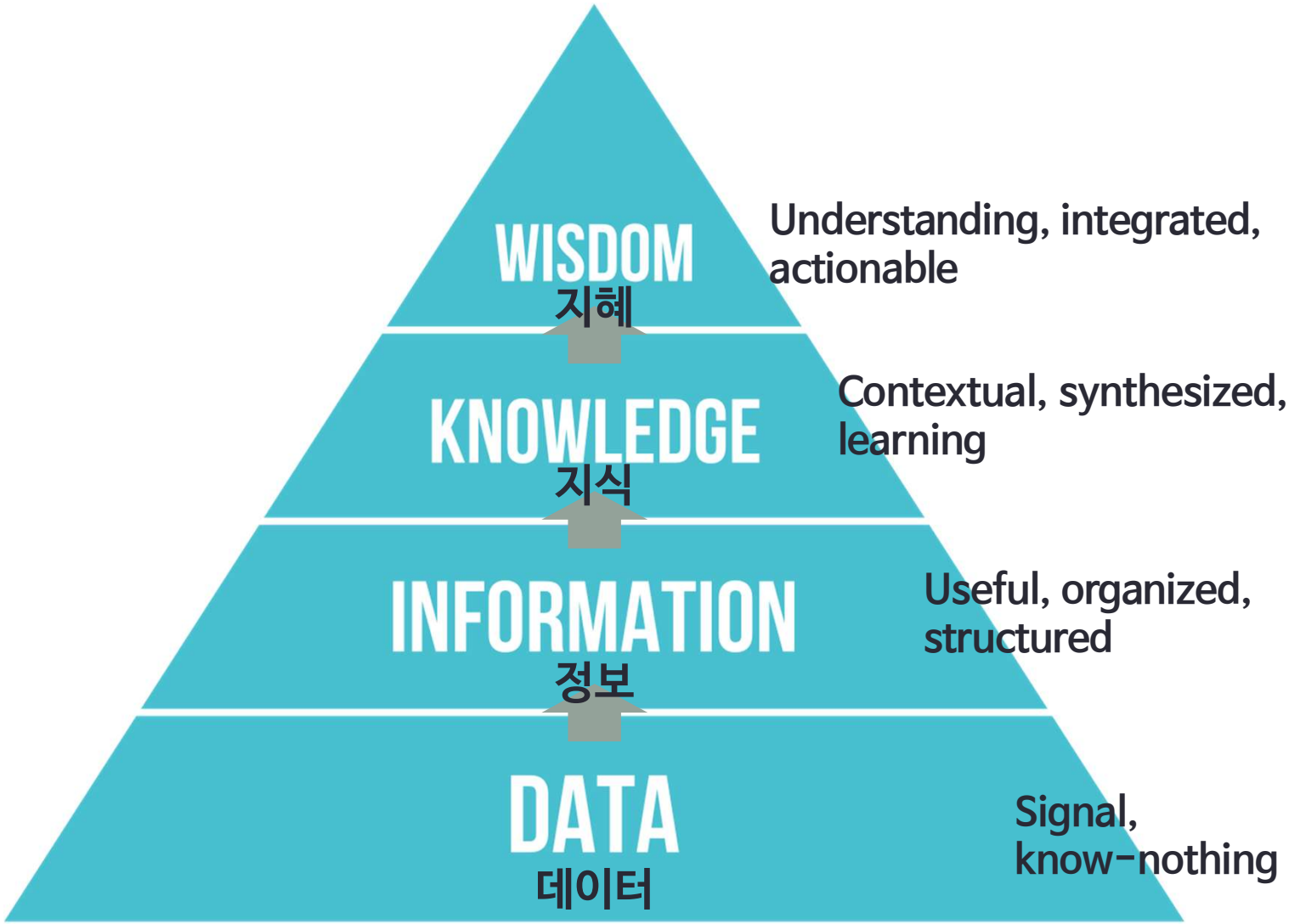
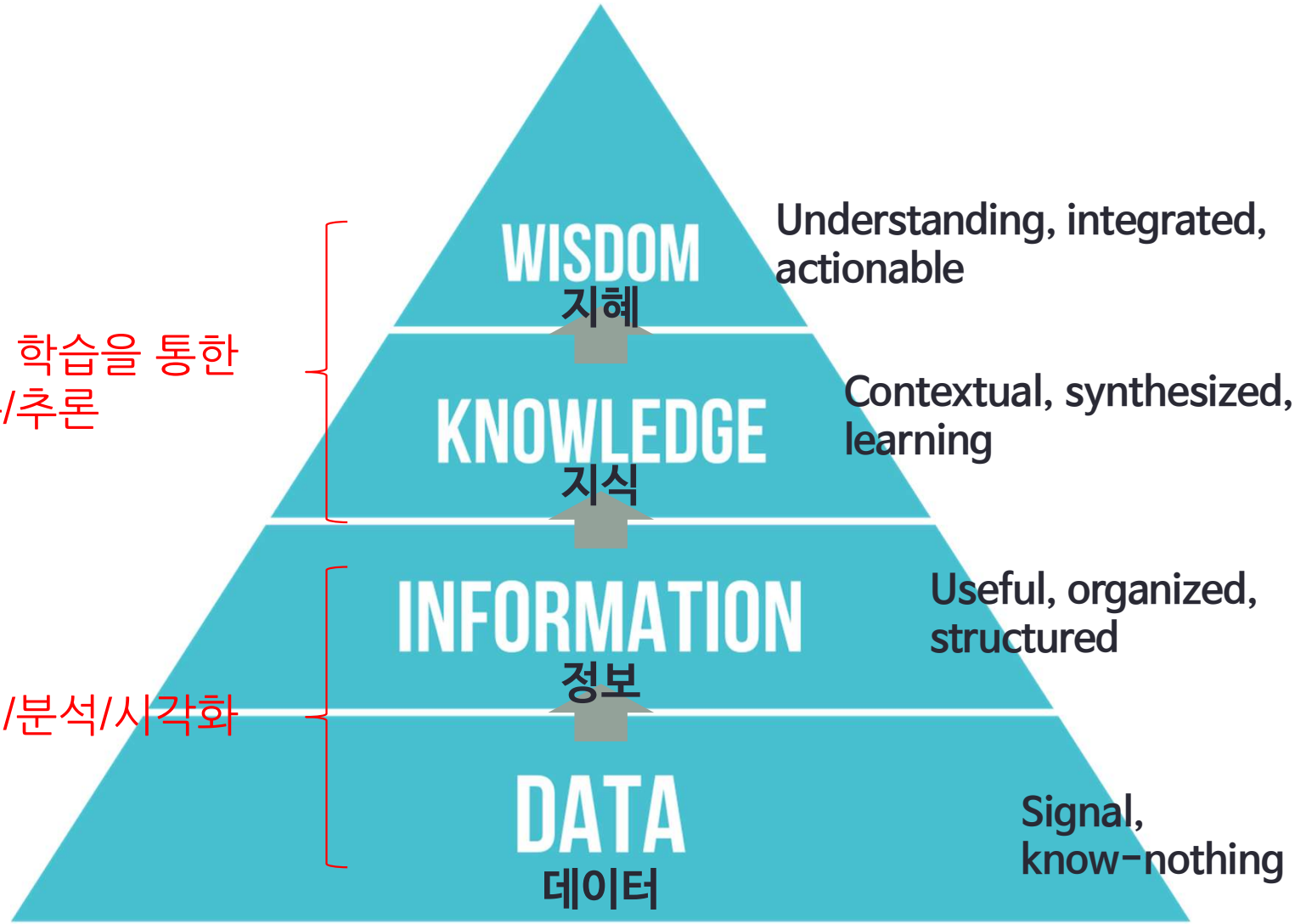
데이터사이언스 (DataScience)

