Fondamenti di Analisi dei Dati

from data analysis to predictive techniques

Prof. Antonino Furnari (antonino.furnari@unict.it)
Corso di Studi in Informatica
Dip. di Matematica e Informatica
Università di Catania

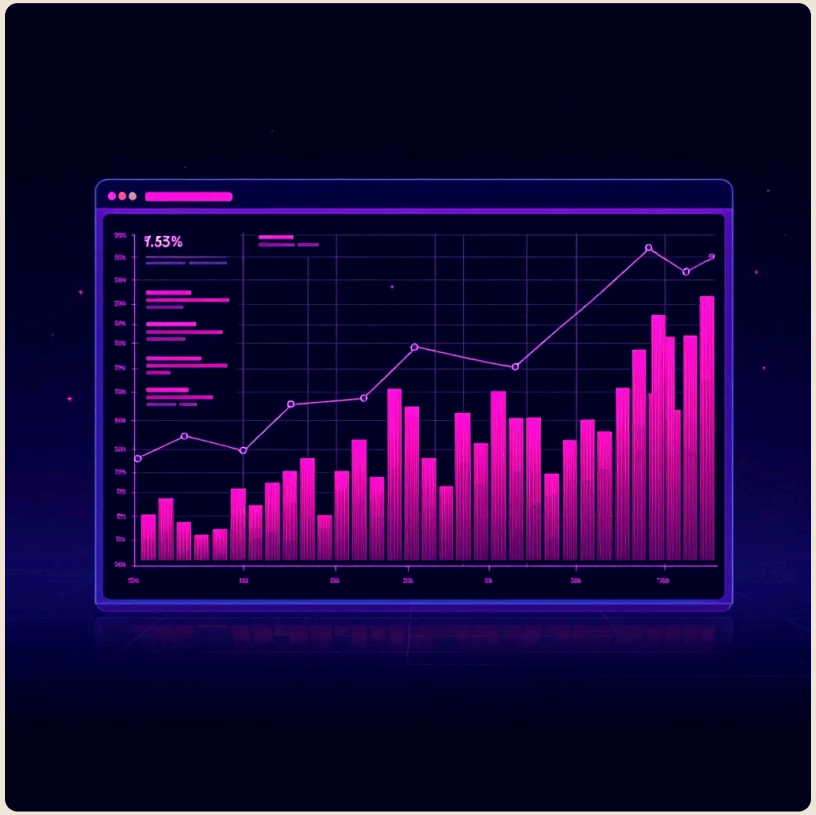
Key Concepts of Data Analysis

What is Data? An Informal Definition

Let's begin with an informal but fundamental definition:

> Data is a **collection of values** gathered with respect to certain **variables** that describe a given **phenomenon**.

This definition introduces key concepts that we will explore in the following sections: values, variables, and observable phenomena.



Observations, Populations, and Samples

Observations

The **fundamental units** with which we measure data. These can be people, automobiles, animals, or any entity we are studying.

Example: A specific student in a class

Population

The **complete set** of all observations we wish to study. This is often a theoretical concept.

Example: All computer science students in Italy

Sample

A **subset of the population** from which we actually collect data. This is what we analyse in practice.

