insideHPC

insideHPC Special Report

Riding the Wave of Machine Learning & Deep Learning

Written by Daniel D. Gutierrez, Managing Editor, insideBIGDATA

BROUGHT TO YOU BY DELL EMC | NVIDIA

Riding the Wave of Machine Learning and Deep Learning

Artificial intelligence (AI) and machine learning — decades-old technologies that are now electrifying the computing industry — for all intents and purposes, seem to be in the process of transforming corporate America. But why is AI so hot right now? Many experts believe it's because, after 50 years of promises that AI was going to solve critical problems, it's finally working.

We now have applications where everyone who uses a mobile phone or search on the Internet realizes something has changed. The reason is a convergence of multiple forces. First, we have a lot more data. Second, the computing power is just amazing. And third, are incredible breakthroughs in some aspects of AI, especially deep learning. Now, all of a sudden, things are actually starting to come together.

With AI, machine learning and deep learning, we're just scratching the surface of what's possible. That's why researchers at top universities worldwide, large enterprises, and a host of startups are rushing to put deep learning to work.

> "I think that deep learning is going to be applied in a lot more applications over the next several years," said Bryan Catanzaro, Ph.D., Vice President, Applied Deep Learning Research at NVIDIA.
>
> "One of the things that's really exciting about deep learning is that it's taking a lot of problems that used to be very specialized and… transferring them into an engineering problem, so that people even without a lot of domain expertise are able to solve hard problems in many different domains and, because of this, deep learning is going to be applied very widely in a lot of surprising places...
>
> I think a lot of domains are really important, especially domains involving text and speech, because that's kind of the primary modalities that humans use to interact with each other. Deep learning is going to give us the ability to understand that and to generate content." [1]

[1] Why is Deep Learning Hot Right now?, NVIDIA, https://youtube/1G0e-mR9a4k

Contents

Widespread adoption of cognitive systems and AI across a broad range of industries will drive worldwide revenues from nearly $8.0 billion in 2016 to more than $47 billion in 2020 and will experience a compound annual growth rate (CAGR) of 55.1% over the 2016-2020 forecast period.[2] This rapidly expanding market has a huge potential across various industry verticals. The major factors driving the AI market globally are the growing number of applications of the technology in various end-user verticals and the growing adoption of AI for the improvement of consumer services. The growth of the AI market is also driven by the development of IT infrastructure and the penetration of smartphones and smart wearables in countries such as India and China.

Widespread adoption of cognitive systems and AI across a broad range of industries will drive worldwide revenues from nearly $8.0 billion in 2016 to more than $47 billion in 2020 and will experience a compound annual growth rate (CAGR) of 55.1% over the 2016-2020 forecast period.[2] This rapidly expanding market has a huge potential across various industry verticals.

The purpose of this white paper is to make a solid case for why it's important to get off the fence now with respect to AI, machine learning and deep learning and to engage a strategic proof-of-concept project today in order to start a path in your organization for adoption of this technology. The specter of established or new technology companies using AI to enter and disrupt their markets is leading many businesses to advance their own AI plans. Delaying these plans will make businesses vulnerable to new tech entrants. Waiting too long may serve to clear the path for competitors to gain a strategic edge with all the technology can deliver. But fortunately, along with the fear of disruption comes the hope of handsome returns.

How Traditional Industries Are Using Machine Learning and Deep Learning to Gain Strategic Business Insights

The impact of AI on business functions will play out differently in different industry sectors. In financial services firms, AI's impact is expected to be felt most strongly in the area of customer interaction, and in manufacturing organizations — product development. Health and life sciences, meanwhile, anticipate the AI impact will be greatest in management decision-making. In general, early AI implementations may help some organizations to boost revenue, their operating efficiency or their margins.

Many enterprises are now embracing an accelerated path for accessing, and processing, and taking advantage of data by using AI — machine learning or deep learning, in other words — becoming an "AI enterprise." If you're still on the fence, it's clear that the time for action is now. Why? Because this is the year we're going to cross into the zettabyte regime in terms of data volume. So, with the amount of data that's at our disposal, coupled with the intended applications of the data, enterprises are looking for new ways to manage it all. The whole idea of drowning in your "data deluge," predicates that enterprises look for new solutions. The motivation is to do more with valuable data assets — start using AI, machine learning and deep learning. The whole mindset is changing from being overwhelmed by the data deluge, to actually being data hungry. AI is opening an insatiable desire for data, and this is happening now.

