MOOR INSIGHTS & STRATEGY

# A PRACTITIONER'S GUIDE TO ARTIFICIAL INTELLIGENCE

AND HOW DELL TECHNOLOGIES CAN HELP THE AI PRACTITIONER

## OBJECTIVES OF THIS GUIDE

Artificial intelligence (AI) is delivering new insights − previously hidden in vast pools of data − to add intelligence to many products and services that ultimately transform the way organizations and machines interact. While human-like intelligence will remain the stuff of fantasy novels and movies for the near future, most organizations should explore incorporating AI into their business, products, and IT projects. Our firm's research concludes that AI can improve productivity of internal applications, increase revenue, reduce costs, and improve products and services with added functionality or communication modes.

This guide provides AI practitioners with an overview of AI and the underlying technologies of machine learning and deep learning. We will outline common applications of this technology, pointing out when and where machine learning and deep learning are typically used. The guide will also outline relevant hardware and software building blocks. Finally, we will explore the technologies and assistance Dell Technologies offers to enable successful AI projects.

## BUSINESS DRIVERS FOR AI

The many possible benefits of AI − increases in productivity, revenues, product improvements, etc. – are creating surging demand for AI technologies. In fact, IDC forecasts total spending in AI will ramp to tens of billions of dollars by the early 2020s.[1] Many drivers fuel this interest which we believe comes down to two primary motivations: 1) to improve operational efficiencies, and 2) to enhance products and services through data-driven insights and use of unstructured data types such as voice and images to enhance human-machine interaction. For simplicity, we call these "Smart Operations" and "Smart Products and Services", respectively. Examples of Smart Operations include e-commerce product recommendation engines, cyber security, customer sales and

[1] IDC forecasts that the total spending for cognitive and AI systems will reach $46B by 2020. https://www.idc.com/getdoc.jsp?containerId=prUS42439617

support chatbots, financial trading, fraud detection, enhanced public safety services and supply chain optimization. Smart Products and Services includes medical diagnosis and treatment, drug discovery, hospital clinical care management, autonomous vehicles, drones, consumer electronics, and threat intelligence and prevention. Moreover, we believe AI will become pervasive and impact virtually every product and business process over the next decade.

## GETTING STARTED

### *SELECT THE RIGHT PROJECTS: CRAWL, WALK, RUN*

Getting started on the right foot with AI requires that you select a project that solves a pressing business problem for which you have access to the requisite data and the right skills set to tackle it. Most enterprises already have a list of projects that could benefit from ML. Start small, perhaps extending your existing Big Data analytics or existing software to include one or two classical ML capabilities, where the ROI is relatively easy to achieve and measure. Meanwhile, team members can get up to speed on DL with the resources outlined below, even if the first project may not require DL skills and technology.

### *TALENT, SKILLS, AND RESOURCES*

Getting the right team and expertise in place is critical for early success. We recommend providing the team with training for SPARK MLlib or SciKit-Learn if they are not already familiar with these tools. This training is critical for projects that will build on existing Big Data infrastructure with classical ML. For those preferring ISVs solutions (see the section on the ML Software Ecosystem), leading software vendors have a rich portfolio of training and documentation available.

For teams embarking on DL projects, adding DL expertise to your existing staff will be needed. First, identify internal talent who possess the aptitude and motivation for learning about AI, especially those with backgrounds in statistics and math. Consider the many online courses that are available from providers such as Coursera, Udemy, and Udacity. Leading AI practitioners such as Andrew Ng, Geoffrey Hinton, and Yoshua Bengio teach some of these highly technical classes.

## Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL)

AI is a branch of computer science dealing with the simulation of intelligent behavior in computers and the capability of a machine to imitate human intelligence. This broad term includes something as simple as being able to recognize an object in an image to the ability to apply reason and ethical values in problem solving − something far beyond current capabilities. Machine Learning (ML) is a field of AI that evolved from the study of pattern recognition and computational learning theory, as opposed to the application of traditional logic-based (If-then-else) algorithms used in earlier AI attempts.

