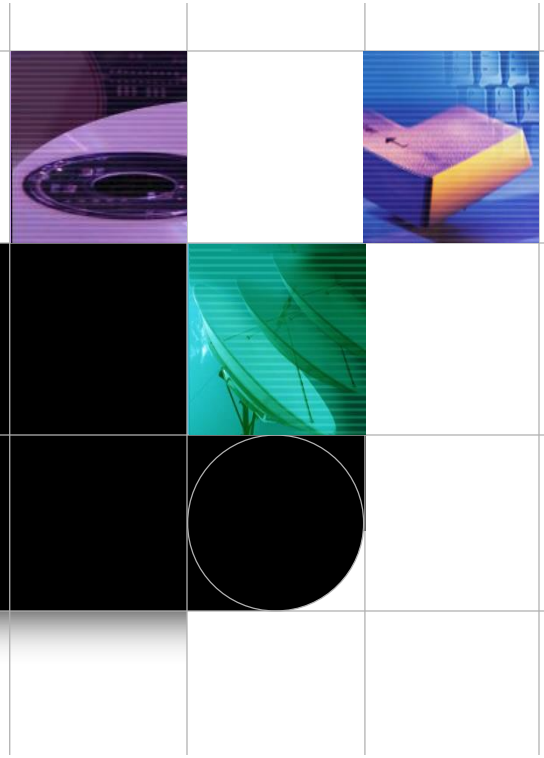
R을 활용한 머신러닝



1장

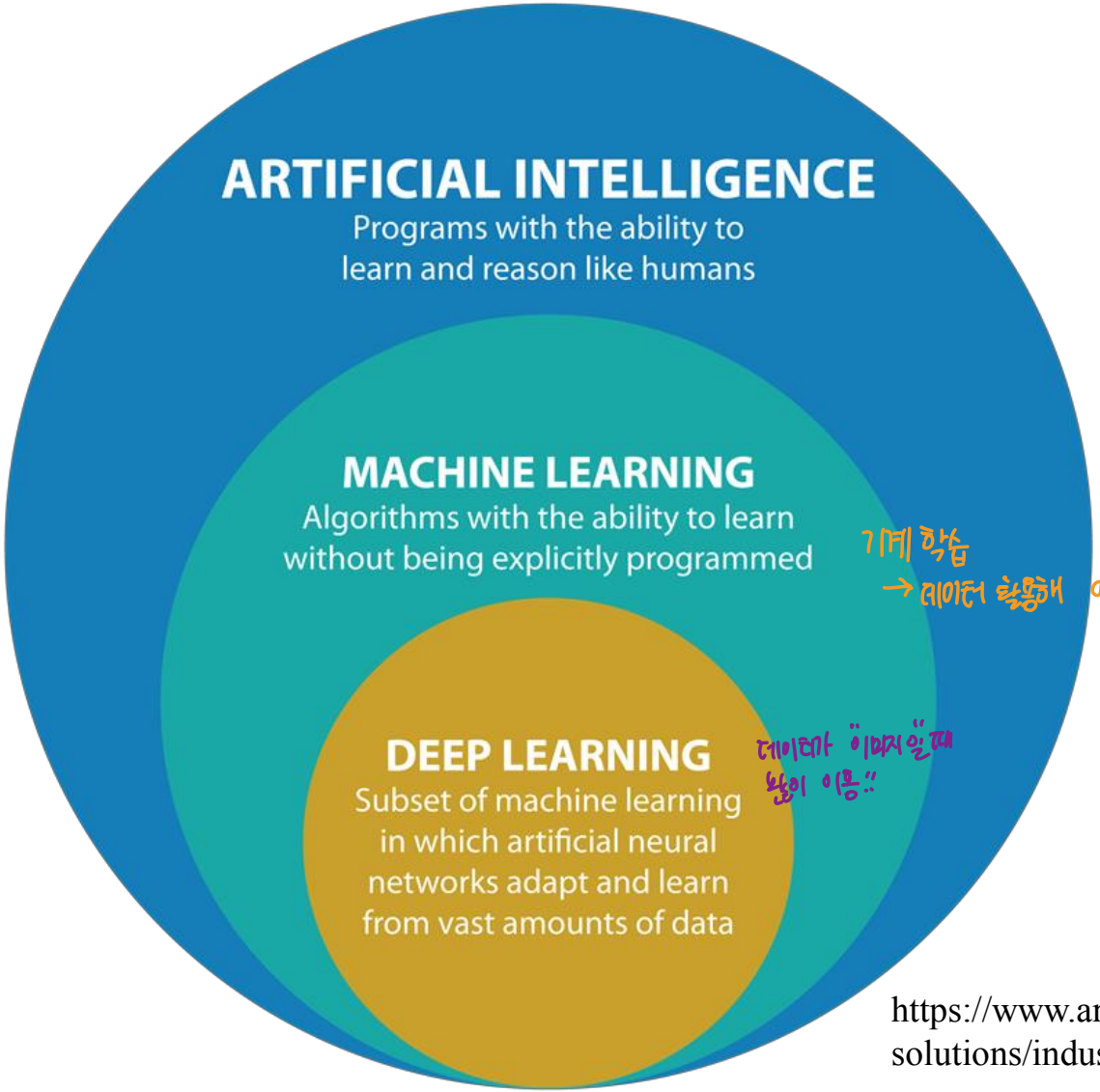
머신러닝 소개

Sejong Oh

Bio Information Technology Lab.

