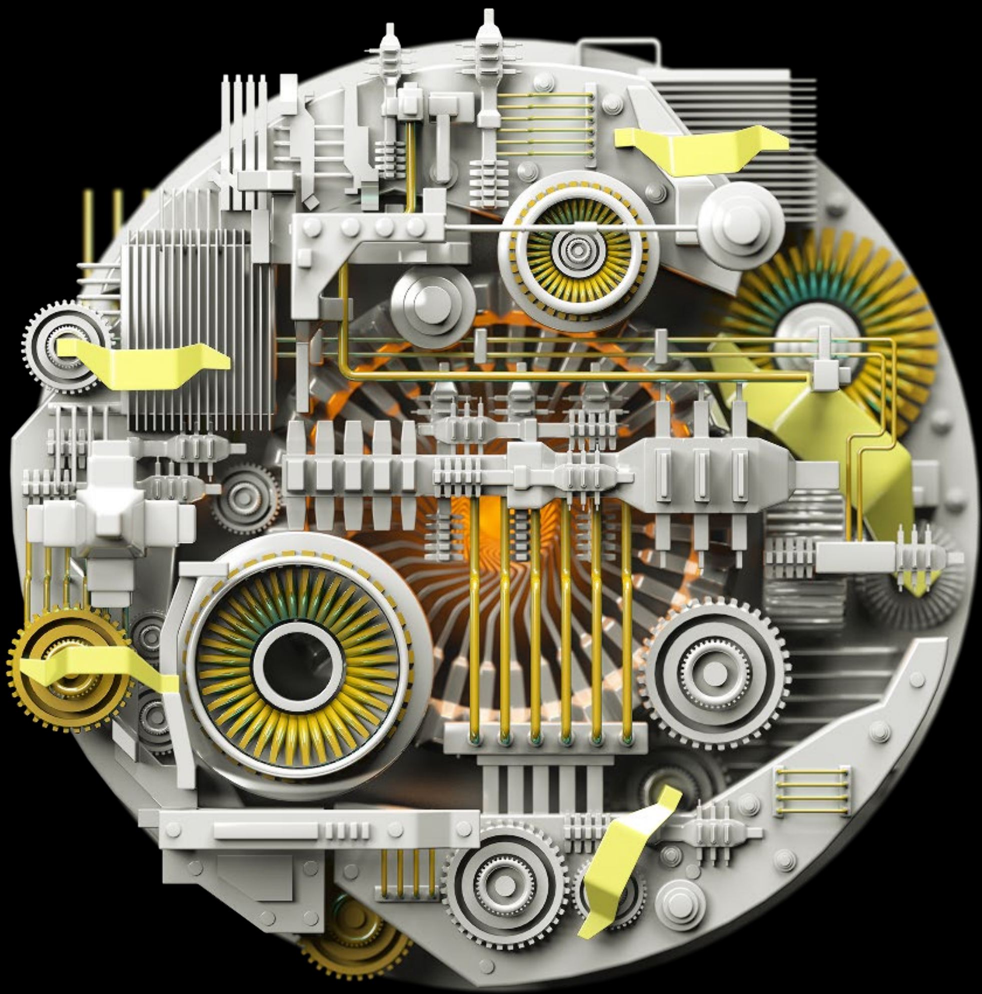
Deloitte.



The part of MLOps less talked about

Successful operationalization of machine learning at scale requires more than just technology

MAKING AN IMPACT THAT MATTERS since 1845

Organizations that implement and enforce MLOps are twice as likely to achieve their goals. They are also about two times more likely to report being prepared for the risks associated with AI, and nearly two times as confident that they can deploy AI initiatives in a trustworthy way.

State of AI in the Enterprise – Edition 4, 5

The need for an operating model approach to machine learning at scale

It's not news that Artificial Intelligence (AI) is fundamentally reshaping how companies run and improve their business, and Machine Learning (ML) is the main contributor to its success[1]. Yet successful operationalization of ML, key to capturing the immense value AI brings, proves challenging according to the State of AI in the Enterprise 4 and again 5[2]. Tackling this challenge has led to the rise of Machine Learning Operations (MLOps).

MLOps is a set of practices that aims to develop, deploy, and maintain machine-learning models in production, reliably and efficiently.

While the tech community has focused on MLOps with countless tools and technologies emerging in recent years[3], there's more to the story: simply having the right technologies is not enough to operationalize ML at scale. Two additional and equally important components deserve attention: People and Processes have emerged as "missing links" in our work with clients.

