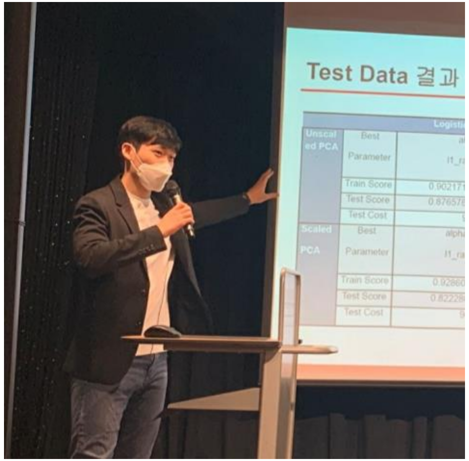
KUBIG
Data Science and Machine Learning

Week 1. Data Science Lifecycle

Introduction and Study Overview



분반장 **조규선**

통계학과 18학번
KUBIG 10기

분반장 **조민제**

경제학과 18학번
KUBIG 11기

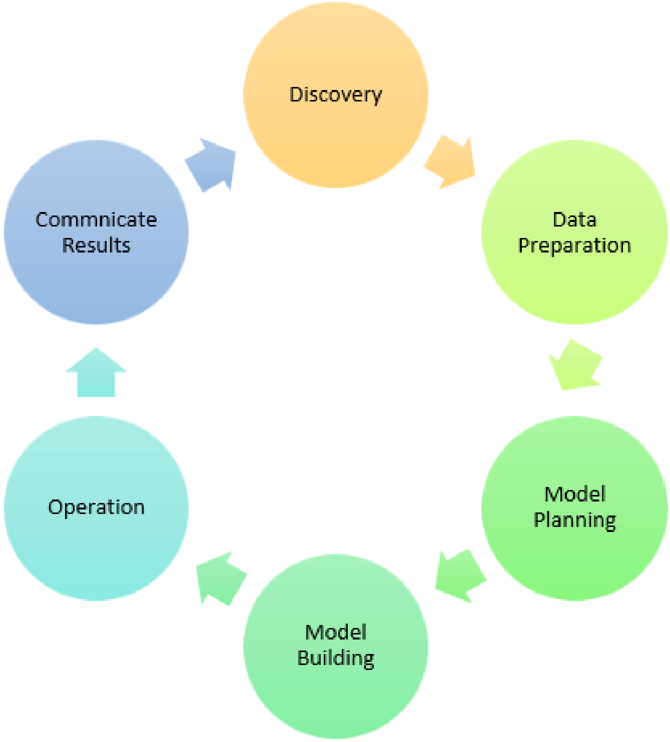
분반장 **조규선**

- 통계학과 18학번
- KUBIG 10기

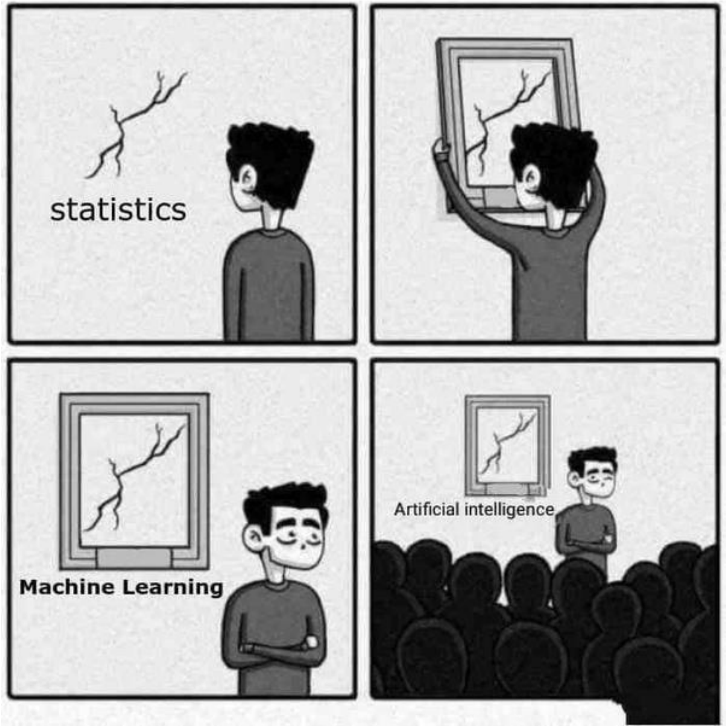
- KUBIG 2020-2학기, 2021-2학기 학회장
- 고려대학교 CVLAB 학부연구생 근무중
- 21-1학기 컴퓨터 비전 분반 진행
- 주요 관심사: 자율주행자동차, Image-to-Image Translation, 3D Depth Estimation

What is this study about?

