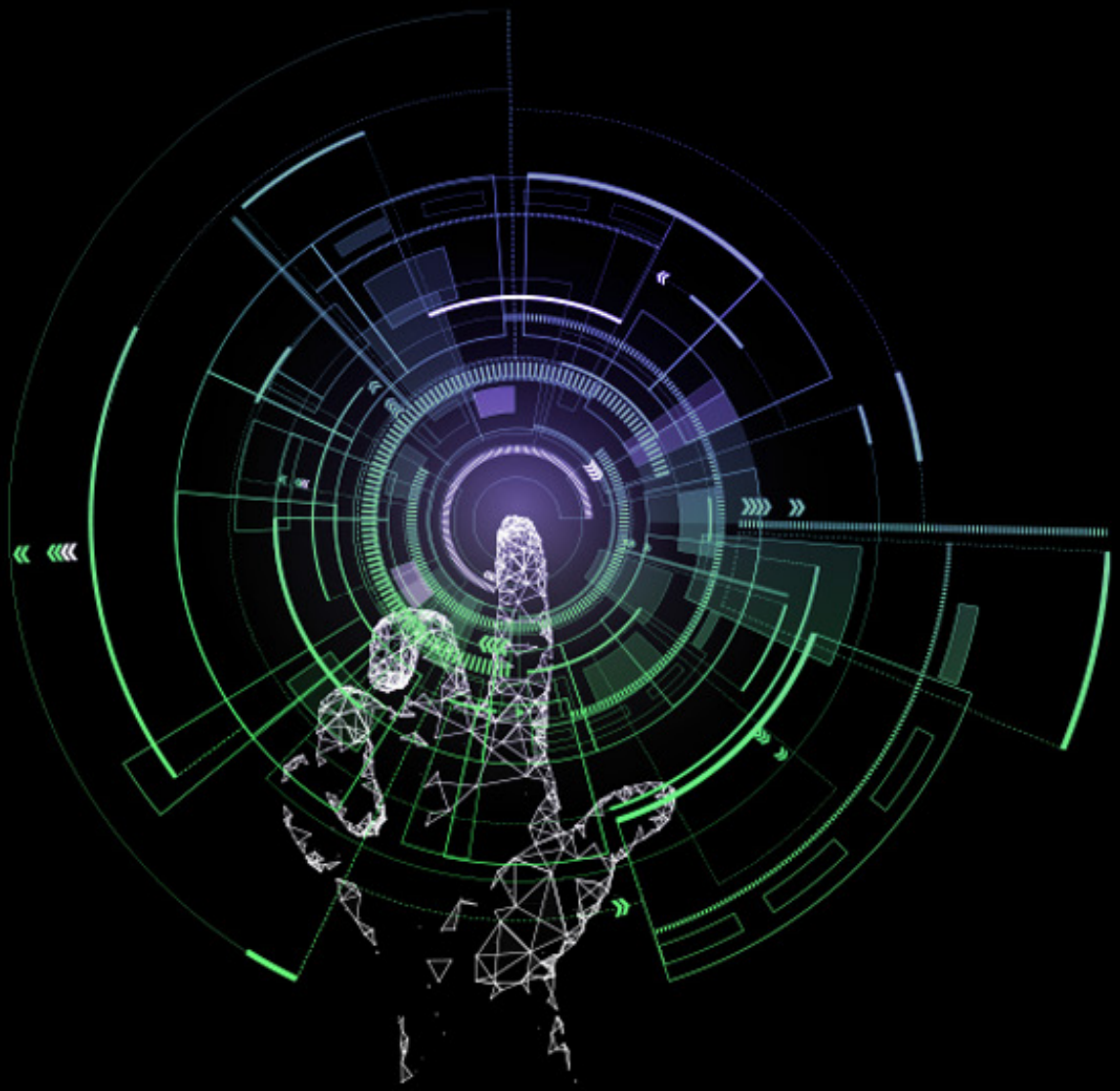
Deloitte.

A report by the Deloitte AI Institute



ML-Oops to MLOps

About the Deloitte AI Institute

The Deloitte AI Institute helps organizations transform with AI through cutting-edge research and innovation, bringing together the brightest minds in AI to help advance human-machine collaboration in the Age of With. The institute was established to advance the conversation and development of AI in order to challenge the status quo. The Deloitte AI Institute collaborates with an ecosystem of industry thought leaders, academic luminaries, start-ups, research and development groups, entrepreneurs, investors, and innovators. This network, combined with Deloitte's depth of applied AI experience, can help organizations transform with AI. The institute covers a broad spectrum of AI focus areas, with current research on ethics, innovation, global advancements, the future of work, and AI case studies.

