



Dark Data: Why What You Don't Know Matters

Mohieddine Rahmouni

► To cite this version:

Mohieddine Rahmouni. Dark Data: Why What You Don't Know Matters. *Technometrics*, 2023, 65 (1), pp.129-131. 10.1080/00401706.2022.2163804 . hal-04044127

HAL Id: hal-04044127

<https://hal.science/hal-04044127>

Submitted on 27 Apr 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Dark Data: Why What You Don't Know Matters

by David J. Hand, Princeton, NJ: Princeton University Press, 2020, xii + 330 pp., ISBN 9780691182377.

Mohieddine Rahmouni

To cite this article: Mohieddine Rahmouni (2023) Dark Data: Why What You Don't Know Matters, Technometrics, 65:1, 129-131, DOI: [10.1080/00401706.2022.2163804](https://doi.org/10.1080/00401706.2022.2163804)

To link to this article: <https://doi.org/10.1080/00401706.2022.2163804>



Published online: 01 Feb 2023.



Submit your article to this journal [↗](#)



View related articles [↗](#)



View Crossmark data [↗](#)

- (2022c), “Philosophy of Mathematics: Classic and Contemporary Studies,” *The Mathematical Intelligencer*, 1–3. [128]
- (2022d), “The Psychology of Mathematics: A Journey of Personal Mathematical Empowerment for Educators and Curious Minds by Anderson Norton,” *World Literature Today*, 96, 79–79. [128]
- (2022e), “All the Math You’ll Ever Need: A Self-Teaching Guide (3rd Ed.),” *Technometrics*, 64, 579–580. [128]
- Pápay, G. (2022), “Mercator’s Geometric Method in the Construction of His Projection from 1569,” *KN-Journal of Cartography and Geographic Information*, 1–7. [127]
- Ronga, F., and Vust, T. (1992), “Stewart Platforms Without Computer,” in *Proceedings of the International Conference on Real, Analytic and Algebraic Geometry*, pp. 197–212. [128]
- Stewart, I. (2021), *What’s the Use? The Unreasonable Effectiveness of Mathematics*, London: Profile Books. [128]

Firdous Ahmad Mala 

Government Degree College Sopore, Baramulla,
Jammu and Kashmir, India
firdousmala@gmail.com



Dark Data: Why What You Don’t Know Matters, by David J. Hand, Princeton, NJ: Princeton University Press, 2020, xii + 330 pp., ISBN 9780691182377.

While I was eagerly reading this valuable and unique book, which I believe fills a huge void in data science through a smooth explanation of Dark Data (DD), and where the author mixes humor with large-scale realistic examples from the real world in a wonderful and attractive artistic way, I came up with the idea of writing a review for it. This book is by David J. Hand, a data expert, and Professor of Statistics at the Department of Mathematics at Imperial College London, who is the author of several books on data science and statistics. In this book, David Hand investigates the ubiquitous phenomenon of DD in the real world. The author has given a taxonomy of 15 types of DD and has stated that he is sure that this is not an exhaustive list. As time goes on, new data sources, new ways of collecting data, and new types of data emerge, all of which bring with them new types of DD. I expect that this book, which does not contain statistical models or equations at all except for some simple tables and figures, will be a reference for many readers who specialize in statistics and machine learning or even those interested in data science and its interpretation on a nontechnical level. Thanks Hand! Using the third person pronoun makes the reader feel that you are addressing him (her) directly while taking him (her) on a wonderful and useful journey into the world of information full of DD that we do not see, such as this universe, which consists of a large part of dark matter. The missing data are just as important as the data we have. The data we can’t see have the potential to mislead us. The author shows how we can upend the traditional way of looking at data analysis. In his book he explores how and why DD arise and demonstrates the different types of DD and what leads to them, in addition to what we can do when we realize that DD are withheld from us, to leverage them to enable better decisions and actions.

