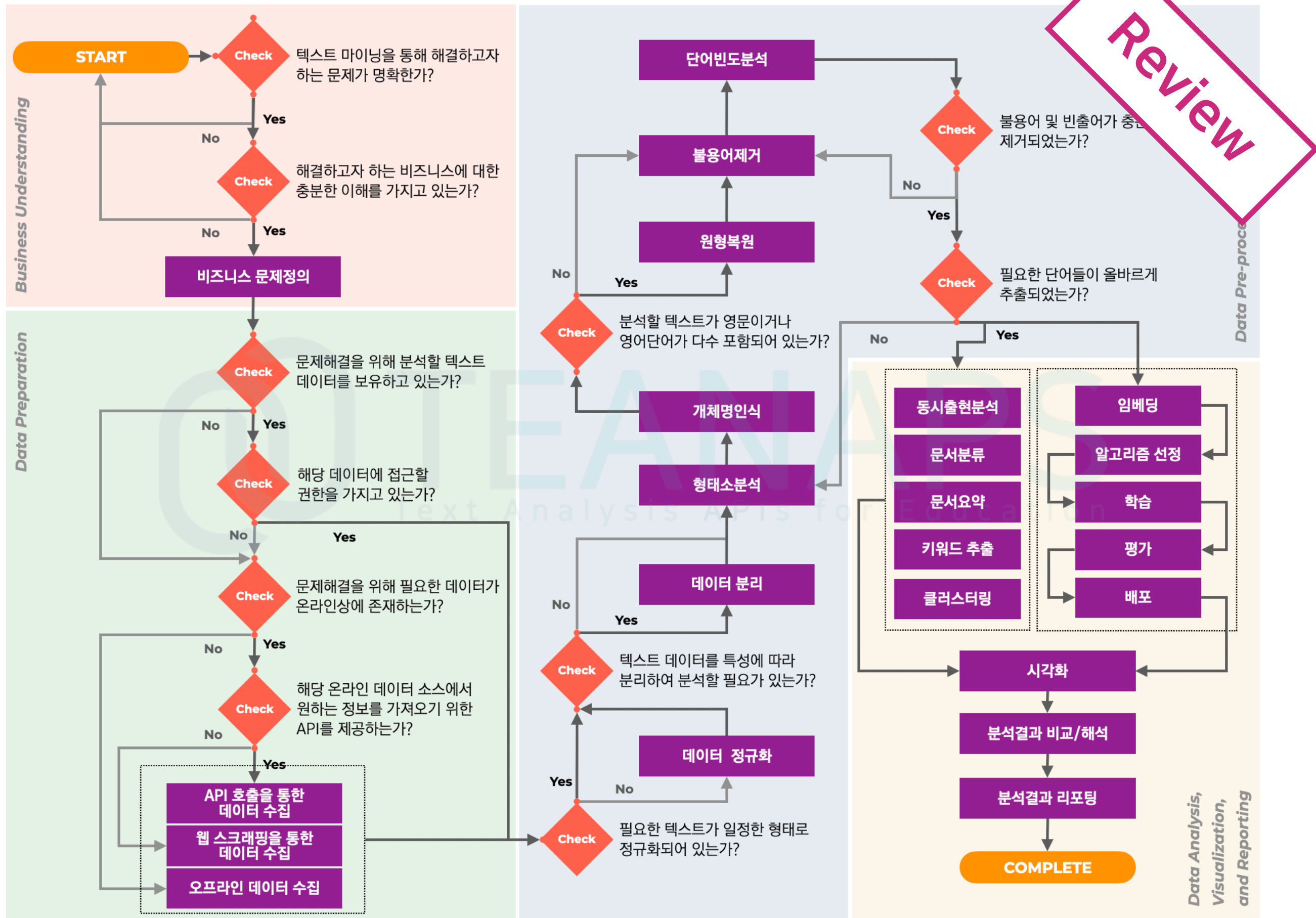


# ADVANCED TEXT MINING

---

by FINGEREDMAN (fingeredman@gmail.com)



WEEK 08

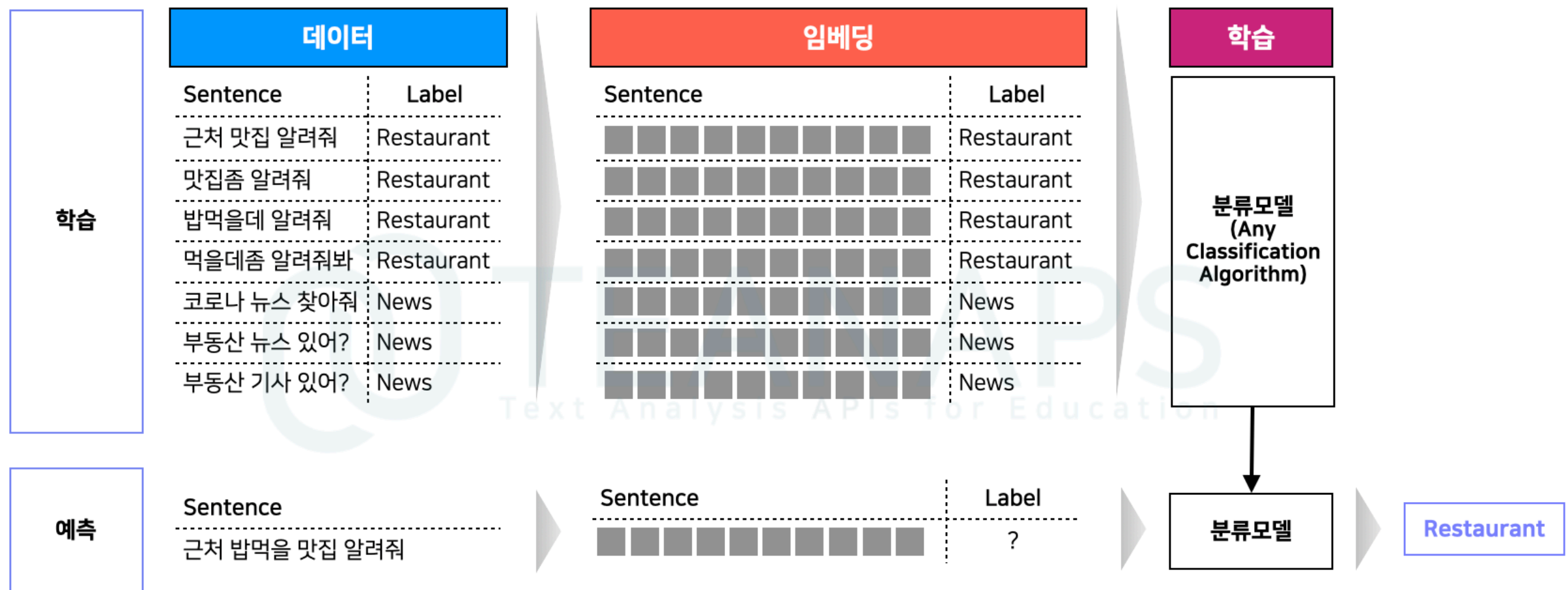
# Classification



# 텍스트를 분류하는 방법

## 텍스트 분류 (Text Classification)

- 문서(문장, 문단 등)를 입력으로 받아 사전에 정의된 클래스(class) 중에 어디에 속하는지 분류하는 과정
- **클래스** (범주, class) : 불연속적인 값(discrete data)으로 정의된 분류단위
- 나이브 베이즈(Naive Bayes), SVM 등 머신러닝부터 RNN, CNN 등 딥러닝으로도 문제해결이 가능함
- 활용 예 : 감성분석(긍정, 중립, 부정), 스팸탐지(스팸, 비스팸), 챗봇 의도분류, 뉴스기사 주제분류 등



