

Introduction to big data

2018.06.29

Jaewoong Kang



DAEGU
UNIVERSITY

목차

- 제2회 빅데이터 스쿨을 준비하며
- 데이터란?
- '빅'데이터란?
- 데이터 사이언티스트가 하는 5가지 업무
- 데이터 분석의 목적과 방법
- 데이터 해석의 중요성
- 데이터 분석을 위해 해야 하는 공부들

제2회 빅데이터 스쿨을 준비하며

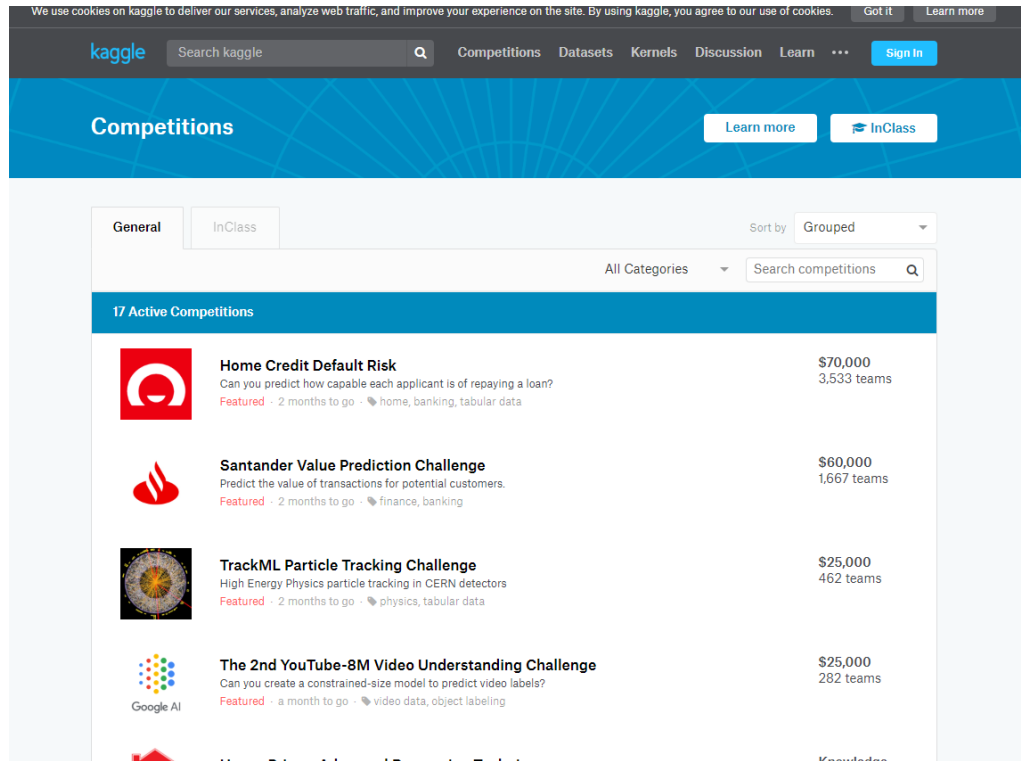
□ Ice Breaking



제2회 빅데이터 스쿨을 준비하며

□ 올해는 Kaggle이다!

- 우리도 남들이 한다는 거 한번 해보자!
- 데이터 사이언티스트가 되기 위해 반드시 한번쯤은 거쳐야 할 관문!



The screenshot shows the Kaggle website's 'Competitions' page. At the top, there is a navigation bar with the Kaggle logo, a search bar, and links for 'Competitions', 'Datasets', 'Kernels', 'Discussion', 'Learn', and 'Sign In'. Below the navigation bar, the 'Competitions' section is highlighted in blue, with 'Learn more' and 'InClass' buttons. The main content area shows '17 Active Competitions' and a list of four featured competitions:

Competition Name	Prize Pool	Number of Teams
Home Credit Default Risk	\$70,000	3,533 teams
Santander Value Prediction Challenge	\$60,000	1,667 teams
TrackML Particle Tracking Challenge	\$25,000	462 teams
The 2nd YouTube-8M Video Understanding Challenge	\$25,000	282 teams

제2회 빅데이터 스쿨을 준비하며

□ Kaggle이란?

