Bringing Big Artificial Intelligence to Life

Ankur Patel
Data Science Program
Saint Peter's University, NJ, USA

***Abstract* –** Data is rampant in the Internet of Things (IoT) Age, during which the exponential growth of data has outpaced the capacity of traditional computing. It has reached maturity in some stages, but it is in adolescence in new stages. With the advent of Big Data, businesses are merging it with big compute and IoT for analytics using Artificial Intelligence (AI). After the raw input from the big data gets cleaned, structured, and unified, AI performs cognitive functions and outputs values for the business. With its ability to analyze massive amounts of data in milliseconds, it can now be processed in "real-time". In data science, the "hypothesis-first" approach has been moved to "data-first" approach. In the near future, big AI will be ruling many different industries and also in software. It is already an emerging asset in health care, finance, management, education, transportation, and manufacturing, and it is revolutionizing the operations. Since computers can solve problems, they compare any information and decide what it signifies. The human brain consists of nerve cells or neurons, which constantly transmit and process information received from the senses in milliseconds. Likewise, deep learning architectures use multiple layers of artificial neural networks on input data to abstract and composite representation. Although mimicking the human thought process is far away, robotics is a growing field of research and design with the goal to recreate human intellect. Following that, the superintelligent machines – as new species – can hold tremendous advantages in mental capability, a vastly superior knowledge base, and the skills to multitask. This paper reviews the current status of the methods – such as artificial intelligence, machine learning, and deep learning – and their automation's result in superintelligence.

I. INTRODUCTION

The growth of data is enormous at every moment, and its size is increasing exponentially. Big data is taking the world due to its growth and large volume. As it continuously grows, it becomes more meaningful and pertinent for big data analytics. As the large and varied data sets are processed and analyzed, information and the patterns are uncovered. That helps the companies make informed business predictions and decisions.

The large sets of data were almost impossible to process using the tools in early 21th century, until Apache Hadoop was built by Yahoo! on top of Google's MapReduce [1]. The open-source Hadoop facilitates multiple computers to network and crunch through the large data using MapReduce, which is a programming model for big data, to solve problems. The data processing methods of reading, performing, and writing the operations back and forth from the cluster are extremely valuable. Besides processing the large data much faster, it also effectively provides fault tolerance, which enables the system to operate properly in the event of failure. Spark operates similarly on huge datasets, but also provides a distributed file system that allows real-time and in-memory processing. The expanded application of big data analytics has resulted in a massive increase in startups that understand and adopt big data [1].

As human capabilities are being replaced or enhanced by machines, the artificial intelligence is what constructs the automation for breakthrough results. Artificial Intelligence, referred to as machine intelligence, is intelligence presented by machines in contrast to natural human intelligence. Understanding decision-making by AI, as researched by Future of Life Institute (FLI) [2], has been funded to manage the growth of technology. As the initiatives of big data mature, companies and organizations combine big data processing and AI to accelerate their business values. Their convergence has developed a significant impact to drive the company's business value. Down to its core, AI describes the dynamic process of a machine to resolve and conclude based on logic. It performs these processes on big data to determine its meaning and relevance for the company.

Machine learning is a subset of AI that without minimal human intervention. It consists of algorithms that take in the data, perform calculations, and deliver the correct answer in the most efficient manner. In the realm of big data, machine learning and AI are used interchangeably. Since it is now possible to process the data streams in real-time, it moves machine learning into the same direction to control real-time data. Instead of depending on the representative data samples, the data itself can mine and find relevant information. Machine learning and AI has moved from research labs to production phase.

The clusters of Hadoop and Spark, as mentioned earlier, can also be leveraged for deep learning. Powerful tools of deep learning are BigDL library, which provides deep learning applications, and Math Kernel Library (MKL), which contains mathematical functions on the basis of machine learning algorithms for optimized performance. Deep learning is the engine that uses these frameworks, among others, to propel the science behind machine learning and AI. Another series of algorithms are neural networks, and they process data like the brain to make sense of the information. The concept of deep learning is simply multiple layers of neural networks nested together, sometimes referred as "deep neural network".

