|| JAI SRI GURUDEV ||

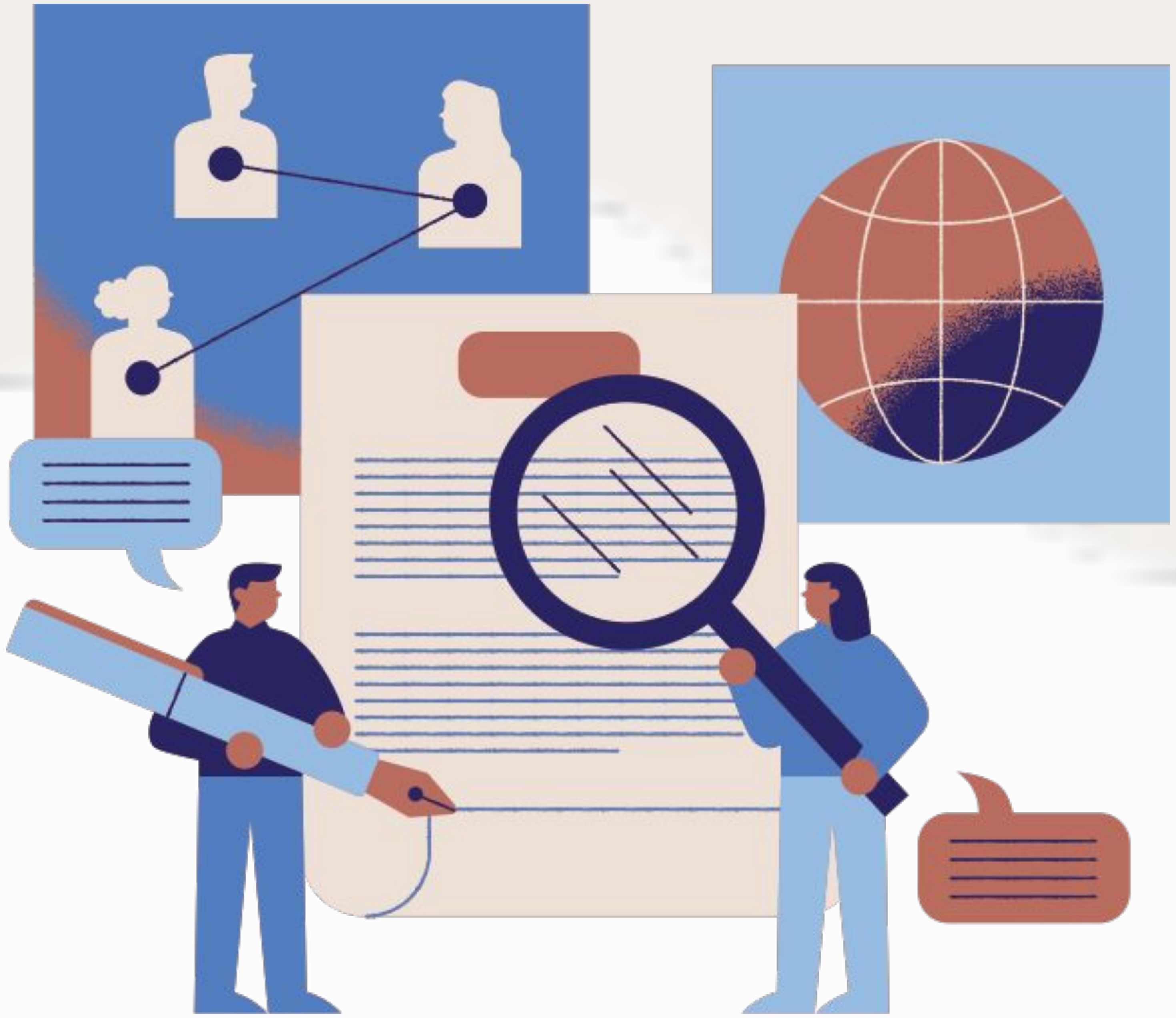
Sri Adichunchanagiri Shikshana Trust®

SJB INSTITUTE OF TECHNOLOGY

AN AUTONOMOUS INSTITUTE UNDER VISVESVARAYA TECHNOLOGICAL UNIVERSITY

Course: Exploratory Data Analytics

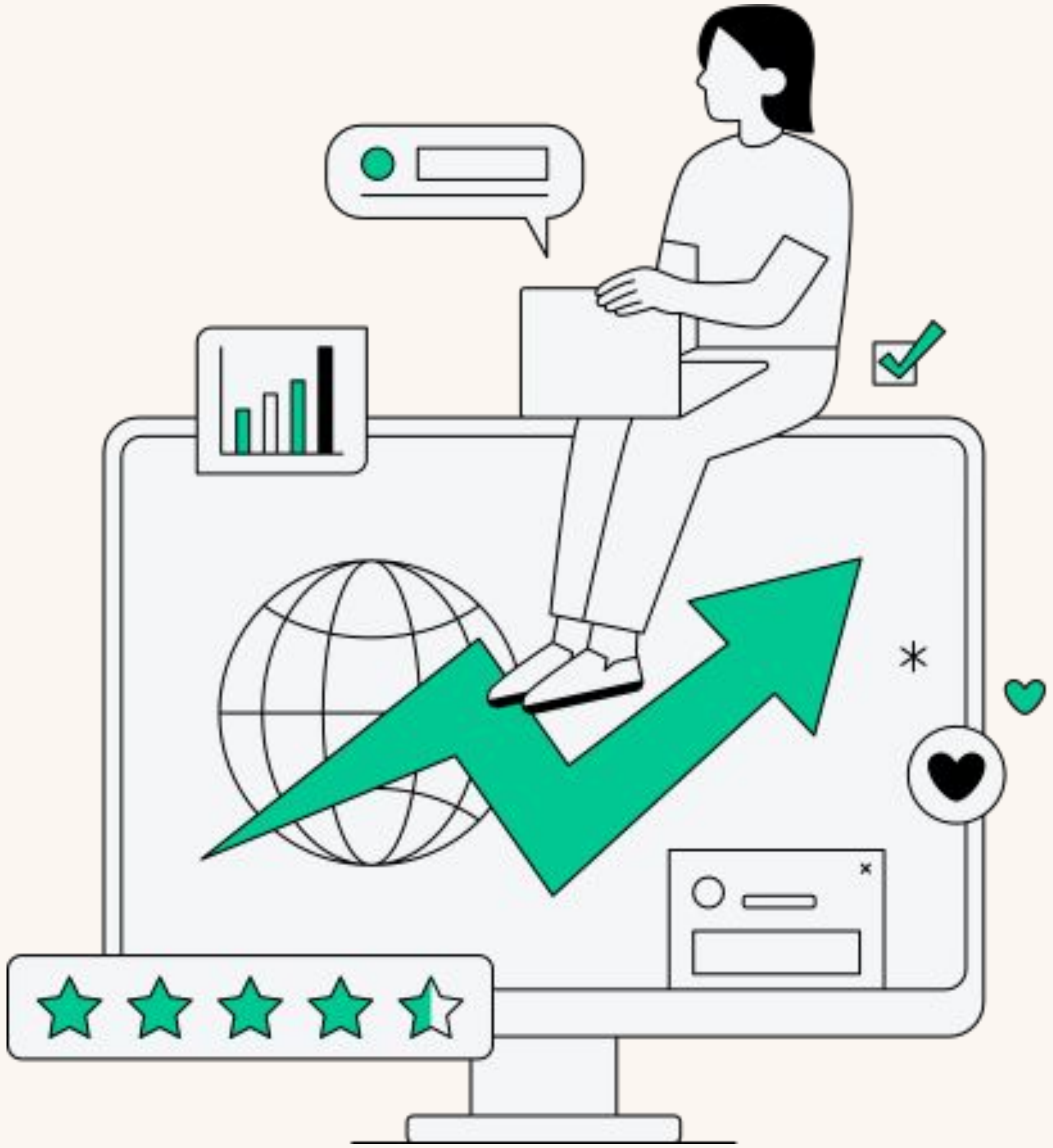
Course Code: 23CSE422



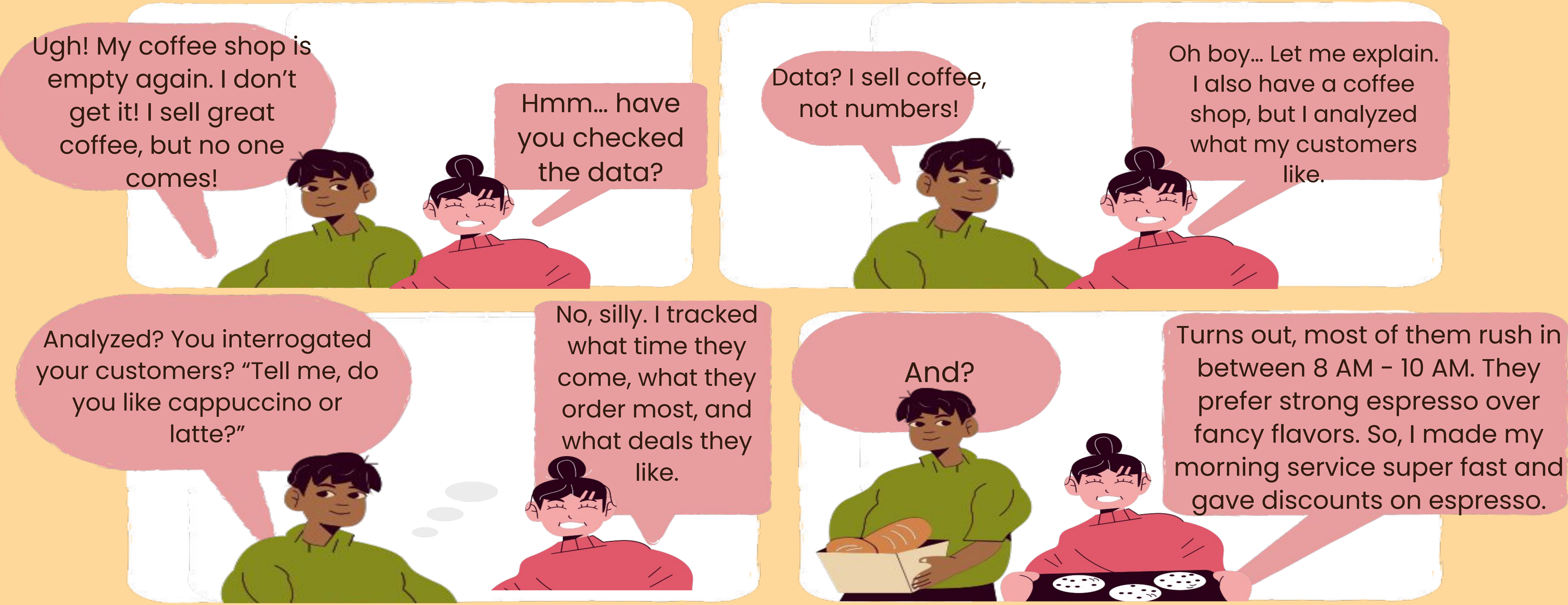
Mrs. Shilpashree S,
Assistant Professor
Department of CS & E

Introduction to EDA

Presented by
Shilpashree S,
Assistant Professor,
Dept of CSE, SJBIT

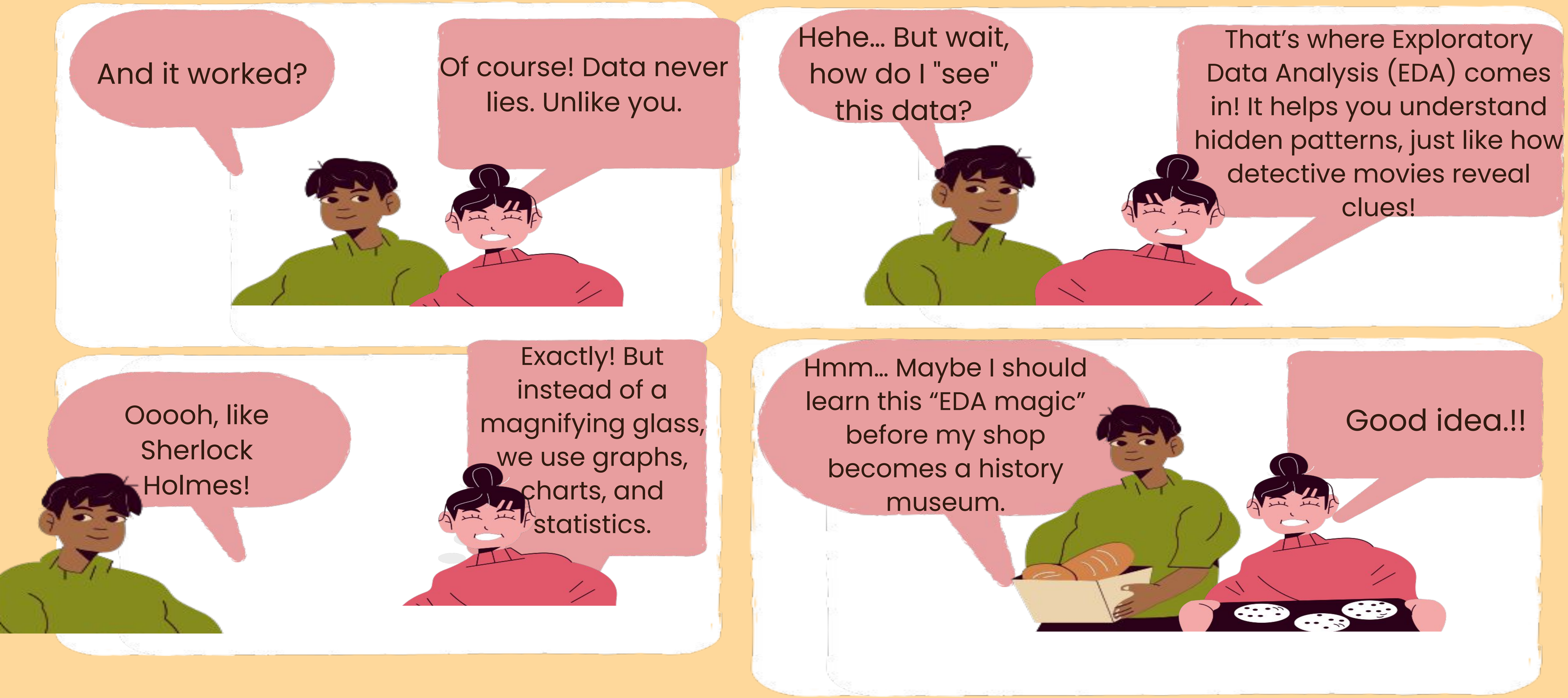


The Secret Recipe for Success: Data!