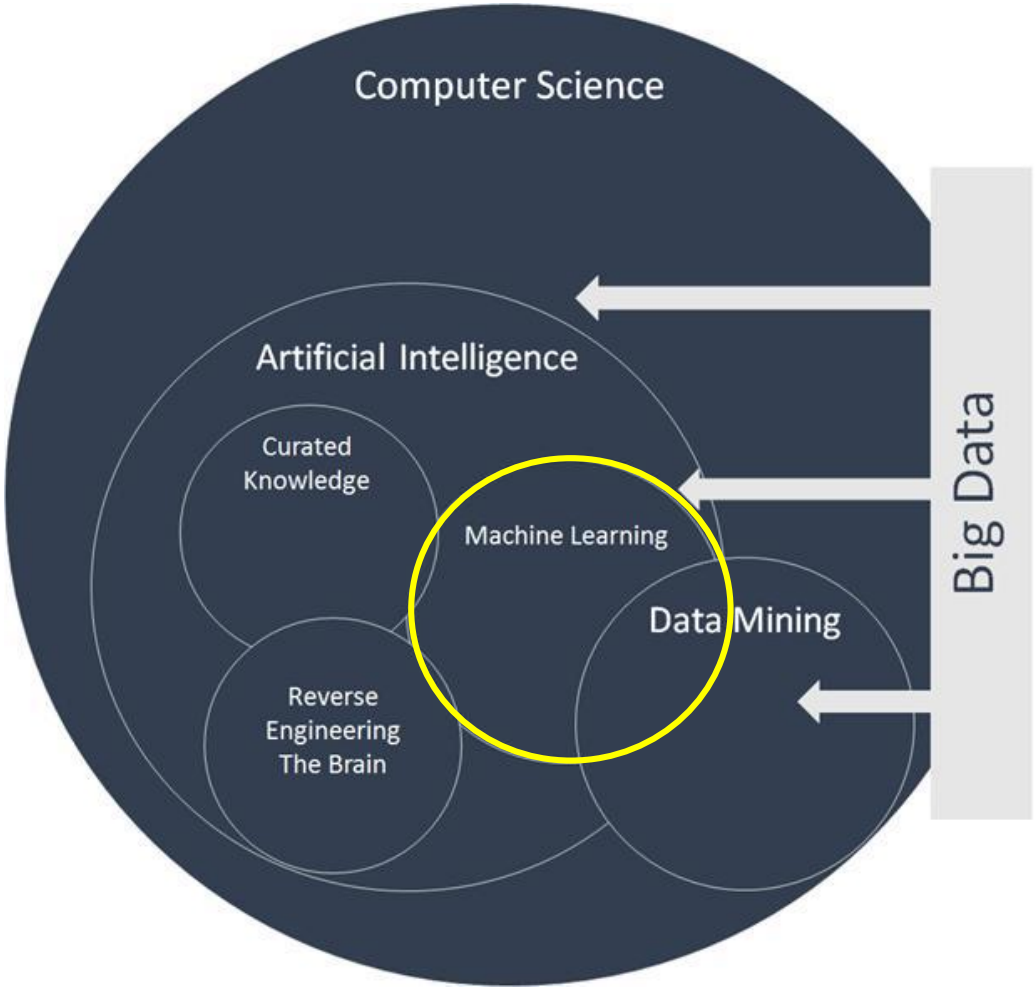
1. 개요

- AI, machine learning, deep learning



https://www.argility.com/argility-ecosystem-solutions/industry-4-0/machine-learning-deep-learning/

