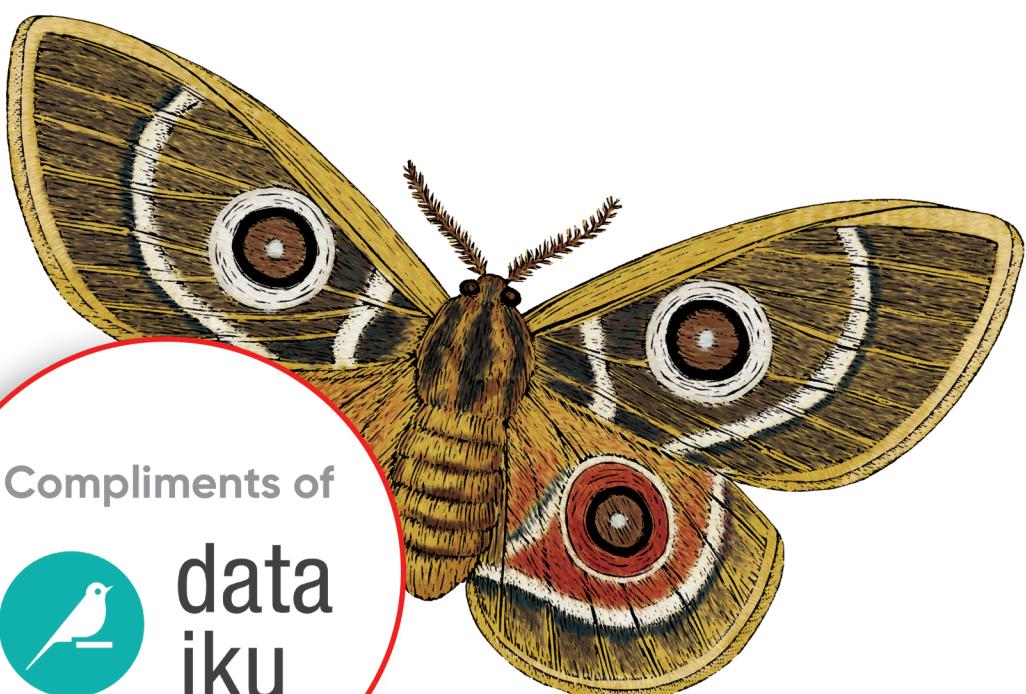


O'REILLY®

Introducing MLOps

How to Scale Machine Learning in the Enterprise



Compliments of
 data
iku

Mark Treveil
& the Dataiku Team

Mastering MLops

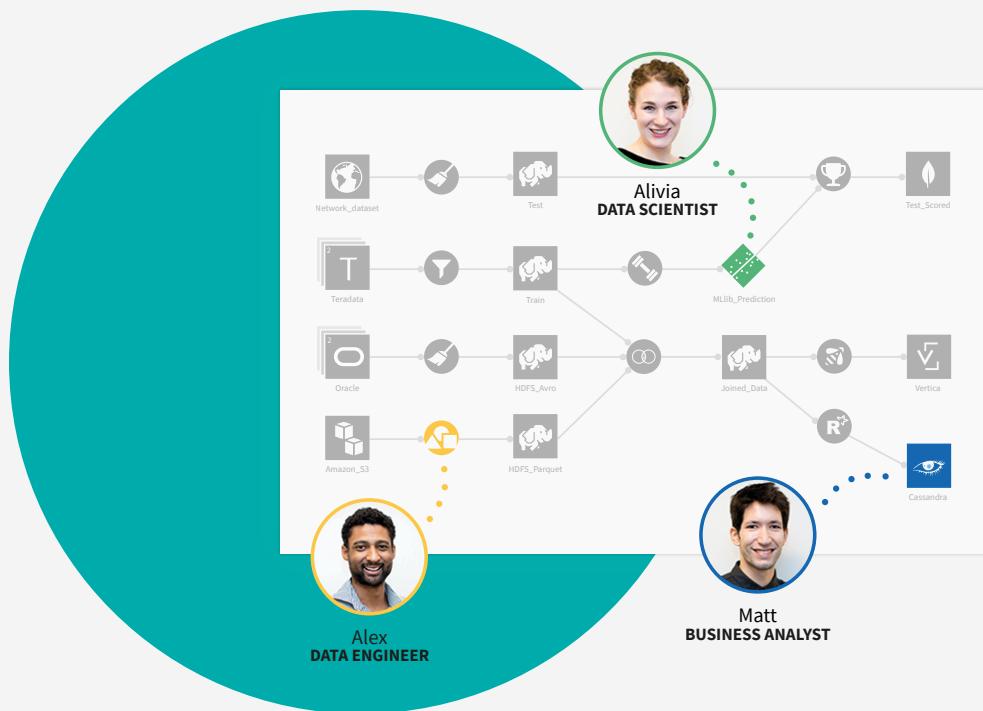
with Dataiku



Dataiku is the only platform that provides one simple, consistent UI for data connection, wrangling, mining, visualization, machine learning, deployment, and model monitoring, all at enterprise scale.

KEY FEATURES FOR A SCALABLE MLOPS STRATEGY INCLUDE:

- 1 **Model input drift detection** that looks at the recent data the model has had to score and statistically compares it with the data on which the model was evaluated.
- 2 Easier creation of **validation feedback loops** via Dataiku Evaluation Recipes to compute the true performance of a saved model against a new validation dataset, plus automated retraining and redeployment.
- 3 **Dashboard interfaces** dedicated to the monitoring of global pipelines.
- 4 ...and more! Go in-depth on all the **features Dataiku** has to offer at [dataiku.com](https://www.dataiku.com)



Introducing MLOps

How to Scale Machine Learning in the Enterprise

Mark Treveil and the Dataiku Team

Beijing • Boston • Farnham • Sebastopol • Tokyo

O'REILLY®

Introducing MLOps

by Mark Treveil, and the Dataiku Team

Copyright © 2020 Dataiku. All rights reserved.

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

Acquisitions Editor: Rebecca Novack

Development Editor: Angela Rufino

Production Editor: Katherine Tozer

Copyeditor: Penelope Perkins

Proofreader: Kim Wimpsett

Indexer: Ellen Troutman-Zaig

Interior Designer: David Futato

Cover Designer: Karen Montgomery

Illustrator: Kate Dullea

December 2020: First Edition

Revision History for the First Edition

2020-11-30: First Release

