

TEXT MINING for PRACTICE

by FINGEREDMAN (fingeredman@gmail.com)

WEEK 07

ML for NLP & Text Mining

HOW?

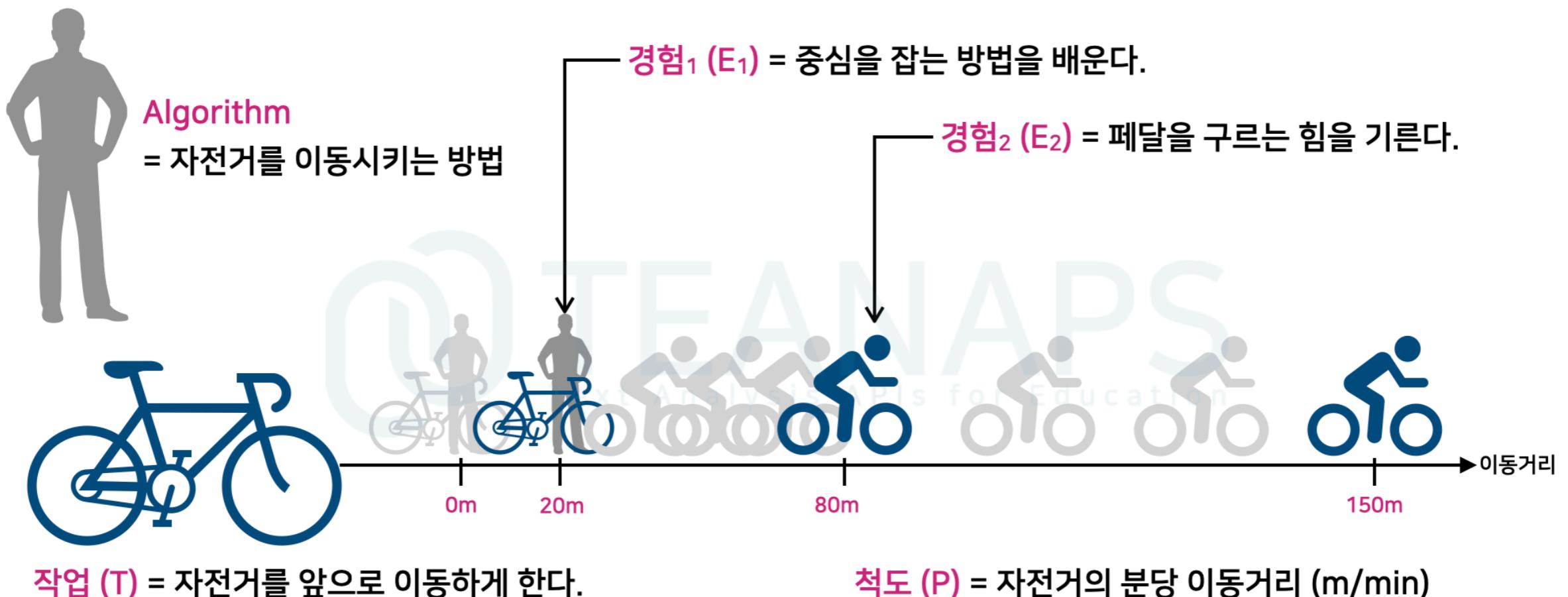
자연어 처리와 텍스트 마이닝에 머신러닝이 어떻게 활용될까?

기계학습

(Machine Learning)

Mitchell의 정의

- "A computer **program** is said to learn from **experience E** with respect to some class of **tasks T** and performance **measure P**, if its performance at tasks in T, as measured by P, improves with experience E"
- "컴퓨터 알고리즘(프로그램)이 작업 T를 수행하고 이 알고리즘의 성능을 척도 P_(performance)로 평가할 수 있다면, 경험 E_(experience)를 통해 P가 개선되는 경우 이 알고리즘은 학습이 되었다고 볼 수 있다."



기계학습

(Machine Learning)

기계학습의 기본개념

- 외부 환경이 사람을 지도(supervise)하는 것과 같이, 기계가 기존에 할 수 없던 것을 가능하도록 하게하는 과정
- 사람과 기계 모두 외부 지도에 따라 매우는 학습단계와 실제 성능을 평가하는 테스트 과정을 통해 학습함
- 기계학습의 유형

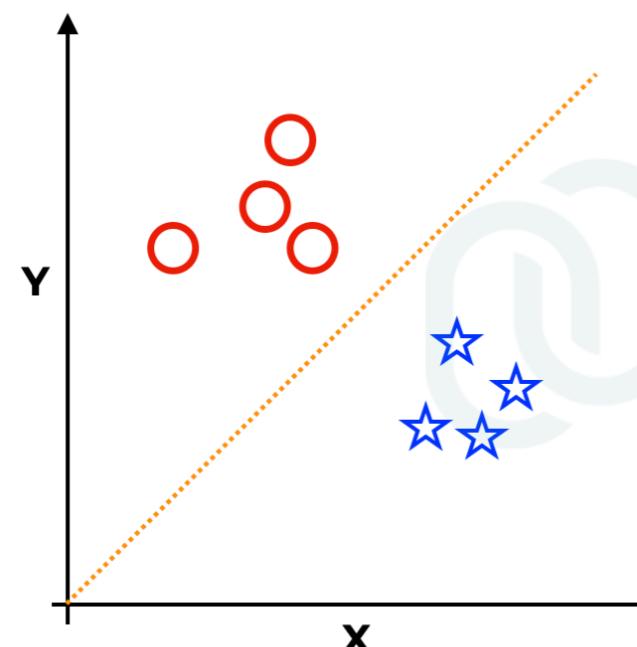
1) 지도학습 (supervised learning) :

입력과 출력을 가지는 데이터로부터 패턴을 추출하여 새로운 입력에 대한 출력을 결정하는 학습방법

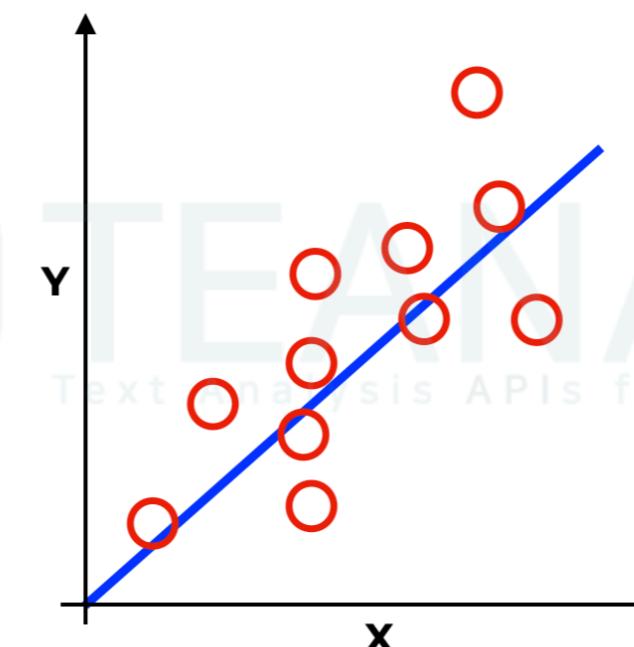
2) 비지도학습 (unsupervised learning) :

출력에 대한 정의가 없는 데이터로부터 의미있는 패턴을 추출하는 학습방법

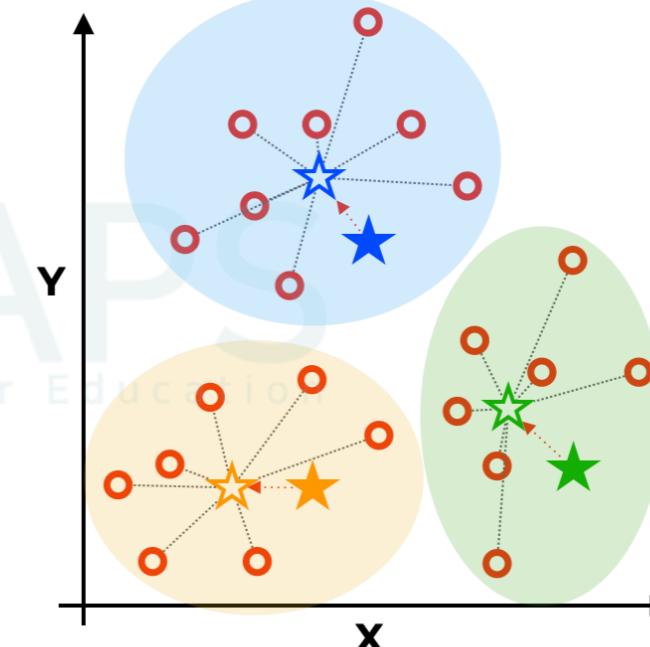
분류 (Classification)



회귀분석 (Regression)

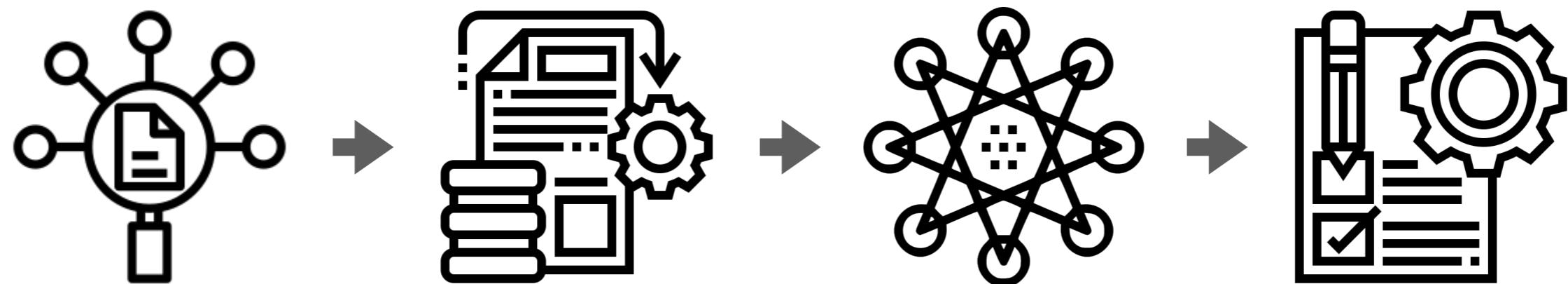


군집화 (Clustering)



기계학습 절차

| 지도학습 기반 기계학습 모델 생성절차



(2) 기계학습 절차: 데이터 준비

자질추출 (Feature Extraction)



- 기계학습에 필요한 자질(변수, feature)을 추출하고 이를 수치로 표현하는 방법
- 과거 기계학습의 성능은 동일한 알고리즘에 대해 특징을 추출하는 방법에 의해 좌우되었으나, 딥러닝(deep learning)의 등장으로 자질을 추출하는 전처리 과정이 자동화되어 거의 사라짐

구분	메시지	특징 (Feature)						
		메시지 길이	URL 여부	특수문자 개수	해외수신 여부	의심단어 개수	광고문자 표시여부	
1	[국제발신] 하루 30만 달 1천만 만원으로 이렇게! http://bit.ly/3f~	40	1	6	1	1	0	
2	팀장님 이보람 선임입니다. 출근하시면 결재 부탁드립니다.	20	0	1	0	0	0	
3	(광고)웰컴박하라 vc⑤47③.co 코드 wc1004 무료수신거부 01084510000	45	1	5	0	1	1	
4	(광고)신한과 함께하는 소중한 미래 따뜻한 금융 [신한]입니다. 2019년에 힘들었던 모든 일들은 다 잊어버리시고, ~	80	0	4	0	0	1	
5	[WEB발신] 갤럭시 노트20/노트20 울트라 사전예약 오늘이 마지막날입니다-!! 구매를 망설이고 ~	75	0	6	0	0	0	
6	(광고)등촌역스톤힐 ★더블역세권 9호선 ~ ★선착순으로 동호수 지정분양 가능 ★인근주변 아파트 시세보다 4~5억 저렴 ★~	120	0	8	0	0	1	
7	[한진택배] 상품 배송 안내 안녕하세요 고객님. ★상품 수령이 편하신 장소를 선택 ~ ①직접수령 ②경비실 ③문앞 ~	110	0	9	0	0	0	

