O'REILLY®

Grayscale Edition

For Sale in the Indian Subcontinent & Select Countries Only*

*Refer Back Cover

Second Edition

Data Science from Scratch

First Principles with Python

SPD®

Joel Grus

SECOND EDITION

Data Science from Scratch

First Principles with Python

Joel Grus

Beijing • Boston • Farnham • Sebastopol • Tokyo

O'REILLY®

SPD®

SHROFF PUBLISHERS & DISTRIBUTORS PVT. LTD.

Mumbai *Bangalore* *Kolkata* *New Delhi*

Data Science from Scratch

by Joel Grus

Copyright © 2019 Joel Grus. All rights reserved. ISBN: 978-1-492-04113-9
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://safari.oreilly.com)*. For more information, contact our corporate/ institutional sales department: (800) 998-9938 or *corporate@oreilly.com*.

Editor: Michele Cronin
Production Editor: Deborah Baker
Copyeditor: Rachel Monaghan
Proofreader: Rachel Head

Indexer: Judy McConville
Interior Designer: David Futato
Cover Designer: Karen Montgomery
Illustrator: Rebecca Demarest

Printing History:
April 2015: First Edition
May 2019: Second Edition

Revision History for the Second Edition 2019-04-10: First Release

See http://oreilly.com/catalog/errata.csp?isbn=9781492041139 for release details.

First Indian Reprint: May 2019

ISBN: 978-93-5213-832-6

The O'Reilly logo is a registered trademark of O'Reilly Media, Inc. *Data Science from Scratch*, Second Edition, the cover image of a rock ptarmigan, and related trade dress are trademarks of O'Reilly Media, Inc.

While the publisher and the author have used good faith efforts to ensure that the information and instructions contained in this work are accurate, the publisher and the author disclaim all responsibility for errors or omissions, including without limitation responsibility for damages resulting from the use of or reliance on this work. Use of the information and instructions contained in this work is at your own risk. If any code samples or other technology this work contains or describes is subject to open source licenses or the intellectual property rights of others, it is your responsibility to ensure that your use thereof complies with such licenses and/or rights.

For sale in the Indian Subcontinent (India, Pakistan, Bangladesh, Sri Lanka, Nepal, Bhutan, Maldives) and African Continent (excluding Morocco, Algeria, Tunisia, Libya, Egypt, and the Republic of South Africa) only. Illegal for sale outside of these countries.

Authorized reprint of the original work published by O'Reilly Media, Inc. All rights reserved. No part of the material protected by this copyright notice may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, nor exported to any countries other than ones mentioned above without the written permission of the copyright owner.

Published by **Shroff Publishers & Distributors Pvt. Ltd.** B-103, Railway Commercial Complex, Sector 3, Sanpada (E), Navi Mumbai 400705 • TEL: (91 22) 4158 4158 • FAX: (91 22) 4158 4141 E-mail:spdorders@ shroffpublishers.com•Web:www.shroffpublishers.com Printed at Jasmine Art Printers Pvt. Ltd., Navi Mumbai.

Table of Contents

Preface to the Second Edition

I am exceptionally proud of the first edition of *Data Science from Scratch*. It turned out very much the book I wanted it to be. But several years of developments in data science, of progress in the Python ecosystem, and of personal growth as a developer and educator have *changed* what I think a first book in data science should look like.

In life, there are no do-overs. In writing, however, there are second editions.

Accordingly, I've rewritten all the code and examples using Python 3.6 (and many of its newly introduced features, like type annotations). I've woven into the book an emphasis on writing clean code. I've replaced some of the first edition's toy examples with more realistic ones using "real" datasets. I've added new material on topics such as deep learning, statistics, and natural language processing, corresponding to things that today's data scientists are likely to be working with. (I've also removed some material that seems less relevant.) And I've gone over the book with a fine-toothed comb, fixing bugs, rewriting explanations that are less clear than they could be, and freshening up some of the jokes.

The first edition was a great book, and this edition is even better. Enjoy!

Joel Grus
Seattle, WA
2019

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 element signifies a tip or suggestion.

This element signifies a general note.

This element indicates a warning or caution.

Using Code Examples

Supplemental material (code examples, exercises, etc.) is available for download at *https://github.com/joelgrus/data-science-from-scratch*.

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: “*Data Science from Scratch*, Second Edition, by Joel Grus (O’Reilly). Copyright 2019 Joel Grus, 978-1-492-04113-9.”

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*.

O’Reilly Online Learning

For almost 40 years, *O’Reilly Media* has provided technology and business training, knowledge, and insight to help companies succeed.

Our unique network of experts and innovators share their knowledge and expertise through books, articles, conferences, and our online learning platform. O’Reilly’s online learning platform gives you on-demand access to live training courses, in-depth learning paths, interactive coding environments, and a vast collection of text and video from O’Reilly and 200+ other publishers. For more information, please visit *http://oreilly.com*.

