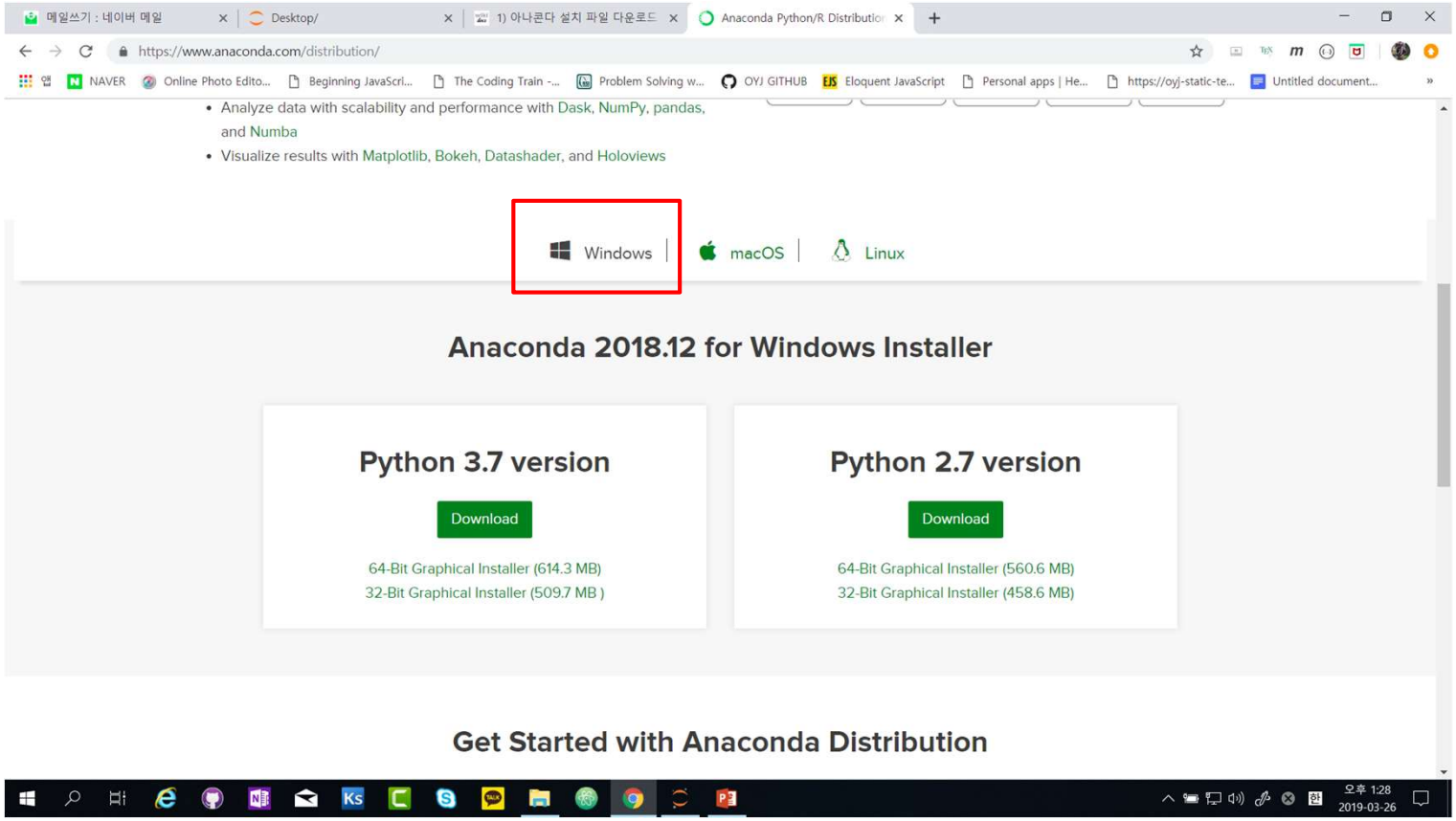
Python 을 활용한 머신러닝 입문

2019. 6

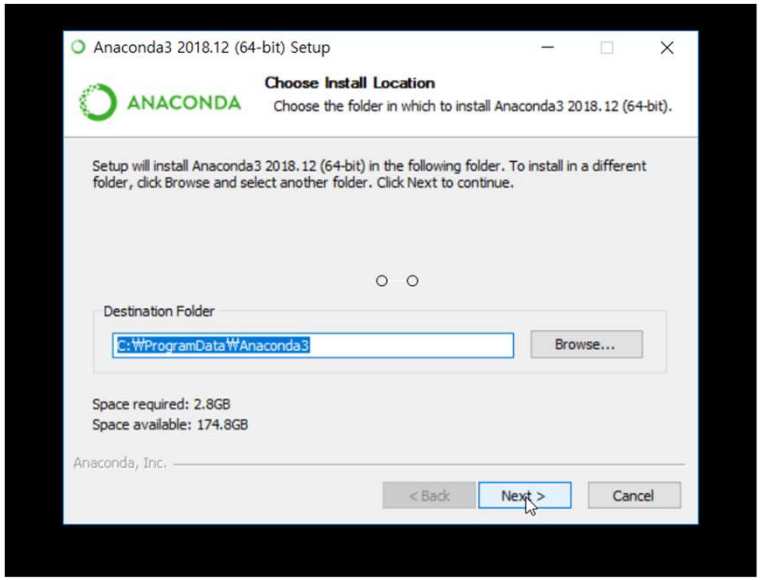
Anaconda Installation

- https://www.anaconda.com/distribution



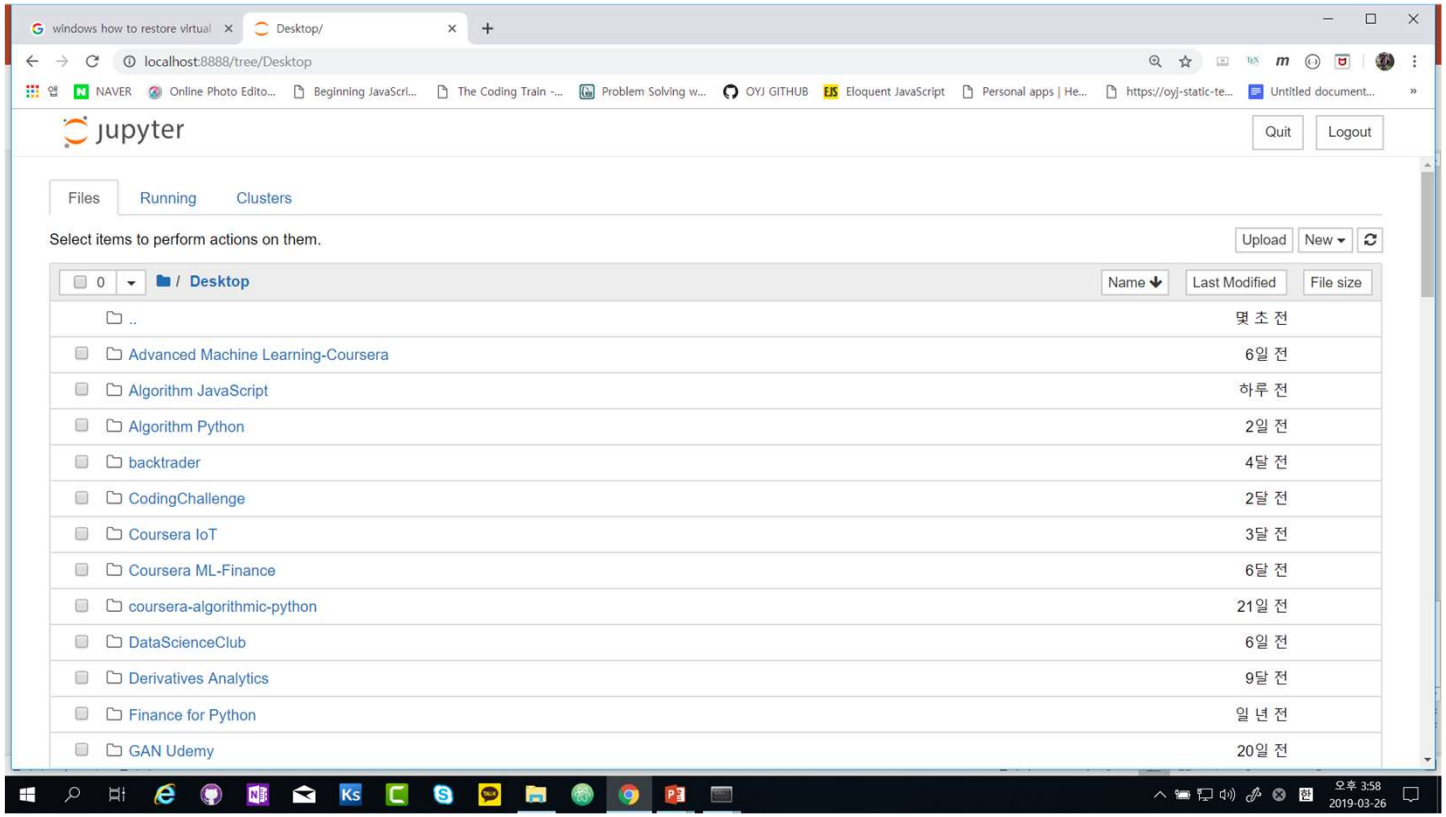
Anaconda 설치


Anaconda3 2018.12 (64-bit) Setup
ANACONDA.
Welcome to Anaconda3 2018.12 (64-bit) Setup
Setup will guide you through the installation of Anaconda3 2018.12 (64-bit).
It is recommended that you close all other applications before starting Setup. This will make it possible to update relevant system files without having to reboot your computer.
Click Next to continue.
Next >
Cancel

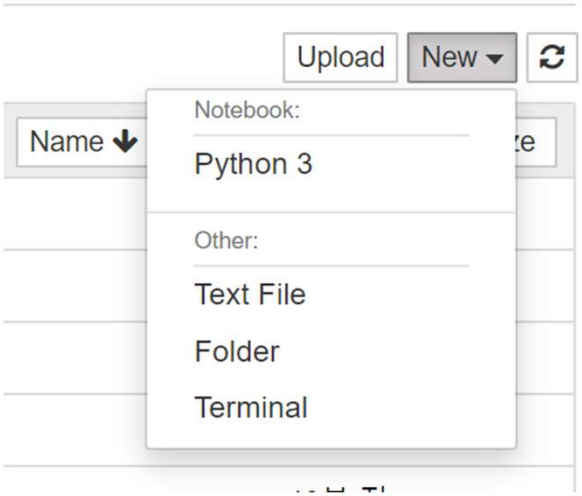

Anaconda3 2018.12 (64-bit) Setup
ANACONDA
Choose Install Location
Choose the folder in which to install Anaconda3 2018.12 (64-bit).
Setup will install Anaconda3 2018.12 (64-bit) in the following folder. To install in a different folder, click Browse and select another folder. Click Next to continue.
Destination Folder
C:₩ProgramData₩Anaconda3
Browse...
Space required: 2.8GB
Space available: 174.8GB
Anaconda, Inc.
< Back
Next >
Cancel

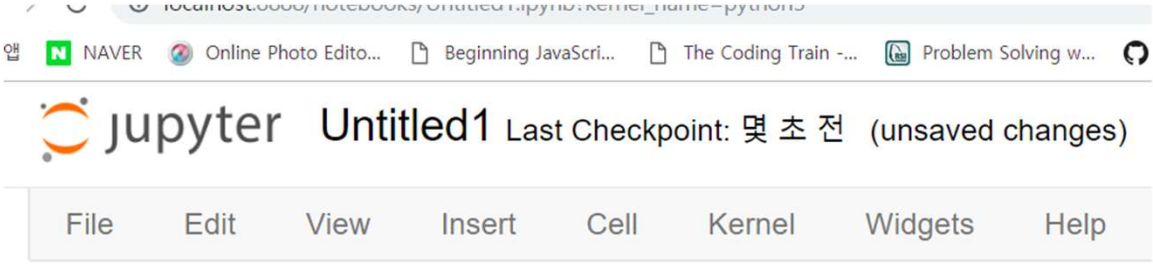
- 개인별 작업 directory 생성
- 패키지 설지 : pip install <package name>
- 설치된 Python package 확인 : pip list