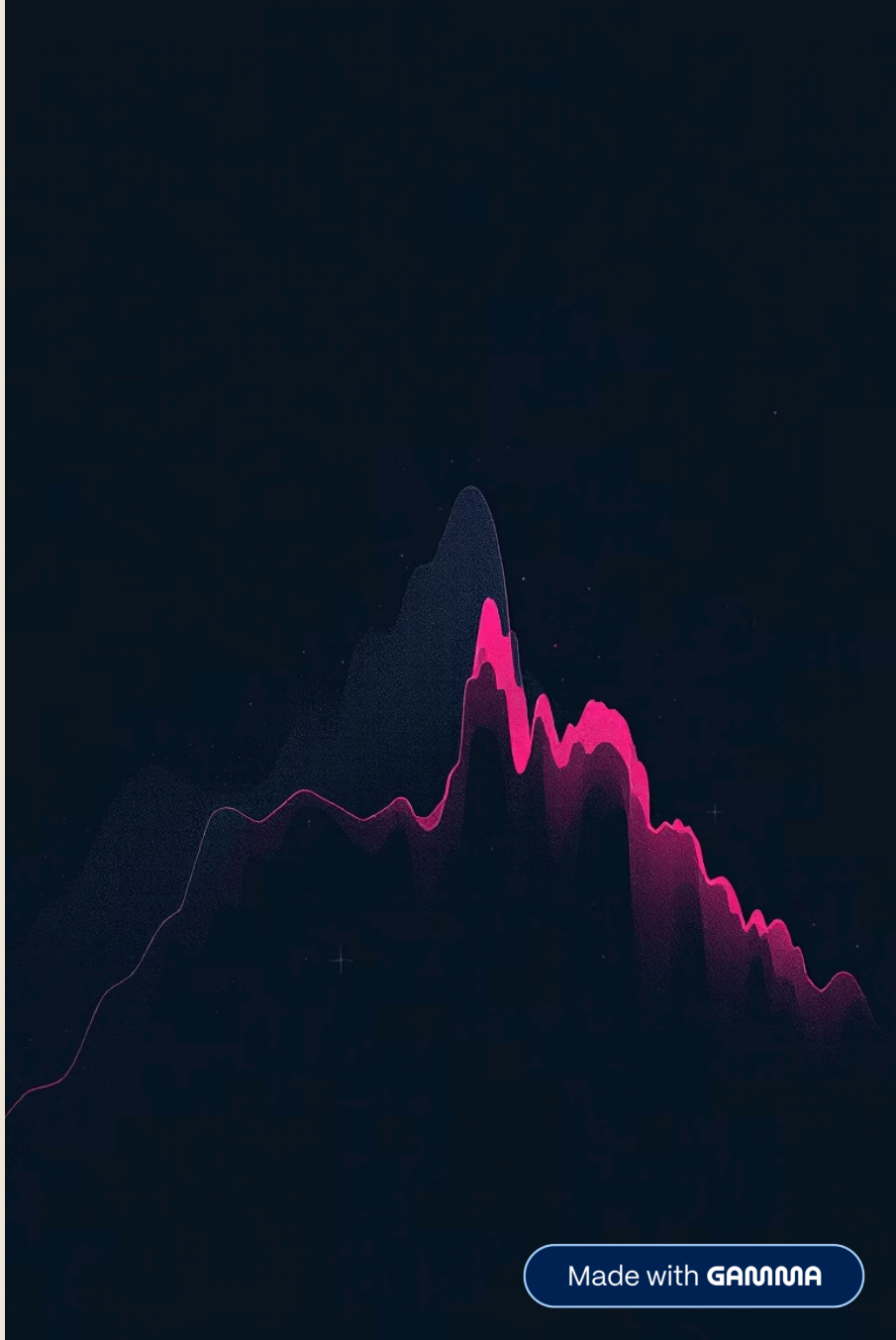
Example: 500 randomly selected students

Observations: The Fundamental Units of Data

In the context of data analysis, an observation is the most elementary unit on which we gather information. Each observation represents a single "subject" or "event" that we are studying.

- **Unit of Study:** These can be people, objects, animals, companies, events or even specific moments.
- **Specificity:** Each observation is unique and distinct from others, even if it shares many characteristics.
- **Basis for Analysis:** The set of all observations constitutes our dataset, from which we will extract insights.

Think of an observation as a single row in a spreadsheet, where each row describes a specific entity with its attributes and values.



The Population: The Complete Set

The population is the complete universe of all observations we wish to study. It is the total group of individuals, objects, or events that our investigation focuses on and to which we want to extend our conclusions.

- **Ideal Scope:** It is often a theoretical concept, so vast that it is practically impossible to observe or measure in its entirety.
- **Research Objective:** It represents the broadest scope of our analysis, defining the context in which our findings will be meaningful.
- **Examples:** All subscribers to a social network, all financial transactions in a country over a year, all the trees in a forest.

The Sample: A Representative Subset

A sample is a **finite and manageable subset** of the population. It is the actual object of our analysis, as the entire population is often too vast or inaccessible.

- **Practicality:** Allows research and analysis to be conducted efficiently and sustainably.
- **Inference:** Conclusions drawn from the sample are extended to the entire population.
- **Representativeness:** It is crucial that the sample is representative of the population to avoid biases and ensure the validity of inferences.

Accurate sample selection is fundamental to the quality of all data analysis.

Practical Example: Forest Health

Let's imagine we want to study the health of trees in the vast Black Forest. Here's how we would apply the concepts of observation, population, and sample in a practical context:

The Observation

A **single Norway spruce**, whose health conditions (height, trunk diameter, leaf colour) are measured.

