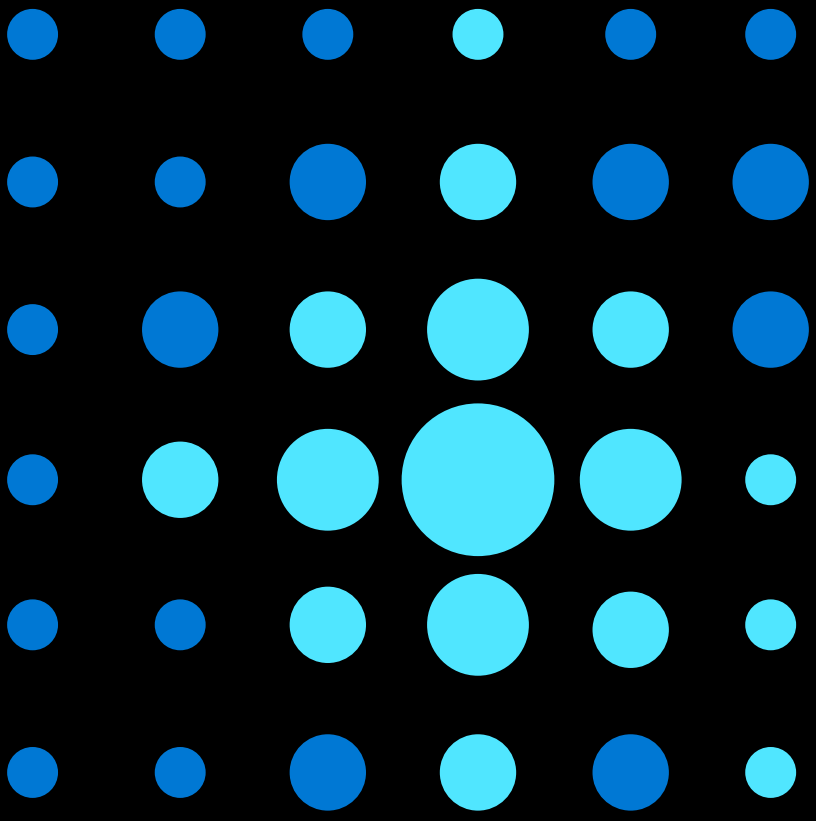
O'REILLY®


# Thoughtful Machine Learning with Python

A TEST-DRIVEN APPROACH

Matthew Kirk

Microsoft Azure



# Build machine learning models easily and quickly

Overcome the complexity of building, training, and deploying machine learning models. Accelerate your path to production, scale on demand, and gain insights from cloud to edge.

Find out more about using Azure Machine Learning service with your favorite open-source tools and frameworks.

**Learn more >**

# Thoughtful Machine Learning with Python

*A Test-Driven Approach*

*Matthew Kirk*

Beijing • Boston • Farnham • Sebastopol • Tokyo

O'REILLY®

**Thoughtful Machine Learning with Python**
by Matthew Kirk



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

**Editors:** Mike Loukides and Shannon Cutt
**Production Editor:** Nicholas Adams
**Copyeditor:** James Fraleigh
**Proofreader:** Charles Roumeliotis
**Indexer:** Wendy Catalano
**Interior Designer:** David Futato
**Cover Designer:** Randy Comer
**Illustrator:** Rebecca Demarest

January 2017: First Edition

**Revision History for the First Edition**
2017-01-10: First Release
2017-10-20: Second Release

See *http://oreilly.com/catalog/errata.csp?isbn=9781491924136* for release details.



978-1-491-08304-7

[LSI]

# Table of Contents

# Foreword

Machine learning is not an entirely new subject, but it has gained more popularity in recent years as organizations accelerate development of AI solutions.

Author Matthew Kirk takes readers through the basics of machine learning, with topics such as neural networks, K-Nearest Neighbors (KNNs), clustering, and other algorithms; applying test-driven development (TDD); exploring techniques for improving ML models; and more. This practical guide features code examples with Python's NumPy, Pandas, Scikit-Learn, and SciPy data science libraries. Kirk brings these learnings full circle, with references to real-world examples and engaging, hands-on exercises.

While this book is not intended to be an exhaustive introduction to machine learning, it is designed to help the readers learn the fundamentals, understand the various machine learning algorithms and their applications, and develop a framework to build machine learning solutions.

