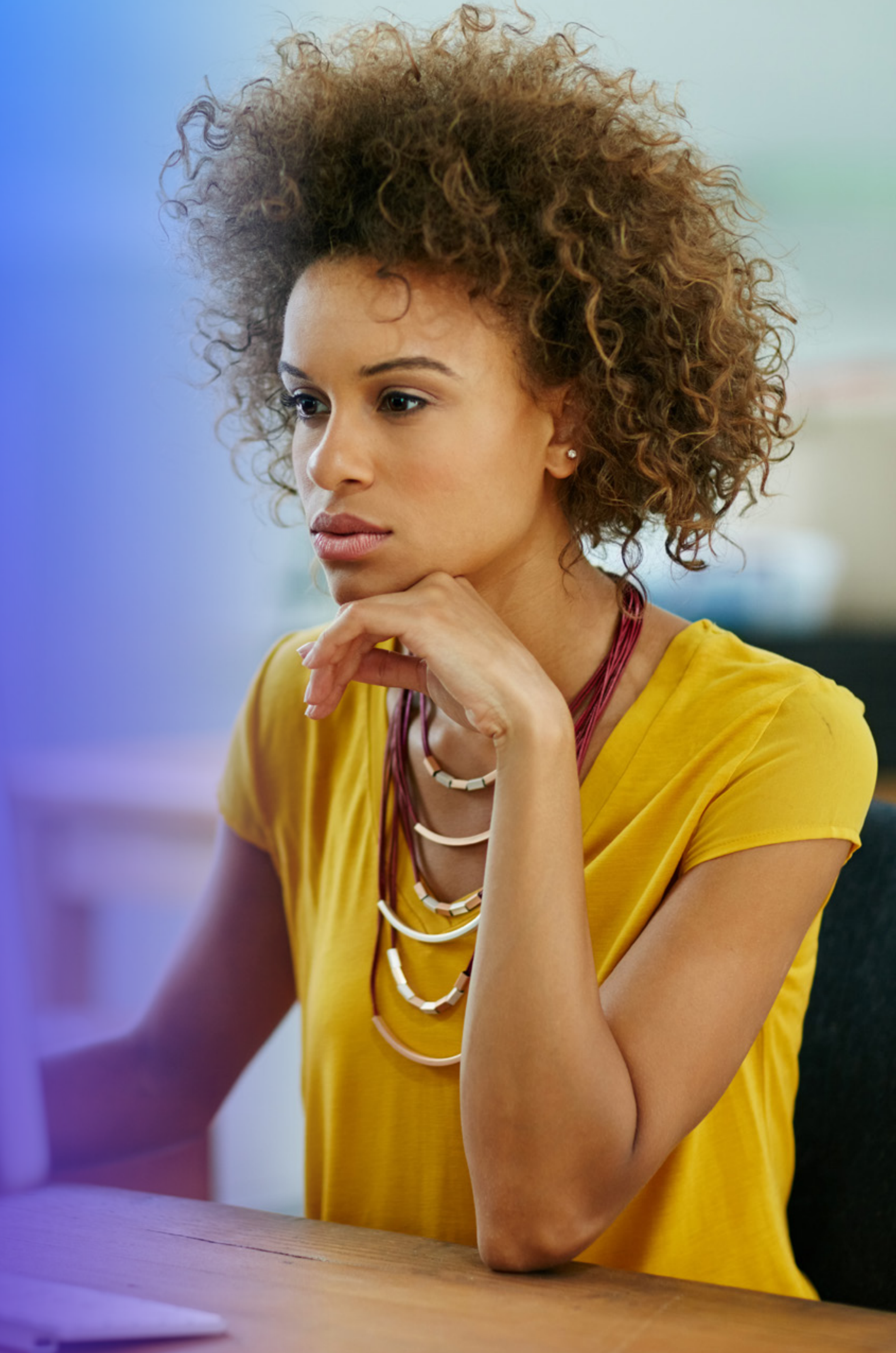


Machine learning at scale

High-performance, low-cost
machine learning for any use case

INTRODUCTION

Solve for machine learning scalability

Machine learning (ML) has emerged as a core technology ingredient for organizations to drive innovation. Today, more than a hundred thousand organizations are leveraging artificial intelligence (AI) and ML solutions and services from Amazon Web Services (AWS) to achieve substantial business results. These businesses span virtually every industry, including financial services, healthcare, media, professional sports, retail, and the industrial sector.

