- **Vision of Computer Department**
- Develop globally competent Computer Engineering Professionals who achieve excellence in an environment conducive for technical knowledge, skills, moral and ethical values with a focus to serve the society.

- **Mission of Computer Department** Supervised learning
- M1 : To provide state of the art infrastructure and facilities for imparting quality education and computer engineering skills for societal benefit.
- M2 : Adopt industry oriented curriculum with an exposure to technologies for building systems & application in computer engineering.
- M3 : To provide quality technical professional as per the industry and societal needs, encourage entrepreneurship, nurture innovation and life skills in consonance with latest interdisciplinary trends.

Introduction to Machine Learning (Course Code: 4350702)

Course Outcomes

a) Describe basic concept of machine learning and its applications

b) Practice Numpy, Pandas, Matplotlib, sklearn library's inbuilt function required to solve machine learning problems

c) Use Pandas library for data preprocessing

d) Apply supervised learning algorithms based on dataset characteristics

e) Apply unsupervised learning algorithms based on dataset characteristics

Unit – I
Introduction to machine learning

Content

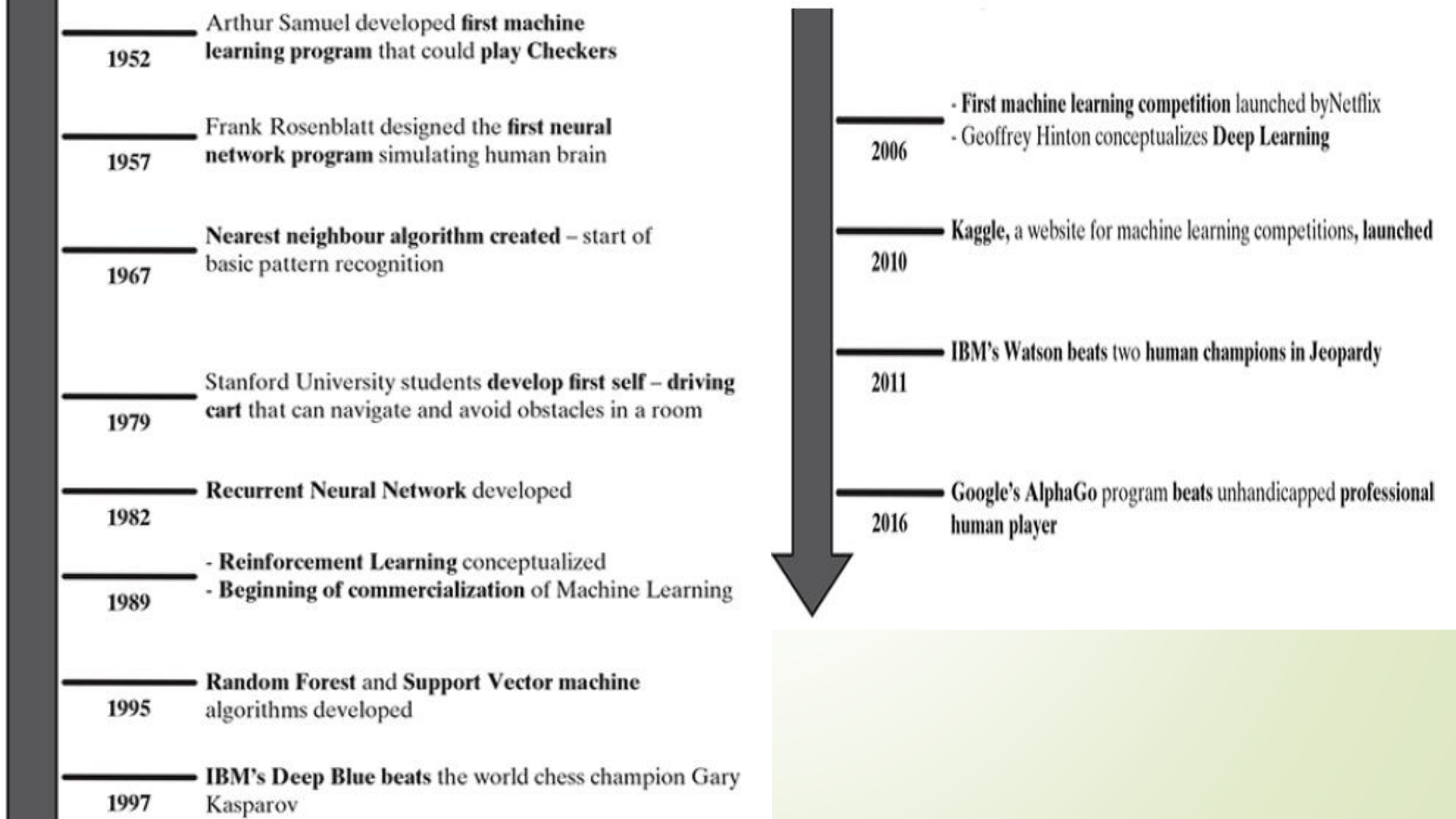
1. Overview of Human Learning and Machine Learning
2. Types of Machine Learning
    1) Supervised Machine Learning
    2) Unsupervised Machine Learning
    3) Reinforcement Learning.
3. Applications of Machine Learning
4. Tools and Technology for Machine Learning

Inception of Machine Learning

- The computer program IBM's Deep Blue and it defeated world chess champion, Gary Kasparov before 20 years.
- That was the time, when people gave serious attention to a fast evolving field in computer science or more specifically artificial intelligence – i.e. machine learning (ML).

Applications of ML in various sphere of life

- ML can recommend toys to toddlers much in the same way as it can suggest a technology book to a geek or a rich title in literature to a writer.
- It predicts the future market to help amateur traders compete with seasoned stock traders.
- It helps an oncologist find whether a tumour is malignant or benign.
- It helps in optimizing energy consumption thus helping the cause of Green Earth.


1952
Arthur Samuel developed first machine learning program that could play Checkers
1957
Frank Rosenblatt designed the first neural network program simulating human brain
1967
Nearest neighbour algorithm created – start of basic pattern recognition
1979
Stanford University students develop first self – driving cart that can navigate and avoid obstacles in a room
1982
Recurrent Neural Network developed
1989
- Reinforcement Learning conceptualized
- Beginning of commercialization of Machine Learning
1995
Random Forest and Support Vector machine algorithms developed
1997
IBM's Deep Blue beats the world chess champion Gary Kasparov
2006
- First machine learning competition launched byNetflix
- Geoffrey Hinton conceptualizes Deep Learning
2010
Kaggle, a website for machine learning competitions, launched
2011
IBM's Watson beats two human champions in Jeopardy
2016
Google's AlphaGo program beats unhandicapped professional human player

Learning

- Learning is typically referred to as the process of gaining information through observation
- In our daily life, we need to carry out multiple activities. It may be a task as simple as walking down the street or doing the homework. Or it may be some complex task like deciding the angle in which a rocket should be launched so that it can have a particular trajectory.
- To do a task in a proper way, we need to have prior information on one or more things related to the task.