- 캐글은 2010년 설립된 예측모델 및 분석 대회 플랫폼이다. 기업 및 단체에서 데이터와 해결과제를 등록하면, 데이터 과학자들이 이를 해결하는 모델을 개발하고 경쟁한다. 2017년 3월 구글에 인수되었다 - 위키백과

□ 즉, 이번 빅데이터의 스쿨은 Not only 수학, But also 코딩!

- Python과 함께하는 데이터 사이언스
- 3일이면 배우는 Python (= 72시간 정도 걸려요...)

데이터란?

□ 데이터의 정의

- The quantities, characters, or symbols on which operations are performed by a computer, being stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media (Google)

□ 데이터의 종류

■ 정형 데이터

- 벡터 또는 행렬의 형태로 표현 가능한 데이터
- Ex) 엑셀 데이터 (table), 그림

■ 비정형데이터

- 그렇지 않은 데이터
- Ex) 자연어

데이터란?

□ 자료

- 의미 없는 기록

□ 정보

- 의미 있는 자료

□ 지식

- 가치 있는 정보

□ 지혜

- 패턴화된 지식



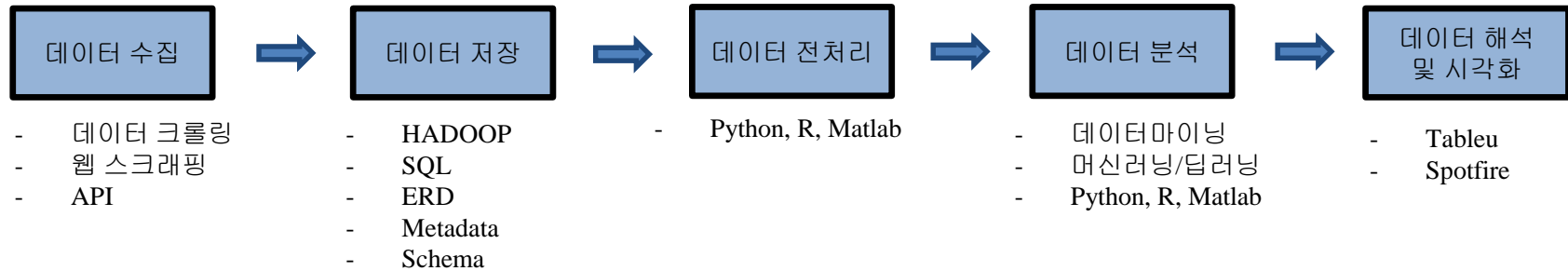
□ TO WHOM?

'빅'데이터란?

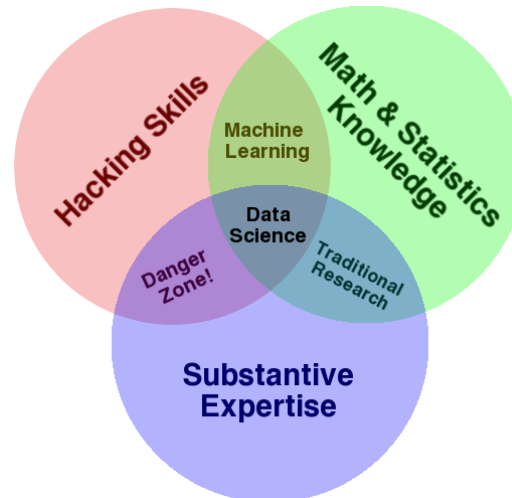
- 빅 데이터란 기존 [데이터베이스](#) 관리도구의 능력을 넘어서는 대량(수십 [테라바이트](#))의 정형 또는 심지어 데이터베이스 형태가 아닌 비정형의 데이터 집합조차 포함한 데이터로부터 가치를 추출하고 결과를 분석하는 기술이다. – 위키피디아
- 데이터의 양(Volume), 데이터 입출력의 속도(Velocity), 데이터 종류의 다양성(Variety), 정확성(Veracity), 가변성(Variability)
- 인공지능과는 다름

