the starting value. In this case the starting value was $19,211, so we say bitcoin lost 34 percent of its value over those thirteen days. This can be a subtle issue, however. One would say that bitcoin lost 34 percent of its value over this period, because when we talk about a loss in value, the starting value is the appropriate comparison. But we would also say that bitcoin was apparently overvalued by 52 percent at the start of December 2017, because when we talk about something being overvalued, the appropriate baseline for comparison is our current best estimate of value. Different ways of reporting give different impressions.

Instead of listing percentage changes, studies in health and medicine often report relative risks. New drivers aged sixteen and seventeen years have some of the highest accident rates on the road. But their accident rates depend on whether they are carrying passengers, and on whom those passengers are. Compared to teen drivers without passengers, the relative risk of a teen driver dying in a crash, per mile driven, is 1.44 for teens carrying a single passenger under the age of twenty-one. This relative risk value simply tells us how likely something is compared to an alternative. Here we see that teen drivers with a young passenger are 1.44 times as likely to be killed as teen drivers without a passenger. This is easily converted to a percentage value. A teen driver with a passenger has a 44 percent higher chance of being killed than a teen driver without a passenger. Carrying an older passenger has the opposite effect on the risk of fatal crashes. Compared to teen drivers without passengers, the relative risk of a teen driver dying in a crash while carrying a passenger over the age of thirty-five is 0.36. This means that the rate of fatal crashes is only 36 percent as high when carrying an older passenger as it is when driving alone.

Relative risks can help conceptualize the impact of various conditions, behaviors, or health treatments. But they sometimes fail to provide adequate context. A worldwide study of alcohol-related disease was reported with stark headlines such as "There's 'No Safe Level of Alcohol,' Major New Study Concludes." In particular, even very modest drinking—a single drink a day—was found to have negative health consequences. That sounds bad for those of us who enjoy a beer or glass of wine with dinner. But let's look more closely.

The press release from *The Lancet,* where the study was published, reports that:

> They estimate that, for one year, in people aged 15–95 years, drinking one alcoholic drink a day increases the risk of developing one of the 23 alcohol-related health problems by 0.5%, compared with not drinking at all.

Scary stuff? To evaluate whether this is a substantial increase or not, we need to know how common are "alcohol-related health problems"—liver cirrhosis, various cancers, some forms of heart disease, self-harm, auto accidents, and other maladies—among nondrinkers. It turns out that these problems are rare among nondrinkers, occurring in less than 1 percent of the nondrinking population in a year. And while a drink per day increases this risk by 0.5 percent, that is 0.5 percent *of the very small baseline rate.* In other words, the relative risk of having one drink a day is 1.005. People who have a drink a day are 1.005 times as likely to suffer "alcohol-related disease" as those who do not drink.

The authors of the study calculated that having a single daily drink would lead to four additional cases of alcohol-related illness per 100,000 people. You would have to have 25,000 people consume one drink a day for a year in order to cause a single additional case of illness. Now the risk of low-level drinking doesn't sound as severe. To provide further perspective, David Spiegelhalter computed the amount of gin that those 25,000 people would drink over the course of the year: 400,000 bottles. Based on this number, he quipped that it would take 400,000 bottles of gin shared across 25,000 people to cause a single additional case of illness.

To be fair, this is the risk from drinking one drink a day; the risk rises substantially for those who consume larger amounts. People who drink two drinks a day have a relative risk of 1.07 (7 percent higher than nondrinkers) and those who drink five drinks a day have a relative risk of 1.37. The main point is that simply reporting a relative risk of a disease isn't enough to assess the effect unless we also know the baseline rate of the disease.

Percentages get even more slippery when we are comparing one percentage figure to another. We can look at the numerical difference

between two percentages, but we can also create a new percentage, reflecting the percentage difference between our percentage values. Even professional scientists sometimes mix up a subtle issue in this regard: the difference between *percentages* and *percentage points*. An example is by far the easiest way to illustrate the difference. Suppose that on January 1, the sales tax increases from 4 percent to 6 percent of the purchase price. This is an increase of 2 percentage points: 6% – 4% = 2%. But it is also an increase of 50 percent: The 6 cents that I now pay on the dollar is 50 percent more than the 4 cents I paid previously.

So the same change can be expressed in very different ways that give substantially different impressions. If I want to make the tax increase sound small, I can say that there has been only a 2 percentage point increase. If I want to make it sound large, I can say that taxes have gone up 50 percent. Whether accomplished by accident or intent, we need to be wary of this distinction.

Another example: On her website, an MD questions the utility of the influenza vaccine. Quoting the abstract of a medical review article,* she writes:

> In the "relatively uncommon" case where the vaccine matched the annual flu variety, the rate of influenza was 4% for the unvaccinated and 1% for the vaccinated, so 3 less people out of every 100 would become ill. In the more common scenario of vaccine/flu mis-match the corresponding numbers were 2% sick among unvaccinated [and] 1% of the vaccinated, thus 1 less person per 100.

