

2022

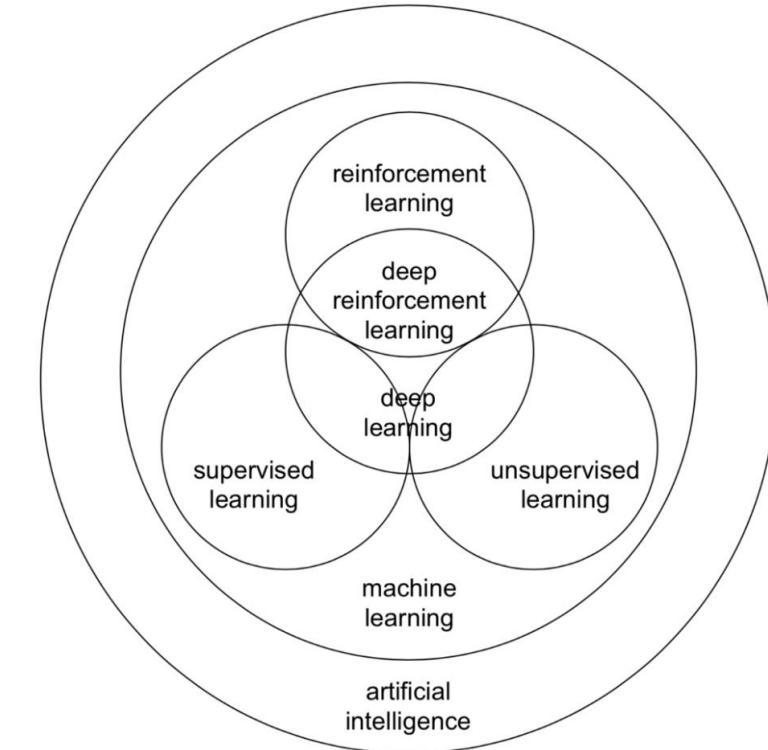
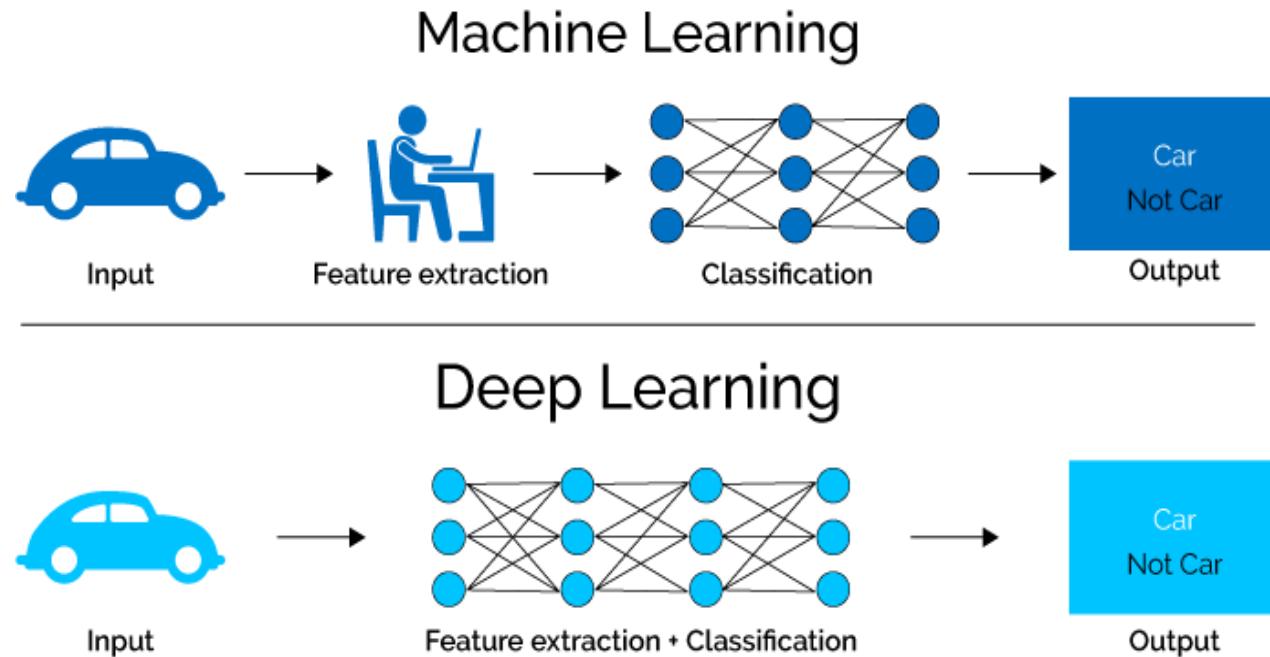
스마트교통 빅데이터 분석

인공지능과 딥러닝 소개



인공지능 소개

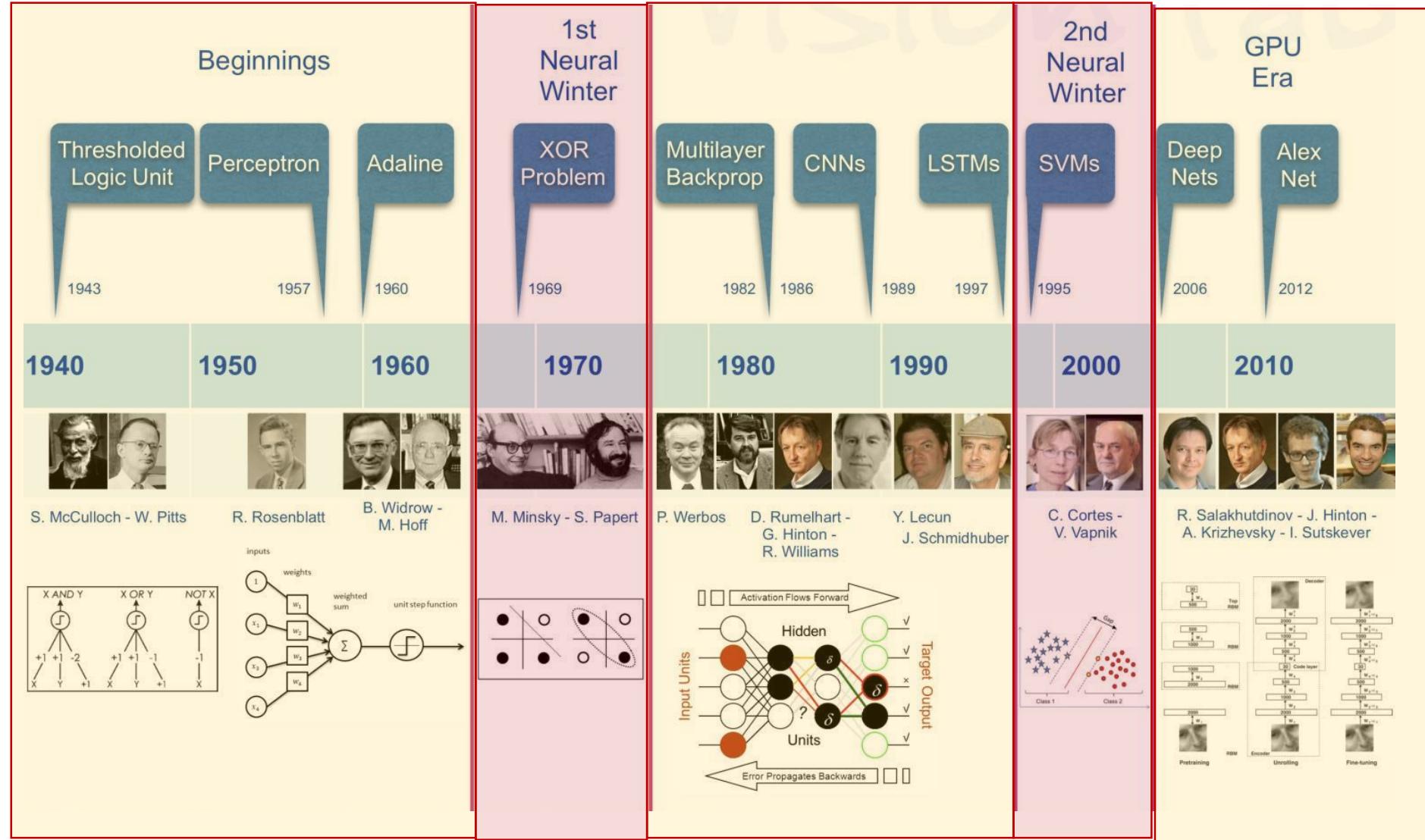
- ❖ 인공지능의 역사에 대하여 이해한다.
 - ✓ 딥러닝과 머신러닝의 관계
 - ✓ 지도학습, 비지도학습, 강화학습 구분
 - ✓ 지도학습을 위한 데이터 특성과 라벨의 중요성을 이해한다.
- ❖ 인공지능을 구현하기 위한 SW 프레임워크
 - ✓ Sklearn
 - ✓ Tensorflow
 - ✓ Pytorch
 - ✓ Keras
 - ✓ 등



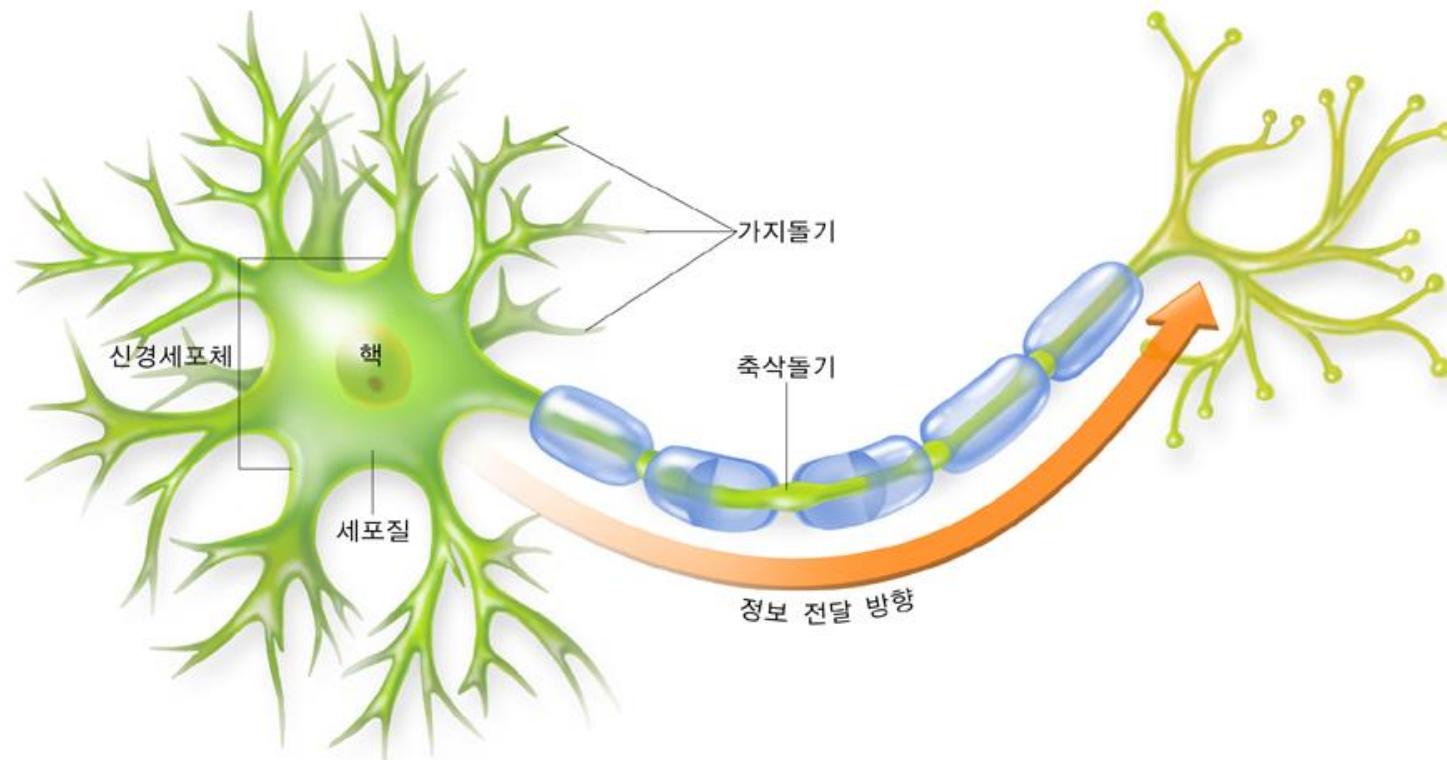
<https://www.xenonstack.com/blog/data-science/log-analytics-deep-machine-learning-ai/>

Yuxi Li, Deep Reinforcement Learning, arXiv, 2018

History of Artificial Intelligence

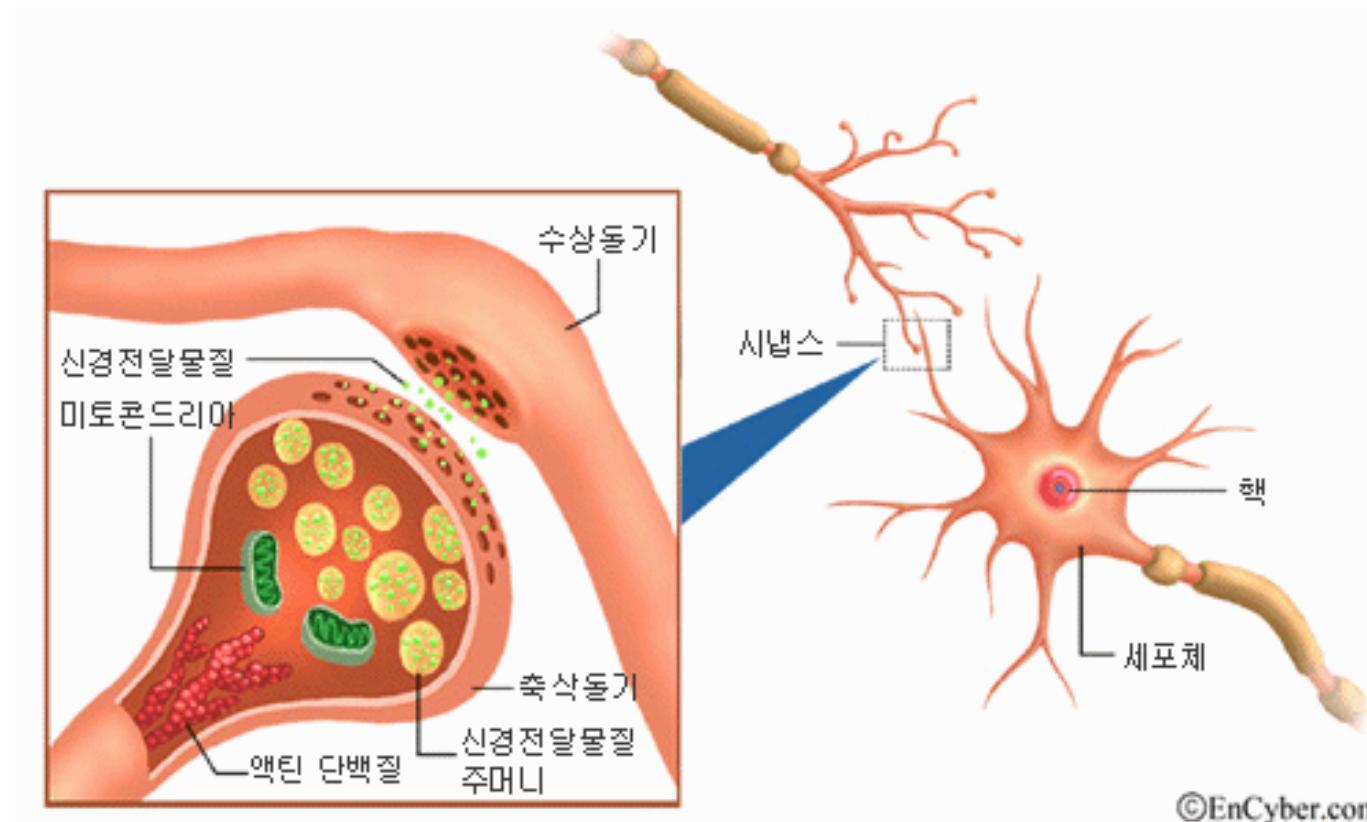


사람의 자극의 전달 과정과 시냅스



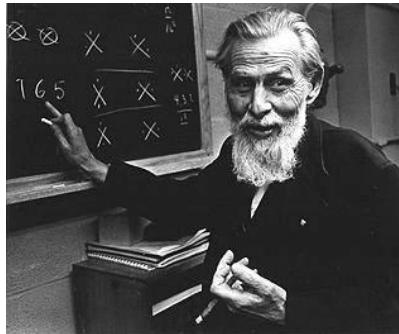
시냅스 (Synapse): 자극의 전달

뉴런의 집합체로 한 뉴런의 축삭돌기 말단과 다음 뉴런의 수상돌기 사이의 연접 부위



❖ 인공신경망 개념 최초 제안 : 맥컬록-피츠 신경망 모델 (1943)

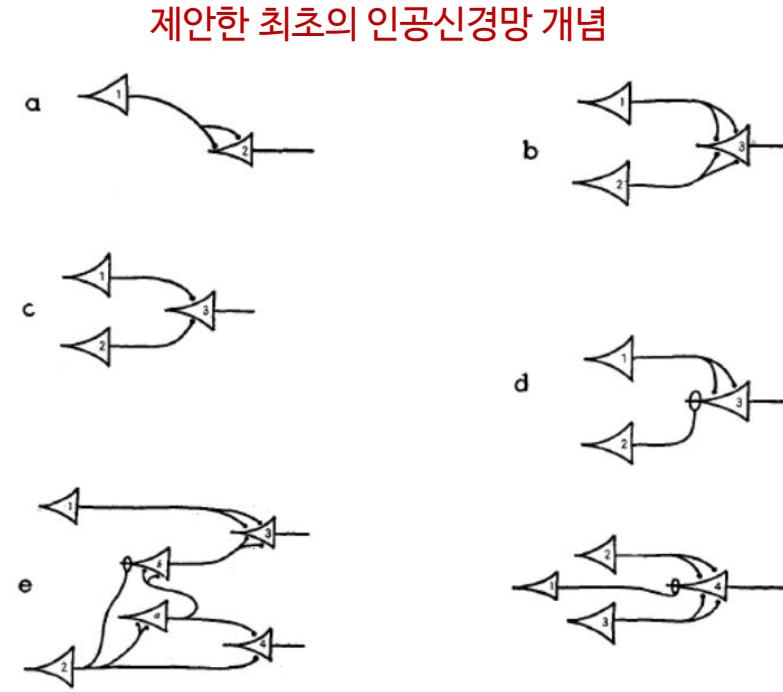
- ✓ McCulloch와 Pitts가 1943년에 ANN 최초 논문이 발표
 - “A logical calculus of the ideas immanent in nervous activity”
- ✓ 인간의 신경 구조를 복잡한 스위치들이 연결된 네트워크로 표현할 수 있다



맥컬록

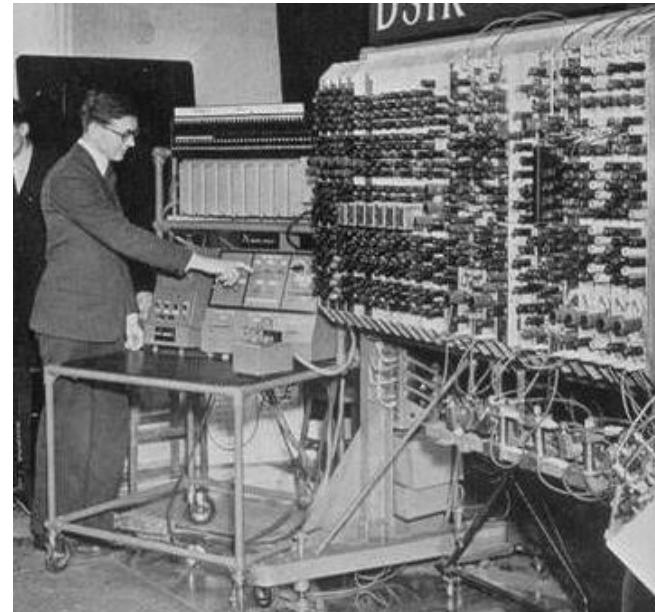
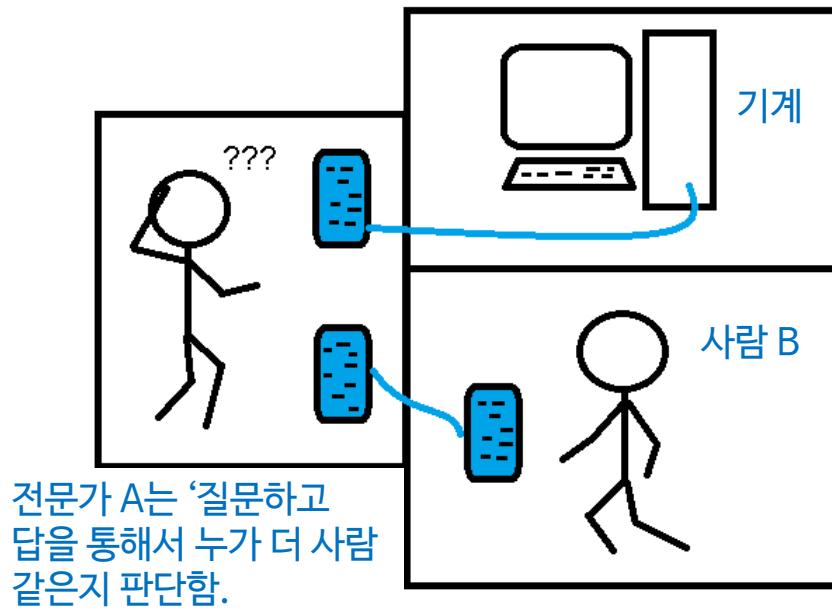


피츠



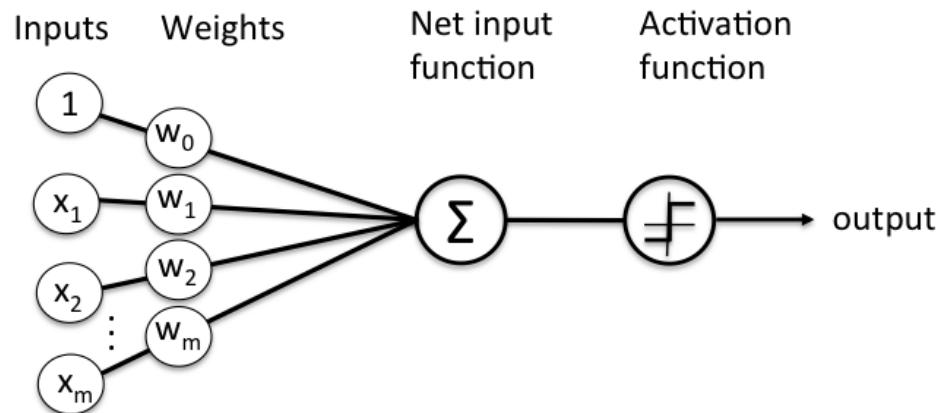
❖ 튜링 테스트 (1950)

- ✓ 앨런 튜링(Turing)은 컴퓨터 과학의 아버지로 불리움.
- ✓ 기계가 사람처럼 생각할 수 있다는 것을 아래 그림처럼 테스트 함.
- ✓ 사람 A가 상태에서 2명과 대화를 했을 때, 기계(Z)가 더 자연스러움.

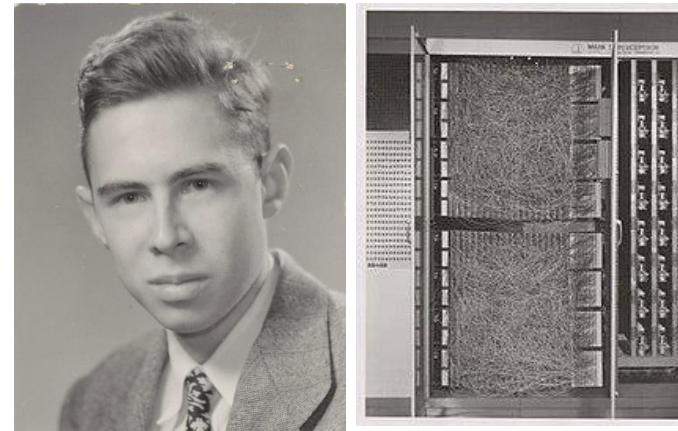


❖ ANN은 퍼셉트론 신경망 설명 : 프랑크 로젠블라트, 1958)

- ✓ Frank Rosenblatt는 퍼셉트론(Perceptron)라는 선형분류 피드포워드 신경망
 - “The perceptron: A probabilistic model for information storage and organization in the brain.” 논문에서 제시
 - 입력과 가중치(weight)들의 곱을 모두 더한 뒤 활성함수(계단함수)로 선형 분류기
- ✓ (문제점) 1개 퍼셉트론으로 XOR 선형 분류를 설명할 수 없음.



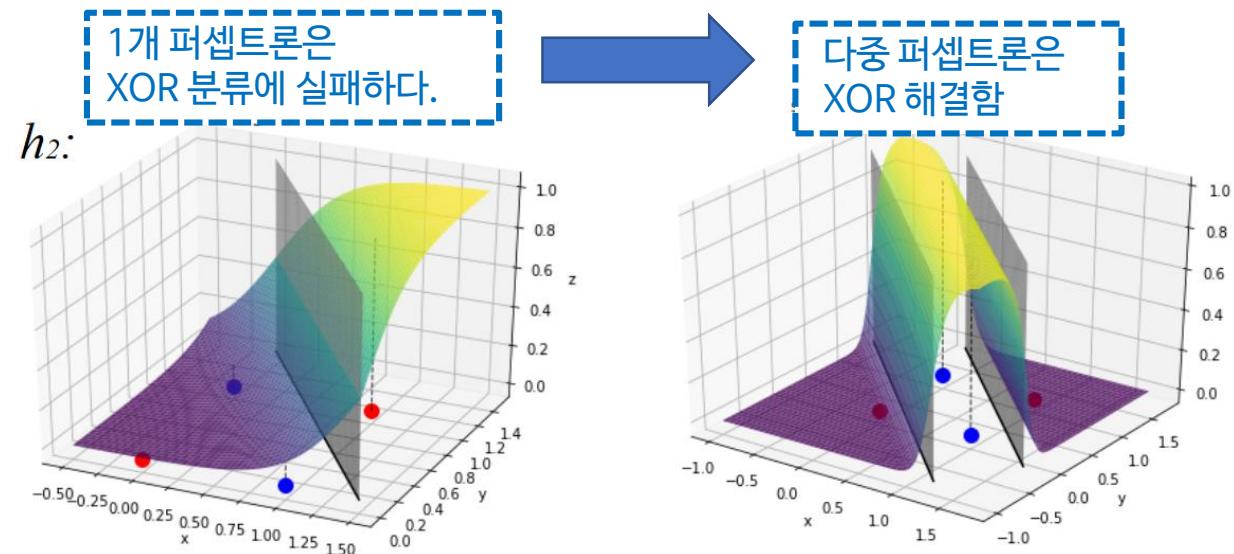
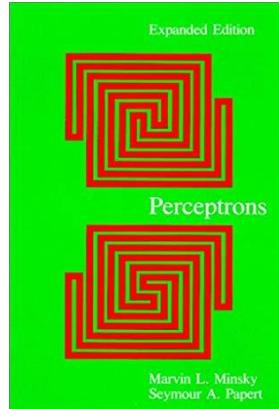
퍼셉트론은 현대의 딥러닝의 기초이다. 그 당시, 퍼셉트론을 통해서 진짜 인간과 같은 인공지능을 만들 수 있다는 기대가 매우 컸다.



로젠블라트 IBM 퍼셉트론 계산기

❖ 퍼셉트론의 무용론 등장으로 1차 인공지능 겨울이 시작된다.

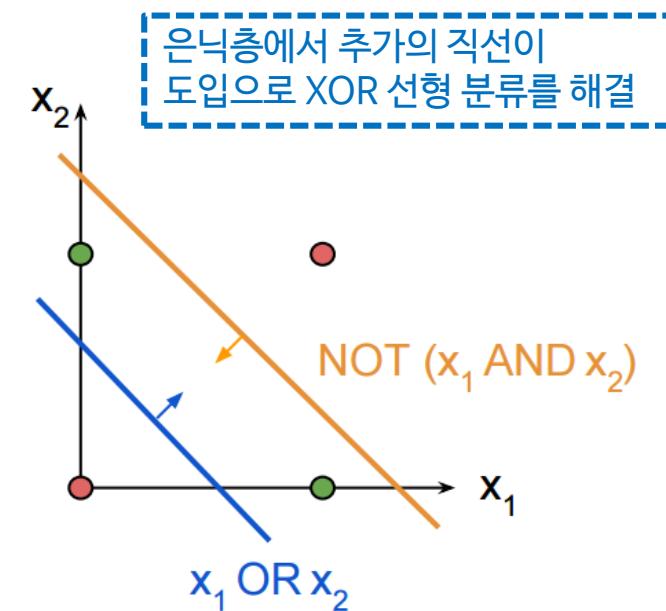
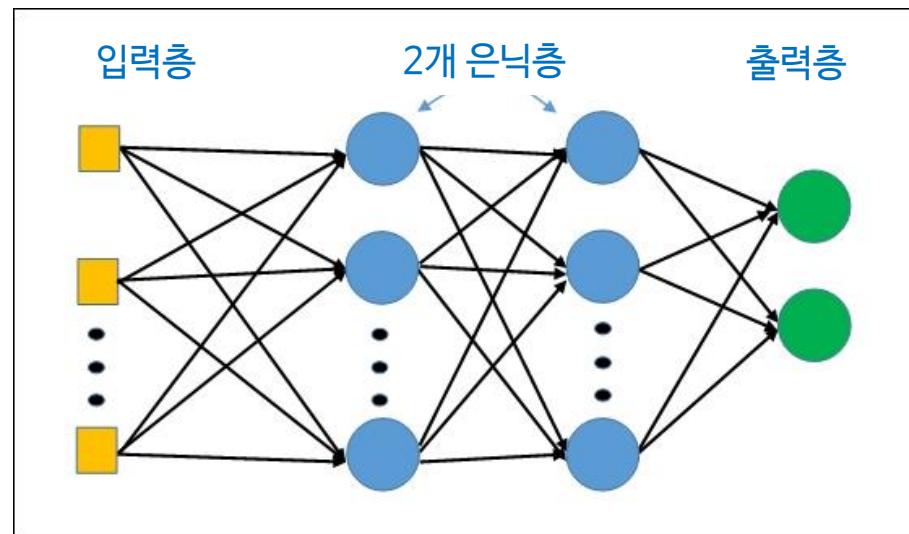
- ✓ 1969년 마빈 민스키는 '퍼셉트론'은 단순 선형 분류기이며, 'XOR' 분류도 할 수 없는 미미한 선형분류기라는 것을 수학적으로 증명함
- ✓ 퍼셉트론 인기가 사그라 들면서 인공지능 1차 암흑기가 도래한다.



