PRAISE FOR *BECOMING A DATA HEAD*

Big Data, Data Science, Machine Learning, Artificial Intelligence, Neural Networks, Deep Learning ... It can be buzzword bingo, but make no mistake, everything is becoming "datafied" and an understanding of data problems and the data science toolset is becoming a requirement for every business person. Alex and Jordan have put together a must read whether you are just starting your journey or already in the thick of it. They made this complex space simple by breaking down the "data process" into understandable patterns and using everyday examples and events over our history to make the concepts relatable.

—**Milen Mahadevan**, President of 84.51°

What I love about this book is its remarkable breadth of topics covered, while maintaining a healthy depth in the content presented for each topic. I believe in the pedagogical concept of "Talking the Walk," which means being able to explain the hard stuff in terms that broad audiences can grasp. Too many data science books are either too specialized in taking you down the deep paths of mathematics and coding ("Walking the Walk") or too shallow in over-hyping the content with a plethora of shallow buzzwords ("Talking the Talk"). You can take a great walk down the pathways of the data field in Alex and Jordan's without fear of falling off the path. The journey and destination are well worth the trip, and the talk.

—**Kirk Borne**, Data Scientist, Top Worldwide Influencer in Data Science

The most clear, concise, and practical characterization of working in corporate analytics that I've seen. If you want to be a killer analyst and ask the right questions, this is for you.

—**Kristen Kehrer**, Data Moves Me, LLC, LinkedIn Top Voices in Data Science & Analytics

THE book that business and technology leaders need to read to fully understand the potential, power, AND limitations of data science.

—**Jennifer L. L. Morgan**, PhD, Analytical Chemist at Procter and Gamble

You've heard it before: "We need to be doing more machine learning. Why aren't we doing more sophisticated data science work?" Data science isn't the magic unicorn that will solve all of your company's problems. *Data Head* brings this idea to life by highlighting when data science is (and isn't) the right approach and the common pitfalls to watch out for, explaining it all in a way that a data novice can understand. This book will be my new "pocket reference" when communicating complicated concepts to non-technically trained leaders.

—**Sandy Steiger**, Director, Center for Analytics and Data Science at Miami University

Individuals and organizations want to be data driven. They say they are data driven. *Becoming a Data Head* shows them how to actually become data driven, without the assumption of a statistics or data background. This book is for anyone,

Table of Contents

List of Tables

List of Illustrations

FIGURE 12.9 Deep learning is a subfield of machine learning, which is a subf...

So forget these false assumptions, and turn yourself into a Data Head. You'll become a more valuable employee and make your organization more successful. This is the way the world is going, so it's time to get with the program and learn more about data and analytics. I think you will find the process—and the reading of *Becoming a Data Head*—more rewarding and more pleasant than you may imagine.

Thomas H. Davenport
Distinguished Professor, Babson College
Visiting Professor, Oxford Saïd Business School
Research Fellow, MIT Initiative on the Digital Economy
Author of *Competing on Analytics, Big Data @ Work*, and *The AI Advantage*

NOTE

[1] Splunk Inc., "The State of Dark Data," 2019, www.splunk.com/en_us/form/thestate-of-dark-data.html .

Introduction

Data is perhaps the single most important aspect to your job, whether you want it to be or not. And you're likely reading this book because you want to be able to understand what it's all about.

To begin, it's worth stating what has almost become cliché: we create and consume more information than ever before. Without a doubt, we are in the age of data. And this age of data has created an entire industry of promises, buzzwords, and products many of which you, your managers, colleagues, and subordinates are or will be using. But, despite the claims and proliferation of data promises and products, data science projects are failing at alarming rates.[1]

To be sure, we're not saying all data promises are empty or all products are terrible. Rather, to truly get your head around this space, you must embrace a fundamental truth: this stuff is complex. Working with data is about numbers, nuance, and uncertainty. Data is important, yes, but it's rarely simple. And yet, there is an entire industry that would have us think otherwise. An industry that promises certainty in an uncertain world and plays on companies' fear of missing out. We, the authors, call this the Data Science Industrial Complex.

