Chapter One

The Seven Types of Data Pitfalls

"You need to give yourself permission to be human."
—*Joyce Brothers*

Data pitfalls. Anyone who has worked with data has fallen into them many, many times. I certainly have. It's as if we've used data to pave the way for a better future, but the road we've made is filled with craters we just don't seem to notice until we're at the bottom looking up. Sometimes we fall into them and don't even know it. Finding out about it much later can be quite humbling.

If you've worked with data before, you know the feeling. You're giving an important presentation, your data is insightful beyond belief, your charts and graphs are impeccable and Tufte-compliant, the build to your grand conclusion is unassailable and awe-inspiring. And then that one guy in the back of the room – the guy with folded arms and furrowed brow – waits until the very end to ask you if you're aware that the database you're working with is fundamentally

flawed, pulling the rug right out from underneath you, and plunging you to the bottom of yet another data pitfall. It's enough to make a poor data geek sweat bullets.

The nature of data pitfalls is that we have a particular blindness to them. It makes sense if you think about it. The human race hasn't needed to work with billions of records of data in the form of zeros and ones until the second half of the last century. Just a couple of decades later, though, our era is characterized by an ever-increasing abundance of data and a growing array of incredibly powerful tools. In many ways, our brains just haven't quite caught up yet.

These data pitfalls don't doom our every endeavor, though. Far from it. We've accomplished great things in this new era of data. We've mapped the human genome and begun to understand the complexity of the human brain, how its neurons interact so as to stimulate cognition. We've charted vast galaxies *out there* and we've come to a better understanding of geological and meteorological patterns *right here* on our own planet. Even in the simpler endeavors of life like holiday shopping, recommendation engines on e-commerce sites have evolved to be incredibly helpful. Our successes with data are too numerous to list.

But our slipups with data are mounting as well. Misuse of data has led to great harm and loss. From the colossal failure of Wall Street quants and their models in the financial crisis of the previous decade to the parable of Google Flu Trends and its lesson in data-induced hubris,[1] our use of data isn't always so successful. In fact, sometimes it's downright disastrous.

Why is that? Simply because we have a tendency to make certain kinds of mistakes time and time again. Noticing those mistakes early in the process is quite easy – just as long as it's someone else who's making them. When I'm the one committing the blunder, it seems I don't find out until that guy in the back of the room launches his zinger.

---

[1] http://gking.harvard.edu/files/gking/files/0314policyforumff.pdf.

And like our good friend and colleague, we're all quite adept at spotting the screw-ups of other people, aren't we? I had an early lesson in this haphazard trade. In my seventh-grade science fair exhibition, a small group of budding student scientists had a chance to walk around with the judges and explain our respective science fair projects while the other would-be blue-ribbon winners listened along. The judges, wanting to encourage dialogue and inquisitiveness, encouraged the students to also ask questions after each presentation. In spite of the noble intention behind this prompting, we basically just used the opportunity to poke holes in the methods and analysis of our competition. Kids can be cruel.

I don't do science fair projects anymore, unlike many other parents at my sons' schools, but I do work with data a lot. And I work with others who work with data a lot, too. In all of my data wrangling, data remixing, data analyzing, data visualizing, and data surmising, I've noticed that there are specific types of pitfalls that exist on the road to data paradise.

In fact, in my experience, I've found that the pitfalls we fall into can be grouped into one of seven categories.

Seven Types of Data Pitfalls

Pitfall 1: Epistemic Errors: How We Think About Data

What can data tell us? Maybe even more importantly, what *can't* it tell us? Epistemology is the field of philosophy that deals with the theory of knowledge – what's a reasonable belief versus what is just opinion. We often approach data with the wrong mind-set and assumptions, leading to errors all along the way, regardless of what chart type we choose, such as:

- Assuming that the data we are using is a perfect reflection of reality
- Forming conclusions about the future based on historical data only
- Seeking to use data to verify a previously held belief rather than to test it to see whether it's actually false

Avoiding epistemic errors and making sure we are thinking clearly about what's reasonable and what's unreasonable is an important foundation for successful data analysis.