# 기계학습 절차: 학습 (Training)

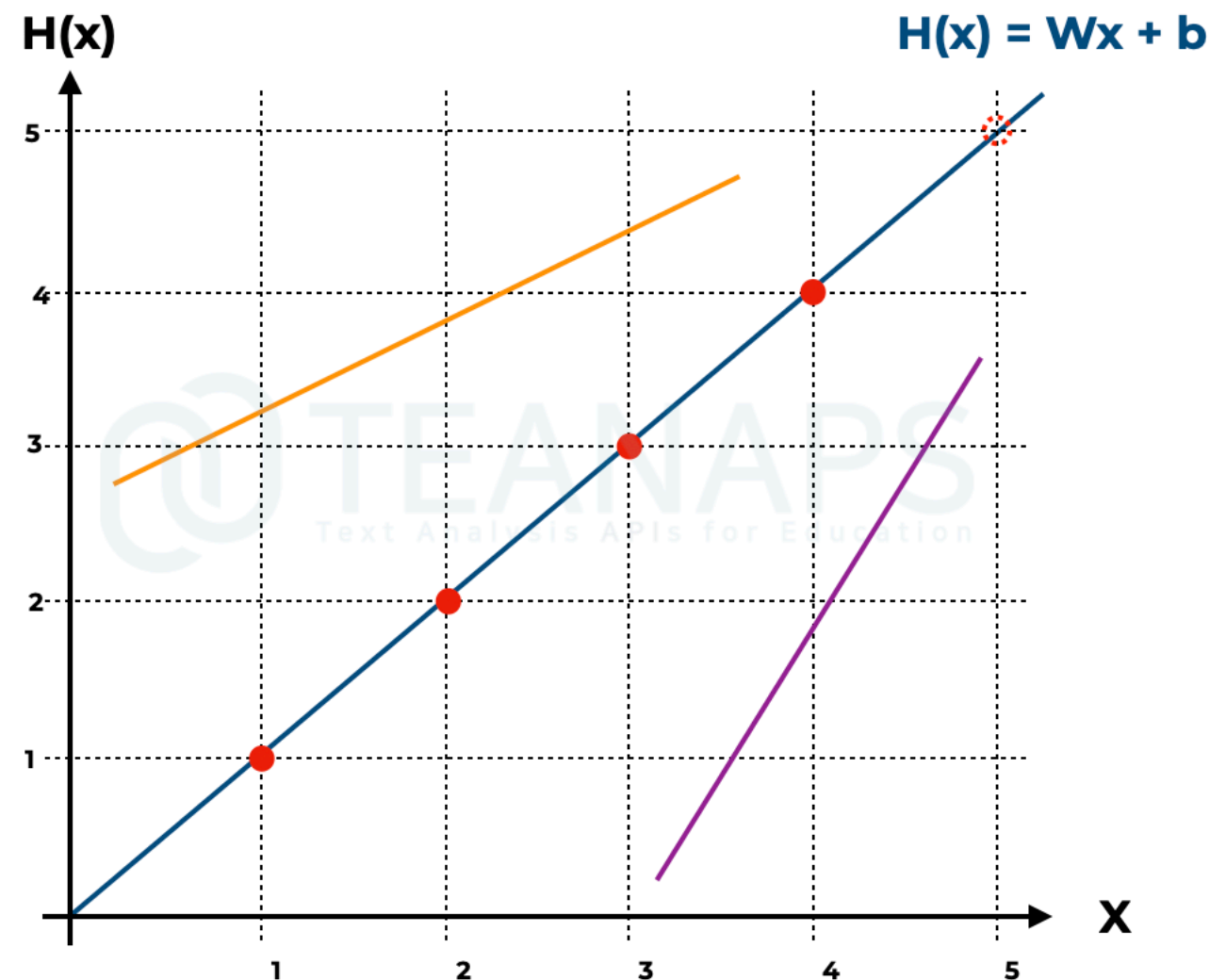
Review

## 기계가 데이터를 학습하는 과정 (Machine Training)

- 학습데이터에 정의된 정보나 규칙을 추상적인 형태로 표현하는 모델을 생성하는 과정
- 학습데이터에 포함된 다양한 정보나 규칙을 모델이 얼마나 잘 표현하는가에 따라 머신러닝 모델의 성능이 좌우됨
- **선형가정 (Linear Hypothesis)** : 학습데이터의 분포를 선형이라 가정하고 학습데이터를 가장 잘 설명한 직선



X	Y
1	1
2	2
3	3
4	4
5	y



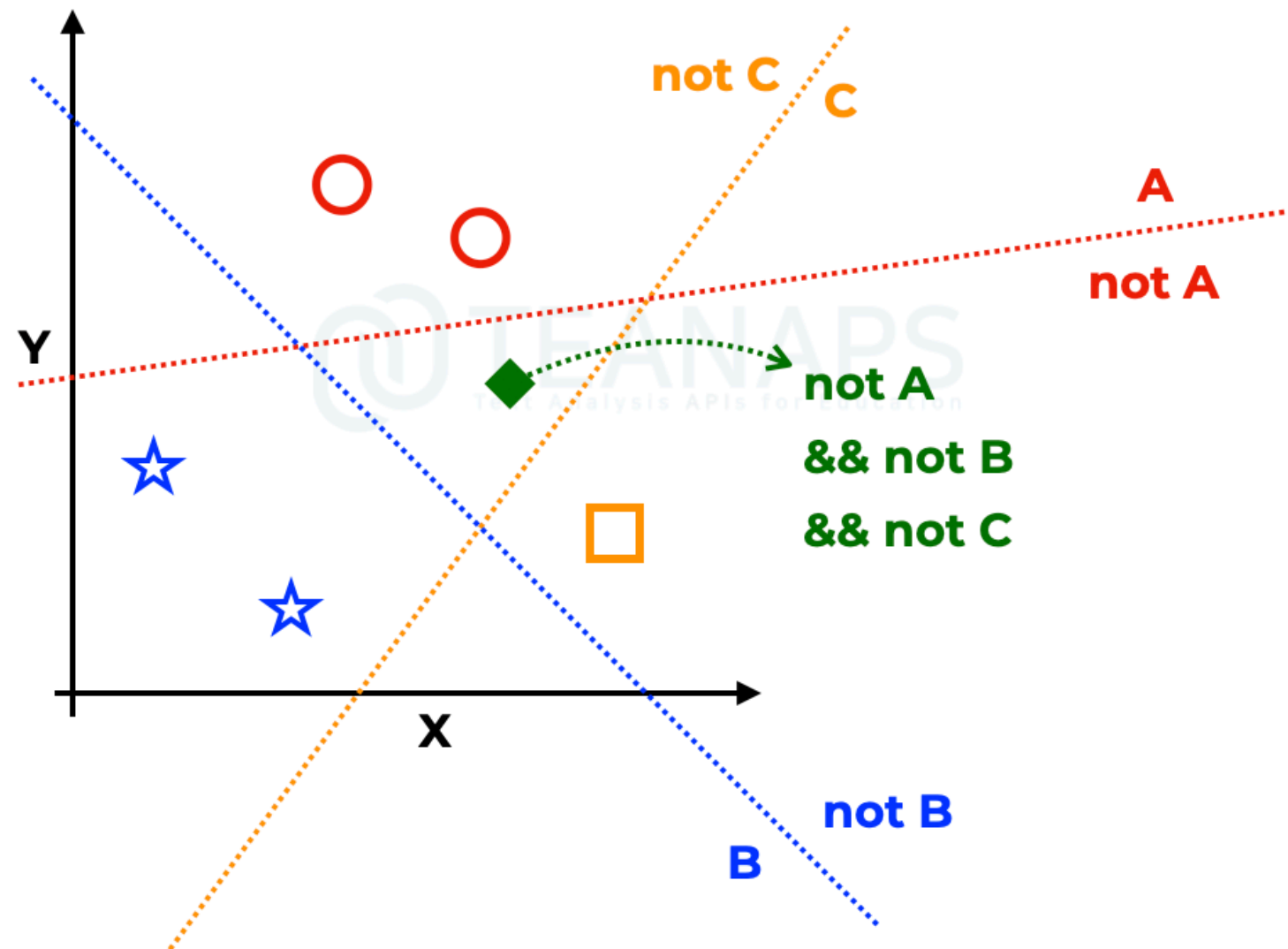
# 기계학습 절차: 학습 (Training)

Review

## 다차원 분류 (Multinomial Classification)

- 3개이상의 범주를 가지는 분류문제
- 다차원 분류 문제는 2진분류(binary classification) 문제를 해결하는 모델을 여러개 활용하여 해결 가능함

X <sub>1</sub>	X <sub>2</sub>	Y
10	5	A
9	5	A
3	2	B
2	4	B
11	1	C



\* 조태호, 모두의 딥러닝(개정판 2판), 길벗출판사.

