

The State of Machine Learning Adoption in the Enterprise



Ben Lorica & Paco Nathan

Stay ahead, solve problems, and learn new skills on the O'Reilly Safari learning platform.



Structured and
unstructured
online learning



Hands-on
applied learning



Multiple formats
and learning modes



In-person events

O'Reilly Safari delivers
multimedia learning
when and where
you need it.



O'REILLY®

oreilly.com/safari

Business innovates with data. Data innovates here.

Get a first look at emerging trends and what you need to make your data strategies and implementations work—at the biggest gathering of data professionals and business managers.

Strata DATA CONFERENCE

PRESENTED BY

O'REILLY®

cloudera®



strataconf.com

The State of Machine Learning Adoption in the Enterprise

Ben Lorica and Paco Nathan

Beijing • Boston • Farnham • Sebastopol • Tokyo

O'REILLY®

The State of Machine Learning Adoption in the Enterprise

by Ben Lorica and Paco Nathan

Copyright © 2018 O'Reilly Media. 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/safari>). For more information, contact our corporate/institutional sales department: 800-998-9938 or corporate@oreilly.com.

Editor: Mac Slocum

Interior Designer: David Futato

Production Editor: Nicholas Adams

Cover Designer: Karen Montgomery

Copyeditor: Octal Publishing, Inc.

Illustrator: Rebecca Demarest

August 2018: First Edition

Revision History for the First Edition

2018-07-27: First Release

The O'Reilly logo is a registered trademark of O'Reilly Media, Inc. *The State of Machine Learning Adoption in the Enterprise*, 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.

978-1-492-04437-6

[LSI]

Table of Contents

The State of Machine Learning Adoption in the Enterprise.....	1
Introduction	1
Survey Respondents	3
Culture and Organization	5
Building Machine Learning Models	6
Evaluating Metrics for Success	8
Closing Thoughts	12

The State of Machine Learning Adoption in the Enterprise

Introduction

As machine learning has become more widely adopted by businesses, O'Reilly set out to survey our audience to learn more about how companies approach this work. Do companies with more experience deploying machine learning in production use methods that differ significantly from organizations that are just beginning? For companies that haven't begun this journey, are there any best practices that might help?

Machine learning use in production dates back about 20 years. Leo Breiman chronicled a sea change in data analytics in the notable 2001 paper "[The Two Cultures](#)," based on the introduction of machine learning models at scale during the dot-com boom. Even so, it took two decades before machine learning moved beyond the specialized "unicorns" of the internet to become a mainstream practice.

Looking at mainstream adoption for machine learning now—especially in light of recent data privacy legislation such as [General Data Protection Regulation \(GDPR\) in Europe](#) and a [similar political movement in California](#)—we wanted to probe current trends, with these questions in particular:

- How experienced are companies with machine learning adoption, in terms of number of years deploying models in production?

- What has the impact been on culture and organization; for example, have job titles changed?
- Who builds machine learning models: internal teams, external consultants, cloud APIs?
- How are decisions and priorities set and by whom within the organization?
- What methodologies apply for developing machine learning; for example, Agile?
- What metrics are used to evaluate success?

Notable findings from the survey include the following:

- Job titles specific to machine learning are already widely used at organizations with extensive experience in machine learning: data scientist (81%), machine learning engineer (39%), and deep learning engineer (20%).
- One in two (54%) respondents who belong to companies with extensive experience in machine learning check for fairness and bias. Overall, 40% of respondents indicated their organizations check for model fairness and bias. As **tutorials and training materials become available**, the number of companies capable of addressing fairness and bias should increase.
- One in two (53%) respondents who belong to companies with extensive experience in machine learning check for privacy (43% across respondents from all companies). The EU's **GDPR** mandates "privacy-by-design" ("inclusion of data protection from the onset of the designing of systems rather than an addition"), which means more companies will add privacy to their machine learning checklist. Fortunately, new regulations coincide with the rise of **tools and methods for privacy-preserving analytics** and machine learning.
- One in two (51%) respondents use internal data science teams to build their machine learning models, whereas use of AutoML services from cloud providers is in low single digits, and this split grows even more pronounced among sophisticated teams. Companies with less extensive experience tend to rely on external consultants.
- Sophisticated teams tend to have data science leads set team priorities and determine key metrics for project success—responsi-

bilities that would typically be performed by product managers in more traditional software engineering.

