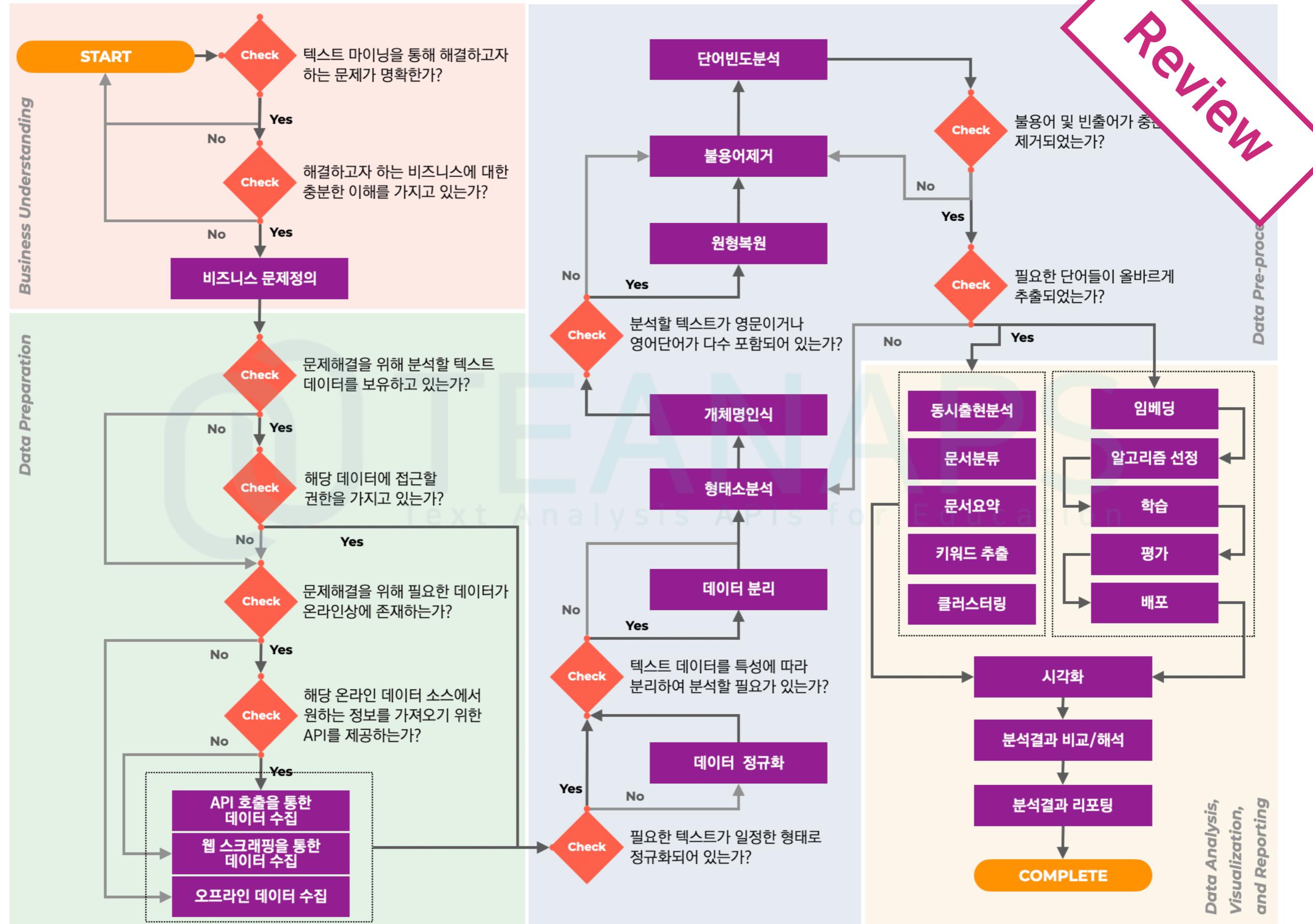


# ADVANCED TEXT MINING

---

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



WEEK 12

# Predictive Analysis

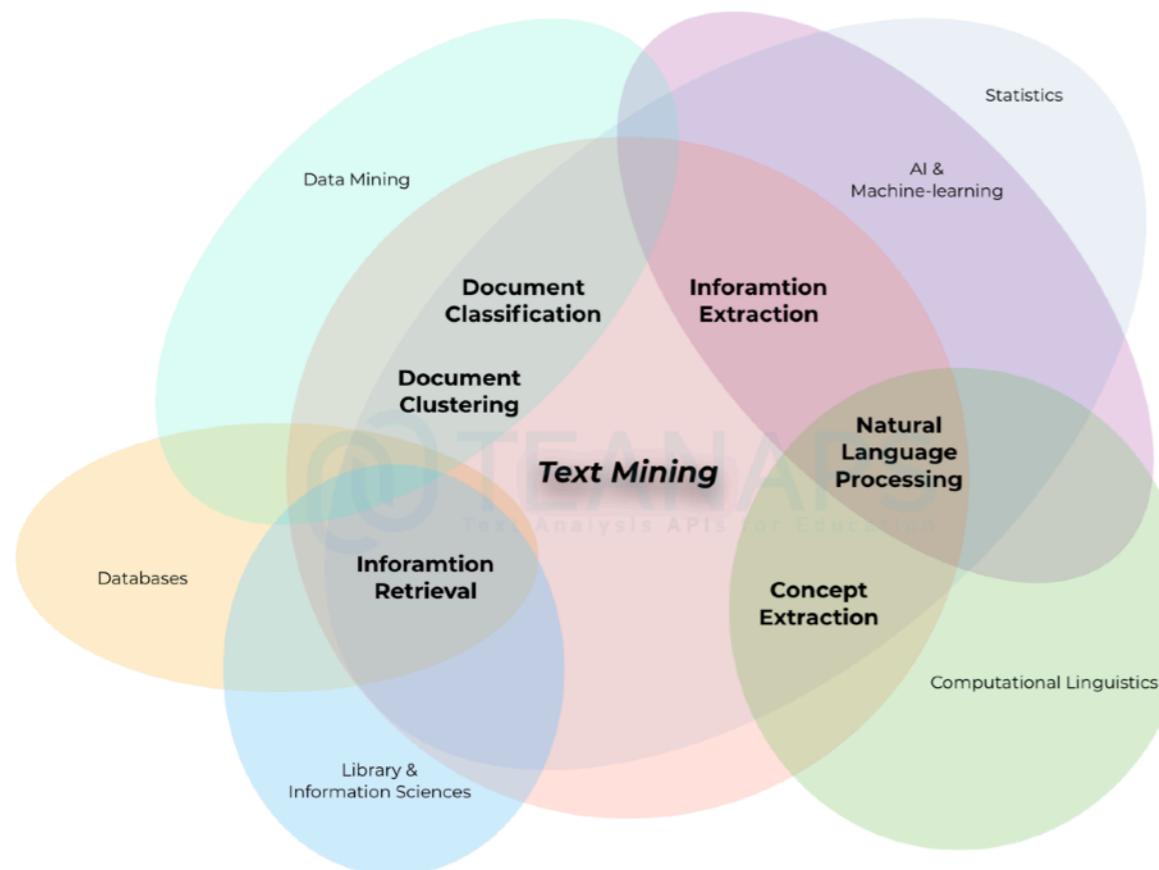


# What is Text Mining

Review

## 텍스트 마이닝 (Text Mining)

- 텍스트 데이터에서 자연어처리(Natural Language Processing, NLP) 기술을 바탕으로 유의미한 패턴 또는 지식을 추출하는 과정
- 언어학과 통계 기반에서 머신러닝을 통해 기계가 언어학적, 통계적 특징을 학습하는 형태로 발전하여 활용됨
- 텍스트 마이닝 유형
  - 1) **설명적 마이닝**(descriptive mining) : 텍스트 집합에 있는 의미나 개념을 찾아내거나 이해를 돋는 형태 (분류, 검색, 여론조사 등)
  - 2) **예측적 마이닝**(predictive mining) : 텍스트에 내포된 정보를 의사결정에 활용하는 형태 (질문 자동답변, 구매 예측, 주가예측, 스팸분류 등)



활용분야	텍스트 마이닝 유형	
	실무	연구
스팸 필터링	검색 (Information Retrieval)	사회동향 분석
이슈 검출/트래킹	분류 (Classification)	소셜미디어 분석
정보검색	군집화 (Clustering)	이슈 트래킹
자살률 예측	웹마이닝 (Web Mining)	온라인 행동 분석
주가 예측	정보추출 (Information Extraction)	연구분야 탐색
소비자 인식 조사	개념추출 (Concept Extraction)	질병관계 예측
경쟁사 분석	자연어처리 (NLP)	정책전략 수립

