

데이터마케팅팀

4팀

황유나
문서영
김지현
위재성
이진모

CONTENTS

1. What is Data Mining?

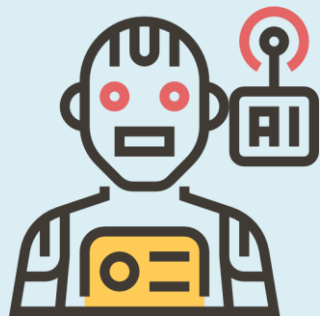
2. What is Modeling?

3. How to Avoid Overfitting

1

What is Data Mining?

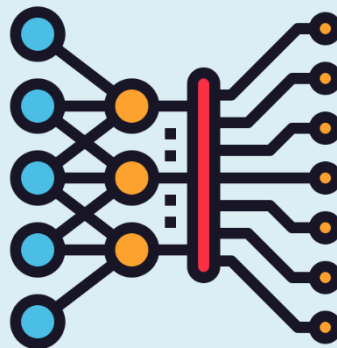
정의 및 접근법



AI



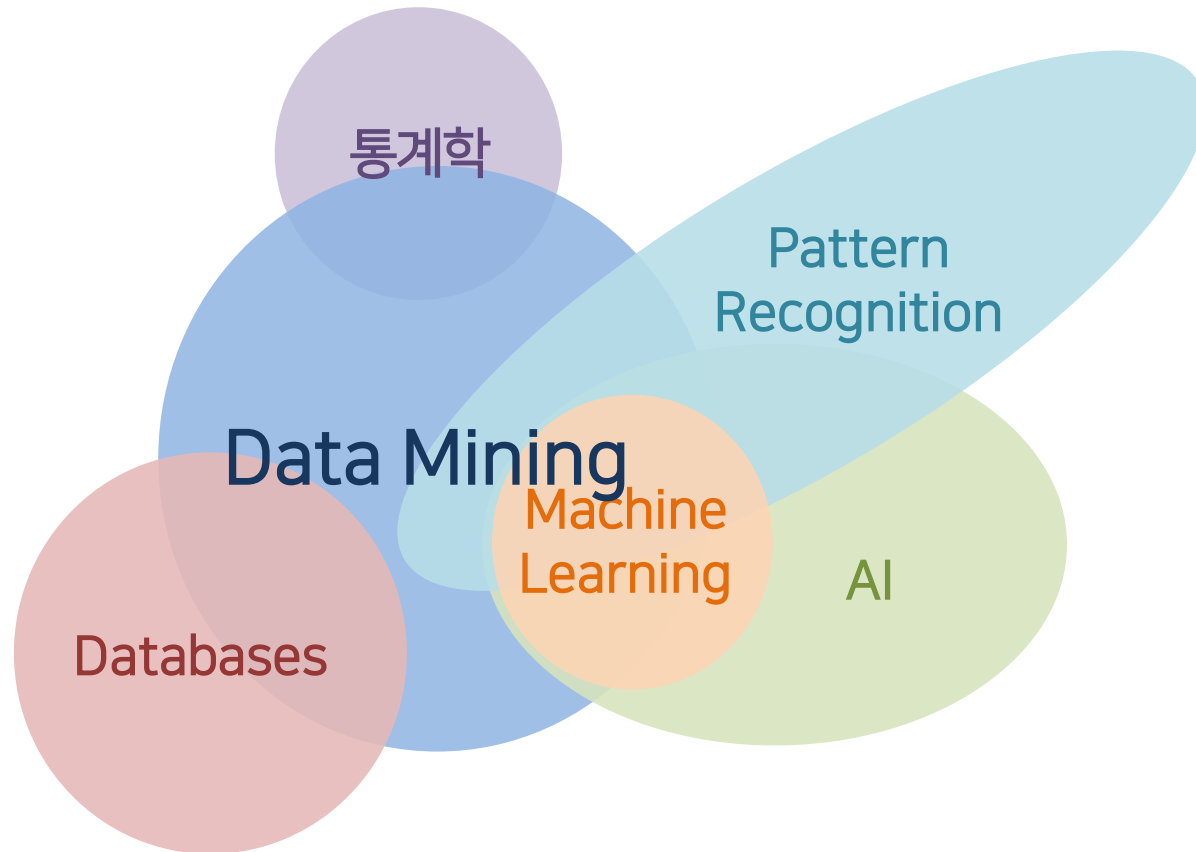
통계학



패턴 인식

정의 및 접근법

여러 학문과 밀접히 맞닿아있어!



정의 및 접근법

여러 학문과 밀접히 맞닿아있어!

Data Mining

통계학

Pattern
Recognition

Machine
Learning

AI



Databases

주요한 인사이트 채굴이 목표!

정의 및 접근법



어떻게 하면 **유의미한 정보**를
도출할 수 있을까?

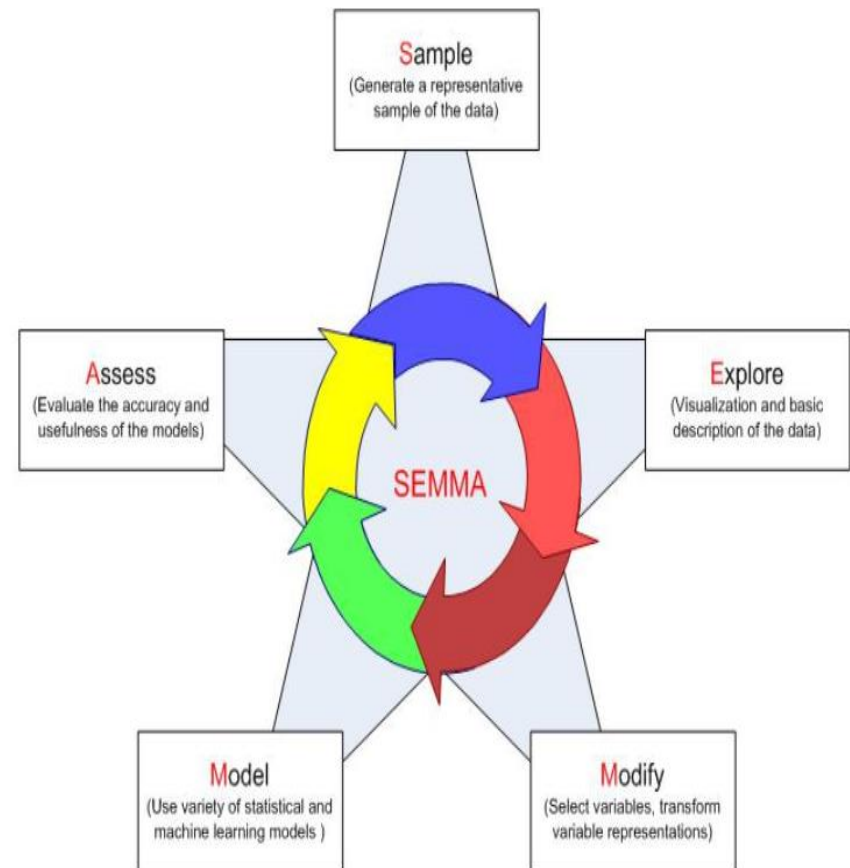
“**‘인사이트’라 불릴 만한 정보를 채굴하는 것이 목표!**”

