MIT SMR CONNECTIONS

MANAGER'S GUIDE

What Leaders Must Know About Data for Machine Learning

ON BEHALF OF: aws machine learning

CONTENTS

What Leaders Must Know About Data to Drive Success With Machine Learning **2**

1. Align machine learning initiatives with business priorities.

2. Create and maintain a comprehensive view of all data assets.

3. Lay the groundwork for data governance.

4. Identify the specific roles required to build a strong data foundation for machine learning.

Data Management Strategy Checklist **5**

Sponsor's Viewpoint: Your Data Strategy Is Key to Machine Learning; a Data Lake Can Help **6**

MIT SMR Connections develops content in collaboration with our sponsors. It operates independently of the *MIT Sloan Management Review* editorial group.

What Leaders Must Know About Data to Drive Success With Machine Learning

Machine learning is taking predictive analytics to the next level to drive tangible business value for a wide array of industries. Algorithms allow credit card companies to detect fraud in real time and help retailers direct offers to the customers most likely to respond. In health care, tools powered by machine learning help doctors transcribe notes more easily so they can focus on patient care. Manufacturers can take in data from sensors on plant equipment and recommend maintenance before malfunctions cause production delays.

But machine learning models are only as good as the data they ingest. "If data is not clean, if it's not accessible, if it isn't stitched together to form a strong foundation, the machine learning and artificial intelligence capabilities built on top of it will have problems," warns Ashok Srivastava, senior vice president and chief data officer at financial software provider Intuit. This can lead to difficulties such as inaccurate insights or inherent bias — factors that can hamper intelligent business decision-making.

Fortunately, businesses can avoid these perils by designing a data management strategy that develops new capabilities, initiatives, and roles around machine learning. This guide aims to share lessons from business leaders and industry experts on how, with the right policies and frameworks in place, data can serve as a strategic corporate asset.

1. Align machine learning initiatives with business priorities.
The first step in creating an enterprise data management strategy is understanding the business's goal for machine learning. For example, Intuit's machine learning initiatives aim to improve customer service by providing personalized recommendations to subscribers of its accounting and tax software programs. An online retailer may plan to use machine learning to create more-effective targeted marketing campaigns, while an automotive manufacturer may be building machine learning systems to predict equipment failures.

Establishing which of a business's strategic priorities have the best potential to be advanced via machine learning provides clarity around which data sets are most important to collect, store, and prepare for analysis.

"Being focused on knowing what data is truly driving your business and matters most is the first piece to a data strategy," says Juan Tello, chief data officer at Deloitte Consulting and principal in its Strategy & Analytics practice. "So, for example, if business priorities are to win more customers and provide more-competitive pricing based on the products a company sells, that requires three critical data domains: customer data, pricing data, and product data. Prioritizing the data strategy on those areas as a starting point will maximize business outcomes. Organizations should also reevaluate and adjust as their business priorities change."

This focus is essential, given the vast volumes of data generated by enterprise applications, connected devices, and customer interactions via the web or social media platforms, to name just a few sources. However, by narrowing the scope for data management to three or four key sources, businesses can focus on those data sets that will deliver the most value.

At Intuit, data management experts meet with the teams that own data to build a catalog of that information, resulting in a robust list of data assets within the company.

2. Create and maintain a comprehensive view of all data assets.
For data to be useful, a business must know it exists. Unfortunately, legacy systems, mergers and acquisitions, and poor data onboarding practices can create silos of unidentified and untagged information.

At Intuit, data management experts "meet with the teams that own data systems or data pipelines, and we start to build a catalog of that information. That means understanding what data they have and how it is stored." The result, says Srivastava, is "a robust list of data assets that we have within the company."

But data troves are constantly evolving as businesses deploy new systems. GE Healthcare offers a perfect example of how to stay ahead of the curve. The manufacturer of diagnostic imaging equipment, which uses machine learning algorithms to improve traditional imaging technologies like CT scanning and X-ray, continuously works with collaborators and partners to inventory and onboard de-identified data. A dedicated team of data specialists receives, processes, and properly catalogs contractually de-identified data sets and then uploads them for use in AI development. This process leads to greater data transparency and availability.

Business leaders must also be held accountable for maintaining a comprehensive view of data assets. At GE Healthcare, chief data officer Derek Danois says, broad communication and transparency are key to building trust: Business units now collaborate so that the company knows the moment a new data set becomes available.

3. Lay the groundwork for data governance.
At the core of every data management strategy is data governance — a set of rules and systems that ensures that data is secure, handled in compliance with applicable regulations, accessible, and useable.

Data security and compliance with privacy laws are table stakes and as such have been the primary drivers of data governance for most enterprises. In addition to guarding against intruders via cybersecurity measures that protect the IT perimeter, businesses must also establish controls that limit how data is accessed, used, and managed by employees. This typically means granting different access levels depending on variables such as role, tenure, and function. Compliance with regulations such as the European Union's GDPR (General Data Protection Regulation) and similar requirements in other jurisdictions means that companies must also be prepared to explain to consumers how their data is being used to make decisions that affect them.

Another key component of data governance is quality: A machine learning model's output depends on the quality of its training data.

At GE Healthcare, for instance, a team of data architects and data scientists evaluates data quality based on a variety of metrics. A medical imaging study might be vetted for standard-of-care parameters (such as slice thickness or scan geometry), field of view (the area of a scanned object), and metadata content requirements. If quality standards are met, GE Healthcare de-identifies or anonymizes the data and establishes a chain of custody that chronicles the data's control, transfer, and analysis, before it's uploaded for use in AI development.