\*\* references

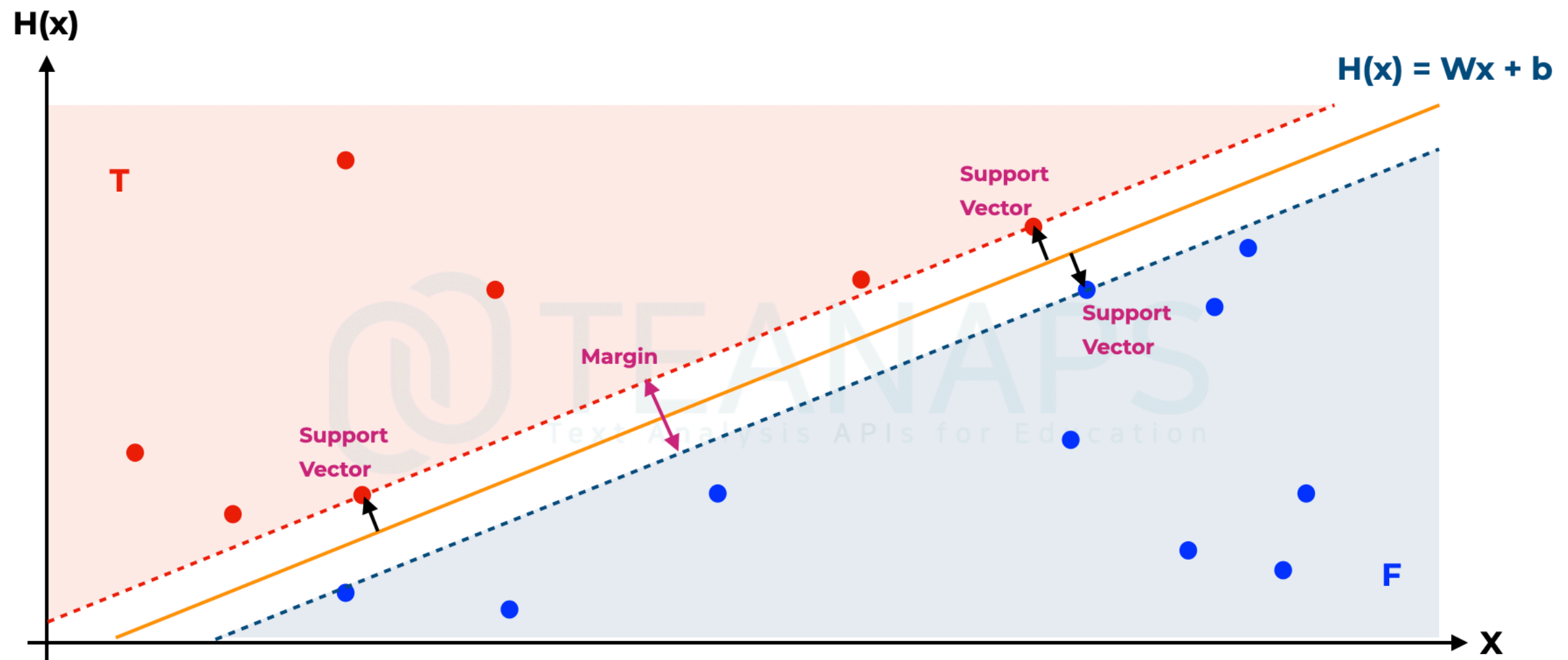
\*\*\* references

# 분류 알고리즘: 서포트 벡터 머신

Review

## 서포트 벡터 머신 (Support Vector Machine, SVM)

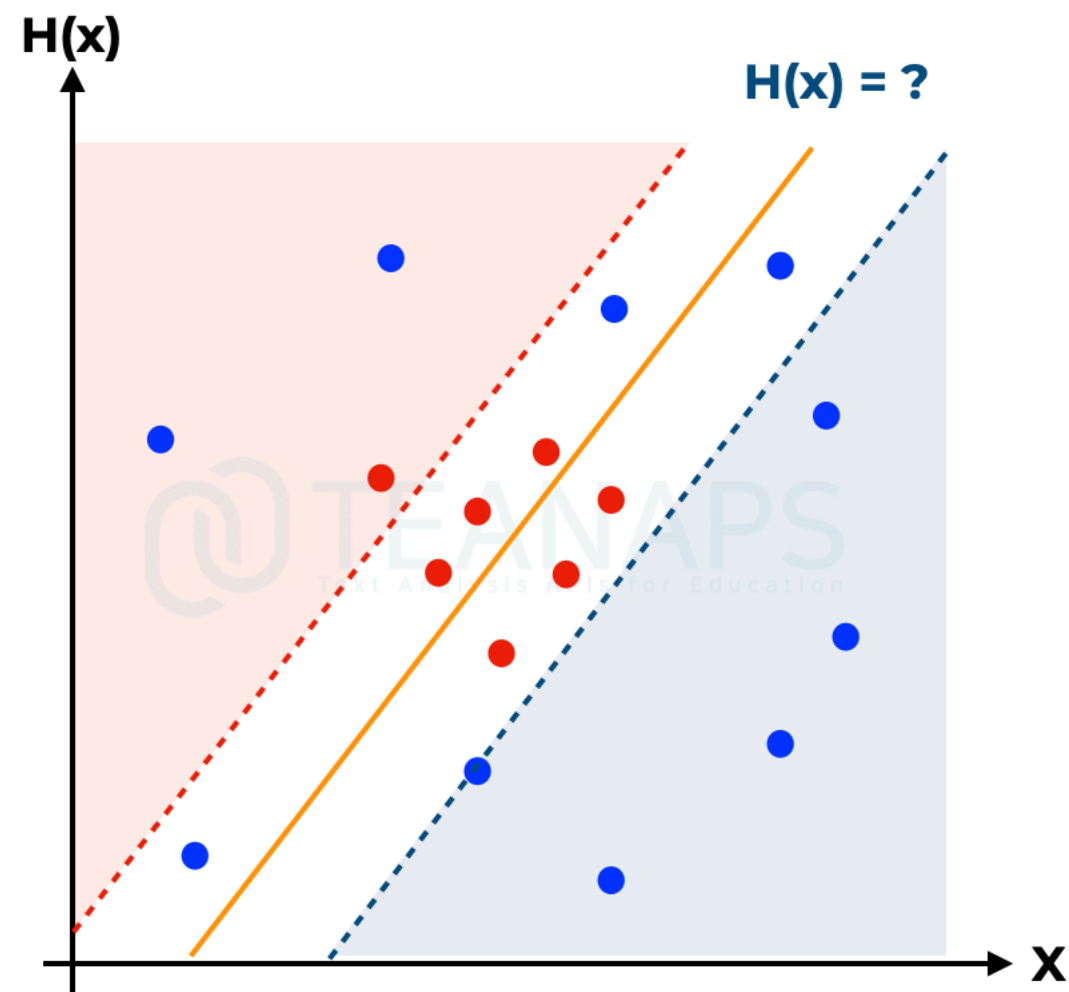
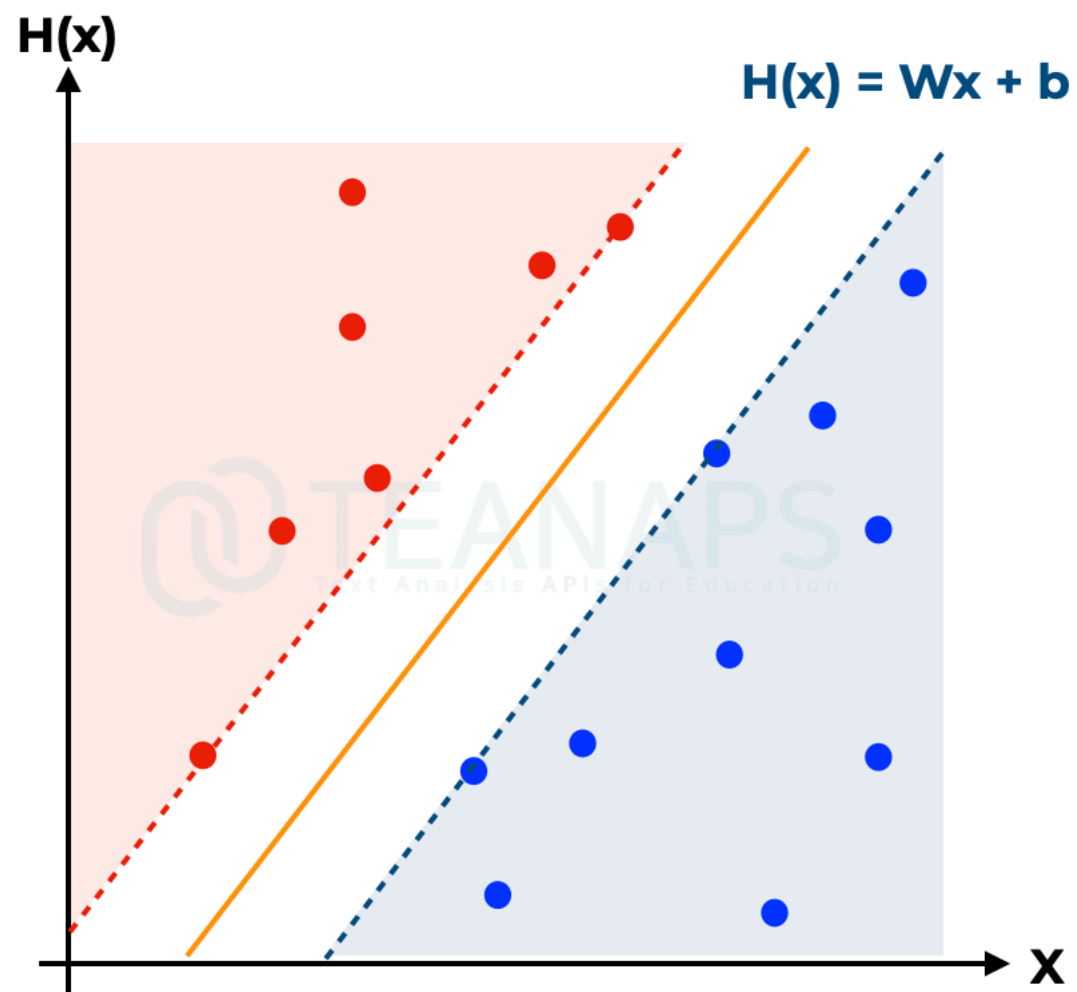
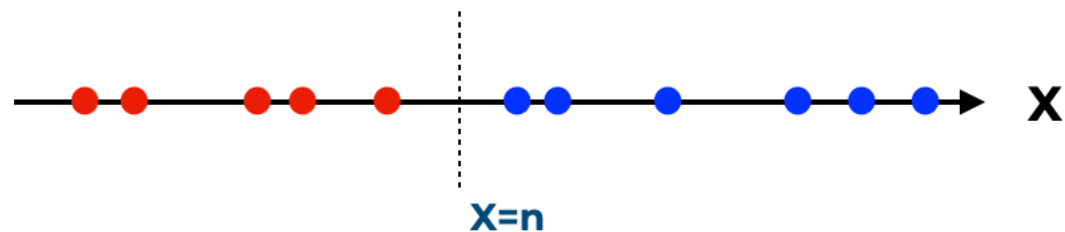
- 지지벡터로 이루어진 초평면과 마진을 최대화 하는 직선으로 선형 분류하는 기계학습 알고리즘
- 데이터를 선형함수로 분류할 수 없더라도 커널함수를 활용해 데이터를 고차원 공간으로 이동한 후 분류 가능함
- **지지벡터** (Support Vector) : 선형분류의 경계에 존재하는 데이터
- **커널함수** (kernel function) : 선형분류를 위해 데이터를 다른 차원으로 표현할 수 있도록 하는 함수



