

Data Engineering Concepts

COMP63301

Dr. Sandra Sampaio

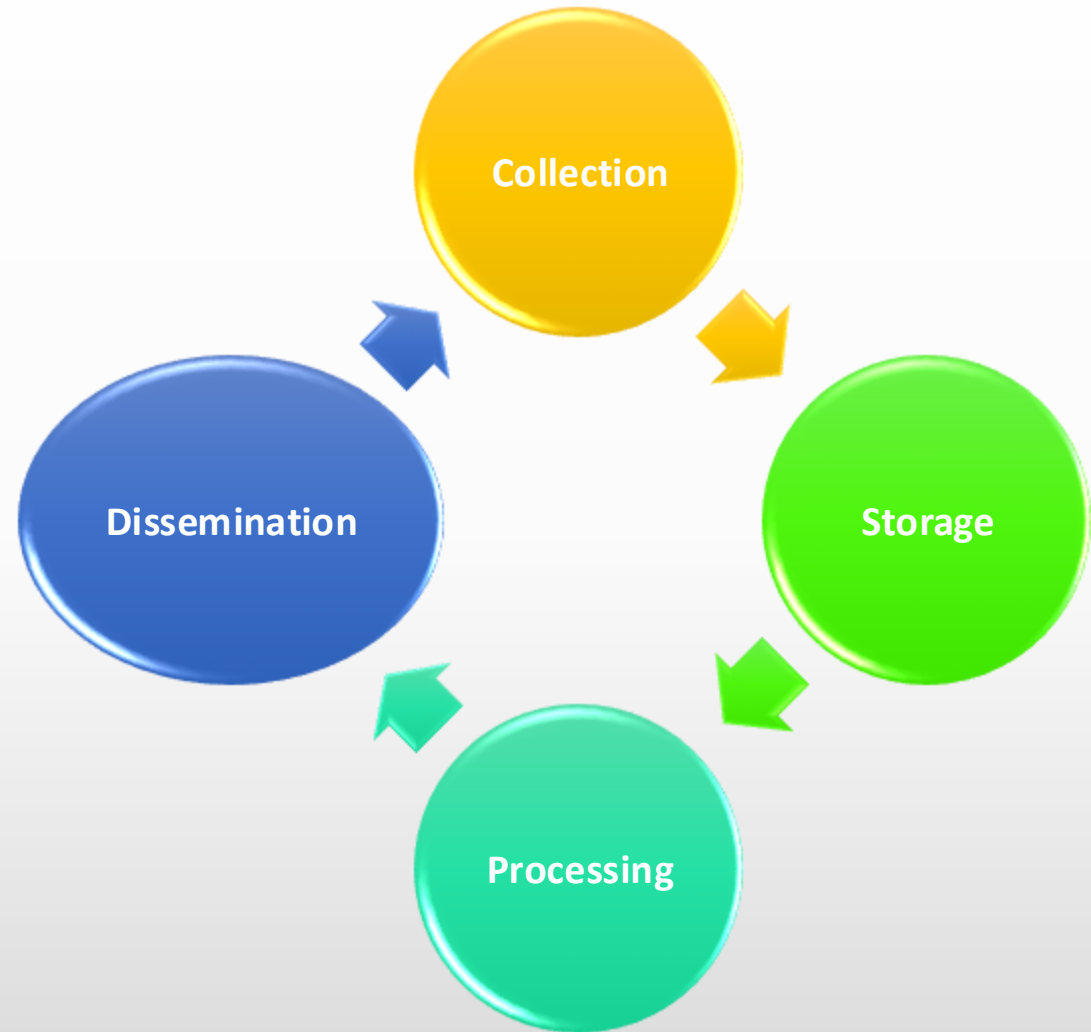


Data Lifecycle, Data Engineering Lifecycle and Understanding Data through Profiling