Fundamentals of Data Science



Statistical Machine Learning

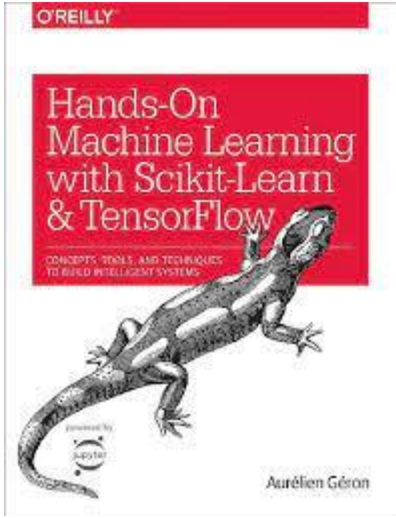


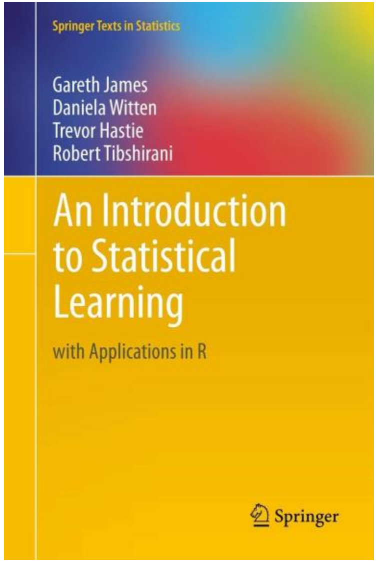
Get the most out of this sessions

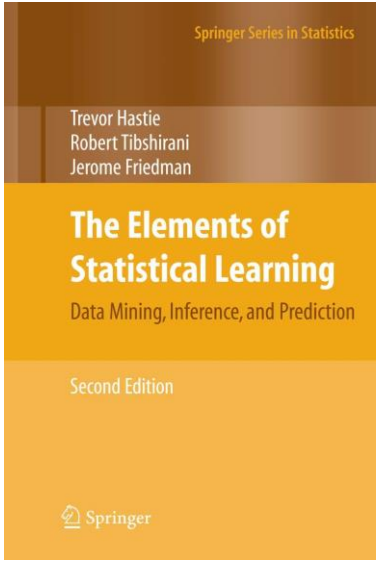
1. Practical Use as well as Understanding the Fundamentals
2. Understanding the Primary Link
3. Ability to Self-Study
4. Teamwork

What will we do?

Books


O'REILLY
Hands-On Machine Learning with Scikit-Learn & TensorFlow
CONCEPTS, TOOLS, AND TECHNIQUES TO BUILD INTELLIGENT SYSTEMS
Aurélien Géron


Springer Texts in Statistics
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R
Springer


Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Second Edition
Springer

Classes (Each Week)

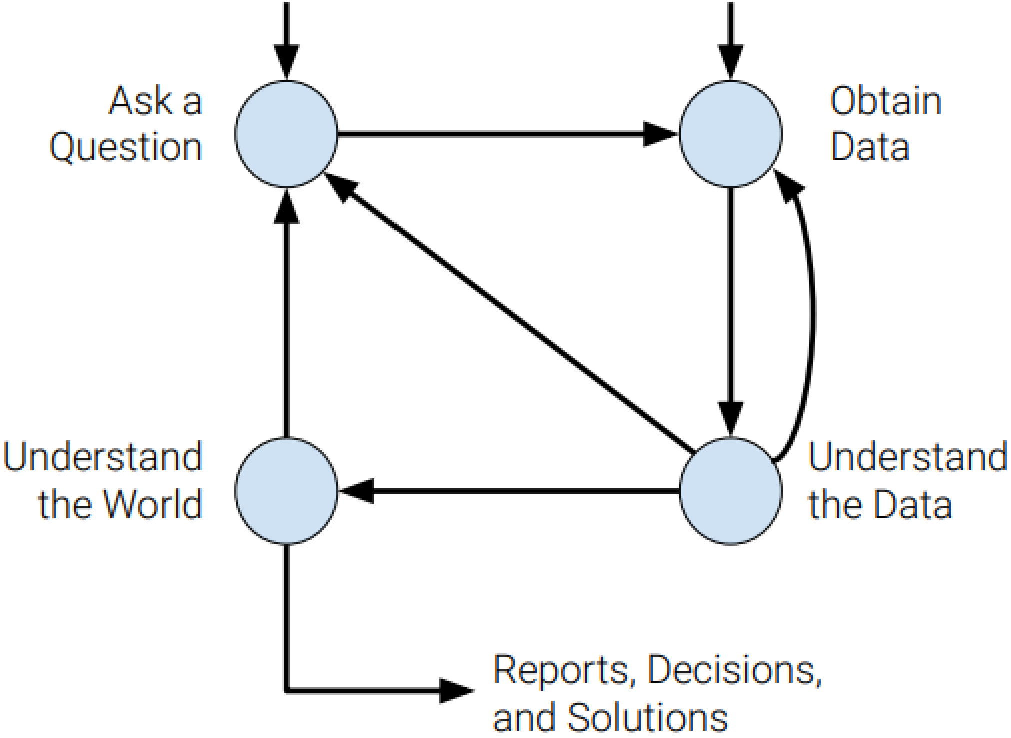
~	Monday	~	Thursday
(Preview) Read Books and Resources	In-Depth Analysis and Explanation Code Review	Group Projects and Homework	Group Project Presentation

Syllabus

Week 1	Data Science Lifecycle
Week 2	Regression and Classification
Week 3	Model Fitting and Validation
Week 4	Regularization
Week 5	Decision Trees and SVM
Week 6	Ensemble Methods
Week 7	Neural Networks

Idea 1. Data Science Lifecycle

Data Science Lifecycle



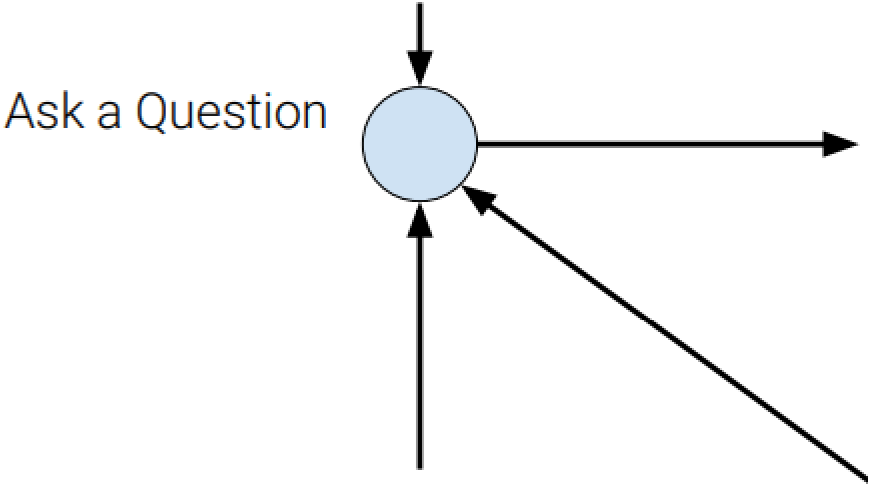
The data science lifecycle is a **high-level description** of the data science workflow.

Note the two distinct entry points!