Her argument appears to be that when mismatched, the influenza vaccine is near to useless because it reduces influenza cases by a mere one person per hundred per year. The MD goes on to say that in lieu of vaccination, "I have a new 'health tonic' this year which I will try out if I feel some illness approaching and let you know [if] it works."

* In addition to our other arguments about the value of the influenza vaccine, the 2 percent influenza attack rate cited here is surprisingly low. The CDC estimates that since 2010, influenza attack rates in the US have ranged from about 3 percent to 11 percent. Part of this discrepancy is explained by the fact that this review paper looks only at studies of adults aged eighteen to sixty-four, whereas influenza rates are much higher in children, and part may be due to the international nature of the studies reviewed. Influenza rates in the US are commonly estimated to be several times higher than 2 percent even in adults.

Does that sound reasonable? If we think of the flu vaccine as helping only one person in a hundred, it may seem that some unspecified "health tonic" could perform as well. But the claim is misleading. First, notice that even in these years when the influenza vaccine is ineffective, it cuts the number of influenza cases among vaccinated individuals in half. Because influenza is relatively infrequent in the population, a 1 *percentage point* decrease in incidence corresponds to a 50 *percent* decrease.

Second, we can put these numbers into context with a well-chosen comparison. Like influenza vaccines in an ineffective year, the use of seatbelts drops the annual risk of injury from "only" about 2 percent to 1 percent.* Would you trust your health to an MD who advocated a "health tonic" in lieu of seatbelts?

Reporting numbers as percentages can obscure important changes in net values. For example, the US incarcerates African Americans at a shockingly high rate relative to members of other groups. African Americans are more than five times as likely as whites to be in jail or prison. In the year 2000, African Americans composed 12.9 percent of the US population but a staggering 41.3 percent of the inmates in US jails. Given that, it would seem to be good news that between the year 2000 and the year 2005, the fraction of the jail population that was African American declined from 41.3 percent to 38.9 percent.

But the real story isn't as encouraging as the percentage figures seem to indicate. Over this period, the number of African Americans in US jails actually *increased* by more than 13 percent. But this increase is obscured by the fact that over the same period, the number of Caucasians in US jails increased by an even larger fraction: 27 percent.

This is an example of a more general issue with comparisons that involve fractions or percentages: Changing denominators obscure changes in numerators. The numerator—the top part of the fraction, here the number of jailed African Americans—increased substantially from 2000 to 2005. But the denominator—the bottom part of the fraction, here the total number of jailed Americans—increased by an

* The National Safety Council estimates that in 2016, there were approximately 4.6 million injuries from automobile accidents that required medical consultation or treatment, corresponding to about 1.4 percent of the US population. The National Highway Traffic Safety Administration estimates that wearing a seatbelt reduces the risk of moderate to critical injury by 50 percent to passengers traveling in the front seat of a passenger car, and by even greater amounts to passengers traveling in light trucks.

even larger proportion over the same period. As a result, African Americans made up a smaller fraction of the incarcerated population in 2005 than they did in 2000. The fact that in 2005 more African Americans were locked up than ever before is obscured by this changing denominator.

Changing denominators wreak havoc on percentages. Consider the following. If the Dow Jones Industrial Average should rise by 10 percent today and then fall 10 percent tomorrow, you might think that it would end up right where it started—but it won't. Suppose the Dow is at 20,000. A 10 percent increase takes it up 2,000 points to 22,000. A subsequent 10 percent decrease, relative to its new higher value of 22,000, drops it 2,200 points to 19,800. It doesn't matter whether it rises and then drops or vice versa; you lose either way. If the Dow starts at 20,000 and loses 10 percent, it falls to 18,000. A 10 percent increase then takes it up to 19,800—the same place it ended up when it rose first and fell second. Be on the lookout for these kinds of perverse consequences of reporting data in terms of percentage changes.

Let's consider more of the strange ways that percentages can behave. In late April 2016, Uber was delivering about 161,000 rides per day in New York City while Lyft was delivering about 29,000 a day. A year later, Uber was delivering 291,000 rides per day and Lyft about 60,000 rides per day. This was an overall increase of 161,000 rides per day, from 190,000 rides per day to 351,000 rides per day. Of those 161,000 extra rides, Lyft contributed about 31,000 of them. Thus Lyft is responsible for about 16 percent of the increase, and Uber for the rest. So far so good.

Over the same period, the number of yellow taxi rides plummeted from 398,000 per day to 355,000. If we look at the total number of rides by yellow taxi, Uber, or Lyft, we see a net increase, from 588,000 to 696,000. That's a net increase of 108,000 rides per day. We already know that Lyft increased its rides by 31,000 per day over this period. So it looks like we could say that Lyft is responsible for $31{,}000 / 108{,}000 \times 100 = 29\%$ of the increase in overall rides.

