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
Joshua Tokle and Stefan Bender
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
Ian Foster and Pascal Heus
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
Huy Vo and Claudio Silva
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
M. Adil Yalçın and Catherine Plaisant
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
Rayid Ghani and Malte Schierholz
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
Evgeny Klochikhin and Jordan Boyd-Graber
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
Jason Owen-Smith
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
Paul P. Biemer
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
Kit T. Rodolfa, Pedro Saleiro, and Rayid Ghani
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
Stefan Bender, Ron S. Jarmin, Frauke Kreuter, and Julia Lane
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
Brian Kim, Christoph Kern, Jonathan Scott Morgan, Clayton Hunter, and Avishek Kumar
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