The structure of this paper is as follows. Section II introduces big data and its statistical power, higher complexity, and analytics. Although big data analytics offers great statistical power, its higher complexity can lead to false discovery. Section III explains the emerging AI. The greater volumes and sources of data are enabling capabilities in AI, as well as evolving the grounds of machine learning and deep learning. The ground concepts will not be explained in much detail in this paper because of their comparably tremendous size. Big data analytics and AI are huge topics themselves, but because of their significant connection, their schemas will be explained. Section IV summarizes the performance of current methods, predicts data science in the future, and glimpses the incoming superintelligence.

II. BIG DATA

Now, the computing environment for big data has expanded to include various systems and networks. As its three major characteristics – volume, velocity, and variety – are expanding, the rate of data creation is accelerating, as seen by the items in Figure 1[7]. Due to scaling up for more powerful servers, the computing resources are provided by clusters for the massive computing units. This is talking about warehouse-sized computer with thousands, if not more, of datacenters. In this section, the power of big data analytic, which includes the frameworks of Apache Spark, will be explained. Additionally, the paper of "Scaling Big Data Mining Infrastructure: The Twitter Experience"[9] and Target, a U.S. retailer, using sales and marketing analytics[7][8] will be discussed as it exemplifies big data analytics really well.

Figure 1: Data Deluge[9]

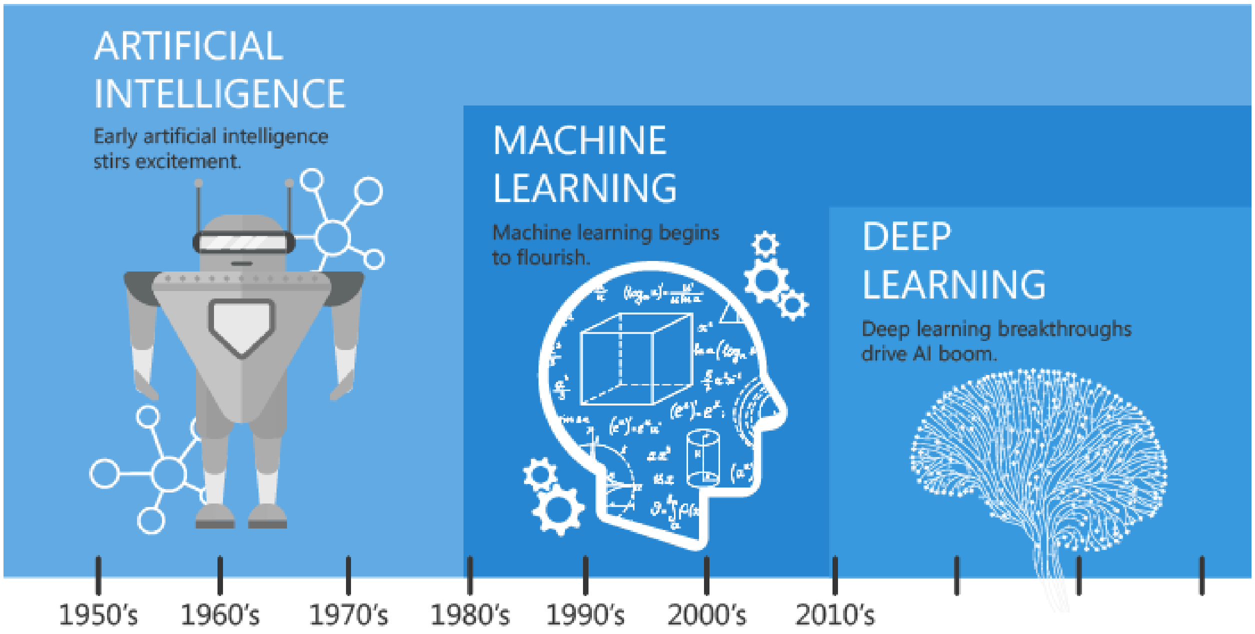
MapReduce is a framework used for simplified data processing on large clusters through 2 tasks – Map and Reduce. A good example to understand the process is making sandwiches. Suppose we want to make 3 sandwiches- ham, turkey, and Italian. A list of the ingredients needed is gathered in the Map phase. The total amount of each ingredient is determined during the Shuffle and Reduce phase. Generally, the input data in Map is stored in the Hadoop File System (HDFS), and the output set after the Reducer process is stored in HDFS.

