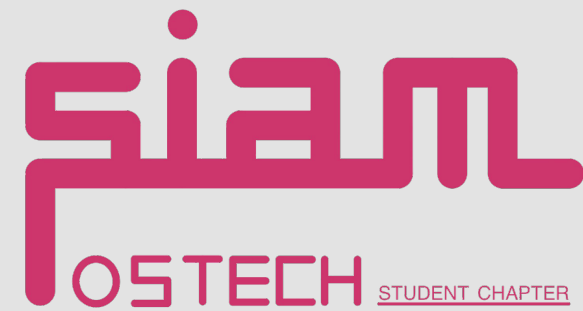


# Chapter 1.

# 한눈에 보는 머신러닝

[왕초보 머신러닝 파티 시즌 2]

POSTECH GSAI 박사과정 이성헌



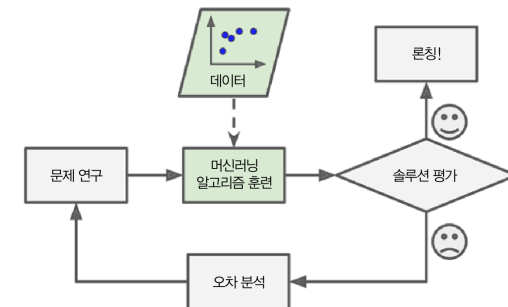
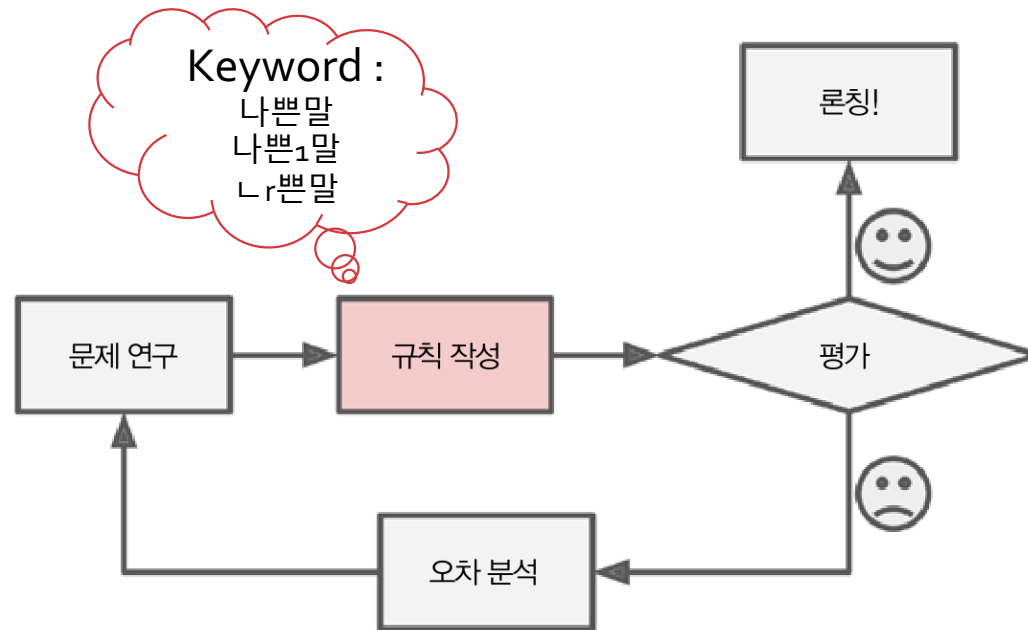
## 1.1 머신러닝이란?

- 머신러닝은 **명시적인 프로그래밍 없이** 컴퓨터가 학습하는 능력을 갖추게 하는 연구 분야다. \_아서 새뮤얼, 1959
- 머신러닝은 문제 푸는 방법을 일일이 가르쳐주는 것이 아니라 **문제집만 던져주면 컴퓨터가 독학하는 것**이다. \_이성헌, 2022



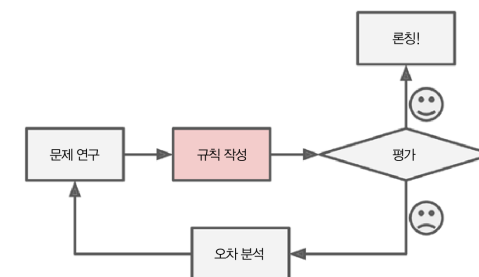
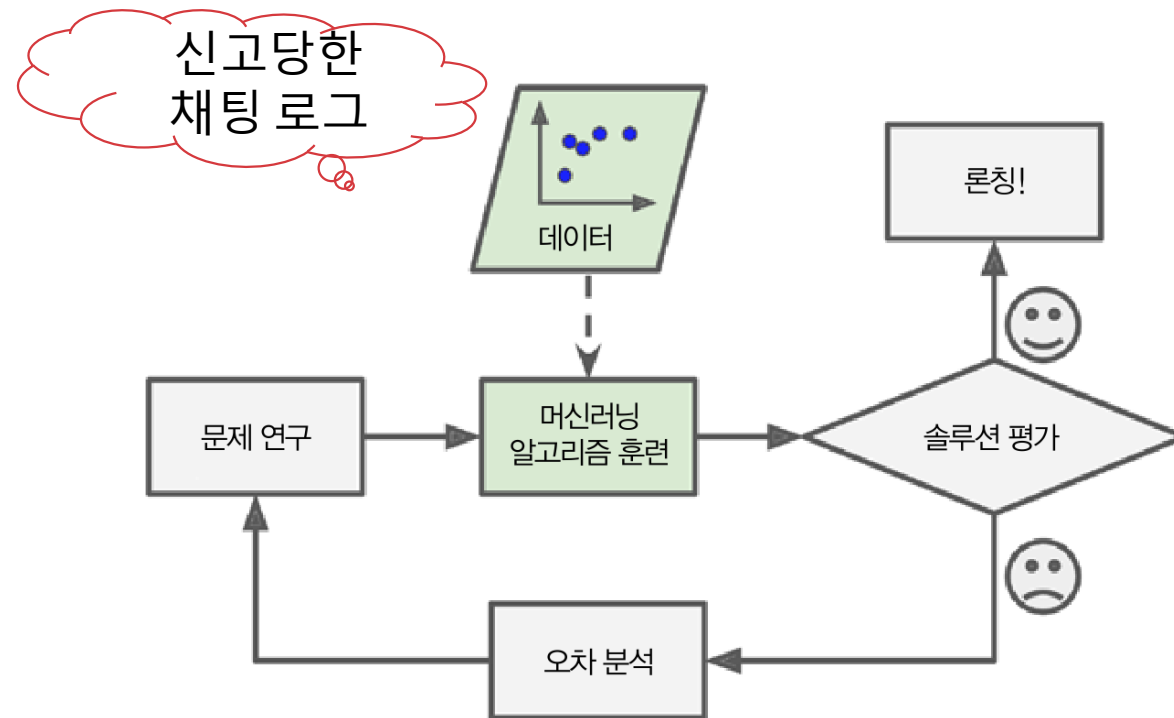
## 1.2 왜 머신러닝을 사용하는가?