But that is odd. We said Lyft is responsible for about 24 percent of the increase in rides that ride-hailing services provide, but now we're saying that Lyft is responsible for about 34 percent of the increase in rides by ride-hailing *or* taxi. How could both things be true? Things get even weirder if we look at Uber's contribution to this increase in

rides. Uber provided an extra 383,000 rides per day by the end of this period. We could say that the increase in Uber rides is responsible for 130,000 / 108,000 × 100 = 120% of the increase in total rides. What does that even mean? How could the Uber increase account for more than 100 percent of the total increase?

Percentage calculations can give strange answers when any of the numbers involved are negative. In general, if we can break the change down into different categories, and any of those categories decrease, we should not talk about percentage contributions to the total change. This is how Governor Scott Walker was able to claim in June 2011 that 50 percent of the nation's job growth had occurred in his home state of Wisconsin. What had actually happened was that a number of US states lost jobs overall, while others gained. These changes nearly balanced out, leaving a net change of only about 18,000 jobs. Wisconsin's net growth of 9,500 jobs was over half the size of the country's net growth, even though only a small fraction of the total jobs created in the country were created in Wisconsin.

GOODHART'S LAW

When scientists measure the molecular weights of the elements, the elements do not conspire to make themselves heavier and connive to sneak down the periodic table. But when administrators measure the productivity of their employees, they cannot expect these people to stand by idly: Employees want to look good. As a result, every time you quantify performance or rank individuals, you risk altering the behaviors you are trying to measure.

At the turn of the twentieth century, Vietnam was a part of the French colonies collectively known as French Indochina. Hanoi was developing into a prosperous modern city, and its sewer system provided European-style sanitation, mainly for the wealthy white neighborhoods around the city. Unfortunately, the sewer system also provided an ideal breeding ground for rats, which not only emerged from the sewers to terrorize the inhabitants but also spread diseases such as the bubonic plague. In an effort to rid the city of these pests, the colonial bureaucracy hired rat catchers to go into the sewers, and eventually offered a bounty on the creatures, paying a small reward for each rat tail delivered to the colonial offices. Many of the inhabi-

tants of Hanoi were pleased to be a part of this enterprise, and rat tails started pouring in.

Before long, however, the inhabitants of Hanoi started to see tailless rats skulking in the sewers. Apparently, the rat hunters preferred not to kill their prey. Better to slice off the tail for bounty and leave the rats to reproduce, guaranteeing a steady supply of tails in the future. Entrepreneurs imported rats from other cities, and even began raising rats in captivity to harvest their tails. The bounty program failed because people did what people always do at the prospect of a reward: They start gaming the system.

The same thing happens even when you aren't offering rewards directly. Take college rankings such as the one compiled by *U.S. News & World Report*. This survey accounts for numerous aspects of colleges, including the acceptance rates for applicants and the average SAT scores of the incoming classes. Once these college rankings began to influence the number of applicants, admissions departments started to play the angles. To lower their acceptance rates and thereby appear more selective, some schools aggressively recruited applications, even from students unlikely to be admitted. Many schools switched to the common application so that candidates could apply by simply checking a box. To get credit for teaching small classes, some universities capped many class sizes at eighteen or nineteen—just below the *U.S. News & World Report* class-size cutoff of twenty. To lift their average SAT scores, schools employed a range of stratagems: not requiring SAT for international applicants who typically have lower scores, bringing in lower-scoring students in the spring semester when their scores are not counted, or even paying admitted students to retake the SAT, with bonuses for substantive increases in score. These efforts to game the system undercut the value of the rankings. The rankings ended up being influenced as much by admissions departments' willingness to chase metrics as they did by the quality of schools' applicants.

This problem is canonized in a principle known as Goodhart's law. While Goodhart's original formulation is a bit opaque,* anthropologist Marilyn Strathern rephrased it clearly and concisely:

* Goodhart originally expressed his law as: "Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes."

> When a measure becomes a target, it ceases to be a good measure.

In other words, if sufficient rewards are attached to some measure, people will find ways to increase their scores one way or another, and in doing so will undercut the value of the measure for assessing what it was originally designed to assess.

We see this in any number of domains. In the sciences, the use of citation metrics to measure journal quality has led editors to game the system. Some pressure authors to include citations to papers in the same journal. Some publish an excess of articles in January, when they have the most time to be cited during the year. Others publish annual summary papers that cite many of the articles published within a year; yet others shift their focus to disciplines or types of articles that tend to attract more citations. All of these perverse behaviors undermine the mission of the journals and the effectiveness of citation measures as indicators of quality.

When the manager of an auto dealership provides bonuses to salespeople for hitting certain volume targets, the sales force will have incentives to offer larger discounts to their customers, in order to make the sale quickly. If salespeople were trying to maximize profits in making each sale, one might be able to use the number of cars sold to approximate how much money each salesperson brings in. But once the number of cars sold becomes a target, salespeople alter their sales strategies, more cars will be sold, and yet more cars sold will not necessarily correspond to higher profits as salespeople offer bigger discounts to make sales and make them quickly.

