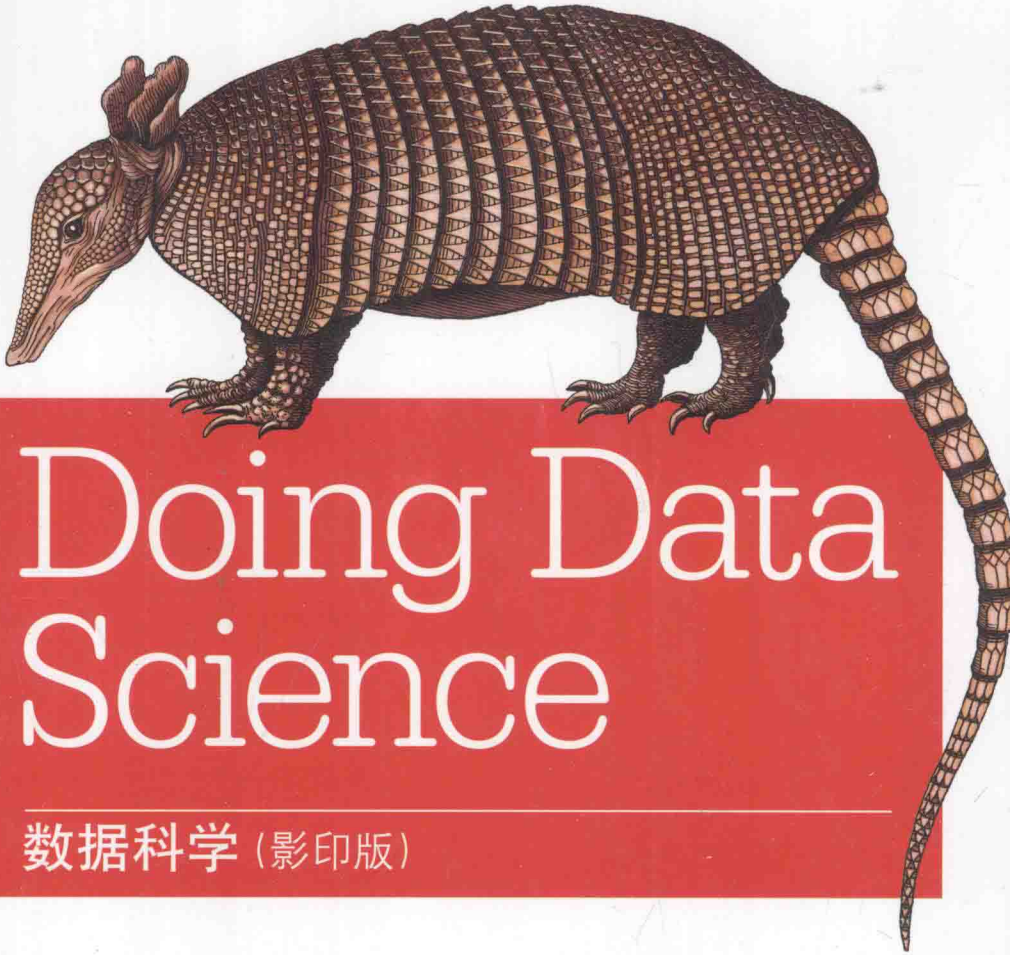
O'REILLY®



Doing Data Science

数据科学（影印版）

東南大學出版社

Rachel Schutt & Cathy O'Neil 著

数据科学（影印版）

Doing Data Science

Rachel Schutt & Cathy O'Neil 著

Beijing · Cambridge · Farnham · Köln · Sebastopol · Tokyo O'REILLY®

O'Reilly Media, Inc.授权东南大学出版社出版

南京 东南大学出版社

图书在版编目（CIP）数据

数据科学：英文/（美）舒特（Schutt, R.），（美）奥尼尔（O'Neil, C.）著. —影印本. —南京：东南大学出版社，2014.9

书名原文：Doing Data Science

ISBN 978-7-5641-4984-0

I. ①数… II. ①舒… ②奥… III. ①数据管理－英文 IV. ①TP274

中国版本图书馆CIP数据核字（2014）第102968号

江苏省版权局著作权合同登记

图字：10-2014-144号

数据科学（影印版）

出版发行：东南大学出版社

地　　址：南京四牌楼2号　　邮编：210096

出 版 人：江建中

网　　址：http://www.seupress.com

电子邮件：press@seupress.com

印　　刷：常州市武进第三印刷有限公司

开　　本：787毫米×980毫米　16开本

印　　张：25.25

字　　数：412千字

版　　次：2014年9月第1版

印　　次：2014年9月第1次印刷

书　　号：ISBN 978-7-5641-4984-0

定　　价：74.00元

本社图书若有印装质量问题，请直接与营销部联系。电话（传真）：025-83791830

In loving memory of Kelly Feeney.