Ugh! My coffee shop is empty again. I don’t get it! I sell great coffee, but no one comes!
Hmm… have you checked the data?
Data? I sell coffee, not numbers!
Oh boy… Let me explain. I also have a coffee shop, but I analyzed what my customers like.
Analyzed? You interrogated your customers? “Tell me, do you like cappuccino or latte?”
No, silly. I tracked what time they come, what they order most, and what deals they like.
And?
Turns out, most of them rush in between 8 AM - 10 AM. They prefer strong espresso over fancy flavors. So, I made my morning service super fast and gave discounts on espresso.

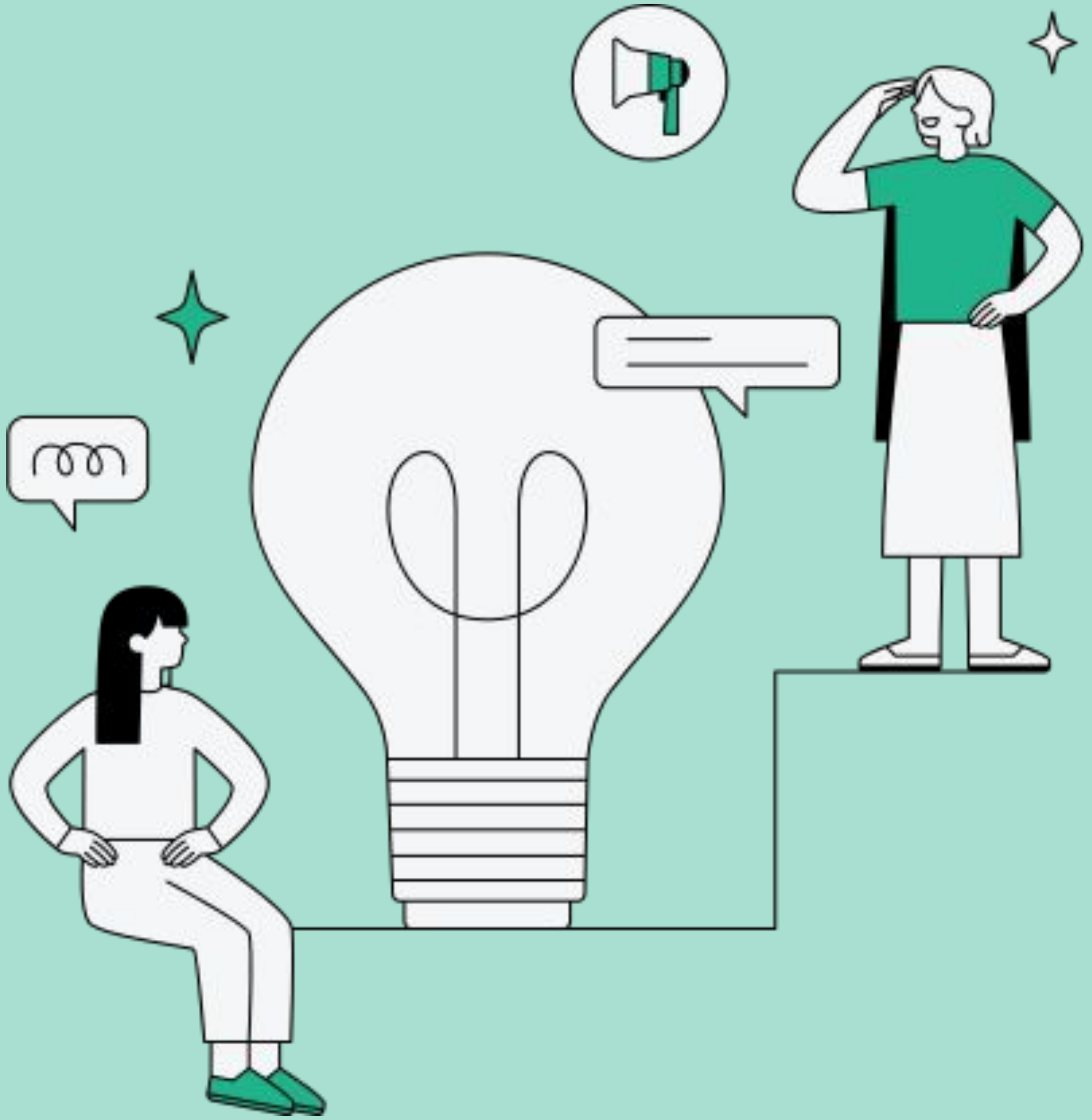
INSPIRED BY mathspace

The Secret Recipe for Success: Data!


And it worked?
Of course! Data never lies. Unlike you.
Hehe... But wait, how do I "see" this data?
That’s where Exploratory Data Analysis (EDA) comes in! It helps you understand hidden patterns, just like how detective movies reveal clues!
Ooooh, like Sherlock Holmes!
Exactly! But instead of a magnifying glass, we use graphs, charts, and statistics.
Hmm... Maybe I should learn this “EDA magic” before my shop becomes a history museum.
Good idea.!!

What is Data

- **Definition:**
  - **Data is a collection of discrete objects, numbers, words, events, facts, measurements, or descriptions.**
- **Sources of Data:**
  - **Collected and stored during events or processes in various disciplines (e.g., biology, economics, engineering).**
- **Purpose of Data:**
  - **To extract useful information and generate knowledge.**



What is Data

Objects 🏀 – A sports store keeps track of footballs, basketballs, and tennis rackets in its inventory.

Numbers 🔢 – A bank records account balances, transaction amounts, and credit scores.

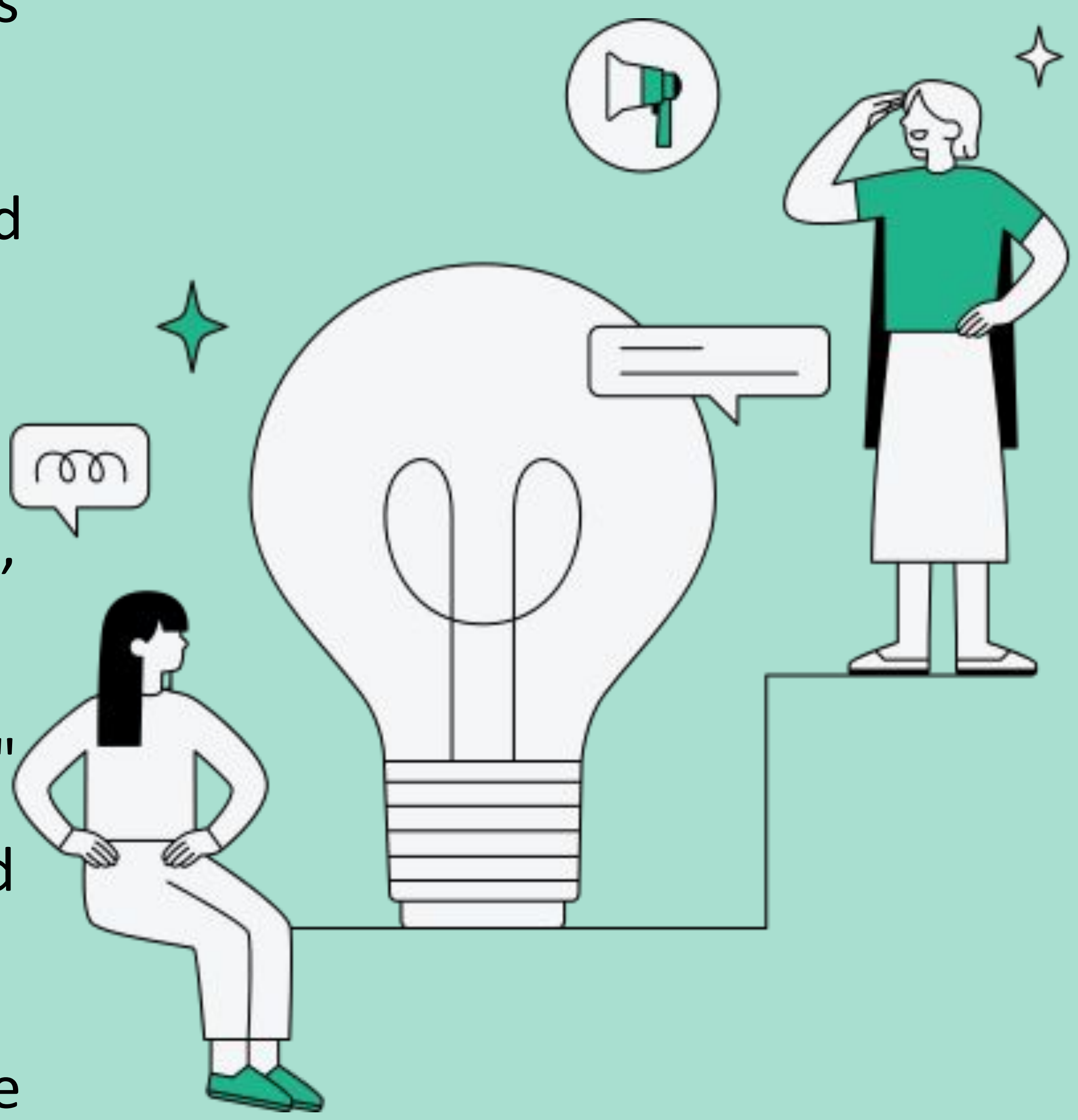
Words 📝 – A chatbot stores customer questions to improve its responses.

Events 🎉 – A ticketing website logs details of concerts, sports matches, and movie showtimes.

Facts 📖 – A weather app notes that "Rain is common in July in Bengaluru."

Measurements 📏 – A fitness tracker records heart rate, step count, and sleep hours.

Descriptions 🗣️ – An online store collects product reviews like "This phone has a great camera but poor battery life."



Sources of Data

1️⃣ Primary Sources (First-hand Data)

This is data collected **directly** from the source for a specific purpose.

✅ **Examples:**

- Surveys & Questionnaires 📝 – Customer feedback collected by a company.
- Experiments 🧪 – Scientists measuring chemical reactions in a lab.
- Observations 👀 – A researcher noting traffic patterns in a city.
- Interviews 🎤 – A journalist collecting quotes from eyewitnesses.

2️⃣ Secondary Sources (Pre-collected Data)

This is data **already collected and published** by someone else.

✅ **Examples:**

- Government Reports 📊 – Census data, economic surveys.
- Research Papers 📚 – Studies published by universities.
- Company Databases 💾 – Sales reports, employee records.
- News Articles 📰 – Data compiled from different events.

3️⃣ Machine-Generated Data (Automated Data)

Data created by **computers, sensors, and devices** without human input.

✅ **Examples:**

- Web Analytics 🌐 – Google tracking website visitors.
- IoT Sensors 📡 – Smartwatches tracking heart rate.
- GPS & Location Data 📍 – Maps tracking real-time traffic.
- Automated Transactions 💰 – ATMs recording withdrawals.

4️⃣ Online & Social Media Data

Massive amounts of data are generated through digital platforms.

✅ **Examples:**

- Social Media Posts 📱 – Tweets, Facebook updates, YouTube comments.
- E-commerce Data 🛒 – Amazon tracking customer purchases.
- Search Engine Queries 🔍 – Google searches providing trends.
- Online Reviews ⭐ – Ratings and feedback on products.

5️⃣ Open Data Sources

Freely available datasets from public institutions and organizations.

