

# Machine Learning

2022. 3. 25

정 준 수 Ph.D

# Machine Learning에서 “예측”이란?

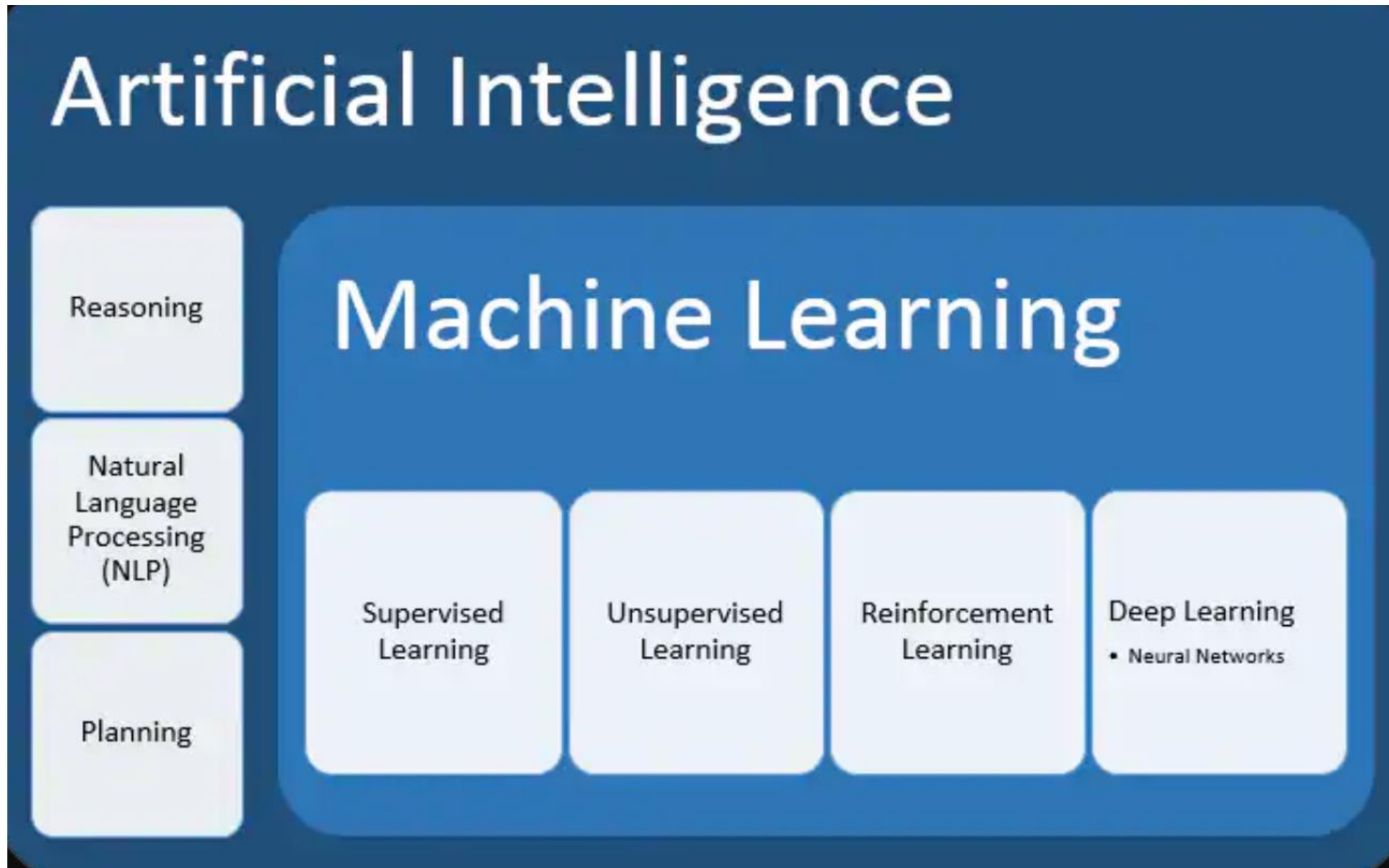
이전에 본적 없는 새로운 데이터에 대한 정확한 출력 예측

# Predictive Analytics

1 무엇을 예측하는가?

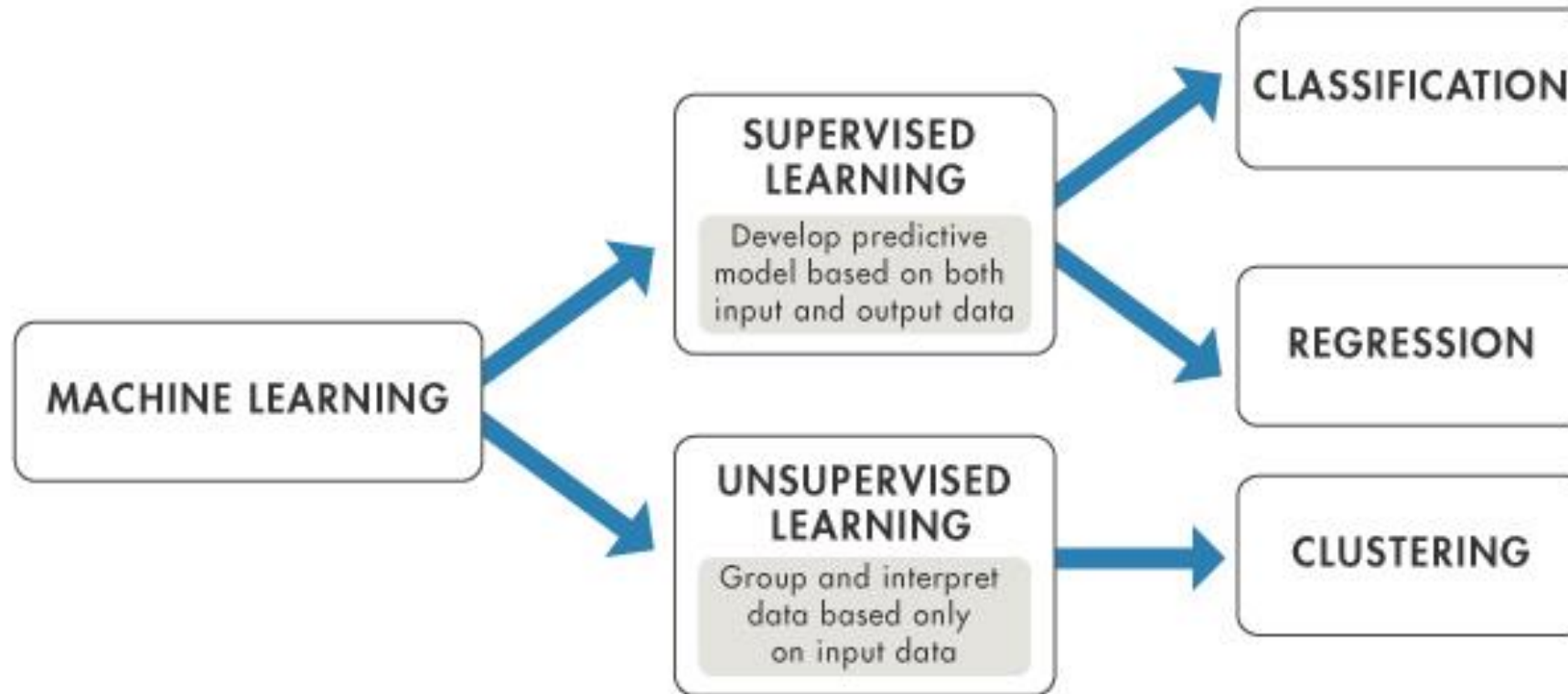
2 무엇을 할 것인가?

# Machine Learning



Machine learning uses two types of techniques:

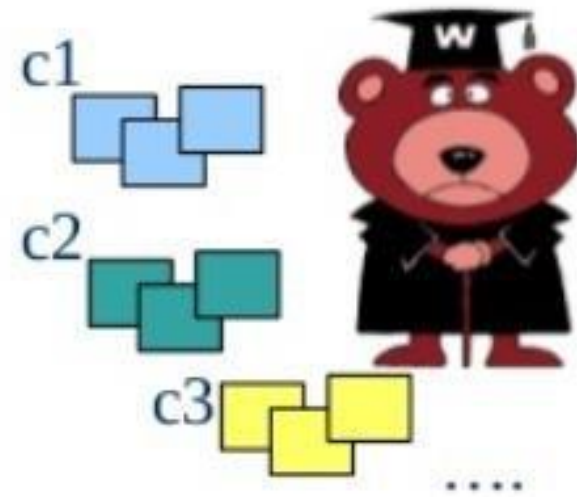
- Supervised learning, which trains a model on known input and output data so that it can predict future outputs,
- Unsupervised learning, which finds hidden patterns or intrinsic structures in input data.



