CONTENTS

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

outcomes all died within two days of admission. If we ignored those with unknown outcomes, we would justifiably conclude there was nothing to worry about, and all patients with trauma recovered. On this basis, we wouldn't treat any incoming trauma cases, expecting them to recover naturally. And then we would be horrified and confused by the fact that more than 11 percent of our patients were dying.

Before I go any further with this story, I want to reassure the reader. My extreme illustration is very much a worst-case scenario—we might reasonably expect things not to be that bad in reality—and Dr. Mirkes and his colleagues are experts on missing data analysis. They are very aware of the dangers and have been developing statistical methods to cope with the problem; I describe similar methods later in this book. But the take-home message from this story is that *things may not be what they seem*. Indeed, if there were a single take-home message from this book, that would be a good approximation to it: while it helps to have lots of data—that is, "big data"—size is not everything. And what you don't know, the data you don't have, may be even more important in understanding what's going on than the data you do have. In any case, as we shall see, the problems of dark data are not merely big-data problems: they also arise with small data sets. They are ubiquitous.

My story about the TARN database may be exaggerated, but it serves as a warning. Perhaps the outcomes of the 19,289 patients were not recorded precisely *because* they'd all died within 30 days. After all, if the outcome was based on contacting the patients 30 days after admission to see how they were, none of those who died would respond to questions. Unless we were aware of this possibility, we'd never record that any patients had died.

This may sound a bit silly, but in fact such situations arise quite often. For example, a model built to determine the prognosis for

patients being given a particular treatment will be based on the outcomes of previous patients who received that treatment. But what if insufficient time had passed for all such previous patients to have reached an outcome? For those patients the eventual outcome would be unknown. A model built just on those for whom the outcome was known could be misleading.

A similar phenomenon happens with surveys, in which *nonresponse* is a source of difficulty. Researchers will typically have a complete list of people from whom they would ideally like answers, but, also typically, not everyone responds. If those who do respond differ in some way from those who do not, then the researchers might have cause to doubt whether the statistics are good summaries of the population. After all, if a magazine carried out a survey of its subscribers asking the single question, Do you reply to magazine surveys? then we could not interpret the fact that 100 percent of those who replied answered yes as meaning that all the subscribers replied to such surveys.

The preceding examples illustrate our first type of dark data. We know that the data for the TARN patients all exist, even if the values aren't all recorded. We know that the people on the survey list had answers, even if they did not give them. In general, we know that there are values for the data; we just don't know what those values are.

An illustration of a different kind of dark data (*DD-Type 2: Data We Don't Know Are Missing*) is the following.

Many cities have problems with potholes in road surfaces. Water gets into small cracks and freezes in the winter, expanding the cracks, which are then further damaged by car tires. This results in a vicious circle, ending with a tire- and axle-wrecking hole in the road. The city of Boston decided to tackle this problem using modern technology. It released a smartphone app which used the internal accelerometer of the phone to detect the

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

course.) This example is an illustration of dark data *DD-Type 3: Choosing Just Some Cases*. Furthermore, that mode of dark data can even be a necessary consequence of collecting data, in which case it illustrates *DD-Type 4: Self-Selection*. The following are my two favorite examples of opposite extremes in terms of significance.

The first is the cartoon showing a man looking at one of those maps which are placed outside railway stations. In the middle of the map is a red dot with a label saying "You are here." "How," thinks the man, "did they know?" They knew because they recognized that *everyone* looking at the red dot had to be in front of the sign. It was a highly selected sample and *necessarily* missed everyone standing elsewhere.

The point is that data can be collected only if there is someone or something—a measuring instrument, for example—there to collect them. And the second extreme manifestation of this is described by the *anthropic principle*, which essentially says that the universe has to be like it is, or we would not be here to observe it. We cannot have data from very different universes because we could not exist in those and so could not collect data from them. This means any conclusions we draw are necessarily limited to our (type of) universe: as with the potholes, there might be all sorts of other things going on which we don't know about.

