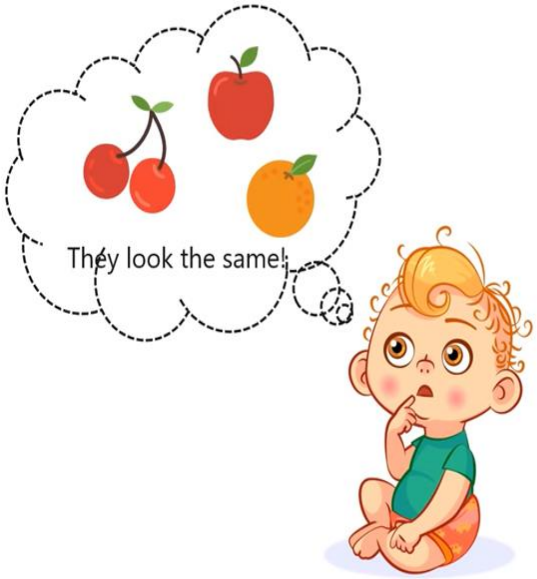
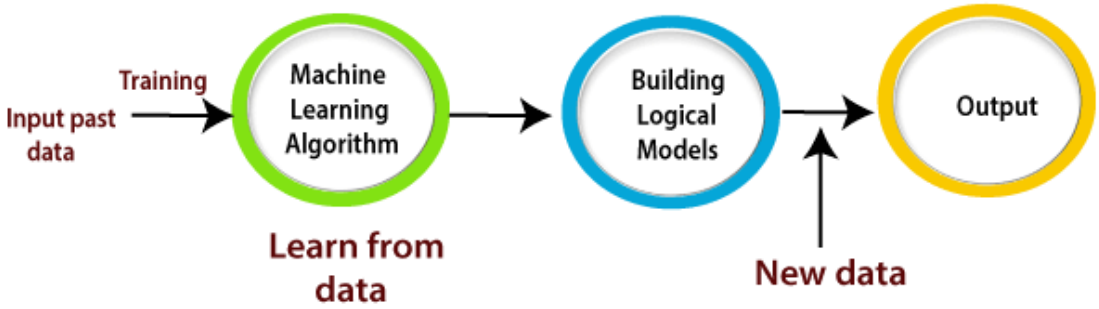
Module No. 1
Introduction to Machine Learning

Dr. Mrunal Rane

Syllabus :

- **Introduction to Machine Learning:**
  - Types of Machine Learning, Issues in Machine Learning,
  - Application of Machine Learning,
  - Steps involved in developing a Machine Learning Application.

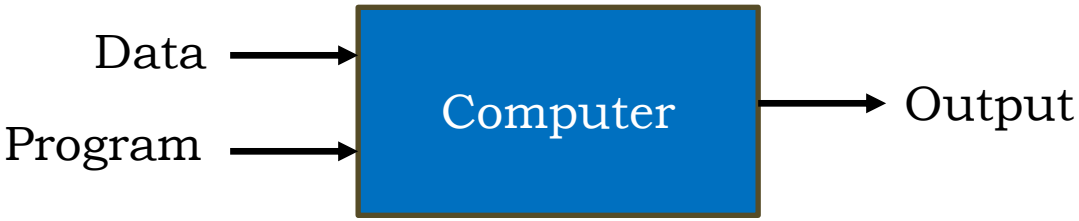

They look the same!


Training
Input past data
Machine Learning Algorithm
Learn from data
Building Logical Models
Output
New data

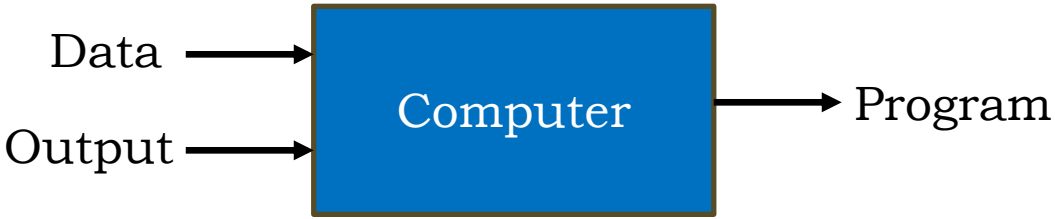
What is Machine Learning?

- “Learning is any process by which a system improves performance from experience.” - Herbert Simon
- Definition by Tom Mitchell (1998): Machine Learning is the study of algorithms that
  - improve their performance P
  - at some task T
  - with experience E.
- A well-defined learning task is given by <P, T, E>

- **Traditional Programming**



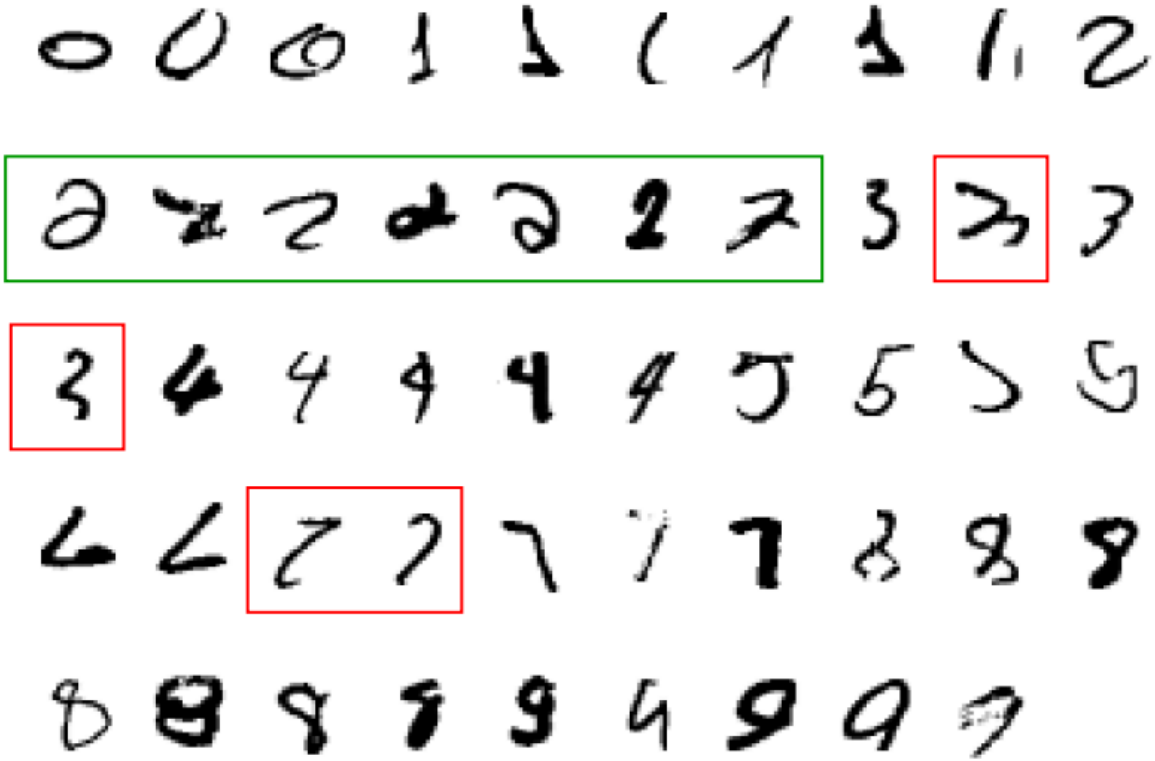
- **Machine Learning**



When Do We Use Machine Learning?

- ML is used when:
  - Human expertise does not exist
    - navigating on Mars
  - Humans can't explain their expertise
    - speech recognition
  - Models must be customized
    - personalized medicine
  - Models are based on huge amounts of data
    - genomics

- A classic example of a task that requires machine learning: It is very hard to say what makes a 2



Some more examples of tasks that are best solved by using a learning algorithm

- Recognizing patterns:
  - Facial identities or facial expressions
  - Handwritten or spoken words
  - Medical images
- Generating patterns:
  - Generating images or motion sequences
- Recognizing anomalies:
  - Unusual credit card transactions
  - Unusual patterns of sensor readings in a nuclear power plant
- Prediction:
  - Future stock prices or currency exchange rates

Learning :

- ***Learning or training is a process by means of which a neural network adapts itself to a stimulus by making proper parameter adjustments, resulting in a desired response.***
- A neural network, made up of a set of interconnected neurons, is related to its environment.
- The definition of learning process implies,
  1. The NN is ***stimulated*** by an environment.
  2. The NN ***undergoes changes*** in its free parameters as a result of stimulation.
  3. The NN ***responds in a new way*** to the environment because the changes have occurred in its internal structure.

Learning . . . :

- Types of learning :

  - ***Structure learning :*** changes in network structure
    also called as ***learning paradigm***

  - ***Parameter learning :*** updating the weights
    also called as ***learning algorithm***

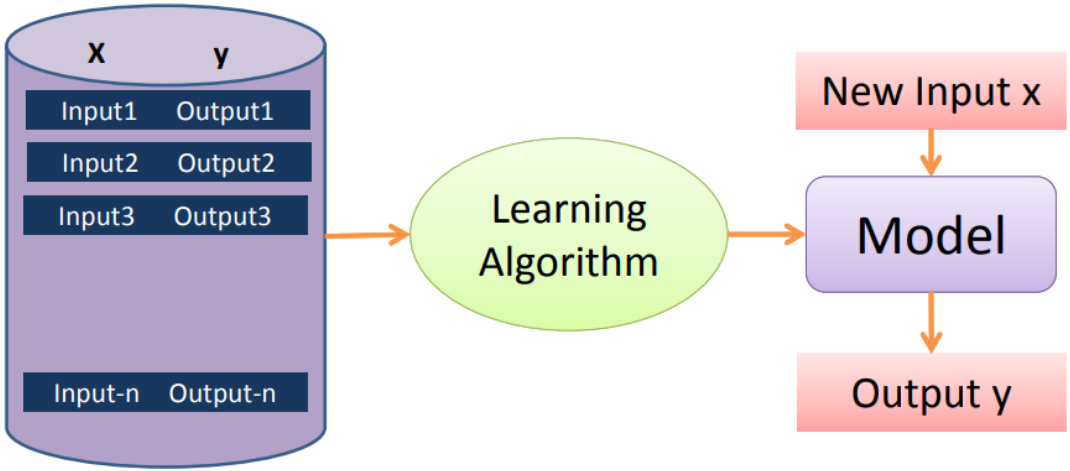
Learning Paradigm :

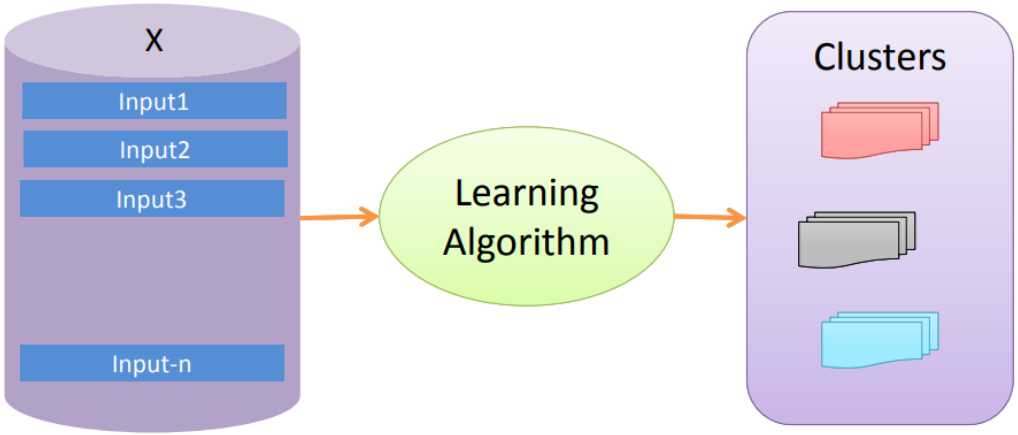
- ***Learning paradigm*** refers to the model of the environment in which the neural network operates.

