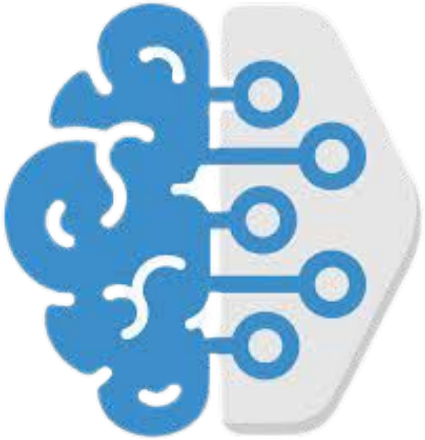
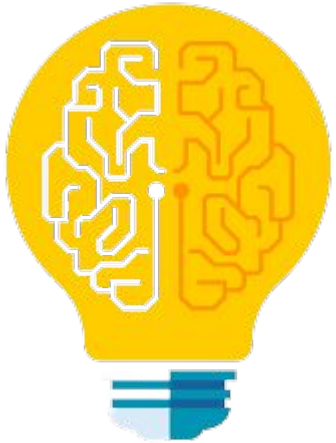




Machine learning : Introduction

Prof.Trupesh Patel,
Assistant Professor, Computer Engineering Department,
trupesh1991@gmail.com

Topics to be discussed

Introduction

Statistical Learning

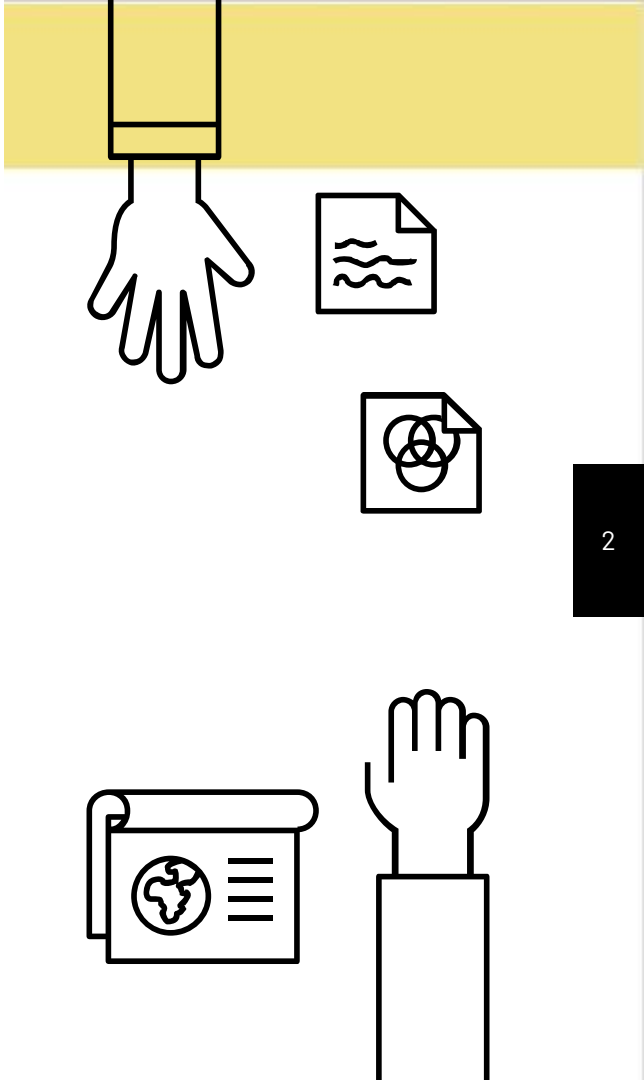
Supervised Learning

Unsupervised Learning

Reinforcement Learning

Linear Algebra basics

Probability basics



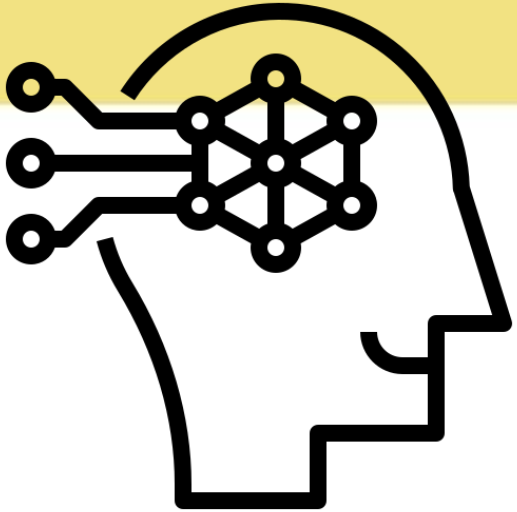
INTRODUCTION

- **Learning :** 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.

