Project Pipeline of Data Science

SIT22009Data Science

Presented by Hyebong Choi

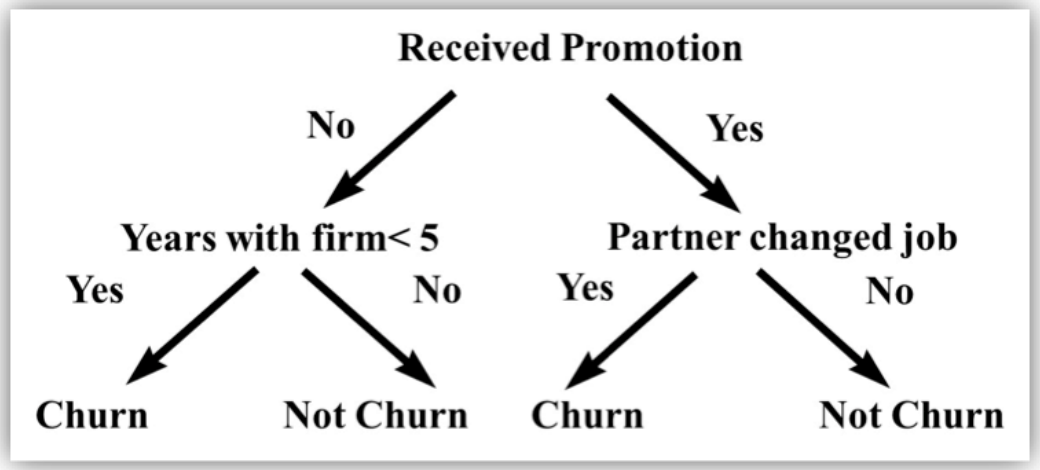
DATA SCIENCE

- **Data Science aims to derive knowledge from big data, efficiently and intelligently**
- **Data Science encompasses the set of activities, tools, and methods that enable data-driven activities in science, business, medicine, and government**
- **Machine Learning(or Data Mining) is again one of the core technologies that enables Data Science**