Data Science Lifecycle (1)

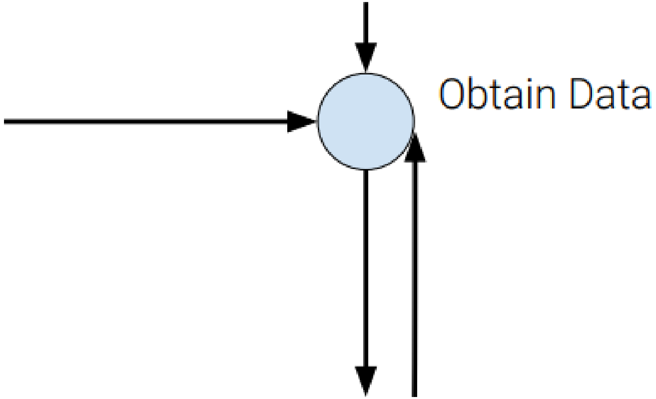
1. Question/Problem Formulation

- What do we want to know?
- What problems are we trying to solve?
- What are the hypotheses we want to test?
- What are our metrics for success?



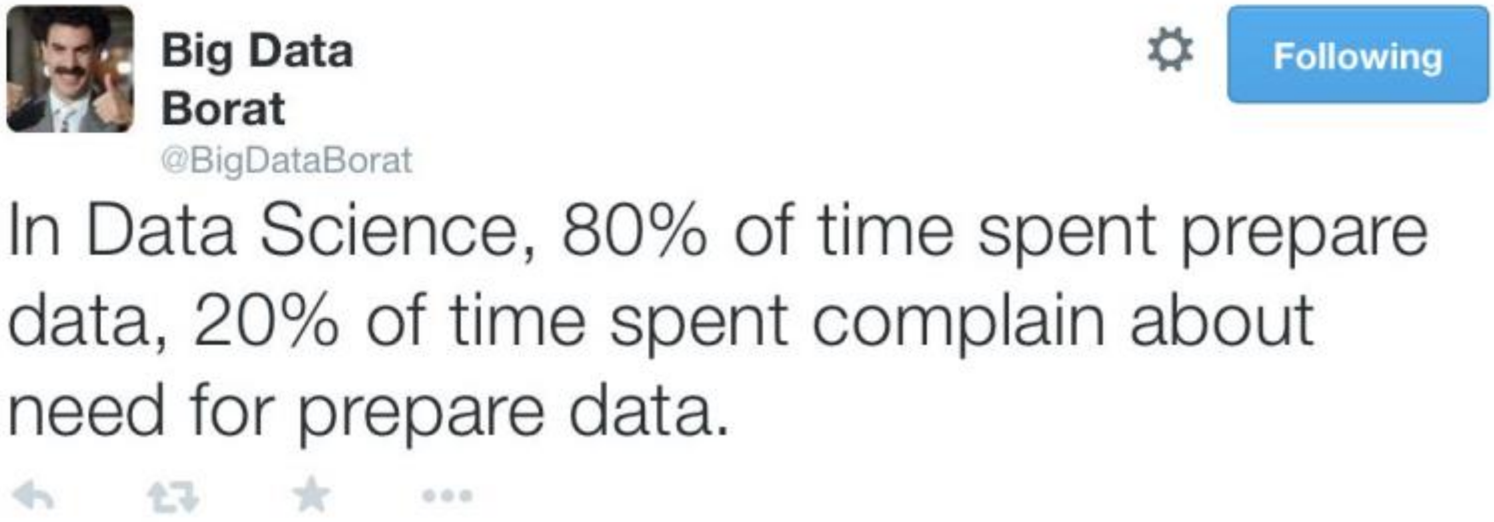
Data Science Lifecycle (2)

2. Data Acquisition and Cleaning



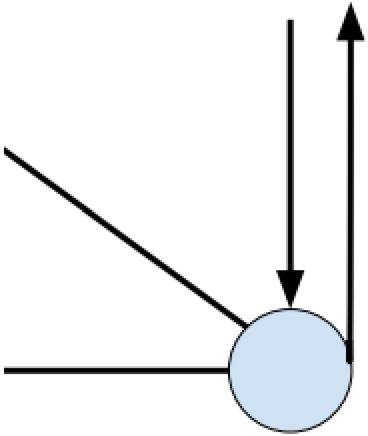
- What data do we have and what data do we need?
- How will we sample more data?
- Is our data representative of the population we want to study?

Data Science Lifecycle (2)



Data Science Lifecycle (3)