Preface

Rachel Schutt

Data science is an emerging field in industry, and as yet, it is not well-defined as an academic subject. This book represents an ongoing investigation into the central question: "What is data science?" It's based on a class called "Introduction to Data Science," which I designed and taught at Columbia University for the first time in the Fall of 2012.

In order to understand this book and its origins, it might help you to understand a little bit about me and what my motivations were for creating the class.

Motivation

In short, I created a course that I wish had existed when I was in college, but that was the 1990s, and we weren't in the midst of a data explosion, so the class couldn't have existed back then. I was a math major as an undergraduate, and the track I was on was theoretical and proof-oriented. While I am glad I took this path, and feel it trained me for rigorous problem-solving, I would have also liked to have been exposed then to ways those skills could be put to use to solve real-world problems.

I took a wandering path between college and a PhD program in statistics, struggling to find my field and place—a place where I could put my love of finding patterns and solving puzzles to good use. I bring this up because many students feel they need to know what they are "going to do with their lives" now, and when I was a student, I couldn't plan to work in data science as it wasn't even yet a field. My advice to students (and anyone else who cares to listen): you don't need to figure it all out now. It's OK to take a wandering path. Who knows what you

might find? After I got my PhD, I worked at Google for a few years around the same time that "data science" and "data scientist" were becoming terms in Silicon Valley.

The world is opening up with possibilities for people who are quantitatively minded and interested in putting their brains to work to solve the world's problems. I consider it my goal to help these students to become critical thinkers, creative solvers of problems (even those that have not yet been identified), and curious question askers. While I myself may never build a mathematical model that is a piece of the cure for cancer, or identifies the underlying mystery of autism, or that prevents terrorist attacks, I like to think that I'm doing my part by teaching students who might one day do these things. And by writing this book, I'm expanding my reach to an even wider audience of data scientists who I hope will be inspired by this book, or learn tools in it, to make the world better and not worse.

Building models and working with data is not value-neutral. You choose the problems you will work on, you make assumptions in those models, you choose metrics, and you design the algorithms.

The solutions to all the world's problems may not lie in data and technology—and in fact, the mark of a good data scientist is someone who can identify problems that *can* be solved with data and is well-versed in the tools of modeling and code. But I do believe that interdisciplinary teams of people that include a data-savvy, quantitatively minded, coding-literate problem-solver (let's call that person a "data scientist") could go a long way.

Origins of the Class

I proposed the class in March 2012. At the time, there were three primary reasons. The first will take the longest to explain.

Reason 1: I wanted to give students an education in what it's like to be a data scientist in industry and give them some of the skills data scientists have.

I was working on the Google+ data science team with an interdisciplinary team of PhDs. There was me (a statistician), a social scientist, an engineer, a physicist, and a computer scientist. We were part of a larger team that included talented data engineers who built the data pipelines, infrastructure, and dashboards, as well as built the experimental infrastructure (A/B testing). Our team had a flat structure.

Together our skills were powerful, and we were able to do amazing things with massive datasets, including predictive modeling, prototyping algorithms, and unearthing patterns in the data that had huge impact on the product.

We provided leadership with insights for making data-driven decisions, while also developing new methodologies and novel ways to understand causality. Our ability to do this was dependent on top-notch engineering and infrastructure. We each brought a solid mix of skills to the team, which together included coding, software engineering, statistics, mathematics, machine learning, communication, visualization, exploratory data analysis (EDA), data sense, and intuition, as well as expertise in social networks and the social space.

To be clear, no one of us excelled at all those things, but together we did; we recognized the value of all those skills, and that's why we thrived. What we had in common was integrity and a genuine interest in solving interesting problems, always with a healthy blend of skepticism as well as a sense of excitement over scientific discovery. We cared about what we were doing and loved unearthing patterns in the data.

I live in New York and wanted to bring my experience at Google back to students at Columbia University because I believe this is stuff they need to know, and because I enjoy teaching. I wanted to teach them what I had learned on the job. And I recognized that there was an emerging data scientist community in the New York tech scene, and I wanted students to hear from them as well.

One aspect of the class was that we had guest lectures by data scientists currently working in industry and academia, each of whom had a different mix of skills. We heard a diversity of perspectives, which contributed to a holistic understanding of data science.

Reason 2: Data science has the potential to be a deep and profound research discipline impacting all aspects of our lives. Columbia University and Mayor Bloomberg announced the Institute for Data Sciences and Engineering in July 2012. This course created an opportunity to develop the theory of data science and to formalize it as a legitimate science.

