O'REILLY®

Introducing MLOps

How to Scale Machine Learning in the Enterprise

Early Release
RAW & UNEDITED
Sponsored by
data iku

Clément Stenac,
Léo Dreyfus-Schmidt,
Kenji Lefèvre, Nicolas Omont
& Mark Treveil

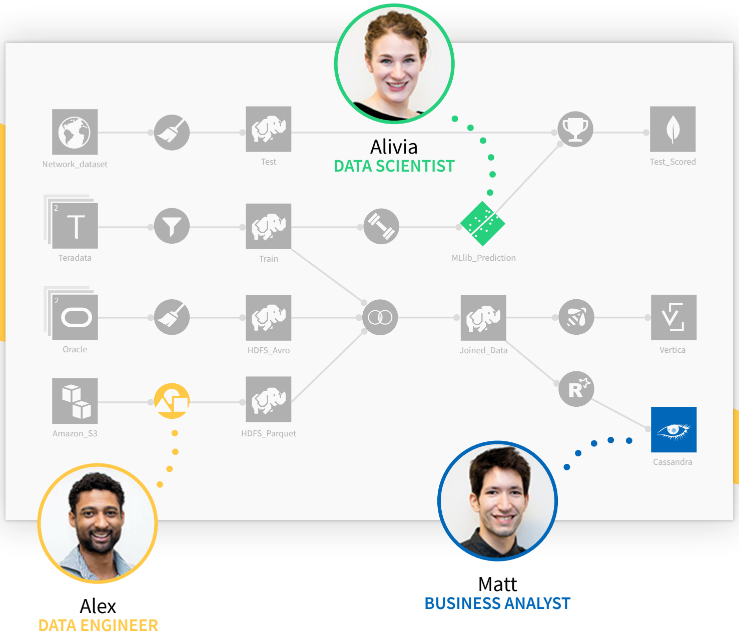
MASTERING MLOps
WITH DATAIKU

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 with the complete data sheet.

→ GET THE DATAIKU DATA SHEET



Introducing MLOps

How to Scale Machine Learning in the Enterprise

With Early Release ebooks, you get books in their earliest form—the author's raw and unedited content as they write—so you can take advantage of these technologies long before the official release of these titles.

Clément Stenac, Léo Dreyfus-Schmidt, Kenji Lefèvre, Nicolas Omont, and Mark Treveil

Beijing • Boston • Farnham • Sebastopol • Tokyo
O'REILLY®

Mastering ModelOps
by Clément Stenac, Léo Dreyfus-Schmidt, Kenji Lefèvre, Nicolas Omont, and Mark Treveil

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*.

Editors: Angela Rufino and Rebecca Novack
Production Editor: Katherine Tozer
Interior Designer: David Futato
Cover Designer: Karen Montgomery
Illustrator: Rebecca Demarest

February 2021: First Edition

Revision History for the Early Release
2020-05-19: First Release

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

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.

978-1-098-11640-8

Table of Contents

CHAPTER 1
Why Now and Challenges

A note for Early Release readers

With Early Release ebooks, you get books in their earliest form—the author's raw and unedited content as they write—so you can take advantage of these technologies long before the official release of these titles.

This will be the first chapter of the final book. If you have comments about how we might improve the content and/or examples in this book, or if you notice missing material within this chapter, please reach out to the author at *mlops@dataiku.com*.

Machine learning operations (MLOps) is quickly becoming a critical component of successful data science project deployment in the enterprise (Figure 1-1). Yet it's a relatively new concept, so why has it seemingly skyrocketed into the data science lexicon overnight? This introductory chapter will delve into what MLOps is at a high level, its challenges, why it's become essential to a successful data science strategy in the enterprise, and — critically — why it is coming to the forefront now.