✅ **Examples:**

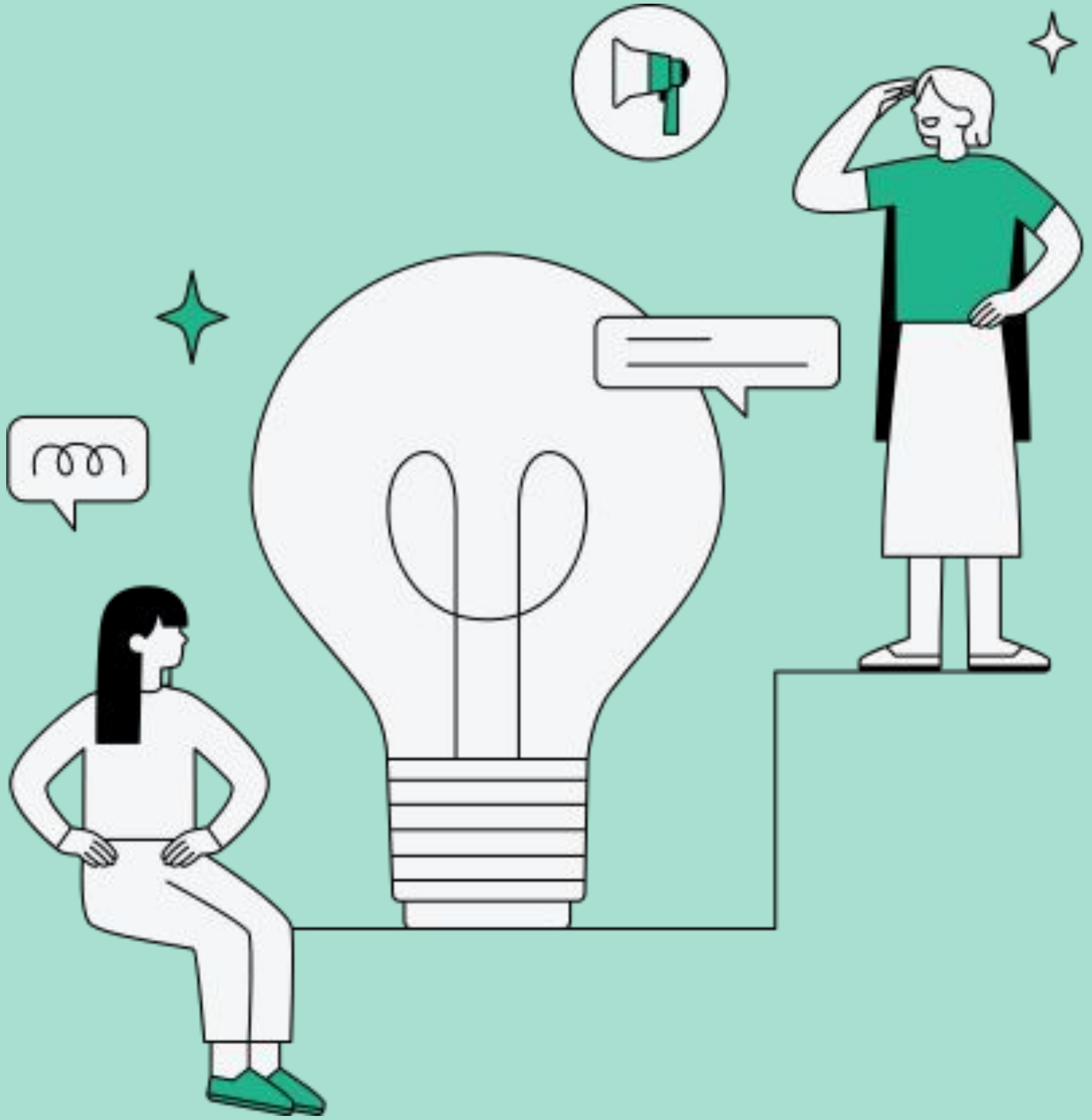
- Government Portals 🌍 – Data from WHO, NASA, World Bank.
- Open-Source Datasets 📂 – Kaggle, UCI Machine Learning Repository.
- Weather & Climate Data 🌦️ – OpenWeather, NOAA records.
- Public APIs 🔗 – Wikipedia, Twitter API for extracting data.

Purpose of Data

- **E-commerce**
  An online store uses product descriptions like "Lightweight, waterproof, and best for running" to help customers make informed choices.
- **Sports Inventory Management**
  A sports store tracks the number of footballs, basketballs, and tennis rackets to avoid overstocking or running out of items.
- **Banking & Finance**
  A bank records account balances and transactions to detect fraud, calculate interest, and offer financial insights.
- **Customer Support (AI Chatbots)**
  A chatbot analyzes customer queries to improve automated responses and provide better support.
- **Event Ticketing**
  A ticketing website records upcoming concerts and sports matches to recommend events based on user preferences.
- **Weather Forecasting**
  A weather app uses historical data to state that "Rain is common in July in Bengaluru," helping farmers and travelers plan accordingly.
- **Fitness & Health**
  A fitness tracker records heart rate, step count, and sleep hours to provide health insights.

What is EDA?

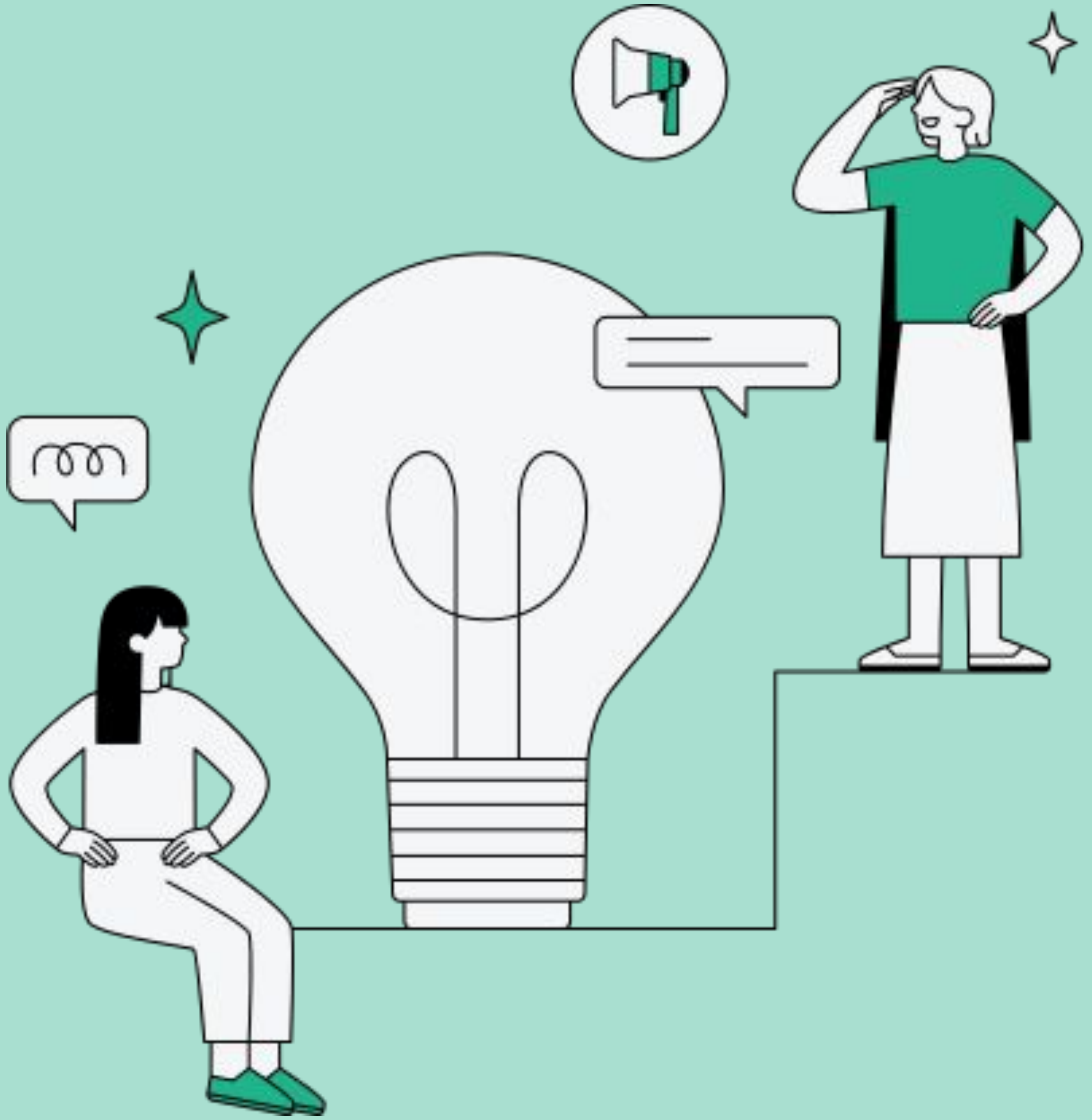
- Definition:
  - EDA is the process of examining datasets to:
    - Discover patterns.
    - Spot anomalies.
    - Test hypotheses.
    - Check assumptions.
- Tools Used:
  - Statistical measures and visualizations.



Learning Goals

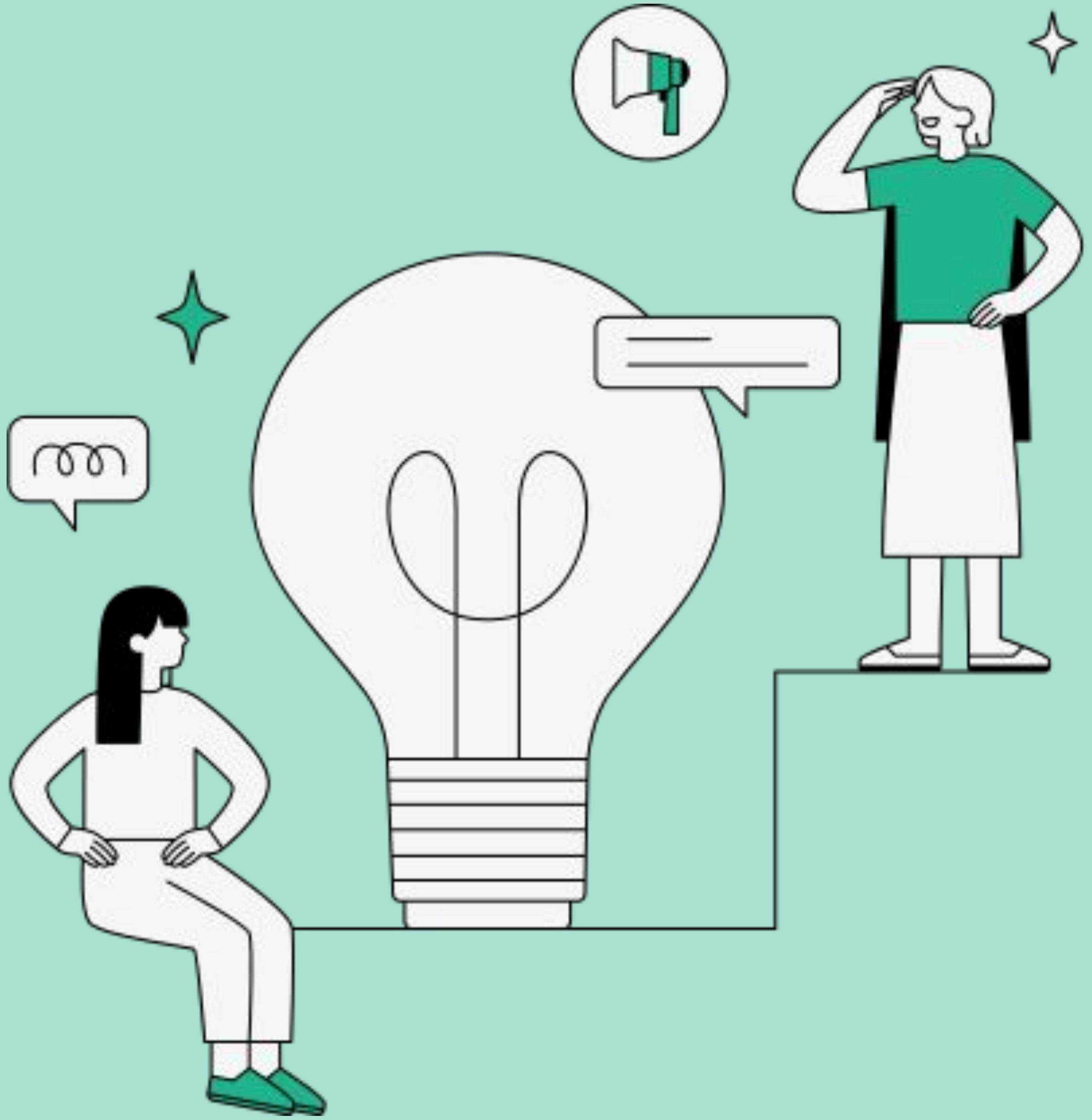
In this chapter, we will learn and revise the following topics:

- Understanding data science
- The significance of EDA
- Making sense of data
- Comparing EDA with classical and Bayesian analysis
- Getting started with EDA



Overview of EDA Stages

1. Data Requirements
2. Data Collection
3. Data Processing
4. Data Cleaning
5. Exploratory Data Analysis
6. Modeling and Algorithm Development
7. Data Product Creation
8. Communication



1. Data Requirements

- Sources of Data: Data can come from multiple sources like sensors, databases, or user interactions.
- Example: Application for monitoring dementia patients' sleep patterns:
- Types of Data Required:
  - Sleep data
  - Heart rate
  - Electro-dermal activity
  - User activity patterns
- Importance: Correct diagnosis requires all these data points to be accurately collected and stored.
- Categorizing Data:
  - Numerical Data: Continuous data like heart rate or sleep duration.
  - Categorical Data: Data like activity patterns or types of sleep stages.
- Data Storage and Dissemination:
  - Define storage formats (e.g., tables, databases).
  - Ensure accessibility and proper dissemination for analysis.

