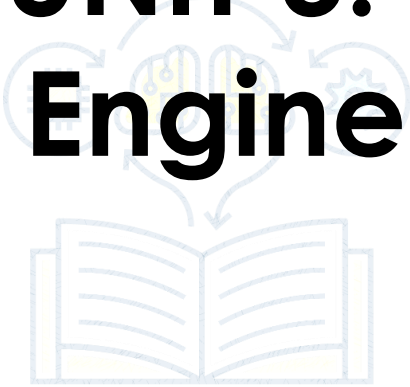


UNIT 3:

Data Engineering

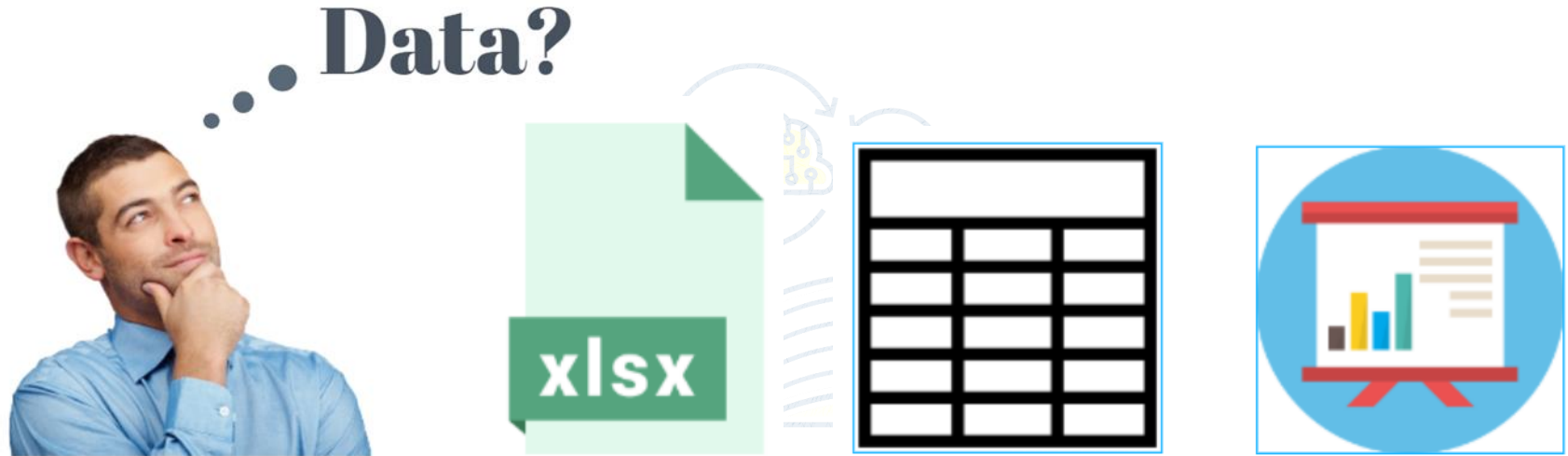


Prepared by Nima Dema

Overview of data and its sources

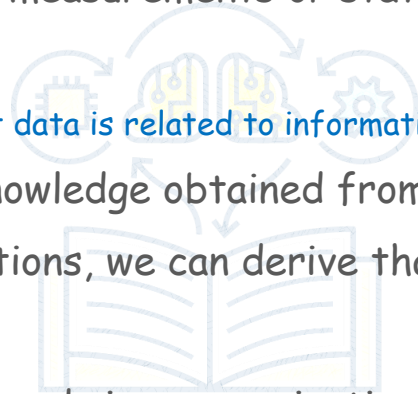


What is Data?



What is Data?

- **Data** are collected observations or measurements represented as text, numbers, multimedia.
- Factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation
 - It indicates everything about data is related to information.
- **Information** is defined as knowledge obtained from investigation, study, or instruction.
- By comparing the two definitions, we can derive that information is more useful than data.
- Information is created and used via communication after the data is studied and analysed.



What is Data?



fieldnotes



Data can be



videos



audio recordings



images



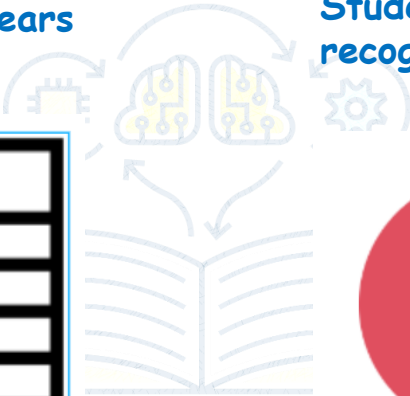
documents

What is Data?

