



# Accelerate Machine Learning Development to Build Intelligent Applications Faster

An IDC White Paper, Sponsored by Amazon Web Services

Author: David Schubmehl



Sponsored by:  
Amazon Web Services

Author: David Schubmehl

September 2020

## Highlights

By 2021, spending on AI applications will exceed  
**\$63 BILLION**

By 2023, growing to over  
**\$96 BILLION**

And by 2025, at least  
**90%**  
of new enterprise application releases will include embedded AI-based functionality, recommendations, or advice

# Accelerate Machine Learning Development to Build Intelligent Applications Faster

## IN THIS WHITE PAPER

Organizations today are experiencing situations and business conditions unlike anything they have seen in the past 50 years. The COVID-19 pandemic and related business conditions are leading companies to implement digital transformation and process reorganization at a rate that would have been unthinkable six months ago. To do this successfully, companies are working to improve and provide better value to their customers, shareholders, and workers. Specifically, they are looking for new ways to increase sales, reduce costs, streamline business processes, and understand their customers better by using various types of automation coupled with the ever-increasing amount of data available to them. Artificial intelligence (AI) and machine learning (ML) are key technologies that help organizations actualize digital transformation. As a result, IDC is seeing that many organizations are accelerating deployments and changing the focus of their AI projects from proof of concept (POC) to return on investment (ROI).

Organizations today are using deep learning (DL) to create models that are providing more accurate predictions, recommendations, and analyses in areas ranging from financial services and accounting to marketing, retail, and supply chains in production on a day-to-day and even minute by minute basis. Deep learning is a type of machine learning based on neural network (NN) algorithms, used to produce more accurate insights, recommendations, and predictions, trained on large amounts of data.

Organizations are using deep learning models to recommend products, predict pricing, recognize images, and improve decision making as well as for a host of other use cases. Until recently, developing deep learning models took significant amounts of time, effort, and knowledge and required expertise in this field. However, vendors such as Amazon Web Services (AWS) have developed services and tools for deep learning, allowing data scientists and developers (and even business analysts) to experiment, develop, test, and deploy deep learning models into production more quickly and easily than ever before.

Cloud-based tools such as managed machine learning services like Amazon SageMaker and preconfigured images like the AWS Deep Learning AMIs (Amazon Machine Images) provide capabilities that handle many of these factors and help developers and their organizations speed deep learning applications to market.

There are numerous deep learning tools and frameworks such as TensorFlow, PyTorch, and Apache MXNet — all have valuable attributes that make them useful in developing the algorithms to build the model. However, there are many factors involved that inhibit the development of deep learning models:

- » Choosing the right deep learning framework for the job at hand
- » Choosing the right deep learning algorithm
- » Adjusting and tuning the deep learning algorithm and data for the most accurate predictions
- » Identifying, locating, and curating training data for deep learning models
- » Having the right amount of compute resources for both model training and generating predictions in production (inferences)
- » Integrating deep learning models into existing enterprise applications
- » Operationalizing models to perform at scale in production

Cloud-based tools such as managed machine learning services like Amazon SageMaker and preconfigured images like the AWS Deep Learning AMIs (Amazon Machine Images) provide capabilities that handle many of these factors and help developers and their organizations speed deep learning applications to market.

For organizations that prefer the ease and convenience of using pre-trained AI services via APIs and services, Amazon Rekognition for images and video, Amazon Lex for chatbot integration, Amazon Polly for text to speech, Amazon Kendra for intelligent search, Amazon Textract for OCR text extraction, Amazon Translate for natural language translation, Amazon Transcribe for speech recognition, and Amazon Comprehend to find relationships in text can accelerate the addition of intelligent capabilities to applications. AWS offers fully managed services such as Amazon Personalize, Amazon Forecast, Amazon CodeGuru, and Amazon Fraud Detector that make the process of building models very easy. These services automatically inspect user data, extract features, select the appropriate algorithms, and then build and deploy the models into business processes via a single API call.

The advantage to developers is that without any machine learning expertise, they can just use these APIs simply and easily without having to go through the entire process of creating their own custom deep learning models. Most developers would be well served to check whether a preexisting API can solve their problem before beginning the process of creating a custom deep learning model.

Other use cases include office applications surfacing related content and suggestions as knowledge workers develop new analysis reports or create new content for a project.

## Situation Overview

### Introduction

The market for deep learning-based artificial intelligence applications has grown rapidly and continues to surge. IDC estimates that spending on AI applications will exceed \$63 billion by 2021, growing to over \$96 billion by 2023, and by 2025, at least 90% of new enterprise application releases will include embedded AI-based functionality, recommendations, or advice. Organizations need to consider the following reasons as to why these systems are important for their future:

- » **Augment human judgment.** The best business cases are about extending human capabilities, not replacing them, by positioning intelligent applications as an extension of human intention. Power tools in the hands of a craftsman is the best analogy. Pricing optimization models are good examples of deep learning in this area. A second example would be an AI imaging application that automatically identifies cancerous tumors by examining radiology images, aiding radiologists.
- » **Accelerate investigation and discovery.** Even the very best human readers cannot comprehend millions of pages of documents in one day. Applications that understand natural language can be applied to this task for both the spoken and the printed word. Deep learning-based natural language systems provide better results than handcrafted, taxonomy-based systems.
- » **Recommend “next best actions” and predict outcomes.** Deep learning-based applications build models using relevant data for recommendations and predictions, which are some of the typical use cases.
- » **Personalize outcomes and recommendations.** Many organizations are beginning to use deep learning models to “personalize” content, predictions, and recommendations to specific customers or prospects. This is especially true with mobile applications where users increasingly expect their devices and applications to “know” their likes, dislikes, and expectations.
- » **Automate organizational knowledge management.** While knowledge management systems have existed for decades, many have failed under the weight of human effort required for ongoing operation. Applying automation to investigation and discovery activities, or developing best practices, is a key benefit. Automatic categorization and theme identification of documents are some of the key use cases of deep learning. Other use cases include office applications surfacing related content and suggestions as knowledge workers develop new analysis reports or create new content for a project.

At the same time as intelligent, AI-enabled applications are beginning to emerge, we are seeing a growing market for deep learning tools and solutions based on open source.

The combination of high-performance compute resources, tremendous amounts of data, and the frameworks and libraries for deep learning is solving problems and challenges without the need to resort to programming.

» **Encapsulate and “systematize” best practices.** This is a variation on the themes previously mentioned about learning from experience; developing deep learning models that replace rule- or heuristics-based systems is a key use case in this area.

As previously mentioned, organizations need to digitally transform themselves to operate during these times of unprecedented change. AI and machine learning are the key catalysts that organizations are using to effect this digital transformation. Large healthcare organizations are examining how deep learning-based, cloud-hosted computer vision applications can help democratize and accelerate “best practice” diagnosis and treatment regimens, no matter where their clients live. Many retailers are using deep learning-based solutions to target very specific audiences and are developing customized personalized profiles for each shopper, resulting in very specific personalized offers for products ranging from lattes to the latest designer clothes and shoes.

Global financial institutions are using intelligent, AI-enabled applications to accelerate, automate, and sometimes, eliminate manual workflows and business processes that handle financial transactions. Manufacturing companies are developing sophisticated predictive maintenance strategies based on IoT and deep learning models. These are just a few of the hundreds of use cases that organizations are beginning to examine as their marketplaces and competition begin to embrace artificial intelligence and deep learning applications.

At the same time as intelligent, AI-enabled applications are beginning to emerge, we are seeing a growing market for deep learning tools and solutions based on open source. A powerful combination of motivated, capable developers; a proven open source community development model; and the need and desire for more efficient, nimble, and responsive software products to provide automated and adaptive methods has led to a growing market segment producing deep learning software libraries and tools. These tools are part of a larger group of technologies that include speech recognition, natural language processing, predictive analytics, advanced analytics, reinforcement learning, neural networks, and supervised and unsupervised machine learning. The endgame is all about making applications smarter by using special libraries containing self-learning algorithms, which when unleashed on a data set can learn and self-program to solve various types of problems. These self-programmed computer algorithms are fueling the emergence of what we at IDC call intelligent applications.

The emergence of tools, frameworks, and libraries that provide services for deep learning is setting the stage for a low-cost enabler of intelligent applications to be built by developers today. Organizations are looking at these services to replace rule- or heuristics-based approaches that must be extensively programmed and maintained today. The combination of high-performance compute resources, tremendous amounts of data, and the frameworks and libraries for deep learning is solving problems and challenges without the need to resort to programming. These deep learning libraries



Improvements in the variety, efficiency, and reliability of machine learning will make these applications more usable and stable and help increase their popularity.

Deep learning is based on neural network algorithms that have seen significant commercial success.

and technologies are being used for an ever-wider array of use cases, from image recognition and disease diagnosis to pricing optimization and product recommendations. Deep learning is a key component of most AI applications and is also being added to enterprise applications. Improvements in the variety, efficiency, and reliability of machine learning will make these applications more usable and stable and help increase their popularity.

## Machine Learning and Deep Learning

To be clear about these topics, it helps to understand how IDC defines machine learning and related processes:

- » **Machine learning** is a subset of AI techniques that enables computer systems to learn from previous experience (i.e., data observations) and improve their behavior for a given task. It is the process of creating a statistical model from various types of data that performs various functions without having to be programmed by a human.
- » **Neural networks, or artificial NNs**, are a subset of ML techniques loosely inspired by biological neural networks. They are usually described as a collection of connected units, called artificial neurons, organized in layers.
- » **Deep learning** is a subset of NNs that makes the computational multilayer NN feasible. Typical DL architectures are deep neural networks (DNNs), convolutional neural networks (CNNs), recurrent neural networks (RNNs), generative adversarial networks (GANs), and so forth.

Deep learning is based on neural network algorithms that have seen significant commercial success. A neural network attempts to mimic a biological brain through interconnected artificial neurons that have various weights assigned to influence how an algorithm arrives at an answer. Deep learning models improve through complex pattern recognition in pictures, text, sounds, and other data that influence the relative weight of each neuron over time (a process called training) to produce more and more accurate insights, recommendations, and predictions.