http://www.oreilly.com/data/free/what-is-data-science.csp

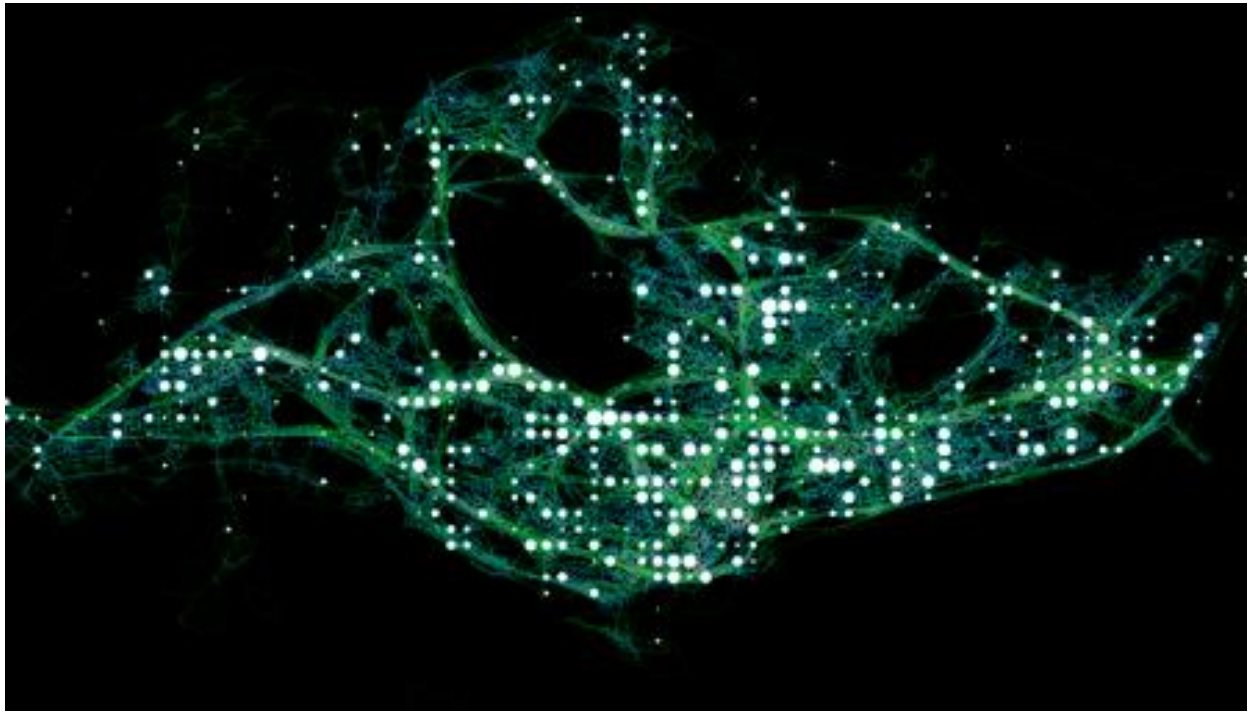
DATA SCIENCE PROJECTS

- **Churn Prediction**
  - Keeping customer royal and not escaping is crucial as much as observing new customers especially for Telco. company, Credit card company
  - Churn: Customers is just about to leave or finish the relationship and subscription the company provides
  - Goal: Reduce Churn Rate by churn prediction
  - escaping customer / incoming customers



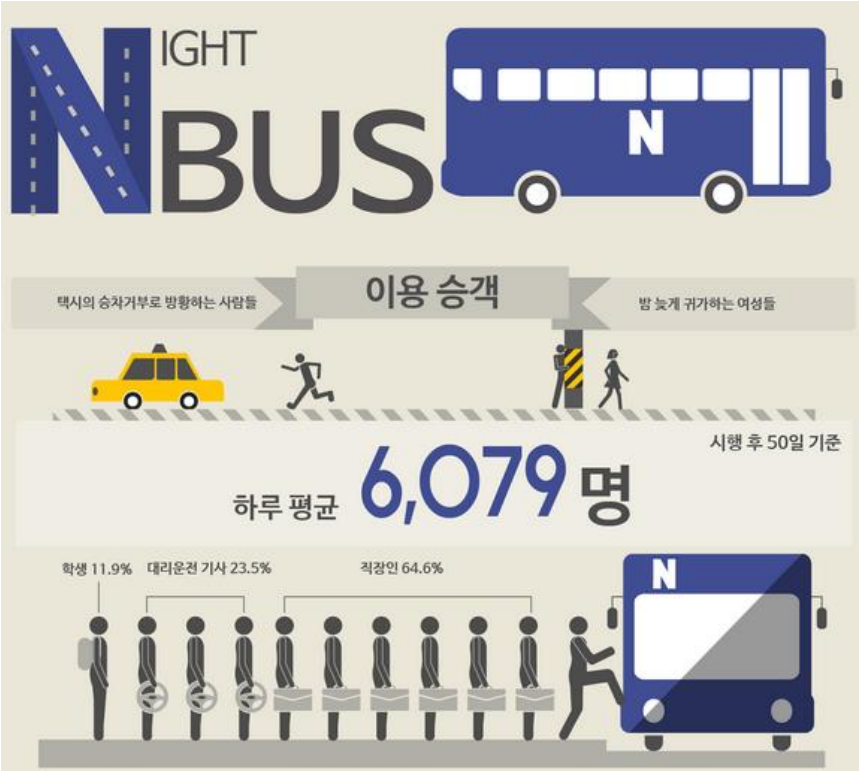
DATA SCIENCE PROJECTS

- **Data Visualization**
  - For authorities, monitoring real-time complex data at a glace would help them make prompt and right decision
  - Singapore's Traffic monitoring system, Subway, bus, taxi, express-way info.



DATA SCIENCE PROJECTS

- **Seoul City-Government, Night-time bus project**
  - Before launching the service, to design the bus route that maximize utilization and citizen's satisfaction given limited budget
  - The government analysed night-time floating population derived from 5 millon taxi-on/off data, and 3 billion phone-calls (KT)