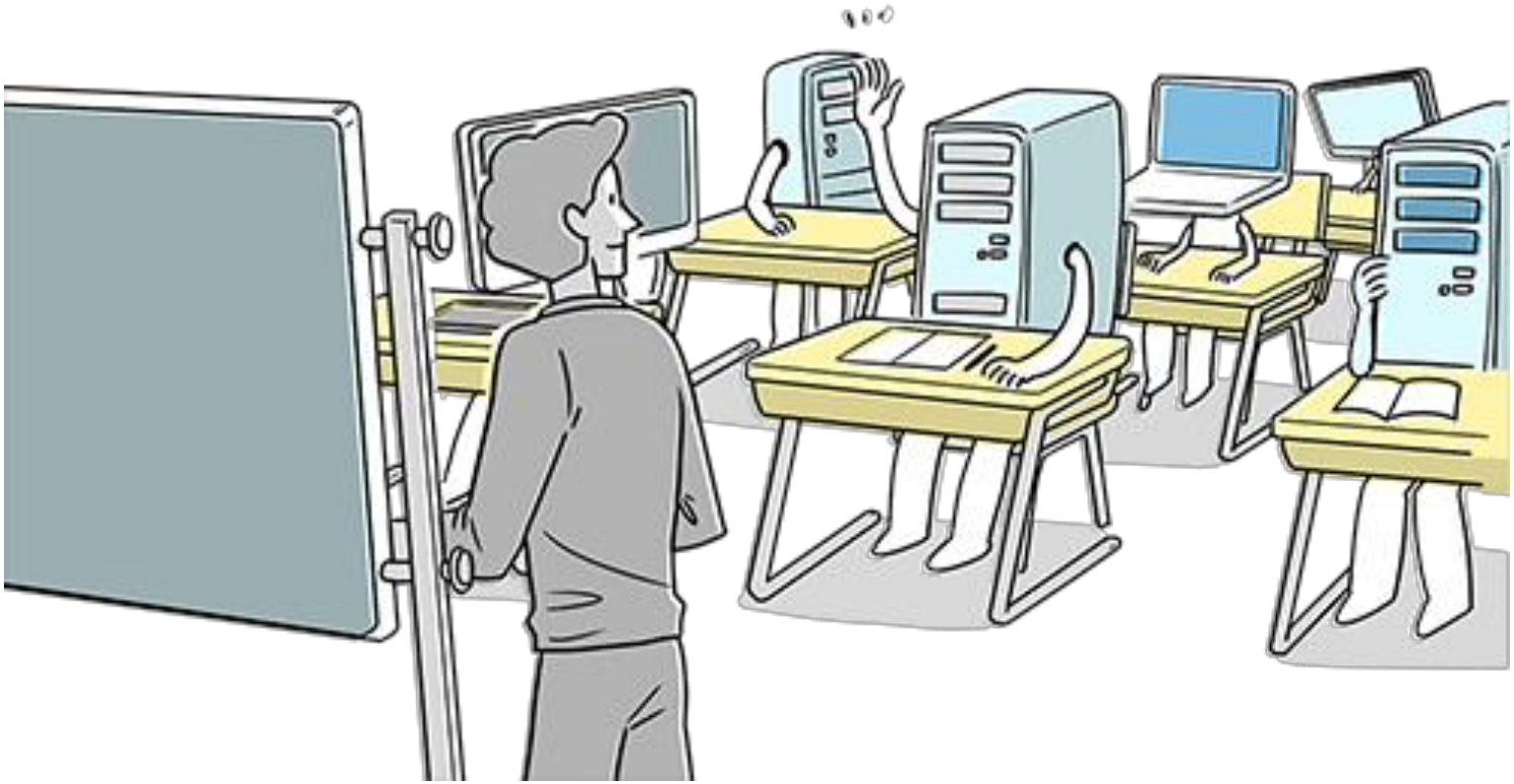
1. 개요



https://www.linkedin.com/pulse/20140916175039-113015482-how-the-buzz-words-fit-into-the-trading-world-ai-machine-learning-and-data-mining

2. Machine learning

- What is machine learning ?



https://www.lexalytics.com/technology/machine-learning

2. Machine learning

- 과거의 경험을 미래의 결정(예측)에 활용하는 소프트웨어를 디자인하고 연구하는 분야
  - 과거의 경험 → 데이터에 반영
  - 과거 데이터로 부터 숨겨진 규칙을 찾아내어 일반화. 이를 미래의 예측에 활용.
  - ex) 주가 예측