- Based on this, generally a neural network is classified as,
  1. Supervised learning
  2. Unsupervised learning
  3. Semi-supervised learning
  4. Reinforcement learning

Broad types of machine learning

Supervised Learning	Unsupervised learning	Reinforcement Learning
– X, y (pre-classified training examples) – Given an observation x, what is the best label for y	– X – Given a set of x's, – cluster or summarize them	– Determine what to do based on rewards and punishments
Semi-supervised Learning		

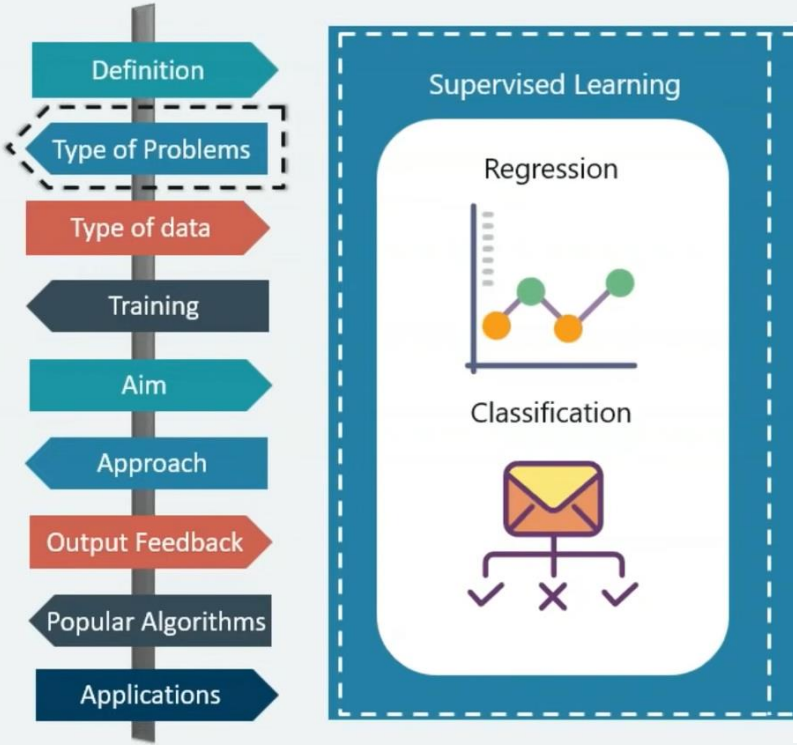

X
y
Input1
Output1
Input2
Output2
Input3
Output3
Input-n
Output-n
Learning Algorithm
New Input x
Model
Output y


X
Input1
Input2
Input3
Input-n
Learning Algorithm
Clusters

- Definition
- Type of Problems
- Type of data
- Training
- Aim
- Approach
- Output Feedback
- Popular Algorithms
- Applications

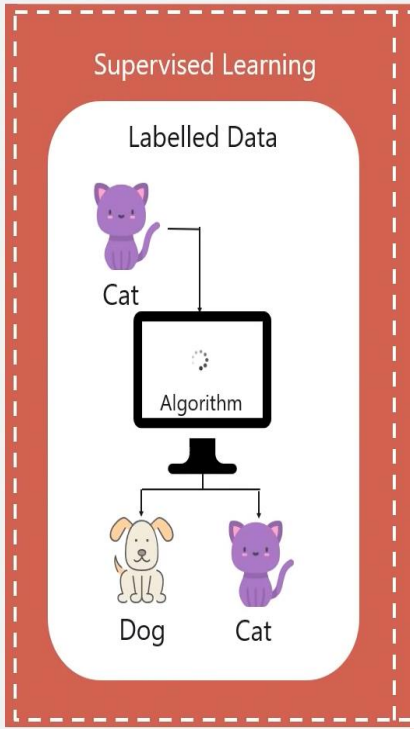
Supervised learning is a method in which we teach the machine using labelled data

Problem Type



Type of data


Definition
Type of Problems
Type of data
Training
Aim
Approach
Output Feedback
Popular Algorithms
Applications

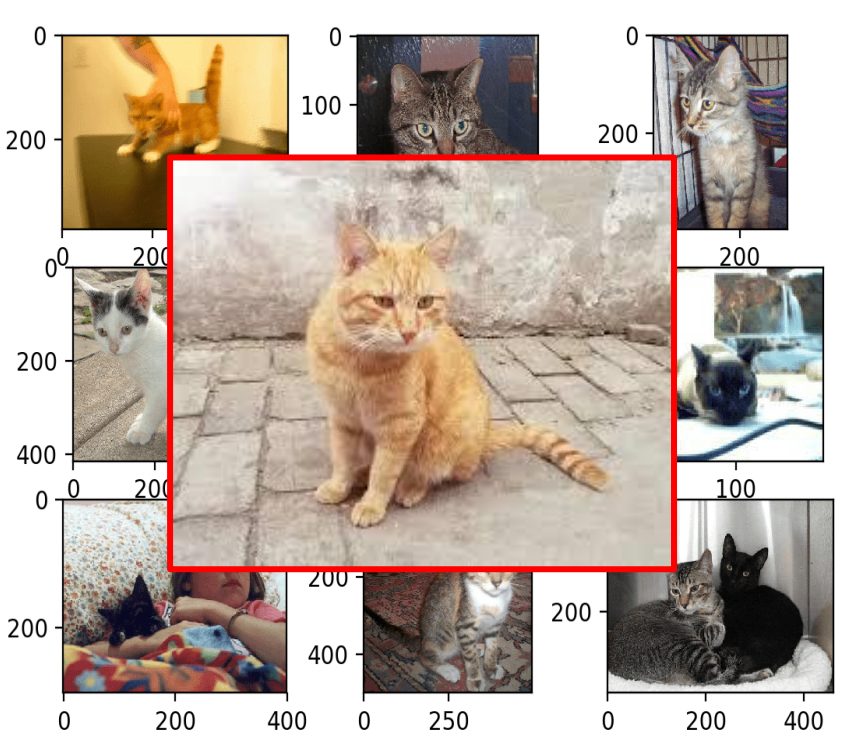

Supervised Learning
Labelled Data
Cat
Algorithm
Dog
Cat

Training

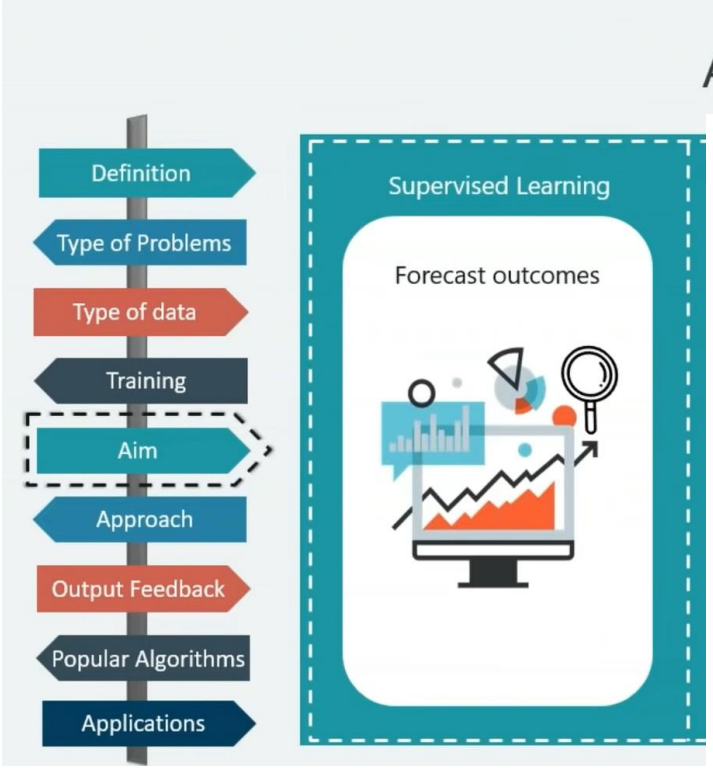
- Definition
- Type of Problems
- Type of data
- Training
- Aim
- Approach
- Output Feedback
- Popular Algorithms
- Applications

Supervised Learning

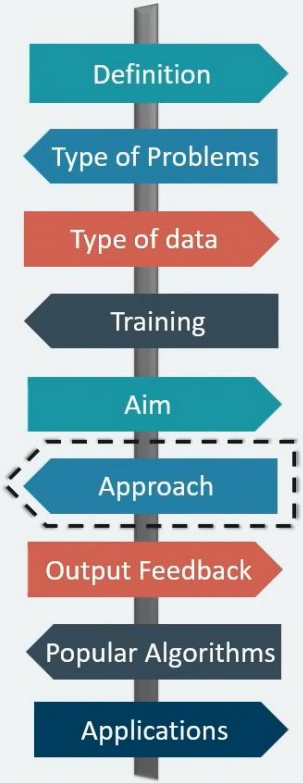
External supervision

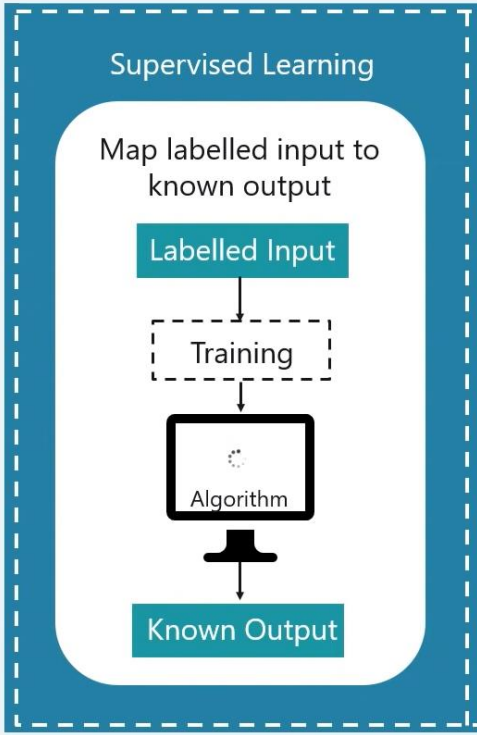


Aim

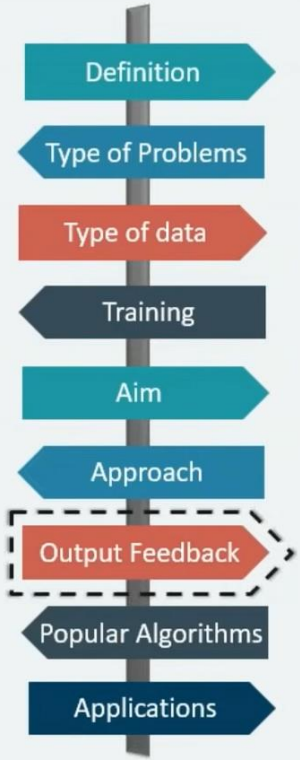

Definition
Type of Problems
Type of data
Training
Aim
Approach
Output Feedback
Popular Algorithms
Applications
Supervised Learning
Forecast outcomes

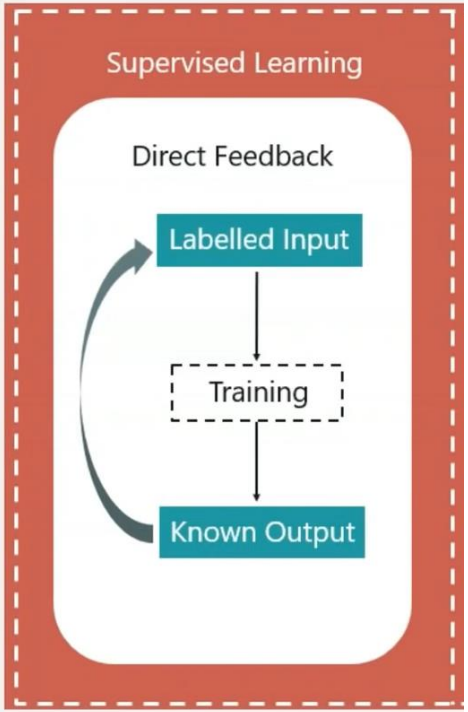
Approach


Definition
Type of Problems
Type of data
Training
Aim
Approach
Output Feedback
Popular Algorithms
Applications

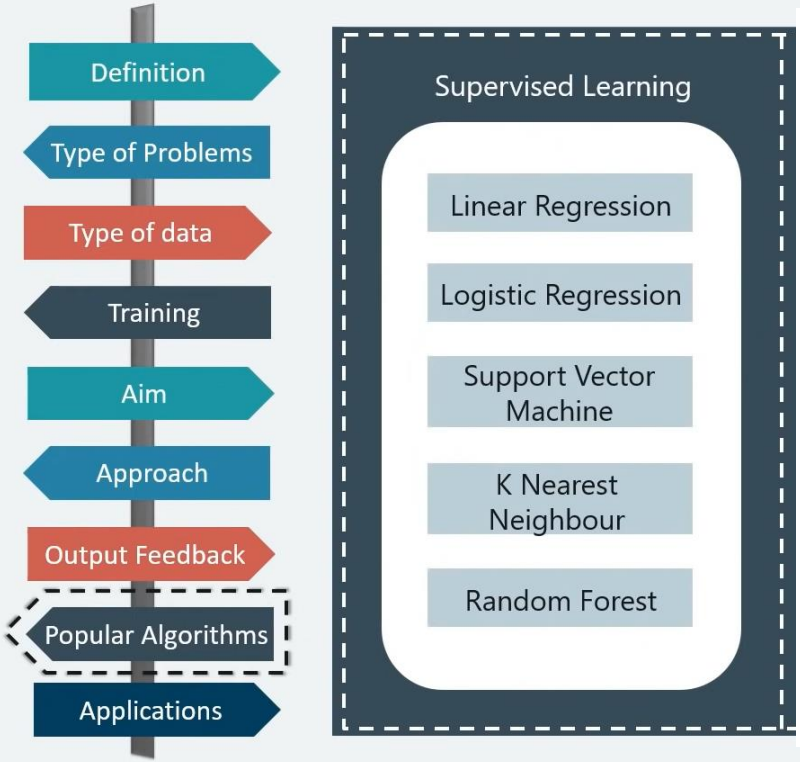

Supervised Learning
Map labelled input to known output
Labelled Input
Training
Algorithm
Known Output

