CLABBY ANALYTICS

Research Report

The Invasion of the Learning Machines

Executive Summary

Last month *Clabby Analytics* attended NVIDIA's Graphics Technology Conference (GTC), an annual event held in San Jose, California. And what we saw at this conference amazed us – the conference was packed with "learning machines". As compared with last year's conference, we saw a great-leap-forward in machine learning software and hardware – and in the use of advanced analytics to explore various scientific principles. Software vendors were demonstrating advances in virtual reality, deep learning, robotics and autonomous machines. Hardware vendors were demonstrating large scale parallel processing engines that can analyze vast amounts of Big Data in record time – and, of course, IBM's Watson cognitive computing environment was represented. Data scientists were sharing the latest advances in aerospace and defense, astronomy and astrophysics, computational biology, computational chemistry, computational physics, computer vision and machine vision, deep learning and artificial intelligence, earth system modelling, embedded technologies, virtual and augmented reality, video image processing, supercomputing and high performance computing (HPC), signal and audio processing, self-driving vehicles, real-time graphics, product and building design, performance optimization, medical imaging, intelligent video analytics and even game development. It was truly amazing!

We are now at a point where machines can generate their own algorithms using machine learning techniques, and where amazingly fast systems can analyze petabytes of information in increasingly smaller time frames. In this Research Report, Clabby Analytics discusses some of the advances that we are now seeing in the area of autonomous machine learning and associated system designs. Given all of the advances that we have seen in machine learning in recent years, <u>we believe that the relationship between man and computer has now changed radically</u>.

About This Report

For six years we've been reporting about advances in business analytics, Big Data analysis, the injection of learning and predictive analytics into application performance management (APM) and infrastructure management environments, and about the role that algorithms are playing in the evolution of intelligent systems.

In short, we've been covering the technologies that are using or serving machine learning. But we've never clearly demonstrated how all of these technologies are working together to usher in a new age in computer systems – the age of intelligent learning machines. This report discusses how these technologies interrelate – and how new, Big Data-driven intelligent machines are now being used to tackle very complex business and science problems.

Machine Learning: A Radical Change

Computer systems can now program themselves. They can analyze data, create their own models, discover patterns, and generate algorithms to analyze those patterns. What this means is that computers are now "intelligent entities", and these intelligent entities are, today, opening new doors for mankind in business and the sciences.

> ***In fact, we expect most of the major scientific breakthroughs in the future to be driven by computers, no longer solely by man…***

The Dawn of "Machine Learning"

Man created the first electronic digital programmable computer in the mid-1940s. First generation models were "programmed" by humans who manually set plugs and rearranged patch cables and switches in order to enable computers to perform desired calculations. Not long after the introduction of these initial systems, humans learned how to store programs in memory – setting in motion a human-centric programming model that has lasted, and will continue to last, for decades. Meanwhile, in the background, academia and industry continuously experimented with the design of "artificially intelligent" computer systems – seeking to create computer hardware and software capable of autonomous intelligent behavior.

As far back as the 1950s, man was already trying to figure out how to enable computer systems to "learn". In 1959, one of the first pioneers in the area of artificial intelligence, Arthur Samuel, demonstrated a machine learning program that could play checkers (the computer would choose which moves to make using "if", "and" and "not" algorithms without being explicitly programed).

> ***The operative word in the preceding sentence is "explicitly". Most of today's programming involves delivering explicit instructions to a computer in order to achieve an expected result. But with machine learning, a computer can be given a model (this is called "supervised learning") and can extrapolate algorithmic solutions using this pre-supplied model; or computers can be given raw data from which patterns can be extrapolated and new models can be created (this is known as "unsupervised learning"). With the ability to use or create models – and then draw conclusions based upon the analysis of vast amounts of data – computer systems are now capable of autonomously analyzing data and drawing valid conclusions. In other words, they are capable of learning through data analysis.***

The Rift in Machine Learning

Computer scientists and theoreticians from a number of disciplines have approached to problem of artificial intelligence in computer systems from different angles. Each group developed its own model to teach computer systems how to learn. Some models mimicked human brain activity (neural networks); other models imitated evolutionary biology (genetic programming); while others used statistics-oriented approaches (driven by probabilistic inference sometimes using expert system designs). Other approaches have included an analogy-driven approach (using kernel or vector machines), and a logic-driven approach (using inverse deduction to arrive at conclusions). Given all of these different approaches, one might expect the field of machine learning to be fragmented. But, despite taking different angles, each approach has delivered hardware/software environments that can behave autonomously and act intelligently.

