

LOGO

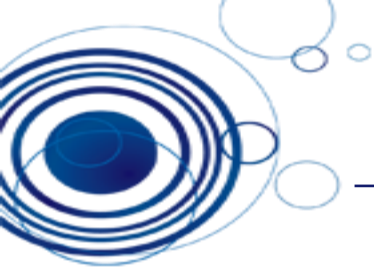
BigData Engineering

14주차: Data Engineering
Workflow Example
(Kaggle)

강의 : 신경섭

11101001110000111110101110010101010011001010011010111101001110000111110101110010101010011001010011010111101001110
11100110000011011101001011101011111010101010010101111100110000011011101001011101011111010101010010101111100110000
110100111001101010101001011011111010100110101001010111010011100110101010100101101111010100110101001010111010011100

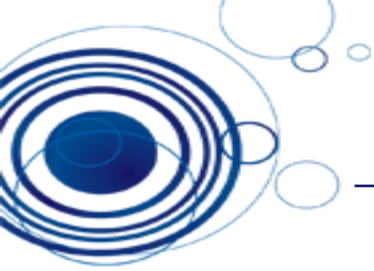




We covered...

- Data management/Visualization by python
 - Numpy, pandas, data acquisition
- Machine learning workflow with data
- EDA (Exploratory Data Analysis)
- Supervised learning
 - k-NN classifier
 - logistic regression based binary classification
 - Support vector machine
 - Decision tree
 - Random Forest
- Model Evaluation

