Learn Something Every Day

Implementation of DS

Data Science in Search Engine

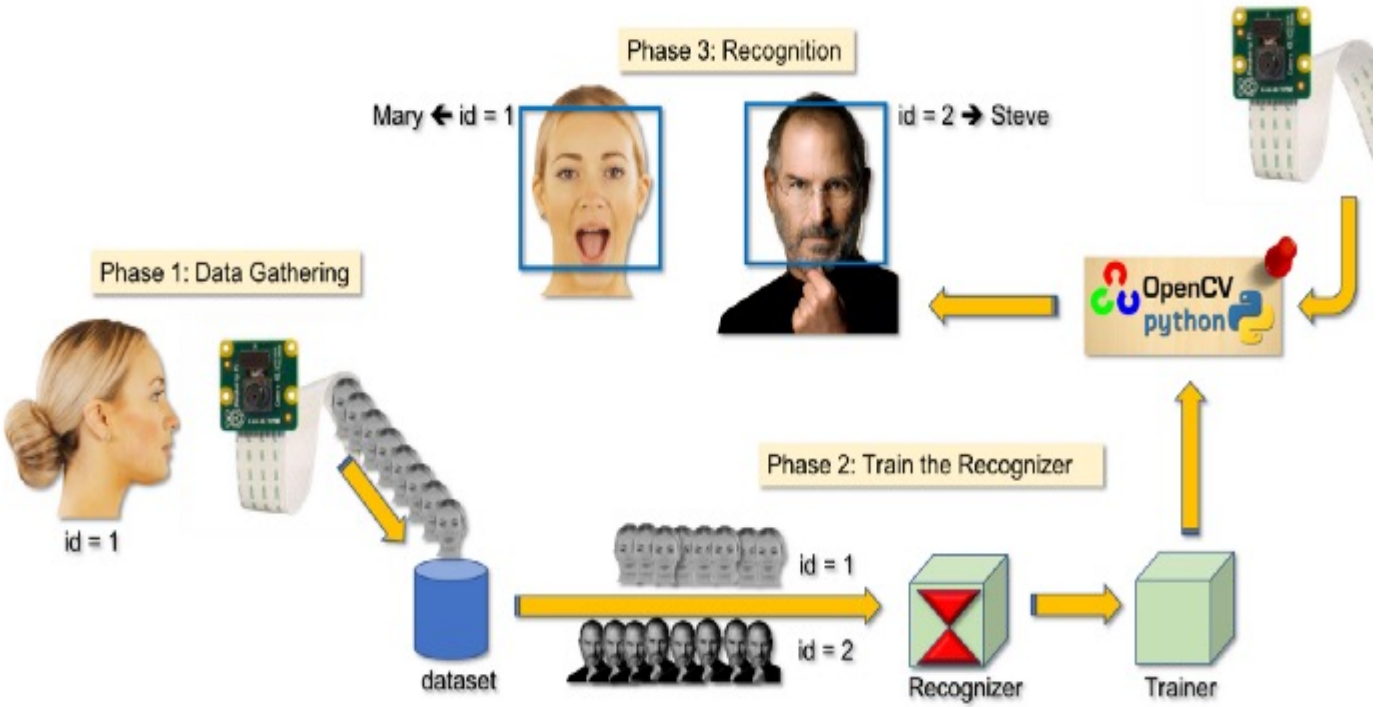
Data Science in Finance

Data Science in Transport

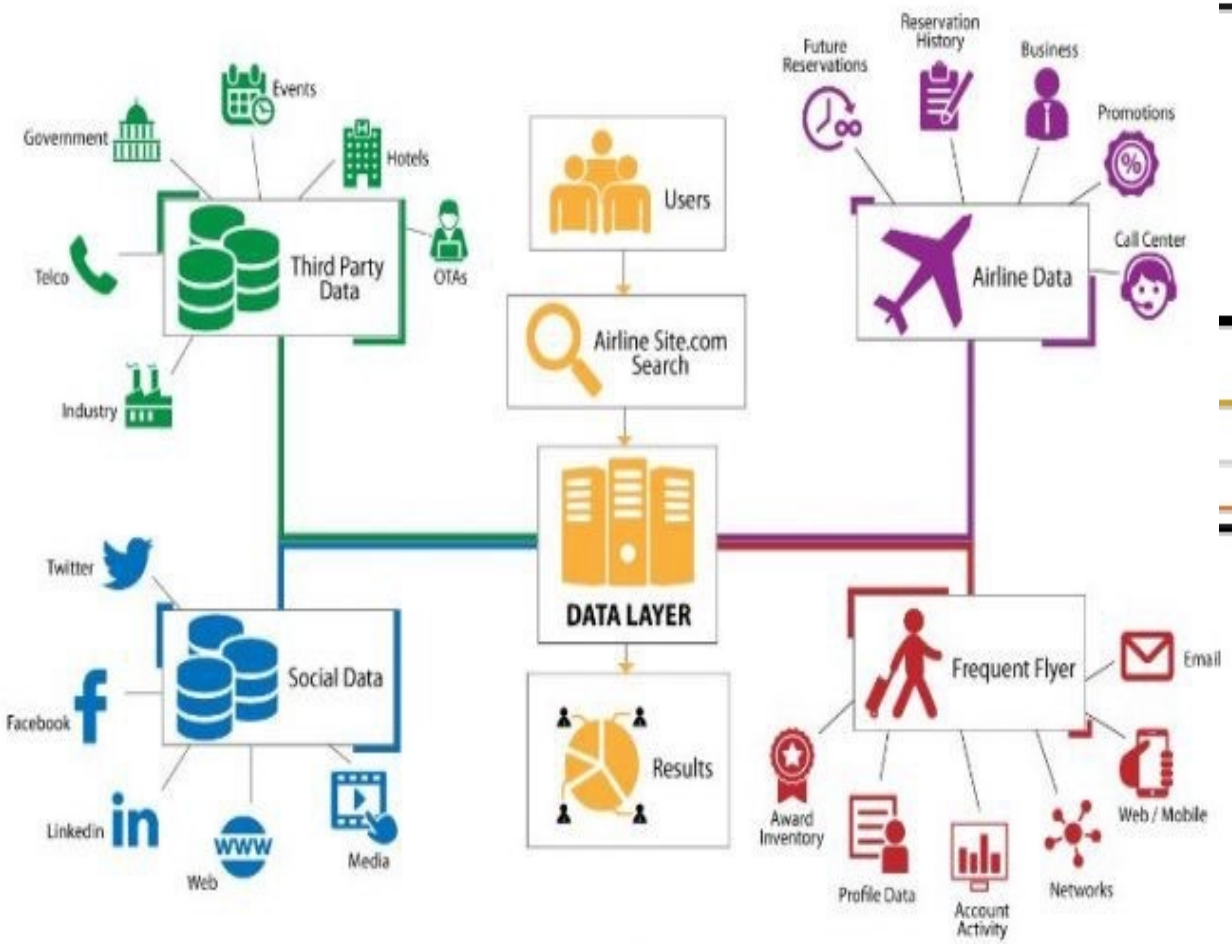
Data Science in E-Commerce

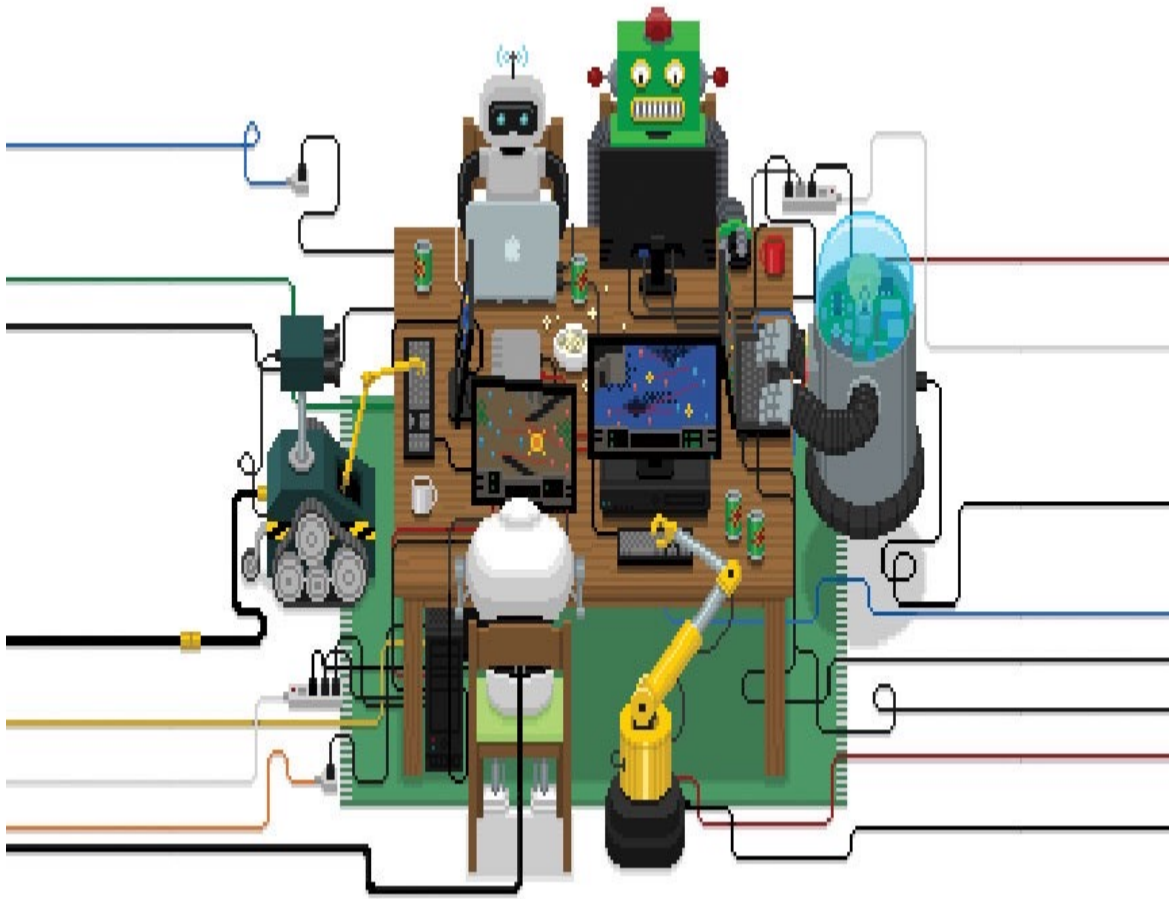
Data Science in Health Care

Data Science in Image recognition



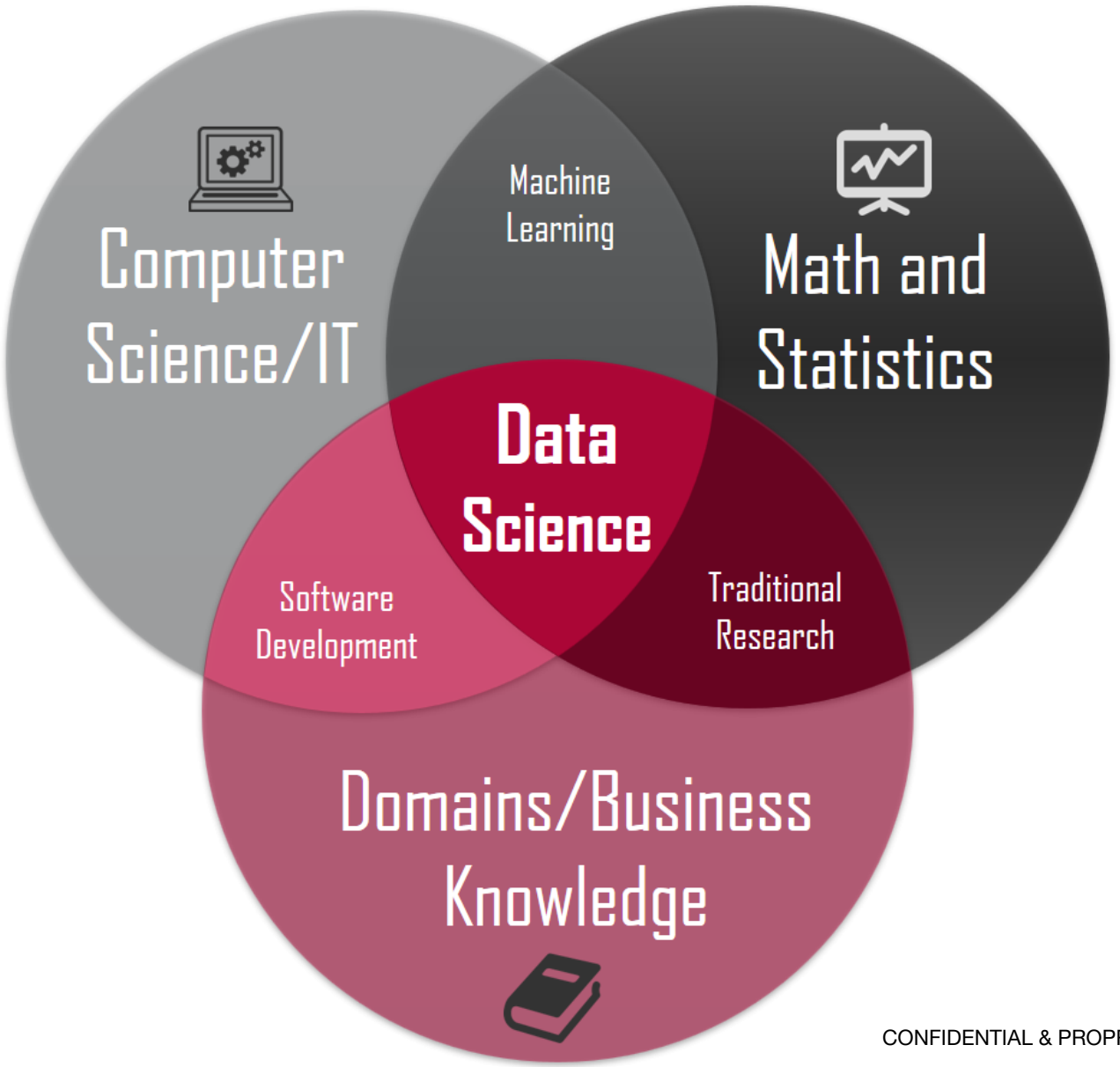
Data Science in Airline Routing Planning

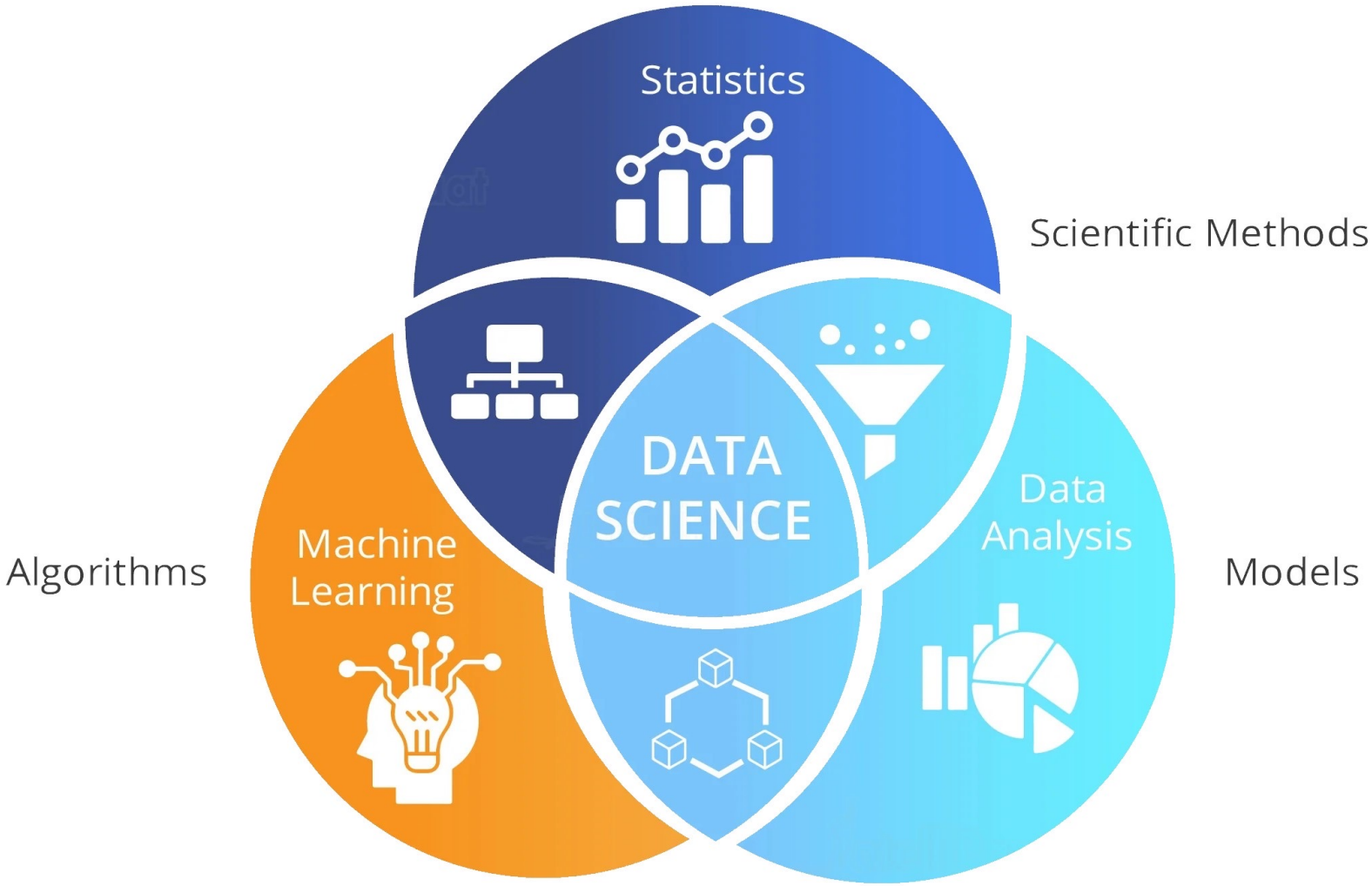




Data Science in Gaming

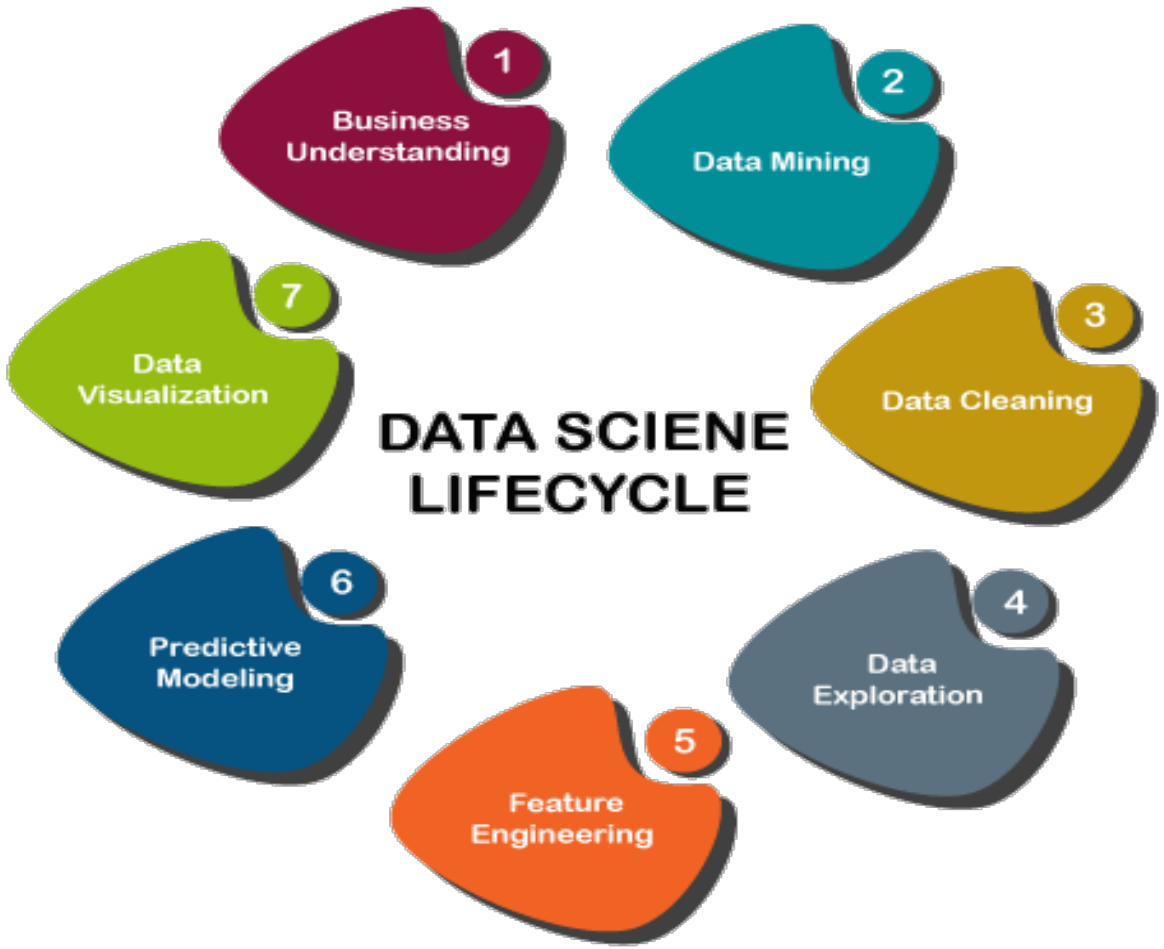
Introduction to DS





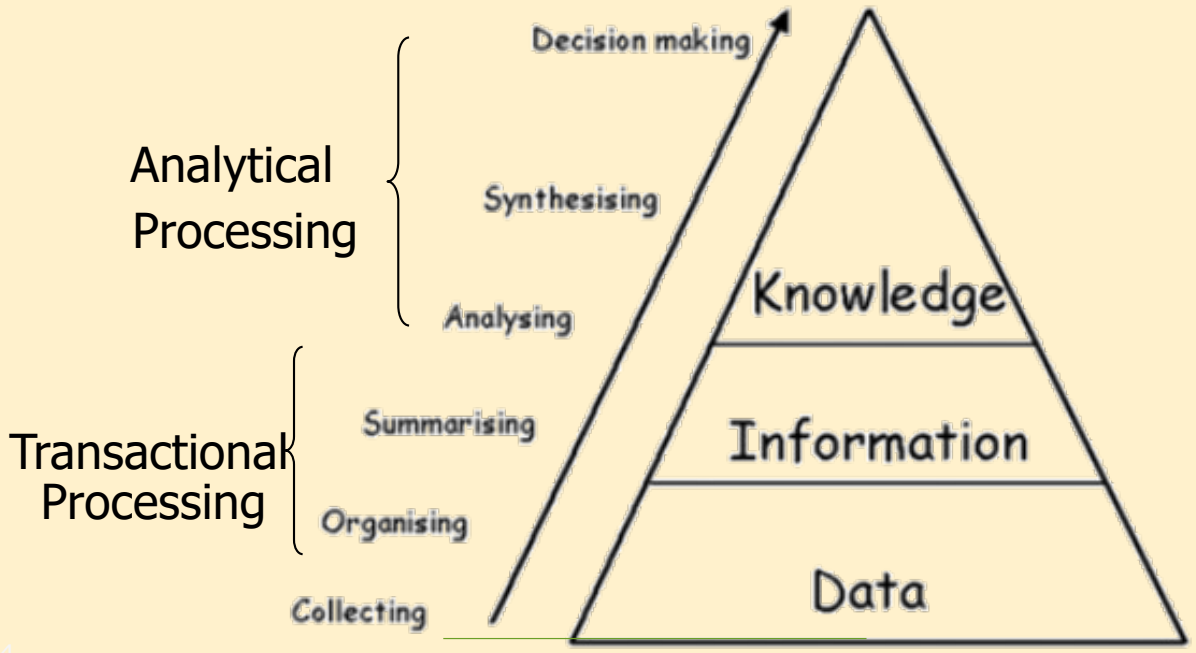
Data Science Basic

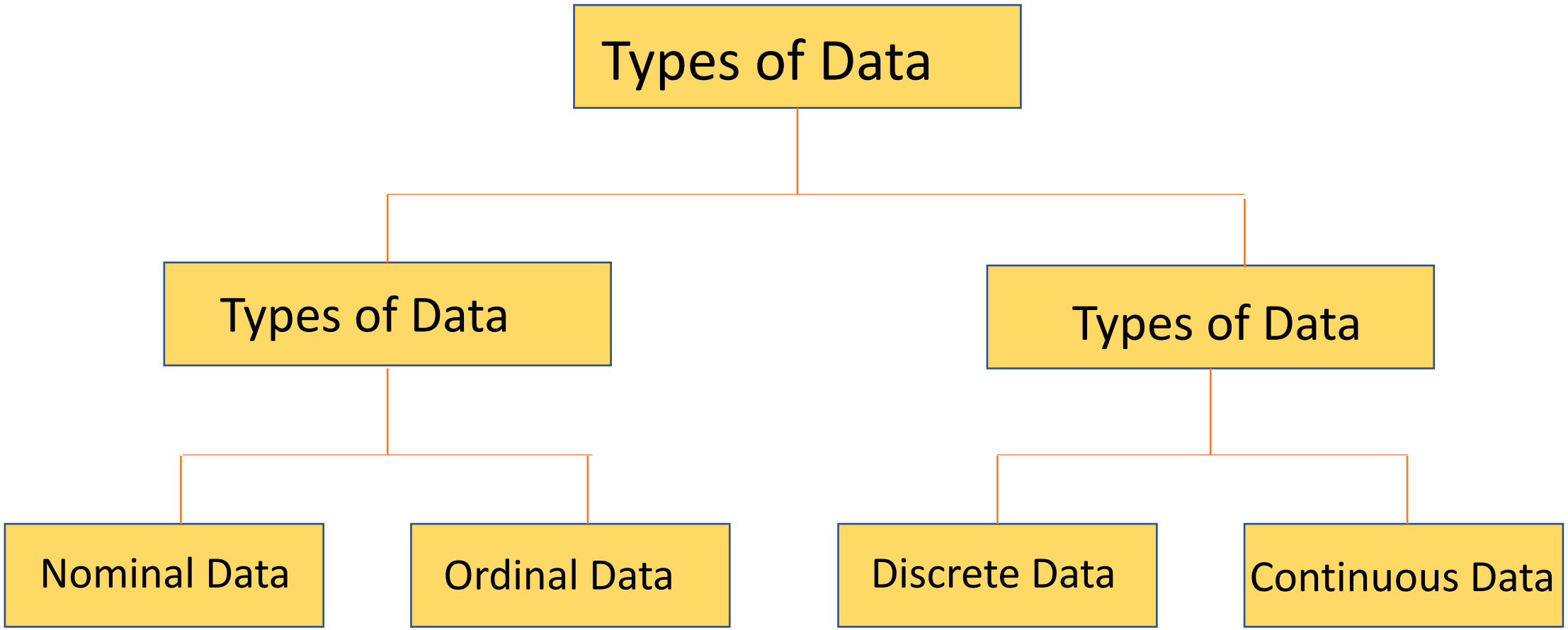
1. Understand the project objectives and requirements from a business perspective, and then convert this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

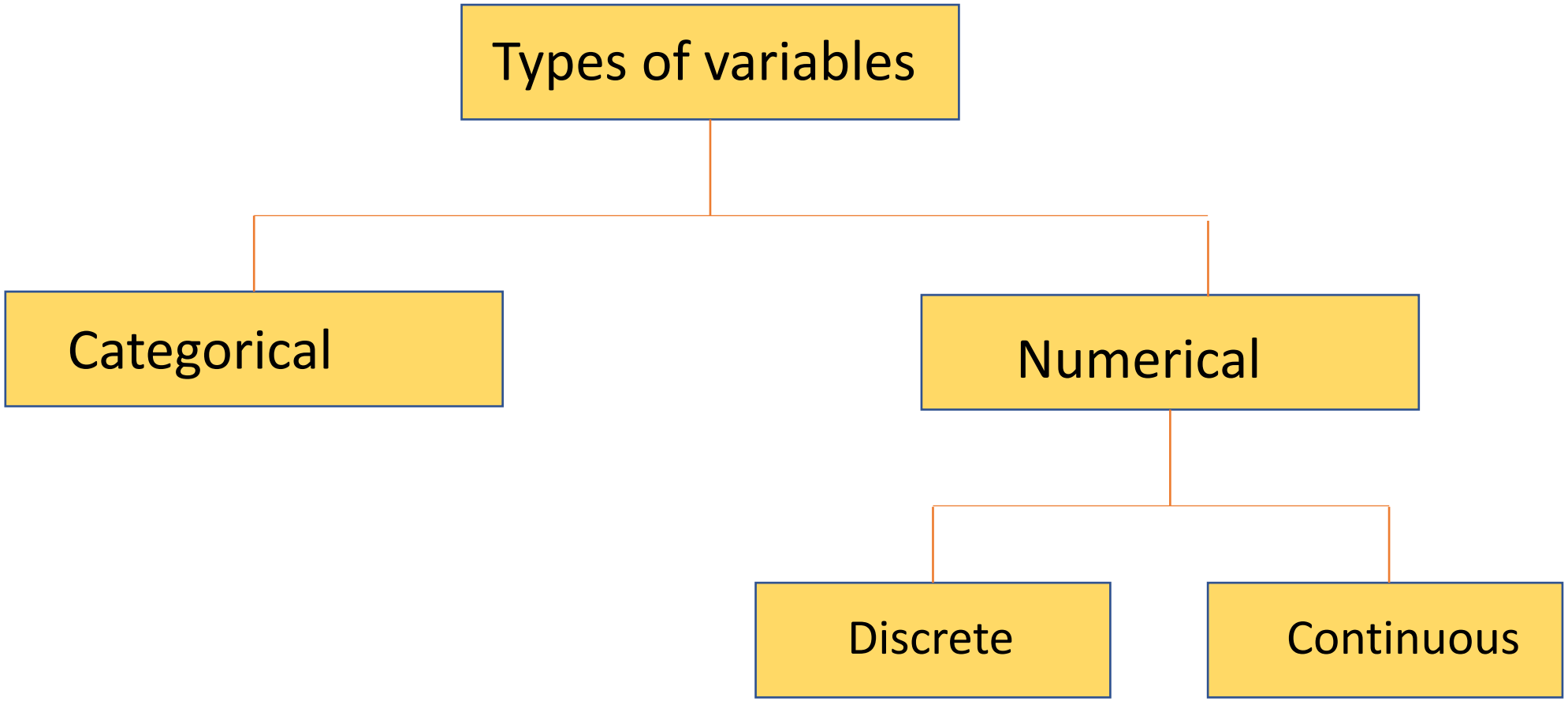


2. Data mining is the process of sorting through large data sets to identify patterns and relationships that can help solve business problems through data analysis

DIKW

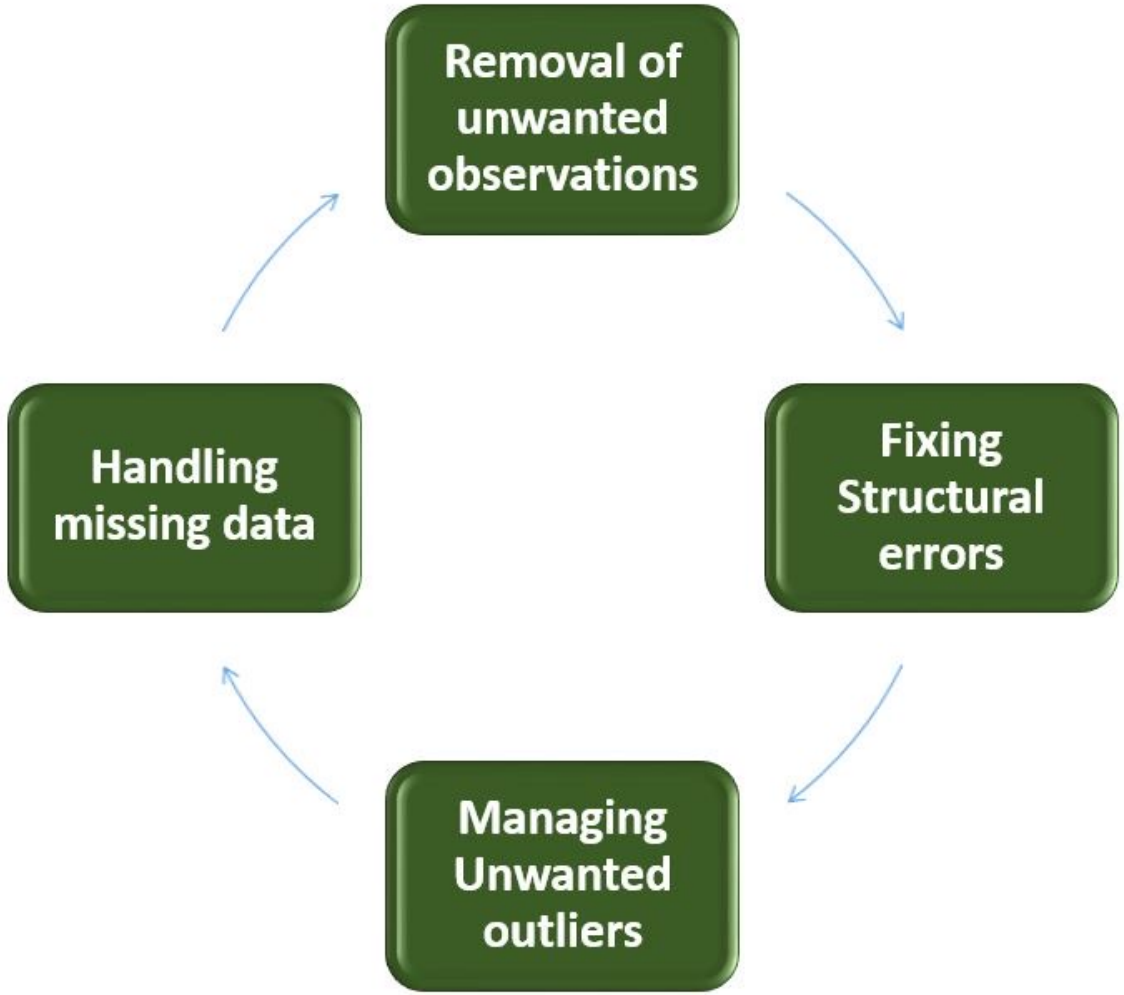






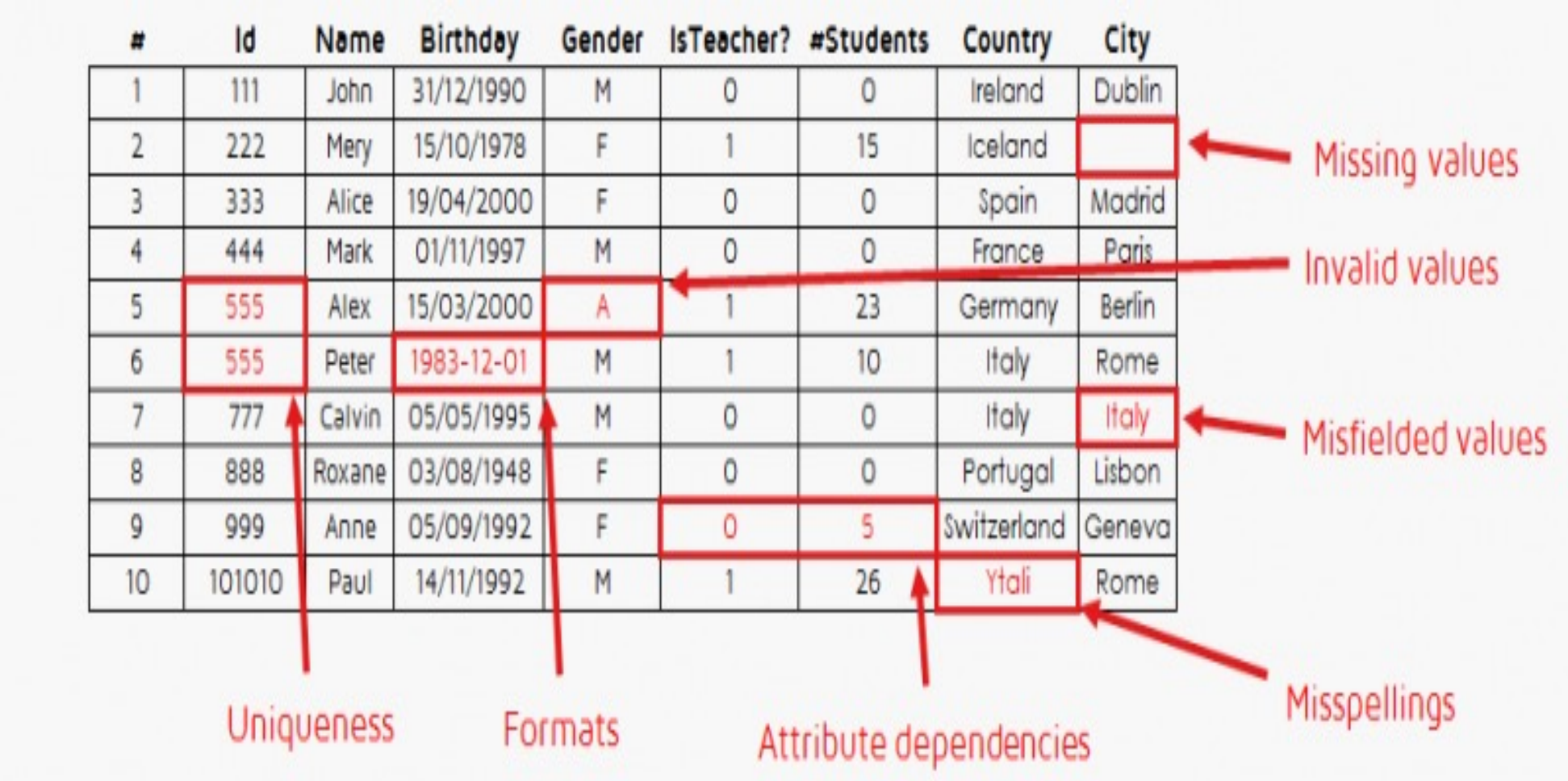
3. Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset

Data Cleaning



DATA CLEANING STEPS

Removing unwanted observations	• Duplicate/ redundant or irrelevant values deletion .
Missing Data handling	• Fixing issue of unknown missing values
Structural error solving	• Fixing problems with mislabeled classes, typesin names of features, same attribute with different name etc.
Outliers Management	• Unwanted values which are not fiting in datasets.



4. Data exploration refers to the initial step in data analysis in which data analysts use data visualization and statistical techniques to describe dataset characterizations, such as size, quantity, and accuracy, in order to better understand the nature of the data.

Steps in the Data Analysis Process

Step 1: Decide on the objectives or Pose a Question

Step 2: What to Measure and How to Measures

Step 3: Data Collection

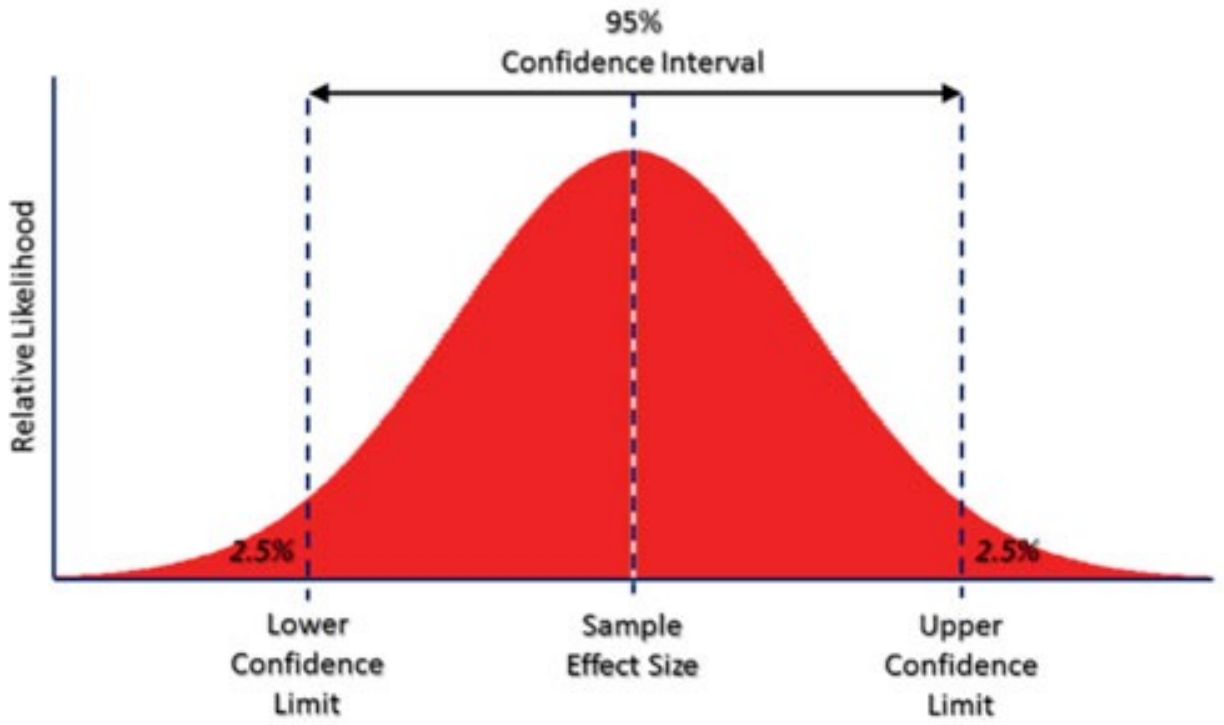
Step 4: Summarizing and Visualizing Data

Step 5: Data Modelling

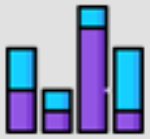
5. Feature engineering is a machine learning technique that leverages data to create new variables that aren't in the training set.

6. Predictive modeling is a mathematical process used to predict future events or outcomes by analyzing patterns in a given set of input data.