- Data is different depending on your area of research/study

Student's performance over the years

Student Attendance system using face recognition



Categories of Data

- **Quantitative Data** - can be expressed as numerical values, counted or compared.



Survey Scale



temperature



weights



count of words



textual descriptions



images



maps



videos

- **Qualitative Data** - Qualitative data is the descriptive and conceptual findings collected through questionnaires, interviews, or observation.

Categories of Data

Quantitative/Numerical Data

- ✓ This data type tries to quantify things and it does by considering numerical values that make it countable in nature.
- **Continuous**
 - Continuous Data can take any value (within a range)
 - Age of person: could be any value (within a range of human age)
 - Height of a person
 - Temperature of the day
- **Discrete**
 - The numerical values which fall under are integers or whole numbers are placed under this category.
 - Discrete data is a count that involves integers — only a limited number of values is possible.
 - Number of student attending ITS307 class, number of family members.

Categories of Data

Qualitative/Categorical Data

- ✓ Qualitative or Categorical Data describes the object under consideration using a finite set of discrete classes. It means that this type of data can't be counted or measured easily using numbers and therefore divided into categories.
- **Nominal**
 - These are the set of values that don't possess a natural ordering.
 - Colour of gho can be considered as nominal data as it is not possible to state black is greater than red.
 - Gender of a person is also an nominal data.
- **Ordinal**
 - These types of values have a natural ordering while maintaining their class of values.
 - Size of clothing brand can be considered as ordinal data (small < medium < large)
 - Grading system ($A > B > C$)
 - Education Level (Degree > High School > Middle School)
 - Satisfactory Rating (extremely agree > agree > disagree)

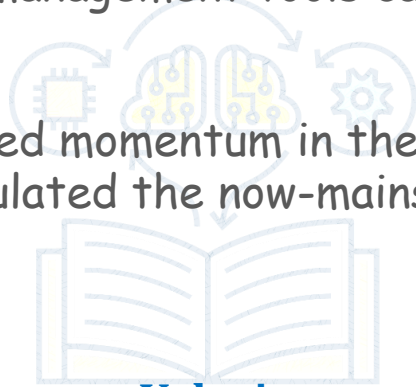
What is Big Data?

- **Big Data** is a collection of data that is huge in volume, yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently.
- Concept of big data gained momentum in the early 2000s when industry analyst Doug Laney articulated the now-mainstream definition of big data as the three V's:

Volume

Velocity

Variety

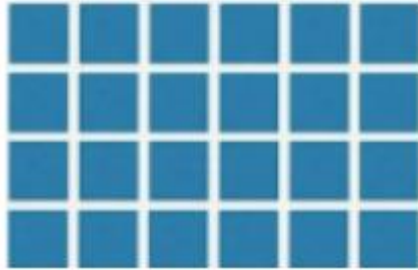


Types of Big Data

unstructured data



Structured data



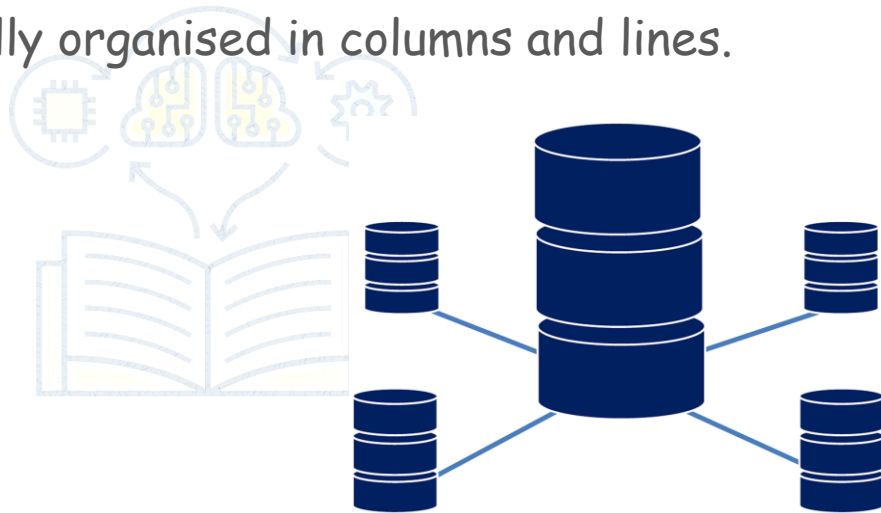
semi-structured data



Types of Big Data

1. Structured Data

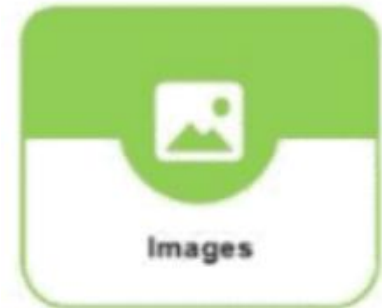
- ✓ Structured data is often referred to as the *easiest* one to manage because it is methodically organised in columns and lines.



Types of Big Data

2. Unstructured Data

- ✓ Unlike structured data, unstructured data does not have a predefined structure.
- ✓ What characterises unstructured data is its qualitative nature and the fact that it is not held or controlled within a mere excel sheet or a column/line database.



Types of Big Data

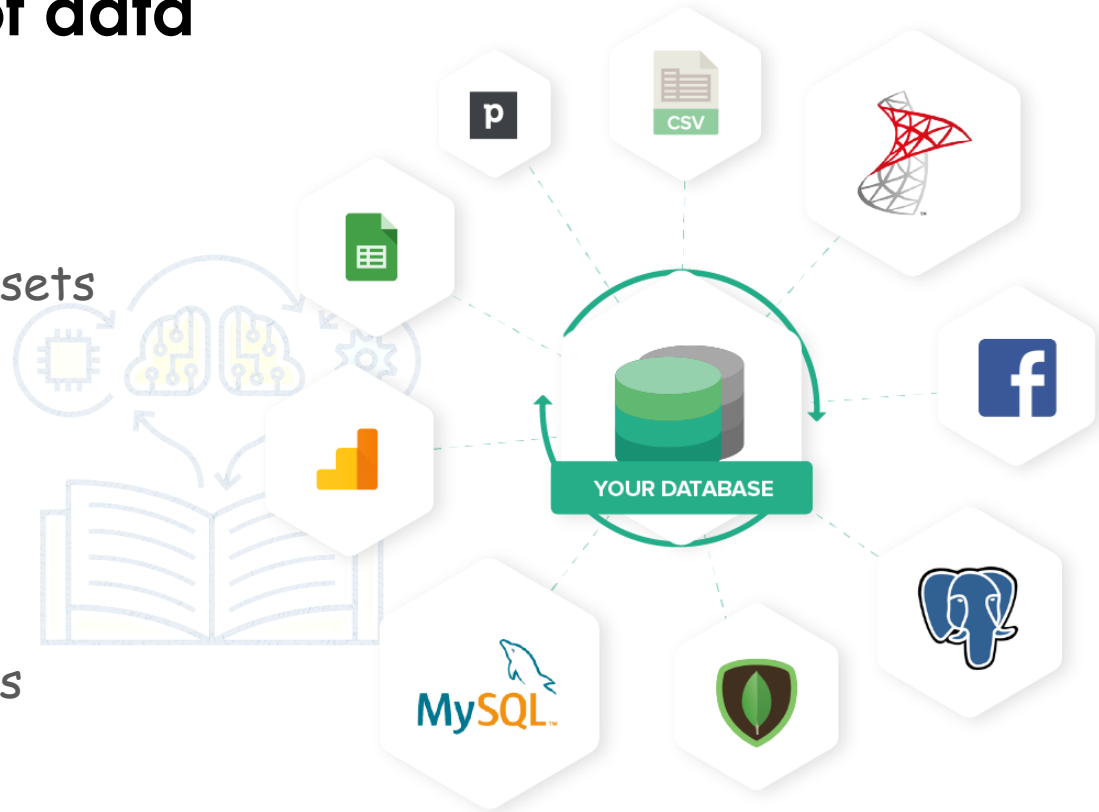
3. Semi-structured Data

- ✓ This is essentially structured and unstructured data combined.
- ✓ Semi-structured data is information that does not reside in a relational database or any other data table, but it contains some organizational properties.