Examples -
Handwriting recognition learning problem
A robot driving learning problem

Prof.Trupesh Patel,
Assistant Professor, Computer Engineering Department,
trupesh1991@gmail.com

Machine learning



"The field of study that gives computers the ability to learn without being explicitly programmed." - Arthur samuel,1959- IBM USA

Prof.Trupesh Patel,
Assistant Professor, Computer Engineering Department,
trupesh1991@gmail.com

How does ML works ?

1. **A Decision Process**: In general, machine learning algorithms are used to make a prediction or classification. Based on some input data, which can be labeled or unlabeled, your algorithm will produce an estimate about a pattern in the data.
2. **An Error Function:** An error function evaluates the prediction of the model. If there are known examples, an error function can make a comparison to assess the accuracy of the model.
3. **A Model Optimization Process:** If the model can fit better to the data points in the training set, then weights are adjusted to reduce the discrepancy between the known example and the model estimate. The algorithm will repeat this “evaluate and optimize” process, updating weights autonomously until a threshold of accuracy has been met.

Prof.Trupesh Patel,
Assistant Professor, Computer Engineering Department,
trupesh1991@gmail.com

Statistics and Machine Learning

- Statistics is the mathematical study of data. Using statistics, an interpretable statistical model is created to describe the data, and this model can then be used to infer something about the data or even to predict values that are not present in the sample data used to create the model.
- The ‘accuracy’ of prediction is not the focus of statistics.
- To avoid overfitting the data using machine learning methods, and indeed often in statistics, the observed data set is tested by separating out a portion of the data (known as the *test set*) to confirm the strength of the model built by the majority of the data (known as the *training set*). Often, a *validation set* within the training set is used to determine the validity of the predictive model before it’s confirmed on the test data set.

Prof.Trupesh Patel,
Assistant Professor, Computer Engineering Department,
trupesh1991@gmail.com

ML Algorithms :

- Neural networks: Neural networks simulate the way the human brain works, with a huge number of linked processing nodes. Neural networks are good at recognizing patterns and play an important role in applications including natural language translation, image recognition, speech recognition, and image creation.
- Linear regression: This algorithm is used to predict numerical values, based on a linear relationship between different values. For example, the technique could be used to predict house prices based on historical data for the area.
- Logistic regression: This supervised learning algorithm makes predictions for categorical response variables, such as“yes/no” answers to questions. It can be used for applications such as classifying spam and quality control on a production line.

Prof.Trupesh Patel,
Assistant Professor, Computer Engineering Department,
trupesh1991@gmail.com

ML Algorithms :

- Clustering: Using unsupervised learning, clustering algorithms can identify patterns in data so that it can be grouped. Computers can help data scientists by identifying differences between data items that humans have overlooked.
- Decision trees: Decision trees can be used for both predicting numerical values (regression) and classifying data into categories. Decision trees use a branching sequence of linked decisions that can be represented with a tree diagram. One of the advantages of decision trees is that they are easy to validate and audit, unlike the black box of the neural network.
- Random forests: In a random forest, the machine learning algorithm predicts a value or category by combining the results from a number of decision trees.

Prof.Trupesh Patel,
Assistant Professor, Computer Engineering Department,
trupesh1991@gmail.com

MI use cases

Speech recognition: It is also known as automatic speech recognition (ASR), computer speech recognition, or speech-to-text, and it is a capability which uses natural language processing (NLP) to translate human speech into a written format. Many mobile devices incorporate speech recognition into their systems to conduct voice search—e.g. Siri—or improve accessibility for texting.

Customer service: Customer service: Online chatbots are replacing human agents along the customer journey, changing the way we think about customer engagement across websites and social media platforms. Chatbots answer frequently asked questions (FAQs) about topics such as shipping, or provide personalized advice, cross-selling products or suggesting sizes for users. Examples include virtual agents on e-commerce sites; messaging bots, using Slack and Facebook Messenger; and tasks usually done by virtual assistants and voice assistants.