For an *in depth* discussion of each of this approaches, and their strengths and weaknesses, we highly recommend that our readers obtain a copy of Professor Pedro Domingo's new book entitled "The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World" – or listen to his YouTube seminar that describes the history of each discipline and how each type of machine learning works). In these works, Professor Domingos describes five differing approaches to machine learning:

1. *Symbolists* – who believe that learning is the inverse of deduction, and model their machine learning systems thusly;
2. *Connectionists* – who believe that machine learning should mimic how the brain works (and thus create neural networks, artificial intelligence systems, etc.);
3. *Evolutionists* – who model machine learning from a genetics/evolutionary biology perspective;
4. *Phaseans* – who use probability/root cause statistics as the basis for their machine learning systems;
5. *Analygizers* – who model machine learning systems by extrapolating similarities.

Each of these approaches processes certain types of applications very well – while each also has its limitations. ***Demonstrations using each type of machine learning could be found at the NVIDIA GTC event***.

Driving Machine Learning – New and Evolving Systems/Software Architectures

In the past, some of the technical challenges that these scientists faced as they sought to build artificial intelligence/machine learning environments have included:

- Not enough processing speed;
- Not enough systems memory;
- Slow networking;
- Slow data delivery; and,
- Immense power consumption.

But, in recent years, new technologies and approaches have addressed each of these problems. New generation computer systems architectures are using multi-core processors; with new memory interfaces; software-defined networking is offloading central processors from having to handle communications tasks; better data migration and management is speeding access to data (as are data handling approaches like Hadoop and Apache Spark) – and new systems are being designed to use power more efficiently (including some direct current DC solutions).

A Closer Look at Today's Learning Machine Environments

Back in 2011, *Clabby Analytics* started writing about how strategically important algorithms were becoming in the field of business analytics. From our perspective, we saw the evolving business analytics marketplace as "a battlefield" in which the companies that "deliver the most effective business analytics solutions using the most effective algorithms" would become champions. (Google, Amazon, Netflix and Walmart are all testaments to the importance business analytics algorithms integrated with business processes now play in creating competitive advantage).

But good algorithms are not the sole drivers of competitive advantage using machine learning. For the past six years, we have been reporting on the arrival of new systems designs that employ field programmable gate arrays (FPGAs), GP GPUs (general purpose graphical processing units) and enterprise class, general purpose CPUs (central processing units that are capable efficiently executing data-intensive applications found in very large memory). Examples of these architectures can be found here, here and here. Additional reports on The Now Factory, VelociData and NextScale are also available (on request).

> ***These new systems designs are using specialized processors to stream data faster (FPGAs); they are using general purposes graphical processing units (GP GPUs) to process data in parallel fashion (to deliver results more quickly); and they are using CPUs for both serial, parallel processing and compute-intensive processing. These systems can also offer massive scale and access to very large memory, enabling users to analyze very large databases in comparatively (to older designs) shorter time frames.***

Machine learning techniques are also being used in infrastructure management, operational analytics, application performance management and security/fraud analysis. Often called "predictive analytics", machine "learners" (algorithms) are being used to examine machine data looking for the root causes of problems or suspicious behaviors – and are being used to trace application activity in order to troubleshoot application behavior as well as to tune application performance.

In infrastructure management, some of the best examples of machine learning programs that we have found include IBM's zAware (a program that takes a snapshot of mainframe behavior when a mainframe is running in an optimized fashion, then compares machine data to that snapshot should performance degrade –isolating anomalies); IBM's Workload Automation product suite (report found here); CA Workload Automation iDash (report found here); IBM's Operational Analytics Log Analysis (report found here); and Virtual Instrument's infrastructure management portfolio (report found here). Notice in these reports how each company simplifies management by using predictive analytics to improve workflow and tune infrastructure performance.

