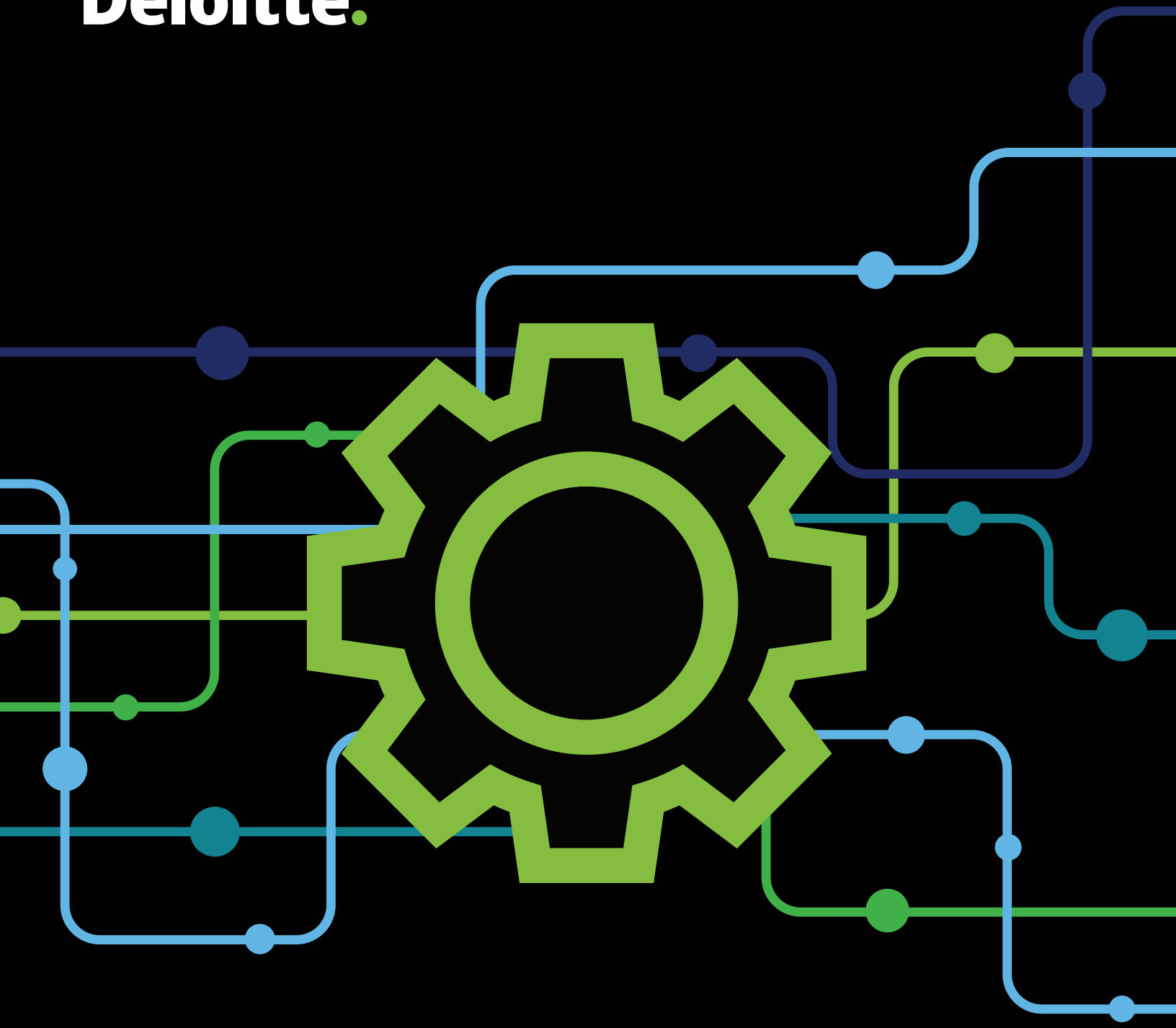