There are many types of frameworks and tools for deep learning. A few of these are:

- » **TensorFlow**, originally developed by Google, is a comprehensive ecosystem of tools for developers, enterprises, and researchers that want to push the state-of-the-art machine learning and build scalable ML-powered applications.
- » **PyTorch** is a Python package originally developed by Facebook for performing scientific computations or building dynamic neural networks, allowing the development of machine learning for more advanced and complex AI tasks.

Amazon Web Services and Microsoft developed Gluon, a deep learning API that allows developers of all skill levels to prototype, build, and train deep learning models.

- » **Apache MXNet** is a scalable training and inference framework from the Distributed (Deep) Machine Learning Community (DMLC) consortium and is an incubating project of the Apache Software Foundation.
- » **Caffe/Caffe2** is a deep learning framework that is optimized for high-speed processing of images.

In addition to these frameworks, there are higher-level tools such as Gluon and Keras. Keras is a high-level neural networks API, written in Python and capable of running on top of TensorFlow and Microsoft Cognitive Toolkit. It was developed with a focus on enabling faster model development.

Amazon Web Services and Microsoft developed Gluon, a deep learning API that allows developers of all skill levels to prototype, build, and train deep learning models. Gluon currently works with the deep learning framework Apache MXNet and gives developers access to a simplified programming framework that allows them to accelerate both the development of neural network-based models and the time required to train them. In addition, Gluon enhances the agility of developers by enabling them to make changes more easily to neural networks and to debug training models faster by incrementally understanding the effects of model training.

## Why Deep Learning?

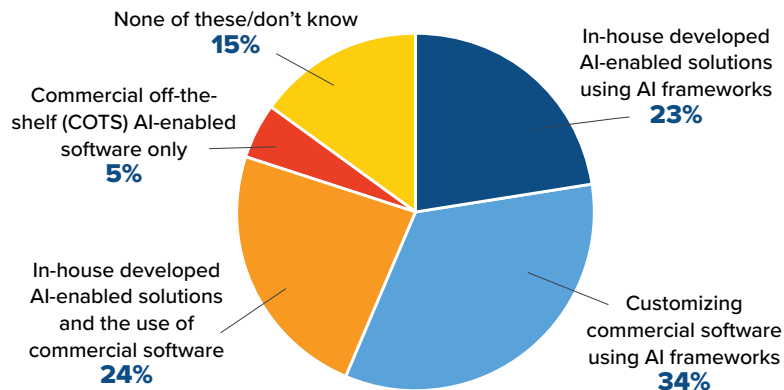
IDC research in 2019 showed that at least 85% of the organizations with over 500 employees surveyed had adopted some type of intelligent, AI-enabled software and the vast majority of those were customizing or developing their own AI/ML models and applications (see Figure 1). In our research, we define AI software frameworks and platforms as:

Artificial intelligence software frameworks and platforms are defined as sets of technologies that can be used to develop applications that answer questions, hold conversations, discover insights, provide recommendations, and/or automate tasks and processes. These frameworks and platforms offer deep learning, machine learning, natural language processing, voice/speech recognition, image/video analysis, and other technologies to develop intelligent, AI-enabled solutions.

**FIGURE 1**

## AI Adoption Status

Q. How would you describe your organization's development and use of intelligent, AI-enabled software/solutions?



n=839

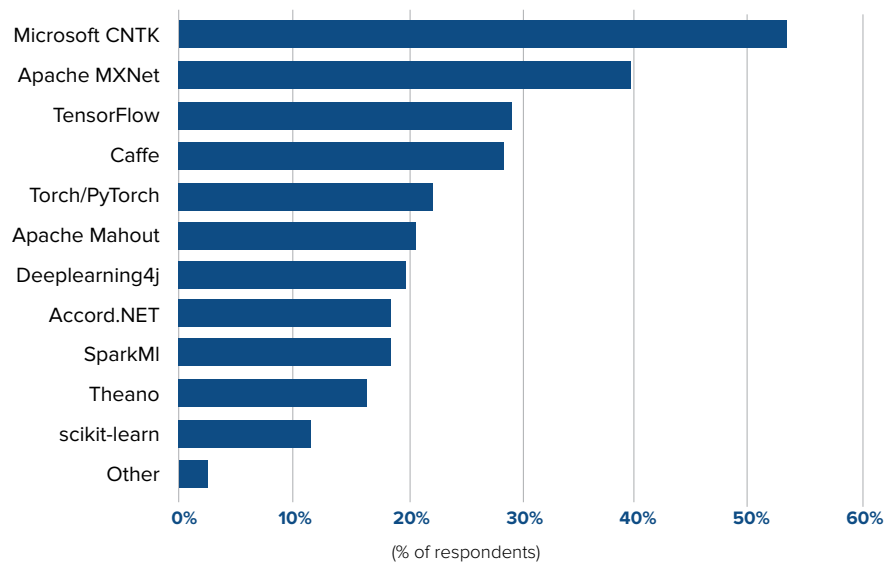
Source: IDC's AI Software Platform Adoption Survey, 2019

Most firms that IDC surveyed are using some amount of open source technologies for their machine learning development (see Figure 2).

