

# Machine Learning

B. Keerthana

Assistant Professor

Computer Science and Engineering (CSD)

GVPCE(A)

# Unit - 1

## **Machine Learning Basics:**

- The Need for Machine Learning.
- Understanding Machine Learning,
- Computer Science, Data Science, Artificial Intelligence,
- Natural Language Processing, Deep Learning,
- Machine Learning Methods, Semi-Supervised Learning, Reinforcement Learning,
- Model Based Learning,
- The CRISP-DM Process Model,
- Building Machine Intelligence, and Real-World Case Study

# The Need for machine learning

## 1. Making data driven decisions:

- The art and science of leveraging your data to get actionable insights and make better decisions is known as making data-driven decisions.
- Fields like operations research, statistics, and management information systems have existed for decades and attempt to bring efficiency to any business or organization by using data and analytics to make data-driven decisions.
- Solutions to problems that cannot be programmed inherently need a different approach where we use the data itself to drive decisions instead of using programmable logic, rules, or code to make these decisions

## 2. Efficiency and Scale:

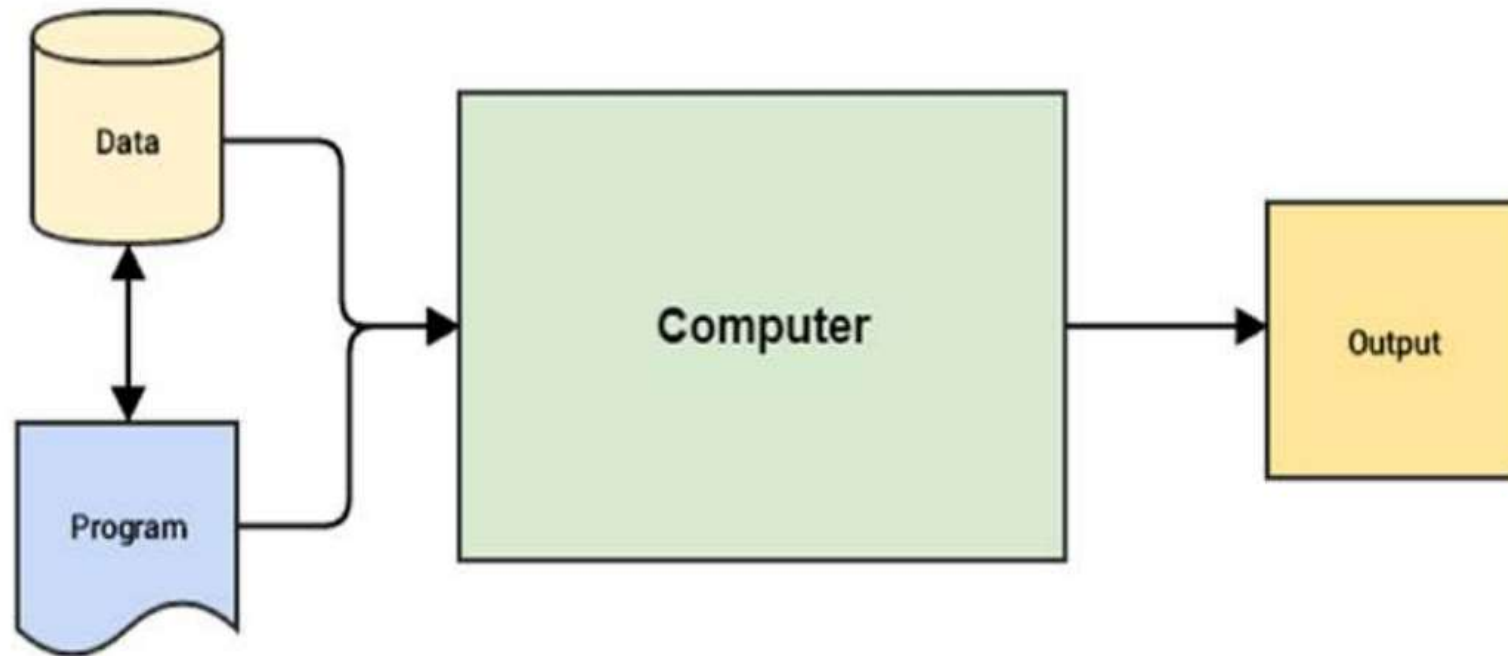
- While getting insights and making decisions driven by data are of paramount importance, it also needs to be done with efficiency and at scale.
- The key idea of using techniques from Machine Learning or artificial intelligence is to **automate processes or tasks by learning specific patterns from the data.**
- We all want computers or machines to tell us “when a stock might rise or fall”, “whether an image is of a computer or a television”, “whether our product placement and offers are the best”, “determine shopping price trends” etc.

### Scale :

A unit is said to be scale efficient when its size of operations is optimal so that any modifications on its size will render the unit less efficient.

### 3. Traditional Programming Paradigm:

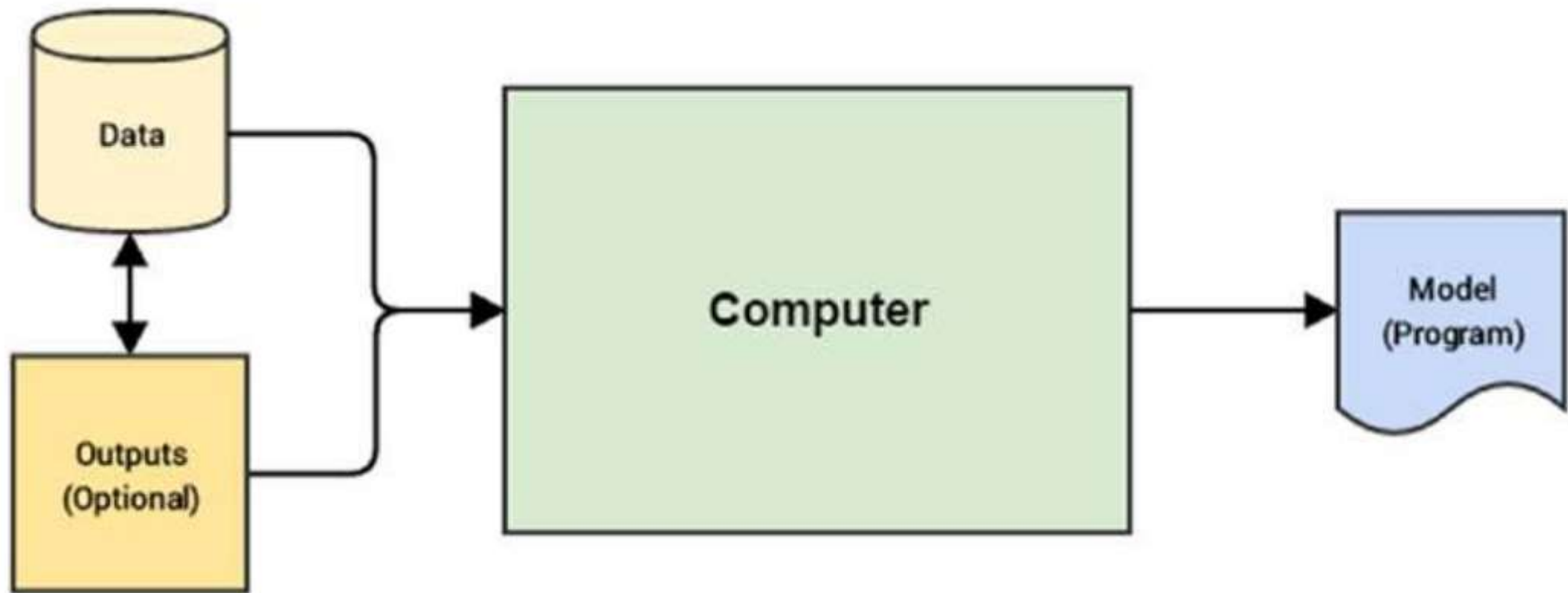
- Traditional programming paradigms basically involve the user or programmer to write a set of instructions or operations using code that makes the computer perform specific computations on data to give the desired results.



*Figure 1-1. Traditional programming paradigm*

# Why Machine Learning?

- The traditional programming paradigm is quite good and human intelligence and domain expertise is definitely an important factor in making data-driven decisions, we need Machine Learning to make faster and better decisions.
- The Machine Learning paradigm tries to take into account data and expected outputs or results if any and uses the computer to build the program, which is also known as a model.
- This program or model can then be used in the future to make necessary decisions and give expected outputs from new inputs.



*Figure 1-2. Machine Learning paradigm*

# Understanding Machine Learning

## Why Make Machines Learn?

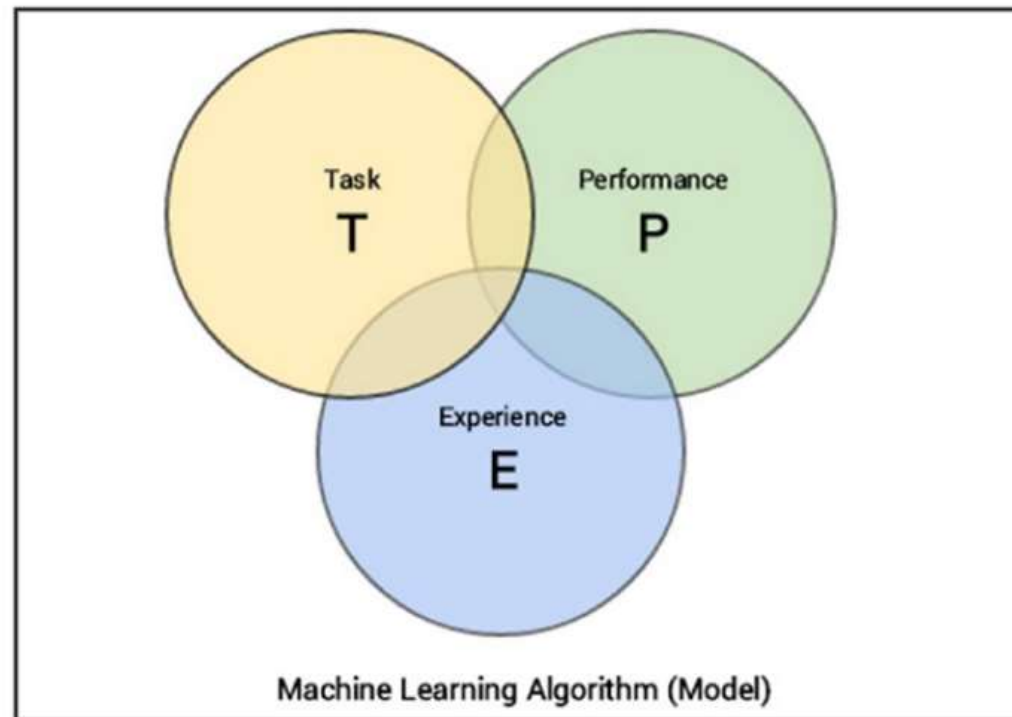
There are several scenarios when it might be beneficial to make machines learn and some of them are:

1. **Lack of sufficient human expertise in a domain** (e.g., simulating navigations in unknown territories or even spatial planets).
2. **Scenarios and behavior can keep changing over time** (e.g., availability of infrastructure in an organization, network connectivity, and so on).
3. **Humans have sufficient expertise in the domain but it is extremely difficult to formally explain or translate this expertise into computational tasks** (e.g., speech recognition, translation, scene recognition, cognitive tasks, and so on).
4. **Addressing domain specific problems at scale with huge volumes of data with too many complex conditions and constraints.**



# Formal Definition

“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  $E$ . ”



# Defining the Task, T

- **Classification or categorization:** A simple example would be classifying animal images into dogs, cats, and zebras.
- **Regression:** stock predictions, House price predictions
- **Translation:** Translating from one language to other
- **Clustering or grouping:** Examples would be grouping similar products, events and entities.
- **Transcriptions:** Examples include speech to text, optical character recognition, images to text, and so on

