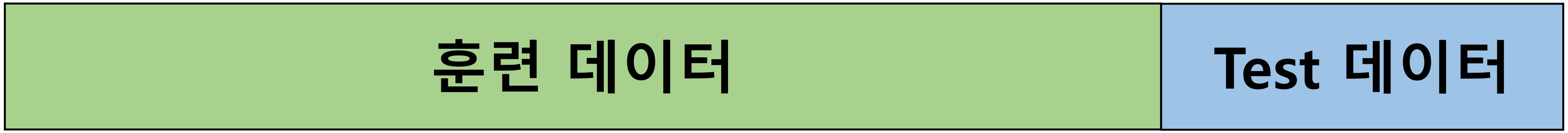
Resampling

Introduction to Statistical Learning

황성원

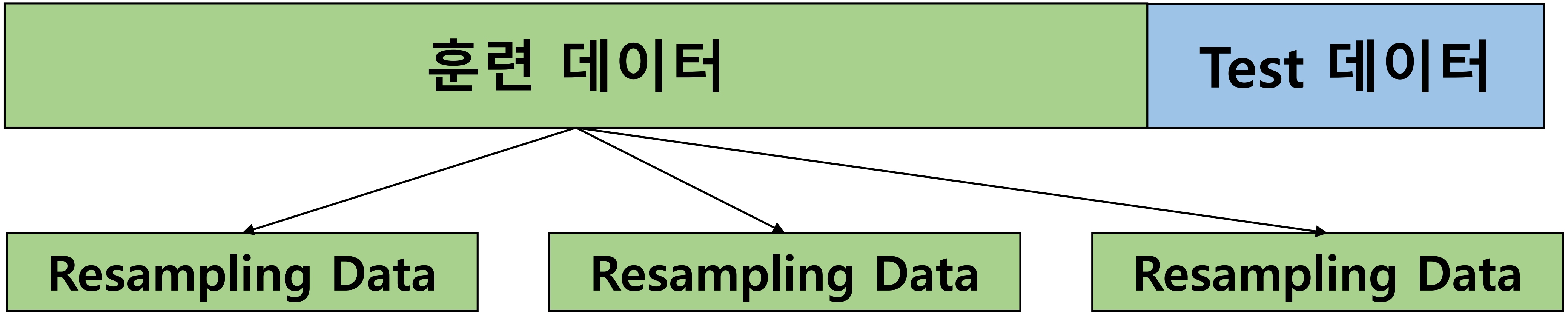
Resampling?

- 기존 방법 -



훈련데이터 → Fitting Model → Test (한 번의 Training 과정)

- Resampling 방법 -



훈련데이터 → Resampling → Fitting Model → Test (여러 번의 Training 과정)

SW

Resampling?

- Resampling 방법 -

훈련 데이터 | Test 데이터

Resampling Data | Resampling Data | Resampling Data

훈련데이터 → Resampling → Fitting Model → Test (**여러 번**의 Training 과정)

↓

계산 부담 증가

↓

컴퓨터 파워 증가로 해결!

Resampling?

- Resampling 방법 -

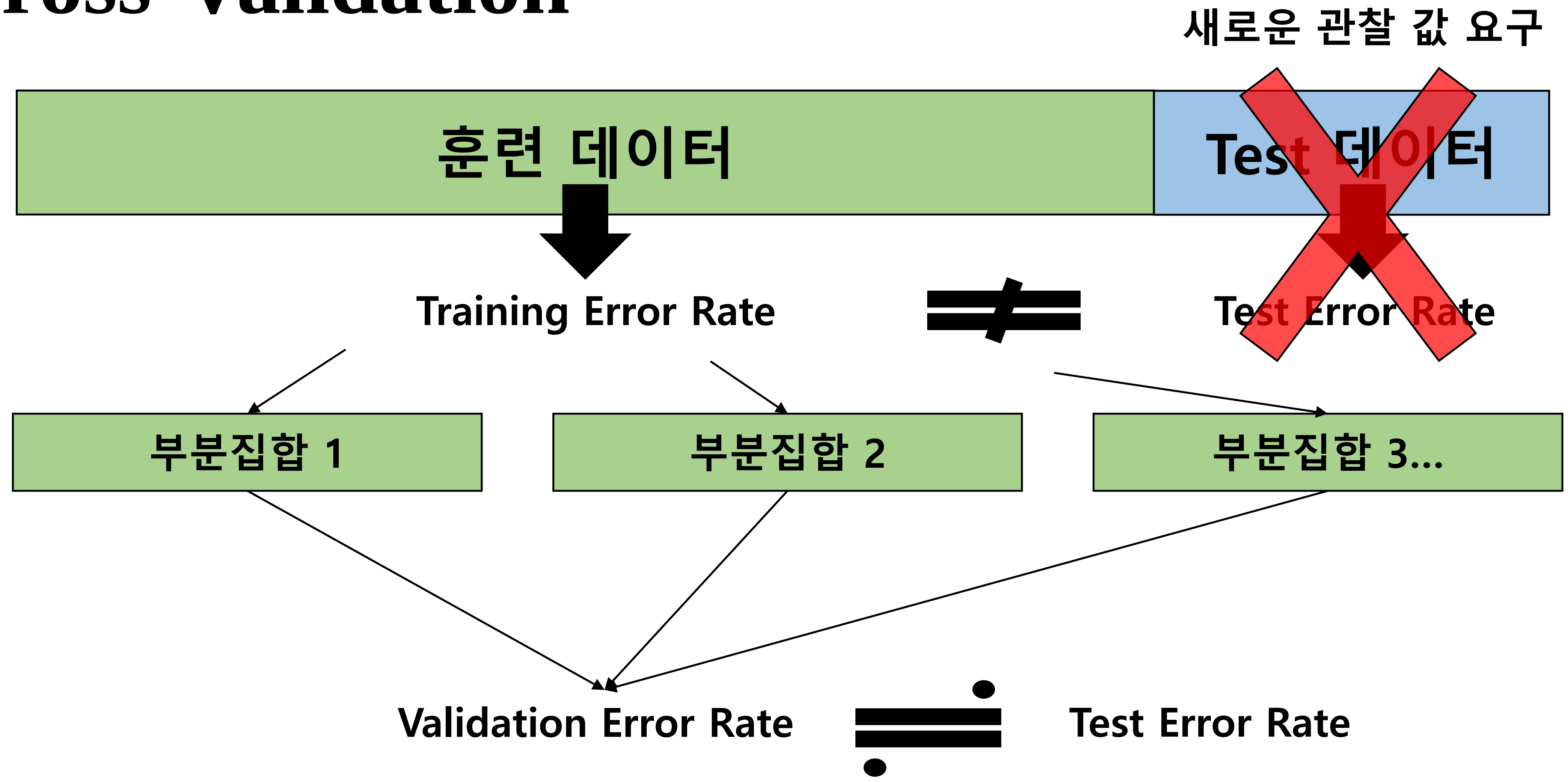
CV (Cross Validation)

