CSE1006
Foundations of Data Analytics

Adla Padma
Assistant Professor
School of Computer Science & Engineering
VIT-AP University

Module - 1
Introduction to Data Analytics

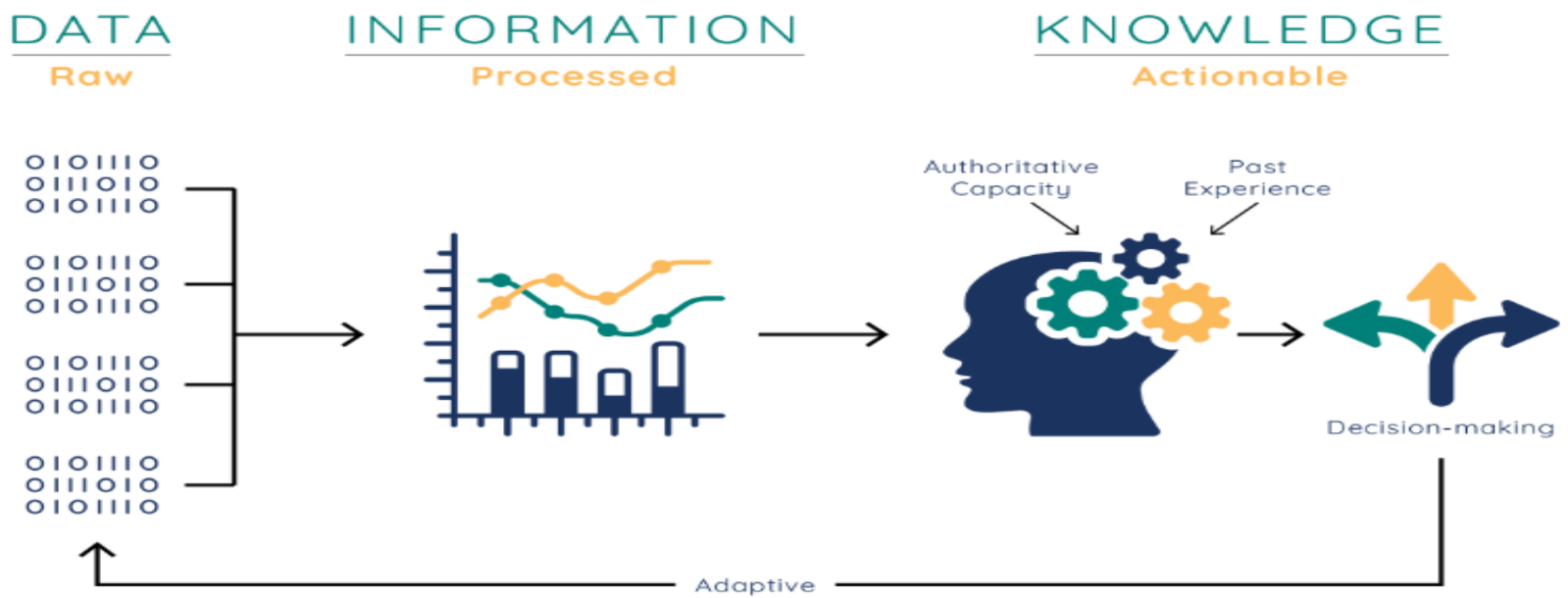
- Data-Information
- Characteristics of data
- Data *munging*
- *Scraping*
- *Sampling*
- *Cleaning*
- Importance of data analytics
- Success Stories

Introduction

- Data is a raw and unorganized fact that is required to be processed to make it meaningful. It can be considered as facts and statistics collected together for reference or analysis.
- Quantitative: Quantitative data refers to numerical information like weight, height, etc.
- Qualitative: Qualitative data refers to non-numeric information like opinions, perceptions, etc.
- Information is defined as structured, organized, and processed data, presented within a context that makes it relevant and useful to the person who needs it. Data suggests that raw facts and figures regarding individuals, places, or the other issue, that is expressed within the type of numbers, letters or symbols.