**FIGURE 2**

## Use of ML Frameworks

Q. Which of the following types of machine learning frameworks are you actively working with?



n = 340

Base = respondents who are actively working with machine learning frameworks

Source: IDC's AI Software Platform Adoption Survey, 2019

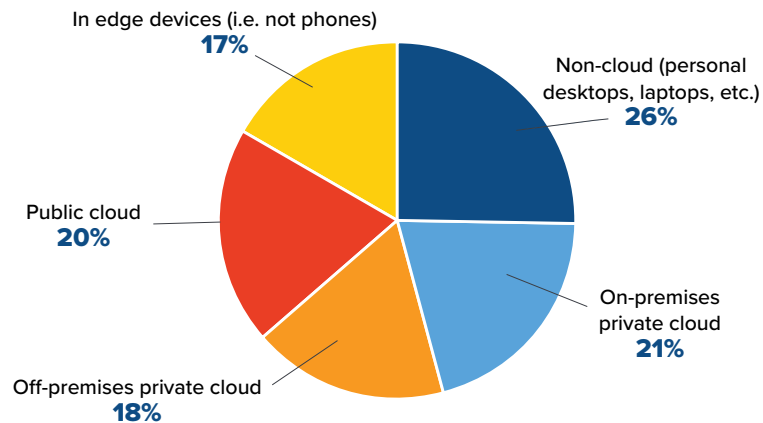


In addition, most organizations are using either private or public cloud environments for machine learning/deep learning model development (see Figure 3).

**FIGURE 3**

## AI/ML Development Locations

*Q. Where are you developing your AI models?*



n=250

Base = all respondents

Source: IDC's AI Frameworks, Tools, Methods Survey, June 2019

According to IDC's 2019 *AI Frameworks, Tools, Methods Survey*, 37.5% of organizations are developing their models off-premises in a cloud environment. IDC believes that public cloud as a development and deployment environment will outweigh on-premises development and deployment by as early as 2023, according to *Worldwide Artificial Intelligence Software Platforms Forecast, 2020–2024* (IDC #US45724520, June 2020). In addition, IDC is seeing more cloud-based projects moving from proof of concept to production.

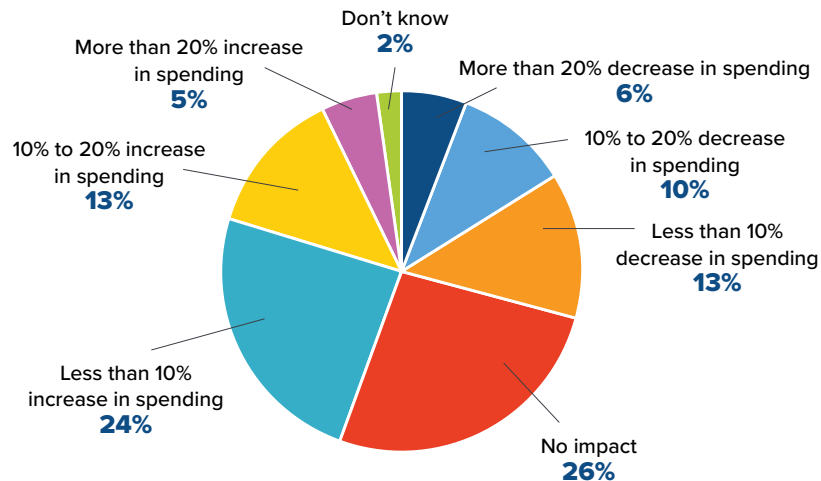
This is consistent with the general trends we see with application deployments for other workload types. Developing cloud-based intelligent, AI-enabled applications that use deep learning is increasingly popular across both the enterprise software market and the application development market. IT organizations are feeling pressure from their management teams, boards, and even customers to find advantages and efficiencies with this new wave of computing and increasingly popular public cloud development and deployment.

Even the COVID-19 pandemic hasn't slowed budgeted IT plans for AI spending. IDC's most recent *COVID-19 Impact on IT Spending Survey* (conducted July 7–20, 2020) indicated that the majority of organizations are either increasing their AI and deep learning spending or leaving it unchanged. Only 16% of respondents on a worldwide basis planned to reduce AI and deep learning spending (see Figure 4).

**FIGURE 4**

## AI Spending Plans

*Q. Compared with your organization's originally budgeted IT spending plans, in 2020 overall, how do you think your organization's actual spending on artificial intelligence will be affected due to COVID-19?*



n=844

Note: The survey was conducted July 7–20, 2020.

Source: IDC's COVID-19 Impact on IT Spending Survey, 2020

Given all of this, organizations should be emphasizing the need to invest in and create AI and deep learning solutions. However, there are significant challenges faced by enterprises wanting to adopt these new technologies:

- » Disparate tools and technologies are freely available, but knowing where to start and which tools or technologies to use can be confusing to organizations and their developers.
- » The lack of integrated development environments for deep learning slows down the cycle of experimentation, development, testing, and production. While these types of tools have been available for languages for years, getting models to production was complicated and time consuming, but this is changing.
- » There is an inability to find and organize enough of the right types of data to build and operate deep learning-based models within their own organizations.
- » The absence of suitable developer skills in deep learning and data science to use these tools makes this more difficult for organizations. Finding data scientists is challenging; finding data scientists with developer skills to use existing deep learning frameworks is even more difficult.
- » API and/or template-based solutions designed for use with prebuilt domains do exist, but locating them and making use of them for a project can be problematic at best.