Bootstrap

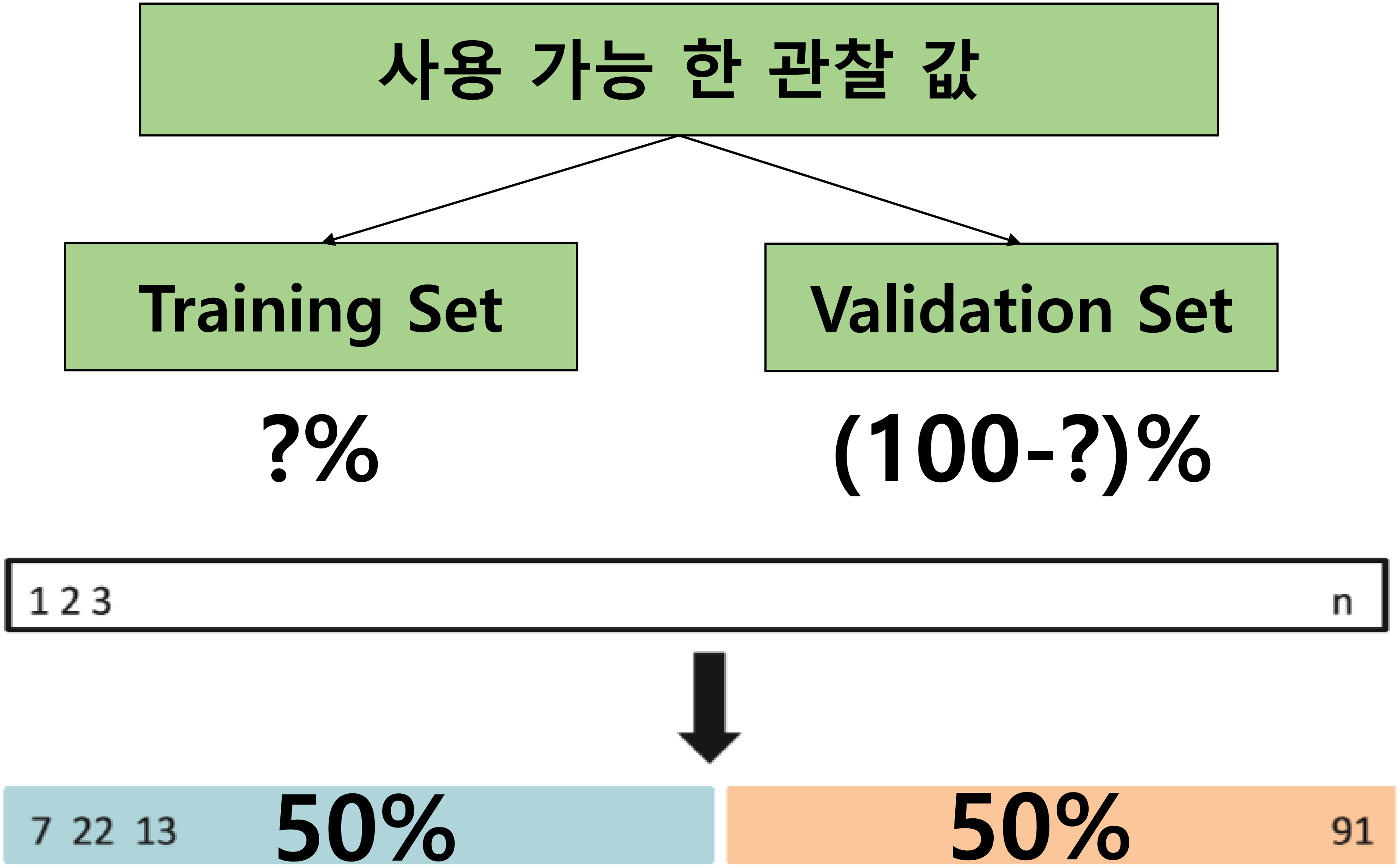
모델의 Performance를 평가 → Model Assessment

모델의 Flexibility를 선택 → Model Selection

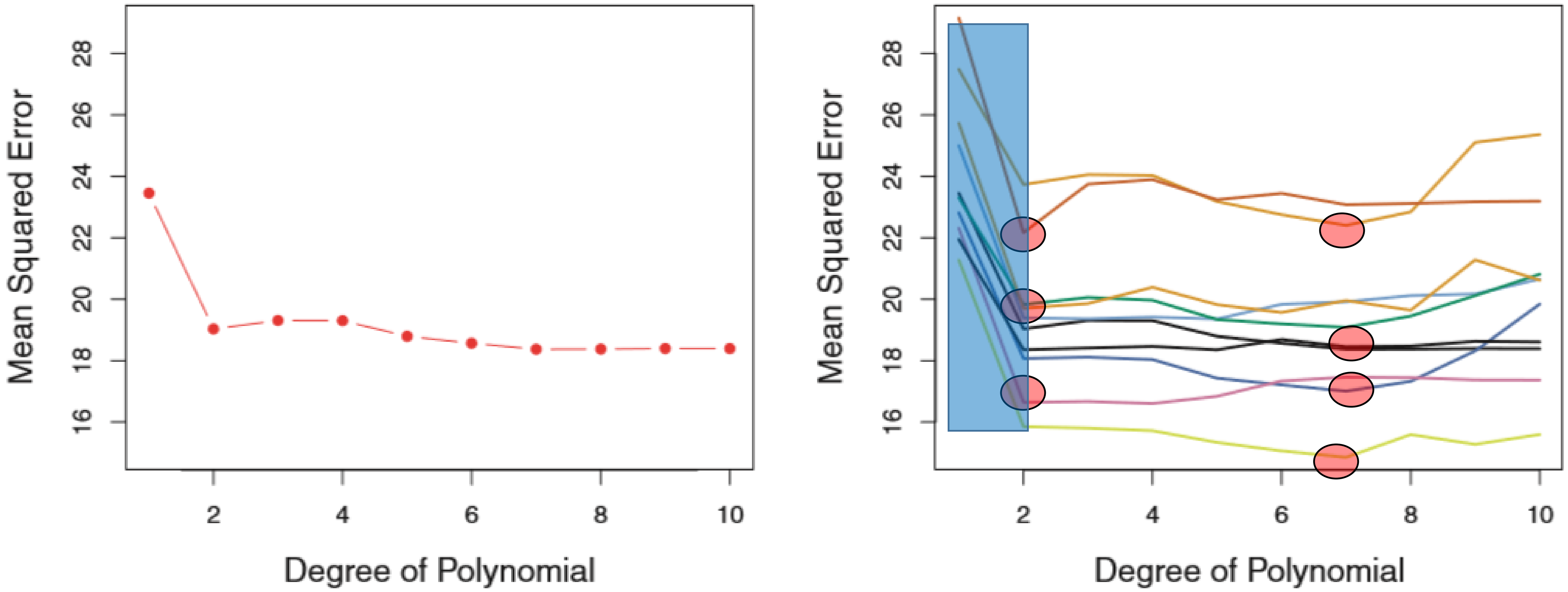
Cross-Validation



Validation Set Approach

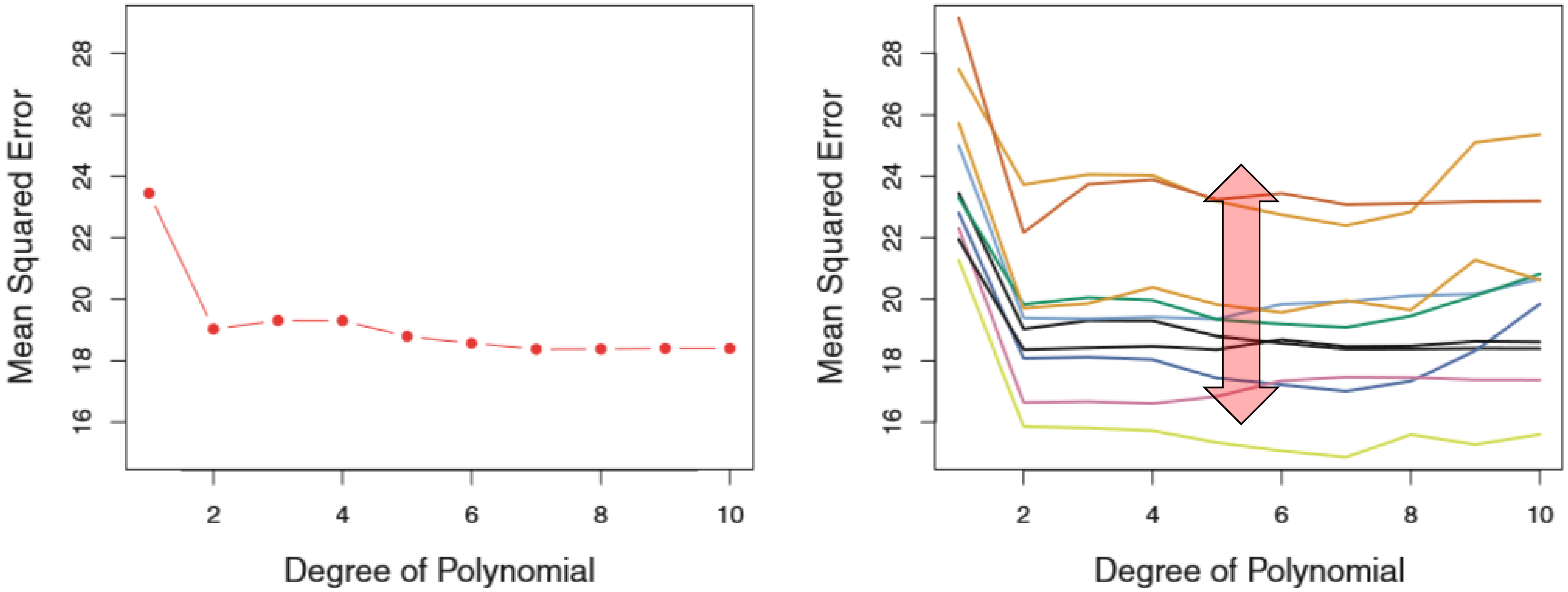


알 수 있는 것! (2 가지)



$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \epsilon$$

단점! (2 가지) – 경우에 따라 변화 심함



$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \epsilon$$

단점! (2 가지) – Training 사이즈 축소 = Performance 감소

사용 가능 한 관찰 값 = Training Set

VS.