See <http://oreilly.com/catalog/errata.csp?isbn=9781492083290> for release details.

The O'Reilly logo is a registered trademark of O'Reilly Media, Inc. *Introducing MLOps*, the cover image, and related trade dress are trademarks of O'Reilly Media, Inc.

The views expressed in this work are those of the authors, and do not represent the publisher's views. While the publisher and the authors have used good faith efforts to ensure that the information and instructions contained in this work are accurate, the publisher and the authors 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.

This work is part of a collaboration between O'Reilly and Dataiku. See our [statement of editorial independence](#).

978-1-492-08330-6

[LSI]

Table of Contents

Preface.....	ix
--------------	----

Part I. MLOps: What and Why

1. Why Now and Challenges.....	3
Defining MLOps and Its Challenges	4
MLOps to Mitigate Risk	7
Risk Assessment	8
Risk Mitigation	9
MLOps for Responsible AI	9
MLOps for Scale	10
Closing Thoughts	11
2. People of MLOps.....	13
Subject Matter Experts	15
Data Scientists	17
Data Engineers	19
Software Engineers	20
DevOps	20
Model Risk Manager/Auditor	21
Machine Learning Architect	21
Closing Thoughts	22
3. Key MLOps Features.....	23
A Primer on Machine Learning	23
Model Development	24
Establishing Business Objectives	24

Data Sources and Exploratory Data Analysis	24
Feature Engineering and Selection	25
Training and Evaluation	26
Reproducibility	26
Responsible AI	26
Productionalization and Deployment	27
Model Deployment Types and Contents	28
Model Deployment Requirements	29
Monitoring	29
DevOps Concerns	30
Data Scientist Concerns	30
Business Concerns	31
Iteration and Life Cycle	32
Iteration	32
The Feedback Loop	33
Governance	34
Data Governance	36
Process Governance	37
Closing Thoughts	38