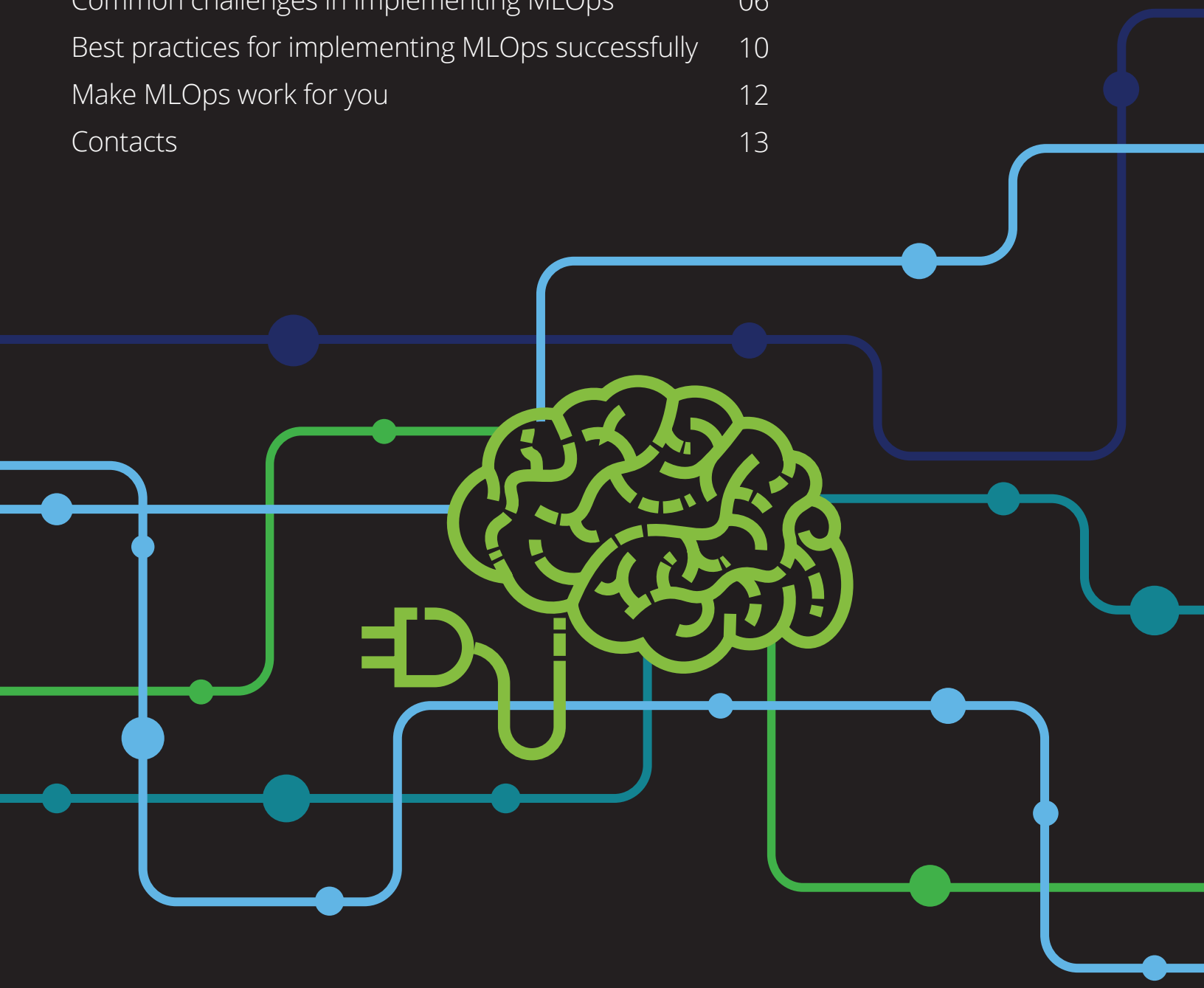
Deloitte.