Prof.Trupesh Patel,
Assistant Professor, Computer Engineering Department,
trupesh1991@gmail.com

MI use cases

Computer vision: This AI technology enables computers to derive meaningful information from digital images, videos, and other visual inputs, and then take the appropriate action. Powered by convolutional neural networks, computer vision has applications in photo tagging on social media, radiology imaging in healthcare, and self-driving cars in the automotive industry.

Recommendation engines: Using past consumption behavior data, AI algorithms can help to discover data trends that can be used to develop more effective cross-selling strategies. This approach is used by online retailers to make relevant product recommendations to customers during the checkout process.
Automated stock trading: Designed to optimize stock portfolios, AI-driven high-frequency trading platforms make thousands or even millions of trades per day without human intervention.

Fraud detection: Banks and other financial institutions can use machine learning to spot suspicious transactions. Supervised learning can train a model using information about known fraudulent transactions. Anomaly detection can identify transactions that look atypical and deserve further investigation.

Prof.Trupesh Patel,
Assistant Professor, Computer Engineering Department,
trupesh1991@gmail.com

Supervised learning

It is defined by its use of labeled datasets to train algorithms that to classify data or predict outcomes accurately.
As input data is fed into the model, it adjusts its weights until the model has been fitted appropriately, which occurs as part of the cross validation process. Supervised learning helps organizations solve for a variety of real-world problems at scale, such as classifying spam in a separate folder from your inbox.

How it works :
Supervised learning uses a training set to teach models to yield the desired output. This training dataset includes inputs and correct outputs, which allow the model to learn over time. The algorithm measures its accuracy through the loss function, adjusting until the error has been sufficiently minimized.

Supervised learning can be separated into two types of problems when data mining—**classification and regression:**

Prof.Trupesh Patel,
Assistant Professor, Computer Engineering Department,
trupesh1991@gmail.com

Supervised learning

Classification uses an algorithm to accurately assign test data into specific categories. It recognizes specific entities within the dataset and attempts to draw some conclusions on how those entities should be labeled or defined. Common classification algorithms are linear classifiers, support vector machines (SVM), decision trees, k-nearest neighbor, and random forest.

Regression is used to understand the relationship between dependent and independent variables. It is commonly used to make projections, such as for sales revenue for a given business. Linear regression, logistical regression, and polynomial regression are popular regression algorithms.

Prof.Trupesh Patel,
Assistant Professor, Computer Engineering Department,
trupesh1991@gmail.com

SVL ALGORITHMS

Prof.Trupesh Patel,
Assistant Professor, Computer Engineering Department,
trupesh1991@gmail.com