Survey Respondents

The survey was sent to people who previously attended one of our [Strata Data](#) or [AI Conferences](#) or have consumed content in related topics through our online properties or our ecommerce channels. The survey was open for two weeks in June 2018, and we received a substantial number of responses from many regions (see [Figure 1](#)), including North America (more than 6,000), Europe (2,000), and Asia (more than 1,700):

Where is your company located? [Total of 11,400+ respondents]

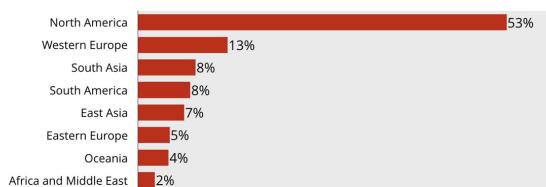


Figure 1

About half of the respondents worked for organizations that were in the early stages of exploring machine learning, whereas the rest had moderate or extensive experience deploying machine learning models to production

For the remainder of this report, we've adopted the following terminology to describe these cohorts from our survey:

Exploring

Respondents who work for organizations that are just beginning to use machine learning.

Early adopter

Respondents who work for organizations that have had machine learning models in production for more than two years.

Sophisticated

Respondents who work for organizations that have had machine learning models in production for more than five years.

It's interesting to note the geographic distribution of respondents versus the maturity of their machine learning adoption. North America and Western Europe led for sophisticated respondents, whereas East Asia and South Asia both showed a noticeable bump among the early adopters, as exhibited in Figures 2 and 3.

What is the stage of ML adoption in your organization?

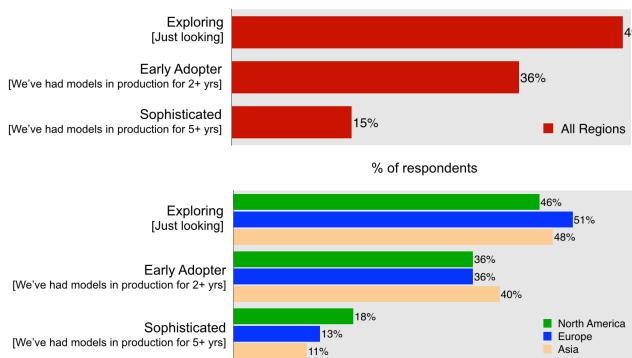


Figure 2

What is the stage of ML adoption in your organization? (By region)

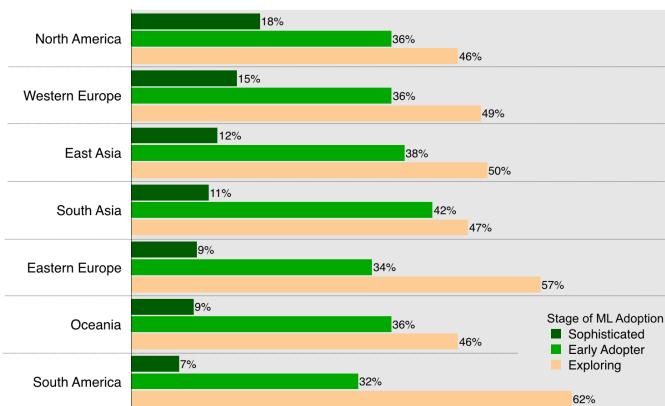


Figure 3

Culture and Organization

Although the term “machine learning” was coined in 1959, the term “data mining” was much more commonly used in the business world in the early 1990s and early 2000s. Around 2008, a group based in the San Francisco Bay Area began using the term “data scientist” to describe practitioners who drew from a variety of disciplines—including systems and data management, machine learning, statistics, and data visualization—to build data products. Many of the job titles listed in Figure 4 are fairly new, but it’s fair to say that for most companies a “data scientist” is someone who specializes in machine learning and statistics, whereas a “data engineer” is someone who specializes in building and maintaining infrastructure and data pipelines. “Machine learning (and deep learning) engineers” are relatively new job titles used to designate data engineers who specialize in building and deploying machine learning models to production. The “data ops” title is also being used, primarily at sophisticated organizations (see Figure 4).

As you can glean from Figure 4, organizations that have more experience deploying machine learning models to production are more likely to use these newer job titles (data scientist, data engineer, machine learning engineer, deep learning engineer). Experienced organizations also use the title “research scientist” to designate individuals tasked with developing sophisticated algorithms.

Which of these titles are used in the organization where you work?

