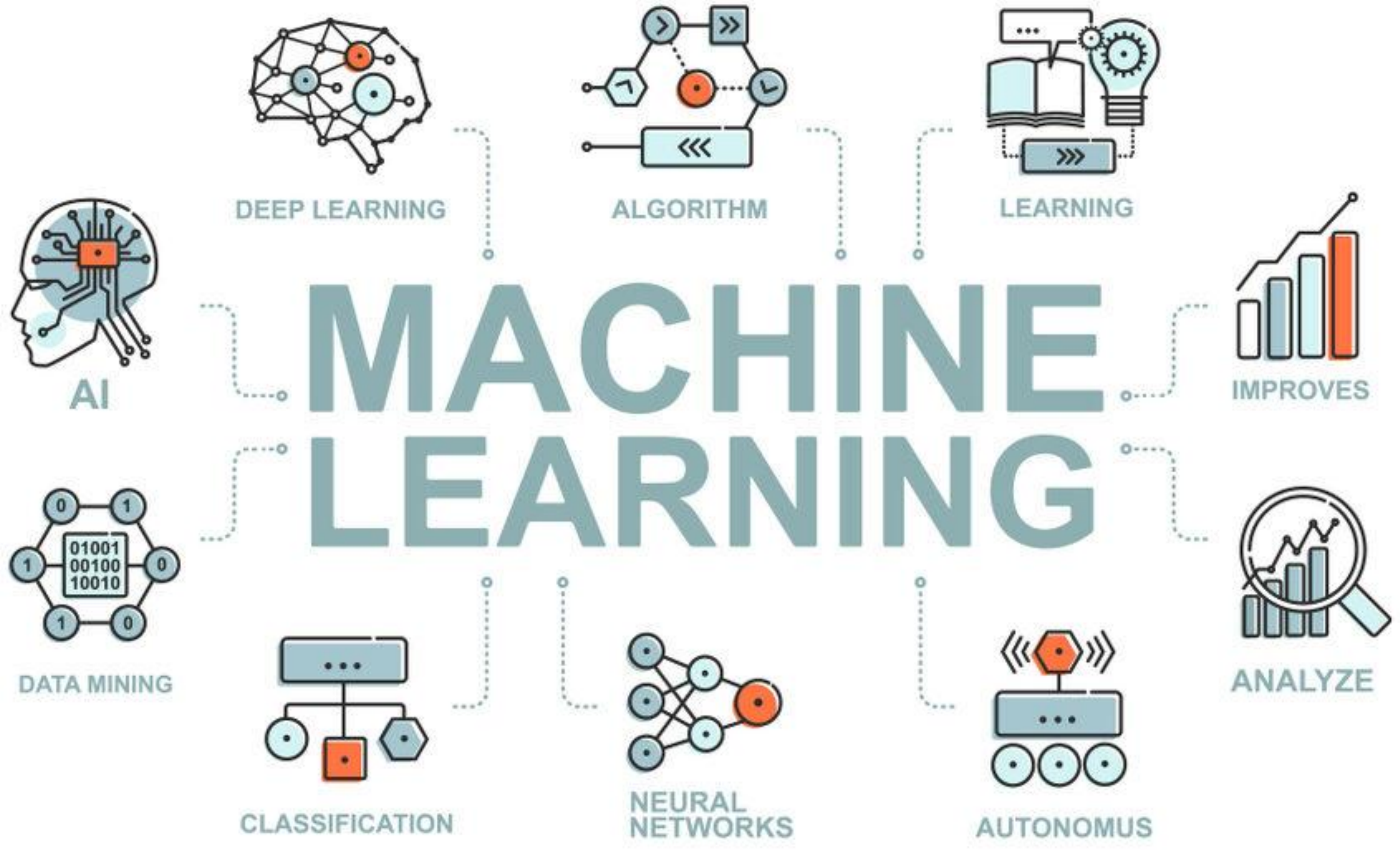
RV College of Engineering

Go, change the world

Machine Learning
UNIT I



Faculty Incharge
Dr Andhe Dharani
Dr S Anupama Kumar

Department of Master of Computer Applications

Introduction to Machine Learning

- *Introduction*
- *Human Learning*
- *Machine Learning*
- *Types of ML*
- *Problems not be solved using ML*
- *Applications of ML*
- *Languages / Tools in ML*
- *Issues in ML*

Introduction

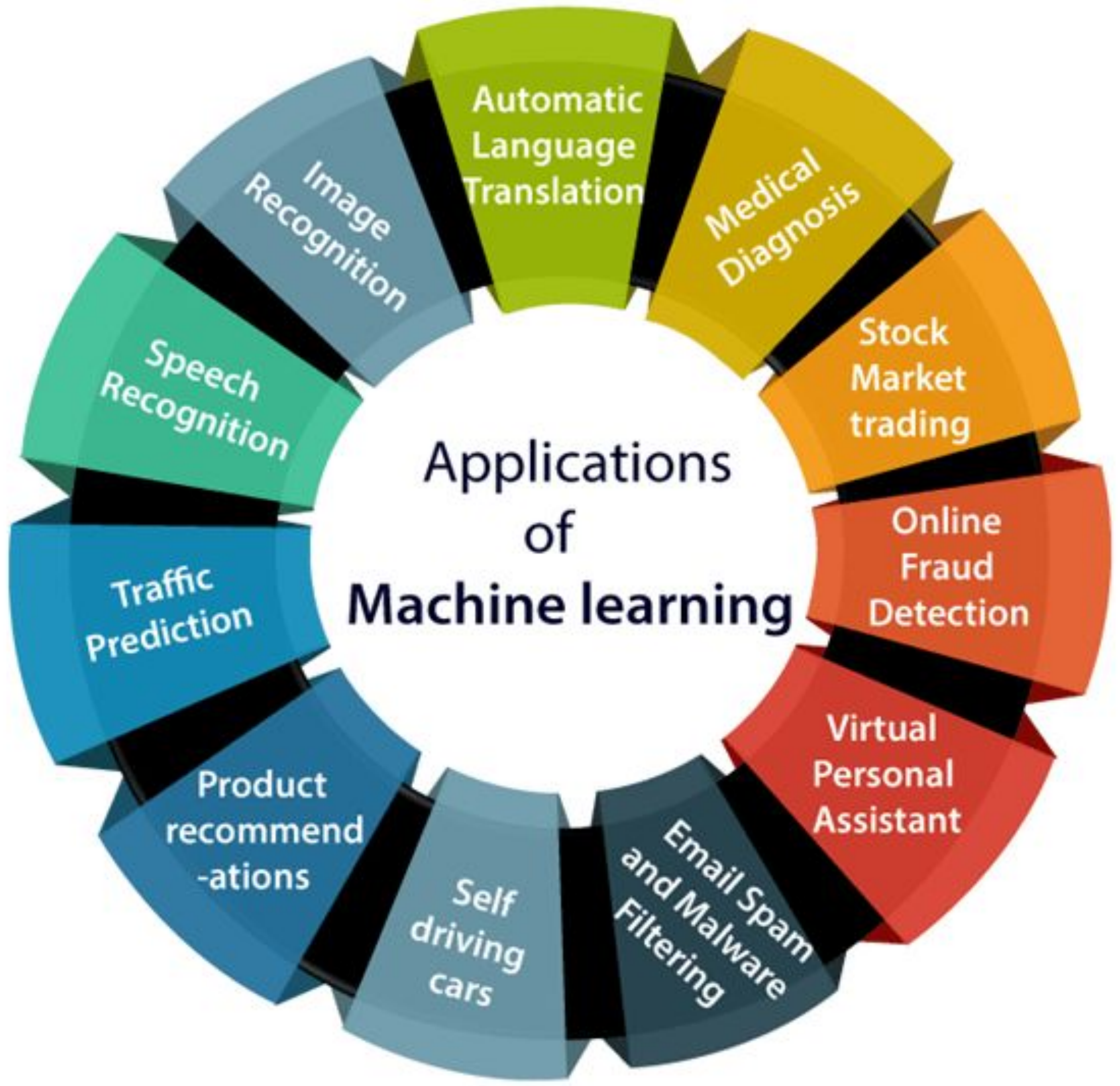
Machine learning – finding its application in almost every sphere of life

Why Machine learning ??

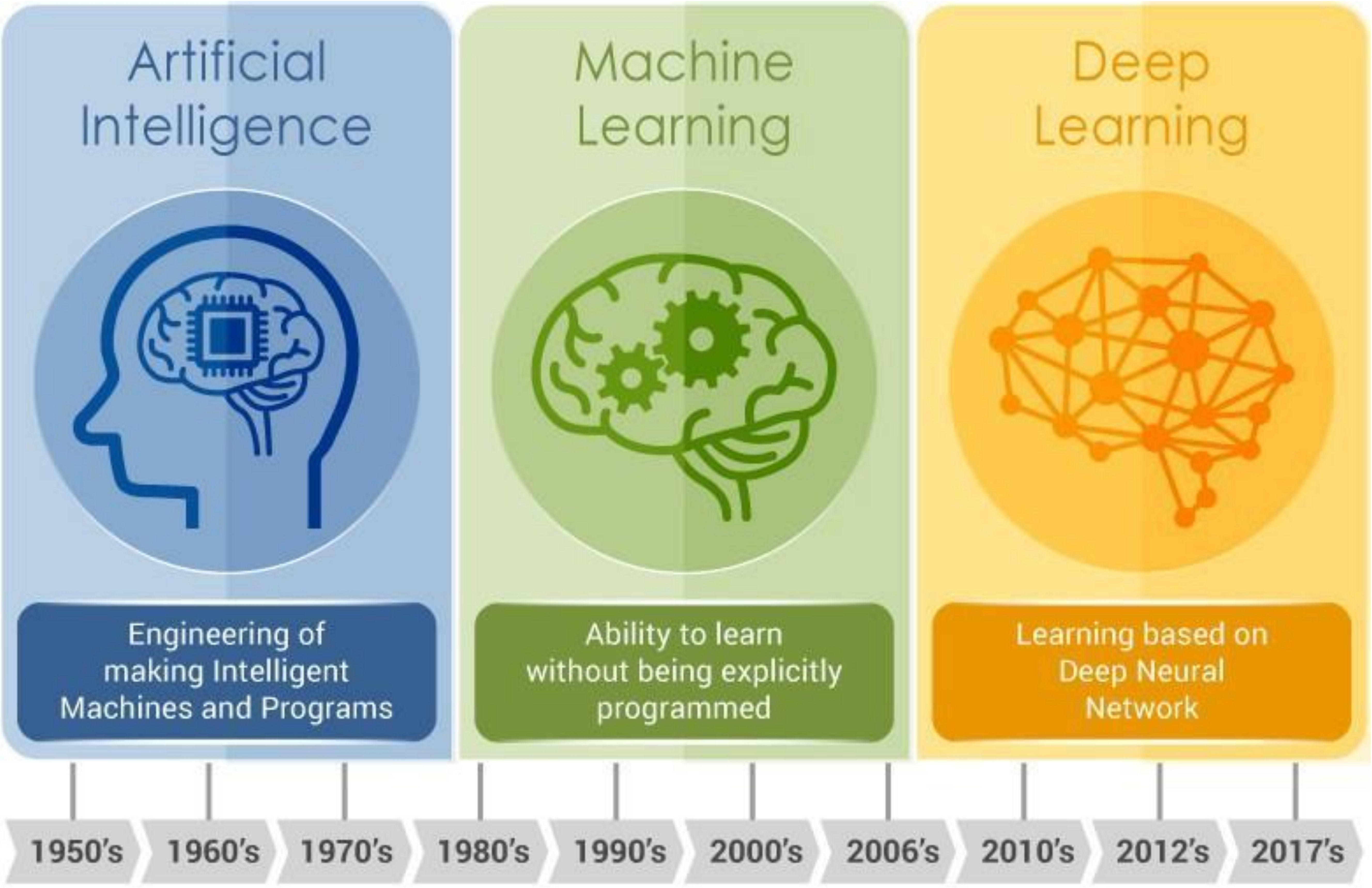
- Develop systems that can automatically adapt and customize themselves
- Discover New knowledge from large databases
- Ability to mimic human an replace monotonous tasks
- Develop systems that are too difficult / expensive to construct

Why Now ??

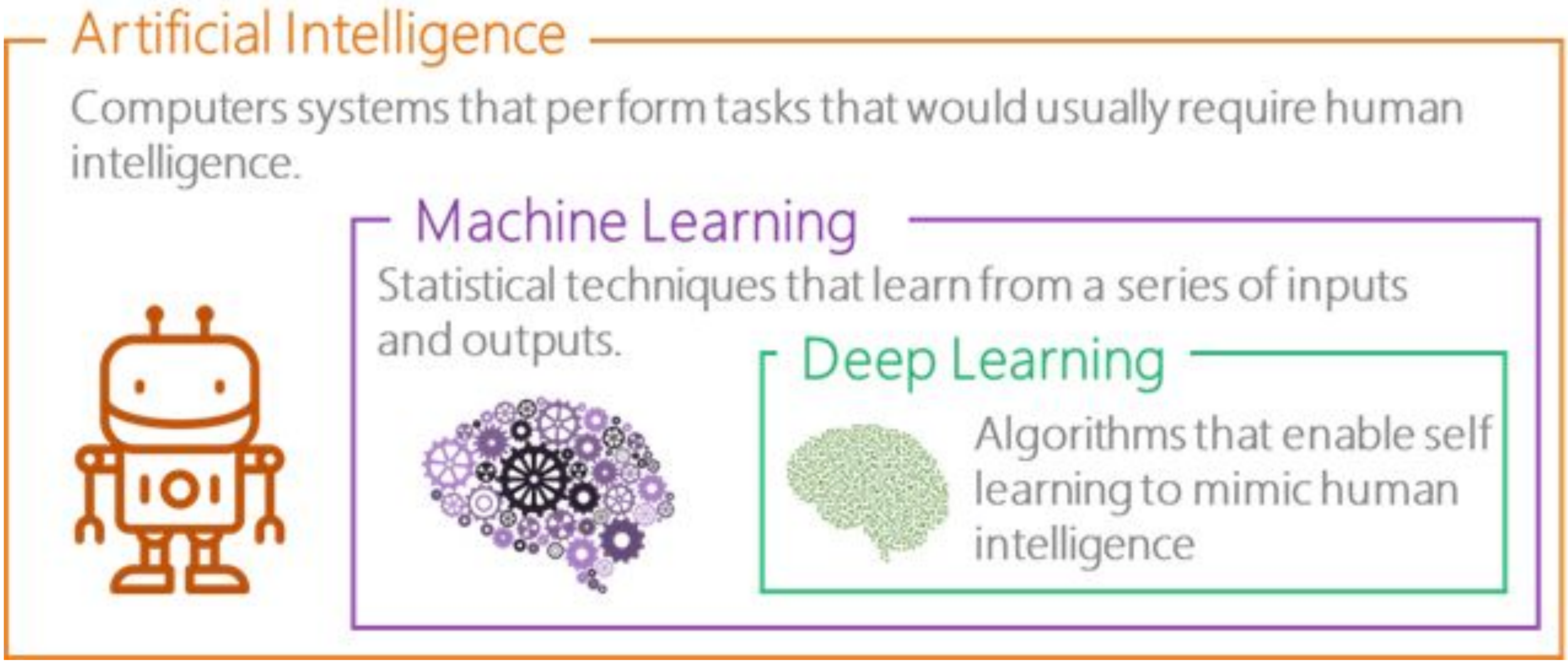
- Flood of Data Mining
- Increasing computational Power
- Growing Progress in Available algorithms and theory developed by researchers
- Increased support from Industries