(2) 기계학습 절차: 데이터 준비

레이블링 (Labeling)

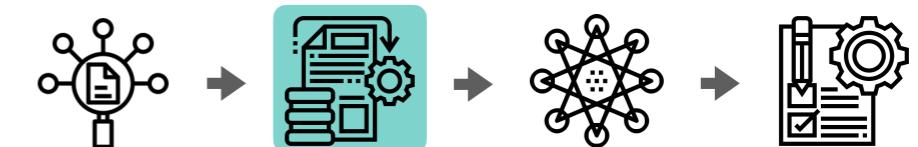


- 준비된 학습데이터에 지도학습을 위한 라벨(label)을 부착하는 과정
- 지도학습을 활용하는 경우, 학습데이터의 양과 레이블링의 정확도가 모델의 성능에 큰 영향을 미칠 수 있음
- 효율적으로 라벨을 부착하는 방법을 찾는 것도 데이터 분석 준비 과정 중 매우 중요한 요소로 작용함

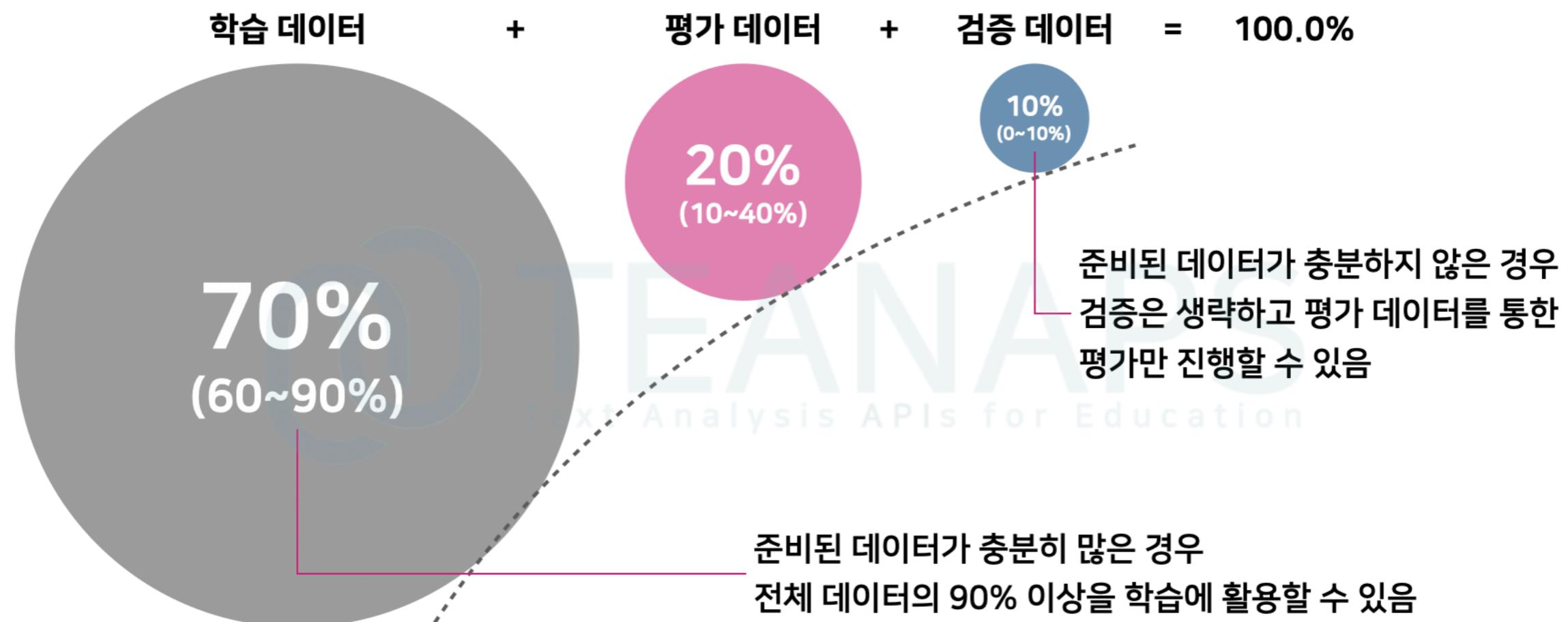
구분	메시지	특징 (Feature)						라벨 (Label)
		메시지 길이	URL 여부	특수문자 개수	해외수신 여부	의심단어 개수	광고문자 표시여부	
1	[국제발신] 하루 30만 달 1천만 만원으로 이렇게! http://bit.ly/3f~	40	1	6	1	1	0	TRUE
2	팀장님 이보람 선임입니다. 출근하시면 결재 부탁드립니다.	20	0	1	0	0	0	FALSE
3	(광고)웰컴박하라 vc⑤47③.co 코드 wc1004 무료수신거부 01084510000	45	1	5	0	1	1	TRUE
4	(광고)신한과 함께하는 소중한 미래 따뜻한 금융 [신한]입니다. 2019년에 힘들었던 모든 일들은 다 잊어버리시고, ~	80	0	4	0	0	1	TRUE
5	[WEB발신] 갤럭시 노트20/노트20 울트라 사전예약 오늘이 마지막날입니다-!! 구매를 망설이고 ~	75	0	6	0	0	0	TRUE
6	(광고)등촌역스톤힐 ★더블역세권 9호선 ~ ★선착순으로 동호수 지정분양 가능 ★인근주변 아파트 시세보다 4~5억 저렴 ★~	120	0	8	0	0	1	TRUE
7	[한진택배] 상품 배송 안내 안녕하세요 고객님. ★상품 수령이 편하신 장소를 선택 ~ ①직접수령 ②경비실 ③문앞 ~	110	0	9	0	0	0	FALSE

(2) 기계학습 절차: 데이터 준비

데이터 분리 (Partitioning)



- 효율적인 학습과 평가를 위해, 준비된 데이터를 학습 데이터, 평가 데이터, 검증 데이터로 분리하는 과정
- **학습 데이터** (training data) : 기계학습 모델의 학습을 위한 데이터, 양은 많을수록 좋음
- **평가 데이터** (test data) : 모델의 학습결과 성능을 평가하기 위한 데이터
- **검증 데이터** (validation data) : 평가를 마친 모델에 대해 마지막 검증을 수행하기 위한 데이터

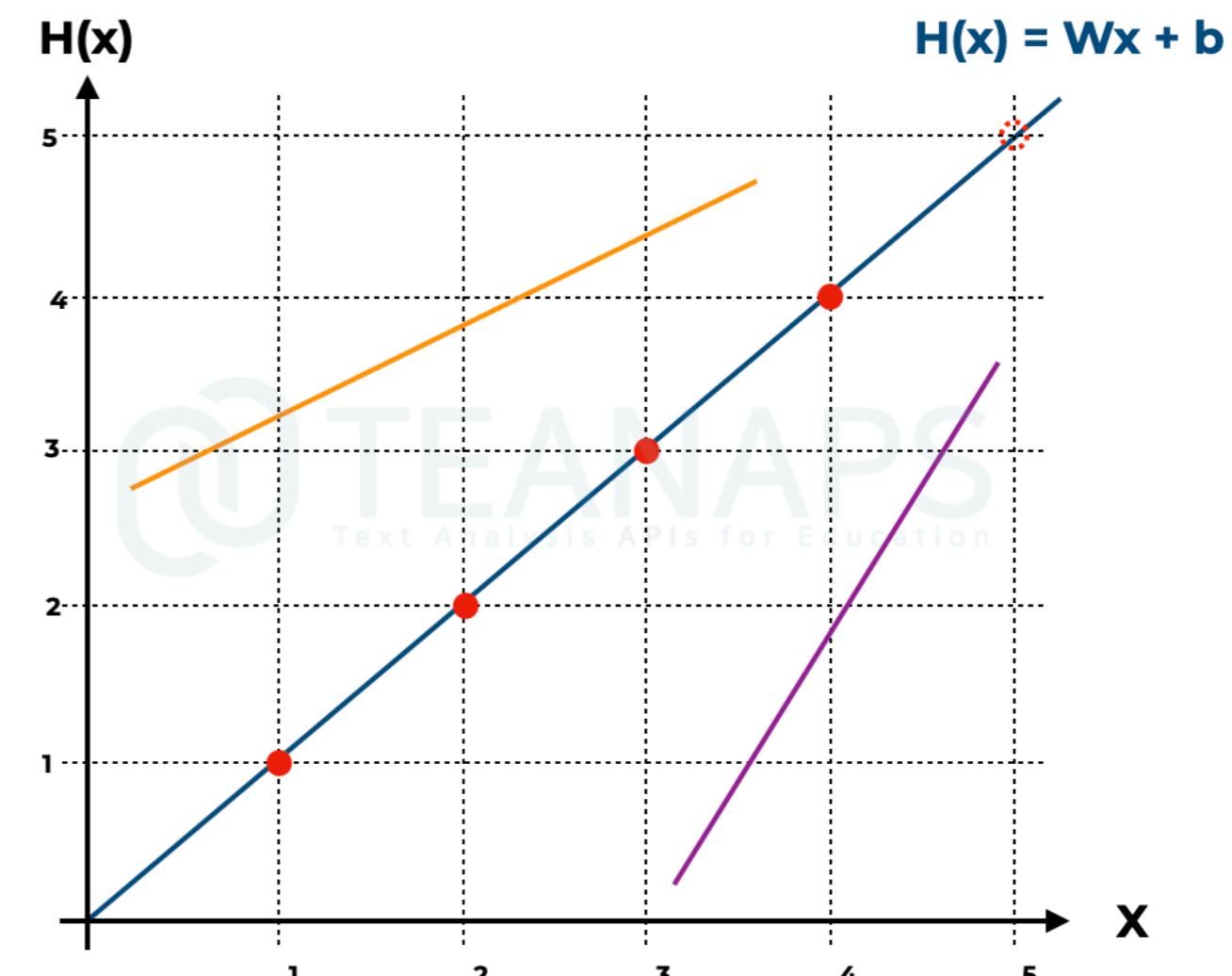
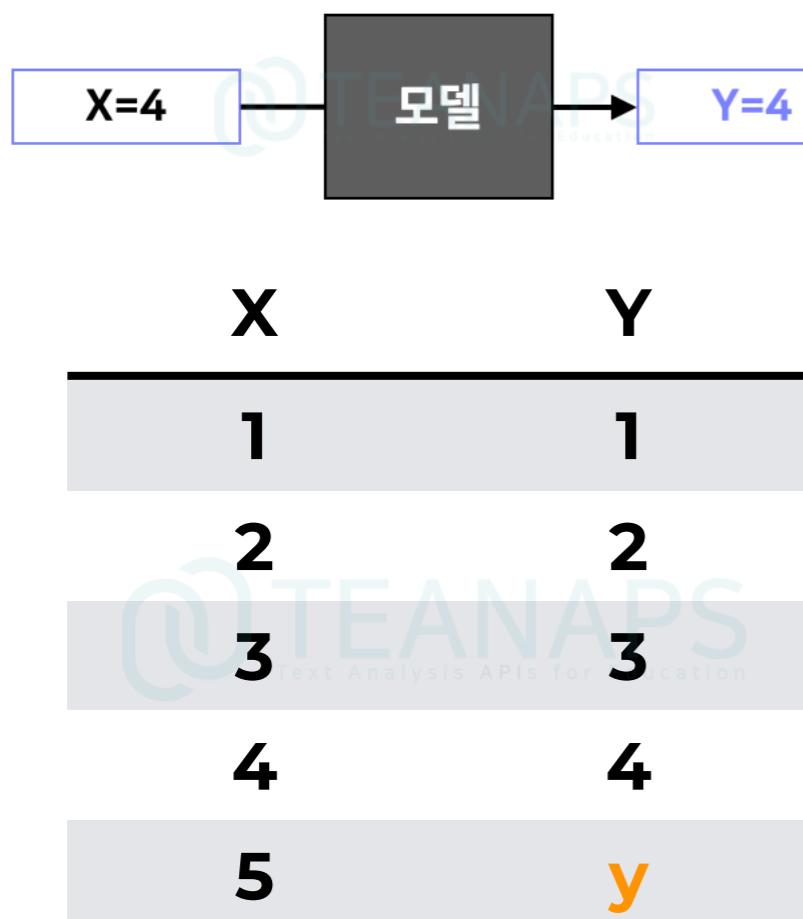