This high-level and non-technical guide shows leaders from business and IT alike how to navigate the complex and continuously evolving MLOps landscape by examining the People, Process, and Technology components in MLOps from a holistic perspective. Against this backdrop, we introduce a maturity assessment that will help in identifying and filling any gaps in reaching the MLOps level you need.

Organizations still struggle to capture lasting value from Machine Learning

Even though much effort and resources have been invested in harnessing the power of AI, most organizations still struggle to retain real, sustainable value from it[4, 5, 6, 7]. Lasting impact will only come from ML models that have been designed, operationalized, and embedded into business processes at scale. The inability to move from ML experimentation to operationalization will only lead to sunk costs. The operationalization of ML solutions should thus not be treated as an afterthought.[8]

The inability to operationalize the ML model will only lead to sunk costs.

How can you ensure that your machine-learning endeavors deliver value and return on your investment?

MLOps: More than just a set of tools and technologies

How does MLOps pave the way for enterprise-grade AI at scale? MLOps overcomes the challenges in operationalizing Machine Learning by encompassing a set of practices around people, processes, and technologies. It enables the continuous experimentation, development, and operation of ML systems. According to our research[2], companies that are successful in their AI journey make heavy use of MLOps.

Inspired by DevOps, a set of practices to bridge the gap between the development and operation of software, MLOps streamlines and standardizes the end-to-end ML lifecycle. MLOps aims to close the loop between ML modeling and its operationalization, thereby optimizing the ML release process.

But make no mistake: It's not just the right tools and technologies. Bridging the gap between the people involved in the ML lifecycle and the appropriate processes is equally important and crucial to success.

Just as AI requires rethinking approaches and processes, ML requires a new approach to the novel elements that make its operationalization so different. Three of the more popular ones are:

- ML code represents only a tiny fraction of real-world ML systems (see Figure 4). The infrastructure surrounding the ML code and containing the model at its heart is vast and complex. Operationalization of ML implies integrating and running the solution with the entire sociotechnical system within which the ML model works, including the workforce, data provisioning, and serving infrastructure.

- Data influences the results of ML-enabled solutions more than the code itself. This strong data dependency requires more tests and validations than "classical" software. There must thus be processes in place that support the validation, monitoring, and continuous iteration of the "data-to-prediction" lifecycle.

- High talent scarcity in the data and ML market requires guardrails, solid practices, and access to education (e.g., through colleagues in cross-disciplinary team setting).

MLOps is designed to address these and other specificities implied by ML and the surrounding environment. It's here to stay.

The three pillars of succesful MLOps

A holistic operating-model approach to the implementation of ML systems at enterprise-scale requires three pillars (see Figure 1):

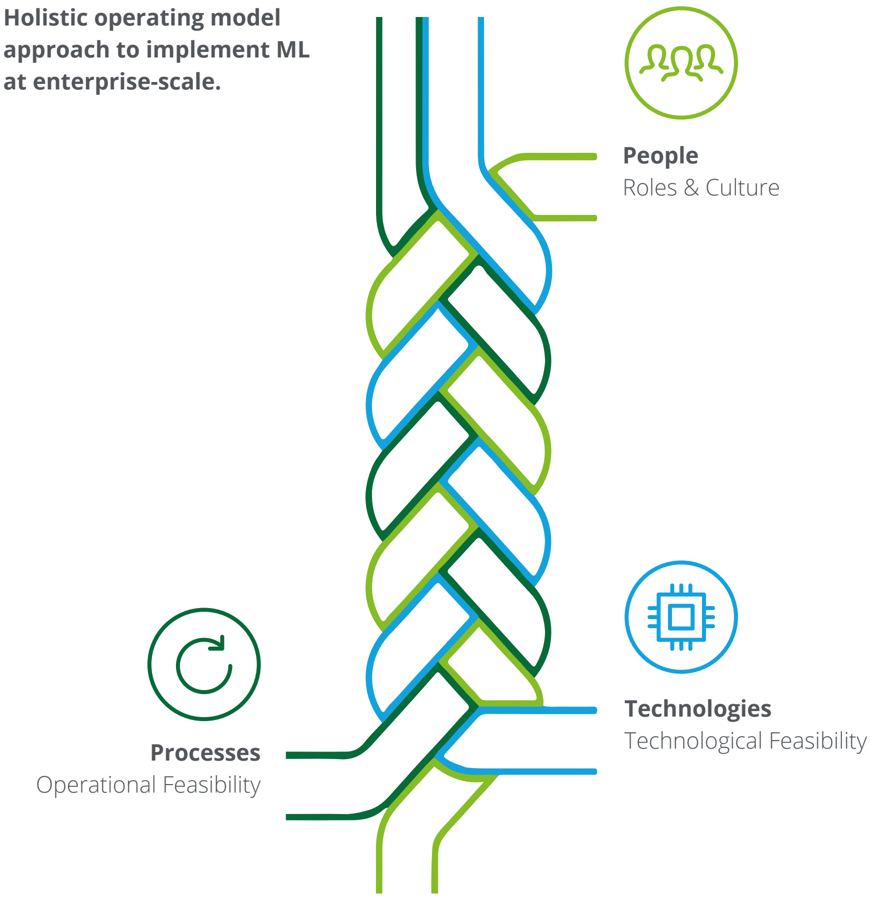
People
Put people at the center with a focus on roles and culture.