Connect

To learn more about the Deloitte AI Institute, please visit www.deloitte.com/us/AIInstitute.

Introduction

Artificial intelligence (AI) and machine learning (ML) are ubiquitous. Market microstructures in the industry are changing in response to evolving AI, ML, and cloud technologies. Businesses need insights driven by data, optimized by AI, built on the cloud and at the point of decision making.

- Businesses continually need a nuanced understanding of their customers and their evolving behaviors to maintain a competitive edge. Plus, competition from nontraditional sources is forcing a revisit of the existing strategic and operational paradigm.
- Massive proliferation of data continues to add more texture to what insights can be gleaned and how decisions can be made. It's predicted that 175 zettabytes of data will exist by 2025.[2]
- In keeping with the proliferation of data, there have been rapid advancements in data storage and computational availability. Adoption of the cloud, on-demand models, open-source computing, and the wide availability of data itself in its full glory of volume, variety, and veracity have enabled big data (streaming and static) to be mined, housed, and analyzed.
- The rise of auto-ML tools and platforms has democratized the effort of data mining and decision sciences, and the breed of citizen developers is growing.

This presents both the need for and the potential to capture continuous insights that can inform business decisions. From smart manufacturing to finance transformation to omnichannel customer experience, across all business functions, AI and ML today need to be adopted widely and operationalized. Organizations can drive stronger business outcomes through human and machine collaboration and achieve scale with speed, data with understanding, decisions with confidence, and outcomes with accountability—the Age of With™.

Of respondents surveyed for the Deloitte's State of AI in the Enterprise, 3rd edition:[1]

64%

believe that **AI enables a competitive advantage**

54%

are spending **4x more than last year** on AI initiatives

74%

plan to integrate AI into all enterprise applications **within three years**

ML-Oops...

The concept of machine learning often conjures images of data scientists designing sophisticated algorithms and writing highly technical code. It is presumed that ML is the forte of elite experts delivering signature models and driving innovation, often in niche areas. It is partly true. Those of us who have been at the center of an Applied AI practice, solving real-world business problems at scale, know that seasoned data scientists emphasize that there is a lot more to machine learning and modeling than the ML code or algorithm.

Success in using ML for business impact hinges on the culmination of data, technique, process, and training. Machine learning cycles involve heavy data that, if unavailable or delayed, becomes a blocker. "Waiting for data" is a common refrain in the corridors of data science.

Another crucial modeling input is feature engineering, an energy-intensive and iterative process. With models developed, the focus turns to training, testing, and deploying solutions. Overall, this cycle is predicated on tools and infrastructure, which can complicate the effort, depending on the tech stacks used by organizations and, often, their other vendors and partners.

When machine learning was a small discipline, locally owned, and contained in divisions and functions by a small group of experts, this entire process happened quietly, even smoothly, and was manageable. As AI and ML started getting to the core of enterprise transformations and bearing expectations of being sustainable at scale, there came the need for them to track to a fully functional development, operationalization, and automation cycle. This is the realm of ML operations (MLOps).

In our own experiences helping clients realize impact from what's possible with ML and translate that insight into sustainable performance, enterprises have faced significant challenges around MLOps due to a number of factors. A few illustrative examples to provide a flavor of some routine issues:.

- A large pharma client had a data science workbench, but could not scale its manufacturing and operations research and models across different products or different business aspects in the same product. Transition of data science models from the development phase into production and operation phases was a gap that was labyrinthine to scale and required an end-to-end MLOps process.

There is a chasm between ML and MLOps that can be tricky to scale, and MLOps can turn into ML-Oops.

- A quick service restaurant chain had a large technology stack with a complex and legacy restaurant system in the back end. To transform its pricing, profitability, and customer experience, it needed support across the full cycle of data integration and management, modernization of its legacy systems onto cloud, data science and AI, and ongoing analytics for profitability and business performance monitoring.
- A health care conglomerate found its data science models losing accuracy faster in the pandemic era, as historical and traditional data was rendered ineffective. We helped the client rebuild its models, redesign the modeling approach itself, and develop robust model operations support.
- A retail client did not have the bandwidth to create and manage a data layer for feature engineering at scale. Creation of a feature store and automated feature engineering significantly reduced the time to scale for model design and development.

