vaseline555@github

Lab Session 1

Seok-Ju Hahn

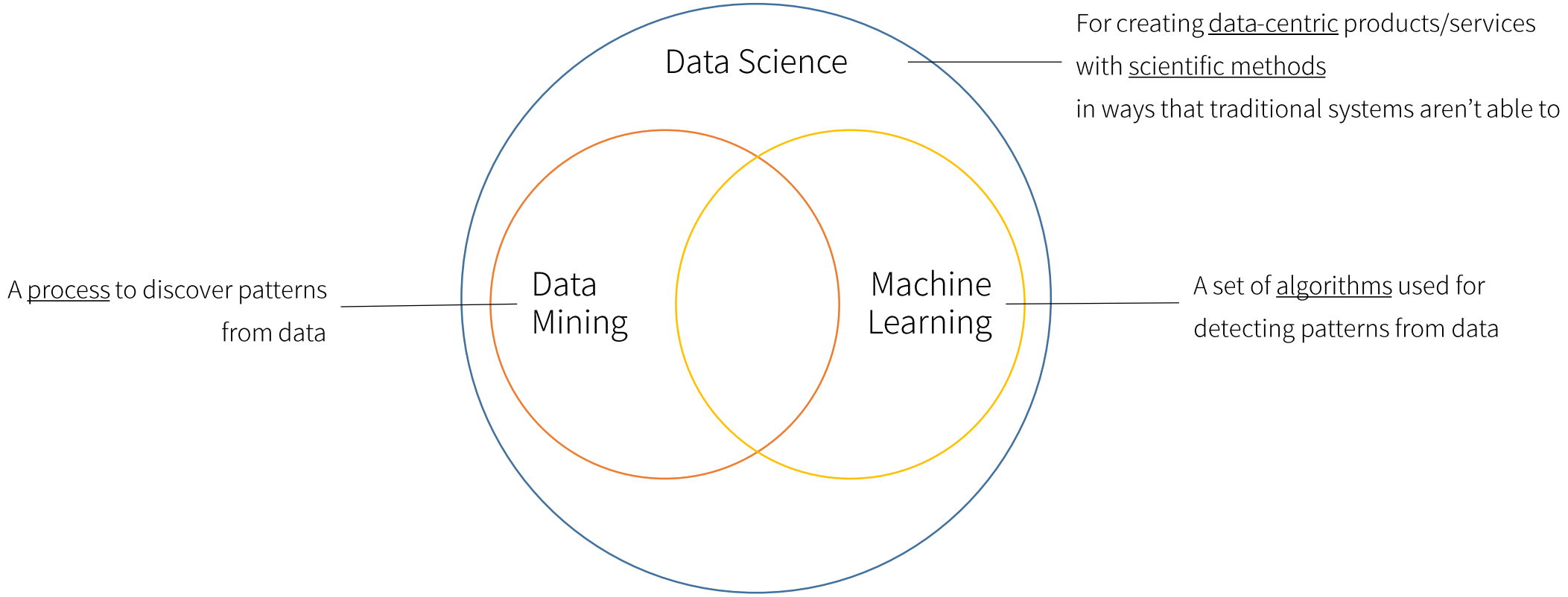
Master course student @ UNIST Data Mining Lab.
(Prof. Junghye Lee)

Contents

- Glossary
- Scientific Method
- Data Mining Workflow
- Example: House Price Prediction

