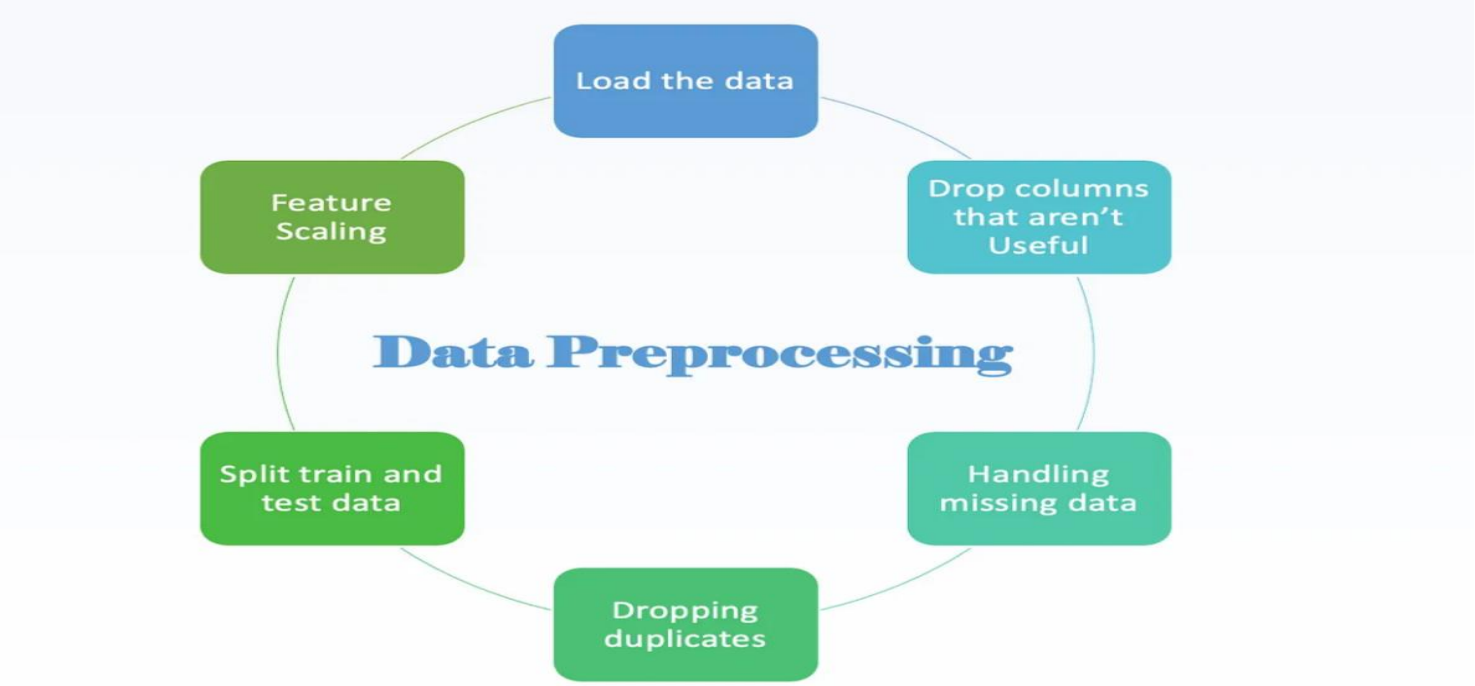
WP4 : Privacy Supporting Technologies

T4.1: Data Preprocessing and Preparation

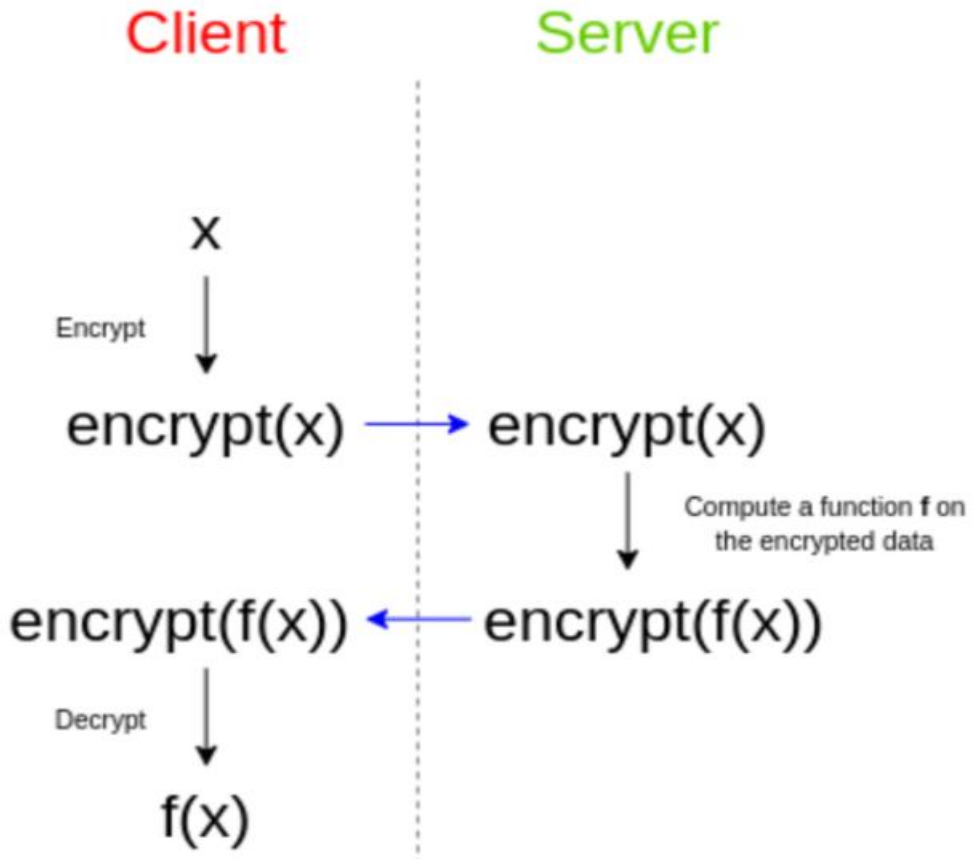
Aims of the Task

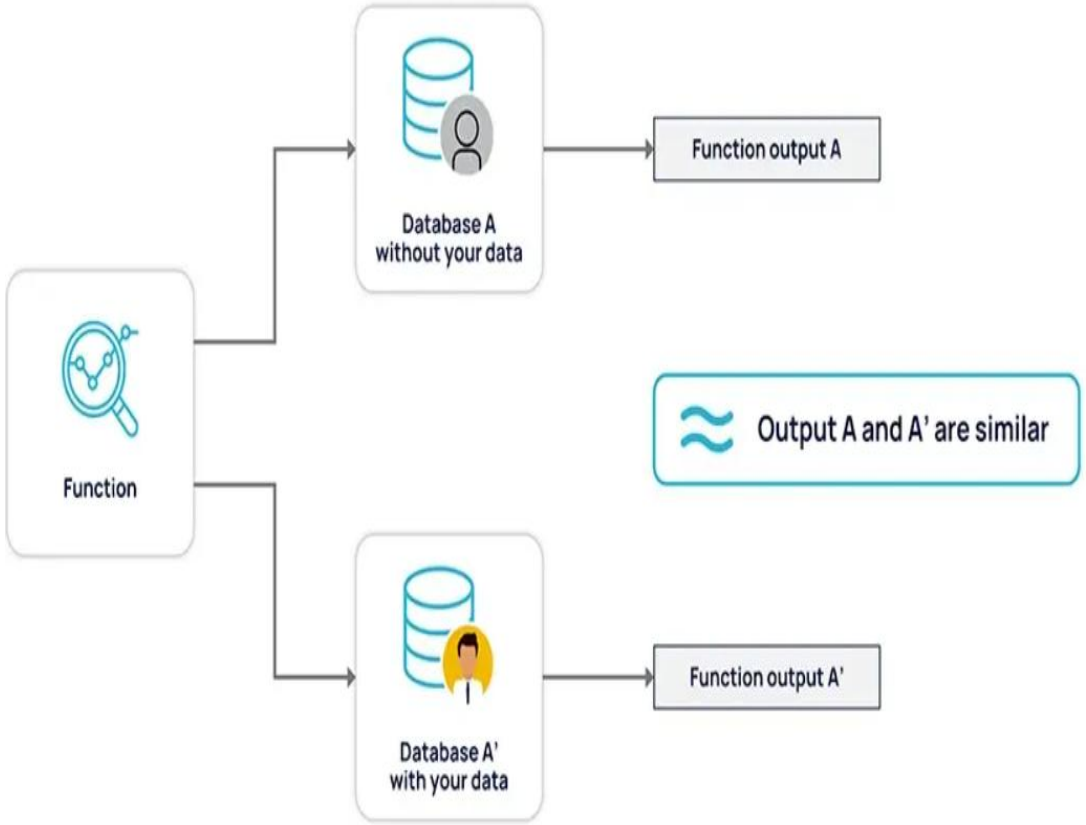
- ***Aim:*** *Develop the data preprocessing and preparation techniques required to format the data in the proper way for the application of the Privacy Preserving (PP) technologies on them and for facilitating any post analysis (e.g., Machine Learning).*