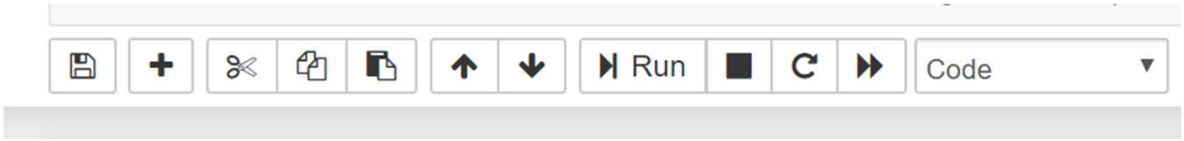
>jupyter notebook



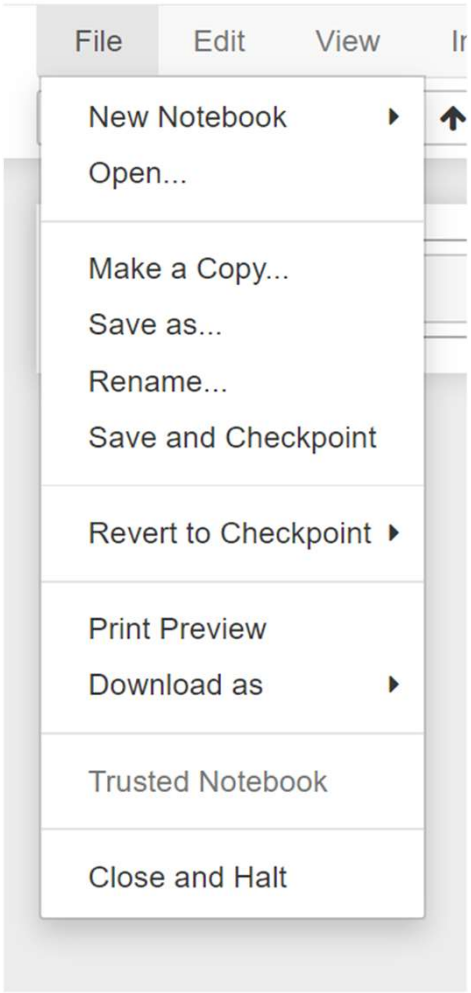
Jupyter Notebook 사용방법

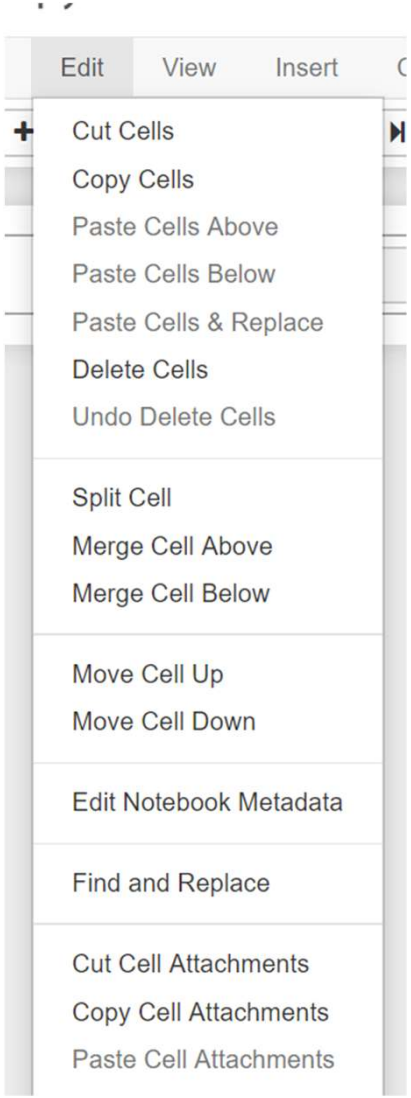


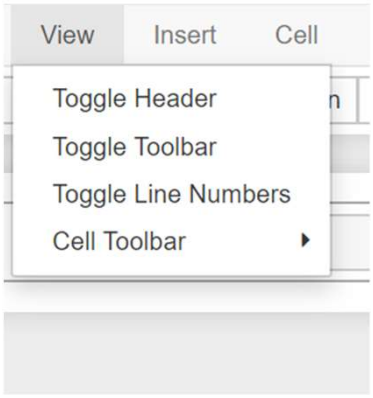




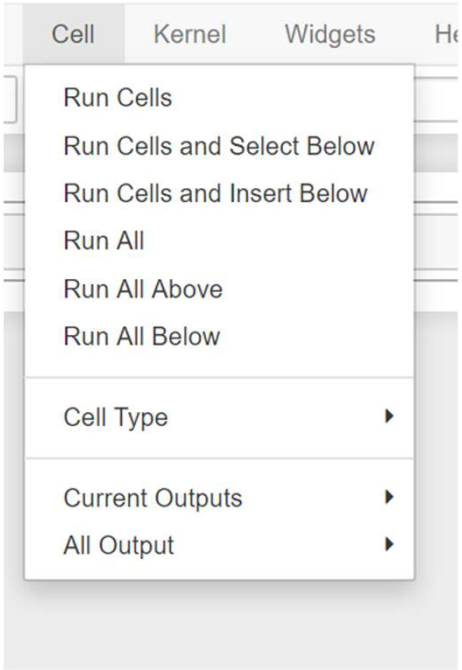
Jupyter Notebook 사용방법

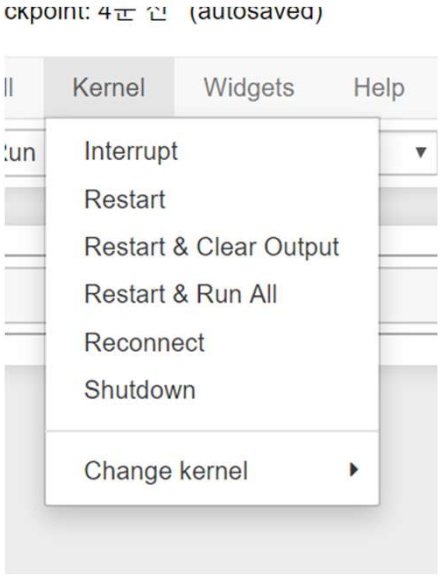


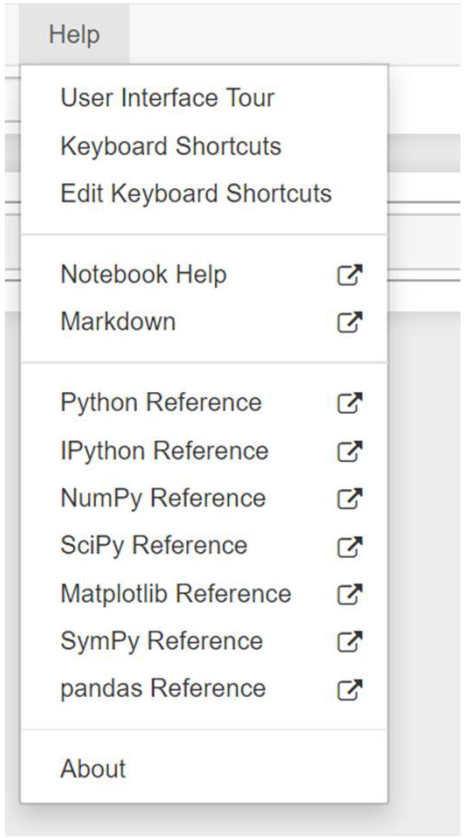




Jupyter Notebook 사용방법

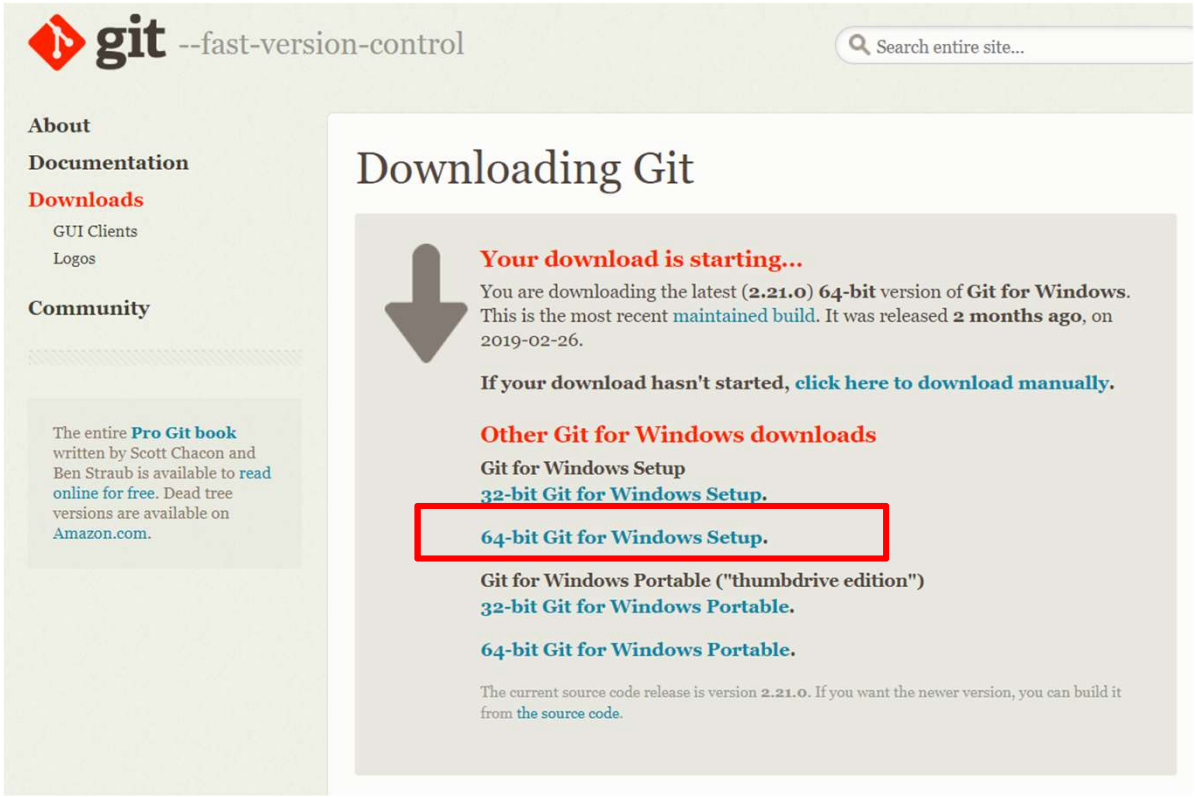






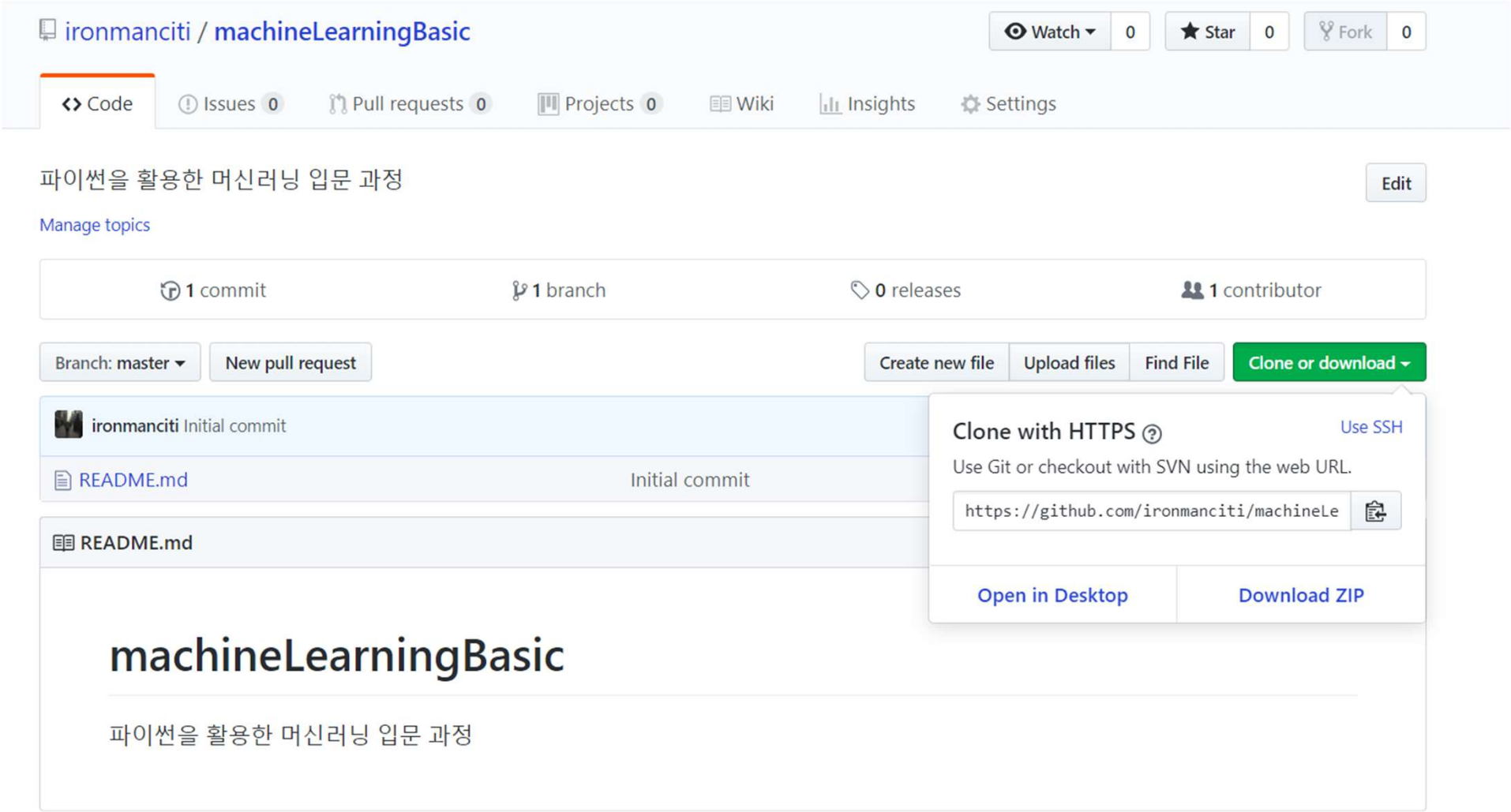
git Installation

https://git-scm.com/download/win



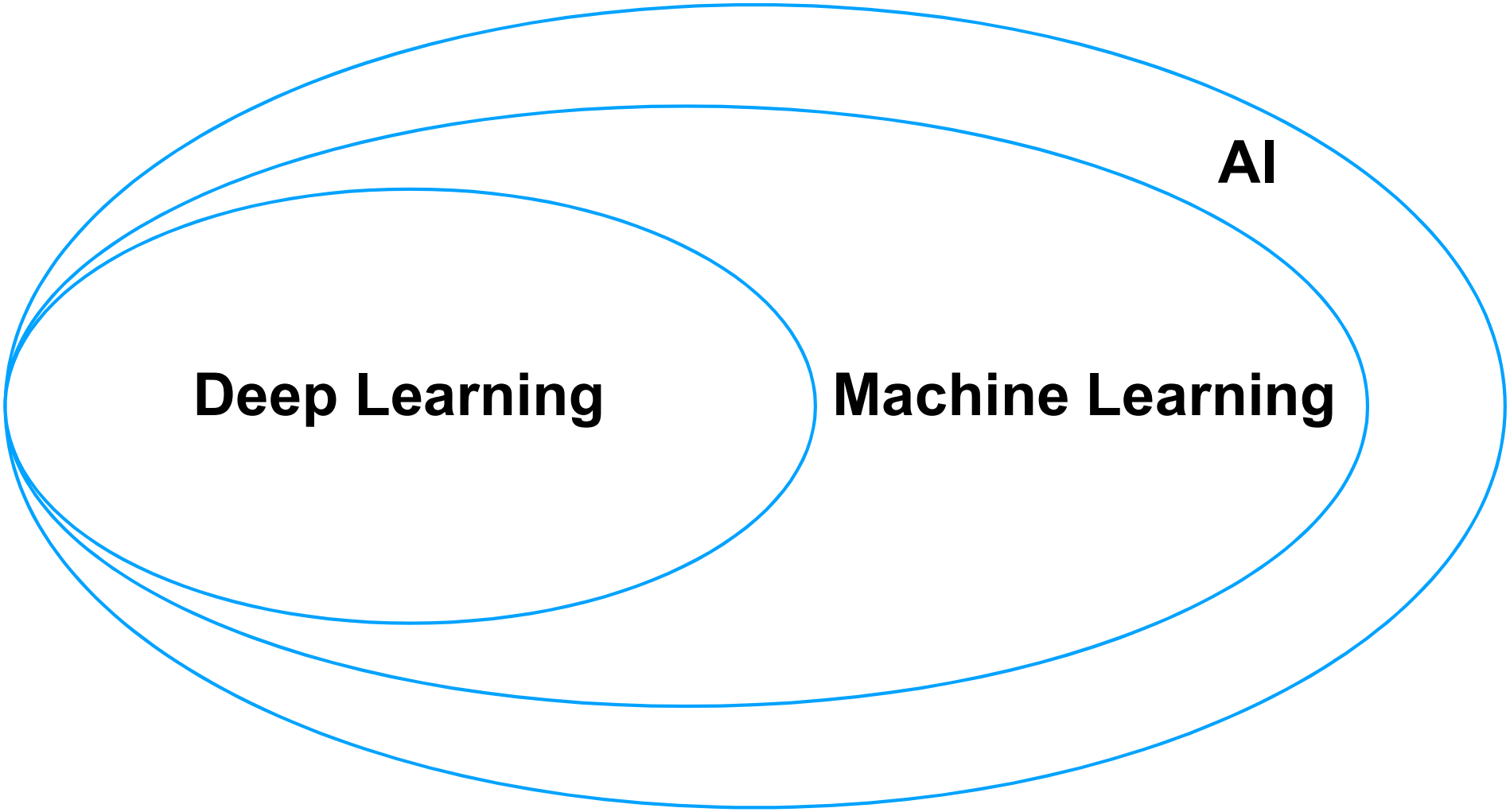
Github Repository

https://github.com/ironmanciti/machineLearningBasic



Machine Learning
개요

AI vs ML vs DL



History of Machine Learning

- **탄생 [1950 년대] : 1958**
  — Frank Rosenblatt 가 인간의 뇌신경을 본떠 Perceptron 고안

- **AI 의 첫번째 Winter [1970 년대] :**
  Marvin Minsky 가 Perceptron 은 XOR 문제를 해결할 수 없음을 수학적으로 증명

- **중흥기 [1980 년대] :** Expert System

- **AI 의 두번째 Winter [1987 – 1993]**
  – Jeffery Hinton (Toronto 대학) back-propagation algorithm 개발

- **IBM Deep Blue 가 Garry Kasparov 에 승리 - 1996**
- **Google Brain 이 최초로 인간 얼굴 인식 - 2012**
- **~ 상업적 대 폭발기**
- **Alphago 이세돌에 승리 – 2016**