Human Learning

- Human learning happens in one of the three ways –

1. either somebody who is an expert in the subject directly teaches us.
2. we build our own notion indirectly based on what we have learnt from the expert in the past
3. we do it ourselves, may be after multiple attempts, some being unsuccessful.

1. Learning under expert guidance

- An infant may inculcate certain traits and characteristics, learning straight from its guardians.
- He calls his hand, a ‘hand’, because that is the information he gets from his parents.
- The sky is ‘blue’ to him because that is what his parents have taught him.
- In school, he starts with basic familiarization of alphabets and digits.
- Then the baby learns how to form words from the alphabets and numbers from the digits.
- Slowly more complex learning happens in the form of sentences, paragraphs, complex mathematics, science, etc

Continue

- Then starts higher studies where the person learns about more complex, application-oriented skills.
- Then starts higher studies where the person learns about more complex, application-oriented skills.
- Then the person starts working as a professional in some field. The professional mentors, by virtue of the knowledge that they have gained through years of hands-on experience, help all new comers in the field to learn on-job.

2. Learning guided by knowledge gained from experts

- An essential part of learning also happens with the knowledge which has been imparted by teacher or mentor at some point of time in some other form/context.
- For example, a baby can group together all objects of same colour even if his parents have not specifically taught him to do so.
- He is able to do so because at some point of time or other his parents have told him which colour is blue, which is red, which is green, etc.

3. Learning by self

- In many situations, humans are left to learn on their own.
- A baby learning to walk through obstacles. He bumps on to obstacles and falls down multiple times till he learns that whenever there is an obstacle, he needs to cross over it.
- He faces the same challenge while learning to ride a cycle as a kid or drive a car as an adult.

WHAT IS MACHINE 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.’

- Note: The first step in any project is defining your problem. Even if the most powerful algorithm is used, the results will be meaningless if the wrong problem is solved.

How do machines learn?

The basic machine learning process can be divided into three

parts.

1. **Data Input:** Past data or information is utilized as a basis for future decision-making
2. **Abstraction:** The input data is represented in a broader way through the underlying algorithm
3. **Generalization:** The abstracted representation is generalized to form a framework for making decisions

Process of machine learning

Process of Human Learning Process

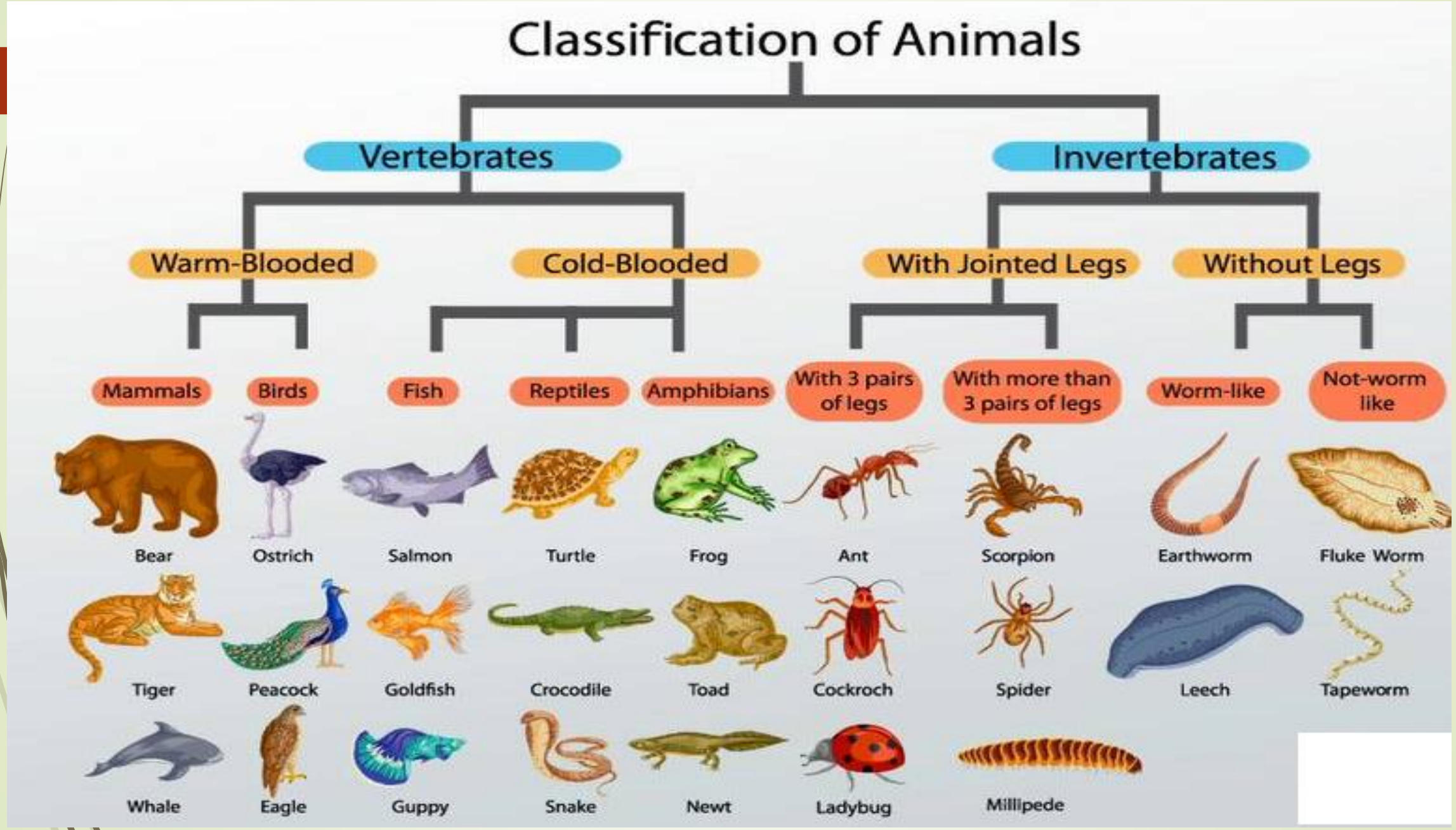
- It is a tendency of many students to try and memorize.
- The kinds of questions which are asked in the examination are pretty much simple and straightforward.
- The questions can be answered by simply writing the same things which have been memorized.
- As the scope gets broader and the questions asked in the examination gets more complex, the strategy of memorizing doesn't work well.
- Thus a better learning strategy needs to be adopted:

    1. to be able to deal with the vastness of the subject matter and the related issues in memorizing it

    2. to be able to answer questions where a direct answer has not been learnt

1. Input data in Process of Learning

- A broad pool of knowledge may consist of all living animals and their characteristics such as whether they live in land or water, whether they lay eggs, whether they have scales or fur or none, etc.
- It is a difficult task for any student to memorize the characteristics of all living animals
- It is better to form basic groups that all living animals belong to.


