

# 1강. 데이터분석 개요

2023.01.03.

양희철

hcyang@cnu.ac.kr

# Introduction

---

## 강의 개요

- Contents
  - Introduction
  - Data preprocessing (데이터 전처리)
  - Data visualization (데이터 시각화)
  - Regression (회귀)
  - Classification (분류)
  - Clustering (군집화)
  - Deep learning

# Introduction

---

## 강의 개요

- Course materials
  - ppt slides
  - 따라하며 배우는 파이썬과 데이터 과학, 천인국·박동규·강영민 저, 생능출판
  - 실습 자료
- Prerequisite
  - Python programming
  - 선형대수
  - 확률 및 통계

# Introduction

---

## 데이터 분류

- Structured or unstructured
- i.i.d. data or non-i.i.d. data (\*i.i.d.: independent and identically distributed)
- Vectorial or non-vectorial data
- Labeled or unlabeled data
- Images, text, languages, time series, graphs, and so on

# Introduction

---

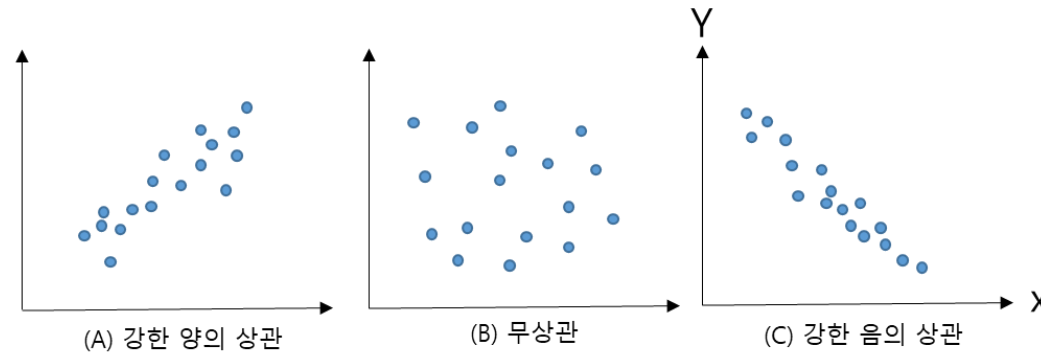
## Regression

- 상관분석(Correlation analysis)
  - 두 연속형 변수 사이 상관관계가 존재하는지를 파악하고, 상관관계의 정도를 확인 하는 것
  - 상관분석에서는 관련성을 파악하는 지표로 상관계수(Correlation coefficient)라는 통계학적 관점에서 선형적 상관도를 확인하여 정도를 파악
- 상관 분석 과정
  - 산점도(Scatter) 두 변수 상관 파악
  - 상관계수 확인
  - 의사결정

# Introduction

## Regression

- 상관분석(Correlation analysis)
  - 상관계수



상관	상관계수
음의 상관	-1.0 ~ -0.7 이면, 강한 음의 상관관계 - 그림C -0.7 ~ -0.3 이면, 뚜렷한 음의 상관관계 -0.3 ~ -0.1 이면, 약한 음의 상관관계
무상관	-0.1 ~ +0.1 이면, 없다고 할 수 있는 상관관계 - 그림 B
양의 상관	+0.1 ~ +0.3 이면, 약한 양의 상관관계 +0.3 ~ +0.7 이면, 뚜렷한 양의 상관관계 +0.7 ~ +1.0 이면, 강한 양의 상관관계 - 그림A

# Introduction

---

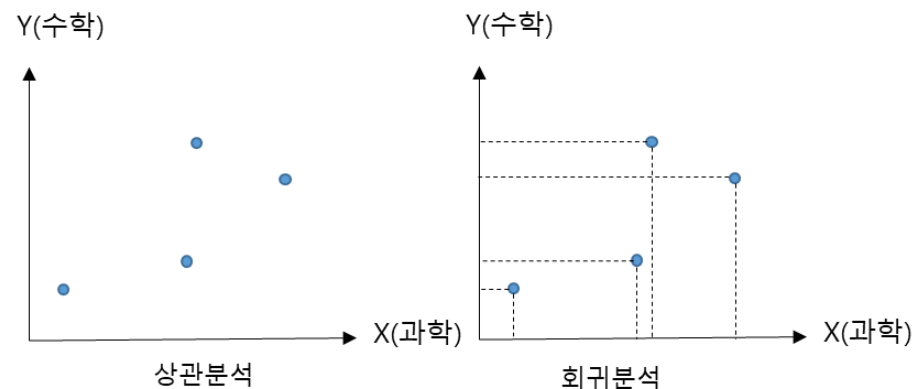
## Regression

- 회귀분석(Regression analysis)
  - 상관분석에서는 두 연속형 변수  $X$ (과학)와  $Y$ (수학)의 상관 정도만 알 수 있고 인과관계는 알 수 없음
  - 회귀분석에서는 두 연속형 변수  $X$ 와  $Y$ 를 독립변수와 종속변수라고 하는 인과관계로 설명할 수 있음
  - '과학 점수가 좋으면 수학점수가 좋을까요?' 와 같이 간단하지만 미래를 예측할 수 있는 머신러닝의 초기 모델이 됨

