소프트웨어와 문제해결

Dr. Young-Woo Kwon

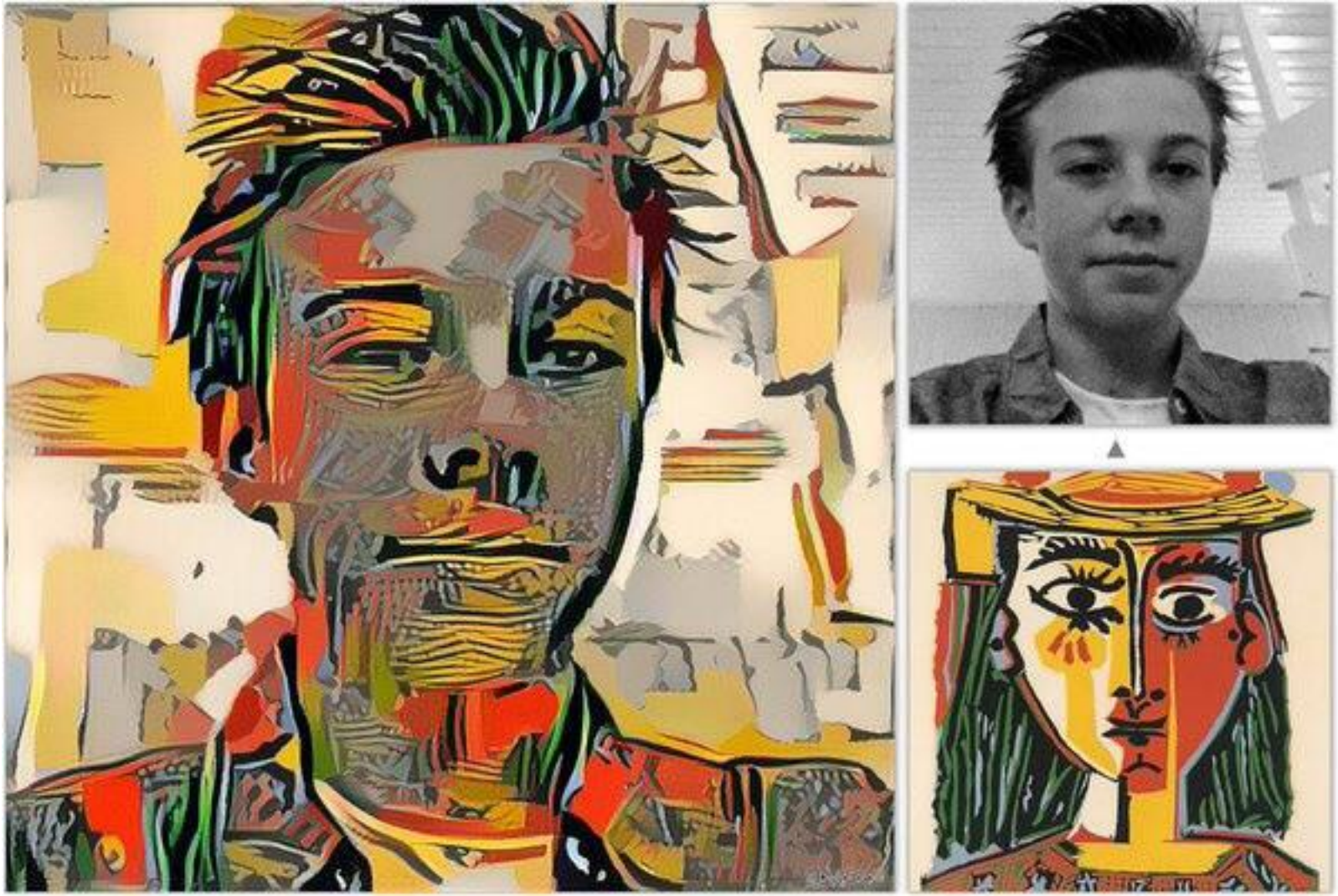
인공지능 (AI), 기계학습, 딥 러닝

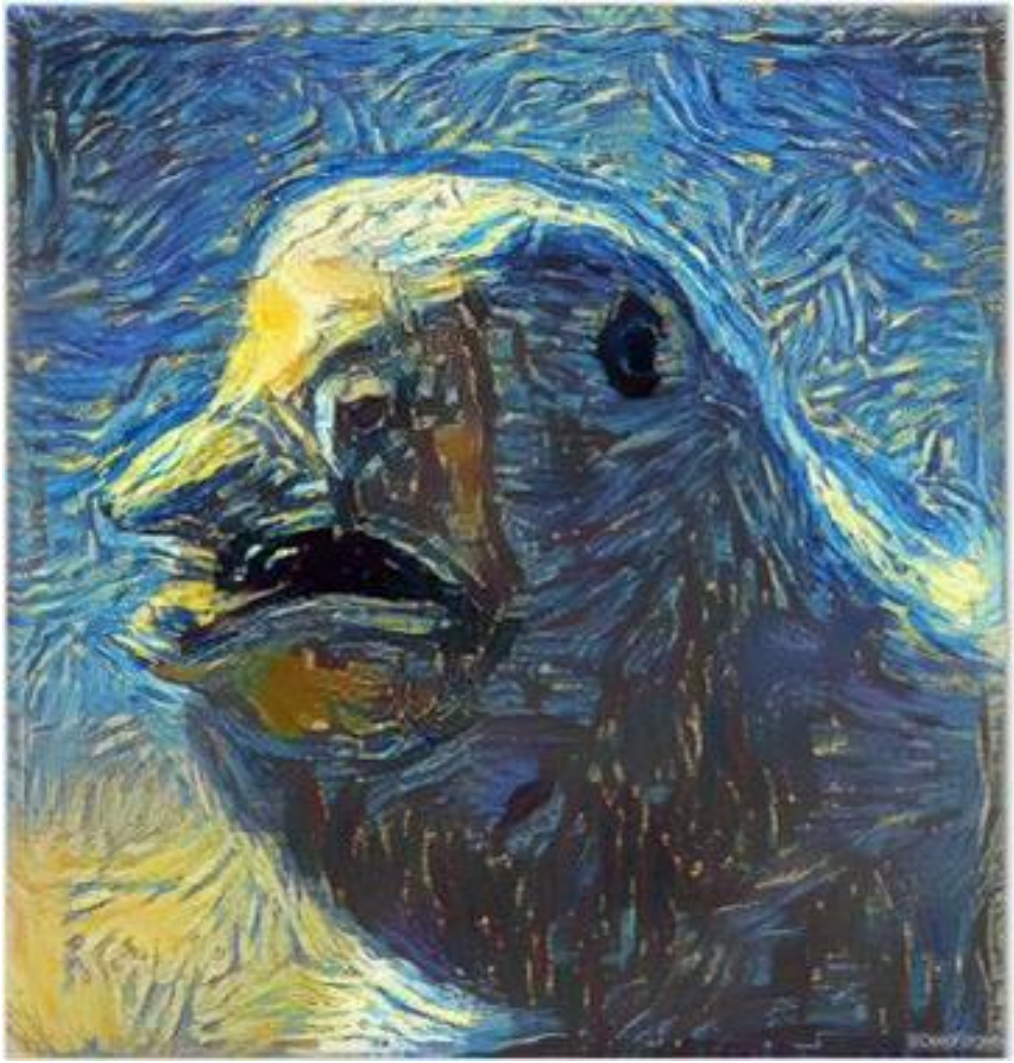

KBA
Korea Baduk Association
Google DeepMind
Challenge Match
8 - 15 March 2016
AlphaGo
ALPHAGO 01:58:22
LEE SEDOL 01:59:27
AlphaGo
Lee Sedol
Google DeepMind
Challenge Match
8 - 15 March 2016

KNU

KNU

마이크로 소프트와 네덜란드 기술자들이 만든
AI"넥스트 렘브란트'가 그린 렘브란트의 초상화





KNU

기계학습(머신러닝) 개요

강아지

머핀

KNU

강아지?

머핀?

source: boredpanda.com

source: boredpanda.com

어려운 계산을 너무나도 쉽게 하는 컴퓨터.

그렇다면 ‘언어처리’ ,‘얼굴인식’ 같이 인간에게 쉬운 문제 역시 쉽게 풀 수 있지 않을까?