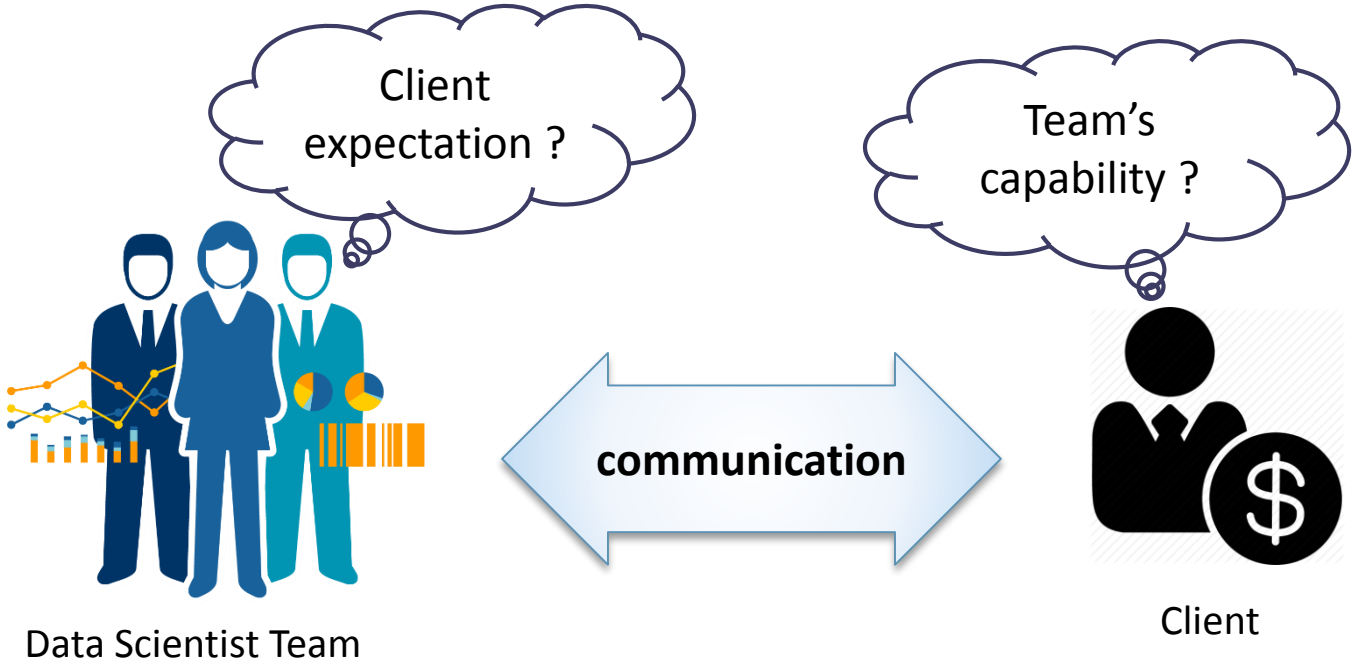


Project Pipeline of Data Science

- Typical Data Science Process
  1. Goal setup
  2. Data Collection
  3. Data Preprocessing (verification, cleaning)
  4. Data Analysis
  5. Evaluation and Revision
  6. Documentation and Presentation of Result
  7. Deployment

Goal Setup

- Setting up a proper goal is the key to the successful project



Goal Setup

- Setting up a proper goal is the key to the successful project
  - understanding of client's domain knowledge and culture
  - setting up proper measures along with the goal (baseline model)

Goal Setup

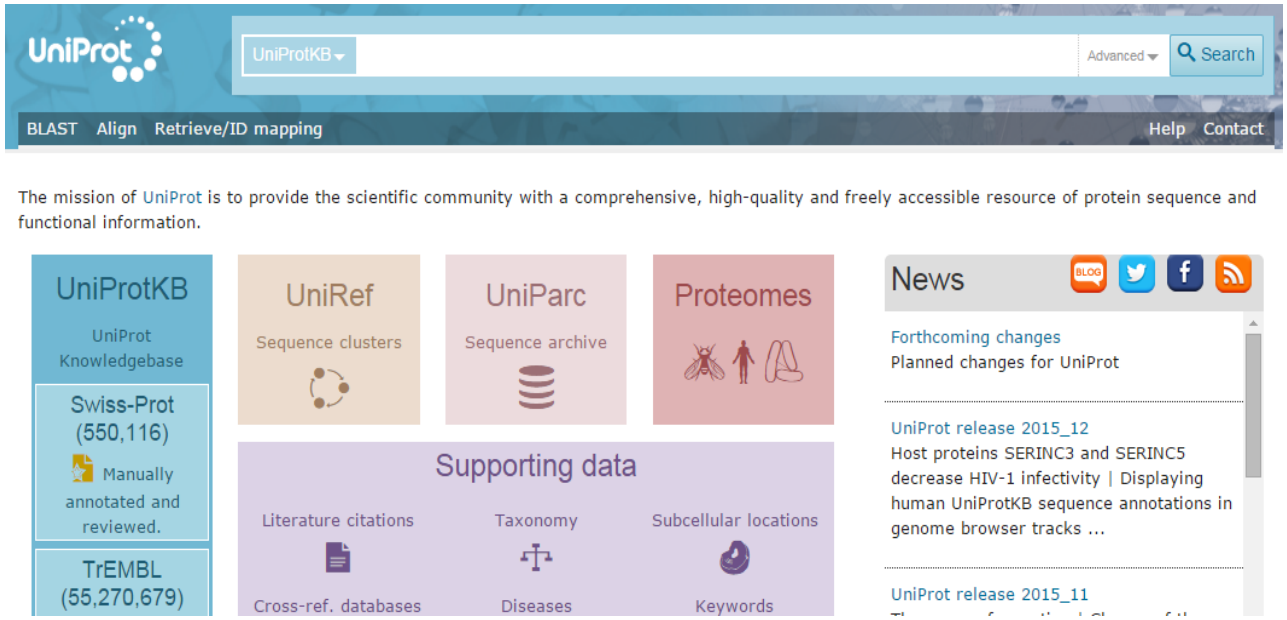
- To Goal should be
  - reasonable
    - considering the capability of DS team and resource (time, men-power, …)
    - considering available data (quality, timeliness, …)
  - clear
    - should be able to state success or failure of the project
    - Something good enough is never clear enough
    - e.g. raise click-through-rate of online advertisement by 10%
    - e.g. improve accuracy of existing detecting system by 5%
  - Measurable
    - How to evaluate the performance of the project
    - Testing scheme