One area that appears to be taking off is how enterprises are now seriously looking at "accelerated analytics" and "AI driven analytics" as a solution. "Accelerated analytics" involves graphics processing unit (GPU)-accelerated solutions, like Kinetica, MapD, SQream DB, BlazingDB, Graphistry, and BlazeGraph. "AI driven analytics" addresses the needs of the enterprise's digital business and of getting access to information. Accelerating the analytics component is one side of the equation, the other side is doing it at a fraction of the cost.

[2] IDC "Cognitive/Artificial Intelligence Systems Spending Guide, Version 2016 H1.

The industry is in a dynamic feedback loop state where a continuing stream of successful proof-of-concept projects drive more and more enterprises to consider AI solutions.

There is a striking importance of evaluating and adopting AI, machine learning and deep learning technologies now or continuing to lose ground to competitors that already are in the process of doing so. The number of examples of AI successfully solving complex problems grows every day. This means that the industry is in a dynamic feedback loop state where a continuing stream of successful proof-of-concept projects drive more and more enterprises to consider AI solutions.

Below, we'll highlight a couple of compelling use case examples focusing on the successful use of AI, machine learning and deep learning. The examples also illustrate how GPUs can be effectively combined with AI technology.

University of Pisa and Scuola Normale Superiore di Pisa

The Competence Centre at University of Pisa and Scuola Normale Superiore di Pisa has been created to respond to the rapidly growing need for cutting-edge infrastructure solutions, allowing university researchers to share and power their work, and visitors to get insights into the latest and most efficient infrastructure technology. Using standards-based architecture—from the latest in cloud computing to HPC solutions and data protection—the Centre can help organizations achieve key performance and strategic goals. The University of Pisa is using deep learning technologies and systems from Dell EMC for DNA sequencing and encoding DNA as an image. The researchers are using Dell EMC PowerEdge C4130 servers, with NVIDIA Tesla P100 accelerators and NVIDIA Deep Learning GPU Training System (DIGITS), for deep learning classification of DNA sequencing data.

Chinese Academy of Sciences

In China, Dell China is collaborating with the Chinese Academy of Sciences on a joint AI and advanced computing laboratory. This lab focuses on research and applications of new computing architectures in the fields of brain information processing and AI. Research conducted in the lab spans cognitive function simulation, deep learning, brain computer simulation, and related new computing systems. The lab also supports the development of brain science and intellect technology research, promoting Chinese innovation and breakthroughs at the forefront of science. In fact, Dell China was recently honored with an "Innovation Award of Artificial Intelligence in Technology & Practice" in recognition of the collaboration.[3]

Machine learning algorithms can dig through mountains of data to ferret patterns that might not otherwise be recognizable. Moreover, machine learning algorithms get better over time, because they learn from their experiences.

Drilling Down into Machine Learning and Deep Learning

This is the year of artificial intelligence and deep learning, when the technology is coming into its own for mainstream businesses. AI-based tools are pouring into the marketplace, but machine learning has been around for decades. So, why does this matter? In short, because we need to gain insights from massive amounts of data, and this process requires systems that exceed human capabilities. Machine learning algorithms can dig through mountains of data to ferret patterns that might not otherwise be recognizable. Moreover, machine learning algorithms get better over time, because they learn from their experiences.

[3] http://en.community.dell.com/techcenter/high-performance-computing/b/general_hpc/archive/2016/11/29/dell-china-receives-ai-innovation-award

Machine learning is a subclass of AI techniques which automate the learning process through algorithms and high-powered data analytics. In recent years, advances in data science, combined with significant increases and declining cost of computing power, have yielded swelling data lakes and increasingly sophisticated analytics that have, for all practical purposes, made machine learning and AI business-ready.

Deep learning in turn, is subclass of machine learning that creates machines that use methods originally inspired by how a cat's brain reacted with light signals and then generalized to mimic the human brain's ability to learn. Until recently, we simply didn't have enough data and processing power to train a machine to learn. Deep neural networks (DNNs) learn at many levels of abstraction, ranging from simple concepts to complex ones. This is what designates the "deep" in deep learning. Each layer in the neural network categorizes some kind of information, refines it, and passes it along to the next layer. Deep learning lets the machine use this progression to build a hierarchical representation.

As an example, the first layer for a facial recognition system might look for simple edges. The next layer might look for collections of edges that form simple shapes like rectangles or circles. The third layer might identify features like eyes, ears and noses. After five or six layers, the neural network can position these features together. The result is a machine that can recognize certain objects or even concepts dependent on the training data set.

