DOMINO

The Complete Guide to Enterprise MLOps Principles

Learn the keys to becoming a model-driven business

WHITEPAPER

Table of Contents

Executive Summary

Today's businesses are investing heavily in data science – spending on software, hardware and services is projected to break the $500 billion mark by 2024, according to IDC. Data science models with machine learning (ML) and artificial intelligence (AI) techniques have proven their worth at forging new revenue streams and upending entire industries. For a model-driven business, new revenue can range from hundreds of millions to billions of dollars.

The leading question is: how does an aspiring enterprise scale its data science program to aim for those rewards? Frankly, scaling data science is not easy, nor can it happen overnight. At successful companies, leaders have built a well-oiled analytical flywheel to create a steady flow of models that can tap this new gold rush. This is an ideal scenario, but the cost of getting it wrong is equally large. Operational expenses can quickly spiral out of control, and significant monetary and brand reputation risks are real consequences of creating bad models or using them in a wrong way.

Domino Data Lab has collaborated with many companies across all industries that have built a revenue-generating data science machine at scale. One thing they all have in common is a holistic approach that looks for efficiencies of scale across all stages of the data science lifecycle. We call this approach Enterprise MLOps. Enterprise MLOps is a set of technologies and best practices that streamline the management, development, deployment, and monitoring of data science models at scale across a diverse enterprise. This whitepaper describes the challenges of scaling data science and what to expect as your organization begins or extends this journey. It shows how incorrectly defining MLOps leads to roadblocks. It provides an operational blueprint for creating your own revenue-generating data science flywheel.

Enterprise MLOps is a set of technologies and best practices that streamline the management, development, deployment, and monitoring of data science models at scale across a diverse enterprise.

Why Scaling Data Science is So Hard

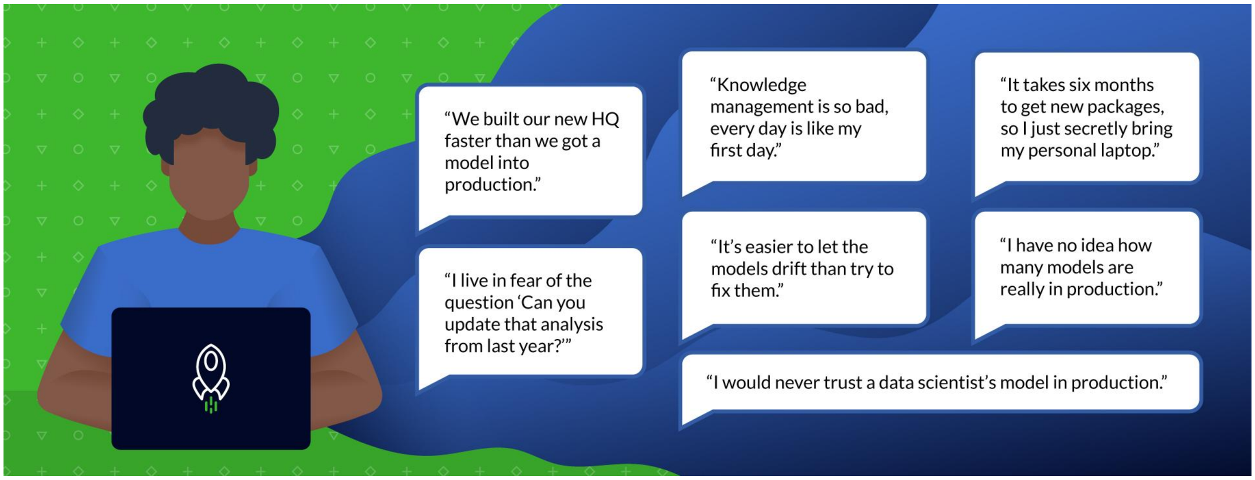
Data science is on an all-time tear. On an organizational level, 62 percent of firms have invested over $50 million in big data and AI, with 17 percent investing more than $500 million, according to a recent survey from New Vantage Partners (NVP). Expectations are just as high as investment levels, with a survey from Data IQ revealing that a quarter of companies expect data science to increase revenue by 11 percent or more. This is a major leap for giant enterprises that are already generating huge cash flows.

