

Chapman & Hall/CRC
Statistics in the Social and Behavioral Sciences Series

BIG DATA AND SOCIAL SCIENCE

**Data Science Methods and Tools
for Research and Practice**

SECOND EDITION



Edited by

**Ian Foster, Rayid Ghani,
Ron S. Jarmin, Frauke Kreuter,
and Julia Lane**



CRC Press
Taylor & Francis Group

A CHAPMAN & HALL BOOK

BIG DATA AND SOCIAL SCIENCE

Chapman & Hall/CRC

Statistics in the Social and Behavioral Sciences Series

Series Editors

Jeff Gill, Steven Heeringa, Wim J. van der Linden, Tom Snijders

Recently Published Titles

Multilevel Modelling Using Mplus

Holmes Finch and Jocelyn Bolin

Applied Survey Data Analysis, Second Edition

Steven G. Heeringa, Brady T. West, and Patricia A. Berglund

Adaptive Survey Design

Barry Schouten, Andy Peytchev, and James Wagner

Handbook of Item Response Theory, Volume One: Models

Wim J. van der Linden

Handbook of Item Response Theory, Volume Two: Statistical Tools

Wim J. van der Linden

Handbook of Item Response Theory, Volume Three: Applications

Wim J. van der Linden

Bayesian Demographic Estimation and Forecasting

John Bryant and Junni L. Zhang

Multivariate Analysis in the Behavioral Sciences, Second Edition

Kimmo Vehkalahti and Brian S. Everitt

Analysis of Integrated Data

Li-Chun Zhang and Raymond L. Chambers

Multilevel Modeling Using R, Second Edition

W. Holmes Finch, Joselyn E. Bolin, and Ken Kelley

Modelling Spatial and Spatial-Temporal Data: A Bayesian Approach

Robert Haining and Guangquan Li

Measurement Models for Psychological Attributes

Klaas Sijtsma and Andries van der Ark

Handbook of Automated Scoring: Theory into Practice

Duanli Yan, André A. Rupp, and Peter W. Foltz

Interviewer Effects from a Total Survey Error Perspective

Kristen Olson, Jolene D. Smyth, Jennifer Dykema, Allyson Holbrook, Frauke Kreuter, and Brady T. West

Statistics and Elections: Polling, Prediction, and Testing

Ole J. Forsberg

Big Data and Social Science: Data Science Methods and Tools for Research and Practice, Second Edition

Ian Foster, Rayid Ghani, Ron S. Jarmin, Frauke Kreuter and Julia Lane

Analyzing Spatial Models of Choice and Judgment, Second Edition

David A. Armstrong II, Ryan Bakker, Royce Carroll, Christopher Hare, Keith T. Poole and Howard Rosenthal

For more information about this series, please visit: <https://www.routledge.com/Chapman--HallCRC-Statistics-in-the-Social-and-Behavioral-Sciences/book-series/CHSTSOBESCI>

Chapman & Hall/CRC
Statistics in the Social and Behavioral Sciences Series

BIG DATA AND SOCIAL SCIENCE

**Data Science Methods and Tools for
Research and Practice**

Second Edition

Edited by

Ian Foster

**University of Chicago
Argonne National Laboratory**

Rayid Ghani

University of Chicago

Ron S. Jarmin

U.S. Census Bureau

Frauke Kreuter

**University of Maryland
University of Mannheim
Institute for Employment Research**

Julia Lane

**New York University
American Institutes for Research**



CRC Press

Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

Second edition published 2021
by CRC Press
6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742

and by CRC Press
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

© 2021 Taylor & Francis Group, LLC

First edition published by CRC Press 2016

CRC Press is an imprint of Taylor & Francis Group, LLC

Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, access www.copyright.com or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. For works that are not available on CCC please contact mpkbookspermissions@tandf.co.uk

Trademark notice: Product or corporate names may be trademarks or registered trademarks and are used only for identification and explanation without intent to infringe.

ISBN: 9780367341879 (hbk)
ISBN: 9780367568597 (pbk)
ISBN: 9780429324383 (ebk)

Typeset in Kerkis
by Nova Techset Private Limited, Bengaluru & Chennai, India

Contents

Preface	xv
Editors	xvii
Contributors	xix
1 Introduction	1
1.1 Why this book?	1
1.2 Defining big data and its value	2
1.3 The importance of inference	4
1.3.1 Description	5
1.3.2 Causation	6
1.3.3 Prediction	7
1.4 The importance of understanding how data are generated	7
1.5 New tools for new data	9
1.6 The book's "use case"	10
1.7 The structure of the book	15
1.7.1 Part I: Capture and curation	15
1.7.2 Part II: Modeling and analysis	17
1.7.3 Part III: Inference and ethics	19
1.8 Resources	20
Part I Capture and Curation	23
2 Working with Web Data and APIs	25
Cameron Neylon	
2.1 Introduction	25
2.2 Scraping information from the web	27
2.2.1 Obtaining data from websites	27
2.2.1.1 Constructing the URL	28
2.2.1.2 Obtaining the contents of the page from the URL	28
2.2.1.3 Processing the HTML response	29