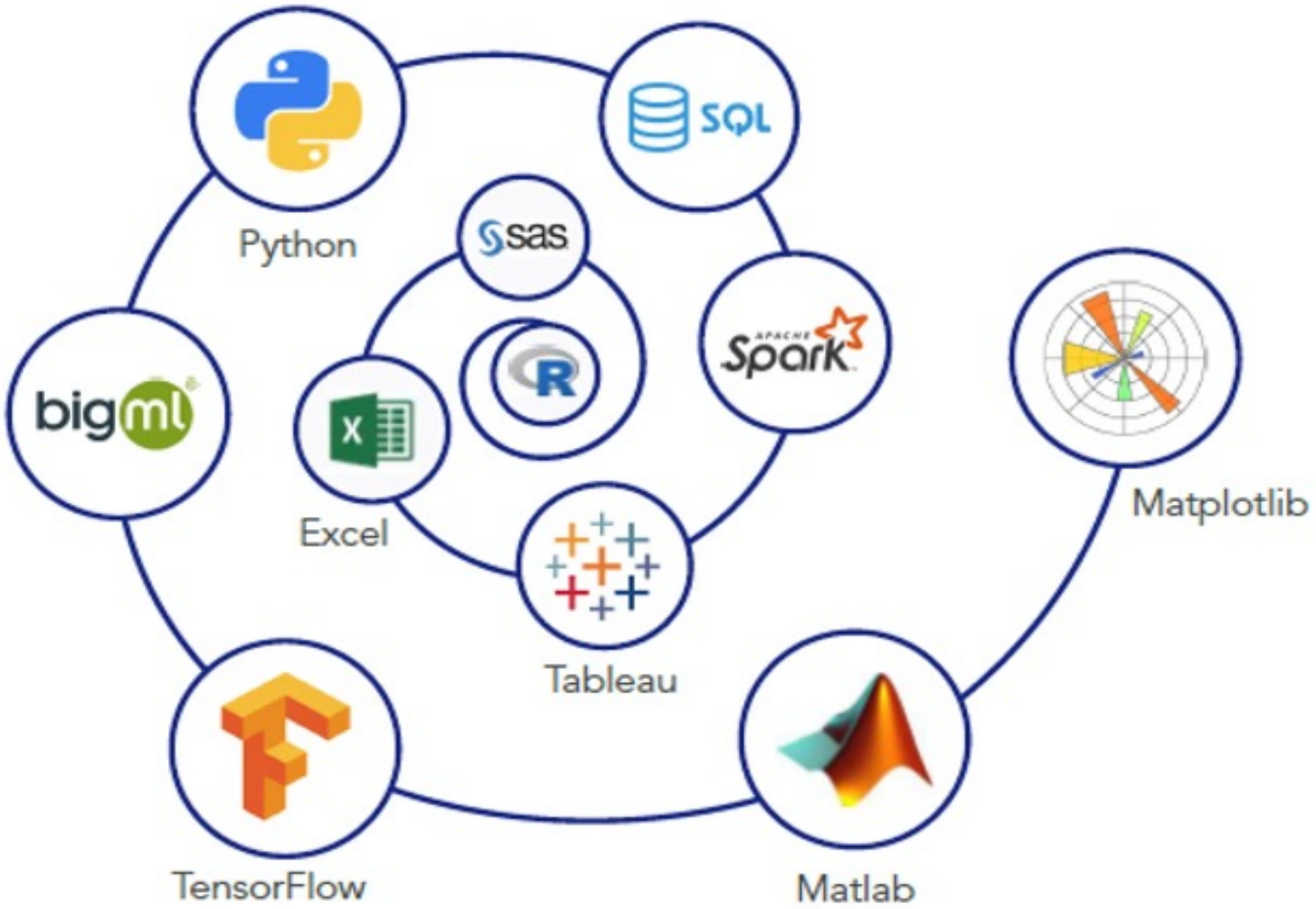
A Confidence Score is a number between 0 and 1 that represents the likelihood that the output of a Machine Learning model is correct and will satisfy a user's request.



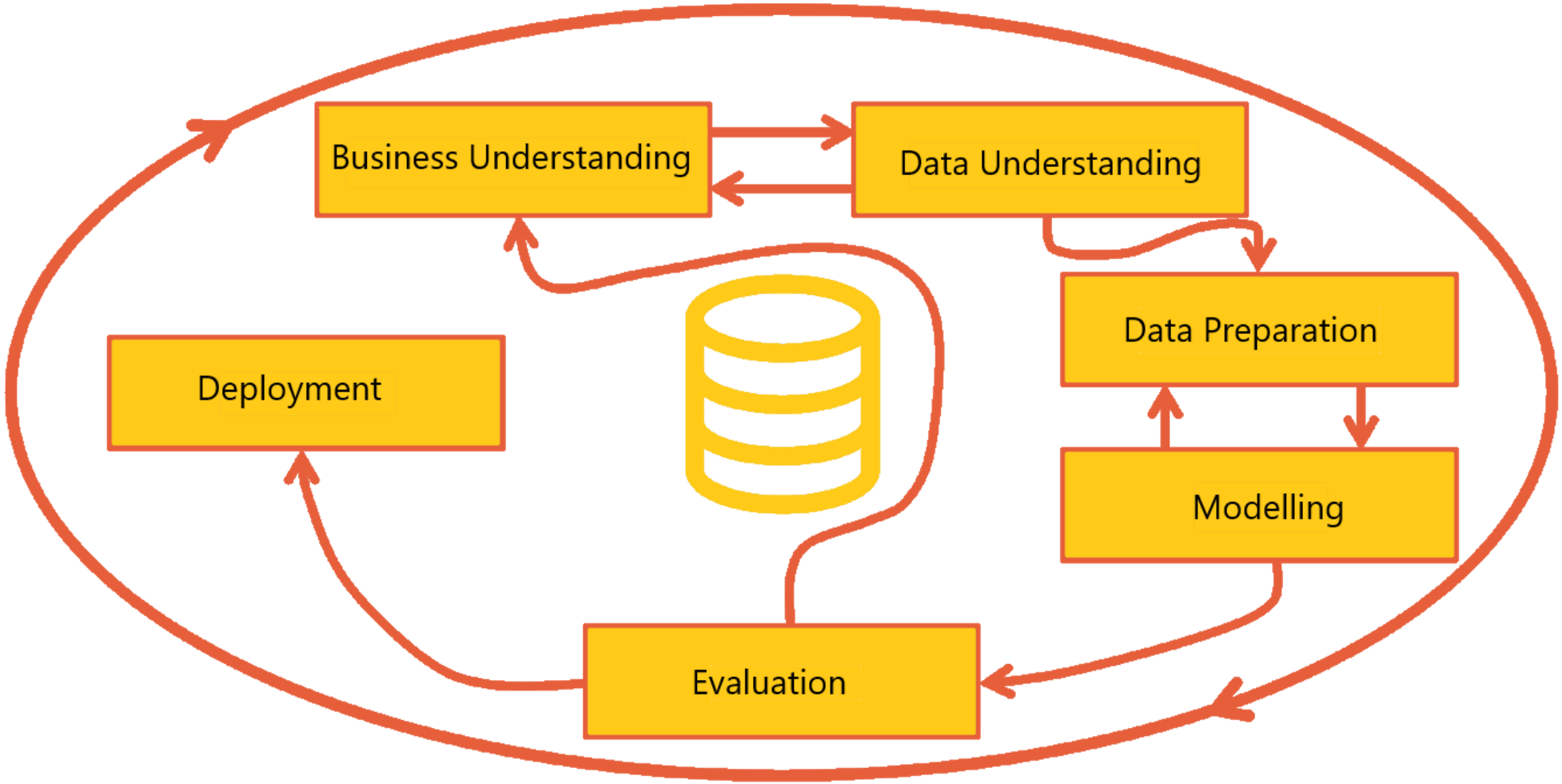
7. Data visualization is the practice of translating information into a visual context, such as a map or graph, to make data easier for the human brain to understand and pull insights.

Chart	Visual	X axis	Y axis	Analysis	Example
Scatter plot/Line Plot		Continuous	Continuous	- Understanding linear, non-linear relationship between two variables - Trend analysis, change in KPI over time	- How does heart rate change with age? - How sales of a company varied over a period of time?
Bar Graph		Categorical /Discrete Continuous	Continuous	- How Y (can be any performance indicator) varies across different categories?	- How sales in 2019 varied for different mobile phone brands? i.e. mobile phone brand is the category and sales is the KPI
Stack Bar Graph		Categorical	Continuous	- Relative comparison of multiple categories within a category	- Comparison of revenue generated by Apple, Samsung & Xiaomi across different products like mobile phone, laptops, television, and headsets
Box Plot		Continuous		- Outlier detection - Analysing data distribution across Median and Inter Quartile Range	- How different sales figures across a year is distributed?
Pie Chart		Categorical& Continuous		- Relative comparison of different categories for one single entity in terms of proportion/percentages	- What percentage of Sales in 2019 is constituted by different products under Apple?
Histogram Plot		Continuous	-	- How distribution of values of x varies across different range buckets?	- Distribution of income across income buckets for developing countries

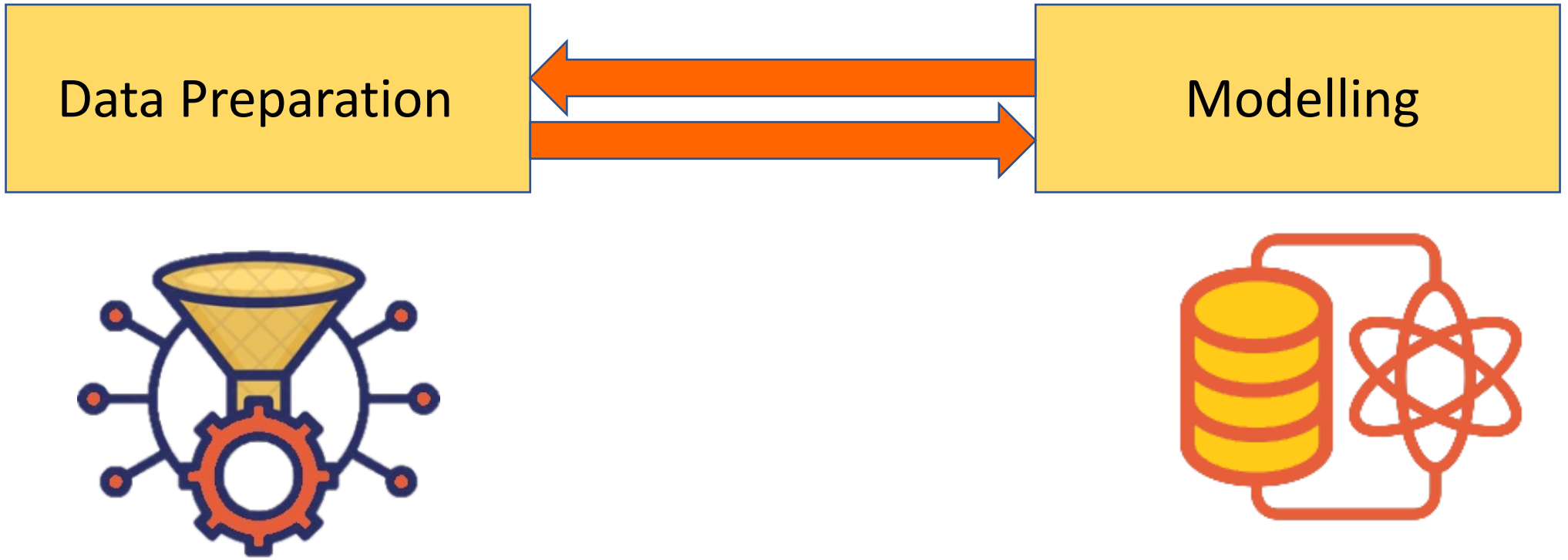
Data Science Tools



Data Science Basics



Modelling in Data Science

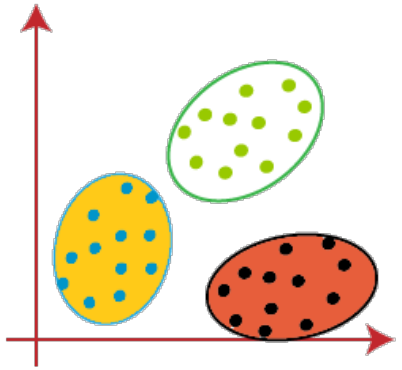


Data Science Task

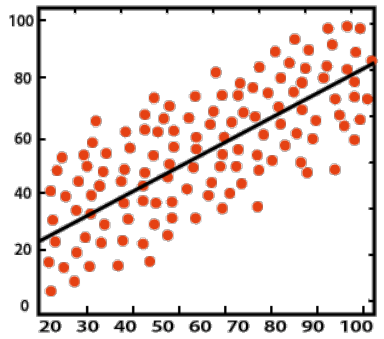
ML Algorithms

- Classification

- Clustering



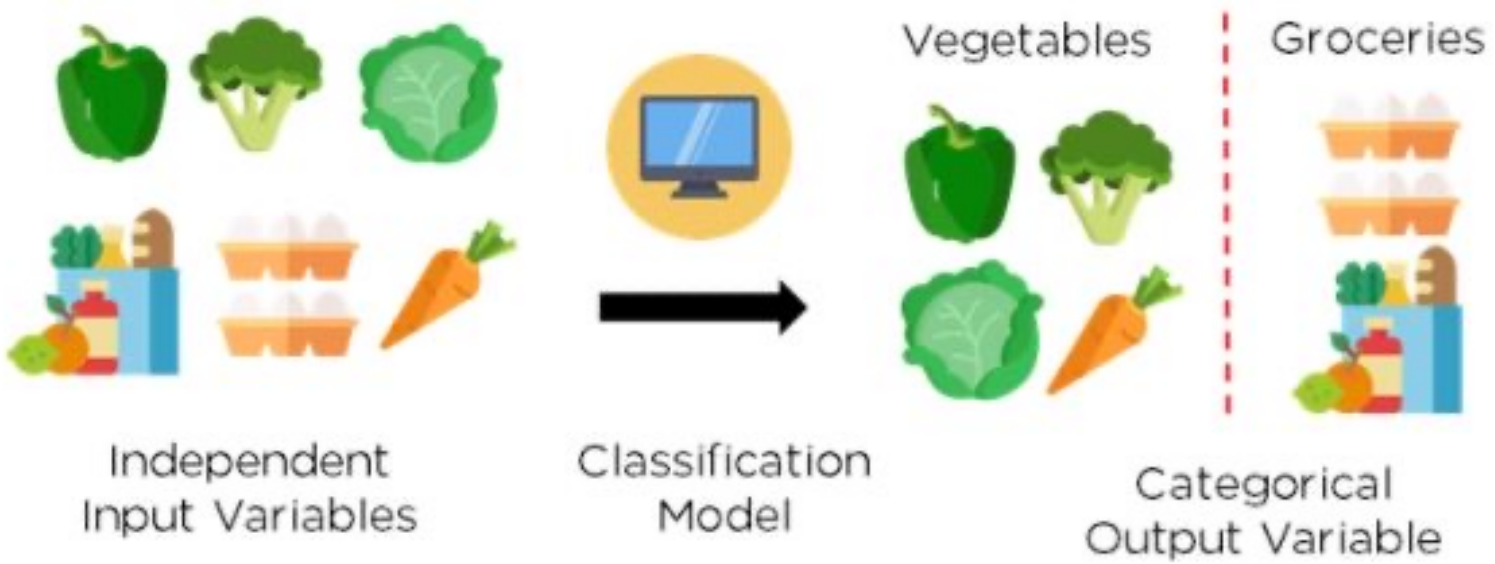
- Regression



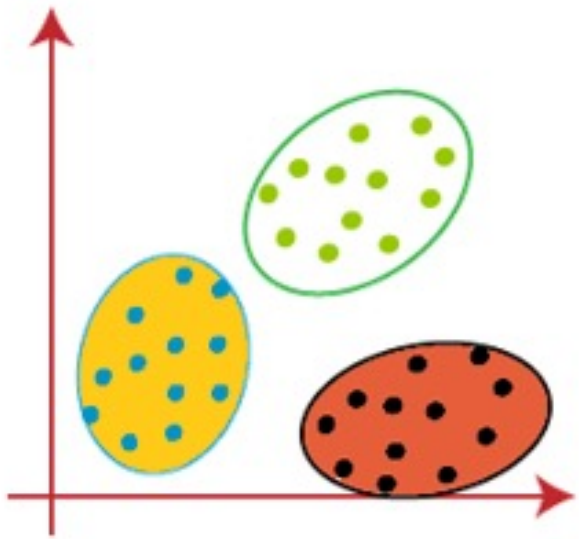
- Association

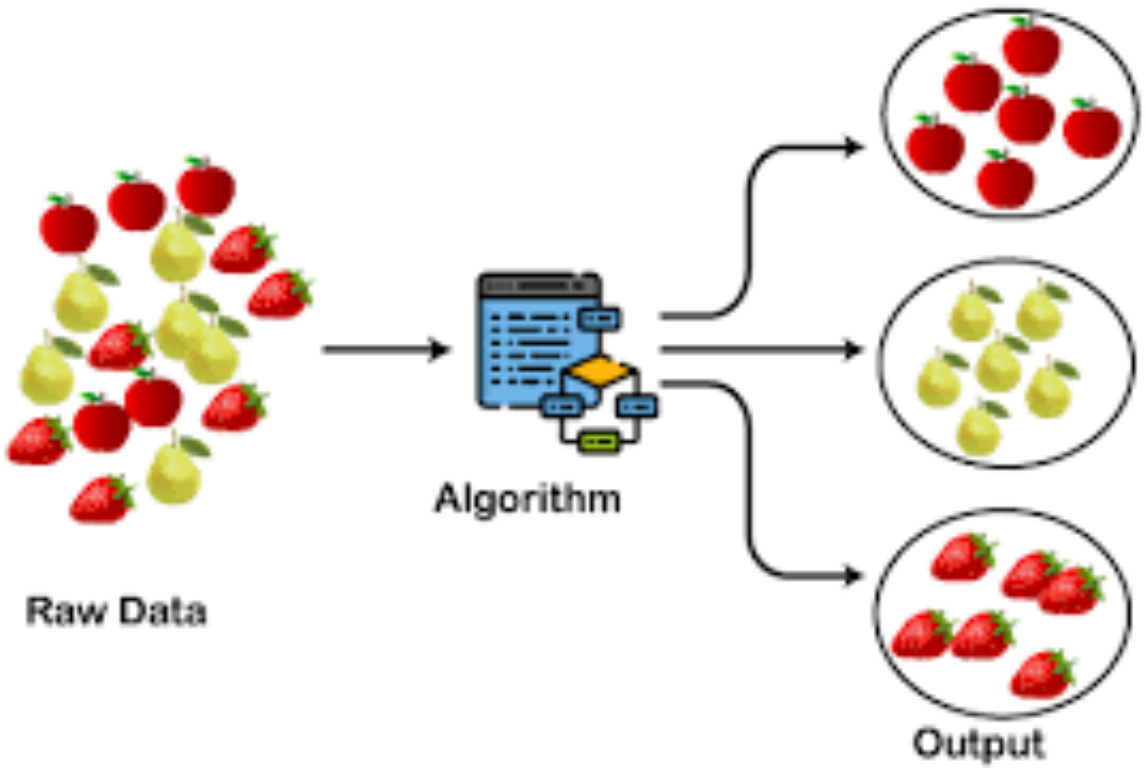
(A, B) ⟶ C
(D, E) ⟶ F
(A, E) ⟶ G

Classification

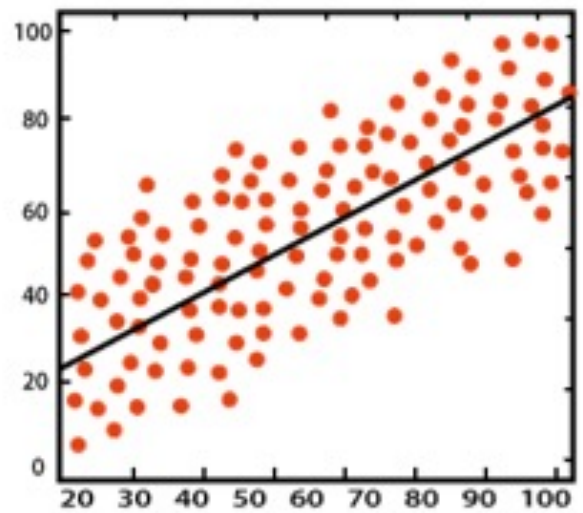


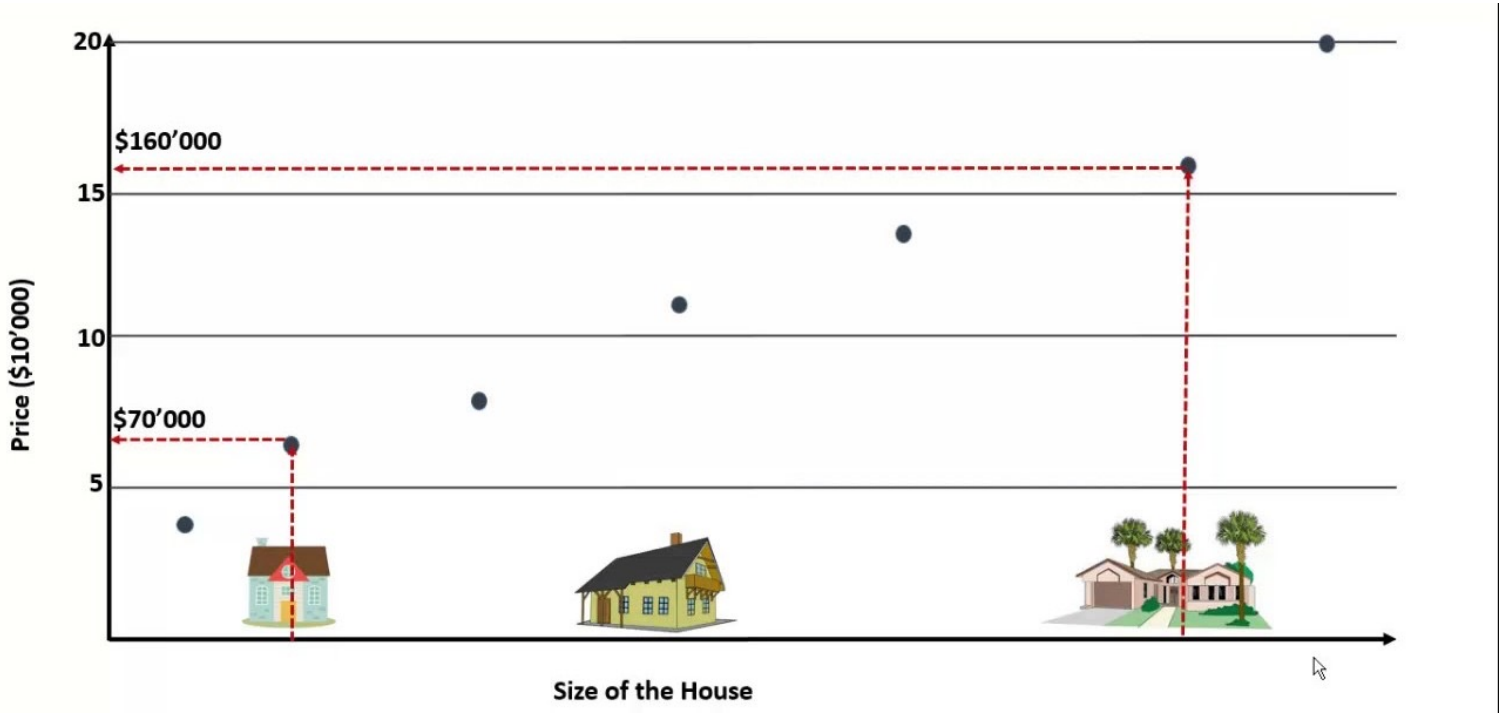
Clustering





Regression





Association

(A, B) ⟶ C

(D, E) ⟶ F

(A, E) ⟶ G

Customer 1

Customer 2

Customer 3

Customer n

THANK YOU!

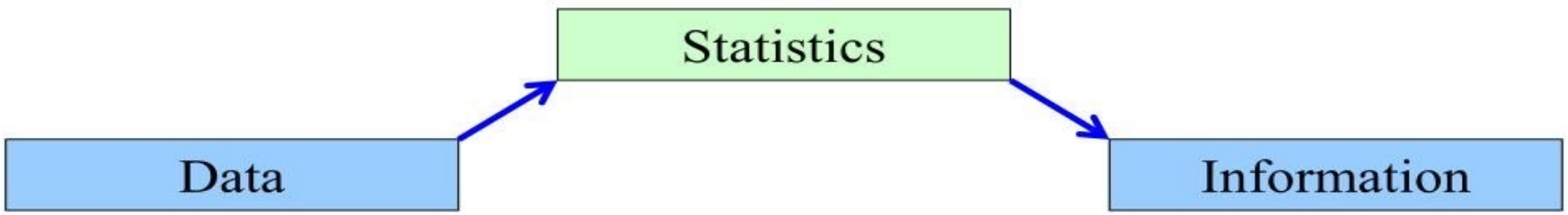
Learn Something Every Day

Basic of Statistics

Statistics

What is Statistics?

"Statistics is a way to get information from data"



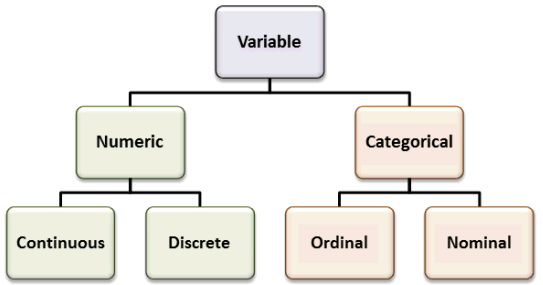
Data: Facts, especially numerical facts, collected together for reference or information.

Information: Knowledge communicated concerning some particular fact.

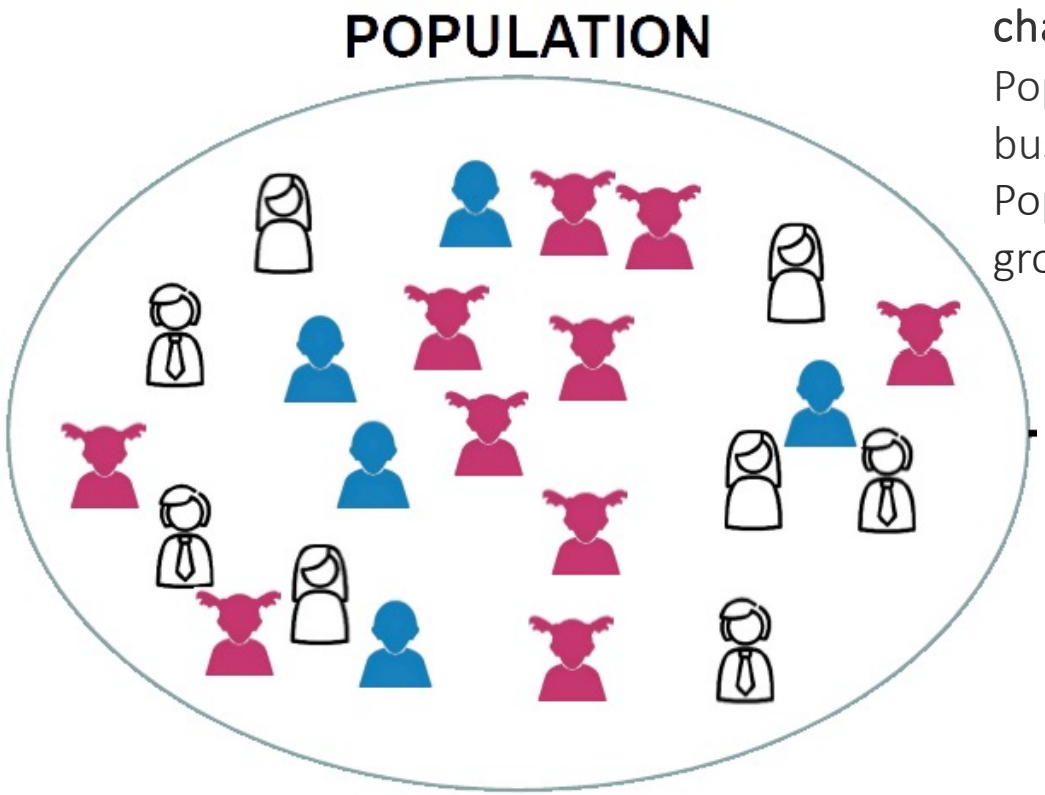
Statistics is a ***tool*** for creating ***new understanding*** from a set of numbers.

Basics of Statistics

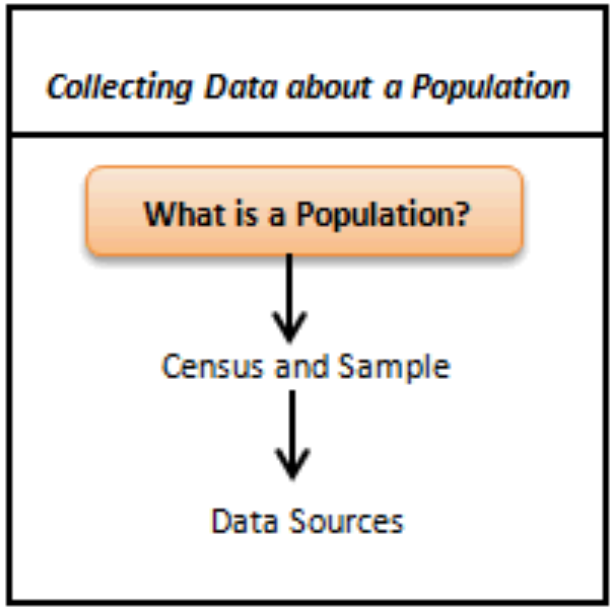
- Population
- Samples
- Variables
- Statistical Model



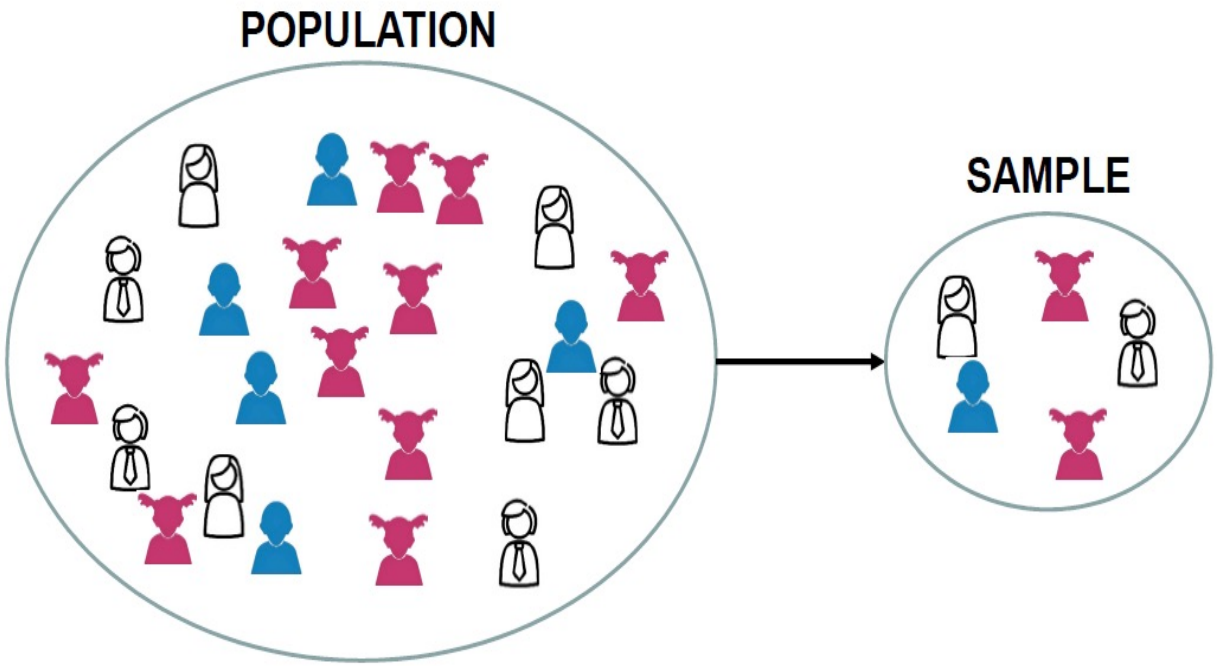
Population



A population is any complete group with at least one characteristic in common. Populations are not just people. Populations may consist of, but are not limited to, people, animals, businesses, buildings, motor vehicles, farms, objects or events. Population contains all the data points from a set of data. It is a group from where we collect the data.



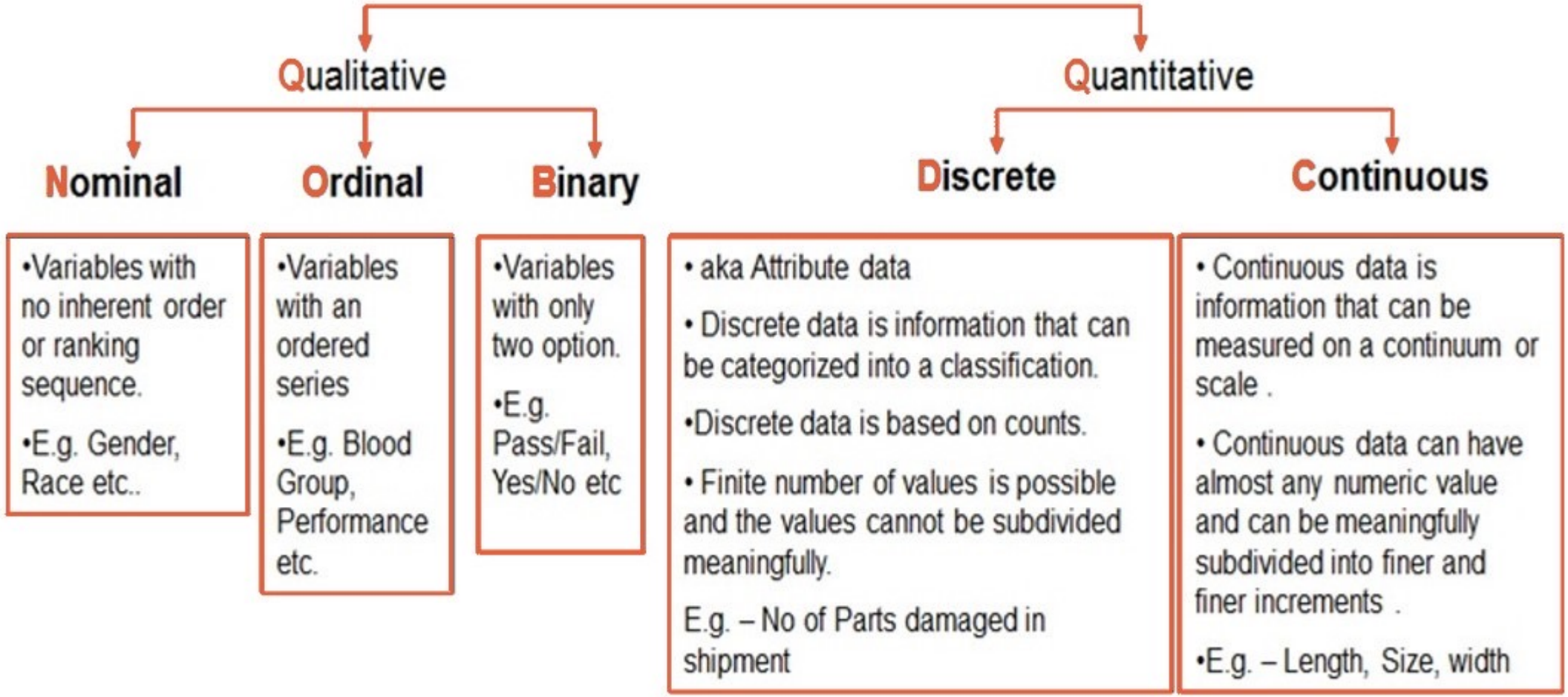
Sample



A sample consists of some observations selected from the population. The sample from the population should be selected such that it has all the characteristics that a population has. Population's measurable characteristics such as mean, standard deviation etc. are called as parameters while Sample's measurable characteristic is known as a statistic.

Cluster Sampling	**Random Sampling**
Stratified Sampling	**Systematical Sampling**

Variables in statistics

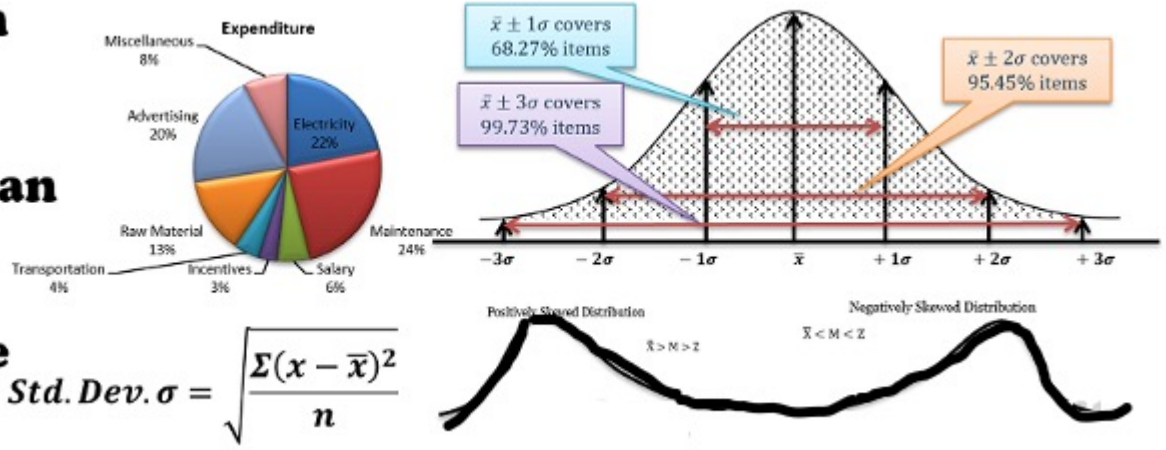


Statistical Model

Mean

Median

Mode



$$Std.Dev.\sigma = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}$$

Statistical modelling is a method of mathematically approximating the world. Statistical models contain variables that can be used to explain relationships between other variables. We use hypothesis testing, confidence intervals etc to make inferences and validate our hypothesis.

A statistical model will have sampling, probability spaces, assumptions and diagnostics etc, to make inferences.

We use statistical models to find insights given a particular set of data. We can conduct modelling on a relatively small set of data just to try and understand the underlying nature of the data.

Different Types of Analysis

Different Types of Analysis

Qualitative Analysis

Quantitative Analysis

Different Types of Analysis

Qualitative Data
(Categorical)

- Gender
- Religion
- Marital status
- Native language
- Social class
- Qualifications
- Type of instruction
- Method of treatment
- Type of teaching approach
- Problem-solving strategy used

Quantitative Data
(Numerical)

- Age
- Height
- Weight
- Income
- University size
- Group size
- Self-efficacy test score
- Percent of lecture attended
- Clinical skills performed
- Number of errors

Qualitative Analysis

Content Analysis	•It is a method for subjective interpretation of content of text through the systematic classification process of coding and identifying themes or patterns
Narrative Analysis	•These approaches typically focus on the lives of individuals as told through their own stories.
Discourse Analysis	•It involves analyzing a naturally occurring language use and types of written texts.
Grounded Theory	•is an approach for theory construction through the analysis of data. it is usually inductive in nature.
Thematic Analysis	•It is a method for identifying and analyzing patterns (themes) in the data by means of thematic codes.

Quantitative Analysis

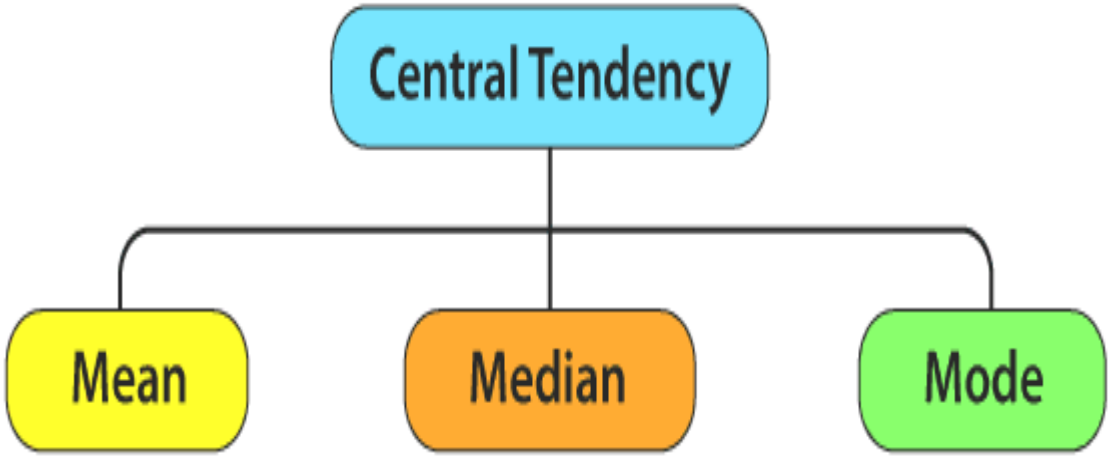
Survey Research	Survey methodology studies the in-depth sampling of individual units from a population and administering data collection techniques on that sample.
Correlation Research	Correlational research is a type of nonexperimental research in which the researcher measures two variables and assesses the statistical relationship between them with little or no effort to control extraneous variables.
Casual-Comparitive Research	In causal-comparative research, the researcher investigates the effect of an independent variable on a dependent variable by comparing two or more groups of individuals.
Experimental Research	Experimental research is research conducted with a scientific approach using two sets of variables. The first set acts as a constant, which you use to measure the differences of the second set.