- 전통적인 접근 방법 : Rule Based Model



## 1.2 왜 머신러닝을 사용하는가?

- 머신러닝 접근 방법 : Data Based Model



## 1.2 왜 머신러닝을 사용하는가?

- ***Rule Based Model***

- 데이터의 패턴을 직접 설계 해야함
- 새로운 패턴을 찾으려면 새 알고리즘을 만들어야 함
- 데이터가 많으면 계산 복잡도가 커질 수 있음
- 문제에 대한 전문적인 지식을 필요로 함
- 모델의 결과에 대해 이론적으로 설명하기 용이함
- 데이터가 적어도 잘 작동하는 모델을 설계할 수 있음

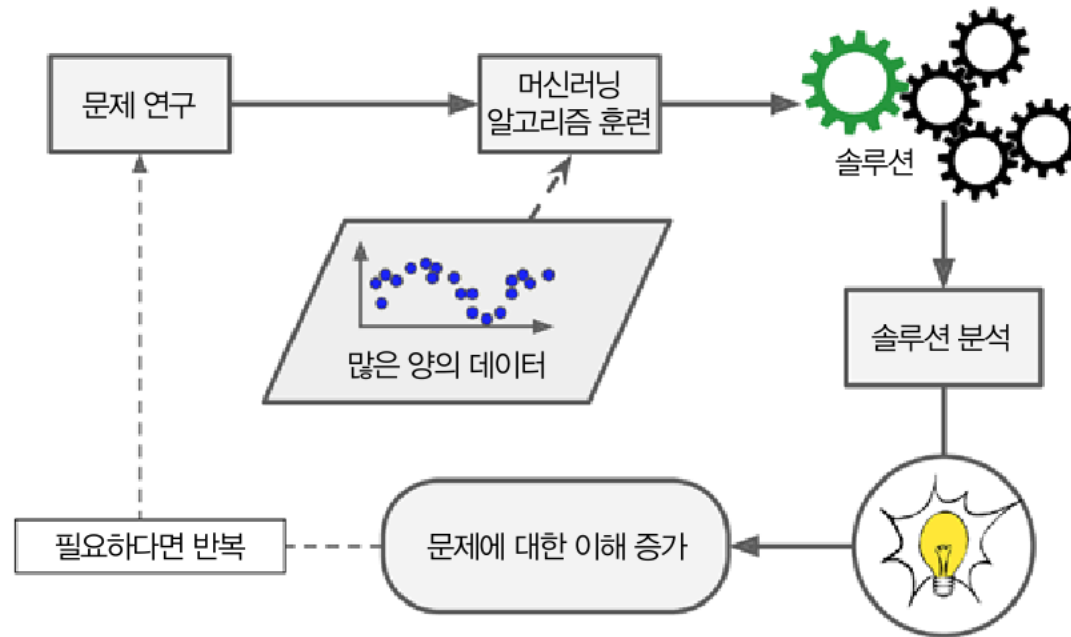
- ***Data Based Model***

- 데이터의 패턴을 스스로 학습함
- 새로운 패턴도 스스로 찾음
- 많은 양의 데이터를 처리하는 것에 특화됨
- 상대적으로 문제에 대한 지식을 덜 요구함
- 모델의 결과에 대해 이론적으로 설명하기 어려움
- 데이터가 적으면 모델의 신뢰도를 얻기 어려움

## 1.2 왜 머신러닝을 사용하는가?

- **Data Mining**

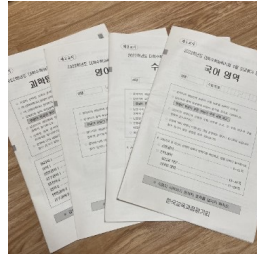
- 데이터로부터 유용한 패턴을 자동적으로 추출하는 기법
- 머신러닝의 솔루션을 분석해 어려운 문제를 더 잘 이해할 수 있음
  - Ex) Random Forest Model의 Feature Importance



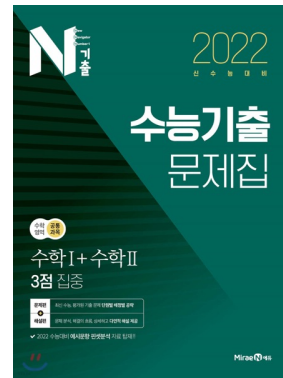
## 1.4 머신러닝 시스템의 종류

- **1.4.1 지도 학습과 비지도 학습**

- 지도학습 : 답지가 있는 문제집으로 공부하는 모델.



