

Data Science From Scratch: First Principles with Python

I am super-excited to announce that the book I've been working on for more than the last year, *Data Science from Scratch: First Principles with Python* is finally available! (buy from [O'Reilly](#), use discount code AUTHD to save some money) (buy from [Amazon](#)).

My experience learning and teaching data science was that there were two primary paths:

1. The Math Path: "So you want to be a data scientist? Sure, the first thing you need to know is *matrix decompositions*. How well do you remember your measure theory?"
2. The Tools Path: "So you want to be a data scientist? Great, here's the most important libraries to know. How well do you know R?"

Although I am myself a "math person", the first approach never resonated with me. The fun of data science for me has always been *working with data*. At the same time, I've never been thrilled with the second approach -- it's a good way to start doing data science without ever really understanding what you're doing.

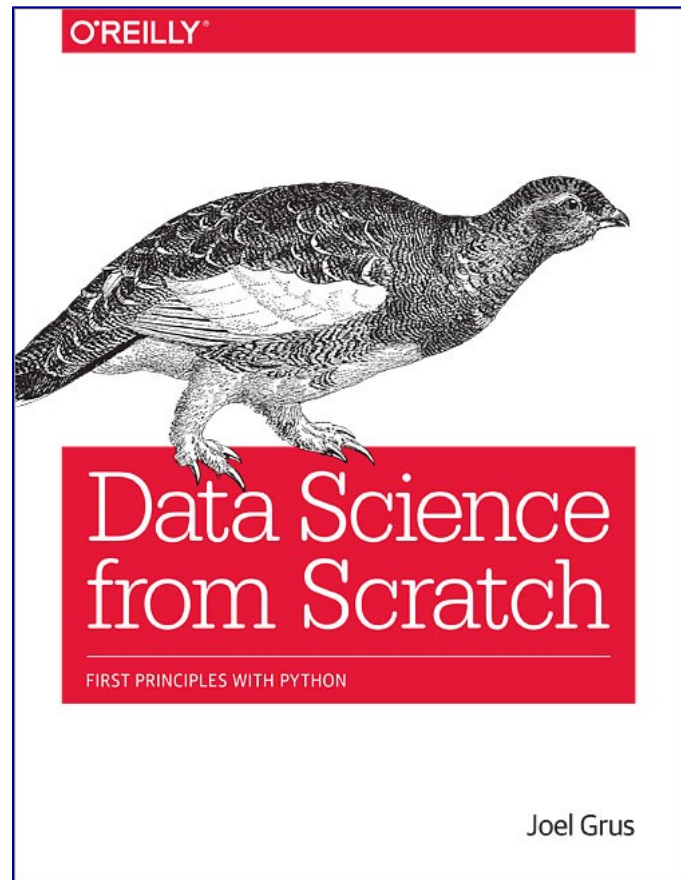
My ideal would be a "[third way](#)" between these approaches:

1. understanding the behavior of the most common tools by working through a solid-but-less-than-textbook-rigorous understanding of the math behind them, and
2. implementing simplified versions of them *from scratch* to understand exactly what it is they're doing.

Hence *Data Science from Scratch*. It's got math, but only as much as is totally necessary. It's got scraping and cleaning and munging. It's got machine learning. It's got databases and MapReduce. Necessarily it doesn't go *deep* into any of these, but I like to think it establishes a broad, solid foundation for someone who knows some math and some programming but is not (necessarily) an expert at either.

Many technical books (I won't name names) explain things in their text and then dump pages of hard-to-follow code at you that you are expected to puzzle through. I spent a lot of time trying to write clean code that illuminated the concepts *on its own* and that *reinforced* the ideas from the text. As is the current fashion these days, all of the code and data is [on GitHub](#), if you'd like to get a sense of what the book is about.

If you are interested in the topic, I encourage you to check it out, write a review, and let me know what you



think! (You can see the full table of contents on the [O'Reilly page](#).)

Posted on: 2015-04-26

Category: [Life, Data Science, Writing](#)

Data Science from Scratch
First Principles with Python

By [Joel Grus](#)

Publisher: O'Reilly Media

Final Release Date: April 2015

Pages: 330

Data science libraries, frameworks, modules, and toolkits are great for doing data science, but they're also a good way to dive into the discipline without actually understanding data science. In this book, you'll learn how many of the most fundamental data science tools and algorithms work by implementing them *from scratch*.

If you have an aptitude for mathematics and some programming skills, author Joel Grus will help you get comfortable with the math and statistics at the core of data science, and with hacking skills you need to get started as a data scientist. Today's messy glut of data holds answers to questions no one's even thought to ask. This book provides you with the know-how to dig those answers out.

