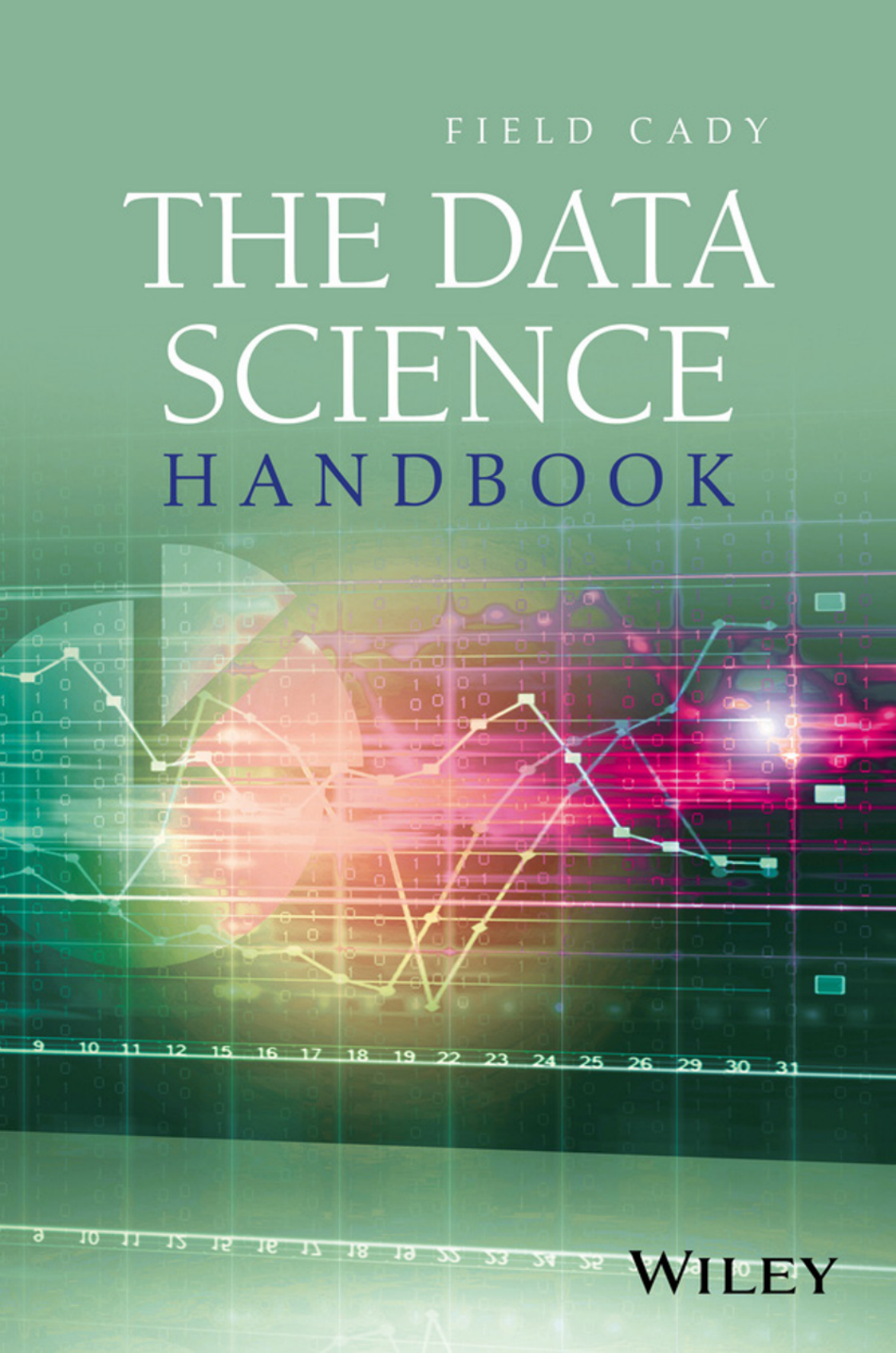
FIELD CADY

THE DATA SCIENCE HANDBOOK



WILEY

The Data Science Handbook

The Data Science Handbook

Field Cady

WILEY

Registered Offices
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

Editorial Office
111 River Street, Hoboken, NJ 07030, USA

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Library of Congress Cataloguing-in-Publication Data