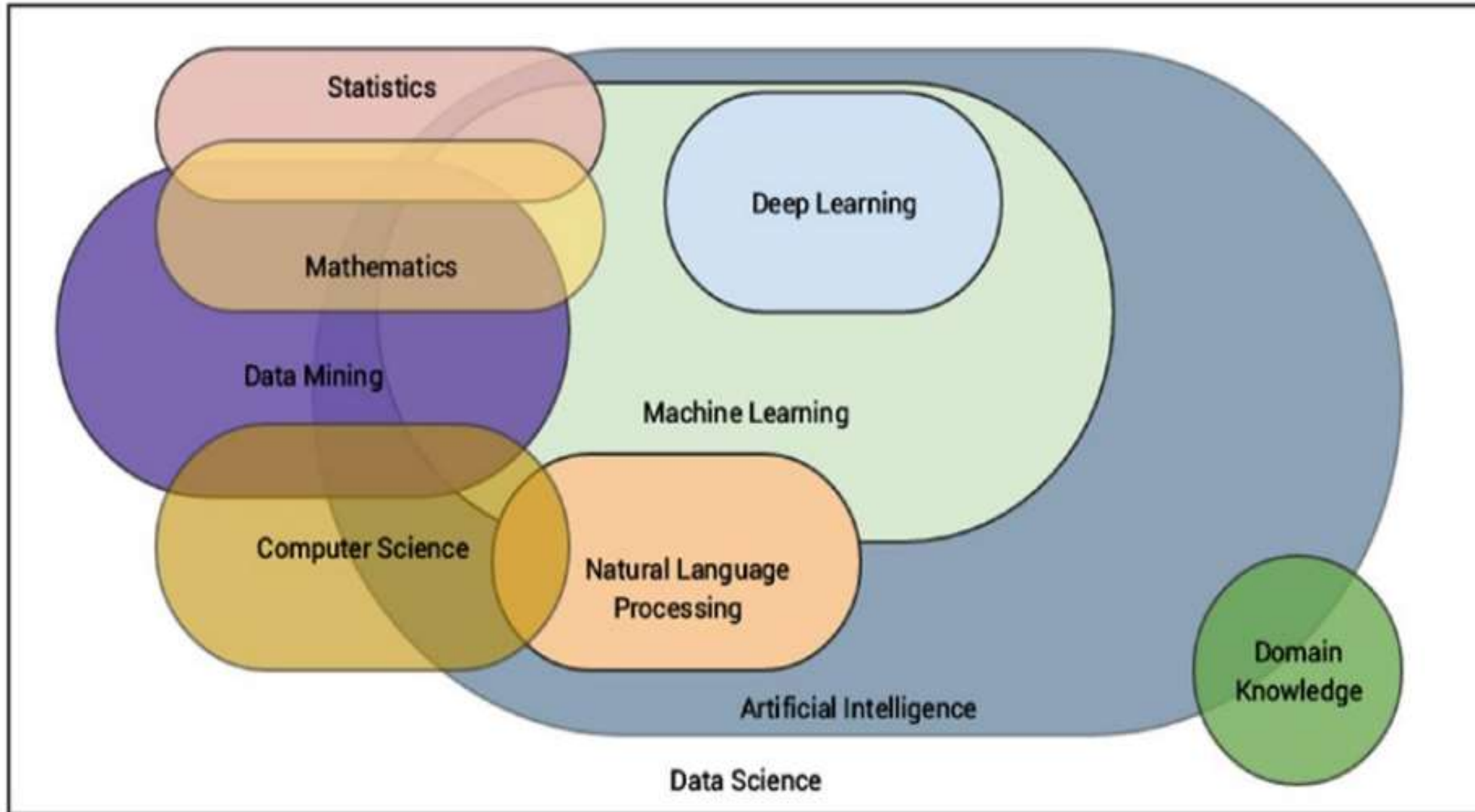
# Defining the Experience, E

- The process of consuming a dataset that consists of data samples or data points such that a learning algorithm or model learns inherent patterns is defined as the **experience, E** which is gained by the learning algorithm.
- Any experience that the algorithm gains is from data samples or data points and this can be at any point of time.
- You can feed it data samples in one go using historical data or even supply fresh data samples whenever they are acquired.

# Defining the Performance, P

- The performance, P, is usually a quantitative measure or metric that's used to see how well the algorithm or model is performing the task, T, with experience, E.
- Typical performance measures include accuracy, precision, recall, F1 score, sensitivity, specificity, error rate, misclassification rate, and many more.
- Performance measures are usually evaluated on training data samples as well as data samples which it has not seen or learned from before, which are usually known as validation and test data samples.

# A Multi-Disciplinary Field



# Computer Science

- The field of computer science (CS) can be defined as the study of the science of understanding computers.
- This involves study, research, development, engineering, and experimentation of areas dealing with understanding, designing, building, and using computers.
- This also involves extensive design and development of algorithms and programs that can be used to make the computer perform computations and tasks as desired.

- There are mainly two major areas or fields under computer science, as follows
  - Theoretical computer science
  - Applied or practical computer science

### **Theoretical computer science:**

- Theoretical computer science is the study of theory and logic that tries to explain the principles and processes behind computation.
- This involves understanding the theory of computation which talks about how computation can be used efficiently to solve problems.
- Theory of computation includes the study of formal languages, automata, and understanding complexities involved in computations and algorithms
- data structures and algorithms are the two fundamental pillars of theoretical CS used extensively in computational programs and functions.

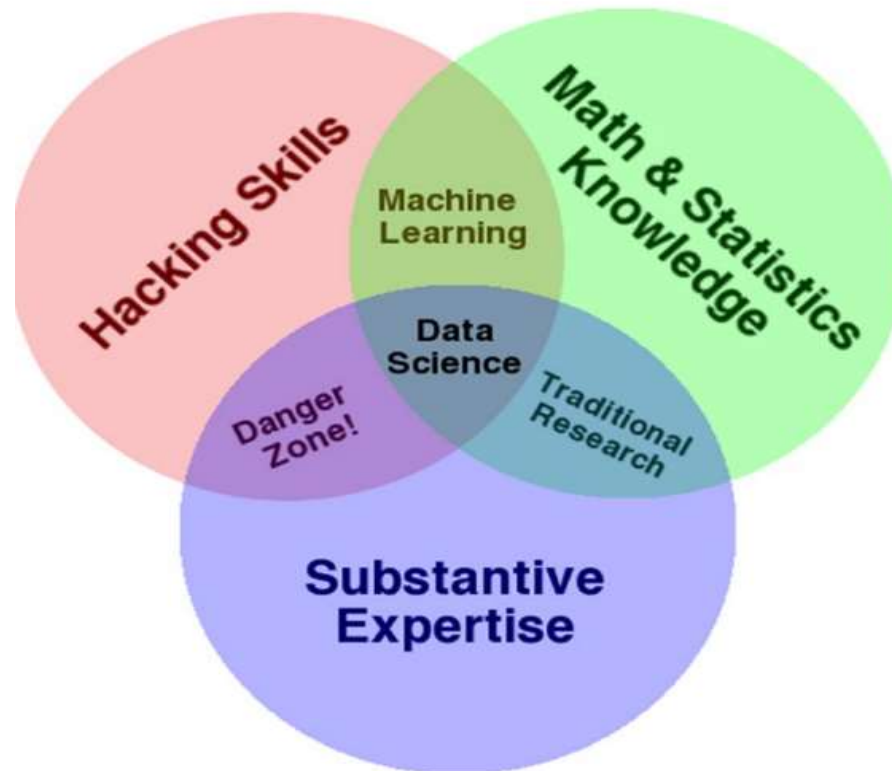
## **Applied or practical computer science:**

- Practical computer science also known as applied computer science is more about tools, methodologies, and processes that deal with applying concepts and principles from computer science in the real world to solve practical day-to-day problems.
- This includes emerging sub-fields like artificial intelligence, Machine Learning, computer vision, Deep Learning, natural language processing, data mining, and robotics and they try to solve complex real-world problems based on multiple constraints and parameters and try to emulate tasks that require considerable human intelligence and experience
- Besides these, we also have well established fields, including computer architecture, operating systems, digital logic and design, distributed computing, computer networks, security, databases, and software engineering.



# Data Science

- Data Science basically deals with principles, methodologies, processes, tools, and techniques to gather knowledge or information from data



# Data Science

Basically there are three major components and Data Science sits at the intersection of them.

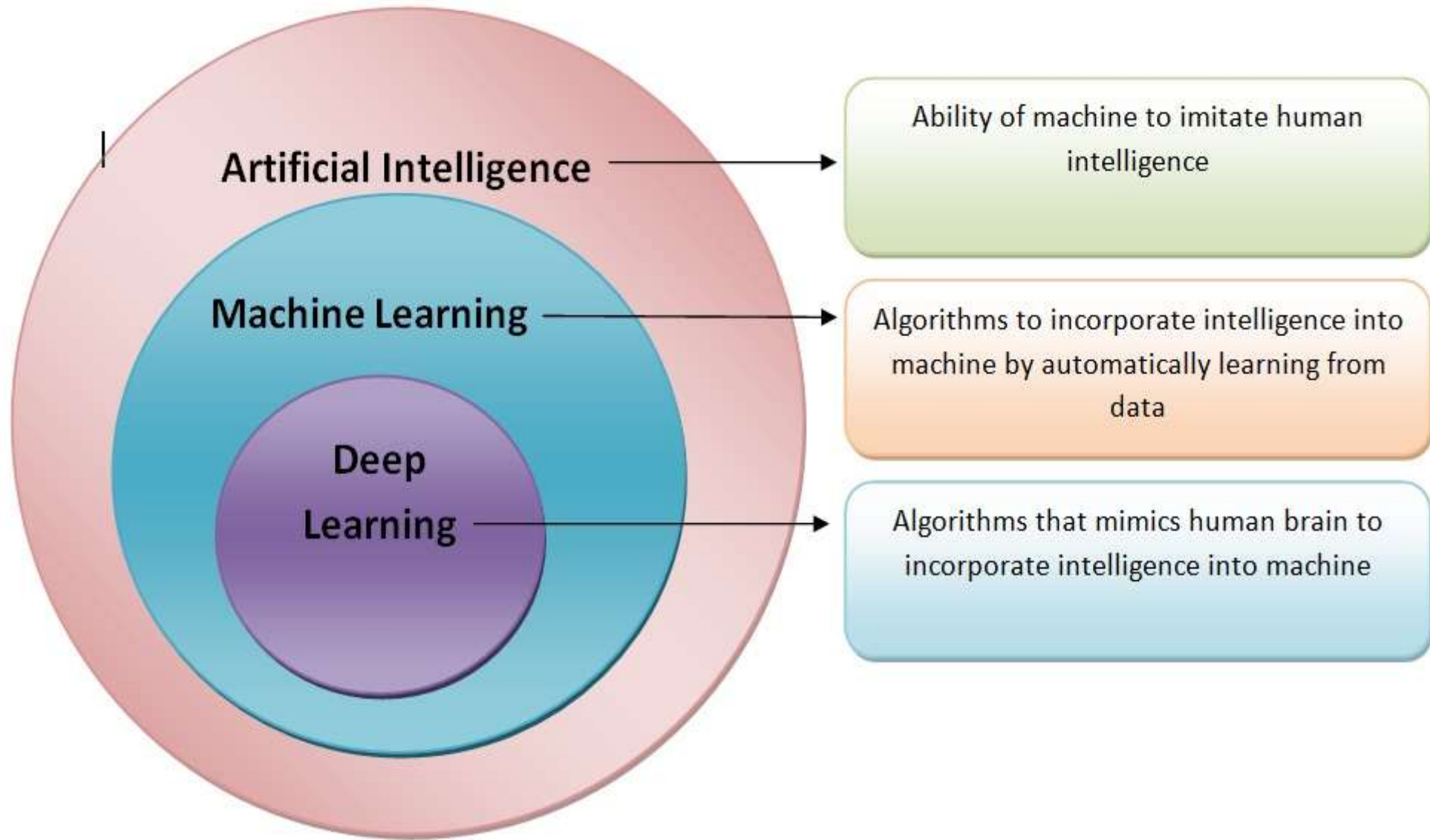
- **Math and statistics** knowledge is all about applying various computational and quantitative math and statistical based techniques to extract insights from data.
- **Hacking skills** basically indicate the capability of handling, processing, manipulating and wrangling data into easy to understand and analyzable formats.
- **Substantive expertise** is basically the actual real-world domain expertise which is extremely important when you are solving a problem because you need to know about various factors, attributes, constraints, and knowledge related to the domain besides your expertise in data and algorithms.

