insideHPC

insideHPC Special Report

Advancing the Financial Services Industry Through Machine Learning

Written by Beth Harlen

BROUGHT TO YOU BY DELL EMC | NVIDIA

Introduction

Competitive advantage, minimized risk, and improved bottom lines are just some of the benefits leading financial services institutions to invest heavily in artificial intelligence. A collection of technologies, artificial intelligence dates back to the 1950s and the idea that a computer could be made to simulate not only human intelligence but also our ability to learn.

Artificial intelligence, or cognitive computing, offered a glimpse at a future where machines could essentially 'learn' and improve without the need for explicit programming at every step along the way. Machine learning is a subclass of AI whereby a computer can take large amounts of data and use it to begin to recognize patterns and, within a set of specific parameters, make predictions on new data — prices within the stock market, for example. It's worth emphasizing that these predictions can take into account variables that people would struggle to apply.

As financial institutions look to be empowered through machine learning, they should first acknowledge the benefits, challenges, and considerations involved. This guide is essential reading for anyone involved in the financial services industry, from those who are beginning to explore the potential of machine learning, to those looking to expand and maximize its use. We will identify and explore the critical challenges within this industry, and the steps you need to take to ensure your institution is ready to leverage machine learning to meet those challenges.

We'll present an overview of the technologies you need to know about, the resources and communities on-hand to offer guidance, and we'll explore a case study highlighting the process of implementing machine learning.

This report will help you...

- Explore key machine learning innovations in financial services
- Understand what's needed to leverage machine learning
- Know what applications and technologies to use
- Learn from organizations' successes
- Connect with the community

CONTENTS

The Financial Services Industry

Trends

From small banks to large financial institutions, there is a growing and significant level of interest in machine learning. Applications range from management and improvement of the customer experience to decision-making on the trading floor. These applications have been fueled by a fall in the cost of computing power and data storage, coupled with a rise in the volume and variety of available data. The proliferation of open source tools and active communities are also enabling machine learning to add value through its ability to evolve and adapt.

Algorithms applied to trading problems can use current daily data, like stock prices, to ensure that opportunities are seized and risks avoided. Increasingly, financial institutions are also recognizing that machine learning is perfectly suited to the automation of repetitive, and often costly processes.

High-frequency trading is a good example of how machine learning has evolved within the financial services industry to take influencing factors, such as social media feeds, then analyze, interpret, and use them to implement trading strategies at a fraction of the time it would take a human trader. Furthermore, algorithms applied to trading problems can use current daily data, like stock prices, to ensure that opportunities are seized and risks avoided. Increasingly, financial institutions are also recognizing that machine learning is perfectly suited to the automation of repetitive, and often costly processes. But to fully appreciate the breath of applications, we should first examine the specific challenges within financial services.

Challenges

Operating within one of the most competitive and highly regulated industries, banks, insurance companies, and other financial institutions face significant challenges. The most critical of these are regulation and compliance, cybersecurity and fraud detection, risk management, and competition.

The large volume of data involved in the area of regulation and compliance makes it an ideal application for machine learning.

Regulation and compliance

Through the application of machine learning, data from large or disparate systems within a financial services institution can be collated, analyzed and interpreted. Central functions, such as the quarterly close, can be handled with greater speed and accuracy than would be humanly possible. As this process can be undertaken in near real-time, any irregularities or issues can be identified early on, and adjustments made accordingly. Regulation and compliance requirements, such as traceability and insight into how decisions are being made, can also be met with ease. The large volume of data involved in this area makes it an ideal application for machine learning.

Cybersecurity and fraud detection

Early detection of internal and external threats can be invaluable in protecting a company's reputation and bottom line. Historically, fraud detection systems have relied on applying available data to a robust set of rules to identify and flag concerns. Through their ability to learn, machine learning-enabled systems go beyond that. These systems are able to take information gathered from previous threats and use it to adapt to and anticipate new threats.

Voice recognition has been successful in preventing fraud — the monitoring of traders' phone conversations to prevent illegal acts, such as insider trading, is a good example of its use.

Fraud related to credit card use or identity theft can be tackled head-on by ensuring any unusual transaction patterns are identified early. This automated monitoring is highly efficient, and the results can be fed back into the algorithm to generate a more accurate profile of the customer. This, in turn, refines the ability to identify suspicious activity in the future.

The same process is used to detect insurance fraud, as information pertaining to previous claims is instantly applied. Voice recognition has also been successful in preventing fraud — the monitoring of traders' phone conversations to prevent illegal acts, such as insider trading, is a good example of its use.

