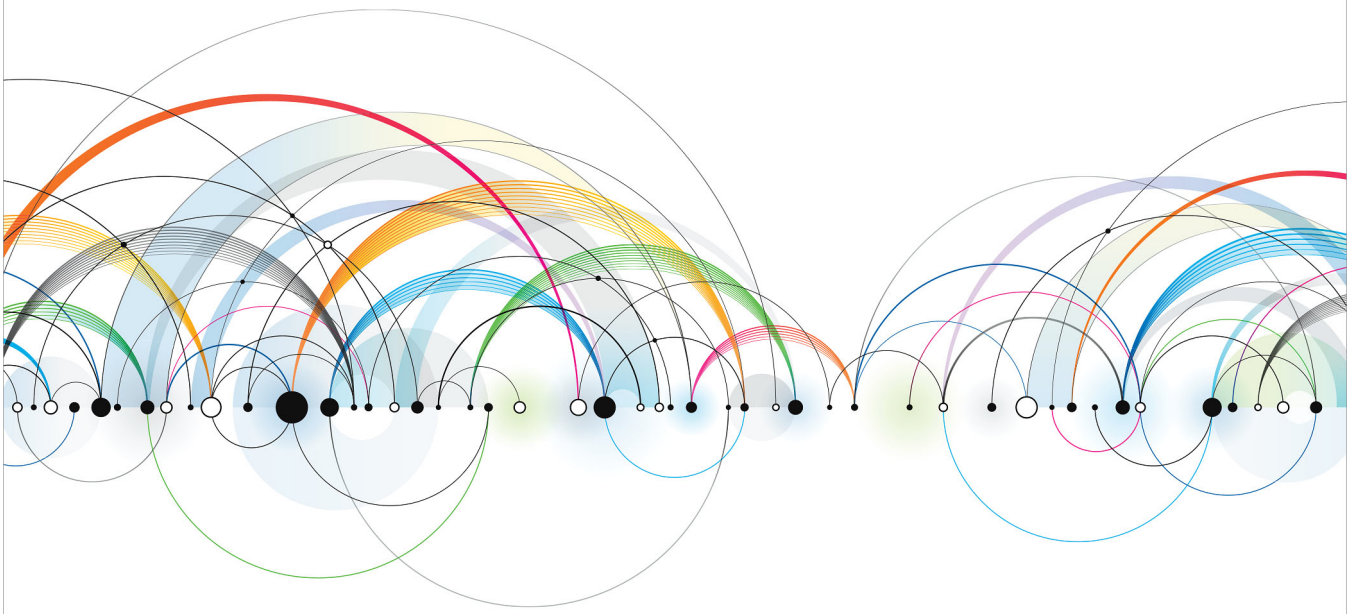
"A must-read resource for anyone who is serious about embracing the opportunity of big data."

—Craig Vaughan, Global Vice President, SAP

Data Science *for* Business

What You Need to Know About Data Mining and Data-Analytic Thinking



Foster Provost & Tom Fawcett

Praise

"A must-read resource for anyone who is serious about embracing the opportunity of big data."

—*Craig Vaughan*
Global Vice President at SAP

"This timely book says out loud what has finally become apparent: in the modern world, Data is Business, and you can no longer think business without *thinking data*. Read this book and you will understand the Science behind thinking data."

—*Ron Bekkerman*
Chief Data Officer at Carmel Ventures

"A great book for business managers who lead or interact with data scientists, who wish to better understand the principles and algorithms available without the technical details of single-disciplinary books."

—*Ronny Kohavi*
Partner Architect at Microsoft Online Services Division

"Provost and Fawcett have distilled their mastery of both the art and science of real-world data analysis into an unrivalled introduction to the field."

—*Geoff Webb*
Editor-in-Chief of *Data Mining and Knowledge Discovery* Journal

"I would love it if everyone I had to work with had read this book."

—*Claudia Perlich*
Chief Scientist of Dstillery and Advertising Research
Foundation Innovation Award Grand Winner (2013)

"A foundational piece in the fast developing world of Data Science. A must read for anyone interested in the Big Data revolution."

—Justin Gapper
Business Unit Analytics Manager
at Teledyne Scientific and Imaging

"The authors, both renowned experts in data science before it had a name, have taken a complex topic and made it accessible to all levels, but mostly helpful to the budding data scientist. As far as I know, this is the first book of its kind—with a focus on data science concepts as applied to practical business problems. It is liberally sprinkled with compelling real-world examples outlining familiar, accessible problems in the business world: customer churn, targeted marking, even whiskey analytics!