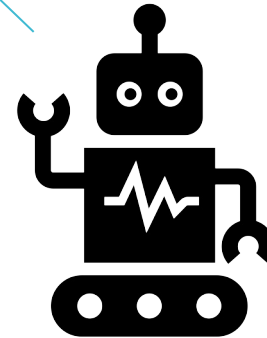
테스트 데이터



훈련 데이터

성능 평가

학습

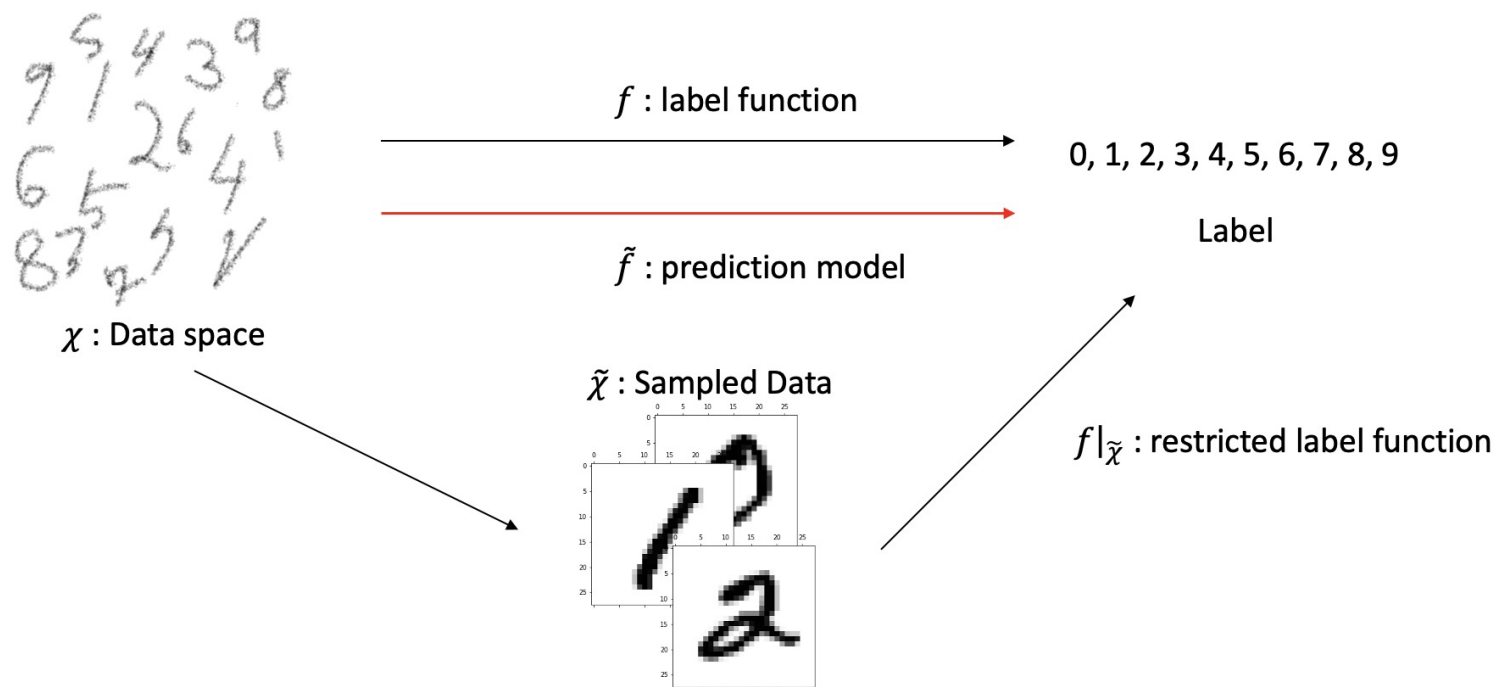


2023년 수능

## 1.4 머신러닝 시스템의 종류

### • 1.4.1 지도 학습과 비지도 학습

- 지도학습 : 답지가 있는 문제집으로 공부하는 모델.
- Prediction function  $\tilde{f}$ 을 최대한  $f$ 에 가깝게 근사시키는 것이 목적.





## 1.4 머신러닝 시스템의 종류

### • 1.4.1 지도 학습과 비지도 학습

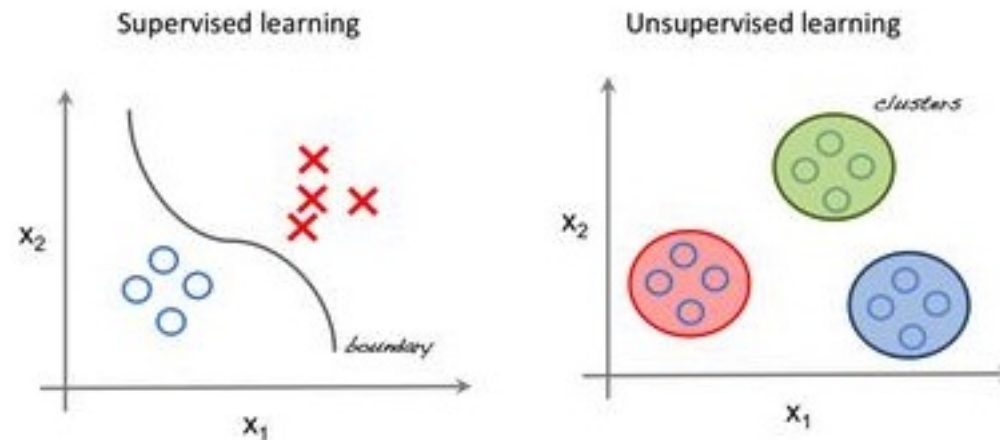
- 비지도학습 : 답지가 없는 문제를 푸는 모델
  - 군집 : 유사한 패턴의 데이터끼리 묶는 문제

<b>ISTJ</b> 세상의 소금형 한번 시작한 일은 끝까지 해내는 사람들	<b>ISFJ</b> 임금 뒤편의 권력형 성실하고 온화하며 협조를 잘하는 사람들	<b>INFJ</b> 예언자형 사람과 관련된 뛰어난 통찰력을 가지고 있는 사람들	<b>INTJ</b> 과학자형 전체적인 부분을 조합하여 비전을 제시하는 사람들
<b>ISTP</b> 백과사전형 논리적이고 뛰어난 상황 적응력을 가지고 있는 사람들	<b>ISFP</b> 성인군자형 따뜻한 감성을 가지고 있는 겸손한 사람들	<b>INFP</b> 잔다르크형 이상적인 세상을 만들어 가는 사람들	<b>INTP</b> 아이디어 뱅크형 비평적인 관점을 가지고 있는 뛰어난 전략가들
<b>ESTP</b> 수완좋은 활동가형 친구, 운동, 음식 등 다양한 활동을 선호하는 사람들	<b>ESFP</b> 사교적인 유형 분위기를 고조시키는 우호적 사람들	<b>ENFP</b> 스파크형 열정적으로 새로운 관계를 만드는 사람들	<b>ENTP</b> 발명가형 풍부한 상상력을 가지고 새로운 것에 도전하는 사람들
<b>ESTJ</b> 사업가형 사무적, 실용적, 현실적으로 일을 많이하는 사람들	<b>ESFJ</b> 친선도모형 친절과 현실감을 바탕으로 타인에게 봉사하는 사람들	<b>ENFJ</b> 언변능숙형 타인의 성장을 도모하고 협동하는 사람들	<b>ENTJ</b> 지도자형 비전을 가지고 사람들을 활력적으로 이끌어가는 사람들

## 1.4 머신러닝 시스템의 종류

- **1.4.1 지도 학습과 비지도 학습**

- 비지도학습 : 답지가 없는 문제를 푸는 모델
  - 군집 : 유사한 패턴의 데이터끼리 묶는 문제
  - 주어진 데이터 셋  $\chi$  에 대하여,  $\chi$ 가 따르는 분포를 추정하는 문제
- K-means clustering, DBSCAN, t-SNE ...



