White Paper

Streamlining the Development of Deep Learning Applications in an AI-Enabled World

Sponsored by: Amazon Web Services

David Schubmehl
February 2018

IN THIS WHITE PAPER

In today's hypercompetitive business environment, organizations are continually seeking to improve and provide better value to their customers, shareholders, and workers. Specifically, they're 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. To that end, many organizations have started to look at deep learning as a method to build real-world artificial intelligence (AI) into their applications and business processes. Machine learning (ML) is the process of creating a statistical model from various types of data that perform various functions without having to be programmed by a human. Deep learning is a type of machine learning based on neural network 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. Recently vendors like Amazon Web Services (AWS) have developed services and tools for deep learning, that allow data scientists and developers to develop and deploy deep learning models more quickly and easily.

There are numerous machine/deep learning tools and frameworks such as TensorFlow, Apache MXNet, Caffe2, and PyTorch – all have valuable attributes that make them useful in developing intelligent applications. However, there are many factors involved that inhibit the development of machine learning applications:

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

February 2018, IDC #US43529118

Cloud-based tools such as managed machine learning platforms 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 machine/deep learning applications to market.

For organizations that prefer the ease and convenience of using pre-trained deep learning models via APIs, services such as Amazon Rekognition for images and video, Amazon Lex for chatbot integration, Amazon Polly for text to speech, Amazon Translate for natural language translation, Amazon Transcribe for speech recognition, and Amazon Comprehend to find relationships in text can accelerate adding intelligent capabilities to applications. The advantage to developers is that they can just use these APIs simply and easily without having to go through the entire process of creating their own custom machine learning models. Most developers would be well served to check if a pre-existing API can solve their problem before beginning the process of creating a custom machine learning model.

SITUATION OVERVIEW

Introduction

The market for machine learning and deep learning-based AI applications has grown rapidly and continues to surge. IDC estimates that spending on machine learning and deep learning solutions will exceed $57 billion by 2021, and by 2026, IDC predicts 75% of all enterprise software will include some aspect of machine/deep learning for predictions, recommendations, or advice. Organizations need to consider the following reasons why these systems are important to their future:

- **Augment human judgment.** The best business cases are about extending human capabilities, not replacing them, by positioning AI-enabled applications as an extension of human intention. Power tools in the hands of a craftsperson 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, providing assistance to radiologists.
- **Accelerate investigation and discovery.** Even the very best human readers can't 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.

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

IDC is beginning to see that organizations are using deep learning applications as a catalyst for business process disruption, digital transformation, and the creation of new economies of scale. 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. Global financial institutions are using 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, machine learning, and deep learning applications.

At the same time as AI-enabled applications are beginning to emerge, we're 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 low-cost or free software products to provide learning abilities for computing systems has led to a growing market segment producing machine learning/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, 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 machine learning and 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 have to be extensively programmed and maintained today. The combination of high-performance compute resources, tremendous amounts of data, and the frameworks and libraries for machine /deep learning is solving problems and challenges without the need to resort to programming. These machine learning/deep learning libraries 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. Machine/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.

Deep Learning

Deep learning is a type of machine learning based on neural network algorithms that has seen significant commercial success. A neural network attempts to mimic biological brains 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:

- **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.
- **Microsoft Cognitive Toolkit** is a Microsoft Research-initiated technology positioned as a unified deep learning toolkit that can be used as a services library for Python or C++ programs or as a standalone machine learning tool.
- **Caffe/Caffe2** is a deep learning framework that is optimized for high-speed processing of images.
- **TensorFlow,** originally developed by Google, is an open source software framework for numerical computation.
- **Torch** is a "GPU first" deep learning framework that supports N-dimensional arrays, with library routines offered to execute a variety of mathematical functions on the array data.
- **PyTorch** is a python package for doing scientific computations or building dynamic neural networks, allowing the development and machine learning for more advanced and complex AI tasks.

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 recently announced 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 more easily make changes to neural networks and to debug training models faster by incrementally understanding the effects of model training.

Why Deep Learning?

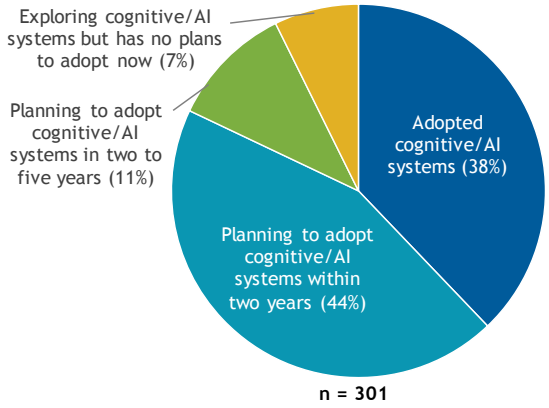
IDC research shows that a full 93% of organizations that have explored cognitive/AI systems have either adopted cognitive/AI systems or plan to adopt such systems within five years (see Figure 1). In our survey we define cognitive/AI systems as the following:

> A set of technologies that use natural language processing, speech recognition, machine learning, knowledge graphs, and other technologies to answer questions, discover insights, and provide recommendations. These systems hypothesize and formulate possible answers based on available evidence, can be trained through the ingestion of vast amounts of content, and automatically adapt and learn from their successes and failures.