The book is unique in that it does not give a cookbook of algorithms, rather it helps the reader understand the underlying concepts behind data science, and most importantly how to approach and be successful at problem solving. Whether you are looking for a good comprehensive overview of data science or are a budding data scientist in need of the basics, this is a must-read."

—Chris Volinsky
Director of Statistics Research at AT&T Labs and Winning
Team Member for the $1 Million Netflix Challenge

"This book goes beyond data analytics 101. It's the essential guide for those of us (all of us?) whose businesses are built on the ubiquity of data opportunities and the new mandate for data-driven decision-making."

—Tom Phillips
CEO of Dstillery and Former Head of
Google Search and Analytics

"Intelligent use of data has become a force powering business to new levels of competitiveness. To thrive in this data-driven ecosystem, engineers, analysts, and managers alike must understand the options, design choices, and tradeoffs before them. With motivating examples, clear exposition, and a breadth of details covering not only the "hows" but the "whys", *Data Science for Business* is the perfect primer for those wishing to become involved in the development and application of data-driven systems."

—Josh Attenberg
Data Science Lead at Etsy

"Data is the foundation of new waves of productivity growth, innovation, and richer customer insight. Only recently viewed broadly as a source of competitive advantage, dealing well with data is rapidly becoming table stakes to stay in the game. The authors' deep applied experience makes this a must read—a window into your competitor's strategy."

—Alan Murray
Serial Entrepreneur; Partner at Coriolis Ventures; Co-Founder Neuehouse

"One of the best data mining books, which helped me think through various ideas on liquidity analysis in the FX business. The examples are excellent and help you take a deep dive into the subject! This one is going to be on my shelf for lifetime!"

—Nidhi Kathuria
Vice President of FX at Royal Bank of Scotland

"An excellent and accessible primer to help businessfolk better appreciate the concepts, tools and techniques employed by data scientists... and for data scientists to better appreciate the business context in which their solutions are deployed."

—Joe McCarthy
Director of Analytics and Data Science at Atigeo, LLC

"In my opinion it is the best book on Data Science and Big Data for a professional understanding by business analysts and managers who must apply these techniques in the practical world."

—Ira Laefsky
MS Engineering (Computer Science)/MBA Information Technology and Human Computer Interaction Researcher formerly on the Senior Consulting Staff of Arthur D. Little, Inc. and Digital Equipment Corporation

"With motivating examples, clear exposition and a breadth of details covering not only the "hows" but the "whys," Data Science for Business is the perfect primer for those wishing to become involved in the development and application of data driven systems."

—Ted O'Brien
Co-Founder / Director of Talent Acquisition at Starbridge
Partners and Publisher of the *Data Science Report*

Data Science for Business

**Foster Provost and Tom Fawcett

Special Edition for Data Science for Business Analytics, Stern School, NYU

Beijing • Boston • Farnham • Sebastopol • Tokyo O'REILLY®

Data Science for Business
by Foster Provost and Tom Fawcett

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 Meghan Blanchette
Production Editor: Christopher Hearse
Proofreader: Kiel Van Horn
Indexer: WordCo Indexing Services, Inc.
Interior Designer: David Futato
Cover Designer: Mark Paglietti
Illustrator: Rebecca Demarest

July 2013: First Edition

Revision History for the First Edition
2013-07-25: First Release
2013-12-19: Second Release
yyyy-mm-dd: Third Release

See *http://oreilly.com/catalog/errata.csp?isbn=9781449361327* for release details.

978-1-449-36132-7

[LSI]

For our fathers.

Table of Contents

Preface

Data Science for Business is intended for several sorts of readers:

- Business people who will be working with data scientists, managing data science–oriented projects, or investing in data science ventures,
- Developers who will be implementing data science solutions, and
- Aspiring data scientists.

This is not a book about algorithms, nor is it a replacement for a book about algorithms. We deliberately avoided an algorithm-centered approach. We believe there is a relatively small set of fundamental concepts or principles that underlie techniques for extracting useful knowledge from data. These concepts serve as the *foundation* for many well-known algorithms of data mining. Moreover, these concepts underlie the analysis of data-centered business problems, the creation and evaluation of data science solutions, and the evaluation of general data science strategies and proposals. Accordingly, we organized the exposition around these general principles rather than around specific algorithms. Where necessary to describe procedural details, we use a combination of text and diagrams, which we think are more accessible than a listing of detailed algorithmic steps.

