Cambridge University Press
978-1-107-13057-9 - Truth or Truthiness: Distinguishing Fact from Fiction by Learning to Think Like a Data Scientist
Howard Wainer
Excerpt
More information

Introduction

The modern method is to count;
The ancient one was to guess.
Samuel Johnson

In the months leading up to Election Day 2012 we were torn between two very different kinds of outcome predictions. On one side were partisans, usually Republicans, telling us about the imminent defeat of President Obama. They based their prognostication on experience, inside information from "experts," and talking heads from Fox News. On the other side, were "the Quants" represented most visibly by Nate Silver, whose predictions were based on a broad range of polls, historical data, and statistical models. The efficacy of the former method was attested to by snarky experts, armed with anecdotes and feigned fervor, who amplified the deeply held beliefs of their colleagues. The other side relied largely on the stark beauty of unadorned facts. Augmenting their bona fides was a history of success in predicting the outcomes of previous elections, and, perhaps even more convincing, was remarkable prior success, using the same methods, in predicting the outcome of a broad range of sporting events.

It would be easy to say that the apparent supporters of an anecdote-based approach to political prediction didn't really believe their own hype, but were just pretending to go along to boost their own paychecks.[1] And perhaps that cynical conclusion was often true. But how

[1] I am thinking of Upton Sinclair's observation that "it is difficult to get someone to understand something if their paycheck depends on their not understanding it."

are we to interpret the behavior of major donors who continued to pour real money into what was almost surely a rat hole of failure? And what about Mitt Romney, a man of uncommon intelligence, who appeared to believe that in January 2013, he was going to be moving into The White House? Perhaps, deep in his pragmatic and quantitative soul, he knew that the presidency was not his destiny, but I don't think so. I believe that he succumbed to that most natural of human tendencies, the triumph of hope over evidence.

We need not reach into the antics of America's right wing to find examples of humanity's frequent preference for magical thinking over empiricism; it is widespread. Renée Haynes (1906–94), a writer and historian, introduced the useful concept of a *boggle threshold*: "the level at which the mind boggles when faced with some new idea." The renowned Stanford anthropologist Tanya Luhrmann (2014) illustrates the boggle threshold with a number of examples (e.g., "A god who has a human son whom he allows to be killed is natural; a god with eight arms and a lusty sexual appetite is weird."). I would like to borrow the term, but redefine it using her evocative phrase, as the place "where reason ends and faith begins."

The goal of this book is to provide an illustrated toolkit to allow us to identify that line – that place beyond which evidence and reason have been abandoned – so that we can act sensibly in the face of noisy claims that lie beyond the boggle threshold.

The tools that I shall offer are drawn from the field of data science. The character of the support for claims made to the right of the boggle threshold we will call their "truthiness."

> Data science is the study of the generalizable extraction of knowledge from data.
>
> *Peter Naur 1960*

> Truthiness is a quality characterizing a "truth" that a person making an argument or assertion claims to know intuitively "from the gut" or because it "feels right" without regard to evidence, logic, intellectual examination, or facts.
>
> *Stephen Colbert, October 17, 2005*

Data science is a relatively recent term coined by Peter Naur but expanded on by statisticians Jeff Wu (in 1997) and Bill Cleveland (in

2001). They characterized data science as an extension of the science of statistics to include multidisciplinary investigations, models and methods for data, computing with data, pedagogy, tool evaluation, and theory. The modern conception is a complex mixture of ideas and methods drawn from many related fields, among them signal processing, mathematics, probability models, machine learning, statistical learning, computer programming, data engineering, pattern recognition and learning, visualization, uncertainty modeling, data warehousing, and high-performance computing. It sounds complicated and so any attempt for even a partial mastery seems exhausting. And, indeed it is, but just as one needn't master solid state physics to successfully operate a TV, so too one can, by understanding some basic principles of data science, be able to think like an expert and so recognize claims that are made without evidence, and by doing so banish them from any place of influence. The core of data science is, in fact, science, and the scientific method with its emphasis on only what is observable and replicable provides its very soul.

This book is meant as a primer on thinking like a data scientist. It is a series of loosely related case studies in which the principles of data science are exemplified. There are only a few such principles illustrated, but it has been my experience that these few can carry you a long way.

Truthiness, although a new word, is a very old concept and has long predated science. It is so well inculcated in the human psyche that trying to banish it is surely a task of insuperable difficulty. The best we can hope for is to recognize that the core of truthiness's origins lies in the reptilian portion of our brains so that we can admit its influence yet still try to curb it through the practice of logical thinking.[2]

Escaping from the clutches of truthiness begins with one simple question. When a claim is made the first question that we ought to ask ourselves is "how can anyone know this?" And, if the answer isn't obvious, we must ask the person who made the claim, "what evidence do you have to support it?"

[2] It is beyond my immediate goals to discuss what sorts of evolutionary pressures must have existed to establish and preserve truthiness. For such an in-depth look there is no place better to begin than Nobel Laureate Danny Kahneman's inspired book *Thinking Fast, Thinking Slow*.