기계학습 절차: (3) 학습 (Training)

기계가 데이터를 학습하는 과정 (Machine Training)



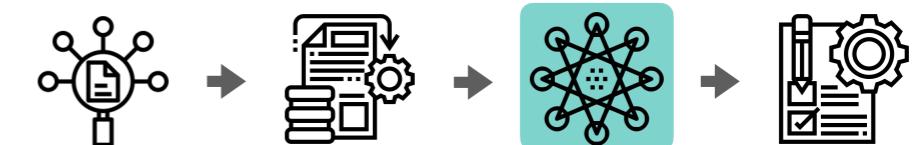
- 학습데이터에 정의된 정보나 규칙을 추상적인 형태로 표현하는 모델을 생성하는 과정
- 학습데이터에 포함된 다양한 정보나 규칙을 모델이 얼마나 잘 표현하는가에 따라 머신러닝 모델의 성능이 좌우됨
- **선형가정** (Linear Hypothesis) : 학습데이터의 분포를 선형이라 가정하고 학습데이터를 가장 잘 설명한 직선



기계학습 절차: (3) 학습

(Training)

파라미터 접근법 (Parametric Approach)



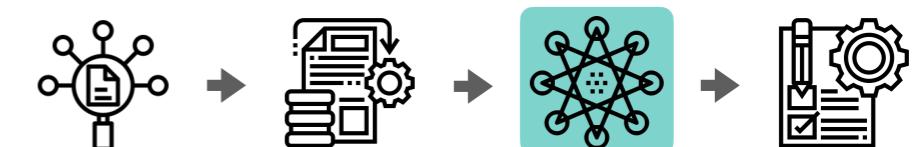
- 입력변수(x)와 목표변수(y) 사이의 복잡한 관계를 어떠한 파라미터 (w)와의 관계로 표현하는 방식
- 정답을 구하기 위한 적절한 파라미터 (w)를 구하고 예측된 값 (y_n)과 정답 (Y)와의 차이 (error, loss)를 계산하여 그 평균을 최소로하는 적절한 파라미터 (w)를 도출하는 과정

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

a	x₁	b	x₂	y_n	Y	Y - y_n
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 _n)	

기계학습 절차: (3) 학습 (Training)

| **파라미터 접근법** (Parametric Approach)

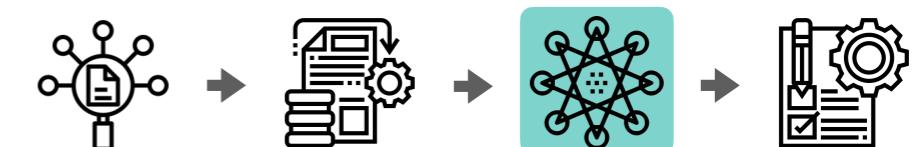


$$y = ax_1 + bx_2$$

a	x ₁	b	x ₂	y _n	Y	Y - y _n
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 _n)	

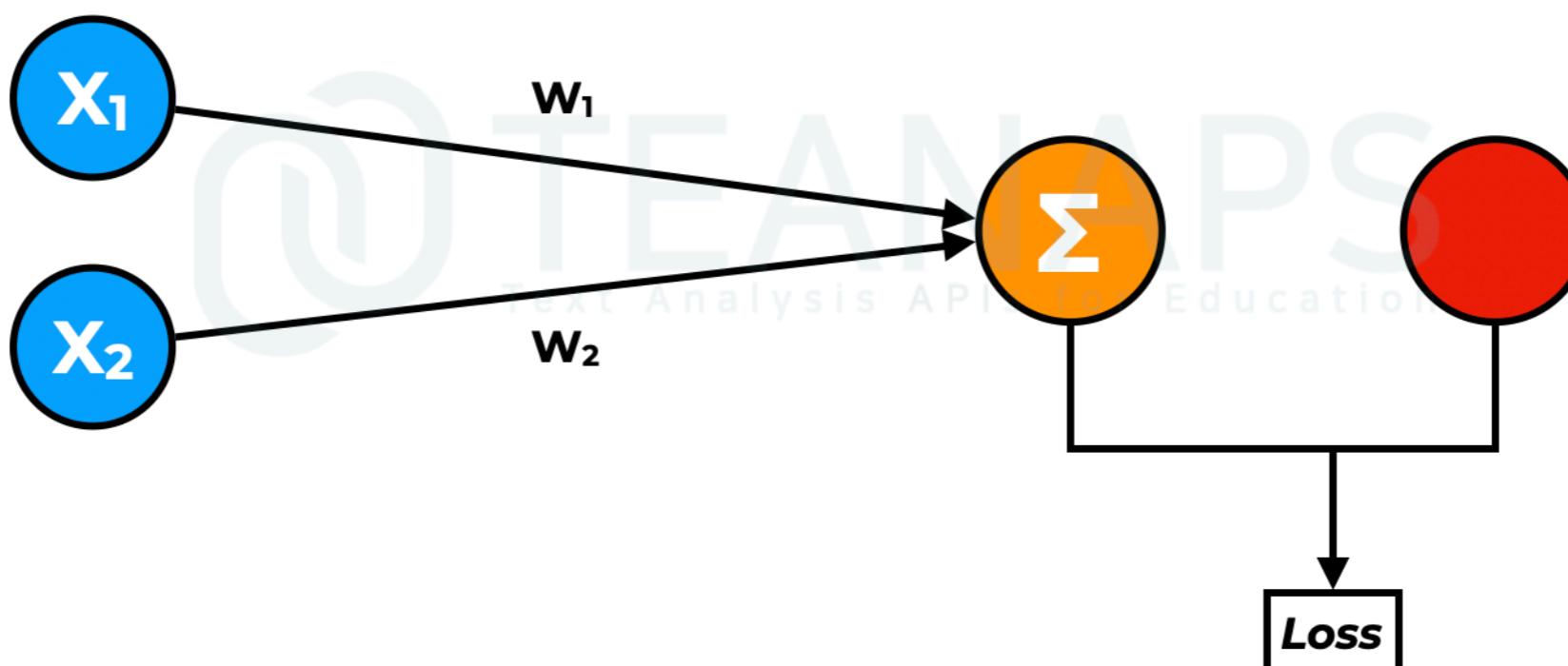
기계학습 절차: (3) 학습 (Training)

| **파라미터 접근법** (Parametric Approach)



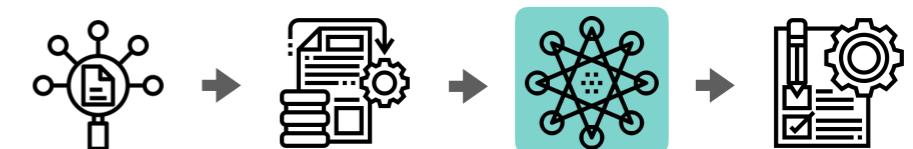
$$Y_n = W_1 \cdot X_1 + W_2 \cdot X_2$$

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

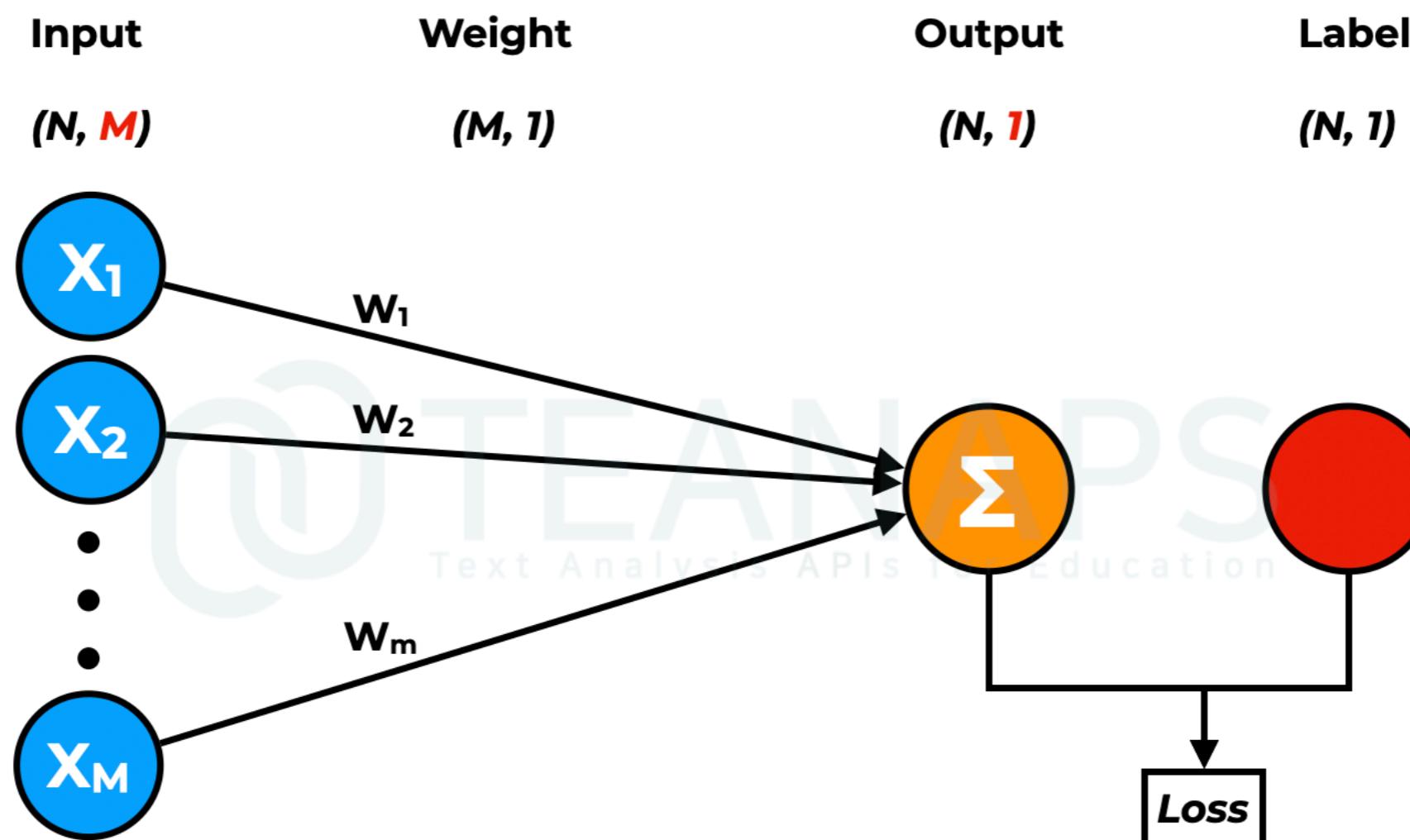


기계학습 절차: (3) 학습 (Training)

| **파라미터 접근법** (Parametric Approach)

