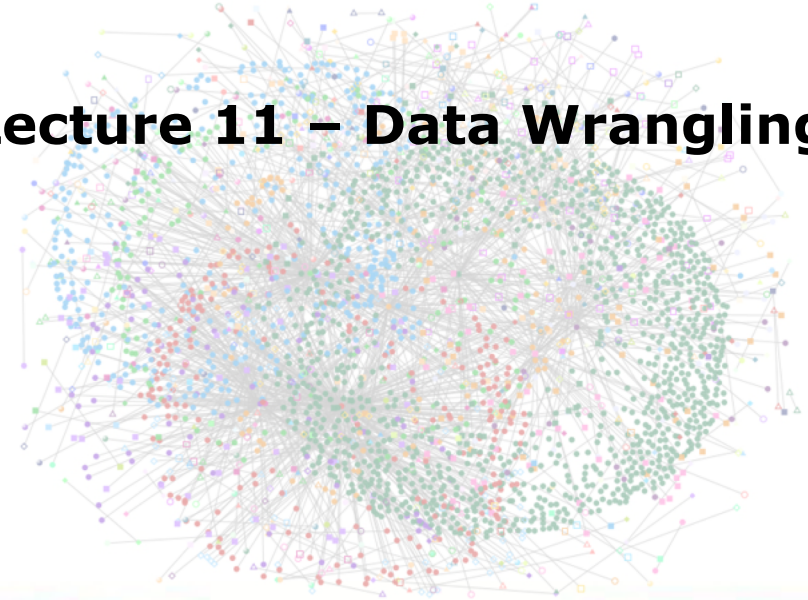
Fakultas Informatika
School of Computing
Telkom University

CTI-3A3
Applied Social Network Analysis

Dr. Warih Maharani

wmaharani.staff.telkomuniversity.ac.id



Lecture 11 – Data Wrangling

Outline

- Data Preparation
- Data Wrangling

Course Learning Outcomes (CLO)

- use software to implement statistical models of social networks to analyzed network formation and evolution;
- use software to simulate the dynamics of networks based on social network models.

Creating networks from data

From Raw Data to Networks

When creating networks from data we need to make a number of design descisions

- How will we collect the data?
- What type of entity (node) to use and how to extract it?
- What type of relationship or interaction do our links represent?
- What time period?
- Directed or undirected links?

How we make these decisions depends on:

- the task we're trying to achieve
- the model and algorithm we are using

Facts

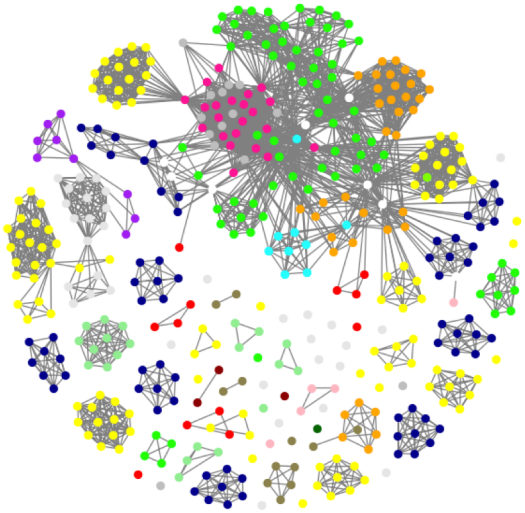
Data preparation takes 60 to 80 percent of the whole analytical pipeline

Various programming languages, frameworks and tools are available for data cleansing and feature engineering

Data wrangling as important add-on to data preprocessing; it's best used within a visual analytics tool to avoid breaking the analysis flow

Visual analytics tools and open source data science components like R, Python, KNIME or RapidMiner are complementary