Risk management

Minimizing the risk that banks and other financial institutions are exposed to is another ideal task for machine learning-enabled systems. These systems can take a holistic approach to vast quantities of structured and non-structured data in order to assess risk and facilitate improved human decision making.

The improved accuracy of the machine learning risk models is partly driven by its ability to collect data from multiple sources in real-time. For example, predictive models can take fluctuating market prices, negative press, and many other factors into account. Based purely on the data, these models improve with each new piece of information.

Another application in this area is investment research. Customer opinions can be garnered from direct communication, but also through the monitoring of mentions across external channels, such as social media. Gaining insight into how the public perceives a particular company, for example, can influence the risk associated with investing in that company — for better or worse.

Competition

This is a competitive industry, and one significant source of competitive edge is through improved customer engagement and retention. That, of course, sounds far easier than it actually is. Savvy consumers have come to expect tailored offers, recommendations, instant responses, and excellent service. Critically, within the financial services industry, customers also require a good level of confidence and trust. This can be achieved in several ways.

It's important to note at this point that the successful use of machine learning is dependent upon several factors, including the integrity of the data, organizational readiness, and the deployment of an appropriate solution suited to the individual use case.

Take customer service, for example. When a customer sends an email complaining about an area of their experience, they expect that the appropriate person will receive that email and be able to promptly deal with the complaint. It's therefore necessary to have a system in place that can analyze the sentiment contained within high volumes of emails, categorize those emails, and deliver them to the correct internal person.

Personal assistants and 'Bank Bots' can also be used to identify complaints, answer customer queries, and offer advice. Machine learning can ensure that advice is based on numerous factors, such as the customer's account activity and risk. It's important to note at this point that the successful use of machine learning is dependent upon several factors, including the integrity of the data, organizational readiness, and the deployment of an appropriate solution suited to the individual use case. Ensuring the correct tools and infrastructure are in place leads us to our next section — Finding a solution.

Finding a Solution

Overview

The successful application of machine learning can depend upon the following variables:

- Cost of ownership/development
- Organizational readiness
- Speed, capacity, accuracy
- Scalability

Most things in life inevitably come down to cost — and the deployment of advanced analytics is no different. Can you afford to replace your legacy systems, both in terms of the compute and the investment of personnel training, as new processes are introduced? As mentioned earlier, high volume, repetitive functions can benefit the most from machine learning — but how prepared is your organization to move those tasks away from employees? Then there's the hardware. Are you making the right choice? Is it scalable? Beyond that, you need to think about the level of advancement you need in the algorithms.

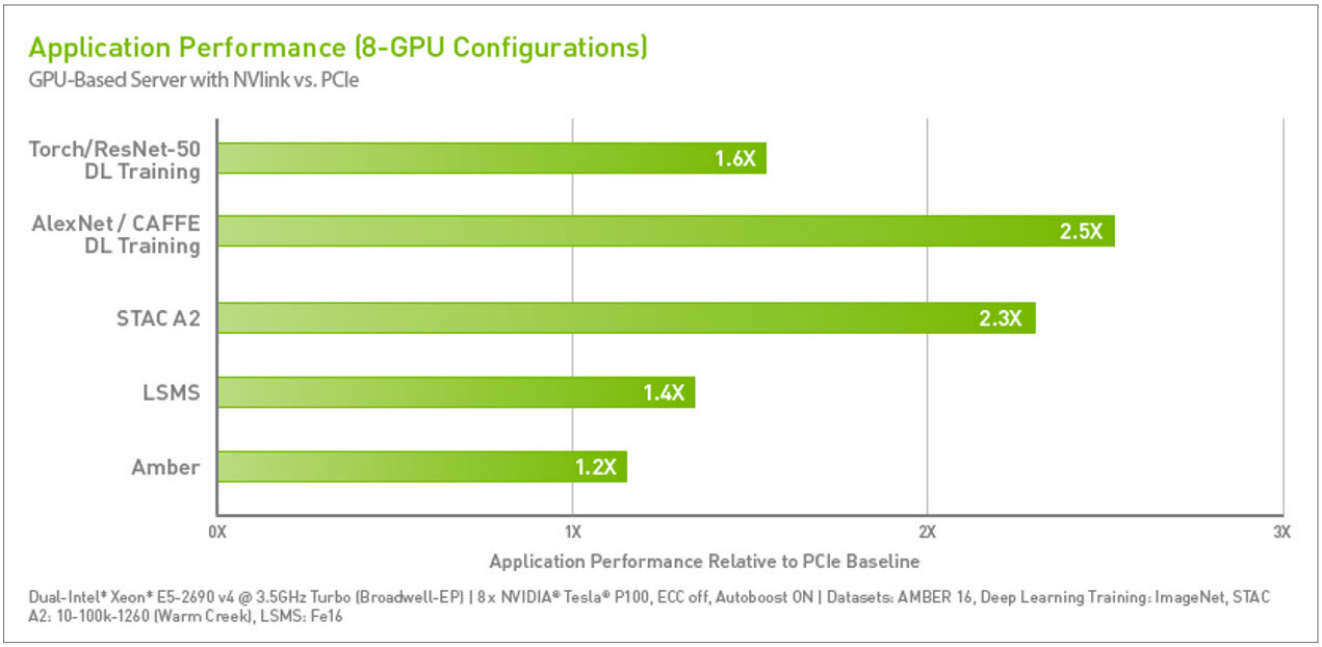
As you can see, there are many points that need to be considered. Further advice can be found in the section — What next? First, we'll take a look at the more technical aspects.