# 분류 알고리즘: 서포트 벡터 머신

Review

## 커널함수 (Kernel Function)

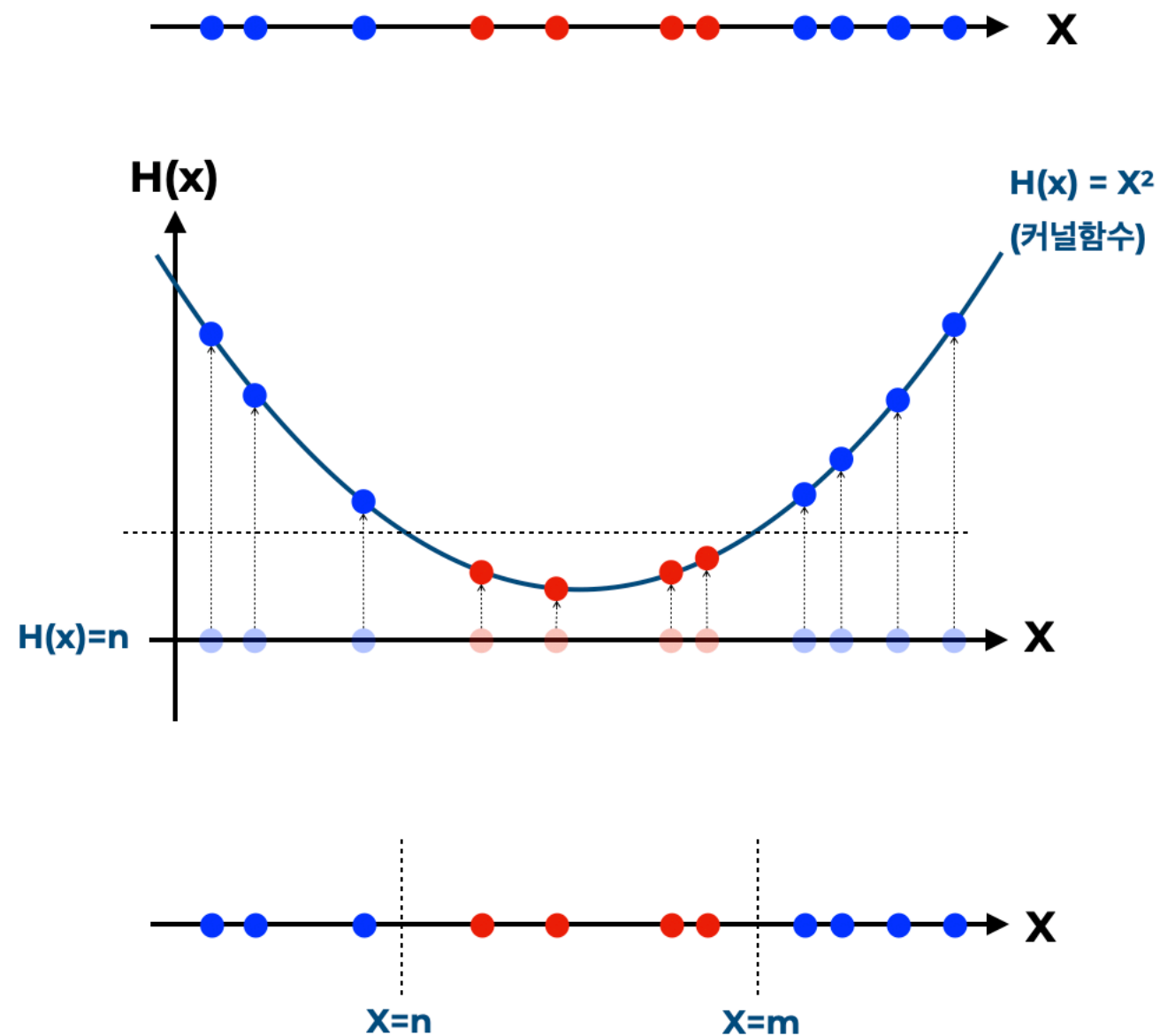
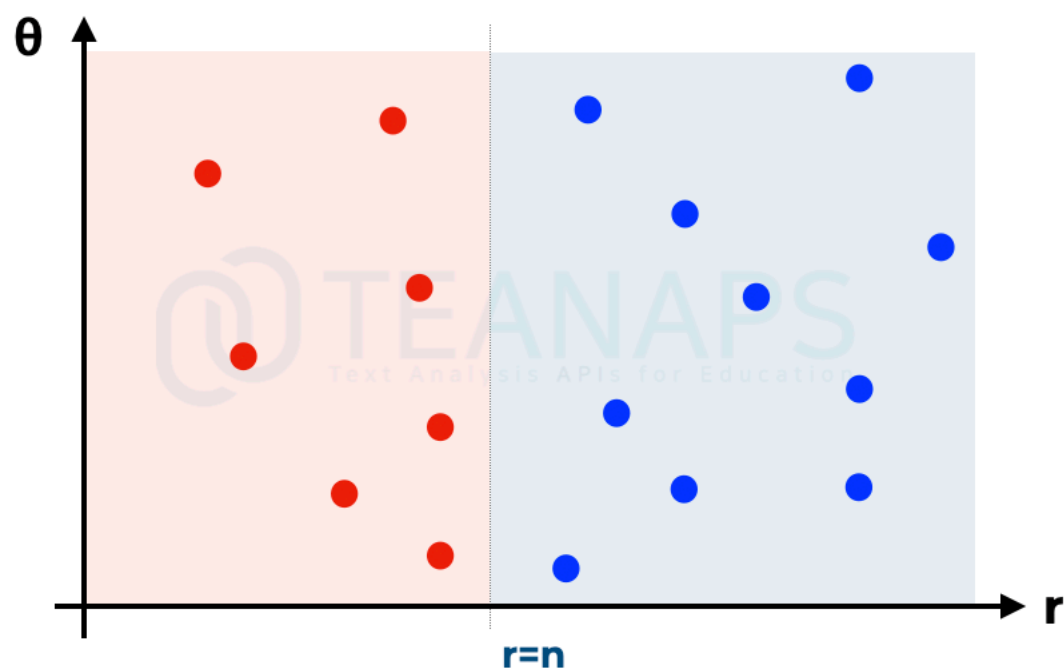
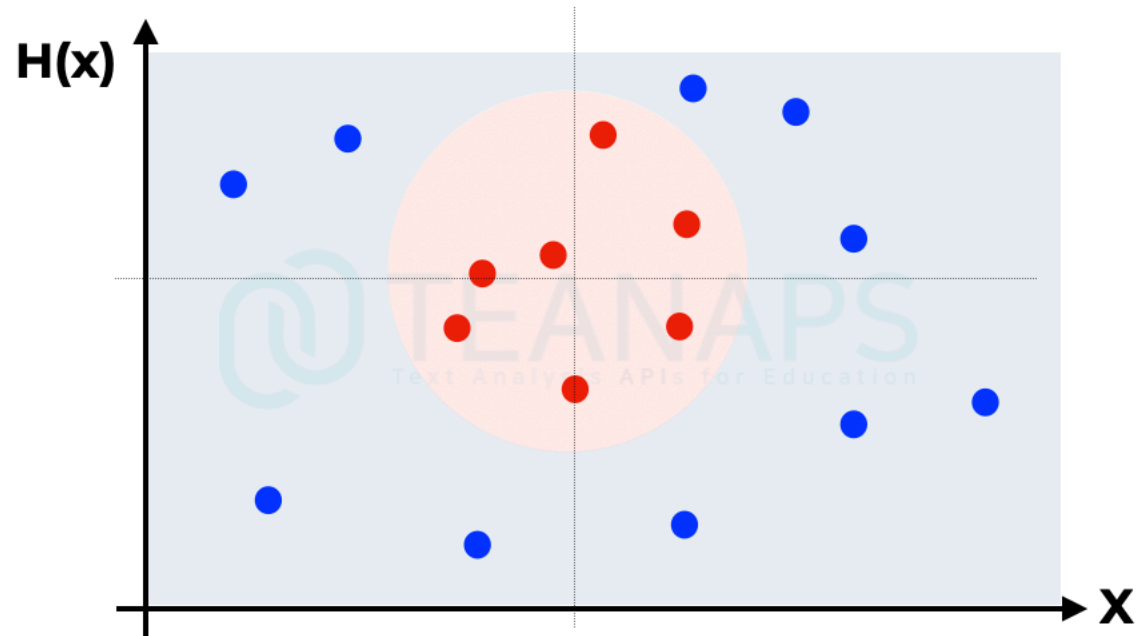




# 분류 알고리즘: 서포트 벡터 머신

Review

## 커널함수 (Kernel Function)



# 분류 알고리즘: 서포트 벡터 머신

Review

## 커널함수 (Kernel Function)

Linear :

$$K(x_1, x_2) = x_1^T x_2$$

Polynomial :

$$K(x_1, x_2) = (x_1^T x_2 + c)^d \quad (c > 0)$$

Sigmoid :

$$K(x_1, x_2) = \tanh(a(x_1^T x_2) + b) \quad (a, b \geq 0)$$

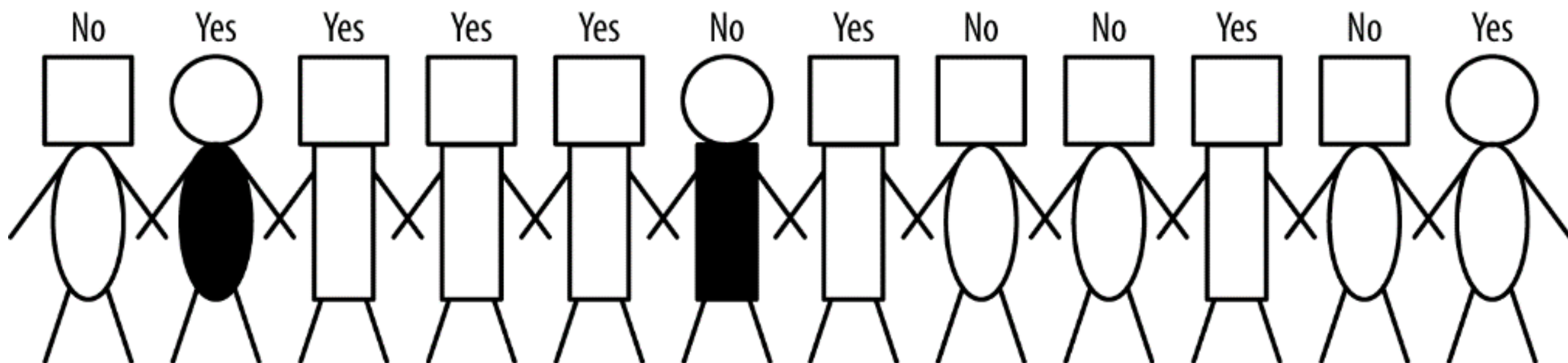
Gaussian :

$$K(x_1, x_2) = \exp \left\{ -\frac{\|x_1 - x_2\|_2^2}{2\sigma^2} \right\} \quad (a, b \geq 0)$$

# 분류 알고리즘: 분류트리

## 분류트리 (Decision Tree)

- 데이터를 주어진 자질(feature)에 따른 의사결정 기준에 따라 분류할 수 있는 규칙을 찾아내는 분류 알고리즘
- 가장 빠르고 간편한 머신러닝 알고리즘 중에 하나로 수치형 데이터와 범주형 데이터 모두를 자질로 활용 가능
- 전처리 과정이 거의 필요하지 않고 모델이 출력을 결정하는 과정을 직관적으로 확인 가능함(Non-blackbox Algorithm)



