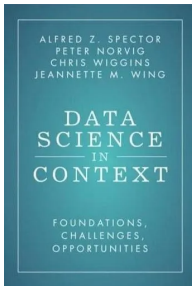
Review of [1]

Data Science in Context:
Foundations, Challenges, Opportunities

Alfred Z. Spector, Peter Norvig,
Chris Wiggins, and **Jeannette M. Wing**

Cambridge University Press, 2022
$39.99, Hardback, 335 pages



Review by

Shoshana Marcus
(shoshana.marcus@kbcc.cuny.edu)
Department of Mathematics and Computer Science
Kingsborough Community College
of the City University of New York

1 Overview

The term *data science* is a buzzword that has garnered much attention in recent years. Do you ever find yourself wondering, *what exactly is data science?* Do you find yourself thinking, *how does data science impact my life?* Do you find yourself considering, *how can data science further improve my life?* Do you find yourself contemplating, *how can I protect myself from the vulnerabilities exposed by data science?* If so, then this book should help you put these new trends into proper perspective!

As this book defines it, "Data science is the study of extracting value from data - value in the form of insights or conclusions." Data science combines advances in computing with the aspirations and methods of statistics and operations research. The process begins with the collection of large data sets, then the data is processed, conclusions are drawn, and the system applies what was learned to new sets of data. This cycle is naturally an ongoing process.

As the authors of this book observe, data science is *transdisciplinary*. A new field has emerged from the interactions between many different disciplines, thus, enabling data science to achieve its theoretical, methodological, and practical results. The field of data science derives primarily from the fields of statistics, operations research, and computing. The nascent field has also been informed by the sciences due to its abundance of applications with datasets so large they are otherwise untenable. The humanities and social sciences have provided perspectives that ensure societal benefits are maximized and harmful effects are kept at bay. This book demonstrates and emphasizes data science's integration of many forms of knowledge, techniques, and modes of thought.

"In this book, four leading experts convey the excitement and promise of data science and examine the major challenges in gaining its benefits and limiting its harm. They offer frameworks for critically evaluating the ingredients and the ethical considerations needed to apply data science productively, illustrated by extensive application examples" (prelude).

In the words of author Peter Norvig (p. 273), "The challenge for machine learning and data science is to build systems that align with society's real needs, and work for everyone. I hope this

book will inspire researchers to develop ideas that contribute to this; will enable developers to build systems that work for the betterment of all; and will empower consumers to know what they can ask for." This book presents a range of perspectives which should hopefully benefit researchers and users of data science alike.

2 Summary of Contents

In the Age of Technology, data science has become pervasive. Data science drives popular software used by billions of people every day, providing new tools, forms of entertainment, economic growth, and potential solutions to difficult and complex problems. These opportunities come with significant societal consequences, raising fundamental questions about issues such as data quality, fairness, privacy, and causation. Through many examples, this book illustrates data science's broad and growing impact, multi-faceted challenges, and its powerful future.

The challenging goal of data science is to build systems that align with society's real needs, and are beneficial to everyone affected. By understanding how things work and exposing vulnerabilities, we are able to increase benefits and reduce risks.

The pace of innovation is astounding and this makes it hard to respond to changes in time. The authors emphasize the ethical concerns of data science throughout this book. Although it is not simple to balance ethical considerations with primary objectives, all data science practitioners bear the responsibility of doing so. As implied by the title, the authors instill the mantra that a successful data scientist considers the context that defines success, expanding the primary focus of problem solving in ways that are effective and efficient. Consumers of data science products and results also play a role in ensuring that ethical considerations are given proper deliberation.

Data science relies heavily on artificial intelligence; big data and machine learning go together. Data science applications rely on algorithms for big data as well as statistical methods for deduction in heuristics. This book is replete with diverse examples across many domains. Some of these applications, such as product recommendation, are used by the population at large. Others applications are used by scientists; the field of computational biology has flourished due to the advances of data science. Scientific modeling problems such as protein folding and genome-wide association studies have become tractable with the tools and models that we have today, equipped to efficiently process enormous sets of data. Data science applications are continually evolving due to growth in data, advances in computational capacity and developments in machine learning.

