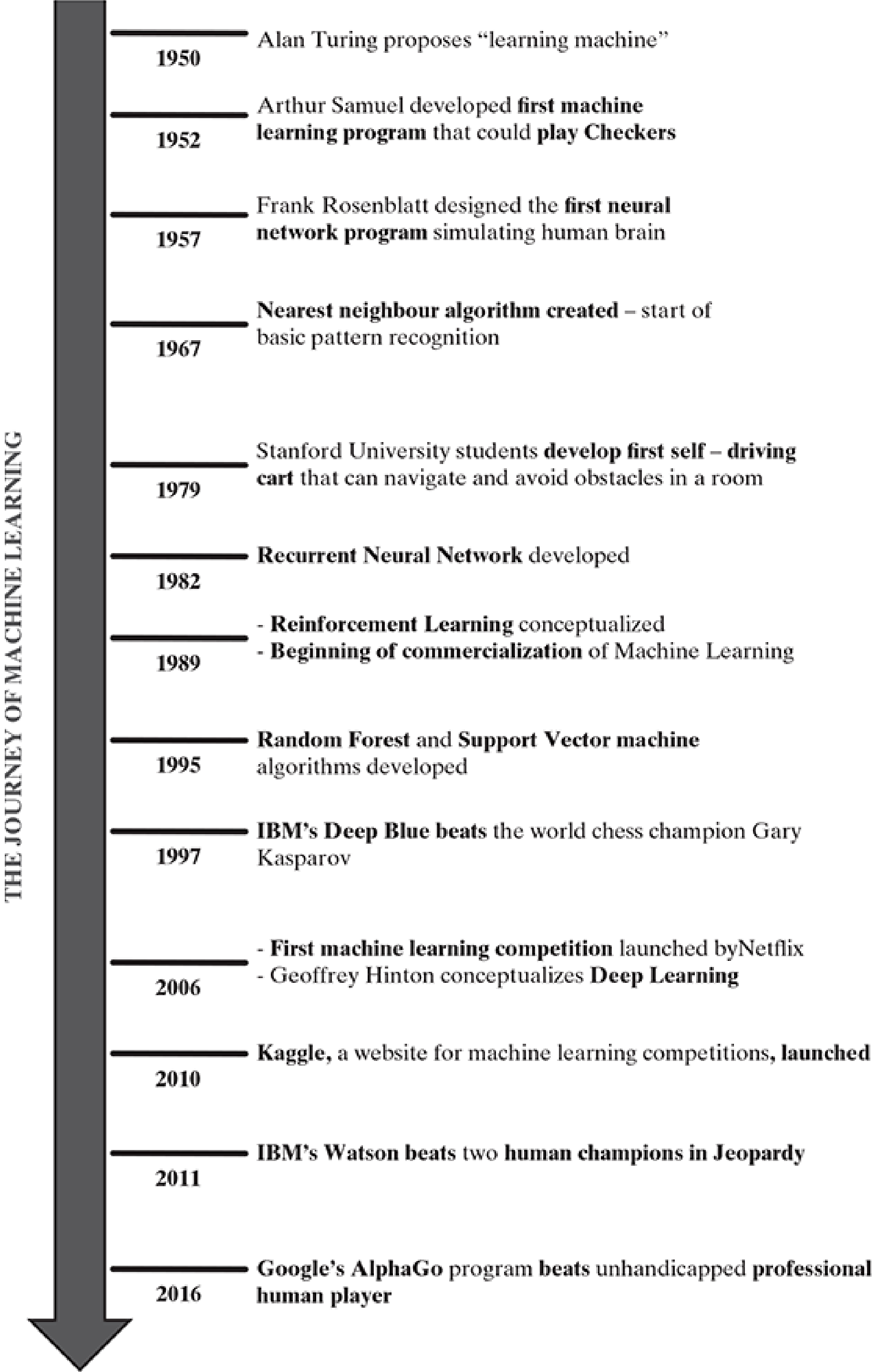
Introduction to Machine Learning

Machine learning Applications

- Machine learning helps Oncologist to find whether a tumour is malignant or benign.
- Medical Diagnosis
- Image Recognition
- Image Captioning
- Tagging images on social media
- Recognizing handwriting
- Speech Recognition (Voice search, Voice Dialing, Appliance control)
- Stock price prediction
- Sentence completion
- Machine Translation
- Chatbot
- Text summarization
- Recommendation Systems (Movie Recommendation, Product Recommendation, Book Recommendation)
- Crop yield Prediction
- Google self-driving car and Google Brain being two most ambitious projects of Google in its journey of innovation in the field of machine learning.
- And many more…….

Evolution of machine learning



THE JOURNEY OF MACHINE LEARNING

1950 Alan Turing proposes “learning machine”

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**

WHAT IS HUMAN LEARNING?

- Learning is referred to as the process of gaining information through observation.
- To do a task in a proper way, we need to have prior information on one or more things related to the task. Also, as we keep learning more or in other words acquiring more information, the efficiency in doing the tasks keep improving.
- For example, information from past rocket launches helps in taking the right precautions and makes more successful rocket launch.

TYPES OF 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, or

(3) we do it ourselves, may be after multiple attempts, some being unsuccessful.

- The first type of learning, we may call, falls under the category of learning directly under expert guidance,
- the second type falls under learning guided by knowledge gained from experts and
- the third type is learning by self or self-learning.

Learning under expert guidance

- The baby is able to learn alphabets, digits, sentences, paragraph, mathematics, science from his teacher who already has knowledge on these areas.
- In all phases of life of a human being, there is an element of guided learning. This learning is imparted by someone, purely because of the fact that he/she has already gathered the knowledge by virtue of his/her experience in that field. So guided learning is the process of gaining information from a person having sufficient knowledge due to the past experience.

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.
- A grown-up kid can select one odd word from a set of words because it is a verb and other words being all nouns. He could do this because of his ability to label the words as verbs or nouns, taught by his English teacher long back.

