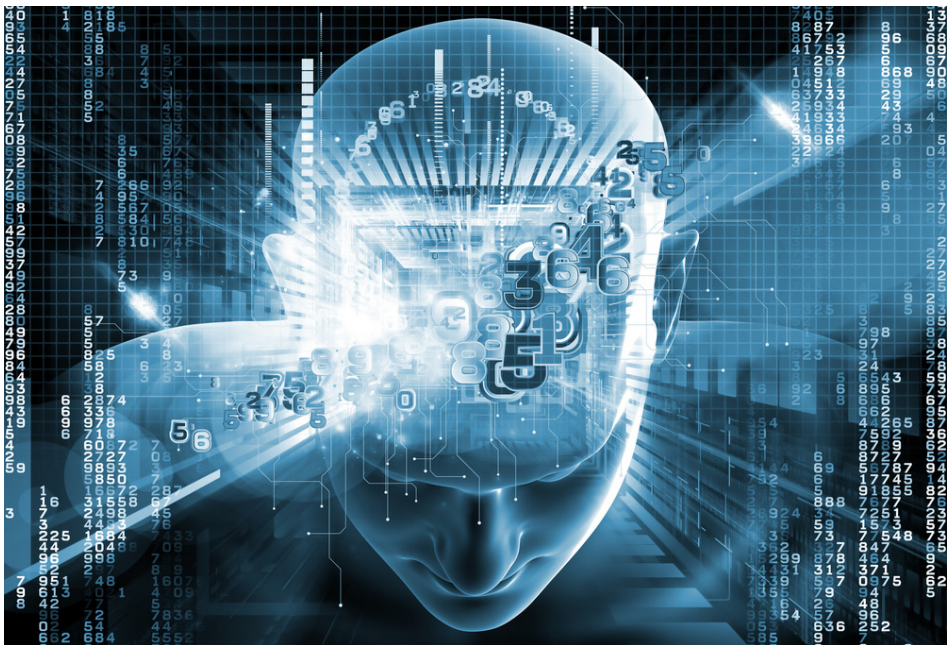
insideBIGDATA

InsideBIGDATA Guide to

Deep Learning & Artificial Intelligence

Written by Daniel D. Gutierrez, Managing Editor, insideBIGDATA



BROUGHT TO YOU BY

NVIDIA

Deep Learning and AI – An Overview

This is the epoch of artificial intelligence (AI), when the technology came into its own for the mainstream enterprise. AI-based tools are pouring into the marketplace, and many well-known names have committed to adding AI solutions to their product mix — General Electric is pushing its AI business called Predix, IBM runs ads featuring its Watson technology talking with Bob Dylan, and CRM giant Saleforce released an AI addition to their products, a system called Einstein that provides insights into what sales leads to follow and what products to make next.

These moves represent years of collective development effort and billions of dollars in terms of investment. There are big pushes for AI in manufacturing, transportation, consumer finance, precision agriculture, healthcare & medicine, and many other industries including the public sector.

AI is becoming important as an enabling technology, and as a result the U.S. federal government recently issued a policy statement, "Preparing for the Future of AI" from the "Subcommittee on Machine Learning and Artificial Intelligence," to provide technical and policy advice on topics related to AI.

> One of the big reasons why AI is on its upward trajectory is the rise of relatively inexpensive compute resources.

Perhaps the biggest question surrounding this new-found momentum is "Why now?" The answer centers on both 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: (i) scale of data, and (ii) scale of computation. It was only recently that technologists have figured out how to scale computation to build deep learning algorithms that can take effective advantage of voluminous amounts of data.

One of the big reasons why AI is on its upward trajectory is the rise of relatively inexpensive compute resources. 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 2006, advanced neural networks have been used to realize methods referred to as Deep Learning. Now, with the adoption of GPUs (the graphics processing unit originally designed 10 years ago for gaming), neural network developers can now run deep learning with compute power required to bring AI to life quickly. Cloud and GPUs are merging as well, with AWS, Azure and Google now offering GPU access in the cloud.