2. Data Collection

- Multiple Sources: Data is gathered from various objects and events, such as sensors, devices, and user interactions.
- Correct Storage Format:
  - Data must be stored in the appropriate format for effective analysis (e.g., CSV, JSON, SQL databases).
  - Formats depend on the type of data and tools used for analysis.
- Transfer to IT Personnel:
  - Collected data must be transferred to the right personnel for further processing and handling.
  - Ensures proper data management and security during storage and transit.
- Types of Tools Used for Data Collection
  - Sensors:
    - IoT devices, wearables, environmental sensors
  - Storage Tools:
    - Databases (SQL/NoSQL), cloud storage systems, local storage

3. Data Processing

- Preprocessing:
  - The initial step in preparing data for analysis.
  - Ensures data is organized, clean, and ready for exploration.
- Importance of Pre-curation:
  - Reduces inconsistencies and errors in data.
  - Enhances the quality and reliability of analysis outcomes.
- Common Tasks in Data Processing
  - Exporting the Dataset:
    - Ensure data is exported correctly from its source (e.g., database, API, files).
  - Organizing Data into Tables:
    - Place data in appropriate tables or data frames for easy access and analysis.
  - Structuring the Data:
    - Arrange data in a logical and analyzable format (e.g., rows for records, columns for attributes).
  - Exporting in the Correct Format:
    - Save the processed data in widely used formats such as CSV, Excel, or JSON for compatibility with analysis tools.

4. Data Cleaning

- What is Data Cleaning?
  - The process of transforming preprocessed data to prepare it for detailed analysis.
  - Ensures the dataset is complete, accurate, and free from errors.
- Key Tasks in Data Cleaning
  - Incompleteness Check:
    - Identify and address missing values in the dataset.
  - Duplicates Check:
    - Detect and remove redundant records.
  - Error Check:
    - Spot inaccuracies or inconsistencies in the data.
  - Missing Value Handling:
    - Fill in or impute missing values to maintain dataset integrity.
- Analytical Techniques for Identifying Anomalies
  - Use statistical and visualization tools to detect data issues like outliers, inconsistencies, or gaps.
- Example of Data Cleaning
  - Outlier Detection: Apply quantitative methods (e.g., z-score, IQR) to identify and handle extreme values that may skew analysis.

5. Exploratory Data Analysis (EDA)

What is EDA?

- A crucial stage in data analysis where the dataset is explored to understand its underlying patterns and insights.
- Involves identifying trends, spotting anomalies, testing hypotheses, and validating assumptions.

Key Aspects of EDA

- Understanding the Data:
  - Gain a clear picture of the dataset's structure, variables, and relationships.
- Data Transformation:
  - Apply necessary techniques (e.g., normalization, scaling) to prepare data for analysis.
- Statistical Measures:
  - Use descriptive statistics to summarize data (e.g., mean, median, mode, variance).
- Why EDA Matters
  - Helps in uncovering hidden patterns and insights.
  - Assists in feature selection for modeling.
  - Provides a foundation for building robust and accurate data models.

6. Modeling and Algorithm

What is Modeling?

- Modeling involves creating mathematical formulas to describe relationships between variables.
- These relationships can show correlation or causation.

Key Concepts

1. Independent and Dependent Variables:
    - Independent Variable: Input factor influencing the outcome.
    - Dependent Variable: Outcome influenced by independent variables.
    - Example:
        - Total Price = Unit Price × Quantity.
        - Total Price = Dependent Variable.
        - Unit Price and Quantity = Independent Variables.
2. The Judd Model:
    - Data = Model + Error.
    - Emphasizes that models include inherent errors or uncertainties.

- Inferential Statistics in Modeling
    - Quantifies relationships between variables.
    - Example: Regression Analysis
        - Helps predict dependent variable outcomes based on independent variables.

7. Data Product

What is a Data Product?

- A data product is a computer software system that:
    - Uses data as input.
    - Produces output based on analysis.
    - Provides feedback to control or improve the environment.
- Key Characteristics:
    - Built upon models developed during data analysis.
    - Designed to automate decision-making or improve user experience.
- Example:
    - Recommendation Model:
        - Inputs: User purchase history.
        - Output: Suggests related items the user is likely to buy.
- Purpose of Data Products:
    - Enhance business intelligence.
    - Drive user engagement and satisfaction.
    - Provide actionable insights.

8. Communication

What is Communication in EDA?

- The stage where results from exploratory data analysis are shared with stakeholders.
- Helps in converting analysis into actionable business intelligence.

Key Aspect:

- Data Visualization:
    - The primary tool for communicating findings.
    - Uses visual aids to present analyzed data clearly and effectively.

Common Visualization Techniques:

- Tables
- Charts
- Summary Diagrams
- Bar Charts

Purpose of Visualization:

- Simplifies complex data for easier interpretation.
- Provides clarity and insights to drive informed decision-making.

Significance of Exploratory Data Analysis (EDA)

What is EDA?

- Exploratory Data Analysis (EDA) is the initial phase in the data mining process.
- It focuses on exploring and understanding datasets before applying more complex statistical models.
- EDA is crucial for generating hypotheses and visualizing the data to grasp patterns, trends, and relationships within the data.
- It serves as a preliminary investigation to decide which modeling techniques or methods are appropriate for the data.

Why is EDA Significant?

- Understanding Data:
  - Without proper exploration, making sense of large datasets with numerous variables becomes difficult.
  - EDA helps uncover hidden insights about the data, including distributions, patterns, and correlations.
  - By performing EDA, we get a clear picture of the ground truth of the data, enabling informed decision-making.
- Data-Driven Decision Making:
  - EDA is essential for making well-informed, data-driven decisions in various fields such as business, healthcare, and finance.
  - By using EDA techniques, we ensure that the dataset is properly understood and processed before moving forward in the analysis.
- Hypothesis Creation:
  - It allows us to form hypotheses or assumptions based on initial findings, which can later be tested with more sophisticated modeling.
  - Hypotheses generated in this phase serve as a foundation for further data mining and predictive analysis.

Significance of Exploratory Data Analysis (EDA)

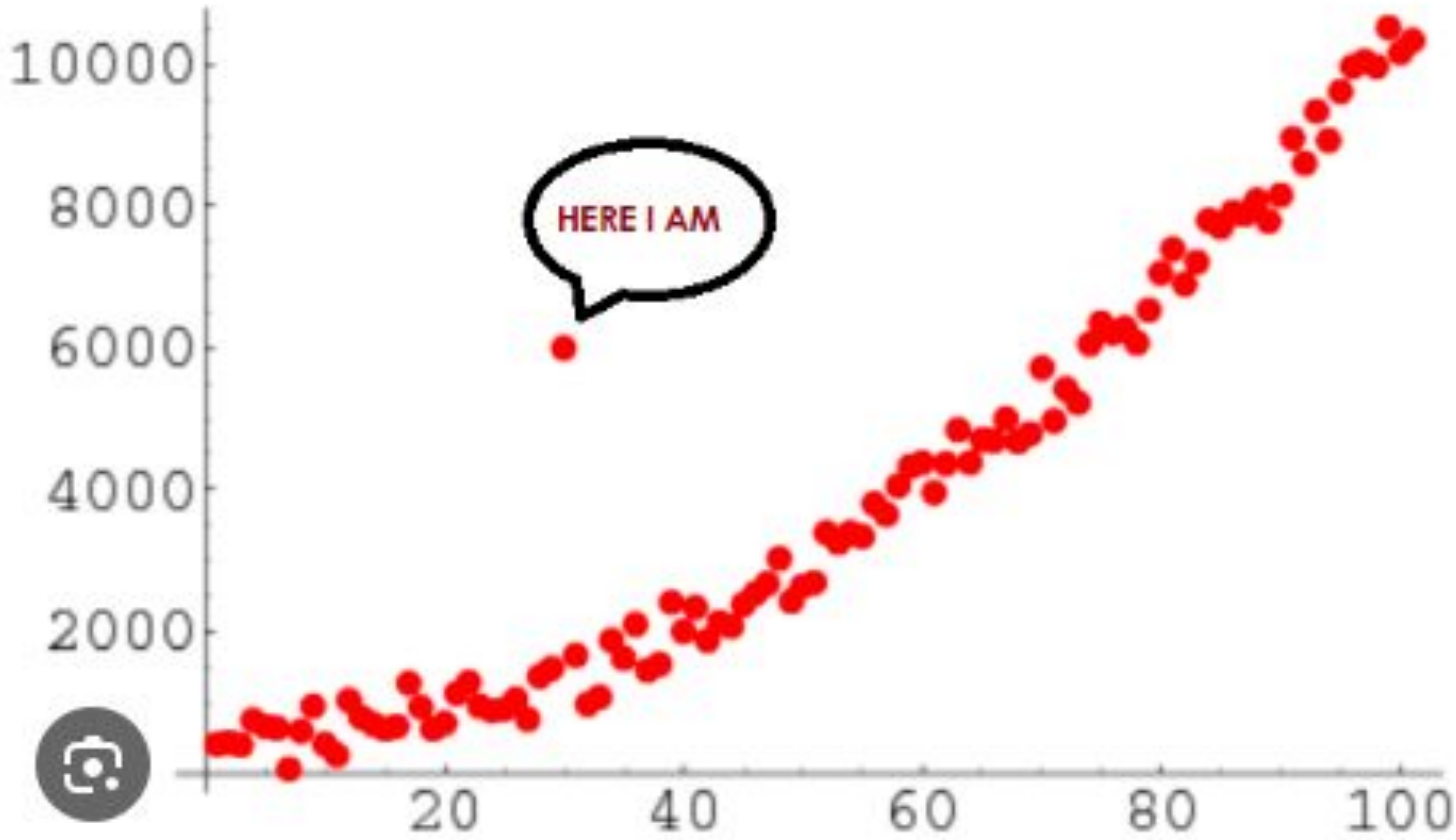
Key Components of EDA

1. Summarizing Data:
   - Descriptive statistics like mean, median, mode, variance, and standard deviation are used to summarize the key characteristics of the data.
   - Summarization helps to detect any outliers or anomalies early on, aiding in further cleaning or transformation tasks.
2. Statistical Analysis:
   - Basic statistical tests and analysis, such as hypothesis testing, correlation analysis, and regression analysis, are conducted during EDA.
   - These tests help to uncover any statistical relationships between variables, supporting further hypotheses for deeper analysis.
3. Data Visualization:
   - Visualization techniques such as bar charts, histograms, box plots, scatter plots, and heatmaps help illustrate the data's distributions, relationships, and patterns.
   - Visualization makes it easier to identify trends, clusters, and outliers, which may be difficult to detect through raw data alone.

Significance of Exploratory Data Analysis (EDA)

Tools for EDA

- Python provides powerful tools for performing EDA:
  - pandas: A library used for data summarization (mean, median, standard deviation, etc.), and manipulation, such as data cleaning and transformation.
  - scipy: A library that supports statistical analysis, hypothesis testing, and more advanced statistical functions.
  - matplotlib & plotly: Libraries that help create a variety of visualizations to reveal the underlying structure of data, such as charts, graphs, and plots.


HERE I AM
10000
8000
6000
4000
2000
20
40
60
80
100

Steps in EDA

4 Different steps involved in data analysis are:

1. Problem Definition
2. Data preparation
3. Data analysis
4. Development and representation of the results

1. Problem Definition: Define the business problem to solve.
   - Define the main objective of the analysis.
   - Identify deliverables.
   - Outline roles and responsibilities.
   - Assess current status of the data.
   - Set a timetable and perform cost/benefit analysis.

   Purpose: The problem definition drives the execution plan for data analysis.
2. Data preparation: Prepare the dataset for analysis.
   - Define data sources and schema.
   - Understand the characteristics of the data.
   - Clean the dataset (remove irrelevant data, handle missing values).
   - Transform and divide data into required chunks.

   Purpose: Ensure data is well-organized and ready for analysis.

