



Computer Science Preparation – M5 – T1 Final

Data processing

Data are units of information such as numbers, text, images, sounds, or observations gathered from events or research. We process and analyze data so we can get useful results from it. A clear goal helps us decide what data to collect and how to analyze it.

A useful way to think about is:

Step	Main Idea
Collect	Get data for a clear purpose.
Process / Analyze	Work with the data to get useful information.
Clean	Find and fix data problems safely.
Transform	Change data into a more useful format for a task.
Explore	Use graphs to understand patterns, comparisons, and unusual values.

Data types:

A **data source** is the place or origin from which we obtain data. Sources can include people, animals, things, places, or events.

We have two data types:

- **Quantitative** – data in form of numbers
- **Qualitative** – data in form of words

Data is classified according to how we have gotten that information:

- **Primary Data**
- **Secondary Data**

Primary Data	Secondary Data
Obtained from the original source.	Does not come directly from the original source; it uses existing data or information based on primary data.
You collect it directly by taking actions, asking people, observing, or running an experiment.	You refer to information that already exists, such as statistical forms, reviews, articles, or documents.
Often current and realistic for your specific question.	Can be cheaper and less time-consuming, but you must check whether the source is trustworthy and the information is correct.

Primary VS Secondary data

Using **secondary data** can be **cheaper** and **less time-consuming** than using primary data, but **primary data** is **current data**, **realistic**. Secondary data is **obtained by referencing** the **primary source** or passing it on to each other. Secondary data can be changed, and not correct, that is why we need to use trustworthy sources.

Data Collection

Information and sources of information will determine the method of collecting information. Each piece of information may **require different collection methods**. Some of those methods are:

- **Interview** – ask people and listen to their answers
- **Survey** - collect answers from many people at the same time
- **Observing** – Watch carefully and record what happens
- **Experiment** – test something by controlling factors and record results
- **Document review** – use existing records
- **Census** - systematically collect and record information about the members of population

Correctness of Information

Selecting incorrect or wrong information can cause a wrong conclusion and a wrong result. It can be dangerous and create problems for us and others. We should follow these 5 steps for checking if the information is correct or not:

- 1. Information Age** - You should check when the information was published. Did the information change or improve in the meantime? You should check if the information is still correct and usable.
- 2. Relevance** - We should check if the information is connected to the problem that we are working on.
- 3. Reliability of the source** - We need to consider if the source is reliable or not. Who published it and can we trust them.
- 4. Accuracy** - We need to check if the collected data is correct or not. Are the people that published that information professionals with a lot of experience?
- 5. Purpose** - We have to verify for what purpose was that data collected for. Can it be used for public campaigns, advertisements, educational purposes...

TERMS



Attribution

Others can copy, distribute, display, perform and remix your work if they credit your name as requested by you



No Derivative Works

Others can only copy, distribute, display or perform verbatim copies of your work



Share Alike

Others can distribute your work only under a license identical to the one you have chosen for your work



Non-Commercial

Others can copy, distribute, display, perform or remix your work but for non-commercial purposes only.

Data Cleaning

The **collected information** can be **incorrect** or **not suitable** for processing. Examples: Incorrect information, no information, information that is out of the scope, units that don't match....

These **errors** are **caused** by several reasons such as: Incomplete information provided by the person that was collecting the information, Typing mistakes, Lack of matching information, Communication problems

We **should correct** the **information** when **errors** are **found if we can correct it**. If we **cannot correct** it and it **doesn't affect processing**, we can **delete** it.

Data Transformation

The **goal** of **data transformation** is to **prepare** the **data** in a **format** that is **ready** for **processing**. The format of data that is ready to be processed in a **spreadsheet program** is a **table** in which each **row** is an **example of data** and each **column** is the **attribute** for that data. We have three ways of transforming data:

- 1. Reducing information** - Reducing the amount of information is selecting only the information that we are interested in. This way it will be easier to process the data as we will have only important information
- 2. The Increasing amount of information** - Sometimes we do not have enough information so we have to find more information and combine it with the existing. This way we are increasing the amount of information. We have to be careful to only import correct and relevant information
- 3. Data Integration** - Data integration is the grouping of data whose values are consistent with the conditions defined by the attribute of interest. For example, we want to organize provinces depending on which low, middle, and high media income per household.