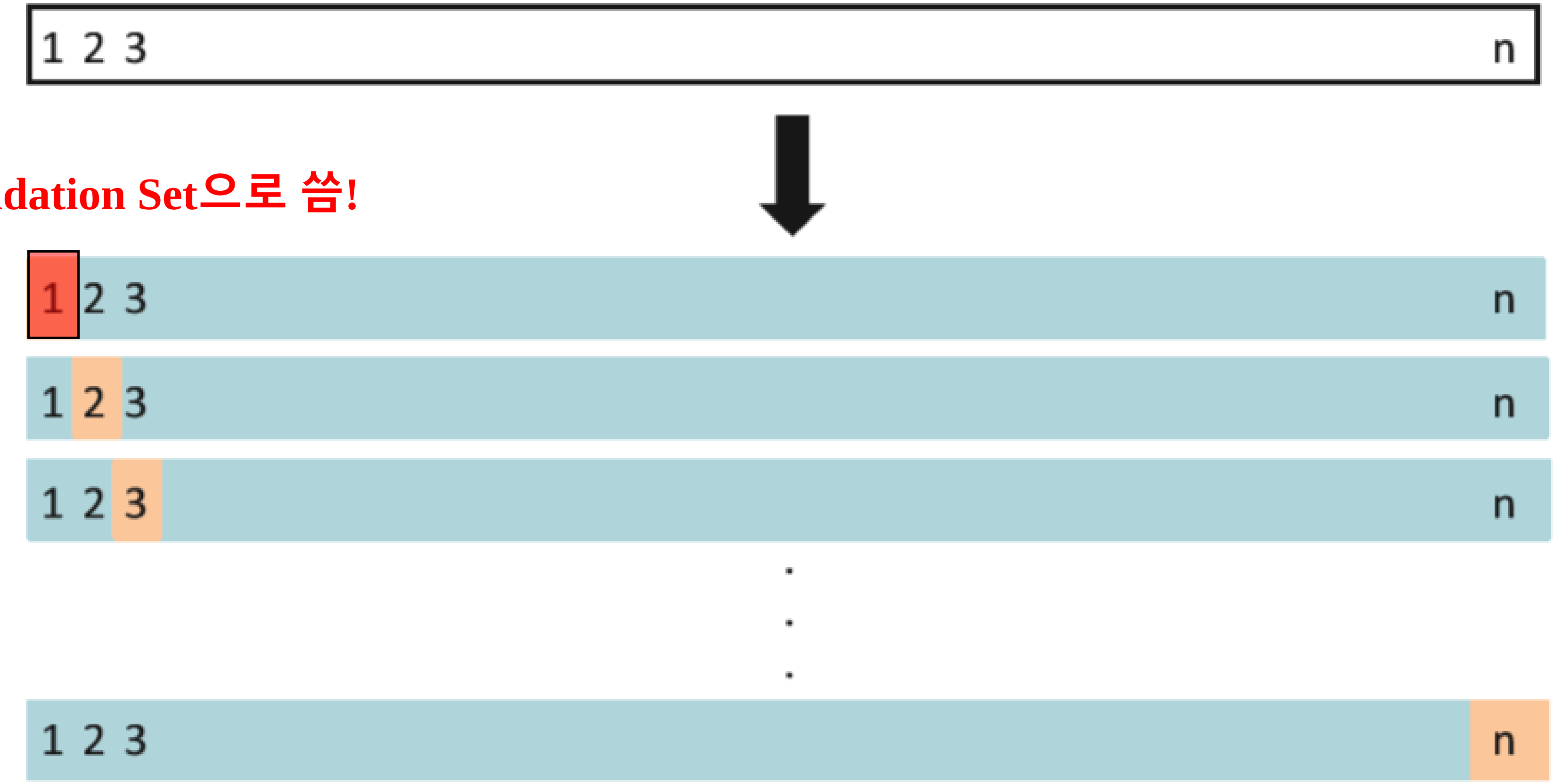
사용 가능 한 관찰 값

Training Set

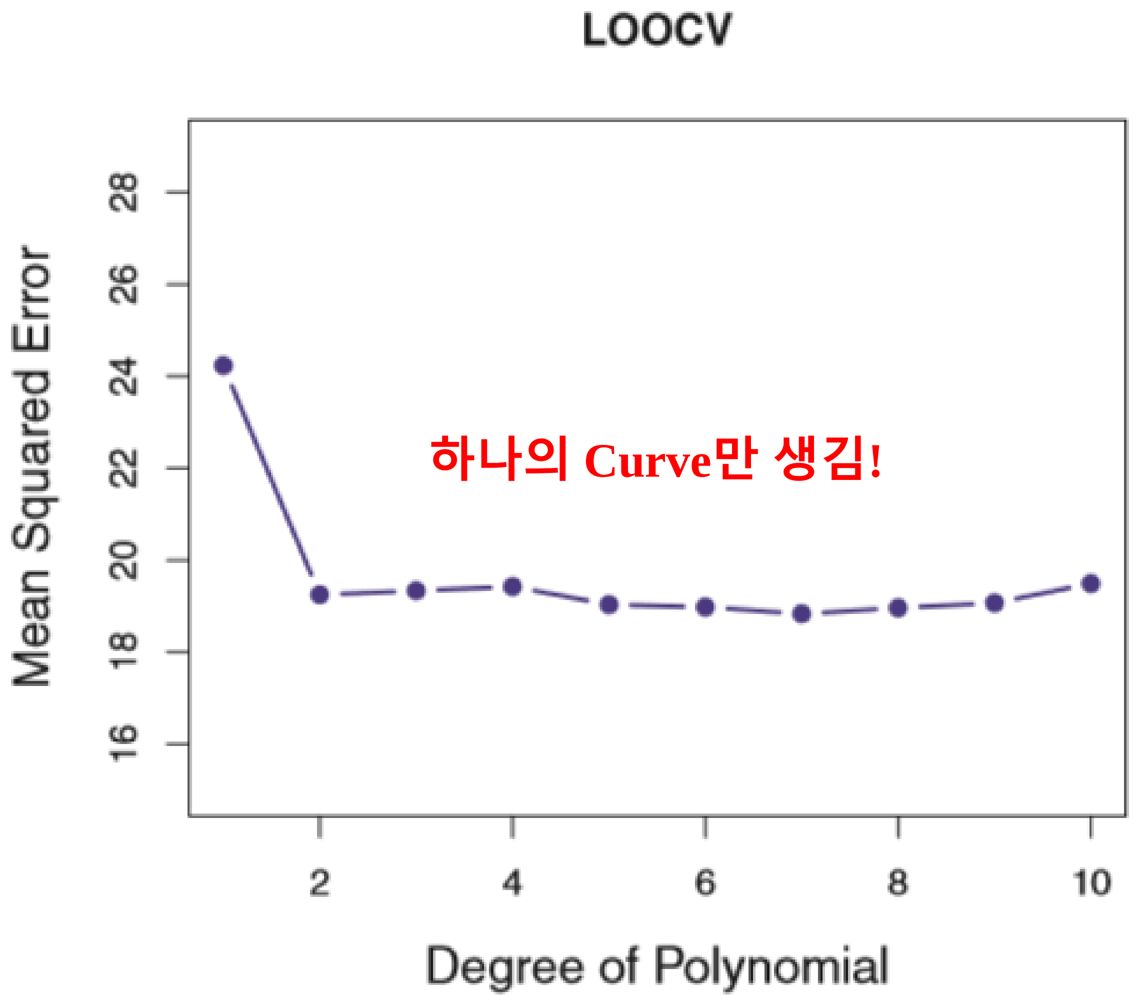
Validation Set

단점 제거 → LOOCV (Leave-One-Out Cross-Validation

하나만 Validation Set으로 씀!