Steps in EDA

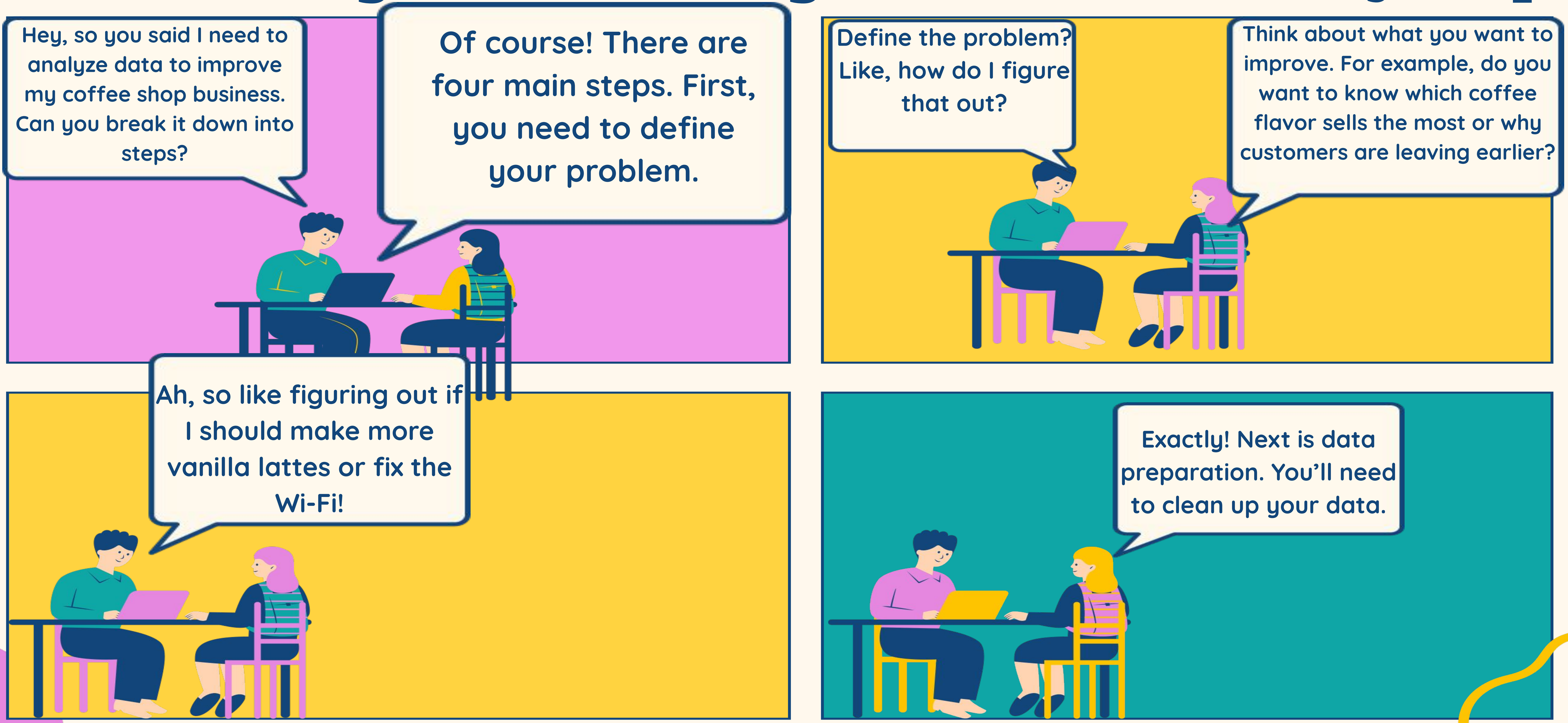
3. Development and representation of the results: Analyze the data using various techniques.
    - Summarize the data.
    - Identify correlations and relationships.
    - Develop and evaluate predictive models.
    - Calculate model accuracies.

- Techniques:
    - Descriptive statistics, correlation, inferential statistics.
    - Mathematical models, grouping, and searching.
- Purpose: Extract meaningful insights and patterns.

4. Development & Representation of Results:

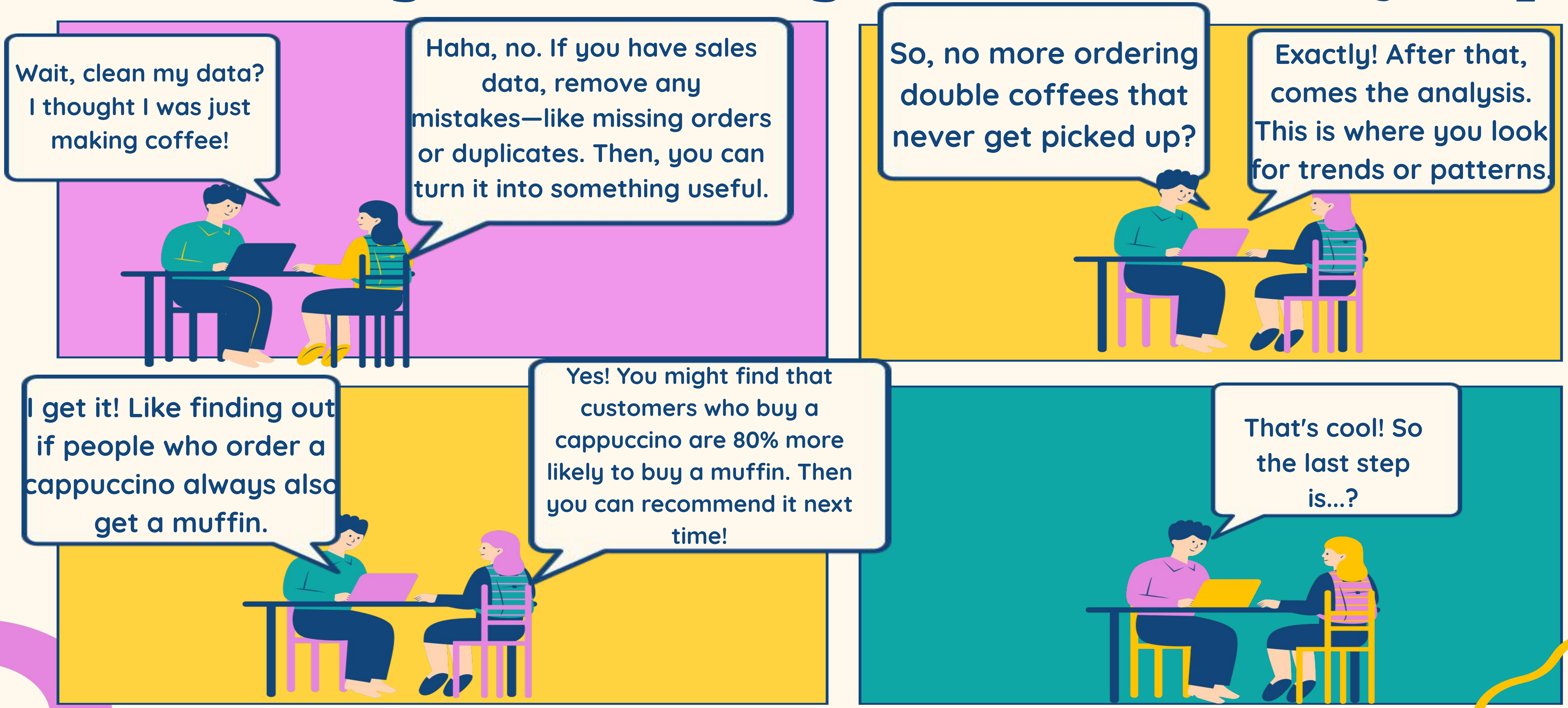
- Objective: Present the findings in an understandable format for stakeholders.
- Key Tasks:
    - Create visualizations: graphs, tables, maps, diagrams.
    - Communicate results effectively for decision-making.
- Types of Graphical Representations:
    - Scatter plots, histograms, box plots, residual plots.
- Purpose: Ensure the results are interpretable and actionable.

Caffeinating Data: Brewing Success with Every Step

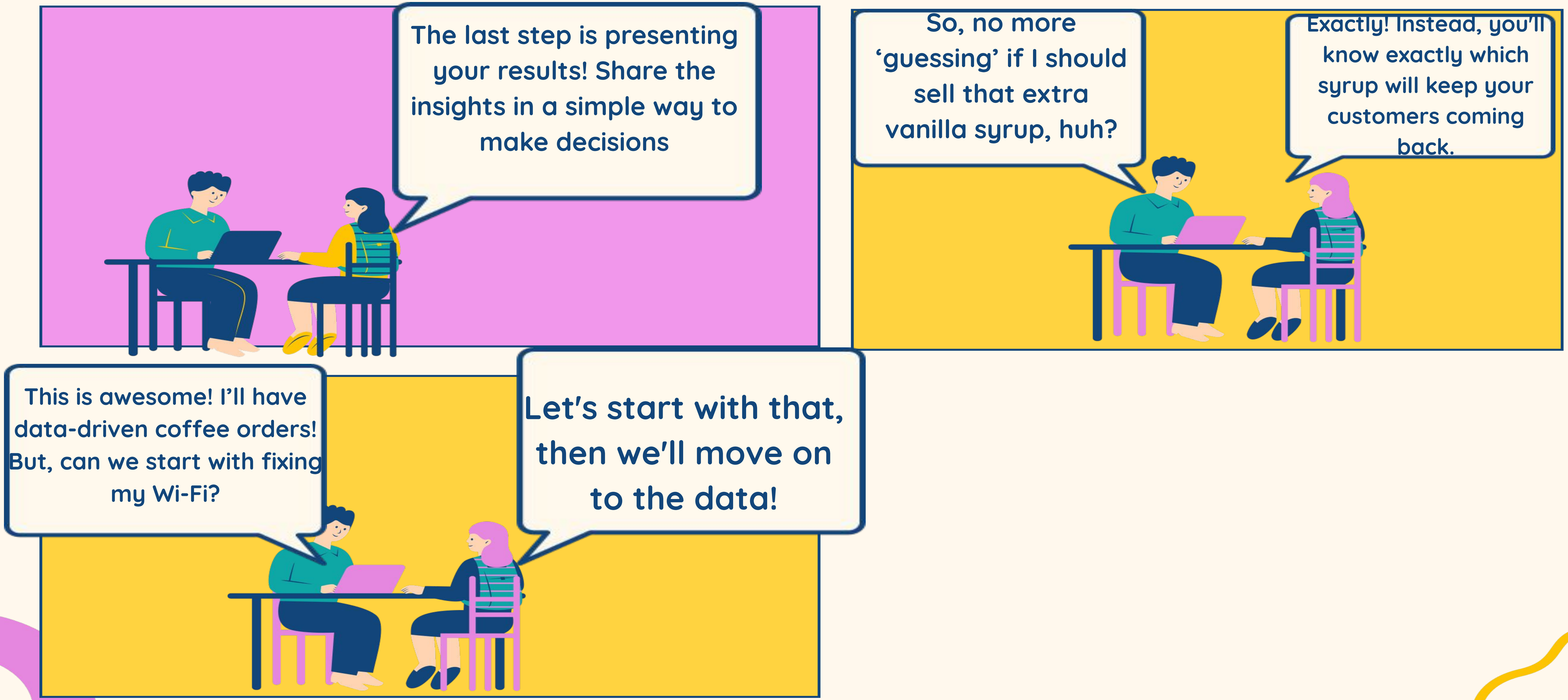

Hey, so you said I need to analyze data to improve my coffee shop business. Can you break it down into steps?
Of course! There are four main steps. First, you need to define your problem.
Define the problem? Like, how do I figure that out?
Think about what you want to improve. For example, do you want to know which coffee flavor sells the most or why customers are leaving earlier?
Ah, so like figuring out if I should make more vanilla lattes or fix the Wi-Fi!
Exactly! Next is data preparation. You'll need to clean up your data.

INSPIRED BY mathspace

Caffeinating Data: Brewing Success with Every Step


Wait, clean my data? I thought I was just making coffee!
Haha, no. If you have sales data, remove any mistakes—like missing orders or duplicates. Then, you can turn it into something useful.
So, no more ordering double coffees that never get picked up?
Exactly! After that, comes the analysis. This is where you look for trends or patterns.
I get it! Like finding out if people who order a cappuccino always also get a muffin.
Yes! You might find that customers who buy a cappuccino are 80% more likely to buy a muffin. Then you can recommend it next time!
That's cool! So the last step is...?

Caffeinating Data: Brewing Success with Every Step


The last step is presenting your results! Share the insights in a simple way to make decisions
So, no more ‘guessing’ if I should sell that extra vanilla syrup, huh?
Exactly! Instead, you'll know exactly which syrup will keep your customers coming back.
This is awesome! I’ll have data-driven coffee orders! But, can we start with fixing my Wi-Fi?
Let's start with that, then we'll move on to the data!