- 전통적 SW
  - 규칙을 인간이 알아내어 알고리즘의 형태로 SW 안에 구현함

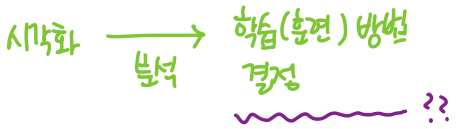
- 머신러닝
  - **규칙을 알아내는 방법**은 인간이 제시
  - 실제 규칙을 알아내는 과정은 머신(?)이 진행함.
  - 머신이 규칙을 알아내는 과정이 '학습(learning)'
    (인간 입장에서는 머신을 '훈련(training)' 시키는 과정)

2. Machine learning

- 머신러닝 분야의 예: 주가 예측
  - 전통적 방법
    - 주가 예측 공식을 인간이 개발하여 SW 로 구현
  - 머신러닝 방법
    - 1) 과거 데이터를 수집. 정리

주가에 영향을 미치는 요인들(X) 실제 주가 (Y)

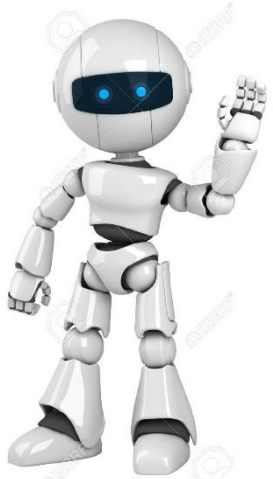
	Country	Salesperson	Order Date	OrderID	Units	Order Amount
2	USA	Fuller	1/01/2011	10392	13	1,440.00
3	UK	Gloucester	2/01/2011	10397	17	716.72
4	UK	Bromley	2/01/2011	10771	18	344.00
5	USA	Finchley	3/01/2011	10393	16	2,556.95
6	USA	Finchley	3/01/2011	10394	10	442.00



    - 2) 학습(훈련) 방법 결정 (regression, decision tree, deep neural network,..)
    - 3) 학습(훈련) 진행
    - 4) 예측모델 도출 (학습방법에 따라 다양한 형태)
    - 5) 주가 예측에 활용

2. Machine learning

- Machine ?



  - SW, Program
  - 학습의 주체가 사람이 아니라는 의미

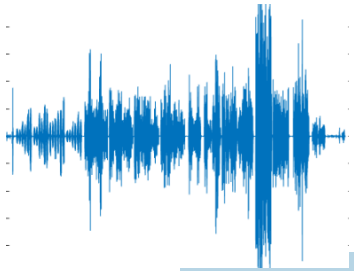
2. Machine learning

- 학습 자료 ?
  - Data

	A	B	C	D	E	F
1	Country	Salesperson	Order Date	OrderID	Units	Order Amount
2	USA	Fuller	1/01/2011	10392	13	1,440.00
3	UK	Gloucester	2/01/2011	10397	17	716.72
4	UK	Bromley	2/01/2011	10771	18	344.00
5	USA	Finchley	3/01/2011	10393	16	2,556.95
6	USA	Finchley	3/01/2011	10394	10	442.00
7	UK	Gillingham	3/01/2011	10395	9	2,122.92
8	USA	Finchley	6/01/2011	10396	7	1,903.80
9	USA	Callahan	8/01/2011	10399	17	1,765.60
10	USA	Fuller	8/01/2011	10404	7	1,591.25
11	USA	Fuller	9/01/2011	10398	11	2,505.60
12	USA	Coghill	9/01/2011	10403	18	855.01
13	USA	Finchley	10/01/2011	10401	7	3,868.60
14	USA	Callahan	10/01/2011	10402	11	2,713.50
15	UK	Rayleigh	13/01/2011	10406	15	1,830.78
16	USA	Callahan	14/01/2011	10408	10	1,622.40
17	USA	Farnham	14/01/2011	10409	19	319.20
18	USA	Farnham	15/01/2011	10410	16	802.00

