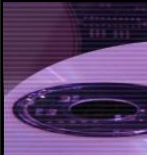
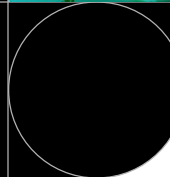
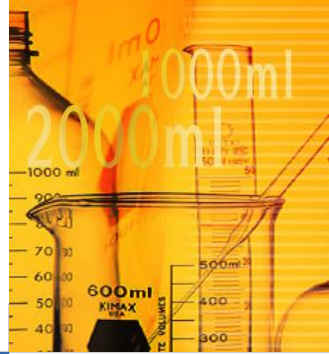


Chapter 3

Machine learning Concept

Sejong Oh

Bio Information Technology Lab.

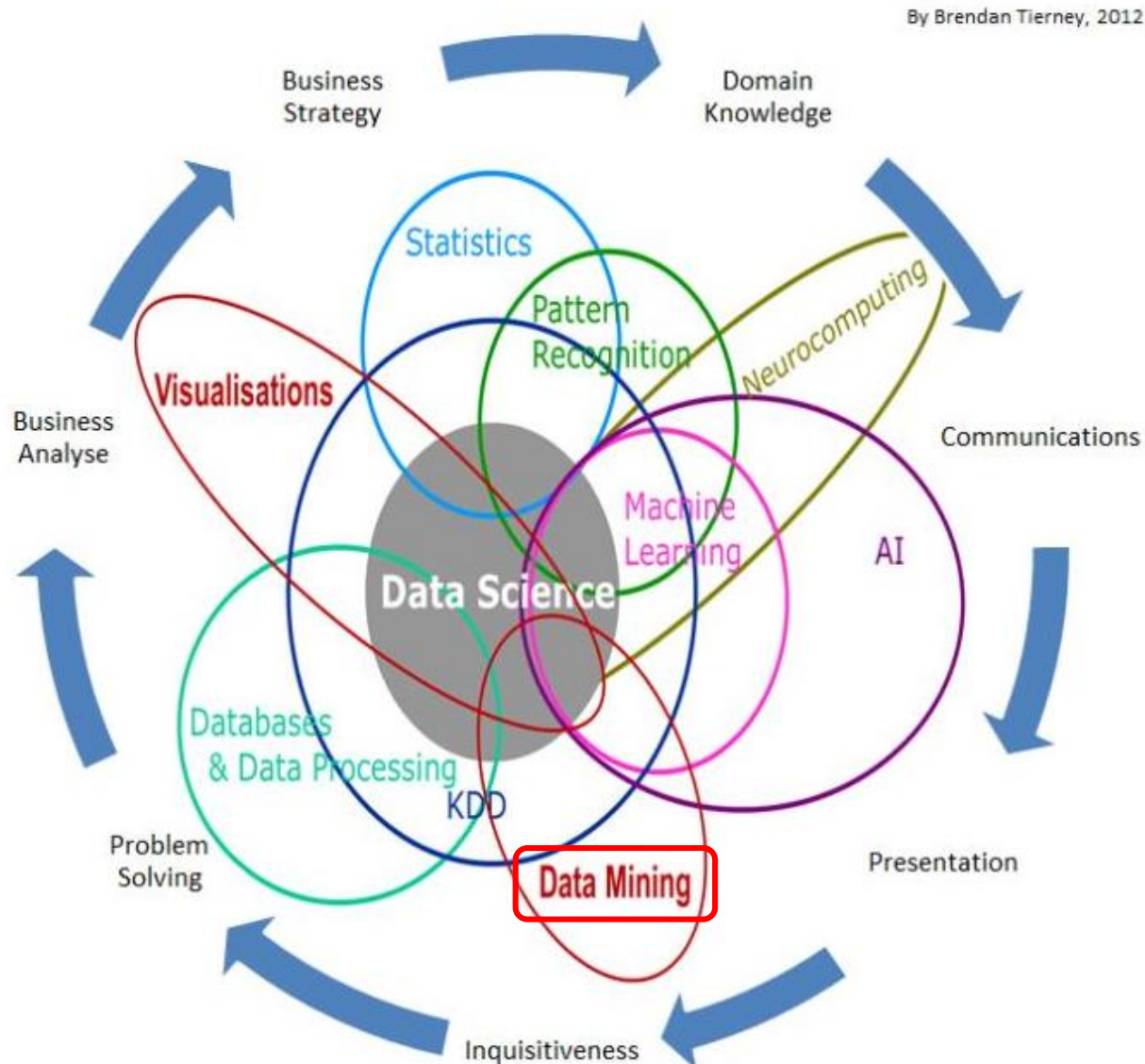


Contents



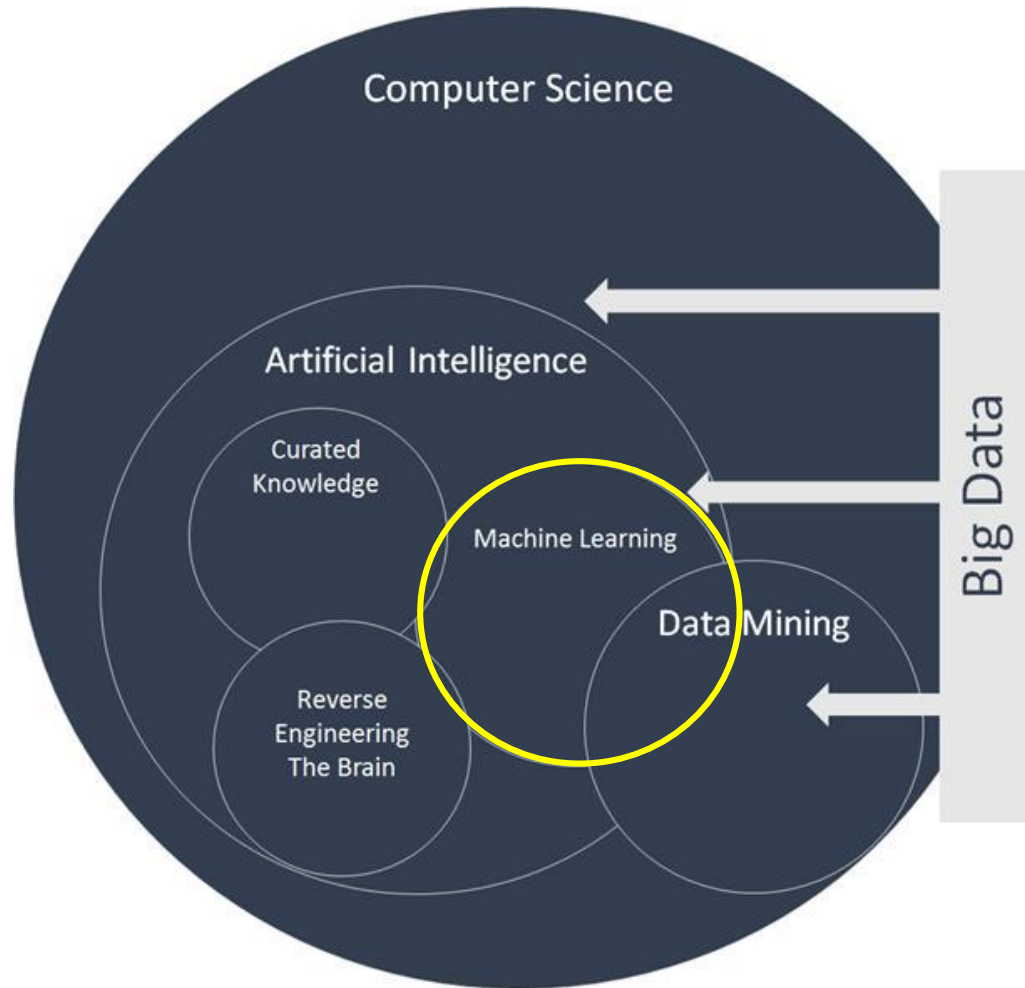
- AI scopes
- Machine learning
- Machine learning areas
- Development process of learning model
- Scikit-learn

1. AI scopes



<http://www.kdnuggets.com/2014/06/data-science-skills-business-problems.html>