2.2.2	Programmatically iterating over the search results	33
2.2.3	Limits of scraping	34
2.3	Application programming interfaces	35
2.3.1	Relevant APIs and resources	35
2.3.2	RESTful APIs, returned data, and Python wrappers	35
2.4	Using an API	37
2.5	Another example: Using the ORCID API via a wrapper	39
2.6	Integrating data from multiple sources	40
2.7	Summary	41
3	Record Linkage	43
	<i>Joshua Tokle and Stefan Bender</i>	
3.1	Motivation	43
3.2	Introduction to record linkage	44
3.3	Preprocessing data for record linkage	49
3.4	Indexing and blocking	51
3.5	Matching	53
3.5.1	Rule-based approaches	54
3.5.2	Probabilistic record linkage	55
3.5.3	Machine learning approaches to record linkage	57
3.5.4	Disambiguating networks	60
3.6	Classification	60
3.6.1	Thresholds	61
3.6.2	One-to-one links	62
3.7	Record linkage and data protection	63
3.8	Summary	64
3.9	Resources	64
4	Databases	67
	<i>Ian Foster and Pascal Heus</i>	
4.1	Introduction	67
4.2	The DBMS: When and why	68
4.3	Relational DBMSs	74
4.3.1	Structured Query Language	76
4.3.2	Manipulating and querying data	76
4.3.3	Schema design and definition	79
4.3.4	Loading data	82
4.3.5	Transactions and crash recovery	83
4.3.6	Database optimizations	84
4.3.7	Caveats and challenges	87
4.3.7.1	Data cleaning	87
4.3.7.2	Missing values	87
4.3.7.3	Metadata for categorical variables	87

4.4	Linking DBMSs and other tools	88
4.5	NoSQL databases	91
4.5.1	Challenges of scale: The CAP theorem	91
4.5.2	NoSQL and key-value stores	92
4.5.3	Other NoSQL databases	94
4.6	Spatial databases	95
4.7	Which database to use?	97
4.7.1	Relational DBMSs	97
4.7.2	NoSQL DBMSs	98
4.8	Summary	98
4.9	Resources	99
5	Scaling up through Parallel and Distributed Computing	101
	<i>Huy Vo and Claudio Silva</i>	
5.1	Introduction	101
5.2	MapReduce	103
5.3	Apache Hadoop MapReduce	105
5.3.1	The Hadoop Distributed File System	105
5.3.2	Hadoop setup: Bringing compute to the data	106
5.3.3	Hardware provisioning	108
5.3.4	Programming in Hadoop	109
5.3.5	Programming language support	111
5.3.6	Benefits and limitations of Hadoop	112
5.4	Other MapReduce Implementations	113
5.5	Apache Spark	114
5.6	Summary	116
5.7	Resources	117
Part II	Modeling and Analysis	119
6	Information Visualization	121
	<i>M. Adil Yalçın and Catherine Plaisant</i>	
6.1	Introduction	121
6.2	Developing effective visualizations	122
6.3	A data-by-tasks taxonomy	127
6.3.1	Multivariate data	129
6.3.2	Spatial data	130
6.3.3	Temporal data	131
6.3.4	Hierarchical data	133
6.3.5	Network data	134
6.3.6	Text data	136

6.4	Challenges	138
6.4.1	Scalability	138
6.4.2	Evaluation	139
6.4.3	Visual impairment	140
6.4.4	Visual literacy	140
6.5	Summary	141
6.6	Resources	141
7	Machine Learning	143
	<i>Rayid Ghani and Malte Schierholz</i>	
7.1	Introduction	143
7.2	What is machine learning?	144
7.3	Types of analysis	147
7.4	The machine learning process	147
7.5	Problem formulation: Mapping a problem to machine learning methods	150
7.5.1	Features	151
7.6	Methods	153
7.6.1	Unsupervised learning methods	154
7.6.1.1	Clustering	154
7.6.1.2	The k -means clustering	155
7.6.1.3	Expectation-maximization clustering	157
7.6.1.4	Mean shift clustering	157
7.6.1.5	Hierarchical clustering	158
7.6.1.6	Spectral clustering	158
7.6.1.7	Principal components analysis	160
7.6.1.8	Association rules	160
7.6.2	Supervised learning	161
7.6.2.1	Training a model	163
7.6.2.2	Using the model to score new data	163
7.6.2.3	The k -nearest neighbor	163
7.6.2.4	Support vector machines	165
7.6.2.5	Decision trees	167
7.6.2.6	Ensemble methods	169
7.6.2.7	Bagging	169
7.6.2.8	Boosting	170
7.6.2.9	Random forests	171
7.6.2.10	Stacking	172
7.6.2.11	Neural networks and deep learning	172
7.6.3	Binary vs. multiclass classification problems	174
7.6.4	Skewed or imbalanced classification problems	175
7.6.5	Model interpretability	176
7.6.5.1	Global vs. individual-level explanations	176

