



Think Like a **Data Scientist**

Tackle the data science process step-by-step

Brian Godsey

Think Like a Data Scientist

TACKLE THE DATA SCIENCE PROCESS STEP-BY-STEP

BRIAN GODSEY



MANNING
SHELTER ISLAND

For online information and ordering of this and other Manning books, please visit www.manning.com. The publisher offers discounts on this book when ordered in quantity. For more information, please contact


Special Sales Department
Manning Publications Co.
20 Baldwin Road
PO Box 761
Shelter Island, NY 11964
Email: orders@manning.com

© 2017 by Manning Publications Co. All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by means electronic, mechanical, photocopying, or otherwise, without prior written permission of the publisher.

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in the book, and Manning Publications was aware of a trademark claim, the designations have been printed in initial caps or all caps.

☞ Recognizing the importance of preserving what has been written, it is Manning's policy to have the books we publish printed on acid-free paper, and we exert our best efforts to that end. Recognizing also our responsibility to conserve the resources of our planet, Manning books are printed on paper that is at least 15 percent recycled and processed without the use of elemental chlorine.

 Manning Publications Co.
20 Baldwin Road
PO Box 761
Shelter Island, NY 11964

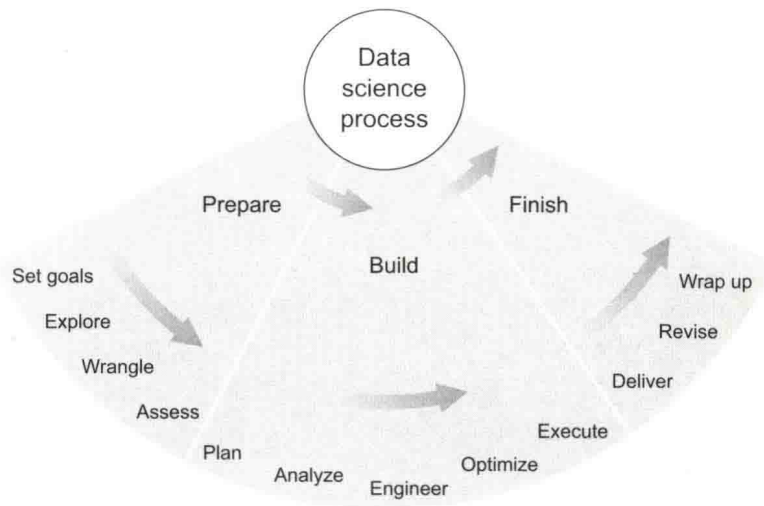
Development editor: Karen Miller
Review editor: Aleksandar Dragosavljević
Technical development editor: Mike Shepard
Project editor: Kevin Sullivan
Copy editor: Linda Recktenwald
Proofreader: Corbin Collins
Typesetter: Dennis Dalinnik
Cover designer: Marija Tudor

ISBN: 9781633430273

Printed in the United States of America

1 2 3 4 5 6 7 8 9 10 – EBM – 22 21 20 19 18 17

The lifecycle of a data science project



This book is organized around the three phases of a data science project:

- The first phase is *preparation*—time and effort spent gathering information at the beginning of a project can spare big headaches later.
- The second phase is *building* the product, from planning through execution, using what you learned during the *preparation* phase and all the tools that statistics and software can provide.
- The third and final phase is *finishing*—delivering the product, getting feedback, making revisions, supporting the product, and wrapping up the project.

Get the eBooks FREE!

(PDF, ePub, and Kindle all included)

We believe that once you buy a book from us, you should be able to read it in any format we have available. To get electronic versions of this book at no additional cost to you, purchase and then register this book at the Manning website following the instructions inside this insert.

That's it!

Thanks from Manning!



*To all thoughtful, deliberate problem-solvers
who consider themselves scientists first
and builders second*

*For everyone everywhere
who ever taught me anything*

preface

In 2012, an article in the *Harvard Business Review* named the role of data scientist “the sexiest job of the 21st century.” With 87 years left in the century, it’s fair to say they might yet change their minds. Nevertheless, at the moment, data scientists are getting a lot of attention, and as a result, books about data science are proliferating. There would be no sense in adding another book to the pile if it merely repeated or repackaged text that is easily found elsewhere. But, while surveying new data science literature, it became clear to me that most authors would rather explain how to use all the latest tools and technologies than discuss the nuanced problem-solving nature of the data science process. Armed with several books and the latest knowledge of algorithms and data stores, many aspiring data scientists were still asking the question: *Where do I start?*

And so, here is another book on data science. This one, however, attempts to lead you through the data science process as a path with many forks and potentially unknown destinations. The book warns you of what may be ahead, tells you how to prepare for it, and suggests how to react to surprises. It discusses what tools might be the most useful, and why, but the main objective is always to navigate the path—the data science process—intelligently, efficiently, and successfully, to arrive at practical solutions to real-life data-centric problems.

acknowledgments

I would like to thank everyone at Manning who helped to make this book a reality, and Marjan Bace, Manning's publisher, for giving me this opportunity.

