O'REILLY®

Doing Data Science

STRAIGHT TALK FROM THE FRONTLINE

Cathy O'Neil & Rachel Schutt

O'REILLY®

Doing Data Science

Now that people are aware that data can make the difference in an election or a business model, data science as an occupation is gaining ground. But how can you get started working in a wide-ranging, interdisciplinary field that's so clouded in hype? This insightful book, based on Columbia University's Introduction to Data Science class, tells you what you need to know.

In many of these chapter-long lectures, data scientists from companies such as Google, Microsoft, and eBay share new algorithms, methods, and models by presenting case studies and the code they use. If you're familiar with linear algebra, probability, and statistics, and have programming experience, this book is an ideal introduction to data science.

Topics include:

- Statistical inference, exploratory data analysis, and the data science process
- Algorithms
- Spam filters, Naive Bayes, and data wrangling
- Logistic regression
- Financial modeling
- Recommendation engines and causality
- Data visualization
- Social networks and data journalism
- Data engineering, MapReduce, Pregel, and Hadoop

Cathy O'Neil, a senior data scientist at Johnson Research Labs, earned a Ph.D. in math from Harvard, and was a postdoc in the math department at MIT and a professor at Barnard College.

Rachel Schutt, Senior VP of Data Science at News Corp, is an adjunct professor of statistics at Columbia University, and a founding member of CU's Education Committee for the Institute for Data Sciences and Engineering.

DATABASES / DATA

US $54.99 **CAN $72.99**

ISBN: 978-1-449-35865-5

9 781449 358655 55499

Twitter: @oreillymedia
facebook.com/oreilly

Doing Data Science

Cathy O'Neil and Rachel Schutt

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

Doing Data Science

by Cathy O'Neil and Rachel Schutt

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://safaribooksonline.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Editors: Mike Loukides and Courtney Nash
Production Editor: Kristen Brown
Copyeditor: Kim Cofer
Proofreader: Amanda Kersey

Indexer: WordCo Indexing Services
Cover Designer: Karen Montgomery
Interior Designer: David Futato
Illustrator: Rebecca Demarest

October 2013: First Edition

Revision History for the First Edition:

2013-10-08: First release

2013-12-13: Second release

2014-10-10: Third release

See *http://oreilly.com/catalog/errata.csp?isbn=9781449358655* for release details.

ISBN: 978-1-449-35865-5

[LSI]

In loving memory of Kelly Feeney.

Table of Contents

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 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 than depth in some cases.

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 Alpaydin (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)
- *The Visual Display of Quantitative Information* by Edward Tufte (Graphics Press)

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 Cathy O'Neil and Rachel Schutt (O'Reilly). Copyright 2014 Cathy O'Neil and Rachel Schutt, 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 plans and pricing for enterprise, government, education, and individuals.

Members 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 hundreds 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.

CHAPTER 1

Introduction: What Is Data Science?

Over the past few years, there's been a lot of hype in the media about "data science" and "Big Data." A reasonable first reaction to all of this might be some combination of skepticism and confusion; indeed we, Cathy and Rachel, had that exact reaction.

And we let ourselves indulge in our bewilderment for a while, first separately, and then, once we met, together over many Wednesday morning breakfasts. But we couldn't get rid of a nagging feeling that there was something *real* there, perhaps something deep and profound representing a paradigm shift in our culture around data. Perhaps, we considered, it's even a paradigm shift that plays to our strengths. Instead of ignoring it, we decided to explore it more.

But before we go into that, let's first delve into what struck us as confusing and vague—perhaps you've had similar inclinations. After that we'll explain what made us get past our own concerns, to the point where Rachel created a course on data science at Columbia University, Cathy blogged the course, and you're now reading a book based on it.

Big Data and Data Science Hype

Let's get this out of the way right off the bat, because many of you are likely skeptical of data science already for many of the reasons we were. We want to address this up front to let you know: *we're right there with you*. If you're a skeptic too, it probably means you have something

useful to contribute to making data science into a more legitimate field that has the power to have a positive impact on society.