Learning by self

- In many situations, humans are left to learn on their own.
- A classic example is 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.
- A lot of things need to be learnt only from mistakes made in the past.

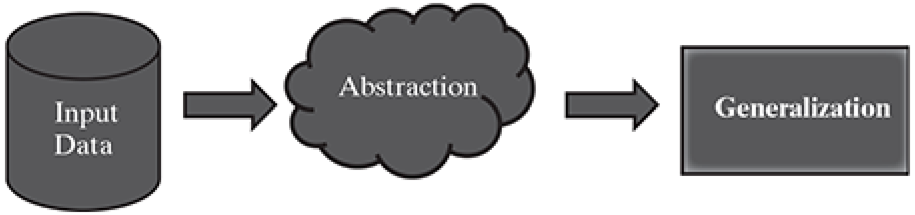
WHAT IS MACHINE LEARNING?

- Tom M. Mitchell, Professor of Machine Learning Department, School of Computer Science, Carnegie Mellon University has defined machine learning as
- **‘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.’**
- A machine can be considered to learn if it is able to gather experience by doing a certain task and improve its performance in doing the similar tasks in the future.

WHAT IS MACHINE LEARNING?

- When we talk about past experience, it means past data related to the task. This data is an input to the machine from some source.
- In context of image classification, E represents the past data with images having labels or assigned classes (for example whether the image is of a class cat or a class dog or a class elephant etc.), T is the task of assigning class to new, unlabelled images and P is the performance measure indicated by the percentage of images correctly classified.

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

Abstraction

- During the machine learning process, knowledge is fed in the form of input data.
- Abstraction helps in deriving a conceptual map based on the input data.
- This map, or a **model** as it is known in the machine learning paradigm, is summarized knowledge representation of the raw data.
- The model may be in any one of the following forms:
    - Computational blocks like if/else rules
    - Mathematical equations
    - Specific data structures like trees or graphs
    - Logical groupings of similar observations

Abstraction (Choice of Model)

- The choice of the model used to solve a specific learning problem is a human task.
- The decision related to the choice of model is taken based on multiple aspects, some of which are listed below:
- 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.
- Nature of the input data: How exhaustive the input data is, whether the data has no values for many fields, the data types, etc.
- Domain of the problem: If the problem is in a business critical domain with a high rate of data input and need for immediate inference, e.g. fraud detection problem in banking domain.

Training the Model

- Once the model is chosen, the next task is to fit the model based on the input data.
- Let’s understand this with an example.
- In a case where the model is represented by a mathematical equation, say
- ‘y = c1 + c2x’ (the model is known as simple linear regression), based on the input data, we have to find out the values of c1 and c2. Otherwise, the equation (or the model) is of no use.
- So, fitting the model, in this case, means finding the values of the unknown coefficients or constants of the equation or the model.
- This process of fitting the model based on the input data is known as training.
- The input data based on which the model is being finalized is known as training data.

Generalization

- The abstraction process, or more popularly training the model, is just one part of machine learning.
- The other key part is to tune up the abstracted knowledge to a form which can be used to take future decisions. This is achieved as a part of generalization.
- This part is quite difficult to achieve.
- This is because the model is trained based on a finite set of data, which may possess a limited set of characteristics. But when we want to apply the model to take decision on a set of unknown data, usually termed as test data, we may encounter two problems:
    1. The trained model is aligned with the training data too much, hence may not portray the actual trend.
    2. The test data possess certain characteristics apparently unknown to the training data.

Decision making based on model

- Hence, a precise approach of decision-making will not work.
- An approximate or heuristic approach, much like gut-feeling-based decision-making in human beings, has to be adopted.
- This approach has the risk of not making a correct decision – quite obviously because certain assumptions that are made may not be true in reality.
- But just like machines, same mistakes can be made by humans too when a decision is made based on intuition or gut-feeling – in a situation where exact reason-based decision-making is not possible.

Well-posed learning problem

- For defining a new problem, which can be solved using machine learning, a simple framework, highlighted below, can be used.
- This framework also helps in deciding whether the problem is a right candidate to be solved using machine learning.
- The framework involves answering three questions:

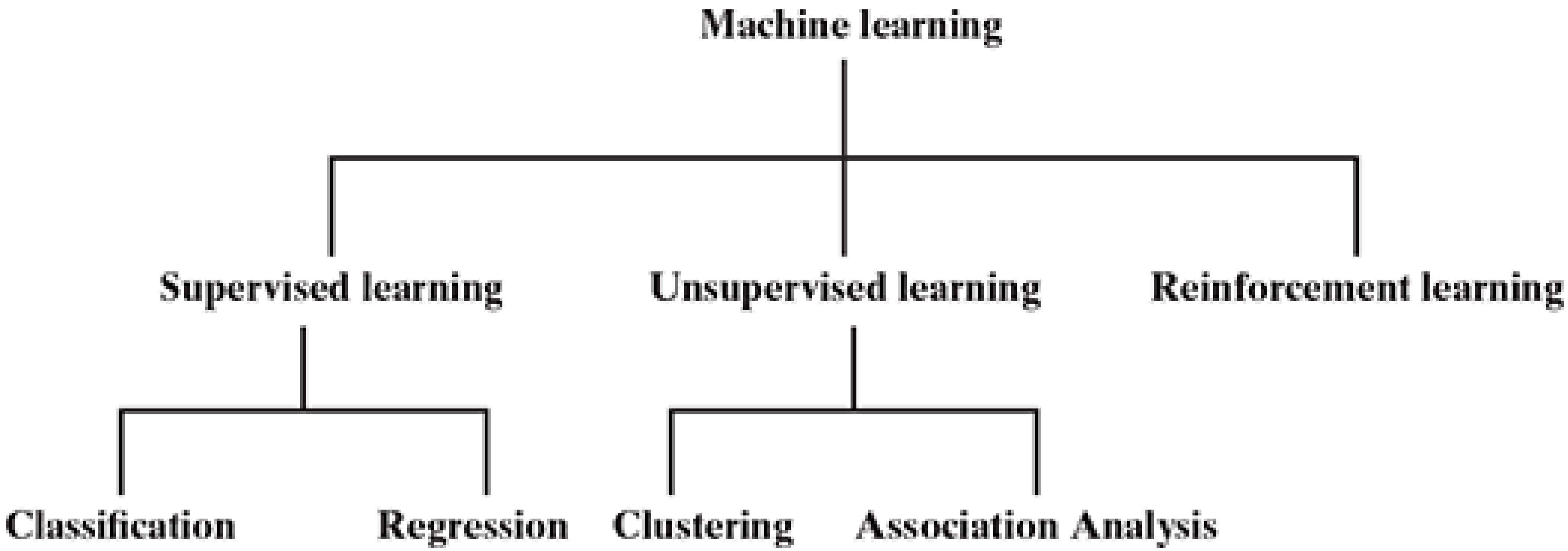
1. What is the problem?
2. Why does the problem need to be solved?
3. How to solve the problem?