The question for enterprises is how best to develop these deep learning models while minimizing the amount of effort and time needed to collect and organize data and then develop accurate predictive and prescriptive models.














**FIGURE 5**

## AWS Deep Learning Architecture

### The AWS ML Stack

Broadest and most complete set of machine learning capabilities

#### AWS AI SERVICES

VISION	SPEECH		TEXT			SEARCH	CHATBOTS	PERSONALIZATION	FORECASTING	FRAUD	DEVELOPMENT	CONTACT CENTERS
												
Amazon Rekognition	Amazon Polly	Amazon Transcribe	Amazon Comprehend	Amazon Translate	Amazon Textract	Amazon Kendra	Amazon Lex	Amazon Personalize	Amazon Forecast	Amazon Fraud Detector	Amazon CodeGuru	Contact Lens <i>For Amazon Connect</i>
		+Medical	+Medical									

#### AWS ML SERVICES

 Amazon SageMaker	Amazon SageMaker Ground Truth	AWS ML Marketplace	AMAZON SAGEMAKER STUDIO IDE							Neo	Augmented AI
			Built-in algorithms	Notebooks	Experiments	Model training & tuning	Debugger	Autopilot	Model hosting	Model Monitor	

#### ML FRAMEWORKS & INFRASTRUCTURE

TensorFlow PyTorch	MXNet	Gluon scikit-learn	Horovod Deep Graphic Library	Keras	Deep Learning AMIs & Containers	GPUs & CPUs	Elastic Inference	Inferentia	FPGA
-----------------------	-------	-----------------------	---------------------------------	-------	------------------------------------	-------------	-------------------	------------	------

Source: Amazon Web Services, 2020

## Considering AWS Machine and Deep Learning Services

AWS has made it much easier for deep learning developers by offering a broad range of tools and services to get deep learning models created and into production. AWS offers tools and capabilities at every level of the stack (refer back to Figure 5).

At the bottom level are open source frameworks and infrastructure resources for developing, implementing, and operating deep learning models at scale. These tools range from infrastructure such as easily usable and expandable GPU, FPGA, and CPU resources; elastic inference and deep learning AMIs; and containers to frameworks such as PyTorch, TensorFlow, and Apache MXNet as well as support for Gluon and Keras. AWS is unique in its approach in supporting all deep learning frameworks, rather than offering a single preferred framework as some of the other cloud service providers do.

AWS is also offering fully managed services for jobs such as personalization with Amazon Personalize, forecasting with Amazon Forecast, fraud detection with Amazon Fraud Detector, AI-powered search with Amazon Kendra, and a host of others.

In the middle level, Amazon SageMaker enables developers and data scientists to build, train, and deploy deep learning models at any scale quickly and easily. Amazon SageMaker removes the complexity that gets in the way of successfully implementing deep learning across use cases and industries. This includes Amazon SageMaker Studio, a fully integrated development environment (IDE) for machine learning, that provides a single web-based visual interface for the complete ML workflow.

Finally, at the top are AWS' ready-to-use AI services and APIs. These include deep learning-based services for vision with Amazon Rekognition, speech with Amazon Polly and Amazon Transcribe, text with Amazon Comprehend, and others. AWS is also offering fully managed services for jobs such as personalization with Amazon Personalize, forecasting with Amazon Forecast, fraud detection with Amazon Fraud Detector, AI-powered search with Amazon Kendra, and a host of others.

## Amazon SageMaker: A Fully Managed Service for Machine Learning

To ease development, training, and deployment of deep learning models, AWS has developed and extended Amazon SageMaker into a complete deep learning development environment. SageMaker is a fully managed service that enables developers and data scientists to build, train, and deploy deep learning models at any scale quickly and easily. SageMaker removes the complexity that holds back developer success. SageMaker includes modules that can be used together or independently to build, train, and deploy deep learning models, such as:

- » **Amazon SageMaker Studio:** This is the first fully integrated development environment for deep learning where developers and data scientists can perform all development steps. Teams and individuals can quickly upload data, create and share new notebooks, train and tune models, move back and forth between steps to adjust experiments, debug and compare results, and deploy and monitor models (all in a single visual interface), making them much more productive.
- » **Amazon SageMaker Autopilot:** Automatically build, train, and tune models with full visibility and control using Amazon SageMaker Autopilot. It incorporates the industry's first automated machine learning capability that gives developers complete control and visibility into how their models were created and what logic was used in creating these models.
- » **Amazon SageMaker Ground Truth:** This fully managed data labeling service makes it easy to build highly accurate training data sets for machine learning. It provides tools and services for labeling data including automatic data labeling, which uses a machine learning model to label first-party data.

Amazon SageMaker requires no setup and provides hosted Jupyter notebooks so that developers can start processing training data sets and developing machine learning and deep learning models immediately.