https://www.myonlinetraininghub.com/excel-tabular-data-format

http://blog.ageha-inc.jp/2015/10/sns-data/

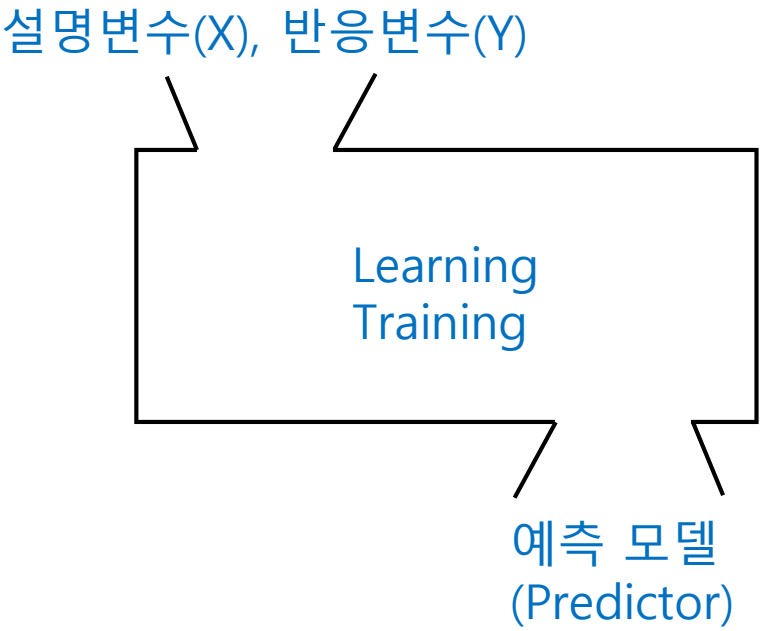


2. Machine learning

- Learning ?
  - 데이터: $(y_i, \boldsymbol{x}_i)$, $i=1,2,3,..,n$
    - 반응변수(response variable) : y_i
    - 설명변수(explanatory variable) : $\boldsymbol{x}_i = (x_{i1}, x_{i2},..,x_{ip})^T$

predictor

  - 반응변수(Y)와 설명변수(X) 간의 관계를 찾는 것 -> 훈련(training)



X : 키,몸무게, 허리둘레,..
Y : 고혈압여부

2. Machine learning

- Learning ?
  - 과거의 주식 변동 데이터를 학습하여 일주일 후의 주가를 예측
  - 건강검진 데이터를 학습하여 간암 발생률 추이를 예측
  - 과거의 대출 및 회수 데이터를 학습하여 대출 신청자가 대출금을 갚을지, 못갚을지를 예측
  - 키, 몸무게 등 정보로 부터 고혈압 여부를 예측
  - 과거 월드컵 경기 데이터를 학습하여 올해의 우승팀을 예측
  - 특정 기업의 10년후 생존 가능성 예측
  - 다양한 사진 정보를 학습하여 특정 사진속에서 사람이 몇 명 있는지 검사
  - 필기체 글씨 판독
  - 이미지 안에서 사람의 성별 구분
  - 음성 인식 (Seri, 빅스비, google)
  - 번역

2. Machine learning

- Learning 방법
  - 다양한 학습 알고리즘들이 존재함
  - KNN, SVM, regression, random forest, deep neural network, …