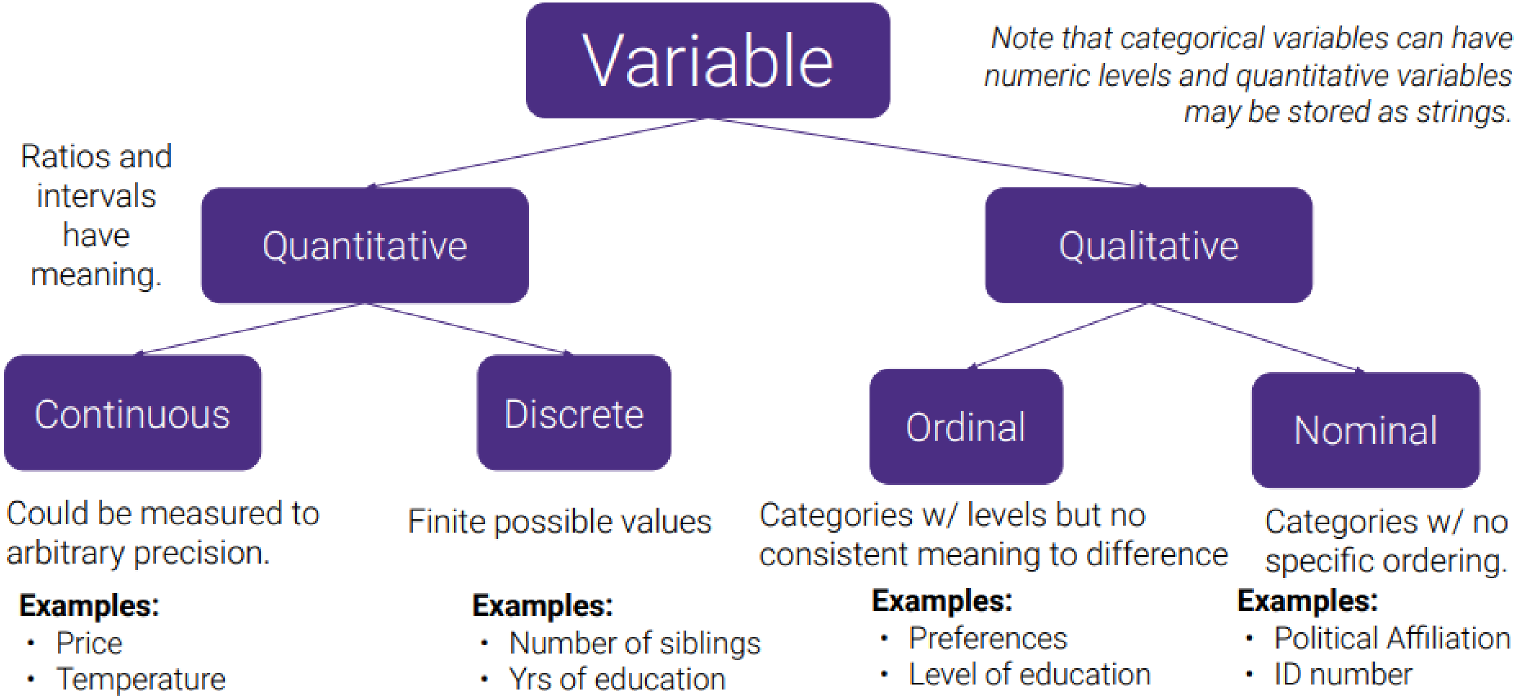
3. Exploratory Data Analysis & Visualization



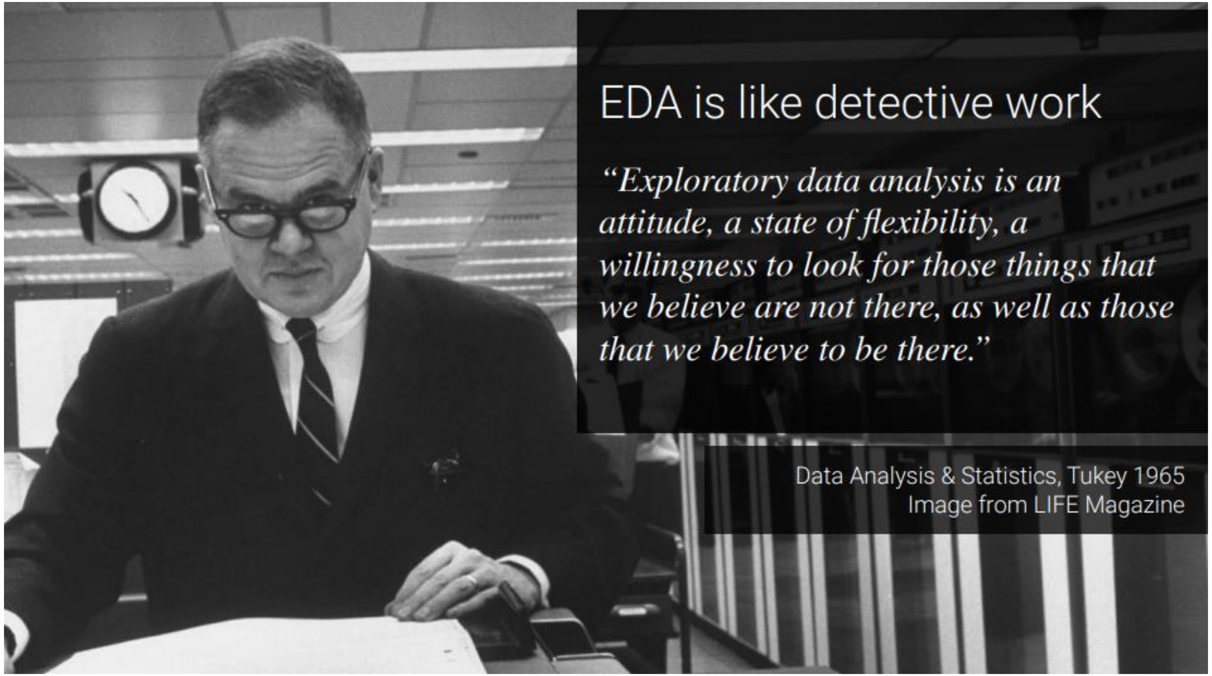
Understand the Data

- How is our data organized and what does it contain?
- Do we already have relevant data?
- What are the biases, anomalies, or other issues with the data?
- How do we transform the data to enable effective analysis?

Data Science Lifecycle (3)



Data Science Lifecycle (3)



Data Science Lifecycle (3)

Key Data Properties to Consider in EDA

- **Structure --** *the "shape" of a data file*
- **Granularity --** *how fine/coarse is each datum*
- **Scope --** *how (in)complete is the data*
- **Temporality --** *how is the data situated in time*
- **Faithfulness --** *how well does the data capture "reality"*

Data Science Lifecycle (3)

- Data Visualization

Data Science Lifecycle (3)

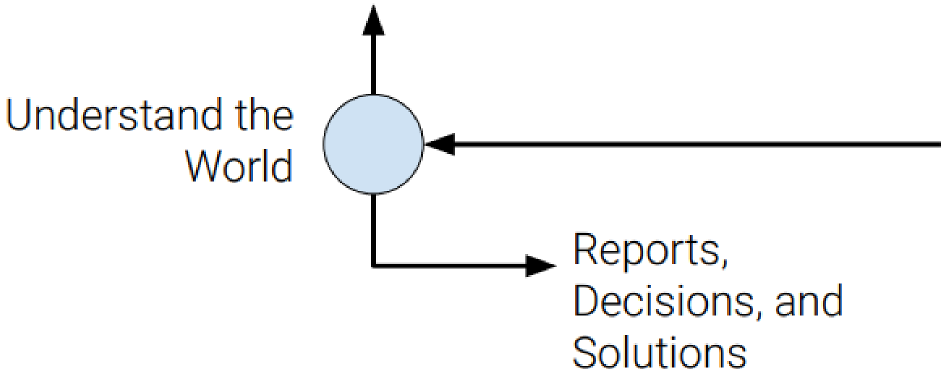
Preprocessing Methods

1. 데이터 정제
2. 데이터 축소
3. 데이터 변환

Data Science Lifecycle (4)

