DELL Technologies

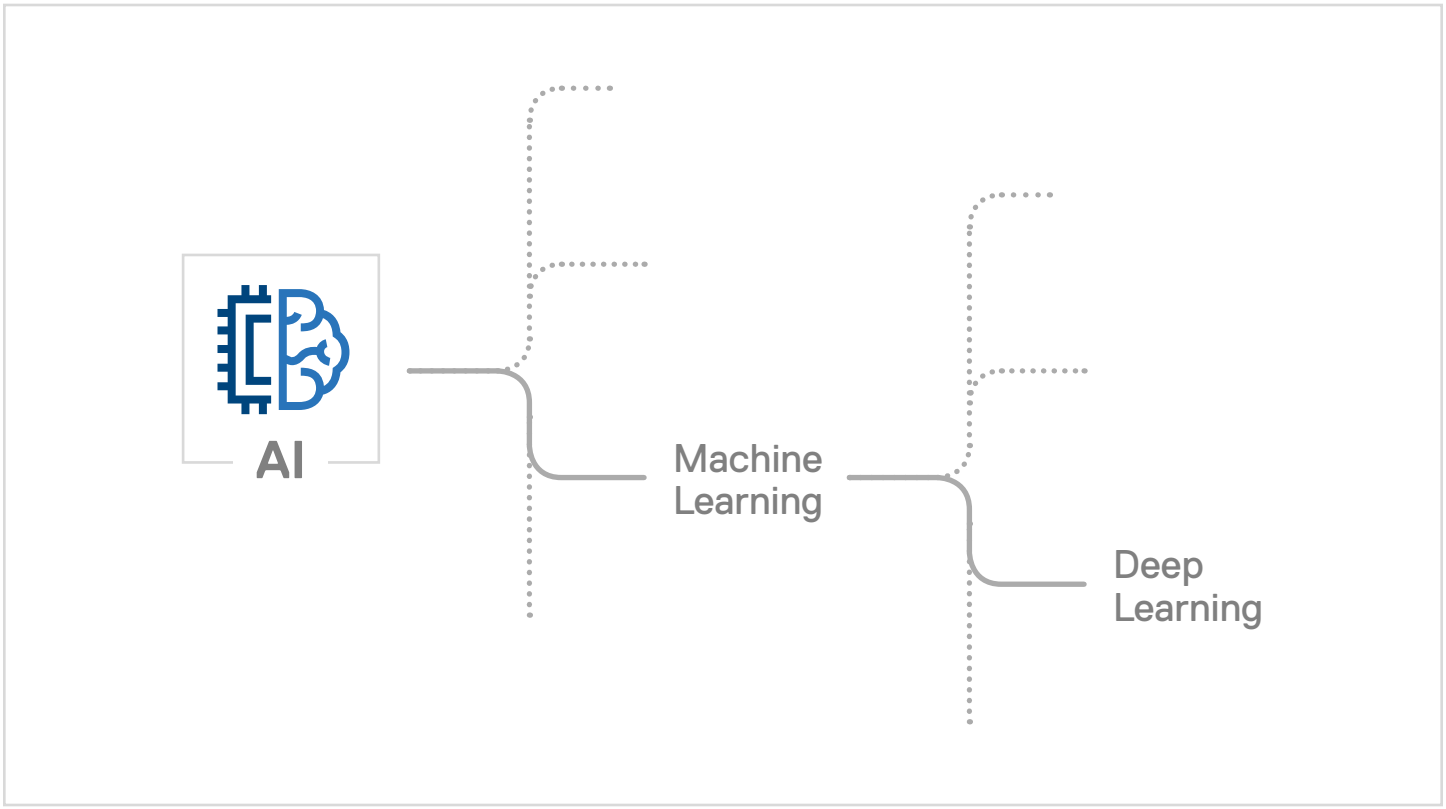
THE SERVER TECHNOLOGY VITAL TO DEEP LEARNING

Introduction

If your organization deals with vast amounts of data of any kind, you are missing out on opportunities to create value and improve customer experience if you are not utilizing deep learning techniques. Deep learning is a subset of machine learning, which in turn is a subset of artificial intelligence (AI). It involves training mathematical models with large amounts of data and then applying them to new data in order to gain fresh insights. Imagine being able to stop fraud in its tracks, for example, with an applied model that detects anomalies in buying patterns of an end user in real time. Or having the ability to recommend the perfect product to a customer browsing your website, keeping the prospect engaged and satisfied.

