Preface

Why We Wrote This Book

We've both spent most of our careers automating things. When we first met and Alfredo didn't know Python, Noah suggested automating one task per week. Automation is a core pillar for MLOps, DevOps, and this book throughout. You should take all the examples and opinions in this book in the context of future automation.

If Noah could summarize how he spent 2000–2020, it was automating just about anything he could, from film pipelines to software installation to machine learning pipelines. As an engineering manager and CTO at startups in the Bay Area, he built many data science teams from scratch. As a result, he saw many of the core problems in getting machine learning to production in the early stages of the AI/ML revolution.

Noah has been an adjunct professor at Duke, Northwestern, and UC Davis in the last several years, teaching topics that primarily focus on cloud computing, data science, and machine learning engineering. This teaching and work experience gives him a unique perspective about the issues involved in the real-world deployment of machine learning solutions.

Alfredo has a heavy ops background from his Systems Administrator days, with a similar passion for automation. It is not possible to build resilient infrastructure without push-button automation. There is nothing more gratifying when disaster situations happen than rerunning a script or a pipeline to re-create what crashed.

When COVID-19 hit, it accelerated a question we both had, which was, "why aren't we putting more models into production?" Noah touched on some of these issues in an article he wrote for Forbes. The summarized premise of the article is that something is wrong with data science because organizations are not seeing returns on their investments.

Later at O'Reilly's "Foo Camp", Noah led a session on "Why can we not be 10X faster at ML in production?" where we had a great discussion with many people, including Tim O'Reilly, Mike Loukides, Roger Magoulas, and others. The result of that discussion was: "Yes, we can go 10X faster." So thanks to Tim and Mike for stirring such a fascinating discussion and getting this book on its way.

Machine learning feels a lot like many other technologies that have appeared in the past several decades. At first, it takes years to get results. Steve Jobs talked about how NeXT wanted to make it 10X faster to build software (and he did). You can watch the interview on YouTube. What are some of the problems with machine learning currently?

- Focus on the "code" and technical details versus the business problem
- Lack of automation
- HiPPO (Highest Paid Person's Opinions)
- Not cloud native
- Lack of urgency to solve solvable problems

Quoting one of the things Noah brought up in the discussion: "I'm anti-elitism across the board. Programming is a human right. The idea that there is some priesthood that is only allowed to do it is just wrong." Similar to machine learning, it is too crucial for technology to lie only in the hands of a select group of people. With MLOps and AutoML, these technologies can go into the public's hands. We can do better with machine learning and artificial intelligence by democratizing this technology. "Real" AI/ML practitioners ship models to production, and in the "real" future, people such as doctors, lawyers, mechanics, and teachers will use AI/ML to help them do their jobs.

How This Book Is Organized

We designed this book so that you can consume each chapter as a standalone section designed to give you immediate help. At the end of each chapter are discussion questions that are intended to spur critical thinking and technical exercises to improve your understanding of the material.

These discussion questions and exercises are also well suited for use in the classroom in a Data Science, Computer Science, or MBA program and for the motivated self-learner. The final chapter contains several case studies helpful in building a work portfolio as an expert in MLOps.

The book is divided into 12 chapters, which we'll break down a little more in the following section. At the end of the book, there is an appendix with a collection of valuable resources for implementing MLOps.

Chapters

The first few chapters cover the theory and practice of both DevOps and MLOps. One of the items covered is how to set up continuous integration and continuous delivery. Another critical topic is Kaizen, i.e., the idea of continuous improvement in everything.

There are three chapters on cloud computing that cover AWS, Azure, and GCP. Alfredo, a developer advocate for Microsoft, is an ideal source of knowledge for MLOps on the Azure platform. Likewise, Noah has spent years getting students trained on cloud computing and working with the education arms of Google, AWS, and Azure. These chapters are an excellent way to get familiar with cloud-based MLOps.

Other chapters cover critical technical areas of MLOps, including AutoML, containers, edge computing, and model portability. These topics encompass many cutting-edge emerging technologies with active traction.

Finally, in the last chapter, Noah covers a real-world case study of his time at a social media startup and the challenges they faced doing MLOps.

Appendixes

The appendixes are a collection of essays, ideas, and valuable items that cropped up in years between finishing *Python for DevOps* (O'Reilly) and this book. The primary way to use them is to help you make decisions about the future.

Exercise Questions

In this book's exercises, a helpful heuristic considers how you can leverage them into a portfolio using GitHub and a YouTube walkthrough of what you did. In keeping with the expression "a picture is worth a thousand words," a YouTube link to a walkthrough of a reproducible GitHub project on a resume may be worth 10,000 words and puts the resume in a new category of qualification for a job.

As you go through the book and exercises, consider the following critical thinking framework.

Discussion Questions

According to Jonathan Haber in *Critical Thinking* (MIT Press Essential Knowledge series) and the nonprofit Foundation for Critical Thinking, discussion questions are essential critical thinking components. The world is in dire need of critical thinking due to the proliferation of misinformation and shallow content in social media. A mastery of the following skills sets an individual apart from the pack:

Intellectual humility
: Recognition of the limits of your knowledge.

Intellectual courage
: The ability to argue for your beliefs even in the face of social pressure.

Intellectual empathy
: The ability to put yourself in the minds of others to understand their position.

Intellectual autonomy
: The ability to think for yourself independently of others.

Intellectual integrity
: The ability to think and argue with the same intellectual standards you expect others to apply to you.

Intellectual perseverance
: The ability to provide evidence that supports your position.

Confidence in reason
: The belief that there are indisputable facts and that reason is the best solution to gain knowledge.

Fairmindedness
: The ability to put in the good-faith effort to treat all viewpoints fairly.

