DARK
DATA

DARK DATA

WHY WHAT YOU DON'T KNOW MATTERS

DAVID J. HAND

PRINCETON UNIVERSITY PRESS
PRINCETON AND OXFORD

Published by Princeton University Press
41 William Street, Princeton, New Jersey 08540
6 Oxford Street, Woodstock, Oxfordshire OX20 1TR

press.princeton.edu

Library of Congress Cataloging-in-Publication Data

Names: Hand, D. J. (David J.), 1950–author.
Title: Dark data : why what you don't know matters / David J. Hand.
Description: Princeton : Princeton University Press, [2020] | Includes bibliographical references and index.
Identifiers: LCCN 2019022971 (print) | LCCN 2019022972 (ebook) | ISBN 9780691182377 (hardback) | ISBN 9780691198859 (ebook)
Subjects: LCSH: Missing observations (Statistics) | Big data.
Classification: LCC QA276 .H3178 2020 (print) | LCC QA276 (ebook) | DDC 519.5—dc23
LC record available at https://lccn.loc.gov/2019022971
LC ebook record available at https://lccn.loc.gov/2019022972

British Library Cataloging-in-Publication Data is available

Editorial: Ingrid Gnerlich and Arthur Werneck
Production Editorial: Karen Carter
Text Design: Leslie Flis
Jacket/Cover Design: Jason Alejandro
Production: Jacquie Poirier
Publicity: Sara Henning-Stout and Katie Lewis
Copyeditor: Barbara Liguori

This book has been composed in Arno

Printed on acid-free paper. ∞

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CONTENTS

Preface xi

PART 1: DARK DATA: THEIR ORIGINS AND CONSEQUENCES

CHAPTER 1: **Dark Data: What We Don't See Shapes Our World** 3

The Ghost of Data 3

So You Think You Have All the Data? 12

Nothing Happened, So We Ignored It 17

The Power of Dark Data 22

All around Us 24

CHAPTER 2: **Discovering Dark Data: What We Collect and What We Don't** 28

Dark Data on All Sides 28

Data Exhaust, Selection, and Self-Selection 31

From the Few to the Many 43

Experimental Data 56

Beware Human Frailties 67

CHAPTER 3: **Definitions and Dark Data: What Do You Want to Know?** 72

Different Definitions and Measuring the Wrong Thing 72

You Can't Measure Everything 80
Screening 90
Selection on the Basis of Past Performance 94

CHAPTER 4: **Unintentional Dark Data: Saying One Thing, Doing Another** 98
The Big Picture 98
Summarizing 102
Human Error 103
Instrument Limitations 108
Linking Data Sets 111

CHAPTER 5: **Strategic Dark Data: Gaming, Feedback, and Information Asymmetry** 114
Gaming 114
Feedback 122
Information Asymmetry 128
Adverse Selection and Algorithms 130

CHAPTER 6: **Intentional Dark Data: Fraud and Deception** 140
Fraud 140
Identity Theft and Internet Fraud 144
Personal Financial Fraud 149
Financial Market Fraud and Insider Trading 153
Insurance Fraud 158
And More 163

CHAPTER 7: **Science and Dark Data: The Nature of Discovery** 167

The Nature of Science 167

If Only I'd Known That 172

Tripping over Dark Data 181

Dark Data and the Big Picture 184

Hiding the Facts 199

Retraction 215

Provenance and Trustworthiness: Who Told You That? 217

PART II:
ILLUMINATING AND USING DARK DATA

CHAPTER 8: **Dealing with Dark Data: Shining a Light** 223

Hope! 223

Linking Observed and Missing Data 224

Identifying the Missing Data Mechanism 233

Working with the Data We Have 236

Going Beyond the Data: What If You Die First? 241

Going Beyond the Data: Imputation 245

Iteration 252

Wrong Number! 256

CHAPTER 9: **Benefiting from Dark Data: Reframing the Question** 262

Hiding Data 262

Hiding Data from Ourselves: Randomized Controlled Trials 263

What Might Have Been 265

Replicated Data 269

Imaginary Data: The Bayesian Prior 276

Privacy and Confidentiality Preservation 278

Collecting Data in the Dark 287

CHAPTER 10: **Classifying Dark Data: A Route through the Maze** 291

A Taxonomy of Dark Data 291

Illumination 298

Notes 307

Index 319

PREFACE

This book is unusual. Most books about data—be they popular books about big data, open data, or data science, or technical statistical books about how to analyze data—are about the data you have. They are about the data sitting in folders on your computer, in files on your desk, or as records in your notebook. In contrast, this book is about data you *don't* have—perhaps data you wish you had, or hoped to have, or thought you had, but nonetheless data you don't have. I argue, and illustrate with many examples, that the missing data are at least as important as the data you do have. The data you cannot see have the potential to mislead you, sometimes even with catastrophic consequences, as we shall see. I show how and why this can happen. But I also show how it can be avoided—what you should look for to sidestep such disasters. And then, perhaps surprisingly, once we have seen how dark data arise and can cause such problems, I show how you can use the dark data perspective to flip the conventional way of looking at data analysis on its head: how hiding data can, if you are clever enough, lead to deeper understanding, better decisions, and better choice of actions.

The question of whether the word *data* should be treated as singular or plural has been a fraught one. In the past it was typically treated as plural, but language evolves, and many people now treat it as singular. In this book I have tried to treat "data" as plural except in those instances where to do so sounded ugly to my ears. Since beauty is said to be in the eye of the beholder, it is entirely possible that my perception may not match yours.

My own understanding of dark data grew slowly throughout my career, and I owe a huge debt of gratitude to the many people who brought me challenges which I slowly realized were dark data problems and who worked with me on developing ways to cope with them. These problems ranged over medical research, the pharmaceutical industry, government and social policy, the financial sector, manufacturing, and other domains. No area is free from the risks of dark data.

Particular people who kindly sacrificed their time to read drafts of the book include Christoforos Anagnostopoulos, Neil Channon, Niall Adams, and three anonymous publisher's readers. They prevented me from making too many embarrassing mistakes. Peter Tallack, my agent, has been hugely supportive in helping me find the ideal publisher for this work, as well as graciously advising me and steering the emphasis and direction of the book. My editor at Princeton University Press, Ingrid Gnerlich, has been a wise and valuable guide in helping me beat my draft into shape. Finally, I am especially grateful to my wife, Professor Shelley Channon, for her thoughtful critique of multiple drafts. The book is significantly improved because of her input.