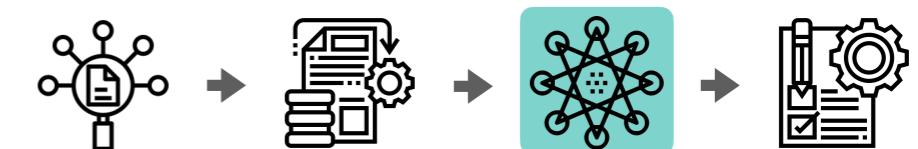


$$Y_n = W_1 \cdot X_1 + W_2 \cdot X_2 + \dots + W_m \cdot X_m$$

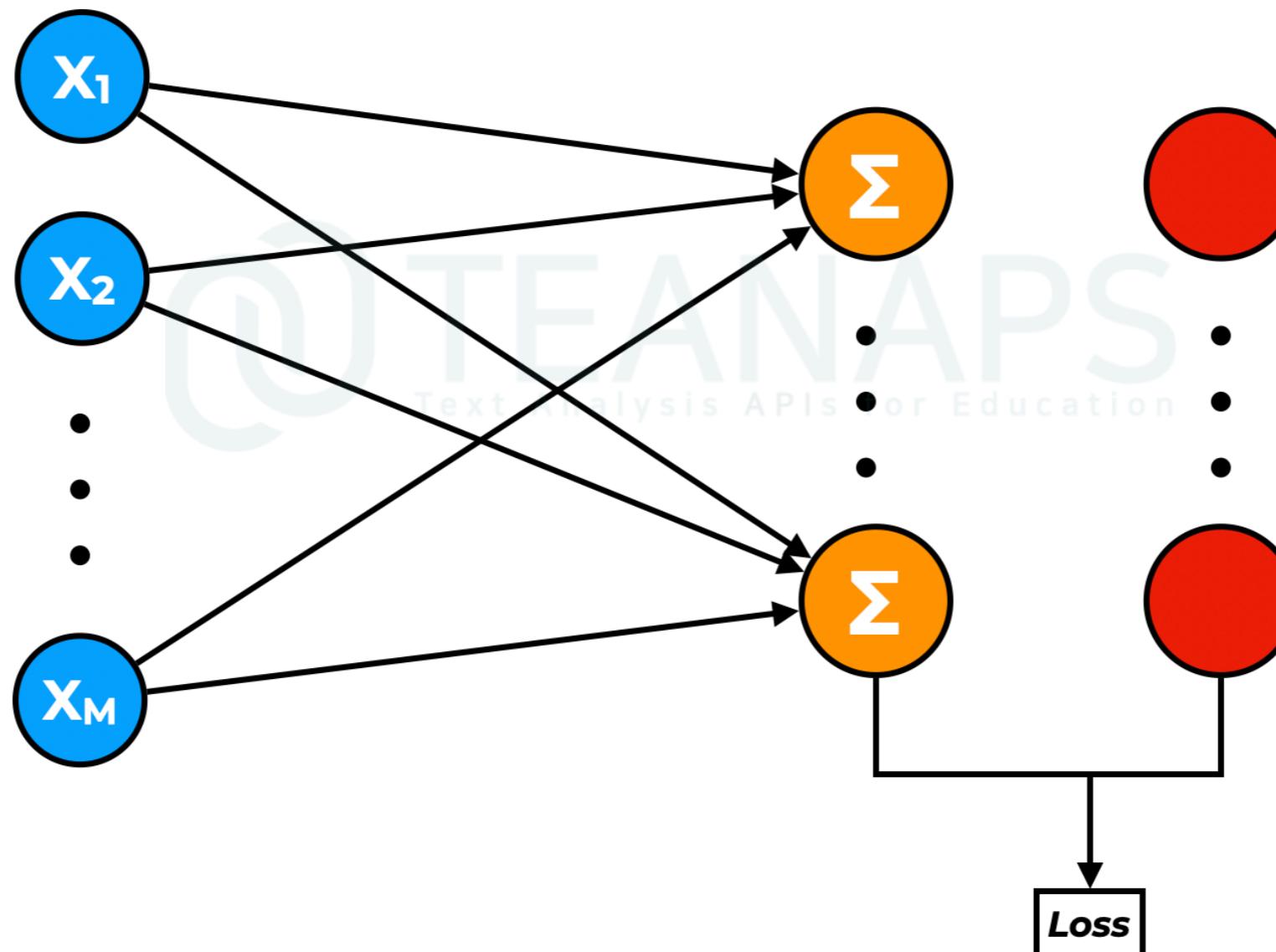


기계학습 절차: (3) 학습 (Training)

파라미터 접근법 (Parametric Approach)



Input	Weight	Output	Label
(N, M)	(M, T)	(N, T)	(N, T)

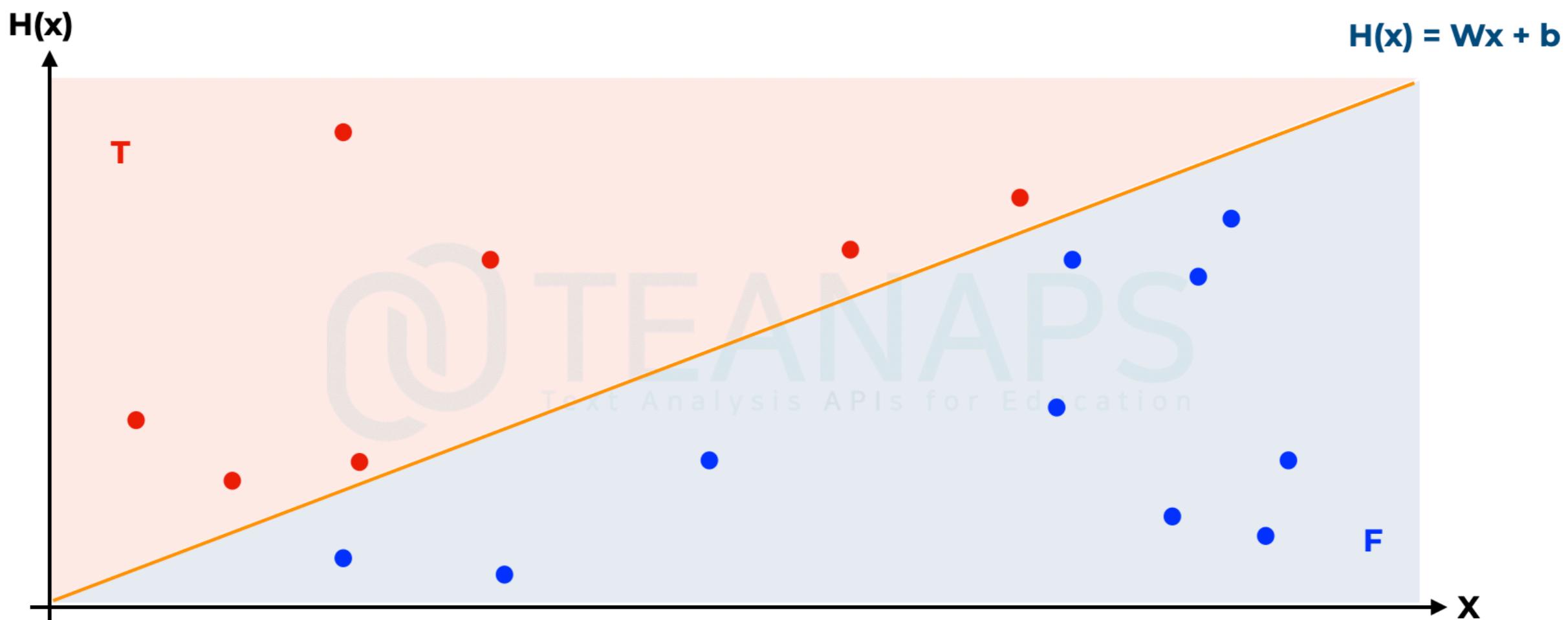


기계학습 절차: (3) 학습 (Training)

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)



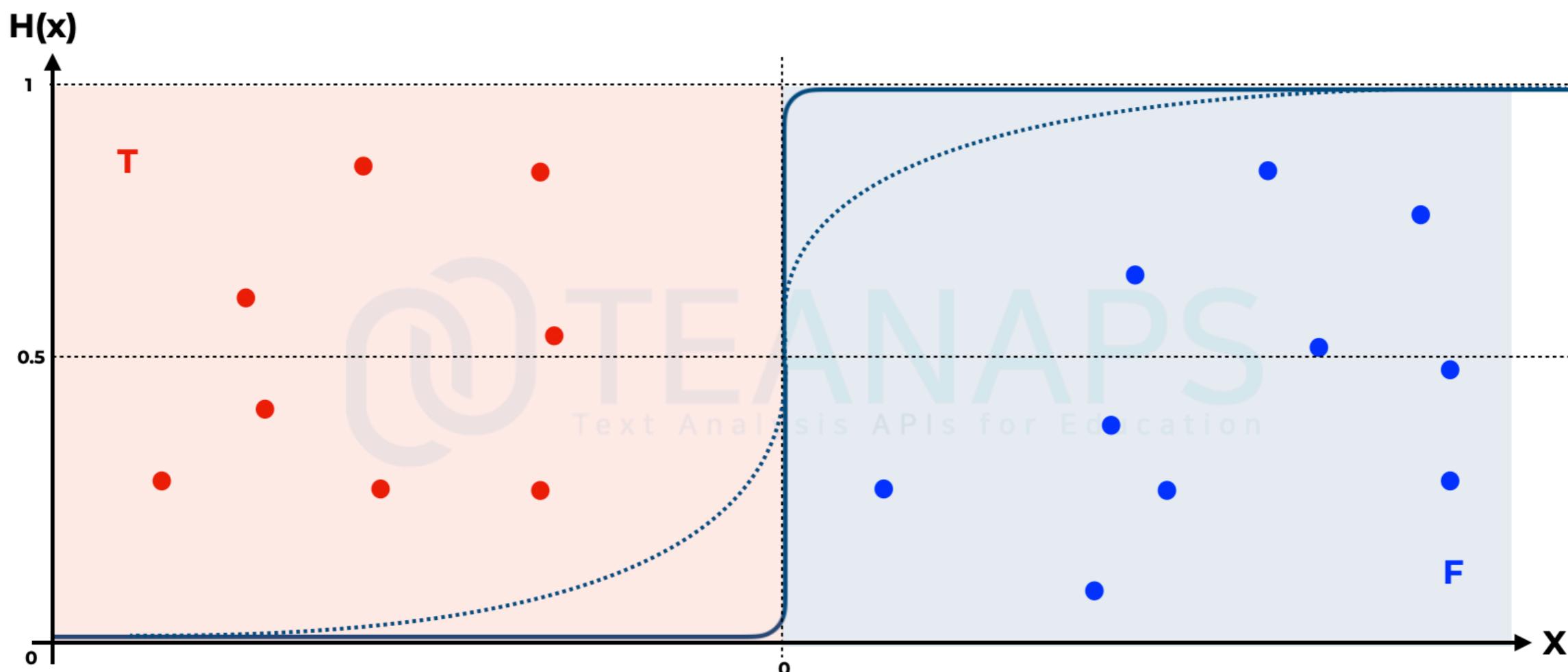
기계학습 절차: (3) 학습 (Training)

| 시그모이드 함수를 통한 로짓변환



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

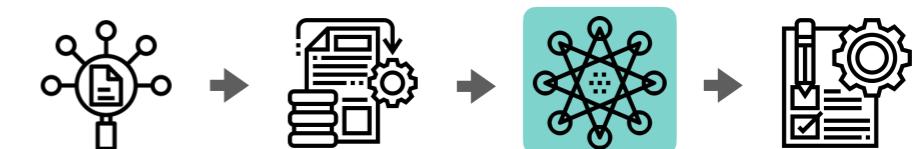
$$H(x) = Wx + b \rightarrow \text{Sigmoid}(H(x)) = \frac{1}{(1 + e^{-H(x)})}$$