하지만, 현실은…

정보

- 정량적 정보

숫자로 측정하고 표현하는 정보

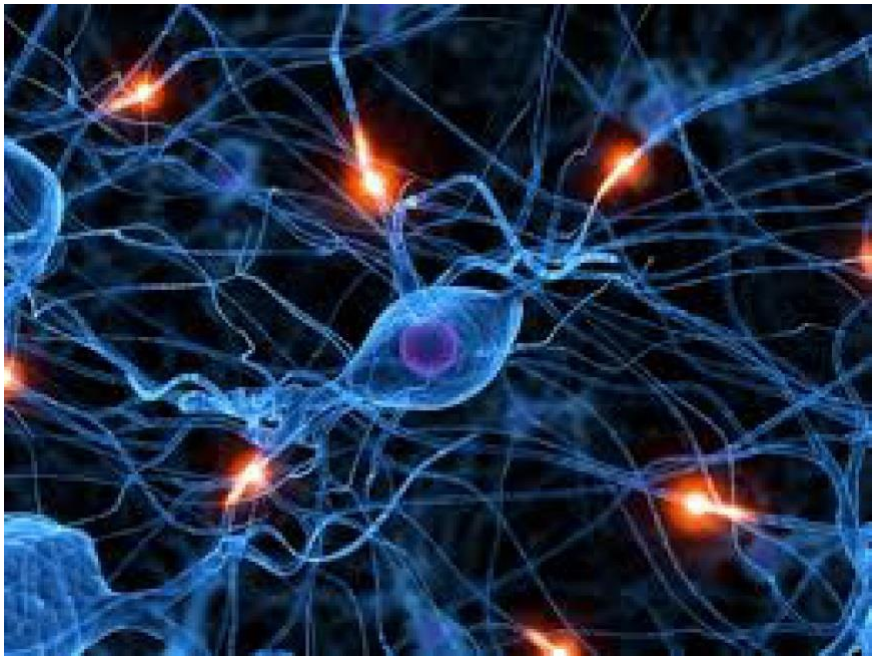
- 비정량적 정보

숫자가 아닌 주관이 들어가는 정보

KNU

직관

완벽한 설명이 어려운 비 정량화된 정보를 통해 이루어지는 행위



시냅스

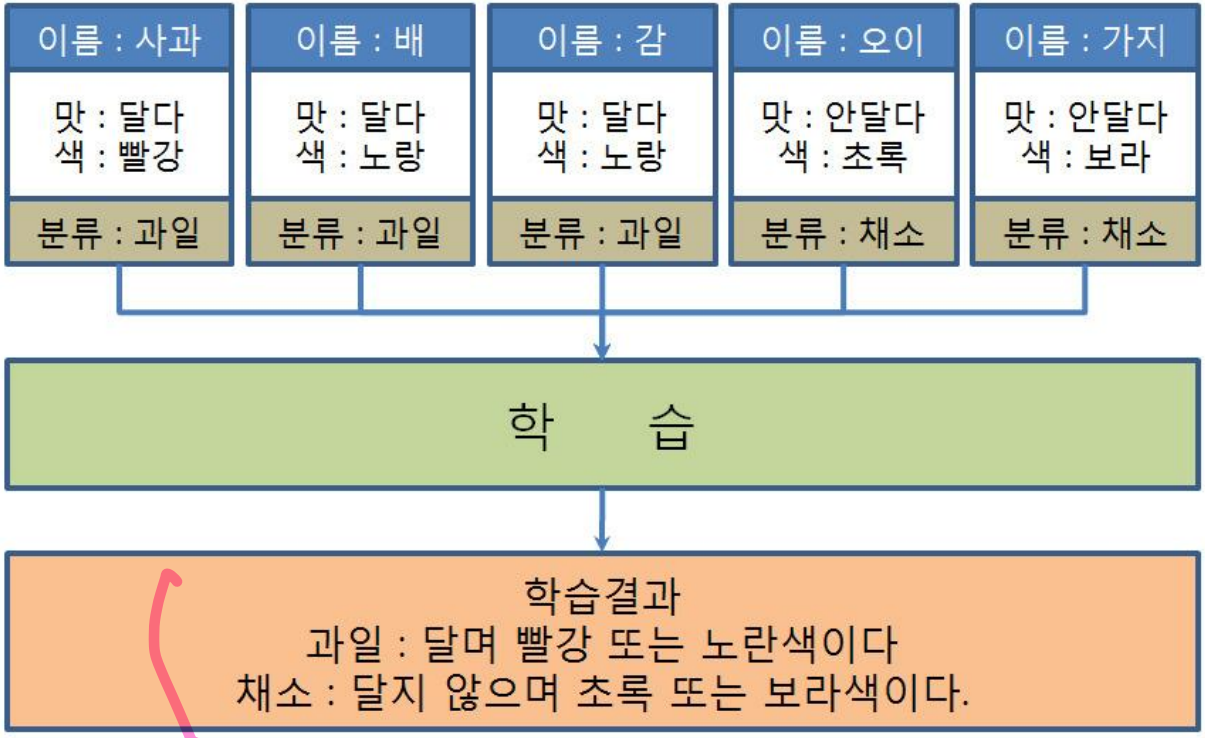
KNU

직관은 어떻게 기를 수 있을까?

경험과 학습

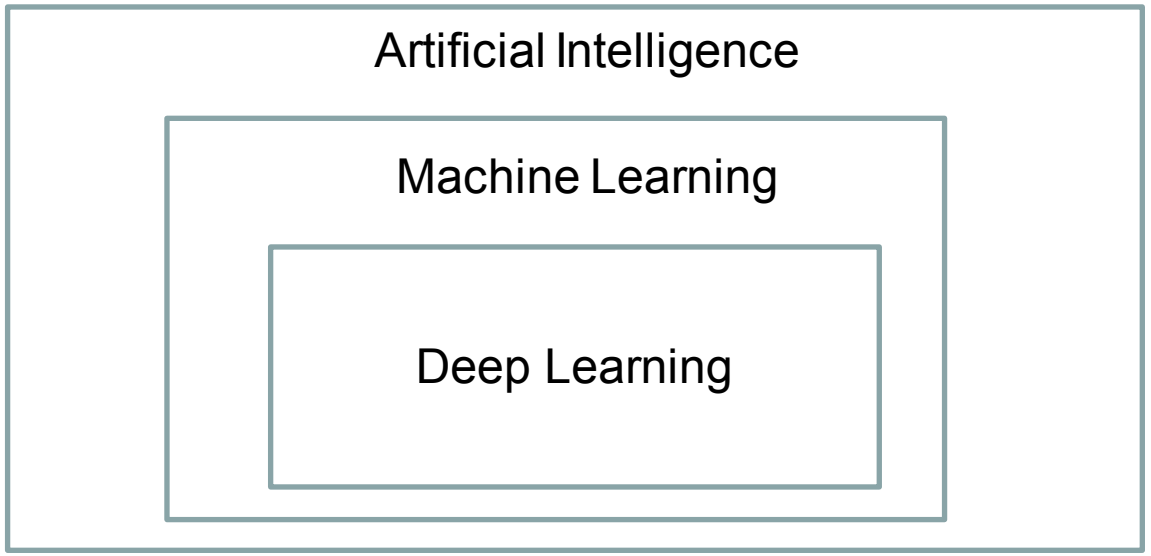
어떻게 컴퓨터가 학습능력을 기를 수 있을까?

컴퓨터의 학습


이름 : 사과
맛 : 달다
색 : 빨강
분류 : 과일
이름 : 배
맛 : 달다
색 : 노랑
분류 : 과일
이름 : 감
맛 : 달다
색 : 노랑
분류 : 과일
이름 : 오이
맛 : 안달다
색 : 초록
분류 : 채소
이름 : 가지
맛 : 안달다
색 : 보라
분류 : 채소
학 습
학습결과
과일 : 달며 빨강 또는 노란색이다
채소 : 달지 않으며 초록 또는 보라색이다.

기계학습

- **기계 학습** 또는 머신 러닝(machine learning)은 인공 지능의 한 분야로, 컴퓨터가 **학습**할 수 있도록 하는 알고리즘과 기술을 개발하는 분야



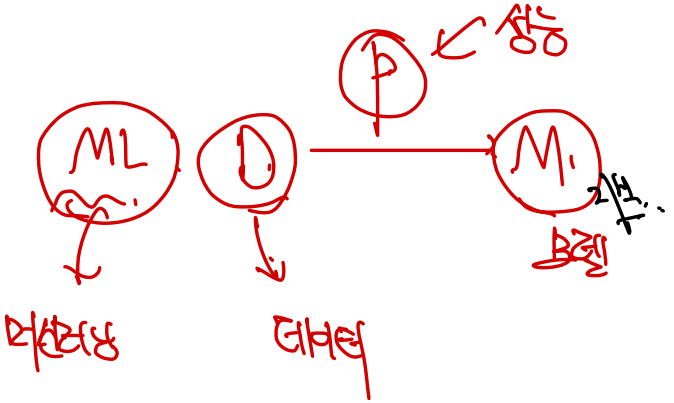
오늘날의 기계 학습

- 인공지능에게 고양이 사진을 주고 이 사진이 고양이라는 걸 알려준다.
- 사진 정보가 **신경망**에 들어가 분석된다.
- 최종 출력값이 고양이가 나올 때까지 **신경망**을 고친다.
- 다른 고양이 사진으로도 학습을 반복해 고양이 얼굴의 일반적인 패턴을 알아차린다.
- 학습이 끝나면 처음 본 고양이 사진도 고양이라고 분류할 수 있다.