Data Preparation



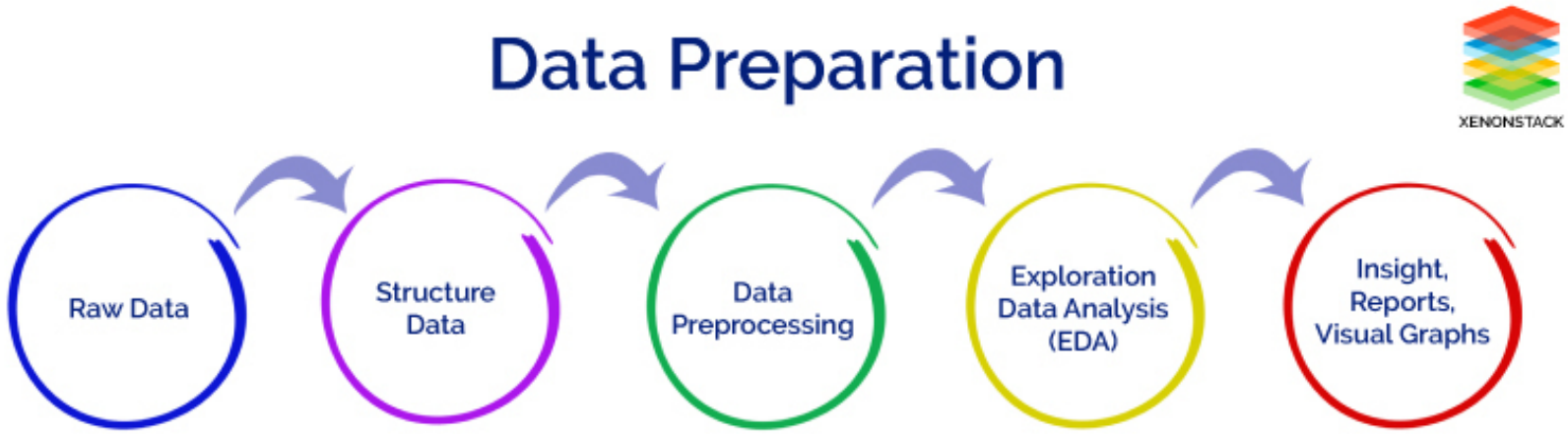
Data Preparation

A key task when you want to build an appropriate analytic model is the **integration and preparation** of data sets from various sources like files, databases, big data storage, sensors or social networks.

This step can take up to 80 percent of the whole analytics project.

Data Preparation = Data Cleaning + Feature Engineering

- Includes data cleansing and feature engineering
- Data preparation cannot be fully automated; at least not in the beginning



Data Cleansing

- Data Cleansing puts data into the right shape and quality for analysis:
  - Basics (select, filter, removal of duplicates, …)
  - Sampling (balanced, stratified, ...)
  - Data Partitioning (create training + validation + test data set, ...)
  - Transformations (normalisation, standardisation, scaling, pivoting, ...)
  - Binning (count-based, handling of missing values as its own group, …)
  - Data Replacement (cutting, splitting, merging, ...)
  - Weighting and Selection (attribute weighting, automatic optimization, ...)
  - Attribute Generation (ID generation, ...)
  - Imputation (replacement of missing observations by using statistical algorithms

Feature Engineering

Feature Engineering selects the right attributes to analyze. You use domain knowledge of the data to select or create attributes that make machine learning algorithms work. Feature Engineering process includes:

- Brainstorming or testing of features
- Feature selection
- Validation of how the features work with your model
- Improvement of features if needed
- Return to brainstorming / creation of more features until the work is done

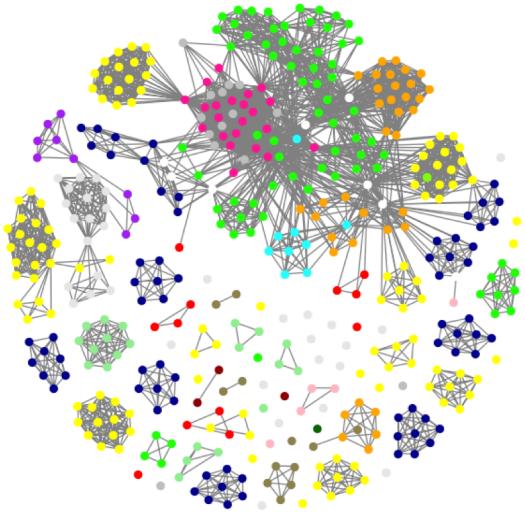
Data Preparation

- Data preparation occurs in different phases of an analytics project:
  - **Data Preprocessing**: Preparation of data directly after accessing it from a data source. Typically realized by a developer or data scientist for initial transformations, aggregations and data cleansing. This step is done before the interactive analysis of data begins. It is executed once.
  - **Data Wrangling**: Preparation of data during the interactive data analysis and model building. Typically done by a data scientist or business analyst to change views on a dataset and for features engineering. This step iteratively changes the shape of a dataset until it works well for finding insights or building a good analytic model.