At around the same time that Goodhart proposed his law, psychologist Donald Campbell independently proposed an analogous principle:

> The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor.

Campbell illustrated his principle with the case of standardized testing in education:

> Achievement tests may well be valuable indicators of general school achievement under conditions of normal teaching aimed at general competence. But when test scores become the goal of the teaching process, they both lose their value as indicators of educational status and distort the educational process in undesirable ways.

If no one knew that you were going to evaluate schools by looking at test scores of students, test scores might provide a reasonable way of measuring schools' effectiveness. But once teachers and administrators recognize that test scores will be used to evaluate their effectiveness, they have every incentive to find ways of raising their students' scores—even at the expense of the quality of their education. They may "teach to the test," for example, rather than teaching critical thinking. The consequences are twofold. Once this happens, test scores lose much of their value as a means of assessing a school's performance. The process of measurement may even actively undermine the quality of education, as time and effort are shifted away from valuable activity and toward the kind of memorization that improves test scores but is nearly useless from an educational perspective.

The same problem occurs in the sciences as well. A few years back, Jevin penned a commentary about the dangers of using quantitative metrics to measure the quality of scientific research. Getting the wrong answer is not the worst thing a metric can do. It's much worse to provide people with incentives to do bad science.

Campbell stresses the social element of this problem. As we mentioned at the start of this section, we don't have to worry about the laws of physics or chemistry gaming the metrics. Hydrogen doesn't care what you think its emission spectra are, or that its highest wavelength band is lower than that of helium. People do care about how they are measured. What can we do about this? If you are in the position to measure something, think about whether measuring it will change people's behaviors in ways that undermine the value of your results. If you are looking at quantitative indicators that others have compiled, ask yourself: Are these numbers measuring what they are intended to measure? Or are people gaming the system and rendering this measure useless?

MATHINESS

Online voters selected *truthiness* as the word of the year in a 2006 survey run by the publishers of the *Merriam-Webster* dictionary. The term, coined in 2005 by comedian Stephen Colbert, is defined as "the quality of seeming to be true according to one's intuition, opinion, or perception without regard to logic, factual evidence, or the like." In its disregard for actual logic and fact, this hews pretty closely to our definition of bullshit.

We propose an analogous expression, *mathiness*. Mathiness refers to formulas and expressions that may look and feel like math—even as they disregard the logical coherence and formal rigor of actual mathematics.

Let's begin with an example. The following formula, known as the VMMC Quality Equation, is apparently much discussed in the area of healthcare quality management.

$$Q = A \times \frac{O + S}{W}$$

Q: Quality
A: Appropriateness
O: Outcomes
S: Service
W: Waste

This looks like a rigorous mathematical way to consider patient care. But what does it actually mean? How would these quantities be measured, and in what units? Why does it take the form that it does? It is not clear whether any of these questions have satisfying answers.

We see mathiness as classic bullshit in the vein of our original definition. These equations make mathematical claims that cannot be supported by positing formal relationships—variables interacting multiplicatively or additively, for example—between ill-defined and impossible-to-measure quantities. In other words, mathiness, like truthiness and like bullshit, involves a disregard for logic or factual accuracy. Also, as in our definition of bullshit, mathiness is often produced for the purpose of impressing or persuading an audience, trad-

ing on the air of rigor conferred by mathematical equations, and throwing in a healthy dose of algebraic shock and awe for good measure. (We don't mean to attribute malice; some snake oil salesmen believe in the restorative powers of their tonics, and some peddlers of mathiness may think they are providing deep insights.)

Most examples of mathiness are not made up at random. Rather, they are designed to express basic truths about a given system. For example, the VMMC Quality Equation represents the idea that higher levels of Appropriateness, Outcomes, and Service all lead to higher Quality, whereas greater Waste decreases Quality.* This information could equally well be expressed in a simple table:

Parameter	Effect on Quality
Appropriateness	+
Outcomes	+
Service	+
Waste	−

All of this is implied by the Quality Equation, but there are many other equations that have the same properties. The formula $Q = S \times \frac{O + A}{W}$ also reflects the qualitative relationship shown in this table, as does $Q = (A + O) \times S - W$. For that matter, so does $Q = \sqrt[W]{A^O + S^O}$. If one is not able to explain why the VMMC Equation is $Q = A \times \frac{O + S}{W}$ and not any of these alternatives, the relationship should not be dignified with an equation in the first place.