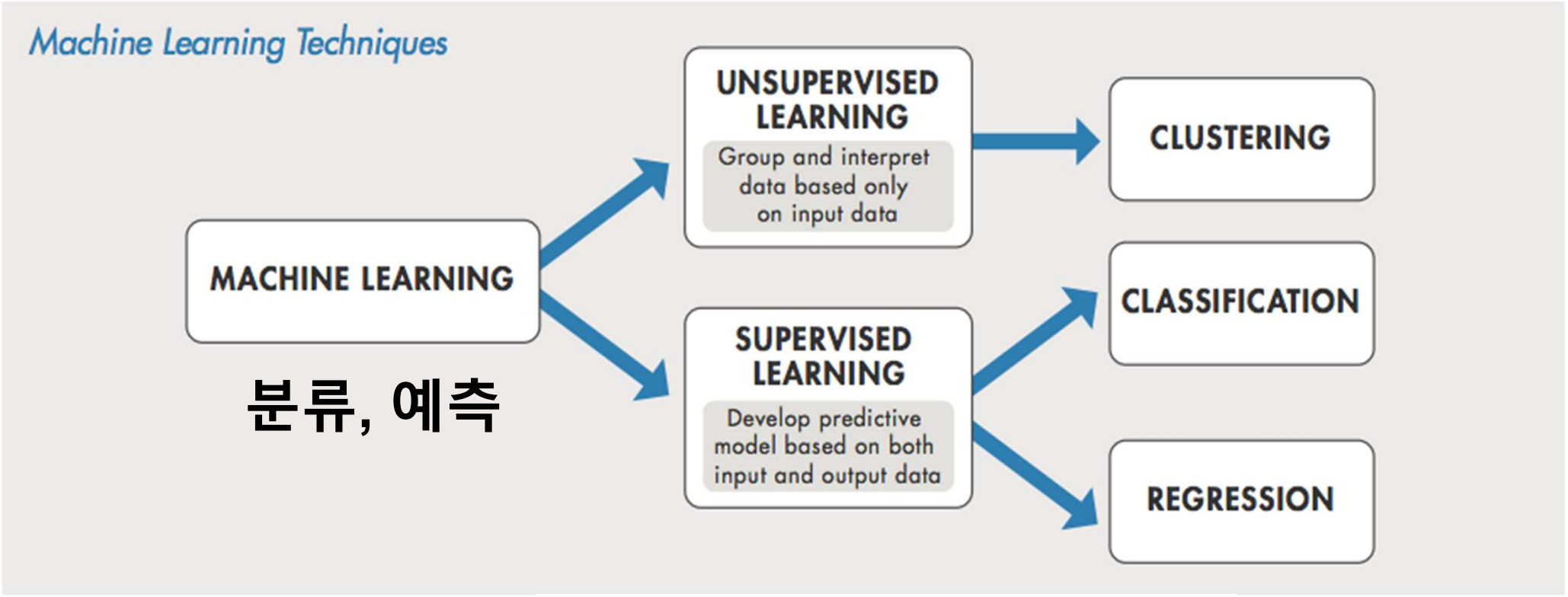
Andrew Ng.
Stanford Univ.
Google/Baidu

HEROs

Jeffery Hinton
Toronto Univ.
Google

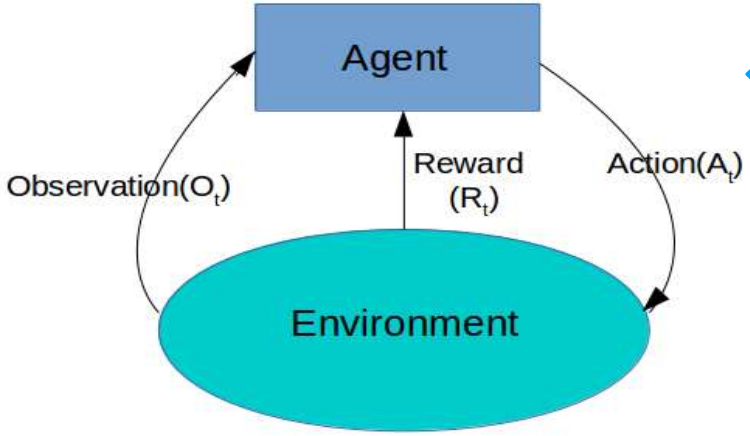
Yan LeCun
New York Univ.
Facebook

Machine Learning 의 종류



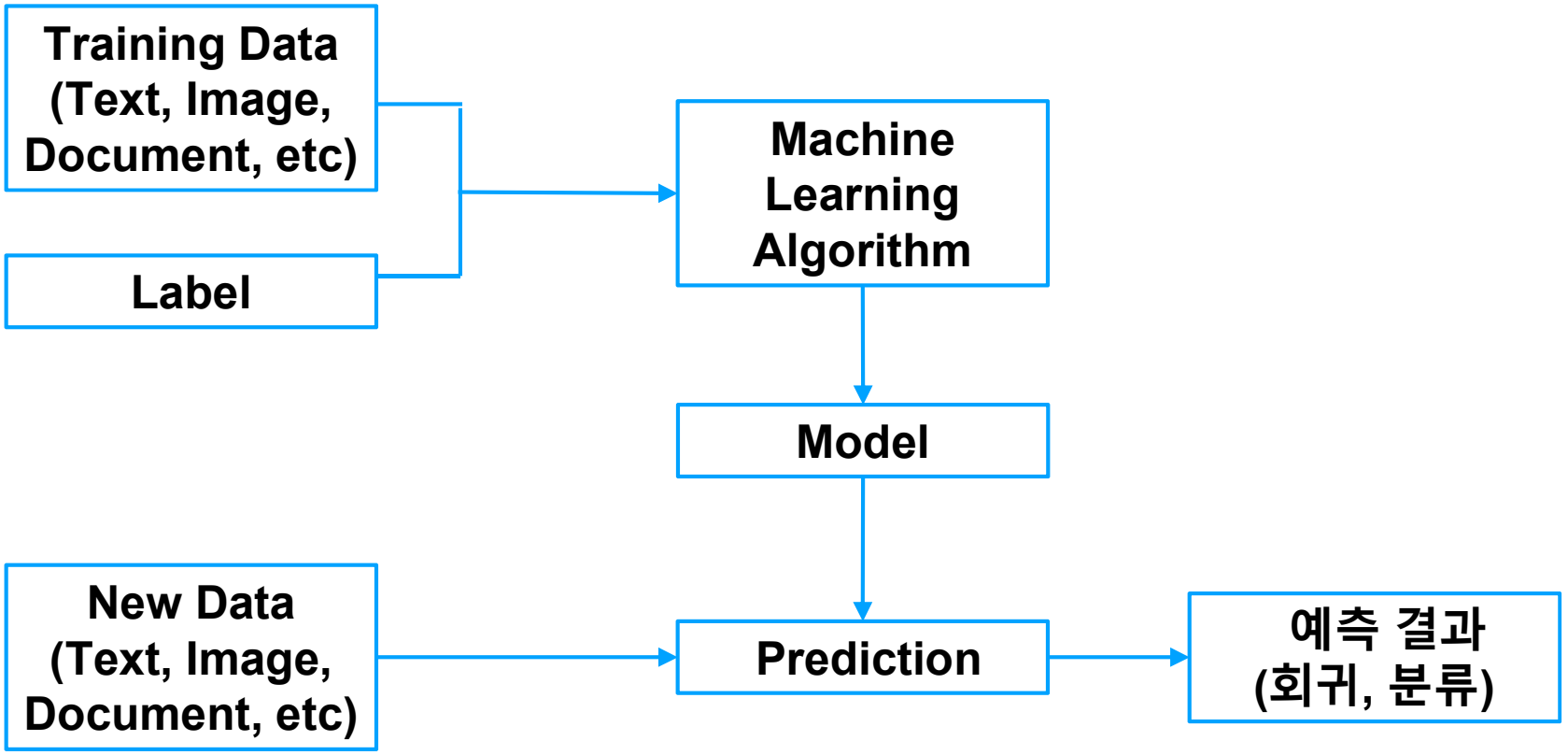
분류, 예측

REINFORCEMENT LEARNING

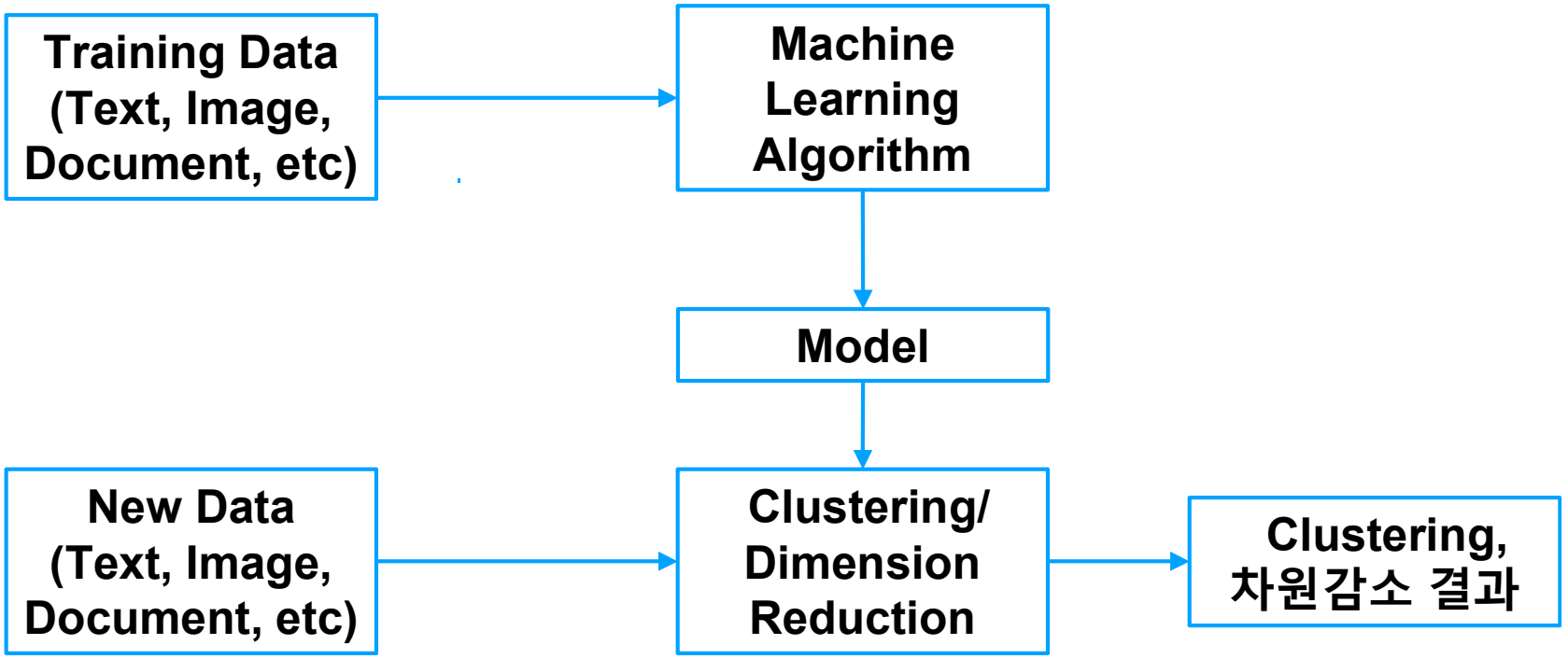


의사결정

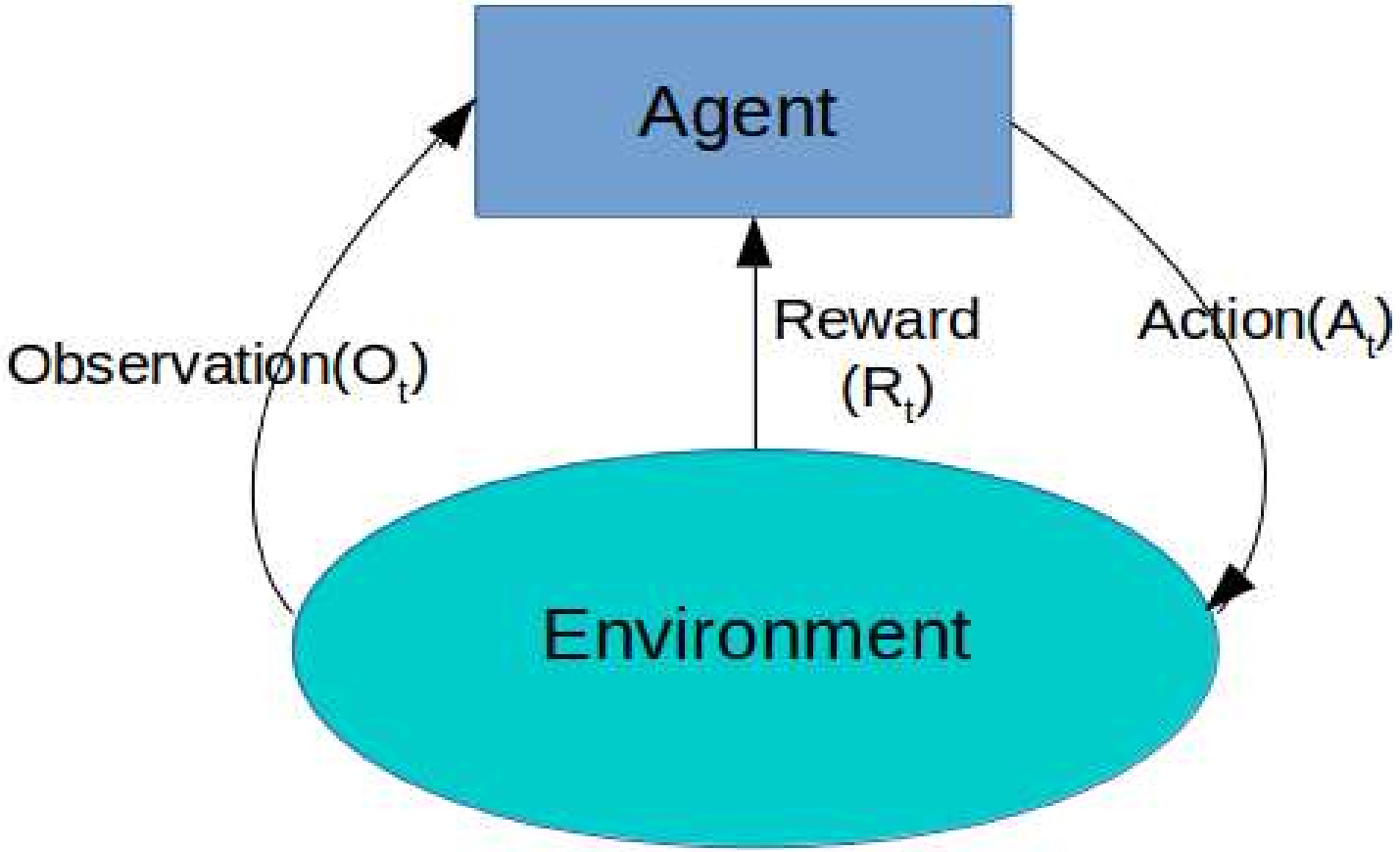
Supervised Learning (지도학습)



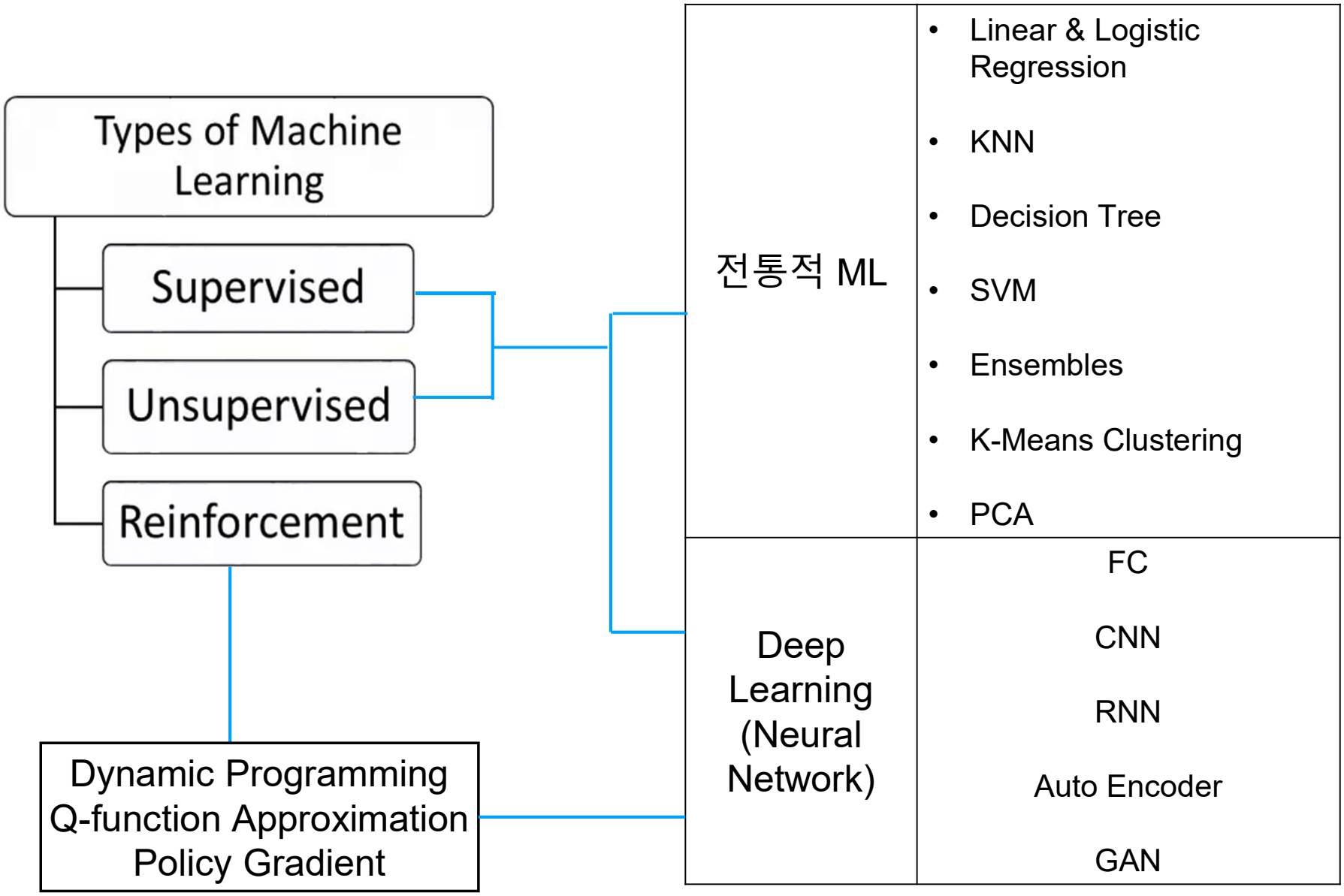
Unsupervised Learning (비지도학습)



Reinforcement Learning (강화학습)



Machine Learning 기법의 종류



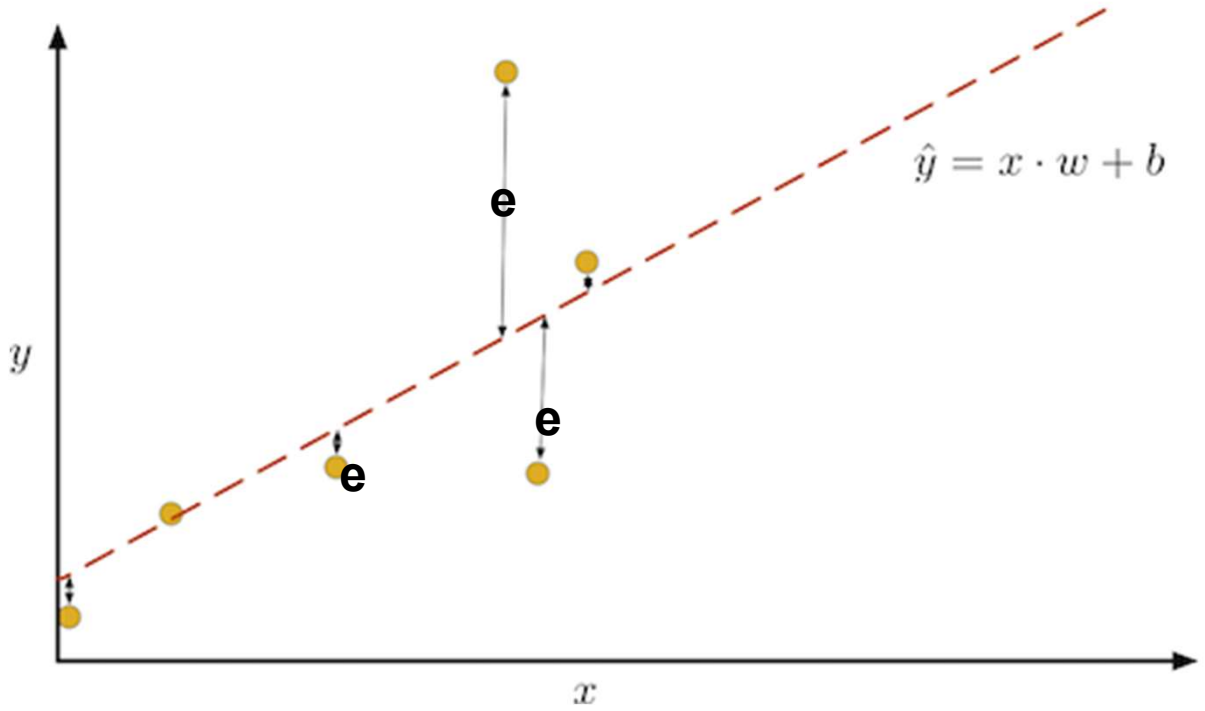
전통적 ML의 기법

종류	용도
Linear / Logistic Regression	선형회귀/분류
KNN (K-Nearest Neighbor)	분류
Decision Tree (결정나무)	분류
SVM (Support Vector Machine)	분류, 회귀
Ensemble (Random Forest, XGBoost, etc)	분류, 회귀
K-Means Clustering (K-평균 군집화)	군집화
PCA (Principal Component Analysis)	차원 축소

Neural Network 의 종류

종류	용도
FC (Fully Connected Neural Network)	분류, 회귀
CNN (Convolutional Neural Network)	Image 인식
RNN (Recurrent Neural Network)	시계열 인식, 자동번역, 감성분석 등
AE (Auto Encoder)	비지도 학습, 차원축소
GAN (Generative Adversarial Nets)	적대적 생성모델, 이미지 위조

Linear Regression



Ex) x : 강우량

y : 실제우산 판매량 (label)

$\bar{y}$: 모델에 의한 예측 값

w, b : 학습에 의한 model parameter

e : 예측 값과 실제 값의 차이

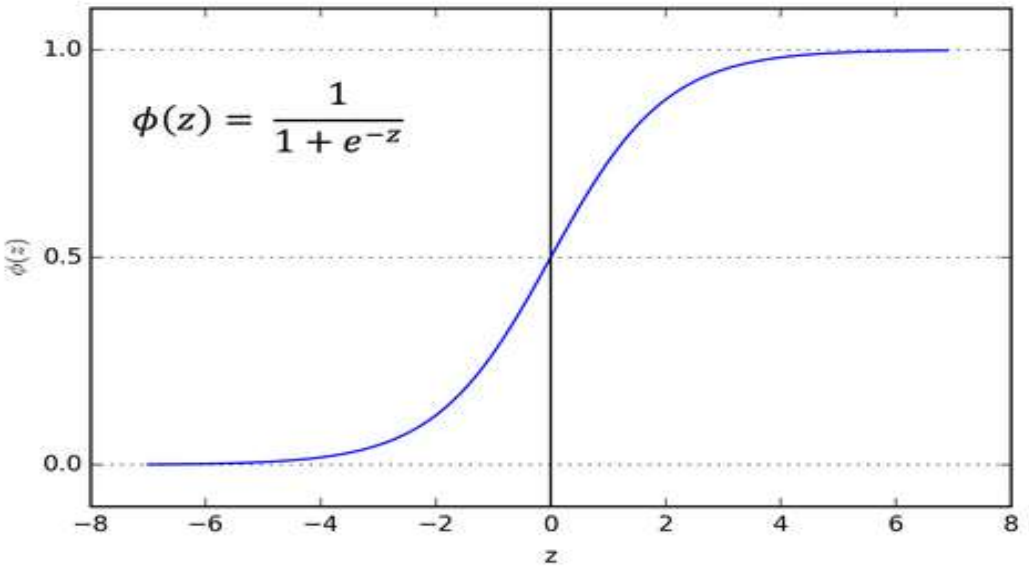
Machine Learning 은 주어진 data 에 대하여 Parameter 를 inference 하는 작업