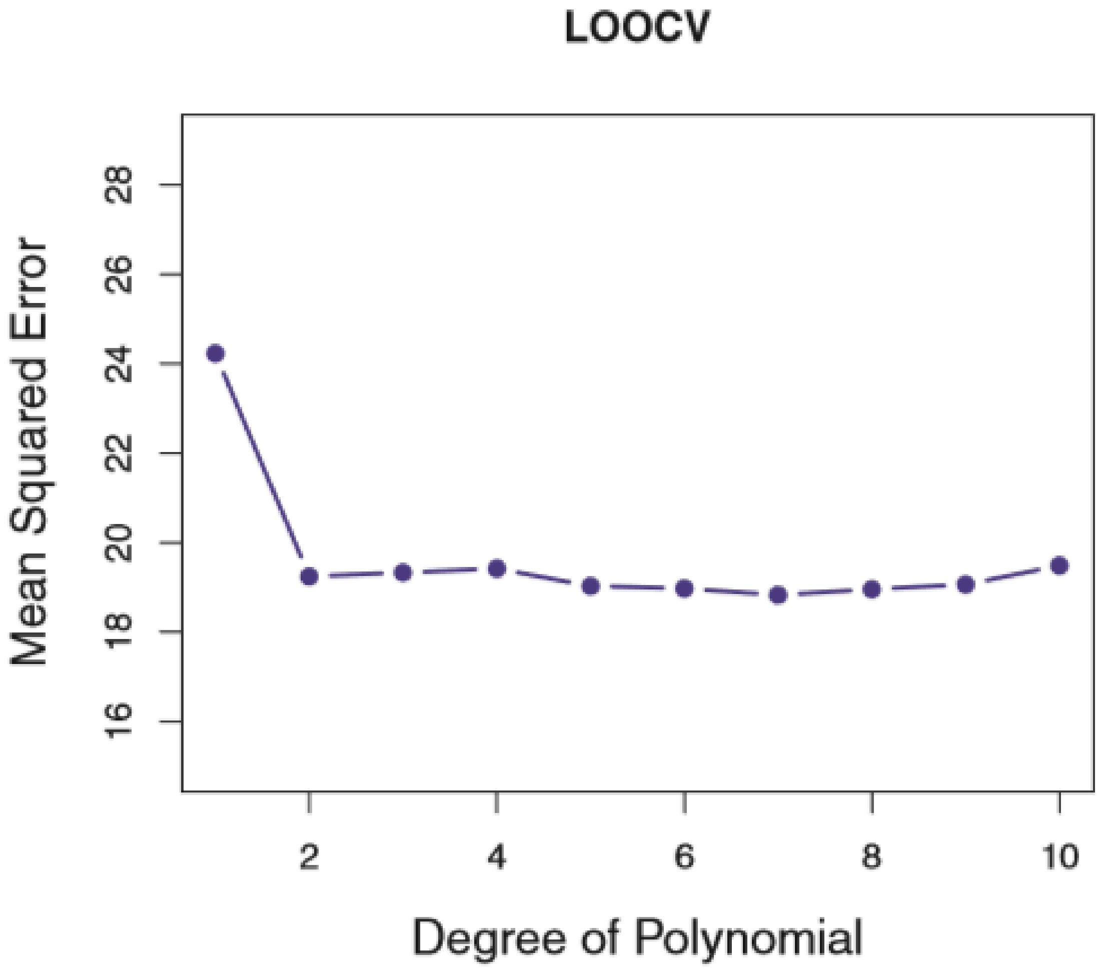


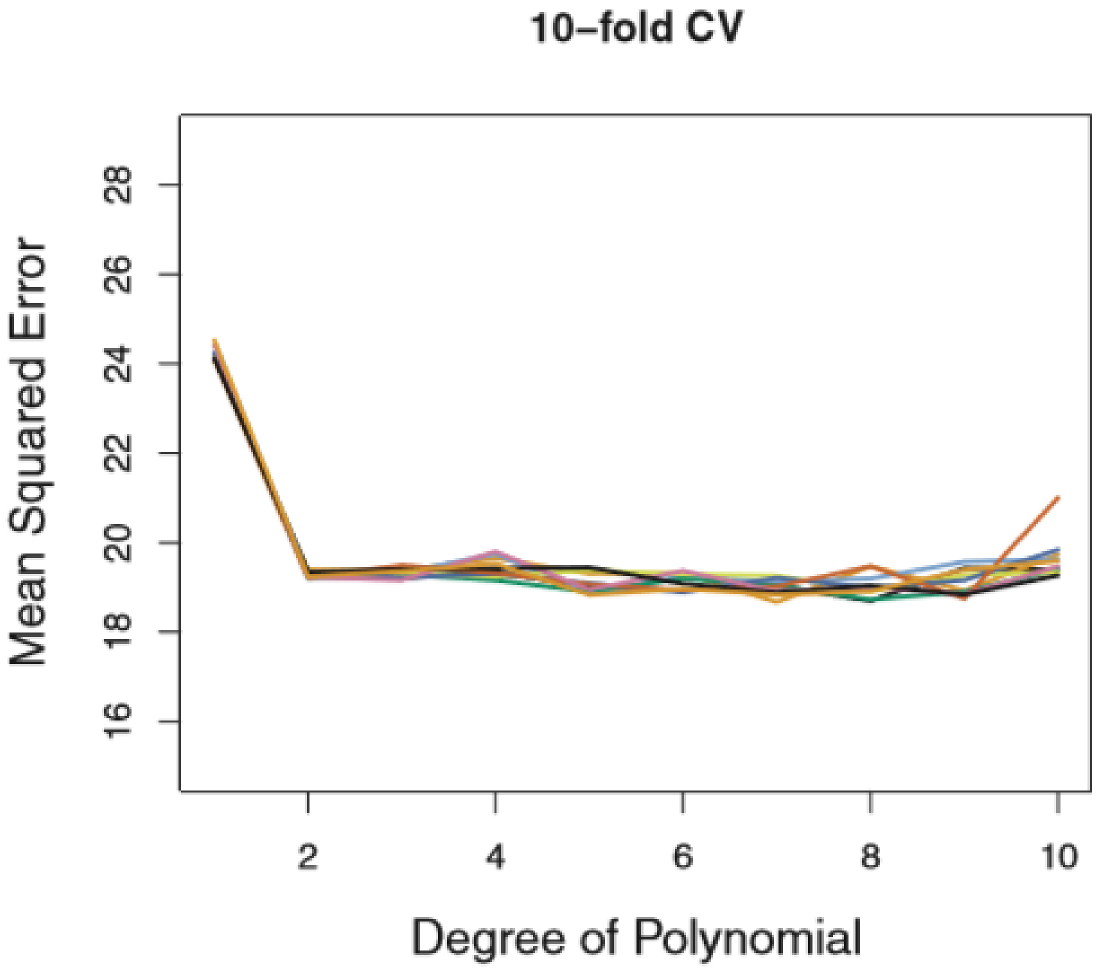
단점 제거 → LOOCV (Leave-One-Out Cross-Validation



$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{MSE}_i$$

정확도는?


LOOCV
Mean Squared Error
16
18
20
22
24
26
28
2
4
6
8
10
Degree of Polynomial


10−fold CV
Mean Squared Error
16
18
20
22
24
26
28
2
4
6
8
10
Degree of Polynomial

SW

특수한 경우의 추정 test MSE: Linear Model Fitting

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \boxed{\mathrm{MSE}_i}$$

Shortcut!

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \boxed{\hat{y}_i}}{1 - \boxed{h_i}}\right)^2$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2}$$

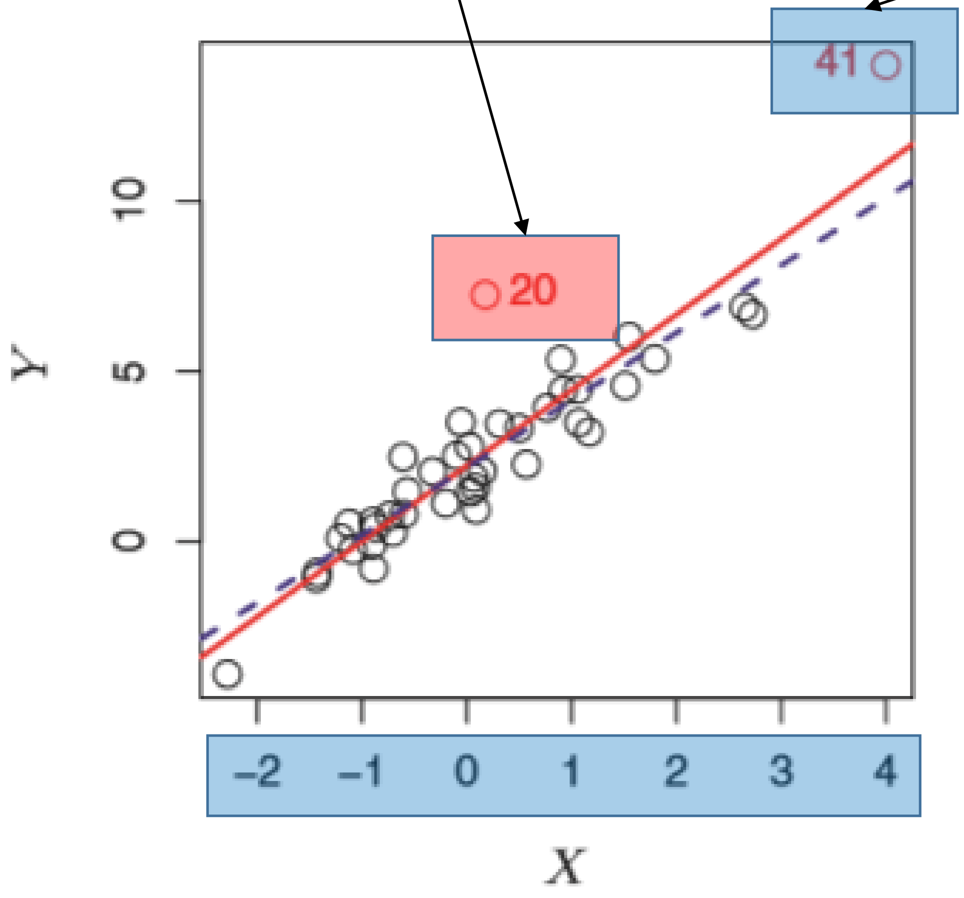
$$\mathrm{MSE}_2 = (y_2 - \boxed{\hat{y}_2})^2$$