Making Sense of Data

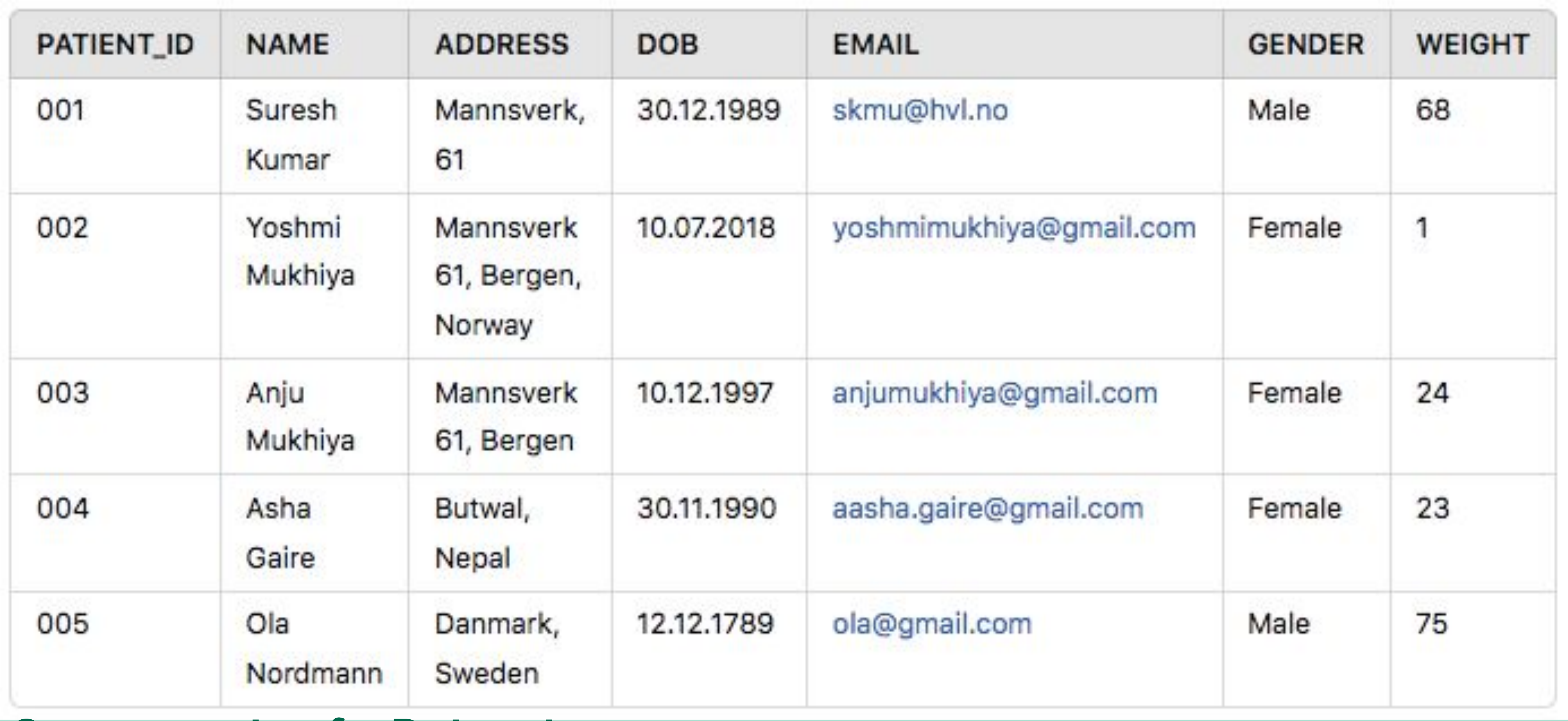
- Importance of Identifying Data Types:
  - Different disciplines store and analyze data for different purposes:
    - Medical researchers: Patients' data.
    - Universities: Students' and teachers' data.
    - Real estate: House and building data.
  - Understanding the type of data is critical for proper analysis.
- Dataset Structure:
  - A dataset consists of multiple observations about specific objects or entities.
  - Example: A hospital dataset containing patient information.
  - Each observation has several variables or features that describe the object.
  - Example: A patient described by Patient_ID, Name, Address, DOB, Email, Weight, Gender.
- Example of Patient Data:
  - Patient Information:
    - PATIENT_ID: Unique identifier for the patient.
    - Name, Address, Date of Birth (DOB), Email, Gender, Weight: Descriptive variables.
- Example record:

"PATIENT_ID = 1001, Name = Yoshmi Mukhiya, Address = Mannsverk 61, 5094, Bergen, Norway, DOB = 10th July 2018, Email = yoshmimukhiya@gmail.com, Weight = 10, Gender = Female"

Making Sense of Data

Dataset Table Example:

- The dataset is stored in tables/schema in database management systems.
- Example Table Structure for Storing Patient Information:



PATIENT_ID	NAME	ADDRESS	DOB	EMAIL	GENDER	WEIGHT
001	Suresh Kumar	Mannsverk, 61	30.12.1989	skmu@hvl.no	Male	68
002	Yoshmi Mukhiya	Mannsverk 61, Bergen, Norway	10.07.2018	yoshmimukhiya@gmail.com	Female	1
003	Anju Mukhiya	Mannsverk 61, Bergen	10.12.1997	anjumukhiya@gmail.com	Female	24
004	Asha Gaire	Butwal, Nepal	30.11.1990	aasha.gaire@gmail.com	Female	23
005	Ola Nordmann	Danmark, Sweden	12.12.1789	ola@gmail.com	Male	75

Key Components of a Dataset:

- Observations: Rows in the dataset (e.g., each patient record).
- Variables: Columns describing the data (e.g., Patient ID, Name, Weight).

Data Types:

- Numerical Data: Quantitative variables like weight, age, etc.
- Categorical Data: Qualitative variables like gender, blood type, etc

Numerical Data

Definition of Numerical Data:

- **Numerical data involves measurable quantities and has a sense of measurement attached to it.**
- **Examples of numerical data:**
    - **Person's age, height, weight, blood pressure, heart rate, temperature, number of teeth, number of bones, number of family members.**
- **Also referred to as quantitative data in statistics.**
- **Types of Numerical Data:**
    - **Numerical data can be broadly classified into two types: Discrete Data and Continuous Data.**
        - **Discrete Data:**
            - **Data that is countable and whose values can be listed out.**
            - **Values are finite and distinct.**
            - **Examples:**
                - **Number of heads in 200 coin flips (values from 0 to 200).**
                - **Number of students in a class.**
            - **Discrete Variable: A variable that can only take distinct, countable values.**
                - **Example: The Rank of a student in a classroom (1, 2, 3, etc.).**
                - **Example: The Country a person belongs to (Nepal, India, Norway, Japan).**

Numerical Data

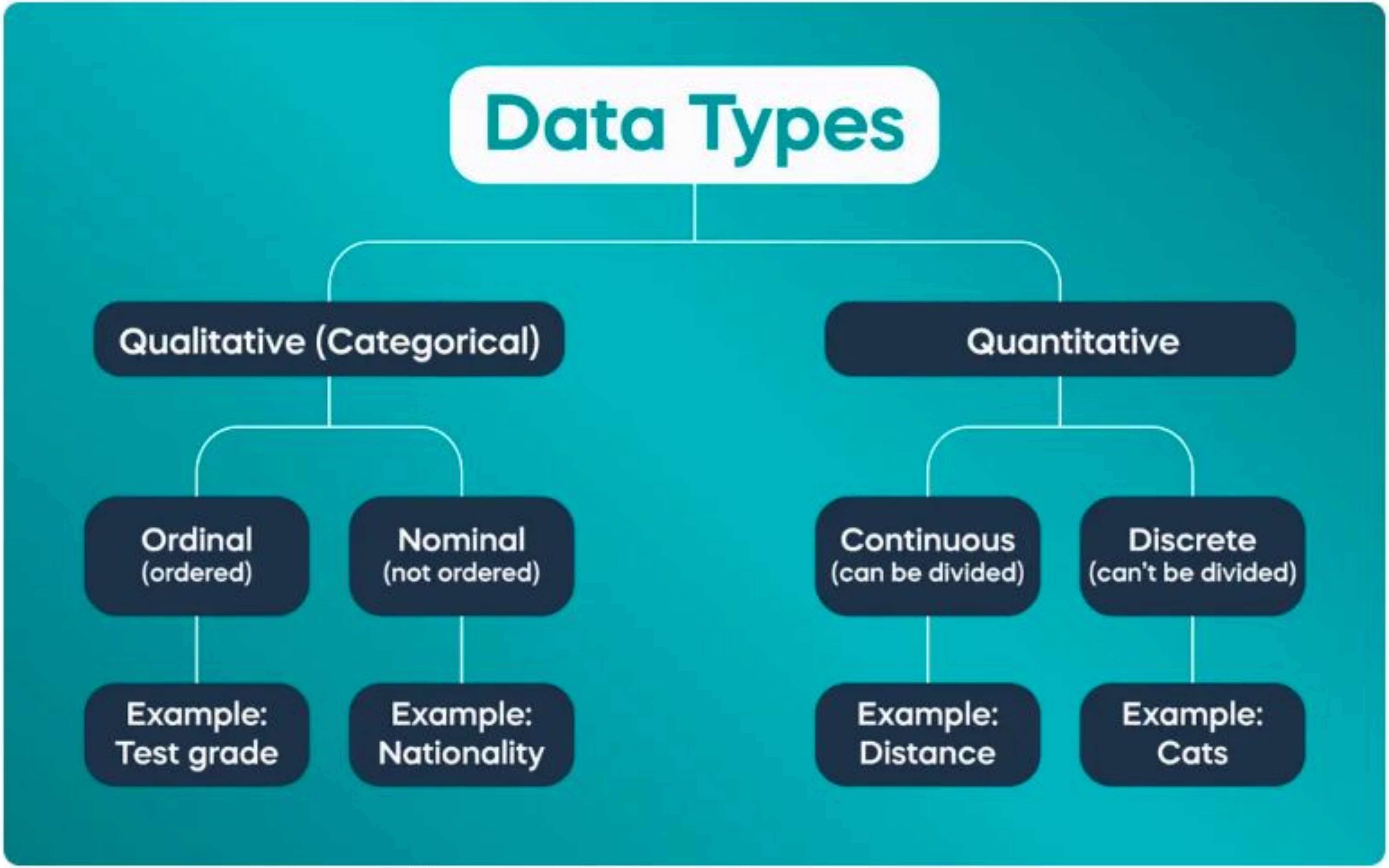
- Continuous Data:
  - Data that can take an infinite number of values within a specific range.
  - The values are not countable and can be measured with higher precision.
  - Examples:
    - Temperature in a city (can have infinite decimal values within a specific range).
    - Weight (can have infinite possible values like 50.1 kg, 50.12 kg, etc.).
  - Continuous Variable: A variable that can take any numerical value within a range.
  - Example: The Temperature of a city today, which could have infinitely many possible values.

Categorical Data

- Categorical data represents the characteristics of an object or phenomenon.
- It involves non-numeric information and is often referred to as qualitative data in statistics.
- Examples of categorical data:
    - Gender: Male, Female, Other, Unknown
    - Marital Status: Annulled, Divorced, Interlocutory, Legally Separated, Married, Polygamous, Never Married, Domestic Partner, Unmarried, Widowed, Unknown
    - Movie Genres: Action, Adventure, Comedy, Crime, Drama, Fantasy, Historical, Horror, Mystery, Philosophical, Political, Romance, Saga, Satire, Science Fiction, Social, Thriller, Urban, Western
    - Blood Type: A, B, AB, O
    - Types of Drugs: Stimulants, Depressants, Hallucinogens, Dissociatives, Opioids, Inhalants, Cannabis
- Categorical Variable:
    - A variable that describes categorical data is called a categorical variable.
    - Categorical variables can take one of a limited number of values or categories.
- Types of Categorical Variables:
    - Binary Categorical Variable:
        - A variable that can take exactly two values.
        - Also referred to as a dichotomous variable.
        - Example: Experiment results (Success, Failure).

Categorical Data

- Polytomous Variable:
    - A variable that can take more than two values.
    - Example: Marital Status (Annulled, Divorced, Married, Widowed, etc.).
- Measurement Scales for Categorical Data:
    - Most categorical datasets follow either Nominal or Ordinal measurement scales.
- Understanding Nominal and Ordinal Scales:
    - Nominal Scale: Categorical variables with no inherent order (e.g., Gender, Movie Genres).
    - Ordinal Scale: Categorical variables with a defined order but no precise numeric difference between categories (e.g., Marital Status categories).



Data Types

Qualitative (Categorical)

Ordinal
(ordered)

Example:
Test grade

Nominal
(not ordered)

Example:
Nationality

Quantitative

Continuous
(can be divided)

Example:
Distance

Discrete
(can't be divided)

Example:
Cats

Measurement Scales

Introduction to Measurement Scales:

- There are four types of measurement scales in statistics: Nominal, Ordinal, Interval, and Ratio.
- These scales are essential in academic research and data analysis, as they help determine the nature of data and the appropriate statistical methods.

Nominal Scale:

- Used for labeling variables without any quantitative value.
- Key Characteristics:
  - Categories are mutually exclusive.
  - No numerical importance or order.
- Examples:
  - Gender: Male, Female, Third gender/Non-binary, Prefer not to answer
  - Languages spoken in a country
  - Biological species
  - Parts of speech in grammar (noun, verb, adjective)
  - Taxonomic ranks in biology (Archea, Bacteria, Eukarya)
- Data Type: Qualitative.

Important Note: Numbers used as labels in nominal measurement have no concrete numerical value or meaning. Arithmetic calculations cannot be performed.