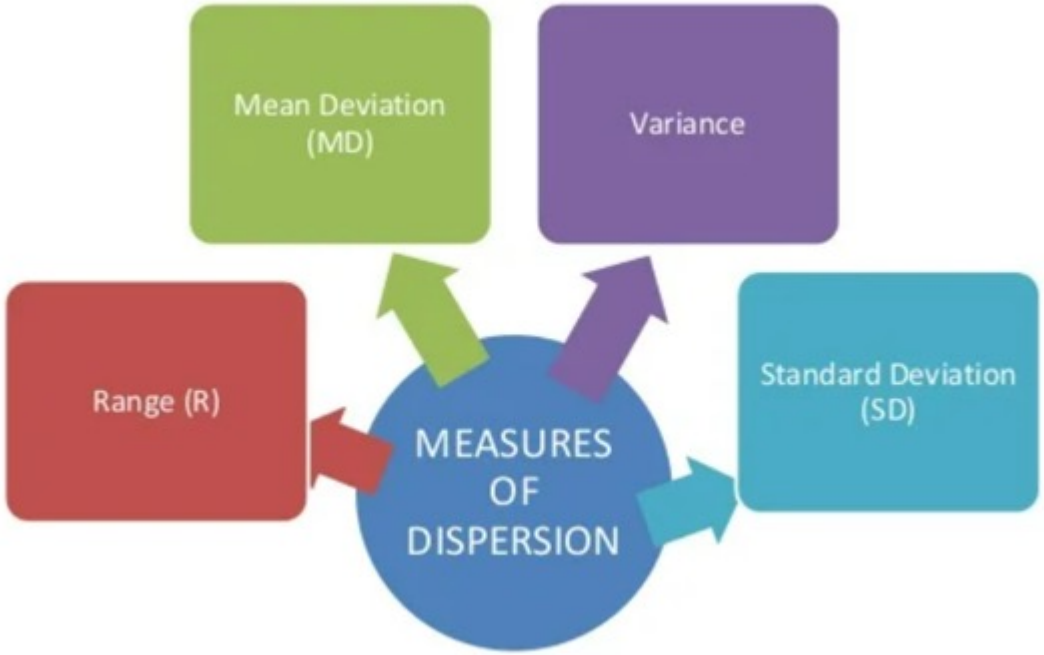
Types of in Statistic Analysis

Statistic Analysis

Descriptive Statistics		Inferential Statistics	
Measures of Central Tendency	Measures of Dispersion	Hypothesis Testing	Regression Analysis
Mean Median Mode	Range Standard Deviation Variance Absolute Deviation	Z test F test T test	Linear Regression

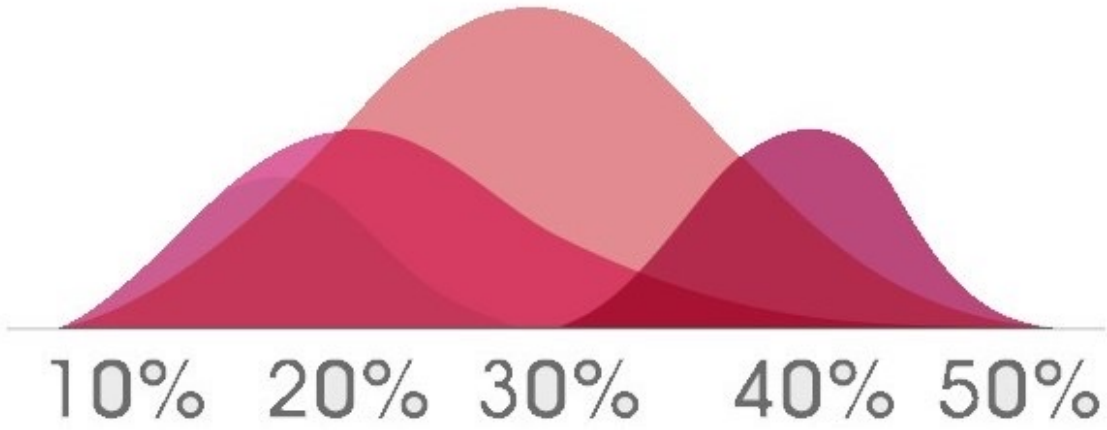
Measures of central tendency and Dispersion

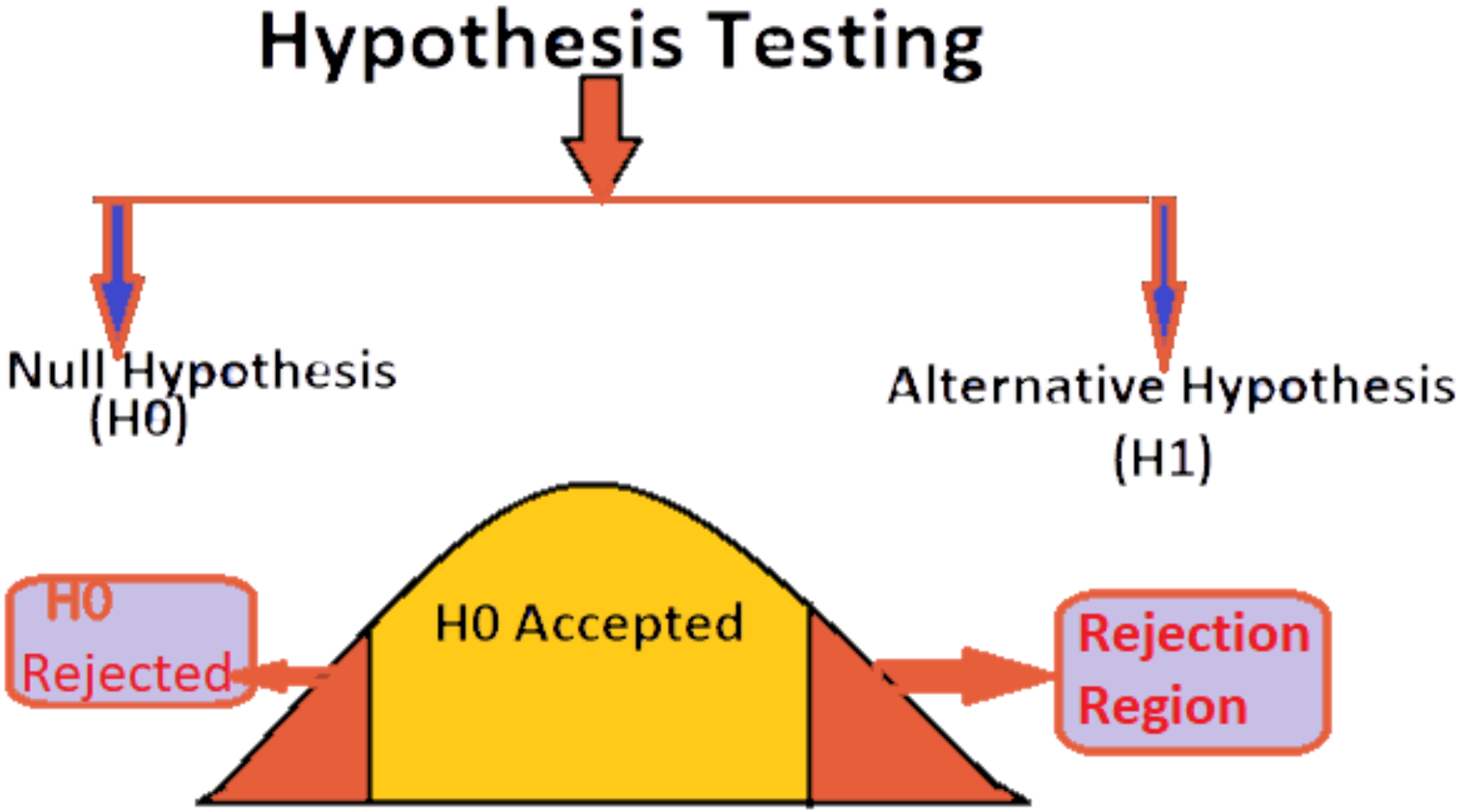




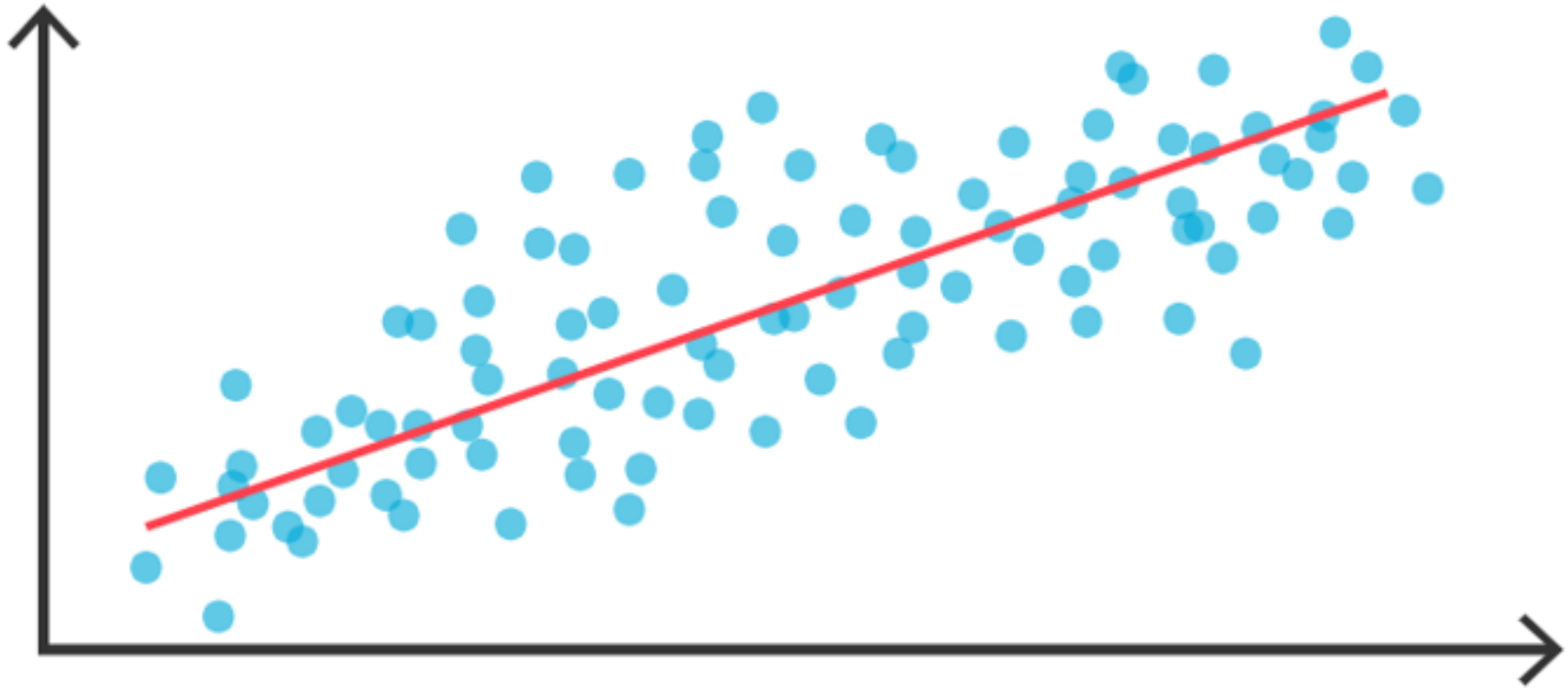
Inferential statistics

studies a sample
of the same data.

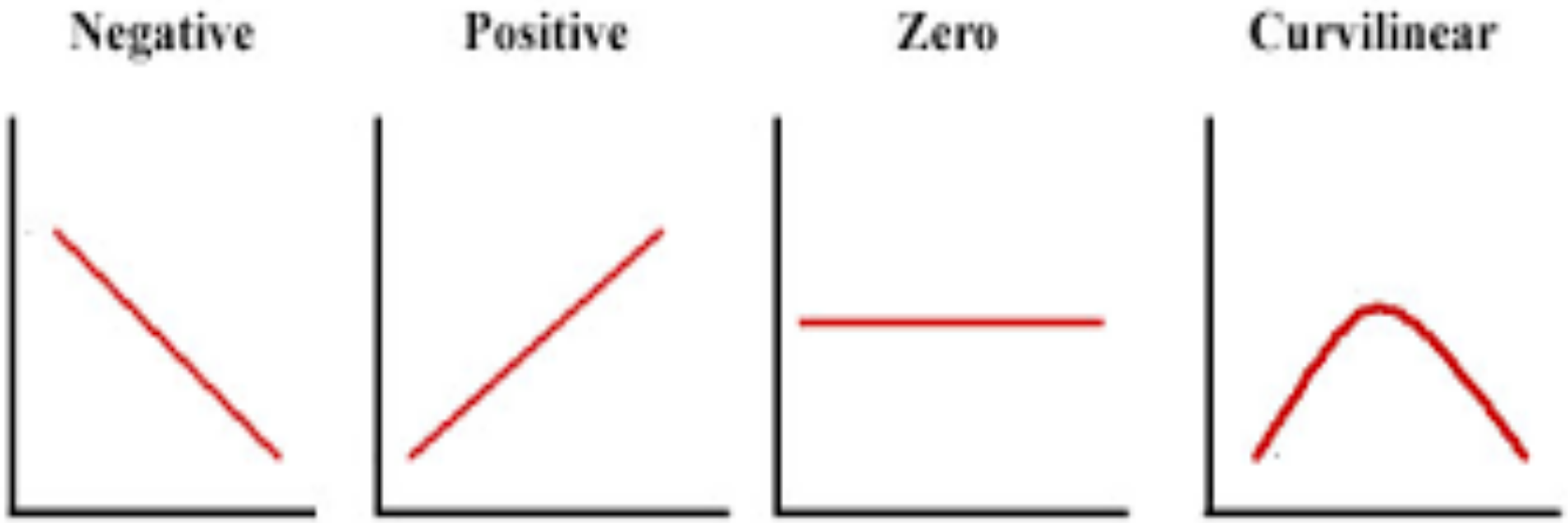




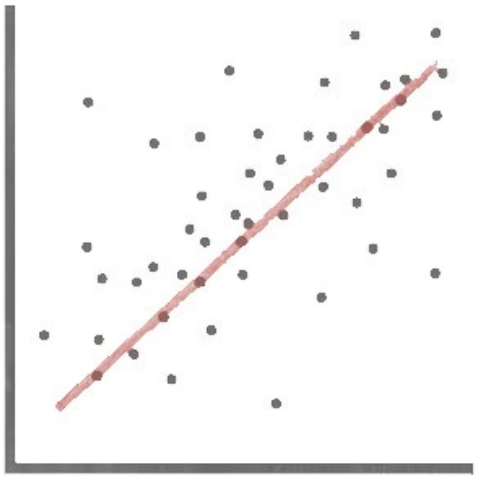
Regression Analysis



Relationship between variables

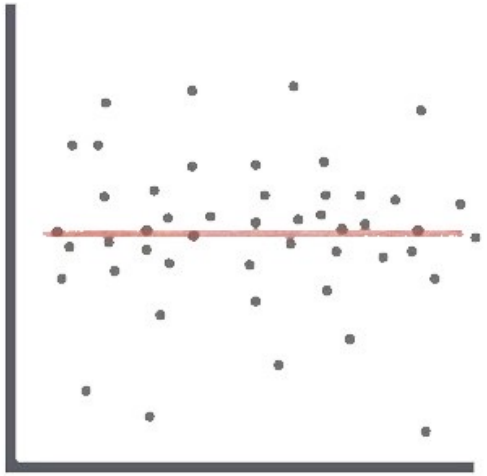


Relationship between variables



Positive Correlation

Negative Correlation

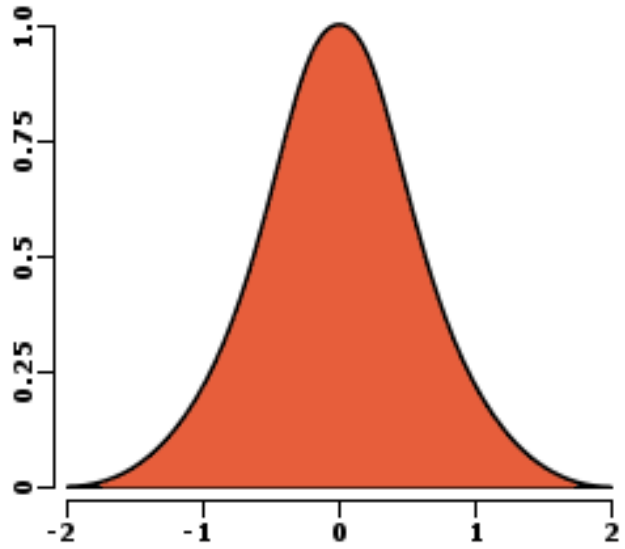


No Correlation

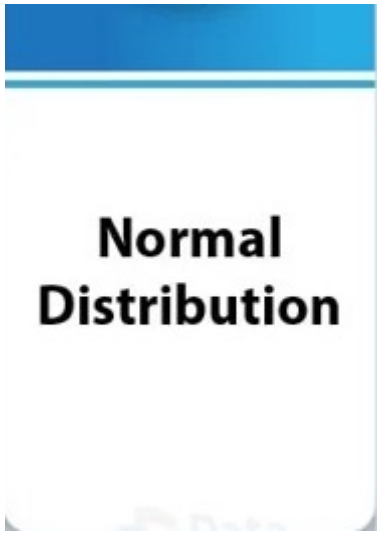
Probability Distribution

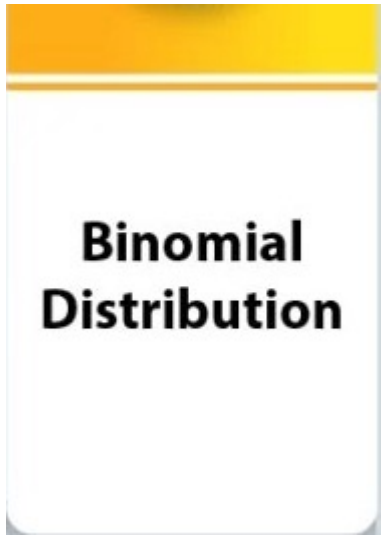
Probability Distribution

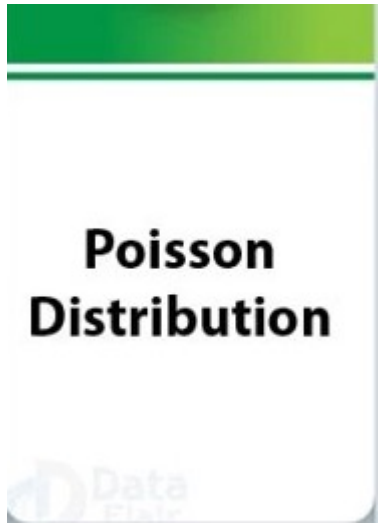
Distribution function is a mathematical expression that describes the probability that a system will take on a specific value or set of values.



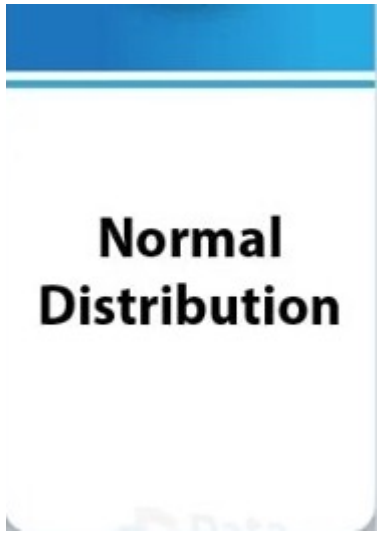
Probability Distribution







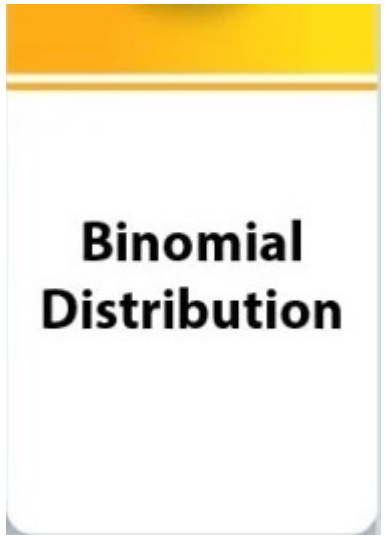
Probability Distribution



Normal Distribution:

The normal distribution is a symmetric probability distribution centered on the mean, indicating that data around the mean occur more frequently than data far from it. the normal distribution is also called Gaussian distribution. The normal distribution curve resembles a bell curve.

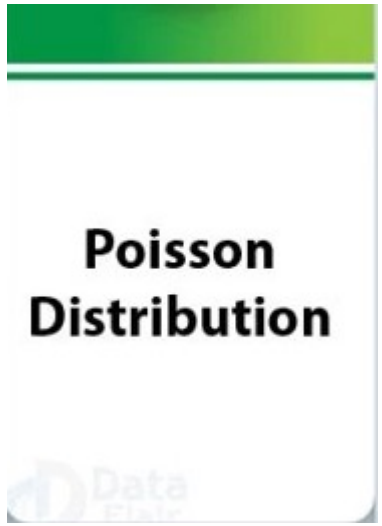
Probability Distribution



Binomial Distribution

Under a given set of factors or assumptions, the binomial distribution expresses the likelihood that a variable will take one of two outcomes or independent values.

Probability Distribution

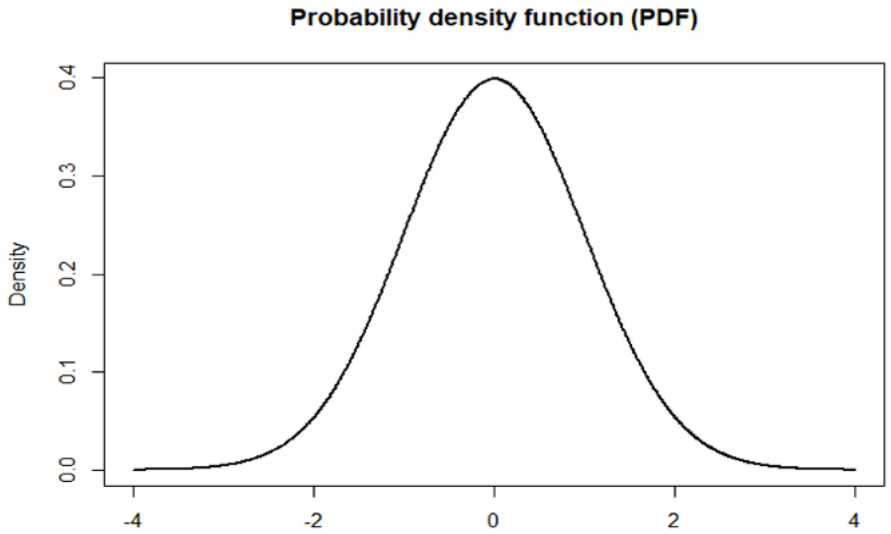


Poisson Distribution:

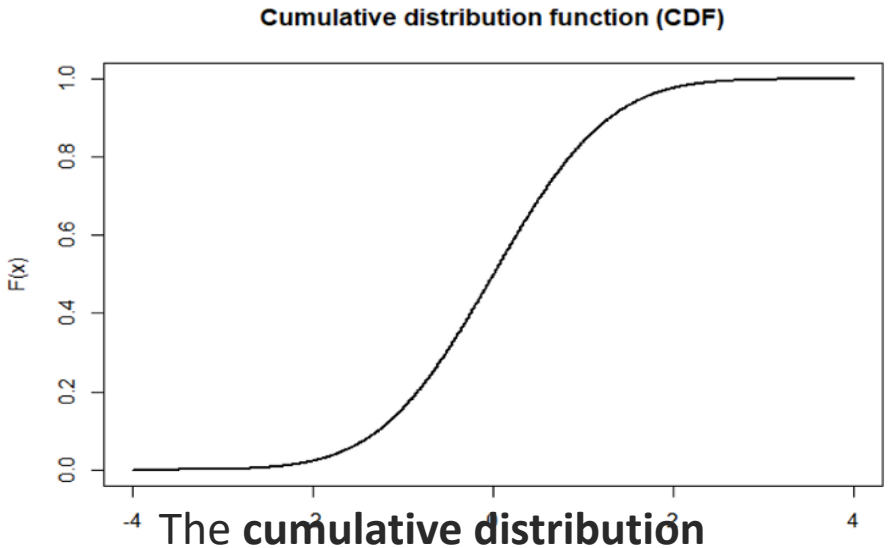
A Poisson distribution is a kind of probability distribution used in statistics to illustrate how many times an event is expected to happen over a certain amount of time. It's also called count distribution.

PDF & CDF

Probability density Function
&
Cumulative Distribution Function



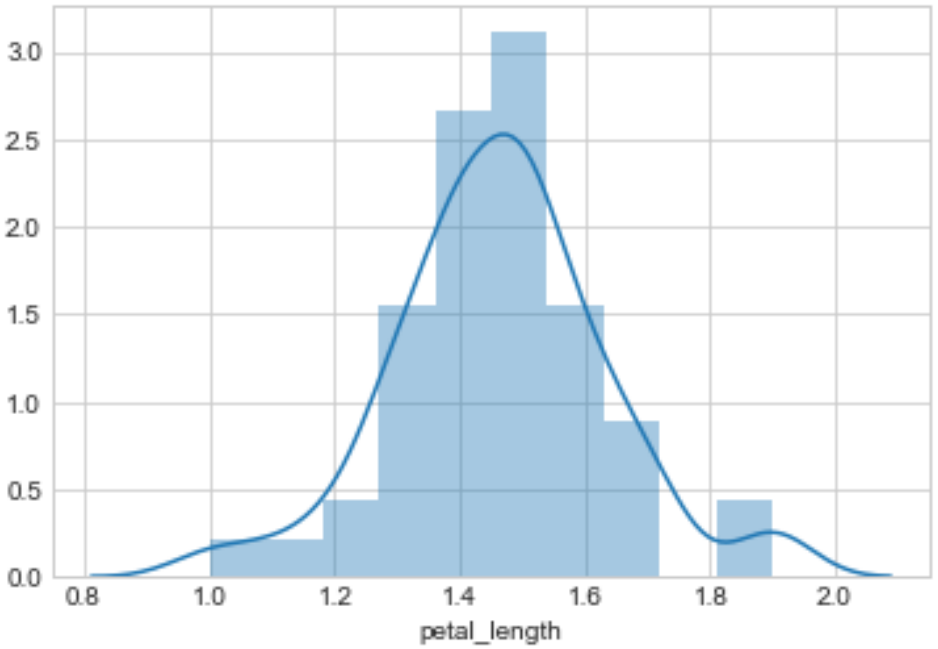
*PDF is a statistical term that describes the probability distribution of the **continues** random variable*



The **cumulative distribution function** is applicable for describing the distribution of random variables either it is continuous or discrete

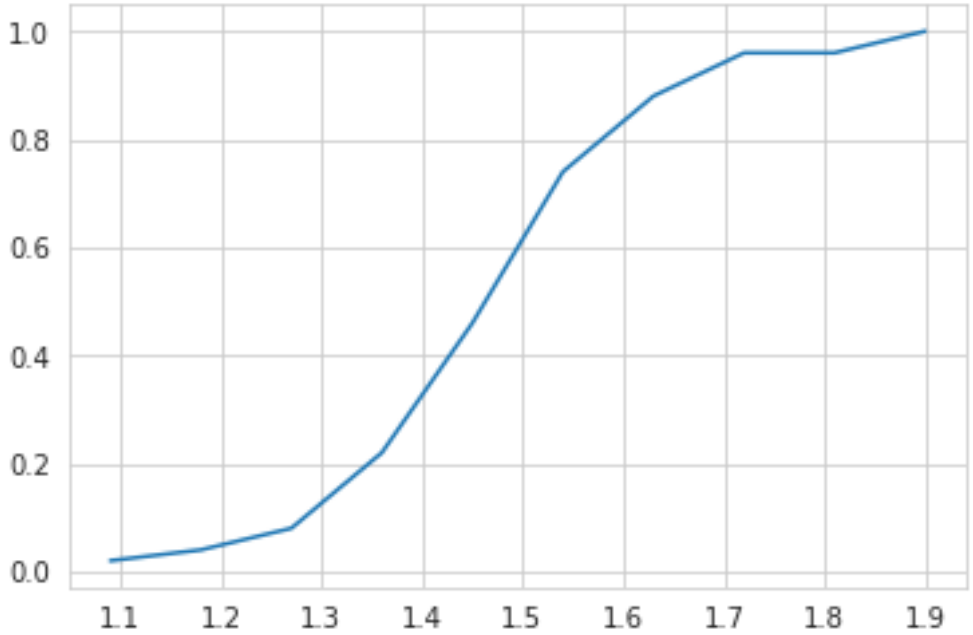
PDF & CDF

PDF On Iris



PDF for [‘species’]== ‘setosa’ on petal length

CDF on Iris

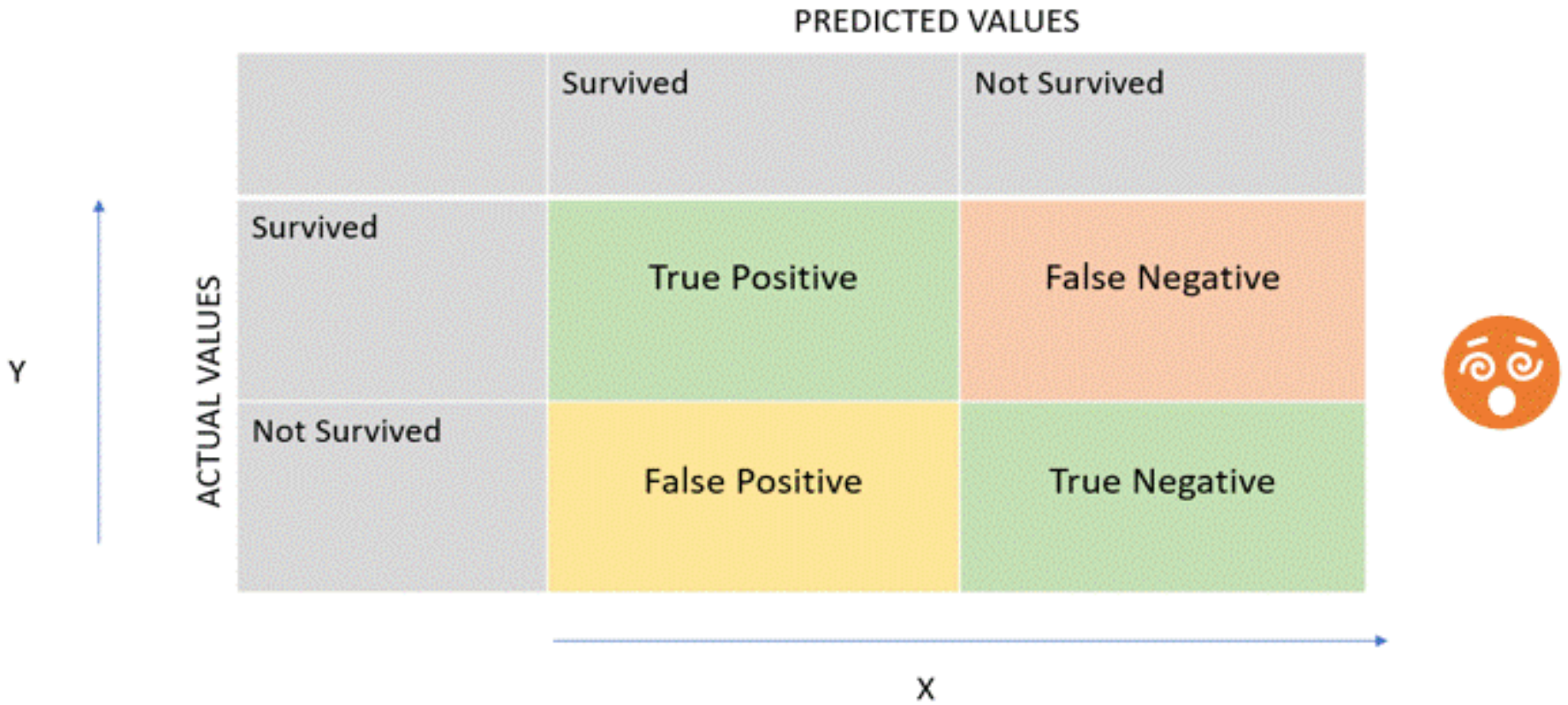


CDf of iris_setosa using petal length

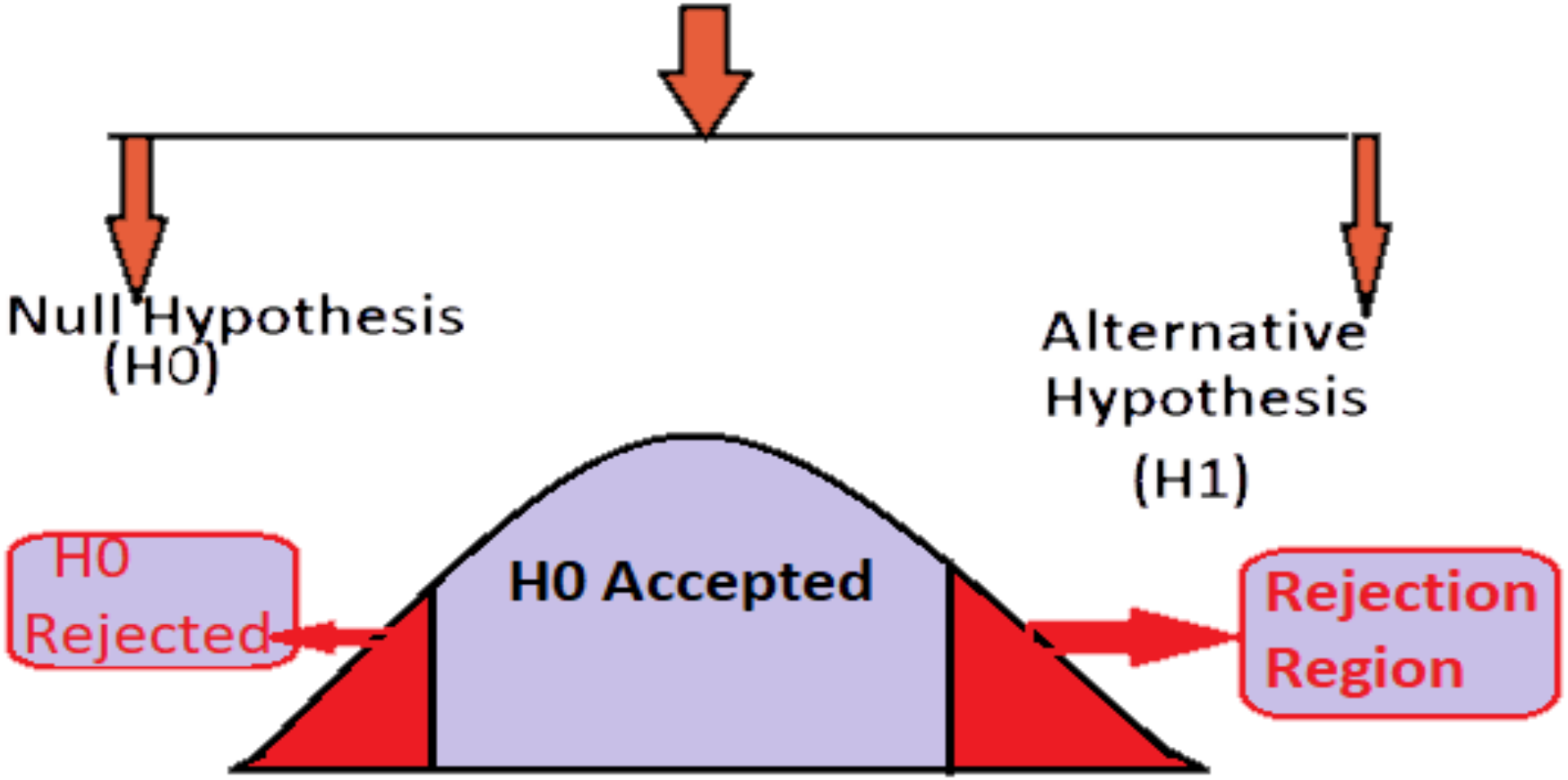
Accuracy

		Actual	
		Positive	Negative
Predicted	Positive	**True Positive**	**False Positive**
	Negative	**False Negative**	**True Negative**

CONFUSING!



Hypothesis Testing



Hypothesis Testing

Null Hypothesis

H_0

A statement about a population parameter.

We test the likelihood of this statement being true in order to decide whether to accept or reject our alternative hypothesis.

Can include =, ≤, or ≥ sign.

Alternative Hypothesis

H_a

A statement that directly contradicts the null hypothesis.

We determine whether or not to accept or reject this statement based on the likelihood of the null (opposite) hypothesis being true.

Can include a ≠, >, or < sign.

Hypothesis Testing

The two types of hypothesis testing are null hypothesis and alternate hypothesis.

Null hypothesis is the initial assumption about an event (also referred to as the ground truth).

Alternate hypothesis is an assumption that counters the initial assumption.

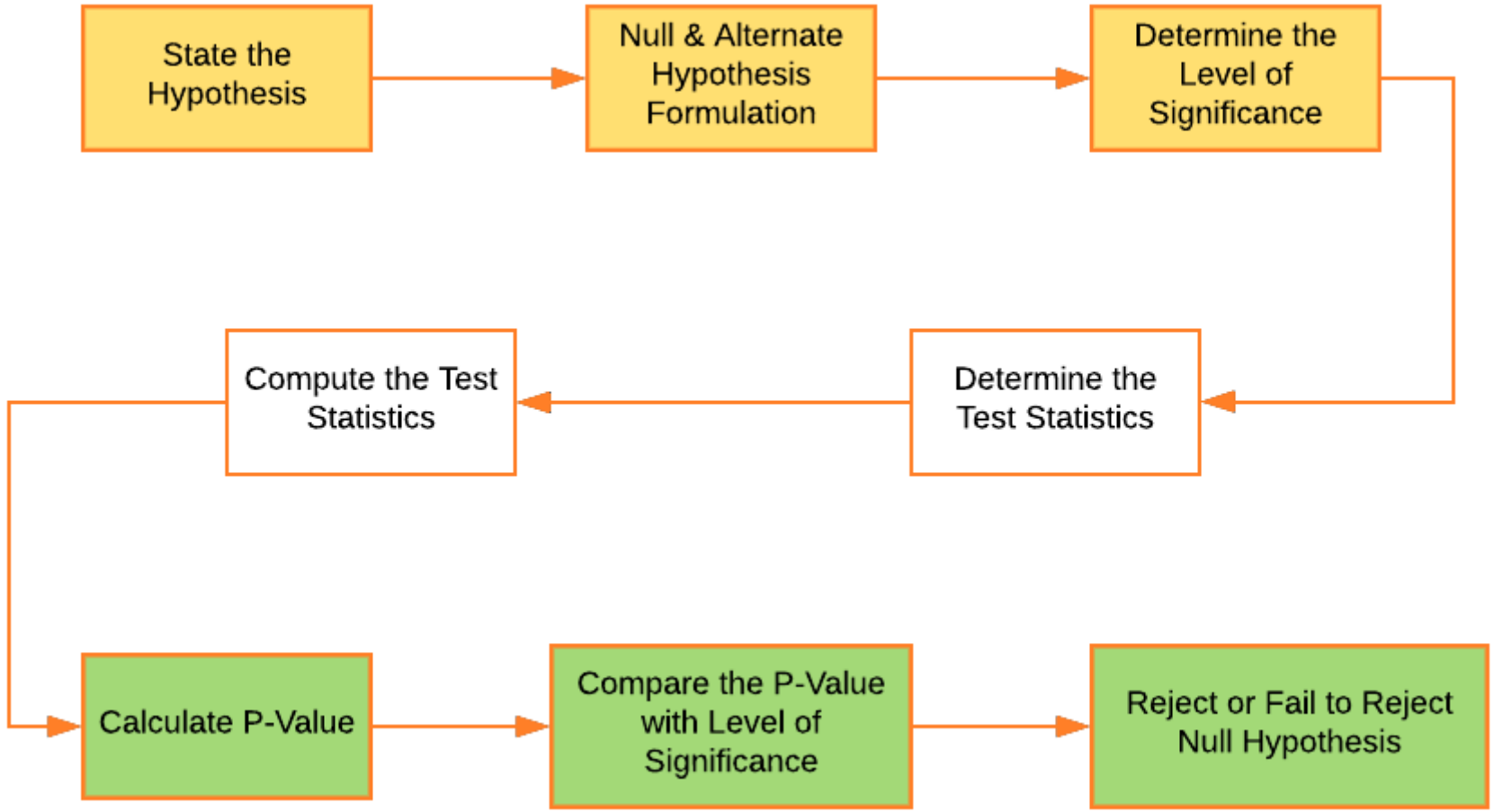
Example of Hypothesis Testing With T test

A ***t*-test** is a statistical test that is used to compare the means of two groups. It is often used in hypothesis testing to determine whether a process or treatment actually has an effect on the population of interest, or whether two groups are different from one another.

- The null hypothesis (H_0) is that the true difference between these group means is zero.
- The alternate hypothesis (H_a) is that the true difference is different from zero.

- A p-value is a statistical measurement used to validate a hypothesis against observed data.
- A p-value measures the probability of obtaining the observed results, assuming that the null hypothesis is true.
- The lower the p-value, the greater the statistical significance of the observed difference.
- A p-value of 0.05 or lower is generally considered statistically significant.
- P-value can serve as an alternative to or in addition to preselected confidence levels for hypothesis testing.

Hypothesis Testing Workflow



THANK YOU!

Learn Something Every Day

Data Preparation

The data collected from the respondents is generally not in the form to be analyzed directly. After the responses are recorded or received, the next stage is that of **preparation of data** i.e. to make the data amenable for appropriate analysis.

Data preparation **includes editing, coding, and data entry** and is the activity that **ensures the accuracy of the data** and their **conversion from raw form** to reduced and classified forms that are more appropriate for analysis. Preparing a descriptive statistical summary is another preliminary step leading to an understanding of the collected data

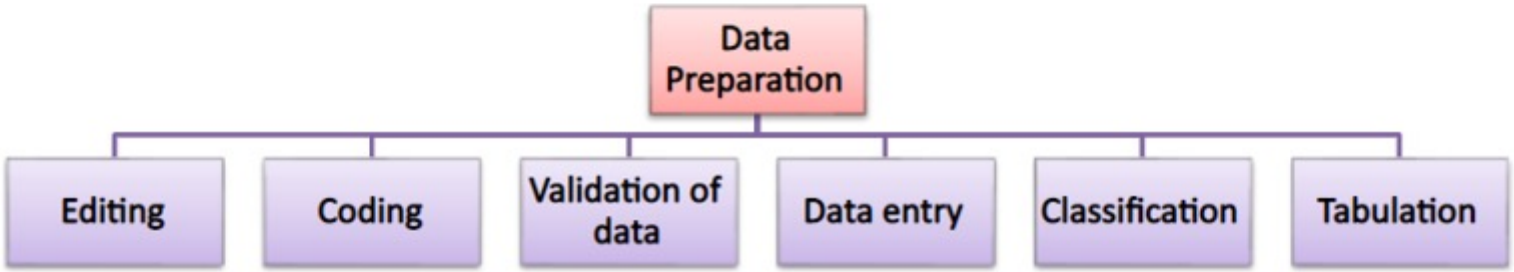
Why Data Preprocessing?