프로세스

CRISP-DM

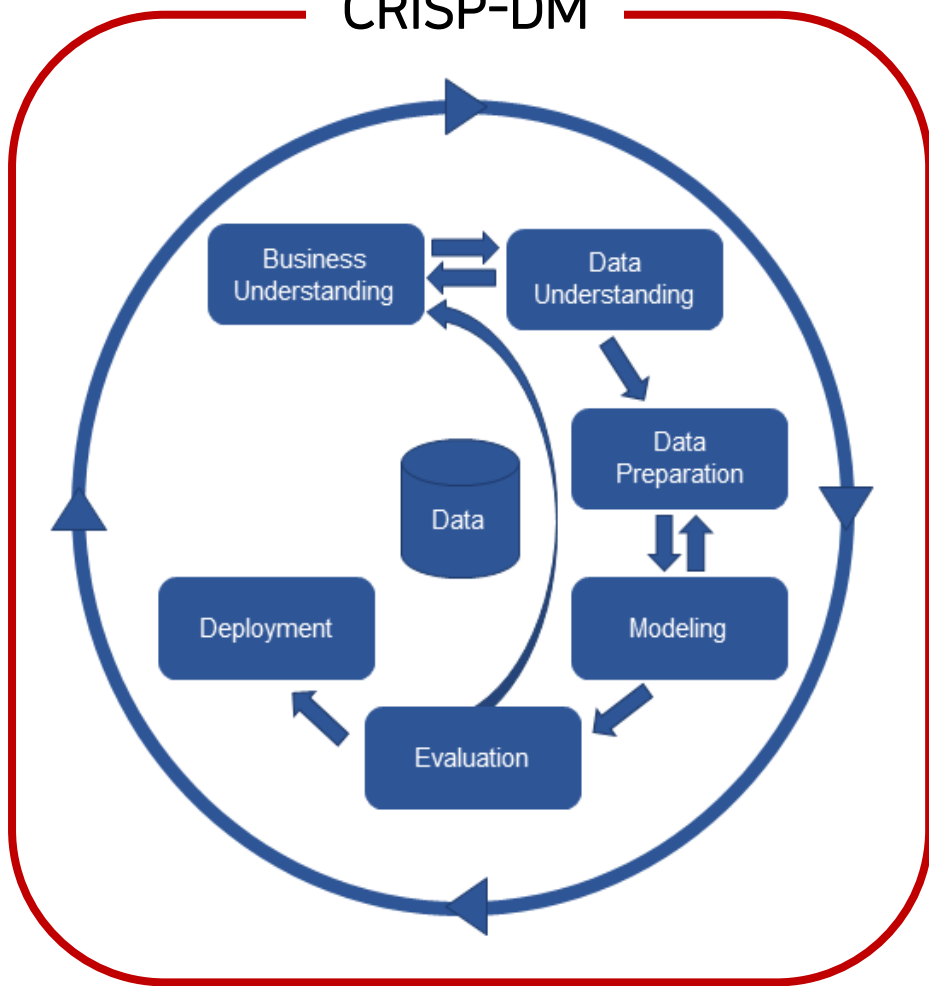


SEMMA

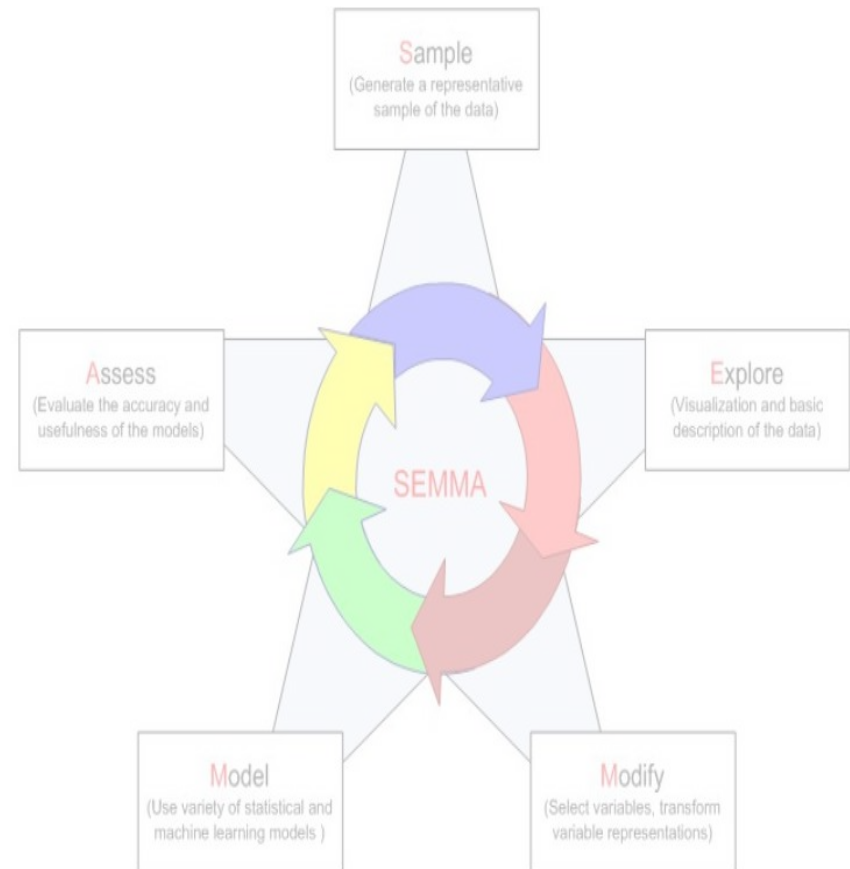


프로세스

CRISP-DM

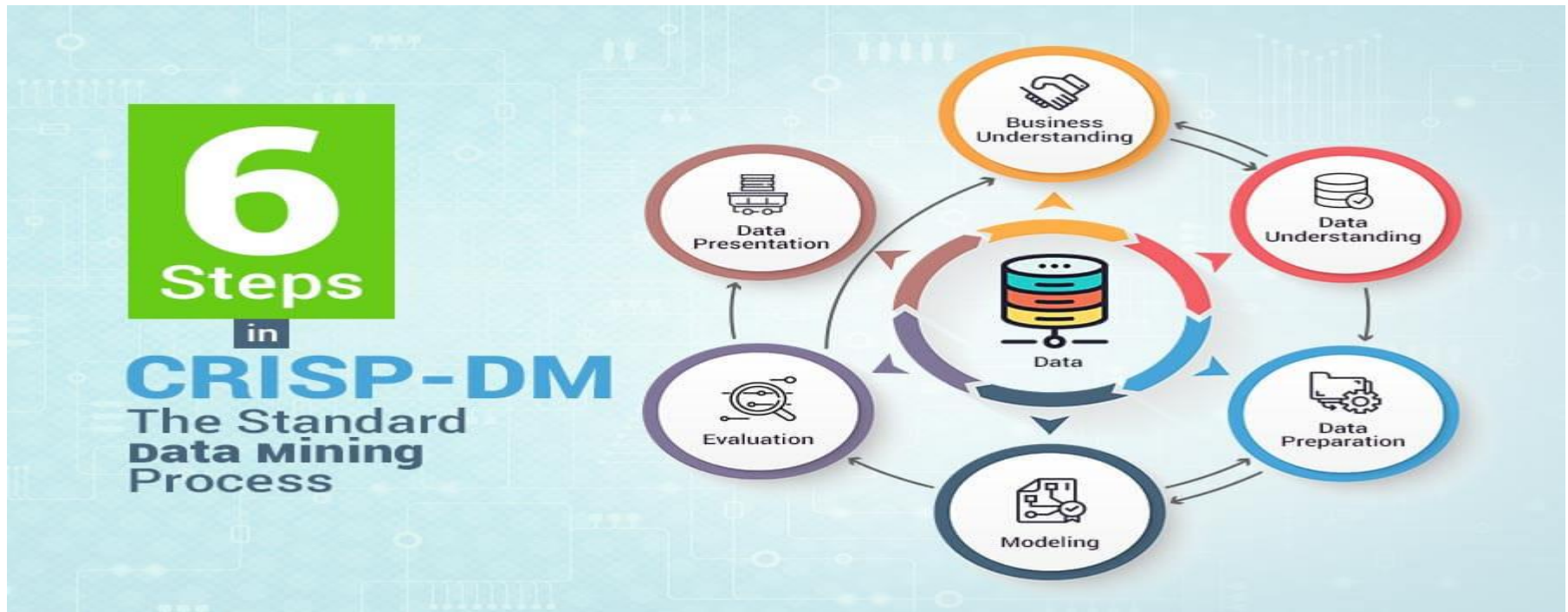


SEMMA



프로세스: CRISP-DM

“ **C**Ross **I**ndustry **S**tandard **P**rocess for
Data **M**ining Framework ”



프로세스: CRISP-DM



프로세스: CRISP-DM

1.

비즈니스 문제 이해

- 비즈니스 상황의 배경지식 쌓기
- 데이터 마이닝 과정의 성공 여부 기준 세우기

2.

데이터 이해
(EDA)

- 시각화를 통해 데이터 직관적 이해 달성
- 변수의미, 변수 간의 관계 파악
- 이상치, 결측치 유무 파악

3.

데이터 준비

- 데이터 전처리 과정
- 모델의 성능 개선에 주요한 역할

프로세스: CRISP-DM

4.

데이터 분석과
모델링

- 머신러닝 / 딥러닝 기법 적용
- 추천, 예측, 해석 등

5.

분석 모델의
평가

- '모델링이 잘 되었는지' 평가
- 범주형 데이터 (misclassification rate)
- 연속형 데이터 (RMSE)

6.

분석 결과의
적용

- 실제 비즈니스 상황에 적용

2

What is Modeling?

정의 및 접근법

사람이 하기 어려운 작업을 기계가 대신 수행할 수 있도록
기계를 학습시키는 일련의 작업



정의 및 접근법



수동적 정보발굴의 어려움 발생

정의 및 접근법

'컴퓨터야, 나 너 믿어도 되지?'