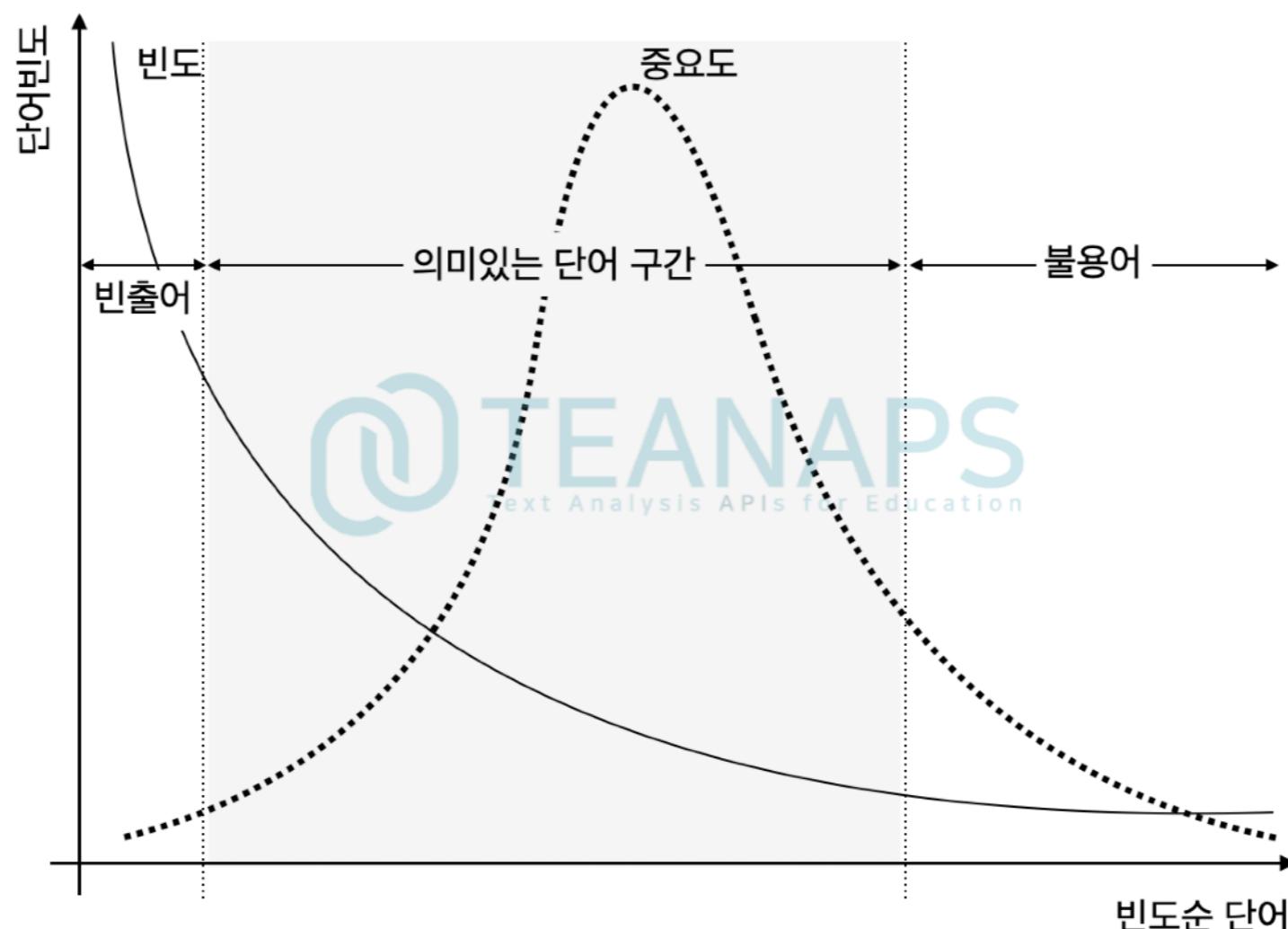
# 단어 가중치: 단어빈도

Review

## 단어빈도 (Term Frequency, TF)

- 특정 단어가 문서에 출현한 횟수로 단어의 특징을 표현하는 가장 간단한 방법
- 간단하지만 가장 빠르게 문서를 표현하고 파악할 수 있으며 기초통계와 같이 분석 전 반드시 거쳐야 하는 과정
- 단어가 너무 희귀한 경우 큰 의미를 부여하기 어려우며, 너무 흔한 경우 의미가 과도하게 부여될 가능성이 있음

$$\text{TermFrequency} = \text{count}(word | document)$$



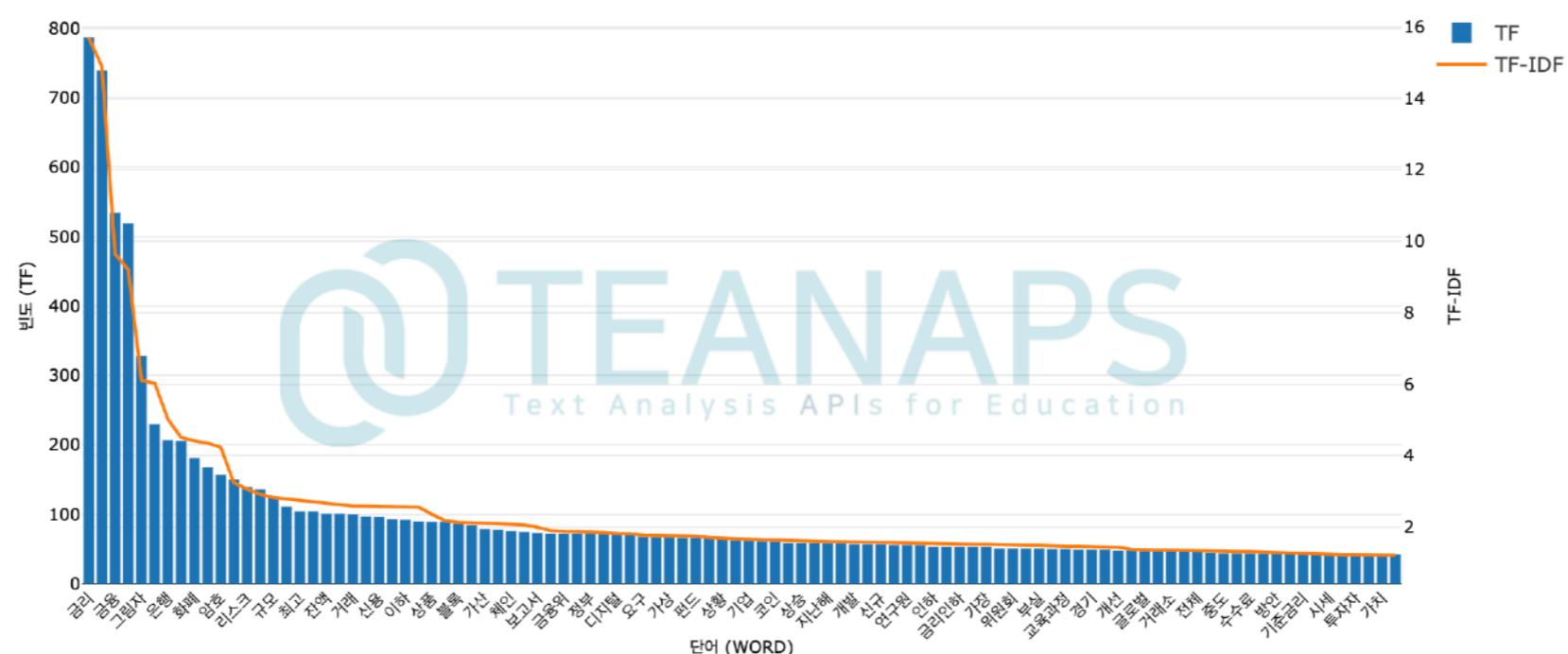
# 단어 가중치: 단어빈도

Review

## TF-IDF (TF-Inverse Document Frequency)

- 역문서빈도 (Inverse Document Frequency, IDF) : 단어가 출현한 문서가 적을수록 단어의 가중치를 낮게 표현하는 방법 (희박성)  
 $IDF = 1 + \log(\text{document}/\text{count(word/document)})/\log(\text{document})$
- TF와 IDF 개념을 통합하여, 단어가 문서에 출현한 횟수와 희박성을 동시에 활용해 가중치를 표현하는 방법  
 $TF - IDF = \text{Frequency} * IDF$
- 지프의 법칙 (Zipf's law)
  - 1) 자연어에 나타나는 단어들을 출현 횟수가 높은 순으로 정렬하면, 단어의 출현 횟수는 순위에 반비례함
  - 2) 가장 사용 빈도가 높은 단어는 두 번째 단어보다 빈도가 약 두 배 높으며, 세 번째 단어보다는 빈도가 세 배 높음

단어빈도 및 TF-IDF (TF &amp; TF-IDF)

\* 김수인, 김재원, and 배휘동, 왜 프로그래밍에는 창의성이 필요하다고 할까요, 2017.5.25., <https://medium.com/elice/>.