There's an important lesson for science here. Your theory might be perfectly sound for your data, but your data will have limits. They might not refer to very high temperatures or long times or vast distances. And if you extrapolate beyond the limits within which your data were collected, then perhaps your theory will break down. Economic theories built on data collected during benign conditions can fail dramatically in recessions, and Newton's laws work fine unless tiny objects or high velocities or

other extremes are involved. This is the essence of *DD-Type 15: Extrapolating beyond Your Data.*

I have a T-shirt with an *xkcd* cartoon with two characters talking to each other. One character says "I used to think correlation implied causation." In the next frame, he goes on to say, "Then I took a statistics class. Now I don't." Finally, the other character says, "Sounds like the class helped," and the first character replies, "Well, maybe."[7]

Correlation simply means that two things vary together: for example, positive correlation means that when one is big then the other is big, and when the first is small, the second is small. That's different from causation. One thing is said to *cause* another if a change in the first induces a change in the second. And the trouble is that two things can vary together without changes in one being the cause of changes in the other. For example, observations over the early years of schooling show that children with a larger vocabulary tend, on average, to be taller. But you wouldn't then believe that parents who wanted taller offspring should hire tutors to expand their vocabulary. It's more likely that there are some unmeasured dark data, a third factor which accounts for the correlation—such as the ages of the children. When the *xkcd* character says, "Well, maybe," he's acknowledging that it's possible that taking the statistics class caused his understanding to change, but maybe there was some other cause. We shall see some striking examples of this situation, characterized by *DD-Type 5: Missing What Matters.*

I've now mentioned several dark data types. But there are more. The aim of this book is to reveal them, to show how they can be identified, to observe their impact, and to show how to tackle the problems they cause—and even how to take advantage of them. They are listed at the end of this chapter, and their content is summarized in chapter 10.

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.

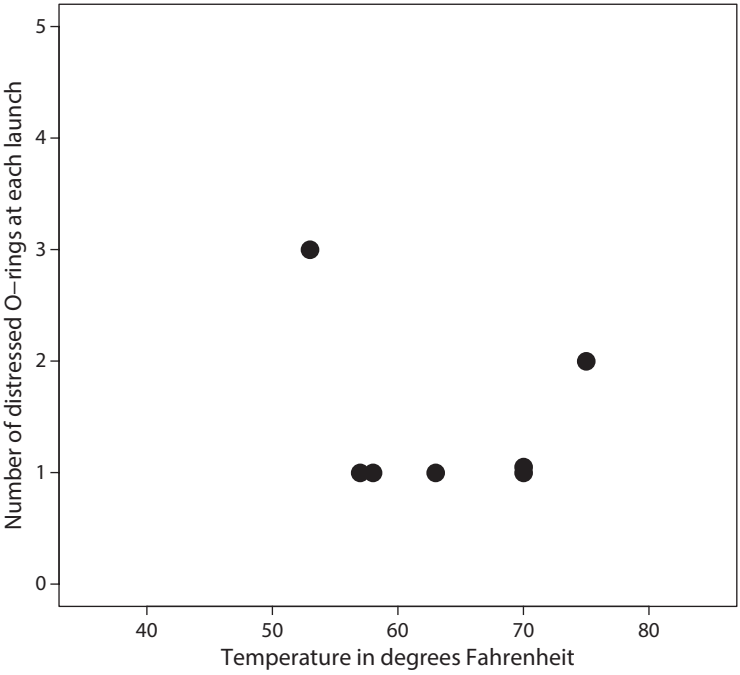
After the disaster, a commission headed by former secretary of state William Rogers drew attention to the fact that flights which had not had any O-rings showing distress had not been included in the diagram discussed at the teleconference (dark data *DD-Type 3: Choosing Just Some Cases* but also *DD-Type 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