SageMaker provides high-performance, scalable machine learning algorithms optimized for speed, scale, and accuracy.

- » **Amazon SageMaker Neo:** This service enables developers to train machine learning models once and run them anywhere in the cloud and at the edge. It optimizes models to run up to twice as fast, with less than a tenth of the memory footprint, with no loss in accuracy.
- » **Amazon Augmented AI:** This service makes it easy to build the workflows required for human review of ML predictions. It brings human review to all developers, removing the undifferentiated heavy lifting associated with building human review systems or managing large numbers of human reviewers.

SageMaker requires no setup and provides hosted Jupyter notebooks so that developers can start processing training data sets and developing machine learning and deep learning models immediately. All it takes is a few clicks in the SageMaker console to create a fully managed notebook workspace. The service takes care of establishing secure network connections to your organization's VPC and launching an Amazon EC2 instance, preloaded with useful libraries for the most popular machine learning and deep learning frameworks such as TensorFlow and PyTorch. Developers can build or import their own notebook or just bring data to one of many prebuilt notebooks designed for common use cases, such as risk modeling, churn prediction, and OCR. To prepare the data from Amazon S3, Amazon Redshift, Amazon DynamoDB, and Amazon RDS for model training, developers can use AWS Glue, Apache Spark, on Amazon EMR for data preprocessing, and Amazon EFS as optional storage for your workspace.

When the application is ready to train, developers simply indicate the type and quantity of Amazon EC2 instances that they need and initiate training with a single click. SageMaker then sets up the distributed compute cluster, performs the training, and tears down the cluster when complete; so organizations only pay for the resources they have used and don't have to worry about the underlying infrastructure. SageMaker seamlessly scales to virtually unlimited nodes; so developers no longer need to worry about all the complexity and lost time involved in making distributed training architectures work.

SageMaker provides high-performance, scalable machine learning algorithms optimized for speed, scale, and accuracy. Developers can choose from supervised algorithms where the correct answers are known during training, and you can instruct the model where it made mistakes. For example, SageMaker includes supervised algorithms such as XGBoost and linear/logistic regression or classification to address recommendation and time series prediction problems. SageMaker also includes support for unsupervised learning (i.e., the algorithms must discover the correct answers on their own), such as with k-means clustering and principal component analysis (PCA), to solve problems such as identifying customer

SageMaker also reduces the amount of time spent tuning deep learning models. It can automatically tune the model by adjusting thousands of different combinations of algorithm parameters to arrive at the most accurate predictions the model can produce.

groupings based on purchasing behavior. In addition, developers can use Amazon Augmented AI to build a “human in the loop” review process that allows humans to aid in low-confidence predictions.

SageMaker also reduces the amount of time spent tuning deep learning models. It can automatically tune the model by adjusting thousands of different combinations of algorithm parameters to arrive at the most accurate predictions the model can produce. This can save days of manual trial-and-error adjustments. SageMaker’s support for hyperparameter tuning automatically identifies the right set of tuning parameters for the models by applying different techniques such as random search or Bayesian search.

After training, SageMaker provides the model artifacts and scoring images to the developer for deployment to EC2 or anywhere else. The developer can then specify the type and number of EC2 instances, and SageMaker takes care of launching the instances, deploying the model, and setting up the HTTPS endpoint for your organization’s application to achieve low-latency/high-throughput inferences. Once in production, SageMaker manages the compute infrastructure to perform health checks, apply security patches, and conduct other routine maintenance, all with built-in Amazon CloudWatch monitoring and logging. Organizations pay for AWS compute and storage resources that the model uses for hosting the Jupyter notebook, training the model, performing predictions, and logging the outputs. Building, training, and hosting are billed by the second, with no minimum fees and no up-front commitments.

Finally, one of the best aspects of SageMaker is its modular architecture. Developers can use any combination of SageMaker’s building, training, and deployment capabilities to fit the organization’s workflow. With SageMaker, developing and deploying a deep learning model can be as straightforward as choosing a notebook template, selecting an algorithm, and then training, testing, and deploying the model using the management service. Amazon also offers a service called SageMaker Neo that enables developers to train machine learning models developed using popular frameworks once and then run them anywhere in the cloud or at the edge. This service optimizes models to run up to twice as fast, with less than a tenth of the memory footprint, with no loss in accuracy, allowing them to be deployed into almost any environment. The bottom line is that SageMaker provides an end-to-end machine learning environment that can significantly accelerate and simplify the process of creating, training, and deploying models into production applications.



Experienced machine learning developers that are already familiar with machine learning frameworks and the tools necessary to build machine learning applications can use these AMIs to deploy applications more quickly.

## AWS Deep Learning AMIs

Amazon also offers its AWS Deep Learning AMIs as another option for enterprises and their developers. Built on top of Amazon EC2, the AWS Deep Learning AMIs offer a full environment for building, training, and running deep learning applications. The AMIs come with the latest versions of pre-installed, open source deep learning frameworks including Apache MXNet and Gluon, TensorFlow, PyTorch, Chainer, Theano, Keras, and Microsoft Cognitive Toolkit. They offer GPU acceleration through preconfigured CUDA and cuDNN drivers. The AMIs also come with popular Python packages and the Anaconda Platform. The instances in the platform also come preconfigured with Jupyter notebooks, which enables the implementation of interactive deep learning models using Python 2.7 or 3.4. In addition to Jupyter, AWS Deep Learning AMIs include other toolkits such as CppLit, PyLint, Pandas, and GraphViz.