# Introduction

## Regression

- 회귀분석(Regression analysis)
  - 회귀분석 변수



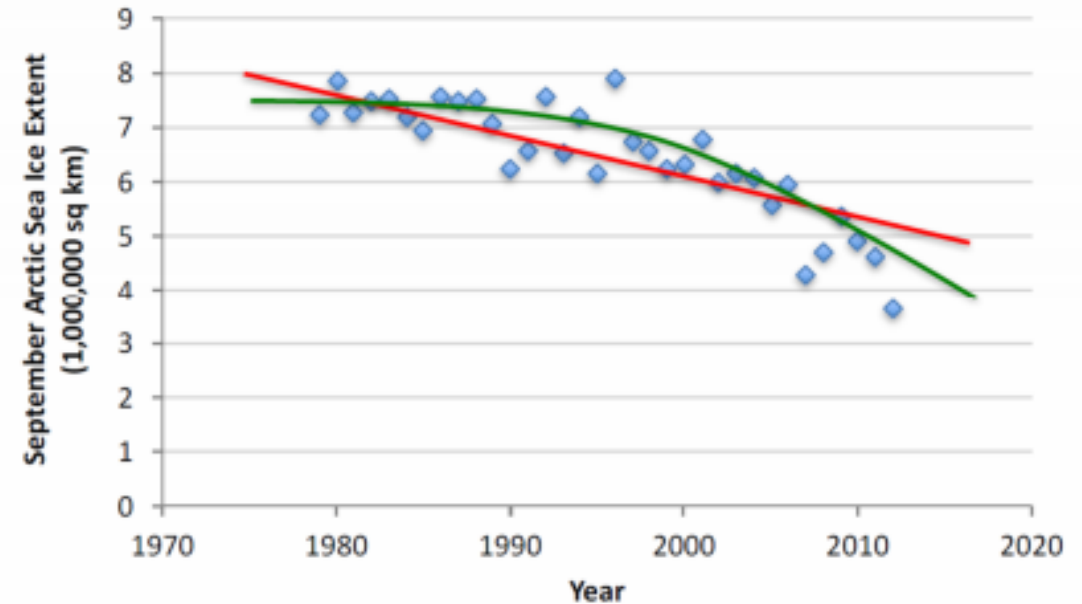
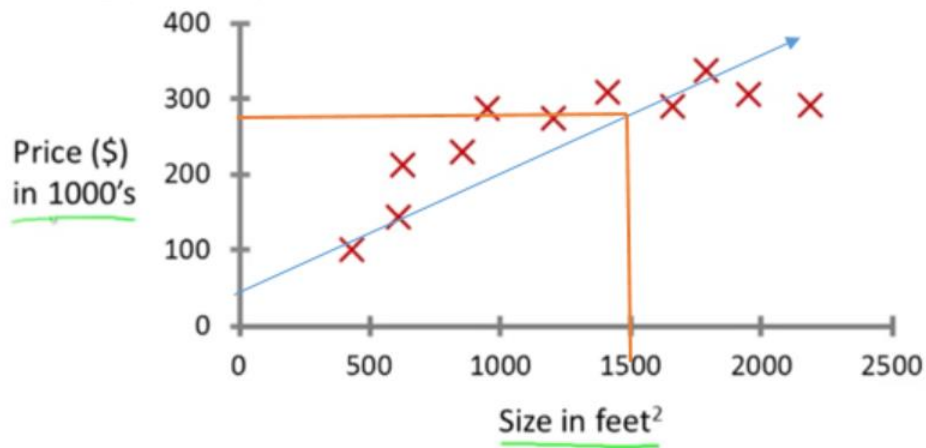
X	Y
독립변수, 설명변수, 원인변수	종속변수, 반응변수, 결과변수 머신러닝(클래스, 라벨)
다른 변수에 영향을 주는 원인	다른 변수에 영향을 받는 결과
Dependent Variable Response	Independent Variable, Predictor Factor



# Introduction

## Regression

- 회귀분석(Regression analysis)
  - 데이터  $x$ 에 대한 결과  $y$  (실수)를 통해 둘 사이의 함수  $f(x)$ 를 학습

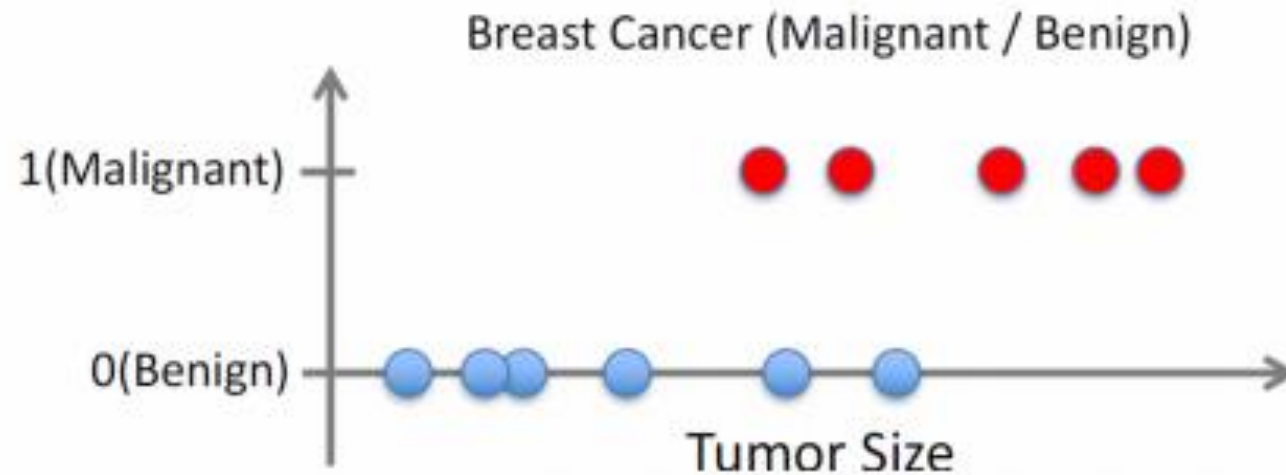


# Introduction

---

## Classification

- 데이터  $x$ 에 대한 결과  $y$  (분류값)를 통해 둘 사이의 분류 함수  $f(x)$ 를 학습
  - Tumor size를 통한 breast cancer 진단