## 1.4 머신러닝 시스템의 종류

- **1.4.1 지도 학습과 비지도 학습**

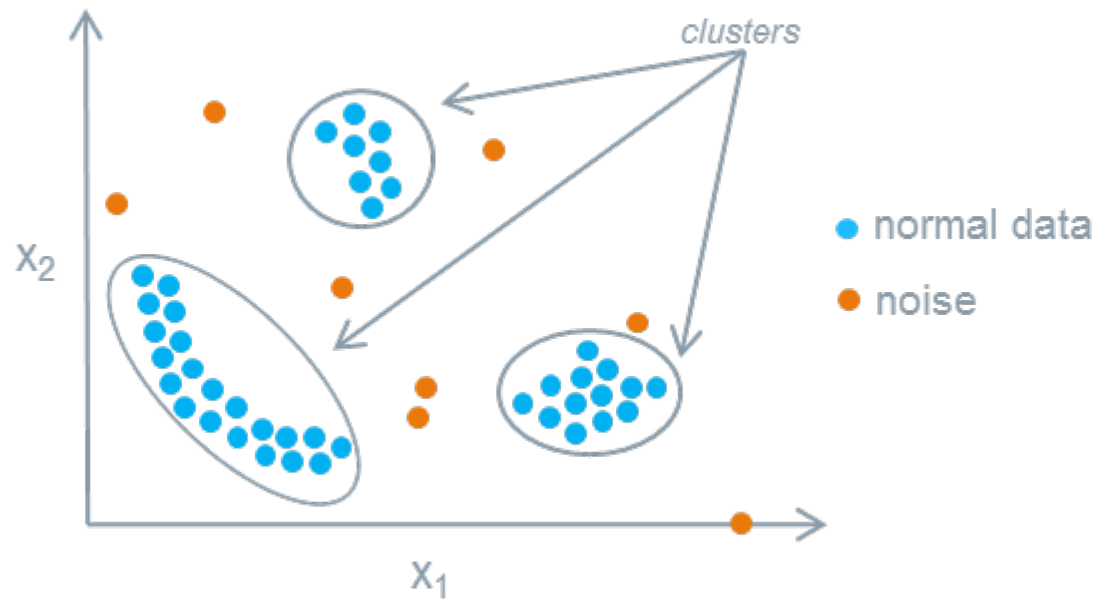
- 비지도학습 : 답지가 없는 문제를 푸는 모델
  - 이상(특이)치탐지 : 정상 데이터에 섞인 소량의 이상(특이)치를 감지



## 1.4 머신러닝 시스템의 종류

- **1.4.1 지도 학습과 비지도 학습**

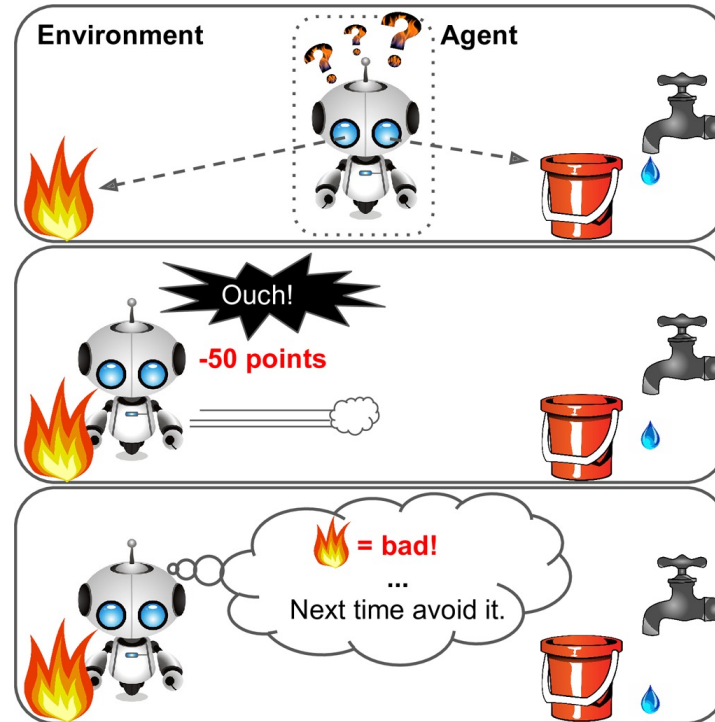
- 비지도학습 : 답지가 없는 문제를 푸는 모델
  - 이상(특이)치탐지 : 정상 데이터에 섞인 소량의 이상(특이)치를 감지



## 1.4 머신러닝 시스템의 종류

- 강화학습 (Reinforcement Learning)

- 에이전트와 환경이 상호작용하며 최적의 행동을 찾아나가는 모델



1 Observe

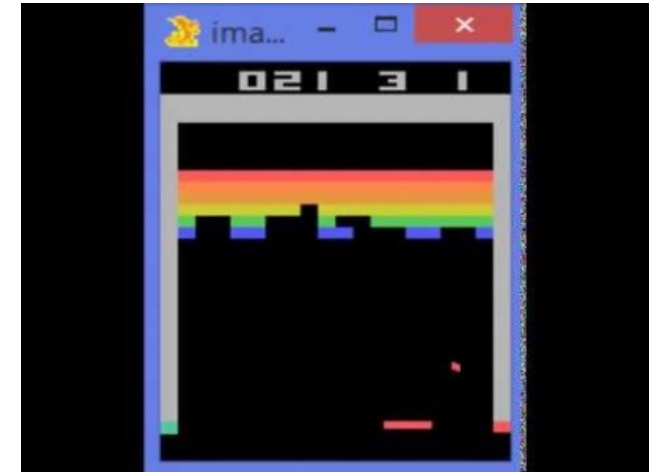
2 Select action using policy

3 Action!

4 Get reward or penalty

5 Update policy (learning step)

6 Iterate until an optimal policy is found

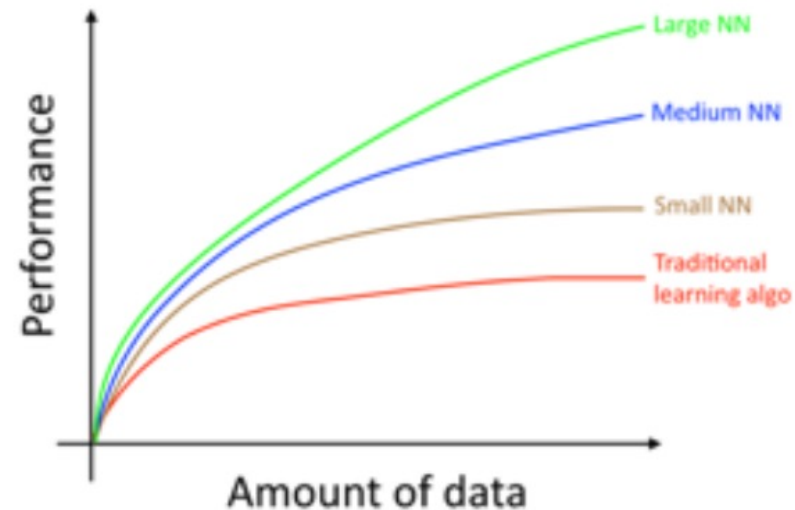


## 1.5 머신러닝의 주요 도전 과제

### • 1.5.1 충분하지 않은 양의 훈련 데이터

- 머신러닝은 데이터가 많을 수록 더 많은 패턴을 더 정확히 학습합니다.
- 문제를 많이 풀어봐야 시험을 잘 칠 수 있는 것과 마찬가지로.
- 그러나, 실전에서 좋은 데이터를 충분히 많이 얻기는 어렵습니다.