Contents

Deep Learning and AI – An Overview 2
- The Difference between AI, Machine Learning and Deep Learning 3
- The Intersection of AI and HPC 3

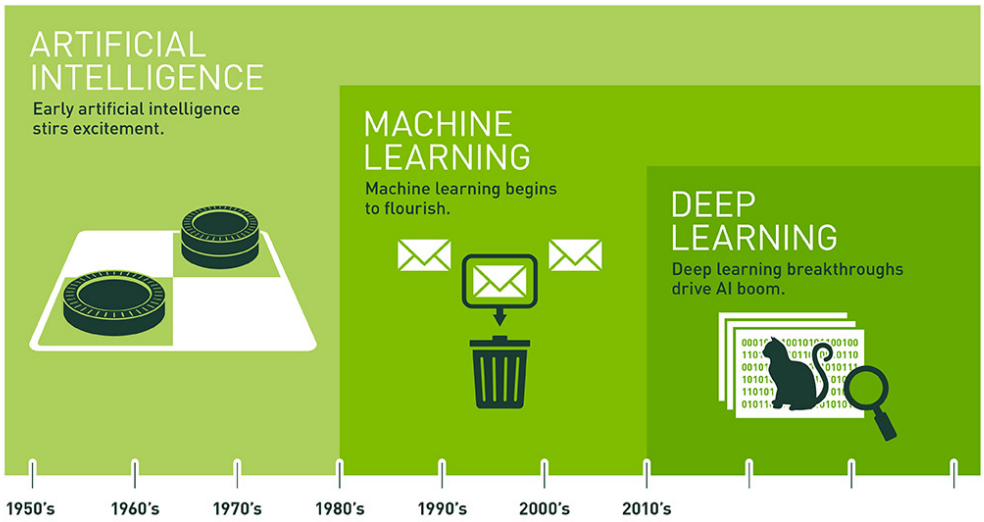
Are AI/Machine Learning/Deep Learning in Your Company's Future? 5
- State of Evaluation 5
- Deployment Results 6
- Technology Perspectives 6

Accelerating Analytics for the Enterprise 7
- Accelerating Machine Learning 7
- Becoming an AI Enterprise 7

Success Stories 8
- AI-Powered Healthcare at Scale 8
- AI-Powered Weather Forecasting 8
- AI Accelerated Cyber Defense 9
- Defending the Planet with AI 9

Summary 9

There are many flavors of AI: neural networks, long short-term memories (LSTM), Bayesian belief networks, etc. Neural networks for AI are currently split between two distinct workloads, training and inference. Commonly, training takes much more compute performance and uses more power, and inference (formerly known as scoring) is the opposite. Generally speaking, leading edge training compute is dominated by NVIDIA GPUs, whereas legacy training compute (before the use of GPUs) by traditional CPUs. Inference compute is divided across the Intel CPU, Xilinx/Altera FPGA, NVIDIA GPU, ASICs like Google TPU and even DSPs.



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

The Difference between AI, Machine Learning and Deep Learning

With all the quickly evolving nomenclature in the industry today, it's important to be able to differentiate between AI, machine learning and deep learning. The simplest way to think of their relationship is to visualize them as a concentric model as depicted in the figure below. 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.

AI has been part of our thoughts and 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. In the long decades since, AI has alternately been heralded as an all-encompassing holy grail, and thrown on technology's bit bucket as a mad conception of overactive academic imaginations. Candidly, until around 2012, it was a bit of both.

Over the past few years, especially since 2015, AI has exploded on the scene. Much of that enthusiasm has to do with the wide availability of GPUs that make parallel processing ever faster, cheaper, and more powerful. It also has to do with the simultaneous one-two punch of practically infinite storage and a flood of data of every stripe including images, video, text, transactions, geospatial data, etc.

On the same trajectory, deep learning has enabled many practical applications of machine learning and by extension the overall field of AI. Deep learning breaks down tasks in ways that make all kinds of machine assists seem possible, even likely. Driverless cars, better preventive healthcare, even better movie recommendations, are all here today or on the horizon. AI is the foundation for the present and the future.

The Intersection of AI and HPC

The intersection of AI and HPC is showing that cognition can be computable in a practical way for real-world applications. It represents a balance of logic processing with numerically intensive computation. It is an area of intense activity in commercial, industrial, government, and academic research settings. The combination of AI and deep learning algorithms with HPC numerically intensive statistical analysis and optimization profoundly impacts the IT industry to influence every aspect of human life.

Recently, Baidu Chief Scientist Andrew Ng posed the question "Why is HPC speeding up machine learning and deep learning research?" He indicated that a lot of progress in AI today is driven by empirical work by running experiments. He also points out that HPC is allowing researchers and developers to be more productive by letting them iterate experiments more quickly using the familiar "idea → experiment (code) → result (test)" cycle. Instead of weeks, months or even years to complete an experiment, you now can reduce the time to days.