Technology

Of the array of technological innovations that have made machine learning possible, GPUs have perhaps had the most impact. GPUs, or graphics processing units, provide significant advantages through the acceleration of applications. Put simply, the analytics and process management you're trying to do can be done much, much faster through the use of GPUs — any GPUs, from small workstations to the more substantial hardware.

Throughput

Throughput is the rate at which data can be processed. It's an important consideration, because it dictates performance. NVIDIA® NVLink™ is a high-bandwidth, energy-efficient interconnect that supports ultra-fast communication between the CPU and GPU, and between GPUs. Next-generation NVIDIA NVLink high-speed interconnect technology delivers 2X the throughput, compared to the previous generation. This enables more advanced model and data parallel approaches for strong scaling to achieve the absolute highest application performance.



Dell EMC PowerEdge C4140 server

What is it?
The Dell EMC PowerEdge C4140 is a next-generation accelerator-optimized, 1U rack server designed for most demanding workloads.

What does it do?
The PowerEdge C4140 server accelerates most demanding workloads. It is ideal for cognitive workloads such as artificial intelligence, machine learning, deep learning ,and for technical computing in industry verticals such as financial services.

Tech specs and features

- Up to four 300W NVIDIA® Tesla® GPU accelerators in just 1U of rack space
- Two Intel® Xeon® Scalable processors, up to 20 cores per processor
- Red Hat® Enterprise Linux operating system
- Memory: Up to 24 DDR4 DIMMs in total; Supports RDIMM /LRDIMM speeds up to 2666, 1.5 TB Max ECC Registered DDR4
- Storage controllers: IDSDM or Internal M.2 Boot Module (2 x M.2) -120G or 240G (mirrored configuration)
- iDRAC9 with Lifecycle Controller
- Accelerators: NVIDIA® Tesla® P40, P100 12GB PCIe, P100 16GB PCIe and NVLinkTM, V100 16GB PCIe and NVLink GPUs
- Power: 2000W, 2400W hot-plug, redundant power supply unit (PSU)

How will it help me?
The PowerEdge C4140 and previous generation C4130 servers offer a combination of flexibility, efficiency and performance in a compact package that reduces cost and management requirements. The balanced architecture ensures that workload requirements are met through a flexible combination of accelerator, processor, memory and low latency I/O.

NVIDIA Volta

What is it?
Characterized as the driving force behind artificial intelligence, NVIDIA Volta is the most advanced GPU architecture in existence.

What does it do?
By pairing NVIDIA CUDA (the company's own parallel computing platform and programming model) and Tensor cores, Volta delivers the performance of an AI supercomputer.

Tech specs and features

- 640 Tensor Cores
- 125 Teraflops per second — an increase of more than 5x the performance of prior generation NVIDIA Pascal architecture
- 300 GB/s interconnect bandwidth — next generation NVIDIA NVLink delivering 2x the throughput of the previous generation
- Over 450 GPU-Accelerated frameworks and applications

How will it help me?
From recognizing speech to training virtual personal assistants, NVIDIA Volta meets the complex challenge of artificial intelligence and machine learning. Models that would consume weeks of computing resources on previous systems can now be trained in a few days. With this dramatic reduction in training time, a whole new world of problems will now be solvable.