- Get a crash course in Python
- Learn the basics of linear algebra, statistics, and probability—and understand how and when they're used in data science
- Collect, explore, clean, munge, and manipulate data
- Dive into the fundamentals of machine learning
- Implement models such as k-nearest Neighbors, Naive Bayes, linear and logistic regression, decision trees, neural networks, and clustering
- Explore recommender systems, natural language processing, network analysis, MapReduce, and databases

Chapter 1 Introduction

The Ascendance of Data

What Is Data Science?

Motivating Hypothetical: DataSciencecenter

Chapter 2 A Crash Course in Python

The Basics

The Not-So-Basics

For Further Exploration

Chapter 3 Visualizing Data

matplotlib

Bar Charts

Line Charts

Scatterplots

For Further Exploration

Chapter 4 Linear Algebra

Vectors

Matrices

For Further Exploration

Chapter 5 Statistics

Describing a Single Set of Data

Correlation

Simpson's Paradox

Some Other Correlational Caveats

Correlation and Causation

For Further Exploration

Chapter 6 Probability

Dependence and Independence

Conditional Probability

Bayes's Theorem
Random Variables
Continuous Distributions
The Normal Distribution
The Central Limit Theorem
For Further Exploration

Chapter 7 Hypothesis and Inference

Statistical Hypothesis Testing
Example: Flipping a Coin
Confidence Intervals
P-hacking
Example: Running an A/B Test
Bayesian Inference
For Further Exploration

Chapter 8 Gradient Descent

The Idea Behind Gradient Descent
Estimating the Gradient
Using the Gradient
Choosing the Right Step Size
Putting It All Together
Stochastic Gradient Descent
For Further Exploration

Chapter 9 Getting Data

stdin and stdout
Reading Files
Scraping the Web
Using APIs
Example: Using the Twitter APIs
For Further Exploration

Chapter 10 Working with Data

Exploring Your Data
Cleaning and Munging
Manipulating Data

Rescaling
Dimensionality Reduction
For Further Exploration

Chapter 11 Machine Learning

Modeling
What Is Machine Learning?
Overfitting and Underfitting
Correctness
The Bias-Variance Trade-off
Feature Extraction and Selection
For Further Exploration

Chapter 12 k-Nearest Neighbors

The Model
Example: Favorite Languages
The Curse of Dimensionality
For Further Exploration

Chapter 13 Naive Bayes

A Really Dumb Spam Filter
A More Sophisticated Spam Filter
Implementation
Testing Our Model
For Further Exploration

Chapter 14 Simple Linear Regression

The Model
Using Gradient Descent
Maximum Likelihood Estimation
For Further Exploration

Chapter 15 Multiple Regression

The Model
Further Assumptions of the Least Squares Model
Fitting the Model

Interpreting the Model
Goodness of Fit
Digression: The Bootstrap
Standard Errors of Regression Coefficients
Regularization
For Further Exploration

Chapter 16 Logistic Regression

The Problem
The Logistic Function
Applying the Model
Goodness of Fit
Support Vector Machines
For Further Investigation

Chapter 17 Decision Trees

What Is a Decision Tree?
Entropy
The Entropy of a Partition
Creating a Decision Tree
Putting It All Together
Random Forests
For Further Exploration

Chapter 18 Neural Networks

Perceptrons
Feed-Forward Neural Networks
Backpropagation
Example: Defeating a CAPTCHA
For Further Exploration

Chapter 19 Clustering

The Idea
The Model
Example: Meetups
Choosing k
Example: Clustering Colors

Bottom-up Hierarchical Clustering
For Further Exploration

Chapter 20 Natural Language Processing

Word Clouds
n-gram Models
Grammars
An Aside: Gibbs Sampling
Topic Modeling
For Further Exploration

Chapter 21 Network Analysis

Betweenness Centrality
Eigenvector Centrality
Directed Graphs and PageRank
For Further Exploration

Chapter 22 Recommender Systems

Manual Curation
Recommending What's Popular
User-Based Collaborative Filtering
Item-Based Collaborative Filtering
For Further Exploration

Chapter 23 Databases and SQL

CREATE TABLE and INSERT
UPDATE
DELETE
SELECT
GROUP BY
ORDER BY
JOIN
Subqueries
Indexes
Query Optimization
NoSQL
For Further Exploration

Chapter 24 MapReduce

Example: Word Count

Why MapReduce?

MapReduce More Generally

Example: Analyzing Status Updates

Example: Matrix Multiplication

An Aside: Combiners

For Further Exploration

Chapter 25 Go Forth and Do Data Science

IPython

Mathematics

Not from Scratch

Find Data

Do Data Science