Measurement Scales

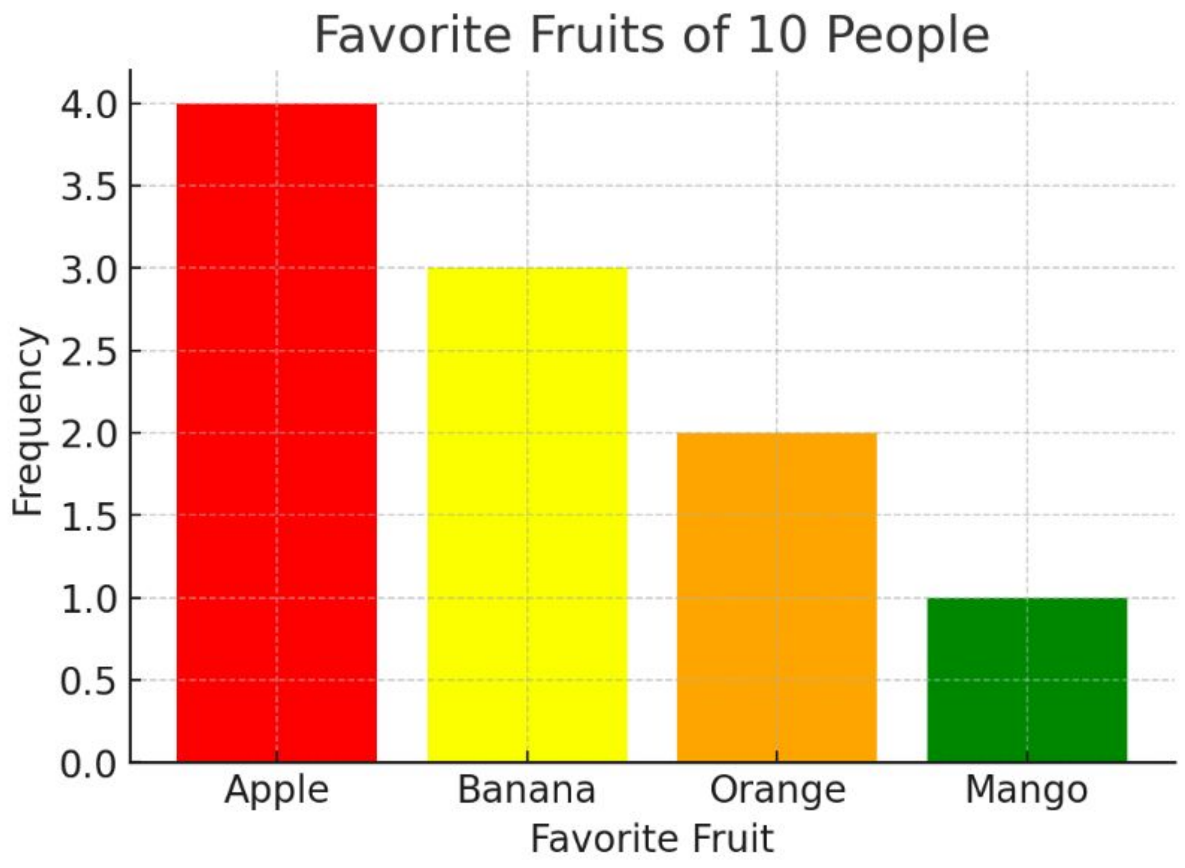
- Why Understanding Nominal Data is Important:
    - Frequency: Rate at which a label occurs in a dataset.
    - Proportion: Calculated by dividing the frequency by the total number of occurrences.
    - Visualization: Nominal data can be visualized using pie charts or bar charts.
- Why This Matters:
    - Understanding the type of data helps in:
        - Deciding which type of computation to perform.
        - Selecting the appropriate model to fit on the data.
        - Determining the visualization method.
- Example: For nominal data, pie charts or bar charts are ideal visualizations.

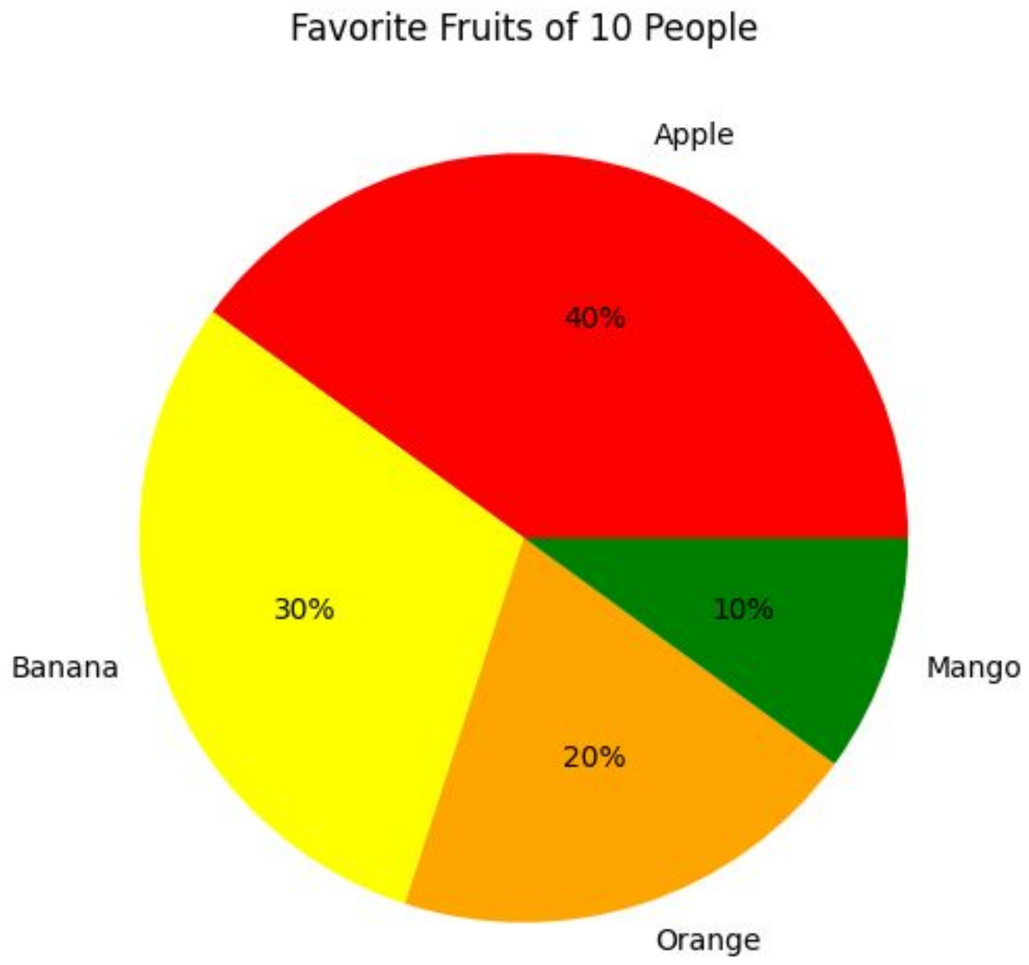
Nominal Scale - Example

Person	Fav Fruit
1	Apple
2	Banana
3	Apple
4	Orange
5	Banana
6	Mango
7	Apple
8	Orange
9	Banana
10	Apple

Fav. Fruit	Frequency
Apple	4
Banana	3
Orange	2
Mango	1

Fav. Fruit	Proportion
Apple	4/10 = 0.4 (40%)
Banana	3/10 = 0.3 (30%)
Orange	2/10 = 0.2 (20%)
Mango	1/10 = 0.1 (10%)



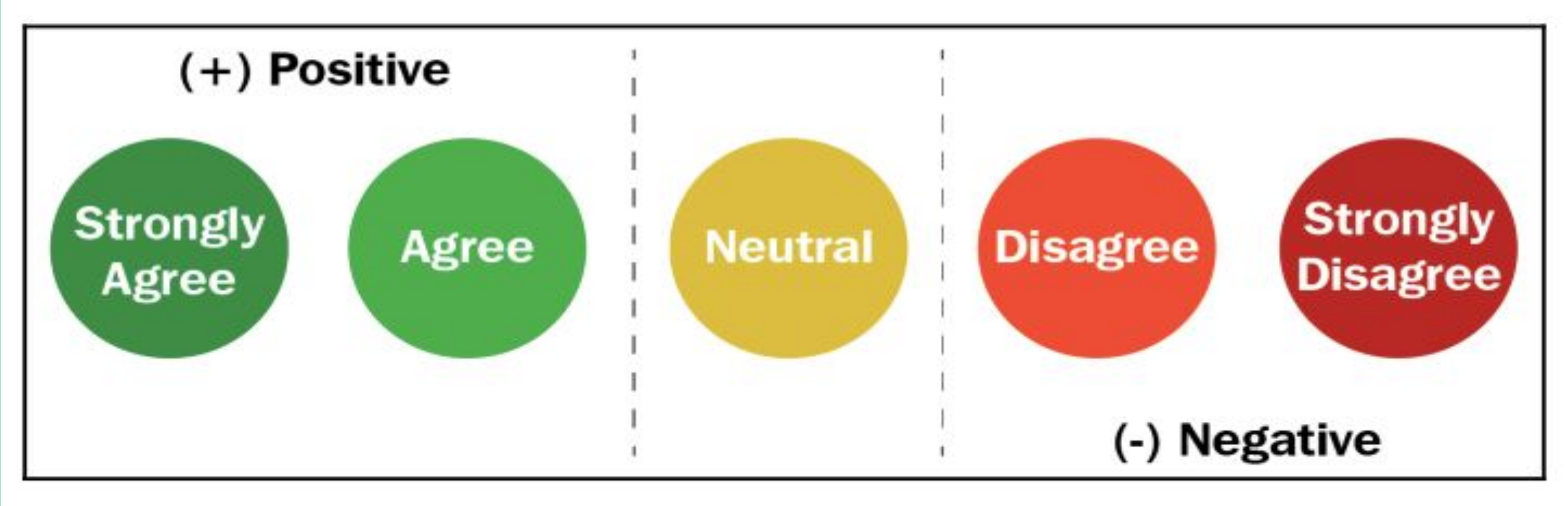


Ordinal Scale

The ordinal scale involves data that have a defined order, but the differences between the values are not meaningful.

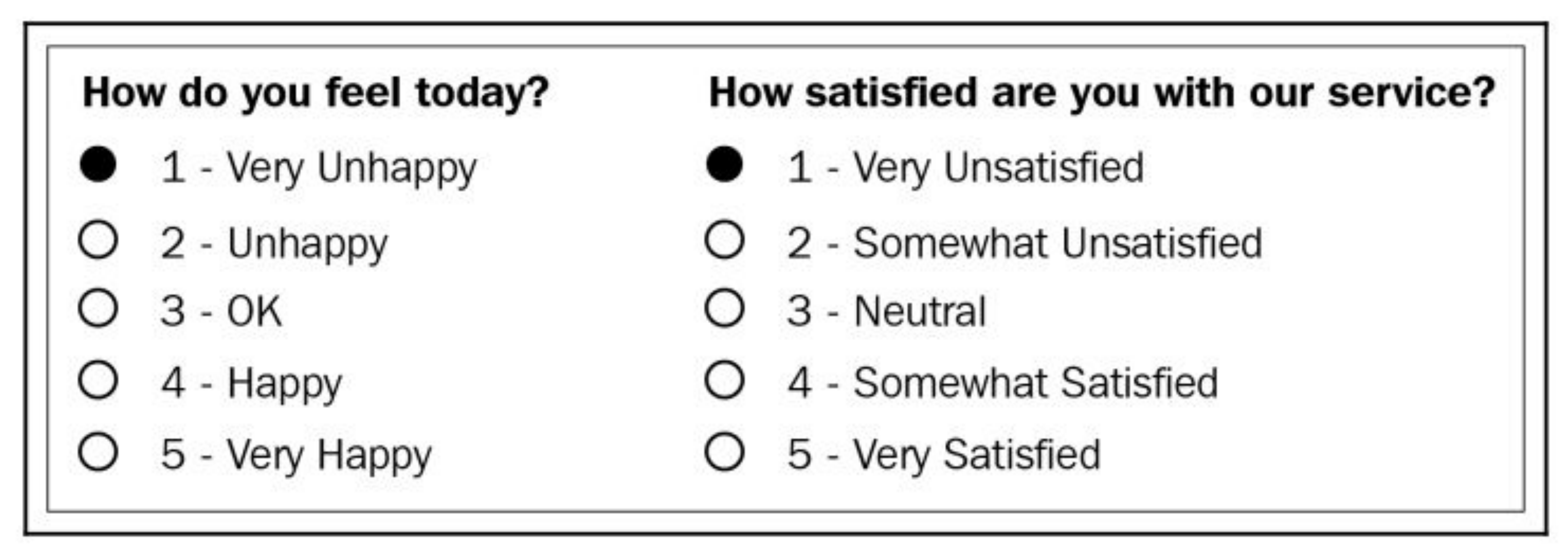
Key Characteristics:

- Order matters: The sequence of values represents a ranking or hierarchy.
- No meaningful difference between consecutive values (i.e., the difference between 1st and 2nd may not be the same as the difference between 3rd and 4th).
- Example: Likert Scale:
    - A common example of an ordinal scale is the Likert scale, used in surveys to measure people's opinion.