Evolution of Machine Learning



AI ML and Dl



DL ⊆ ML ⊆ AI

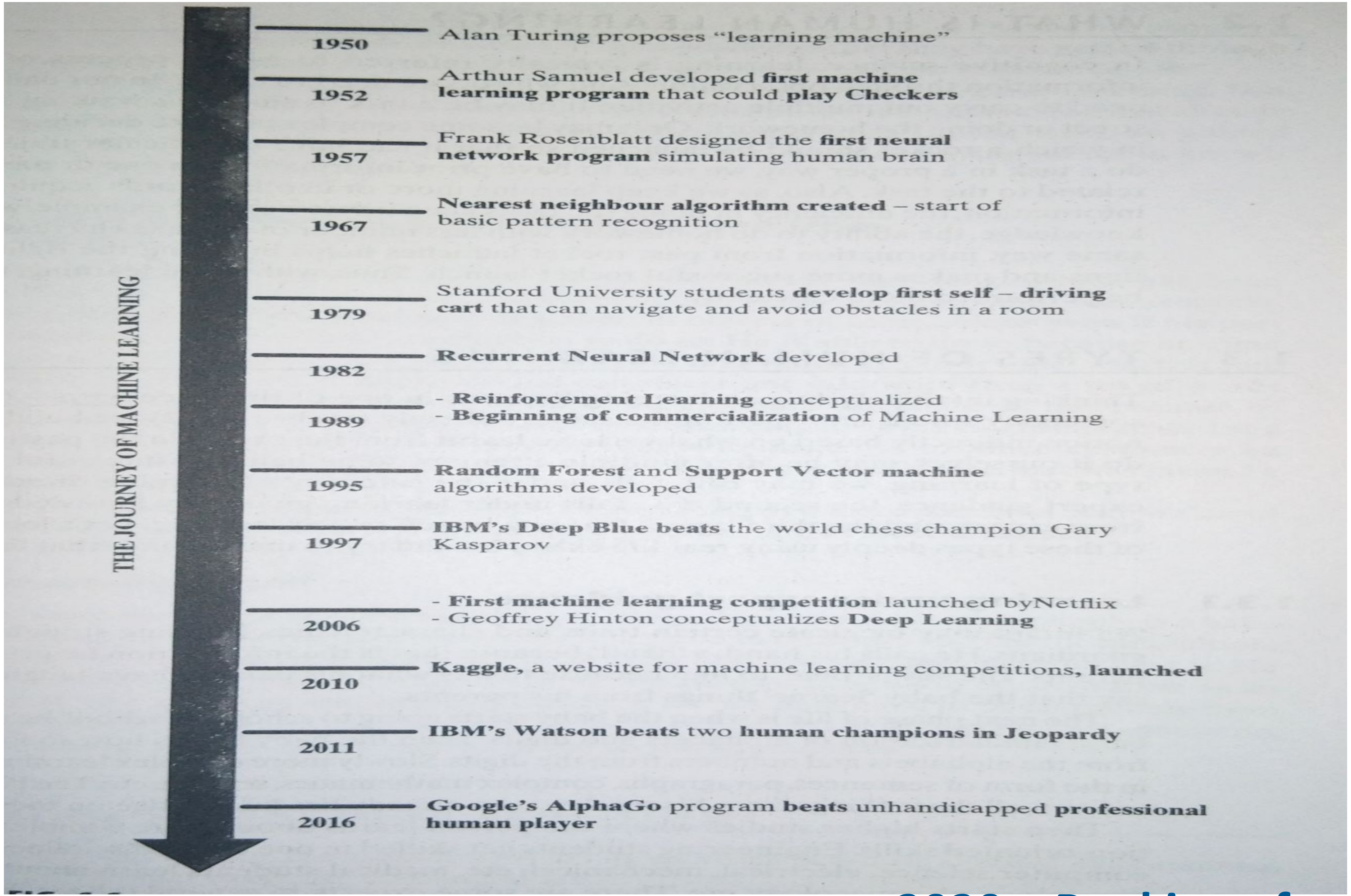
AI ML and Dl

	Artificial intelligence	Machine learning	Deep learning
What is it	Intelligence demonstrated by machines	A subset of artificial intelligence	A subset of machine learning
What does it use	It studies ways to build programs so that machines can solve problems	It provides systems the ability to automatically learn and improve from experience	It imitates the workings of the human brain in processing data so the system can create patterns
Where is it used	Siri, Tesla, Alexa, Netflix, Face detection and recognition, Recommendation algorithms, Google maps	Virtual assistants, Traffic predictions, Social media with people you may know suggestions, Medical diagnosis	Self-driving cars, Visual recognition, Virtual assistants, Financial fraud detection

Why Machine Learning ?

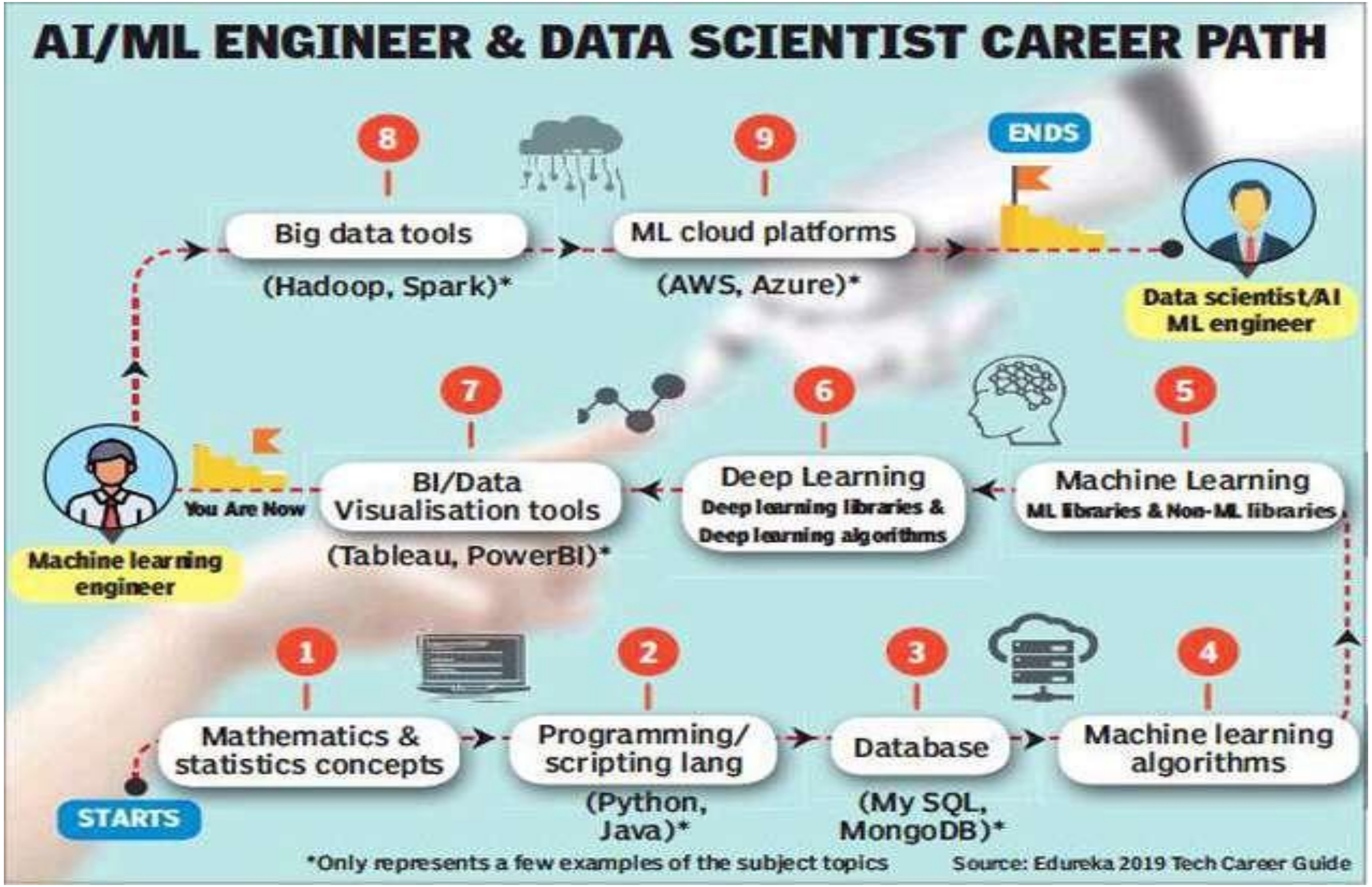
- **No need for human experts**
  - industrial/manufacturing control
  - mass spectrometer analysis, drug design, astronomic discovery
- **Black-box human expertise**
  - face/handwriting/speech recognition
  - driving a car, flying a plane
- **Rapidly changing phenomena**
  - credit scoring, financial modeling
  - diagnosis, fraud detection
- **Need for customization/personalization**
  - personalized news reader
  - movie/book recommendation

Evolution of Machine Learning

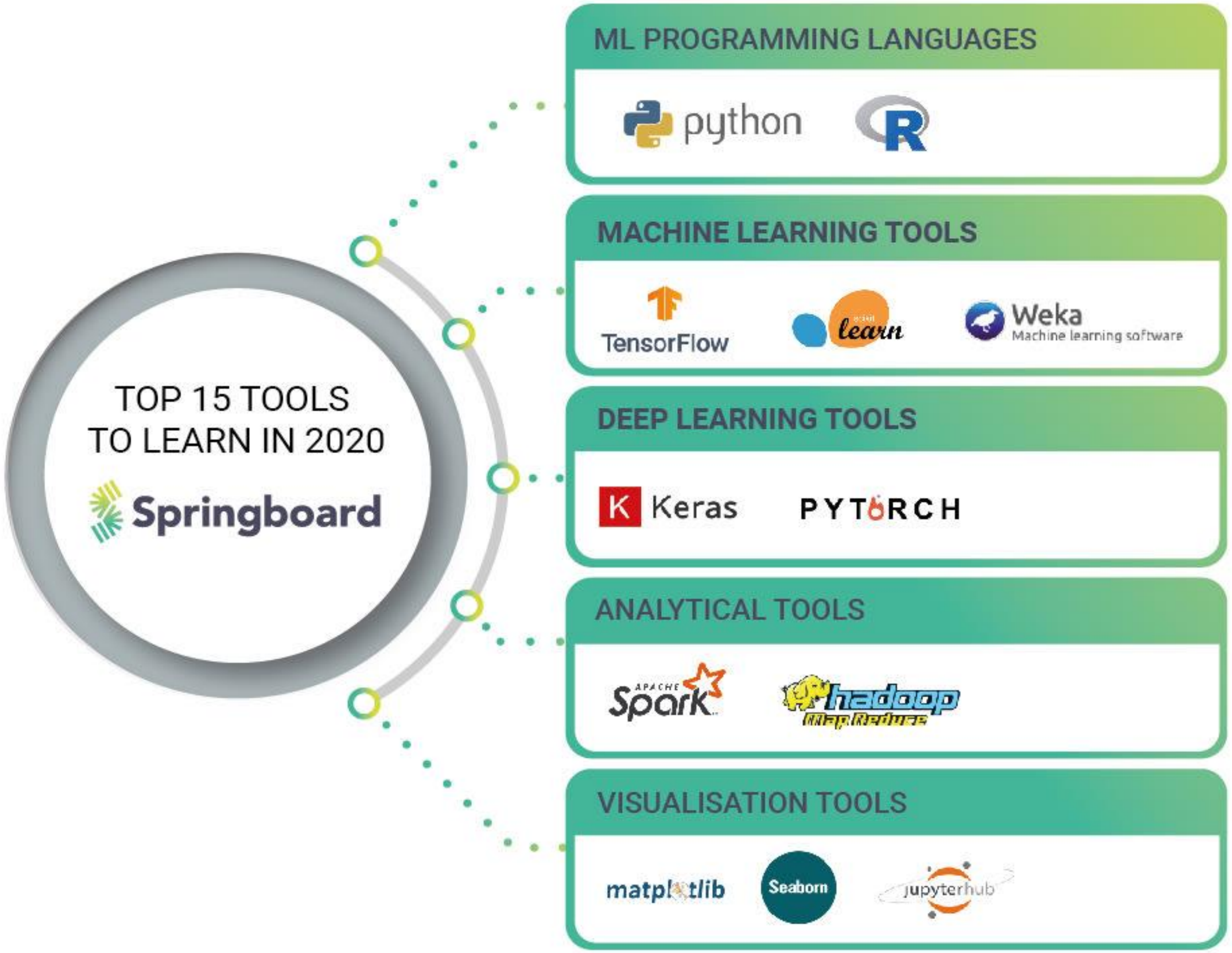


- 2020 – Banking - fraud

Career Oppurtunities



Tools to Learn



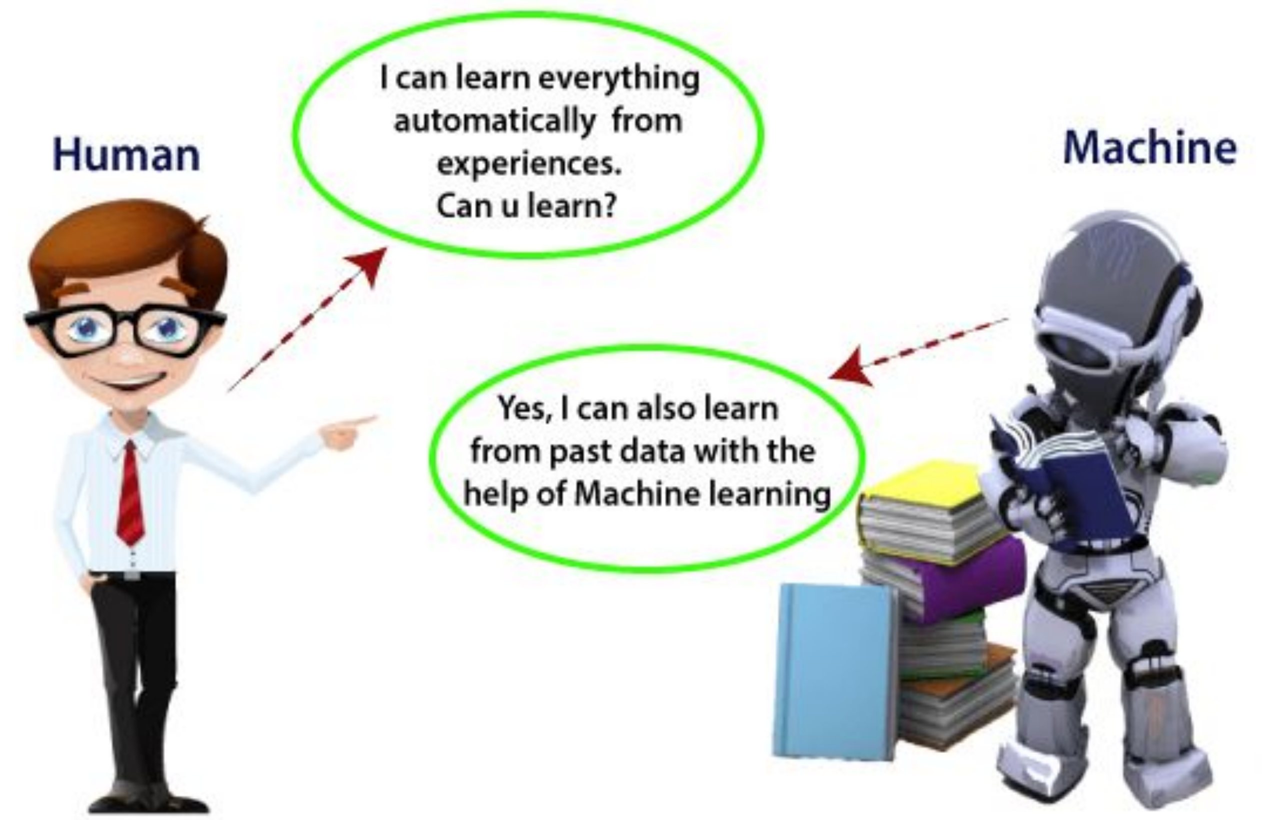
What is Machine Learning

A branch of **artificial intelligence** that provide computers with the ability to learn without being explicitly programmed