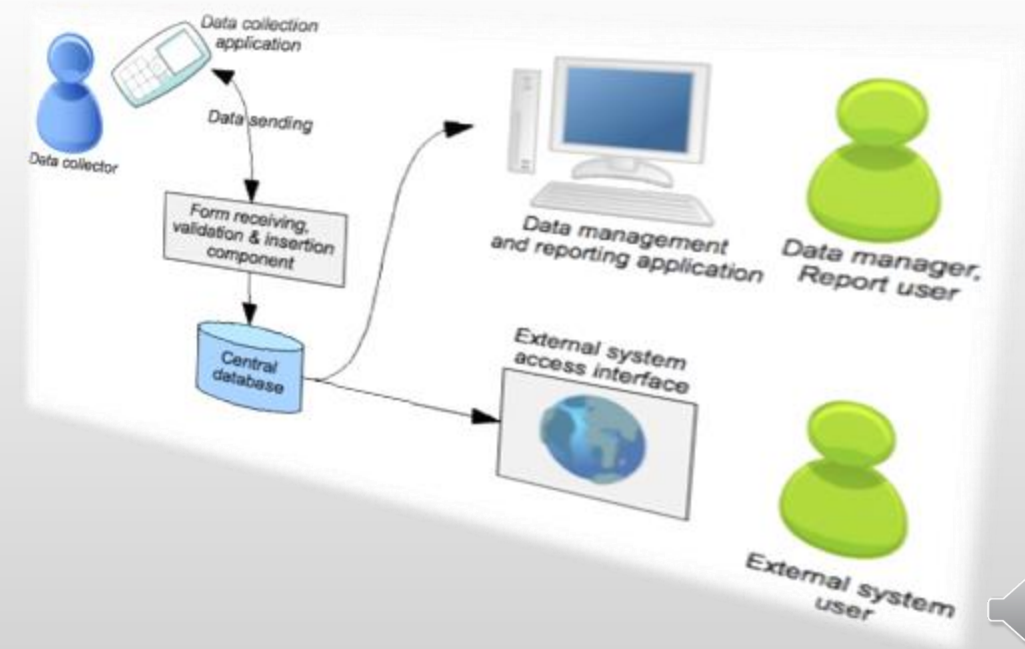
What Is the Data Lifecycle?

- Data Lifecycle (DL) refers to **the phases through which data moves in an organization**. The different phases include how the organization collects, stores, processes and disseminates their *key data*.
- Key data defines the most critical or important data elements that support the organization's specific business processes.



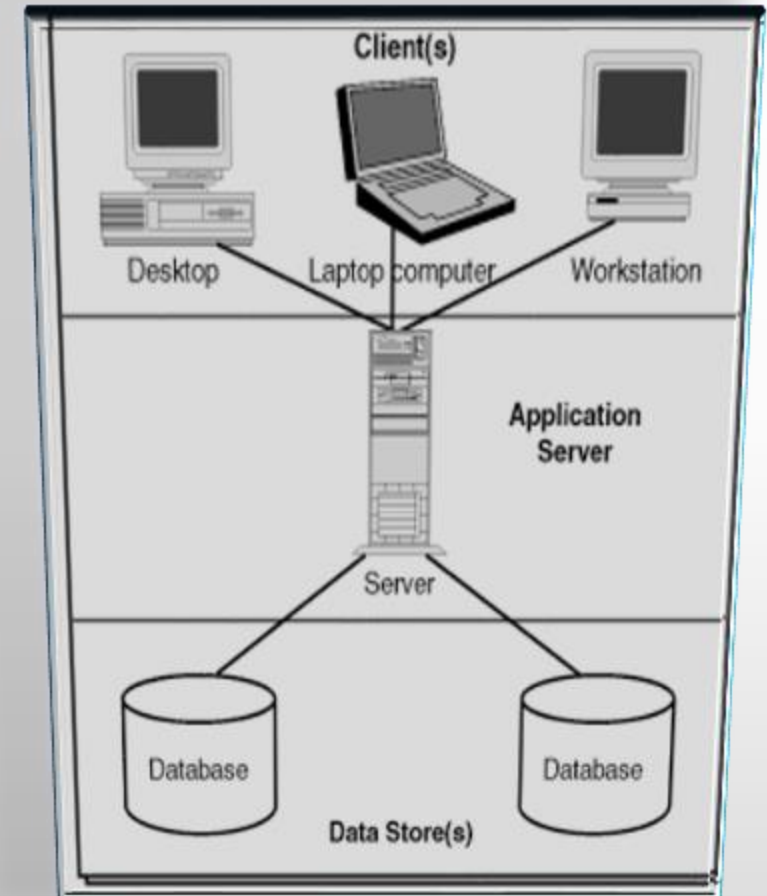
Collection

- Data collection methods are wide and varied.
- Any one method of collection is not inherently better than any other; i.e., **each has pros and cons** that must be weighed up in view of a rich and complex context.
- Examples of **primary** data collection methods:
 - Surveys
 - Interviews
 - Observations
 - Unobtrusive Methods
 - Experimentation