- These responses represent a ranked order, but the difference between, say, "Strongly Agree" and "Agree" isn't numerically defined.

Ordinal Scale



How do you feel today?

● 1 - Very Unhappy

○ 2 - Unhappy

○ 3 - OK

○ 4 - Happy

○ 5 - Very Happy

How satisfied are you with our service?

● 1 - Very Unsatisfied

○ 2 - Somewhat Unsatisfied

○ 3 - Neutral

○ 4 - Somewhat Satisfied

○ 5 - Very Satisfied

- **Characteristics of Ordinal Data:**
  - **Order of Values: Values are ranked or ordered (1st, 2nd, 3rd, etc.).**
  - **Median: The median can be used to measure central tendency.**
  - **Mean: The average (mean) is not suitable, as the data is not uniformly spaced.**
- **Why Understanding Ordinal Data is Important:**
  - **Knowing that your data follows an ordinal scale helps you choose the right statistical analysis and visualization.**
  - **For ordinal data, you should avoid calculations like averages and focus on the median to understand central tendency.**

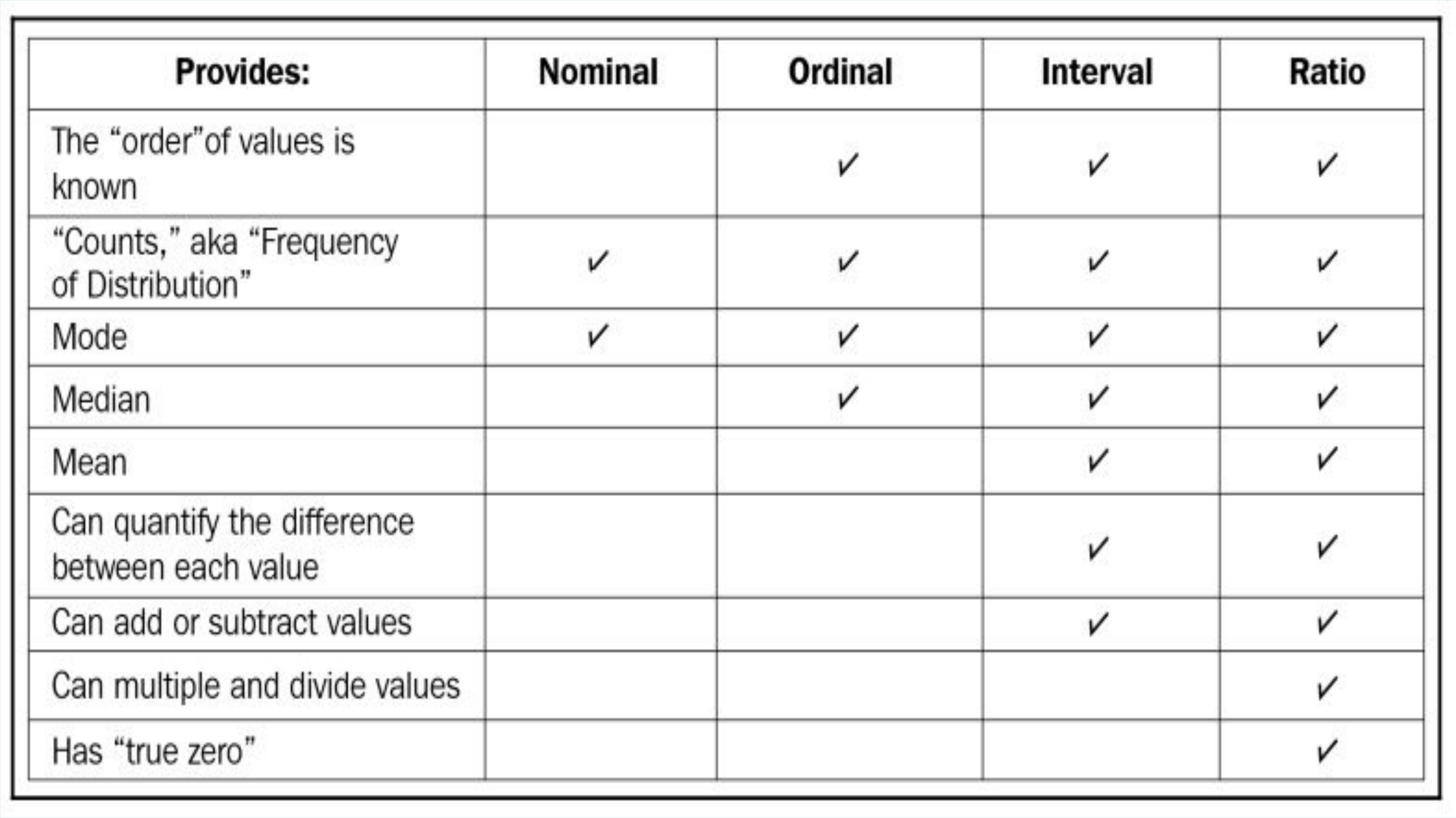
Interval

- The interval scale has both order and exact differences between values, but it does not have an absolute zero.
- Key Characteristics:
    - Order: Values have a specific rank.
    - Exact differences: The difference between any two consecutive values is meaningful.
    - No Absolute Zero: The scale doesn't have a true zero point, meaning you cannot make ratios (e.g., you cannot say 20°C is twice as hot as 10°C).
- Examples:
    - Temperature (in Celsius or Fahrenheit): The difference between 10°C and 20°C is the same as between 20°C and 30°C, but 0°C does not represent the absence of temperature.
    - Calendar Dates: The difference between January 1st and January 2nd is meaningful, but the zero point (start of the calendar) is arbitrary.
- Statistical Measures:
    - You can compute mean, median, and mode for interval data.
    - Standard deviation can also be calculated.

Ratio

- Definition:
  - The ratio scale has order, exact differences, and also includes an absolute zero. This allows for the full range of mathematical operations.
- Key Characteristics:
  - Order and Exact Differences: As in the interval scale, the values have a defined order, and the differences between them are meaningful.
  - Absolute Zero: There is a true zero, meaning zero indicates the complete absence of the measured attribute. This allows for meaningful ratios.
- Examples:
  - Mass: Zero mass means no mass at all.
  - Length: Zero length means no length at all.
  - Time Duration: Zero time means no elapsed time.
  - Energy, Volume, Electrical Current: Zero indicates complete absence of the quantity.
- Statistical Measures:
  - All basic operations (addition, subtraction, multiplication, division) are valid for ratio scales.
  - Mean, median, mode, and standard deviation can all be calculated.
  - Coefficient of Variation can also be computed.

Ratio



Provides:	Nominal	Ordinal	Interval	Ratio
The “order”of values is known		✓	✓	✓
“Counts,” aka “Frequency of Distribution”	✓	✓	✓	✓
Mode	✓	✓	✓	✓
Median		✓	✓	✓
Mean			✓	✓
Can quantify the difference between each value			✓	✓
Can add or subtract values			✓	✓
Can multiple and divide values				✓
Has “true zero”				✓

Comparing EDA, Classical, and Bayesian Data Analysis

Classical Data Analysis

- Relies heavily on pre-defined models.
- Assumes data follows a specific distribution or pattern.
- Focus is on hypothesis testing and statistical inference.

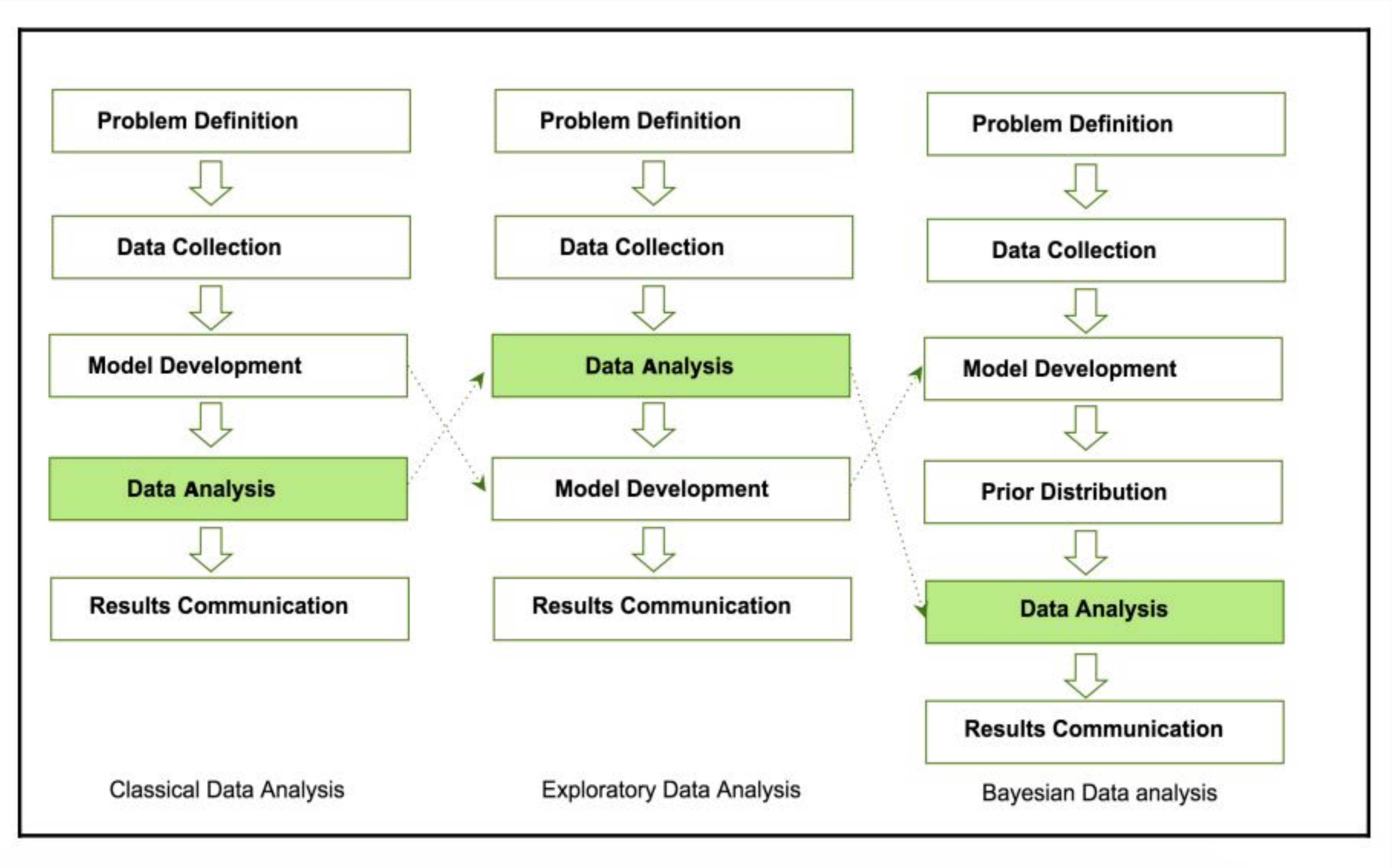
Exploratory Data Analysis (EDA)

- Focuses on understanding the data's structure, patterns, and anomalies.
- Uses visualizations (e.g., histograms, scatter plots) extensively.
- Avoids imposing deterministic or probabilistic models initially.
- Aims to explore data and generate hypotheses rather than confirm them.

Bayesian Data Analysis:

- Integrates prior beliefs (prior probability distributions) with the data.
- Focuses on updating these priors using evidence to compute the posterior probability distribution.
- Provides a probabilistic framework for decision-making.
- Useful in scenarios with limited data or when prior knowledge is crucial.

Comparing EDA with classical and Bayesian analysis



Getting started with EDA

Why Python is the main tool for data analysis?

- Popularity and Adoption
- Ease of Use
- Rich Ecosystem of Libraries
- Versatility
- Active Community Support
- Cross-Platform Compatibility

Essential Python Tools and Packages for Data Analysis

Python Programming Basics

- Core concepts: Variables, strings, and data types
- Logic flow: Conditionals and functions
- Data structures: Sequences, collections, and iterations
- File handling: Reading and writing files
- Object-Oriented Programming (OOP): Classes, objects, and methods

NumPy

- Array creation: Initialize and copy arrays, divide arrays
- Array operations: Perform mathematical and logical operations
- Advanced indexing: Array selection and manipulation
- Multi-dimensional arrays: Work with matrices and tensors
- Linear algebra: Use built-in linear algebraic functions
- NumPy functions: Leverage a wide range of pre-defined functions

pandas

- DataFrames: Create and manipulate DataFrame objects
- Data subsetting: Indexing and filtering data
- Arithmetic and mapping: Apply operations across datasets
- Index management: Optimize and control DataFrame indexing
- Visual analysis: Customize DataFrame styling for better insights

Essential Python Tools and Packages for Data Analysis

Matplotlib

- Plotting basics: Load and visualize linear datasets
- Customization: Adjust axes, grids, labels, titles, and legends
- Exporting plots: Save visualizations in various formats

SciPy

- Statistical analysis: Use SciPy’s statistical packages
- Descriptive statistics: Summarize data distributions
- Inference and analysis: Perform hypothesis testing and data modeling