The Population

All the trees present in the Black Forest, representing the entire universe of study for our investigation.

The Sample

A **group of 500 trees** randomly selected from different areas of the Black Forest, on which we will carry out our measurements.

Variables: Capturing Characteristics

Formal Definition

A variable is a function that **maps each observation** to a specific value:

$$X : \Omega \to S$$
$$\omega \mapsto x$$

Where Ω is the population and S is the set of possible values.

> Given the population of all people currently living in the world Ω, we define a variable H to collect the heights of the observed people ω:
>
> $$H : \Omega \to \mathbb{R}$$
> $$\omega \mapsto h$$
>
> Given an observation $\omega^{(1)}$, we may obtain $H(\omega) = 180$. We often say that H assumes the value 180 and write: $H = 180$.

Note: we use a capital letter X to denote the variable, while a lowercase letter x to represent the value assumed by the variable. For instance, we could write $X = x \in S$.

Types of Variables

Variables are classified based on the nature of the values they can take. Understanding these classifications is fundamental as it dictates the types of statistical analyses that can be applied.

Qualitative Variables

Describe qualities or characteristics that cannot be numerically measured or ordered in a meaningful way. They represent categories.

- **Example:** Colour (e.g., red, blue, green)
- **Example:** Sex (e.g., male, female)
- **Example:** Nationality (e.g., Italian, French, German)

Quantitative Variables

Describe quantities, meaning they can be numerically measured and ordered. These variables have a numerical value.

- **Example:** Height (e.g., 175 cm)
- **Example:** Weight (e.g., 70 kg)
- **Example:** Age (e.g., 30 years)

Discrete Variables

These are quantitative variables that can only take a finite number of values or a countably infinite number of values. They are typically whole numbers that result from counting.

- **Example:** Number of children in a family (e.g., 0, 1, 2, 3)
- **Example:** Number of cars in a parking lot

Continuous Variables

These are quantitative variables that can take any value within a given range. They often result from measurements and can have decimal values.

- **Example:** Height (e.g., 175.5 cm, 175.52 cm)
- **Example:** Temperature (e.g., 25.3°C)
- **Example:** Time taken to complete a task

Scalar Variables

Represent a single value at each observation. They are characterized by magnitude only and are the most common type of variable encountered.

- **Example:** Height (e.g., 180 cm)
- **Example:** Age (e.g., 35 years)
- **Example:** Temperature (e.g., 22°C)

Multidimensional Variables

Composed of multiple components or values for each observation, often representing different aspects of a single entity. They are characterized by both magnitude and direction, or by multiple distinct measurements.

- **Example:** Coordinates (e.g., latitude and longitude)
- **Example:** RGB Color (e.g., (255, 0, 0) for red)
- **Example:** Financial Portfolio (e.g., [Stock A, Stock B, Bond C])

Examples of Variables: Discrete, Continuous, and Multidimensional

To better illustrate the concepts of population, observation, and variable, let us consider some practical examples covering different types of variables.

Discrete Scalar Variables

We want to assess whether a coin is fair. The population is the set of all possible tosses. An observation is a specific toss. A discrete scalar variable X records the outcome, for example $S = \{heads, tails\}$. If we toss a coin, we might obtain $X = tails$.

Continuous Scalar Variables

We want to study the heights in centimetres of students in this class. Our population is the set of all students. We can use a continuous scalar variable X to record the heights, where $S = \mathbb{R}_+$. If we select a student, we might obtain $X = 175$.

Multidimensional Continuous Variables

We want to study the positions of all cars in the world. Our population is the set of all cars. We could use the variable X to indicate a car's latitude and longitude coordinates. The set of possible values could be $S = \mathbb{R}^2$. Once a car is selected, we might have $X = (37, 15)$.

Measurement Scales: Understanding Data Types

Nominal

Non-orderable values. Example: **Marital Status**. Categories such as "Married" or "Single" are distinct but cannot be meaningfully ranked as "greater" or "lesser".

Ordinal

Orderable values, non-significant differences. Example: **Level of Satisfaction**. There is a clear order (e.g., "slightly satisfied" to "very satisfied"), but the interval between one level and another is not necessarily equal.

Interval

Significant differences, but without an absolute zero. Example: **Temperature** (in Celsius or Fahrenheit). The difference between 20°C and 10°C is significant, but 0°C does not signify an absence of temperature, and ratios cannot be formed (e.g., 20°C is not "twice as warm" as 10°C).

