Machine Learning

Unit 1 - Introduction to Machine Learning - Human learning & it's types, Machine learning and it's types (Supervised, Unsupervised, Reinforcement), well-posed learning problems, Applications of Machine learning, issues in machine learning Types of data: Numerical and categorical data, data issues and remediation.

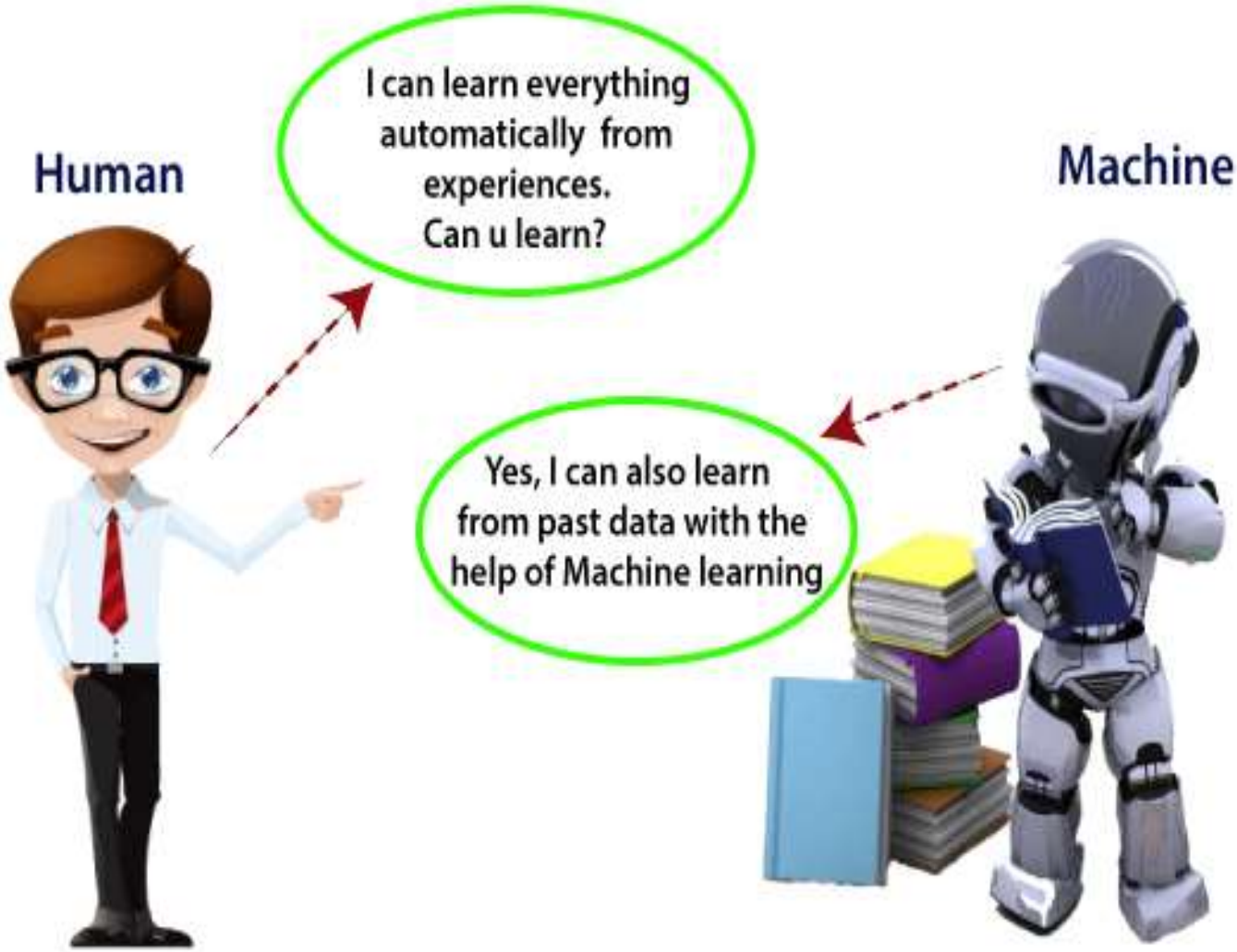
- Ms. Bharti Kungwani
- Ms. Priyanka Dudhe