7.7	Evaluation	178
7.7.1	Methodology	178
7.7.1.1	In-sample evaluation	178
7.7.1.2	Out-of-sample and holdout set	179
7.7.1.3	Cross-validation	179
7.7.1.4	Temporal validation	180
7.7.2	Metrics	181
7.8	Practical tips	185
7.8.1	Avoiding leakage	185
7.8.2	Machine learning pipeline	187
7.9	How can social scientists benefit from machine learning?	187
7.10	Advanced topics	189
7.11	Summary	191
7.12	Resources	191
8	Text Analysis	193
	<i>Evgeny Klochikhin and Jordan Boyd-Graber</i>	
8.1	Understanding human-generated text	193
8.2	How is text data different than “structured” data?	194
8.3	What can we do with text data?	194
8.4	How to analyze text	196
8.4.1	Initial processing	197
8.4.1.1	Tokenization	197
8.4.1.2	Stop words	198
8.4.1.3	The N -grams	198
8.4.1.4	Stemming and lemmatization	199
8.4.2	Linguistic analysis	199
8.4.2.1	Part-of-speech tagging	199
8.4.2.2	Order matters	200
8.4.3	Turning text data into a matrix: How much is a word worth?	200
8.4.4	Analysis	201
8.4.4.1	Use case: Finding similar documents	202
8.4.4.2	Example: Measuring similarity between documents	203
8.4.4.3	Example code	203
8.4.4.4	Augmenting similarity calculations with external knowledge repositories	203
8.4.4.5	Evaluating “find similar” methods	205
8.4.4.6	The F score	206
8.4.4.7	Examples	206
8.4.4.8	Use case: Clustering	206
8.4.5	Topic modeling	208
8.4.5.1	Inferring “topics” from raw text	209
8.4.5.2	Applications of topic models	212

8.5	Word embeddings and deep learning	214
8.6	Text analysis tools	215
8.6.1	The natural language toolkit	216
8.6.2	Stanford CoreNLP	216
8.6.3	The MALLET	217
8.6.3.1	Spacy.io	217
8.6.3.2	Pytorch	217
8.7	Summary	218
8.8	Resources	218
9	Networks: The Basics	221
	<i>Jason Owen-Smith</i>	
9.1	Introduction	221
9.2	What are networks?	222
9.3	Structure for this chapter	224
9.4	Turning data into a network	224
9.4.1	Types of networks	225
9.4.2	Inducing one-mode networks from two-mode data	227
9.5	Network measures	230
9.5.1	Reachability	231
9.5.2	Whole-network measures	232
9.5.2.1	Components and reachability	232
9.5.2.2	Path length	233
9.5.2.3	Degree distribution	236
9.5.2.4	Clustering coefficient	236
9.5.2.5	Centrality measures	238
9.6	Case study: Comparing collaboration networks	241
9.7	Summary	246
9.8	Resources	246
	Part III Inference and Ethics	249
10	Data Quality and Inference Errors	251
	<i>Paul P. Biemer</i>	
10.1	Introduction	251
10.2	The total error paradigm	252
10.2.1	The traditional model	253
10.2.1.1	Types of errors	254
10.2.1.2	Column error	257
10.2.1.3	Cell errors	258
10.3	Example: Google Flu Trends	260

10.4	Errors in data analysis	261
10.4.1	Analysis errors despite accurate data	261
10.4.2	Noise accumulation	262
10.4.3	Spurious correlations	262
10.4.4	Incidental endogeneity	263
10.4.5	Analysis errors resulting from inaccurate data	264
10.4.5.1	Variable (uncorrelated) and correlated error in continuous variables	264
10.4.5.2	Extending variable and correlated error to categorical data	266
10.4.5.3	Errors when analyzing rare population groups	267
10.4.5.4	Errors in correlation analysis	269
10.4.5.5	Errors in regression analysis	273
10.5	Detecting and compensating for data errors	275
10.5.1	TablePlots	276
10.6	Summary	279
10.7	Resources	280
11	Bias and Fairness	281
	<i>Kit T. Rodolfa, Pedro Saleiro, and Rayid Ghani</i>	
11.1	Introduction	281
11.2	Sources of bias	282
11.2.1	Sample bias	282
11.2.2	Label (outcome) bias	283
11.2.3	Machine learning pipeline bias	283
11.2.4	Application bias	285
11.2.5	Considering bias when deploying your model	286
11.3	Dealing with bias	286
11.3.1	Define bias	286
11.3.2	Definitions	287
11.3.3	Choosing bias metrics	289
11.3.4	Punitive example	291
11.3.4.1	Count of false positives	291
11.3.4.2	Group size-adjusted false positives	291
11.3.4.3	False discovery rate	292
11.3.4.4	False positive rate	292
11.3.4.5	Tradeoffs in metric choice	292
11.3.5	Assistive example	294
11.3.5.1	Count of false negatives	294
11.3.5.2	Group size-adjusted false negatives	294
11.3.5.3	False omission rate	295
11.3.5.4	False negative rate	295
11.3.6	Special case: Resource-constrained programs	296