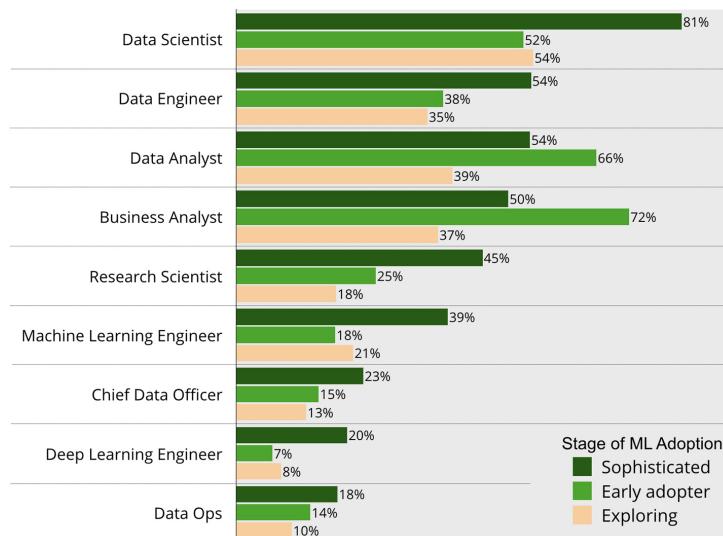
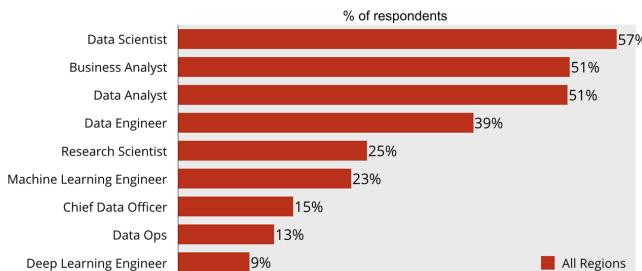


Figure 4

Building Machine Learning Models

About half of the respondents stated that machine learning models were built by their data science teams (Figure 5). However, that number rises considerably as organizations gain more experience. One of eight (12%) of those who belonged to organizations that are just beginning to explore machine learning stated that they relied on external consultants, whereas three out of four (73%) of those who belonged to the most sophisticated companies relied on their inter-

nal data science teams. Not many respondents belonged to organizations that rely on [AutoML](#) services offered by cloud providers.

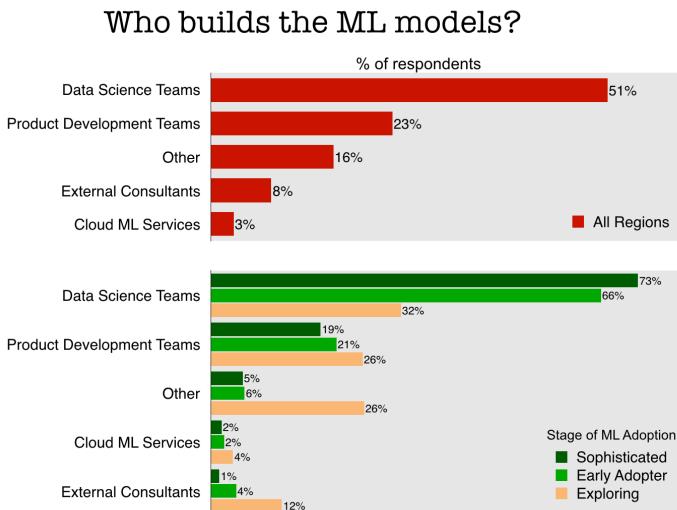


Figure 5

Machine learning adoption within businesses has reached a stage of maturity at which methodologies are being debated. As machine learning becomes more widely used, many organizations are [adapting the processes](#) they've used in software development also to build data products—examples include [Agile methodology](#) and [Kanban](#). Even so, some experts have pointed out [failures in that approach and how to work around them](#). One emerging theme is to compare and contrast the retrospective learnings from machine learning in production against the conventional wisdom of applying Agile. Also, the product management role has [not yet been clearly defined](#) for machine learning based products. We probed these questions to see how decisions and priorities are set, and by whom, within an organization.

At this stage, nearly one in three (32%) respondents used “No methodology” (see [Figure 6](#)), so we can say that Agile isn’t necessarily a default approach. Meanwhile, the sophisticated organizations clearly led on using “Other” methodologies. Although we don’t have a breakdown for what those other approaches are, that would be an interesting point to explore in future research.

What kind of methodology do you use with ML work?