MLOps

The key to ML model production challenges

MAKING AN IMPACT THAT MATTERS since 1845



Introduction

Organizations worldwide are turning to machine learning (ML) to reshape business models, reimagine business processes, disrupt industries, and create lasting competitive advantage.

At least, that's the idea. In reality, the vast majority of these machine learning projects fail to reach production at scale. Most remain stuck at the stage of model development, validation, and tuning to meet success criteria—never reaching the point where the model is rolled out into production. This happens for a variety of reasons: technology constraints and alignment, knowledge or skill gaps, organizational culture, limited use of development best practices, and more. No matter the reason, it can be enormously frustrating for organizations.

MLOps can help these organizations overcome the issues blocking the progress of their machine learning projects. MLOps is a unique combination of machine learning processes, DevOps principles, and Agile methodologies that streamlines the machine learning lifecycle and helps organizations overcome obstacles to putting machine learning models into production. Used effectively, MLOps can enable organizations to generate significant value.

But adopting MLOps isn't just about choosing technology and running with it. Embracing MLOps involves changing processes and organizational cultures, and doing so in a way that aligns with and supports the organization's overall vision.

Deloitte has helped companies in several industries tailor MLOps to their unique needs. We've worked with these companies to implement the process and cultural changes needed to test, iterate, and refine machine learning models more effectively—and put them into production.

In this report, we explore some of the challenges that organizations face in putting machine learning models into production—and we share lessons learned and insights into best practices that can help you and your team successfully implement MLOps in your organization.

MLOps defined

MLOps is a relatively recent concept that has arisen alongside organizations' growing embrace of artificial intelligence (AI), machine learning (ML), and analytics. It combines the machine learning processes with DevOps workflow and Agile methodology.

DevOps emerged in the late 2000s as a way to improve and accelerate software development by using iterative, continuous, and collaborative development cycles based on Agile principles. New techniques and approaches were introduced to traditional software development, including self-service configuration, automated provisioning, continuous build and solution integration, automated release management, and incremental testing. By standardizing and automating much of the work of software development, deployment, and management, DevOps transformed how IT teams released and managed software—and enabled them to work more efficiently, deliver more quickly, and improve software quality overall. It didn't take long for DevOps to become one of the dominant approaches to software development.

Today, most ML development is where software development was in the 2000s—slow, inefficient, and frustratingly hard to get into production. DevOps has proven inadequate to the challenges of ML projects, because it was developed with software in mind. In the world of traditional software, code defines the outcome; with ML code and data define the outcome, and DevOps isn't set up to deal with data. As well, traditional software development doesn't involve the repeated experimentation of ML development, which introduces a host of different elements and challenges. Something new was needed to support ML development—and that's MLOps.

MLOps aims to do for ML what DevOps did for software: embrace standardization, automation, and collaboration to speed up the ML model development life cycle and make it easier and faster to put ML models into production at scale. In some ways, MLOps has an advantage over DevOps, in

that the ML lifecycle is intrinsically more agile than the software lifecycle, which should make its adoption and implementation smoother at most organizations.

With MLOps, IT and data science teams work in close collaboration to build, accelerate, and implement model development, refinement, deployment, monitoring, and more, harnessing best practices and working in an agile way. MLOps can help teams move ML models into production and—perhaps more importantly—ensure the models continue delivering value post-production, which is something that challenges many organizations. MLOps achieves this by facilitating continuous model experimentation, refinement, and deployment, supported by an active, efficient data pipeline available to data scientists, data engineers, and data modellers.

The benefits of MLOps

MLOps is the key to enabling organizations to realize the full potential of machine learning and use ML to make better, faster decisions.

- **Drives business value.** MLOps drives business value by fast-tracking the experimentation process and development pipeline, improving the quality of models moved into production and making it easier to scale production models.
- **Better collaboration among teams.** The continuous development, testing, deployment, monitoring, and retraining at the heart of MLOps brings IT and data science teams into constant contact with a shared focused on delivering better models, improving collaboration overall.
- **Faster progress to production.** MLOps can greatly shorten development lifecycles and facilitate faster, more reliable, and more efficient model deployment, operations, and maintenance.
- **More accurate models.** MLOps enables IT and data science teams to continuously monitor the performance of ML models and overcome the issue of 'model drift'—a tendency for predictive models to grow less accurate over time. MLOps' standardized processes allow teams to constantly realign ML models as the business and customer data used by the model expands and evolves.

MLOps helps grocery chain deliver the personalized content and experience customers want

A North American grocery chain wanted to improve the functionality of the mobile and web apps that supported its sizeable loyalty program, by providing more user-centric content such as recipes and food-related articles. Unfortunately, the company's initial attempt failed to deliver. The new content wasn't tailored to users, and customers complained that the apps' new content was irrelevant—sometimes wildly so, such as when vegan customers received beef recipes.

The company wanted to improve the recommendation engine to better deliver personalized, customized content and advice by drawing on a vast array of shopper and transaction data (e.g., customers' browsing history, past purchases, likely purchases, banner affinity), with business rules layered on top.

The challenge

The principal challenge was size, as millions of Canadians were members of this loyalty program and users of the program apps. The resulting web traffic was astonishingly high, with 10,000 requests per second. The models used for the recommendation engine didn't just need to be very accurate and reliable—they needed to be very fast and error-free. Responses needed to be delivered to customers in milliseconds. And new models had to be deployed seamlessly, with no impact on the user experience.