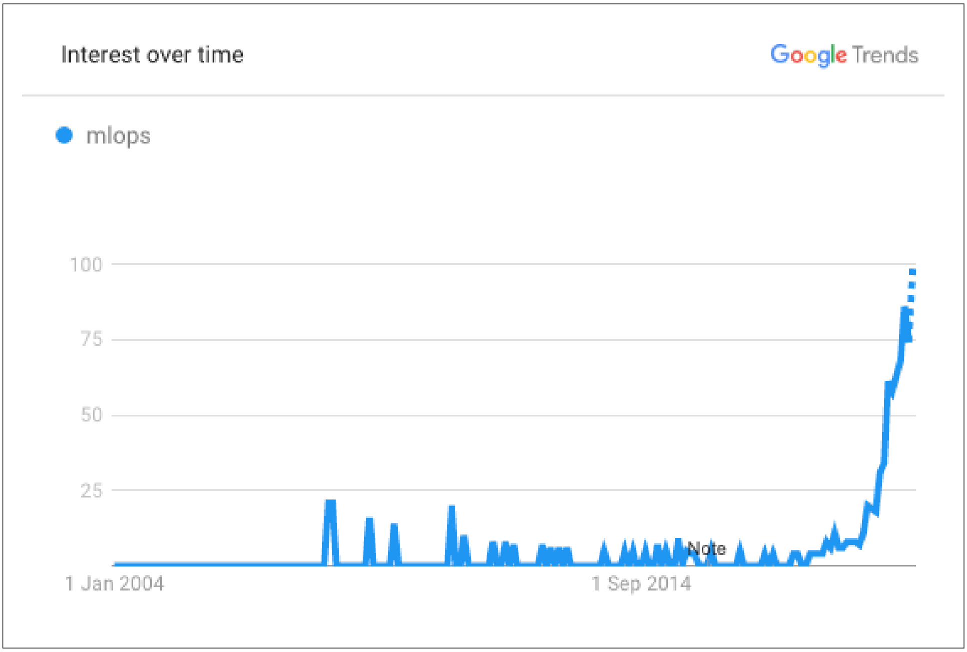


Figure 1-1. The exponential growth of MLOps. This represents only the growth of MLOps, not the parallel growth of the term ModelOps (subtle differences explained in the sidebar MLOps vs. ModelOps vs. AIOps).

MLOps vs. ModelOps vs. 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 and published, 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 lifecycle and will thus use 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 but 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.

Defining MLOps and Its Challenges

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

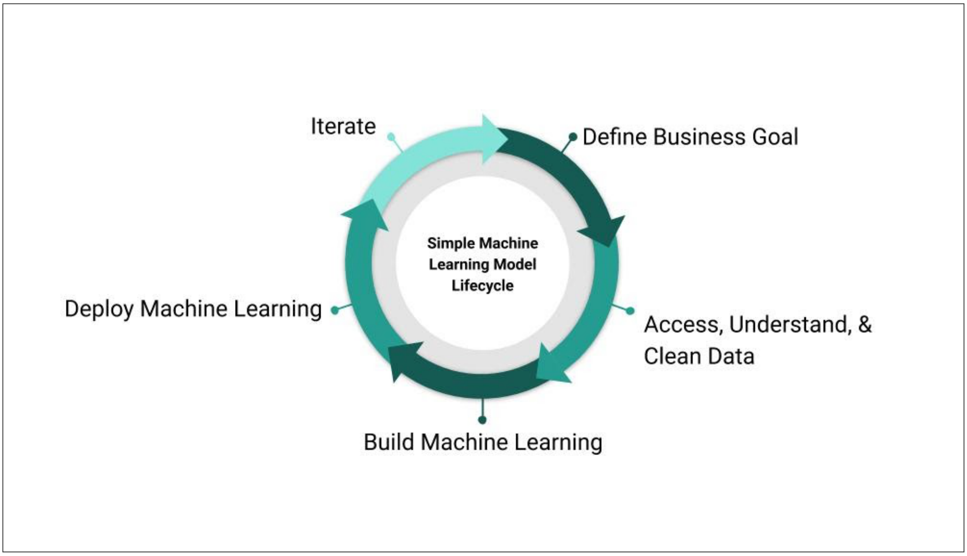


Figure 1-2. A simple representation of the machine learning model lifecycle, which often underplays the need for MLOps; compare to Figure 3, which is a more realistic representation of how the machine learning model lifecycle plays out in today's organizations, which are complex in terms of needs as well as tooling.

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, models become more critical, and in parallel, managing model risks becomes more important at the top level.

The reality of the machine learning lifecycle in an enterprise setting is much more complex (Figure 1-3). There are three key reasons that managing machine learning lifecycles at scale are 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 lifecycle 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 operationalization 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 when suddenly, data scientists have to manage models they did not create.