Classification of Animals
Vertebrates
Invertebrates
Warm-Blooded
Cold-Blooded
With Jointed Legs
Without Legs
Mammals
Birds
Fish
Reptiles
Amphibians
With 3 pairs of legs
With more than 3 pairs of legs
Worm-like
Not-worm like
Bear
Ostrich
Salmon
Turtle
Frog
Ant
Scorpion
Earthworm
Fluke Worm
Tiger
Peacock
Goldfish
Crocodile
Toad
Cockroch
Spider
Leech
Tapeworm
Whale
Eagle
Guppy
Snake
Newt
Ladybug
Millipede

Grouping (Concept Map)

1. Invertebrate: Do not have backbones and skeletons
2. Vertebrate
    1. Fishes: Always live in water and lay eggs
    2. Amphibians: Semi-aquatic i.e. may live in water or land; smooth skin; lay eggs
    3. Reptiles: Semi-aquatic like amphibians; scaly skin; lay eggs; cold-blooded
    4. Birds: Can fly; lay eggs; warm-blooded
    5. Mammals: Have hair or fur; have milk to feed their young; warm-blooded

This makes it easier to memorize as the scope now reduces to know the animal groups that the animals belong to.
In machine learning paradigm, the vast pool of knowledge is available from the data input

2. *Abstraction*

- During the machine learning process, knowledge is fed in the

form of input data. However, the data cannot be used in the original shape and form.

- The map, or a **model** is the summarized knowledge representation of the raw data.
- The model may be in any one of the following forms
  1. Computational blocks like if/else rules
  2. Mathematical equations
  3. Specific data structures like trees or graphs
  4. Logical groupings of similar observations

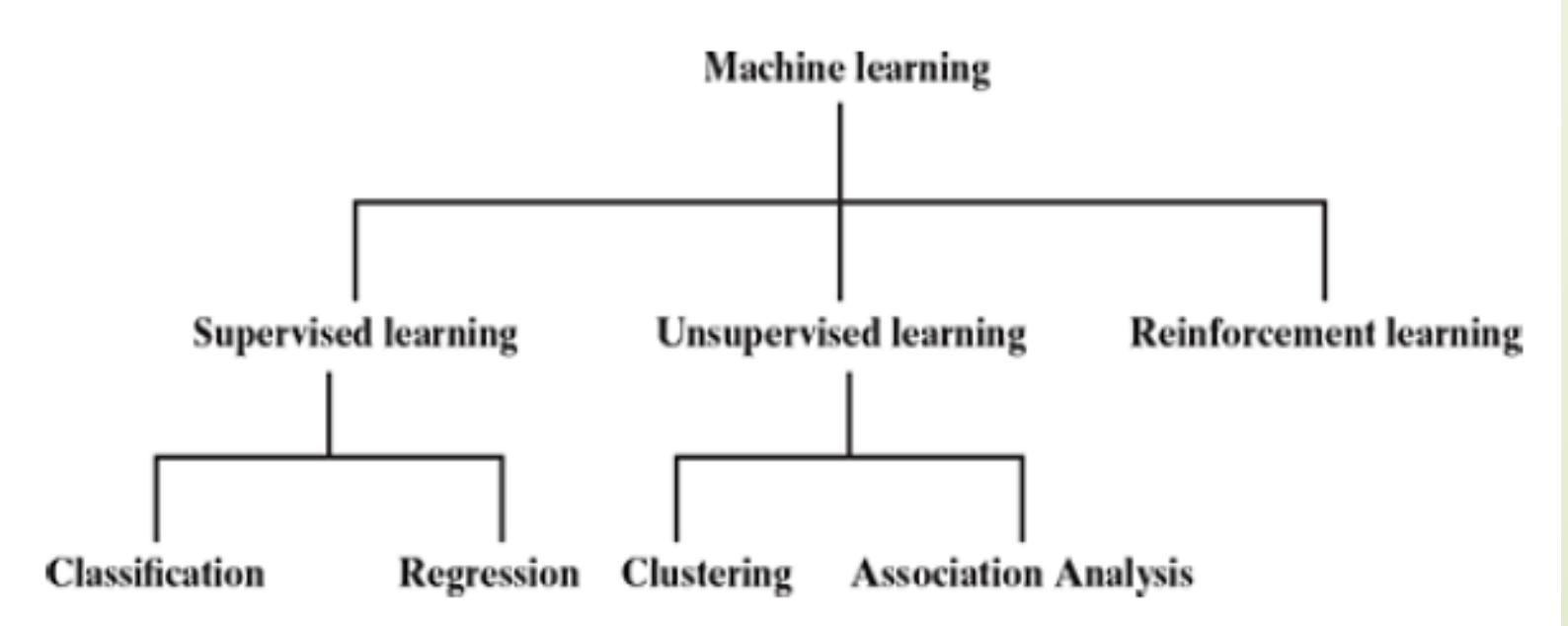
- The decision related to the choice of model is taken based on multiple aspects, some of which are listed below:
  1. The type of problem to be solved: Whether the problem is related to forecast or prediction, analysis of trend, understanding the different segments or groups of objects, etc.
  2. Nature of the input data: How exhaustive the input data is, whether the data has no values for many fields, the data types, etc.
  3. Domain of the problem: If the problem is in a business critical domain with high rate of data input and need for immediate inference, e.g. fraud detection problem in banking domain.

3. *Generalization*

- Genearalization is to tune up the abstracted knowledge to a form which can be used to take future decisions.

Types of ML

- Machine learning can be classified into three broad categories:
  1. Supervised learning – Also called predictive learning. A machine predicts the class of unknown objects based on prior class-related information of similar objects.
  2. Unsupervised learning – Also called descriptive learning. A machine finds patterns in unknown objects by grouping similar objects together.
  3. Reinforcement learning – A machine learns to act on its own to achieve the given goals.