The so-called Trust Equation, laid out in a 2005 book of the same title, provides us with another illustration of mathiness. According to the Trust Equation,

$$\text{Trust} = \frac{\text{Credibility} + \text{Reliability} + \text{Authenticity}}{\text{Perception of self-interest}}$$

As with the VMMC Equation, the general direction of the relationships seems right. Trust increases with credibility, reliability, and

* In more technical language, we suspect that equations with false mathiness are typically intended to express something about the signs of the partial derivatives of a function, to an audience not necessarily used to thinking about partial derivatives. For example, the main information in the VMMC Equation is probably that Quality can be represented as a function $Q = f(A,O,S,W)$ with $df/dA > 0$, $df/dO > 0$, $df/dS > 0$, and $df/dW < 0$. Beyond that, the functional form may be largely arbitrary.

authenticity, and decreases with perception of self-interest. So far, so good. But again there are any number of other equations that work this way as well. And the Trust Equation has some very specific implications for how trust works. Perhaps the strongest of these occurs because the sum of the other terms is *divided* by the perception of self-interest. This means that (provided the sum of the other terms is positive), the degree of trust becomes very large as the perception of self-interest gets very small. If one could eliminate the perception of self-interest entirely, trust should be infinitely strong! But the world doesn't work like that. Suppose I flip a coin to tell me what stock to buy. The coin is inarguably free from self-interest, so according to the equation I should trust it infinitely. In reality, why would I trust a random device rather than an expert predictor?

Second, notice that credibility, reliability, and authenticity are added together in the top part of the equation. That means that no matter how big or small credibility, reliability, and authenticity are, a one-unit increase in reliability has the same effect as a one-unit increase in authenticity. It also means that trust can be high even if one or two of these quantities is zero. If authenticity is high and perceived self-interest is low, the equation predicts that trust should be high even when the speaker has no credibility and is entirely unreliable. Again, this seems like an unintended consequence of expressing some general trends as a specific formula in order to create the perception of mathiness.

A somewhat more technical matter is the issue of units. Readers may remember learning dimensional analysis in high school. This is the process of tracking the units in a calculation to make sure that one's answer is in the right currency.

For example,* if a cyclist goes 45 miles in 3 hours, we don't just divide 45 by 3 and say that her speed is 15; we need to state the units of measurement. We treat the units much as we treat the numbers themselves, writing them out and simplifying, if necessary: 45 miles / 3 hours = 15 miles/hour.

We all know that the numerical values on each side of an equation

* A slightly more complicated example: In the Guinness Gastropod Championship, a snail covered a 13-inch course in 140 seconds. How fast was it going?

$$\frac{13 \text{ in.}}{140 \text{ sec.}} \times \frac{1 \text{ ft.}}{12 \text{ in.}} \times \frac{1 \text{ mile}}{5280 \text{ ft.}} \times \frac{60 \text{ sec.}}{1 \text{ min.}} \times \frac{60 \text{ min.}}{1 \text{ hr.}} = 0.0053 \text{ mile/hr.}$$

have to be the same. The key to dimensional analysis is that the units have to be the same as well. This provides a convenient way to keep careful track of units when making calculations in engineering and other quantitative disciplines, to make sure one is computing what one thinks one is computing.

When an equation exists only for the sake of mathiness, dimensional analysis often makes no sense. Let's look at an example. Every January, news outlets run or rerun a story about how scientists have found that "Blue Monday"—the third Monday of the month—is the saddest day of the year. This claim, advanced in various press releases, is based upon a formula proposed by a part-time tutor at Cardiff University named Cliff Arnall:

$$\frac{(W + D - d) * T^{Q}}{M * N_a}$$

The press release explains that W represents the weather, d debt, T time since Christmas, Q time since abandoning New Year's resolutions, M low levels of motivation, and N_a the need to take action. (It is unclear what D represents in the formula.)

By now you may be able to pick this example apart. For starters, it is unclear how "weather," "low levels of motivation," or "the need to take action" would be quantified. It gets worse. Any quantities that are added or subtracted have to be in the same units. We can't sensibly add miles and hours, for example, let alone mental exertion and Reuben sandwiches. But look at the Blue Monday equation. The first term, $(W + D - d)$, is presumably some sort of difference between weather and debt. What units could possibly apply to both quantities? Another rule is that when you raise a quantity to a power, that power has to be a dimensionless number. It makes sense to talk about time raised to the power of two, but not about time raised to the power of two milliseconds or two parakeets. But Q is not dimensionless; it is in units of time—and thus the expression T^{Q} makes no sense. Even if we could resolve these issues, we would run into any number of problems already discussed. We might think that sadness should increase with the need to take action and with low levels of motivation, but both quantities are in the denominator, meaning that sadness decreases with the need to take action and with low levels of motivation.

Martin Seligman, former president of the American Psychological Association, developed a famous formula for happiness: $H = S + C + V$, where S is a "set point" representing one's innate predisposition toward happiness, C is one's circumstances, and V is those aspects of life under voluntary control. Again, think about units. Even if set point, circumstances, and aspects under voluntary control were quantifiable, what could be the common unit in which all three are measured? And even if such a common unit exists, would this also be the appropriate unit to measure happiness, as implied by the equation? If Seligman had claimed that happiness *was a function* of S, C, and V, we might have been willing to go along. But to claim that it is the *mathematical sum* of these three quantities is an exercise in mathiness.

