

---

# Springer Series in the Data Sciences

## Series Editors

David Banks, Duke University, Durham, NC, USA

Jianqing Fan, Department of Financial Engineering, Princeton University, Princeton, NJ, USA

Michael Jordan, University of California, Berkeley, CA, USA

Ravi Kannan, Microsoft Research Labs, Bangalore, India

Yurii Nesterov, CORE, Universite Catholique de Louvain, Louvain-la-Neuve, Belgium

Christopher Ré, Department of Computer Science, Stanford University, Stanford, USA

Ryan J. Tibshirani, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, USA

Larry Wasserman, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, USA

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

© Springer Nature Switzerland AG 2021

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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

Eugene, USA

Suitland, USA

Jeffrey C. Chen

Edward A. Rubin

Gary J. Cornwall

# Contents

<b>Preface</b>	vii
<b>1 An Introduction</b>	<b>1</b>
1.1 Why we wrote this book	2
1.2 What we assume	2
1.3 How this book is structured	3
<b>2 The Case for Programming</b>	<b>5</b>
2.1 Doing visual analytics since the 1780s	5
2.2 How does programming work?	7
2.3 Setting up R and RStudio	8
2.3.1 Installing R	8
2.3.2 Installing RStudio	9
2.3.3 DIY: Running your first code snippet	10
2.4 Making the case for open-source software	11
<b>3 Elements of Programming</b>	<b>13</b>
3.1 Data are everywhere	13
3.2 Data types	14
3.2.1 numeric	14
3.2.2 character	14
3.2.3 logical	14
3.2.4 factor	16
3.2.5 date	16
3.2.6 The class function	16
3.3 Objects in R	17
3.4 R's object classes	18
3.4.1 vector	18
3.4.2 matrix	18
3.4.3 data.frame	19
3.4.4 list	20
3.4.5 The class function, v2	21
3.4.6 More classes	22
3.5 Packages	22
3.5.1 Base R and the need to extend functionality	22
3.5.2 Installing packages	22
3.5.3 Loading packages	23
3.5.4 Package management and pacman	23
3.6 Data input/output	24
3.6.1 Directories	24

3.6.2	Load functions	26
3.6.3	Datasets	26
3.7	Finding help	27
3.7.1	Help function	27
3.7.2	Google and online communities	27
3.8	Beyond this chapter	27
3.8.1	Best practices	27
3.8.2	Further study	28
3.9	DIY: Loading solar energy data from the web	29
<b>4</b>	<b>Transforming Data</b>	<b>33</b>
4.1	Importing and assembling data	34
4.1.1	Loading files	35
4.2	Manipulating values	38
4.2.1	Text manipulation functions	39
4.2.2	Regular Expressions (RegEx)	40
4.2.3	DIY: Working with PII	43
4.2.4	Working with dates	44
4.3	The structure of data	45
4.3.1	Matrix or data frame?	45
4.3.2	Array indexes	45
4.3.3	Subsetting	46
4.3.4	Sorting and re-ordering	47
4.3.5	Aggregating data	48
4.3.6	Reshaping data	49
4.4	Control structures	51
4.4.1	If statement	52
4.4.2	For-loops	53
4.4.3	While	55
4.5	Functions	56
4.6	Beyond this chapter	57
4.6.1	Best practices	57
4.6.2	Further study	58
<b>5</b>	<b>Record Linkage</b>	<b>61</b>
5.1	Edward Kennedy, Bill de Blasio, and Bayerische Motoren Werke	61
5.2	How does record linkage work?	62
5.3	Pre-processing the data	63
5.4	De-duplication	66
5.5	Deterministic record linkage	67
5.6	Comparison functions	70
5.6.1	Edit distances	70
5.6.2	Phonetic algorithms	71
5.6.3	New tricks, same heuristics	73
5.7	Probabilistic record linkage	74
5.8	Data privacy	76
5.9	DIY: Matching people in the UK-UN sanction lists	77
5.10	Beyond this chapter	80
5.10.1	Best practices	80
5.10.2	Further study	81
<b>6</b>	<b>Exploratory Data Analysis</b>	<b>83</b>
6.1	Visually detecting patterns	83
6.2	The gist of EDA	85