Yet while money is flowing strong, results have not been so rosy. Consider the quotes shown below, which Domino has heard from data science stakeholders across large enterprises while in the early phase of scaling data science. Their negativity is mirrored in five conclusions from a recent survey by Wakefield Research and Domino Data Lab about why initiatives are falling short of expectations:

- Short-term investment thwarts growth expectations
- The role of data science is unclear
- More revenue requires better models
- Unimproved models bring higher risk
- Organizations must clear obstacles to achieve goals

Such obstacles entail both technical and cultural components to scaling model velocity. The collision of expectations for AI and significant obstacles to achieving data science at scale is equally ominous: 75 percent of executives responding to a recent survey by Accenture believe that their companies will most likely go out of business if they can't scale data science successfully within the next five years.

How can successful model-driven businesses surmount these obstacles for positive operational and financial rewards? Businesses can take the technical principles of MLOps and apply them to the entire data science lifecycle, not just the last mile. Additionally, they should consider how the same efficiencies can apply to processes and people. Enterprise MLOps captures this set of technologies and best practices.



Redefining MLOps for Enterprise Scalability

Most data scientists are familiar with a technical concept called MLOps, or machine learning operations. MLOps is relatively new in the AI world, dating to 2015 in a paper called, "Hidden Technical Debt in Machine Learning Systems," written by two teams at Google. The authors argued that it is dangerous to think that rapidly building complex prediction systems results in "quick wins as coming for free." The paper offered ideas for avoiding massive maintenance costs in real-world systems. Hence the genesis of MLOps.

Some data science stakeholders see MLOps as "DevOps in the context of AI." A standard definition of MLOps is somewhat fluid and evolving in the data science community. It is instructive to take a closer look at shortcomings in how some perceive what MLOps can do and effectively redefine MLOps for enterprise scalability.

A common definition of MLOps pigeonholes it as solving only a backend problem for data science. Frequently, articles in the press and from software vendors say the purpose of MLOps is getting models into production and maintaining them in a streamlined, semi-automatic manner using microservices and CI/CD principles. While partly accurate, this definition is also somewhat incoherent and even myopic in applicability.

In Scaling Data Science, R&D is Harder than Production

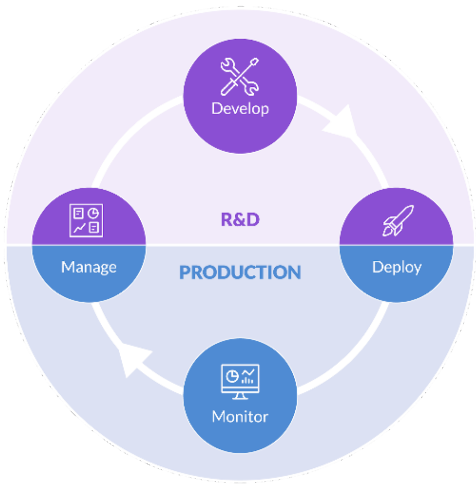
One reason for the prevalence of narrowly defining MLOps is limited product capabilities. It is far easier for platforms to solve problems of scale in the data science lifecycle's back-end production than it is to solve similar problems in the R&D front end.

In the back end, the model is already built and packaged as a file. The model is supported by a data pipeline and is often wrapped in a container. While it is not at all trivial to build tools to scale and standardize the deployment of models, the variables are finite and more easily controlled once understood.

Front-end R&D has vastly more variables to manage. Many are not easy to understand or control. The process of developing a model may require significant iterations before the true scope of variables becomes apparent. Some variables also involve human stakeholders, which brings unpredictability with more "cooks in the kitchen." Achieving success in the front-end R&D relies on collaborative work of teams of data scientists, other experts, and the business.

Failure to scale the front-end R&D half of the data science lifecycle will stymie the ROI that leaders hope to realize for the business. Many of these non-technical barriers can be overcome with best practices and tools that are built on MLOps principles by using an Enterprise MLOps platform.

To understand why, consider the scope of the data science lifecycle that applies to every organization attempting to deploy models at scale. The lifecycle has two parts: Part I is the "front end": research and development of the model idea, experimentation, iteration, and validation up to the point of deployment. Part II is the "back end": deployment and subsequent activities of monitoring model performance, addressing potential model drift, and otherwise managing proper use of the model and retraining for continuous improvement.