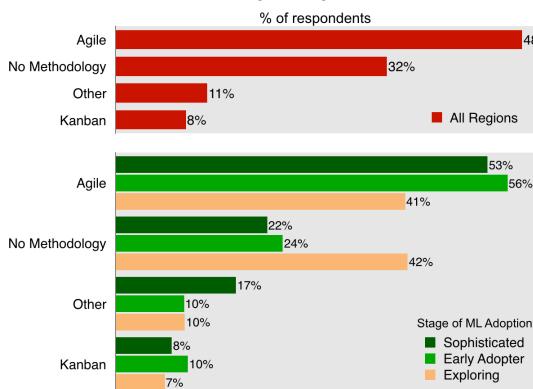


Figure 6

Evaluating Metrics for Success

In terms of evaluating key metrics for project success, generally in software development, that role would be performed by a product manager. Based on survey responses, as shown in [Figure 7](#), that's not entirely the case for machine learning. We can say that less experienced organizations placed more emphasis on product managers or executives determining the criteria for project success—as one might expect. In the more experienced organizations, that leadership aspect shifted toward data science leads.

Who determines the key metrics for project success?

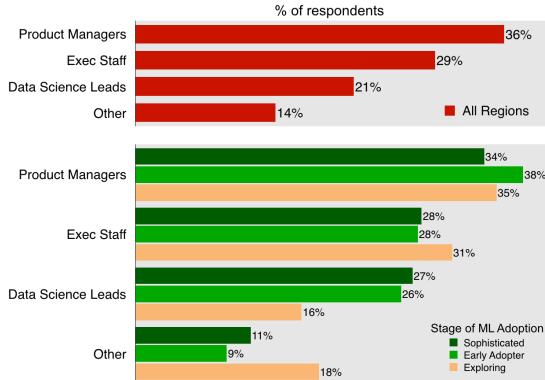


Figure 7

Similarly, you might expect that team priorities are defined primarily by product managers or executives, as in more traditional software development. As Figure 8 illustrates, clearly the product managers led on this, whereas, again, we see that for the more sophisticated organizations those decisions tended to shift over to data science leads. Although the numbers are not decisive, they indicate how machine learning adoption introduces challenges for project management and team leadership that diverge from the standard practices of software engineering.

Who sets the team priorities?

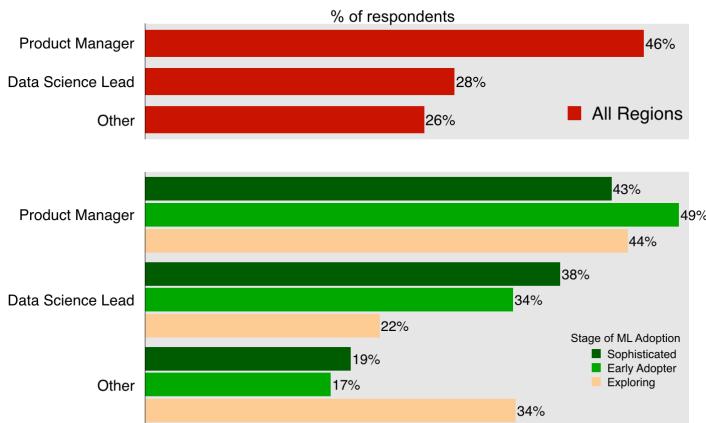


Figure 8

For most organizations, machine learning is used to enable some level of automation. Imagine going from a rule-based fraud detection system to a complex model that was automatically learned from examples. But machine learning makes sense only when there are clear benefits (improve decision making or operational efficiency, higher revenue or engagement, etc.). Because they are more likely to build models in-house and they have had more experience deploying models to production, respondents who belong to the most sophisticated organizations are likely to use multiple metrics.

In particular, one in four (26%) organizations—particularly those with extensive experience deploying machine learning—are checking for **possible bias in models**, as shown in [Figure 9](#). There is a growing interest in ethics among data professionals and researchers. Academics are organizing **conferences** and writing papers that are being translated into **best practices and checklists for industry practitioners**. But we are very much in the early days, and this remains an active area for machine learning researchers.

What are the metrics used to evaluate success?

