

PNU Industrial Data Science

Data Analytics Intro.

Class Orientation

- Lectured by
 - Hyerim Bae, PhD.
 - Tel: 2733
 - E-mail: hrbae@pusan.ac.kr
- Class Time and methods
 - Mon.&Wed. 15:00-16:15 (official)
 - 월요일: 전주에 올린 강의 동영상을 학습하고 월요일 수업시간에는 질의응답
 - 수요일: 실습동영상을 통해서 실습 수행후 과제 제출
- Text book
 - Class materials
 - Online contents will be uploaded
- Evaluation
 - Mid. Exam, Final Exam, Homework, class attendance, etc.
 - The method and the rate of evaluation may be altered considering progress of the class.
- Lecture website: <http://plato.pusan.ac.kr>
- Things to prepare: Google account for using colab

Class schedule

Week	Topic	Practice	etc
1 (09/01, 03)	산업 데이터 과학 개요	Python 기본 실습 1	
2 (09/08, 10)	데이터 시각화	Python 기본 실습 2	
3 (09/15, 17)	데이터 품질 관리	Python을 이용한 데이터 처리 실습 1	
4 (09/22, 24)	데이터 성능	Python을 이용한 데이터 처리 실습 2	
5 (09/29, 10/1)	회귀분석	중고차 가격 데이터를 이용한 가격 예측 실습	
6 (10/6, 8)	K-means 클러스터링	금융 상품 데이터를 이용한 상품 군집화 실습	
7 (10/13, 15)	의사결정나무	자동차부품 데이터를 이용한 불량 판별 실습	
8 (10/20, 22)	중간고사		
9 (10/27, 29)	Navie Bayes	자동차부품 데이터를 이용한 불량 판별 실습	
10 (11/3, 5)	인공신경망	부산항 미세먼지 데이터를 이용한 미세먼지 예측 실습	
11 (11/10, 12)	CNN	제조 이미지 데이터를 이용한 이미지 판별 실습	
12 (11/17, 19)	시계열 분석	해운운임지수 데이터를 이용한 시계열 예측 실습	
13 (11/24, 26)	텍스트마이닝	웹 크롤링을 이용한 네이버 뉴스 크롤링 구현 및 Word Cloud 실습	
14 (12/1, 3)	강화학습	OpenGym을 이용한 강화학습 실습	
15 (12/8, 10)	기말고사		

Contents

산업데이터과학은 산업현장에서 수집된 데이터를 분석하는데 필요한 기초 소양을 강의합니다.

01

What is DS

02

Induction & Deduction

03

Applications

04

Scope of the class



What is data mining and data science?



Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use.



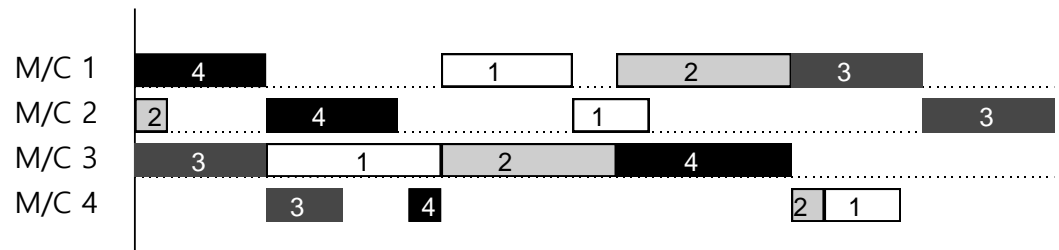
Data science is a "concept to unify statistics, data analysis, machine learning, domain knowledge and their related methods" in order to "understand and analyze actual phenomena" with data. It uses techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, domain knowledge and information science. Turing award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge.

What is IDS (Industrial Data Science) ?

- What can we do using DS for a manufacturing company?
 - Classification/Prediction
 - Correlation
 - Clustering
 - Rule finding
- Using DS for industrial purpose
 - Productivity
 - Economic cost saving
 - Quality improvement

job#	order of Machines
job 1	3-1-2-4
job 2	2-3-1-4
job 3	3-4-1-2
job 4	1-2-4-3

job#	Processing time
job 1	4-3-2-2
job 2	1-4-4-1
job 3	3-2-3-3
job 4	3-3-1-4



What is A, and what is B?