KNU

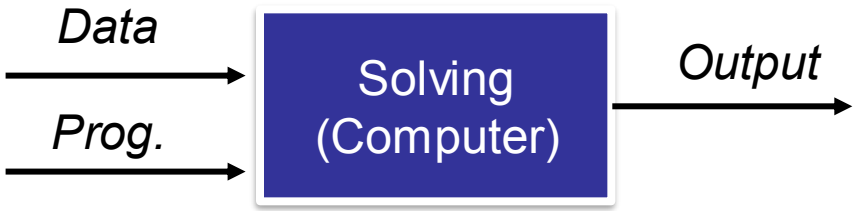
기계학습의 특징

- 기계학습의 정의
  - "환경(E)과의 상호작용을 통해서 축적되는 경험적인 데이터(D)를 바탕으로 지식, 모델(M)을 자동으로 구축하고 스스로 성능(P)을 향상하는 시스템 (Mitchell, 1997)
  - ML: D $\xrightarrow{P}$ M

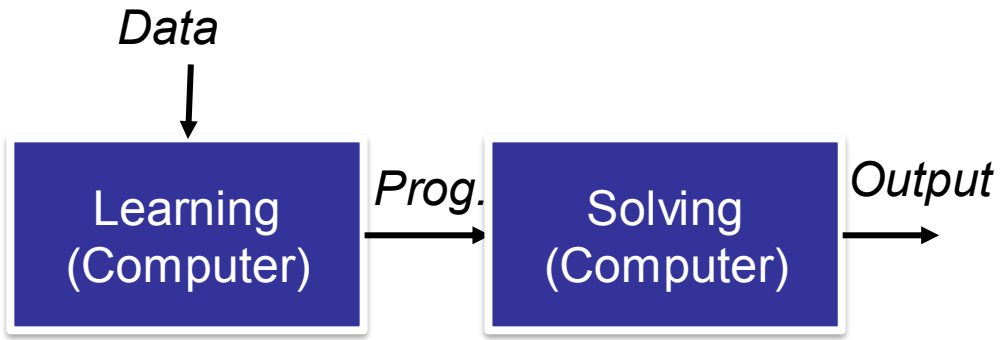


기계학습 vs. 프로그래밍

- 컴퓨터 프로그램
  - 사람이 알고리즘 설계 및 코딩
  - 주어진 문제(데이터)에 대한 답 출력



- 기계학습 프로그램
  - 사람이 코딩
  - 기계가 알고리즘을 자동으로 프로그래밍
  - 데이터에 대한 프로그램을 추력



기계학습 예

- 폐암 수술 환자의 생존율 예측하기

기존 환자 데이터 입력

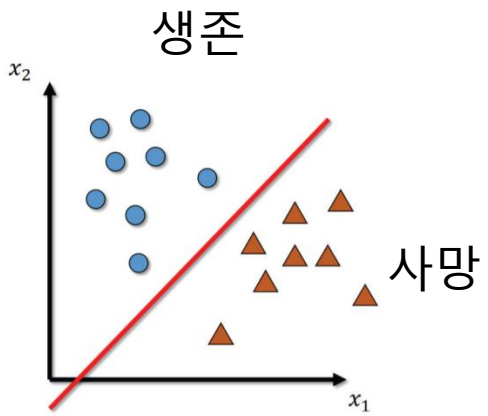
- 진료기록 1: 사망
- 진료기록 2: 생존
- 진료기록 3: 사망

.

.

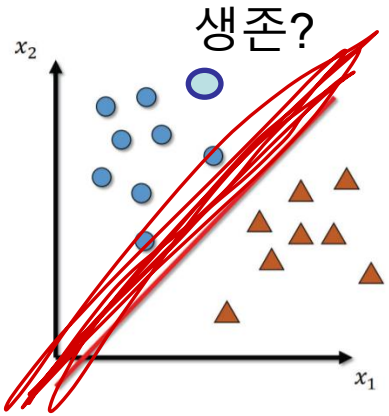
.

머신러닝으로 학습



규칙 발견

새로운 환자 예측



관찰하면 데이터를 학습

기계학습의 종류

- 지도 학습과 비지도 학습
  - 지도 학습: 훈련 데이터에 레이블(답)을 표기
    - 분류에 활용
  - 비지도 학습: 훈련 데이터에 레이블이 없이 학습
    - 클러스터링(군집), 시각화, 차원 축소, 연관 규칙 학습, 이상치 탐지
  - 준지도 학습: 일부만 레이블이 있음
    - 구글 포토 서비스
  - 강화 학습환경을 관찰해서 행동하고 결과로 보상 또는 벌점을 받으면서 최상의 전략을 학습

기계 학습의 주요 도전 과제

- 충분하지 않은 양의 훈련 데이터
- 대표성이 없는 훈련 데이터
- 낮은 품질의 데이터
- 관련 없는 특성
- 훈련 데이터 과대적합
- 훈련 데이터 과소적합

충분하지 않고 품질 데이터
낮은 품질의 데이터
대표성 ↓
관련 X 특성

→ 실제환경 정확도↓

학습데이터에 대해 과하게 학습하여
실제 데이터에 대한 오차가 증가한
현상이다.