# Artificial Intelligence

- Artificial Intelligence is composed of two words “**Artificial**” and “**Intelligence**”, where Artificial defines “**man-made**,” and intelligence defines “**thinking power**”, hence AI means “**a man-made thinking power**.”

## Definition:

"It is a branch of computer science by which we can create intelligent machines which can behave like a human, think like humans, and able to make decisions”.



# Importance of AI?

There are few main reason to learn AI:

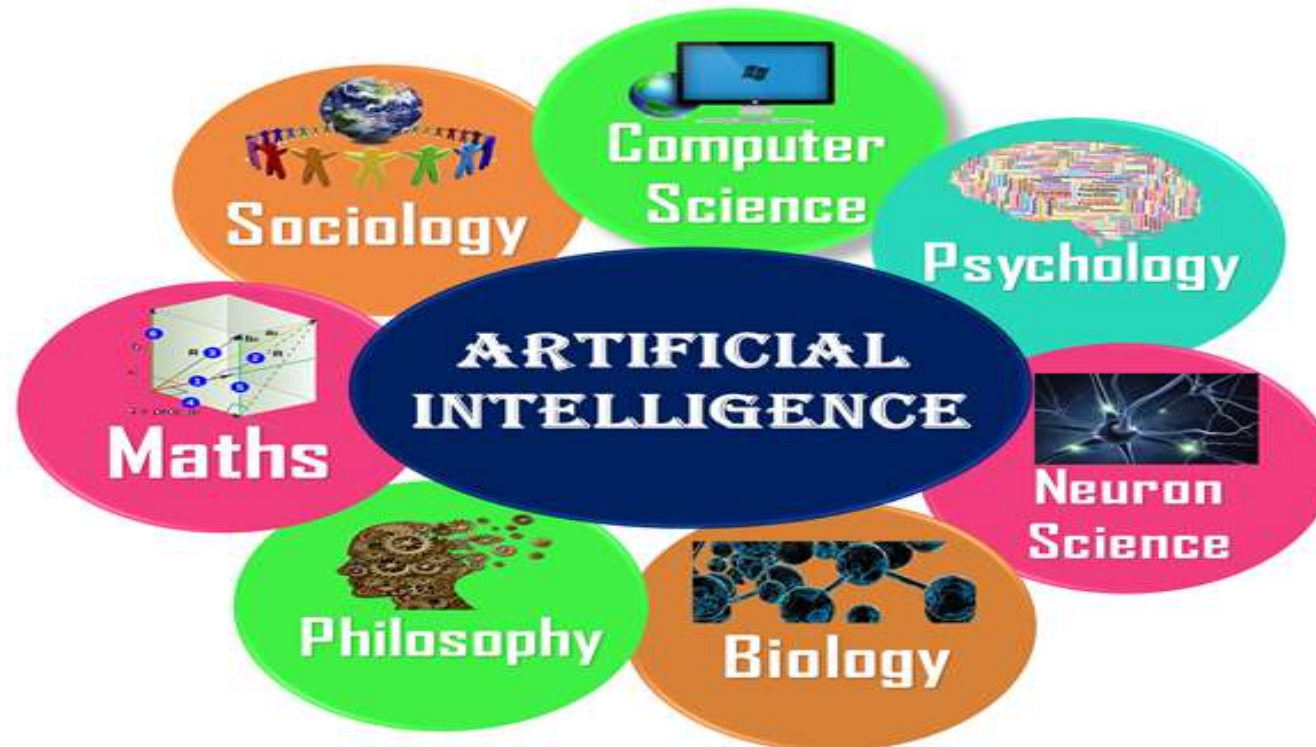
- With the help of AI, we can create such software which can solve real-world problems very easily and with accuracy such as health issues, marketing, traffic issues, etc.
- We can also create our personal virtual Assistant, such as Cortana, Google Assistant, Siri, etc.
- We can build Robots which can work in an environment where survival of humans can be at risk.

# Goals of AI:

1. Replicate human intelligence
2. Solve Knowledge-intensive tasks
3. An intelligent connection of perception and action
4. Building a machine which can perform tasks that requires human intelligence such as:
  1. Proving a theorem
  2. Playing chess
  3. Plan some surgical operation
  4. Driving a car in traffic
5. Creating some system which can exhibit intelligent behavior, learn new things by itself, demonstrate, explain, and can advise to its user.

# What Comprises to Artificial Intelligence?

- To create the AI, first we should know that how intelligence is composed, so the Intelligence is an intangible part of our brain which is a combination of “*Reasoning, learning, problem-solving, language, understanding,*” etc.
- To achieve the above factors for an AI machine requires the following disciplines:



# Advantages of AI:

- High Accuracy with less errors
- High speed and reliability
- Useful for risky areas
- Digital Assistant
- Useful as public utility

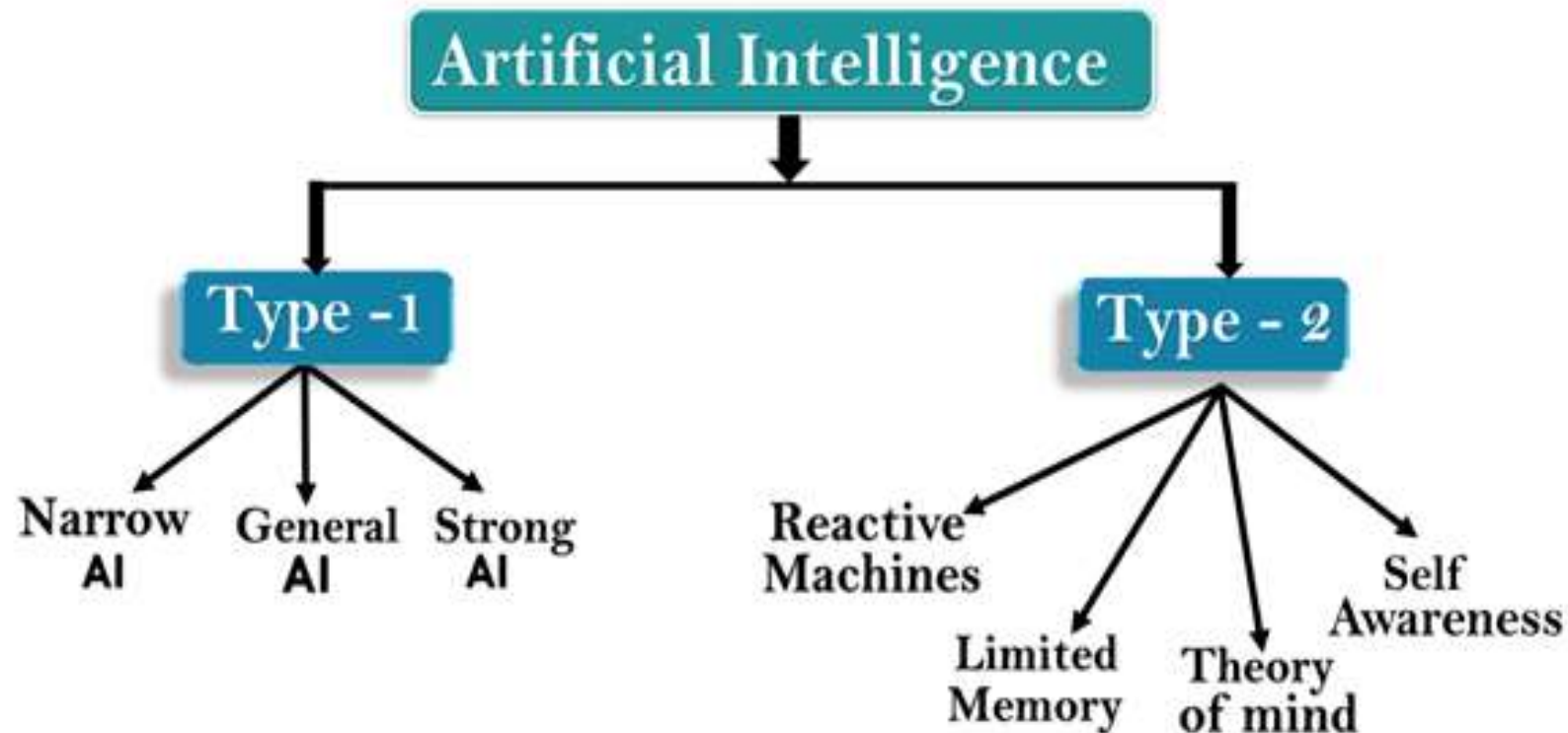


# Disadvantages of AI:

- High cost
- Can't think out of box
- No feelings and emotions
- Increase dependency on machines
- No original creativity

# Types of AI:

- Artificial Intelligence can be divided in two types based on capabilities and functionalities



# AI type 1: Based on Capability

## 1. **Weak or Narrow AI:**

- > only trained for one specific task
- > It can fail in unpredictable ways if it goes beyond its limits.

Example: Chat bots, Apple Siris,

## 2. **General AI:**

-> could perform any intellectual task with efficiency like a human.

-> The worldwide researchers are now focused on developing machines with General AI

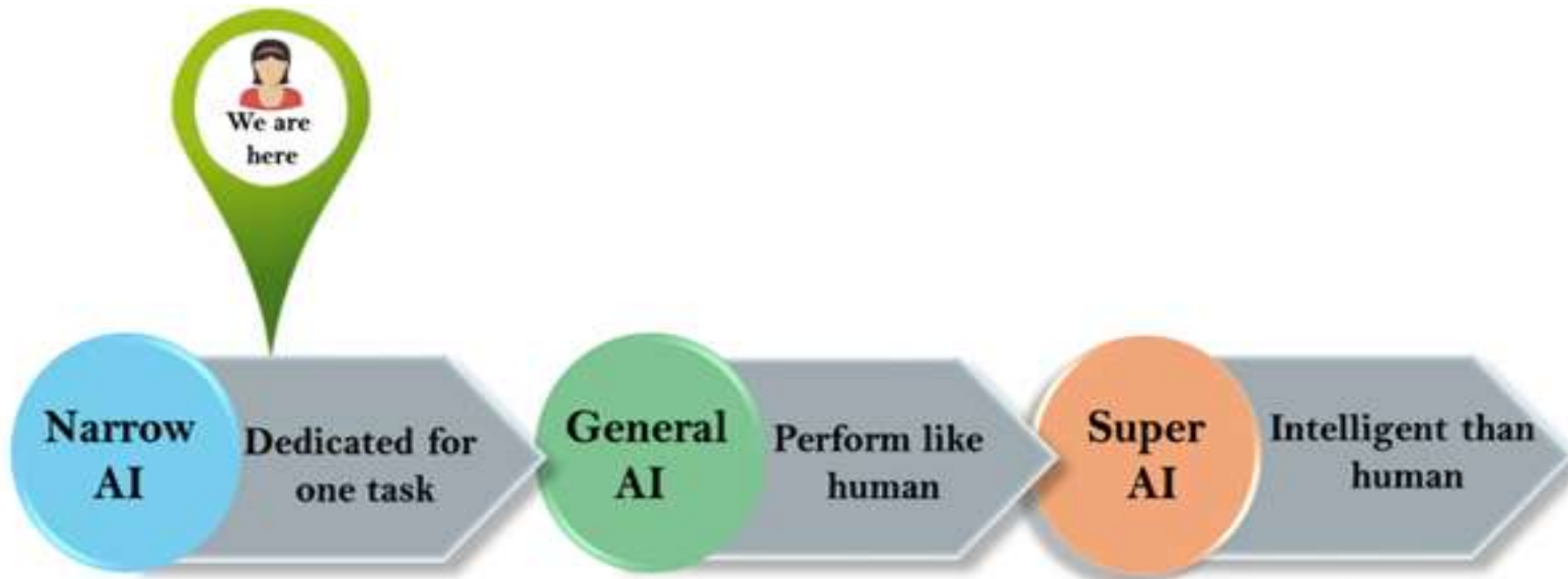
Example: Robot

# AI type 1: Based on Capability

## 3. Super AI:

-> Machines can perform any task better than human with cognitive properties.

