DATA SCIENCE IN CONTEXT

Data science is the foundation of our modern world. It underlies applications used by billions of people every day, providing new tools, forms of entertainment, economic growth, and potential solutions to difficult, complex problems. These opportunities come with significant societal consequences, raising fundamental questions about issues such as data quality, fairness, privacy, and causation.

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

The authors' far-ranging exploration of these complex issues will stimulate data science practitioners and students, as well as humanists, social scientists, scientists, and policy makers, to study and debate how data science can be used more effectively and more ethically to better our world.

Alfred Z. Spector is a Visiting Scholar at MIT, who has innovated in large-scale, networked computing systems and provided research leadership to many. After beginning his career on the Carnegie Mellon faculty, he founded Transarc, and then led IBM Software Research. He was later Vice President of Research and Special Initiatives at Google and then CTO at Two Sigma Investments.

Peter Norvig is a Distinguished Education Fellow at Stanford's Human-Centered Artificial Intelligence Institute and a research director at Google; previously he directed Google's core search algorithms group and NASA's AI efforts.

Chris Wiggins is an Associate Professor of Applied Mathematics at Columbia University and the Chief Data Scientist at *The New York Times*. At Columbia he is a founding member of the executive committee of the Data Science Institute, and a member of the Department of Applied Physics and Applied Mathematics as well as the Department of Systems Biology, and is affiliated faculty in Statistics.

Jeannette M. Wing is the Executive Vice President for Research and Professor of Computer Science at Columbia University and was the inaugural Avanessians Director of its data science institute. She is known for her research contributions in security and privacy, programming languages, and concurrent and distributed systems.

"This book provides an important view of the contextual landscape for data science: the context of related fields of statistics, visualization, optimization, and computer science; the context of a broad range of applications, together with an Analysis Rubric; the context of societal impacts from dependability, to understandability, to ethical and legal questions. These are critically important factors for any practitioner of data science to understand, and for others to be aware of in evaluating the use of data science."

— Daniel Huttenlocher, Massachusetts Institute of Technology,
co-author of *The Age of AI*

"As data science becomes a crucial element in momentous decisions of war and peace, as well as commerce and innovation, it is vital that it rests on sound foundations. This book is an important step forward in that regard, illuminating the context in which data science is practiced. It is essential reading for both data scientists and decision makers."

— James Arroyo, OBE, Director of the Ditchley Foundation

"Data science touches every aspect of our modern lives. This book digs into the practical, legal, and ethical challenges that result. It is the only book that's comprehensive in its consideration of the thorny issues arising from the broad application and unprecedented growth of data science. If you do data science, you should read this book."

— Michael D. Smith, John H. Finley, Jr. Professor of Engineering
and Applied Sciences, Harvard University

"This book will be essential reading for all data scientists and data teams. The self-contained text explains what students and practitioners need to know to use data science more effectively and ethically. It draws on the authors' years of experience and offers practical insights into data science that complement other books that focus on specific techniques. I'll be referencing and recommending this book for many years to come."

— Ben Lorica, Gradient Flow, host of *The Data Exchange* podcast

DATA SCIENCE IN CONTEXT

Foundations, Challenges, Opportunities

ALFRED Z. SPECTOR
Massachusetts Institute of Technology

PETER NORVIG
Stanford University, California

CHRIS WIGGINS
Columbia University, New York

JEANNETTE M. WING
Columbia University, New York

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781009272209
DOI: 10.1017/9781009272230

First published 2023

Printed in the United Kingdom by TJ Books Limited, Padstow Cornwall

A catalogue record for this publication is available from the British Library.

ISBN 978-1-009-27220-9 Hardback

Contents

Figures

Tables

Preface

Combine unprecedented scientific and engineering advances in computing with the aspirations, methods, and advances in statistics and operations research, and we get the field of data science, which broadly aims to extract insights or conclusions from data. Data science has come into existence due to rapidly increasing capabilities to collect, process, and learn from data, and then to apply what was learned with near- and long-term benefits.

Even though the term "data science" only began to be used widely circa 2010, it has had enormous effects on science, engineering, commerce, and society at large, and the field has explosively grown in vitality and impact by almost any metric: Educational programs in the field are blossoming, as is employment. Social networks, online shopping, streaming entertainment, internet search, new cancer treatments, many scientific discoveries, and semi-automated driving are not solely due to data science, but it plays a huge and central role in each. Most household name companies, whether in technology, pharma, logistics, finance, education, or another field entirely, are heavily based on data science techniques.

However, as the field has grown, so have public concerns about it, including, but not limited to the following:

- Economic and fairness impacts on people and institutions
- Potential and actual misuse of personal data
- Effects on harmony and governance
- Power consumption
- General mistrust

It seems every day we hear of a new concern garnering attention, whether well- or ill-founded. Perhaps this is unsurprising, as data science impacts so many aspects of life. Most innovations, no matter how good they are, have unintended consequences.

