Table of Contents

Preface. xiii

1. Introduction: What Is Data Science?. 1

Big Data and Data Science Hype 1
Getting Past the Hype 3
Why Now? 4
Datafication 5
The Current Landscape (with a Little History) 6
Data Science Jobs 10
A Data Science Profile 10
Thought Experiment: Meta-Definition 13
OK, So What Is a Data Scientist, Really? 14
In Academia 14
In Industry 15

2. Statistical Inference, Exploratory Data Analysis, and the Data Science Process. 17

Statistical Thinking in the Age of Big Data 17
Statistical Inference 18
Populations and Samples 19
Populations and Samples of Big Data 21
Big Data Can Mean Big Assumptions 24
Modeling 26
Exploratory Data Analysis 34
Philosophy of Exploratory Data Analysis 36
Exercise: EDA 37
The Data Science Process 41
A Data Scientist's Role in This Process 43

Thought Experiment: How Would You Simulate Chaos? 44
Case Study: RealDirect 46
How Does RealDirect Make Money? 47
Exercise: RealDirect Data Strategy 48

3. Algorithms. 51
Machine Learning Algorithms 52
Three Basic Algorithms 54
Linear Regression 55
k-Nearest Neighbors (k-NN) 71
k-means 82
Exercise: Basic Machine Learning Algorithms 86
Solutions 86
Summing It All Up 91
Thought Experiment: Automated Statistician 92

4. Spam Filters, Naive Bayes, and Wrangling. 93
Thought Experiment: Learning by Example 93
Why Won't Linear Regression Work for Filtering Spam? 95
How About k-nearest Neighbors? 96
Naive Bayes 98
Bayes Law 98
A Spam Filter for Individual Words 99
A Spam Filter That Combines Words: Naive Bayes 101
Fancy It Up: Laplace Smoothing 103
Comparing Naive Bayes to k-NN 105
Sample Code in bash 105
Scraping the Web: APIs and Other Tools 106
Jake's Exercise: Naive Bayes for Article Classification 108
Sample R Code for Dealing with the NYT API 110

5. Logistic Regression. 113
Thought Experiments 114
Classifiers 115
Runtime 116
You 117
Interpretability 117
Scalability 117
M6D Logistic Regression Case Study 118
Click Models 118
The Underlying Math 120

Estimating α and β 122
Newton's Method 124
Stochastic Gradient Descent 124
Implementation 124
Evaluation 125
Media 6 Degrees Exercise 128
Sample R Code 129

6. Time Stamps and Financial Modeling........................ 135
Kyle Teague and GetGlue 135
Timestamps 137
Exploratory Data Analysis (EDA) 138
Metrics and New Variables or Features 142
What's Next? 142
Cathy O'Neil 144
Thought Experiment 144
Financial Modeling 145
In-Sample, Out-of-Sample, and Causality 146
Preparing Financial Data 148
Log Returns 149
Example: The S&P Index 151
Working out a Volatility Measurement 153
Exponential Downweighting 155
The Financial Modeling Feedback Loop 156
Why Regression? 158
Adding Priors 159
A Baby Model 159
Exercise: GetGlue and Timestamped Event Data 162
Exercise: Financial Data 164

7. Extracting Meaning from Data........................... 165
William Cukierski 165
Background: Data Science Competitions 166
Background: Crowdsourcing 167
The Kaggle Model 170
A Single Contestant 170
Their Customers 172
Thought Experiment: What Are the Ethical Implications of a Robo-Grader? 174
Feature Selection 176
Example: User Retention 177

Filters 181
Wrappers 181
Embedded Methods: Decision Trees 184
Entropy 186
The Decision Tree Algorithm 187
Handling Continuous Variables in Decision Trees 188
Random Forests 190
User Retention: Interpretability Versus Predictive Power 192
David Huffaker: Google's Hybrid Approach to Social Research 193
Moving from Descriptive to Predictive 194
Social at Google 196
Privacy 196
Thought Experiment: What Is the Best Way to Decrease Concern and Increase Understanding and Control? 197

8. Recommendation Engines: Building a User-Facing Data Product at Scale.. 199
A Real-World Recommendation Engine 200
Nearest Neighbor Algorithm Review 202
Some Problems with Nearest Neighbors 202
Beyond Nearest Neighbor: Machine Learning Classification 204
The Dimensionality Problem 206
Singular Value Decomposition (SVD) 207
Important Properties of SVD 208
Principal Component Analysis (PCA) 209
Alternating Least Squares 211
Fix V and Update U 212
Last Thoughts on These Algorithms 213
Thought Experiment: Filter Bubbles 213
Exercise: Build Your Own Recommendation System 214
Sample Code in Python 214