THE DATA SCIENCE INDUSTRIAL COMPLEX

It's a problem for everyone involved. Businesses endlessly pursue products that will do their thinking for them. Managers hire analytics professionals who really aren't. Data scientists are hired to work in companies that aren't ready for them. Executives are forced to listen to technobabble and pretend to understand. Projects stall. Money is wasted.

Meanwhile, the Data Science Industrial Complex is churning out new concepts faster than our ability to define and articulate the opportunities (and problems) they create. Blink, and you'll miss one. When your authors started working together, *Big Data* was all the rage. As time went on, *data science* became the hot new topic. Since then, *machine learning*, *deep learning*, and *artificial intelligence* have become the next focus.

To the curious and critical thinkers among us, something doesn't sit well. Are the problems really new? Or are these new definitions just rebranding old problems?

The answer, of course, is yes to both.

But the bigger question we hope you're asking yourself is, *How can I think and speak critically about data?*

Let us show you how.

By reading this book, you'll learn the tools, terms, and thinking necessary to navigate the Data Science Industrial Complex. You'll understand data and its challenges at a deeper level. You'll be able to think critically about the data and results you come across, and you'll be able to speak intelligently about all things data.

In short, you'll become a *Data Head*.

WHY WE CARE

Before we get into the details, it's worth discussing why your authors, Alex and Jordan, care so much about this topic. In this section, we share two important examples of how data affected society at large and impacted us personally.

The Subprime Mortgage Crises

We were fresh out of college when the subprime mortgage crisis hit. We both landed jobs in 2009 for the Air Force, at a time when jobs were hard to find. We were both lucky. We had an in-demand skill: working with data. We had our hands in data every single day, working to operationalize research from Air Force analysts and scientists into products the government could use. Our hiring would be a harbinger of the focus the country would soon place on the types of roles we filled. As two data workers, we looked on the mortgage crisis with interest and curiosity.

The subprime mortgage crises had a lot of contributing factors behind it.[2] In our attempt to offer it up as an example here, we don't want to negate other factors. However, put simply, we see it as a major data failure. Banks and investors created models to understand the value of mortgage-backed collateralized debt obligations (CDOs). You might remember those as the investment vehicles behind the United States' market collapse.

Mortgage-backed CDOs were thought to be a safe investment because they spread the risk associated with loan default across multiple investment units. The idea was that in a portfolio of mortgages, if only a few went into default, this would not materially affect the underlying value of the entire portfolio.

And yet, upon reflection we know that some fundamental underlying assumptions were wrong. Chief among them were that default outcomes were independent events. If Person A defaults on a loan, it wouldn't impact Person B's risk of default. We would all soon learn defaults functioned more like dominoes where a previous default could predict further defaults. When one mortgage defaulted, the property values surrounding the home dropped, and the risk of defaults on those homes increased. The default effectively dragged the neighboring houses down into a sinkhole.

Assuming independence when events are in fact connected is a common error in statistics.

But let's go further into this story. Investment banks created models that overvalued these investments. A model, which we'll talk about later in the book, is a deliberate oversimplification of reality. It uses assumptions about the real world in an attempt to understand and make predictions about certain phenomena.

And who were these people who created and understood these models? They were the people who would lay the groundwork for what today we call the data scientist. Our kind of people. Statisticians, economists, physicists—folks who did machine learning, artificial intelligence, and statistics. They worked with data. And they were smart. Super smart.

And yet, something went wrong. Did they not ask the correct questions of their work? Were disclosures of risk lost in a game of telephone from the analysts to the decision makers, with uncertainty being stripped away piece by piece, giving an illusion of a perfectly predictable housing market? Did the people involved flat out lie about results?

More personal to us, how could we avoid similar mistakes in our own work?