Glossary

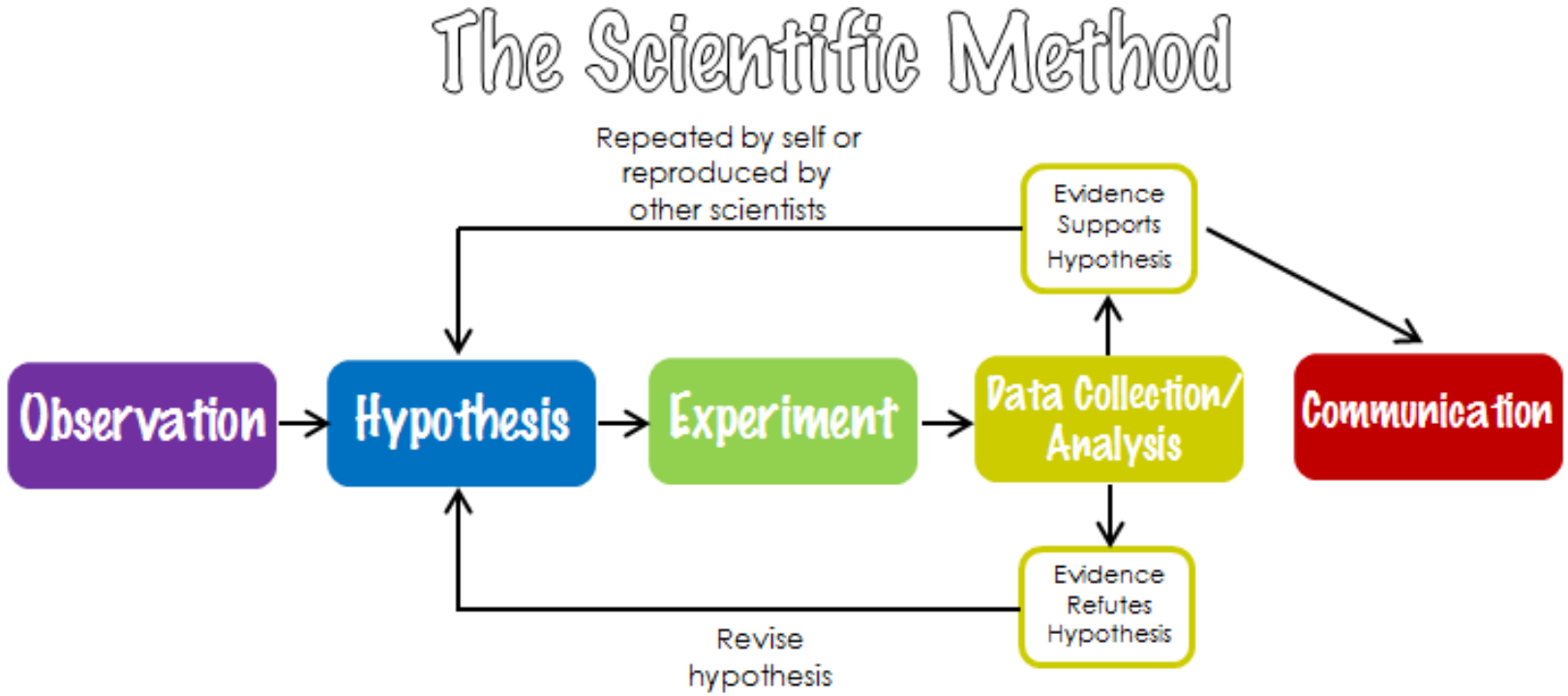
- Data science? Machine Learning? Data Mining?



Data scientists build a **data** products/services in **scientific** way using **machine learning algorithms** on the process called **data mining**.

Scientific Method

- It is data '**science**'; we need scientific method!
- You've already learned in a middle/high school...
- See wiki: https://en.wikipedia.org/wiki/Scientific_method#Process



Scientific Method

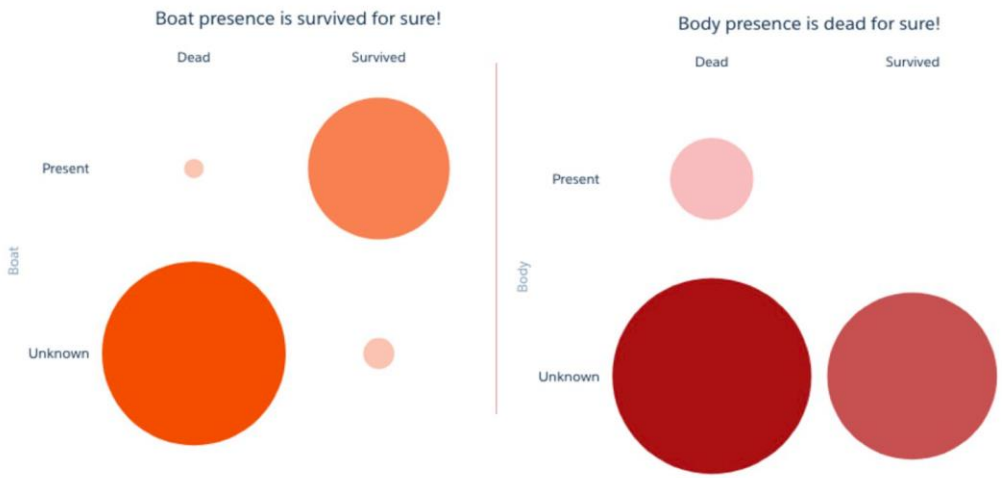
- Observation
  - Formulation of a question, collect and inspect data thoroughly.
  - ex) How can I predict house price?