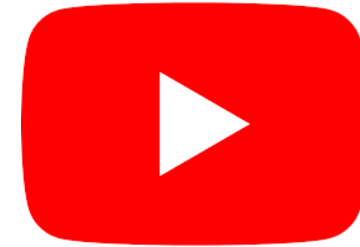
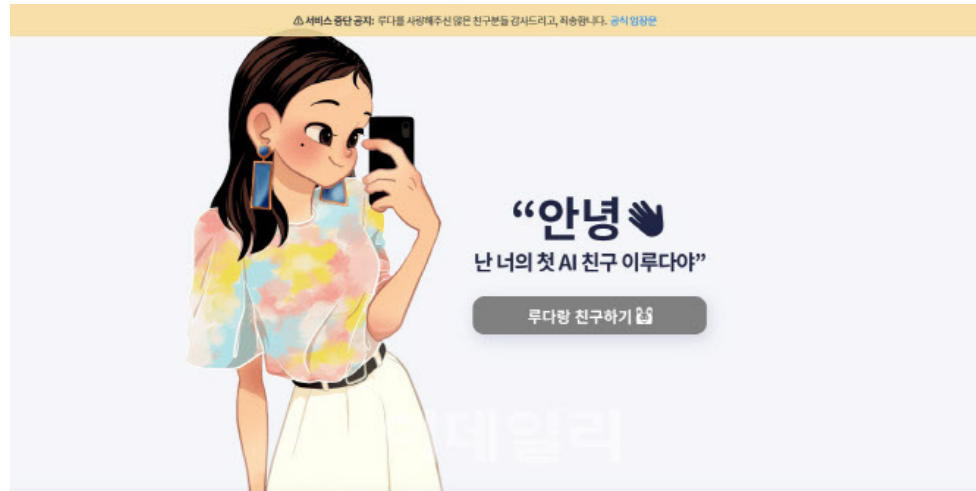
Trend #1: Scale driving Deep Learning progress



## 1.5 머신러닝의 주요 도전 과제

### • 1.5.2 대표성 없는 훈련 데이터

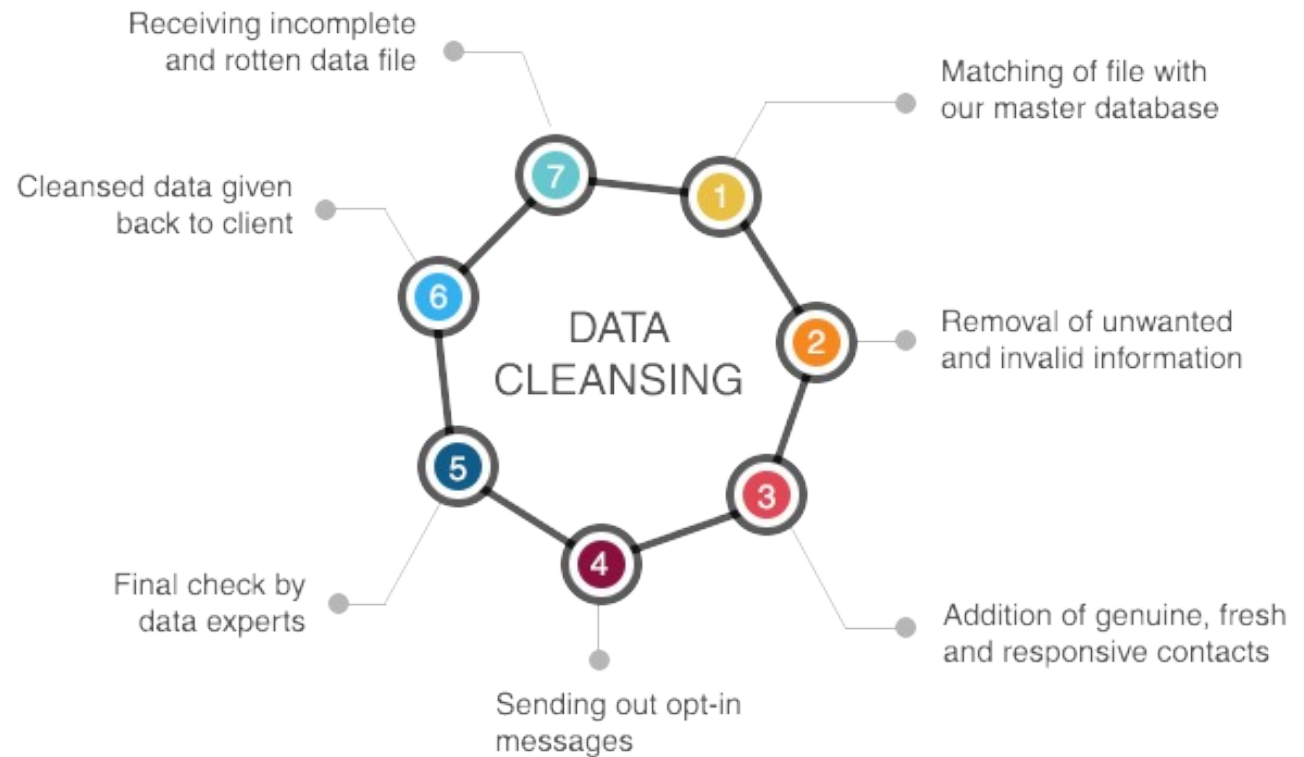
- 훈련 데이터가 모집단을 잘 대표해야 정확한 모델을 얻을 수 있습니다.
- 지엽적인 문제만 풀어서는 좋은 성적을 얻기 힘든 것과 마찬가지로.
- 그러나, 생각보다 편향된 데이터가 수집되기가 쉽습니다.



## 1.5 머신러닝의 주요 도전 과제

### • 1.5.3 낮은 품질의 데이터

- 훈련 데이터에 에러, 이상치, 노이즈 등이 적어야 잘 작동합니다.
- 오류가 많은 문제집을 풀면 좋은 성적을 얻기 힘든 것과 마찬가지로.
- 훈련 데이터의 품질을 정제하고 가공하는 것을 Data Cleansing이라고 합니다.