9. Data Visualization and Fraud Detection........................ 217
Data Visualization History 217
Gabriel Tarde 218
Mark's Thought Experiment 219
What Is Data Science, Redux? 220
Processing 221
Franco Moretti 221

A Sample of Data Visualization Projects 222
Mark's Data Visualization Projects 227
New York Times Lobby: Moveable Type 227
Project Cascade: Lives on a Screen 230
Cronkite Plaza 231
eBay Transactions and Books 232
Public Theater Shakespeare Machine 234
Goals of These Exhibits 235
Data Science and Risk 235
About Square 236
The Risk Challenge 237
The Trouble with Performance Estimation 240
Model Building Tips 244
Data Visualization at Square 248
Ian's Thought Experiment 249
Data Visualization for the Rest of Us 250
Data Visualization Exercise 251

10. Social Networks and Data Journalism........................ 253
Social Network Analysis at Morning Analytics 254
Case-Attribute Data versus Social Network Data 254
Social Network Analysis 255
Terminology from Social Networks 256
Centrality Measures 257
The Industry of Centrality Measures 258
Thought Experiment 259
Morningside Analytics 260
How Visualizations Help Us Find Schools of Fish 262
More Background on Social Network Analysis from a Statistical Point of View 263
Representations of Networks and Eigenvalue Centrality 264
A First Example of Random Graphs: The Erdos-Renyi Model 265
A Second Example of Random Graphs: The Exponential Random Graph Model 266
Data Journalism 269
A Bit of History on Data Journalism 269
Writing Technical Journalism: Advice from an Expert 270

11. Causality.. 273
Correlation Doesn't Imply Causation 274

Asking Causal Questions 274
Confounders: A Dating Example 275
OK Cupid's Attempt 276
The Gold Standard: Randomized Clinical Trials 279
A/B Tests 280
Second Best: Observational Studies 283
Simpson's Paradox 283
The Rubin Causal Model 285
Visualizing Causality 286
Definition: The Causal Effect 287
Three Pieces of Advice 289

12. Epidemiology. 291
Madigan's Background 291
Thought Experiment 292
Modern Academic Statistics 293
Medical Literature and Observational Studies 293
Stratification Does Not Solve the Confounder Problem 294
What Do People Do About Confounding Things in Practice? 295
Is There a Better Way? 296
Research Experiment (Observational Medical Outcomes Partnership) 298
Closing Thought Experiment 303

13. Lessons Learned from Data Competitions: Data Leakage and Model Evaluation. 305
Claudia's Data Scientist Profile 306
The Life of a Chief Data Scientist 306
On Being a Female Data Scientist 307
Data Mining Competitions 307
How to Be a Good Modeler 309
Data Leakage 309
Market Predictions 310
Amazon Case Study: Big Spenders 310
A Jewelry Sampling Problem 311
IBM Customer Targeting 311
Breast Cancer Detection 312
Pneumonia Prediction 313
How to Avoid Leakage 315
Evaluating Models 315

Accuracy: Meh 317
Probabilities Matter, Not 0s and 1s 317
Choosing an Algorithm 320
A Final Example 321
Parting Thoughts 322

14. Data Engineering: MapReduce, Pregel, and Hadoop............ 323
About David Crawshaw 324
Thought Experiment 325
MapReduce 326
Word Frequency Problem 327
Enter MapReduce 330
Other Examples of MapReduce 332
What Can't MapReduce Do? 333
Pregel 333
About Josh Wills 334
Thought Experiment 334
On Being a Data Scientist 334
Data Abundance Versus Data Scarcity 335
Designing Models 335
Economic Interlude: Hadoop 335
A Brief Introduction to Hadoop 336
Cloudera 337
Back to Josh: Workflow 337
So How to Get Started with Hadoop? 338

15. The Students Speak.. 339
Process Thinking 339
Naive No Longer 341
Helping Hands 342
Your Mileage May Vary 344
Bridging Tunnels 347
Some of Our Work 347

16. Next-Generation Data Scientists, Hubris, and Ethics............ 349
What Just Happened? 349
What Is Data Science (Again)? 350
What Are Next-Gen Data Scientists? 352
Being Problem Solvers 352
Cultivating Soft Skills 353
Being Question Askers 354

Being an Ethical Data Scientist 356
Career Advice 361

Index. 363