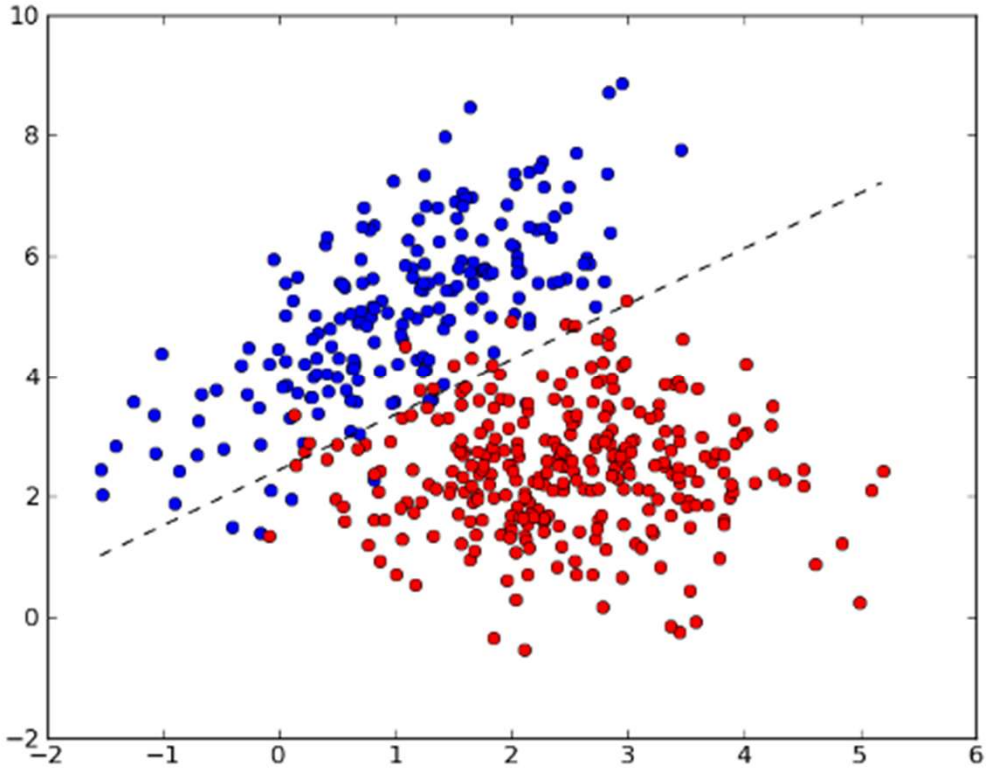
Logistic Regression

$$\Phi(z) = \frac{1}{1 + e^{-z}}$$

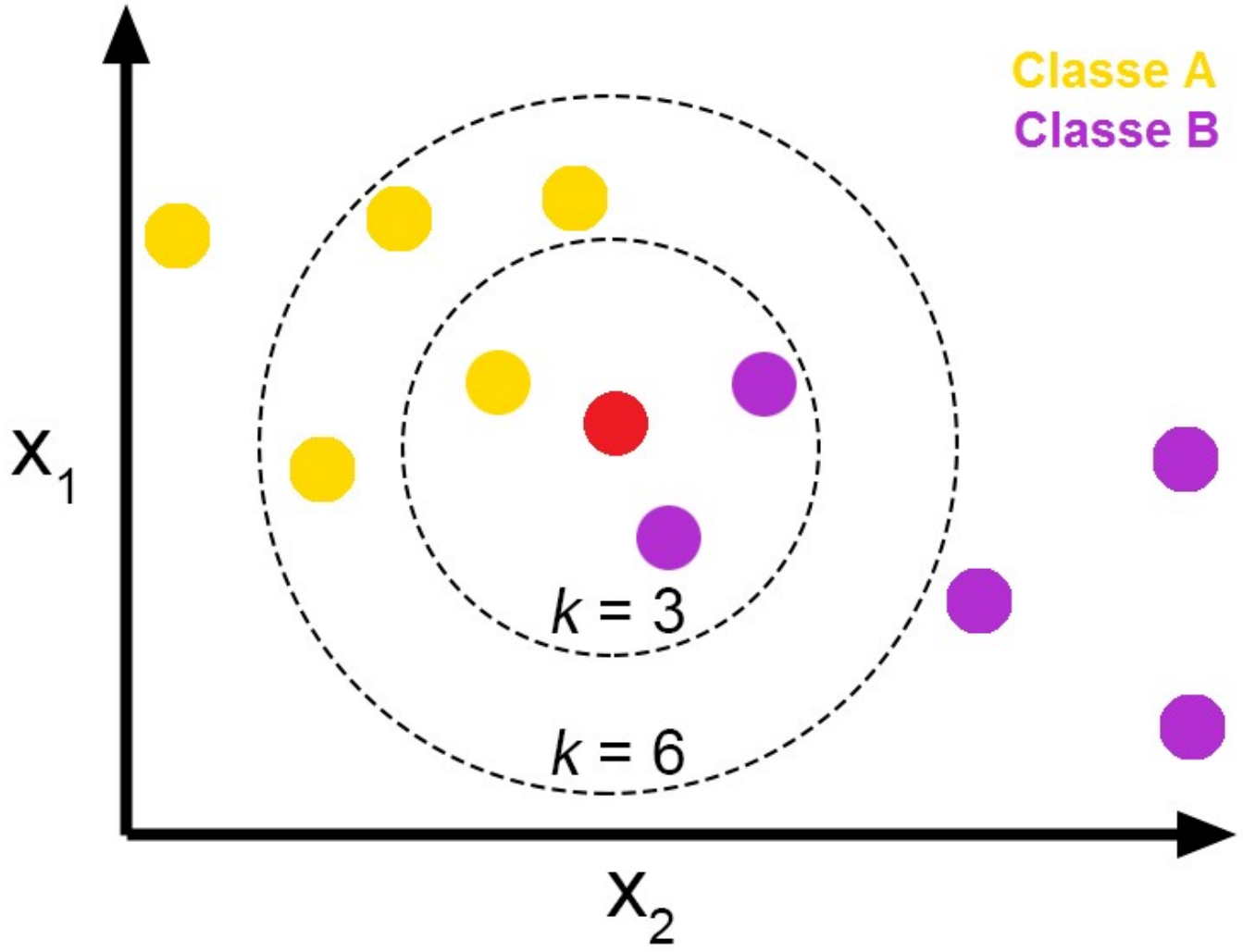
z : wx + b

If Φ(z) > 0.5 → blue
< 0.5 → red





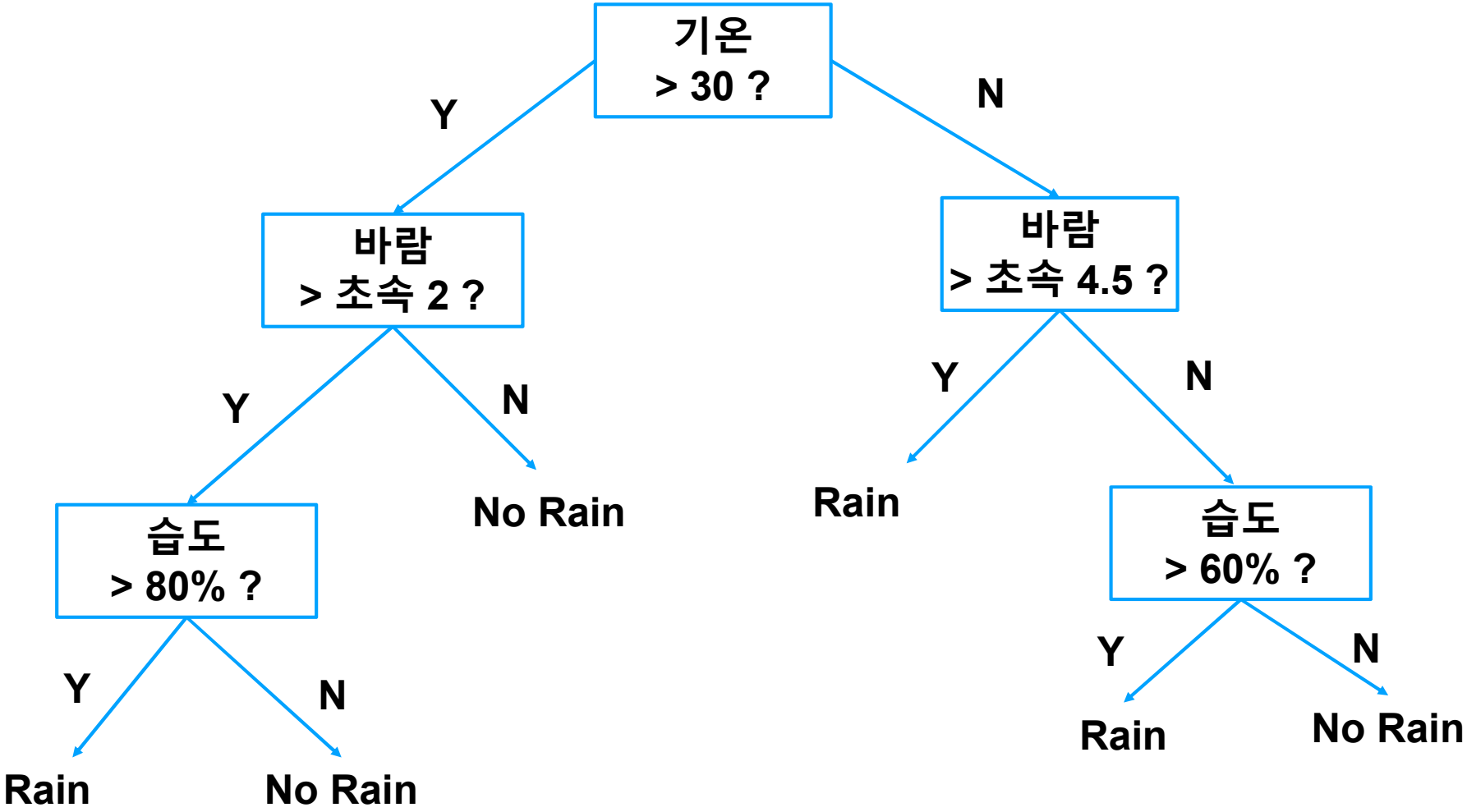
KNN (K-Nearest Neighbor)



If K=3 → Class B

If K=6 → Class A

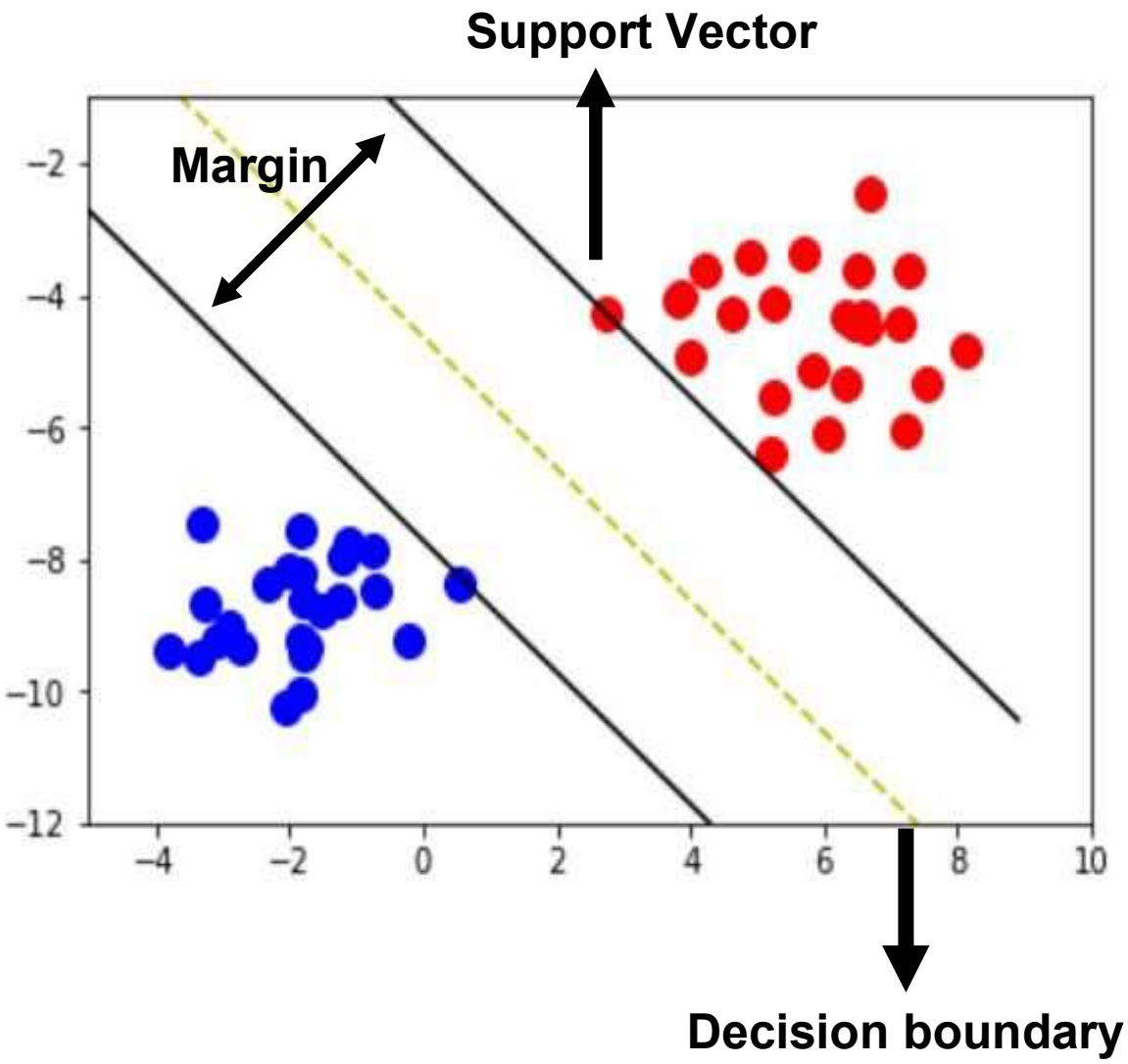
Decision Tree
(결정나무)



SVM (Support Vector Machine)

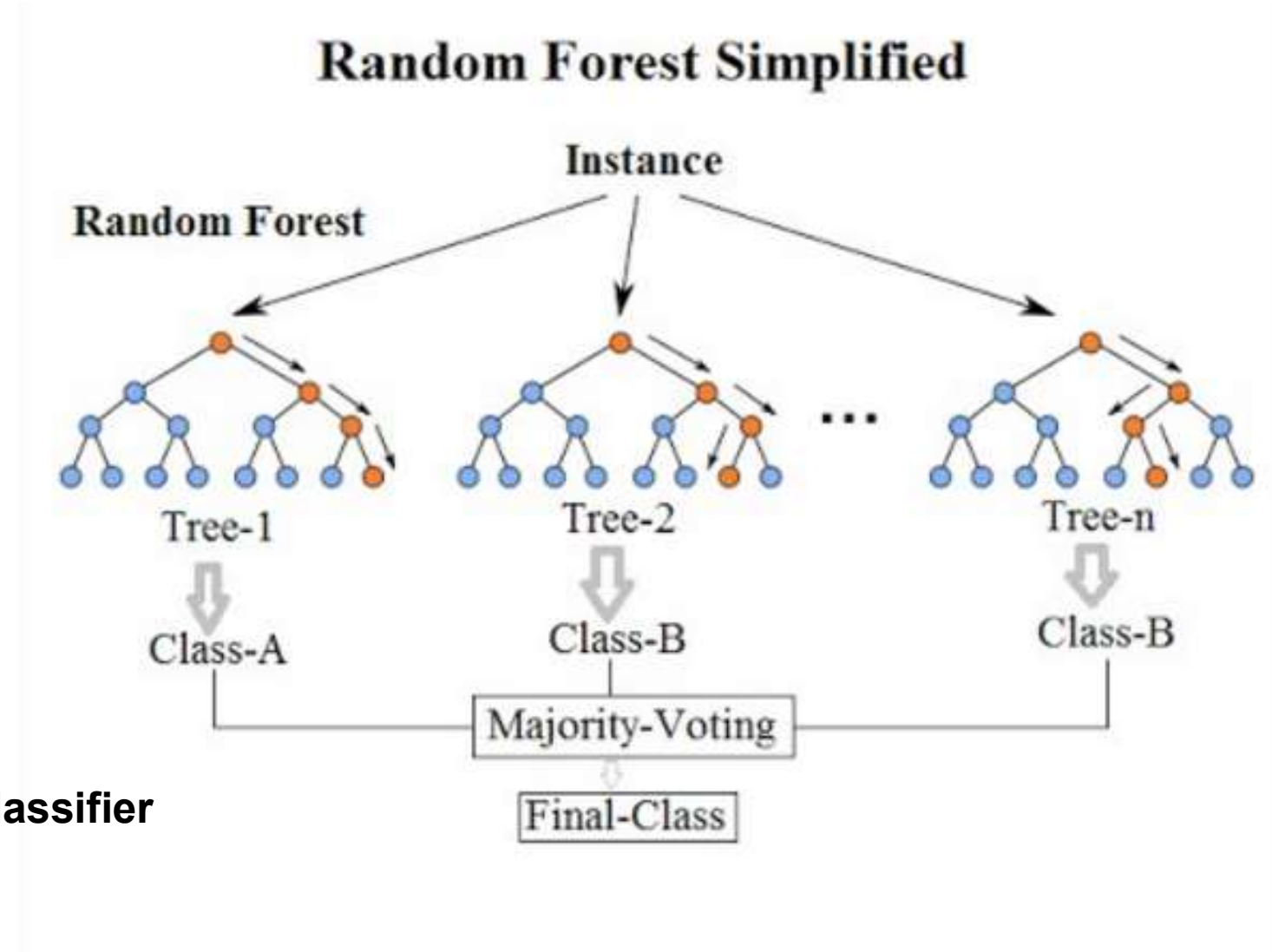
$$f(x) = sgn(w.x + b)$$

Yhat = 0 if f(x) < 0
1 if f(x) >= 0



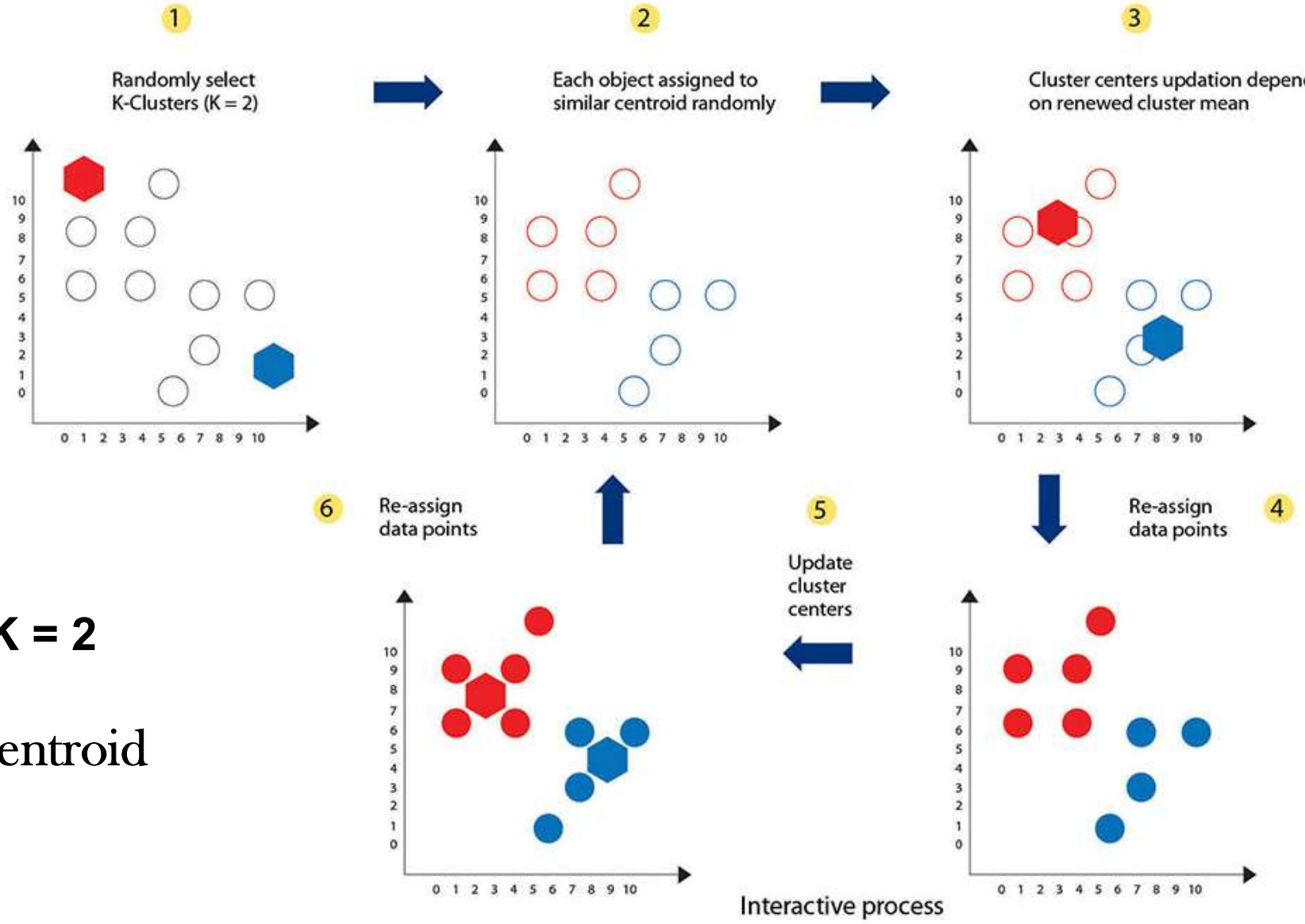
Ensemble
Random Forest

Weak classifiers



Strong classifier

K-Means Clustering
(K-평균 군집화)