As the biggest and most widely used software, Apache, which includes Apache

Hadoop in the collection of numerous softwares, is an open source web server that runs on 67% of the webservers globally [6]. In 2009, Apache Spark project designed a unified system for distributed data processing. It can be looked at as an enhanced version of MapReduce that is used as a data-sharing structure called Resilient Distributed Data (RDD), as it performs 100 times faster for a few applications with in-memory primitives and up to 10 times when accessing. Previously, separate engines were needed to process a range of distributed workloads, and now, they can be run as libraries with a common engine. Using a unified application processing interface, or API, for processing different tasks will provide efficiency. A great example is smartphone: it combines the functions of camera, cellphone, and GPS so we can use the one device with 3 functions. The API is the software intermediary that allows the mentioned application programs to run, interact with each other, and share data. They dictate how the programs tap into the big web services – social networks such as Facebook or Twitter or utilities such as Google Maps or Dropbox[11]. Developers get attracted by the best set of APIs for operating on large datasets, which Apache Spark offers – RDDs, DataFrames, and DataSets across languages of Scale, Java, Python, and R. Because of Spark's wide range of applications, companies from various different industries use it. Spark is used for interactive queries, real-time stream processing, and scientific applications. It is a unified data-processing engine that is deployed in diverse environments, and the libraries' details are open-source[1].

The paper of "Scaling Big Data Mining Infrastructure: The Twitter Experience" provides knowledge of big data mining infrastructure that Jimmy Lin, an Assistant Professor who spent an extended sabbatical from 2010 to 2012 at Twitter, and Dmitriy Ryaboy, who was a tech lead first then an engineering manager of infrastructure at Twitter, wrote from their Twitter experience. The two topics described were schemas and heterogeneity. Firstly, schemas alone are insufficient for data scientists to get an overall understanding of the data. Scribe is a log-transport mechanism that Twitter uses for "aggregating high volumes of streaming log data in a robust, fault-tolerant, distributed manner"[9]. As the core of the log-collection strategy, it is an open source software created by Facebook to collect logs from thousands of web servers[12]. Another major challenge was the heterogeneity of the various components that must be integrated for production workflow. As Jimmy Lin mentions in a talk[9], data cleaning and data munching take 80% of the time, and since that takes data scientists' majority of their time, he shares his experience for future data scientists. He says, "Schemas aren't enough! We need a data discovery service!" Finding data easier, rather than through a complex path, is done by Data Access Layer (DAL), a loader that knows via metadata how to access it. Then, the heterogonous components are synchronized by "plumbing" so the data runs smoothly. The components are wired together via different channels, and the quality must be good to prevent clogging. A successful big data analytics platform is achieved by balancing speed, efficiency, flexibility, scalability, robustness, etc.

Target used big data analytics to drive new revenue after analyzing customer behavior. Their statisticians determined Target's profits depending on three main life-event situations, as following. 1. New products bought after Marriage; 2. New products and spending habits changed after Divorce; 3. Many new things with urgency after pregnancy [7]. The most lucrative of these time-event situations was pregnancy. Target was able to predict the customer's pregnancy state from their behavior towards products. Such analytics resulted in the retailer's knowledge for segmenting the customers and offering specific coupons. Besides the support in marketing, it also helps with the retailer's management of the inventory, according to the time of demand. Since the expansion of big data causes changed in markets, companies find creative ways to comply and reach their business needs.