\*\* references

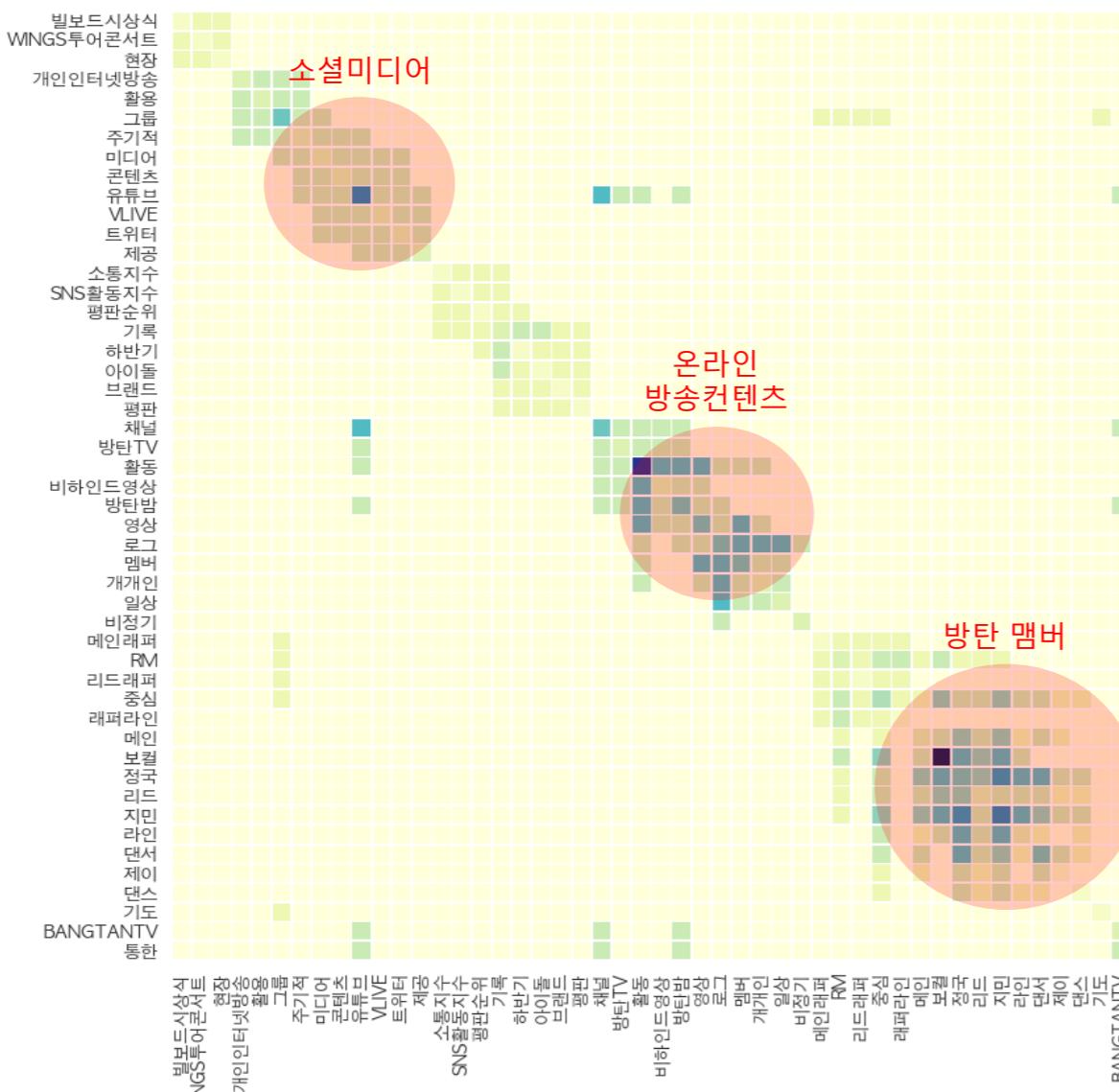
\*\*\* references

# 단어 가중치: 네트워크 중심성

Review

## 동시출현 분석 (Co-word Analysis)

- 문서에 서로 다른 두 단어의 동시출현 횟수와 네트워크 중심성을 통해 단어의 특징을 표현하는 방법
- 두 단어 사이의 동시출현을 연관성의 척도로 취급하고, 그 관계를 네트워크 중심성으로 표현하여 가중치를 계산함
- **연관어** (공기어, Co-word) : 하나의 문서에서 함께 출현하여 서로 밀접한 의미관계를 가지는 단어



**표1 '아쿠르트 아줌마' 연관어 변화**

아쿠르트 아줌마는 여전히 '아쿠르트'와의 연관도가 가장 높지만 2016년 들어 '커피' 및 '크림치즈' 제품 연관어와 '10일'이라는 키워드가 등장. 아쿠르트 아줌마는 '배달하는' 역할에서 맛난 제품을 위해 '만나고' '찾고' '발견하는' 대상으로 변화 중.

2013년		2014년		2015년		2016년		
No.	연관어	언급 비중	No.	연관어	언급 비중	No.	연관어	언급 비중
1	아쿠르트	21.3%	1	아쿠르트	26.3%	1	아쿠르트	26.6%
2	먹다	4.9%	2	건강	4.5%	2	집	4.7%
3	아침	4.4%	3	아침	4.0%	3	아침	4.4%
4	엄마	4.2%	4	집	3.6%	4	맛	3.9%
5	집	3.5%	5	제품	3.4%	5	먹다	3.4%
6	오다	2.8%	6	엄마	3.3%	6	사다	2.8%
7	사다	2.7%	7	맛	2.7%	7	주다	2.8%
8	주다	2.5%	8	같다	2.6%	8	다니다	2.7%
9	구입하다	2.4%	9	우유	2.6%	9	엄마	2.6%
10	아이	2.4%	10	주다	2.2%	10	우유	2.1%
11	아쿠르트 주다	2.3%	11	먹다	2.2%	11	만나다	2.1%
12	배달하다	2.3%	12	만나다	2.0%	12	제품	2.0%
13	수입	2.3%	13	사다	1.9%	13	사진	2.0%
14	다니다	2.1%	14	알다	1.9%	14	나오다	2.0%
15	얼려먹다	2.0%	15	배달하다	1.8%	15	팔다	1.9%
16	살다	2.0%	16	다니다	1.8%	16	지나가다	1.8%
17	제품	2.0%	17	하루야채	1.7%	17	하나	1.7%
18	세븐	1.8%	18	나누다	1.7%	18	판매	1.7%
19	가다	1.8%	19	지나가다	1.6%	19	일하다	1.6%
20	자녀	1.8%	20	세븐	1.5%	20	오다	1.6%
21	만나다	1.8%	21	수입	1.5%	21	찾다	1.6%
22	마시다	1.7%	22	찾다	2.3%	22	음료	1.5%
23	유산균	1.7%	23	노인	1.4%	23	마시다	1.4%
24	일하다	1.7%	24	마시다	1.4%	24	길	1.4%
...		...		...		...		
29	팔다	1.4%	29	물다	1.3%	29	배달하다	1.3%
...		...		...		...		
29	구입하다	1.0%						

상승 키워드      하락 키워드      신규 키워드

\* 전병진, 신한은행 파이낸스로 시작하는 데이터분석: 텍스트 마이닝 기초, 2018.12.12.

\*\* 백경혜(DBR), "매력을 소비하는 나는 덕후! 즐거움을 위해 기꺼이 지갑을 연다", 2017.1., [http://dbr.donga.com/article/view/1203/article\\_no/7935/](http://dbr.donga.com/article/view/1203/article_no/7935/).