K = 2

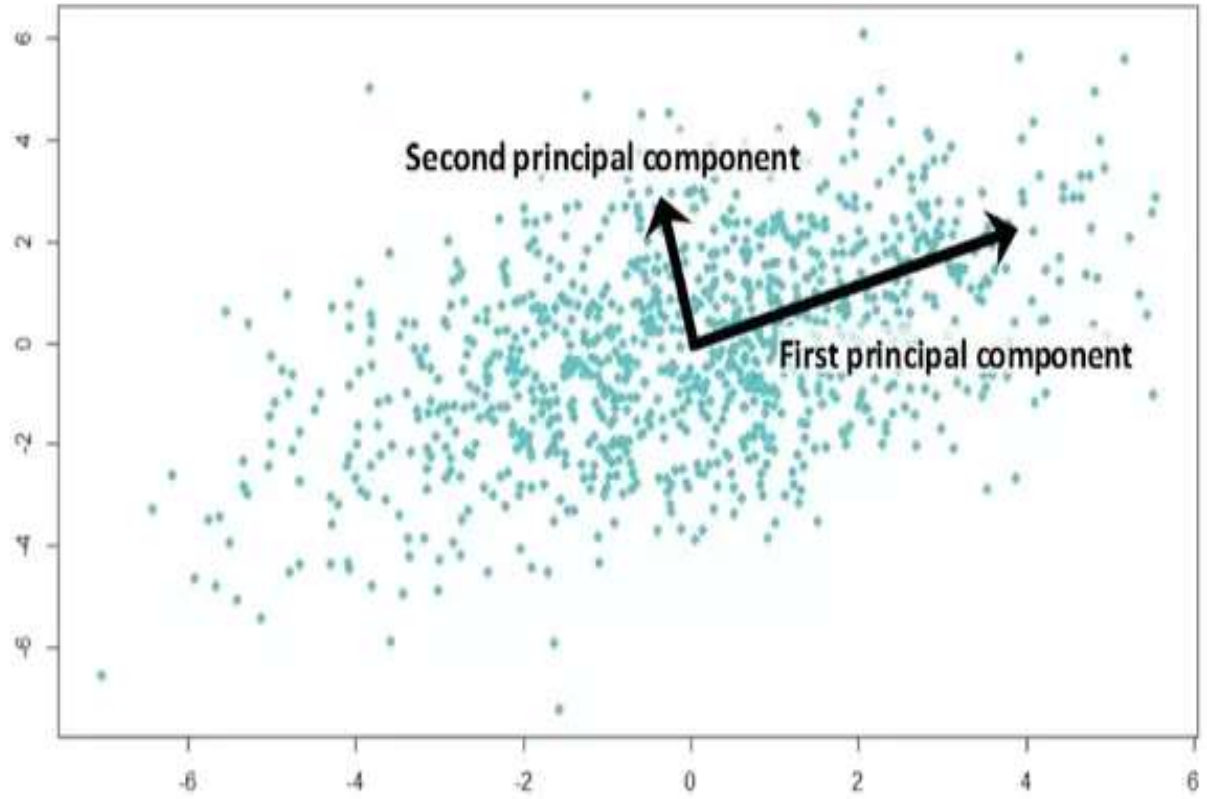
: centroid

PCA (Principal Component Analysis)
(주성분 분석)

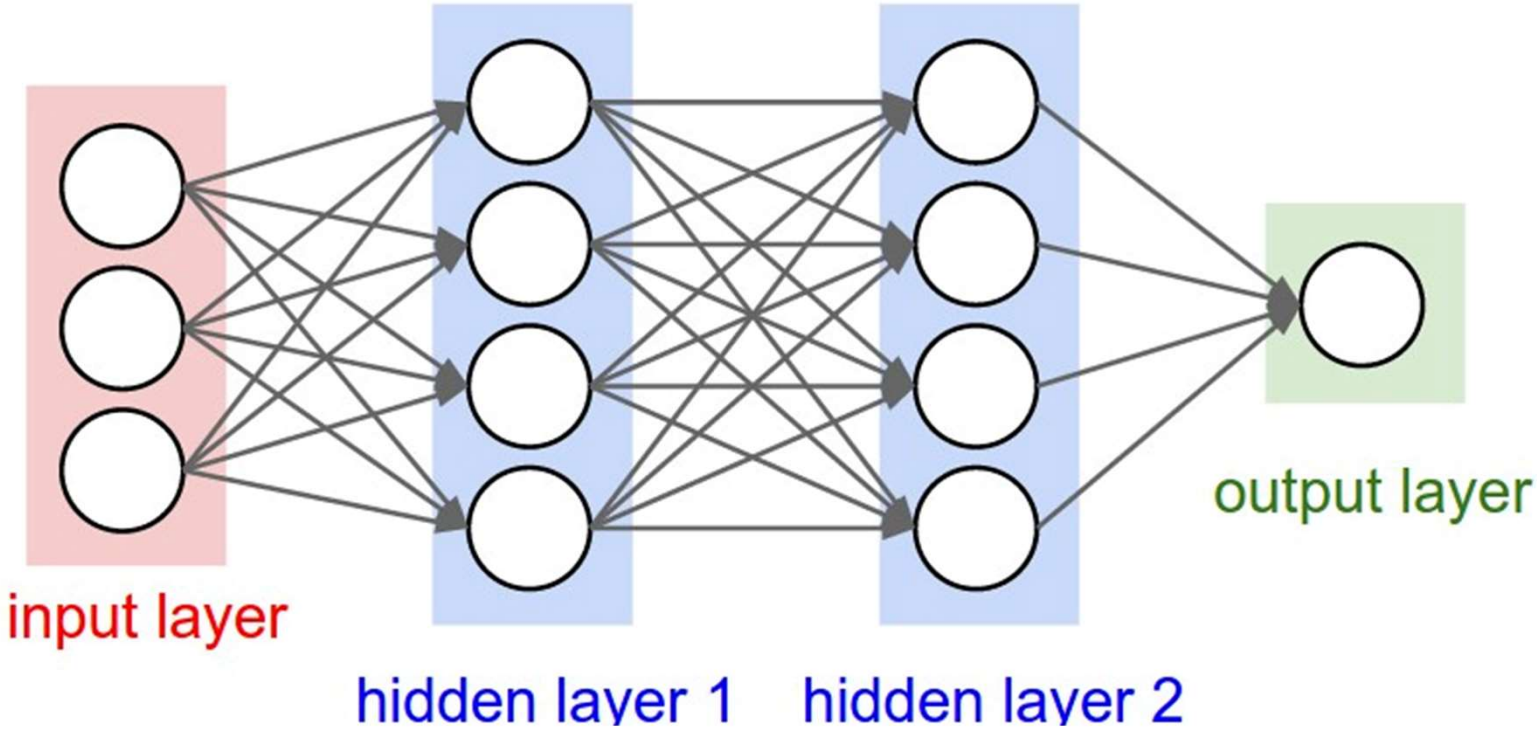
- 선형대수학의 SVD(Singluar Value Decomposition) 이용
- 어떤 m X n 행렬 A 는 다음과 같은 형태의 세가지 행렬의 곱으로 분해할 수 있다.

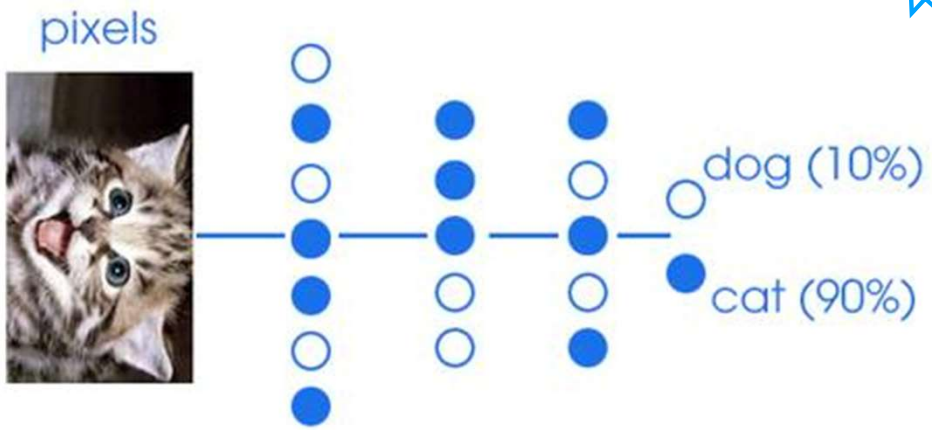
$$A_{mXn} = U_{nXn} \sum_{nXm} V^T_{mXm}$$

U : nXn 직교행렬
V : mXm 직교행렬
Σ : nXm 직사각대각행렬



Fully Connected Neural Network

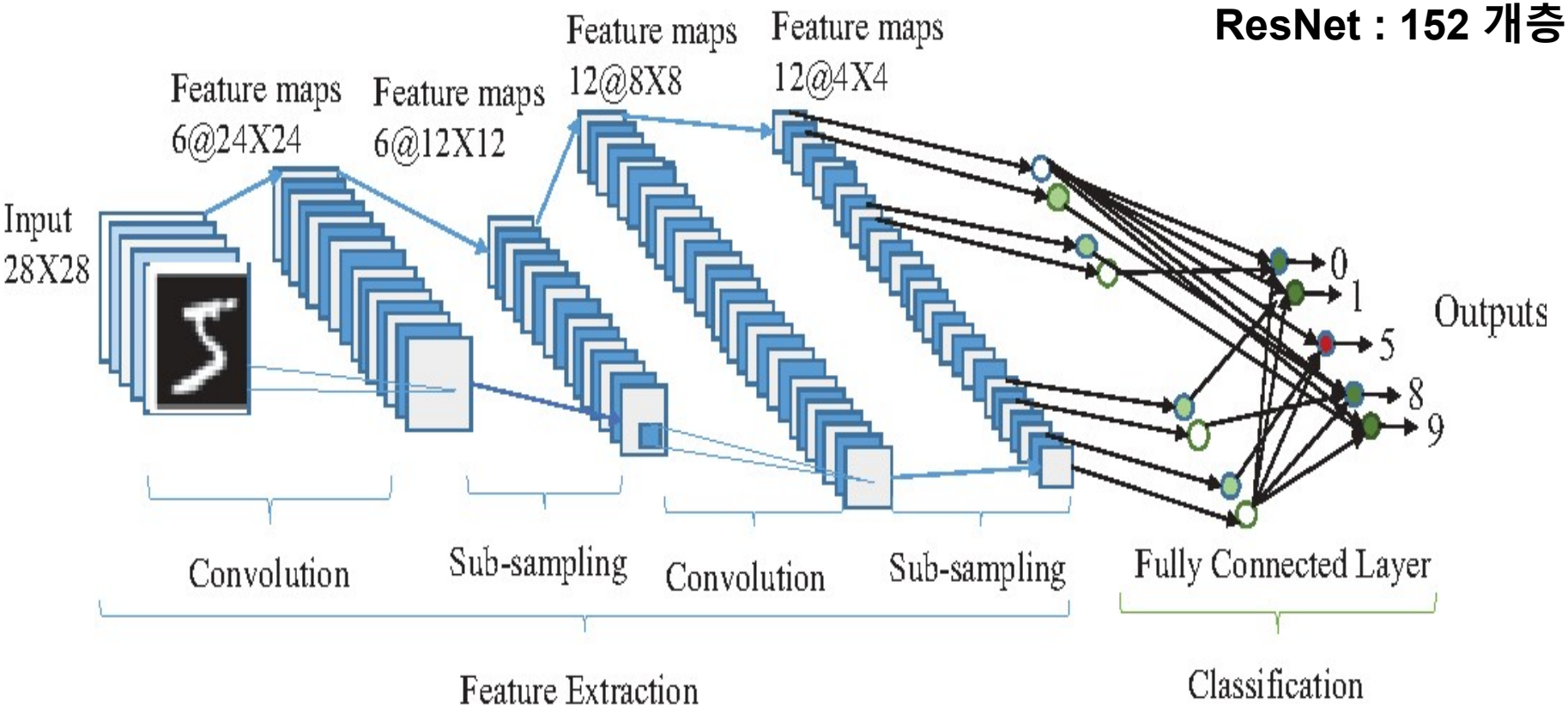




HOW DEEP ?

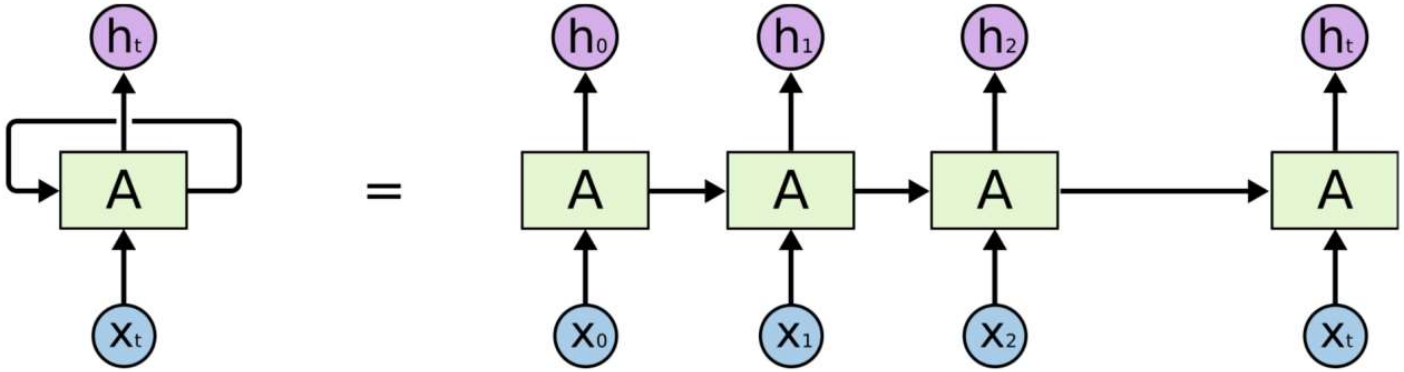
CNN (Convolutional Neural Network, 합성곱 신경망)

LeNet : 5 개층
ResNet : 152 개층

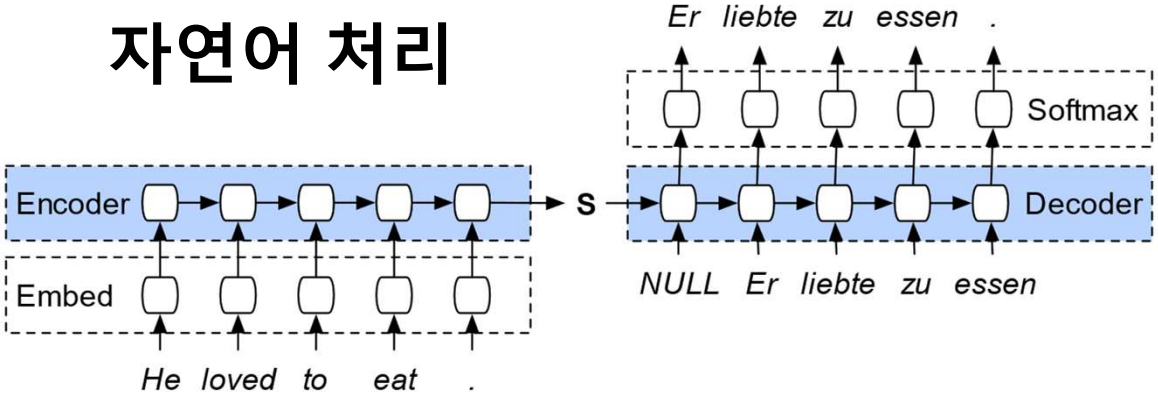


입력 **합성곱** **풀링** **완전연결**

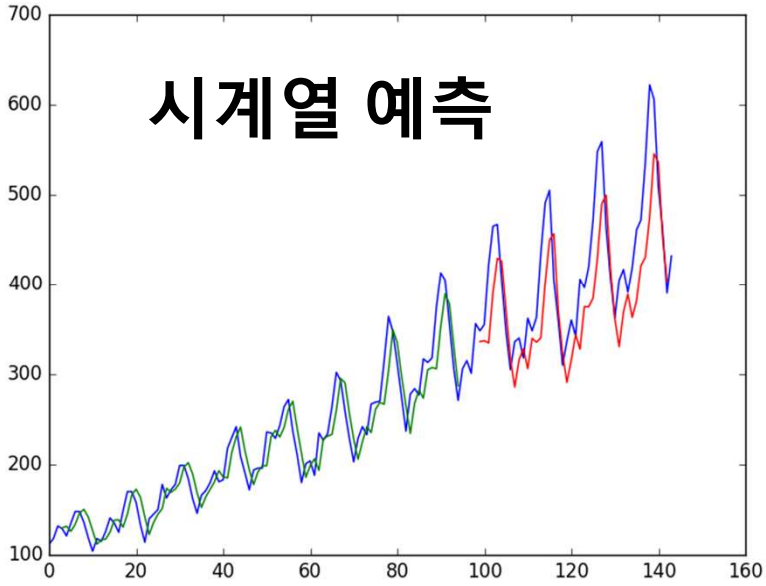
RNN
(Recurrent Neural Network, 순환 신경망)