Data science's juxtaposition of opportunities and challenges gave rise to this book. By illustrating and exploring the complex issues, we aim to provide both

students and practitioners the ability to use data science more effectively and more ethically. We offer a method for critically evaluating data science's applicability to particular problems, an extensive list of examples, and a detailed discussion of the technical, societal, and ethical challenges that data scientists must navigate.

Part I begins by delineating the field, explaining its historical roots in statistics, operations research, and computing. We then start a thread on ethical considerations in applying data science, which continues through later parts of the book.

Part II then describes more than 30 data science applications with three goals:

- Explaining aspects of how these applications work
- Illustrating the complexities in making them work *well*
- Introducing an Analysis Rubric that practitioners can apply to tease out those complexities when applying data science to new problems

Motivated by this Analysis Rubric, Part III delves into the technical, contextual, and societal challenges of data science, including privacy, security, the complexity in setting objectives, and many ethical issues.

Part IV describes societal concerns with the unintended consequences of data science, and then makes recommendations for ameliorating some of them. We summarize our major points in Chapter 20.

Our journey will intermingle topics in data science, its technological underpinnings, and related fields. In part, this is because data science arose through the confluence of diverse technical, scientific, and commercial advances. We believe this breadth is needed to explain how data science has become so important, how it solves problems, and what challenges exist.

As an example, topics like the growth in power of computation and computer security may not seem to be primarily data science topics. But vast computing capability makes data science feasible, while security issues force us to temper our enthusiasm with a deep consideration of risk. We were guided in choosing topics for this book by a desire to enhance our and our readers' understanding of data science and its future.

We do not duplicate textbooks on the theory and application of data science techniques, but instead address the breadth of data science, a field in which the revolutionary growth in computing coupled with advances in statistics and operations research is changing almost all aspects of society. We believe this material can be the basis for a full course, though we recommend adding supplemental case studies and analyses. We also believe this book provides important perspectives that are a useful addition to courses that focus on statistical, operations research, or computational techniques. We hope it will also be useful to technically oriented professionals wanting to apply data science to new problems. Finally, we have tried to make the material accessible to non-experts, particularly in public policy or

business, who are interested in the benefits and challenges at the confluence of data science, technology, and society.

This is a fast-moving field, and we anticipate providing additional commentary, questions, and updates on the book's website, DataScienceInContext.com.

Alfred Z. Spector
Peter Norvig
Chris Wiggins
Jeannette M. Wing

Acknowledgments

We are deeply grateful to many who have given us specific advice and comments on this manuscript: Alfred Aho, James Arroyo, Robin Berjon, David Blei, Aaron Brown, Sarah Cen, Brenda Dietrich, Ulfar Erlingsson, Kaylee Fisher, Ben Fried, Alon Halevy, Mark Hansen, Reece Hirsch, George Hripcsak, Daniel Huttenlocher, Jon Kleinberg, David Konerding, Rhonda Kost, Aleksander Madry, Preston McAfee, Vikram Modi, Nicola Phillips, Calton Pu, Alexander Rodriguez, Roni Rosenfeld, Daniel M. Russell, Thomas Sakmar, Teymour Shahabi, David Shaw, James Shinn, Asher Spector, Benjamin Spector, Emily Spector, Billie-Grace Ward, Martin Wattenberg, Peter Weinberger, the spring 2022 students in MIT's class 6.S978, and several anonymous reviewers. We benefited greatly from the often extensive feedback we received, in part because the book covers such a broad collection of topics.

We are also very grateful to the following five people who improved the prose in this book: Nikki Ho-Shing worked on the transcription of co-author Alfred's 2015 talk, "Opportunities and Perils in Data Science," which was the book's initial seed (Spector, 2016); Cindy Bloch assisted in many ways – particularly by corralling our roughly 400 citations; Tom Galloway made two passes over our prose, suggesting thousands of improvements for readability, consistency, and comprehensibility; and Lauren Cowles, our editor at Cambridge University Press, and Geoff Amor, our freelance copy editor, had invaluable structural and editorial advice.

Co-author Jeannette thanks Armen Avanessians who endowed the directorship of the Data Science Institute at Columbia University, giving her the opportunity to explore data science in its breadth. We all thank the institutions at which we have taught and worked for giving us opportunities to explore and solve challenging problems.

Finally, we recognize the impact of our own teachers, and we cherish all the interactions we have had with students, colleagues, seminar attendees, clients, and user communities. We could never have written this book without them.

Any shortcomings in the book are due to us, not to any who have assisted us.