Modernized servers with the latest in accelerator technologies can get you where you need to go. Deep learning models can be mammoth, as you will learn in this eBook, and accelerators such as graphics processing units (GPUs) and field programmable gate arrays (FPGAs) are essential in making sure your infrastructure can keep up and remain relevant. In partnership with frameworks like TensorFlow and workbooks like Jupyter Notebook, accelerators can facilitate improvements in training the model, decision speed and accuracy, customer retention, and cross-sell and upsell opportunities. Read on to learn how you can ride the deep learning wave to extraordinary results.

Deep learning is a branch of machine learning, which in turn is a branch of AI.



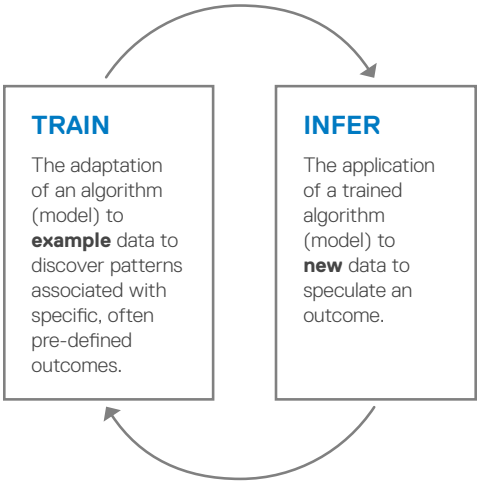
Deep Learning

The deep learning world roughly can be broken down into two pursuits: training and inferencing. Training involves giving the computer instructions on how to learn to do something; in other words, teaching the model to learn and develop accuracy. Inferencing is about running the model on brand new data in an application. You input the model into software code and gain some insight at the edge or point of sale. Training and inferencing are iterative processes. As your model comes across new data, it must be continually refreshed and retrained. Some companies retrain their models as frequently as daily to compete with the ever-evolving threat of fraud.

If your organization isn't already pursuing deep learning, you may be wondering how it applies to your business. Sample use cases include:

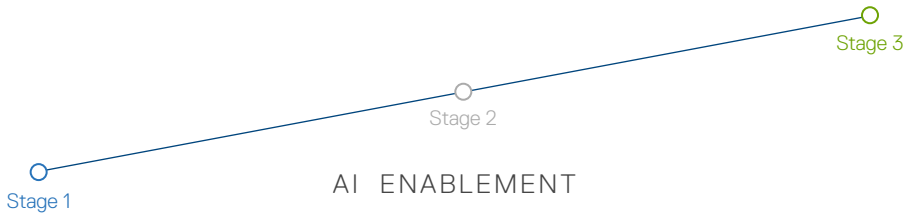
- Computer vision
  - » Facial recognition/image recognition/object detection
  - » Drones
  - » Self-driving cars
- Product recommendation engines
- Natural Language Processing
  - » Text suggestions on phones
  - » Voice recognition when a customer calls in or uses a smart home device

Training and Inferencing[1]



Iterative process to adapt model to changes in data characteristics due to external factors.

Forrester finds that "58% of senior business purchase influencers said that their firm is implementing, planning to implement, or interested in implementing computer vision in [2020]."[2] The **most AI-advanced organizations** (categorized as "Stage 3") are pursuing initiatives related to deep learning, according to recent research by ESG: 64% of these companies are developing, deploying, and tuning AI models in production for natural language processing, while 58% are engaged in the same activities for image classification.[3] Stage 3 organizations derive major business benefits from their involvement with AI.