- **Step 1: What is the Problem?**
- Informal description of the problem, e.g. I need a program that will prompt the next word as and when I type a word.
- Use Tom Mitchell’s machine learning formalism stated above to define the T, P, and E for the problem.
- For given example:
  - Task (T): Prompt the next word when I type a word.
  - Experience (E): A corpus of commonly used English words and phrases together.
  - Performance (P): The number of correct words prompted considered as a percentage (which in machine learning paradigm is known as learning accuracy).
- **Assumptions -** Create a list of assumptions about the problem.
- **Similar problems**
- What other problems have you seen or can you think of that are similar to the problem that you are trying to solve?

- **Step 2: Why does the problem need to be solved?**
- What is the motivation for solving the problem? What requirement will it fulfil?
- Consider the benefits of solving the problem.
- How will the solution to the problem be used and the life time of the solution is expected to have?

- **Step 3: How would I solve the problem?**
- Try to explore how to solve the problem manually.
- Detail out step-by-step data collection, data preparation, and program design to solve the problem.
- Collect all these details and update the previous sections of the problem definition, especially the assumptions.

TYPES OF MACHINE LEARNING



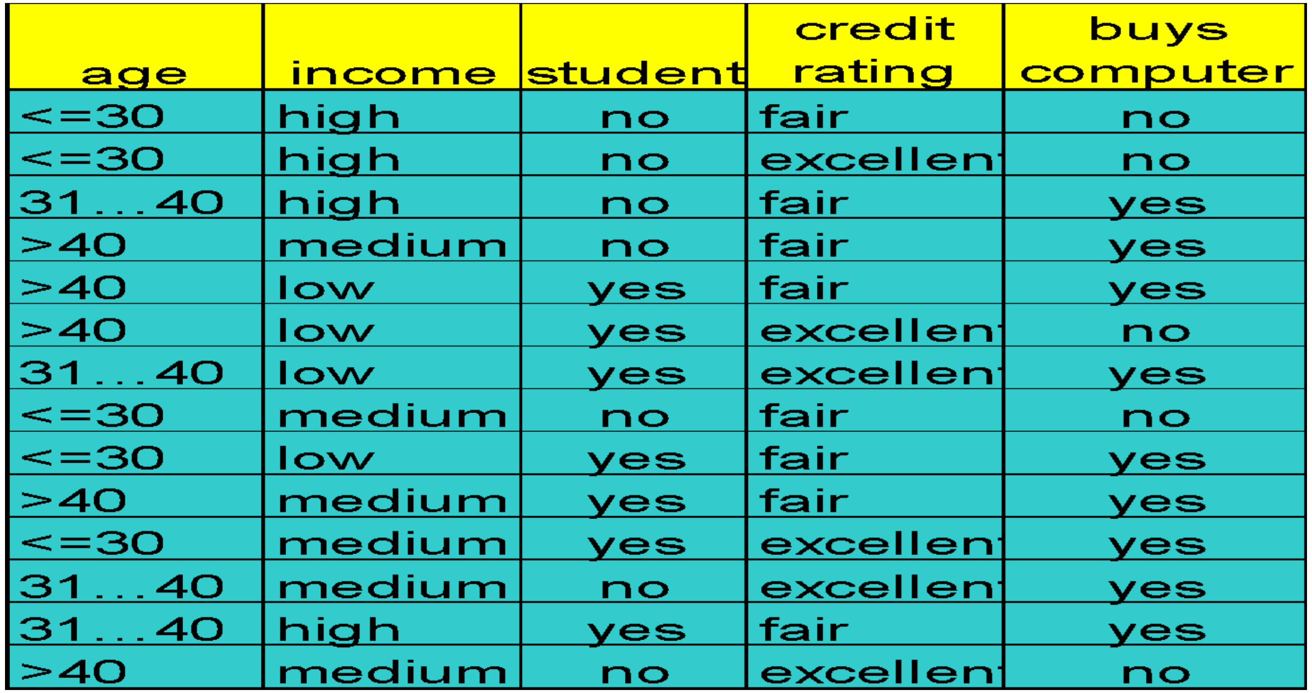
TYPES OF MACHINE LEARNING

- 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

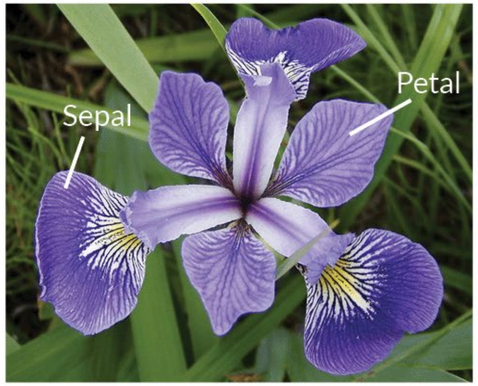
- The major motivation of supervised learning is to learn from past information.
- So what kind of past information does the machine need for supervised learning?
- Example-1: Predict whether the customer will buy computer or not?



