Introduction to Data Analytics

Dealing with Different Types of Data

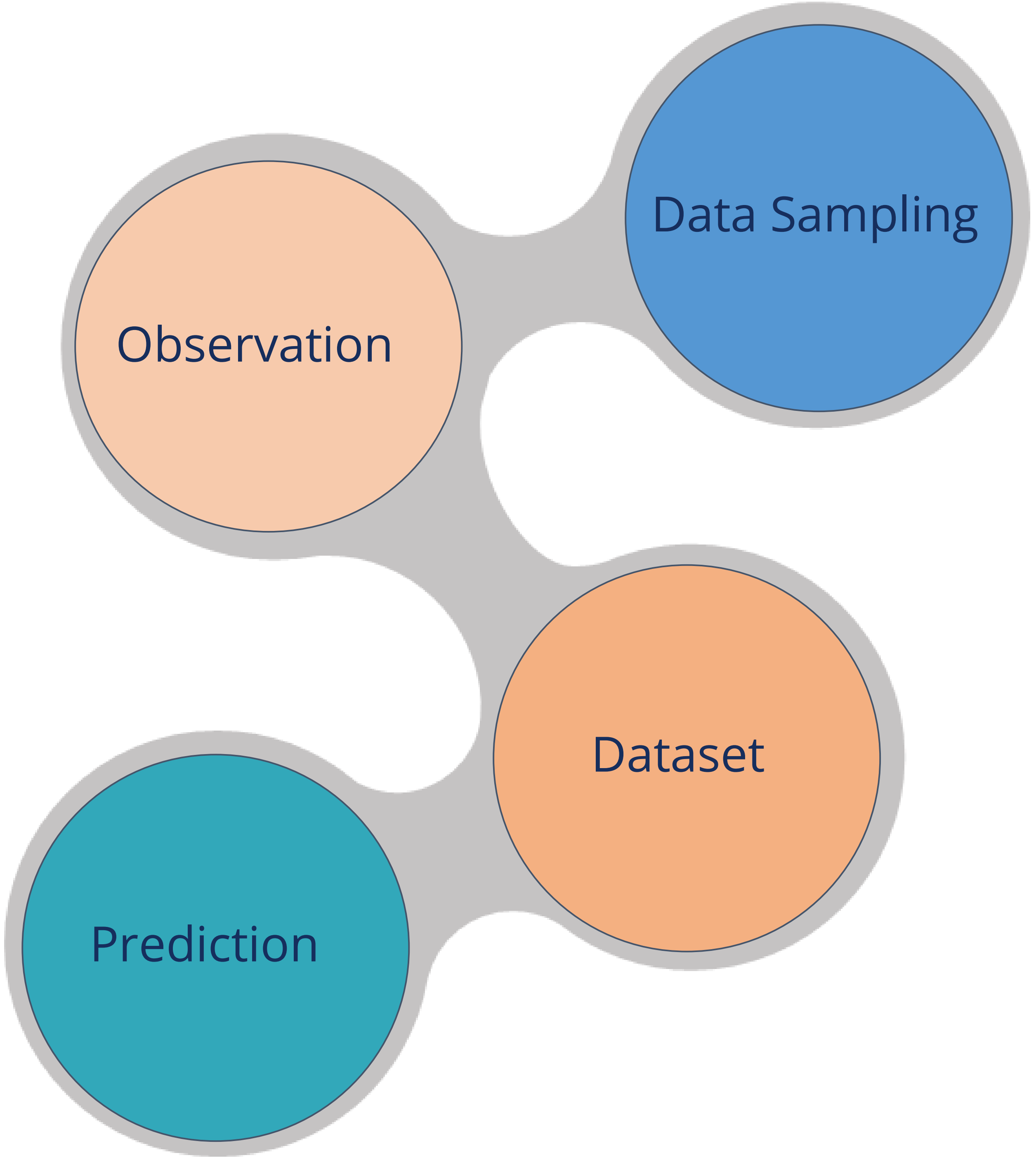
Learning Objectives

By the end of this lesson, you will be able to:

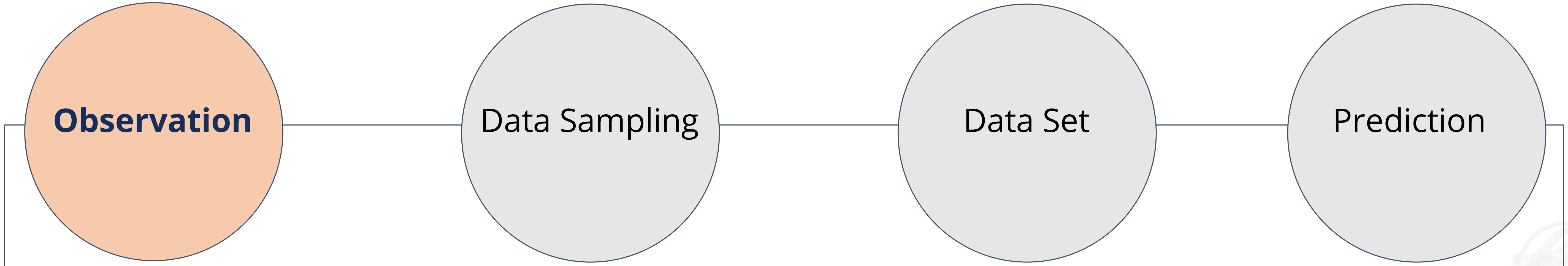
- List the terminologies used in data analytics
- Describe the types of data
- Explain the levels of measurement

Terminologies in Data Analytics

Terminologies in Data Analytics



Terminologies in Data Analytics



Observation

Data Sampling

Data Set

Prediction

- Observation is a single row or a record of data from the database.
- Any data can be assumed as a set of observations.