Data Preprocessing

- *The data preprocessing takes into account the requirements of the privacy tools*





Data Preprocessing

Feature	Correlation with Target	Description
Age	0.02	Weakly correlated
Income	0.68	Strongly correlated
Height	0.03	Weakly correlated
Weight	0.05	Weakly correlated
Spending Score	0.72	Strongly correlated
Location Index	0.01	Weakly correlated

Data Preprocessing

Feature	Correlation with Age	Correlation with Income	Correlation with Spending Score
Age	1.0	0.5	0.4
Income	0.5	1.0	0.8
Spending Score	0.4	0.8	1.0

Private Identifiable Information Extraction

- ***Aim:*** *To extract Private Identifiable Information (PIIs)*

  ✓Regular expressions (REGEX) and Named Entity Recognition (NER) are used

Private Identifiable Information Extraction

Name	Contact Information
John Doe	john.doe@example.com, 555-1234
Jane Smith	jane.smith@workplace.com, 555-5678
Alice Johnson	alice.j@example.org, 555-8765

Name	Extracted Email	Extracted Phone Number
John Doe	john.doe@example.com	555-1234
Jane Smith	jane.smith@workplace.com	555-5678
Alice Johnson	alice.j@example.org	558-8765

Private Identifiable Information Extraction

- **Sample Unstructured Text Data**
  - ✓ "John Doe lives in New York and his date of birth is 1985-12-15.“
  - ✓ "Jane Smith, born on 1992-06-22, works in Chicago."

Text	Extracted Entities
John Doe lives in New York and his date of birth is 1985-12-15.	Name: John Doe, Location: New York, Date of Birth: 1985-12-15
Jane Smith, born on 1992-06-22, works in Chicago.	Name: Jane Smith, Date of Birth: 1992-06-22, Location: Chicago

PII Type	Count Extracted	Example Extracted Data
Names	2	John Doe, Jane Smith
Dates of Birth	2	1985-12-15, 1992-06-22
Locations	2	New York, Chicago

Private Identifiable Information Extraction

- **Example DICOM Metadata:**
  - ✓ In a real-world DICOM file, the metadata might contain fields like the following, indicating PIIs:
    - **Patient's Name**: John Doe
    - **Patient ID**: 12345
    - **Patient's Birth Date**: 1980-01-01
    - **Study Date**: 2024-11-01
    - **Study Description**: Chest X-ray

Thank you for your attention

Find out more:

https://encrypt-project.eu/

encrypt-project

@encrypt_project

encrypt