Data Collection & Analysis of Emerging Artificial Intelligence & Machine Learning Techniques and Technologies

***Abstract*—**This paper is a review of the literature regarding the topic of emerging artificial intelligence (AI) and machine learning (ML) techniques and technologies. The basics of AI and ML will be covered as well as where it is/may be headed in the future. By compiling data on the topic from different sources, we hope to be able reach greater insight and understanding of said topic. Initial research has revealed that both AI and ML techniques and technologies have been expanding more and more as machines are beginning to solve more complex problems. Even global issues such as Covid-19 are now being eased with AI technologies. As the potential capabilities of AI and ML are being recognized, more support and research has gone into the fields as well. However, there are often compatibility issues between the many different ML technologies that have been developed. Training more complex models also takes more time and data, which can create issues in and of itself. The data can be poisoned to cause strange behavior in an AI or even cause it to be easily manipulated by a malicious party. These are just some of the problems that must be solved when creating AI and ML models in the future.

Keywords—Artificial Intelligence (AI), Machine Learning (ML), Learning Techniques.

I. INTRODUCTION

AI is short for artificial intelligence and was originally designed to follow a specific set of rules that would tell the computer what to do next depending on the current scenario. The biggest problem with this is the more complex the task and scenarios the code gets exponentially larger. The best solution to this problem is not to tell the computer what to do but to teach it what to do. This is accomplished by marching learning. Machine learning is a way of coding a computer so that it takes in data and begins to process that data in order to formulate a result. Even with this technique AI is still narrow, meaning it has a shallow skill field unlike a human. A human can drive a car then switch to writing a poem or playing chess. While a computer that can play chess does not even register what a car is, the upside is that the computer will never lose at chess. Computers right now are masters of one while humans are the jack of all trades (p. 4).

Machine learning replaces the basic if-then instructions of the rule-based system and uses a system of data analysis and pattern recognition. The key word being data, as without data the machine cannot learn and if it cannot learn it cannot act. Often when machine learning is used, as more data is given to the machine, its output will become better and more accurate. There are four main learning methods that are used in AI programs today: unsupervised learning, reinforcement learning, deep learning, and generative adversarial learning, but each of these may have subsets that extend further (p. 5).

Machine Learning vs Artificial Intelligence

While many people and even corporations use the terms "machine learning" and "artificial intelligence" interchangeably, they actually have separate meanings. A machine learning algorithm generally focuses on solving a single, specific problem through pattern recognition. Once an algorithm has been trained to solve the problem, it cannot solve a different problem without being completely retrained first.

Artificial intelligence is more of a broad term that machine learning is only a subset of. Technically, artificial intelligence should be more generalist in its problem solving capabilities, similar to the human brain. As the name suggests, artificial intelligence should be, or at least appear to be, intelligent. This means that an AI would be able to utilize abstract thinking and new strategies to solve numerous different problems. Currently, AI is considered "weak" or "narrow" as it can only imitate intelligence. Even if an AI is taking different inputs and returning correct outputs, it does not mean that it has an understanding of what it is doing; it just does.