We had many questions and could only speculate the answers, but one thing was clear—this was a large-scale data disaster at work. And it wouldn't be the last.

The 2016 United States General Election

On November 8, 2016, the Republican candidate, Donald J. Trump, won the general election of the United States beating the assumed front-runner and Democratic challenger, Hillary Clinton. For the political pollsters this came as a shock. Their models hadn't predicted his win. And this was supposed to be the year for election prediction.

In 2008, Nate Silver's FiveThirtyEight blog—then part of *The New York Times*—had done a fantastic job predicting Barack Obama's win. At the time, pundits were skeptical that his forecasting algorithm could accurately predict the election. In 2012, once again, Nate Silver was front and center predicting another win for Barack Obama.

By this point, the business world was starting to embrace data and hire *data scientists*. The successful prediction by Nate Silver of Barack Obama's reelection only reinforced the importance and perhaps oracle-like abilities of forecasting with data. Articles in business magazines warned executives to adopt data or be swallowed by a data-driven competitor. The Data Science Industrial Complex was in full force.

By 2016, every major news outlet had invested in a prediction algorithm to forecast the general election outcome. The vast, vast majority of them by and large suggested an overwhelming victory for the Democratic candidate, Hillary Clinton. Oh, how wrong they were.

Let's contrast how wrong they were as we compare it against the subprime mortgage crisis. One could argue that we learned a lot from the past. That interest in data science would give rise to avoiding past mistakes. Yes, it's true: since 2008—and 2012—news organizations hired data scientists, invested in polling research, created data teams, and spent more money ensuring they received good data.

Which begs the question: with all that time, money, effort, and education—what happened?[3]

Our Hypothesis

Why do data problems like this occur? We assign three causes: hard problems, lack of critical thinking, and poor communication.

First (as we said earlier), this stuff is complex. Many data problems are fundamentally difficult. Even with lots of data, the right tools and techniques, and the smartest analysts, mistakes happen. Predictions can and will be wrong. This is

not a criticism of data and statistics. It's simply reality.

Second, some analysts and stakeholders stopped thinking critically about data problems. The Data Science Industrial Complex, in its hubris, painted a picture of certainty and simplicity, and a subset of people drank the proverbial "Kool-Aid." Perhaps it's human nature—people don't want to admit they don't know what is going to happen. But a key part of thinking about and using data correctly is recognizing wrong decisions can happen. This means communicating and understanding risks and uncertainties. Somehow this message got lost. While we'd hope the tremendous progress in research and methods in data and analysis would sharpen everyone's critical thinking, it caused some to turn it off.

The third reason we think data problems continue to occur is poor communication between data scientists and decision makers. Even with the best intentions, results are often lost in translation. Decision makers don't speak the language because no one bothered to teach data literacy. And, frankly, data workers don't always explain things well. There's a communication gap.

DATA IN THE WORKPLACE

Your data problems might not bring down the global economy or incorrectly predict the next president of the United States, but the context of these stories is important. If miscommunication, misunderstanding, and lapses in critical thinking occur when the world is watching, they're probably happening in your workplace. In most cases, these are micro failures subtly reinforcing a culture without data literacy.

We know it's happened in our workplace, and it was partly our fault.