Pitfall 2: Technical Traps: How We Process Data

Once we've decided to use data to help solve a particular problem, we have to gather it, store it, join it with other data sets, transform it, clean it up, and get it in the right shape. Doing so can result in:

- Dirty data with mismatching category levels and data entry typos
- Units of measurement or date fields that aren't consistent or compatible
- Bringing together disparate data sets and getting nulls or duplicated rows that skew analysis

These steps can be complex and messy, but accurate analysis depends on doing them right. Sometimes the truth contained within data gets "lost in translation," and it's possible to plow ahead and make decisions without even knowing we're dealing with a seriously flawed data set.

Pitfall 3: Mathematical Miscues: How We Calculate Data

Working with data almost always involves calculations – doing math with the quantitative data we have at our disposal:

- Summing at various levels of aggregation
- Calculating rates or ratios
- Working with proportions and percentages
- Dealing with different units

These are just a few examples of how we take data fields that exist and create new data fields out of them. Just like in grade school, it's very possible to get the math wrong. These mistakes can be quite

costly – an error of this type led to the loss of a $125 million Mars orbiter in 1999.[2] That was more like falling into a black hole than a pitfall.

Pitfall 4: Statistical Slipups: How We Compare Data

"There are lies, damned lies, and statistics." This saying usually implies that someone is fudging the numbers to mislead others, but we can just as often be lying to ourselves when it comes to statistics. Whether we're talking about descriptive or inferential statistics, the pitfalls abound:

- Are the measures of central tendency or variation that we're using leading us astray?
- Are the samples we're working with representative of the population we wish to study?
- Are the means of comparison we're using valid and statistically sound?

These pitfalls are numerous and particularly hard to spot on the horizon, because they deal with a way of thinking that even experts can get wrong sometimes. "Simple random samples" can be anything but simple to get right, and just ask a data guru to explain what a "p-value" means in layman's terms sometime.

Pitfall 5: Analytical Aberrations: How We Analyze Data

Analysis is at the heart of every data working endeavor. It's the means by which we draw conclusions and make decisions. There are many people who have "analyst" in their job title, but in truth, data analysis is a task that virtually everyone performs at one point or another. Data analysis has reached new heights, but we can also sink to new lows, like:

[2] http://www.cnn.com/TECH/space/9909/30/mars.metric.02/.

- Over-fitting our models to the historical data
- Missing important signals in the data
- Extrapolating or interpolating in ways that don't make sense
- Using metrics that don't really matter at all

Was it really reasonable to assume search trends could let us accurately predict the number of people who will get the flu, even while search algorithms are constantly changing and the searching population reacts to inputs like media hype and search engine recommendations?

Pitfall 6: Graphical Gaffes: How We Visualize Data

These are the mistakes that are most commonly noticed and talked about. Why? Because they're the visual ones. They're there for all to see and gaze upon in horror. You know the ones I'm talking about: dizzying pie charts with dozens of slices, misleading bar charts with y-axes that start at half of the maximum value. Luckily, these pitfalls are well documented, and can be identified by asking a handful of questions:

- Did we choose a sufficiently suitable chart type for the task at hand?
- If a point is being made, is it shown clearly, or do we have to strain to see it?
- Are we making use of rules of thumb without being unduly limited by them?

Sure, getting the chart type perfectly right is useless if we've fallen into one of the first five pitfalls above, but what a shame it is when we successfully execute on the whole routine up until this point only to botch the landing.

Pitfall 7: Design Dangers: How We Dress up Data

As humans, we really appreciate good design. We drive to work in well-designed automobiles, with all of the controls in the right place, and sit at our desk in ergonomic chairs that conform gracefully to

the contours of our bodies. Why would we want to sit there and open our browser to look at some garish infographic or clunky data dashboard? Design matters.

- Do our color choices confuse our audience or do they make things clearer to them?
- Have we used our creativity to judiciously embellish charts, or have we missed out on a great opportunity to include aesthetic components that add value?
- Are the visual objects we have created easy to interact with, or do they befuddle the user?

Getting these design elements right can actually mean the difference between our audience paying close attention to our message and totally ignoring us and paying attention to something else instead.

