

# Contents

|                                                        |      |
|--------------------------------------------------------|------|
| Preface                                                | xv   |
| Editors                                                | xvii |
| Contributors                                           | xxi  |
| 1 Introduction                                         | 1    |
| 1.1 Why this book?                                     | 1    |
| 1.2 Defining big data and its value                    | 3    |
| 1.3 Social science, inference, and big data            | 4    |
| 1.4 Social science, data quality, and big data         | 7    |
| 1.5 New tools for new data                             | 9    |
| 1.6 The book's "use case"                              | 10   |
| 1.7 The structure of the book                          | 13   |
| 1.7.1 Part I: Capture and curation                     | 13   |
| 1.7.2 Part II: Modeling and analysis                   | 15   |
| 1.7.3 Part III: Inference and ethics                   | 16   |
| 1.8 Resources                                          | 17   |
| I Capture and Curation                                 | 21   |
| 2 Working with Web Data and APIs                       | 23   |
| Cameron Neylon                                         |      |
| 2.1 Introduction                                       | 23   |
| 2.2 Scraping information from the web                  | 24   |
| 2.2.1 Obtaining data from the HHMI website             | 24   |
| 2.2.2 Limits of scraping                               | 30   |
| 2.3 New data in the research enterprise                | 31   |
| 2.4 A functional view                                  | 37   |
| 2.4.1 Relevant APIs and resources                      | 38   |
| 2.4.2 RESTful APIs, returned data, and Python wrappers | 38   |
| 2.5 Programming against an API                         | 41   |

|        |                                                             |     |
|--------|-------------------------------------------------------------|-----|
| 2.6    | Using the ORCID API via a wrapper . . . . .                 | 42  |
| 2.7    | Quality, scope, and management . . . . .                    | 44  |
| 2.8    | Integrating data from multiple sources . . . . .            | 46  |
| 2.8.1  | The Lagotto API . . . . .                                   | 46  |
| 2.8.2  | Working with a corpus . . . . .                             | 52  |
| 2.9    | Working with the graph of relationships . . . . .           | 58  |
| 2.9.1  | Citation links between articles . . . . .                   | 58  |
| 2.9.2  | Categories, sources, and connections . . . . .              | 60  |
| 2.9.3  | Data availability and completeness . . . . .                | 61  |
| 2.9.4  | The value of sparse dynamic data . . . . .                  | 62  |
| 2.10   | Bringing it together: Tracking pathways to impact . . . . . | 65  |
| 2.10.1 | Network analysis approaches . . . . .                       | 66  |
| 2.10.2 | Future prospects and new data sources . . . . .             | 66  |
| 2.11   | Summary . . . . .                                           | 67  |
| 2.12   | Resources . . . . .                                         | 69  |
| 2.13   | Acknowledgements and copyright . . . . .                    | 70  |
| 3      | <b>Record Linkage</b> . . . . .                             | 71  |
|        | <i>Joshua Tokle and Stefan Bender</i> . . . . .             |     |
| 3.1    | Motivation . . . . .                                        | 71  |
| 3.2    | Introduction to record linkage . . . . .                    | 72  |
| 3.3    | Preprocessing data for record linkage . . . . .             | 76  |
| 3.4    | Indexing and blocking . . . . .                             | 78  |
| 3.5    | Matching . . . . .                                          | 80  |
| 3.5.1  | Rule-based approaches . . . . .                             | 82  |
| 3.5.2  | Probabilistic record linkage . . . . .                      | 83  |
| 3.5.3  | Machine learning approaches to linking . . . . .            | 85  |
| 3.5.4  | Disambiguating networks . . . . .                           | 88  |
| 3.6    | Classification . . . . .                                    | 88  |
| 3.6.1  | Thresholds . . . . .                                        | 89  |
| 3.6.2  | One-to-one links . . . . .                                  | 90  |
| 3.7    | Record linkage and data protection . . . . .                | 91  |
| 3.8    | Summary . . . . .                                           | 92  |
| 3.9    | Resources . . . . .                                         | 92  |
| 4      | <b>Databases</b> . . . . .                                  | 93  |
|        | <i>Ian Foster and Pascal Heus</i> . . . . .                 |     |
| 4.1    | Introduction . . . . .                                      | 93  |
| 4.2    | DBMS: When and why . . . . .                                | 94  |
| 4.3    | Relational DBMSs . . . . .                                  | 100 |
| 4.3.1  | Structured Query Language (SQL) . . . . .                   | 102 |
| 4.3.2  | Manipulating and querying data . . . . .                    | 102 |
| 4.3.3  | Schema design and definition . . . . .                      | 105 |