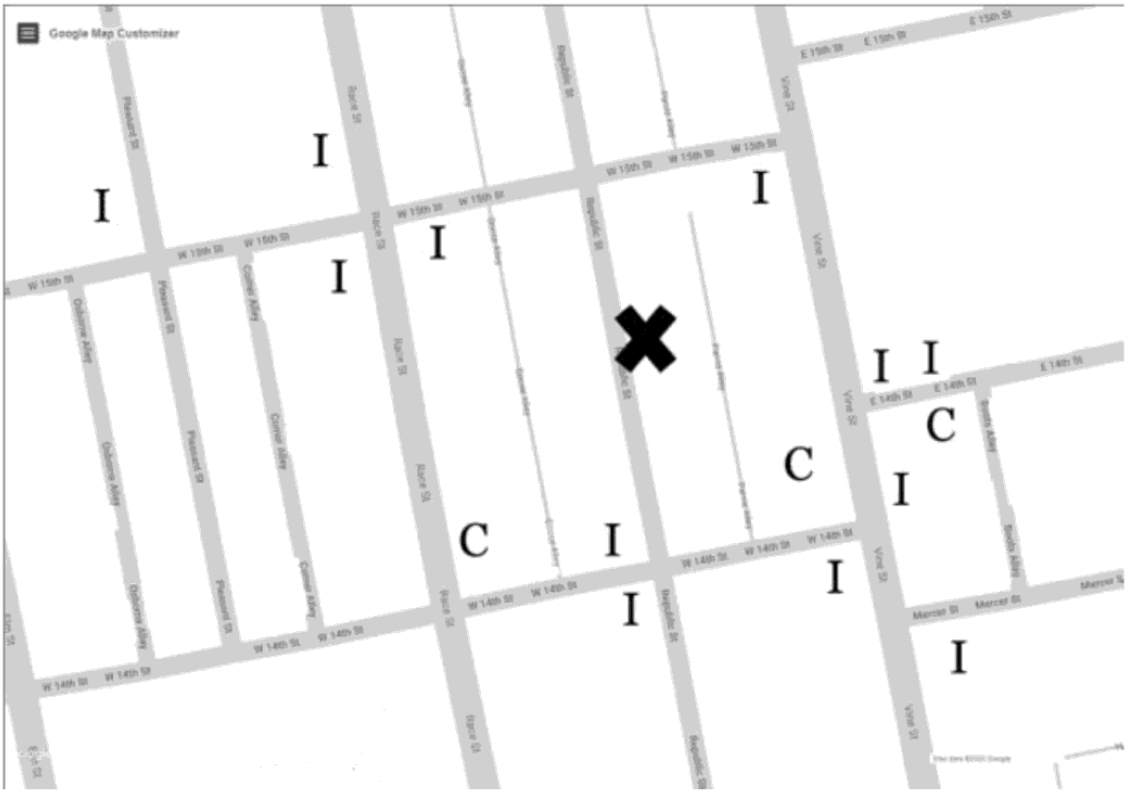
The Boardroom Scene

Fans of science fiction and adventure movies know this scene all too well: The hero is faced with a seemingly unsurmountable task and the world's leaders and scientists are brought together to discuss the situation. One scientist, the nerdiest among the group, proposes an idea dropping esoteric jargon before the general barks, "Speak English!" At this point, the viewer receives some exposition that explains what was meant. The idea of this plot point is to translate what is otherwise mission-critical information into something not just our hero—but the viewer—can understand.

We've discussed this movie trope often in our roles as researchers for the federal government. Why? Because it never seemed to unfold this way. In fact, what we saw early in our careers was often the opposite of this movie moment.

