Text Mining

Lecture 01

Lecture Overview

Keungoui Kim

awekim@handong.edu

Handong Global University, 2024-2.

Class Overview

Lecturer

- Keungoui Kim
  - School of Applied Artificial Intelligence
  - Office hours
    - 10:00a.m. ~ 11:00a.m., Wednesday
    - Location: Room 306, Ebenezer Hall
  - Contact
    - email: awekim@handong.edu

Course Descriptions

- Text mining
  - Covering knowledge and techniques needed for “analyzing” texts

- Recommended prerequisite knowledge
  - Introduction to Big Data or Data science

  → Basic knowledge of the data analysis

- Regardless of the prerequisite knowledge, anyone who is willing to conduct data analysis in social science and learn natural language processing is welcome.

Course Objectives

- To learn the basic text mining skills and related theories
- Practice overall text analysis procedures and steps
- Simply speaking, this course will help students learn and get familiar with data analysis focused on text data

Weekly Schedule

Week	Contents	
Week 1	Introduction to Text Mining	
Week 2	Intermediate RPython Programming	Assignment 1
Week 3	Text Mining Principles	
Week 4	Text Exploration	Assignment 2
Week 5	Text Pre-processing I	Assignment 3
Week 6	Text Pre-processing II	Assignment 4
Week 7	Text Quantification	
Week 8	Midterm Exam	Covers week 1 - 7

Weekly Schedule

Week	Contents	
Week 9	Text Similarity - Proposal presentation	
Week 10	Text Network Analysis	Assignment 5
Week 11	Sentiment Analysis	
Week 12	Topic Modelling	Assignment 6
Week 13	Text Embedding I	
Week 14	Text Embedding II	
Week 15	Final Presentation	
Week 16	Final Exam	Covers week 1 - 15

Evaluation

- Evaluation
  - Attendance: 10%. Three lates = 1 absence (-1 pts)
  - Team Assignment: 30%
  - Team Project: 20%
  - Midterm exam: 20%
  - Final exam: 20%

  - Absolute evaluation

Announcement

- 100% contact lecture
  - You can either attend ZOOM or come to the classroom.

- A laptop (notebook) is required.

- Midterm and final exams will be conducted offline.
  - No excuses.

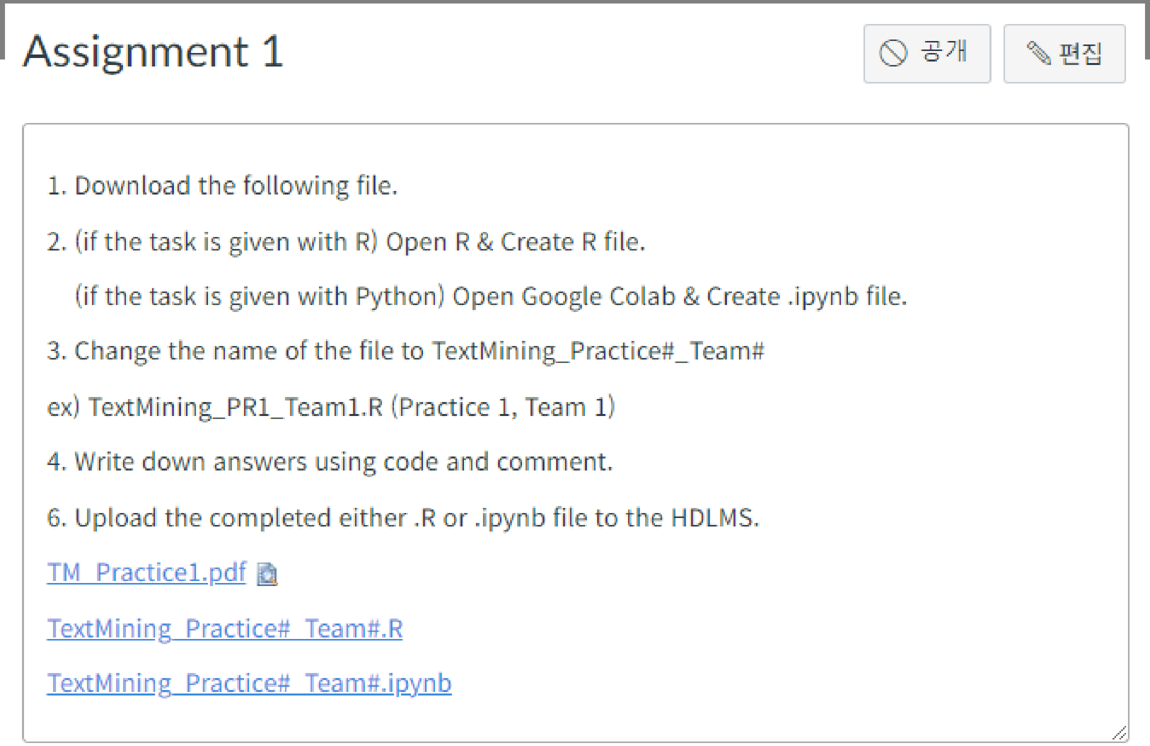
- In this lecture, R will be mainly used and Python will be used as supplementary
  - RPython?

