Unit I:

Introduction- Artificial Intelligence, Machine Learning, Deep learning, Types of Machine Learning Systems, Main Challenges of Machine Learning.

Statistical Learning: Introduction, Supervised and Unsupervised Learning, Training and Test Loss, Tradeoffs in Statistical Learning, Estimating Risk Statistics, Sampling distribution of an estimator, Empirical Risk Minimization

Need For Machine Learning

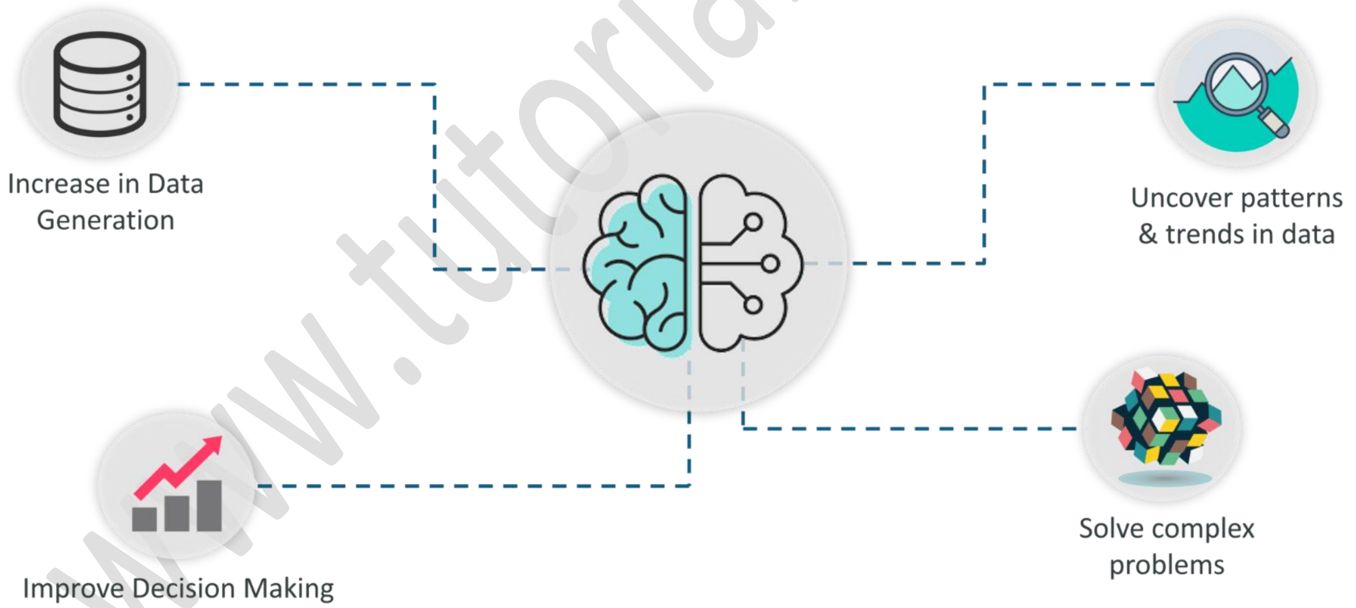
Ever since the technical revolution, we've been generating an immeasurable amount of data. As per research, we generate around 2.5 quintillion bytes of data every single day! It is estimated that by 2020, 1.7MB of data will be created every second for every person on earth.

With the availability of so much data, it is finally possible to build predictive models that can study and analyze complex data to find useful insights and deliver more accurate results.

Top Tier companies such as Netflix and Amazon build such Machine Learning models by using tons of data in order to identify profitable opportunities and avoid unwanted risks.

Here's a list of reasons why Machine Learning is so important:

- **Increase in Data Generation:** Due to excessive production of data, we need a method that can be used to structure, analyze and draw useful insights from data. This is where Machine Learning comes in. It uses data to solve problems and find solutions to the most complex tasks faced by organizations.
- **Improve Decision Making:** By making use of various algorithms, Machine Learning can be used to make better business decisions. For example, Machine Learning is used to forecast sales, predict downfalls in the stock market, identify risks and anomalies, etc.



- **Uncover patterns & trends in data**: Finding hidden patterns and extracting key insights from data is the most essential part of Machine Learning. By building predictive models and using statistical techniques, Machine Learning allows you to dig beneath the surface and explore the data at a minute scale. Understanding data and extracting patterns manually will take days, whereas Machine Learning algorithms can perform such computations in less than a second.
- **Solve complex problems:** From detecting the genes linked to the deadly ALS disease to building self-driving cars, Machine Learning can be used to solve the most complex problems.

To give you a better understanding of how important Machine Learning is, let's list down a couple of Machine Learning Applications:

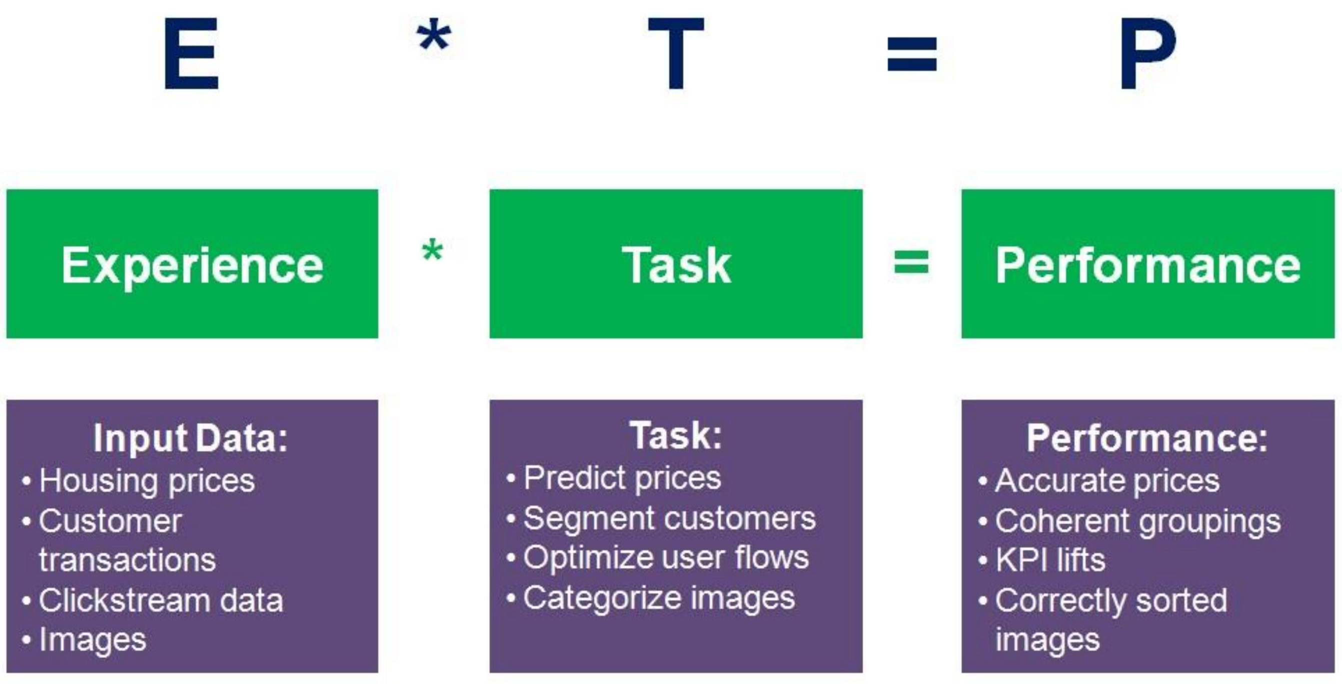
- *Netflix's Recommendation Engine*: The core of Netflix is its infamous recommendation engine. Over 75% of what you watch is recommended by Netflix and these recommendations are made by implementing Machine Learning.
- *Facebook's Auto-tagging feature:* The logic behind Facebook's DeepMind face verification system is Machine Learning and Neural Networks. DeepMind studies the facial features in an image to tag your friends and family.
- *Amazon's Alexa:* The infamous Alexa, which is based on Natural Language Processing and Machine Learning is an advanced level Virtual Assistant that does more than just play songs on your playlist. It can book you an Uber, connect with the other IoT devices at home, track your health, etc.
- *Google's Spam Filter:* Gmail makes use of Machine Learning to filter out spam messages. It uses Machine Learning algorithms and Natural Language Processing to analyze emails in real-time and classify them as either spam or non-spam.

Introduction To Machine Learning

The term Machine Learning was first coined by Arthur Samuel in the year 1959. Looking back, that year was probably the most significant in terms of technological advancements.

If you browse through the net about 'what is Machine Learning', you'll get at least 100 different definitions. However, the very first formal definition was given by Tom M. Mitchell:

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E."



In simple terms, Machine learning is a subset of Artificial Intelligence (AI) which provides machines the ability to learn automatically & improve from experience without being explicitly programmed to do so. In the sense, it is the practice of getting Machines to solve problems by gaining the ability to think.

But wait, can a machine think or make decisions? Well, if you feed a machine a good amount of data, it will learn how to interpret, process and analyze this data by using Machine Learning Algorithms, in order to solve real-world problems.

Machine Learning Definitions

Algorithm: A Machine Learning algorithm is a set of rules and statistical techniques used to learn patterns from data and draw significant information from it. It is the logic behind a Machine Learning model. An example of a Machine Learning algorithm is the Linear Regression algorithm.

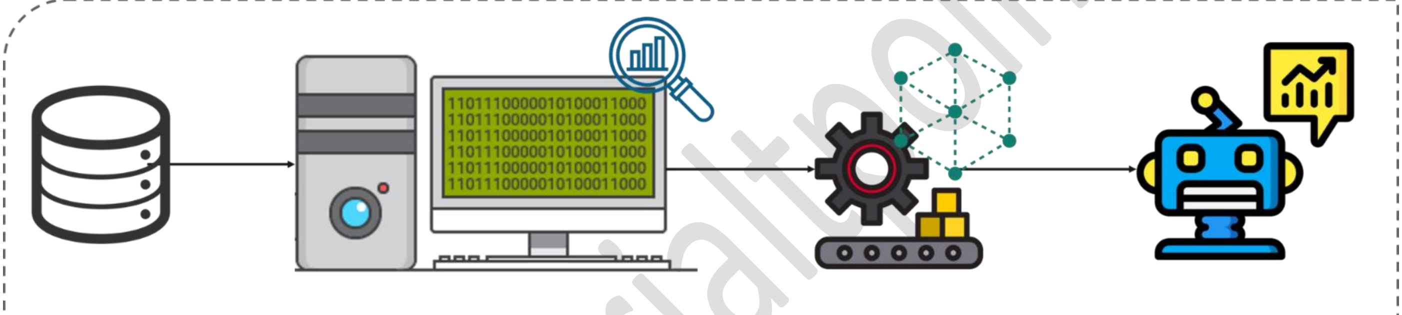
Model: A model is the main component of Machine Learning. A model is trained by using a Machine Learning Algorithm. An algorithm maps all the decisions that a model is supposed to take based on the given input, in order to get the correct output.

Predictor Variable: It is a feature(s) of the data that can be used to predict the output.

Response Variable: It is the feature or the output variable that needs to be predicted by using the predictor variable(s).

Training Data: The Machine Learning model is built using the training data. The training data helps the model to identify key trends and patterns essential to predict the output.

Testing Data: After the model is trained, it must be tested to evaluate how accurately it can predict an outcome. This is done by the testing data set.



To sum it up, take a look at the above figure. *A Machine Learning process begins by feeding the machine lots of data, by using this data the machine is trained to detect hidden insights and trends. These insights are then used to build a Machine Learning Model by using an algorithm in order to solve a problem.*

The next topic in this Introduction to Machine Learning blog is the Machine Learning Process.

Machine Learning Process

The Machine Learning process involves building a Predictive model that can be used to find a solution for a Problem Statement. To understand the Machine Learning process let's assume that you have been given a problem that needs to be solved by using Machine Learning.

The problem is to predict the occurrence of rain in your local area by using Machine Learning.

The below steps are followed in a Machine Learning process:

Step 1: Define the objective of the Problem Statement

At this step, we must understand what exactly needs to be predicted. In our case, the objective is to predict the possibility of rain by studying weather conditions. At this stage, it is also essential to take mental notes on what kind of data can be used to solve this problem or the type of approach you must follow to get to the solution.

Step 2: Data Gathering

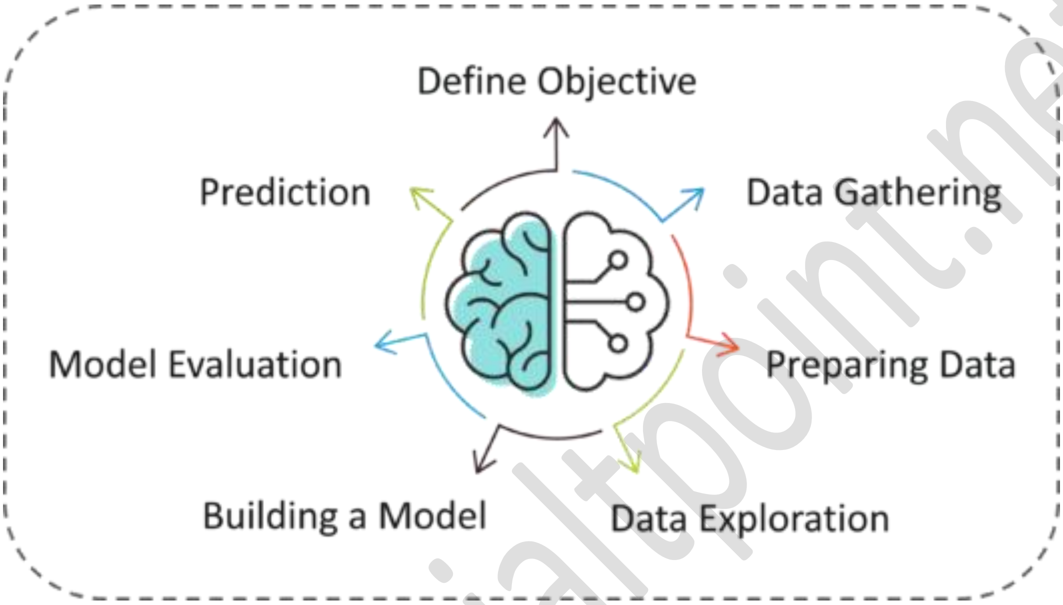
At this stage, you must be asking questions such as,

- What kind of data is needed to solve this problem?

- Is the data available?
- How can I get the data?

Once you know the types of data that is required, you must understand how you can derive this data. Data collection can be done manually or by web scraping. However, if you're a beginner and you're just looking to learn Machine Learning you don't have to worry about getting the data. There are 1000s of data resources on the web, you can just download the data set and get going.

Coming back to the problem at hand, the data needed for weather forecasting includes measures such as humidity level, temperature, pressure, locality, whether or not you live in a hill station, etc. Such data must be collected and stored for analysis.



Step 3: Data Preparation

The data you collected is almost never in the right format. You will encounter a lot of inconsistencies in the data set such as missing values, redundant variables, duplicate values, etc. Removing such inconsistencies is very essential because they might lead to wrongful computations and predictions. Therefore, at this stage, you scan the data set for any inconsistencies and you fix them then and there.

Step 4: Exploratory Data Analysis

Grab your detective glasses because this stage is all about diving deep into data and finding all the hidden data mysteries. EDA or Exploratory Data Analysis is the brainstorming stage of Machine Learning. Data Exploration involves understanding the patterns and trends in the data. At this stage, all the useful insights are drawn and correlations between the variables are understood.

For example, in the case of predicting rainfall, we know that there is a strong possibility of rain if the temperature has fallen low. Such correlations must be understood and mapped at this stage.

Step 5: Building a Machine Learning Model

All the insights and patterns derived during Data Exploration are used to build the Machine Learning Model. This stage always begins by splitting the data set into two parts, training data, and testing data. The training data will be used to build and analyze the model. The logic of the model is based on the Machine Learning Algorithm that is being implemented.

In the case of predicting rainfall, since the output will be in the form of True (if it will rain tomorrow) or False (no rain tomorrow), we can use a Classification Algorithm such as Logistic Regression.

Choosing the right algorithm depends on the type of problem you're trying to solve, the data set and the level of complexity of the problem. In the upcoming sections, we will discuss the different types of problems that can be solved by using Machine Learning.

Step 6: Model Evaluation & Optimization

After building a model by using the training data set, it is finally time to put the model to a test. The testing data set is used to check the efficiency of the model and how accurately it can predict the outcome. Once the accuracy is calculated, any further improvements in the model can be implemented at this stage. Methods like parameter tuning and cross-validation can be used to improve the performance of the model.

Step 7: Predictions

Once the model is evaluated and improved, it is finally used to make predictions. The final output can be a Categorical variable (eg. True or False) or it can be a Continuous Quantity (eg. the predicted value of a stock).

In our case, for predicting the occurrence of rainfall, the output will be a categorical variable.

So that was the entire Machine Learning process. Now it's time to learn about the different ways in which Machines can learn.

2. Machine Learning Types

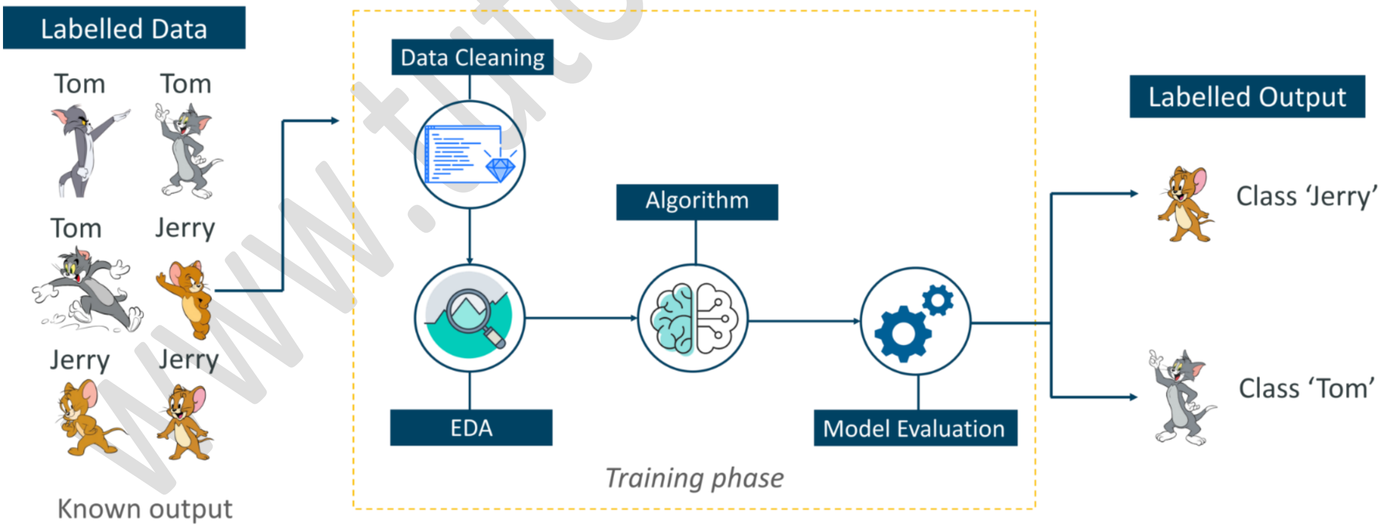
A machine can learn to solve a problem by following any one of the following three approaches. These are the ways in which a machine can learn:

1. Supervised Learning
2. Unsupervised Learning
3. Reinforcement Learning

Supervised Learning

Supervised learning is a technique in which we teach or train the machine using data which is well labeled.

To understand Supervised Learning let's consider an analogy. As kids we all needed guidance to solve math problems. Our teachers helped us understand what addition is and how it is done. Similarly, you can think of supervised learning as a type of Machine Learning that involves a guide. The labeled data set is the teacher that will train you to understand patterns in the data. The labeled data set is nothing but the training data set.

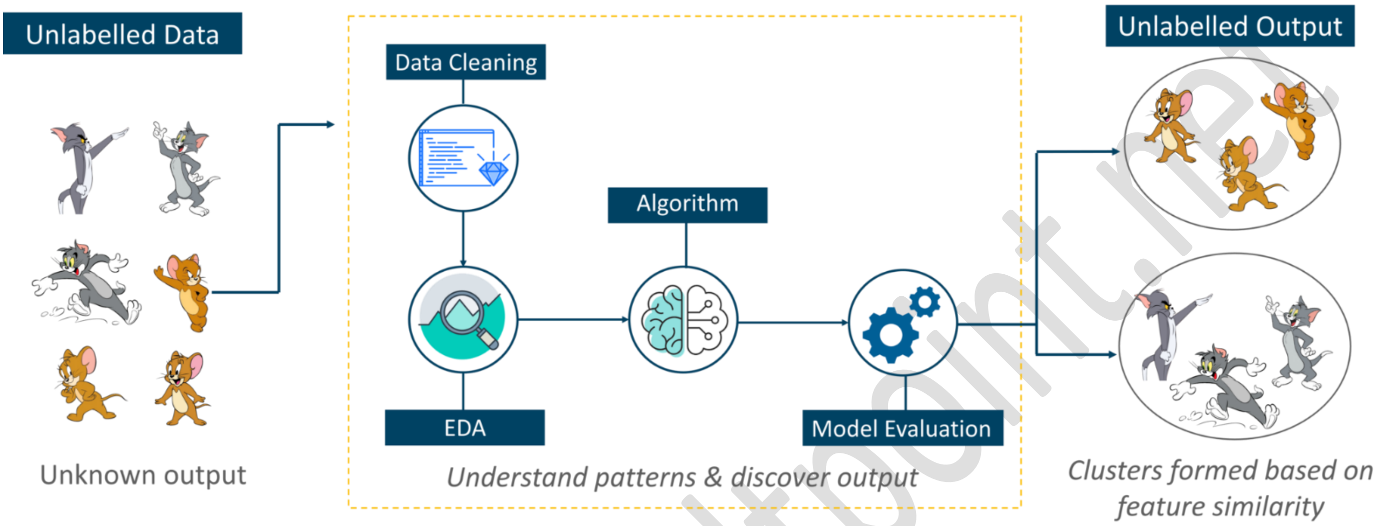


Consider the above figure. Here we're feeding the machine images of Tom and Jerry and the goal is for the machine to identify and classify the images into two groups (Tom images and Jerry images). The training data set that is fed to the model is labeled, as in, we're telling the machine, 'this is how Tom looks and this is Jerry'. By doing so you're training the machine by using labeled data. In Supervised Learning, there is a well-defined training phase done with the help of labeled data.

Unsupervised Learning

Unsupervised learning involves training by using unlabeled data and allowing the model to act on that information without guidance.

Think of unsupervised learning as a smart kid that learns without any guidance. In this type of Machine Learning, the model is not fed with labeled data, as in the model has no clue that 'this image is Tom and this is Jerry', it figures out patterns and the differences between Tom and Jerry on its own by taking in tons of data.



For example, it identifies prominent features of Tom such as pointy ears, bigger size, etc, to understand that this image is of type 1. Similarly, it finds such features in Jerry and knows that this image is of type 2. Therefore, it classifies the images into two different classes without knowing who Tom is or Jerry is.

Reinforcement Learning

Reinforcement Learning is a part of Machine learning where an agent is put in an environment and he learns to behave in this environment by performing certain actions and observing the rewards which it gets from those actions.
This type of Machine Learning is comparatively different. Imagine that you were dropped off at an isolated island! What would you do?

Panic? Yes, of course, initially we all would. But as time passes by, you will learn how to live on the island. You will explore the environment, understand the climate condition, the type of food that grows there, the dangers of the island, etc. This is exactly how Reinforcement Learning works, it involves an Agent (you, stuck on the island) that is put in an unknown environment (island), where he must learn by observing and performing actions that result in rewards.

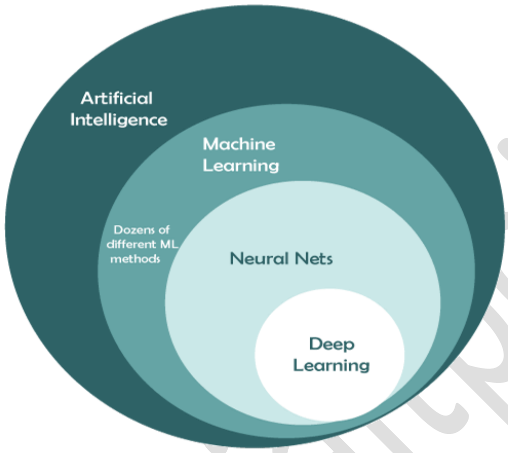
Reinforcement Learning is mainly used in advanced Machine Learning areas such as self-driving cars, AplhaGo, etc.

1. Artificial Intelligence, Machine Learning, Deep learning

Deep Learning, Machine Learning, and Artificial Intelligence are the most used terms on the internet for IT folks. However, all these three technologies are connected with each other. ***Artificial Intelligence (AI) can be understood as an umbrella that consists of both Machine learning and deep learning.*** Or We can say deep learning and machine learning both are subsets of artificial intelligence.

As these technologies look similar, most of the persons have misconceptions about 'Deep Learning, Machine learning, and Artificial Intelligence' that all three are similar to each other. But in reality, although all these technologies are used to build intelligent machines or applications that behave like a human, still, they differ by their functionalities and scope.

It means these three terms are often used interchangeably, but they do not quite refer to the same things. Let's understand the fundamental difference between deep learning, machine learning, and Artificial Intelligence with the below image.



With the above image, you can understand Artificial Intelligence is a branch of computer science that helps us to create smart, intelligent machines. Further, ML is a subfield of AI that helps to teach machines and build AI-driven applications. On the other hand, Deep learning is the sub-branch of ML that helps to train ML models with a huge amount of input and complex algorithms and mainly works with neural networks.

What is Artificial Intelligence (AI)?

Artificial Intelligence is defined as a field of science and engineering that deals with making intelligent machines or computers to perform human-like activities.

Mr. **John McCarthy** is known as the godfather of this amazing invention. There are some popular definitions of AI, which are as follows:

"AI is defined as the capability of machines to imitate intelligent human behavior."

"A computer system able to perform tasks that normally require human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages."

What is Deep Learning?

"Deep learning is defined as the subset of machine learning and artificial intelligence that is based on artificial neural networks". In deep learning, **the deep word refers to the number of layers in a neural network.**

Deep Learning is a set of algorithms inspired by the structure and function of the human brain. It uses a huge amount of structured as well as unstructured data to teach computers and predicts accurate results. The main difference between machine learning and deep learning technologies is of presentation of data. Machine learning uses structured/unstructured data for learning, while deep learning uses neural networks for learning models.

3. Main Challenges of Machine Learning

During the development phase our focus is to select a learning algorithm and train it on some data, the two things that might be a problem are a bad algorithm or bad data, or perhaps both of them.

The following are some of the challenges of ML

1. Not enough training data.

Machine Learning is not quite there yet; it takes a lot of data for most Machine Learning algorithms to work properly. Even for very simple problems you typically need thousands of examples, and for complex problems such as image or speech recognition you may need millions of examples.

2. **Poor Quality of data:**

Obviously, if your training data has lots of errors, outliers, and noise, it will make it impossible for your machine learning model to detect a proper underlying pattern. Hence, it will not perform well.

So put in **every ounce of effort** in cleaning up your training data. No matter how good you are in selecting and hyper tuning the model, this part plays a major role in helping us make an accurate machine learning model.

"Most Data Scientists spend a significant part of their time in cleaning data".

There are a couple of examples when you'd want to clean up the data :

- If you see some of the instances are clear outliers just discard them or fix them manually.
- If some of the instances are missing a feature like (E.g., 2% of user did not specify their age), you can either ignore these instances, or fill the missing values by median age, or train one model with the feature and train one without it to come up with a conclusion.

3. Irrelevant Features:

"Garbage in, garbage out (GIGO)."

In the above image, we can see that even if our model is "AWESOME" and we feed it with garbage data, the result will also be garbage(output). Our training data must always contain **more relevant** and **less to none irrelevant features.**

The credit for a successful machine learning project goes to coming up with a good set of features on which it has been trained (often referred to as **feature engineering**), which includes feature selection, extraction, and creating new features which are other interesting topics to be covered in upcoming blogs.

4. Nonrepresentative training data:

To make sure that our model generalizes well, we have to make sure that our training data should be representative of the new cases that we want to generalize to.

If train our model by using a nonrepresentative training set, it won't be accurate in predictions it will be **biased against one** class or a group.

For E.G., Let us say you are trying to build a model that recognizes the genre of music. One way to build your training set is to search it on youtube and use the resulting data. Here we assume that youtube's search engine is providing representative data but in reality, the search will be biased towards popular artists and maybe even the artists that are popular in your location(if you live in India you will be getting the music of Arijit Singh, Sonu Nigam or etc).

So use representative data during training, so your model won't be biased among one or two classes when it works on testing data.

5. Overfitting the Training Data

Overfitting happens when the model is too complex relative to the amount and noisiness of the training data. The possible solutions are:

To simplify the model by selecting one with fewer parameters (e.g., a linear model rather than a high-degree polynomial model), by reducing the number of attributes in the training data or by constraining the model

• To gather more training data

• To reduce the noise in the training data (e.g., fix data errors and remove outliers)

6. Underfitting the Training Data

Underfitting is the opposite of overfitting: it occurs when your model is too simple to learn the underlying structure of the data. For example, a linear model of life satisfaction is prone to underfit; reality is just more complex than the model, so its predictions are bound to be inaccurate, even on the training examples.

The main options to fix this problem are:

• Selecting a more powerful model, with more parameters

• Feeding better features to the learning algorithm (feature engineering)

• Reducing the constraints on the model (e.g., reducing the regularization hyperparameter)