1. AI scopes



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

1. AI scopes

Artificial Intelligence

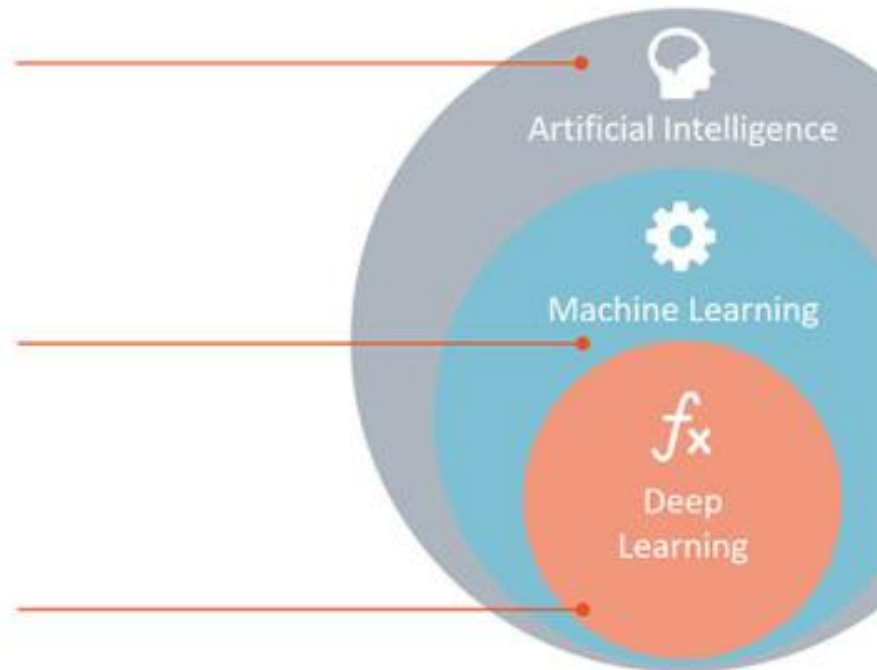
Any technique which enables computers to mimic human behavior.

Machine Learning

Subset of AI techniques which use statistical methods to enable machines to improve with experiences.

Deep Learning

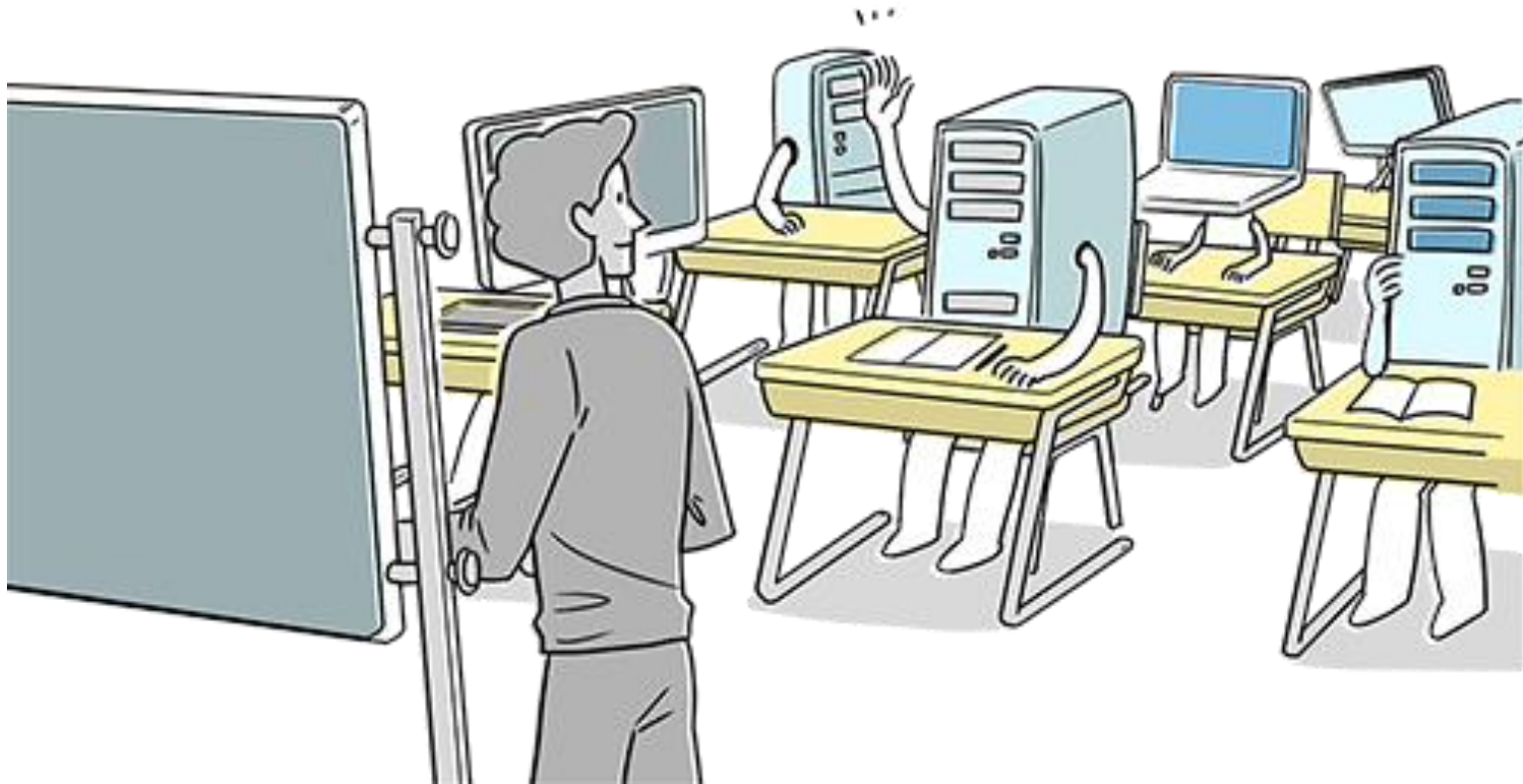
Subset of ML which make the computation of multi-layer neural networks feasible.