The relevance and impact of ML are expected to accelerate. According to IDC, by 2025, global spending on AI will reach $204 billion.[1]

[1] "Investment in Artificial Intelligence Solutions Will Accelerate as Businesses Seek Insights, Efficiency, and Innovation, According to a New IDC Spending Guide," IDC, 2021

Amidst the successes and growth, however, challenges to widespread ML adoption persist. Many organizations—enticed by the multitude of potential benefits—have grown frustrated by slow progress and a lack of return on their ML investments. For these organizations to reach their goals, they must find ways to put models into production faster and at a lower cost, ultimately scaling the technology to produce results across the entire business.

In this eBook, we'll explore the major barriers to ML scalability and success. Then we'll demonstrate how solutions and services from AWS can help virtually any organization overcome those challenges—and leverage ML to drive innovation and achieve tangible business results.

Top 5 barriers to achieving machine learning results at scale:

1 Data processing

2 Acquiring data science skills

3 Responsible use of machine learning

4 Expensive infrastructure

5 Lack of development tools and MLOps

Examine the barriers to machine learning success

For many organizations, ML has proven difficult to scale, leading to a lack of progress and frustration with the technology.

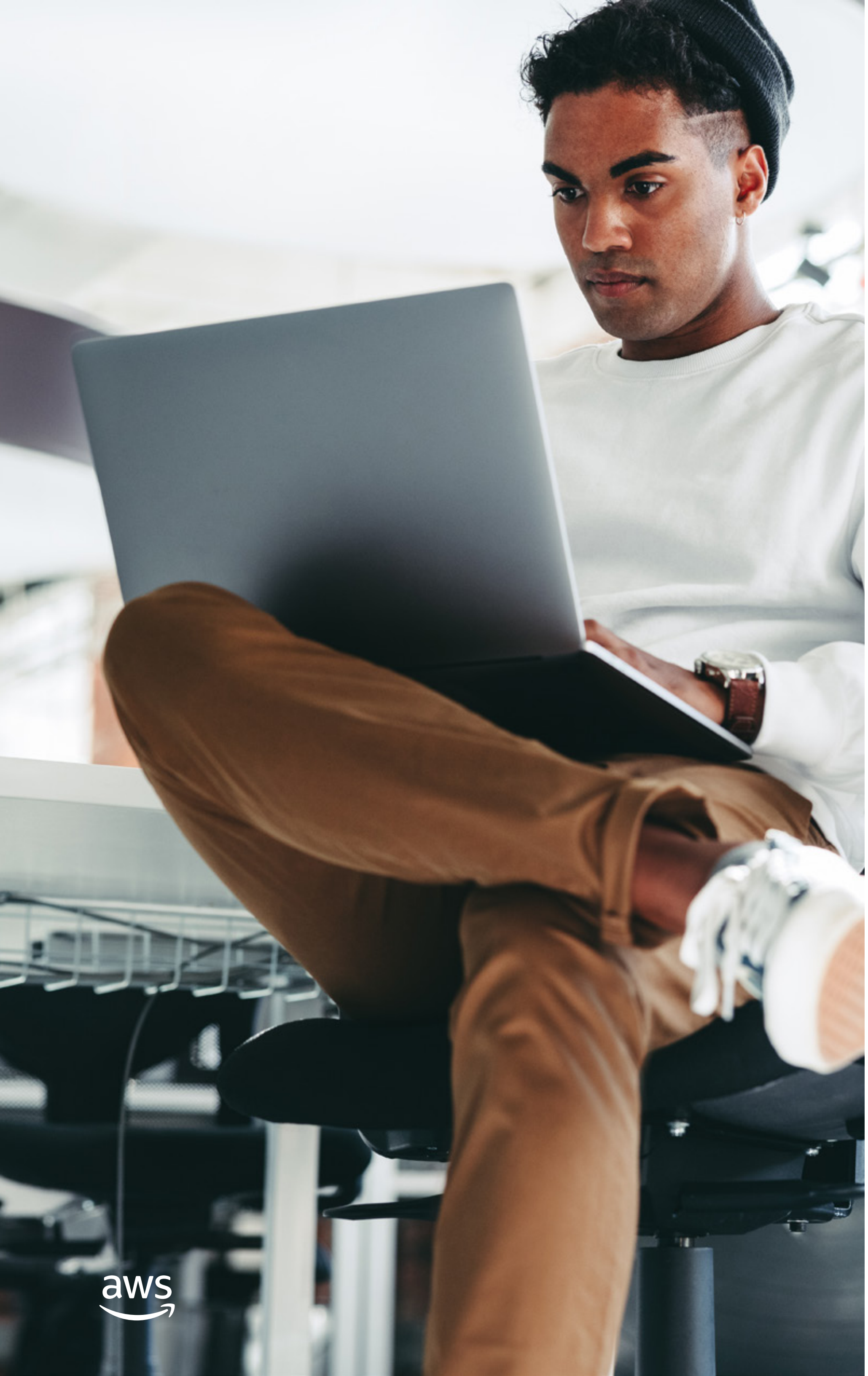
With the right services, solutions, tools, and processes, any organization can achieve success with ML and scale it across their business. But determining what those solutions are—and how best to implement them—starts with examining and understanding the barriers that must be overcome.