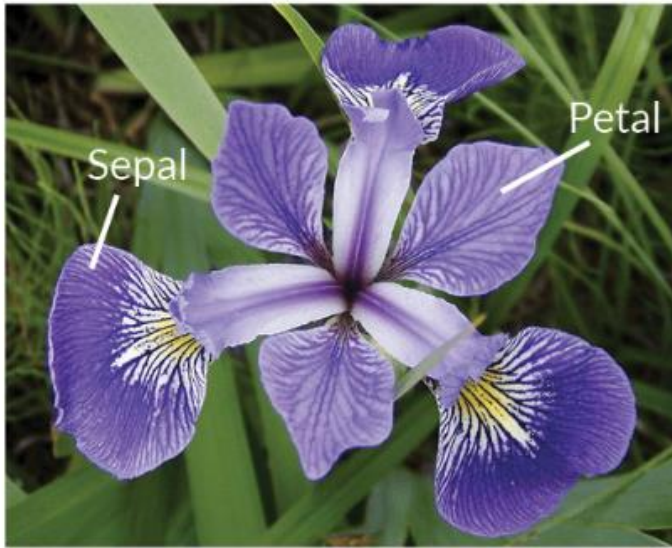


# Introduction

---

## Classification

- 데이터  $x$ 에 대한 결과  $y$  (분류값)를 통해 둘 사이의 분류 함수  $f(x)$ 를 학습
  - 붓꽃(iris) 이미지를 통한 종 분류



**Iris Versicolor**



**Iris Setosa**



**Iris Virginica**

# Introduction

---

## Classification

- 데이터  $x$ 에 대한 결과  $y$  (분류값)를 통해 둘 사이의 분류 함수  $f(x)$ 를 학습
  - 이미지 분류

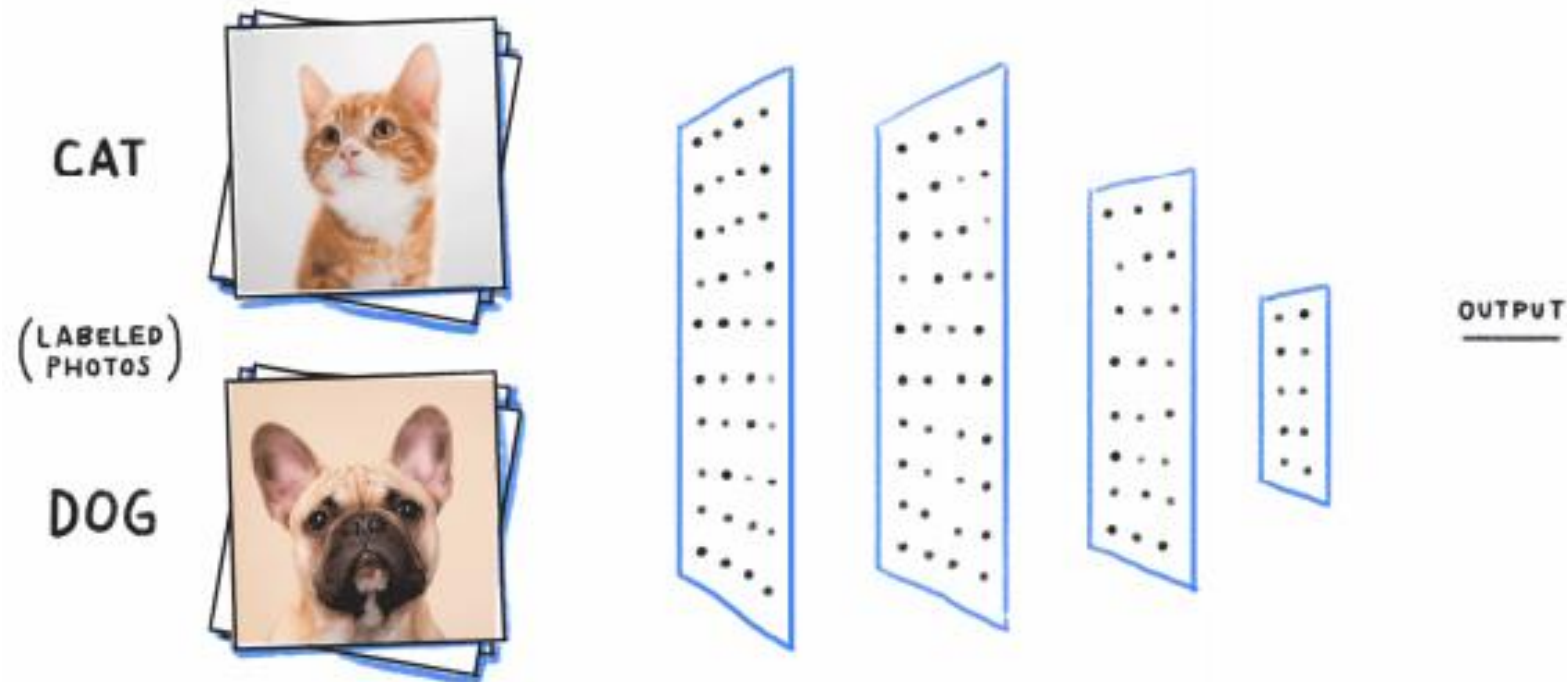


# Introduction

---

## Classification

- 데이터  $x$ 에 대한 결과  $y$  (분류값)를 통해 둘 사이의 분류 함수  $f(x)$ 를 학습
  - Deep neural networks를 통한 이미지 분류

