SECOND EDITION

Data Science from Scratch

First Principles with Python

Joel Grus

Beijing • Boston • Farnham • Sebastopol • Tokyo

O'REILLY®

Data Science from Scratch
by Joel Grus

Copyright © 2019 Joel Grus. All rights reserved.

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://oreilly.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Editor: Michele Cronin
Production Editor: Deborah Baker
Copy Editor: Rachel Monaghan
Proofreader: Rachel Head
Indexer: Judy McConville
Interior Designer: David Futato
Cover Designer: Karen Montgomery
Illustrator: Rebecca Demarest

April 2015:	First Edition
May 2019:	Second Edition

Revision History for the Second Edition

2019-04-10:	First Release

See *http://oreilly.com/catalog/errata.csp?isbn=9781492041139* for release details.

The O'Reilly logo is a registered trademark of O'Reilly Media, Inc. *Data Science from Scratch*, Second Edition, the cover image of a rock ptarmigan, and related trade dress are trademarks of O'Reilly Media, Inc.

While the publisher and the author have used good faith efforts to ensure that the information and instructions contained in this work are accurate, the publisher and the author disclaim all responsibility for errors or omissions, including without limitation responsibility for damages resulting from the use of or reliance on this work. Use of the information and instructions contained in this work is at your own risk. If any code samples or other technology this work contains or describes is subject to open source licenses or the intellectual property rights of others, it is your responsibility to ensure that your use thereof complies with such licenses and/or rights.

978-1-492-04113-9

[LSI]

Table of Contents

Preface to the Second Edition. xi

Preface to the First Edition. xv

1. Introduction. 1

The Ascendance 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 10
Topics of Interest 11
Onward 12

2. A Crash Course in Python. 13

The Zen of Python 13
Getting Python 14
Virtual Environments 14
Whitespace Formatting 15
Modules 17
Functions 17
Strings 18
Exceptions 19
Lists 20
Tuples 21
Dictionaries 22
defaultdict 23

Counters 24
Sets 24
Control Flow 25
Truthiness 26
Sorting 27
List Comprehensions 27
Automated Testing and assert 28
Object-Oriented Programming 29
Iterables and Generators 31
Randomness 32
Regular Expressions 33
Functional Programming 34
zip and Argument Unpacking 34
args and kwargs 35
Type Annotations 36
How to Write Type Annotations 38
Welcome to DataSciencester! 40
For Further Exploration 40

3. Visualizing Data. 41
matplotlib 41
Bar Charts 43
Line Charts 46
Scatterplots 47
For Further Exploration 49

4. Linear Algebra. 51
Vectors 51
Matrices 55
For Further Exploration 58

5. Statistics. 59
Describing a Single Set of Data 59
Central Tendencies 61
Dispersion 63
Correlation 64
Simpson’s Paradox 67
Some Other Correlational Caveats 68
Correlation and Causation 69
For Further Exploration 69

6. Probability. 71
Dependence and Independence 71
Conditional Probability 72
Bayes's Theorem 74
Random Variables 75
Continuous Distributions 76
The Normal Distribution 77
The Central Limit Theorem 80
For Further Exploration 82

7. Hypothesis and Inference. 83
Statistical Hypothesis Testing 83
Example: Flipping a Coin 83
p-Values 86
Confidence Intervals 88
p-Hacking 89
Example: Running an A/B Test 90
Bayesian Inference 91
For Further Exploration 94

8. Gradient Descent. 95
The Idea Behind Gradient Descent 95
Estimating the Gradient 96
Using the Gradient 99
Choosing the Right Step Size 100
Using Gradient Descent to Fit Models 100
Minibatch and Stochastic Gradient Descent 102
For Further Exploration 103

9. Getting Data. 105
stdin and stdout 105
Reading Files 107
  The Basics of Text Files 107
  Delimited Files 108
Scraping the Web 110
  HTML and the Parsing Thereof 110
  Example: Keeping Tabs on Congress 112
Using APIs 115
  JSON and XML 115
  Using an Unauthenticated API 116
  Finding APIs 117
Example: Using the Twitter APIs 117

Getting Credentials 118
For Further Exploration 122

10. Working with Data. 123
Exploring Your Data 123
Exploring One-Dimensional Data 123
Two Dimensions 125
Many Dimensions 127
Using NamedTuples 129
Dataclasses 131
Cleaning and Munging 132
Manipulating Data 134
Rescaling 136
An Aside: tqdm 138
Dimensionality Reduction 140
For Further Exploration 145

11. Machine Learning. 147
Modeling 147
What Is Machine Learning? 148
Overfitting and Underfitting 149
Correctness 151
The Bias-Variance Tradeoff 154
Feature Extraction and Selection 155
For Further Exploration 157

12. k-Nearest Neighbors. 159
The Model 159
Example: The Iris Dataset 161
The Curse of Dimensionality 164
For Further Exploration 168