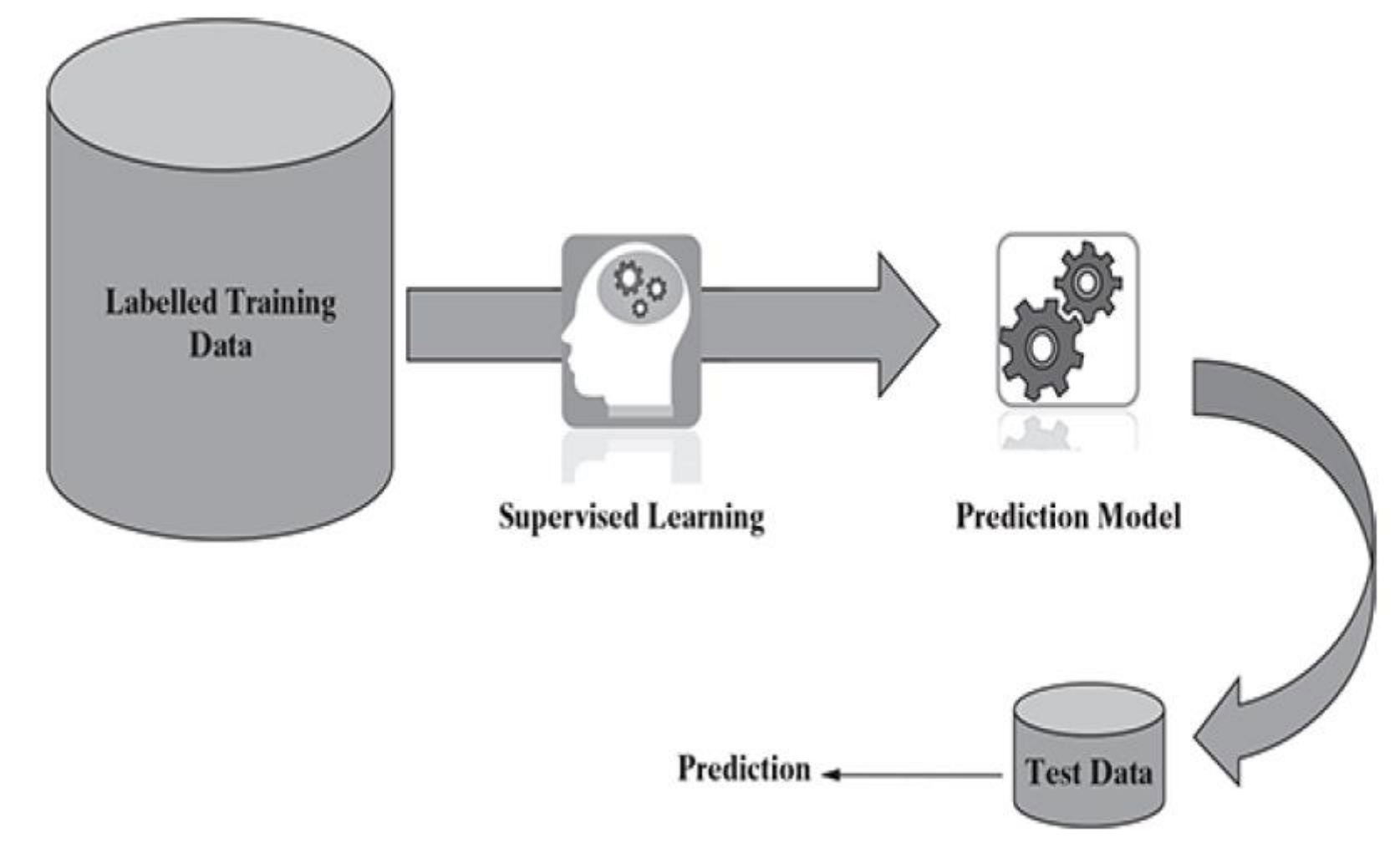
Supervised learning

- Supervised learning: Machine learns from past information which is an experience in terms of ML.
- E.g. A machine is getting images of different objects as input and the task is to segregate the images by either shape or colour of the object.
- how can the machine know what is round shape, or triangular shape or whether it is blue or green in colour?
- A machine needs the basic information to be provided to it.
- This basic input, or the experience in the paradigm of machine learning, is given in the form of **training data** .

Training Data

- Training data is the past information on a specific task.
- i.e. on different aspects or features on a number of images, along with a tag on whether the image is round or triangular, or blue or green in colour.
- The tag is called ' **label'** and we say that the training data is labelled in case of supervised learning.
- Based on the training data, the machine builds a predictive model that can be used on test data to assign a label for each record in the test data.