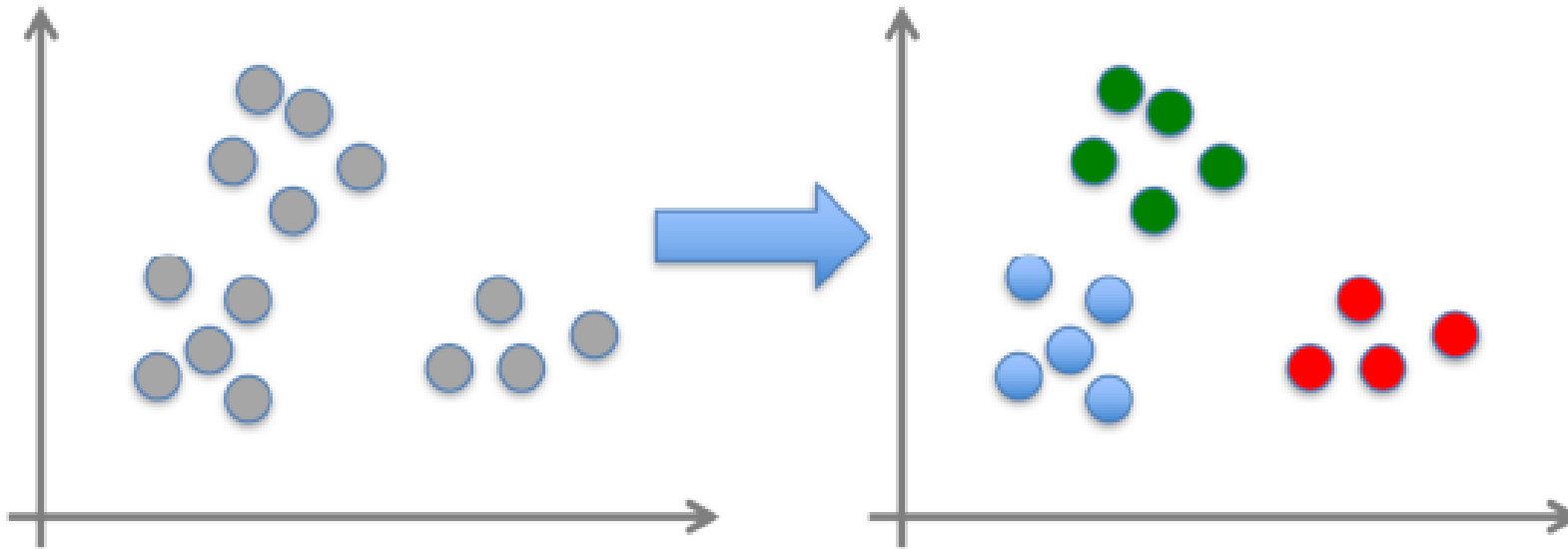


# Introduction

---

## Clustering

- 데이터 사이의 숨겨진 구조를 밝혀 비슷한 데이터들을 군집화

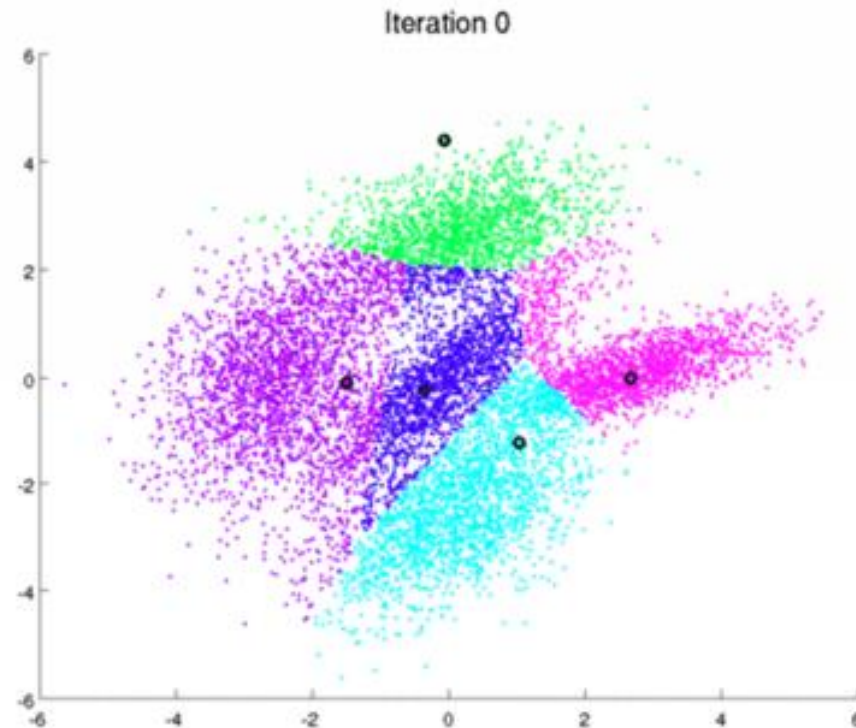


# Introduction

---

## Clustering

- 데이터 사이의 숨겨진 구조를 밝혀 비슷한 데이터들을 군집화
  - K-means clustering

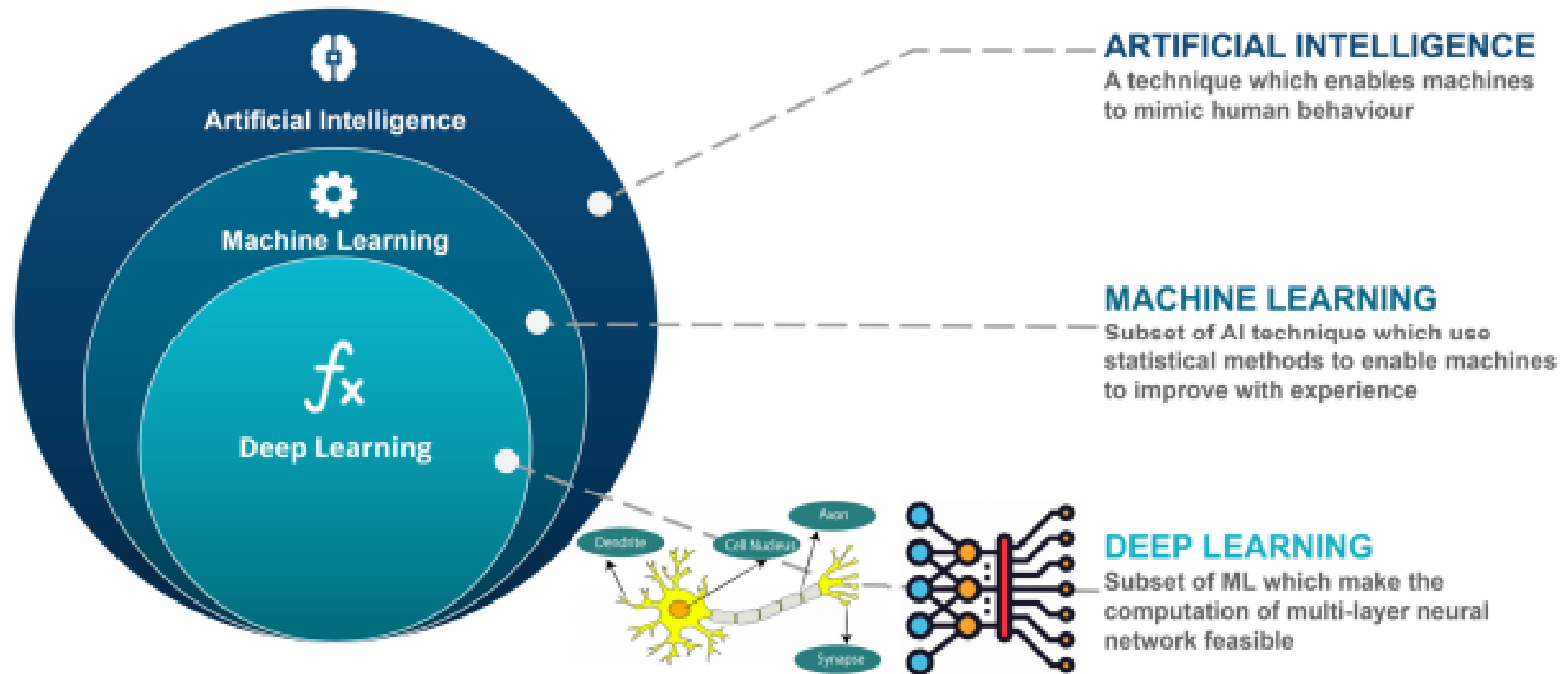




# Introduction

---

## Machine learning





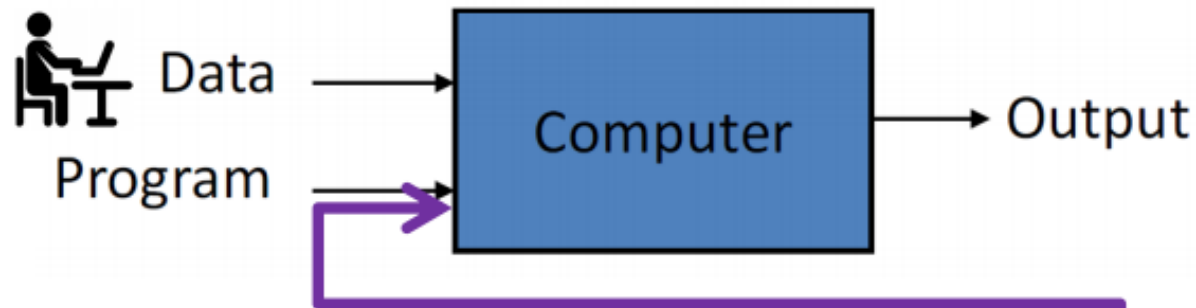
# Introduction

---

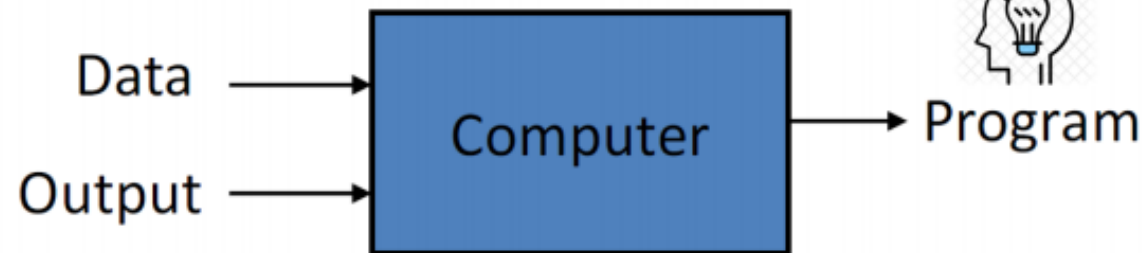