Humans learn the numbers 0 through 9 at about three to five years of age with the help of parents and teachers. Computers can learn numbers with a CPU, recognizing as many as a hundred numbers in a day. With GPUs, on the other hand, they can do this in an hour. That's because GPUs[4] perform many calculations at once or in parallel and, once the system is trained with GPUs, scientists, researchers and enterprise practitioners can put that learning to work.

Deep learning already has countless applications ranging from autonomous vehicles, medical diagnostics and robotics, to visual perception (such as image and video classification), speech recognition, language translation and text recognition, just to name a few.

"We've had a lot of success in deep learning and industrial applications with supervised learning solutions," said Yoshua Bengio, Ph.D., Professor of Computer Science and Operations Research at Université de Montréal.

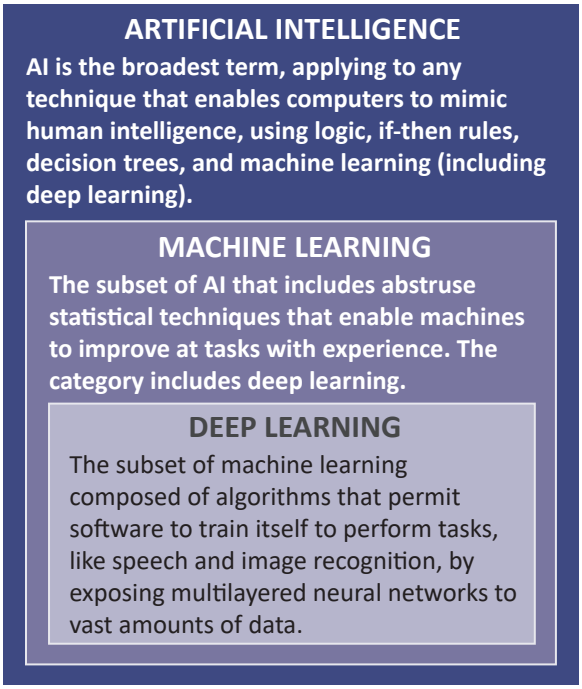
"With supervised learning, we can guide the computer, take it by the hand and tell it, for this example you should do this, for this example you should do that, and it requires a lot of human labor. It works for many data sets that we're doing well on now for images and speech, and even some are natural language.

But if we really aim for AI, we need to be able to let the computer learn more by itself. We can still give it these human labeled examples, but it should be able to use a lot more data. A two-year-old child hasn't been supervised that much, and he or she was able to learn just by observing and experimenting with the world. That's what we're trying to aim with unsupervised learning." [5]

[4] https://www.youtube.com/watch?v=-P28LKWTzrI
[5] Why is Deep Learning Hot Right now?, NVIDIA, https://youtube/1G0e-mR9a4k

Differentiating between AI, machine learning and deep learning

With all the quickly evolving jargon in the industry today, it's important to be able to differentiate between AI, machine learning and deep learning. The easiest way to think of their relationship is to visualize them as a concentric model, as depicted in the figure to the right, with each term defined. Here, AI — the idea that came first — has the largest area, followed by machine learning — which blossomed later and is shown as a subset of AI. Finally, deep learning — which is driving today's AI explosion — fits inside both.