So, what is eyebrow-raising about Big Data and data science? Let's count the ways:

1. There's a lack of definitions around the most basic terminology. What is "Big Data" anyway? What does "data science" mean? What is the relationship between Big Data and data science? Is data science the science of Big Data? Is data science only the stuff going on in companies like Google and Facebook and tech companies? Why do many people refer to Big Data as crossing disciplines (astronomy, finance, tech, etc.) and to data science as only taking place in tech? Just how *big* is big? Or is it just a relative term? These terms are so ambiguous, they're well-nigh meaningless.
2. There's a distinct lack of respect for the researchers in academia and industry labs who have been working on this kind of stuff for years, and whose work is based on decades (in some cases, centuries) of work by statisticians, computer scientists, mathematicians, engineers, and scientists of all types. From the way the media describes it, machine learning algorithms were just invented last week and data was never "big" until Google came along. This is simply not the case. Many of the methods and techniques we're using—and the challenges we're facing now—are part of the evolution of everything that's come before. This doesn't mean that there's not new and exciting stuff going on, but we think it's important to show some basic respect for everything that came before.
3. The hype is crazy—people throw around tired phrases straight out of the height of the pre-financial crisis era like "Masters of the Universe" to describe data scientists, and that doesn't bode well. In general, hype masks reality and increases the noise-to-signal ratio. The longer the hype goes on, the more many of us will get turned off by it, and the harder it will be to see what's good underneath it all, if anything.
4. Statisticians already feel that they are studying and working on the "Science of Data." That's their bread and butter. Maybe you, dear reader, are not a statistican and don't care, but imagine that for the statistician, this feels a little bit like how identity theft might feel for you. Although we will make the case that data science is *not* just a rebranding of statistics or machine learning but rather

a field unto itself, the media often describes data science in a way that makes it sound like as if it's simply statistics or machine learning in the context of the tech industry.

5. People have said to us, "Anything that has to call itself a science isn't." Although there might be truth in there, that doesn't mean that the term "data science" *itself* represents nothing, but of course what it represents may not be science but more of a craft.

Getting Past the Hype

Rachel's experience going from getting a PhD in statistics to working at Google is a great example to illustrate why we thought, in spite of the aforementioned reasons to be dubious, there might be some meat in the data science sandwich. In her words:

> It was clear to me pretty quickly that the stuff I was working on at Google was different than anything I had learned at school when I got my PhD in statistics. This is not to say that my degree was useless; far from it—what I'd learned in school provided a framework and way of thinking that I relied on daily, and much of the actual content provided a solid theoretical and practical foundation necessary to do my work.
>
> But there were also many skills I had to acquire on the job at Google that I *hadn't* learned in school. Of course, my experience is specific to me in the sense that I had a statistics background and picked up more computation, coding, and visualization skills, as well as domain expertise while at Google. Another person coming in as a computer scientist or a social scientist or a physicist would have different gaps and would fill them in accordingly. But what is important here is that, as individuals, we each had different strengths and gaps, yet we were able to solve problems by putting ourselves together into a data team well-suited to solve the data problems that came our way.

Here's a reasonable response you might have to this story. It's a general truism that, whenever you go from school to a real job, you realize there's a gap between what you learned in school and what you do on the job. In other words, you were simply facing the difference between academic statistics and industry statistics.

We have a couple replies to this:

- Sure, there's is a difference between industry and academia. But does it really have to be that way? Why do many courses in school have to be so intrinsically out of touch with reality?

- Even so, the gap doesn't represent simply a difference between industry statistics and academic statistics. The general experience of data scientists is that, at their job, they have access to a *larger body of knowledge and methodology*, as well as a process, which we now define as the *data science process* (details in Chapter 2), that has foundations in both statistics and computer science.

Around all the hype, in other words, there is a ring of truth: this *is* something new. But at the same time, it's a fragile, nascent idea at real risk of being rejected prematurely. For one thing, it's being paraded around as a magic bullet, raising unrealistic expectations that will surely be disappointed.

Rachel gave herself the task of understanding the cultural phenomenon of data science and how others were experiencing it. She started meeting with people at Google, at startups and tech companies, and at universities, mostly from within statistics departments.

From those meetings she started to form a clearer picture of the new thing that's emerging. She ultimately decided to continue the investigation by giving a course at Columbia called "Introduction to Data Science," which Cathy covered on her blog. We figured that by the end of the semester, we, and hopefully the students, would know what all this actually meant. And now, with this book, we hope to do the same for many more people.