Supervised Learning



Examples of supervised learning

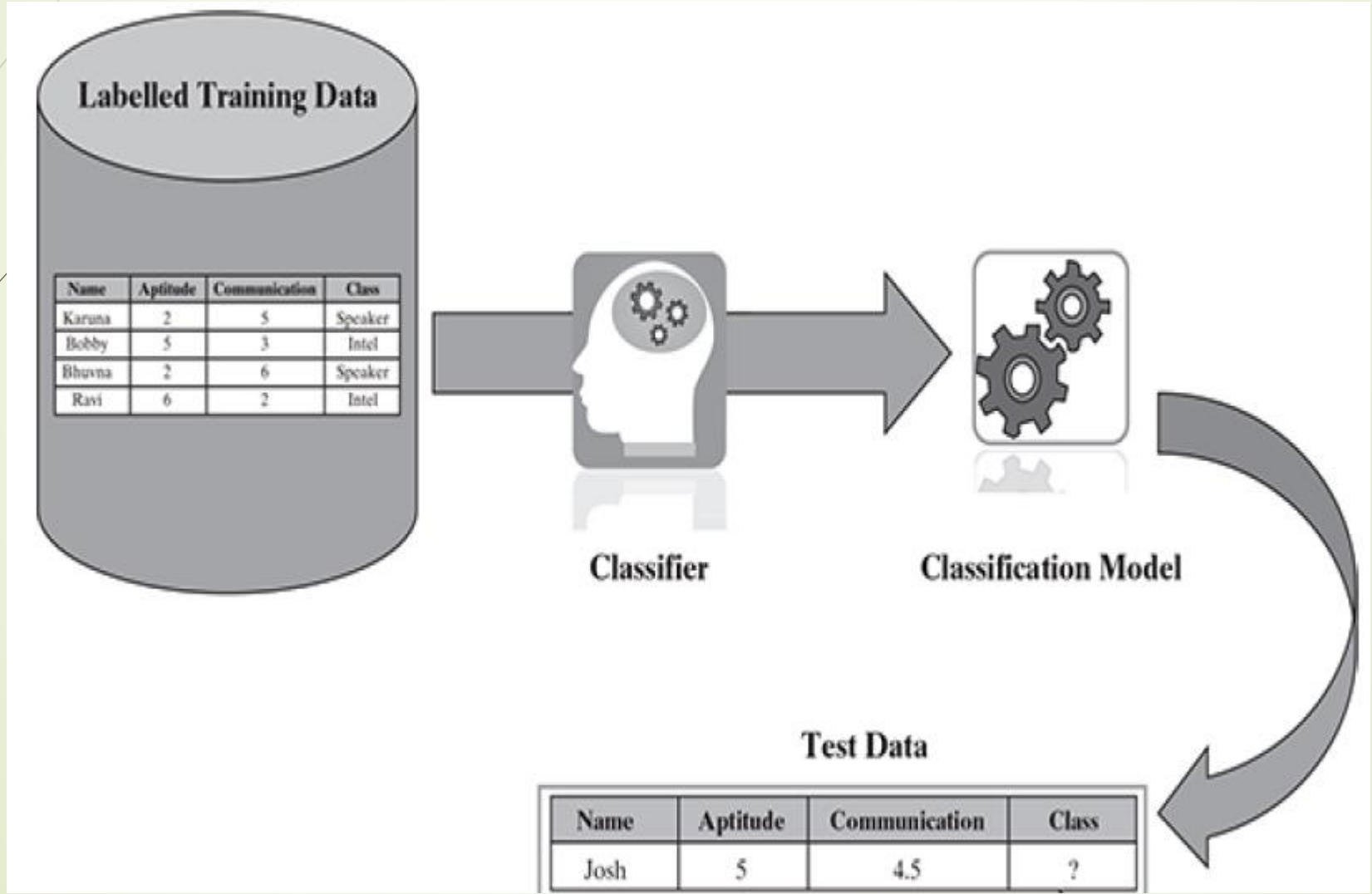
- Predicting the results of a game
- Predicting whether a tumour is malignant or benign
- Predicting the price of domains like real estate, stocks, etc.
- Classifying texts such as classifying a set of emails as spam or non-spam

- Note: When we are trying to predict a categorical or nominal variable, the problem is known as a **classification** problem.
  - Whereas when we are trying to predict a real-valued variable, the problem falls under the category of **regression.**

Classification

- How to segregate the images of objects based on the shape .
  - If the image is of a round object, it is put under one category, while if the image is of a triangular object, it is put under another category.
- The image is placed in a category based on past data called **training data.**
- The category in which the machine should put an image of unknown category, also called a **test data.**
- The whole problem revolves around assigning a label or category or class to a test data based on the label or category or class information, thus its called classification problem.

Classification



Popular machine learning algorithms

- Naïve Bayes
- Decision tree
- k-Nearest Neighbour

Typical classification problems include:

- Image classification
- Prediction of disease
- Win–loss prediction of games
- Prediction of natural calamity like earthquake, flood, etc.
- Recognition of handwriting

Regression

- In linear regression, the objective is to predict numerical features like real estate or stock price, temperature, marks in an examination, sales revenue, etc.
- In case of linear regression, a straight line relationship is ‘fitted’ between the predictor variables and the target variables

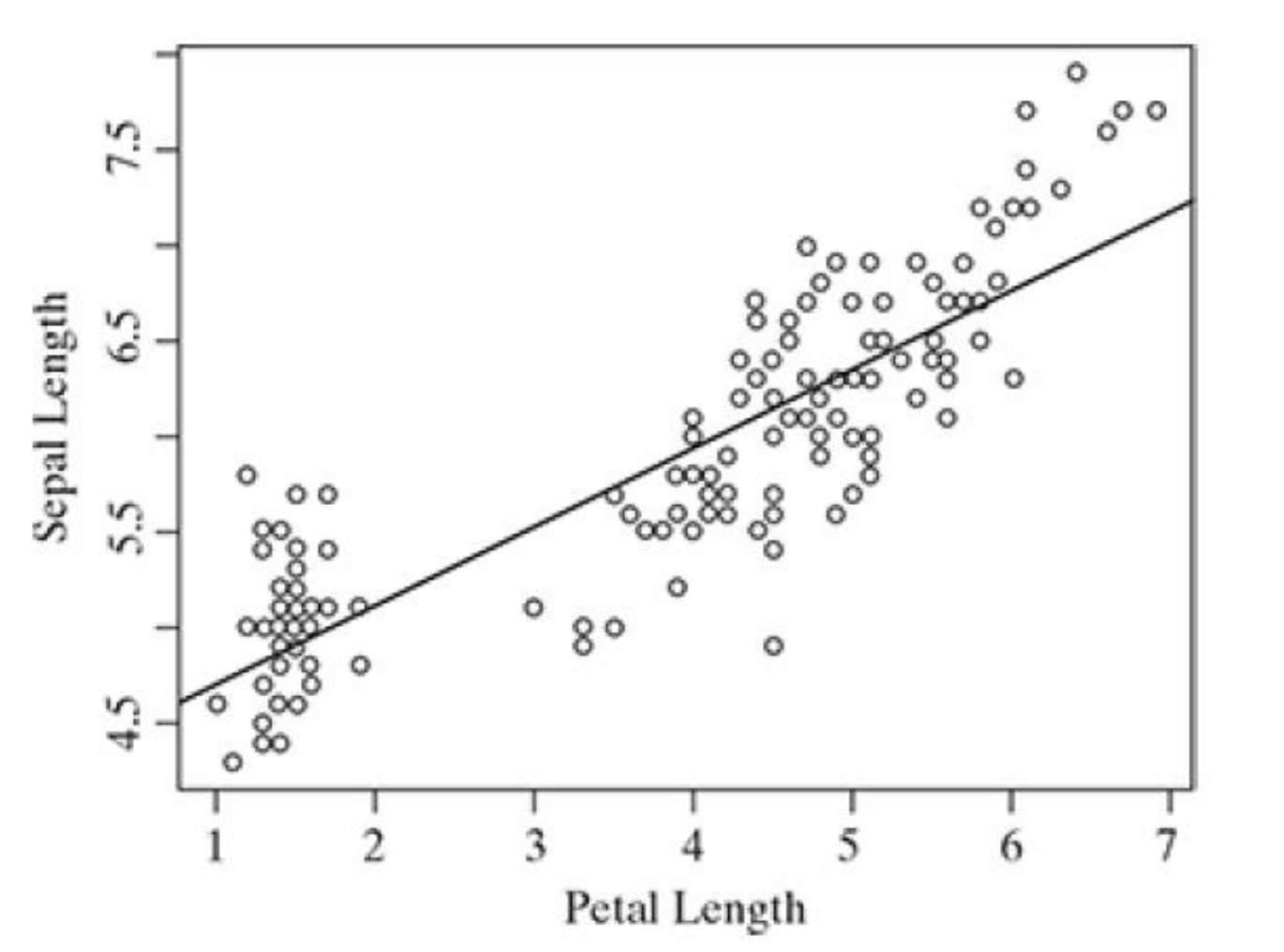
Example

- Yearly budgeting exercise of the sales managers. They have to give sales prediction for the next year based on sales figure of previous years.
- A typical linear regression model can be represented in the form –

$$y = \alpha + \beta x$$

where ‘x’ is the predictor variable and ‘y’ is the target variable.