Output Feedback





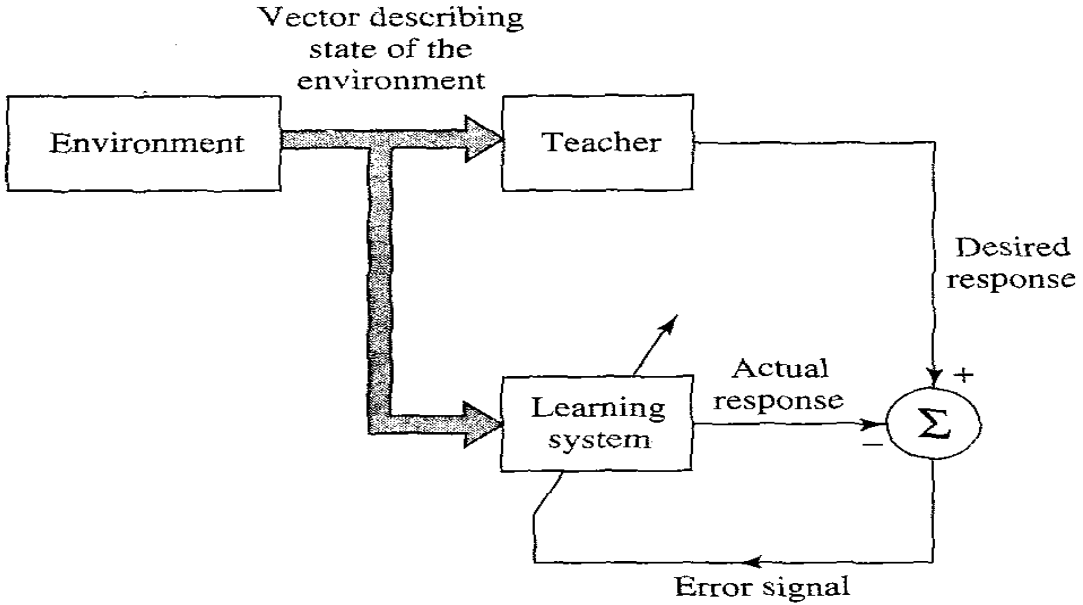
Popular Algorithms



Applications

Supervised Learning :

- The error minimization in this kind of training requires a supervisor or a teacher.



- **Advantages:**
- Since supervised learning work with the labelled dataset so we can have an exact idea about the classes of objects.
- These algorithms are helpful in predicting the output on the basis of prior experience.
- **Disadvantages:**
- These algorithms are not able to solve complex tasks.
- It may predict the wrong output if the test data is different from the training data.
- It requires lots of computational time to train the algorithm.

Applications of Supervised Learning

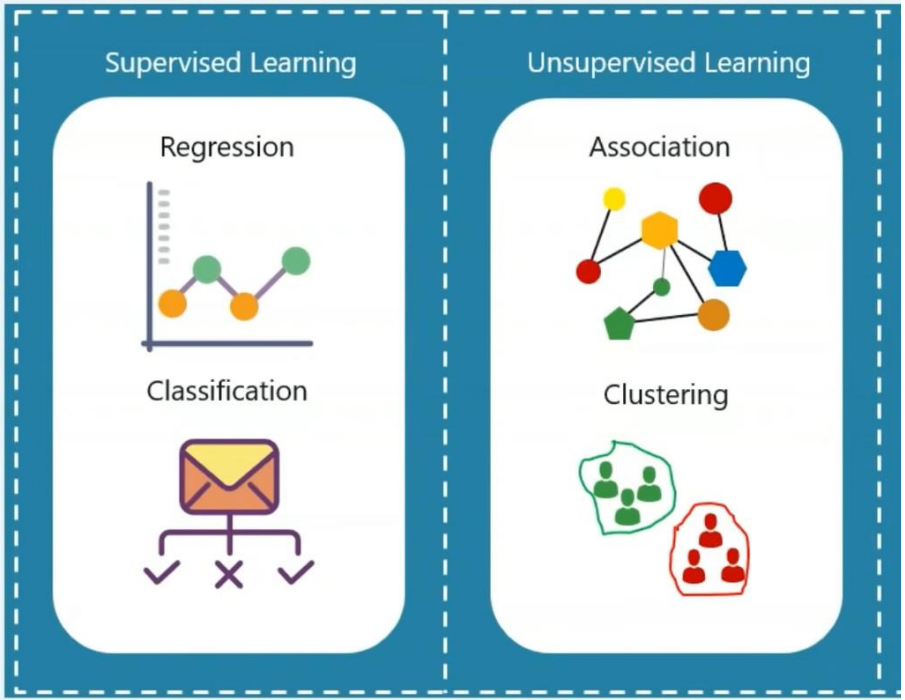
- Image Segmentation
- Medical Diagnosis
- Fraud Detection
- Spam detection
- Speech Recognition

- Definition
- Type of Problems
- Type of data
- Training
- Aim
- Approach
- Output Feedback
- Popular Algorithms
- Applications

Supervised learning is a method in which we teach the machine using labelled data

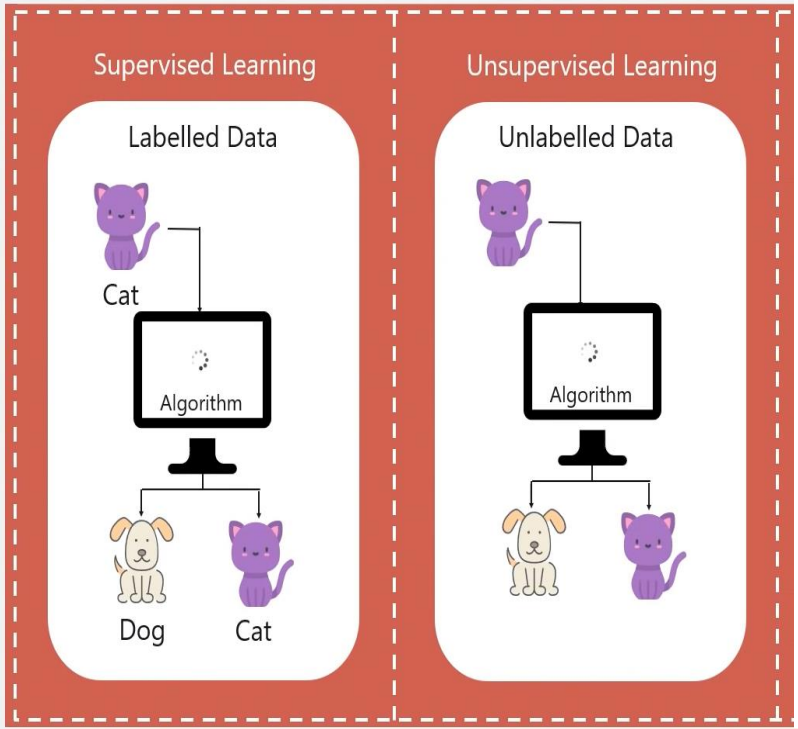
In unsupervised learning the machine is trained on unlabelled data without any guidance

Problem Type

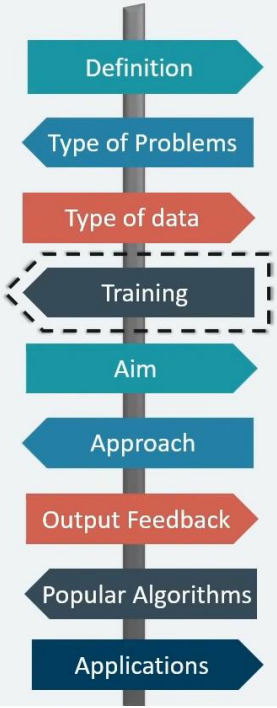


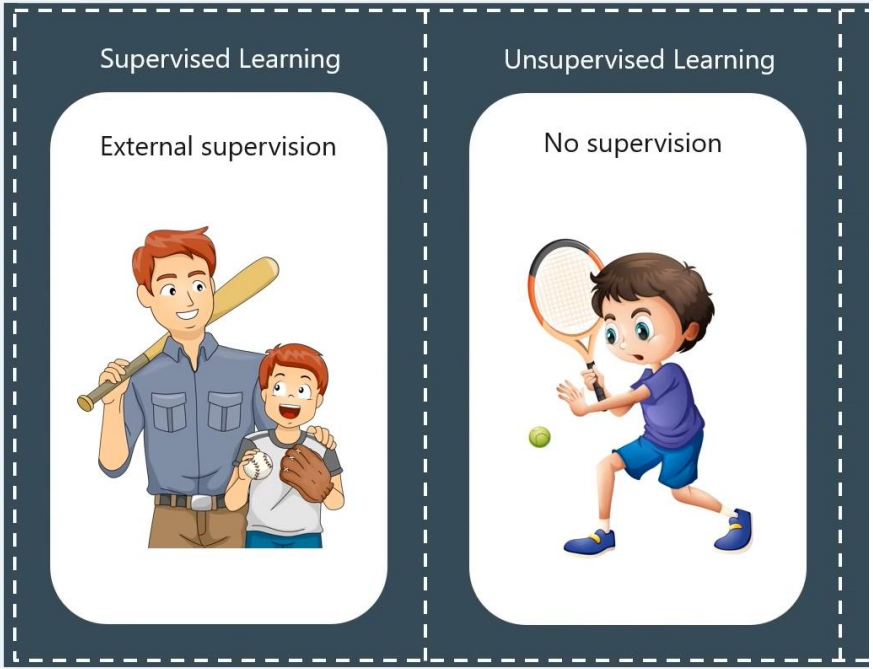
Type of data

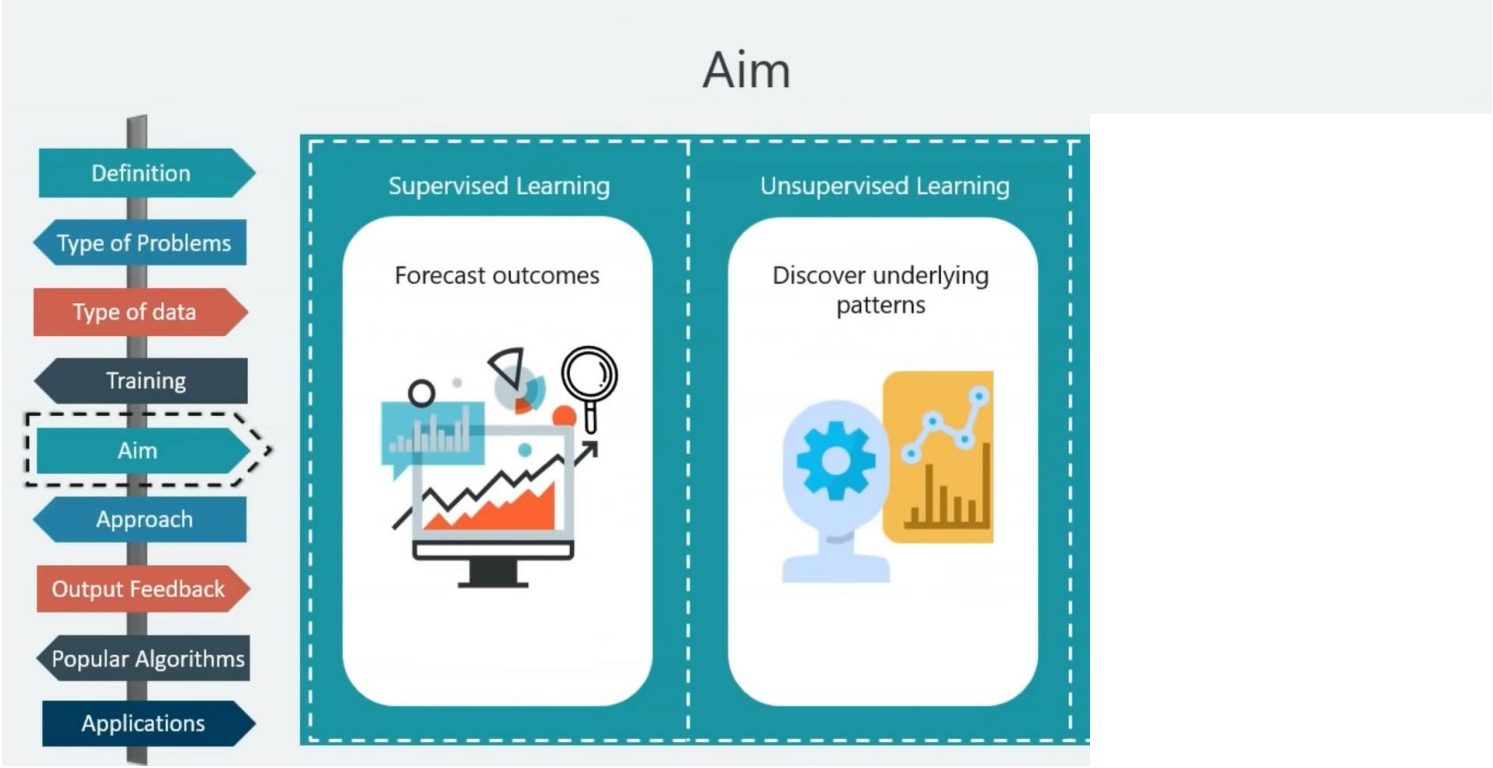

Definition
Type of Problems
Type of data
Training
Aim
Approach
Output Feedback
Popular Algorithms
Applications