Imperial College, London

Chapter 1

DARK DATA

●

What We Don't See Shapes Our World

The Ghost of Data

First, a joke.

Walking along the road the other day, I came across an elderly man putting small heaps of powder at intervals of about 50 feet down the center of the road. I asked him what he was doing. "It's elephant powder," he said. "They can't stand it, so it keeps them away."

"But there are no elephants here," I said.

"Exactly!" he replied. "It's wonderfully effective."

Now, on to something much more serious.

Measles kills nearly a 100,000 people each year. One in 500 people who get the disease die from complications, and others suffer permanent hearing loss or brain damage. Fortunately, it's rare in the United States; for example, only 99 cases were reported in 1999. But a measles outbreak led Washington to declare a statewide emergency in January 2019, and other states also reported dramatically increased numbers of cases.[1] A similar pattern was reported elsewhere. In Ukraine, an outbreak resulted in over 21,000 cases by mid-February 2019.[2] In Europe there were 25,863 cases in 2017, but in 2018 there were over 82,000.[3] From

1 January 2016 through the end of March 2017, Romania reported more than 4,000 cases and 18 deaths from measles.

Measles is a particularly pernicious disease, spreading undetected because the symptoms do not become apparent until some weeks after you contract it. It slips under the radar, and you have it before you even know that it's around.

But the disease is also preventable. A simple vaccination can immunize you against the risk of contracting measles. And, indeed, national immunization programs of the kind carried out in the United States have been immensely successful—so successful in fact that most parents in countries which carry out such programs have never seen or experienced the terrible consequences of such preventable diseases.

So, when parents are advised to vaccinate their children against a disease they have neither seen nor heard of any of their friends or neighbors having, a disease which the Centers for Disease Control and Prevention announced was no longer endemic in the United States, they naturally take the advice with a pinch of salt.

Vaccinate against something which is not there? It's like using the elephant powder.

Except that, unlike the elephants, the risks are still there, just as real as ever. It's merely that the information and data these parents need to make decisions are missing, so that the risks have become invisible.

My general term for the various kinds of missing data is *dark data*. Dark data are concealed from us, and that very fact means we are at risk of misunderstanding, of drawing incorrect conclusions, and of making poor decisions. In short, our ignorance means we get things wrong.

The term "dark data" arises by analogy with the dark matter of physics. About 27 percent of the universe consists of this mysterious substance, which doesn't interact with light or other electromagnetic radiation and so can't be seen. Since dark matter can't be seen, astronomers were long unaware of its existence. But then observations of the rotations of galaxies revealed that the more distant stars were not moving more slowly than stars nearer the center, contradicting what we would have expected from our understanding of gravity. This rotational anomaly can be explained by supposing that galaxies have more mass than appears to be the case judging from the stars and other objects we can see through our telescopes. Since we can't see this extra mass, it has been called dark matter. And it can be significant (I almost said "it can matter"): our home galaxy, the Milky Way, is estimated to have some ten times as much dark matter as ordinary matter.

Dark data and dark matter behave in an analogous way: we don't see such data, they have not been recorded, and yet they can have a major effect on our conclusions, decisions, and actions. And as some of the later examples will show, unless we are aware of the possibility that there's something unknown lurking out there, the consequences can be disastrous, even fatal.

The aim of this book is to explore just how and why dark data arise. We shall look at the different kinds of dark data and see what leads to them. We shall see what steps we can take to avoid dark data's arising in the first place. We shall see what we can do when we realize that dark data are obscured from us. Ultimately, we shall also see that if we are clever enough, we can sometimes take advantage of dark data. Curious and paradoxical though that may seem, we can make use of ignorance and the dark data perspective to enable better decisions and take better actions. In practical terms, this means we can lead healthier lives, make more money, and take lower risks by judicious use of the unknown. This doesn't mean we should hide information from others (though, as we shall also see, deliberately concealed data is one common kind of dark data). It is much more subtle than that, and it means that everyone can benefit.

Dark data arise in many different shapes and forms as well as for many different reasons, and this book introduces a taxonomy of such reasons, the *types* of dark data, labeled *DD-Type x,* for "Dark Data-Type x." There are 15 *DD-Types* in all. My taxonomy is not exhaustive. Given the wealth of reasons for dark data, that would probably be impossible. Moreover, any particular example of dark data might well illustrate the effect of more than one *DD-Type* simultaneously—*DD-Types* can work together and can even combine in an unfortunate synergy. Nonetheless, an awareness of these *DD-Types,* and examination of examples showing how dark data can manifest, can equip you to identify when problems occur and protect you against their dangers. I list the *DD-Types* at the end of this chapter, ordered roughly according to similarity, and describe them in more detail in chapter 10. Throughout the book I have indicated some of the places when an example of a particular *Type* occurs. However, I have deliberately not tried to do this in an exhaustive way—that would be rather intrusive.

To get us going, let's take a new example.

In medicine, trauma is serious injury with possible major long-term consequences. It's one of the most serious causes of "life years lost" through premature death and disability, and is the commonest cause of death for those under age 40. The database of the Trauma Audit and Research Network (TARN) is the largest medical trauma database in Europe. It receives data on trauma events from more than 200 hospitals, including over 93 percent of the hospitals in England and Wales, as well as hospitals in Ireland, the Netherlands, and Switzerland. It's clearly a very rich seam of data for studying prognoses and the effectiveness of interventions in trauma cases.

Dr. Evgeny Mirkes and his colleagues from the University of Leicester in the UK looked at some of the data from this database.[4] Among the 165,559 trauma cases they examined, they found 19,289 with unknown outcomes. "Outcome" in trauma research means whether or not the patient survives at least 30 days after the injury. So the 30-day survival was unknown for over 11 percent of the patients. This example illustrates a common form of dark data—our *DD-Type 1: Data We Know Are Missing.* We know these patients had some outcome—we just don't know what it was.