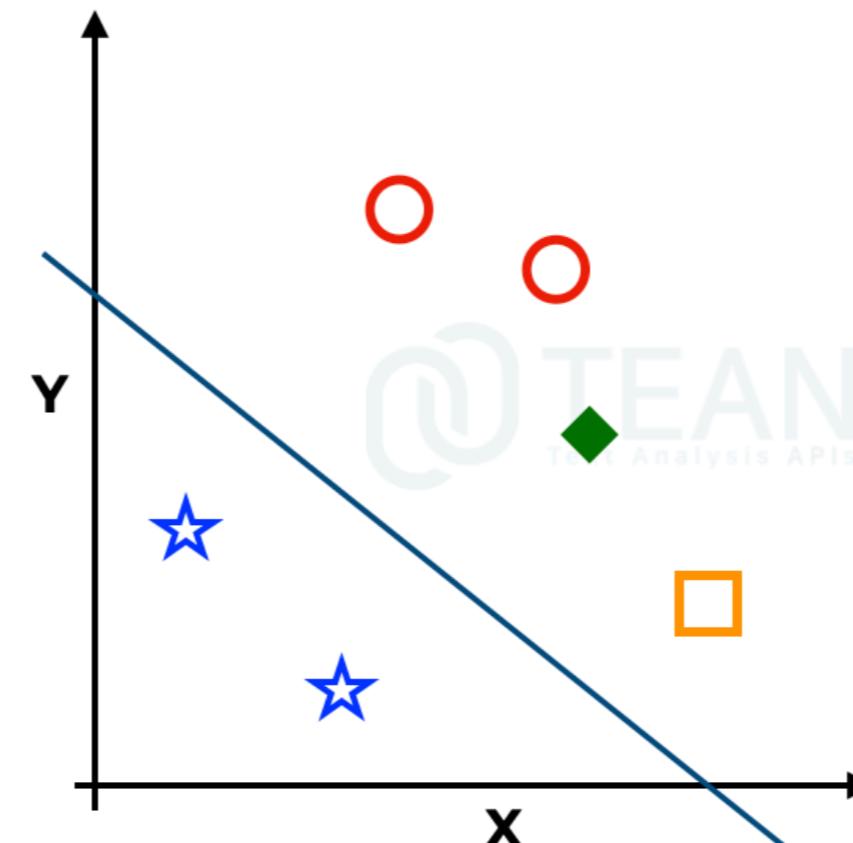
기계학습 절차: (3) 학습 (Training)

다차원 분류 (Multinomial Classification)

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



X ₁	X ₂	Y
10	5	A
9	5	A
3	2	B
2	4	B
11	1	C



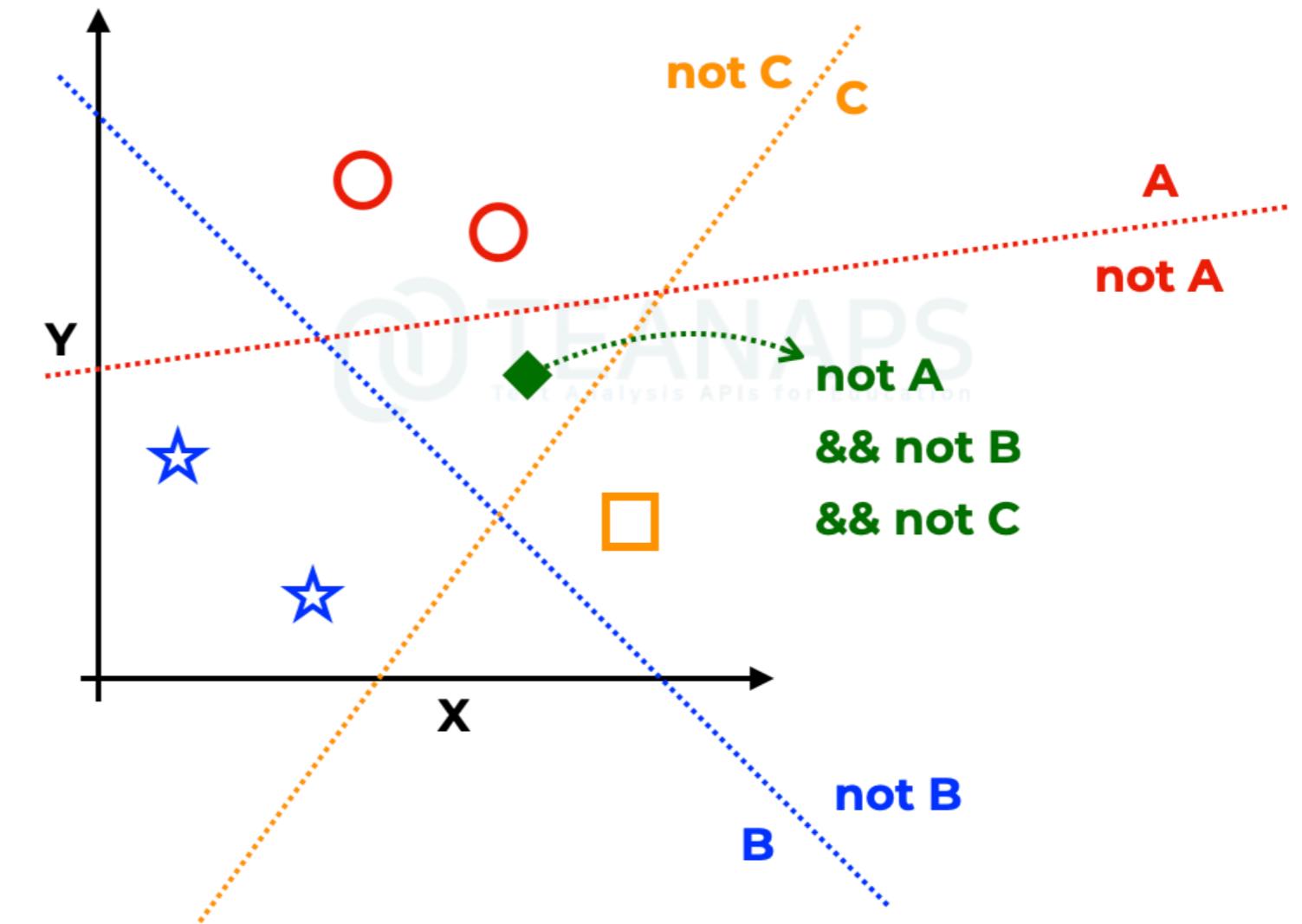
기계학습 절차: (3) 학습 (Training)

다차원 분류 (Multinomial Classification)

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



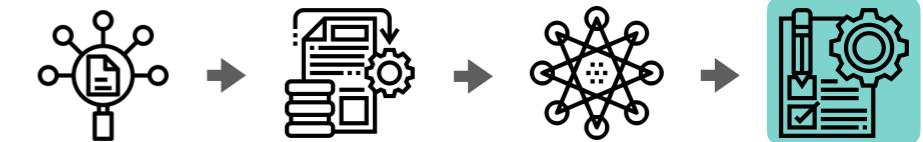
X ₁	X ₂	Y
10	5	A
9	5	A
3	2	B
2	4	B
11	1	C



기계학습 절차: (4) 평가

(Validation)

2진분류 (Binary Classification) 모델을 평가하는 방법



- 모델의 분류 결과와 실제 정답과의 비교를 통해 모델의 성능을 평가할 수 있음
- **정확도 (Accuracy)** : 전체 분류결과 중 정답과 일치하는 분류결과의 비율
- **재현율 (Recall)** : 정답이 TRUE인 경우의 수 중 모델이 정답과 동일하게 분류한 결과의 비율
- **정밀도 (Precision)** : 모델이 TRUE로 분류한 결과 중 정답과 일치하는 분류결과의 비율
- **F1-score** : 재현율과 정밀도의 조화평균으로 두 가지 평가지표의 특성을 균등하게 반영할 수 있음

		정답	
		TRUE	FALSE
분류 결과	TRUE	TRUE Positive (TP)	FALSE Positive (FP)
	FALSE	FALSE Negative (FN)	TRUE Negative (TN)

1종오류 (Type1 Error)
 2종오류 (Type2 Error)

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

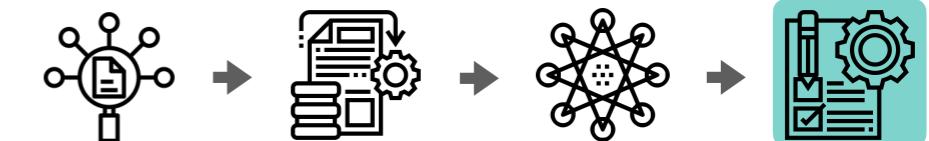
$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{F1-score} = \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

기계학습 절차: (4) 평가

(Validation)

다중분류 (Multi-class Classification) 모델을 평가하는 방법



- 다중분류 모델의 각 범주(Class)에 대해 2진분류와 동일한 평가를 진행하고, 각 범주 별 평가 결과의 평균값을 통해 전체 모델의 성능을 평가할 수 있음

정답			정답			정답			정답			
	A	Not A	B	TP	FP	C	TP	FP	D	TP	FP	
분류	A	TP	FP	B	TP	FP	C	TP	FP	D	TP	FP
결과	Not A	FN	TN	Not B	FN	TN	Not C	FN	TN	Not D	FN	TN
범주			Accuracy			Recall			Precision			
Class A			89			88			92			
Class B			100			100			100			
Class C			83			84			78			
Class D			65			62			67			
최종 모델평가			84.25			83.5			84.25			