### 자질 (feature)

1. 머리 모양 : 네모, 동그라미
2. 몸 모양 : 네모, 동그라미
3. 몸 색상 : 검정색, 흰색



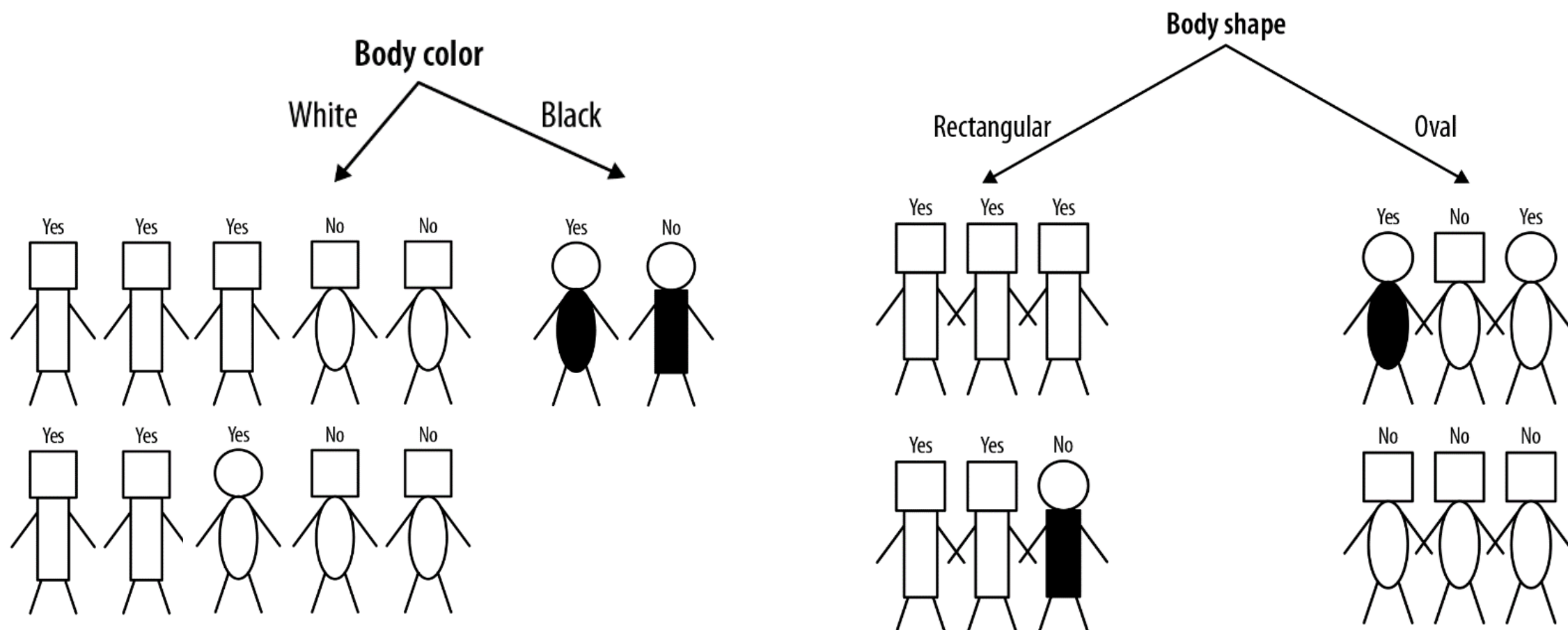
### 분류하고자 하는 범주 (class)

감염병 증상여부 : YES, NO

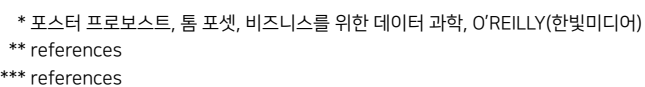
# 분류 알고리즘: 분류트리

## 분류트리 (Decision Tree)

- **엔트로피 (Entropy)** : 변수에 대한 분할을 평가하는 순수도 척도(무질서도)로, 노드에서 포함된 모든 클래스에 대하여 특정 클래스의 레코드의 비율과 이 값에 로그를 취한 값을 곱의 합으로 표현함
- **정보 증가량 (Information Gain)** : 새로 추가된 정보에 따른 엔트로피의 변화량으로, 부모가 자식보다 얼마나 더 순수한지(엔트로피가 감소하는지) 측정하는 척도