시스템경영공학부
이지환 교수

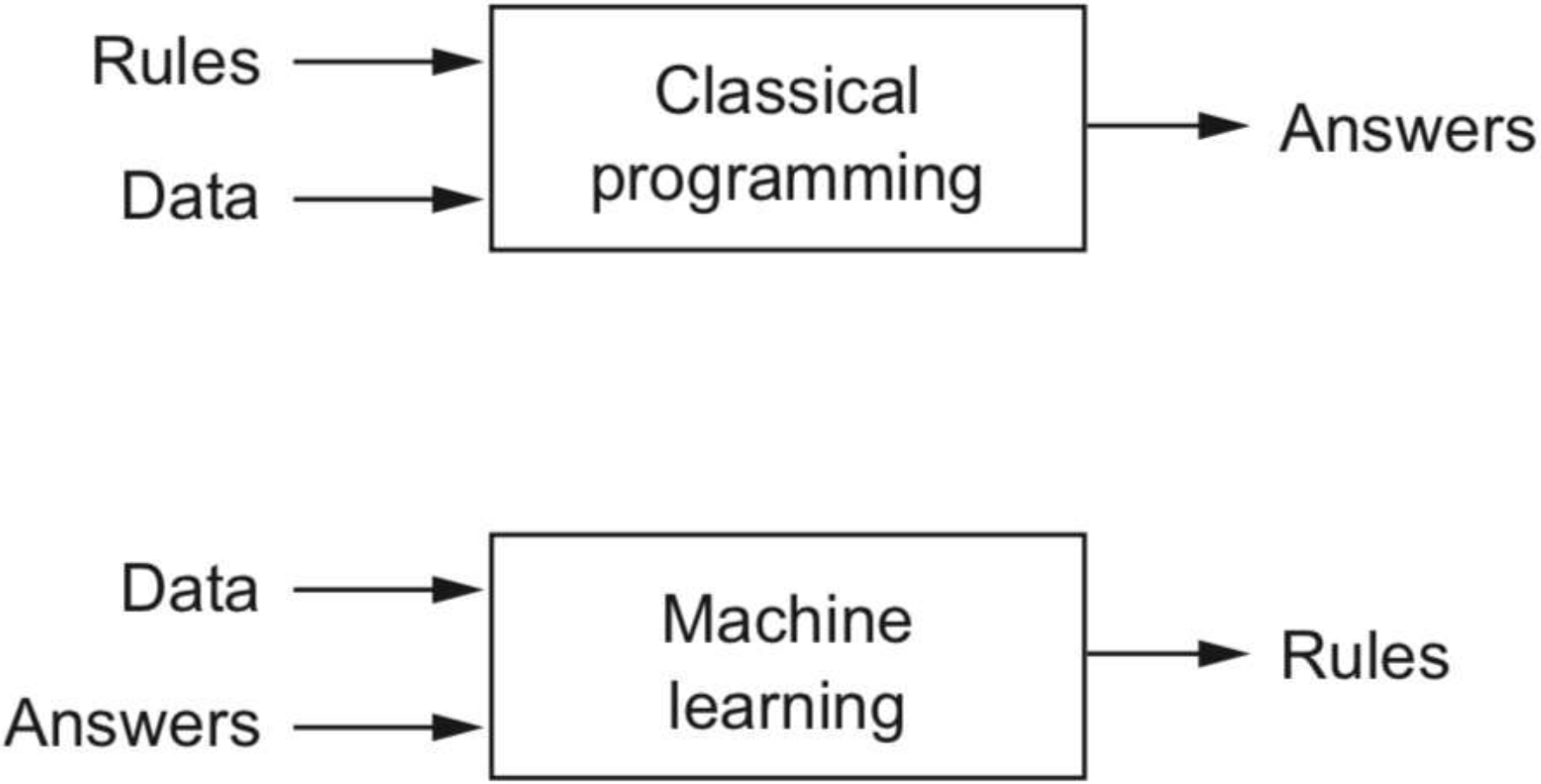
데이터로부터 지혜를 얻기까지..



데이터로부터 지혜를 얻기까지..

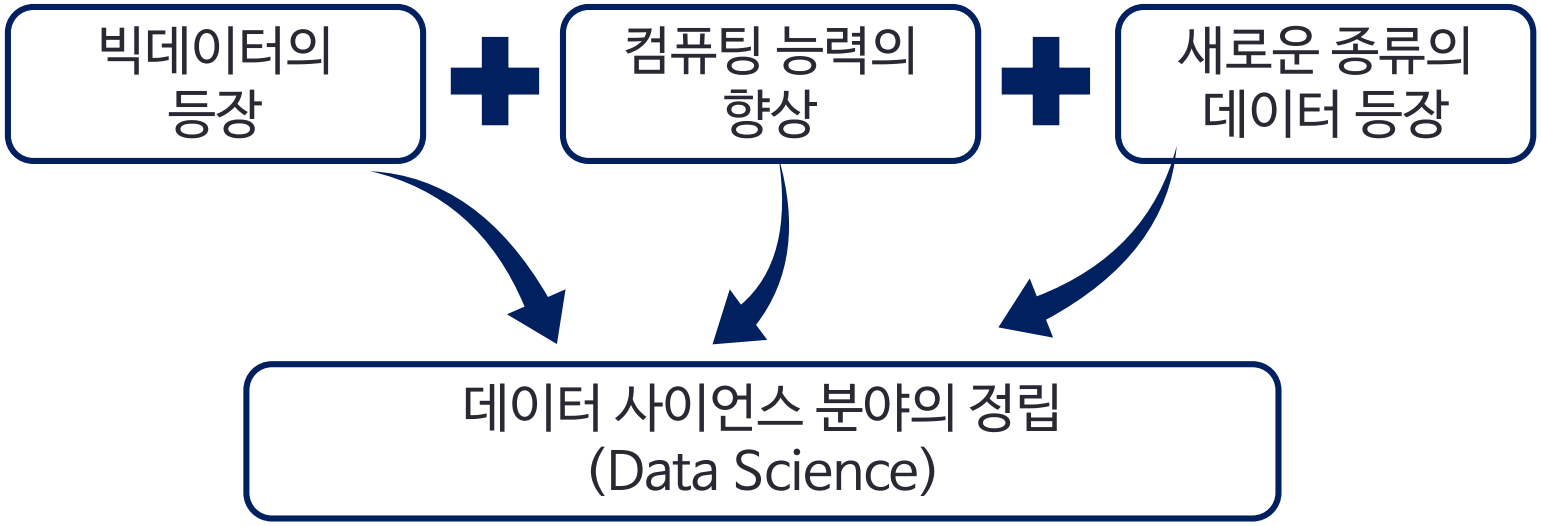


데이터 기반 문제해결


Rules
Data
Classical programming
Answers
Data
Answers
Machine learning
Rules

데이터 사이언스

- **데이터 과학**(data science)이란 정형, 비정형 형태를 포함한 다양한 데이터로부터 지식과 인사이트를 추출하는데 과학적 방법론, 프로세스, 알고리즘, 시스템을 동원하는 융합분야다.

