contents

preface xiii
acknowledgments xiv
about this book xvi
about the authors xviii
about the cover illustration xx

1 ***Data science in a big data world*** 1

1.1 Benefits and uses of data science and big data 2

1.2 Facets of data 4

Structured data 4 ▪ *Unstructured data* 5
Natural language 5 ▪ *Machine-generated data* 6
Graph-based or network data 7 ▪ *Audio, image, and video* 8
Streaming data 8

1.3 The data science process 8

Setting the research goal 8 ▪ *Retrieving data* 9
Data preparation 9 ▪ *Data exploration* 9
Data modeling or model building 9 ▪ *Presentation and automation* 9

1.4 The big data ecosystem and data science 10

Distributed file systems 10 ▪ *Distributed programming framework* 12 ▪ *Data integration framework* 12

Machine learning frameworks 12 ▪ NoSQL databases 13
Scheduling tools 14 ▪ Benchmarking tools 14
System deployment 14 ▪ Service programming 14
Security 14

1.5 An introductory working example of Hadoop 15

1.6 Summary 20

2 ***The data science process 22***

2.1 Overview of the data science process 22

Don't be a slave to the process 25

2.2 Step 1: Defining research goals and creating a project charter 25

Spend time understanding the goals and context of your research 26
Create a project charter 26

2.3 Step 2: Retrieving data 27

Start with data stored within the company 28 ▪ Don't be afraid to shop around 28 ▪ Do data quality checks now to prevent problems later 29

2.4 Step 3: Cleansing, integrating, and transforming data 29

Cleansing data 30 ▪ Correct errors as early as possible 36
Combining data from different data sources 37
Transforming data 40

2.5 Step 4: Exploratory data analysis 43

2.6 Step 5: Build the models 48

Model and variable selection 48 ▪ Model execution 49
Model diagnostics and model comparison 54

2.7 Step 6: Presenting findings and building applications on top of them 55

2.8 Summary 56

3 ***Machine learning 57***

3.1 What is machine learning and why should you care about it? 58

Applications for machine learning in data science 58
Where machine learning is used in the data science process 59
Python tools used in machine learning 60

3.2 The modeling process 62

Engineering features and selecting a model 62 ▪ Training your model 64 ▪ Validating a model 64 ▪ Predicting new observations 65

3.3 Types of machine learning 65

Supervised learning 66 ▪ Unsupervised learning 72

3.4 Semi-supervised learning 82

3.5 Summary 83

4 *Handling large data on a single computer 85*

4.1 The problems you face when handling large data 86

4.2 General techniques for handling large volumes of data 87

Choosing the right algorithm 88 ▪ Choosing the right data structure 96 ▪ Selecting the right tools 99

4.3 General programming tips for dealing with large data sets 101

Don't reinvent the wheel 101 ▪ Get the most out of your hardware 102 ▪ Reduce your computing needs 102

4.4 Case study 1: Predicting malicious URLs 103

Step 1: Defining the research goal 104 ▪ Step 2: Acquiring the URL data 104 ▪ Step 4: Data exploration 105 Step 5: Model building 106

4.5 Case study 2: Building a recommender system inside a database 108

Tools and techniques needed 108 ▪ Step 1: Research question 111 ▪ Step 3: Data preparation 111 Step 5: Model building 115 ▪ Step 6: Presentation and automation 116

4.6 Summary 118

5 *First steps in big data 119*

5.1 Distributing data storage and processing with frameworks 120

Hadoop: a framework for storing and processing large data sets 121 Spark: replacing MapReduce for better performance 123

5.2 Case study: Assessing risk when loaning money 125

Step 1: The research goal 126 ▪ Step 2: Data retrieval 127
Step 3: Data preparation 131 ▪ Step 4: Data exploration & Step 6: Report building 135

5.3 Summary 149

6 *Join the NoSQL movement 150*

6.1 Introduction to NoSQL 153

ACID: the core principle of relational databases 153
CAP Theorem: the problem with DBs on many nodes 154
The BASE principles of NoSQL databases 156
NoSQL database types 158

6.2 Case study: What disease is that? 164

Step 1: Setting the research goal 166 ▪ Steps 2 and 3: Data retrieval and preparation 167 ▪ Step 4: Data exploration 175
Step 3 revisited: Data preparation for disease profiling 183
Step 4 revisited: Data exploration for disease profiling 187
Step 6: Presentation and automation 188

6.3 Summary 189

7 *The rise of graph databases 190*

7.1 Introducing connected data and graph databases 191

Why and when should I use a graph database? 193

7.2 Introducing Neo4j: a graph database 196

Cypher: a graph query language 198

7.3 Connected data example: a recipe recommendation engine 204

Step 1: Setting the research goal 205 ▪ Step 2: Data retrieval 206
Step 3: Data preparation 207 ▪ Step 4: Data exploration 210
Step 5: Data modeling 212 ▪ Step 6: Presentation 216

7.4 Summary 216

8 *Text mining and text analytics 218*

8.1 Text mining in the real world 220

8.2 Text mining techniques 225

Bag of words 225 ▪ Stemming and lemmatization 227
Decision tree classifier 228

8.3 Case study: Classifying Reddit posts 230

Meet the Natural Language Toolkit 231 ▪ Data science process overview and step 1: The research goal 233 ▪ Step 2: Data retrieval 234 ▪ Step 3: Data preparation 237 ▪ Step 4: Data exploration 240 ▪ Step 3 revisited: Data preparation adapted 242 ▪ Step 5: Data analysis 246 ▪ Step 6: Presentation and automation 250

8.4 Summary 252

9 ***Data visualization to the end user 253***

9.1 Data visualization options 254

9.2 Crossfilter, the JavaScript MapReduce library 257

Setting up everything 258 ▪ Unleashing Crossfilter to filter the medicine data set 262

9.3 Creating an interactive dashboard with dc.js 267

9.4 Dashboard development tools 272

9.5 Summary 273

appendix A Setting up Elasticsearch 275
appendix B Setting up Neo4j 281
appendix C Installing MySQL server 284
appendix D Setting up Anaconda with a virtual environment 288

index 291