Ratio

True zero and significant ratios. Example: **Height**. A zero indicates the absence of height (0 cm), and ratios are significant (e.g., 200 cm is twice 100 cm).

Dataset and Design Matrix

A dataset is organised as a **table** where **columns represent variables and rows represent observations**.

ID	Mathematics	Geography	English	Physics	Chemistry	Notes
x001	8	9	30	8	10	✓
x038	9	7	27	6	-	Missing
x002	6	-1	18	5	6	Error?
x012	7	7	25	4	10	✓

Missing values are common in real data and require specific management strategies.

Data Collection: Key Methods

Surveys

Structured tools for collecting information from individuals or groups using predefined questions.

Format: online forms, paper questionnaires, telephone interviews

Experiments

Data collection in a **controlled** environment to test causal relationships between variables.

Main type: randomised controlled trials

Observational Data

Recording data as it naturally occurs, without researcher intervention.

Advantages: captures real-world behaviours

Online Sources

Platforms such as Kaggle, UCI Repository, social media APIs and web scraping.

Benefits: rapid access, diverse datasets, ideal for prototyping

Data Collection Method: Surveys

Surveys gather data by asking questions to a sample of individuals, collecting information on opinions, behaviors, and attitudes across various fields.

Types of Survey Formats

Online Forms

Digital questionnaires distributed via email or web, offering broad reach and automated data collection.

Paper Questionnaires

Traditional print-based surveys distributed in person or via mail, suitable for populations with limited digital access.

Telephone Interviews

Live conversations over the phone, allowing for direct interaction and clarification of questions.

Key Advantages of Surveys

Cost-Effective

Collects large amounts of data at low cost, especially with online tools.

Broad Reach

Reaches diverse populations and large sample sizes across wide geographical areas.

Standardized

Consistent questions ensure data uniformity for easy comparison and analysis.

Versatile

Adapts to collect various information types, from facts to opinions and attitudes.

Real-World Examples

Customer Satisfaction

Businesses use surveys to gauge happiness and identify improvement areas.

Market Research

Companies conduct surveys to understand consumer preferences and market trends.

Data Collection Method: Experiments

Experiments are powerful research methods designed to establish cause-and-effect relationships by manipulating variables and observing their impact.

Controlled Experiments

Researchers manipulate an independent variable, controlling others, to observe its effect on a dependent variable, allowing for strong causal inferences.

Randomised Controlled Trials (RCTs)

Participants are randomly assigned to experimental or control groups, minimizing bias and serving as the "gold standard" for evaluating interventions.

Key Advantages

Experiments offer high control over variables, enabling precise measurement and the determination of clear cause-and-effect relationships.

Real-World Example

Medicine

Testing the efficacy of new drugs or vaccines through RCTs to compare patient outcomes against placebos or existing treatments.

Data Collection Method: Observational Data

Observational data involves collecting information by watching behaviors, situations, or phenomena in their natural settings, without direct intervention. It's valuable for understanding complex interactions and natural processes.

Observational Studies

Researchers observe and record phenomena in their natural state without intervention, gathering rich descriptive data.

Naturalistic Observation

Observing subjects in their natural environment without manipulation, ideal for authentic behavior and high ecological validity.

Structured Observation

Researchers create a specific situation to observe participant reactions, introducing some control to focus on variables.

Real-World Examples

Anthropology

Ethnographic studies observing cultural practices and social interactions within a community over an extended period.

Ecology

Tracking animal migration patterns or predator-prey dynamics in their natural habitats without interference.

Psychology

Observing classroom interactions between teachers and students to understand learning behaviors and social dynamics.

Business

Analyzing customer shopping habits in a retail store to optimize product placement and marketing strategies.

Loading Data in Python: Titanic Dataset

We will use the famous **Titanic dataset** to learn data analysis techniques. It contains information about passengers and their survival.

Key Characteristics

- Categorical and numerical variables
- Presence of **missing values**
- 891 observations, 12 variables
- Understandable historical context

- **PassengerId**

  Unique identifier

- **Survived**

  Survival (0=No, 1=Yes)

- **Pclass**

  Ticket class (1st, 2nd, 3rd)

- **Age, Sex, Fare**

  Demographic information

Data Wrangling: Cleaning and Preparation

Real-world data is never perfectly clean. **Data wrangling** transforms raw data into a structured and usable format.