"A people bow but B people shake hands



- A is Asian B is western












- A with similar colored clothes, B with different colored clothes



Data Science and Machine Learning

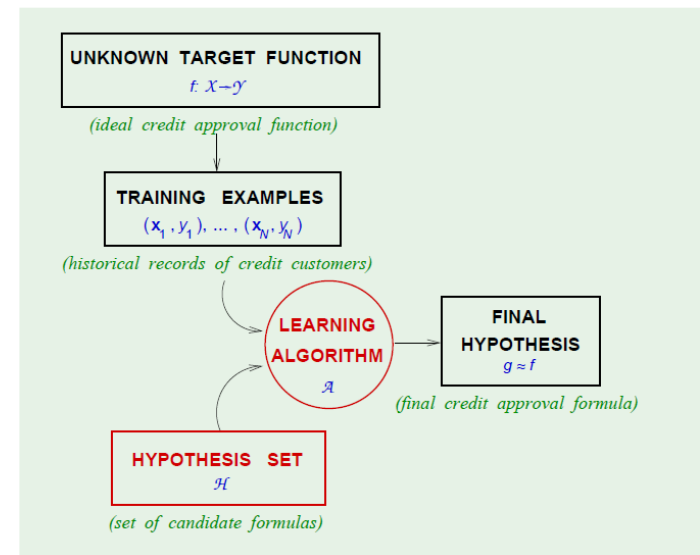
- Make machine learn

- Input      
- Output {"bow(1)", "Shake hands(0)"}
- Data { , 1 }, { , 0 }, { , 0 }

- Finding a function which can predict the output from a new input

Formalization:

- Input: \mathbf{x} (customer application)
 - Output: y (good/bad customer?)
 - Target function: $f : \mathcal{X} \rightarrow \mathcal{Y}$ (ideal credit approval formula)
 - Data: $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)$ (historical records)
- ↓ ↓ ↓
- Hypothesis: $g : \mathcal{X} \rightarrow \mathcal{Y}$ (formula to be used)



What is (Machine) learning?