Let me offer four examples:

1. Speaking to your fetus in utero is important to the child's development.
2. Having your child repeat kindergarten would be a good idea.
3. Sex with uncircumcised men is a cause of cervical cancer in women.
4. There are about one thousand fish in that pond.

Ideas that lean on truthiness are sometimes referred to as "rapid ideas," for they only make sense if you say them fast.[3] Let us take a slower look at each of these claims in turn.

Claim 1: Talk to Your Fetus

Let us start with a plan to try to gather the kind of evidence necessary to make such a claim, and then try to imagine how close, in the real world, anyone could get to that ideal study. In order to know the extent of the effect any treatment has on a fetus, we have to compare what happens with that treatment with what would have happened had the fetus not had the treatment. In this situation we must compare the child's development after having regular conversations with its mother with how it would have developed had there been only silence. Obviously the same fetus cannot have both conditions. The problem of assessing the value of an action by comparing its outcome with that of a counterfactual isn't likely to have a solution. Instead we'll have to retreat to making such inferences based on averages within groups, in which we have one group of fetuses subject to the action of interest (being spoken to) and another group in which the alternative was tried (the comparison group). If the two groups were formed through a random process, it becomes plausible to believe that what was observed in the comparison (control) group would have been observed in the treatment group had that group had the control condition.

Next, what is the treatment? How much time is spent conversing with the fetus? What is being discussed? Is it OK to nag? Or instruct? Or is just cooing permissible? And what is the alternative condition? Is

[3] The tendency of stupid ideas to seem smarter when they come at you quickly is known in some quarters as the "Dopeler Effect."

it complete silence? Or just no talk directed solely at the fetus? Does the language matter? What about the correctness of syntax and grammar?

And finally, we need a dependent variable. What is meant by "the child's development"? Is it their final adult height? Or is it the speed with which they acquire language? Their general happiness? What? And how do we measure each child and so be able to make the comparison? And when? Is it at birth? At age one? Five? Twenty?

It seems sensible when confronted with claims like this to ask at least some of these questions. The answers will allow you to classify the claim as based on evidence or truthiness.

I have yet to hear anyone who makes such a claim provide any credible evidence.

Claim 2: Repeat Kindergarten

The same issues that arise in assessing the evidentiary support for the efficacy of fetal conversations repeat themselves here. How would the child do if not held back? What are the dependent variables that reflect the success of the intervention? Could there ever have been an experiment in which some randomly chosen children were held back and others not? And if this unlikely scenario had actually been followed, how was success judged? If it were on something trivial like height or age in first grade, children held back would be taller and older than those who progressed normally, but that isn't what we care about. We want to know whether the children would be happier if their progress is delayed. Are they reading better in sixth grade than they would have had they not been held back?

It isn't hard to construct a credible theory to support repeating a grade – if a child can't add integers, it makes little sense to move them forward into a class where such skill is assumed, but such decisions are rarely so cut and dried. It is more likely a quantitative decision: "Is this child's skill too low to be able to manage at the next level?" This is knowable, but it requires gathering of evidence. We might display the results of such a study as a graph in which a child's math score in kindergarten is plotted on the horizontal axis and her math score in grade one on the vertical axis. This tells us the relation between performances in the two grades, but it does not tell us about the efficacy of repeating kindergarten. For that we need to know the counterfactual event of what the child's score would have been had she repeated kindergarten. We would need to know how scores compared the first time taking the test with the second time, that is, how she did in first grade after repeating and how she would have done in first grade had she not repeated.

Again, it is possible to construct such an experiment, based on average group performance and random assignment, but the likelihood that any such experiment has ever been performed is small.

Try to imagine the response to your asking about what sort of evidence was used to support a teacher's recommendation that your child should repeat kindergarten. The response would be overflowing with truthiness and rich with phrases like "in my experience" or "I deeply feel."

Claim 3: Male Circumcision as a Cause of Cervical Cancer

This example was brought to my attention by a student in STAT 112 at the University of Pennsylvania. Each student was asked to find a claim in the popular media and design a study that would produce the necessary evidence to support that claim. Then they were to try to guess what data were actually gathered and judge how close those were to what would be required for a credible conclusion.

The student recognized that the decision to have a baby boy circumcised was likely related to social variables that might have some connection with cervical cancer. To eliminate this possibility, she felt that a sensible experiment that controlled for an unseen connection would need to randomly assign boys to be circumcised or not. She also recognized that women's choice of sex partner might have some unintended connection and so suggested that the matings between men and women should also be done at random. Once such a design was carried out, there would be nothing more to do than to keep track of all of the women in the study for thirty or forty years and count up the frequency of cervical cancer on the basis of the circumcision status of their sex partner. Of course, they would need to keep the same partner for all that time, or we would not have an unambiguous connection to the treatment.