The book does not presume a sophisticated mathematical background. However, by its very nature the material is somewhat technical—the goal is to impart a significant understanding of data science, not just to give a high-level overview. In general, we have tried to minimize the mathematics and make the exposition as "conceptual" as possible.

Colleagues in industry comment that the book is invaluable for helping to align the understanding of the business, technical/development, and data science teams. That observation is based on a small sample, so we are curious to see how general it truly is (see Chapter 5!). Ideally, we envision a book that any data scientist would give to his collaborators from the development or business teams, effectively saying: if you really

want to design/implement top-notch data science solutions to business problems, we all need to have a common understanding of this material.

Colleagues also tell us that the book has been quite useful in an unforeseen way: for preparing to interview data science job candidates. The demand from business for hiring data scientists is strong and increasing. In response, more and more job seekers are presenting themselves as data scientists. Every data science job candidate should understand the fundamentals presented in this book. (Our industry colleagues tell us that they are surprised how many do not. We have half-seriously discussed a follow-up pamphlet "Cliff's Notes to Interviewing for Data Science Jobs.")

Our Conceptual Approach to Data Science

In this book we introduce a collection of the most important fundamental concepts of data science. Some of these concepts are "headliners" for chapters, and others are introduced more naturally through the discussions (and thus they are not necessarily labeled as fundamental concepts). The concepts span the process from envisioning the problem, to applying data science techniques, to deploying the results to improve decision-making. The concepts also undergird a large array of business analytics methods and techniques.

The concepts fit into three general types:

1. Concepts about how data science fits in the organization and the competitive landscape, including ways to attract, structure, and nurture data science teams; ways for thinking about how data science leads to competitive advantage; and tactical concepts for doing well with data science projects.
2. General ways of thinking data-analytically. These help in identifying appropriate data and consider appropriate methods. The concepts include the *data mining process* as well as the collection of different *high-level data mining tasks*.
3. General concepts for actually extracting knowledge from data, which undergird the vast array of data science tasks and their algorithms.

For example, one fundamental concept is that of determining the similarity of two entities described by data. This ability forms the basis for various specific tasks. It may be used directly to *find* customers similar to a given customer. It forms the core of several *prediction* algorithms that estimate a target value such as the expected resource usage of a client or the probability of a customer to respond to an offer. It is also the basis for *clustering* techniques, which group entities by their shared features without a focused objective. Similarity forms the basis of *information retrieval*, in which documents or webpages relevant to a search query are retrieved. Finally, it underlies several common algorithms for *recommendation*. A traditional algorithm-oriented book might present each of these tasks in a different chapter, under different

names, with common aspects buried in algorithm details or mathematical propositions. In this book we instead focus on the unifying concepts, presenting specific tasks and algorithms as natural manifestations of them.

As another example, in evaluating the utility of a pattern, we see a notion of *lift*—how much more prevalent a pattern is than would be expected by chance—recurring broadly across data science. It is used to evaluate very different sorts of patterns in different contexts. Algorithms for targeting advertisements are evaluated by computing the lift one gets for the targeted population. Lift is used to judge the weight of evidence for or against a conclusion. Lift helps determine whether a co-occurrence (an association) in data is interesting, as opposed to simply being a natural consequence of popularity.

We believe that explaining data science around such fundamental concepts not only aids the reader, it also facilitates communication between business stakeholders and data scientists. It provides a shared vocabulary and enables both parties to understand each other better. The shared concepts lead to deeper discussions that may uncover critical issues otherwise missed.

To the Instructor

This book has been used successfully as a textbook for a very wide variety of data science and business analytics courses. Historically, the book arose from the development of Foster's multidisciplinary Data Science and Business Analytics classes at the Stern School at NYU, starting in the fall of 2005.[1] The original class was nominally for MBA students and MSIS students, but drew students from schools across the university. The most interesting aspect of the class was not that it appealed to MBA and MSIS students, for whom it was designed. More interesting, it also was found to be very valuable by students with strong backgrounds in machine learning and other technical disciplines. Part of the reason seemed to be that the focus on fundamental principles and other issues besides algorithms was missing from their curricula.