Example: Terminator movie.



# AI type 2: Based on functionality

## 1. Reactive Machines:

- Reactive machines do not store memories or past experiences for future actions.
- Only focus on current scenarios and react on it as per possible best action.
- Example: IBM's Deep Blue, AlphaGo

## 2. Limited Memory:

- Limited memory machines can store past experiences or some data for a short period of time.
- Example: Self-driving cars. These cars can store recent speed of nearby cars, the distance of other cars, speed limit, and other information to navigate the road

# AI type 2: Based on functionality

## **3. Theory of Mind:**

- Should understand the human emotions, people, beliefs, and be able to interact socially like humans.
- This type of AI machines are still not developed, but researchers are making lots of efforts.

## **4. Self – Awareness:**

- These machines will be super intelligent, and will have their own consciousness, sentiments, and self-awareness
- Self-Awareness AI does not exist in reality still

# Natural Language Processing

- Natural language processing (NLP) is the ability of a computer program to understand human language as it is spoken and written -- referred to as natural language. It is a component of artificial intelligence

## How does natural language processing work?

- NLP enables computers to understand natural language as humans do.
- Whether the language is spoken or written, natural language processing uses artificial intelligence to take real-world input, process it, and make sense of it in a way a computer can understand.
- Just as humans have different sensors -- such as ears to hear and eyes to see -- computers have programs to read and microphones to collect audio

# Natural Language Processing

There are two main phases to NLP

- 1. Data preprocessing:** Data preprocessing involves preparing and "cleaning" text data for machines to be able to analyze it.. There are several ways this can be done, including:
  - > Tokenization: text is broken down into smaller units
  - > Stop word: common words are removed from text
  - > Part-of-speech tagging: words are marked based on the part-of speech



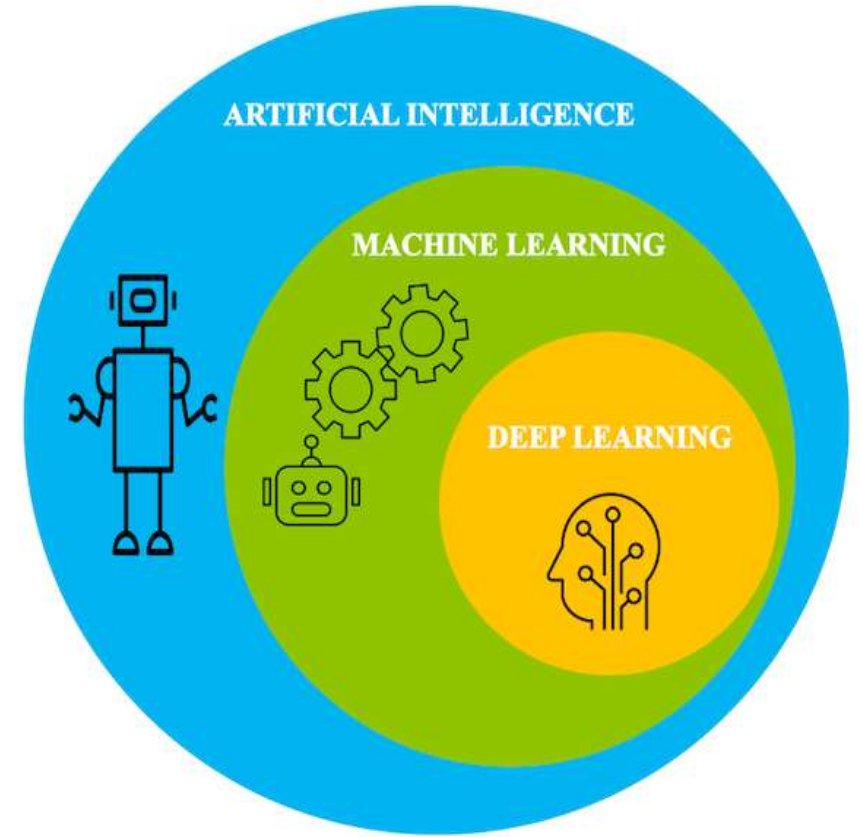
# Natural Language Processing

## 2. Algorithm Development:

- > Rules-based system
- > Machine learning-based system

# Deep Learning

- Deep learning is a subdomain of machine learning.
- With accelerated computational power and large data sets, deep learning algorithms are able to self-learn hidden patterns within data to make predictions.
- Deep learning architecture contains a computational unit that allows modeling of nonlinear functions called *perceptron*.



# Deep learning and human brain

- Generally, how a "neuron" in a human brain transmits electrical pulses throughout our nervous system, the perceptron receives a list of input signals and transforms them into output signals.
- The perceptron aims to understand data representation by stacking together many layers, where each layer is responsible for understanding some part of the input.
- A network of these perceptron mimics how neurons in the brain form a network, so the architecture is called neural networks (or artificial neural networks).

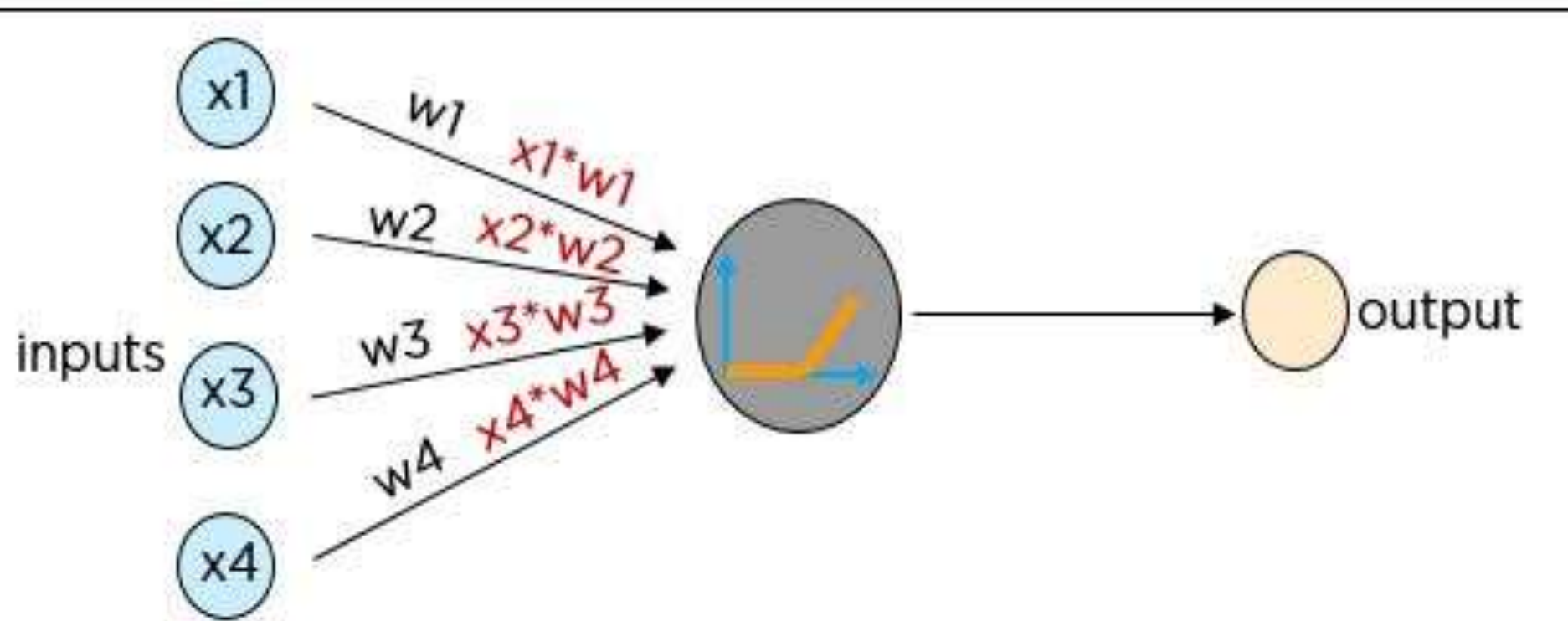
# What are Neural Networks?

- A neural network is a system modeled on the human brain, consisting of an input layer, multiple hidden layers, and an output layer.
- Data is fed as input to the neurons. The information is transferred to the next layer using appropriate weights and biases.
- The output is the final value predicted by the artificial neuron.

# What are Neural Networks?

Each neuron in a neural network performs the following operations:

- The product of each input and the weight of the channel it is passed over is found
- The sum of the weighted products is computed, which is called the weighted sum
- A bias value of the neuron is added to the weighted sum
- The final sum is then subjected to a particular function known as the *activation function*.



$$x_1 * w_1 + x_2 * w_2 + x_3 * w_3 + x_4 * w_4 + \text{bias} \rightarrow \text{final sum}$$

Activation function ( final sum)

# Activation function?

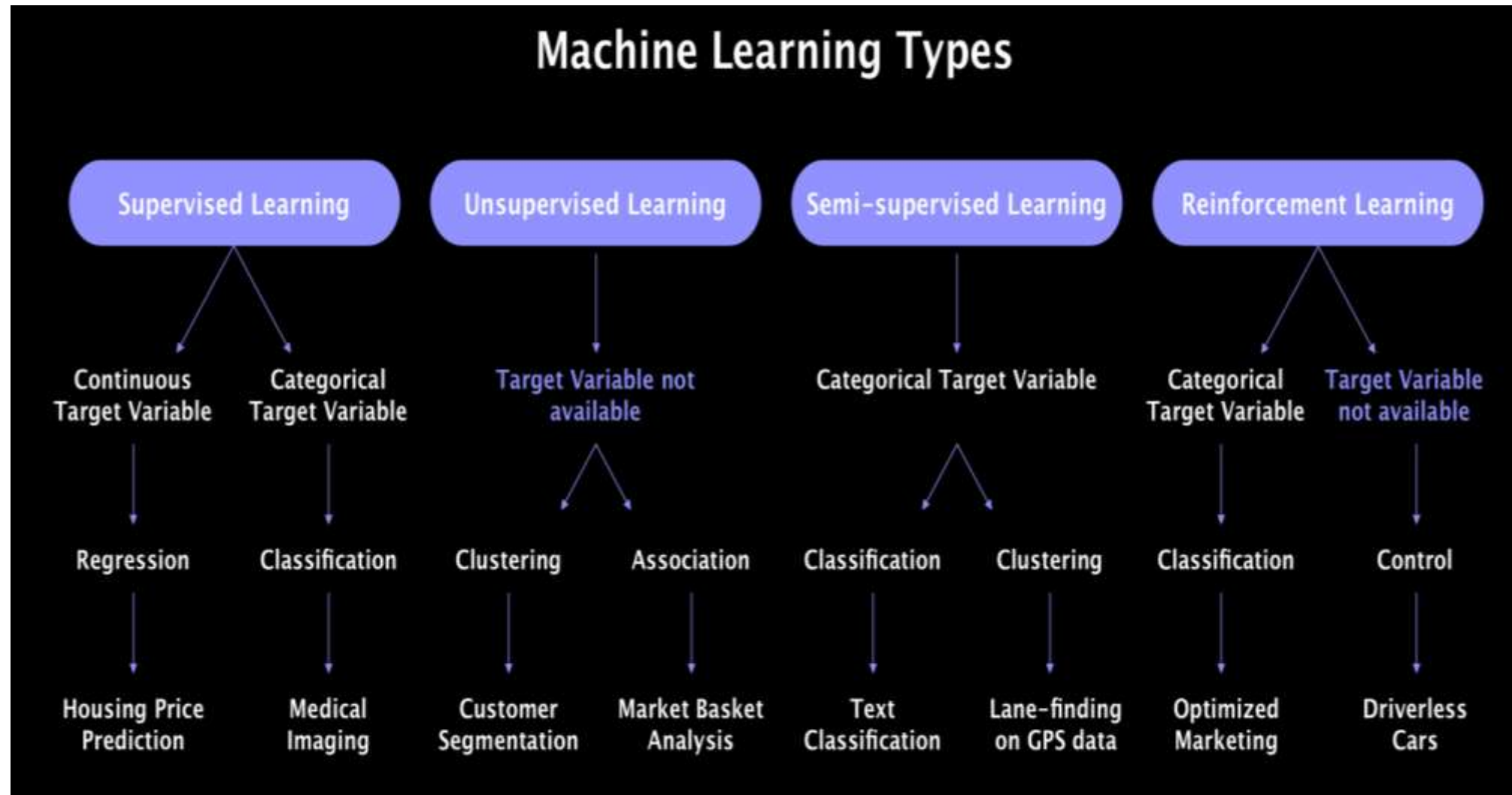
- Activation function decides, whether a neuron should be activated or not by calculating weighted sum and further adding bias with it.
- In a neural network, we would update the weights and biases of the neurons on the basis of the error at the output. This process is known as *back-propagation*.
- Activation functions make the back-propagation possible since the *gradients* are supplied along with the error to update the weights and biases.