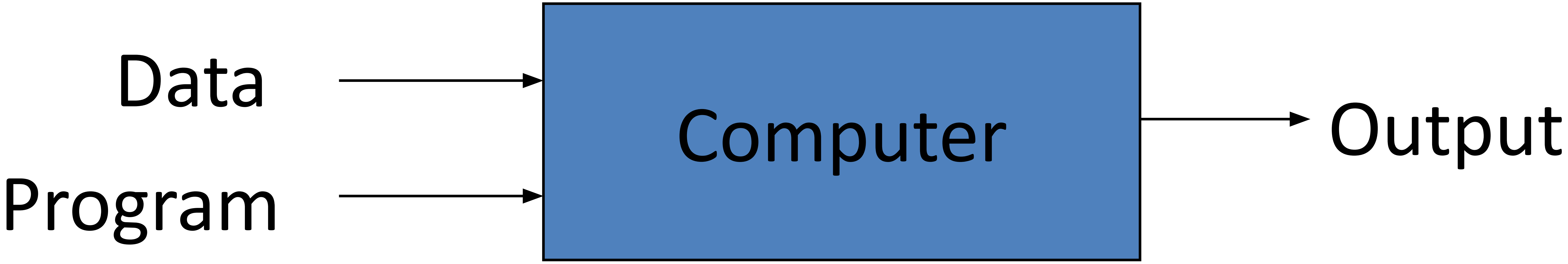
Simple definition: Machine Learning is a program which can learn on it's own from the data

In conventional programming we explicitly write what a program should do.

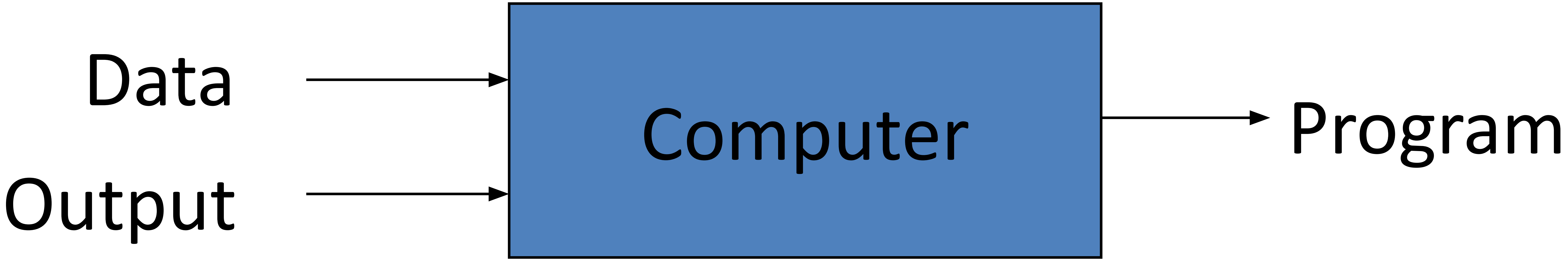
In machine learning an algorithm is given data and it learns the relation on it's own



Conventional Programming



Machine Learning



HUMAN LEARNING

In Cognitive Science – Process of gaining Information through observation

Types

- Learning under Expert Guidance

- Learning Guided by Knowledge gained from experts

- Learning by Self

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" – Tom M Mitchell

- Do machine learn – if so how?
- Which problem is well-posed learning problem?
- What are the important features that are required to well define a learning problem

HOW DO MACHINES LEARN

Data Input – Past data or information is utilized as a basis for future decision making

Abstraction - The input data is represented in a brooder way through the underlying algorithm

Generalization – The abstracted representation is generalized to form a framework for decision making

Data Input – vast pool of knowledge is available from the data input; features considered, labels, type values

Abstraction – helps in deriving conceptual map – model as known in ML

- Computational blocks like if/else rules
- Mathematical equations
- Specific data structures like tree or graphs
- Logical grouping of similar observations
- Choice of model based on multiple aspects eg –
- Type of problem to be solved – prediction, analysis of trends
- Nature of input data
- Domain of the problem – critical domain eg fraud detection

Well-Posed Learning Problem

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 – need a program that will prompt the next word as and when I type a word

Formalism

Task (T) : Prompt the next word when I type a word

Experience (E) : A corpus of commonly used English words and phrases

Performance (P) : Number of correct words prompted considered as percentage – in ML – learning accuracy

Assumptions – Create a list of assumptions about the problem

Similar problems – What other problems seen similar trying to solve?

Step 2 – why need to be solved?

Motivation – long standing business issues etc

Solution Benefits – articulated to sell the project

Solution Use – life time expected

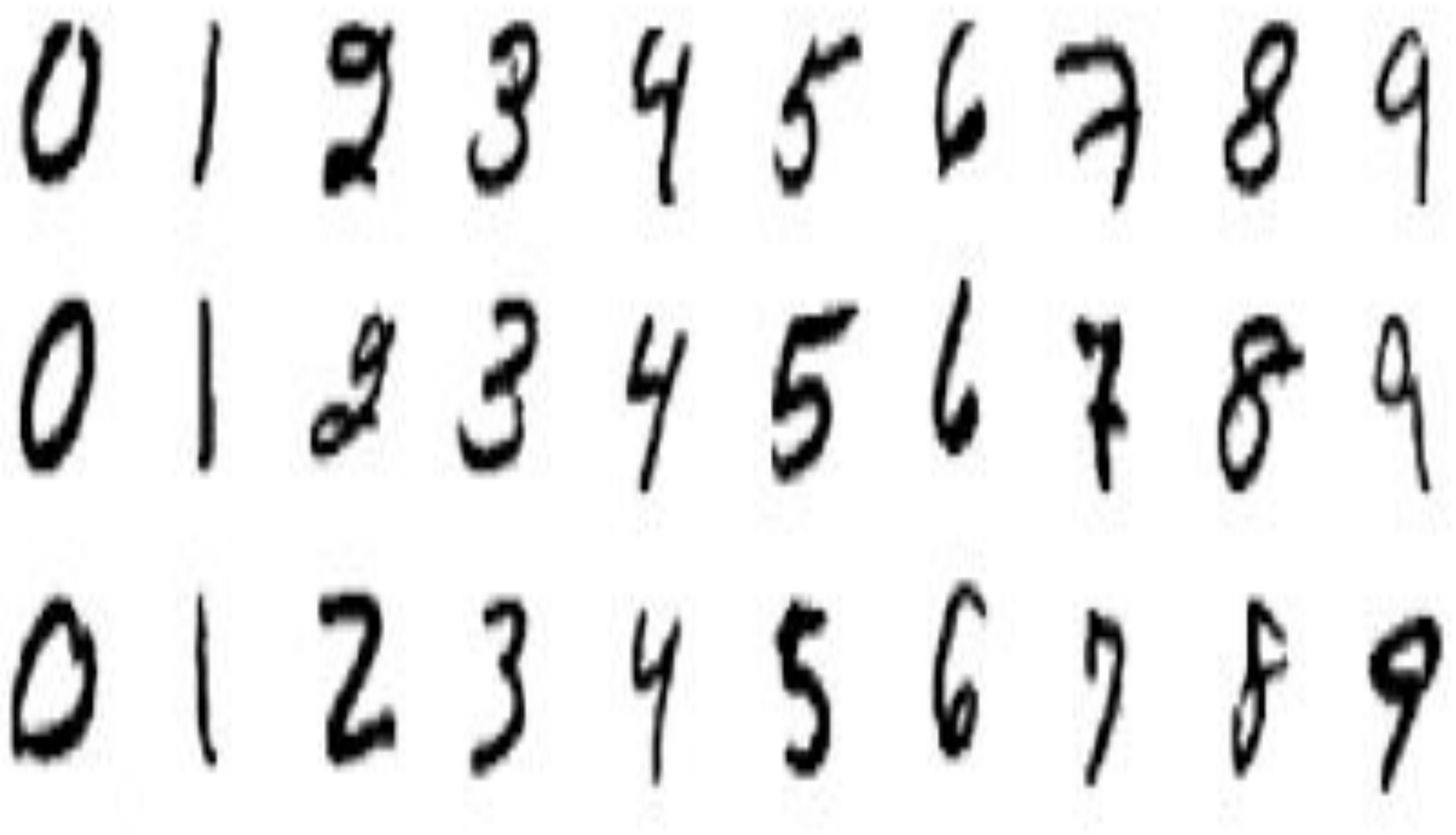
Step 3 – how would I solve the problem?

Explore to solve manually

Detail out step-by step data collection, data preparation and program design; collect all previous section details – including assumptions

A Handwriting Recognition Learning Problem:

- Task T : Recognizing and Classifying Handwritten words within images
- Performance Measure P : Percentage of words correctly classified
- Training Experience E : A database /dataset of handwritten words with classifications



Justify the following as Well posed Problem and write the steps in detail:

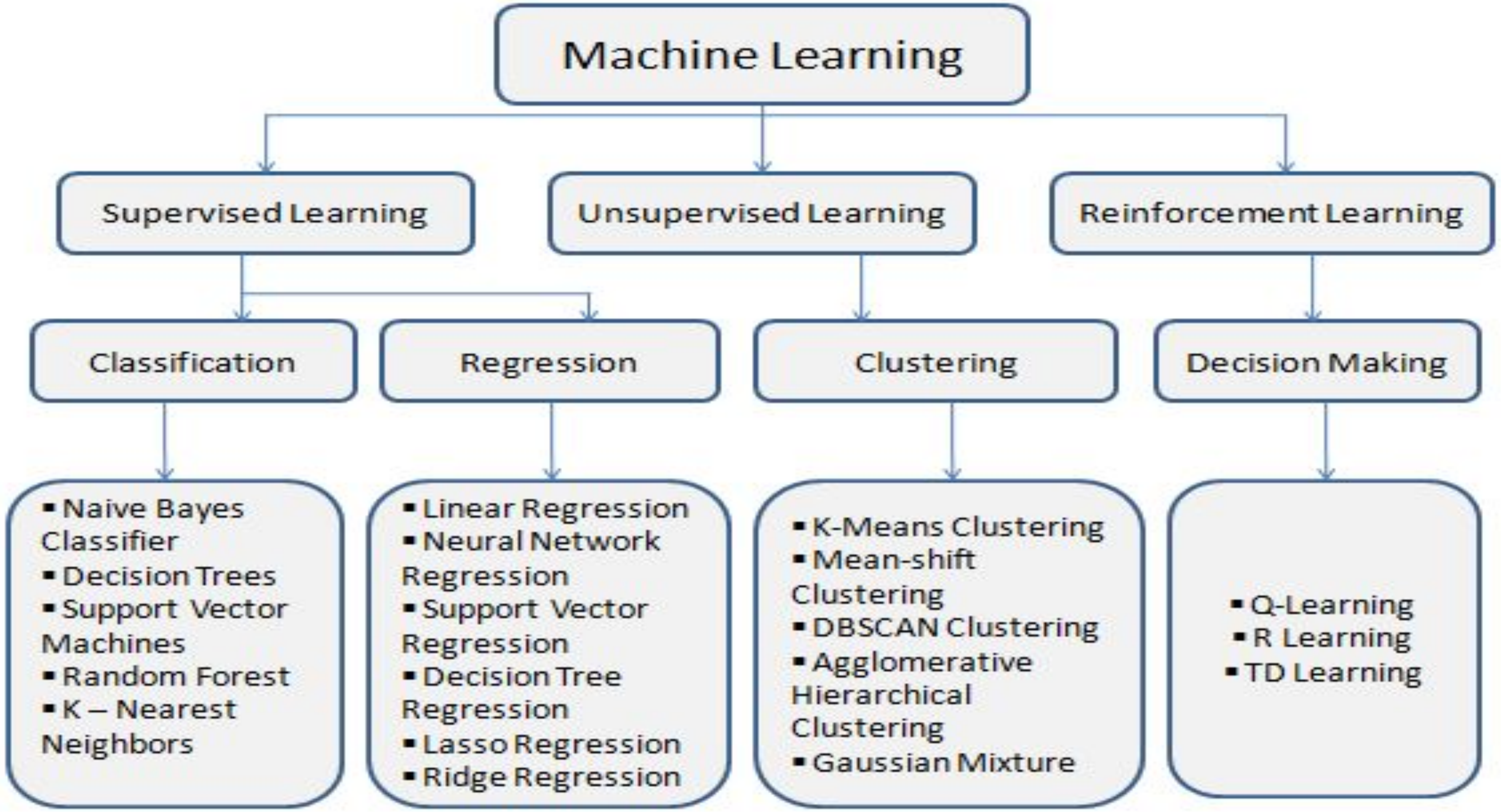
(i)Spam Detection

(ii)Face Recognition Problem

(iii)Drug analysis

(iv)Fraud detection

Types of Machine learning



SUPERVISED LEARNING

Also called predictive learning.

A machine predicts the class of unknown objects based on prior class-related information of similar objects.

The major motivation of supervised learning is to learn from past information.

A machine needs the basics information to be provided to it. This basic input is given in the form of training data.

Training data is the past information on specific task.