Reason 3: I kept hearing from data scientists in industry that you can't teach data science in a classroom or university setting, and I took that on as a challenge. I thought of my classroom as an incubator of data

science teams. The students I had were very impressive and are turning into top-notch data scientists. They've contributed a chapter to this book, in fact.

Origins of the Book

The class would not have become a book if I hadn't met Cathy O'Neil, a mathematician-turned-data scientist and prominent and outspoken blogger on *mathbabe.org*, where her "About" section states that she hopes to someday have a better answer to the question, "What can a nonacademic mathematician do that makes the world a better place?" Cathy and I met around the time I proposed the course and she was working as a data scientist at a startup. She was encouraging and supportive of my efforts to create the class, and offered to come and blog it. Given that I'm a fairly private person, I initially did not feel comfortable with this idea. But Cathy convinced me by pointing out that this was an opportunity to put ideas about data science into the public realm as a voice running counter to the marketing and hype that is going on around data science.

Cathy attended every class and sat in the front row asking questions, and was also a guest lecturer (see Chapter 6). As well as documenting the class on her blog, she made valuable intellectual contributions to the course content, including reminding us of the ethical components of modeling. She encouraged me to blog as well, and so in parallel to her documenting the class, I maintained a blog (*http://columbiadatascience.com/blog/*) to communicate with my students directly, as well as capture the experience of teaching data science in the hopes it would be useful to other professors. All Cathy's blog entries for the course, and some of mine, became the raw material for this book. We've added additional material and revised and edited and made it much more robust than the blogs, so now it's a full-fledged book.

What to Expect from This Book

In this book, we want to both describe and prescribe. We want to *describe* the current state of data science by observing a set of top-notch thinkers describe their jobs and what it's like to "do data science." We also want to *prescribe* what data science could be as an academic discipline.

Don't expect a machine learning textbook. Instead, expect full immersion into the multifaceted aspects of data science from multiple points of view. This is a survey of the existing landscape of data science—an attempt to map this emerging field—and as a result, there is more breadth in some cases than depth.

This book is written with the hope that it will find itself into the hands of someone—you?—who will make even more of it than what it is, and go on to solve important problems.

After the class was over, I heard it characterized as a holistic, humanist approach to data science—we did not just focus on the tools, math, models, algorithms, and code, but on the human side as well. I like this definition of humanist: "a person having a strong interest in or concern for human welfare, values, and dignity." Being humanist in the context of data science means recognizing the role your own humanity plays in building models and algorithms, thinking about qualities you have as a human that a computer does not have (which includes the ability to make ethical decisions), and thinking about the humans whose lives you are impacting when you unleash a model onto the world.

How This Book Is Organized

This book is organized in the same order as the class. We'll begin with some introductory material on the central question, "What is data science?" and introduce the data science process as an organizing principle. In Chapters 2 and 3, we'll begin with an overview of statistical modeling and machine learning algorithms as a foundation for the rest of the book. Then in Chapters 4–6 and 8 we'll get into specific examples of models and algorithms in context. In Chapter 7 we'll hear about how to extract meaning from data and create features to incorporate into the models. Chapters 9 and 10 involve two of the areas not traditionally taught (but this is changing) in academia: data visualization and social networks. We'll switch gears from prediction to causality in Chapters 11 and 12. Chapters 13 and 14 will be about data preparation and engineering. Chapter 15 lets us hear from the students who took the class about what it was like to learn data science, and then we will end by telling you in Chapter 16 about what we hope for the future of data science.

How to Read This Book

Generally speaking, this book will make more sense if you read it straight through in a linear fashion because many of the concepts build on one another. It's also possible that you will need to read this book with supplemental material if you have holes in your probability and statistics background, or you've never coded before. We've tried to give suggestions throughout the book for additional reading. We hope that when you don't understand something in the book, perhaps because of gaps in your background, or inadequate explanation on our part, that you will take this moment of confusion as an opportunity to investigate the concepts further.

How Code Is Used in This Book

This isn't a how-to manual, so code is used to provide examples, but in many cases, it might require you to implement it yourself and play around with it to truly understand it.

Who This Book Is For

Because of the media coverage around data science and the characterization of data scientists as "rock stars," you may feel like it's impossible for you to enter into this realm. If you're the type of person who loves to solve puzzles and find patterns, whether or not you consider yourself a quant, then data science is for you.

This book is meant for people coming from a wide variety of backgrounds. We hope and expect that different people will get different things out of it depending on their strengths and weaknesses.

- Experienced data scientists will perhaps come to see and understand themselves and what they do in a new light.
- Statisticians may gain an appreciation of the relationship between data science and statistics. Or they may continue to maintain the attitude, "that's just statistics," in which case we'd like to see that argument clearly articulated.
- Quants, math, physics, or other science PhDs who are thinking about transitioning to data science or building up their data science skills will gain perspective on what that would require or mean.

- Students and those new to data science will be getting thrown into the deep end, so if you don't understand everything all the time, don't worry; that's part of the process.
- Those who have never coded in R or Python before will want to have a manual for learning R or Python. We recommend *The Art of R Programming* by Norman Matloff (No Starch Press). Students who took the course also benefitted from the expert instruction of lab instructor, Jared Lander, whose book *R for Everyone: Advanced Analytics and Graphics* (Addison-Wesley) is scheduled to come out in November 2013. It's also possible to do all the exercises using packages in Python.
- For those who have never coded at all before, the same advice holds. You might also want to consider picking up *Learning Python* by Mark Lutz and David Ascher (O'Reilly) or Wes McKinney's *Python for Data Analysis* (also O'Reilly) as well.

Prerequisites

We assume prerequisites of linear algebra, some probability and statistics, and some experience coding in any language. Even so, we will try to make the book as self-contained as possible, keeping in mind that it's up to you to do supplemental reading if you're missing some of that background. We'll try to point out places throughout the book where supplemental reading might help you gain a deeper understanding.

Supplemental Reading

This book is an overview of the landscape of a new emerging field with roots in many other disciplines: statistical inference, algorithms, statistical modeling, machine learning, experimental design, optimization, probability, artificial intelligence, data visualization, and exploratory data analysis. The challenge in writing this book has been that each of these disciplines corresponds to several academic courses or books in their own right. There may be times when gaps in the reader's prior knowledge require supplemental reading.

Math

- *Linear Algebra and Its Applications* by Gilbert Strang (Cengage Learning)

- *Convex Optimization* by Stephen Boyd and Lieven Vandenberghe (Cambridge University Press)
- *A First Course in Probability* (Pearson) and *Introduction to Probability Models* (Academic Press) by Sheldon Ross

Coding

- *R in a Nutshell* by Joseph Adler (O'Reilly)
- *Learning Python* by Mark Lutz and David Ascher (O'Reilly)
- *R for Everyone: Advanced Analytics and Graphics* by Jared Lander (Addison-Wesley)
- *The Art of R Programming: A Tour of Statistical Software Design* by Norman Matloff (No Starch Press)
- *Python for Data Analysis* by Wes McKinney (O'Reilly)

Data Analysis and Statistical Inference

- *Statistical Inference* by George Casella and Roger L. Berger (Cengage Learning)
- *Bayesian Data Analysis* by Andrew Gelman, et al. (Chapman & Hall)
- *Data Analysis Using Regression and Multilevel/Hierarchical Models* by Andrew Gelman and Jennifer Hill (Cambridge University Press)
- *Advanced Data Analysis from an Elementary Point of View* by Cosma Shalizi (under contract with Cambridge University Press)
- *The Elements of Statistical Learning: Data Mining, Inference and Prediction* by Trevor Hastie, Robert Tibshirani, and Jerome Friedman (Springer)

Artificial Intelligence and Machine Learning

- *Pattern Recognition and Machine Learning* by Christopher Bishop (Springer)
- *Bayesian Reasoning and Machine Learning* by David Barber (Cambridge University Press)
- *Programming Collective Intelligence* by Toby Segaran (O'Reilly)
- *Artificial Intelligence: A Modern Approach* by Stuart Russell and Peter Norvig (Prentice Hall)

- *Foundations of Machine Learning* by Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar (MIT Press)
- *Introduction to Machine Learning (Adaptive Computation and Machine Learning)* by Ethem Alpaydim (MIT Press)

Experimental Design

- *Field Experiments* by Alan S. Gerber and Donald P. Green (Norton)
- *Statistics for Experimenters: Design, Innovation, and Discovery* by George E. P. Box, et al. (Wiley-Interscience)

Visualization

- *The Elements of Graphing Data* by William Cleveland (Hobart Press)
- *Visualize This: The FlowingData Guide to Design, Visualization, and Statistics* by Nathan Yau (Wiley)

About the Contributors

The course would not have been a success without the many guest lecturers that came to speak to the class. While I gave some of the lectures, a large majority were given by guests from startups and tech companies, as well as professors from Columbia University. Most chapters in this book are based on those lectures. While generally speaking the contributors did not write the book, they contributed many of the ideas and content of the book, reviewed their chapters and offered feedback, and we're grateful to them. The class and book would not have existed without them. I invited them to speak in the class because I hold them up as role models for aspiring data scientists.

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.

`Constant width`
: Used for program listings, as well as within paragraphs to refer to program elements such as variable or function names, databases, data types, environment variables, statements, and keywords.

`Constant width bold`
: Shows commands or other text that should be typed literally by the user.

`Constant width italic`
: Shows text that should be replaced with user-supplied values or by values determined by context.

This icon signifies a tip, suggestion, or general note.

This icon indicates a warning or caution.

Using Code Examples

Supplemental material (datasets, exercises, etc.) is available for download at *https://github.com/oreillymedia/doing_data_science*.

This book is here to help you get your job done. In general, if example code is offered with this book, you may use it in your programs and documentation. You do not need to contact us for permission unless you're reproducing a significant portion of the code. For example, writing a program that uses several chunks of code from this book does not require permission. Selling or distributing a CD-ROM of examples from O'Reilly books does require permission. Answering a question by citing this book and quoting example code does not require permission. Incorporating a significant amount of example code from this book into your product's documentation does require permission.

We appreciate, but do not require, attribution. An attribution usually includes the title, author, publisher, and ISBN. For example: "*Doing Data Science* by Rachel Schutt and Cathy O'Neil (O'Reilly). Copyright 2014 Rachel Schutt and Cathy O'Neil, 978-1-449-35865-5."

If you feel your use of code examples falls outside fair use or the permission given above, feel free to contact us at *permissions@oreilly.com*.

Safari® Books Online

Safari Books Online is an on-demand digital library that delivers expert content in both book and video form from the world's leading authors in technology and business.

Technology professionals, software developers, web designers, and business and creative professionals use Safari Books Online as their primary resource for research, problem solving, learning, and certification training.

Safari Books Online offers a range of product mixes and pricing programs for organizations, government agencies, and individuals. Subscribers have access to thousands of books, training videos, and prepublication manuscripts in one fully searchable database from publishers like O'Reilly Media, Prentice Hall Professional, Addison-Wesley Professional, Microsoft Press, Sams, Que, Peachpit Press, Focal Press, Cisco Press, John Wiley & Sons, Syngress, Morgan Kaufmann, IBM Redbooks, Packt, Adobe Press, FT Press, Apress, Manning, New Riders, McGraw-Hill, Jones & Bartlett, Course Technology, and dozens more. For more information about Safari Books Online, please visit us online.

How to Contact Us

Please address comments and questions concerning this book to the publisher:

O'Reilly Media, Inc.
1005 Gravenstein Highway North
Sebastopol, CA 95472
800-998-9938 (in the United States or Canada)
707-829-0515 (international or local)
707-829-0104 (fax)

We have a web page for this book, where we list errata, examples, and any additional information. You can access this page at *http://oreil.ly/doing_data_science*.

To comment or ask technical questions about this book, send email to *bookquestions@oreilly.com*.

For more information about our books, courses, conferences, and news, see our website at *http://www.oreilly.com*.

Find us on Facebook: *http://facebook.com/oreilly*

Follow us on Twitter: *http://twitter.com/oreillymedia*

Watch us on YouTube: *http://www.youtube.com/oreillymedia*

Acknowledgments

Rachel would like to thank her Google influences: David Huffaker, Makoto Uchida, Andrew Tomkins, Abhijit Bose, Daryl Pregibon, Diane Lambert, Josh Wills, David Crawshaw, David Gibson, Corinna Cortes, Zach Yeskel, and Gueorgi Kossinetts. From the Columbia statistics department: Andrew Gelman and David Madigan; and the lab instructor and teaching assistant for the course, Jared Lander and Ben Reddy.

Rachel appreciates the loving support of family and friends, especially Eran Goldshtein, Barbara and Schutt, Becky, Susie and Alex, Nick, Lilah, Belle, Shahed, and the Feeneys.

Cathy would like to thank her family and friends, including her wonderful sons and husband, who let her go off once a week to blog the evening class.

We both would like to thank:

- The brain trust that convened in Cathy's apartment: Chris Wiggins, David Madigan, Mark Hansen, Jake Hofman, Ori Stitelman, and Brian Dalessandro.
- Our editors, Courtney Nash and Mike Loukides.
- The participants and organizers of the IMA User-level modeling conference where some preliminary conversations took place.
- The students!
- Coppelia, where Cathy and Rachel met for breakfast a lot.

We'd also like to thank John Johnson and David Park of Johnson Research Labs for their generosity and the luxury of time to spend writing this book.

Table of Contents