Processes
To ensure operational feasibility you need to have the right processes in place.

Technologies
Use appropriate technologies to ensure technological feasibility.

Bear in mind that these pillars are not mutually exclusive. Together they form an interdependent system that, when integrated, serves as a foundation for successful ML system operationalization at scale.

Fig. 1 – Intertwinement of People, Process, Tech in MLOps setting



Put People at the center

The results from our study on the State of AI in the Enterprise, Edition 5, show that establishing clear structures and (re)defining roles, supplemented by an open culture, result in improved outcomes (see Figure 2).

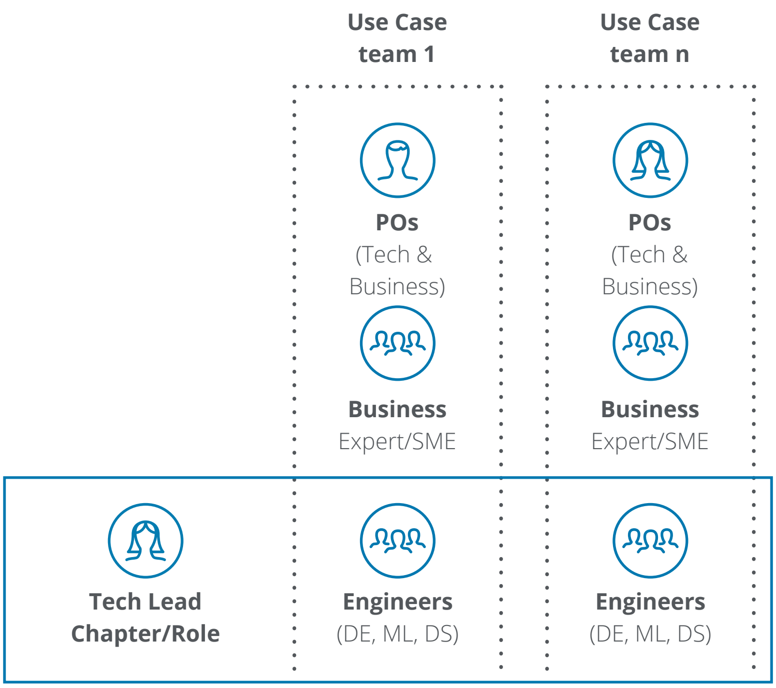
Business Experts/SMEs define the project goal on the basis of the business problem. They are specialists in their field.

Data Engineers build, prepare and run data pipelines that serve as a foundation for ML training.

Data Scientists translate the business problem into ML by using available data.

ML Engineers focus on developing, deploying and monitoring ML solutions and its surrounding infrastructure.

Fig. 2 – Interdisciplinary teams capable in end-to-end delivery



Have the right Processes in place

"If you're not changing how you work, you might be leaving value on the table". How you redesign your operations to accommodate the unique demands of ML will impact how successful deployment at scale will be[2] (see Figure 3).

Modeling
Starts with scoping the business problem and data acquisition, and ends with the training and evaluation of carefully selected ML model(s).

Validation
After sufficient testing of the different ML pipeline components, additionally model performance is validated, thereby bridging the gap between experimentation and operationalization.

Operations
Machine learning pipelines are deployed to serve the customer. Once running, ML models and respective components are then monitored and validated against specific KPIs.

Use of the appropriate Technology

While companies invest a lot into ML, they often neglect the surrounding system landscape that the ML model depends upon to deliver value. This includes, among other things, automating data/feature engineering pipelines, using orchestration tools as well as monitoring models in production, essential to detecting and analyzing unexpected behavior before it causes major problems.

Note that this article focuses on the people and process components, as countless technologies, used in the setting of MLOps, have already been covered extensively.