전통적 문제해결

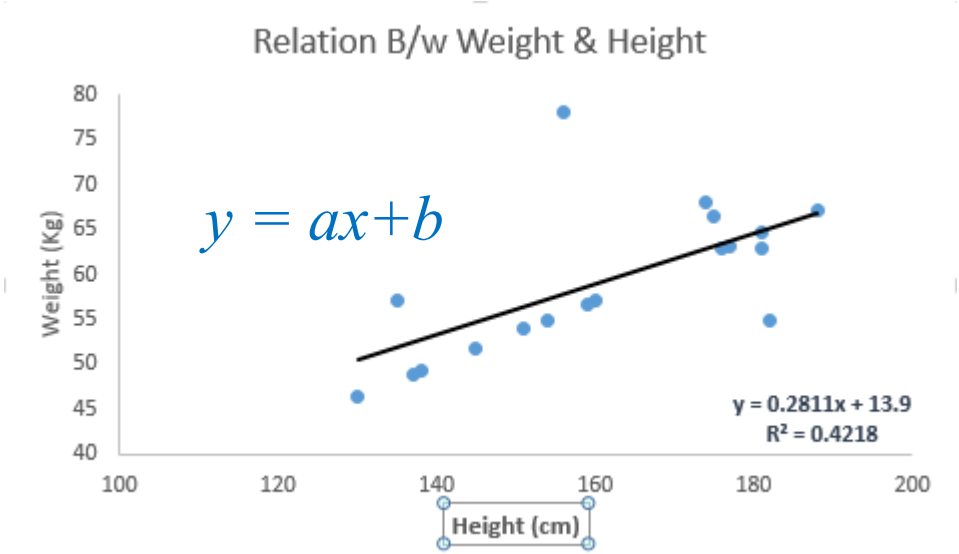
인간 분석자가 데이터를 연구하여 어떤 원리나 이론을 도출

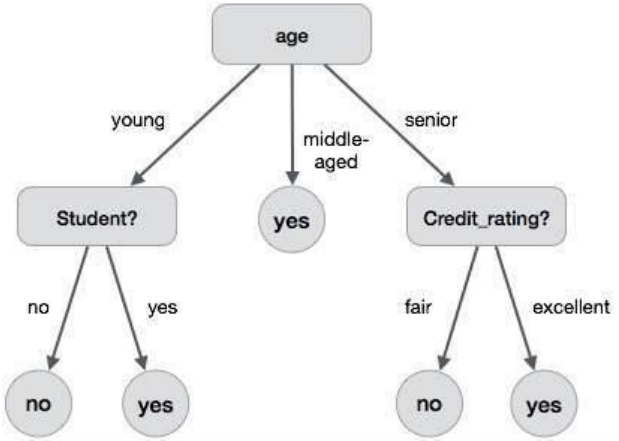
Machine learning

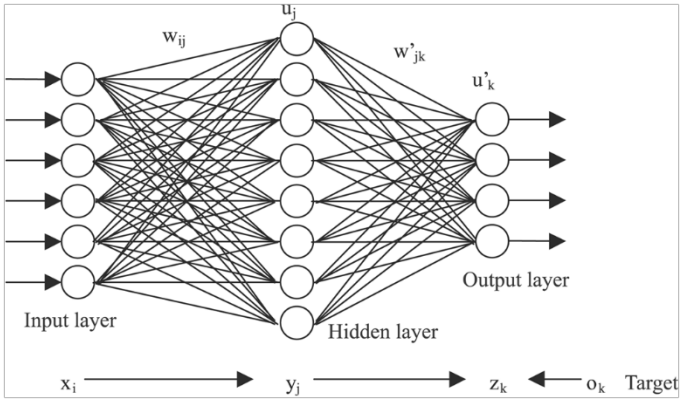
데이터와 학습 방법을 제시하고 프로그램 스스로 원리나 이론을 도출하도록 함