Data Preprocessing

- **Data Preprocessing** is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis
- Includes:
  - Data Cleaning
  - Data Integration
  - Data Transformation
  - Data Reduction

Data Wrangling



Data Wrangling

- **Data Wrangling** is a technique that is executed at the time of making an interactive model. In other words, it is used to convert the raw data into the format that is convenient for the consumption of data.
- This technique is also known as **Data Munging**.
- This method also follows certain steps such as after extracting the data from different data sources, sorting of data using the certain algorithms are performed, decompose the data into a different structured format and finally store the data into another database.

The need of Data Wrangling

Data Wrangling is an **important aspect** of implementing the model.

Therefore, data is converted to the **proper feasible format** before applying any model to it. By performing filtering, grouping, and selecting appropriate data accuracy and performance of the model could be increased.

Another concept is that when time-series data has to be handled every algorithm is executed with different aspects. Therefore Data Wrangling is used to convert the time series data into the required format of the applied model. In simple words, the complex data is transformed into a usable format for performing analysis on it.

Why is Data Preparation Important?

- **Inaccurate data (missing data) –** There are many reasons for missing data such as data is not continuously collected, a mistake in data entry, technical problems with biometrics, and much more, which requires proper Data Preparation.
- **The presence of noisy data (erroneous data and outliers) –** The reasons for the existence of noisy data could be a technological problem of gadget that gathers data, a human mistake during data entry and much more.
- **Inconsistent data –** The presence of inconsistencies are due to the reasons such that existence of duplication within data, human data entry, containing mistakes in codes or names, i.e., violation of data constraints and much more necessitate Data Preparation and analysis.

Why is Data Wrangling Important?

- To handle the issue of **Data Leakage**
- Data Leakage is responsible for the cause of an invalid model due to the over-optimization of the applied model
- Data Leakage is the term used when the data from outside, i.e., not part of the training dataset is used for the learning process of the model. This additional learning of information by the applied model will disapprove of the computed estimated performance of the model
- Data Leakage can be demonstrated in many ways
  - The Leakage of data from test dataset to the training data set.
  - Leakage of computed correct prediction to the training dataset.
  - Leakage of future data into the past data.
  - Usage of data outside the scope of the applied algorithm

How is Data Preprocessing performed?

Data Preprocessing is carried out to remove the cause of unformatted real-world data

How missing data can be handled during Data Preparation:

- **Ignoring the missing record –** It is the simplest and efficient method for handling the missing data. But, this method should not be performed at the time when the number of missing values is immense or when the pattern of data is related to the unrecognized primary root of the cause of the statement problem.
- **Filling the missing values manually –** This is one of the best-chosen methods of Data Preparation process. But there is one limitation that when there are large data set, and missing values are significant then, this approach is not efficient as it becomes a time-consuming task.
- **Filling using computed values –** The missing values can also be occupied by computing mean, mode or median of the observed given values. Another method could be the predictive values in Data Preprocessing is that are computed by using any Machine Learning or Deep Learning tools and algorithms. But one drawback of this approach is that it can generate bias within the data as the calculated values are not accurate concerning the observed values.

Deal with the **inconsistent data**

How is Data Wrangling performed?

- Data Wrangling is conducted to minimize the effect of Data Leakage while executing the model.
- In other words if one considers the complete data set for normalization and standardization, then the cross-validation is performed for the estimation of the performance of the model leads to the beginning of data leakage.