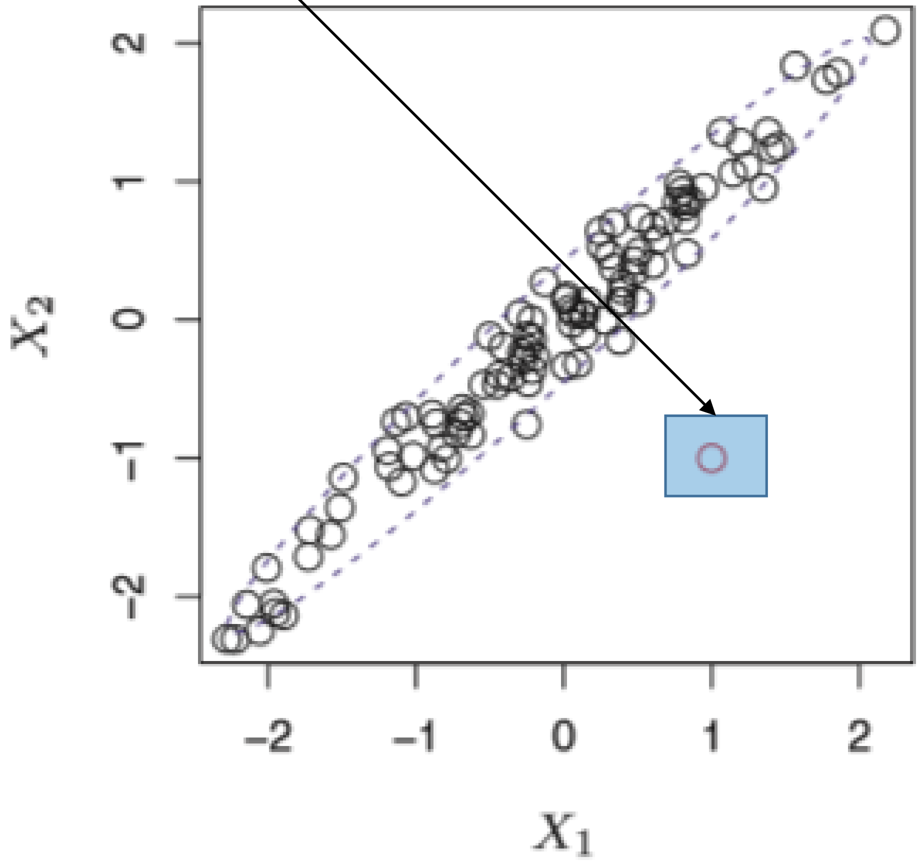
(x_2, y_2)를 포함한 모든 데이터에서 Fitting한 Model에서 예측된 값!

(x_2, y_2)를 제외한 나머지 데이터에서 Fitting한 Model에서 예측된 값!

SW

복습: Outlier VS. Leverage point

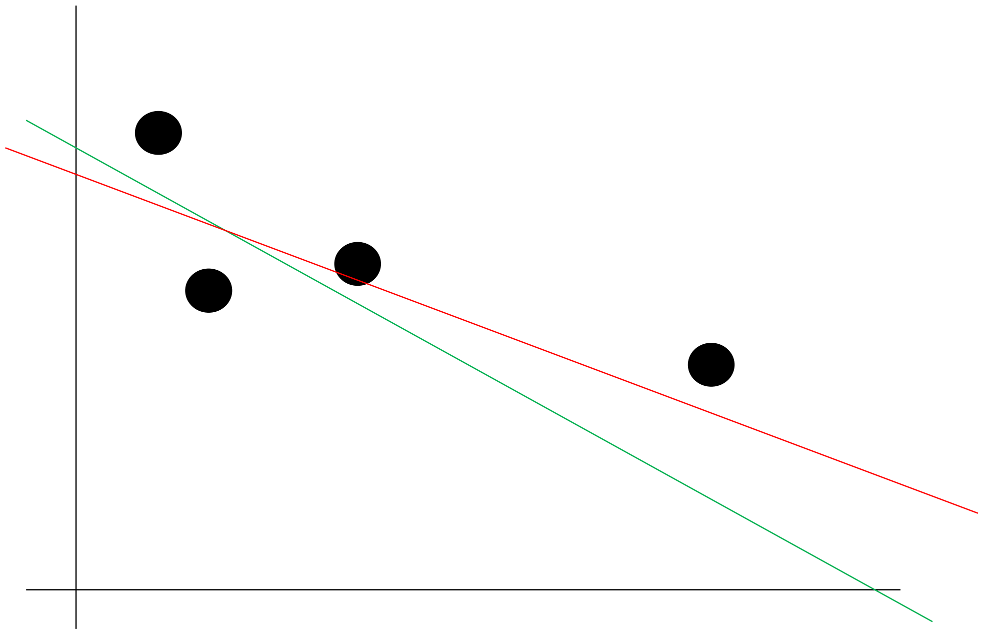


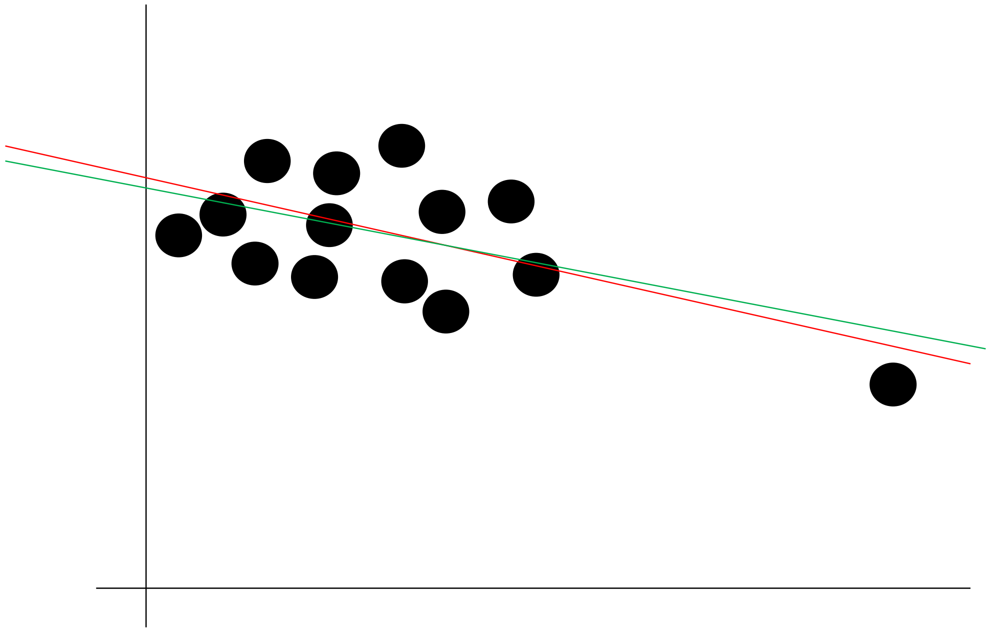


Leverage statistic $$h_i = \frac{1}{n} + \boxed{\frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2}}$$

SW

복습: Outlier VS. Leverage point





Leverage statistic $$h_i = \boxed{\frac{1}{n}} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2}$$

SW

Why shortcut?

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{MSE}_i$$

Shortcut!

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{y_i - \boxed{\hat{y}_i}}{1 - h_i} \right)^2$$

$$\mathrm{MSE}_2 = (y_2 - \boxed{\hat{y}_2})^2$$

(x_2, y_2)를 포함한 모든 데이터에서 Fitting한 Model에서 예측된 값!