Defining MLOps strictly from the back-end perspective is shortsighted. For how can data scientists properly retrain a model without solving for the intricacies of the complex R&D processes at the heart of data science? It is impossible to have one without the other.

> "In regulated industries like pharmaceuticals, biopharmaceuticals, and finance, MLOps is actually from the inception of the model all the way through into production. We need to view MLOps that way in the future. I don't think most industries think about it that way yet."
>
> ***-John K. Thompson, Global Head, Advanced Analytics & AI, CSL Behring***

The same is true for benefits. Obtaining speed, security, automation, tracking, versioning, continuous integration, and reproducibility are often cited as byproducts of the back-end definition of MLOps. For a large enterprise to scale data science, achieving those identical benefits for the front-end processes is equally mandatory. Therefore, the functional definition of MLOps must apply to the entire data science lifecycle.

As a side note, the narrow definition of MLOps was more of an accident by well-meaning software developers and cloud computing companies. They looked at the model deployment and maintenance dilemma and saw a nail that fit their hammer. Most of their products addressed the back end of the lifecycle, as were the ideas and tools of standard software development … a "DevOps for AI models."

> An Enterprise MLOps platform also removes the non-technical barriers by enabling best practices for collaboration, knowledge sharing, and project management.

Understanding what a large enterprise needs to scale data science is what should guide the required platform definition. An Enterprise MLOps platform also removes the non-technical barriers by enabling best practices for collaboration, knowledge sharing, and project management.

This expanded definition of MLOps – Enterprise MLOps – is a vital foundation of successfully building a safe, governed, revenue-generating data science flywheel at scale.

"We see a lot of organizations talking about MLOps as referring to models in production, monitoring, and maintaining them. However, MLOps is not just about managing models in production. It's about the whole business of getting people to collaborate around the entire process. It has been interesting to see that evolve recently in a way that actually speaks more to the value it provides."

-Matt Aslett, Research Director, AI and Data at 451 Research, a part of S&P Global Market Intelligence

Examples of Barriers to Scale R&D for Models

Silos. Stovepipe solutions for development using different tool stacks prevent collaboration and increase the burden on IT support.

Resources. Inability to access the right compute and tools for the task.

Governance. Ungoverned usage of compute both in-cloud and on-premises.

Software. Lack of software environment management leads to an inability to share, collaborate, and time-travel to past work states.

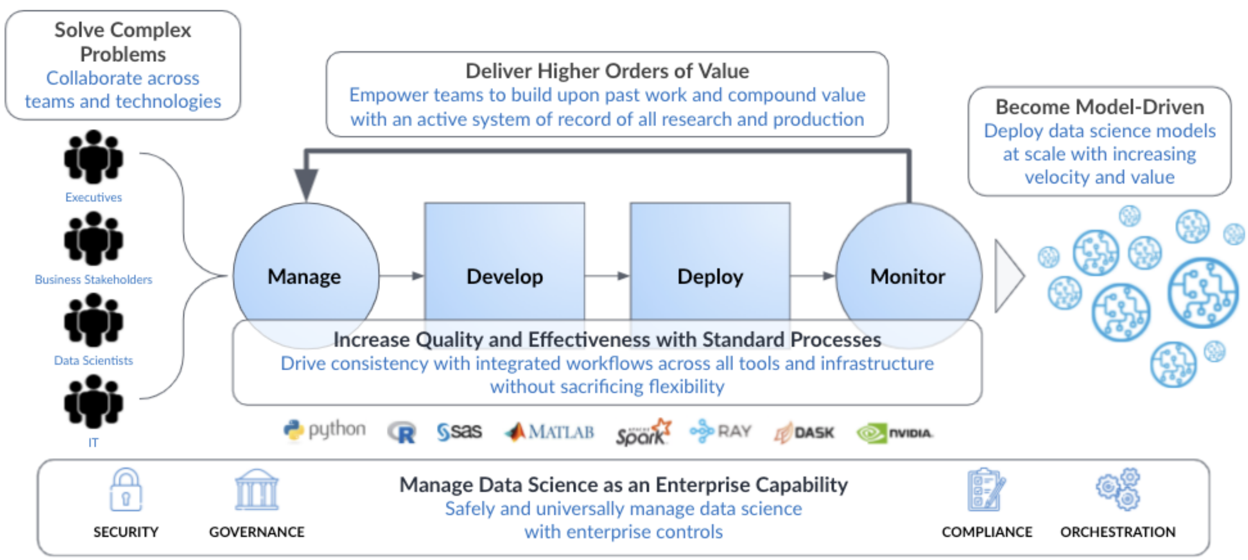
Visibility. No repeatable processes for project management and progress tracking make the flywheel unmanageable.