수동적 정보발굴의 어려움 발생



컴퓨터의 '자동화 기술'

모델링을 통한 데이터의 경향성/패턴
자동 파악

정의 및 접근법

<기계학습>

컴퓨터가 알아서 규칙을 발견하고 데이터의 경향성을 설명해주어야 함

기계학습 문제 3요소

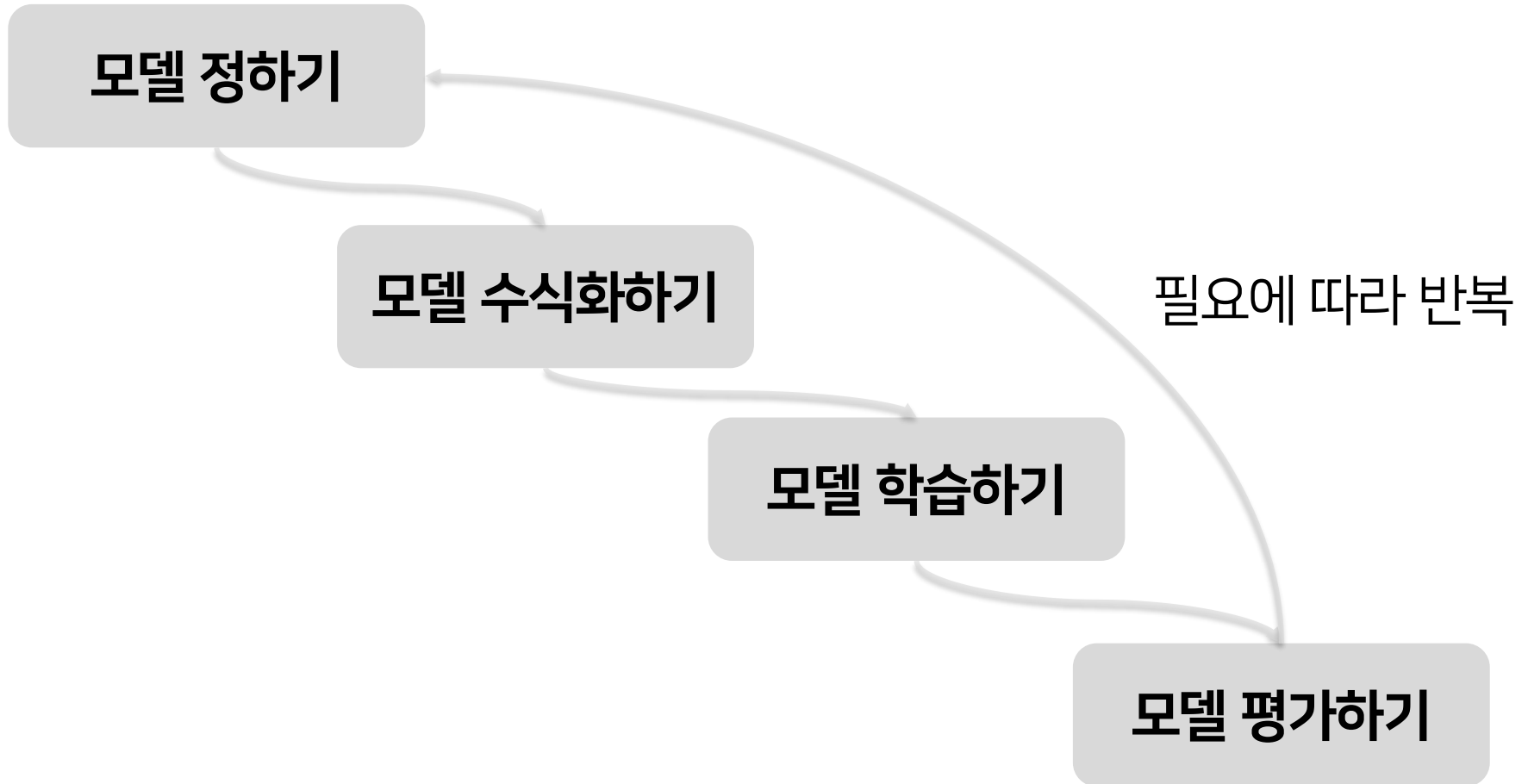
경험 (E, Experience) -> 데이터

학습 (T, Train / learn) -> 모델 적합

성능 평가 (P, Performance measure)

성능 개선의
여지가 있나?