Today, there are two approaches to ML. The first and perhaps most widely understood approach uses statistical analysis modeling, now often called “classical” ML. This approach includes algorithmic tools such as linear regression (fitting a line or curve in n-space), support vector machines (binary classifiers), graph algorithms (minimizing cost functions), and Bayesian analysis (estimating probability distributions). These algorithms are well suited to numerical data and can often produce excellent results in less time while consuming far fewer resources than Deep Learning (DL). Resource allocation and scheduling, predictive analytics, predictive maintenance, recommendation engines, trend discovery, sales forecasting, and product pricing are a few typical use cases of classical ML.

The second approach is Deep Learning (DL), which has stoked much publicity and hype in recent years. DL is a now widely used for computer vision, speech recognition, natural language processing, social network processing, image processing and classification, vision-guided systems, and financial market modeling. DL involves building neural networks that learn from data to discern patterns and draw conclusions. The concept of neural networks has been around for decades, but access to data and extensive compute capability now allows more neural nodes and deep layers to produce strikingly accurate results. Developing a deep neural network (DNN) requires massive training data sets and acceleration hardware such as Graphics Processing Units (GPUs) or Application-Specific Integrated Circuits (ASICs). These tools are needed to parallelize the extreme computational load of the training process.

## A Deep Learning Overview

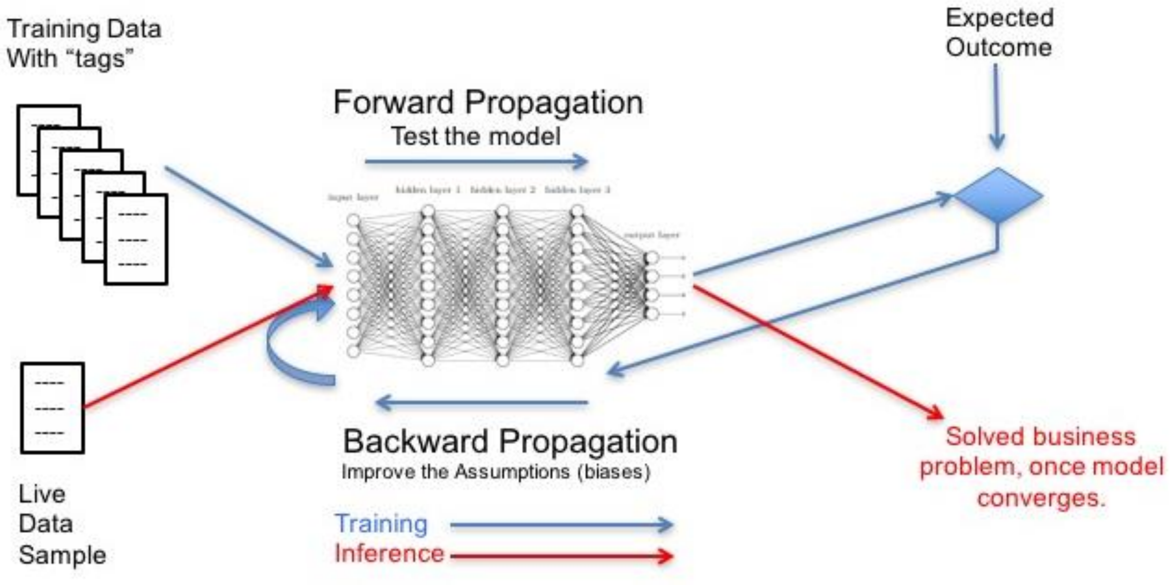
Deep Learning includes simple Deep Neural Networks (DNN), convolutional neural networks (CNN), deep belief networks (DBN), recurrent neural networks (RNN), generative adversarial networks (GAN), and others. Each DL variant is designed to excel in a specific data type. DNNs, often referred to as feed-forward multi-level perceptrons, are general purpose. The other variants were developed to handle complex problems for which classical DNNs become unwieldy or too large for practical use. CNNs were developed for images, RNNs for time series data such as audio and natural languages, and DBNs and GANs for unsupervised learning and gaming.

