Introduction

The quantity of data that is collected, processed, and employed has exploded in this millennium. Many organizations now collect more data in a month than the total stored in the Library of Congress. With the goal of gaining insight and drawing conclusions from this vast sea of information, data science has fueled many of the vast benefits brought by the Internet and provided the business models that pay many of its costs.

Beyond the marvels of present data science applications, there are even greater breakthroughs on the horizon: semi-autonomous cars and trucks, and perhaps even fully autonomous ones; widespread precision medicine leading to longer and healthier lives; transformative improvements to education and the pursuit of science; new ways to pursue the humanities; and evolution in the workplace. There are new data science applications brewing in almost every field of human endeavor.

However, no new technology arrives without complications: Some of the complications are technological, based on challenges in both developing algorithms and then perfecting software and computer systems. For example, with so much data and processing capability, there are inevitably security, privacy, and reliability challenges. And, if applications of data science become as omnipresent as predicted, society needs the technologists to ensure they are rock solid.

Some complications are broader, relating to the very premise of using data to valuable effect. With mountains of data and correlations becoming available, we need to learn to cut through them to ascertain fundamental truths, not erroneous associations which may obfuscate the truth. Deeper risks arise when using data in decision systems; as new applications become available, and we can predict and optimize many outcomes, we must decide what we are really trying to achieve.

Some challenges are truly fundamental, as data science may change the operation of our society and impact our own humanness. We must come to grips with limitations on how much mechanistic advice and control we are willing to act on or even receive. As these systems alter our jobs and our socio-political systems, we will need to understand their effects and adjust in ways that we do not yet understand. Data science is affecting us already, and it may even challenge our notions of ourselves as the intelligent masters of our world.

https://doi.org/10.1017/9781009272230.002 Published online by Cambridge University Press

Because of these very broad impacts, data science as a field has led to entirely new research agendas outside of its foundational fields of computer science, statistics, and operations research. Data science is also changing many other disciplines (e.g., how we think about and practice political science, but many more). There are also growing transdisciplinary relationships between data science and many of the humanities and social sciences.

This book's holistic approach to data science leads us through these topics:

- Part I, Data Science, provides a unifying definition of data science and sets forth the field's goals. It then provides a historical perspective on how data science arose from its foundational fields (statistics, operations research, and computing, metaphorically illustrated in Figure 1) and describes its relationship to the sciences, social sciences, and humanities. The historical story is an exciting one due to exceedingly rapid progress that has changed the course of technology, many domains of applicability, and even our society writ large.

Figure 1 This metaphorical braid shows the integration of the foundational fields, labeled S (statistics), OR (operations research), and C (computing).

- Part II, Applying Data Science, presents examples of data science applications from the domains of technology, commerce, science, medicine, and more. Based on our detailed exposition of six applications, the chapter develops a seven-element Analysis Rubric to help us analyze the relative ease or difficulty of applying data science to other applications. We then review 26 more applications against the Analysis Rubric. Some of these are straightforward; others gnarly but feasible; yet others nearly impossible. Almost all have unintended consequences that require care and thought. Figure 2 illustrates this part's flow.

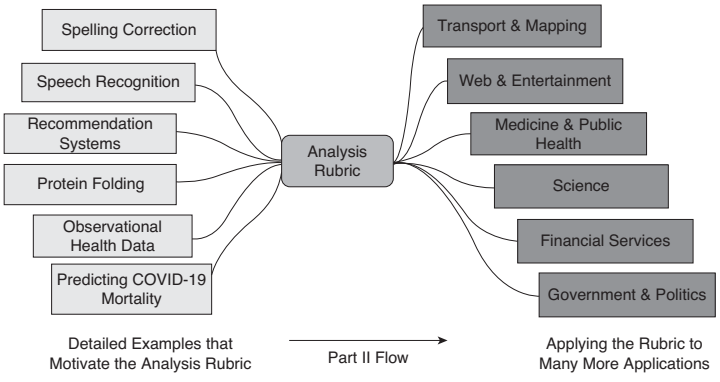

Figure 2 Part II introduces six applications, uses them to induce an Analysis Rubric, and then demonstrates its applications.

https://doi.org/10.1017/9781009272230.002 Published online by Cambridge University Press

- Part III, Challenges in Applying Data Science, builds off the seven elements in the Analysis Rubric to present the technical, contextual, and societal challenges in making data science work well (this is illustrated in Figure 3). With care, users of data science can often navigate many of these challenges effectively. However, some are perilous and very difficult to resolve, implying that, in some cases, data science is simply not the right tool for the job. Part III is quite clear about the risks of the misapplication of technology.