Terminologies in Data Analytics

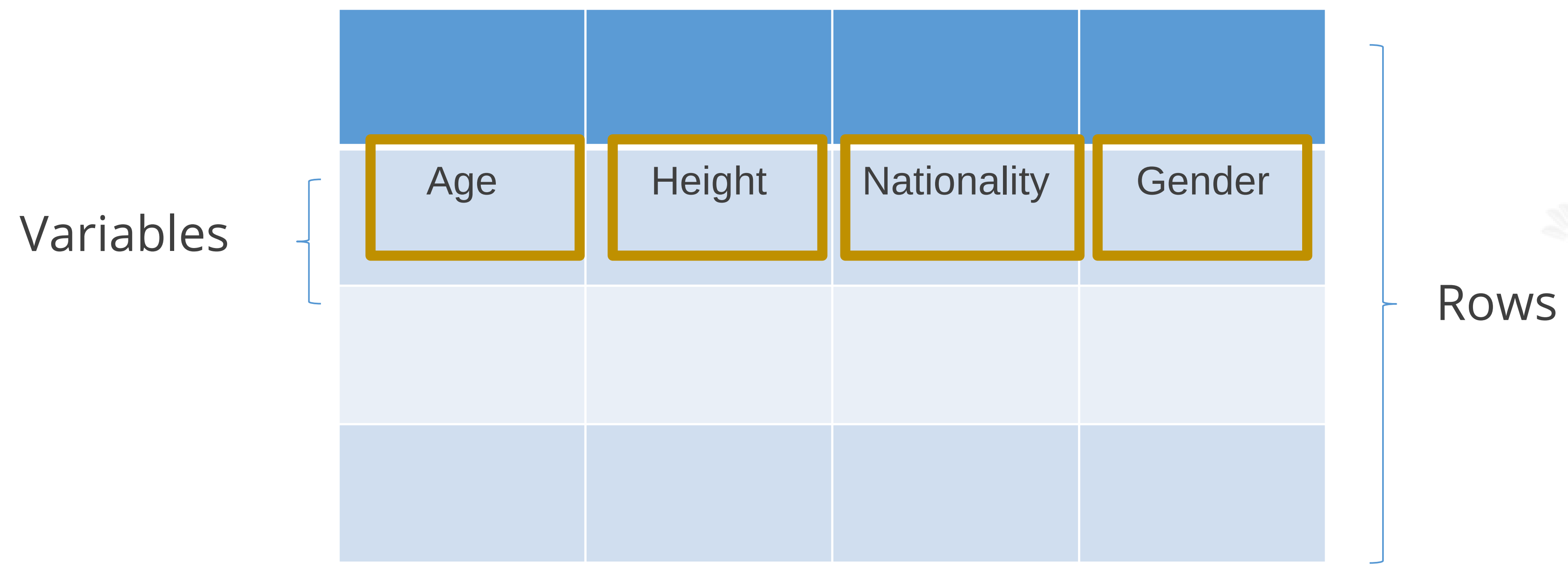
Observation

Data Sampling

Data Set

Prediction

Database Table



Age	Height	Nationality	Gender

Observation is the unit of analysis on which the measurements are taken.

It is also known as a case, record, pattern, or row.

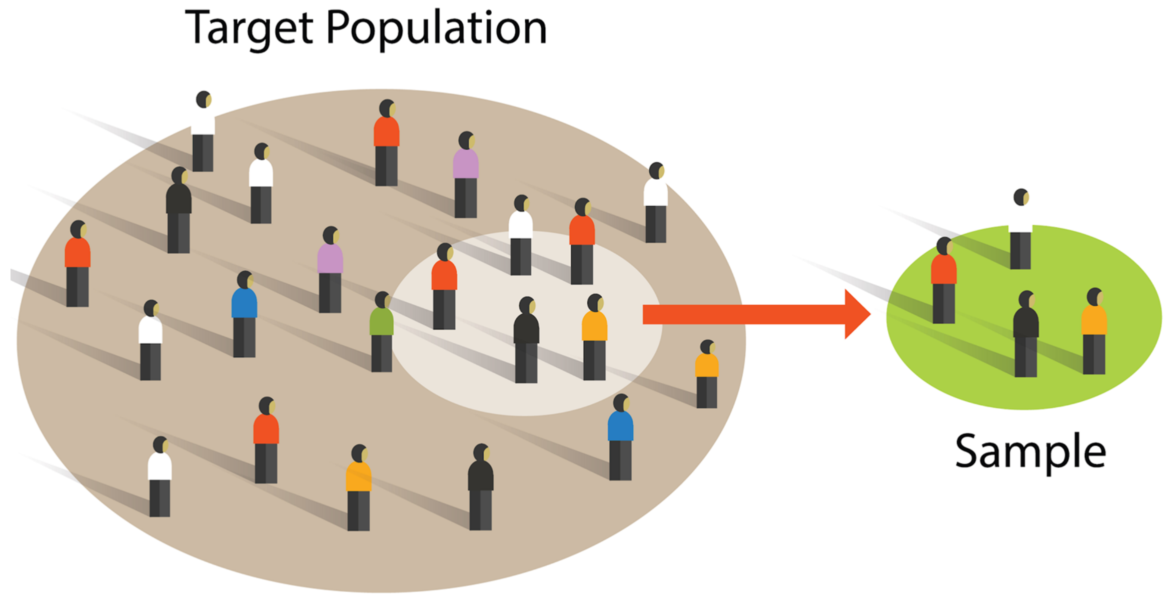
Terminologies in Data Analytics

Observation

Data Sampling

Data Set

Prediction



- Data sampling is a statistical analysis technique used to select, manipulate, and analyze a representative subset of data points.
- Data sampling identifies patterns and trends in the larger data set.