NEURAL NETWORK

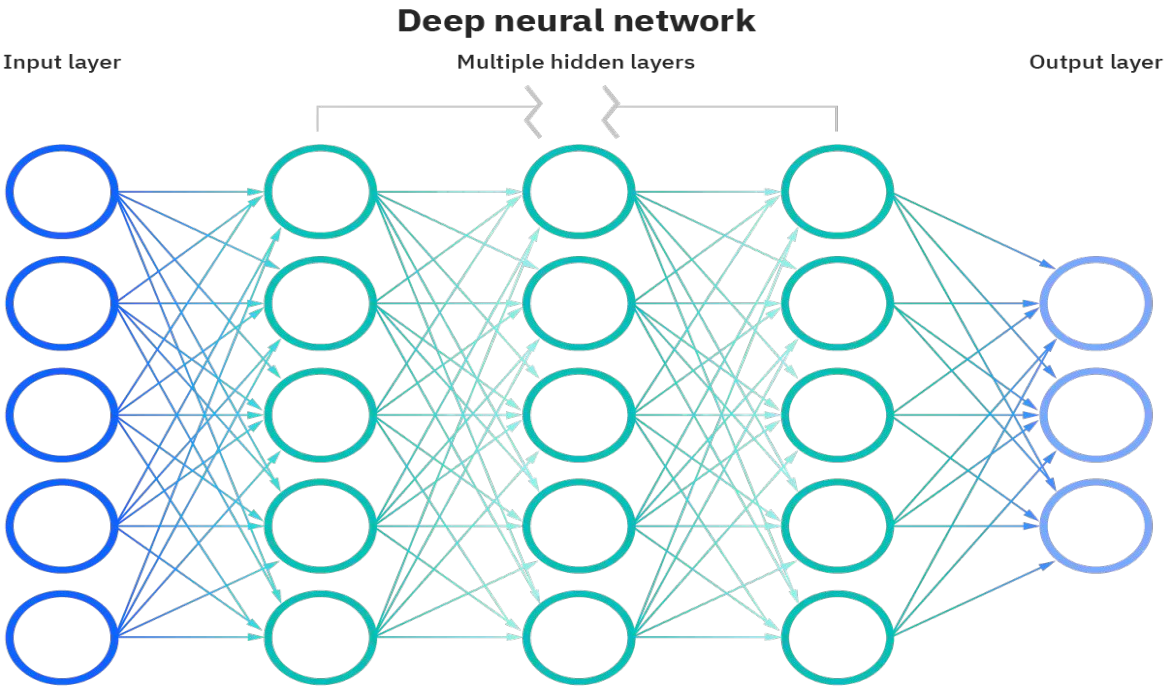
Primarily leveraged for deep learning algorithms, neural networks process training data by mimicking the interconnectivity of the human brain through layers of nodes. Each node is made up of inputs, weights, a bias (or threshold), and an output.

If that output value exceeds a given threshold, it “fires” or activates the node, passing data to the next layer in the network. Neural networks learn this mapping function through supervised learning, adjusting based on the loss function through the process of gradient descent.

When the cost function is at or near zero, we can be confident in the model’s accuracy to yield the correct answer.

Prof.Trupesh Patel,
Assistant Professor, Computer Engineering Department,
trupesh1991@gmail.com

NEURAL NETWORK



Prof.Trupesh Patel,
Assistant Professor, Computer Engineering Department,
trupesh1991@gmail.com

Source : https://www.ibm.com/cloud/learn/neural-networks

NAIVE BAYES

Naive Bayes is classification approach that adopts the principle of class conditional independence from the Bayes Theorem.
This means that the presence of one feature does not impact the presence of another in the probability of a given outcome, and each predictor has an equal effect on that result.
There are three types of Naïve Bayes classifiers: Multinomial Naïve Bayes, Bernoulli Naïve Bayes, and Gaussian Naïve Bayes.
This technique is primarily used in text classification, spam identification, and recommendation systems.

Prof.Trupesh Patel,
Assistant Professor, Computer Engineering Department,
trupesh1991@gmail.com

LINEAR REGRESSION

Linear regression is used to identify the relationship between a dependent variable and one or more independent variables and is typically leveraged to make predictions about future outcomes.

When there is only one independent variable and one dependent variable, it is known as simple linear regression.

As the number of independent variables increases, it is referred to as multiple linear regression.

For each type of linear regression, it seeks to plot a line of best fit, which is calculated through the method of least squares. However, unlike other regression models, this line is straight when plotted on a graph.

Prof.Trupesh Patel,
Assistant Professor, Computer Engineering Department,
trupesh1991@gmail.com

LOGISTIC REGRESSION

While linear regression is leveraged when dependent variables are continuous, logistic regression is selected when the dependent variable is categorical, meaning they have binary outputs, such as "true" and "false" or "yes" and "no."

While both regression models seek to understand relationships between data inputs, logistic regression is mainly used to solve binary classification problems, such as spam identification.

Prof.Trupesh Patel,
Assistant Professor, Computer Engineering Department,
trupesh1991@gmail.com

SUPPORT VECTOR MACHINE

A support vector machine is a popular supervised learning model developed by Vladimir Vapnik, used for both data classification and regression.
That said, it is typically leveraged for classification problems, constructing a hyperplane where the distance between two classes of data points is at its maximum.
This hyperplane is known as the decision boundary, separating the classes of data points (e.g., oranges vs. apples) on either side of the plane.