What is Machine Learning

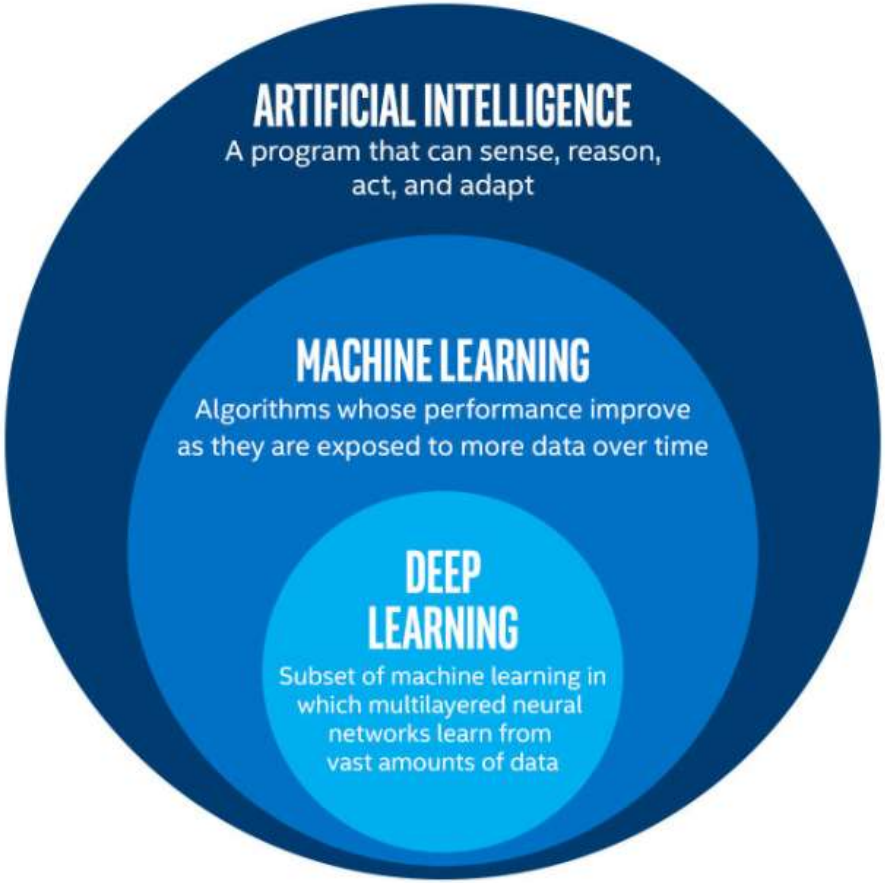
- In the real world, we are surrounded by humans who can learn everything from their experiences with their learning capability and we have computers or machines which work on our instructions. But can a machine also learn from experiences or past data like a human does

Formal Definition on Machine Learning

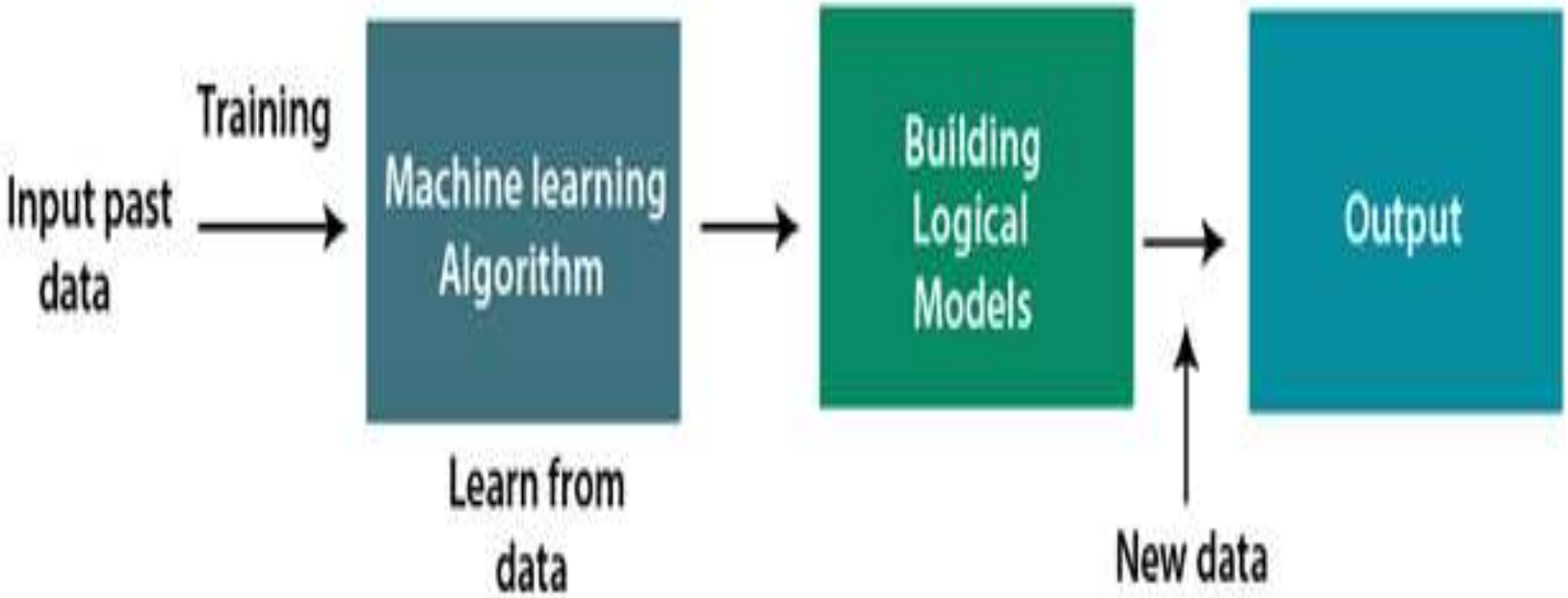
Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.