Stage 3 organizations:

- are **7.8x** more likely than Stage 1 to say AI has been very effective at driving value.
- are **2x** more likely to tie 10% or more revenue to AI initiatives.
- are **2x** more likely to experience a time to value shorter than expectations.
- see a **19%** average improvement in decision speed with AI.
- see a **21%** average improvement in decision accuracy with AI.

To keep pace with advanced AI organizations, embrace technologies that enable deep learning capabilities. Per Forrester, "The algorithms, amount of data, and number of iterations necessary to train a good model will only get more intense. Don't make data science and AI engineering teams beg [infrastructure and operations] for AI infrastructure, or your enterprise will fall behind." [4]

ESG AI Maturity Stages

Stage 1

(**42%** of organizations in the study): Low levels of automation, very limited use of accelerators, and/or little to no converged/HCI-based infrastructure for AI.

Stage 2

(**33%** of organizations): Moderate levels of automation, some use of accelerators, and/or some converged/HCI-based infrastructure for AI.

Stage 3

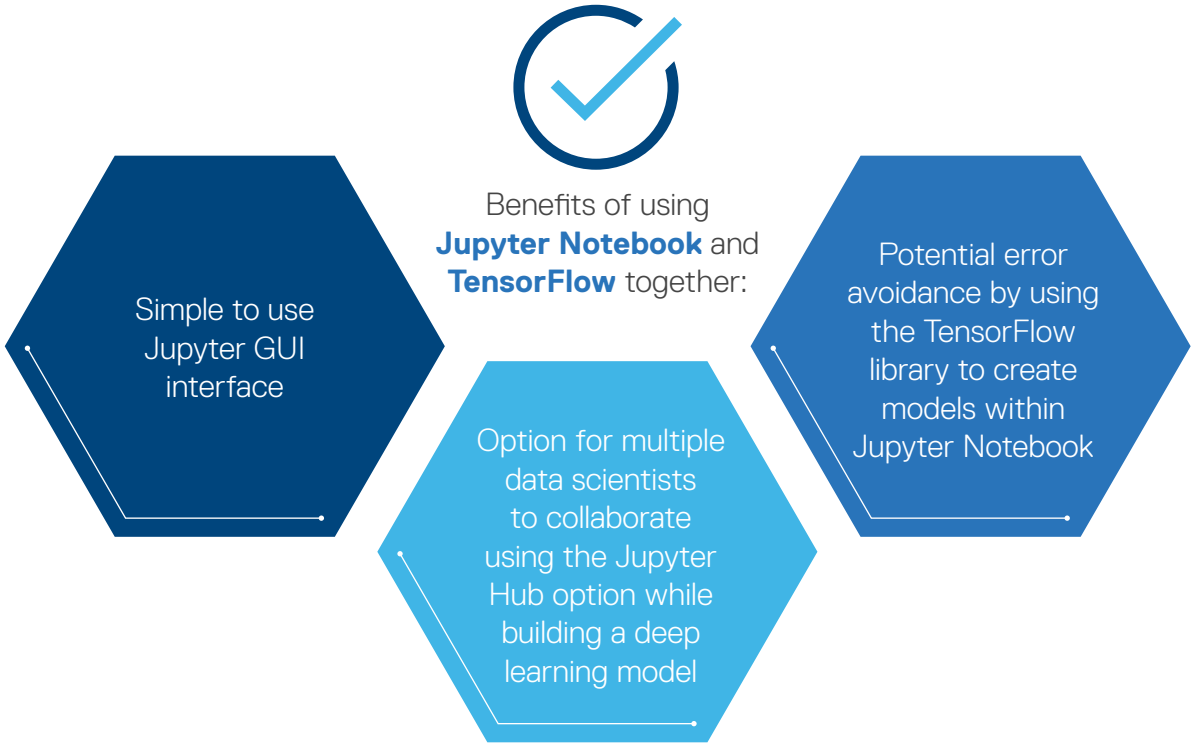
(**24%** of organizations): High levels of automation, broad use of accelerators, and/or high use of converged/HCI-based infrastructure for AI.