Terminologies in Data Analytics

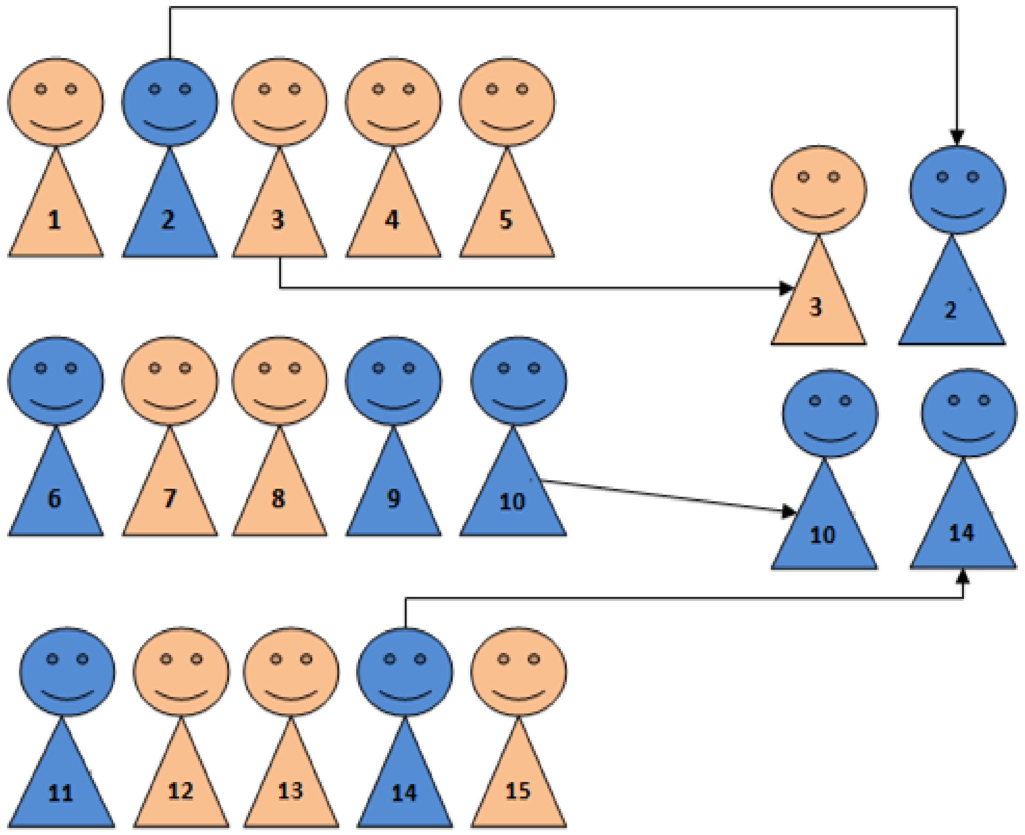
Observation — **Data Sampling** — Data Set — Prediction

- If a sample is randomly selected with 1 or *n* observations, then *n* is the sample size.

- The chart explains the sampling process where a few people are randomly sampled from a group of population.

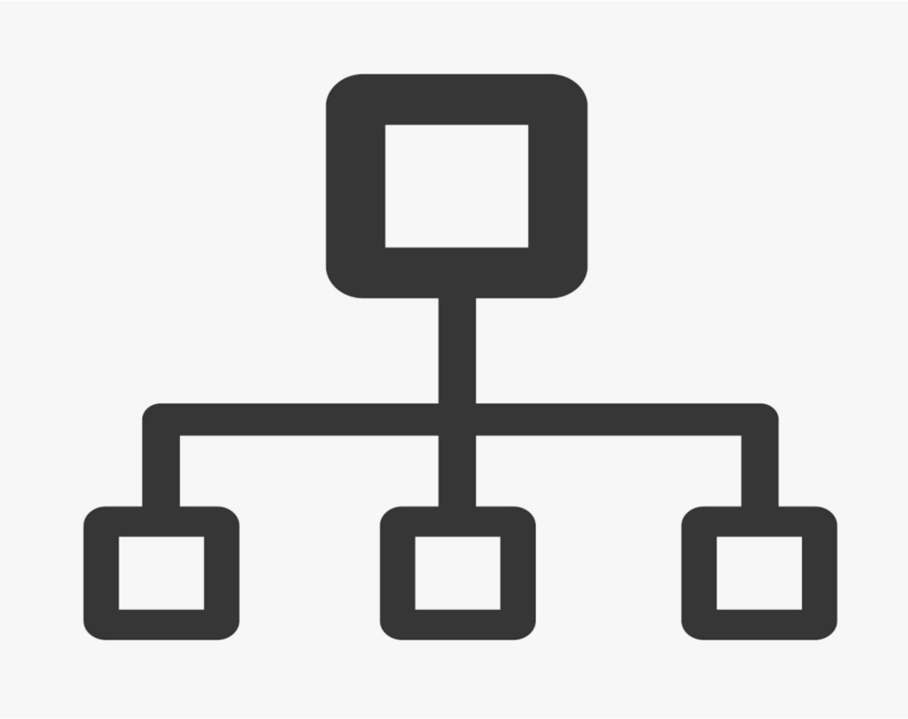
- Data sampling is cost effective and surveys only the representative sample.

- It enables data scientists, predictive modelers, and data analysts to produce accurate findings.