|       |                                                                              |     |
|-------|------------------------------------------------------------------------------|-----|
| 4.3.4 | Loading data . . . . .                                                       | 107 |
| 4.3.5 | Transactions and crash recovery . . . . .                                    | 108 |
| 4.3.6 | Database optimizations . . . . .                                             | 109 |
| 4.3.7 | Caveats and challenges . . . . .                                             | 112 |
| 4.4   | Linking DBMSs and other tools . . . . .                                      | 113 |
| 4.5   | NoSQL databases . . . . .                                                    | 116 |
| 4.5.1 | Challenges of scale: The CAP theorem . . . . .                               | 116 |
| 4.5.2 | NoSQL and key-value stores . . . . .                                         | 117 |
| 4.5.3 | Other NoSQL databases . . . . .                                              | 119 |
| 4.6   | Spatial databases . . . . .                                                  | 120 |
| 4.7   | Which database to use? . . . . .                                             | 122 |
| 4.7.1 | Relational DBMSs . . . . .                                                   | 122 |
| 4.7.2 | NoSQL DBMSs . . . . .                                                        | 123 |
| 4.8   | Summary . . . . .                                                            | 123 |
| 4.9   | Resources . . . . .                                                          | 124 |
| 5     | Programming with Big Data . . . . .                                          | 125 |
|       | <i>Huy Vo and Claudio Silva</i>                                              |     |
| 5.1   | Introduction . . . . .                                                       | 125 |
| 5.2   | The MapReduce programming model . . . . .                                    | 127 |
| 5.3   | Apache Hadoop MapReduce . . . . .                                            | 129 |
| 5.3.1 | The Hadoop Distributed File System . . . . .                                 | 130 |
| 5.3.2 | Hadoop: Bringing compute to the data . . . . .                               | 131 |
| 5.3.3 | Hardware provisioning . . . . .                                              | 134 |
| 5.3.4 | Programming language support . . . . .                                       | 136 |
| 5.3.5 | Fault tolerance . . . . .                                                    | 137 |
| 5.3.6 | Limitations of Hadoop . . . . .                                              | 137 |
| 5.4   | Apache Spark . . . . .                                                       | 138 |
| 5.5   | Summary . . . . .                                                            | 141 |
| 5.6   | Resources . . . . .                                                          | 143 |
| II    | Modeling and Analysis . . . . .                                              | 145 |
| 6     | Machine Learning . . . . .                                                   | 147 |
|       | <i>Rayid Ghani and Malte Schierholz</i>                                      |     |
| 6.1   | Introduction . . . . .                                                       | 147 |
| 6.2   | What is machine learning? . . . . .                                          | 148 |
| 6.3   | <i>The machine learning process</i> . . . . .                                | 150 |
| 6.4   | Problem formulation: Mapping a problem to machine learning methods . . . . . | 151 |
| 6.5   | Methods . . . . .                                                            | 153 |
| 6.5.1 | Unsupervised learning methods . . . . .                                      | 153 |
| 6.5.2 | <i>Supervised learning</i> . . . . .                                         | 161 |

|         |                                                                    |     |
|---------|--------------------------------------------------------------------|-----|
| 6.6     | Evaluation . . . . .                                               | 173 |
| 6.6.1   | Methodology . . . . .                                              | 173 |
| 6.6.2   | Metrics . . . . .                                                  | 176 |
| 6.7     | Practical tips . . . . .                                           | 180 |
| 6.7.1   | Features . . . . .                                                 | 180 |
| 6.7.2   | Machine learning pipeline . . . . .                                | 181 |
| 6.7.3   | Multiclass problems . . . . .                                      | 181 |
| 6.7.4   | Skewed or imbalanced classification problems . . . . .             | 182 |
| 6.8     | How can social scientists benefit from machine learning? . . . . . | 183 |
| 6.9     | Advanced topics . . . . .                                          | 185 |
| 6.10    | Summary . . . . .                                                  | 185 |
| 6.11    | Resources . . . . .                                                | 186 |
| 7       | <b>Text Analysis</b> . . . . .                                     | 187 |
|         | <i>Evgeny Klochikhin and Jordan Boyd-Graber</i>                    |     |
| 7.1     | Understanding what people write . . . . .                          | 187 |
| 7.2     | How to analyze text . . . . .                                      | 189 |
| 7.2.1   | Processing text data . . . . .                                     | 190 |
| 7.2.2   | How much is a word worth? . . . . .                                | 192 |
| 7.3     | Approaches and applications . . . . .                              | 193 |
| 7.3.1   | Topic modeling . . . . .                                           | 193 |
| 7.3.1.1 | Inferring topics from raw text . . . . .                           | 194 |
| 7.3.1.2 | Applications of topic models . . . . .                             | 197 |
| 7.3.2   | Information retrieval and clustering . . . . .                     | 198 |
| 7.3.3   | Other approaches . . . . .                                         | 205 |
| 7.4     | Evaluation . . . . .                                               | 208 |
| 7.5     | Text analysis tools . . . . .                                      | 210 |
| 7.6     | Summary . . . . .                                                  | 212 |
| 7.7     | Resources . . . . .                                                | 213 |
| 8       | <b>Networks: The Basics</b> . . . . .                              | 215 |
|         | <i>Jason Owen-Smith</i>                                            |     |
| 8.1     | Introduction . . . . .                                             | 215 |
| 8.2     | Network data . . . . .                                             | 218 |
| 8.2.1   | Forms of network data . . . . .                                    | 218 |
| 8.2.2   | Inducing one-mode networks from two-mode data . . . . .            | 220 |
| 8.3     | Network measures . . . . .                                         | 224 |
| 8.3.1   | Reachability . . . . .                                             | 224 |
| 8.3.2   | Whole-network measures . . . . .                                   | 225 |
| 8.4     | Comparing collaboration networks . . . . .                         | 234 |
| 8.5     | Summary . . . . .                                                  | 238 |
| 8.6     | Resources . . . . .                                                | 239 |