Team Announcement

- Team
  - Assignments
  - Project

Team Assignment

- Team Assignment
  - Do your assignment with your teammates → **Honor Code**
  - Submit .R file to HDLMS individually (**Follow the guideline**)
  - Assignment will be evaluated for each team



Assignment 1

1. Download the following file.

2. (if the task is given with R) Open R & Create R file.

(if the task is given with Python) Open Google Colab & Create .ipynb file.

3. Change the name of the file to TextMining_Practice#_Team#

ex) TextMining_PR1_Team1.R (Practice 1, Team 1)

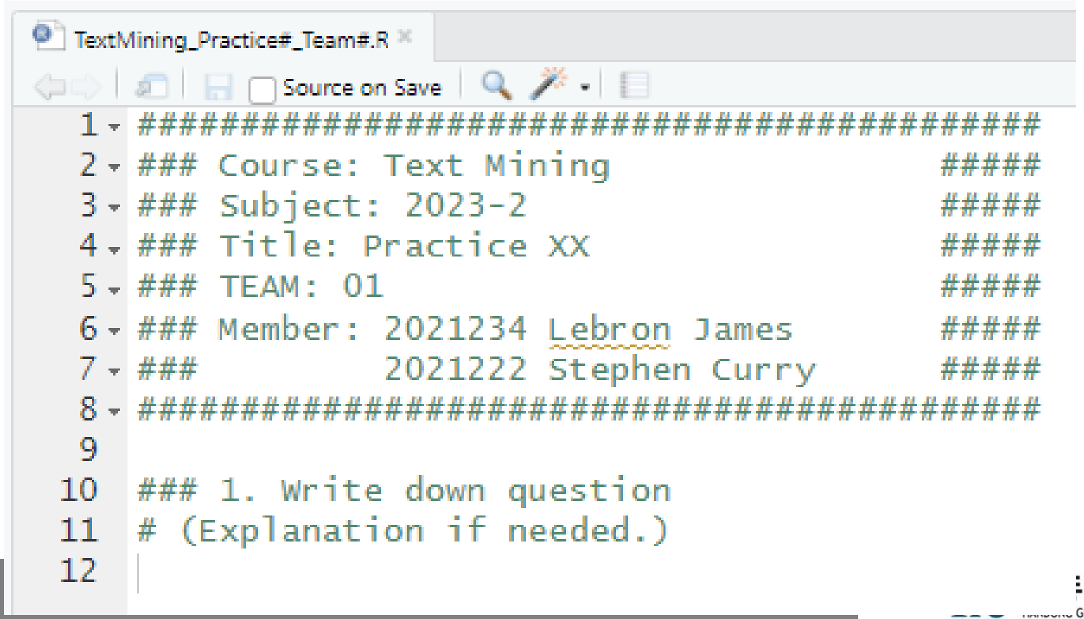
4. Write down answers using code and comment.

6. Upload the completed either .R or .ipynb file to the HDLMS.

TM_Practice1.pdf

TextMining_Practice#_Team#.R

TextMining_Practice#_Team#.ipynb



□ Source on Save

```
############################################
### Course: Text Mining                #####
### Subject: 2023-2                    #####
### Title: Practice XX                 #####
### TEAM: 01                           #####
### Member: 2021234 Lebron James       #####
###         2021222 Stephen Curry      #####
############################################

### 1. Write down question
# (Explanation if needed.)