Prof.Trupesh Patel,
Assistant Professor, Computer Engineering Department,
trupesh1991@gmail.com

K-NEAREST NEIGHBOUR

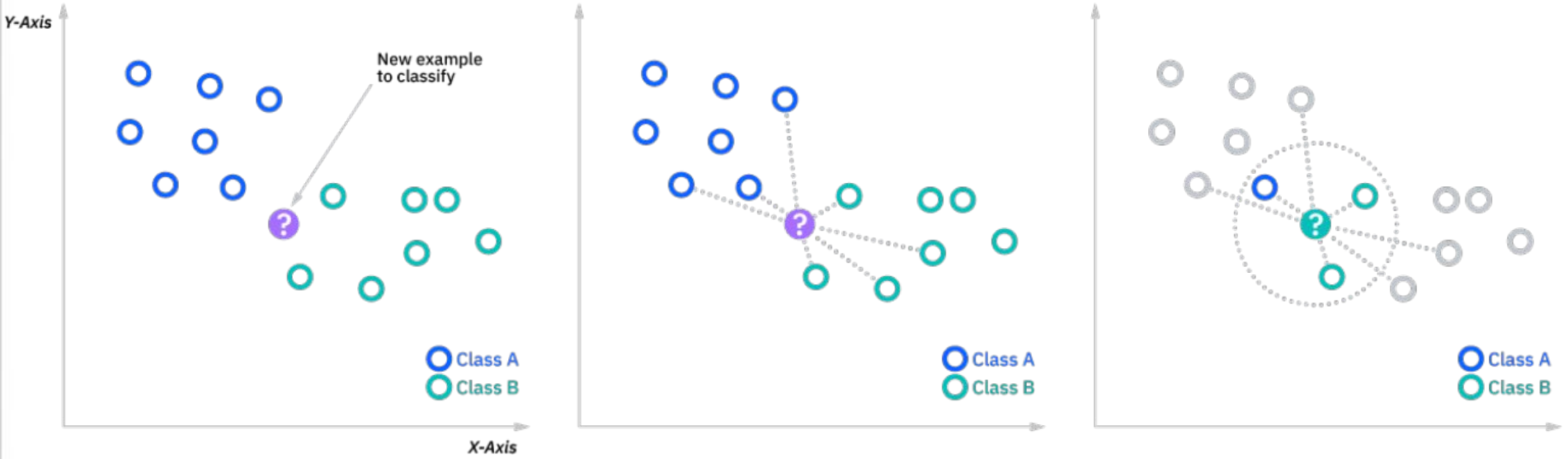
K-nearest neighbor, also known as the KNN algorithm, is a non-parametric algorithm that classifies data points based on their proximity and association to other available data. This algorithm assumes that similar data points can be found near each other.

As a result, it seeks to calculate the distance between data points, usually through Euclidean distance, and then it assigns a category based on the most frequent category or average.

Its ease of use and low calculation time make it a preferred algorithm by data scientists, but as the test dataset grows, the processing time lengthens, making it less appealing for classification tasks. KNN is typically used for recommendation engines and image recognition.

Prof.Trupesh Patel,
Assistant Professor, Computer Engineering Department,
trupesh1991@gmail.com