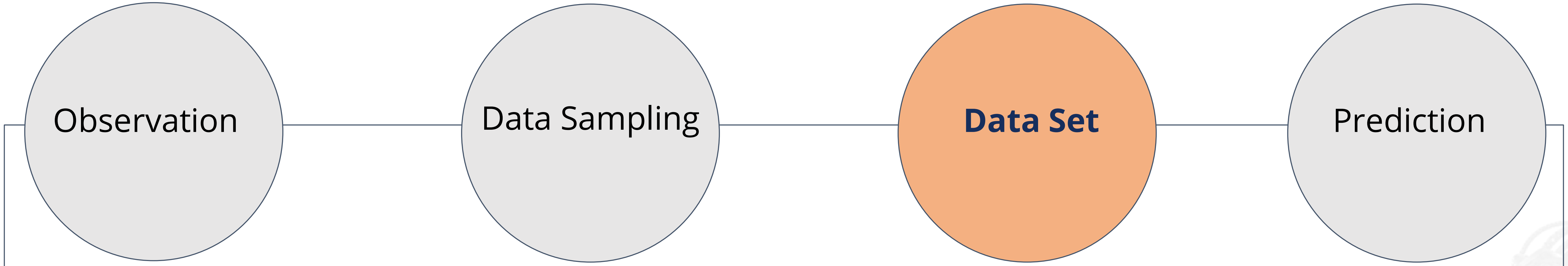
Terminologies in Data Analytics

Observation | Data Sampling | **Data Set** | Prediction

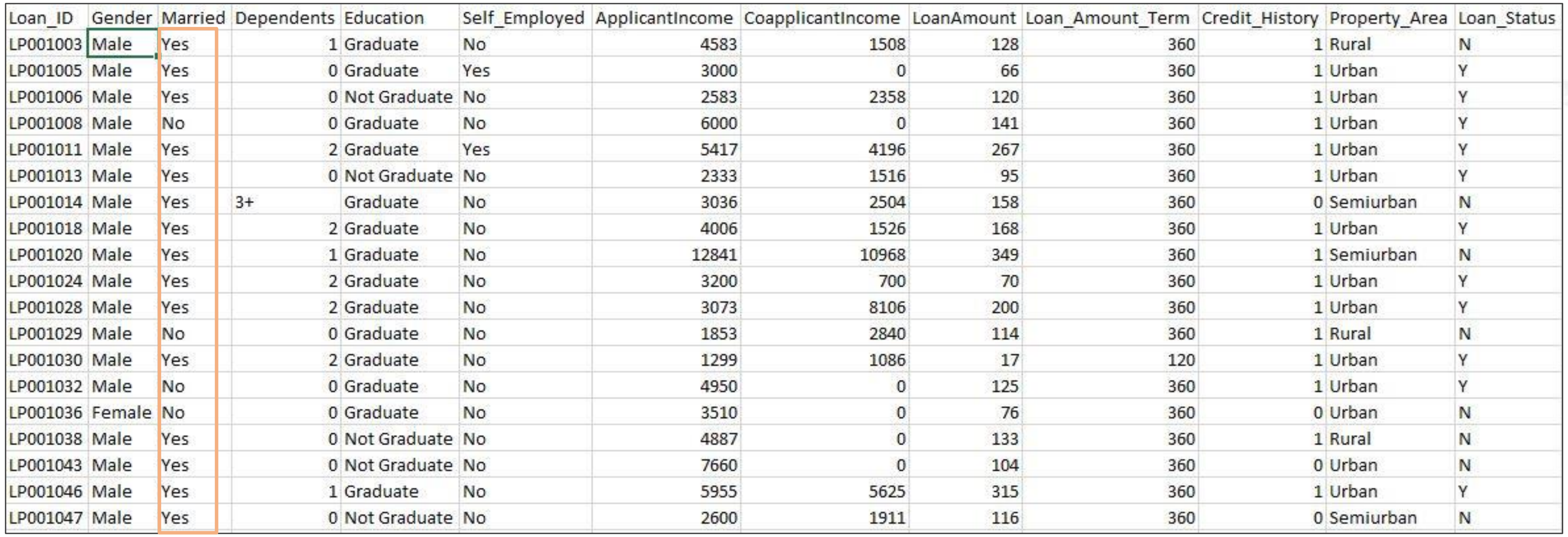


- Data set is a collection of data or the total data captured about a particular use case.
- It can hold information such as medical, insurance, and loan approval records.
- It is not limited to numbers and texts and may include collections of images or videos.

Terminologies in Data Analytics



The table represents loan data with attributes such as loan ID, borrower's gender, education, employment status, credit history, loan amount, and property details.



Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	N
LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	Y
LP001006	Male	Yes	0	Not Graduate	No	2583	2358	120	360	1	Urban	Y
LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	Y
LP001011	Male	Yes	2	Graduate	Yes	5417	4196	267	360	1	Urban	Y
LP001013	Male	Yes	0	Not Graduate	No	2333	1516	95	360	1	Urban	Y
LP001014	Male	Yes	3+	Graduate	No	3036	2504	158	360	0	Semiurban	N
LP001018	Male	Yes	2	Graduate	No	4006	1526	168	360	1	Urban	Y
LP001020	Male	Yes	1	Graduate	No	12841	10968	349	360	1	Semiurban	N
LP001024	Male	Yes	2	Graduate	No	3200	700	70	360	1	Urban	Y
LP001028	Male	Yes	2	Graduate	No	3073	8106	200	360	1	Urban	Y
LP001029	Male	No	0	Graduate	No	1853	2840	114	360	1	Rural	N
LP001030	Male	Yes	2	Graduate	No	1299	1086	17	120	1	Urban	Y
LP001032	Male	No	0	Graduate	No	4950	0	125	360	1	Urban	Y
LP001036	Female	No	0	Graduate	No	3510	0	76	360	0	Urban	N
LP001038	Male	Yes	0	Not Graduate	No	4887	0	133	360	1	Rural	N
LP001043	Male	Yes	0	Not Graduate	No	7660	0	104	360	0	Urban	N
LP001046	Male	Yes	1	Graduate	No	5955	5625	315	360	1	Urban	Y
LP001047	Male	Yes	0	Not Graduate	No	2600	1911	116	360	0	Semiurban	N