Example

- Temperature Readings: Numbers representing temperature throughout the day, such as “72°F”, “68°F”, “75°F”.
- Student Grades: A list of numerical scores obtained by students on a test, like “85”, “92”, “78”.
- Stock Prices: Daily closing prices of a company’s stock, such as “$50.25”, “$48.90”, “$52.10”.



Characteristics of data

Accurate: Error free, Measured correctly, Entered correctly

Complete: Has all important facts or data fields needed to achieve goals.

For example in Student records system: Name, SSN, GPA, payments, course grades

Flexible: Can be used for a variety of purposes, easily transformed to another use and transferable to another application or use

Reliable: Always accurate/available – not just sometimes

Relevant: Important to decision maker and Remove non-relevant items from screens

Timely: Delivered when it is needed, Maximizes value and Relevant in time

Data Analytics

- Data analytics is the process of collecting, transforming, and organizing data to discover patterns, make predictions, and drive decision making.
- Data analysis refers to the process of examining datasets to draw conclusions or make predictions based on historical information. It focus on specific problem and techniques involved are statistical techniques, such as hypothesis testing, regression analysis, and data visualization
- Data analytics involves not only analyzing past trends but also utilizing advanced tools and techniques, such as machine learning algorithms and statistical models for predictive modeling and decision-making purposes. It focus on broder application area.

Steps involved in data analytics

Data Collection: Gathering data from various sources such as databases, sensors, surveys, and online sources.

Data Cleaning: Preparing data by removing errors, duplicates, and inconsistencies to ensure accuracy and quality.

Data Transformation: Converting data into a suitable format for analysis, often involving normalization, aggregation, and other preprocessing steps.

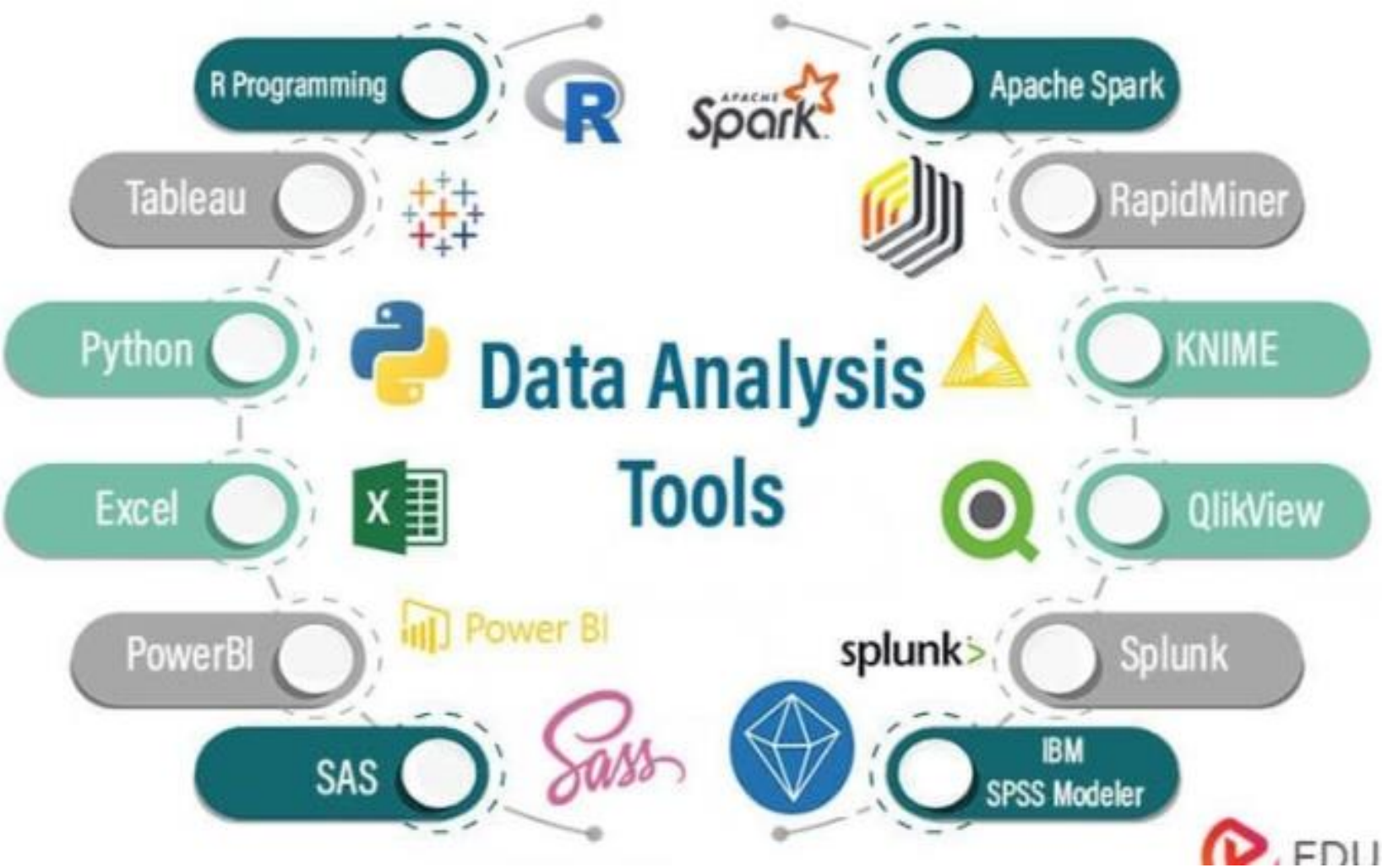
Exploratory Data Analysis (EDA): Using statistical methods and visualization tools to explore data and understand its main characteristics and underlying patterns.