## Machine learning

- 전통적 프로그래밍과 머신러닝의 차이

### ▶ Traditional Programming



### ▶ Machine Learning (ML)

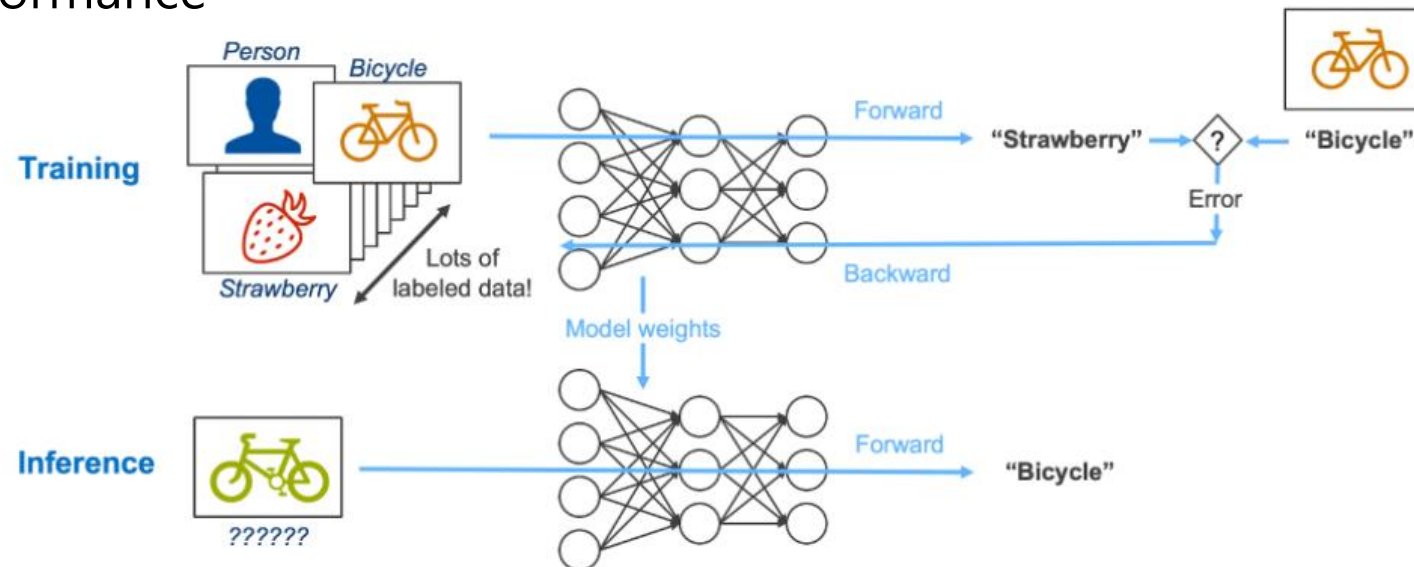


# Introduction

---

## Machine learning

- Training (learning): With training data, extract a model (approximate input-output relationship (function))
- Test (inference): With the trained model, apply the model in practice with test data, and measure the performance



# Introduction

---