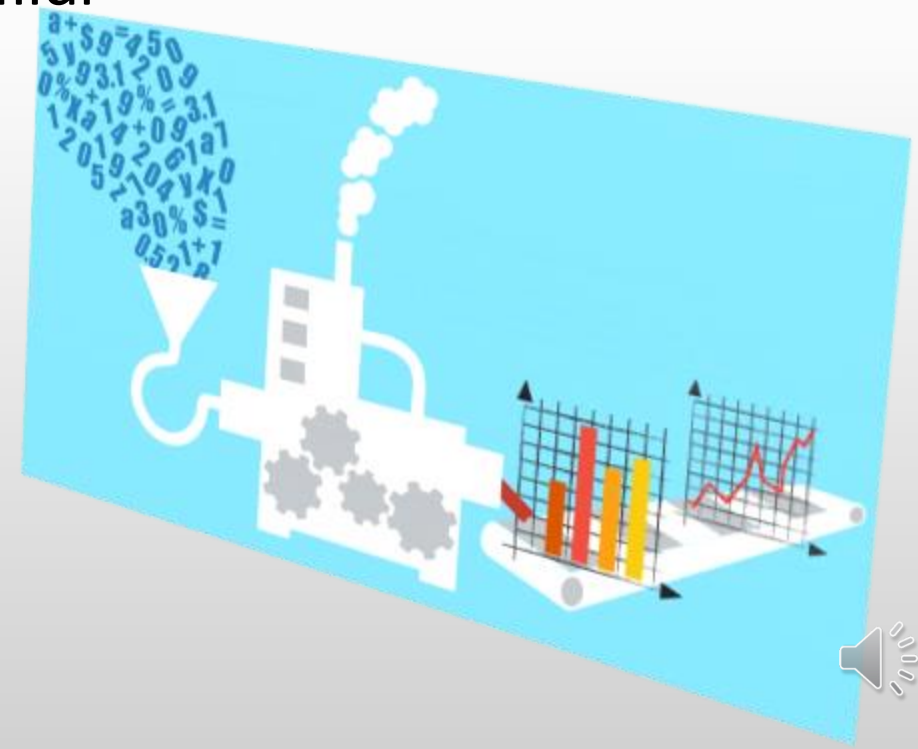
Storage

- It is about placing data in a container resource for data objects for future access.
- Prior to storage, **data modelling** may be required, if the data is to be accessed via a given application. For example, a single CSV file may need to be transformed into one or multiple relational tables, before it can be stored in a relational database.
- Regarding storage media, electronic data is stored not in boxes or files, but in a variety of media that challenges accountability, for example:
 - servers,
 - laptop computers,
 - handheld devices,
 - old back-up tapes,
 - CD-ROM disks,
 - thumb drives,
 - cellular phones and pagers,
 - etc.



Processing

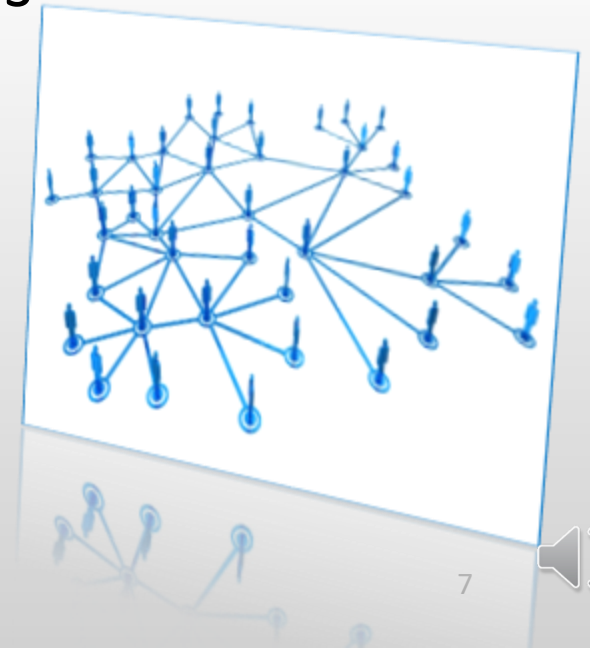
- It is the use of **automated** or **semi-automated methods** to process commercial or scientific data.
- Typically, this uses relatively simple, **repetitive activities to process large volumes** of similar information.