ARTIFICIAL INTELLIGENCE
Early artificial intelligence stirs excitement.
MACHINE LEARNING
Machine learning begins to flourish.
DEEP LEARNING
Deep learning breakthroughs drive AI boom.
1950's
1960's
1970's
1980's
1990's
2000's
2010's

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

Figure 2: Artificial Intelligence, Machine Learning, & Deep Learning[10]

III. ARTIFICIAL INTELLIGENCE & SUBSETS

Artificial Intelligence (AI) is an area of computer science that is used to construct intelligence in machines, so they to act and react like humans. It was inspired by reverse engineering, such that machines can process, reason, and decide from the information, similar to neurons in the human brain. Using the networks of neurons in neuroscience to engineer the models, the machines execute those models and the "thinking" process. Ever wonder how the e-commerce websites recommend quite accurate products or services based on your personal habits and tastes? Or how Chatbots or similar programs assist as virtual customer service? It all falls into a reason why online shopping got so popular and addictive. Yes, it is AI that automatically predicts and suggests to the customers after analyzing abundant information collected by e-tailers over years. Apple's Siri and Amazon's Alexa are perfect examples, since everyone has surely experienced their AI.

The intelligent machines that were created to work and act like humans have intelligence in different areas. They have the abilities of understanding the natural language, learn and adapt problem-solving, perception, modeling, robot ability when combining the ones above and even structured games. They must have access to the information to perform knowledge engineering, which is a core part of AI research. It analyzes the data mathematically with machine learning algorithms, a well-defined branch of computer science. Machine learning -a core part of AI itself, as seen in Figure 2- is a methodology that is supervised for classification, while mostly unsupervised for pattern recognition. As AI is unsupervised, it analyzes the visual inputs with sub-categories of facial objects and gesture recognition, and it also does motion planning and mapping for itself. Its techniques are pervasive and too many to list here, but a few will be written here to get a good understanding.

AI encompasses a diversity of disciplines to form a valid solution. Figure 3 lists its different disciplines with a small description [13], which gives a general but better understanding of AI.

Robotics	Make objects travel in space
Algorithm Theory	Construct efficient algorithms
Statistics	Analyze past results, derive useful results, predict the future
Psychology	Model the human brain’s functions
Software Engineering	Create endurable solutions that sustain the test of time
Computer Science	Enforce the software solutions in practice
Mathematics	Execute complex mathematical operations
Control Theory	Create feed-forward and feedback systems
Information Theory	Encode, decode, compress, and represent information
Graph Theory	Represent an optimized model’s hierarchies of different points in space
Physics	Model the real world
Computer Graphics & Image Processing	Process and visualize images and videos

Figure 3: Artificial Intelligence Disciplines [13]

For example, robotics requires complex level math to understand what’s happening. It is based on control theory, where you control the object’s movement by the feedback gathered from the loop. Speech recognition system, for example, can use speech synthesis to convert internal data into sounds that humans emit or understand. A recently common method for understanding language is natural language processing (NLP). Using NLP, machines are able to understand when and how humans speak. It is commonly used for text mining, machine translation, and automated Q&A. Its algorithm is based on machine learning algorithms- it automatically learns by tuning and evaluating the data samples and makes a statistical inference. There are some open-source libraries that provide the algorithmic building block for real-life applications of NLP. They are: Apache OpenNLP, which is a machine learning toolkit with tokenizer, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, coreference resolution, and others; Natural Language Toolkit (NLTK), which has modules for processing text, classification, tokenizing, stemming, tagging, parsing, and others; Standard NLP, which provides tools for part-of-speech tagging, named entity recognizer, coreference resolution system, sentiment analysis and others; and MALLET, a Java package that provides Latent Dirichlet Allocation, document classification, clustering, topic modeling, information extraction, and others [5]. That blog particle also provides numerous useful books, tutorials, videos, and courses for NLP. Such automatic calculations have also added to the growth of automobile industry and self-driving cars – Tesla, Cadillac, Mercedes, BMW, and more. The AI agent in the cars process and sense its surroundings using lots of physics, especially when it comes to moving objects. Even before self-driving cars, the mean shift algorithm, which is a hierarchical clustering algorithm, were built into the car detection software where the alarming sounds or emergency brake is applied in certain situations [13]. Now, the central computer system in the self-driving cars analyzes all the data from their sensors by running machine learning and making predictions, and they administrate the steering, accelerating, and braking. It must occupy highly detailed maps of street features, and also be able to read real-time surroundings. Furthermore, although it’s running yet, the vehicle-to-vehicle (V2V) technology will enable the autonomous vehicles to communicate important details. It will also be tied tightly to the IoT to implement real-time vehicle health.