STAC benchmark results

The STAC-A2 Benchmark suite is the industry standard for testing technology stacks used for compute-intensive analytic workloads involved in pricing and risk management.

Stack under test:

- STAC-A2 Pack for CUDA (Rev C)
- Red Hat Enterprise Linux Server release 7.3
- 4x NVIDIA Tesla P100 GPU Accelerator card
- 2x Intel Xeon E5-2690v4 @ 2.60GHz
- Dell EMC PowerEdge C4130 Server

The STAC Benchmark Council performed STAC-A2 (the technology benchmark standard based on financial market risk analysis) Benchmark tests on a stack consisting of the STAC-A2 pack for CUDA (Rev C) on a Dell EMC PowerEdge C4130 server with 4 x NVIDIA Tesla P100 GPU cards and 2 x Intel Xeon E5-2690v4 CPUs. Designed by quants and technologists from some of the world's largest banks, STAC-A2 reports the performance, scaling, quality, and resource efficiency of any technology stack that is able to handle the workload (Monte Carlo estimation of Heston-based Greeks for a path-dependent, multi-asset option with early exercise).

When the report was published on August 1, 2017, the solution set several new records:

Highest space efficiency (STAC-A2.β2.HPORTFOLIO.SPACE_EFF)

- 1.98x the efficiency of the previously tested system with 4 x P100 GPUs (NVDA161102)
- 1.95x the efficiency of the previous record holder (INTC170503)

Highest energy efficiency (STAC-A2.β2.HPORTFOLIO.ENERG_EFF)

- 12% higher than the previously tested system with 4 x P100 GPUs (NVDA161102)

Highest throughput in the portfolio benchmark (STAC-A2.β2.HPORTFOLIO.SPEED)

- 25.2 options per second

In *HPC Applications Performance on V100,* testing concluded that 'the C4130 server with NVIDIA Tesla V100 GPUs demonstrates exceptional performance for HPC applications that require faster computational speed and highest data throughput.'

Fastest performance in warm and cold runs of the large problem size Greeks benchmark

- 12.7 seconds (STAC-A2.β2.GREEKS.10-100k-1260.TIME.WARM)
- 13.5 seconds (STAC-A2.β2.GREEKS.10-100k-1260.TIME.COLD)

The full report can be read here.

Further benchmark results

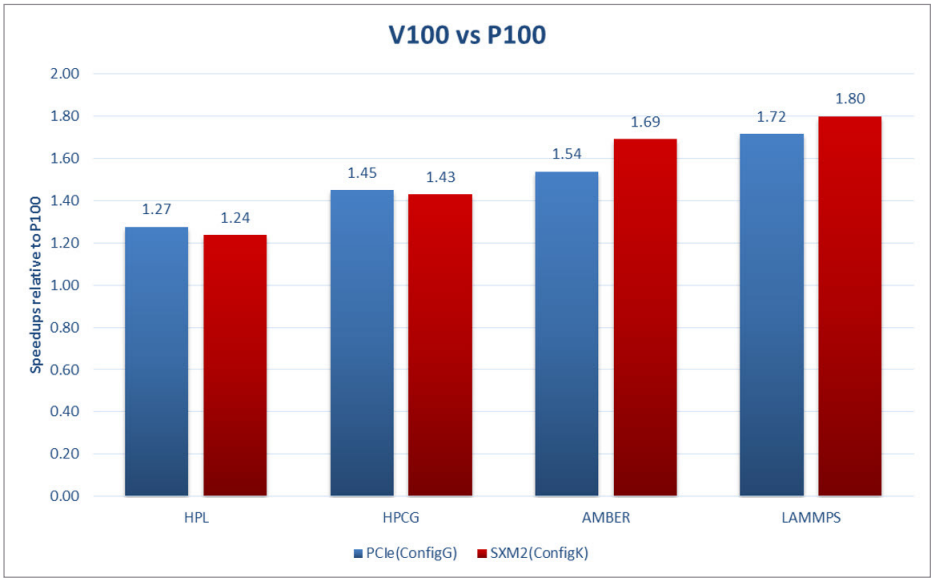
Dell EMC's HPC Innovation Lab has recently published two blog posts introducing the NVIDIA Tesla Volta-based V100 GPU, and presenting the initial benchmark results of those GPUs on four different HPC benchmarks, as well as a comparative analysis against previous generation Tesla P100 GPUs. While the focus is on HPC applications, they make for interesting reading in terms of the technological performance of the GPUs when dealing with speed and throughput.