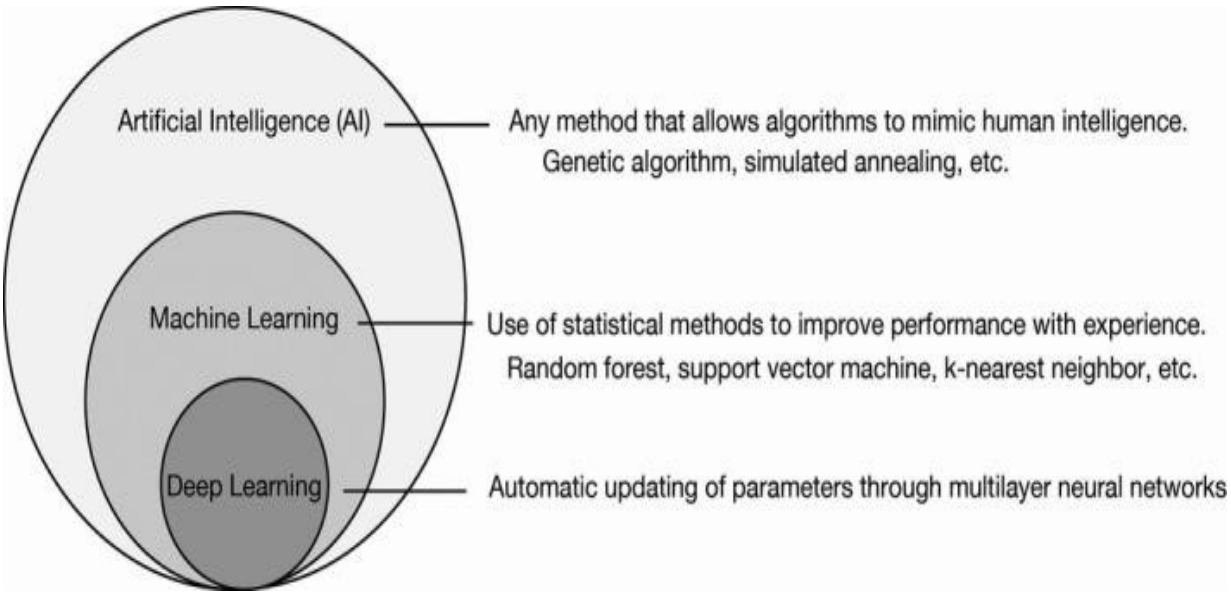

Figure 1

The relationship between AI, ML, and DL (Shimizu et al., 2020)

A popular case for describing weak AI is the Chinese room argument, which states that just because something appears intelligent does not mean it actually has a true understanding or consciousness. For a modern example, a language translation algorithm may be able to translate sentences with decent accuracy, but it has no understanding of the meaning behind the words it translates. In the future, "strong" AI could potentially be made self-aware and be intellectually on par with humans. However, this level of AI is still quite far off if it is even possible at all.

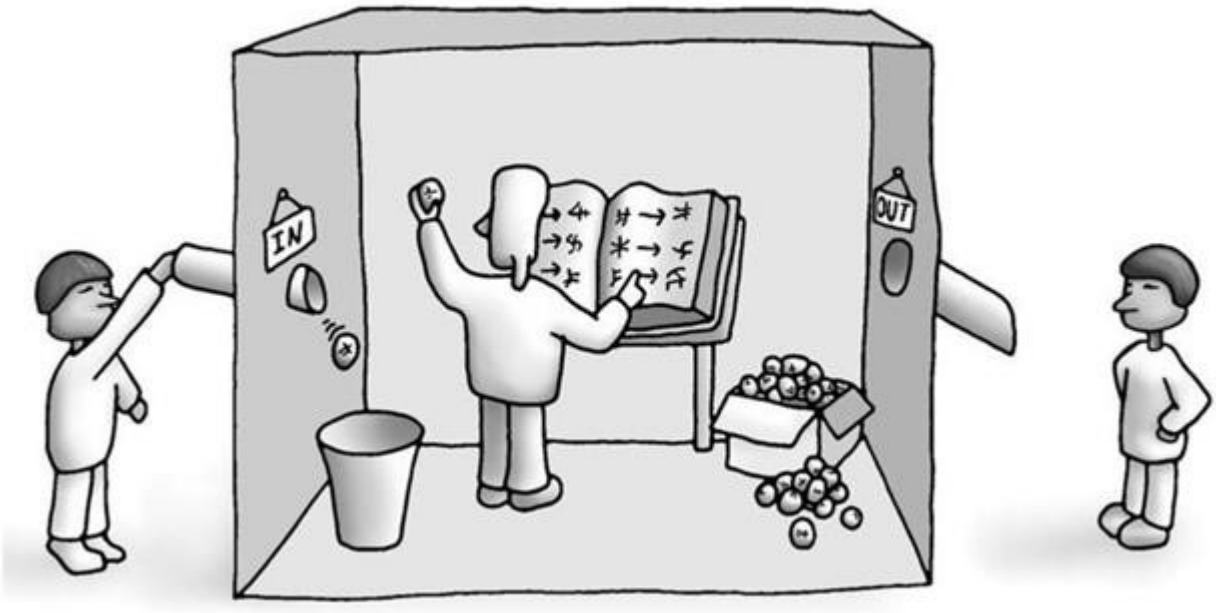

Figure 2
The Chinese room argument (Wikicomms, as cited in Karbhari, 2019)