|          |                                                                                |     |
|----------|--------------------------------------------------------------------------------|-----|
| III      | Inference and Ethics                                                           | 241 |
| 9        | Information Visualization                                                      | 243 |
|          | <i>M. Adil Yalçın and Catherine Plaisant</i>                                   |     |
| 9.1      | Introduction                                                                   | 243 |
| 9.2      | Developing effective visualizations                                            | 244 |
| 9.3      | A data-by-tasks taxonomy                                                       | 249 |
| 9.3.1    | Multivariate data                                                              | 249 |
| 9.3.2    | Spatial data                                                                   | 251 |
| 9.3.3    | Temporal data                                                                  | 252 |
| 9.3.4    | Hierarchical data                                                              | 255 |
| 9.3.5    | Network data                                                                   | 257 |
| 9.3.6    | Text data                                                                      | 259 |
| 9.4      | Challenges                                                                     | 259 |
| 9.4.1    | Scalability                                                                    | 260 |
| 9.4.2    | Evaluation                                                                     | 261 |
| 9.4.3    | Visual impairment                                                              | 261 |
| 9.4.4    | Visual literacy                                                                | 262 |
| 9.5      | Summary                                                                        | 262 |
| 9.6      | Resources                                                                      | 263 |
| 10       | Errors and Inference                                                           | 265 |
|          | <i>Paul P. Biemer</i>                                                          |     |
| 10.1     | Introduction                                                                   | 265 |
| 10.2     | The total error paradigm                                                       | 266 |
| 10.2.1   | The traditional model                                                          | 266 |
| 10.2.2   | Extending the framework to big data                                            | 273 |
| 10.3     | Illustrations of errors in big data                                            | 275 |
| 10.4     | Errors in big data analytics                                                   | 277 |
| 10.4.1   | Errors resulting from volume, velocity, and variety, assuming perfect veracity | 277 |
| 10.4.2   | Errors resulting from lack of veracity                                         | 279 |
| 10.4.2.1 | Variable and correlated error                                                  | 280 |
| 10.4.2.2 | Models for categorical data                                                    | 282 |
| 10.4.2.3 | Misclassification and rare classes                                             | 283 |
| 10.4.2.4 | Correlation analysis                                                           | 284 |
| 10.4.2.5 | Regression analysis                                                            | 288 |
| 10.5     | Some methods for mitigating, detecting, and compensating for errors            | 290 |
| 10.6     | Summary                                                                        | 295 |
| 10.7     | Resources                                                                      | 296 |

## 11 Privacy and Confidentiality

*Stefan Bender, Ron Jarmin, Frauke Kreuter, and Julia Lane*

|                                  |     |
|----------------------------------|-----|
| 11.1 Introduction                | 299 |
| 11.2 Why is access important?    | 303 |
| 11.3 Providing access            | 305 |
| 11.4 The new challenges          | 306 |
| 11.5 Legal and ethical framework | 308 |
| 11.6 Summary                     | 310 |
| 11.7 Resources                   | 311 |

## 12 Workbooks

*Jonathan Scott Morgan, Christina Jones, and Ahmad Emad*

|                                  |     |
|----------------------------------|-----|
| 12.1 Introduction                | 313 |
| 12.2 Environment                 | 314 |
| 12.2.1 Running workbooks locally | 314 |
| 12.2.2 Central workbook server   | 315 |
| 12.3 Workbook details            | 315 |
| 12.3.1 Social Media and APIs     | 315 |
| 12.3.2 Database basics           | 316 |
| 12.3.3 Data Linkage              | 316 |
| 12.3.4 Machine Learning          | 317 |
| 12.3.5 Text Analysis             | 317 |
| 12.3.6 Networks                  | 318 |
| 12.3.7 Visualization             | 318 |
| 12.4 Resources                   | 319 |

## Bibliography

## Index

321

349