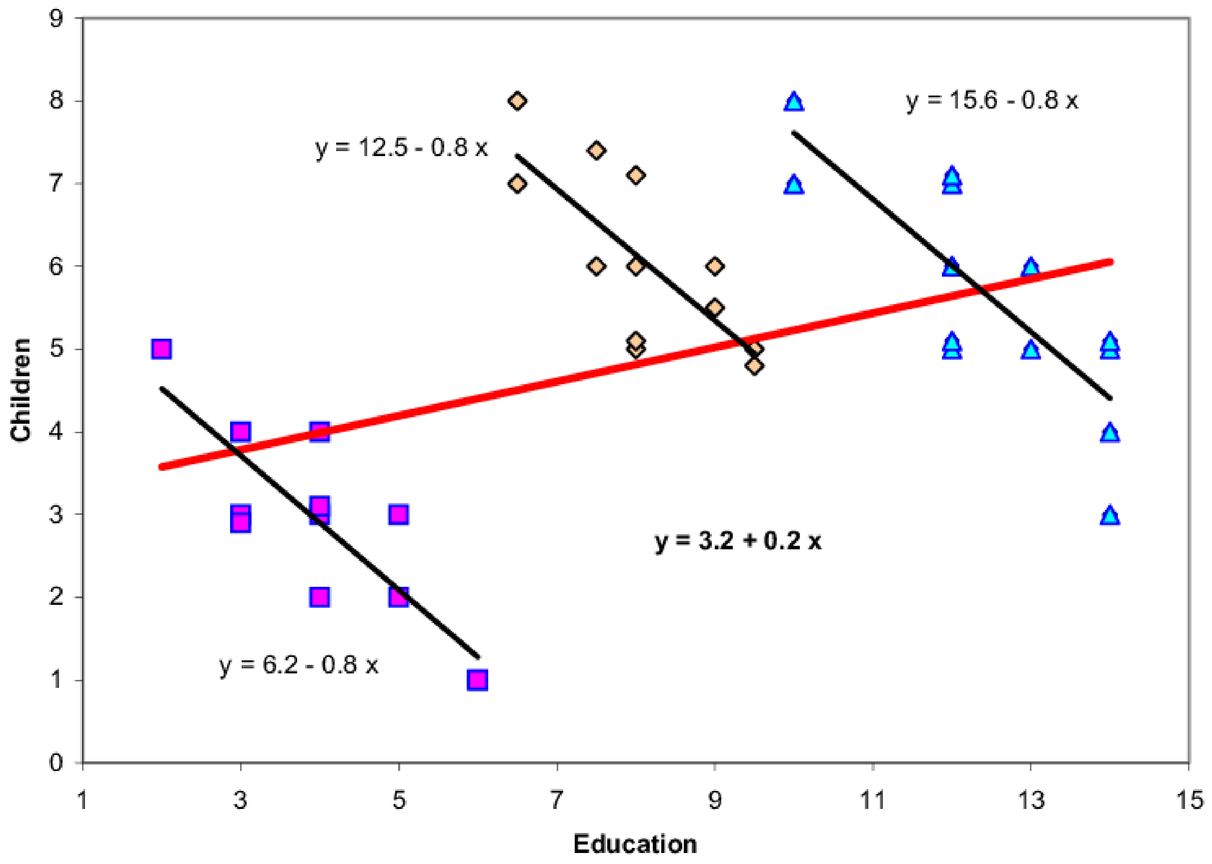
Terminologies in Data Analytics

Observation

Data Sampling

Data Set

Prediction

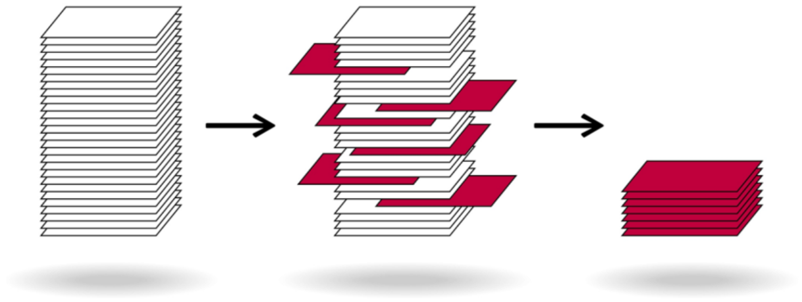


- The goal of prediction is to move from *what has happened* to providing the best assessment of *what will happen*.

- In the graph, linear prediction technique is used to predict the number of children within different education levels.

Types of Data

Types of Data



Structured Data

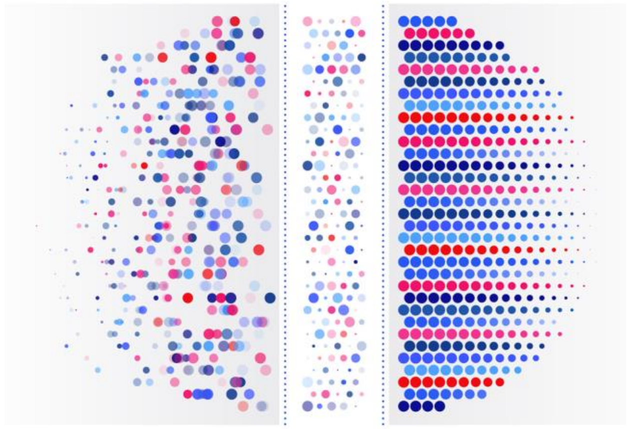
It is the data that is processed, stored, and retrieved in a fixed format.

Example: Employee details, job positions, and salaries.

Unstructured Data

It is the type of data that lacks any specific form or structure.

Example: Email



Semi-Structured Data

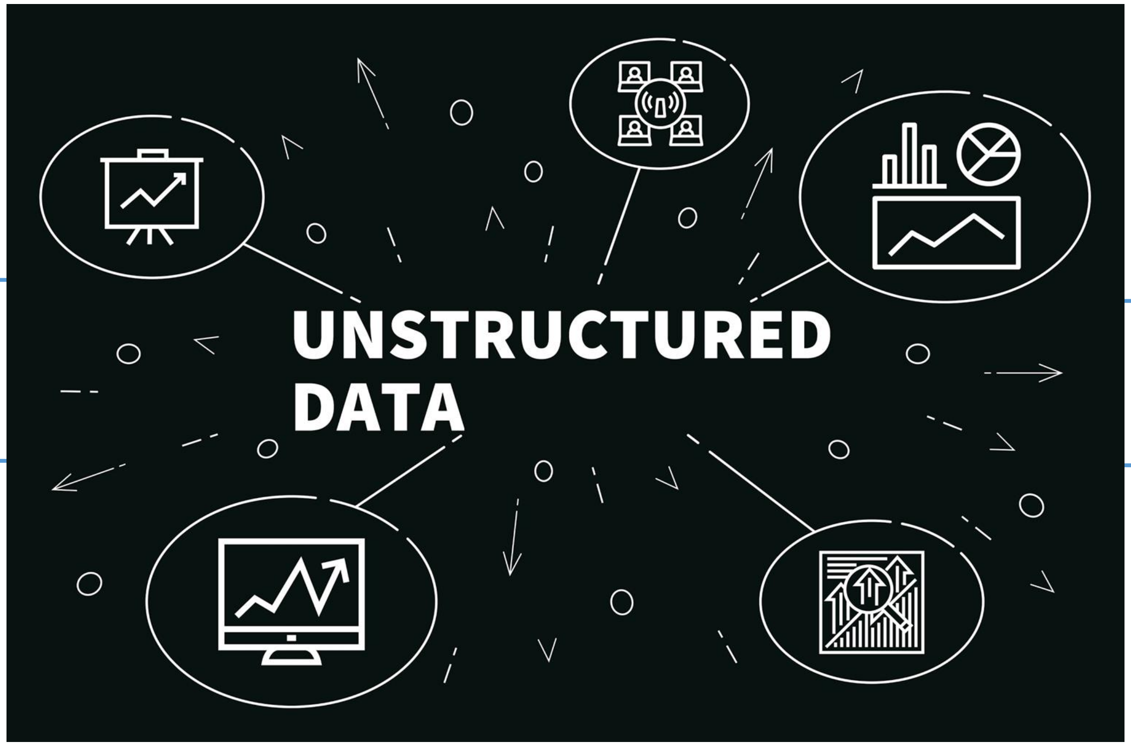
It is the data type containing both structured and unstructured data.

Example: CSV and JSON documents

Analyzing Unstructured Data

About 80% of business data is unstructured.

Unstructured information is text-heavy and contains data such as dates, numbers, and facts.



Internally generated information is considered *unstructured* as the intelligence doesn’t fit neatly into a database.

Unstructured data is primarily used for BI and analytics but not for transaction processing applications.

Analyzing Unstructured Data

Retailers and manufacturers analyze unstructured data to:

- Improve customer relationship management processes
- Enable targeted marketing
- Perform sentiment analysis on product reviews

The line between unstructured and semi-structured data is not clearly defined.

Unstructured data has some level of structure in it.