모델링의 과정



학습 목적에 따른 분류

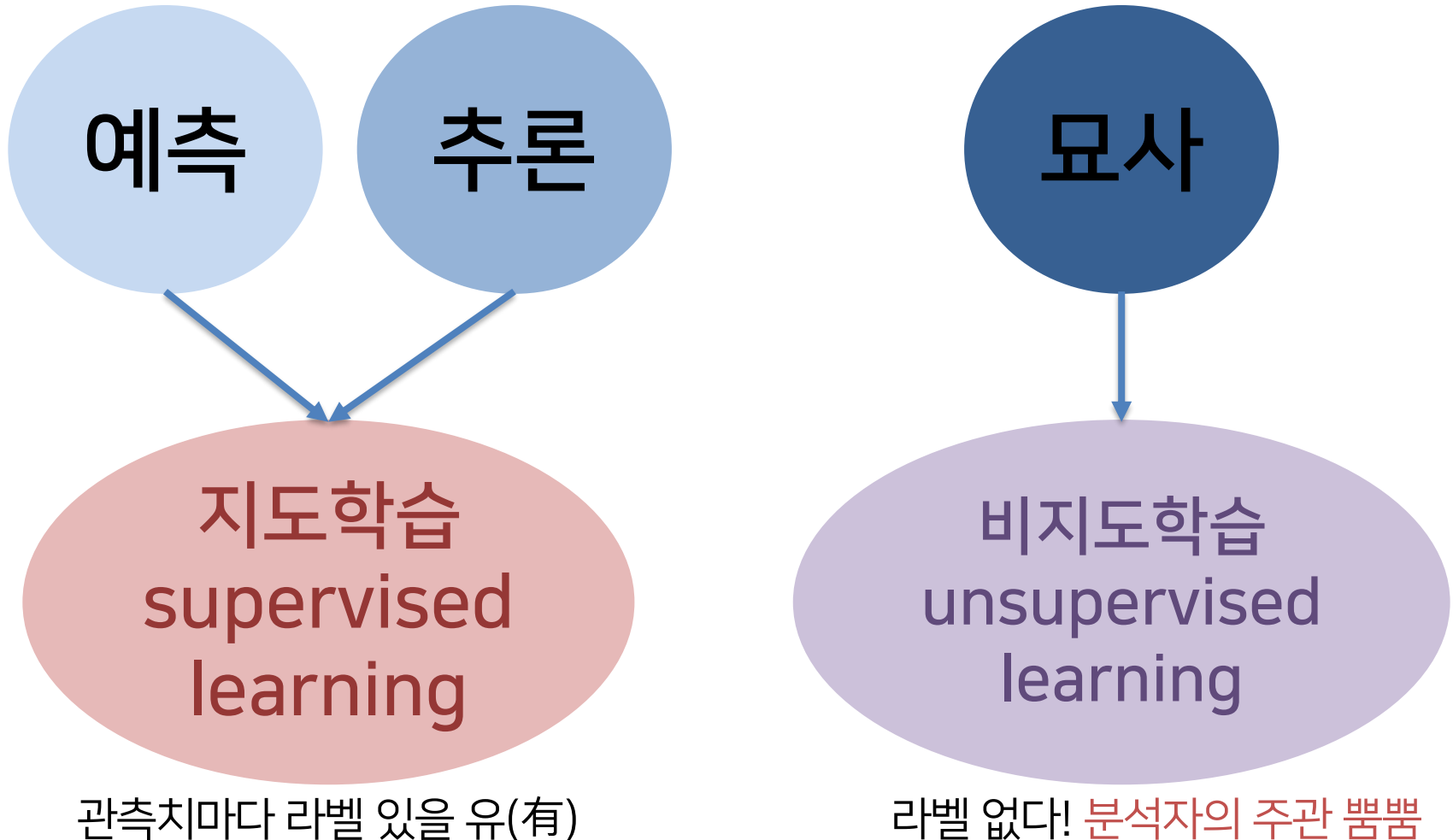


예측

추론

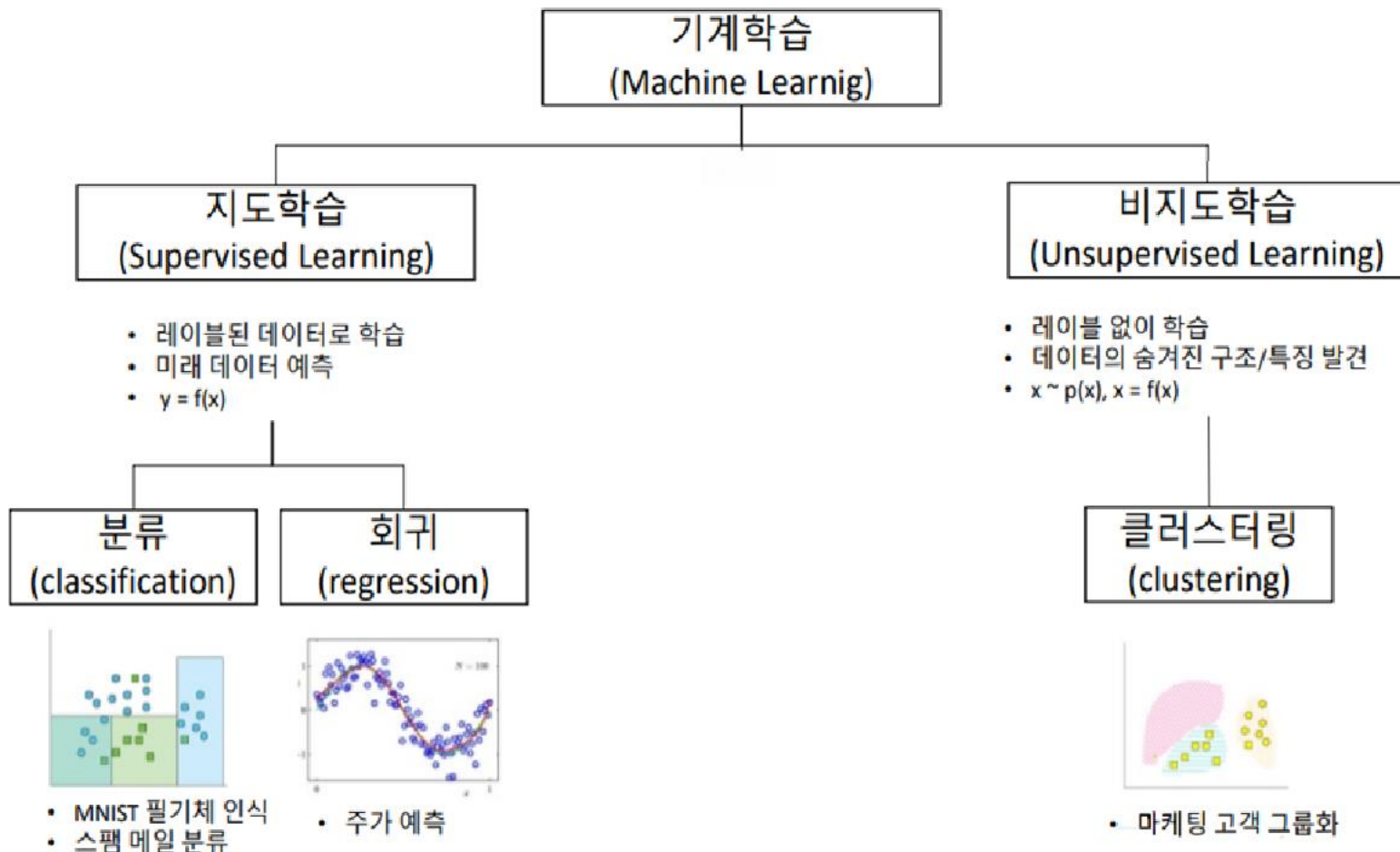
묘사

| 학습 목적에 따른 분류



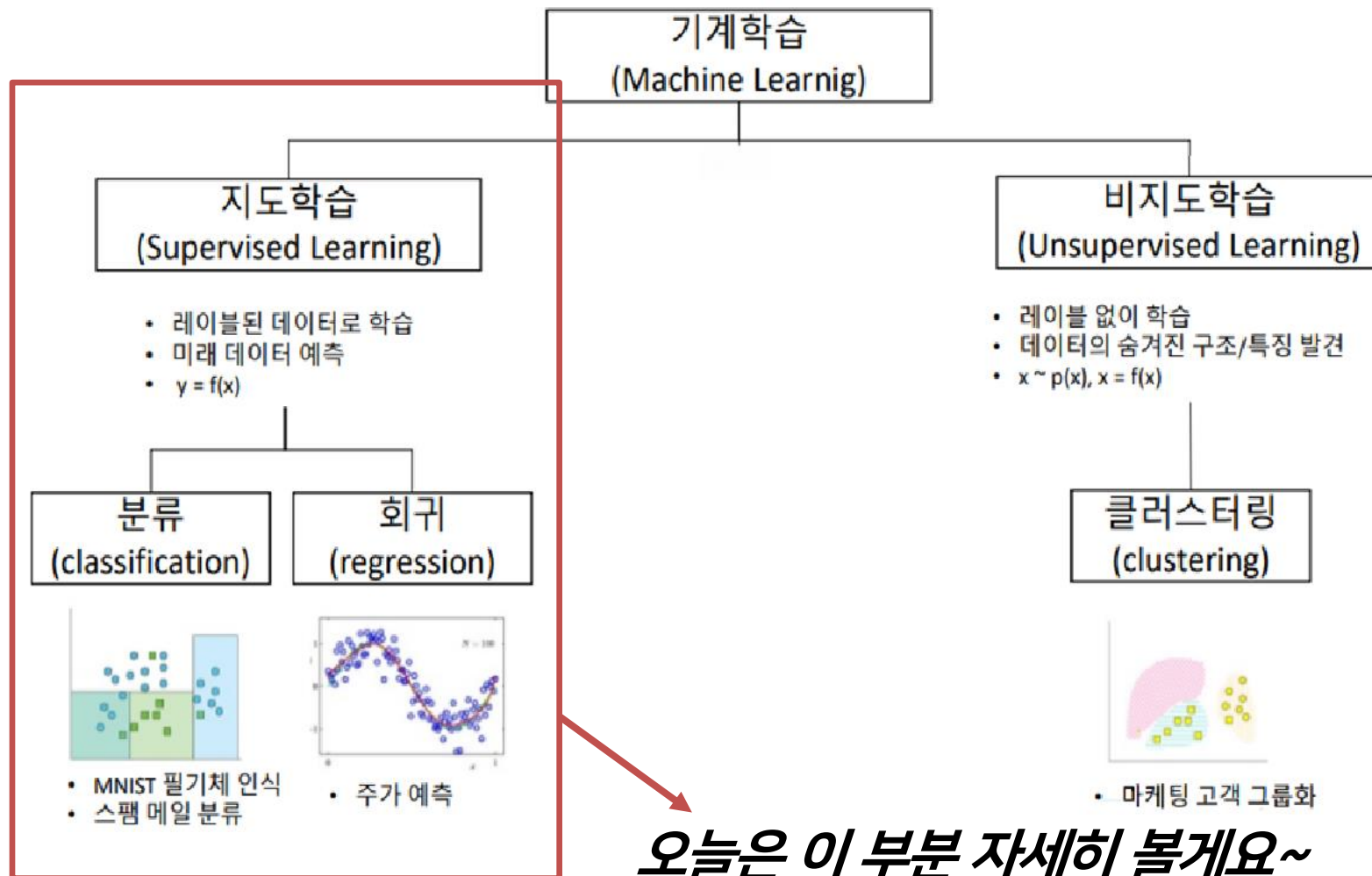
학습 목적에 따른 분류

큰 그림 그려볼까...?



학습 목적에 따른 분류

큰 그림 그려볼까...?



지도학습 (supervised learning)

무엇을 바탕으로
무엇을 예측/추론 하는가?



지도학습 (supervised learning)

학습 데이터 정보 이용해서

Bedrooms	Sq. feet	Neighborhood	Sale price
3	2000	Normaltown	\$250,000
2	800	Hipsterton	\$300,000
2	850	Normaltown	\$150,000
1	550	Normaltown	\$78,000
4	2000	Skid Row	\$150,000

[training data]

테스트 데이터에서
구하고자 하는 값 유추!

Bedrooms	Sq. feet	Neighborhood	Sale price
3	2000	Hipsterton	???

[test data]

지도학습 (supervised learning)



Bedrooms	Sq. feet	Neighborhood	Sale price
3	2000	Normaltown	\$250,000
2	800	Hipsterton	\$300,000
2	850	Normaltown	\$150,000
1	550	Normaltown	\$78,000
4	2000	Skid Row	\$150,000

[training data]

독립변수
Feature

특정 주택 내 방의 개수,
면적, 주택이 속한 동네

Bedrooms	Sq. feet	Neighborhood	Sale price
3	2000	Hipsterton	???

[test data]

종속변수
Target

주택 가격

지도학습 (supervised learning)

학습 데이터 정보 이용해서

Bedrooms	Sq. feet	Neighborhood	Sale price
3	2000	Normaltown	\$250,000
2	800	Hipsterton	\$300,000
2	850	Normaltown	\$150,000
1	550	Normaltown	\$78,000
4	2000	Skid Row	\$150,000

[training data]

종속변수

독립변수

테스트 데이터에서
구하고자 하는 값 유추!

Bedrooms	Sq. feet	Neighborhood	Sale price
3	2000	Hipsterton	???

[test data]

종속변수

지도학습 (supervised learning)

함수식 세우듯이!

$Y=f(X)$ 에 대하여 입력변수(X)와 출력변수(Y)의 **관계** 모델링

- **회귀(regression)** : 입력변수 X 에 대해서 연속형 출력변수 Y 를 예측
- **분류(classification)** : 입력변수 X 에 대해서 이산형 출력변수 Y 를 예측
어느 카테고리에 속하는지 (class)

지도학습 (supervised learning)

행렬로써 표현하자면~

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

또는 설명변수! ~ 각 관측치의 경향성에 대해 '설명'하니까!

↑
독립변수 p개와 관측치 n개로 구성된 input data

지도학습 (supervised learning)

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

$$Y = f(X) + \epsilon$$

실제 수학식, 그러니까 모델의 형태 알 수 없어!

지도학습 (supervised learning)

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

$$Y = f(X) + \epsilon$$

실제 수학적식, 그러니까 모델의 형태 알 수 없어!

➡ **“추정” 하자!**
우리의 모델이 실제와 최대한 유사하도록.

편향-분산 트레이드 오프(Bias-Variance Tradeoff)

구한 예측치가 실제값과의 차이가 작을수록 좋은 모델!

* 회귀분석의 관점) **MSE(Mean Squared Error)**를 최소화

MSE

$$\begin{aligned}
 E[(y - \hat{f})^2] &= E[y^2 + \hat{f}^2 - 2y\hat{f}] \\
 &= E[y^2] + E[\hat{f}^2] - E[2y\hat{f}] \\
 &= \text{Var}[y] + E[y]^2 + \text{Var}[\hat{f}] + E[\hat{f}]^2 - 2fE[\hat{f}] \\
 &= \text{Var}[y] + \text{Var}[\hat{f}] + (f - E[\hat{f}])^2 \\
 &= \sigma^2 + \text{Var}[\hat{f}] + \text{Bias}[\hat{f}]^2
 \end{aligned}$$

Irreducible
Error

Reducible
Error

편향-분산 트레이드 오프(Bias-Variance Tradeoff)

MSE

=

Irreducible
Error

+

Reducible
Error



표본 추출로 인해 발생하는 Irreducible error (δ^2)는 차치하고,
Reducible error ($\text{Bias} + \text{Var}(\hat{f})$)을 최소화 하는 모델 설계가 목표!