Jupyter Notebook and TensorFlow: Enabling Your Deep Learning Work

Packaged together, the Jupyter Notebook platform and the TensorFlow framework offer you an ideal jumping-off point for your deep learning explorations and implementations. These two open-source options offer you a solution for seamlessly researching and building your models.

Jupyter Notebook is a web GUI where the data scientist does the work: building the model, scripting, using libraries, and executing. It is heavily used on the training side of deep learning during the process of creating models. Developers use Jupyter Notebook on the back end to help encode models.

Jupyter Notebook and TensorFlow make solid partners. TensorFlow is a deep learning framework or library that is incorporated into Jupyter Notebook as a resource. TensorFlow is the framework you will use to build your model. A specific benefit? "[You] can . . . use open-source libraries like TensorFlow to prototype and build them very fast, hence avoiding writing error-prone algorithms from scratch."[5] This framework offers easy model building and training, robust production anywhere, and powerful experimentation for research without giving up speed or performance.[6]



The next step is to pair these platforms with powerful accelerators in modernized servers needed to train, retrain, and run the deep learning models you create.

Jupyter Notebook & **TensorFlow** + Modernized **Server infrastructure** =

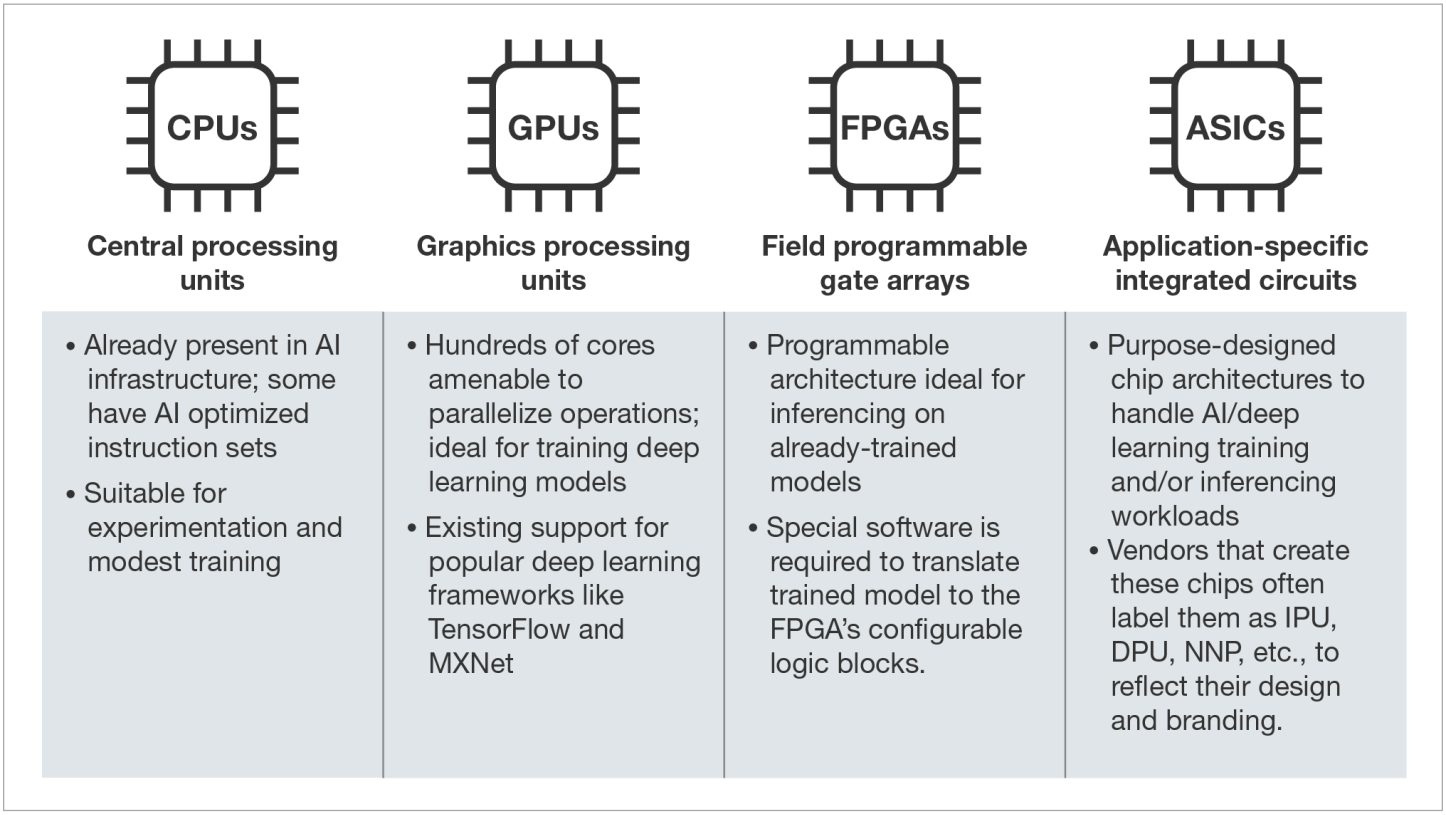
OPTIMIZED DEEP LEARNING ENVIRONMENT

Accelerators: Taking Deep Learning from Theory to Mainstream

Without the parallel processing speed provided by the server technology known as accelerators, deep learning would be destined to languish as just a theory.[7] ESG acknowledges the important role accelerators play in AI enablement, classifying them as one of the three essential compute technologies for an organization to be considered AI-advanced.[8] The complex and ever-updating neural networks essential to deep learning require immense processing power, and accelerators are a requirement to participate.

The two best-known players in this space are GPUs and FPGAs,[9] although other competitors do exist and continue to gain traction. Both accelerators utilize parallel processing, a way of breaking up and running program tasks on multiple microprocessors, to reduce processing time. However, there are significant differences in the two options.