    At first, I need house price data with its location, some demographics of residents, ...
- Hypothesis
  - Assumption based on knowledge obtained in observation stage
  - You can make a statistical assumption or plausible assumption based on domain-specific knowledge.
  - In data mining process, the hypothesis is represented as the combination of features.
  - ex) I think median income in a certain population group is a key factor to predict house price;

    People with high income may tend to buy an expensive house.
  - CAUTION: hindsight bias (사후 편향): Please be cautious if you have data leakage before modeling!

*Hindsight bias

- Knowing things you shouldn’t know
- Data leakage (or label leakage)
  - The accidental presence of information in the training data.
  - Model relies on information not available at scoring time.
  - In the original data of Titanic survival prediction, there exist ‘boat’ and ‘body’ features. (http://campus.lakeforest.edu/frank/FILES/MLFfiles/Bio150/Titanic/TitanicMETA.pdf)



Scientific Method

- Experiment
  - Determination of logical consequences of the hypothesis
    & investigation of whether the real world behaves as assumed by the hypothesis
  - ex) Train a machine learning algorithm to check whether the median income
    raises prediction accuracy of house price.
- Analysis
  - Determination of the next action based on the result of the experiment
  - Statistical analysis is frequently used for checking the strength of evidence
  - Reinforce/discard hypothesis to answer the question you made in the observation stage
  - ex) Test a machine learning algorithm with unseen data.

Data Mining Workflow

- Observation & Hypothesis (1: **Data Preparation**)
  - Collecting & Storing data (← nowadays a main role of a *data engineer*)
  - Exploratory data analysis (EDA)
  - Cleansing data (munging, imputation, removing duplicated values, feature scaling...)

Data Mining Workflow

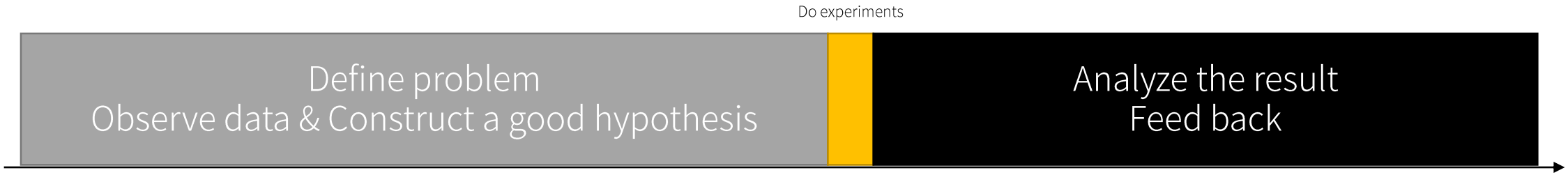
- Observation & Hypothesis (1: **Data Preparation**)
  - Collecting & Storing data (← nowadays a main role of a *data engineer*)
  - Exploratory data analysis (EDA)
  - Cleansing data (munging, imputation, removing duplicated values, feature scaling...)
- Experiment (2: **Modeling**)
  - Selection of appropriate machine learning algorithms
  - Training & evaluating & tuning models

Data Mining Workflow

- Observation & Hypothesis (1: **Data Preparation**)
  - Collecting & Storing data (← nowadays a main role of a *data engineer*)
  - Exploratory data analysis (EDA)
  - Cleansing data (munging, imputation, removing duplicated values, feature scaling...)
- Experiment (2: **Modeling**)
  - Selection of appropriate machine learning algorithms
  - Training & evaluating & tuning models
- Analysis (3: **Evaluation & Feedback**)
  - Testing models
  - Delivering the result to stakeholders
  - Deploying & monitoring services/systems or releasing products (← nowadays a main role of a *data engineer*)

Data Mining Workflow

- Overall summary



- It is **NOT** meaning that you don't have to know algorithms like you've learned in the class
- In fact, well-made algorithms are already ready for you.
- You SHOULD think about an appropriate algorithms for your problem.
  - In other words, you should know many algorithms as far as you can, so that can tackle various real-world problems!
  - And then, you may start to process your data in the form suitable for the model you select.
- If you know many, but still not being satisfied with them, then you CAN make your own algorithm! (→ ☎ Data Mining Lab. @ UNIST)

Observation: House Price Prediction

- Problem definition
  - Suppose we are now in 1990s, and you are the CEO of a construction company in California.
  - You are going to build an apartment complex somewhere in California, but where?
  - Where you can redeem a capital you invested.
  - If you build luxurious penthouses in Eonyang-eup, can you retrieve your investment...?
  - Thus, the important thing is to predict the future housing price accurately.
- Checking existing method
  - Existing method is to ask for consultation to experts in real-estate fields. (Delphi method)
  - But it is time-consuming, inaccurate, and expensive.
  - Then, it seems reasonable to use data mining approach in prediction of house price!
- Collect related data
  - Thankfully, suppose there already exists a good data; California Housing Prices dataset.
  - http://lib.stat.cmu.edu/datasets/houses.zip

Observation: House Price Prediction

- Problem definition
  - Suppose we are now in 1990s, and you are the CEO of a construction company in California.
  - You are going to build an apartment complex somewhere in California, but where?
  - Where you can redeem a capital you invested.
  - If you build luxurious penthouses in Eonyang-eup, can you retrieve your investment...?
  - Thus, the important thing is to predict the future housing price accurately.