Dissemination

The most popular dissemination methods include the following:

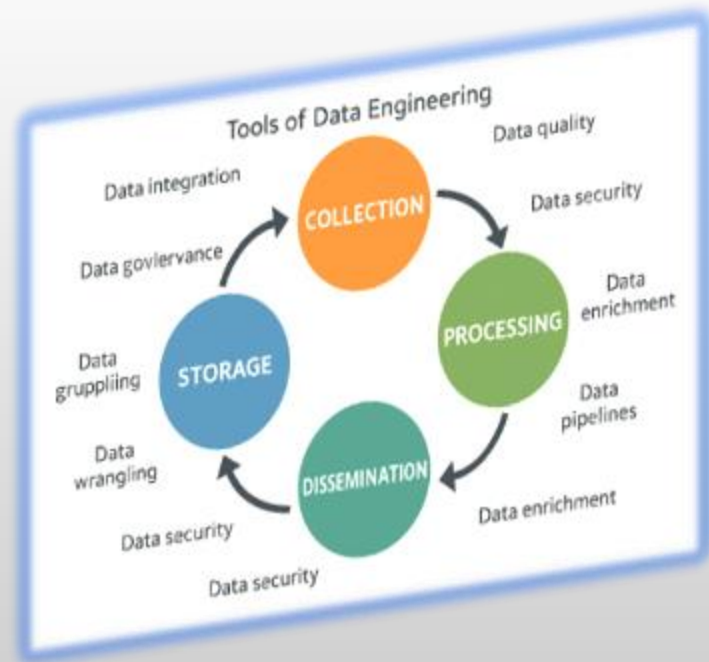
- Digital Content for Sale
- List/Reports/Queries
- Inter/Intra Net
- Business Intelligence
- Forecasting
- New Product/Service
- Knowledge Creation/Mining
- Web Forms
- PDA
- Personalized Emails
- Publications
- File Transfer
- Conversations
- Etc.



Data Engineering and the Data Lifecycle

Data Engineering is to data what Software Engineering is to software.

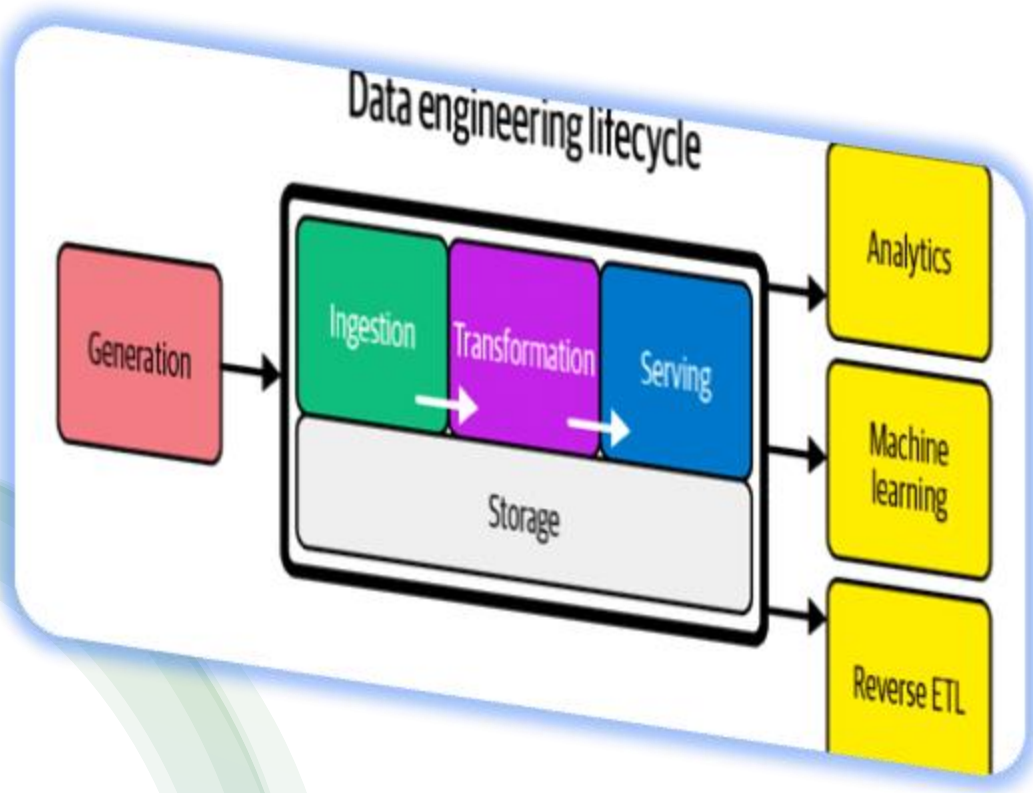
Just as software evolves through development stages, data too follows a lifecycle—from its creation and maturation to its eventual archiving or deletion.