The discovery of GPUs' suitability for deep learning was a bit of a happy accident, as they were originally designed to perform the complex math operations needed for complex graphics. However, it was found that they also work wonderfully for accelerating compute for AI as well. They are best suited for the training branch of deep learning. FPGAs had early origins in networking and telecommunications, and are ideal for inferencing tasks. They have programmable logic blocks that can be optimized to run already trained models at top speed.



AI Chips Vary in Silicon Architectures[10]

Read on to learn how modern Dell EMC PowerEdge servers optimized with accelerators can assist you on your deep learning journey.

The Best Dell EMC PowerEdge Servers for Deep Learning

When you opt for PowerEdge servers, the experts agree you are in good hands.

ChannelPro
Best Server Hardware: **Gold Winner**

IT Brand Pulse
2019 Market Leader: **Rackmount Servers**

For deep learning servers, you will be focused on the product's ability to integrate some combination of GPUs and/or FPGAs to take advantage of their parallel processing powers.

Good: PowerEdge R740

Deep learning workloads like language translation and object detection require scalable resources to deliver fast insights. The PowerEdge R740 offers an optimal mix of accelerator cards, storage, compute power and expandable memory to drive complex deep learning data sets. With single-width and double-width GPU and FPGA options, R740 can accelerate results based on your unique requirements. The R740 features Intel Deep Learning Boost for inferencing in vector neural networks. The R740 can scale up to three 300W or six 150W GPUs, or up to three double-width or four single-width FPGAs.[11]

Better: PowerEdge R7525

The PowerEdge R7525 is a rack server that delivers powerful performance and flexible configurations. The R7525 has the option for up to 24 NVMe drives and can house up to 6 GPUs for high throughput and computing. The number of NVMe drives and GPU customization create flexibility that allows companies to balance throughput and computing power for deep learning. Powered by 2nd Gen AMD EPYC processors, the R7525 has twice as many cores as previous generations. With double the CPU cores, multiple GPU options, and plentiful NVMe drives, the R7525 produces an amazing blend for any deep learning applications.[12]