# Supervised Vs. Unsupervised

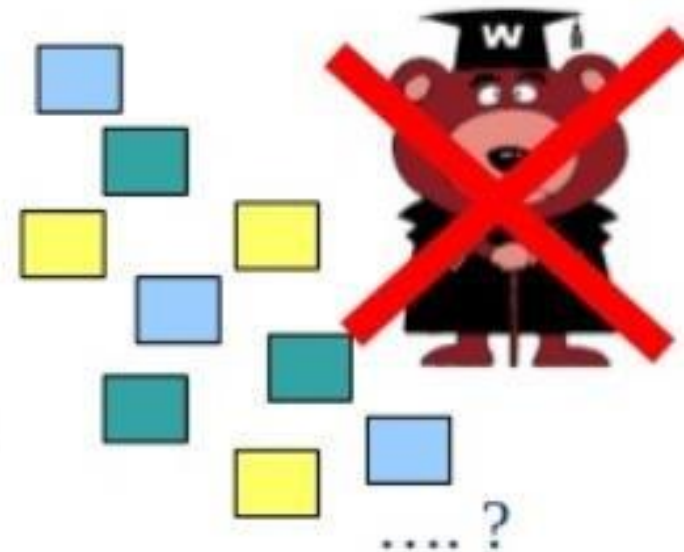
- **Supervised**

- **knowledge of output** - learning with the presence of an “expert” / teacher
  - data is **labelled** with a class or value
  - **Goal:** predict class or value label
    - e.g. Neural Network, Support Vector Machines, Decision Trees, Bayesian Classifiers ....

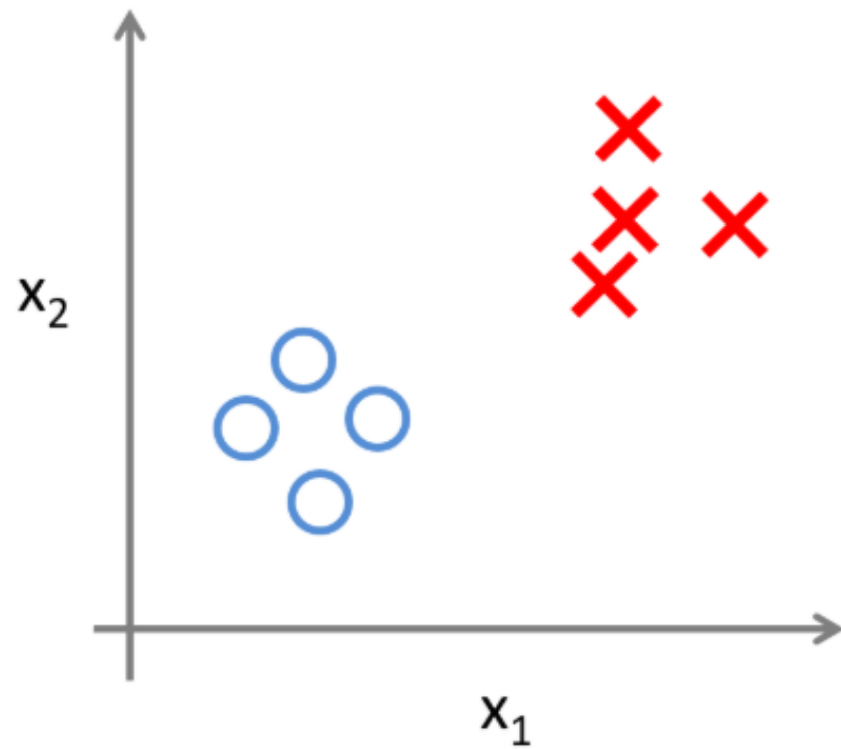


- **Unsupervised**

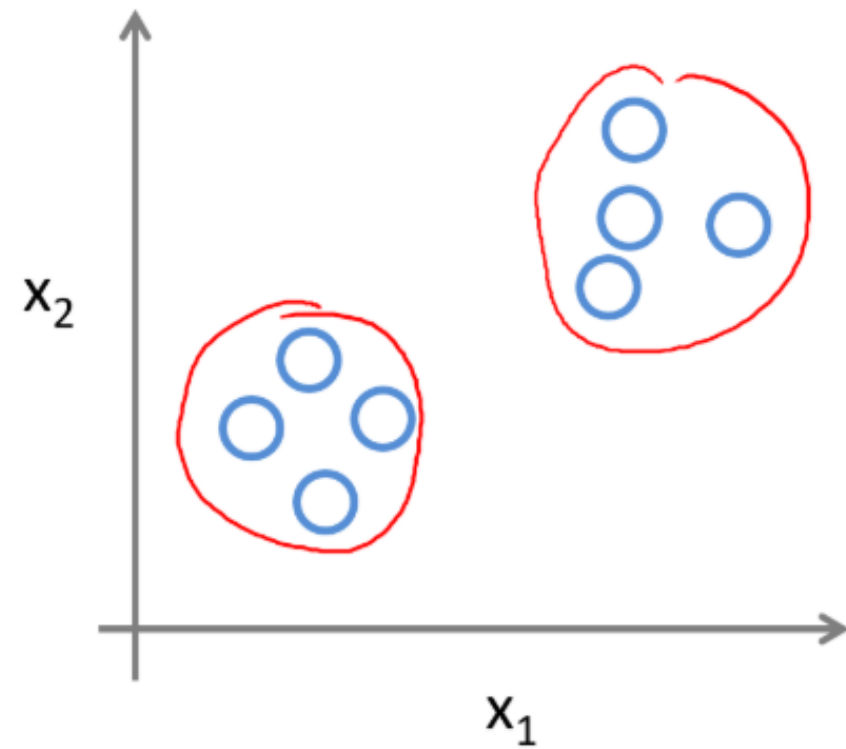
- **no knowledge of output** class or value
  - data is **unlabelled** or value un-known
  - **Goal:** determine data patterns/groupings
- Self-guided learning algorithm
  - (internal self-evaluation against some criteria)
  - e.g. k-means, genetic algorithms, clustering approaches ...