- Data Engineering is a discipline focused on managing the entire lifecycle of data.
- Throughout this lifecycle, data is generated, collected, transformed, disseminated, and stored using a wide array of devices, techniques, tools, and systems.
- The choices made at each stage—driven by downstream use cases—determine whether raw data becomes a high-quality, valuable asset.
- Data Engineering is the discipline that guides these choices, ensuring that data is not only usable but also trustworthy, secure, and impactful.



Data Engineering



Data Engineering involves “the development, implementation, and maintenance of systems and processes that *take in raw data and produce high-quality, consistent information that supports downstream use cases, such as analysis and machine learning*”.

“A data engineer *manages the data engineering lifecycle*, beginning with getting data from source systems and ending with serving data for use cases, such as analysis or machine learning”.



Data Profiling for Data Engineering

- Data Profiling is an essential process carried out at different phases of the Data Engineering Lifecycle (DEL), aimed at **assessing the quality of data** and **inform downstream transformations**.
- Data profiling may begin during the Data Generation and Ingestion phases of the DEL, to allow understanding of the **structure and quality of incoming data** to be obtained.
- However, data profiling is most predominant just before the Data Transformation phase, when data analysis is required for **understanding data types and distributions, missing values, duplicated, outliers and relationships between features or fields**.



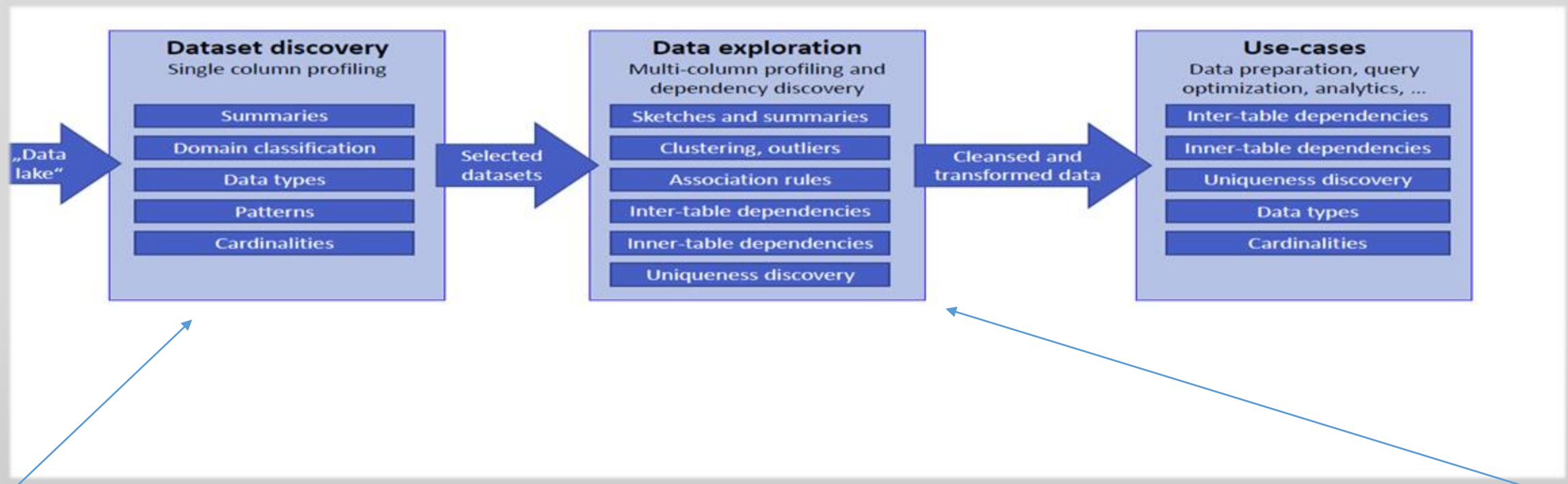
Data Profiling Defined

- The term **Data Profiling** refers to the process of applying available methods to efficiently analyse a dataset.
- This process typically involves the **scanning** of a dataset for the **collection of metadata**, e.g., data types, data value patterns, completeness of a table column, keys and foreign keys of a relational table, etc¹.