Best: PowerEdge C4140

The C4140 is truly an accelerator-optimized deep learning powerhouse. It is the only platform that can support NVlink proprietary technology. This is NVIDIA technology that enables the GPUs to talk directly with each other at a speed of up to 25 Gbps, optimizing throughput at a phenomenal rate. GPU-to-CPU communication is also boosted in the C4140, with no switches between the GPU and CPU modules. All four GPUs are put in the front of the server to maximize access to cooler air. GPUs are cleverly overlapped to improve the thermal profile for unlimited unthrottled performance.[13]

Dell EMC Ready Solutions for AI: Deep Learning with NVIDIA

Ready to get started with deep learning but need some assistance in setting up your platform? Dell EMC has developed an architecture for deep learning that provides a complete, validated, and supported solution. Technologies were carefully selected across our portfolio and this solution provides details on the design choice for each component.

Dell EMC created the Data Science Provisioning Portal based on Jupyter notebook and supporting open source frameworks such as TensorFlow.[14] We put our top processors on the job—in fact, one of the configuration options includes this eBook's highest recommended server, the PowerEdge C4140. Watch the video to learn more about the portal and visit the website to learn more about Dell EMC AI solutions.

Accelerate your Deep Learning Initiatives with ProConsult Advisory Services

Create a strategy and roadmap for implementing deep learning and accelerators such as GPUs and FPGAs with ProConsult Advisory Services.

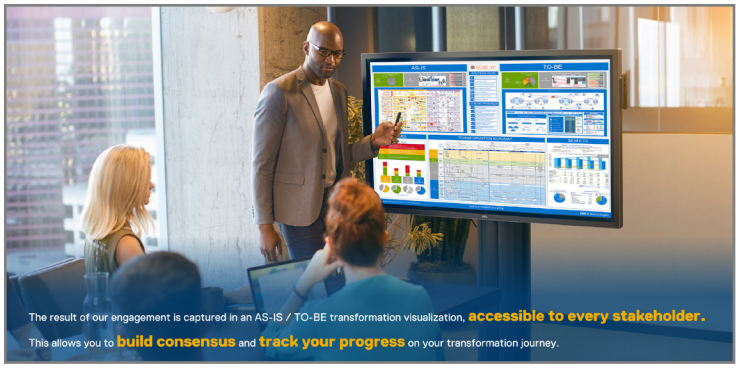
Dell Technologies Consultants can help with an unbiased, end-to-end assessment, addressing internal and external obstacles to help you create a plan for your future state. Dell Technologies ProConsult Advisory Services facilitate a plan for beneficial and lasting change. The methodology can help you realize the business benefits of deep learning transformation faster, more reliably, and with lower risk. The services are designed to help assess and plan transformations that achieve measurable outcomes aligned to your corporate vision and strategy.

Dell Technologies Consulting can help you:

- Secure participation and support from key stakeholders and aligned cross-functional teams sharing a single vision and guiding principles.
- Leverage their experiences with what "good" looks like, including best practice guidance.
- Prepare to immediately execute on findings and recommendations.

Learn more about Dell Technologies ProConsult Services.

Contact a Dell Technologies Consulting Services Expert today.



Ramp Up Your Deep Learning Operations with Flex on Demand for PowerEdge Servers

PowerEdge servers and Flex on Demand combine the industry's best-selling servers with innovative consumption-based payment programs. This solution** provides customers with elastic infrastructure, maximum flexibility, and public cloud-like economics for data centers with dynamic computing requirements and unpredictable workload fluctuations. While one approach has been to support new projects using CAPEX, there are more choices now to help you support and manage your innovative AI projects.

Flex on Demand allows you to pay for technology as you use it, and provides immediate access to buffer capacity and payments that adjust to match your actual usage. For Intel-based PowerEdge servers, CPU utilization can be measured by the hour, so you can avoid the cost of over provisioning. To learn more, visit the Dell Technologies Flex on Demand site.