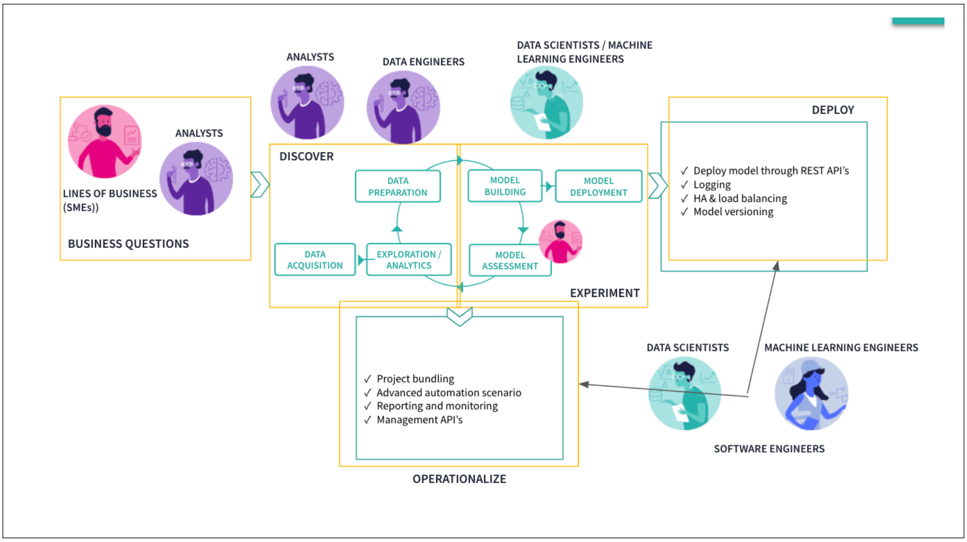


Figure 1-3. The realistic picture of a machine learning model lifecycle inside an average organization today, which involves many different people with completely different skill sets and who are often using entirely different tools.

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.

The end-to-end service lifecycle (build-test-release).

- Prioritizing continuous delivery as well as high quality.

Yet there is one critical difference between MLOps and DevOps that makes the latter not immediately transferable to data science teams: deploying software code in production is fundamentally different than deploying machine learning models into production. While software code is relatively static ("relatively" because many modern SaaS companies do have DevOps teams that can iterate quite quickly and deploy in production multiple times per day), data is always changing, which means machine learning models are constantly learning and adapting — or not, as the case may be — to new inputs. The complexity of this environment, including the fact that machine learning models are made up of both code as well as data, is what makes MLOps a new and unique discipline.

What About DataOps?

To add to the complexity of MLOps vs. DevOps, there is also DataOps, a term introduced in 2014 by IBM. DataOps seeks to provide business-ready data that is quickly available for use, with a large focus on data quality and metadata management. For example, if there's a sudden change in data that a model relies on, a DataOps system would alert the business team to deal more carefully with the latest insights, and the data team would be notified to investigate the change or revert a library upgrade and rebuild the related partition.

The rise of MLOps, therefore, intersects DataOps at some level, though MLOps goes a step further and brings even more robustness through additional key features (discussed in more detail in chapter 3, MLOps: Key Features).

As was the case with DevOps and later DataOps, until recently, teams have been able to get by without defined and centralized MLOps processes mostly because — at an enterprise level — they weren't deploying machine learning models into production at a large enough scale. Now, the tables are turning and teams are increasingly looking for ways to formalize a multi-stage, multi-discipline, multi-phase process with a heterogeneous environment and a framework for MLOps best practices, which is no small task. Part II of this book (***MLOps: How***) will provide this guidance.

MLOps to Mitigate Risk

MLOps is important to any team that has even one model in production, as depending on the model, continuous performance monitoring and adjusting is essential.

Think about a travel site whose pricing model would require top-notch MLOps to ensure that the model is continuously delivering business results.

However, MLOps really tips the scales as critical for risk mitigation when a centralized team (with unique reporting of its activities, meaning that there can be multiple such teams at any given enterprise) has more than a handful of operational models. At this point, it becomes difficult to have a global view of the states of these models without some standardization.

Pushing machine learning models into production without MLOps infrastructure is risky for many reasons, but first and foremost because fully assessing the performance of a machine learning model can often only be done in the production environment. Why? Because prediction models are only as good as the data they are trained on, which means the training data must be a good reflection of the data encountered in the production environment. If the production environment changes, then the model performance is likely to decrease rapidly.