11.4	Mitigating bias	296
11.4.1	Auditing model results	297
11.4.2	Model selection	297
11.4.3	Other options for mitigating bias	299
11.5	Further considerations	300
11.5.1	Compared to what?	300
11.5.2	Costs to both errors	301
11.5.3	What is the relevant population?	301
11.5.4	Continuous outcomes	302
11.5.5	Considerations for ongoing measurement	302
11.5.6	Equity in practice	303
11.5.7	Additional terms for metrics	304
11.6	Case studies	305
11.6.1	Recidivism predictions with COMPAS	306
11.6.2	Facial recognition	307
11.6.3	Facebook advertisement targeting	309
11.7	Aequitas: A toolkit for auditing bias and fairness in machine learning models	310
11.7.1	Aequitas in the larger context of the machine learning pipeline.	311
12	Privacy and Confidentiality	313
	<i>Stefan Bender, Ron S. Jarmin, Frauke Kreuter, and Julia Lane</i>	
12.1	Introduction	313
12.2	Why is access important?	319
12.2.1	Validating the data-generating process	319
12.2.2	Replication	320
12.2.3	Building knowledge infrastructure	320
12.3	Providing access	321
12.3.1	Statistical disclosure control techniques	321
12.3.2	Research data centers	323
12.4	Non-tabular data	323
12.5	The new challenges	326
12.6	Legal and ethical framework	328
12.7	Summary	329
12.8	Resources	331
13	Workbooks	333
	<i>Brian Kim, Christoph Kern, Jonathan Scott Morgan, Clayton Hunter, and Avishek Kumar</i>	
13.1	Introduction	333
13.2	Notebooks	334
13.2.1	Databases	334
13.2.2	Dataset Exploration and Visualization	334

13.2.3 APIs	335
13.2.4 Record Linkage	335
13.2.5 Text Analysis	336
13.2.6 Networks	336
13.2.7 Machine Learning—Creating Labels	336
13.2.8 Machine Learning—Creating Features	337
13.2.9 Machine Learning—Model Training and Evaluation	337
13.2.10 Bias and Fairness	337
13.2.11 Errors and Inference	338
13.2.12 Additional workbooks	338
13.3 Resources	338
Bibliography	341
Index	381



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Preface

The class on which this book is based was created in response to a very real challenge: how to introduce new ideas and methodologies about economic and social measurement into a workplace focused on producing high-quality statistics. Since the publication of the first edition, we have been fortunate to train more than 450 participants in the Applied Data Analytics classes, resulting in increased data analytics capacity, in terms of both human and technical resources. What we have learned in delivering these classes has greatly influenced the second edition. We also have added a new chapter on Bias and Fairness in Machine Learning as well as reorganized some of the chapters.

As with any book, there are many people to be thanked. The Coleridge Initiative team at New York University, the University of Maryland, and the University of Chicago have been critical in shaping the format and structure—we are particularly grateful to Clayton Hunter, Jody Derezinski Williams, Graham Henke, Jonathan Morgan, Drew Gordon, Avishek Kumar, Brian Kim, Christoph Kern, and all of the book chapter authors for their contributions to the second edition.

We also thank the critical reviewers solicited from CRC Press and everyone from whom we received revision suggestions online, in particular Stas Kolenikov, who carefully examined the first edition and suggested updates. We owe a great debt of gratitude to the copyeditor, Anna Stamm; the project manager, Arun Kumar; the editorial assistant, Vaishali Singh; the project editor, Iris Fahrner; and the publisher, Rob Calver, for their hard work and dedication.



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Editors

Ian Foster, PhD, is a professor of computer science at the University of Chicago as well as a senior scientist and distinguished fellow at Argonne National Laboratory. His research addresses innovative applications of distributed, parallel, and data-intensive computing technologies to scientific problems in such domains as climate change and biomedicine. Methods and software developed under his leadership underpin many large national and international cyberinfrastructures. He is a fellow of the American Association for the Advancement of Science, the Association for Computing Machinery, and the British Computer Society. He earned a PhD in computer science from Imperial College London.



Prof. Rayid Ghani is a professor in the Machine Learning Department (in the School of Computer Science) and the Heinz College of Information Systems and Public Policy at Carnegie Mellon University. His research focuses on developing and using Machine Learning, AI, and Data Science methods for solving high impact social good and public policy problems in a fair and equitable way across criminal justice, education, healthcare, energy, transportation, economic development, workforce development and public safety. He is also the founder and director of the “Data Science for Social Good” summer program for aspiring data scientists to work on data mining, machine learning, big data, and data science projects with social impact. Previously Prof. Ghani was a faculty member at the University of Chicago, and prior to that, served as the Chief Scientist for Obama for America (Obama 2012 Campaign).





Ron S. Jarmin, PhD, is the deputy director at the U.S. Census Bureau. He earned a PhD in economics from the University of Oregon and has published in the areas of industrial organization, business dynamics, entrepreneurship, technology and firm performance, urban economics, Big Data, data access and statistical disclosure avoidance. He oversees the Census Bureau's large portfolio of data collection, research and dissemination activities for critical economic and social statistics including the 2020 Decennial Census of Population and Housing.