No problem, you might think—let's just analyze the 146,270 patients for whom we do know the outcome and base our understanding and prognoses on those. After all, 146,270 is a big number—within the realm of medicine it's "big data"—so surely we can be confident that any conclusions based on these data will be right.

But can we? Perhaps the missing 19,289 cases are very different from the others. After all, they were certainly different in that they had unknown outcomes, so it wouldn't be unreasonable to suspect they might differ in other ways. Consequently, any analysis of the 146,270 patients with known outcomes might be misleading relative to the overall population of trauma patients. Thus, actions taken on the basis of such analysis might be the wrong actions, perhaps leading to mistaken prognoses, incorrect prescriptions, and inappropriate treatment regimes, with unfortunate, even fatal, consequences for patients.

To take a deliberately unrealistic and extreme illustration, suppose that all 146,270 of those with known outcomes survived and recovered without treatment, but the 19,289 with unknown jolt of a car being driven over a pothole and then used GPS to automatically transmit the location of the hole to the city authorities.

Wonderful! Now the highway maintenance people would know exactly where to go to repair the potholes.

Again, this looks like an elegant and cheap solution to a real problem, built on modern data analytic technology—except for the fact that ownership of cars and expensive smartphones is more likely to be concentrated in wealthier areas. Thus, it's quite likely that potholes in poorer areas would not be detected, so that their location would not be transmitted, and some areas might never have their potholes fixed. Rather than solving the pothole problem in general, this approach might even aggravate social inequalities. The situation here is different from that in the TARN example, in which we knew that certain data were missing. Here we are unaware of them.

The following is another illustration of this kind of dark data. In late October 2012, Hurricane Sandy, also called "Superstorm Sandy,"[5] struck the Eastern Seaboard of the United States. At the time it was the second most costly hurricane in U.S. history and the largest Atlantic hurricane on record, causing damage estimated at $75 billion, and killing more than 200 people in eight countries. Sandy affected 24 U.S. states, from Florida to Maine to Michigan to Wisconsin, and led to the closure of the financial markets owing to power cuts. And it resulted, indirectly, in a surge in the birth rate some nine months later.

It was also a triumph of modern media. The physical storm Hurricane Sandy was accompanied by a Twitter storm of messages describing what was going on. The point about Twitter is that it tells you what and where something is happening as it happens, as well as who it's happening to. The social media platform is a way to keep up in real time as events unfold. And that's exactly what occurred with Hurricane Sandy. Between 27 October and 1 November 2012, there were more than 20 million tweets about it. Clearly, then, we might think, this is ideal material from which to get a continuously evolving picture of the storm as it develops, identifying which areas have been most seriously affected, and where emergency relief is needed.

But later analysis revealed that the largest number of tweets about Sandy came from Manhattan, with few tweets coming from areas like Rockaway and Coney Island. Did that mean that Rockaway and Coney Island were less severely affected? Now it's true that subways and streets of Manhattan were flooded, but it was hardly the worst-hit region, even of New York. The truth is, of course, that those regions transmitting fewer tweets may have been doing so not because the storm had less impact but simply because there were fewer Twitter users with fewer smartphones to tweet them.

In fact, we can again imagine an extreme of this situation. Had any community been completely obliterated by Sandy, then no tweets at all would have emerged. The superficial impression would be that everybody there was fine. Dark data indeed.

As with the first type of dark data, examples of this second kind, in which we don't know that something is missing, are ubiquitous. Think of undetected fraud, or the failure of a crime-victim survey to identify that any murders have been committed.

You might have a sense of déjà vu about those first two types of dark data. In a famous news briefing, former U.S. Secretary of Defense Donald Rumsfeld nicely characterized them in a punchy sound bite, saying "there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns—the ones we don't know we don't know."[6] Rumsfeld attracted considerable media ridicule for that convoluted statement, but the criticism was unfair. What he said made very good sense and was certainly true.

But those first two types are just the beginning. In the next section we introduce some of the other types of dark data. These, and others described later, are what this book is all about. As you will see, dark data have many forms. Unless we are aware that data might be incomplete, that observing something does not mean observing everything, that a measurement procedure might be inaccurate, and that what is measured might not really be what we want to measure, then we could get a very misleading impression of what's going on. Just because there's no one around to hear that tree fall in the forest doesn't mean that it didn't make a noise.

So You Think You Have All the Data?

The customer arrives at the supermarket checkout with a full shopping cart. The laser scans the barcode of each item, and the till emits its electronic beep as it adds up the total cost. At the end of this exercise, the customer is presented with the overall bill and pays. Except that's not really the end. The data describing the items bought and the price of each are sent to a database and stored. Later, statisticians and data scientists will pore over the data, extracting a picture of customer behavior from details of what items were bought, which items were bought together, and indeed what sort of customer bought the items. Surely there's no opportunity for missing data here? Data of the transaction have to be captured if the supermarket is to work out how much to charge the customer—short of a power cut, register failure, or fraud, that is.

Now it seems pretty obvious that the data collected are all the data there are. It's not just *some* of the transactions or details of just *some* of the items purchased. It's *all* the transactions made by *all* the customers on *all* the items in that supermarket. It is, as is sometimes simply said, "data = all."

But is it really? After all, these data describe what happened *last* week or *last* month. That's useful, but if we are running the supermarket, what we probably really want to know is what will happen tomorrow or next week or next month. We really want to know who will buy what when, and how much of it they will buy in the future. What's likely to run out if we don't put more on the shelves? What brands will people prefer to buy? We really want data that have not been measured. Dark data *DD-Type 7: Changes with Time* describes the obscuring nature of time on data.

Indeed, beyond that complication, we might want to know how people *would have behaved* had we stocked different items, or arranged them differently on the shelves, or changed the supermarket opening times. These are called *counterfactuals* because they are contrary to fact—they are about what would have happened if what actually happened hadn't. Counterfactuals are dark data *DD-Type 6: Data Which Might Have Been*.

Needless to say, counterfactuals are of concern not just to supermarket managers. You've taken medicines in the past. You trusted the doctor who prescribed them, and you assumed they'd been tested and found to be effective in alleviating a condition. But how would you feel if you discovered that they hadn't been tested? That no data had been collected on whether the medicines made things better? Indeed, that it was possible they made things worse? Or that even if they had been tested and found to help, the medicines hadn't been compared with simply leaving the condition alone, to see if they made it get better more quickly than natural healing processes? Or the medicines hadn't been compared with other ones, to see if they were more effective than familiar alternatives? In the elephant powder example, a comparison with doing nothing would soon reveal that *doing nothing was just as effective at keeping the elephants away* as putting down the heaps of powder. (And that, in turn could lead to the observation that there were actually no elephants to be kept away.)