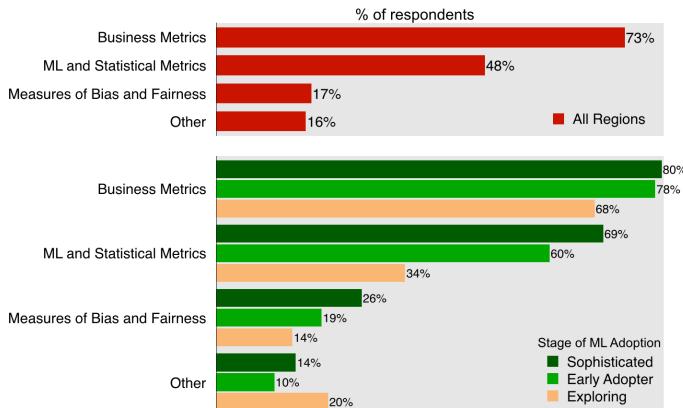


Figure 9

There are other important considerations that organizations are beginning to address. There is growing awareness among consumers, companies, and regulators about the importance of data privacy and security. The EU's GDPR mandates privacy by design ("inclusion of data protection from the onset of the designing of systems rather than an addition"). One of the more exciting trends we've been following is the growing interest in privacy-preserving analytic methods for business intelligence, analytics, and machine learning. This includes techniques and tools like differential privacy, homomorphic encryption, federated learning, hardware enclaves, and more. As companies begin to distinguish themselves by taking a stronger stance on privacy, analytics and machine learning need to keep pace.

In addition, as machine learning models become more widely deployed and used, interest in transparency, interpretability, and explainability continues to grow among data professionals. How does a model "work"? What is it basing its output on? Organizations are beginning to grapple with the trade-off between accuracy and interpretability, and, fortunately, some tools and best practices are beginning to emerge.

As Figure 10 illustrates, organizations that have extensive experience in deploying machine learning models have much more robust model-building checklists than their peers:

- One in two (54%) respondents who belong to sophisticated companies check for fairness and bias, compared to one in three (32%) companies that are just starting out. The good news is that **tutorials and training materials are becoming available**, and the number of companies capable of addressing fairness and bias should increase.
- Three out of four (74%) respondents who belong to sophisticated companies already check for transparency.

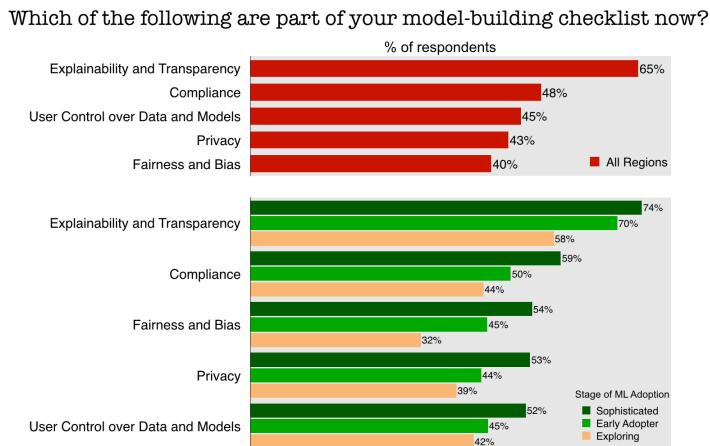


Figure 10

Closing Thoughts

Looking at overall themes from the survey responses, it's interesting to see how the more sophisticated companies differ from those that are still exploring. Here are some of the things that sophisticated companies do:

- Use more specialized roles such as data scientist and data engineer in lieu of older roles such as business analyst.
- Use their internal data science teams to build machine learning models.

- Employ much more robust model-building checklists than their peers, already including checks for machine learning model transparency and data privacy.
- Have their data science leads set team priorities and determine key metrics for project success instead of product managers.

These points indicate some of the key learnings that derive from deploying machine learning in production, and also where other companies should focus as they begin their journey of machine learning adoption. Meanwhile, the jury's out about whether Agile methodology is entirely appropriate for machine learning.

About the Authors

Ben Lorica is the chief data scientist of O'Reilly Media, and program director of Strata Data Conference and the Artificial Intelligence Conference. He has applied business intelligence, data mining, machine learning, and statistical analysis in a variety of settings, including direct marketing, consumer and market research, targeted advertising, text mining, and financial engineering.

Paco Nathan is known as a “player/coach,” with core expertise in data science, natural language processing, machine learning, and cloud computing. He has over 35 years of tech industry experience, ranging from Bell Labs to early-stage start-ups. He is the cochair of JupyterCon; the host of Executive Briefings at The AI Conf and Strata Data; evangelist for Computable; and advisor for Amplify Partners, Deep Learning Analytics, and Recognai. His recent roles include Director, Learning Group at O'Reilly Media and Director, Community Evangelism at Databricks and Apache Spark. He was cited in 2015 as one of the “Top 30 People in Big Data and Analytics” by Innovation Enterprise.