These seven pitfalls are like seven deadly sins – any one of them can make or break our data-working endeavor. But there's no sense in fearing them. We'd like to learn how to recover quickly when we find ourselves at the bottom of one, or, even better, learn to avoid them altogether. How do we do that?

Avoiding the Seven Pitfalls

When we come across a pitfall on a particular path in the real world, we'd like to think that there is a nice, helpful sign pointing it out to us and warning us of the danger, like the one on the Coal Creek Falls trail near my home in Bellevue, Washington (Figure 1.1).

But with data pitfalls, such helpful warning signs don't typically exist, do they? It's up to us to know these cognitive, procedural, and communicative pitfalls well and to know why it's so easy to fall into one. Awareness and mindfulness are the keys. If we aren't familiar with these nasty traps – what they look like, how to spot them, their telltale signs – then we're much more likely to fall into them. That much is obvious.

FIGURE 1.1 An ominous warning sign of a pitfall on the path to Coal Creek Falls in Bellevue, Washington.

But merely knowing about them often isn't enough. Even the sagest of data experts falls into these well-hidden traps from time to time. We need some helpful tips and trusty guides to help us along the way.

Starting in the following chapter, "Epistemic Errors," we'll begin collecting practical tips that will help us avoid each of the seven pitfalls so that we can remain on the straight and narrow data highway. By the end of our discussion of the seventh pitfall in Chapter 8, "Design Dangers," we'll have a full checklist that we can use to serve as a kind of trail map for our journey.

"I've Fallen and I Can't Get Up"

The fact is, though, we don't often have time to run through a comprehensive checklist before forging ahead on our data journey. The pressing demands of business and the fast-paced environments in

which we operate routinely present us with shortened deadlines and the need to produce insights from data in less time than we really need.

In these cases, we may have no choice but to press ahead, but at least we can use the "Avoiding Data Pitfalls Checklist" that we will present in the final chapter as a postmortem tool to identify our particular propensities, and to find out which pitfalls we find ourselves falling into time and again.

And it's going to happen. I promise that you will fall into one or more of these pitfalls in the very near future. So will your colleagues. So will I. I probably fell into more than one of them in this book itself. As a species, we're still learning how to change the way we think to suit this relatively new medium.

On an evolutionary scale, interacting with large spreadsheets and databases is not just new, it's brand new. Anatomically modern humans first appear in the fossil record around 195,000 years ago in Africa, and pioneering computer scientist Alan Turing set out the idea of the modern computer in his seminal 1936 paper, roughly 80 years ago.[3] That means we've been acclimating to the computing era for a grand total of 0.04% of human history. That's the fraction of a day that occurs in the last 35 seconds, between 11:59:25 p.m. and 12:00:00 a.m.

Okay, then it's going to happen. So how do we react when it does? We should see these mistakes as an unavoidable step in the process of developing a keen sense of navigation.

Do you remember learning about bloodletting, the ill-conceived practice of withdrawing blood from a patient to treat illness and disease? In classrooms around the world, the youth of our era scoff at the folly of this barbaric practice as they are taught about it year after year. But it was a common medical technique for 2,000 years, from antiquity until the late nineteenth century.

---

[3] https://www.cs.virginia.edu/~robins/Turing_Paper_1936.pdf

Just like our forebears, our generation makes many boneheaded mistakes on a routine basis that future generations will find baffling. It's my hope that falling into data pitfalls will be among those human propensities that our progeny find inexplicable in future generations.

So what happens when we find ourselves in the bottom of a nasty data pitfall? What do we do then? Our inclination is to pretend it never happened, cover up the mistake, and hope no one finds out. This is the opposite of what we should do:

- First, try to get out: fix your mistake.
- Second, put a notch on your checklist next to the pitfall into which you fell.
- Third, tell everyone about what happened.

This process, as self-flagellating as it sounds, will help us all grow the muscle of effective data working. To ensure others can follow this process as well, we'll need to refrain from vilifying those who fall into data pitfalls. Remember, it not only could have been you, it *will* be you sometime down the road.