Frauke Kreuter, PhD, is a professor at the University of Maryland in the Joint Program in Survey Methodology, professor of Statistics and Methodology at the University of Mannheim and head of the Statistical Methods group at the Institute for Employment Research in Nuremberg, Germany. She is the founder of the International Program in Survey and Data Science, co-founder of the Coleridge Initiative, fellow of the American Statistical Association (ASA), and recipient of the WSS Cox and the ASA Links Lecture Awards. Her research focuses on data quality, privacy, and the effects of bias in data collection on statistical estimates and algorithmic fairness.



Julia Lane, PhD, is a professor at the NYU Wagner Graduate School of Public Service. She is also an NYU Provostial Fellow for Innovation Analytics. She co-founded the Coleridge Initiative as well as UMETRICS and STAR METRICS programs at the National Science Foundation, established a data enclave at NORC/University of Chicago, and co-founded the Longitudinal Employer-Household Dynamics Program at the U.S. Census Bureau and the Linked Employer Employee Database at Statistics New Zealand. She is the author/editor of 10 books and the author of more than 70 articles in leading journals, including *Nature* and *Science*. She is an elected fellow of the American Association for the Advancement of Science and a fellow of the American Statistical Association.

Contributors

Stefan Bender

Deutsche Bundesbank
Frankfurt, Germany

Paul P. Biemer

RTI International
Raleigh, NC, USA
University of North Carolina
Chapel Hill, NC, USA

Jordan Boyd-Graber

University of Maryland
College Park, MD, USA

Pascal Heus

Metadata Technology North America
Knoxville, TN, USA

Clayton Hunter

New York University
New York, NY, USA

Christoph Kern

University of Mannheim
Mannheim, Germany

Brian Kim

University of Maryland
College Park, MD, USA

Evgeny Klochikhin

Parkofon Inc.
Washington, DC, USA

Avishek Kumar

Intuit AI

Cameron Neylon

Curtin University
Perth, Australia

Jason Owen-Smith

University of Michigan
Ann Arbor, MI, USA

Catherine Plaisant

University of Maryland
College Park, MD, USA

Kit T. Rodolfa

Carnegie Mellon University
Pittsburgh, PA, USA

Pedro Saleiro

Feedzai
California, USA

Malte Schierholz

Institute for Employment Research (IAB)
Nuremberg, Germany

Jonathan Scott Morgan

University of Mannheim
Mannheim, Germany

Claudio Silva

New York University
New York, NY, USA

Joshua Tokle

Amazon
Seattle, WA, USA

Huy Vo

City University of New York
New York, NY, USA

M. Adil Yalçın

Keshif
Washington, DC, USA

Chapter 1

Introduction

1.1 Why this book?

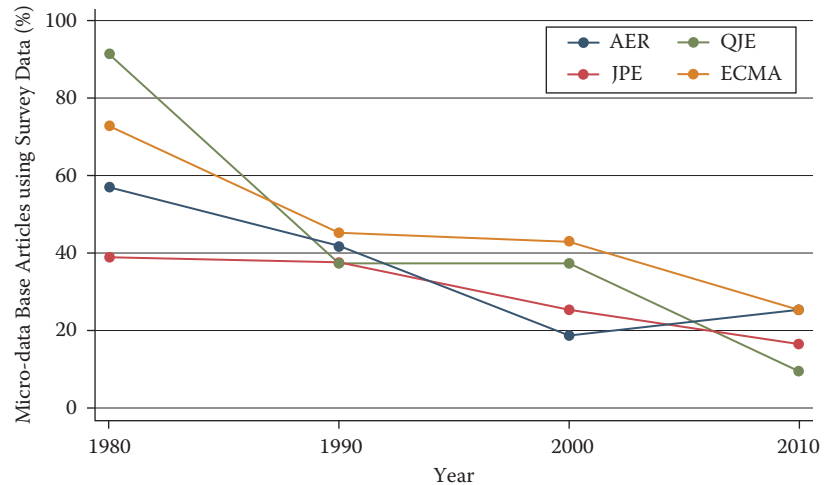
The world has changed for empirical social scientists. The new types of “big data” have generated an entire new research field—that of data science. That world is dominated by computer scientists who have generated new ways of creating and collecting data, developed new analytical techniques, and provided new ways of visualizing and presenting information. The results have been to change the nature of the work that social scientists do.

Social scientists have been enthusiastic in responding to the new opportunity. Python and R are becoming as, and hopefully more, well-known as SAS and Stata—indeed, the 2018 Nobel Laureate in Economics, Paul Romer, is a Python convert (Kopf, 2018). Research also has changed. Researchers draw on data that are “found” rather than “made” by federal agencies; those publishing in leading academic journals are much less likely today to draw on preprocessed survey data (Figure 1.1). Social science workflows can become more automated, replicable, and reproducible (Yarkoni et al., 2019).