(x_2, y_2)를 제외한 나머지 데이터에서 Fitting한 Model에서 예측된 값!

SW

참고 링크: http://efavdb.com/leave-one-out-cross-validation/

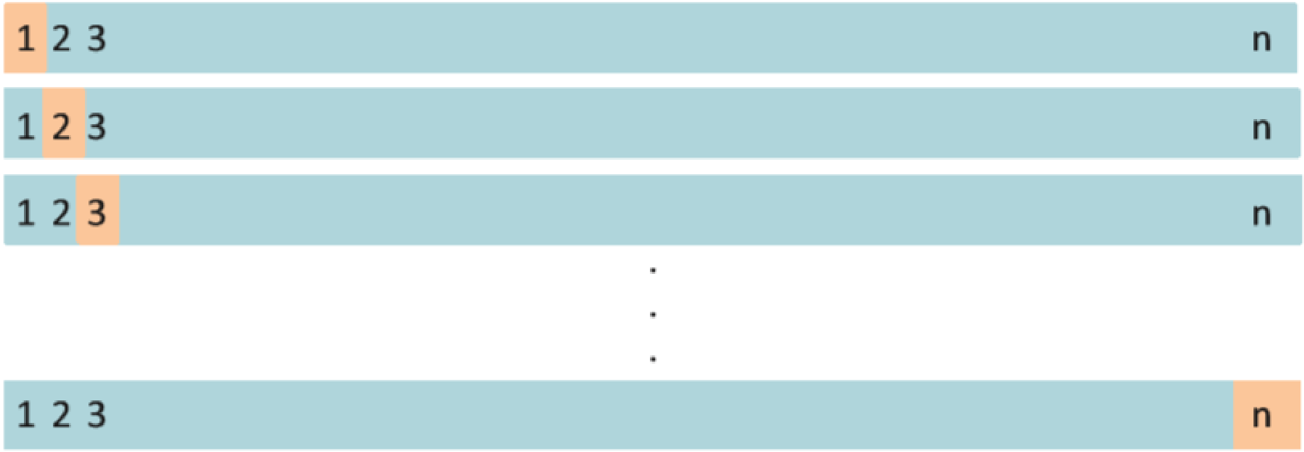
Why shortcut?

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{MSE}_i$$

Shortcut!