Supervised Learning
Labelled Data
Cat
Algorithm
Dog
Cat
Unsupervised Learning
Unlabelled Data
Algorithm

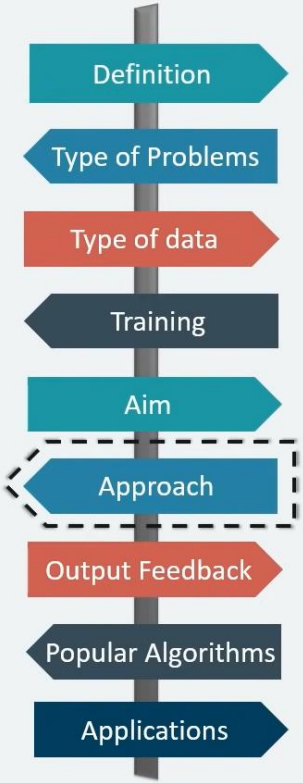
Training

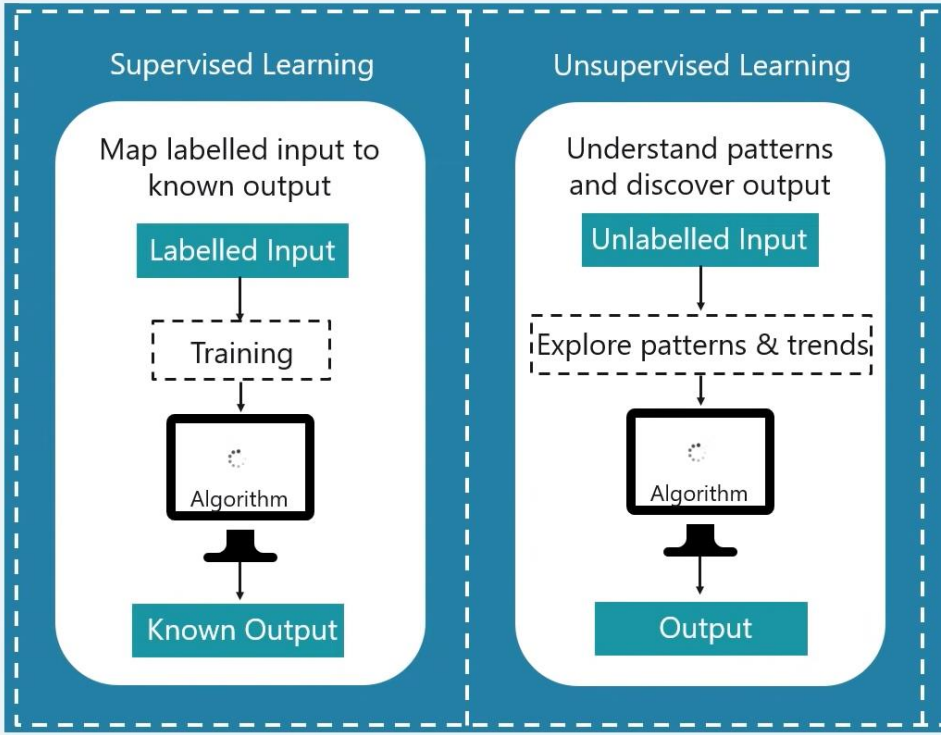

Definition
Type of Problems
Type of data
Training
Aim
Approach
Output Feedback
Popular Algorithms
Applications


Supervised Learning
External supervision
Unsupervised Learning
No supervision

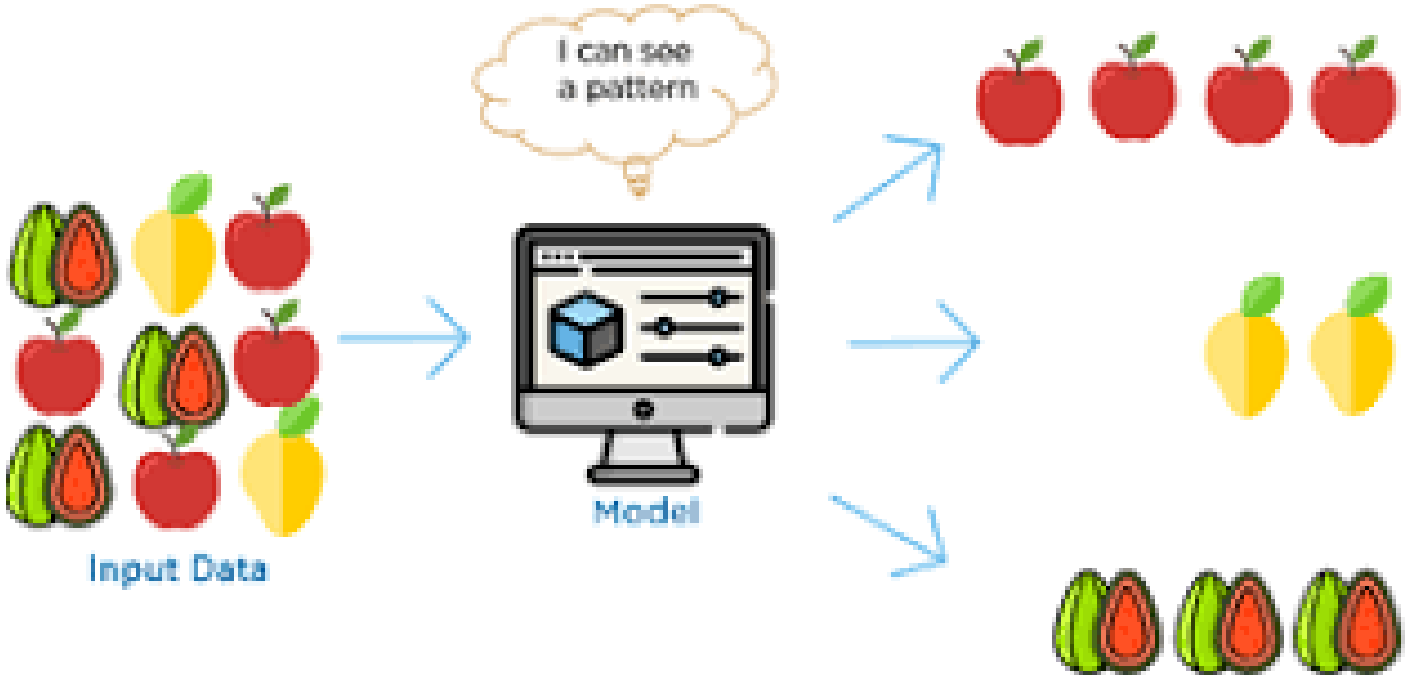

Aim
Definition
Type of Problems
Type of data
Training
Aim
Approach
Output Feedback
Popular Algorithms
Applications
Supervised Learning
Forecast outcomes
Unsupervised Learning
Discover underlying patterns

Approach

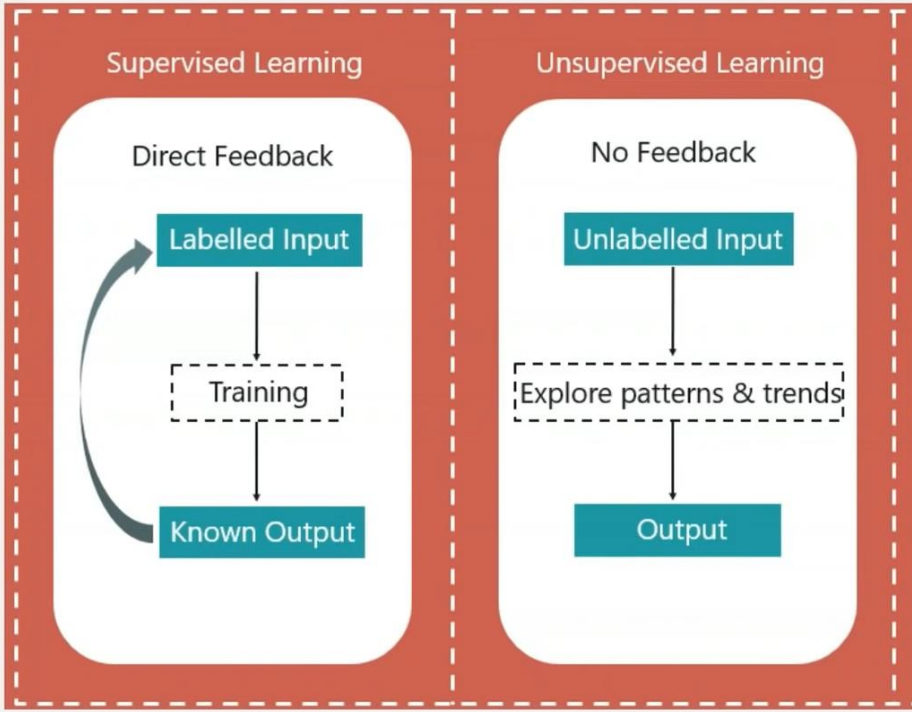

Definition
Type of Problems
Type of data
Training
Aim
Approach
Output Feedback
Popular Algorithms
Applications


Supervised Learning
Map labelled input to known output
Labelled Input
Training
Algorithm
Known Output
Unsupervised Learning
Understand patterns and discover output
Unlabelled Input
Explore patterns & trends
Algorithm
Output

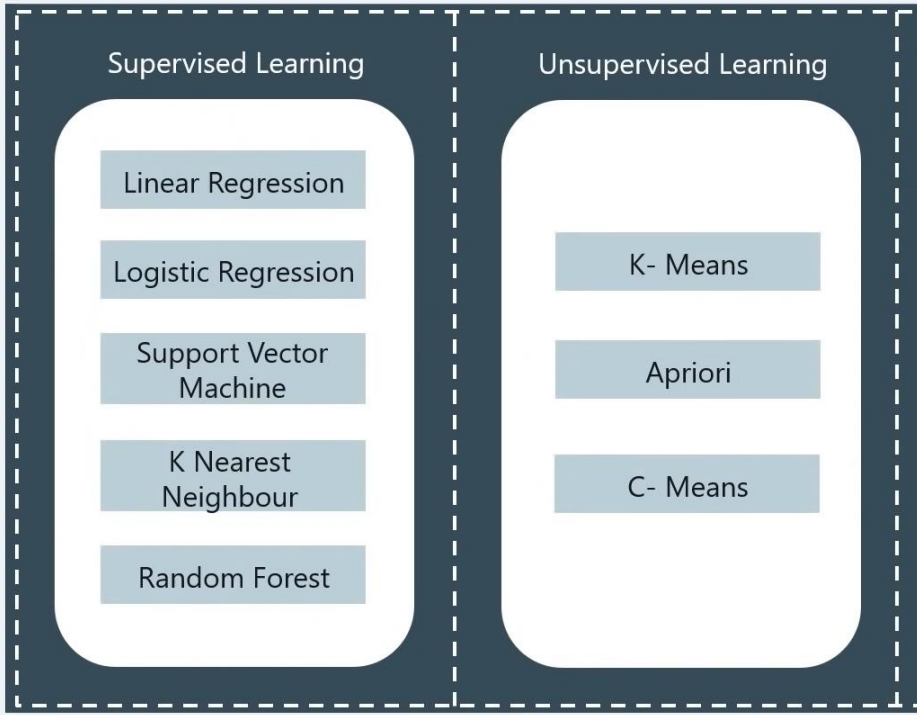
Unsupervised learning approach



Output Feedback



Popular Algorithms



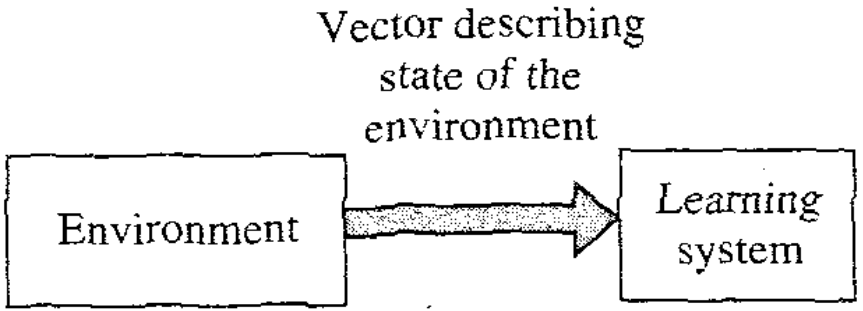
Applications

- Definition
- Type of Problems
- Type of data
- Training
- Aim
- Approach
- Output Feedback
- Popular Algorithms
- Applications

Unsupervised Learning :

- There is no external teacher or critic to oversee the learning process.
- The input vectors of similar types are grouped without the use of training data to specify how a member of each group looks or to which group a member belongs.

- Once the network becomes tuned to the statistical irregularities of the input data, it develops the ability to form internal representations for the encoding features of the input and thereby to create a new class automatically.



- **Advantages:**
  - These algorithms can be used for complicated tasks compared to the supervised ones because these algorithms work on the unlabeled dataset.
  - Unsupervised algorithms are preferable for various tasks as getting the unlabeled dataset is easier as compared to the labelled dataset.