Why Now?

We have massive amounts of data about many aspects of our lives, and, simultaneously, an abundance of inexpensive computing power. Shopping, communicating, reading news, listening to music, searching for information, expressing our opinions—all this is being tracked online, as most people know.

What people might not know is that the "datafication" of our offline behavior has started as well, mirroring the online data collection revolution (more on this later). Put the two together, and there's a lot to learn about our behavior and, by extension, who we are as a species.

It's not just Internet data, though—it's finance, the medical industry, pharmaceuticals, bioinformatics, social welfare, government, education, retail, and the list goes on. There is a growing influence of data in most sectors and most industries. In some cases, the amount of data

collected might be enough to be considered "big" (more on this in the next chapter); in other cases, it's not.

But it's not only the massiveness that makes all this new data interesting (or poses challenges). It's that the data itself, often in real time, becomes the building blocks of data *products*. On the Internet, this means Amazon recommendation systems, friend recommendations on Facebook, film and music recommendations, and so on. In finance, this means credit ratings, trading algorithms, and models. In education, this is starting to mean dynamic personalized learning and assessments coming out of places like Knewton and Khan Academy. In government, this means policies based on data.

We're witnessing the beginning of a massive, culturally saturated feedback loop where our behavior changes the product and the product changes our behavior. Technology makes this possible: infrastructure for large-scale data processing, increased memory, and bandwidth, as well as a cultural acceptance of technology in the fabric of our lives. This wasn't true a decade ago.

Considering the impact of this feedback loop, we should start thinking seriously about how it's being conducted, along with the ethical and technical responsibilities for the people responsible for the process. One goal of this book is a first stab at that conversation.

Datafication

In the May/June 2013 issue of *Foreign Affairs*, Kenneth Neil Cukier and Viktor Mayer-Schoenberger wrote an article called "The Rise of Big Data". In it they discuss the concept of datafication, and their example is how we quantify friendships with "likes": it's the way everything we do, online or otherwise, ends up recorded for later examination in someone's data storage units. Or maybe multiple storage units, and maybe also for sale.

They define datafication as a process of "taking all aspects of life and turning them into data." As examples, they mention that "Google's augmented-reality glasses datafy the gaze. Twitter datafies stray thoughts. LinkedIn datafies professional networks."

Datafication is an interesting concept and led us to consider its importance with respect to people's intentions about sharing their own data. We are being datafied, or rather our actions are, and when we "like" someone or something online, we are intending to be datafied,

or at least we should expect to be. But when we merely browse the Web, we are unintentionally, or at least passively, being datafied through cookies that we might or might not be aware of. And when we walk around in a store, or even on the street, we are being datafied in a completely unintentional way, via sensors, cameras, or Google glasses.

This spectrum of intentionality ranges from us gleefully taking part in a social media experiment we are proud of, to all-out surveillance and stalking. But it's all datafication. Our intentions may run the gamut, but the results don't.

They follow up their definition in the article with a line that speaks volumes about their perspective:

> Once we datafy things, we can transform their purpose and turn the information into new forms of value.

Here's an important question that we will come back to throughout the book: who is "we" in that case? What kinds of *value* do they refer to? Mostly, given their examples, the "we" is the modelers and entrepreneurs making money from getting people to buy stuff, and the "value" translates into something like increased efficiency through automation.

If we want to think bigger, if we want our "we" to refer to people in general, we'll be swimming against the tide.

The Current Landscape (with a Little History)

So, what is data science? Is it new, or is it just statistics or analytics rebranded? Is it real, or is it pure hype? And if it's new and if it's real, what does that mean?

This is an ongoing discussion, but one way to understand what's going on in this industry is to look online and see what current discussions are taking place. This doesn't necessarily tell us what data science is, but it at least tells us what other people think it is, or how they're perceiving it. For example, on Quora there's a discussion from 2010 about "What is Data Science?" and here's Metamarket CEO Mike Driscoll's answer:

> Data science, as it's practiced, is a blend of Red-Bull-fueled hacking and espresso-inspired statistics.
>
> But data science is not merely hacking—because when hackers finish debugging their Bash one-liners and Pig scripts, few of them care about non-Euclidean distance metrics.
>
> And data science is not merely statistics, because when statisticians finish theorizing the perfect model, few could read a tab-delimited file into R if their job depended on it.
>
> Data science is the civil engineering of data. Its acolytes possess a practical knowledge of tools and materials, coupled with a theoretical understanding of what's possible.