The solution

Given the need for scale and speed, the grocer deployed an MLOps approach that was anchored to a customer-centric perspective and focused on production first, starting with a simple model and iterating from there. The models were designed to measure and improve the customer experience, and key performance indicators (KPIs) focused on customers. The MLOps pipeline allowed models to be iterated, enhanced, and replaced quickly and reliably. A/B testing enabled the team to rapidly test new models and identify their impact, and extensive monitoring ensure acceptable performance was maintained. Following the implementation, customers' negative feedback subsided—and the grocer has a platform on which to build future enhancements.

Common challenges in implementing MLOps

MLOps can make a real and positive impact on organizations' efforts to get ML models into production and delivering value. However, implementing MLOps can be challenging—and our experience in helping clients to adopt MLOps has given us insights into why some organizations stumble along their journey.

Implementing MLOps in isolation

Working in silos never works. And when different teams within an organization work on MLOps solutions in isolation from one another, both they and the organization can fail to realize the significant benefits of a collaborative, integrated solution.

Taking a siloed approach to MLOps can result in many inefficiencies. In large organizations especially, implementing MLOps in isolation from the rest of the enterprise can falter because the teams involved are using different versions of the same tools and technologies—or different tools and technologies entirely. This can greatly complicate efforts to develop and productionize ML models. Teams and individuals can find themselves in working 'technology silos,' unable to collaborate with colleagues effectively. System incompatibilities interfere with efforts share data. Inefficiencies proliferate, and costs rise.

As well, taking an isolated approach to MLOps can result in disconnects between the MLOps implementation and the organization's overall business strategy and priorities, as teams move forward without consensus on the end goal for the project. It can impede data governance and create obstacles to sharing data and insights. It can lead to duplication of effort and competing solutions to the same problems, wasting team's time and effort and it can make it even more difficult to progress ML models beyond the proof-of-concept stage.

MLOps skill gaps

MLOps implementation is truly a team effort. The skillsets needed to develop an effective ML model and to operationalize and maintain that model are quite different—and it's rare to find one person who can do it all.

Data scientists' skillsets are focused on mathematics, probability, statistics, and a variety of modelling techniques. When an organization needs a model built, it turns to its data scientists. But it takes very different skillsets to operationalize that model, converting it into a form where it can be put into a production environment and enable users to interact with it to obtain insights, make decisions, and generate business value. To do that, organizations need people with software engineering and DevOps skillsets—people who understand programming, networks, and IT architecture, and who can determine whether a model is practical, viable, and capable of delivering value once put into production.

Organizations that attempt to develop and productionize ML models by relying on a single person or a small, lean team to be data scientists, data engineers, and software engineers at the same time soon run into difficulty. Team members become overstretched, and the gaps in their knowledge and skillsets increase the risk of error and slow progress. Developing an AI talent strategy that includes recruiting and retaining individuals with MLOps knowledge and skills can be essential to the successful implementation of MLOps overall.

Staying aligned with evolving regulatory requirements

As the use of AI and machine learning have become increasingly widespread, governments and regulatory bodies around the world have responded with rules and legislations designed to protect individuals' privacy and personal data. As technologies evolve and organizations discover new uses for individuals' data, these data and privacy laws are continuously evolving—and can vary depending on a variety of factors, from industry to jurisdiction to operational area.

In addition, the rising use of AI has led to a growing recognition that AI must be used responsibly, and that personal data needs to be collected, shared, and used in a fair and ethical manner. Data ethics, or "responsible AI," calls on organizations and their AI teams to consider the human impact of collecting, sharing, and using data, whether that data is sourced internally or externally from partners, open sources, or third-party vendors. It means providing individuals with a say in what data they share and how it can be used—and respecting and meeting their wishes. Data ethics involves grounding data-related decisions in the organization's brand values and a clear understanding of the potential financial and reputational impact of a data misstep. And it means understanding and complying with applicable regulatory requirements for data collection, storage, and usage.

Ensuring ML models reflect and align with applicable regulations, and that data is being used in fair and ethical ways, can be challenging, requiring a tailored approach. Not aligning models with applicable regulations and ensuring data is being used ethically isn't really an option, however, as it leaves organizations open to financial, operational, and reputational risks related to potential misuse of data.

Pharmaceutical company leverages vendor MLOps platform and framework to scale and build sustainable Machine Learning solutions

A Canadian pharmaceutical company had launched a data-driven business transformation initiative. The goal of the project was to enable to organization to move away from its traditional, physician-centric model and become a patient-centric, insight-driven organization that leveraged data insights to make business decisions.