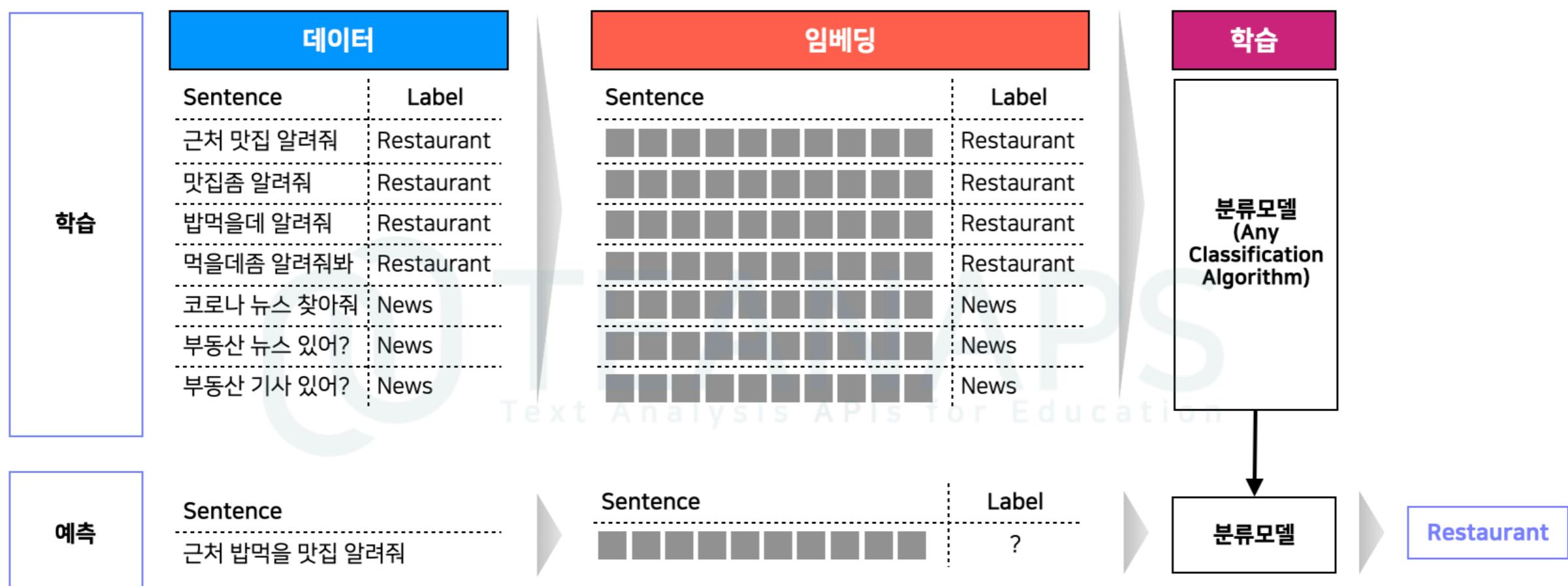
\*\*\* references

# 텍스트를 분류하는 방법

Review

## 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(descret 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



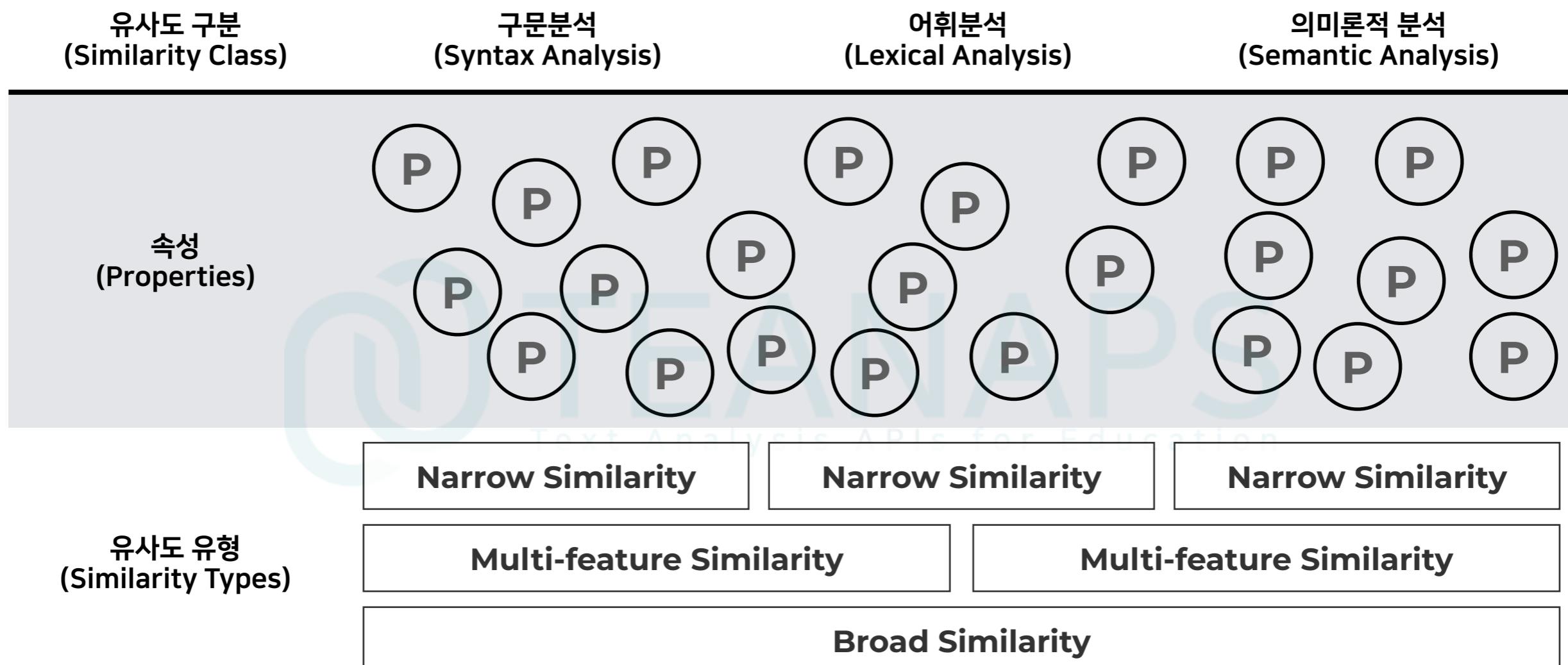
# 문서 유사도

(Document Similarity)

Review

## 유사도 (Similarity)

- 서로 다른 두 객체 사이의 공통점을 통해 서로 공유하는 속성의 수에 따라 증가하는 유사한 정도
- 서로 공유하는 속성은 그 기능에 따라서 매우 많이 존재할 수도 있으며 없을 수도 있음

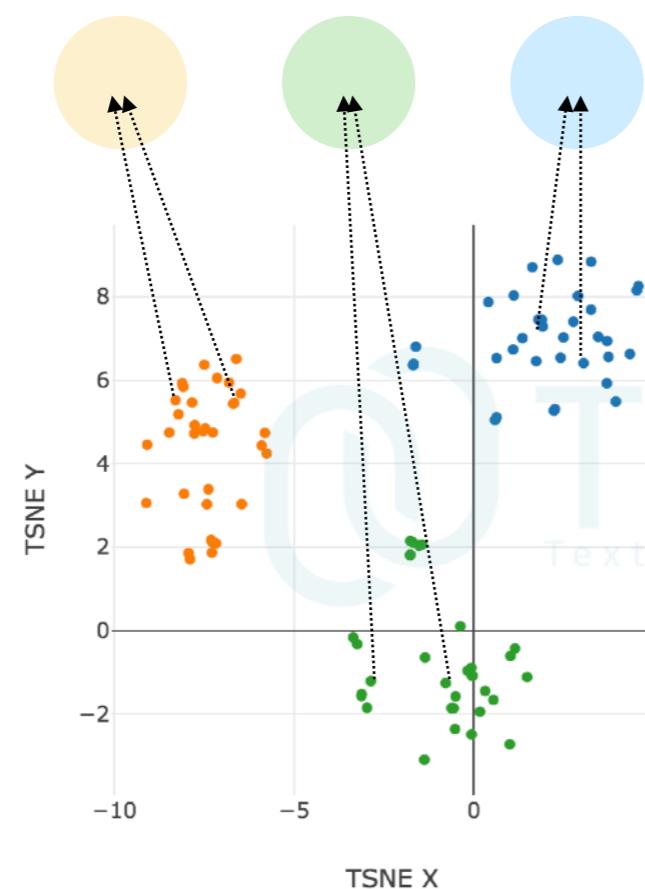


# 문서 군집화 (Document Clustering)

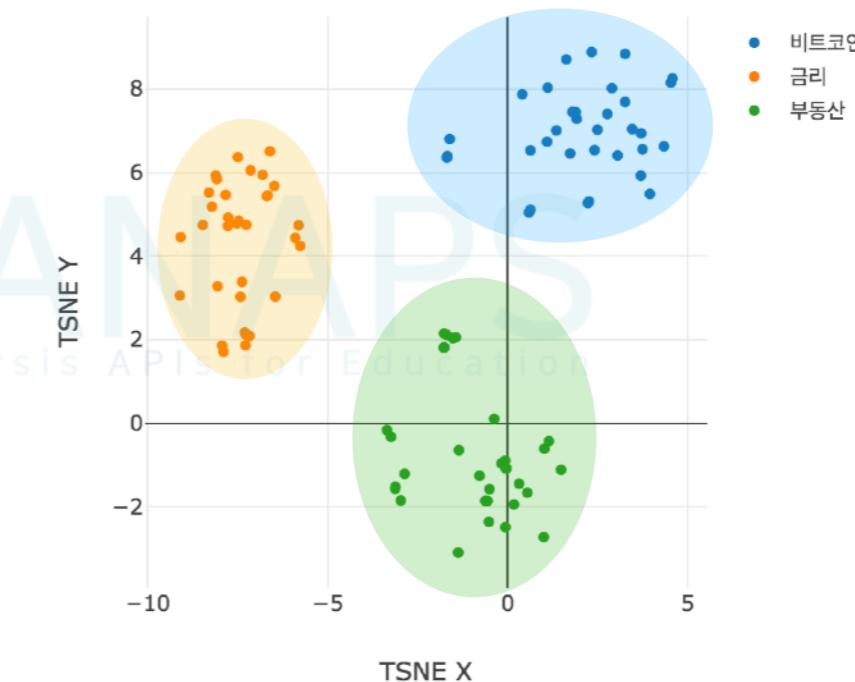
Review

## 군집화 (Clustering)

- 데이터 간에 발견되는 자연스러운 그룹(군집)을 발견하고 주제를 제시하는 분석방법
- 분류와 군집화의 차이점
  - 1) **분류** (Classification) : 문서의 집합을 이미 정해진 유형의 개수와 속성에 따라서 분류하는 방법
  - 2) **군집화** (Clustering) : 사전에 유형의 개수와 속성이 알려지지 않은 상태로 그룹을 발견하는 방법



K-Means Clustering Graph - label



# 군집화 알고리즘: 토픽 모델링

Review

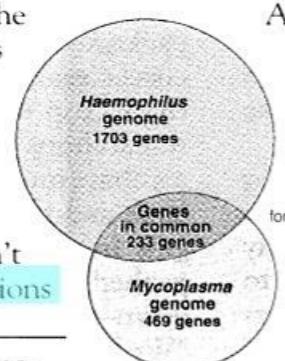
## 토픽모델링 (Topic Modeling)

- 뉴스, 블로그, 웹페이지, 기사 등 구조화되지 않은 방대한 문서(비정형데이터)에서 주제를 찾아내기 위한 방법
- 맥락과 관련된 단서들을 이용하여 유사한 의미를 가진 단어들을 클러스터링하는 방식으로 주제를 추출하며, 같은 맥락에서 나타날 가능성이 있는 단어들을 그룹화함
- 활용범위 : 문서 요약, 검색 등

## Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,\* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions



**Stripping down.** Computer analysis yields an estimate of the minimum modern and ancient genomes.

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains

Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



\* 승민, 텍스트 마이닝, 2018., 도서출판 청람.