The book consists of 10 chapters and is divided into two parts. The first part includes seven chapters on DD, their origins,

and their consequences. I enjoyed reading the first chapter about what makes up our world of DD that we don’t see. The book begins explaining Ghost Data with a joke about using the “elephant powder” and the example of being vaccinated against something that doesn’t exist. The term DD arose by analogy with dark matter in physics. About a third of the universe is made of this mysterious substance, which does not interact with light or other electromagnetic radiation and therefore cannot be seen. Because dark matter cannot be seen, astronomers were unaware of its existence for a long time. So, the data we don’t have may be more important than the data we do have in understanding what’s going on. DD are not only associated with big data but also arise with small data sets. Problems as well occur with opinion polls, where the lack of response is a source of doubt as to whether the statistics are good summaries of the community. Economic theories built on data collected during benign circumstances can fail dramatically in recessions. Also, Newton’s laws work well unless small objects, high speeds, or otherwise are involved. Thus, missing data are crucial to understanding what is going on. And since the number of potential causes of DD is essentially unlimited, knowing what kind of thing to watch out for can be very useful in helping to avoid mistakes and missteps. Thus, the functionality of DD types described in this book provides a classification that can help protect against errors including those caused by ignoring the distinction between data we know are missing and data we don’t know are missing.

Chapter two examines the primary methods for creating datasets and the DD challenges associated with each method. There is no doubt that different techniques of data collection can lead to different types of DD. Data that we don’t know are missing are rogue data because we generally wouldn’t have reason to suspect them. For data that we know to be missing, it is possible to explore the effect of excluding them from the calculations. The author has investigated the two types of data collection methods described at the beginning of this chapter: situations in which “all” data were recorded, and situations in which a sample of data was used show different types of DD can arise in each situation. The third mode of data collection is the experimental mode, where treatments and conditions to which objects (or people) are subjected vary in a fully controlled manner. However, several complexities emerge from the Hawthorne Effect, which is the way people should act differently if they know they’re being noticed. Sometimes decisions about what data to collect and what the results of analyses mean depend on past experiences, but these analyses can be subject to unconscious biases. Such as the availability bias which arises from the tendency to judge the probability of an event based on how easily an example can be recalled. Confirmation bias is another related example. Whereas the base rate fallacy and availability bias arise from ignoring the data that describes a population. In confirmation bias people actively, albeit unconsciously, seek data that are not adequately representative of the population. They look for information that supports their point of view and tend to ignore data that do not support it. The complementary aspect of confirmation bias is that people also tend to forget hard evidence if it contradicts their initial beliefs. Other examples cited by the author of how people draw inappropriate conclusions because they ignore part of the data (perhaps unconsciously) are negative bias, compliance bias, belief bias, and bizarreness effect.

Chapter Three then explores some of the additional intricacies of DD that can be applied in a multitude of situations. It is known that the usefulness of the data depends on the correct data collected without misrepresentation or distortion. But these terms are so vulnerable to potential DD risks that a comprehensive enumeration of the relevant risks is impossible. The author examines the data we aim to collect, and Chapter 4 considers how well we can achieve this, either way from the perspective of the dangers of DD. Through a few examples, the book shows how a basic type of DD arises from using inappropriate definitions or from not knowing what we're talking about. The data are intended for the process for which they were collected and may not be ideal for other purposes. Inappropriate definitions effectively obscure data of interest. Also, changing definitions over time can lead to changes to the data collected. But not everything can be measured as there always remain countless other properties that cannot be determined. These other characteristics are inevitable dark statements, with implications. Thus, the researchers need to be clear about the question they are asking, as whether the data are dark or not will depend on that question. The data they need to collect, the analysis they will undergo, and the answer they will get depend on what they want to know.

In Chapter Four, the book looks at how we are misled, even if we are sure of what we want to know. For example, statistical summaries and data analysis facilitate the assimilation of data, however, summarizing means sacrificing details and obfuscating the data. It doesn't tell us everything about the data. By obscuring the data, summaries may hide important information, and we should be alert to this. The catch is that we need to choose the summarization statistic(s) carefully to answer the question we want to ask. This approximation is not "wrong." Rather, it hides the details.