<https://rapidminer.com/blog/artificial-intelligence-machine-learning-deep-learning/>

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
 - 1959년 Arthur Samuel 의 논문에 처음 등장 (Some Studies in Machine Learning Using the Game of Checkers)
 - "컴퓨터가 명시적으로 프로그램되지 않고도 학습할 수 있도록 하는 연구 분야 "
 - 과거의 경험 → 데이터에 반영
 - 과거 데이터로 부터 숨겨진 규칙을 찾아내어 일반화. 이를 미래의 예측에 활용.
 - ex) 주가 예측
- 전통적 SW 개발
 - 규칙을 인간이 알아내어 알고리즘의 형태로 SW 안에 구현함
- 머신 러닝
 - 규칙을 알아내는 방법은 인간이 제시
 - 실제 규칙을 알아내는 과정은 머신(?)이 진행함.
 - 머신이 규칙을 알아내는 과정이 '학습(learning)'
(인간 입장에서는 머신을 '훈련(training)' 시키는 과정)

허리가
아프면
비가온다

2. Machine learning

- 머신러닝 분야의 예: 주가 예측


- 전통적 방법

- 주가 예측 공식을 인간이 개발하여 SW 로 구현

- 머신러닝 방법

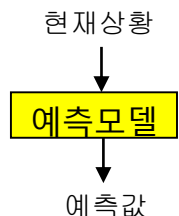
- 1) 과거 데이터를 수집. 정리

주가에 영향을 미치는 요인들(X) 실제 주가 (y)



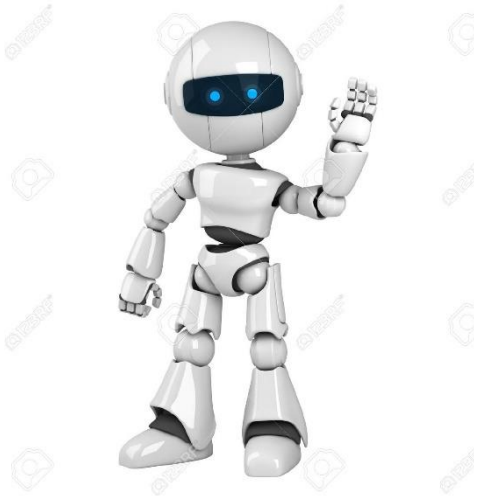
	Country	Salesperson	Order Date	OrderID	Units	Order Amount
1	USA	Fuller	1/01/2011	10392	13	1,440.00
2	UK	Gloucester	2/01/2011	10397	17	716.72
3	UK	Bromley	2/01/2011	10771	18	344.00
4	USA	Finchley	3/01/2011	10393	16	2,556.95
5	USA	Finchley	3/01/2011	10394	10	442.00

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



2. Machine learning

- Machine ?



- SW, Program
- 학습의 주체가 사람이 아니라는 의미

2. Machine learning

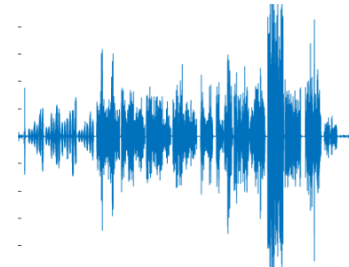
- 학습 자료 ?
 - Data

	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, \mathbf{x}_i) , $i=1,2,3,\dots,n$

- 반응변수(response variable) : y_i

- 설명변수(explanatory variable) : $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})$

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

설명변수(X), 반응변수(y)



X : 키, 몸무게, 허리둘레, ...
Y : 고혈압 여부

2. Machine learning

- Learning ?

$$y = f(x)$$

- 지금까지는 $f()$ 와 x 를 알 때 y 를 구하는 일을 함
- 머신러닝에서는 x, y 를 알 때 $f()$ 를 알아내고자 함

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 10년 후 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