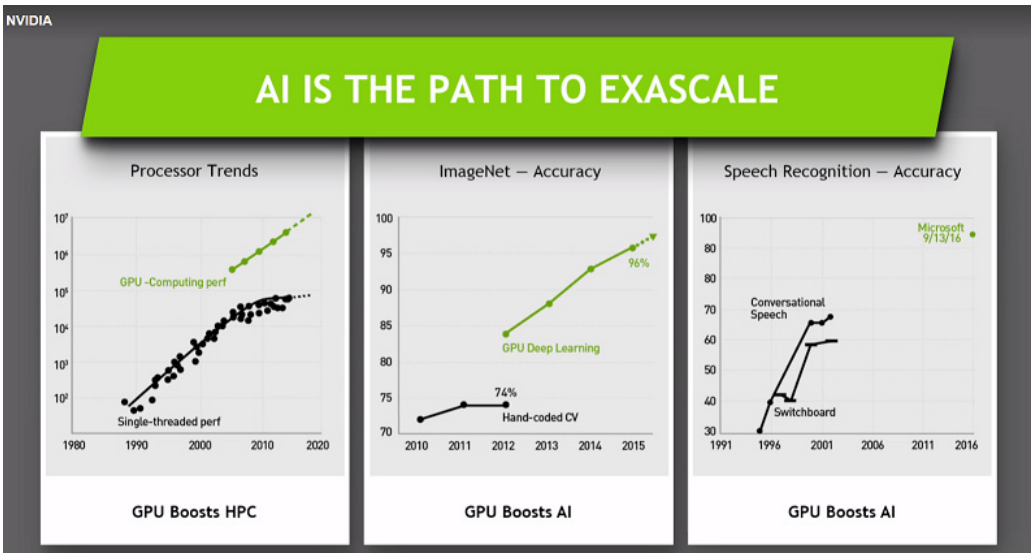
> The future of supercomputing will integrate computational science and AI, and that AI supercomputing is our best path forward to exascale computing.

Here are some other observations on the coupling of AI and HPC:

- Computation science and data science are both critical for future HPC. Specifically, AI is the path toward *exascale* computing (computing systems capable of at least one exaFLOPS). At the SC16 conference, there were numerous examples from a diverse set of domains such as healthcare, weather, astronomy, etc. where AI is being used to solve traditional HPC problems. We're already seeing ways for deep learning to help HPC solve problems.
- Deep learning is a super-computing challenge as well as a supercomputing opportunity. In fact, future supercomputers should be designed for both computational science as well as data science. The future of supercomputing will integrate computational science and AI, and that AI supercomputing is our best path forward to exascale computing.
- Once you've trained an accurate model, the next step is deployment, or figuring out how to serve large neural networks across a base of users. Many feel that HPC brings the benefits of advanced AI to users at scale.
- An interesting statistic involving the use of AI for speech recognition shows that using 10x more data can lower relative error rates by 40%. HPC allows training models on ever larger data sets, and is therefore a key part of progress in AI.

Not too long ago, HPC was the exclusive domain of companies like Cray and IBM. But earlier this year, at the *NVIDIA* GPU Technology Conference 2016, NVIDIA CEO Jen-Hsun Huang used his keynote address to introduce the DGX-1, touted as the world's first "AI supercomputer in a box" costing around $130,000. This launch was groundbreaking for the field of AI.

In the final analysis, the success of organizations to fully capitalize on these technologies may boil down to finding enough skilled workers. Many companies are working to determine how to recruit the right talent and build the right organization to succeed in an AI-driven economy. Maybe some practitioners who grew up doing machine learning didn't grow up with an HPC background. The solution may be to sit down and learn it yourself, or maybe partner with people that have this skillset.



Are AI/Machine Learning/Deep Learning in Your Company's Future?

In this section we'll discuss the results of the recent "insideHPC insideBIGDATA AI/Deep Learning Survey 2016" underwritten by NVIDIA. In November 2016, insideHPC Media LLC conducted this audience survey to get readers thoughts about how they see AI, machine learning and deep learning for their own companies. Here, we'll provide some of the survey results including numeric results, data visualization, and interpretation of the results.