Policy also has changed. The Foundations of Evidence-based Policy Act, which was signed into law in 2019, requires agencies to utilize evidence and data in making policy decisions (Hart, 2019). The Act, together with the Federal Data Strategy (Office of Management and Budget, 2019), establishes both Chief Data Officers to oversee the collection, use of, and access to many new types of data and a learning agenda to build the data science capacity of agency staff.

In addition, the jobs have changed. The new job title of “data scientist” is highlighted in job advertisements on CareerBuilder.com and Burningglass—supplanting the demand for statisticians, economists, and other quantitative social scientists if starting salaries are useful indicators. At the federal level, the Office of Personnel Management has created a new data scientist job title.

The goal of this book is to provide social scientists with an understanding of the key elements of this new science, the value of the



Note: “Pre-existing survey” data sets refer to micro surveys such as the CPS or SIPP and do not include surveys designed by researchers for their study. Sample excludes studies whose primary data source is from developing countries.

Figure 1.1. Use of pre-existing survey data in publications in leading journals, 1980–2010 (Chetty, 2012)

tools, and the opportunities for doing better work. The goal is also to identify the many ways in which the analytical toolkits possessed by social scientists can enhance the generalizability and usefulness of the work done by computer scientists.

We take a pragmatic approach, drawn on our experience of working with data to tackle a wide variety of policy problems. Most social scientists set out to solve a real world social or economic problem: they frame the problem, identify the data, conduct the analysis, and then draw inferences. At all points, of course, the social scientist needs to consider the ethical ramifications of their work, particularly respecting privacy and confidentiality. The book follows the same structure. We chose a particular problem—the link between research investments and innovation—because that is a major social science policy issue, and one in which social scientists have been addressing the use of big data techniques.

1.2 Defining big data and its value

There are almost as many definitions of big data as there are new types of data. One approach is to define big data as [anything too big](#)

to fit onto your computer. Another approach is to define it as data with high volume, high velocity, and great variety. We choose the description adopted by the American Association of Public Opinion Research: “The term ‘Big Data’ is an imprecise description of a rich and complicated set of characteristics, practices, techniques, ethical issues, and outcomes all associated with data” (Japiec et al., 2015).

► This topic will be discussed in more detail in [Chapter 5](#).

The value of the new types of data for social science is quite substantial. Personal data have been hailed as the “new oil” of the 21st century (Greenwood et al., 2014). Policymakers have found that detailed data on human beings can be used to reduce crime (Lynch, 2018), improve health delivery (Pan et al., 2017), and better manage cities (Glaeser, 2019). Society can gain as well—much cited work shows data-driven businesses are 5% more productive and 6% more profitable than their competitors (Brynjolfsson et al., 2011). Henry Brady provides a succinct overview when he says, “Burgeoning data and innovative methods facilitate answering previously hard-to-tackle questions about society by offering new ways to form concepts from data, to do descriptive inference, to make causal inferences, and to generate predictions. They also pose challenges as social scientists must grasp the meaning of concepts and predictions generated by convoluted algorithms, weigh the relative value of prediction versus causal inference, and cope with ethical challenges as their methods, such as algorithms for mobilizing voters or determining bail, are adopted by policy makers” (Brady, 2019).

Example: New potential for social science

The billion prices project is a great example of how researchers can use new web-scraping techniques to obtain online prices from hundreds of websites and thousands of webpages to build datasets customized to fit specific measurement and research needs in ways that were unimaginable 20 years ago (Cavallo and Rigobon, 2016); other great examples include the way in which researchers use text analysis of political speeches to study political polarization (Peterson and Spirling, 2018) or of Airbnb postings to obtain new insights into racial discrimination (Edelman et al., 2017).

Of course, these new sources come with their own caveats and biases that need to be considered when drawing inferences. We will cover this later in the book in more detail.

But most interestingly, the new data can change the way we think about behavior. For example, in a study of environmental

effects on health, researchers combine information on public school cafeteria deliveries with children's school health records to show that simply putting water jets in cafeterias reduced milk consumption and also reduced childhood obesity (Schwartz et al., 2016). Another study which sheds new light into the role of peers on productivity finds that the productivity of a cashier increases if they are within eyesight of a highly productive cashier but not otherwise (Mas and Moretti, 2009). Studies such as these show ways in which clever use of data can lead to greater understanding of the effects of complex environmental inputs on human behavior.

New types of data also can enable us to study and examine small groups—the tails of a distribution—in a way that is not possible with small data. Much of the interest in human behavior is driven by those tails, such as health care costs by small numbers of ill people (Stanton and Rutherford, 2006) or economic activity and employment by a small number of firms (Evans, 1987; Jovanovic, 1982).

Our excitement about the value of new types of data must be accompanied by a recognition of the lessons learned by statisticians and social scientists from their past experience with surveys and small scale data collection. The next sections provide a brief overview.