# 추출요약과 생성요약

Review

## 문서 요약 (Document Summarization)

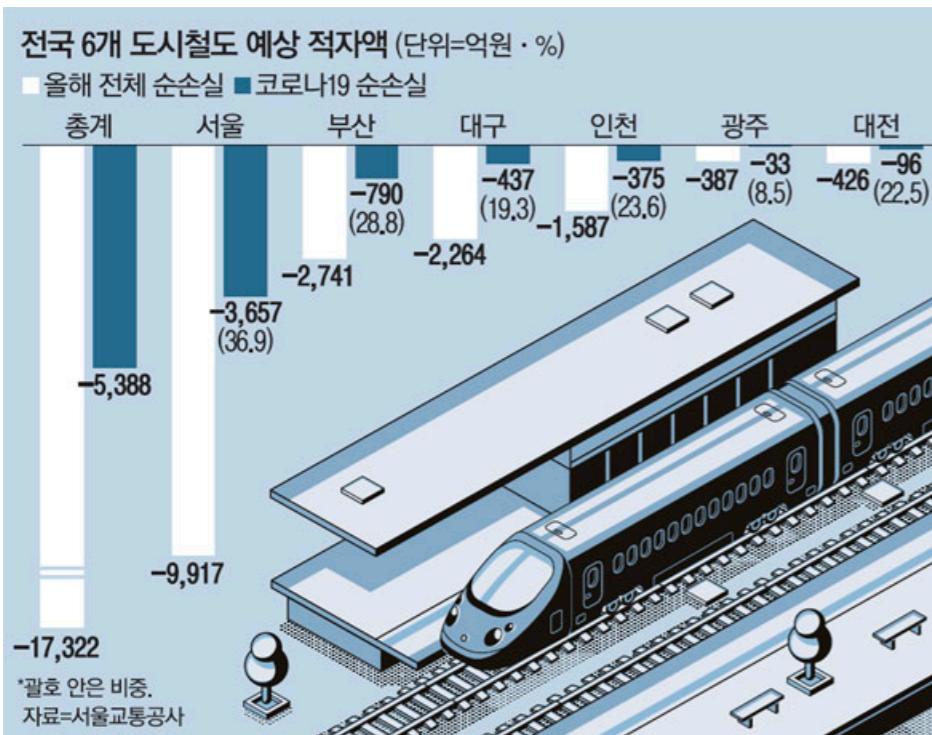
- 문서를 목적에 맞게 축약된 형태의 문서로 표현하는 방법
- 문서의 복잡도를 줄이면서도 필요한 정보는 유지하고 강조하여 표현하는 것이 중요함
- **추출요약** (Extractive Summarization) : 문서에 존재하는 단어나 구, 문장을 그대로 활용하여 요약하는 방법
- **생성요약** (Abstractive Summarization) : 문서의 내용을 요약하여 표현한 새로운 문서를 작성하는 방법

**매일경제**  
코로나로 승객 29% 감소...지하철 적자 1.7조

A29면 TOP | 기사입력 2020.11.15. 오후 5:49 | 기사원문 | 스크랩 | 본문듣기 | 설정

공감 댓글

오늘보 가 둘 드립니다



서울을 비롯한 전국 6개 도시철도의 올해 예상 적자액이 무려 1조7000억원에 달하는 것으로 나타났다. 도시철도의 만년 적자 원인인 '무임승차 손실' 보전 문제가 해결되지 않고 있는 가운데 코로나19 감염 우려에 따른 승객 감소가 겹친 탓이다. 코로나19발 도시철도 경영 위기가 현실화하면서 낮은 요금 체계 개선, 중앙정부의 법정 무임수송 손실분 보전 등 대책을 촉구하는 목소리가 높지만 정부와 지방자치단체 대응이 미온적이라는 비판이 나온다.

15일 서울교통공사에 따르면 서울, 부산, 대구, 인천, 광주, 대전 등 6개 도시철도의 올해 예상 당기순 손실은 1조7322억원으로 집계됐다. 보유한 노선이 가장 많은 서울이 9917억원으로 예상 적자 폭이 가장 커으며, 부산(2741억원) 대구(2264억원) 인천(1587억원) 대전(426억원) 광주(387억원)가 뒤를 이었다. 전체 예상 적자 가운데 코로나19에 따른 승객 감소가 불러온 순손실액만 5388억원으로 30%가 넘는 비중을 차지했다.

올해 전국 도시철도의 예상 적자액이 이토록 심각한 이유는 무임승차 손실을 철도 운영기관이 온전히 떠안아왔고, 코로나19로 인한 큰 폭의 승객 감소가 더해졌기 때문이다. 예상 적자 폭이 가장 큰 서울 교통공사는 2016년부터 지난해까지 연간 3000억원대 무임수송 손실이 발생해 적자가 지속적으로 쌓여 왔다.

\* 나의 큰 O는 log x야(티스토리), 자동 요약 기법의 연구 동향 정리, <https://bab2min.tistory.com/625?category=673750/>.

\*\* 최현재(매일경제), 코로나로 승객 29% 감소...지하철 적자 1.7조, 2020.11.15., <https://news.naver.com/main/read.nhn?mode=LSD&mid=shm&sid1=103&oid=009&aid=0004695804/>.

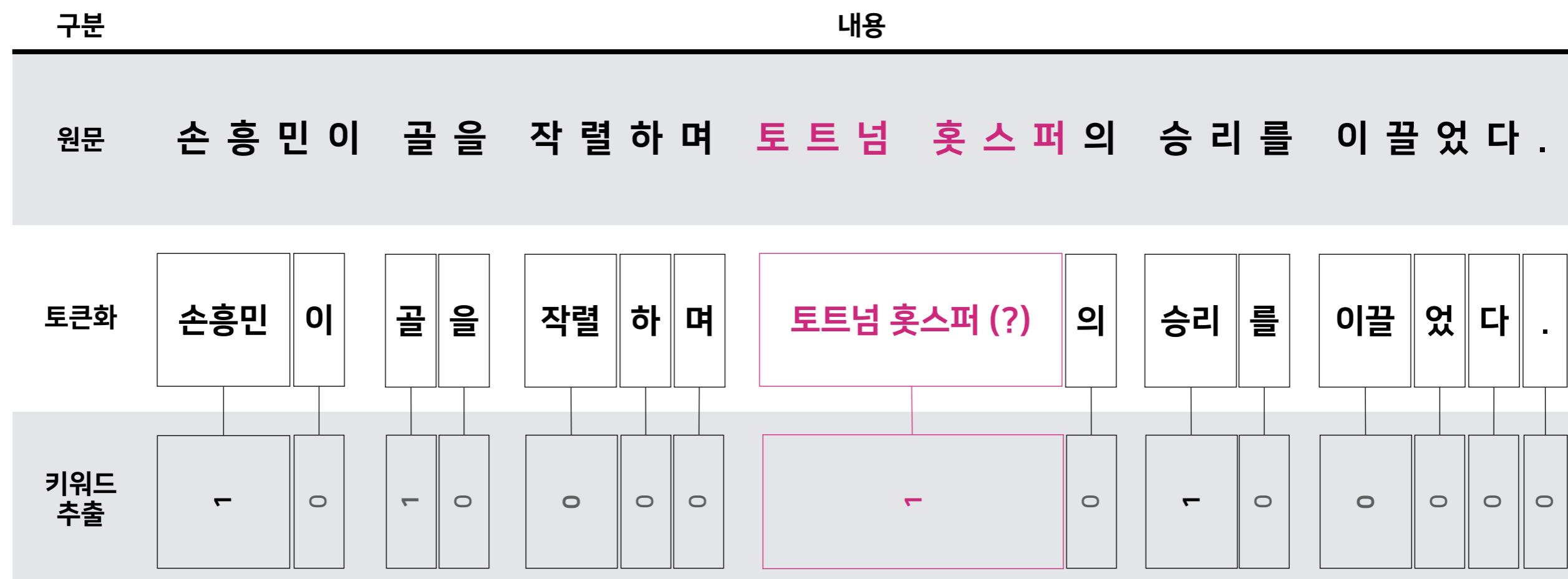
\*\*\* references

# 키워드 추출 알고리즘

Review

## 키워드 추출 (Keyword Extraction)

- 문서의 주제를 가장 잘 설명하는 키워드를 자동으로 식별하는 작업
- 키워드란 하나의 단어 뿐만 아니라 구(phrase) 단위까지를 의미하며, 영문 표현으로 Keyword, Key-phrase, Key-term, Key-segment 등이 모두 같은 의미를 가짐 (주로 명사구 단위를 취급함)
- 키워드 추출은 텍스트 마이닝, 정보검색 (information retrieval, IR) 및 자연어처리 (natural language processing, NLP) 분야에서 오랫동안 중요한 문제로 인식되어 왔으며 다양한 키워드 추출 알고리즘이 제안됨 (TF-IDF, TextRank, PKEA, RKEA, ...)