All DL methods use a cascade of multiple layers of non-linear processing units (see Figure 1), called nodes or neurons, to implement feature extraction and transformation. In supervised learning, the most common type of learning used today, the network is typically trained with “tagged” or “labeled” data. Tagging the dataset can be a tedious and costly process, a fact that steers many projects towards classical ML tools. Once trained to an acceptable level of accuracy, the neural network is now ready to infer the attribute(s) of new (untagged) data samples

The classic example is to identify whether a cat appears in an image. The image is segmented into groups of pixels and each segment is then fed into the initial layer of nodes in the neural network, wherein each segments’ pixel values are multiplied by the nodes’ weight vectors and convoluted with adjacent data to improve feature recognition. Initially, the weights and output result are random, guessing whether the input contains an image of a cat or not.

The known correct answer is then used to adjust the weights in the network layers in reverse order, right to left, in a process called back propagation. The weights of each node are adjusted to better predict to the now-known answer. While the cat example is instructive, imagine the added complexity required to not only identify a human face, but determining whose face is present in an image, or processing a video of a crowded airport or sporting venue for security screening.

## FIGURE 1: DEVELOPING AND USING A DEEP NEURAL NETWORK



**There are two stages of DL: training and inference. Training is the process that calculates the deep neural network parameters (biases and weights). Inference is the subsequent use of the trained network model for evaluation of new data.**

*Source: Moor Insights & Strategy*

Let's look at the function inside each neuron. A neuron calculates a weighted sum of the input vector, or multi-dimensional tensor, applies a bias, and then applies a non-linear transformation function to produce the output tensor. This becomes the input into all nodes of the successive layer. The training process successively refines the weight tensor in each neuron to continually produce more accurate outputs. The non-linear activation function is a simple way to maximize the effect of certainty around a result (-1 or +1) and to minimize the effect of uncertainty (near 0). Common activation transformations include the sigmoid ("S" shaped) logistic and hyperbolic tangent functions to produce data in the range of (0 to 1) and (-1 to +1), respectively, and the rectified linear function (max(n,0)).

While DL is a powerful tool, we note that it's a "black box" methodology. You cannot easily determine exactly why a DNN or CNN produced a certain result; therefore, DL may not be appropriate for certain applications where the decision process must be transparent and auditable. The study of "simulated" data sets with known characteristics can provide insight into how the model evaluates new data, but this process can be labor and compute intensive.

## The Machine Learning Ecosystem

There are three sources of ML software to build and manage your ML project: Open Source code, ML API cloud services (MLaaS), and traditional ISV solutions. Software selection should be based on the type of data you have and the problem you want to solve.

### *Open Source and ISV Tools for Classical ML*

Python R, SciKits-Learn and Apache SPARK MLib are excellent resources for classical ML projects.  They provide state-of-the-art open source tools, complete with examples, documentation, tutorials, and offer active user communities for support. These tools provide an ideal starting point for numerical data and can produce quick results. They are easy to learn and do not require a great deal of data nor hardware acceleration.

For organizations with an ISV license and support agreements in place, robust ML solutions from enterprise-grade software companies like SAP, SAS, and Oracle provide the needed tools and training as well as support and implementation services. These companies also offer a range of DL libraries and even cloud API services. We are also seeing new ISV's enter the market for data analytics that are based on ML such as DataRobot for predictive modeling, DataBricks for business analytics, Domino Data Labs for group collaboration of models, and BigML for easy to use ML for numerical data.

Finally, adding new technologies such as large GPU farms to a datacenter presents new challenges to infrastructure management. Bright Computing offers extensive toolsets to help manage these new heterogeneous clusters. For tools that help manage and optimize the ML application development lifecycle, BitFusion.io offers a platform that includes GPU and FPGA virtualization and sharing.