Part II. MLOps: How

4. Developing Models.....	41
What Is a Machine Learning Model?	42
In Theory	42
In Practice	43
Required Components	44
Different ML Algorithms, Different MLOps Challenges	45
Data Exploration	46
Feature Engineering and Selection	47
Feature Engineering Techniques	47
How Feature Selection Impacts MLOps Strategy	48
Experimentation	49
Evaluating and Comparing Models	51
Choosing Evaluation Metrics	51
Cross-Checking Model Behavior	53
Impact of Responsible AI on Modeling	53
Version Management and Reproducibility	56
Closing Thoughts	58

5. Preparing for Production.....	59
Runtime Environments	60
Adaptation from Development to Production Environments	60
Data Access Before Validation and Launch to Production	62
Final Thoughts on Runtime Environments	62
Model Risk Evaluation	63
The Purpose of Model Validation	63
The Origins of ML Model Risk	64
Quality Assurance for Machine Learning	64
Key Testing Considerations	65
Reproducibility and Auditability	66
Machine Learning Security	67
Adversarial Attacks	68
Other Vulnerabilities	68
Model Risk Mitigation	69
Changing Environments	70
Interactions Between Models	70
Model Misbehavior	71
Closing Thoughts	72
6. Deploying to Production.....	73
CI/CD Pipelines	73
Building ML Artifacts	75
What's in an ML Artifact?	75
The Testing Pipeline	75
Deployment Strategies	77
Categories of Model Deployment	77
Considerations When Sending Models to Production	78
Maintenance in Production	79
Containerization	79
Scaling Deployments	81
Requirements and Challenges	83
Closing Thoughts	84
7. Monitoring and Feedback Loop.....	85
How Often Should Models Be Retrained?	86
Understanding Model Degradation	89
Ground Truth Evaluation	89
Input Drift Detection	91
Drift Detection in Practice	92
Example Causes of Data Drift	93
Input Drift Detection Techniques	93

The Feedback Loop	95
Logging	96
Model Evaluation	97
Online Evaluation	99
Closing Thoughts	103
8. Model Governance.....	105
Who Decides What Governance the Organization Needs?	105
Matching Governance with Risk Level	107
Current Regulations Driving MLOps Governance	108
Pharmaceutical Regulation in the US: GxP	109
Financial Model Risk Management Regulation	109
GDPR and CCPA Data Privacy Regulations	110
The New Wave of AI-Specific Regulations	111
The Emergence of Responsible AI	112
Key Elements of Responsible AI	113
Element 1: Data	113
Element 2: Bias	114
Element 3: Inclusiveness	115
Element 4: Model Management at Scale	116
Element 5: Governance	116
A Template for MLOps Governance	117
Step 1: Understand and Classify the Analytics Use Cases	118
Step 2: Establish an Ethical Position	118
Step 3: Establish Responsibilities	119
Step 4: Determine Governance Policies	120
Step 5: Integrate Policies into the MLOps Process	121
Step 6: Select the Tools for Centralized Governance Management	122
Step 7: Engage and Educate	123
Step 8: Monitor and Refine	124
Closing Thoughts	125

Part III. MLOps: Real-World Examples

9. MLOps in Practice: Consumer Credit Risk Management.....	129
Background: The Business Use Case	129
Model Development	130
Model Bias Considerations	131
Prepare for Production	131
Deploy to Production	132
Closing Thoughts	133

10. MLOps in Practice: Marketing Recommendation Engines.....	135
The Rise of Recommendation Engines	135
The Role of Machine Learning	136
Push or Pull?	136
Data Preparation	137
Design and Manage Experiments	138
Model Training and Deployment	138
Scalability and Customizability	140
Monitoring and Retraining Strategy	140
Real-Time Scoring	140
Ability to Turn Recommendations On and Off	141
Pipeline Structure and Deployment Strategy	141
Monitoring and Feedback	142
Retraining Models	142
Updating Models	143
Runs Overnight, Sleeps During Daytime	143
Option to Manually Control Models	143
Option to Automatically Control Models	144
Monitoring Performance	144
Closing Thoughts	145
11. MLOps in Practice: Consumption Forecast.....	147
Power Systems	147
Data Collection	149
Problem Definition: Machine Learning, or Not Machine Learning?	151
Spatial and Temporal Resolution	151
Implementation	153
Modeling	153
Deployment	155
Monitoring	156
Closing Thoughts	157
Index.....	159

Preface

We've reached a turning point in the story of machine learning where the technology has moved from the realm of theory and academics and into the "real world"—that is, businesses providing all kinds of services and products to people across the globe. While this shift is exciting, it's also challenging, as it combines the complexities of machine learning models with the complexities of the modern organization.