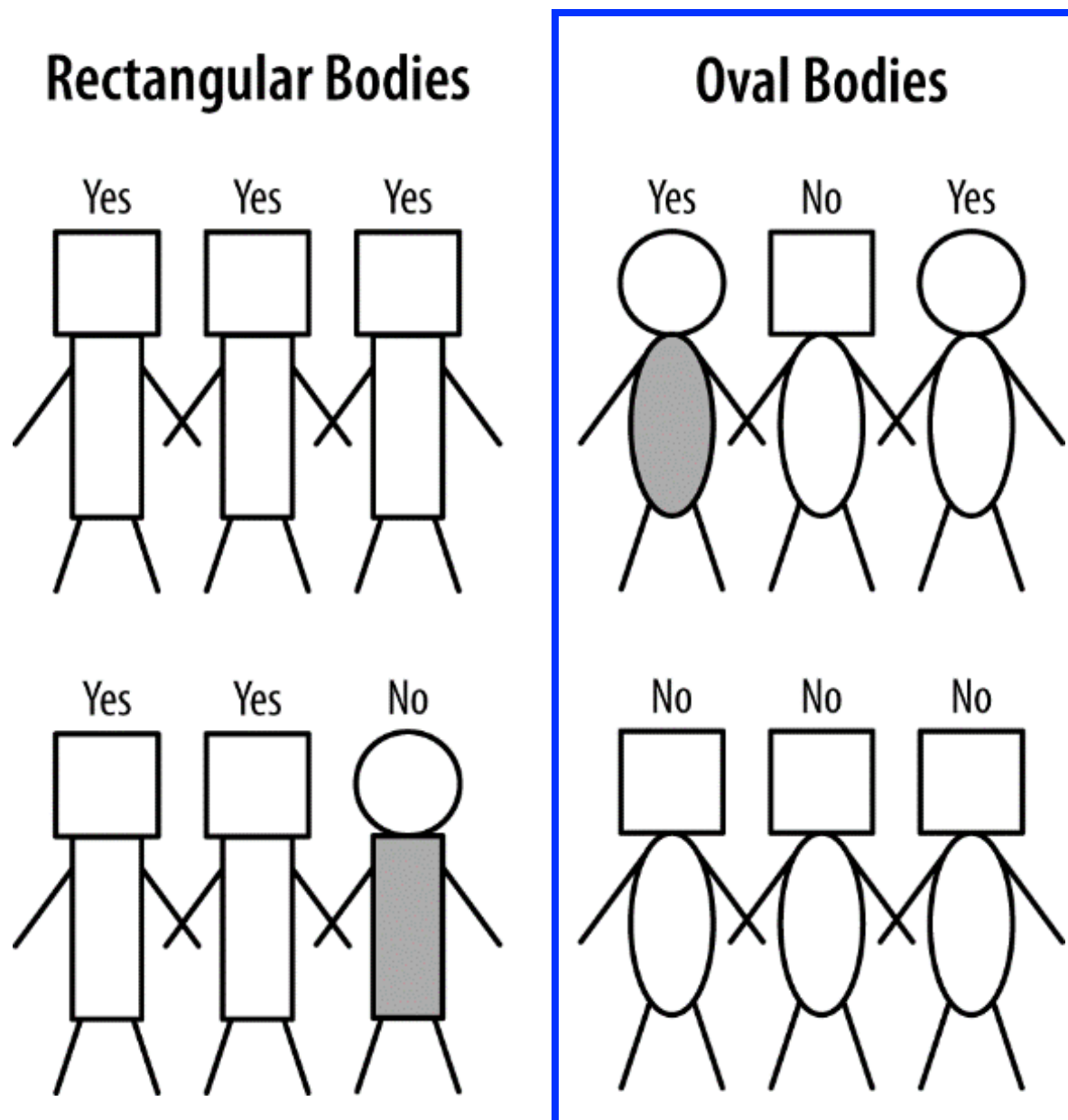


## 분류트리 (Decision Tree)

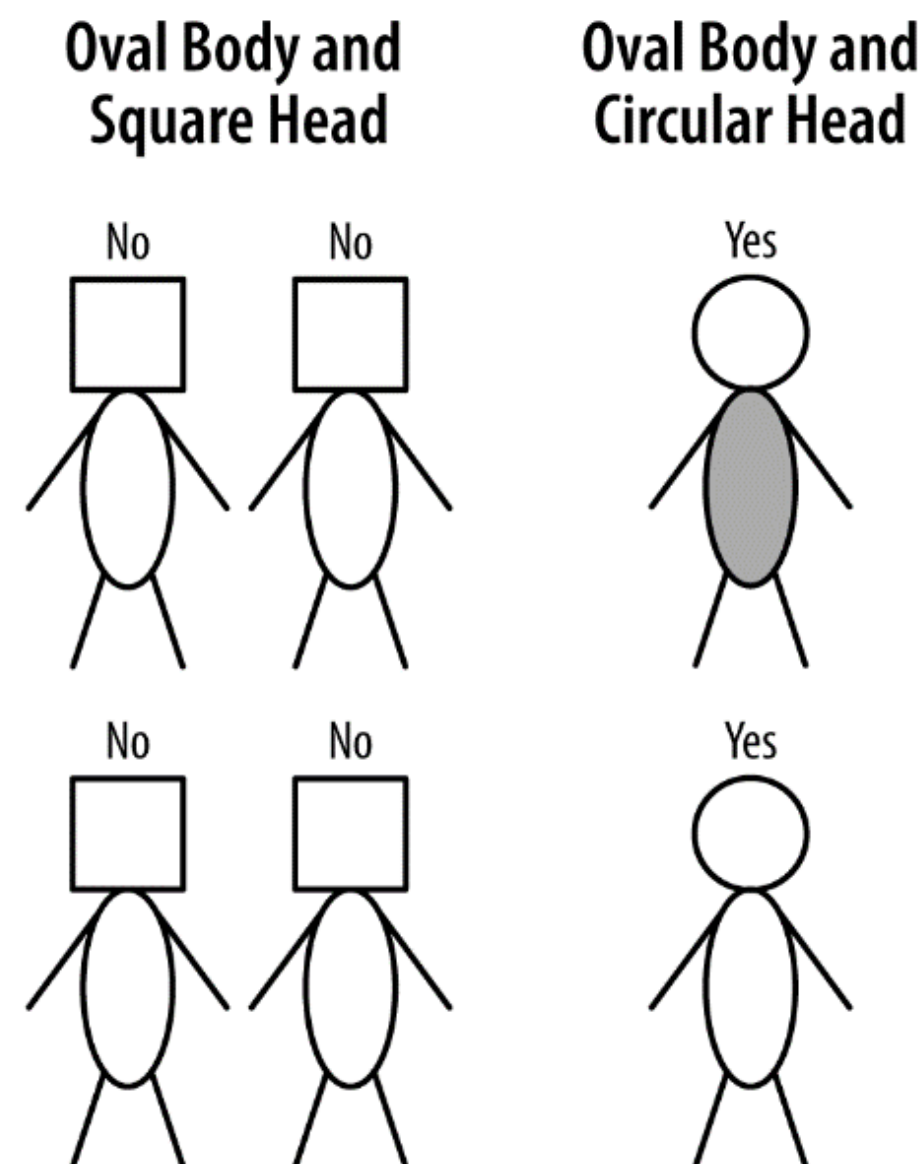


# 분류 알고리즘: 분류트리

## 1. 몸의 모양에 따라서 분류



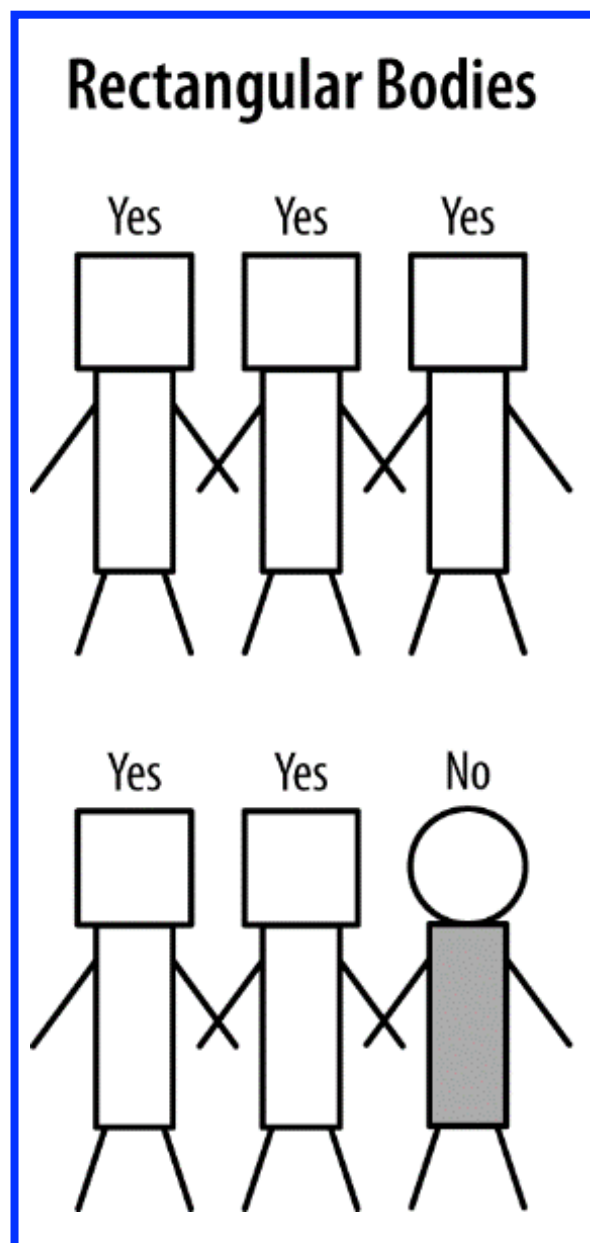
## 2-1. 머리 모양에 따라서 분류



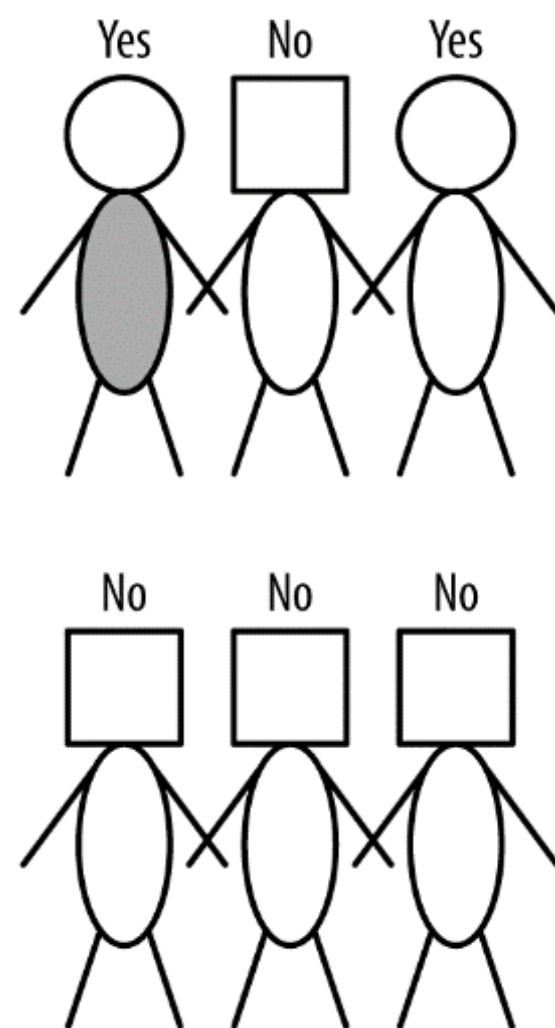


# 분류 알고리즘: 분류트리

## 1. 몸의 모양에 따라서 분류

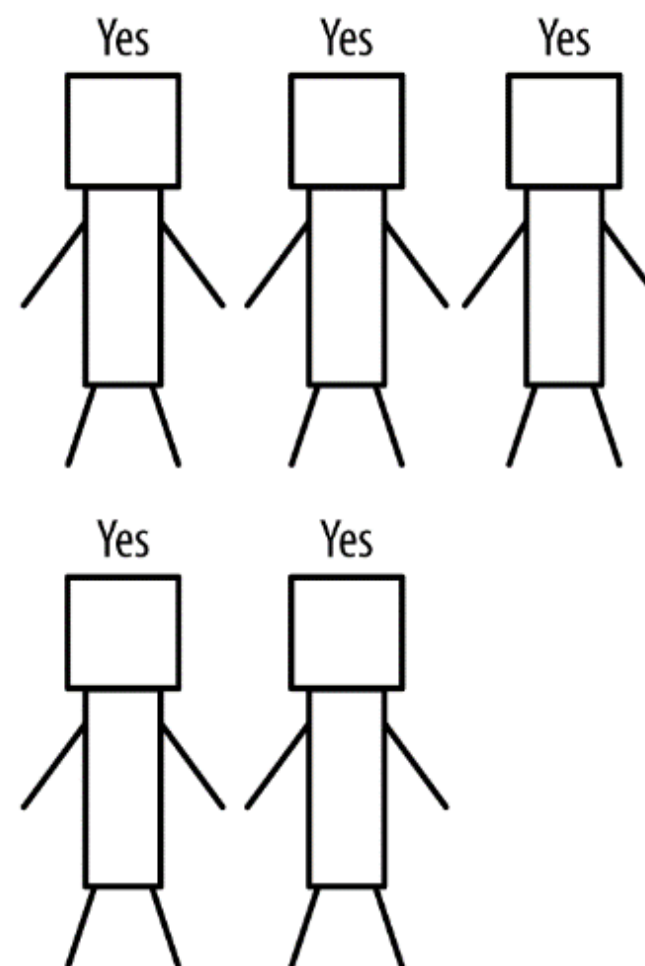


## Oval Bodies



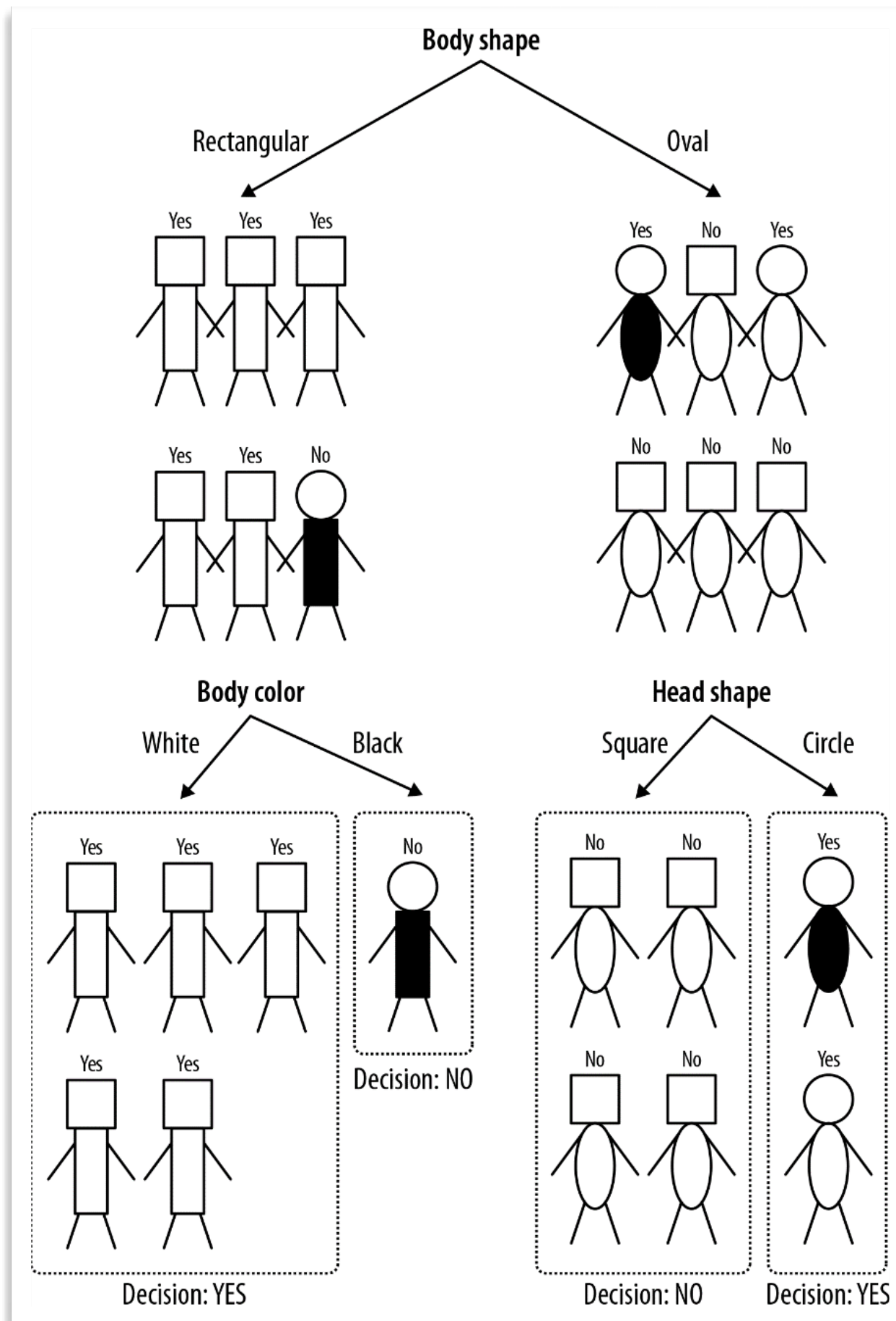
## 2-2. 몸의 색깔에 따라서 분류

### Rectangular Body and White



### Rectangular Body and Gray



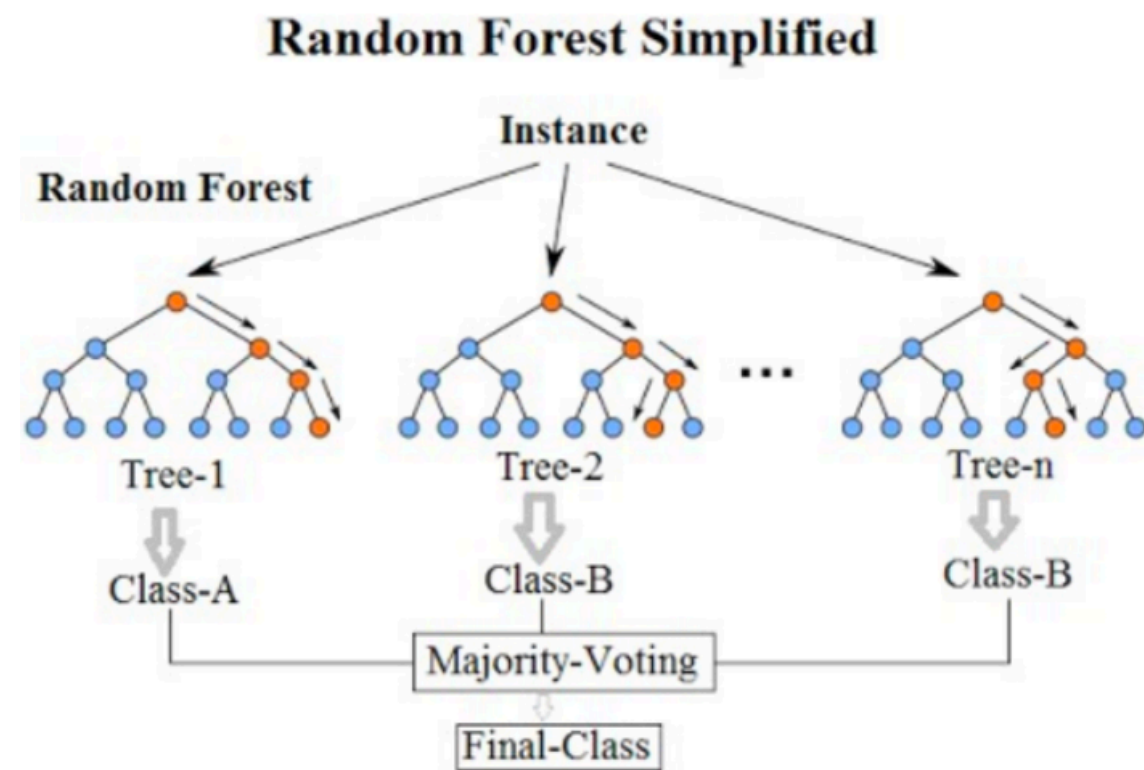
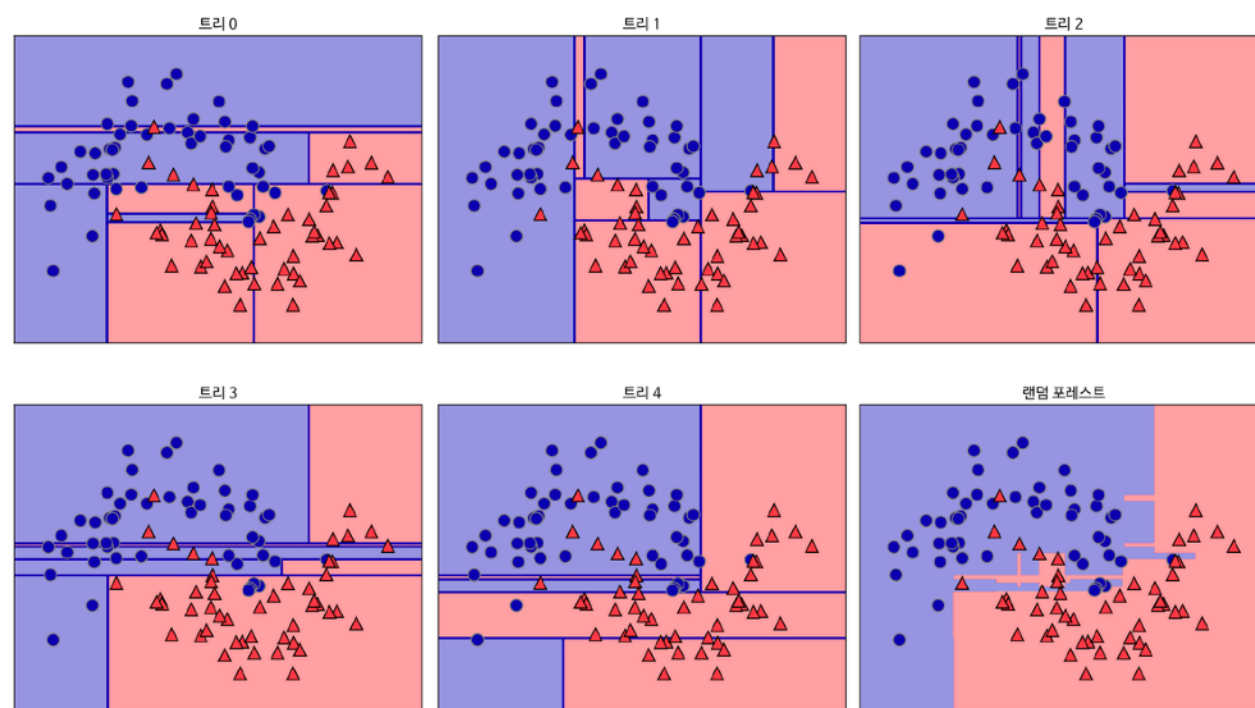