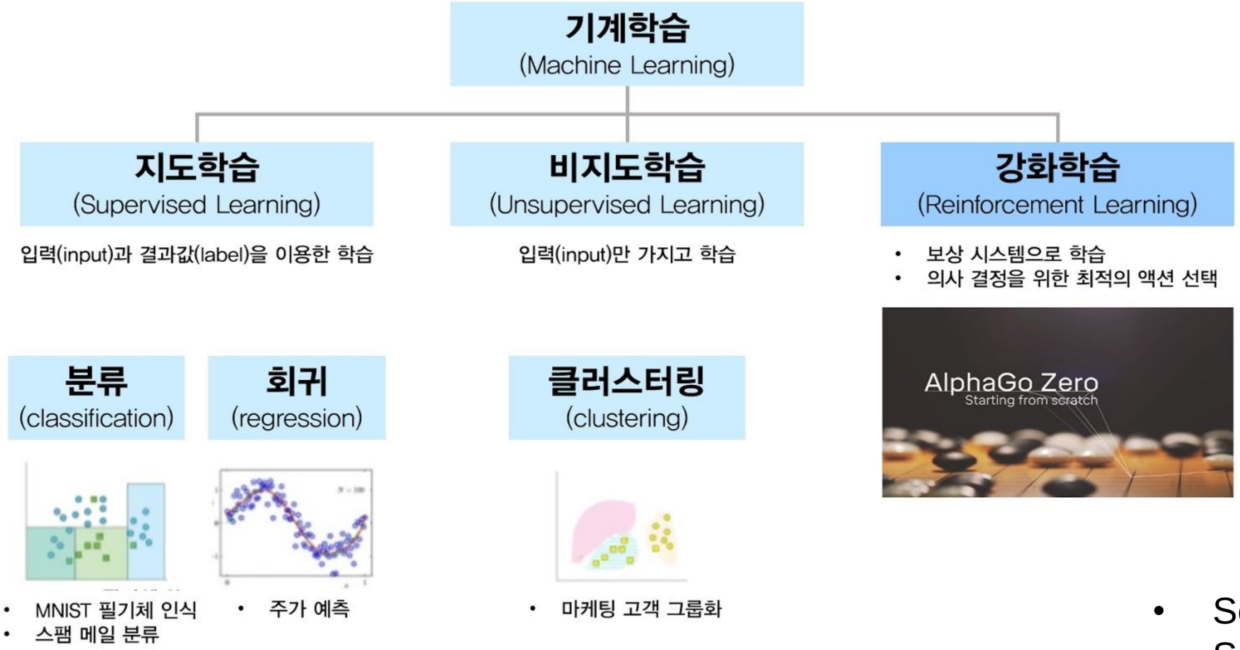
4. Prediction and Inference

- What does the data say about the world?
- Does it answer our questions or accurately solve the problem?
- How robust are our conclusions and can we trust the predictions?



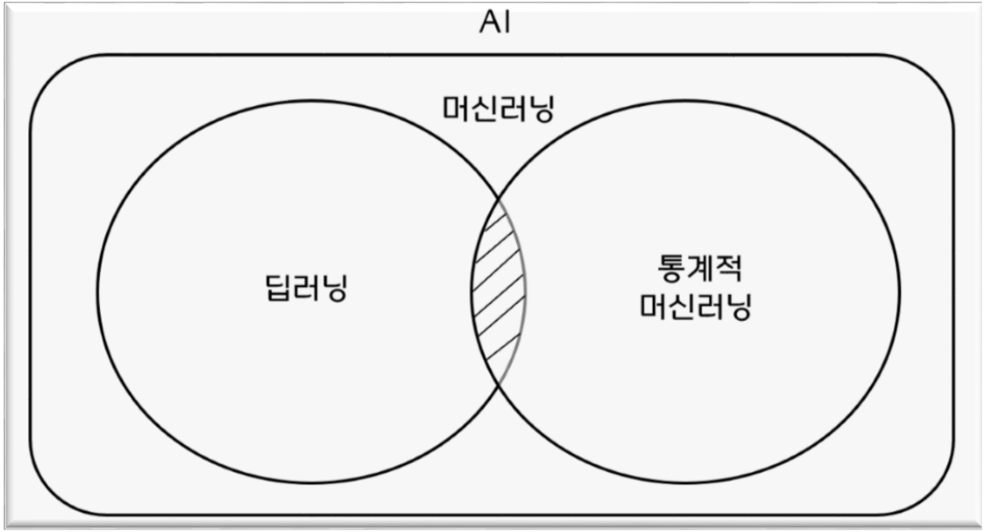
Data Science Lifecycle (4)

Machine Learning



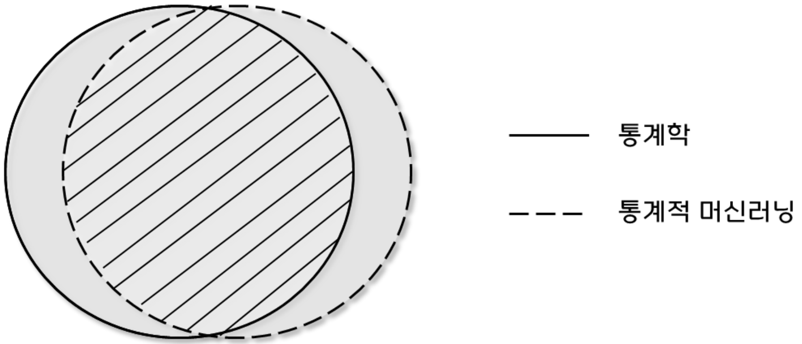
- Semi-Supervised Learning
- Self-Supervised Learning