Handling Missing Values

Removing rows, filling with means, or marking as "unknown"

Type Conversion

Converting columns to the correct type (e.g., Pclass from integer to categorical)

Feature Engineering

Creating derived variables (e.g., FamilySize = SibSp + Parch)

Filtering and Selection

Focusing on specific subsets (e.g., only adult passengers)

Wide vs. Long Format: A Practical Example

Data can be structured in two main formats, each with its advantages depending on the analysis to be performed. Understanding the difference is fundamental for data preparation.

Wide Format

In "Wide" format, each variable has its own column and each row represents a unique observation. This is often the default format for relational databases and many traditional statistical analyses.

Student ID	Mathematics Score	Science Score
S001	90	85
S002	75	80
S003	88	92

Here, each student (observation) is a unique row, and scores for different subjects are separate variables in columns.

Long Format

In "Long" format, the variables that were columns in wide format are "stacked" into one or more new columns. Each row now represents a single data point for a specific variable. This format is often preferred for visualisations or time series analysis.

Student ID	Subject	Score
S001	Mathematics	90
S001	Science	85
S002	Mathematics	75
S002	Science	80
S003	Mathematics	88
S003	Science	92

In this case, each row describes a student's score for a single subject, making the dataset longer but potentially more flexible for certain operations.

The Data Analysis Workflow

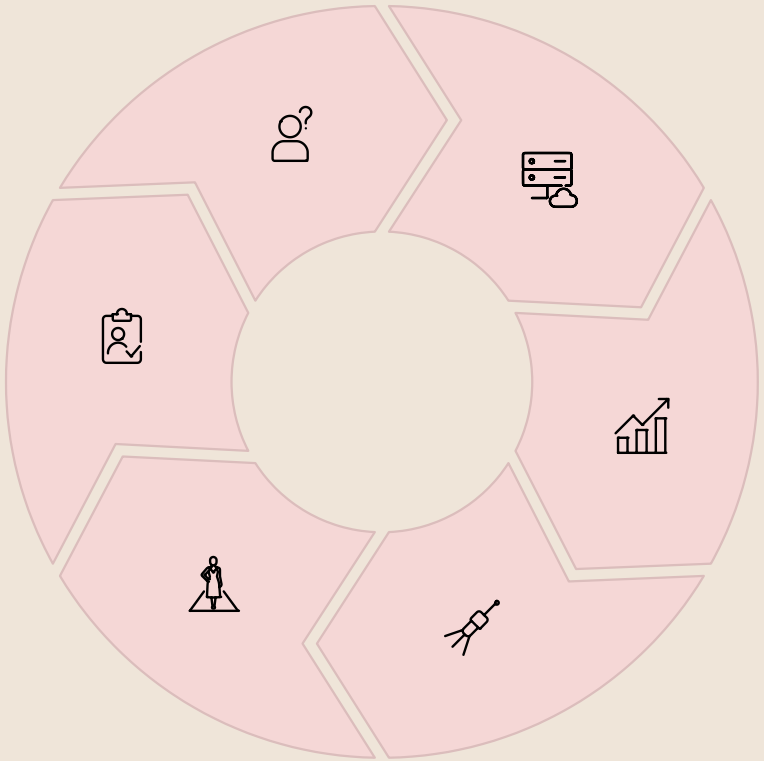
Data analysis is the **process** of inspecting, cleaning, transforming, and modelling data to discover useful information and support decision-making.