age	income	student	credit rating	buys computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Example 2: Classify iris plant

The data set contains 3 classes, where each class refers to a type of iris plant



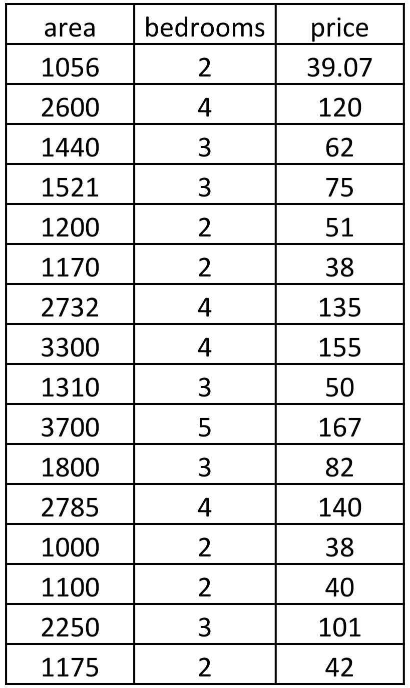
Iris Versicolor

Iris Setosa

Iris Virginica

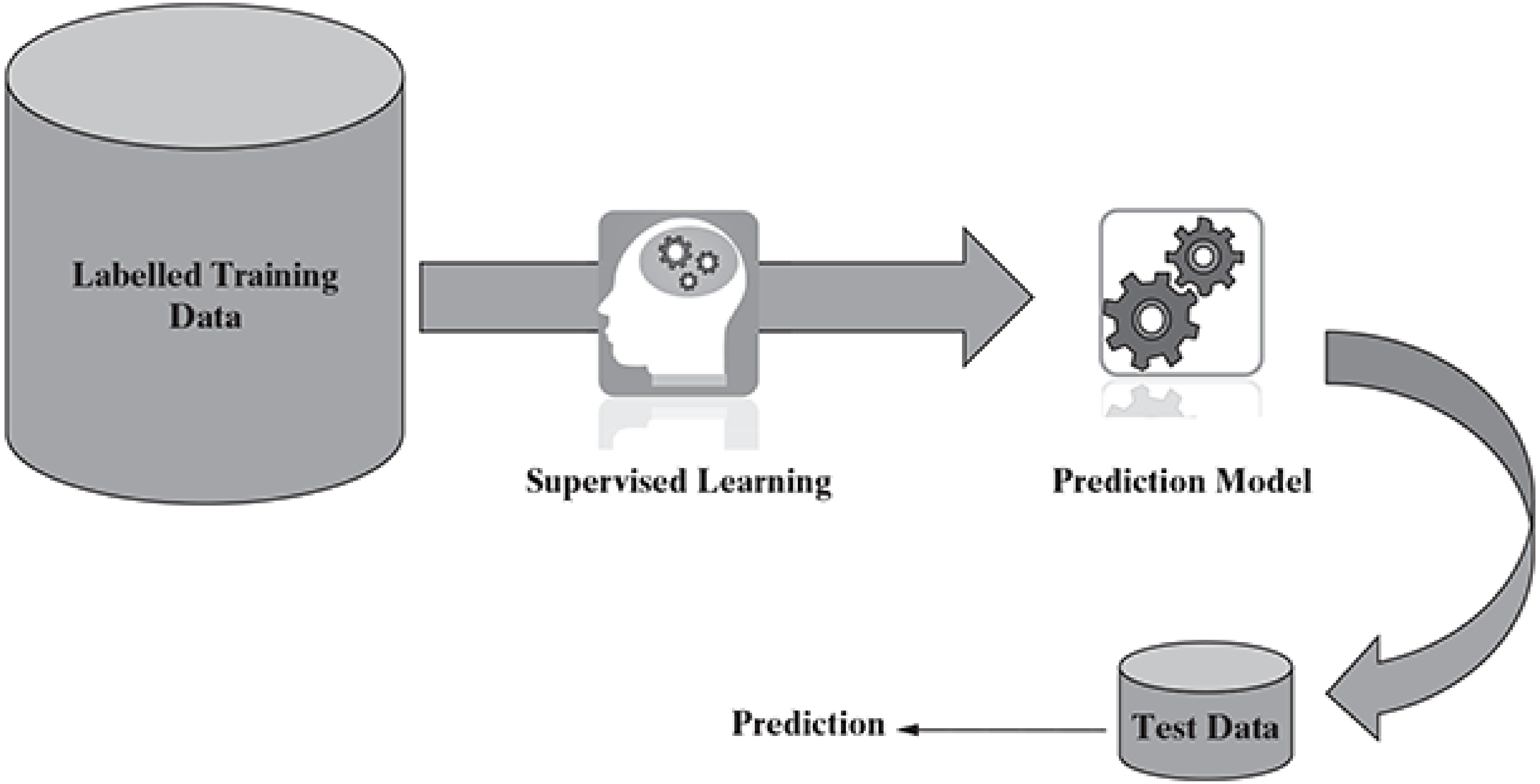
Sepal-length	Sepal-width	Petal-length	Petal-width	class
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3.0	1.4	0.2	Iris-setosa
7.0	3.2	4.7	1.4	Iris-versicolor
6.4	3.2	4.5	1.5	Iris-versicolor
4.9	2.5	4.5	1.7	Iris-virginica
7.3	2.9	6.3	1.8	Iris-virginica
6.7	2.5	5.8	1.8	Iris-virginica