# Machine Learning Methods

- Machine learning methods are classified under some categories such as
  1. Methods based on the amount of human supervision in the learning process
    - a. Supervised learning
    - b. Unsupervised learning
    - c. Semi-supervised learning
    - d. Reinforcement learning
  2. Methods based on the ability to learn from incremental data samples
    - a. Batch learning
    - b. Online learning
  3. Methods based on their approach to generalization from data samples
    - a. Instance based learning
    - b. Model based learning

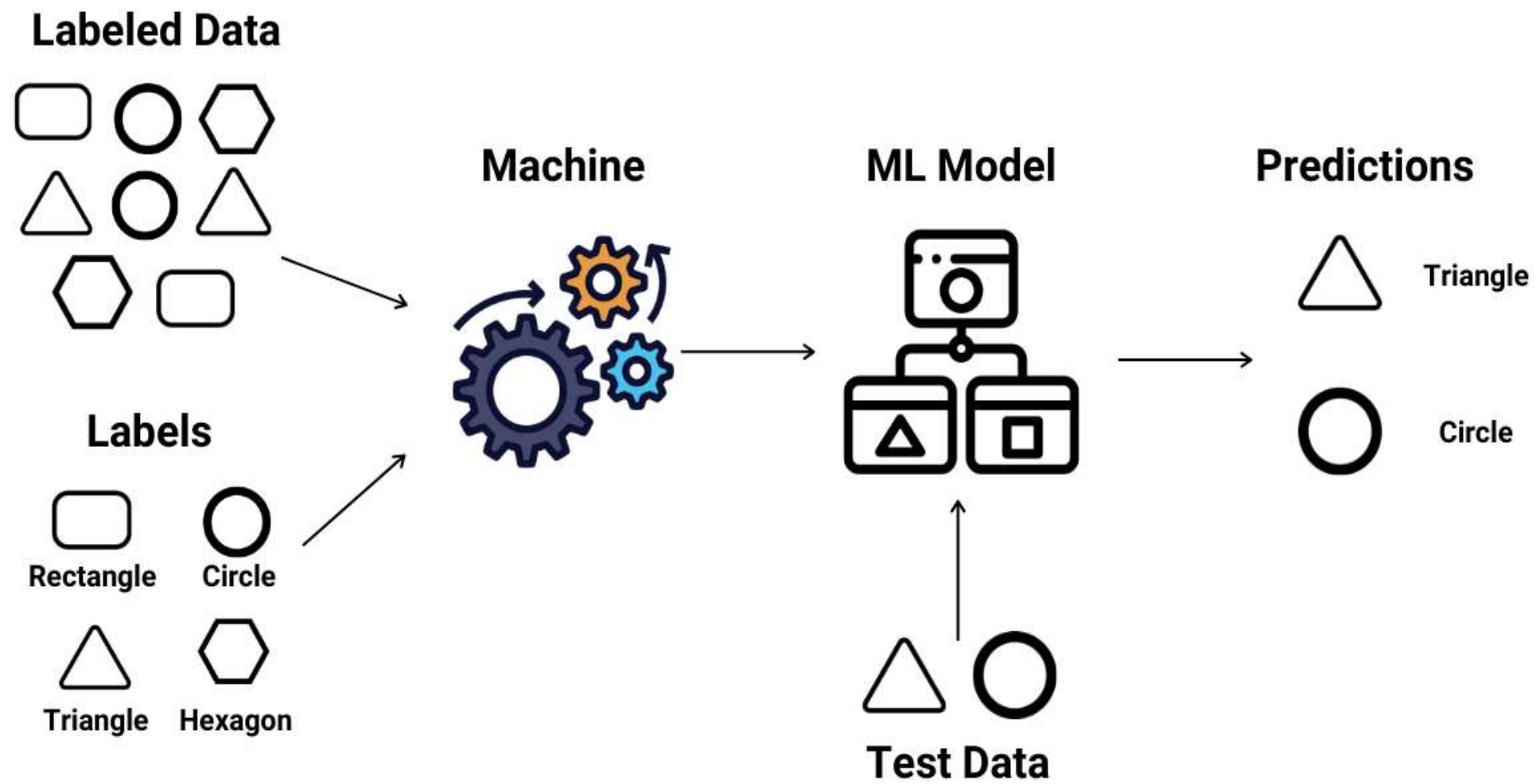


# 1. Methods based on the amount of human supervision in the learning process



# 1. Supervised Learning

- Supervised learning is used to identify the relationship between the input and output variables and then use it to map new unlabelled data.
- Types:
  1. classification: used to predict a categorical or nominal variable  
Algorithms: KNN, Logistic Regression, SVM, Decision Trees, Random Forest, Naïve Bayes etc.
  2. Regression: used to predict real – valued or continuous variable.  
Algorithms: Linear regression, SVR, Decision Trees, Random Forest etc.



# Classification:

- Classification is a task that requires the use of machine learning algorithms that learn how to assign a class label to examples from the problem domain. An easy to understand example is classifying emails as “*spam*” or “*not spam*.”
- In this classification we have many other types some of them are:
  - Binary classification
  - Multi – class classification

# Binary classification

- In machine learning, binary classification is a supervised learning algorithm that categorizes new observations into one of **two** classes.
- The following are a few binary classification applications, where the 0 and 1 columns are two possible classes for each observation:

Application	Observation	0	1
Medical Diagnosis	Patient	Healthy	Diseased
Email Analysis	Email	Not Spam	Spam
Financial Data Analysis	Transaction	Not Fraud	Fraud
Marketing	Website visitor	Won't Buy	Will Buy
Image Classification	Image	Hotdog	Not Hotdog

- Popular algorithms that can be used for binary classification include:
  - Logistic Regression, k-Nearest Neighbors, Decision Trees, Support Vector Machine, Naive Bayes

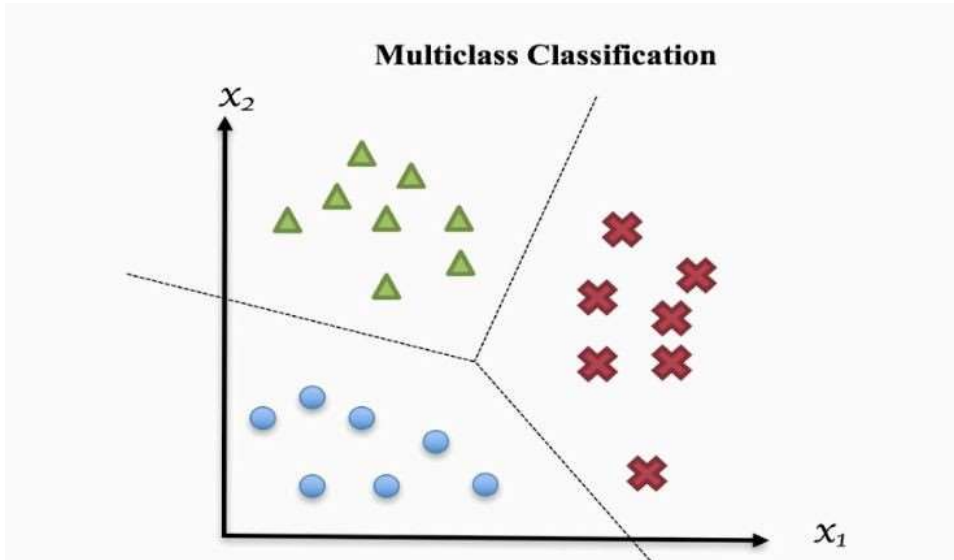
# Evaluation of binary classifiers

- If the model successfully predicts the patients as positive, this case is called *True Positive (TP)*.
- If the model successfully predicts patients as negative, this is called *True Negative (TN)*.
- The binary classifier may misdiagnose some patients as well. If a diseased patient is classified as healthy by a negative test result, this error is called *False Negative (FN)*.
- Similarly, If a healthy patient is classified as diseased by a positive test result, this error is called *False Positive (FP)*.
- After obtaining these values, we can compute the **accuracy score** of the binary classifier as follows:

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

# Multi class classification

- Multi class classification is the task of classifying elements into different classes.



Examples include:

- Face classification, Plant species classification, Optical character recognition.
- Popular algorithms that can be used for multi-class classification include:  
k-Nearest Neighbors, Decision Trees, Naive Bayes, Random Forest, Gradient Boosting.

# Real Life examples for supervised ML

- Text categorization
- Face Detection
- Signature recognition
- Customer discovery
- Spam detection
- Weather forecasting
- Predicting housing prices based on the prevailing market price
- Stock price predictions, among others



# Regression

- Regression is a supervised learning algorithm which helps in finding the correlation between variables and enables us to predict the continuous output variable based on the one or more predictor variables.
- It is mainly used for **prediction, forecasting, time series modeling, and determining the causal-effect relationship between variables.**
- Some examples are:
  - Prediction of rain using temperature and other factors,
  - Determining Market trends,
  - Prediction of road accidents due to rash driving.

# Types of Linear Regression

Linear regression can be further divided into two types of the algorithm:

- **Simple Linear Regression:**

If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

- **Multiple Linear regression:**

If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

## 2. Unsupervised Learning

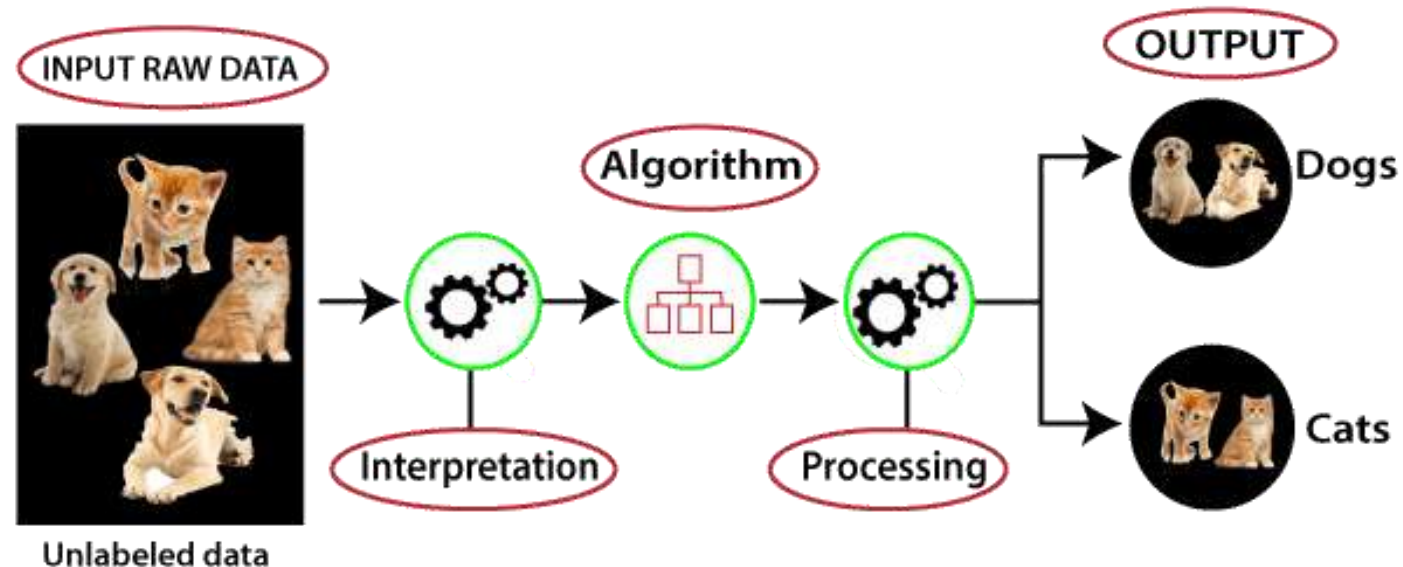
- In which only have the input data to feed to the model but no corresponding output data. There are 2 categories:

1. Clustering: group or organize similar objects together.

Algorithms: K means, DBSCAN, Mean- shift algorithm etc.