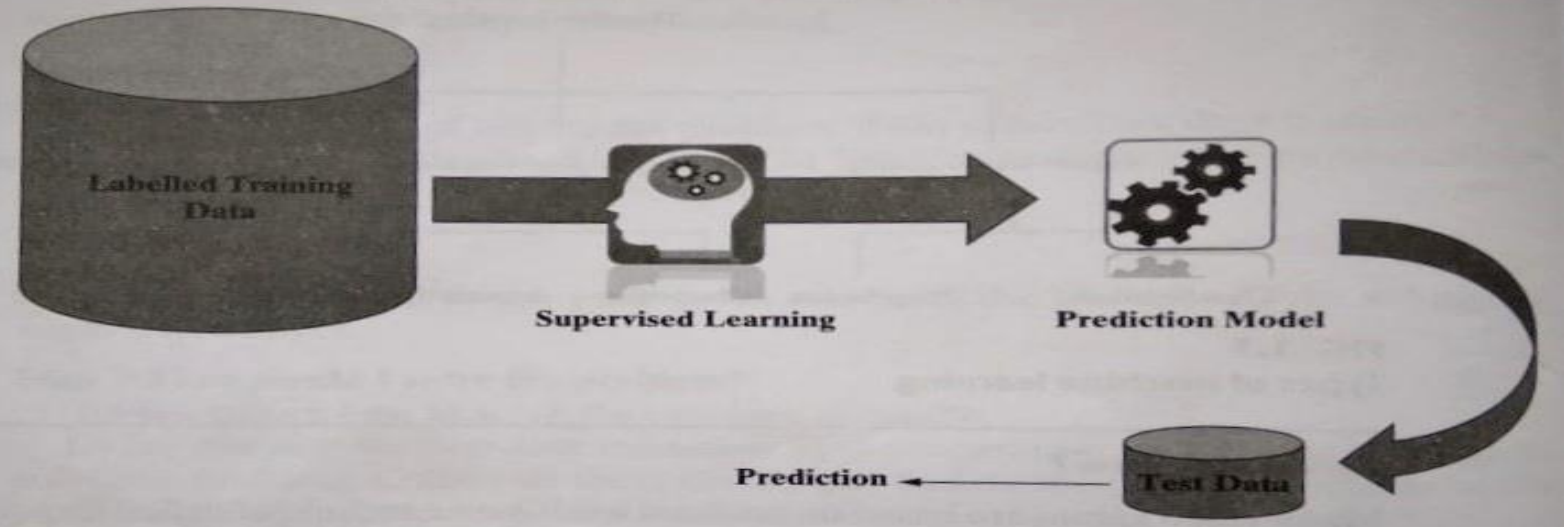
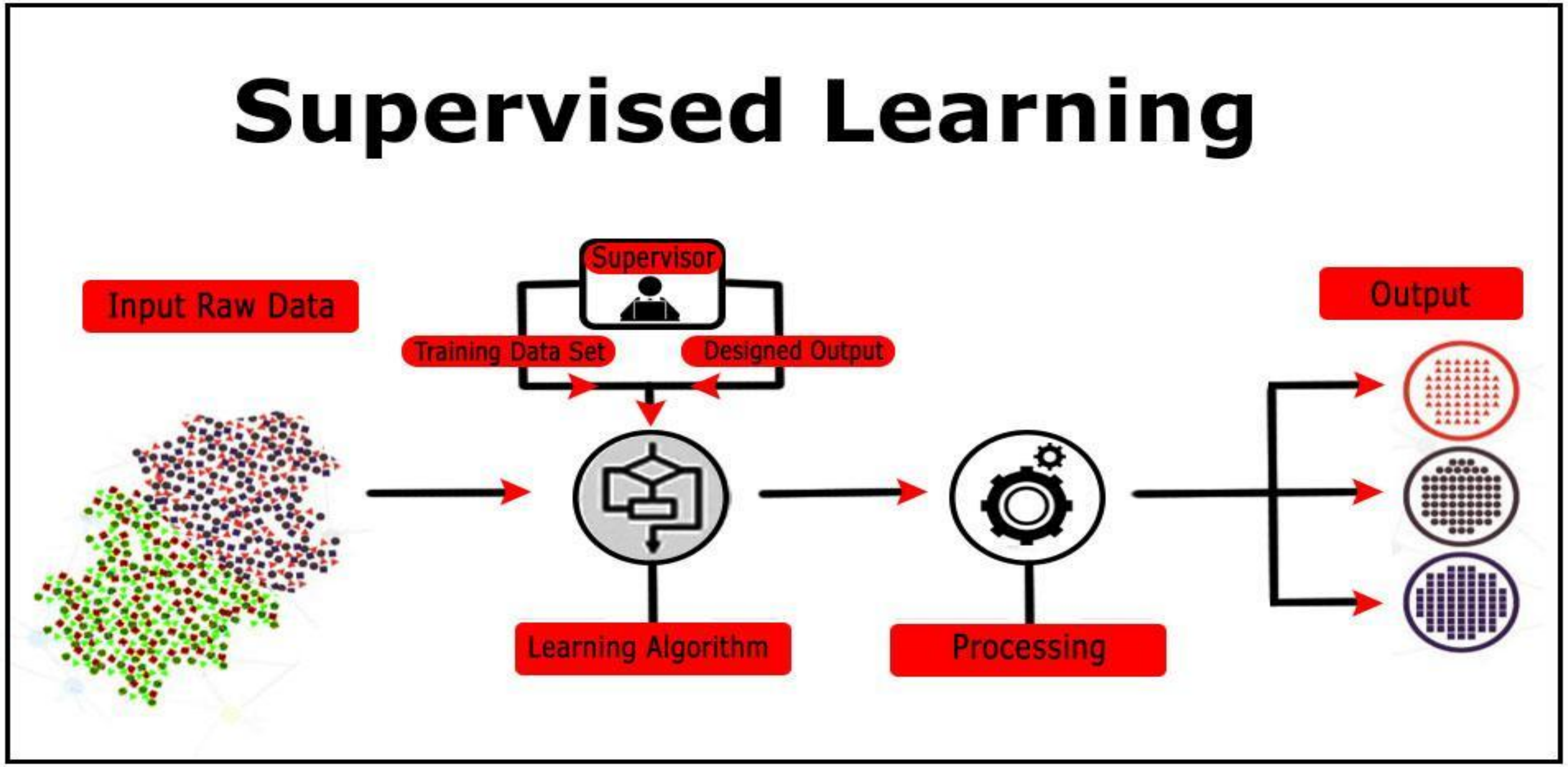


FIG. 1.4
Supervised learning

Some examples of Supervise learning are:

- Predicting the results of a game
- Classifying text such as classifying a set of e-mails as a spam or non-spam



Some examples of Supervise learning are:

- Predicting the results of a game
- Classifying text such as classifying a set of e-mails as a spam or non-spam

A - Classification

In which category the machine should put an data of unknown category, also called a **test data**, depends on the information it gets from the past data that is training data.

Assigning a label or category or class to a test data based on label or category or class information that is imparted by the training data.

Typical classification problems

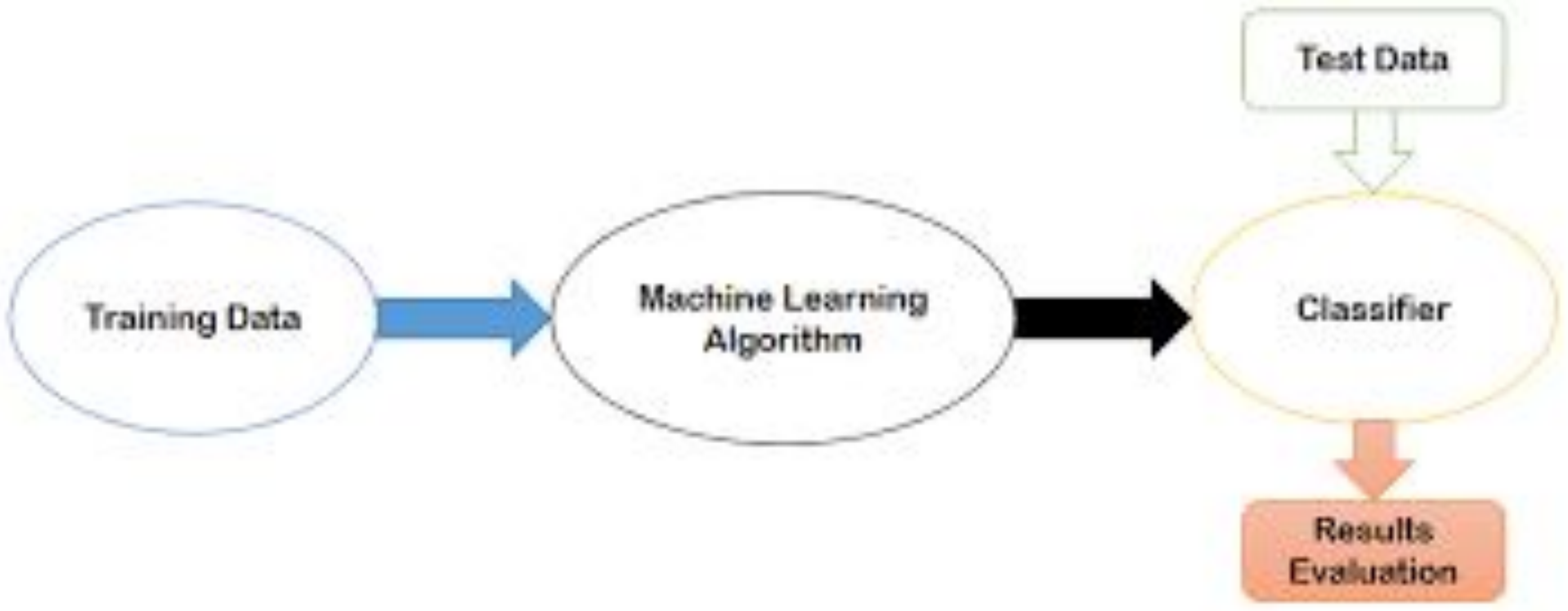
-Image classification

-Prediction of disease

-Win-loss prediction of games

-Prediction of natural calamity like earthquake, flood, etc.

-Recognition of handwriting



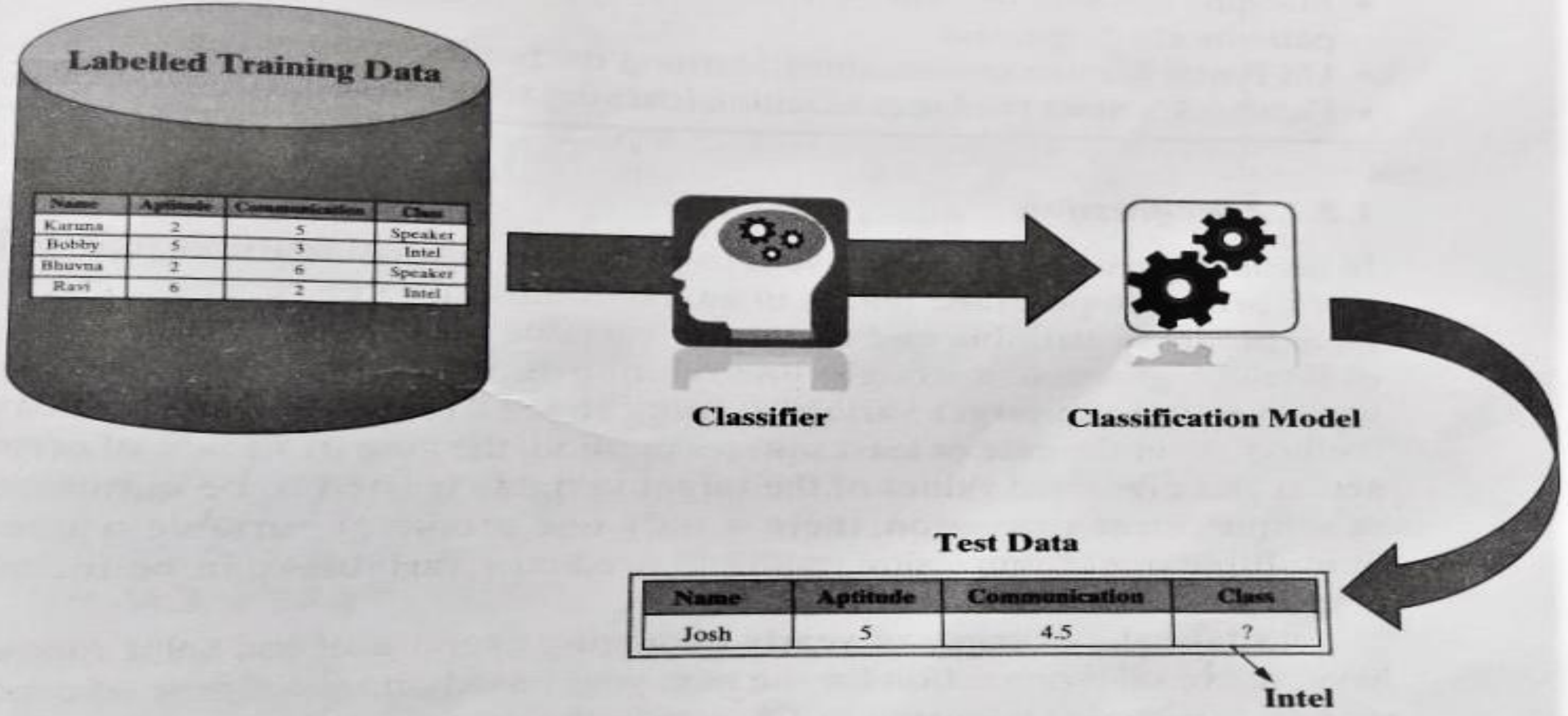


FIG. 1.5
Classification

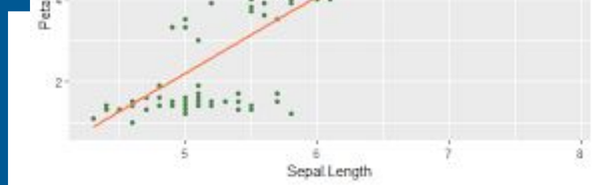
B - Regression

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.