Another major risk factor is that machine learning model performance is often very sensitive to the production environment it is running in, including the versions of software and operating systems they use. They tend not to be buggy in the classic software sense, because most weren't written by hand but rather were machine-generated. Instead, the problem is they are often built on a pile of open-source software (e.g., libraries — like Scikit-Learn — to Python to Linux), and having versions of this software in production that match those that the model was verified on is critically important.

Ultimately, pushing models into production is not the final step of the machine learning lifecycle and is, in fact, far from it. It's often just the beginning of monitoring its performance and ensuring that it behaves as expected. As more data scientists start pushing more machine learning models into production, MLOps becomes critical in mitigating the potential risks, which (depending on the model) can be devastating for the business.

MLOps for Responsible AI

A responsible use of machine learning (more commonly referred to as Responsible AI) covers three main dimensions:

Accountability: Ensuring that machine learning models are designed and behave in ways aligned with their purpose. Note that for publicly-traded companies in the United States, this is related to the notion of full disclosure.

Sustainability: Establishing the continued reliability of machine learning models in their operation as well as execution.

- Governability: Centrally controlling, managing, and auditing machine learning capabilities in the enterprise.

These principles may seem obvious, but it's important to consider that machine learning models lack the transparency of traditional imperative code. In other words, it is much harder to understand what features are used to determine a prediction, which in turn can make it much harder to demonstrate that models comply with the necessary regulatory or internal governance requirements.

The reality is that introducing automation vis-à-vis machine learning models shifts the fundamental onus of accountability from the bottom of the hierarchy to the top. That is, decisions that were perhaps previously made by individual contributors who operated within a margin of guidelines (for example, what the price of a given product should be or whether or not a person should be accepted for a loan) are now being made by a model. The person responsible for the automated decisions of said model is likely a data team manager or even executive, and that brings the concept of Responsible AI even more to the forefront.

Given the previously discussed risks as well as these particular challenges and principals, it's easy to see the interplay between MLOps and Responsible AI — teams must have good MLOps principles to practice Responsible AI, and Responsible AI necessitates MLOps strategies.

MLOps for Scale

MLOps isn't just important because it helps mitigate the risk of machine learning models in production, but it is also an essential component to scaling machine learning efforts (and in turn benefiting from the corresponding economies of scale). Going from one or a handful of models in production to tens, hundreds, or thousands that have a positive business impact will require MLOps discipline.

Good MLOps practices will help teams at a minimum:

Keep track of versioning, especially with experiments in the design phase.

Understand if retrained models are better than the previous versions (and promoting models to production that are performing better).

- Ensure (at defined periods — daily, monthly, etc.) that model performance is not degrading in production.

Closing Thoughts

Key features will be discussed at length in Chapter 3, but the point here is that these are not optional practices — they are essential tasks for not only efficiently scaling data science and machine learning at the enterprise level, but also doing it in a way that doesn't put the business at risk. Teams that attempt to deploy data science without proper MLOps practices in place will face issues with model quality, continuity, or worse — they will introduce models that have a real, negative impact on the business (e.g., a model that makes biased predictions that reflect poorly on the company).

MLOps is also, at a higher level, a critical part of transparent strategies for machine learning. Upper management and the C-suite should be able to understand as well as data scientists what machine learning models are deployed in production and what effect they're having on the business. Beyond that, they should arguably be able to drill down to understand the whole data pipeline behind those machine learning models. MLOps, as described in this book, can provide this level of transparency and accountability.

CHAPTER 2

People of Model Ops

A note for Early Release readers

With Early Release ebooks, you get books in their earliest form—the author's raw and unedited content as they write—so you can take advantage of these technologies long before the official release of these titles.

This will be the second chapter of the final book. If you have comments about how we might improve the content and/or examples in this book, or if you notice missing material within this chapter, please reach out to the author at *mlops@dataiku.com*.

Even though machine learning models are primarily built by data scientists, it's a misnomer that only data scientists can benefit from robust MLOps processes and systems. In fact, MLOps is an essential piece of Enterprise AI strategy and affects everyone working — or benefiting from — the machine learning model lifecycle.

This chapter will cover the roles each of these people play in the machine learning lifecycle, who they should ideally be connected and working together with under a top-notch MLOps program to achieve the best possible results from machine learning efforts, and what MLOps requirements they may have. Before diving into the details.