# 분류 알고리즘: 랜덤 포레스트

## 랜덤 포레스트 (Random Forest)

- 복원추출을 통해 여러개의 의사결정 나무를 만들고 각 모델의 예측에 대해 다수결의 원칙(voting)을 적용해 최종 분류결과를 결정하는 앙상블(ensemble) 기반 분류 알고리즘
- 대부분의 경우 의사결정 나무보다 과적합(overfitting)이 덜하고 더 높은 성능을 보임
- 각 자질(feature)의 중요도(importance)를 도출해낼 수 있어 자질추출 과정에서도 활용할 수 있음



\* 이형주, 정건희, 랜덤포레스트를 이용한 대설피해액에 대한 범주형 예측 및 개선방안 검토, 2019.5.

\*\* references

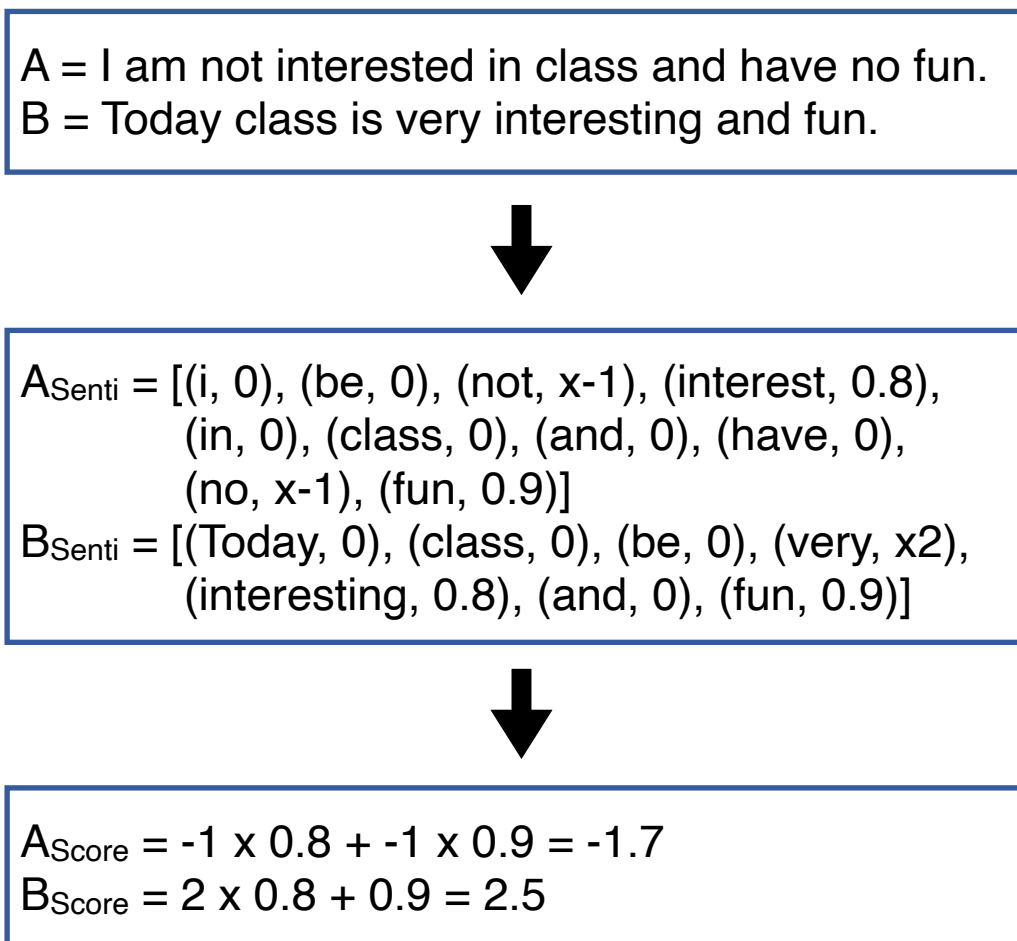
\*\*\* references

# 문서의 감성수준을 평가하는 방법

## 감성분석 (Sentiment Analysis)

- 문장이 의미하는 감성의 극성을 판별하거나 그 수준을 점수로 매기는 분석방법
- 텍스트 데이터를 수치 데이터로 바꾸는 가장 좋은 방법 중 하나
- 단순히 사전(말뭉치) 기반으로 감성수준을 판별하는 방법과 머신러닝을 활용해 판별하는 방법이 있음