Accomplishing this goal required the organization to have a much more comprehensive view of the patient, however. The company engaged Deloitte to help them achieve this. The Deloitte team would help the company leverage its existing data and available third-party data and establish partnerships with other organizations to access additional data. The data from these various sources would then be pulled together, and Machine Learning Models would be developed to support decision-making.

The challenge

The team began the work using the company's existing IT environment and infrastructure, but a challenge soon arose. The company's people, processes, and technology were more than sufficient for day-to-day reporting and business intelligence needs; however, they weren't set up in an agile way that could support machine learning development work. The existing IT infrastructure wasn't able to support rapid ingestion of third-party data or speedy integration of internal and external data sets. That meant that it could take weeks or even months to stand up the environment needed for a machine learning use case—and there were multiple use cases.

The solution

To expedite the process and deliver speed to value, the company focused on creating its own the machine learning environment and establishing standardization to technology and tools.

MLOps principles was used to expedite the end-to-end development and deployment lifecycle of Machine Learning solutions in an agile way. Data from multiple sources—comprising several hundred data tables and tens of thousands of data columns and fields—were ingested, integrated, and cleansed quickly. Within the first year of the initiative, 10 machine learning models were developed and deployed into three front-end visual applications, along with an end-to-end data engineering pipeline.

Manulife Financial Corporation uses MLOps to standardize analytics processes enterprise-wide

A global insurer headquartered in North America with operations and customers around the world wanted to improve the speed and scale of its AI/ML delivery, improve cross-functional collaboration, and better link AI/ML solutions to business outcomes.

The challenge
Not enough of insurer's analytical solutions made it to production—and those that did required significant resources to get to that point. Traditional approaches clearly weren't working.

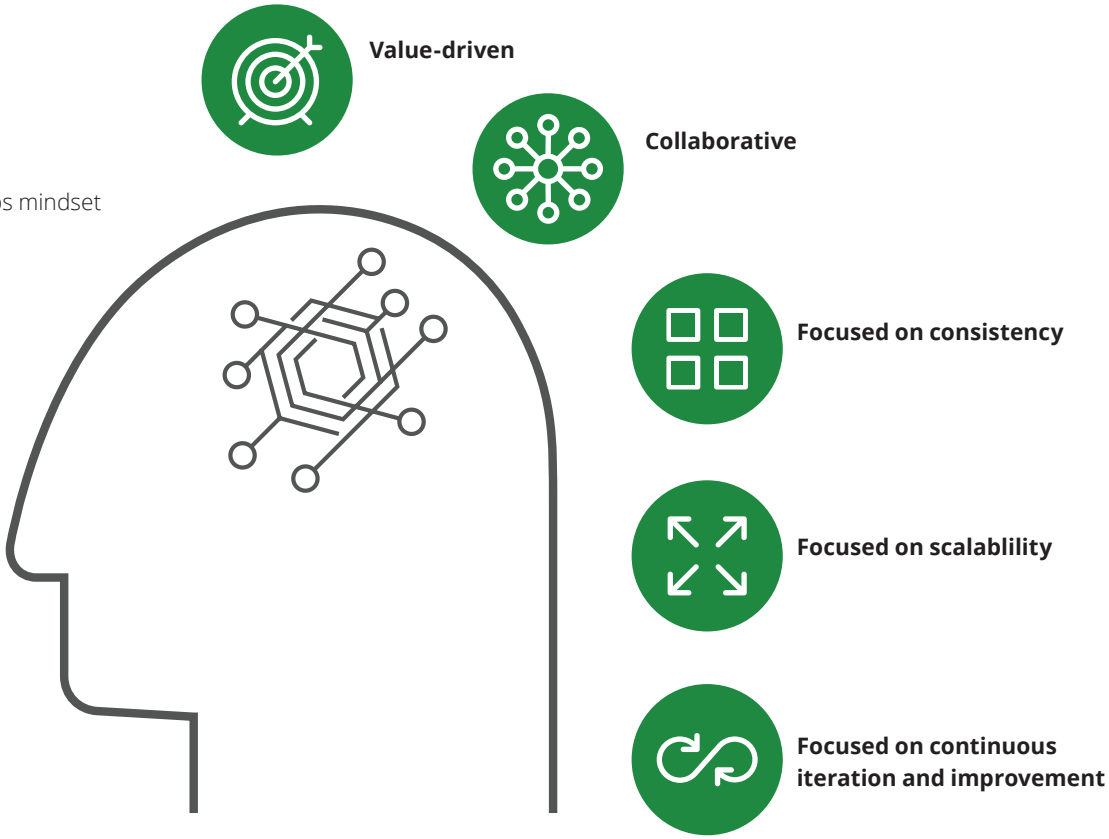
The solution
To industrialize and professionalize AI and advanced analytics across the enterprise and significantly accelerate model deployment relative to industry, the global insurer created an enterprise framework including tactical guidelines for incorporating MLOps into the AI/ML development and deployment lifecycle.