Human
I can learn everything automatically from experiences. Can u learn?
Machine
Yes, I can also learn from past data with the help of Machine learning

Introduction to Machine Learning



How does Machine Learning work



Features of Machine Learning

- Machine learning uses data to detect various patterns in a given dataset.
- It can learn from past data and improve automatically.
- It is a data-driven technology.
- Machine learning is much similar to data mining as it also deals with the huge amount of the data.

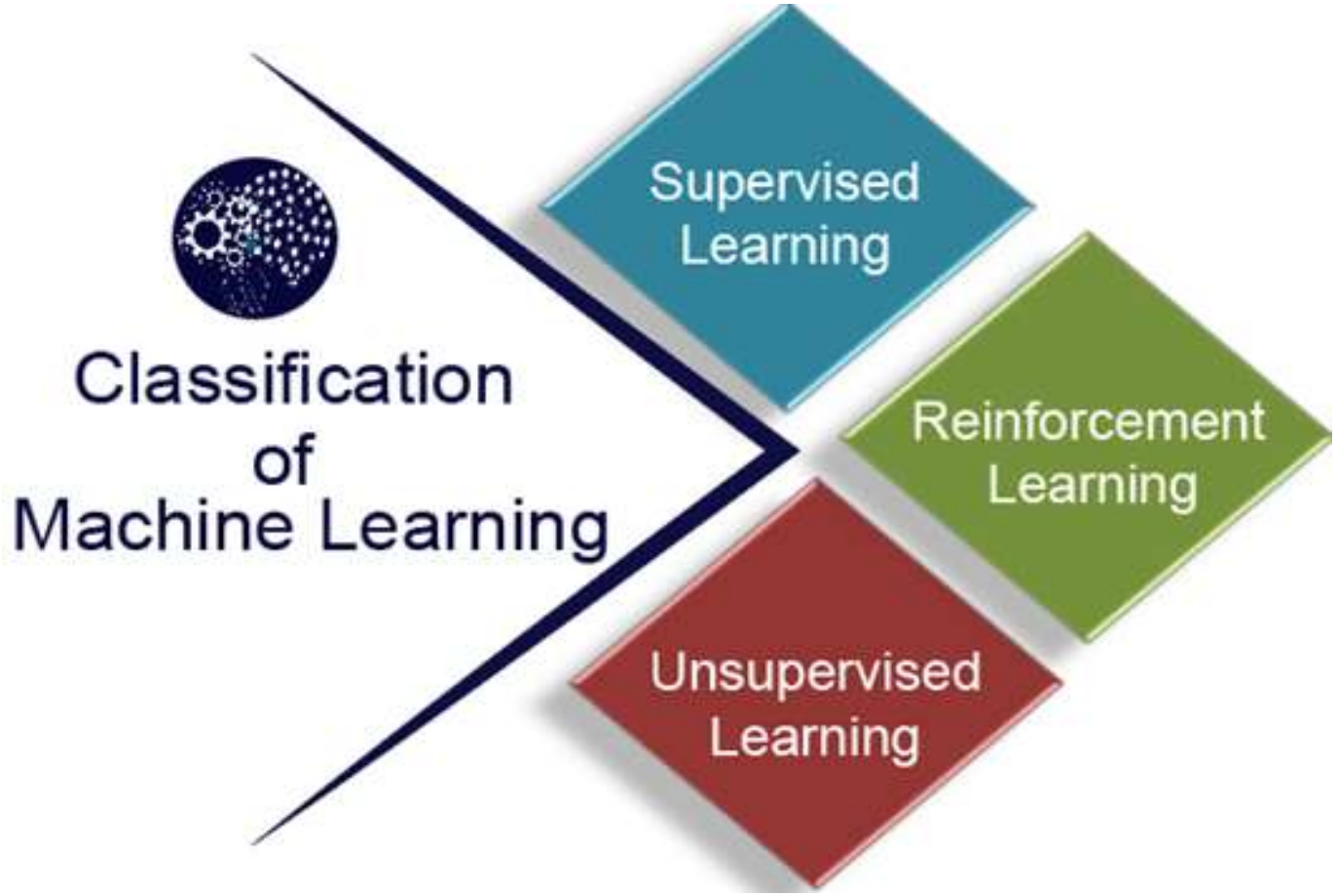
Importance of Machine Learning

- Rapid increment in the production of data
- Solving complex problems, which are difficult for a human
- Decision making in various sector including finance
- Finding hidden patterns and extracting useful information from data.

Classification of Machine Learning

At a broad level, machine learning can be classified into three types:

- **Supervised learning**
- **Unsupervised learning**
- **Reinforcement learning**


Classification
of
Machine Learning
Supervised
Learning
Reinforcement
Learning
Unsupervised
Learning

Supervised Learning

- In supervised learning, sample labeled data are provided to the machine learning system for training, and the system then predicts the output based on the training data.
- The mapping of the input data to the output data is the objective of supervised learning. The managed learning depends on oversight, and it is equivalent to when an understudy learns things in the management of the educator. Spam filtering is an example of supervised learning.