Modeling and Algorithms: Applying mathematical models, algorithms, and machine learning techniques to analyze data and make predictions or classifications.

Data Visualization: Creating charts, graphs, and dashboards to present the results of data analysis in an understandable and visually appealing way.

Interpretation and Reporting: Interpreting the results of the analysis, drawing conclusions, and making recommendations based on the insights gained.

Data Analytics Tools



Measures of centrality or central tendency

Measures of centrality, or central tendency are statistical measures that describe the center or distribution of a dataset

1. Mean (Arithmetic Average)

Definition: Mean is the sum of all the data points divided by the number of points. It is used to find the average value in a dataset.

- Suppose you have the following data set of test scores: 60, 70, 80, 90, and 100.
- The mean is calculated as:

$$\text{Mean} = \frac{60 + 70 + 80 + 90 + 100}{5} = \frac{400}{5} = 80$$

- The mean score is 80.

Median (Middle Value)

Definition: The median is the middle value when the data points are arranged in ascending or descending order. If there is an even number of observations, the median is the average of the two middle values.

- Consider the following data set: 12, 18, 22, 26, 30.
- The median is 22 because it is the middle value.
- If the data set is 12, 18, 22, 26, 30, 34, then the median is the average of 22 and 26:

$$\text{Median} = \frac{22 + 26}{2} = 24$$

Mode (Most Frequent Value)

Definition: The mode is the value that appears most frequently in a dataset. A dataset may have one mode, more than one mode (bimodal or multimodal), or no mode at all.

- Suppose you have the following set of numbers: 4, 4, 5, 5, 5, 6, 7, 7.
- The mode is 5 because it appears most frequently.
- If you have the data set 1, 2, 2, 3, 3, 4, then the modes are 2 and 3 (bimodal).

Range

Definition: It represents the difference between the highest and lowest values in the data.

Formula:

$$\text{Range} = \text{Maximum Value} - \text{Minimum Value}$$

Example:

- Consider the following dataset of exam scores: 58, 62, 75, 80, 85, 90.
- The maximum value is 90, and the minimum value is 58.
- The range is calculated as:

$$\text{Range} = 90 - 58 = 32$$

- So, the range of this dataset is 32.

Variance

Definition: In statistics, variance measures how far a set of numbers are spread out from their mean value.

