
Table of Contents

Preface to the Second Edition.....	xi
Preface to the First Edition.....	xv
1. Introduction.....	1
The Ascendancy of Data.....	1
What Is Data Science?.....	1
Motivating Hypothetical: DataSciencester.....	2
Finding Key Connectors.....	3
Data Scientists You May Know.....	6
Salaries and Experience.....	8
Paid Accounts.....	11
Topics of Interest.....	11
Onward.....	13
2. A Crash Course in Python.....	15
The Zen of Python.....	15
Getting Python.....	16
Virtual Environments.....	16
Whitespace Formatting.....	17
Modules.....	19
Functions.....	20
Strings.....	21
Exceptions.....	21
Lists.....	22
Tuples.....	23
Dictionaries.....	24
defaultdict.....	25

Counters	26
Sets	26
Control Flow	27
Truthiness	28
Sorting	29
List Comprehensions	30
Automated Testing and assert	30
Object-Oriented Programming	31
Iterables and Generators	33
Randomness	35
Regular Expressions	36
Functional Programming	36
zip and Argument Unpacking	36
args and kwargs	37
Type Annotations	38
How to Write Type Annotations	41
Welcome to DataSciencester!	42
For Further Exploration	42
3. Visualizing Data.....	43
matplotlib	43
Bar Charts	45
Line Charts	49
Scatterplots	50
For Further Exploration	52
4. Linear Algebra.....	55
Vectors	55
Matrices	59
For Further Exploration	62
5. Statistics.....	63
Describing a Single Set of Data	63
Central Tendencies	65
Dispersion	67
Correlation	68
Simpson's Paradox	71
Some Other Correlational Caveats	72
Correlation and Causation	73
For Further Exploration	73

6. Probability.....	75
Dependence and Independence	75
Conditional Probability	76
Bayes's Theorem	78
Random Variables	79
Continuous Distributions	80
The Normal Distribution	81
The Central Limit Theorem	84
For Further Exploration	86
7. Hypothesis and Inference.....	87
Statistical Hypothesis Testing	87
Example: Flipping a Coin	87
p-Values	90
Confidence Intervals	92
p-Hacking	93
Example: Running an A/B Test	94
Bayesian Inference	95
For Further Exploration	99
8. Gradient Descent.....	101
The Idea Behind Gradient Descent	101
Estimating the Gradient	102
Using the Gradient	105
Choosing the Right Step Size	106
Using Gradient Descent to Fit Models	106
Minibatch and Stochastic Gradient Descent	108
For Further Exploration	109
9. Getting Data.....	111
stdin and stdout	111
Reading Files	113
The Basics of Text Files	113
Delimited Files	115
Scraping the Web	116
HTML and the Parsing Thereof	117
Example: Keeping Tabs on Congress	118
Using APIs	121
JSON and XML	121
Using an Unauthenticated API	122
Finding APIs	123
Example: Using the Twitter APIs	124

Getting Credentials	124
For Further Exploration	128
10. Working with Data.....	129
Exploring Your Data	129
Exploring One-Dimensional Data	129
Two Dimensions	131
Many Dimensions	133
Using NamedTuples	135
Dataclasses	137
Cleaning and Munging	138
Manipulating Data	140
Rescaling	142
An Aside: tqdm	144
Dimensionality Reduction	146
For Further Exploration	151
11. Machine Learning.....	153
Modeling	153
What Is Machine Learning?	154
Overfitting and Underfitting	155
Correctness	157
The Bias-Variance Tradeoff	160
Feature Extraction and Selection	161
For Further Exploration	163
12. k-Nearest Neighbors.....	165
The Model	165
Example: The Iris Dataset	167
The Curse of Dimensionality	170
For Further Exploration	174
13. Naive Bayes.....	175
A Really Dumb Spam Filter	175
A More Sophisticated Spam Filter	176
Implementation	178
Testing Our Model	180
Using Our Model	181
For Further Exploration	183
14. Simple Linear Regression.....	185
The Model	185

Using Gradient Descent	189
Maximum Likelihood Estimation	190
For Further Exploration	190
15. Multiple Regression.....	191
The Model	191
Further Assumptions of the Least Squares Model	192
Fitting the Model	193
Interpreting the Model	195
Goodness of Fit	196
Digression: The Bootstrap	196
Standard Errors of Regression Coefficients	198
Regularization	200
For Further Exploration	202
16. Logistic Regression.....	203
The Problem	203
The Logistic Function	206
Applying the Model	208
Goodness of Fit	209
Support Vector Machines	210
For Further Investigation	214
17. Decision Trees.....	215
What Is a Decision Tree?	215
Entropy	217
The Entropy of a Partition	219
Creating a Decision Tree	220
Putting It All Together	223
Random Forests	225
For Further Exploration	226
18. Neural Networks.....	227
Perceptrons	227
Feed-Forward Neural Networks	230
Backpropagation	233
Example: Fizz Buzz	235
For Further Exploration	238
19. Deep Learning.....	239
The Tensor	239
The Layer Abstraction	242

The Linear Layer	243
Neural Networks as a Sequence of Layers	246
Loss and Optimization	247
Example: XOR Revisited	250
Other Activation Functions	251
Example: FizzBuzz Revisited	252
Softmaxes and Cross-Entropy	253
Dropout	256
Example: MNIST	257
Saving and Loading Models	261
For Further Exploration	262
20. Clustering.....	263
The Idea	263
The Model	264
Example: Meetups	266
Choosing k	268
Example: Clustering Colors	269
Bottom-Up Hierarchical Clustering	271
For Further Exploration	277
21. Natural Language Processing.....	279
Word Clouds	279
n-Gram Language Models	281
Grammars	284
An Aside: Gibbs Sampling	286
Topic Modeling	288
Word Vectors	293
Recurrent Neural Networks	301
Example: Using a Character-Level RNN	304
For Further Exploration	308
22. Network Analysis.....	309
Betweenness Centrality	309
Eigenvector Centrality	314
Matrix Multiplication	314
Centrality	316
Directed Graphs and PageRank	318
For Further Exploration	320
23. Recommender Systems.....	321
Manual Curation	322

Recommending What's Popular	322
User-Based Collaborative Filtering	323
Item-Based Collaborative Filtering	326
Matrix Factorization	328
For Further Exploration	333
24. Databases and SQL	335
CREATE TABLE and INSERT	335
UPDATE	338
DELETE	339
SELECT	340
GROUP BY	342
ORDER BY	345
JOIN	346
Subqueries	348
Indexes	349
Query Optimization	349
NoSQL	350
For Further Exploration	350
25. MapReduce	351
Example: Word Count	352
Why MapReduce?	353
MapReduce More Generally	354
Example: Analyzing Status Updates	355
Example: Matrix Multiplication	357
An Aside: Combiners	359
For Further Exploration	359
26. Data Ethics	361
What Is Data Ethics?	361
No, Really, What Is Data Ethics?	362
Should I Care About Data Ethics?	362
Building Bad Data Products	363
Trading Off Accuracy and Fairness	364
Collaboration	365
Interpretability	365
Recommendations	366
Biased Data	367
Data Protection	368
In Summary	368
For Further Exploration	369

27. Go Forth and Do Data Science.....	371
IPython	371
Mathematics	372
Not from Scratch	372
NumPy	372
pandas	372
scikit-learn	373
Visualization	373
R	373
Deep Learning	374
Find Data	374
Do Data Science	375
Hacker News	375
Fire Trucks	375
T-Shirts	375
Tweets on a Globe	376
And You?	376
Index.....	377