1개의 퍼셉트론은 XOR 문제에서 빨강과 파랑색을 구분하는 초평면을 만들수 없다.

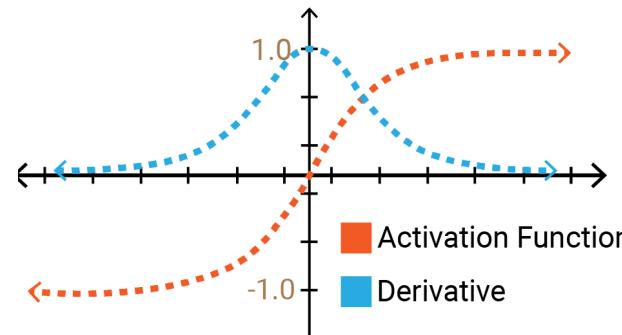
❖ 다층 퍼셉트론과 역전파 알고리즘 등장

- ✓ 다층 퍼셉트론은 전방향(feed-forward) 신경망으로 중간에 은닉층을 추가함
- ✓ 기존의 퍼셉트론이 선형 분류기라는 한계에 의해 XOR 문제를 해결할 수 없었다면,
- ✓ 다층 퍼셉트론은 은닉층(hidden layer)라는 중간 레이어를 추가로 XOR 문제를 해결
- ✓ (문제점) 다층 퍼셉트론은 은닉층의 추가로 신경망을 훈련에 많은 어려움이 있다.

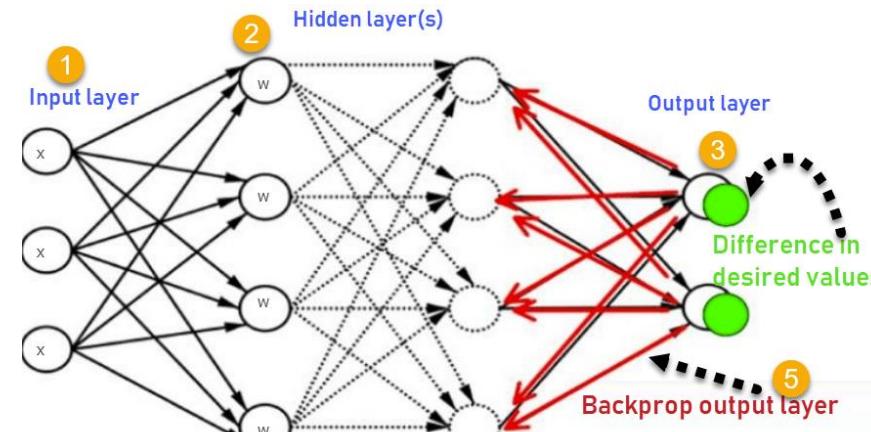


❖ 다층 퍼셉트론과 역전파 알고리즘 등장

- ✓ 1986년 McClelland, James L., David E. Rumelhart, and Geoffrey E. Hinton은 Backpropagation Algorithm을 제안해서 이 문제를 해결
- ✓ 오류 역전파 알고리즘은 Feedforward 연산 이후, 오차를 후방(Backward)으로 다시 보내 줌으로써, 많은 노드를 가진 MLP라도 최적의 가중치와 Bias를 학습할 수 있다.



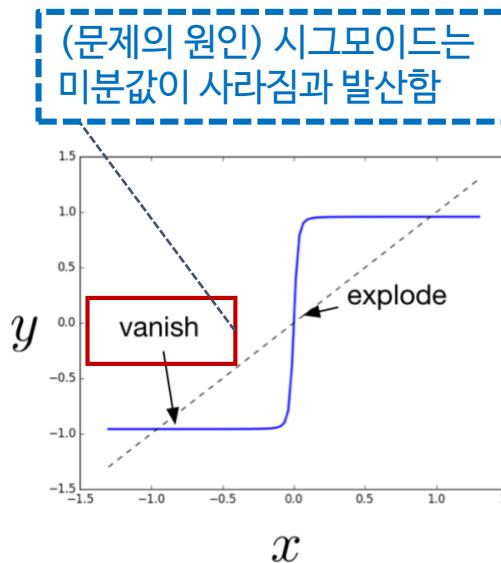
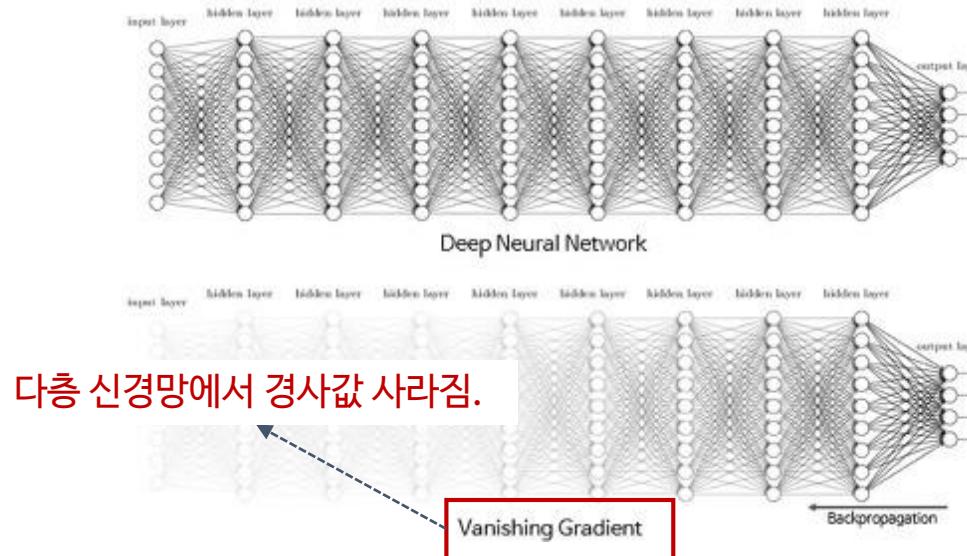
활성함수로 Sigmoid 함수와
미분 가능함.



역전파 알고리즘

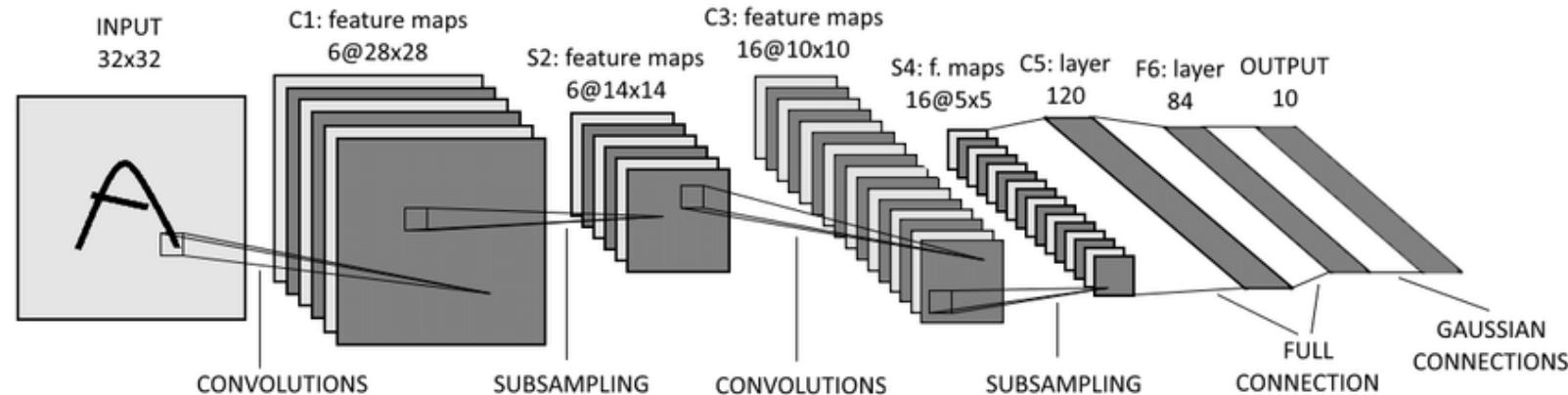
❖ 다층 퍼셉트론의 경사 발산과 소멸 문제 등장(1997)

- ✓ RNN에서 VGP(Vanishing Gradient Problem) 문제 (1993)
- ✓ LSTM(Long Short Term Memory)로 VGP 해결(1997), Hochreiter
- ✓ 이시기는 기계학습 SVM, Random Forest 등이 큰 인기가 있음.



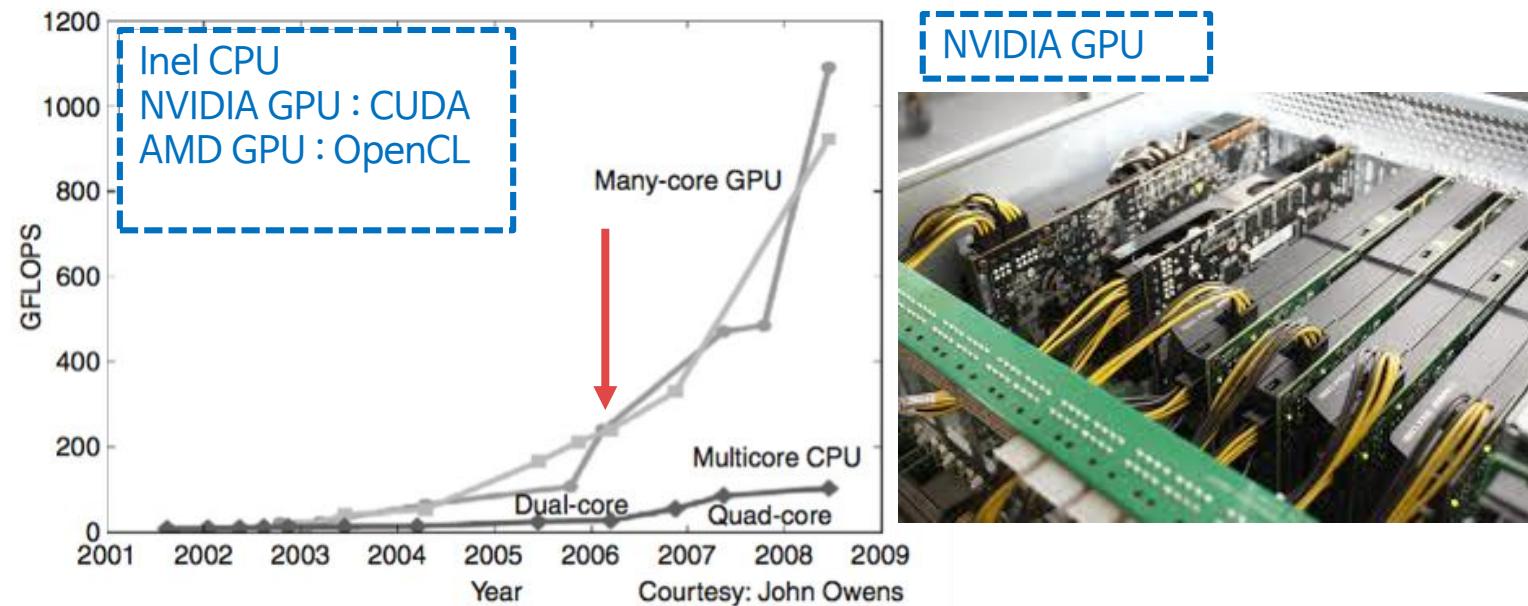
❖ 초기 합성곱 신경망(CNN)인 LeNet 신경망 성과(1998)

- ✓ 1998년 얀 루큰(Lecun)은 MLP와 역전파를 MNIST 이미지에 적용 성공
- ✓ LeNet-5라는 현대의 CNNs의 시초 제안



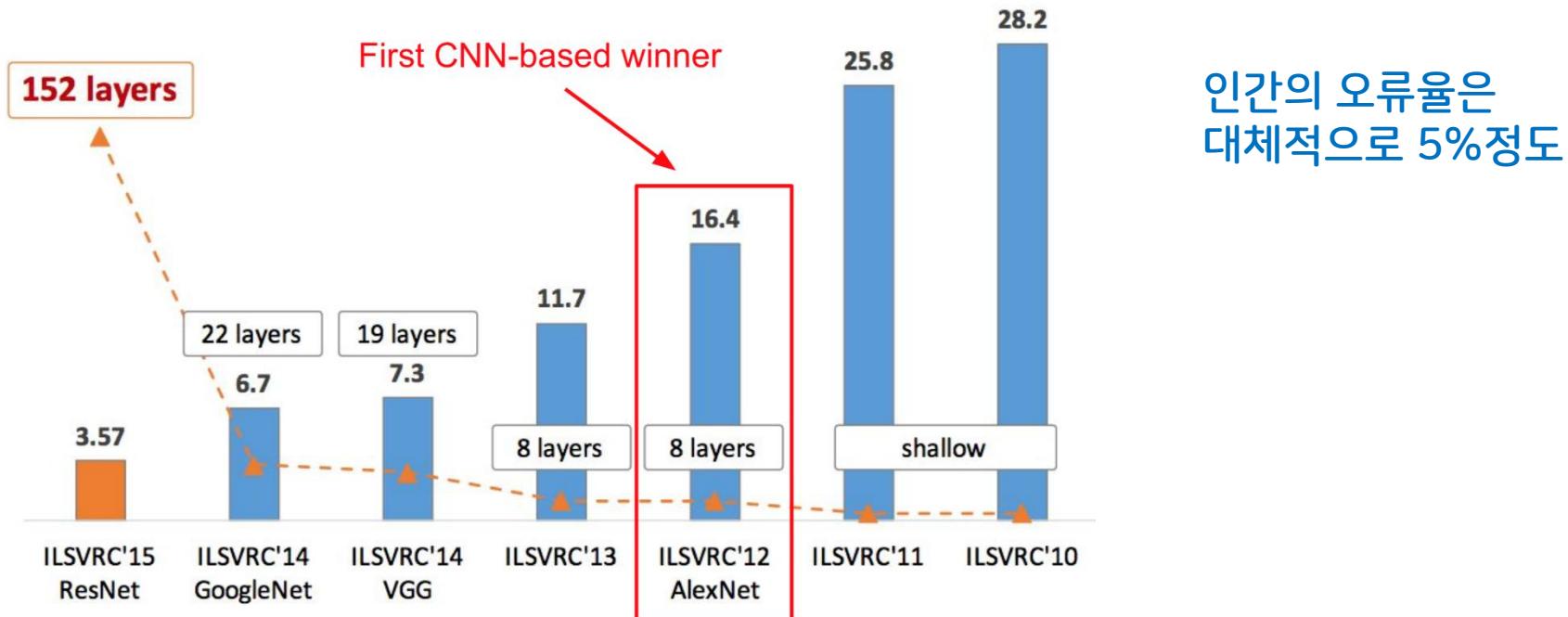
❖ 인공지능 2차 겨울 10년간 지속됨 (1993~2006)

- ✓ 과적합문제(Overfitting)
- ✓ Vanishing Gradient (다층 신경망 역전파 과정에서 기울기 값이 사라짐)
- ✓ 계산이 너무 느리다 (Too slow)
 - GPU, TPU, Xeon Phi 등 가속기 컴퓨팅 등장이 아직 안됨



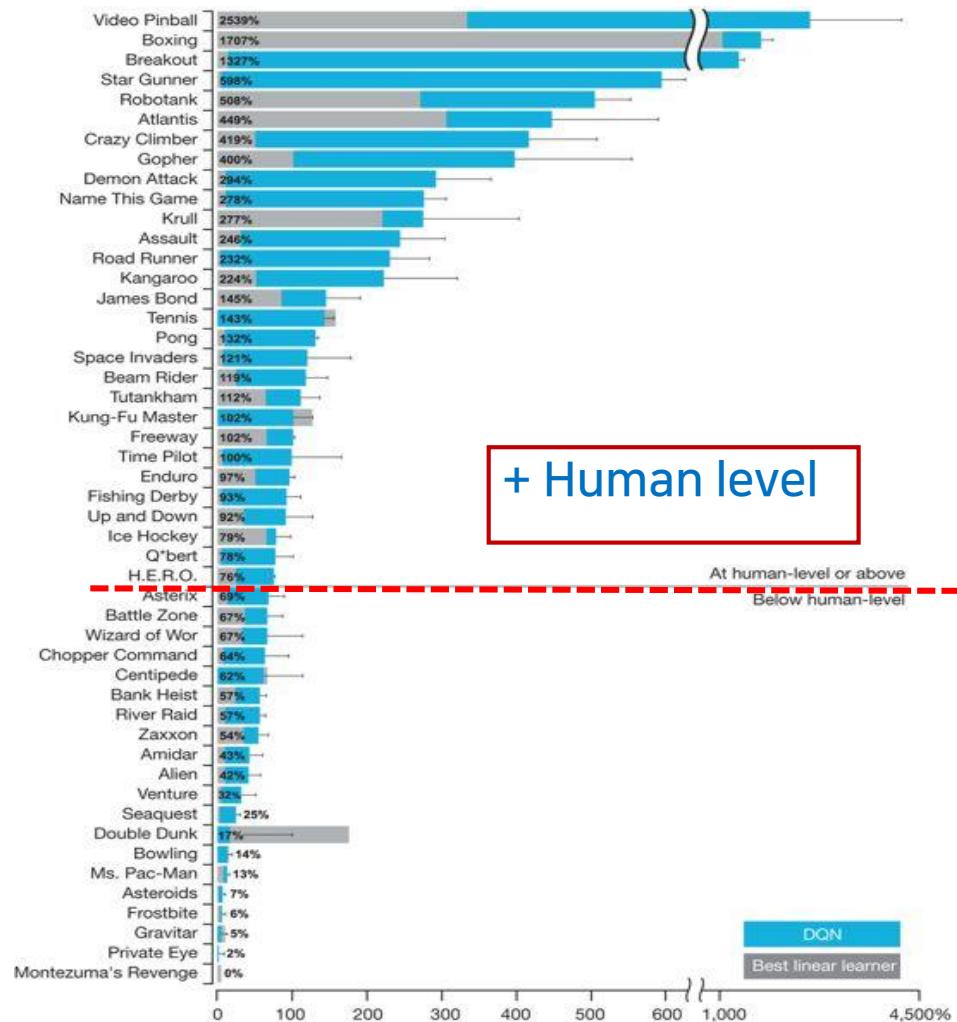
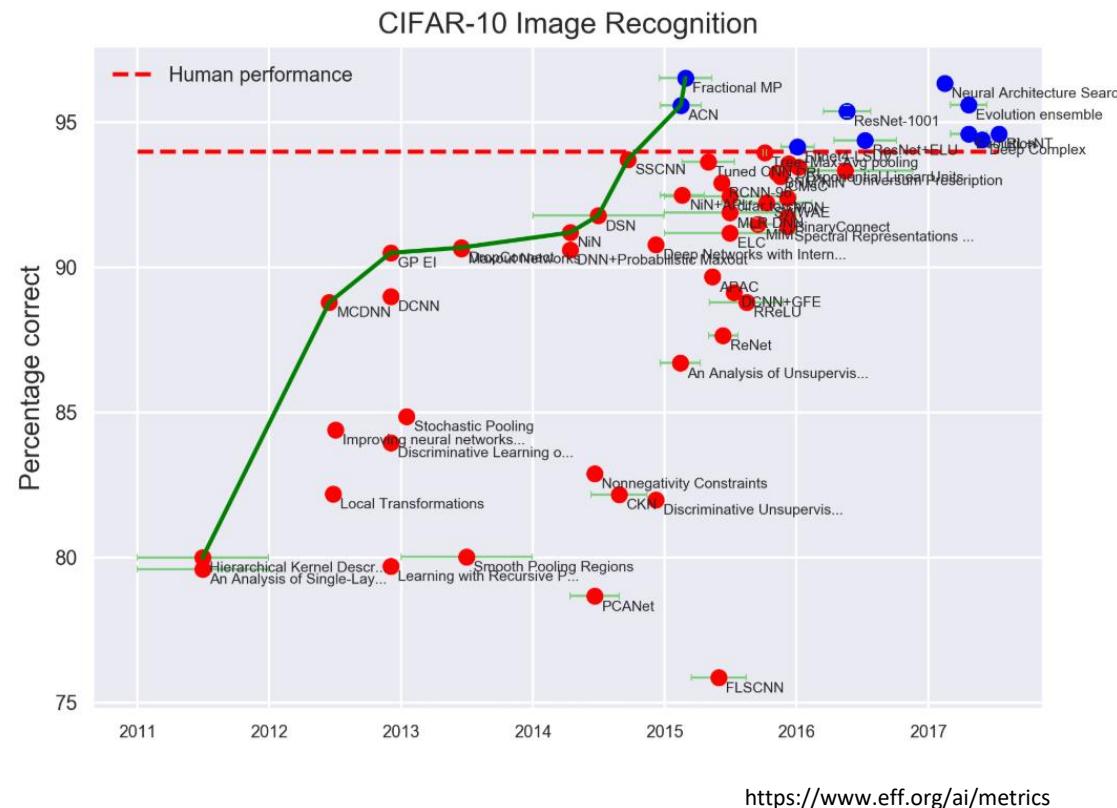
❖ 신경망 대신 딥러닝으로 인공지능 분야의 부활 (2006~)

- ✓ 드롭아웃 층(dropout layer) 도입으로 과적합 문제를 해결
- ✓ ReLU(Rectified Linear Unit) 활성함수 도입으로 기울기 사라짐 문제 해결
- ✓ GPU 컴퓨팅과 고속 최적화 알고리즘 등장
- ✓ ILSVRC에서 오류율을 15%로 획기적으로 낮춤



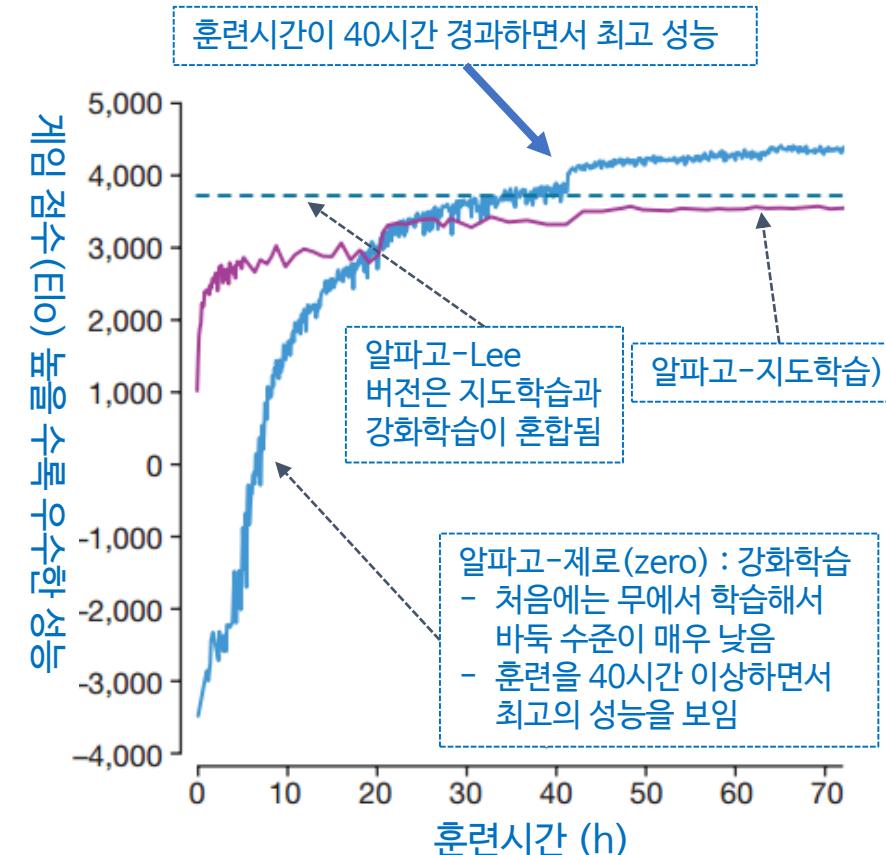
ANN의 역사 (12): 이미지 및 비전

구글 딥마인드, Nature 2015



알파고(AlphaGo)는 3개 버전

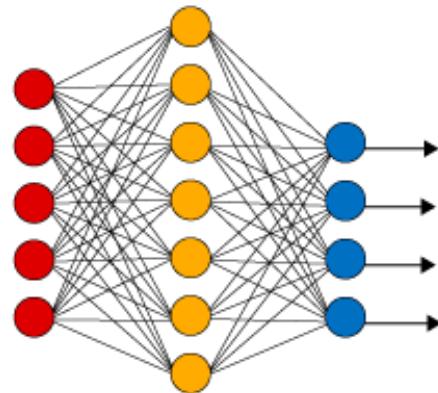
- ① 알파고-지도학습(바둑 기보)
- ② 알파고-리 (이세돌 9단)
- ③ 알파고-제로 (강화학습)



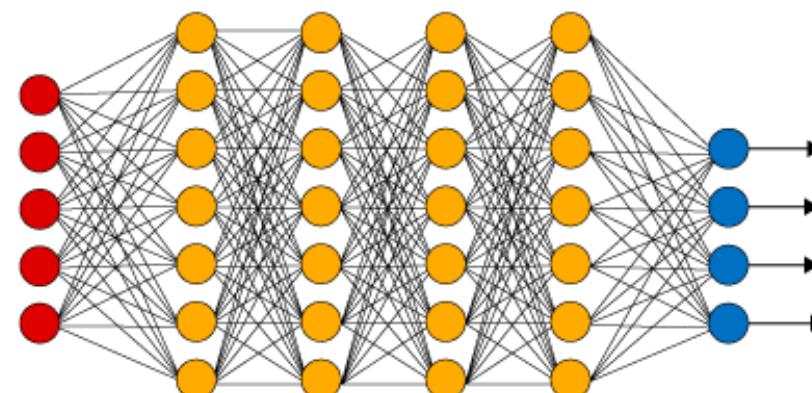
자료: Mastering the game of Go without human knowledge , David Silver, et al. Nature(2017)

딥러닝은 은닉층 2개 이상으로 구성된 인공신경망으로 정의함.