6.3	Visualizing distributions . . . . .	87
6.3.1	Skewed variables . . . . .	92
6.4	Exploring missing values . . . . .	94
6.4.1	Encodings . . . . .	94
6.4.2	Missing value functions . . . . .	95
6.4.3	Exploring missingness . . . . .	96
6.4.4	Treating missingness . . . . .	98
6.5	Analyzing time series . . . . .	103
6.6	Finding visual correlations . . . . .	105
6.6.1	Visual analysis on high-dimensional datasets . . . . .	108
6.7	Beyond this chapter . . . . .	109
<b>7</b>	<b>Regression Analysis</b>	<b>113</b>
7.1	Measuring and predicting the preferences of society . . . . .	113
7.2	Simple linear regression . . . . .	114
7.2.1	Mean squared error . . . . .	116
7.2.2	Ordinary least squares . . . . .	117
7.2.3	DIY: A simple hedonic model . . . . .	118
7.3	Checking for linearity . . . . .	121
7.4	Multiple regression . . . . .	123
7.4.1	Non-linearities . . . . .	124
7.4.2	Discrete variables . . . . .	125
7.4.3	Discontinuities . . . . .	127
7.4.4	Measures of model fitness . . . . .	128
7.4.5	DIY: Choosing between models . . . . .	129
7.4.6	DIY: Housing prices over time . . . . .	132
7.5	Beyond this chapter . . . . .	137
<b>8</b>	<b>Framing Classification</b>	<b>139</b>
8.1	Playing with fire . . . . .	139
8.1.1	FireCast . . . . .	139
8.1.2	What’s a classifier? . . . . .	140
8.2	The basics of classifiers . . . . .	141
8.2.1	The anatomy of a classifier . . . . .	141
8.2.2	Finding signal in classification contexts . . . . .	142
8.2.3	Measuring accuracy . . . . .	142
8.3	Logistic regression . . . . .	146
8.3.1	The social science workhorse . . . . .	146
8.3.2	Telling the story from coefficients . . . . .	147
8.3.3	How are coefficients learned? . . . . .	148
8.3.4	In practice . . . . .	148
8.3.5	DIY: Expanding health care coverage . . . . .	150
8.4	Regularized regression . . . . .	156
8.4.1	From regularization to interpretation . . . . .	158
8.4.2	DIY: Re-visiting health care coverage . . . . .	158
8.5	Beyond this chapter . . . . .	161
<b>9</b>	<b>Three Quantitative Perspectives</b>	<b>163</b>
9.1	Descriptive analysis . . . . .	164
9.2	Causal inference . . . . .	165
9.2.1	Potential outcomes framework . . . . .	166
9.2.2	Regression discontinuity . . . . .	167
9.2.3	Difference-in-differences . . . . .	172

9.3	Prediction . . . . .	174
9.3.1	Understanding accuracy . . . . .	175
9.3.2	Model validation . . . . .	180
9.4	Beyond this chapter . . . . .	182
<b>10</b>	<b>Prediction</b> . . . . .	<b>185</b>
10.1	The role of algorithms . . . . .	185
10.2	Data science pipelines . . . . .	187
10.3	K-Nearest Neighbors ( <i>k</i> -NN) . . . . .	189
10.3.1	Under the hood . . . . .	190
10.3.2	DIY: Predicting the extent of storm damage . . . . .	192
10.4	Tree-based learning . . . . .	195
10.4.1	Classification and Regression Trees (CART) . . . . .	196
10.4.2	Random forests . . . . .	201
10.4.3	In practice . . . . .	203
10.4.4	DIY: Wage prediction with CART and random forests . . . . .	204
10.5	An introduction to other algorithms . . . . .	210
10.5.1	Gradient boosting . . . . .	211
10.5.2	Neural networks . . . . .	212
10.6	Beyond this chapter . . . . .	215
<b>11</b>	<b>Cluster Analysis</b> . . . . .	<b>217</b>
11.1	Things closer together are more related . . . . .	217
11.2	Foundational concepts . . . . .	218
11.3	<i>k</i> -means . . . . .	219
11.3.1	Under the hood . . . . .	219
11.3.2	In Practice . . . . .	221
11.3.3	DIY: Clustering for economic development . . . . .	223
11.4	Hierarchical clustering . . . . .	226
11.4.1	Under the hood . . . . .	227
11.4.2	In Practice . . . . .	229
11.4.3	DIY: Clustering time series . . . . .	230
11.5	Beyond this chapter . . . . .	234
<b>12</b>	<b>Spatial Data</b> . . . . .	<b>237</b>
12.1	Anticipating climate impacts . . . . .	237
12.2	Classes of spatial data . . . . .	239
12.3	Rasters . . . . .	239
12.3.1	Raster files . . . . .	241
12.3.2	Rasters and math . . . . .	242
12.3.3	DIY: Working with raster math . . . . .	242
12.4	Vectors . . . . .	244
12.4.1	Vector files . . . . .	244
12.4.2	Converting points to spatial objects . . . . .	245
12.4.3	Coordinate Reference Systems . . . . .	246
12.4.4	DIY: Converting coordinates into point vectors . . . . .	248
12.4.5	Reading shapefiles . . . . .	249
12.4.6	Spatial joins . . . . .	250
12.4.7	DIY: Analyzing spatial relationships . . . . .	252
12.5	Beyond this chapter . . . . .	256