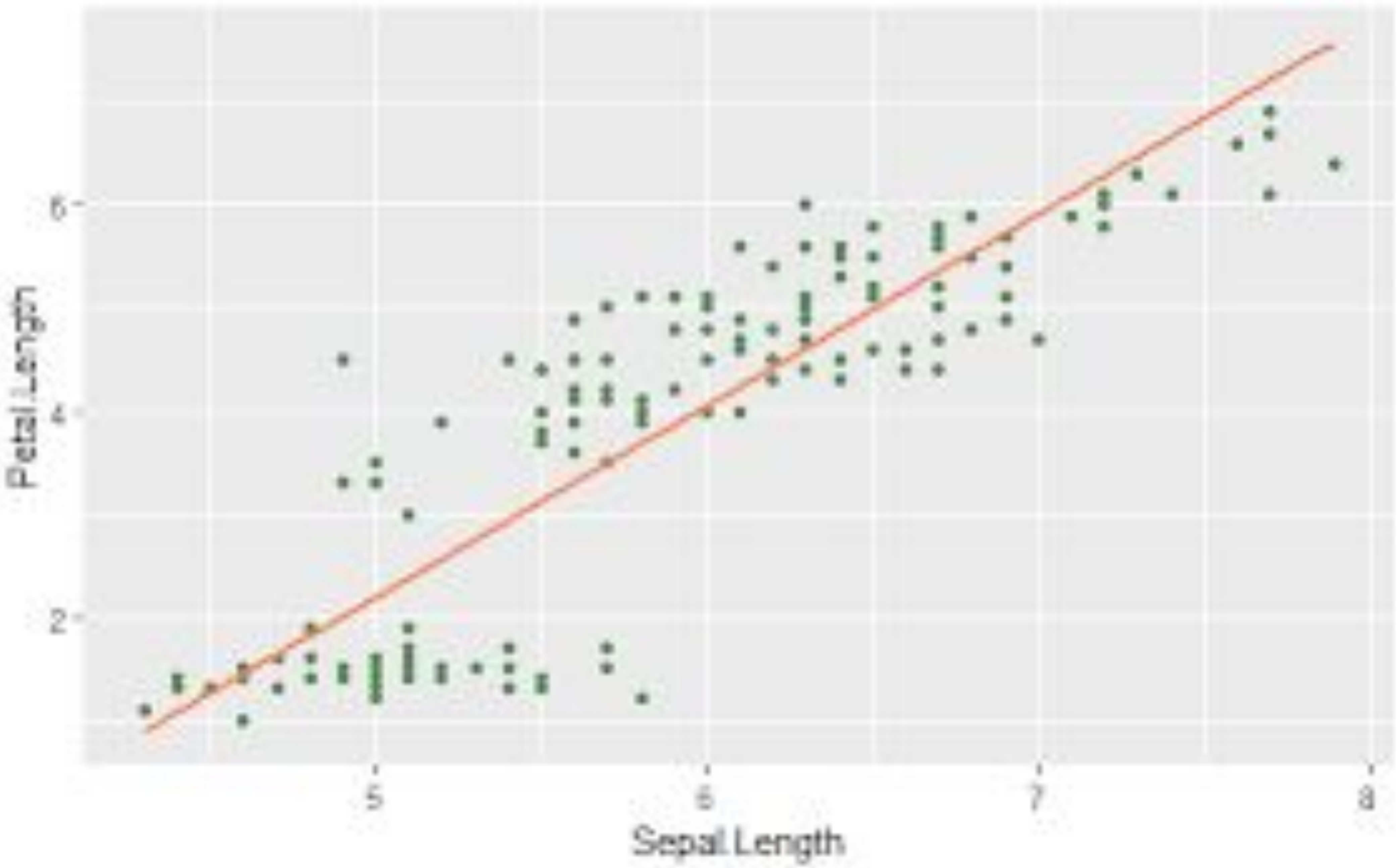
Typical applications of regression,

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



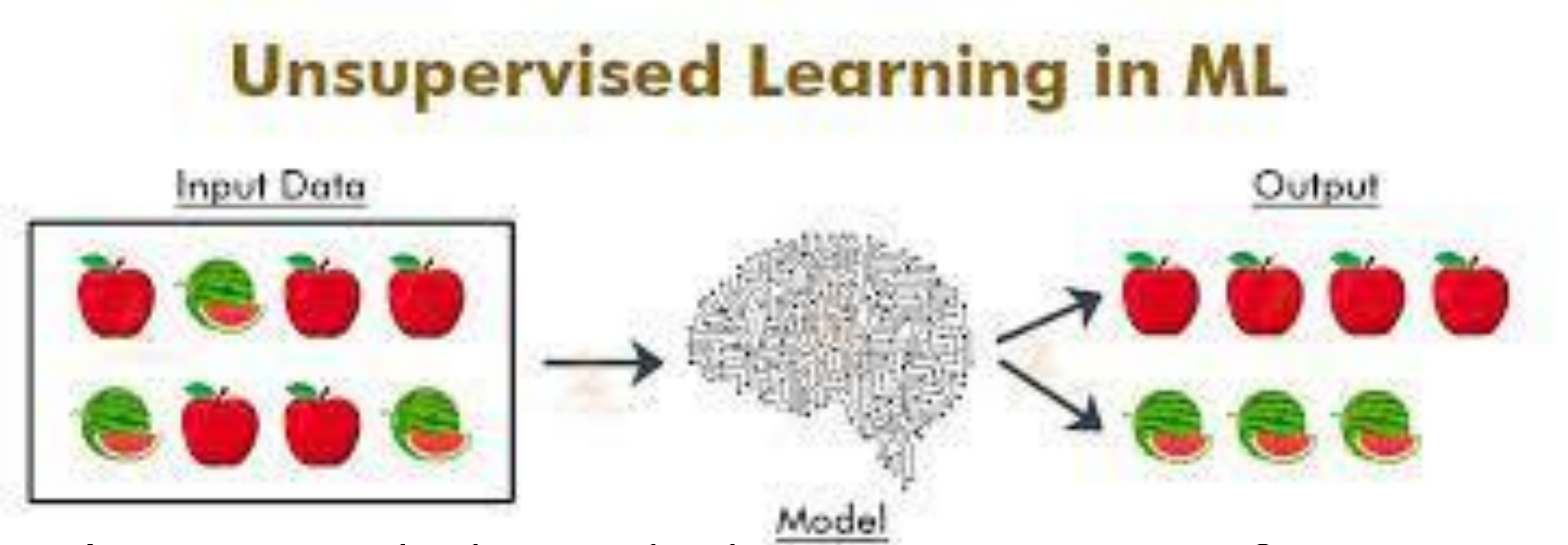
A typical linear regression model can be represented in the form-

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



UNSUPERVISED LEARNING

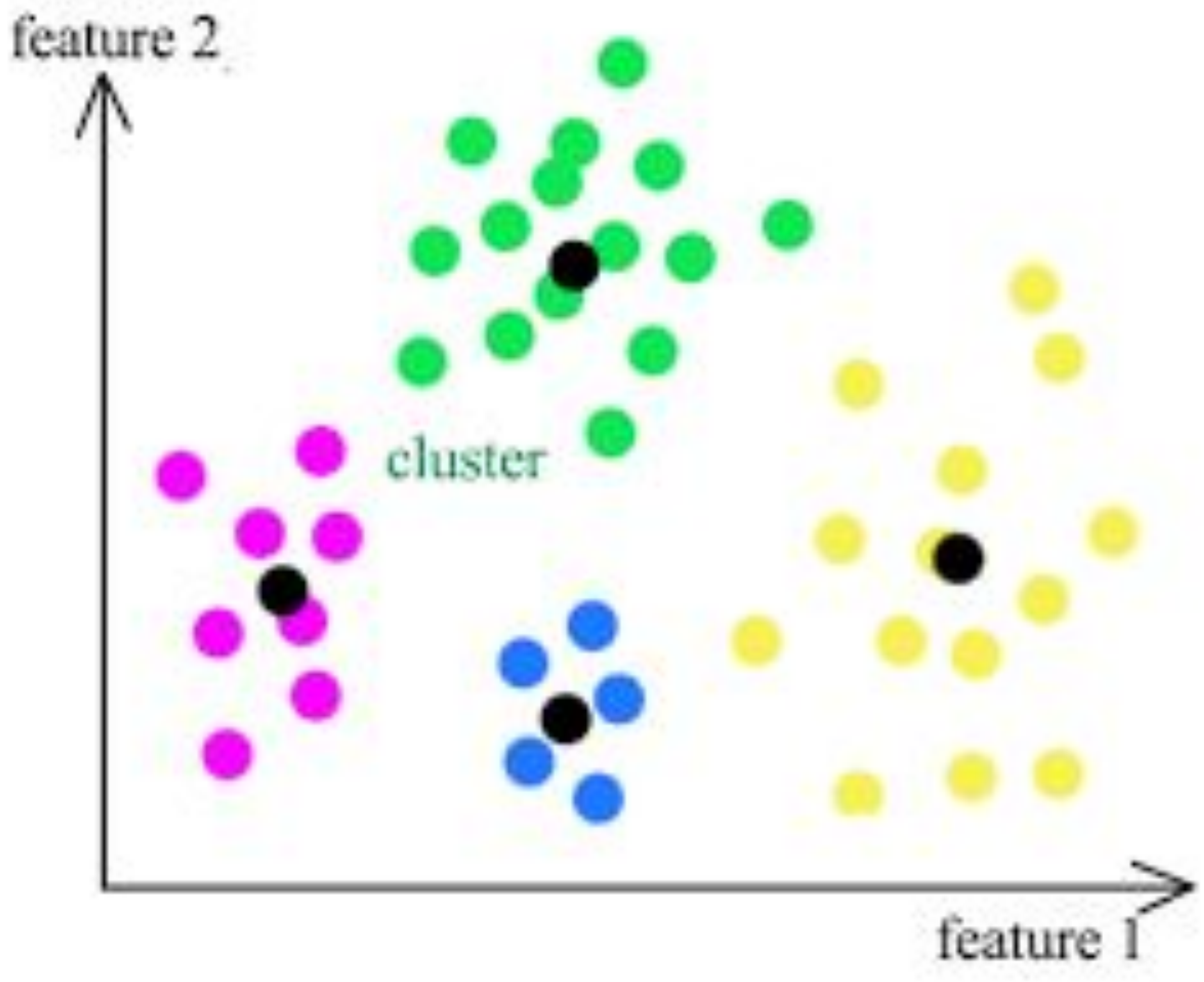
- There is no labelled training data to learn from and no prediction to be made.
- The objective to take a dataset as input and try to find natural grouping or patterns within the data elements or records.



- 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.

A - Clustering

- To group or organize similar objects together.
- The objective of clustering to discover the intrinsic grouping of unlabelled data and form clusters.
- Different measures of similarity can be applied for clustering.
  - Most commonly adopted similarity measure is distance.



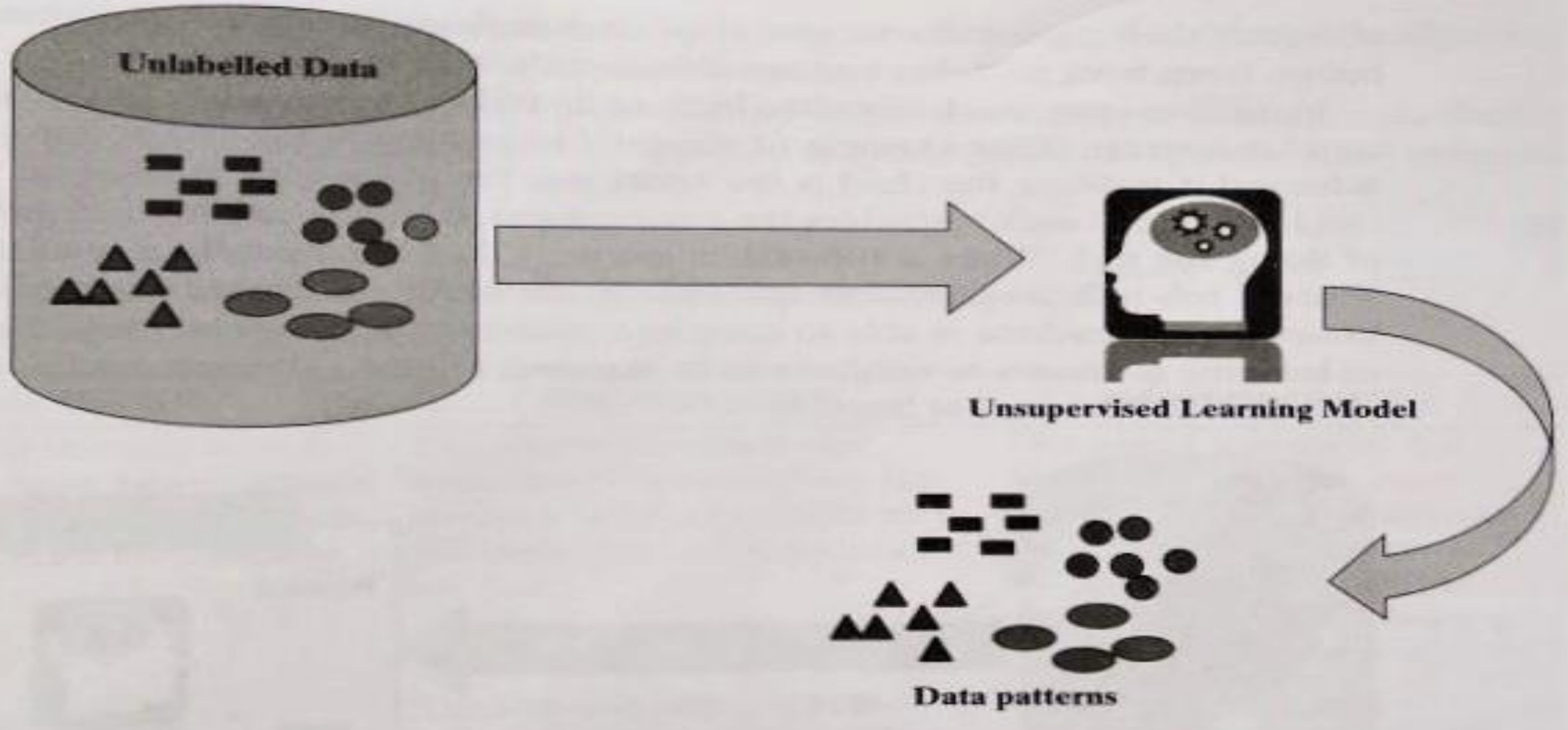
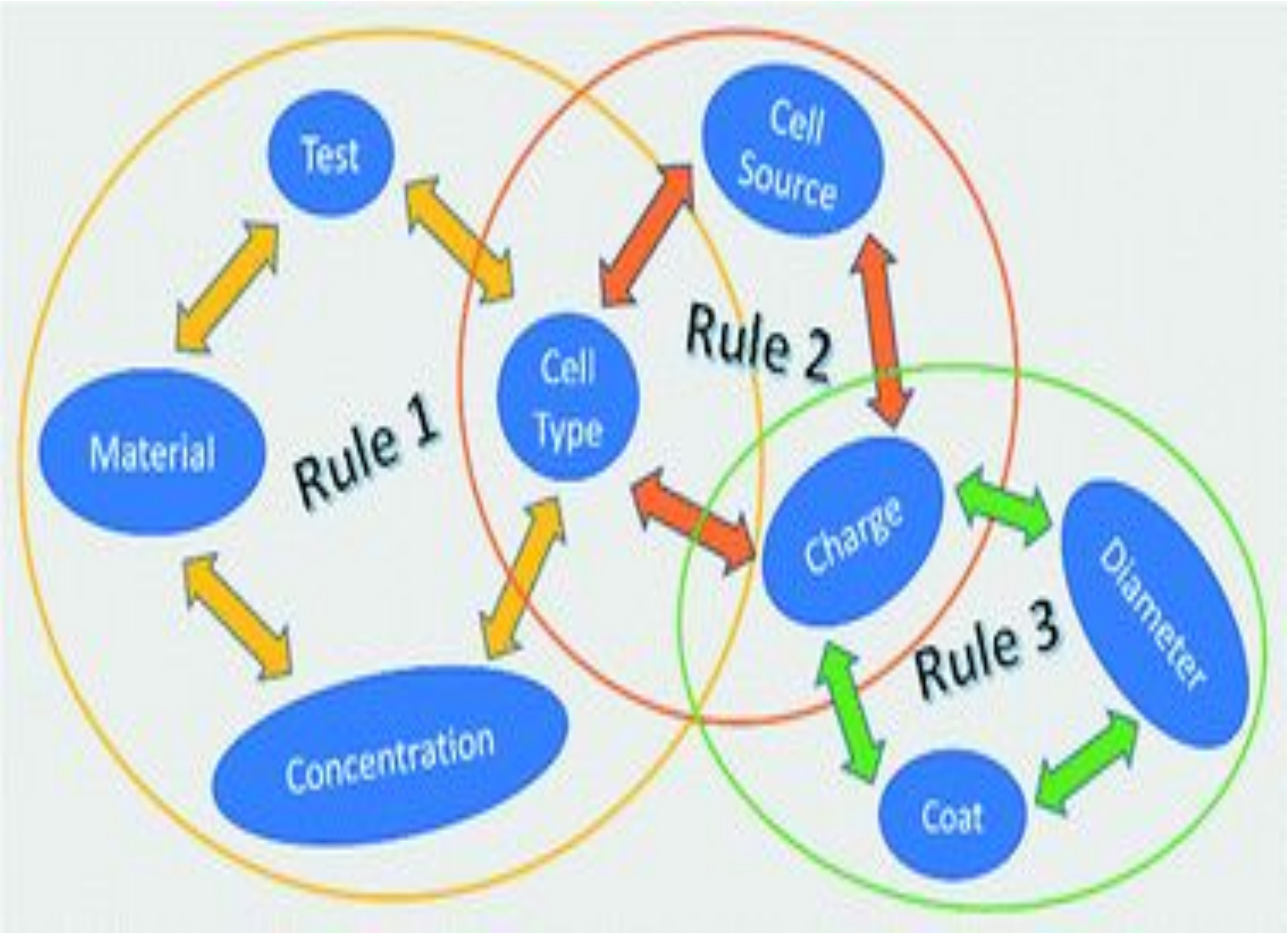


FIG. 1.8
Unsupervised learning

B - Association Analysis

One more variant of unsupervised learning is association analysis, the association between data elements is identified.