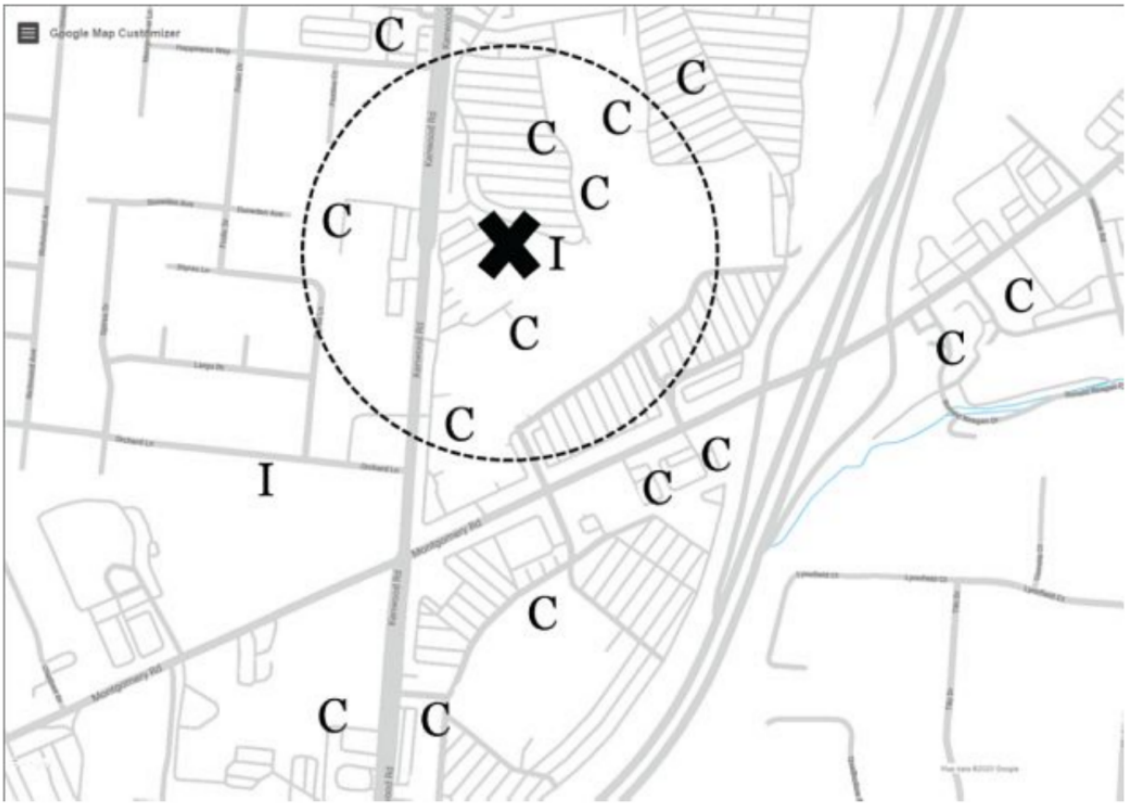
We presented our work to blank stares, listless head nodding, and occasional heavy eyelids. We watched as confused audiences seemed to receive what we were saying without question. They were either impressed by how smart we seemed or bored because they didn't get it. No one demanded we repeat what was said in a language everyone could understand. We saw something unfold that was dramatically different. It often unfolded like this:

Us: "Based on our supervised learning analysis of the binary response



Over the Rhine neighborhood, Cincinnati, Ohio

Next, look at the data in the following image. This area includes a large shopping mall, and most restaurants in the area are chains. When asked to predict chain or independent, the majority choose (C). But we love when someone chooses (I) because it highlights several important lessons.



Kenwood Towne Centre, Cincinnati, Ohio

During this thought experiment, everyone creates a slightly different *algorithm* in their head. Of course, everyone looks at the markers surrounding the point of interest, X, to understand the neighborhood, but at some point, you must decide when a restaurant is too far away to influence your prediction. At one extreme (and we see it happen), someone looks at the restaurant's single closest neighbor, in this case an independent restaurant, and bases their prediction on it: “The nearest neighbor to X is an (I), so my prediction is (I).”

Most people, however, look at several neighboring restaurants. The second image shows a circle surrounding the new restaurant containing its seven nearest neighbors. You probably chose a different number, but we chose 7, and 6 out of the 7 are (C) chains, so we'd predict (C).

So What?

If you understand the restaurant example, you're well on your way to becoming a Data Head. Let's reveal what you learned, little by little:

- You performed *classification* by predicting the *label* (chain or independent) on a new restaurant by *training* an algorithm using a set of data (restaurants' location and their chain/independent label).
- This is precisely *machine learning*! You just didn't build the algorithm on a computer—you used your head.
- Specifically, this is a type of machine learning called *supervised learning*. It was “supervised” because you knew the existing restaurants were (C) chain or (I) independent. The *labels* directed (i.e., supervised) your thinking about how