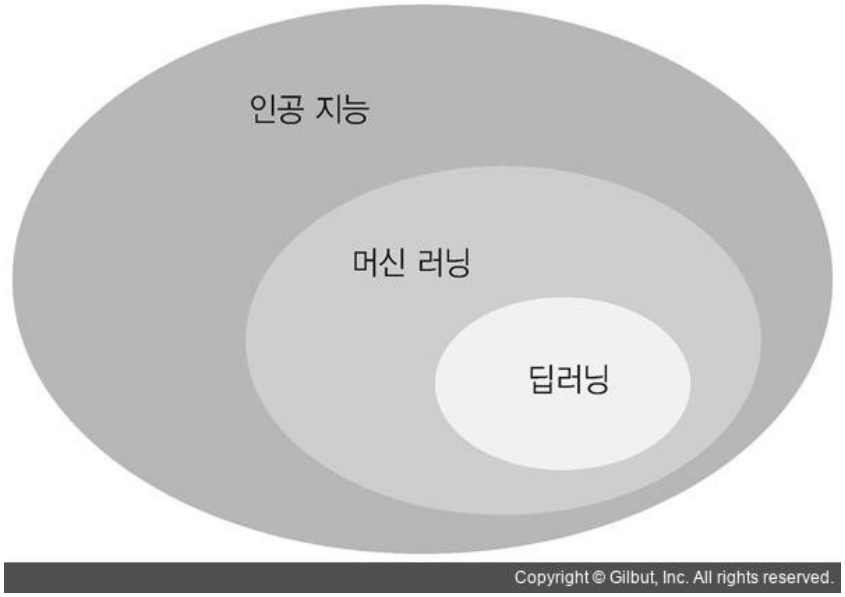


머신러닝, 딥러닝, 인공지능

- 딥러닝
  - Extract patterns from data using neural network
  - 인공신경망 구조를 사용하여 사용하여 데이터로부터 패턴을 추출

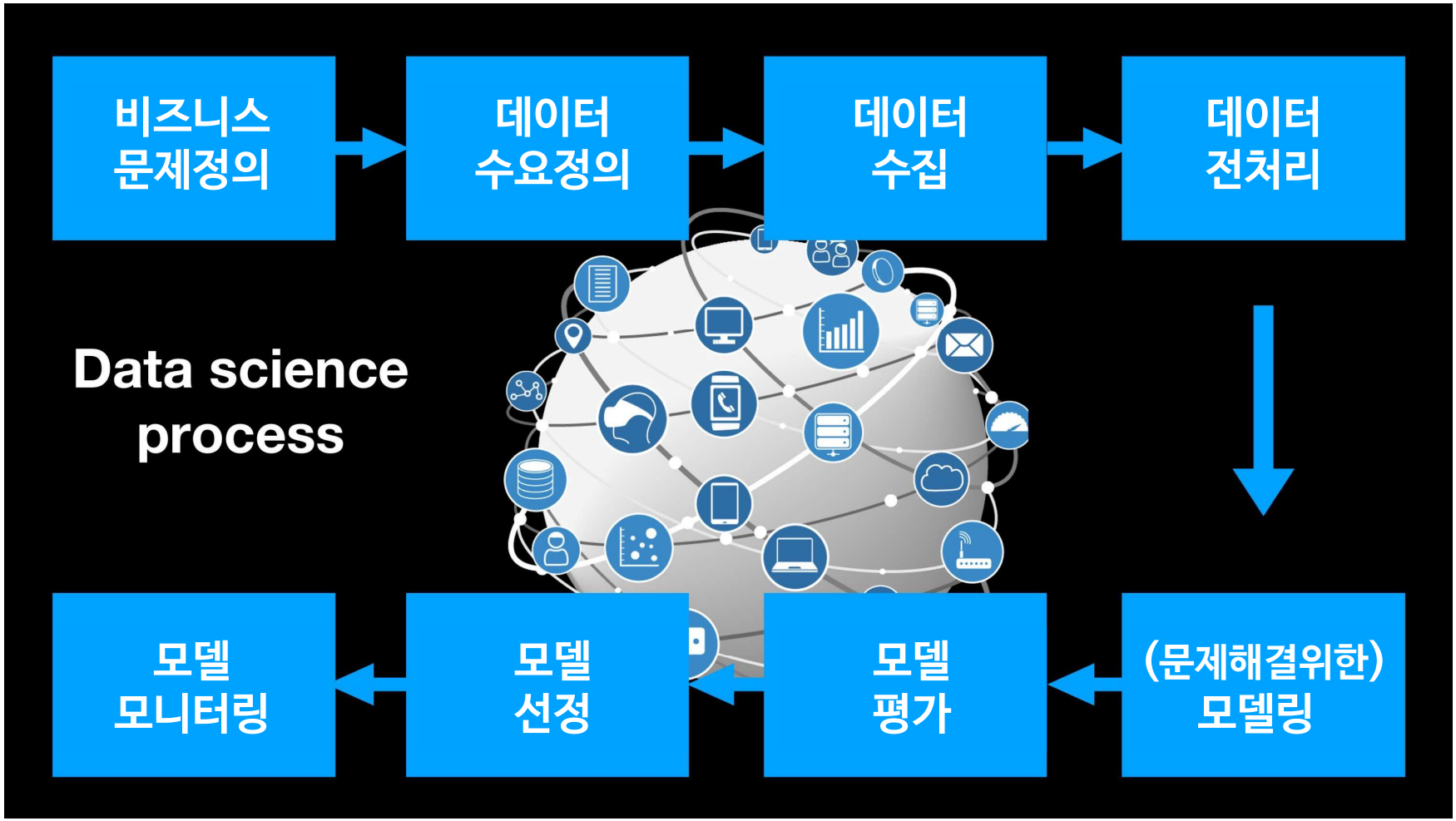
- 머신러닝
  - Ability to learn without explicitly being programmed
  - 컴퓨터가 알아서 데이터로부터 규칙을 발견하도록 하는 것

- 인공지능
  - Any techniques that enables computers to mimic human behavior
  - 컴퓨터로 하여금 사람의 생각과 행동을 모사(mimic)하도록 하는 모든 기술



목표

- 머신러닝 기법을 적용하는 일반적인 절차의 습득



데이터사이언스 프로세스

시스템경영공학부
이지환 교수

데이터사이언스 주요 프로세스

- 특성, 관측점
- 종속변수, 독립변수

- 가설
- 패러미터

- 학습의 의미

- 예측
- 추론

데이터사이언스 주요 프로세스

- 특성, 관측점
- 종속변수, 독립변수

- 가설
- 패러미터

- 학습의 의미

- 예측
- 추론

데이터

- 과거에 일어난 사실들을 특정 형식에 맞게 기록해놓은 자료

- 데이터의 예시
  - 어떤 기업내 사원 50명에 대한 데이터를 다음과 같이 수집하였다

	경력	연봉($)
1	0.3	40000
2	0.5	48000
3	5	70000
49	11	120000
50	6	65000

데이터

- 특성(Features)
  - 데이터의 특성을 구분 지어 설명할 수 있는 것 (=변수,열)

- 관측점(Observation)
  - 특성에 따라 기록되어있는 서로 다른 객체들 (=행, data point)

	경력	연봉($)
1	0.3	40000
2	0.5	48000
3	5	70000
49	11	120000
50	6	65000

독립변수와 종속변수

	경력	연봉($)
1	0.3	40000
2	0.5	48000
3	5	70000
49	11	120000
50	6	65000

- 종속변수(dependent variable)
  - 독립변수에 영향을 받는 특성
  - (예시) 연봉
  - (소문자) y
  - A.k.a target variable, output variable, label

- 독립변수(independent variable)
  - 종속변수에 영향을 주는 것으로 여겨지는 특성들
  - (예시) 경력
  - (대문자) **X**
  - a.k.a predictor, input variable, regressor

데이터를 가지고 해볼 수 있는 질문들: 예측

- 예측 (Prediction)
  - 독립변수가 주어져 있는 상태에서, 종속변수의 값을 추측
  - Predict the outcome for new data point.
  - (예시) 경력이 3년, 6년, 9년인 사람에게 각각 얼마의 연봉을 주어야 할까?