Chapter 5 looks at an entirely different class of ways in which DD can be created: games, feedback, and information asymmetry. For example, bubbles in financial markets do not reflect any real fundamental changes in value, but rather arise from a lack of critical assessment of the fundamental value of an asset: the mistaken belief that the fundamental value has increased. While the underlying value is one of the factors that influence stock prices, the bottom line is what others want to pay. On the topic of feedback, one of the main psychological drivers behind bubbles is confirmation bias. Which leads us to subconsciously search for information that supports our point of view and ignore data that do not support it. Deliberately creating incorrect and misleading information in this way may be more harmful than simply disguising the truth as DD.

Chapter 6 presents examples of intended dark statements such as the story of "The Man Who Sold the Eiffel Tower" where the victim could not tell people he had fallen for a trust scam and kept it a secret. Thus, there were layers of deception, each of which concealed the truth. The key to scams is to hide information about the real situation: that is, data hiding. But such deceptions also often depend on the human mind's tendency to make snap judgments rather than take the trouble to seriously weigh the evidence and carefully consider data. Because they are so diverse, a variety of different strategies are needed to tackle DD through sophisticated statistical approaches or through machine learning and data mining tools.

Chapter 7 is about science and DD where the author considers the foundation of the practice of science to be the Popperian (after Karl Popper) concept of "testability" or "falsifiability." The basic idea is to come up with a possible explanation — a theory, conjecture, or hypothesis — for the phenomenon being studied, and then test that explanation by seeing how well its results or predictions match what's going on. The paraphrase is done in the researcher's terms, he (she) knows what the unseen data should look like if his (her) theory is correct, then the experiments produce data that can be matched with the predictions. If the theory's prediction does not match reality, as indicated by the data, then the theory is replaced, modified, or expanded so that it also predicts the new theory. This is an example of the type of data extrapolation. So, the basic process of science is first to test theories against obscure, observable data, whereby a mismatch between the theory and that data leads to the theory being rejected or modified. We are supposed to realize that the mismatch may have another explanation. If theory and data do not match, it may be because something is wrong with the data. Data are always at risk of errors, measurement inaccuracies, sample distortions, and a host of other issues.

Among the examples cited by the author is how a lack of data misled scientist and chemist Linus Pauling about determining the structure of DNA in the mid-20th century. A Nobel Prize winner in chemistry, Pauling was initially not yet ready to accept that he was wrong until he later looked at Crick and Watson's structure and examined X-ray images. Further data gathering also led to the discovery that the universe is not only expanding but is expanding at an increasing rate. If we look at data, especially big data, in sufficiently different ways, it is possible to get to some realities. Economist Ronald Coase has interpreted this situation, saying that if we torture data long enough, they will confess. But like other confessions extracted under torture, the confession may not represent the truth.

The first part of the book explores the different ways DD can cause problems. In Part Two, the author looks at how DD can be discovered, allowed, and beyond that, how we can make use of DD. Chapter 8 looks at ways in which to investigate the shadows to discern what is hidden there, and ways in which problems might be mitigated. It outlines the ideas, tools, methods, and strategies that have been developed to guide us to the right answers, even if they are shrouded in a cloud of ignorance. The bulk of the chapter discusses situations where data are missing, and then at the end of the chapter moves on to discuss data that we can see, but that may be misleading. Regardless of the cause of the problem, a key element of the solution is awareness of what could go wrong. Perhaps this is especially important in situations where the data themselves cannot give a hint that something untoward is going on.

The key to dealing with DD is gaining an understanding of the cause of data loss. We need to explore the relationship between the data, which are observed or not, and whether an item is missing. This might give us an idea of what kind of values the missing elements could have, which in turn might allow us to compensate for them. Because different types of deficiencies require different types of solutions. This means that we need to be able to identify the category of any given missing data problem — and if we get it wrong, our conclusions may be

wrong. There are different strategies for doing this. Perhaps the most basic strategy is to use expertise about the area the data describe. Working with the data we must identify the mechanism of missing data gives us a solid handle on how to deal with the problem. But this requires a complex level of understanding, so various simpler methods are often adopted. These methods, which are often straightforward, are widely available in statistical software packages.