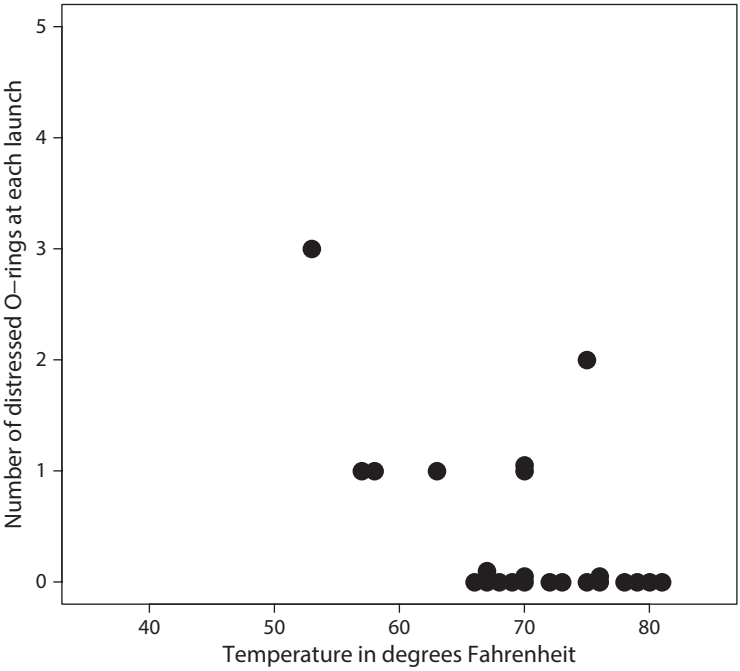


FIGURE 1. (a) Data examined in the *Challenger* prelaunch teleconference; (b) complete data.

when the temperature was less than 65°F experienced some O-ring distress, but only four of the 21 launches which took place at higher temperatures led to O-ring distress. The figure shows that the lower the temperature, the greater the risk. And worse, the predicted launch temperature was way below any previously experienced (*DD-Type 15: Extrapolating beyond Your Data*).

The missing data are crucial to understanding what is going on.

There is an interesting corollary to this story. Although it took months for the official report to arrive at its conclusion, Morton Thiokol's stock price crashed by 11.86 percent on the day of the disaster. Based on Morton Thiokol price movements prior to the incident, stock price changes of even as much as 4 percent would be rare. The stock prices of other companies involved in constructing the shuttle launch vehicle also slumped, but by far less. It was as if the market knew what was responsible for the crash. Dark data again?

The Power of Dark Data

That last example shows just how catastrophically badly wrong things can go when we fail to allow for dark data. It seems that they represent a real threat. But in fact the picture is not all gloomy. It turns out that an understanding of dark data can be used to our advantage, in a sort of data science judo. We can do this in several ways, as I describe in the second part of the book. Here's one way.

I introduce so-called randomized controlled trials in chapter 2, and in chapter 9 I return to look at them from a different perspective. In a medical context, in the simplest such trial two treatments are compared by giving one treatment to one group of people and the other to another group. However, there is a

risk. If the researchers know which treatment has been given to which people, this knowledge could influence the study. The researchers might be tempted to treat one of the groups more carefully than the other. For example, if the study aimed to compare an untested new treatment with the standard treatment, the researchers might monitor the former group for side effects more closely (perhaps subconsciously) or take more care in their measurement of possible outcomes. To overcome this potential bias, in such studies the treatment allocation is concealed from the researchers (*DD-Type 13: Intentionally Darkened Data*). The term *blinded* is used, to indicate that the data are dark.

Another familiar way in which dark data are used to advantage is in sample surveys. We might want to know the opinions of the people in a town, or of those who buy a company's products, and perhaps it is prohibitively expensive to ask them all. Certainly it is likely to be a time-consuming exercise, and opinions might change over its course. An alternative to asking everyone is to ask just some of them. The opinions of the others, the ones you don't ask, will be dark data. This might look like a high-risk strategy—it clearly resembles the TARN example. But it turns out that by using judicious methods of selecting which people you approach, you can get accurate and reliable answers—more quickly and cheaply than by attempting to approach everyone.

Yet a third way to use dark data to advantage is through so-called smoothing of data. As we shall see in chapter 9, this is equivalent to revealing unobserved and unobservable kinds of dark data (*DD-Type 14: Fabricated and Synthetic Data*), and it enables more accurate estimates and better predictions.

We shall explore other uses of dark data in chapter 9, where we will see that they often have exotic names. Some of them are widely applied in areas such as machine-learning and artificial intelligence.

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

INDEX