데이터 (과거)

	경력	연봉($)
1	0.3	40000
2	0.5	48000
3	5	70000
49	11	120000
50	6	65000

예측 (일어나지 않음)

3	??
6	??
9	??

데이터를 가지고 해볼 수 있는 질문들: 추론

- 추론 (Inference)
  - 독립변수와 종속변수간의 설명가능한 관계를 파악
  - Learn about data generation process
  - (예시) 1년 연봉이 오를때마다 20,000$씩 연봉이 오른다.

데이터 (과거)

	경력	연봉($)
1	0.3	40000
2	0.5	48000
3	5	70000
49	11	120000
50	6	65000

예측 (일어나지 않음)

3	??
6	??
9	??

데이터 사이언스 주요 프로세스

- 특성, 관측점
- 종속변수, 독립변수

- 가설
- 패러미터

- 학습의 의미

- 예측
- 추론

모형(Model)

- {예측}과 {추론}을 위해서는 데이터를 활용하여 {모형}을 만들어야 한다

- 모형(Model)
  - 독립변수와 종속변수의 관계에 대한 가설을 수학적으로 표현한 것
  - A machine learning model can be a mathematical representation of a relationship between X and y

- 모형의 종류
  - X와 y의 관계에 대한 가설에 따라 다양한 종류의 모형 존재
    - 회귀모형
    - 로지스틱회귀모형
    - K-Nearest Neighbor
    - Support Vector Machine
    - Random Forest
    - 딥뉴럴네트워크

모형의 예시

- 같은 데이터라 할 지라도 다양한 X와 y의 관계에 대한 다양한 모형을 생각해 볼 수 있음

가설1: 연봉이 경력에 비례하여 증가할 것이다

	경력	연봉($)
1	0.3	40000
2	0.5	48000
3	5	70000
49	11	120000
50	6	65000

가설2: 연봉이 경력의 제곱에 비례하여 증가할 것이다

가설n: 연봉이 경력의 n차식으로 표현될 것이다.

모형의 예시

- 같은 데이터라 할 지라도 다양한 X와 y의 관계에 대한 다양한 모형을 생각해 볼 수 있음

	경력	연봉($)
1	0.3	40000
2	0.5	48000
3	5	70000
49	11	120000
50	6	65000

가설1: 연봉이 경력에 비례하여 증가할 것이다

$$y = \beta_0 + \beta_1 x$$

가설2: 연봉이 경력의 제곱에 비례하여 증가할 것이다

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

가설n: 연봉이 경력의 n차식으로 표현될 것이다.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_n x^n$$

패러미터 (parameter)

- 모델안에서 독립변수(X)와 종속변수(y)간의 관계를 표현하기 위해 조절할 수 있는 매개변수

	경력	연봉($)
1	0.3	40000
2	0.5	48000
3	5	70000
49	11	120000
50	6	65000

가설1: 연봉이 경력에 비례하여 증가할 것이다

$$y = \beta_0 + \beta_1 x$$

패러미터

$$\beta_0, \beta_1$$

가설2: 연봉이 경력의 제곱에 비례하여 증가할 것이다

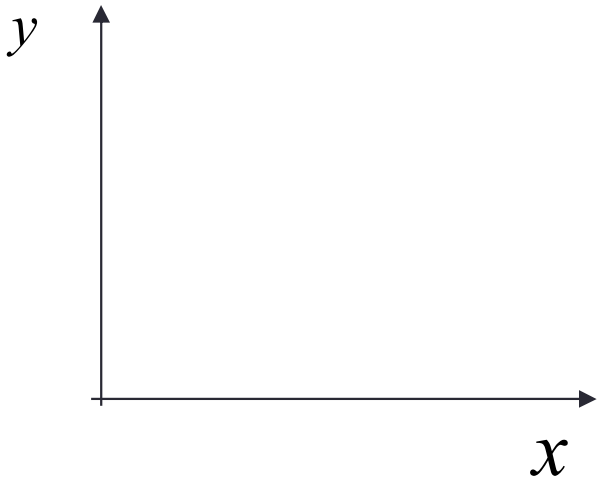
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 \Rightarrow \beta_0, \beta_1, \beta_2$$

가설n: 연봉이 경력의 n차식으로 표현될 것이다.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_n x^n \Rightarrow \beta_0, \beta_1, \ldots, \beta_n$$

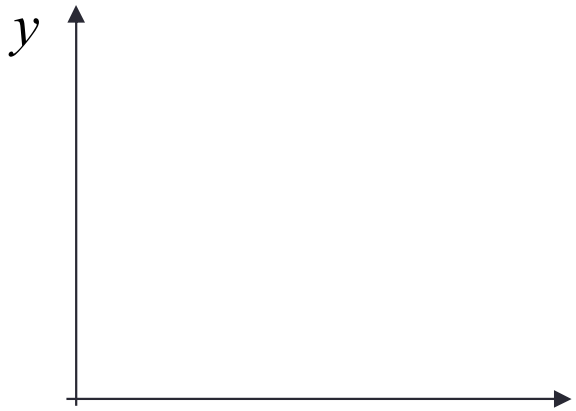
패러미터 (parameter)의 역할

- 패러미터를 어떻게 잡느냐에 따라 X와 y의 관계를 다르게 표현할 수 있다!



$\beta_0 = 2$

$\beta_1 = 0$



$\beta_0 = 2$

$\beta_1 = 1.5$

$\beta_0 = 2$

$\beta_1 = -0.5$

데이터 사이언스 주요 프로세스

- 특성, 관측점
- 종속변수, 독립변수

- 가설
- 패러미터

- 학습의 의미

- 예측
- 추론

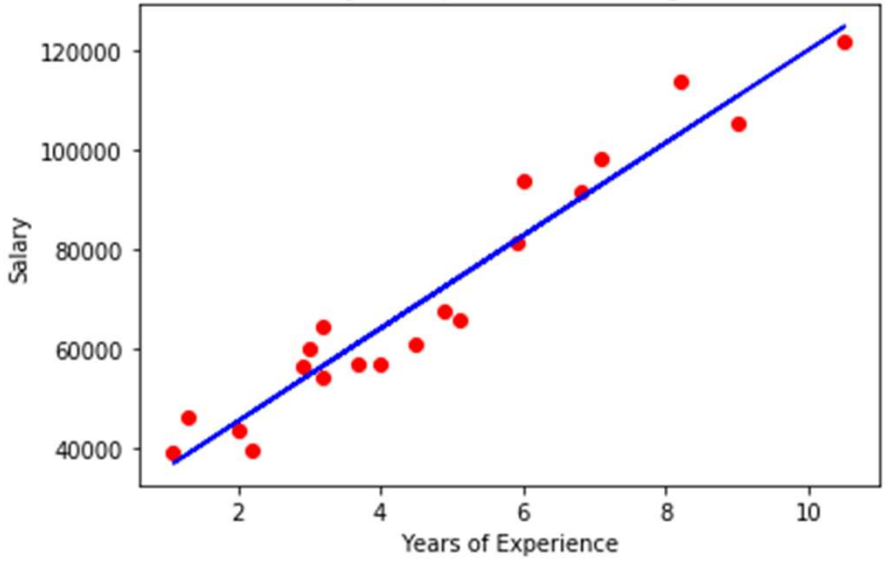
모델의 학습(Learning)