- **Disadvantages:**
  - The output of an unsupervised algorithm can be less accurate as the dataset is not labelled, and algorithms are not trained with the exact output in prior.
  - Working with Unsupervised learning is more difficult as it works with the unlabelled dataset that does not map with the output.

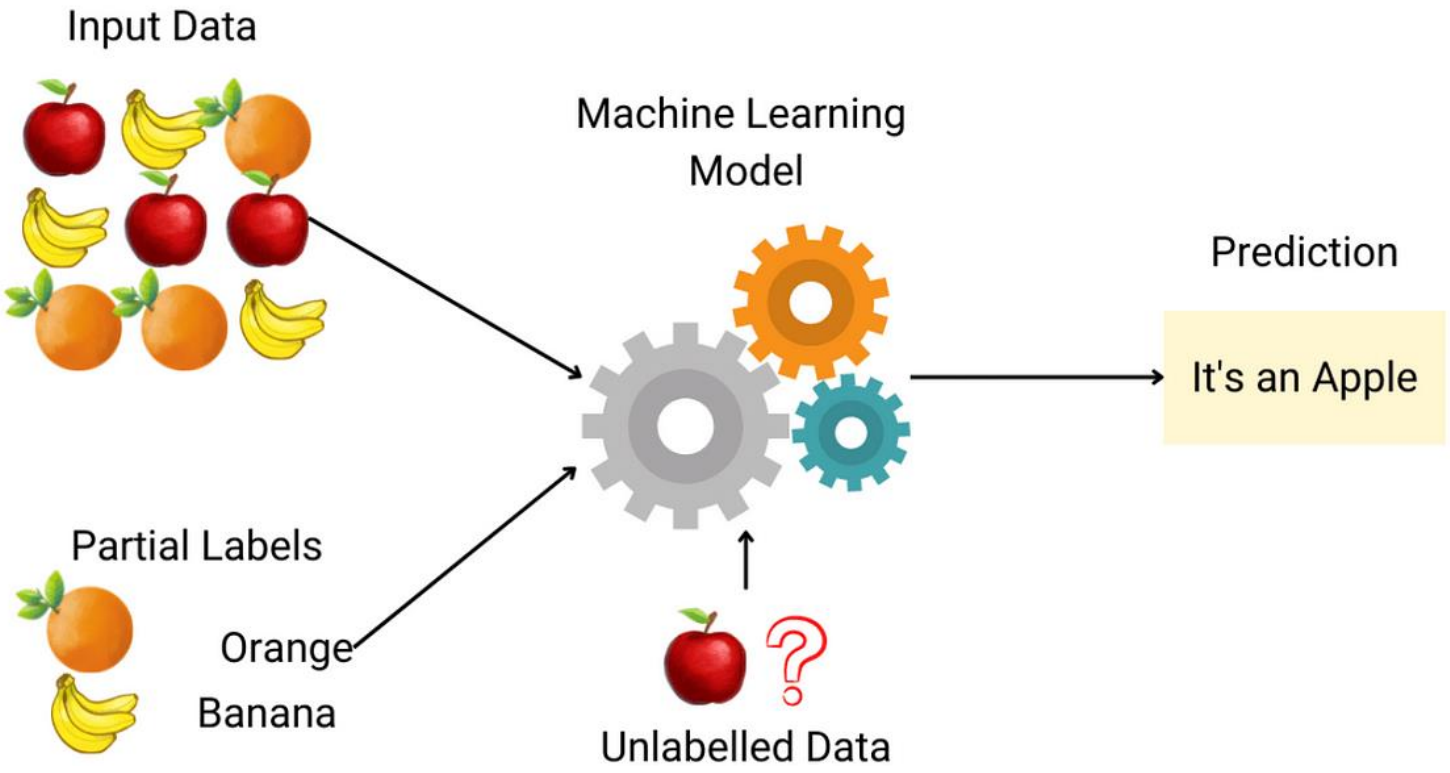
Applications of Unsupervised Learning

- Network Analysis: Plagiarism, copyright
- Recommendation Systems
- Anomaly Detection: fraudulent transactions
- Singular Value Decomposition

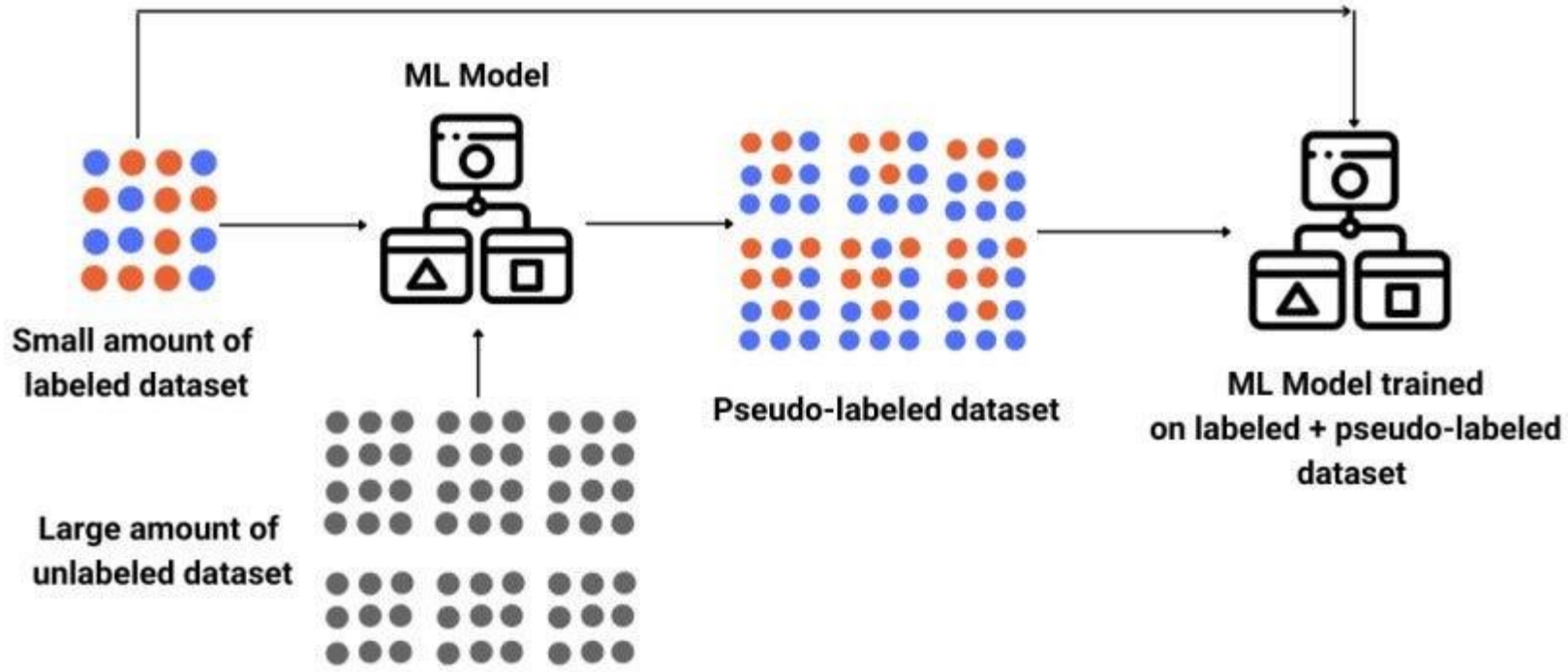
Comparison :

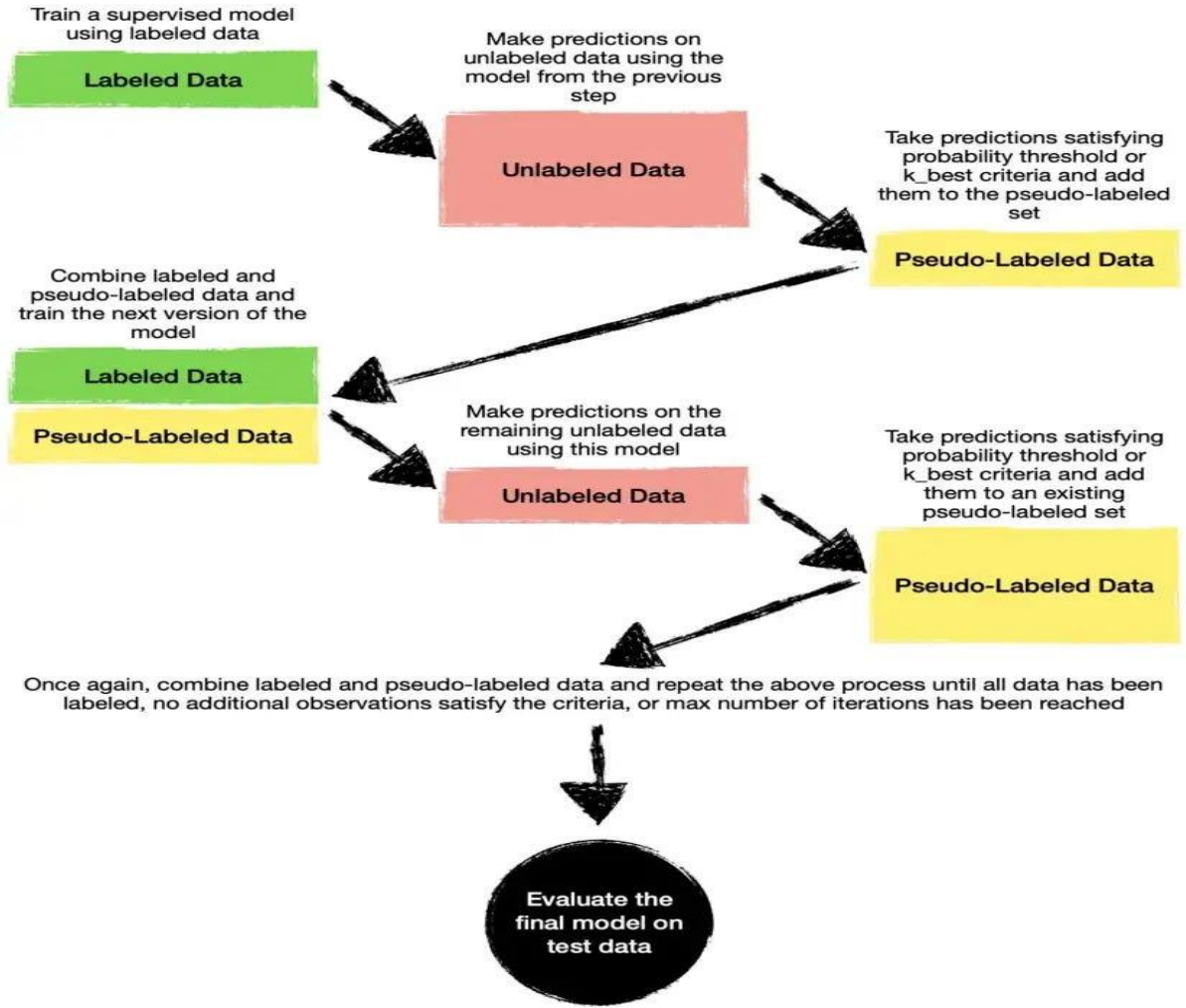
Parameter	Supervised learning	Unsupervised learning
Input data	Uses known input data	Uses unknown data as input
I/O data pair	Available	Output may not be paired with the input
Computational Complexity	High	Low
No. of classes	Known	Unknown (may form new classes)
Accuracy	Highest	Moderate
Categories	**Classification** and **Regression** problems	**Clustering** and **Associations** problems
Feed back from user	Provided	Not provided

Semi-Supervised Learning



Semi-supervised learning use-case




Train a supervised model using labeled data
Labeled Data
Make predictions on unlabeled data using the model from the previous step
Unlabeled Data
Take predictions satisfying probability threshold or k_best criteria and add them to the pseudo-labeled set
Pseudo-Labeled Data
Combine labeled and pseudo-labeled data and train the next version of the model
Labeled Data
Pseudo-Labeled Data
Make predictions on the remaining unlabeled data using this model
Unlabeled Data
Take predictions satisfying probability threshold or k_best criteria and add them to an existing pseudo-labeled set
Pseudo-Labeled Data
Once again, combine labeled and pseudo-labeled data and repeat the above process until all data has been labeled, no additional observations satisfy the criteria, or max number of iterations has been reached
Evaluate the final model on test data

Advantages

- Simple and easy to understand algorithm
- Highly efficient
- It is used to solve drawbacks of supervised and unsupervised learning algorithm

Disadvantages

- Iteration results may not be stable
- Not suitable for network level data
- Accuracy is low

Applications

- Speech Analysis
- Web content classification
- Protein sequence classification
- Text document