Data Science Lifecycle (4)



Data Science Lifecycle (4)

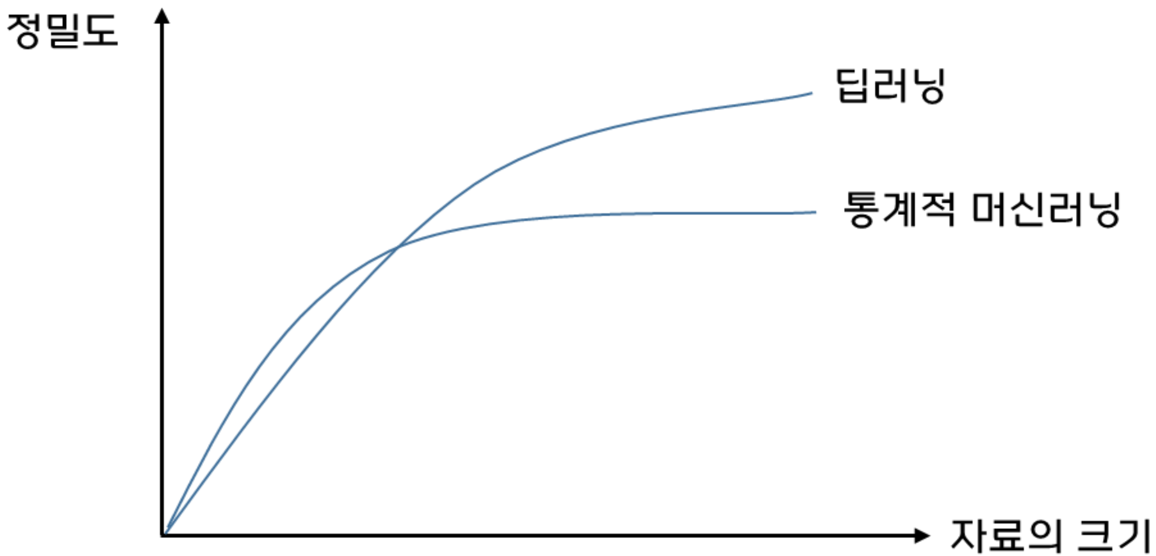
전통적인 통계학	통계적 머신러닝
• 규칙의 통계적 추론에 중점 (전문적인 통계적, 수학적 지식) • 자료의 특성(다변량, 시계열, 범주형 등)에 따라 분석.	• 규칙의 일반화에 중점 • 목적변수의 관측여부에 따라 지도학습, 비지도학습으로 분석



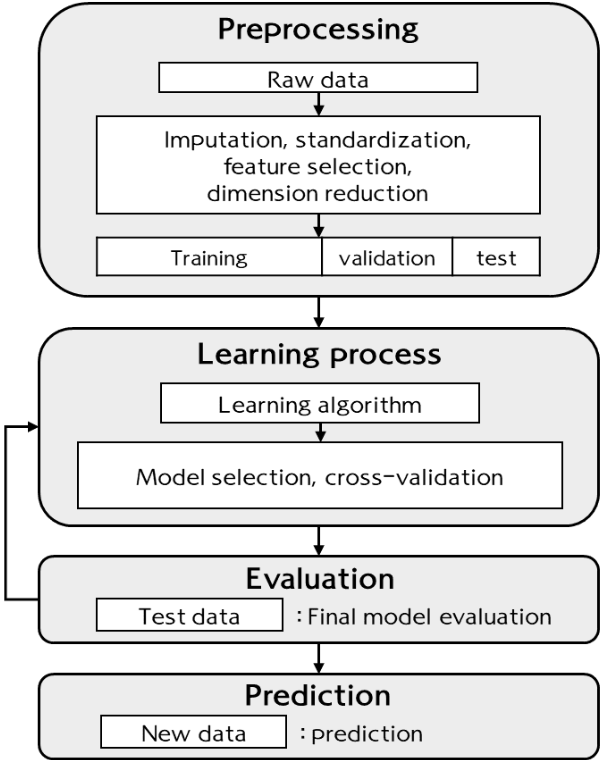
Data S

구분	통계적 머신러닝	딥러닝
데이터 크기	중/소 크기	빅데이터
분석자료 형태	2차원 텐서	2차원 텐서이상
강점을 갖는 자료	정형화된 자료	비정형자료
특성변수	특성변수를 만들어야 함	특성변수가 만들어짐
특성변수의 정규화 및 표준화	선택	필요
모형	매우 많음	기본적으로 3 개의 모형
최적화	일반적으로 전체 데이터 사용	배치데이터
해석여부	해석이 쉬움 (단, SVM과 boosting 제외)	어렵거나 불가능
하드웨어	중급	고성능(GPU 요구)
실행요구시간	최대 시간 단위	최대 주단위 시간