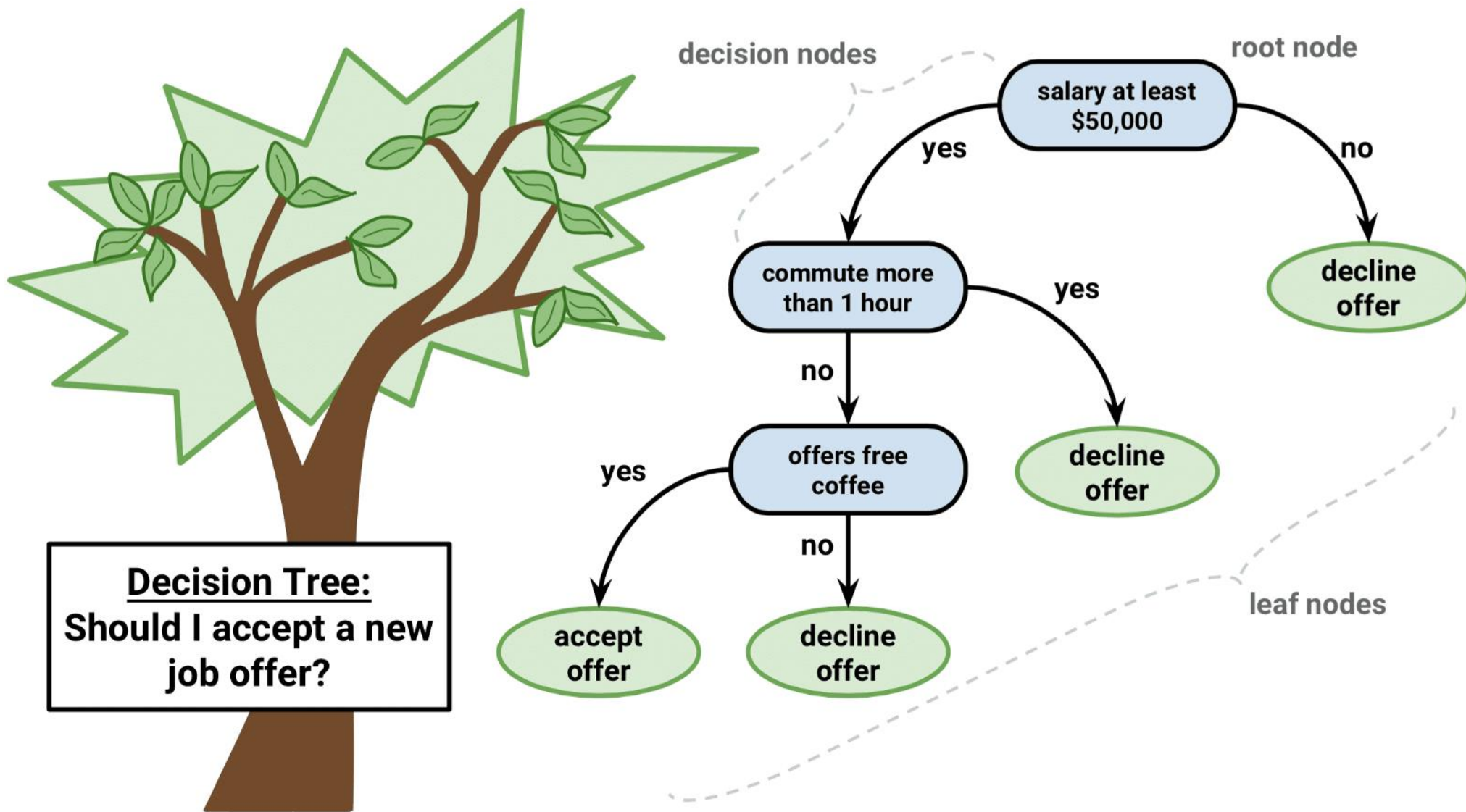
## Supervised Learning



## Unsupervised Learning



# Decision Tree





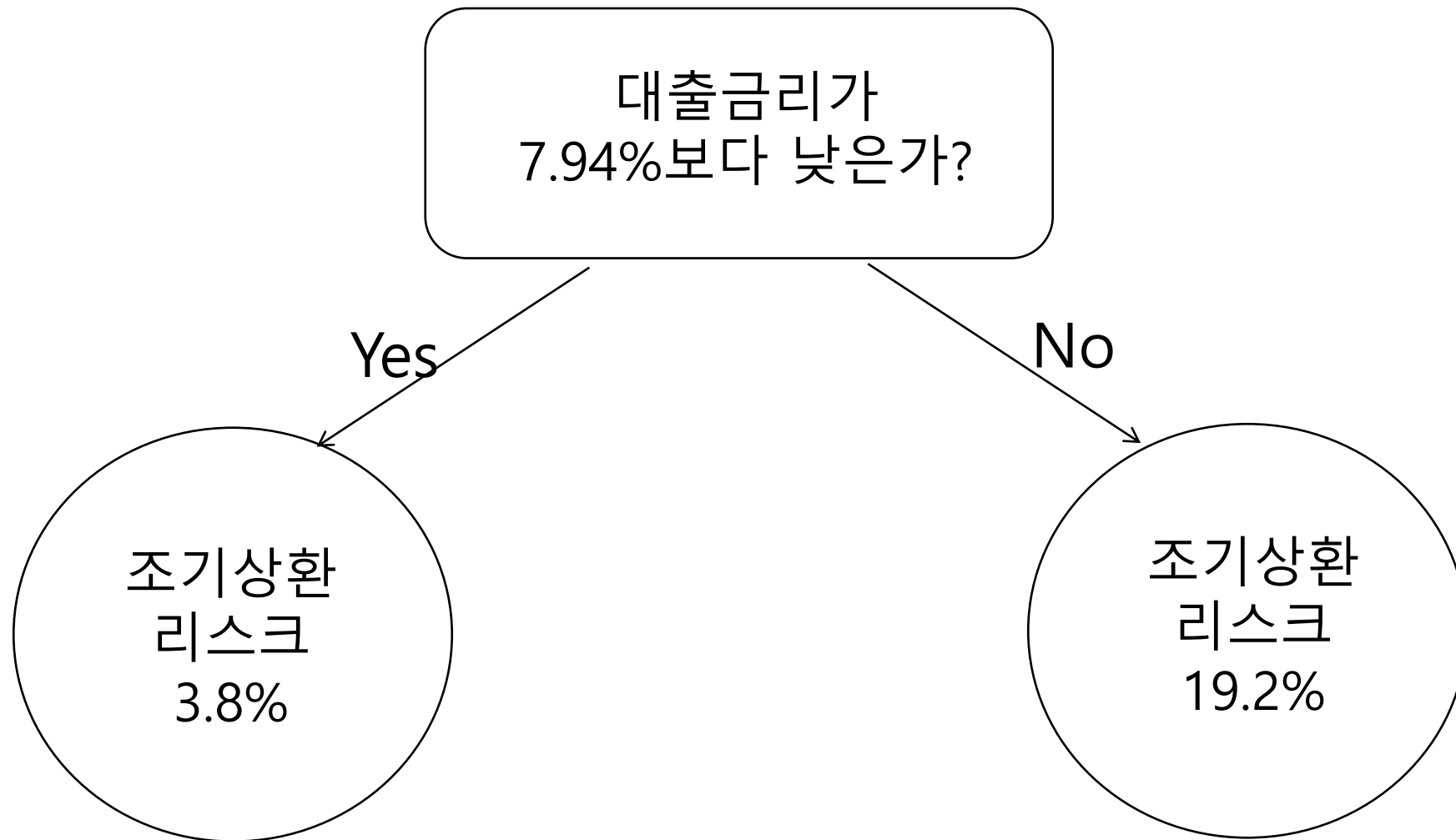
# 예측 분석 응용: 이탈 모델링으로 고객 이탈 방지하기

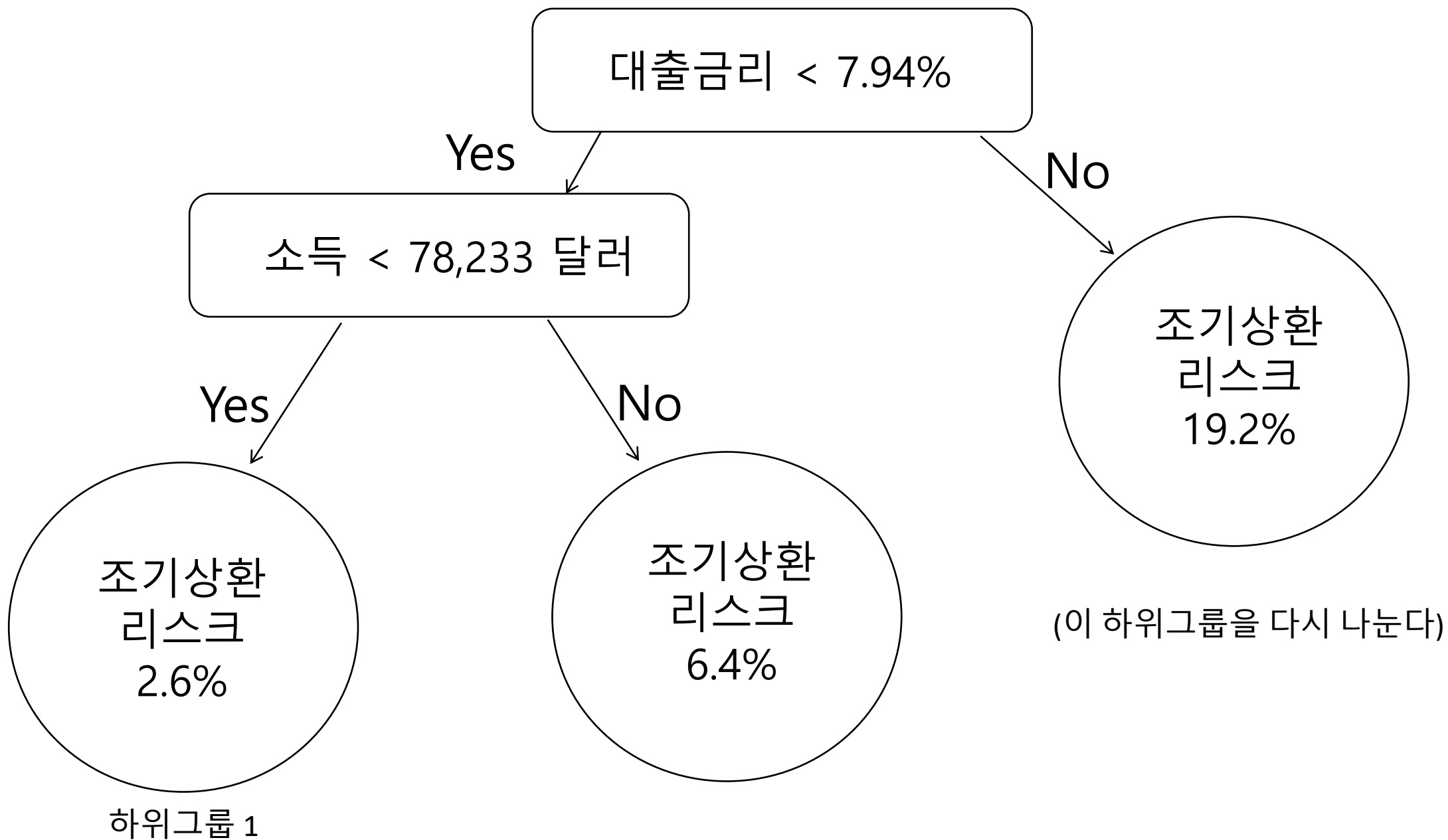
## 1 무엇을 예측하는가?