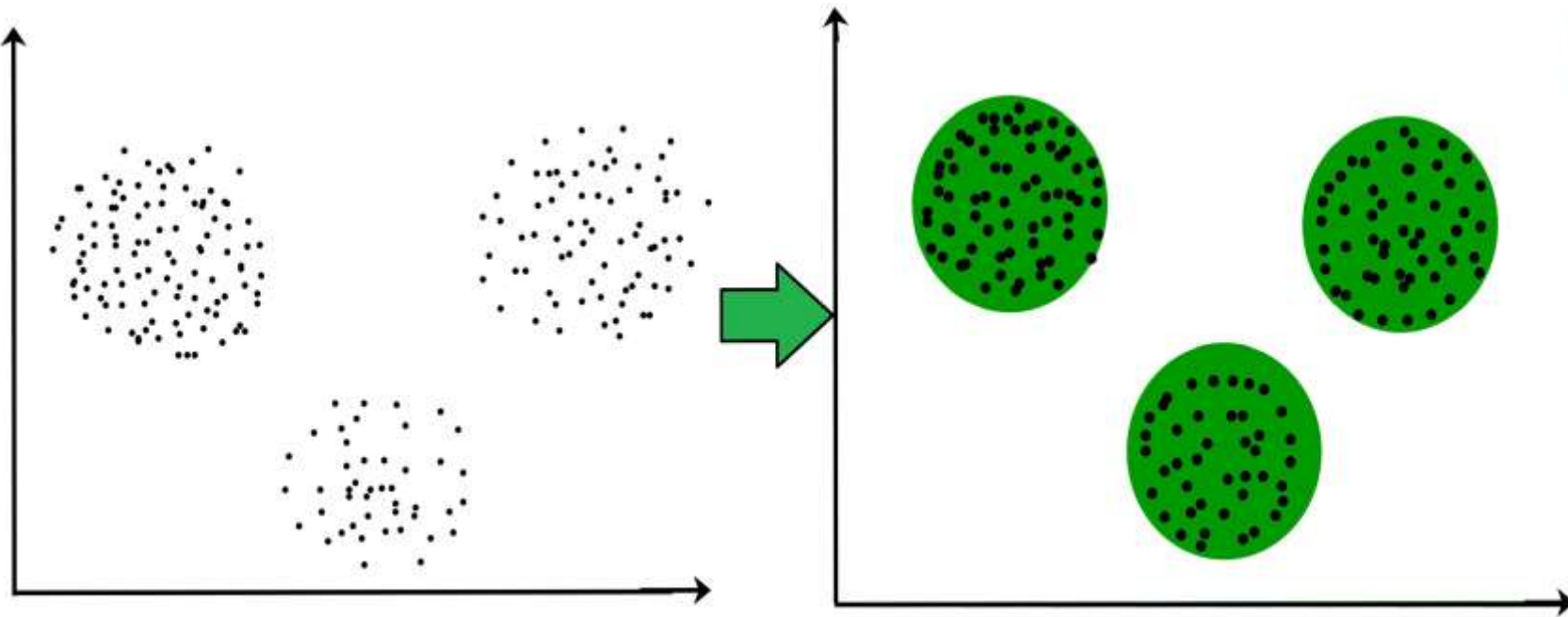
2. Association: the association between data elements is identified.

Algorithms: Apriori and FP growth



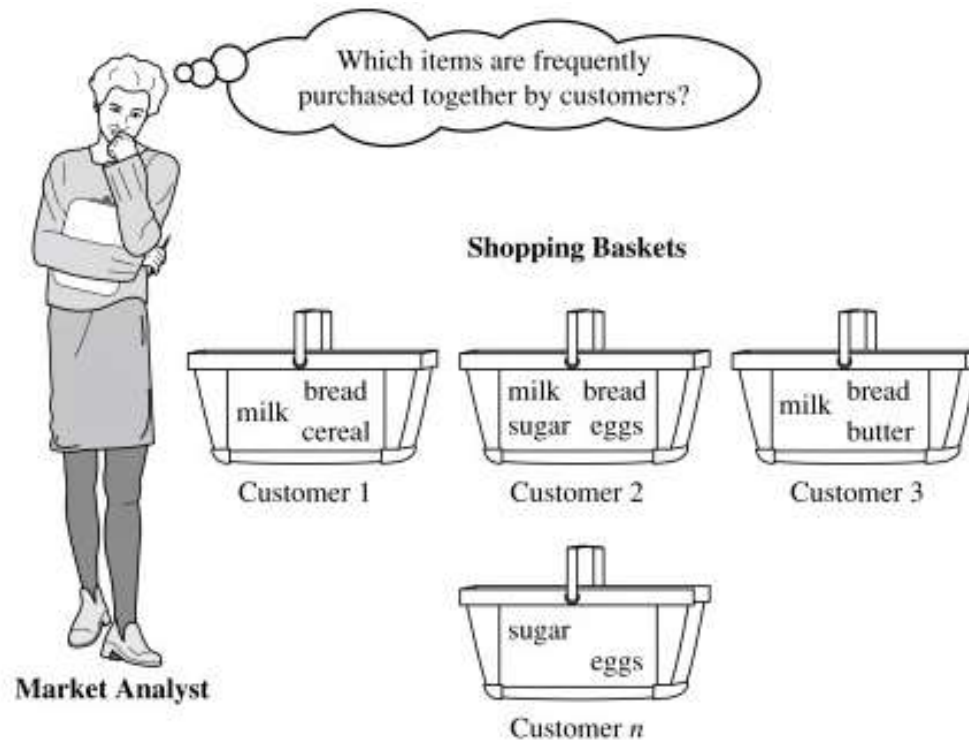
# Clustering :

- Clustering is a way of grouping the data points into different clusters, consisting of similar data points.
- The objects with the possible similarities remain in a group that has less or no similarities with another group.

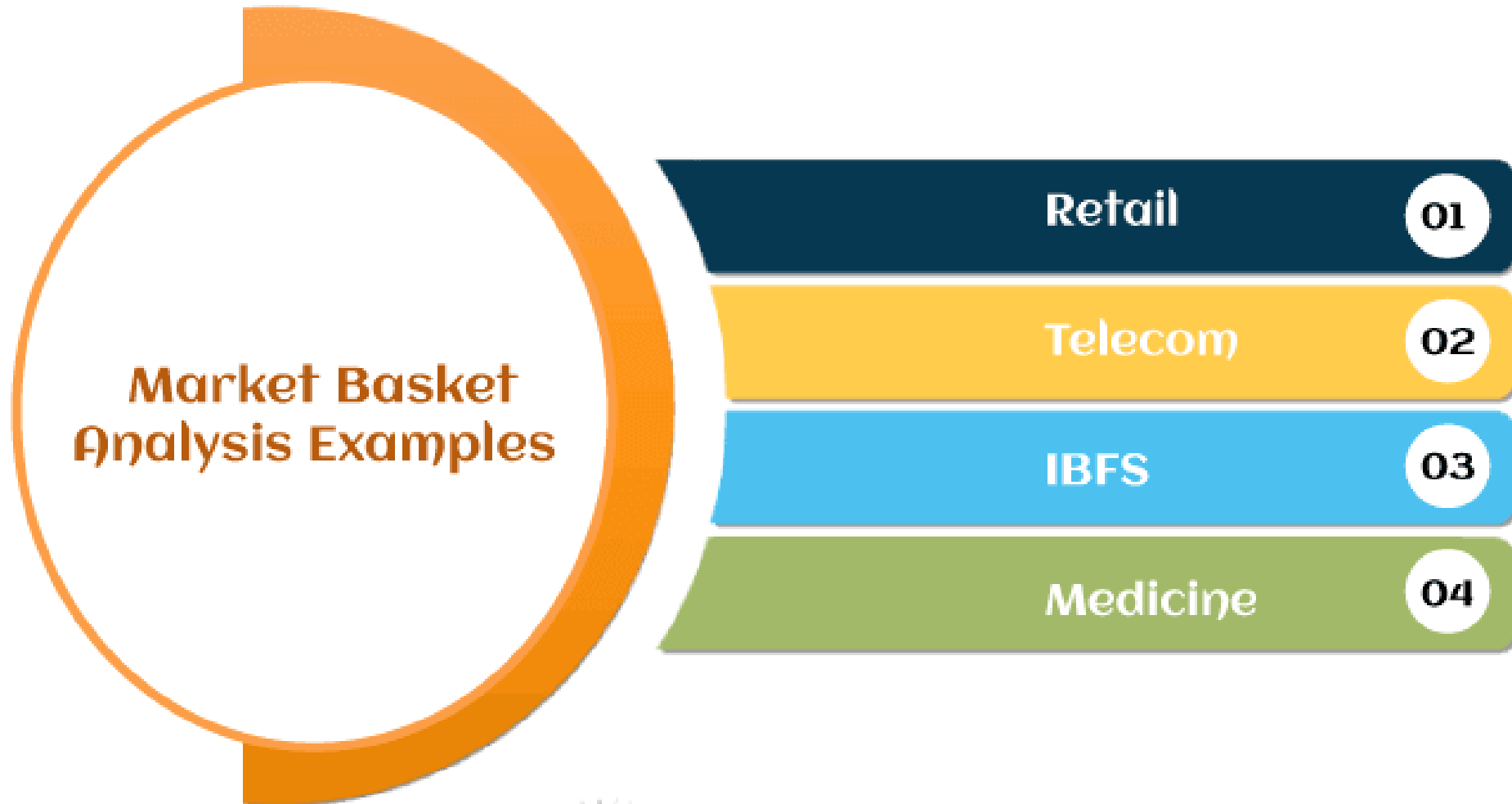


# Association – Market Basket Analysis

- Market basket analysis is a data mining technique used by retailers to increase sales by better understanding customer purchasing patterns. It involves analyzing large data sets, such as purchase history, to reveal product groupings and products that are likely to be purchased together.



# Examples for market basket analysis



# Real Life examples for Unsupervised ML

- Audience segmentation.
- Customer personality investigation.
- Anomaly detection (for example, to detect bot activity)
- Pattern recognition (grouping images, transcribing audio)
- Inventory management (by conversion activity or by availability)

### 3. Semi supervised learning

- The most basic disadvantage of any **Supervised Learning** algorithm is that the dataset has to be hand-labeled either by a Machine Learning Engineer or a Data Scientist.
- This is a very *costly process*, especially when dealing with large volumes of data. The most basic disadvantage of any **Unsupervised Learning** is that it's **application spectrum is limited**.
- To counter these disadvantages, the concept of **Semi-Supervised Learning** was introduced. In this type of learning, the algorithm is trained upon a combination of labeled and unlabeled data.



### 3. Semi supervised learning (cont..)

- Typically, this combination will contain a very small amount of labeled data and a very large amount of unlabeled data.
- **The basic procedure involved is that first, the programmer will cluster similar data using an unsupervised learning algorithm and then use the existing labeled data to label the rest of the unlabeled data.**
- The typical use cases of such type of algorithm have a common property among them – The acquisition of unlabeled data is relatively cheap while labeling the said data is very expensive.