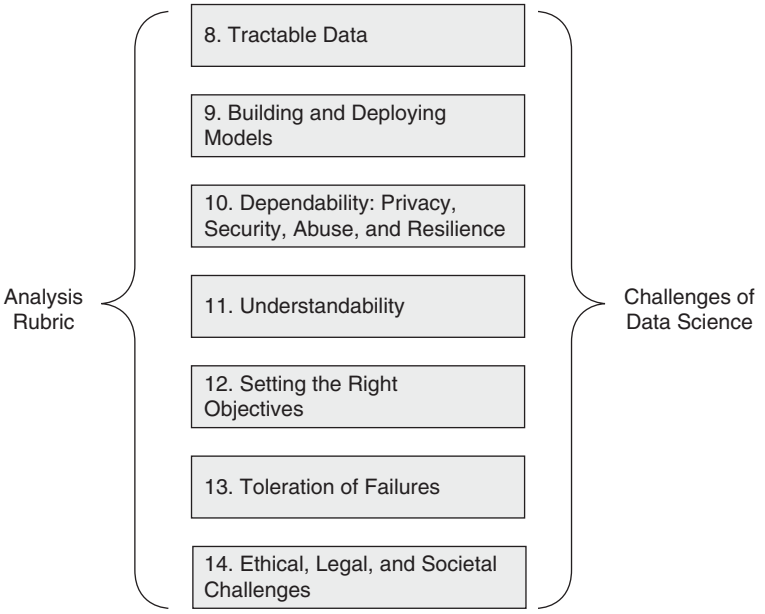

Figure 3 The Analysis Rubric's seven elements motivate the challenges in Chapters 8 to 14.

- Part IV, Addressing Concerns, describes many societal concerns regarding data science and its applications – concerns which in turn are influenced by Part III's challenges. It then discusses some approaches for mitigating these concerns while still allowing us to reap the rewards. In some cases, we make prescriptive proposals: For example, we recommend increasing data science education at the secondary school level and above, even if this means reprioritizing a little of the current mathematics curriculum and substituting more probability, statistics, and computing. In other areas, we only set forth some considerations that decision makers should take into account. Part IV's flow is described in Figure 4.

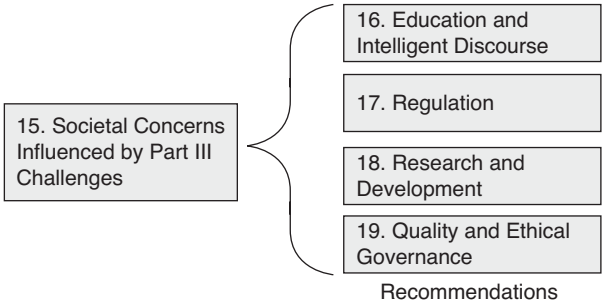

Figure 4 Part IV's summary of societal concerns motivate the recommendations in Chapters 16 to 19.

- An ethics thread flows through the book, with a focused section near the end of each part. Data science must consider ethical matters carefully because many data science applications have significant societal consequences and often rely on personal data to create computational models. As illustrated in Figure 5, the thread starts by defining ethical principles relevant to data science and then reviews some of the Part II applications in light of those principles. While most chapters of the book (and particularly those in Part III) present ethics-related issues, the ethics thread augments these discussions with the organizational challenges of balancing incentives and governance to achieve good outcomes. The ethics thread concludes in Part IV, which ends with three recommendations.

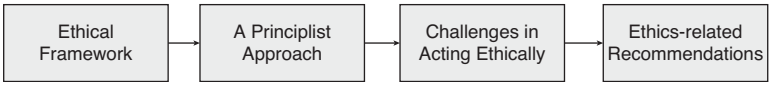

Figure 5 This figure illustrates the flow of the ethics thread, which spans Parts I to IV.

While the chapters build on each other, data science courses will vary in what examples from Part II they emphasize. Some readers may choose to omit Part IV (which bridges from the *challenges* of Part III to societal *concerns*), while others may wish to omit some of the technical details in Part II and Part III.

In all, this book's broad perspective on the field of data science aims to educate readers about the data science applications they regularly use, to apply that understanding to new applications, to more fully recognize the challenges inherent in data science, and to educate and catalyze thoughtful analysis, debate, and action to make data science ever more beneficial.

https://doi.org/10.1017/9781009272230.002 Published online by Cambridge University Press