데이터 사이언티스트가 하는 5가지 업무

□ 데이터 사이언스의 5가지 순서

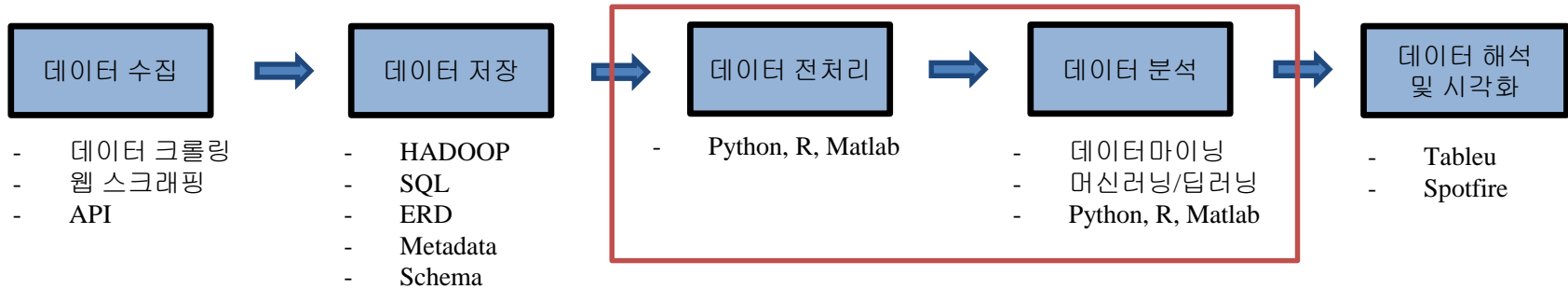


□ 데이터 사이언스 벤 다이어그램

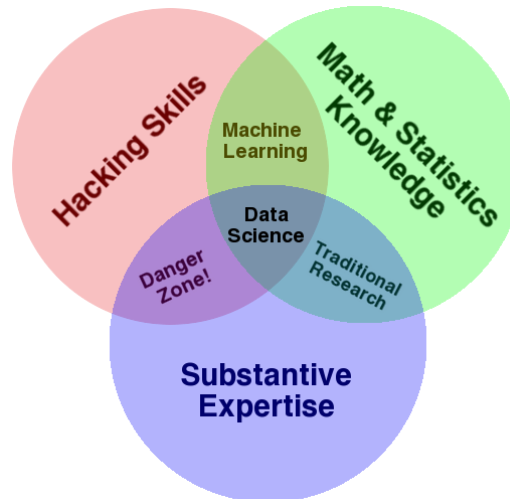


데이터 사이언티스트가 하는 5가지 업무

□ 데이터 사이언스의 5가지 순서



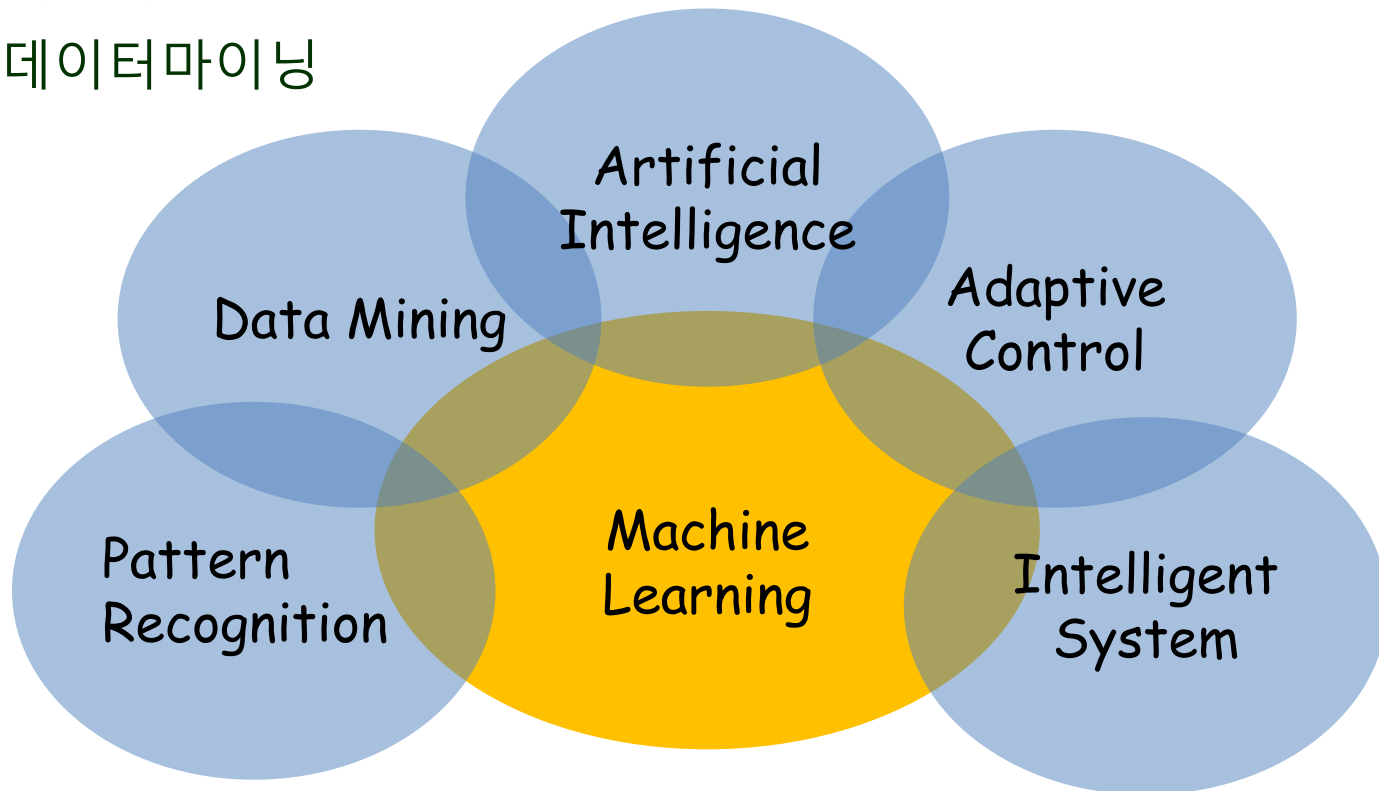
□ 데이터 사이언스 벤 다이어그램



데이터 분석의 목적과 방법

□ 데이터 분석의 방법

- 모델링
- 머신러닝
- 데이터마이닝



데이터 분석의 목적과 방법

- Classification
 - Voice/Face/Fingerprint/Iris/DNA/Signature recognition, Recommendation, Spam filter, Credit card fraud detection
- Regression
 - Loan application analysis, Marketing, Stock market prediction
- Clustering
 - Web-search, Document & information retrieval, Machine translation
- Dimension Reduction
- Strategy Learning
 - Game, Marketing
- Association
 - POS analysis

데이터 분석의 목적과 방법

□ Classification -1

- Each given data has its own class or label
- Once a query is given, the system should tell the class of the query

- Example: Door gate

Permitted
Persons



Query:

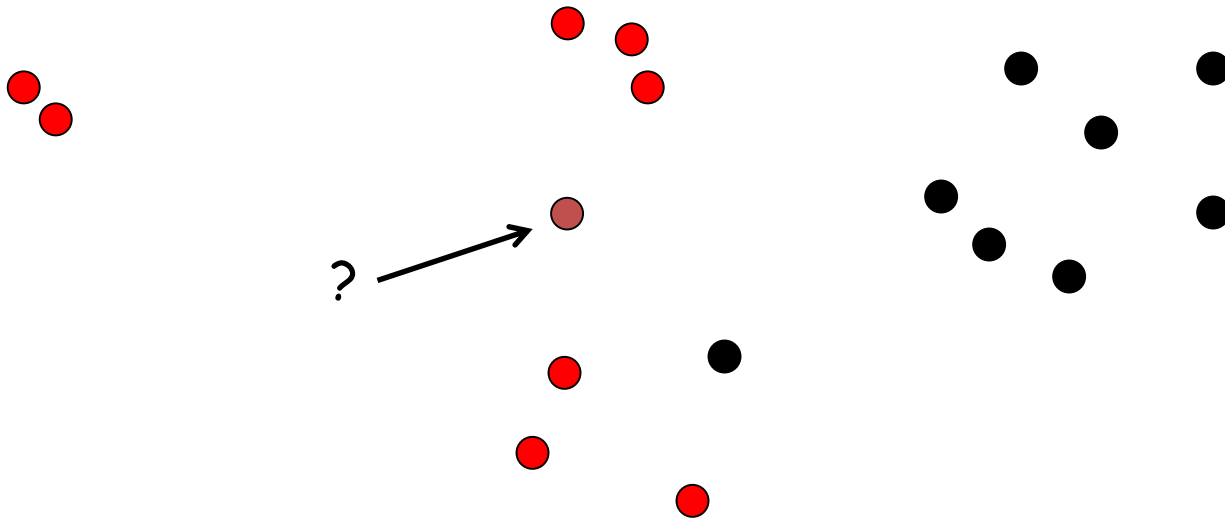


Permitted or Not?

데이터 분석의 목적과 방법

□ Classification - 2

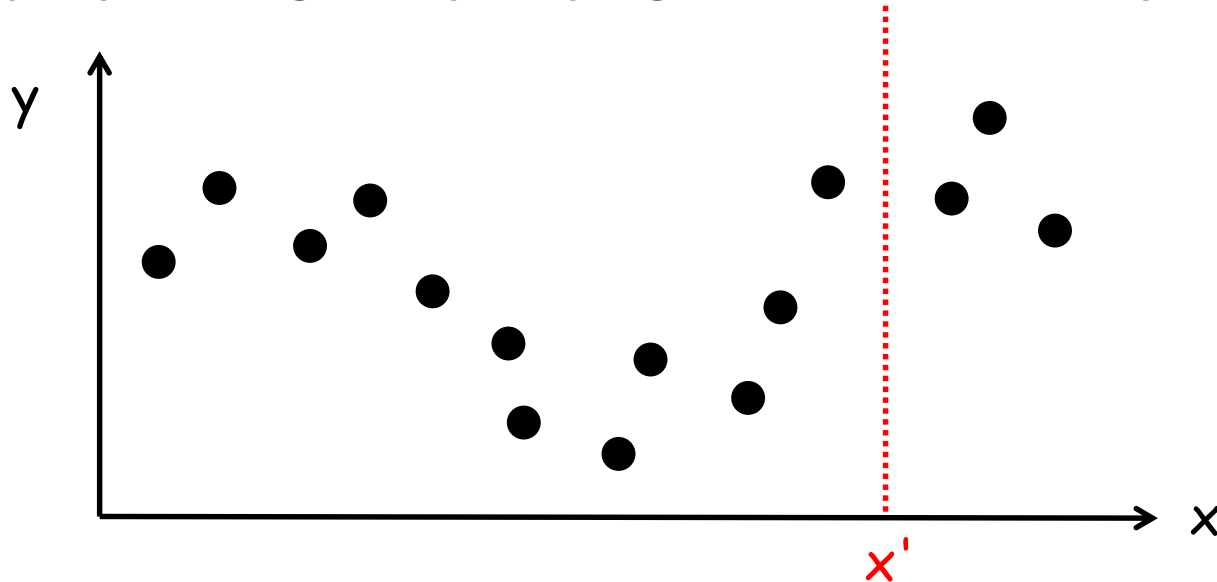
- A set of labeled data is given
- Your program should find the boundary between labels
- If a query is given, your program should answer the label