Table 2-1. provides an overview:

Role	Role in Machine Learning Model Lifecycle	MLOps Requirements
Subject Matter Experts	• Provide business questions, goals, or KPIs around which machine learning models should be framed. • Continually evaluate and ensure that model performance aligns with or resolves the initial need.	• Easy way to understand deployed model performance in business terms. • Mechanism or feedback loop for flagging model results that don't align with business expectations.
Data Scientists	• Build models that address the business question or needs brought by subject matter experts. • Deliver operationalizable models so that they can be properly used in the production environment and with production data. • Assess model quality (of both original and tests) in tandem with subject matter experts to ensure they answer initial business questions or needs.	• Automated model packaging and delivery for quick and easy (yet safe) deployment to production. • Ability to develop tests to determine the quality of deployed models and to make continual improvements. • Visibility into the performance of all deployed models (including side-by-side for tests) from one central location. • Ability to investigate data pipelines of each model to make quick assessments and adjustments regardless of who originally built the model.
Data Engineers	• Optimize the retrieval and use of data to power machine learning models.	• Visibility into performance of all deployed models. • Ability to see the full details of individual data pipelines to address underlying data plumbing issues.
Software Engineers	• Integrate machine learning models in the company's applications and systems. • Ensure that machine learning models work seamlessly with other non-machine learning-based applications.	• Versioning and automatic tests. • The ability to work in parallel on the same application.
DevOps	• Conduct and build operational systems and test for security, performance, availability. • Continuous Integration/Continuous Delivery (CI/CD) pipeline management.	• Seamless integration of MLOps into the larger DevOps strategy of the enterprise. • Seamless deployment pipeline.
Model Risk Managers / Auditors	• Minimize overall risk to the company as a result of machine learning models in production. • Ensure compliance with internal and external requirements before pushing machine learning models to production.	• Robust — likely automated — reporting tools on all models (currently or ever in production), including data lineage.

Role	Role in Machine Learning Model Lifecycle	MLOps Requirements
Machine Learning Architects	• Ensure a scalable and flexible environment for machine learning model pipelines, from design to development and monitoring. • Introduce new technologies when appropriate that improve machine learning model performance in production.	• High-level overview of models and their resources consumed. • Ability to drill down into data pipelines to assess and adjust infrastructure needs.

Subject Matter Experts

The first profile to consider as part of MLOps efforts are the subject matter experts; after all, the machine learning model lifecycle starts and ends with them. While the data-oriented profiles (data scientist, engineer, architect, etc.) have expertise across many areas, one where they tend to lack is a deep understanding of the business and the problems or questions at hand that need to be addressed using machine learning.

Role in the Machine Learning Model Lifecycle

Subject matter experts usually come to the table — or at least, they *should* come to the table — with clearly defined goals, business questions, and/or key performance indicators (KPIs) that they want to achieve or address. In some cases, they might be extremely well defined (e.g., "In order to hit our numbers for the quarter, we need to reduce customer churn by 10%" or "We're losing $N per quarter due to unscheduled maintenance, how can we better predict downtime?") In other cases, less so (e.g., "Our service staff needs to better understand our customers to upsell them" or "How can we get people to buy more widgets?").

In organizations with healthy processes, starting the machine learning model lifecycle with a more well-defined business question isn't necessarily always an imperative, or even an ideal scenario. Working with a less-defined business goal can be a good opportunity for subject matter experts to work directly with data scientists upfront to better frame the problem and brainstorm possible solutions before even beginning any data exploration or model experimentation.

Without this critical starting point from subject matter experts, other data professionals (particularly data scientists) risk starting the machine learning lifecycle process trying to solve problems or provide solutions that don't serve the larger business. Ultimately, this is detrimental not only to the subject matter experts who need to partner with data scientists and other data experts to build solutions, but to data scientists themselves who might struggle to provide larger value (data teams might in turn see trust in — or budgets for — data initiatives fall). Business decision modeling methodologies can be applied to formalize the business problems to be solved and frame the role of machine learning in the solution.

Business Decision Modeling

Decision modeling creates a business blueprint of the decision-making process, allowing subject matter experts to directly structure and describe their needs. Decision models can be helpful because they put machine learning in context for subject matter experts, allowing them to integrate with business rules as well to fully understand decision contexts and the potential impact of model changes.

MLOps strategies that include a component of business decision modeling for subject matter experts can be an effective tool for ensuring real-world machine learning model results are properly contextualized for those that don't have deep knowledge of how the underlying models themselves work. Further reading on building decision requirement models.