- Finding ' f ' such that

$$Y = f(X)$$

rule

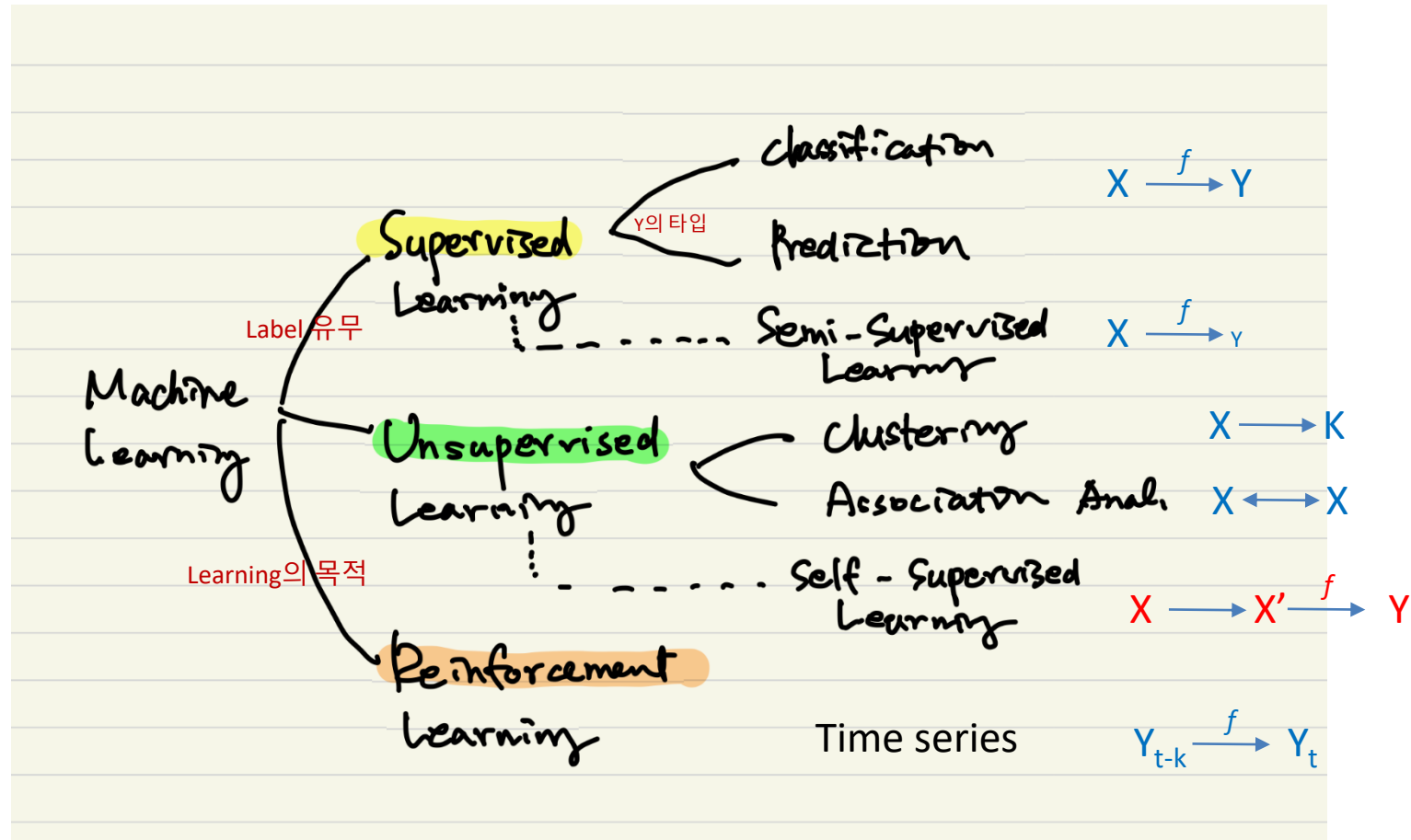
pattern

knowledge

- We use X and Y to find ' f '

Learning method

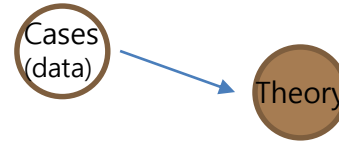
- Traditional learning



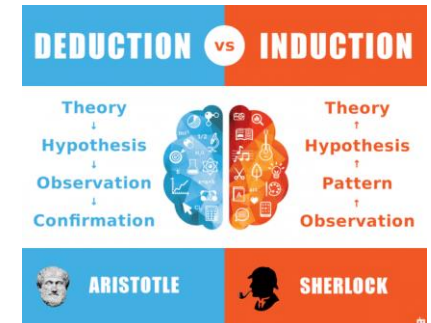
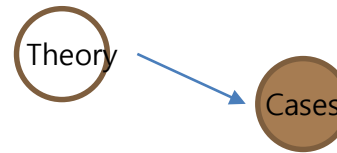
- Deep learning