### [ 사전기반 감성분석 ]



### [ 감성사전 예시 ]

Word	Polarity	Weight
not	-	negation
no	-	negation
...	...	...
interest	+	0.8
fun	+	0.9
...	...	...
sorry	-	0.9
sad	-	0.8
...	...	...

# 문서의 감성수준을 평가하는 방법

## 상용/연구용 감성사전 종류

- Linguistic Inquiry and Word Count (LIWC) - <http://www.liwc.net/>
- MPQA Subjectivity Cues Lexicon - <http://www.cs.pitt.edu/>
- SentiWordNet - <http://sentiwordnet.isti.cnr.it/>
- KOSAC - <http://word.snu.ac.kr/kosac/lexicon.php>

ngram	freq	COMP	NEG	NEUT	None	POS	max.value	max.prop
싸구려/NNG	1	0	1	0	0	0	NEG	1
싸구려/NNG;로/JKB	1	0	1	0	0	0	NEG	1
싸구려/NNG;로/JKB;둔갑/NNG	1	0	1	0	0	0	NEG	1
싸늘/XR	1	0	1	0	0	0	NEG	1
싸늘/XR;하/XSA	1	0	1	0	0	0	NEG	1
싸움/NNG	1	0	1	0	0	0	NEG	1
싸움/NNG;을/JKO	1	0	1	0	0	0	NEG	1
싸움/NNG;을/JKO;일으키/VV	1	0	1	0	0	0	NEG	1
써먹/VV	1	0	1	0	0	0	NEG	1
써먹/VV;지/EC	1	0	1	0	0	0	NEG	1
써먹/VV;지/EC;못하/VX	1	0	1	0	0	0	NEG	1
기대/NNG;되/XSV	2	0	0	0	0	1	POS	1
기대/NNG;를/JKO	2	0	0	0	0	1	POS	1
기대/NNG;하/XSV	2	0	0	0	0	1	POS	1
기량/NNG	2	0	0	0	0	1	POS	1
기뻐하/VV	2	0	0	0	0	1	POS	1
기회/NNG;를/JKO;주/VV	2	0	0	0	0	1	POS	1
길/VA	2	0	0	0	0	1	POS	1
꼭/MAG;필요/NNG	2	0	0	0	0	1	POS	1
꼭/MAG;필요/NNG;하/XSA	2	0	0	0	0	1	POS	1
꿈/VV	2	0	0	0	0	1	POS	1
꿈히/VV	2	0	0	0	0	1	POS	1
꽃/NNG;을/JKO	2	0	0	0	0	1	POS	1
꽃/NNG;을/JKO;피우/VV	2	0	0	0	0	1	POS	1

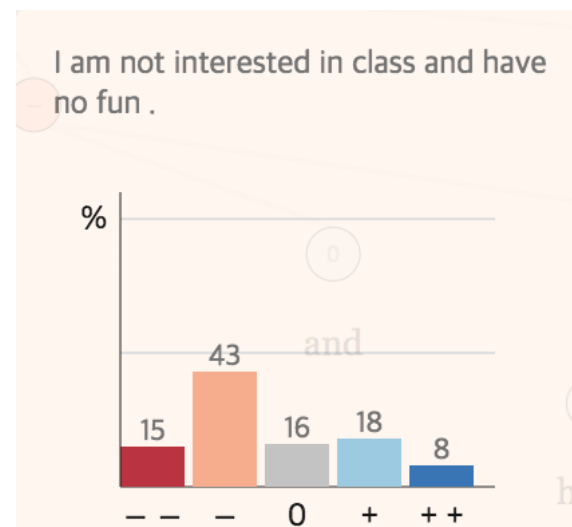
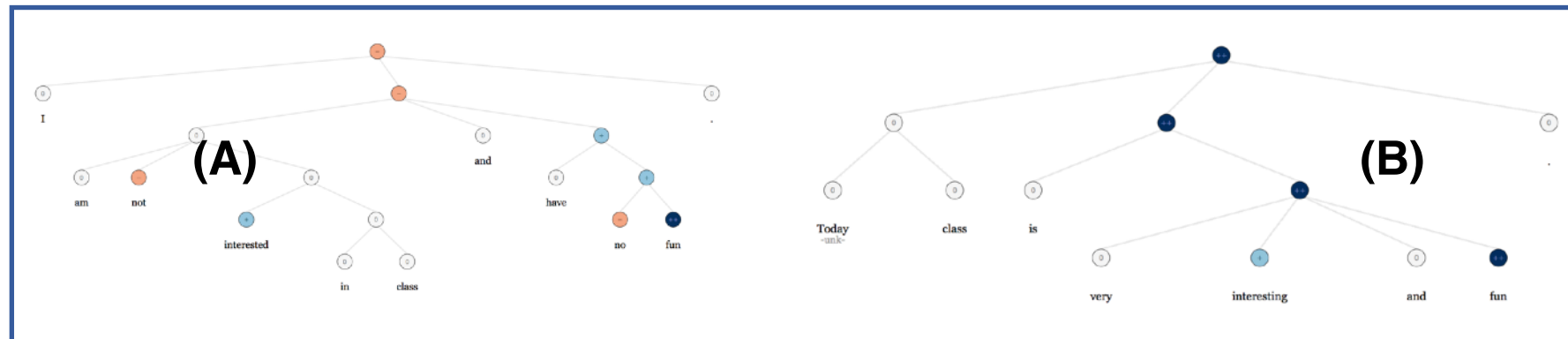
\* references  
\*\* references  
\*\*\* references

ngram	freq	COMP	NEG	NEUT	None	POS	max.value	max.prop
싸구려/NNG	1	0	1	0	0	0	NEG	1
싸구려/NNG;로/JKB	1	0	1	0	0	0	NEG	1
싸구려/NNG;로/JKB;둔갑/NNG	1	0	1	0	0	0	NEG	1
싸늘/XR	1	0	1	0	0	0	NEG	1
싸늘/XR;하/XSA	1	0	1	0	0	0	NEG	1
싸움/NNG	1	0	1	0	0	0	NEG	1
싸움/NNG;을/JKO	1	0	1	0	0	0	NEG	1
싸움/NNG;을/JKO;일으키/VV	1	0	1	0	0	0	NEG	1
써먹/VV	1	0	1	0	0	0	NEG	1
써먹/VV;지/EC	1	0	1	0	0	0	NEG	1
써먹/VV;지/EC;못하/VX	1	0	1	0	0	0	NEG	1
기대/NNG;되/XSV	2	0	0	0	0	1	POS	1
기대/NNG;를/JKO	2	0	0	0	0	1	POS	1
기대/NNG;하/XSV	2	0	0	0	0	1	POS	1
기량/NNG	2	0	0	0	0	1	POS	1
기빠하/VV	2	0	0	0	0	1	POS	1
기회/NNG;를/JKO;주/VV	2	0	0	0	0	1	POS	1
길/VA	2	0	0	0	0	1	POS	1
꼭/MAG;필요/NNG	2	0	0	0	0	1	POS	1
꼭/MAG;필요/NNG;하/XSA	2	0	0	0	0	1	POS	1
꼽/VV	2	0	0	0	0	1	POS	1
꼽히/VV	2	0	0	0	0	1	POS	1
꽃/NNG;을/JKO	2	0	0	0	0	1	POS	1
꽃/NNG;을/JKO;피우/VV	2	0	0	0	0	1	POS	1

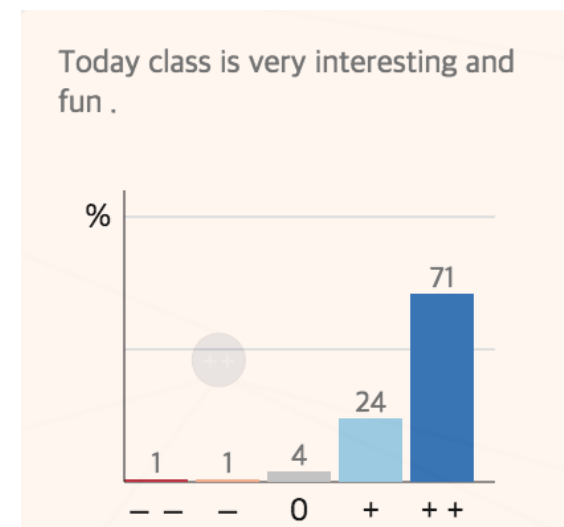
# 문서의 감성수준을 평가하는 방법

[ 머신러닝(딥러닝) 기반 감성분석 ]

A = I am not interested in class and have no fun.  
B = Today class is very interesting and fun.



A<sub>Score</sub> = -2  
B<sub>Score</sub> = +2



E.O.D

Contact

 <http://www.teanaps.com>

 [fingeredman@gmail.com](mailto:fingeredman@gmail.com)