Returning to the notion of "data=all," in other contexts the notion that we might have "all" the data is *clearly* nonsensical. Consider your weight. This is easy enough to measure—just hop on your bathroom scale. But if you repeat the measurement, even very soon afterward, you might find a slightly different result, especially if you try to measure it to the nearest ounce or gram. All physical measurements are subject to potential inaccuracies as a result of measurement error or random fluctuations arising from very slight changes in the circumstances (*DD-Type 10: Measurement Error and Uncertainty*). To get around this problem, scientists measuring the magnitude of some phenomenon—the speed of light, say, or the electric charge of the electron—will take multiple measurements and average them. They might take 10 measurements, or 100. But what they obviously cannot do is take "all" the measurements. There is no such thing as "all" in this context.

A different type of dark data is illustrated when you ride on London's red buses: you will know that more often than not they are packed with passengers. And yet data show that the occupancy of the average bus is just 17 people. What can explain this apparent contradiction? Is someone manipulating the figures?

A little thought reveals that the answer is simply that more people are riding on the buses when they are full—that's what "full" means. The consequence is that more people see a full bus. At the opposite extreme, an empty bus will have no one to report that it was empty. (I'm ignoring the driver in all this, of

Nothing Happened, So We Ignored It

A final example illustrates that dark data can have disastrous consequences and that they are not especially a problem of large data sets.

Thirty years ago, on 28 January 1986, 73 seconds into its flight and at an altitude of 9 miles, the space shuttle *Challenger* experienced an enormous fireball caused by one of its two booster rockets and broke up. The crew compartment continued its trajectory, reaching an altitude of 12 miles, before falling into the Atlantic. All seven crew members, consisting of five astronauts and two payload specialists, were killed.

A later presidential commission found that NASA middle managers had violated safety rules requiring data to be passed up the chain of command. This was attributed to economic pressures, making it very important that the launch schedule should be maintained: the launch date had already slipped from January 22nd to the 23rd, then to the 25th, and then to the 26th. Since temperature forecasts for that day suggested an unacceptably low temperature, the launch was again rescheduled, for the 27th. Countdown proceeded normally until indicators suggested a hatch lock had not closed properly. By the time that was fixed the wind was too strong, and again the launch was postponed.

On the night of January 27th, a three-hour teleconference was held between Morton Thiokol, which was the company that made the booster rockets, NASA staff at the Marshall Space Flight Center, and people from the Kennedy Space Center. Larry Wear, of the Marshall Center, asked Morton Thiokol to check the possible impact of low temperatures on the solid rocket motors. In response, the Morton Thiokol team pointed out that the O-rings would harden in low temperatures.

The O-rings were rubber-like seals, with a cross-section diameter of about a quarter of an inch, which fitted in the joint around the circumference between each of the four rocket motor segments. The solid rocket boosters were 149 feet high and 38 feet in circumference. Under launch conditions, the 0.004 inch gap that the O-rings normally sealed typically opened to a maximum of 0.06 inch: just six one-hundredths of an inch. And during launch this larger gap remained open for just six-tenths of a second.

Robert Ebeling of Morton Thiokol had been concerned that at low temperatures the hardening of the O-rings meant they would lose their ability to create an effective seal between segments when the gaps expanded by that 0.056 inch for that 0.6 second. At the teleconference Robert Lund, vice president of Morton Thiokol, said that the O-ring operating temperature must not be less than the previous lowest launch temperature, 53°F. Extensive, sometimes heated, discussion ensued, both in the conference and off-line in private conversations. Eventually, Morton Thiokol reconsidered and recommended launch.

Precisely 58.79 seconds after the launch a flame burst from the right solid rocket motor near the last joint. This flame quickly grew into a jet which broke the struts joining the solid rocket motor to the external fuel tank. The motor pivoted, hitting first the Orbiter's wing and then the external fuel tank. The jet of flame then fell onto this external tank containing the liquid hydrogen and oxygen fuel. At 64.66 seconds the tank's surface was breached, and 9 seconds later *Challenger* was engulfed in a ball of flame and broke into several large sections.[8]

One thing we have to remember is that space travel is all about risk. No mission, even under the very best of circumstances, is a risk-free enterprise: the risk cannot be reduced to zero. And there are always competing demands.

Furthermore, as with any incident like this, the notion of "cause" is complicated. Was it due to violation of safety rules, undue pressure put on managers because of economic considerations, other consequences of budget tightening, or perhaps media pressure following the fact that the launch of the previous shuttle, *Columbia*, had been delayed seven times, each delay greeted with press ridicule? For example, here's Dan Rather's script for the evening news on Monday, January 27th, following the four delays to the *Challenger* launch: "Yet another costly, red-faces-all-around space-shuttle-launch delay. This time a bad bolt on a hatch and a bad-weather bolt from the blue are being blamed." Or was it a consequence of political pressure. After all, there was significantly more interest in this launch than earlier launches because it carried an "ordinary person," Christa McAuliffe, a teacher, and the president's State of the Union address was scheduled for the evening of January 28th.

In such situations, multiple factors typically come together. Complex and obscure interactions can lead to unexpected consequences. But in this case there was another factor: dark data.

After the disaster, a commission headed by former secretary of state William Rogers drew attention to the fact that flights which had not had any O-rings showing distress had not been included in the diagram discussed at the teleconference (dark data *DD-Type 3: Choosing Just Some Cases* but also *DD-Type 2: Data We Don't Know Are Missing*). The report said (p. 146): "The managers compared as a function of temperature the flights for which thermal distress of O-rings had been observed—not the frequency of occurrence based on all flights."[9] And that's the giveaway: *data from some flights were not included in the analysis*. My earlier examples have shown the sorts of problems leaving out some of the data can lead to.