Like mentioned earlier, the concepts of the neural networks in the human brain are used to model and process similarly in computing.

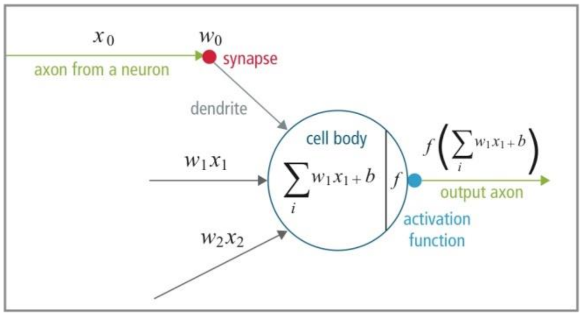


Figure 4: Neuron’s mathematical model [18]

Figure 4 shows the inner-workings happening of 100 billion neurons in the brain [18][15], or the artificial neural network (ANN) of 1 million neurons built by Sparse Evolution Training

(SET) [16], for example. In simple words, a neuron communicates with other neurons in the interface in the condition that the weighted input signals' summation is high enough to transmit. In computing and healthcare terms, large ANNs built on even supercomputers now can "hypothetically lead to better healthcare and affordable personalized medicine for all of us," says Dr. Decebal Mocano. As studied by The Nutritional Immunology and Molecular Medicine (NIMML), "a leading lab of at the Biocomplexity Institute of Virginia Tech and BioTherapeutics", the AI algorithms can build a synthetic population of patients with certain problems, create a personalized model of drug combinations, and determine the best treatment solution. Their new computational pipeline to predict efficacy on their model for Clostridium difficile infection uses machine learning to translate preclinical research on animal models to identify effective treatments and predict optimal doses of drugs [16]. Like hospitals already use AI for efficiency in procedures and treatments, there is also a growth in usage of AI and machine learning in therapeutics and other healthcare areas.

Of course, with the wide range of tools and techniques in AI and due to its increasing growth, it is expected to encounter many problems, such as "issues in parsing, text generation, understanding spoken language, search routines, automatic programming, semantic network models of memory, automatic deduction, vision, learning and inductive inference, planning and problem solving, robotics, and so forth" [3]. With the growing field of AI, its research has also increased; for example, Google supports research publications to advance the future of computer science, including AI. If the research makes significant headway, it would result in plenty of productive outcomes. Teachers, for example, can be helped or supported by AI, which score their teaching abilities if it can process and evaluate natural speech automatically. Since it can store a tremendous amount of data, or better yet, be connected to internet and a cloud server, it can take the role of teachers with all that memory of knowledge.

VI. CONCLUSION

The exponential growth of data along with the expansion of technology has enabled numerous abilities for companies and organizations in many industries. The increasing growth rate of data in machines extends beyond data centers and into the cloud in a hybrid environment. Using the previously mentioned methodologies, the preprocessing and the analytical systems of big data are now in high performance computing. Artificial Intelligence outperforms human scientists and mathematicians, and it can perform complex activities that humans would need professional expertise for. Using big data, it has revolutionized the world of business, healthcare, aviation, and academia among many others. With the rage in the topics and subtopics talked about in this paper, the data and its calculations for the tasks must use current and efficient concepts.