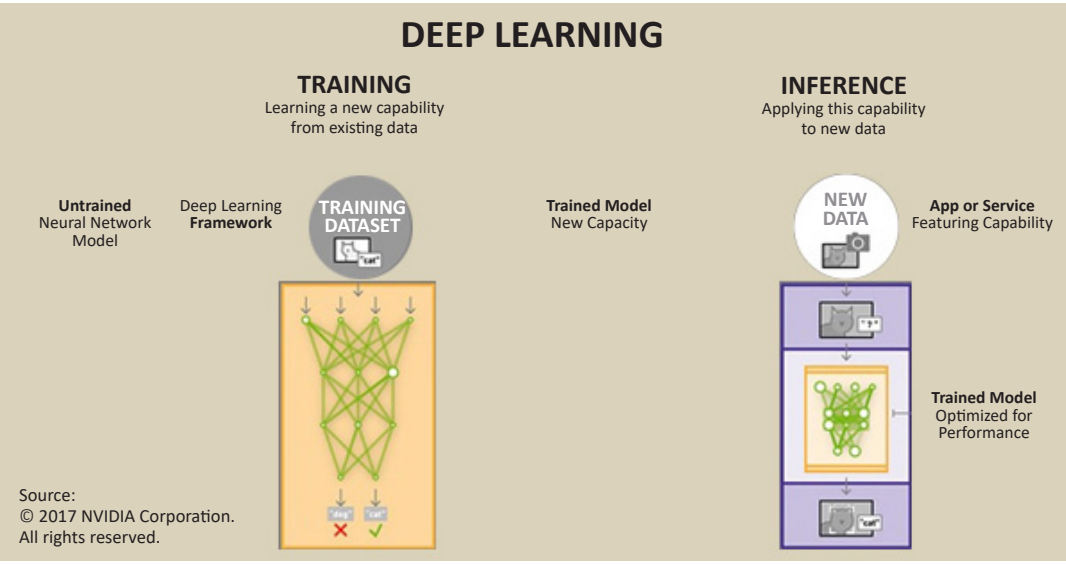


Credit: Fortune

Machine learning takes some of the core ideas of AI and focuses them on solving real-world problems with neural networks designed to mimic our own decision-making. Deep learning focuses even more narrowly on a subset of machine learning tools and techniques, and applies them to solving just about any problem which requires "thought" — human or artificial.

Machine learning is well-suited for problem domains typically found in the enterprise, like making predictions with supervised learning methods (e.g. regression and classification), and knowledge discovery with unsupervised methods (e.g. clustering). Deep learning is an area of machine learning that has achieved significant progress in certain application areas that include pattern recognition, image classification, natural language processing (NLP), autonomous driving, and so on. Machine learning techniques like random forests[6] and gradient boosting[7] often perform better in the enterprise problem space than deep learning.

Deep learning attempts to learn multiple levels of features of large data sets with multi-layer neural networks and make predictive decisions for the new data. This indicates two phases in deep learning: first, the neural network is "trained" with a large number of input data; second, the trained neural network is used for "inference" designed to make predictions with new data. Due to the large number of parameters and training set size, the training phase requires tremendous amounts of computation power. The figure below summarizes the roles of training versus inferencing in deep learning.



Inference is where capabilities learned during deep learning training are put to work.

[6] https://en.wikipedia.org/wiki/Random_forest [7] https://en.wikipedia.org/wiki/Gradient_boosting

Deep learning "why now?"

Perhaps the biggest question surrounding the upsurge momentum for AI is "Why now?" The answer centers upon the opportunity that AI represents, as well as the reality of how many companies are afraid to miss out on potential benefit.

Two key drivers of AI progress today are: scale of data, and scale of computation. It was only recently that technologists figured out how to scale computation to build deep learning algorithms that can take effective advantage of voluminous amounts of data.

AI has been part of our thoughts and has been slowly evolving in academic research labs since a group of computer scientists first defined the term at the Dartmouth Conferences in 1956 and provided the genesis of the field of AI. As an extension, machine learning techniques like artificial neural networks were widely used in the 1980s and early 1990s; but for various reasons, their popularity diminished in the late 1990s.

More recently, neural networks have had a major resurgence. A central factor for why their popularity waned is because a neural network is a computationally expensive algorithm. Today, computers have become fast enough to run large-scale neural networks. Since 2010, advanced neural networks have been used to realize methods referred to as deep learning.

One of the big reasons why AI is on its upward trajectory is the rise of relatively inexpensive compute resources. Now, with the adoption of GPUs (the graphics processing unit originally designed 10 years ago for gaming), neural network developers can run deep learning with the compute power required to bring AI to life quickly. The "Enabling Hardware Technologies" section to follow highlights important options for implementing AI applications.

What Does It Take to Get Started?

In this section, we'll start off by providing a handy five-step enterprise AI strategy designed to ensure your early AI deployment projects are a success. We'll also highlight several hardware and software tracks that will assist you along the way.

1. **Get familiar with the technology –** Take the time to become familiar with what modern AI can do. A good way to acquire this knowledge is to develop a close partnership with your chosen technology vendor.
2. I**dentify the problems you want AI to solve –** Once you're up to speed on the basics, the next step for any business is to begin exploring different ideas. Think about how you can add AI capabilities to your existing products and services. More importantly, your company should have goals in mind of specific use cases in which AI could solve business problems or provide demonstrable value.
3. **Prioritize concrete value –** Next, you need to assess the potential business and financial value of the various possible AI implementations you've identified. It's easy to get lost in "pie in the sky" AI discussions and it's important to tie your initiatives directly to business value.
4. **Acknowledge the internal capability gap –** There's a stark difference between what you want to accomplish and what you have the organizational ability to actually achieve within a given time frame. A business should know what it's capable of and what it's not from a technological and business process perspective before launching into a full-blown AI implementation.
5. **Bring in experts and set-up a pilot project –** Once your business is ready from an organizational and technological standpoint, then it's time to start building and integrating. The most important factors here are to start small, have project goals in mind, and especially to be aware of what you know and what you don't know about AI. This is where bringing in outside experts or AI consultants from a technology partner can be invaluable.