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2$$

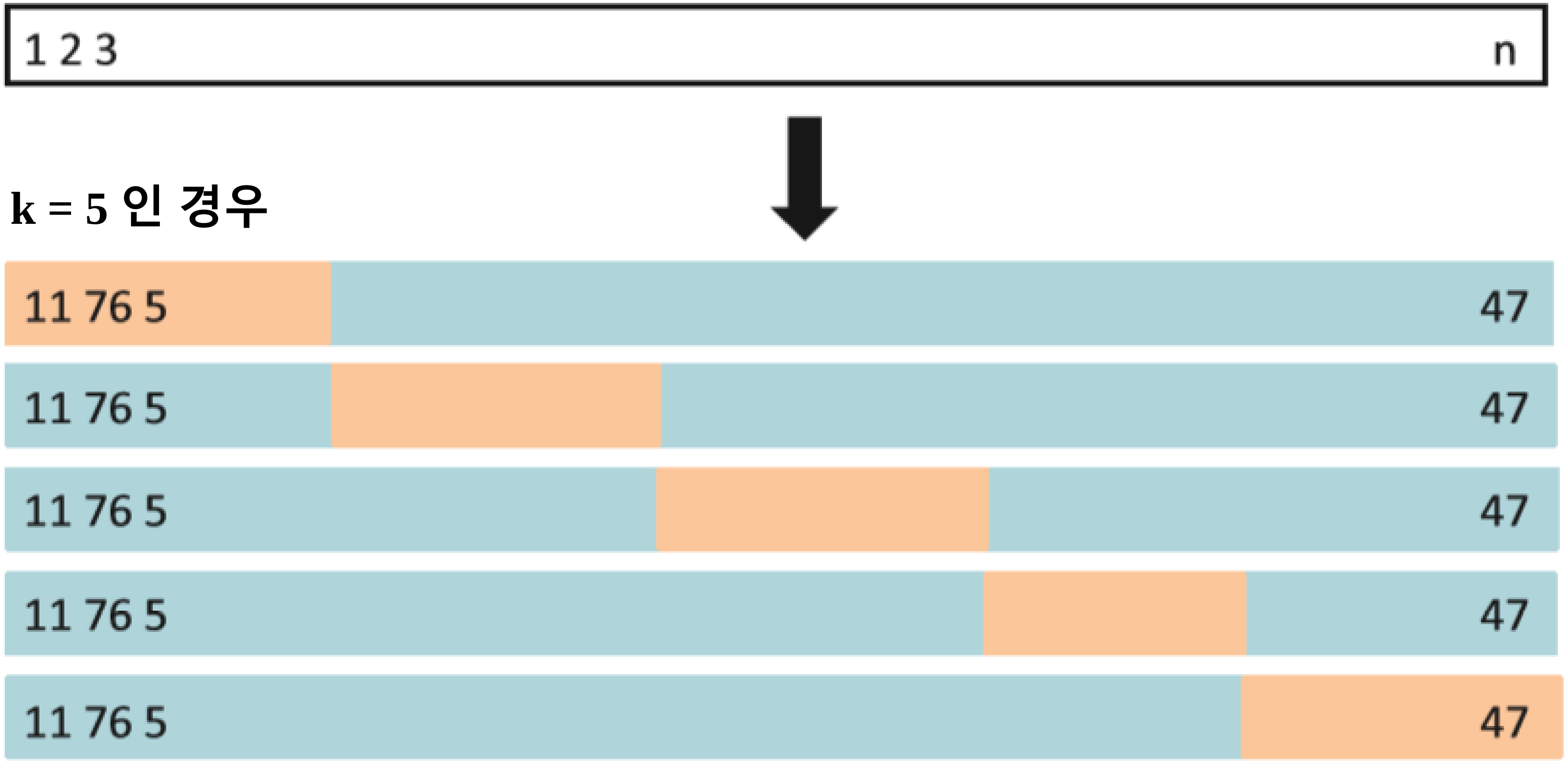
1 2 3 **한 Set만 Fitting 하면 된다!** n



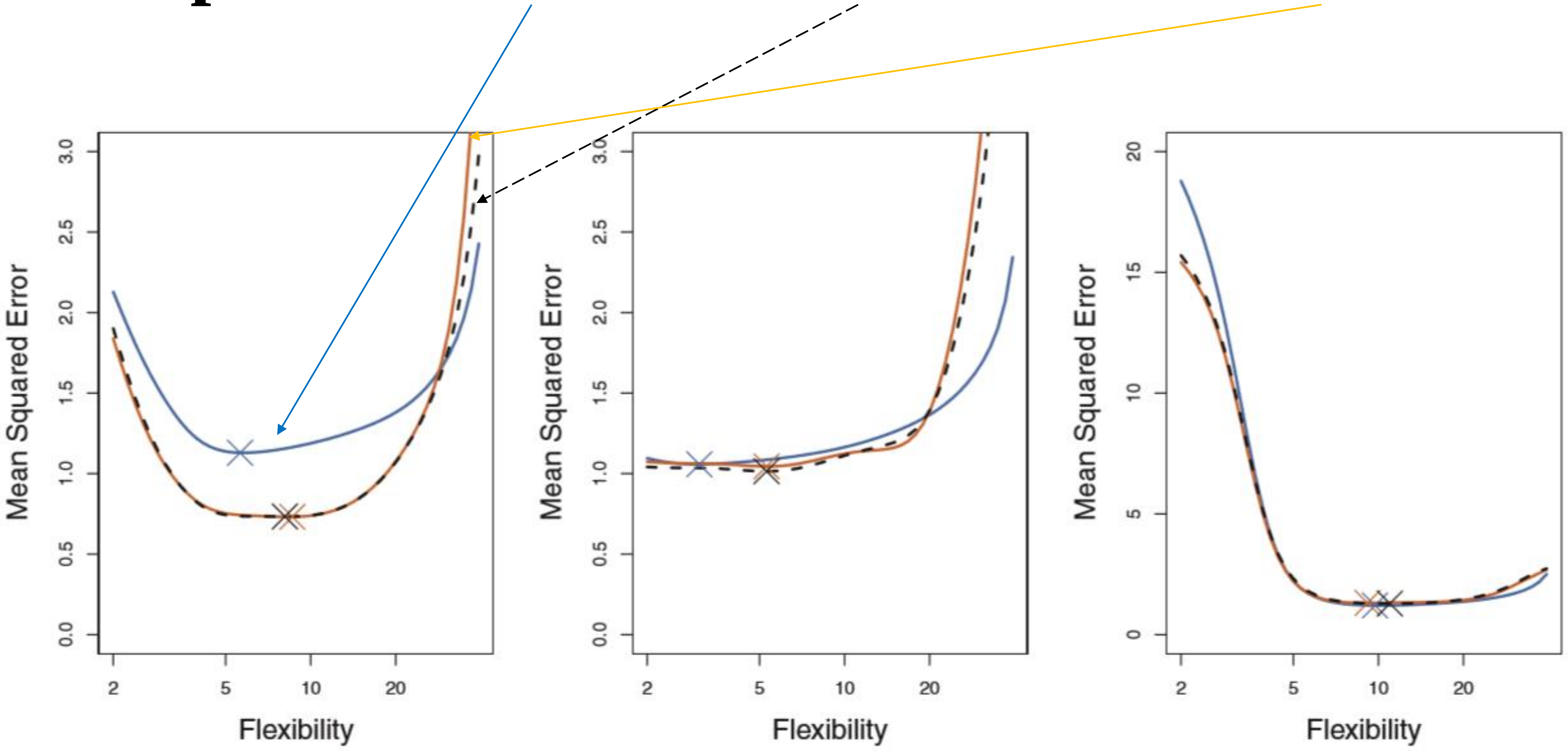
LOOCV는 Linear Model 아닌 경우, 계산 량 너무 크다!