When processes are not designed optimally, they can cause long lead times that render models ineffective when they get into production. Adverse effects of data quality can amplify when models are in production. Inadequate monitoring and maintenance of models leads to faster degradation and decay. Governance and multilevel metrics vary across business units and functions, leading to unclear and suboptimal linkage to business outcomes.

Organizations that want to scale AI and ML across all areas must focus on implementing a set of standards and a framework to create production-capable AI and ML building blocks.

It is not enough to focus on sophisticated model development. It is also imperative to focus on building foundations of processes that are reliable and repeatable. It will not be possible to industrialize machine learning if the reliance is on a few talented practitioners in niche techniques and technologies; industrialization will require the coming together of a varied mix of talent and technologies.

The AI and ML needs of the enterprise are too big and too complex for any small group with niche skills and bespoke models to run. It requires method, process, and the art of the organization.

...to MLOps

MLOps drives this through the entire life cycle of ML models, from design to implementation to management.

If enterprises develop only a few models for limited product lines in project cycles of a few months, they will see limited value in AI and ML adoption. Sustainable impact will come from a portfolio of machine learning models that are designed, productionized, automated, operationalized, and embedded into ongoing business functions at scale for enterprise-level use.

MLOps is a process, in classic Lean Six Sigma parlance. It is not dependent on a few experts, niche use, bespoke designs, or custom development.

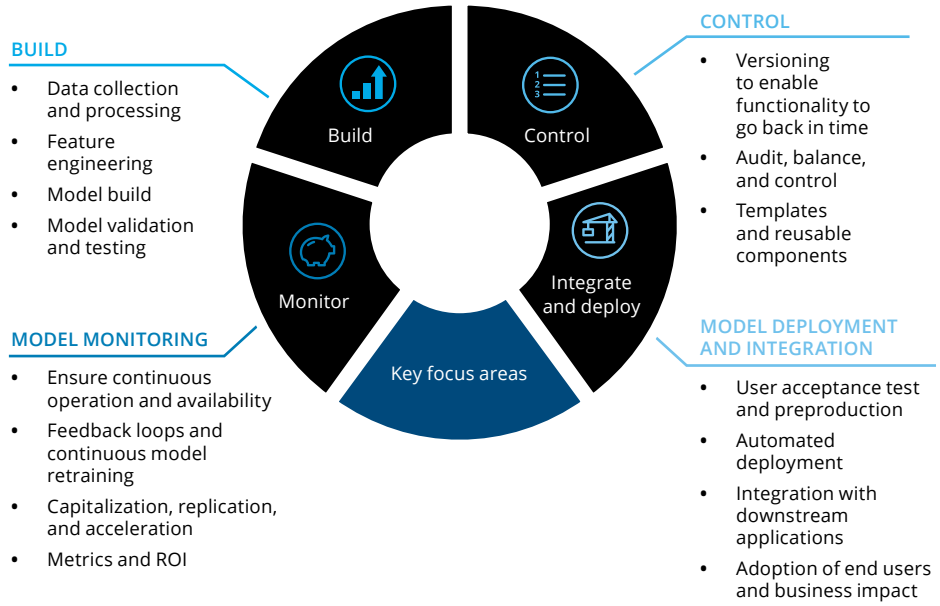
MLOps builds on DevOps

MLOps aims to achieve the core principles of DevOps: automation (as opposed to siloed custom dev); deployment (proliferation, as opposed to one-time use); process (integration, testing, and releasing); and infrastructure considerations.

That said, MLOps builds on and goes beyond DevOps:

- **Core team structures.** For MLOps to be successful, data science and ML modelers need to be in lockstep with MLOps engineers, data engineers, and process experts. It requires a diverse and cross-functional team much more complex than DevOps.

Operationalizing MLOps



- **Experimentation.** ML models are iterative and involve many experiments in their development phase. They also need to stay tuned to the evolving core business issues they are trying to solve—pricing strategies, customer behaviors, competitive intelligence, omnichannel, and industry- and domain-specific issues like the future of work or consumerism.
- **Versatility of testing.** In addition to the standard unit and integration testing, ML testing needs to validate ML models and retrain them.
- **Production and training are subject to change in business fundamentals.** When models are in production, a lot can change. Data profiles will evolve and affect downstream processes, and revalidations of critical assumptions and parameters need to be incorporated.