Fig. 3 – MLOps Phases – Based on [10] and adapted for the purpose of this document

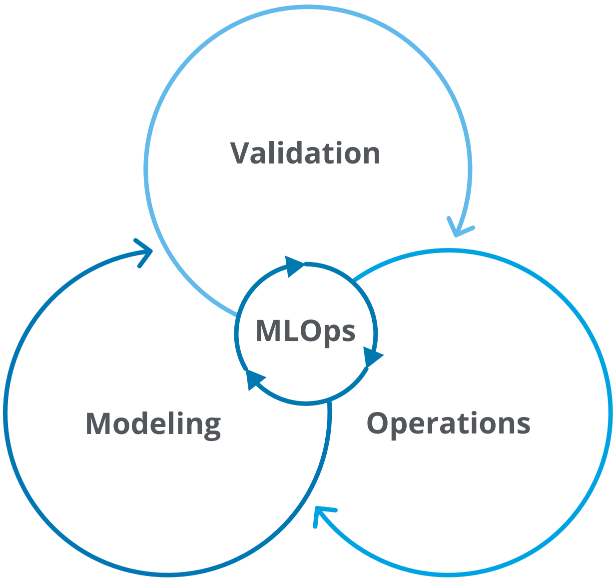
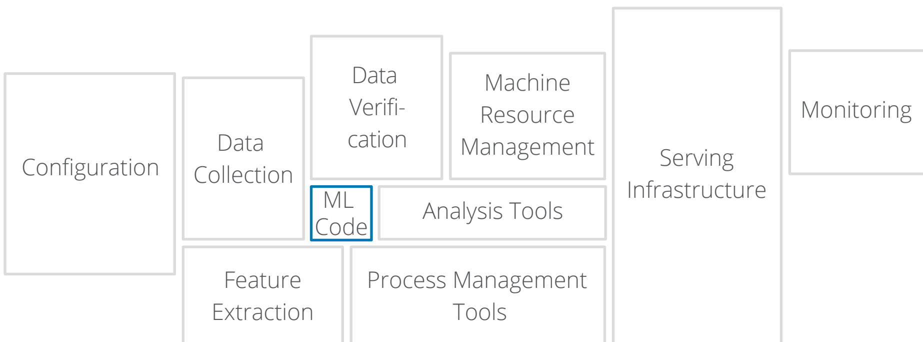


Fig. 4 – ML System Components – Based on [11] and adapted for the purpose of this document



Starting your journey to MLOps

Don't miss out on the potential of ML unlocked by MLOps. Having seen what it is and why it exists, let's take the next step to see who is affected by MLOps, as well as where and how to start the journey.

Who should care about MLOps? Any company that wants to capitalize on AI!
Regardless of their experience with ML and which industry they operate in, all companies will benefit from applying MLOps practices to ML projects aligned with their people, processes, and technologies. It ensures reproducibility and transparency, and enables fast pivoting in response to changing objectives.

Where to start with MLOps? Start by identifying what is missing!
The idea is not to start from scratch but to capitalize on the status-quo in your organization. Any organization wishing to implement MLOps must ask itself several questions:

- What works in our organization?
- Where should we start?
- Which actions should we take?

We have designed a maturity assessment to help paint an accurate and complete picture of any gaps by using the operating model pillars as follows:

People
Assess whether you have the right talent working on ML and its surrounding environment. Find out if you need to hire new talent or upskill existing employees. Clarify what is required to set up a sound team structure, capable of end-to-end delivery, and propagate an open culture to facilitate collaboration and knowledge exchange. High inter-/intra team collaboration will lead to better team dynamics and efficiencies, and greater impact.

Processes
The assessment of gaps in processes is crucial. MLOps requires reliable data pipelines, automated workflows/orchestration, and agility. Pinpoint what is needed to follow standards and best practices that improve reliability in data provisioning, time to market, and adaptation speed to change.

Technologies
Identify gaps early in the process by assessing your technological landscape and evaluating it against novel requirements of ML use cases. Clarify how to fill gaps by adapting to new ML requirements. Choosing the right technologies, including feature stores, orchestration and ML pipeline tools, will create a more efficient ML system.

Fig. 5 – Maturity Level Assessment and main characteristics for each category

	MLOps Level 0	MLOps Level 1	MLOps Level 2
People	Homogenous, completely decoupled, teams (e.g., data scientist team, Software Engineer team etc.)	First signs of interdisciplinary teams (e.g., data scientists paired with data engineers)	Highly interdisciplinary teams capable in the end-to-end delivery of ML solutions
Process	Manual experimentation and deployment with low collaboration between roles/teams	Orchestrated experimentation and continuous training pipeline with data and model validation	Standard, highly automatized, process intertwining modeling, validation and ops
Technologies	No use of enabling technologies such as orchestration, versioning etc.	Feature – and metadata stores, (data) version control and first orchestration techniques	Orchestration tools, CI/CD/CT enabling capabilities, Infrastructure as Code, observability techniques

Knowing your maturity is interesting, but not actionable. Our maturity assessment helps lay out a roadmap towards efficient and effective MLOps implementation, suited to your needs. The speed at which your MLOps capabilities improve largely depends on the starting point with regard to people, process, and technology, as well as organizational flexibility. By identifying quick wins along the MLOps lifecycle, efficient MLOps capability can evolve to support powerful ML systems. Growing your MLOps capabilities along with your ML systems will help identify and mitigate risks linked to the latter.