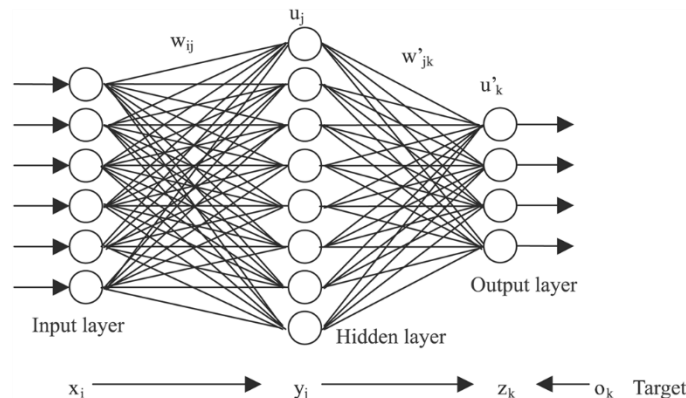
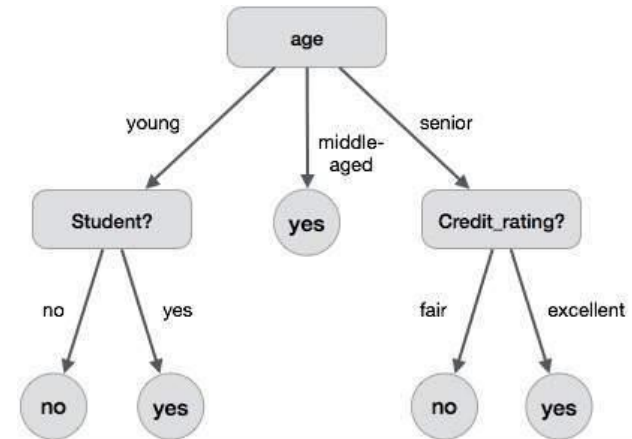
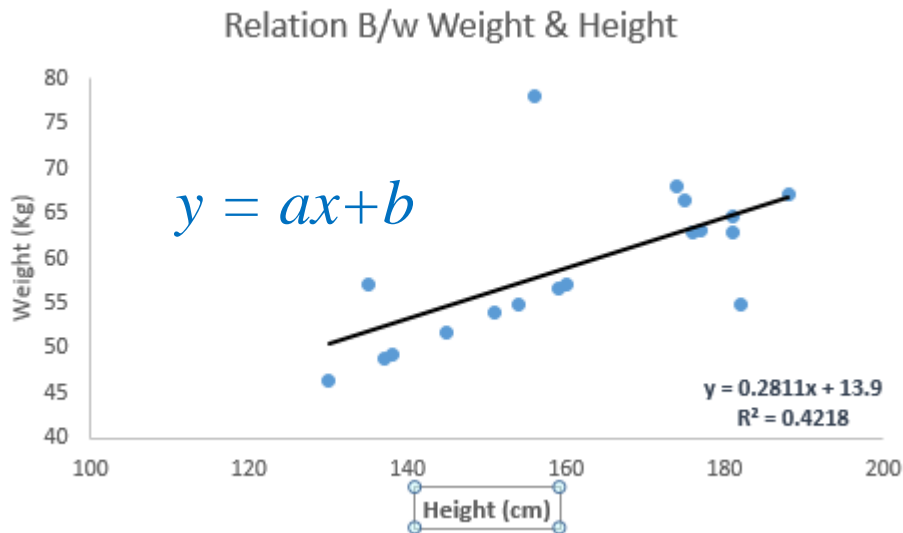
인간 분석자가 데이터를 연구하여
어떤 원리나 이론을 도출

Machine learning

데이터와 학습 방법을 제시하고
프로그램 스스로 원리나 이론을
도출하도록 함

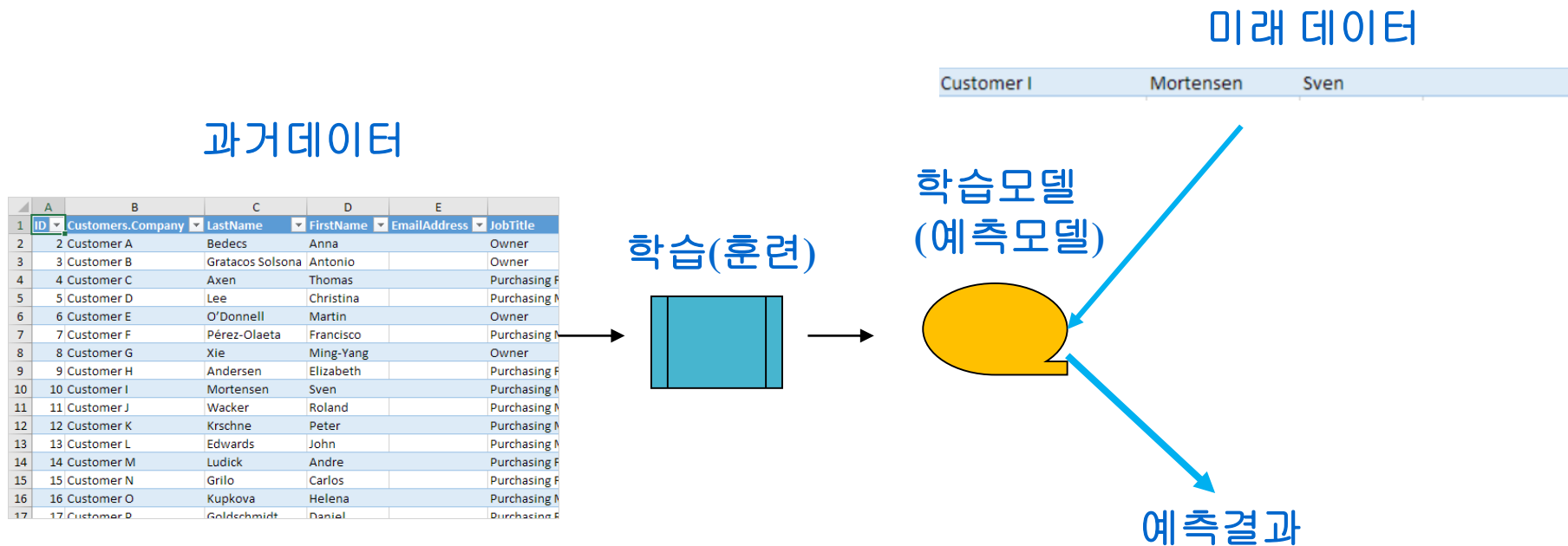
2. Machine learning

- 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



2. Machine learning

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

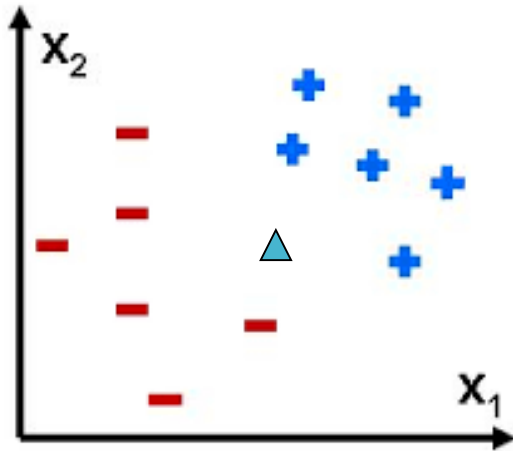
3. Machine learning areas

- Machine learning 분류

- 지도학습 (supervised learning) 설명변수(X), 반응변수(y) 존재
 - 회귀(regression) y 가 수치형 (주가, 기온,..)
 - 분류(classification) 등 y 가 범주형 (정상인/환자, 남/녀, ..)
- 비지도학습(unsupervised learning) 설명변수(X)만 존재
 - 군집화(clustering)
- 강화학습(Reinforcement learning)