One difficulty, as organizations move from experimenting with machine learning to scaling it in production environments, is maintenance. How can companies go from managing just one model to managing tens, hundreds, or even thousands? This is not only where MLOps comes into play, but it's also where the aforementioned complexities, both on the technical and business sides, appear. This book will introduce readers to the challenges at hand, while also offering practical insights and solutions for developing MLOps capabilities.

Who This Book Is For

We wrote this book specifically for analytics and IT operations team managers, that is, the people directly facing the task of scaling machine learning (ML) in production. Given that MLOps is a new field, we developed this book as a guide for creating a successful MLOps environment, from the organizational to the technical challenges involved.

How This Book Is Organized

This book is divided into three parts. The first is an introduction to the topic of MLOps, diving into how (and why) it has developed as a discipline, who needs to be involved to execute MLOps successfully, and what components are required.

The second part roughly follows the machine learning model life cycle, with chapters on developing models, preparing for production, deploying to production, monitoring, and governance. These chapters cover not only general considerations, but

MLOps considerations at each stage of the life cycle, providing more detail on the topics touched on in [Chapter 3](#).

The final part provides tangible examples of how MLOps looks in companies today, so that readers can understand the setup and implications in practice. Though the company names are fictitious, the stories are based on real-life companies' experience with MLOps and model management at scale.

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.

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/intro-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

We would like to thank the entire Dataiku team for their support in developing this book, from conception to completion. It's been a true team effort and, like most things we do at Dataiku, rooted in fundamental collaboration between countless people and teams.

Thanks to those who supported our vision from the beginning of writing this book with O'Reilly. Thanks to those who stepped in to help with writing and editing. Thanks to those who provided honest feedback (even when it meant more writing and rewriting and re-writing). Thanks to those who were internal cheerleaders and, of course, those who helped us promote the finished product to the world.

PART I

MLOps: What and Why

CHAPTER 1

Why Now and Challenges

Machine learning operations (MLOps) is quickly becoming a critical component of successful data science project deployment in the enterprise (Figure 1-1). It's a process that helps organizations and business leaders generate long-term value and reduce risk associated with data science, machine learning, and AI initiatives. Yet it's a relatively new concept; so why has it seemingly skyrocketed into the data science lexicon overnight? This introductory chapter delves into what MLOps is at a high level, its challenges, why it has become essential to a successful data science strategy in the enterprise, and, critically, why it is coming to the forefront now.

MLOps Versus ModelOps Versus AIOps

MLOps (or ModelOps) is a relatively new discipline, emerging under these names particularly in late 2018 and 2019. The two—MLOps and ModelOps—are, at the time this book is being written, largely being used interchangeably. However, some argue that ModelOps is more general than MLOps, as it's not only about machine learning models but any kind of model (e.g., rule-based models). For the purpose of this book, we'll be specifically discussing the machine learning model life cycle and will thus use the term "MLOps."

AIOps, though sometimes confused with MLOps, is another topic entirely and refers to the process of solving operational challenges through the use of artificial intelligence (i.e., AI for DevOps). An example would be a form of predictive maintenance for network failures, alerting DevOps teams to possible problems before they arise. While important and interesting in its own right, AIOps is outside the scope of this book.