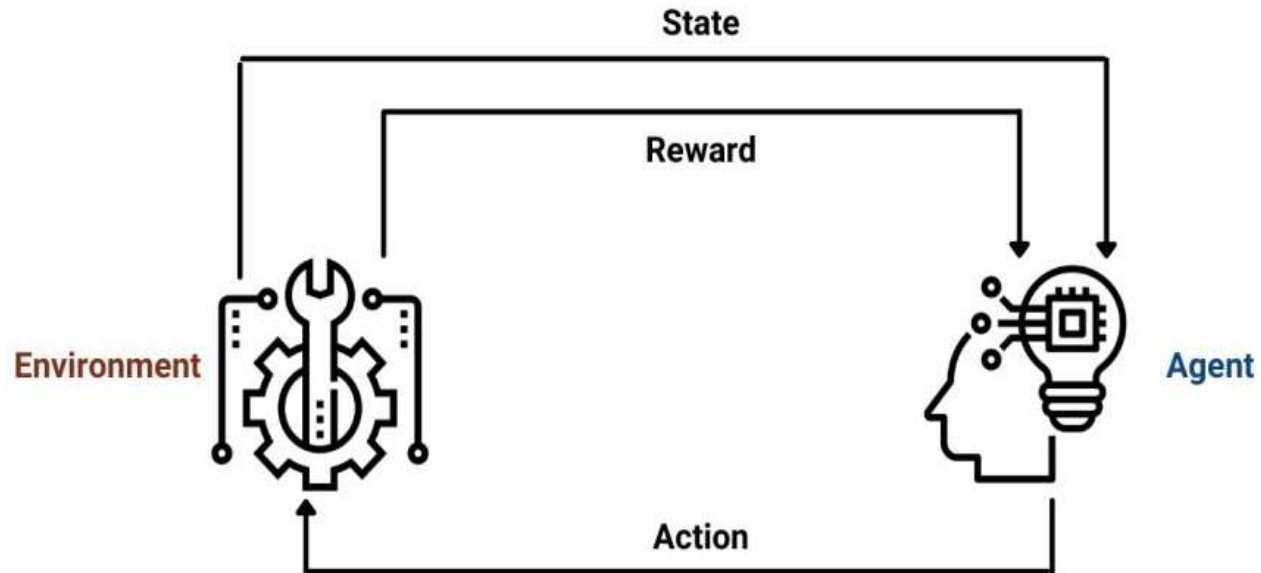
### 3. Semi supervised learning (cont..)

Intuitively, one may imagine the three types of learning algorithms as:

- **Supervised learning** where a student is under the supervision of a teacher at both home and school,
- **Unsupervised learning** where a student has to figure out a concept himself and
- **Semi-Supervised learning** where a teacher teaches a few concepts in class and gives questions as homework which are based on similar concepts.

# 4. Reinforcement Learning

- Reinforcement Learning is a feedback-based Machine learning technique in which an agent learns to behave in an environment by performing the actions and seeing the results of actions. For each good action, the agent gets positive feedback, and for each bad action, the agent gets negative feedback or penalty.
- Algorithms: Q – learning, Sarsa

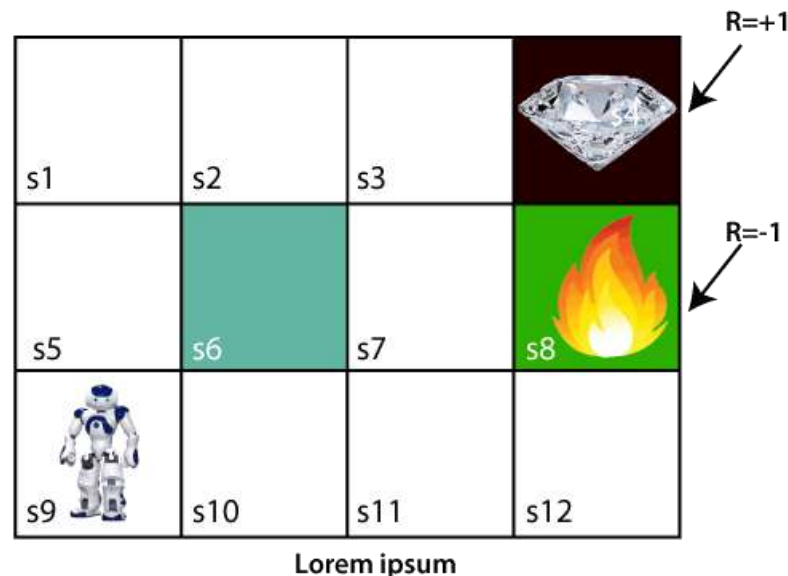


# Terms used in Reinforcement Learning

- **Agent():** An entity that can perceive/explore the environment and act upon it.
- **Environment():** A situation in which an agent is present or surrounded by. In RL, we assume the stochastic environment, which means it is random in nature.
- **Action():** Actions are the moves taken by an agent within the environment.
- **State():** State is a situation returned by the environment after each action taken by the agent.
- **Reward():** A feedback returned to the agent from the environment to evaluate the action of the agent.
- **Policy():** Policy is a strategy applied by the agent for the next action based on the current state.
- **Value():** It is expected long-term return with the discount factor and opposite to the short-term reward.


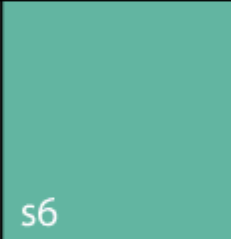


# How does Reinforcement Learning Work?

- To understand the working process of the RL, we need to consider two main things:
- **Environment:** It can be anything such as a room, maze, football ground, etc.
- **Agent:** An intelligent agent such as AI robot.
- Let's take an example of a maze environment that the agent needs to explore.  
Consider an image:



- the agent is at the very first block of the maze. The maze is consisting of an  $S_6$  block, which is a **wall**,  $S_8$  a **fire pit**, and  $S_4$  a **diamond block**.
- The agent cannot cross the  $S_6$  block, as it is a solid wall. If the agent reaches the  $S_4$  block, then get the **+1 reward**; if it reaches the fire pit, then gets **-1 reward point**. It can take four actions: **move up, move down, move left, and move right**.
- The agent can take any path to reach to the final point, but he needs to make it in possible fewer steps. Suppose the agent considers the path **S9-S5-S1-S2-S3**, so he will get the +1-reward point.

- The agent will try to remember the preceding steps that it has taken to reach the final step. To memorize the steps, it assigns 1 value to each previous step. Consider the below step:
- Now, the agent has successfully stored the previous steps assigning the 1 value to each previous block.

<b>V=1</b> s1	<b>V=1</b> s2	<b>V=1</b> s3	 s4
<b>V=1</b> s5	 s6	s7	 s8
 <b>V=1</b> s9	s10	s11	s12

# Applications of Reinforcement Learning

- Robotics for industrial automation.
- Business strategy planning
- Machine learning and data processing
- It helps you to create training systems that provide custom instruction and materials according to the requirement of students.
- Aircraft control and robot motion control



# Examples in Reinforcement Learning

- **Personalized product recommendation system:** Personalize / customize what products need to be shown to individual users to realize maximum sale; This would be something ecommerce portals would love to implement to realize maximum click-through rates on any given product and related sales, on any given day
- **Customized action in video games** based on reinforcement learning; AI agents use reinforcement learning to coordinate actions and react appropriately to new situations through a series of rewards

# Examples in Reinforcement Learning

- **RL in healthcare** can be used to **recommend different treatment options**. While supervised learning models can be used to predict whether a person is suffering from a disease or not, RL can be used to predict treatment options given a person is suffering from a particular disease.
- **RL can be used for NLP use cases** such as text summarization, question & answers, machine translation.
- **AI-powered stock buying/selling**: While supervised learning algorithms can be used to predict the stock prices, it is the reinforcement learning which can be used to decide whether to buy, sell or hold the stock at given predicted price.

## 2. Methods based on the ability to learn from incremental data samples

### 1. **Batch or Offline learning:**

- Offline learning refers to situations where the program is not operating and taking in new information in real-time
- So the model doesn't keep learning over a period of time continuously with the new data. Once the training is complete the model stops learning.
- We can always train the model on new data but then we would have to add new data samples along with the older historical training data and again re-build the model using this new batch of data.
- Used in applications where data patterns remain constant and don't have sudden concept drifts (e.g., Netflix recommendation system)

## **2. Online Learning:**

- Online learning is ideal for machine learning systems that receive data as a continuous flow and need to be able to adapt to rapidly changing conditions.
- More computational power is required because of the continuous feed of data that leads to continuous refinement.
- Harder to implement and control because the production model changes in real-time according to its data feed.
- Used in applications where new data patterns are constantly required (e.g., weather prediction tools)

### 3. Methods based on their approach to generalization from data samples

#### 1. **Instance Based Learning:**

- **instance-based learning** are the systems that learn the training examples by heart and then generalizes to new instances based on some similarity measure
- It is also known as **memory-based learning** or **lazy-learning**.
- Example: K-Nearest Neighbours.
- Some of the instance-based learning algorithms are :
  - 1.K Nearest Neighbor (KNN)
  - 2.Self-Organizing Map (SOM)
  - 3.Learning Vector Quantization (LVQ)
  - 4.Locally Weighted Learning (LWL)

## **2. Model Based Learning:**

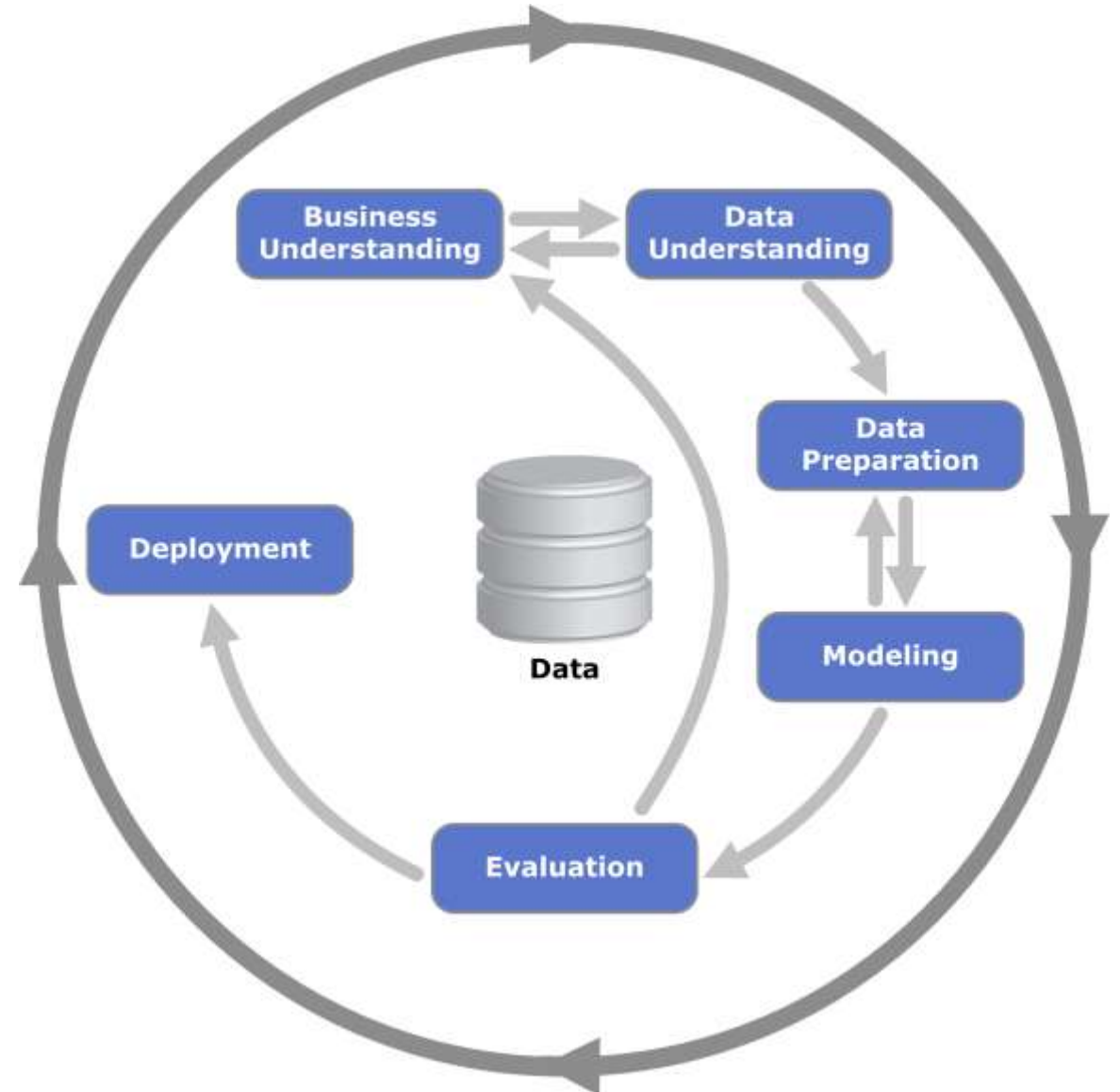
- The model based learning methods are a more traditional ML approach toward generalizing based on training data.
- Typically an iterative process takes place where the input data is used to extract features and models are built based on various model parameters (known as hyperparameters).
- These hyperparameters are optimized based on various model validation techniques to select the model that generalizes best on the training data and some amount of validation and test data (split from the initial dataset).
- Finally, the best model is used to make predictions or decisions as and when needed.