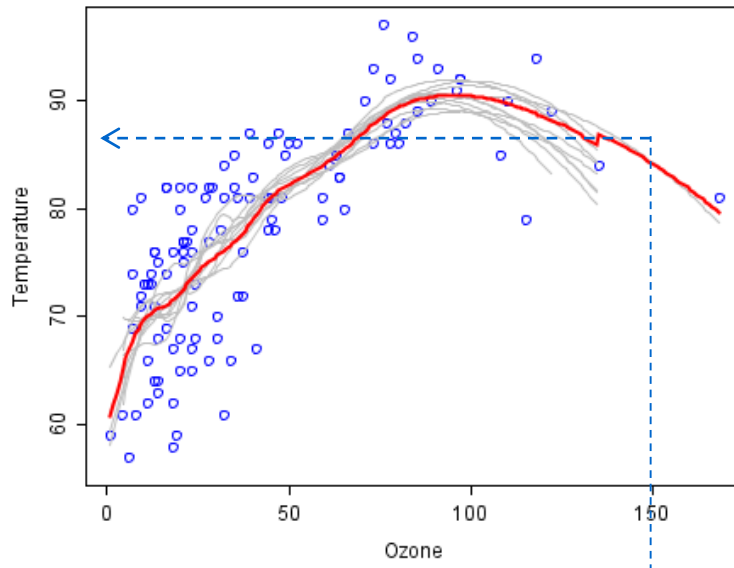
* Deep learning 은 지도학습 방법에 해당

3. Machine learning areas



classification

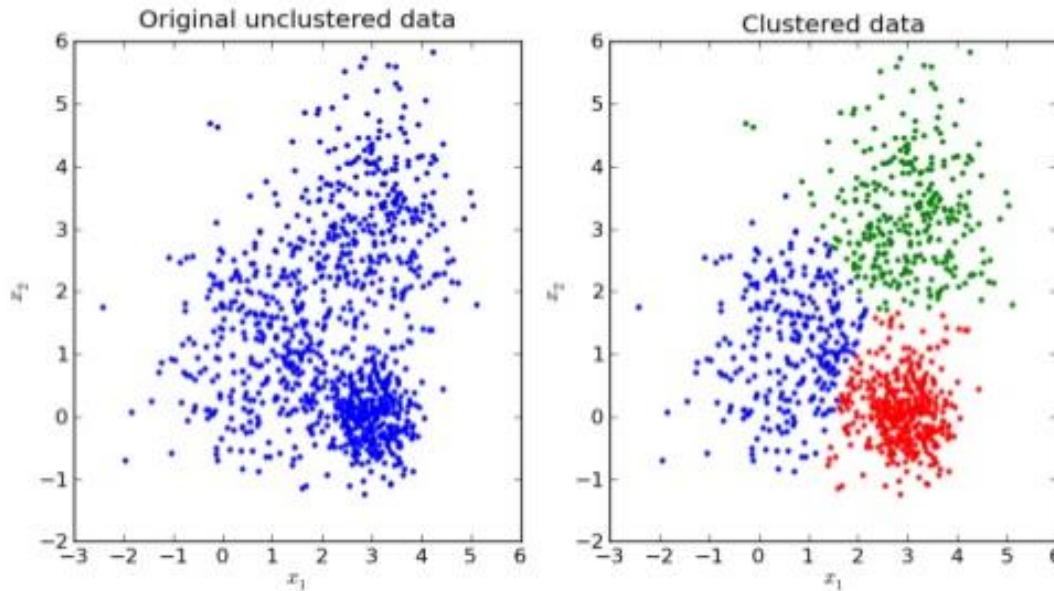
질병진단
문자인식
이미지분류



regression

주가예측
오존농도에 따른 기온예측

3. Machine learning areas



clustering

고객 세분화
비정상거래 탐지

3. Machine learning areas

- Reinforcement learning

- 행동심리학에서 영감을 받았으며, 어떤 환경 안에서 정의된 에이전트가 현재의 상태를 인식하여, 선택 가능한 행동들 중 보상을 최대화하는 행동 혹은 행동 순서를 선택하는 방법 (장기, 바둑,..)
- <https://www.youtube.com/watch?v=SH3bADiB7uQ>