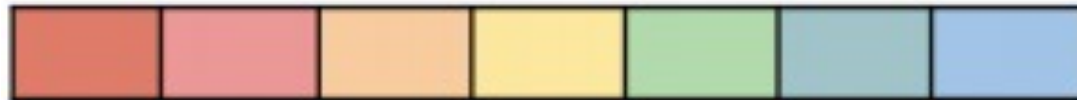


## 1.5 머신러닝의 주요 도전 과제

- 1.5.4 **관련 없는 특성**

- 훈련 데이터에 쓸모 없는 특성은 적고,  
좋은 특성이 많아야 좋은 모델을 얻습니다.
- “쓰레기를 넣으면 쓰레기가 나온다.”
- 특성 공학 (Feature engineering)

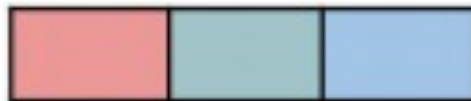
All Features



Feature Selection



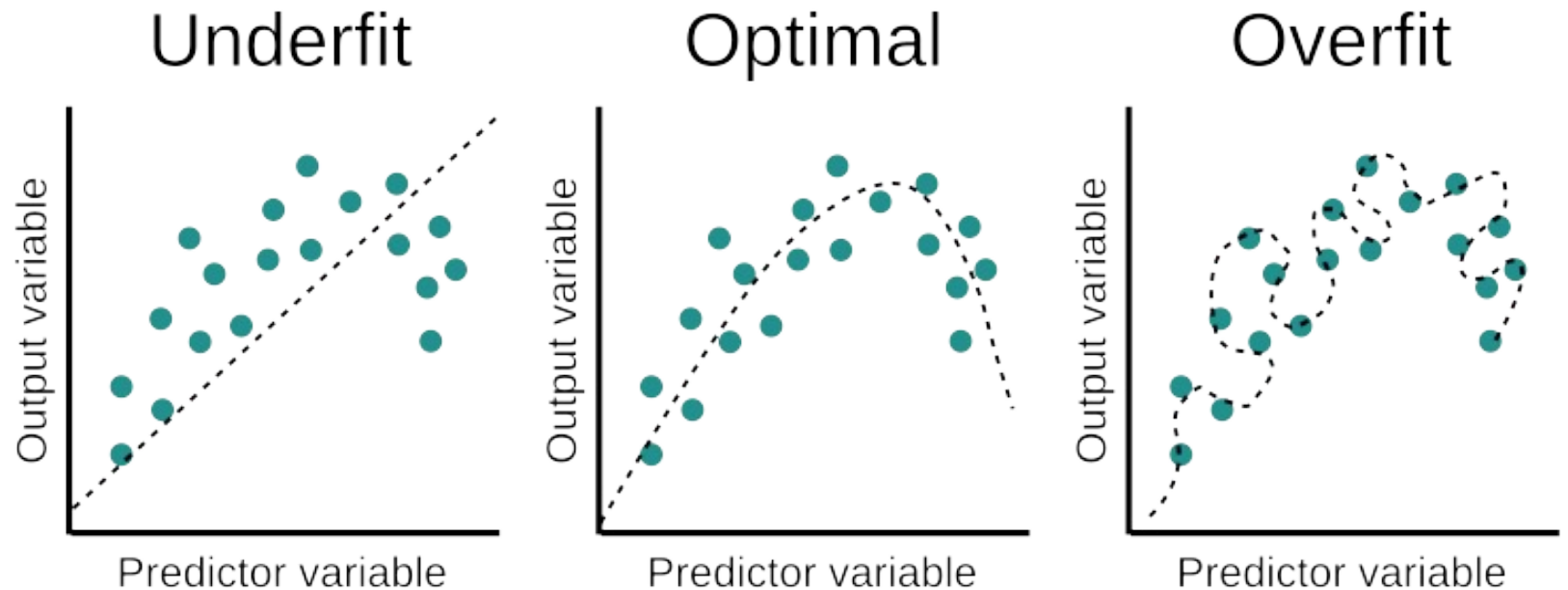
Final Features



## 1.5 머신러닝의 주요 도전 과제

- 1.5.5 **훈련 데이터 과대적합 (Overfitting)**

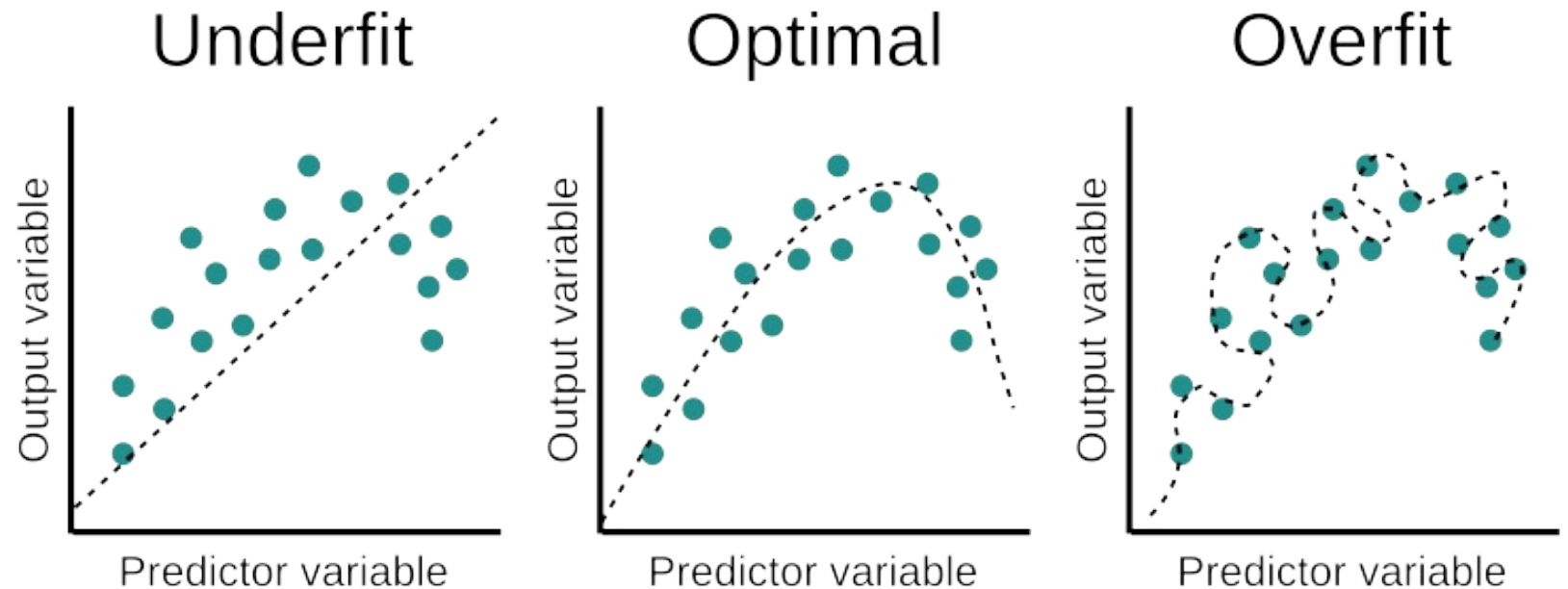
- 모델이 훈련 데이터에 대하여 성급한 일반화를 하지 않도록 유의해야 합니다.
- “우리 아이가 내신 점수는 잘 받는데, 모의고사만 치면 점수가 떨어져요”
- 오버피팅을 해결하는 것도 중요하지만,  
오버피팅이 일어났는지 판단하는 것도 중요.