편향-분산 트레이드 오프(Bias-Variance Tradeoff)

MSE

=

Irreducible
Error

+

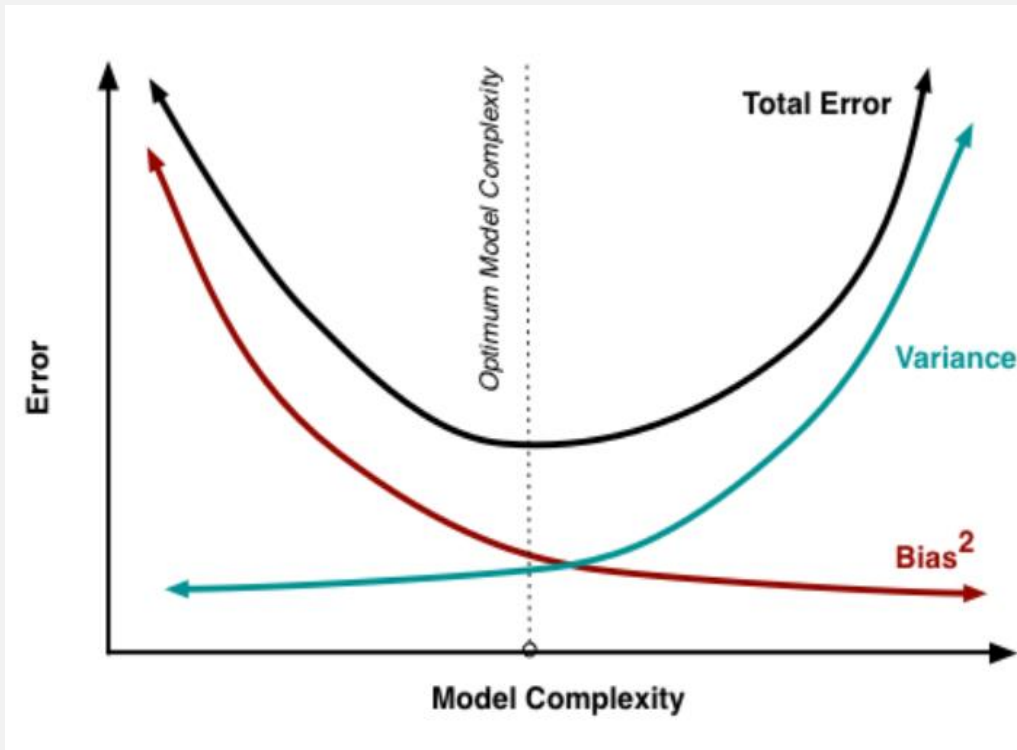
Reducible
Error



표본 추출로 인해 발생하는 Irreducible error (δ^2)는 차치하고,
Reducible error ($\text{Bias} + \text{Var}(\hat{f})$)을 최소화 하는 모델 설계가 목표!

- * **Bias**는 편향, 즉 추정된 f 와 \hat{f} 의 차이!
- * **$\text{Var}(\hat{f})$** 은 모델의 분산으로, 매번 다른 표본 추출마다 얼마나 다양한 형태를 나타내는 정도!

편향-분산 트레이드 오프(Bias-Variance Tradeoff)



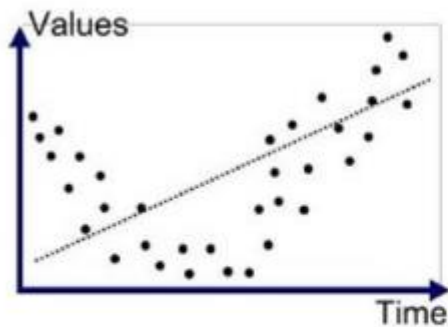
하지만,

Bias와 Variance를
동시에 원하는 수준으로
줄이기 어려움.

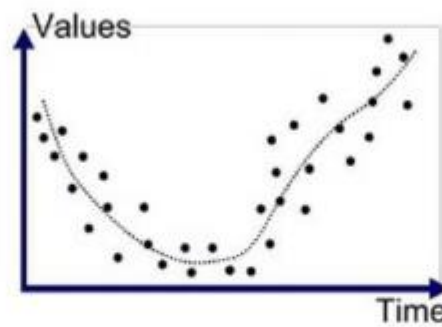
편향-분산 트레이드 오프(Bias-Variance Tradeoff)



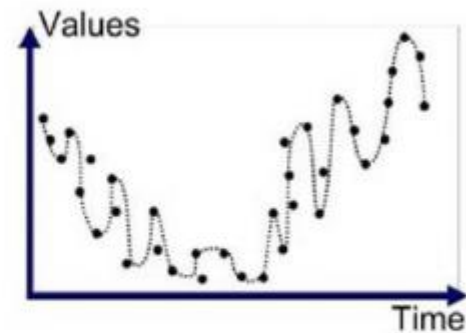
모델의 복잡도 (model complexity)



Underfitted

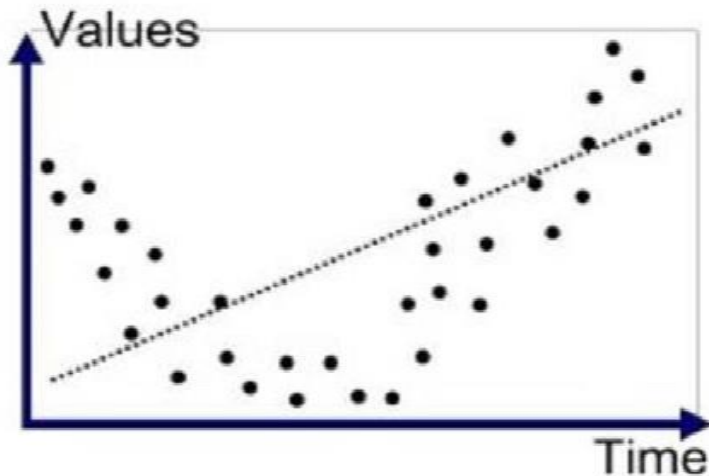


Good Fit/Robust



Overfitted

모델의 복잡도 (model complexity)

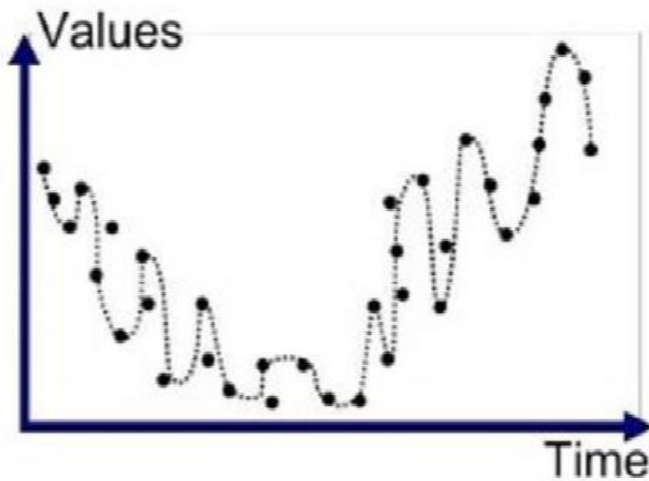


Underfitted

High bias, low variance
이해와 해석 Good!

높은 편향
정확한 예측은 Bad!

모델의 복잡도 (model complexity)



Low bias, High variance

주어진 데이터에서는 정확한 예측
Good!

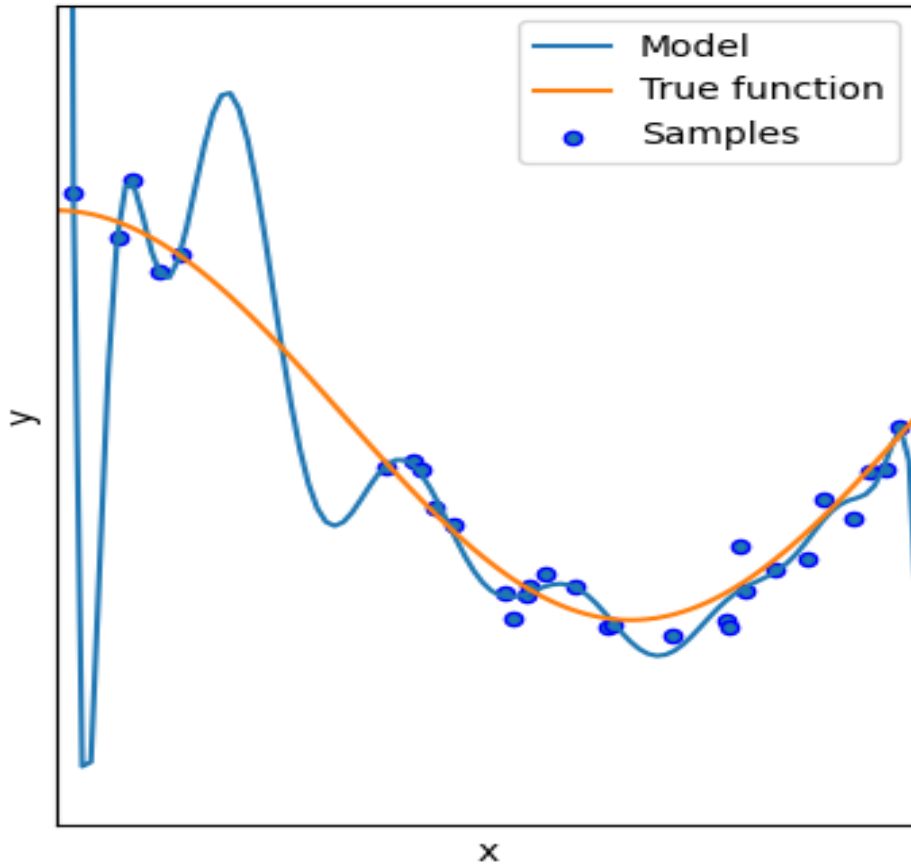
높은 분산

일반화된 해석력은 Bad!