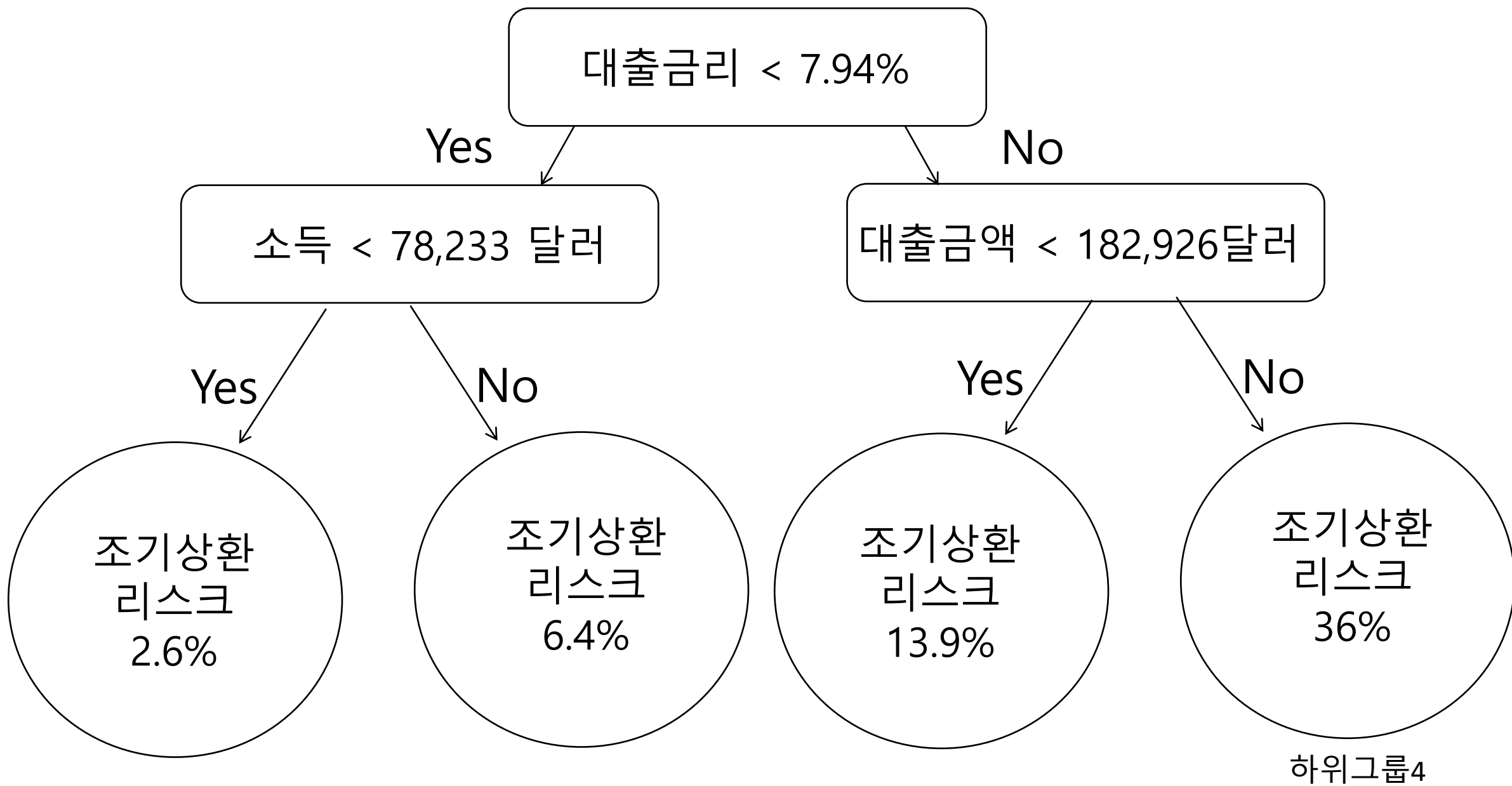
어느 고객이 떠나갈 것인가.

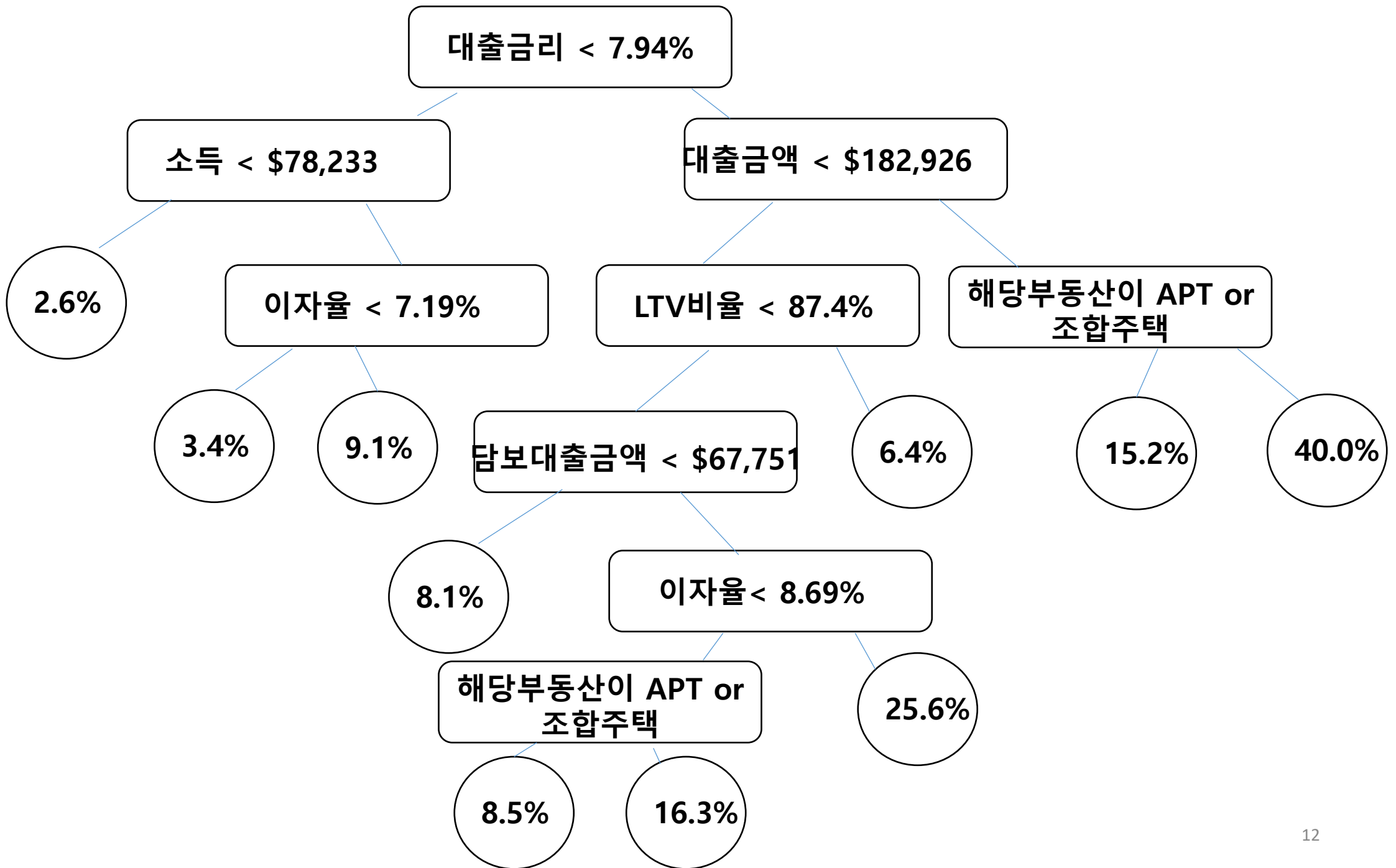
## 2 무엇을 할 것인가?

떠날 위기에 있는 고객들을 타겟으로 한 고객 유지 마케팅을 수행한다.









### 만약(IF):

부동산 담보대출 금액이 67,751 달러와 같거나 그보다 더 많고 182,926 달러보다 작다.

### 그리고(AND):

이자율이 8.69%와 같거나 그보다 더 높다.

### 그리고(AND):

부동산 자산가치 대비 대출금액의 비율이 87.4% 보다 작다.

### 그러면(THEN):

조기상환 확률은 25.6% 이다.

인간이 의사결정을 할 때에 이성과 감정의 조화를 잘 이뤄가면서 의사결정을 할 것이라고 생각하겠지만, 인간은 결국 감정의 동물이라는 것이 임상실험을 통해 의학적으로 밝혀진 사실입니다. 즉 감정에 기반하여 잘못된 의사결정을 할 수 있는 위험이 항상 있다는 것입니다. 하지만 감정을 이기는 것이 있습니다. 바로 데이터입니다. **흘러가는 무형의 시간은 데이터로 그 모습을 남깁니다.** 즉 **데이터는 시간의 다른 이름입니다.** 이 데이터를 수집 · 축적 · 분석하여 감정을 이기는 무기로 사용함으로써 우리는 감정이 아닌 데이터에 기반하여 합리적인 의사결정을 할 수 있게 됩니다.