## Machine learning

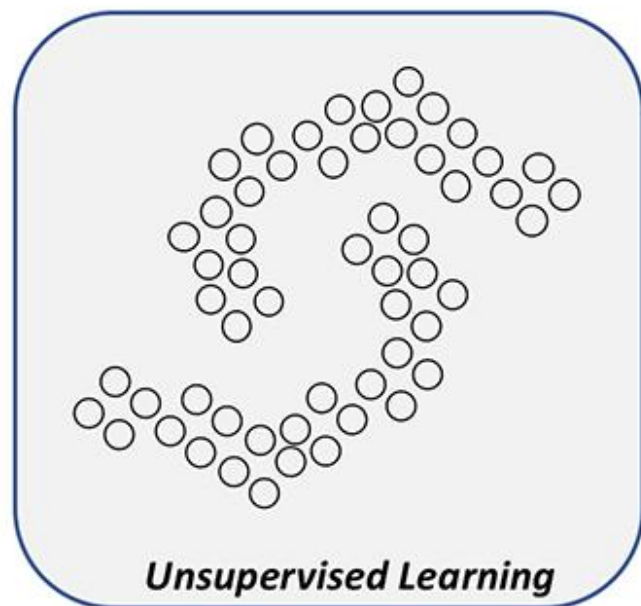
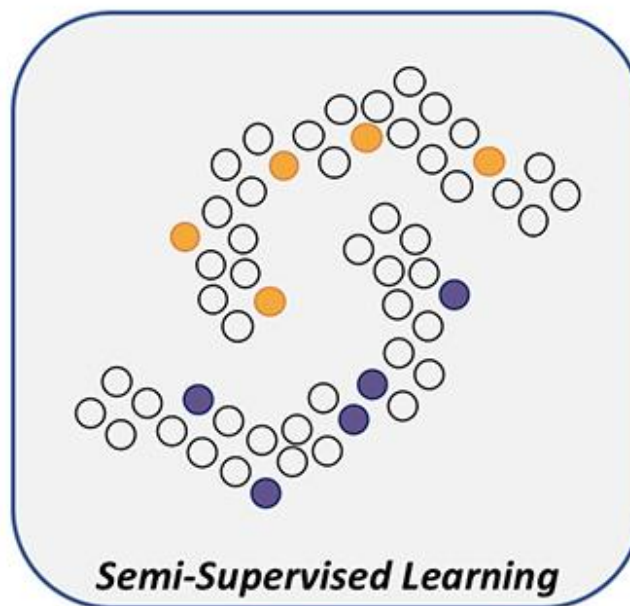
- 지도학습(Supervised learning)
  - Given: training data + desired outputs (labels)
  - ex) regression, classification
- 비지도학습(Unsupervised learning)
  - Given: training data (without desired outputs)
  - ex) clustering
- \*준지도학습(Semi-supervised learning)
  - Given: training data + a few desired outputs
- 강화학습(Reinforcement learning)
  - Given: Rewards from sequence of actions