- 모델: X와 y에 대한 수학적인 관계
- 학습: 주어진 데이터를 가장 잘 나타내는 패러미터를 찾아나가는 과정

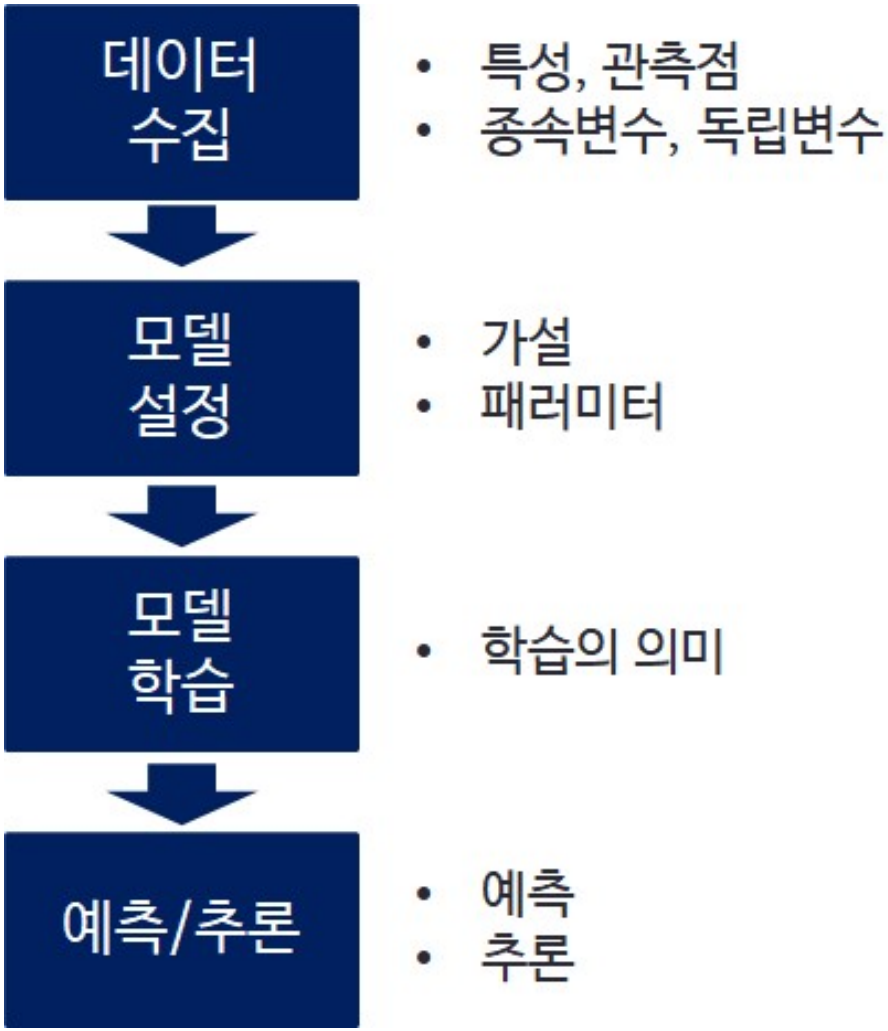
$$y = \beta_0 + \beta_1 x$$

$$\beta_0, \beta_1$$

	경력	연봉($)
1	0.3	40000
2	0.5	48000
3	5	70000
49	11	120000
50	6	65000



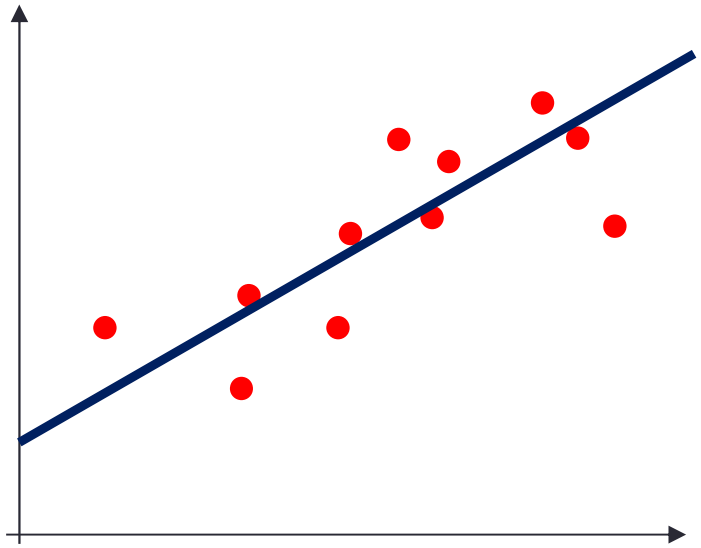
데이터 사이언스 주요 프로세스

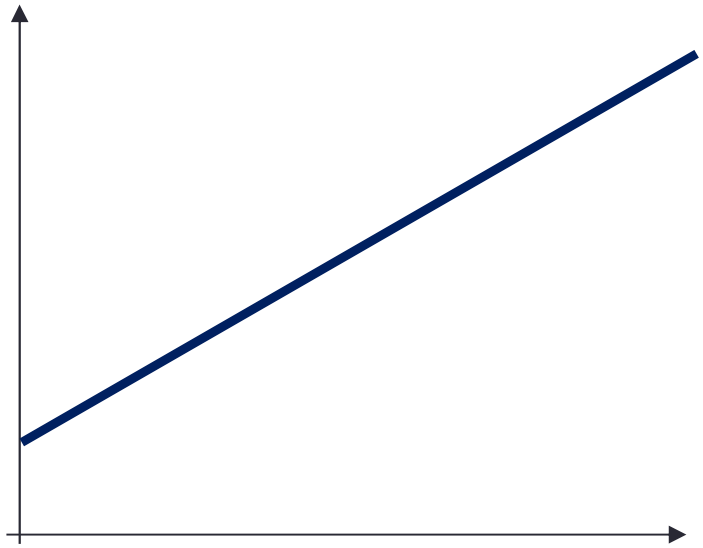


모델의 이용: 예측(Prediction)

- 새로운 데이터의 x값이 주어질때, 모델에서 학습된 패러미터를 이용하여 y값을 예측
- 경력이 4년, 6년, 8년인 직원의 연봉은 각각 얼마인가?