A few illustrative and representative MLOps best practices deployed at our clients

MLOps principles	Core components	Client scenario
End-to-end design and delivery	• ML current-state assessment • Establishment of desired outcomes • Evaluation of client's vendors and alliances • Design of future-state MLOps—user experience, architecture, operating model	We supported a pharma giant in its AI and analytics vision to be a data-driven organization that can more quickly and cost-effectively discover, develop, and deliver medicines and vaccines. A key outcome was to reduce the time to analytical insights from real-world data from four months to two weeks.
Versioning	• ML model • Code and configurations • Data parameters • ML hyperparameters • Environment	With effective versioning standards, a life sciences client with several hundred models can now perform gradual, staged deployment. With MLOps tools, it can also version-control data and code, in addition to ML model components.
Testing	• Model specification is unit-tested • ML model training pipeline is integration-tested • Out-of-time ML model validations • ML model staleness test (in production) • Testing ML • Testing nonfunctional requirements (security, fairness, interpretability)	Automation testing has helped a health care client quickly discover problems in early phases of development. It also helped the client reduce testing time, in turn reducing overall response time for any new change in model, code, or data.
Automation	• Data engineering pipeline • ML model training pipeline • Hyperparameter and parameter selection	With automated CI/CD pipelines, a retail client can train, build, and deploy ML and data pipelines daily (if not hourly), update them in minutes, and redeploy on thousands of servers simultaneously.
Deployment	• Containerization of ML stack • REST API • On-premises, cloud, or edge	A framework of Dockers and Kuberflow deployments enabled a pharma client to build environments once and ship its training and deployment quickly and easily at any time. The client can easily reproduce the working environment and orchestrate ML pipelines on Kubernetes.
Monitoring	• ML model decay • Numerical stability • Computational performance of the ML model	Our end-to-end design and development of model operations for a consumer client highlights the degradation of model behavior well before time. With dedicated, centralized dashboards, the client is able to monitor all global pipelines.

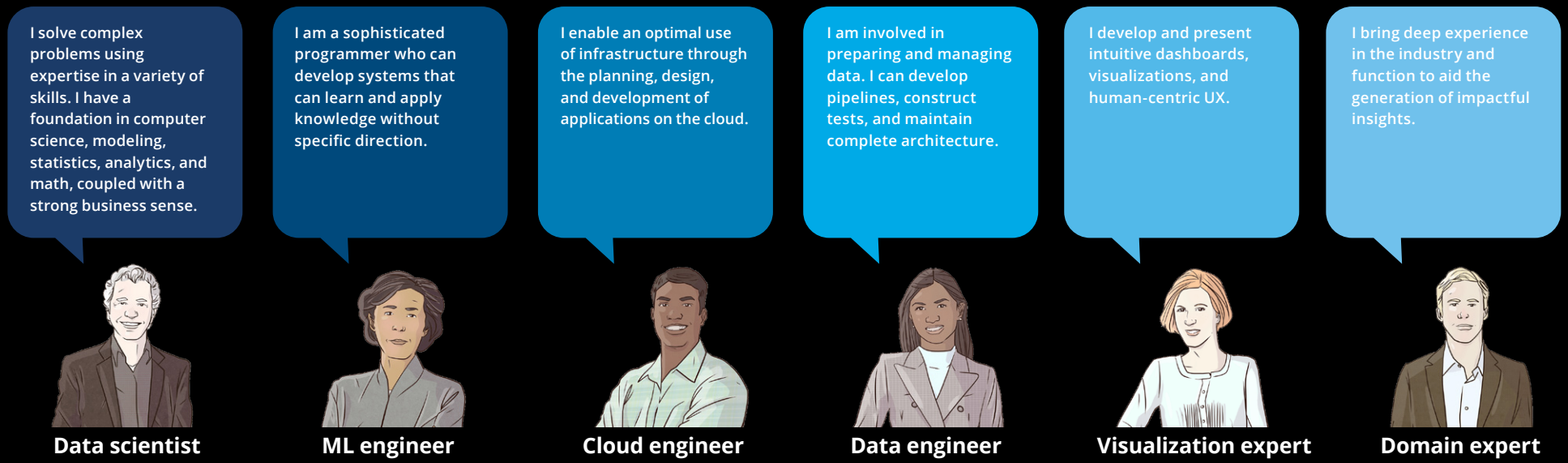
How to make the journey

The path to MLOps and more effective ML development and deployment hinges on selecting the right people, processes, technologies, and operating models with a clear linkage to business issues and outcomes.