Under the umbrella of this framework, key actions, templates and leading practices for each stage of the AI/ML development lifecycle were identified and documented to communicate leading practices for developing and delivering insights through multi-disciplinary and cross-functional teams. To activate and support the deployment of this framework, KPIs were also developed to ensure AI initiatives were clearly linked to business objectives and MLOps capabilities.

A framework or process on its own rarely delivers results. To facilitate the adoption of MLOps framework and accelerate change in the organization, a detailed adoption strategy and roadmap was created by the insurer. That 'people/ framework implementation' phase included the creation of an MLOps coaching community, training for all AI/ML team members (hundreds of data scientists and engineers), further refinement of the strategic and operational KPIs, and the creation of several governance mechanisms including a minimum viable canvas to hold each AI/ML team member accountable for adopting at least one change to bring the standardized MLOps leading practices to life.

In addition to benefits such as speed to market, throughput and efficient operations, the adoption and implementation of the framework has sparked a cultural shift across data science practitioners and stakeholder groups toward becoming a leading employer in the space.

Figure 1: The MLOps mindset



Best practices for implementing MLOps successfully

MLOps can be a game-changer in enabling organizations to move ML models past the experimentation and proof-of-concept stage and establish a strong capability to develop new models and put them into production at scale. But as we've seen, it can be challenging for some organizations to capitalize on the potential benefits of MLOps.

In recent years, Deloitte has built its own internal MLOps capability and helped our clients in a variety of industries successfully implement MLOps. Our experience and the lessons learned along the way have given us insights into a series of best practices that can help other companies bring MLOps to life in their own organizations.

Make sure you have the right people

MLOps is definitely a team sport. Organizations need data scientists who can build effective ML models—and developers and ML engineers who can put those models into production and maintain them. If key skillsets are lacking, it can take far longer to move models into production—and once they are there, those models may perform poorly, creating additional problems and costs.

Having the right combination of knowledge and skillsets is one thing. It's also very important that the team members are able to communicate and collaborate effectively in order to execute MLOps successfully. Data scientists, data engineers, ML engineers, domain experts, and DevOps specialists need to work side-by-side in order to build, test, evaluate, iterate, and refine ML models, speed them into production, and keep them operating at peak performance.

Some organizations may be challenged to find the people they need to fully staff their MLOps team, whether because of budget constraints or a tight talent market. Cross-training team members in missing skillsets can help close some skills gaps, but it should be regarded as a short-term measure at most. Alternatively, organizations may consider working with external, third-party MLOps specialists, who can deploy proven solutions to help internal MLOps teams accelerate progress through collaboration, co-delivery, and knowledge transfer.

Instill a shared MLOps vision and mindset

Organizations shouldn't undertake MLOps in a vacuum. There should be a clear purpose for implementing MLOps that aligns with the organization's business strategy and priorities—a shared vision and goal that shapes all MLOps activities.

It's also important to develop a shared mindset that instils the "MLOps way" throughout the organization. The MLOps mindset is characterized by being value-driven, collaborative, and focused on consistency, scalability, continuous iteration and improvement.

Export credit agency accelerates digital transformation with MLOps

An export credit agency was looking to establish an MLOps practice for developing, deploying, and managing advanced analytics solutions.

The challenge

The agency is in the midst of a digital transformation and on a journey to becoming a "customer-obsessed" organization, and it recognizes that a critical part of this journey is developing a robust data and advanced analytics environment, with MLOps enabled across people, processes, tools and technologies. The organization wanted to establish an MLOps practice that meets current and future demands for building advanced analytics solutions, solving existing business problems, and unlocking new opportunities.

The solution

To develop a robust data and advanced analytics environment that would support the agency's ambitions, the engagement focus was to create both high-level and detailed solution architectures across key domains, including application, data, infrastructure, and security. The agency formulated a strategy to implement and operationalize MLOps principles by designing and building the relevant platform along with the communications and change plans.