인공지능의 발전이 인간의 일자리를 위협할 것이라는 예측이 나오고 있습니다. 하지만, 인공지능의 발전이 인간과 인공지능 간의 경쟁을 야기한다고 보기보다는, 인공지능을 활용하는 자와 인공지능을 활용할 줄 모르는 자의 경쟁을 야기한다고 보는 것이 적절한 예측일 것입니다. 즉 일자리를 위협받는 사람은 모든 사람이 아니고 인공지능을 활용할 줄 모르는 사람입니다. 데이터 애널리틱스 분야에서도 마찬가지로 예측을 할 수 있습니다. 앞으로는 데이터를 활용하는 자와 데이터를 활용할 줄 모르는 자의 경쟁이 야기될 것이고, 데이터를 활용하는 자가 절대적인 우위를 점하게 될 것입니다. 그러므로 이제는 데이터 애널리틱스가 자신의 전공분야가 아니더라도 기본개념 정도는 반드시 알고 있어야 하는 시대가 되었습니다.

# 모델학습 데이터의 구성

예시) 현 시점으로부터 과거 3년 동안의 분기별 이탈 고객의 데이터를 가지고 다음 분기 예측

2022년 1분기	2022년 2분기	2022년 3분기	2022년 4분기
2021년 1분기	2021년 2분기	2021년 3분기	2021년 4분기
2020년 1분기	2020년 2분기	2020년 3분기	2020년 4분기
2019년 1분기	2019년 2분기	2019년 3분기	2019년 4분기

모델학습 데이터

예측 분기



# 학습 모델의 평가

예시) 현 시점으로부터 과거 3년 동안의 분기별 이탈 고객의 데이터를 가지고 다음 분기 예측

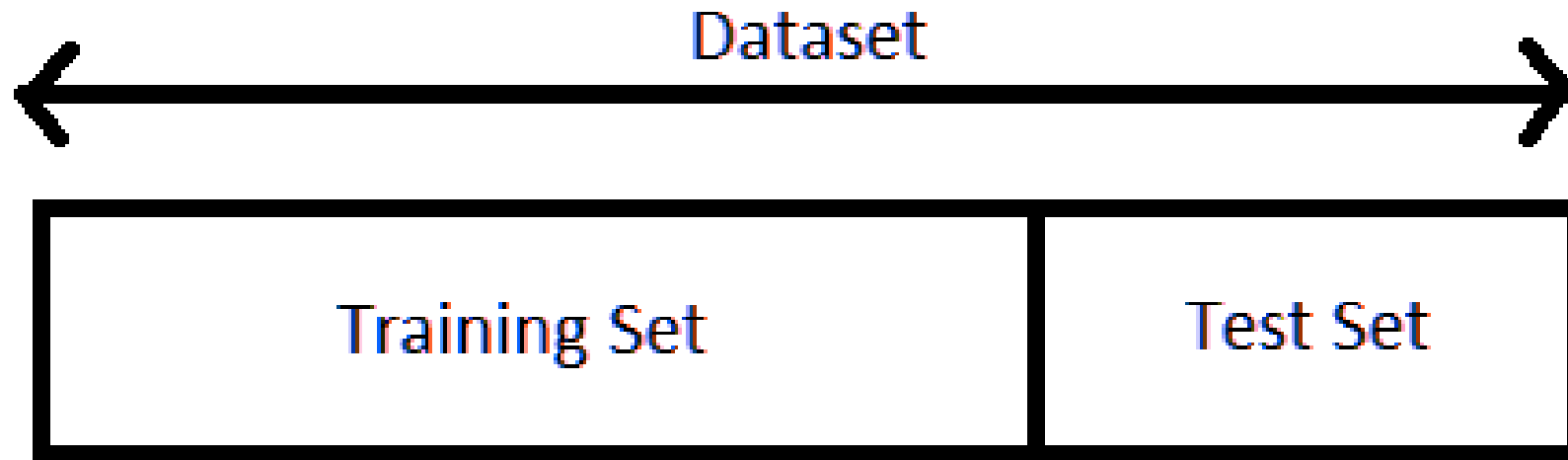
2022년 1분기	2022년 2분기	2022년 3분기	2022년 4분기
2021년 1분기	2021년 2분기	2021년 3분기	2021년 4분기
2020년 1분기	2020년 2분기	2020년 3분기	2020년 4분기
2019년 1분기	2019년 2분기	2019년 3분기	2019년 4분기

모델학습 데이터

모델 평가 (테스트)분기





예측 분기

# 모델 학습 데이터의 준비



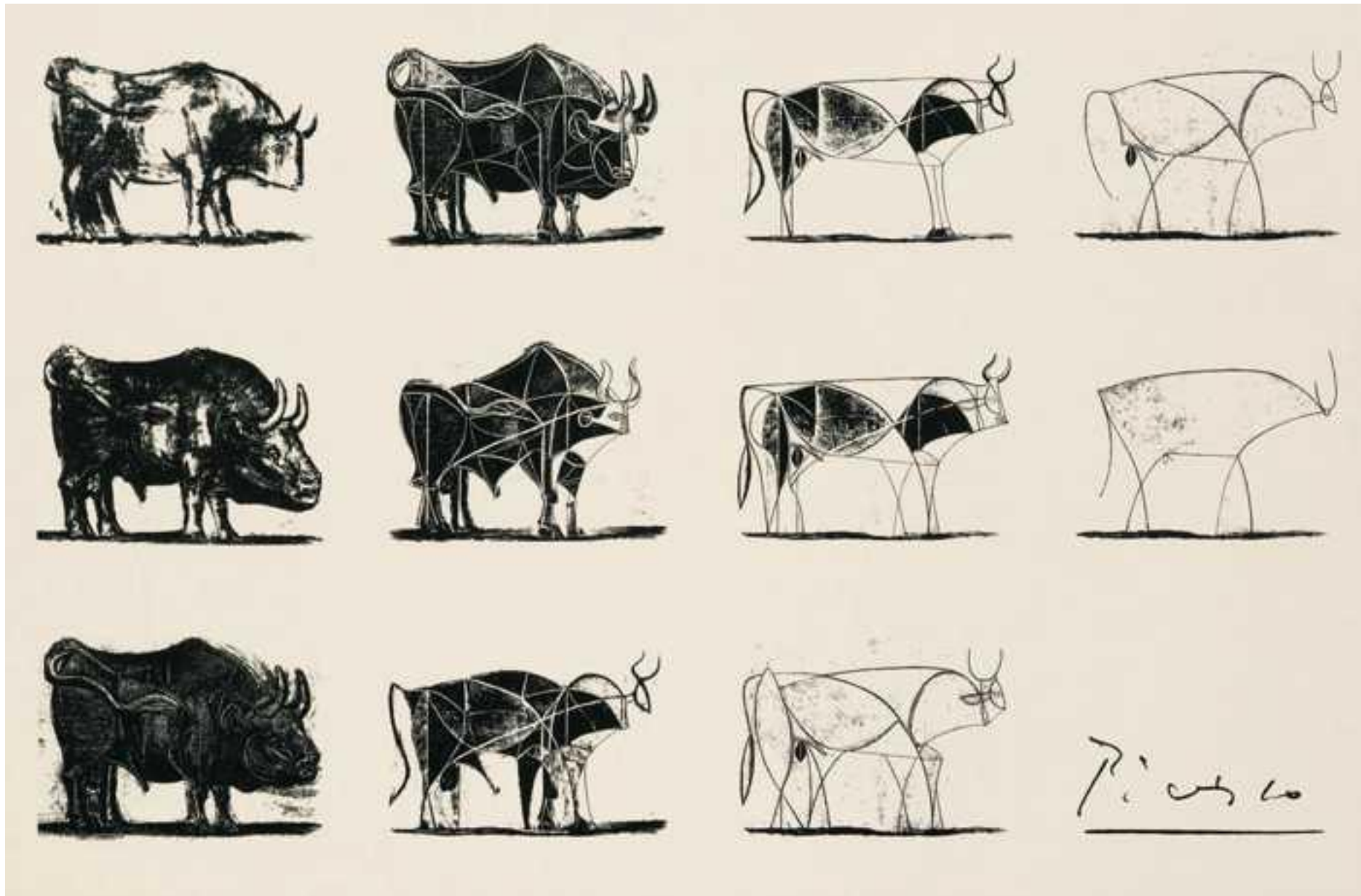
# Confusion Matrix for Binary Classification

		True Class		1
		Positive	Negative	
2	Positive	TP	FP	
	Negative	FN	TN	
3	Predicted Class			4

		Actual Values	
		1	0
Predicted Values	1	<p><b>TRUE POSITIVE</b></p> 	<p><b>FALSE POSITIVE</b></p>  <p><b>TYPE 1 ERROR</b></p>
	0	<p><b>FALSE NEGATIVE</b></p>  <p><b>TYPE 2 ERROR</b></p>	<p><b>TRUE NEGATIVE</b></p> 

## 판별분석의 라벨(Label)이 3개인 경우

개 ->개	개 -> 고양이	개 -> 코끼리
고양이 -> 개	고양이 -> 고양이	고양이 -> 코끼리
		코끼리 -> 코끼리











보·이·는·대·로·**밀**·지·마·라!

대/반/전/ 음악추리쇼



Mnet tvN 공동 방송

10월 22일 목요일 | 밤 9시 40분

6명 중에는 음치도 있고 노래를 잘하는 실력자(정상)도 있다.

번호 [ 1, 2, 3, 4, 5, 6 ]

정답 : [음치, 음치, 음치, 음치, 정상, 정상] 가 있다.

