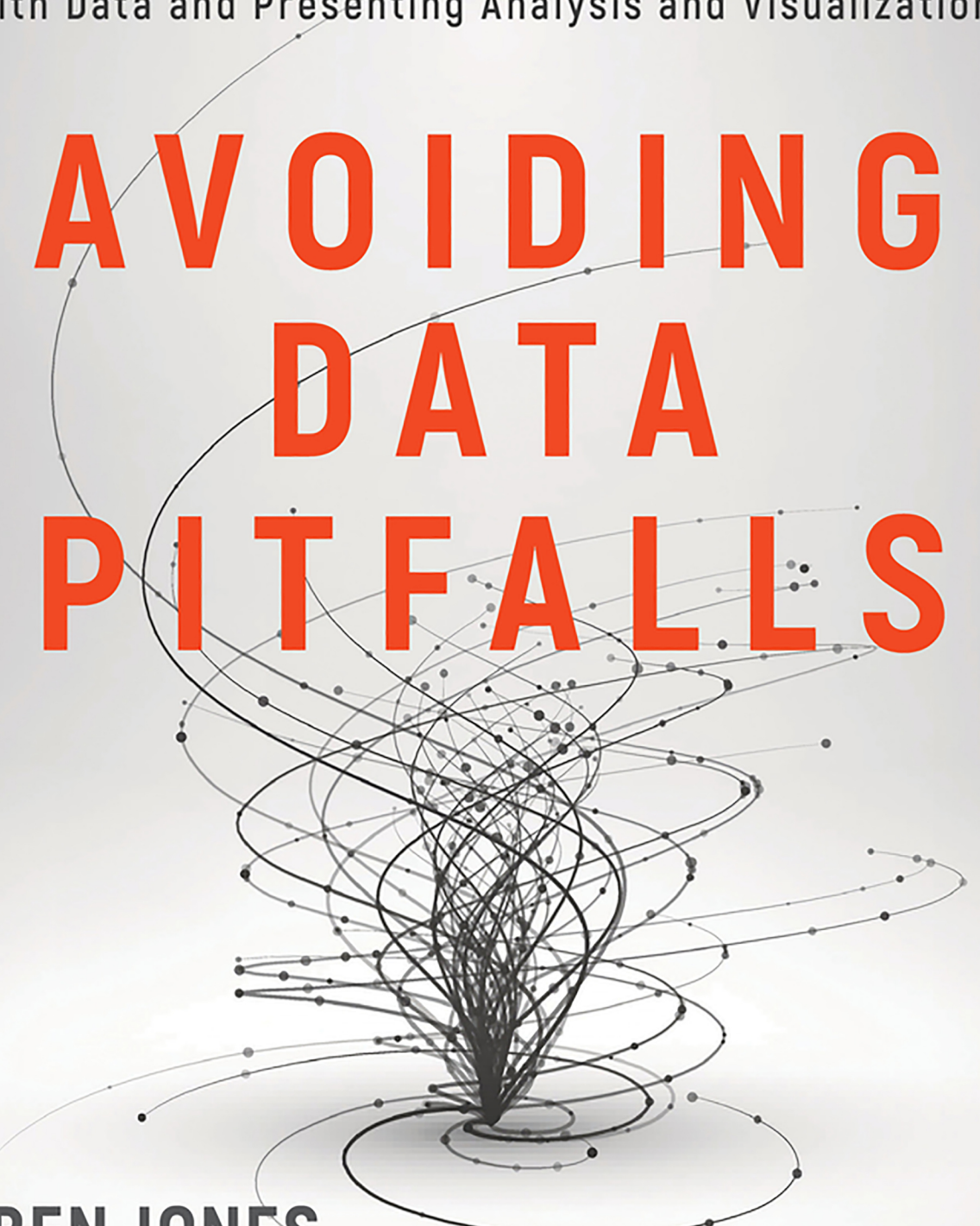


How to Steer Clear of Common Blunders When Working  
with Data and Presenting Analysis and Visualizations

# AVOIDING DATA PITFALLS



**BEN JONES**

Founder and CEO, Data Literacy

**WILEY**



# Avoiding Data Pitfalls



# Avoiding Data Pitfalls

How to Steer Clear of Common Blunders  
When Working with Data and Presenting  
Analysis and Visualizations

Ben Jones

WILEY

Copyright © 2020 Ben Jones. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.

Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 646-8600, or on the Web at [www.copyright.com](http://www.copyright.com). Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at [www.wiley.com/go/permissions](http://www.wiley.com/go/permissions).

**Limit of Liability/Disclaimer of Warranty:** While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993, or fax (317) 572-4002.

Wiley publishes in a variety of print and electronic formats and by print-on-demand. Some material included with standard print versions of this book may not be included in e-books or in print-on-demand. If this book refers to media such as a CD or DVD that is not included in the version you purchased, you may download this material at <http://booksupport.wiley.com>. For more information about Wiley products, visit [www.wiley.com](http://www.wiley.com).

### ***Library of Congress Cataloging-in-Publication Data***

Names: Jones, Ben, 1978– author.

Title: Avoiding data pitfalls : how to steer clear of common blunders when working with data and presenting analysis and visualizations / Ben Jones.

Description: Hoboken, New Jersey : John Wiley & Sons, Inc., [2020] | Includes index.

Identifiers: LCCN 2019033232 (print) | LCCN 2019033233 (ebook) | ISBN 9781119278160 (paperback) | ISBN 9781119278191 (adobe pdf) | ISBN 9781119278177 (epub)

Subjects: LCSH: Information visualization. | Quantitative research.

Classification: LCC QA76.9.I52 J6639 2020 (print) | LCC QA76.9.I52 (ebook) | DDC 001.4/226—dc23

LC record available at <https://lcn.loc.gov/2019033232>

LC ebook record available at <https://lcn.loc.gov/2019033233>

Cover Design: Wiley

Cover Image: © deomis/iStock.com

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

# Contents

|  |           |
|--|-----------|
| Preface  | ix        |
| Chapter 1 <b>The Seven Types of Data Pitfalls</b>      | <b>1</b>  |
| Seven Types of Data Pitfalls                           | 3         |
| Pitfall 1: Epistemic Errors: How We Think About Data   | 3         |
| Pitfall 2: Technical Traps: How We Process Data        | 4         |
| Pitfall 3: Mathematical Miscues: How We Calculate Data | 4         |
| Pitfall 4: Statistical Slipups: How We Compare Data    | 5         |
| Pitfall 5: Analytical Aberrations: How We Analyze Data | 5         |
| Pitfall 6: Graphical Gaffes: How We Visualize Data     | 6         |
| Pitfall 7: Design Dangers: How We Dress up Data        | 6         |
| Avoiding the Seven Pitfalls                            | 7         |
| “I’ve Fallen and I Can’t Get Up”                       | 8         |
| Chapter 2 <b>Pitfall 1: Epistemic Errors</b>           | <b>11</b> |
| How We Think About Data                                | 11        |
| Pitfall 1A: The Data-Reality Gap                       | 12        |
| Pitfall 1B: All Too Human Data                         | 24        |
| Pitfall 1C: Inconsistent Ratings                       | 32        |
| Pitfall 1D: The Black Swan Pitfall                     | 39        |
| Pitfall 1E: Falsifiability and the God Pitfall         | 43        |
| Avoiding the Swan Pitfall and the God Pitfall          | 44        |

|           |  |            |
|-----------|--|------------|
| Chapter 3 | <b>Pitfall 2: Technical Trespasses</b>             | <b>47</b>  |
|           | How We Process Data                                | 47         |
|           | Pitfall 2A: The Dirty Data Pitfall                 | 48         |
|           | Pitfall 2B: Bad Blends and Joins                   | 67         |
| Chapter 4 | <b>Pitfall 3: Mathematical Miscues</b>             | <b>74</b>  |
|           | How We Calculate Data                              | 74         |
|           | Pitfall 3A: Aggravating Aggregations               | 75         |
|           | Pitfall 3B: Missing Values                         | 83         |
|           | Pitfall 3C: Tripping on Totals                     | 88         |
|           | Pitfall 3D: Preposterous Percents                  | 93         |
|           | Pitfall 3E: Unmatching Units                       | 102        |
| Chapter 5 | <b>Pitfall 4: Statistical Slipups</b>              | <b>107</b> |
|           | How We Compare Data                                | 107        |
|           | Pitfall 4A: Descriptive Debacles                   | 109        |
|           | Pitfall 4B: Inferential Infernos                   | 131        |
|           | Pitfall 4C: Slippery Sampling                      | 136        |
|           | Pitfall 4D: Insensitivity to Sample Size           | 142        |
| Chapter 6 | <b>Pitfall 5: Analytical Aberrations</b>           | <b>148</b> |
|           | How We Analyze Data                                | 148        |
|           | Pitfall 5A: The Intuition/Analysis False Dichotomy | 149        |
|           | Pitfall 5B: Exuberant Extrapolations               | 157        |
|           | Pitfall 5C: Ill-Advised Interpolations             | 163        |
|           | Pitfall 5D: Funky Forecasts                        | 166        |
|           | Pitfall 5E: Moronic Measures                       | 168        |
| Chapter 7 | <b>Pitfall 6: Graphical Gaffes</b>                 | <b>173</b> |
|           | How We Visualize Data                              | 173        |
|           | Pitfall 6A: Challenging Charts                     | 175        |
|           | Pitfall 6B: Data Dogmatism                         | 202        |
|           | Pitfall 6C: The Optimize/Satisfice False Dichotomy | 207        |