K-NEAREST NEIGHBOUR

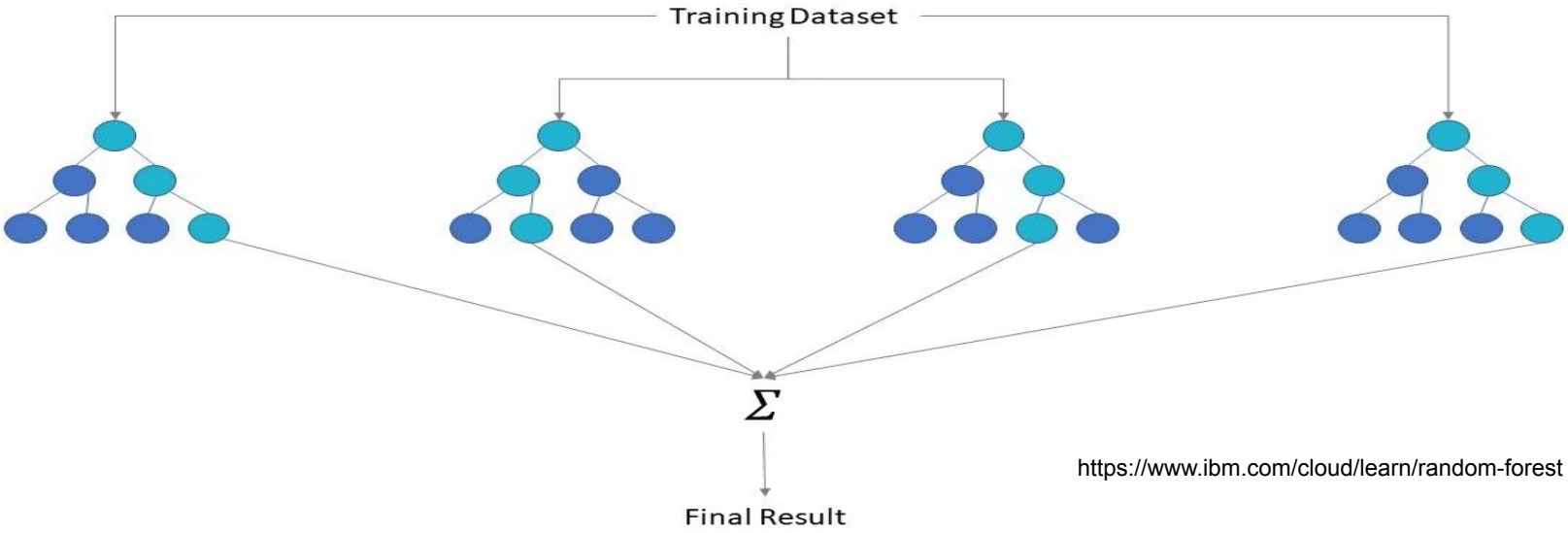


Prof.Trupesh Patel,
Assistant Professor, Computer Engineering Department,
trupesh1991@gmail.com

https://www.ibm.com/in-en/topics/knn

RANDOM FOREST

Random forest is another flexible supervised machine learning algorithm used for both classification and regression purposes. The "forest" references a collection of uncorrelated decision trees, which are then merged together to reduce variance and create more accurate data predictions.



https://www.ibm.com/cloud/learn/random-forest

SUPERVISED LEARNING : EXAMPLES

- Image- and object-recognition: Supervised learning algorithms can be used to locate, isolate, and categorize objects out of videos or images, making them useful when applied to various computer vision techniques and imagery analysis.
- Predictive analytics: A widespread use case for supervised learning models is in creating predictive analytics systems to provide deep insights into various business data points. This allows enterprises to anticipate certain results based on a given output variable, helping business leaders justify decisions or pivot for the benefit of the organization.
- Customer sentiment analysis: Using supervised machine learning algorithms, organizations can extract and classify important pieces of information from large volumes of data—including context, emotion, and intent—with very little human intervention. This can be incredibly useful when gaining a better understanding of customer interactions and can be used to improve brand engagement efforts.

Prof.Trupesh Patel,
Assistant Professor, Computer Engineering Department,
trupesh1991@gmail.com

SUPERVISED LEARNING : EXAMPLES

- Spam detection: Spam detection is another example of a supervised learning model. Using supervised classification algorithms, organizations can train databases to recognize patterns or anomalies in new data to organize spam and non-spam-related correspondences effectively.

Prof.Trupesh Patel,
Assistant Professor, Computer Engineering Department,
trupesh1991@gmail.com

UNSUPERVISED LEARNING

- Unsupervised learning, uses machine learning algorithms to analyze and cluster unlabeled datasets.
- These algorithms discover hidden patterns or data groupings without the need for human intervention.
- Its ability to discover similarities and differences in information make it the ideal solution for exploratory data analysis, cross-selling strategies, customer segmentation, and image recognition.
- Unsupervised learning models are utilized for three main tasks—clustering, association, and dimensionality reduction. Below we'll define each learning method and highlight common algorithms and approaches to conduct them effectively.

Prof.Trupesh Patel,
Assistant Professor, Computer Engineering Department,
trupesh1991@gmail.com

UNSUPERVISED LEARNING : Clustering

- Clustering is a data mining technique which groups unlabeled data based on their similarities or differences.
- Clustering algorithms are used to process raw, unclassified data objects into groups represented by structures or patterns in the information.
- Clustering algorithms can be categorized into a few types, specifically exclusive, overlapping, hierarchical, and probabilistic.