II. LITERATURE REVIEW

Unsupervised Learning

Unsupervised learning takes data that is unlabeled and uncategorized, and the algorithm finds similarities in them and sorts the data into groups. For example photos of dogs and houses and cars could be showed to a computer and it would begin to recognize differences in the photos and start to place each one on the correct category even without know what the photo is of, just like a human could be shown picture without know what it was of and place the photos in categories. The more data given to algorithms greatly increases the accuracy of the categories. This type of learning can help find anomalies in irregular data groups, such as credit card purchases. A person might purchase goods without thinking of a pattern. However, patterns do form and a computer can notify when outliers occur, such as multiple large purchases close together or a purchase from a location that is not near the others (p. 5).

Reinforcement Learning

Reinforcement learning takes information learned from the environment to train the machine. Just as a human learns from either pain or pleasure the computer does the same. As the computer attempts actions it takes outcomes as either bad or good and will make minor adjustments to slowly perfect the action. As more data is received the better and better the outcome gets. When a computer is set to play a game such as brick breaker it is told that points are good and losing lives are bad. Over time the computer learns more and more techniques that increase the score and decrease the lives lost. This learning technique is useful for finding new and different strategies for scenarios like games that have a set environment and easily definable goals (p. 5).

Deep Learning

Deep learning (DL) uses neural networks, which are inspired by biological neurons, to interpret data and to learn. The network takes in data and passes it through each neuron and eventually gets to an output layer. Some neural networks can contain hundreds of thousands of neurons and tons of layers to them. The algorithm learns by tracking the links between each neuron and changing the weight for different outcomes. Neural networks can use different learning methods of machine learning such as unsupervised if the data is unlabeled or reinforcement if the data is received from an environment. For image recognition each pixel of an image is passed through the neural network. This learning is used not only in image recognition but also in the medical field to predict outcomes (pp. 5-6).

Generative Adversarial Learning

Generative adversarial learning puts two computers using neural networks against one another in different scenarios. One example of this is Synthetic data recognition, where AI attempts to duplicate given data while another tries to recognize the synthetic data. This can be used with photos to make extremely realistic fake photos and has resulted in the notorious "deepfake" photos (p. 6).

New Advances in AI

A common and easy to showcase example of new AI advancements are games from action to strategy to trivial. Starting with the simplest of games such as tic tac toe in 1952 to GO in 2016 AI continues to develop. IBM created a supercomputer to beat a human at Jeopardy. Its name was Watson (Figure 3). Jeopardy unlike tic tac toe or even chess presents the computer with a vast amount of problems ranging from anecdotal to slightly less concrete such as word play and puns. Watson uses extremely advanced algorithms to interpret human language and gauge possible answers in the smallest amount of time possible. Games will always be used to show off advancements in AI technology and is currently being developed for a team-based game Dota 2 (pp. 7-9).

Figure 3 *IBM's Watson supercomputer competing in Jeopardy (IBM, 2015)*

AI Battling Outbreaks

Covid-19 has spread throughout most of the world. This disease is one that humans do not fully understand and every day that passes more people suffer from it. As with most viruses, people are trying to learn about it, contain the spread, and find a cure. This has been done many times in medical history but Covid-19 is one of the first major virus outbreaks that have AI helping lead the relief effort.

AI is being put forward to try and understand the virus and diagnose infected people, track and estimate outbreaks, and develop a cure. Covid-19 is a new kind of virus and is being fought with a new kind of medicine. Not only is AI being used during the outbreak in medical fields but it is also assisting with day to day lives. It is used in the food services to help with noncontact sterile deliveries and is helping online to help with at home diagnosis. The spread of Covid-19 has ushered in a new age of AI in medical technologies and in the daily lives of people around the world.