The report went on: "In such a comparison [that is, using the limited set of data presented], there is nothing irregular in the distribution of O-ring 'distress' over the spectrum of joint temperatures at launch between 53 degrees Fahrenheit and 75 degrees Fahrenheit, " meaning: there is no apparent relationship between temperature and number of O-rings showing distress. However, "when the entire history of flight experience is considered, including 'normal' flights with no erosion or blow-by, the comparison is substantially different"; that is, if you include all the data, you get a different picture. In fact, flights which took place at higher temperatures were much more likely to show no problems, and these were the dark data not shown in the plot. But if the higher the temperature, the less the chance of a problem, then, conversely, the lower the temperature, the greater the chance of a problem. And the ambient temperature was predicted to be just 31°F.

This section of the report concluded: "Consideration of the entire launch temperature history indicates that the probability of O-ring distress is increased to *almost a certainty* if the temperature of the joint is less than 65[°F]." (my italics)

The situation is graphically illustrated in the two diagrams in Figure 1. Figure 1(a) shows the diagram discussed at the teleconference. This is a plot of the number of distressed O-rings on each launch plotted against launch temperature in degrees Fahrenheit. So, for example, at the lowest launch temperature in the past, 53°F, three of the O-rings experienced distress, and at the highest launch temperature in the past, 75°F, two of the O-rings experienced distress. There is no clear relationship between launch temperature and the number of distressed O-rings.

However, if we add the missing data—showing the launches which led to no O-ring distress, we obtain Figure 1(b). The pattern is now very clear. In fact, *all* the launches which occurred

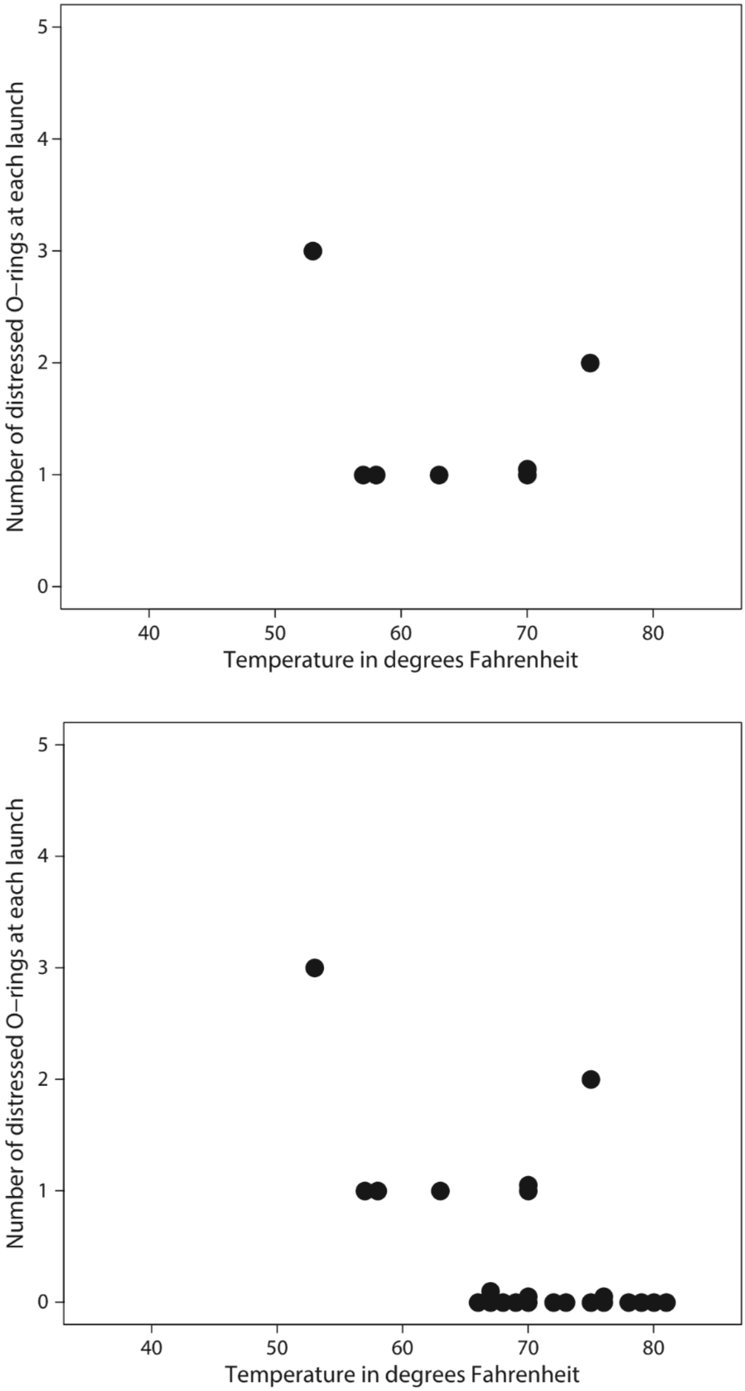

FIGURE 1. (a) Data examined in the *Challenger* prelaunch teleconference; (b) complete data.

All around Us

We've seen that dark data are ubiquitous. They can arise anywhere and everywhere, and one of their most dangerous aspects is that, by definition, we may not know that they are *not* there. It means we have to be constantly on the alert, asking ourselves, *what are we missing?*

Are we failing to notice large amounts of fraud because the police catch the inept criminals while the really good ones escape unnoticed? Bernie Madoff established his firm Bernard L. Madoff Investment Securities LLC in 1960 but wasn't arrested until 2008, and sentenced (to 150 years in prison) in 2009, when he was already 71—he almost got away with it.

Are we not noticing many potentially curable sick people simply because the more severe cases are obvious, but the less severe don't show so many symptoms?

Are the social networks established by modern social media dangerous simply because they reflect what we already know and believe, not challenging us because they don't show us facts or events outside our comfort zone?

Perhaps worse still, the descriptions people choose to post on social media may give us a false impression of how wonderful everyone else's life is, casting us into depression because in contrast our lives have so many obstacles.

We tend to think of data as numerical. But data don't have to be just numbers. And that means that dark data also don't have to be numerical. The following is an example in which the crucial missing information is a single letter.