$$y = 30000 + 5000x$$

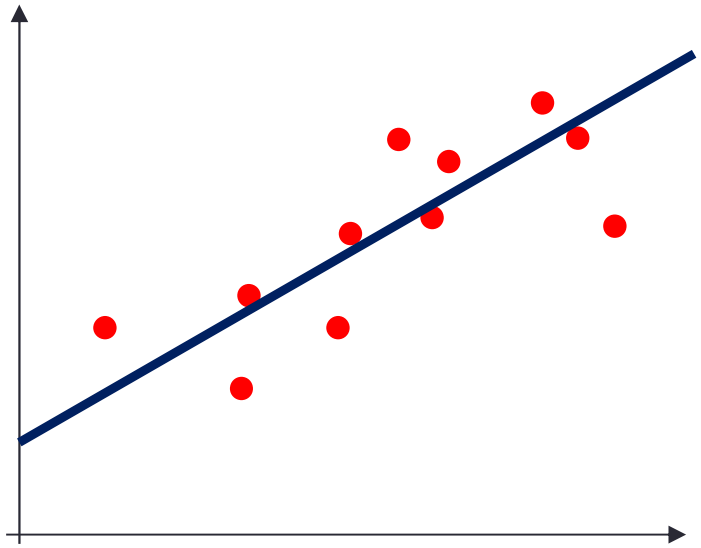


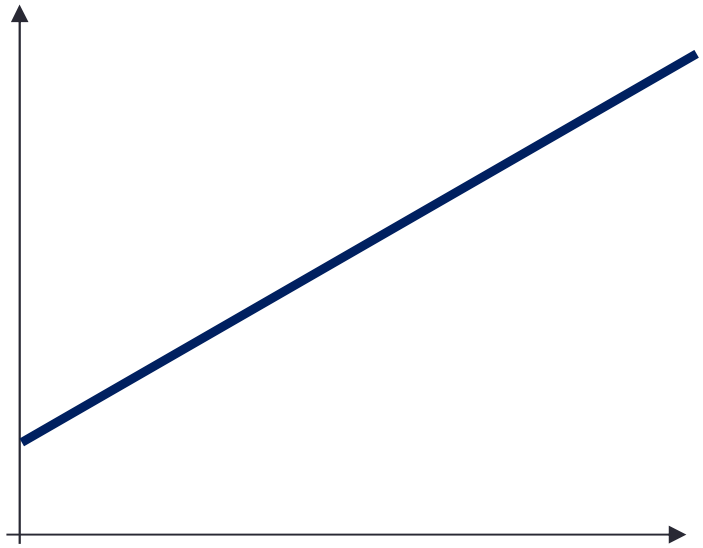


모델의 이용: 예측(Prediction)

- 새로운 데이터의 x값이 주어질때, 모델에서 학습된 패러미터를 이용하여 y값을 예측
- 경력이 4년, 6년, 8년인 직원의 연봉은 각각 얼마인가?

$$y = 30000 + 5000x$$

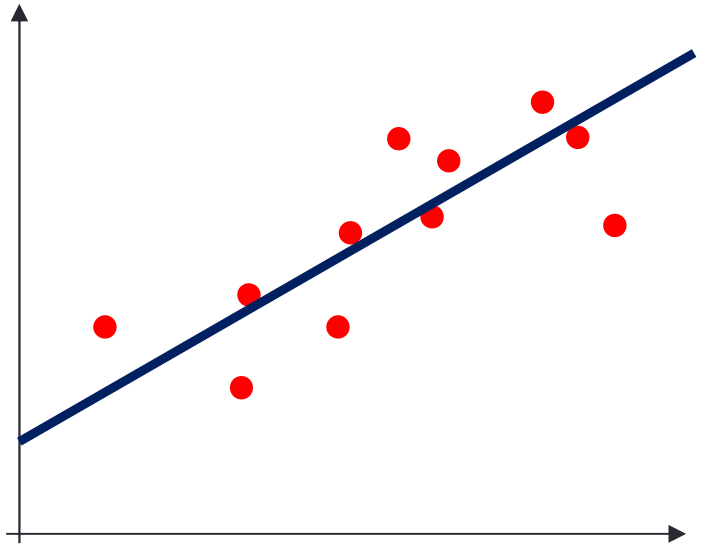




모델의 이용: 추론(Prediction)

- y값을 예측하는데 있어 X의 중요한 특성(feature)이 무엇인지, 그리고 그 방향성에 대하여 설명하는 것
- 데이터의 {생성과정을 설명}할 수 있는 명확한 모델이 있을때 가능
  - 모든 모델에서 추론이 가능하진 않음

$$y = 30000 + 5000x$$



추론을 통해
설명가능한 것

회사에 처음 입사시 연봉이 30,000$이다.

1년 지날때마다 5,000$씩 인상된다.

머신러닝 모형의 분류

시스템경영공학부
이지환 교수

머신러닝 모형의 분류

- 예측해야하는 값이 실수인가?
  - (Yes) 회귀(Regression)
  - (No) 분류(Classification)

- 정답을 알고있는 데이터인가?
  - (Yes) 지도학습(Supervised Learning)
  - (No) 비지도학습(Unsupervised Learning)

- 패러미터로 모델을 표현할 수 있는가?
  - (Yes) Parametric Method
  - (No) Non-parametric Method

머신러닝 모형의 분류

- 종속변수가 숫자형 값인가? 범주형 값인가?
  - (숫자형) 회귀(Regression)
  - (범주형) 분류(Classification)

- 정답을 알고있는 데이터인가?
  - (Yes) 지도학습(Supervised Learning)
  - (No) 비지도학습(Unsupervised Learning)

- 패러미터로 모델을 표현할 수 있는가?
  - (Yes) Parametric Method
  - (No) Non-parametric Method

변수의 종류

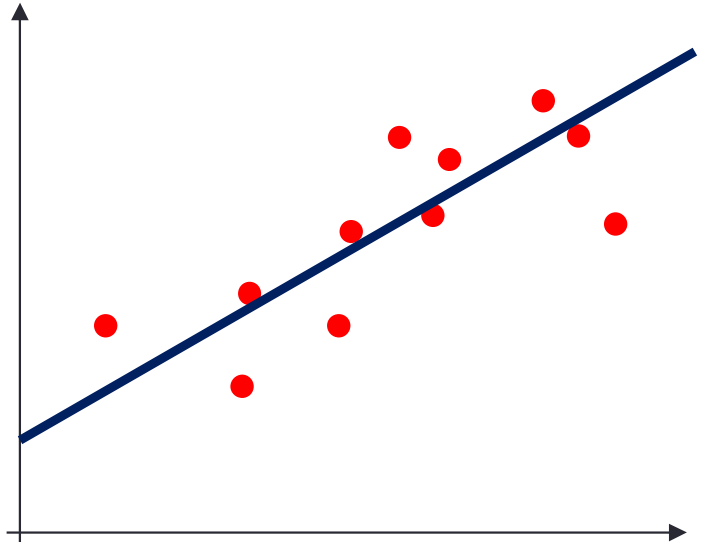
- 숫자형 변수 (Numerical Variable)
  - 실수나 정수로 표현할 수 있음
    - 예시) 나이: {56, 24,…}, 온도: {27.2,24.0,…},
  - 값의 차이가 의미 있음

- 범주형 변수 (Categorical Variable)
  - 실수로 표현할 수 없음
  - 분절된 서로 다른 값을 가짐
    - 예시) 성별: {남자, 여자} // 브랜드: {현대, 도요타, 기아, 포드} // 결혼여부: {yes, no}
  - (설령) 서로 다른 값에 숫자를 부여한다 해도 숫자간의 차이는 무의미

회귀(Regression)

- 종속변수가 숫자형인 경우
- 경력이 8년인 사람의 연봉은 얼마인가? → 실수값을 예측

	경력	연봉($)
1	0.3	40000
2	0.5	48000
3	5	70000
49	11	120000
50	6	65000



분류(Classification)