→일반적인 데이터에 대한 정확도↓

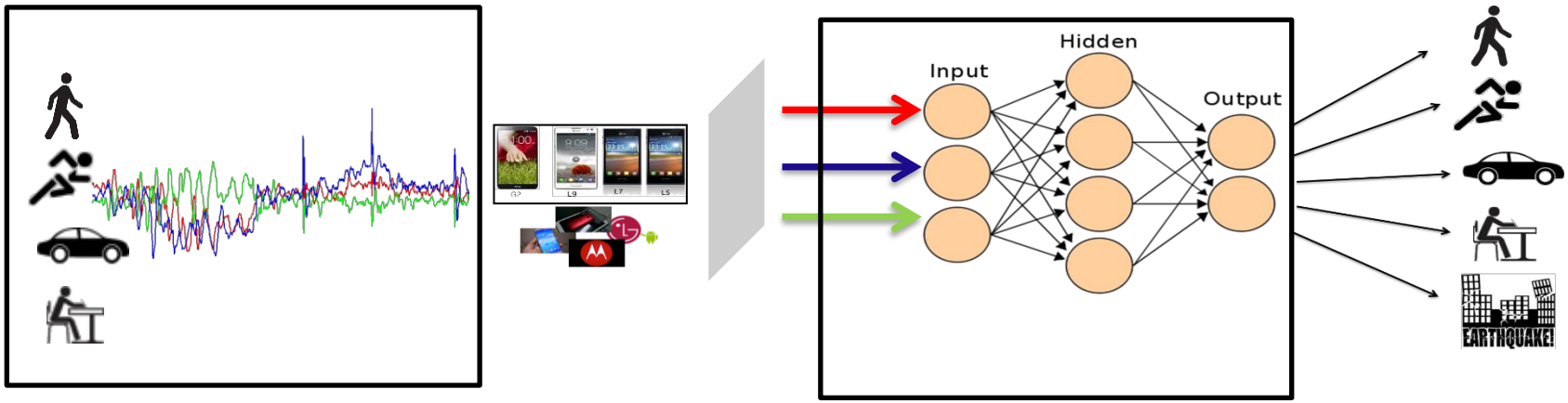
데이터량 대표성 품질 관련 X 특성

학습데이터 단순화

모델이 너무 단순해서
데이터의 내재된 구조를 학습하지 못할 때 발생
→파라미터가 더 많은 복잡한 모델을 선택해 줘야 한다

6만원짜리 한뼘 센서 달았더니…지진 관측소로 떠오른 SKT 기지국

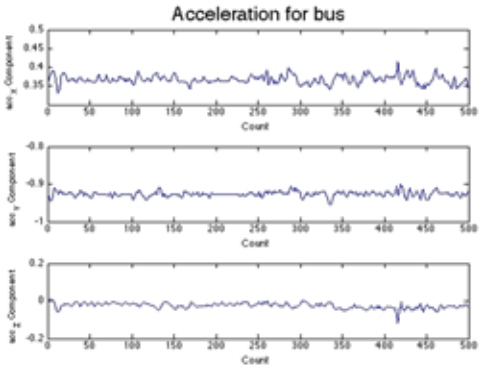
MyShake:
스마트폰기반의 지진 조기경보

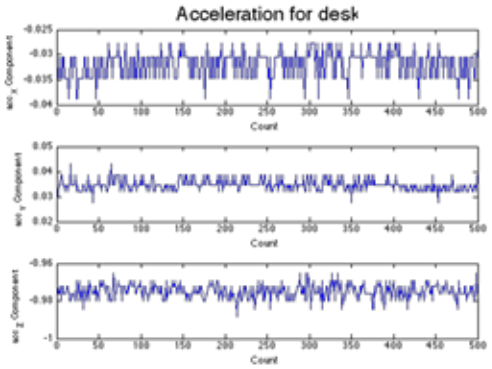


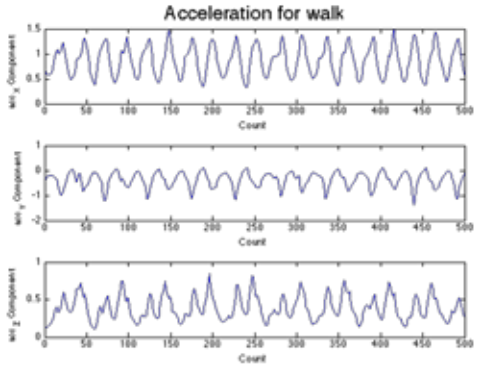
움직임 학습

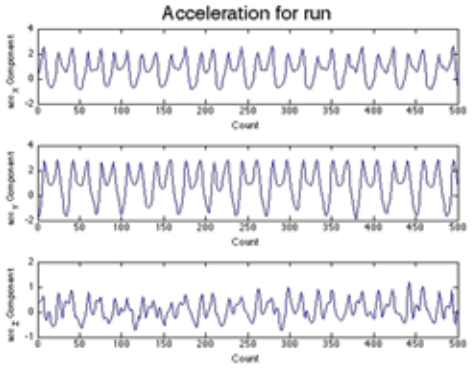
지진 감지(분류)

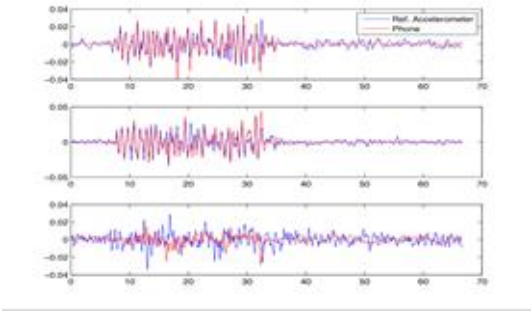
신경망을 통한 학습

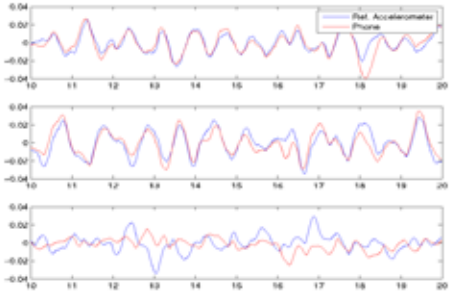

Acceleration for bus


Acceleration for desk


Acceleration for walk

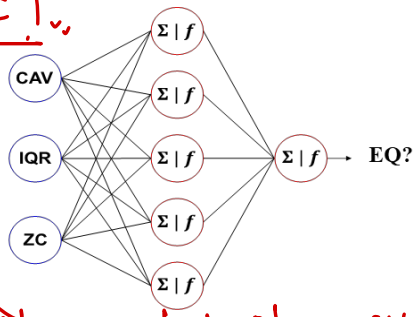

Acceleration for run

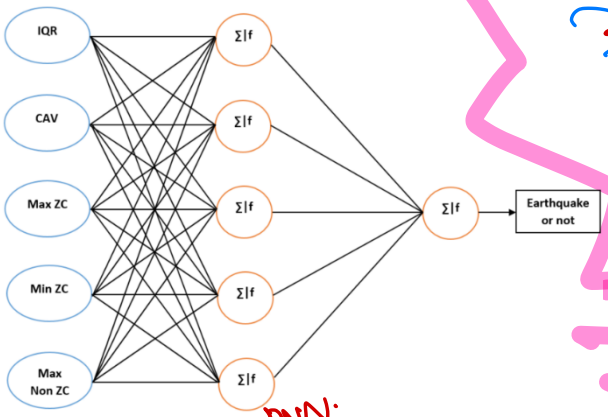




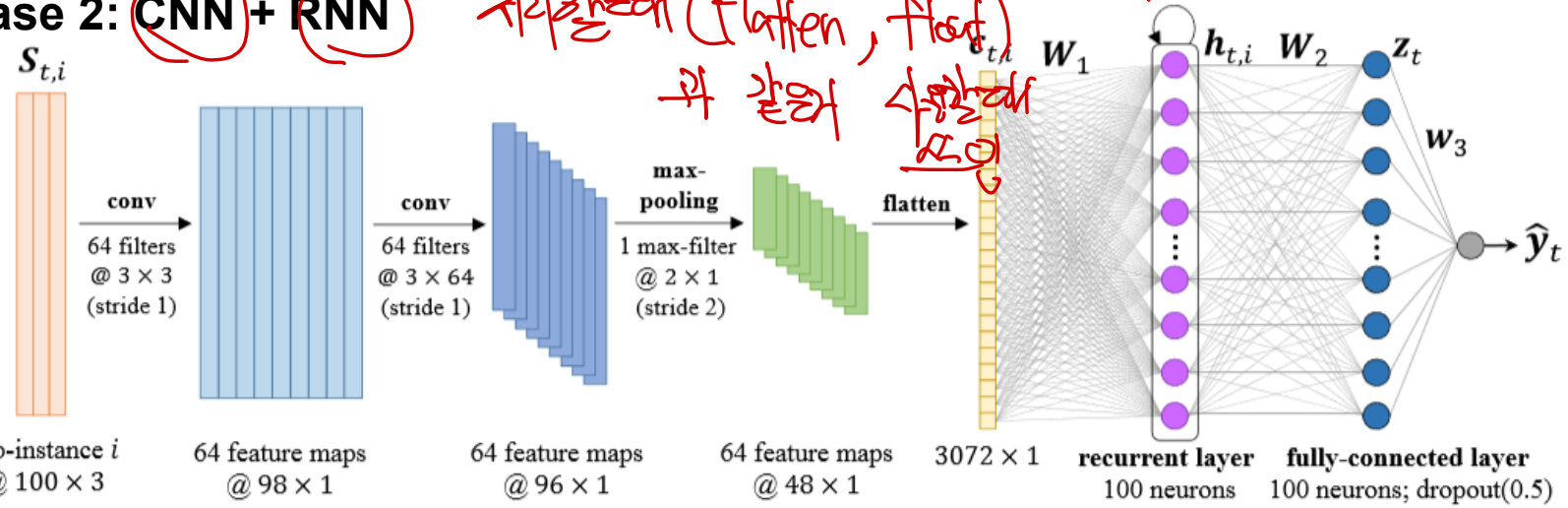
지진 감지 모델의 발전

Phase 1: Artificial Neural Network





Phase 2: CNN + RNN



도전 과제 1

1) 충분하지 않은 양의 훈련 데이터

- 지진이 많이 발생하는 지역은 미국, 일본, 대만이며 한반도는 지진이 많이 발생하지 않음

→ **한반도 지형을 반영한 지진 데이터 부족!**

- 스마트폰에서 지진을 감지하려면 스마트폰에서 감지된 지진이 많아야 함

→ **지진 전문 관측계에서 기록된 데이터가 대부분임!**

도전 과제 2

2) 대표성이 없는 훈련 데이터

- 지진으로 인한 피해는 규모 5.0 이상에서 발생하기 시작하지만 국내에서 규모 5.0 이상의 지진은 별로 없으며 규모 2.0 ~ 4.0 사이의 지진이 대부분임