A. 단순 인공신경망(Shallow)
1개의 은닉층과 7개의 뉴런으로 구성



B. 딥러닝 구조
4개의 은닉층과 4x7(28)개의 뉴런으로 구성



● Input Layer

● Hidden Layer

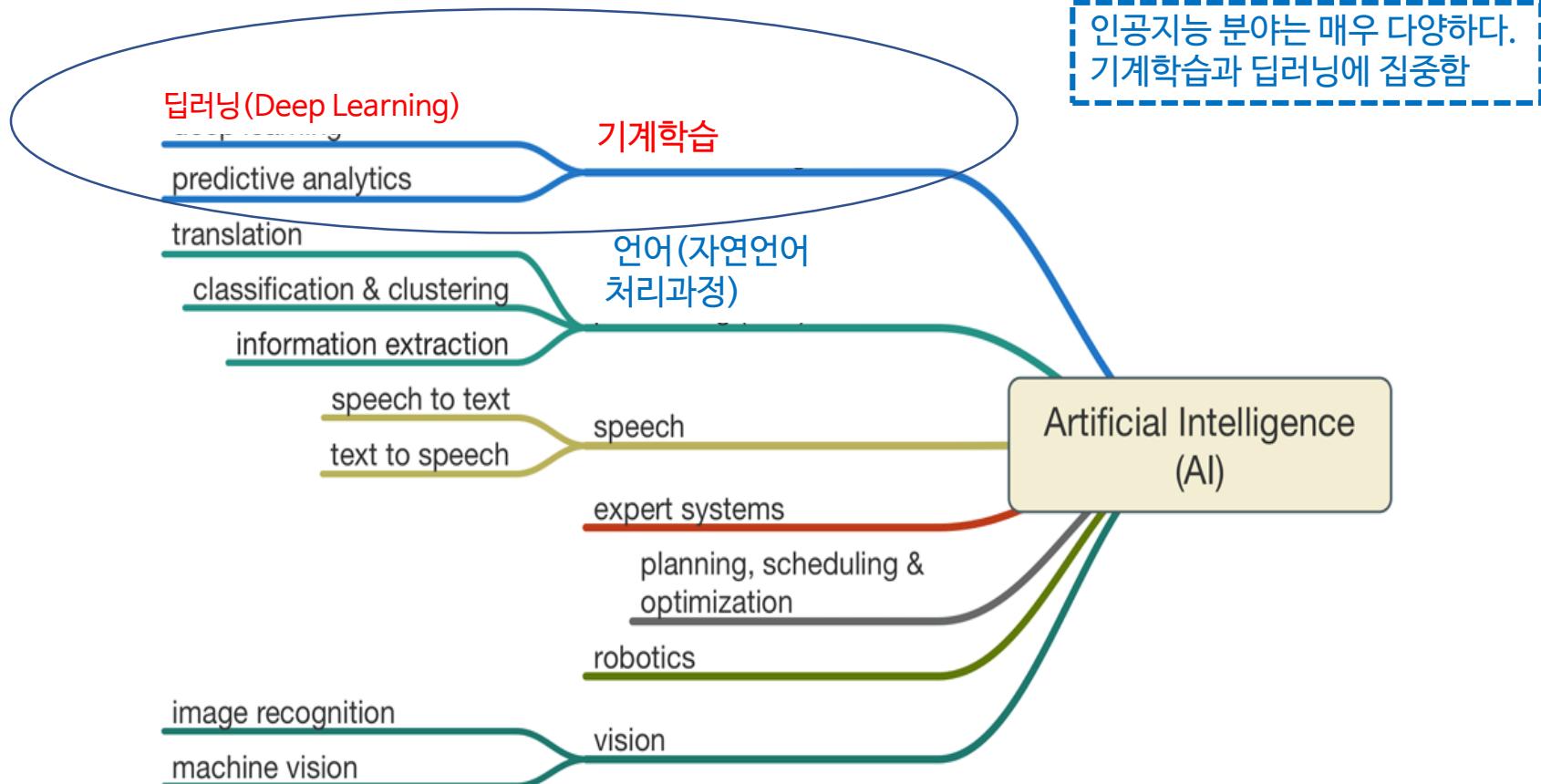
● Output Layer

자료: <https://www.xenonstack.com/blog/data-science/log-analytics-deep-machine-learning-ai/>

(퀴즈) 단순 인공신경망 II(Shallow)

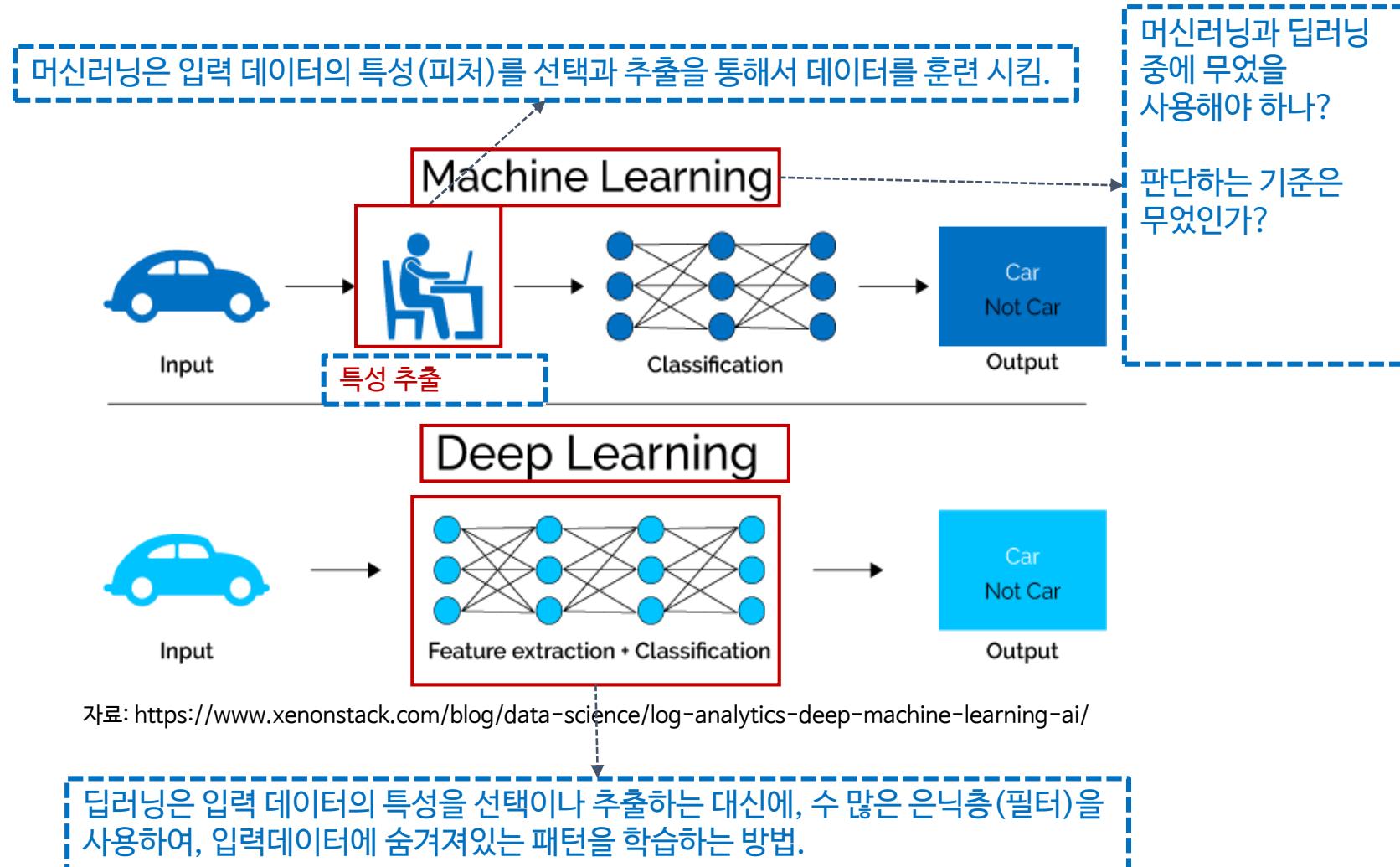
만일, 1개의 은닉층과 28개의 뉴런으로 구성되었다면, B와 비교하면 무엇이 다를까?

인공지능의 분류 (1)



자료: <https://www.eyerys.com/articles/paving-roads-artificial-intelligence-its-either-us-or-them>

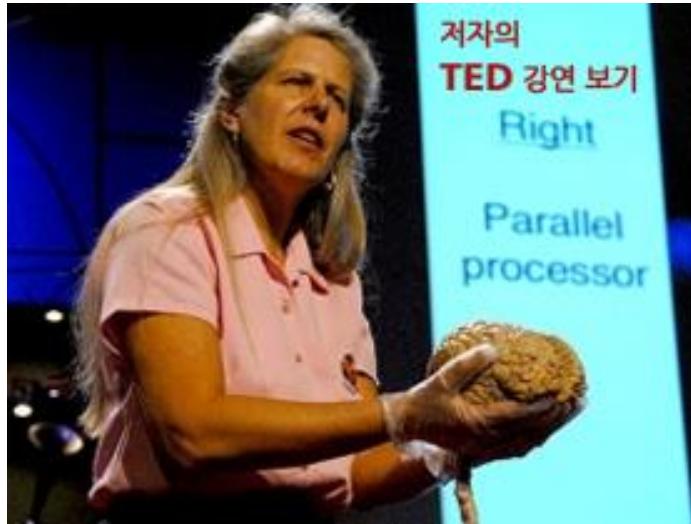
인공지능의 분류 (2)



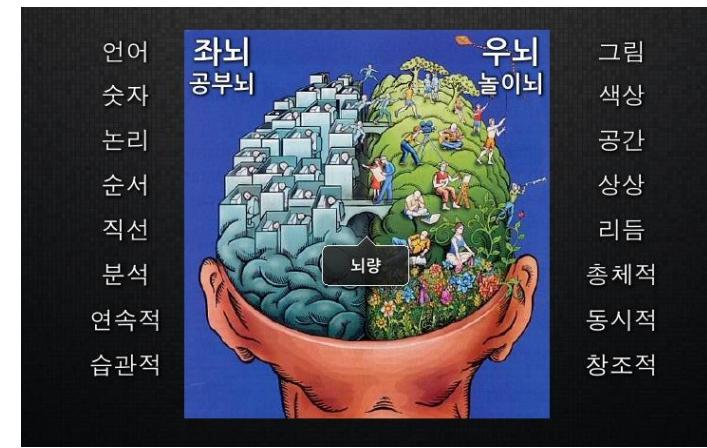
뇌와 기억 (2) : 긍정의 뇌

J. 테이러(Taylor) 교수:
하버드대 뇌과학자 뇌졸중 체험기

뇌량은 좌뇌와 우뇌의
신경세포들을 서로 연결해
주는 신경다발



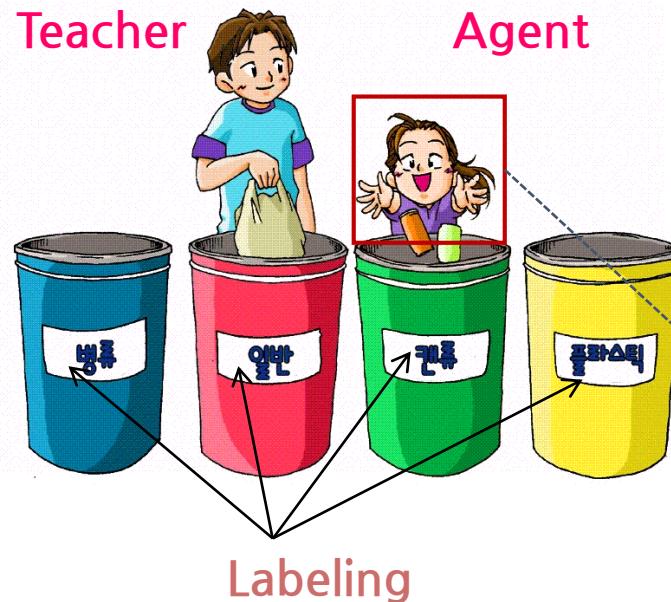
좌뇌와 우뇌의 차이는 왜 생길까?
상과 벌



Supervised Learning (지도학습)

- ① 입력(x), 출력(y)
- ② Labeling
- ③ Prediction

빅데이터 : 원본 데이터
훈련데이터+테스트 데이터로 나눔
비율은 7:3정도가 일반적임.



훈련데이터와 테스트 데이터

기존 훈련 데이터로 학습을 완료한 이후에,
새로운 테스트 데이터(쓰레기)를 누구의
도움도 없이 혼자서 분류를 한다면, 분류의
정확도는 얼마나 될까요?

지도학습과 레이블(Label) (2)

대표적인 딥러닝(머신러닝) 기반 지도학습을 위한 기초 데이터 제공함.
MNIST 데이터는 훈련용으로 60,000장 이미지와, 테스트용으로 10,000장

label = 5



label = 0



label = 4



label = 1



label = 9



label = 2



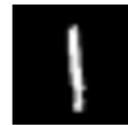
label = 1



label = 3



label = 1



label = 4



label = 3



label = 5



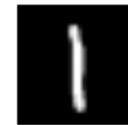
label = 3



label = 6



label = 1



label = 7



label = 2



label = 8



label = 6

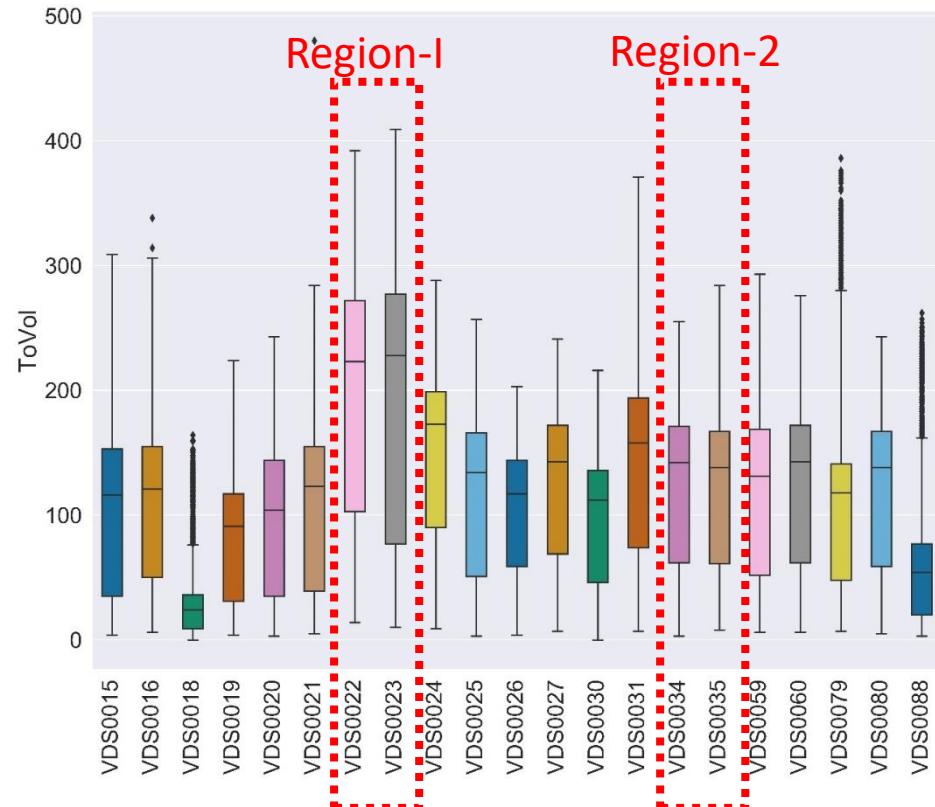


label = 9

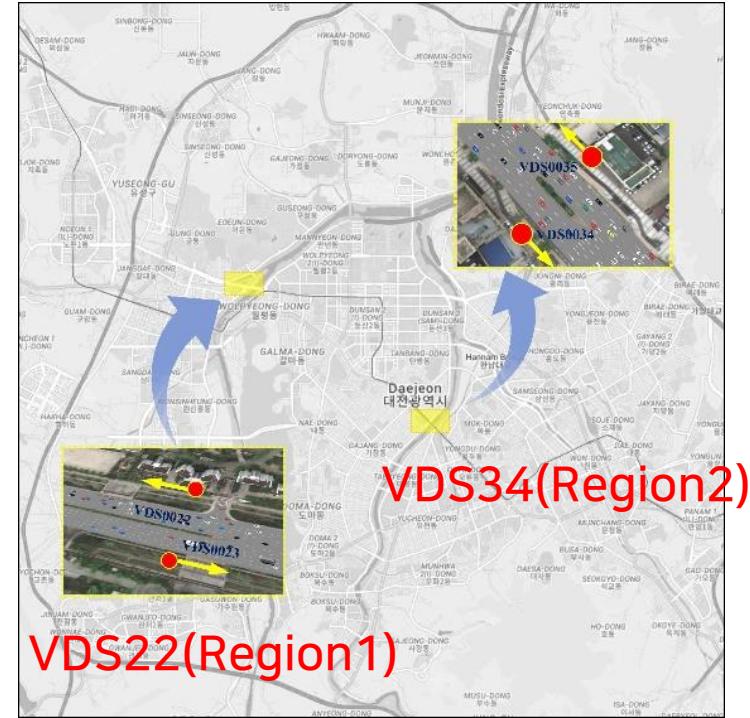


모든 손글씨 이미지
마다 라벨링이 되어
있음.
이미지 9와 라벨 '9'는
항상 쌍으로 저장됨

장단기메모리(LSTM) 교통 적용(1)

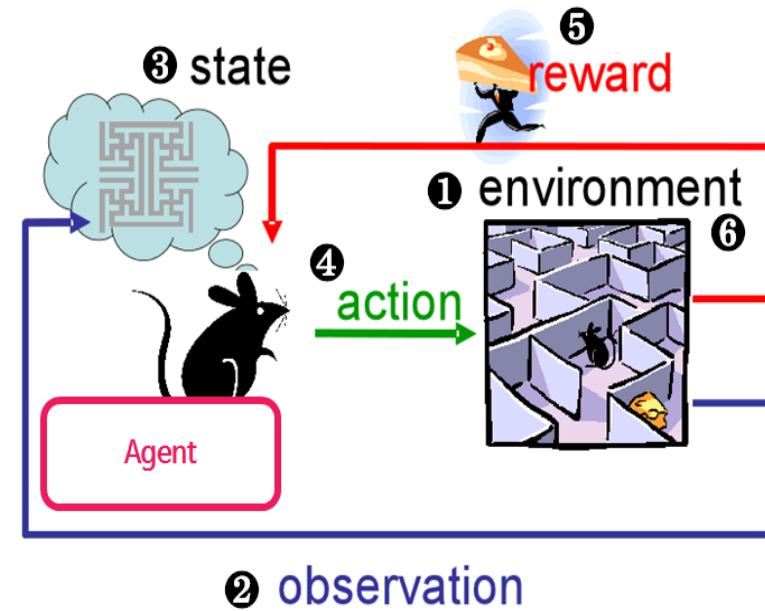
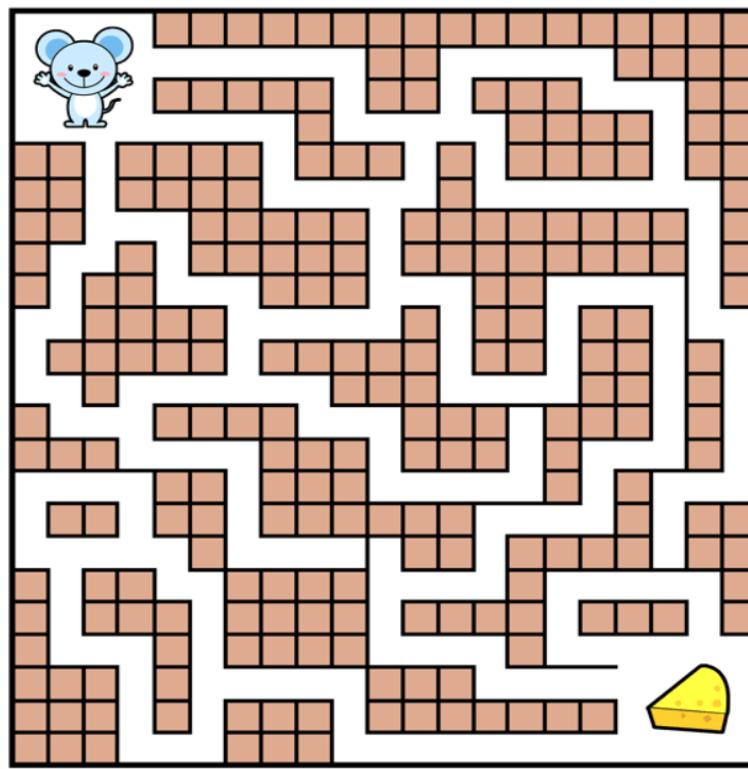


#18 and #22 are the highest and lowest traffic volume in the area, respectively. In this regard, we classify the area into three types of traffic flow such as high (#22), and low (# 18) densities for taking into consideration.



강화학습 : 쥐가 치즈를 먹을 수 있을까?

마르코프 결정 프로세스 (MDP) : 지금 이순간 최적의 행동을 선택한다



큐(Q)에게 상담을 해보자.

큐(Q)는 미로의 구조를 알고 있다고 가정!

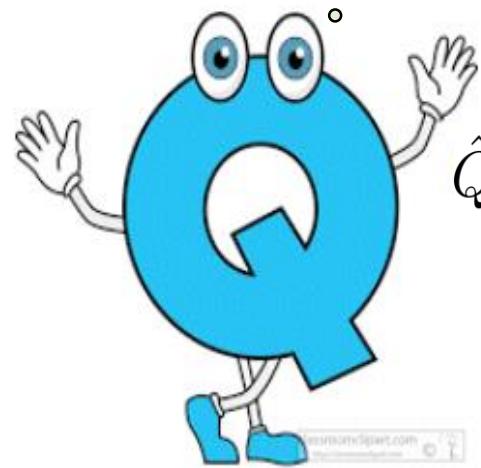
바로 직전 행동(t)

상태(위치, 좌표)

새로운 행동(t+1)을 알려줌

보상을 최대!

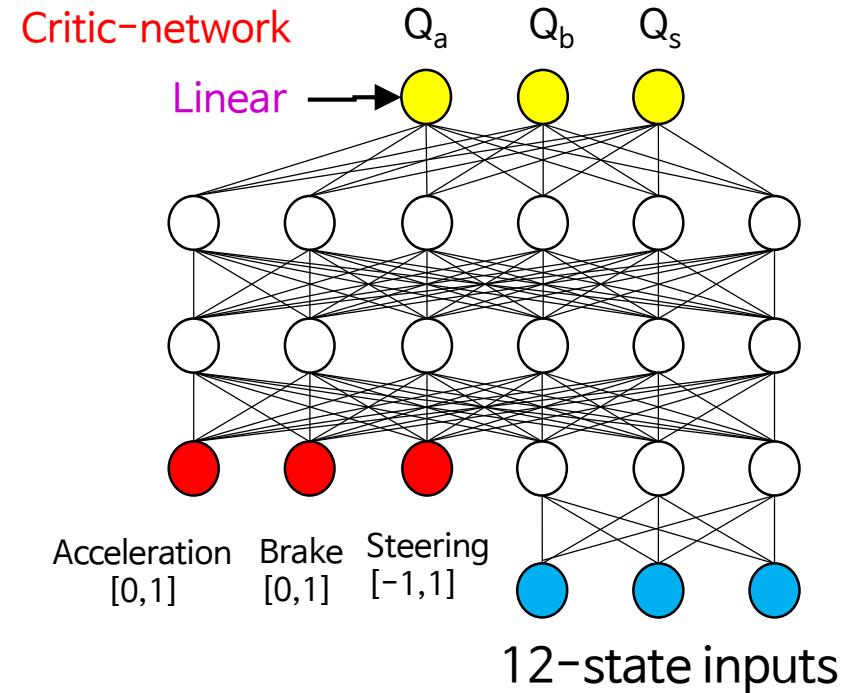
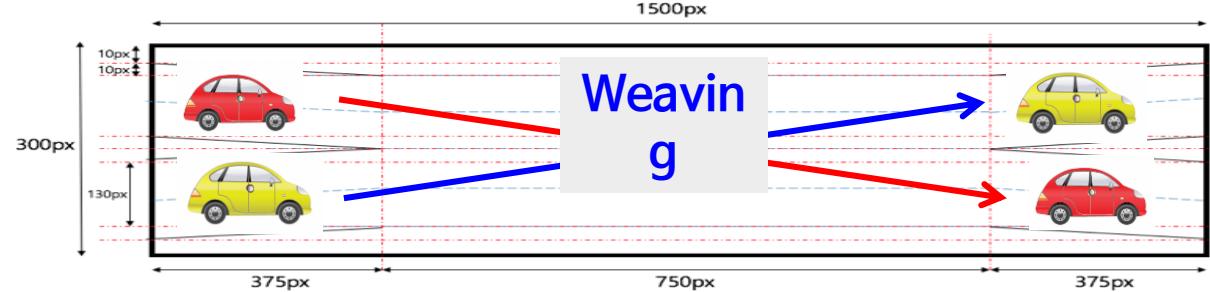
$$\hat{Q}(s, a) \leftarrow r + \max_{a'} \hat{Q}(s', a')$$



Weaving 구간 (카이스트교) : 차선 변경이 빈번히 발생하는 위험도로 구간

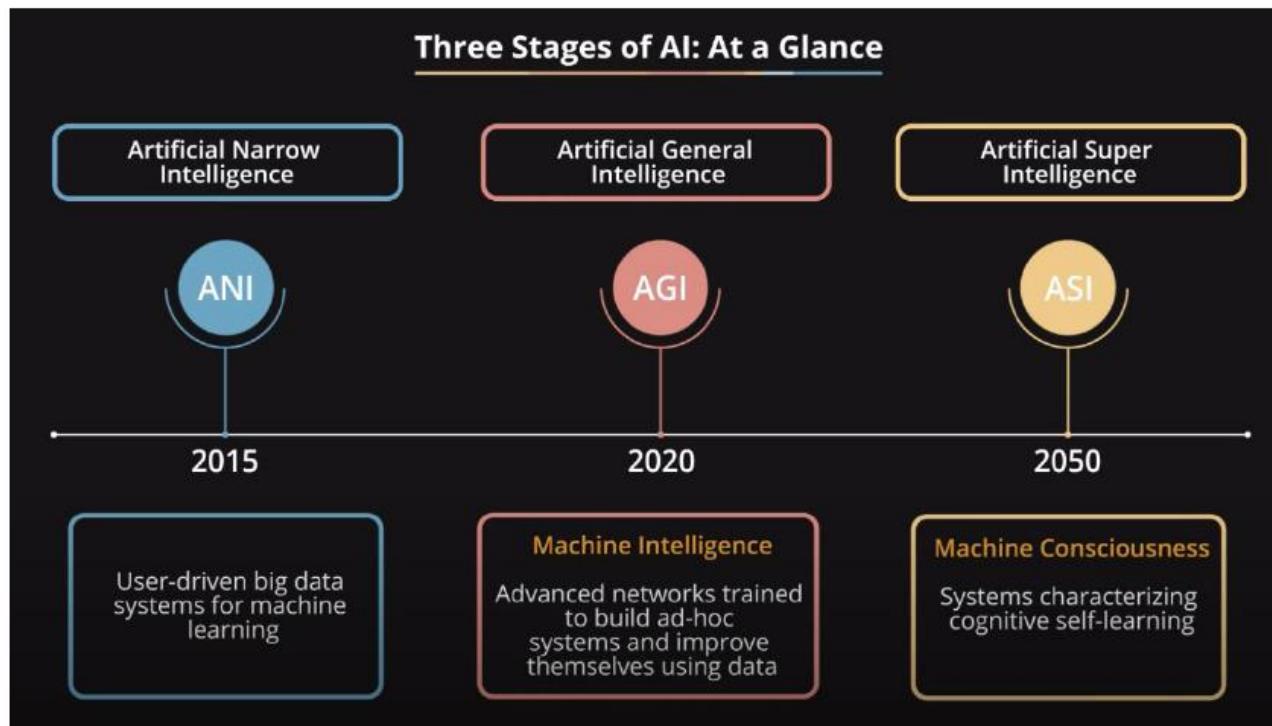


강화학습을 위한 환경(Environment) 설정



❖ 인공지능

- ✓ 현재 특정분야에 적용되는 약 인공지능 단계 (Weak AI)
- ✓ 향후 인간의 전체 지적 활동을 모방하는 일반적 인공지능 (AGI)
- ✓ 강 인공지능 까지 발전할 수 있을 것으로 전망 (Strong AI)



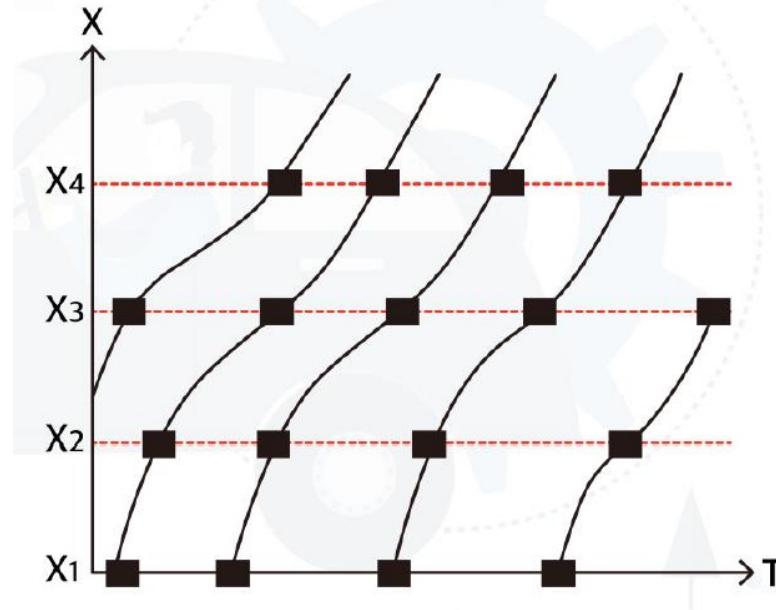
교통 데이터 소개

❖ 고정위치 관측

- ✓ 도로상의 고정된 위치에서 지나가는 차량들을 관측
- ✓ 고정된 시각에 일정 구간의 교통 상태를 관측
- ✓ 항공 사진 촬영

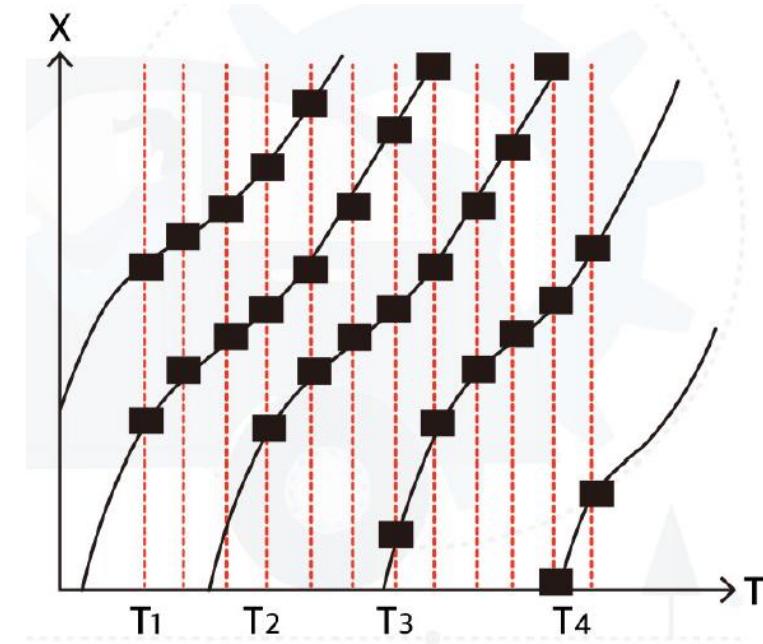
❖ 관측가능한 데이터는

- ✓ 통과 차량수, 차량 사이의 시간 간격, 차량 속도



• 연속 시간 관측

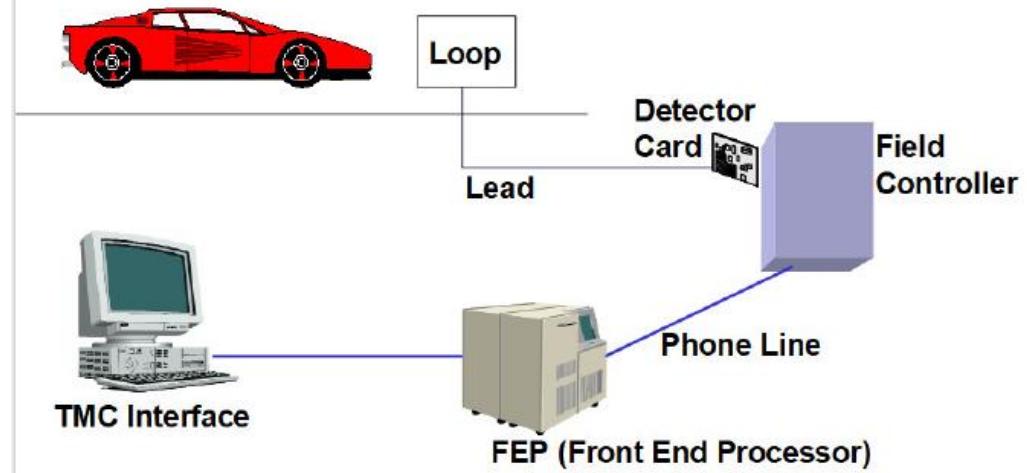
- ✓ 일정 구간내의 교통상태를 연속된 시간동안 관측
- ✓ 교통 카메라
- ✓ 관측데이터