In that spirit, let's take a look at the five greatest challenges to driving widespread adoption and business results with ML.

1 Data processing

Data processing is very time-consuming, typically comprising about 80 percent of an ML project. Further, ML models are built on an enormous foundation of data from multiple modalities—tabular, text, audio, video, and others—which need to be managed differently. There are many disparate tools for processing structured data, and individual teams will have their own preferred approach. This makes it difficult for organizations to centralize their efforts into a single method for creating data pipelines.

Also, unstructured data must be properly cleaned and labeled before it can be made usable for ML. But setting up data labeling workflows, validating label quality, and managing labelers can be a time-, cost-, and resource-intensive process—especially when skilled ML developers and data scientists are hard to find.



2 Acquiring data science skills

Unfortunately, a shortage of data science professionals makes it difficult to acquire new talent or train existing talent for ML development. Without proper training or skilled data scientists to pick up the slack, developers often struggle to make effective use of ML. Providing developers access to pretrained models and fully managed solutions can help bridge this gap.

To truly scale ML, however, use of the technology must be expanded throughout the organization and made available for everyone, including business analysts and leaders with no discrete ML skills. Engaging more users with tools that fit their skill sets frees organizations from the burden of relying solely on highly skilled and hard-to-find data scientists, ultimately enabling greater ML scalability and results. But, if the organization isn't careful, it can also lead to the technology being used in irresponsible ways that may jeopardize business integrity.

3 Responsible use of machine learning

The explosion of datasets and models has caused rapid growth in the number of people interacting or building with ML. From data security and privacy to bias mitigation in data, ML comes with an inherent responsibility for all who harness its benefits. Putting practices in place to support the responsible use of ML within organizations is essential. Unfortunately, organizations already experiencing prohibitive factors, such as rising costs, may choose to forgo these efforts—and wind up compounding their issues when they ultimately face the consequences of irresponsible ML use.

4 Expensive infrastructure

With the increased use of ML comes more requirements for compute, storage, and networking. This can lead to burdens of time, cost, and resources—especially for organizations that choose to house and manage their ML infrastructure on premises. As organizations push the boundaries of ML complexity—creating models that use billions of parameters to make thousands of predictions—these problems can escalate exponentially if left unchecked.

Costs can be controlled by only procuring the amount of infrastructure that is needed for an organization's ML workloads. But this can prove difficult, as infrastructure requirements drastically change throughout the ML lifecycle. For example, moving ML workloads to production can account for up to 90 percent of the overall operational budget.

5 Lack of development tools and MLOps

Due to the relative newness and rapidly changing nature of ML, most organizations don't have standard processes for ML development. Most also lack an integrated set of securely connected ML tools—such as integrated development environments (IDEs), debuggers, profilers, and solutions for collaboration, workflows, and project management.

Instead, teams are forced to rely on disparate, disconnected tools for ML development. This makes it difficult to scale ML throughout the organization, as business analysts, developers, and data scientists will struggle to collaborate, deliver results at the speed the business demands, and involve non-technical teams in the process. By adopting ML operations (MLOps) processes and standardizing ML development, organizations can move faster and more efficiently toward achieving success with ML at scale.