- While efficient scanning of a dataset is an important issue, our focus is on the kind of metadata that needs to be collected to **identify the presence of data quality issues and how to best model the data for future storage**.

¹ *Felix Naumann. 2014. Data profiling revisited. SIGMOD Rec. 42, 4 (December 2013), 40–49. DOI:<https://doi.org/10.1145/2590989.2590995>*



General Data Profiling Workflow



Single column profiling encompasses the **analysis of values in a single column**. Examples typically range from simple counts and aggregation functions to analysis of data distributions and the discovery of patterns.

Multicolumn profiling refers to the set of activities that can be applied to a single column but that allows for the **analysis of inter-value dependencies across columns**, which can result from the application of association rules, clustering and outlier detection algorithms.



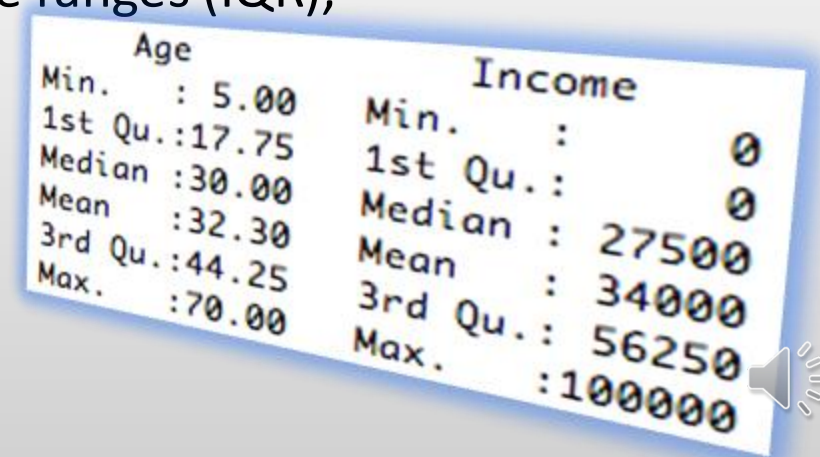
Data Exploration

- Some data scientists claim that Data Exploration is a separate process that follows Data Profiling.
- Data Exploration allows understanding of the meaning, distribution, and relationships within the data to be obtained, and involves the use of visualisations (e.g., histograms, scatterplots, etc.), data summarisations (e.g., mean, median, standard deviation, etc.), discovery of patterns, trends, and correlations.
- However, we will consider **Data Exploration as a part of the Data Profiling Process**, as the figure in the previous slide suggests.



Descriptive Data Summarization (DDS) as Data Profiling Metadata

- DDS is obtained by applying techniques that identify the **typical properties** of a dataset and highlight which data values should be treated as noise or outliers.
- Thus, DDS provides an **overall picture** of your data.
- Examples of DDS metadata include the following:
 - Measures of the **central tendency** of data, e.g., mean (average), median, mode and midrange.
 - Measures of **data dispersion**: quartiles, interquartile ranges (IQR), and variance.