Data in the real world is dirty

- incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
- noisy: containing errors or outliers
- inconsistent: containing discrepancies in codes or names

No quality data, no quality mining results!

- Quality decisions must be based on quality data
- Data warehouse needs consistent integration of quality data



Major Tasks in Data Preprocessing

Data cleaning

Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies

Data integration

Integration of multiple databases, data cubes, or files

Data transformation

Normalization and aggregation

Data reduction

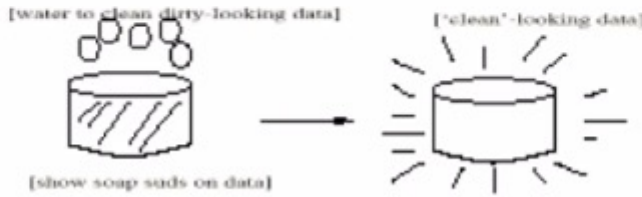
Obtains reduced representation in volume but produces the same or similar analytical results

Data discretization

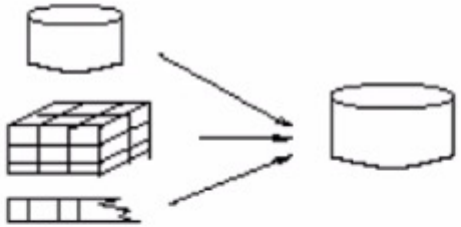
Part of data reduction but with particular importance, especially for numerical data

Forms of data preprocessing

Data Cleaning



Data Integration



Data Transformation

-2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data Reduction

Data Cleaning

Data cleaning tasks

- Fill in missing values
- Identify outliers and smooth out noisy data
- Correct inconsistent data

Missing Data

Data is not always available

- E.g., many tuples have no recorded value for several attributes, such as customer income in sales data

Missing data may be due to

- equipment malfunction
- inconsistent with other recorded data and thus deleted
- data not entered due to misunderstanding
- certain data may not be considered important at the time of entry
- not register history or changes of the data

Missing data may need to be inferred.

Missing Data

Data is not always available

- E.g., many tuples have no recorded value for several attributes, such as customer income in sales data

Missing data may be due to

- equipment malfunction
- inconsistent with other recorded data and thus deleted
- data not entered due to misunderstanding
- certain data may not be considered important at the time of entry
- not register history or changes of the data

Missing data may need to be inferred.

How to Handle Missing Data?

Ignore the tuple: usually done when class label is missing (assuming the tasks in classification—not effective when the percentage of missing values per attribute varies considerably)

Fill in the missing value manually: tedious + infeasible?

Use a global constant to fill in the missing value: e.g., “unknown”, a new class?!

Use the attribute mean to fill in the missing value

Use the most probable value to fill in the missing value: inference-based such as Bayesian formula or decision tree

Noisy Data

Noise: random error or variance in a measured variable

Incorrect attribute values may due to

- faulty data collection instruments
- data entry problems
- data transmission problems
- technology limitation
- inconsistency in naming convention

Other data problems which requires data cleaning

- duplicate records
- incomplete data
- inconsistent data

How to Handle Noisy Data?

Binning method:

- first sort data and partition into (equi-depth) bins
- then smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.

Clustering

- detect and remove outliers

Combined computer and human inspection

- detect suspicious values and check by human

Regression

- smooth by fitting the data into regression functions

Simple Discretization Methods: Binning

Equal-width (distance) partitioning:

- It divides the range into *N* intervals of equal size: uniform grid
- if *A* and *B* are the lowest and highest values of the attribute, the width of intervals will be: $W = (B-A)/N$.
- The most straightforward
- But outliers may dominate presentation
- Skewed data is not handled well.

Equal-depth (frequency) partitioning:

- It divides the range into *N* intervals, each containing approximately same number of samples
- Good data scaling
- Managing categorical attributes can be tricky.

Binning Methods for Data Smoothing

* Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
* Partition into (equi-depth) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
* Smoothing by bin means:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
* Smoothing by bin boundaries:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

Data Integration

Data integration:

- combines data from multiple sources into a coherent store

Schema integration

- integrate metadata from different sources
- Entity identification problem: identify real world entities from multiple data sources, e.g., A.cust-id ≡ B.cust-#

Detecting and resolving data value conflicts

- for the same real world entity, attribute values from different sources are different
- possible reasons: different representations, different scales, e.g., metric vs. British units

Handling Redundant Data

Redundant data occur often when integration of multiple databases

- The same attribute may have different names in different databasesCareful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Data Transformation

Smoothing: remove noise from data

Aggregation: summarization, data cube construction

Generalization: concept hierarchy climbing

Normalization: scaled to fall within a small, specified range

- min-max normalization
- z-score normalization
- normalization by decimal scaling

Data Reduction Strategies

Warehouse may store terabytes of data: Complex data analysis/mining may take a very long time to run on the complete data set

Data reduction

- Obtains a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results

Data reduction strategies

- Data cube aggregation
- Dimensionality reduction
- Numerosity reduction
- Discretization and concept hierarchy generation

Data Cube Aggregation

The lowest level of a data cube

- the aggregated data for an individual entity of interest
- e.g., a customer in a phone calling data warehouse.

Multiple levels of aggregation in data cubes

- Further reduce the size of data to deal with

Reference appropriate levels

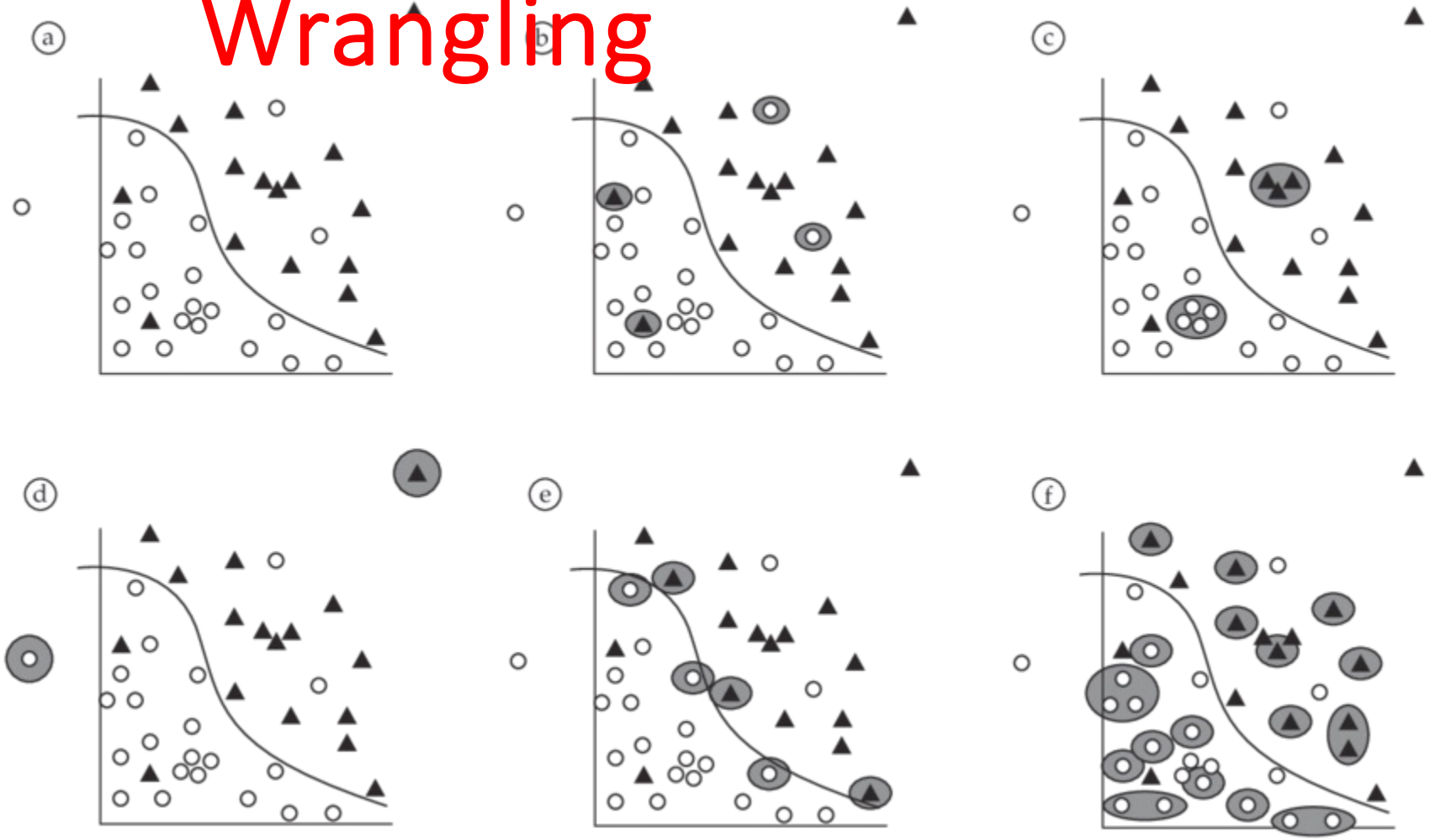
- Use the smallest representation which is enough to solve the task

Data Wrangling

- What Is Data Wrangling?
- Importance of Data Wrangling
- Benefits of Data Wrangling
- Data Wrangling Tools
- Data Wrangling Examples

Data Wrangling is the process of removing errors and combining complex data sets to make them more accessible and easier to analyze.

Need of Data Wrangling

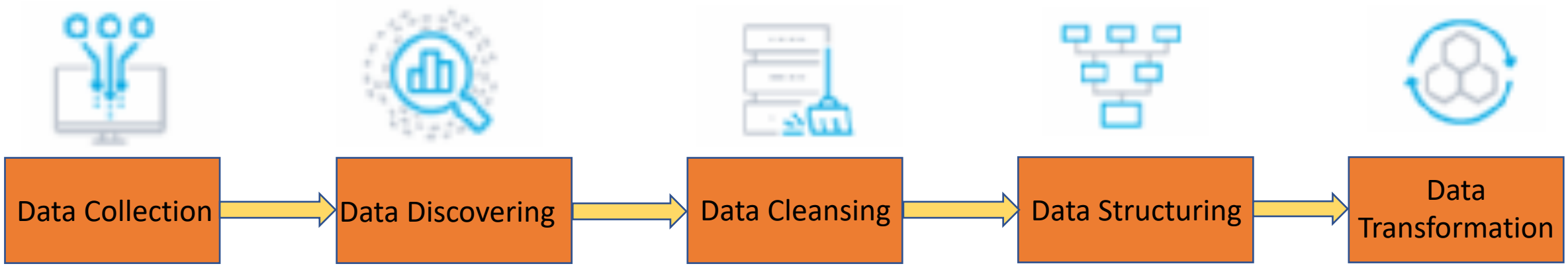


Importance of Data Wrangling

- Making raw data usable. Accurately wrangled data guarantees that quality data is entered into the downstream analysis.
- Getting all data from various sources into a centralized location so it can be used.
- Piecing together raw data according to the required format and understanding the business context of data
- Automated data integration tools are used as data wrangling techniques that clean and convert source data into a standard format that can be used repeatedly according to end requirements. Businesses use this standardized data to perform crucial, cross-data set analytics.
- Cleansing the data from the noise or flawed, missing elements
- Data wrangling acts as a preparation stage for the data mining process, which involves gathering data and making sense of it.
- Helping business users make concrete, timely decisions

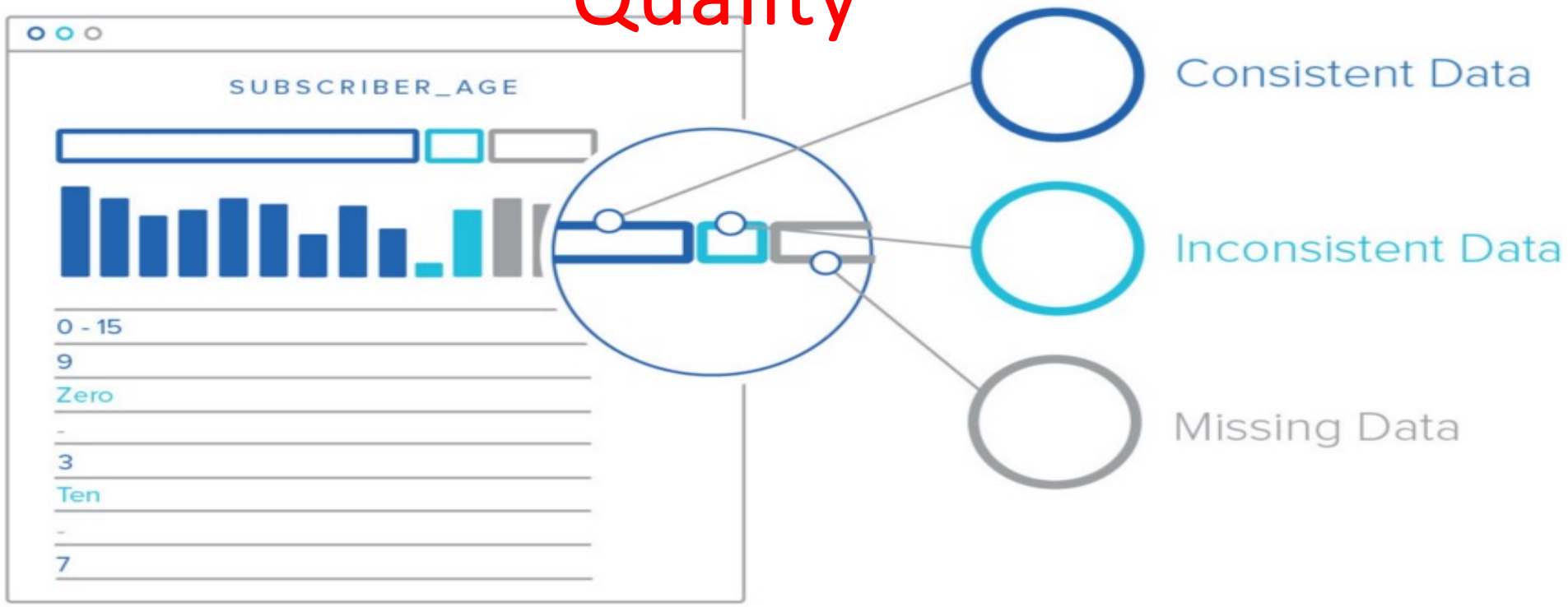
Data Preparation

Data preparation is the process of preparing raw data so that it is suitable for further processing and analysis.



DATA COLLECTION

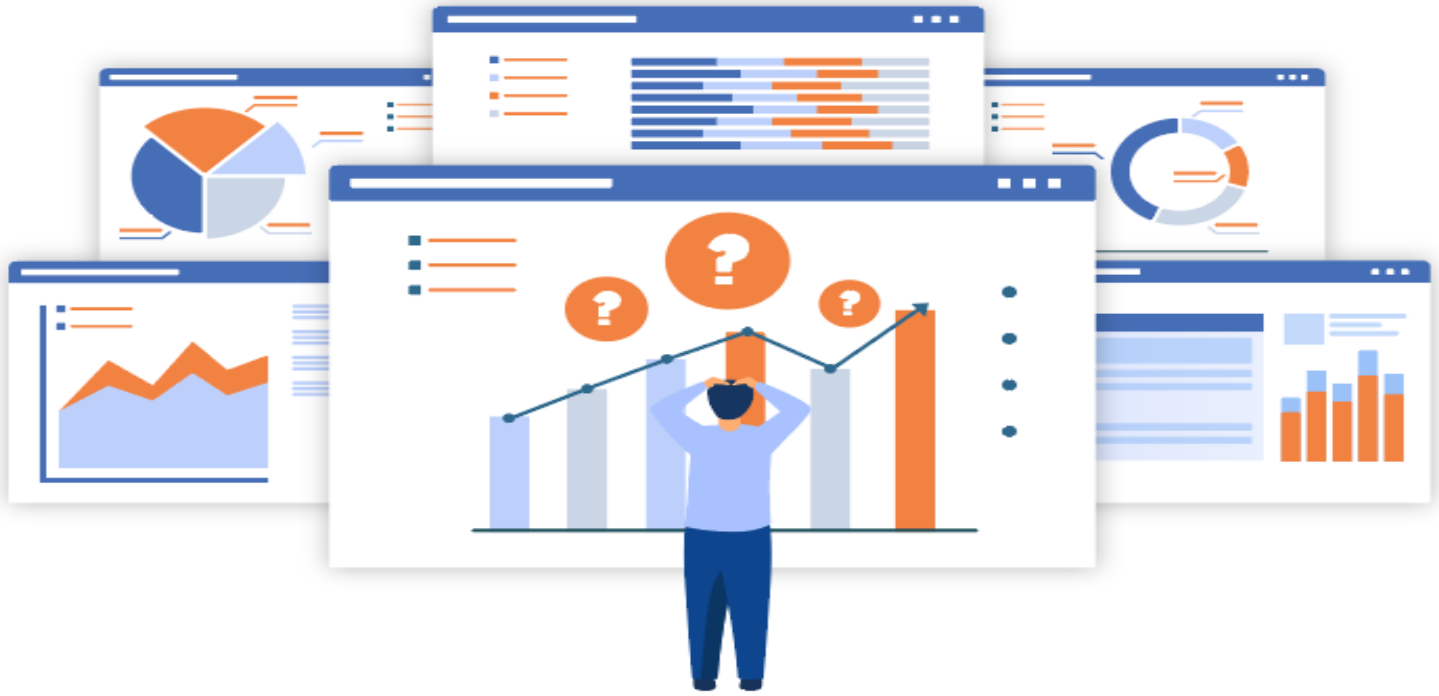
Data Quality



Data Quality includes examining data accuracy, consistency, completeness, and relevance.

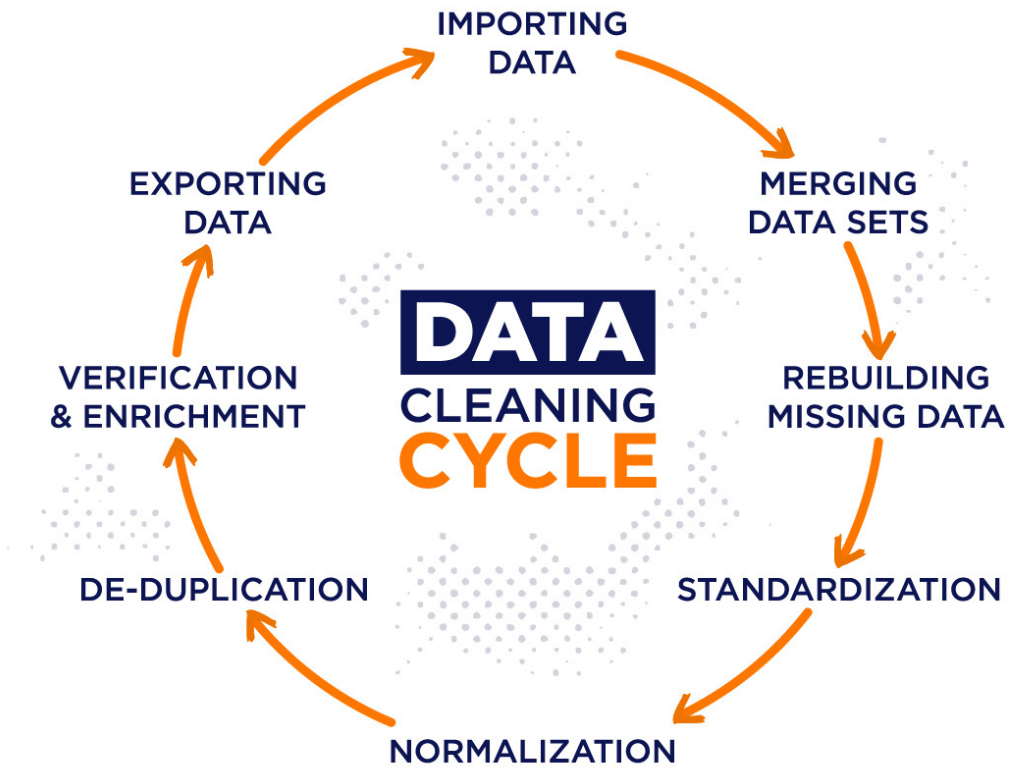
Data Discovery

The second step in the Data Wrangling process is **Discovery.** This is an all-encompassing term for understanding or getting familiar with your data.



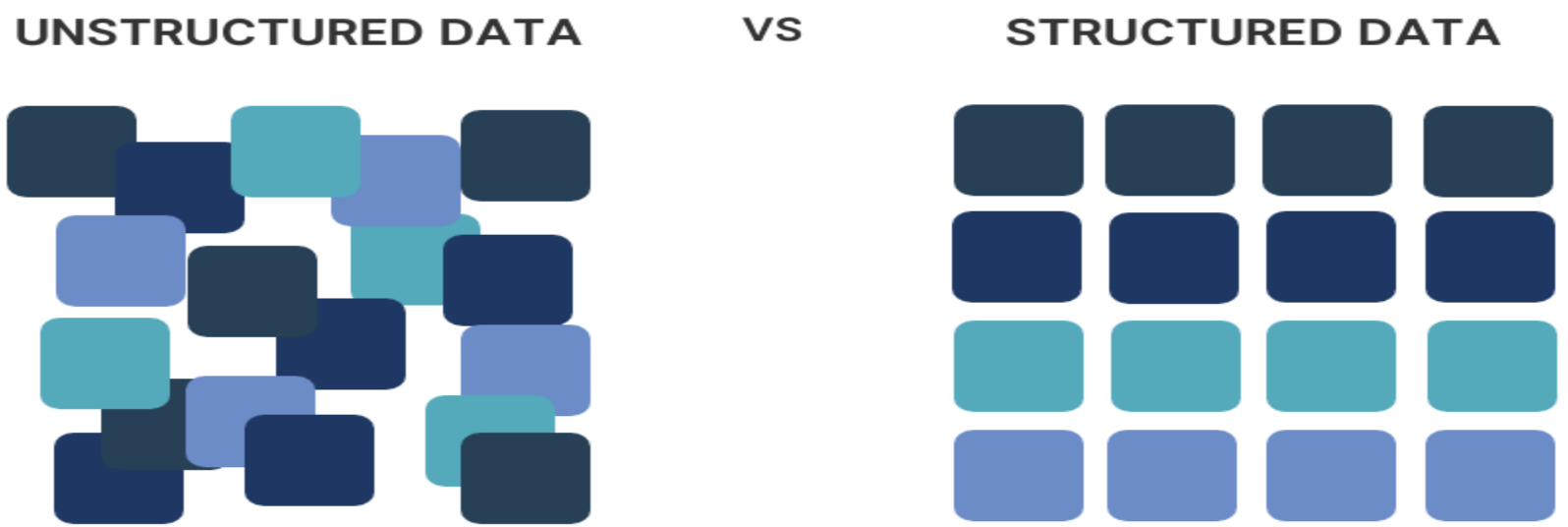
Data Cleaning

Data Cleaning involves Tackling Outliers, Making Corrections, Deleting Bad Data completely, etc. This is done by applying algorithms to tidy up and sanitize the dataset.



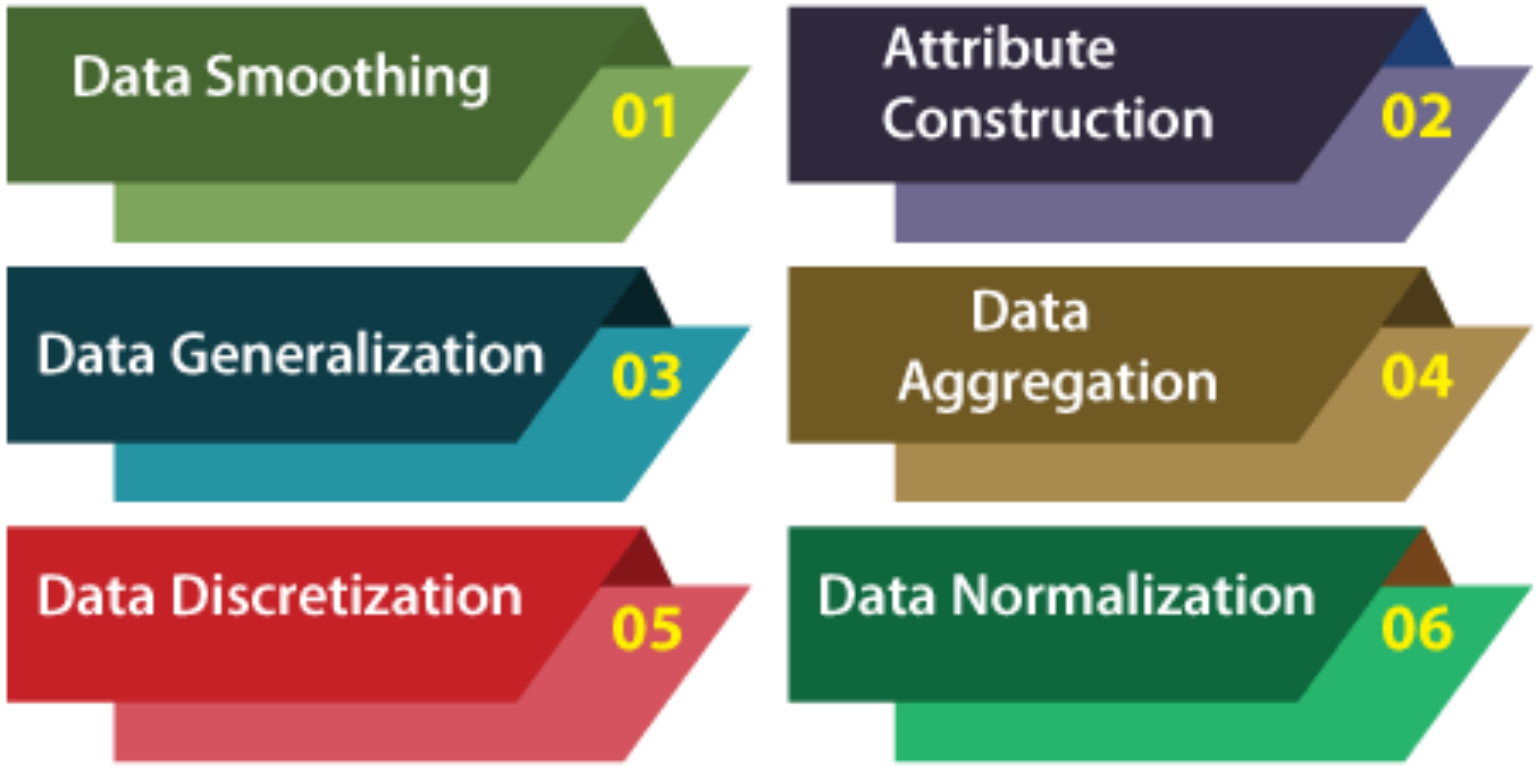
Data Structuring

When raw data is collected, it's in a wide range of formats and sizes. It has no definite structure, which means that it lacks an existing model and is completely disorganized. It needs to be restructured to fit in with the **Analytical Model** deployed by your business, and giving it a structure allows for better analysis.



Structured data	Unstructured data
Structured data stands for information that is highly organized, factual, and to-the-point.	Unstructured data doesn't have any predefined structure to it and comes in all its diversity of forms.
Quantitative	Qualitative
Data warehouses Relational databases	Data lakes Non-relational databases
Several predetermined formats	A huge array of formats

Data Transformation



THANK YOU!

Learn Something Every Day

DATA VISUALIZATION?

What is Data Visualization

- Visual display of quantitative information
- Mapping data to visual elements
- Encoding data with size, shape, color...
- Storytelling / narrative elements

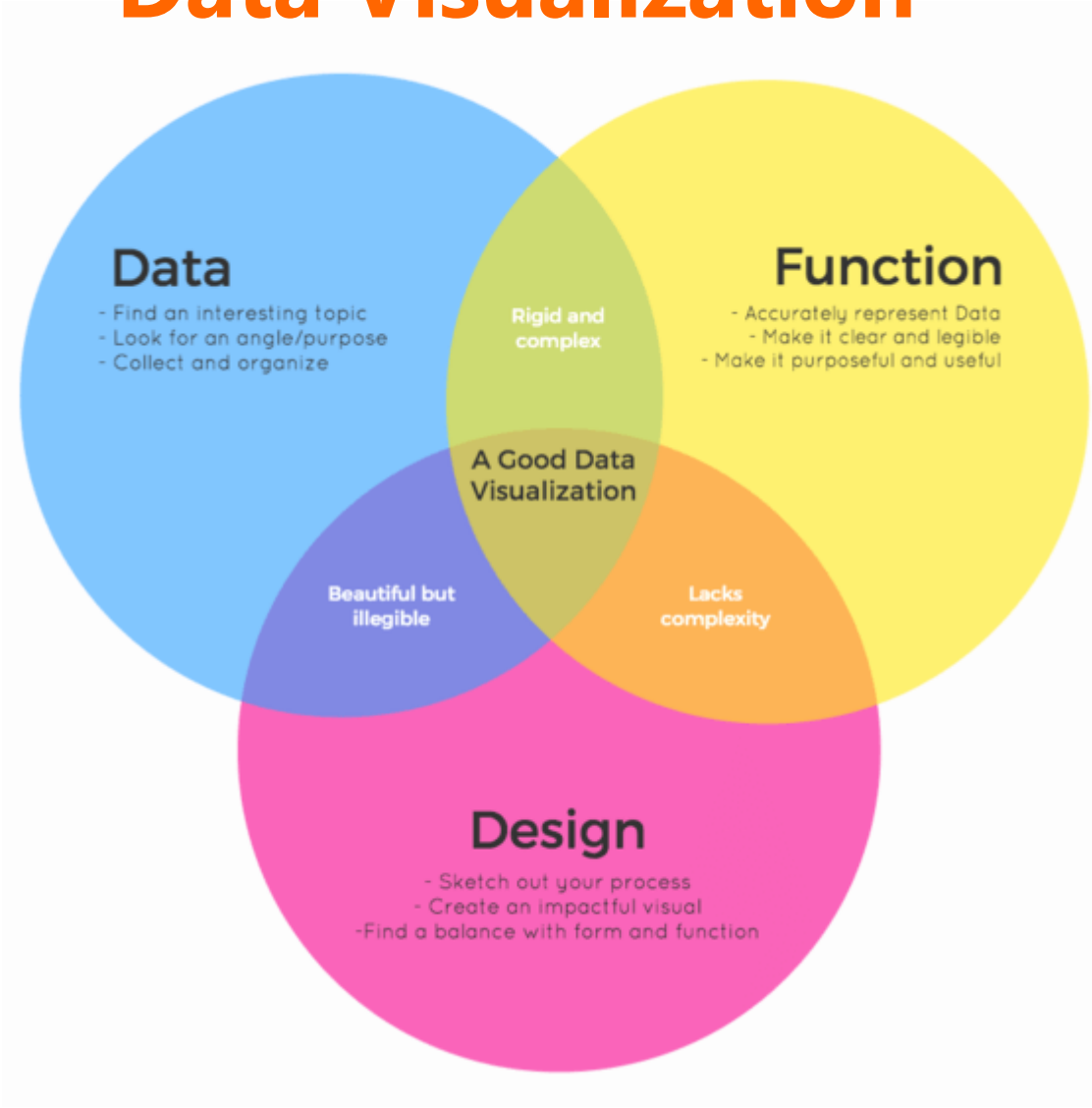
What is Data Visualization

- **Data visualization** is the visual presentation of data or information.
- The goal of data visualization is to communicate data or information clearly and effectively to readers.
- Typically, data is visualized in the form of a chart, infographic, diagram, map and more.

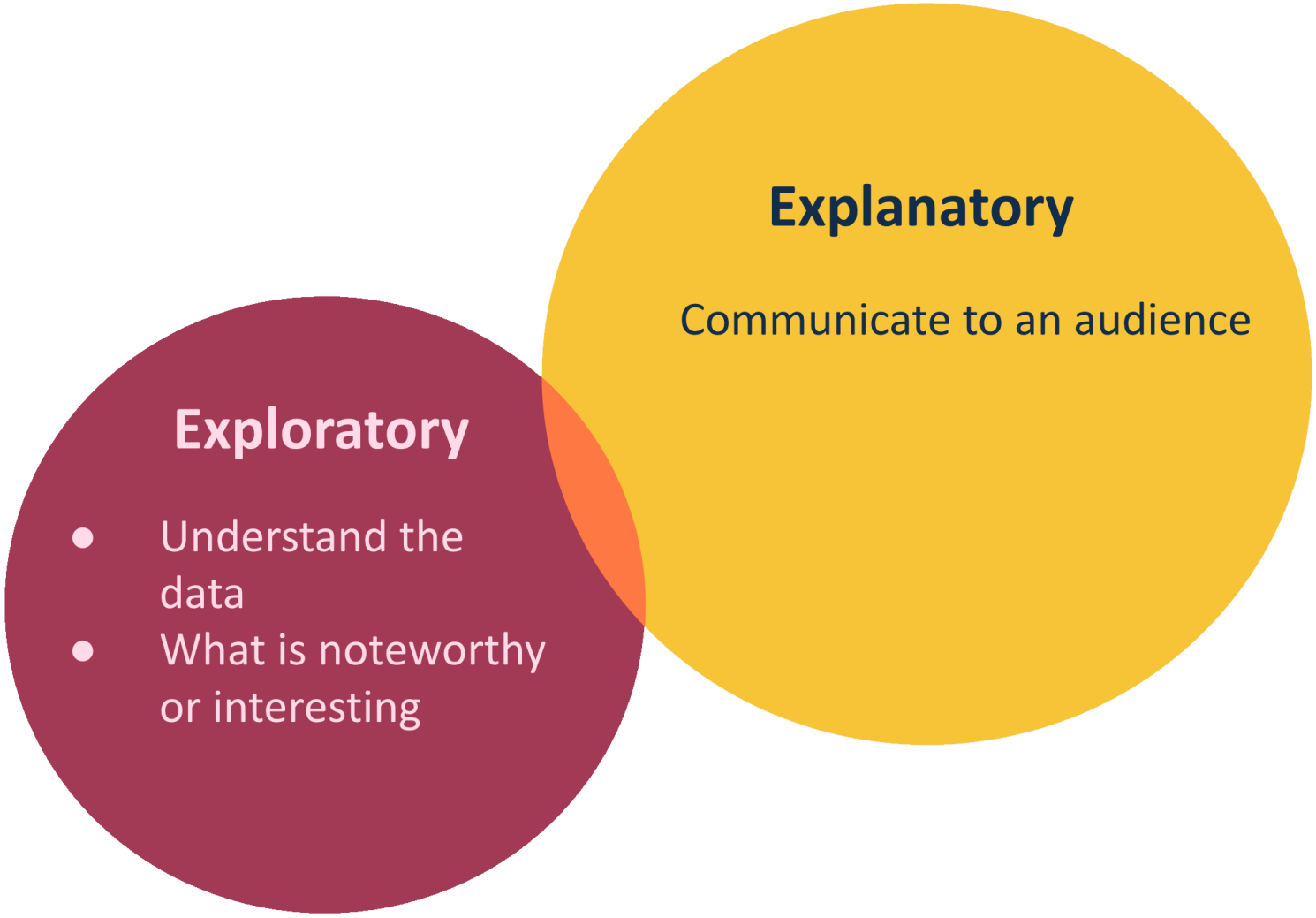
How does it help?

•Identify trends and outliers
•Tell a story within the data
•Reinforce an argument or opinion
•Highlight an important point in a set of data

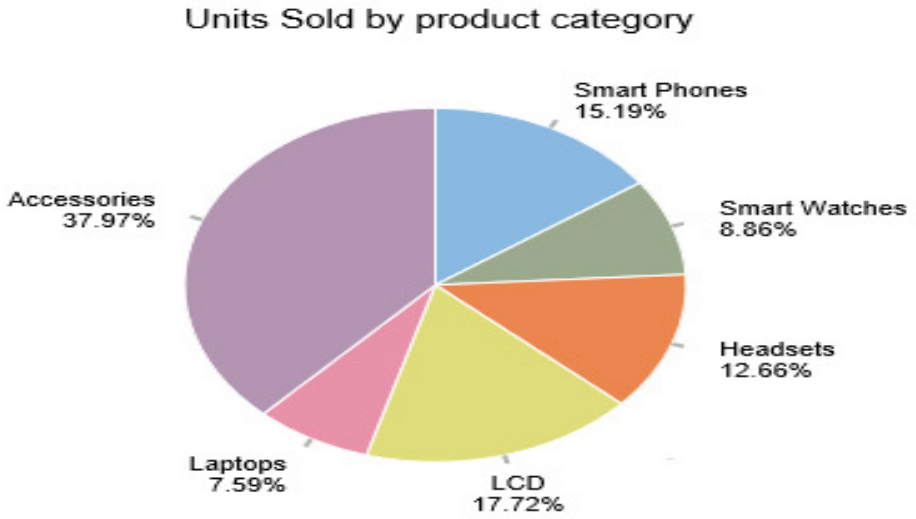
What makes good Data Visualization

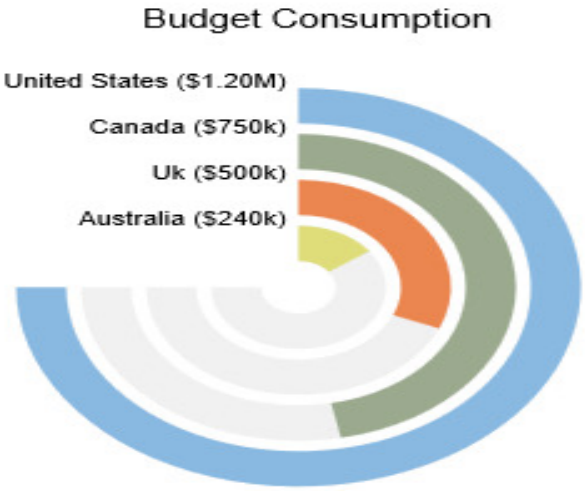


Types of Data Analysis



Exploratory Data Analysis

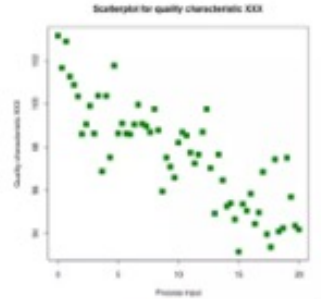


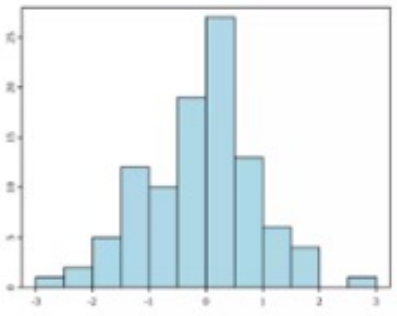


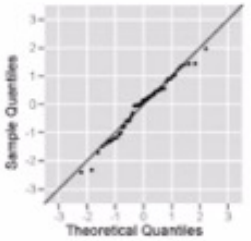
Exploratory data visualization

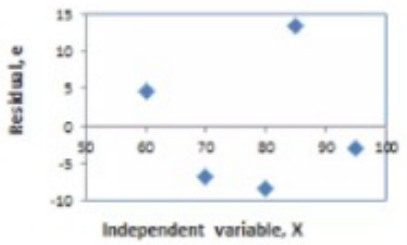
Graphical

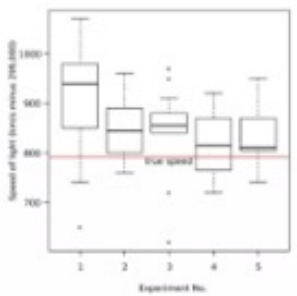
- Scatter plots
- Histograms
- Probability plots
- Residual plots
- Box plots
- Block plots

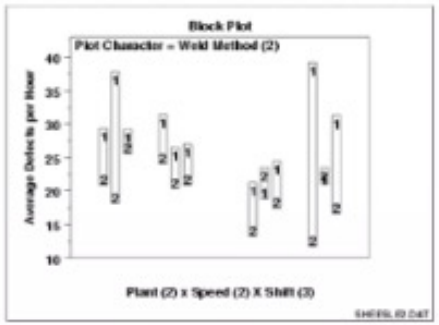












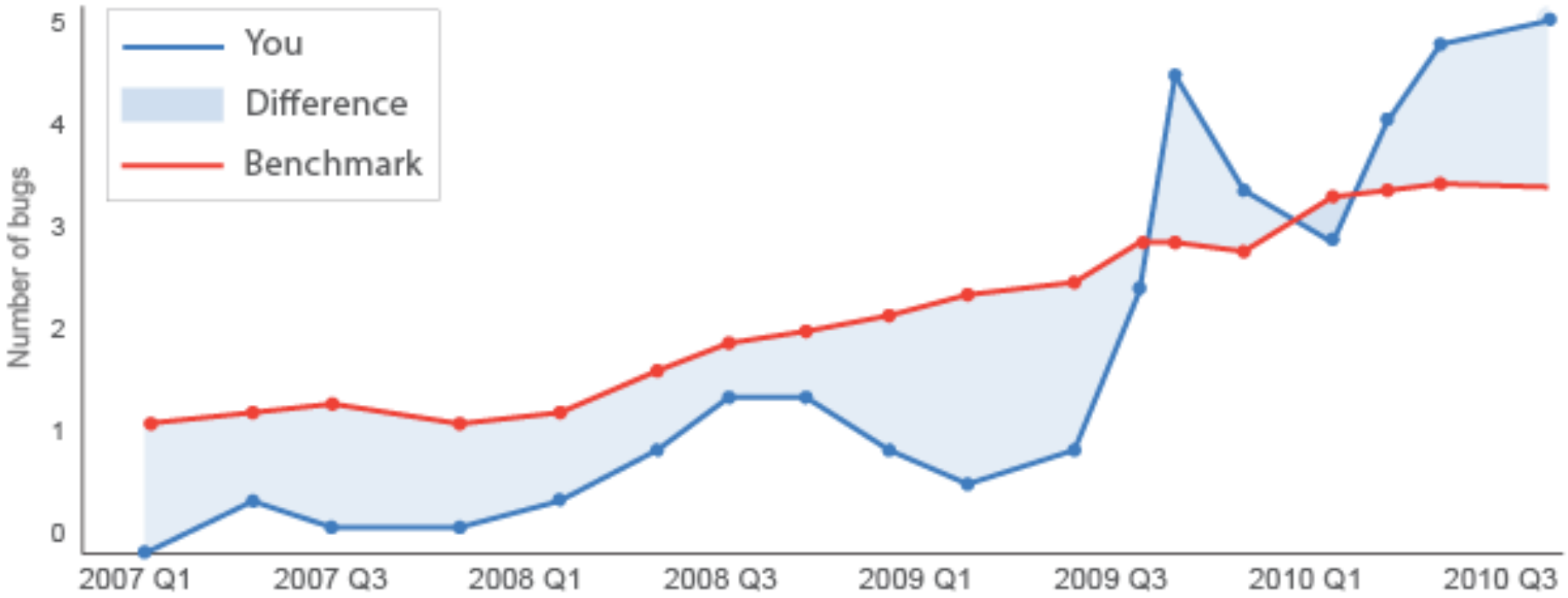
Exploratory data visualization

Graphical analysis procedures:

- Testing assumptions
- Model selection
- Model validation
- Estimator selection
- Relationship identification
- Factor effect determination
- Outlier detection

MUST USE for deriving insights from data

Explanatory Data Analysis



Explanatory data visualization

- Design
- Engineering
- Journalism

Data Visualization in Data Science	Graph	Bar Graph
		Stack Bar Graph
	Plot	Scatter Plot & Line Plot
		Box Plot
		Histogram Plot
	Chart	Pie Chart

Libraries required for data visualization in Python

1. Matplotlib

Matplotlib is one of the best python data visualization libraries for generating powerful yet simple visualization. It is a 2-D plotting library that can be used in various ways, including Python, iPython sheets, and Jupyter notebooks.