```

Team Project

- Text mining research
  - Research project using text data
  - Use all the techniques covered during the class
  - Any topics that are related to your major or interest are welcome
  - Proposal presentation: Week 9 – Tuesday (2 slides)
  - Final presentation: Week 15
  - 20 minutes of presentation
- Evaluation
  - Novelty (topic & data)
  - Text pre-processing & analysis
  - Implication
  - Delivery (presentation & communication)

Big Data & Text

Background

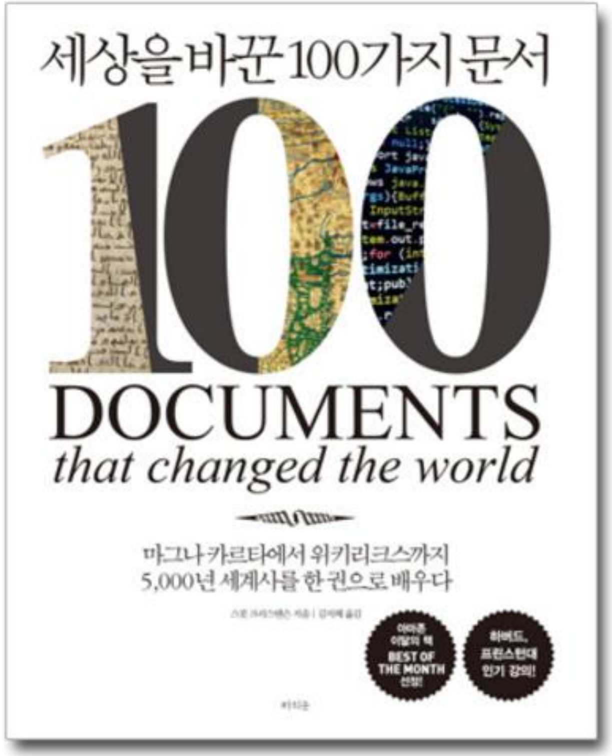
- Digitization
  - Converting data into a digital format
  - “format”
- Digitalization
  - Transforming business process to digital business process
  - “process”

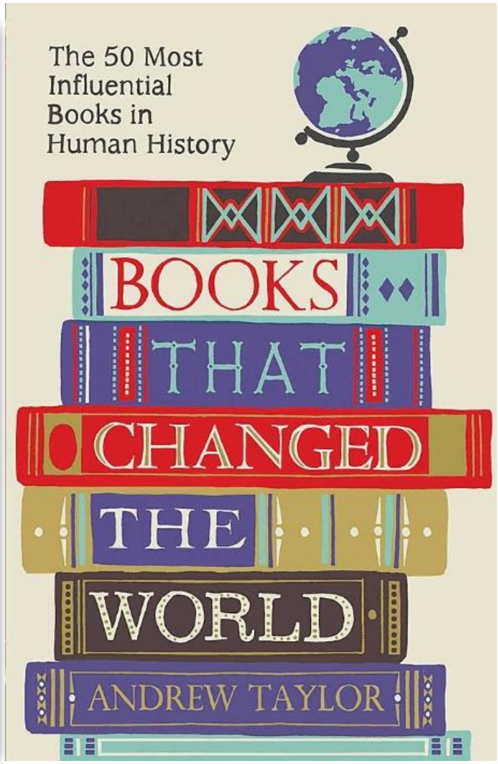
- Data
  - Structured data: data frame, database, etc.
  - Unstructured data: audio, video, text, etc.

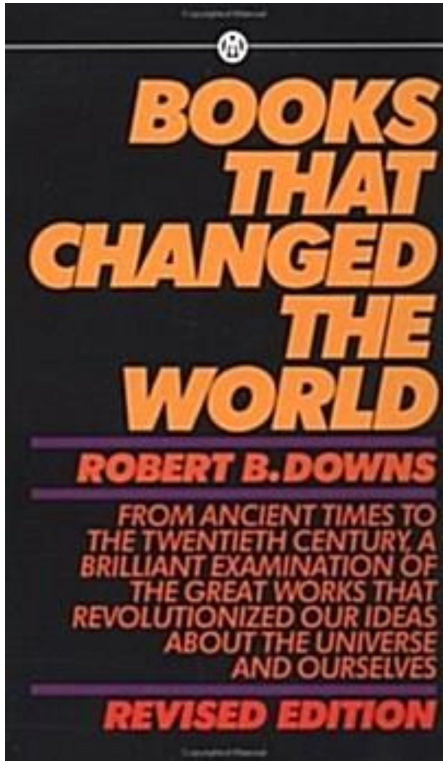
Text

- Text
  - Book
  - Historical records
  - Love letter
  - Text message

  → Anything that is written
- Why do we use text?
  - To remember
  - To understand
  - To think
  - To improve
- Some texts do change the world.