Define

The specific research question

Collect

The necessary data to answer



Evaluate

And iteratively update the analysis

Clean

And format the collected data

Model

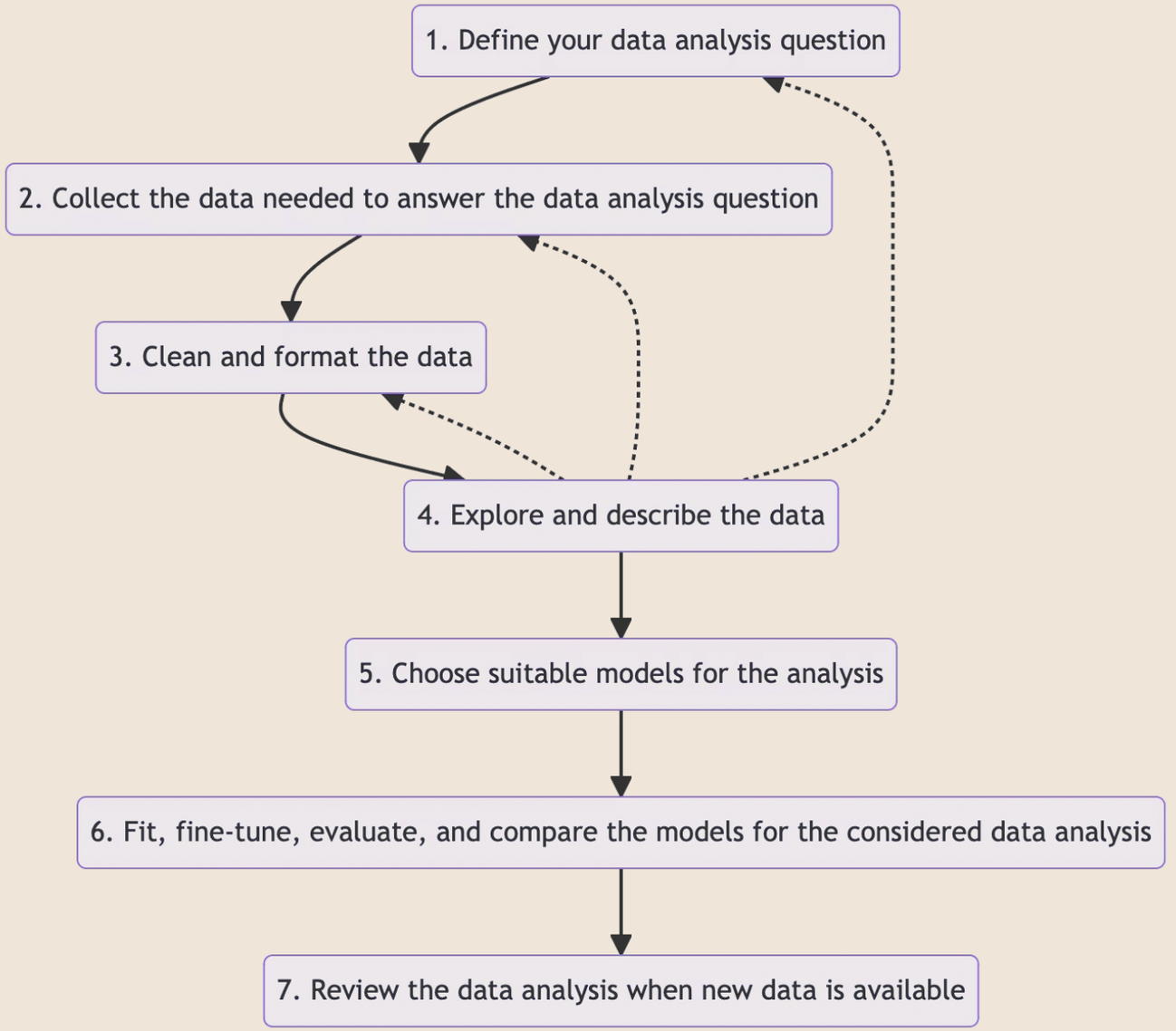
Choose and apply appropriate techniques

Explore

And describe patterns in the data

Remember: data analysis is a non-linear process. It is often necessary to revisit previous steps to refine the approach.

A Non-Linear Process



Practical Example: Customer Review Analysis

Let's follow a practical example where a data analyst aims to improve product quality and customer satisfaction by analysing reviews. This illustrates the iterative workflow of data analysis.

Define the Question

Clarify the specific objective of the analysis.

Collect and Cleanse Data

Acquire reviews and prepare them for analysis (duplicate removal, corrections).

Explore and Model

Utilise techniques such as word clouds and sentiment analysis to identify initial themes.

Evaluate and Iterate

Analyse the results to refine insights and reapply the model for clarity.

Continuous Monitoring

Implement a system for constantly analysing new reviews.

This iterative approach allows the analyst to refine insights and adapt strategies over time. This workflow will guide our journey, applying each phase to real datasets, starting with the Titanic dataset.

Conclusions and Next Steps

We Have Explored:

Fundamental Concepts

Observations, populations, samples, and variables

Practical Management

Data loading, cleaning, and transformation

Methodology

The iterative workflow of data analysis

In upcoming lectures, we will delve deeper into visual data exploration, descriptive statistics, and initial modelling techniques.

References

- Heumann, Christian, and Michael Schomaker Shalabh. Introduction to statistics and data analysis. Springer International Publishing Switzerland, 2016.