Key Features

- It supports various types of graphical representation, including line graphs, bar graphs, and histograms.
- It can work with the NumPy arrays and border SciPy stack.
- It has a huge amount of plots for understanding trends and making correlations.

Pros And Cons

- Interactive platform
- Versatile library

Libraries required for data visualization in Python

2. Plotly

The most popular data visualization library in Python is Plotly, which delivers an interactive plot and is easily readable to beginners. It is widely used for handling financial, geographical, statistical, and scientific data.

Key Features

- Its robust API functions effectively in both local and web browser modes.
- It is an interactive, open-source, and high-level visualization library.
- It can be viewed in Jupyter notebooks, standalone HTML files, or even hosted online.

Pros And Cons

- Offers contour plots, dimension chars, and dendrograms.
- Allows 40 unique chart and plot types
- Difficult to use

Libraries required for data visualization in Python

3. Seaborn

Seaborn is the best python library for data visualization, which offers a variety of visualized patterns. It is designed to work more compatible with Pandas data form and is widely used for statistical visualization.

Key Features

- It performs the necessary mapping and aggregation to form information visuals.
- It is integrated to explore and understand data in a better and more detailed way.
- It offers a high level of a crossing point for creating beautiful and informative algebraic graphics.

Pros And Cons

- Much more visually appealing representation
- Switch to any other data format
- Limited customizable options

Libraries required for data visualization in Python

4. Ggplot

GGplot is another popular data visualization library in Python, known as the python implementation of graphics grammar. It refers to the map of the data, with its aesthetic attributes including color, shape, and geometric objects like points and bars.

Key Features

- It allows you to build informative visualization substantially with improved representations
- It is integrated with Panda to store data in a data frame.
- It is based on ggplot2, an R programming language plotting system.

Pros And Cons

- Documentation is simple and easy to follow.
- Save method to discuss and exhibit plots
- Not suitable for creating highly customized graphics.

Libraries required for data visualization in Python

5. Bokeh

Bokeh is another interactive python library for data visualized for modern web browsers. This is best suitable for developing interactive plots and dashboards for complex or streaming data assets.

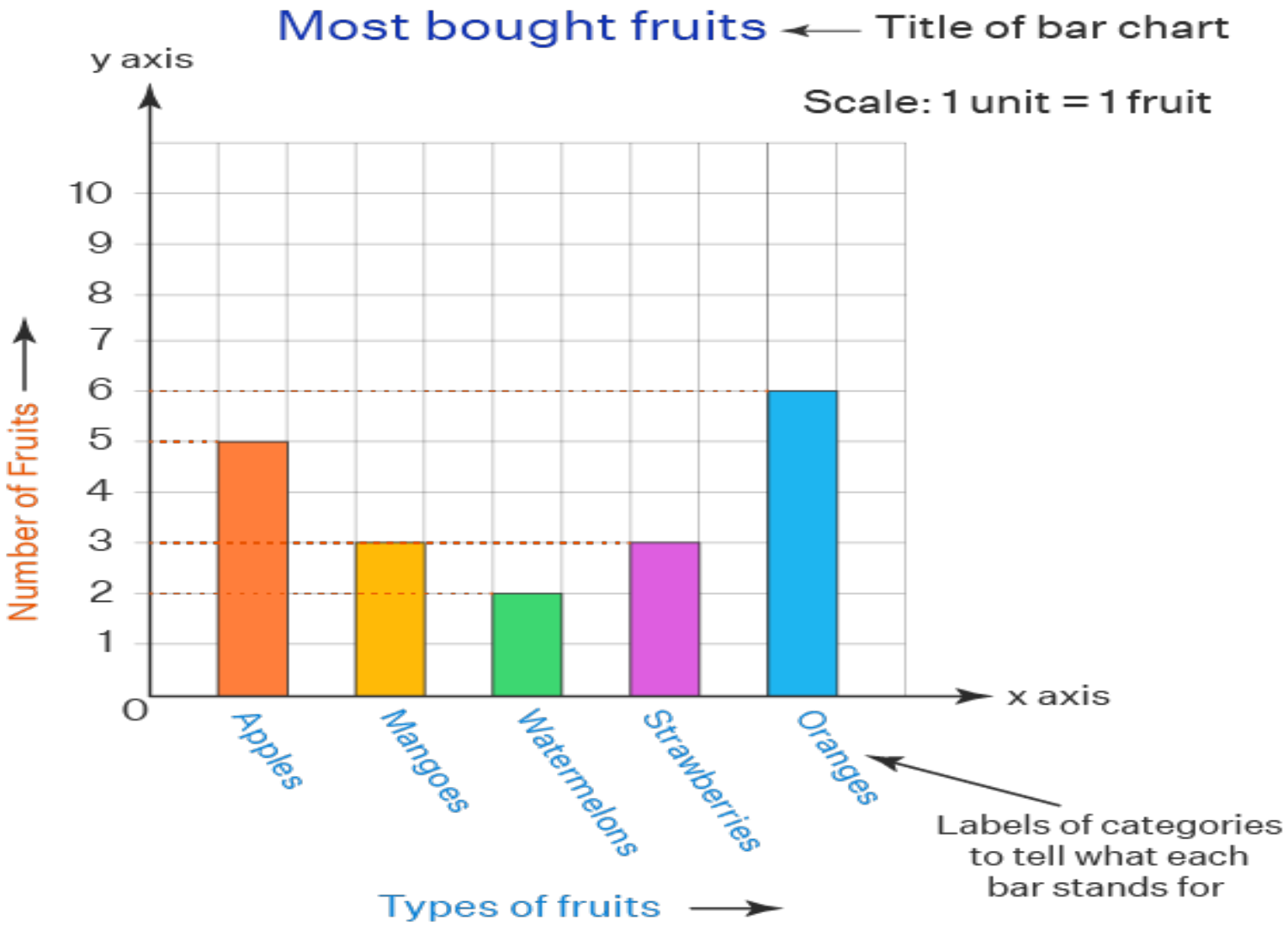
Key Features

•It has a broad range of intuitive graphs which can be leveraged to form solutions.
•It is well-known for creating custom-made visualizations.
•It includes various generation and plot chart methods, including box plots, bar plots, and histograms.

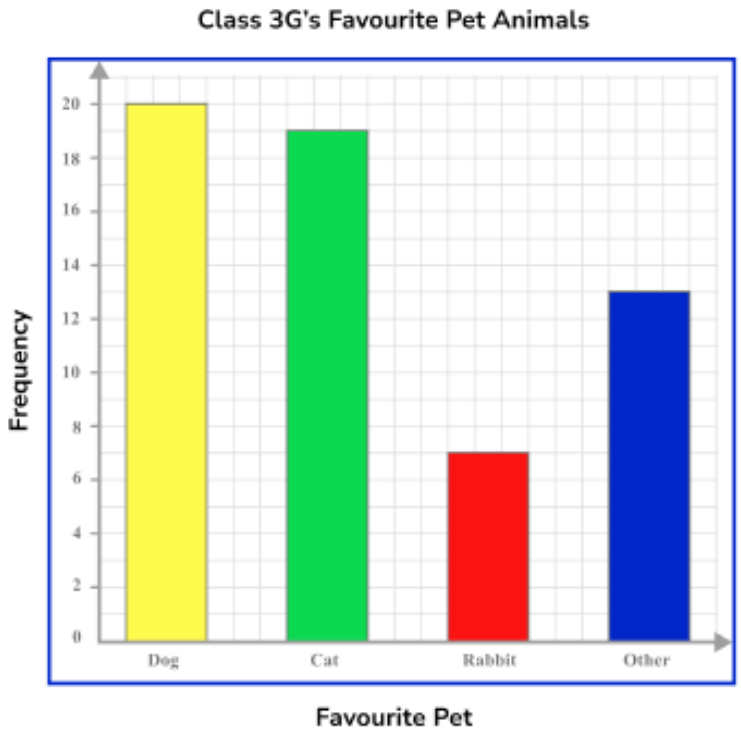
Pros And Cons

•Highest level of control for the rapid creation of charts
•Many graphs with fewer codes and higher resolution
•No pre-set defaults, and users have to define them each time.

Bar Graph



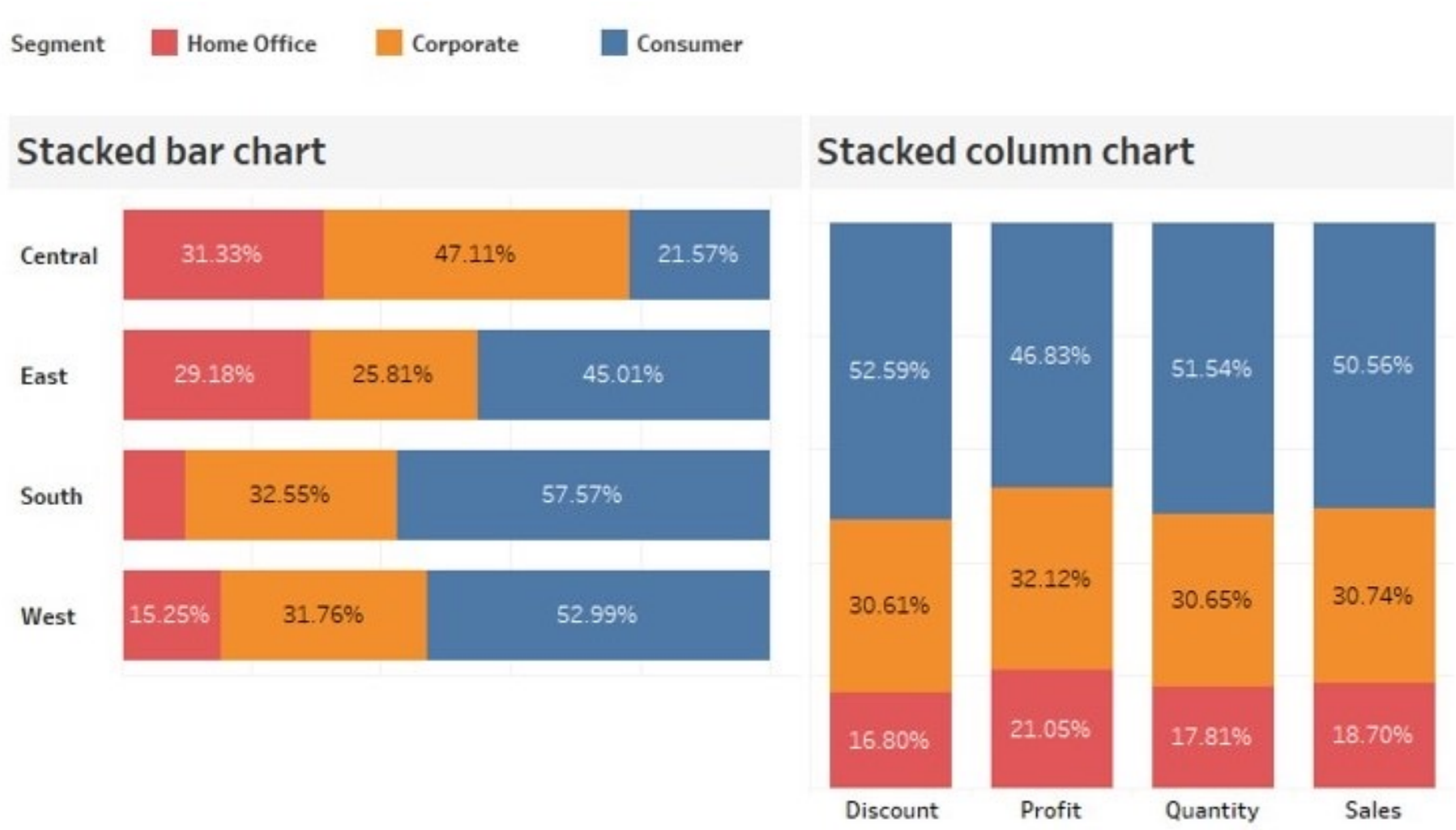
A **bar chart** represents a data set by using vertical or horizontal bars. The larger the bar, the higher the value for the individual category.



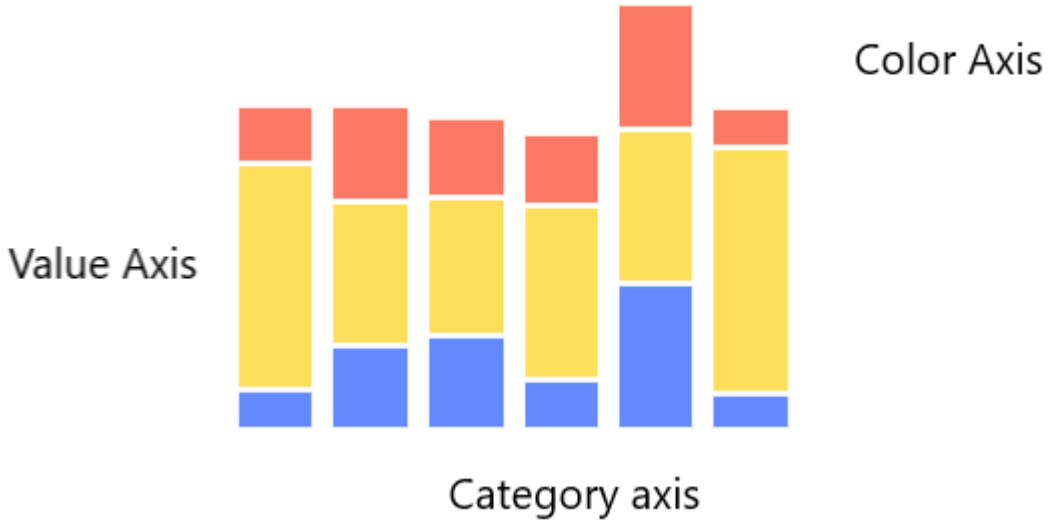
To draw a **bar chart** we need the following:

- A pair of axes. Usually the horizontal axis is labelled with the categories of the data set and the vertical axis is the frequency. Your axes must be labelled.
- The frequencies need to be labelled on the vertical axis in equal intervals.
- The bars need to have equal gaps between them as it is representing discrete data.
- The bars need to be of equal width.
- The chart needs a title.

Stacked Bar Graph

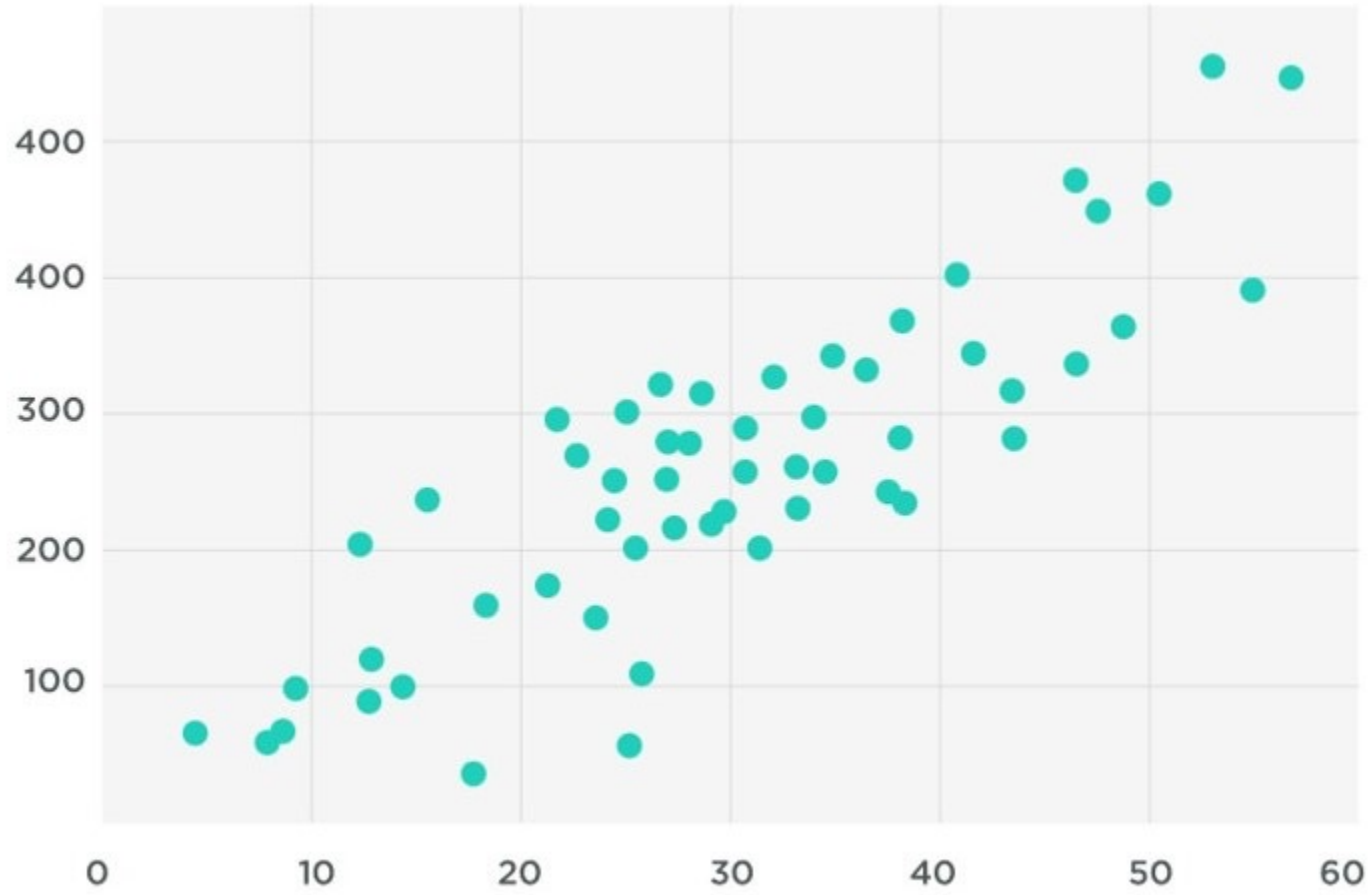


Stack Bar Graph

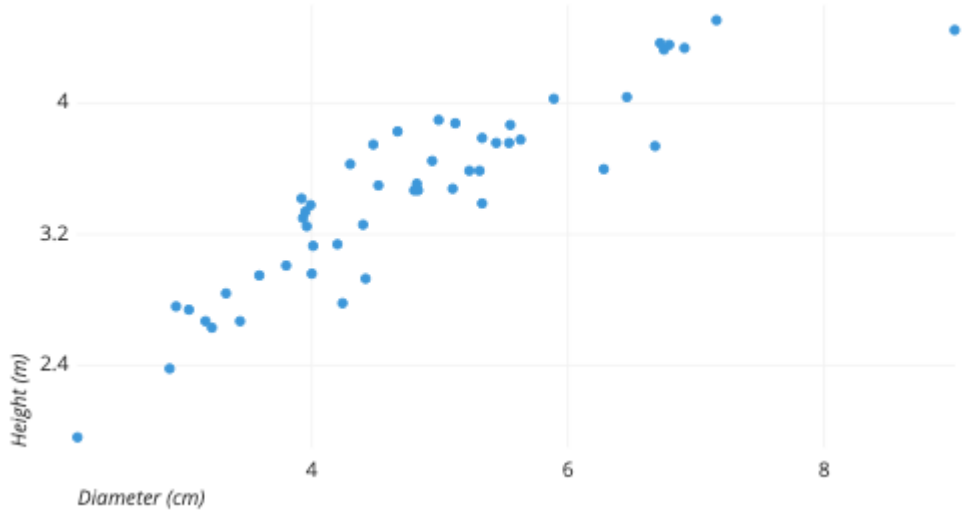


Stack Bar Graph is a form of bar chart that shows the composition and comparison of a few variables, either relative or absolute, over time. Also called a stacked bar or column chart, they look like a series of columns or bars that are stacked on top of each other.

Scatter Plot

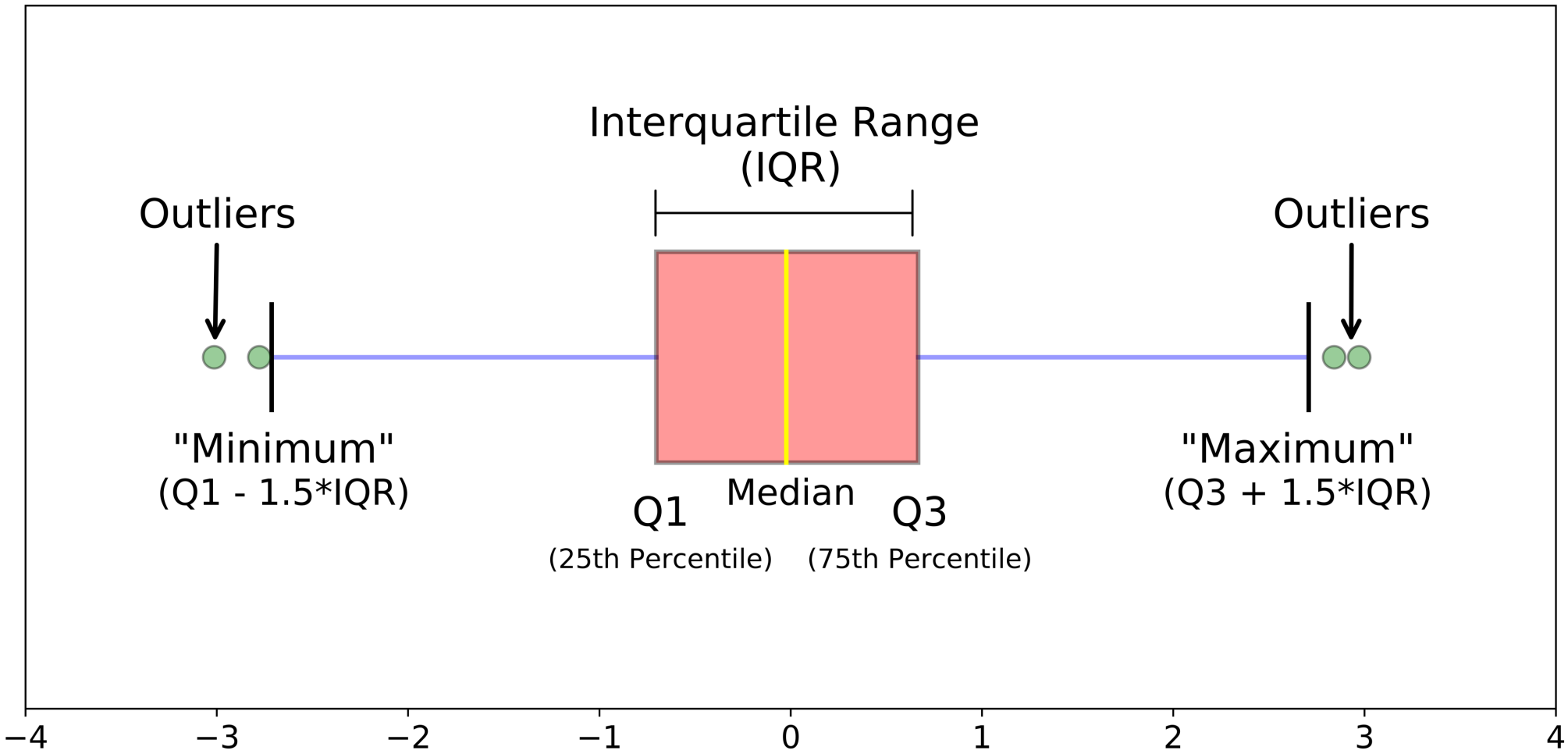


Scatter Plot

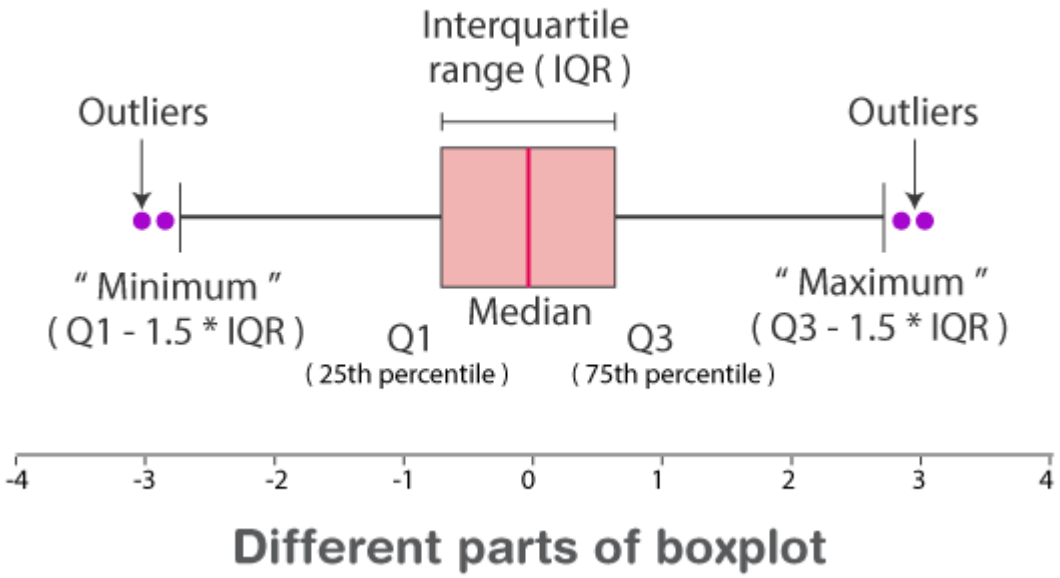


A scatter plot uses dots to represent values for two different numeric variables. The position of each dot on the horizontal and vertical axis indicates values for an individual data point. Scatter plots are used to observe relationships between variables.

Box Plot



Box Plot

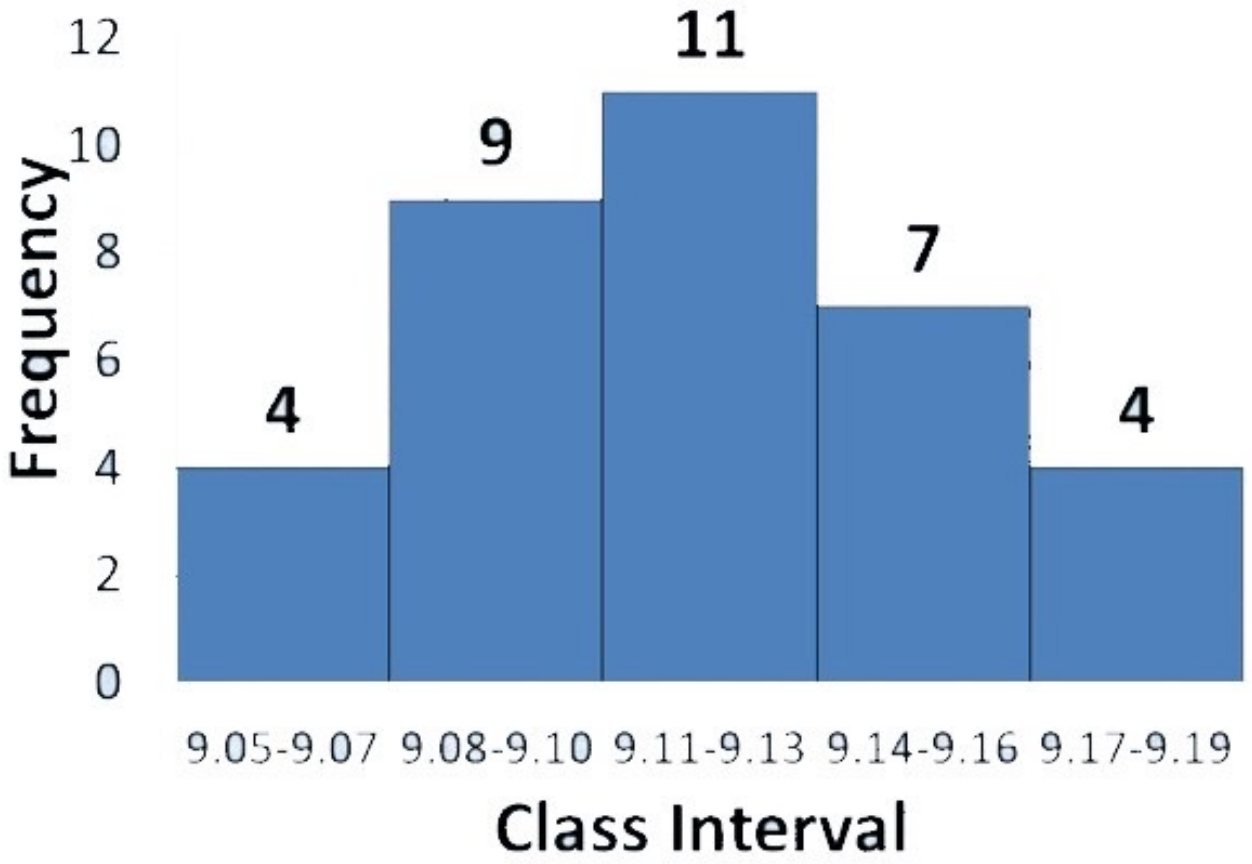


Different parts of boxplot

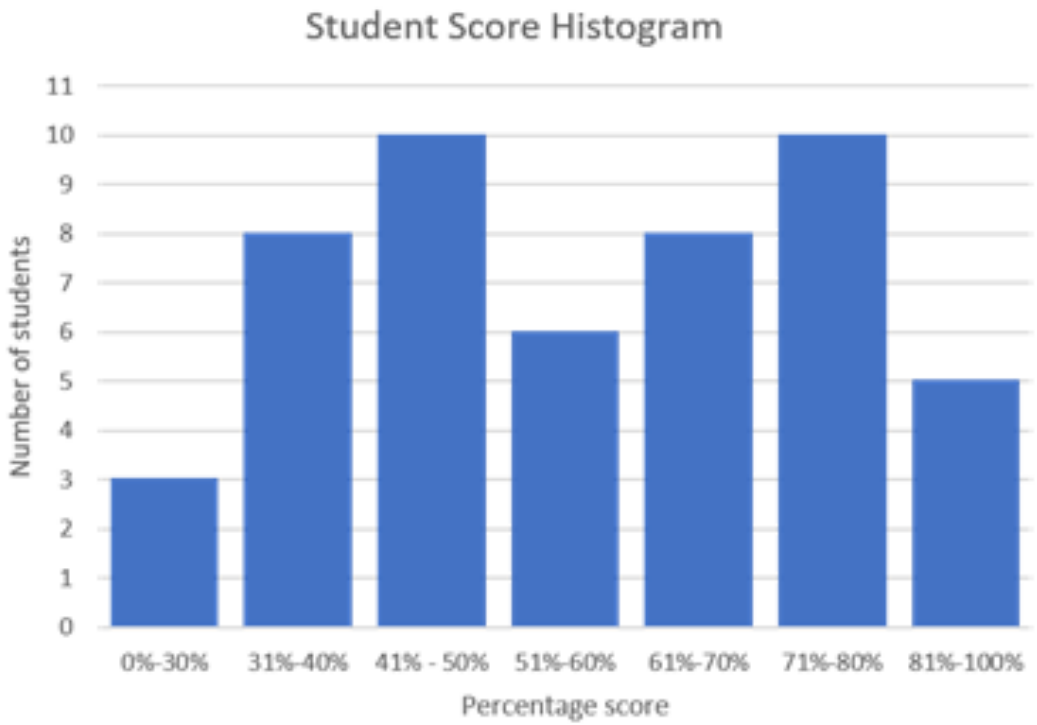
A box plot is a chart that shows data from a five-number summary including one of the measures of central tendency.
It helps to find out how much the data values vary or spread out with the help of graphs.

Minimum: The minimum value in the given dataset
First Quartile (Q1): The first quartile is the median of the lower half of the data set.
Median: The median is the middle value of the dataset, which divides the given dataset into two equal parts. The median is considered as the second quartile.
Third Quartile (Q3): The third quartile is the median of the upper half of the data.
Maximum: The maximum value in the given dataset.
Apart from these five terms, the other terms used in the box plot are:
Interquartile Range (IQR): The difference between the third quartile and first quartile is known as the interquartile range. (i.e.) IQR = Q3-Q1
Outlier: The data that falls on the far left or right side of the ordered data is tested to be the outliers. Generally, the outliers fall more than the specified distance from the first and third quartile.
(i.e.) Outliers are greater than Q3+(1.5 . IQR) or less than Q1-(1.5 . IQR).

Histogram

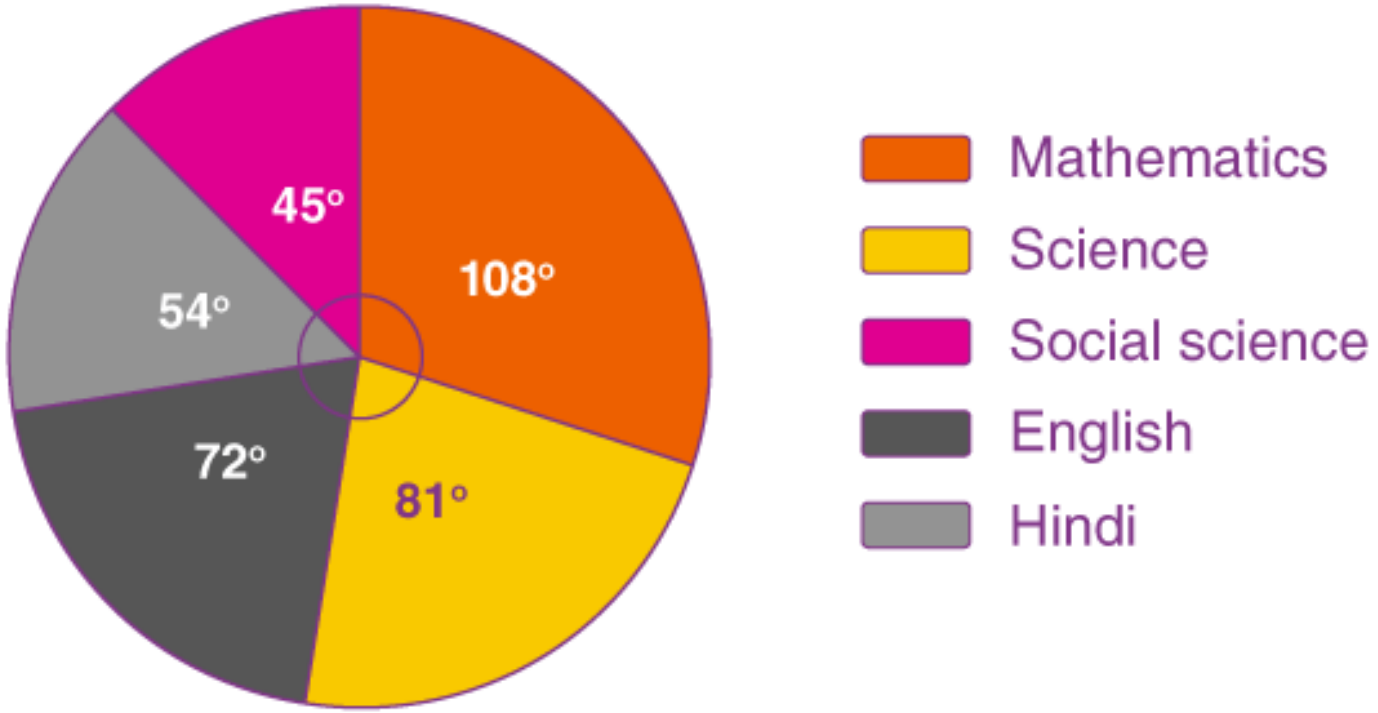


Histogram

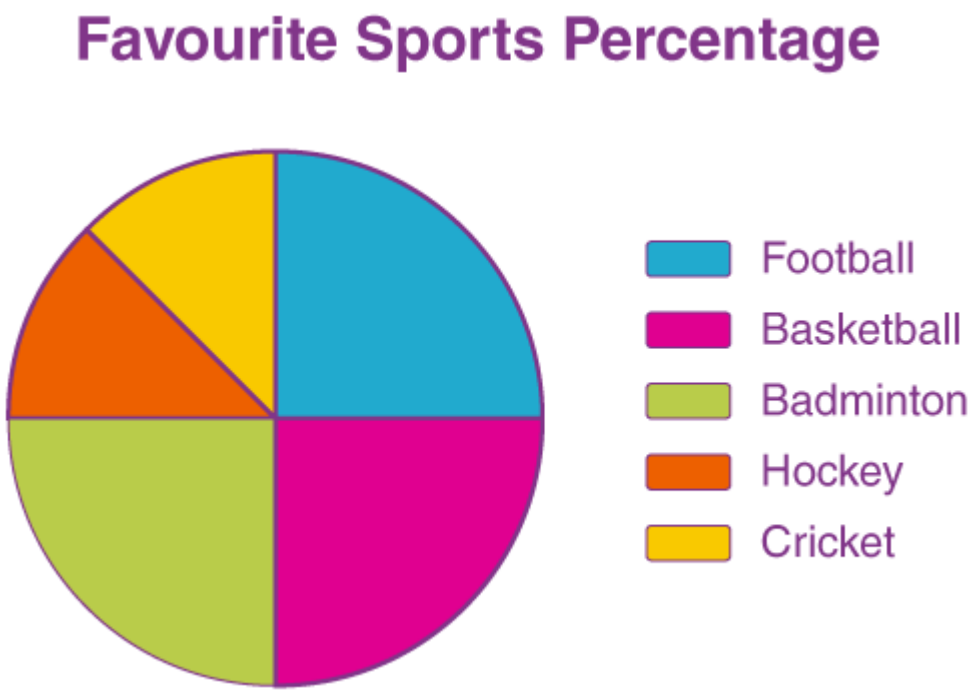


A histogram is a graph used to represent the frequency distribution of a few data points of one variable. Histograms often classify data into various “bins” or “range groups” and count how many data points belong to each of those bins.