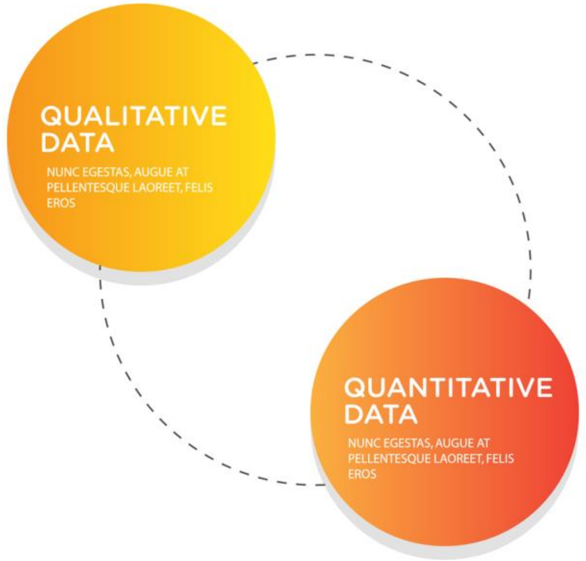
Qualitative and Quantitative Data

Qualitative and Quantitative Data

Qualitative Data

Data in which classification of objects is based on attributes and properties.

Example: Softness of skin etc.



QUALITATIVE DATA

NUNC EGESTAS, AUGUE AT PELLENTESQUE LAOREET, FELIS EROS

QUANTITATIVE DATA

NUNC EGESTAS, AUGUE AT PELLENTESQUE LAOREET, FELIS EROS

Quantitative Data

Data can be measured and expressed numerically.

Example: Your height and shoe size.

Qualitative and Quantitative Data

Qualitative Data

- Data collection is unstructured.
- It asks *why.*
- It cannot be computed as it is non-statistical.
- It develops initial understanding and defines the problem.

Quantitative Data

- Data collection is structured.
- It is all about *how much* or *how many.*
- It is statistical and is about numbers.
- It recommends the final course of action.

Subgroups of Qualitative Data

Qualitative Data

Nominal data

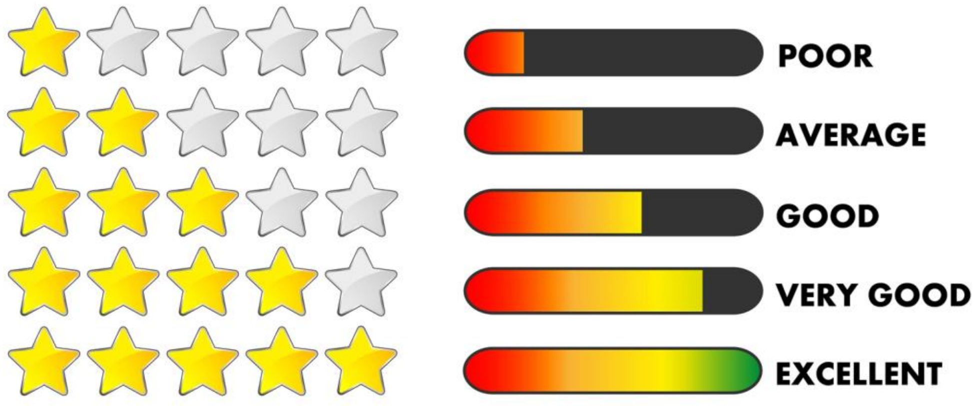
Unordered data to which an order is assigned in relation to other named categories

Example: Grade classification like pass or fail for student's test results.

Ordinal data

Ordered data that is assigned to categories in a ranked fashion

Example: Feedback to a product with 1–5 ranking.



Subgroups of Quantitative Data

Quantitative Data

Discrete data

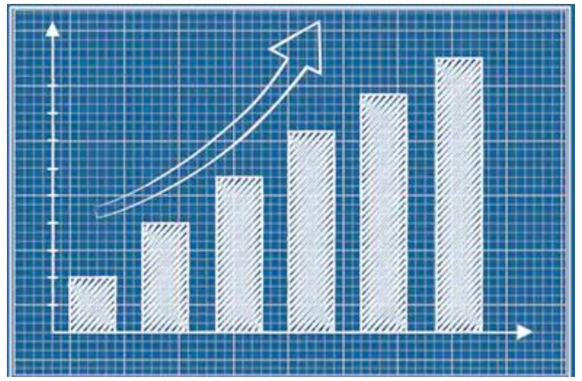
It can only take certain values.

Example: The number of students in a class

Continuous data

It can take any value within a specified range.