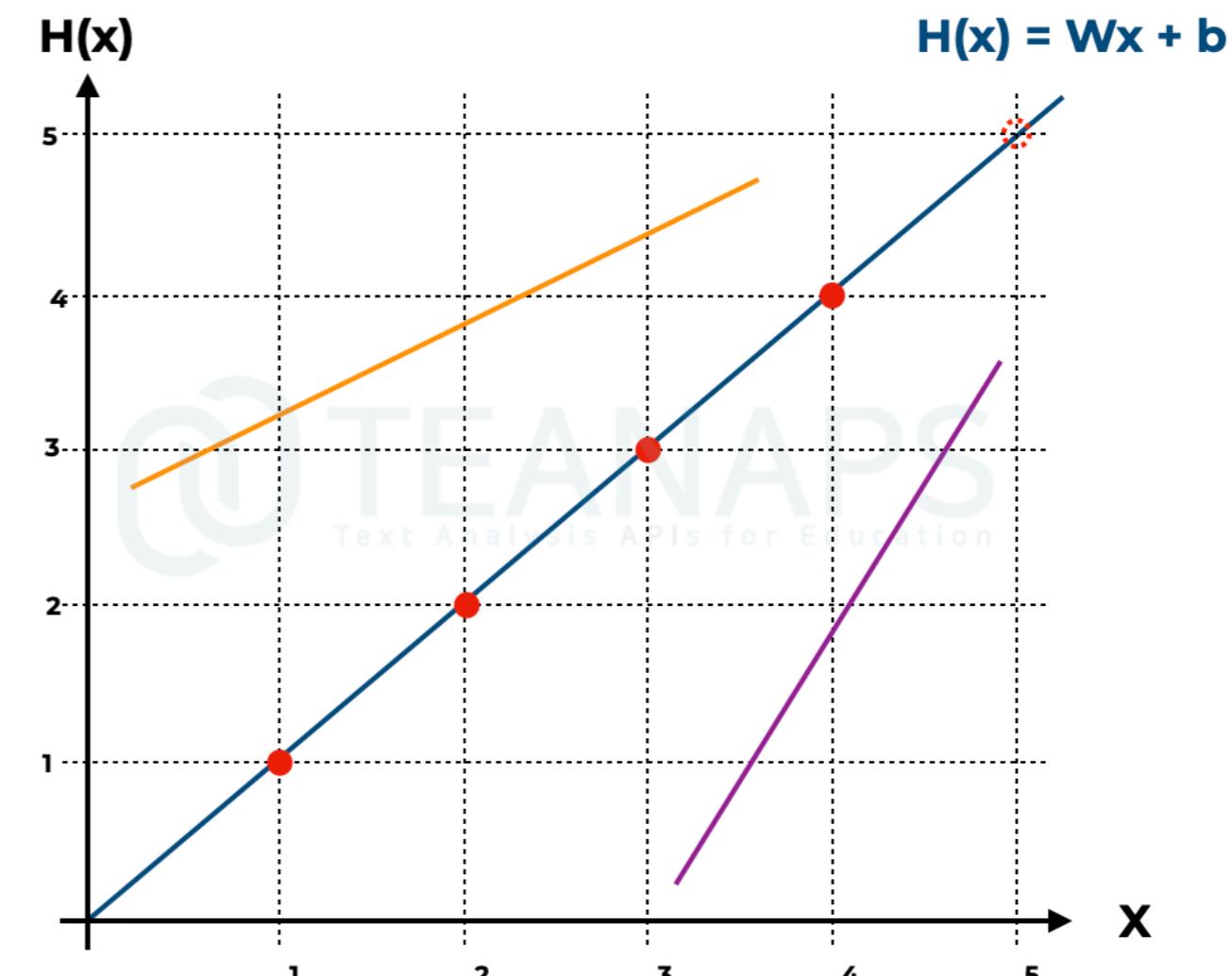
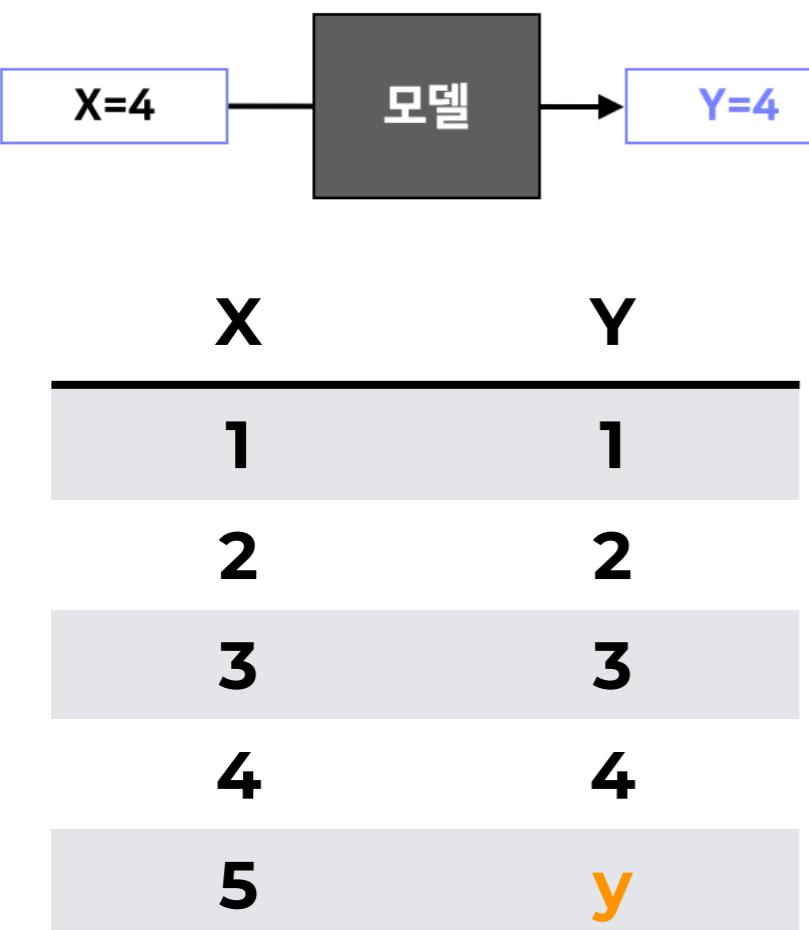
# 기계학습 절차: 학습

(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



# 기계학습 절차: 학습

(Training)

Review

## 파라미터 접근법 (Parametric Approach)

- 입력변수(x)와 목표변수(y) 사이의 복잡한 관계를 어떠한 파라미터 (w)와의 관계로 표현하는 방식
- 정답을 구하기 위한 적절한 파라미터 (w)를 구하고 예측된 값 ( $y_n$ )과 정답 (Y)와의 차이 (error, loss)를 계산하여 그 평균을 최소로 하는 적절한 파라미터 (w)를 도출하는 과정

$$y = \mathbf{a}x_1 + \mathbf{b}x_2$$

<b>a</b>	<b>x<sub>1</sub></b>	<b>b</b>	<b>x<sub>2</sub></b>	<b>y<sub>n</sub></b>	<b>Y</b>	<b>Y - y<sub>n</sub></b>
0			1	$y_1$	2	$2 - y_1$
?	1	?	2	$y_2$	6	$6 - y_2$
1			1	$y_3$	4	$4 - y_3$
1.5			1	$y_4$	5	$5 - y_4$
					Avg(Y - y <sub>n</sub> )	