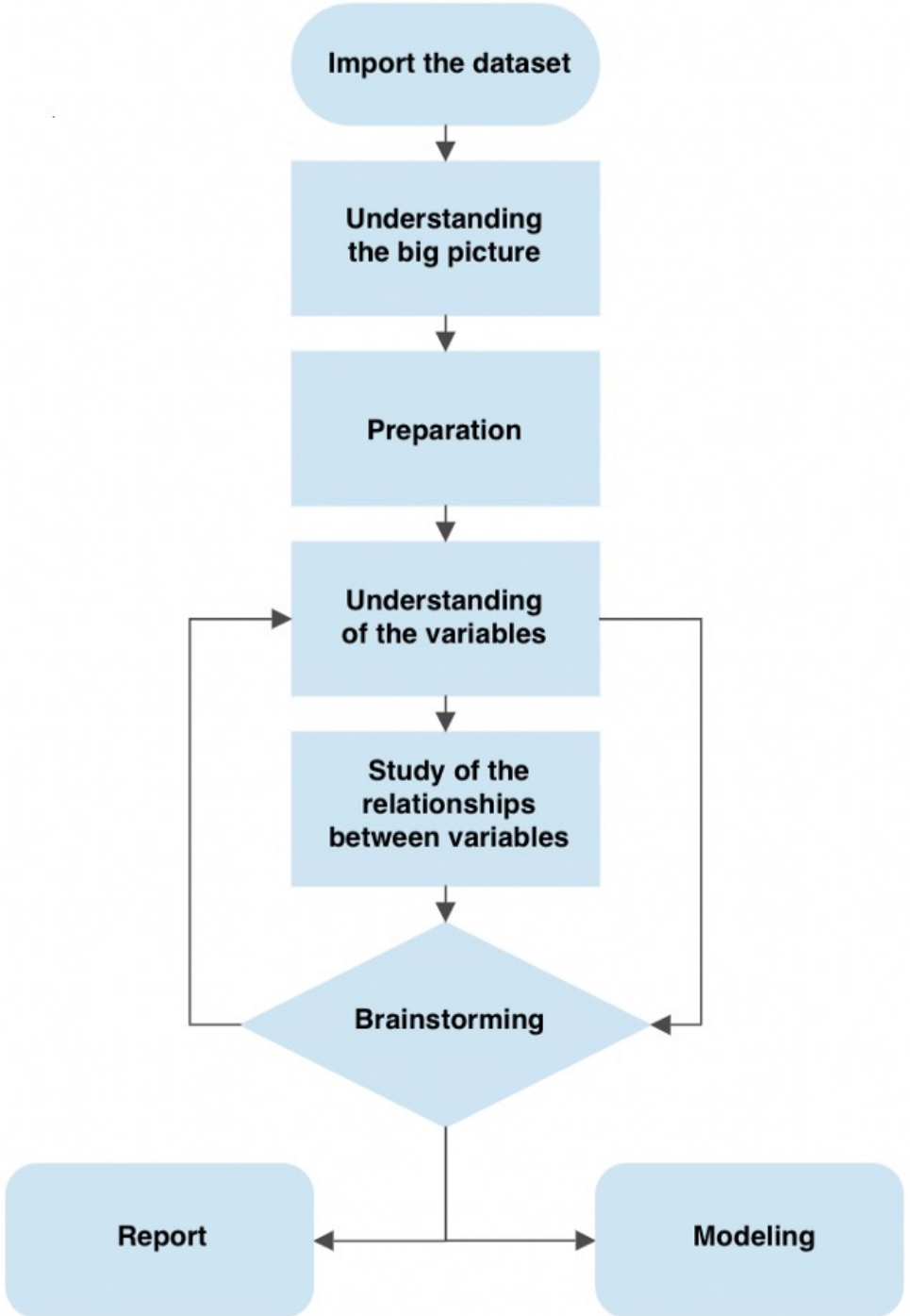
Pie Chart



Pie Chart

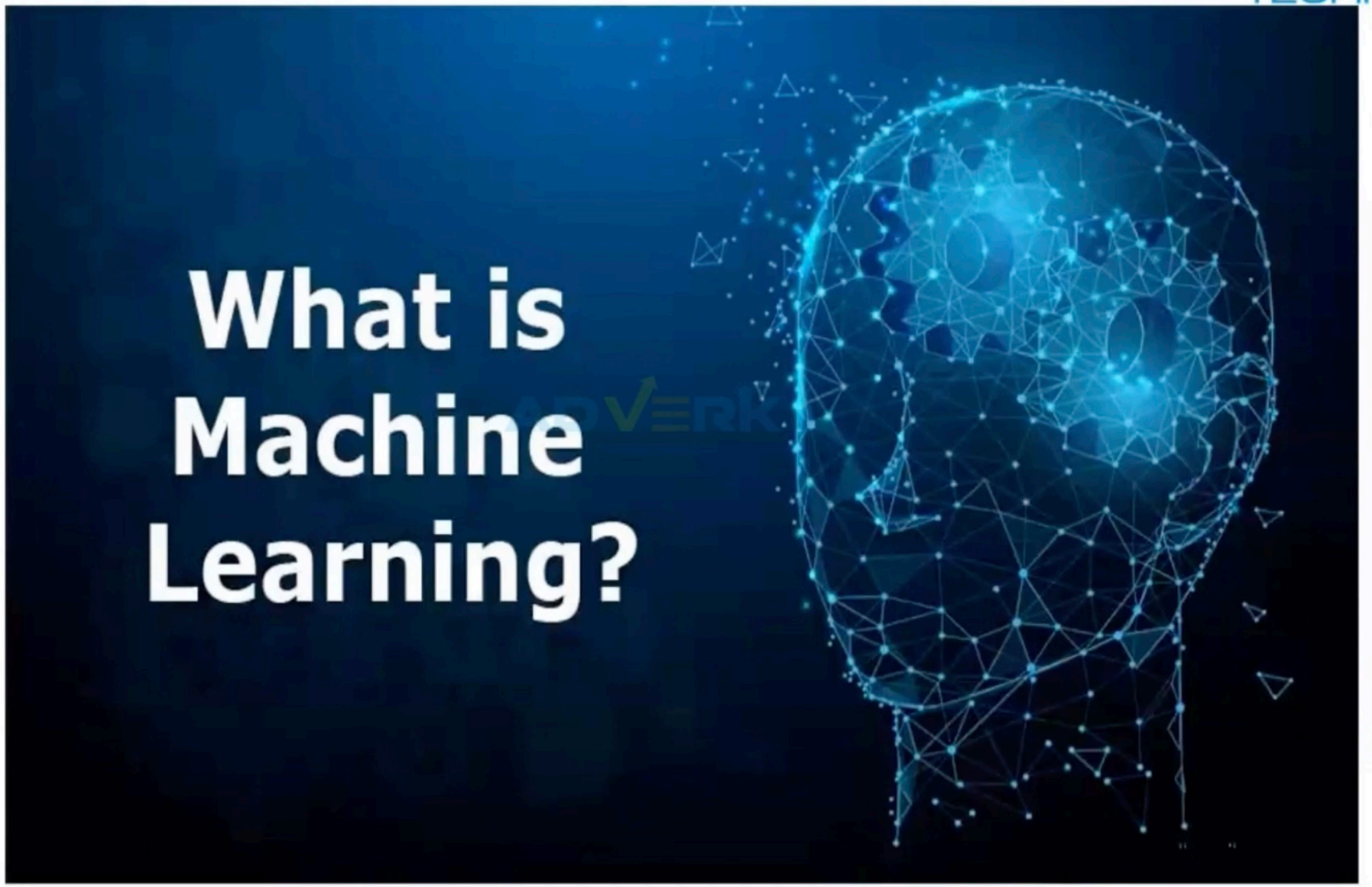


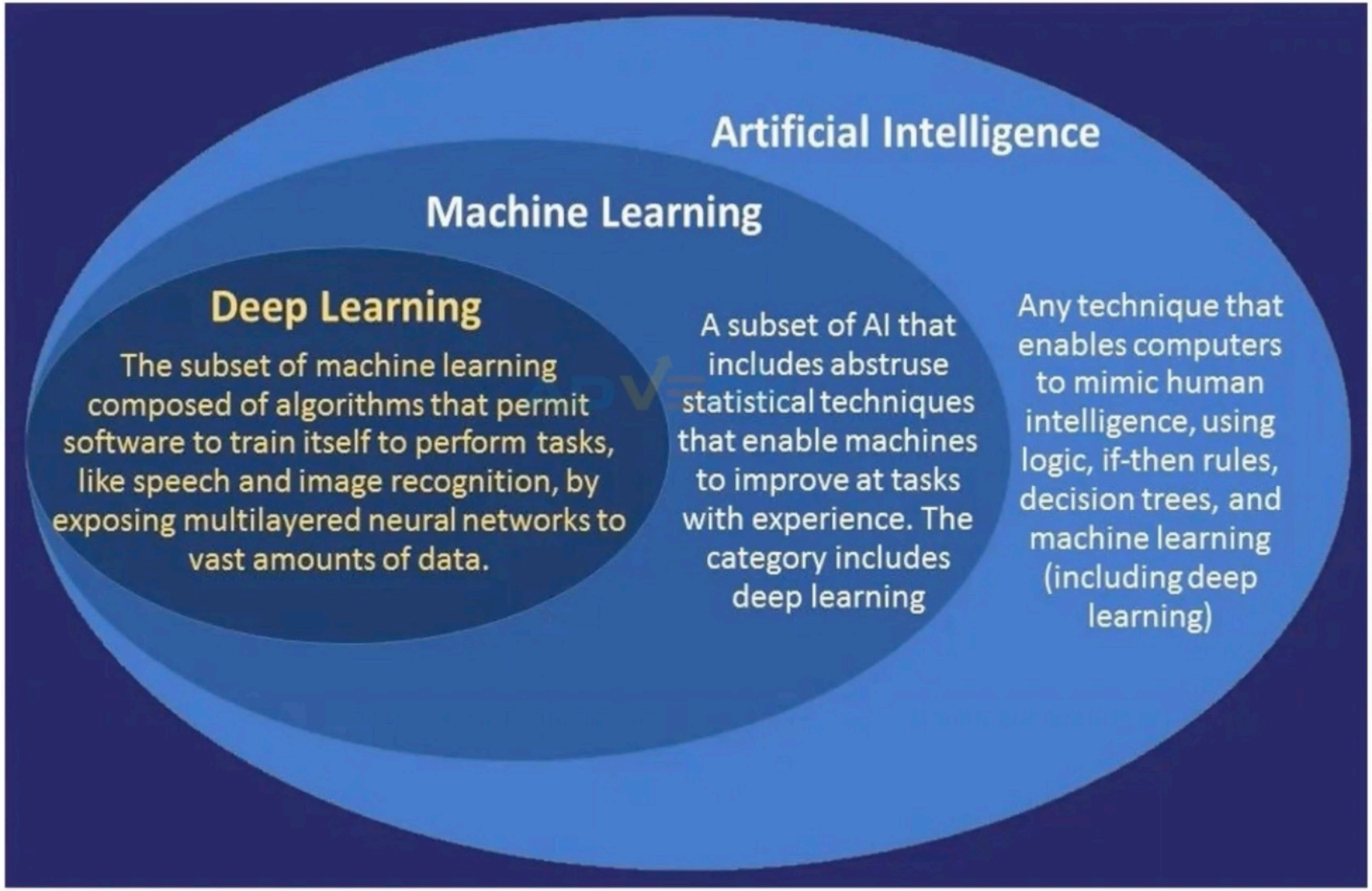
A pie chart is a type of graph that represents the data in the circular graph. The slices of pie show the relative size of the data, and it is a type of pictorial representation of data. A pie chart requires a list of categorical variables and numerical variables. Here, the term “pie” represents the whole, and the “slices” represent the parts of the whole.

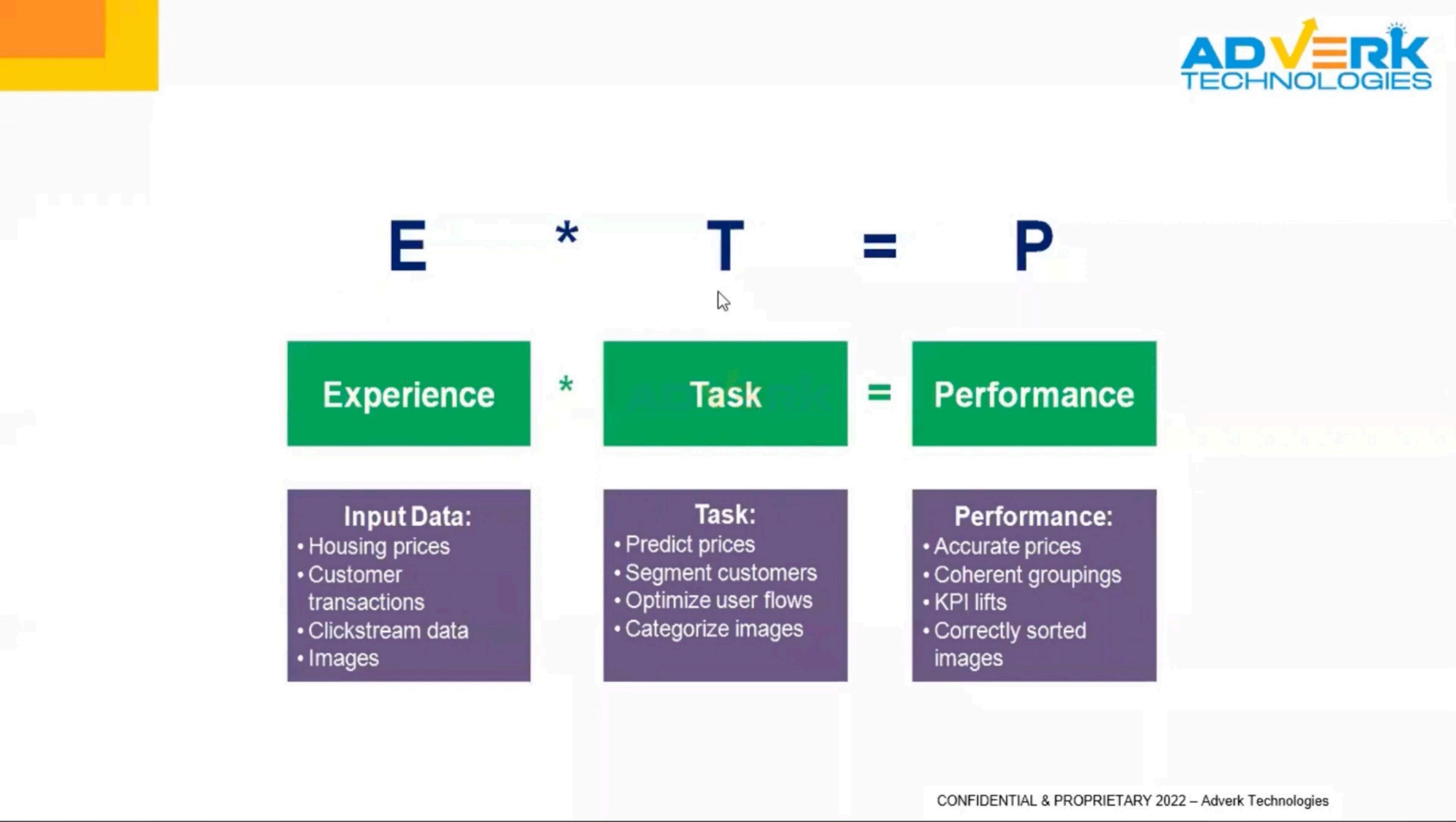


THANK YOU!

ADVERK TECHNOLOGIES



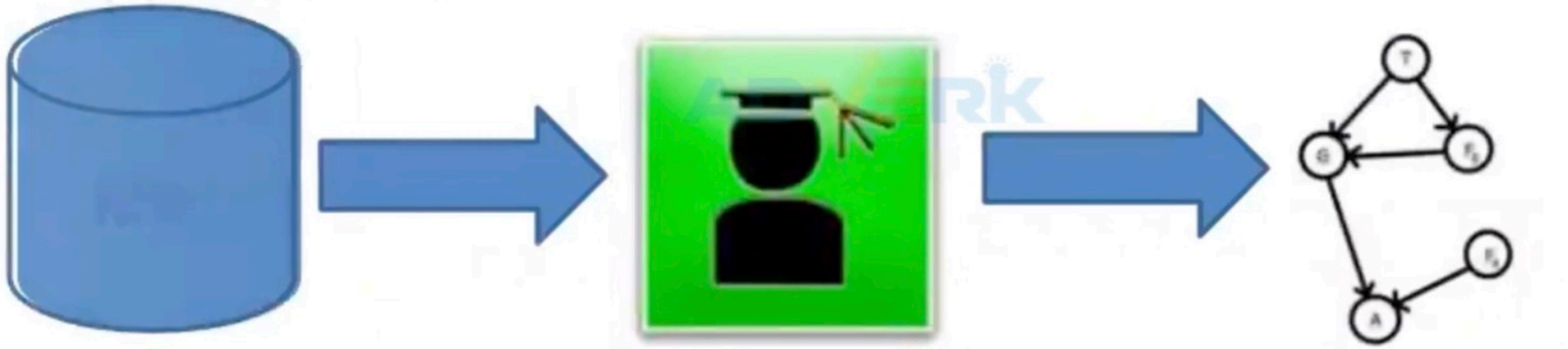




Traditional Programming

Machine Learning

How Machine Learn



Database
Training Set

Learner
Inducer
Induction Algorithm

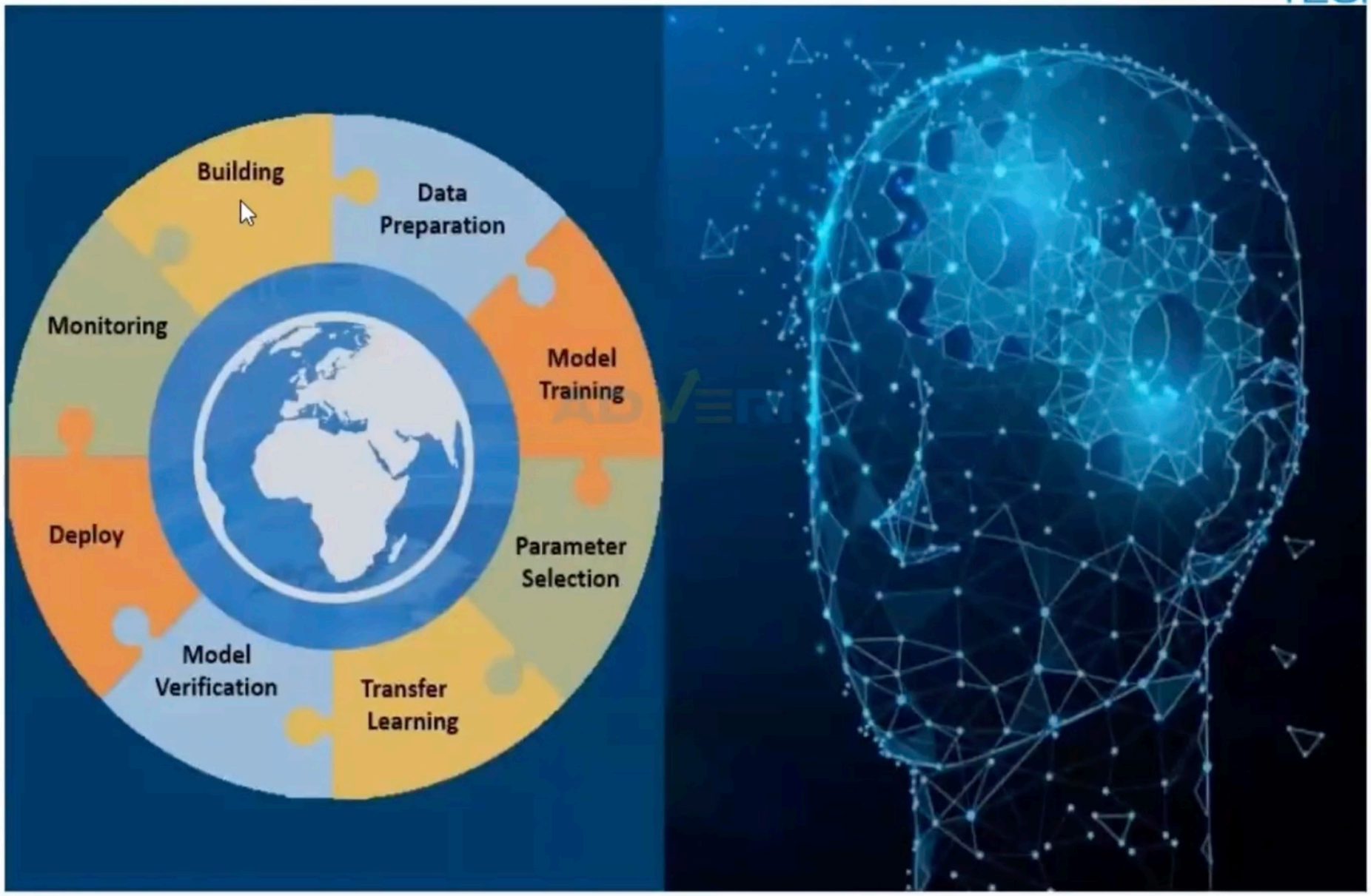
Classifier
Classification Model

How Machine Learn

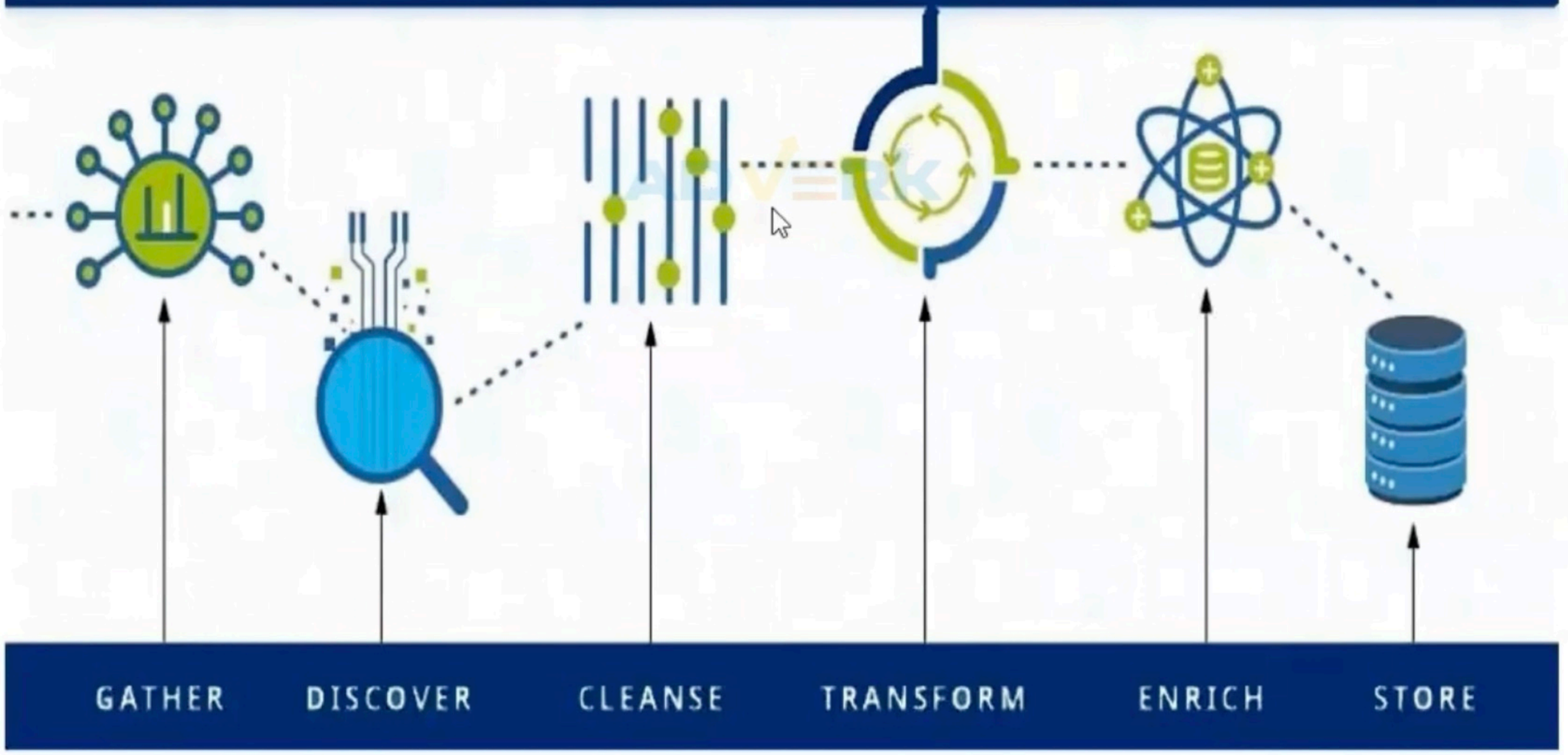
Database
Training Set

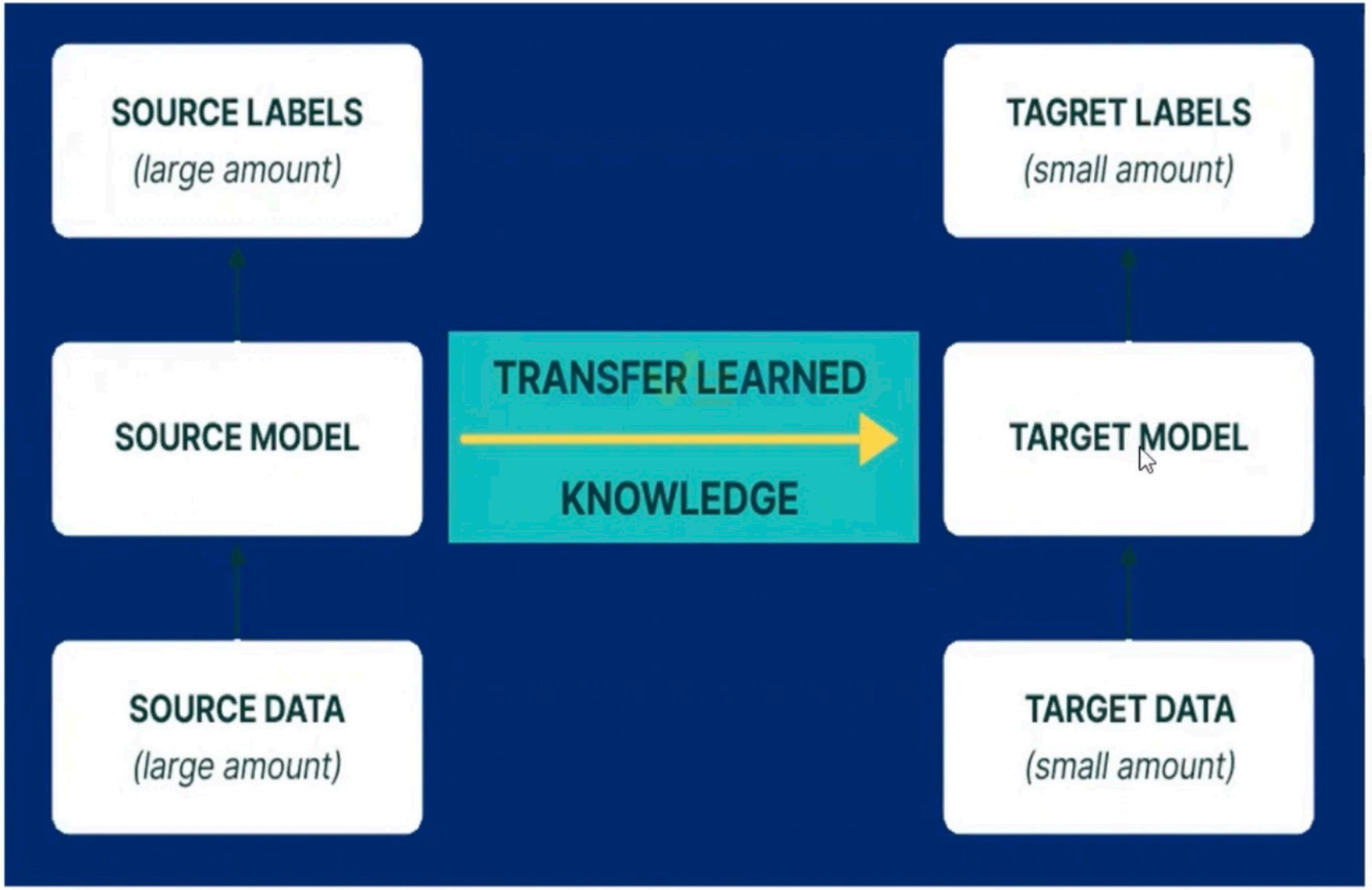
Learner
Inducer
Induction Algorithm

Classifier
Classification Model



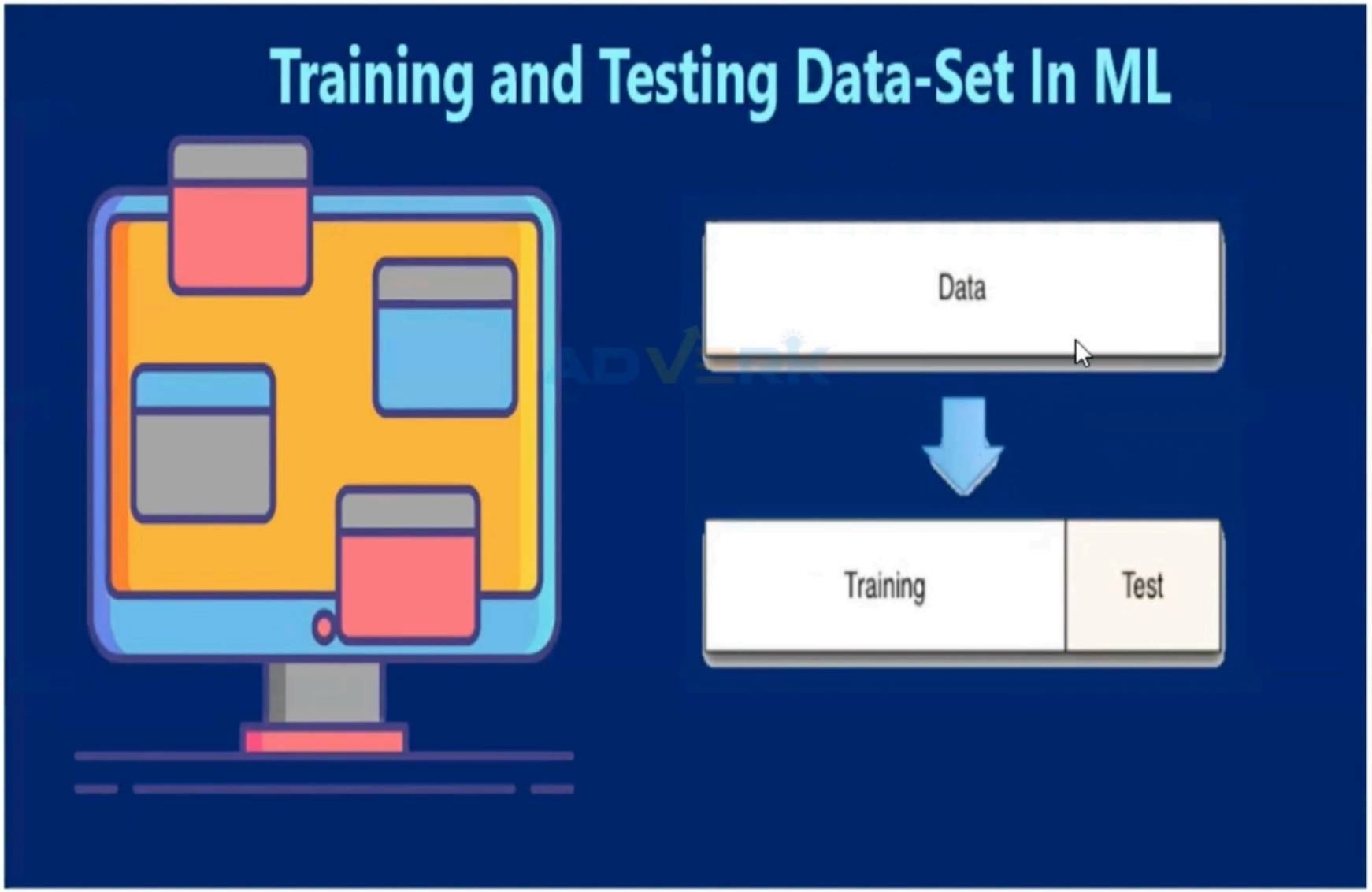
3rd Step: Data Preparation

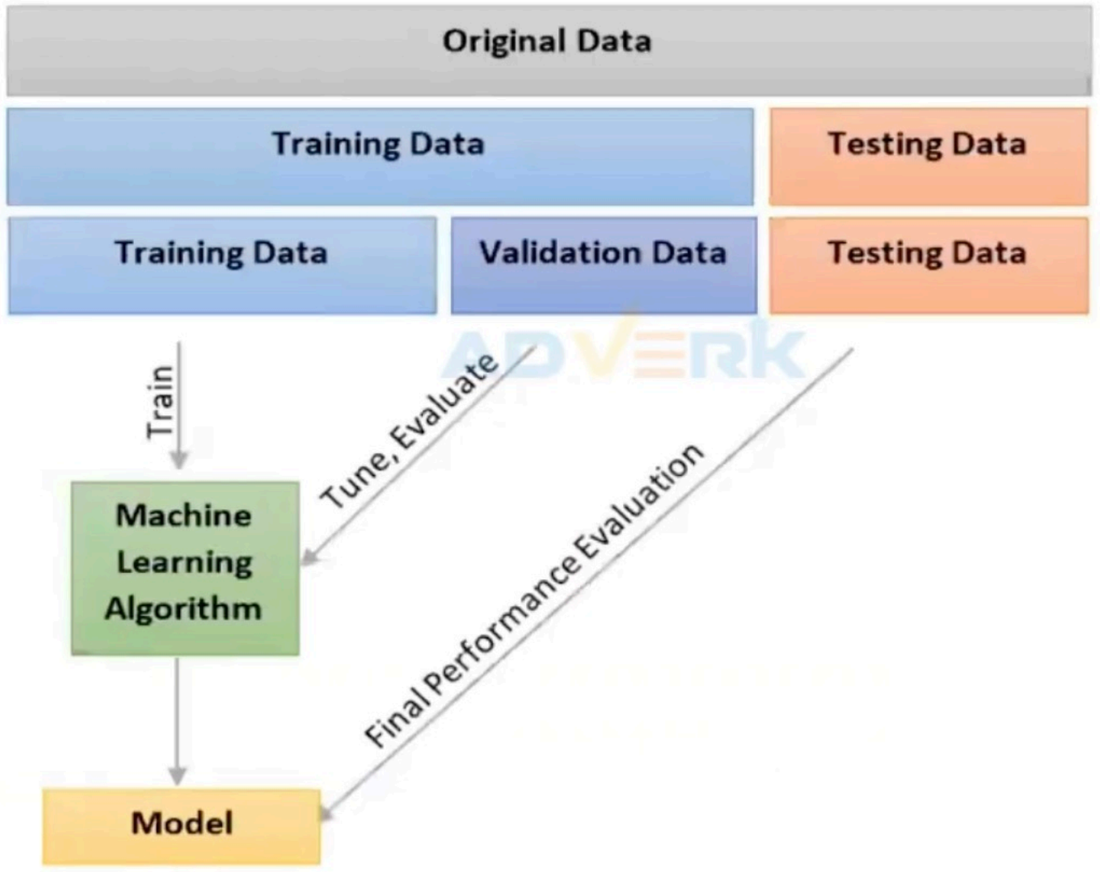




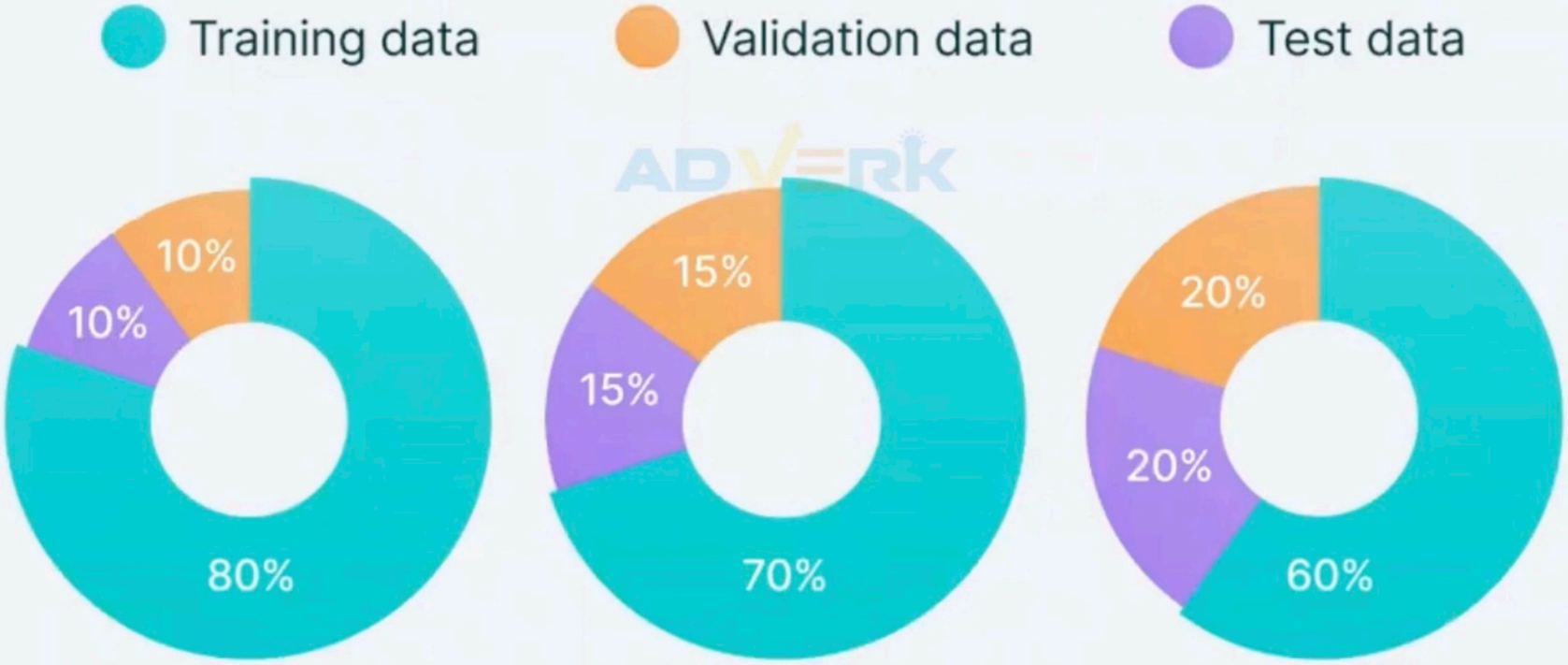
Machine learning Model Validation Testing