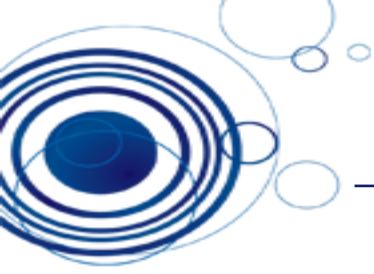




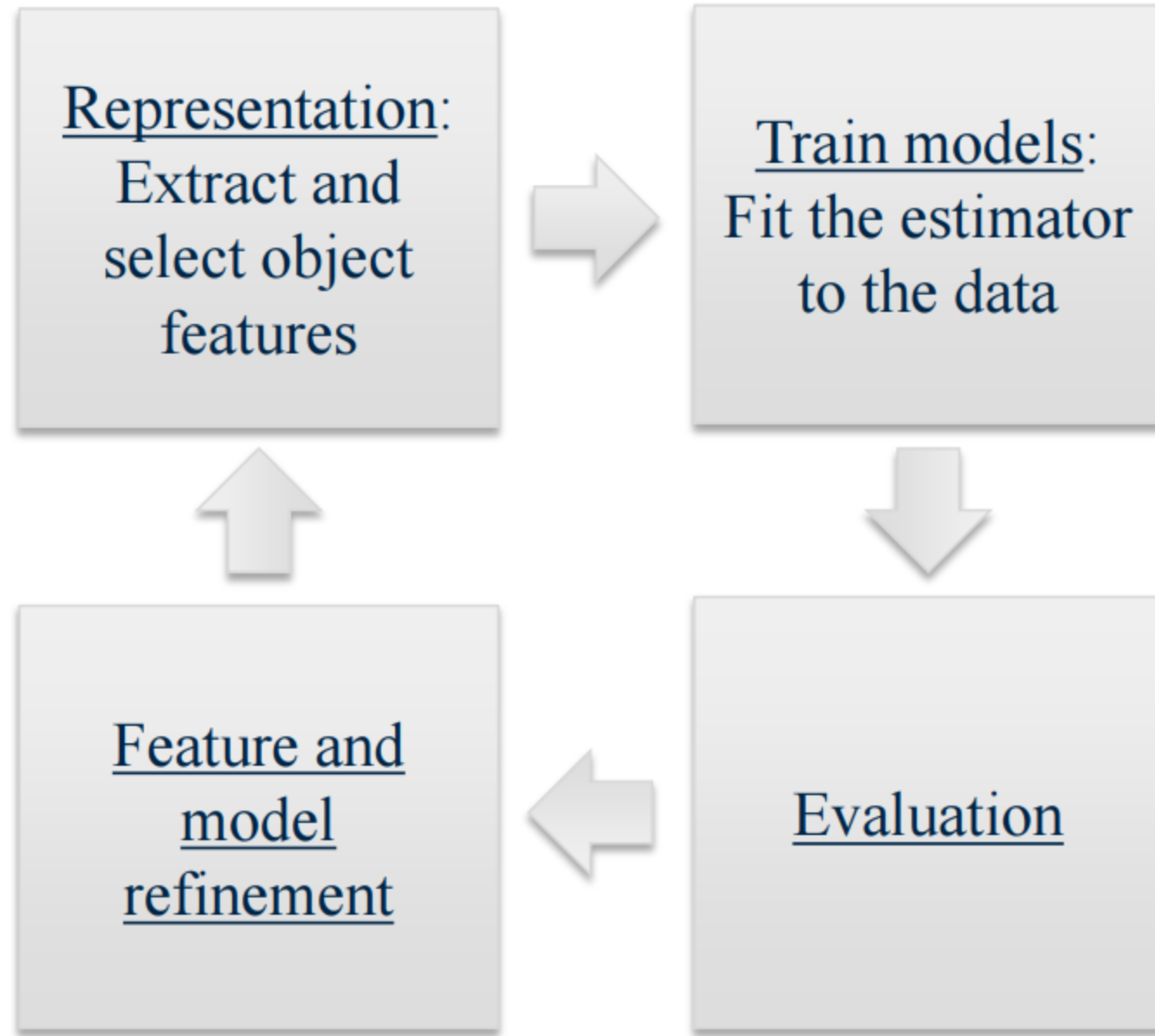
Course Evaluation

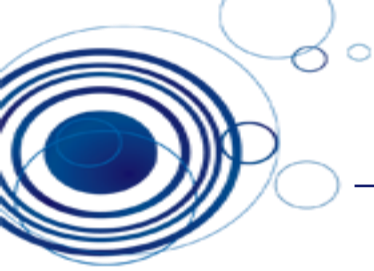
- Attendance : 10%
- Homework : 40%
- Final Project : 50%
 - 다양한 데이터셋을 시각화하여 분석하는 프로젝트
 - 평가방법 : 데이터의 시각화, 데이터의 분석 등
 - Copy는 지양하고 창의적으로 만들어보시기 바랍니다.
 - Jupyter notebook file, readme file, pdf file, dataset file 네가지 압축하여 제출 (파일명: 학번_이름.zip)
 - Readme 파일: 데이터출처, 추가 패키지 설치 필요한 경우 설명 작성
 - pdf 파일: jupyter notebook 단순 변환
 - 데이터셋 참조
 - <https://brunch.co.kr/@data/10>
 - <https://brunch.co.kr/@jowlee/118>





Represent / Train / Evaluate / Refine Cycle

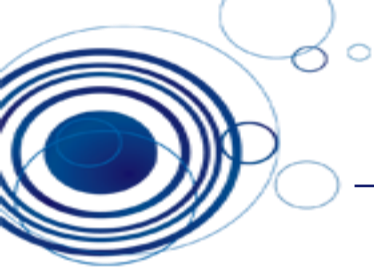




Sci-kit Learn Library reference

- <https://www.datacamp.com/community/blog/scikit-learn-cheat-sheet>

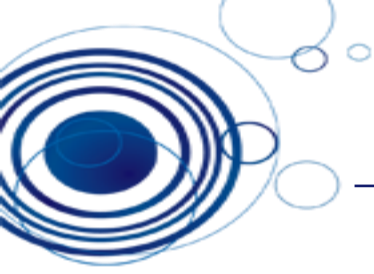




Kaggle competition

- <https://www.kaggle.com/c/titanic/overview>
- Material
 - <https://www.kaggle.com/ash316/eda-to-prediction-dietanic>





Further Study

- Ready to study “data engineering”
- Study with these materials
 - <https://kaggle-kr.tistory.com/32>