Subject matter experts have a role to play not only at the beginning of the machine learning model lifecycle, but the end (post-production) as well. Oftentimes, to understand if a machine learning model is performing well or as expected, data scientists need subject matter experts to close the feedback loop — traditional metrics (accuracy, precision, recall, etc.) are not enough.

For example, data scientists could build a simple churn prediction model that has very high accuracy in a production environment; however, marketing does not manage to prevent anyone from churning. From a business perspective, that means the model didn't work, and that's important information that needs to make its way back to those building the machine learning model so that they can find another possible solution — e.g., introducing uplift modeling that helps marketing better target potential churners who might be receptive to marketing messaging.

Role In and Needs From MLOps

Given their role in the machine learning model lifecycle, it's critical when building MLOps processes for subject matter experts to have an easy way to understand deployed model performance in business terms. That is, not just model accuracy, precision, and recall, but its results or impact on the business process identified upfront. In addition, when there are unexpected shifts in performance, subject matter experts need a scalable way through MLOps processes for flagging model results that don't align with business expectations.

On top of these explicit feedback mechanisms, more generally, MLOps should be built in a way that increases transparency for subject matter experts. That is, they should be able to use MLOps processes as a jumping-off point for exploring the data pipelines behind the models, understanding what data is being used, how it's being transformed and enhanced, and what kind of machine learning techniques are being applied.

For subject matter experts who are also concerned with compliance of machine learning models with regulations (either internal or external), MLOps serves as an additional way to bring transparency and understanding to these processes. This includes being able to dig into individual decisions made by a model to understand why the model came to that decision — this should be complementary to statistical and aggregated feedback.

Ultimately, MLOps is most relevant for subject matter experts as a feedback mechanism and a platform for communication with data scientists about the models they are building. However, there are other MLOps needs as well — specifically around transparency, which ties up into Responsible AI — that are relevant for subject matter experts and make them an important part of the MLOps picture.

Data Scientists

The needs of data scientists are the most critical ones to consider when building an MLOps strategy. To be sure, they have a lot to gain; data scientists at most organizations today are often dealing with siloed data, processes, and tools, making it difficult to effectively scale their efforts. MLOps is well positioned to change this.

Role in the Machine Learning Model Lifecycle

Though most see data scientists' role in the machine learning model lifecycle as strictly the model building portion, it is actually — or at least, it should be — much wider. From the very beginning, data scientists need to be involved with subject matter experts, understanding and helping to frame business problems in such a way that they can build a viable machine learning solution.

The reality is that this very first, critical step in the machine learning model lifecycle is often the hardest. It's challenging particularly for data scientists first and foremost because it's not where their training lies; that is, both formal and informal data science programs in universities and online heavily emphasize technical skill and not necessarily skills for communicating effectively with subject matter experts from the business side of the house who usually are not intimately familiar with machine learning techniques. Once again, business decision modeling techniques can help here.

It's also a challenge because it can take time — for data scientists who want to dive in and get their hands dirty, spending weeks framing and outlining the problem before getting started on solving it can be torture. To top it all off, data scientists are often siloed (physically, culturally, or both) from the core of the business and from subject matter experts, so they simply don't have the organizational infrastructure that facilitates easy collaboration between these profiles. Robust MLOps systems can help address some of these challenges.

After overcoming the first hurdle, depending on the organization, the project might get handed off to either data engineers or analysts to do some of the initial data gathering, preparation, and exploration. In some cases, data scientists themselves manage these parts of the machine learning model lifecycle. But in any case, data scientists step back in when it comes time to build, test, robustify, and then deploy the model.

Following deployment, data scientists' roles include constantly assessing model quality to ensure the way it's working in production answers initial business questions or needs. The underlying question in many organizations is often whether data scientists monitor only the models they have had a hand in building, or if there is one person who handles all monitoring. In the former scenario, what happens when there is staff turnover? In the latter scenario, building good MLOps practices is critical, as the person monitoring also needs to quickly be able to jump in and take action should the model drift and start negatively affecting the business. If they weren't the ones who built it, how can MLOps make this process seamless?

Operationalization and MLOps

Throughout 2018 and the beginning of 2019, operationalization was the key buzzword when it came to machine learning model lifecycles and AI in the enterprise. Put simply, operationalization of data science is the process of pushing models to production and measuring their performance against business goals. So how does operationalization fit into the MLOps story? MLOps takes operationalization one step further, encompassing not just the push to production but the maintenance of those models — and the entire data pipeline — in production.