The **audience sample** consisted of 130 readers from a wide variety of industries — 21% academic/scientific research, 20% technology service provider, 18% manufacturer of technology products/software, 9% government, 6% financial/banking, 5% energy, 4% healthcare/pharma, and others.

The variety of **job description** also was varied — 15% scientific researcher, 10% IT consultant/integrator, 9% data scientist, 9% non-IT executive management, 8% IT management, 8% data architect, 8% marketing/business analyst, 6% system admin, 5% data analyst, 5% big data engineer, and others.

Do you feel AI/machine learning/deep learning will have an impact on the work you are doing?

> In terms of stage of adoption, the survey found that vast majority (98%) are evaluating or already have adopted AI, machine learning and deep learning.

State of Evaluation

The survey worked to gauge the degree to which an impact would be seen by engaging AI, machine learning and deep learning. A significant 92% felt that their business will be impacted and indicated some level of importance that AI will have on the work they're doing, either now or in the next three years.

In terms of **stage of adoption,** the survey found that vast majority (98%) are evaluating or already have adopted AI, machine learning and deep learning — 22% actively evaluating, 17% currently testing, 26% currently using in production, and another 33% indicated they are interested in the technology.

In terms of **perceived benefit** in adopting AI, machine learning and deep learning, 29% indicated they expected to more quickly turn data into knowledge and uncover patterns in large data sets to reveal new insights, 20% expected to more effectively and visualize data, 19% expected to accelerate the speed of analytics, 19% wanted to scale to turn big data into actionable knowledge, and 10% expected it would help manage the data deluge that many enterprises experience today.

When asked about what industries will see the **most impact** from AI, machine learning and deep learning, respondents indicated — 16% healthcare, 13% finance/insurance, 11% transportation, 11% IT, 10% energy, 10% manufacturing, 10% academic research, 9% retail, 7% government, and others.

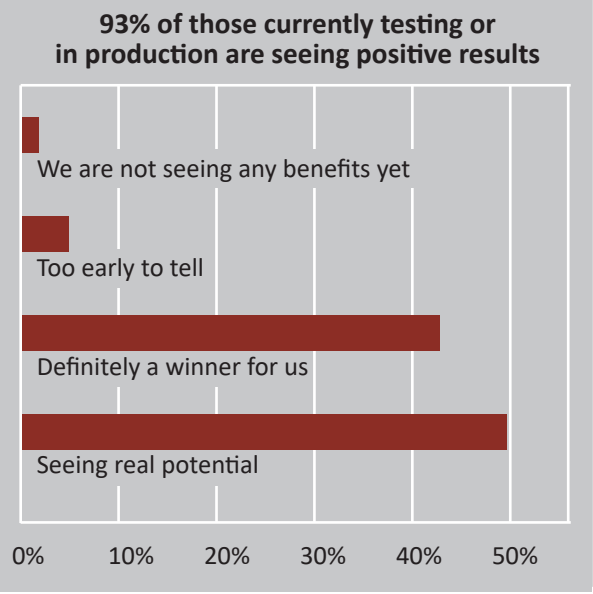
Readers were asked about any perceived impediments for adopting AI, machine learning and deep learning. 40% indicated lack of appropriate skill-sets.

Deployment Results

For those respondents who already have engaged AI, machine learning and deep learning, the survey found that 93% of those currently testing or in production are seeing positive results.

Readers also were asked an important question about any **perceived impediments** for adopting AI, machine learning and deep learning — 40% indicated lack of appropriate skill-sets, 22% technology complexity, 16% trust of the technology, 10% cost is prohibitively high, and others.

What are the results you are seeing with AI/machine learning/deep learning?



Technology Perspectives

The survey also worked to uncover perspectives for how new, enabling technologies would affect the ability to adopt AI, machine learning and deep learning. For instance, 61% of respondents indicated that GPUs are having a positive impact — 35% say that GPUs represent the Holy Grail for increased compute resources, and 26% believe GPUs may have at minimum a modest effect on compute resources.

24% felt that HPC translates into machine learning progress: the faster we can train our networks, the more iterations we can make on our data sets and models, and the more iterations we make, the more we advance our machine learning.

The survey also inquired about the **effect of HPC intersecting with AI,** machine learning and deep learning. 28% of respondents felt that HPC technologies will allow them to scale computationally to build deep learning algorithms that can take advantage of high volumes of data, 25% felt that HPC techniques will allow the training of deep neural networks which tend to be computationally intensive, and 24% felt that HPC translates into machine learning progress: the faster we can train our networks, the more iterations we can make on our data sets and models, and the more iterations we make, the more we advance our machine learning.