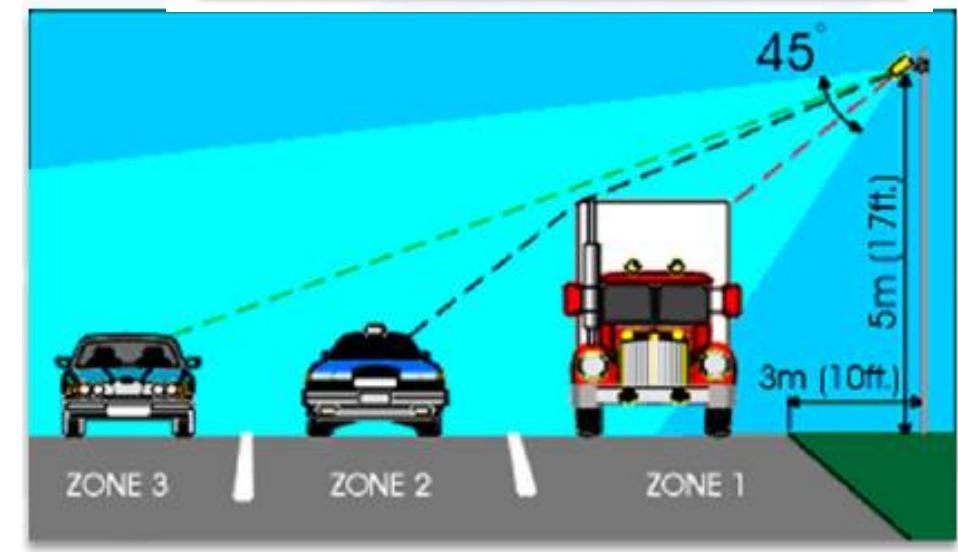
루프 검지기



- 교통량 산정
- 도로 용량 분석
- 차량 통과 속도 감시



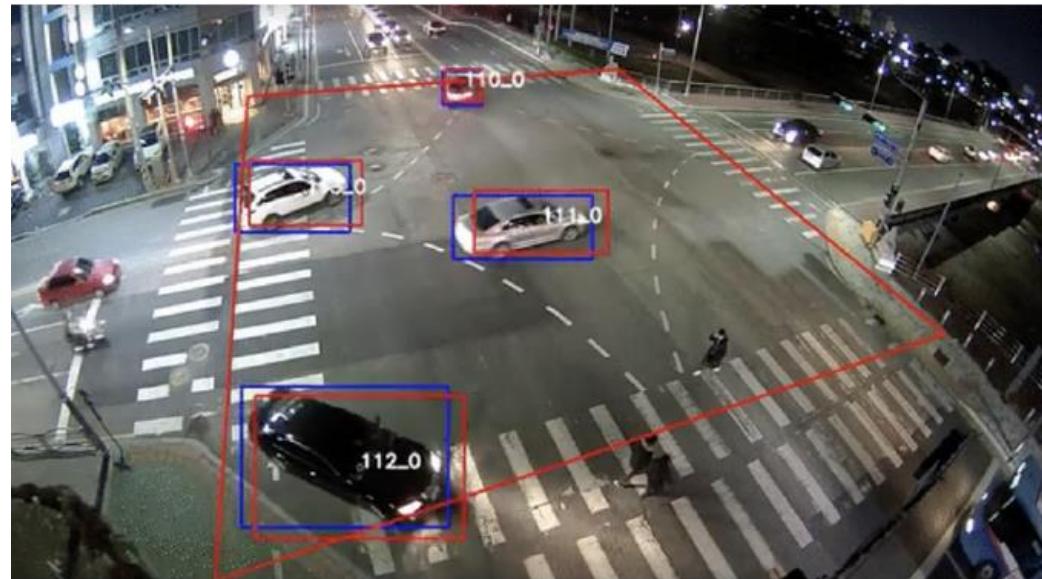
교통 레이다



영상 검지기

RSU(Road Side Unit)

도로를 운행하는 차량에 설치된 단말기와 WAVE 무선통신을 수행 차량 단말기에서 전송하는 각종 정보를 수집 저장하여 센터로 전송하는 기능

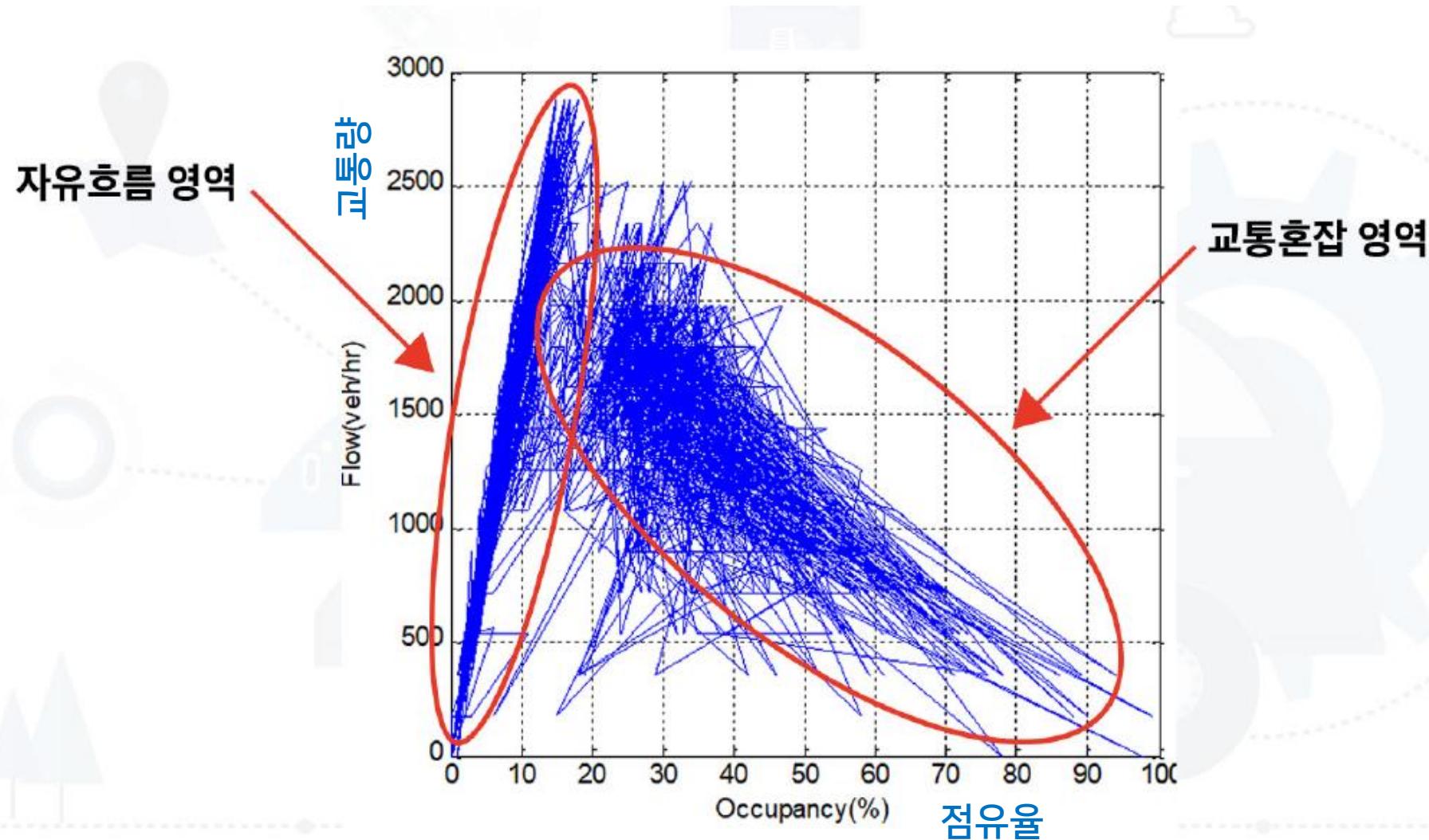


VDS 차량검지기



속도
차량별 교통량
점유률

교통 데이터의 해석



❖ 도시 교통 문제의 원인

- ✓ 교통혼잡은 온실가스, 미세먼지 원인
- ✓ 도시문제는 지속 가능성에 대한 글로벌한 도전

❖ 도시문제 해결을 위한 핵심 키워드는 스마트 시티(Smart city)



스마트 교통 시스템

❖ 빅데이터 및 활용 과정

- ✓ 분석, 해석의 자동화 수단으로 빅데이터의 축적에 따라 주목받고 투자가 집중
- ✓ 데이터 수집과 저장, 분석, 시각화 단계를 거쳐 최종 제품이나 서비스에 적용

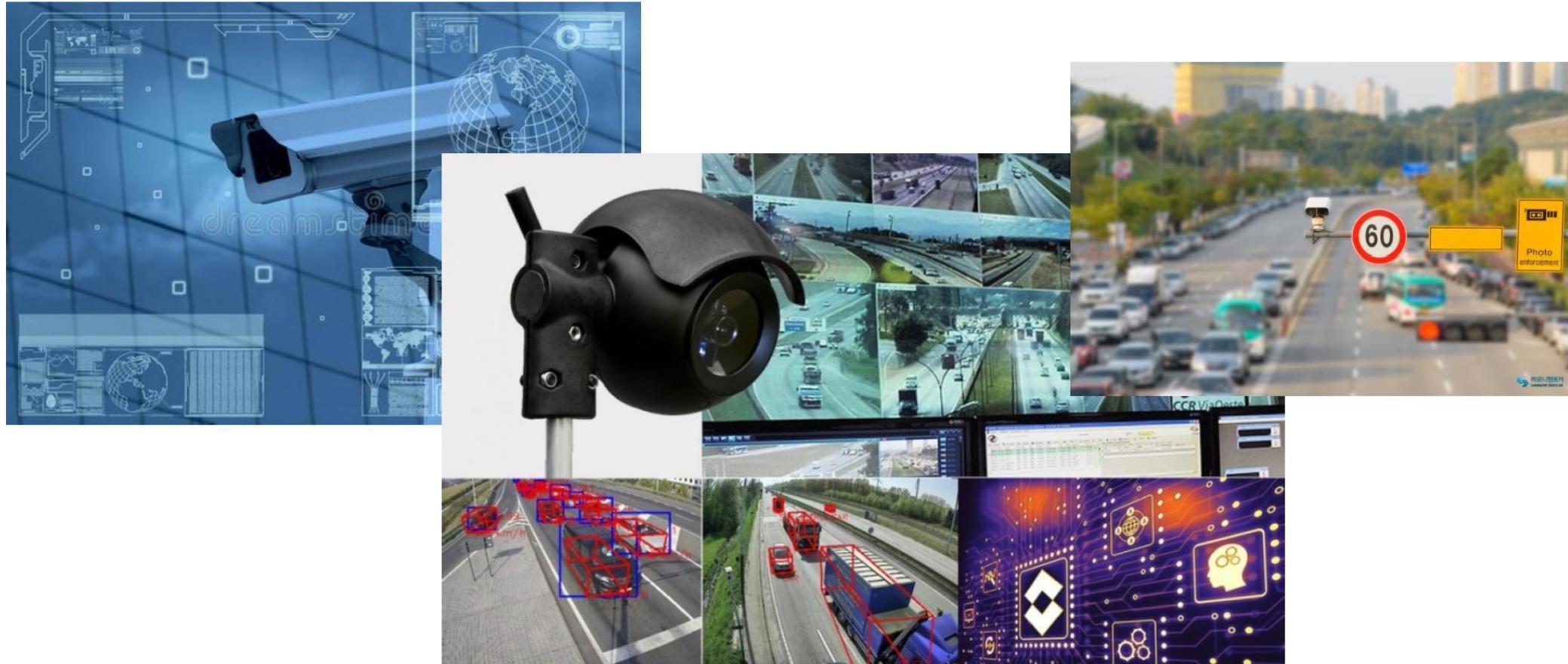


❖ 인공지능

- ✓ AI (학습을 통한 빅데이터 분석)는 인간의 인지, 예측, 분석을 보조하고
- ✓ 장기적으로 스스로 지식, 아이디어의 창출 활동을 함으로써 대리인 역할을 수행

도시 교통분야의 인공지능 활용

- ❖ 기존 CCTV 영상을 분석하여 교통정보 생성 및 정지차량, 역주행차량, 보행자 등 돌발상황을 검지하여 교통상황실 운영자에게 알림



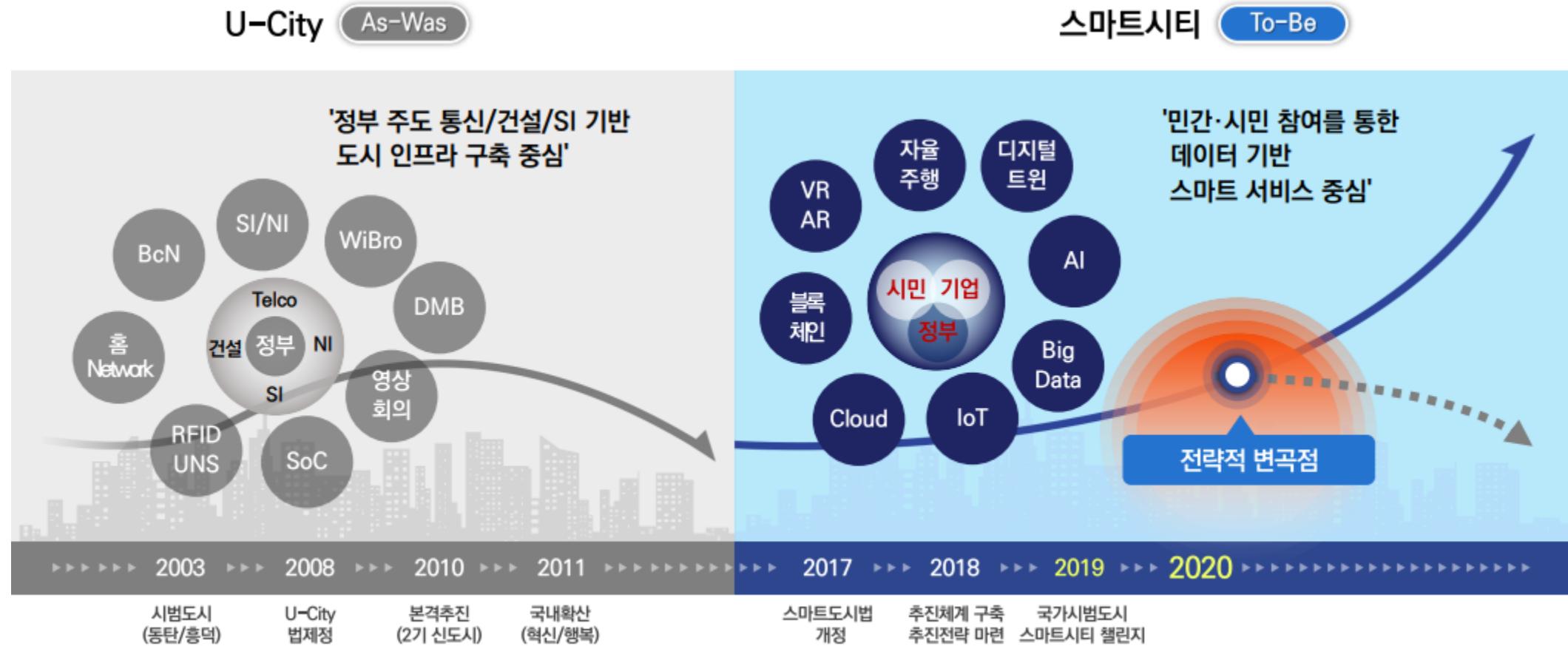
❖ 네덜란드 지능형 교통체계 : 스마트 신호등 도입 이후 교통체증의 약 20%가 감소

- ✓ 6년간 신호등과 도시 곳곳에 달린 카메라를 통해 수집한 빅데이터를 AI가 분석한 자료로 최적의 신호를 적절한 시간과 장소에 보냄

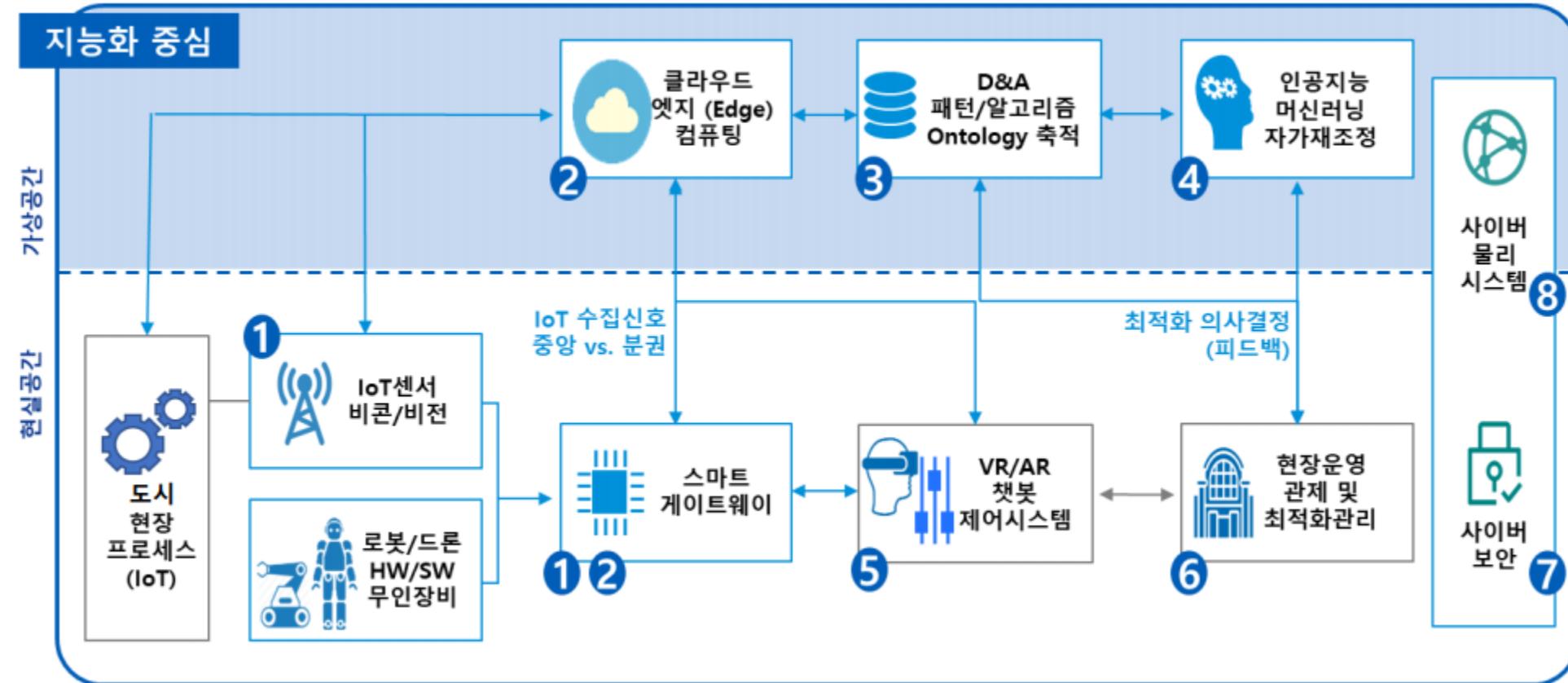


- ❖ 세계는 도시에서 스마트시티로 전환중으로, IoT의 활성화 덕분에 정부는 실시간으로 데이터 수집을 하여 인공지능의 기능으로 통합가능
- ❖ 현재 스마트 시티 계획은 주차관리 뿐 아니라 교통량의 향상을 수반하는 인공지능의 응용을 활용하고 있으며, 장차 자율주행 자동차의 안전한 융합으로 구성될 것
- ❖ AI 기반 Smart City
 - ✓ 첨단 정보통신기술(ICT)을 이용해 교통문제, 환경문제, 주거문제, 시설 비효율 등을 해결하여
 - ✓ 시민들이 편리하고 쾌적한 삶을 누릴 수 있도록 한 '똑똑한 도시'를 뜻한다.
- ❖ Smart City는 혁신, 자율성, 역동성의 특성을 갖춘 도시를 의미한다
 - ✓ 첨단 정보통신기술로 인해 발전한 다양한 유형의 전자데이터 수집 센서를 사용해서
 - ✓ 능동적으로 정보를 취득하고, 이를 자산과 리소스를 효율적으로 관리하는데 사용하는 도시지역

구현지속 가능한 도시운영효율 및 도시 경쟁력 강화를 목표



- ❖ 도시현상에 대한 이해를 바탕으로 AI, IoT, Digital Twin, Blockchain 등 융합 기술로 구현함



스마트 시티 실증도시



부산 Eco Delta City

“로봇 등 산업육성으로 혁신생태계가 조성되는 미래 수변도시”

4차산업혁명에 대응하고 산업육성을 위한 5대 클러스터 조성

(서비스) 로봇활용, 배움-일-놀이, 도시관리 지능화, 스마트워터, 제로에너지, 스마트교육&리빙, 헬스, 모빌리티, 안전, 스마트공원



세종 5-1 생활권

“인공지능(AI)기반 도시로 시민의 일상을 바꾸는 스마트시티”

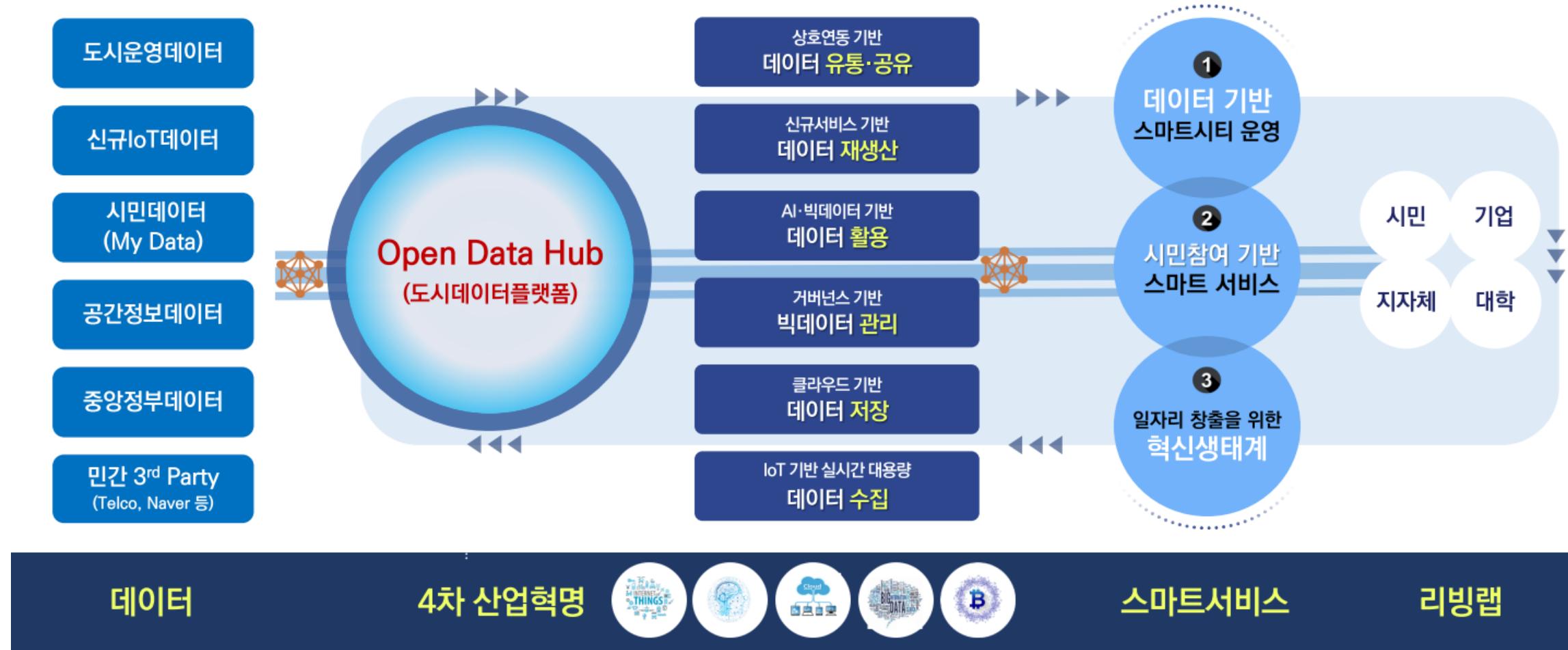
소유차 제한구역 설정, BRT 중심으로 직주 근접(용도혼합) 등

(서비스) 모빌리티, 헬스케어, 교육, 에너지-환경, 거버넌스, 문화-쇼핑, 일자리

2021년 국토부 스마트시티 실증 도시 1단계 예산 지원 사업

- SPC 설립 후 이를 통한 사업 발주 진행될 예정이고 부산 EDC는 한화에너지컨소시엄이, 세종5-1은 LG 컨소시엄이 주관이 되어 SPC 설립 후 진행 예정
- AI 데이터센터
- 디지털트윈
- 스마트IOT
- 사이버보안
- 스마트혁신단지
- 스마트교통
- 스마트에너지
- 스마트헬스케어
- 스마트안전

도시데이터플랫폼은 '오픈 데이터 허브'로 구축

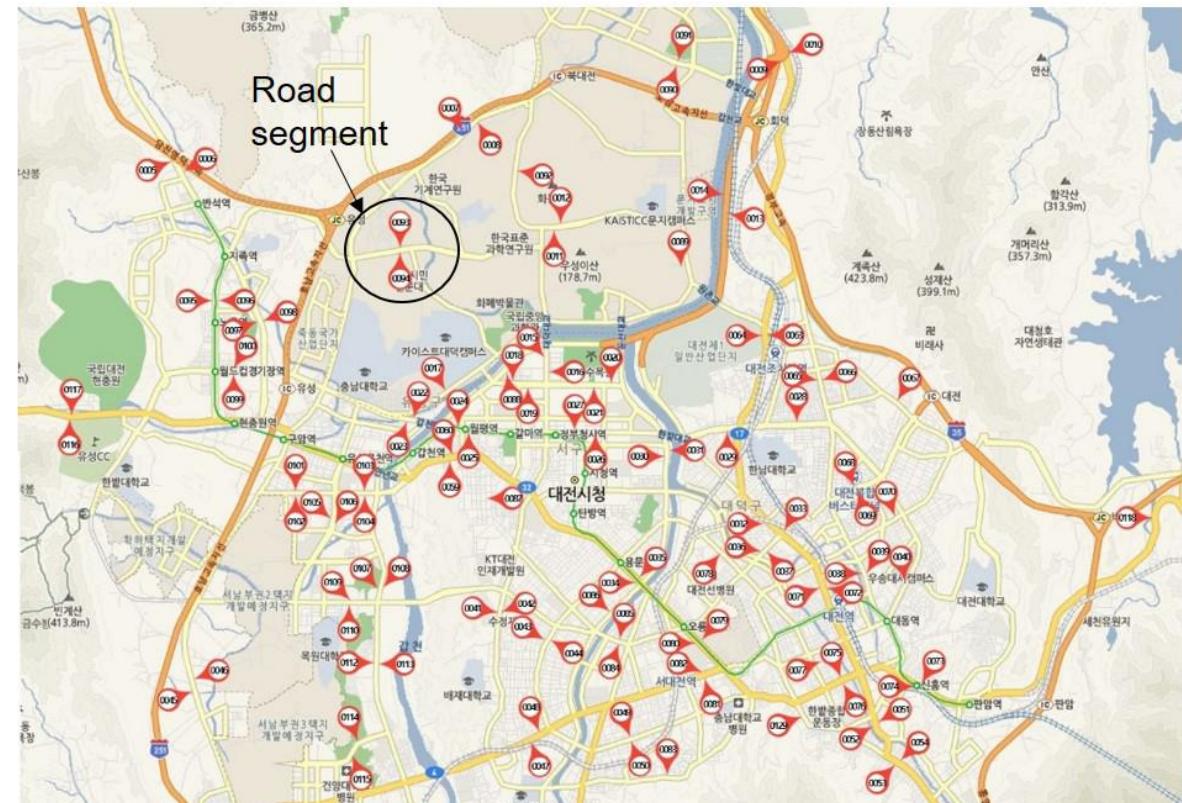
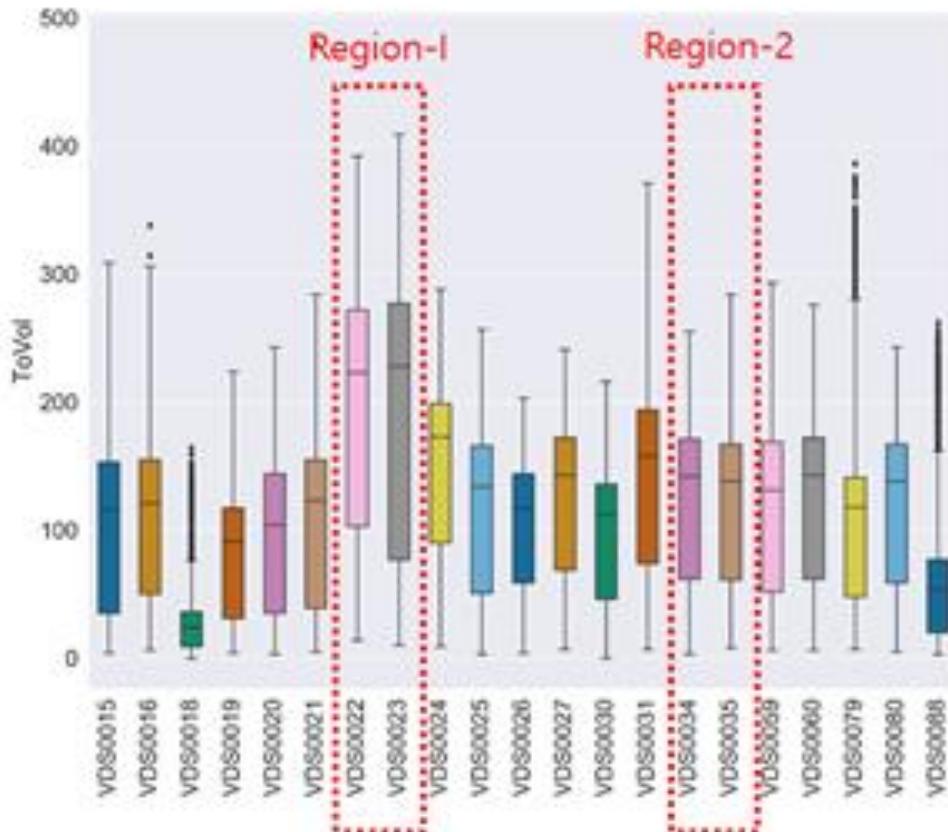