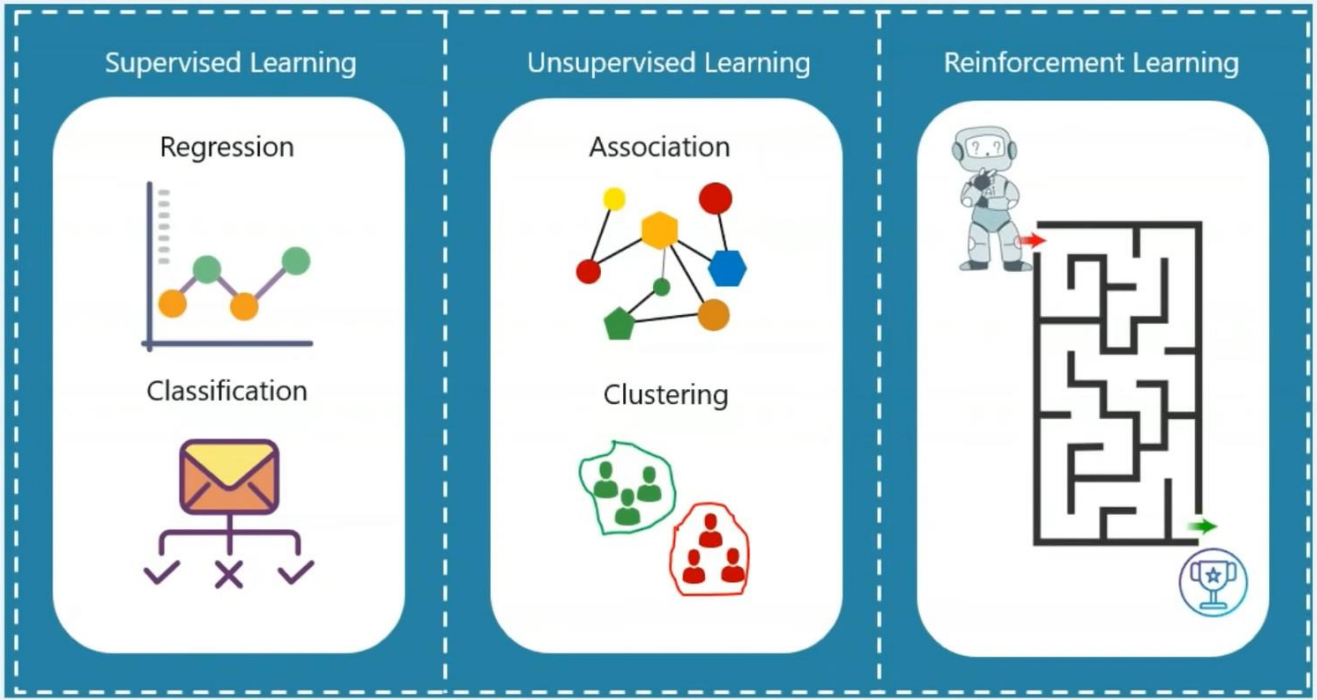
- Definition
- Type of Problems
- Type of data
- Training
- Aim
- Approach
- Output Feedback
- Popular Algorithms
- Applications

Supervised learning is a method in which we teach the machine using labelled data

In unsupervised learning the machine is trained on unlabelled data without any guidance

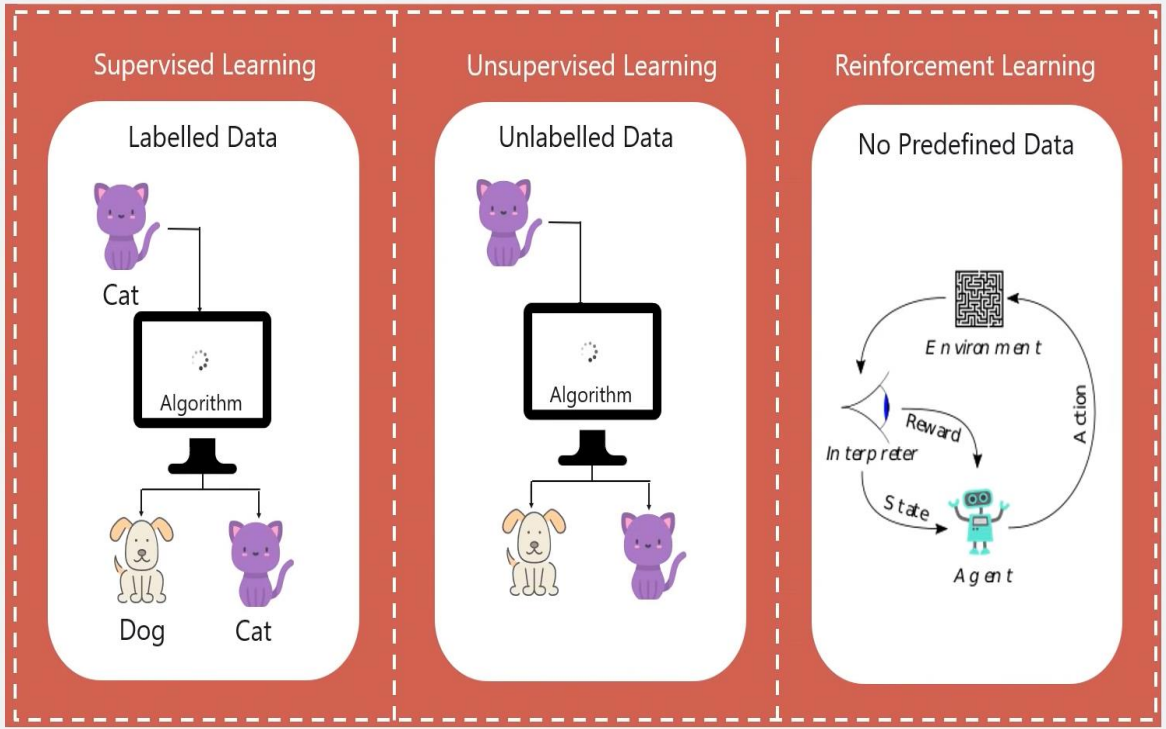
In Reinforcement learning an agent interacts with its environment by producing actions & discovers errors or rewards

Problem Type



Type of data


Definition
Type of Problems
Type of data
Training
Aim
Approach
Output Feedback
Popular Algorithms
Applications


Supervised Learning
Labelled Data
Cat
Algorithm
Dog
Cat
Unsupervised Learning
Unlabelled Data
Algorithm
Reinforcement Learning
No Predefined Data
Environment
Reward
Interpreter
Action
State
Agent

Training

- Definition
- Type of Problems
- Type of data
- Training
- Aim
- Approach
- Output Feedback
- Popular Algorithms
- Applications

Supervised Learning

External supervision

Unsupervised Learning

No supervision

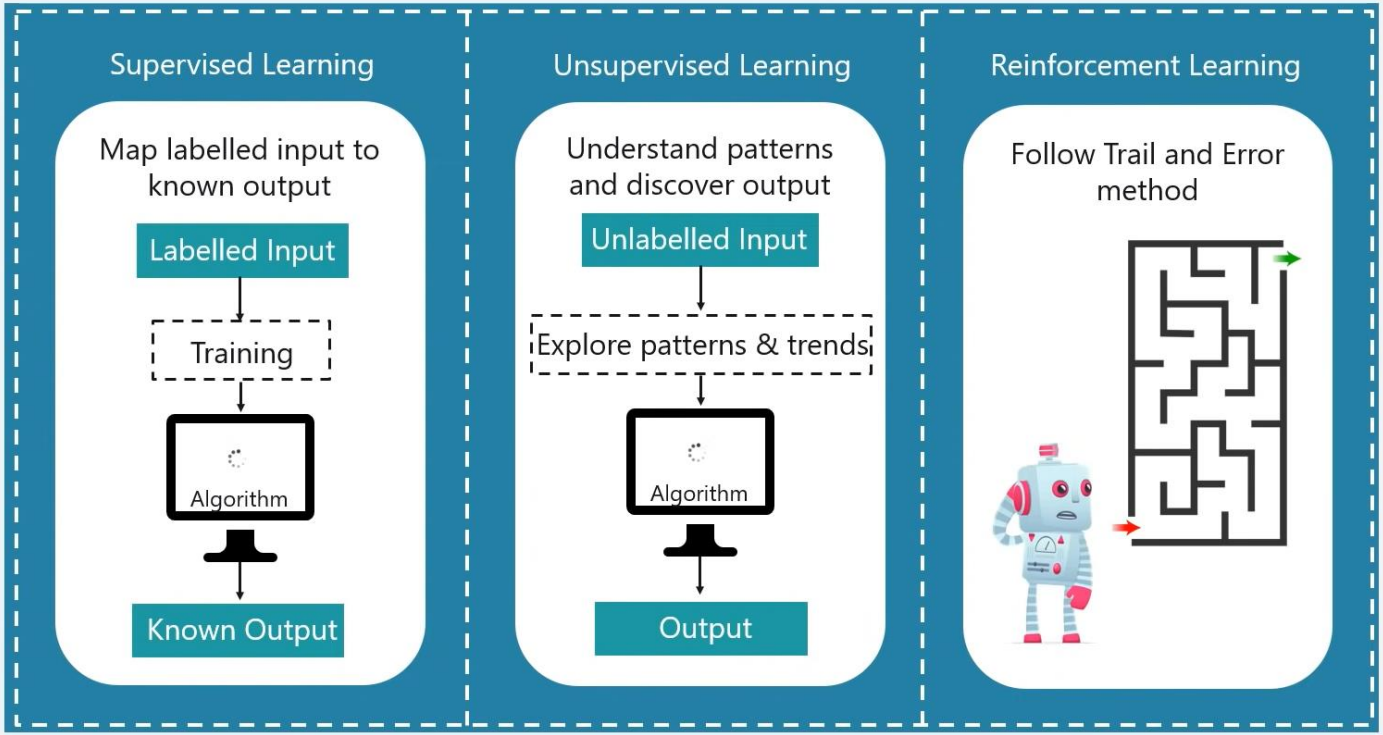
Reinforcement Learning

No supervision


Aim
Definition
Type of Problems
Type of data
Training
Aim
Approach
Output Feedback
Popular Algorithms
Applications
Supervised Learning
Forecast outcomes
Unsupervised Learning
Discover underlying patterns
Reinforcement Learning
Learn series of action

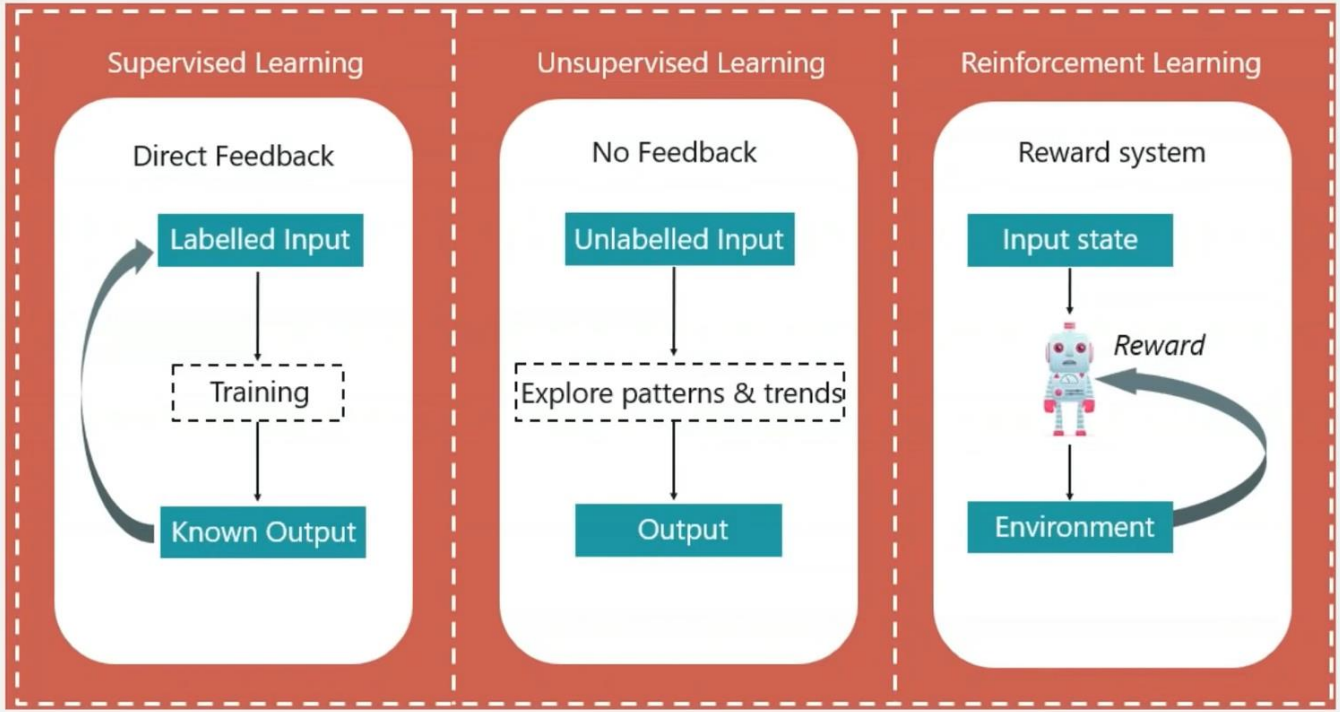
Approach

- Definition
- Type of Problems
- Type of data
- Training
- Aim
- Approach
- Output Feedback
- Popular Algorithms
- Applications