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://bit.ly/data-science-from-scratch-2e*.

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

First, I would like to thank Mike Loukides for accepting my proposal for this book (and for insisting that I pare it down to a reasonable size). It would have been very easy for him to say, "Who's this person who keeps emailing me sample chapters, and how do I get him to go away?" I'm grateful he didn't. I'd also like to thank my editors, Michele Cronin and Marie Beaugureau, for guiding me through the publishing process and getting the book in a much better state than I ever would have gotten it on my own.

I couldn't have written this book if I'd never learned data science, and I probably wouldn't have learned data science if not for the influence of Dave Hsu, Igor Tatarinov, John Rauser, and the rest of the Farecast gang. (So long ago that it wasn't even called data science at the time!) The good folks at Coursera and DataTau deserve a lot of credit, too.

I am also grateful to my beta readers and reviewers. Jay Fundling found a ton of mistakes and pointed out many unclear explanations, and the book is much better (and much more correct) thanks to him. Debashis Ghosh is a hero for sanity-checking all of my statistics. Andrew Musselman suggested toning down the "people who prefer R to Python are moral reprobates" aspect of the book, which I think ended up being pretty good advice. Trey Causey, Ryan Matthew Balfanz, Loris Mularoni, Núria Pujol, Rob Jefferson, Mary Pat Campbell, Zach Geary, Denise Mauldin, Jimmy O'Donnell, and Wendy Grus also provided invaluable feedback. Thanks to everyone who read the first edition and helped make this a better book. Any errors remaining are of course my responsibility.

I owe a lot to the Twitter #datascience commmunity, for exposing me to a ton of new concepts, introducing me to a lot of great people, and making me feel like enough of an underachiever that I went out and wrote a book to compensate. Special thanks to Trey Causey (again), for (inadvertently) reminding me to include a chapter on linear algebra, and to Sean J. Taylor, for (inadvertently) pointing out a couple of huge gaps in the "Working with Data" chapter.

Above all, I owe immense thanks to Ganga and Madeline. The only thing harder than writing a book is living with someone who's writing a book, and I couldn't have pulled it off without their support.

O'REILLY®

Data Science from Scratch

To really learn data science, you should not only master the tools—data science libraries, frameworks, modules, and toolkits—but also understand the ideas and principles underlying them. Updated for Python 3.6, this second edition of *Data Science from Scratch* shows you how these tools and algorithms work by implementing them from scratch.

If you have an aptitude for mathematics and some programming skills, author Joel Grus will help you get comfortable with the math and statistics at the core of data science, and with the hacking skills you need to get started as a data scientist. Packed with new material on deep learning, statistics, and natural language processing, this updated book shows you how to find the gems in today's messy glut of data.

- **Get a crash course in Python**
- **Learn the basics of linear algebra, statistics, and probability—and how and when they're used in data science**
- **Collect, explore, clean, munge, and manipulate data**
- **Dive into the fundamentals of machine learning**
- **Implement models such as k-nearest neighbors, Naive Bayes, linear and logistic regression, decision trees, neural networks, and clustering**
- **Explore recommender systems, natural language processing, network analysis, MapReduce, and databases**

Joel Grus is a research engineer at the Allen Institute for Artificial Intelligence. Previously he worked as a software engineer at Google and as a data scientist at several startups. He lives in Seattle, where he regularly attends data science happy hours. He blogs infrequently at *joelgrus.com* and tweets all day long at *@joelgrus*.

"Joel takes you on a journey from being data-curious to getting a thorough understanding of the bread-and-butter algorithms that every data scientist should know."

—Rohit Sivaprasad
Engineer, Facebook

"I've recommended *Data Science from Scratch* to analysts and engineers wanting to make the jump into machine learning. It's the best tool for understanding the fundamentals of the discipline."

—Tom Marthaler
Engineering Manager, Amazon

"Translating data science concepts into code is hard. Joel's book makes it much easier."

—William Cox
Machine Learning Engineer, Grubhub

DATA / DATA SCIENCE

For sale in the Indian Subcontinent (India, Pakistan, Bangladesh, Nepal, Sri Lanka, Bhutan, Maldives) and African Continent (excluding Morocco, Algeria, Tunisia, Libya, Egypt, and the Republic of South Africa) only. Illegal for sale outside of these countries.

ISBN : 978-93-5213-832-6

9 789352 138326

Second Edition/2019/Paperback/English

SPD®

MRP: ₹ 1,000.00

Twitter: @oreillymedia
facebook.com/oreilly

SHROFF PUBLISHERS & DISTRIBUTORS PVT. LTD.