This is an evolved state and very much possible in the **Age of With,** in which human-machine collaboration through next-gen assets and platforms predict what is possible and translate the insight into trustworthy performance. Companies invest in bringing AI practitioners and data scientists together into a practice while also investing in preconfigured solutions. Business and domain experts can build use cases around signature issues. The data science experts can drive innovation in machine learning models. Data and ML engineers can use auto-ML tools to stitch together quick ML models.

Talent needed

Creating and sustaining models at scale requires people with capabilities across data science, ITOps, and UX who work seamlessly toward a common goal.



Aligned people with common goals

Adopting a coherent and cohesive inclusive effort to bring people together for a common goal is key. Identifying and clarifying roles and driving collaboration across teams through multilevel governance is necessary.

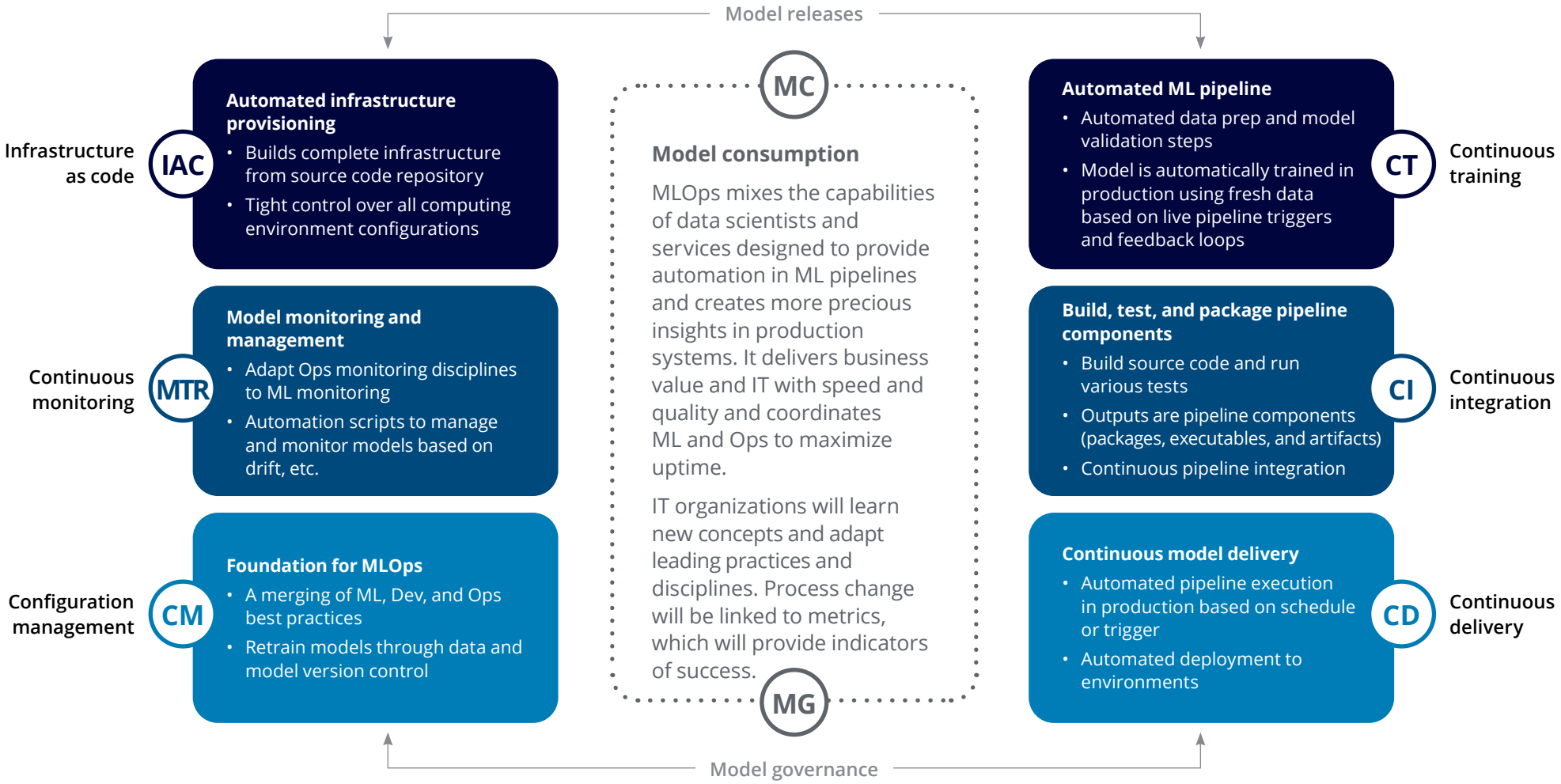
In addition to these core roles, the data and MLOps governance framework must include business program managers, finance and technology, legal counsel, enterprise and model risk, and the enterprise data office and audit.