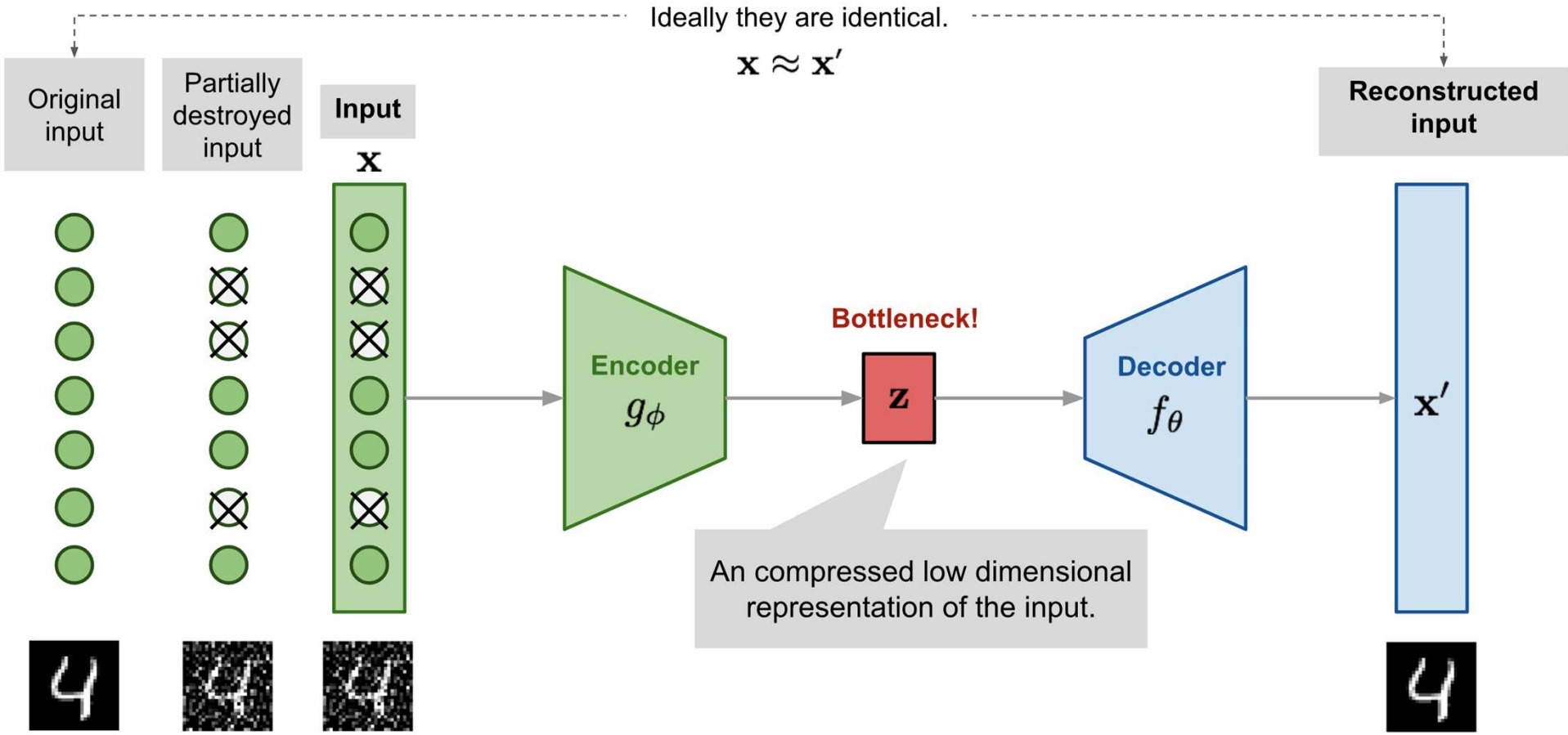
자연어 처리



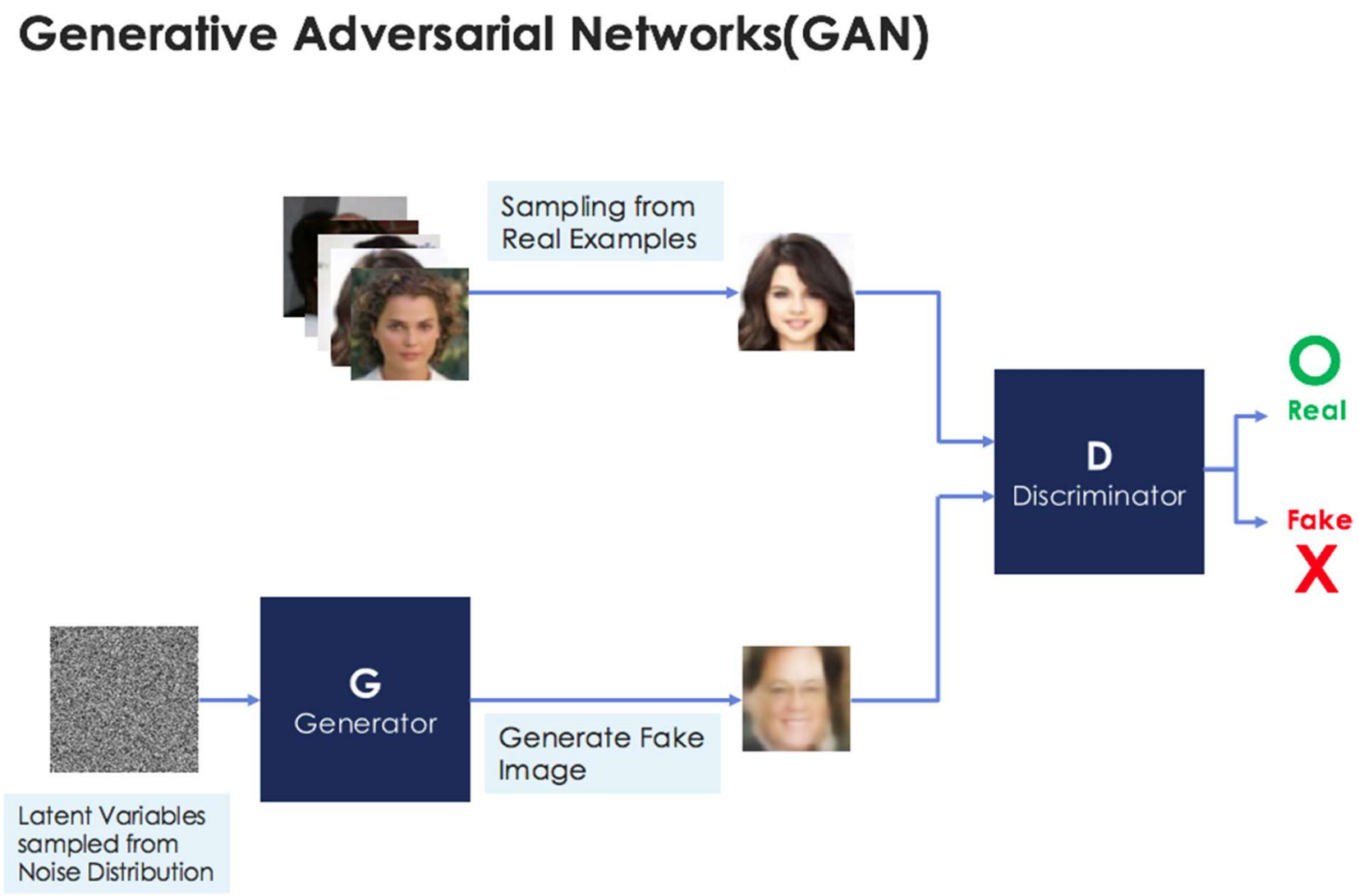
시계열 예측

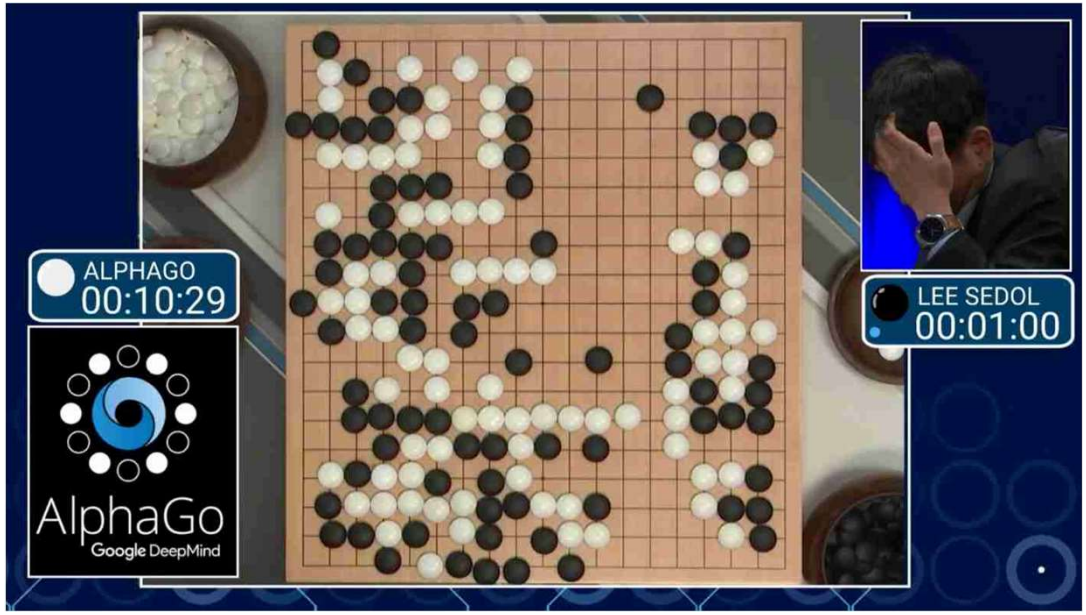


Autoencoder
(특성추출, 차원감소)



GAN(적대적 생성모델)


Generative Adversarial Networks(GAN)
Sampling from
Real Examples
D
Discriminator
Real
Fake
G
Generator
Generate Fake
Image
Latent Variables
sampled from
Noise Distribution

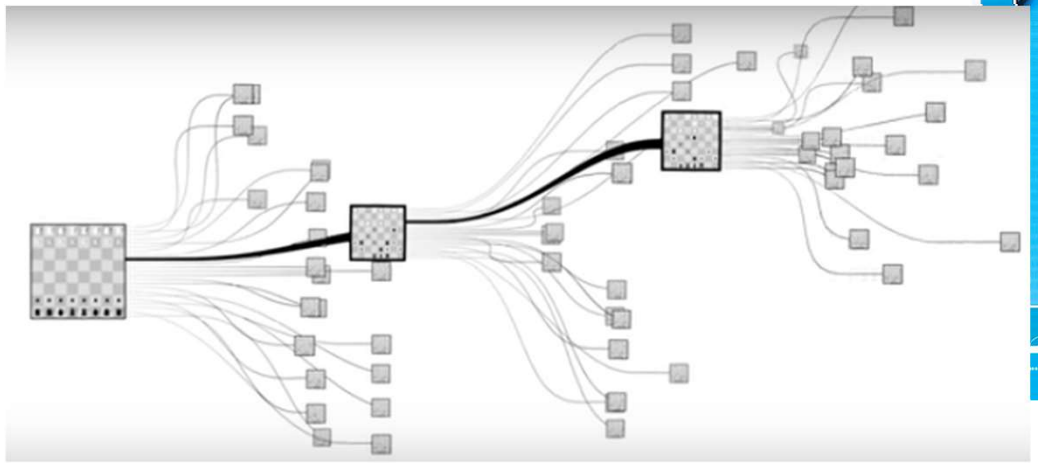


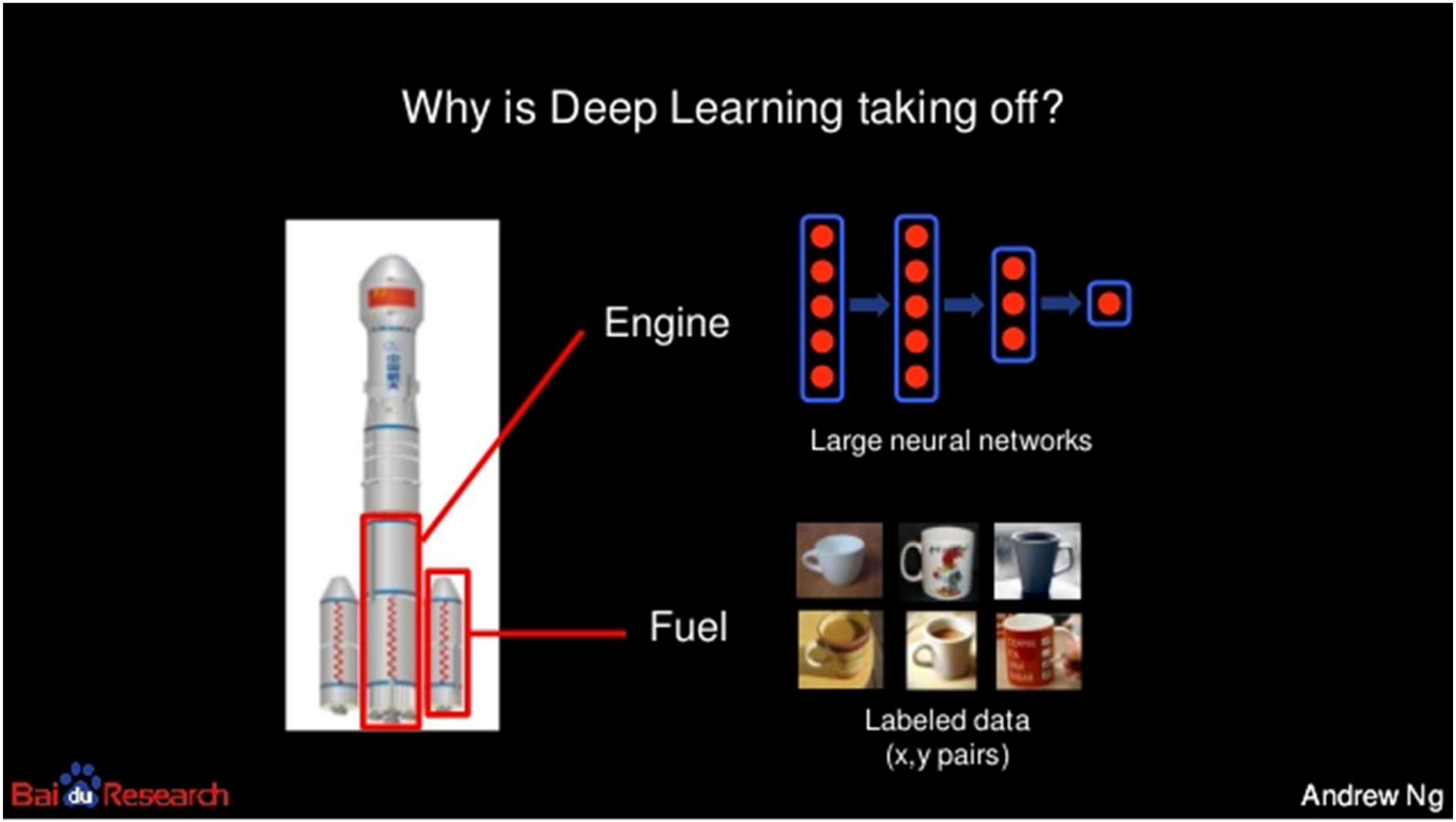
⇨ **Alphago-Lee**

Alphago-Master

⇩

Deep Learning 과
강화학습의 만남

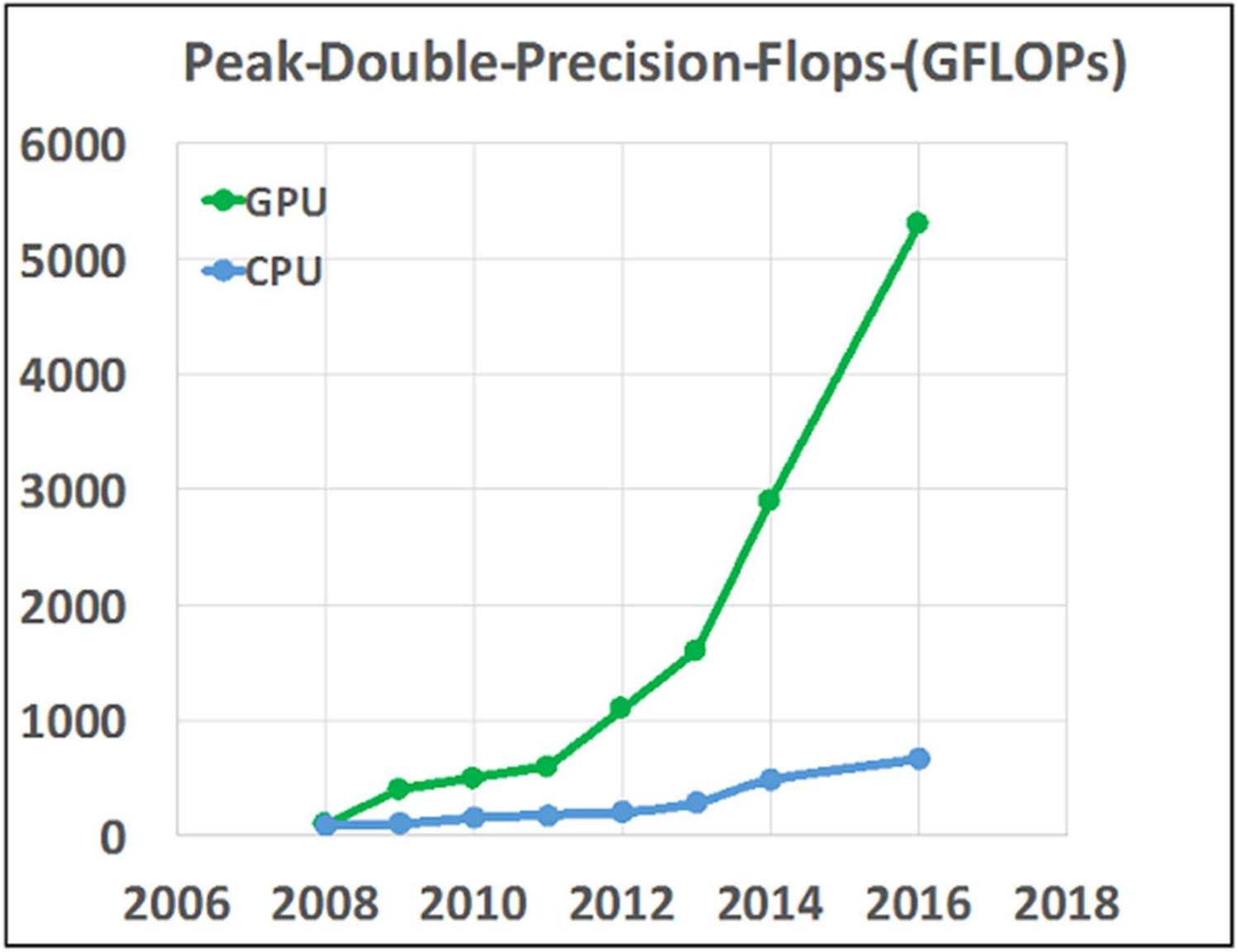




Rocket Engine : NVIDIA + Deep Learning Algorithm

Fuel : Data (25,000 pictures for cat)