Example: Predict Sepal length based on Petal length



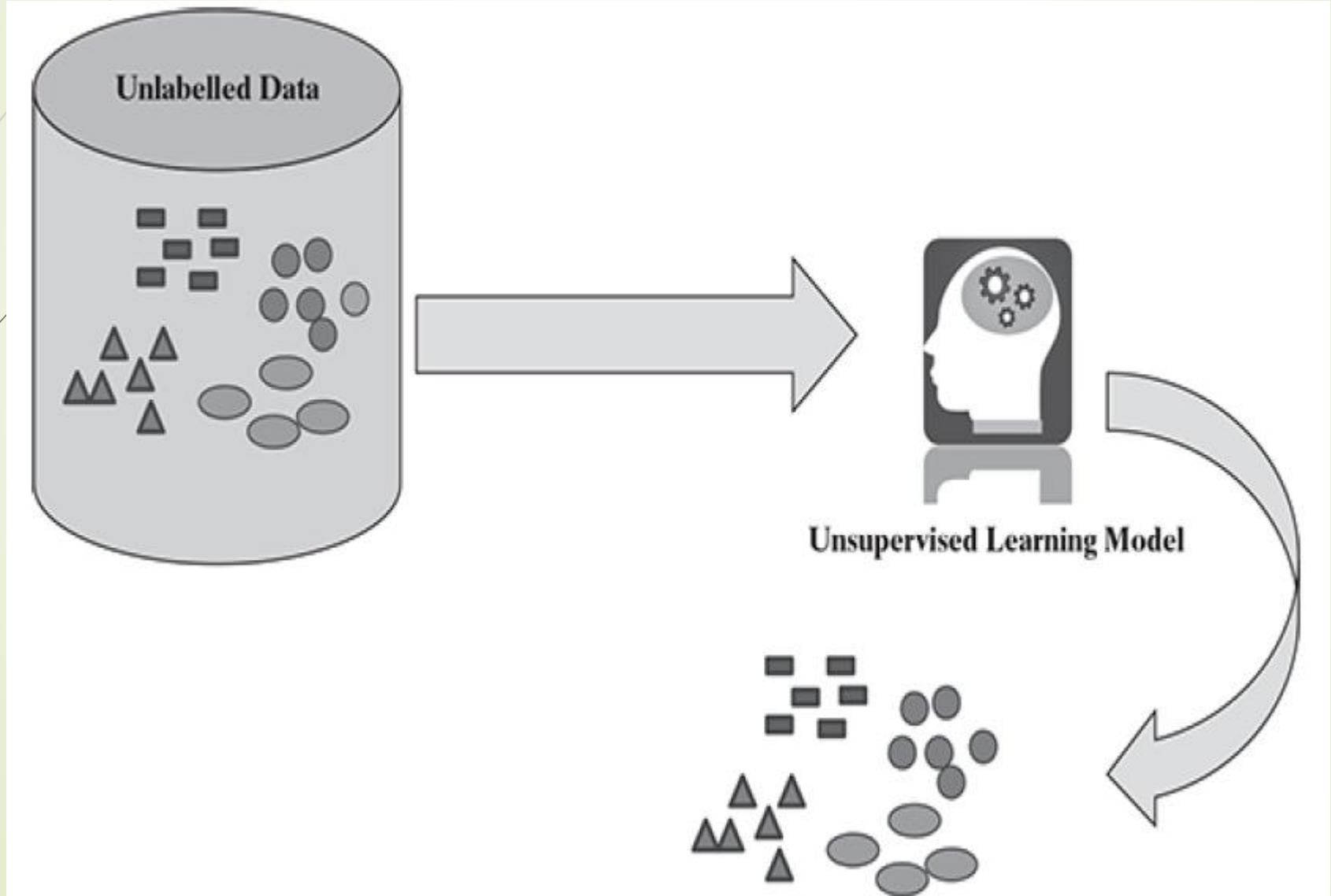
Applications of Regression

- Demand forecasting in retails
- Sales prediction for managers
- Price prediction in real estate
- Weather forecast
- Skill demand forecast in job market

Unsupervised learning

- Unlike supervised learning, in unsupervised learning, there is no labelled training data to learn from and no prediction to be made.
- In unsupervised learning, the objective is to take a dataset as input and try to find natural groupings or **patterns** within the data elements or records.
- Unsupervised learning is often termed as **descriptive model** and the process of unsupervised learning is referred as **pattern discovery** or **knowledge discovery**.

Unsupervised Learning



Clustering

Clustering is the main type of unsupervised learning. It intends to group or organize similar objects together.