Machine Learning Is No Longer Just for Experts

by Josh Schwartz

OCTOBER 26, 2016

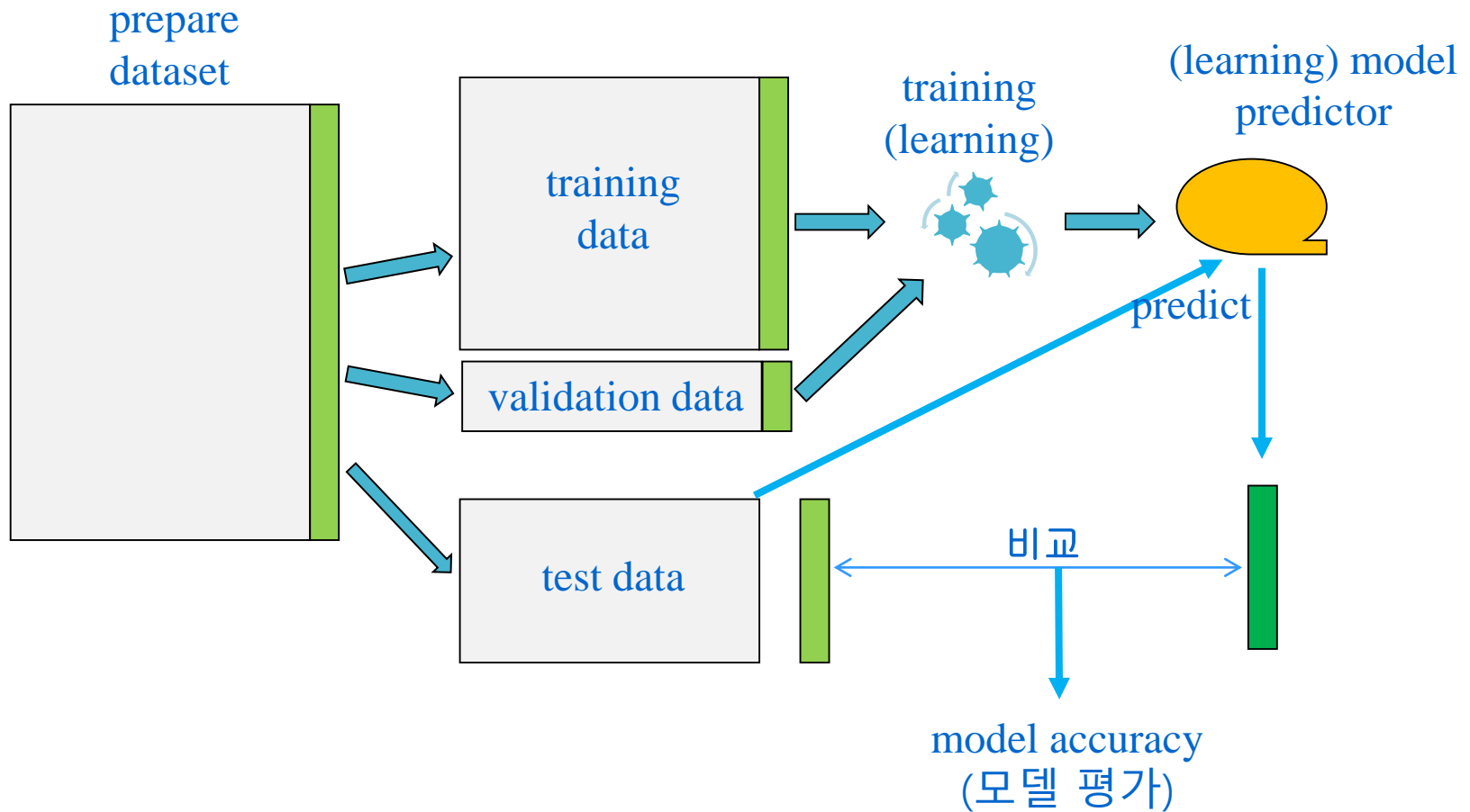
 SAVE  SHARE  COMMENT 4  TEXT SIZE  PRINT \$8.95 BUY COPIES



<https://hbr.org/2016/10/machine-learning-is-no-longer-just-for-experts>

4. 학습모델 개발 과정

- Classification, regression



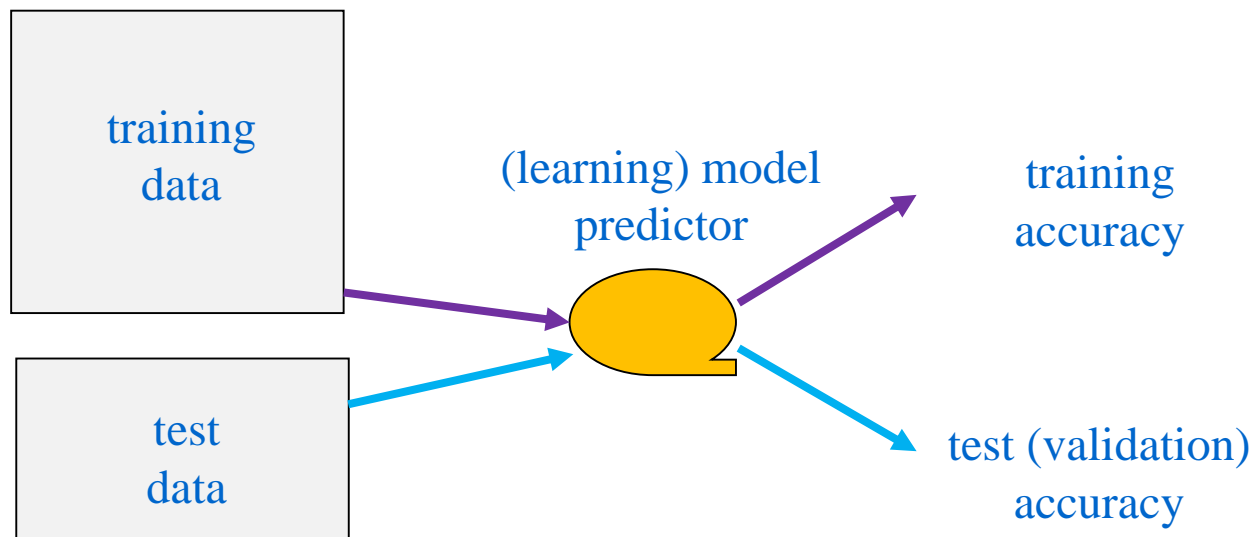
4. 학습모델 개발 과정

- Training data
 - 과거 데이터의 역할
- Validation data
 - 학습(훈련)과정에서 만들어지는 모델을 평가하는데 사용
 - 더 나은 모델을 만드는데 기여
 - 학습방법에 따라 필요치 않은 경우도 있음
- Test data
 - 미래 데이터의 역할
 - 미래 데이터는 없으므로 학습에 사용하지 않은 일부 데이터를 미래의 데이터로 간주
 - 미래 예측시 모델이 어느 정도의 성능을 보일지를 판단하는 자료