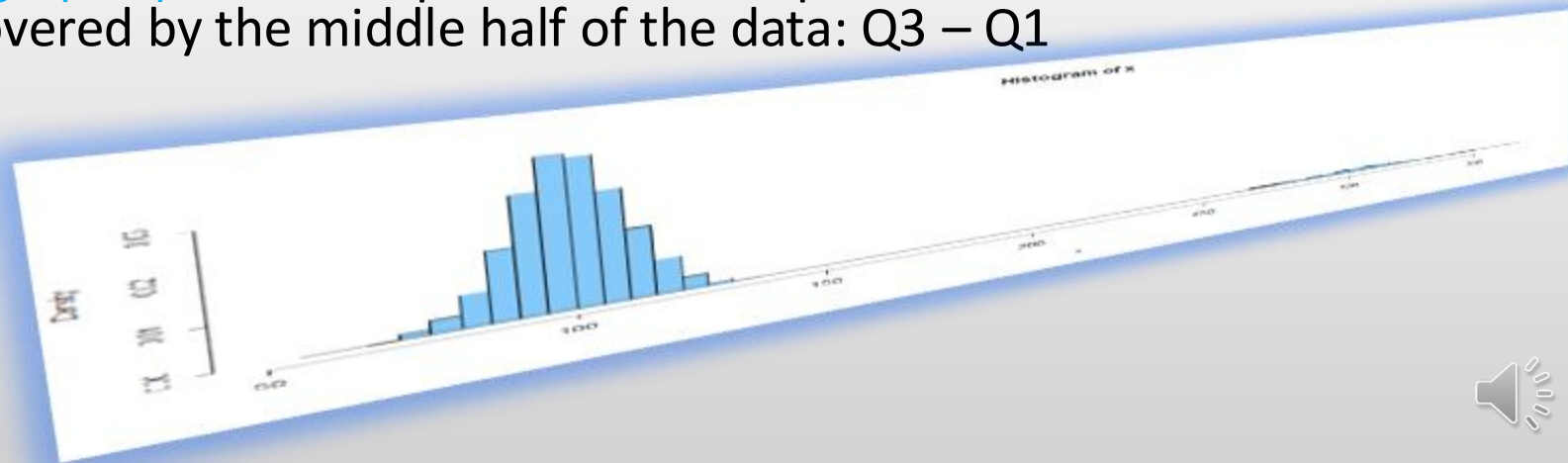


Age		Income	
Min.	: 5.00	Min.	: 0
1st Qu.	: 17.75	1st Qu.	: 0
Median	: 30.00	Median	: 27500
Mean	: 32.30	Mean	: 34000
3rd Qu.	: 44.25	3rd Qu.	: 56250
Max.	: 70.00	Max.	: 100000

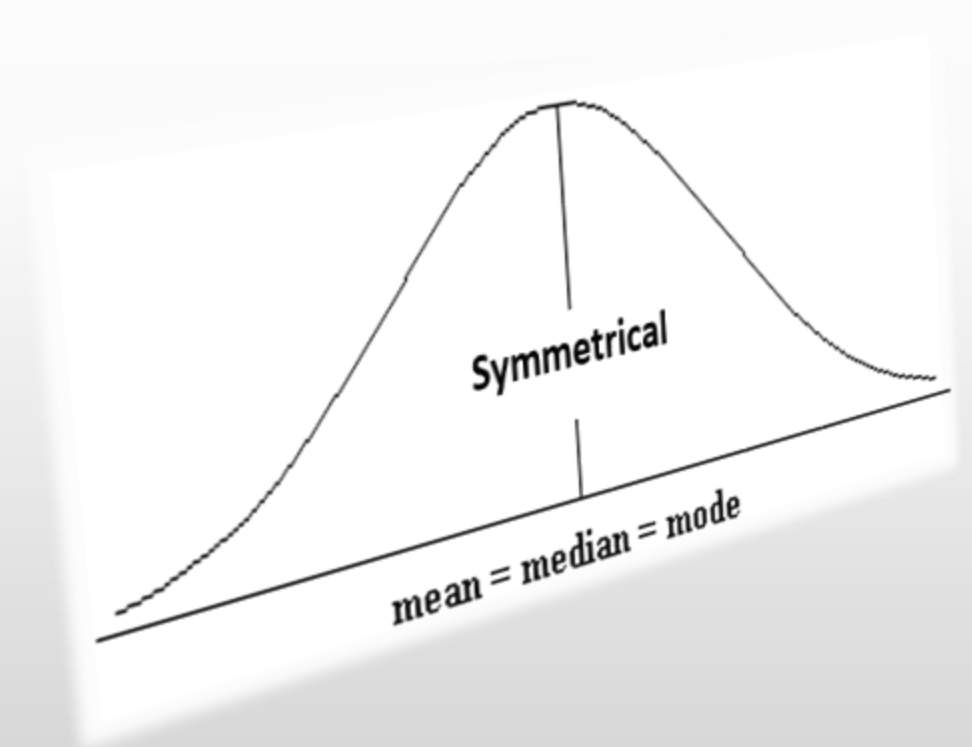
Most Popular DDS Measures

- **Range**: $\max() - \min()$.
- **Kth percentile**: the value x_i having the property that k percent of the data entries lie at or below x_i .
- **Median**: it is the 50th percentile.
- **Quartiles**: the most used percentiles:
 - 1st quartile (Q1): the 25th percentile
 - 2nd quartile (Q2): median (the 50th percentile)
 - 3rd quartile (Q3): the 75th percentile.
- **Inter-quartile range (IQR)**: it is a simple measure of spread that gives the range covered by the middle half of the data: $Q3 - Q1$