AI Smart 교통 접근 전략

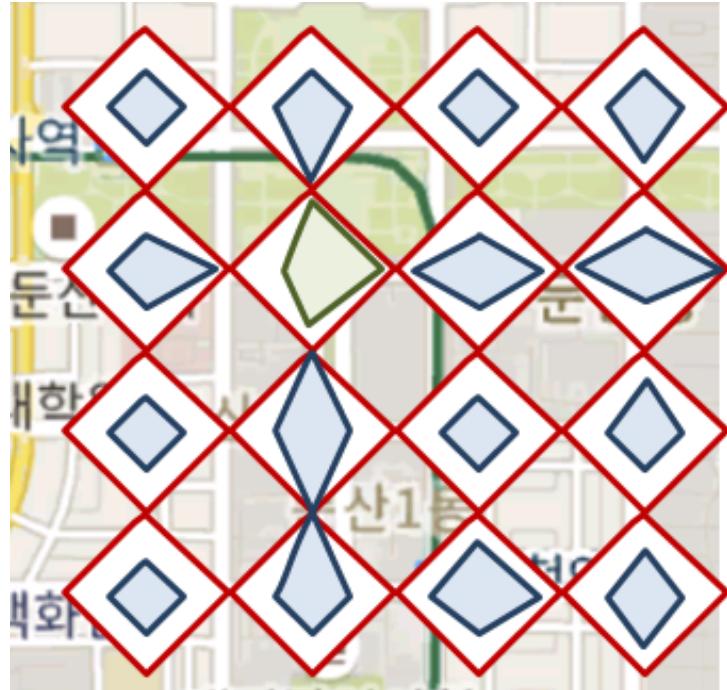
교통 장기 예측

- ❖ 대전시 VDS 교통데이터를 이용한 교통혼잡 지역에서 교통흐름 예측

- ✓ VDS(속도, 교통량, 점유율)을 장단기메모리(LSTM)을 이용한 예측 모델

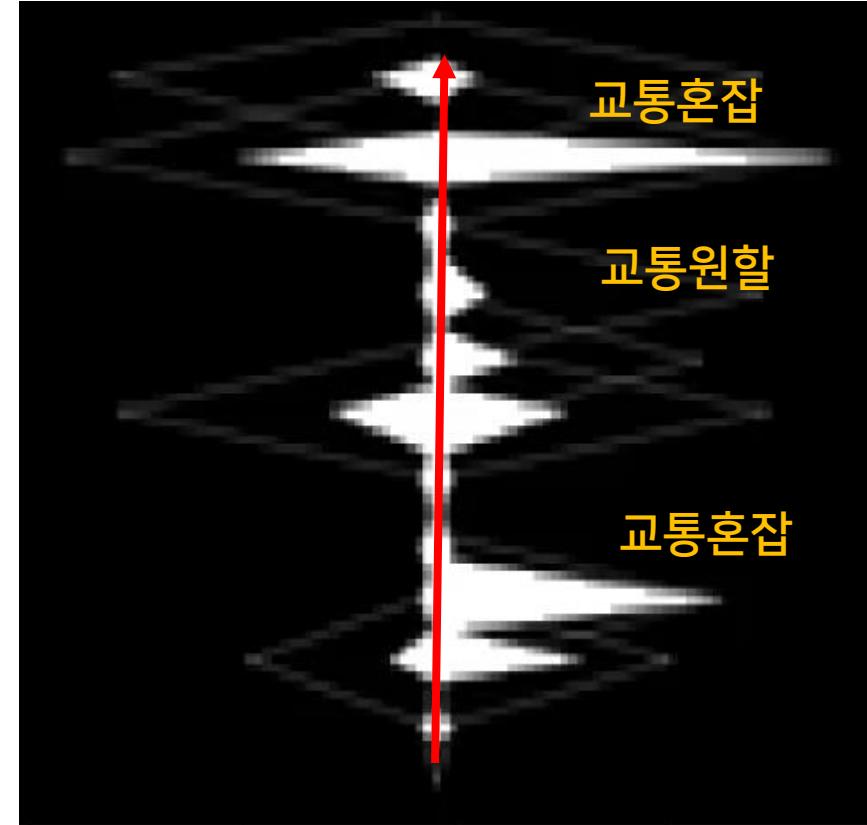


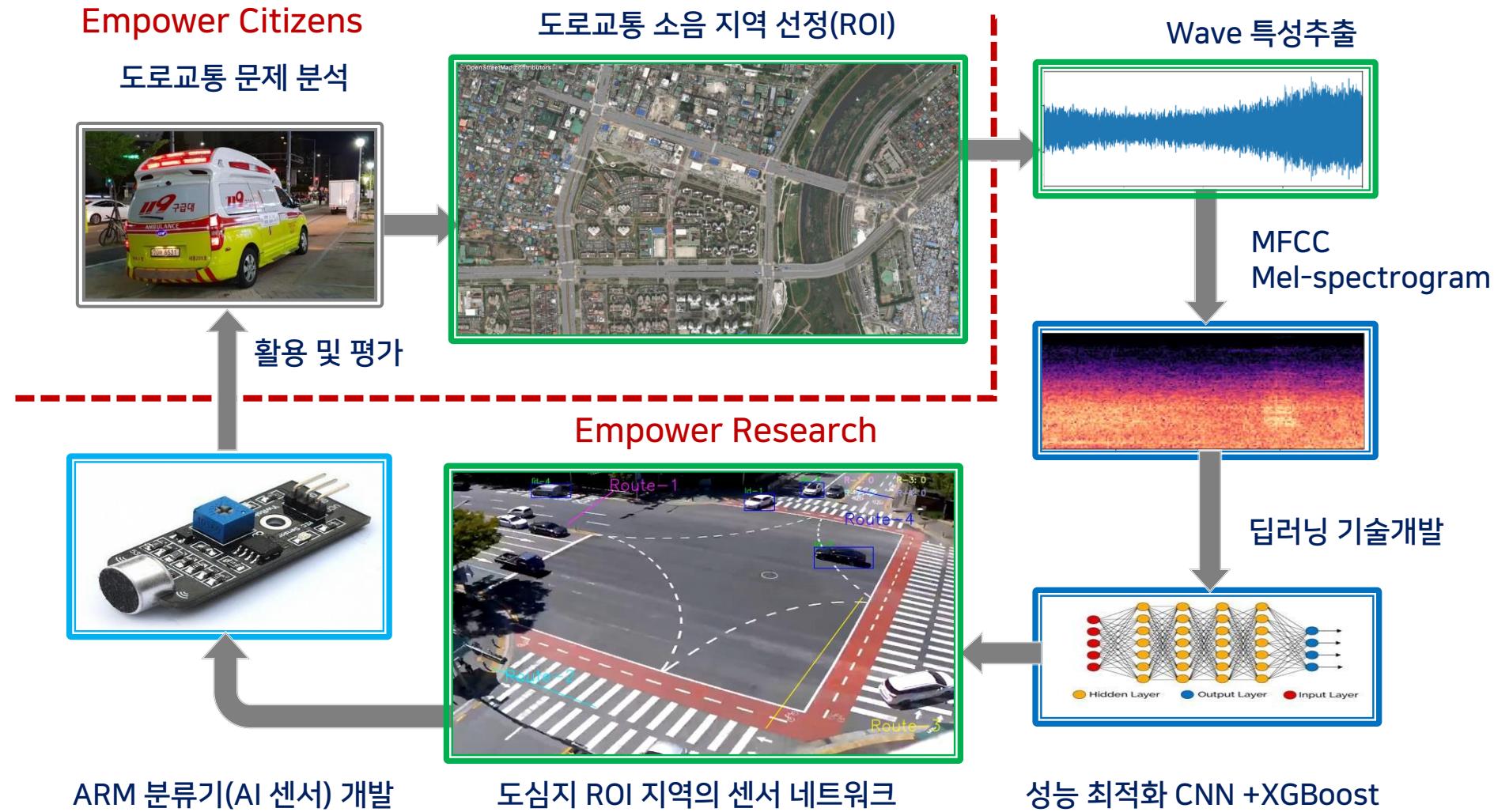
- ❖ 대전시 대덕대로(갤러리아 백화점~과학관) RSE 데이터 활용
- ❖ RSE 데이터 기반 교통 혼잡 데이터 분석으로 AI CNN 모데러 적용 가능함



대전시 중앙과학관 방면

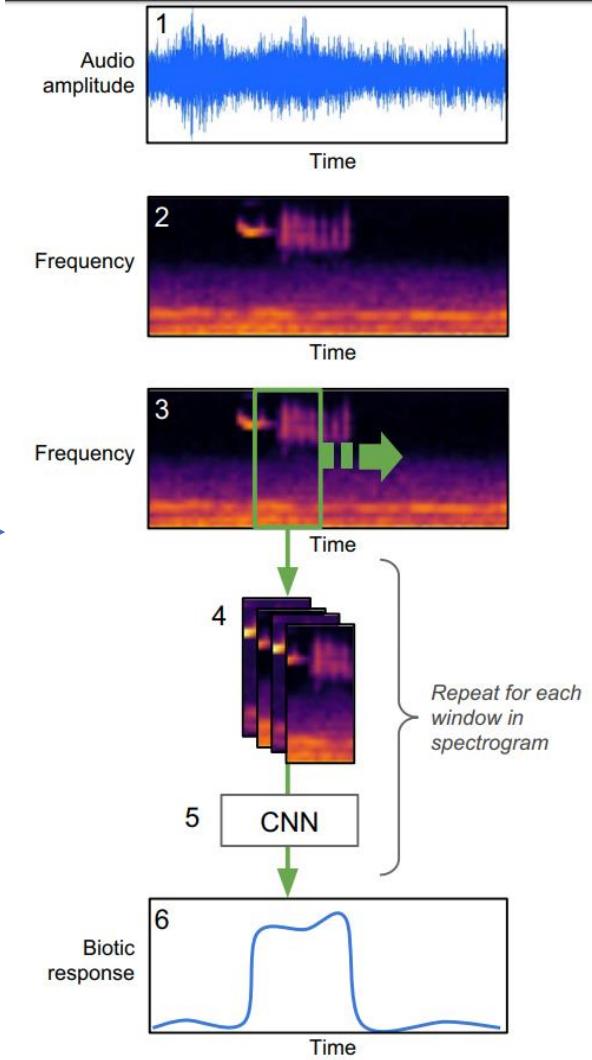
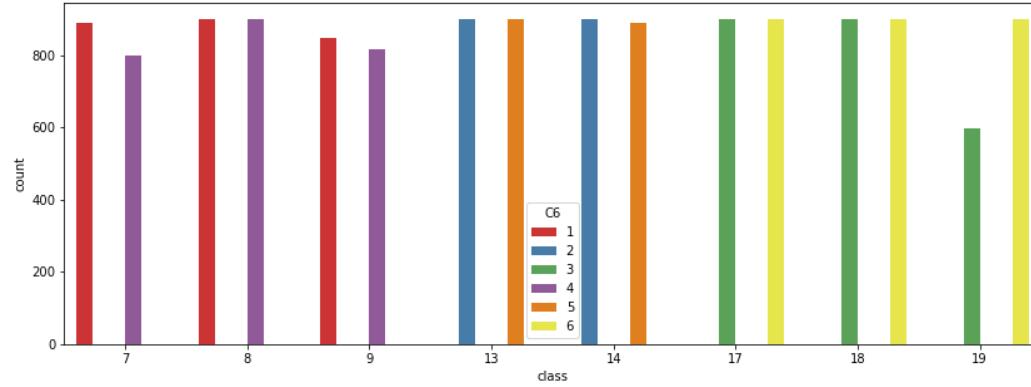
대전시 갤러기아 백화점





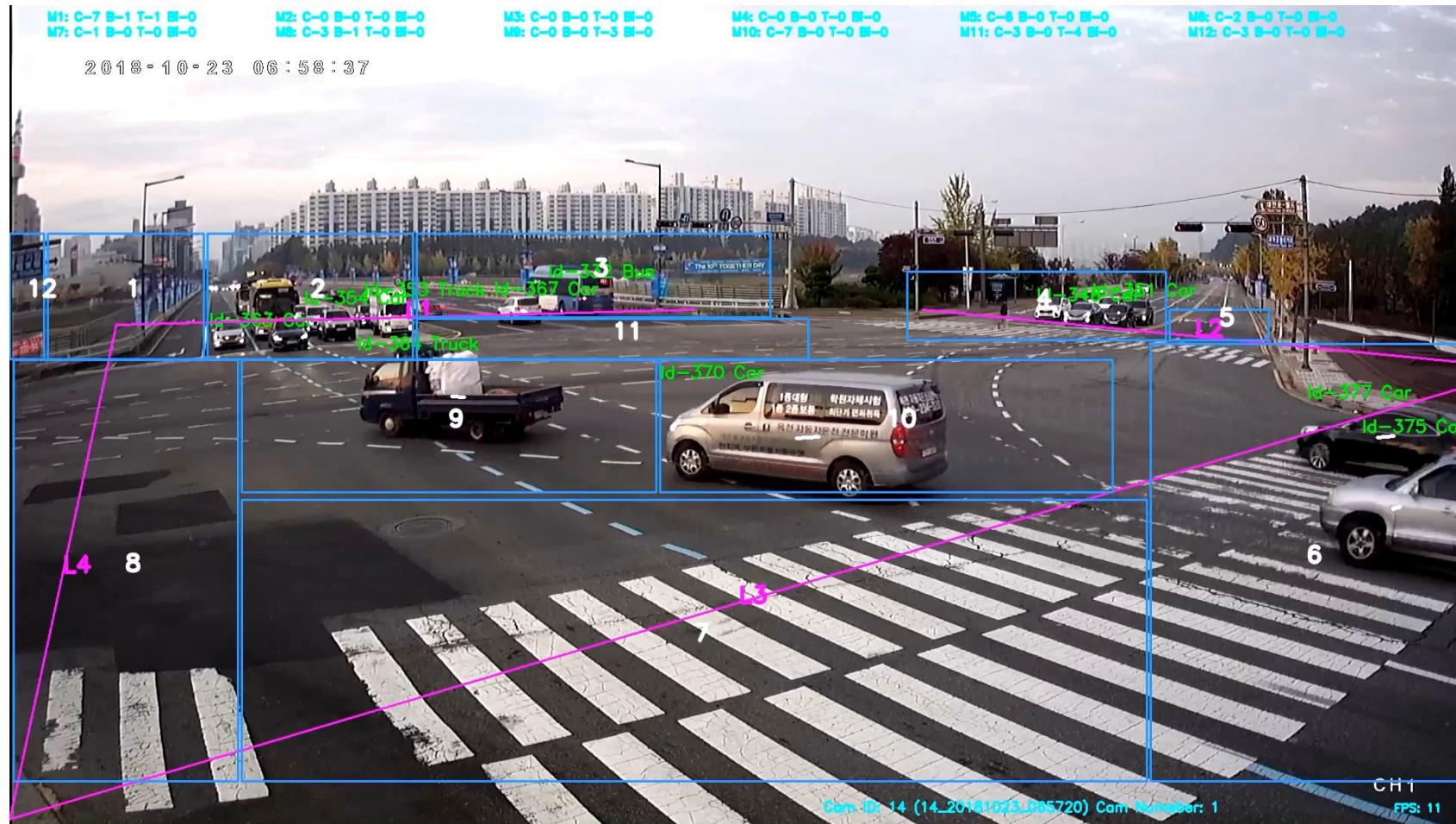
대전시 둔산대로 소음 측정 및 영상 분석 기술

- 소음 빅데이터 자체 수집 : DJ_TrafficSound14K (대덕대로)



대전시 CCTV 영상 분석으로 교통량 추정 기술

❖ 세계 최고 AI 컨퍼런스인 CVPR20의 'AI City Challenge' 경진대회 참가 및 (7위, KISTI)



데이터 사이언스 분야에 자주 사용되는 소프트웨어

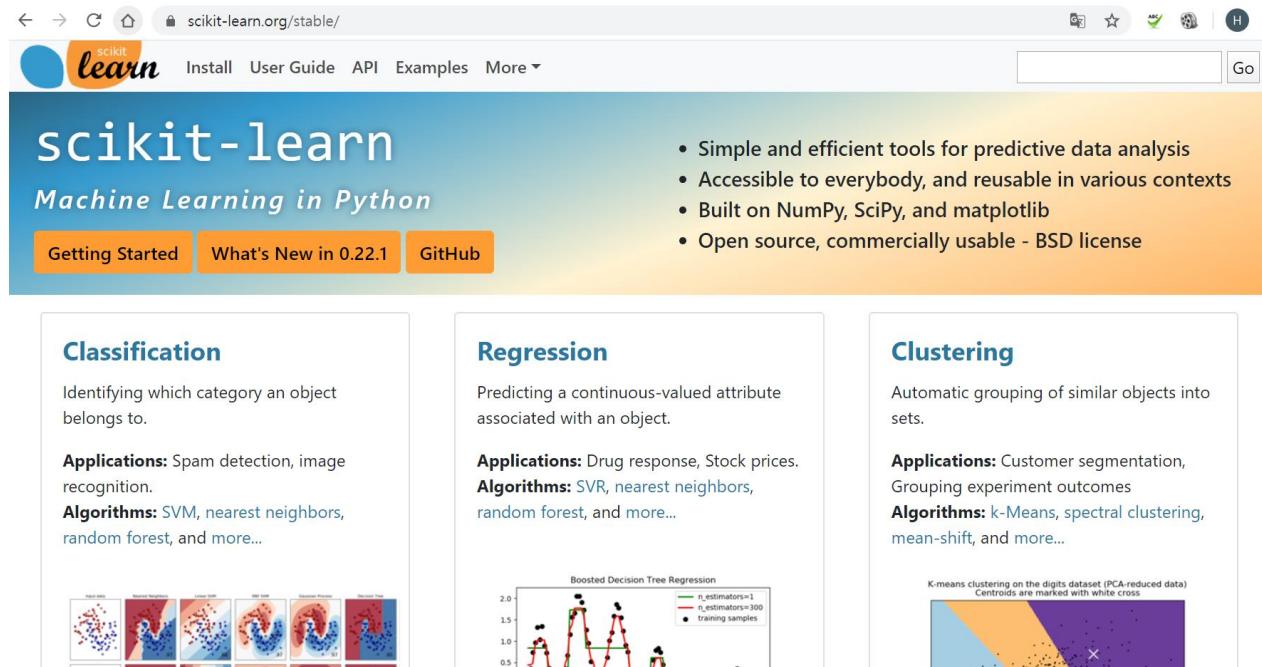
■ 데이터 사이언스 분야에 자주 사용되는 패키지

| 패키지명 | 주 요 기 능 |
|---------------------|--|
| pandas | 테이블 형태의 데이터를 다루는 데이터프레임(DataFrame) 자료 처리 |
| numpy | 수치 해석, 특히 선형 대수(linear algebra)의 다차원 배열, 벡터 연산 |
| matplotlib | 각종 그래프나 차트 등을 그리는 시각화 기능 |
| scipy | 고급 수학 함수, 수치 미적분, 미분 방정식 계산, 최적화, 신호 처리 |
| seaborn | matplotlib 패키지에서 지원하지 않는 고급 통계 차트 등 시각화 기능 |
| statsmodels | 통계 및 회귀 분석이나 시계열 분석 |
| scikit-learn | 대부분의 머신러닝 모델 제공, 파이썬으로 머신러닝을 공부하는데 최적의 패키지임 |
| tensorflow* | 신경망 모형 등 딥러닝 모델 제공 |

* 아나콘다 작업환경에는 데이터 사이언스 패키지들이 기본적으로 설치되어 있음. tensorflow는 추가 설치 필요

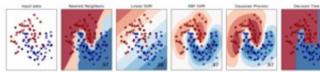
❖ 파이썬 머신러닝 중에서 가장 많이 사용되는 라이브러리

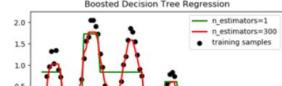
- ✓ 예측 데이터 분석을 위해 간단하고 효과적인 툴 제공
- ✓ Numpy, SciPy, Matplotlib을 기반으로 구성
- ✓ 아나콘다를 설치하면 기본적으로 사이킷런까지 설치가 완료 됨
 - \$ conda install scikit-learn



The screenshot shows the official website for scikit-learn at scikit-learn.org/stable/. The header includes the scikit-learn logo, navigation links for Install, User Guide, API, Examples, and More, and a search bar. The main content area has a blue gradient background with the text "scikit-learn" and "Machine Learning in Python". Below this, there are three main sections: "Classification", "Regression", and "Clustering", each with a brief description, applications, algorithms, and a small visual example. A large orange callout box highlights the following features:

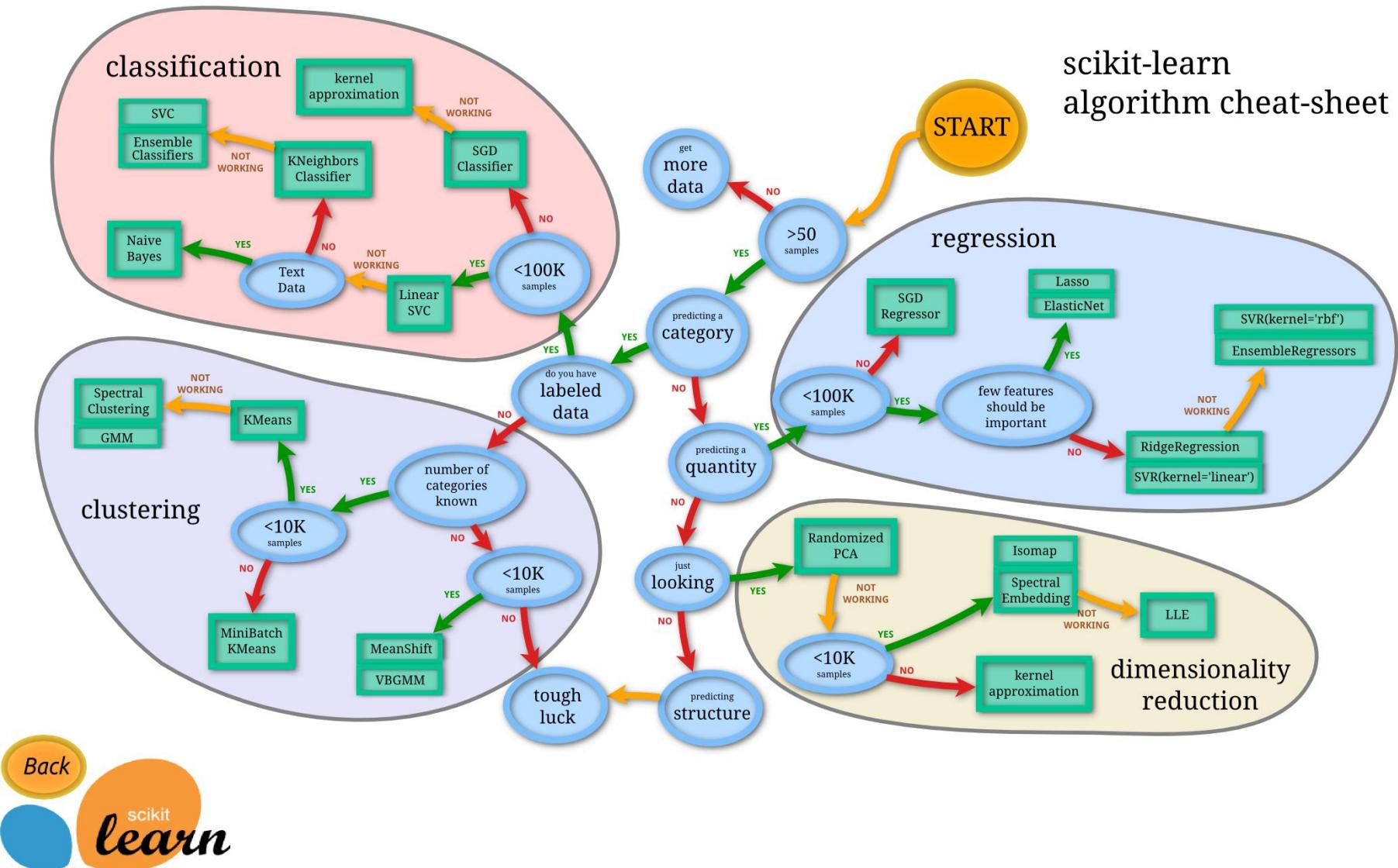
- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification
Identifying which category an object belongs to.
Applications: Spam detection, image recognition.
Algorithms: SVM, nearest neighbors, random forest, and more...


Regression
Predicting a continuous-valued attribute associated with an object.
Applications: Drug response, Stock prices.
Algorithms: SVR, nearest neighbors, random forest, and more...


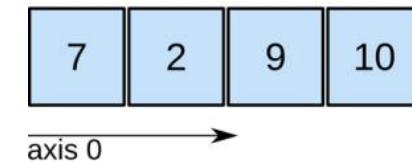
Clustering
Automatic grouping of similar objects into sets.
Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: k-Means, spectral clustering, mean-shift, and more...