** Train : 50~75%. Test : 10~30%, validation: 나머지

4. 학습모델 개발 과정

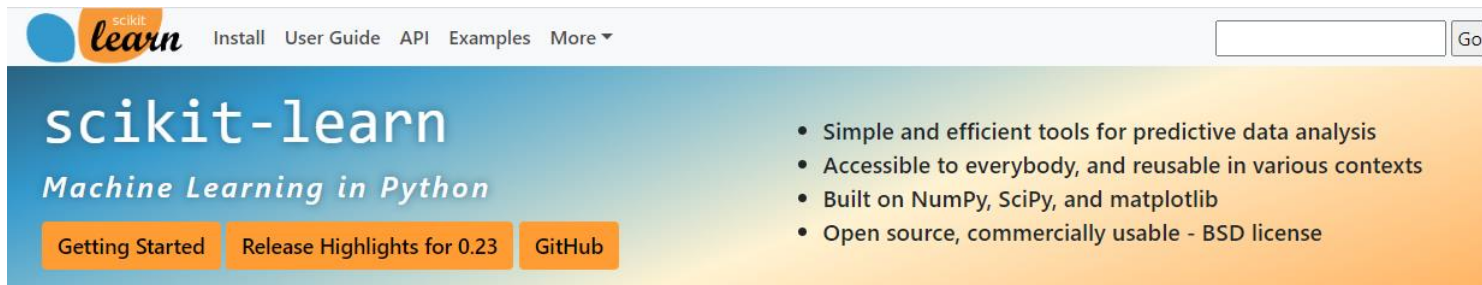
- Training accuracy vs Test accuracy




- Training accuracy : 모델이 과거의 데이터를 얼마나 잘 설명할 수 있는지를 보여줌.
- Test accuracy : 모델이 미래의 데이터를 얼마나 잘 예측할 수 있는지를 보여줌
- 일반적으로 Training accuracy > Test accuracy

5. scikit-learn

- Machine learning library in python
- <https://scikit-learn.org/stable/>
- Open source
- Published in 2007
- 분류, 회귀, 차원축소 등 머신러닝 관련 알고리즘들을 구현
- 데이터 전처리, hyper-parameter tuning, 모델 평가 기능도 제공
- Install scikit-learn
 - Python console 에서 `install scikit-learn`



5. scikit-learn

 [Install](#) [Use](#)

[Prev](#) [Up](#) [Next](#)

scikit-learn 1.1.2
[Other versions](#)

Please [cite us](#) if you use the software.

API Reference
sklearn.base: Base classes and utility functions
sklearn.calibration: Probability Calibration
sklearn.cluster: Clustering
sklearn.compose: Composite Estimators
sklearn.covariance: Covariance Estimators
sklearn.cross_decomposition: Cross decomposition
sklearn.datasets: Datasets
sklearn.decomposition: Matrix Decomposition
sklearn.discriminant_analysis: Discriminant Analysis
sklearn.dummy: Dummy estimators
sklearn.ensemble: Ensemble Methods

sklearn.exceptions: Exceptions and warnings
sklearn.experimental: Experimental
sklearn.feature_extraction: Feature Extraction
sklearn.feature_selection: Feature Selection
sklearn.gaussian_process: Gaussian Processes
sklearn.impute: Impute
sklearn.inspection: Inspection
sklearn.isotonic: Isotonic regression
sklearn.kernel_approximation: Kernel Approximation
sklearn.kernel_ridge: Kernel Ridge Regression
sklearn.linear_model: Linear Models
sklearn.manifold: Manifold Learning
sklearn.metrics: Metrics
sklearn.mixture: Gaussian Mixture Models
sklearn.model_selection: Model Selection
sklearn.multiclass: Multiclass

sklearn.multioutput: Multioutput regression and classification
sklearn.naive_bayes: Naive Bayes
sklearn.neighbors: Nearest Neighbors
sklearn.neural_network: Neural network models
sklearn.pipeline: Pipeline
sklearn.preprocessing: Preprocessing and Normalization
sklearn.random_projection: Random projection
sklearn.semi_supervised: Semi-Supervised Learning
sklearn.svm: Support Vector Machines
sklearn.tree: Decision Trees
sklearn.utils: Utilities
Recently deprecated

5. scikit-learn

모듈	설명
<code>sklearn.datasets</code>	내장된 예제 데이터 세트
<code>sklearn.preprocessing</code>	다양한 데이터 전처리 기능 제공 (변환, 정규화, 스케일링 등)
<code>sklearn.feature_selection</code>	특징(feature)을 선택할 수 있는 기능 제공
<code>sklearn.feature_extraction</code>	특징(feature) 추출에 사용
<code>sklearn.decomposition</code>	차원 축소 관련 알고리즘 지원 (PCA, NMF, Truncated SVD 등)
<code>sklearn.model_selection</code>	교차 검증을 위해 데이터를 학습/테스트용으로 분리, 최적 파라미터를 추출하는 API 제공 (GridSearch 등)
<code>sklearn.metrics</code>	분류, 회귀, 클러스터링, Pairwise에 대한 다양한 성능 측정 방법 제공 (Accuracy, Precision, Recall, ROC-AUC, RMSE 등)
<code>sklearn.pipeline</code>	특징 처리 등의 변환과 ML 알고리즘 학습, 예측 등을 묶어서 실행할 수 있는 유틸리티 제공
<code>sklearn.linear_model</code>	선형 회귀, 릿지(Ridge), 라쏘(Lasso), 로지스틱 회귀 등 회귀 관련 알고리즘과 SGD(Stochastic Gradient Descent) 알고리즘 제공
<code>sklearn.svm</code>	서포트 벡터 머신 알고리즘 제공
<code>sklearn.neighbors</code>	최근접 이웃 알고리즘 제공 (k-NN 등)
<code>sklearn.naive_bayes</code>	나이브 베이즈 알고리즘 제공 (가우시안 NB, 다항 분포 NB 등)
<code>sklearn.tree</code>	의사 결정 트리 알고리즘 제공
<code>sklearn.ensemble</code>	앙상블 알고리즘 제공 (Random Forest, AdaBoost, GradientBoost 등)
<code>sklearn.cluster</code>	비지도 클러스터링 알고리즘 제공 (k-Means, 계층형 클러스터링, DBSCAN 등)

<https://makeit.tistory.com/132>

6. pandas

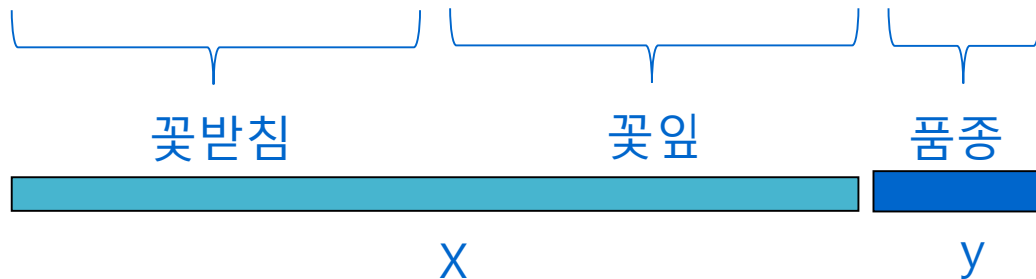
- Pandas is an open source library that is used to analyze data in Python.
- It takes in data, like a CSV or SQL database, and creates an object with rows and columns called a **data frame**.
- Pandas is typically imported with the alias `pd`
- Install pandas
 - Python console에서 `install pandas`
- Numpy 의 배열은 주로 숫자를 저장
- Pandas 의 data frame 은 문자, 숫자 컬럼을 함께 저장 가능