MACHINE LEARNING MODEL DEPLOYMENT

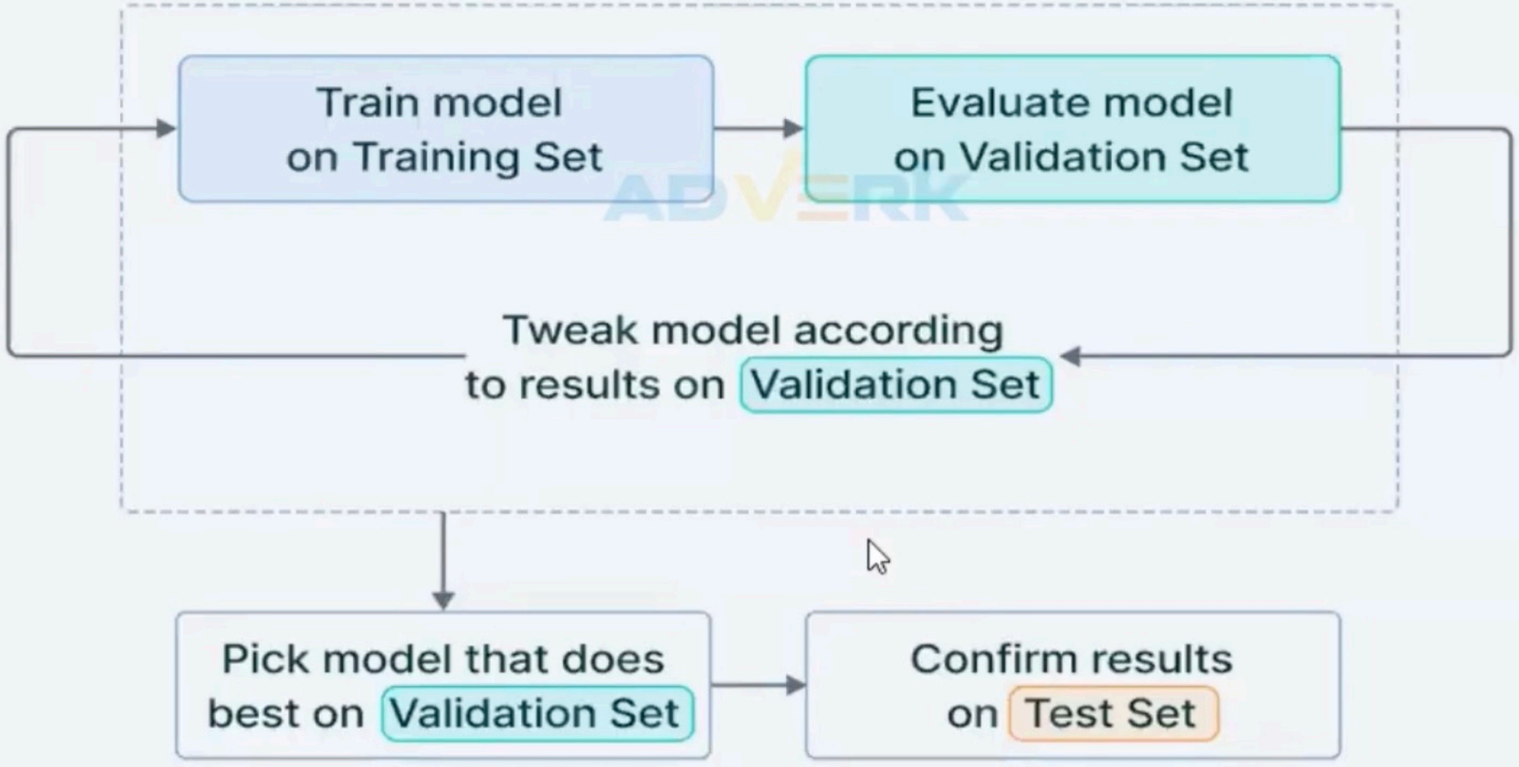




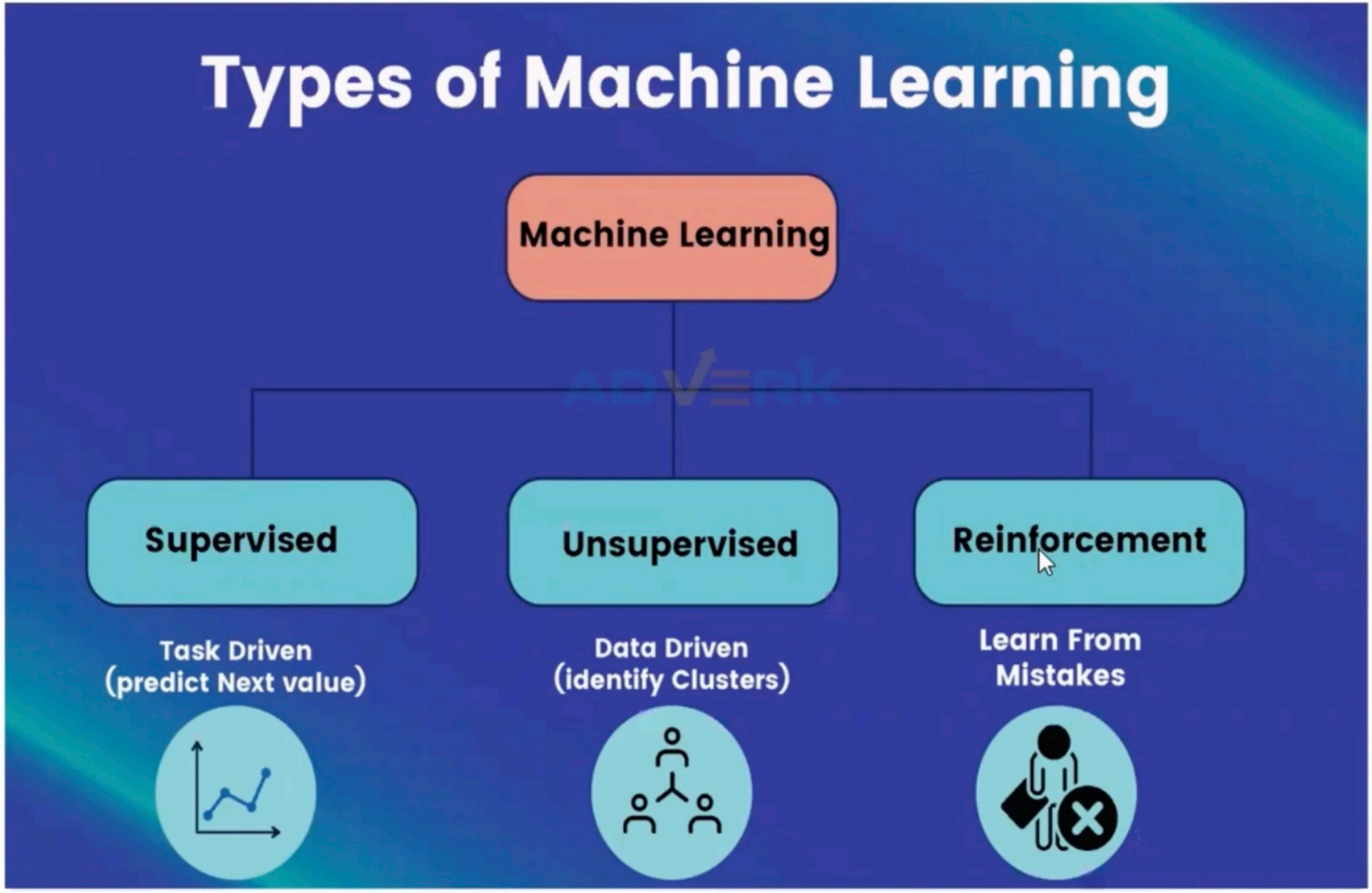
Data Training Needs

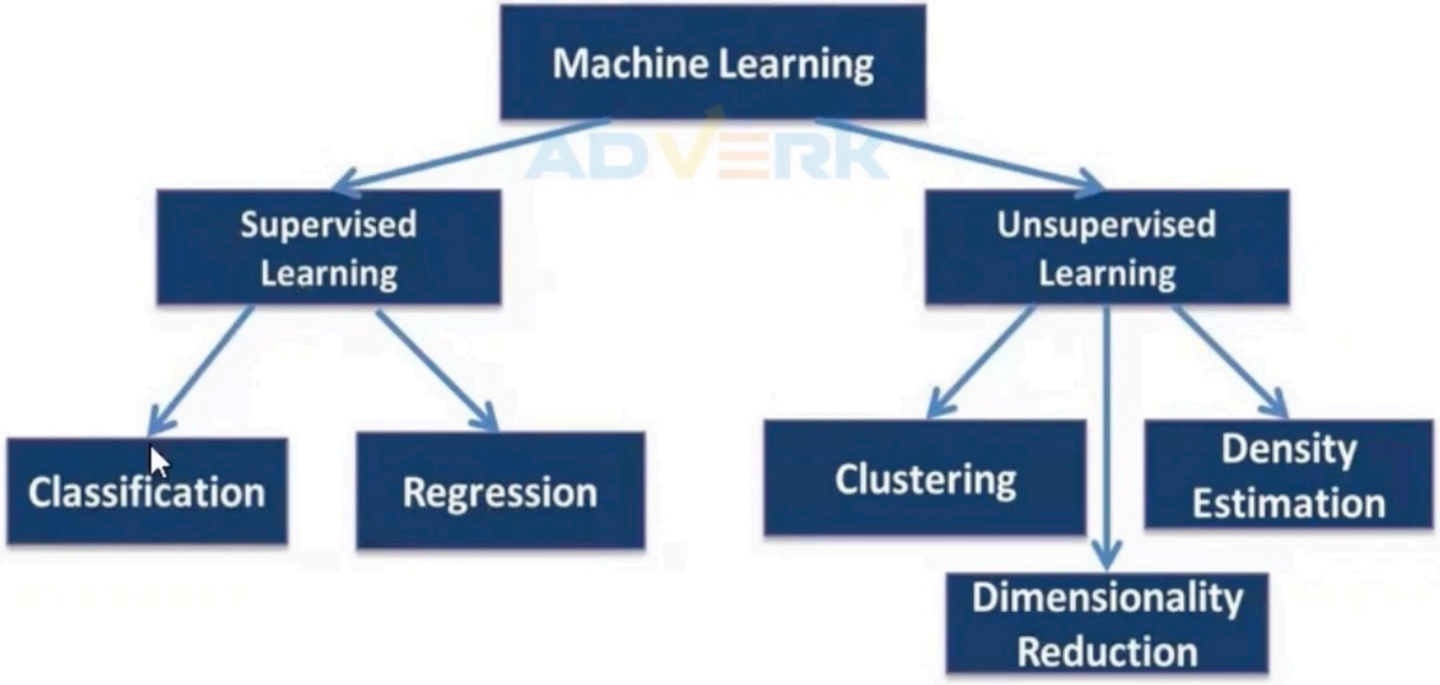


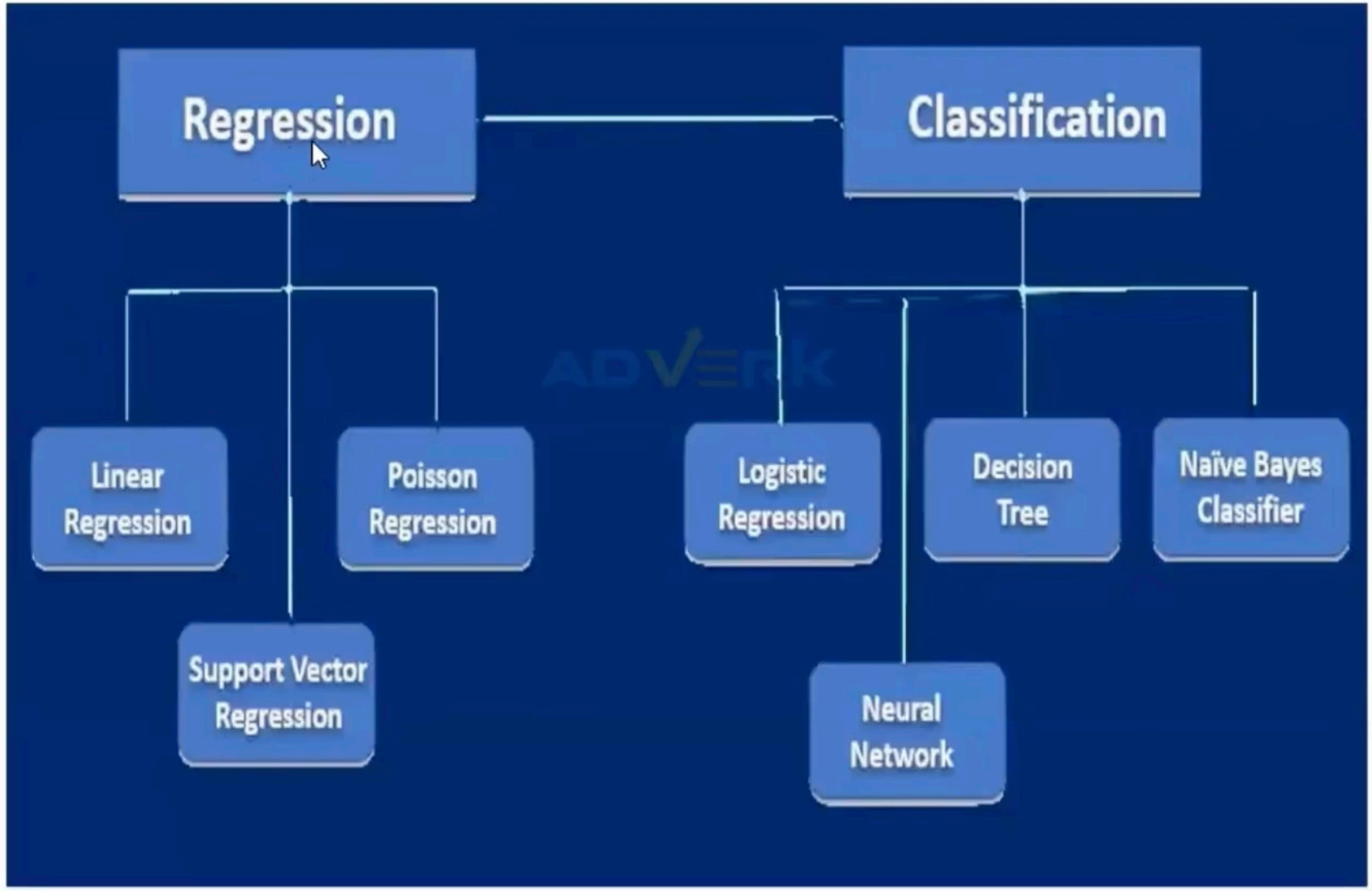
Training data/validation/test

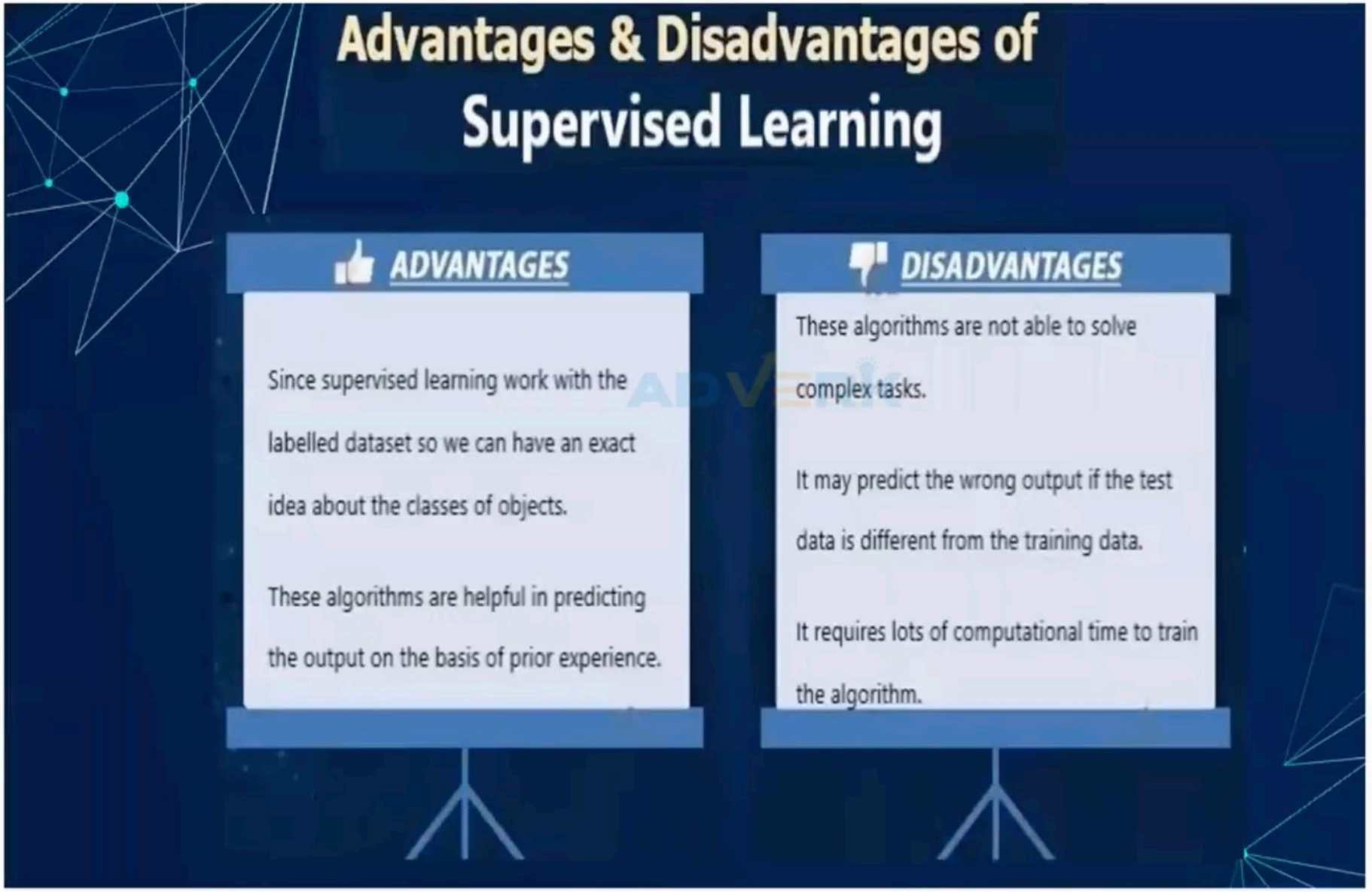


Types of Machine Learning









Applications of Supervised Learning

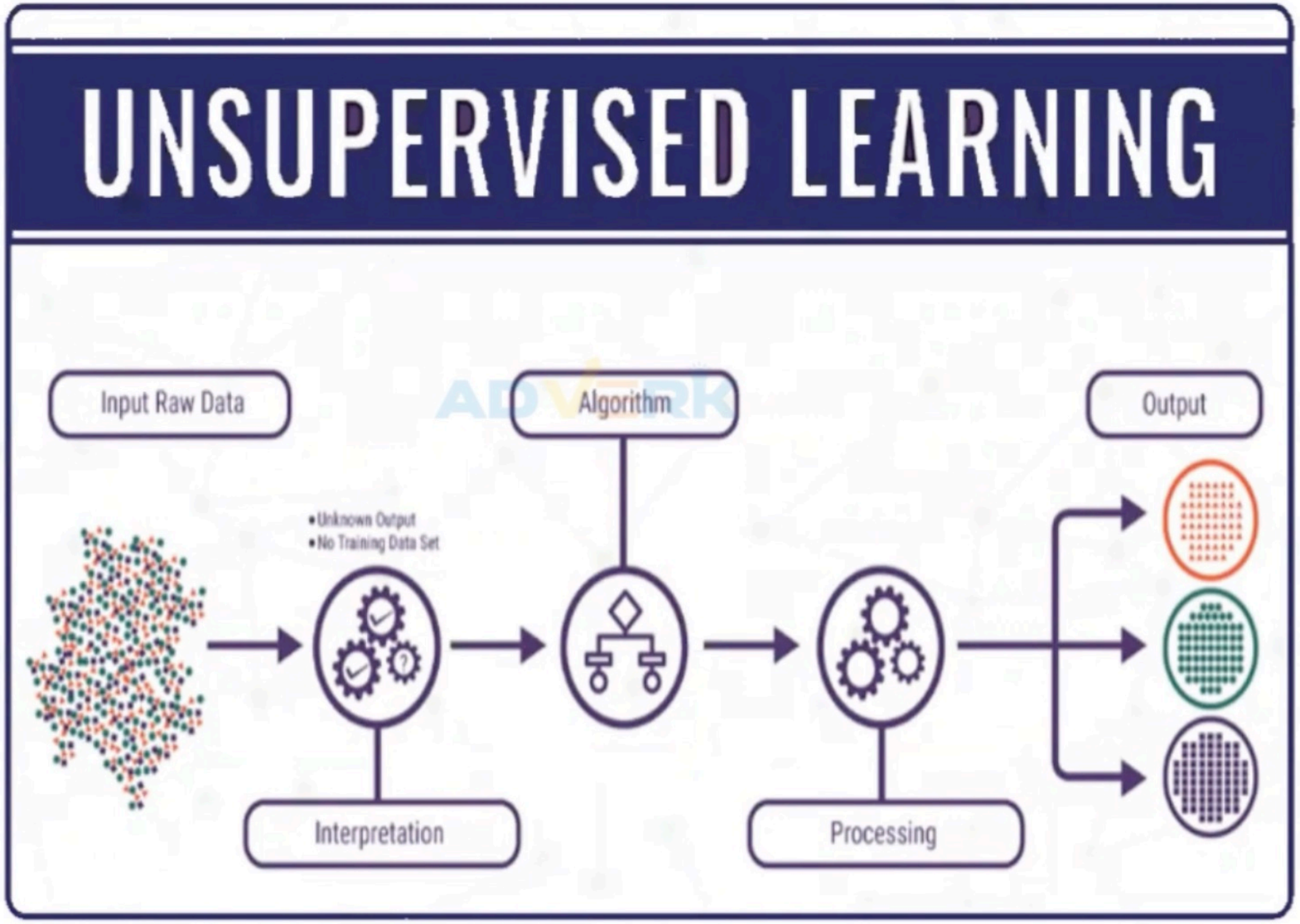
Image Segmentation - Supervised Learning algorithms are used in image segmentation. In this process, image classification is performed on different image data with pre-defined labels.

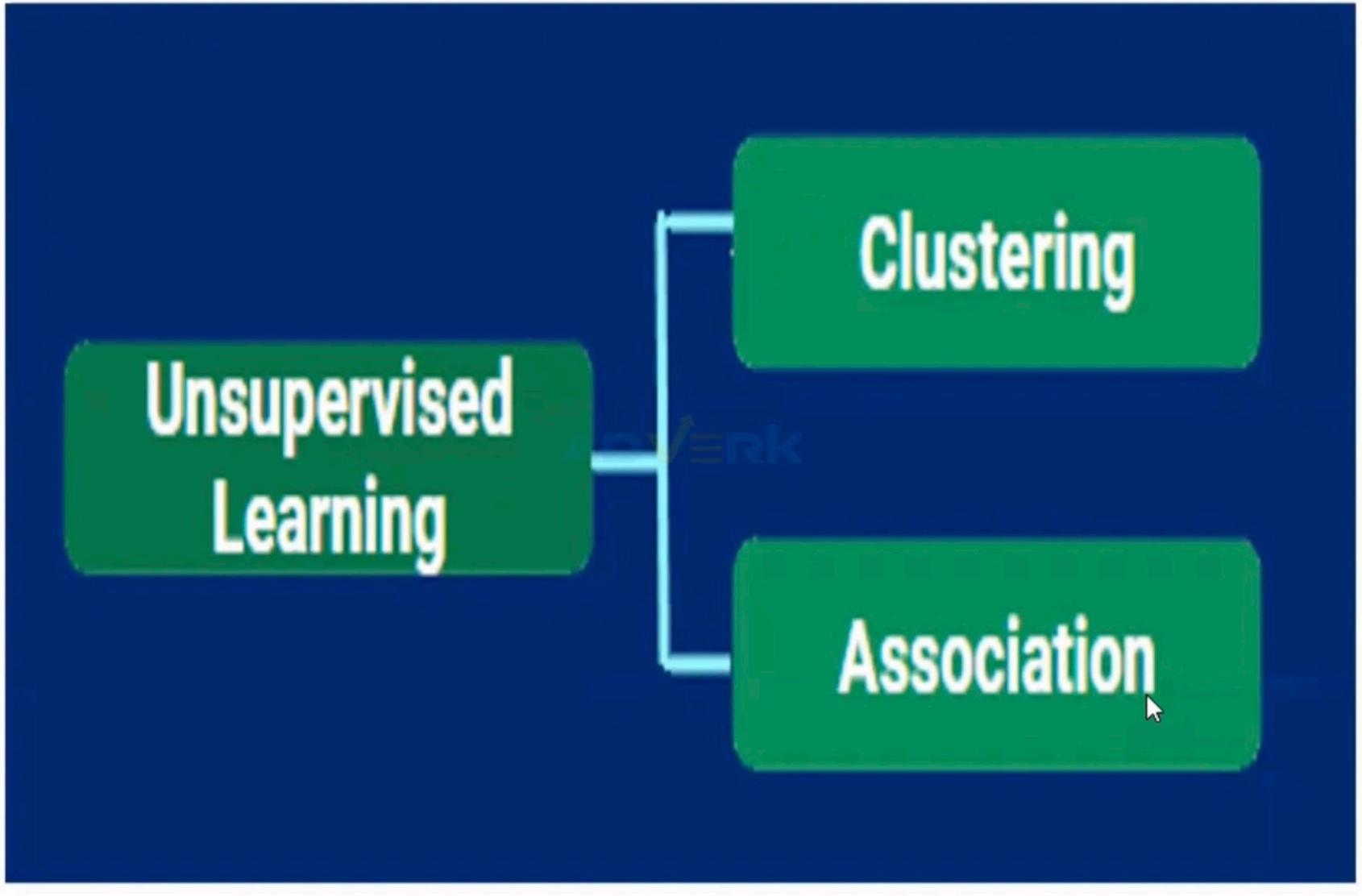
Medical Diagnosis - Supervised algorithms are also used in the medical field for diagnosis purposes. It is done by using medical images and past labelled data with labels for disease conditions. With such a process, the machine can identify a disease for the new patients.

Fraud Detection - Supervised Learning classification algorithms are used for identifying fraud transactions, fraud customers, etc. It is done by using historic data to identify the patterns that can lead to possible fraud.

Spam detection - In spam detection & filtering, classification algorithms are used. These algorithms classify an email as spam or not spam. The spam emails are sent to the spam folder.

Speech Recognition - Supervised learning algorithms are also used in speech recognition. The algorithm is trained with voice data, and various identifications can be done using the same, such as voice-activated passwords, voice commands, etc.





Advantages & Disadvantages of Unsupervised Learning

ADVANTAGES

These algorithms can be used for complicated tasks compared to the supervised ones because these algorithms work on the unlabeled dataset.

Unsupervised algorithms are preferable for various tasks as getting the unlabeled dataset is easier as compared to the labelled dataset.

DISADVANTAGES

The output of an unsupervised algorithm can be less accurate as the dataset is not labelled, and algorithms are not trained with the exact output in prior.

Working with Unsupervised learning is more difficult as it works with the unlabelled dataset that does not map with the output.

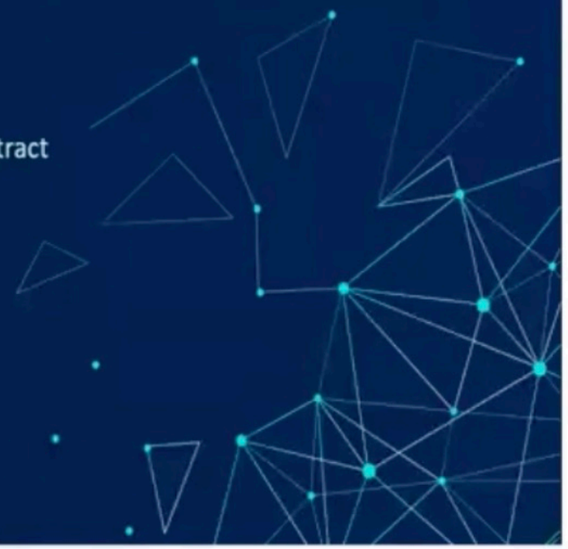
Applications of Unsupervised Learning

Network Analysis: Unsupervised learning is used for identifying plagiarism and copyright in document network analysis of text data for scholarly articles.

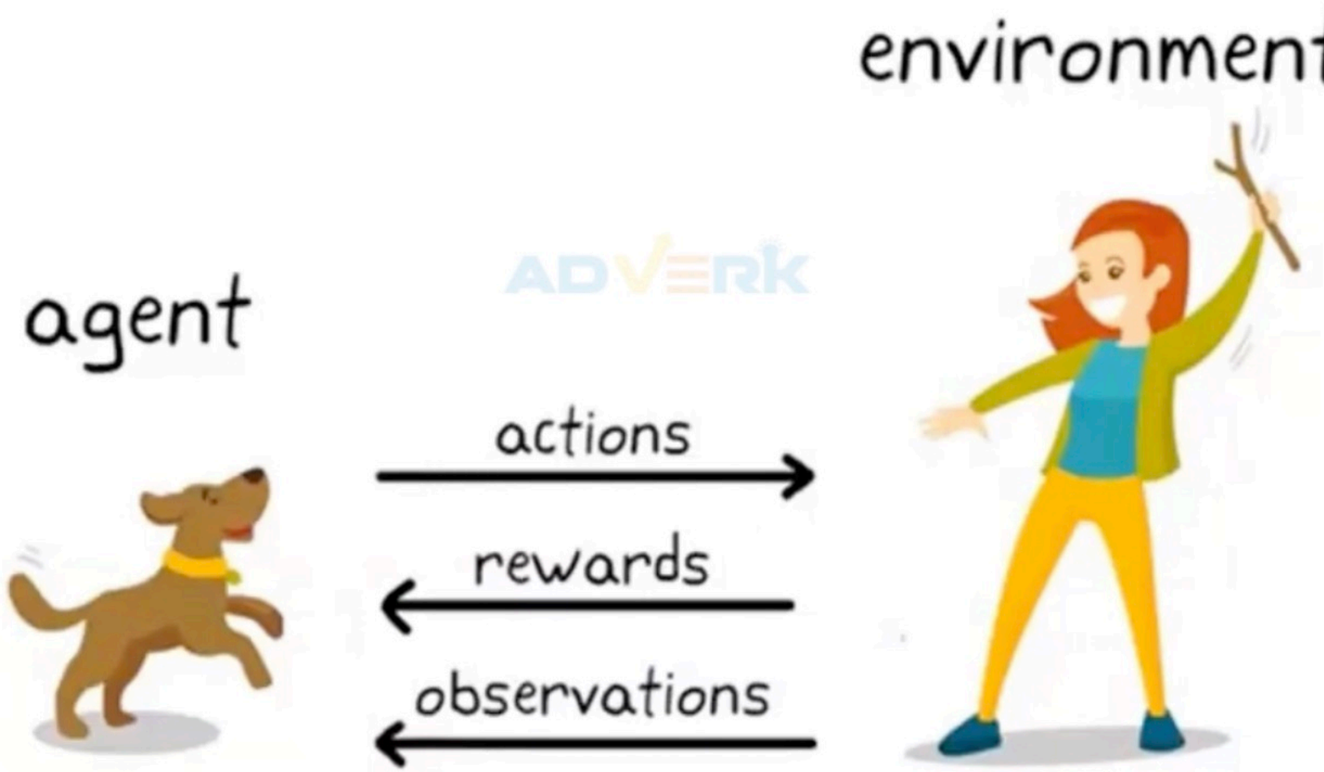
Recommendation Systems: Recommendation systems widely use unsupervised learning techniques for building recommendation applications for different web applications and e-commerce websites.

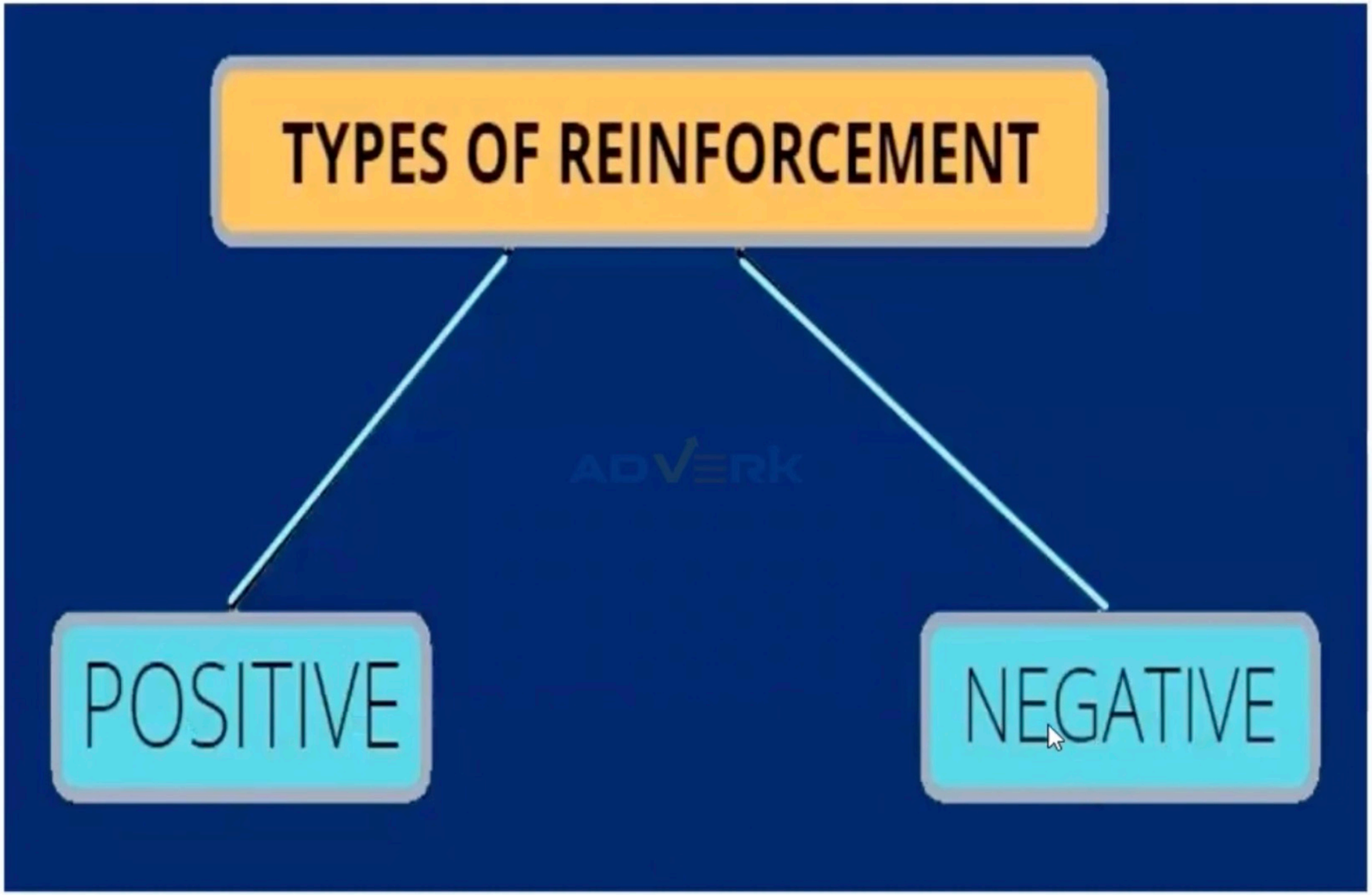
Anomaly Detection: Anomaly detection is a popular application of unsupervised learning, which can identify unusual data points within the dataset. It is used to discover fraudulent transactions.

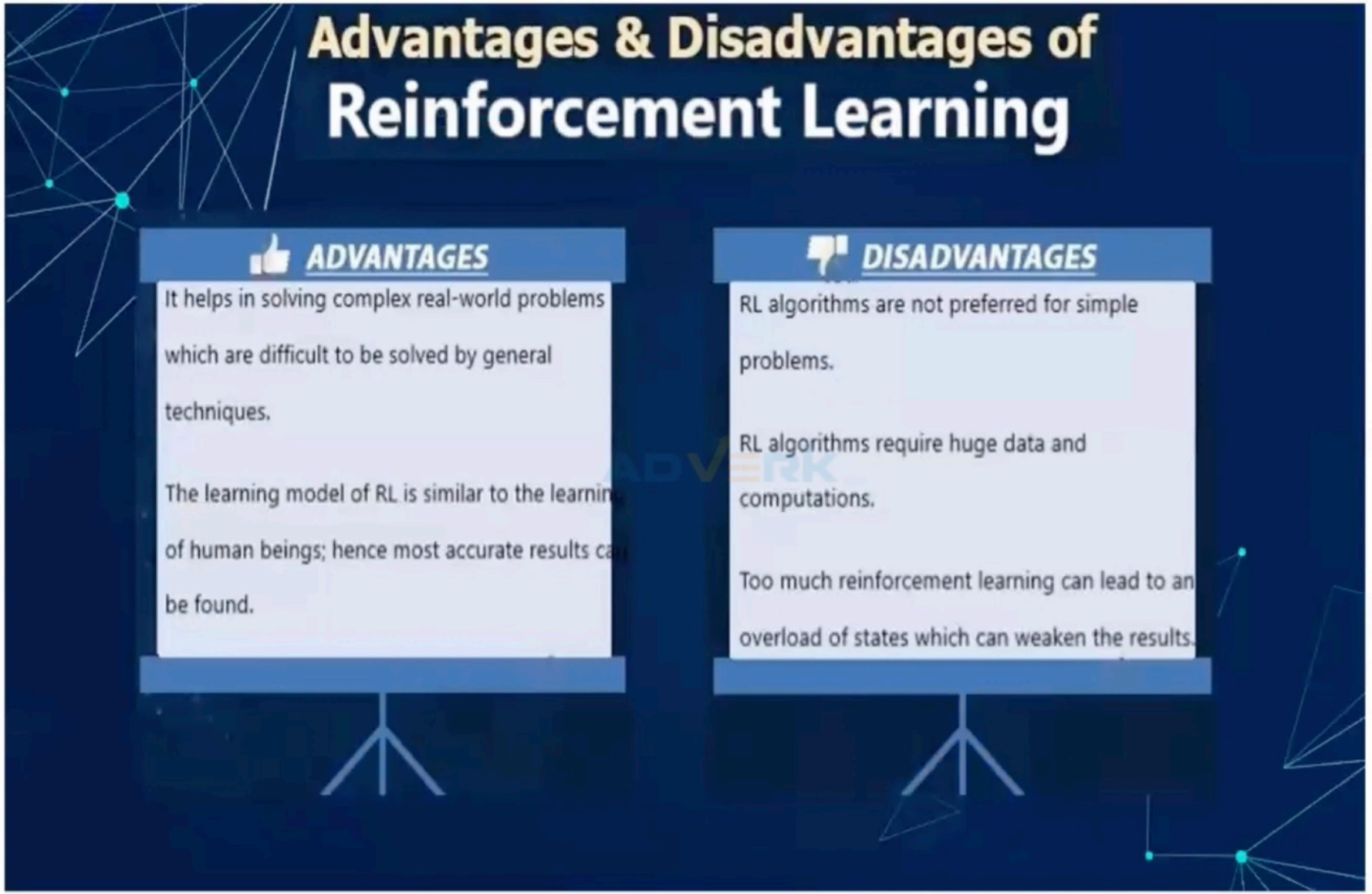
Singular Value Decomposition: Singular Value Decomposition or SVD is used to extract particular information from the database. For example, extracting information of each user located at a particular location.



REINFORCEMENT LEARNING







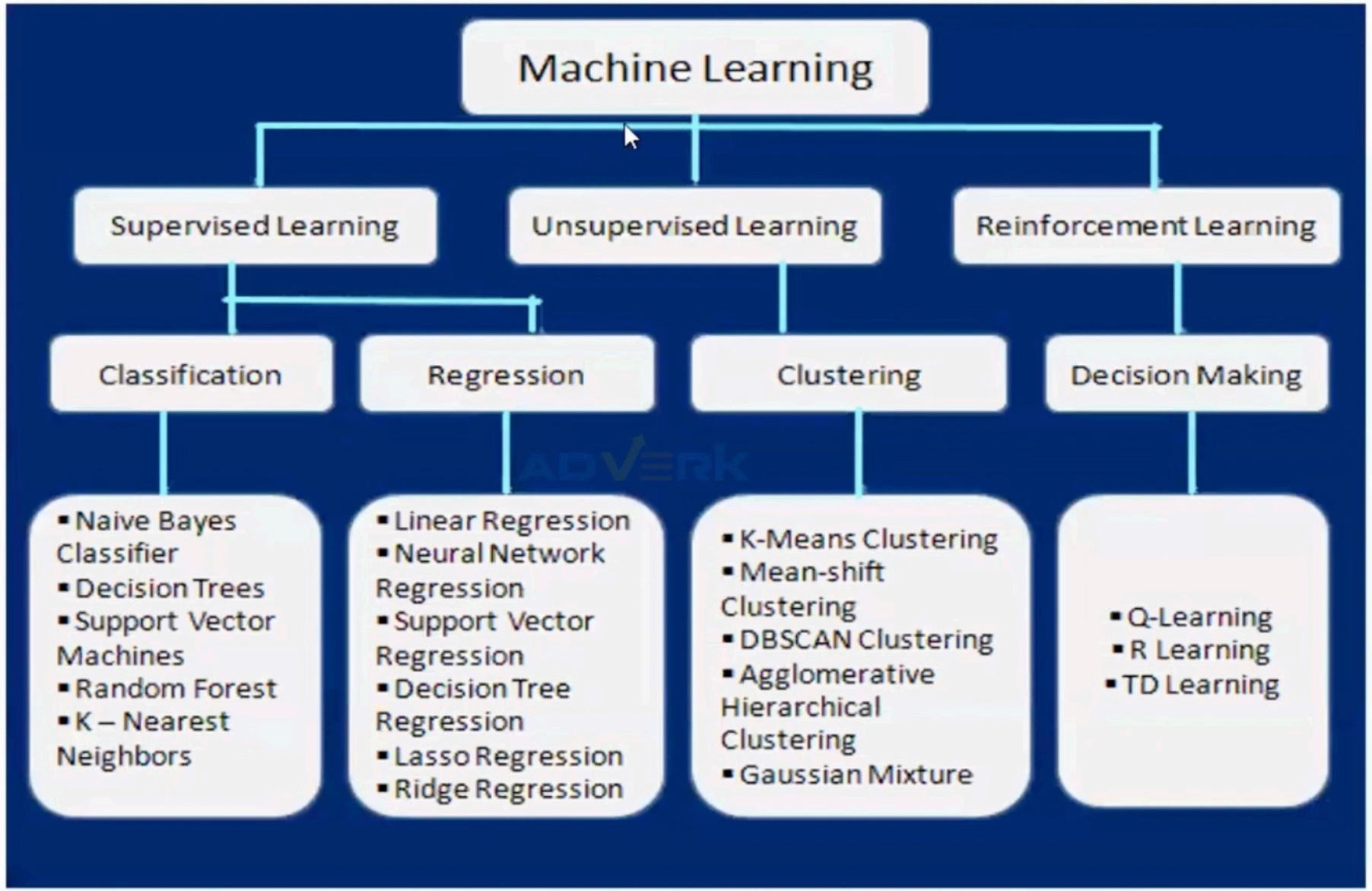
Applications of Reinforcement Learning

Video Games: RL algorithms are much popular in gaming applications. It is used to gain super-human performance. Some popular games that use RL algorithms are **AlphaGO** and **AlphaGO Zero**.

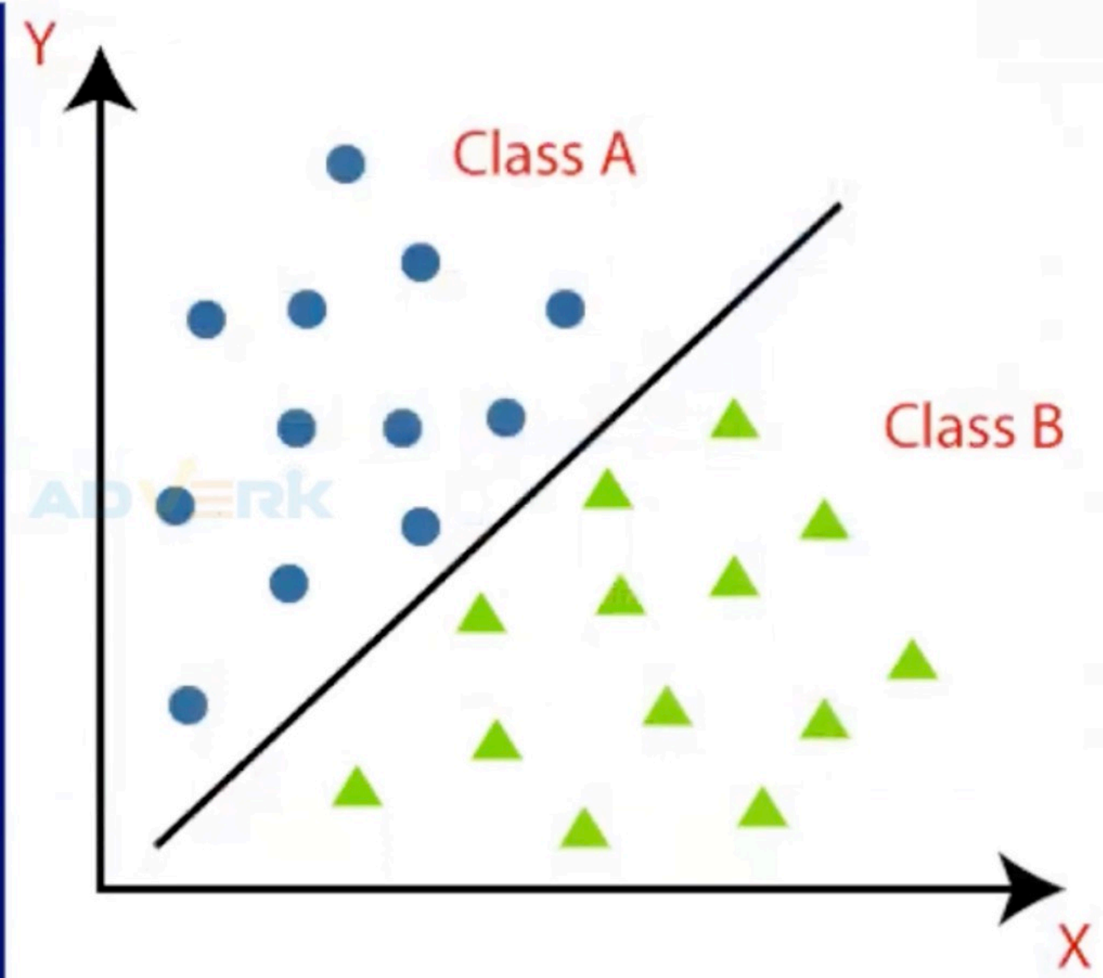
Resource Management: The "Resource Management with Deep Reinforcement Learning" paper showed that how to use RL in computer to automatically learn and schedule resources to wait for different jobs in order to minimize average job slowdown.

Robotics: RL is widely being used in Robotics applications. Robots are used in the industrial and manufacturing area, and these robots are made more powerful with reinforcement learning. There are different industries that have their vision of building intelligent robots using AI and Machine learning technology.

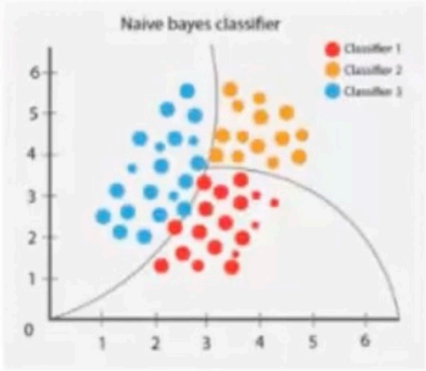
Text Mining: Text-mining, one of the great applications of NLP, is now being implemented with the help of Reinforcement Learning by Salesforce company.