Is it supervised learning? or unsupervised learning?

Observation: House Price Prediction

- Problem definition
  - Suppose we are now in 1990s, and you are the CEO of a construction company in California.
  - You are going to build an apartment complex somewhere in California, but where?
  - Where you can redeem a capital you invested.
  - If you build luxurious penthouses in Eonyang-eup, can you retrieve your investment...?
  - Thus, the important thing is to predict the future housing price accurately.

Is it supervised learning? or unsupervised learning?

Then, is it classification? regression? clustering?

Observation: House Price Prediction

- Problem definition
  - Suppose we are now in 1990s, and you are the CEO of a construction company in California.
  - You are going to build an apartment complex somewhere in California, but where?
  - Where you can redeem a capital you invested.
  - If you build luxurious penthouses in Eonyang-eup, can you retrieve your investment...?
  - Thus, the important thing is to predict the future housing price accurately.

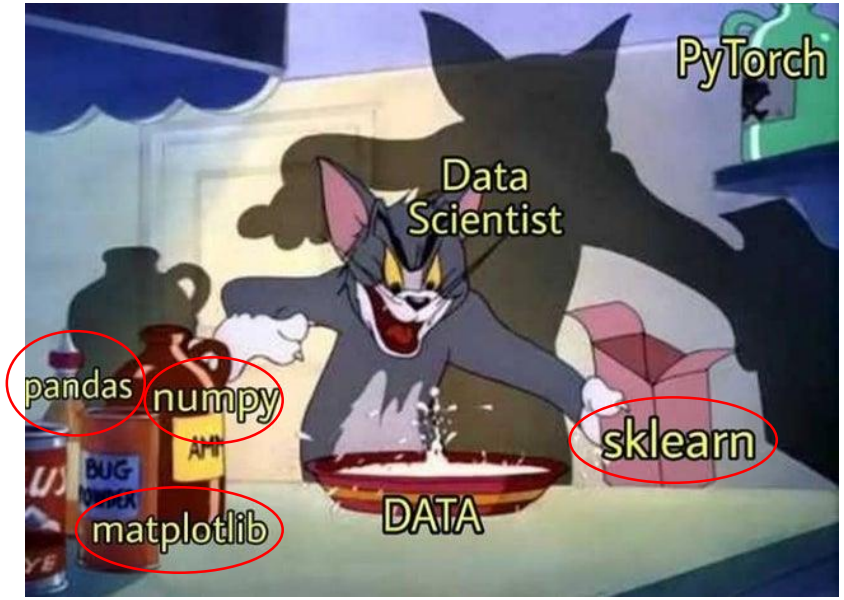
Is it supervised learning? or unsupervised learning?

Then, is it classification? regression? clustering?

What if you just want to know the degree of expensiveness of house price?
(e.g. cheap / so-so / expensive)

Hypothesis: House Price Prediction

- Hypothesis
  - In collected data, suppose we already completed filtering out relevant predictors according to our hypothesis.
- Munging data
  - Load data
  - Handle columns and rows of data
  - Split training and test data
- Exploratory data analysis (EDA)
  - Data visualization
- Pre-process data
  - Data cleansing (handling missing values, handling categorical features, feature scaling)
- Let’s code!
  - Remember popular libraries
  - NumPy, Pandas, MatPlotLib, Scikit-Learn



Experiment: House Price Prediction

- Modeling on training data
  - Select algorithms
  - Fit algorithms on data
  - Evaluation of the trained model using cross-validation (internal validation)
  - Tune hyper-parameters of the model
- Test a generalization performance of the model
  - Check performance of the model using unseen test data

Analysis: House Price Prediction

- Analyze the result
  - Did predictors you provide as a means of verifying your hypothesis work well?
  - If it is, tidy up results, make a presentation file, and deliver the result to stakeholders!
  - If it is not, collect more data or change your hypothesis and problem definition.
  - In high probability, a machine learning model you choose has no fault.
    (Yes, it is developed by genius!)

Questions?

Thank You