3

How to avoid Overfitting

Overfitting [과적합 문제]



**복잡한 모델일수록,
즉 모델의 파라미터가
 많아질수록**



**“과적합 발생
위험 증가!”**

관측치 지점 하나하나 다 연결한 모습 보이냐?

Overfitting [과적합 문제]

과적합 발생 위험 증가!

우리에게 주어진 데이터에 대해서만
완벽히 설명하는 모델 설계

↓ 새로운 데이터가 들어온다면?

새로운 데이터에 대한 설명력 확보하지 못함!

모델의 '재사용성'은 이렇게 날라가고...!

Overfitting [과적합 문제]

따라서,

과적합을 방지하기 위해
우리가 설계한 **모델의 성능**을 평가할 때

조금 더 객관적일 필요가 있다!

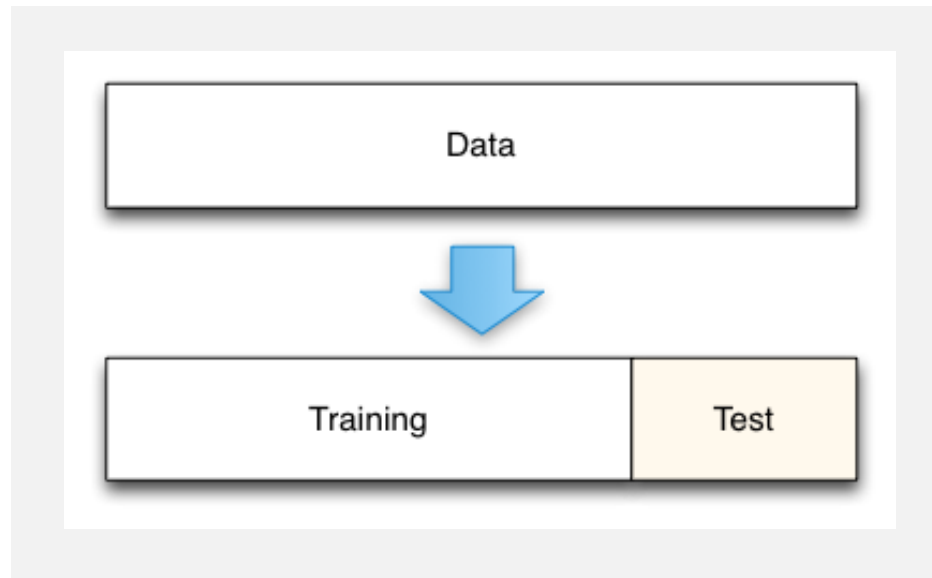
*우리에겐 주어진 데이터에 대해서는
완벽히 설명하는 모델 설계*

✓ *새로운 데이터가 들어온다면?*
새로운 데이터 (**검증 데이터**)에 대해서

새로운 데이터에 대한 모델의 성능을 평가할 수 없음!
어떻게 반응할지 궁금하다!

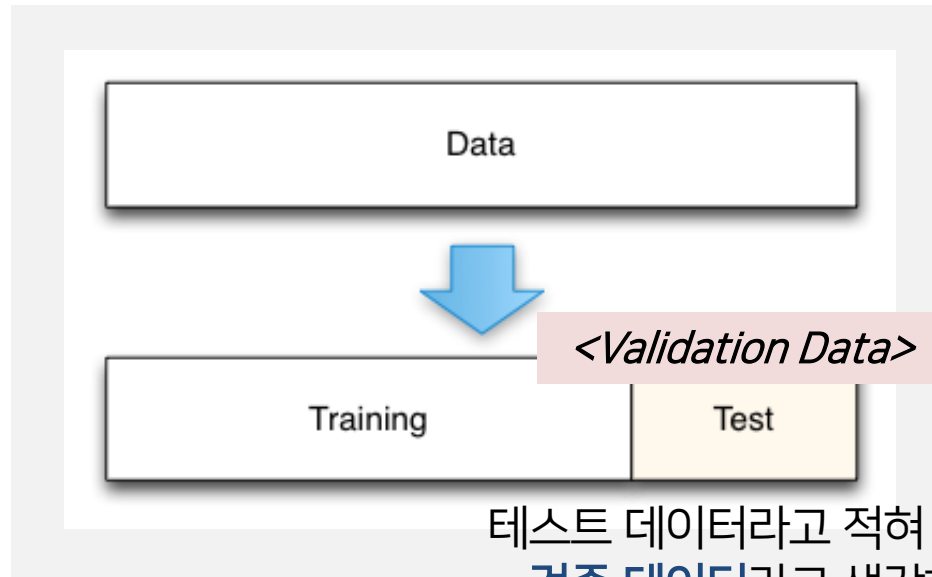
모델의 '재사용성'은 이렇게 날라가고...!

Train-test split [검증데이터 분할]



실제 데이터와 모델을 통해 예측한 데이터의 차이,
즉 **error**를 줄이는 것이 우리의 목표!

Train-test split [검증데이터 분할]



테스트 데이터라고 적혀 있지만
검증 데이터라고 생각하자!

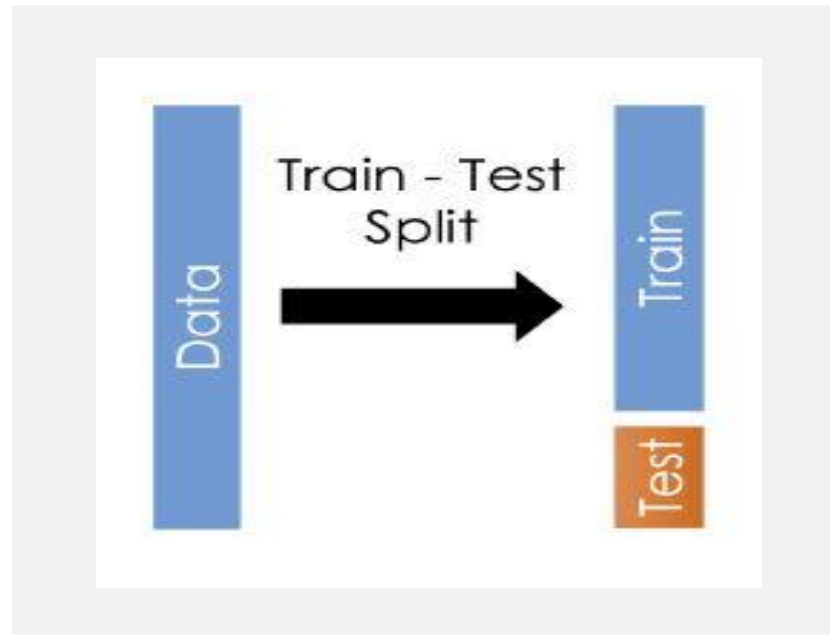
“

학습 데이터의 일부를

”

검증 데이터로 삼아 모델의 성능을 평가해보자!

Train-test split [검증데이터 분할]

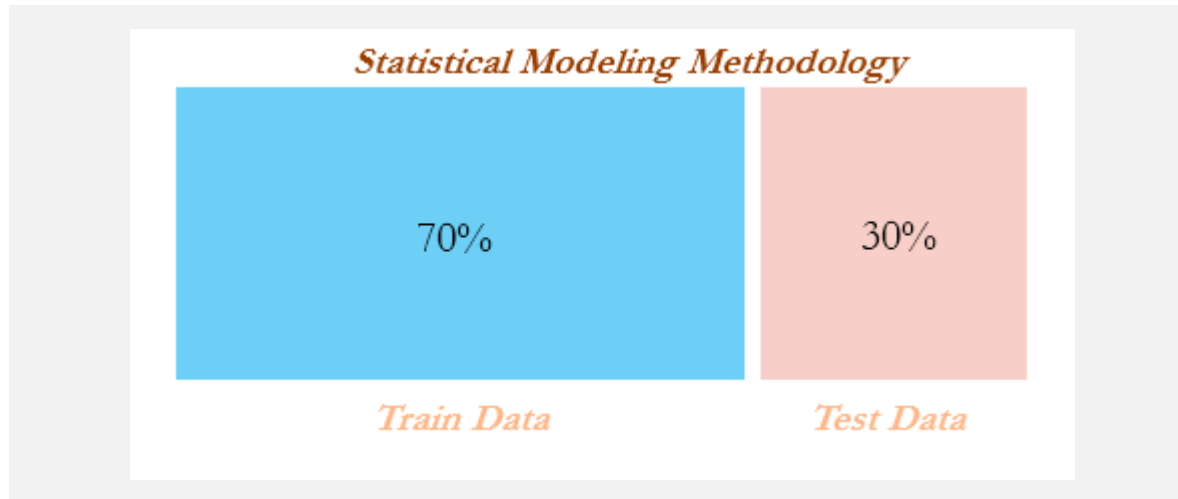


검증 데이터 분할 시,

“What is the most proper ratio?”