누가 음치인지 겉모습만 보고 맞춰야 한다.

감으로 예측을 한다.

예측 : [음치, 음치, 정상, 정상, 정상, 정상]

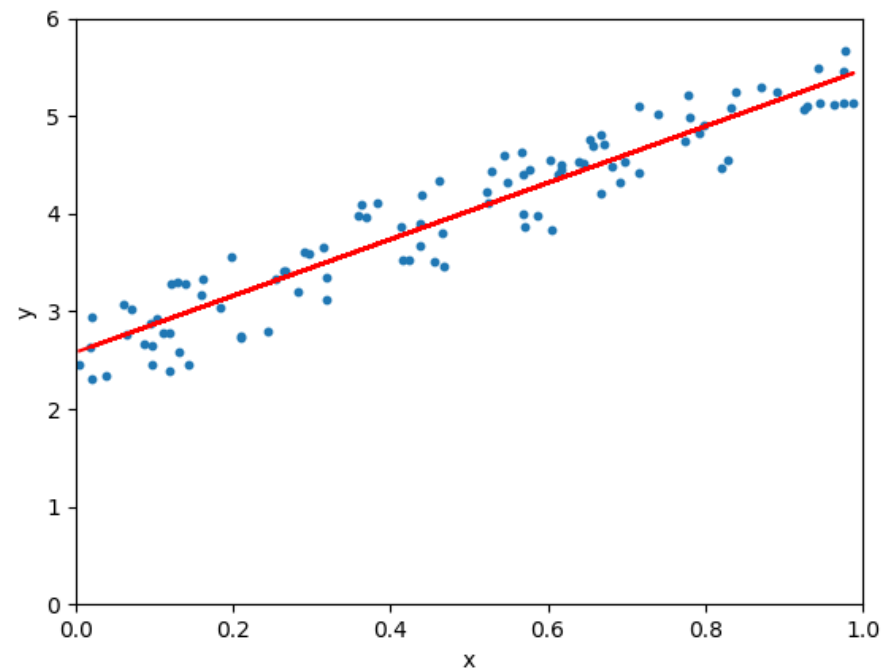
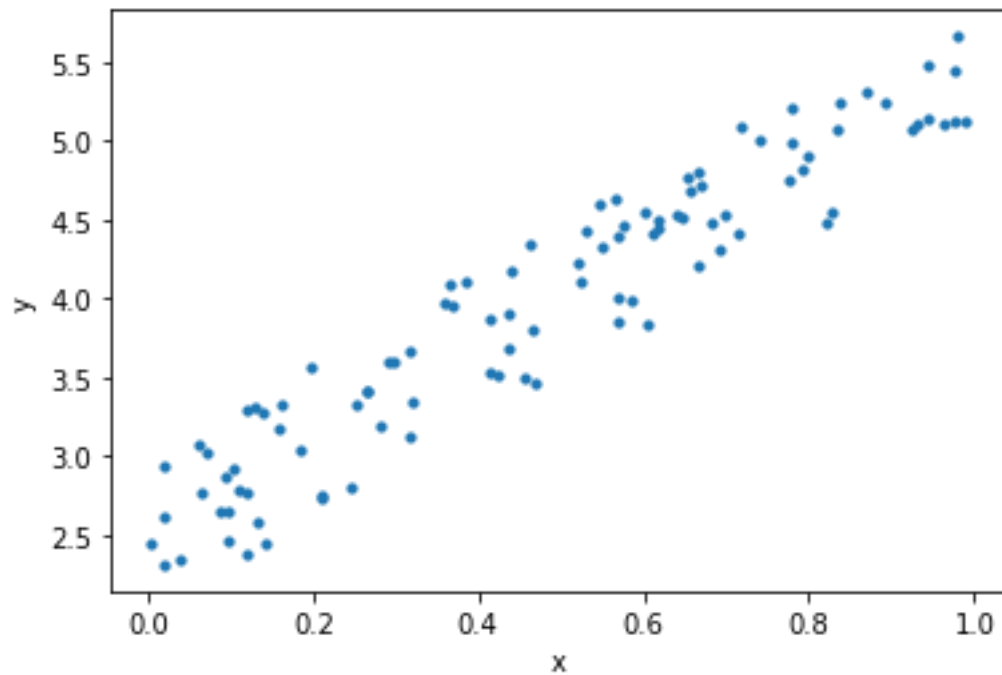
다음 항목들도 모델을 평가하는 데 매우 중요한 지표로 활용된다.

■ Precision(정밀도) – 모델이 비정상으로 분류한 스마트카 중에서 실제 비정상인 스마트카 비율

■ Recall(재현율) – 실제 비정상인 스마트카 중에서 모델이 비정상 스마트카로 분류한 비율

■ F1-Score – Precision과 Recall의 조화 평균

# 회기(Regression) 예측분석



# 회기(Regression) 예측분석(예: 집가격 예측)

## Housing Prices



House 1

1 room

\$150K



House 2

2 rooms

\$200K



House 3

3 rooms

???