Perhaps we aren't supposed to take this equation-building approach literally. Why can't we use these equations as a metaphor or a mode of expression? Mathematical equations are useful precisely for their precision of expression—but all of these examples fail to deliver this precision. You make a false promise when you suggest that happiness is an arithmetic sum of three things, if all you mean is that it increases with each of them. It is like promising chia-seed-and-spelt waffles sweetened with wildflower honey; topped with blackberry compote, a lingonberry reduction, and house-whipped sweet cream—but delivering an Eggo with artificial maple syrup. You can claim your description wasn't meant to be taken literally, but why not say what you mean in the first place?

We do not know why so many writers are prone to invent equations that layer a veneer of mathiness atop their qualitative arguments. Obviously it impresses some of the people, some of the time, but why aren't these authors embarrassed to be caught out by the rest of us? Perhaps without a clear understanding of everything a mathematical equation entails, they fail to see the scale at which their faux equations underdeliver.

ZOMBIE STATISTICS

We close this chapter with a cautionary note. Look in sufficiently nerdy corners of the Internet, and you can find T-shirts, coffee mugs, mouse pads, and any number of other paraphernalia emblazoned with a slogan along the lines of "78.4% of statistics are made up on the

spot." The precise number varies—53.2 percent, 72.9 percent, 78.4 percent, 84.3 percent—but the joke, of course, is that this particular statistic is among the made-up numbers it purports to describe.

Like most good jokes, this quip carries an element of truth. Without knowing the source and context, a particular statistic is worth little. Yet numbers and statistics appear rigorous and reliable simply by virtue of being quantitative, and have a tendency to spread. And as a result, there are a lot of *zombie statistics* out there. Zombie statistics are numbers that are cited badly out of context, are sorely outdated, or were entirely made up in the first place—but they are quoted so often that they simply won't die.

By way of example: We both spend a good deal of time studying the way that scientists make use of the scientific literature. Not a month goes by without someone repeating to us an old trope that 50 percent of scientific articles are never read by anyone. But where does this "50 percent" figure come from? No one seems to know.

Management professor Arthur Jago became curious about this number and set out to verify it. Jago first saw the claim, unsupported by any citation, in a *Chronicle of Higher Education* commentary. The author of that commentary pointed him to an article in *Smithsonian*. The citation given for this claim in *Smithsonian* was incorrect, but Jago was able to track it to a 2007 article in *Physics World*. The author of that article did not have a citation; the figure had been added by a journal editor late in the publication process. This editor, once reached, thought he perhaps got the material from a set of 2001 course notes. The instructor who wrote those course notes did not have a citation and provided only the unencouraging remark that "everything in those notes had a source, but whether I cross-checked them all before banging the notes out, I doubt."

Though this trail went dead, eventually Jago tracked the 50 percent statistic back to a pair of papers published in 1990 and 1991 in the journal *Science*. The original claim was that over 50 percent of papers go *uncited,* not unread. It's a big difference; a paper may be accessed in the library or downloaded from the Internet hundreds of times for every citation it receives.

Even with this correction, the 50 percent figure is inaccurate. First, this number represents the fraction of papers uncited after four years, not the fraction that will remain uncited forever. In some fields, such

as mathematics, citations scarcely begin to accrue after four years. Second, this statistic is derived by looking at citation records in the Clarivate Web of Science. That database covers only a subset of journals, varies widely in how thoroughly it covers different fields, and does not index some of the most important types of publications—books in the humanities and conference proceedings in engineering. Third, to arrive at the 50 percent figure, the author considered citations to all types of articles published in scientific journals, including letters to the editor, news articles, opinion pieces, book reviews, even obituaries and errata. Saying that many book reviews and obituaries are uncited is very different from saying that many scientific articles are uncited.

Although numerous scholars challenged the 50 percent figure, they failed to stop its spread. Even Eugene Garfield—the father of citation analysis and founder of the Science Citation Index that was used to arrive at this number—also attempted to correct the record. The horse was out of the barn, however, and there was no getting it back. Garfield was resigned to this, and he responded to the error pessimistically, quoting Peggy Thomasson and Julian C. Stanley:

> The uncritical citation of disputed data by a writer, whether it be deliberate or not, is a serious matter. Of course, knowingly propagandizing unsubstantiated claims is particularly abhorrent, but just as many naïve students may be swayed by unfounded assertions presented by a writer who is unaware of the criticisms. Buried in scholarly journals, critical notes are increasingly likely to be overlooked with the passage of time, while the studies to which they pertain, having been reported more widely, are apt to be rediscovered.

This quotation refers to any sort of unsubstantiated claim, but simple figures and statistics are particularly prone to this type of spread. They leave context behind, and whatever halfhearted caveats were initially in place become lost as source after source repeats the basic number with none of the background.

In this chapter, we have seen that although numbers may seem to be pure facts that exist independently from any human judgment, they are heavily laden with context and shaped by decisions—from how

Passy, Jacob. "Another Adverse Effect of High Home Prices: Fewer Babies." MarketWatch. June 9, 2018. https://www.marketwatch.com/story/another-adverse-effect-of-high-home-prices-fewer-babies-2018-06-06.