Critical applications of association analysis include market basket analysis and recommender system.



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

FIG. 1.9

Market basket analysis

Reinforcement learning

- A machine learns act on its own to achieve the given goals.
- Learn from their past mistakes – Eg - Babies
- Machine often learn to do tasks automatically.
- Subtask is accomplished successfully - a reward is given
- Subtask is not executed correctly - no reward is given.
- Example of reinforcement learning is self-driving cars.

Supervised Learning – ***Train Me!***

Unsupervised Learning – ***I am self sufficient in learning***

Reinforcement Learning – ***My life My rules! (Hit & Trial)***

DIFFERENCES

Supervised Learning	Unsupervised Learning	Reinforcement Learning
Task Driven	Data Driven	React on environment
Classes / Labels available	Model has to find pattern	Reward – if classification correct else punishment
Model is built on training data	Unknown and unlabelled data set - records to be grouped	Model learns and updates itself
Performance evaluated – based on misclassifications done based on predicted and actual values	Difficult to measure whether the model did something useful. Homogeneity of records is the only measure	Evaluated by means of reward function
Two types - classification and regression	Two types - clustering and association	no such types

Supervised Learning	Unsupervised Learning	Reinforcement Learning
Simplest to understand	More difficult to understand and implement than supervised learning	Most complex to understand and apply
	Standard algorithms	
Naïve Bayes k –Nearest Neighbor(kNN) Decision tree Linear regression Logistsic regression SVM	k- Means Principal Component Analysis(PCA) Self Organizing Maps Apriori algorithm DBSCAN	Q-learning Sarsa
	Practical Applications Include	
Handwriting recognition Stock market prediction Disease prediction Fraud detection	Market basket analysis Recommender systems Customer Segmentation	Self driving cars Intelligent Robots AlphaGo Zero

Problems not to be solved using ML

- Machine learning should not be applied to tasks in which humans are very effective or frequent human intervention is needed. For example air traffic control.
- For very simple tasks which can be implemented using traditional programming paradigms, there is no need of machine learning. For example, formula based applications like calculator engine, dispute tracking application.
- Machine learning should be used only when the business process has some lapses. If the task is already optimized, machine learning will not serve to justify the return the return on investment.
- For situations where training data is not sufficient, machine learning cannot be used effectively, because with small data sets, the impact of bad data is exponentially worse.
- For the quality of prediction or recommendation to be good, the training data should be sizeable.

Banking and finance

Fraudulent transactions are spotted and prevented right at the time of occurrence.

Demotivated customers

Customer churn reducing

Insurance

Data intensive

Two major areas in the insurance industry where machine learning is used are risk prediction during new customer on boarding and claims management.

Healthcare

Wearable device

Alert systems

Machine learning along with computer vision also plays a crucial role in disease diagnosis from medical imaging.

And many more

STATE OF THE ART LANGUAGES/TOOLS IN ML

Python

Python has very strong libraries for advanced mathematical functionalities(NumPy)

Algorithms and mathematical tools (SciPy)

Numerical plotting(matplotib)

Machine learning library which has various classification, regression and clustering algorithms embedded in it.

R

Statistical computing and data analysis ; Simple programming language with huge set of libraries available.

Matlab

Licenced commercial software ; Robust support for wide range of numerical computing

Supports statistical functions

SAS

Statistical Analysis System ; Developed in C by SAS Institute

Other languages/tools

SPSS(Statistical Package for the Social Sciences) ; Julia

ISSUES IN MACHINE LEARNING

- Level of research and kind of use of machine learning tools and technologies varies drastically from country to country.
- The biggest fear and issue - privacy and the breach of it.

Preparing to Model

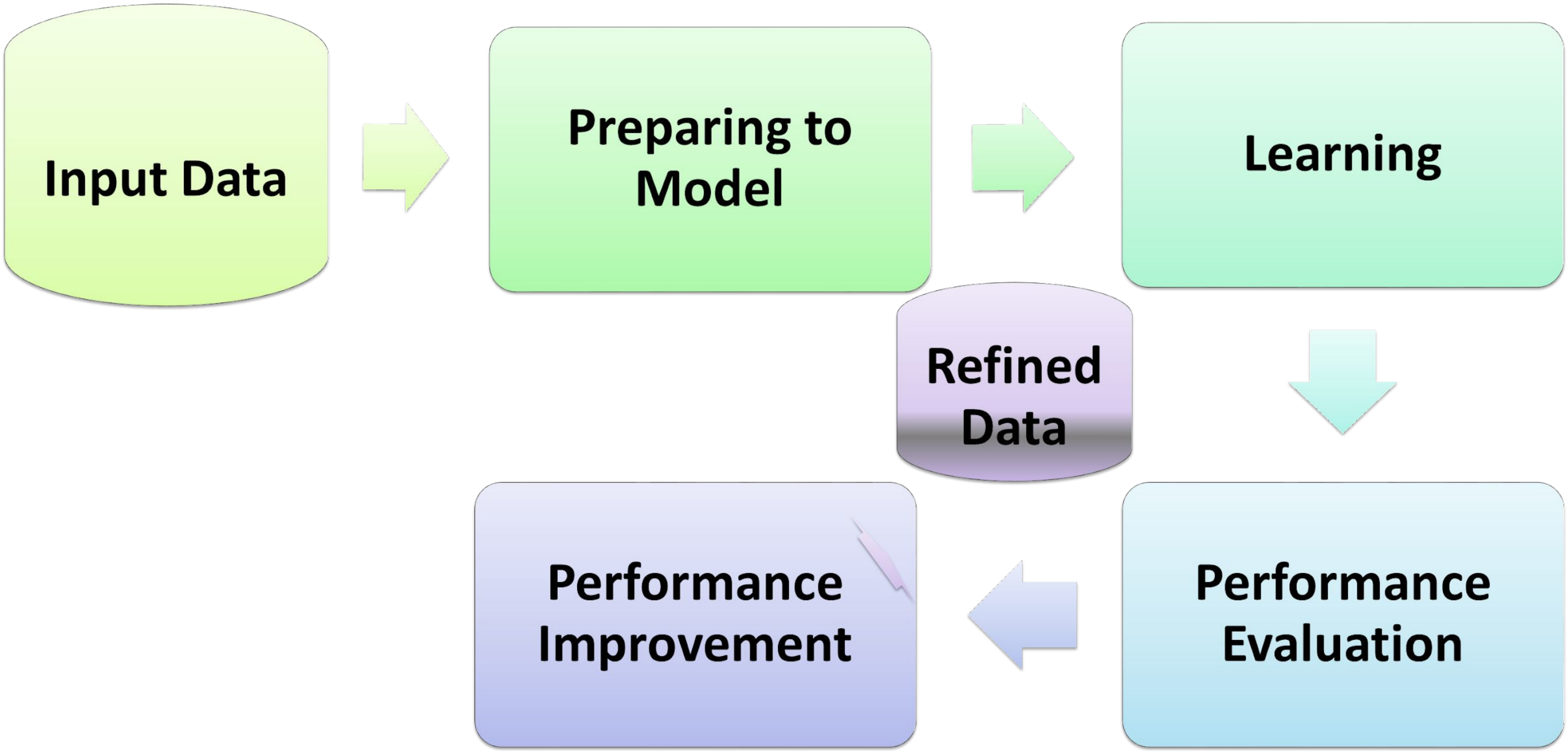
- ***ML Activities***
- ***Basic Datatypes in ML***
- ***Exploring Structure of Data***
- ***Data Quality & Remediation***
- ***Data Pre-processing***

Machine Learning Activities

Preparation activities once the input data comes into the ML system

- Understand the type of data in the i/p data set
- Explore the data – understand the nature & Quality
- Explore the relationships amongst the elements
- Find potential issues in data
- Do necessary remediation – missing data values etc.
- Apply pre-processing steps, as necessary
- Data is prepared – learning tasks start off

DETAILED PROCESS OF ML



Step	Step name	Activities involved

Step 1	Preparing to Model	• Understand the type of data in the i/p data set
		• Explore the data – understand the nature & Quality
		• Explore the relationships amongst the elements
		• Find potential issues in data
		• Remediate data, if needed – missing data values
		• Apply pre-processing steps, as necessary
		  • Dimensionality Reduction
		  • Feature Subset selection

Step 2	Learning	• Data partitioning / holdout ; Model selection; Cross-validation
Step 3	Performance Evaluation	Examine the model performance, eg, confusion matrix in case of classification
Step 4	Performance improvement	• Tuning the model ; Ensemble ; Bagging; Boosting

BASIC TYPES OF DATA IN ML

Data set –

Collection of related info or records

Data objects – representing the entity

Information – entity or some subject area

Attribute – in data field represents a characteristic or feature of a data object

- Machine learning literature uses it as ***Feature***
- Data warehousing – ***Dimensions***
- Statisticians – ***Variable***
- Data mining / Data base professionals use word ***Attribute***

Data types – classification two types

- ***Qualitative and Quantitative***

BASIC TYPES OF DATA IN ML

Student data set:

Roll Number	Name	Gender	Age
129/011	Mihir Karmarkar	M	14
129/012	Geeta Iyer	F	15
129/013	Chanda Bose	F	14
129/014	Sreenu Subramanian	M	14
129/015	Pallav Gupta	M	16
129/016	Gajanan Sharma	M	15

Student performance data set:

Roll Number	Maths	Science	Percentage
129/011	89	45	89.33%
129/012	89	47	90.67%
129/013	68	29	64.67%
129/014	83	38	80.67%
129/015	57	23	53.33%
129/016	78	35	75.33%

FIG. 2.2 Examples of data set

Roll Number	Name	Gender	Age
129/011	Mihir Karmarkar	M	14
129/012	Geeta Iyer	F	15

FIG. 2.3 Data set records and attributes

BASIC TYPES OF DATA IN ML

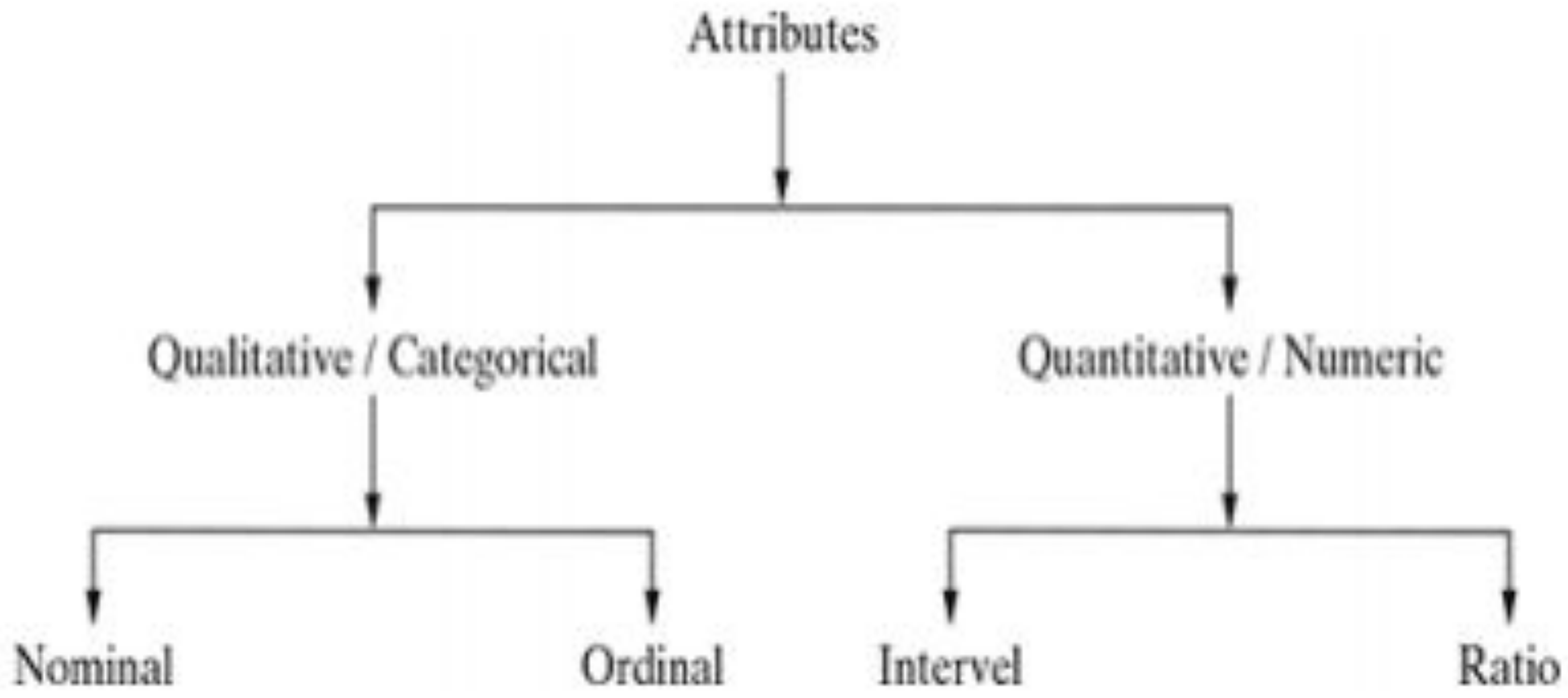


FIG. 2.4 Types of data

Qualitative Data type

Also known as ***categorical data***

Attributes are Qualitative or categorical

Describe a feature of object without giving an actual size or quantity

Representation can be divided into groups

If integers are used – represent computer codes for the categories – as opposed to measurable

0 for small drink; 1 for medium; 2 for large

Two types – ***Nominal and Ordinal data***

Qualitative – Nominal Data

Nominal Data – *"Relating to Names"* – has no numeric value, but named value

Blood Group – A,B,O, AB

Nationality – Indian, American

Gender – Male, Female

Special type of Nominal data – **dichotomous** – two labels – Eg. pass/fail

Binary Attributes – 0 or 1 – 0 typically means absent 1 means present ; ***Symmetric*** if both states are equal –Gender – Male -0 ; female – 1

Asymmetric if the outcomes are not equal – medical test -1 HIV positive ; 0 negative

Qualitative –Ordinal Data

Ordinal Data

same as nominal data plus naturally ordered

Customer satisfaction – very happy, happy, unhappy

Grades – A,B,C etc

Can be obtained from discretization of numeric quantities by splitting the value range into finite number of ordered categories

Can mathematical operations be performed on qualitative data?? If so which ones??

Quantitative Data type

- Also known as numeric data
- Relates to quantity of an object
- Can be measured
- Eg – Marks – can be measured on a scale of measurement

Two types

- ***Interval Data***
- ***Ratio Data***

Quantitative – Interval Data

Interval Data **–** numeric data – order is known and also the exact difference between values is also known

Example – Celsius data – 20°C is equal to five plus 15°C

Calendar data – 2018 – 2021 – 3 years

Doesn't have true zero value

hence only addition and subtraction can be applied – ratio cannot be applied - cannot say the temp 40° C means it is twice as hot as 20°c

Ratio Data

Represents numeric data – exact value can be measured

Eg- height, weight, age, salary etc.