Last, she noted that in the United States about twelve thousand women a year are diagnosed with cervical cancer (out of about 155 million women), or about one case for each thirteen thousand women. So the study would probably need at least half a million women in each of the two experimental groups to allow it to have enough power to detect what is likely to be a modest effect.

Once she had prepared this list of desiderata, she realized that such an experiment was almost certainly never done. Instead, she guessed that someone asked a bunch of women with cervical cancer about the status of their companions and found an overabundance of uncircumcised men. This led her to conclude that the possibilities of alternative explanations were sufficiently numerous and likely to allow her to dismiss the claim.

Is there nothing between a full-randomized experiment and a naïve data collection? In situations where the full experiment is too difficult to perform, there are a number of alternatives, like a case-control study that could provide some of the credibility of a full-randomized experiment, with a vastly more practical format.

Modern science is a complex edifice built on techniques that may not be obvious or even understandable to a layperson. How are we to know that the claims being made are not using credible methods of which we are unaware? I will return to this shortly after illustrating it in the next example.

Claim 4: Counting Fish in a Pond

"There are about one thousand fish in that pond." How could anyone know that? Did they put a huge net across the pond, capture all the fish, and count them? That sounds unlikely. And so, we may doubt the accuracy of the estimate. But perhaps some scientific methods allow such an estimate. Though it is important to maintain a healthy skepticism it is sensible to ask the person making the claim of one thousand fish what supporting evidence she might have. Had we done so, she might have responded, "We used the method of 'capture-recapture.'" Such jargon requires clarification. And so she expands, "Last week we came here and caught 100 fish, tagged them, and threw them back. We allowed a week to pass so that the tagged fish could mix in with the others and then we

Cambridge University Press
978-1-107-13057-9 - Truth or Truthiness: Distinguishing Fact from Fiction by Learning to Think Like a Data Scientist
Howard Wainer
Excerpt
More information

returned and caught another 100 fish and found that 10 of them were tagged. The calculation is simple, 10% of the fish we caught were tagged, and we know that in total, 100 were tagged. Therefore there must be about 1,000 fish in the pond."

The use of capture-recapture procedures can be traced back at least to 1783, when the famous French polymath Pierre-Simon Laplace used it to estimate the population of France.[4] This approach is widely used for many purposes; one is to estimate the number of illegal aliens in the United States.

Coda

The lesson to be learned from these four examples is that skepticism is important, but we must keep an open mind to the possibilities of modern data science. The more we know about it, the better we can design *gedanken* experiments that could yield the evidence that would support the claims made. If we can't imagine one that could work, or if whatever we imagine is unlikely to be practical, we should keep our skepticism, but ask for an explanation, based on science not anecdotes, from the person making the claim. The credibility of the methodology is what tells us how likely the claim is to be on the truthiness side of the boggle threshold.

This book has three parts:

I. **How to think like a data scientist** has, as its centerpiece, a beautiful theory of causality that is used to describe some methods of thinking about claims. In each situation, I illustrate the approach with a real-life claim and its supporting evidence. The questions examined range widely from the causes of happiness; the relation between virtuosos in both music and track; how much has fracking in Oklahoma affected the frequency of earthquakes in that state; and even how to evaluate experimental evidence the collection of which has been censored by death.

II. **How data scientists communicate to themselves and others.** I begin with some theory about the importance of empathy and

[4] Amoros 2014.

effective communication, and then narrow the focus to the communication of quantitative phenomena. The topics include communicating the genetic risks of cancer, the media's use of statistical methods, and the mapping of moral statistics.

III. **The application of these tools of thinking and communicating to the field of education.** Among the topics explored are the surprising trends in student performance over the past few decades, the point of teacher tenure in public schools, and what might have motivated the College Board in 2014 to institute three changes to the SAT.

In each section of this book a series of case studies describe some of the deep ideas of modern data science and how they can be used to help us defeat deception. The world of ideas is often divided into two camps: the practical and the theoretical. Fifty years of experience have convinced me that nothing is so practical as a good theory. The problems associated with making causal inferences lie at the very core of all aspects of our attempts to understand the world we live in, and so there is really no other way to begin than with a discussion of causal inference. This discussion focuses on a very good theory indeed, one that has come to be called "Rubin's Model for Causal Inference" after the Harvard statistician Donald Rubin, who first laid it out forty years ago.

Chapters 1 and 2 provide a brief warm-up, so that, in Chapter 3, we can turn our attention to the rudiments of Rubin's Model and show how it can be used to clarify a vexing chicken-and-egg question. It does this by guiding us to the structure of an experiment, the results of which can settle the issue. In Chapter 4 I expand the applicability of Rubin's Model and show how it casts light into dark corners of scientific inquiry in ways that are surprising. In Chapter 5, we continue on this same tack, by using the fundamental ideas of Rubin's Model to help us design experiments that can answer questions that appear, literally, beyond the reach of empirical solution. After this, the story ebbs and flows, but always with conviction borne of facts. I strive to avoid the passionate intensity that always seems to accompany evidence-starved truthiness.