Lineage. No lineage between production assets and their R&D work make retraining and rebuilding fraught with error.

Enterprise MLOps in Practice: Four Case Studies

The core capabilities of an Enterprise MLOps platform provide an operational blueprint to scale data science for model-driven companies. The capabilities cover four phases of the full data science lifecycle: manage, develop, deploy and monitor. Focusing on the provision of capabilities for the entire lifecycle will help a model-driven business to avoid the mistakes and issues that are common to entities following the more limited definition of MLOps. This section describes how four large enterprises have successfully used Enterprise MLOps to operationalize data science at scale. Each case study describes how a company leveraged Enterprise MLOps capabilities at scale for a particular phase of the data science lifecycle.

How Enterprise MLOps Scales Data Science



1. Managing the Data Science Lifecycle

Manage is the first phase of the data science lifecycle. A significant objective of this phase is breaking down knowledge silos that keep data scientists from collaborating. Because data scientists often work independently with a variety of tools, there are no standard ways of working, which compromises governance, auditability, reproducibility, and so forth.

Strong project management capabilities are also essential for scalability in this phase. They enable governance and collaboration by large teams of stakeholders and facilitate audit and review processes.

To illustrate how one enterprise met the challenges of the Manage phase, consider the experience of SCOR, the world's fourth largest reinsurance company. SCOR helps clients control and manage risk—natural risks, climate risks, health risks, geopolitical risks, cyber risks, and many others. And they help people rebuild when adversity occurs.

"Our success is deeply rooted in our ability to understand an issue and collaborate with others to solve a problem." noted Antoine Ly, Head, Data Science. In recent years, SCOR's Research & Development division was renamed the Knowledge team and organized with chapters dedicated to specific communities of expertise to reflect key focus areas of the company. In an effort to improve efficiencies in the Manage phase, SCOR focused on collaborative best practices. They gathered their best technical experts from across regions to develop templates and strategies for using Python and R. These experts looked across all the different projects and developed a strong skeleton that could be reused from one project to another and brought a systematic approach to the early phases of creating an application or API.

SCOR leveraged new technologies to ensure the enterprise would achieve the kind of knowledge sharing they envisioned. Ly said, "To share knowledge, practices, and code, we have to share tools. To this end, we've implemented a multi-cloud strategy and platform. For example, we launched a dashboarding initiative. We're expanding on some existing dashboards to monitor the data and control some of the different models so that other markets can take advantage of them. We've already made one developed in Australia available to the European and US markets. We've also used the platform to extend the use of an API developed in Europe to make it available worldwide."

Essential Enterprise MLOps Functionality: Manage Phase

Project management. For total control and access of all resources including a master repository, status tracking, project arc, authorizations, information access, and portfolio snapshot.

Enterprise-wide knowledge management. Automates keeping everyone in the loop and avoids the embarrassing and wasteful scenario of a data scientist being forced to start with an empty coding console. Having the ability to search for prior art is a best practice for efficiency. Capabilities include sharable access and the ability to organize and track work.

Governance of technical resources. If governance fails, the burden of managing an MLOps platform can quickly become overwhelming. Capabilities include cost tracking and controls, user/role-based permissions for access to information, tools and compute resources; and intelligent processes for efficient use of people and resources.

2. Developing Models for Business Use Cases

Model building is the core function of data science work. Development includes access to tools and infrastructure such as powerful compute resources, high-value and sensitive data, and the latest open-source tools and packages to support diverse experiments. The flexibility and agility enabled by Enterprise MLOps principles is essential both to making data scientists productive and for innovation.