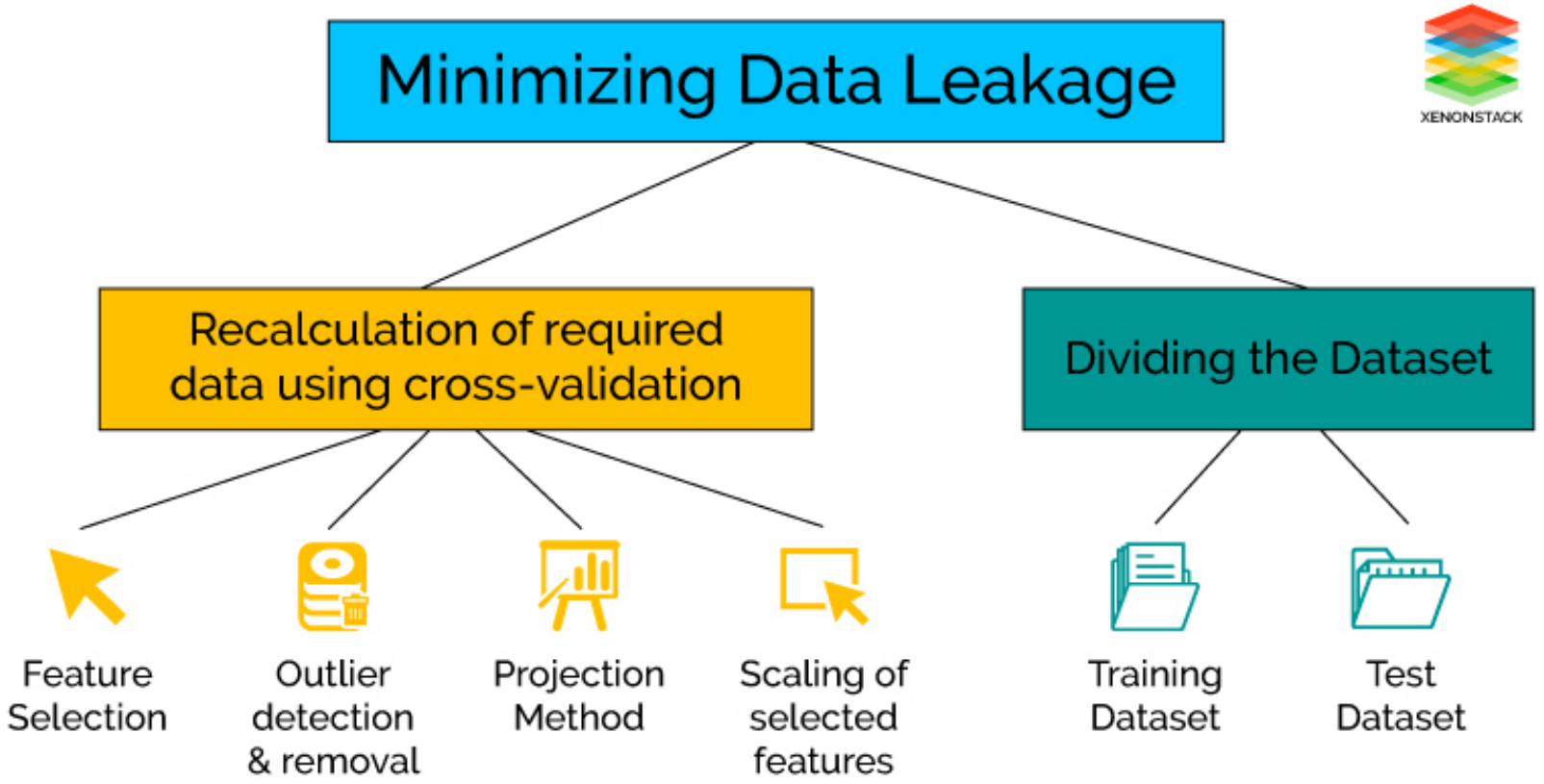
How is Data Wrangling performed?

The effect of Data Leakage could be minimized by recalculating for the required Data Preparation during the cross-validation process that includes feature selection, outliers detection, and removal, projection methods, scaling of selected features and much more

Another solution for better Data Preparation is that dividing the complete dataset into a training data set that is used to train the model and validation dataset which is used to evaluate the performance and accuracy of the applied model

But, the selection of the model is made by looking at the results of the test data set in the cross-validation process. This conclusion will not always be valid as the sample of test data set could vary, and the performance of different models are evaluated for the particular type of test dataset. Therefore, while selecting the best model test error is overfitting

How is Data Wrangling performed?



Data Preparation vs Data Wrangling

- **Data Preprocessing** steps are performed before the Data Wrangling.
- In this case, Data Preprocessing data is prepared exactly after receiving the data from the data source. In this initial transformations, **Data Cleaning** or any aggregation of data is performed. It is executed once.
  - For example, we have data where one attribute has three variables, and we have to convert them into three attributes and delete the special characters from them. The concept of Data Preparation steps performed before applying any iterative model and will be executed once in the project.

Data Preparation vs Data Wrangling

On the other hand, **Data Wrangling** is performed during the iterative analysis and model building. This concept at the time of feature engineering. The conceptual view of the dataset changes as different models are applied to achieve a good analytic model.