사이킷런(scikit-learn) 활용 예제



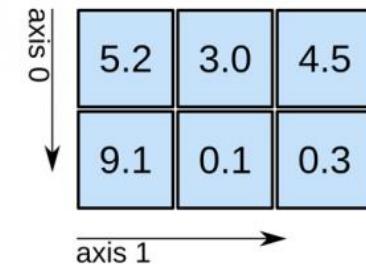


1D array



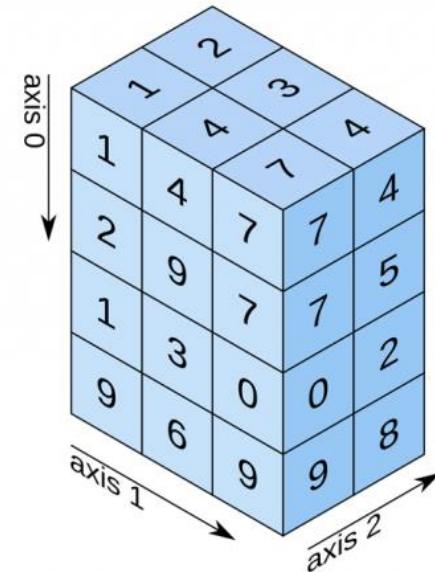
shape: (4,)

2D array



shape: (2, 3)

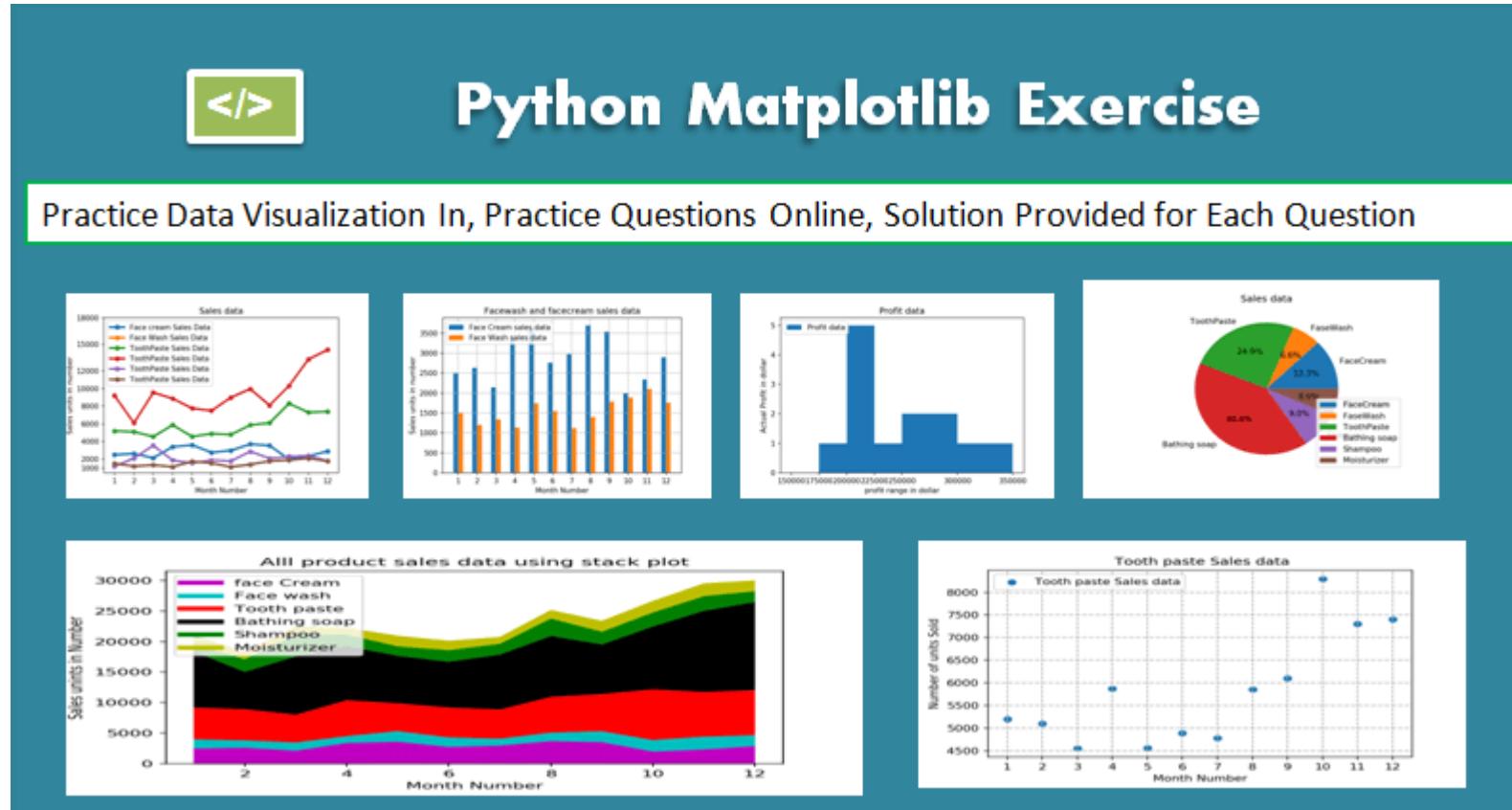
3D array



shape: (4, 3, 2)

Python Matplotlib Exercise

Practice Data Visualization In, Practice Questions Online, Solution Provided for Each Question

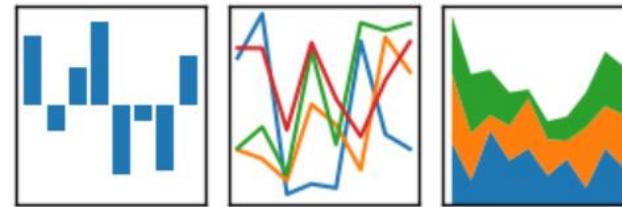


The image displays six Matplotlib plots arranged in a grid:

- Sales data:** Line chart showing Sales units in Number over Month Number (1-12) for Face cream, Face wash, Toothpaste, Bathing soap, Shampoo, and Moisturizer.
- Facewash and facecream sales data:** Bar chart showing Sales units in Number over Month Number (1-12) for Face Cream sales data and Face Wash sales data.
- Profit data:** Histogram showing Active Profit in dollar over profit range in dollar.
- Sales data:** Pie chart showing Sales data distribution for FaceCream, FaceWash, ToothPaste, Bathing soap, Shampoo, and Moisturizer.
- All product sales data using stack plot:** Stacked area chart showing Sales units in Number over Month Number (1-12) for Face Cream, Face wash, Tooth paste, Bathing soap, Shampoo, and Moisturizer.
- Tooth paste Sales data:** Scatter plot showing Number of units Sold over Month Number (1-12) for Tooth paste Sales data.

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



Python Pandas

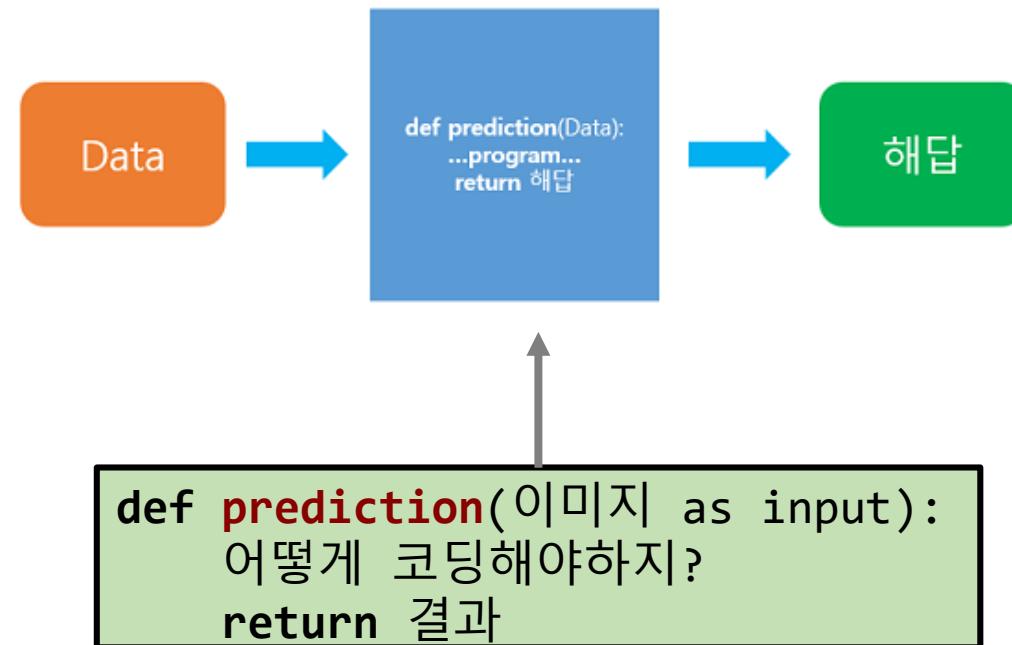


DataFrame Basics

머신러닝 시작하기

❖ 기존의 프로그래밍 접근 방법

- ✓ Ex) 주어진 사진으로부터 고양이 사진인지 강아지 사진인지 판별하는 일.



❖ 기존 프로그래밍의 한계에 대한 해결책: 학습(Train)

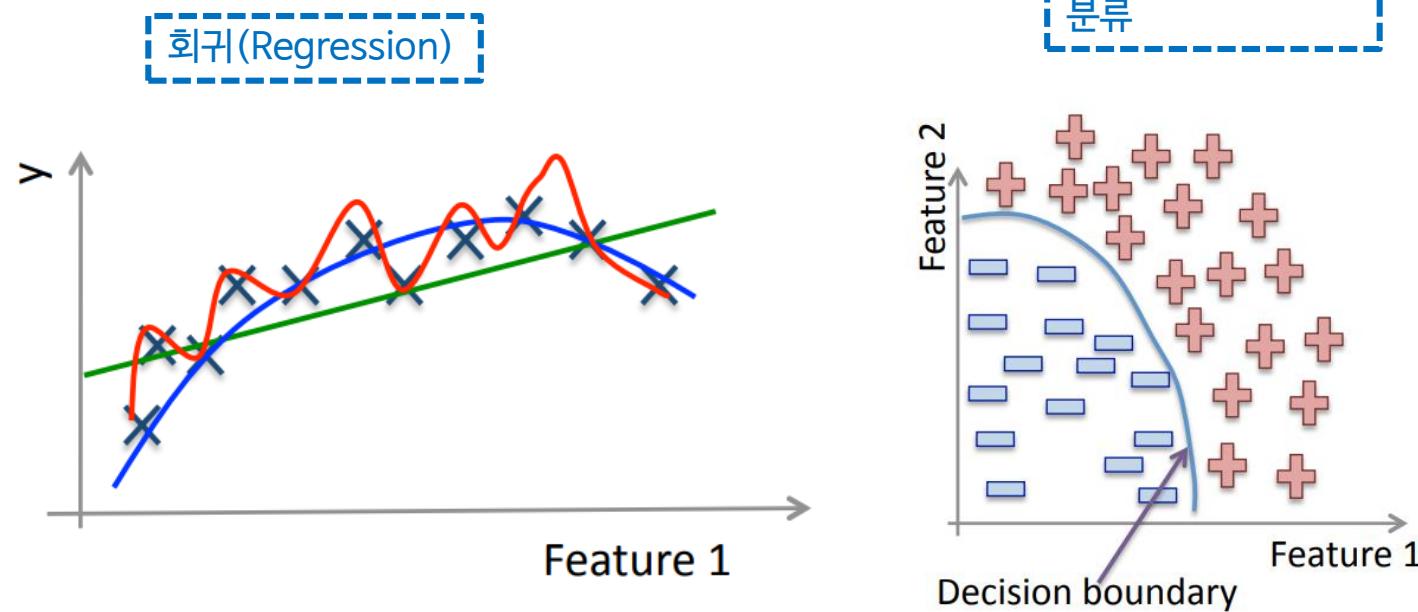
- ✓ 머신 러닝은 주어진 데이터로부터 규칙성 또는 패턴을 찾는 것에 초점이 맞추어져 있다.
- ✓ 주어진 데이터로부터 규칙성을 찾는 과정을 우리는 학습(training)이라고 합니다.
- ✓ 일단 규칙성을 발견해내면, 그 후에 들어오는 새로운 데이터에 대해서 발견한 규칙성을 기준으로 정답을 찾아내는데, 이는 기존의 프로그래밍 방식으로 접근하기 어려웠던 문제의 해결책이 됨



- ❖ 실제 모델 훈련 및 평가하기
- ❖ 데이터를 훈련용, 검증용, 테스트용 이렇게 세 가지로 분리
 - ✓ 테스트 데이터는 20%~30% 정도
 - ✓ 검증용 데이터 약 10%정도는 과적합을 판단하거나 하이퍼파라미터의 조정을 위한 용도



- ❖ 지도학습 알고리즘
- ❖ Support Vector Machines, neural networks,
- ❖ decision trees, K-nearest neighbors, naive Bayes
- ❖ 레이블이 있어야 함. 누가 레이블을 만드나?



❖ 특징

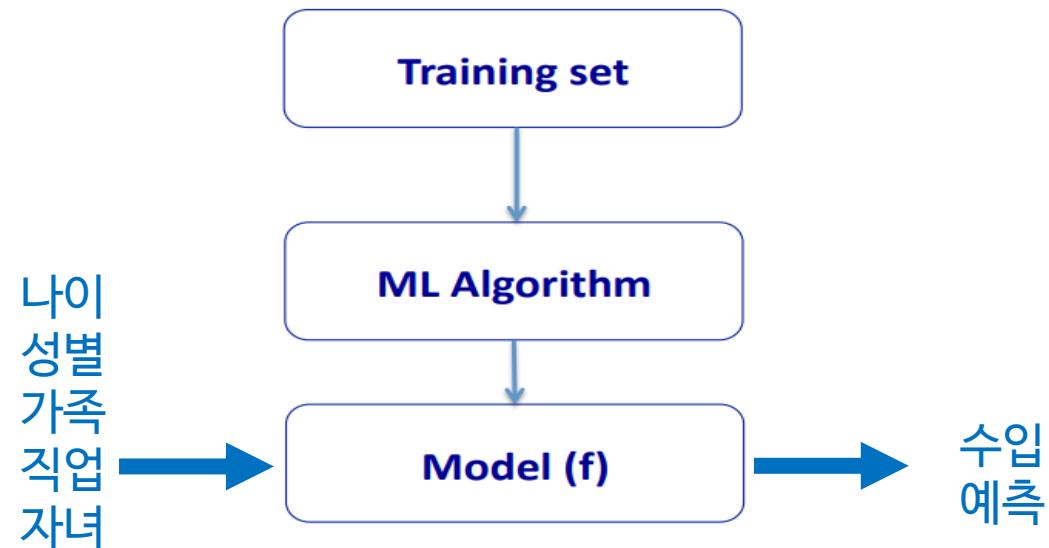
- ✓ 파이썬 머신러닝 중에서 가장 많이 사용되는 라이브러리
- ✓ 예측 데이터 분석을 위해 간단하고 효과적인 툴 제공
- ✓ Numpy, SciPy, Matplotlib을 기반으로 구성
- ✓ 공개 소프트웨어, 상업용을 사용 불가 - BSD 라이선스
- ✓ 최근 텐서플로우, 케라스, 파이토치 등 딥러닝 전문 라이브러리와 경쟁 중

❖ 설치

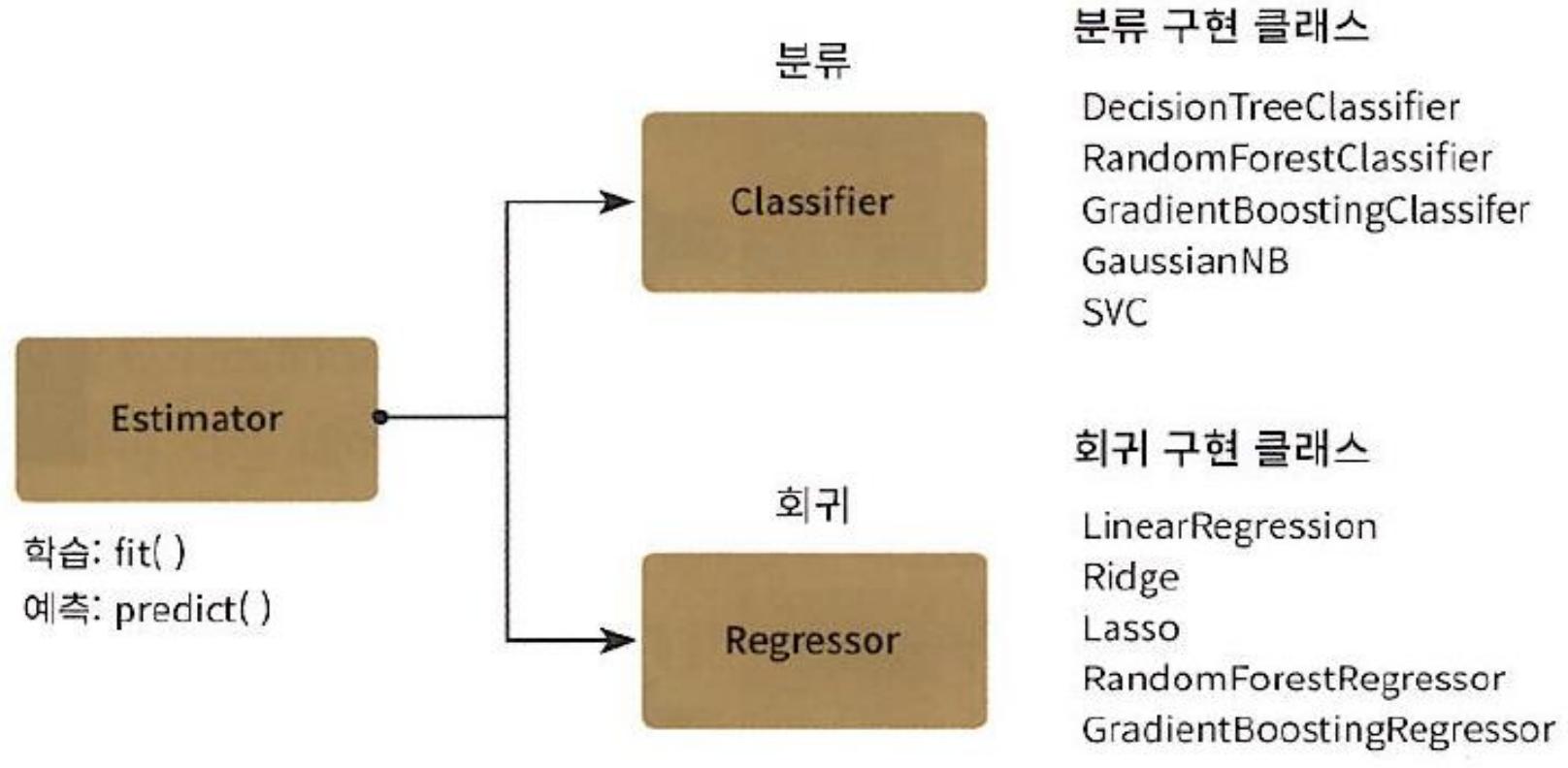
- ✓ 아나콘다를 설치하면 기본적으로 사이킷런까지 설치가 완료 됨
- ✓ 가능하면 conda로 설치를 권장
 - `$ conda install scikit-learn`
- ✓ 버전을 확인
 - `$ print(sklearn.__version__)`

❖ 사이킷런 제공해야 하는 모듈을 생각해보면

- ✓ 피처 처리 (feature processing): 피처의 가공, 변경, 추출
- ✓ 머신러닝 학습/테스트/예측 수행
- ✓ 모델 평가



사이킷런 Estimator 구분



<그림: 파이썬 머신러닝 완벽가이드, 위키북스, 2019>

❖ 지도학습의 분류와 회귀

- ✓ 학습은 fit()으로 하고
- ✓ 예측은 predict()를 이용

❖ Estimator란?

- ✓ regressor + classifier 클래스로 지칭

❖ 비지도학습에서는

- ✓ 차원축소, 클러스터링, 피처추출
- ✓ fit()와 transform()을 적용
 - 여기서 fit()은 지도학습의 fit()과 다름.
 - 입력데이터 형태에 맞춰 데이터를 변환하기 위한 사전 구조를 맞추는 작업
 - 실제 fit()로 사전 구조를 맞추면 이후 입력데이터의 차원변환 등 transform()로 처리
- ✓ fit_transform()도 제공함.

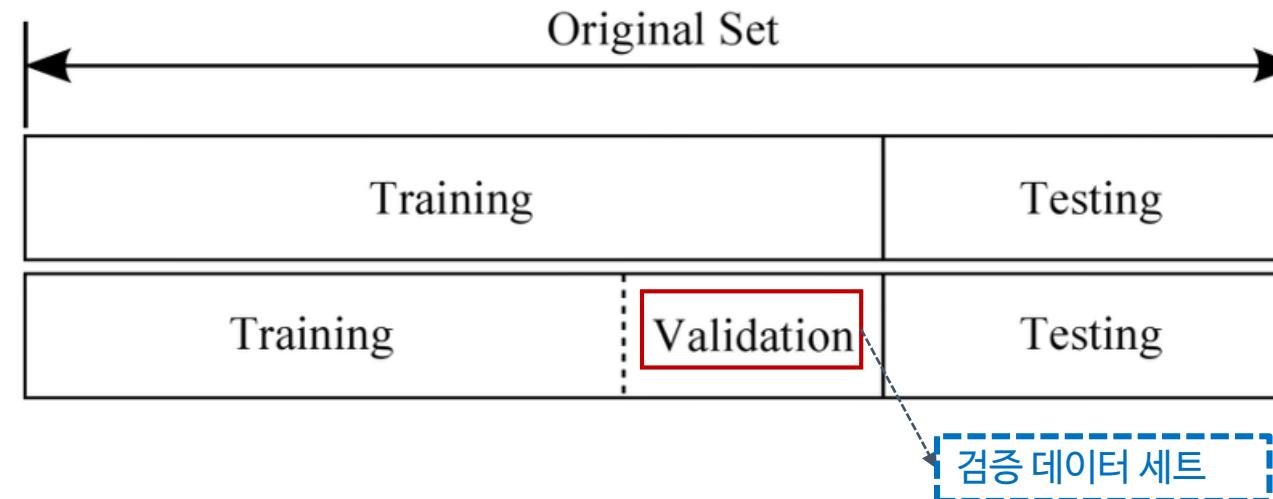
❖ 과적합은 모델이 학습데이터에만 과도하게 최적화 됨

- ✓ 실제 다른 데이터를 테스트하면 정확도학 과도하게 떨어지는 현상이 발생
- ✓ 고정된 학습과 테스트 데이터로 평가하다 보면, 편향되게 모델을 유도

❖ 교차검증이 필요.

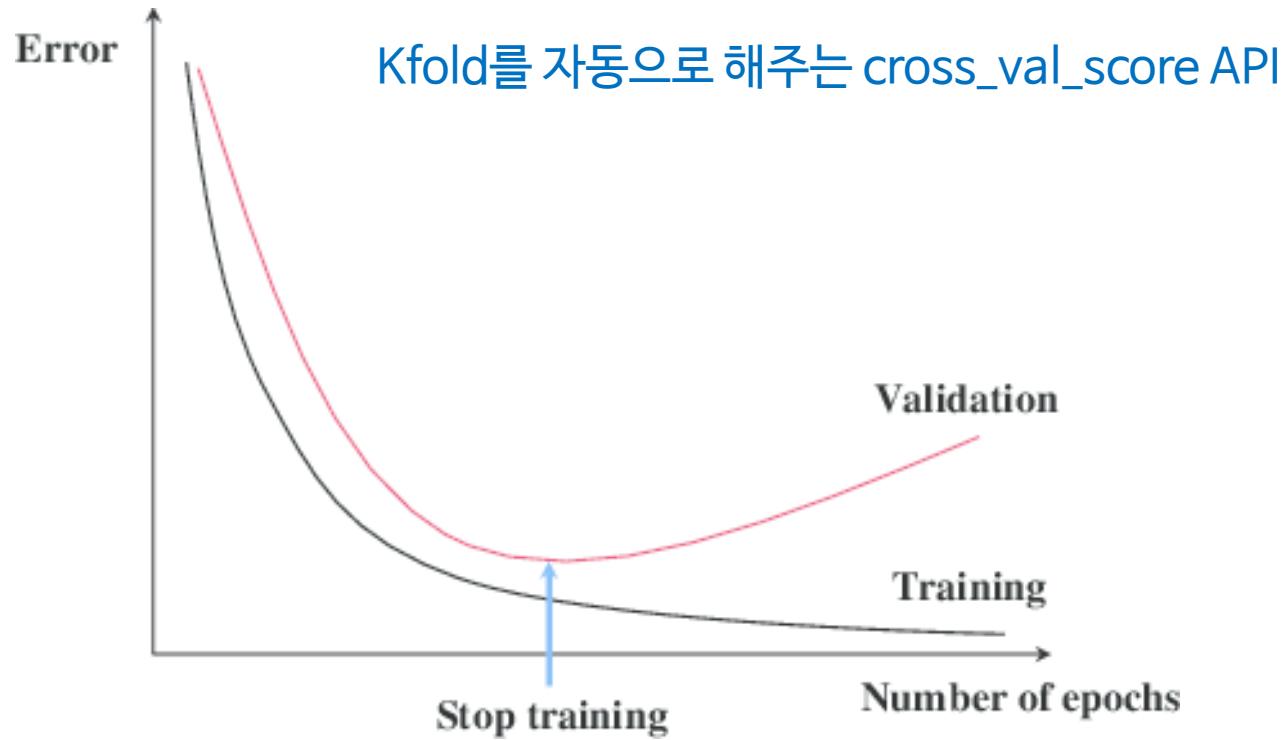
- ✓ 학습/검증/테스트

(0.7/0.1/0.2) 혹은 (0.6/0.1/0.3)



교차검증





```
sklearn.model_selection.cross_val_score(estimator, X, y=None, groups=None, scoring=None,  
cv=None, n_jobs=None, verbose=0, fit_params=None, pre_dispatch='2*n_jobs', error_score=nan)
```

교차 검증 개수 (디폴트=5개)

- ❖ 하이퍼 파라미터(Hyper-parameter)는 무엇인가?

- ✓ 참고, 신경망에서 Learning Rate, Batch size, # Layers 등

- ❖ Hyper-parameter tuning 방법

- ✓ Grid Search
 - ✓ Random Search
 - ✓ Bayesian Optimization

- ❖ GridSearchCV API

- ✓ 균일한 그리드 사용
 - ✓ 촘촘한 파라미터의 최적값을 구함
 - ✓ 가장 기초적이며 많은 계산시간을 요구한다.
 - ✓ High Throughput 계산에 적합

- ❖ 사용자가 직접 적용(지정)해야 하는 변수 :
- ❖ 하이퍼파라미터(초매개변수) : 값에 따라서 모델의 성능에 영향을 주는 변수
- ❖ 학습률(Learning rate), 은닉층 수, 뉴런 수, Batch size 등
- ❖ 매개변수는 (Random)
- ❖ 가중치와 편향과 같은 학습을 통해 바뀌어져가는 변수
- ❖ 하이퍼파라미터 최적화 (Tuning)
- ❖ 모델 최적화를 통해서 진행

❖ 피처 스케일링(feature scaling)

- ✓ 서로 다른 변수의 값 범위를 일정하게 맞추는 작업
- ✓ 표준화(Standardization)
 - 데이터 피처 각각이 평균이 0이고 분산이 1인 가우시안 분포로 변환

$$x_i_new = \frac{x_i - mean(x)}{stddev(x)}$$

- ✓ 정규화(Normalization)
 - 서로 다른 피처의 크기를 통일하기 위해 크기를 변환해주는 개념
 - 값이 0과 1 사이로 변환하는 것

$$x_i_new = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

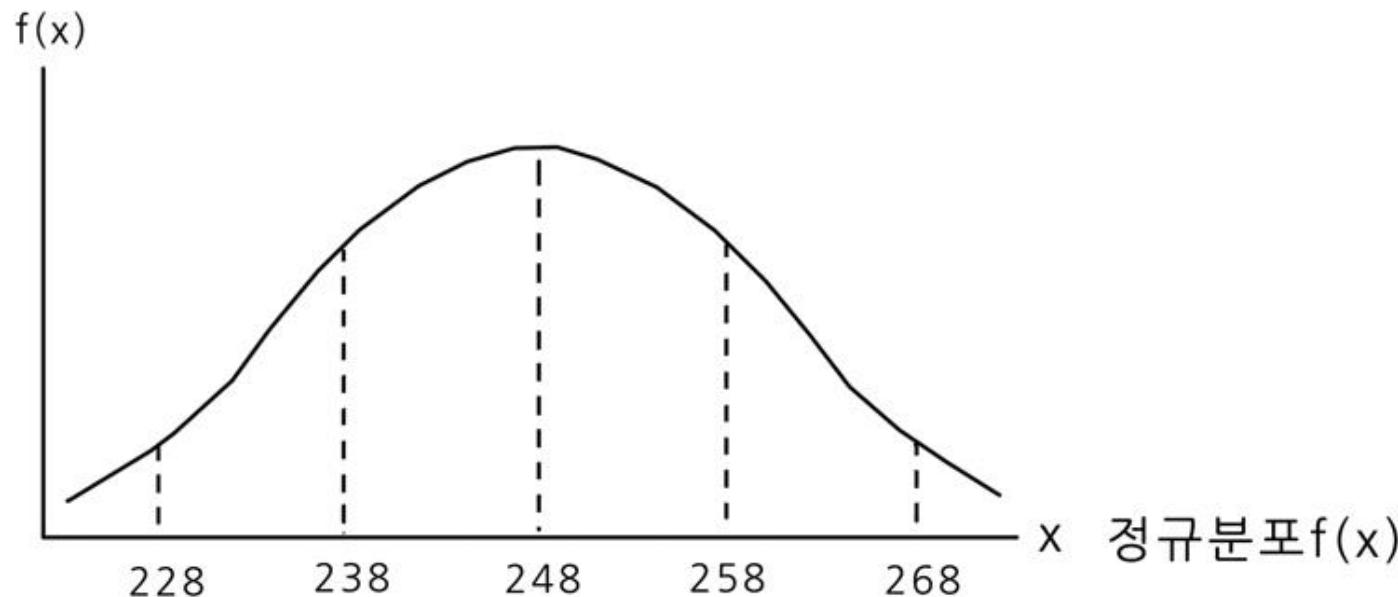
- ❖ 사이킷런에서 제공하는 정규화 함수 Normalizer
- ❖ (벡터) 정규화

- ✓ 선형대수의 정규화 개념이 적용
- ✓ 개별 벡터의 크기에 맞추기 위해 변환
- ✓ 개별 벡터를 모든 피처 벡터의 크기로 나눔

$$x_i_new = \frac{x_i}{\sqrt{x_i^2 + y_i^2 + z_i^2}}$$

❖ StandardScaler는 표준화를 쉽게 지원하는 클래스

- ✓ 가우시안 분포로 RBF 커널을 이용한 SVM, 선형회귀, 로지스틱 회귀는 가우시안 분포를 가지고 있다고 가정하고 구현됨
- ✓ 따라서 표준화를 적용하는 것이 성능향상 예측에 도움이 됨



모델 평가

❖ 분류 문제에서 가장 일반적으로 사용되는 평가 척도는 정확도(Accuracy)이다.