There are three versions of the AWS Deep Learning AMI. The first is a Conda-based AMI with separate Python environments for deep learning frameworks created using Conda — a popular open source package and environment management tool. The second is a Base AMI with GPU drivers and libraries for developers to deploy their own customized deep learning models. These are preconfigured environments that allow developers the freedom and flexibility to use the setup and tools they need to accomplish their desired goals with less work and aggravation. The last is the AMI with Source Code for developers that want pre-installed deep learning frameworks and their source code in a shared Python environment. Experienced machine learning developers that are already familiar with machine learning frameworks and the tools necessary to build machine learning applications can use these AMIs to deploy applications more quickly.

## NVIDIA GPU-Powered Deep Learning Models

Amazon Web Services offers NVIDIA GPU-powered instances for both machine learning training and inference. With the launch of Amazon EC2 P3 instances in 2017, AWS was the first cloud provider to introduce instances optimized for machine learning training in the cloud with powerful NVIDIA V100 Tensor Core GPUs, allowing customers to reduce machine learning training time from days to minutes. For GPU-powered inference, in 2019 AWS introduced Amazon EC2 G4 instances featuring NVIDIA T4 Tensor Core GPUs to provide the most cost-effective GPU instances in the cloud for running machine learning inference.

Amazon EC2 P3 instances deliver high-performance compute in the cloud, with up to eight NVIDIA V100 Tensor Core GPUs and up to 100Gbps of networking throughput for machine learning applications. These instances deliver up to one

Amazon EC2 G4 instances deliver the industry's most cost-effective and versatile GPU instance for deploying machine learning models in production. G4 instances provide the latest-generation NVIDIA T4 GPUs, up to 100Gbps of networking throughput, and up to 1.8TB of local NVMe storage.

petaflop of mixed-precision performance per instance to significantly accelerate machine learning applications. Amazon EC2 P3 instances have been proven to reduce machine learning training times from days to minutes.

NVIDIA V100 Tensor Core is the first Tensor Core GPU brought to market, and it is built to accelerate AI, high-performance computing (HPC), data science, and graphics. It's powered by NVIDIA Volta architecture, comes in 16GB and 32GB configurations, and offers the performance of up to 32 CPUs in a single GPU. Data scientists, researchers, and engineers can now spend less time optimizing memory usage and more time designing the next AI breakthrough.

Amazon EC2 G4 instances deliver the industry's most cost-effective and versatile GPU instance for deploying machine learning models in production. G4 instances provide the latest-generation NVIDIA T4 GPUs, up to 100Gbps of networking throughput, and up to 1.8TB of local NVMe storage. G4 instances are offered in different instance sizes with access to one GPU or multiple GPUs and different amounts of vCPU and memory — giving developers the flexibility to pick the right instance size for their applications. G4 instances are optimized for machine learning application deployments (inference), such as image classification, object detection, recommendation engines, automated speech recognition, and language translation that push the boundary on AI innovation and latency.

The NVIDIA T4 GPU accelerates diverse cloud workloads, including high-performance computing, deep learning training and inference, machine learning, data analytics, and graphics. Based on the NVIDIA Turing architecture and packaged in an energy-efficient 70W, small PCIe form factor, T4 is optimized for mainstream computing environments and features multi-precision Turing Tensor Cores and new RT Cores. Combined with accelerated containerized software stacks from NGC, T4 delivers revolutionary performance at scale.

## Challenges and Opportunities

COVID-19 is leading every organization to examine and understand its business processes and methods of doing business. The reasons for this are many, but the bottom line is that organizations that are not open to changing the way they do business may not survive this incredibly complex business cycle. Organizations need to understand where AI and deep learning technology will deliver the best business benefits. They also need to understand what skill sets are needed to build and deploy intelligent, AI-enabled applications. Finally, organizations need to reassess what tools, infrastructure, and environments are needed to put these intelligent, AI-enabled applications to use, especially with all the changes that have occurred in the development and deployment of deep learning models over the past two years.

The need and desire for better (and simpler) tools, quicker time to market, and efficiency are key concerns in the market for intelligent, AI-enabled applications.

Organizations need guidance about what types of tools and technologies can help them develop intelligent, AI-enabled applications. They also need to understand when, why, and how these applications will be most effective in their organizations. In addition, organizations need to measure the effectiveness of these applications to determine return on investment for future projects that will include deep learning.