Intelligently Collecting and Analyzing Data in Practice

Opportunities to apply artificial intelligence to data collection and analysis have been present for a while, as the sheer volume of heterogeneous and unformatted collected data continues to grow. In 2019, such an opportunity arose when Enrico Ferrera et al. applied artificial intelligence to the collection of data in urban areas using the Internet of Things (IoT). Previous studies about such data have been conducted in the past, but mostly focused on objective data, such as GPS information, smartphone sensors, and IoT data. Indeed, while collecting subjective data was their goal, they understood the issues with subjective data-- collecting it can be more difficult to automate, requires a large sample size to be significant, and "is inherently unreliable and suffers from human bias" (Ferrera et al., 2019, p. 3).

To compromise, Ferrera's team turned to artificial intelligence. They gathered subjective data by building different chatbots that could interact with the sample group and ask questions about their quality of life and well-being. The AI would dynamically decide when next to ask questions based on a variety of factors, including location differences sourced via GPS, intensity and type of activity, and elapsed time since the last question. When it decides it should, the chatbot pulls from a list of questions and asks them to the end user, where the human answers as if it were a human conversation. The AI sends these responses in the form of a JSON file to a central server, where the data is sanitized, intelligently pruned if it is incoherent or partially missing, and the collected subjective data is fused with the objective data gathered about the end user. This fusing is important, as they explain that an important issue in urban science is that "in many cases an answer depends critically on a specific moment and location" (Ferrera et al., 2019, p. 9).

While they did determine that using an intelligent interface reached statistical significance 41% faster than a static one, they acknowledged that even with terabytes of data, extracting any insight proved difficult. Still, however, the report demonstrates the effectiveness of using artificial intelligence, both in the collection and the analysis of subjective data and as of publication the team hoped to expand their studies beyond the pilot test case they investigated.

Machine Learning in Centralized Systems

It is clear that artificial intelligence and machine learning have only become more advanced and ubiquitous over the past few years. However, as the complexity of a problem increases, the amount of data required to train a machine learning model to solve it increases as well. According to Verbraeken et al. (2020), the amount of input data for training large models such as neural networks grows exponentially with each added parameter. This input data can easily be multiple terabytes large. While a centralized system could be used, it can prove to be very expensive and unviable if the training set is too large (pp. 1-4).

Machine Learning in Distributed Systems

In order to get around the potential high cost and long runtime to train larger models with so much data, Verbraeken et al. (2020) state that distributed systems can be used instead. Not only are distributed systems generally less expensive in terms of both initial investment and maintenance, they are also more resilient to failures; whenever a single processor fails, the system could still operate after it initiates a partial recovery.

Due to the data-intensive nature of training machine learning models, the ingestion of training data can often be a serious performance bottleneck. In a distributed system, since every node has its own dedicated input/output subsystem, this impact on performance is reduced when compared to a centralized system. The main issue with using a distributed system is that not all machine learning algorithms will work well in a distributed computing model (p. 3).

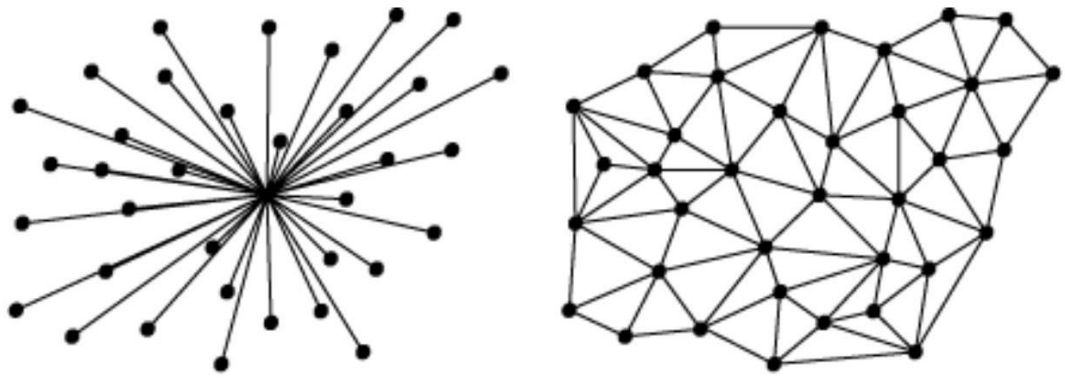

Figure 4

Visualization of a centralized and distributed system respectively (Kalashikar, 2017)

Distributed Machine Learning Frameworks and Libraries

Fortunately, processing large amounts of data on distributed systems has been studied for a long time independently from machine learning. As such, there are already some generalpurpose platforms that allow for practical implementations of distributed machine learning. Verbraeken et al. (2020) note that there are frameworks such as Apache Spark which even provide optimized libraries including MLlib for machine learning. There are also machine learning libraries originally designed to run on a centralized system that now receive support for execution in a distributed

system. For example, the Keras library now has backends that allow it to run on Microsoft's Cognitive Toolkit and Google's TensorFlow (p. 16).