The Arctic expeditions of 1852, 1857, and 1875 were stocked with a supply of Allsopp's Arctic Ale, an ale with an especially low freezing point prepared by brewer Samuel Allsopp. Alfred Barnard sampled the beer in 1889, describing it as "of a nice brown colour, and of a vinous, and at the same time, nutty flavor, and as sound as on the day it was brewed. . . . Owing to the large amount of unfermented extract still remaining in it, it must be considered as an extremely valuable and nourishing food."[10] Just the sort of thing you need to sustain you on Arctic expeditions.

In 2007 a bottle of the 1852 batch came up for sale on eBay, with a reserve price of $299. Or at least that was the aim. In fact the vendor, who had had the bottle for 50 years, misspelled the beer's name—he missed one of the *p*'s in Allsopp. As a consequence, the item didn't show up in the searches carried out by most vintage beer enthusiasts, so that there were just two bids. The winning bid, for $304, was from 25-year-old Daniel P. Woodul. Aiming to appraise the value of the bottle, Woodul immediately relisted it on eBay, but this time with the correct spelling. This time there were 157 bids, with the winning one being for $503,300.

That missing *p* clearly mattered, to the tune of some half a million dollars.* This shows that missing information can have significant consequences. In fact, as we shall see, a mere half-million-dollar loss is nothing compared with the losses that other missing data situations have led to. Indeed, missing data can wreck lives, destroy companies, and (as with the *Challenger* disaster) can even lead to death. In short, missing data matter.

In the case of Allsopp's Arctic Ale, a little care would have avoided the problem. But while carelessness is certainly a common cause of dark data, there are many others. The painful fact

*In fact it turned out that the winning bid was a practical joke, and the bidder had no intention of paying. But Woodul is nevertheless doubtless still sitting on a tidy profit: a private collector from Scotland recently auctioned a bottle from the 1875 expedition for £3,300 (~$4,300).

is that data can be dark for a tremendously wide variety of reasons, as we shall see in this book.

It is tempting to regard dark data as simply synonymous with data which could have been observed but which for some reason were not. That is certainly the most obvious kind of dark data. The missing salary levels in a survey in which some people refused to divulge how much they were paid are certainly dark data, but so also are the salary levels for those who do not work and hence do not have a salary level to divulge. Measurement error obscures true values, data summaries (such as averages) hide the details, and incorrect definitions misrepresent what you want to know. More generally still, any unknown characteristic of a population can be thought of as dark data (statisticians often refer to such characteristics as *parameters*).

Since the number of possible causes of dark data is essentially unlimited, knowing what *sort* of thing to keep an eye open for can be immensely useful in helping avoid mistakes and missteps. And that is the function of the *DD-Types* described in this book. These are not basic causes (like failure to include the final outcome for patients who have been in a study for only a short time) but provide a more general taxonomy (like the distinction between data we know are missing and data we don't know are missing). An awareness of these *DD-Types* can help in protecting against mistakes, errors, and disasters arising from ignorance about what you do not know. The *DD-Types*, which are introduced in this book, and which are summarized in chapter 10, are as follows:

DD-Type 1: Data We Know Are Missing
DD-Type 2: Data We Don't Know Are Missing
DD-Type 3: Choosing Just Some Cases
DD-Type 4: Self-Selection

DD-Type 5: Missing What Matters
DD-Type 6: Data Which Might Have Been
DD-Type 7: Changes with Time
DD-Type 8: Definitions of Data
DD-Type 9: Summaries of Data
DD-Type 10: Measurement Error and Uncertainty
DD-Type 11: Feedback and Gaming
DD-Type 12: Information Asymmetry
DD-Type 13: Intentionally Darkened Data
DD-Type 14: Fabricated and Synthetic Data
DD-Type 15: Extrapolating beyond Your Data

Chapter 2

DISCOVERING DARK DATA

What We Collect and What We Don't

Dark Data on All Sides

Data do not exist de novo. They have not been there since the beginning of time, sitting around just waiting for someone to come along and analyze them. Rather, someone has to collect them in the first place. And—as you doubtless expected—different ways of collecting data can lead to different kinds of dark data.

This chapter looks at the three fundamental ways that data sets are created, along with the dark data challenges associated with each method. The next chapter then explores some further dark data complications that can apply in many situations.

The three fundamental strategies for creating data sets are as follows:

1. Collect data on *everyone* or *everything* for the objects you are interested in.

   For human populations, this is what censuses strive to do. Likewise, stock-taking exercises aim to determine the details of everything in the warehouse or wherever. The annual stock take at London Zoo lasts about a week, revealing that (in 2018) there were 19,289 animals from

Data Exhaust, Selection, and Self-Selection

The computer has revolutionized all aspects of our lives. Some of these ways are obvious, like the word-processing software I am using to write this book, or the travel booking system I use when I buy an air ticket. Others are concealed, like the computers controlling the brakes and engine of a car, or those inside an elaborate printer or photocopying machine.

But whether the role of the computer is obvious or not, in all cases the machine takes in data—measurements, signals, commands, or other kinds of data—and processes them to make a decision or carry out some operation. Then, once the operation has been completed, the processing could stop. But often it does not. Often those data are stored, sent to a database, and retained. They are spin-off data, data exhaust, which can be examined later to gain understanding, improve systems, and decipher what happened if things went wrong. Black box recorders on aircraft are classic examples of this sort of system.

When data like this describe humans, they are often called *administrative data*.[2] The particular strength of administrative data is that they actually tell you *what people do*, not (as can be the case with surveys, for example) *what people say they do*. They tell you what people bought, where they bought it, what they ate, what web searches they made, and so on. Administrative data, it is claimed, get you nearer to social reality than exercises involving asking people what they did or how they behave. This has led to the accumulation of giant databases describing our behavior by governments, corporations, and other organizations. There is no doubt that these databases represent a great resource, a veritable gold mine of potential value enabling all sorts of insights to be gained into human behavior. From those insights we can improve decision-making, enhance corporate efficiency, and devise better public policy—provided, of course, that those insights are accurate and have not been contaminated by the impact of dark data. Moreover, there are privacy risks which arise when data we would like to keep dark become known to others. We'll return to issues of privacy at the end of the section, but let's look first at unsuspected dark data.