CPU vs GPU Performance

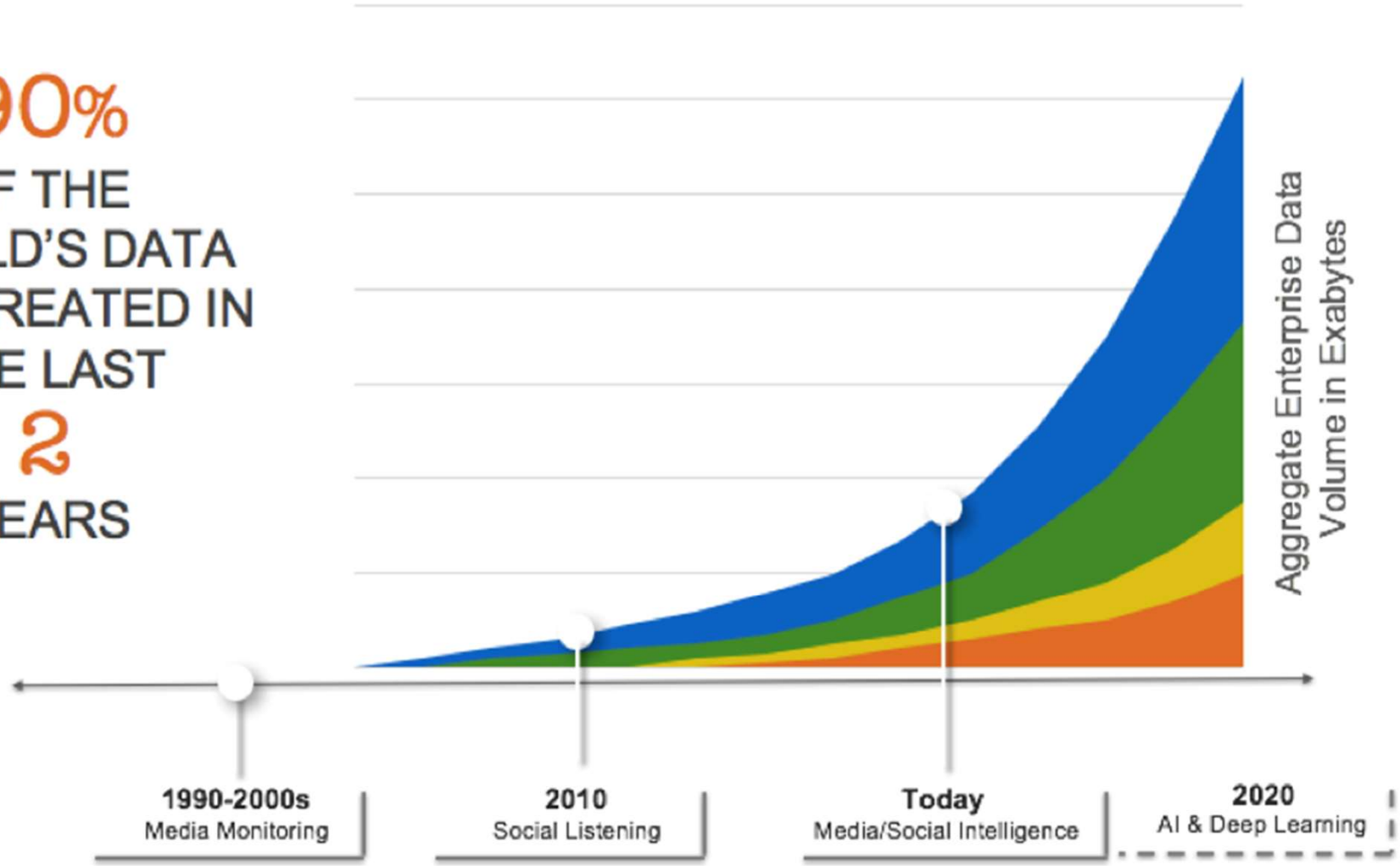


90%

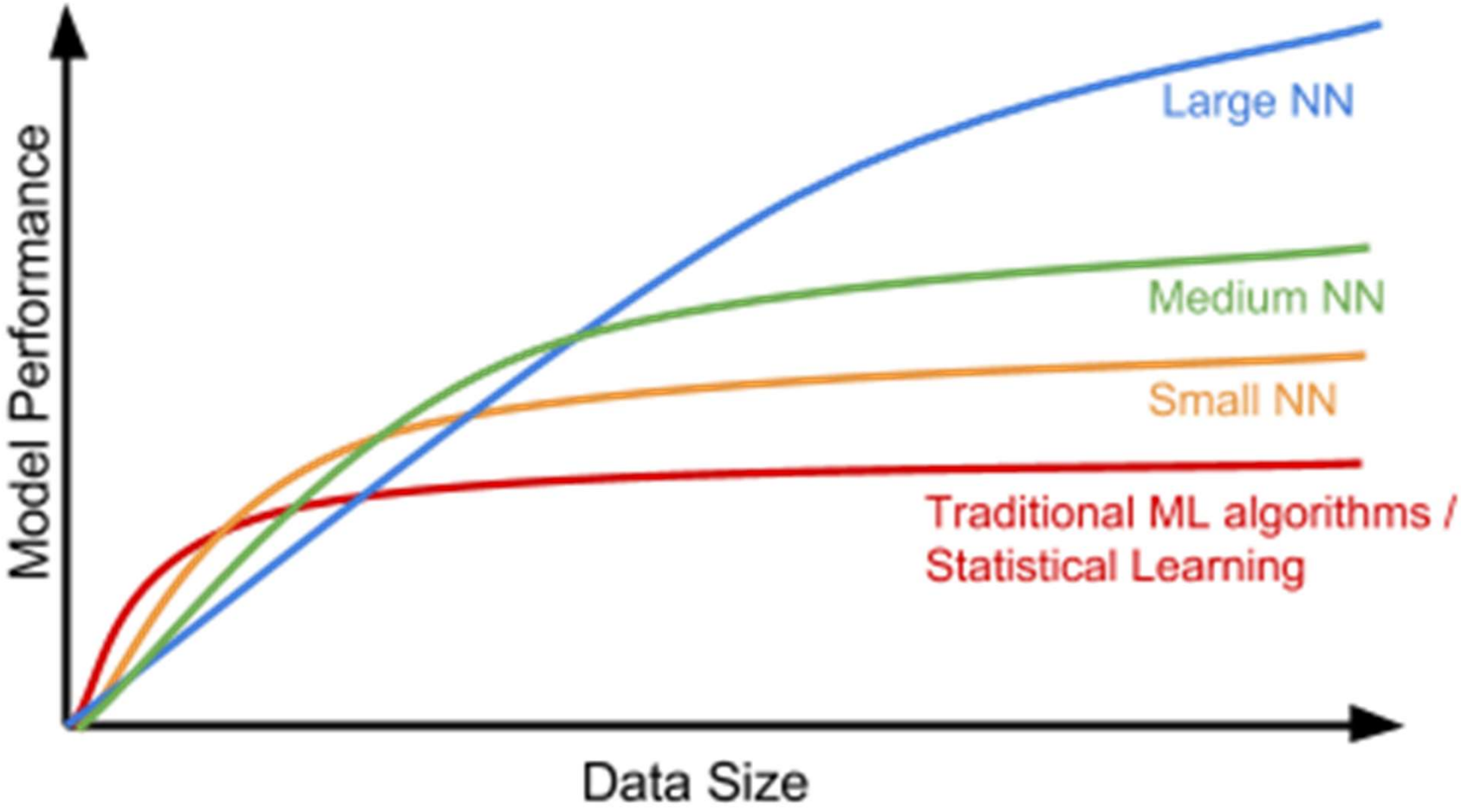
OF THE WORLD'S DATA WAS CREATED IN THE LAST

2

YEARS

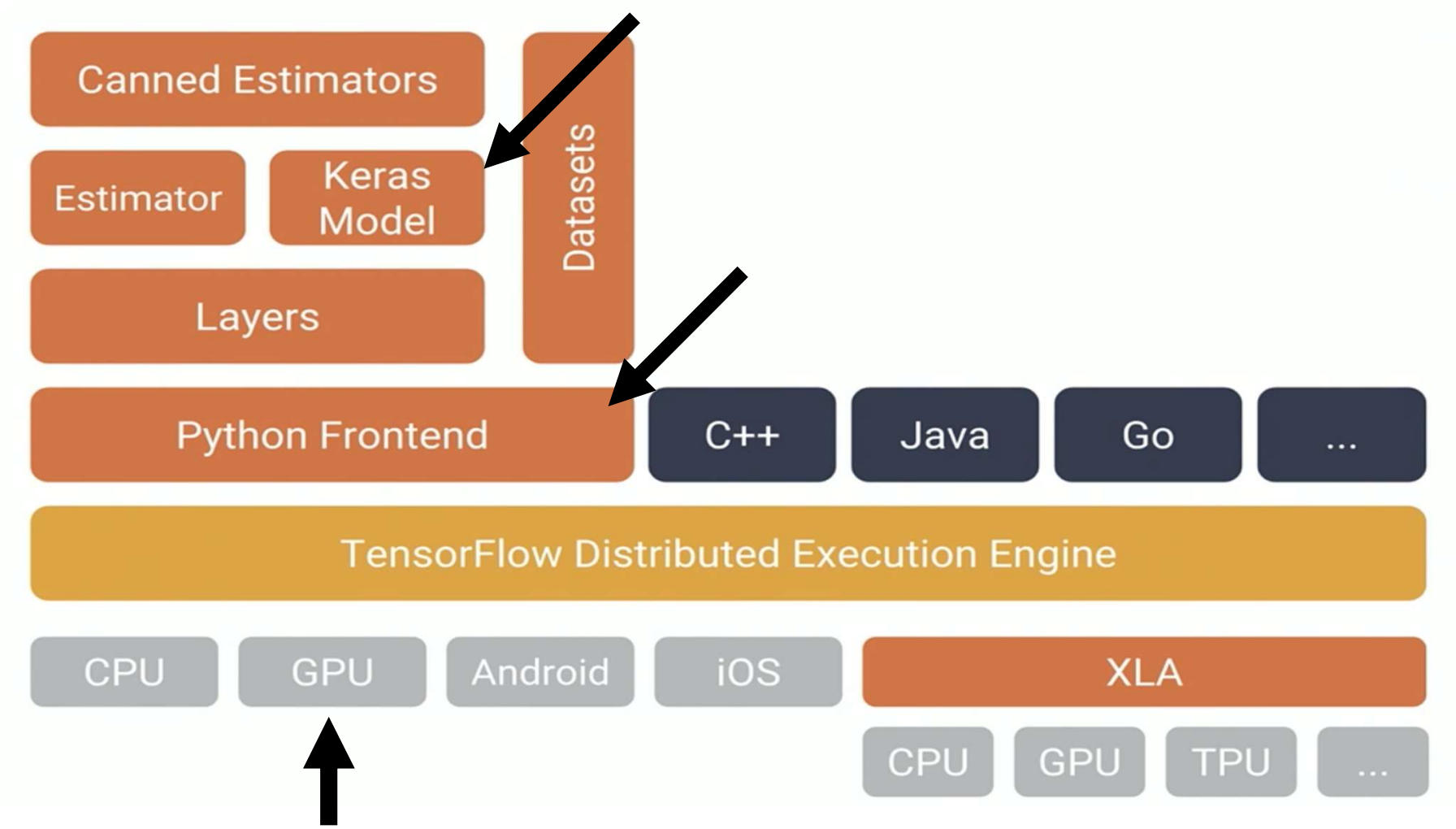


전통적 Machine Learning vs. Deep Neural Network

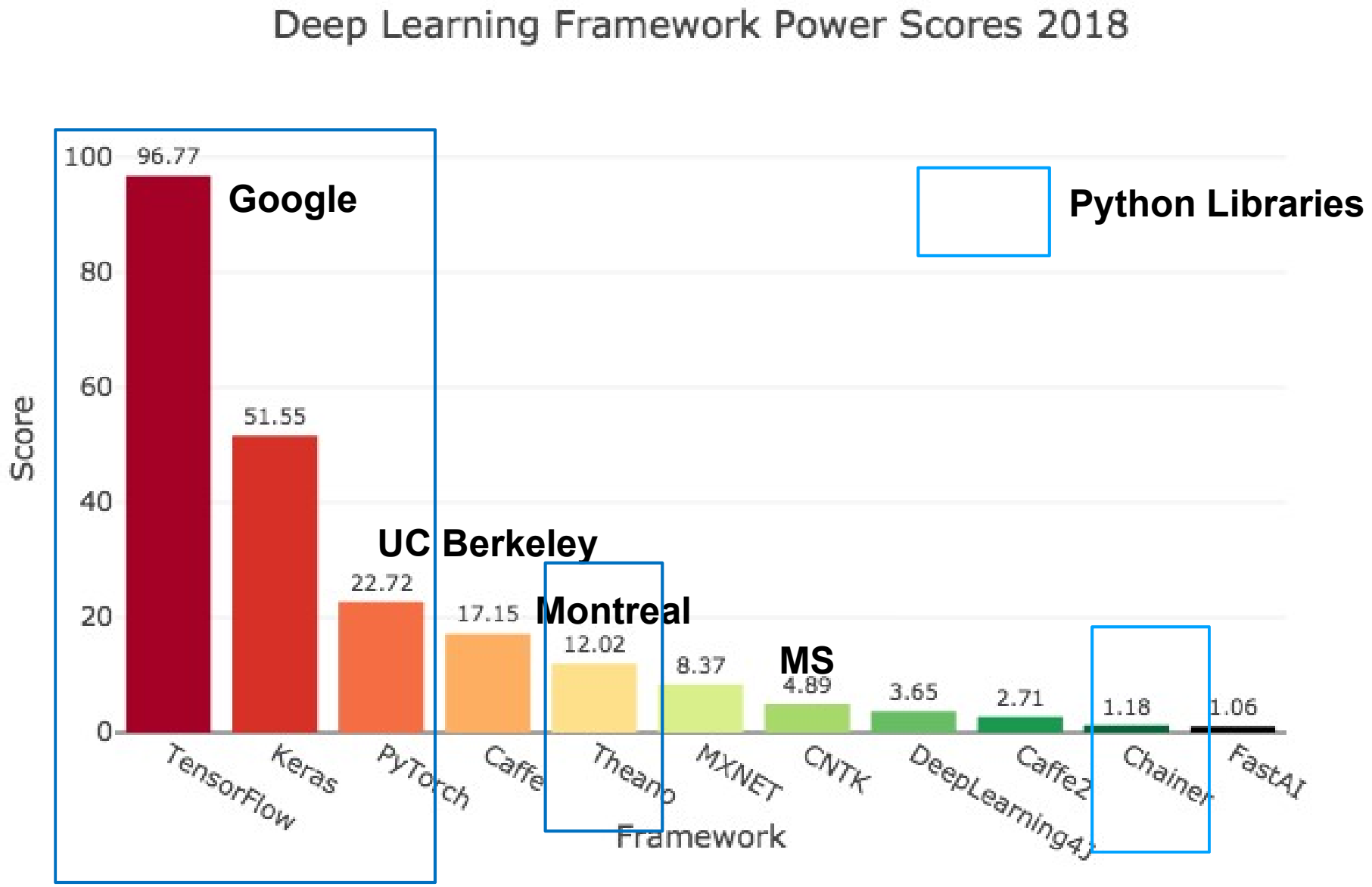


Google open source (2015년 11 월)

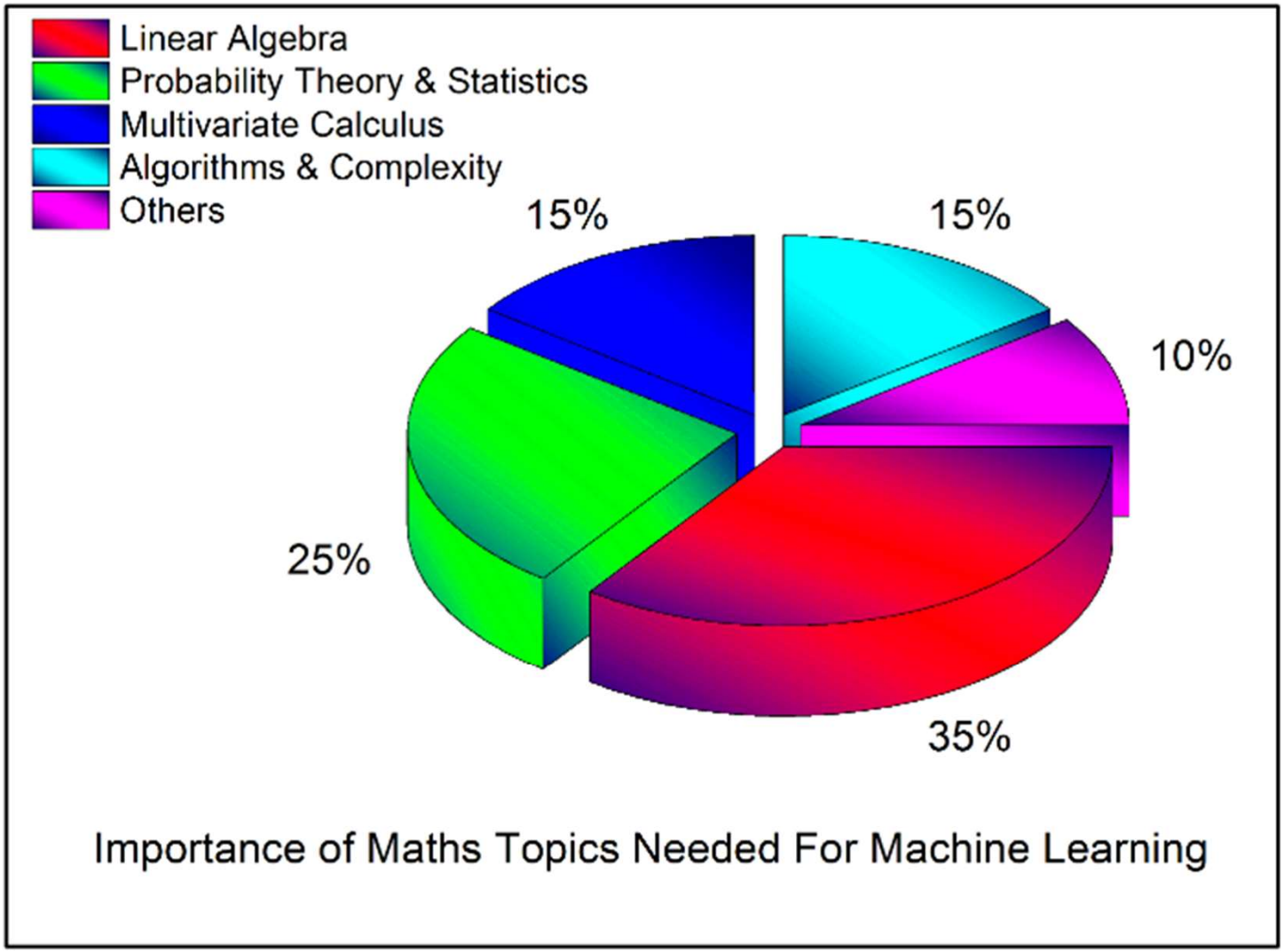
Version 2.1 : Keras 공식 포함



Kaggle 사용 순위


Deep Learning Framework Power Scores 2018
Score
100
80
60
40
20
0
96.77
51.55
22.72
17.15
12.02
8.37
4.89
3.65
2.71
1.18
1.06
TensorFlow
Keras
PyTorch
Caffe
Theano
MXNET
CNTK
DeepLearning4j
Caffe2
Chainer
FastAI
Framework
Google
UC Berkeley
Montreal
MS
Python Libraries

Machine Learning 학습에 필요한 수학 지식



Linear Algebra (선형대수학)

We have

$$a_1 = f(W_{11}x_1 + W_{12}x_2 + W_{13}x_3 + b_1)$$

$$a_2 = f(W_{21}x_1 + W_{22}x_2 + W_{23}x_3 + b_2)$$

etc.