Train-test split [검증데이터 분할]

“What is the most proper ratio?”



학습 데이터 : 검증 데이터를
7:3 혹은 **8:2** 비율로 두면
오버피팅 상황과 언더피팅 상황을 적절히 피해!

Train-test split [검증데이터 분할]

근데 이렇게 **단일한** 검증 데이터셋만을 구성한다면
좋은 성능을 기대하기란 어렵다!

Statistical Modeling Methodology

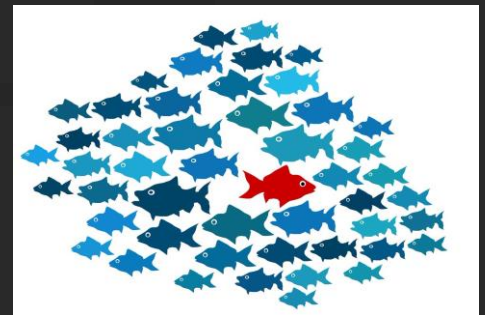
→ **검증 데이터를 여러 개 만들어내**
이들 데이터들을 종합적으로 평가해내는
방법을 활용하자.

Train Data

Test Data

학습 데이터 : 검증 데이터를
7:3 혹은 8:2 비율로 두면

오버피팅 상황과 언더피팅 상황을 적절히 피해
(빨강 물고기만으로 검증 데이터가 구성될 수도)



K-fold CV [K-fold 교차검증]

주의

우리나라에서 만들어서 앞에 K 붙은 거 아님

Estimation 1	Test	Train	Train	Train	Train
Estimation 2	Train	Test	Train	Train	Train
Estimation 3	Train	Train	Test	Train	Train
Estimation 4	Train	Train	Train	Test	Train
Estimation 5	Train	Train	Train	Train	Test

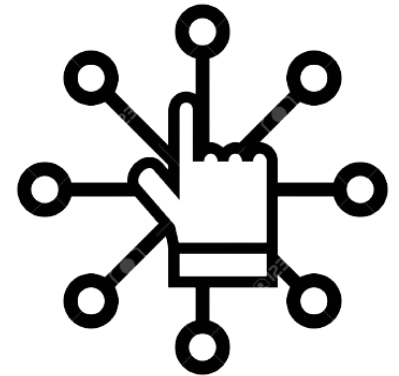
전체 데이터를 **k개의 그룹(fold)으로** 나눈 후
한 개의 데이터셋을 검증 데이터셋으로,
나머지 k-1개의 데이터셋을 학습 데이터셋으로 사용

K-fold CV [K-fold 교차검증]

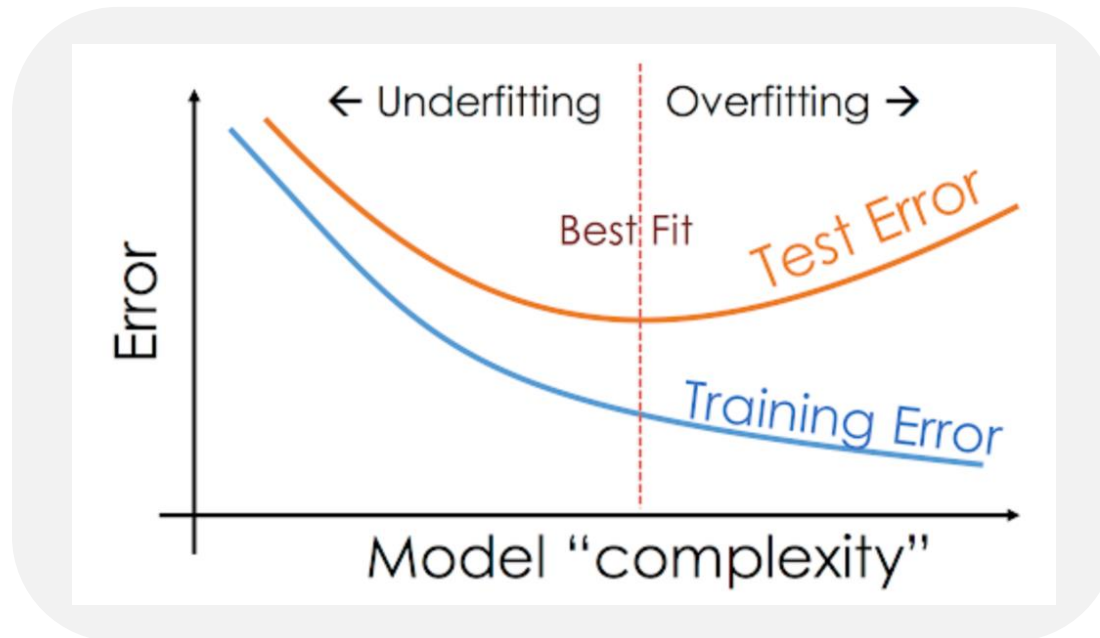
k 번 반복되는 과정에서

모델링에 활용되는 변수 선택,

좋은 성능을 내는 hyperparameter 탐색

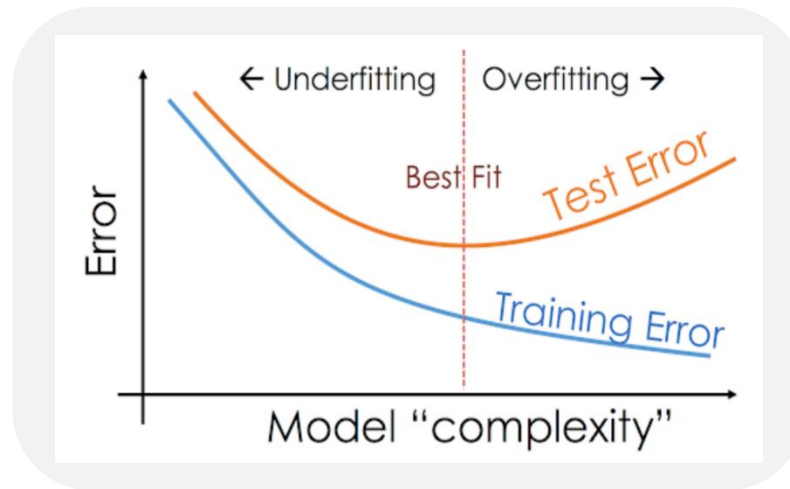


차원의 저주 [Curse of Dimensionality]



*모델의 복잡도가 높아짐에 따라
검증 데이터 error는 다시 증가하는 모습을 보이는데...*

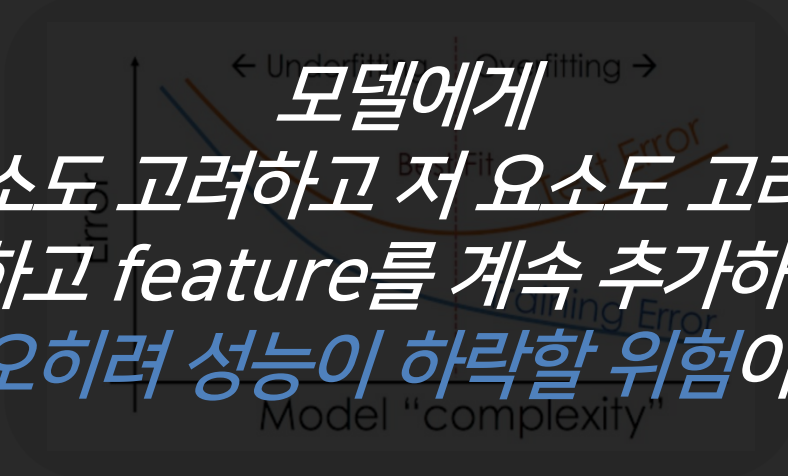
차원의 저주 [Curse of Dimensionality]



“변수(feature)가 너무 많아!”

모델의 복잡도가 높아짐에 따라
검증 데이터 error는 다시 증가하는 모습을 보이는데...

차원의 저주 [Curse of Dimensionality]

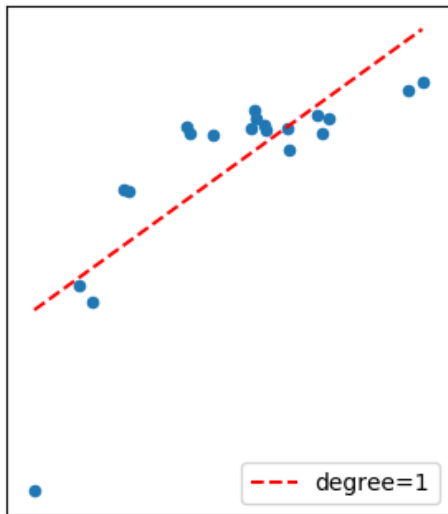


모델에게
'이 요소도 고려하고 저 요소도 고려하렴~'
하고 *feature*를 계속 추가하면
오히려 성능이 하락할 위험이!!