## 1.5 머신러닝의 주요 도전 과제

- 1.5.6 **훈련 데이터 과소적합 (Underfitting)**

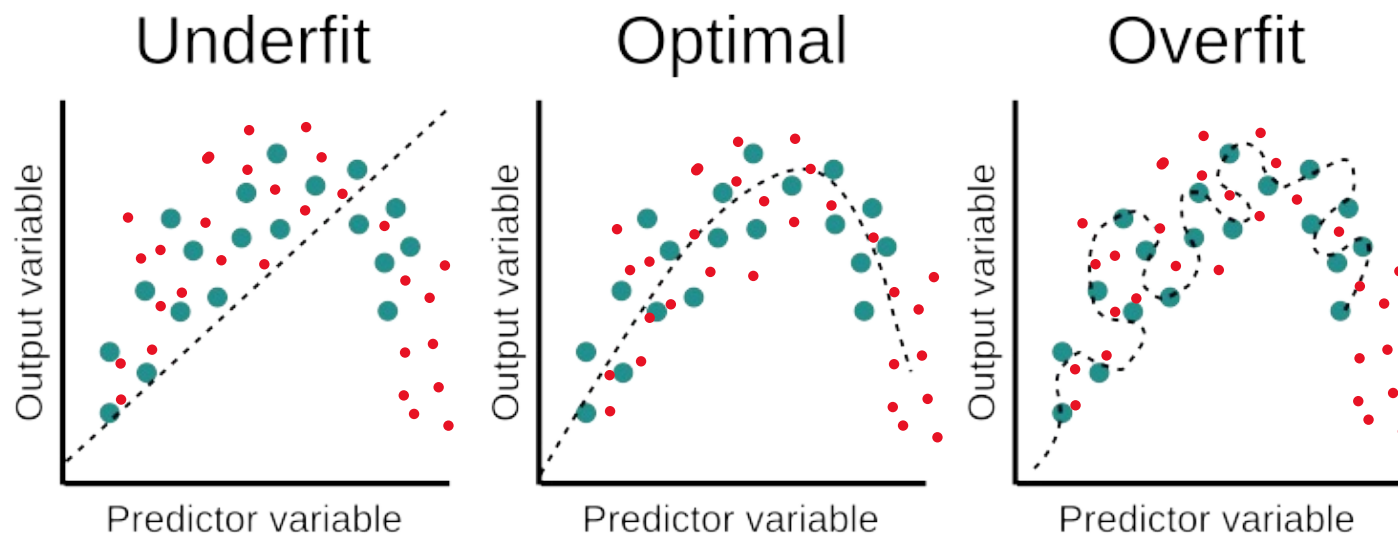
- 모델은 데이터의 복잡도를 충분히 반영할만큼 복잡해야 합니다.
- “작년 물수능만 보고 준비하면 올해 불수능에서 큰코다칩니다”
- 더 복잡한 모델을 선택하거나 더 좋은 특성을 제공해줍니다.



## 1.6 테스트와 검증

### 훈련시킨 모델이 얼마나 좋은지 어떻게 판단할까?

- 훈련 데이터로 오차를 판단해볼까? (NO!)
- 실제 서비스에 바로 론칭해볼까? (NO!)
- 테스트 데이터를 따로 만들어두고 모델을 평가한다! (YES!)

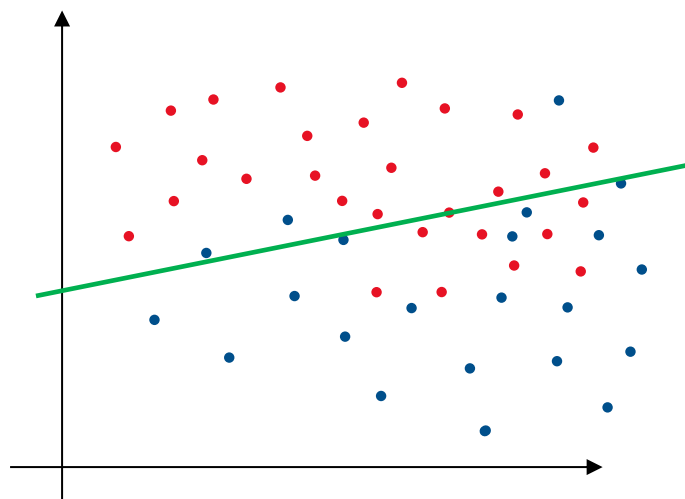


• Test data

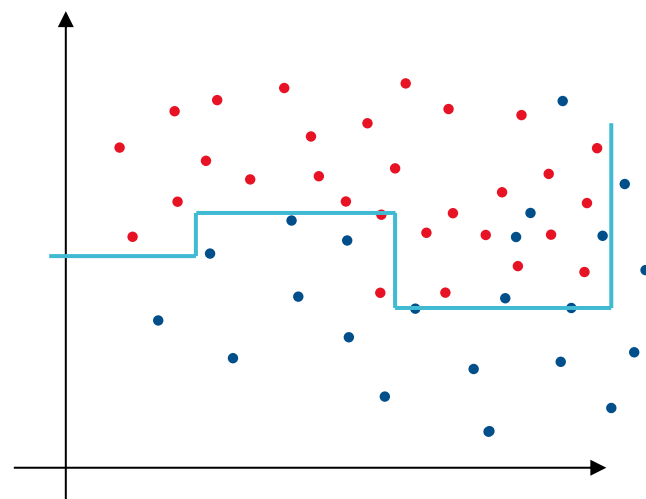
## 1.6 테스트와 검증

### 1.6.1 하이퍼파라미터 튜닝과 모델 선택

- 모델 A를 사용할까? 모델 B를 사용할까?