One obvious high-level gap is that administrative data do indeed tell you what people actually do—which is useful, unless you actually want to explore what people think and feel. Discovering that a population of people in a particular corporation are unhappy with the way things are going might be just as important as noting how they behave under the constraints and imperatives of the corporation's daily activities with their boss looking over their shoulder. To discover how they feel we would have to actively elicit data from them—perhaps in a survey, for example. Different kinds of data collection strategies are suited to answering different kinds of questions—and have different kinds of dark data challenges.

My own first serious exposure to dark data was in the area of consumer banking: the world of credit cards, debit cards, personal loans, auto finance, mortgages, and so on. Credit card transaction data involve giant data sets, with millions of customers making billions of plastic card transactions each year. Around 35 billion Visa card transactions were made between June 2014 and June 2015, for instance.[3] Whenever a purchase is made with a credit card, details of the amount spent, its currency, the vendor, the date and time of the transaction, and many other items of information are recorded (in fact, 70–80 items of information). Much of this information has to be collected so that the transaction can be made and the appropriate account charged; it's a necessary part of the operation, so that omitting these details is unlikely or even impossible. For example, the transaction could not take place without knowledge of how much to charge or who to charge it to. But other items of data might not be critical to the operation, so it is possible they might not be recorded. For example, omitting invoice numbers, detailed product codes, and unit prices would not interfere with the operation. Clearly this is an example of our first dark data type: *DD-Type 1: Data We Know Are Missing*.

Worse still, at least from a dark data perspective, while some customers will use a credit card for their purchases, others might use cash. This would mean that, as a record of *all* purchases and transactions, the credit card database would have unseen swaths of dark data, arising because of *DD-Type 4: Self-Selection*. Moreover, there are multiple credit card operators. The data from one operator may not be representative of the entire population of credit card holders, and certainly not of the entire population altogether. So, while holding great promise, administrative data might well have dark data shortcomings which are not obvious at first glance.

The particular problem I was presented with was a request to construct a "scorecard"—a statistical model for predicting whether an applicant was likely to default with repayments which could be used to guide the decision about whether the bank should give him or her a loan. I was supplied with a large data set giving the application-form details of previous customers, along with the outcome indicating whether or not those previous customers had actually defaulted.

In essence the exercise was straightforward. I needed to find what patterns of characteristics distinguished those customers who had defaulted from those who had not. Then future applicants could be categorized by determining whether they were more similar to the defaulters or the nondefaulters.

The trouble was that the bank wanted to make predictions about *all* future applicants. The data given to me were surely unlike the population of future applicants, because my data had already gone through a selection process. Presumably the previous customers had been given a loan because they were thought to be good risks according to some earlier mechanism—either a previous statistical model or perhaps a bank manager's subjective opinion. Those previously thought to be bad risks would not have been given a loan, so I knew nothing about whether they would actually have defaulted. Indeed, I had no idea how many applicants had previously been declined and not made it into my data set. In short, the data given to me were a distorted sample, subject to an unknown extent of selection or selectivity bias, and any statistical model built on this distorted data set could be very misleading when applied to the overall population of potential future applicants.

In fact, the problem was even worse than that. It actually had multiple layers of dark data. Consider the following:

Who actually applied? In the past the bank might have mailed potential customers asking if they would like a loan. Some would have replied they did want a loan, and others would not have replied. The data would include only those who had felt motivated to reply to the initial mailshot, and this might depend on how it was worded, how much was offered, the interest rate, and a host of other factors about which I knew nothing. The ones who had not replied would represent dark data.

Who received an offer? Those who replied would have been evaluated, and some of those would have been offered a loan, while others would not. But since I didn't know on what basis this offer had been made, I was presented with more dark data.

Who took up the offer? In addition to the preceding two selection processes, of those who had been offered a loan some would have taken it up, while others wouldn't have—introducing yet another layer of dark data.

Adding all these layers together made it very unclear how the data I was given related to the problem to be solved, which was to build a model to evaluate new applications. The multiple layers of dark data could mean that the sample I had, with all the known good/bad actual outcomes, was completely different from the population to which the bank would like to apply the model. Ignoring the dark data could be disastrous. (The bank still exists, so I suppose my model was not that bad!)

Administrative data are ubiquitous—just think of all the databases storing information about you relating to education, work, health, hobbies, purchases, financial transactions, mortgages, insurance, travel, web searches, social media, and so on. Up until very recently, in most of these cases your data were stored automatically, without your knowing about it and having a say in it. The European Union's General Data Protection Regulation (GDPR) has changed that—as you doubtless realize because of all the invitations to check boxes saying you understand and give permission for personal data about you to be recorded by websites. But occasionally you can have a say in other ways as well. (The protection of data of U.S. residents is regulated by both federal and state laws, varying by sector.)

In 2013 the UK National Health Service (NHS) launched a scheme whereby medical data would be copied from family doctor records each month and merged with hospital records in the national Health and Social Care Information Centre (HSCIC).

doubled, from 8,000 to 16,300.[5] There are various theories about why this might be happening, but one is that overstretched police forces are taking too long to answer. Another is that mobile phones are automatically generating such calls, perhaps as the buttons are accidentally pressed in a pocket or handbag.

If that last theory were the sole cause, we might expect the problem not to arise, or at least be less serious, in the United States, where the emergency call number 911 uses two different digits (it's 999 in the UK). But the rate of such calls is rising in America also. Records over three months from the Lincoln Emergency Communications Center illustrate the sort of change, with the percentage of abandoned incoming calls increasing from 0.92 percent to 3.47 percent from April through June 2013.

Abandoned calls are a clear case of *DD-Type 1: Data We Know Are Missing*. In contrast, a wonderful example of *DD-Type 2: Data We Don't Know Are Missing* was given by Mike Johnston in his column *The Online Photographer*.[6] He wrote: "I have to chuckle whenever I read yet another description of American frontier log cabins as having been well crafted or sturdily or beautifully built. The much more likely truth is that 99.9 percent of frontier log cabins were horribly built—it's just that all of those fell down. The few that have survived intact were the ones that were well made. That doesn't mean all of them were." Since there is no record of all the many log cabins which have collapsed and decayed, these are dark data.