Microsoft designed Azure Machine Learning service to provide a platform to build, train, and deploy machine learning models easily from cloud to edge. We hope you enjoy the book and consider Azure Machine Learning to accelerate your path to developing high-quality models and AI solutions.

*— Bharat Sandhu*
*Director, Azure AI Platform*
*Microsoft*

# Preface

I wrote the first edition of *Thoughtful Machine Learning* out of frustration over my coworkers' lack of discipline. Back in 2009 I was working on lots of machine learning projects and found that as soon as we introduced support vector machines, neural nets, or anything else, all of a sudden common coding practice just went out the window.

*Thoughtful Machine Learning* was my response. At the time I was writing 100% of my code in Ruby and wrote this book for that language. Well, as you can imagine, that was a tough challenge, and I'm excited to present a new edition of this book rewritten for Python. I have gone through most of the chapters, changed the examples, and made it much more up to date and useful for people who will write machine learning code. I hope you enjoy it.

As I stated in the first edition, my door is always open. If you want to talk to me for any reason, feel free to drop me a line at *matt@matthewkirk.com*. And if you ever make it to Seattle, I would love to meet you over coffee.

## 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 general note.

## Using Code Examples

Supplemental material (code examples, exercises, etc.) is available for download at *http://github.com/thoughtfulml/examples-in-python*.

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: "*Thoughtful Machine Learning with Python* by Matthew Kirk (O'Reilly). Copyright 2017 Matthew Kirk, 978-1-491-92413-6."

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 Safari


*Safari* (formerly Safari Books Online) is a membership-based training and reference platform for enterprise, government, educators, and individuals.

Members have access to thousands of books, training videos, Learning Paths, interactive tutorials, and curated playlists from over 250 publishers, including O'Reilly Media, Harvard Business Review, Prentice Hall Professional, Addison-Wesley Professional, Microsoft Press, Sams, Que, Peachpit Press, Adobe, 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, and Course Technology, among others.

For more information, please visit *http://oreilly.com/safari*.

## 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/thoughtful-machine-learning-with-python*.

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

I've waited over a year to finish this book. My diagnosis of testicular cancer and the sudden death of my dad forced me take a step back and reflect before I could come to grips with writing again. Even though it took longer than I estimated, I'm quite pleased with the result.

I am grateful for the support I received in writing this book: everybody who helped me at O'Reilly and with writing the book. Shannon Cutt, my editor, who was a rock and consistently uplifting. Liz Rush, the sole technical reviewer who was able to make it through the process with me. Stephen Elston, who gave helpful feedback. Mike Loukides, for humoring my idea and letting it grow into two published books. Alexey Porotnikov who helped me extensively with the Python coding examples.

I also want to give special thanks to Alexey Porotnikov (*https://github.com/alpo*) who painstakingly helped me convert all these examples from Ruby to Python and also from Python 2 to Python 3. Seriously, thank you!

I'm grateful for friends, most especially Curtis Fanta. We've known each other since we were five. Thank you for always making time for me (and never being deterred by my busy schedule).

To my family. For my nieces Zoe and Darby, for their curiosity and awe. To my brother Jake, for entertaining me with new music and movies. To my mom Carol, for letting me discover the answers, and advising me to take physics (even though I never have). You all mean so much to me.

To the Le family, for treating me like one of their own. Thanks to Liliana for the Lego dates, and Sayone and Alyssa for being bright spirits in my life. For Martin and Han for their continual support and love. To Thanh (Dad) and Kim (Mom) for feeding me more food than I probably should have, and for giving me multimeters and books on opamps. Thanks for being a part of my life.

To my grandma, who kept asking when she was going to see the cover. You're always pushing me to achieve, be it through Boy Scouts or owning a business. Thank you for always being there.

To Sophia, my wife. A year ago, we were in a hospital room while I was pumped full of painkillers…and we survived. You've been the most constant pillar of my adult life. Whenever I take on a big hairy audacious goal (like writing a book), you always put your needs aside and make sure I'm well taken care of. You mean the world to me.

Last, to my dad. I miss your visits and our camping trips to the woods. I wish you were here to share this with me, but I cherish the time we did have together. This book is for you.