Example-3: House price prediction



area	bedrooms	price
1056	2	39.07
2600	4	120
1440	3	62
1521	3	75
1200	2	51
1170	2	38
2732	4	135
3300	4	155
1310	3	50
3700	5	167
1800	3	82
2785	4	140
1000	2	38
1100	2	40
2250	3	101
1175	2	42

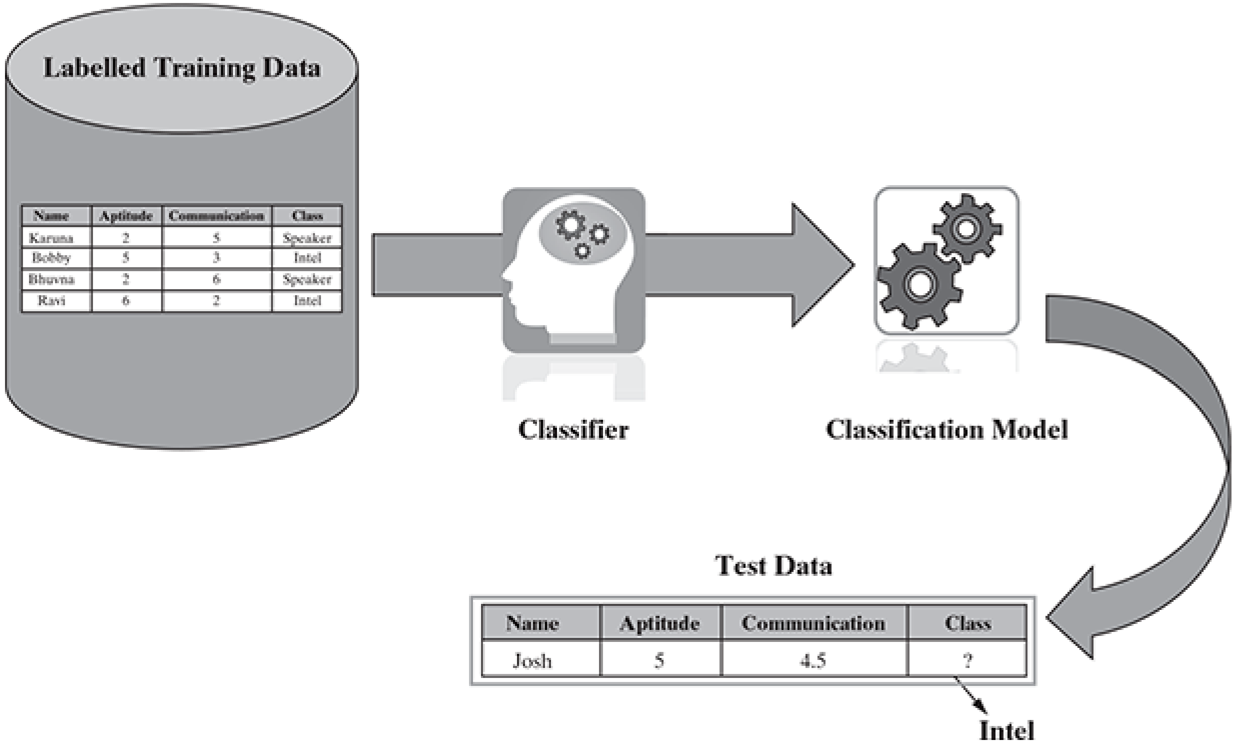
Supervised learning



Supervised Learning

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

Classification



Name	Aptitude	Communication	Class
Karuna	2	5	Speaker
Bobby	5	3	Intel
Bhuvna	2	6	Speaker
Ravi	6	2	Intel

Name	Aptitude	Communication	Class
Josh	5	4.5	?

- There are number of popular machine learning algorithms which help in solving classification problems.
- To name a few, Naïve Bayes, Decision tree, and k-Nearest Neighbour algorithms are adopted by many machine learning practitioners.

Classification Examples:

- In context of banking domain is identifying potential fraudulent transactions.
- Image classification
- Prediction of disease
- 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.
- The underlying predictor variable and the target variable are continuous in nature.
- In case of linear regression, a straight line relationship is ‘fitted’ between the predictor variables and the target variables, using the statistical concept of least squares method.
- As in the case of least squares method, the sum of square of error between actual and predicted values of the target variable is tried to be minimized.
- In case of simple linear regression, there is only one predictor variable.
- In case of multiple linear regression, multiple predictor variables can be included in the model.
- 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.

Regression

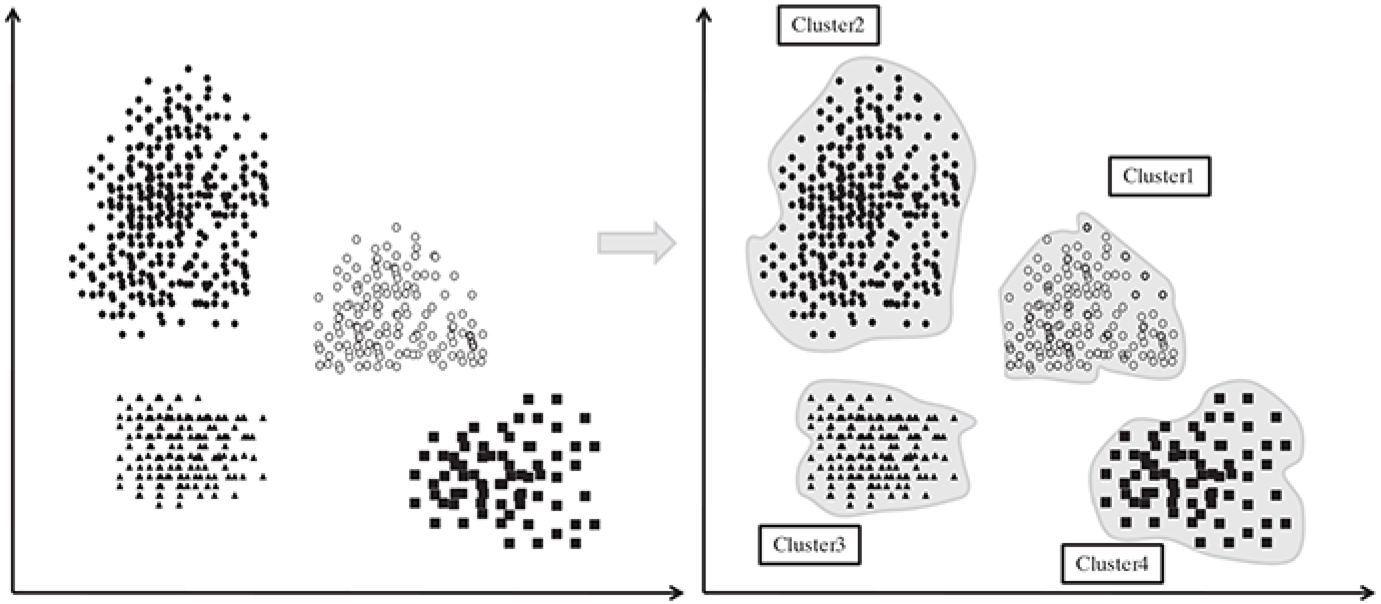
- Typical applications of regression can be seen in
  - Demand forecasting in retails
  - Sales prediction for managers
  - Price prediction in real estate
  - Weather forecast