Driscoll then refers to Drew Conway's Venn diagram of data science from 2010, shown in Figure 1-1.

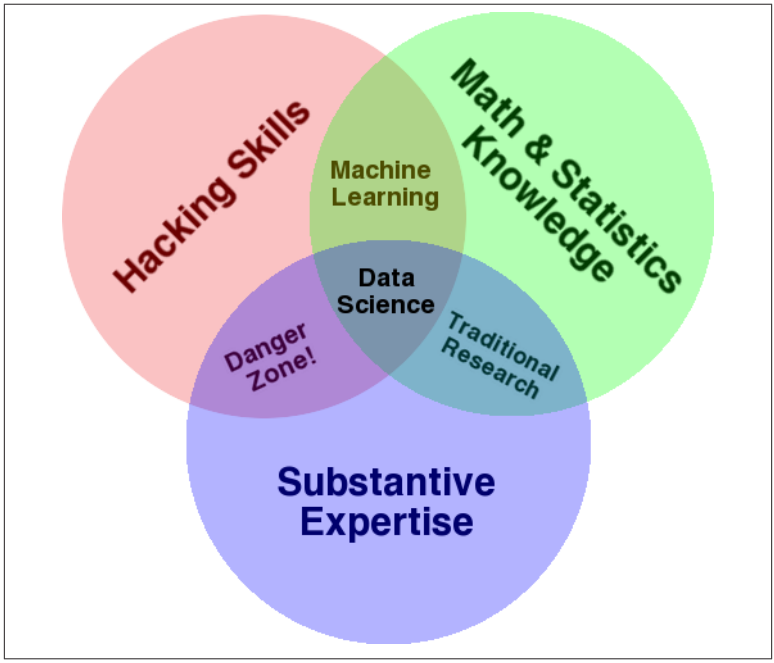

Figure 1-1. Drew Conway's Venn diagram of data science

He also mentions the sexy skills of data geeks from Nathan Yau's 2009 post, "Rise of the Data Scientist", which include:

- Statistics (traditional analysis you're used to thinking about)
- Data munging (parsing, scraping, and formatting data)

- Visualization (graphs, tools, etc.)

But wait, is data science just a bag of tricks? Or is it the logical extension of other fields like statistics and machine learning?

For one argument, see Cosma Shalizi's posts here and here, and Cathy's posts here and here, which constitute an ongoing discussion of the difference between a statistician and a data scientist. Cosma basically argues that any statistics department worth its salt does all the stuff in the descriptions of data science that he sees, and therefore data science is just a rebranding and unwelcome takeover of statistics.

For a slightly different perspective, see ASA President Nancy Geller's 2011 Amstat News article, "Don't shun the 'S' word", in which she defends statistics:

> We need to tell people that Statisticians are the ones who make sense of the data deluge occurring in science, engineering, and medicine; that statistics provides methods for data analysis in all fields, from art history to zoology; that it is exciting to be a Statistician in the 21st century because of the many challenges brought about by the data explosion in all of these fields.

Though we get her point—the phrase "art history to zoology" is supposed to represent the concept of A to Z—she's kind of shooting herself in the foot with these examples because they don't correspond to the high-tech world where much of the data explosion is coming from. Much of the development of the field is happening in industry, not academia. That is, there are people with the job title data scientist in companies, but no professors of data science in academia. (Though this may be changing.)

Not long ago, DJ Patil described how he and Jeff Hammerbacher—then at LinkedIn and Facebook, respectively—coined the term "data scientist" in 2008. So that is when "data scientist" emerged as a job title. (Wikipedia finally gained an entry on data science in 2012.)

It makes sense to us that once the skill set required to thrive at Google—working with a team on problems that required a hybrid skill set of stats and computer science paired with personal characteristics including curiosity and persistence—spread to other Silicon Valley tech companies, it required a new job title. Once it became a pattern, it deserved a name. And once it got a name, everyone and their mother wanted to be one. It got even worse when *Harvard Business Review* declared data scientist to be the "Sexiest Job of the 21st Century".