1.3 The importance of inference

It is critically important to be able to use data to generalize from the data source to the population. That requirement exists, regardless of the data source. Statisticians and social scientists have developed methodologies for survey data to overcome problems in the data-generating process. A guiding principle for survey methodologists is the total survey error framework, and statistical methods for weighting, calibration, and other forms of adjustment are commonly used to mitigate errors in the survey process. Likewise for “broken” experimental data, techniques such as propensity score adjustment and principal stratification are widely used to fix flaws in the data-generating process.

If we take a look across the social sciences, including economics, public policy, sociology, management, (parts of) psychology, and the like, their scientific activities can be grouped into three categories with three different inferential goals: Description, Causation, and Prediction.

1.3.1 Description

The job of many social scientists is to provide descriptive statements about the population of interest. These could be univariate, bivariate, or even multivariate statements.

Usually, descriptive statistics are created based on census data or sample surveys to create some summary statistics such as a mean, a median, or a graphical distribution to describe the population of interest. In the case of a census, the work ends there. With sample surveys, the point estimates come with measures of uncertainties (standard errors). The estimation of standard errors has been established for most descriptive statistics and common survey designs, even complex ones that include multiple layers of sampling and disproportional selection probabilities (Hansen et al., 1993; Valliant et al., 2018).

Example: Descriptive statistics

The Census Bureau's American Community Survey (ACS) "helps local officials, community leaders, and businesses understand the changes taking place in their communities. It is the premier source for detailed population and housing information about our nation" (<https://www.census.gov/programs-surveys/acs>). The summary statistics are used by planners to allocate resources—but it's important to pay attention to the standard errors, particularly for small samples. For example, in one county (Autauga) in Alabama, with a total population of about 55,000, the ACS estimates that 139 children under age 5 live in poverty—plus or minus 178! So the plausible range is somewhere between 0 and 317 (Spielman and Singleton, 2015).

Proper inference from a sample survey to the population usually depends on (1) knowing that everyone from the target population has had the chance to be included in the survey and (2) calculating the selection probability for each element in the population. The latter does not necessarily need to be known prior to sampling, but eventually a probability is assigned for each case. Obtaining correct selection probabilities is particularly important when reporting totals (Lohr, 2009). Unfortunately in practice, samples that begin as probability samples can suffer from a high rate of nonresponse. Because the survey designer cannot completely control which units respond, the set of units that ultimately respond cannot be considered to be a probability sample. Nevertheless, starting with a probability sample provides some degree of assurance that a sample

will have limited coverage errors (nonzero probability of being in the sample).

1.3.2 Causation

Identifying causal relationships is another common goal for social science researchers (Varian, 2014). Ideally, such explanations stem from data that allow causal inference: typically randomized experiments or strong non-experimental study designs. When examining the effect of X on Y , knowing how cases have been selected into the sample or dataset is much less important for estimating causal effects than they are for descriptive studies, e.g., population means. What is important is that all elements of the inferential population have a chance to be selected for the treatment (Imbens and Rubin, 2015). In the debate about probability and non-probability surveys, this distinction often is overlooked. Medical researchers have operated with unknown study selection mechanisms for years, e.g., randomized trials that enroll very select samples.

Example: New data and causal inference

If the data-generating process is not understood, resources can be badly misallocated. Overreliance on, for example, Twitter data, in targeting resources after hurricanes can lead to the misallocation of resources towards young internet-savvy people with cell phones and away from elderly or impoverished neighborhoods (Shelton et al., 2014). Of course, all data collection approaches have had similar risks. Bad survey methodology is what led the *Literary Digest* to incorrectly call the 1936 election for Landon, not Roosevelt (Squire, 1988). Inadequate understanding of coverage, incentive, and quality issues, together with the lack of a comparison group, has hampered the use of administrative records—famously in the case of using administrative records on crime to make inferences about the role of death penalty policy in crime reduction (Donohue and Wolfers, 2006).

In practice, regardless of how much data are available, researchers must consider at least two things: (1) how well the results generalize to other populations (Athey and Imbens, 2017) and (2) whether the treatment effect on the treated population is different than the treatment effect on the full population of interest (Stuart, 2010). New methods to address generalizability are under development (DuGoff et al., 2014). While unknown study selection probabilities usually make it difficult to estimate population causal effects, as long as we are able to model the selection process there is no reason not to do causal inference from so-called non-probability data.

1.3.3 Prediction

Forecasting or prediction tasks. The potential for massive amounts of data to improve prediction is undeniable. However, just like the causal inference setting, it is of the utmost importance that we know the process that has generated the data, so that biases due to unknown or unobserved systematic selection can be minimized. Predictive policing is a good example of the challenges. The criminal justice system generates massive amounts of data that can be used to better allocate police resources—but if the data do not represent the population at large, the predictions could be biased and, more importantly, the interventions assigned using those predictions could harm society.