Crossing the chasm

ML-based systems are constantly evolving to boost not only business efficiency but to allow for new business models. But current slipshod approaches lead to risky experimenting with ML solutions, as well as the operation and maintenance of ML systems for costly one-off activities. In fact, every second respondent in our recent survey[2] reported AI related risks and lack of maintenance or ongoing support of ML solutions as highly challenging.

MLOps is key to capitalizing on the promise of AI, enabling business to use ML in a structured, de-risked and efficient way. However, there are many MLOps pitfalls; from experimentation and continuously deploying the model to keeping it working as intended.

Some of these pitfalls come to light when asking questions like:

- Do you have the right people working on ML projects?
- Are your ML systems deployed quick enough to create ROI?
- Can you update and operate dozens to hundreds of ML models and ensure they are trustworthy and working reliably?

While these are just examples, leaving these questions unaddressed could risk falling short of expected results or worse.

MLOps cannot exist in a vacuum. There is a cultural shift that the top of business, jointly with technical decision makers, must understand and push for to ensure the organization makes full use of MLOps, and thereby AI at an enterprise level.

01. European Parliament, Economic impacts of artificial intelligence (AI), https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/637967/EPRS_BRI(2019)637967_EN.pdf, 2019
02. Deloitte, Deloitte's State of AI in the Enterprise, 5th Edition report, https://www2.deloitte.com/content/dam/Deloitte/us/Documents/deloitte-analytics/us-ai-institute-state-of-ai-fifth-edition.pdf, 2022
03. Chip Huyen, What I learned from looking at 200 machine learning tools, https://huyenchip.com/2020/06/22/mlops.html, 2022
04. Venture Beat, Why do 87% of data science projects never make it into production?, https://venturebeat.com/ai/why-do-87-of-data-science-projects-never-make-it-into-production/, 2019
05. Deloitte,State of AI in the enterprise, 4th edition, https://www2.deloitte.com/us/en/insights/focus/cognitive-technologies/state-of-ai-and-intelligent-automation-in-business-survey.html, 2021
06. Forbes, AI Stats News, https://www.forbes.com/sites/gilpress/2020/01/13/ai-stats-news-only-146-of-firms-have-deployed-ai-capabilities-in-production/?sh=551600fc2650, 2020
07. Gartner, Nearly Half of CIOs Are Planning to Deploy Artificial Intelligence, https://www.gartner.com/en/newsroom/press-releases/2018-02-13-gartner-says-nearly-half-of-cios-are-planning-to-deploy-artificial-intelligence, 2018
08. A. B. Kolltveit and J. Li, "Operationalizing machine learning models – A systematic literature review." 2022
09. World Economic Forum, Perfecting the art of the whole organisation with machine learning, https://www.weforum.org/agenda/2022/04/perfecting-the-art-of-the-whole-organisation-with-machine-learning/, 2022
10. Gartner, MLOps Framework to Operationalize Machine Learning Projects, https://www.gartner.com/document/4018020?ref=solrAll&refval=349924345, 2022
11. Google Cloud, MLOps: Continuous delivery and automation pipelines in machine learning, https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning, 2020

Contacts

Dr. Björn Bringmann
Managing Director
AI Institute – The Garage
Tel: +49 89 29036 6131
bbringmann@deloitte.de

Nadja Stark
Senior Manager
AI Institute – The Garage
Tel: +49 30 25468 3628
nastark@deloitte.de

Ali Aras Gündüz
Senior Consultant
AI Institute – The Garage
Tel: +49 30 25468 5241
aguenduez@deloitte.de

Karl-Eduard Berger
AI & Data Science Lead
Analytics and Information Management
Tel: +33 158 379452
kaberger@deloitte.fr

Alexandre Pinot
Partner
Technology Strategy and Architecture
Tel: +33 140 884910
apinot@deloitte.fr

Alexandre Breboin
Consultant
Analytics and Information Management
Tel: +33 155 614749
abreboin@deloitte.fr

Visit the Deloitte AI Institute:
www.deloitte.com/de/de/pages/innovation/solutions/ai-institute.html

Issue 01/2023