2. Machine learning

- Learning 의 결과는
  - (learning) model
  - 어떤 방법으로 학습을 시켰는가에 따라 model 의 형태는 다양함

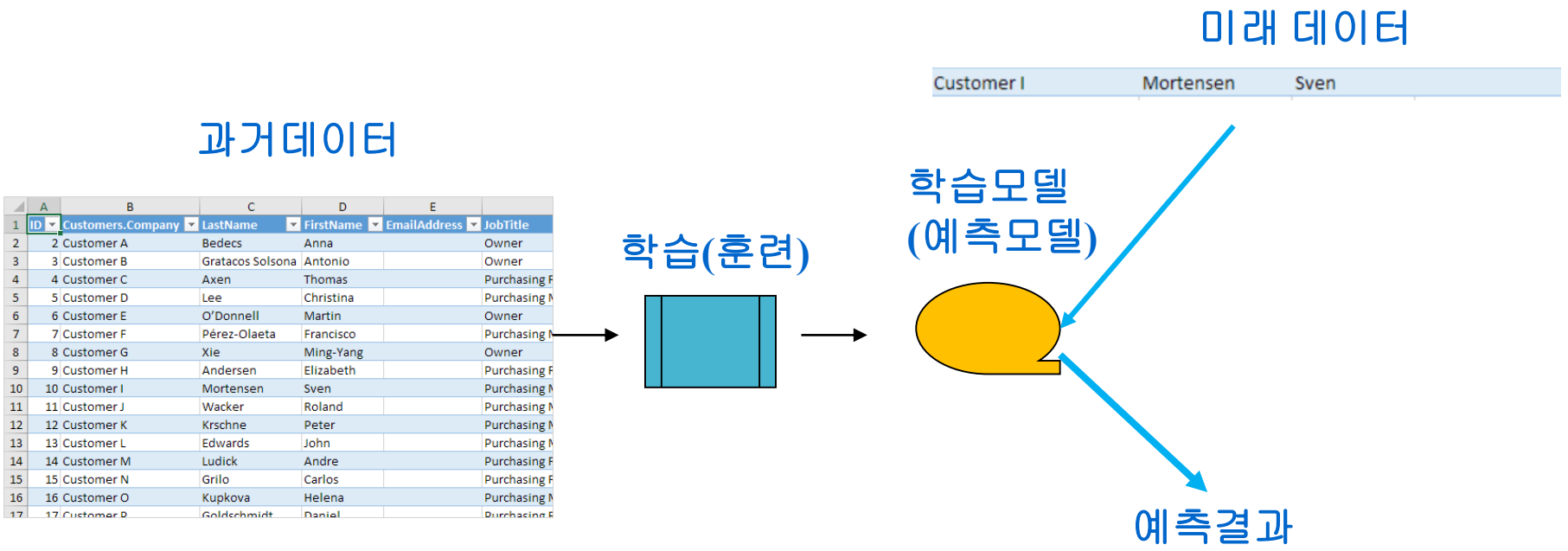






2. Machine learning

- 정리: Machine learning 은
  - 과거의 축적된 데이터를 학습하여 미래를 예측하는 기술
  - 주가 예측, 질병진단, 스팸 필터링, 이미지 분류, 번역, …
  - 얼마나 정확한 모델을 만드느냐가 관건
  - 학습 데이터가 많을 수록 유리



	A	B	C	D	E	
1	ID	Customers.Company	LastName	FirstName	EmailAddress	JobTitle
2	2	Customer A	Bedecs	Anna		Owner
3	3	Customer B	Gratacos Solsona	Antonio		Owner
4	4	Customer C	Axen	Thomas		Purchasing F
5	5	Customer D	Lee	Christina		Purchasing M
6	6	Customer E	O'Donnell	Martin		Owner
7	7	Customer F	Pérez-Olaeta	Francisco		Purchasing M
8	8	Customer G	Xie	Ming-Yang		Owner
9	9	Customer H	Andersen	Elizabeth		Purchasing F
10	10	Customer I	Mortensen	Sven		Purchasing M
11	11	Customer J	Wacker	Roland		Purchasing M
12	12	Customer K	Krschne	Peter		Purchasing M
13	13	Customer L	Edwards	John		Purchasing M
14	14	Customer M	Ludick	Andre		Purchasing F
15	15	Customer N	Grilo	Carlos		Purchasing F
16	16	Customer O	Kupkova	Helena		Purchasing M
17	17	Customer P	Goldschmidt	Daniel		Purchasing F

2. Machine learning

- 정리: Machine learning 의 목표
  - 주어진 자료를 가장 잘 설명하는 모형을 찾는 것이 최종 목표가 아님
  - 새로운 설명변수의 값이 주어졌 때, 정확한 예측값을 주는 모형을 찾는 것이 목적 (과거 현상을 잘 설명하기 보다는 미래의 자료를 잘 예측할 수 있어야 함)

3. Machine learning areas

- Machine learning 분류
  - 지도학습 (supervised learning)
    - 회귀(regression)
    - 분류(classification) 등

설명변수(X), 반응변수(Y) 존재

Y 가 수치형 (주가, 기온,..)

Y 가 범주형 (정상인/환자, 남/녀, ..)

  - 비지도학습(unsupervised learning)
    - 군집화(clustering)

설명변수(X)만 존재

*'반응 변수'가 없어서 알아서 학습 결과 보고

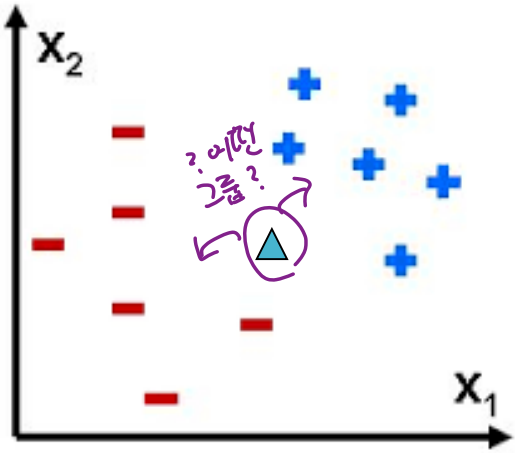
  - 강화학습(Reinforcement learning)