At NYU we now use the book in support of a variety of data science–related programs: the original MBA and MSIS programs, undergraduate business analytics, NYU/Stern's MS in Business Analytics program, executive education, and as the Introduction to Data Science for NYU's MS in Data Science. In addition, the book has been adopted by well over 100 other universities for programs in at least 22 countries (and counting), in business schools, in data science programs, in computer science programs, and for more general introductions to data science.

1 Of course, each author has the distinct impression that he did the majority of the work on the book.

The books' website gives pointers on how to obtain helpful instructional material, including lecture slides, sample homework questions and problems, example project instructions based on the frameworks from the book, exam questions, and more.

We keep an up-to-date list of known adopters on the book's website. Click *Who's Using It* at the top.

Other Skills and Concepts

There are many other concepts and skills that a practical data scientist needs to know besides the fundamental principles of data science. These skills and concepts will be discussed in Chapter 1 and Chapter 2. The interested reader is encouraged to visit the book's website for pointers to material for learning these additional skills and concepts (for example, scripting in Python, Unix command-line processing, datafiles, common data formats, databases and querying, big data architectures and systems like MapReduce and Hadoop, data visualization, and other related topics).

Sections and Notation

In addition to occasional footnotes, the book contains boxed "sidebars." These are essentially extended footnotes. We reserve these for material that we consider interesting and worthwhile, but too long for a footnote and too much of a digression for the main text.

Technical Details Ahead — A note on the starred sections

The occasional mathematical details are relegated to optional "starred" sections. These section titles will have asterisk prefixes, and they will be preceded by a paragraph rendered like this one. Such "starred" sections contain more detailed mathematics and/or more technical details than elsewhere, and these introductory paragraph explains its purpose. The book is written so that these sections may be skipped without loss of continuity, although in a few places we remind readers that details appear there.

Constructions in the text like (Smith and Jones, 2003) indicate a reference to an entry in the bibliography (in this case, the 2003 article or book by Smith and Jones); "Smith and Jones (2003)" is a similar reference. A single bibliography for the entire book appears in the endmatter.

In this book we try to keep math to a minimum, and what math there is we have simplified as much as possible without introducing confusion. For our readers with technical backgrounds, a few comments may be in order regarding our simplifying choices.

1. We avoid Sigma (Σ) and Pi (Π) notation, commonly used in textbooks to indicate sums and products, respectively. Instead we simply use equations with ellipses like this:

   $$f(x) = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n$$

   In the technical, "starred" sections we sometimes adopt Sigma and Pi notation when this ellipsis approach is just too cumbersome. We assume people reading these sections are somewhat more comfortable with math notation and will not be confused.
2. Statistics books are usually careful to distinguish between a value and its estimate by putting a "hat" on variables that are estimates, so in such books you'll typically see a true probability denoted p and its estimate denoted $\hat{p}$. In this book we are almost always talking about estimates from data, and putting hats on everything makes equations verbose and ugly. Everything should be assumed to be an estimate from data unless we say otherwise.
3. We simplify notation and remove extraneous variables where we believe they are clear from context. For example, when we discuss classifiers mathematically, we are technically dealing with decision predicates over feature vectors. Expressing this formally would lead to equations like:

   $$\hat{f}_R(\mathbf{x}) = x_{\text{Age}} + 0.7 \times x_{\text{Balance}} + 60$$

   Instead we opt for the more readable:

   $$f(\mathbf{x}) = Age + 0.7 \times \text{Balance} + 60$$

   with the understanding that $\mathbf{x}$ is a vector and *Age* and *Balance* are components of it.

We have tried to be consistent with typography, reserving fixed-width typewriter fonts like `sepal_width` to indicate attributes or keywords in data. For example, in the text-mining chapter, a word like *'discussing'* designates a word in a document while `discuss` might be the resulting token in the data.

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

Throughout the book we have placed special inline tips and warnings relevant to the material. They will be rendered differently depending on whether you're reading paper, PDF, or an ebook, as follows:

A sentence or paragraph typeset like this signifies a tip or a suggestion.

This text and element signifies a general note.

Text rendered like this signifies a warning or caution. These are more important than tips and are used sparingly.