\$250K



House 4

4 rooms

\$300K

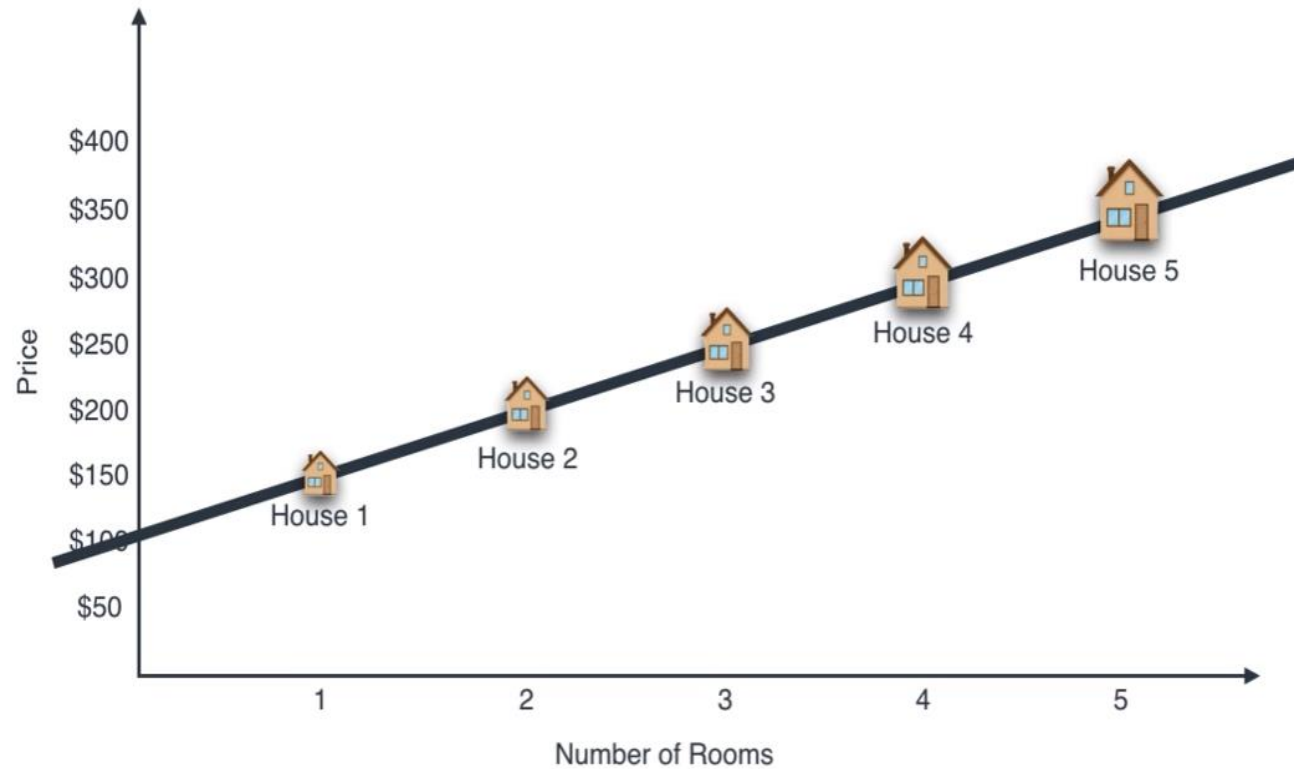


House 5

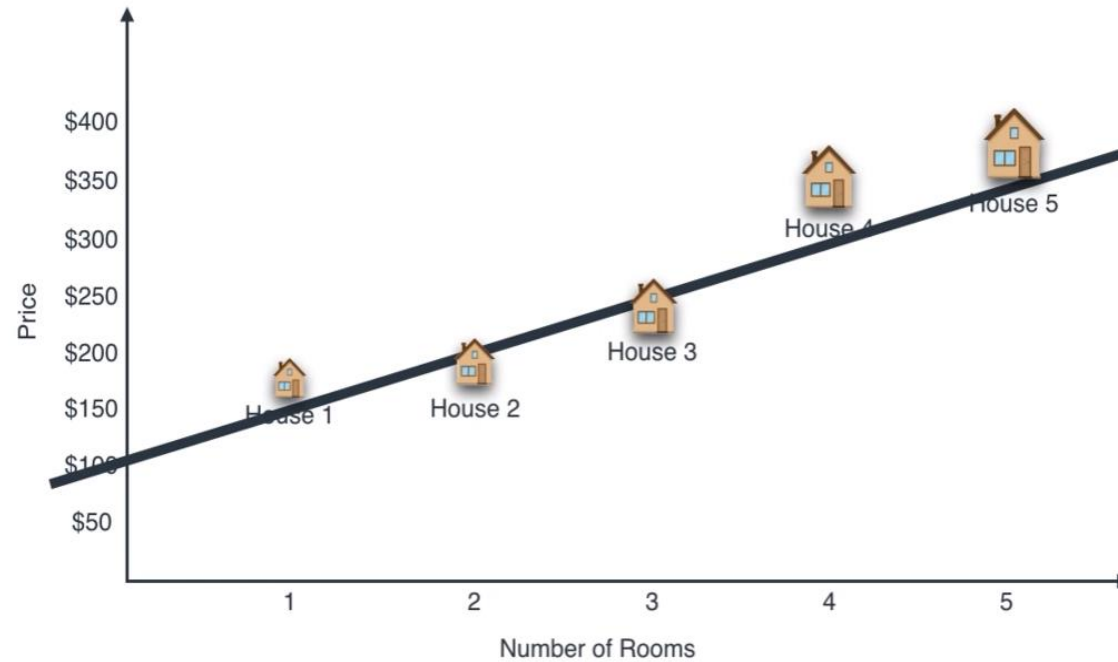
5 rooms

\$350K

# 회기(Regression) 예측분석(예: 집가격 예측)



# 회기(Regression) 예측분석(예: 집가격 예측)



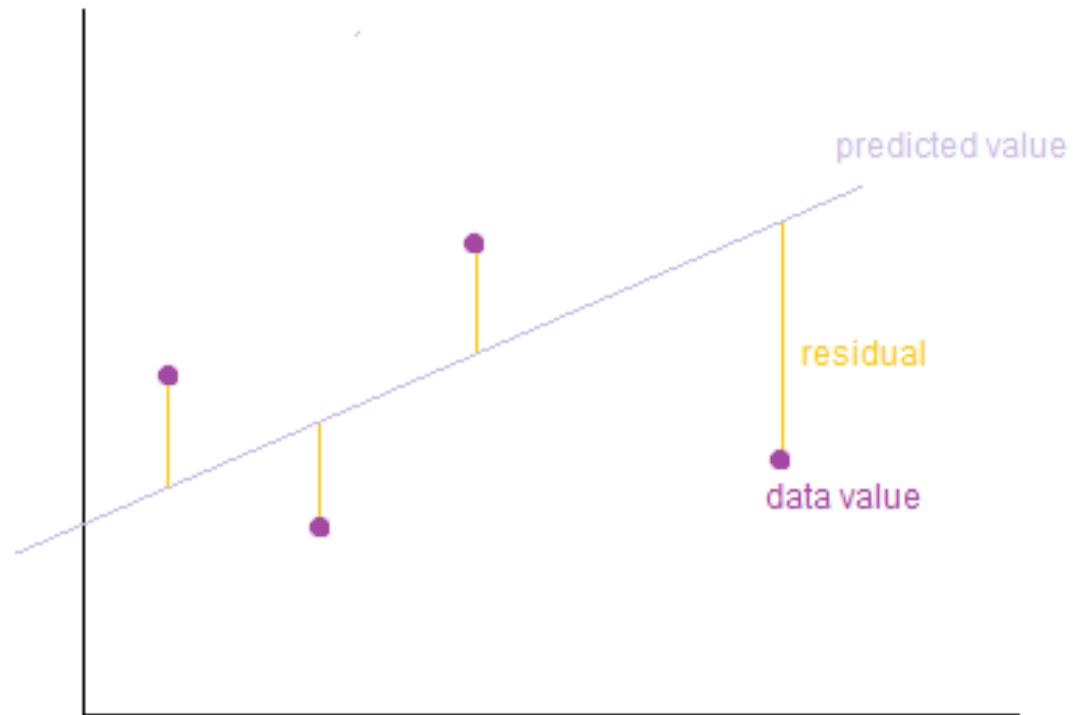
# 회기(Regression) 예측분석의 Metrics

## Error metrics for regression predictive modeling

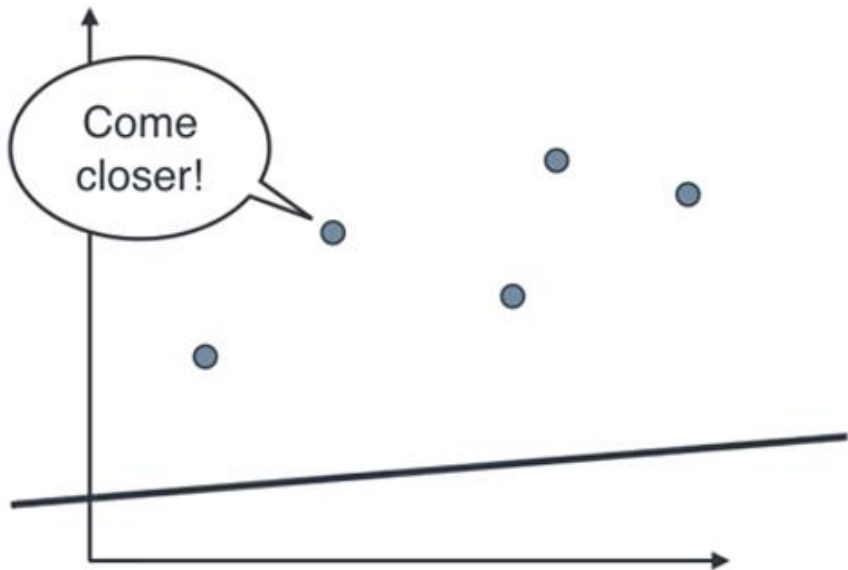
- Regression predictive modeling are those problems that involve predicting a numeric value.
- Metrics for regression involve calculating an error score to summarize the predictive skill of a model.
- How to calculate and report mean squared error, root mean squared error, and mean absolute error.



# 회기(Regression) 예측 분석 모형의 성능 평가



# Linear regression algorithm



**Step 1:** Start with a random line

**Step 2:** Pick a large number. **1000**  
(number of repetitions, or epochs)

**Step 3:** Pick a small number. **0.01**  
(learning rate)

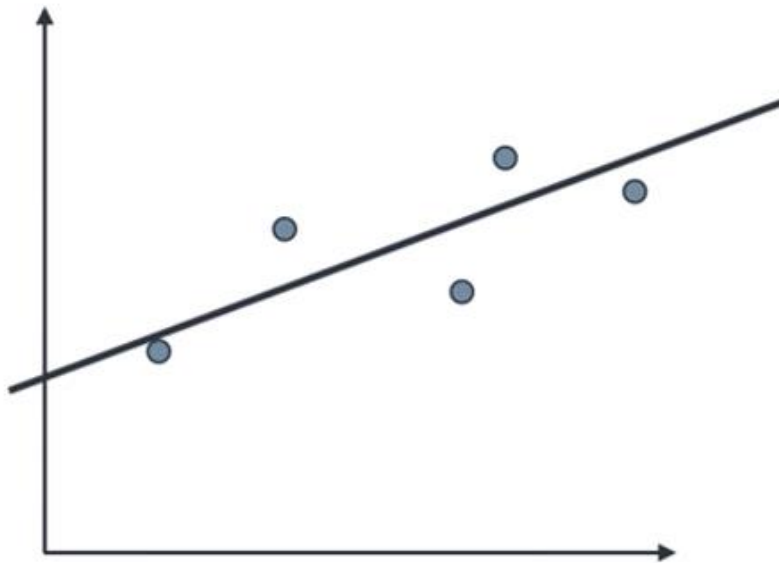
**Step 4:** (repeat **1000** times)

- Pick random point
- Add  $(\text{learning rate}) \times (\text{vertical distance}) \times (\text{horizontal distance})$  to **slope**
- Add  $(\text{learning rate}) \times (\text{vertical distance})$  to **y-intercept**



<https://www.youtube.com/watch?v=wYPUhge9w5c>

# Linear regression algorithm



**Step 1:** Start with a random line

**Step 2:** Pick a large number. **1000**  
(number of repetitions, or epochs)

**Step 3:** Pick a small number. **0.01**  
(learning rate)

**Step 4:** (repeat **1000** times)

- Pick random point
- Add  $(\text{learning rate}) \times (\text{vertical distance}) \times (\text{horizontal distance})$  to **slope**
- Add  $(\text{learning rate}) \times (\text{vertical distance})$  to **y-intercept**