<b>13</b>	<b>Natural Language</b>	<b>259</b>
13.1	Transforming text into data	260
13.1.1	Processing textual data	260
13.1.2	TF-IDF	262
13.1.3	Document similarities	263
13.1.4	DIY: Basic text processing	263
13.2	Sentiment Analysis	266
13.2.1	Sentiment lexicons	267
13.2.2	Calculating sentiment scores	267
13.2.3	DIY: Scoring text for sentiment	269
13.3	Topic modeling	271
13.3.1	A conceptual base	271
13.3.2	How do topics models work?	272
13.3.3	DIY: Finding topics in presidential speeches	273
13.4	Beyond this chapter	280
13.4.1	Best practices	280
13.4.2	Further study	281
<b>14</b>	<b>The Ethics of Data Science</b>	<b>283</b>
14.1	An emerging debate	283
14.2	Bias	284
14.2.1	Sampling bias	285
14.2.2	Measurement bias	287
14.2.3	Prejudicial bias	289
14.3	Fairness	289
14.3.1	Score-based fairness	290
14.3.2	Accuracy-based fairness	290
14.3.3	Other considerations	291
14.4	Transparency and Interpretability	291
14.4.1	Interpretability	292
14.4.2	Explainability	293
14.5	Privacy	295
14.5.1	An evolving landscape	295
14.5.2	Privacy strategies	295
14.6	Beyond this chapter	297
<b>15</b>	<b>Developing Data Products</b>	<b>299</b>
15.1	Meeting people where they are	299
15.2	Designing for impact	301
15.2.1	Identify a user need	301
15.2.2	Size up the situation	302
15.2.3	Build a lean “V1”	303
15.2.4	Test and evaluate its impact, then iterate	303
15.3	Communicating data science projects	304
15.3.1	Presentations	304
15.3.2	Written reports	306
15.4	Reporting dashboards	308
15.5	Prediction products	311
15.5.1	Prioritization and targeting lists	311
15.5.2	Scoring engines	311
15.6	Continuing to hone your craft	313
15.7	Where to next?	315

<b>16 Building Data Teams</b>	<b>317</b>
16.1 Establishing a baseline	317
16.2 Operating models	320
16.2.1 Center of excellence	320
16.2.2 Hack teams	321
16.2.3 Consultancy	323
16.2.4 Matrix organizations	324
16.3 Identifying roles	326
16.3.1 The manager	326
16.3.2 Analytics roles	326
16.3.3 Data product roles	327
16.3.4 Titles in the civil service system	328
16.4 The hiring process	328
16.4.1 Job postings and application review	328
16.4.2 Interviews	329
16.5 Final thoughts	330
 <b>Appendix A: Planning a Data Product</b>	 <b>331</b>
Key Questions	331
 <b>Appendix B: Interview Questions</b>	 <b>335</b>
Getting to know the candidate	335
Business acumen	335
Project experience	335
Whiteboard questions	336
Statistics	336
Causal inference	337
Estimation versus prediction	337
Machine learning	338
Model evaluation	339
Communication and visualization	339
Programming	340
Take-home questions	341
 <b>References</b>	 <b>343</b>
 <b>Index</b>	 <b>357</b>