Chapter 9 deals with how to make use of DD. I see caution as the main message of this book. And there are ways we can use DD to our advantage, provided we know what we're doing and act very carefully. That is, there are ways we can transform the seeming ambiguity of DD to enable us to gain greater understanding, make better predictions, and choose more effective courses of action. We can do this by strategically ignoring pieces of data and by deliberately throwing them into the shadows. Random selection is also critical. It makes the process unmanipulable and not subject to intentional or unconscious biases. Simulation is an alternative strategy for investigating what might happen. Overall, this chapter has turned the concept of DD on its head. DD are often problematic: they hide things from us that we want to know, which can lead to distorted analyses and misunderstandings. This chapter described how data masking can be valuable and how it can lead to improved estimates and better decisions.


In addition to the examples in the previous chapters, Chapter 10 sought to classify the types of DD with examples of each of them, for use in real and practical situations. The taxonomy of DD-Types of dark data is summarized 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.

Funding

This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [Project No. GRANT2220].

ORCID

Mohieddine Rahmouni  <http://orcid.org/0000-0003-0248-4398>

Mohieddine Rahmouni 
Applied College in Abqaiq, King Faisal University,
Al-Ahsa 31982, Saudi Arabia
ESSECT, University of Tunis, Tunis, Tunisia
malrhmonie@kfu.edu.sa



Vedic Mathematics: A Mathematical Tale from the Ancient Veda to Modern Times, by Giuseppe Dattoli, Silvia Licciardi, and Marcello Artoli, World Scientific, 2021, 232 pp., \$68 (HB), ISBN 9789811221552.

Vedic mathematics is primarily a collection of methods, called the *sutras*, that aim at making one faster at numerical computations. These formulas are claimed to have been retrieved from the Vedas (Khare 2006); however, there is considerable disagreement on this issue (Dani 2001; Plofker 2008).

The Vedas are religious texts, originating in ancient India, composed in Sanskrit, and considered the oldest and holiest scriptures of Hinduism.

The authors, at the very outset, confess the fact that none of them is well-versed in Vedic mathematics. As luck would have it, the primary author, *Giuseppe Dattoli* happened to visit India for a conference during which one of his Indian friends gifted him a book, *Vedic Mathematics: Sixteen Simple Mathematical Formulae from the Vedas*, of a revered learned Indian monk, *His Holiness Jagadguru Shankaracharya Swami Bharati Krishna Tirtha Maharaj*. This book, first published in 1965, is considered to be the oldest on this subject. It has been the author's only encounter with Vedic mathematics.

As prerequisites, the authors claim that the book calls for knowledge of the calculus and a fair disposition in the art of computing on the part of readers. However, even a first reading of the book makes readers realize that knowledge of the basics of linear congruences, matrices, and recurrence relations will make the reading of the text more enjoyable and meaningful.

In its six chapters, the book, on account of various historical snippets, attempts to trace a brief history of mathematics from the ancient Vedic sources to the current times.

Chapter 1, *Mixing Up Ancient and Modern*, the shortest chapter of the book, starts with an engaging discussion on how Pythagoras' theorem was treated in the works of Euclid and that of ancient Indian Vedic texts. The authors find the Vedic treatment more effective from a practical point of view, despite its evident shortcoming in matters of rigor. Using the proof of Pythagoras' theorem, as given in the Vedic texts, the authors go on to demonstrate how Euclid's theorems could be derived as its consequences, contrary to the Greeks' practice of deducing the former from the latter. The authors also bring to light that the Babylonians were well-versed in the solution of a quadratic equation, and the calculation of cube roots.

They also show how, despite being similar in principle, the search for a divisibility criterion is, in practice, unlike the process of factorization of polynomials.

Chapter 2, *Divisibility Criteria, Osculator Numbers and Roots*, contains discussion on the divisibility tests for primes up to 23. A comparison is drawn between the modern-day techniques using polynomials and the Vedic mathematical techniques for divisibility testing. The Vedic approach makes use of the Osculator Number (ON), the Reduced Number (RN), and the Dividend Number (DDN). Examples, wherever needed, have been supplied for better and practical understanding. A separate section, *Osculator and Numbers*, has been dedicated to how geometry's osculating circle is connected to the divisibility of numbers.