5 AWS machine learning solutions

1. Simplified data processing
2. No-code/low-code solutions
3. Responsible machine learning
4. Flexible infrastructure
5. Development tools and MLOps

Achieve machine learning success with AWS

Now you can overcome ML challenges, accelerate your ML journey, and reach your business goals faster by using cloud services designed specifically for ML.

1 Simplified data processing

Amazon SageMaker provides a single service for business users and ML practitioners to access, label, and process large amounts of structured data (tabular data) and unstructured data (photo, video, and audio) for ML. You have the choice to use low-code data processing tools and Jupyter notebooks, or shop for datasets on **AWS Marketplace**.

2 No-code/low-code solutions

To open ML to a broad range of users, **Amazon SageMaker Canvas** offers a visual point-and-click interface to generate predictions. You can easily access data from the cloud and on-premises data sources and automatically generate predictions without having to write a single line of code. SageMaker also comes with hundreds of built-in algorithms and pre-built ML solutions that you can deploy with just a few clicks. Additionally, you can pick from 25+ API-based AI services for top ML use cases.

3 Responsible machine learning

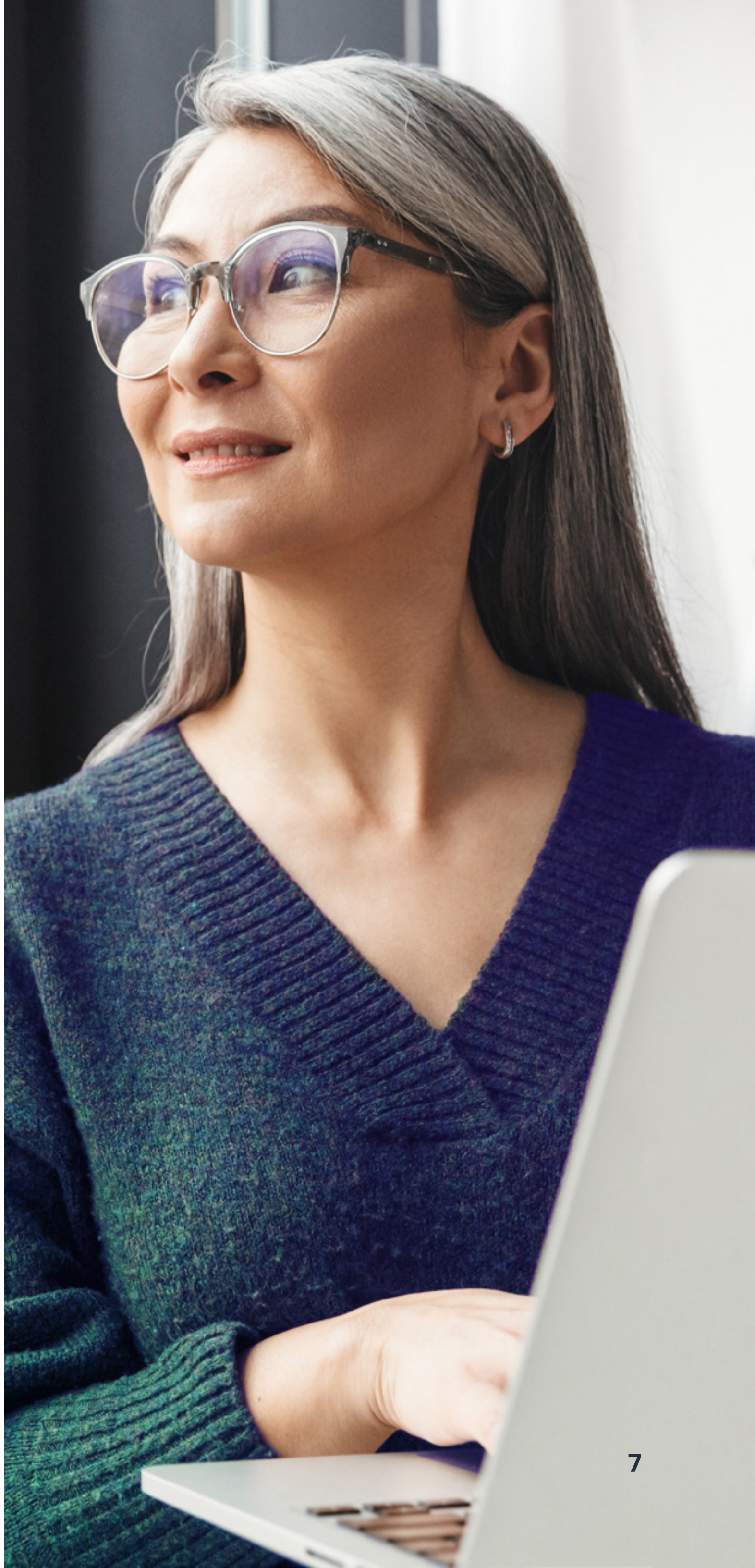
SageMaker provides bias detection, explainability, security, and governance features to help you support responsible use of ML and offers transparency to your business stakeholders and customers. SageMaker detects potential bias during data preparation, after model training, and in your deployed model and includes feature importance graphs that help you explain model predictions and produce reports for stakeholders.