Names: Cady, Field, 1984- author.
Title: The data science handbook / Field Cady.
Description: Hoboken, NJ: John Wiley & Sons, Inc., 2017. | Includes bibliographical references and index.
Identifiers: LCCN 2016043329 (print) | LCCN 2016046263 (ebook) | ISBN 9781119092940 (cloth) | ISBN 9781119092933 (pdf) | ISBN 9781119092926 (epub)
Subjects: LCSH: Databases--Handbooks, manuals, etc. | Statistics--Data processing--Handbooks, manuals, etc. | Big data--Handbooks, manuals, etc. | Information theory--Handbooks, manuals, etc.
Classification: LCC QA76.9.D32 C33 2017 (print) | LCC QA76.9.D32 (ebook) | DDC 005.74--dc23
LC record available at https://lccn.loc.gov/2016043329

Cover image: Сергей Хакимуллин/Gettyimages
Cover design by Wiley

Printed in the United States of America

Set in 10/12pt Warnock by SPi Global, Chennai, India

10 9 8 7 6 5 4 3 2 1

To my wife, Ryna. Thank you honey, for your support and for always believing in me.

Contents

Preface

This book was written to solve a problem. The people who I interview for data science jobs have sterling mathematical pedigrees, but most of them are unable to write a simple script that computes Fibonacci numbers (in case you aren't familiar with Fibonacci numbers, this takes about five lines of code). On the other side, employers tend to view data scientists as either mysterious wizards or used-car salesmen (and when data scientists can't be trusted to write a basic script, the latter impression has some merit!). These problems reflect a fundamental misunderstanding, by all parties, of what data science is (and isn't) and what skills its practitioners need.

When I first got into data science, I was part of that problem. Years of doing academic physics had trained me to solve problems in a way that was long on abstract theory but short on common sense or flexibility. Mercifully, I also knew how to code (thanks, Google™ internships!), and this let me limp along while I picked up the skills and mindsets that actually mattered.

Since leaving academia, I have done data science consulting for companies of every stripe. This includes web traffic analysis for tiny start-ups, manufacturing optimizations for Fortune 100 giants, and everything in between. The problems to solve are always unique, but the skills required to solve them are strikingly universal. They are an eclectic mix of computer programming, mathematics, and business savvy. They are rarely found together in one person, but in truth they can be learned by anybody.

A few interviews I have given stand out in my mind. The candidate was smart and knowledgeable, but the interview made it painfully clear that they were unprepared for the daily work of a data scientist. What do you do as an interviewer when the candidate starts apologizing for wasting your time? We ended up filling the hour with a crash course on what they were missing and how they could go out and fill the gaps in their knowledge. They went out, learned what they needed to, and are now successful data scientists.

I wrote this book in an attempt to help people like that out, by condensing data science's various skill sets into a single, coherent volume. It is hands-on

and to the point: ideal for somebody who needs to come up to speed quickly or solve a problem on a tight deadline. The educational system has not yet caught up to the demands of this new and exciting field, and my hope is that this book will help you bridge the gap.

Field Cady
September 2016
Redmond, Washington

1

Introduction: Becoming a Unicorn

"Data science" is a very popular term these days, and it gets applied to so many things that its meaning has become very vague. So I'd like to start this book by giving you the definition that I use. I've found that this one gets right to the heart of what sets it apart from other disciplines. Here goes:

> Data science means doing analytics work that, for one reason or another, requires a substantial amount of software engineering skills.

Sometimes, the final deliverable is the kind of thing a statistician or business analyst might provide, but achieving that goal demands software skills that your typical analyst simply doesn't have. For example, a dataset might be so large that you need to use distributed computing to analyze it or so convoluted in its format that many lines of code are required to parse it. In many cases, data scientists also have to write big chunks of production software that implement their analytics ideas in real time. In practice, there are usually other differences as well. For example, data scientists usually have to extract features from raw data, which means that they tackle very open-ended problems such as how to quantify the "spamminess" of an e-mail.

It's very hard to find people who can construct good statistical models, hack quality software, and relate this all in a meaningful way to business problems. It's a lot of hats to wear! These individuals are so rare that recruiters often call them "unicorns."

The message of this book is that it is not only possible but also relatively straightforward to become a "unicorn." It's just a question of acquiring the particular balance of skills required. Very few educational programs teach all of those skills, which is why unicorns are rare, but that's mostly a historical accident. It is perfectly reasonable for a single person to have the whole palette of abilities, provided they're willing to ignore the traditional boundaries between different disciplines.

This book aims to teach you everything you'll need to know to be a competent data scientist. My guess is that you're either a computer programmer

The Data Science Handbook, First Edition. Field Cady.

looking to learn about analytics or more of a mathematician trying to bone up on their coding. You might also be a businessperson who needs the technical skills to answer your business questions or simply an interested layman. Whoever you are though, this book will teach you the concepts you need.