Prof.Trupesh Patel,
Assistant Professor, Computer Engineering Department,
trupesh1991@gmail.com

UNSUPERVISED LEARNING : Association rules

- An association rule is a rule-based method for finding relationships between variables in a given dataset.
- These methods are frequently used for market basket analysis, allowing companies to better understand relationships between different products.
- Understanding consumption habits of customers enables businesses to develop better cross-selling strategies and recommendation engines.
- Examples of this can be seen in Amazon's "Customers Who Bought This Item Also Bought" or Spotify's "Discover Weekly" playlist.
- While there are a few different algorithms used to generate association rules, such as Apriori, Eclat, and FP-Growth, the Apriori algorithm is most widely used.

Prof.Trupesh Patel,
Assistant Professor, Computer Engineering Department,
trupesh1991@gmail.com

UNSUPERVISED LEARNING : Dimensionality Reduction

While more data generally yields more accurate results, it can also impact the performance of machine learning algorithms (e.g. overfitting) and it can also make it difficult to visualize datasets.

Dimensionality reduction is a technique used when the number of features, or dimensions, in a given dataset is too high.

It reduces the number of data inputs to a manageable size while also preserving the integrity of the dataset as much as possible.

It is commonly used in the preprocessing data stage, and there are a few different dimensionality reduction methods that can be used, such as: Principal component analysis.

Prof.Trupesh Patel,
Assistant Professor, Computer Engineering Department,
trupesh1991@gmail.com

UNSUPERVISED LEARNING : Applications

- News Sections: Google News uses unsupervised learning to categorize articles on the same story from various online news outlets. For example, the results of a presidential election could be categorized under their label for “US” news.
- Computer vision: Unsupervised learning algorithms are used for visual perception tasks, such as object recognition.
- Medical imaging: Unsupervised machine learning provides essential features to medical imaging devices, such as image detection, classification and segmentation, used in radiology and pathology to diagnose patients quickly and accurately.
- Anomaly detection: Unsupervised learning models can comb through large amounts of data and discover atypical data points within a dataset. These anomalies can raise awareness around faulty equipment, human error, or breaches in security.

Prof.Trupesh Patel,
Assistant Professor, Computer Engineering Department,
trupesh1991@gmail.com

UNSUPERVISED LEARNING : Applications

- Customer personas: Defining customer personas makes it easier to understand common traits and business clients' purchasing habits. Unsupervised learning allows businesses to build better buyer persona profiles, enabling organizations to align their product messaging more appropriately.
- Recommendation Engines: Using past purchase behavior data, unsupervised learning can help to discover data trends that can be used to develop more effective cross-selling strategies. This is used to make relevant add-on recommendations to customers during the checkout process for online retailers.

Prof.Trupesh Patel,
Assistant Professor, Computer Engineering Department,
trupesh1991@gmail.com