# Model vs Instance based learning

Model based learning	Instance based learning
Prepare the data for model training	Prepare the data for model training
Train model from training data to estimate model parameters i.e discover patterns	Do not train model
Store the model in suitable form	There is no model to store
Can throw away input/training data after model training	Input/training data must be kept since each query uses part or full set of training observations
Storing models generally requires less storage	Storing training data generally requires more storage

# CRISP-DM Process Model

The **C**Ross Industry Standard Process for **D**ata **M**ining (*CRISP-DM*) is a process model that serves as the base for a data science process.





# CRISP-DM Process Model

- It has six sequential phases:
  1. Business understanding – What does the business need?
  2. Data understanding – What data do we have / need? Is it clean?
  3. Data preparation – How do we organize the data for modeling?
  4. Modeling – What modeling techniques should we apply?
  5. Evaluation – Which model best meets the business objectives?
  6. Deployment – How do stakeholders access the results?

# Business Understanding

- The *Business Understanding* phase focuses on understanding the objectives and requirements of the project.
1. **Determine business objectives:** You should first “thoroughly understand, from a business perspective, what the customer really wants to accomplish
  2. **Assess situation:** Determine resources availability, project requirements, assess risks and conduct a cost-benefit analysis.
  3. **Determine data mining goals:** In addition to defining the business objectives, you should also define what success looks like from a technical data mining perspective.
  4. **Produce project plan:** Select technologies and tools and define detailed plans for each project phase

# Data understanding

- It drives the focus to identify, collect, and analyze the data sets that can help you accomplish the project goals.
- 1. Collect initial data:** Acquire the necessary data and (if necessary) load it into your analysis tool.
  - 2. Describe data:** Examine the data and document its surface properties like data format, number of records, or field identities.
  - 3. Explore data:** Dig deeper into the data. Query it, visualize it, and identify relationships among the data.
  - 4. Verify data quality:** How clean/dirty is the data? Document any quality issues.

# Data Preparation

- 1.Select data:** Determine which data sets will be used and document reasons for inclusion/exclusion.
- 2.Clean data:** Often this is the lengthiest task. Without it, you'll likely fall victim to garbage-in, garbage-out. A common practice during this task is to correct, impute, or remove erroneous values.
- 3.Construct data:** Derive new attributes that will be helpful. For example, derive someone's body mass index from height and weight fields.
- 4.Integrate data:** Create new data sets by combining data from multiple sources.
- 5.Format data:** Re-format data as necessary. For example, you might convert string values that store numbers to numeric values so that you can perform mathematical operations.

# Modeling

- 1. Select modeling techniques:** Determine which algorithms to try (e.g. regression, neural net).
- 2. Generate test design:** Pending your modeling approach, you might need to split the data into training, test, and validation sets.
- 3. Build model:** As glamorous as this might sound, this might just be executing a few lines of code like “`reg = LinearRegression().fit(X, y)`”.
- 4. Assess model:** Generally, multiple models are competing against each other, and the data scientist needs to interpret the model results based on domain knowledge, the pre-defined success criteria, and the test design.

# Evaluation

*Evaluation* phase looks more broadly at which model best meets the business and what to do next. This phase has three tasks:

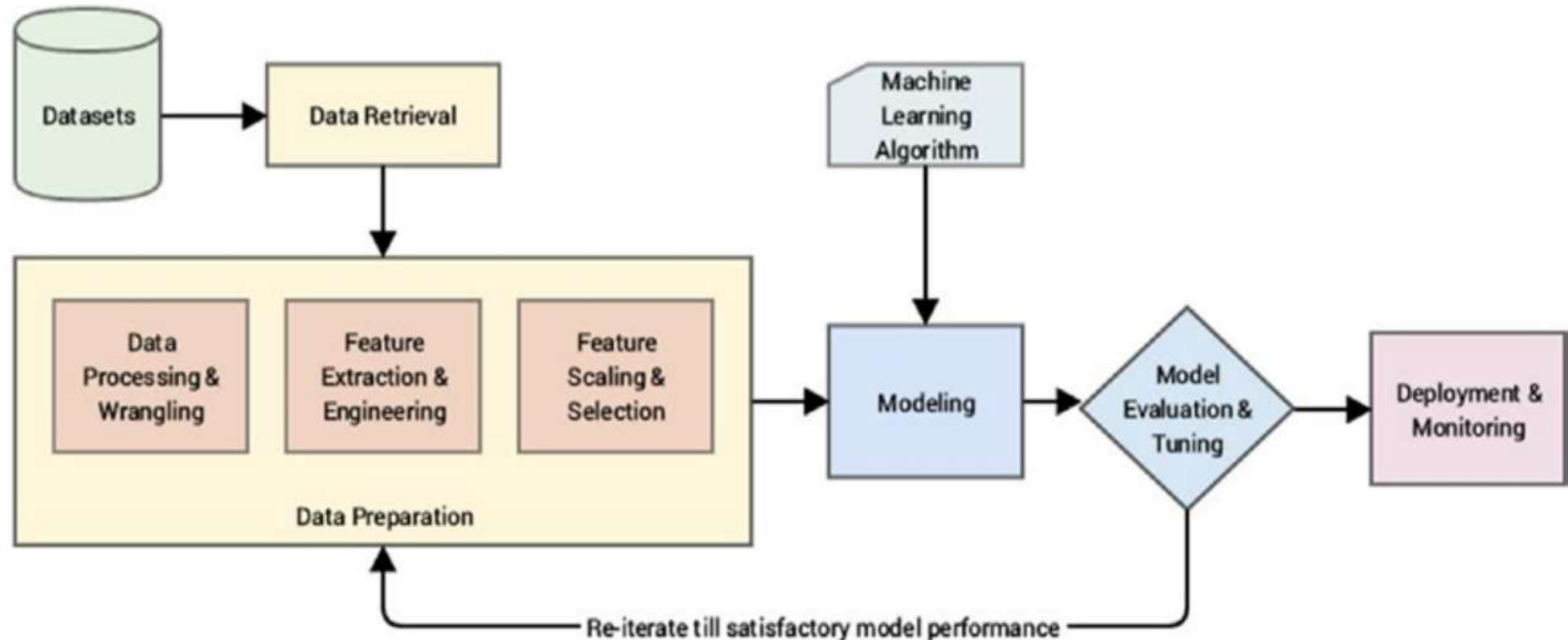
- 1.Evaluate results:** Do the models meet the business success criteria?  
Which one(s) should we approve for the business?
- 2.Review process:** Review the work accomplished. Was anything overlooked? Were all steps properly executed? Summarize findings and correct anything if needed.
- 3.Determine next steps:** Based on the previous three tasks, determine whether to proceed to deployment, iterate further, or initiate new projects.

# Deployment

- A model is not particularly useful unless the customer can access its results.
- 1. Plan deployment:** Develop and document a plan for deploying the model.
  - 2. Plan monitoring and maintenance:** Develop a thorough monitoring and maintenance plan to avoid issues during the operational phase (or post-project phase) of a model.
  - 3. Produce final report:** The project team documents a summary of the project which might include a final presentation of data mining results.
  - 4. Review project:** Conduct a project retrospective about what went well, what could have been better, and how to improve in the future.

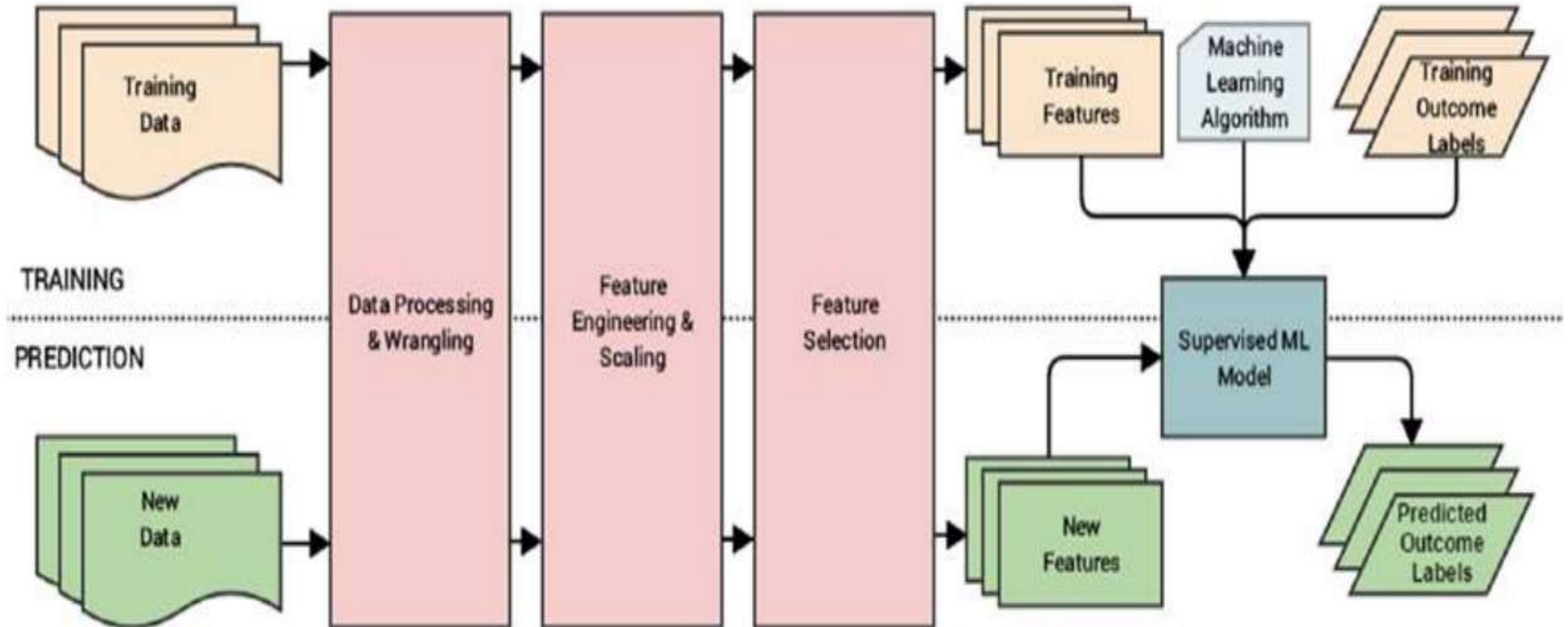
# Building Machine Intelligence

- **Machine Learning Pipe Lines:** A Machine Learning pipeline will mainly consist of elements related to data retrieval and extraction, preparation, modeling, evaluation, and deployment.





- Supervised Machine Learning Pipeline



- Unsupervised Machine Learning Pipeline:

