Chapter 2: Data Preprocessing and Exploration

Dr. Rochdi Boudjehem

جامعة 8 ماي 1945 قالمة
UNIVERSITE 8 MAI 1945 GUELMA

2024 - 2025

2nd-Year Professional

Computer Science Department - University of 8 Mai 1945 Guelma

Table of contents

II Chapter 2: Data Preprocessing and Exploration

1. Data types and formats

Warning

Before digging into **Data Mining process** and **techniques**, it is imperative to understand the different **Data Types** to be able to handle them correctly.

Fundamental

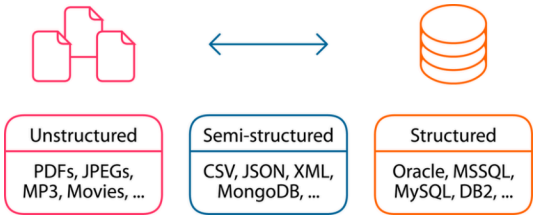
Data comes in various types and formats.

Understanding these types and formats is essential for effective preprocessing.

Definition

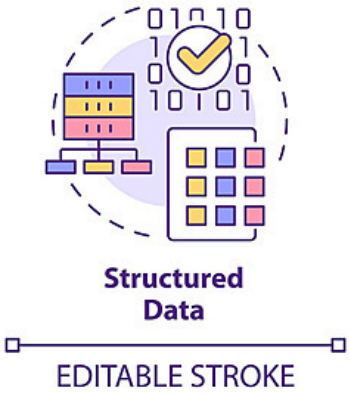
The most common types of data include:

1. Structured Data
2. Unstructured Data
3. Semi-Structured Data



1.1. Structured Data

Definition



This type of data is organized in a **tabular format**, with:

- **rows** representing **records** and
- **columns** representing **attributes**.

Example

Examples include **relational databases**, **Excel Spreadsheets**, and **CSV files**.

Note

Structured data is easy to analyze using traditional Data Mining techniques.

1.2. Unstructured Data

Definition

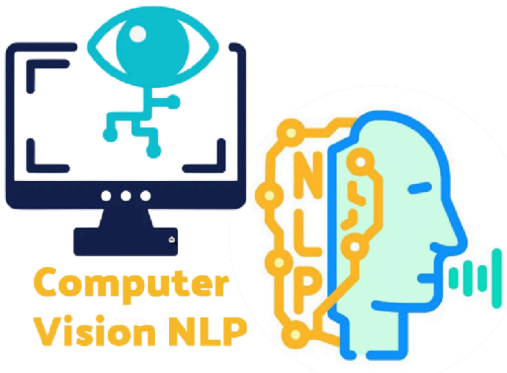
Unstructured data does not have a **predefined format**.

Example

Unstructured data could include:

- **audio** , **video**
- **text**, or **images**,

Note

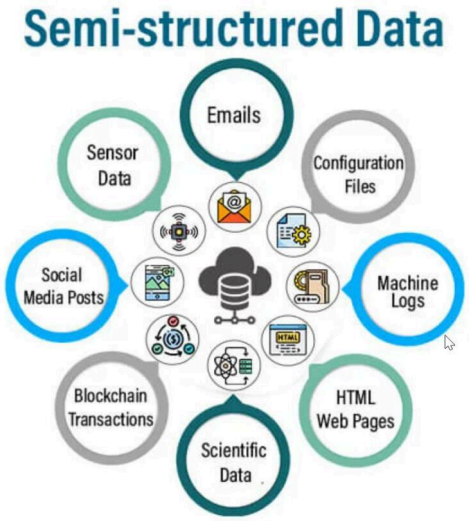


Analyzing unstructured data requires specialized techniques such as ***NLP*** p.16 for ***text*** or **Computer Vision** for ***images***.

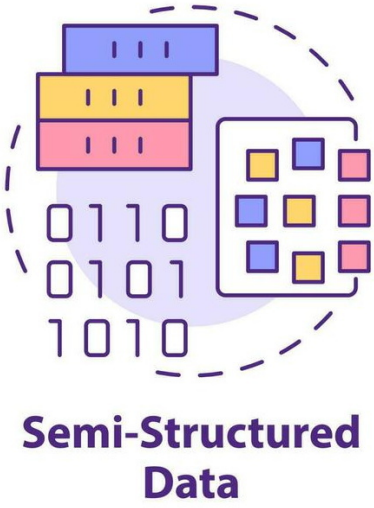
1.3. Semi-Structured Data

Definition

Semi-structured data maintains some degree of organization but does not follow the strict structure of traditional structured data, like databases.



Extra



It is more adaptable and tolerant of different data formats because it does not have a set schema or data model.

Example

Examples include ***JSON*** p.16,***XML*** p.16 , and markup languages, which are commonly used in web applications.

Advice

Understanding the type and format of the data is crucial for selecting appropriate preprocessing techniques.

For example, **text data** may require **tokenization** and ***stemming*** [p.15], while **numerical data** may require **normalization** or **scaling**.

2. Data Preprocessing Process

Fundamental

This stage is the **initial** step in the **DM process**.

It ensures that the data is properly prepared for analysis, making it easier for DM algorithms to correctly detect patterns.

This stage plays a crucial role in enhancing the overall **quality** and **reliability** of the **data**.

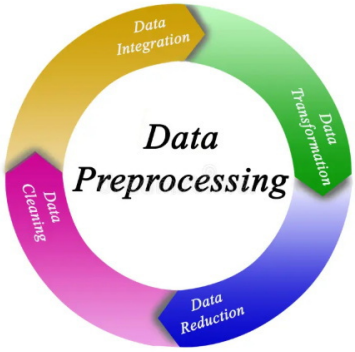
The preprocessing phase typically includes tasks such as :

- handling missing values,
- removing noise,
- normalizing data,
- and selecting relevant features.

Method

The "Data Preprocessing" process involves four stages:

1. Data Cleaning
2. Data integration,
3. Data Transformation,
4. Data Reduction,



2.1. Data Cleaning

Definition

Data cleaning is the process of **detecting** and **correcting :**

- **errors**,
- **inconsistencies**,
- and **inaccuracies** in the **data**.

This step is essential for ensuring the quality of the data and improving the accuracy of the analysis.

Method

Common data cleaning tasks include:

1. Handling Missing Values
2. Removing Noise
3. Correcting Inconsistencies

a) Handling Missing Values

Reminder

Missing values can occur due to :

- data entry errors,
- sensor malfunctions,
- or other issues.

Method

There are several strategies for handling missing values, including:

1. **Removing records** with missing values (if the number of such records is small).
2. Filling in missing values using statistical methods such as **mean**, **median**, or **mode**.
3. Using advanced techniques such as **imputation** or **predictive modeling** to **estimate missing values**.

b) Removing Noise

Definition

Noise refers to **random errors** or **outliers** in the data that can distort the analysis.

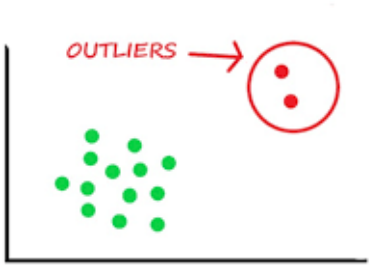
Techniques for removing noise include:

i) Smoothing

is creating an approximating function that attempts to capture important patterns in the data, while leaving out noise or other fine-scale structures/rapid phenomena.,

ii) Outlier Detection

: is the identification of rare items, events or observations which deviate significantly from the majority of the data and do not conform to a well defined notion of normal behavior



c) Correcting Inconsistencies

Inconsistent data can arise due to differences in:

- data entry formats,
- units of measurement,
- or coding schemes.

For example:

- dates may be recorded in different formats,
  - **Example:** DD/MM/YYYY vs. MM/DD/YYYY
- or categorical variables may have inconsistent labels.
  - **Example**: "Male" vs. "M".

These inconsistencies must be resolved to ensure the data is accurate and reliable.

2.2. Data Integration

Definition

Data integration involves combining data from multiple sources into a unified dataset.

This is often necessary when working with data from different departments, systems, or organizations.

Common challenges in data integration include:

1. Schema Integration
2. Entity Resolution

a) Schema Integration

Different datasets may use different schemas or attribute names, making it difficult to combine them.

Definition

Schema integration involves resolving these differences by mapping attributes and standardizing formats.

b) Entity Resolution

Entity Resolution involves identifying and merging records that refer to the same entity (e.g., the same customer or product) across different datasets.

Techniques for entity resolution include ***record linkage*** p.15 and ***deduplication*** p.15.

2.3. Data Transformation

Once the data is integrated, it may need to be transformed into a suitable format for analysis.

Common data transformation techniques include:

1. Normalization
2. Aggregation
3. Encoding Categorical Variables

a) Normalization

This involves **scaling numerical** data to a **standard** range (***e.g., 0 to 1***) to ensure that all attributes contribute equally to the analysis.

Normalization is particularly important for algorithms that are sensitive to the scale of the data, such as **k-means clustering** or **gradient descent**.

b) Aggregation

This involves **summarizing** data at a **higher level** of granularity

Example: calculating monthly sales from daily sales data). Aggregation can:

- help to reduce the complexity of the data
- and highlight important trends.

c) Encoding Categorical Variables

Many Data Mining algorithms require **numerical input**, so **categorical variables** (e.g., gender, product category) must be encoded as numbers.

Common encoding techniques include ***one-hot encoding*** p.15 and ***label encoding*** p.15.

- **Example**: gender 0/1 instead of M/F
- **Example**: 0/1/2 instead of Tall/Medium/Short

2.4. Data Reduction

In this stage, **Data Reduction** techniques are used to **reduce** the **size** of the **dataset** while preserving its essential characteristics.

This is particularly important when working with **large datasets**, as it can significantly reduce the computational cost of the analysis.

Common data reduction techniques include:

1. Dimensionality Reduction
2. Sampling
3. Discretization

a) Dimensionality Reduction

This involves reducing the number of attributes in the dataset by:

- selecting the most relevant features
- or creating new features that capture the essential information.

Techniques for dimensionality reduction include *PCA* [p.16] and **Feature Selection** algorithms.

b) Sampling

This involves selecting a subset of the data for analysis.

Sampling can be:

- random (e.g., selecting records at random) or
- stratified (e.g., ensuring that each subgroup in the data is represented proportionally).

c) Discretization

This involves converting continuous numerical data into discrete intervals or categories.

For example, age can be discretized into ranges such as 0–18, 19–35, and 36–50.

Discretization can simplify the analysis and make it easier to interpret the results.

3. Exploratory Data Analysis (EDA)

Definition

EDA is the process of **summarizing** and **visualizing** the data to gain **insights** and **identify patterns**.

EDA is an essential step in the data mining process. playing a key role in **determining** the **appropriate algorithms** and **techniques** to be **applied**.

Method

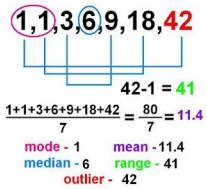
Common EDA techniques include:

1. Summary Statistics
2. Data Visualization
3. Correlation Analysis

3.1. Summary Statistics

This involves calculating **descriptive statistics** such as :

- **the mean**,
- **the median**,
- **the standard deviation**,
- or **quartiles** to summarize the data.