Absolute zero is available for ratio data

Added , subtracted, multiplied or divided – ***Yes***

Central tendency –measured by mean, mode, median - ***Yes***

Methods of dispersion – standard deviation - ***Yes***

Quantitative

- Discrete Attribute
  - Has only a finite or countably infinite set of values
    - E.g., Profession, Roll No, Rank of students
  - Sometimes, represented as integer variables
  - Note: Binary attributes are a special case of discrete attributes
- Continuous Attribute
  - Has real numbers as attribute values
    - E.g., temperature, height, or weight
  - Practically, real values can only be measured and represented using a finite number of digits
  - Continuous attributes are typically represented as floating-point variables

EXPLORING STRUCTURE OF DATA

In case of std data set – data dictionary available for reference

Data dictionary – metadata repository

Detailed information plus description

if data dictionary not available – use standard library function of the ML tool

Standard data set from UCI Machine learning repository is used (University of California, Irvine)

EXPLORING STRUCTURE OF DATA

mpg	cylinder	displace-ment	horse-power	weight	accel-eration	model year	origin	car name
18	8	307	130	3504	12	70	1	Chevrolet chev-elle malibu
15	8	350	165	3693	11.5	70	1	Buick skylark 320
18	8	318	150	3436	11	70	1	Plymouth satellite
16	8	304	150	3433	12	70	1	Amc rebel sst
17	8	302	140	3449	10.5	70	1	Ford torino
15	8	429	198	4341	10	70	1	Ford galaxie 500
14	8	454	220	4354	9	70	1	Chevrolet impala
14	8	440	215	4312	8.5	70	1	Plymouth fury iii
14	8	455	225	4425	10	70	1	Pontiac catalina
15	8	390	190	3850	8.5	70	1	Amc acbassador dpl

Title: Auto-Mpg Data

Sources: (a) Origin: This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University

Number of Instances: 398

Number of Attributes: 9 including the class attribute

Attribute Information:

1. mpg: continuous
2. cylinders: multi-valued discrete
3. displacement: continuous
4. horsepower: continuous
5. weight: continuous
6. acceleration: continuous
7. model year: multi-valued discrete
8. origin: multi-valued discrete
9. car name: string (unique for each instance)

A - Exploring Numerical Data

Two most effective mathematical plots to explore numerical data – **box plot and histogram**

Understanding the Central tendency- understand the central point of a set of data.

- ***MEAN*** – sum of all data values divided by the count of data elements – shifts drastically even due to small number of outliers
- ***MEDIAN*** - value of the element appearing in the middle of an ordered list
- Impacted by data values appearing in beginning or end of range – close to min or max values
  - Especially sensitive to outliers – unusually high / low values
- Given e.g. – for mean /median- horsepower – not available, mpg, weight, acceleration – low deviation; cylinders, displacement, origin -high deviation

Understanding the data spread

Looking closely at attributes – granular view of the data spread

Dispersion of data

Position of the different data values

Measuring the data dispersion

Variance

Standard deviation

$$s^2 = \frac{\sum (x - \bar{x})^2}{n-1}$$

SD = √s

$$\text{Variance}(x) = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2$$, where x is the variable or attribute whose variance is to be measured and n is the number of observations or values of variable x.

Standard deviation of a data is measured as follows:

$$\text{Standard deviation}(x) = \sqrt{\text{Variance}(x)}$$

Consider the data values of two attributes

1. Attribute 1 values : 44, 46, 48, 45, and 47
2. Attribute 2 values : 34, 46, 59, 39, and 52

Calculate the Mean, Median, Variance and Standard Deviation

$$\text{Variance} = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2$$

$$= \frac{44^2 + 46^2 + 48^2 + 45^2 + 47^2}{5} - \left(\frac{44 + 46 + 48 + 45 + 47}{5}\right)^2$$

$$= \frac{1936 + 2116 + 2304 + 2025 + 2209}{5} - \left(\frac{230}{5}\right)^2 = \frac{10590}{5} - (46)^2 = 2$$

For attribute 2,

$$\text{Variance} = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2$$

$$= \frac{34^2 + 46^2 + 59^2 + 39^2 + 52^2}{5} - \left(\frac{34 + 46 + 59 + 39 + 52}{5}\right)^2$$

$$= \frac{1156 + 2116 + 3481 + 1521 + 2704}{5} - \left(\frac{230}{5}\right)^2 = \frac{10978}{5} - (46)^2 = 79.6$$

Standard Déviation
Attribute 1 = 1.41
Attribute 2 = 8.88

Measuring Data Values positions

First quartile or Q1 -- first half of the data is divided into two halves so that each half consists of one quarter of the data set, that median is called first Quartile

Q2 : Median is called as Second Quartile

Third quartile or Q3 : if the second half of the data is divided into two halves, then that median of the second half is known as third quartile or Q3

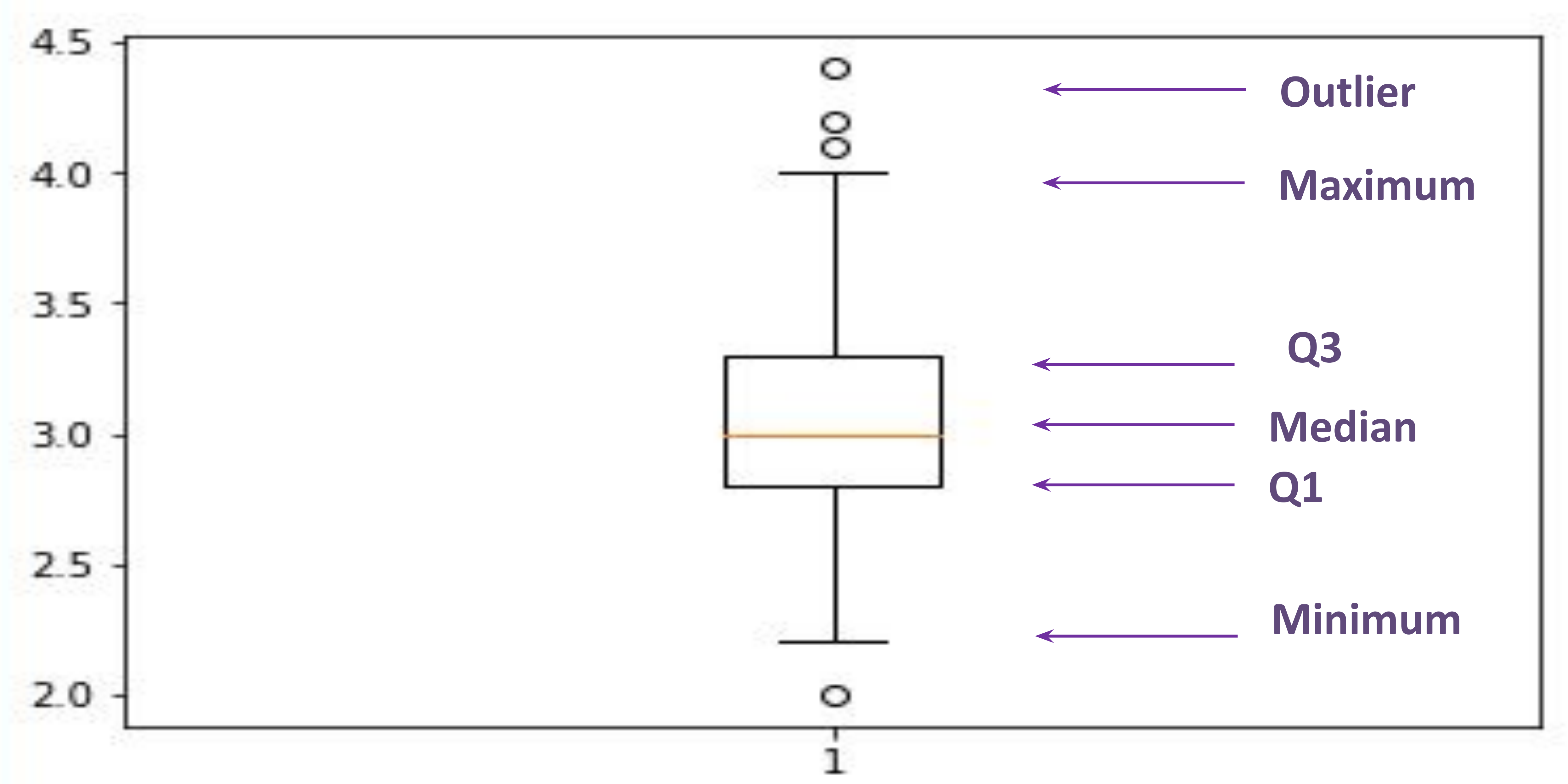
Difference – between quarters reason for attribute displacement

	Cylinders	Displacement	Origin
Minimum	3	68	1
Q1	4	104.2	1
Median	4	148.5	1
Q3	8	262	2
Maximum	8	455	3

36.2
44.3
113.5
193

Box Plots

Effective mechanism to et a one-shot view and understand the nature of data using minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum



IQR – Inter Quartile Range

Outliers – values that lie outside the 1.5 |X| IQR limits

Draw a box plot for

12, 5, 22, 30, 7, 36, 14, 42, 15, 53, 25

Step 1 – arrange in ascending order

Step 2 - Find the median or middle value that splits the set of data into two equal groups. If there is no one middle value, use the average of the two middle values as the median

Step 3. Find the median for the lower half of the data set

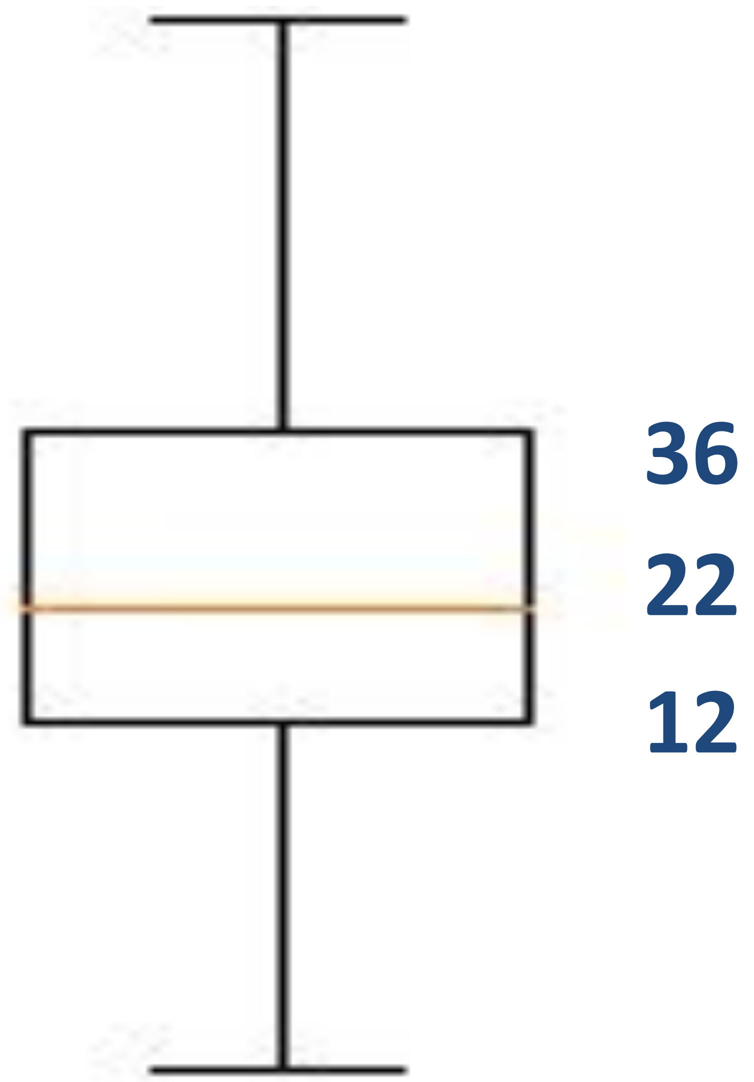
Step 4. Find the median for the upper half of the data set.

5, 7, 12, 14, 15, 22, 25, 30, 36, 42, 53

lower quartile (12) median (22) upper quartile (36)

Draw a box plot for

Step 5. Use these five values to construct a box plot: minimum, lower quartile, median, upper quartile, maximum - minimum and maximum whiskers by calculating the IQR



IQR = Q3 - Q1 = **24**

MAX = Q3 + 1.5 * IQR = **72**

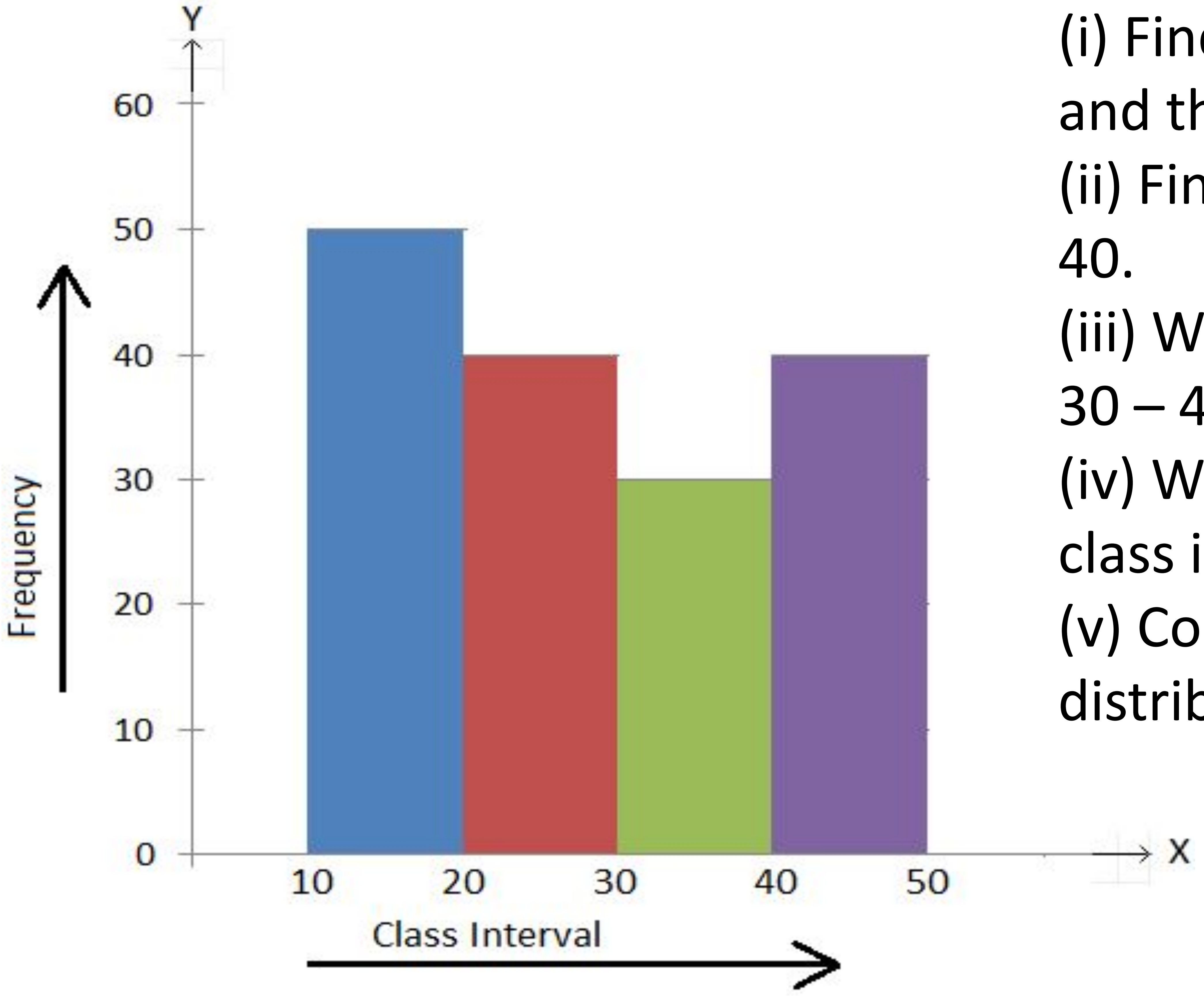
MIN = Q1 - 1.5 * IQR = **- 24**

Draw a box plot for

- 7, 3, 35, 14, 9, 7, 8, 12, 2
- 12, 5, 22, 30, 7, 36, 14, 42, 15, 53, 25
- 76, 57, 63, 66, 72, 73, 75, 70, 75, 79, 57, 58, 66, 67, 68, 70, 76, 70, 78, 67, 69, 70
- 76, 57, 63, 66, 72, 73, 75, 70, 75, 52, 89, 57, 58, 66, 67, 68, 70, 76, 70, 78, 67, 69, 70