|           |                                   |     |
|-----------|-----------------------------------|-----|
| Chapter 8 | Pitfall 7: Design Dangers         | 212 |
|           | How We Dress up Data              | 212 |
|           | Pitfall 7A: Confusing Colors      | 214 |
|           | Pitfall 7B: Omitted Opportunities | 222 |
|           | Pitfall 7C: Usability Uh-Ohs      | 227 |
| Chapter 9 | Conclusion                        | 237 |
|           | Avoiding Data Pitfalls Checklist  | 241 |
|           | The Pitfall of the Unheard Voice  | 243 |
|           | Index                             | 247 |



# Preface

There's a passage in *The Dhammapada*, which is a collection of sayings of the Buddha in verse form, which has this to say:

*If you see someone wise, who can steer you away from the wrong path, follow that person, as you would one who can reveal hidden treasures. Only good can come out of it. Let them admonish, or instruct, or restrain you from what is wrong.*

— The Dhammapada, Verse 76

Most ancient wisdom texts have a similar exhortation to find a wise person and follow their advice. If you listen to their words carefully, you'll avoid horrible mistakes and the associated pain and discomfort that these mistakes will cause in your life. We all need a mentor, guide, or guru from time to time.

Unfortunately, I may not be that wise person you're looking for. I'd say I'm more like a person who has taken his bumps and bruises while searching for such a wise individual myself. So I'm more like the person in the words to a favorite song of mine:

*No, I don't claim to be a wise man, a poet or a saint. I'm just another man who's searching for a better way.*

— Jon Bon Jovi, "Bang a Drum"

I figured it just might be helpful, however, to capture and convey the types of mistakes I've made, and seen others make, many times over the course of my data working career. I've worked with data in manufacturing and transactional environments, on the shop floor, in the boardroom, and in newsrooms all over the country, and I've trained and taught people how to work with data in corporate conference rooms, virtual chat rooms, and in the halls of academia, from coast to coast.

But who, exactly, am I trying to help with this book? Well, me, for one. Every time I write a blog post, record a tutorial, or craft a presentation, I find myself going back to these materials over and over. Whenever I do, I stop and wonder how the past version of myself was so much smarter than the current version of me, and I tip my hat for the favor.

And I hope it will be helpful for you. If you're starting out in your data journey, you'll come across many of these pitfalls, I promise you. My hope is that you'll be more likely to recognize them for what they are when you see them – sometimes from the top and other times from the bottom.

And if you've been working with data for some time, you'll read a section here or there, and you'll nod knowingly, glancing down at a scar or two that you earned by falling headfirst into the pit with me. And your brow may furrow when you read about other pitfalls, a sinking feeling coming over you that you may have made that mistake without recognizing it. If so, know that I feel your pain.

It's really important, though, that we learn to pick ourselves up and dust off our jeans, clean off those scuff marks, ice any bruises we may have suffered, and carry on, a bit wiser for the experience.

Equal in importance is that we show others the same grace. It's just too easy to get it wrong, and it's bound to happen. Even experts fall into data pitfalls on a fairly regular basis. Just like a well-worn path that marks the best route across a terrain, the more we admit and talk about our shortcomings, the less likely others are to fall into the same trap we fell into. We'll be leaving warning signs for them.

You and I may have to swallow our pride when we raise our hand and share the mea culpa, but we can think of it as a gift for future generations. They may shake their heads and wonder how we messed up so badly, but you and I will know that the only reason they've evolved as far as they have is that we paved the way for them with our blunders.

What's more important than the protection of my ego or yours is the evolution of our species into an efficient and effective data working force for the planet. We're not there, yet. Not even close.

I'd like to dedicate this book to my father, Richard Jones. My father had a truly wonderful mind, and, like the rest of us, he fell into many pitfalls in his life. When I confronted him about some things he had done that were particularly hurtful to me, he admitted it, and he apologized. I'll never forget that gift he gave me. It was very liberating.

I've been much better at admitting my own failures ever since – personal and professional. Thank you, Dad. I love you.

I signed the contract to write this book right before his health took a turn for the worse. He passed away of glioblastoma half a year later, and I lost all sight of the reason for wanting to write this book in the first place. My editors and the team at John Wiley & Sons were very gracious to me during this process. It took a while – four years to be exact – but eventually I rediscovered my passion for this topic and finished the book.

On the positive side, all that extra time meant that I was able to discover even more pitfalls to add to these chapters, mostly by falling into them.

So I hope this book proves to be helpful for you. May you stride forward onto the road paved with data, alert to pitfalls around every turn. May you show yourself compassion whenever you get it wrong, and may you share the lessons of mistakes made with others around you. Ultimately, may you reach great heights on the

pathway, uncovering new treasures, solving pressing problems, and unlocking growth you never thought you'd find.

And if you bump into a truly wise data guru, will you give me their number? I'd appreciate it – I have a lot more to learn from them.

## 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,<sup>1</sup> 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.

---

<sup>1</sup><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.<sup>2</sup> 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:

---

<sup>2</sup><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.<sup>3</sup> 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.

---

<sup>3</sup>[https://www.cs.virginia.edu/~robins/Turing\\_Paper\\_1936.pdf](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.



## Chapter Two

# Pitfall 1: Epistemic Errors

“Everybody gets so much information all day long that they lose their common sense.”

—Gertrude Stein

### How We Think About Data

Epistemology is the branch of philosophy that deals with the nature, origin, and scope of our knowledge. It comes from the Greek words *episteme* (knowledge) and *logos* (word/speech) – knowledge speech, or, in other words, talking about knowledge.

Let’s talk about knowledge as it relates to working with data. But first, why spend any time on this at all? A fair question, since clearly practitioners of many disciplines ignore the fundamental principles underlying their area of focus. Take driving an automobile, for example. Most drivers can’t explain how the internal combustion

engine works, or how the batteries in their electric vehicle function. But that doesn't stop them from driving around town, does it?

But working with data isn't like driving a car in this respect. It's more like cooking. In order to do it well, we need a working knowledge of the ways heat is transferred to food over time and how different flavors combine to produce a final result. We can't just start throwing ingredients together at random and hope to make a great dish. That's just one of the things I learned living with roommates in college, by the way.

But that's what happens when we start cooking things up with data before we have an understanding of the basic principles of knowledge. Epistemology is our data cookbook of sorts. Let's see what's in it.

### Pitfall 1A: The Data-Reality Gap

The first epistemic principle to embrace is that there is always a gap between our data and the real world. We fall headfirst into a pitfall when we forget that this gap exists, that our data isn't a perfect reflection of the real-world phenomena it's representing. Do people really fail to remember this? It sounds so basic. How could anyone fall into such an obvious trap?

I'm not exaggerating when I say that I fail to avoid this trap almost every single time. The first pitfall is a gaping hole and pretty much every one falls right into it at first.

It works like this: I get my hands on some data and I perform some analysis, but I don't stop to think about where the data came from, who collected it, what it tells me, and, importantly, what it doesn't tell me.

It's easy when working with data to treat it as reality itself rather than data collected about reality. Here are some examples:

- It's not crime, it's *reported* crime.
- It's not the outer diameter of a mechanical part, it's the *measured* outer diameter.

- It's not how the public feels about a topic, it's how *people who responded to the survey are willing to say they feel*.

You get the picture. This distinction may seem like a technicality, and sometimes it might be (the number of home runs Hank Aaron "reportedly" hit?), but it can also be a very big deal. Let's see some examples of how we can fall into this pitfall.

### *Example 1: All the Meteorites We Don't See*

The Meteoritical Society provides data for 34,513 meteorites that struck the surface of the earth between 2,500 BCE and 2012.<sup>1</sup> If you and I were to take this figure and run with it, we might make a number of incorrect assumptions that stem from the first data pitfall we're considering together.