When data science teams cannot get access to the infrastructure they need, they create ad-hoc workarounds that involve building and maintaining their own local infrastructure. This might include unsecured laptops, local servers, and unmanaged cloud environments. The multiple stacks of data science tools and bespoke hardware for each team slows data scientists down, frustrates IT support teams, creates significant support costs, and increases operational and security risks.

Obstacles like these were slowing efforts by a Fortune 500 global financial services leader to scale its data science practice across the enterprise. The company's Analytics Center of Excellence made it a priority to reduce overall time for development of models. An

Enterprise MLOps platform was adopted to streamline collaborative development of models by teams spread across different geographies.

> The flexibility and agility enabled by Enterprise MLOps principles is essential both to making data scientists productive and for innovation.

With the adoption of Enterprise MLOps, the financial services company enabled faster development at scale without compromising IT security requirements. Developmental improvements with Enterprise MLOps include:

- Shared resources. The company was able to combine SAS, R, and Python in common projects shared across teams.
- No more silos. The platform allowed for centralized, IT-blessed, self-service of right-sized compute. Server silos were eliminated.
- Single point of access to data via the centralized platform for better governance and security.
- Environment management was centralized and shareable, which removed the need to manage siloed dependencies.

"Enterprise MLOps helps us bring our distributed team together in a collaborative way so we can operationalize data science, at scale, across the company," says a Senior IT Architect at the financial services company.

Benefits of Enterprise MLOps

✓ **Capacity** - Increased throughput and capabilities of data science teams

✓ **Quality** - Reusability, knowledge management and collaboration

✓ **Governance** - Tools and infrastructure options for any use case

✓ **Operations** - Enable standards and governance across teams

✓ **Management** - Align personnel, infrastructure costs, and project prioritization with business value

Essential Enterprise MLOps Functionality: Develop Phase

Access to data. Easy, secure, permission-based access to data sources and feature stores are foundational.

Environment access. Entails managing a shareable, versioned collection of environments that define the exact specifications of software, drivers, frameworks, and integrated development environments (IDEs) a data scientist will use to do their work.

Tool flexibility. Giving data scientists the flexibility to use a wide variety of open source and proprietary IDEs and tools, all within a common project.

Flexible compute access. Self-service access (without requiring IT support) to CPU and GPUs, high memory nodes and clusters.

Code versioning. To incorporate full-featured native or integrated code repositories.

Experiment tracking. Supports experimental and iterative processing of tracking experiments, with speed, reproducibility, collaboration, and model audits.

Job scheduling. For hyperparameter search, long ETL jobs, report generation, and other use cases.

3. Deploying Models for Production

Deploy is the phase for operationalizing models at scale. The best model in the world means nothing if it is not moved into production to improve actual business processes.

Traditionally, the Deploy phase is the sweet spot of MLOps. But for many organizations deployment can be an onerous process requiring constant reinvention of steps requiring close oversight and assistance by IT support staff. Data scientists should not have to rely on IT or software developers to deploy every model they create! And models are not the only thing deployed that adds value to a business. Web apps and reports are two examples of other data products created by data science teams.

Finding a robust Enterprise MLOps solution that could improve model deployment was an important goal for Moody's Analytics, the New York-based company that supplies expertise and tools such as data, models, software, and professional services to help customers grow efficiently and manage financial risk. "The cost of improving or replacing a model was too high," said Jacob Grotta, General Manager of Banking Operating Unit. In a competitive industry, the company needed a standardized way to deploy models.

Using the technology and principles of Enterprise MLOps for model deployment, data scientists were able to develop an API and share it with customers for beta testing within a few days. They used feedback to make adjustments and redeploy the model almost instantly, eventually embedding it into a product release. "Rather than taking a year, the process took a couple of months, and the cost of deployment was much lower," Grotta said. That's a 6X performance acceleration.

Moody's Analytics can now efficiently deliver customized models for risk and other analytics that help run large-scale enterprises, and cost-effectively deploy them according to customer preferences, either on premises, in the cloud, or as SaaS. Data scientists at banks can now deploy new models on their own, saving time and adding value to the business.

> Our Enterprise MLOps platform "accelerates our speed to delivery, providing a much faster and better return on our modeling investment."
>
> ***– Jacob Grotta, General Manager of Banking Operating Unit at Moody's Analytics***

Essential Enterprise MLOps Functionality: Deploy Phase

Flexible hosting. Enables deployment of model APIs into a variety of hosting infrastructures without the data scientist needing to wait for IT to provision a stack of resources.

Seamless process. Provides seamless deployment of APIs, web apps, and other data products that are permissioned and accessible, including senior and executive stakeholders.

Flexibility for unique use cases. Allows packaging models so they are easily consumed by external systems to help deliver business value.

Data pipeline. Supports complex data input flows to orchestrate and manage the data pipeline.

4. Monitoring the Model Portfolio for Ongoing Performance

Monitoring is about keeping track of model performance, ensuring that models continuously learn, ensuring they are continuously rebuilt (CI/CD), and preventing model drift or even the improper use of models. While these objectives may seem obvious, many (if not most) enterprises that fail to scale models in production are falling short in the Monitor phase because they are disengaged with systematically ensuring performance.

To implement monitoring at scale, Enterprise MLOps needs to integrate a strong model maintenance plan. The risks of ignoring responsibilities of monitoring pose real consequences from bad models or their improper use – including significant monetary and brand reputation risks. Model maintenance should make it easy to trace the history of models and easily reproduce them in follow up experiments, tuning, and re-validation.

Improving its model monitoring capability was one motivation for Topdanmark's move to Enterprise MLOps. Topdanmark is a large European insurance company based in Denmark. It infuses data science across its operations to provide consumers with a better, faster insurance experience.

> "AI-enabled companies develop the skills, processes, and technical systems to build global learning loops that turn individual knowledge and local insights into an ever-increasing flow of collective wisdom that everyone in the organization shares and contributes to."
>
> *—McKinsey & Company, Winning with AI is a State of Mind*

The company adopted Enterprise MLOps technology and practices to get insights into how models are performing in real time, and to detect data and model drift once models are in production. "Data drift can have a critical impact on predictions and ultimately, our business," said Stig Pedersen, the company's head of Machine Learning. He noted that their new approach, "saves us significant time previously spent on maintenance and investigation, and enables us to monitor model performance in real time and compare it to our expectations." In one case, they were able to automatically detect drift that had previously taken three months to identify manually.

Desirable MLOps Functionality for the Monitor Phase

Pipeline verification. Enables testing and deploying of scoring pipelines via CI/CD principles.

Idea testing. Facilitates A/B testing of model versions in production and track results to inform business decisions.

Asset repository. Provides a model repository for all deployed assets (model APIs and otherwise) across the enterprise with metrics to gauge health, usage, and history of models.

Integrated monitoring. Seamlessly integrates model monitoring after deployment.

Easy re-iteration. Enables model retraining/rebuilding with full history and context of original modeling work and previous versions intact and easily consumable.

Conclusion

AI models are the anchor of modern businesses and the focus of big investments to generate significant new revenue. Achieving this aspiration requires more than just a team of smart data scientists. It also entails use of modern principles for creating and managing the production and enterprise deployment of models at scale. The technology and principles of Enterprise MLOps ensure that model performance is consistently and reliably tied to a set of standards for data science excellence – including the flexibility to switch between data science tools and infrastructure on demand. Enterprise MLOps also integrates disparate tools, teams, and data science artifacts to establish visibility, repeatability, and reproducibility of the full data science lifecycle for every use case. Enterprise MLOps allows a model-driven business to power an analytical flywheel, which lets leaders act dynamically and decisively to leverage valuable insights and harvest an ever-growing flow of collective wisdom. If your organization aspires to these rewards, we invite you to read our Guide to Enterprise MLOps and Forrester Consulting's "The Total Economic Impact™ of the Domino Enterprise MLOps Platform" for more information about how and why to create your own analytical flywheel and achieve real breakthroughs in data science learning and scale.

About Domino

Domino powers model-driven businesses with its leading Enterprise MLOps platform that accelerates the development and deployment of data science work while increasing collaboration and governance. More than 20 percent of the Fortune 100 count on Domino to help scale data science, turning it into a competitive advantage. Founded in 2013, Domino is backed by Sequoia Capital and other leading investors. For more information, visit dominodatalab.com.