Though they are distinct, MLOps might be considered the new operationalization. That is, where many of the major hurdles for businesses to operationalize have disappeared, MLOps is the next frontier and presents the next big challenge for machine learning efforts in the enterprise.

Role In and Needs From MLOps

All of the questions in the previous section lead directly here: data scientists' needs when it comes to MLOps. Starting from the end of the process and working backwards, MLOps must provide data scientists with visibility into the performance of all deployed models as well as any models being A/B tested. But taking that one step further, it's not just about monitoring — it's also about action. Top-notch MLOps should also allow data scientists the flexibility to select winning models from tests and easily deploy them.

Transparency is an overarching theme in MLOps, so it's no surprise that it's also a key need for data scientists. The ability to drill down into data pipelines and to make quick assessments and adjustments (regardless of who originally built the model) is critical. Automated model packaging and delivery for quick and easy (yet safe) deployment to production is another important point for transparency, and it's a crucial component of MLOps, especially to bring data scientists together to a place of trust with software engineers and DevOps teams.

In addition to transparency, perhaps another theme for mastering MLOps — especially when it comes to meeting the needs of data scientists — is pure efficiency. In an enterprise setting, agility and speed matter. It's true for DevOps, and the story for MLOps is no different. Of course, data scientists can deploy, test, and monitor models in an ad-hoc fashion. But they will lose enormous amounts of time re-inventing the wheel with every single machine learning model, and that will never add up to scalable machine learning processes for the organization.

Data Engineers

Data pipelines are at the core of the machine learning model lifecycle, and data engineers are, in turn, at the core of data pipelines. Because data pipelines can be abstract and complex, data engineers have a lot of efficiencies to gain from MLOps.

Role in the Machine Learning Model Lifecycle

In large organizations, managing the flow of data itself outside of the application of machine learning models is a full-time job. Depending on the technical stack and organizational structure of the enterprise, data engineers might, therefore, be more focused on databases themselves than on pipelines (especially if the company is leveraging data science and machine learning platforms that facilitate the visual building of pipelines by other data practitioners, like business analysts).

Ultimately, despite these slight variations in the role by an organization, the role of data engineers in the lifecycle is to optimize the retrieval and use of data to eventually power machine learning models. Generally, this means working closely with business teams, particularly subject matter experts, to identify the right data for the project at hand and possibly also prepare it for use. On the other end, they work closely with data scientists as well to resolve any data plumbing issues that might cause a model to behave undesirably in production.

Role In and Needs From MLOps

Given data engineers' central role in the machine learning model lifecycle, underpinning both the building and monitoring portions, MLOps can bring significant efficiency gains. Data engineers will namely require not only visibility into the performance of all models deployed in production, but the ability to take it one step further and directly drill down into individual data pipelines to address any underlying issues.

Ideally, for maximum efficiency for the data engineer profile (and for others as well - including data scientists), MLOps must not consist of simple monitoring, but be a bridge to underlying systems for investigating and tweaking machine learning models.

Software Engineers

It would be easy to exclude classical software engineers from MLOps consideration, but it is crucial from a wider organizational perspective to consider their needs to build a cohesive enterprise-wide strategy for machine learning.

Role in the Machine Learning Model Lifecycle

Software engineers usually aren't building machine learning models, but on the other hand, most organizations are not *only* producing machine learning models, but classic software and applications as well. It's important that software engineers and data scientists work together to ensure the functioning of the larger system. After all, machine learning models aren't just stand-alone experiments; the machine learning code, training, testing, and deployment has to fit into the CI/CD pipelines that the rest of the software is using.

For example, consider a retail company that has built a machine learning-based recommendation engine for their website. The machine learning model was built by the data scientist, but to integrate it into the larger functioning of the site, software engineers will necessarily need to be involved. Similarly, software engineers are responsible for the maintenance of the website as a whole, and a large part of that includes the functioning of the machine learning models in production.

Role In and Needs From MLOps

Given this interplay, software engineers need MLOps to provide them with model performance details as a larger picture of software application performance for the enterprise. MLOps is a way for data scientists and software engineers to speak the same language and have the same baseline understanding of how different models deployed across the silos of the enterprise are working together in production.

Other important features for software engineers include versioning, in order to be sure of what they are currently dealing with; automatic tests, in order to be as sure as possible that what they are currently dealing with is working; and the ability to work in parallel on the same application (thanks to a system that allows branches and merges like Git).