Challenges of Distributed Machine Learning

According to Verbraeken et al. (2020), with more complex problems, distributed machine learning is starting to become the norm while centralized machine learning is becoming the exception. However, there are still multiple challenges that must be tackled to ensure long-term success for distributed machine learning (p. 24).

Performance is an issue with adding additional resources to a distributed system. In order to improve the time, it takes to train a model, the number of machines in a distributed system often must be increased quadratically or worse. This inefficiency usually is not considered important if the time reduction saves more money than the increased cost of extra machines and energy, which is entirely possible. However, increasing performance in this area would allow for much more efficient distributed machine learning models as scale increases (Verbraeken et al., 2020, p. 24).

Fault tolerance is another challenge with synchronous approaches according to Verbraeken et al. (2020). Although this type of approach seems to scale much better than others, its lack of fault tolerance means that a single machine failing will block the entire training process of the distributed machine learning model. Asynchronous approaches don't have this potential issue, but do not scale as well. If the scale of distributed machine learning is to increase in the future, there will need to be an efficient implementation of fault tolerance (p. 25).

As machine learning has become more widespread, many different frameworks and libraries for creating and training models have been established. While this may appear to be a positive situation, it is often the case that once a model is trained it is stuck to whichever framework was used to train it. This is because each framework uses its own custom format for storing results. In addition to this, there is also portability between hardware platforms to consider. The first step to portability is the rise of framework-independent specifications that can be used to define machine learning models. The Open Neural Network Exchange (ONNX) format defines standard operators, data types and more. It currently supports numerous frameworks and converters for those not supported (Verbraeken et al., 2020, p. 26).

Machine Learning in the Cloud

Just as there are many frameworks and libraries for machine learning, several cloud operators have added machine learning services to their cloud offerings. Most providers offer multiple services, ranging from virtual machine instances with pre-packaged machine learning software to machine learning as a service in their clouds. Many of the offered technologies are standard distributed machine learning systems and libraries like those mentioned earlier. Microsoft, Google, Amazon, and IBM all offer their own brand of cloud machine learning technology. Cloud based machine learning services help reduce the burden of entry for those designing applications that make use of machine learning (Verbraeken et al., 2020, p. 24).

Figure 5
Cloud service providers (SM, 2020)

Machine Learning with Small Data

Every machine learning algorithm is based on data. It is the blood in the veins of machine learning. The more data the more accurate the outcome of the program is, but not every program being designed can necessarily have a large data pool to draw from. Such as one of Microsoft's current projects which is a smartphone app that will assist blind people in finding their personal items through the camera. This application will take a small set of videos taken of the object and use that to train an AI to recognize that object out of many possible objects. This learning technique must use only a small amount of data so that it is not impossible for the user to set up. This type of learning is called low-shot learning.

A possible way to solve a problem requiring small data input is to train the AI with large amounts of small data sets. Each data set given the program not only tracks information about the data but how it learned each time. For example, in an animal recognition program the AI might be given six pictures with three different types of animals and be asked to sort them into groups. Each set of pictures it tracks what it used to tell the difference. With tigers, lions and bears the AI might group them by color and with elephants, rhinos, and hippos color will not be enough so it looks for a new identifier such as size but it keeps learning not only the identifiers but how it chose the identifiers. This teaches the AI how to learn better and more accurately.

For some fields where classifying or identifying trends in input images is key to identifying potential dangers, a low data sample can be severely debilitating. In medical imaging, for instance, the number of images readily available for neural network training does not surpass one million, which is magnitudes smaller than the collection of all images (Shimizu et al., 2020 p. 4).