In *HPC Applications Performance on V100,* authored by Frank Han, Rengan Xu, and Nishanth Dandapanthula, testing concluded that 'the C4130 server with NVIDIA Tesla V100 GPUs demonstrates exceptional performance for HPC applications that require faster computational speed and highest data throughput. Applications like HPL, HPCG benefit from the additional PCIe links between CPU and GPU that are offered by Dell PowerEdge C4130 configuration G. On the other hand, applications like AMBER and LAMMPS were boosted with C4130 configuration K, owing to P2P access, higher bandwidth of NVLink and higher CUDA core clock speed. Overall, a PowerEdge C4130 with Tesla V100 GPUs performs 1.24x to 1.8x faster than a C4130 with P100 for HPL, HPCG, AMBER and LAMMPS.'

The full post can be read here.

Deep Learning on V100, by the same authors, concluded that 'V100 is more than 40% faster than P100 in FP32 and more than 100% faster in FP16. This demonstrates the performance benefits when the V100 tensor cores are used.
In the future work, we will evaluate different data type combinations in FP16 and study the accuracy impact with FP16 in deep learning training. We will also evaluate TensorFlow with FP16 once support is added into the software. Finally, we plan to scale the training to multiple nodes with these frameworks.'

The full post can be read here.



Frameworks

The ease with which the machine learning application can be built and run is first determined by the framework you choose. Here is a roundup of five of the best-known.

Caffe

Developed by a mix of Berkeley AI Research and community contributors, Caffe has, according to its website, been 'made with expression, speed and modularity in mind'.

The good news is that the set of pre-trained models don't require coding to implement, and GPU training is supported out-of-the-box. The bad news is that multi-GPU training is only partially supported. The number of third-party packages used by Caffe can lead to version skew.

Platform	Linux, Mac OS X, Windows
Language	C++
Interface	Python, MATLAB
OpenMP support	YES
OpenCL support	In development
CUDA support	YES
Pre-trained models	YES
Community support	YES
Open source	YES

Tensorflow

Developed for different language understanding and perceptual tasks by the team at Google Brain, Tensorflow is the open source framework behind Google services like Gmail, and Google Search. It's ideal for second-order gradient differentiation, and computational graph visualizations, but developers may find it has limited debugging capabilities.

Platform	Linux, Mac OS X, Windows
Language	C++, Python
Interface	Python (Keras), C/C++, Java, Go, R
OpenMP support	NO
OpenCL support	Roadmapped
CUDA support	YES
Pre-trained models	YES
Community support	YES
Open source	YES

Torch

Used by companies like Facebook and Twitter, Torch originated at New York University (NYU) in 2002. Coded in a programming language called Lua, it's one of the simplest machine learning frameworks to set up and deploy. A bonus is that Lua is a user-friendly language that boasts a vast repository of sample code. Another plus is that Torch is available on mobile platforms, while a drawback is that learning materials can be hard to come by.

Platform	Linux, Mac OS X, Windows, Android, iOS
Language	Lua, C
Interface	Lua, LuaJIT, C, CUDA, Utility library for C++/OpenCL
OpenMP support	YES
OpenCL support	Third party
CUDA support	YES
Pre-trained models	YES
Community support	YES
Open source	YES

Microsoft Cognitive Toolkit (CNTK)

Cognitive Toolkit (CNTK) is an open-source machine learning framework from Microsoft. It's one of the most dynamic frameworks available, with support for algorithms such AS CNN, LSTM, RNN, Sequence-to-Sequence and Feed Forward. This also happens to be the only public toolkit that can scale GPUs beyond a single machine. The lack of OpenCL support is a drawback, however.

Platform	Windows, Linux, (OSX via Docker on roadmap)
Language	C++
Interface	Python, C++, Command line, BrainScript (.NET on roadmap)
OpenMP support	YES
OpenCL support	NO
CUDA support	YES
Pre-trained models	YES
Community support	YES
Open source	YES

Apache Mahout

Apache Mahout began as a no-cost open source project by the Apache Software Foundation. Deployed on top of Hadoop using the MapReduce paradigm, the goal was to develop free distributed or scalable machine learning frameworks. While this is a robust framework, it should be noted that it's still under development.

Platform	Cross-platform
Language	Jave, Scala
Interface	TBD
OpenMP support	YES
OpenCL support	In development
CUDA support	YES
Pre-trained models	YES
Community support	YES
Open source	YES

These are five of the frameworks, but you may still be wondering how to choose between them. For in-depth advice on all aspects of technologies involved in machine learning, contact details for experts at Dell EMC and NVIDIA can be found on page 11.