→ **감지하고자 하는 목표와 훈련 데이터가 다름!**

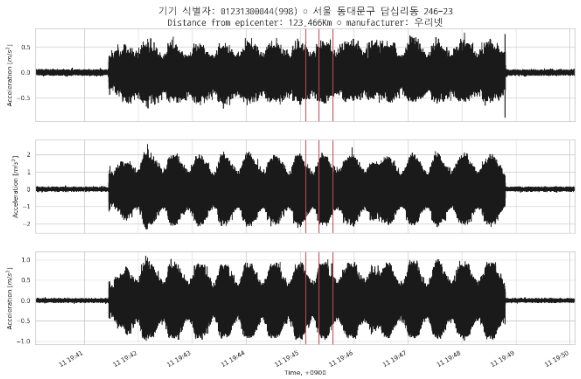
도전 과제 3

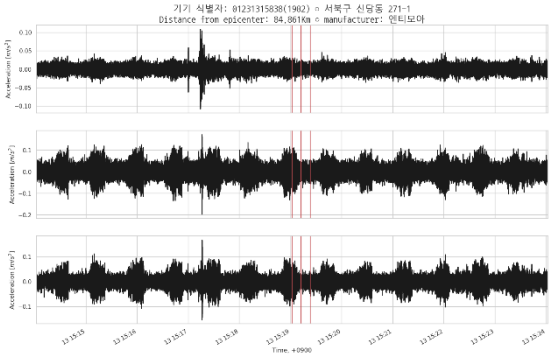
3) 낮은 품질의 데이터

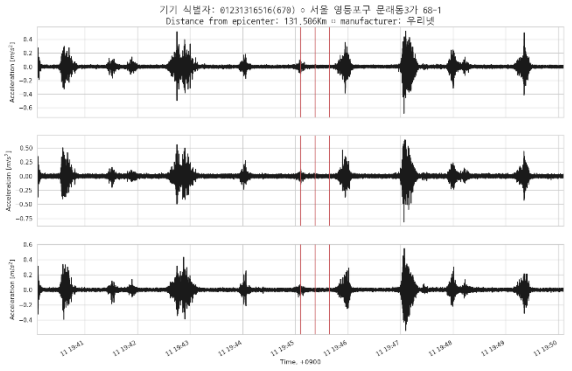
– 스마트폰은 수 많은 종류의 진동에 노출되어 있어 데이터의 품질이 일정하지 않음

→ **낮은 품질의 데이터로 학습**

데이터의 품질이 낮음







이상 데이터

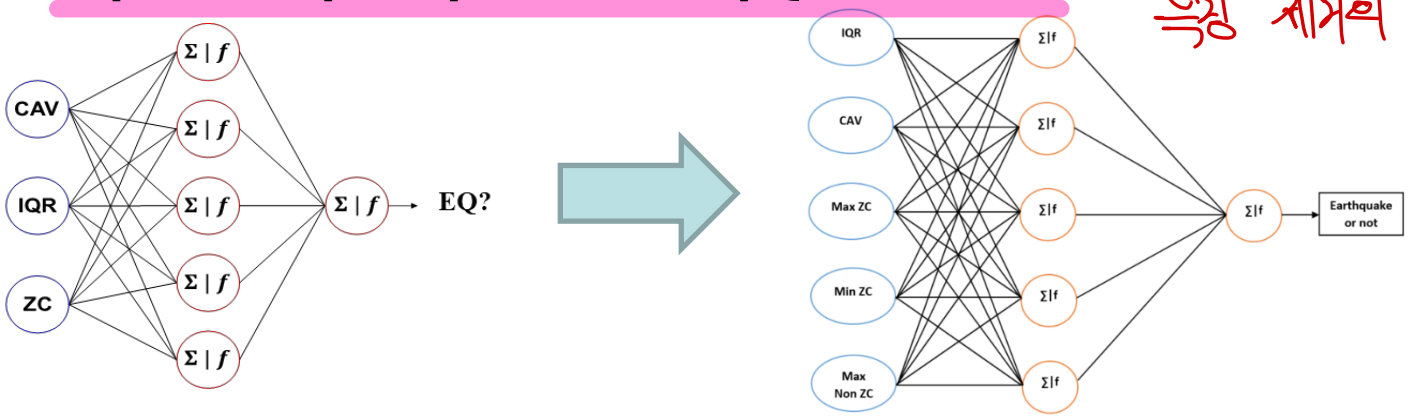
KNU

도전 과제 4

4) 관련이 없는 특성

- 초기의 ANN은 3개의 특성(feature)를 사용하였는데, 이는 사람 움직임과 지진을 구분하기 위함 이었음
- 하지만 사람 움직임이 없는 환경에서는 이러한 특성이 필요 없음

→ 사람 움직임과 관련된 특성 제거 필요



도전 과제 5

5) 훈련 데이터 과소적합

- 초기의 ANN (특성 3개 사용 모델) 모델은 정확도가 70% 수준으로 모델이 단순하여 훈련데이터에서도 낮은 정확도를 보였음

→딥러닝 기반의 모델 제시 (CNN + RNN)

도전 과제 6

6) 훈련 데이터 과대적합

– 딥러닝 기반의 모델 (정확도: 99.9%)를 만들었으나 실제 지진을 감지 못 하는 사례 발생

→ **모델이 전체 지진 데이터를 대표하지 못 함**

Table 1: Performance of the models on different window settings of the earthquake data. Where WS and FS means earthquake window size and feature window size respectively.

Settings WS-FS	Model	TP	TN	FP	FN	Accuracy	Precision	Recall
10-4	CRNN	26148	21819	7	15	99.95	99.97	99.94
8-6	CRNN-LSTM	11211	21790	4	3	99.98	99.96	99.97

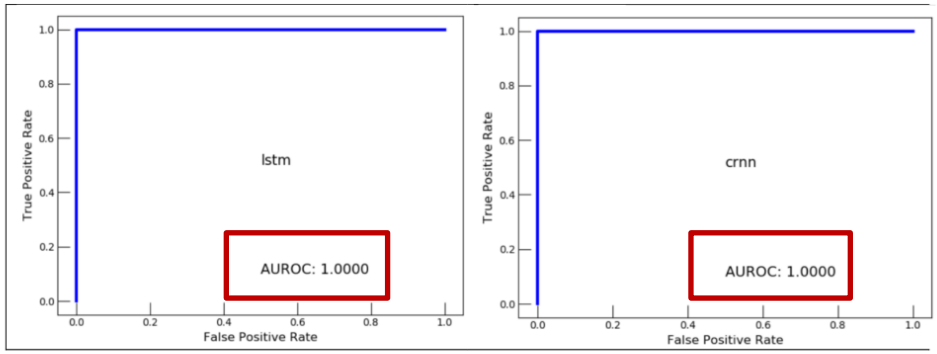
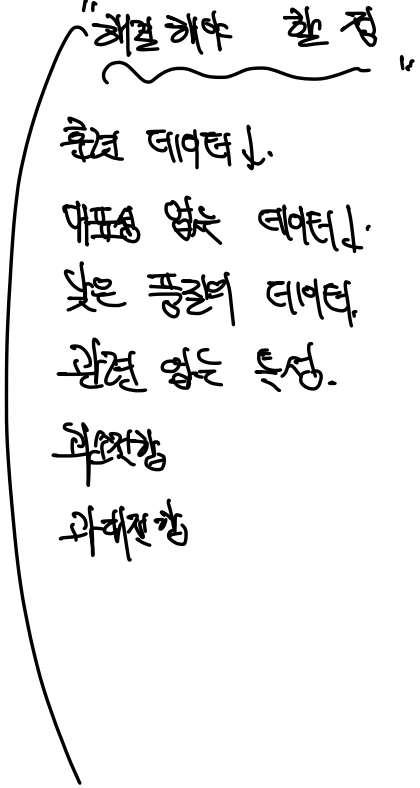


Figure 1: ROC curves of the models.



기계학습 경험 해보기

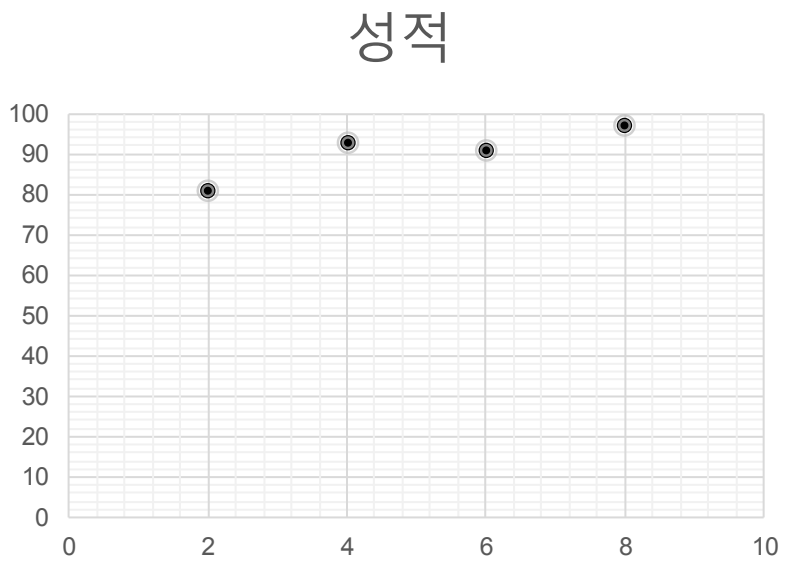
선형 회귀 (Linear Regression)