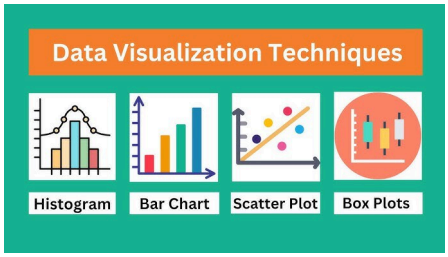


3.2. Data Visualization

Visualization techniques such as:

- **histograms**,
- **scatter plots**, and
- **box plots**

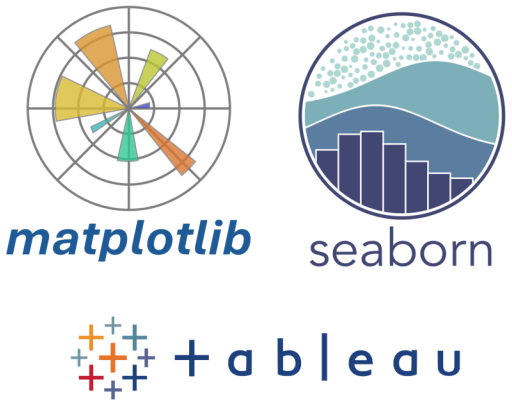
can help to identify **trends**, **outliers**, and **relationships** between **variables**.



Tools like:

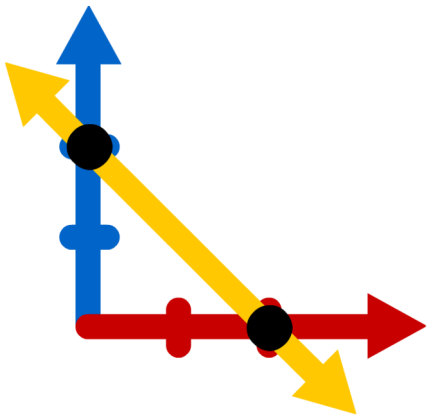
- **Matplotlib**,
- **Seaborn**, and
- **Tableau**

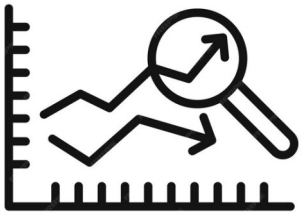
are commonly used for data visualization.



3.3. Correlation Analysis

This involves measuring the **strength** and **direction** of the **relationship** between **variables**.





Correlation analysis can help to identify potential **predictors** for **predictive modeling**.

3.4. Example

Conclusion

Data preprocessing and exploration are essential steps in the Data Mining process, as they ensure the quality and suitability of the data for analysis. By cleaning, transforming, and reducing the data, and by using exploratory techniques to gain insights, professionals can uncover meaningful patterns and build accurate models. Effective preprocessing and exploration lay the foundation for successful Data Mining and enable organizations to make data-driven decisions.

Glossary

Data deduplication

Data deduplication is a process that eliminates excessive copies of data and significantly decreases storage capacity requirements.

Deduplication can be run as an inline process as the data is being written into the storage system and/or as a background process to eliminate duplicates after the data is written to disk.

Label Encoding

Label Encoding is a technique that is used to convert categorical columns into numerical ones so that they can be fitted by machine learning models which only take numerical data.

When working with datasets, we often encounter categorical data, which needs to be converted into numerical format for machine learning algorithms to process. For example, a column representing car brands ("Toyota", "Honda", "Ford") or colors ("Red", "Blue", "Green") is categorical data for Cars Dataset. One common method to achieve this is **Label Encoding.**

One Hot Encoding

One Hot Encoding is a ***method for converting categorical variables into a binary format.*** It creates new columns for each category where **1** means the category is present and **0** means it is not. The primary purpose of One Hot Encoding is to ensure that categorical data can be effectively used in machine learning models.

Record Linkage

Record Linkage is the process in which records or units from different data sources are joined together into a single file using non-unique identifiers, such as names, date of birth, addresses and other characteristics

Stemming

Stemming is a text preprocessing technique in natural language processing (NLP). Specifically, it is the process of reducing inflected form of a word to one so-called “stem,” or root form, also known as a “lemma” in linguistics.1 It is one of two primary methods

Abbreviation

JSON: JavaScript Object Notation
NLP: Natural Language Processing
PCA: Principal Component Analysis
XML: Extensible Markup Language