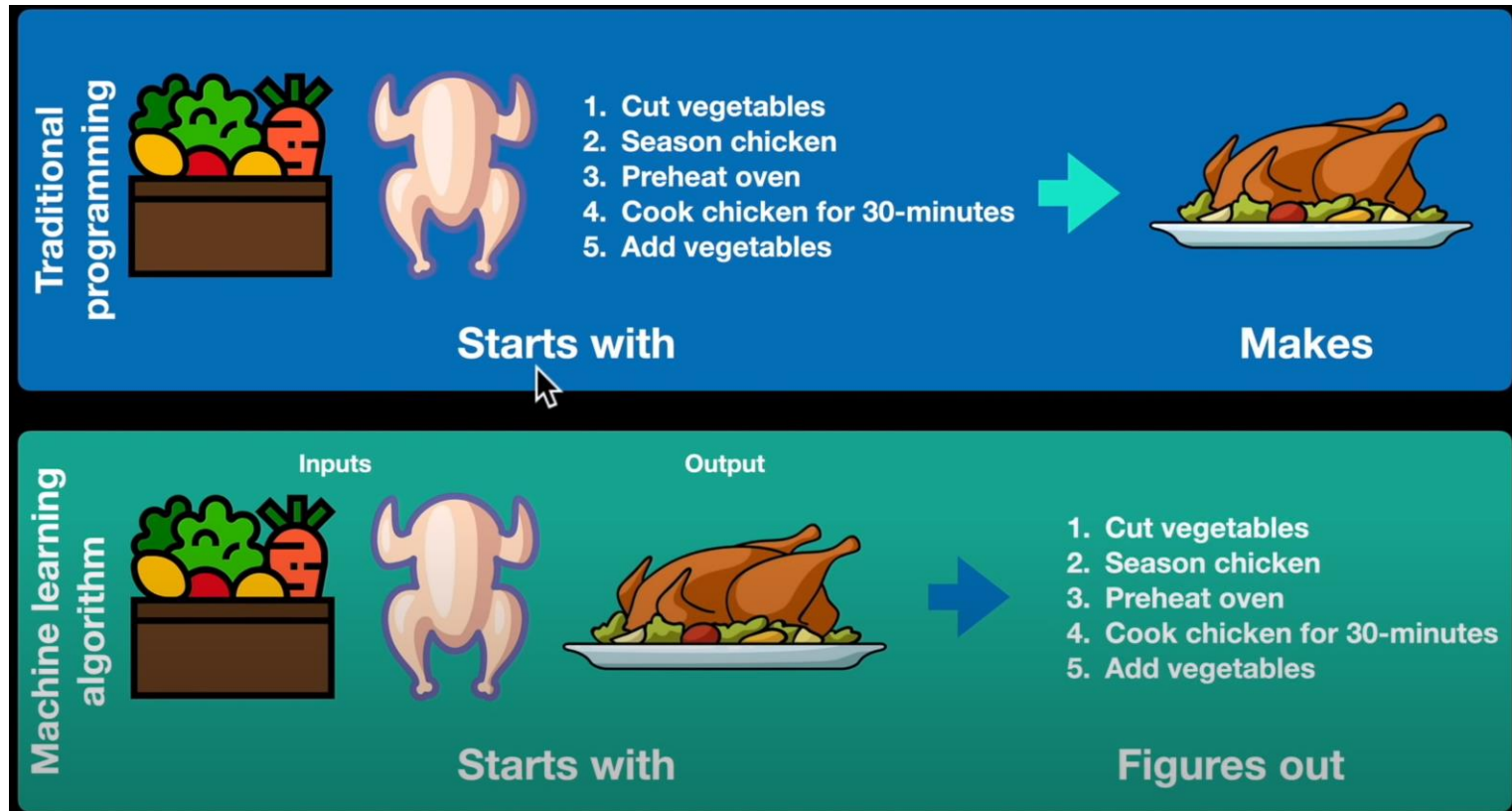
Inductive vs. Deductive

- Induction(귀납): Specific to General
 - A dies, B dies, C dies, ...
 - Everybody dies.



- Deduction(연역): General to specific
 - Every man dies. Socrates is a man.
 - Socrates dies.





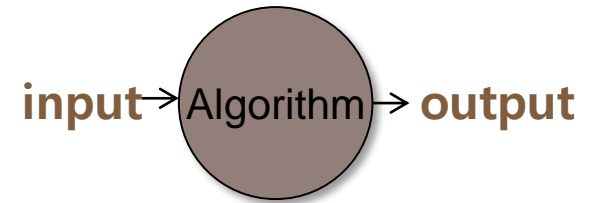
https://www.youtube.com/watch?v=pHiMN_gy9mk&t=358s

Backgrounds

- Large volume of data
 - Created
 - Stored
- Computing power
- S/W package
- Competition among enterprises

Basic concepts

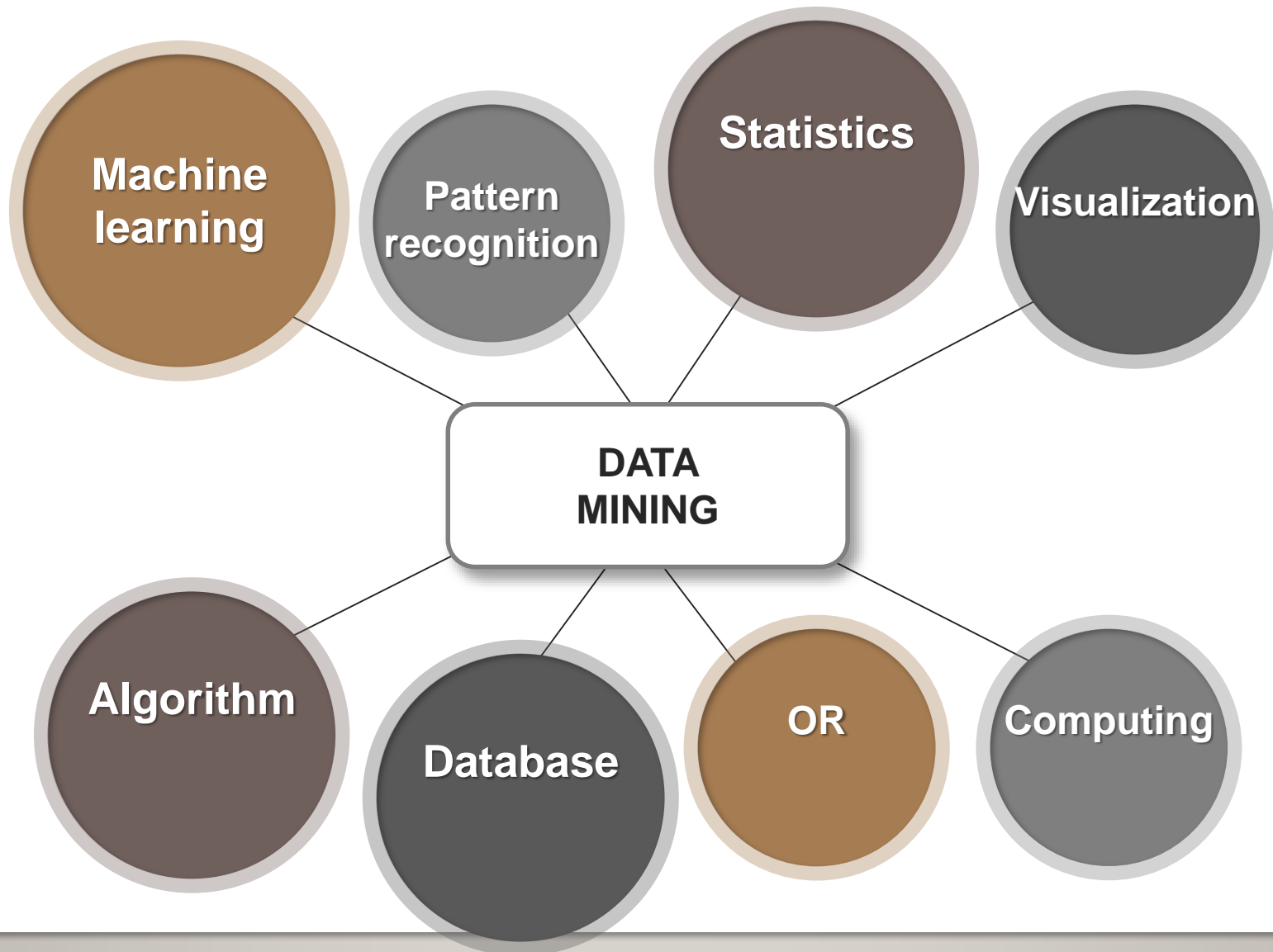
- Algorithm
 - A procedure to implement a particular technique
- Independent variable vs. Dependent variable
 - Predictor, input variable, attribute
 - Response, output variable, outcome variable
- Parameter (Coefficient) estimation
- Error (Residual)
 - Train error
 - Test error
- Case, Observation
- Supervised vs. Unsupervised