데이터 분석의 목적과 방법

□ Regression

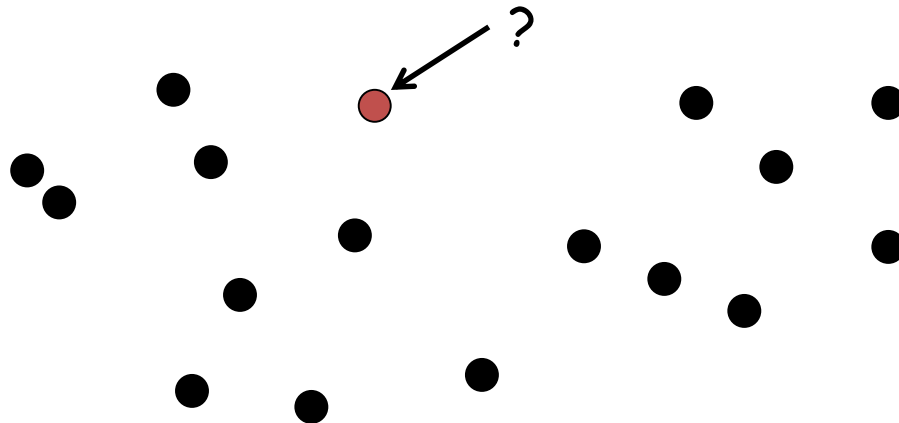
- A set of (\mathbf{x}, y) 's is given. (\mathbf{x} is a vector, y is a real number)
- Your program should find the functional relation between \mathbf{x} and y
- If a query, \mathbf{x}' , is given, your program should answer y for \mathbf{x}'



데이터 분석의 목적과 방법

□ Clustering

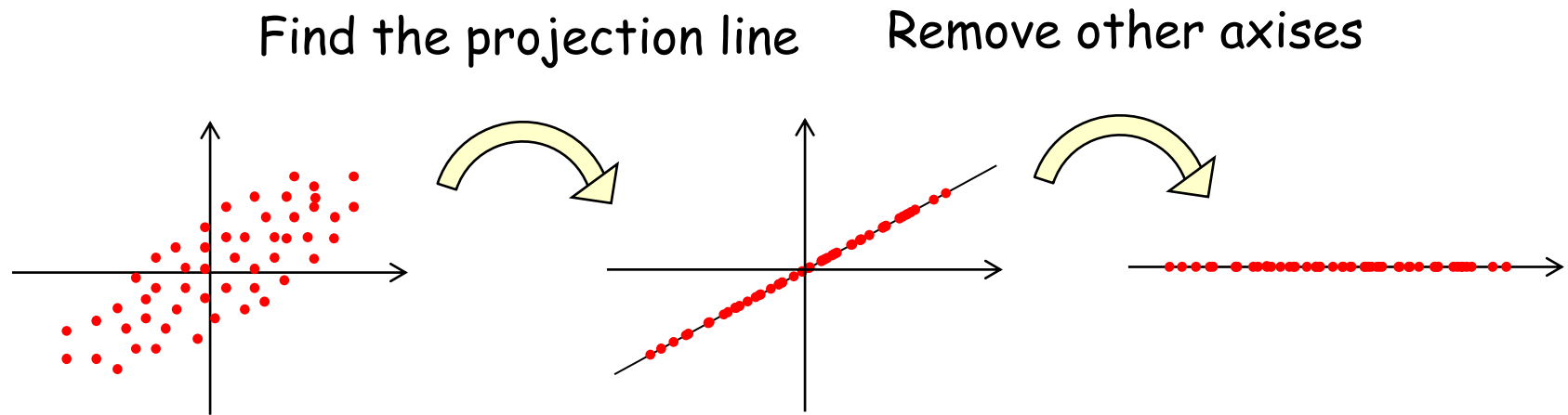
- Unlabeled data is given.
- Your program should group the data (Finding hidden structure of data)
- If a query is given, your program should determine the group in which the query belongs to



데이터 분석의 목적과 방법

□ Dimension Reduction

- A set of unlabeled data is given.
- Your program should reduce the dimension of data by minimizing the loss of information



데이터 분석의 목적과 방법

□ Strategy Learning

- A game is given, but how to win is not known. You may know that you win or not at the end of game
- Your program should learn how to win

□ Association

- A set of items is given
- Your program should find items which appear together

t1: Beef, Chicken, Milk
t2: Beef, Cheese
t3: Cheese, Boots
t4: Beef, Chicken, Cheese
t5: Beef, Chicken, Clothes, Cheese, Milk
t6: Chicken, Clothes, Milk
t7: Chicken, Milk, Clothes