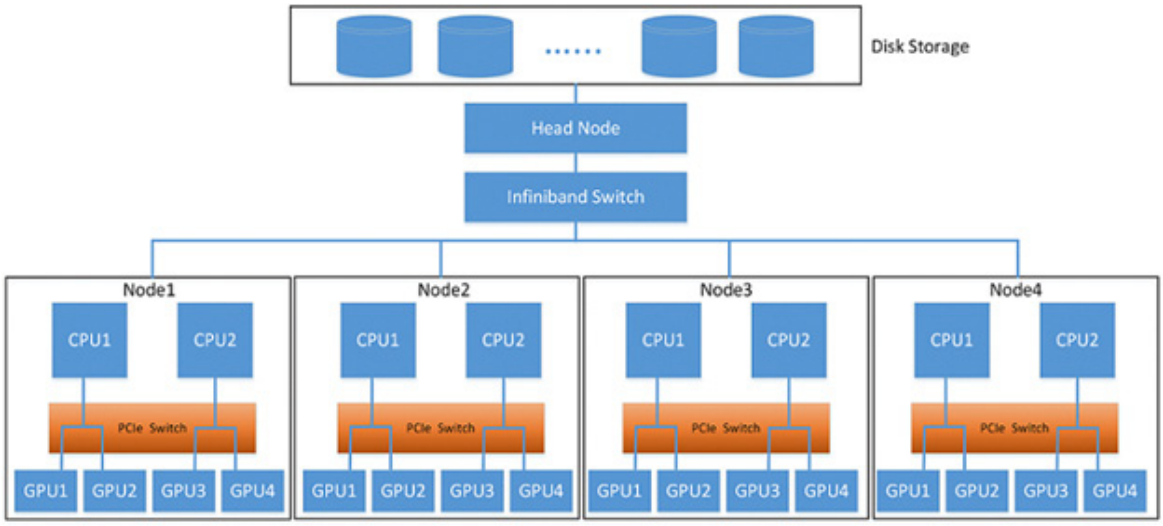
Enabling Hardware Technologies

GPU acceleration is a significant advantage for addressing the needs of deep learning training and inferencing (see the "Differentiating between AI, machine learning and deep learning" sidebar, page 5). The GPU is a massively parallel architecture that employs thousands of small but efficient cores designed to accelerate computational intensive tasks.

For example, the NVIDIA® Tesla® P100™ GPU uses the Pascal™ architecture to deliver very high performance (higher performance than hundreds of slower commodity nodes) for HPC and hyper-scale workloads. In PCIe-based servers, NVIDIA Tesla P100 delivers around 4.7 and 9.3 TeraFLOPS of double and single precision performance, respectively. With NVLink™-optimized servers, NVIDIA Tesla P100 delivers around 5.3 and 10.6 TeraFLOPS of double and single precision performance, respectively. The high compute capability and high memory bandwidth make GPUs an ideal candidate to accelerate deep learning applications, especially when powered with NVIDIA's Deep Learning software development kit (SDK) that includes CUDA® Deep Neural Network library (cuDNN), a GPU-accelerated library of primitives for deep neural networks, TensorRT™, a high performance neural network inference engine for production deployment of deep learning applications, and CuBLAS a fast GPU-accelerated implementation of the standard basic linear algebra subroutines.

Technologies for AI compute include extreme GPU computing with the Dell EMC PowerEdge C4130 Server with support for NVIDIA Tesla P100 GPU accelerator and NVLink High-Speed Interconnect (NVIDIA Tesla P100 SXM2 accelerators); and also the Dell EMC PowerEdge R730 Server with the Tesla P4 accelerator option. As a result, the performance of deep learning frameworks using NVIDIA' Tesla's P100 GPU and Dell EMC's PowerEdge C4130 server architecture is extra-ordinarily high. The Figure below shows a testing cluster that includes one head node, which is Dell EMC's PowerEdge R630 Server, and four compute nodes, which are Dell EMC's PowerEdge C4130 servers. All nodes are connected by an InfiniBand EDR network, sharing disk storage through Network File System (NFS). Each compute node has two CPUs and four NVIDIA Tesla P100 GPUs. All of the four compute nodes have the same configurations.

