# **Working with Data**

## **What is Data?**

Data refers to any information, facts, or figures that can be collected, stored and analyzed. It can be in various forms such as numbers, text, images, or multimedia. In the context of computing and technology, data often to digital information that computers can process and manipulate.

## **Aspects of data**

## **Formats:**

**Structured data:** Organized into a predefined format, often stored in databases or spreadsheets. Imagine data that's neatly organized, like in a table or spreadsheet. It's easy for computers to understand because it follows a clear pattern.

**Unstructured data:** Doesn’t have a predefined format. This is data that doesn't follow a specific format. It could be text documents, photos, or videos things that don't fit neatly into rows and columns.

**Semi-structured data:** Falls between structured and unstructured data. Somewhere in between, like files with a little bit of structure but not fully organized. For example, think of data stored in JSON or XML files.

## **Scope:**

**Big Data**: Large volumes of data that require specialized tools and techniques for processing and analyses. When we have a LOT of data—so much that regular methods struggle to handle it. Big data needs special tools and techniques to manage and analyze effectively.

**Small Data:** Relatively smaller datasets that can be managed and analyzed with simpler methods.

## **Biases:**

**Selection Bias:** When certain groups or types of data are overrepresented or underrepresented. This happens when some types of data are favored or ignored, skewing our understanding of the whole picture.

**Sampling Bias:** When the sample used to collect data doesn’t accurately represent the entire population. If the data we collect doesn't represent the whole group accurately, we might make incorrect assumptions.

**Cultural Bias:** When data reflects cultural norms or prejudices inherent in the society where it was collected. Sometimes, the data reflects the biases or norms of the culture it comes from, which can affect how we interpret it.

**Quality:** Data quality refers to the accuracy, completeness, consistency, and reliability of the data. Poor data quality can cause incorrect analysis and decision-making. Just like how we want our information to be accurate and reliable, we want our data to be the same. If the data isn't good quality, it can lead to mistakes in our analysis and decisions.

**Data Quality:** Just like how we want our information to be accurate and reliable, we want our data to be the same. If the data isn't good quality, it can lead to mistakes in our analysis and decisions

Parsing Data with Regular Expressions Regular expressions (regex) are patterns used to match character combinations in strings. They are widely used for parsing and manipulating textual data. Here’s how they work:

* **Pattern matching:** Regex allows you to define patterns to search for within text data.
* **Tokenization:** You can use regex to split text into tokens based on certain criteria.
* **Data extraction:** Regex can help extract specific information from structured or semi-structured text data.
* **Data cleaning:** Regex can be used to find and replace or remove unwanted characters or patterns from text data.

## **Obtaining web data**

1. **Beautiful Soup:** Python library for parsing HTML and XML documents. Helps extract data from web pages by navigating HTML structure It's like a tool that helps us dissect web pages, pulling out the information we want by understanding the structure of the page.
2. **Pandas:** Python library for data manipulation and analysis. Read HTML tables directly from web pages into dataframes. This is another tool that's great at organizing data, including data we get from web pages. It's like a magic wand for turning messy web data into neat tables we can work with.

## **Benefits**

Saves time by automating web data extraction Seamlessly integrates web data with data analysis workflows provides flexibility in handling and analyzing different types of web data. Using these tools saves us time because they can do the boring stuff of collecting and organizing data from the web for us. They also make it easy to mix this web data with the other data we have, so we can analyze it all together. Plus, they're flexible, meaning they can handle all sorts of different web data without breaking a sweat.