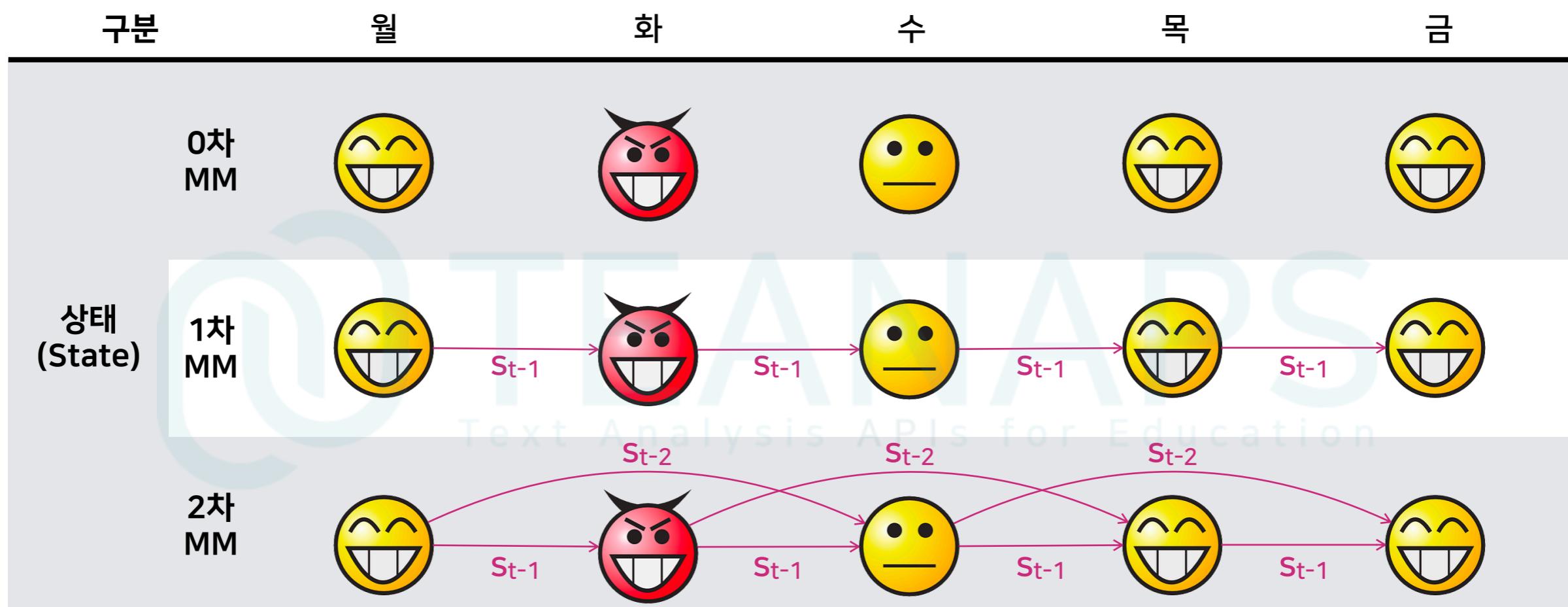
HOW?

자연어처리 대부분의 문제는
순서에 따른 분류 문제!!!

순차 레이블링 (Sequence Labeling)

마르코프 모델 (Markov Model, Markov Chain)

- 순차적으로 출현하는 상태를 예측하는 문제에서 과거와 현재의 상태(state)가 주어졌을 때, 미래 상태가 과거에 출현한 N개 상태에서만 영향을 받는다는 가정을 바탕으로 한 순차적 확률분포 모델
- 0차 마르코프 가정 (Markov Assumption) : $P(S_t|S_{t-N}, \dots, S_{t-2}, S_{t-1}) = P(S_t)$
- 1차 마르코프 가정 (1st Order Markov Assumption, bigram model) : $P(S_t|S_{t-N}, \dots, S_{t-2}, S_{t-1}) = P(S_t|S_{t-1})$
- 2차 마르코프 가정 (2nd Order Markov Assumption) : $P(S_t|S_{t-N}, \dots, S_{t-2}, S_{t-1}) = P(S_t|S_{t-1}, S_{t-2})$

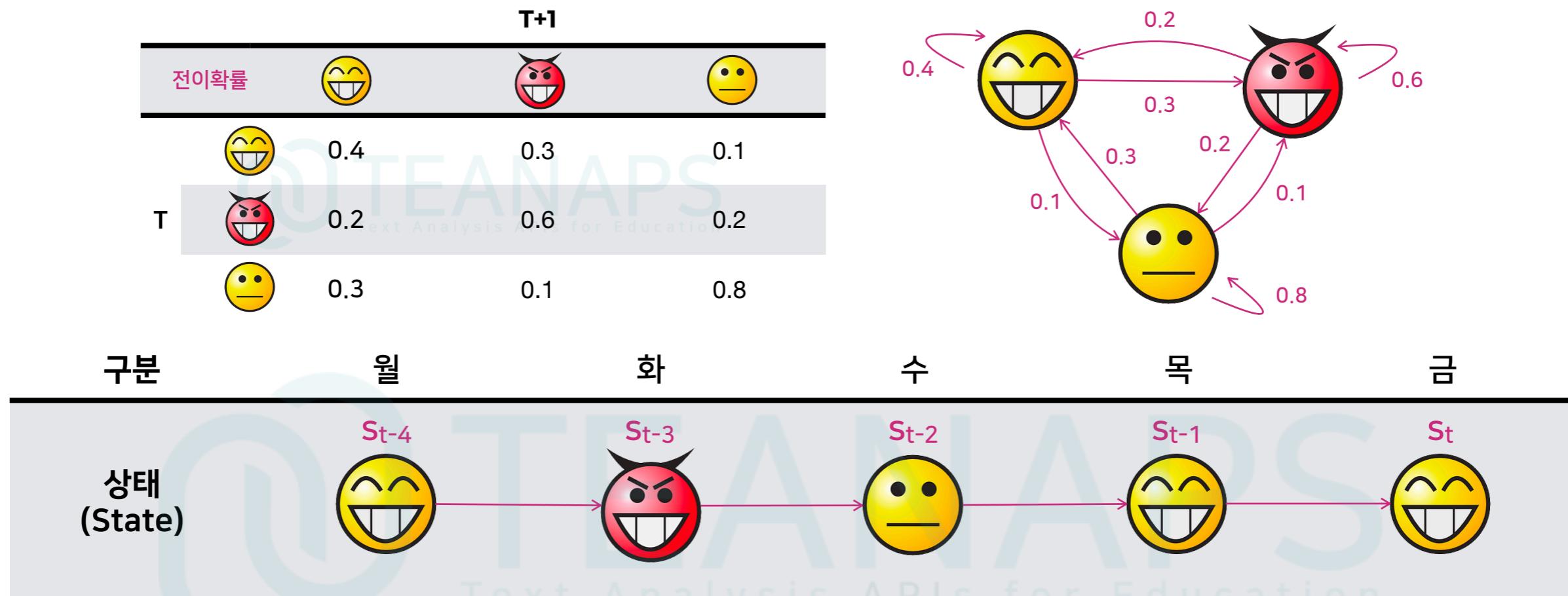


순차 레이블링

(Sequence Labeling)

은닉 마르코프 모델 (Hidden Markov Model, HMM)

- 순차적으로 출현하는 상태를 예측하는 문제에서 상태를 직접 관찰할 수 없는 경우, 미래 상태가 오직 상태의 영향을 받는 과거에 관찰된 N개 현상에서만 영향을 받는다(독립가정)는 가정을 바탕으로 한 순차적 확률분포 모델



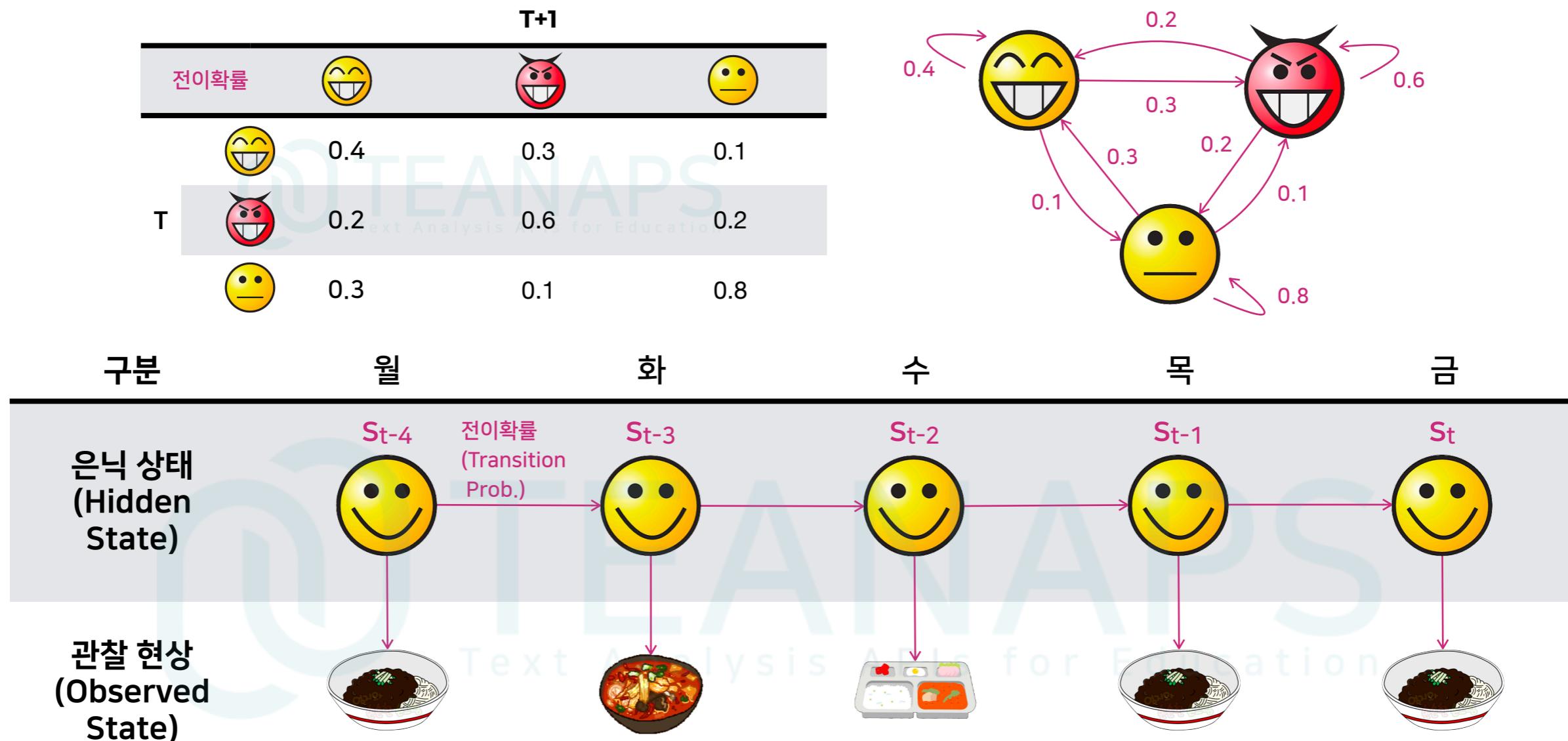
$$\begin{aligned}
 P(S_t | S_{t-4}, S_{t-3}, S_{t-2}, S_{t-1}) &= \prod P(S_{t-4}) P(S_{t-3} | S_{t-4}) P(S_{t-2} | S_{t-3}) P(S_{t-1} | S_{t-2}) P(S_t | S_{t-1}) \\
 &= \prod 0.4 \times 0.3 \times 0.2 \times 0.3 \times 0.4 = 0.00288\pi = 2.88 \times 10^{-3}\pi \quad (1\text{차 마르코프 가정})
 \end{aligned}$$

순차 레이블링

(Sequence Labeling)

은닉 마르코프 모델 (Hidden Markov Model, HMM)

- 순차적으로 출현하는 상태를 예측하는 문제에서 상태를 직접 관찰할 수 없는 경우, 미래 상태가 오직 상태의 영향을 받는 과거에 관찰된 N개 현상에서만 영향을 받는다(독립가정)는 가정을 바탕으로 한 순차적 확률분포 모델