DevOps

MLOps was born out of DevOps principles, but that doesn't mean they can be run in parallel as completely separate and siloed systems.

Role in the Machine Learning Model Lifecycle

DevOps teams have two primary roles in the machine learning model lifecycle: first, they are the people conducting and building operational systems as well as tests to ensure security, performance, and availability of machine learning models. Secondly, they are responsible for CI/CD pipeline management. Both of these roles require tight collaboration with data scientists as well as data engineers and data architects. Tight collaboration is, of course, easier said than done, but that is where MLOps can add value.

Role In and Needs From MLOps

For DevOps teams, MLOps needs to be integrated into the larger DevOps strategy of the enterprise, bridging the gap between traditional CI/CD and modern machine learning. That means systems that are fundamentally complementary and that allow DevOps teams to automate tests for machine learning just as they can automate tests for traditional software.

Model Risk Manager/Auditor

In certain industries (particularly the financial services sector), the model risk management (MRM) function is crucial for regulatory compliance. But it's not only highly-regulated industries that should be concerned or that should have a similar function; MRM can protect companies in any industry from catastrophic loss introduced by poorly performing machine learning models. What's more, audits play a role in many industries and can be labor-intensive, which is where MLOps comes into the picture.

Role in the Machine Learning Model Lifecycle

When it comes to the machine learning model lifecycle, model risk managers play the critical role of analyzing not just model outcomes, but the initial goal and business questions machine learning models seek to resolve to minimize overall risk to the company. They should be involved along with subject matter experts at the very beginning of the lifecycle to ensure that an automated, machine learning-based approach in and of itself doesn't present risk.

And, of course, they have a role to play in monitoring — their more traditional place in the model lifecycle — to ensure that risk stays at bay once models are in production. In between conception and monitoring, MRM also is a factor post-model development and pre-production, ensuring initial compliance with internal and external requirements.

Role In and Needs From MLOps

MRM professionals and teams have a lot to gain from MLOps, as their work is often painstakingly manual, and as teams often use different tools, standardization can offer a huge leg-up in the speed at which auditing and risk management can occur.

When it comes to specific MLOps needs, robust reporting tools on all models — whether they are currently in production or have been in production in the past — is the primary one. This reporting should include not just performance details, but the ability to see data lineage. Automated reporting adds an extra layer of efficiency for MRM and audit teams in MLOps systems and processes.

Machine Learning Architect

Traditional data architects are responsible for understanding the overall enterprise architecture and ensuring that it meets the requirements for data needs from across the business. They generally play a role in defining how data will be stored and consumed.

Today, demands on architects are much greater, and they often have to be knowledgeable not only on the ins and outs of data storage and consumption, but on how machine learning models work in tandem. This adds a lot of complexity to the role and increases their responsibility in the MLOps lifecycle, and it's why in this section, we have called them machine learning architects instead of the more traditional data architect title.

Role in the Machine Learning Model Lifecycle

Machine learning architects play a critical role in the machine learning model lifecycle, ensuring a scalable and flexible environment for model pipelines. In addition, data teams need their expertise to introduce new technologies (when appropriate) that improve machine learning model performance in production. It is for this reason that the data architect title isn't enough; they need to have an intimate understanding of machine learning — not just enterprise architecture — to play this key role in the machine learning model lifecycle.

This role requires collaboration across the enterprise, from data scientists and engineers to DevOps and software engineers. Without a complete understanding of the needs of each of these people and teams, machine learning architects cannot properly allocate resources to ensure optimal performance of machine learning models in production.

Role In and Needs From MLOps

When it comes to MLOps, machine learning engineers' role is about having a centralized view of resource allocation. As they have a strategic, tactical role, they need an overview of the situation to identify bottlenecks and use that information to find long-term improvements. Their role is one of pinpointing possible new technology or infrastructure for investment, not necessarily operational quick fixes that don't address the heart of the scalability of the system.

Closing Thoughts

MLOps isn't just for data scientists; a diverse group of experts across the organization have a role to play not only in the machine learning model lifecycle, but the MLOps strategy as well. In fact, each person — from the subject matter expert on the business side to the most technical machine learning architect — plays a critical part in the maintenance of machine learning models in production. This is ultimately important not only to ensure the best possible results from machine learning models (good results generally lead to more trust in machine learning-based systems as well as increased budget to build more), but perhaps more pointedly, to protect the business from the risks outlined in Chapter 1.