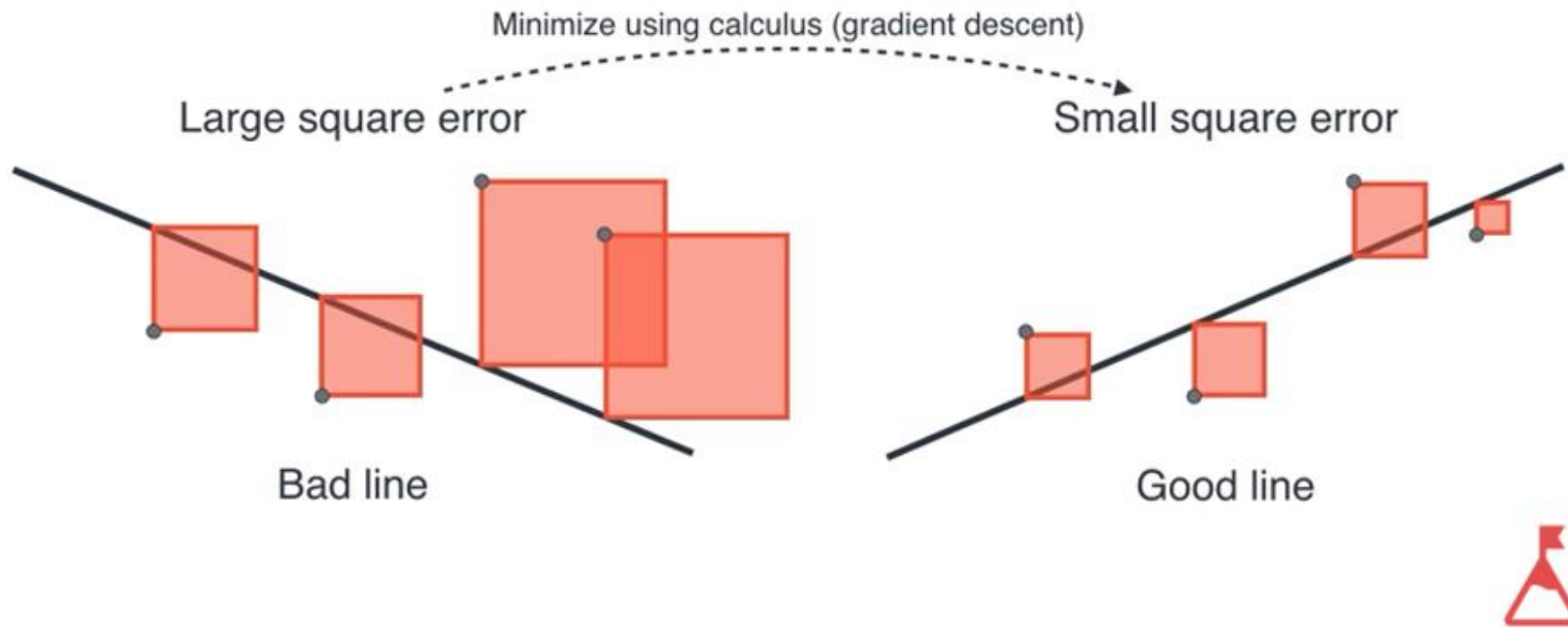


<https://www.youtube.com/watch?v=wYPUhge9w5c>

# Square error

Same as the square trick!

Minimize using calculus (gradient descent)

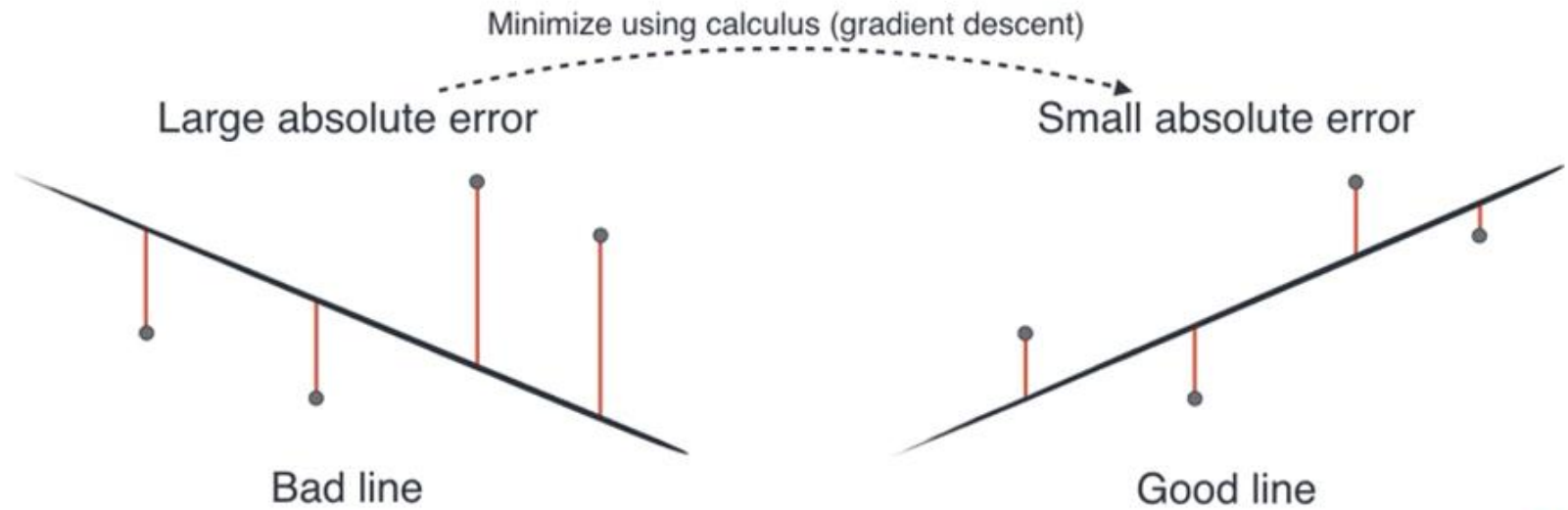


<https://www.youtube.com/watch?v=wYPUhge9w5c>

# Absolute error

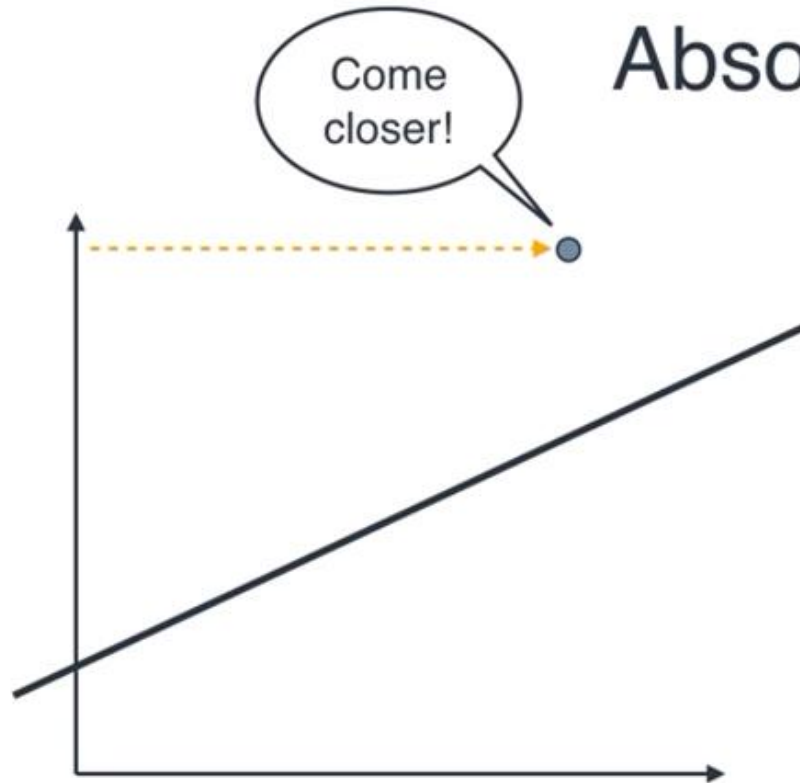
Develop an absolute trick!

Minimize using calculus (gradient descent)



<https://www.youtube.com/watch?v=wYPUhge9w5c>

# Absolute trick



## Step 1:

Pick a small number (learning rate)

## Step 2:

- If point is above the line:

- Add (learning rate) x (horizontal distance) to slope
- Add (learning rate) to y-intercept

- If point is below the line:

- Subtract (learning rate) x (horizontal distance) to slope
- Subtract (learning rate) to y-intercept



<https://www.youtube.com/watch?v=wYPUhge9w5c>

정 준 수 / Ph.D ( jsjeong@hansung.ac.kr )

- 前) 삼성전자 연구원
- 前) 삼성의료원 (삼성생명과학연구소)
- 前) 삼성SDS (정보기술연구소)
- 現) (사)한국인공지능협회, AI, 머신러닝 강의
- 現) 한국소프트웨어산업협회, AI, 머신러닝 강의
- 現) 서울디지털재단, AI 자문위원
- 現) 한성대학교 교수(겸)
- 전문분야: Computer Vision, 머신러닝(ML), RPA
- <https://github.com/JSJeong-me/>