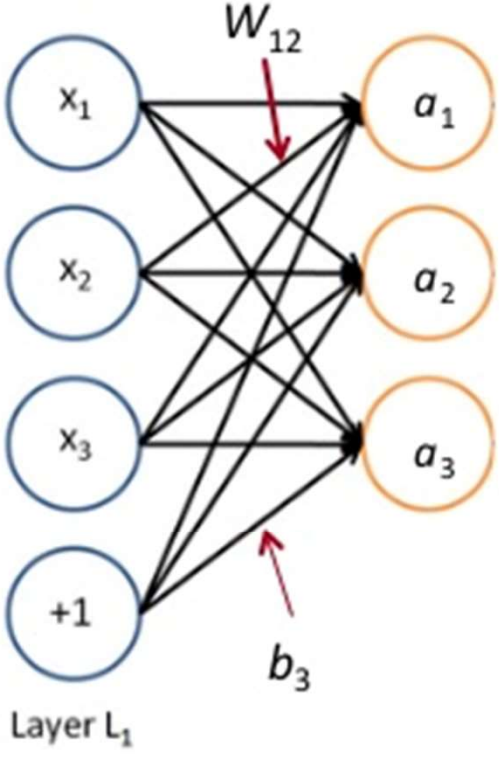
In matrix notation

$$z = Wx + b$$

$$a = f(z)$$

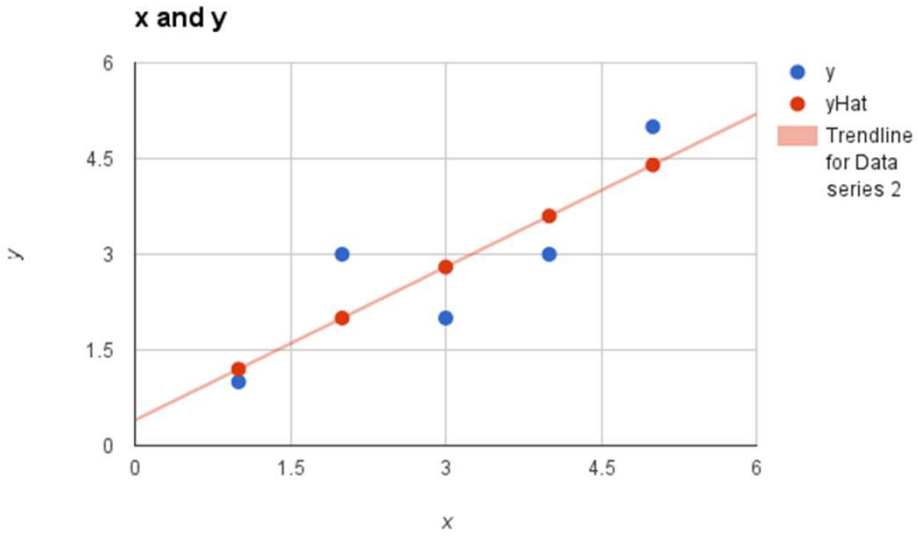
where f is applied element-wise:

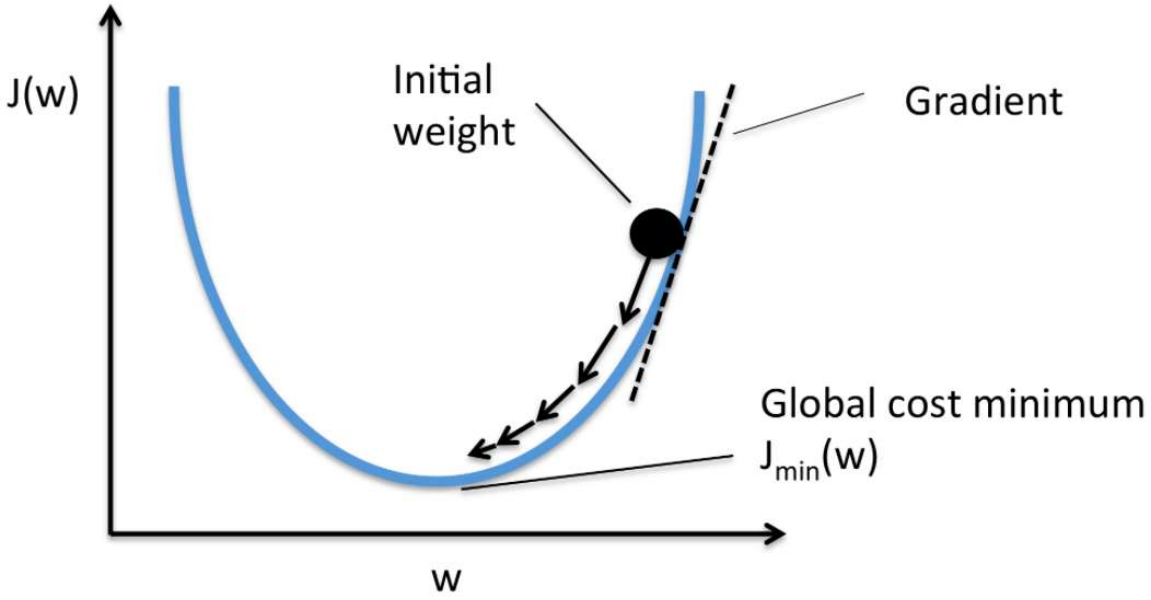
$$f([z_1, z_2, z_3]) = [f(z_1), f(z_2), f(z_3)]$$

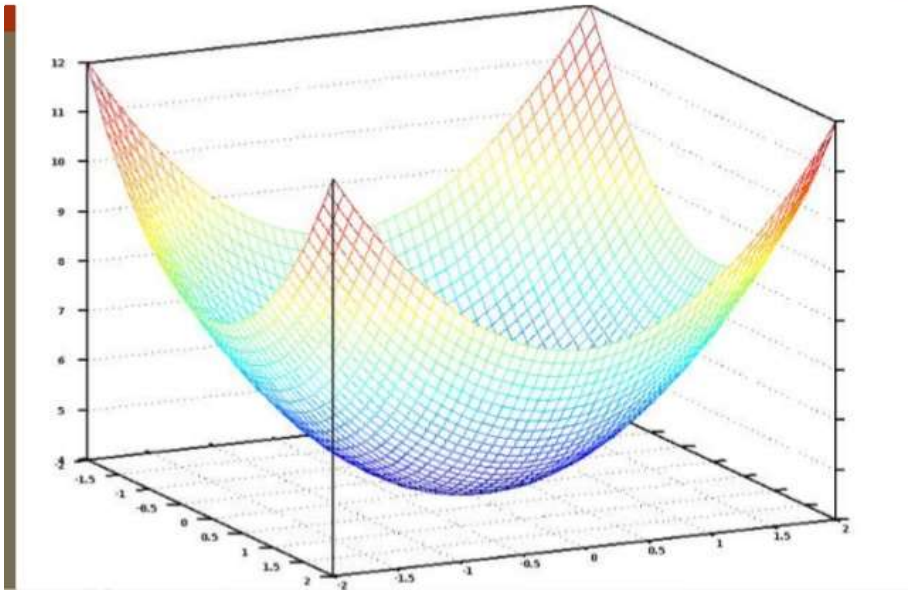


Multivariate Calculus (다변수미분)

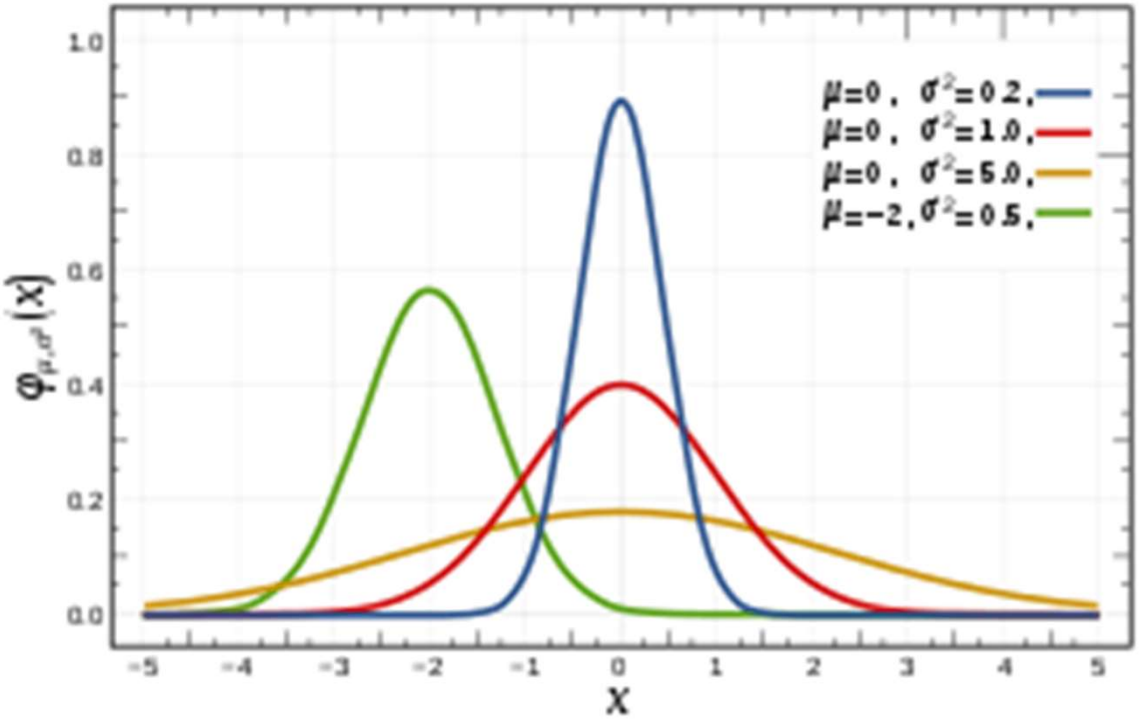
$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$$

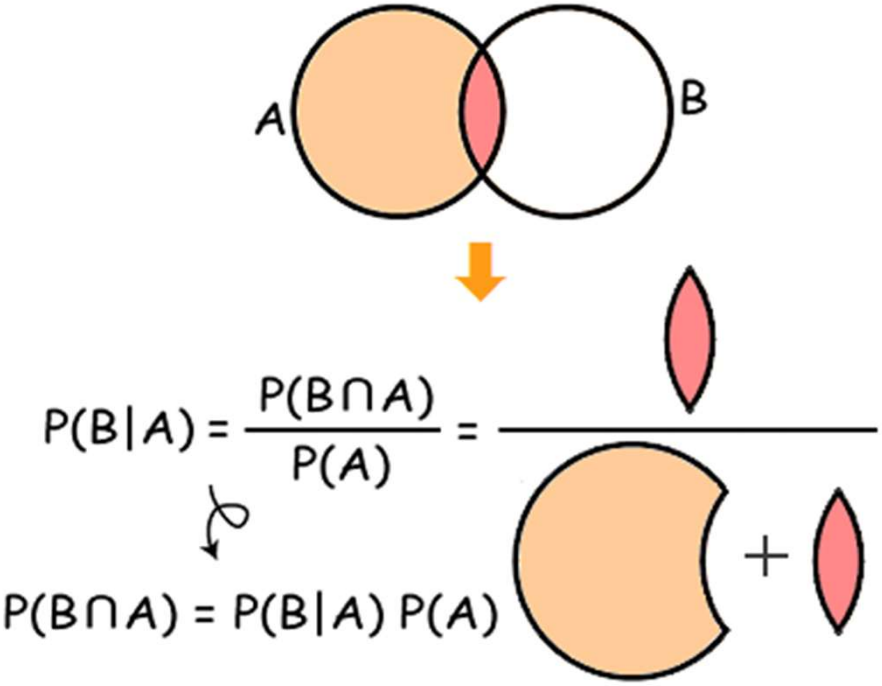




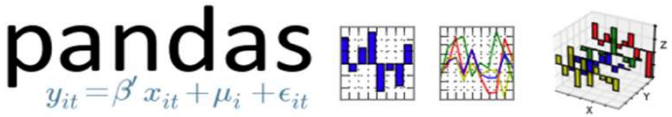


Probability/Statistics
(확률/통계)

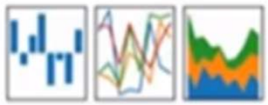




Python Libraries – ML 관련



1. **Scientifics Computing** Libraries



Pandas
(Data structures & tools)

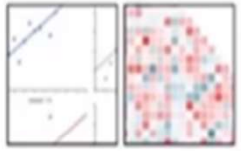
NumPy
(Arrays & matrices)

SciPy
(Integrals, solving differential equations, optimization)

2. Visualization Libraries

Matplotlib

(plots & graphs, most popular)



Seaborn

(plots : heat maps, time series, violin plots)

3. **Algorithmic** libraries

Scikit-learn
(Machine Learning : regression, classification,…)

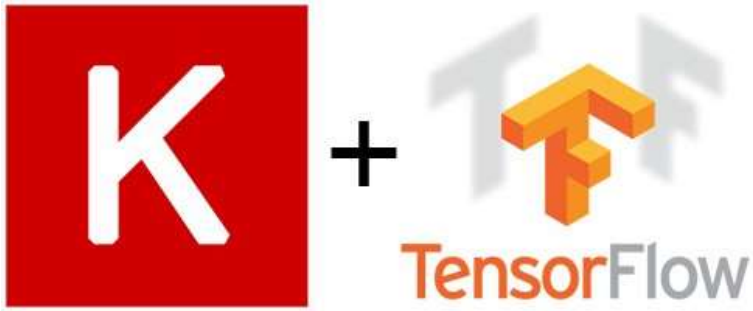
Statsmodels
(Explore data, estimate statistical models, and perform statistical tests.)

Machine Learning

전통적 Machine Learning Tool:

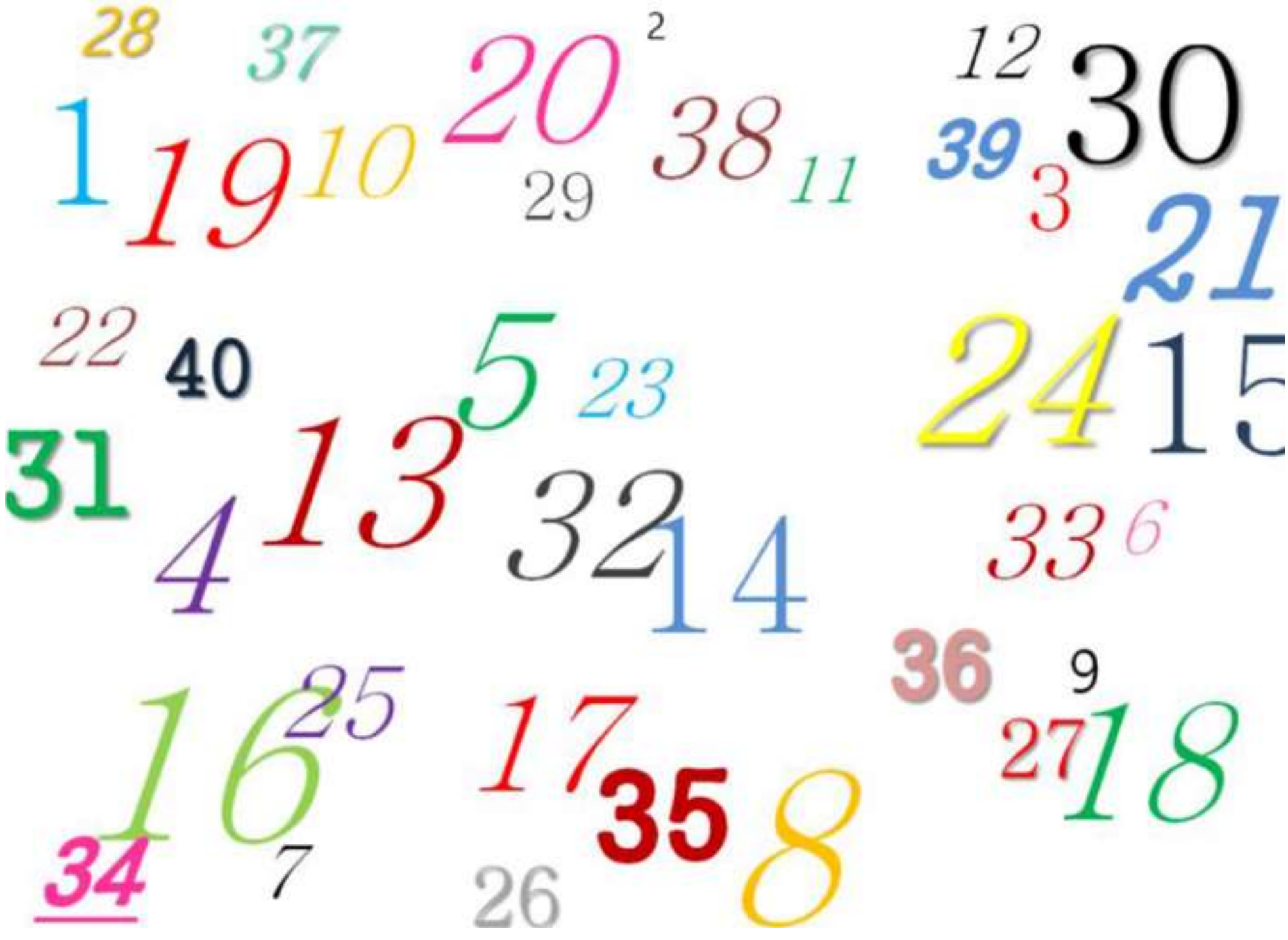
- 벤치마크용 데이터셋 예제
- 데이터 전처리(preprocessing)
- 지도 학습(Supervised learning)
- 비지도 학습(Unsupervised learning)
- 모형 평가 및 선택 (evaluation and selection)

Deep Learning



Deep Learning Tool:

- Pre-train model, Sample Dataset 제공
- Deep Learning 에 필요한 각종 함수 제공
- GPU support
- 각종 language 지원 API 제공


28
37
20
2
12
30
1
19
10
38
11
39
3
29
21
22
40
5
23
24
15
31
4
13
32
14
33
6
36
9
16
25
17
27
18
35
8
34
7
26


28
37
1
19
10
20
2
38
11
29
12
30
39
3
21
22
40
31
4
13
5
23
32
14
24
15
33
6
36
9
27
18
16
25
34
7
17
35
8
26

과정 SCOPE

- Crash Course – Numpy, Pandas, Matplotlib
- Linear Regression (선형회귀)
- non-Linear Regression (비선형회)
- KNN (K-Nearest Neighbor)
- Decision Tree
- Logistic Regression
- SVM (Support Vector Machine)
- Random Forest

- K-Means Clustering
- Tensorflow 소개
- Keras
- Simple Neural Network
- CNN (Convolutional Neural Network)
- RNN
- 실습문제

Crash Course

1. Numpy
2. Pandas
3. Matplotlib

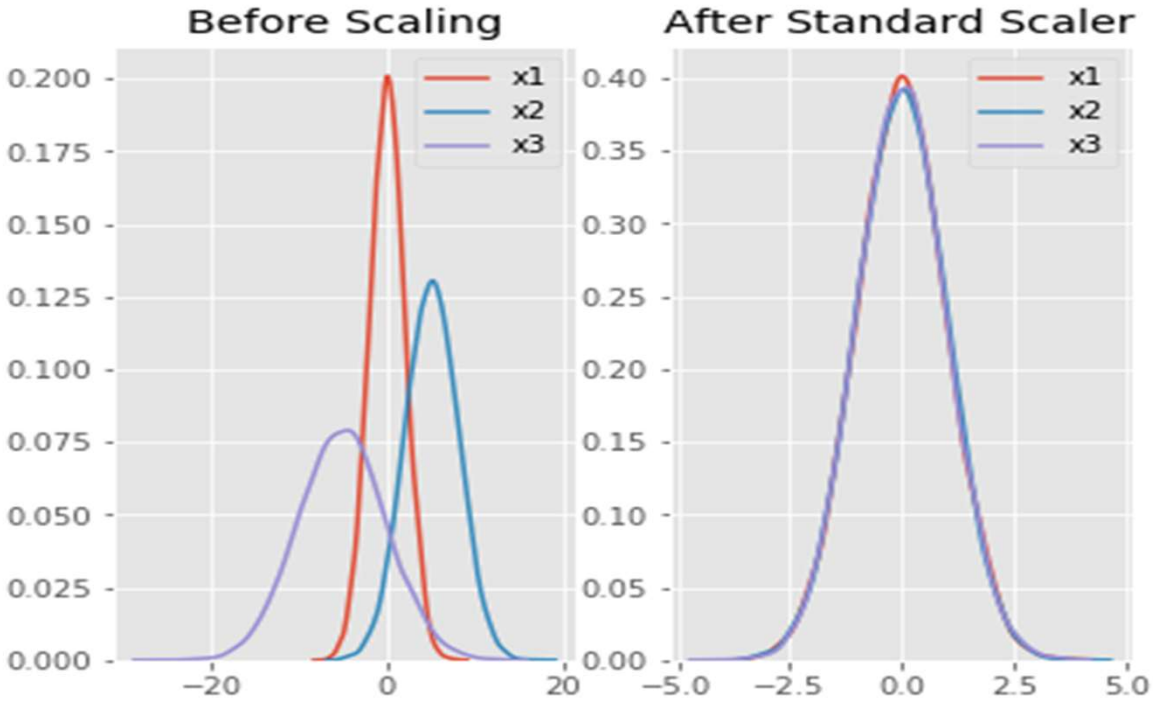
Feature Scaling

- Raw data 를 전 처리하여 input data 의 구간을 표준화.
- Standard Scaling :

  $z = (x - \mu) / s$ (μ : 평균, s : 표준편차)

- Minmax Scaling :

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$



실습: sklearn 을 이용한 scaling

- Standard Scaling

- MinMax Scaling