### *Machine Learning in the Cloud: ML as a Service (MLaaS)*

MLaaS providers such as Google, Amazon AWS, Microsoft and Clarifai offer cloud-based ML services that provide DL tools and pre-trained neural networks such as APIs for voice, text, image, video, and natural language processing. Using MLaaS can be an easy avenue for the novice to begin experimentation and even production deployment for cloud-hosted applications. However, the MLaaS cloud-based delivery model may be inappropriate for projects requiring on-premises solutions, perhaps due to security concerns or high costs of transferring massive training datasets to the cloud.  If one uses MLaaS tools, it may be also be difficult or impossible to transition that model and application to an on-premises solution, as many of these APIs simply have no on-premises equivalent. In addition, simple pre-trained neural networks available as a service may not meet the needs of many AI projects.

## Deep Learning Frameworks

Nearly all projects that build in-house Deep Learning networks now use "DL Frameworks": open source tools sponsored by universities and the big search, e-Commerce, and social networking companies. Each framework has its supporters and often favors a data type or use model for which it was originally developed for internal use.  Figure 2 summarizes the most popular DL frameworks.

## Figure 2: Popular Deep Learning  Frameworks

| Framework | Sponsor(s) | Language(s) | Notes |
|---|---|---|---|
| CAFFE (2) | Berkeley, Facebook | C++, Python | Popular and efficient, especially for deployment on mobile devices |
| MS Cognitive Toolkit (CNTK) | Microsoft | C++, Python | High performance, particularly for RNNs for Natural Language Processing (NLP) |
| MXNet | Apache<br>Amazon | R, Python, Julia, Scala | Comprehensive, cross platform tools, especially if developing for Alexa, good documentation and examples |
| TensorFlow | Google | Python | The most widely used framework, easy to use with good documentation, especially good for CNNs |
| Torch | Facebook | Lua, Python | Popular in research applications due to flexibility, good for CNNs |
| Neon | Intel | Python | Known for performance and scalability, Nervana's Neon will support Intel Neon Engine in 2018 |
| Skymind | VCs | Java | Skymind provide deep learning libraries for Java users on Apache Hadoop and SPARK |
| BigDL | Intel<br>Apache | Python<br>Scala | Distributed (CPU based) deep learning library for Apache SPARK and Hadoop on Intel processors |
| PaddlePaddle | Baidu | Python | Broad set of tools for CNNs and RNNs, targeting datacenter, IOT, and mobile apps |

**Popular DL frameworks are, in general, available as open-source tools and libraries that greatly simplify and standardize the development of DL networks.**

*Source: Moor Insights & Strategy*

In addition to the frameworks listed above, Keras, developed by Google pioneer François Chollet, and the recently announced Gluon, a collaboration between Microsoft and Amazon AWS, recently emerged to provide a higher-level interface on top of these frameworks to speed development. In fact, Keras is already the second fastest growing toolset for DL, after

Google TensorFlow. While these tools can improve productivity, the AI practitioner will still require a solid understanding of the fundamentals of DL. Please see the bibliography for sources of more detailed information on DL.

## MACHINE LEARNING HARDWARE

The hardware required for classical ML depends on the phase of the project and the type of ML algorithms used. A good laptop, workstation, or a virtual machine on a standard server should suffice for the initial development phase as the data scientist begins to develop and test the model. In fact, moderate sized ML model can be developed and deployed on well-equipped 1- and 2-socket workstations equipped with sufficient memory. For larger production installations, a high performance 2-socket server with plenty of memory and storage should be sufficient.

As mentioned earlier, training a neural network is compute-intensive. In fact, the time required to complete a single training run could take days or weeks, even on a fast CPU cluster, or many hours or days on a single GPU. Fortunately, most of the DL frameworks outlined above now provide optimized libraries with decent scalability for multiple GPUs. The depth and breadth of the neural network and the criticality of time-to-production determine the number of such servers required to perform the training. Note that scaling to multiple servers, each with multiple accelerators, can greatly increase the complexity of development and hardware deployment and management.