Let's look more closely to get a better understanding of the impact of falling into this pitfall.

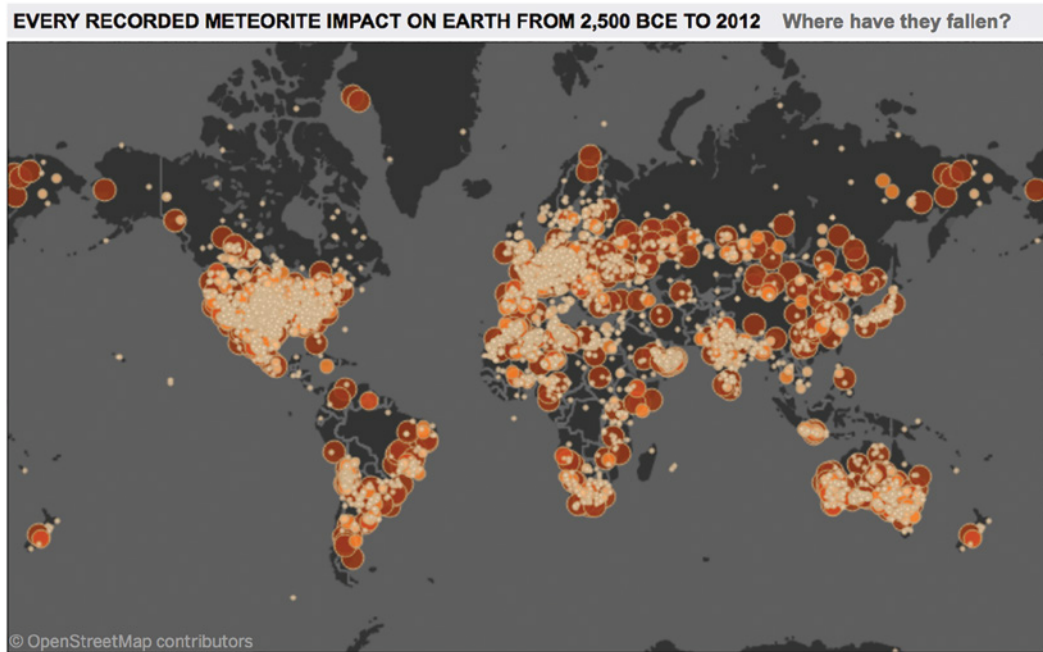
A friend of mine, Ramon Martinez, created a map (Figure 2.1) depicting where each of these 34,513 meteorites had struck the surface of the earth.

What do you notice about the data now that we're looking at it on a map? Doesn't it seem uncanny that meteorites are so much more likely to hit the surface of the earth where there's land, as opposed to where there's ocean? And what about areas like the Amazon (not the one in Seattle), or Greenland, or parts of Central Africa? Is there some kind of shield over those areas, or some deity protecting those areas from damage? We humans are great at coming up with bullshit theories like this, aren't we?

The explanation is obvious, and Ramon actually gives it to us right in the title at the top of the visualization: "Every Recorded Meteorite Impact." In order for a meteorite to be in the database, it had to be recorded. And in order for it to be recorded, it had to be observed

---

<sup>1</sup> <https://www.lpi.usra.edu/meteor/metbull.php>.



**FIGURE 2.1** Meteorite strikes by Ramon Martinez.

Source: Ramon Martinez, [https://public.tableau.com/profile/ramon.martinez#!/vizhome/meteorite\\_fall\\_on\\_earth/Meteoritefallonearthvisualized](https://public.tableau.com/profile/ramon.martinez#!/vizhome/meteorite_fall_on_earth/Meteoritefallonearthvisualized).

by someone. Not just anyone, but someone who knew whom to tell about it. And the person they told had to be faithful in carrying out their job of recording it. That's much more likely to occur in areas of higher population density of developed nations.

The map, then, isn't showing us where meteorites are more likely to strike the earth. It's telling us where meteorites are more likely to have fallen (in the past), and were observed by someone who reported it to someone who recorded it faithfully.

Now, that's a mouthful, isn't it? And you may roll your eyes and say it's all just a bunch of technicalities. But think again about the 34,513 figure. If we began with this figure, and if we assumed like I did at first that experts or enthusiasts observe and record every single meteorite strike no matter where it falls, then we'd have a pretty inaccurate idea of how often this kind of event actually occurs on the planet.