BOOKS THAT CHANGED THE WORLD
ROBERT B. DOWNS
FROM ANCIENT TIMES TO THE TWENTIETH CENTURY, A BRILLIANT EXAMINATION OF THE GREAT WORKS THAT REVOLUTIONIZED OUR IDEAS ABOUT THE UNIVERSE AND OURSELVES
REVISED EDITION

Text as Data

- From a computer's perspective, text is a "digitally formatted symbol"
  - Computer recognizes "text" itself only with the "text"
  - Man uses "context" to interpret or understand the "text"

- Context
  - The text in which a word or passage appears and which affects its meaning; also the words and social setting which surrounds a spoken word or passage [Wikipedia]

Twitter

Twitter is a waste of time 견해

It's about responsibility. I think they (players) are responsible for their actions, responsible for what they said on Twitter. I don't understand it, to be honest with you. I don't know why anybody can be bothered with that kind of stuff. How do you find the time to do that? There are a million things you can do in your life without that. Get yourself down to the library and read a book. Seriously. It is a waste of time.

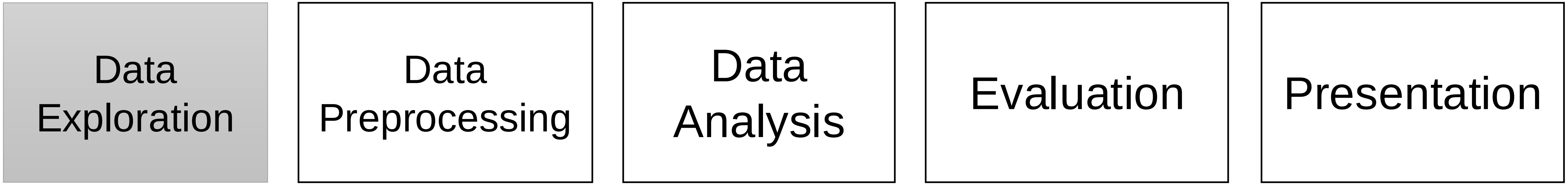
Two Analytic Methods for Text Data

- Qualitative method
  - Read & understand
  - Analyze and write comments
  - Understanding the “context”
  - Most valid approach for understanding the meaning of text
  - Not applicable in a large text data set

- Quantitative method
  - Finding patterns
  - Finding a relationship between words
  - Applicable in a large text data set

Data Analysis Overview

- Typical data analysis steps



Data Exploration

Data Preprocessing

Data Analysis

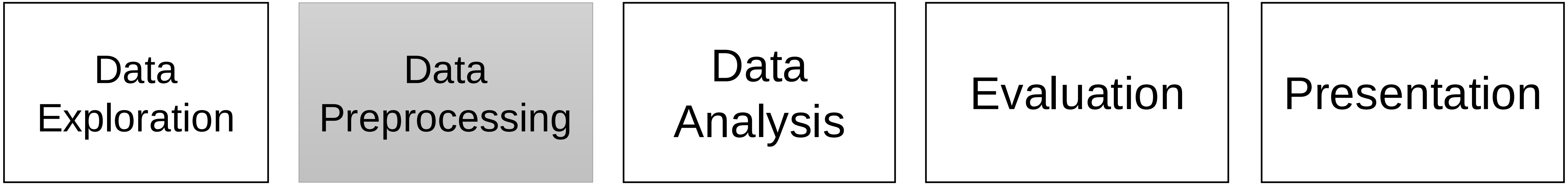
Evaluation

Presentation

- Data Exploration: Understanding data & verifying data
  - number of examples and variables
  - types of variables
  - distribution of each variable, etc.
  - consistency and quality: errors, outliers, missing values