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

- Consider the following data points: 4, 8, 6, 5, 3.
- Step 1: Calculate the mean ($\bar{x}$):

$$\bar{x} = \frac{4 + 8 + 6 + 5 + 3}{5} = \frac{26}{5} = 5.2$$

- Step 2: Calculate each squared difference from the mean:

$$(4 - 5.2)^2 = 1.44, \quad (8 - 5.2)^2 = 7.84, \quad (6 - 5.2)^2 = 0.64,$$

- Step 3: Sum the squared differences: $(5 - 5.2)^2 = 0.04, \quad (3 - 5.2)^2 = 4.84$

$$1.44 + 7.84 + 0.64 + 0.04 + 4.84 = 14.8$$

- Step 4: Calculate the variance:

$$s^2 = \frac{14.8}{5 - 1} = \frac{14.8}{4} = 3.7$$

- So, the sample variance is 3.7.

Standard Deviation

It shows how much variation in the data.

A low standard deviation indicates that the data points tend to be very close to the mean, whereas high standard deviation indicates that the data is spread out over a large range of values.

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

Using the variance from the previous example (3.7), the standard deviation is:

$$s = \sqrt{3.7} \approx 1.92$$

Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
40	M	ATA	140	289	0	Normal	172	N	0	Up	0
49	F	NAP	160	180	0	Normal	156	N	1	Flat	1
37	M	ATA	130	283	0	ST	98	N	0	Up	0
48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
54	M	NAP	150	195	0	Normal	122	N	0	Up	0
39	M	NAP	120	339	0	Normal	170	N	0	Up	0

a) What is the mean of the attribute Age? (2M)
b) What is the median of the attribute Cholesterol? (2M)
c) What is the mode of the attribute ST_Slope? (2M)
d) What is the Range of the attribute MaxHR? (2M)
e) What is the standard deviation of the attribute RestingBP? (2M)

Data munging

In data analysis, data munging refers to the process of cleaning and transforming raw data into a usable format for analysis.

Data Cleaning:

- Removing or correcting errors
- Handling missing values

Data Transformation:

- **Normalizing data**: Scaling data to a standard range or distribution.

Normalization techniques

Normalization techniques in data analytics are methods used to scale and transform data so that it fits within a certain range, making it easier to analysis.

1. Min-Max Normalization (Feature Scaling)

Min-Max normalization scales the data to fit within a specific range, typically [0, 1] or [-1, 1].

Formula:

$$v_i' = \frac{v_i - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A.$$

Example

Min-max normalization. Suppose that the minimum and maximum values for the attribute *income* are \$12,000 and \$98,000, respectively. We would like to map *income* to the range [0.0, 1.0]. By min-max normalization, a value of \$73,600 for *income* is transformed to $\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$. ■

2. Z-Score Normalization (Standardization)

Z-score normalization transforms the data so that it has a mean of 0 and a standard deviation of 1. This technique is useful when the data follows a normal distribution.

Formula:

$$Z = \frac{X - \mu}{\sigma}$$

•μ is the mean of the data & σ is the standard deviation of the data.

Example: For a dataset with a mean of 50 and a standard deviation of 10, if a data point is 60, the z-score would be:

$$Z = \frac{60 - 50}{10} = 1$$

Similarlly find datapoint 210, 510

a value of 70 would be

$$X' = \frac{70 - 50}{10} = 2$$

3. Decimal Scaling Normalization

Decimal scaling normalizes data by moving the decimal point of values. The number of decimal points moved depends on the maximum absolute value of the data.

Formula:

$$X' = \frac{X}{10^j}$$

Where j is the smallest integer such that $\max(|X'|) < 1$

Example-1:

CGPA	Formula	CGPA Normalized after Decimal scaling
2	2/10	0.2
3	3/10	0.3

Example-2:

Salary bonus	Formula	CGPA Normalized after Decimal scaling
400	400 / 1000	0.4
310	310 / 1000	0.31

S no.	Original data	Decimal scaling
1	2598	0.2598
2	3145	0.3145
3	3658	0.3658
4	4597	0.4597
5	5830	0.583
6	6950	0.695
7	7128	0.7128
8	8962	0.8962
9	8763	0.8763

Age	%Fat
23	9.5
23	26.5
27	7.8
27	17.8
39	31.4
41	25.9
47	27.4

a) Normalize the values of the attribute Age using min-max normalization with new minimum as 0 and new maximum as 1.

b) Normalize the values of the attribute FAT using z-score normalization

c) Multiply each value of the attribute FAT with 10. Next, normalize them using decimal scale normalization

Scraping

- Scraping, or web scraping, is a technique in data analytics used to extract large amounts of data from websites.
- This data can be analyzed to uncover patterns, trends, and insights that are valuable for decision-making.