Common sources of data

- ✓ Relational Databases
- ✓ Flat files and XML datasets
- ✓ APIs and Web services
- ✓ Web Scraping
- ✓ Data Streams and Feeds



Data Sources – Relational Databases

- ✓ Organizations have internal applications to support them in managing their day to day business activities, customer transactions, human resources activities and their workflows.
- ✓ These systems use relational databases such as;

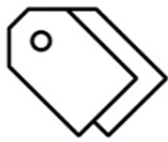


Data Sources – Relational Databases

- ✓ Data stored in databases and data warehouses can be used as a source for analysis.

Example; Data from retail transactions systems can be used to analyse sales in different regions.

Example; Data from customer relationship management system can be used for making sales projections.



Customer relationship
management system



Sales projections

Data Sources – Flat File and XML Datasets

- ✓ External to the organization, there are other publicly and privately available datasets.

Example; Government organization releasing demographic and economic datasets on an ongoing basis.

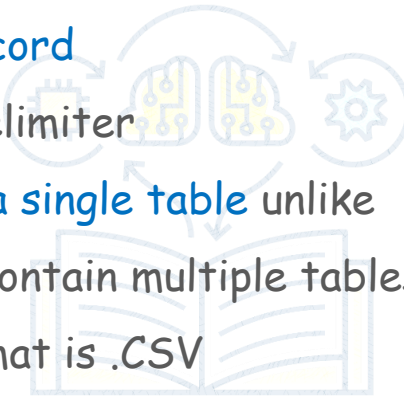
Example; Companies that sell point-of-sale data or Financial or weather data; Which business can use to define strategy, predict demand and make decisions related to distribution or marketing promotion.

- ✓ Such datasets are typically made available as flat files, spreadsheet files or XML documents.

Data Sources – Flat File and XML Datasets

Flat files:

- ✓ Store data in **plain text** format
- ✓ Each **line, or row, is one record**
- ✓ Each value separated by delimiter
- ✓ Data in a flat file **maps to a single table** unlike relational databases that contain multiple tables.
- ✓ Most common flat file format is **.CSV**



```
"OrderID", "CustomerID",  
"01", "001", "06/06/2021"  
"02", "369", "06/06/2021"  
"03", "151", "06/06/2021"  
"04", "014", "06/06/2021"  
"05", "061", "06/06/2021"  
"06", "220", "06/06/2021"
```

Data Sources – Flat File and XML Datasets

Spreadsheet files:

- ✓ Special type of flat files that also organize data in a **tabular - rows and columns**.
- ✓ Can contain multiple worksheets, and each worksheet can map to a different table.
- ✓ Although data in spreadsheets is in plain text, the files can be stored in custom formats, formatting, formulas etc.
- ✓ MS .XLS or .XLSX, google sheet, apple numbers etc



Data Sources – APIs and Web Services

- ✓ Many data providers and websites provide APIs and web services which multiple users or applications can interact with and obtain data for processing or analysis.
- ✓ APIs and Web Services typically listen for incoming requests, which can be in the form of web requests from users and return data in plain text, XML, HTML, JSON or media files.



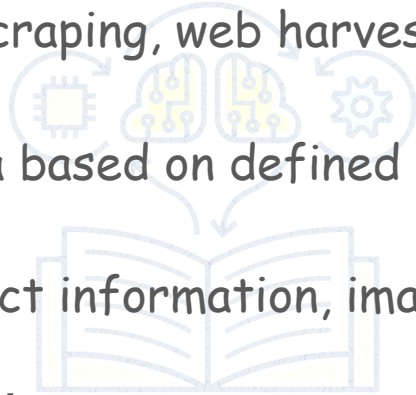
Twitter and Facebook APIs
for customer sentiment analysis



Stock Market APIs
for trading and analysis

Data Sources – Web Scraping

- ✓ Extract relevant data from unstructured sources.
- ✓ Also known as Screen scraping, web harvesting and web data extraction.
- ✓ Downloads specific data based on defined parameters
- ✓ Can extract text, contact information, images, videos, product items and more.



Web scraping tools:

- ☐ BeautifulSoup
- ☐ Scrapy

Data Sources

- ✓ Data sources can be internal or external to the organization.
 - **Primary:** refers to information obtained directly from the source;
 - Data from organization's CRM, HR or workflow applications.
 - Data you gather directly through surveys, interviews, discussions, observations, etc.
 - **Secondary:** refers to information retrieved from existing sources;
 - External databases
 - Research articles, publication, training materials, internet searches or financial records available as public data.
 - **Third-party:** refers to data purchased from aggregators who collected data from various sources and combine it for the purpose of selling it.



THANK YOU 😊