**Payment solutions provided by Dell Financial Services L.L.C. (DFS) or its affiliate or designee, subject to availability and may vary in certain countries. Where available offers may be changed without notice.

Customer Story: Rozetta Japan

A particularly exciting branch of deep learning is exploring the way humans communicate: automated translation. Rozetta Japan has made massive strides in the field of automated translation services since it launched in 2014. Its incorporation of AI into its offerings is differentiating it from the competition and giving it a distinctive value proposition.

Rozetta's deep learning-based translation service, the T-400 2.0, delivers translation accuracy of 95%. To arrive at this lofty benchmark, the company needed the right servers equipped with the right accelerators for the huge undertaking. It turned to Dell EMC, whose C-Series servers and NVIDIA Tesla V100 GPUs perfectly fit the task at hand. The end result? Mitsuru Tosaka, manager of the Machine Translation Business Division at Rozetta, states, "We harnessed DL to tailor our translation services to make them faster and more accurate."

Excitingly, the business benefits have been significant. CMO Terufumi Watanabe says, "The release of [the T-400 2.0] has delivered a 4-5 times increase in the revenue of the company."

Read more about Rozetta Japan's experience with deep learning and PowerEdge servers.

Terufumi Watanabe, CMO, Rozetta Japan

3X
increase in learning speeds

4-5X
increase in revenue

Automatic translations with
95%
accuracy

Conclusion

Given the vast number of ways to apply this AI technology, your organization cannot afford to ignore the benefits of deep learning. Realize the opportunities created through speedy and targeted decision making by investing in the right infrastructure and platforms to make deep learning a reality for your company. Accelerator server technology is the key to your success, along with finding the right frameworks to partner with for maximum productivity and accuracy. To learn more about how Dell Technologies can help you get started or further your journey with deep learning, contact a Dell Technologies sales rep or visit DellTechnologies.com/Servers.

1 https://downloads.dell.com/manuals/common/retail_analytics_malong_poweredge.pdf

2 https://go.forrester.com/blogs/will-computer-vision-rule-the-enterprise/
Base: 1,314 to 2,264 global business purchase influencers at the director level or above. Source: Forrester Analytics Global Business Technographics® Priorities and Journey Survey, 2019.

3 https://www.dellemc.com/en-us/collaterals/unauth/analyst-reports/products/servers/esg-three-transformational-compute-technologies-verified-to-accelerate-ai-and-business-value-en.pdf

4 "AI Deep Learning Workloads Demand A New Approach To Infrastructure," Forrester Research, Inc., May 4, 2018. https://www.forrester.com/report/AI%20Deep%20Learning%20Workloads%20Demand%20A%20New%20Approach%20To%20Infrastructure/-/E-RES142531

5 https://towardsdatascience.com/tensorflow-for-absolute-beginners-28c1544fb0d6

6 https://www.tensorflow.org/about/?

7 https://www.dellemc.com/en-us/collaterals/unauth/brochures/products/ready-solutions/poweredge_accelerators_brochure.pdf

8 https://www.dellemc.com/en-us/collaterals/unauth/analyst-reports/products/servers/esg-three-transformational-compute-technologies-verified-to-accelerate-ai-and-business-value-en.pdf

9 https://blog.dellemc.com/en-us/fpgas-vs-gpus-tale-two-accelerators/

10 "AI Deep Learning Workloads Demand A New Approach To Infrastructure," Forrester Research, Inc., May 4, 2018. https://www.forrester.com/report/AI%20Deep%20Learning%20Workloads%20Demand%20A%20New%20Approach%20To%20Infrastructure/-/E-RES142531

11 https://www.dell.com/en-us/work/shop/povw/poweredge-r740

12 https://www.dell.com/en-us/work/shop/povw/poweredge-r7525

13 https://www.dell.com/en-us/work/shop/povw/poweredge-c4140

14 https://www.dellemc.com/resources/en-us/asset/white-papers/solutions/h17354-dl-architecture-guide.pdf