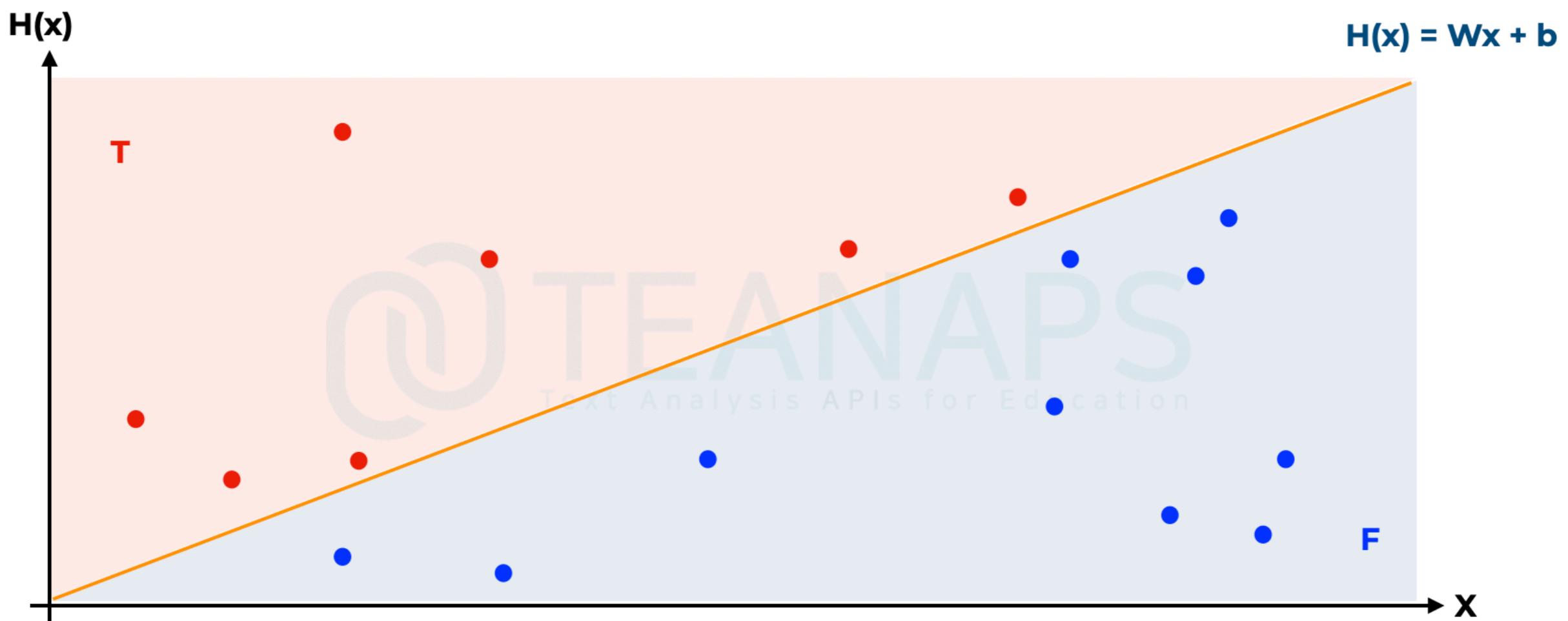
# 기계학습 절차: 학습

(Training)

Review

## 선형회귀로 분류문제를 해결하는 방법

- 선형 함수를 활성화 함수를 통해 로짓 함수로 변환하여 선형가정의 문제를 로짓가정의 문제로 변환할 수 있음
- **로짓가정** (Logistic Hypothesis) : 학습데이터의 분포를 로짓이라 가정하고 학습데이터를 가장 잘 설명한 직선
- **활성화 함수** (activation function) : 선형함수를 입력으로 활성화/비활성화 여부를 결정하여 출력하는 함수  
(계단함수(Step Function), 시그모이드(Sigmoid), 하이퍼볼릭 탄젠트(tanh), Relu)



# 기계학습 절차: 학습

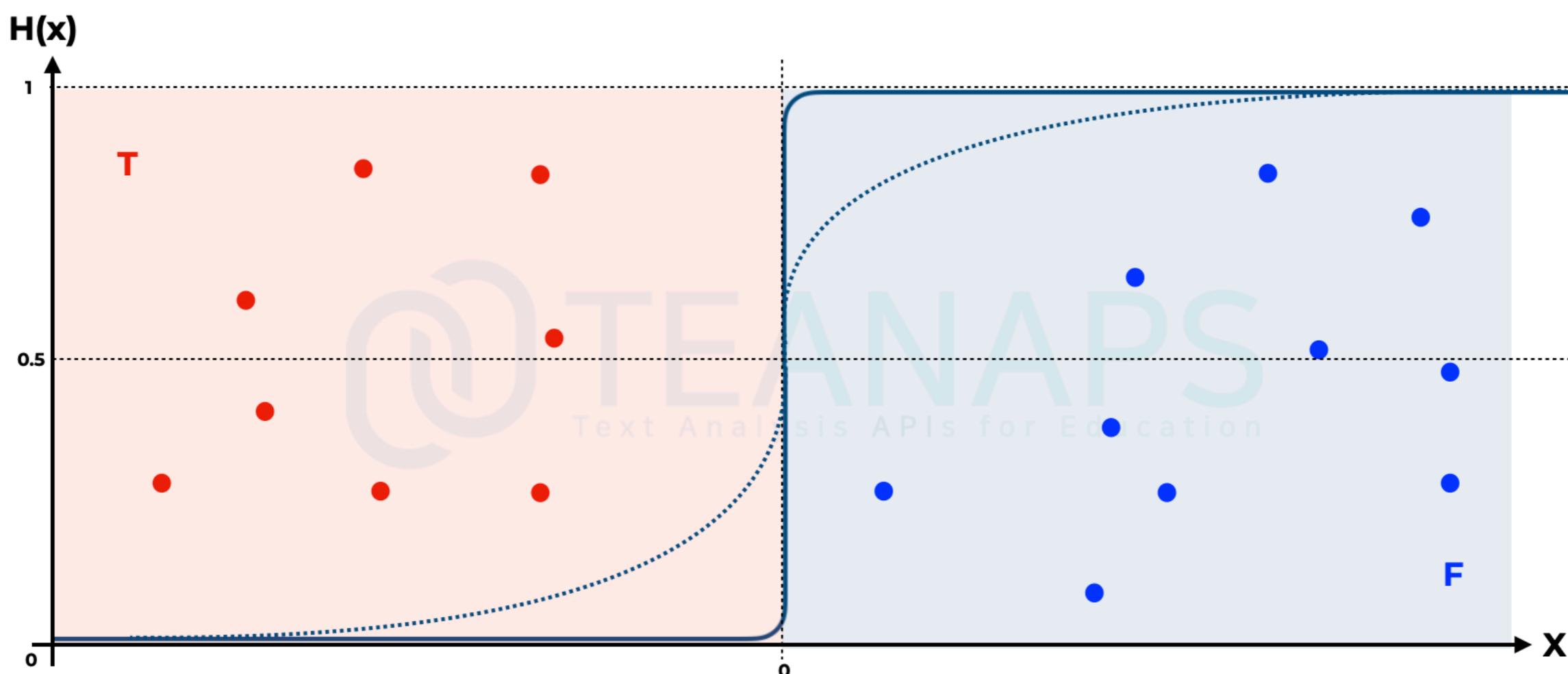
(Training)

Review

## 시그모이드 함수를 통한 로짓변환

$$\text{Sigmoid}(g) = \frac{1}{(1 + e^{-g})}$$

$$H(x) = Wx + b \rightarrow \text{Sigmoid}(H(x)) = \frac{1}{(1 + e^{-H(x)})}$$



# 기계학습 절차: 학습

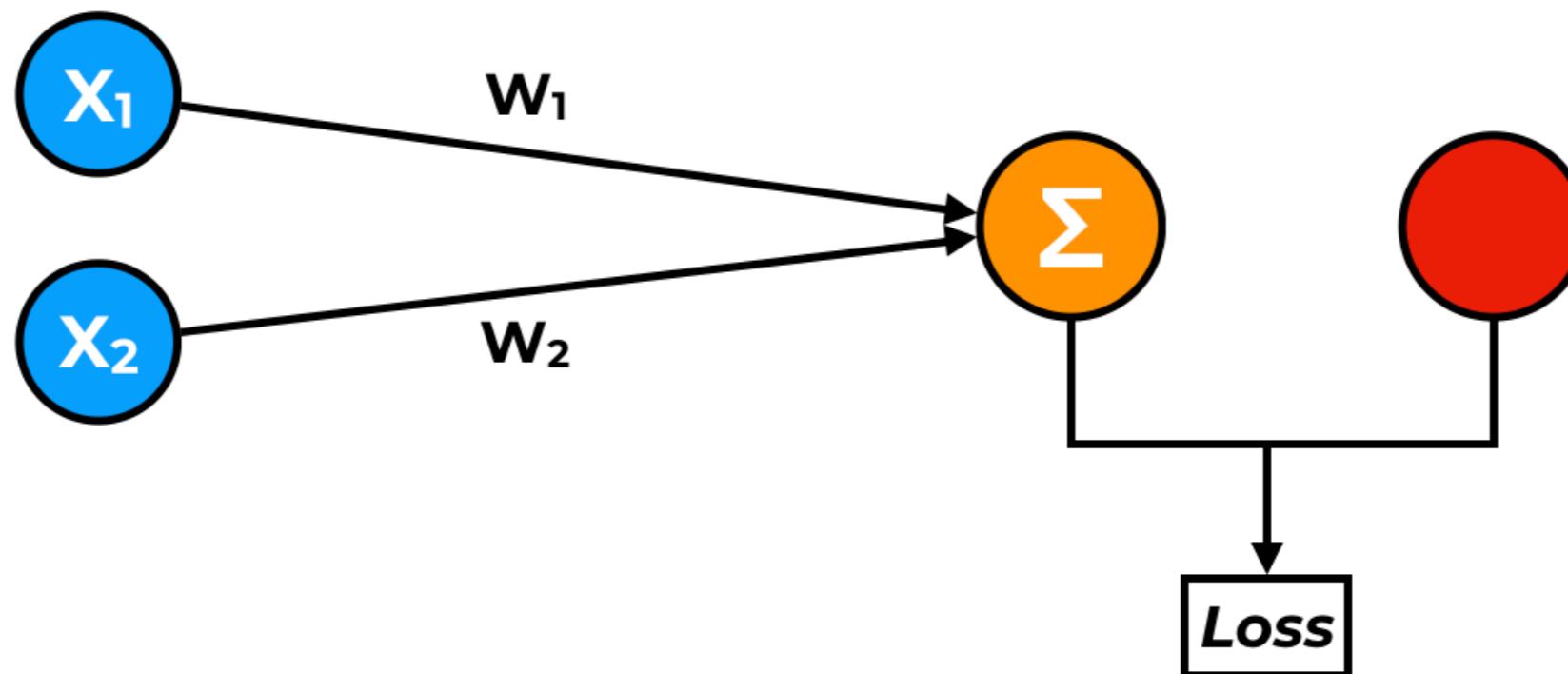
(Training)