4 Flexible infrastructure

SageMaker offers you the ideal combination of high-performance and low-cost infrastructure available in a fully managed service. For example, **AWS Trainium** is a custom ML chip designed by AWS specifically for training deep learning applications, such as image classification, semantic search, translation, voice recognition, natural language processing (NLP), and recommendation engines, to deliver the best performance for training in the cloud. **AWS Inferentia**, Amazon's first custom silicon, is designed to accelerate deep learning workloads and drive down the total cost of inference.

5 Development tools and MLOps

SageMaker offers an integrated set of ML tools to prepare data and build, train, and deploy ML models through **Amazon SageMaker Studio**. It also provides integrated capabilities for MLOps, which help your teams improve productivity. Purpose-built tools for MLOps help you automate and standardize processes across the ML lifecycle, so you can easily train, test, troubleshoot, deploy, and govern ML models at scale to produce models faster while maintaining performance in production.



Simplify machine learning at scale with SageMaker

To maintain focus on your core business objectives, avoid the struggle of building your own ML solution. Instead, offload the heavy lifting to **SageMaker**, which provides high-performance, cost-effective, and scalable ML capabilities to implement an ML environment across your entire business. Regardless of your organization's level of ML skills and experience, your teams can use SageMaker to prepare data and build, train, and deploy ML models for virtually any use case. With SageMaker, your organization can access a broad set of purpose-built ML capabilities under one unified visual user interface.

How does Amazon deliver packages so quickly?

Take a **virtual tour** of an Amazon Fulfillment Center to find out. Discover how Amazon uses a "symphony of machine learning" to help fulfill, sort, and deliver packages in record time.

The top 4 benefits you can achieve with SageMaker:

1. Enable a wider range of people to innovate with ML through a choice of tools—IDEs for data scientists and no-code interface for business analysts.
2. Access, label, and process large amounts of structured data (tabular data) and unstructured data (photo, video, and audio) for ML.
3. Reduce training time from hours to minutes with optimized infrastructure. Achieve up to 10 times better team productivity with purpose-built tools.
4. Automate and standardize MLOps practices across your organization to build, train, deploy, and manage models at scale.

INTUIT

Intuit empowers smarter financial decisions with machine learning

Intuit began its ML journey with just one model that empowered its customers to get the most out of their tax deductions.

Since then, ML models have become a core part of Intuit's business, and the company has seen a massive expansion of the number of ML models it uses—from fraud detection to customer service, personalization, and development of new product features.

In 2020 alone, Intuit increased the number of models deployed across its platform by over 50 percent. Intuit turned to Amazon SageMaker to develop and deploy hundreds of models at scale. Using SageMaker, Intuit modernized its ML platform, saved tax filers over 25,000 hours by utilizing self-help tools and cutting expert review time in half, and improved customer confidence.

Watch the video ›

Process machine learning data at scale

SageMaker helps with both structured and unstructured data processing. Your ML practitioners can prepare data in fully managed Jupyter notebooks, where they can visually browse, discover, and connect to Apache Spark data processing environments running on **Amazon EMR**. They can also interactively query, explore, and visualize data. And they can run Spark jobs to build end-to-end data preparation and ML workflows.

You can also use **Amazon SageMaker Data Wrangler** to prepare structured data with a no-code visual interface. SageMaker Data Wrangler contains over 300 built-in data transformations, so you can quickly normalize, transform, and combine features without having to write any code.