Steps:

1.Identify the Target Website: Determine which website(s) you need to extract data from.

2.Inspect the Web Page: Use browser developer tools to understand the structure of the web page (HTML, CSS).

3.Write the Scraper: Use a programming language (such as Python) and libraries (like BeautifulSoup, Scrapy, or Selenium) to write scripts that can navigate and extract data from the web pages.

4.Handle Data Storage: Store the scraped data in a structured format, such as CSV, JSON, or a database.

5.Data Cleaning and Processing: Clean and process the extracted data to make it suitable for analysis.

Data cleaning

Data cleaning, is the process of detecting and fixing issues such as missing data, duplicate entries, incorrect data types, Noisy data, outliers, and inconsistent formatting.

Data cleaning operations:

1. Handling Missing Data
2. Removing Duplicates
3. Correcting Data Types
4. Removing Outliers
5. Smooth noisy data

RNO	Age	DOB	Salary
1	30	21-10-2002	2000000
2	37	22-10-1985	5000000
3		23-11-2002	-20
4	200	24-12-2000	30000
5	88	"25-10-2002"	20000
6	50	26-01-1990	100000
6	50	26-01-1990	100000

Handling Missing Data

•Some rows in the dataset have missing values.

Cleaning Action:

1. Remove rows with missing data if the dataset is large and the missing data is minimal.

EmployeeID	Name	Email
101	Alice	alice@example.com
102	Bob	None
103	Charlie	charlie@example.com
104	David	None

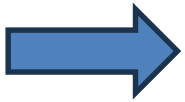
Remove rows

EmployeeID	Name	Email
101	Alice	alice@example.com
103	Charlie	charlie@example.com

2. Impute/Replace missing values using the mean, median, or mode of the column

Filling Missing Data with Mean

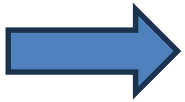
CustomerID	Age
1	20
2	nan
3	30
4	50



CustomerID	Age
1	20
2	33
3	30
4	50

Filling Missing Data with Median

CustomerID	Age
1	20
2	nan
3	30
4	50



CustomerID	Age
1	20
2	30
3	30
4	50

3. Filling Missing Data Using Forward/Backward Fill:

- **Scenario**: If the data is time-series or ordered, missing values can be filled based on the previous or next value in the sequence.
- **Example**: A dataset of daily temperatures where some days have missing entries.

Removing Duplicates

•**Problem**: Some rows are exact duplicates.

•**Cleaning Action**:

- **Remove** duplicate rows to ensure each transaction is counted only once.

Correcting Data Types

•**Problem**: Some columns have incorrect data types.

•**Example**: A dataset where the "Date" column is stored as a string instead of a date.

•**Cleaning Action**:

- Convert a "Date" column from string format to datetime format for accurate time-based analysis.

Removing Outliers

•**Problem**: Some data points are far outside the expected range.

•**Example**: A dataset containing customer ages where a few entries show unrealistic ages, like 200 years.

•**Cleaning Action**:

- **Remove** or **correct** outliers based on domain knowledge or statistical methods.

Smooth noisy data

- Smoothing noisy data is a process in data analytics where noise is reduced or eliminated from a dataset to make the underlying patterns or trends more visible.
- Noisy data contains random fluctuations, errors, or irrelevant information.

Ex: Salary="−10" (an error)

- Smoothing helps in enhancing the clarity and interpretability of data.
- Binning method is used to smoothing data or to handle noisy data.
- In this method, the data is first sorted and then the sorted values are distributed into a number of bins.