> ***Also, this report discusses our view on the future of infrastructure and mainframe/distributed systems management: in short, we believe that systems will take a greater role in managing themselves – taking much of the management burden for systems troubleshooting and performance tuning off of the shoulders of humans (while lowering computer systems management costs dramatically).***

As for application performance management, *Clabby Analytics* has covered a wealth of vendors that offer predictive analytics facilities in their application performance management product sets. One of the deepest/broadest APM environments can be found in IBM's Performance Management portfolio (report found here), where predictive self-learning analytics using machine-learning algorithms are used to automate the process of identifying "normal" and "abnormal" behavior. As a result, anomalies are easily identified without requiring manual threshold setting. With early warnings of impending problems, IBM reports that outages can be reduced by as much as 50% – and administrative efficiency and application quality are improved. Machine learning plays a big role in reducing these outages. Other reports on APM tools that have been augmented with machine learning include this report on Ruxit (found here); this report on Aternity (found here); this report on Dynatrace (found here); and this overview of the APM marketplace in general (our Computerworld article can be found here).

In security, we found a comprehensive machine learning-based predictive analytics environment as part of IBM's Smarter Counter Fraud initiative (report found here). This initiative integrates various sophisticated security and analytics tools – joining them together with IBM infrastructure offerings and linking them tightly with Big Data databases such that large volumes of data can be analyzed quickly. This is a huge step in the right direction to counter fraud because it puts in place a more effective, integrated architecture across which multiple tools can work together in concert to help identify and overcome fraud. Machine learning plays a central role as part of this initiative. This report on advanced security methods and tools also discusses the use of predictive analytics to improve systems security.

> ***In short, machine learning in the commercial computing market has blossomed. There are a wealth of new systems designs, infrastructure management, operations management, application performance management and security tools that make heavy use of machine learning to troubleshoot problems, to improve performance and to look for security anomalies. It is reasonable to expect the use of machine learning to expand in all of these markets over time as new algorithms are developed to further improved machine learning analysis and performance.***

A Closer Look at the Machine Learning Systems Architectures on Display at the GTC

The exhibition/demo floor at the Graphics Technology Conference was rich with demonstrations of computer hardware and software. We saw vector machines, deep learning machines, virtual reality environments and many more examples of systems designs that can support the analysis of Big Data to drive business and scientific outcomes.

From a systems perspective, the new generation machine learning environments need to be fast. They need to take mountains of data, look for patterns, and make sense of that data. Several years ago, when individual microprocessor technology performance peaked in the 5 GHz range, computer makers started to focus their efforts on optimizing systems designs in order to continue to drive processing performance. The number of cores per socket were increased; the number of threads per core increased; the amount of main memory that could be addressed increased (as did on-chip cache); memory bandwidth speed increased and new input/output methods were introduced (CAPI, QPI, PCIe Gen3, NVLink and others).

At the GTC, several IBM POWER-based servers were on display – as were several systems designed using GP GPU processors from NVIDIA. Both types of processors handle large numbers of threads per core – and both support access to very large memory and underlying, FLASH-based fast storage.

> ***What we found particularly interesting is that GPUs, formerly focused only on graphics performance, are taking on a more general purpose role. GPUs excel at handling parallel processing – but new facilities are now being added to GPU chips to enable those chips to handle more computationally-oriented tasks. Could it be that the microprocessor market (that is dominated by Intel, ARM and POWER RISC designs) is going to see a new source of competition in the form of general purpose GPUs?***

In recent years, computer makers have also focused on the scaling of distributed systems. Numerous powerful new high performance computing (HPC), super-computing and hyperscale system configurations were on display at the GTC. These configurations have become more and more powerful every year as new microprocessors and faster, intelligent software-defined communications links have been introduced – leading to greater scale and faster computing speed. Several vendors showed their HPC and hyperscale designs on the exhibition floor.

Hybrid/accelerated systems" designs started to come to market around four years ago, combining graphical GPUs with general purpose CPUs (usually Intel and POWER processors) and with FPGAs for streaming data at line speed and other uses. With new accelerator system designs, businesses and scientific communities can now pour through very large volumes of Big Data faster than ever before. Several hybrid accelerated systems were on display at this year's GTC.