The Results

From fraud detection to a deeper level of customer engagement, many aspects within financial services have benefited from the application of machine learning. This case study from Mastercard is just one example.

Case Study: Mastercard

"Securing sensitive data: how to navigate the new world of compliance" is a white paper that, as the title suggests, delves into how Mastercard achieved the latest compliance standards and ensured optimum data management and security. Data security is big business, in fact International Data Corporation (IDC) forecast that worldwide revenues for security-related hardware, software and services will jump from $73.7 billion in 2016 to $101.6 billion in 2020. As highlighted in the white paper, 'evolving compliance standards demand a proactive capability for data security. For all the data available to potential hackers, there is an equal amount of fraud protection and risk measurement available with a compliant approach. In other words, technology has created data, created the ability for hackers to intrude, but has also strengthened the defences, if managed properly.'

Shirley Inscoe, an Aite banking analyst, is quoted as saying: "When it comes to using big-data technologies for fraud prevention and info-security…the big data rubber is hitting the road. Using Hadoop or similar technology can store and analyze data much more efficiently and in ways that were not previously possible. Processing data is much more efficient using these methods, dramatically reducing both cost and processing time."

"A handful of large FIs [financial institutions] have advanced big-data projects underway, leveraging Hadoop and machine-learning technologies to combine customer data across products and channels. The results are impressive: faster detection of merchant data compromise, more effective transactional fraud analytics, and fewer frustrated customer as a result of false positive declines."

– Shirley Inscoe, banking analyst, Aite

What Next?

The benefits of employing machine learning algorithms are clear. If, at this point, you're sold on the concept but in need of guidance, the following resources are at your disposal:

Community

The nature of machine learning makes it a fast-paced and innovative area. As the boundaries of capabilities are continually being pushed, a large and vocal community has formed to facilitate development, and offer support. Online resources range from white papers to forum blog posts, all containing opinions, debates, user experiences and expert advice. However, if you have a question that needs answering or are simply seeking out opinions, search engines can only provide so much.

By joining one of the many community resources, you'll be in direct contact with thought-leaders, technical experts and fellow users. There are even community resources that compile and summarize community resources. This blog post is a good example of a round-up of machine learning communities.

Dell EMC Customer Solution Center

The Customer Solution Centers are trusted environments where world-class IT experts collaborate with you to share best practices, facilitate in-depth discussions of effective business strategies, and help your company become more successful and competitive. This is where you collaborate with subject matter experts, industry leaders and entrepreneurs.

Whether it's a technical briefing, architectural design session or proof of concept, you'll receive the appropriate expertise and resources at every level. You'll receive a glimpse into the solutions of the future, and be able to test your solution against your current business objectives, industry use-cases and future scalability needs — so you can move forward with confidence.

Dell EMC Machine Learning Knowledge Center

Dell EMC has created a haven where thought leaders and industry experts can gather to share their expertise and offer guidance to anyone who's keen to take advantage of deep learning. This community resource includes everything from expert commentary and infographics, to technical blogs, and news from around the Web. It's helpful that everything is laid out in an intuitive way, meaning that resources and connections to the community are easy to find.

NVIDIA Deep Learning Institute

The NVIDIA Deep Learning Institute (DLI) offers hands-on training for developers, data scientists, and researchers looking to solve the world's most challenging problems with deep learning.

Through self-paced online labs and instructor-led workshops, DLI provides training on the latest techniques for designing, training, and deploying neural networks across a variety of application domains. Students will explore widely used open-source frameworks as well as NVIDIA's latest GPU-accelerated deep learning platforms.

Talk to us

Still have questions? Get in touch with the experts at Dell EMC and NVIDIA to discover how to apply deep learning to your organization.

Visit dell.com/hpc and/or contact your Dell EMC local representative or authorized reseller.

Further Reading

Dell EMC Customer Solution Center

Dell EMC Machine Learning Knowledge Center

Dell Launches Artificial Intelligence, Machine Learning Solution

Dell EMC PowerEdge C4130 Rack Server

How NVLink Will Enable Faster, Easier Multi-GPU Computing

Machine Learning Communities

NVIDIA® NVLink™

NVIDIA Volta

OpenCL Caffe

Securing sensitive data: how to navigate the new world of compliance

STAC-A2 Report

Tony Parkinson, Dell EMC & Nick Curcuru, Mastercard- Dell EMC World 2017

Top 10 Machine Learning Frameworks