Schaffer, Jonathan. "The Metaphysics of Causation." Stanford Encyclopedia of Philosophy. 2016. https://plato.stanford.edu/entries/causation-metaphysics/.

Shoda, Y., W. Mischel, and P. K. Peake. "Predicting Adolescent Cognitive and Self-regulatory Competencies from Preschool Delay of Gratification: Identifying Diagnostic Conditions." *Developmental Psychology* 26 (1990): 978.

Sies, Helmut. "A New Parameter for Sex Education." *Nature* 332 (1988): 495.

Sumner, P., S. Vivian-Griffiths, J. Boivin, A. Williams, C. A. Venetis, A. Davies et al. "The Association between Exaggeration in Health Related Science News and Academic Press Releases: Retrospective Observational Study." *British Medical Journal* 349 (2014): g7015.

Tucker, Jeff. "Birth Rates Dropped Most in Counties Where Home Values Grew Most." Zillow. June 6, 2018. https://www.zillow.com/research/birth-rates-home-values-20165/.

Vigen, Tyler. "Spurious Correlations." 2015. http://www.tylervigen.com/spurious-correlations.

Watts, T. W., G. J. Duncan, and H. Quan. "Revisiting the Marshmallow Test: A Conceptual Replication Investigating Links between Early Delay of Gratification and Later Outcomes." *Psychological Science* 29 (2018): 1159–77.

Zoldan, Ari. "40-Year-Old Stanford Study Reveals the 1 Quality Your Children Need to Succeed in Life." *Inc.* February 1, 2018.

CHAPTER 5: NUMBERS AND NONSENSE

Binder, John. "2,139 DACA Recipients Convicted or Accused of Crimes against Americans." *Breitbart*. September 5, 2017. http://www.breitbart.com/big-government/2017/09/05/2139-daca-recipients-convicted-or-accused-of-crimes-against-americans/.

Bogaert, A. F., and D. R. McCreary. "Masculinity and the Distortion of Self-Reported Height in Men." *Sex Roles* 65 (2011): 548.

Campbell, D. T. "Assessing the Impact of Planned Social Change." *Evaluation and Program Planning* 2 (1979): 67–90.

Camper, English. "How Much Pappy Van Winkle Is Left after 23 Years in a Barrel?" Alcademics. January 15, 2014. http://www.alcademics.com/2014/01/how-much-pappy-van-winkle-is-left-after-23-years-in-a-barrel-.html.

Center for Science in the Public Interest. "Caffeine Chart." December 2016. https://cspinet.org/eating-healthy/ingredients-of-concern/caffeine-chart.

Centers for Disease Control and Prevention. "Disease Burden of Influenza." 2018. https://www.cdc.gov/flu/about/disease/burden.htm.

Cimbala, John M., and Yunus A. Çengel. "Dimensional Analysis and Modeling," Section 7-2: "Dimensional Homogeneity." In *Essential of Fluid Mechanics: Fundamentals and Applications*. New York: McGraw-Hill, 2006.

Drozdeck, Steven, and Lyn Fisher. *The Trust Equation*. Logan, Utah: Financial Forum Publishing, 2005.

Ellenberg, Jordan. *How Not to Be Wrong: The Power of Mathematical Thinking*. New York: Penguin Press, 2014.

Garfield, Eugene. "I Had a Dream . . . about Uncitedness." *The Scientist*. July 1998.

Goodhart, Charles. "Problems of Monetary Management: The U.K. Experience." In *Inflation, Depression, and Economic Policy in the West,* edited by Anthony S. Courakis, 111–46. Lanham, MD: Rowman & Littlefield.

Gordon, Dr. Deborah. *2015 Flu Season*. https://www.drdeborahmd.com/2015-flu-season.

Hamilton, D. P. "Publishing by—and for?—the Numbers." *Science* 250 (1990): 1331–32.

———. "Research Papers: Who's Uncited Now?" *Science* 251 (1991): 25.

Heathcote, Elizabeth. "Does the Happiness Formula Really Add Up?" *Independent*. June 20, 2010. https://www.independent.co.uk/life-style/health-and-families/features/does-the-happiness-formula-really-add-up-2004279.html.

Hines, Nick. "The Amount of Scotch Lost to the Angel's Share Every Year Is Staggering." *Vinepair*. April 11, 2017. https://vinepair.com/articles/what-is-angels-share-scotch/.

Howell, Elizabeth. "How Many Stars Are in the Universe?" *Space.com*. May 18, 2017. https://www.space.com/26078-how-many-stars-are-there.html.

International Whaling Commission. "Population (Abundance) Estimates." 2018. https://iwc.int/estimate.

Jago, Arthur G. "Can It Really Be True That Half of Academic Papers Are Never Read?" *Chronicle of Higher Education*. June 1, 2018.