Additionally, **Amazon SageMaker Ground Truth Plus** can be used to build high-quality ML training datasets at a lower cost—and without having to build labeling applications or manage a labeling workforce on your own.

AWS customers are achieving massive scale in data preparation:

- **The NFL,** in collaboration with AWS, developed the Digital Athlete program, which uses ML to track and identify risks coming from helmet collisions. This requires labeling hours of video footage so that computer vision models can be trained on SageMaker and then track helmet collisions and detect impact during games.
- **Postis** created a scalable system with the power to run heavy ML workloads and support its global growth using AWS. Postis now serves more than 200 customers in 25 countries, including leading companies such as Ikea, Carrefour, Auchan, and Intersport.
- **Aurora**, a leader in self-driving vehicle technology, trains ML and cloud-based simulation workloads using AWS, processing trillions of data points each day. The company is scaling to complete up to 12 million physics-based driving simulations, building on the petabytes of data it collects during real-world road tests.

THOMSON REUTERS

Thomson Reuters accelerates research with Amazon SageMaker

Thomson Reuters—the world's leading source of news and information for professional markets—accelerated research and development of NLP solutions with cost savings and flexibility using Amazon SageMaker.

Thomson Reuters developed an internal platform to apply ML at scale with AWS. The platform enables its developers and data scientists to quickly gain new insights from real-time and historical data in a fully managed and secure environment. It saves developers and data scientists countless hours of coding by providing all the components used for ML in a single toolset. This helps the company put models into production faster, with much less effort and at a lower cost.

Read the success story ›

Make machine learning available to more users

SageMaker enables all users—including business analysts with no coding or ML experience and ML practitioners from beginners to experts—to generate predictions and transform their businesses with ML.

For data scientists and developers who prefer to write code in Python, SageMaker offers fully managed Jupyter notebook environments available through the SageMaker Studio IDE. For builders who prefer more automation, Amazon SageMaker Autopilot automatically builds, trains, and tunes ML models without any loss of visibility or control. When projects need to get fast-tracked, Amazon SageMaker JumpStart offers hundreds of pre-built algorithms, models, and solutions for the most common use cases, which can be deployed in just a few clicks.

For line-of-business analysts supporting finance, marketing, and operations, SageMaker Canvas offers a visual point-and-click interface to generate accurate ML predictions without requiring any ML experience or having to write a single line of code. SageMaker Canvas helps business analysts support common use cases such as churn prediction, forecasting, and pricing recommendations. For developers who prefer not to create their own models in SageMaker, they can use any of 25+ AI services from AWS for all top ML use cases, including text and documents, chatbots, speech, vision, search, business processes, code and DevOps, and even industry-specific services for healthcare and industrial.

To ensure accessibility for all, AWS features include:

- AWS AI services
- Amazon SageMaker Autopilot
- Amazon SageMaker Canvas
- Amazon SageMaker JumpStart
- Amazon SageMaker Studio

Freddy's

Freddy's orders up insights two times faster with Amazon SageMaker

Freddy's Frozen Custard & Steakburgers, a fast-casual restaurant chain headquartered in Wichita, Kansas, turned to data science to find a better way to evaluate the quality of its restaurants. Leveraging the accessibility of Domo AutoML powered by Amazon SageMaker Autopilot, Freddy's built ML models to optimize staffing levels in its restaurants without having to hire ML experts.

Read the success story ›

Foster responsible machine learning

Responsible use of ML is key to achieving tangible benefits that scale across the business. AWS is committed to developing fair and accurate AI and ML services and helping organizations transform responsible AI from theory into practice with purpose-built tools and guidance.

To use ML in a responsible manner, ML models need to be built with transparency, fairness, and security in mind. **Amazon SageMaker Clarify** provides bias detection across the ML workflow and includes feature importance graphs. These explain model predictions and produce reports to support internal presentations while also identifying issues with models to enable course correction.

To help your organization meet security criteria applicable to ML workloads, SageMaker includes solutions for encryption, private network connectivity, authorization, authentication, monitoring, and auditability.