There are three approaches to performing smoothing –

Smoothing by bin means : In smoothing by bin means, each value in a bin is replaced by the mean value of the bin.
Smoothing by bin median : In this method each bin value is replaced by its bin median value.
Smoothing by bin boundary : In smoothing by bin boundaries, the minimum and maximum values in a given bin are identified as the bin boundaries. Each bin value is then replaced by the closest boundary value.

Approach:

1.Sort the array of a given data set.
2.Divides the range into N intervals, each containing the approximately same number of samples(Equal-depth partitioning).
3.Store mean/ median/ boundaries in each row.

Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

Partition using equal frequency approach:

- 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 median:

- 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

The following data is used to represent the weight of sampled 15 elephants in a zoo

2005, 1949, 1299, 1398, 1599, 2200, 4198, 5200, 2400, 6201, 4201, 3201, 4201, 2703, 2054

Smooth the given data using the following methods with bin value is 5

(a) Smoothing by bin means

(b) Smoothing by bin medians

(c) Smoothing by bin boundaries

11,13,13,15,15,16,19,20,20,20,21,21,22,23,24,30,40,45,45,45,71,

72,73,75

Smooth the given data using the following methods with bin value is 4

(a) Smoothing by bin means

(b) Smoothing by bin medians

(c) Smoothing by bin boundaries

Data Sampling

- Sampling in data analytics refers to the process of selecting a subset of data from a larger dataset to make inferences or gain insights about the entire population.
- Sampling is particularly useful when dealing with large datasets, where analyzing the entire dataset might be impractical, time-consuming, or expensive.
- By analyzing a representative sample, you can estimate the characteristics of the whole population with a certain level of confidence.

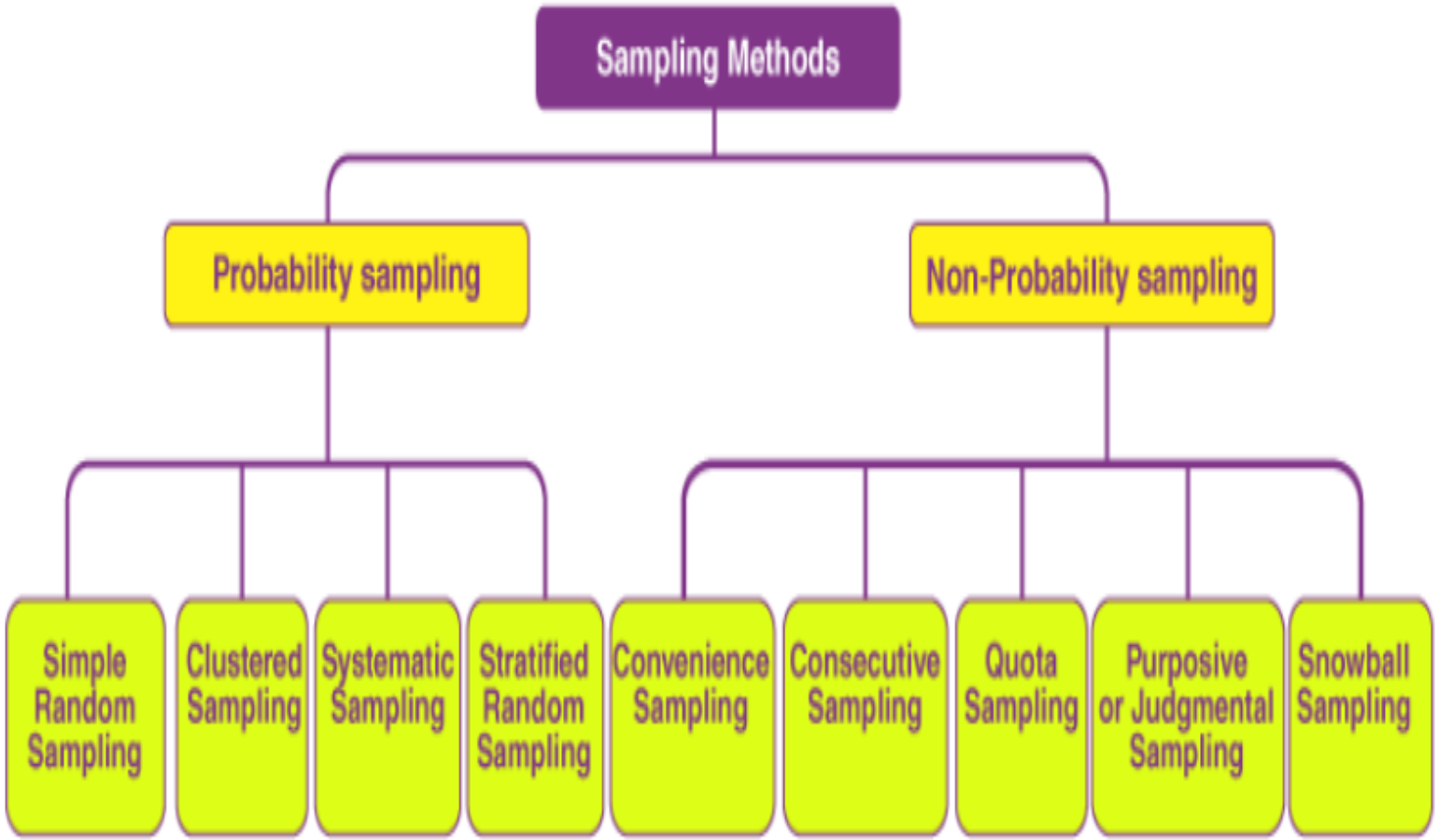
Why Sampling is Important:

1.Efficiency: Analyzing a sample is faster and less resource-intensive than analyzing the entire dataset.

2.Cost-Effective: Reduces the costs associated with data collection, storage, and processing.

3.Feasibility: Sometimes, it's impossible to collect data from the entire population (e.g., surveying all citizens of a country), so sampling is necessary.

4.Accuracy: With proper sampling techniques, you can achieve accurate results that are representative of the whole population.


Sampling Methods
Probability sampling
Non-Probability sampling
Simple Random Sampling
Clustered Sampling
Systematic Sampling
Stratified Random Sampling
Convenience Sampling
Consecutive Sampling
Quota Sampling
Purposive or Judgmental Sampling
Snowball Sampling

Sampling Methods

- Probability Sampling: In this technique, the samples are chosen at random, and each sample has a known probability of being selected. There are different kinds of probability sampling techniques.

- Non-probability Sampling: In this technique, the odds of observations being selected is not defined and selection is carried out by the subjective judgment of the researchers.

Probabilistic Sampling is further classified as

Simple Random Sampling

sample(x, size, replace = FALSE, prob = NULL)

Ex: sample(1:20000, 500, replace = TRUE, prob = c(0.90, 0.10))

A sample of size 10 from the numbers 1 to 10 can be generated as below.

```
> sample(1:10,10)
 [1] 8 2 9 10 3 7 6 5 1 4
```

Stratified Sampling

In stratified sampling, the population is divided into smaller subgroups based on some common factors that best describe the entire population like age, sex, income, etc.

Ex:
```
library(dplyr)
set.seed(1)
sample_iris <- iris %>%
group_by(Species) %>%
sample_n(10)
```

Systematic Sampling

In systematic sampling, individuals are chosen at fixed intervals from the population data. To create a sample of size n from a population of size p fixed interval(k) is taken as p/n.

i.e, k=p/n i.e, for a population of size 1000, to create a sample of size 100 (1000/100), every 10th item from any random starting point can be chosen to be included in the sample.

Cluster Sampling

A cluster sampling is generally used in cases where the population data is geographical in nature or when there are some predefined groups within the population based on demographics, habits, background, etc.

For example, suppose that an organization wants to analyze the side effects of a drug across the United States, in this case, a two-stage cluster sampling can be performed by first dividing the entire population into cities(where each city data has the details about the side effects of the drug for all the patients) and then randomly selecting patients within these cities to be included in the sample.