Jefferson, T., C. Di Pietrantonj, A. Rivetti, G. A. Bawazeer, L. A. Al-Ansary, and E. Ferroni. "Vaccines for Preventing Influenza in Healthy Adults." *Cochrane Library* (2010). https://doi.org/10.1002/14651858.CD001269.pub6.

The Keyword (blog). "Our Latest Quality Improvements for Search." Ben Gomes. Google. April 25, 2017. https://blog.google/products/search/our-latest-quality-improvements-search/.

Kutner, Max. "How to Game the College Rankings." *Boston*. August 26, 2014.

"*The Lancet*: Alcohol Is Associated with 2.8 Million Deaths Each Year Worldwide." Press release. American Association for the Advancement of Science. August 23, 2018. https://www.eurekalert.org/pub_releases/2018-08/tl-tla082218.php.

Molinari, N-A. M., I. R. Ortega-Sanchez, M. L. Messonnier, W. W. Thompson, P. M. Wortley, E. Weintraub, and C. B. Bridges. "The Annual Impact of Seasonal Influenza in the US: Measuring Disease Burden and Costs." *Vaccine* 25 (2007): 5086–96.

National Highway Traffic Safety Administration. "Seat Belts." 2016. https://www.nhtsa.gov/risky-driving/seat-belts.

National Safety Council. "NSC Motor Vehicle Fatality Estimates." 2017. https://www.nsc.org/portals/0/documents/newsdocuments/2017/12-month-estimates.pdf.

NCD Risk Factor Collaboration. "A Century of Trends in Adult Human Height." *eLife* 5 (2016): e13410.

Pease, C. M., and J. J. Bull. *Think Critically*. Ebook. Biology for Business, Law and Liberal Arts (Bio301d) course, University of Idaho. https://bio301d.com/scientific-decision-making/.

Reuter, P. "The (Continued) Vitality of Mythical Numbers." *The Public Interest* 75 (1984): 135.

Silversin, J., and G. Kaplan. "Engaged Physicians Transform Care." Presented at the 29th Annual National Forum on Quality Improvement in Health Care. Slides at http://app.ihi.org/FacultyDocuments/Events/Event-2930/Presentation-15687/Document-12690/Presentation_Q6_Engaged_Physicians_Silversin.pdf.

Spiegelhalter, David. "The Risks of Alcohol (Again)." *Medium*. August 24, 2018. https://medium.com/wintoncentre/the-risks-of-alcohol-again-2ae8cb006a4a.

Tainer, H. A., et al. "Science, Citation, and Funding." *Science* 251 (1991): 1408–11.

Tefft, B. C., A. F. Williams, and J. G. Grabowski. "Teen Driver Risk in Relation to Age and Number of Passengers, United States, 2007–2010." *Traffic Injury Prevention* 14 (2013): 283–92.

Todd W. Schneider (blog). "Taxi, Uber, and Lyft Usage in New York City." Schneider, Todd. April 5, 2016. http://toddwschneider.com/posts/taxi-uber-lyft-usage-new-york-city/.

"Truthiness." Dictionary.com. http://www.dictionary.com/browse/truthiness.

"Use this Equation to Determine, Diagnose, and Repair Trust." *First Round Review*. 2018. http://firstround.com/review/use-this-equation-to-determine-diagnose-and-repair-trust/.

Van Noorden, Richard. "The Science That's Never Been Cited." *Nature* 552 (2017): 162–64.

Vann, M. G. "Of Rats, Rice, and Race: The Great Hanoi Rat Massacre, an Episode in French Colonial History." *French Colonial History* 4 (2003): 191–203.

Welsh, Ashley. "There's 'No Safe Level of Alcohol,' Major New Study Concludes." CBS News. August 23, 2018. https://www.cbsnews.com/news/alcohol-and-health-no-safe-level-of-drinking-major-new-study-concludes/.

West, Jevin. "How to Improve the Use of Metrics: Learn from Game Theory." *Nature* 465 (2010): 871–72.

CHAPTER 6: SELECTION BIAS

Aldana, S. G. "Financial Impact of Health Promotion Programs: A Comprehensive Review of the Literature." *American Journal of Health Promotion* 15 (2001): 296–320.

Baicker, K., D. Cutler, and Z. Song. "Workplace Wellness Programs Can Generate Savings." *Health Affairs* 29 (2010): 304–11.

Carroll, Aaron E. "Workplace Wellness Programs Don't Work Well. Why Some Studies Show Otherwise." *The New York Times*. August 6, 2018.

Chapman, L. S. "Meta-Evaluation of Worksite Health Promotion Economic Return Studies: 2005 Update." *American Journal of Health Promotion* 19 (2005): 1–11.

they are calculated to the units in which they are expressed. In the next chapter we will look at some of the ways that *collections* of numbers are assembled and interpreted—and how such interpretations can be misleading. Without digging into any of the formal mathematics, we will explore the problems associated with picking a representative sample to study. We will look at how unrepresentative samples lead to unjustified conclusions, and can even be used by disingenuous authors to mislead their audiences.