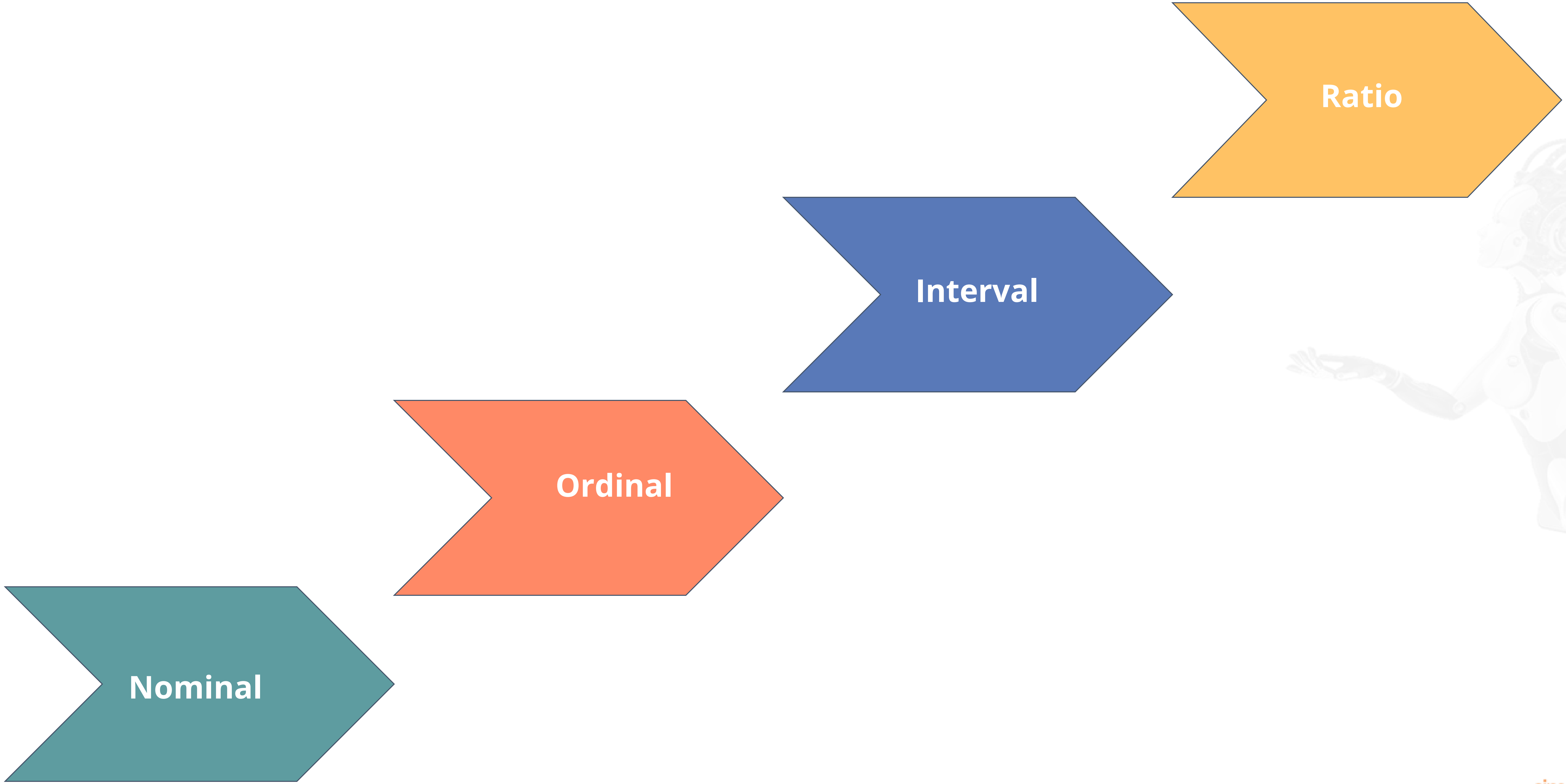
Example: Share price of a company



Data Levels of Measurement

Data Levels of Measurement

It is a classification that describes the nature of information within the values assigned to variables.



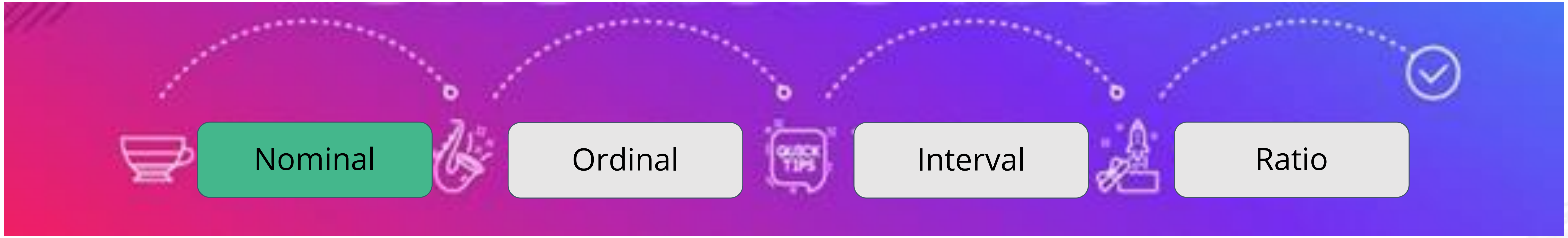
Nominal

Ordinal

Interval

Ratio

Data Levels of Measurement



Nominal

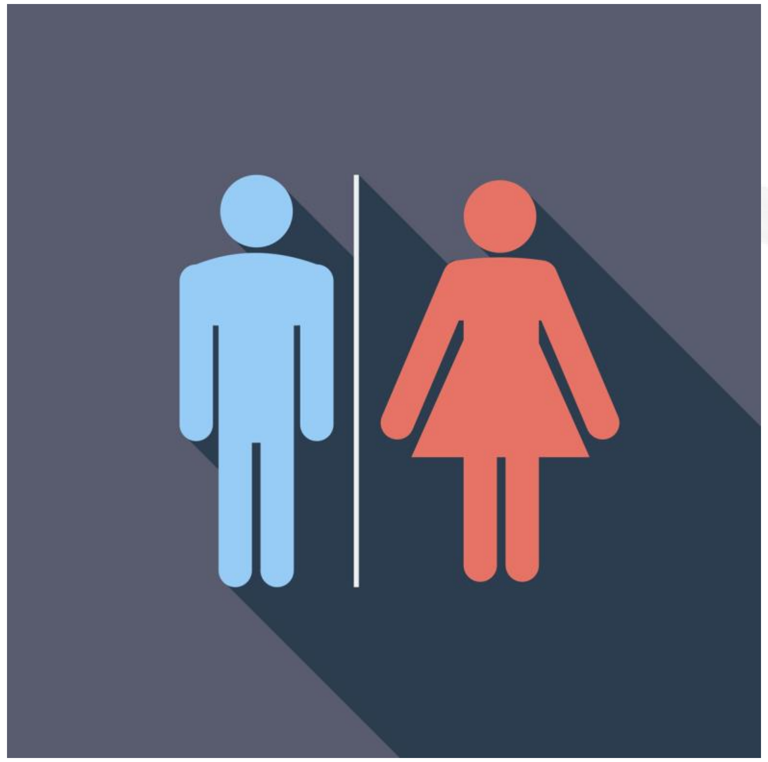
Ordinal

Interval

Ratio

- In nominal level of measurement, numbers in the variable are used to classify data.
- At this level, words, letters, and alphanumeric symbols can be used.
- Example: People in female gender category are classified as F and those in male gender are category classified as M.

M



F

Data Levels of Measurement

Nominal

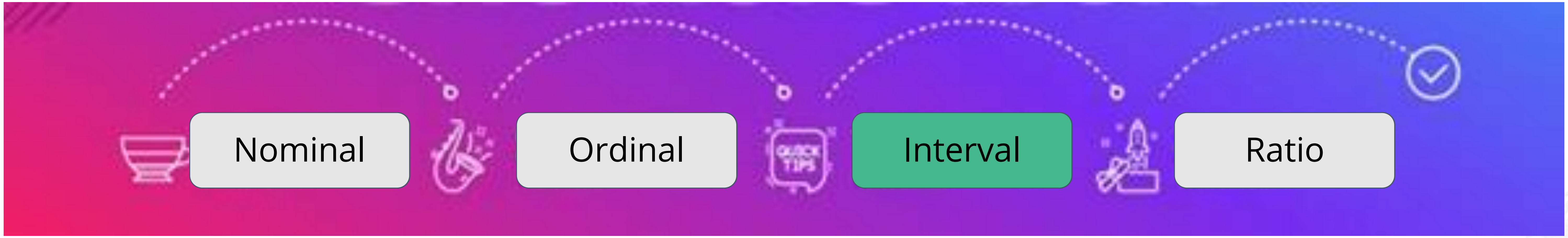
Ordinal

Interval

Ratio

- Ordinal level of measurement depicts ordered relationship among the variable’s observations.
- It indicates an order of the measurements.
- Example: A student with 100% score is assigned the first rank, another student with 95% score would be assigned the second rank, and so on.