13. Naive Bayes. 169
A Really Dumb Spam Filter 169
A More Sophisticated Spam Filter 170
Implementation 172
Testing Our Model 174
Using Our Model 175
For Further Exploration 177

14. Simple Linear Regression. 179
The Model 179

Using Gradient Descent 183
Maximum Likelihood Estimation 184
For Further Exploration 184

15. Multiple Regression. 185
The Model 185
Further Assumptions of the Least Squares Model 186
Fitting the Model 187
Interpreting the Model 189
Goodness of Fit 190
Digression: The Bootstrap 190
Standard Errors of Regression Coefficients 192
Regularization 194
For Further Exploration 196

16. Logistic Regression. 197
The Problem 197
The Logistic Function 200
Applying the Model 202
Goodness of Fit 203
Support Vector Machines 204
For Further Investigation 208

17. Decision Trees. 209
What Is a Decision Tree? 209
Entropy 211
The Entropy of a Partition 213
Creating a Decision Tree 214
Putting It All Together 217
Random Forests 219
For Further Exploration 220

18. Neural Networks. 221
Perceptrons 221
Feed-Forward Neural Networks 224
Backpropagation 227
Example: Fizz Buzz 229
For Further Exploration 232

19. Deep Learning. 233
The Tensor 233
The Layer Abstraction 236

The Linear Layer 237
Neural Networks as a Sequence of Layers 240
Loss and Optimization 241
Example: XOR Revisited 244
Other Activation Functions 245
Example: FizzBuzz Revisited 246
Softmaxes and Cross-Entropy 247
Dropout 250
Example: MNIST 250
Saving and Loading Models 255
For Further Exploration 256

20. Clustering. 257
The Idea 257
The Model 258
Example: Meetups 260
Choosing k 262
Example: Clustering Colors 263
Bottom-Up Hierarchical Clustering 265
For Further Exploration 271

21. Natural Language Processing. 273
Word Clouds 273
n-Gram Language Models 275
Grammars 278
An Aside: Gibbs Sampling 280
Topic Modeling 282
Word Vectors 287
Recurrent Neural Networks 295
Example: Using a Character-Level RNN 298
For Further Exploration 301

22. Network Analysis. 303
Betweenness Centrality 303
Eigenvector Centrality 308
Matrix Multiplication 308
Centrality 310
Directed Graphs and PageRank 312
For Further Exploration 314

23. Recommender Systems. 315
Manual Curation 316

Recommending What's Popular 316
User-Based Collaborative Filtering 317
Item-Based Collaborative Filtering 320
Matrix Factorization 322
For Further Exploration 327

24. Databases and SQL. 329
CREATE TABLE and INSERT 329
UPDATE 332
DELETE 333
SELECT 334
GROUP BY 336
ORDER BY 339
JOIN 340
Subqueries 342
Indexes 343
Query Optimization 343
NoSQL 344
For Further Exploration 344

25. MapReduce. 345
Example: Word Count 346
Why MapReduce? 347
MapReduce More Generally 348
Example: Analyzing Status Updates 349
Example: Matrix Multiplication 351
An Aside: Combiners 353
For Further Exploration 353

26. Data Ethics. 355
What Is Data Ethics? 355
No, Really, What Is Data Ethics? 356
Should I Care About Data Ethics? 356
Building Bad Data Products 357
Trading Off Accuracy and Fairness 357
Collaboration 359
Interpretability 359
Recommendations 360
Biased Data 361
Data Protection 362
In Summary 362
For Further Exploration 362

27. Go Forth and Do Data Science. 363
IPython 363
Mathematics 364
Not from Scratch 364
NumPy 364
pandas 364
scikit-learn 365
Visualization 365
R 365
Deep Learning 366
Find Data 366
Do Data Science 366
Hacker News 367
Fire Trucks 367
T-Shirts 367
Tweets on a Globe 368
And You? 368

Index. 369

Preface to the Second Edition

I am exceptionally proud of the first edition of *Data Science from Scratch*. It turned out very much the book I wanted it to be. But several years of developments in data science, of progress in the Python ecosystem, and of personal growth as a developer and educator have *changed* what I think a first book in data science should look like.

In life, there are no do-overs. In writing, however, there are second editions.

Accordingly, I've rewritten all the code and examples using Python 3.6 (and many of its newly introduced features, like type annotations). I've woven into the book an emphasis on writing clean code. I've replaced some of the first edition's toy examples with more realistic ones using "real" datasets. I've added new material on topics such as deep learning, statistics, and natural language processing, corresponding to things that today's data scientists are likely to be working with. (I've also removed some material that seems less relevant.) And I've gone over the book with a fine-toothed comb, fixing bugs, rewriting explanations that are less clear than they could be, and freshening up some of the jokes.

The first edition was a great book, and this edition is even better. Enjoy!

> Joel Grus
> Seattle, WA
> 2019

Conventions Used in This Book

The following typographical conventions are used in this book:

Italic
: Indicates new terms, URLs, email addresses, filenames, and file extensions.