Overall, there was great speedup and scalability[8] in neural network training when multiple NVIDIA Tesla P100 GPUs were used in Dell EMC's PowerEdge C4130 server and multiple server nodes were used. The training speed increased and the training time decreased as the number of NVIDIA Tesla P100 GPUs increased. It is clear that Dell EMC's PowerEdge C4130 server cluster is a powerful tool for significantly speeding up neural network training.



Testing Cluster for Deep Learning

[8] Deep Learning Performance with NVIDIA Tesla's P100 GPUs, http://en.community.dell.com/techcenter/high-performance-computing/b/general_hpc/archive/2016/11/11/deep-learning-performance-with-p100-gpus

Enterprise Software Tools for AI

There are three exemplary members of the AI software stack available as deep learning frameworks: Caffe, MXNet and TensorFlow.

- **Caffe** is a well-known and widely used deep learning framework which was developed by the Berkeley Vision and Learning Center (BVLC) and by community contributors. It focuses more on the image classification problem and it supports multiple GPUs within a node.
- **MXNet,** jointly developed by collaborators from multiple universities and companies, is a lightweight, portable and flexible deep learning framework designed for both efficiency and flexibility. This framework scales to multiple GPUs within a node and across nodes.
- **TensorFlow,**developed by Google's Brain team, is a library for numerical computation using data flow graphs. TensorFlow also supports multiple GPUs and can scale to multiple nodes.

NVIDIA DIGITS puts the power of deep learning into the hands of engineers and data scientists. DIGITS can be used to rapidly train highly accurate deep neural networks (DNNs) for image classification, segmentation and object detection tasks. DIGITS simplifies common deep learning tasks such as managing data, designing and training neural networks on multi-GPU systems, monitoring performance in real time with advanced visualizations, and selecting the best performing model. The system is completely interactive, so that data scientists can focus on designing and training networks rather than programming and debugging.

Summary

Many companies are moving decisively to develop capabilities based on AI, machine learning and deep learning. In time-honored business fashion, the motivation is a combination of fear and hope. Competitive pressures are spurring companies on, and there is a sense of urgency amongst many enterprise thought leaders about not falling behind.

AI, machine learning and deep learning are transforming the entire world of technology, but these technologies are only making headway now due to the proliferation of data and the investments being made in storage, compute and analytics solutions. Much of this progress is due to the ability of learning algorithms to spot patterns in larger and larger amounts of data.

At first glance, when looking out over the global business landscape, some companies might be considered as "under-investing" in computer systems for AI. But maybe this observation could be more of a long-term investment strategy, since you have to do a lot of work to make AI solutions more productive. It is possible that some companies are still learning how to make more long-term rather than short-term investments in technology. Whether you ascribe to a long-term or short-term technology investment philosophy, there is a sense of urgency permeating in the enterprise world right now to get under way with AI. First steps should include the "Five-step enterprise AI strategy" in the previous section, but early efforts shouldn't be limited to these action items. The evolution toward becoming an AI enterprise should be transformative with touch points across the organization. Making this transformation as seamless as possible will take commitment, understanding and strong relationships with technology partners.

AI is an amazing tool set that is helping people create exciting applications and creating new ways to service customers, cure diseases, prevent security threats, and much more. Rapid progress continues to unlock more and more opportunities for enterprises and scientific research where AI can make a big impact. Many believe that the real world potential for AI is highly promising. Speaking at a 2016 AI conference in London, Microsoft's Chief Envisioning Officer, Dave Coplin observed "This technology will change how we relate to technology. It will change how we relate to each other. I would argue that it will even change how we perceive what it means to be human." Apparently, the best is still to come.

Additional Resources

- Game-changing Extreme GPU computing with The Dell PowerEdge C4130
- Deep Learning Performance with NVIDIA Tesla P100 GPUs
- NVIDIA Deep Learning
- NVIDIA Deep Learning Blogs
- GPU Accelerated NAMD
- Dell HPC
- Dell Blueprint for HPC
- New Tech and Systems Tuning in the Dell HPC Innovation Lab
- Capitalizing on machine learning—from life sciences to financial services