This book is not comprehensive. Data science is too big an area for any person or book to cover all of it. Besides, the field is changing so fast that any "comprehensive" book would be out-of-date before it came off the presses. Instead, I have aimed for two goals. First, I want to give a solid grounding in the big picture of what data science is, how to go about doing it, and the foundational concepts that will stand the test of time. Second, I want to give a "complete" skill set, in the sense that you have the nuts-and-bolts knowledge to go out and do data science work (you can code in Python, you know the libraries to use, most of the big machine learning models, etc.), even if particular projects or companies might require that you pick up a new skill set from somewhere else.

1.1 Aren't Data Scientists Just Overpaid Statisticians?

Nate Silver, a statistician famous for accurate forecasting of US elections, once famously said: "I think data scientist is a sexed-up term for statistician." He has a point, but what he said is only partly true. The discipline of statistics deals mostly with rigorous mathematical methods for solving well-defined problems. Data scientists spend most of their time getting data into a form where statistical methods could even be applied. This involves making sure that the analytics problem is a good match to business objectives, extracting meaningful features from the raw data and coping with any pathologies of the data or weird edge cases. Once that heavy lifting is done, you can apply statistical tools to get the final results, although, in practice, you often don't even need them. Professional statisticians need to do a certain amount of preprocessing themselves, but there is a massive difference in degree.

Historically, data science emerged as a field independently from statistics. Most of the first data scientists were computer programmers or machine learning experts who were working on Big Data problems. They were analyzing datasets of the kind that statisticians don't touch: HTML pages, image files, e-mails, raw output logs of web servers, and so on. These datasets don't fit the mold of relational databases or statistical tools, so for decades, they were just piling up without being analyzed. Data science came into being as a way to finally milk them for insights.

In 20 years, I suspect that statistics, data science, and machine learning will blur into a single discipline. The differences between them are, after all, really

just a matter of degree and/or historical accident. But in practical terms, for the time being, solving data science problems requires skills that a normal statistician does not have. In fact, these skills, which include extensive software engineering and domain-specific feature extraction, constitute the overwhelming majority of the work that needs to be done. In the daily work of a data scientist, statistics plays second fiddle.

1.2 How Is This Book Organized?

This book is organized into three sections. The first, The Stuff You'll Always Use, covers topics that, in my experience, you will end up using in almost any data science project. They are core skills, which are absolutely indispensable for data science at any level.

The first section was also written with an eye toward people who need data science to answer a specific question but do not aspire to become full-fledged data scientists. If you are in this camp, then there is a good chance that Part I of the book will give you everything you need.

The second section, Stuff You Still Need to Know, covers additional core skills for a data scientist. Some of these, such as clustering, are so common that they almost made it into the first section, and they could easily play a role in any project. Others, such as natural language processing, are somewhat specialized subjects that are critical in certain domains but superfluous in others. In my judgment, a data scientist should be conversant in all of these subjects, even if they don't always use them all.

The final section, Stuff That's Good to Know, covers a variety of topics that are optional. Some of these chapters are just expansions on topics from the first two sections, but they give more theoretical background and discuss some additional topics. Others are entirely new material, which does come up in data science, but which you could go through a career without ever running into.

1.3 How to Use This Book?

This book was written with three use cases in mind:

1) You can read it cover-to-cover. If you do that, it should give you a self-contained course in data science that will leave you ready to tackle real problems. If you have a strong background in computer programming, or in mathematics, then some of it will be review.
2) You can use it to come quickly up to speed on a specific subject. I have tried to make the different chapters pretty self-contained, especially the chapters after the first section.

3) The book contains a lot of sample codes, in pieces that are large enough to use as a starting point for your own projects.

1.4 Why Is It All in Python™, Anyway?

The example code in this book is all in Python, except for a few domain-specific languages such as SQL. My goal isn't to push you to use Python; there are lots of good tools out there, and you can use whichever ones you want.

However, I wanted to use one language for all of my examples. This keeps the book readable, and it also lets readers follow the whole book while only knowing one language. Of the various languages available, there are two reasons why I chose Python:

1) Python is the most popular language for data scientists. R is its only major competitor, at least when it comes to free tools. I have used both extensively, and I think that Python is flat-out better (except for some obscure statistics packages that have been written in R and that are rarely needed anyway).
2) I like to say that for any task, Python is the second-best language. It's a jack-of-all-trades. If you only need to worry about statistics, or numerical computation, or web parsing, then there are better options out there. But if you need to do all of these things within a single project, then Python is your best option. Since data science is so inherently multidisciplinary, this makes it a perfect fit.

As a note of advice, it is much better to be proficient in one language, to the point where you can reliably churn out code that is of high quality, than to be mediocre at several.