Maintaining consistently high levels of data quality calls for continuous monitoring of metrics and key performance indicators such as accuracy, timeliness, consistency, and integrity — a process that can become overwhelming, according to Tello. Using AI-powered data quality tools can accelerate the ability to manage and govern data, he says. Enterprise master data management software can also ease the burden by creating a single master reference source for all critical business data, thereby reducing redundancies and the likelihood of errors.

4. Identify the specific roles required to build a strong data foundation for machine learning.

An explosion of new data science job titles has raised questions regarding who is responsible for which tasks within a machine learning practice. A well-thought-out organizational structure can make sense of this landscape by clarifying roles and delineating responsibilities.

> **"The business owners who are making decisions on a daily basis are some of the most important contributors to our overall data strategy"**
>
> ASHOK SRIVASTAVA, INTUIT

According to Peter Nichol, director of IT portfolio management for research and development at Regeneron Pharmaceuticals, some of the key roles required to execute a data management strategy include the following:

- **Chief digital/data officer:** Oversees all digital functions, provides support and leadership, and articulates a strategy for data governance that's consistent across the company.
- **Data scientist:** Creates tools or processes based on machine learning and applies them to well-defined business problems.
- **Decision scientist:** Uses expertise in technology, math, and statistics, along with business domain knowledge, to enable informed decision-making.
- **Compliance/legal team member:** Handles privacy, compliance, data rights, and regulatory aspects impacting a business.

Ancillary positions include data management specialist, business intelligence specialist, and data architect.

But there's also a place for sales executives, HR managers, and chief marketing officers in machine learning initiatives. "The business owners who are making decisions on a daily basis are some of the most important contributors to our overall data strategy," says Intuit's Srivastava.

That's because business leaders possess domain knowledge — an in-depth understanding of the relevant data within the enterprise, the processes that generate useful data, what data might be useful for a model, and how different variables might impact a model's output. Without this guidance, businesses risk creating machine learning applications that don't deliver useful results.

Looking Forward

Machine learning has the potential to improve results in nearly every aspect of business. But to harness it, businesses need a data management strategy that will continuously improve the quality, integrity, access, and security of data. ●

DATA MANAGEMENT STRATEGY CHECKLIST

Keep the following practices in mind to successfully design and execute a data management strategy in support of machine learning:

[✓] Establish rules and processes around how data is sourced, managed, accessed, and used across the business.

[✓] Ascertain which data sets are driving the business and how they can be used to help solve problems, generate revenue, and deliver customer benefits.

[✓] Inventory known data assets, classify them, and organize them in a data catalog.

[✓] Meet with the teams that own and operate data systems to better understand what data they have and how it is stored.

[✓] Understand where your data comes from, who has access, and how it can be used.

[✓] Establish internal security precautions (such as provisioning user access), as well as external safeguards (such as anonymizing data), to protect sensitive data.

[✓] Create access controls that set limitations around how data is accessed and how it might be used.

[✓] Design processes and systems to ensure that data created is accurate and useful.

[✓] Identify specific roles required to build a strong data foundation, including chief digital officer, data scientist, decision scientist, and compliance team member.

SPONSOR'S VIEWPOINT

Your Data Strategy Is Key to Machine Learning; a Data Lake Can Help

Machine learning success is highly dependent on having relevant and high-quality data. Without a proper data strategy in place, machine learning initiatives fail to scale. Worse yet, if the machine learning models are informed by bad data, the results they generate may be misleading — or even incorrect.

The right data strategy for machine learning should aim to break down silos, enabling your IT teams to easily, quickly, and securely access and collect the data they need. While modern data strategies take many forms, data lakes are becoming an increasingly popular core component of the most efficient models. Data lakes offer more agility and flexibility than traditional data management systems, allowing organizations to manage multiple data types from a wide variety of sources and to store the data — whether structured or unstructured — in a centralized repository. Once stored, the data can be leveraged by many types of analytics and machine learning services faster and more efficiently than with traditional, siloed approaches. Data lake architectures also enable multiple groups within the organization to benefit from analyzing a consistent pool of data that spans the entire business. For help developing a more holistic data strategy that includes data lakes, interact with the AWS Data Flywheel.

Amazon's ML Solutions Lab program can also help you build the right data strategy. The Amazon ML Solutions Lab pairs your team with Amazon machine learning experts to prepare data, build and train models, and put models into production. It combines hands-on educational workshops with brainstorming sessions and advisory professional services to help you essentially work backward from business challenges and then go step-by-step through the process of developing solutions based on machine learning. Moreover, one of our machine learning partners can also help you build the right data strategy for your machine learning initiatives. AWS Machine Learning Competency Partners have demonstrated relevant expertise and offer a range of services and technologies to help you create intelligent solutions for your business, from enabling data science workflows to enhancing applications with AI services. Learn more at aws.ai.

About Amazon Web Services

AWS offers the broadest and deepest set of machine learning and AI services. On behalf of our customers, we are focused on solving some of the toughest challenges that hold back machine learning from being in the hands of every developer. Tens of thousands of customers are already using AWS for their machine learning efforts. You can choose from fully managed AI services for computer vision, language, recommendations, forecasting, fraud detection, and search; or Amazon SageMaker to quickly build, train, and deploy machine learning models at scale. SageMaker Studio offers the first fully integrated development environment for machine learning. You can also build custom models with support for all of the popular open-source frameworks. Our capabilities are built on the most comprehensive cloud platform, optimized for machine learning with high-performance computing and no compromises on security and analytics. Learn more at aws.ai.