I'd also like to thank Mike Shepard for evaluating the technical aspects of the book, and the reviewers who contributed helpful feedback during development of the manuscript. Those reviewers include Casimir Saternos, Clemens Baader, David Krief, Gavin Whyte, Ian Stirk, Jenice Tom, Łukasz Bonenberg, Martin Perry, Nicolas Boulet-Lavoie, Pouria Amirian, Ran Volkovich, Shobha Iyer, and Valmiky Arquissandas.

Finally, I extend special thanks to my teammates, current and former, at Unoceros and Panopticon Labs for providing ample fodder for this book in many forms: experiences and knowledge in software development and data science, fruitful conversations, crazy ideas, funny stories, awkward mistakes, and most importantly, willingness to indulge my curiosity.

about this book

Data science still carries the aura of a new field. Most of its components—statistics, software development, evidence-based problem solving, and so on—descend directly from well-established, even old, fields, but data science seems to be a fresh assemblage of these pieces into something that is new, or at least *feels* new in the context of current public discourse.

Like many new fields, data science hasn't quite found its footing. The lines between it and other related fields—as far as those lines matter—are still blurry. Data science may rely on, but is not equivalent to, database architecture and administration, big data engineering, machine learning, or high-performance computing, to name a few.

The core of data science doesn't concern itself with specific database implementations or programming languages, even if these are indispensable to practitioners. The core is the interplay between data content, the goals of a given project, and the data-analytic methods used to achieve those goals. The data scientist, of course, must manage these using any software necessary, but which software and how to implement it are details that I like to imagine have been abstracted away, as if in some distant future reality.

This book attempts to foresee that future in which the most common, rote, mechanical tasks of data science are stripped away, and we are left with only the core: applying the scientific method to data sets in order to achieve a project's goals. This, the process of data science, involves software as a necessary set of tools, just as a traditional scientist might use test tubes, flasks, and a Bunsen burner. But, what

matters is what's happening on the inside: what's happening to the data, what results we get, and why.

In the following pages, I introduce a wide range of software tools, but I keep my descriptions brief. More-comprehensive introductions can always be found elsewhere, and I'm more eager to delve into what those tools can do for you, and how they can aid you in your research and development. Focus always returns to the key concepts and challenges that are unique to each project in data science, and the process of organizing and harnessing available resources and information to achieve the project's goals.

To get the most out of this book, you should be reasonably comfortable with elementary statistics—a college class or two is fine—and have some basic knowledge of a programming language. If you're an expert in statistics, software development, or data science, you might find some parts of this book slow or trivial. That's OK; skip or skim sections if you must. I don't hope to replace anyone's knowledge and experience, but I do hope to supplement them by providing a conceptual framework for working through data science projects, and by sharing some of my own experiences in a constructive way.

If you're a beginner in data science, welcome to the field! I've tried to describe concepts and topics throughout the book so that they'll make sense to just about anyone with some technical aptitude. Likewise, colleagues and managers of data scientists and developers might also read this book to get a better idea of how the data science process works from an inside perspective.

For every reader, I hope this book paints a vivid picture of data science as a process with many nuances, caveats, and uncertainties. The power of data science lies not in figuring out what *should* happen next, but in realizing what *might* happen next and eventually finding out what *does* happen next. My sincere hope is that you enjoy the book and, more importantly, that you learn some things that increase your chances of success in the future.

Roadmap

The book is divided into three parts, representing the three major phases of the data science process. Part 1 covers the preparation phase:

- Chapter 1 discusses my process-oriented perspective of data science projects and introduces some themes and concepts that are present throughout the book.
- Chapter 2 covers the deliberate and important step of setting good goals for the project. Special focus is given to working with the project's customer to generate practical questions to address, and also to being pragmatic about the data's ability to address those questions.
- Chapter 3 delves into the exploration phase of a data science project, in which we try to discover helpful sources of data. I cover some helpful methods of data

discovery and data access, as well as some important things to consider when choosing which data sources to use in the project.

- Chapter 4 gives an overview of data wrangling, a process by which “raw,” unkempt, or unstructured data is brought to heel, so that you can make good use of it.
- Chapter 5 discusses data assessment. After you’ve discovered and selected some data sources, this chapter explains how to perform preliminary examinations of the data you have, so that you’re more informed while making a subsequent project plan, with realistic expectations of what the data can do.