Using these criteria, evaluate the discussion questions in each chapter.

Origin of Chapter Quotes

By Noah

I graduated college in late 1998 and spent a year training to play professional basketball in the minor leagues in the United States or Europe while working as a personal trainer. My backup plan was to get a job in IT. I applied to be a Systems Administrator at Caltech in Pasadena and got a Mac IT expert position on a fluke. I decided the risk/reward ratio of being a low-paid professional athlete wasn't worth it and accepted the job offer.

To say Caltech changed my life is an understatement. At lunch, I played ultimate frisbee and heard about the Python programming language, which I learned so I would "fit in" with my ultimate frisbee friends, who were staff or students at Caltech. Later, I worked directly for Caltech's administration and was the personal Mac expert at Caltech for Dr. David Baltimore, who got the Nobel Prize in his 30s. I interacted with many famous people in many unexpected ways, which boosted my self-confidence and grew my network.

I also had many Forrest Gump–style random encounters with people who would later do incredible things in AI/ML. Once, I had dinner with Dr. Fei-Fei Li, head of AI at Stanford, and her boyfriend; I remember being impressed that her boyfriend spent the summer writing a video game with his dad. I was highly impressed and thought, "Who does that kind of thing?" Later, I set up a mail server under the famous physicist Dr. David Goodstein's desk because he kept getting grief from IT about hitting his mailbox storage limits. These experiences are where I acquired a taste for building "shadow infrastructure." Because I worked directly for the administration, I got to flaunt the rules if there was a good reason for it.

One of the people I randomly met was Dr. Joseph Bogen, a neurosurgeon and visiting professor at Caltech. Of all the people I met at Caltech, he had the most profound impact on my life. One day I responded to a help desk call to come to his house to fix his computer, and later this turned into a weekly dinner at his home with him and his wife, Glenda. From around 2000 until the day he died, he was a friend and mentor.

At the time, I was very interested in artificial intelligence, and I remember a Caltech Computer Science professor telling me it was a dead field and I shouldn't focus on it. Despite that advice, I came up with a plan to be fluent in many software programming languages by 40 years old and writing artificial intelligence programs by then. Lo and behold, my plan worked out.

I can clearly say I wouldn't be doing what I am doing today if I didn't meet Joe Bogen. He blew my mind when he told me he did the first hemispherectomy, removing half of a brain, to help a patient with severe epilepsy. We would talk for hours about the origins of consciousness, the use of neural networks in the 1970s to figure out who would be an Air Force pilot, and whether your brain contained "two of you," one in each hemisphere. Above all, what Bogen gave me was a sense of confidence in my intellect. I had severe doubts up until that point about what I could do, but our conversations were like a master's degree in higher-level thinking. As a professor myself, I think about how big of an impact he had on my life, and I hope to pay it forward to other students I interact with, both as a formal teacher or someone they meet. You can read these quotes yourself from an archive of Dr. Bogen's Caltech home page and his biography.

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/paiml/practical-mlops-book*.

If you have a technical question or a problem using the code examples, please send email to *bookquestions@oreilly.com*.

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 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 generally do not require, attribution. An attribution usually includes the title, author, publisher, and ISBN. For example: "*Practical MLOps* by Noah Gift and Alfredo Deza (O'Reilly). Copyright 2021 Noah Gift and Alfredo Deza, 978-1-098-10301-9."

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 more than 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, 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, 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 *https://oreil.ly/practical-mlops*.

Email *bookquestions@oreilly.com* to comment or ask technical questions about this book.

For news and information about our books and courses, visit *http://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

From Noah

As mentioned earlier, without Mike Loukides inviting me to Foo Camp and having a great discussion with Tim O'Reilly and me, this book wouldn't be here. Next, I would like to acknowledge Alfredo, my coauthor. I have had the pleasure of writing five books, two for O'Reilly and three self-published, in a little over two years with Alfredo, and this is primarily due to his ability to embrace work and get things done. An appetite for hard work is perhaps the best talent, and Alfredo has this skill in abundance.

Our editor, Melissa Potter, did tremendous work getting things into shape, and the book before she edited and afterward are almost two different books. I feel lucky to have worked with such a talented editor.

Our technical editors, including Steve Depp, Nivas Durairaj, and Shubham Saboo, played a crucial role in giving us great feedback about where to zig and when to zag. Many enhancements are particularly due to Steve's thorough feedback. Also, I wanted to thank Julien Simon and Piero Molino for enhancing our book with real-world thoughts on MLOps.

I want to thank my family, Liam, Leah, and Theodore, for giving me the space to finish this book on a tight deadline in the middle of a pandemic. I am also looking forward to reading some of the books they write in the future. Another big group of thanks goes out to all the former students I taught at Northwestern, Duke, UC Davis, and other schools. Many of their questions and feedback made it into this book.

My final thanks go out to Dr. Joseph Bogen, an early pioneer in AI/ML and Neuroscience. If we didn't bump into each other at Caltech, there is zero chance I would be a professor or that this book would exist. His impact was that big on my life.

From Alfredo

I'm absolutely thankful for my family's support while writing this book: Claudia, Efrain, Ignacio, and Alana—your support and patience were essential to get to the finish line. Thanks again for all the opportunities to work with you, Noah; this was another incredible ride. I value your friendship and our professional relationship.

Thanks to Melissa Potter (without a doubt the best editor I've worked with) for her fantastic work. Our technical editors did great by finding problems and highlighting places that needed refinement, always a hard thing to do well.

Also extremely grateful for Lee Stott's help with Azure. The Azure content wouldn't be as good without it. And thanks to Francesca Lazzeri, Mike McCoy, and everyone else I contacted at Microsoft about the book. You were all very helpful.