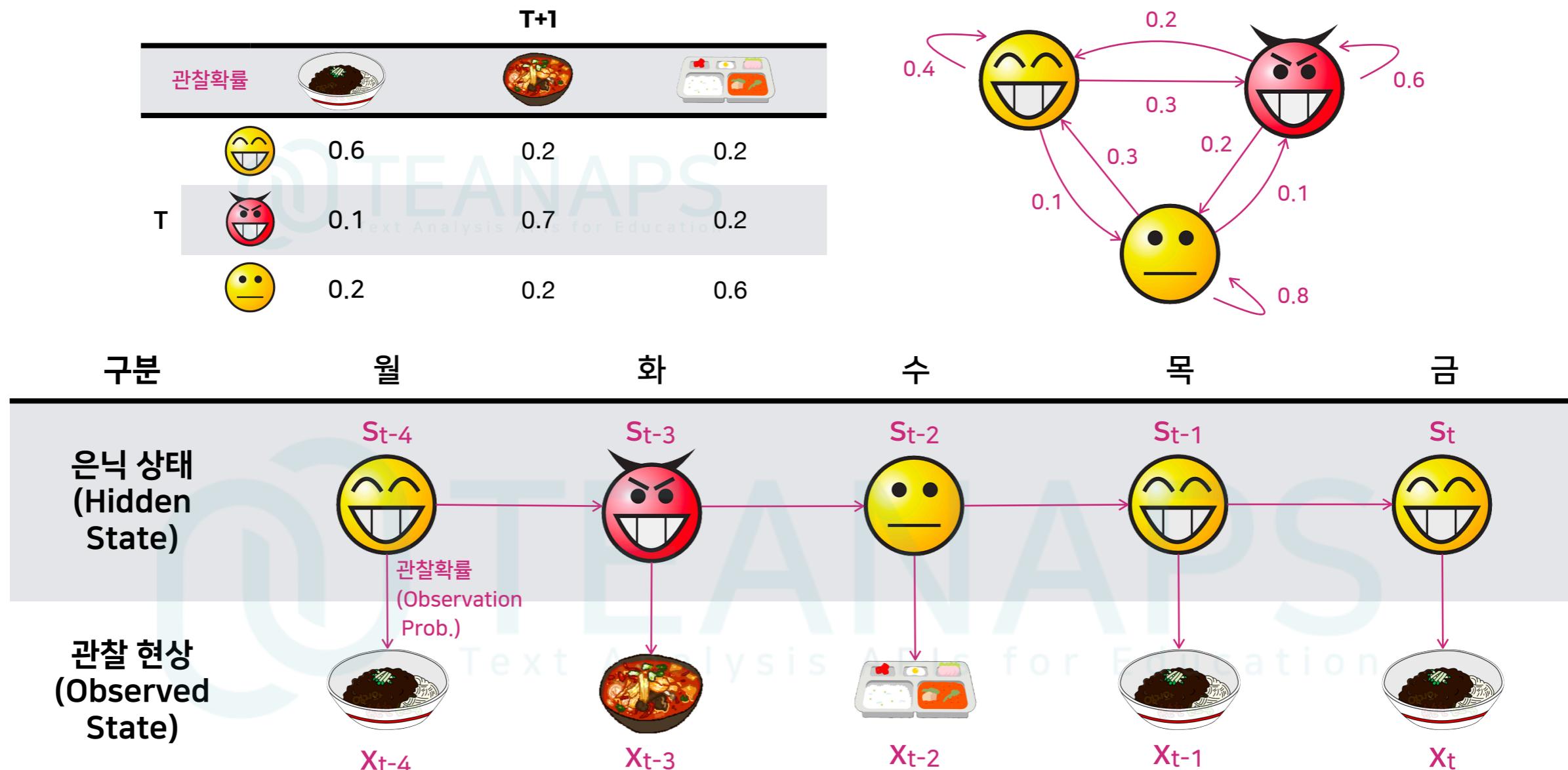


순차 레이블링

(Sequence Labeling)

은닉 마르코프 모델 (Hidden Markov Model, HMM)

- 순차적으로 출현하는 상태를 예측하는 문제에서 상태를 직접 관찰할 수 없는 경우, 미래 상태가 오직 상태의 영향을 받는 과거에 관찰된 N개 현상에서만 영향을 받는다(독립가정)는 가정을 바탕으로 한 순차적 확률분포 모델



순차 레이블링

(Sequence Labeling)

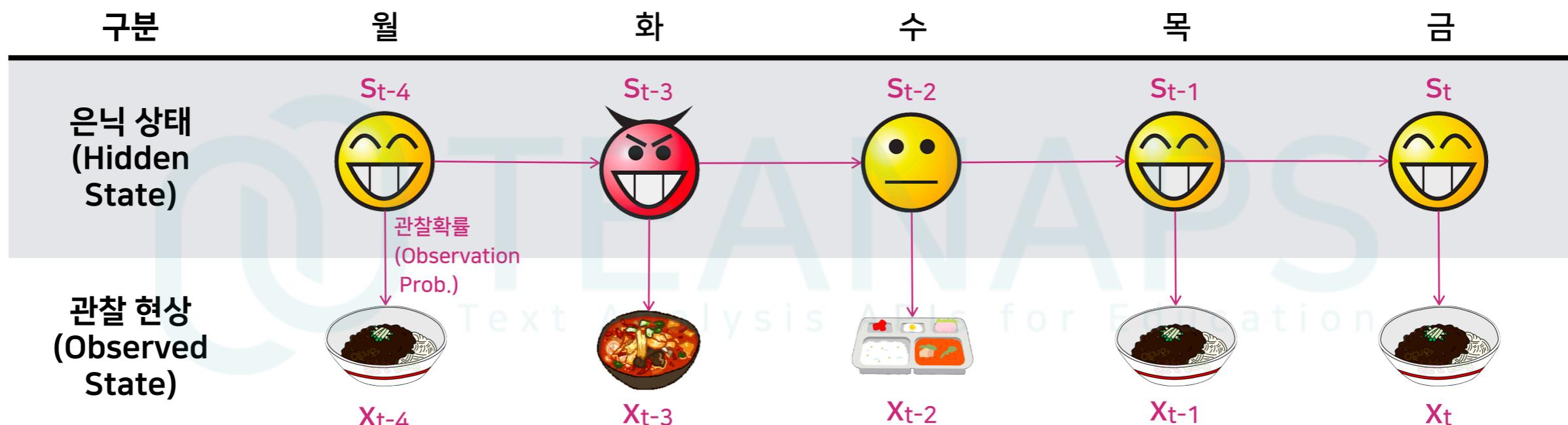
은닉 마르코프 모델 (Hidden Markov Model, HMM)

$$P_T = P(S_t | S_{t-N}, \dots, S_{t-2}, S_{t-1}) = P(S_t | S_{t-1}) \quad (\text{1차 마르코프 가정})$$

$$P_{O|T} = P(X_t | S_{t-N}, X_{t-N}, \dots, S_{t-2}, X_{t-2}, S_{t-1}, X_{t-1}) = P(X_t | S_t) \quad (\text{관찰확률 반영})$$

$$\begin{aligned} P_{HMM} &= \underset{s_{t-N,n}}{\operatorname{argmax}} P(x_{t-N,n}, s_{t-N,n}) = \underset{s_{t-N,n}}{\operatorname{argmax}} P(x_{t-N,n}, s_{t-N,n}) P(s_{t-N,n}) \\ &\approx \underset{s_{t-N,n}}{\operatorname{argmax}} \sum_{t=1}^n P(x_t | s_t) \underset{\substack{\text{관찰확률} \\ \text{초기확률}}}{P(s_t | s_{t-1})} \quad (\text{1차 마르코프 가정, 독립가정 반영}) \end{aligned}$$

Independance Assumption



순차 레이블링

(Sequence Labeling)

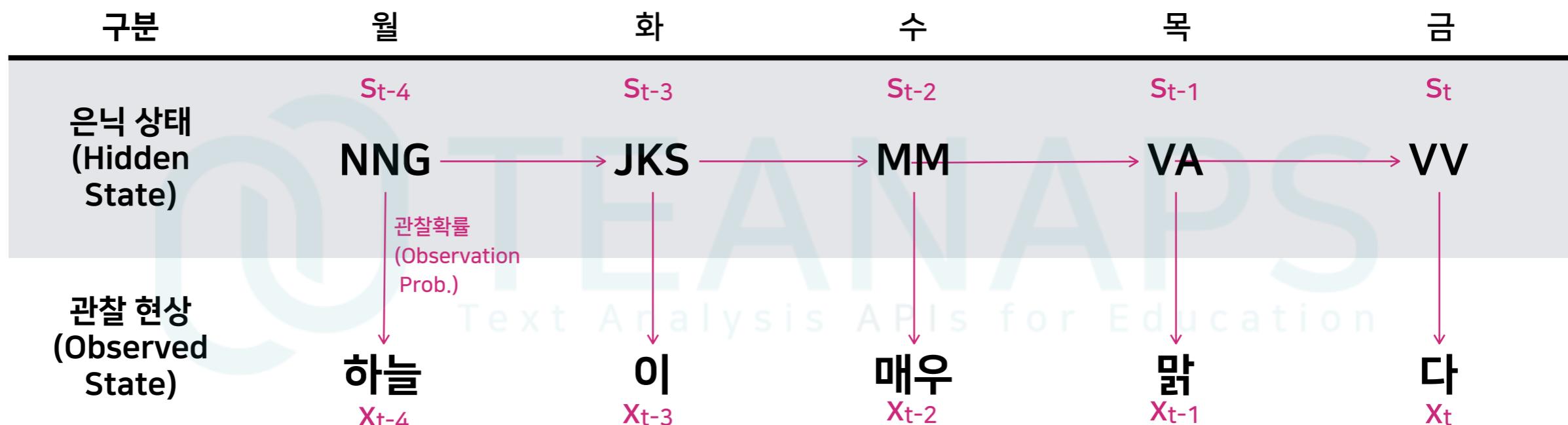
은닉 마르코프 모델 (Hidden Markov Model, HMM)

$$P_T = P(S_t | S_{t-N}, \dots, S_{t-2}, S_{t-1}) = P(S_t | S_{t-1}) \quad (\text{1차 마르코프 가정})$$

$$P_{O|T} = P(X_t | S_{t-N}, X_{t-N}, \dots, S_{t-2}, X_{t-2}, S_{t-1}, X_{t-1}) = P(X_t | S_t) \quad (\text{관찰확률 반영})$$