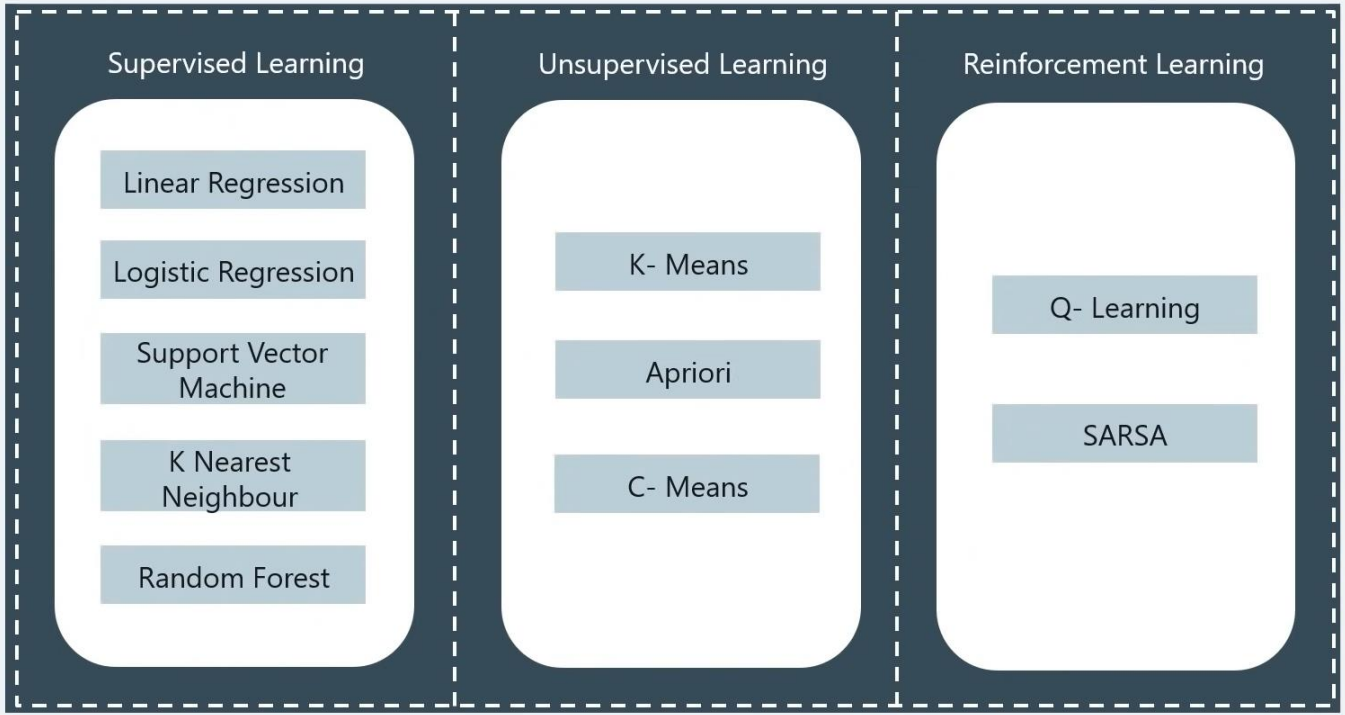


Output Feedback

- Definition
- Type of Problems
- Type of data
- Training
- Aim
- Approach
- Output Feedback
- Popular Algorithms
- Applications



Popular Algorithms



Applications

- Definition
- Type of Problems
- Type of data
- Training
- Aim
- Approach
- Output Feedback
- Popular Algorithms
- Applications

Supervised Learning

Risk Evaluation

Forecast Sales

Unsupervised Learning

Recommendation Systems

Anomaly Detection

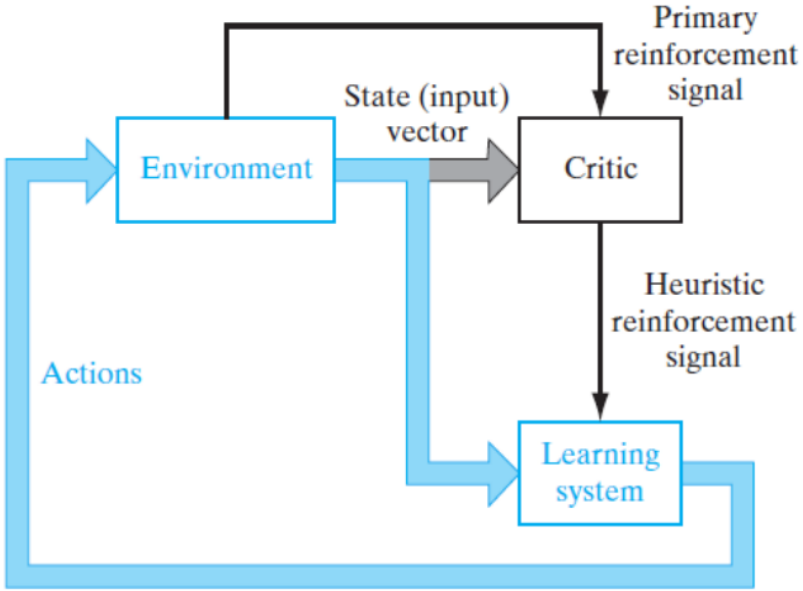
Reinforcement Learning

Self driving cars

Gaming

Reinforcement Learning :

- It is similar to supervised learning, but instead of correct target output values, less information may be available.



Reinforcement Learning . . . :

- It is difficult to perform for two reasons,

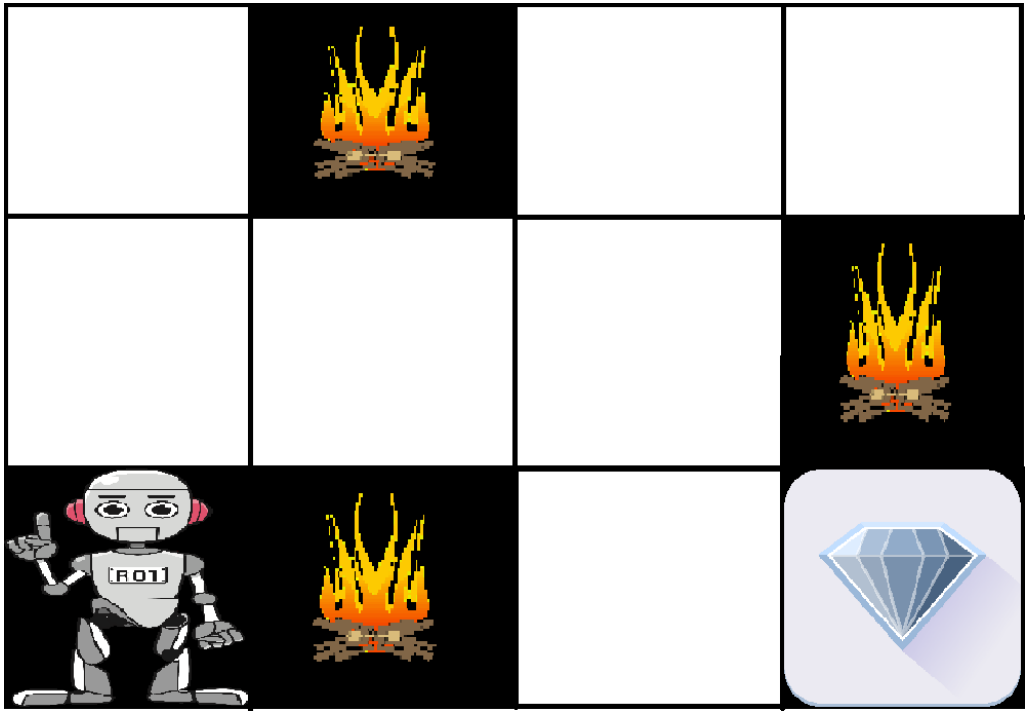
1. There is ***no teacher*** to provide a desired response at each step of learning process.

2. The learning system must be able to ***assign credit and blame*** individually to each action in the sequence of time steps that lead to the final outcome, while the primary reinforcement may only evaluate the outcome.

Reinforcement Learning . . . :

- The name “reinforcement learning” originates from behavioral psychology, but it could just as well be called “interactive learning.”
- Reinforcement learning is often used to teach agents, such as robots, to learn a given task. The agent learns by taking *actions* in the environment and receiving *observations* from this environment

Reinforcement Learning



Reinforcement Learning

- **Positive Reinforcement Learning:** Positive reinforcement learning specifies increasing the tendency that the required behaviour would occur again by adding something. It enhances the strength of the behaviour of the agent and positively impacts it.
- **Negative Reinforcement Learning:** Negative reinforcement learning works exactly opposite to the positive RL. It increases the tendency that the specific behaviour would occur again by avoiding the negative condition.

	REINFORCEMENT	PUNISHMENT
POSITIVE Add stimulus	Add pleasant stimulus to **increase** the behaviour	Add unpleasant stimulus to **decrease** the behaviour
NEGATIVE Remove stimulus	Remove unpleasant stimulus to **increase** the behaviour	Remove pleasant stimulus to decrease the behaviour

Reinforcement Learning: Real-world Use cases

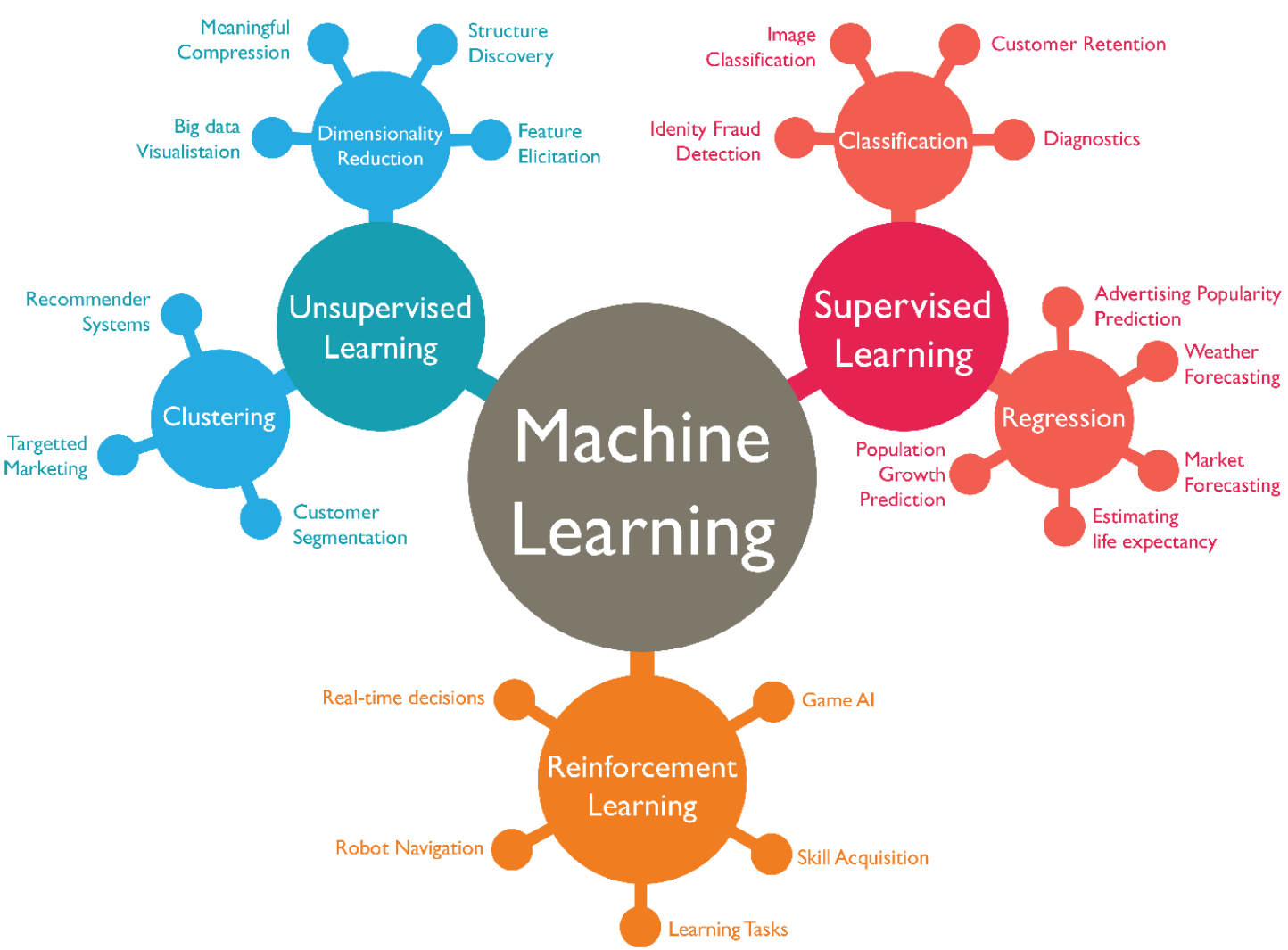
- Video Games
- Resource Management
- Robotics
- Text Mining

Advantages

- It helps in solving complex real-world problems which are difficult to be solved by general techniques.
- The learning model of RL is similar to the learning of human beings; hence most accurate results can be found.
- Helps in achieving long term results.