Data Collection

- Client may provide the data
  - More important is data verification
    - close communication is needed
  - May require additional data collection process
- To check license, privacy, and confidentiality issue

Data Collection

- Open Data Project
  - Uniprot
    - provide protein sequence and functional information
    - http://www.uniprot.org/



Data Collection

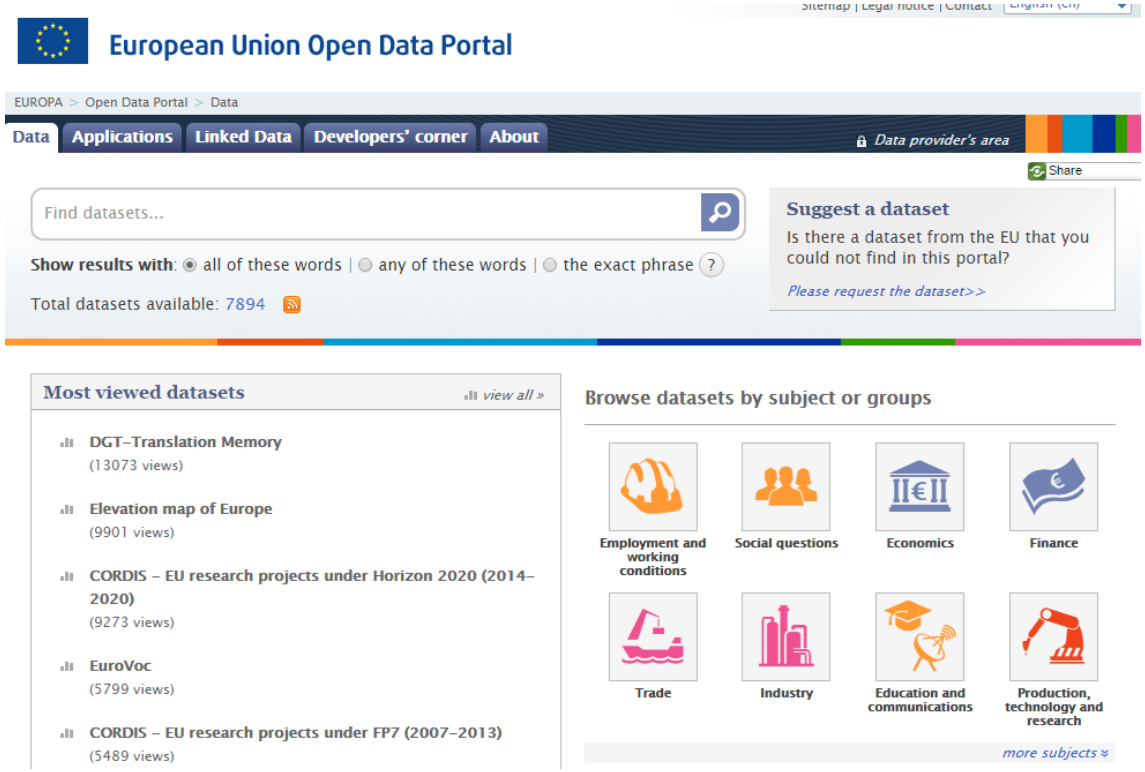
- Open Data Project
  - GDELT Project
    - Daily world-wide events
    - http://gdeltproject.org/