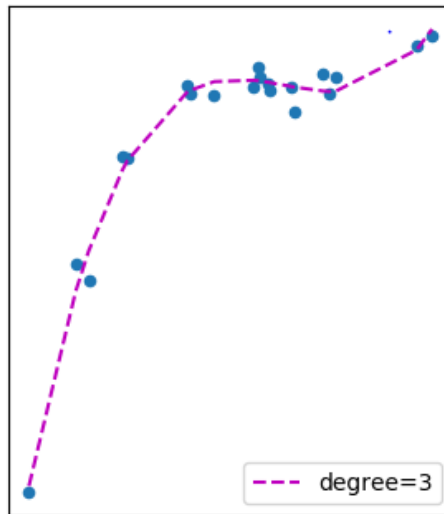
→ **오버피팅 문제 야기**
변수 (*feature*)가 너무 많
아!"

모델의 복잡도가 높아짐에 따라
검증 데이터 *error*는 다시 증가하는 모습을 보이는데...

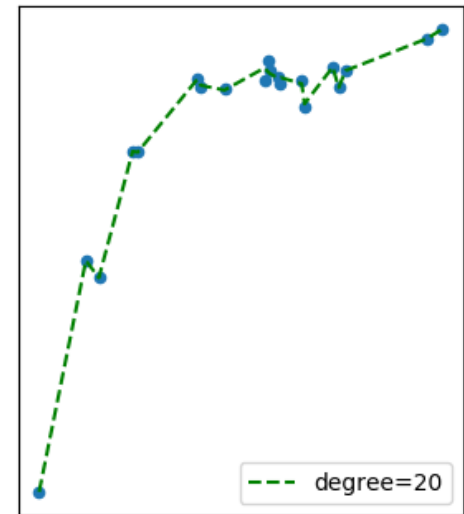
차원의 저주 [Curse of Dimensionality]



Underfit
High Bias
Low Variance



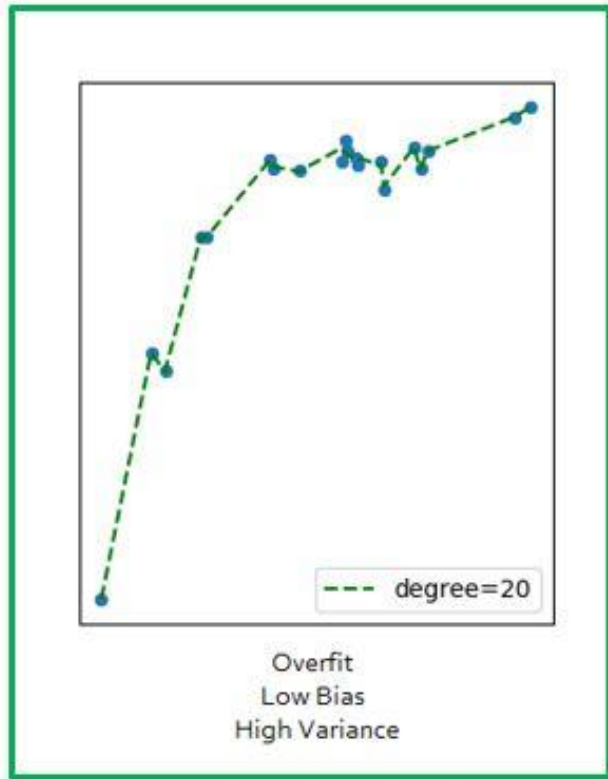
Correct Fit
Low Bias
Low Variance



Overfit
Low Bias
High Variance

이런거 어떻게 해석할건데..?

차원의 저주 [Curse of Dimensionality]



고차원 모델,
즉 **overfitting**할 위험이 높은 모델은

- 1) 관측치들의 경향성 파악
- 2) 1을 바탕으로 해석

하기 어렵다는 단점 보유

이런거 어떻게 해석할건데..?

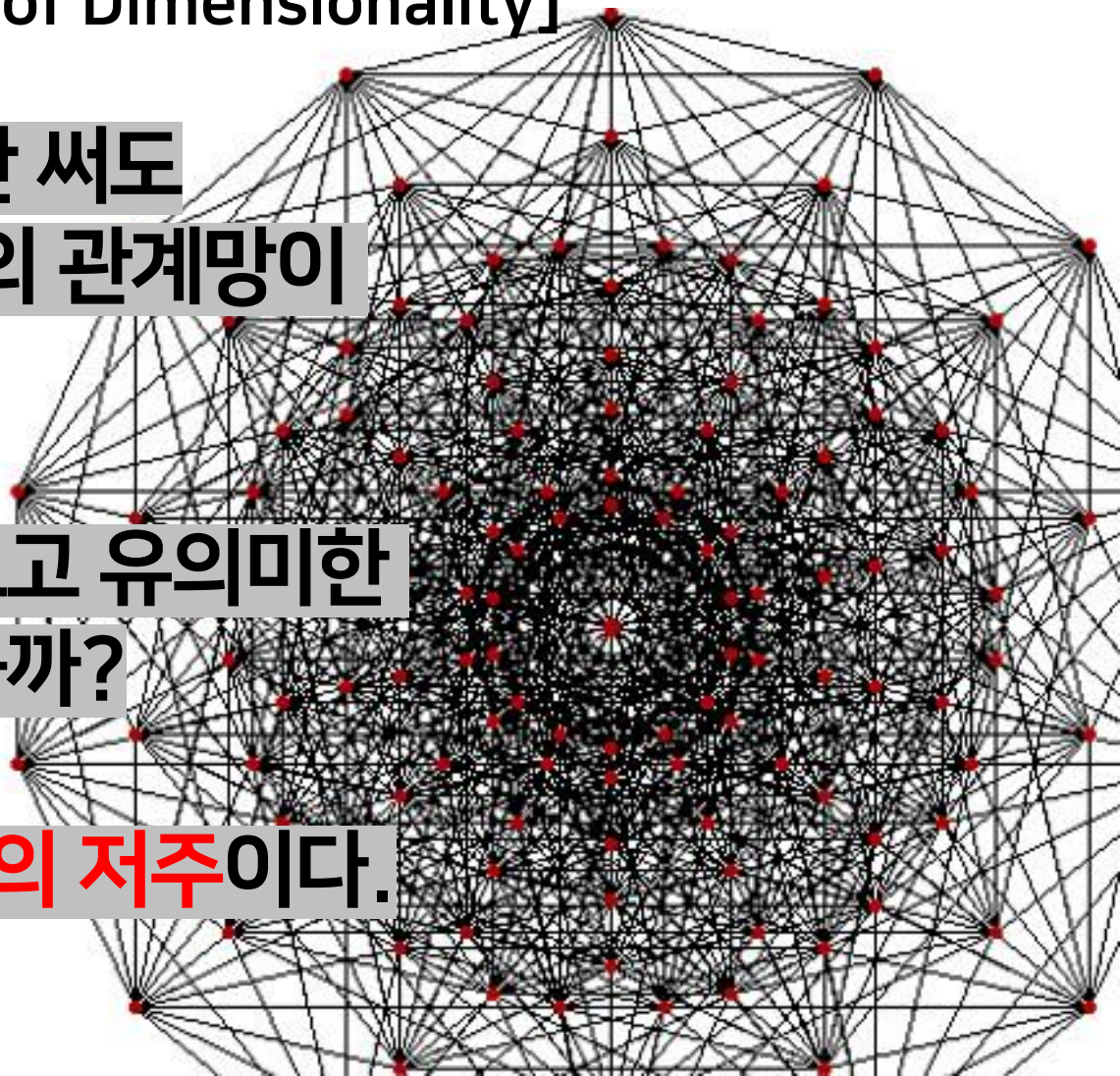
다음 장 살짝 공포 주의

차원의 저주 [Curse of Dimensionality]

독립변수를 8개만 써도
이와 같은 8차원의 관계망이
그려진다.

과연 이 관계를 보고 유의미한
해석을 할 수 있을까?

이것이 바로 차원의 저주이다.



자세한 내용은 회귀/선대/딥팀 교안 참고하시고~!

차원의 저주 [Curse of Dimensionality]

따라서, 너무 적지도 많지도 않은 **적절한 변수 개수를 설정**해야 하는데...

몇 가지 방법 소개해드립니다...!

1. Feature Selection

EX) Forward Selection,
Stepwise Selection

자세한 내용은 회귀/선대/딥팀 교안 참고하시고~!

차원의 저주 [Curse of Dimensionality]

따라서, 너무 적지도 많지도 않은 적절한 변수 개수를 설정해야 하는데...

몇 가지 방법 소개해드릴게요...

1. Feature Selection

EX) Forward Selection,
Stepwise Selection

2. Feature Extraction

EX) Principal Component
Analysis(PCA, 주성분 분석)

자세한 내용은 회귀/선대/딥팀 교안 참고하시고~!

차원의 저주 [Curse of Dimensionality]

따라서, 너무 적지도 많지도 않은 적절한 변수 개수를 설정해야 하는데...

몇 가지 방법 소개해드릴름...!

1. Feature Selection

EX) Forward Selection,
Stepwise Selection

2. Feature Extraction

EX) Principal Component
Analysis(PCA, 주성분 분석)

3. Early Stopping

이건 변수 선택의 문제라기보다
사용자 지정 accuracy 지점에 도달했을 때 모델의 학습 과정을 멈추는 것!



THANK YOU