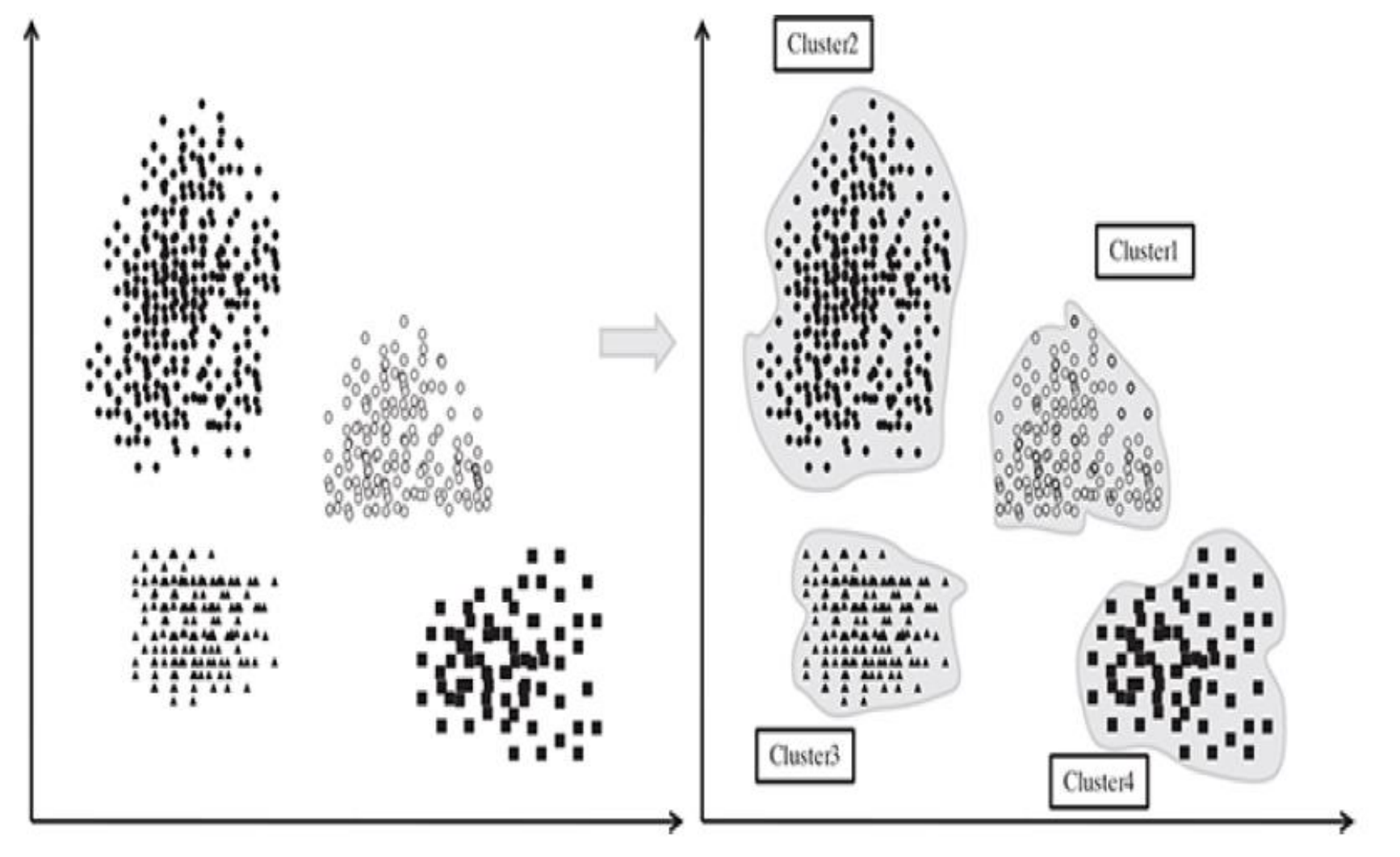
Objects belonging to the same cluster are quite similar

to each other while objects belonging to different clusters are

quite dissimilar.

Hence, the objective of clustering to discover the intrinsic grouping of unlabelled data and form clusters

Distance based Clustering



Association analysis

- The association between data elements is identified.
- From past transaction data in a grocery store, it may be observed that most of the customers who have bought item A, have also bought item B and item C or at least one of them.
- This means that there is a strong association of the event ‘purchase of item A’ with the event ‘purchase of item B’, or ‘purchase of item C’.
- Identifying these sorts of associations is the goal of association analysis.
- This helps in boosting up sales pipeline, hence a critical input for the sales group.
- Critical applications of association analysis include **market basket analysis and recommender systems**.

TransID	Items Bought
1	{Butter, Bread}
2	{Diaper, Bread, Milk, Beer}
3	{Milk, Chicken, Beer, Diaper}
4	{Bread, Diaper, Chicken, Beer}
5	{Diaper, Beer, Cookies, Ice cream}
…	…

Market Basket transactions

Frequent itemsets → (Diaper, Beer)

