Springer Series in the Data Sciences

Springer Series in the Data Sciences focuses primarily on monographs and graduate level textbooks. The target audience includes students and researchers working in and across the fields of mathematics, theoretical computer science, and statistics. Data Analysis and Interpretation is a broad field encompassing some of the fastest-growing subjects in interdisciplinary statistics, mathematics and computer science. It encompasses a process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision making. Data analysis has multiple facets and approaches, including diverse techniques under a variety of names, in different business, science, and social science domains. Springer Series in the Data Sciences addresses the needs of a broad spectrum of scientists and students who are utilizing quantitative methods in their daily research. The series is broad but structured, including topics within all core areas of the data sciences. The breadth of the series reflects the variation of scholarly projects currently underway in the field of machine learning.

More information about this series at http://www.springer.com/series/13852

Jeffrey C. Chen • Edward A. Rubin •
Gary J. Cornwall

Data Science for Public Policy

Springer

Jeffrey C. Chen
Bennett Institute for Public Policy
University of Cambridge
Cambridge, UK

Edward A. Rubin
Department of Economics
University of Oregon
Eugene, OR, USA

Gary J. Cornwall
Department of Commerce
Bureau of Economic Analysis
Suitland, MD, USA

ISSN 2365-5674 ISSN 2365-5682 (electronic)
Springer Series in the Data Sciences
ISBN 978-3-030-71351-5 ISBN 978-3-030-71352-2 (eBook)
https://doi.org/10.1007/978-3-030-71352-2

Mathematics Subject Classification: 98B82, 62-01, 62-04, 62-07, 62H11, 62H12, 62J05, 62J07, 68T50, 68-01, 91-01, 91C20, 68N01, 68N15, 62P20, 62P25, 62P12

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To our families:

Maya and Olivia

Michelle, Theo and Clara

Danielle and Claire

Preface

Public policy has long relied on the empirical traditions of econometrics and causal inference. When new regulations, laws, and new programs are created, policy analysts and social scientists draw on experimental and quasi-experimental methods to assess if these policy implementations have any effect. This empirical evidence is then added to a body of knowledge that identifies causes and effects, which in turn inform future policies that can improve people's lives. The feedback loop can take years (if not decades), but with the advent of data science, we are presented with the opportunity to increase the velocity and precision of decisions. While data science is integrated into many companies of private industry, its role in government is still evolving and being clarified.

Policy analysts strive to understand the measured effects of decisions, but analysts can conflate this measurement with predicting who will most benefit from a policy. Understanding an effect is the pursuit of knowledge, looking back into the annals of history to offer an explanation that satisfies the *whys*. In contrast, being able to anticipate what will happen to program participants is a skill that de-risks decisions, providing operational knowledge about *who* and *what*. These distinctions are not well known, yet government agencies are investing in their data science capabilities.

From working in public policy and government operations, we have had the opportunity to work on data science in this special context. When empowered by the government apparatus, data scientists can unlock the possibility of precision policy. Not only can data science inform policy makers and decision makers to arrive at on-point decisions, but they can develop machine learning applications that have transformative potential. Machine learning can surface patterns in complex datasets, enable precise targeting and prioritization use cases, and inform if not automate decisions. The possibilities and shortcomings have not been documented and illustrated for a public and social sector audience. This textbook introduces aspiring and veteran public servants to core concepts and aims to be the springboard for building data science capacity.

This book has been shaped by a number of brilliant and generous people. Dr. Kyle Bradbury, Dr. Ching-Ling Chen, Dr. Ying-Chih Chen, Christopher Eshleman, Dr. Tyrone W. A. Grandison, Aya Hamano, Annabel Jouard, Artem Kopelev, Dr. Jacob Model, Wade Petty, Alice Ramey, Robin Thottungal, and Levi Weible read many chapters of our text, suggesting ways to simplify complex ideas into a form that is appropriate for policy and strategy audiences. We are incredibly grateful for their contributions. Lastly, our formative experiences in government were enriched by dedicated experts, namely, Jeffrey Tyrens (New York City Mayor's Office of Operations), Jeffrey Roth (New York City Fire Department), Dr. Howard Friedman (United Nations Population Fund), Dr. Curt Tilmes (NASA—Goddard Space Flight Center), Dr. Dennis Fixler, and Dr. Benjamin Bridgman (U.S. Bureau of Economic Analysis).

Cambridge, UK — Jeffrey C. Chen

Eugene, USA — Edward A. Rubin

Suitland, USA — Gary J. Cornwall

Contents