Intel recently announced the Nervana Engine, a DL ASIC, which may offer an alternative to GPUs for training neural networks. Dell is an early adopter and provider of Nervana Engine. While initial performance claims are impressive, the Nervana ecosystem will require time to develop and mature.

For inference processing, the type of processor or accelerator needed is a function of the bandwidth and latency requirements of the deployment environment. A simple query for text or images can be run on an Intel Xeon or even a mobile CPU. However, a GPU, ASIC, or FPGA may be required when the input data is a high-resolution stream of data and/or when the required latency is measured in a few milliseconds. A GPU is fast, but requires more power, while an FPGA offers excellent power efficiency and the flexibility of re-programmability, albeit at the expense of requiring hardware development expertise.

## DELL TECHNOLOGIES PLATFORMS AND ASSISTANCE FOR AI PROJECTS

When Dell acquired EMC, along with the Virtustream cloud provider, the company gained much more than the leading enterprise storage vendor.  Dell found a rich vein of ML expertise in EMC that could improve the ability to service this fast-growing market. This section will cover the Dell server and workstation hardware for ML; the new “Ready Bundles” that make it easier to deploy medium to large-scale ML training and applications; solutions for enterprise class cloud and hybrid off-premises/on-premises via Virtustream; and lastly the consulting services and training capabilities of the combined company.

### HARDWARE FOR CLASSICAL ML

Standard 2-socket servers or workstations will meet most projects’ needs here, as tools such as SPARK MLlib and Intel BigDL do not require GPU acceleration. The Dell PowerEdge R740 or R740xd could be ideal ML platforms. Note that the heftier power supply in the R740xd provides a level of future-proofing for potential DL projects, supporting three 300-watt or six 150-watt accelerators and eight to sixteen disk drives in a 2U chassis.

For a desk side platform, the Dell Precision 5820 and 7920 towers offer cost effective workstation solutions for development and moderate size deployments with one to two high performance Intel Xeon CPUs, up to three GPUs, up to ten 3.5” storage devices, and 3TB of DDR4 DRAM.

### HARDWARE FOR DEEP LEARNING

To develop and train a production-sized DNN, a system like the new Dell PowerEdge C4140 should support your needs for a scalable single server solution. Each C4140 supports up to four 300W NVIDIA Tesla V100 (Volta) GPUs interconnected on the NVIDIA NVLINK 2.0 fabric, all in a thermally efficient 1U package supporting up to 24 DDR4 DIMMs. Eight or more C4140’s can be clustered with Infiniband for building larger scale models, delivering up to 500 PetaFlops of ML performance. For developers, the Dell PowerEdge T640 and T7920 workstations provides a desk side or rackable platform that supports up to four or three state-of-the-art GPUs, respectively, for developing performance-demanding applications.

### Machine and Deep Learning “Ready Bundles”

Dell Technologies has recently announces pre-configured “Ready Bundles” for building ML and DL at scale to simplify the configuration process, lower costs, and speed deployment of distributed multi-node ML/DL clusters. These offerings are relatively new in the ML market and help Dell stand out as a vendor who understands the science and practice of ML. These bundles are intended for customers with limited AI expertise, but who are seeking competitive advantage by developing in-house AI applications that will scale with their needs and capability.  Each Ready Bundle includes preconfigured servers, accelerators, storage, networking, and software for a specific ML use cases. Discuss these new solutions with your Dell account team to learn more.

### Enterprise Cloud Services

Virtustream incorporates enterprise managed private cloud solutions with their cloud provider infrastructure and managed services capabilities. The offering provides a secure and managed AI/ML service for enterprise customers needing this in the cloud. Virtustream leverages the suite of Dell EMC hardware and ancillary components to support development in Google Tensorflow.