1.5 Example Code and Datasets

This book is rich in example code, in fairly long chunks. This was done for two reasons:

1) As a data scientist, you need to be able to read longish pieces of code. This is a nonoptional skill, and if you aren't used to it, then this will give you a chance to practice.
2) I wanted to make it easier for you to poach the code from this book, if you feel so inclined.

You can do whatever you want with the code, with or without attribution. I release it into the public domain in the hope that it can give some people a small leg up. You can find it on my GitHub page at www.github.com/field-cady.

The sample data that I used comes in two forms:

1) Test datasets that are built into Python's scientific libraries
2) Data that is pulled off the Internet, from sources such as Yahoo and Wikipedia. When I do this, the example scripts will include code that pulls the data.

1.6 Parting Words

It is my hope that this book not only teaches you how to do nut-and-bolts data science but also gives you a feel of how exciting this deeply interdisciplinary subject is. Please feel free to reach out to me at www.fieldcady.com or field.cady@gmail.com with comments, errata, or any other feedback.

Part 1

The Stuff You'll Always Use

The first section of this book covers core topics that everybody doing data science should know. This includes people who are not interested in being professional data scientists, but need to know just enough to solve some specific problem. These are the subjects that will likely arise in every data science project you do.

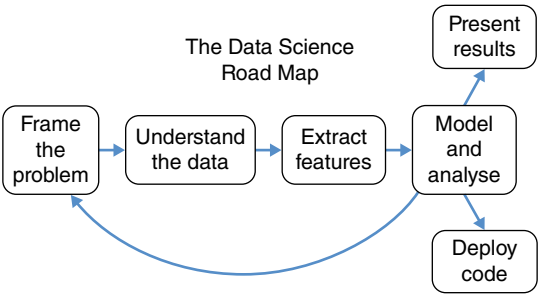
The Data Science Handbook, First Edition. Field Cady.

2

The Data Science Road Map

In this chapter, I will give you a high-level overview of the process of data science. I will focus on the different stages of data science work, including common pain points, key things to get right, and where data science parts ways from other disciplines.

The process of solving a data science problem is summarized in the following figure, which I called the Data Science Road Map.



The first step is always to frame the problem: understand the business use case and craft a well-defined analytics problem (or problems) out of it. This is followed by an extensive stage of grappling with the data and the real-world things that it describes, so that we can extract meaningful features. Finally, these features are plugged into analytical tools that give us hard numerical results.

Before I go into more detail about the different stages of the roadmap, I want to point out two things.

The first is that "Model and Analyze" loops back to framing the problem. This is one of the key features of data science that differentiate it from traditional software engineering. Data scientists write code, and they use many of the same tools as software engineers. However, there is a tight feedback loop between data science work and the real world. Questions are always being reframed as new insights become available, and, as a result, data scientists

The Data Science Handbook, First Edition. Field Cady.

must keep their code base extremely flexible and always have an eye toward the real-world problem they are solving. Ideally, you will follow the loop back many times, constantly refining your methods and producing new insights.

The second point is that there are two different (although not mutually exclusive) ways to exit the road map: presenting results and deploying code. My friend Michael Li, a data scientist who founded The Data Incubator, likened this to having two different types of clients: humans and machines. They require distinct skill sets and modifications to every stage of the data science road map.

If your clients are humans, then usually you are trying to use available data sources to answer some kind of business problem. Examples would be the following:

- Identifying leading indicators of spikes in the price of a stock, so that people can understand what causes price spikes
- Determining whether customers break down into natural subtypes and what characteristics each type has
- Assessing whether traffic to one website can be used to predict traffic to another site.

Typically, the final deliverable for work such as this will be a PowerPoint slide deck or a written report. The goal is to give business insights, and often these insights will be used for making key decisions. This kind of data science also functions as a way to test the waters and see whether some analytics approach is worth a larger follow-up project that may result in production software.

If your clients are machines, then you are doing something that blends into software engineering, where the deliverable is a piece of software that performs some analytics work. Examples would be the following:

- Implementing the algorithm that chooses which ad to show to a customer and training it on real data
- Writing a batch process that generates daily reports based on company records generated that day, using some kind of analytics to point out salient patterns

In these cases, your main deliverable is a piece of software. In addition to performing a useful task, it had better work well in terms of performance, robustness to bad inputs, and so on.

Once you understand who your clients are, the next step is to determine what you'll be doing for them. In the next section, I will show you how to do this all-important step.

2.1 Frame the Problem

The difference between great and mediocre data science is not about math or engineering: it is about asking the right question(s). Alternately, if you're trying