데이터 분석의 목적과 방법

- Supervised Learning
 - Classification, Regression
 - All given data are labeled
- Semi-supervised Learning
 - Classification, Clustering
 - Some data is labeled and some are not
- Unsupervised Learning
 - Clustering, Dimension Reduction, Association
 - Data is not labeled
- Reinforcement Learning
 - Strategy Learning
 - Reward is given to your behaviors

데이터 분석의 목적과 방법

□ Parametric

- A global model is used to describe data
- Your program should estimate the parameters of the global model and answers based on the found model

□ Semiparametric

- A small number of local models are used to describe data
- Your program should estimate the parameters of the local models and answers based on the found models

□ Nonparametric

- Non model based approach
- Your program keeps all the given data and answer based on them
- If a query is given, your program find a small number of closest data instances and answer by combining those
- Aka lazy/memory-based/case-based/instance-based learning

데이터 분석의 목적과 방법

□ Non-meta (Ordinary) Learning

- Methods to learn given data

□ Meta Learning

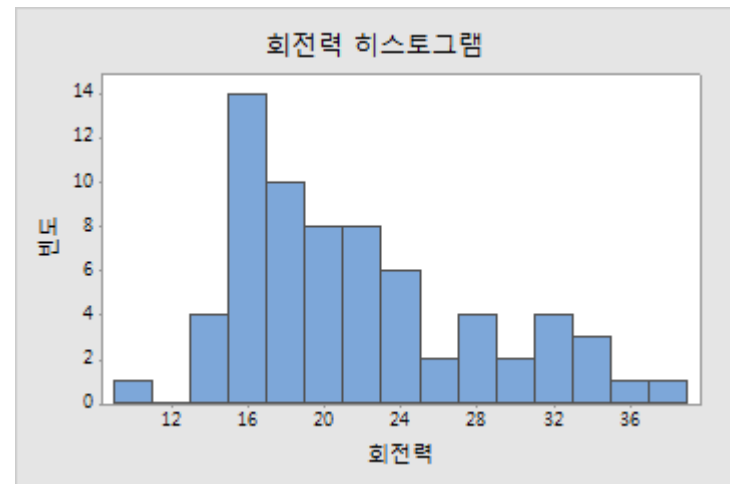
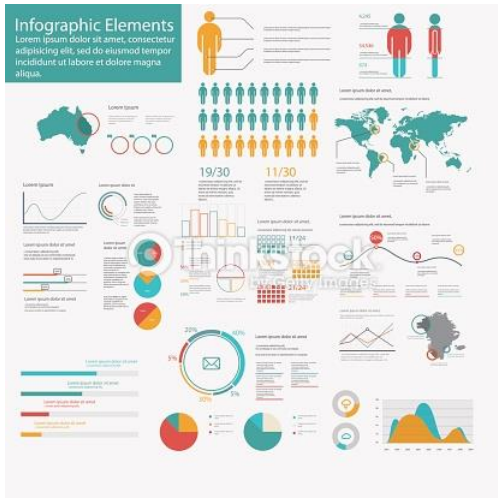
- Methods to learn how to learn
- Boosting, Inductive transfer, ...

데이터 해석의 중요성

□ 데이터 분석이란? DSS! (Decision Support System)

□ 보기 좋은 떡이 맛도 좋다

- 시각화 스킬은 데이터 분석에서 가장 중요함
- 마케팅이 회사 수익의 30%를 올려주듯이 시각화 역시 설득력을 높여주는 좋은 수단
- 데이터 시각화 분야는 아직 개척 중인 전도유망한 분야



데이터 분석을 위해 해야 하는 공부들

□ 수학

- 선형대수/수치선형대수
- 미분적분학
- 수치최적화
- 확률/통계
- 해석학/위상수학

□ 프로그래밍

- Python, R, Matlab

□ 영어 (제일 중요)

참조

- Mye Sohn – Knowledge Engineering lecture
- JH, Lee – Machine Learning lecture
- JS, Lee – Data Mining lecture