Unsupervised 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.
- Therefore, unsupervised learning is often termed as **descriptive model** and the process of unsupervised learning is referred as **pattern discovery** or **knowledge discovery**.
- One critical application of unsupervised learning is customer segmentation.

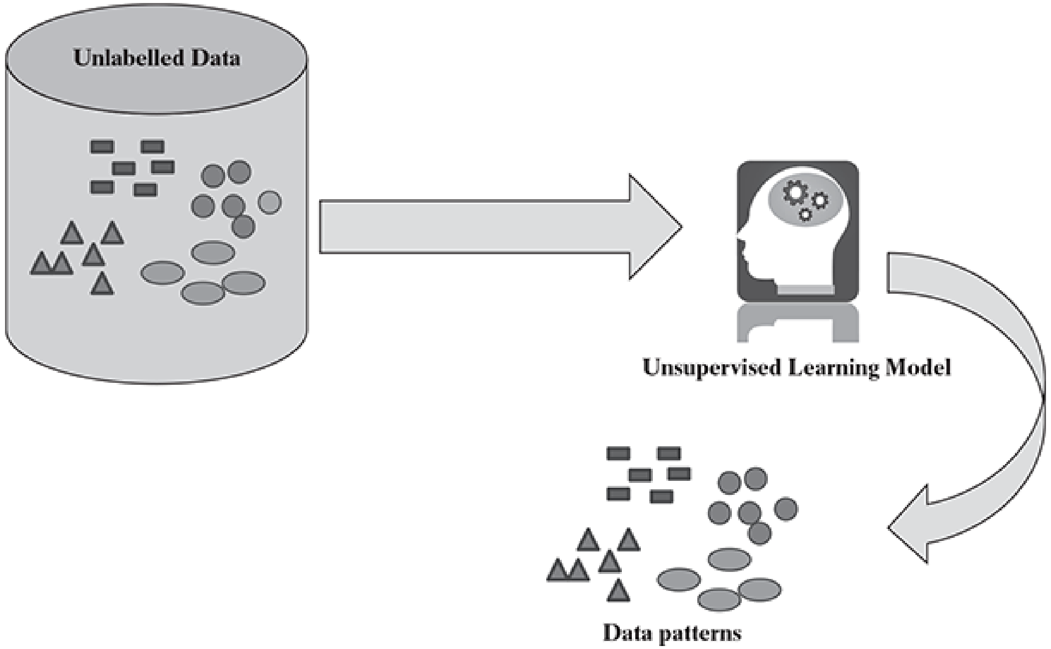
Unsupervised Learning - Clustering

- Clustering is the main type of unsupervised learning.
- It intends to group or organize similar objects together.
- For that reason, 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, as shown in below figure.



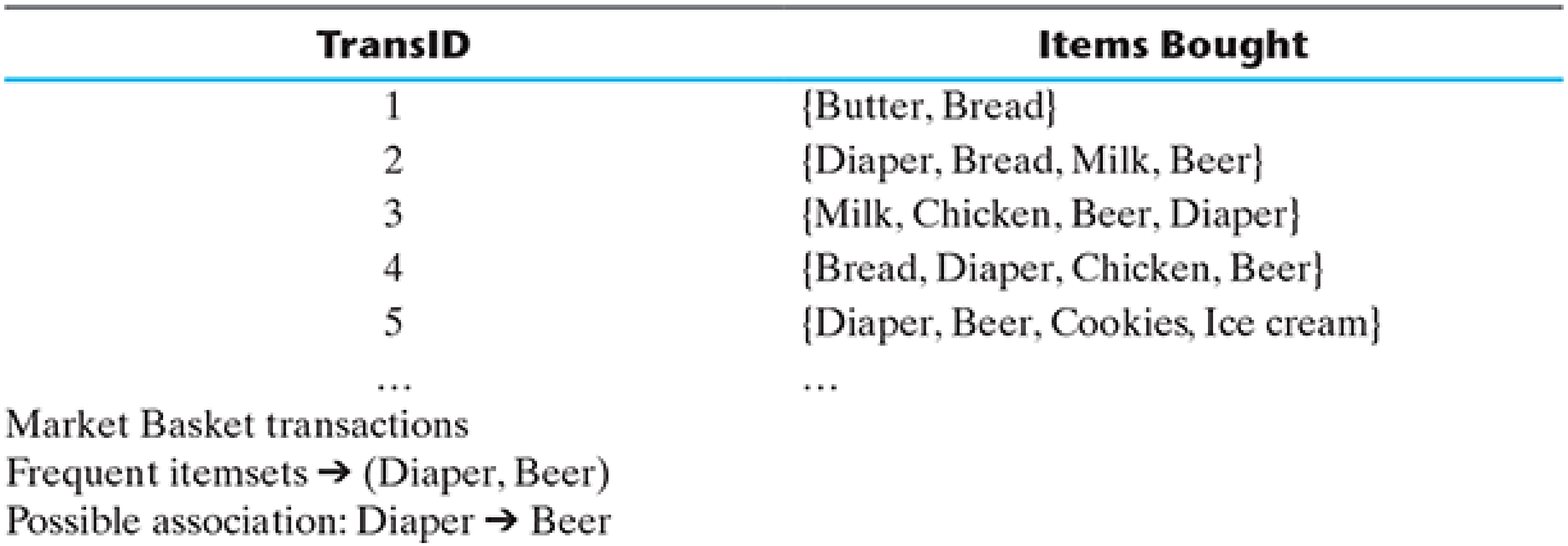
Unsupervised Learning – Association Analysis