# Introduction

---

## Machine learning

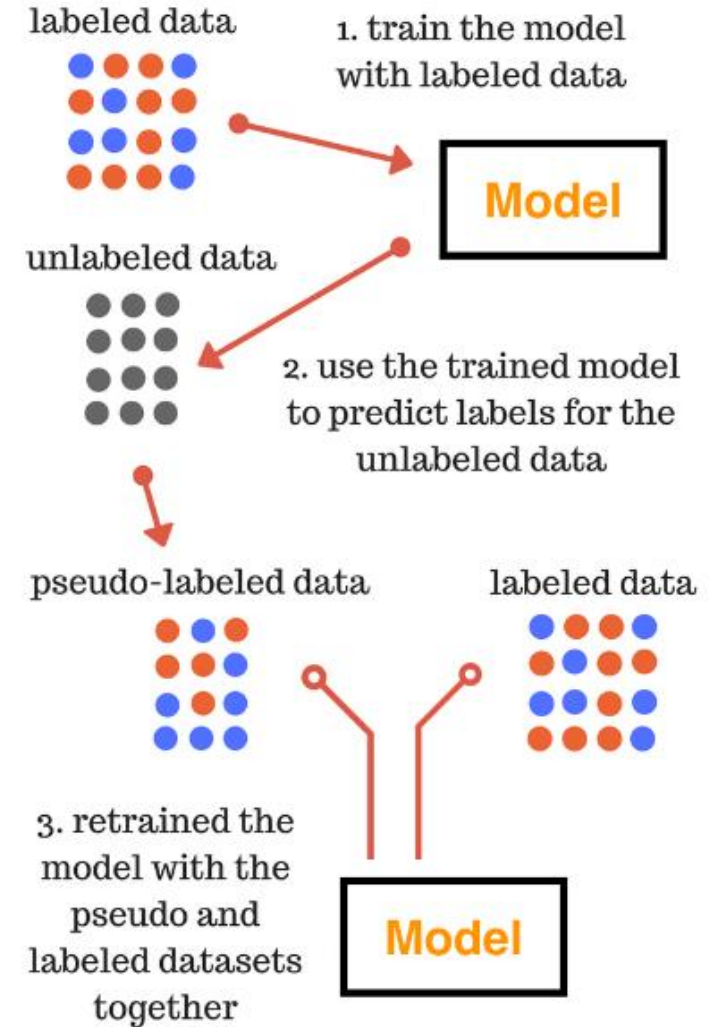
- Semi-supervised learning
  - Labeled data와 unlabeled data가 모두 사용되는 머신러닝 기법
  - 데이터를 수집하는 '데이터 레이블링' 작업에 소요되는 자원과 비용 감소



# Introduction

## Machine learning

- Semi-supervised learning
  - Proxy-label method: labeled data로 학습된 모델을 이용해 unlabeled data에 label을 달아주는 기법

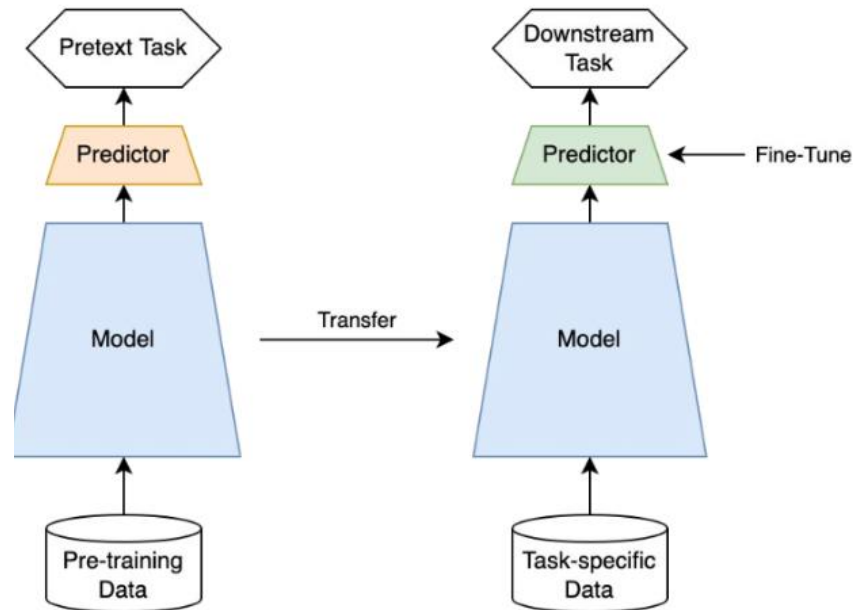


# Introduction

---

## Machine learning

- Self-supervised learning
  - Unlabeled dataset으로부터 good representation을 얻고자 하는 학습 방법으로 비지도 학습의 일종
  - 모델 스스로 task를 정해서 지도학습 방식으로 모델을 학습

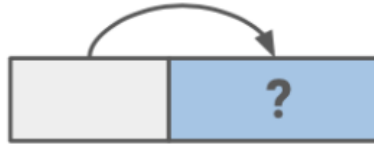


# Introduction

---

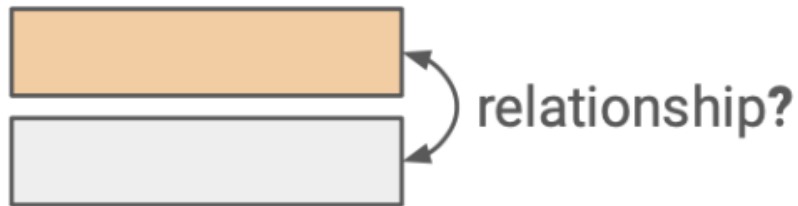
## Machine learning

- Self-supervised learning
  - Self-prediction: 개별 data sample에 대해, sample 내의 한 부분을 통해서 다른 부분을 예측하는 task 수행