Finally, the AI platform-as-a-service market is already crowded and is becoming more competitive with every passing day. The need and desire for better (and simpler) tools, quicker time to market, and efficiency are key considerations in the market for intelligent, AI-enabled applications. There are numerous established and emerging vendors addressing and providing services and solutions within this space at a very wide range of capabilities. As such, Amazon Web Services faces the challenge of continuing as a leader in this market and will need to maintain an aggressive pace of engineering and innovation. Although AWS is productizing machine learning/deep learning services as the foundation of its solutions, this approach is not new to this market. What is new is that managed services such as SageMaker and the AWS Deep Learning AMIs combine numerous deep learning tools, frameworks, and technologies into a single integrated platform that provides significant productivity enhancements for organizations and developers. AWS needs to keep providing this level of innovation and expertise in this emerging market. Given the current state of innovation in deep learning and machine learning, it is important that these types of services remain framework agnostic and continue to accommodate the latest and greatest techniques and algorithms into the development process.

## Conclusion

Critical success factors related to machine learning/deep learning implementation are related to people, process, and technologies. Traditionally, emerging technical solutions require sharp and motivated developers that like to live on the cutting edge of technology. However, cloud vendors are finding ways to democratize the development and use of AI and deep learning technologies to promote wider use and deployment within enterprises. The key is to quickly develop successful models and products based on deep learning. Some factors that can assist with this are:

- » **Quick start packages/development tools.** Some vendors offer templates, sample data, and sample code to help jump-start developer productivity. With managed services such as Amazon SageMaker Studio, data scientists and developers (and even nondevelopers) can be even more productive than they could be with just templates and sample code.

Organizations should make sure the intelligent, AI-enabled solution that they are building will be able to help achieve the desired business outcome and/or address the issues that the solution is planned to overcome utilizing deep learning.

- » **Assistance with data.** Some vendors are either providing curated third-party data sets or evaluating doing so to assist developers with creating the kinds of cutting-edge predictive and prescriptive machine learning models that customers are looking for. For those that have their own data, many cloud vendors are now also offering data curation and integration services that make creating well-formulated data sets easier. Amazon SageMaker Ground Truth provides data labeling services for exactly this purpose. AWS also hosts a number of open public data sets on a variety of topics (geospatial, environmental, genomics, life sciences, regulatory, and statistical data) as well as general-purpose data for machine learning such as images, web crawls, news, and web video.
- » **Education.** Providing training and courses on how developers can best make use of these tools allows developers to get up and running without having to do everything themselves. With machine learning/deep learning services such as Amazon SageMaker, a little education can make nontraditional developers and data scientists productive and enable them to build and deploy their own deep learning models. Amazon offers a full suite of training and education options for machine learning developers, such as self-paced digital training, classroom training, and even AWS certification in machine learning. In addition, AWS has made available tools such as a deep learning-enabled video camera for developers to learn about building AI applications called AWS DeepLens, a deep learning-enabled model race car called AWS DeepRacer and even an ML-enabled keyboard called AWS DeepComposer. Designed specifically to educate developers, these tools include tutorials, sample code, and training data that can be used to get started on building AI models, all without having to write a single line of code.
- » **Consulting and advisory services.** These services will help developers become productive and will help them with challenges related to the kinds of data that they are consuming. An example of this is the new AI and machine learning competency for the AWS Partner Network. Amazon is certifying partners in machine learning/deep learning with this program today. In addition, AWS has created the Amazon Machine Learning Solutions Lab to help organizations develop AI applications more quickly and easily. The ML Solutions Lab pairs enterprise teams with AWS machine learning experts to prepare data, build and train models, and put models into production.

Great intelligent, AI-enabled applications require both advanced technology and solid design judgment. Organizations should make sure the intelligent, AI-enabled solution that they are building will be able to help achieve the desired business outcome and/or address the issues that the solution is planned to overcome utilizing deep learning. Engage in-house subject matter experts, the right

stakeholders, and consulting partners to help develop the right use cases to align with the desired business outcome. Make sure to include past project experiences in the organization's design thinking approach and, if available, include predefined use cases that have been developed for peers within the organization's industry to help develop the optimal use cases for the desired outcome. This process should involve continuous innovation and prototyping until the right use cases have been developed.

There are a wide variety of tools and libraries available, and it is not always clear which services or libraries are best for the use cases or jobs that developers should accomplish to successfully develop intelligent, AI-enabled applications.

Offerings such as pre-trained and managed AI services for natural language understanding and comprehension, computer vision, enterprise search, personalization, fraud detection, demand forecasting etc. allow any developer to build intelligent apps without any machine learning expertise. Amazon SageMaker and the AWS Deep Learning AMLs provide the ways and means for developers to become more productive and deploy deep learning models as well as support services such as data curation, integration, and management to solve a wide range of challenges that are difficult to solve with traditional coding methods and address the organization's business needs. Organizations should be evaluating tools and services like these as they begin to develop and deploy intelligent, AI-enabled applications using deep learning models.

#### IDC Global Headquarters

5 Speen Street  
Framingham, MA 01701  
USA  
508.872.8200  
Twitter: @IDC  
idc-insights-community.com  
www.idc.com

#### Copyright Notice

External Publication of IDC Information and Data — Any IDC information that is to be used in advertising, press releases, or promotional materials requires prior written approval from the appropriate IDC Vice President or Country Manager. A draft of the proposed document should accompany any such request. IDC reserves the right to deny approval of external usage for any reason.

Copyright 2020 IDC. Reproduction without written permission is completely forbidden.

#### About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications and consumer technology markets. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.