- 종속변수가 범주형(Categorical Variable)인 경우
  - 다음과 같은 50명의 기존 데이터가 있다

	나이	성별	직업	브랜드
1	24	남	군인	삼성
2	23	여	디자이너	애플
3	40	남	디자이너	애플
49	50	여	회사원	삼성
50	27	여	요리사	삼성

  - 나이, 성별, 직업이 주어진 어떤사람이 {삼성,애플} 두 핸드폰중 무엇을 고를지 예측하고 싶다.
    - 브랜드는 전형적인 범주형 변수로 분류문제에 해당됨

모델의 분류

- 예측해야하는 값이 실수인가?
  - (Yes) 회귀(Regression)
  - (No) 분류(Classification)

- 정답을 알고있는 데이터인가?
  - (Yes) 지도학습(Supervised Learning)
  - (No) 비지도학습(Unsupervised Learning)

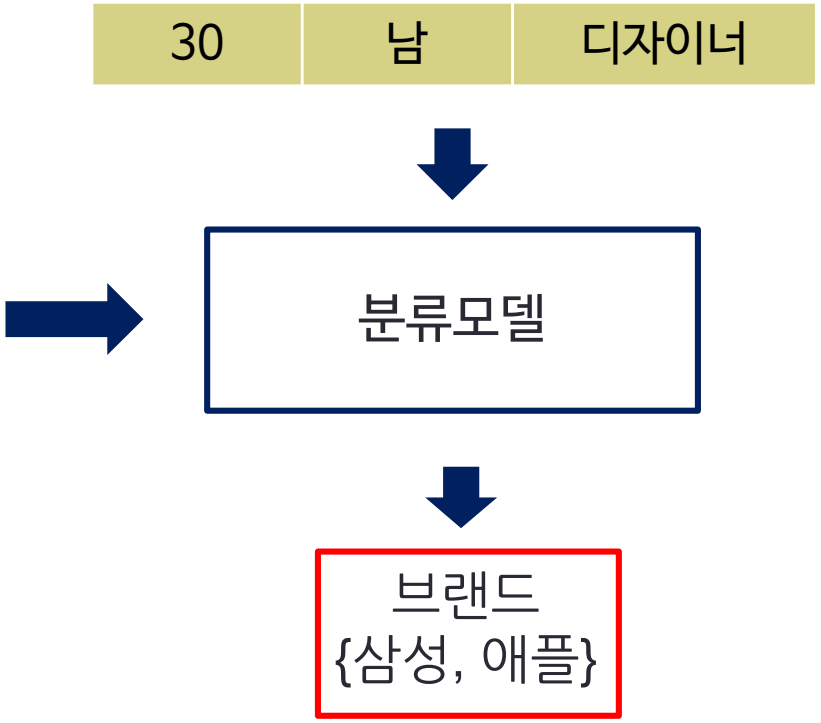
- 패러미터로 모델을 표현할 수 있는가?
  - (Yes) Parametric Method
  - (No) Non-parametric Method

지도학습 (Supervised Learning)

- 학습하는 데이터에 예측하고자 하는 값이 이미 존재하는 경우
- 예시) 나이, 성별, 직업을 통해 핸드폰 브랜드를 예측하고 싶다.
  - 데이터에 실제 사람들이 어떤 브랜드를 선택했는지 있다.

	나이	성별	직업	브랜드
1	24	남	군인	삼성
2	23	여	디자이너	애플
3	40	남	디자이너	애플
49	50	여	회사원	삼성
50	27	여	요리사	삼성

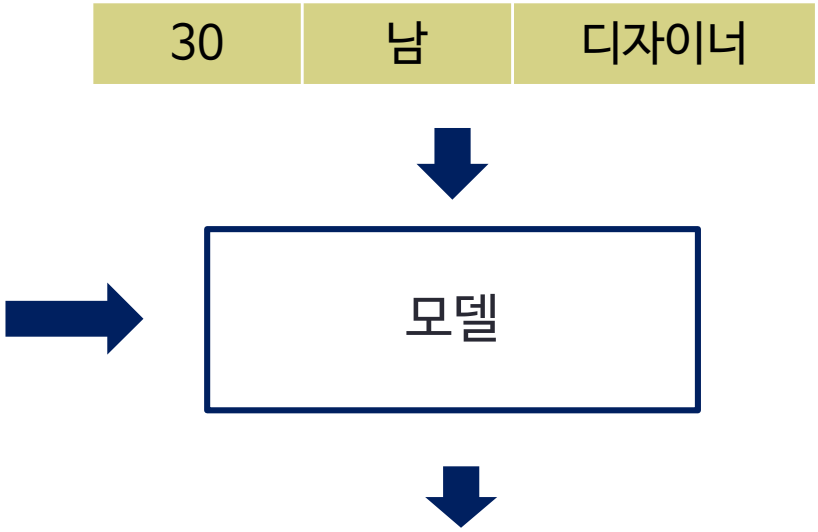
레이블



비지도학습 (Unsupervised Learning)

- 학습하는 데이터내에 예측하고자 하는 값이 없는 경우
- 예시) 나이, 성별, 직업의 유사도에 따라 3개의 그룹으로 나누고 싶다.

	나이	성별	직업
1	24	남	군인
2	23	여	디자이너
3	40	남	디자이너
49	50	여	회사원
50	27	여	요리사



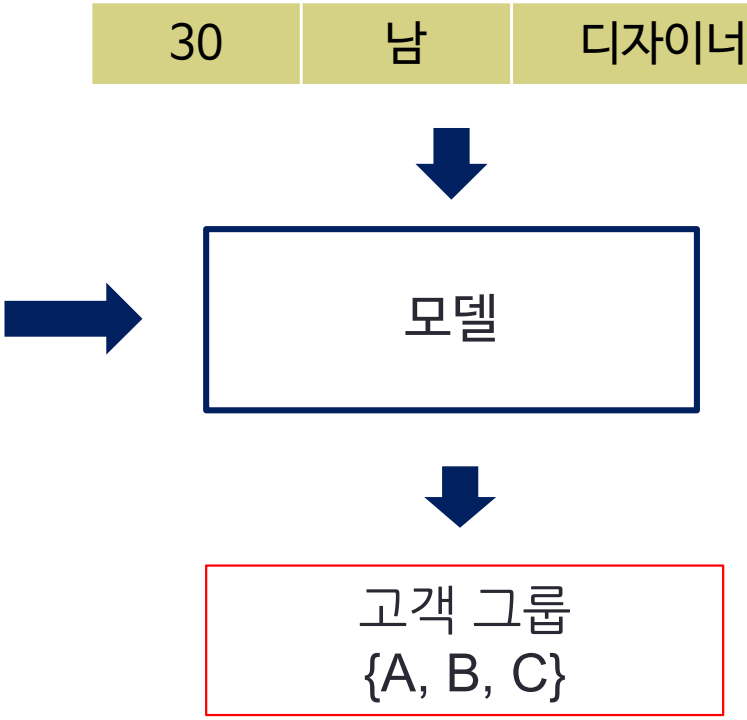
비지도학습 (Unsupervised Learning)

- 학습하는 데이터내에 예측하고자 하는 값이 없는 경우
- 예시) 나이, 성별, 직업의 유사도에 따라 3개의 그룹으로 나누고 싶다.
  - 데이터를 설명하는 내재적인 구조를 학습해야 함

- 군집분석(Clustering Analysis)