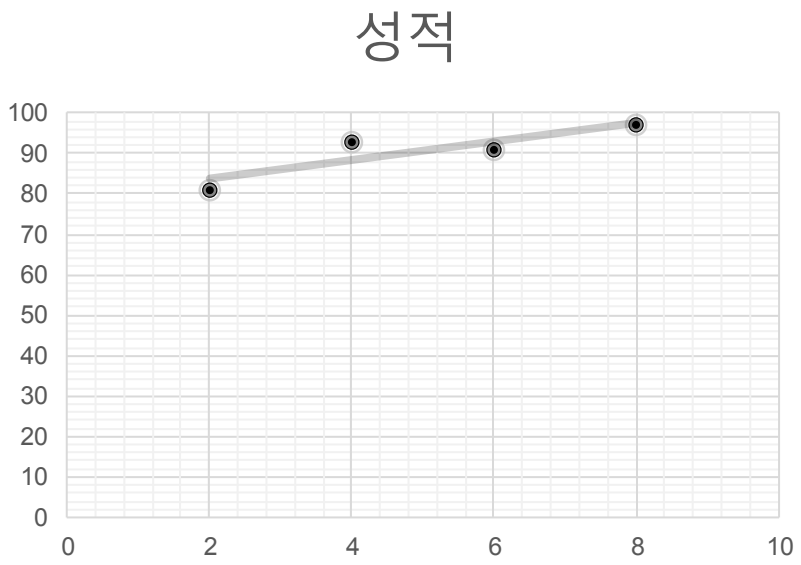
- “학생들의 중간고사 성적이 다 다르다“
- “학생들의 중간고사 성적이 **[정보]에 따라** 다 다르다“
- 정보: x(독립변수), 성적: y(종속변수)으로 두고 x와 y의 관계를 찾음

선형 회귀

선형회귀 ?

공부한 시간	2시간	4시간	6시간	8시간
성적	81점	93점	91점	97점





추세선 추가

$y = ax + b$

KNU

선형 회귀

- 추세선(예측선)을 구하는 방법
  - 최소 제곱근을 사용

$$a = \frac{\sum_{i=1}^{n}(x - mean(x)) \times (y - mean(y))}{\sum_{i=1}^{n}(x - mean(x))^2}$$

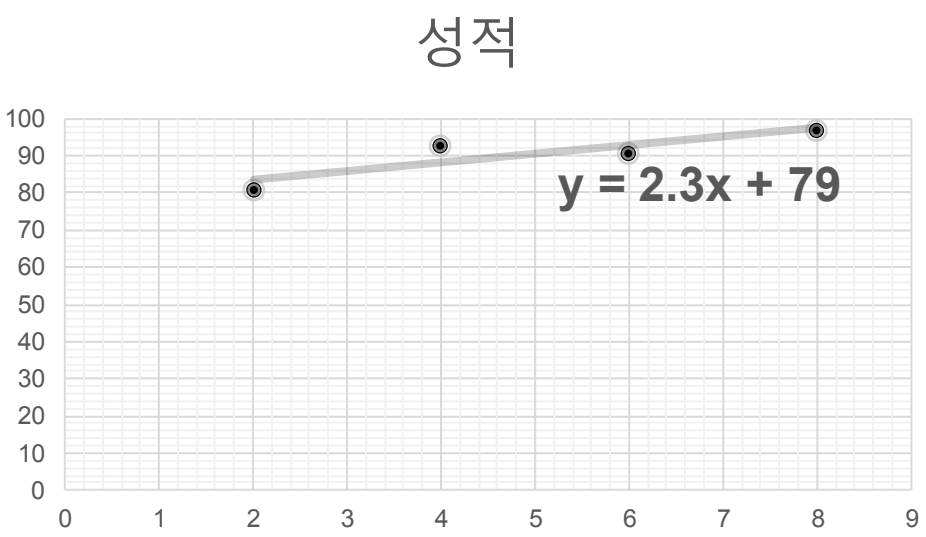
$$b = mean(y) - (mean(x) \times a)$$

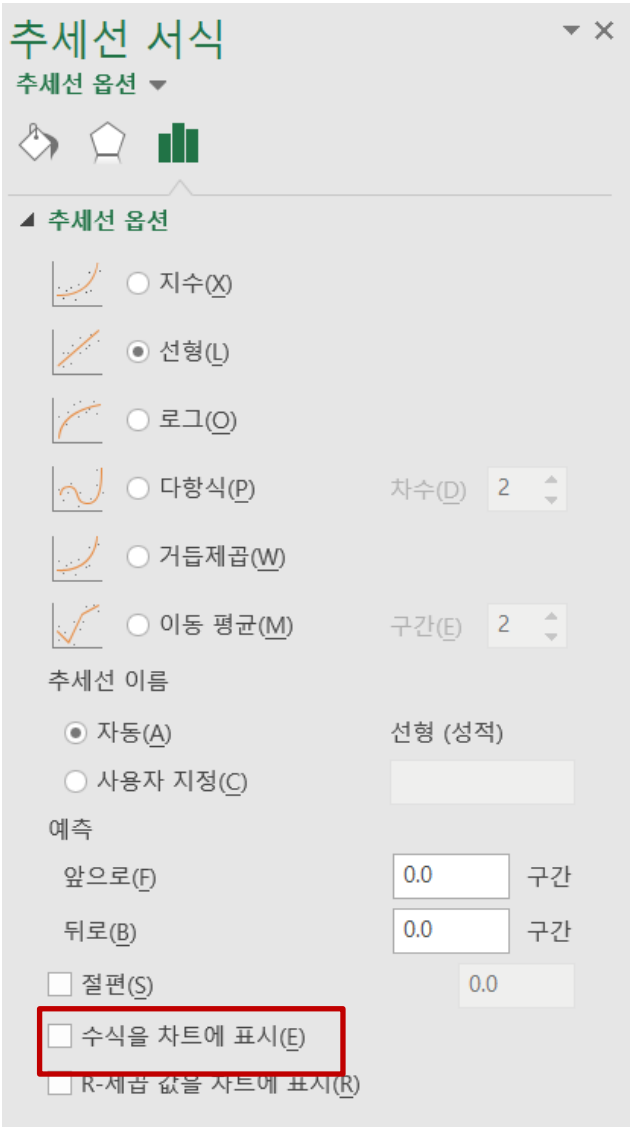
  - 위 공식을 사용하여 a, b를 구하고 예측값을 계산하시오.

공부한 시간	2시간	4시간	6시간	8시간
성적	81점	93점	91점	97점

KNU

엑셀을 활용하기


성적
y = 2.3x + 79
100
90
80
70
60
50
40
30
20
10
0
0
1
2
3
4
5
6
7
8
9


추세선 서식
추세선 옵션
추세선 옵션
지수(X)
선형(L)
로그(O)
다항식(P)
차수(D)
2
거듭제곱(W)
이동 평균(M)
구간(E)
2
추세선 이름
자동(A)
선형 (성적)
사용자 지정(C)
예측
앞으로(F)
0.0
구간
뒤로(B)
0.0
구간
절편(S)
0.0
수식을 차트에 표시(E)
R-제곱 값을 차트에 표시(R)

KNU

오차 계산하기

- 새로운 추세선인 y = 3x + 76을 사용해서 예측값을 계산하시오
- 앞에서 계산한 추세선과 지금의 추세선 중 어느 것이 정확한지 어떻게 비교를 할 수 있을까?
  - 평균 제곱근 오차를 사용하여 오차 계산

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - y_i)^2}$$

(실제값 – 예측값)의 제곱의 평균

다중 선형 회귀

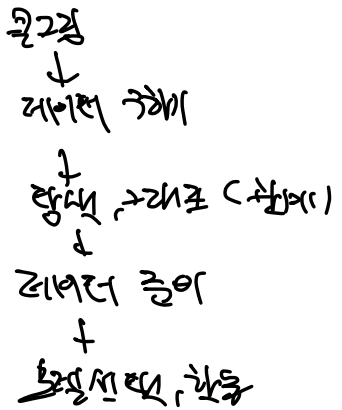
- 오차를 줄이기 위한 방법

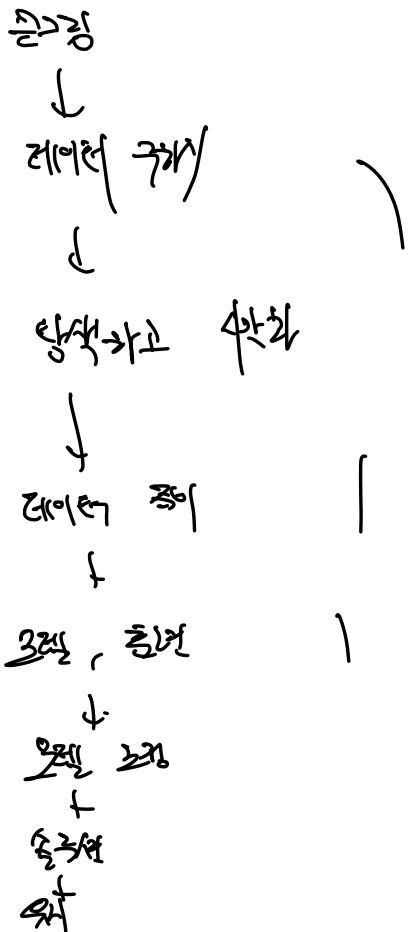
공부한 시간	2시간	4시간	6시간	8시간
성적	81점	93점	91점	97점