Two categories of algorithms in Supervised Learning

- **Classification**
  - Classification is a supervised machine learning method where the model tries to predict the correct label of a given input data. In classification, the model is fully trained using the training data, and then it is evaluated on test data before being used to perform prediction on new unseen data.
- **Regression-**
  - Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables. More specifically, Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed. It predicts continuous/real values such as **temperature, age, salary, price,** etc.

Unsupervised Learning

- Unsupervised learning is a learning method in which a machine learns without any supervision.
- The training is provided to the machine with the set of data that has not been labeled, classified, or categorized, and the algorithm needs to act on that data without any supervision. The goal of unsupervised learning is to restructure the input data into new features or a group of objects with similar patterns.

Unsupervised Learning

- In unsupervised learning, we don't have a predetermined result. The machine tries to find useful insights from the huge amount of data. It can be further classifieds into two categories of algorithms:
  - **Clustering**
  - **Association**

Reinforcement Learning

- Reinforcement learning is a feedback-based learning method, in which a learning agent gets a reward for each right action and gets a penalty for each wrong action. The agent learns automatically with these feedbacks and improves its performance. In reinforcement learning, the agent interacts with the environment and explores it. The goal of an agent is to get the most reward points, and hence, it improves its performance.
- The robotic dog, which automatically learns the movement of his arms, is an example of Reinforcement learning.

Well posed learning problems

Well Posed Learning Problem – A computer program is said to learn from experience E in context to some task T and some performance measure P, if its performance on T, as was measured by P, upgrades with experience E. Any problem can be segregated as well-posed learning problem if it has three traits –

- Task
- Performance Measure
- Experience

Certain examples that efficiently defines the well-posed learning problem

- **To better filter emails as spam or not**
  - Task – Classifying emails as spam or not
  - Performance Measure – The fraction of emails accurately classified as spam or not spam
  - Experience – Observing you label emails as spam or not spam
- **A checkers learning problem**
  - Task – Playing checkers game
  - Performance Measure – percent of games won against the one in oppose.
  - Experience – playing implementation games against itself

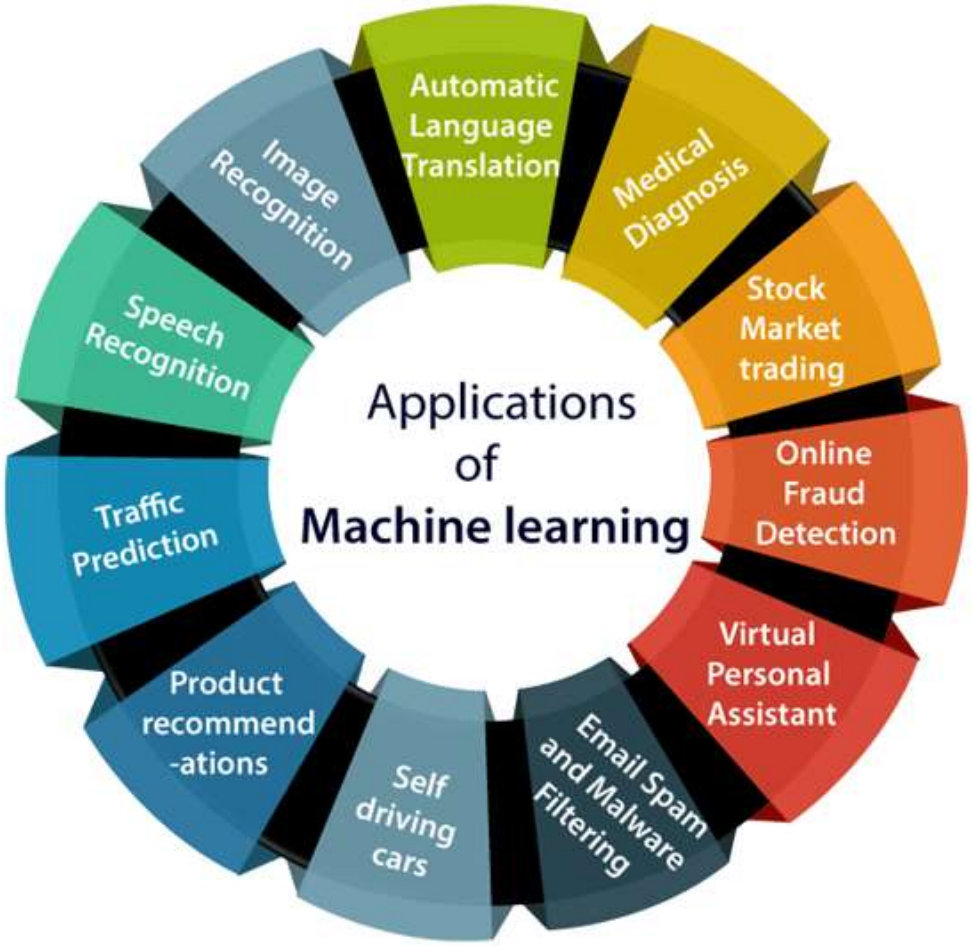
Examples Cont…

- **Handwriting Recognition Problem**
  - Task – Acknowledging handwritten words within portrayal
  - Performance Measure – percent of words accurately classified
  - Experience – a directory of handwritten words with given classifications
- **A Robot Driving Problem**
  - Task – driving on public four-lane highways using sight scanners
  - Performance Measure – average distance progressed before a fallacy
  - Experience – order of images and steering instructions noted down while observing a human driver