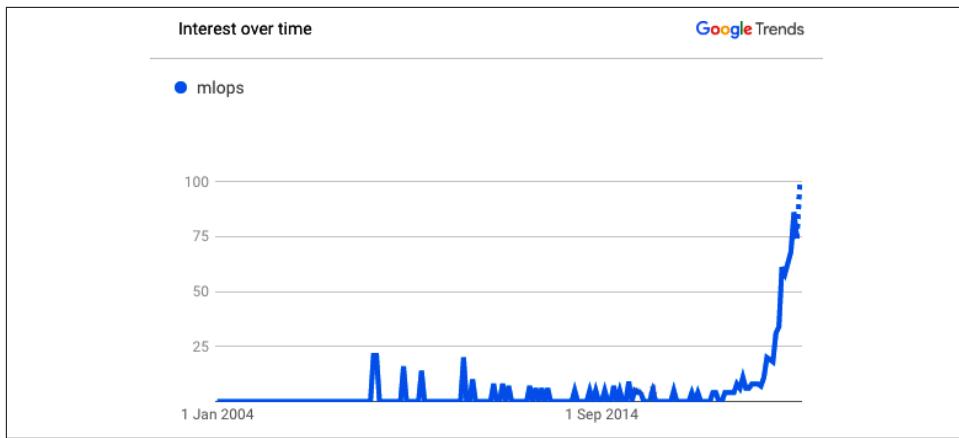


Figure 1-1. Representation of the exponential growth of MLOps (not the parallel growth of the term “ModelOps”)

Defining MLOps and Its Challenges

At its core, MLOps is the standardization and streamlining of machine learning life cycle management (Figure 1-2). But taking a step back, why does the machine learning life cycle need to be streamlined? On the surface, just looking at the steps to go from business problem to a machine learning model at a very high level, it seems straightforward.

For most traditional organizations, the development of multiple machine learning models and their deployment in a production environment are relatively new. Until recently, the number of models may have been manageable at a small scale, or there was simply less interest in understanding these models and their dependencies at a company-wide level. With decision automation (that is, an increasing prevalence of decision making that happens without human intervention), models become more critical, and, in parallel, managing model risks becomes more important at the top level.

The reality of the machine learning life cycle in an enterprise setting is much more complex, in terms of needs and tooling (Figure 1-3).

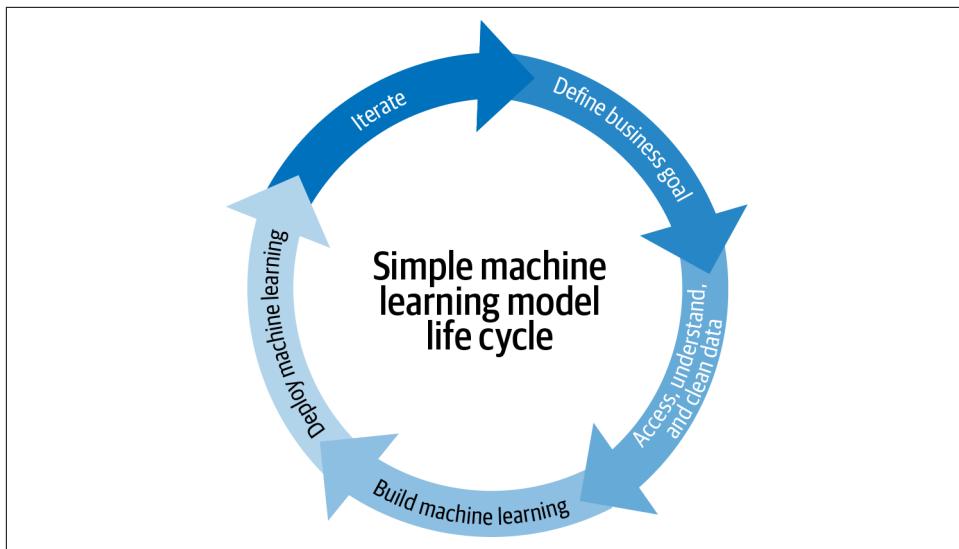


Figure 1-2. A simple representation of the machine learning model life cycle, which often underplays the need for MLOps, compared to Figure 1-3

There are three key reasons that managing machine learning life cycles at scale is challenging:

- There are many dependencies. Not only is data constantly changing, but business needs shift as well. Results need to be continually relayed back to the business to ensure that the reality of the model in production and on production data aligns with expectations and, critically, addresses the original problem or meets the original goal.
- Not everyone speaks the same language. Even though the machine learning life cycle involves people from the business, data science, and IT teams, none of these groups are using the same tools or even, in many cases, share the same fundamental skills to serve as a baseline of communication.
- Data scientists are not software engineers. Most are specialized in model building and assessment, and they are not necessarily experts in writing applications. Though this may start to shift over time as some data scientists become specialists more on the deployment or operational side, for now many data scientists find themselves having to juggle many roles, making it challenging to do any of them thoroughly. Data scientists being stretched too thin becomes especially problematic at scale with increasingly more models to manage. The complexity becomes exponential when considering the turnover of staff on data teams and, suddenly, data scientists have to manage models they did not create.