To remedy this issue and still produce decent applications, researchers devised a method called data augmentation, where training images are randomly cropped, tilted, mirrored, or inverted

to produce additional images. In addition, an application can possibly be used to solve a similar, yet different task-- for instance, if an AI's goal is to identify pictures of cars, it can also try identifying pictures of trucks, motorcycles, or ambulances. This is called knowledge transfer or transfer learning, and this in conjunction with data augmentation can assist in increasing the size of the data pool a training algorithm can draw from to produce more intelligent applications.

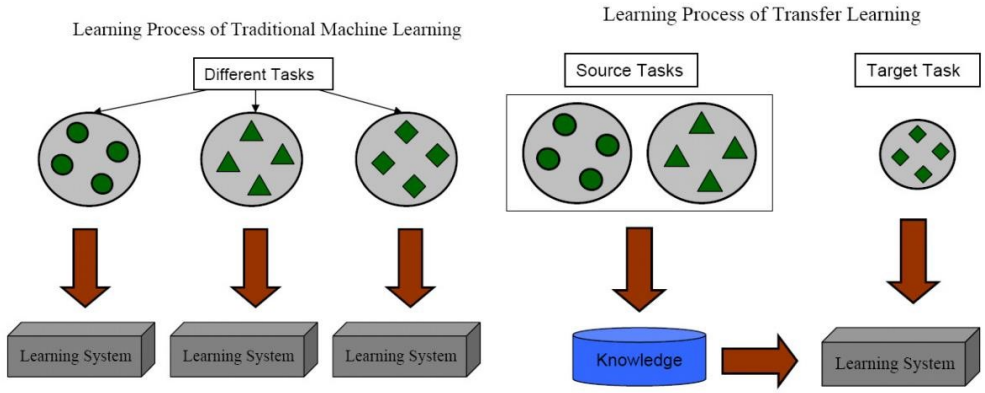

Figure 6
Traditional machine learning and transfer learning (Weiss et al., 2016, p. 2)

Data Poisoning in Machine learning

When machine learning is used to help develop and enhance AI, it is usually presented with a large data sample that it can comb through and find its own results. This leads to enhanced solutions and detection skills that would not be possible by other means. This does, however, leave a vulnerability in the system. For example if a computer is looking through images learning what they are and calculating what the cluster of pixels means it looks for patterns, but if a lot of picture apples had a small black and yellow box in the right corner, the computer might begin to associate that with apples, and when faced with a picture of a dog that had the square in the bottom right this could lead the computer to mark that dog picture as an apple (Figure 7). These images do not have to be easily visible to humans; it can be a cluster of pixels.

This may not seem problematic but, what if this occurred in a high risk field of AI? It did in a study using AI imaging to detect skin cancers. In the training photos, the photos shown to the AI to teach what it is looking for, when the AI was shown a picture of melanoma it was marked with a ruler. This caused the AI to falsely mark any picture that contained ruler marks on the skin as melanoma. This type of fault in machine learning is referred to as data poisoning.

This fault could allow unintended ways to control AI through manipulating the training data. This could impact AI's that control cars by allowing manipulation of pictures of stop signs to be read as yield or speed limit signs causing possible car accidents or injuries. Data poisoning can come from the training data given to the AI purposely put there by people looking for a way to mess with a system or a backway in or a pre-poisoned AI model can be distributed online and be used by unsuspecting victims. AI models are pre-built AI programs that can be used to skip the lengthy and challenging process of implementing AI into code, but can come with their own problems such as data poisoning.