FIGURE 1

Cognitive/AI Systems Adoption Status

Q. Which of the following best reflects your organization's status regarding cognitive/AI systems?



n = 301

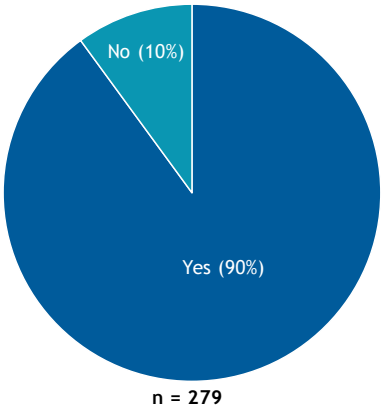
Source: IDC's *Cognitive/AI Adoption Survey,* June 2017

Furthermore, nearly all of these organizations show strong interest in the use of open source deep/machine learning software as a key enabler of their cognitive/AI investments (see Figure 2).

FIGURE 2

Planned Use of Open Source

Q. Are you using or planning to use open source cognitive/AI or machine learning toolkits?



n = 279

Base = respondents planning to deploy cognitive/AI in five years or that have already deployed

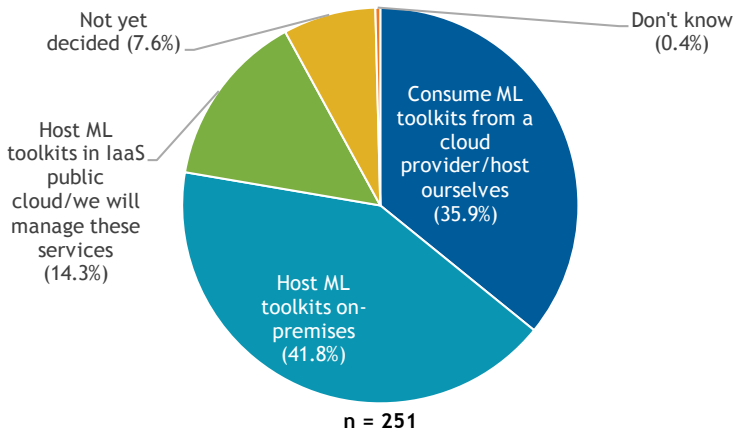
Source: IDC's *Cognitive/AI Adoption Survey,* June 2017

Exactly half of the survey participants cite cloud as the optimal deployment location for their open source machine learning applications (see Figure 3).

FIGURE 3

Open Source Toolkit Deployment Approach

Q. You indicated you are using or plan to use open source machine learning toolkits. Which of the following best describes how that will be done?



n = 251

Base = respondents using or planning to use open source machine learning toolkits

Source: IDC's *Cognitive/AI Adoption Survey,* June 2017

Long term, there will be additional shifts to public cloud infrastructure, as 48% of the survey participants are considering shifting more of their on-premises machine learning frameworks over to run in public cloud infrastructure. This is consistent with general trends we see with application deployments for other workload types. Developing AI-enabled applications that use machine learning and 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.

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.
- Lack of integrated development environments for machine 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 that is changing as is documented below.
- Absence of suitable developer skills in machine 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 machine 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.

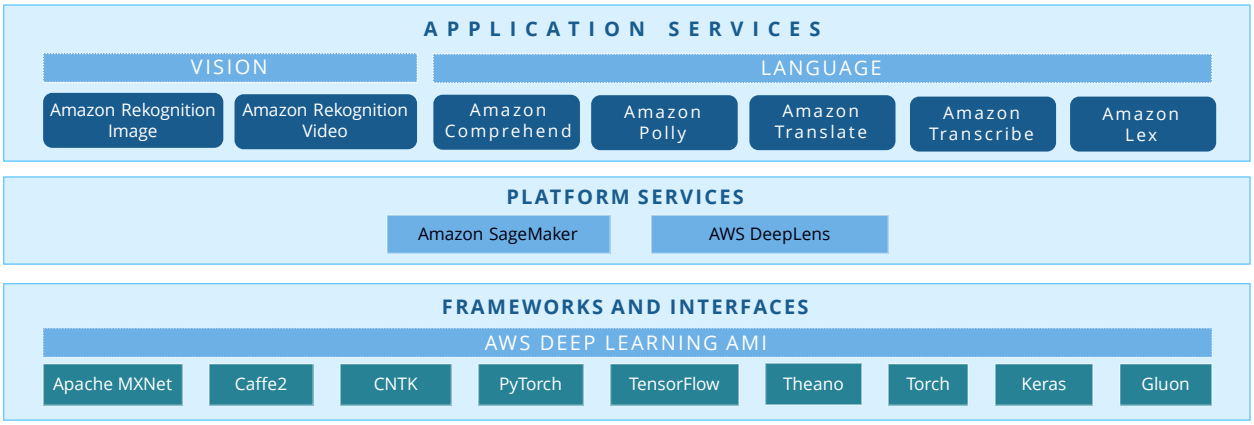
The question for enterprises is how best to develop these machine learning models while minimizing the amount of effort and time needed to develop accurate predictive and prescriptive models.