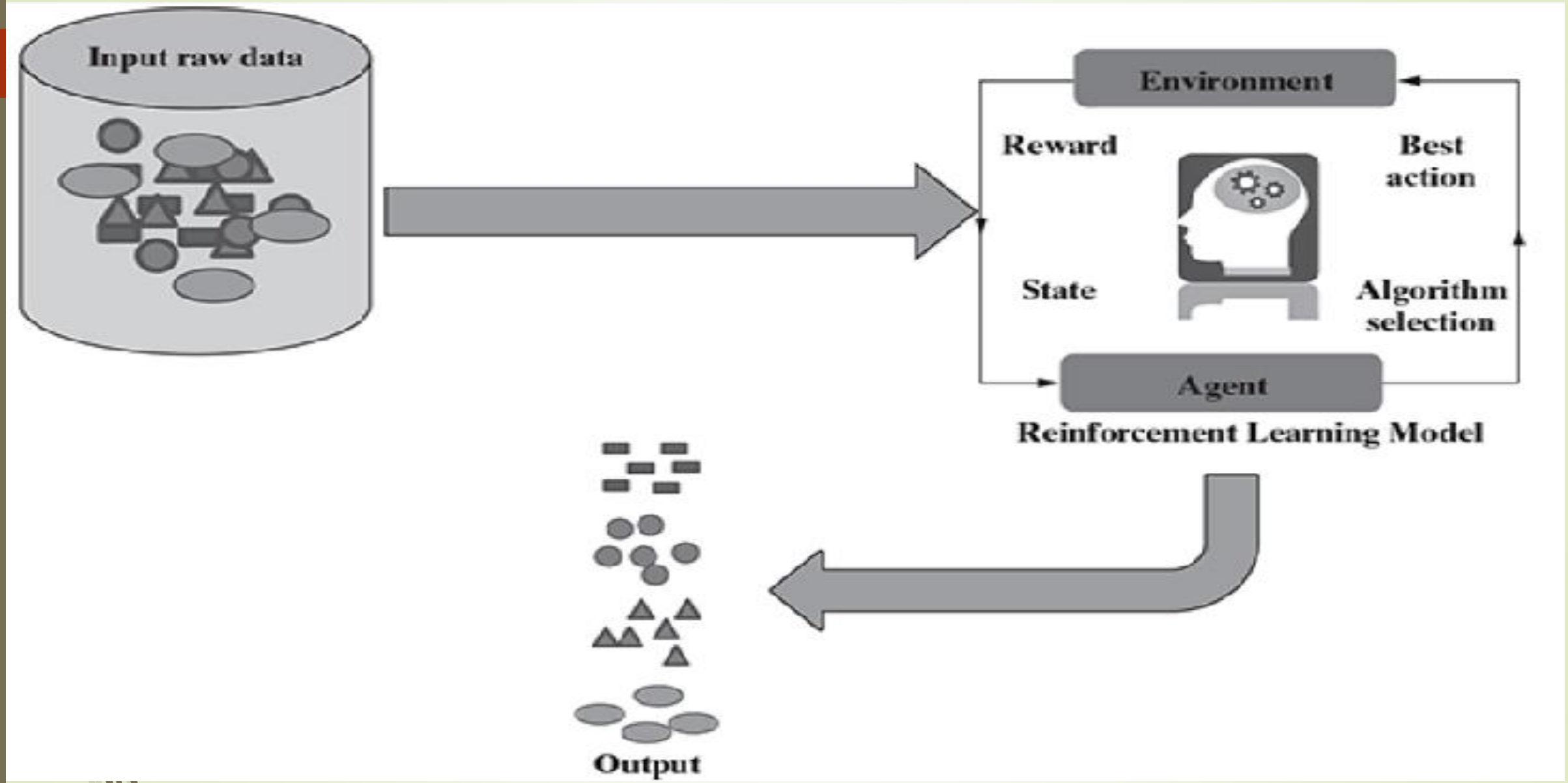
Possible association: Diaper → Beer

Market basket analysis

Reinforcement learning

- Machines learn to do tasks autonomously.
- E.g. We have seen babies learn to walk without any prior knowledge of how to do it.
- First they notice somebody else walking around.
- They understand that legs have to be used, one at a time, to take a step.
- While walking, sometimes they fall down hitting an obstacle, whereas other times they are able to walk smoothly avoiding bumpy obstacles.
- When they are able to walk overcoming the obstacle, their parents are elated and appreciate the baby with loud claps / or may be a chocolates.
- When they fall down while circumventing an obstacle, obviously their parents do not give claps or chocolates.
- Slowly a time comes when the babies learn from mistakes and are able to walk with much ease.

- The action tried to be achieved is walking, the child is the agent and the place with hurdles on which the child is trying to walk resembles the environment.
- It tries to improve its performance of doing the task.
- When a sub-task is accomplished successfully, a reward is given.
- When a sub-task is not executed correctly, obviously no reward is given.
- This continues till the machine is able to complete execution of the whole task.
- This process of learning is known as reinforcement learning.


Input raw data
Environment
Reward
Best action
State
Algorithm selection
Agent
Reinforcement Learning Model
Output

Example of reinforecement Learning

- One contemporary example of reinforcement learning is self-driving cars.
- The critical information which it needs to take care of are speed and speed limit in different road segments, traffic conditions, road conditions, weather conditions, etc.
- The tasks that have to be taken care of are start/stop, accelerate/decelerate, turn to left / right, etc.

Comparison – supervised, unsupervised, and reinforcement learning

SUPERVISED	UNSUPERVISED	REINFORCEMENT
This type of learning is used when you know how to classify a given data, or in other words classes or labels are available.	This type of learning is used when there is no idea about the class or label of a particular data. The model has to find pattern in the data.	This type of learning is used when there is no idea about the class or label of a particular data. The model has to do the classification – it will get rewarded if the classification is correct, else get punished.
Labelled training data is needed. Model is built based on training data.	Any unknown and unlabelled data set is given to the model as input and records are grouped.	The model learns and updates itself through reward/ punishment.
The model performance can be evaluated based on how many misclassifications have been done based on a comparison between predicted and actual values.	Difficult to measure whether the model did something useful or interesting. Homogeneity of records grouped together is the only measure.	Model is evaluated by means of the reward function after it had some time to learn.

SUPERVISED	UNSUPERVISED	REINFORCEMENT
There are two types of supervised learning problems – classification and regression.	There are two types of unsupervised learning problems – clustering and association.	No such types.
Simplest one to understand.	More difficult to understand and implement than supervised learning.	Most complex to understand and apply.
Standard algorithms include • Naïve Bayes • *k*-nearest neighbour (kNN) • Decision tree • Linear regression • Logistic regression • Support Vector Machine SVM), etc.	Standard algorithms are • *k*-means • Principal Component Analysis (PCA) • Self-organizing map (SOM) • Apriori algorithm • DBSCAN etc.	Standard algorithms are • Q-learning • Sarsa

SUPERVISED	UNSUPERVISED	REINFORCEMENT
Practical applications include • Handwriting recognition • Stock market prediction • Disease prediction • Fraud detection, etc.	Practical applications include • Market basket analysis • Recommender systems • Customer segmentation, etc.	Practical applications include • Self-driving cars • Intelligent robots • AlphaGo Zero (the latest version of DeepMind's AI system playing Go)

Applications of Machine learning

Banking and finance:

- The fraudulent transactions are spotted and prevented right at the time of occurrence.
- Customers of a bank are often offered lucrative proposals by other competitor banks. Machine learning helps in preventing or at least reducing the customer churn.

Insurance:

- Risk prediction during new customer onboarding and claims management.

Healthcare:

- Wearable device data form a rich source for applying machine learning and predict the health conditions
- In case there is some health issue which is predicted by the learning model, immediately the person is alerted to take preventive action.

LANGUAGES/TOOLS IN MACHINE LEARNING

- Python: Python is one of the most popular, open source programming language widely adopted by machine learning community.
- Python has very strong libraries for advanced mathematical functionalities (NumPy), algorithms and mathematical tools (SciPy) and numerical plotting (matplotlib)
- **scikitlearn** has various classification, regression, and clustering algorithms embedded in it.

R

- R is a language for statistical computing and data analysis.
- It is an open source language, extremely popular in the academic community – especially among statisticians and data miners.
- R is a very simple programming language with a huge set of libraries available for different stages of machine learning.
  - plyr/dplyr (for data transformation), caret ('Classification and Regression Training' for classification), RJava (to facilitate integration with Java), tm (for text mining), ggplot2 (for data visualization).

Matlab

- MATLAB (matrix laboratory) support for a wide range of numerical computing.
- MATLAB provides extensive support of statistical functions and has a huge number of machine learning algorithms in-built.
- It also has the ability to scale up for large datasets by parallel processing on clusters and cloud.

SAS

- SAS (earlier known as ‘Statistical Analysis System’) is a licenced commercial software which provides strong support for machine learning functionalities.
- SAS is a software suite comprising different components.
- The basic data management functionalities are embedded in the Base SAS component.
- Whereas the other components like SAS/INSIGHT, Enterprise Miner, SAS/STAT, etc. help in specialized functions related to data mining and statistical analysis.

Queries ???