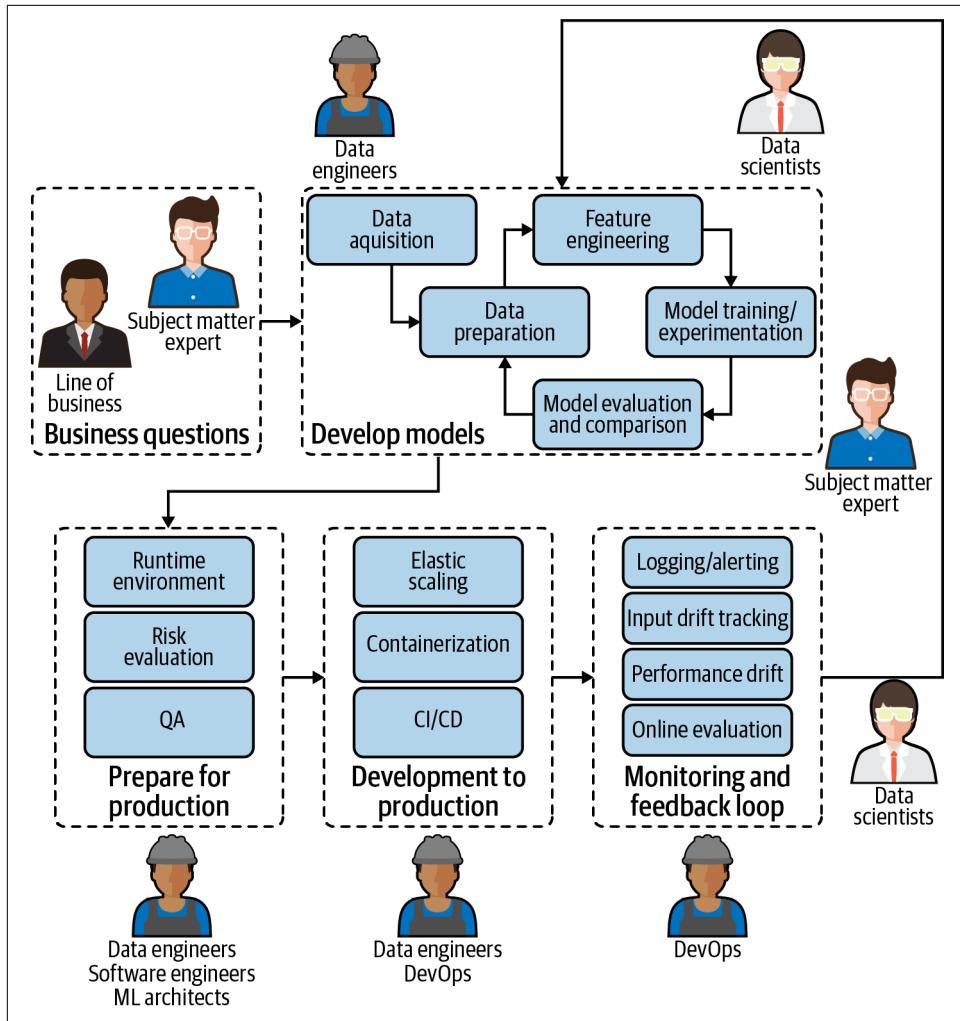


Figure 1-3. The realistic picture of a machine learning model life cycle inside an average organization today, which involves many different people with completely different skill sets and who are often using entirely different tools.

If the definition (or even the name MLOps) sounds familiar, that's because it pulls heavily from the concept of DevOps, which streamlines the practice of software changes and updates. Indeed, the two have quite a bit in common. For example, they both center around:

- Robust automation and trust between teams
- The idea of collaboration and increased communication between teams