DD-Type 2: Data We Don't Know Are Missing is particularly deceptive because we will generally have no reason to suspect it. Suppose, for example, we read, as I did in the *Times* (London) of 29 December 2017, that "the number of sexual assaults allegedly carried out by taxi drivers on passengers has risen by a fifth in three years, according to police figures." The immediate and superficial explanation is that more such offenses are being committed. But there is an alternative explanation, arising from dark data. This is simply that while the rate of commission of such offenses is remaining constant the rate of *reporting* of the offenses is increasing. Hitherto concealed dark data may be becoming visible as a result of changing social mores and societal norms. There's a general moral there: if you see a sudden step change in a time series of values, it could be because the underlying reality has changed, but it could also be because the data collection procedure has changed. This is a manifestation of *DD-Type 7: Changes with Time*.

A more elaborate example of *DD-Type 2: Data We Don't Know Are Missing* and *DD-Type 7: Changes with Time* working in tandem is illustrated by the performance of investment funds. The population of such funds is dynamic: new funds are set up, and old ones die. And, unsurprisingly, it is generally the underperforming funds which die, leaving just the ones which do well. Superficially, if we do not somehow take those that drop out into account, on average such funds will appear to do well.

Although individual funds which have dropped out because they performed badly will be excluded from an index showing overall or average performance, it might be possible to look back and obtain data on those funds. This would change them from *DD-Type 2: Data We Don't Know Are Missing* to *DD-Type 1: Data We Know Are Missing*, and it would then be possible to explore the impact of excluding them from the calculations. A 2006 study by Amy Barrett and Brent Brodeski showed that "purging of the weakest funds from the Morningstar database boosted apparent returns on average by 1.6 percent *per year* over the 10-year period [from 1995 to 2004]."[7] And in a study published in 2013, Todd Schlanger and Christopher Philips of Vanguard looked at the performance of funds including and excluding closed funds over 5-, 10-, and 15-year periods.[8] The differences were striking, with the performance of those excluding the closed funds over 15 years being almost double that when they were included. This study also revealed the magnitude of dark data in this context: only 54 percent of funds lasted the full 15-year period.

The phenomenon also affects more familiar financial indexes such as the Dow Jones Industrial Average and the S&P 500. Companies which perform poorly drop out of these indexes, so that only those which do relatively well contribute to the final performance value. This is fine if you happened to have invested in those companies which went on to do well, but not so fine otherwise. And since it's very difficult (some would say impossible) to tell which companies are going to go on to do well and which aren't, the index performance is deceptive.

Having cautioned about so-called survivor bias in financial indexes, it is worth noting that things can be more complicated. Taking hedge funds as an example, certainly, poorly performing funds are likely to close and not be included in the data, but so also are funds at the opposite end of the spectrum: exceptionally strongly performing funds are likely to close to new investors. Likewise, strongly performing companies can split and so drop out of a share index. Dark data can work in mysterious ways.

Additionally, for reasons we shall explore in chapter 3, there is a good chance funds which have performed exceptionally well in the past will nosedive in the future owing to the phenomenon of "regression to the mean." This means purchasers of funds need to look very carefully at how past performance is evaluated. As in other walks of life, investors need to ask themselves if the truth is being disguised by invisible dark data.

Survivor bias is always a potential problem for things which change over time. In the world of startups we tend to hear more about the successes than the failures—even though the majority of such companies fail. Some researchers put this failure rate as low as 50 percent, while others put it as high as 99 percent. Of course, it partly depends on the time period you are considering (one year, 50 years?) and how you define "failure." Take the social networking site Bebo, for example. Launched in 2005, at one stage Bebo was the most popular social networking site in the UK, with nearly 11 million users. In 2008 it was bought by AOL for $850 million. So, over a three-year horizon Bebo was hugely successful. But then the number of users started to fall, partly as they shifted to Facebook, and in 2010 AOL sold Bebo to Criterion Capital Partners. A computer glitch damaged its reputation, and in 2013 Bebo filed for Chapter 11 bankruptcy protection. Later in 2013 the original founders, Michael and Xochi Birch, bought the company back for $1 million. So is this a success or a failure? And what about Lehman Brothers? This firm was founded in 1850 and became the fourth largest investment bank in the United States—until it filed for bankruptcy in 2008, that is. Like Bebo, the company came to a sticky end, albeit over a longer time interval. But was it a success or a failure?

In the startup world people would naturally *like to* hear about the success stories more than about the failure stories, simply because they are trying to emulate the successes and not the failures. But this situation reveals another kind of dark data. What entrepreneurs should be looking for are characteristics which *distinguish between* successes and failures, not simply characteristics which happen to have been associated with successes. Characteristics of the latter kind might also be associated with the failures. Moreover, even if the characteristics are associated with successes more than failures, there is no guarantee they are causal.

The wonderful comic website *xkcd* has a cartoon about survivor bias.[9] The character is advising us never to stop buying lottery tickets, describing how he lost and lost time after time but kept buying tickets, even taking extra jobs to earn money to buy more tickets. And he eventually succeeded (if "succeeded" is the right word). What we don't see are the gamblers who poured fortunes into lottery tickets but died without winning.

In general, administrative data have immense potential to do good, provided we appreciate the dark data risks. But there is a complementary aspect which might be less positive that is leading to increasing concern.

From our individual perspective the data exhaust retained in an administrative data database is a *data shadow*. It consists of the traces we leave from sending emails or texts, tweeting, posting a comment on YouTube, swiping credit cards, using travel cards, making phone calls, updating a social media app, logging onto a computer or iPad, taking cash from an ATM, driving past a car license plate recognition camera, and so on endlessly, in often unsuspected ways. While such data can indeed be aggregated to benefit society, they also inevitably reveal a huge amount about each of us as individuals, our likes and dislikes, and our habits and behaviors. The data relating to us as individuals can be used to our benefit—guiding us toward products or events which might interest us, facilitating travel, and generally smoothing life out for us. But they can also be used to manipulate behavior. Authoritarian regimes can exert considerable control over us if they know detailed patterns of our lives. In a way this is inevitable: the downside of giving out information so that we can be assisted is that . . . we give out information.

Because of increasing concern about data shadows, services exist which will minimize our shadow. Or, from the perspective of this book, services exist to switch off the light on data, rendering them dark. Basic steps include deactivating all social media accounts (Facebook, Twitter, etc.), deleting old email accounts, deleting search results, using false information for accounts we