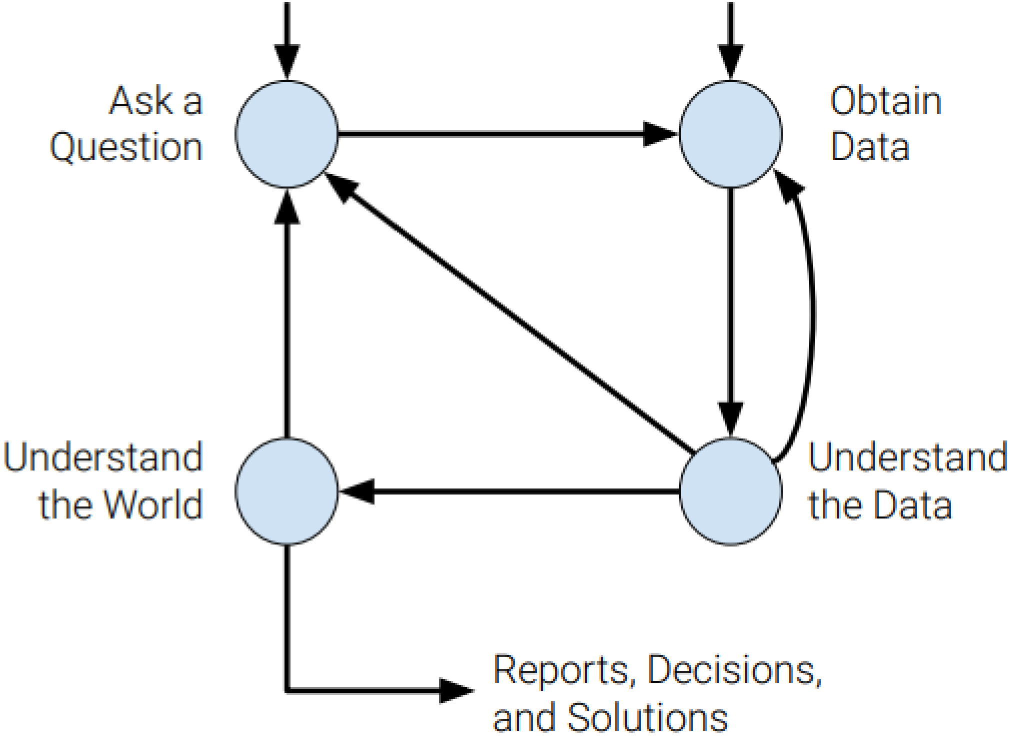
Data Science Lifecycle (4)



Data Science Lifecycle (4)



Data Science Lifecycle

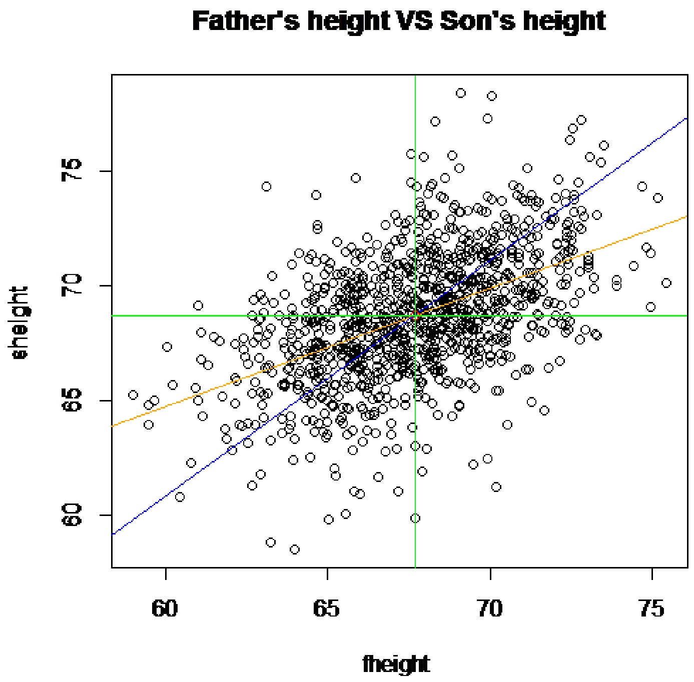


The data science lifecycle is a **high-level description** of the data science workflow.

Note the two distinct entry points!

Idea 2. Linear Regression

What is Regression?



Linear Regression

- Linearity?

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \epsilon_i$$

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \epsilon_i$$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_p X_i^p + \epsilon_i$$

Linear Model

- Linearity? $\longrightarrow$ Linear Model

$$Y_i \overset{iid}{\sim} (\mu(\mathbf{X}),\ \sigma) \qquad \text{where} \quad E[Y] = \mu(\mathbf{X})$$

$$\mu(\mathbf{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

$$= \mathbf{X}\ \boldsymbol{\beta}$$

Linear Regression Model

- Least Square Estimator

$$\sum\epsilon_i^2 = \sum(Y_i - \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi})^2$$

$$\frac{\partial}{\partial\beta_0}\sum(Y_i - \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi})^2 \overset{set}{=} 0$$

$$\frac{\partial}{\partial\beta_1}\sum(Y_i - \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi})^2 \overset{set}{=} 0$$

.

.

.

$$\frac{\partial}{\partial\beta_\mathrm{p}}\sum(Y_i - \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi})^2 \overset{set}{=} 0$$