Data science revolves around a virtuous cycle of increasing usage generating more data that improves quality and garners more usage. The data scientist needs to focus on effective data gathering, modeling, and application, as well as error correction. As this book explains, the route-finding software we have come to rely on (e.g., Waze and Google Maps) are not simply performing the operations research task of finding the shortest path while taking speed limits into account. These sophisticated applications are data driven since their solutions dynamically adapt based on factors that are constantly in flux, such as the current road conditions and historical delays (like rush hour). We would also like the algorithms to take the safety of drivers into account and to consider the privacy of homeowners on quiet streets. Drivers have low tolerance for small errors, such as directing a user to travel on a closed road or to drive the wrong way on a one-way street. When these applications are embedded in self-driving cars, we have even lower tolerance of failure. Errors have extreme legal, ethical, and financial risks.

Spelling correction is one of the first applications discussed at length in this book. Intuiting

the word intended by the typist, when errors are allowed, is not as simple a problem as it seems at first glance. Not only do we expect spelling correction in documents we type, but we rely on these tools to use search engines on the World Wide Web effectively. Speech recognition is more complex and nuanced than spelling correction, and thus is not as simple to formulate or to solve. Music recommendation systems take this to a whole new level and introduce many other complicating factors. It becomes difficult to even intuit the users' preferences and to clearly define the problem. In music recommendation systems, the users' reactions are tracked - how you listen to music, skip around, etc. This brings us to consider privacy concerns.

The Covid-19 pandemic and vaccination controversies form a case study throughout this book. This is a topic that many readers can easily relate to due to its recentness and the abundance of uncertainty and controversy involved. Covid-19 mortality predictions were vastly disparate from actual occurrences. This is certainly humbling and can be understood within context. The data were insufficient and erroneous, the necessary models are inherently complex, the disease is rapidly changing by its nature, and this complexity was exacerbated by feedback phenomena that were catalyzed in part by government actions.

Consumers have come to depend on recommendation systems because the Web has grown so large and no individual can sift through all the information. This benefits the sellers as well since they can reach much larger markets and do more business, which in turn has encouraged many sellers to embark on e-commerce. Similarly, in the financial sector, we have all come to rely on data science to detect fraudulent activity in our checking accounts and credit cards, as well as to aid analysts in predicting stock price fluctuations.

This book introduces the Analysis Rubric for data science, which delineates the major considerations for evaluating data science's suitability for a proposed application. Table 1 delineates the elements of the Analysis Rubric and its application to spelling correction. Spelling correction readily meets the needs surfaced in the Analysis Rubric; this explains why it performs quite well in applications. Speech recognition meets the needs of the Analysis Rubric almost as well, even though it is a more complex problem. Music recommendation is quite challenging for data science, and this is indicated by the Analysis Rubric since the objectives are not even clear. Covid-19 mortality prediction systems did not fare well; applying the Analysis Rubric highlights some of the major interferences.

Tractable data	One can easily procure an appropriate corpus of online text.
Technical Approach	A basic version is relatively simple to code.
Dependability	Privacy and security are not major issues; care must be taken to prevent an attacker from spamming the system with incorrect spellings (perhaps to promote their brand name).
Understandability	Users don't really care how a spelling corrector works, nor does the system need to understand the cause of a spelling error.
Clear Objectives	The clear goal is providing the correct spelling of the word the user meant to type.
Toleration of Failures	Users are accepting if the system does not correct a word's spelling or guesses incorrectly.
Legal, Ethical concerns	Do not seem to pose problems.

Table 1: Analysis Rubric applied to spelling correction (adapted from pages 63-64)

This book devotes a separate chapter to each element of the Analysis Rubric and surveys the challenges in that aspect of applying data science properly. First there is the need for sufficient data of reliable quality. In many applications, gathering data is difficult amidst privacy concerns. This is a fundamental problem in medical research. Additionally, abusers of systems can intentionally misrepresent data. It is important to enforce the most restrictive data access rights as possible to ensure data retains its integrity.

It is interesting to hear anecdotally, in the authors' experience advising companies, that many companies are excited to employ data science because they have the initial data. Yet, they are apprehensive about the mathematical complexity of building a machine learning model. Over time, they often realize that model building may be easier than establishing and maintaining a data pipeline.

Designing and deploying dependable models requires much ingenuity. It is challenging to derive truthful insights and conclusions. The inherent uncertainty poses a significant problem. For some inputs, even the best experts disagree on *the* correct output, so any model will necessarily disappoint some of the time. Another problem is that the world is constantly changing. A system trained on historical data may no longer perform well in the future. It is vital to continuously monitor a deployed system to watch out for any unexpected changes, and to correct for them by updating the model. A third problem is that it can be difficult to specify what we want to optimize in an application. The designer must be careful that inductive bias does not creep in when software makes assumptions to infer generalization from the training data. This can happen easily when hypothesis drives the design of a system.