- 4시간 공부 → 93점
- 6시간 공부 → 91점 ???
- 제 3의 요소가 있음 (학원? 과외?)
- $y = a_1 x_1 + a_2 x_2 + b$

딥러닝과 데이터 다루기

1. 큰 그림 보기
2. 데이터 구하기
3. 탐색하고 시각화하기
4. 데이터 준비하기
5. 모델 선택하고 훈련시키기
6. 모델 조정
7. 솔루션 제시
8. 유지보수





큰 그림 보기

- 문제 파악하기
  - 캘리포니아 인구조사 데이터를 사용해 캘리포니아 주택 가격 모델 만들기
- 데이터 파악하기
  - 캘리포니아의 구역마다 인구, 중간 소득, 중간 주택 가격 정보를 담고 있음
- 목표 및 대상 설정하기
  - 타겟: 부동산, 은행, 투자자 등
  - 모델이 활용되는 분야 설정

데이터 가져오기

- 데이터 구조 훑어보기

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
0	-122.23	37.88	41.0	880.0	129.0	322.0	126.0	8.3252	452600.0	NEAR BAY
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0	1138.0	8.3014	358500.0	NEAR BAY

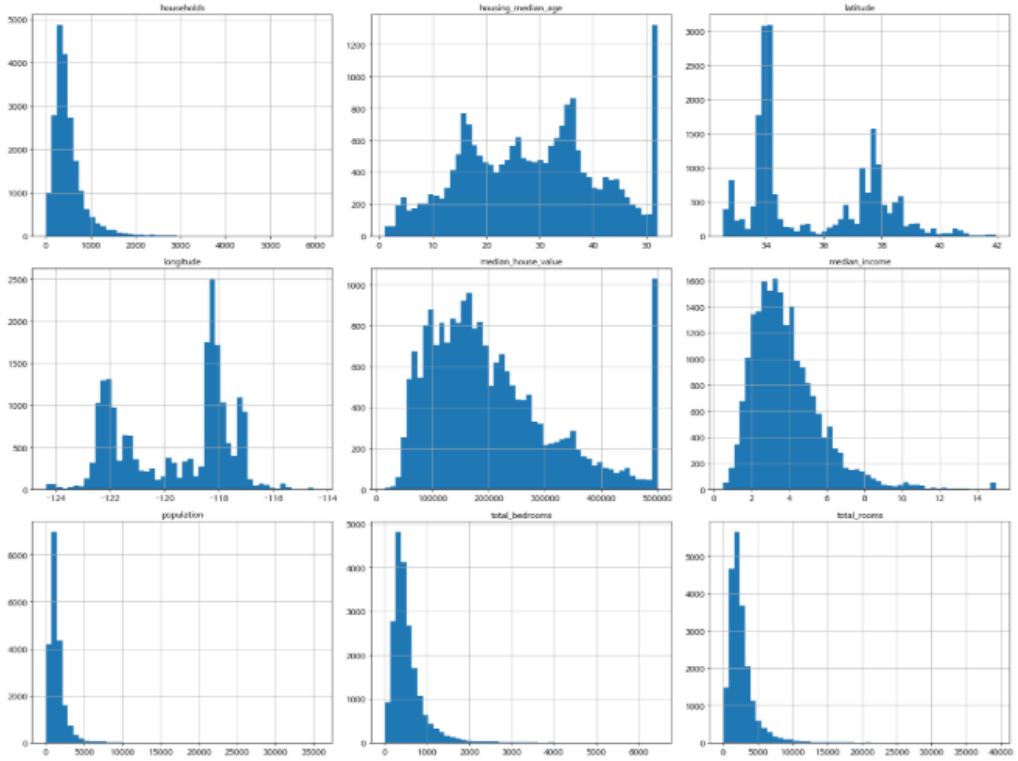
소득 정보

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
count	20640.000000	20640.000000	20640.000000	20640.000000	20433.000000	20640.000000	20640.000000	20640.000000	20640.000000
mean	-119.569704	35.631861	28.639486	2635.763081	537.870553	1425.476744	499.539680	3.870671	206855.816909
std	2.003532	2.135952	12.585558	2181.615252	421.385070	1132.462122	382.329753	1.899822	115395.615874
min	-124.350000	32.540000	1.000000	2.000000	1.000000	3.000000	1.000000	0.499900	14999.000000
25%	-121.800000	33.930000	18.000000	1447.750000	296.000000	787.000000	280.000000	2.563400	119600.000000
50%	-118.490000	34.260000	29.000000	2127.000000	435.000000	1166.000000	409.000000	3.534800	179700.000000
75%	-118.010000	37.710000	37.000000	3148.000000	647.000000	1725.000000	605.000000	4.743250	264725.000000
max	-114.310000	41.950000	52.000000	39320.000000	6445.000000	35682.000000	6082.000000	15.000100	500001.000000