**“Intra-sample” prediction**

- Contrastive learning: batch 내의 data sample이 주어졌을 때, 그들 사이의 관계를 예측하는 task 수행



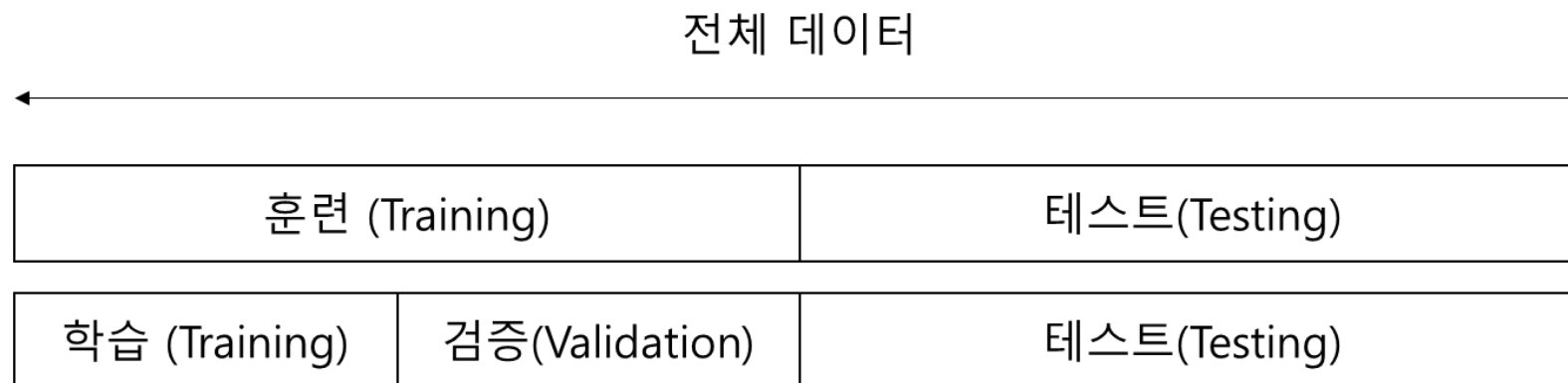
**“Inter-sample” prediction**

# Introduction

---

## 모델 평가

- 데이터 분석 모델이 완성되었다면 전체 데이터를 모델 생성을 위한 훈련용, 평가를 위한 테스트용 두 가지로 분할하여 사용
- 특별한 경우 훈련용, 검증용, 테스트용으로 분리해서 사용하기도 함





# Introduction

---

## 모델 평가

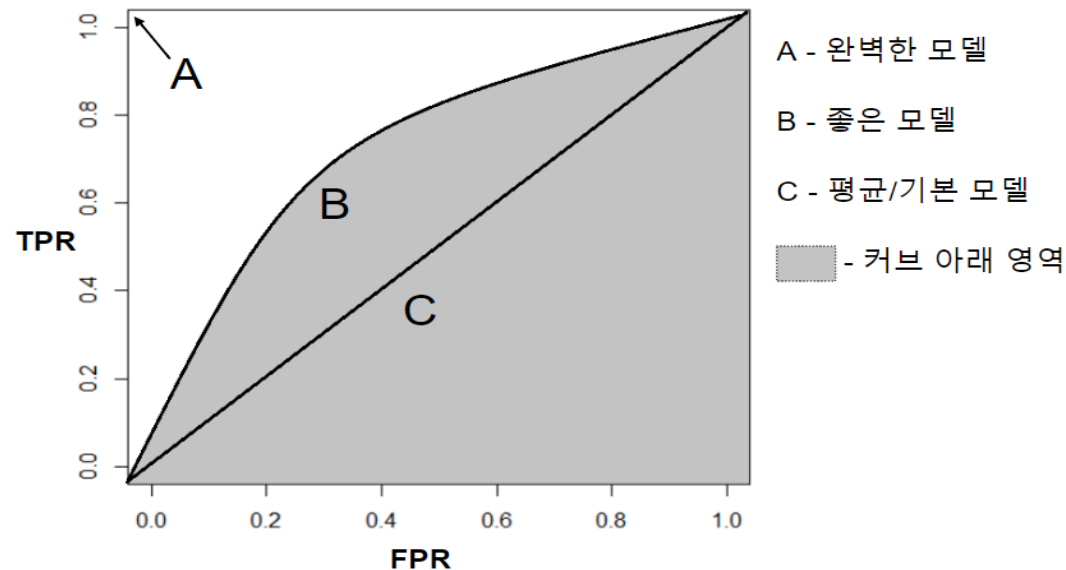
- 데이터 분석 모델을 평가하는 요소로 얼마나 정확하게 예측한 결과와 실제 정답이 일치하는지 검증
- 정확도, 정밀도, 재현율 등을 평가
  - 정확도(accuracy):  $(TP+TN)/(TP+FN+FP+TN)$
  - 정밀도:  $TP/(TP+FP)$
  - 재현율(recall):  $TP/(TP+FN)$

		정답	
		참(True)	거짓(False)
예측	예측/정답 참(True)	TP	FP
	거짓(False)	FN	TN

# Introduction

## 모델 평가

- ROC(Receiver Operating Characteristic)
  - y축에 TP, x축에 FP 수치를 배치해 두 수치의 균형을 살펴보는 머신러닝 평가를 위한 시각화 모델
  - ROC 커브 아래 면적인 AUC(Area Under Curve)가 1에 가까워질수록, 모델이 Y를 예측하는 정확도가 높은 모델

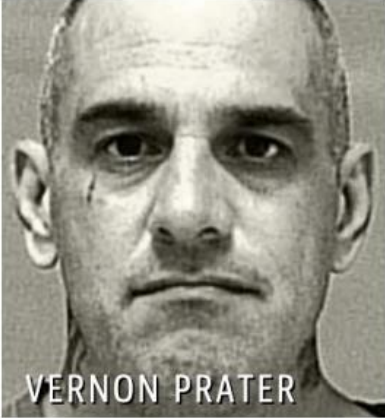



# Introduction

---

## 모델 평가

- 모델의 fairness 평가
  - 최근 데이터 분석 모델의 fairness에 대한 평가를 중요하게 생각하는 추세

	
VERNON PRATER	BRISHA BORDEN
<b>Prior Offenses</b> 2 armed robberies, 1 attempted armed robbery	<b>Prior Offenses</b> 4 juvenile Misdemeanors
LOW RISK 3	HIGH RISK 8

# Introduction

---

## 모델 평가

- 모델의 fairness 평가
  - 최근 데이터 분석 모델의 fairness에 대한 평가를 중요하게 생각하는 추세

A thermometer was labelled “gun” only for a dark-skinned hand



# Introduction

---

## 모델 평가

- 모델의 fairness 평가
  - 최근 데이터 분석 모델의 fairness에 대한 평가를 중요하게 생각하는 추세

A faculty member has been asking how to stop Zoom from removing his head when he uses a virtual background. We suggested the usual plain background, good lighting etc, but it didn't work. I was in a meeting with him today when I realized why it was happening.



# Introduction

---

## 모델 평가

- 모델의 fairness 평가
  - 최근 데이터 분석 모델의 fairness에 대한 평가를 중요하게 생각하는 추세