Data Levels of Measurement



Nominal

Ordinal

Interval

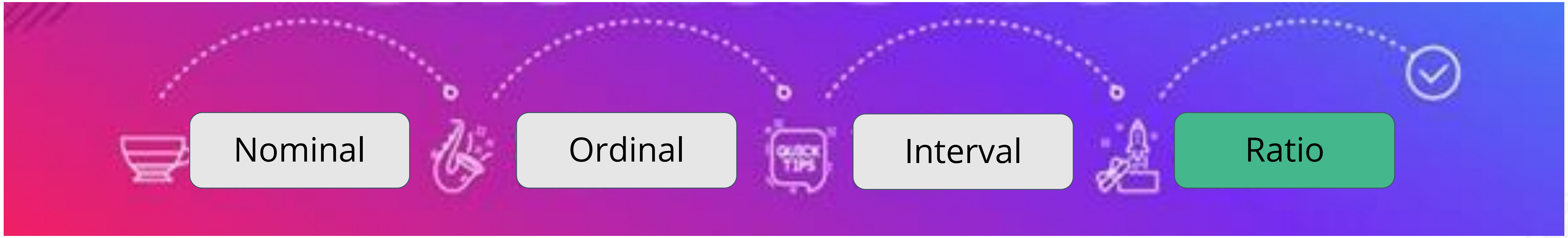
Ratio

- The interval level of measurement classifies and orders the measurements.
- It also specifies that the distances between each interval on the scale are equivalent.
- Example: Temperature in centigrade where the distance between 80 degrees and 100 degrees is same as the distance between 1000 degrees and 1020 degrees.

Temperature in centigrade

80**°C -** 100**°C** **=** 1000**°C -** 1020**°C**

Data Levels of Measurement



Nominal

Ordinal

Interval

Ratio

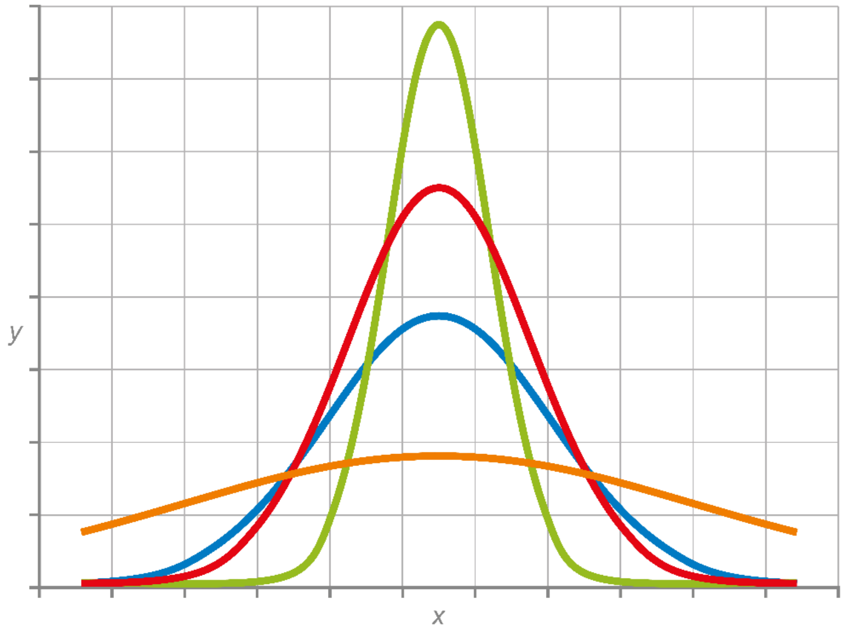
- In the ratio level of measurement, observations can have a value of zero.
- Although properties of ratio measurement are similar to the interval level of measurement, the zero in scale makes it different from the other levels of measurement.

Note: The nominal level classifies data, while the ordinal level indicates an order of measurements.
The interval level and the ratio level of measurements provide the same level of measurement.

Normal Distribution of Data

Normal Distribution of Data

- Normal distribution is also known as Gaussian distribution or Bell curve.
- Most of the natural phenomena and occurrences follow Bell curve.
- It is denser at the center and has equal mean, median, and mode values.



- It is the most important probability distribution in statistics.
- It is a perfectly symmetric bell-shaped distribution curve with only one peak.
- It is continuous and have tails that are asymptotic.

Statistical Parameters

Basic Statistical Parameters

Mean

- Mean is the average of all data points for a given set of data.
- It is used to derive the central tendency of the data.
- It is measured by adding all data points and dividing the sum by the number of data points.

Variance

- Variance is the sum of the squares of differences between all numbers and means divided by the number of data points.
- It gives a measure of how the data distributes itself about the mean.
- It looks at all the data points and then determines their distribution.

Standard Deviation

- Standard deviation is the square root of variance and shows the extent to which data varies from the mean.
- It shows how tightly data points are clustered around the mean.
- It is more concrete and gives the exact distances from the mean.

Basic Statistical Parameters: Example

Dataset x = {1;2;3;4;5;6}

$$\mu = \Sigma \frac{x_i}{n}$$

Mean = (1+2+3+4+5+6)/6 = 3.5

$$\sigma^2 = \frac{\Sigma(X_i - \mu)^2}{N}$$

Variance = $[(1-3.5)^2+(2-3.5)^2+(3-3.5)^2+(4-3.5)^2+(5-3.5)^2+(6-3.5)^2]/6 = 2.917$

$$\mathrm{SD} = \sqrt{\frac{\sum |x-\mu|^2}{N}}$$

Standard deviation = √2.917 = 1.708

Key Takeaways

- Structured data, unstructured data, and semi-structured data are the three types of data.
- Nominal, ordinal, interval, and ratio are four data levels of measurement.
- Normal distribution of data is the most important probability distribution in statistics.
- Mean, variance and standard deviation are the basic statistical parameters.