✓ 정확도(Accuracy)

- $7/10 = 0.7 \rightarrow 70\%$

✓ 재현율(Recall)

- $6/7 = 0.857 \rightarrow 85.7\%$

✓ 정밀도(Precision)

- $6/(2+6) = 0.75 \rightarrow 75\%$

✓ f1 값

- $2*(\text{정밀도} \times \text{재현율}) / (\text{정밀도} + \text{재현율}) = 0.7999$

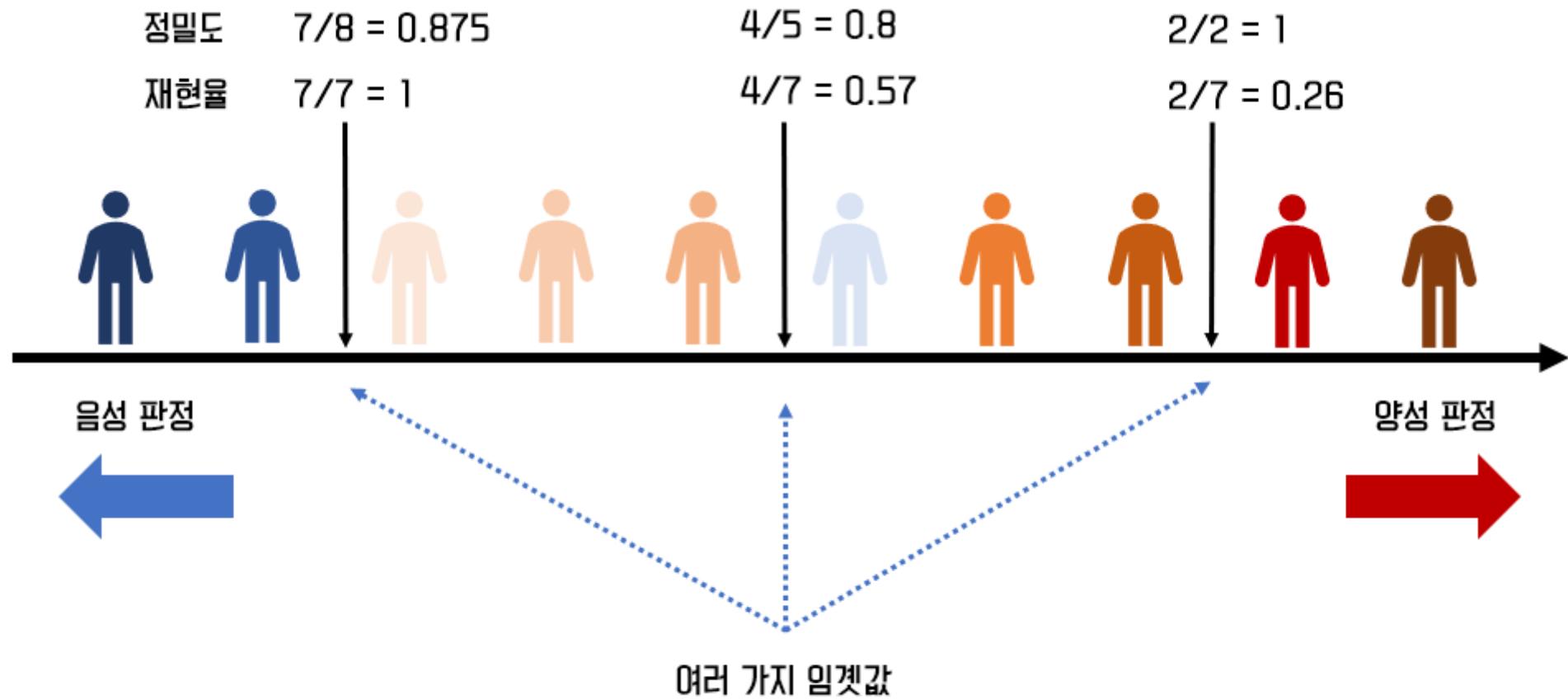
| 실제 상황 (ground truth) | 예측 결과 (predict result) | |
|-------------------------|--|--|
| | Positive | Negative |
| Positive | TP(true positive) 옳은 검출 6명 | FN(false negative) 검출되어야 할 것이 검출되지 않았음 1명 |
| Negative | FP(false positive) 틀린 검출 2명 | TN(true negative) 검출되지 말아야 할 것이 검출되지 않았음 1명 |

$$\text{정확도} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{정밀도} = \frac{TP}{TP + FP}$$

$$\text{재현율} = \frac{TP}{TP + FN}$$

코로나 검사 결과



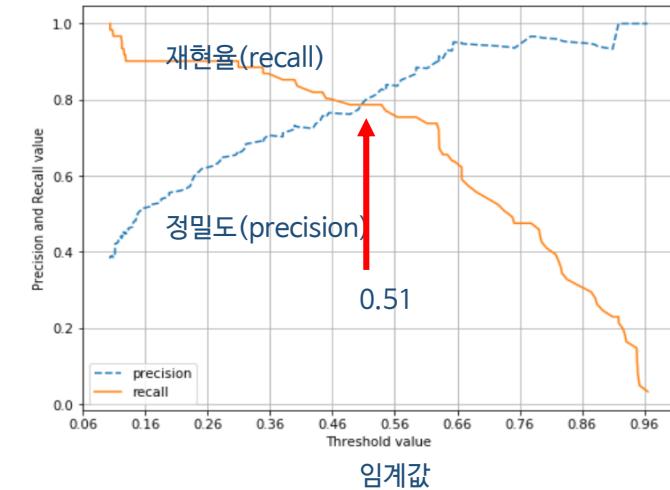
❖ ROC (Receiver Operation Characteristic Curve)

- ✓ 수신자 판단 곡선
- ✓ x 축: FPR(False Positive Rate)
 - TNR(True Negative Rate)은 특이성(Specificity)라고 부름
 - $FPR = FP / (FP + TN) = 1 - TNR = 1 - \text{특이성}$
- ✓ y 축: TPR(True Positive Rate)
 - TPR은 recall(재현율)이다, 민감도(Sensitivity)라고 부름
- ✓ ROC 곡선이 직선일 경우 ($AUC=0.5$)
 - 모델 예측 성능은 떨어짐

❖ 분류의 판단 기준은 FPR과 TPR 변화를 면적 지표로 사용

❖ AUC (Area Under Curve)

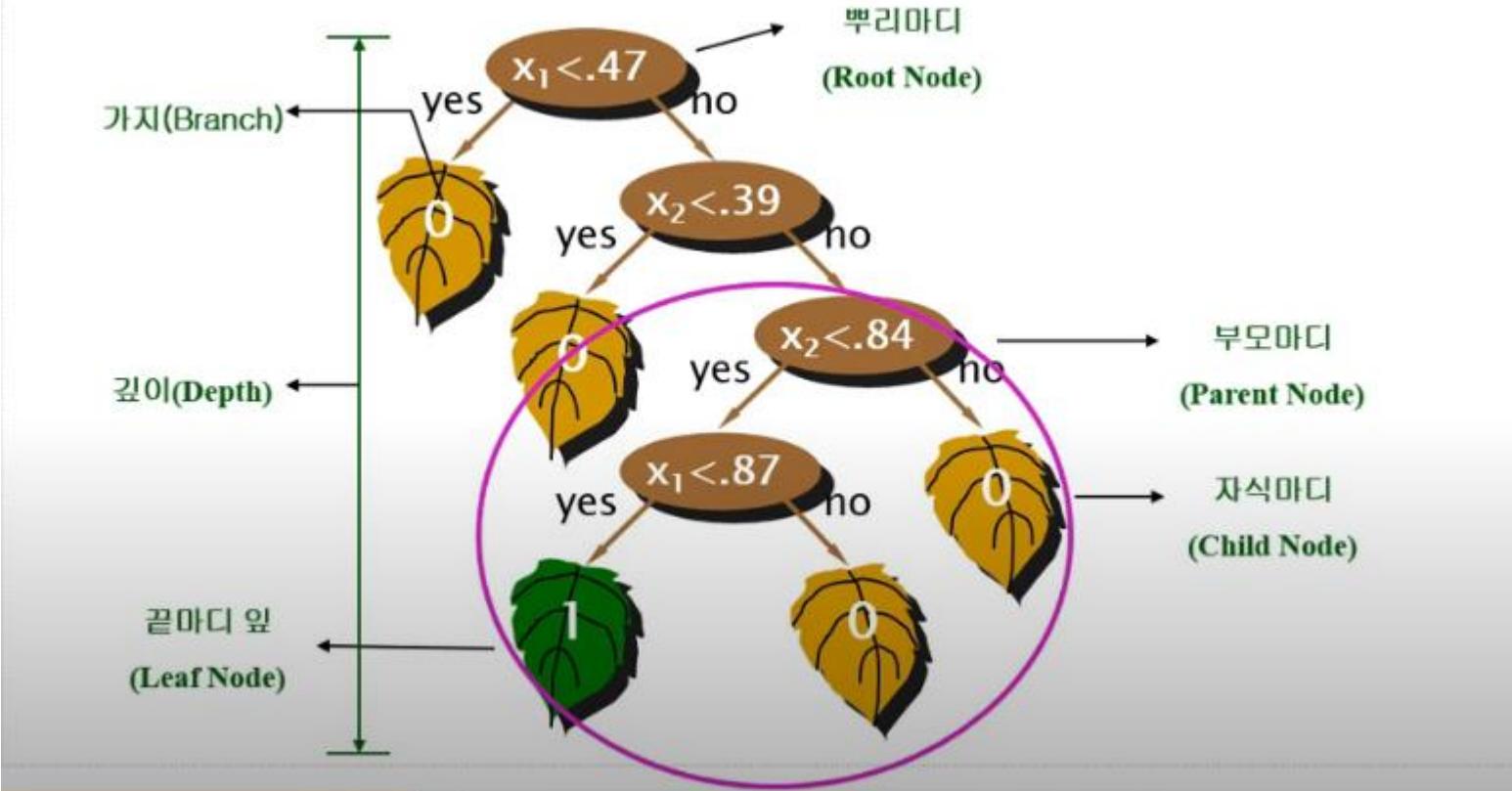
- ✓ 가운데 대각선 ROC 커브의 경우, 면적은 반. 즉 0.5이다.
- ✓ AUC 면적이 1이면 성능이 좋다.



의사결정나무

의사결정나무 구성요소

■ 의사결정나무 구성요소



https://velog.io/@nooooooh_042/%EC%9D%98%EC%82%AC%EA%B2%BD%EC%A0%95%EB%82%98%EB%AC%B4

결정 트리?**

- 결정 트리는 분류 문제를 해결하기 위하여 가장 많이 사용된 지도(supervised) 기계학습 알고리즘의 한 종류
- 물론 회귀(Regression)에도 사용 가능함
- 결정 트리의 결과는 나무 같은 간단한 그래프으로 설명이 되어짐
- 그래서 블랙박스가 아닌 실제로 일어난 일들을 볼수 있음.

결정 트리 용어

루트 노드 - 부모 노드로 알려짐. 데이터셋의 길이와 모든 가지가 여기서 출발 함.

브랜치 - 가지는 루트 노드의 서브 노드로 나눈다. 물론 데이터도 나눔

결정 노드 - 결정노드들은 서브노드를 더 깊게 나눔. 더 이상 나눌 것이 없으면 리프노드임.

리프노드: 더이상 나눌 수 없는 노드.

알고리즘 - '지니'(Gini)는 경제학에서 불평등 지수를 나타냄. 0이 가장 평등하며, 1이 불평등하다.

- 즉, 데이터가 다양한 값을 가질 경우 평등(0)하며
- 특정 값으로 쓸릴경우 불평등(1에 가까움)
- 엔트로피는 무질서도를 나타내며, 무질서도(혼잡도)는 서로 다른 값이 섞여 있으면 높다. 혼잡도가 높으면 1, 적으면 0

$$\text{지니 지수: } G = 1 - \sum_i^c p_i^2 , \quad 0 \leq G \leq 1/2$$

$$\text{엔트로피 지수: } E = - \sum_i^c p_i \log_2 p_i , \quad 0 \leq E \leq 1$$

❖ 기본적으로 gini를 사용 Entropy (엔트로피) 불순도를 사용해도됨

- ✓ 엔트로피는 분자의 무질서도를 측정하는 것 (열역학의 개념)
- ✓ 문자가 안정되고 질서 정연하면 엔트로피는 0에 가까움
- ✓ 정보이론에서 모든 메시기가 동일할때 에트로피는 0에 가까움

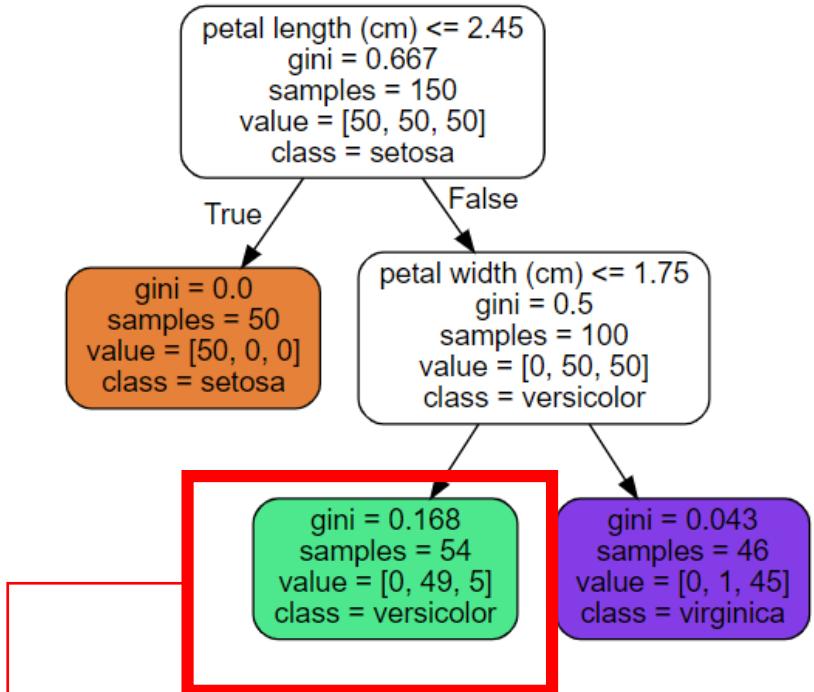
❖ 엔트로피 정의

$$H_i = - \sum_{k=1}^n p_{i,k} \log_2(p_{i,k})$$

❖ 머신러닝에서

- ✓ 어떤 세트가 한 클래스의 샘플만 있는 경우는 엔트로피가 0이다.
- ✓ Gini와 엔트로피 중 어떤 것을 사용해도 실제로는 큰 차이가 없다.
- ✓ Gini 불순도가 계산이 조금 더 빠르기 때문에 기본값으로 좋다.

붓꽃의 결정트리의 엔트로피 계산



$$-\frac{49}{54} \log_2 \left(\frac{49}{54} \right) - \frac{5}{54} \log_2 \left(\frac{5}{54} \right) \approx 0.445.$$

- ❖ 장점

- ✓ 데이터의 전처리가 거의 필요없다.
- ✓ 특성 스케일을 맞추거나, 원점에 맞추는 작업은 필요가 없다.

- ❖ 노드의 속성은 얼마나 많은 훈련 샘플이 적용되었는지 헤아린 것이다.

- ✓ 100개의 훈련 샘플의 꽃의 길이가 2.45cm보다 길고(깊이1, 오른쪽)
 - 그 중에 54개 샘플이 1.75보다 짧고 (깊이2, 왼쪽)

- ❖ 노드의 value 속성은 노드에서 각 클래스에 있는 훈련 샘플 개수다.

- ✓ 맨 오른쪽 아래 노드 Setosa=0, Versicolor=1, Virginica=45개

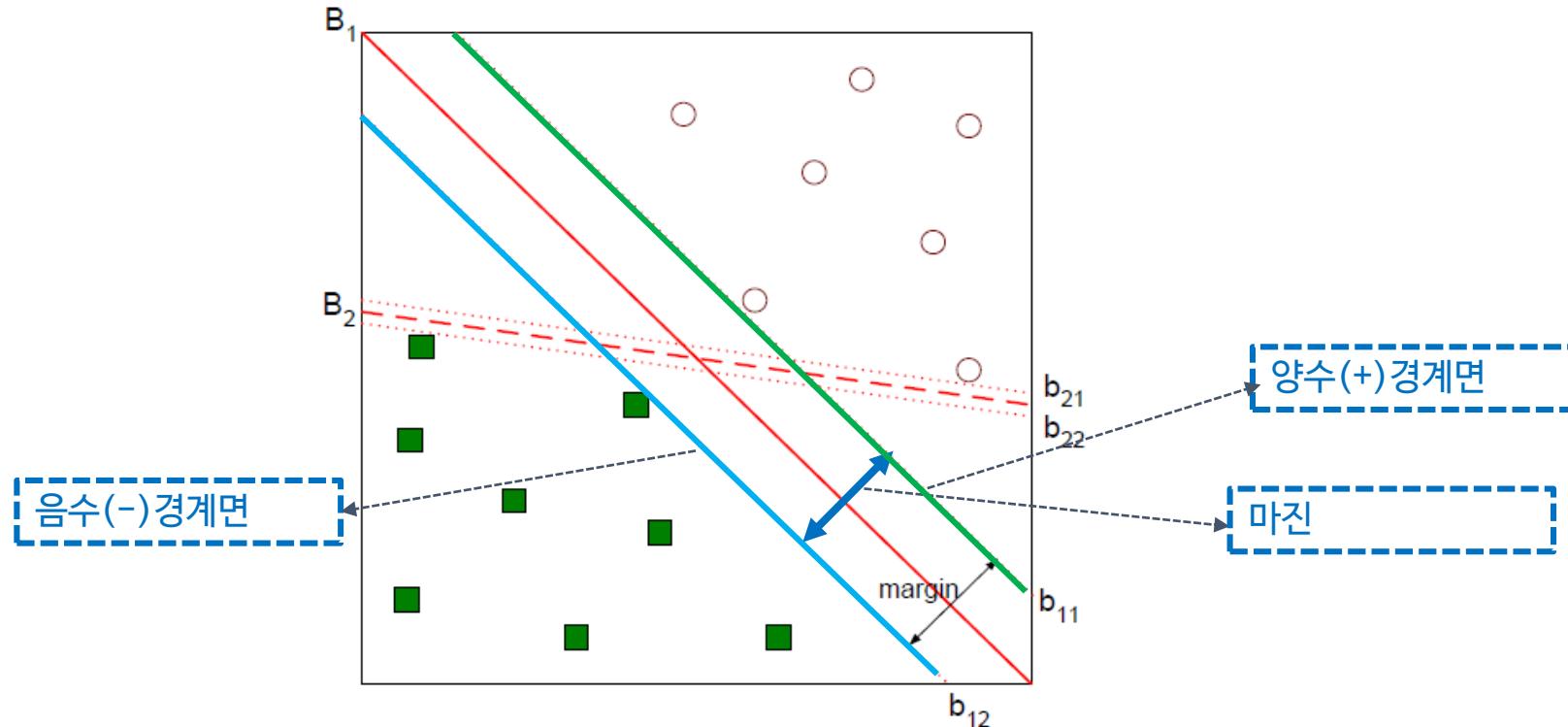
- ❖ 결정트리는 화이트 모델로 직관적이고 이해하기 쉽다.

- ❖ 램덤 포레스트나 신경망은 블랙박스 모델이다.

SVM

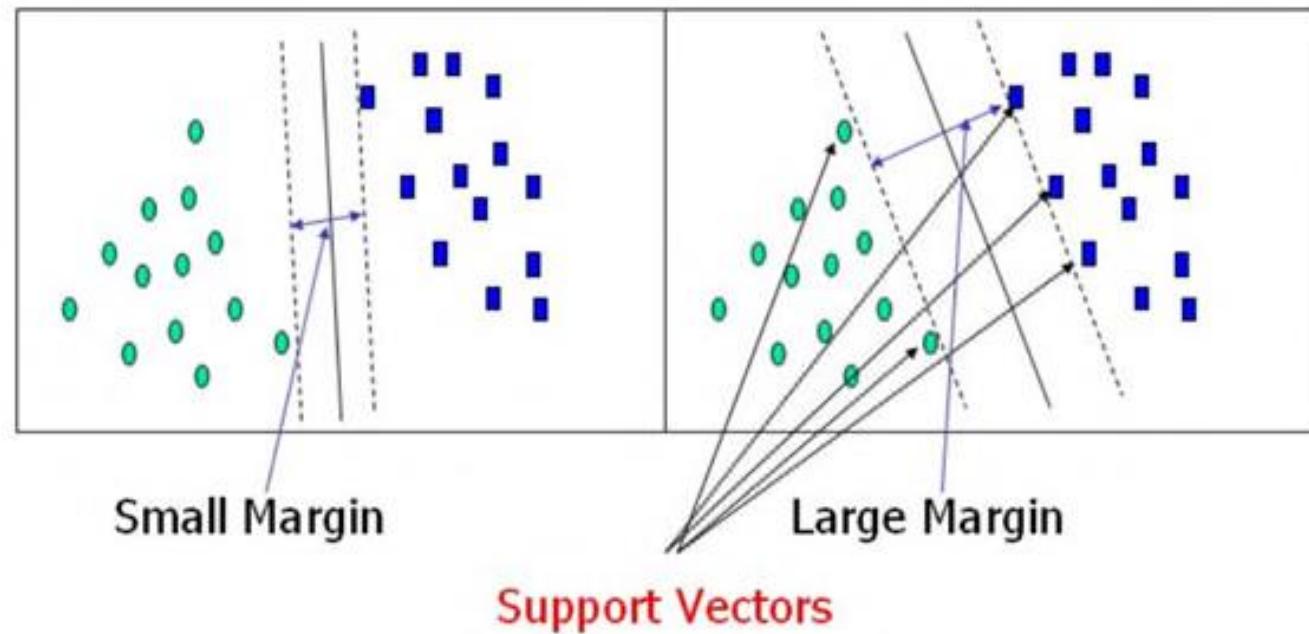
마진(margin)이란 무엇인가?

- ❖ 결정 경계를 정하기 위해 마진(margin)을 도입하자.
 - ✓ 마진은 두 경계면 사이의 거리이며, 이 거리가 최대일때 좋은 결정 경계를 얻음.
- ❖ SVM의 목적은 마진을 최대화하는 경계면을 찾는 것

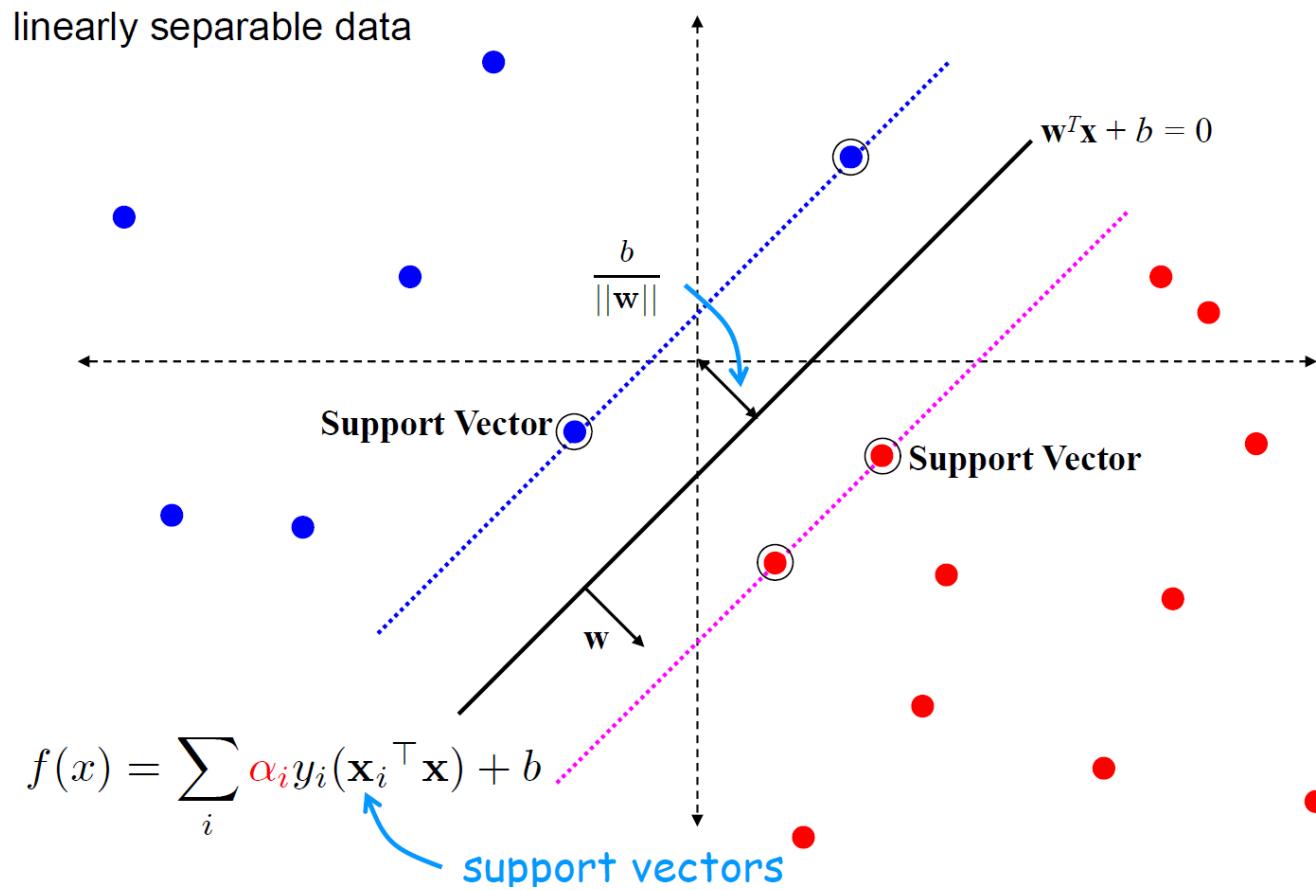


❖ SVM 정리

- ✓ 직관적으로 자료를 군집별로 잘 분리하는 초평면은 가장 가까운 훈련용 자료까지의 거리가 큰 경우
- ✓ 최대 마진을 가지는 선형판별에 기초하며, 속성들 간의 의존성을 고려하지 않는 방법
- ✓ 마진이 가장 큰 초평면을 분류기로 사용할 때, 새로운 자료에 대한 오분류가 가장 낮다

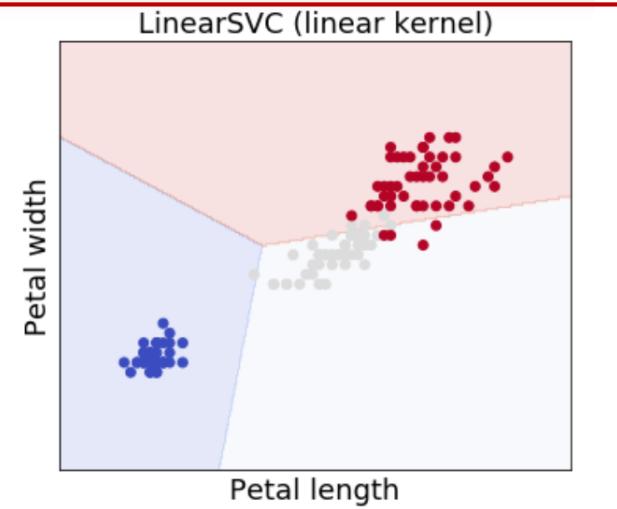
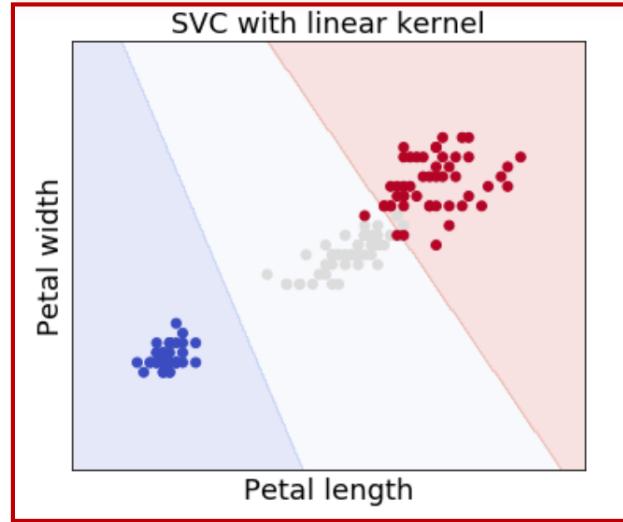


- ❖ 알파는 해당 벡터(x)가 경계선을 정하는 샘플이라는 뜻으로 “서포트 벡터 ”라고 부름

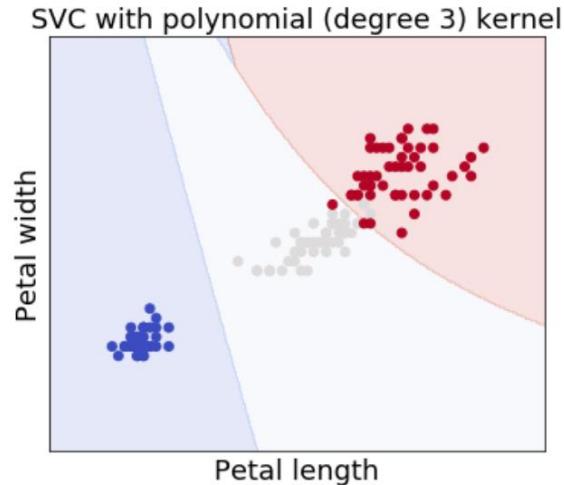
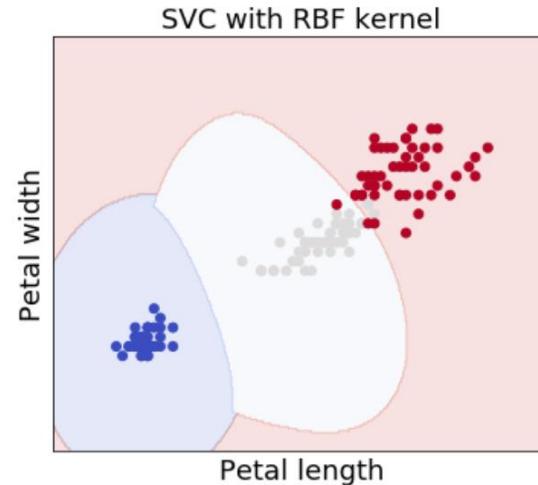


붓꽃에 SVM.SVC 적용하기 (7)

IRIS PETAL (꽃잎) 데이터를 SVM 4개의 다른 종류로 분류

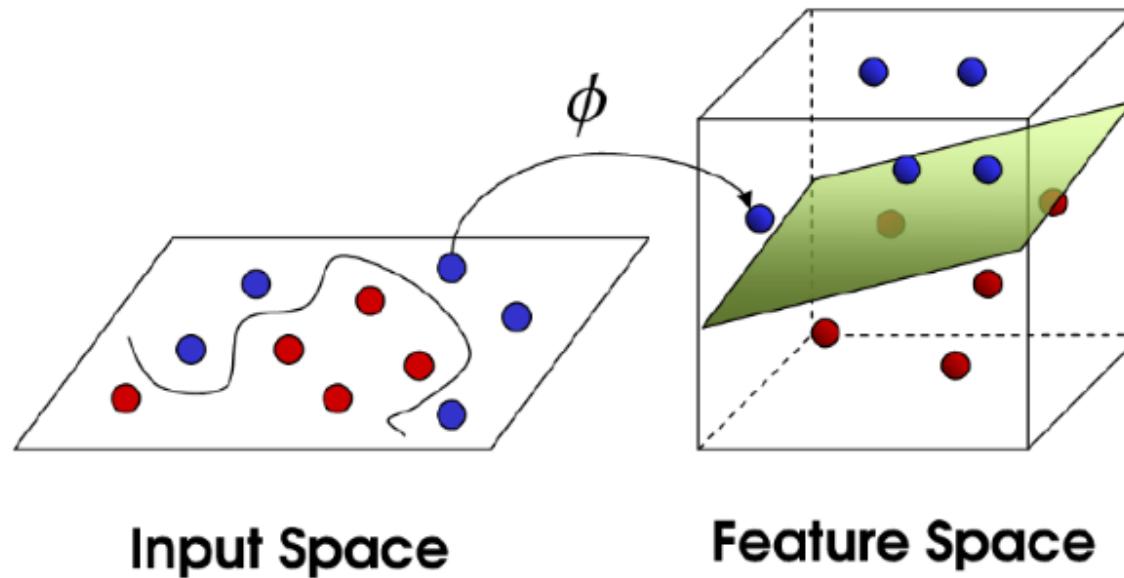


Linear
분류 성능을
비교하면,
어느 것이
더 우수한가?