Data Analysis Overview

- Typical data analysis steps



Data Exploration

Data Preprocessing

Data Analysis

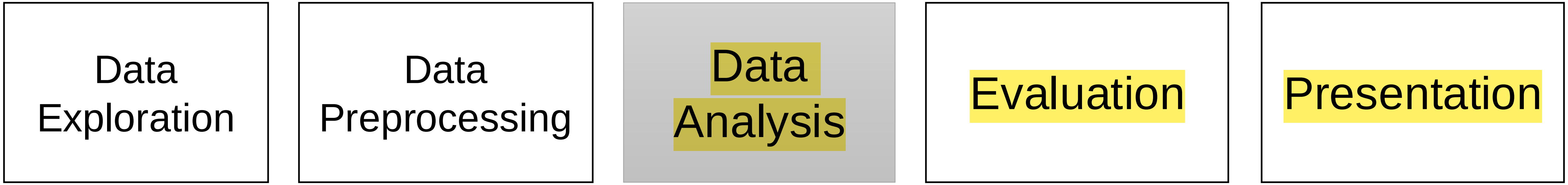
Evaluation

Presentation

- Data Preprocessing: Data cleaning & processing
    - remove outliers
    - handle missing values
    - remove irrelevant variables
    - join data
    - feature extractions

Data Analysis Overview

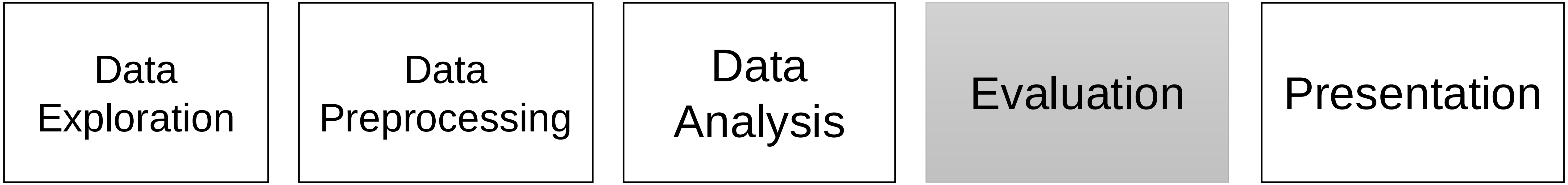
- Typical data analysis steps



- Data Analysis
  - Select an appropriate data analytic method for the project goal
  - Supervised Method: Classification, Regression, prediction, fraud detection, recommendation, …
  - Unsupervised Method: Clustering, Dimensionality reduction, …

Data Analysis Overview

- Typical data analysis steps



Data Exploration

Data Preprocessing

Data Analysis

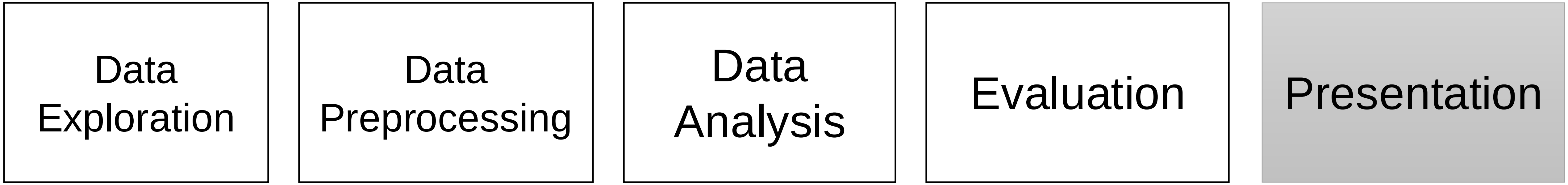
Evaluation

Presentation

- Evaluation
  - Internal review: inside the project team, on a weekly or bi-weekly basis
  - External review: with project client, early stages such as goal setup, data verification

Data Analysis Overview

- Typical data analysis steps



Data Exploration

Data Preprocessing

Data Analysis

Evaluation

Presentation

- Presentation
  - Visualization
  - Delivering the key message

Analyzing Text Data

- Conducting text data analysis
  - Conducting data analysis with text data
  - For text mining, programming and analytic skills focused on text data are needed.
  - In other words, we should be able to understand text as data and do relevant and necessary tasks

  - Data exploration → text exploration
  - Data preprocessing → text preprocessing
  - Data analysis → text analysis
  - Evaluation & presentation → text-centered evaluation and presentation

Required Mind Set

- Avoid efficiency
  - No rule of efficiency works
  - Do all the work with your own effort

- Avoid the illusion of knowing
  - Practice with your own hand
  - Try to explain what you know in your own words

- Redefining the definition of "effort"

- Importance of insight
  - We are learning techniques but "comprehension" matters