Achieve responsible and secure machine learning with SageMaker Clarify:

- Gain greater visibility into data and models to identify and limit bias
- Detect potential bias throughout the entire workflow
- Explore feature-importance graphs to help explain model predictions

BUNDESLIGA

Bundesliga scores higher fan engagement with Amazon SageMaker

The Deutsche Fußball Liga (DFL) GmbH, responsible for organizing and marketing German professional football, set out to create a more engaging experience for Bundesliga fans around the world by uncovering game insights during football matches.

Bundesliga Match Facts, powered by AWS, give viewers information on the difficulty of a shot, the performance of their favorite players, and an exploration of offensive and defensive trends of their team. Using Amazon SageMaker Clarify, the DFL can now interactively explain the key components of the Bundesliga Match Facts insights predictions to improve its ML models and ultimately deliver higher-quality game insights to fans.

Read the success story ›

Improve cost-efficiency with purpose-built machine learning tools

As your use of ML grows, so will your infrastructure requirements. To prevent your costs from becoming prohibitive, you'll need tools and processes that allow you to dynamically match your spend to your specific compute, storage, and networking needs throughout the ML lifecycle. You'll also need to maximize productivity and efficiency, enabling your developers to avoid wasted time and duplicative efforts and to put models into production quickly.

By using services and tools that are purpose-built for ML, you can achieve speed, scale, and cost-efficiency that go far beyond general-purpose and on-premises solutions.

Throughout the ML lifecycle—including labeling, data preparation, feature engineering, training, hosting, monitoring, and workflows—your team can use a single visual inference in **SageMaker Studio**. This provides you with greater control over your infrastructure spend. Furthermore, it can improve your data science team's productivity by up to 10 times and enable them to develop models in weeks instead of months.[2]

[2] Lowering total cost of ownership for machine learning and increasing productivity with Amazon SageMaker

AWS customers are achieving massive scale, productivity, and cost-efficiency with purpose-built tools from AWS:

- **Vanguard** has fully automated the setup of its ML environments and can now deploy ML models 20 times faster.
- **AstraZeneca** can deploy new ML environments in five minutes versus one month to generate insights that improve research and development and accelerate the commercialization of new therapeutics.
- **NerdWallet** reduced training costs by about 75 percent, even while increasing the number of models it trained.
- **Zendesk** reduced ML inference costs by 90 percent by deploying thousands of models per endpoint using SageMaker multi-model endpoints.
- **Mueller Water Products** used SageMaker to improve leak detection performance. One of its customers estimates the solution will save it $8 million over five years.

Learn more about accelerating training and development of ML models ›

Scale machine learning across your business with MLOps

MLOps practices help you streamline the ML lifecycle by automating and standardizing ML workflows. With standardized MLOps processes in place, your teams can get models into production faster and collaborate more effectively. Over time, MLOps can help you reach your ultimate goal—scaling ML adoption and using ML to improve results across the entire organization.

Amazon SageMaker delivers the capabilities, automation, standardization, and centralization you need to make MLOps a reality for your organization. **Purpose-built MLOps tools** within SageMaker allow you to easily train, test, troubleshoot, deploy, and govern ML models at scale. This helps improve the productivity of your data scientists and ML engineers while maintaining model performance in production.

With the purpose-built MLOps tools provided by SageMaker, you can:

- Create repeatable training workflows to accelerate model development
- Catalog ML artifacts centrally for model reproducibility and governance
- Integrate ML workflows with continuous integration and continuous delivery (CI/CD) pipelines for faster time to production
- Continuously monitor data and models in production to maintain quality

Learn more about SageMaker for MLOps ›

It's time to embrace machine learning

By using purpose-built development and data tools, MLOps, no-code ML, infrastructure, and solutions focused on responsible use of data and models on a fully managed service, you can propel many more models from concept to production in a repeatable way for less cost.

Amazon SageMaker outpaces time-consuming, difficult, and expensive self-managed ML platforms to help you:

- Reduce total cost of ownership by 54 percent
- Achieve more than 10 times greater productivity
- Perform over 100 billion predictions per month
- Cut data labeling costs by 40 percent
- Accelerate model training by up to 50 percent through more efficient use of GPUs

And with 22 compliance programs (including PCI, HIPAA, SOC, 1/2/3, FedRAMP, and ISO), AWS can help you gain the swiftness and security that powers your business into the future.

Learn more about SageMaker for high-performance, low-cost ML development at scale ›