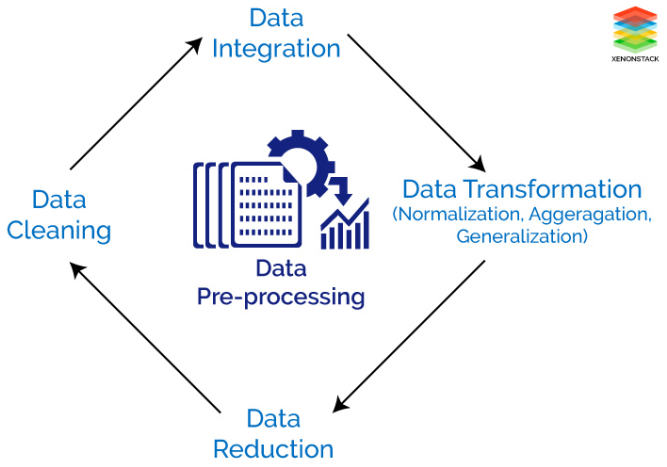
- For example, we have data containing 30 attributes where two attributes are used to compute another attribute, and that computed feature is used for further analysis. In this way, the data could be changed according to the requirement of the applied model, and Data Preparation can be effective.

Tasks of Data Preparation

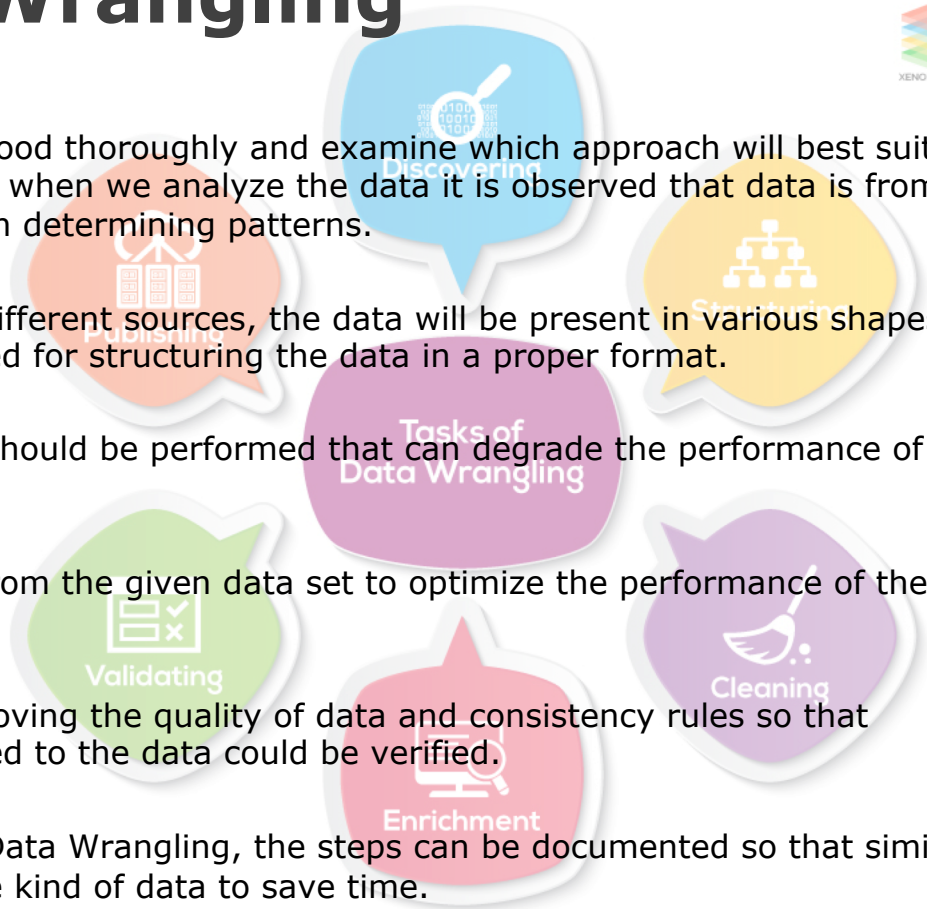
- **Data Cleaning**
  - In this step, the primary focus is on handling missing data, noisy data, detection, and removal of outliers minimizing duplication, and computed biases within the data.
- **Data Integration**
  - This process is used when data is gathered from various data sources and data are combined to form consistent data. This consistent data after performing data cleaning is used for Data Preparation and analysis.
- **Data Transformation**
  - This step is used to convert the raw data into a specified format according to the need of the model.
- **Data Reduction**
  - After the transformation and scaling of data duplication, i.e., redundancy within the data is removed and efficiently organize the data during Data Preparation.