In terms of the **primary enablers and drivers** for adopting AI, machine learning and deep learning, the survey found — 29% said the recent surge of data such as images, text, speech has created a need to understand complex data that previously was not machine understandable and searchable, 23% said the need by data scientists and researchers to solve problems of increasing difficulty, 23% said the need to scale up computational resources to meet the demands of the large availability of data, and others.

Accelerating Analytics for the Enterprise

The field of AI has seen a significant increase in interest in the past couple of years, including an upswing in the number of large enterprises running their own experiments with AI systems. The vast majority of AI projects in enterprises is still at the experimentation stage where about half of large enterprises are experimenting with "smart computing" projects. These organizations are examining problems and deciding whether AI can be applied to find solutions. These efforts are still in the early stages. But 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 to the data deluge many companies are seeing.

"Accelerated Analytics" involve GPU-accelerated solutions, like Kinetica, MapD, SQream Technologies, BlazingDB, Graphistry, and BlazeGraph. These solutions allow you to access and interact with data in milliseconds. It's no longer just about Hadoop and Spark with analysts having to wait a minute and then go on to the next query.

Solutions from Graphistry, a visual graph company, offer the ability to look at security logs. You can point Graphistry at a Splunk implementation and quickly go through hundreds of thousands of security alerts for the day. Instead of just doing a report, you can actually visualize trouble spots and keep dissecting and going deeper from a visual standpoint. Many enterprises are using the graph analytic approach to understand the data so you get the correlations with 100x more data.

"AI driven analytics" addresses the needs of the enterprise's digital business and 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, or "stronge scale."

There's a need for GPU acceleration since doubling of traditional processors every 18 to 24 months (per Moore's Law) isn't happening anymore. GPUs accelerate in-memory databases. For example, Kinetica is 50-100x faster on queries with large data sets than traditional x86 databases. Further, noted blogger, Mark Litwintschik has benchmarked MapD vs. major CPU systems and found it to be between 74x to 3500x faster than CPU DBs.

Accelerating Machine Learning

GPUs have proven their value for machine learning, offering orders-of-magnitude speedups on dense and sparse data. They define the current performance limits for machine learning but have limited model capacity. Academic researchers are working on methods to mitigate that challenge and achieve linear speedups with GPUs on commodity networks. One compelling project is BIDMach from U.C Berkeley, a machine-learning toolkit similar to MLlib for Spark that has demonstrated a two-orders-of-magnitude gain over other systems when run on GPU hardware. BIDMach makes full use of both CPU and GPU acceleration. BIDMach often outperforms cluster systems on problems where models fit in GPU memory and defines the performance limit for small-model problems. Still, Scaling BIDMACH is not an easy task. With two-orders-of-magnitude more computation at each node, a proportionate increase in network bandwidth is needed to avoid stifling performance. Alternately, we need to find radical ways to reduce the load on the network.

> The whole mindset is changing from being overwhelmed by the data deluge, to actually being data hungry.

Becoming an AI Enterprise

Enterprises also appear to be embracing the idea of businesses accessing, and processing, and taking advantage of data by using AI machine learning or deep learning, in other words — becoming an "AI enterprise." 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.

Success Stories

This section highlights a number of compelling use case examples focusing on the use of AI and deep learning for the solution of important problems across a wide spectrum of domains. The examples illustrate how GPUs can be effectively combined with AI technology. The visualization below show the rapid growth in number of organizations engaged with AI and deep learning in just two years' time.

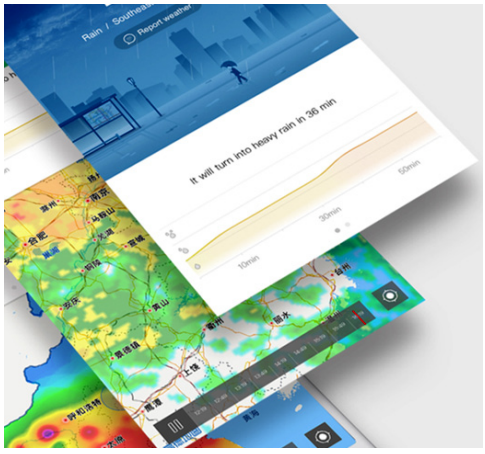
AI-Powered Healthcare at Scale

A number of use case examples of AI can be seen in the healthcare field. The following examples demonstrate AI's flexibility across many problem domains.