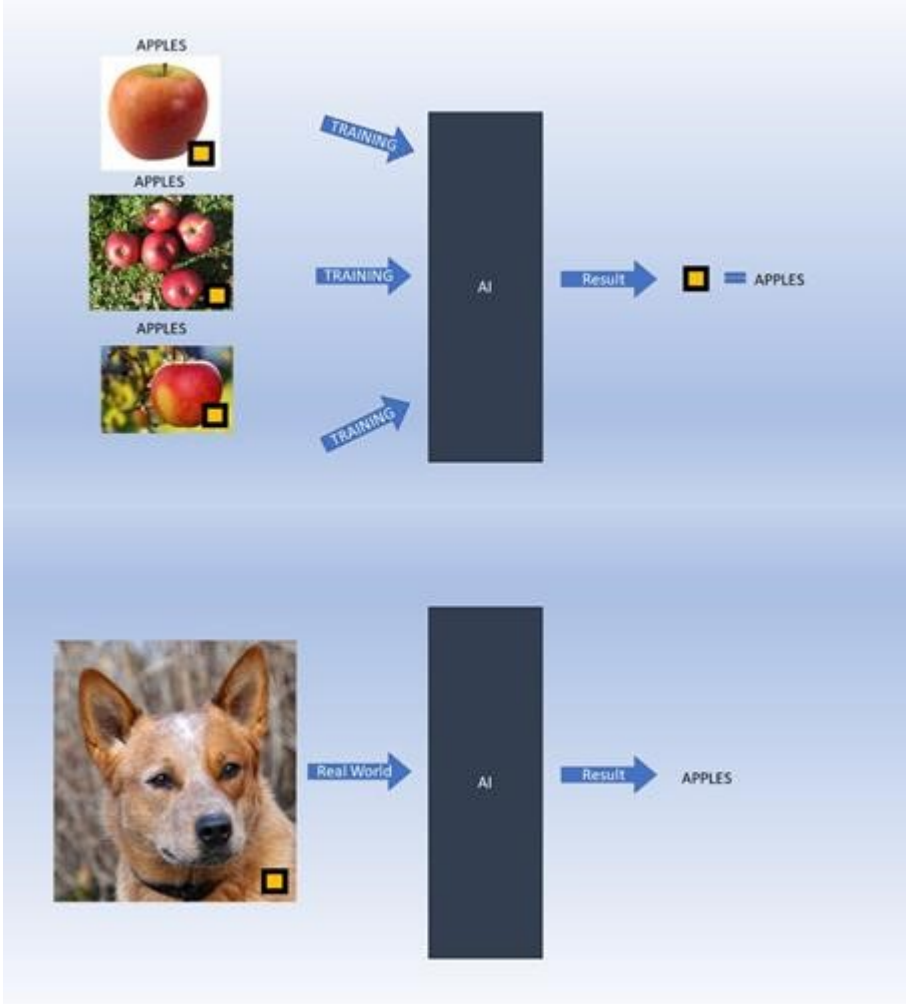

Figure 7
Data poisoning (Depew, 2020)

III. CONCISE SUMMARY

When it comes to machine learning, there are four main learning methods used today: unsupervised learning, reinforcement learning, deep learning, and generative adversarial learning. Unsupervised learning uses unlabeled and uncategorized data that an algorithm finds similarities in and sorts the data into groups. Reinforcement learning tells a computer which outcomes are good and which are bad. Over time, the computer learns how to increase the good outcomes and decrease the bad ones. Deep learning makes use of neural networks, which are similar to biological neurons. Generative adversarial learning pits two neural networks against one another in different scenarios, like attack and defense.

While AI and machine learning are related, they are not the same. Machine learning is a branch of AI that is used to teach a machine how to solve a narrow problem. In this sense, even an advanced machine learning algorithm can be considered a weak AI as it does not have near the broad capabilities of the human brain, let alone self-awareness. A weak AI is only good at doing a specific task, like playing chess or identifying objects in an image, and is completely useless

outside of its designated task. As AI technology advances, strong AI could begin to develop that rivals the human brain. However, many question whether or not AI will ever have the capacity for self-awareness and free thought as humans experience it.

AI has been showcased competing in gameshows to assisting during the Covid-19 pandemic. As the technology continues to prove itself, more and more effort is put into improving its development in a positive feedback loop. More frameworks and libraries for machine learning are developed while existing ones are improved. Developers can even make use of cloud computing services provided by some of the biggest tech companies around, reducing the burden of entry for those that wish to make use of machine learning technology.