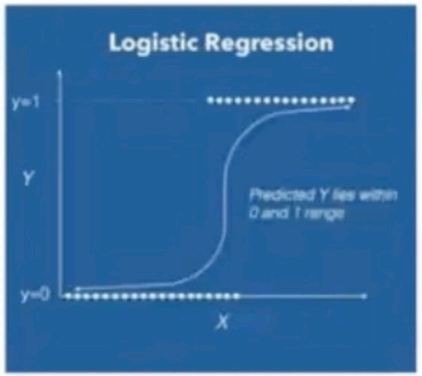


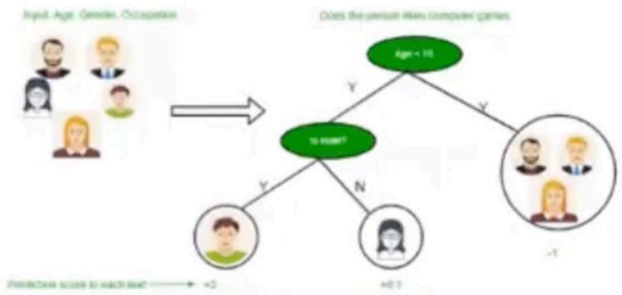
Classification



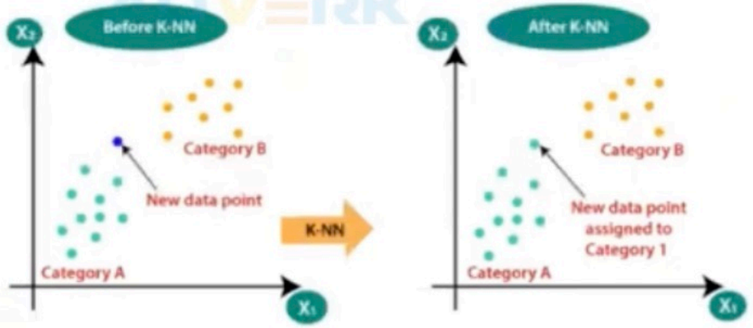
Classification Algorithms



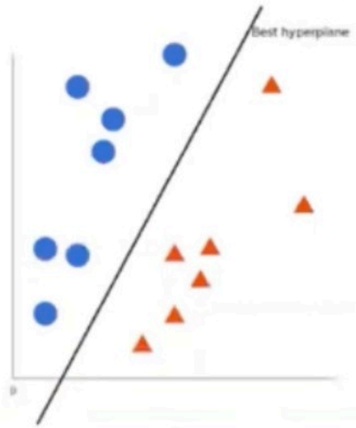




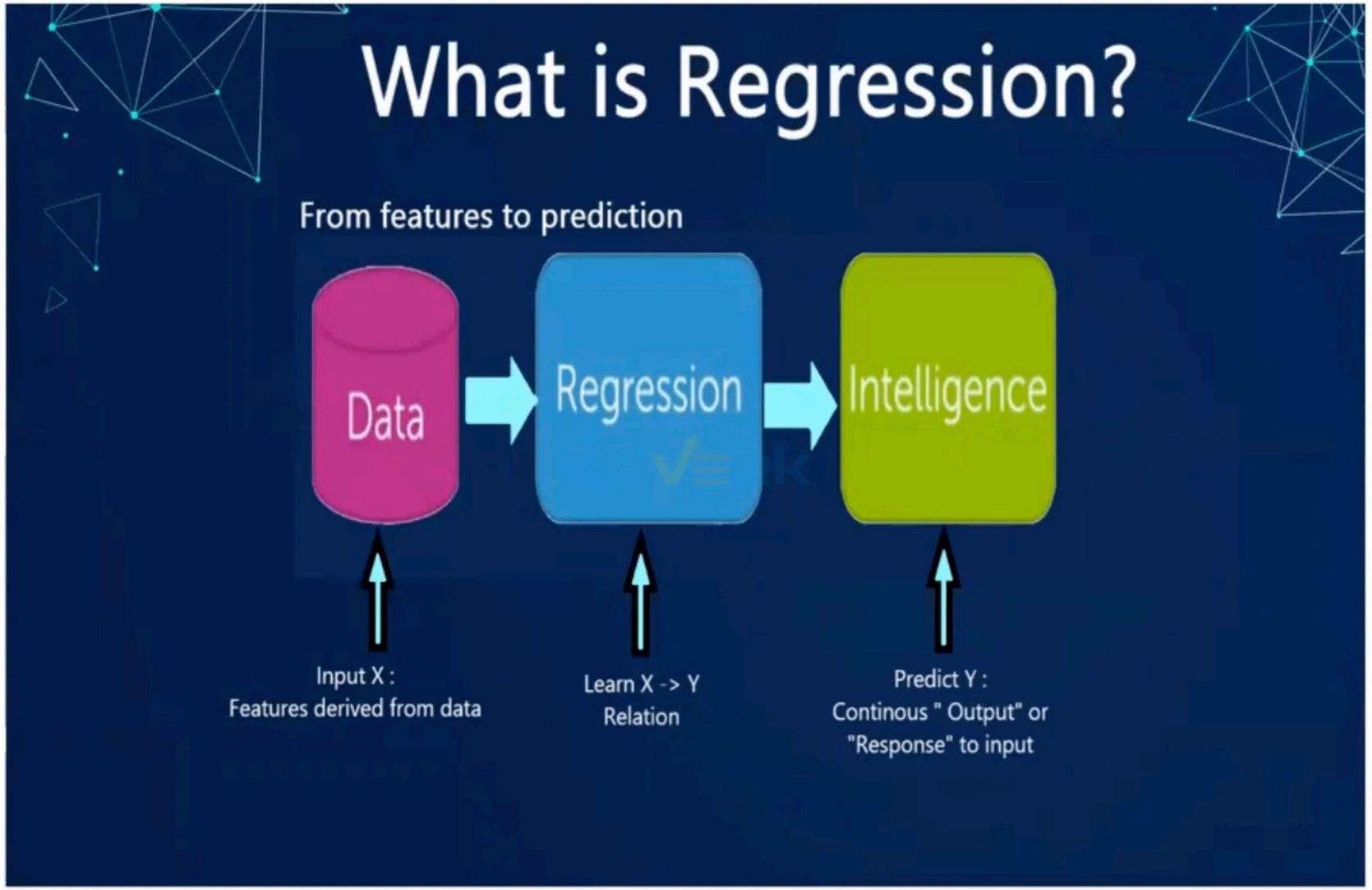
Decision Tree

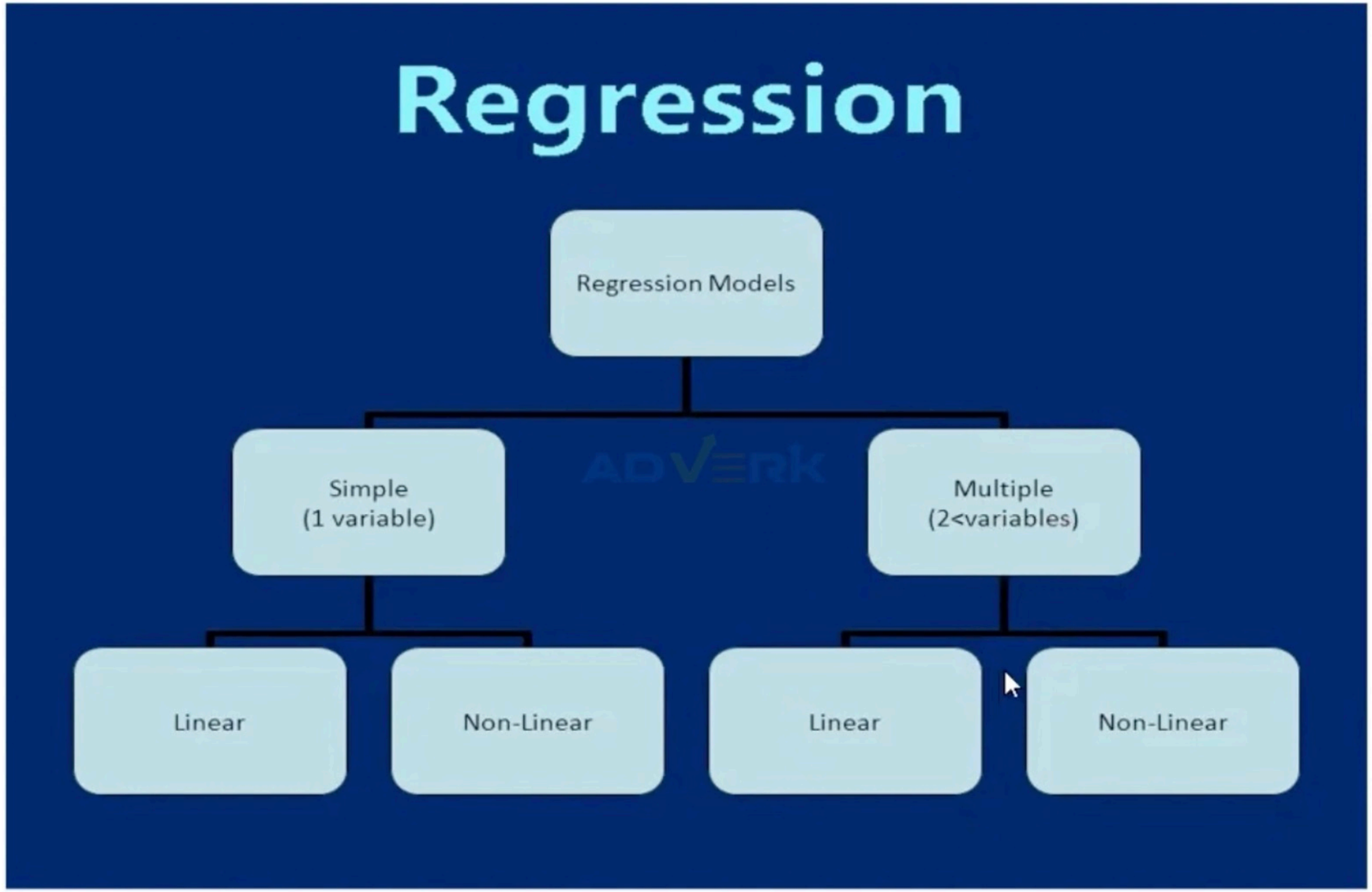


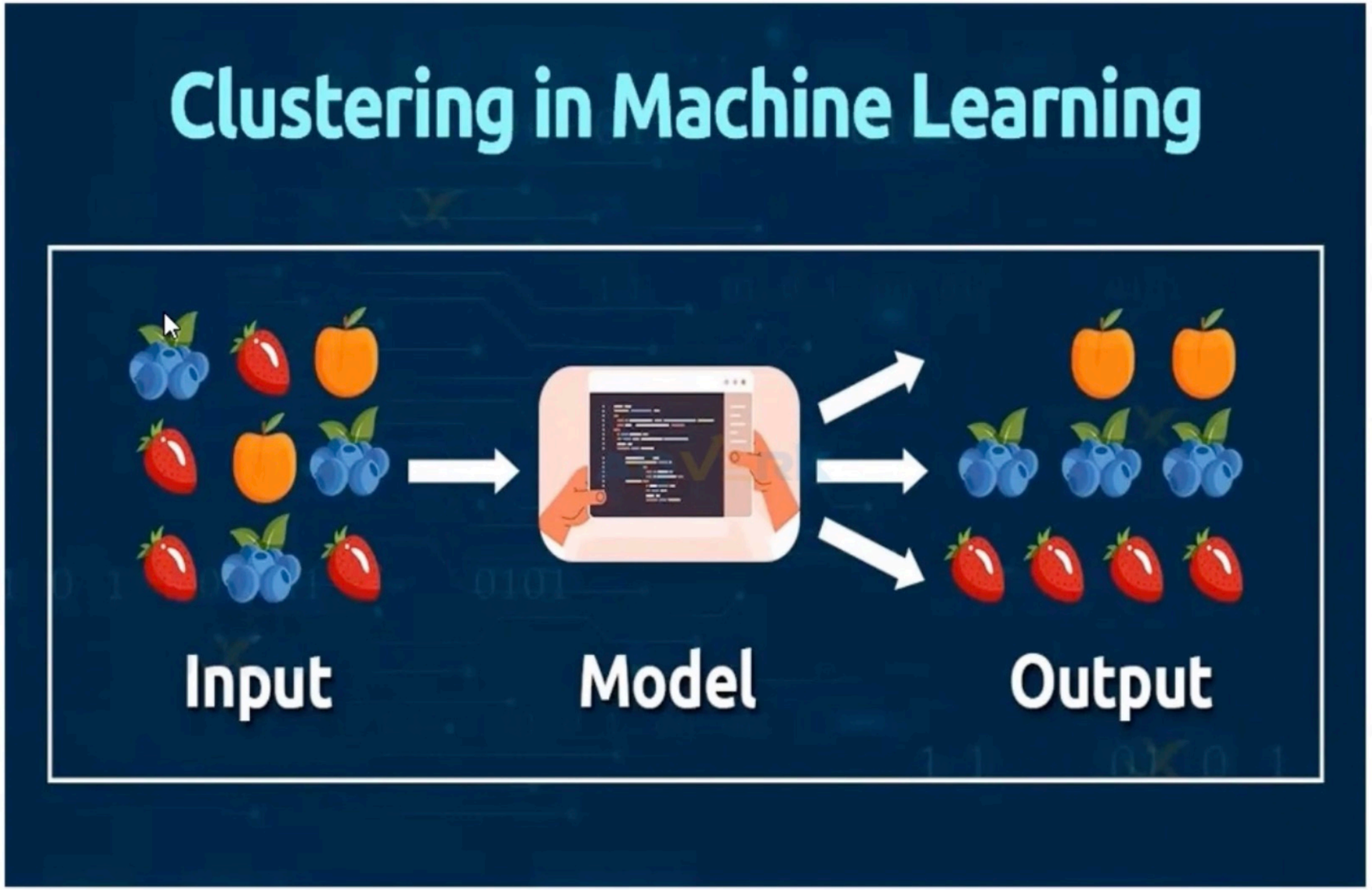
K-Nearest Neighbors

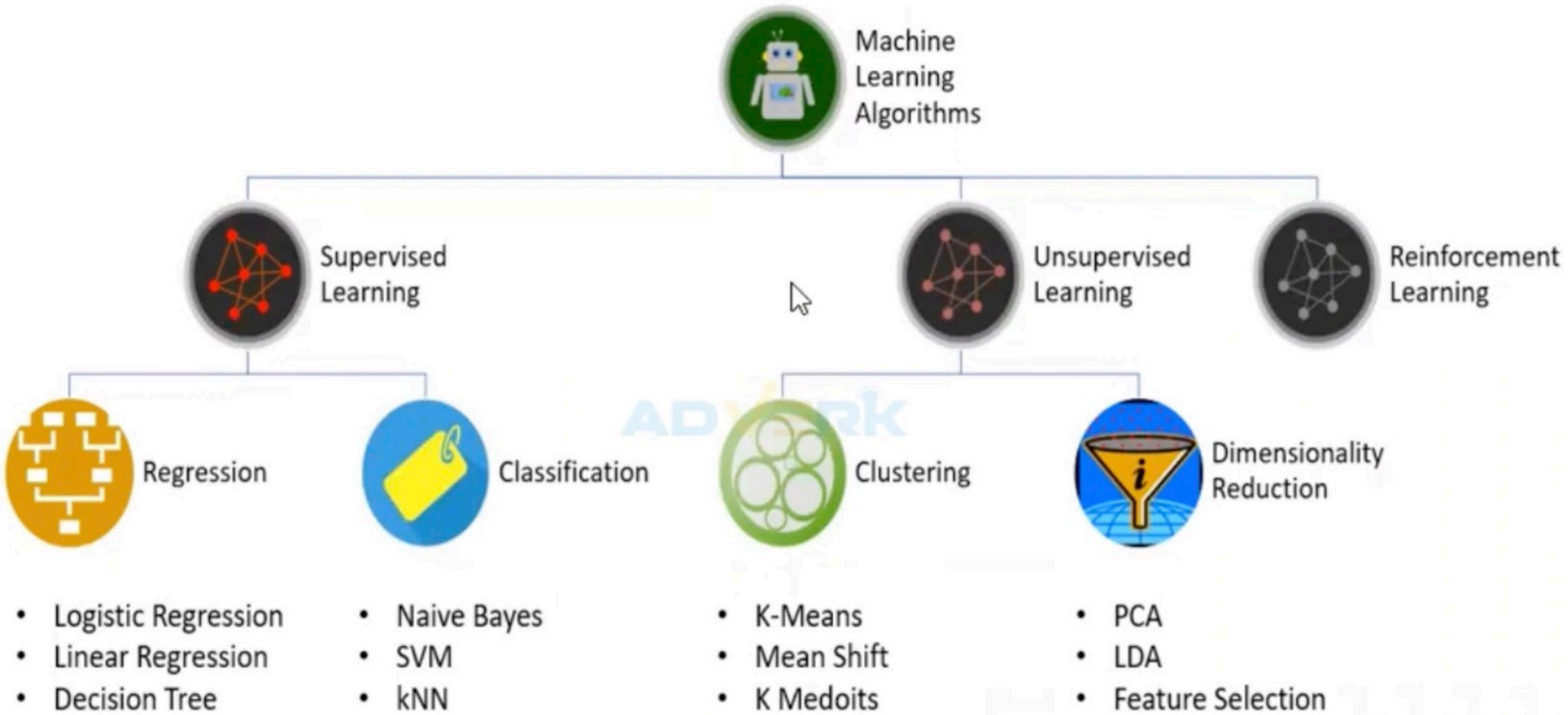


Support Vector Machines

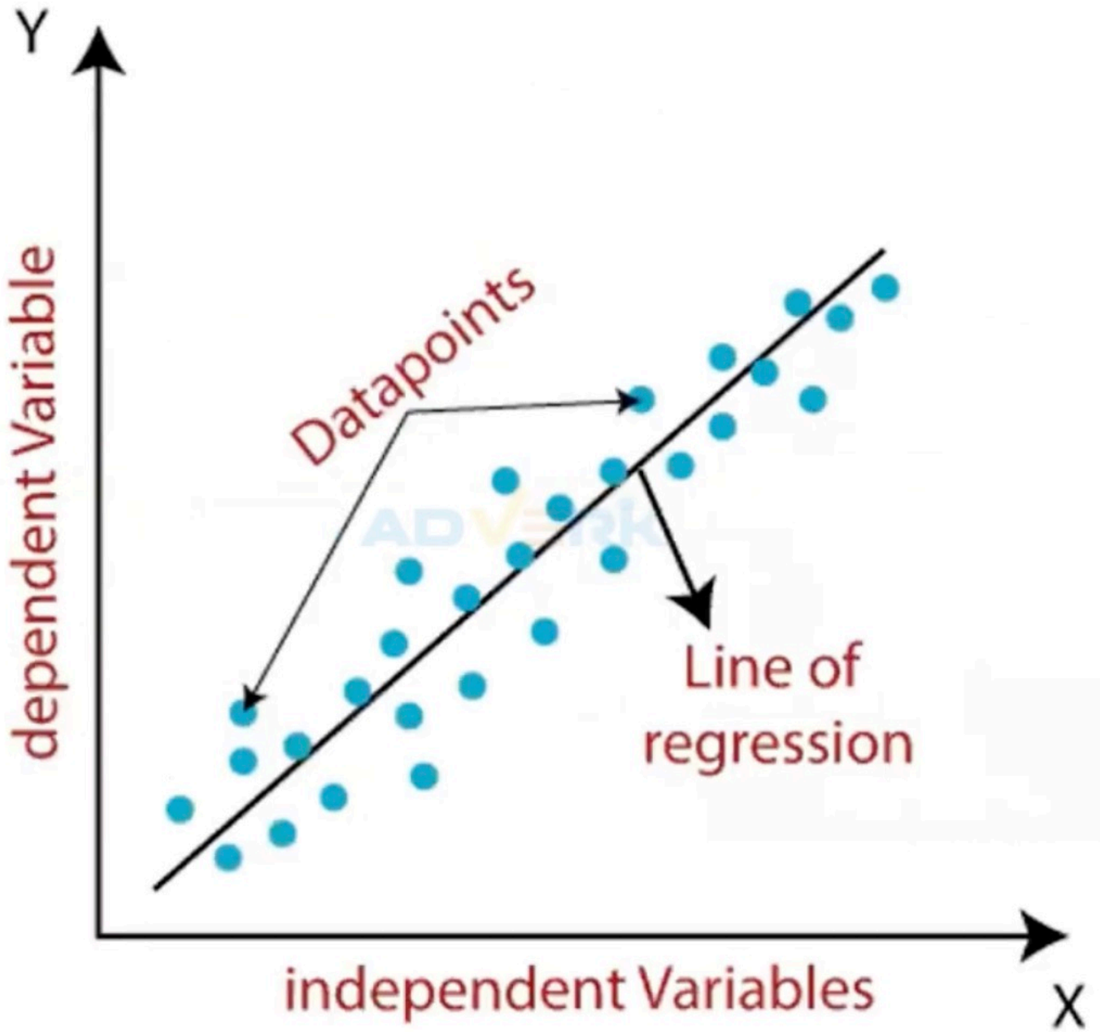




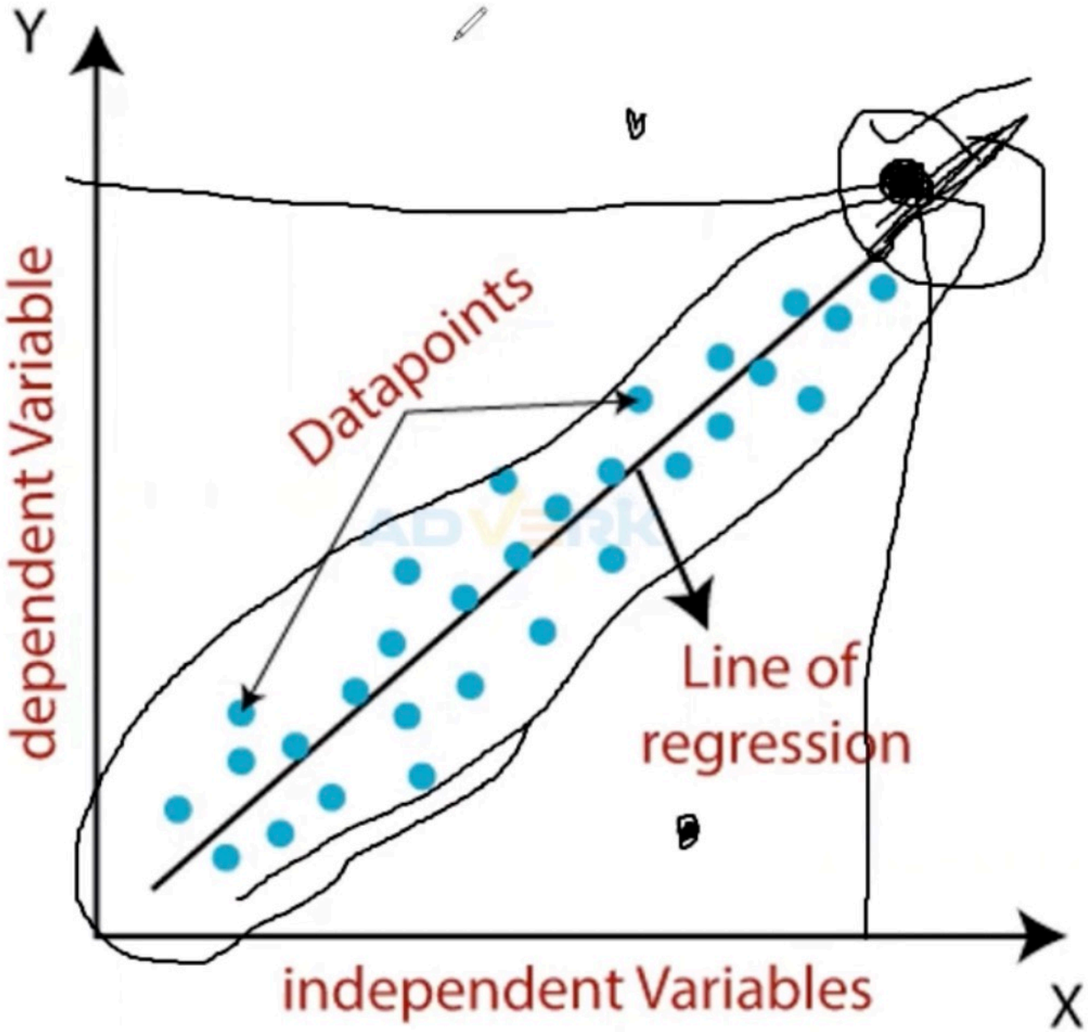




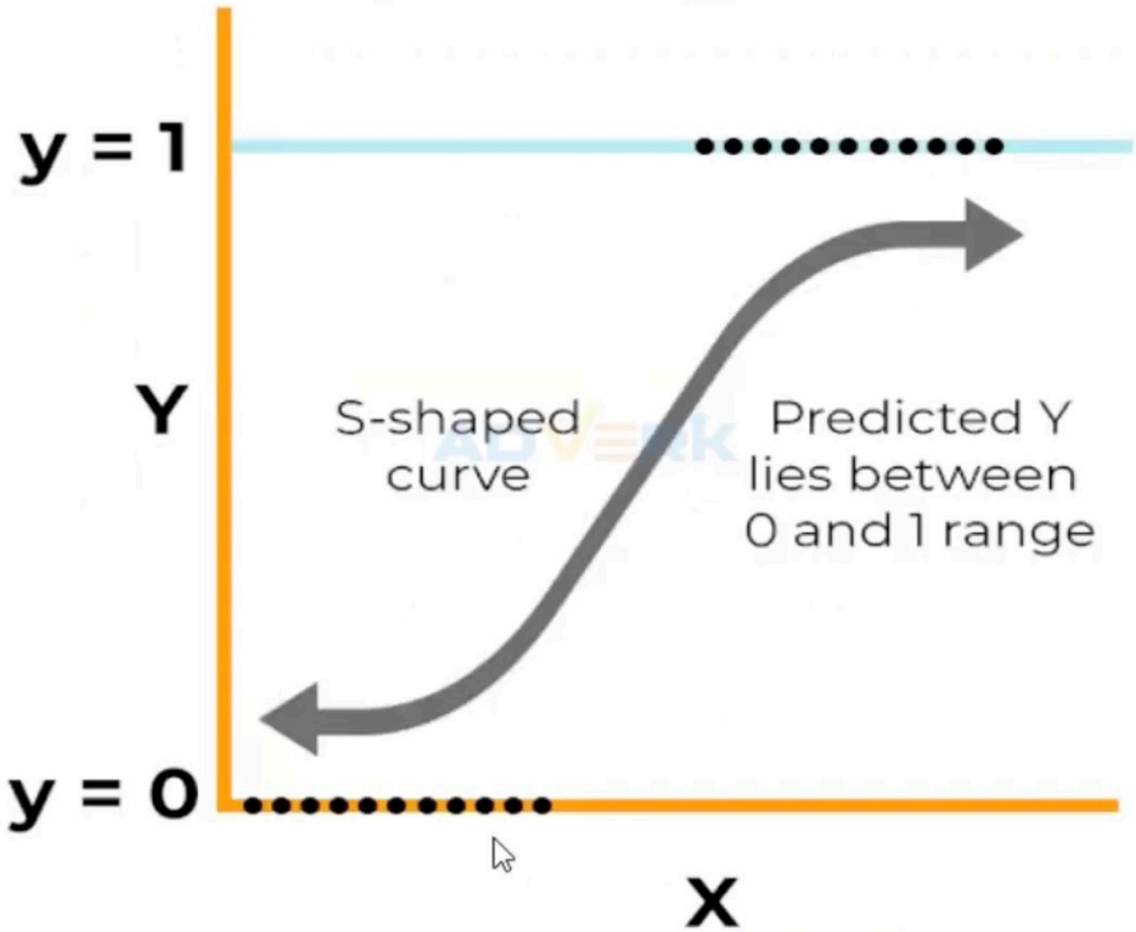
Linear Regression

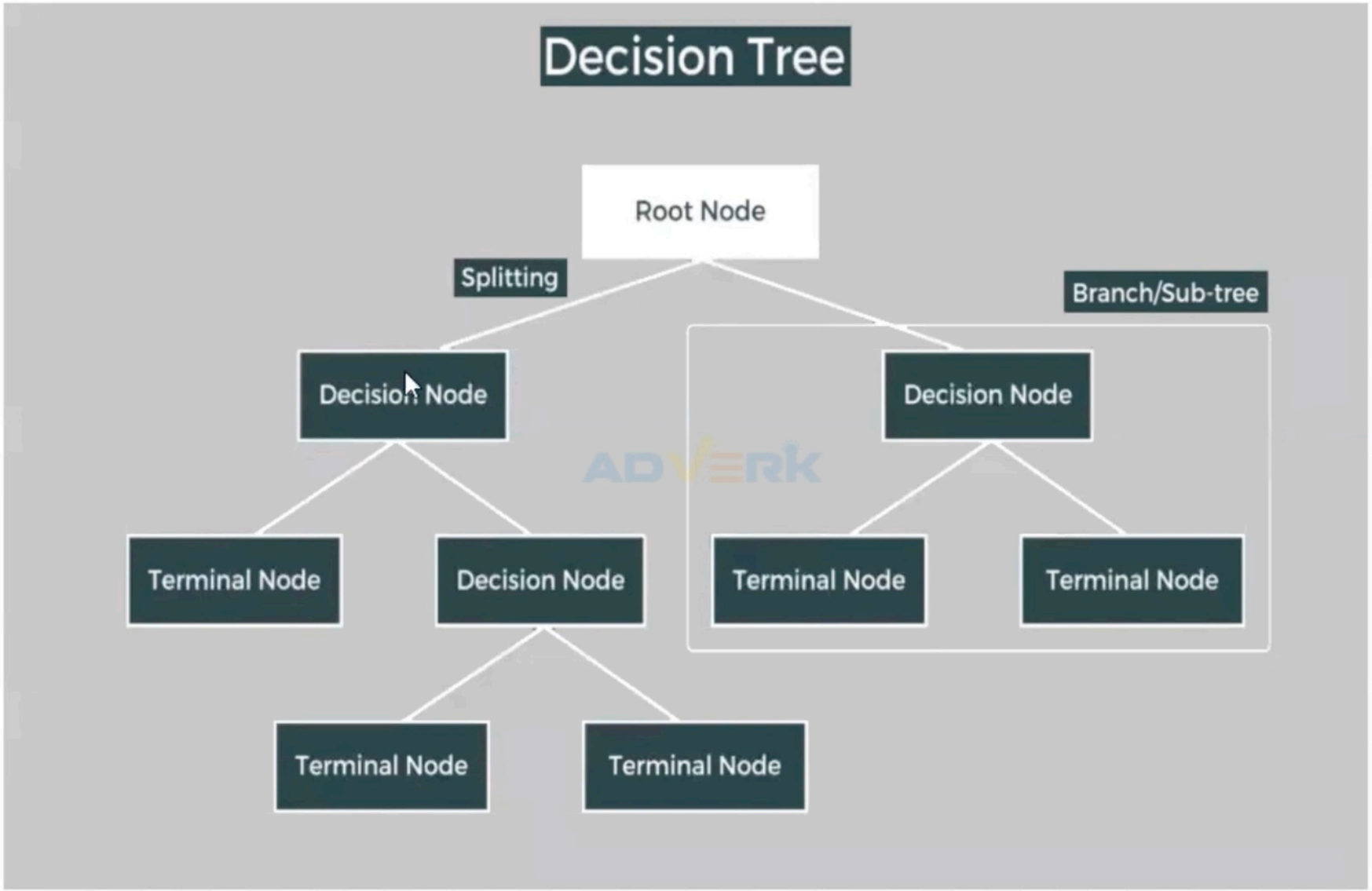


Linear Regression

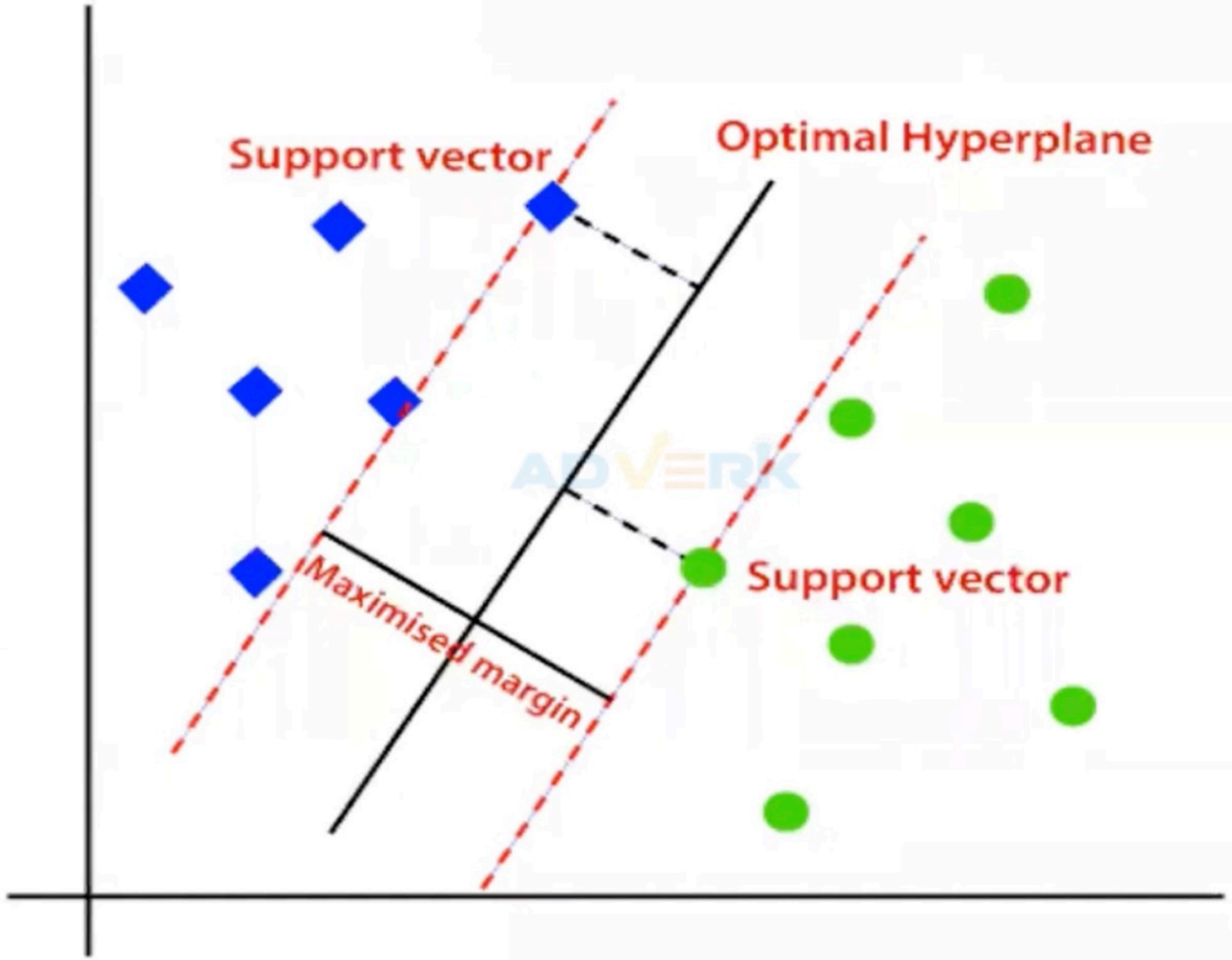


Logistic Regression

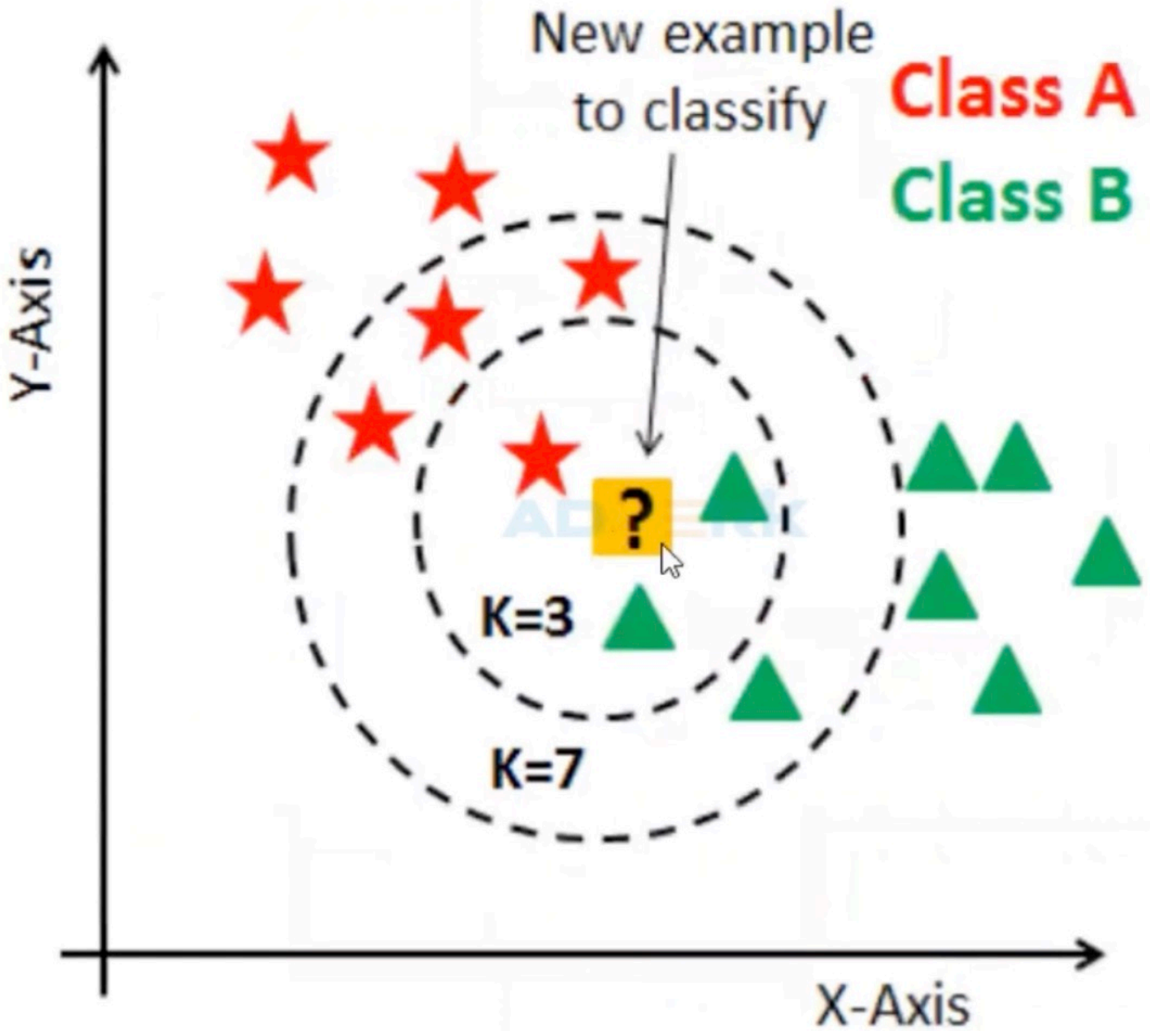




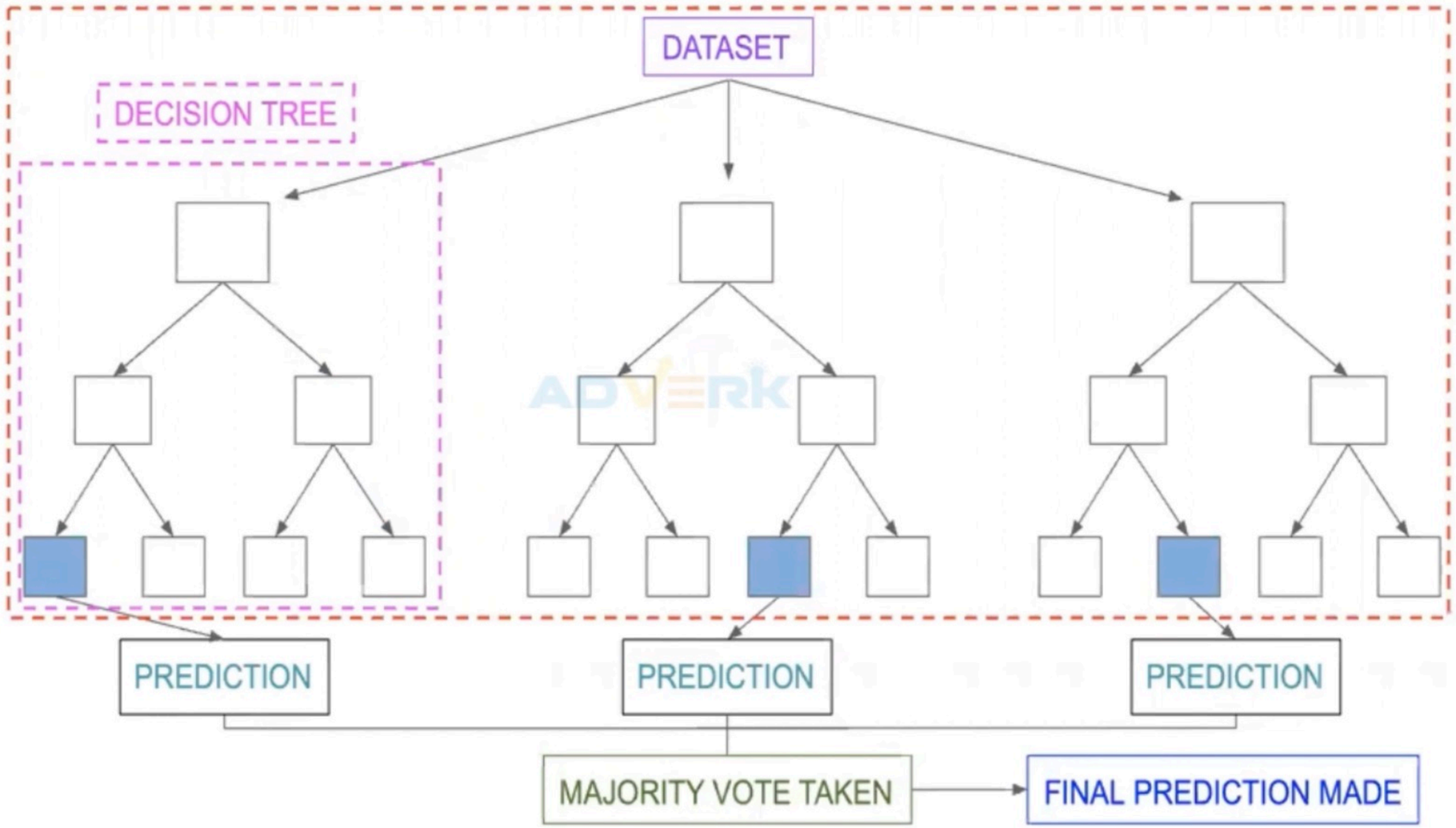
Support Vector Machine



K-Nearest Neighbors



Random Forest



Gradient Boosting