The Role of the Social Scientist in Data Science

Both LinkedIn and Facebook are social network companies. Oftentimes a description or definition of data scientist includes hybrid statistician, software engineer, and social scientist. This made sense in the context of companies where the product was a *social* product and still makes sense when we're dealing with human or user behavior. But if you think about Drew Conway's Venn diagram, data science problems cross disciplines—that's what the substantive expertise is referring to.

In other words, it depends on the context of the problems you're trying to solve. If they're social science-y problems like friend recommendations or people you know or user segmentation, then by all means, bring on the social scientist! Social scientists also do tend to be good question askers and have other good investigative qualities, so a social scientist who also has the quantitative and programming chops makes a great data scientist.

But it's almost a "historical" (historical is in quotes because 2008 isn't that long ago) artifact to limit your conception of a data scientist to someone who works only with online user behavior data. There's another emerging field out there called computational social sciences, which could be thought of as a subset of data science.

But we can go back even further. In 2001, William Cleveland wrote a position paper about data science called "Data Science: An action plan to expand the field of statistics."

So data science existed before data scientists? Is this semantics, or does it make sense?

This all begs a few questions: can you define data science by what data scientists *do*? Who gets to define the field, anyway? There's lots of buzz and hype—does the media get to define it, or should we rely on the practitioners, the self-appointed data scientists? Or is there some actual authority? Let's leave these as open questions for now, though we will return to them throughout the book.

Data Science Jobs

Columbia just decided to start an Institute for Data Sciences and Engineering with Bloomberg's help. There are 465 job openings in New

York City alone for data scientists last time we checked. That's a lot. So even if data science isn't a real field, it has *real* jobs.

And here's one thing we noticed about most of the job descriptions: they ask data scientists to be experts in computer science, statistics, communication, data visualization, *and* to have extensive domain expertise. Nobody is an expert in everything, which is why it makes more sense to create teams of people who have different profiles and different expertise—together, as a team, they can specialize in all those things. We'll talk about this more after we look at the composite set of skills in demand for today's data scientists.

A Data Science Profile

In the class, Rachel handed out index cards and asked everyone to profile themselves (on a relative rather than absolute scale) with respect to their skill levels in the following domains:

- Computer science
- Math
- Statistics
- Machine learning
- Domain expertise
- Communication and presentation skills
- Data visualization

As an example, Figure 1-2 shows Rachel's data science profile.

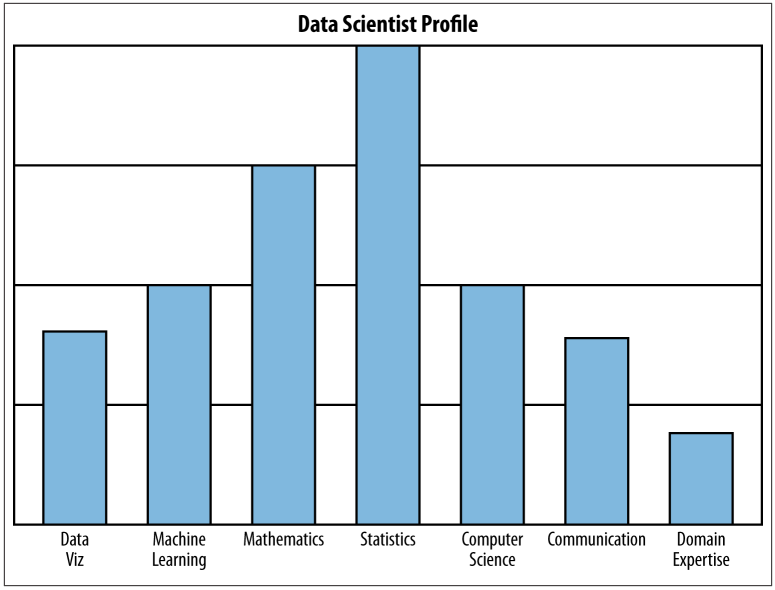

Figure 1-2. Rachel's data science profile, which she created to illustrate trying to visualize oneself as a data scientist; she wanted students and guest lecturers to "riff" on this—to add buckets or remove skills, use a different scale or visualization method, and think about the drawbacks of self-reporting

We taped the index cards to the blackboard and got to see how everyone else thought of themselves. There was quite a bit of variation, which is cool—lots of people in the class were coming from social sciences, for example.

Where is your data science profile at the moment, and where would you like it to be in a few months, or years?

As we mentioned earlier, a data science team works best when different skills (profiles) are represented across different people, because nobody is good at everything. It makes us wonder if it might be more worthwhile to define a "data science team"—as shown in Figure 1-3—than to define a data scientist.

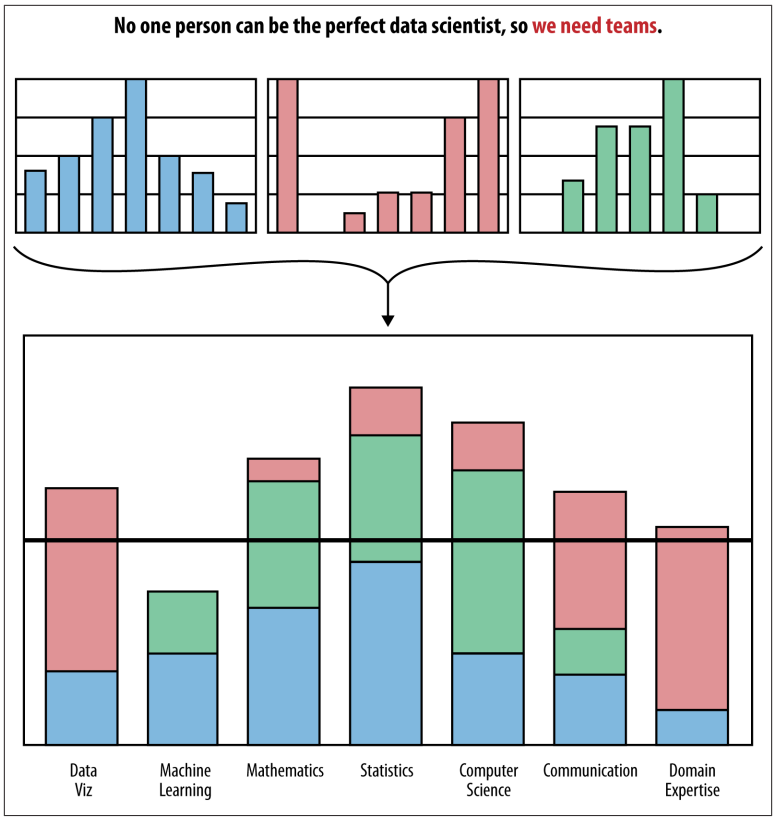

Figure 1-3. Data science team profiles can be constructed from data scientist profiles; there should be alignment between the data science team profile and the profile of the data problems they try to solve

Thought Experiment: Meta-Definition

Every class had at least one thought experiment that the students discussed in groups. Most of the thought experiments were very open-ended, and the intention was to provoke discussion about a wide variety of topics related to data science. For the first class, the initial thought experiment was: *can we use data science to define data science?*

The class broke into small groups to think about and discuss this question. Here are a few interesting things that emerged from those conversations:

Start with a text-mining model.
: We could do a Google search for "data science" and perform a text-mining model. But that would depend on us being a *usagist* rather than a *prescriptionist* with respect to language. A usagist would let the masses define data science (where "the masses" refers to whatever Google's search engine finds). Would it be better to be a prescriptionist and refer to an authority such as the *Oxford English Dictionary*? Unfortunately, the *OED* probably doesn't have an entry yet, and we don't have time to wait for it. Let's agree that there's a spectrum, that one authority doesn't feel right, and that "the masses" doesn't either.

So what about a clustering algorithm?
: How about we look at practitioners of data science and see how *they* describe what they do (maybe in a word cloud for starters)? Then we can look at how people who claim to be other things like statisticians or physicists or economists describe what they do. From there, we can try to use a clustering algorithm (which we'll use in Chapter 3) or some other model and see if, when it gets as input "the stuff someone does," it gives a good prediction on what field that person is in.

Just for comparison, check out what Harlan Harris recently did related to the field of data science: he took a survey and used clustering to define subfields of data science, which gave rise to Figure 1-4.