Once implemented, MLOps enabled the organization to accelerate discovery of enterprise data and drive informed alignment among diverse stakeholders. In addition, the project created iterative business value throughout every initiative, which helped to facilitate the agency's large-scale digital transformation.

Deploy clear, agile processes

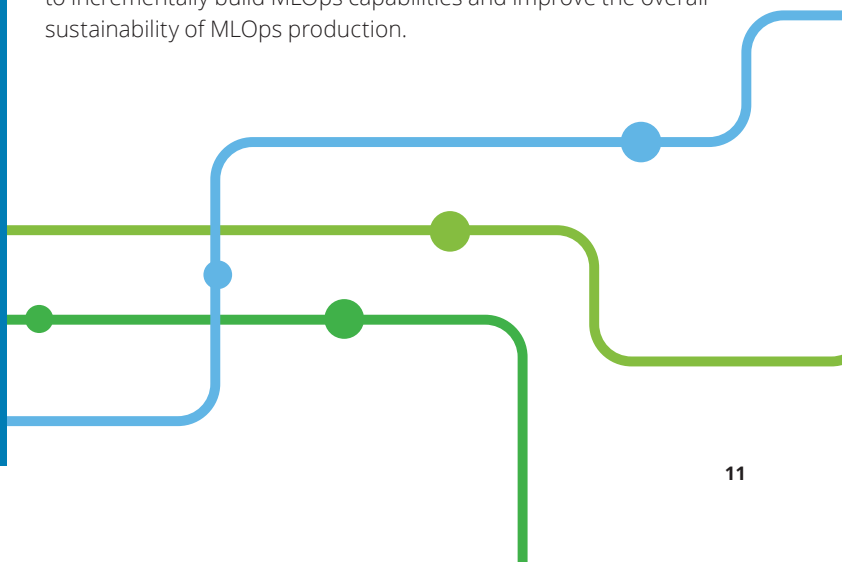
For MLOps to deliver on its potential, it's vital that organizations use well-defined agile processes to minimize obstacles, maintain project momentum, and continuously strive towards delivering the best possible product—at speed.

Agile processes deliver significant benefits to any MLOps implementation. Continuous monitoring and improvement helps ensure high quality ML products and avoids the perils of model drift. Agile makes it easier to involve stakeholders and other key people at the right time of the model development cycle, and feedback can be rapidly incorporated and re-evaluated thanks to the rapid, iterative nature of the agile methodology. This encourages collaboration, communication, and the exchange of ideas, and accelerates model improvement overall while reducing the likelihood of costly delays. Finally, the robust, relevant metrics characteristic of agile processes make it easier for MLOps teams and their organizations to gauge project performance and control overall project costs.

Standardize your technology tools

It's important that organizations ensure their MLOps teams are aligned to a shared vision and goal for MLOps projects and embody a shared mindset around how the work gets done. It's also ideal for organizations to ensure—to a realistically possible extent—that the technology tools used across the organization is consistent, standardized, and uniform. Where this isn't possible, then it's important to understand where different tools are used, why, and what's required to ensure compatibility.

From an MLOps perspective, a common set of technology tools makes it easier for the team to communicate and collaborate with each other, stakeholders, and end users. It makes it easier to tailor delivery frameworks and processes for the organization's current (and future) state. It maximizes opportunities to build reusable data assets, code pipelines, MLOps artifacts, and use cases to reduce the time and cost involved in developing models and facilitating scaling up of production models. It enables organizations to better identify opportunities to automate certain processes, and it allows them to incrementally build MLOps capabilities and improve the overall sustainability of MLOps production.



Make MLOps work for you

MLOps can enable your organization to harness the power of machine learning to stay ahead of the curve in business environment that's constantly changing, sometimes in surprising, unexpected ways.

But as many organizations have discovered, MLOps can be challenging to implement and execute successfully. Deloitte has learned about making MLOps work, both by implementing ourselves and by helping our clients do the same thing. The best practices we've outlined in this report should be seen as a springboard to help your organization re-examine its MLOps efforts and make key changes to get your efforts on track and delivering meaningful business value.

If you'd like to discuss how to bring MLOps to your organization, or how to be more successful at using MLOps to produce models at scale, we're happy to help. Please contact your local Deloitte professionals, or one of the individuals listed here.