CONSIDERING AWS MACHINE AND DEEP LEARNING SERVICES

AWS has made it much easier for machine/deep learning developers by offering a broad range of tools and services to get deep learning models created and into production. These tools range from infrastructure like easily usable GPU services through Amazon Elastic Compute Cloud (EC2) P3 instances, which take advantage of NVIDIA Volta architecture, to frameworks such as TensorFlow and Apache MXNet, platforms like Spark and Amazon EMR, and the newly announced Amazon SageMaker. They also include deep learning-based services such as Amazon Comprehend, Amazon Transcribe, and Amazon Rekognition. The architecture for these services can be seen in Figure 4. 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.

FIGURE 4

AWS Deep Learning Architecture



Source: AWS, 2018

Amazon SageMaker: A Fully Managed Service for Machine Learning

To ease development, training, and deployment of machine learning models, AWS has recently created Amazon SageMaker. SageMaker is a fully managed service that enables developers and data scientists to quickly and easily build, train, and deploy machine learning models at any scale, removing all the barriers that typically slow down developers who want to use machine learning today. SageMaker removes the complexity that holds back developer success with each of these steps. SageMaker includes modules that can be used together or independently to build, train, and deploy machine learning models.

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 machine learning and deep learning frameworks like TensorFlow and Apache MXNet. 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 that 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. You 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 like identifying customer groupings based on purchasing behavior.

SageMaker also reduces the amount of time spent tuning deep learning models. It can automatically tune your model by adjusting thousands of different combinations of algorithm parameters to arrive at the most accurate predictions the model is capable of producing. This can save days of manual trial-and-error adjustments.

After training, SageMaker provides the model artifacts and scoring images to the developer for deployment to EC2 or anywhere else. The developer then can 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.

SageMaker eliminates code updates to incorporate new models. Amazon SageMaker also includes built-in A/B testing capabilities to help users test models and experiment with different versions to achieve the best results.

Finally, one of the best aspects of SageMaker is its modular architecture. Developers can use any combination of its building, training, and hosting capabilities to fit the organization's workflow. With

SageMaker, developing and deploying a machine 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. 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.

AWS Deep Learning AMIs

Amazon also offers its AWS Deep Learning AMIs as another option for enterprises and their developers. Built on top of EC2, the AWS Deep Learning AMIs offer a full environment for building, training, and running deep learning applications. The AMIs come with pre-installed, open source deep learning frameworks including Apache MXNet and Gluon, TensorFlow, Microsoft Cognitive Toolkit, Caffe, Caffe2, Torch, PyTorch, and Keras. 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 who 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 as a way to deploy applications more quickly.

CHALLENGES AND OPPORTUNITIES

Today, many organizations that weren't early adopters of machine learning are unsure as to where the technology will deliver the best business benefits. They're also unsure about what skill sets are needed to build and deploy AI-enabled applications. Finally, organizations are still trying to determine the right tools, infrastructure, and environments that are needed to put these AI-enabled applications to use.

Organizations need guidance about what types of tools and technologies can help them develop 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 (ROI) 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 concerns in the market for 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/deep learning services as the foundation of its solutions, this

approach is also not new to this market. What is new is that managed services like 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.

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 develop successful models and products based on deep learning quickly. 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 like Amazon SageMaker, data scientists and developers (and even nondevelopers) can be even more productive than they could be with just templates and sample code.
- **Assistance with data.** Some vendors are either providing curated third-party data sets or are 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. AWS also hosts a number of open public data sets on a variety of topics (geospatial, environmental, genomics, life science, 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/deep learning services like Amazon SageMaker, a little education can make nontraditional developers and data scientists productive and able to build and deploy their own deep learning models. For example, AWS offers self-paced training and education in machine learning. It's even released a deep learning–enabled videocamera for developers to learn about building AI applications called DeepLens.
- **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're consuming. An example of this is the new AI and machine learning competency for the AWS Partner Network. Amazon is certifying partners in machine/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 Solutions Lab pairs enterprise teams with AWS machine learning experts to prepare data, build and train models, and put models into production.

Great AI-enabled applications require both advanced technology and solid design judgment. Organizations should make sure the AI-enabled solution that they're building will be able to help achieve the desired business outcome and/or address the issues that it 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's not always clear which services or libraries are best for the use cases or jobs that developers should accomplish to successfully develop AI-enabled applications. Offerings such as Amazon SageMaker and the AWS Deep Learning AMIs provide the ways and means for developers to become more productive and deploy deep learning models, as well as supporting 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 AI-enabled applications using deep learning models.

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.

Global Headquarters

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