In addition to computer system performance and scalability issues, another challenge that has impeded system performance over the past several years has been network performance. In this report we discussed how a memory management architecture (RDMA), when used with fast Mellanox 40 Gbps adapters and switches, is having a major positive performance impact on application and database performance. But at the GTC we also learned that Mellanox is working on new software-defined networking algorithms that can alleviate the need for CPUs to handle a lot of packet traffic – and by so doing, help further improve overall system performance.

Finally, at the GTC (at a conference within the conference known as the OpenPOWER Summit) we learned that Google and RackSpace are working on a new 48 volt DC system design that has the potential to lower system power consumption by 30%.

> ***New system designs are making it possible to process Big Data exponentially faster than previous generation designs. Accordingly, system performance challenges in processing speed, access to large memory, slow networking, slow data delivery and power consumption are all being addressed. Next generation system designs are making it possible for systems to expeditiously analyze vast amounts of Big Data – leading to advances in machine learning-driven self management, autonomous security, and to the advancement of the sciences. We can't wait to see what next year's designs will look like!***

Scientific Advances

The first thing conference attendees notice when entering the San Jose Convention Center to attend the GTC conference is poster boards. Yes, low tech poster boards. About a hundred of them, organized by topic (such as HPC or robotics) – displayed in rows (like the stacks in a library). A closer look at these poster boards shows a description of a scientific computing problem – and how that problem was solved using particular processor/system designs. (Examples: A GPU-Accelerated Statistical Method to Identify Differential Genetic Dependencies, or N-Body Simulation of Binary Star Mass Transfer Using NVIDIA GPUs, …).

These posters were fascinating. In short, they stated a science or business computing problem; they articulated how that problem was analyzed; they showed in graphic terms how the problem was solved (and specific algorithms were shared); and then they wrapped with a conclusion (a restatement of the problem, how the result that was achieved, and what was the end result). Essentially, these posters (about 100 of them) represented case studies from numerous scientific fields of endeavor. And they demonstrated that scientists are making heavy use of machine learning techniques and algorithms to progress the sciences.

> ***It was clear to us that learning machines are now becoming indispensable in several sciences. Last year we learned that a computer system (Eve) was given the task of screening mountains of data to help discover a cure for a particular strain of malaria. Using artificial intelligence machine learning techniques, Eve was able to help isolate a compound that could be used to fight this malaria strain. Could it be that learning machines might someday be able to do the jobs of computational chemists in drug discovery at a fraction of the cost? Will learning machines someday become indispensable in other sciences? After attending the GTC and viewing the poster boards, there are clear indications that this will one day becocme the case…***

Wherefore Art Thou, Watson?

It has been several years since *Clabby Analytics* has taken a close look at IBM's Watson cognitive computing environment. In 2014, we wrote that IBM had planned to invest $1B in cognitive computing – and the company revealed that 30% of the IBM Research budget and 2000 people will be devoted to the effort. We also noted that IBM was focused on giving Watson the power to: 1) "reason" using a decision tree approach that could will enable Watson to look for evidence to support hypotheses to form strategic arguments; 2) "see" by looking at related images, analyzin them and looking for anomalies using reasoning in a complex domain; and, 3) "empathize" could give Watson the ability to look at an individual's linguistic footprint (emails, tweets, facebook posts etc.) to create a personality profile for "yourself and others", allowing users to customize their interactions with that person (also called "personality analytics").