Data Collection

- Open Data Project
  - Public Government Data
    - data.go.kr
    - data.seoul.go.kr

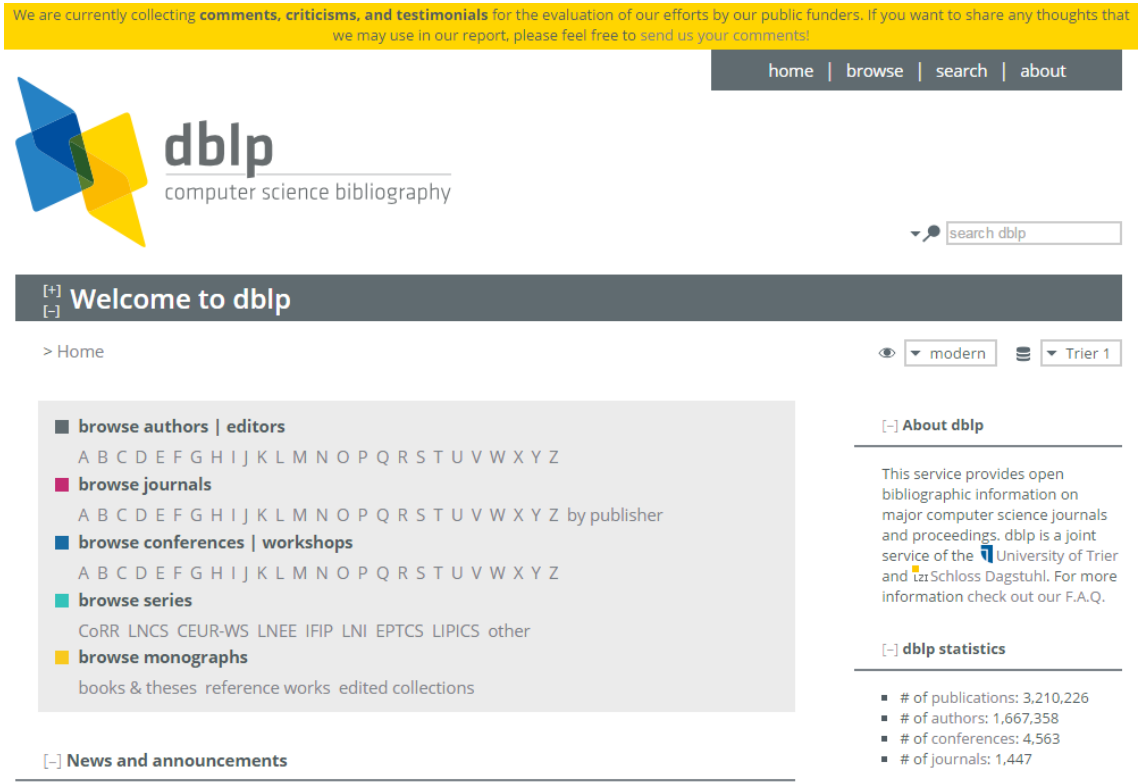
Data Collection

- Open Data Project
  - EU Open Data Portal
    - https://open-data.europa.eu/en/data/



Data Collection

- Open Data Project
  - DBLP
    - Computer science publications, authors, citations, etc.
    - http://dblp.uni-trier.de/