Example: Learning from the flu

“Five years ago [in 2009], a team of researchers from Google announced a remarkable achievement in one of the world’s top scientific journals, *Nature*. Without needing the results of a single medical check-up, they were nevertheless able to track the spread of influenza across the US. What’s more, they could do it more quickly than the Centers for Disease Control and Prevention (CDC). Google’s tracking had only a day’s delay, compared with the week or more it took for the CDC to assemble a picture based on reports from doctors’ surgeries. Google was faster because it was tracking the outbreak by finding a correlation between what people searched for online and whether they had flu symptoms.” . . .

“Four years after the original *Nature* paper was published, *Nature News* had sad tidings to convey: the latest flu outbreak had claimed an unexpected victim: Google Flu Trends. After reliably providing a swift and accurate account of flu outbreaks for several winters, the theory-free, data-rich model had lost its nose for where flu was going. Google’s model pointed to a severe outbreak but when the slow-and-steady data from the CDC arrived, they showed that Google’s estimates of the spread of flu-like illnesses were overstated by almost a factor of two.

The problem was that Google did not know—could not begin to know—what linked the search terms with the spread of flu. Google’s engineers weren’t trying to figure out what caused what. They were merely finding statistical patterns in the data. They cared about correlation rather than causation” (Harford, 2014).

1.4 The importance of understanding how data are generated

The costs of realizing the benefits of the new types of data are nontrivial. Even if data collection is cheap, the costs of cleaning,

► This topic will be discussed in more detail in Section 1.5.

curating, standardizing, integrating, and using the new types of data are substantial. In essence, just as with data from surveys, data still need to be processed—cleaned, normalized, and variables coded—but this needs to be done at scale. But even after all of these tasks are completed, social scientists have a key role in describing the quality of the data. This role is important, because most data in the real world are noisy, inconsistent, and exhibit missing values. Data quality can be characterized in multiple ways (see Christen, 2012a; National Academies of Sciences, Engineering, and Medicine and others [2018]), such as:

- **Accuracy:** How accurate are the attribute values in the data?
- **Completeness:** Are the data complete?
- **Consistency:** How consistent are the values in and between different database(s)?
- **Timeliness:** How timely are the data?
- **Accessibility:** Are all variables available for analysis?

In the social science world, the assessment of data quality has been integral to the production of the resultant statistics. That has not necessarily been easy when assessing new types of data. A good example of the importance of understanding how data are generated arose in one of our classes a few years ago, when class participants were asked to develop employment measures for ex-offenders in the period after they were released from prison (Kreuter et al., 2019).

For people working with surveys, the definition was already pre-constructed: in the Current Population Survey (CPS), respondents were asked about their work activity in the week covering the 12th of the month. Individuals were counted as employed if they had at least one hour of paid work in that week (with some exceptions for family and farm work). But the class participants were working with administrative records from the Illinois Department of Employment Security and the Illinois Department of Corrections (Kreuter et al., 2019). Those records provided a report of all jobs in every quarter that each individual held in the state; when matched to data about formerly incarcerated individuals, it could provide rich information about their employment patterns. A group of class participants produced [Figure 1.2](#)—the white boxes represent quarters in which an individual does not have a job and the blue boxes represent quarters in which an individual does have a job.

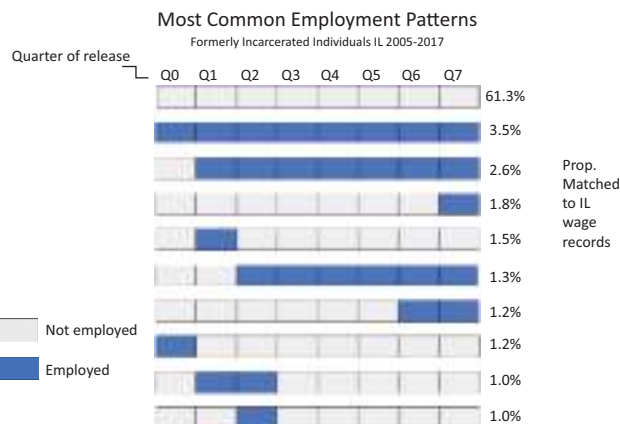


Figure 1.2. Most common employment patterns, formerly incarcerated individuals in Illinois, 2005–2017

A quick look at the results is very interesting. First, the participants present an entirely new dynamic way of looking at employment—not just the relatively static CPS measure. Second, the results are a bit shocking. More than 61% of Illinois exoffenders do not have a job in any of the eight quarters after their release. Only 3.5% have a job in all of the quarters. This is where social scientists and government analysts can contribute—because they know how the data are generated. The matches between the two agencies have been conducted on (deidentified) Social Security numbers (SSNs). It is likely that there are several gaps in those matches. First, agency staff know that the quality of SSNs in prisons is quite low, so that may be one reason for the low match rate. Second, the matches are only to Illinois jobs, and many formerly incarcerated individuals could be working across state lines (if allowed). Third, the individuals may be attending community college, or accepting social assistance, or reincarcerated. More data can be used to examine these different possibilities—but we believe this illustrates the value that social scientists and subject matter experts provide to measuring the quality issues we highlight at the beginning of this section.

1.5 New tools for new data

The new data sources that we have discussed frequently require working at scales for which the social scientist's familiar tools are