Histogram

- Another plot for effective visualization of numeric attributes
- Distribution of numeric data into series of intervals – bins.
- Different shapes –based on nature of data
- Difference between histogram and box plot
  - Focus of histogram is to plot ranges of data values (bins), umber of elements – data distribution; size of each bar – will vary
  - Box plot – divide data elements into 4 equal portions, each portion contains equal no of data elements



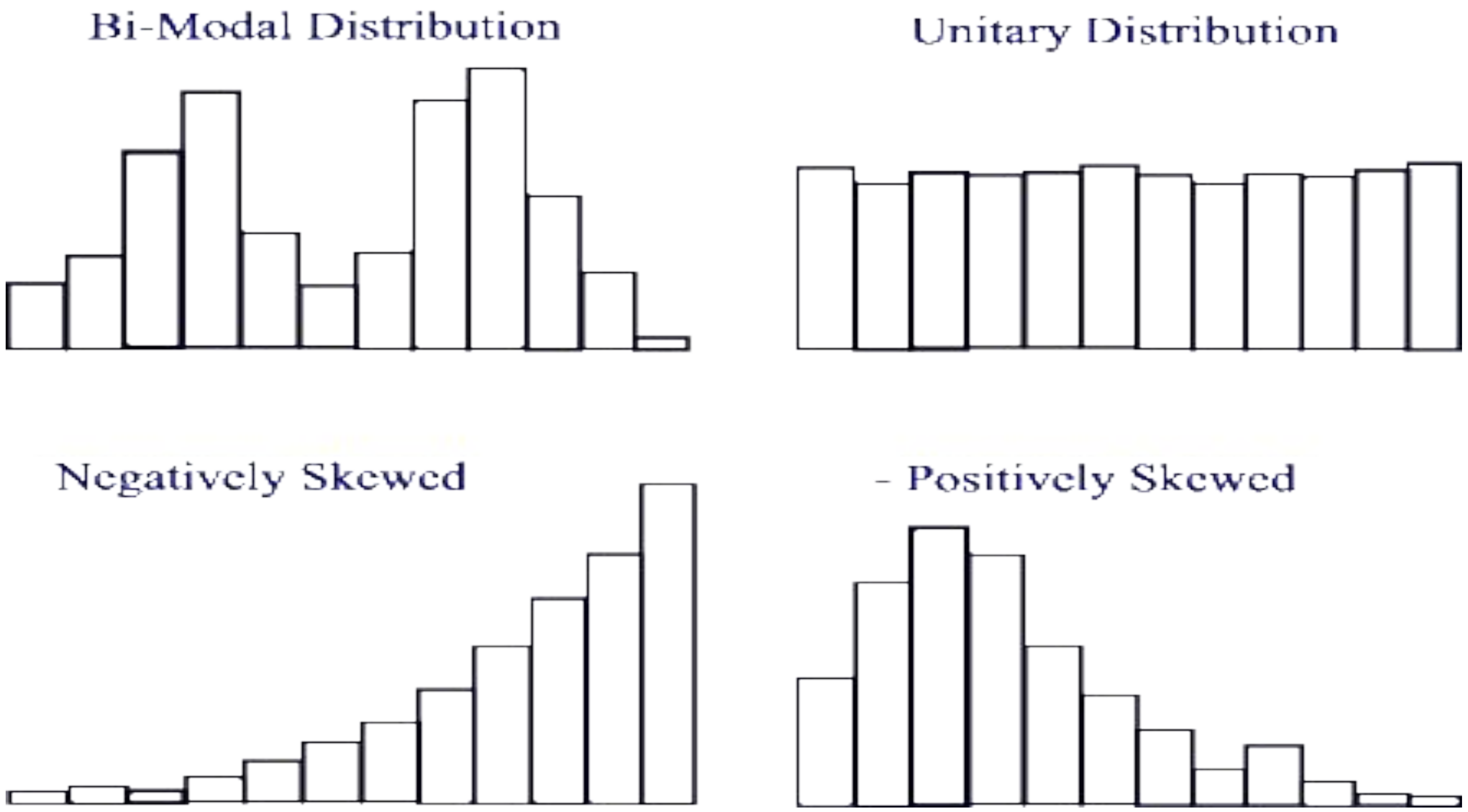
(i) Find the class intervals having the greatest and the least frequencies.

(ii) Find the class interval whose frequency is 40.

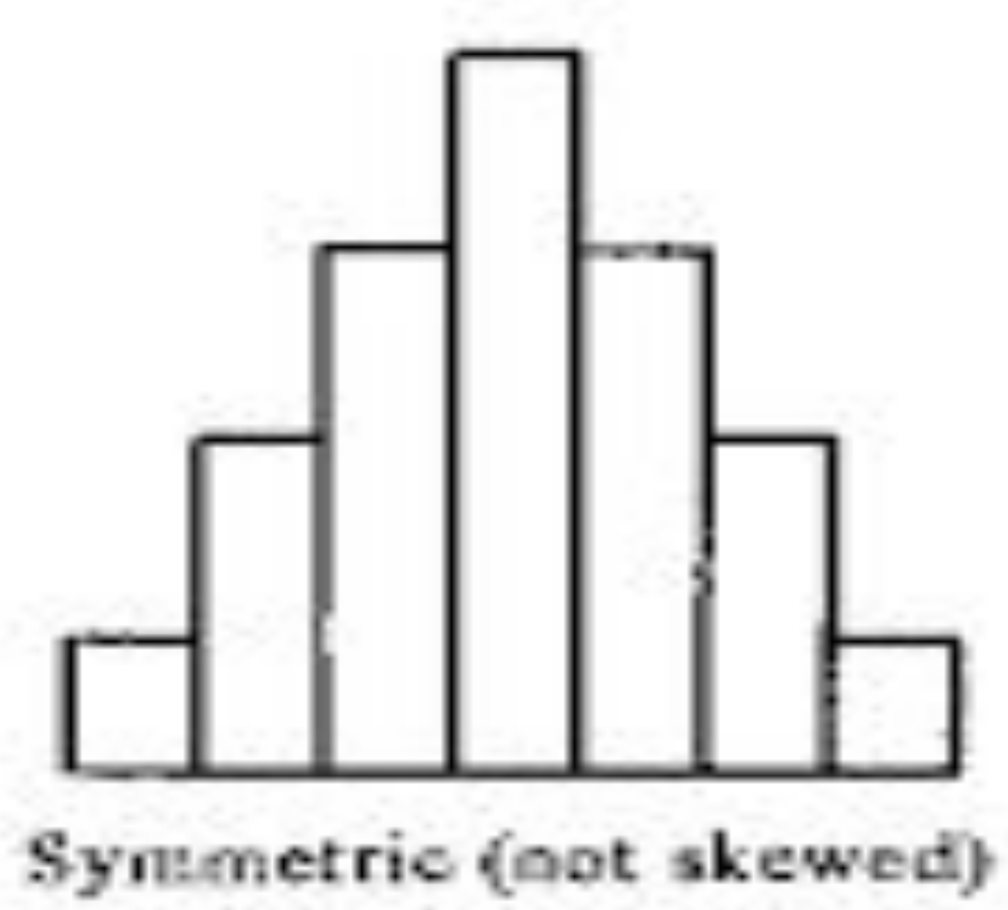
(iii) What is the frequency of the class interval 30 – 40?

(iv) What is the cumulative frequency of the class interval 30 – 40?

(v) Construct the frequency table of the distribution



Histogram is uniform – all values are equally likely to occur]



B - Exploring Categorical Data

Multiple ways – for numeric data compared to categorical data

Count – data elements fall in the category

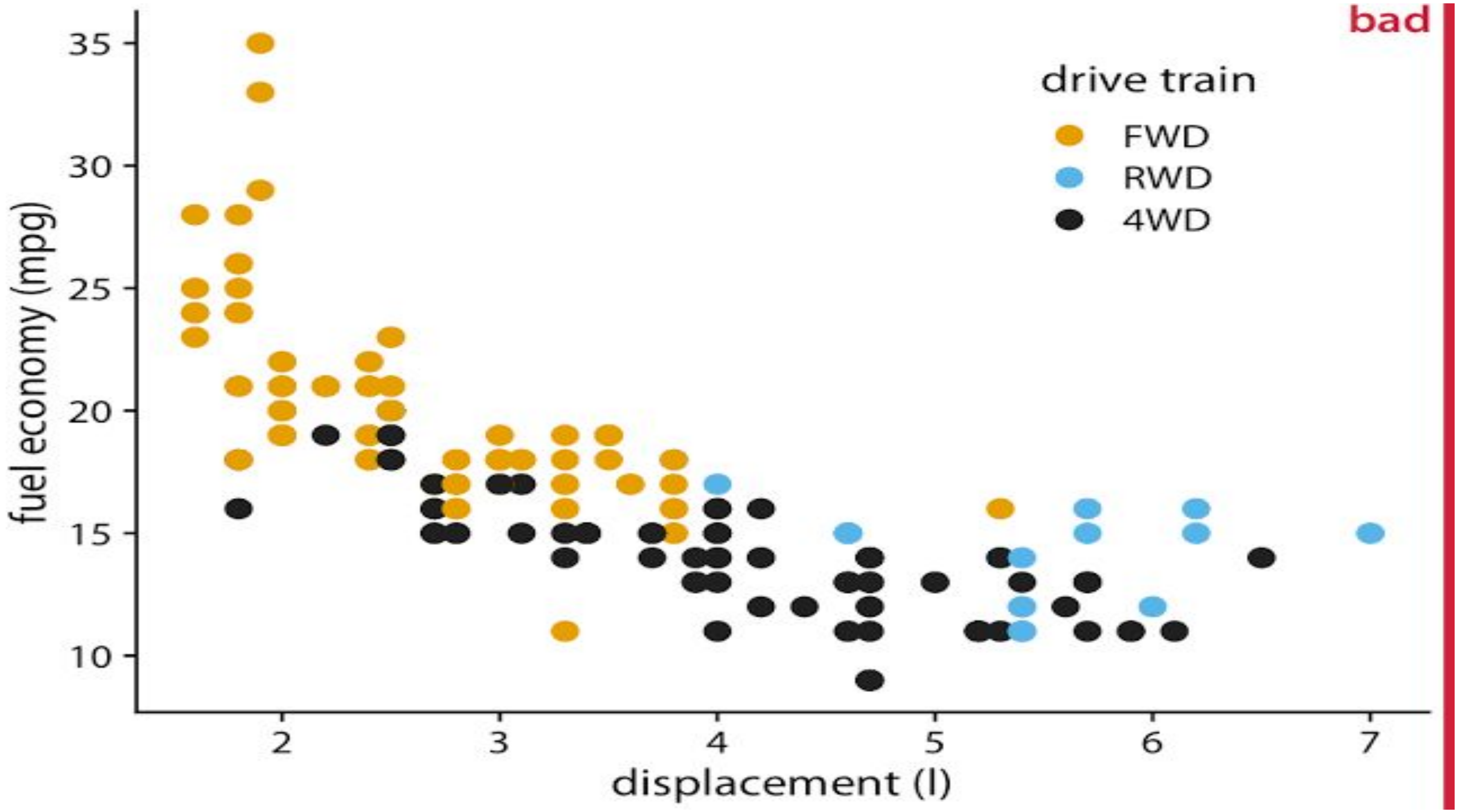
Proportion / percentage – count belonging in that category

Mode – frequency of the data value which is highest

Exploring relationship between variables

- Scatter plot
  - Helps visualizing bivariate relationship – relationship between two variables
  - 2D plot – points or dots are drawn provided by values of attributes – attr1 – x axis ; attr2 – y axis
- Two way cross-tabulations
  - Cross tabulations or contingency tables – relationship of two categorical attributes in a concise way.
  - Matrix format

Scatter plot



Two-Way tabulation

Cylinder/ Model year	70	71	72	73	74	75	76	77	78	79	80	81
3	0	0	1	1	0	0	0	1	0	0	1	0
4	7	13	14	11	15	12	15	14	17	12	25	23
5	0	0	0	0	0	0	0	0	1	1	1	0
6	4	8	0	8	7	12	10	5	12	6	2	7
8	18	7	13	20	5	6	9	8	6	10	0	1

Success of ML – **quality of Data – DATA QUALITY**

Right quality – better prediction

Two types of problems – flaws in data -

- Data elements without a value or missing value
- Outliers – data with surprisingly different values

Factors leading to data quality issues –

Incorrect sample set selection

Sample set from festival – predict sales in features

Errors in data collection – outliers / missing values

Wrongly recording of data – 20.67 may be 206.7 / 2.067

Issues in data quality, previous, need to be remediate

First one can be remediated by proper sampling technique

However human errors are bound to happen

Handling errors

Handling Outliers

Remove outliers – simplest approach

Imputation - by mean / medain

Capping – values outside 1.5 |X| IQR – cap them – lower limits with 5th percentile; upper limits with 95th percentile

Handling Missing values

Eliminate records having missing values

Imputing missing values

Estimate Missing values

Dimensionality Reduction

Last 2 decades

high-dimensional datasets with 20,000 or more features

Wide-spread adoption of social networking – text/image classification

High dimensional datasets – high amount of computational space and time

Not all are useful – degrade the performance of ML

ML algos – better performance if no of features is reduced

Also easier to understand the model

Dimensionality reduction – refers to techniques of reducing the dimensions of data by combining the original attributes and creating new

Common approaches – PCA (Principal Component Analysis) and SVD (singular Value Decomposition)

-Cont

Feature Subset Selection

Or Feature selection – both for supervised and unsupervised learning

Try to find out optimal subset of entire features

- Reduces computational cost – without major impact on learning accuracy

Feature – playing insignificant role in classification or grouping – data instances

- Irrelevant features are eliminated – final feature subset

Features – redundant – if other features are more or less same – small no to be selected without causing negative impact to learn model accuracy

END OF UNIT I

Sample programming using python & R on data sets