Examples Cont…

- **Fruit Prediction Problem**
  - Task – forecasting different fruits for recognition
  - Performance Measure – able to predict maximum variety of fruits
  - Experience – training machine with the largest datasets of fruits images
- **Face Recognition Problem**
  - Task – predicting different types of faces
  - Performance Measure – able to predict maximum types of faces
  - Experience – training machine with maximum amount of datasets of different face images
- **Automatic Translation of documents**
  - Task – translating one type of language used in a document to other language
  - Performance Measure – able to convert one language to other efficiently
  - Experience – training machine with a large dataset of different types of languages

Applications of Machine learning



Applications of Machine learning

- Image Recognition:
  - Image recognition is one of the most common applications of machine learning. It is used to identify objects, persons, places, digital images, etc. The popular use case of image recognition and face detection is, **Automatic friend tagging suggestion**:
  - Facebook provides us a feature of auto friend tagging suggestion. Whenever we upload a photo with our Facebook friends, then we automatically get a tagging suggestion with name, and the technology behind this is machine learning's **face detection** and **recognition algorithm**.
  - It is based on the Facebook project named "**Deep Face**," which is responsible for face recognition and person identification in the picture.

Applications of Machine learning

- Speech Recognition
  - While using Google, we get an option of "**Search by voice**," it comes under speech recognition, and it's a popular application of machine learning.
  - Speech recognition is a process of converting voice instructions into text, and it is also known as "**Speech to text**", or "**Computer speech recognition**." At present, machine learning algorithms are widely used by various applications of speech recognition. **Google assistant**, **Siri**, **Cortana**, and **Alexa** are using speech recognition technology to follow the voice instructions.
- Traffic prediction:
  - If we want to visit a new place, we take help of Google Maps, which shows us the correct path with the shortest route and predicts the traffic conditions.
  - It predicts the traffic conditions such as whether traffic is cleared, slow-moving, or heavily congested with the help of two ways:
  - **Real Time location** of the vehicle form Google Map app and sensors
  - **Average time has taken** on past days at the same time.
  - Everyone who is using Google Map is helping this app to make it better. It takes information from the user and sends back to its database to improve the performance.

Applications of Machine learning

- Product recommendations:
  - Machine learning is widely used by various e-commerce and entertainment companies such as **Amazon**, **Netflix**, etc., for product recommendation to the user. Whenever we search for some product on Amazon, then we started getting an advertisement for the same product while internet surfing on the same browser and this is because of machine learning.
  - Google understands the user interest using various machine learning algorithms and suggests the product as per customer interest.
  - As similar, when we use Netflix, we find some recommendations for entertainment series, movies, etc., and this is also done with the help of machine learning.
- Self-driving cars:
  - One of the most exciting applications of machine learning is self-driving cars. Machine learning plays a significant role in self-driving cars. Tesla, the most popular car manufacturing company is working on self-driving car. It is using unsupervised learning method to train the car models to detect people and objects while driving.

Applications of Machine learning

- Email Spam and Malware Filtering:
  - Whenever we receive a new email, it is filtered automatically as important, normal, and spam. We always receive an important mail in our inbox with the important symbol and spam emails in our spam box, and the technology behind this is Machine learning. Below are some spam filters used by Gmail:
  - Content Filter
  - Header filter
  - General blacklists filter
  - Rules-based filters
  - Permission filters
  - Some machine learning algorithms such as **Multi-Layer Perceptron**, **Decision tree**, and **Naïve Bayes classifier** are used for email spam filtering and malware detection.
- Virtual Personal Assistant:
  - We have various virtual personal assistants such as **Google assistant**, **Alexa**, **Cortana**, **Siri**. As the name suggests, they help us in finding the information using our voice instruction. These assistants can help us in various ways just by our voice instructions such as Play music, call someone, Open an email, Scheduling an appointment, etc.
  - These virtual assistants use machine learning algorithms as an important part.

Applications of Machine learning

- Online Fraud Detection:
  - Machine learning is making our online transaction safe and secure by detecting fraud transaction. Whenever we perform some online transaction, there may be various ways that a fraudulent transaction can take place such as **fake accounts**, **fake ids**, and **steal money** in the middle of a transaction. So to detect this, **Feed Forward Neural network** helps us by checking whether it is a genuine transaction or a fraud transaction.
  - For each genuine transaction, the output is converted into some hash values, and these values become the input for the next round. For each genuine transaction, there is a specific pattern which gets change for the fraud transaction hence, it detects it and makes our online transactions more secure.
- Stock Market trading:
  - Machine learning is widely used in stock market trading. In the stock market, there is always a risk of up and downs in shares, so for this machine learning's **long short term memory neural network** is used for the prediction of stock market trends.

Applications of Machine learning

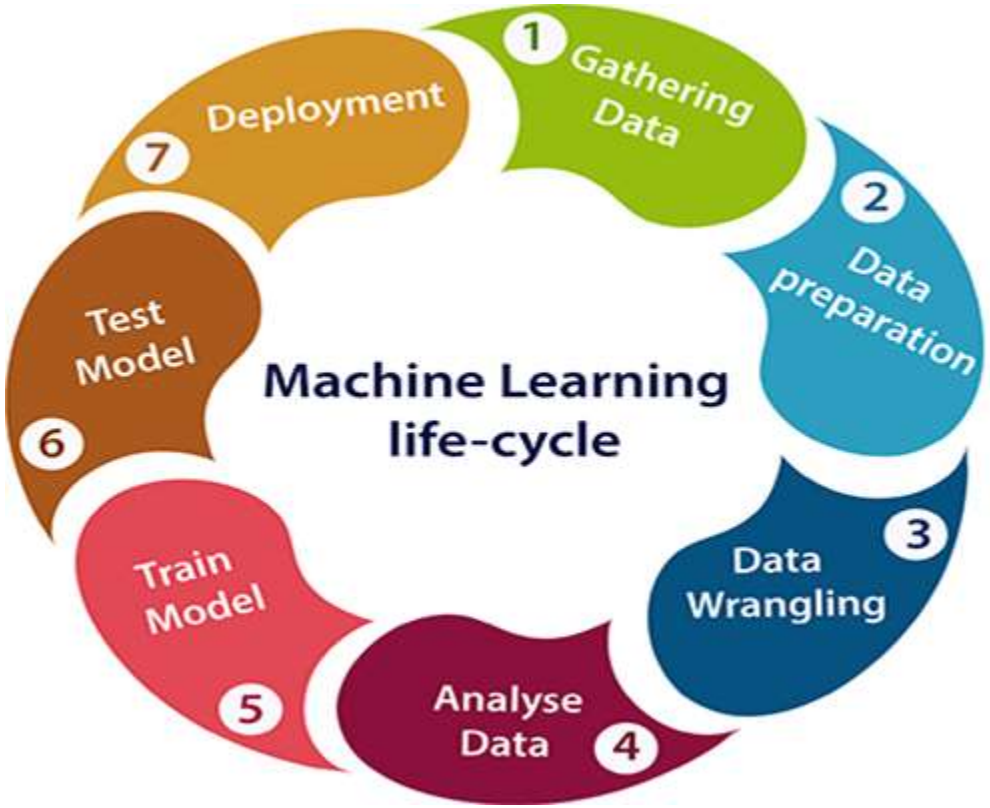
- Medical Diagnosis:
  - In medical science, machine learning is used for diseases diagnoses. With this, medical technology is growing very fast and able to build 3D models that can predict the exact position of lesions in the brain.
  - It helps in finding brain tumors and other brain-related diseases easily.
- Automatic Language Translation:
  - Nowadays, if we visit a new place and we are not aware of the language then it is not a problem at all, as for this also machine learning helps us by converting the text into our known languages. Google's GNMT (Google Neural Machine Translation) provide this feature, which is a Neural Machine Learning that translates the text into our familiar language, and it called as automatic translation.
  - The technology behind the automatic translation is a sequence to sequence learning algorithm, which is used with image recognition and translates the text from one language to another language.

Machine learning Life cycle

- Machine learning has given the computer systems the abilities to automatically learn without being explicitly programmed. But how does a machine learning system work? So, it can be described using the life cycle of machine learning. Machine learning life cycle is a cyclic process to build an efficient machine learning project. The main purpose of the life cycle is to find a solution to the problem or project.

- Machine learning life cycle involves seven major steps, which are given below:
  - **Gathering Data**
  - **Data preparation**
  - **Data Wrangling**
  - **Analyse Data**
  - **Train the model**
  - **Test the model**
  - **Deployment**

Machine learning Life cycle



Gathering Data:

- Data Gathering is the first step of the machine learning life cycle. The goal of this step is to identify and obtain all data-related problems.
- In this step, we need to identify the different data sources, as data can be collected from various sources such as **files**, **database**, **internet**, or **mobile devices**. It is one of the most important steps of the life cycle. The quantity and quality of the collected data will determine the efficiency of the output. The more will be the data, the more accurate will be the prediction.
- This step includes the below tasks:
- **Identify various data sources**
- **Collect data**
- **Integrate the data obtained from different sources**
- By performing the above task, we get a coherent set of data, also called as a **dataset**. It will be used in further steps.

Data preparation

- After collecting the data, we need to prepare it for further steps. Data preparation is a step where we put our data into a suitable place and prepare it to use in our machine learning training.
- In this step, first, we put all data together, and then randomize the ordering of data.
- This step can be further divided into two processes:
- **Data exploration:**

  It is used to understand the nature of data that we have to work with. We need to understand the characteristics, format, and quality of data. A better understanding of data leads to an effective outcome. In this, we find Correlations, general trends, and outliers.
- **Data pre-processing:**
  Now the next step is preprocessing of data for its analysis.

Data Wrangling

- **Data wrangling** is the process of cleaning and converting raw data into a useable format. It is the process of cleaning the data, selecting the variable to use, and transforming the data in a proper format to make it more suitable for analysis in the next step. It is one of the most important steps of the complete process. Cleaning of data is required to address the quality issues.
- It is not necessary that data we have collected is always of our use as some of the data may not be useful. In real-world applications, collected data may have various issues, including:
- **Missing Values**
- **Duplicate data**
- **Invalid data**
- **Noise**
- So, we use various filtering techniques to clean the data.
- It is mandatory to detect and remove the above issues because it can negatively affect the quality of the outcome.

Data Analysis

- Now the cleaned and prepared data is passed on to the analysis step. This step involves:
- **Selection of analytical techniques**
- **Building models**
- **Review the result**
- The aim of this step is to build a machine learning model to analyze the data using various analytical techniques and review the outcome. It starts with the determination of the type of the problems, where we select the machine learning techniques such as **Classification**, **Regression**, **Cluster analysis**, **Association**, etc. then build the model using prepared data, and evaluate the model.
- Hence, in this step, we take the data and use machine learning algorithms to build the model.

Train Model

- Now the next step is to train the model, in this step we train our model to improve its performance for better outcome of the problem.
- We use datasets to train the model using various machine learning algorithms. Training a model is required so that it can understand the various patterns, rules, and, features.

Test Model

- Once our machine learning model has been trained on a given dataset, then we test the model. In this step, we check for the accuracy of our model by providing a test dataset to it.
- Testing the model determines the percentage accuracy of the model as per the requirement of project or problem.

Deployment

- The last step of machine learning life cycle is deployment, where we deploy the model in the real-world system.
- If the above-prepared model is producing an accurate result as per our requirement with acceptable speed, then we deploy the model in the real system. But before deploying the project, we will check whether it is improving its performance using available data or not. The deployment phase is similar to making the final report for a project.

Data in Machine Learning

- **Data** is a crucial component in the field of Machine Learning. It refers to the set of observations or measurements that can be used to train a machine-learning model. The quality and quantity of data available for training and testing play a significant role in determining the performance of a machine-learning model. Data can be in various forms such as numerical, categorical, or time-series data, and can come from various sources such as databases, spreadsheets, or APIs. Machine learning algorithms use data to learn patterns and relationships between input variables and target outputs, which can then be used for prediction or classification tasks.
  - Data is typically divided into two types:
    - Labeled data
    - Unlabeled data

Data in Machine Learning

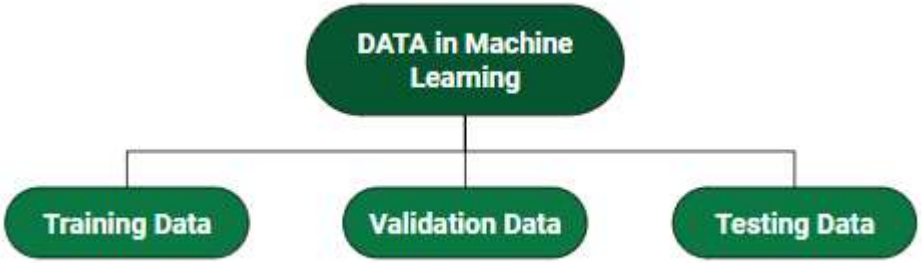
- Labeled data includes a label or target variable that the model is trying to predict, whereas unlabeled data does not include a label or target variable. The data used in machine learning is typically numerical or categorical. Numerical data includes values that can be ordered and measured, such as age or income. Categorical data includes values that represent categories, such as gender or type of fruit.
- Data can be divided into training and testing sets. The training set is used to train the model, and the testing set is used to evaluate the performance of the model. It is important to ensure that the data is split in a random and representative way. Data preprocessing is an important step in the machine learning pipeline. This step can include cleaning and normalizing the data, handling missing values, and feature selection or engineering.

Data in Machine Learning

- **DATA:** It can be any unprocessed fact, value, text, sound, or picture that is not being interpreted and analyzed. Data is the most important part of all Data Analytics, Machine Learning, and Artificial Intelligence. Without data, we can't train any model and all modern research and automation will go in vain. Big Enterprises are spending lots of money just to gather as much certain data as possible.
- **Example:** Why did Facebook acquire WhatsApp by paying a huge price of $19 billion?
- The answer is very simple and logical – it is to have access to the users' information that Facebook may not have but WhatsApp will have. This information about their users is of paramount importance to Facebook as it will facilitate the task of improvement in their services.
- **INFORMATION:** Data that has been interpreted and manipulated and has now some meaningful inference for the users.
- **KNOWLEDGE:** Combination of inferred information, experiences, learning, and insights. Results in awareness or concept building for an individual or organization.