❖ 선형으로 분리 되지 않을 경우

- ✓ 원공간(Input Space)의 데이터를 선형분류가 가능한 고차원 공간(Feature Space)으로 매핑한 뒤 두 범주를 분류하는 초평면을 찾는다.



Random Forest

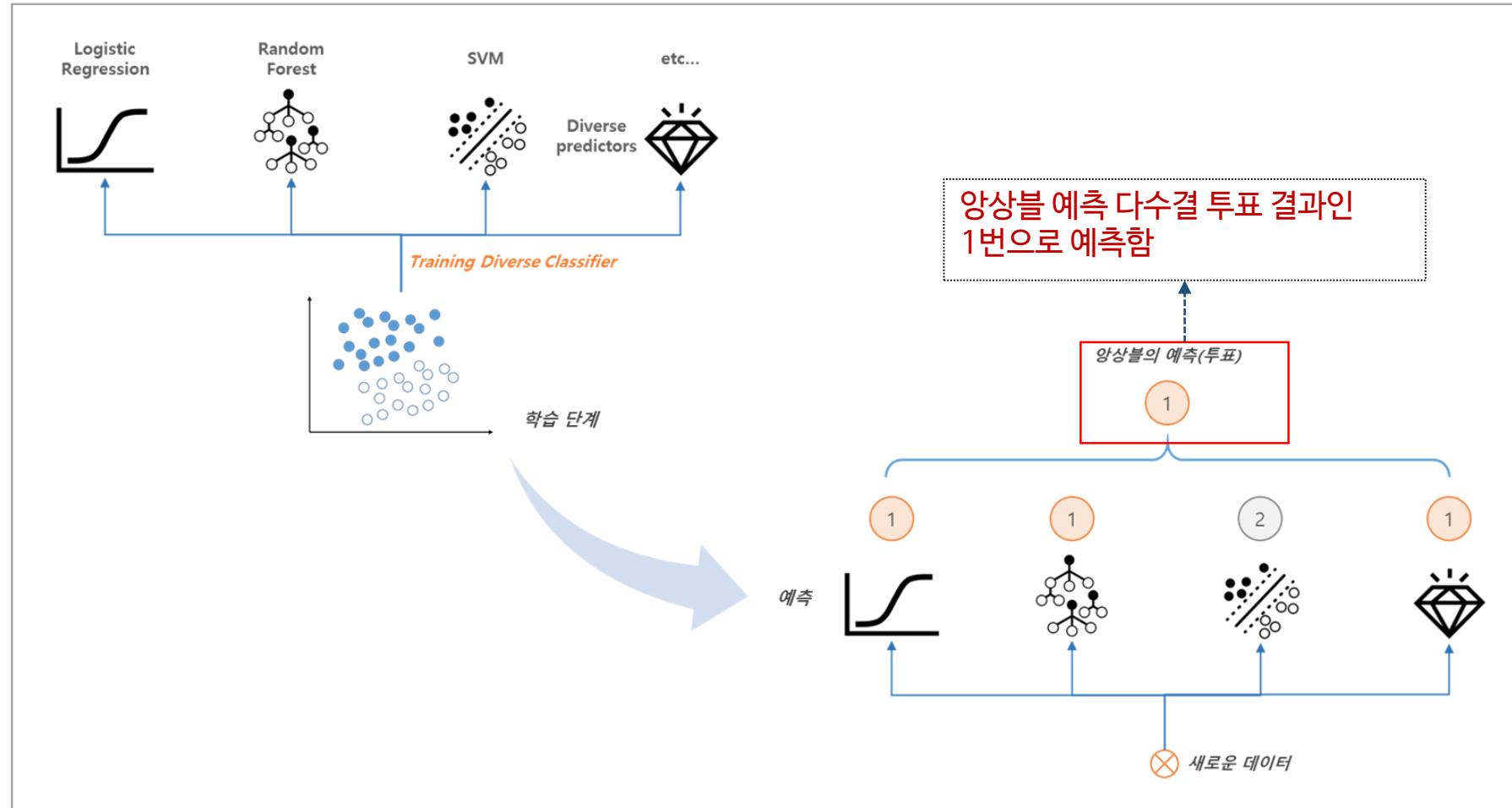
❖ 양상블 학습 개요

- ✓ 여러 개의 분류기를 생성하고 그 예측을 결합. 보다 정확한 최종 예측을 도달함
 - 대중의 지혜(wisdom of the crowd)
 - 단일 분류기 보다 신뢰성이 높은 예측 값을 얻는 것이 핵심
- ✓ 딥러닝이 뛰어난 분야는 비정형 데이터인 이미지, 영상, 음성 분야
- ✓ 양상블이 뛰어난 분야는 정형 데이터 분류, 램덤포레스트와, 그레디언트 부스팅
- ✓ 대표 알고리즘
 - XGboost, 훨씬 빠른 LightGBM, 메타 모델을 수립하는 스태킹(Stacking)
- ✓ 양상블 학습은 일련의 예측기(분류, 회귀)로부터 예측을 수집

❖ 양상블 학습 유형

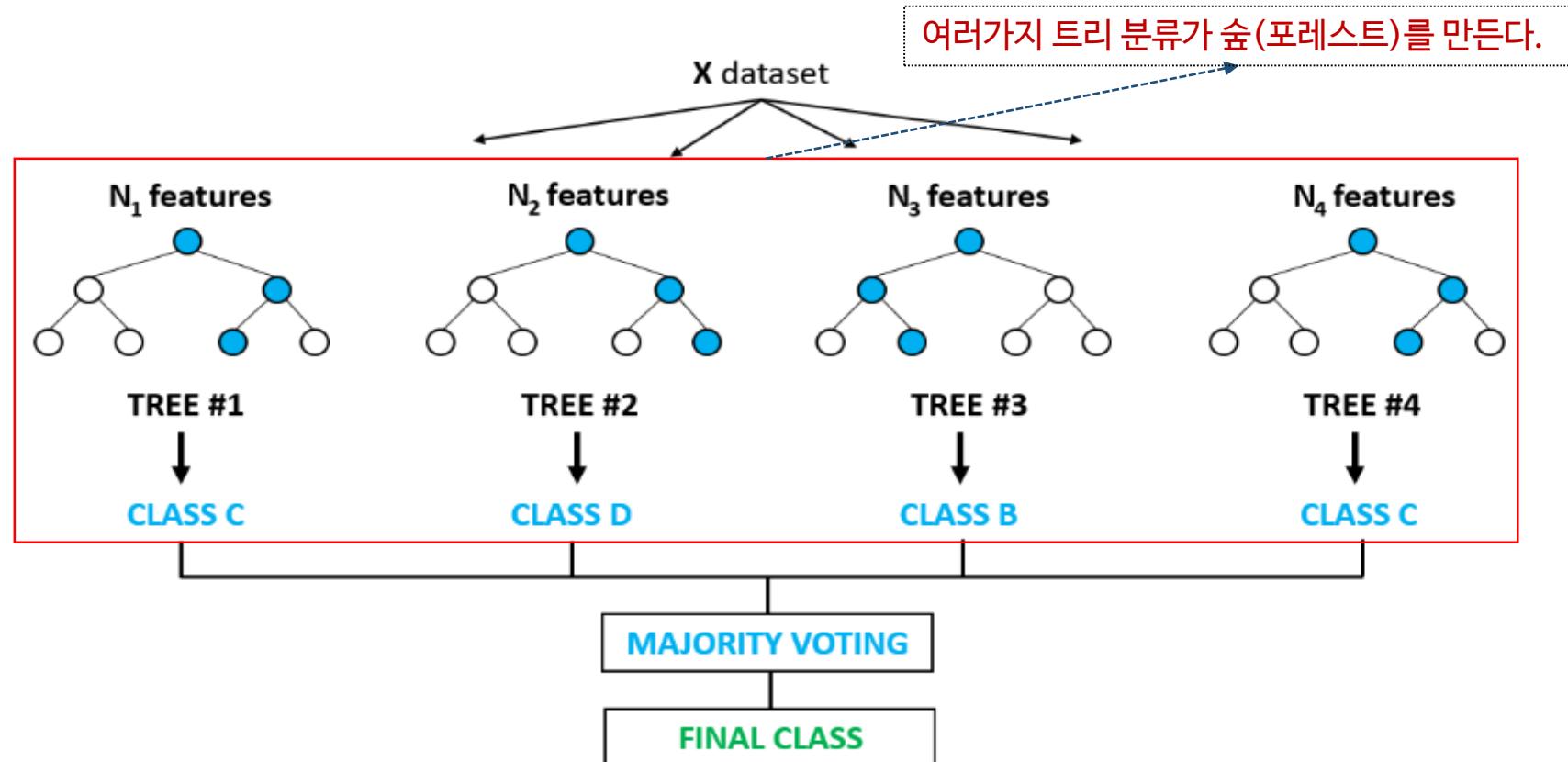
- ✓ 보팅(Voting) : 여러 개의 분류기가 투표를 통해 최종 예측
 - 서로 다른 알고리즘(knn, svm, kmeans)을 가진 분류기를 결합하는 방식
- ✓ 배깅(Bagging) : 여러 개의 분류기가 투표를 통해 최종 예측
 - 분류기는 같은 유형의 알고리즘이지만, 데이터 샘플링을 서로 다르게 학습
 - 대표적인 배깅 방식은 램덤 포레스트 알고리즘이다.
 - 부트스트래핑(Bootstrapping) 분할 방식을 사용, 즉, 원본 학습 데이터를 샘플링해서 추출하는 방식
 - 배깅은 교차 검증과 달리 데이터 셋트 간의 중첩을 허용하는 것이 차이점이다.
- ✓ 부스팅(Boosting)
 - 앞에서 학습한 분류기가 예측이 틀린 데이터에 대해서 올바른 예측을 할 수 있도록 다음 분류기에는 가중치 (weight)를 부여하면서 학습과 예측을 진행하는 것
- ✓ 스태킹(Stacking)
 - 여러 가지 다른 모델의 예측 결과 값을 다시 학습 데이터로 만들어서 다른 모델(메타모델)로 재학습시켜 결과를 예측하는 방법

❖ 다수결 투표 결과로 정해지는 직접 투표(Hard voting) 분류기



❖ 랜덤 포레스트 개요

- ✓ 배깅의 대표적 알고리즘은 랜덤 포레스트
- ✓ 여러 개의 결정 트리 분류기가 전체 데이터에서 배깅 방식으로 보팅 함

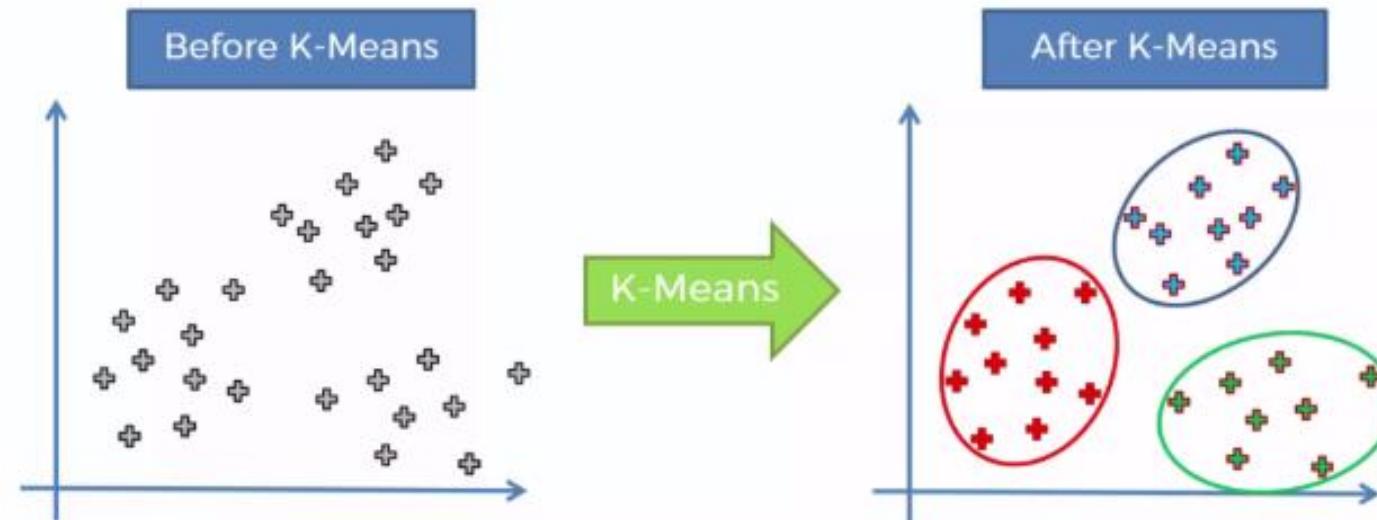


k-NN 알고리즘

K Means

❖ K-Means는 비지도학습(Unsupervised Learning)

- ✓ 주어진 데이터를 k 개의 클러스터(군집)로 묶는 알고리즘
- ✓ "k"는 각 데이터 점들의 서로에 대한 유사성을 기초로 한 고정된 수(k)의 군집을 찾는다는 것을 의미함
- ✓ 각 클러스터간의 거리 차이의 분산을 최소화

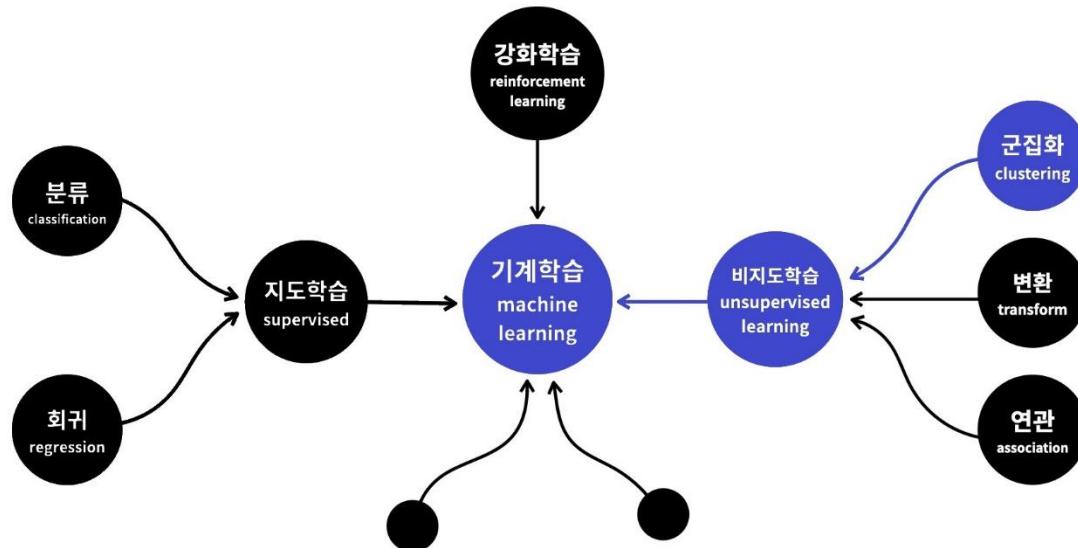


❖ K-means 알고리즘은 K-NN 알고리즘과 유사하면서도 다르다.

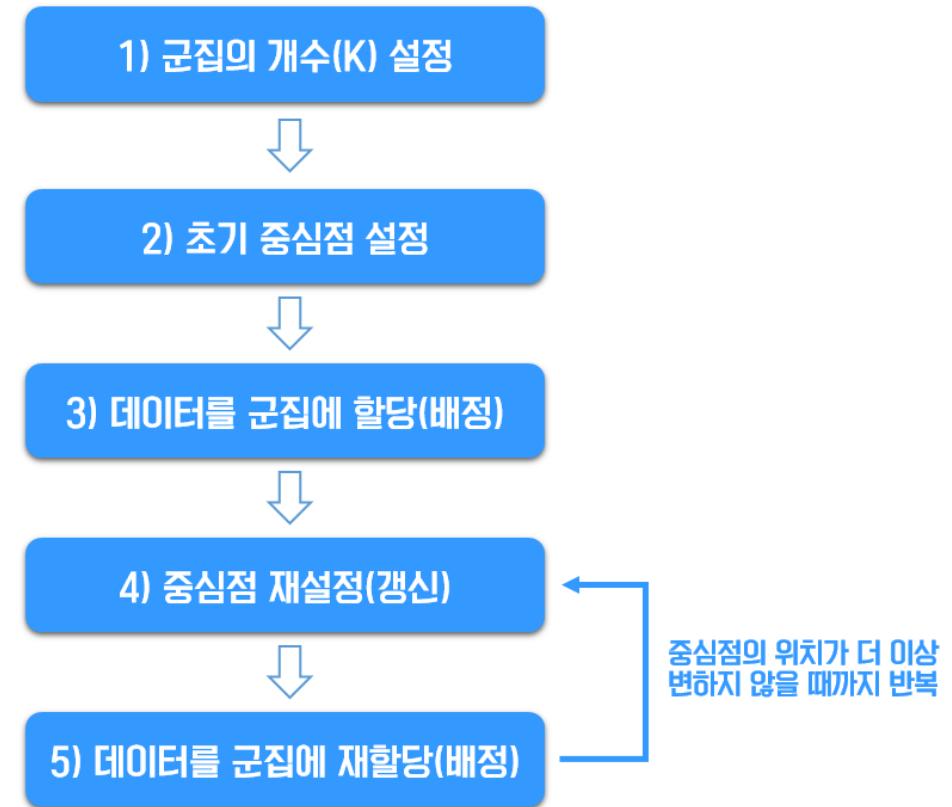
- ✓ 둘은 모두 K개의 점을 지정하여 거리를 기반으로 구현되는 알고리즘

❖ 차이점

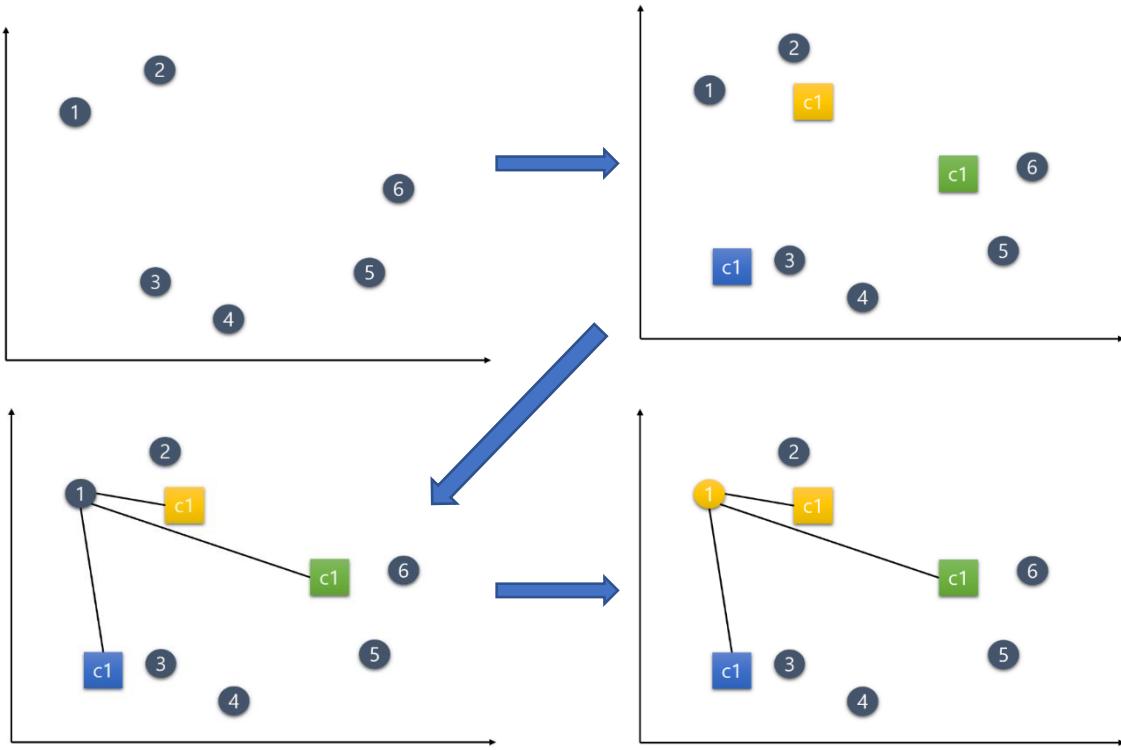
- ✓ 지도학습에 속하는 K-NN 알고리즘은 분류를 위한 알고리즘이다.
- ✓ K-means는 비지도학습 방법으로 군집화(Clustering) 알고리즘이다.



❖ K-Means 알고리즘 단계



K-Means 군집화 과정 (예제)

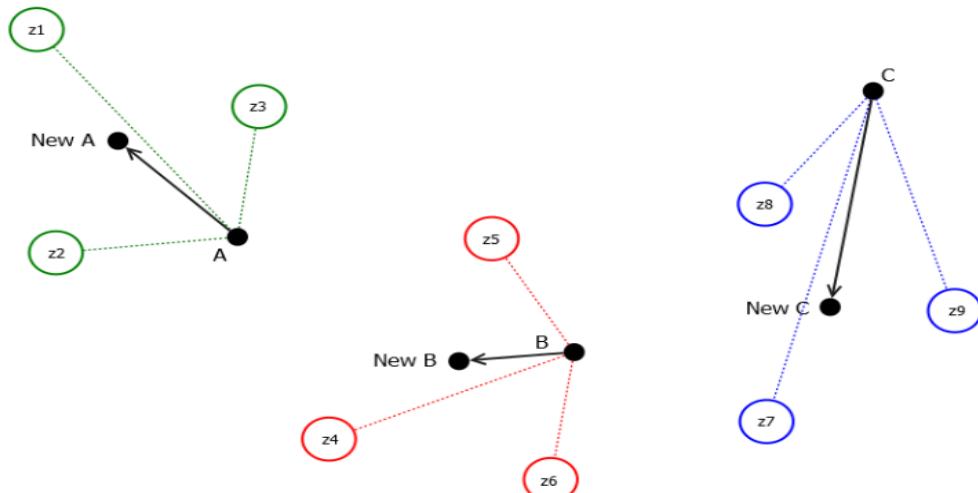


1번 데이터의 경우 c1 중심점과 가장 가까우므로 노란색으로 바뀐다.

K-Means 개요 및 알고리즘의 3단계

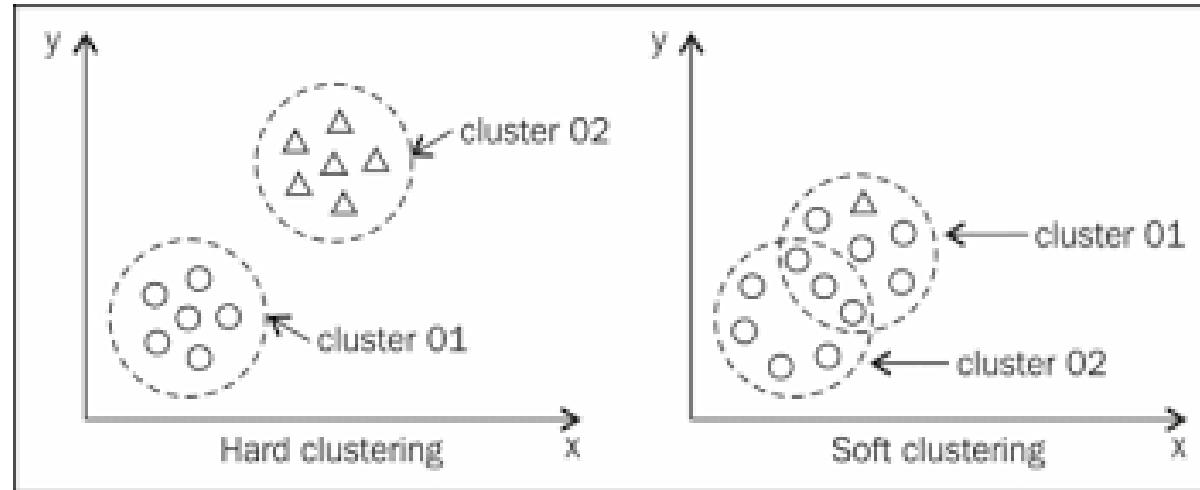
$$J = \sum_{k=1}^K \sum_{i \in C_k} d(x_i, \mu_k)$$

군집 개수
중심점 개수
군집에 속하는 데이터 수
 $d(x_i, \mu_k) = \|x_i - \mu_k\|^2$
2개 데이터 사이의 거리



❖ 군집 분석 방법

- ✓ 하드 클러스터링 : 하나의 데이터가 정확히 하나의 군집에 할당하는 것
- ✓ 소프트 클러스터링 : 하나의 데이터가 다수의 군집에 할당하는 것



❖ K-Means의 최상의 모델 평가를 위해서는

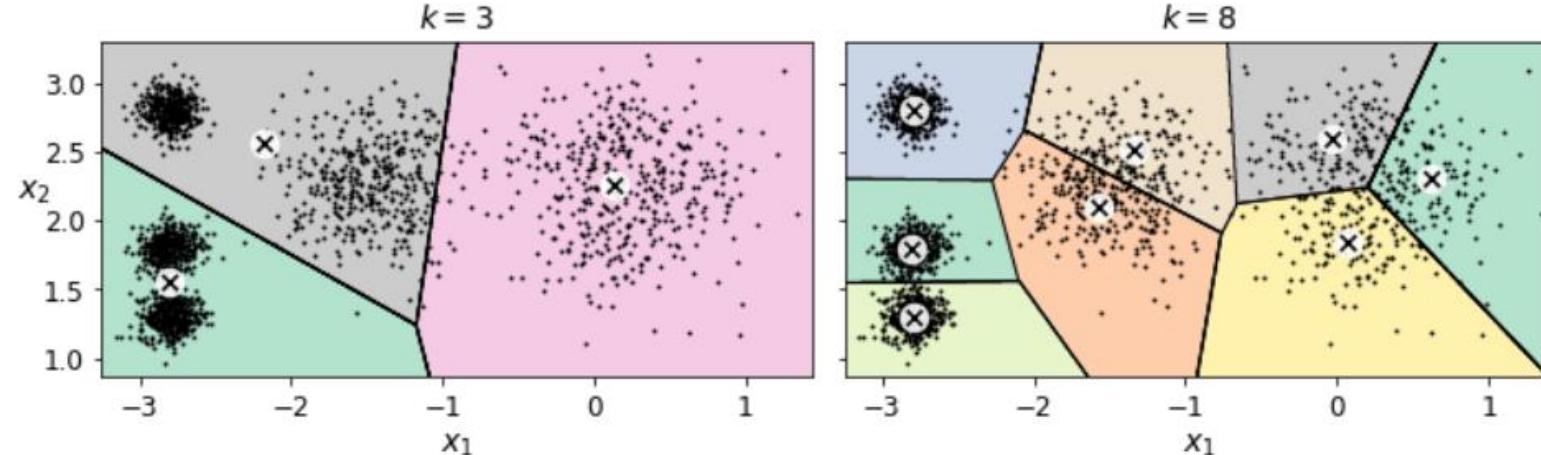
- ✓ 비지도학습, 레이블이 없는 관계로 어렵다
- ✓ 하지만 중심점에 대해 거리를 알고 있어서 Inertia를 사용

❖ Inertia

- ✓ 각각의 데이터와 가장 가까운 중심점 제곱의 거리

Inertia ~ 최적의 모델을 주는 k는

Inertia를 최소화 하는 k를 찾는 거으 쉬지가 안다 k가 많으 수록 inertia는 잡다



2022

Korea Institute of Science
and Technology Information

TRUST
KISTI