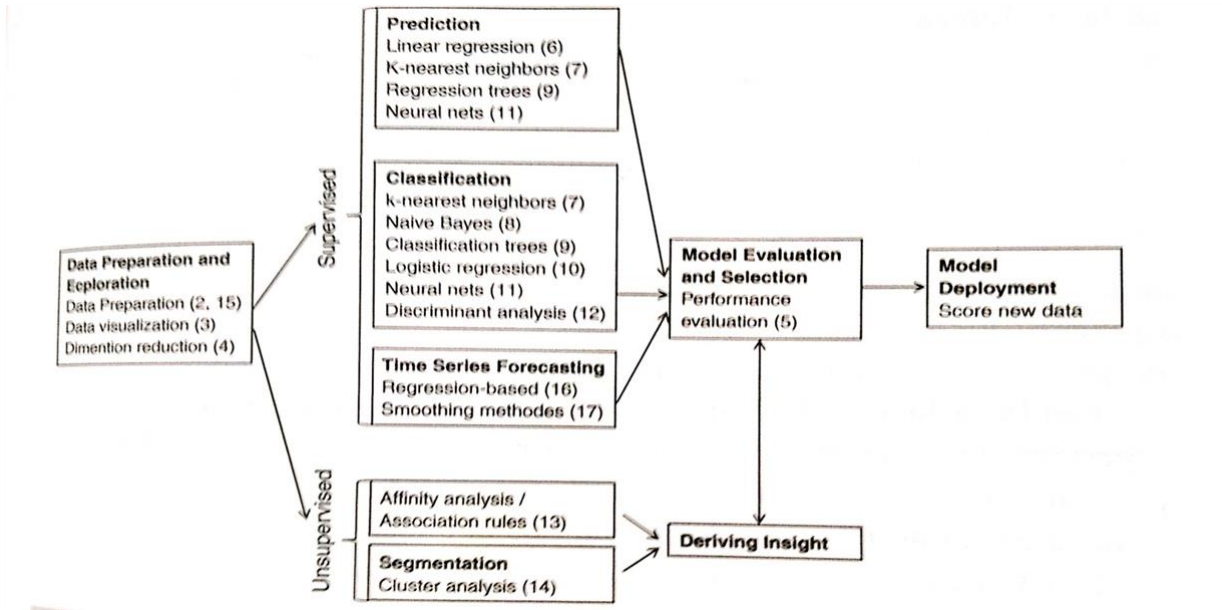
Applications of DA

- Web page analysis
- Recommender system
- MBA (Market basket analysis)
- Bio. And Med. Data analysis
- Finance
- Tele communications
- Manufacturing

DA elements



DA topics in this class



Deep learning
Reinforcement learning



Productivity
Economic cost saving
Quality improvement