How do we split data in Machine Learning?

- **Training Data:** The part of data we use to train our model. This is the data that your model actually sees(both input and output) and learns from.
- **Validation Data:** The part of data that is used to do a frequent evaluation of the model, fit on the training dataset along with improving involved hyperparameters (initially set parameters before the model begins learning). This data plays its part when the model is actually training.
- **Testing Data:** Once our model is completely trained, testing data provides an unbiased evaluation. When we feed in the inputs of Testing data, our model will predict some values(without seeing actual output). After prediction, we evaluate our model by comparing it with the actual output present in the testing data. This is how we evaluate and see how much our model has learned from the experiences feed in as training data, set at the time of training.



Different Forms of Data

- **Numeric Data** : If a feature represents a characteristic measured in numbers , it is called a numeric feature.
- **Categorical Data :** A categorical feature is an attribute that can take on one of the limited , and usually fixed number of possible values on the basis of some qualitative property . A categorical feature is also called a nominal feature.
- **Ordinal Data** : This denotes a nominal variable with categories falling in an ordered list . Examples include clothing sizes such as small, medium , and large , or a measurement of customer satisfaction on a scale from “not at all happy” to “very happy”.

Properties of Data

- **Volume:** Scale of Data. With the growing world population and technology at exposure, huge data is being generated each and every millisecond.
- **Variety:** Different forms of data – healthcare, images, videos, audio clippings.
- **Velocity:** Rate of data streaming and generation.
- **Value:** Meaningfulness of data in terms of information that researchers can infer from it.
- **Veracity:** Certainty and correctness in data we are working on.
- **Viability**: The ability of data to be used and integrated into different systems and processes.
- **Security**: The measures taken to protect data from unauthorized access or manipulation.
- **Accessibility**: The ease of obtaining and utilizing data for decision-making purposes.
- **Integrity**: The accuracy and completeness of data over its entire lifecycle.
- **Usability**: The ease of use and interpretability of data for end-users.

Issues of using data in Machine Learning:

- **Data quality:** One of the biggest issues with using data in machine learning is ensuring that the data is accurate, complete, and representative of the problem domain. Low-quality data can result in inaccurate or biased models.
- **Data quantity:** In some cases, there may not be enough data available to train an accurate machine learning model. This is especially true for complex problems that require a large amount of data to accurately capture all the relevant patterns and relationships.
- **Bias and fairness:** Machine learning models can sometimes perpetuate bias and discrimination if the training data is biased or unrepresentative. This can lead to unfair outcomes for certain groups of people, such as minorities or women.
- **Overfitting and underfitting:** Overfitting occurs when a model is too complex and fits the training data too closely, resulting in poor generalization to new data. Underfitting occurs when a model is too simple and does not capture all the relevant patterns in the data.
- **Privacy and security:** Machine learning models can sometimes be used to infer sensitive information about individuals or organizations, raising concerns about privacy and security.
- **Interpretability:** Some machine learning models, such as deep neural networks, can be difficult to interpret and understand, making it challenging to explain the reasoning behind their predictions and decisions.

Data remediation

- Data remediation is the process of cleansing, organizing and migrating data so that it's properly protected and best serves its intended purpose. There is a misconception that data remediation simply means deleting business data that is no longer needed. It's important to remember that the key word "remediation" derives from the word "remedy," which is to correct a mistake. Since the core initiative is to correct data, the data remediation process typically involves replacing, modifying, cleansing or deleting any "dirty" data.

Data remediation terminology

These are common terms related to data remediation that you should get acquainted with.

1. **Data Migration** – The process of moving data between two or more systems, data formats or servers.
2. **Data Discovery** – A manual or automated process of searching for patterns in data sets to identify structured and unstructured data in an organization's systems.
3. **ROT** – An acronym that stands for redundant, obsolete and trivial data. According to the Association for Intelligent Information Management, ROT data accounts for nearly 80 percent of the unstructured data that is beyond its recommended retention period and no longer useful to an organization.
4. **Dark Data** – Any information that businesses collect, process and store, but do not use for other purposes. Some examples include customer call records, raw survey data or email correspondences. Often, the storing and securing of this type of data incurs more expense and sometimes even greater risk than it does value.
5. **Dirty Data** – Data that damages the integrity of the organization's complete dataset. This can include data that is unnecessarily duplicated, outdated, incomplete or inaccurate.
6. **Data Overload** – This is when an organization has acquired too much data, including low-quality or dark data. Data overload makes the tasks of identifying, classifying and remediating data laborious.
7. **Data Cleansing** – Transforming data in its native state to a predefined standardized format.
8. **Data Governance** – Management of the availability, usability, integrity and security of the data stored within an organization

Stages of data remediation

1. Assessment
2. Organizing and segmentation
3. Indexation and classification
4. Migrating
5. Data cleansing

ThankYou