집값 정보

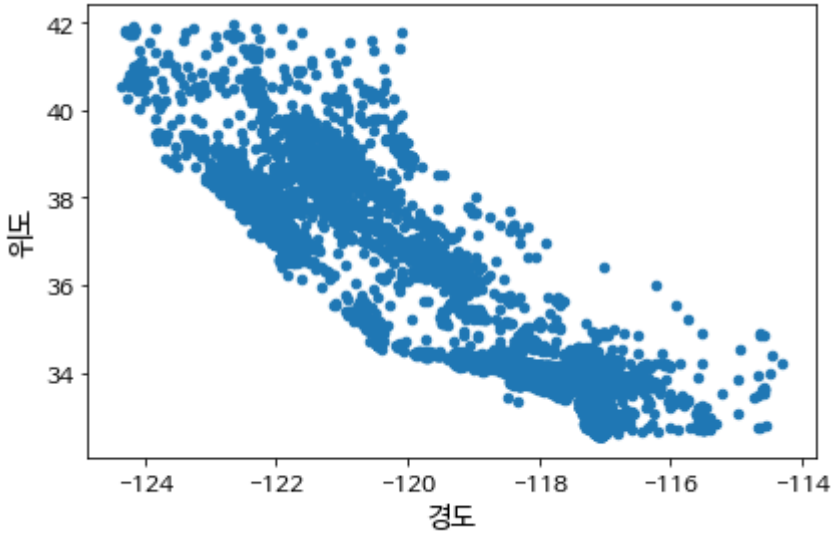
데이터 가져오기

- 히스토그램을 통한 데이터 검토하기

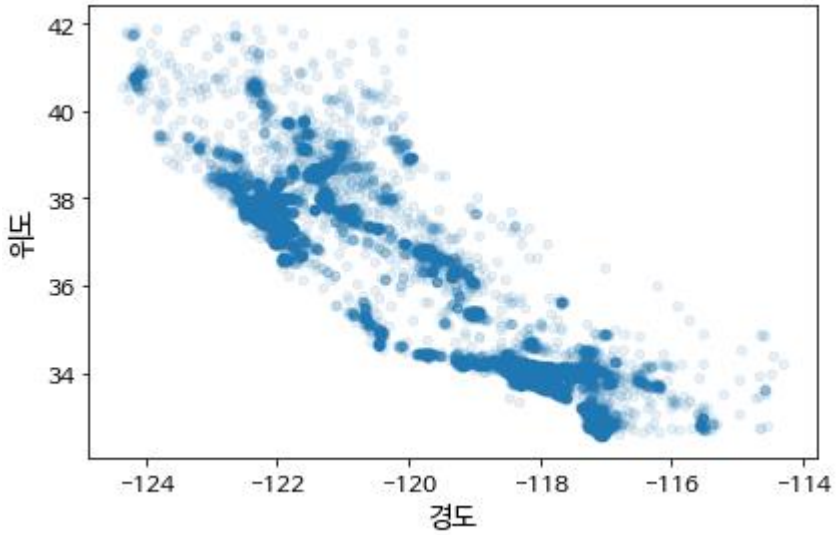


KNU

데이터 탐색과 시각화

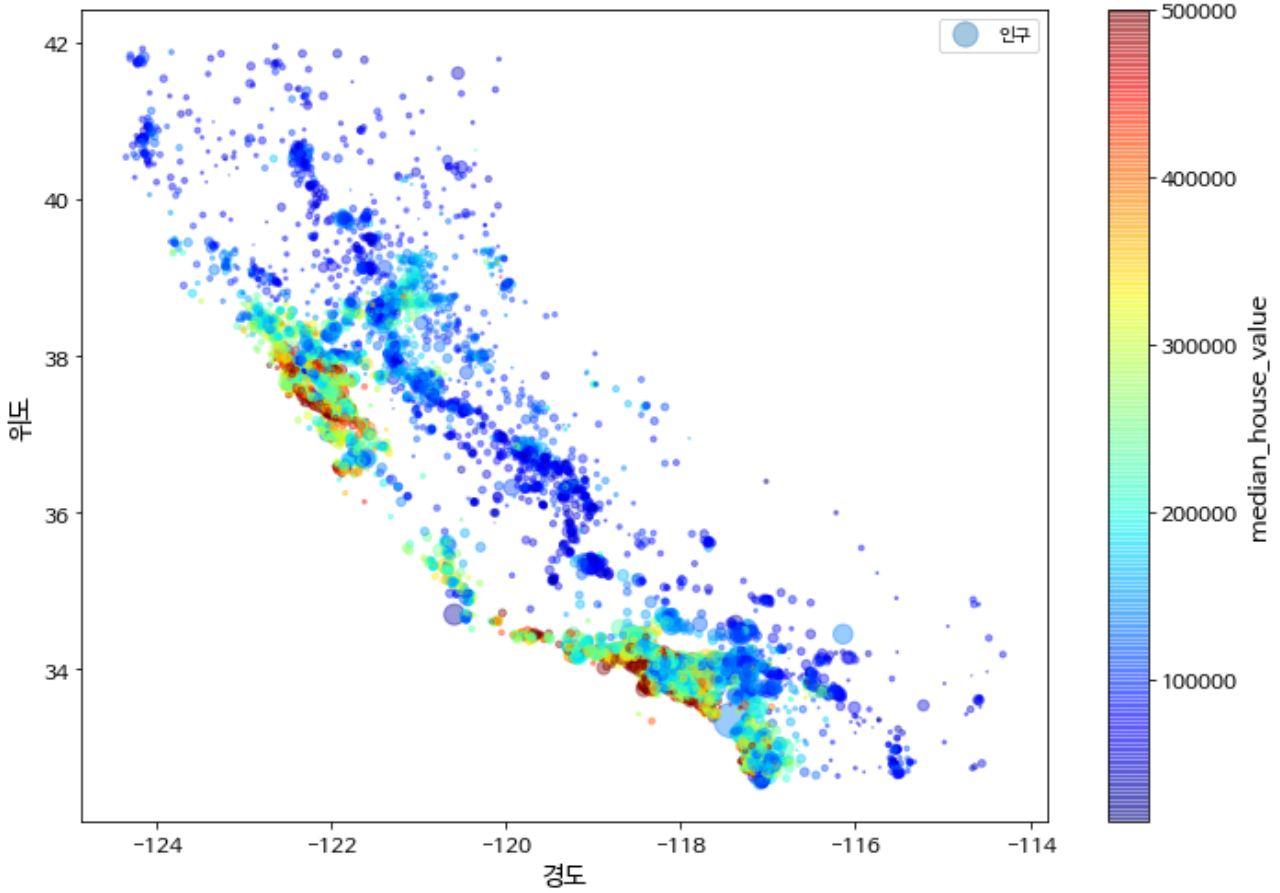


각 구역을 점으로 나타내기



투명도를 사용하여 밀집된 지역 부각시키기

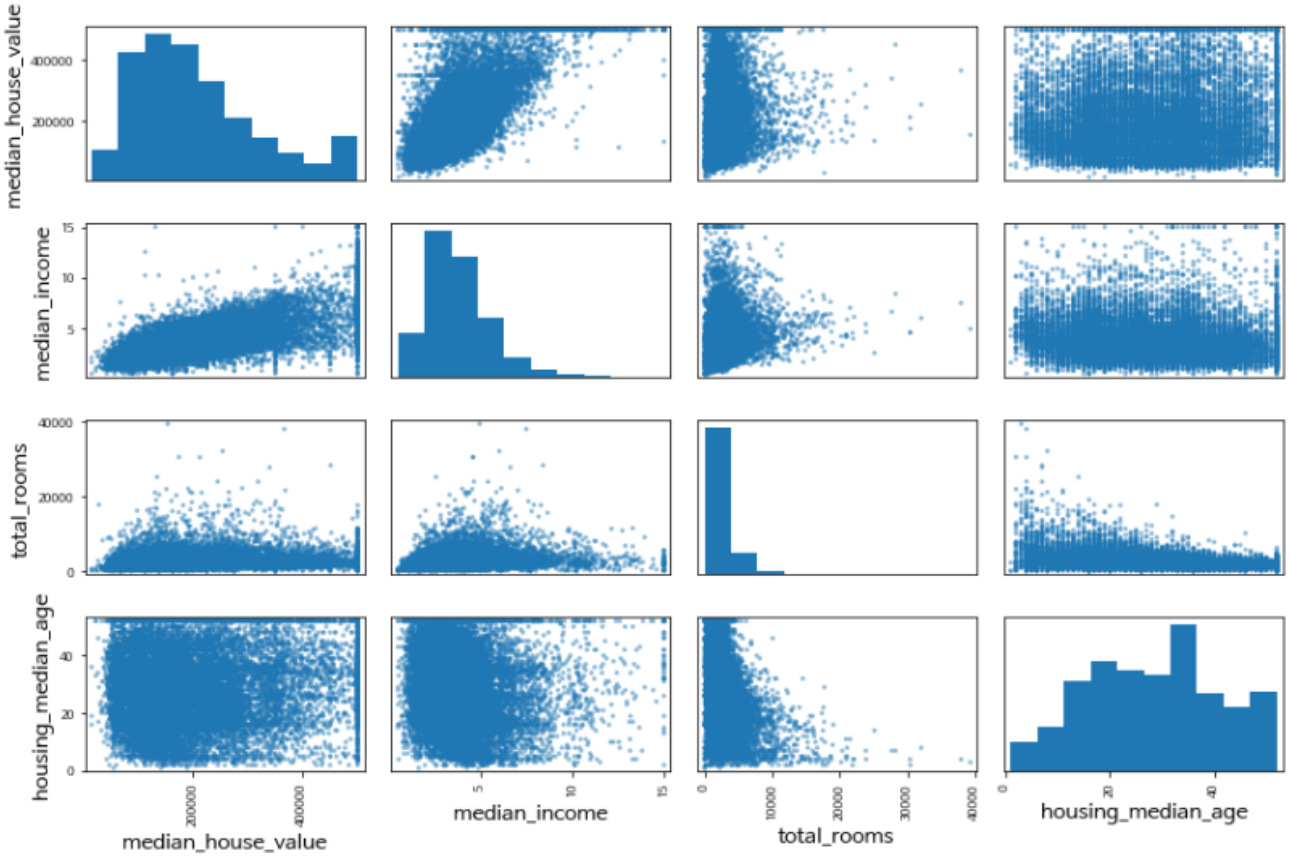
데이터 탐색과 시각화



구역별로 집값을 서로 다른 색으로 표현하기

데이터 탐색과 시각화

- 상관관계 찾기



데이터 준비하기

- 데이터 정제
  - 값이 없는 경우 삭제
  - 필요 없는 특성 값 삭제
  - 누락된 값 대체
  - 데이터 변환하기 (텍스트 -> 숫자)
- 스케일링 하기
  - Min-max 스케일링 (정규화)
  - 표준화

나머지…

- 모델 선택과 훈련
  - 80% 훈련, 20% 테스트 및 검증
- 모델 조정
- 솔루션 제시
- 유지보수

실습 및 과제

- 앞에서 배운 **"딥러닝 & 데이터 다루기"** 를 활용하여 우리 주변에서 접할 수 있는 문제 중에서 딥러닝을 이용하여 해결할 수 있는 문제를 찾으시오 *(문제 정의 단계)*.
- 해당 문제를 해결하기 위해서 필요한 데이터 셋을 정의하고 해당 데이터 셋이 문제 해결에 어떻게 사용될지 설명하시오 *(데이터 파악하기 단계)*
- 실제 데이터를 찾아보고 찾은 데이터 셋에 원하는 정보가 포함되어 있는지 조사하시오 *(데이터 준비 및 탐색하기)*
  - 예) 공공 데이터
- 데이터를 찾을 수 없다면 필요로 하는 데이터를 구체화하고 데이터 생산 및 배포의 주체가 누가 되어야 하며 어떠한 노력을 기울여야 하는지 기술하고, 데이터 공유로 인한 득과 실에 대해서 논하시오

KNU