Data Transformation

- **Normalization –** In this method, numerical data is converted into the specified range, i.e., between 0 and one so that scaling of data can be performed.
- **Aggregation –** The concept can be derived from the word itself, this method is used to combine the features into one. For example, combining two categories can be used to form a new group.
- **Generalization –** In this case, lower level attributes are converted to a higher standard.



Tasks of Data Wrangling



Discovering

- Firstly, data should be understood thoroughly and examine which approach will best suit. For example: if have weather data when we analyze the data it is observed that data is from one area and so primary focus is on determining patterns.

Structuring

- As the data is gathered from different sources, the data will be present in various shapes and sizes. Therefore, there is a need for structuring the data in a proper format.

Cleaning

- Cleaning or removing of data should be performed that can degrade the performance of the analysis.

Enrichment

- Extract new features or data from the given data set to optimize the performance of the applied model.

Validating

- This approach is used for improving the quality of data and consistency rules so that transformations that are applied to the data could be verified.

Publishing

- After completing the steps of Data Wrangling, the steps can be documented so that similar steps can be performed for the same kind of data to save time.

How Data Wrangling improves Data Analytics?

- With the advancement in the technology and generation of data, data is collected from various sources. Therefore, in the Data Preparation process managing data in different formats Data Wrangling is necessary.
- As the simple Data Preparation and analysis methods alone are not feasible for the complex problem statement, Data Wrangling is introduced which simplifies the analysis process of a complex issue.
- In this way, Data Wrangling is used for improving the analysis process of complex problems during Data Preparation.

Data Wrangling vs ETL

- Data Wrangling technology is used by business analysts, users engaged in business, and managers.
- On the other hand, ETL (Extract, Transform, and Load) is employed by IT Professionals. They receive the requirements from business people and then they use ETL tools to deliver the data in a required format.
- Data Wrangling is used to analyze the data that was gathered from different data sources. It is designed specially to handle diverse and complex data of any scale. But in the case of ETL, it can handle structured data that was originated from different databases or operating systems.
- The primary task of the Data Wrangling method is to manage the newly generated data from various sources for the analysis process whereas the goal of ETL is to extract, transform and load the data into the central enterprise Data Warehouse for performing analysis process using business applications.

Data Preprocessing Tools

Data Preprocessing in R

- R a framework that consists of various packages that can be used for Data Preprocessing like dplyr etc.

Data Preprocessing in Weka

- Weka is a software that contains a collection of Machine Learning algorithms for the Data Mining process. It consists of Data Preprocessing tools that are used before applying Machine Learning algorithms.

Data Preprocessing in RapidMiner

- RapidMiner is an open-source **Predictive Analytics Platform** for Data Mining process. It provides efficient tools for performing the exact Data Preprocessing process.

Data Preprocessing in Python

- Python is a programming language that provides various libraries that are used for Data Preprocessing.

Data Wrangling Tools

- **Data Wrangling in Tabula**
  - Tabula is a tool that is used to convert the tabular data present in pdf into a structured form of data, i.e., spreadsheet.
- **Data Wrangling in OpenRefine**
  - OpenRefine is open-source software that provides a friendly Graphical User Interface (GUI) that helps to manipulate the data according to your problem statement and makes Data Preparation process simpler. Therefore, it is highly useful software for the non-data scientist.
- **Data Wrangling in R**
  - R is an important programming language for the data scientist. It provides various packages like dplyr, tidyr, etc. for performing data manipulation.
- **Data Wrangling using Data Wrangler**
  - Data Wrangler is a tool that is used to convert real-world data into the structured format. After the conversion, the file can be imported into the required application like Excel, R, etc. Therefore, less time will be spent on formatting data manually.
- **Data Wrangling in CSVKit**
  - CSVKit is a toolkit that provides the facility of conversion of CSV files into different formats like CSV to JSON, JSON to CSV, and much more. It makes the process of data wrangling easy.
- **Data Wrangling using Python with Pandas**
  - Python is a language with Pandas library. This library helps the data scientist to deal with complex problems efficiently and makes Data Preparation process efficient.
- **Data Wrangling using Mr. Data Converter**
  - Mr. Data Converter is a tool that takes Excel file as an input and converts the file into required formats. It supports the conversion of HTML, XML, and JSON format.

Let's practice in Gephi

- Lab: wmaharani.staff.telkomuniversity.ac.id

THANK YOU