In 2014, we also wrote that IBM Watson Research would be exploring the use of:

- *SyNAPSE Neurosynaptic Systems* – these systems will replace current architectures that are not optimized to handle new computing paradigms and data types (sensor data for example) for the kind of learning developing in cognitive computing. These systems are "brain-inspired" with the goal of reducing the power and time required to compute on very large data sets;

- *Mockingbird* – a cognitive learning and messaging system based around identifying ideas that are relevant to a given community – particularly through social media such as Twitter and on-line discussion forums;
- *Glimpse* – Using contextual learning and expansion of concepts to help the SME collect and share data and use evidence and analytics to make new discoveries (operating at the discovery level of cognitive computing);
- *Piazza* – a deep contextual search that uses not only keywords but is able to identify relevant information in a search based on a relationship to the search query rather than just the keywords themselves. Use cases include an intellectual property search or discovery of evidence in legal investigations; And,
- *MOOV* – a tool that enables people to make complex decisions effectively, taking into account trade-offs and combining visualizations with analytics. For example, a retailer could use MOOV to design an optimal promotion plan that could maximize sales volume while maintaining desired margin and revenue targets

> ***Watson is an architecture that combines several of the approaches described on page 3 to deliver analytics and machine learning results. It is a different type of learning machine than the scientific and business machines describe in previous sections of this report – and, accordingly, our Watson coverage warrants an update. Expect a Clabby Analytics report on the current state of Watson to be issued next month.***

Summary Observations

When we think of machine learning, we think about how humans learn. We humans are constantly flooded with mountains of data (sound, sight, colors, haptics, smells, …) – and we have to figure out what this data means and decide what to do about it. Machine learning mimics our learning patterns to a degree. Computers can evaluate huge amounts of structured and unstructured data, and can now scale to an extent to which they can analyze petabytes and even exabytes of data (see this 10 exabyte system example). Computers can now find patterns in data, create models for analysis, and create algorithms to dig deeply into large databases. They can then present a most likely correct answer to a given query. Like us, computers "think". Accordingly, machine learning computers represent a new tool for advancing our causes in business and the sciences.

For decades, programming has been the major bottleneck in building computing solutions. Programs instruct the computer what to do; they need to link with other programs to flow business processes; they need to be able to talk with highly varied underlying infrastructure in order to communicate – the process of building enterprise class applications is often a daunting task for enterprises as well as for independent software vendors (ISVs). Programmers have had to learn multiple programming languages; they've had to learn to support multiple device types; and they've had to write *highly-explicit* code in order to achieve desired results. With machine learning, code does not have to be explicit. Using machine learning techniques, a lot of the burden of human programming can be eliminated.

As stated on page 3, there are five approaches to machine learning that involve inverse deduction, backpropagation, genetic programming, probabilistic inference and kernel point vector machines. Individually, these approaches can be used to solve specific types of problems; collectively, they can be combined to solve other types of problems (example: IBM Watson).

From a business point-of-view, it is important that enterprise executives understand the impact that machine learning will have on competitiveness, on operations management and on security. Enterprises that have failed to identify the trend toward machine learning – and have failed to recognize that new generation systems designs are creating major performance advantages – risk be usurped by competitors who are exploiting these technologies. Enterprise executives should also take time to understand the new computer architectures that support machine learning Big Data processing – because these hybrid, accelerated systems offer more expeditious, more scalable and more cost efficient alternatives to traditional computing systems.

From a science perspective, scientists who are unaware of the advantages of machine learning will be restrained by the need to write their own algorithms to solve complex problems while fellow scientists using machine learning techniques solve the same problems exponentially faster. A new generation of learning machines is upon us – providing us with new tools to advance businesses, sciences and technologies. We are no longer the only advanced learning entities on the planet – machines now have intelligence.

> ***Over the past few years, humankind has crossed over the artificial intelligence/machine learning threshold with the arrival more advanced analytical algorithms, with the maturing of different machine learning techniques, and with the availability of extremely powerful, hybrid (accelerated) systems. Computers, with increasing regularity, are now being used to analyze vast amounts of data and generate and execute self-generated algorithms. Machines have learned independently analyze data and write program with minimal or no human intervention. Scientists and enterprises need to prepare for an invasion of learning machines – learning machines have arrived and they are here to stay!***

Clabby Analytics
http://www.clabbyanalytics.com
Telephone: 001 (207) 239-1211

Clabby Analytics is an independent technology research and analysis organization. Unlike many other research firms, we advocate certain positions — and encourage our readers to find counter opinions —then balance both points-of-view in order to decide on a course of action. Other research and analysis conducted by Clabby Analytics can be found at: www.ClabbyAnalytics.com.