Holistic Measures



- Main measures of **central tendency**:
 - Mean (avg).
 - Median.
 - Mode.
 - Midrange.

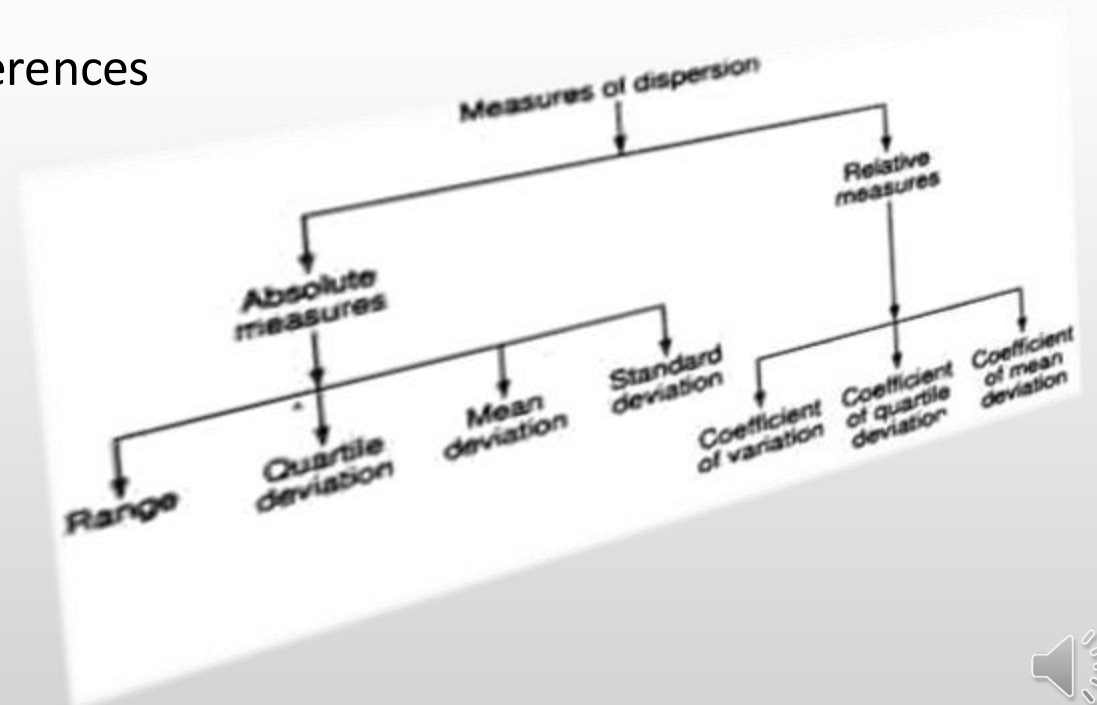


- Main measures of **data dispersion**:

- Quartiles.
- Inter-quartile range (IQR).
- Variance
 - It is the average of the squared differences from the Mean.
- Standard deviation
 - It is the square root of the Variance.

- Both Variance and Standard deviation are zero when there is no spread, that is when all observations have the same value, otherwise it is greater than zero.

- They are both algebraic measures, and so their computation is **scalable in large DBs**.



DDS Measures for Skewed Data

- No **single** numerical measure of spread, such as IQR, is very useful for describing **skewed distributions** of data, as the spreads of two sides of this type of distribution are unequal.
- It is more informative to obtain the two quartiles **Q1** and **Q3** along with the **median** in this case. However, these contain no information about the endpoints (e.g., tails) of the data distribution.
- Adding the lowest (**min**) and the highest (**max**) data values to Q1, Q3 and median (i.e., **the five-number summary**) can help obtain a fuller summary of the shape of the distribution.

Skewed Data are characterised by their uneven distribution of data values.