- As a part of association analysis, the association between data elements is identified.
- Example: market basket analysis
- 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.



Unsupervised Learning – Association Analysis

- Market Basket Analysis



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}
…	…

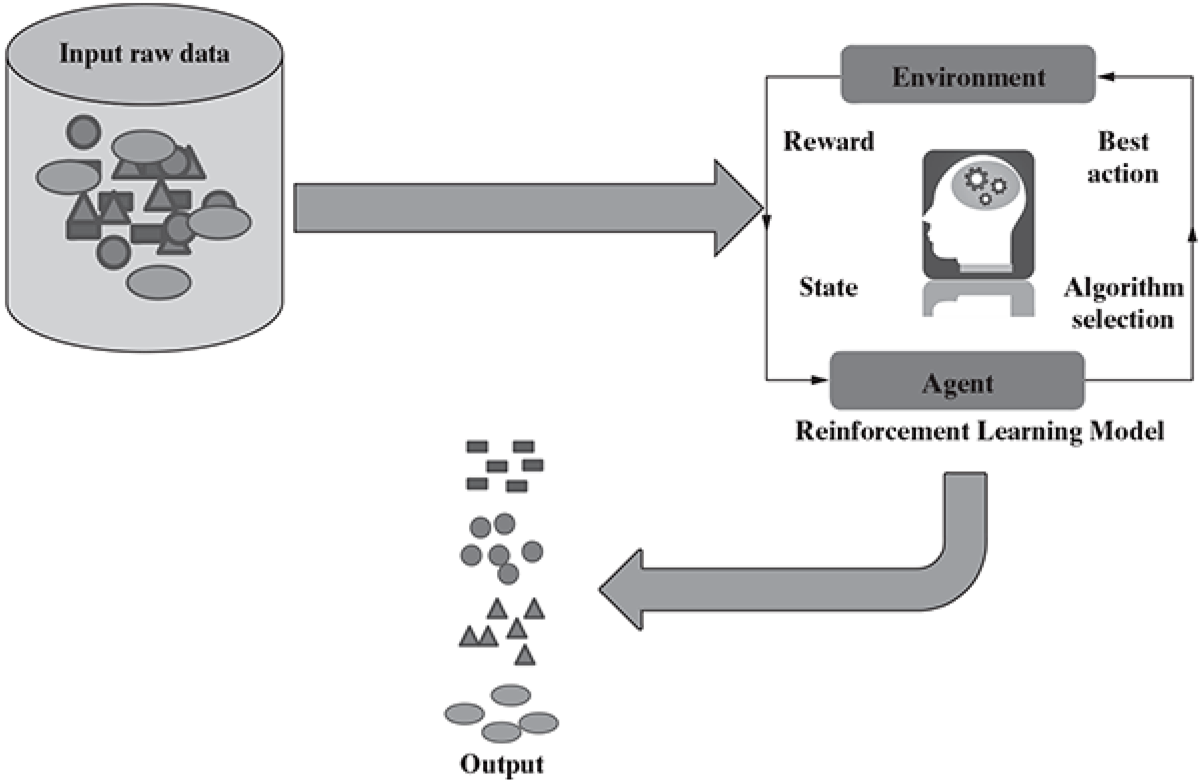
Market Basket transactions

Frequent itemsets ➔ (Diaper, Beer)

Possible association: Diaper ➔ Beer

Reinforcement learning

- Reinforcement learning is a machine learning training method based on rewarding desired behaviors and/or punishing undesired ones.
- A reinforcement learning agent is able to perceive and interpret its environment, take actions and learn through trial and error.