This brings us back to the introduction of artificial superintelligence, and how it speculates the world with a cognitive ability that is superior to a human's. The technology of virtual assistants and self-driving cars, for example, are still in the early days of development. Nick Bostrum argues that machine intelligence surpasses that of humans, and that it could replace humans as the dominant lifeform[17]. That has been seen in several Hollywood movies as well, which foreshows its chances; a famous quote of "I'll be back" can also be seen as indicative.

REFERENCES:

[1] van Rijmenam, Mark. “A Short History Of Big Data.” *Datafloq - Connecting Data and People*, Jan. 2016, datafloq.com/read/big-data-history/239.

[2] *Future of Life Institute*, Jolene Creighton Https://Futureoflife.org/Wp-Content/Uploads/2015/10/FLI_logo-1.Png, futureoflife.org/team/?cn-reloaded=1.

[3] Freedle, Roy O. *Artificial Intelligence and the Future of Testing*. Routledge, 2016.

[4] Kersting, Kristian, and Ulrich Meyer. “From Big Data to Big Artificial Intelligence?” *SpringerLink*, Springer, 31 Jan. 2018, link.springer.com/article/10.1007/s13218-017-0523-7.

[5] Kiser, Matt. “Introduction to Natural Language Processing (NLP).” *Algorithmia Blog*, 2 Jan. 2018, blog.algorithmia.com/introduction-natural-language-processing-nlp/.

[6] Wpbeginner.com. (2019). *What is: Apache*. [online] Available at: https://www.wpbeginner.com/glossary/apache/.

[7] Dietrich, David, et al. *Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. Wiley, 2015.

[8] C. Duhigg, The Power of Habit: Why We Do What We Do in Life and Business, New York: Random House, 2012.

[9] Lin, Jimmy, and Dmitry Ryaboy. *Scaling Big Data Mining Infrastructure: The Twitter Experience*. SIGKDD, 2013, *Scaling Big Data Mining Infrastructure: The Twitter Experience*.

[10] Ray, Tanmoy. “Demystifying Neural Networks, Deep Learning, Machine Learning, and Artificial Intelligence.” *Stoodnt Blog*, 8 June 2018, www.stoodnt.com/blog/ann-neural-networks-deep-learning-machine-learning-artificial-intelligence-differences/.

[11] Proffitt, Brian (September 19, 2013). "What APIs Are And Why They're Important". Readwrite. Retrieved 28 October 2015.

[12] A. Thusoo, Z. Shao, S. Anthony, D. Borthakur, N. Jain, J. Sarma, R. Murthy, and H. Liu. Data warehousing and analytics infrastructure at Facebook. In SIGMOD, 2010.

[13] Nagy, Zsolt. Artifiicial Intelligence and Machine Learning Fundamentals; Develop Real-World Applications Powedered by the Latest AI Advances. PACKT Publishing, 2018

[14] Team, Editorial. “How Artificial Intelligence Will Disrupt Online Shopping!!!....” *Finextra Research*, Finextra, 19 July 2018, www.finextra.com/blogposting/15571/how-artificial-intelligence-will-disrupt-online-shopping.

[15] Castrounis, Alex. “Artificial Intelligence, Deep Learning, and Neural Networks Explained.” *InnoArchiTech Provides Education, Writing, and Speaking Services Focused on Leveraging State of the Art Advanced Analytics to Transform Data into Value | Founded by Alex Castrounis*, InnoArchiTech, www.innoarchitech.com/artificial-intelligence-deep-learning-neural-networks-explained/.

[16] “New AI Method Increases the Power of Artificial Neural Networks.” *Phys.org - News and Articles on Science and Technology*, Eindhoven University of Technology, phys.org/news/2018-06-ai-method-power-artificial-neural.html.

[17] Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2017.

[18] Hijazi, Samer, et al. “Using Convolutional Neural Networks for Image Recognition.” *Cadence Design Systems, Inc*, 2015.