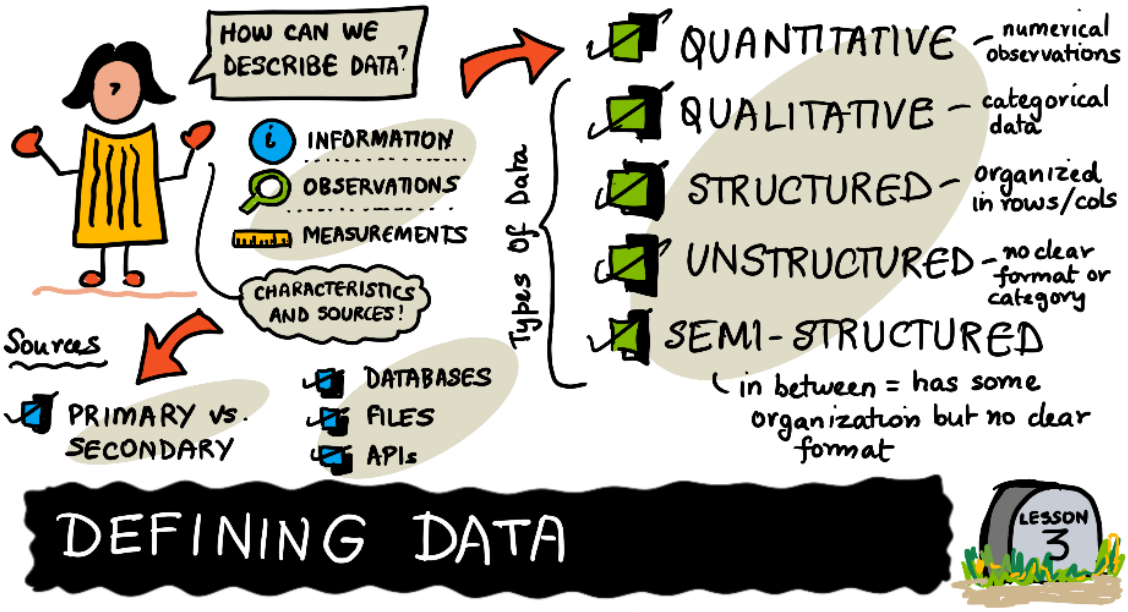
Defining Data



Defining Data - *Sketchnote by @nitya*

Data is facts, information, observations and measurements that are used to make discoveries and to support informed decisions. A data point is a single unit of data with in a dataset, which is collection of data points. Datasets may come in different formats and structures, and will usually be based on its source, or where the data came from. For example, a company's monthly earnings might be in a spreadsheet but hourly heart rate data from a smartwatch may be in JSON format. It's common for data scientists to work with different types of data within a dataset.

This lesson focuses on identifying and classifying data by its characteristics and its sources.

Pre-Lecture Quiz

How Data is Described

Raw Data

Raw data is data that has come from its source in its initial state and has not been analyzed or organized. In order to make sense of what is happening with a dataset, it needs to be organized into a format that can be understood by humans as well as the technology they may use to analyze it further. The structure of a dataset describes how it's organized and can be classified at structured, unstructured and semi-structured. These types of structure will vary, depending on the source but will ultimately fit in these three categories.

Quantitative Data

Quantitative data is numerical observations within a dataset and can typically be analyzed, measured and used mathematically. Some examples of quantitative data are: a country's population, a person's height or a company's quarterly earnings. With some additional analysis, quantitative data could be used to discover

seasonal trends of the Air Quality Index (AQI) or estimate the probability of rush hour traffic on a typical work day.

Qualitative Data

Qualitative data, also known as categorical data is data that cannot be measured objectively like observations of quantitative data. It's generally various formats of subjective data that captures the quality of something, such as a product or process. Sometimes, qualitative data is numerical and wouldn't be typically used mathematically, like phone numbers or timestamps. Some examples of qualitative data are: video comments, the make and model of a car or your closest friends' favorite color. Qualitative data could be used to understand which products consumers like best or identifying popular keywords in job application resumes.

Structured Data

Structured data is data that is organized into rows and columns, where each row will have the same set of columns. Columns represent a value of a particular type and will be identified with a name describing what the value represents, while rows contain the actual values. Columns will often have a specific set of rules or restrictions on the values, to ensure that the values accurately represent the column. For example imagine a spreadsheet of customers where each row must have a phone number and the phone numbers never contain alphabetical characters. There may be rules applied on the phone number column to make sure it's never empty and only contains numbers.

A benefit of structured data is that it can be organized in such a way that it can be related to other structured data. However, because the data is designed to be organized in a specific way, making changes to its overall structure can take a lot of effort to do. For example, adding an email column to the customer spreadsheet that cannot be empty means you'll need figure out how you'll add these values to the existing rows of customers in the dataset.

Examples of structured data: spreadsheets, relational databases, phone numbers, bank statements

Unstructured Data

Unstructured data typically cannot be categorized into rows or columns and doesn't contain a format or set of rules to follow. Because unstructured data has less restrictions on its structure it's easier to add new information in comparison to a structured dataset. If a sensor capturing data on barometric pressure every 2 minutes has received an update that now allows it to measure and record temperature, it doesn't require altering the existing data if it's unstructured. However, this may make analyzing or investigating this type of data take longer. For example, a scientist who wants to find the average temperature of the previous month from the sensors data, but discovers that the sensor recorded an "e" in some of its recorded data to note that it was broken instead of a typical number, which means the data is incomplete.

Examples of unstructured data: text files, text messages, video files

Semi-structured

Semi-structured data has features that make it a combination of structured and unstructured data. It doesn't typically conform to a format of rows and columns but is organized in a way that is considered structured and may follow a fixed format or set of rules. The structure will vary between sources, such as a well defined hierarchy to something more flexible that allows for easy integration of new information. Metadata are

indicators that help decide how the data is organized and stored and will have various names, based on the type of data. Some common names for metadata are tags, elements, entities and attributes. For example, a typical email message will have a subject, body and a set of recipients and can be organized by whom or when it was sent.

Examples of semi-structured data: HTML, CSV files, JavaScript Object Notation (JSON)

Sources of Data

A data source is the initial location of where the data was generated, or where it "lives" and will vary based on how and when it was collected. Data generated by its user(s) are known as primary data while secondary data comes from a source that has collected data for general use. For example, a group of scientists collecting observations in a rainforest would be considered primary and if they decide to share it with other scientists it would be considered secondary to those that use it.

Databases are a common source and rely on a database management system to host and maintain the data where users use commands called queries to explore the data. Files as data sources can be audio, image, and video files as well as spreadsheets like Excel. Internet sources are a common location for hosting data, where databases as well as files can be found. Application programming interfaces, also known as APIs allow programmers to create ways to share data with external users through the internet, while the process of web scraping extracts data from a web page. The lessons in Working with Data focus on how to use various data sources.

Conclusion

In this lesson we have learned:

- What data is
- How data is described
- How data is classified and categorized
- Where data can be found

🚀 Challenge

Kaggle is an excellent source of open datasets. Use the dataset search tool to find some interesting datasets and classify 3-5 datasets with this criteria:

- Is the data quantitative or qualitative?
- Is the data structured, unstructured, or semi-structured?

Post-Lecture Quiz

Review & Self Study

- This Microsoft Learn unit, titled Classify your Data has a detailed breakdown of structured, semi-structured, and unstructured data.

Assignment

Classifying Datasets