In order to be adopted by society, a data science application must demonstrate its dependability. It often takes more time and effort to make a system dependable than to collect the data and build the algorithmic model. A chapter of this book addresses four key aspects of dependability: privacy, security, resistance to abuse, and resilience.

A chapter of this book explores the understandability necessary in data science. In some applications, it is not sufficient for the algorithm to make accurate predictions; the users want to understand what is going on. Understandability is important to the developers of the application (to help them do their jobs), to the users (to trust the application), to the general public (for assurance that the application will not harm society), to the regulators (for compliance and accountability), and to the scientific community (to reproduce results or to extend them).

The authors contend that data science done poorly obfuscates facts and truth. In the words of the authors (p. 186), "Every dataset has a story to tell, but we need to tease out that story by clearly explaining where the data came from and what it says, putting it in the context of other studies of the same phenomenon, making sure we distinguish correlation and causation, and presenting the story in a way that is not prone to cognitive biases and will not lead to misunderstanding."

An essential and non-trivial component in using data science effectively is in setting the right objectives. Often, the objectives seem clear at the beginning of project development, but later in the process, team members realize that objectives are not as clear as they initially seemed to be. This difficulty in setting objectives regularly arises from data science being applied to really complex and hard problems.The text goes through several examples of data science projects to demonstrate the challenge in establishing clear objectives that are acceptable across parties involved.

Traditionally, software developers strive to eliminate bugs and achieve certainty that the software computes the correct answer. In data science, additional challenges get in the way and errors are likely to perpetually occur. As such, the scenarios must inherently be tolerant of some failures,

to a reasonable degree. If not, perhaps the setting is not amenable to a data science solution. We must remember, it is often difficult to agree upon *the* correct answer to a data science problem.

Interestingly, it is mentioned near the end of this book that the United States is one of many countries that are racing to outdo each other at data science and artificial intelligence. Governments are investing heavily in these efforts. This is reminiscent of the race to the Moon during the Cold War. Competition builds momentum and drives progress.

The key ideas of this book are summarized in this fitting paragraph (adapted from p. 229). "Data science has been successfully applied in many applications and it will be applied to many more. New techniques, greater computational power, and creativity will combine to make currently impossible and impractical applications feasible. Individuals and institutions dependent on data science for their success are likely to become even more so." The societal concerns were minimal a few decades ago, when data science was in its infancy. These concerns have grown substantially and are garnering increasing attention at both the business and the political level.

The concluding section of this book provides pragmatic recommendations. The authors present a clear argument that the prevalence of data science and the diversity of its manifestations beg for increased educational opportunities in data science at all levels. Thus, data science should be woven into existing curricula at primary, secondary, and post-secondary levels, as a mandatory and essential component. Author Alfred Spector ran a seminar at MIT based on this book in Spring 2023. The companion website DataScienceInContext.com includes class materials replete with lecture slides, so that the contents of this book can be brought to other educational settings. Although it is complicated to put in place, more legally mandated regulation is certainly in order. It seems necessary and prudent to update existing laws to encompass the societal changes that have been brought about by the proliferation of data science applications.

3 Opinion

This book is accessible to all who are interested in learning about the developing field of data science. It should open discussions and guide each one of us to new realizations. Understanding what data science is and how it is useful should stimulate those who develop the tools, those who use the tools, and policy makers, to understand how data science can continue to be used more effectively as well as more ethically across many domains. Consumers of data science products and their outputs should feel empowered to know what they can ask for.

In the short interim since this book was published, ChatGPT has emerged. Now generative AI tools have become widely available and are now used in a wide variety of settings, to improve quality of output as well as to boost productivity. Policy makers are grappling with the ethical implications of these new tools. Users are trying to understand how much they can trust these tools and to develop an appropriate level of reliance on them. The topics of this book certainly relate to the range of issues that arise in data-driven AI. Thus, this book has become increasingly more relevant to a broader audience now that generative AI tools have become widely available.

The contents of this book are well organized, providing many tables and charts to organize ideas. This book is well written and technically correct, as evidenced by the abundance of references. Technical terms mentioned in this book are first defined and then thoroughly explained, making no assumptions about the reader's familiarity with these terms. Although no formal background in any discipline is necessary, a rudimentary familiarity with statistics makes the book that much easier to read.

The four authors of this book are committed to promoting data science education. It is this commitment which inspired them to write this book. It is my hope that the book is successful in making the field of data science approachable and understandable to all!