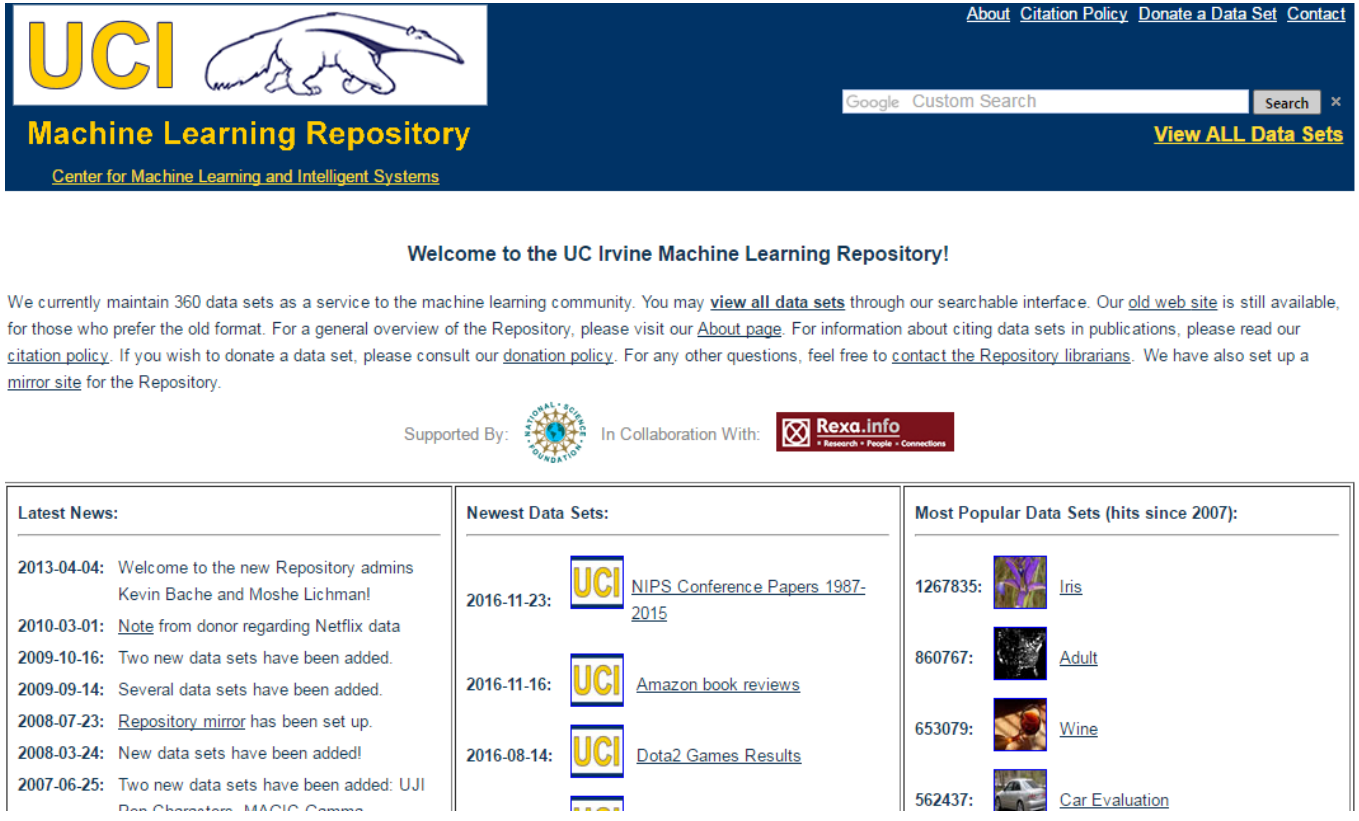


Data Collection

- Open Data Project
  - Weather Underground
    - Real-time and Historical World-wide Weather information
    - http://www.wunderground.com/

Data Collection

- Open Data Repository
  - Welcome to the UC Irvine Machine Learning Repository!
    - 360 data sets as a service to the machine learning community.
    - http://archive.ics.uci.edu/ml/



Data Verification and Preprocessing

- Data Exploration
  - To understand the outline of data
    - number of examples and variables
    - types of variables
    - distribution of each variable, etc.
- Data Verification
  - To check consistency and quality
    - errors, outliers, missing values
  - How data was collected, by handwriting? sensor?
    - measured value, temperature, humidity, …
    - calculated value, heat index, discomfort index, … derived from measured values

Data Verification and Preprocessing

- Data Preprocessing
  - Data Cleaning
    - remove outliers
    - handling missing values
    - remove irrelevant variables
  - Data Processing
    - Joining data
    - Feature extractions

Data Analysis

- Perform the actual Data Analysis
  - Supervised Method
    - Classification, Regression, prediction, fraud detection, recommendation, …
  - Unsupervised Method
    - Clustering, Dimensionality reduction, …
- To choose an appropriate method for the **project goal**

Evaluation and Revision

- Check if the result meets the objective set up in the earlier stage
- Internal review
  - inside project team, weekly or bi-weekly basis
- External review
  - with project client
  - Early stages such as goal setup, data verification, frequently

Documentation and Presentation

- To prove that the project achieve the goal
- Result should be reproducible by client side team
  - Documentation should be clear and sufficient
  - Minimizing additional effort for project maintenance

References

- Practical Data Science with R, by Nina Zumel and John Mount
- R을 이용한 데이터 분석 실무, 서민구, 길벗