Disadvantages

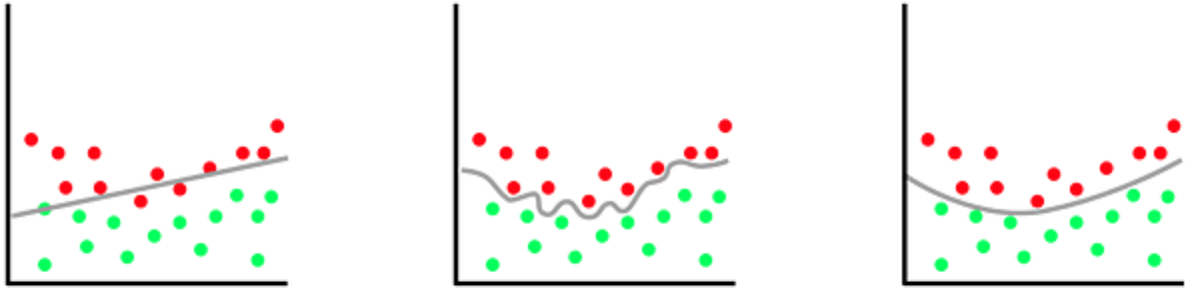
- RL algorithms are not preferred for simple problems.
- RL algorithms require huge data and computations.
- Too much reinforcement learning can lead to an overload of states which can weaken the results.


Machine Learning
Unsupervised Learning
Meaningful Compression
Structure Discovery
Dimensionality Reduction
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Image Classification
Customer Retention
Classification
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

Issues in Machine Learning :

- Inadequate training data.
- Poor quality of data (noisy, unclean).
- Non-representative training data.
- Overfitting and Underfitting

Generalization



Issues in Machine Learning :

- Monitoring and Maintenance
  - Different results – different actions – data change
  - Code editing
- Data drift (bad recommendation)
- Machine learning is a complex process
  - Algorithm selection
  - Hyperparameter tuning
- Slow implementation
  - Resource intensive algorithms
  - Complex model architectures

- Lack of skilled resources
- Imperfections in the algorithm when data grows
  - Scalability challenges
  - Increased sensitivity

How to Overcome Issues in Machine Learning :

1. ***Enhance Data Quality:*** Rigorous preprocessing, bias detection, and additional data collection can improve data quality and mitigate biases.
2. ***Address Underfitting:*** Employ more complex models, increase training iterations, and include relevant features to overcome underfitting issues.
3. ***Combat Overfitting***: Use model simplification techniques, apply regularization, and validate with cross-validation to counter overfitting challenges.
4. ***Manage Complexity:*** Stay updated with evolving techniques, continuously learn, and adopt best practices to navigate the complexities of the field.

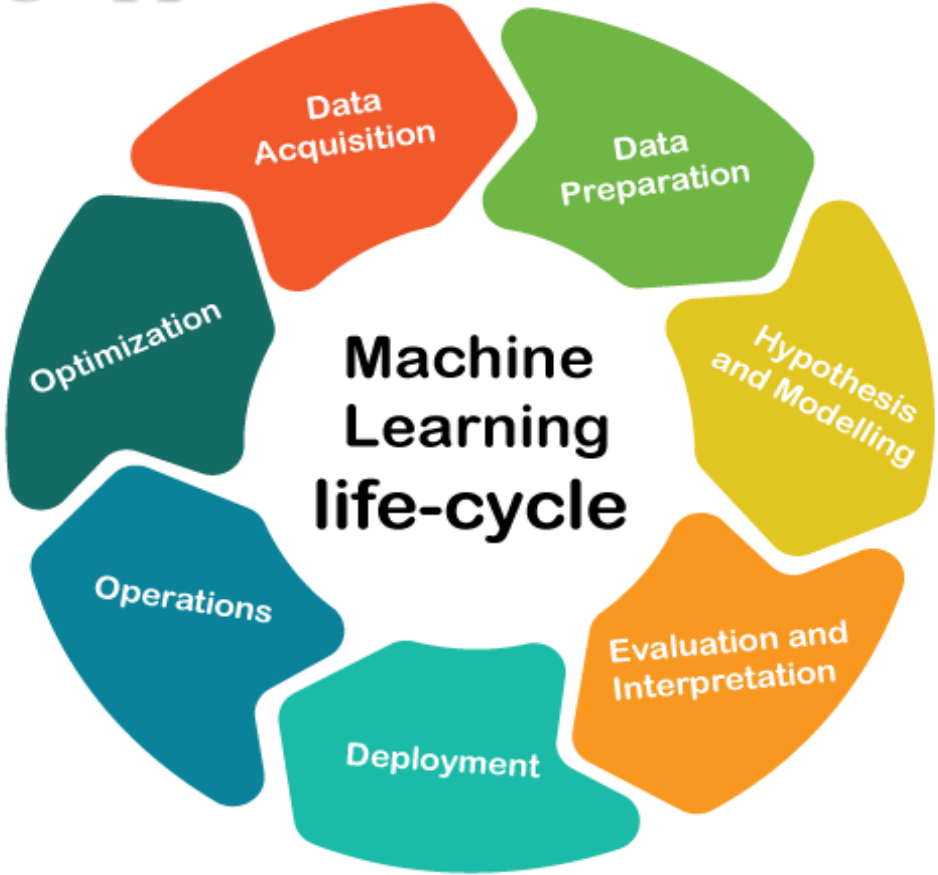
How to Overcome Issues in Machine Learning :

5. ***Tackle Limited Data:*** Implement data augmentation, leverage transfer learning, and consider synthetic data generation to mitigate limitations posed by insufficient data.
6. ***Optimize Implementation:*** Opt for efficient coding practices, utilize parallel processing, and leverage hardware acceleration to overcome slow implementation.
7. ***Adapt to Data Growth:*** Adjust algorithms to handle larger datasets and choose models designed to scale effectively as data expands.

Applications of Machine Learning :

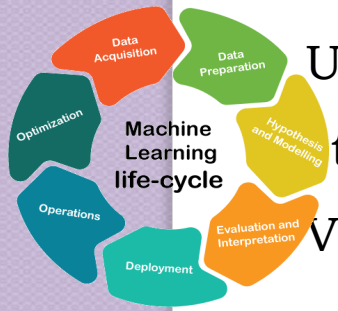
- Social Media Features
- Product Recommendations
- Image Recognition
- Sentiment Analysis
- Automating Employee Access Control
- Marine Wildlife Preservation
- Regulating Healthcare Efficiency and Medical Services
- Language Translation
- Banking Domain

Steps involved in developing a Machine Learning Application :



Steps involved in developing a Machine Learning Application :

- ***Data Acquisition :***
  - The quantity & quality of your data dictate how accurate our model is.
  - The outcome of this step is generally a representation of data which we will use for training.



Using pre-collected data, by way of datasets from Kaggle, UCI, tc.

Visualize data to help detect relevant relationships between variables or class imbalances (bias alert!), or perform other exploratory analysis.

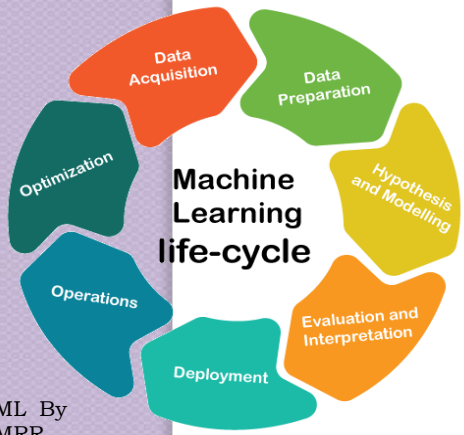
Steps involved in developing a Machine Learning Application :

- ***Data Preparation :***
  - Wrangle data and prepare it for training.
  - Clean the data (*remove duplicates, correct errors, deal with missing values, normalization, data type conversions, etc.*)
  - Randomize data, which erases the effects of the particular order in which we collected and/or otherwise prepared our data
  - Split into training and evaluation sets

Steps involved in developing a Machine Learning Application :

Choose a Model

Different algorithms are for different tasks; choose the right one



Train the Model

-The goal of training is to answer a question or make a prediction correctly as often as possible.

-Linear regression example: algorithm would need to learn values for m (or W) and b (x is input, y is output).

-Each iteration of process is a training step.

Evaluate the Model

-Uses some metric or combination of metrics to "measure" objective performance of model

-Test the model against previously unseen data

-This unseen data is meant to be somewhat representative of model performance in the real world, helps tune the model (as opposed to test data, which does not)

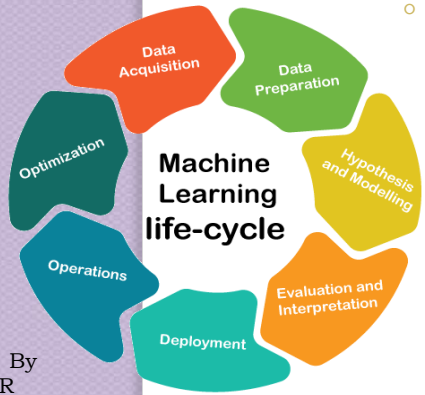
-Good train/eval split? 80/20, 70/30, or similar, depending on domain, data availability, dataset particulars, etc.

Steps involved in developing a Machine Learning Application :

- ***Parameter Tuning***
  - This step refers to hyperparameter tuning, which is an "artform" as opposed to a science
  - Tune model parameters for improved performance
  - Simple model hyperparameters: number of training steps, learning rate, initialization values and distribution, etc.
- ***Make Predictions***
  - Using further (test set) data which have, until this point, been withheld from the model (and for which class labels are known), are used to test the model; a better approximation of how the model will perform in the real world

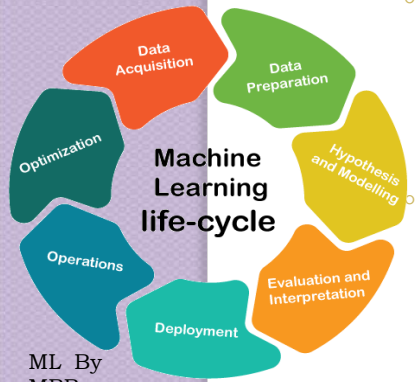
Steps involved in developing a Machine Learning Application :

- ***Deployment :***
  - The last step of machine learning life cycle is deployment, where the model is deployed in the real-world system.
  - If the model is producing an accurate result as per the requirements with acceptable speed, then the model is deployed in the real system.
  - But before deploying the project, it is necessary to check whether it is improving its performance using available data or not.



Steps involved in developing a Machine Learning Application :

- ***Monitoring and Maintenance :***
  - After deploying the model to production, it needs to be constantly monitored, for model metrics, hardware and software performance, and customer satisfaction, and improve the system.
  - The monitoring is done completely automatically, and the professionals are notified about the anomalies, reduced model and system performance, and bad customer reviews.
  - After a reduced performance alert is received, it needs to assess the issues and try to train the model on new data or make changes to model architectures. It is a continuous process.
  - In rare cases, the complete machine learning life cycle is revamped to improve the data processing and model training techniques, update new software and hardware, and introduce a new framework for continuous integration.



Generalization :

- **Generalization:** Generalization is a term that usually refers to a Machine Learning models ability to perform well on the new unseen data. After being trained on a training set, a model can digest new data and can be able to make accurate predictions.
- The main success of the model is the ability of the model to generalize well. If the model has been trained too well on the training data, it will be difficult for the model to generalize.
- Factors that influence generalization:
  - the size of the training set.
  - the architecture of the NN.
  - the complexity of the problem at hand.
- Over-fitting (overtraining): when the NN learns too many I/O examples it may end up memorizing the training data.

Overfitting:

- Overfitting happens when a model learns the detail and noise in the training data very well.
- It negatively affects the performance of the model.
- The main reason behind overfitting is using non-linear methods used in machine learning algorithms as they build non-realistic data models.
- Overfitting can be overcome by using linear and parametric algorithms in the machine learning models.

- **Methods to reduce overfitting:**
  - Increase training data in a dataset.
  - Reduce model complexity by simplifying the model by selecting one with fewer parameters
  - Early stopping during the training phase
  - Reduce the noise
  - Reduce the number of attributes in training data.
  - Constraining the model.

Underfitting:

- Underfitting occurs when the model is too simple to understand the base structure of the data. This generally happens when the data is limited into the data set, and we try to build a linear model with non-linear data.
- In such scenarios, the complexity of the model destroys, and rules of the machine learning model become too easy to be applied on this data set, and the model starts doing wrong predictions as well.
- **Methods to reduce Underfitting:**
  - Increase model complexity
  - Remove noise from the data
  - Train on increased and better features
  - Reduce the constraints
  - Increase the number of epochs to get better results.

Issues in machine learning

- How much training data is sufficient?
- Which algorithms perform best for which types of problems and representations?
- What algorithms exist for learning general target functions from specific training examples?
- In what settings will particular algorithms converge to the desired function, given sufficient training data?
- When and how can prior knowledge held by the learner guide the process of generalizing from examples?

- Can prior knowledge be helpful even when it is only approximately correct?
- What specific functions should the system attempt to learn?
- Can this process itself be automated?
- How can the learner automatically alter its representation to improve its ability to represent and learn the target function?