ex) 알파고, 게임

```
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

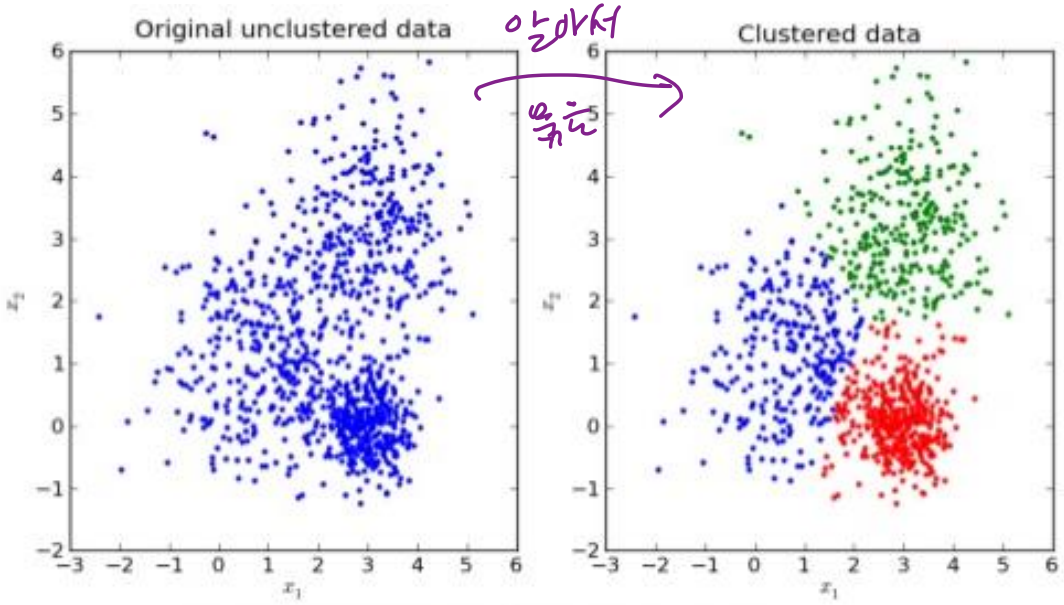
* Deep learning 은 지도학습 방법에 해당

3. Machine learning areas



classification

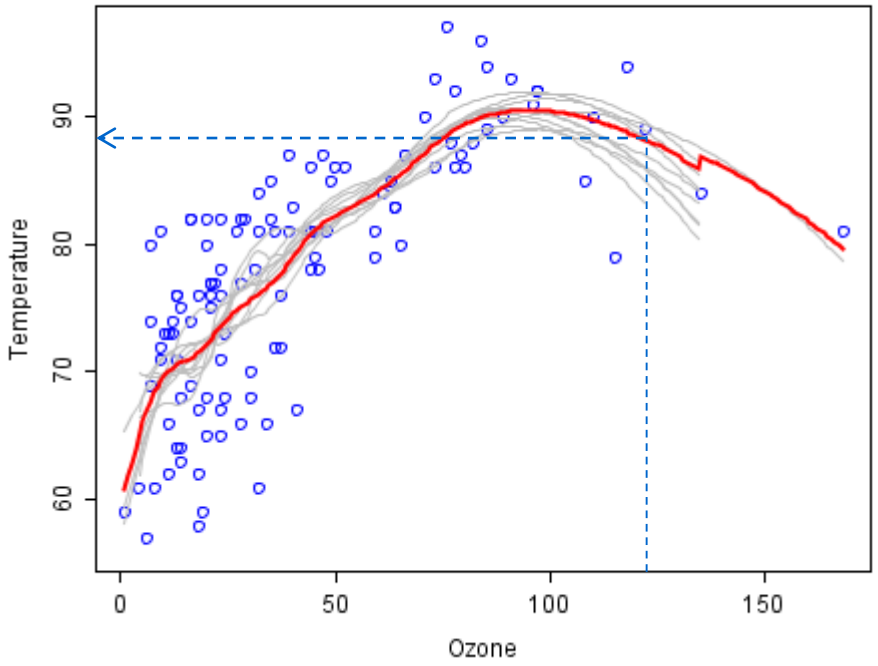
질병진단
문자인식
이미지분류



clustering

고객 세분화
비정상거래 탐지

3. Machine learning areas



regression

주가예측
오존농도에 따른 기온예측

용어정리

설명변수 (feature)

종속변수 (label)

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

- 특징(feature) / 레이블(label)
- 관측값(observation)
- 변수(variable)

- 학습알고리즘(learning algorithm)
- 모형, 모델(model)

- 회귀/분류/군집

용어정리

- 혼동행렬(confusion matrix)