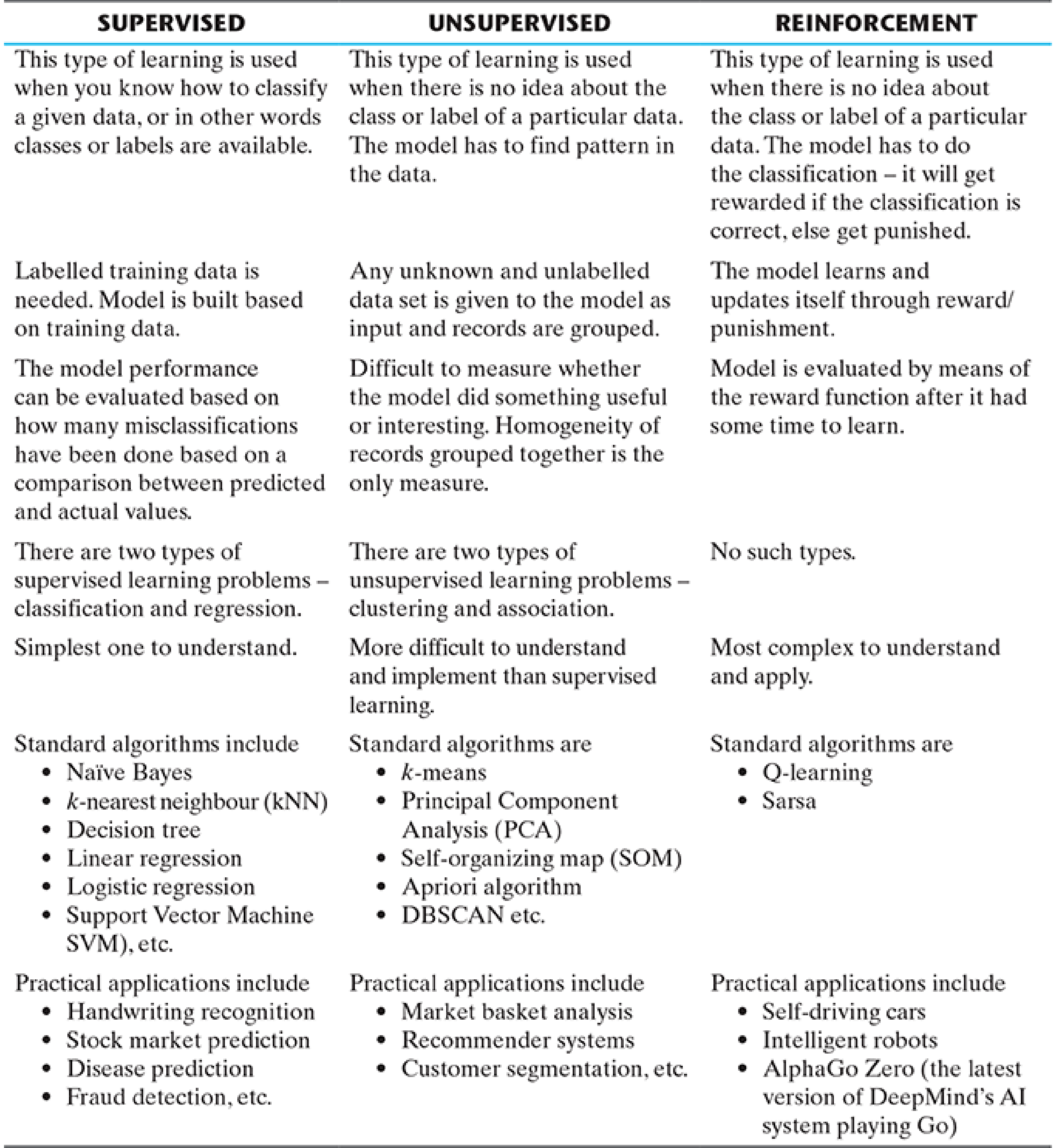


How does reinforcement learning work?

- In reinforcement learning, developers devise a method of rewarding desired behaviors and punishing negative behaviors.
- This method assigns positive values to the desired actions to encourage the agent and negative values to undesired behaviors.
- This programs the agent to seek long-term and maximum overall reward to achieve an optimal solution.
- These long-term goals help prevent the agent from stalling on lesser goals.
- With time, the agent learns to avoid the negative and seek the positive.
- This learning method has been adopted in artificial intelligence (AI) as a way of directing unsupervised machine learning through rewards and penalties.
- Example: Self-driving car

Reinforcement learning

- Reinforcement learning differs from supervised learning in a way that in supervised learning the training data has the answer key with it so the model is trained with the correct answer itself.
- In reinforcement learning, there is no answer but the reinforcement agent decides what to do to perform the given task. In the absence of a training dataset, it is bound to learn from its experience.



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.
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
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**
- Customers of a bank are often offered lucrative proposals by other competitor banks.
- Sometimes customers get demotivated by the poor quality of services of the banks and shift to competitor banks.
- Machine learning helps in preventing or at least reducing the customer churn. Both descriptive and predictive learning can be applied for reducing customer churn.
- Using descriptive learning, the specific pockets of problem, i.e. a specific bank or a specific zone or a specific type of offering like car loan, may be spotted where maximum churn is happening.
- Using predictive learning, the set of vulnerable customers who may leave the bank very soon, can be identified. Proper action can be taken to make sure that the customers stay back.
- In the banking industry, fraudulent transactions, especially the ones related to credit cards, are extremely prevalent. The models work on a real-time basis, i.e. the fraudulent transactions are spotted and prevented right at the time of occurrence.

Applications of Machine Learning

- **Insurance**
- Two major areas in the insurance industry where machine learning is used are risk prediction during new customer onboarding and claims management.
- During customer onboarding, based on the past information the risk profile of a new customer needs to be predicted. Based on the quantum of risk predicted, the quote is generated for the prospective customer.
- When a customer claim comes for settlement, past information related to historic claims along with the adjustor notes are considered to predict whether there is any possibility of the claim to be fraudulent.

Applications of Machine Learning

- **Healthcare**
- Wearable device data form a rich source for applying machine learning and predict the health conditions of the person real time.
- In case there is some health issue which is predicted by the learning model, immediately the person is alerted to take preventive action.
- Machine learning along with computer vision also plays a crucial role in disease diagnosis from medical imaging.

Language/Tools in Machine Learning

- Python
  - Python has very strong libraries for advanced mathematical functionalities (NumPy), algorithms and mathematical tools (SciPy) and numerical plotting (matplotlib). Built on these libraries, there is a machine learning library named **scikit-learn**, which 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. Some of the libraries standing out in terms of popularity are 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).

Language/Tools in Machine Learning

- **Matlab**
  - MATLAB also 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 another licensed commercial software which provides strong support for machine learning functionalities.

Issues in Machine Learning

- The biggest fear and issue arising out of machine learning is related to privacy and the breach of it.
- The primary focus of learning is on analyzing data, both past and current, and coming up with insight from the data.
- This insight may be related to people and the facts revealed might be private enough to be kept confidential.