6. pandas

- Example dataset : iris.csv

실습에 사용되는
데이터셋은 자료실에
게시되어 있음

	A	B	C	D	E
1	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
2	5.1	3.5	1.4	0.2	setosa
3	4.9	3	1.4	0.2	setosa
4	4.7	3.2	1.3	0.2	setosa
5	4.6	3.1	1.5	0.2	setosa
6	5	3.6	1.4	0.2	setosa
7	5.4	3.9	1.7	0.4	setosa
8	4.6	3.4	1.4	0.3	setosa
9	5	3.4	1.5	0.2	setosa
10	4.4	2.9	1.4	0.2	setosa
11	4.9	3.1	1.5	0.1	setosa



6. pandas

03.pandas_test.py

```
# Pandas test

import pandas as pd

# read dataset
iris = pd.read_csv('d:/data/iris.csv')

iris                                # view data
with pd.option_context('display.max_rows', None, 'display.max_columns',
None):
    print(iris)                    # view all data

iris.head()                        # view head of data frame
iris.head(10)                      # view head of data frame
iris.tail()                        # view tail of data frame
iris.tail(15)                      # view tail of data frame

iris.shape                         # dimension
type(iris)                        # type
iris.columns                       # column names
iris.columns[:4]                  # column names
```

6. pandas

```
iris['Species']           # get column by name
iris[['Sepal.Width', 'Sepal.Length']] # get column by name

iris.iloc[90,4]           # indexing cell
iris.iloc[50,0]           # indexing cell

iris.iloc[10:50,0:4]      # slice row/col
iris.iloc[10:50,:]        # slice row/col

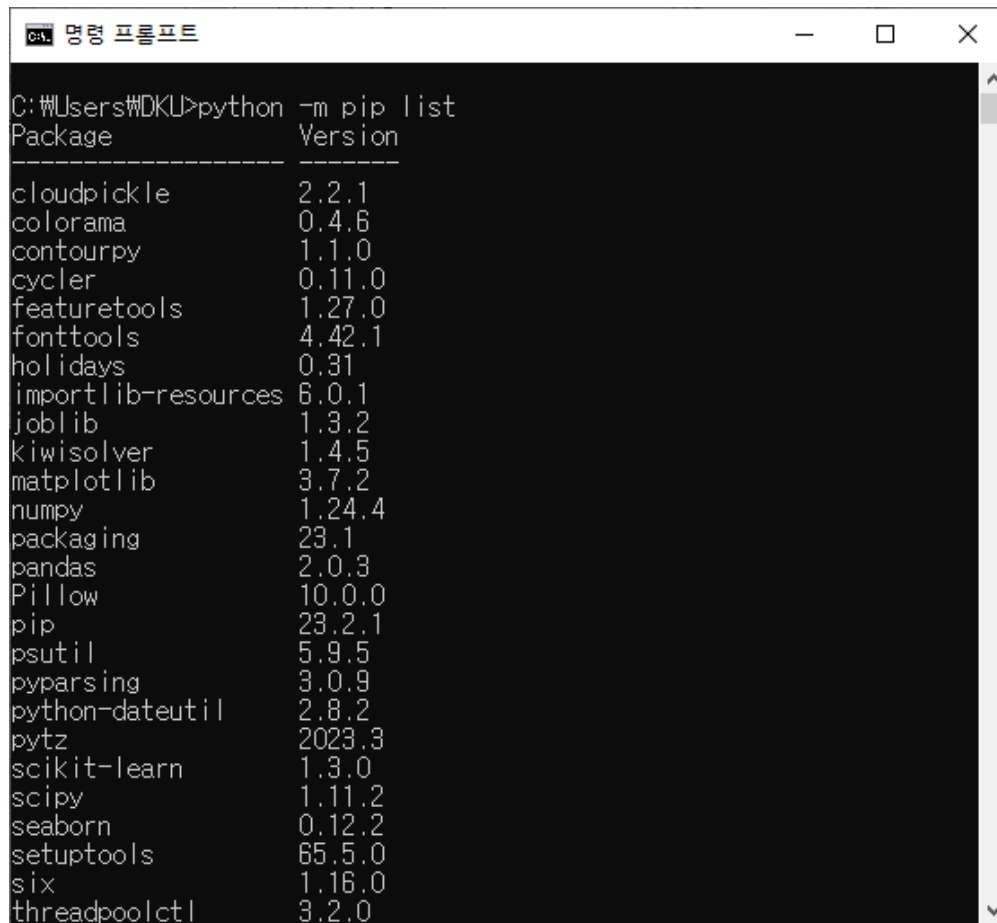
iris.loc[10:50, 'Sepal.Width'] # slice row/col
iris.loc[10:50, ['Sepal.Width', 'Sepal.Length']] # slice row/col

# slicing rows by condition
iris.loc[iris['Species']=='setosa',:]
```


Note.

- Check installed modules (library)

```
pip list
```



```
C:\Users\WDKJ>python -m pip list
Package            Version
-----
cloudpickle        2.2.1
colorama           0.4.6
contourpy          1.1.0
cyclor             0.11.0
featuretools       1.27.0
fonttools          4.42.1
holidays          0.31
importlib-resources 6.0.1
joblib             1.3.2
kiwisolver         1.4.5
matplotlib         3.7.2
numpy              1.24.4
packaging          23.1
pandas             2.0.3
Pillow             10.0.0
pip               23.2.1
psutil             5.9.5
pyparsing          3.0.9
python-dateutil    2.8.2
pytz               2023.3
scikit-learn       1.3.0
scipy              1.11.2
seaborn            0.12.2
setuptools         65.5.0
six                1.16.0
threadpoolctl      3.2.0
```