	나이	성별	직업	고객그룹
1	24	남	군인	A
2	23	여	디자이너	B
3	40	남	디자이너	A
49	50	여	회사원	C
50	27	여	요리사	A

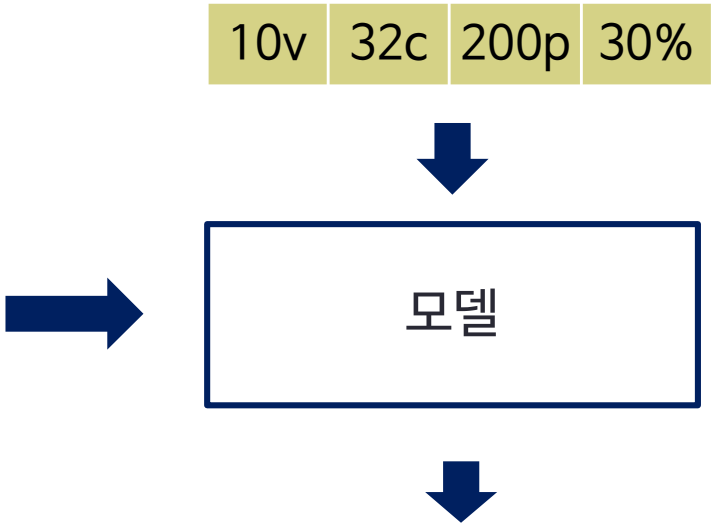
없던 변수



비지도학습 (Unsupervised Learning)

- 학습하는 데이터내에 예측하고자 하는 값이 없는 경우
- 예시) 시간에 따라 기록된 5,000개의 공정 데이터가 존재한다. 새로운 데이터가 기존의 데이터에서 얼마나 벗어나 있는지 측정하고자 한다.

	전압	온도	압력	습도
1	xx	xx	xx	xx
2	xx	xx	xx	xx
3	xx	xx	xx	xx
4999	xx	xx	xx	xx
5000	xx	xx	xx	xx

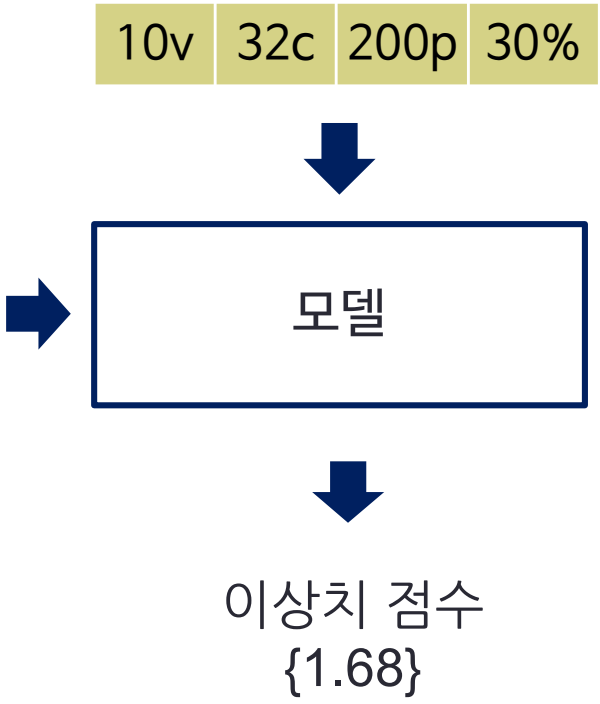


비지도학습 (Unsupervised Learning)

- 학습하는 데이터내에 예측하고자 하는 값이 없는 경우
- 예시) 시간에 따라 기록된 5,000개의 공정 데이터가 존재한다. 새로운 데이터가 기존의 데이터에서 얼마나 벗어나 있는지 측정하고자 한다.

- 이상치 탐지(anomaly detection)

	전압	온도	압력	습도	이상치 점수
1	xx	xx	xx	xx	0.05
2	xx	xx	xx	xx	0.04
3	xx	xx	xx	xx	0.2
4999	xx	xx	xx	xx	0.01
5000	xx	xx	xx	xx	0.2



지도학습 vs 비지도 학습

- 지도학습
  (Supervised Learning)
- 데이터: (X,y)
  - X 데이터
  - y 레이블(정답)
- 목표
  - X → y의 관계를 맵핑하는 함수를 찾는 것
- 예시
  - 회귀, 분류, CNN, 객체탐지, Sequence 모델

- 비지도학습
  (Unsupervised Learning)
- 데이터: X
  - X 데이터
  - y는 존재하지 않음
- 목표
  - 데이터를 설명하는 내재적인 구조를 학습
- 예시
  - 군집분석(Clustering)
  - 차원축소 (Dimensionality Reduction)
  - 이상치 탐지

모델의 분류

- 예측해야하는 값이 실수인가?
  - (Yes) 회귀(Regression)
  - (No) 분류(Classification)

- 정답을 알고있는 데이터인가?
  - (Yes) 지도학습(Supervised Learning)
  - (No) 비지도학습(Unsupervised Learning)

- 패러미터로 모델을 표현할 수 있는가?
  - (Yes) Parametric Method
  - (No) Non-parametric Method

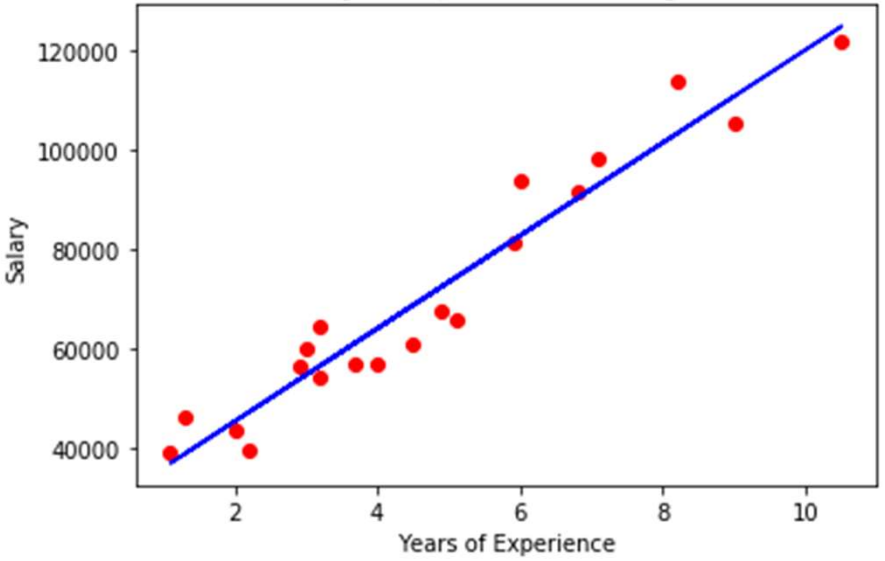
Parametric Model

- X와 y사이의 함수적 형태를 가정
- 독립변수와 종속변수의 관계를 고정된 수의 패러미터로 표현할 수 있음

$$y = \beta_0 + \beta_1 x$$

$$\beta_0, \beta_1$$

	경력	연봉($)
1	0.3	40000
2	0.5	48000
3	5	70000
49	11	120000
50	6	65000



Non-Parametric Model

- 패러미터의 수가 고정되지 않고 데이터에 따라 달라짐
- X와 y 사이의 함수적 형태를 가정하지 않음

- 예시) KNN Regression
  - (주변에서 가장 가까운 K개의 데이터를 y값을 평균내어 예측)