Review

| 파라미터 접근법 (Parametric Approach)

$$Y_n = W_1 \cdot X_1 + W_2 \cdot X_2$$

Input	Weight	Output	Label
$(N, 2)$	$(2, 1)$	$(N, 1)$	$(N, 1)$



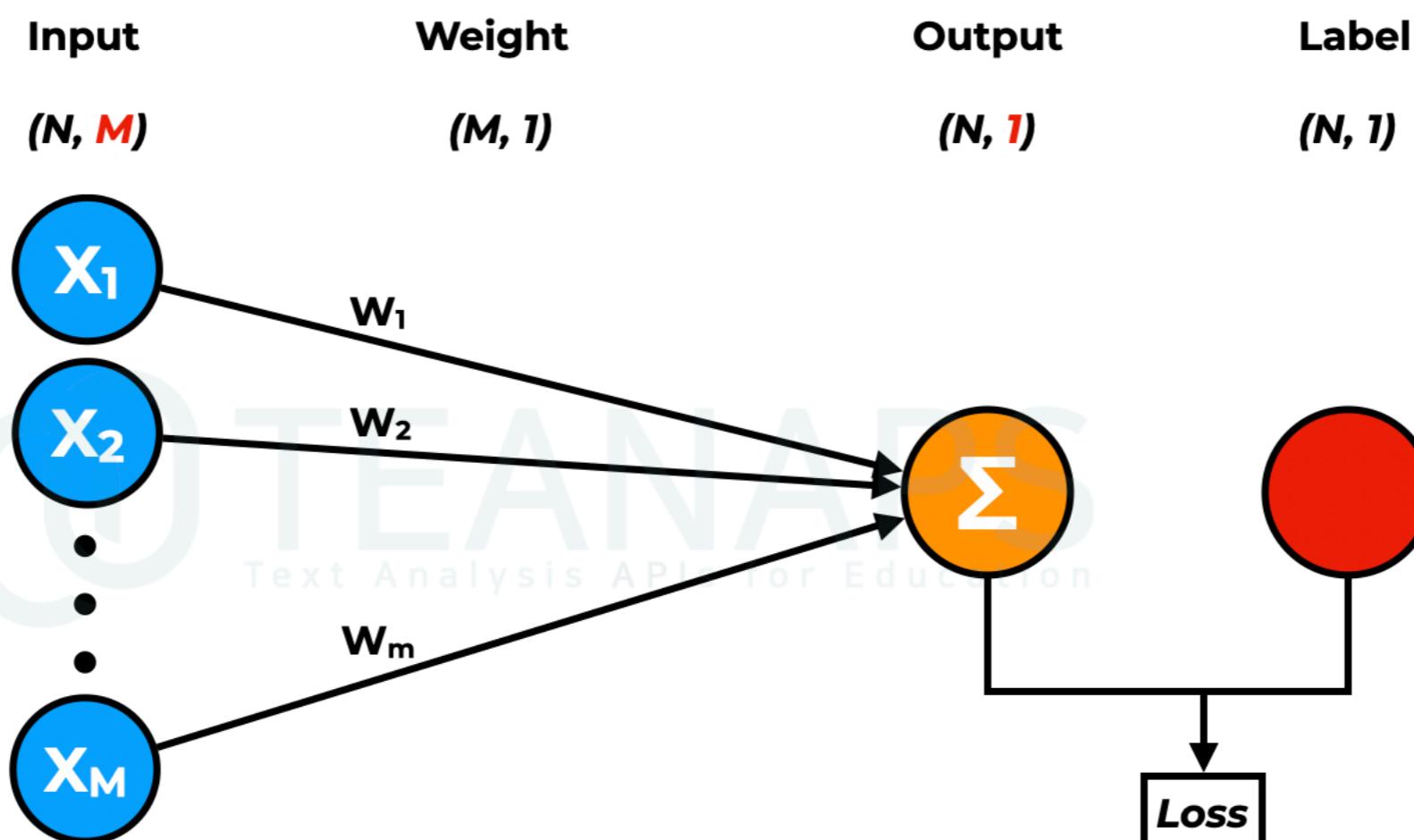
# 기계학습 절차: 학습

(Training)

Review

| 파라미터 접근법 (Parametric Approach)

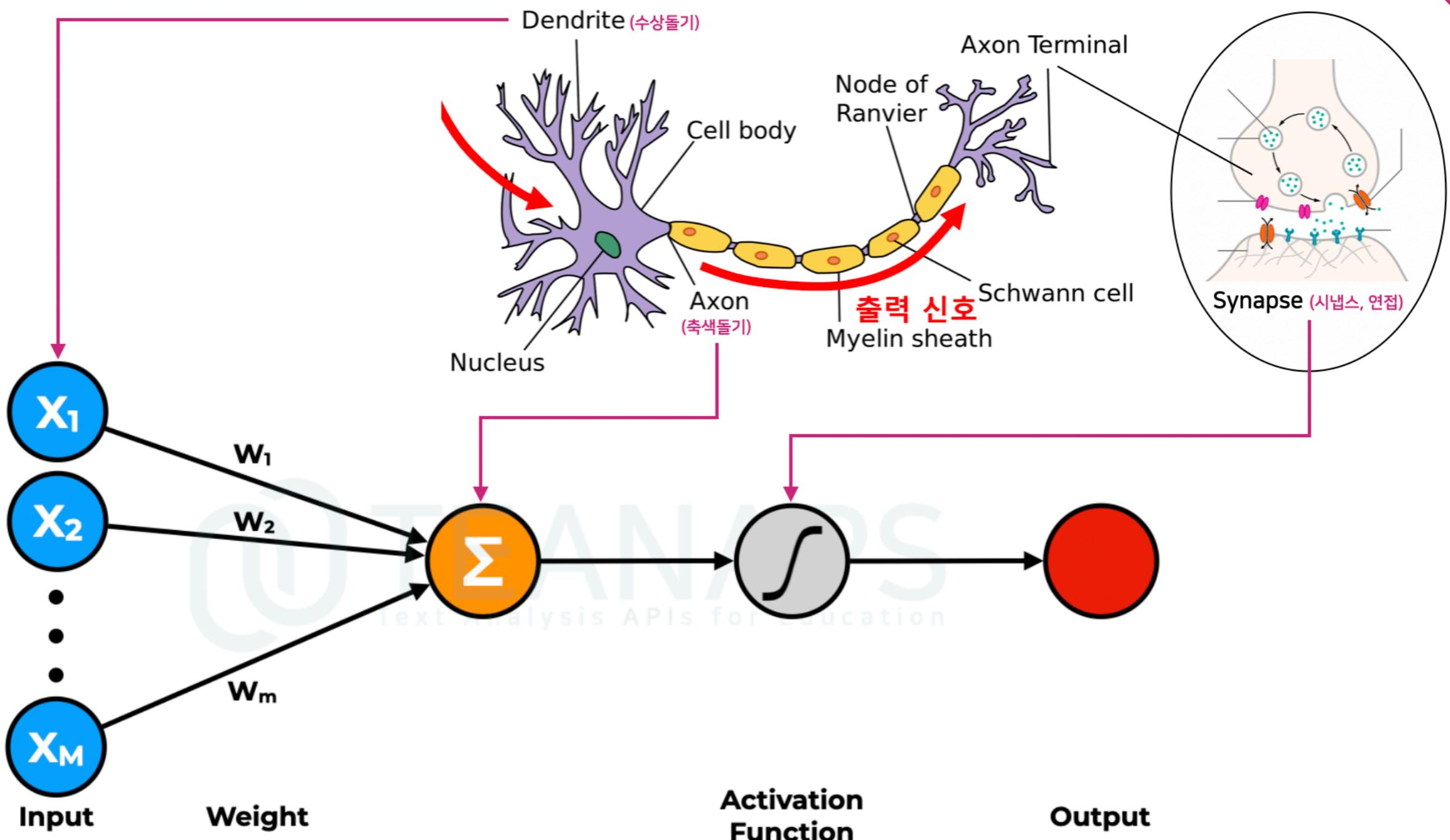
$$Y_n = W_1 \cdot X_1 + W_2 \cdot X_2 + \dots + W_m \cdot X_m$$



# 딥러닝 (Deep Learning)

Review

## 퍼셉트론 (Perceptron)



\* 미프로박재우의원, 유전성 말초신경병증 시술, [https://stems88.cafe24.com/?page\\_id=14798/](https://stems88.cafe24.com/?page_id=14798/).

\*\* 위키피디아, Neurotransmission, <https://en.wikipedia.org/wiki/Neurotransmission/>.

\*\*\* references

# E.O.D

## Contact

-  <http://www.teanaps.com>
-  [fingeredman@gmail.com](mailto:fingeredman@gmail.com)