Linear Regression Model

- Least Square Estimator

```
> summary(model.a<-lm(exp~income+ factor(Region)))

Call:
lm(formula = exp ~ income + factor(Region))

Residuals:
    Min      1Q  Median      3Q     Max
-77.624 -26.431  -8.821  19.391 174.548

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     21.94531   60.05982   0.365   0.7165
income           0.05337    0.01169   4.566 3.84e-05 ***
factor(Region)2  1.21498   20.02606   0.061   0.9519
factor(Region)3 -0.44452   20.91222  -0.021   0.9831
factor(Region)4 49.92487   19.78310   2.524   0.0152 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Linear Regression Model

- Error term?
    - Mean 0
    - Identical, Independent
    - Normal?

Linear Regression Model

- Likelihood function

Definition (Likelihood)

For $X_1, \cdots, X_n \overset{iid}{\sim} f_X(x;\theta)$, where θ denotes a parameter of interest. The likelihood function is

$$L(\theta;\mathbf{X}) = L(\theta; X_1, \cdots, X_n) = \prod_{i=1}^{n} f_X(X_i;\theta)$$

Linear Regression Model

- Maximum Likelihood Estimator

Definition (Maximum likelihood estimator, MLE)

For $X_1, \cdots, X_n \overset{iid}{\sim} f_X(x;\theta)$, the MLE of θ is

$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}}\, L(\theta;\mathbf{x}).$$

which is equivalent to maximize the logarithm of $L(\theta;\mathbf{x})$ which we call the log-likelihood

$$\ell(\theta;\mathbf{x}) = \log L(\theta;\mathbf{x}).$$

Linear Regression Model

- Maximum Likelihood Estimator

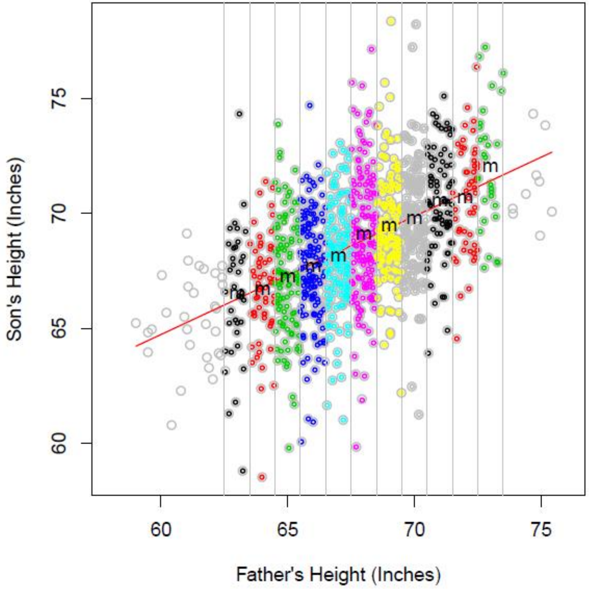
방금 뭐가 지나간거죠….ㅠㅠㅠ

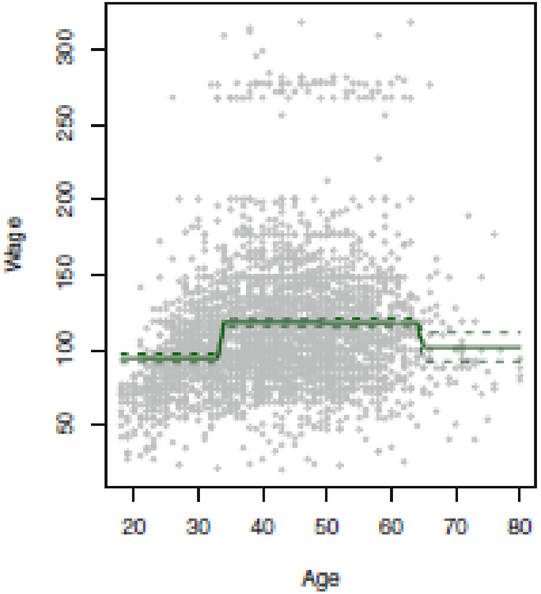
Loss Function

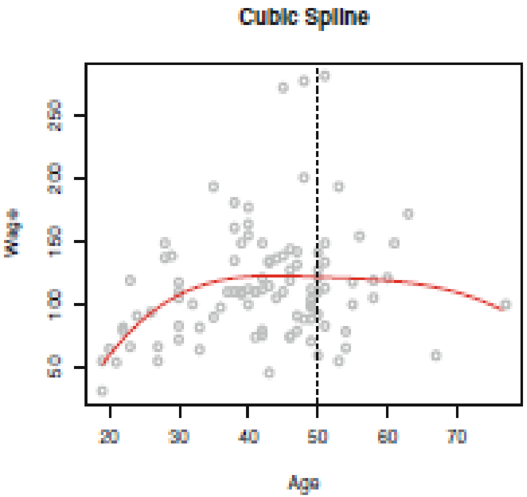
Information Theory

Maximum Likelihood Estimator

Other Regression





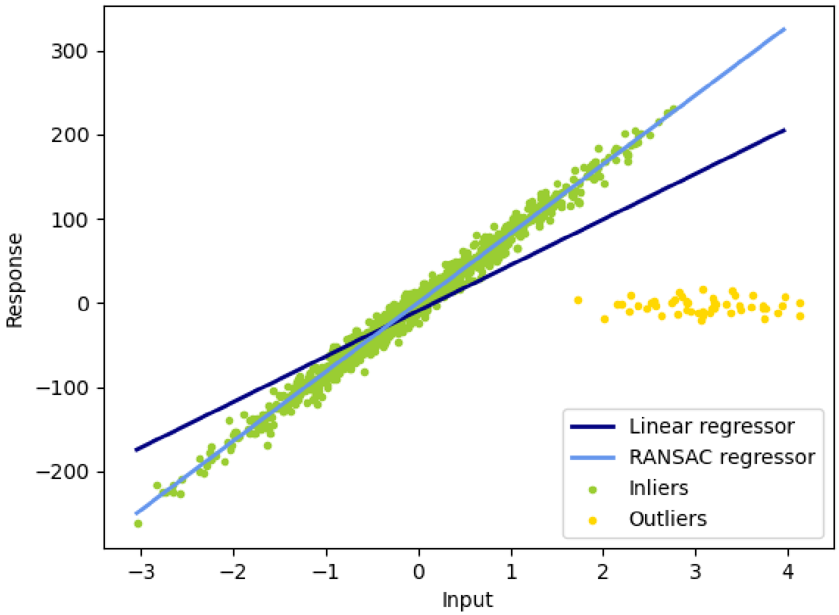


Regression Models vs Outliers

- M-Estimation
- LTS
- DPM

- RANSAC

Regression Models vs Outliers



1. 학습데이터에서 작은 크기의 임의효본을 뽑는다.
2. OLS 추정치를 구하고 추정된 모형에 전체 학습데이터를 적용하여 잔차를 구한다.
3. 잔차의 중위수를 구한 후, 각 잔차의 MAD를 구하여 MAD가 작은 관측치만 모은다. (consensus set)
4. 반복

Why Statistics?

Statistics is important

(a) LSGANs.

(b) Regular GANs.

(c) LSGANs.

(d) Regular GANs.

Reference

자료

21-1 COSE471 데이터과학 - 김진규 교수님

19-2 STAT424 통계적 머신러닝 - 박유성 교수님

교재

파이썬을 이용한 통계적 머신러닝 (2020) – 박유성

ISLR (2013) - G. James, D. Witten, T. Hastie, R. Tibshirani

The elements of Statistical Learning (2001) - J. Friedman, T. Hastie, R. Tibshirani

Hands on Machine Learning (2017) - Aurelien Geron