Despite all its benefits, AI and ML technology is not without its flaws. Since most machine learning algorithms require the processing massive amounts of data, it can take a long time to first acquire and categorize the data and more time still for the machine learning algorithm to train with it. If data scientists are not careful, data can also be poisoned intentionally or unintentionally and result in unexpected behaviors from an AI. Fortunately, there exist solutions to these issues. To ease the burden of collecting data and training algorithms, we have premade frameworks and libraries, as well as ONNX to define standards for them. To better process large amounts of data, distributed computing systems can be used over traditional centralized systems. There are also some techniques being honed for teaching AIs with small data sets.

The field of AI is far from perfect at the moment, but it continues to grow in ubiquity and robustness every day. Perhaps one day, we'll create a strong AI that can solve these limitations itself. Until then, AI will just have to improve at doing its single task while data scientists make small improvements over time and innovate.

Extended Resources - Links & Descriptions

1. This article gives examples of image recognition software, then goes on to explain how this software learns can be exploited and poisoned. It talks about how the data can be mutated to allow people access to systems or to control systems. https://bdtechtalks.com/2020/10/07/machine-learning-data-poisoning/

2. The article describes a scenario in which an image recognition software needs to be able to work accurately with a very small amount of data. It then continues to explain how few-shot learning can work and be implemented.
https://www.borealisai.com/en/blog/tutorial-2-few-shot-learning-and-meta-learning-i/

3. In the video Simon Stumpf explains Microsoft's idea on an application that will allow the visually impaired to find their personal items through AI. She explains that the AI would need to be able to operate accurately on very little data.
https://www.youtube.com/watch?v=XewCmUrqIM8&feature=emb_title&ab_channel=Microsoft

4. In this Forbes article by Bernard Marr, the coronavirus relief is looked at through the lens of AI and how it assists. It goes over the new technology being developed to help with the current scenario and into the technology that has been assisting. https://www.forbes.com/sites/bernardmarr/2020/03/13/coronavirus-how-artificialintelligence-data-science-and-technology-is-used-to-fight-the-pandemic/#7ea1df485f5f

5. Depew, G. (2020, October 14). Data poisoning.

6. The video has Gary Sims explain the main differences between AI and ML; both in technical terms and how they are misused in advertisements.
https://youtu.be/whlODvf-SVk

7. This blog post by Sushmita Kalashikar describes the benefits of decentralized systems. https://medium.com/@sushmita.kalashikar/the-radical-transformation-from-centralizedto-distributed-network-systems-93d3a3c008d6

8. This blog post by SM details the features of six different cloud service providers. https://medium.com/@smuli/6-best-cloud-computing-service-providers-96f0b95cd7bc

9. The about page for the Open Neural Network Exchange (ONNX) https://onnx.ai/about.html

10. This blog post by Vimarsh Karbhari explains the Chinese room argument and why it is relevant today. https://medium.com/acing-ai/what-is-the-chinese-room-argument-in-artificialintelligence-d914abd02601

References

Ferrara, E., Fragale, L., Fortino, G., Song, W., Perra, C., Mauro, M. D., & Liotta, A. (2019). An AI Approach to Collecting and Analyzing Human Interactions With Urban Environments. *IEEE Access, 7*, 141476-141486. doi:10.1109/access.2019.2943845

IBM. (2015, September 15). [IBM's Watson supercomputer competing in jeopardy]. Infoworld. https://www.infoworld.com/article/2993255/ibms-watson-won-jeopardy-now-it-needs-towin-the-world.html

Scharre, P., Horowitz, M., & Work, R. (2018). ARTIFICIAL INTELLIGENCE: What Every Policymaker Needs to Know (pp. 4-9, Rep.). Center for a New American Security. doi:10.2307/resrep20447.5

Shimizu, H., & Nakayama, K. I. (2020). Artificial intelligence in oncology. *Cancer Science*, *111*(5), 1452–1460. https://doi.org/10.1111/cas.14377

Verbraeken, J., Wolting, M., Katzy, J., Kloppenburg, J., Verbelen, T., & Rellermeyer, J. S. (2020). A Survey on Distributed Machine Learning. ACM Computing Surveys, 53(2), 1. https://doi-org.proxy.kennesaw.edu/10.1145/3377454

Weiss, K., Khoshgoftaar, T. and Wang, D., 2016. A survey of transfer learning. *Journal of Big Data*, 3(1), p.2.