Using Examples

In addition to being an introduction to data science, this book is intended to be useful in discussions of and day-to-day work in the field. Answering a question by citing this book and quoting examples does not require permission. We appreciate, but do not require, attribution. Formal attribution usually includes the title, author, publisher, and ISBN. For example: "*Data Science for Business* by Foster Provost and Tom Fawcett (O'Reilly). Copyright 2013 Foster Provost and Tom Fawcett, 978-1-449-36132-7."

If you feel your use of 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 two web pages for this book, where we list errata, examples, and any additional information. You can access the publisher's page at *http://oreil.ly/data-science* and the authors' page at *http://www.data-science-for-biz.com*.

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

For more information about O'Reilly Media's books, courses, conferences, and news, see their 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

Thanks to all the many colleagues and others who have provided invaluable ideas, feedback, criticism, suggestions, and encouragement based on discussions and many prior draft manuscripts. At the risk of missing someone, let us thank in particular: Panos Adamopoulos, Manuel Arriaga, Josh Attenberg, Solon Barocas, Ron Bekkerman, Enrico Bertini, Josh Blumenstock, Ohad Brazilay, Aaron Brick, Jessica Clark, Nitesh Chawla, Brian d'Alessandro, Peter Devito, Vasant Dhar, Jan Ehmke, Theos Evgeniou, Justin Gapper, Tomer Geva, Daniel Gillick, Shawndra Hill, Nidhi Kathuria, Ronny Kohavi, Marios Kokkodis, Tom Lee, Philipp Marek, David Martens, Sophie Mohin, Lauren Moores, Alan Murray, Nick Nishimura, Balaji Padmanabhan, Jason Pan, Claudia Perlich, Gregory Piatetsky-Shapiro, Tom Phillips, Kevin Reilly, Maytal Saar-Tsechansky, Evan Sadler, Galit Shmueli, Roger Stein, Nick Street, Kiril Tsemekhman, Akhmed Umyarov, Craig Vaughan, Chris Volinsky, Wally Wang, Geoff Webb, Debbie Yuster, and Rong Zheng. We would also like to thank more generally the students from Foster's classes, Data Mining for Business Analytics, Practical Data Science, Data Analytics, Introduction to Data Science, and the Data Science Research Seminar. Questions and issues that arose when using prior drafts of this book provided substantive feedback for improving it.

Thanks to all the colleagues who have taught us about data science and about how to teach data science over the years. Thanks especially to Maytal Saar-Tsechansky, Claudia Perlich, Shawndra Hill, and Vasant Dhar. Maytal graciously shared with Foster her notes for her data mining class many years ago. The classification tree example in Chapter 3 (thanks especially for the "bodies" visualization) is based mostly on her idea and example; her ideas and example were the genesis for the visualization comparing the partitioning of the instance space with trees and linear discriminant functions in Chapter 4, the "Will David Respond" example in Chapter 6 is based on her example, and probably other things long forgotten. Claudia has taught companion sections of Data Mining for Business Analytics/Introduction to Data Science along with Foster for the past few years, and has taught him much about data science in the process (and beyond). Shawndra helped Foster with putting together his new kind of data mining class over a decade ago. And way back in the 1990s Vasant taught the first data mining course for a business audience, and invited Foster (then an industry data scientist) to guest lecture about real-world data mining applications.

Thanks to David Stillwell, Thore Graepel, and Michal Kosinski for providing the Facebook Like data for some of the examples. Thanks to Nick Street for providing the cell nuclei data and for letting us use the cell nuclei image in Chapter 4. Thanks to David Martens for his help with the mobile locations visualization. Thanks to Chris Volinsky for providing data from his work on the Netflix Challenge. Thanks to Sonny Tambe for early access to his results on big data technologies and productivity. Thanks to Patrick Perry for pointing us to the bank call center example used in Chap-

ter 12. Thanks to Geoff Webb for the use of the Magnum Opus association mining system.

Thanks especially to our editor Mike Loukides, who shared our vision for a different sort of book, and the entire O'Reilly team for helping us to make it a reality.

Most of all we thank our families for their love, patience and encouragement.

A great deal of open source software was used in the preparation of this book and its examples. The authors wish to thank the developers and contributors of:

- Python and Perl
- Scipy, Numpy, Matplotlib, and Scikit-Learn
- Weka
- The Machine Learning Repository at the University of California at Irvine (Bache & Lichman, 2013)

Finally, we encourage readers to check our website for updates to this material, new chapters, errata, addenda, and accompanying slide sets.

—Foster Provost and Tom Fawcett