Model A



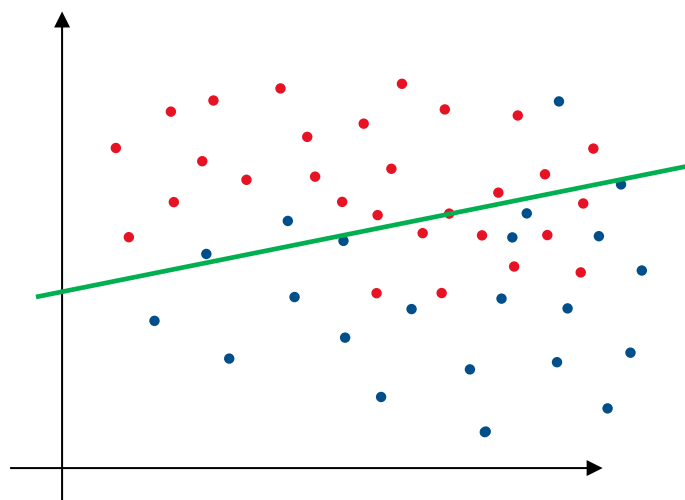
Model B

Test set으로 두 모델의 일반화 오차를 평가해서 선택

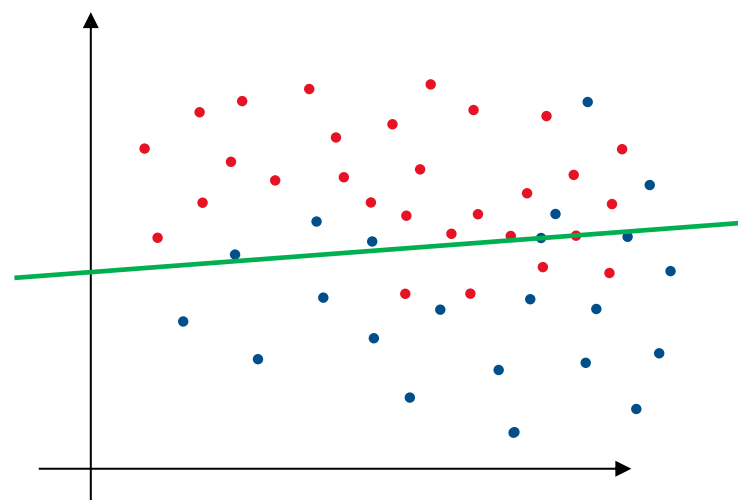
## 1.6 테스트와 검증

### 1.6.1 하이퍼파라미터 튜닝과 모델 선택

- 모델 A의 파라미터를 어떻게 결정할까?



Model A (with L2 norm)



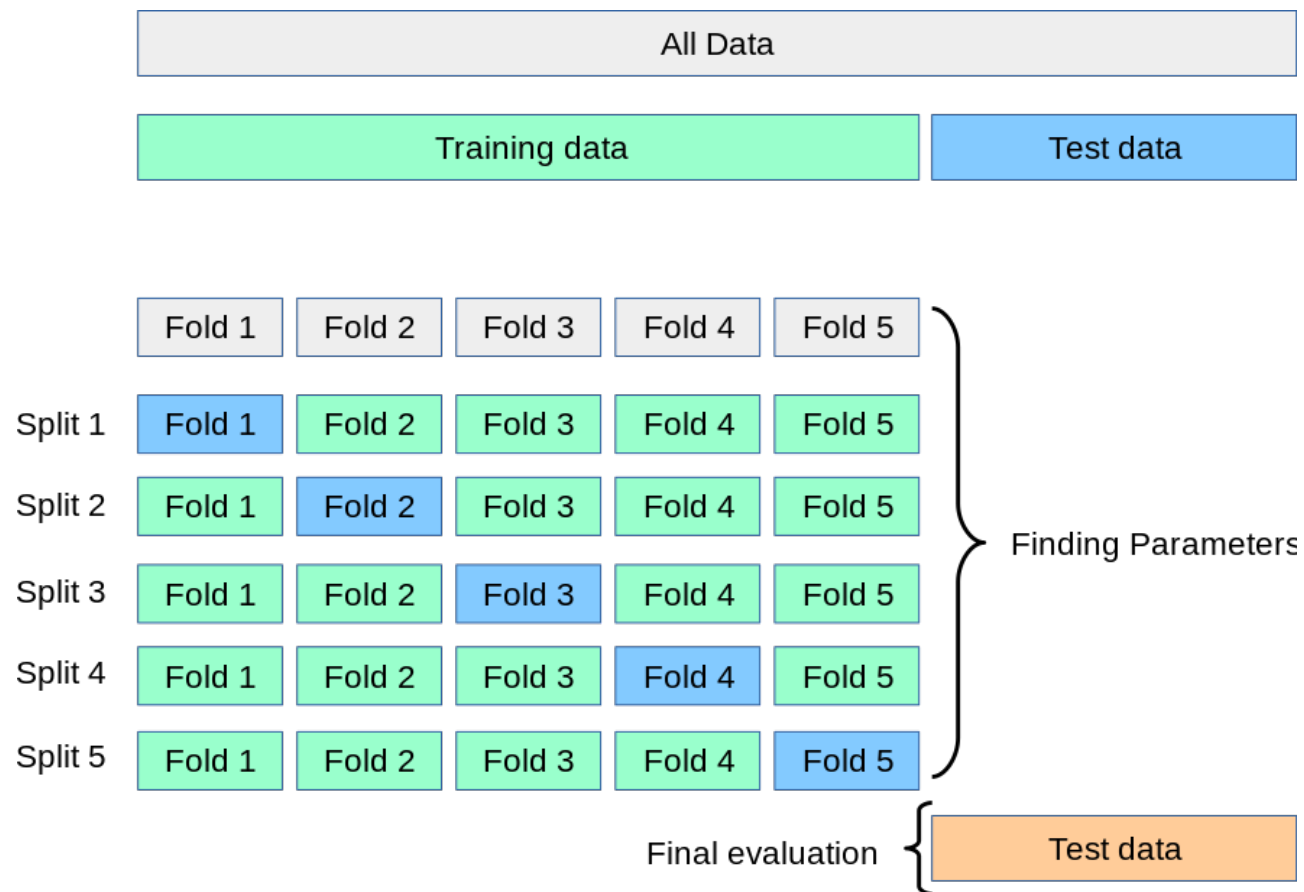
Model A (with L1 norm)

Validation set( $\subset$  Train set)으로 평가 (Cross-Validation)

## 1.6 테스트와 검증

### 1.6.1 하이퍼파라미터 튜닝과 모델 선택

- K-fold Cross Validation (CV)



# Pipeline Summary