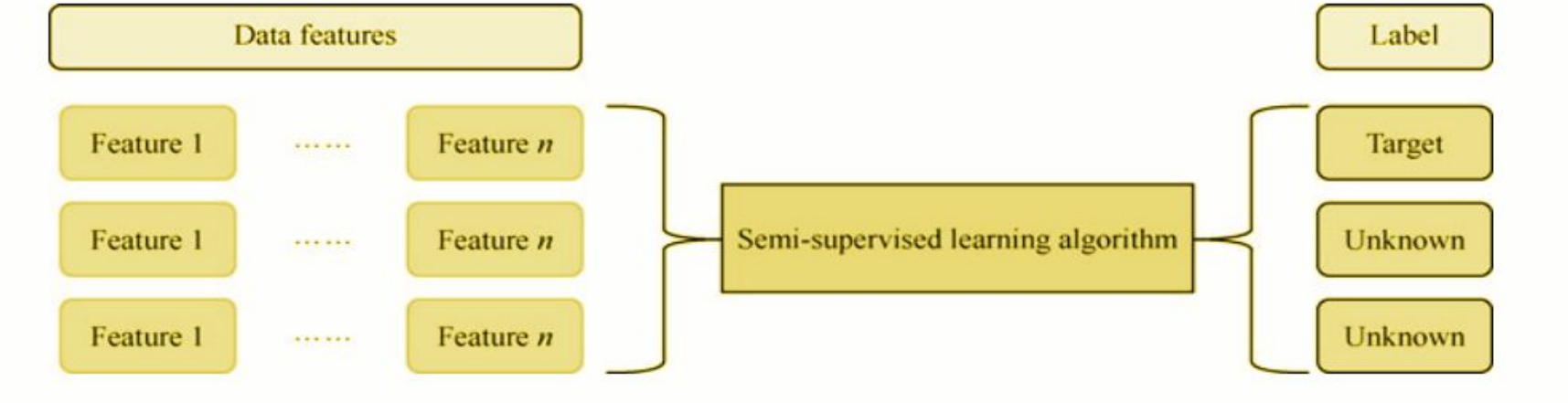
UNSUPERVISED LEARNING : challenges

- Computational complexity due to a high volume of training data
- Longer training times
- Higher risk of inaccurate results
- Human intervention to validate output variables
- Lack of transparency into the basis on which data was clustered

Prof.Trupesh Patel,
Assistant Professor, Computer Engineering Department,
trupesh1991@gmail.com

SEMI SUPERVISED LEARNING

- **SEMI SUPERVISED LEARNING = SUPERVISED + UNSUPERVISED LEARNING**
- It allow the learner to automatically utilize a large amount of unlabeled data to assist the learning of a small amount of labeled data.
- Traditional supervised learning algorithms need to learn from a large number of labeled training samples to build a model for predicting the labels of new samples.



Prof.Trupesh Patel,
Assistant Professor, Computer Engineering Department,
trupesh1991@gmail.com

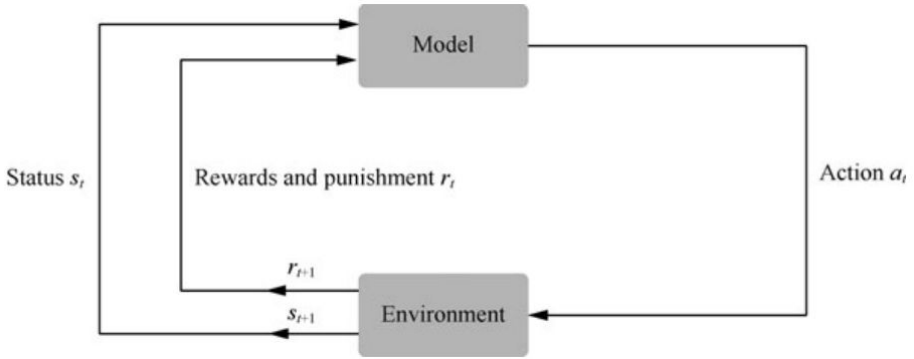
REINFORCEMENT LEARNING

- Reinforcement learning is mainly used to solve multi-step decision-making problems, such as **Go game**, **video games**, and **visual navigation**.
- **Taking Go** as an example. It takes about **10,170 operations** to exhaust the results of the game (there are only 1080 atoms in the universe). So, for a given and common situation, it is difficult to find the perfect move.
- Another characteristic of the multi-step decision problem is that it is easy to define a reward function to evaluate whether the task has been completed. The reward function of Go can be defined as whether to win the game; the reward function of electronic games can be defined as the score.
- The goal of reinforcement learning is to find an action strategy to maximize the value of the reward function.

Prof.Trupesh Patel,
Assistant Professor, Computer Engineering Department,
trupesh1991@gmail.com

REINFORCEMENT LEARNING

- Two most important parts of a reinforcement learning algorithm are the **model** and the **environment**.
- In different environments, the model can determine its own actions, and different actions may have different effects on the environment. Still, in the case of solving test questions, the computer can give the answer randomly, and the teacher will give a score based on the answer given. But if the situation is only limited to this case, it is impossible for the computer to learn how to solve the question, because the teacher's grading does not contribute to the training process.
- In this case, the importance of status and rewards and punishments are highlighted. A higher test score can make the teacher satisfied and then give the computer a certain reward.
- On the contrary, a lower test score may incur penalties.



Prof.Trupesh Patel,
Assistant Professor, Computer Engineering Department,
trupesh1991@gmail.com