Automation and efficiency in the process; tools and technology stack woven into the process

MLOps aspires to deliver:

- Reusable plugins and frameworks, automated data preparation and collaboration, and versioning of models so a data scientist can reuse or accelerate use cases based on models created in an as-is state
- Identification of an ML pipeline that feeds into applications, portals, enterprise analytics platforms, and databases
- Cross-pollination and a continuous feedback loop into data stores and feature stores, as well as building and designing automated learnings into repositories
- Model maintenance, with a cadence for data updates, management of variables, scheduling, and deployment of models from anywhere
- Monitoring model drift and model performance for all in production; notifications and alerts on events in the ML life cycle; centralized interfaces and dashboards to monitor ML pipelines

MLOps framework



To sustain impact and outcomes, there needs to be an adjustment of operating models and pull-through of a service catalog of AI solutions through the continuum of Applied AI and Managed AI.

The way forward

Aligning metrics and measures to vision: It is necessary to establish a vision at the outset, then assess readiness. What cannot be measured cannot be improved (a timeless Lean Six Sigma adage). Standards for how to design and measure the efficiency and effectiveness of MLOps are evolving and need to be integrated into performance management.

"The boat matters more than the rowing"[3] – Focus on making process and system robust: MLOps sits at the intersection of skills and process. It pulls together a range of skills and relies on automation, workflows, and systems to drive impact on a sustained basis.

Design for innovation and change: Process-centricity can sometimes obscure that innovation is at the core of AI and ML. The MLOps framework should promote innovation such that the ML itself stays relevant and future-ready. Further, in Black Swan[4] events like a global pandemic, some established processes will be rendered ineffective and dysfunctional. It is important to provide a forum and empower data scientists, AI practitioners, and ML champions to explore, to innovate, and to stay at the cutting edge of this fast-evolving discipline.

Change management: Given that MLOps requires many teams, it also necessitates consumption of models developed by others. This is not easy to implement and requires change management. Model consumers are concerned with the quality and reliability of models not built by them. Different units tend to build their own data science teams and create their own AI setup. This duplicates efforts and causes redundancies, and worse, the best-in-class that exists in the organization might not be known or might not get leveraged.

MLOps is central to industrialized AI.

As AI and ML proliferate across all industries and sectors and are adopted enterprisewide, machine learning and AI models need to be **explainable** in their construct, **trustworthy** in their genesis and underlying data, **measurable** in their impact, **sustainable** in their outcomes, **scalable** in their design, and **self-correcting** in their behavior.

ML is just like any other powerful tool. When used correctly, it can help build. On the flip side, incorrect deployment leads to damage. A major advantage of AI and ML capabilities is speed of analysis and insight on a huge scale, but if misdirected, models can cause suboptimal and even bad decisions at the same speed and scale. To avoid this, or what we call ML-Oops, we need to embed MLOps into all our AI and ML efforts at scale at the design phase itself.

Contacts/Author

Rohit Tandon
Managing Director – Strategy and Analytics
Deloitte Consulting LLP
General Manager – ReadyAI
rotandon@deloitte.com

Sanghamitra Pati
Managing Director – Strategy and Analytics
Deloitte Consulting LLP
Applied AI
spati@deloitte.com

Endnotes

1. Deloitte AI Institute, AI Survey.
2. Tom Coughlin, "175 Zettabytes by 2025," *Forbes,* November 27, 2018.
3. Deloitte internal research.
4. Rolf Dobelli, *The Art of Thinking Clearly* (Farrar, Strous, and Giroux, 2013).
5. Nassim Nicholas Taleb, *The Black Swan: The Impact of the Highly Improbable* (Random House, 2007).

Deloitte.