- **AI platform to accelerate cancer research –** To speed advances in the fight against cancer, the Cancer Moonshot initiative unites the Department of Energy (DOE), the National Cancer Institute (NCI) and other agencies with researchers at Oak Ridge, Lawrence Livermore, Argonne, and Los Alamos National Laboratories. NVIDIA is collaborating with the labs to help accelerate their AI framework called CANDLE as a common discovery platform, with the goal of achieving 10X annual increases in productivity for cancer researchers. This exascale framework will make it possible for scientists and researchers to use deep learning as well as computational sciences to address the urgent challenges of fighting cancer. The new NVIDIA DGX SATURNV supercomputer will help in this regard.
- **Accelerating drug discoveries with AI –** New drugs typically take 12 to 14 years and $2.6 billion to bring to market. BenevolentAI is using GPU deep learning to bring new therapies to market quickly and more affordably. They've automated the process of identifying patterns within large amounts of research data, enabling scientists to form hypotheses and draw conclusions quicker than any human researcher could. For example, using the NVIDIA DGX-1 AI supercomputer, two potential drug targets for Alzheimer's were identified in less than one month.
- **AI advances the fight against breast cancer –** Breast cancer is the second leading cause of cancer death for women worldwide. Genomic tests help doctors determine a cancer's aggressiveness so they can prescribe appropriate treatment. But testing is expensive, tissue-destructive, and takes 10 to 14 days. Case Western Reserve is using GPU deep learning to develop an automated assessment of cancer risk at 1/20 the cost of current genomic tests.
- **AI predicts and prevents disease –** GPU deep learning is giving doctors a life-saving edge by identifying high-risk patients before diseases are diagnosed. Icahn School of Medicine at Mount Sinai built an AI-powered tool, "Deep Patient," based on NVIDIA GPUs and the CUDA programming model. Deep Patient can analyze a patient's medical history to predict nearly 80 diseases up to 1 year prior to onset.

AI-Powered Weather Forecasting

Weather forecasting involves processing vast amounts of data to derive predictions that can save lives and protect property. Colorful Clouds is using GPU computing and AI to process, predict, and communicate weather and air-quality conditions quickly through a new generation forecasting and reporting tool which, unlike traditional tools, provides individual location-based real-time forecasts with extremely high accuracy. Moving from CPUs to GPUs was able to speed the processing of data by 30-50x.



AI Accelerated Cyber Defense

Our daily life, economic vitality, and national security depend on a stable, safe and resilient cyberspace. But attacks on IT systems are becoming more complex and relentless, resulting in loss of information and money and disruptions to essential services. Accenture's dedicated cyber security lab uses NVIDIA GPUs, CUDA libraries, and machine learning to accelerate the analysis and visualization of 200M-300M alerts daily so analysts can take timely action.

Defending the Planet with AI

The U.S. government's Asteroid Grand Challenge seeks to identify asteroid threats to human populations. The team at NASA Frontier Development Labs picked up the challenge by employing GPU powered AI & deep learning to identify threats and their unique characteristics. The resulting "Deflector Selector" achieved a 98% success rate in determining which technology produced the most successful deflection.

Summary

AI is transforming the entire world of technology, but AI isn't new. It has been around for decades, but AI technologies are only making headway now due to the proliferation of data and the investments being made in storage, compute and analytics technologies. Much of this progress is due to the ability of learning algorithms to spot patterns in larger and larger amounts of data.

In this guide, we've taken a high-level view of AI and deep learning in terms of how it's being used and what technological advances have made it possible. We also explained the difference between AI, machine learning and deep learning, and examined the intersection of AI and HPC.

> Organizations are making significant inroads in using AI to solve real-life problems, especially in the enterprise.

We presented the results of a recent insideBIGDATA survey, "insideHPC / insideBIGDATA AI/Deep Learning Survey 2016," to see how well these new technologies are being received. The results showed that organizations are making significant inroads in using AI to solve real-life problems, especially in the enterprise.

Finally, we took a look at a number of high-profile use case examples showing the effective use of AI in a variety of problem domains. We strived to highlight real examples where people are getting impacted by AI, and the fact that early adopters have been very successful in what they set out to achieve where it did not take a lot for this success to be attained. This is an incentive for organizations to try out AI and deep learning to solve their problems.

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.

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.