### Consulting Services and Training

Dell Technologies offers fixed scope and project-based services to help their customers get up to speed in Machine Learning. Dell has developed and deployed state-of-the-art solutions in almost every industry vertical, particularly in healthcare, financial services, telecommunications, utilities, and media entertainment. The company’s practice is comprised of experienced data scientists who can help design models and algorithms and prepare data for processing in popular frameworks.

## Conclusions and Next Steps

Many industry leaders proclaim we are entering the era of AI, where machines can be trained to do tasks which traditional procedural programming models are unable to tackle. While the technology can seem daunting, a stepwise, practical approach to embracing AI in the enterprise does not have to be intimidating. Solid ROIs can be achieved with relatively straight-forward extensions to existing Big Data infrastructure already in place in most enterprises.

Furthermore, enterprises and government agencies would be well served to examine how AI can transform their products and services with Smart Products and Services and their internal business processes with Smart Operations. These areas can drive significant new sources of revenue and cost savings or improve public service and safety. With the combined expertise, technology, and best practices that resulted from the merger of Dell and EMC, we believe the new company is well positioned to help their clients on this journey. In particular, Dell's new ML and DL Ready Bundles are a notable step to establish the company as a leader in this new phase of IT.

Enterprise and government IT staff should embrace ML and prepare themselves to plan and implement these new learning computing models to further their organizations best interests, as well as further their own careers as IT, data science, and computer science professionals.

For a list of additional resources, see Appendix A.

# APPENDIX A

## ADDITIONAL RESOURCES

Here are some links to additional information, which may be helpful for AI practitioners:

1. An introduction to DNNs and deep learning can be found here: http://neuralnetworksanddeeplearning.com/index.html
2. Here's an excellent tutorial on deep learning which goes into more detail: https://cambridgespark.com/content/tutorials/deep-learning-for-complete-beginners-recognising-handwritten-digits/index.html
3. What is "convolution" and why it's so important in deep learning: http://timdettmers.com/2015/03/26/convolution-deep-learning/#more-192
4. Microsoft has provided an instructive "cheat sheet" to help select the right ML algorithm, and while this article is in the context of Microsoft products, the concepts can be applied to virtually any ML software: https://docs.microsoft.com/en-us/azure/machine-learning/studio/algorithm-choice
5. A short, but deep look at a few ML case studies with code samples: https://towardsdatascience.com/applied-deep-learning-part-2-real-world-case-studies-1bb4b142a585
6. A good source for ML application trends by industry (high level) https://www.techemergence.com
7. Two divergent views on chatbots: https://techcrunch.com/2016/05/29/why-do-chatbots-suck/ https://www.forbes.com/sites/mnewlands/2017/07/29/manychats-chatbots-are-getting-400-roi-heres-how-you-can-too/#2805ab357583
8. A good resource on deep learning and how GPUs make it all possible: https://developer.nvidia.com/deep-learning
9. Easy over Hard: A Case Study on Deep Learning. A scholarly research paper explores tuned support vector machines (SVMs) vs. CNNs with surprising results: https://arxiv.org/pdf/1703.00133.pdf

## IMPORTANT INFORMATION ABOUT THIS PAPER

### AUTHOR

Karl Freund, Senior Analyst, Moor Insights & Strategy

### PUBLISHER

Patrick Moorhead, Founder, President, and Principal Analyst, Moor Insights & Strategy

### INQUIRIES

Contact us if you would like to discuss this report, and Moor Insights & Strategy will respond promptly.

### CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in context, displaying author's name, author's title, and "Moor Insights & Strategy". Non-press and non-analysts must receive prior written permission by Moor Insights & Strategy for any citations.

### LICENSING

### DISCLOSURES

This paper was commissioned by DELL TECHNOLOGIES, Inc. Moor Insights & Strategy provides research, analysis, advising, and consulting to many high-tech companies mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

### DISCLAIMER

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Moor Insights & Strategy disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Moor Insights & Strategy and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Moor Insights & Strategy provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.