Part 2 covers the building phase:

- Chapter 6 shows how to develop a plan for achieving a project’s goals based on what you’ve learned from exploration and assessment. Special focus is given to planning for uncertainty in future outcomes and results.
- Chapter 7 takes a detour into the field of statistics, introducing a wide variety of important concepts, tools, and methods, focusing on their principal capabilities and how they can help achieve project goals.
- Chapter 8 does the same for statistical software; the chapter is intended to arm you with enough knowledge to make informed choices when choosing software for your project.
- Chapter 9 gives a high-level overview of some popular software tools that are not specifically statistical, but that might make building and using your product easier or more efficient.
- Chapter 10 brings chapters 7, 8, and 9 together by discussing the execution of your project plan, given the knowledge gained from the previous detours into statistics and software, while considering some hard-to-identify nuances as well as the many pitfalls of dealing with data, statistics, and software.

Part 3 covers the finishing phase:

- Chapter 11 looks at the advantages of refining and curating the form and content of the product to concisely convey to the customer the results that most effectively solve problems and achieve project goals.
- Chapter 12 discusses some of the things that can happen shortly after product delivery, including bug discovery, inefficient use of the product by the customer, and the need to refine or modify the product.
- Chapter 13 concludes with some advice on storing the project cleanly and carrying forward lessons learned in order to improve your chances of success in future projects.

Exercises are included near the end of every chapter except chapter 1. Answers and example responses to these exercises appear in the last section of the book, before the index.

Author Online

Purchase of *Think Like a Data Scientist* includes free access to a private web forum run by Manning Publications where you can make comments about the book, ask technical questions, and receive help from the author and from other users. To access the forum and subscribe to it, point your web browser to www.manning.com/books/think-like-a-data-scientist. This page provides information on how to get on the forum once you're registered, what kind of help is available, and the rules of conduct on the forum.

Manning's commitment to our readers is to provide a venue where a meaningful dialog between individual readers and between readers and the author can take place. It is not a commitment to any specific amount of participation on the part of the author, whose contributions to the AO forum remain voluntary (and unpaid). We suggest you ask the author challenging questions, lest his interest stray!

About the author

Brian Godsey, PhD, worked for nearly a decade in academic and government roles, applying mathematics and statistics to fields such as bioinformatics, finance, and national defense, before changing focus to data-centric startups. He led the data science team at a local Baltimore startup—seeing it grow from seed to series A funding rounds and seeing the product evolve from prototype to production versions—before helping launch two startups, Unoceros and Panopticon Labs, and their data-centric products.

about the cover illustration

The figure on the cover of *Think Like a Data Scientist* is captioned “A soldier of the Strelitz guards under arms,” or *Soldat du corps des Strelits sous les armés*. The Strelitz guards were part of the Muscovite army in Czarist Russia through the eighteenth century. The illustration is taken from Thomas Jefferys’ *A Collection of the Dresses of Different Nations, Ancient and Modern*, published in London between 1757 and 1772. The title page states that these are hand-colored copperplate engravings, heightened with gum arabic. Thomas Jefferys (1719–1771) was called “Geographer to King George III.” He was an English cartographer who was the leading map supplier of his day. He engraved and printed maps for government and other official bodies and produced a wide range of commercial maps and atlases, especially of North America. His work as a mapmaker sparked an interest in local dress customs of the lands he surveyed and mapped; they are brilliantly displayed in this four-volume collection.

Fascination with faraway lands and travel for pleasure were relatively new phenomena in the eighteenth century, and collections such as this one were popular, introducing both the tourist and the armchair traveler to the inhabitants of other countries. The diversity of the drawings in Jefferys’ volumes speaks vividly of the uniqueness and individuality of the world’s nations centuries ago. Dress codes have changed, and the diversity by region and country, so rich at one time, has faded away. It is now often hard to tell the inhabitant of one continent from another. Perhaps, trying to view it optimistically, we have traded a cultural and visual diversity for a more varied personal life—or a more varied and interesting intellectual and technical life.

At a time when it is hard to tell one computer book from another, Manning celebrates the inventiveness and initiative of the computer business with book covers based on the rich diversity of national costumes from centuries ago, brought back to life by Jefferys' pictures.

brief contents

PART 1 PREPARING AND GATHERING DATA AND KNOWLEDGE1

- 1 ■ Philosophies of data science 3
- 2 ■ Setting goals by asking good questions 19
- 3 ■ Data all around us: the virtual wilderness 37
- 4 ■ Data wrangling: from capture to domestication 67
- 5 ■ Data assessment: poking and prodding 84

PART 2 BUILDING A PRODUCT WITH SOFTWARE AND STATISTICS105

- 6 ■ Developing a plan 107
- 7 ■ Statistics and modeling: concepts and foundations 129
- 8 ■ Software: statistics in action 166
- 9 ■ Supplementary software: bigger, faster, more efficient 201
- 10 ■ Plan execution: putting it all together 215

PART 3 FINISHING OFF THE PRODUCT AND WRAPPING UP237

- 11 ■ Delivering a product 239
- 12 ■ After product delivery: problems and revisions 256
- 13 ■ Wrapping up: putting the project away 274