Contacts

Ian Scott
Partner, Chief Data Scientist
Deloitte Canada

Kevin Laven
Partner, Data Science
Deloitte Canada

Audrey Ancion
Leader, AI Institute
Deloitte Canada

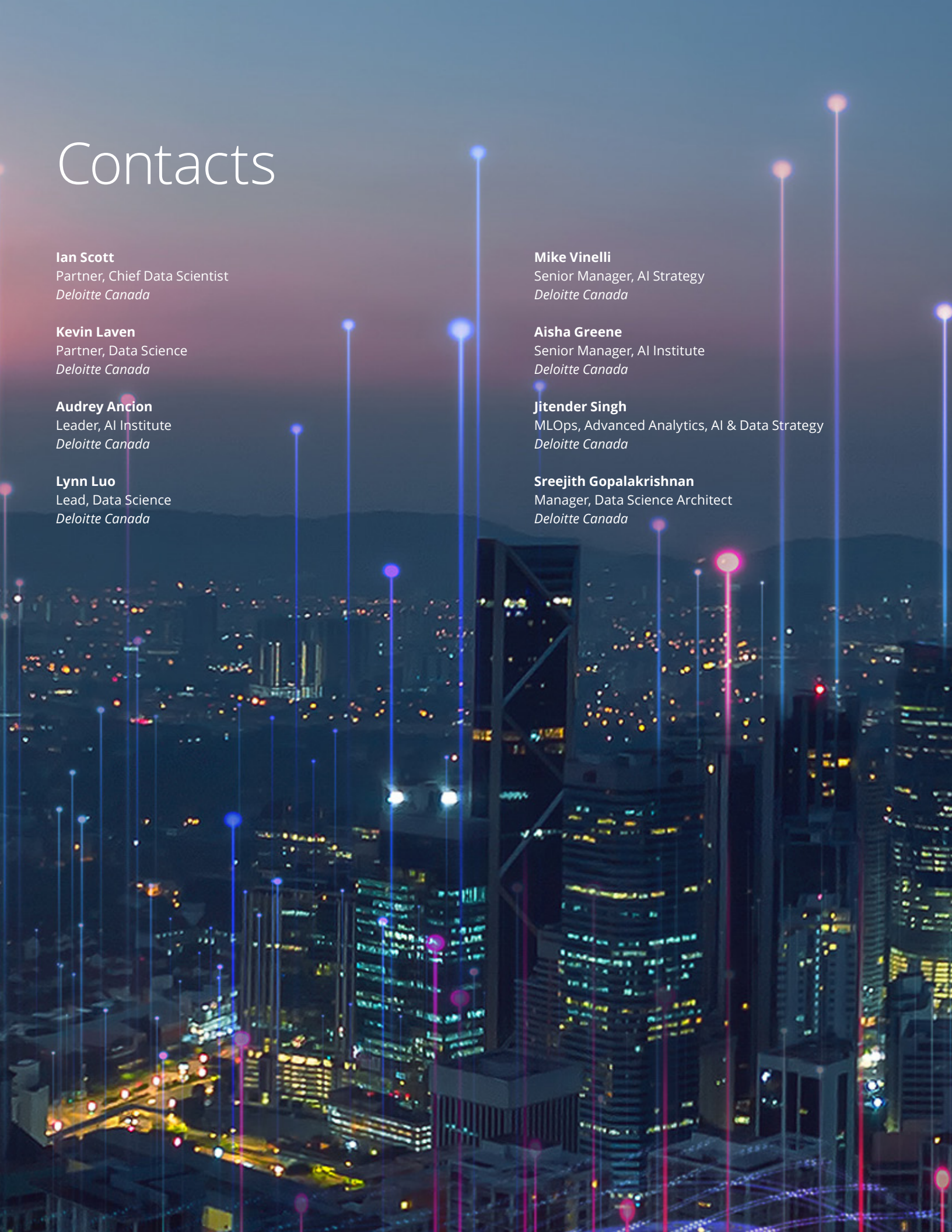
Lynn Luo
Lead, Data Science
Deloitte Canada

Mike Vinelli
Senior Manager, AI Strategy
Deloitte Canada

Aisha Greene
Senior Manager, AI Institute
Deloitte Canada

Jitender Singh
MLOps, Advanced Analytics, AI & Data Strategy
Deloitte Canada

Sreejith Gopalakrishnan
Manager, Data Science Architect
Deloitte Canada



Deloitte.

www.deloitte.ca