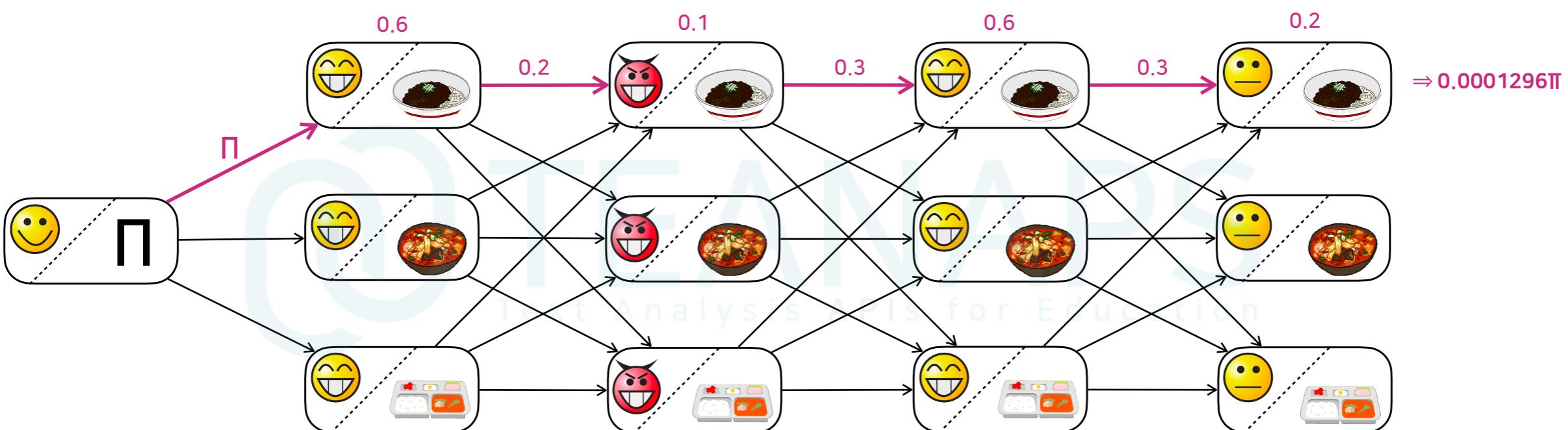
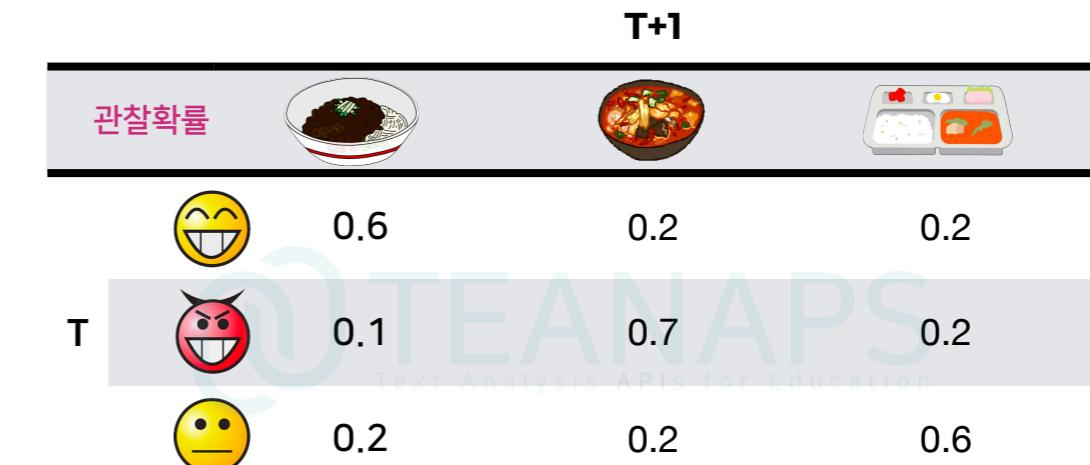
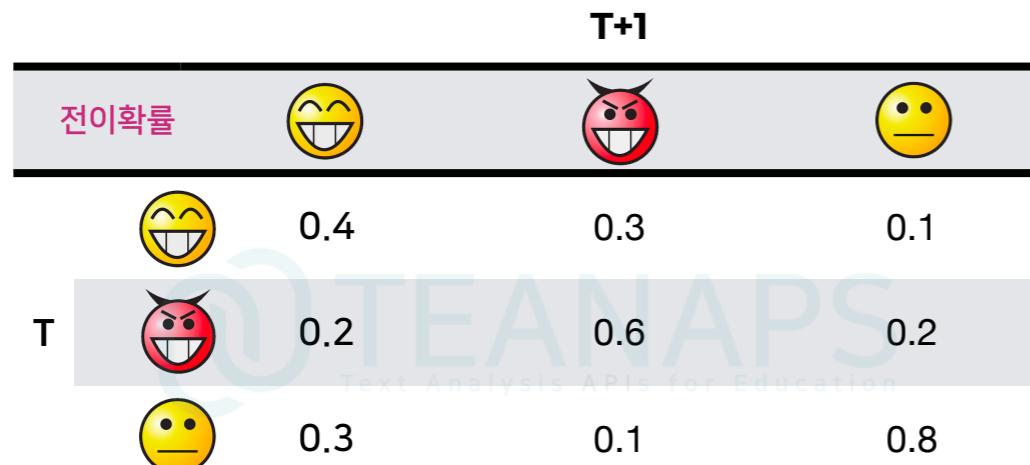
$$\begin{aligned} P_{HMM} &= \underset{S_{t-N,n}}{\operatorname{argmax}} P(x_{t-N,n}, S_{t-N,n}) = \underset{S_{t-N,n}}{\operatorname{argmax}} P(x_{t-N,n}, S_{t-N,n}) P(S_{t-N,n}) \\ &\approx \underset{S_{t-N,n}}{\operatorname{argmax}} \sum_{t=1}^n P(X_t | S_t) \underset{\substack{\text{관찰확률} \\ \text{초기확률}}}{P(S_t | S_{t-1})} \quad (\text{1차 마르코프 가정, 독립가정 반영}) \end{aligned}$$

Independance Assumption



순차 레이블링 (Sequence Labeling)

순차 레이블링 경로 분석 (Sequence Labeling Path Analysis)



순차 레이블링

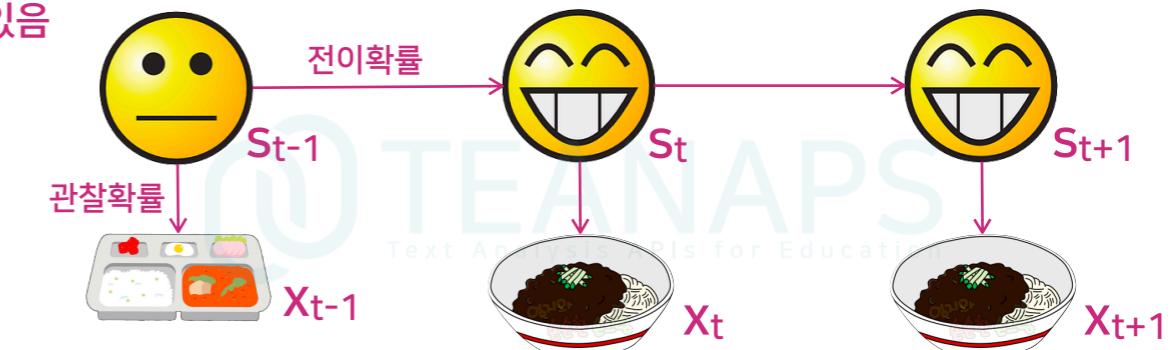
(Sequence Labeling)

HMM & MEMM (Maximum Entropy MM) & CRFs (Conditional Random Fields)

(HMM) "독립가정"에 의해 단어-단어 간 관계(문맥)을 반영하지 못한다는 단점이 있음

$$P_{\text{HMM}} \approx \operatorname{argmax} \prod_{t=1}^n P(x_t|s_t) P(s_t|s_{t-1})$$

(1차 마르코프 가정, 독립가정 반영)



(MEMM) 현재 단어에서 추출된 자질(feature)로 상태(State)를 예측함 ('90 중반)

$$P_{\text{MEMM}}(y, x) = \frac{1}{Z(x)} \exp \left\{ \sum_{i=1}^n \lambda_i f_i(x, y) \right\}$$

$$Z(x) = \sum_y \exp \left\{ \sum_{i=1}^n \lambda_i f_i(x, y) \right\}$$

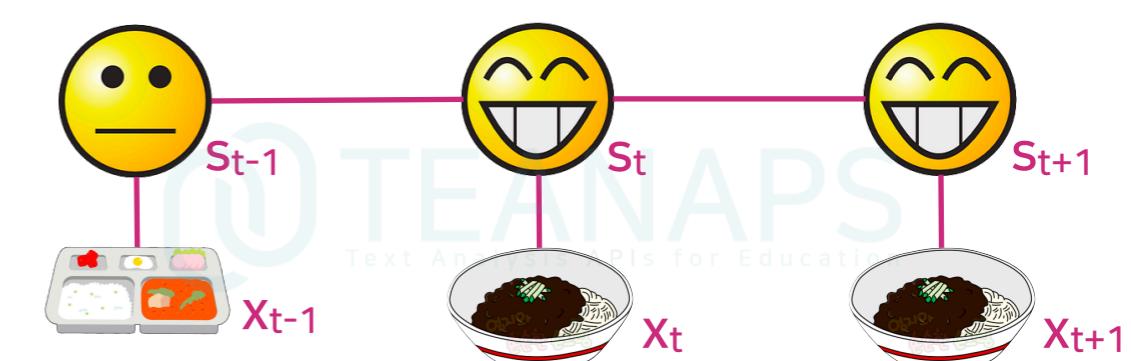
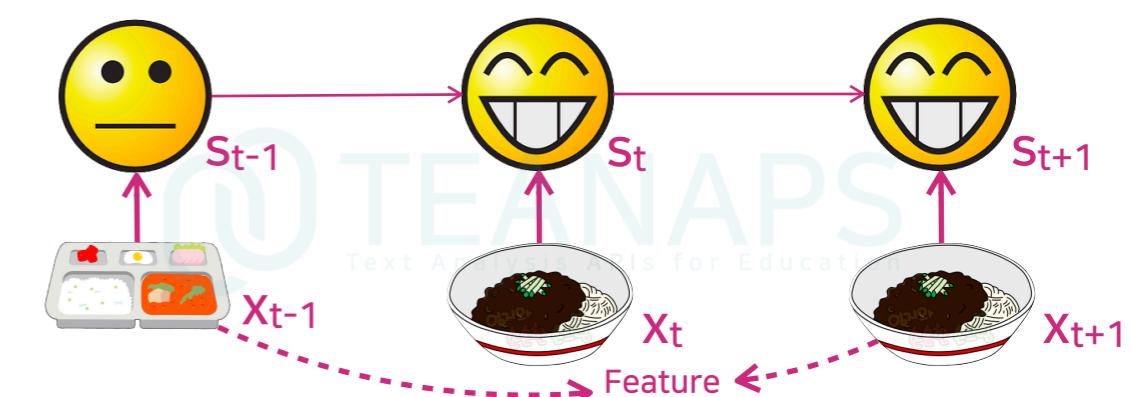
(CRFs) 전이확률을 계산할 때 앞뒤 순서를 모두 고려하여 상태(State)를 예측함 ('90 후반)

$$F_j(y, x) = \sum_{i=1}^n f_i(y_{i-1}, y_i, x, i)$$

$$P(y|x, \lambda) = \frac{1}{Z(x)} \exp \sum_j \lambda_j f_j(x, y)$$

Global Feature Function

Global Normalization



순차 레이블링: RNN

RNN을 활용한 순차 레이블링

구분

내용 (B: Beginner, I: Inner, O: Outer)

원문 손 흥 민 이 골 을 작 렬 하 며 토 트 넘 홋 스 퍼 의 승 리 를 이 꼴 었 다 .

띄어쓰기	손	흥	민	이	골	을	작	렬	하	며	토	트	넘	홋	스	퍼	의	승	리	를	이	꼴	었	다	.
------	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

토큰화	손흥민	이	골	을	작렬	하	며	토트넘	홋스퍼	의	승리	를	이끌	었	다	.
-----	-----	---	---	---	----	---	---	-----	-----	---	----	---	----	---	---	---

품사태깅	NNP	JKS	NNG	JKO	NNG	XSV	EC	NNG	NNG	JKG	NNG	JKO	VV	EP	EF	SF
------	-----	-----	-----	-----	-----	-----	----	-----	-----	-----	-----	-----	----	----	----	----

개체명인식	손	흥	민	이	골	을	작	렬	하	며	토	트	넘	홋	스	퍼	의	승	리	를	이	꼴	었	다	.
	B-PER	I-PER	I-PER	O	O	O	O	O	O	O	B-ORG	I-ORG	I-ORG	I-ORG	I-ORG	I-ORG	O	O	O	O	O	O	O	O	

E.O.D

Contact

-  <http://www.teanaps.com>
-  fingeredman@gmail.com