따라서, 더욱 일반적인 형태가 필요!

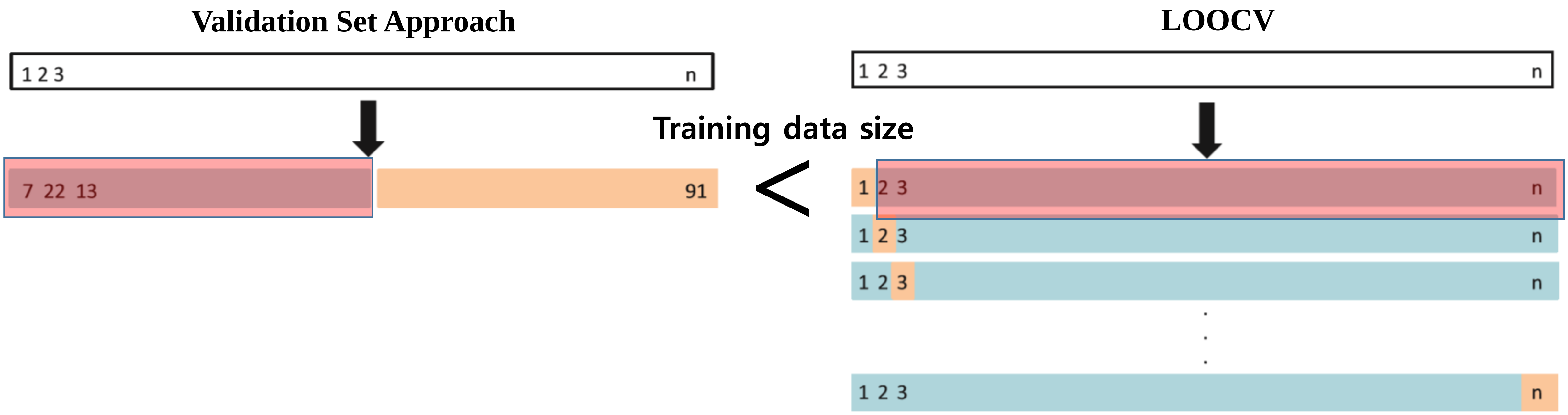
→ **k-Fold Cross-Validation**



Comparison: True VS. LOOCV VS. 10-fold CV

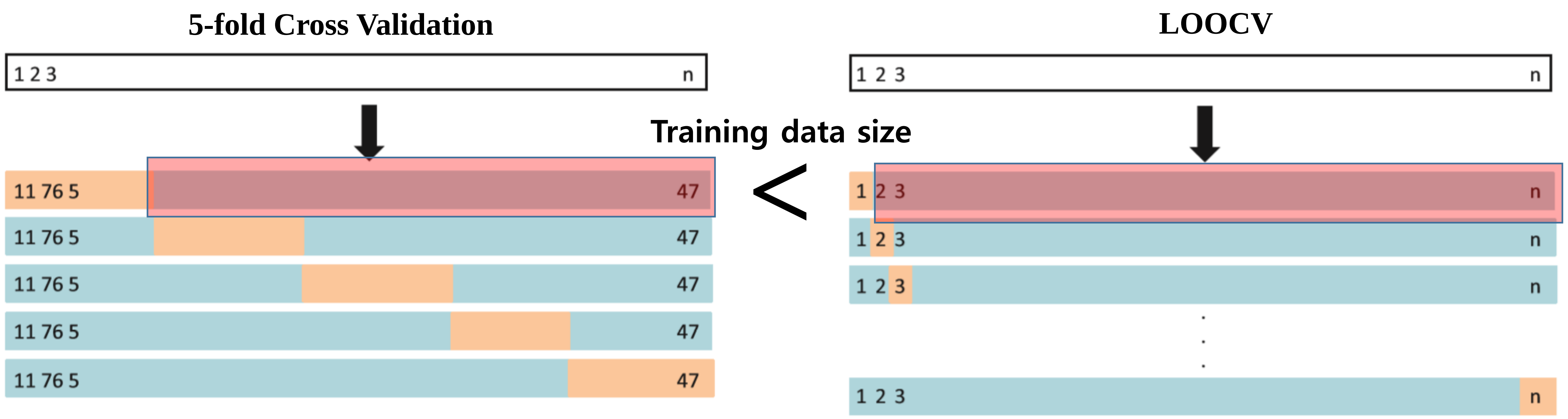


Bias / Variance Trade-off



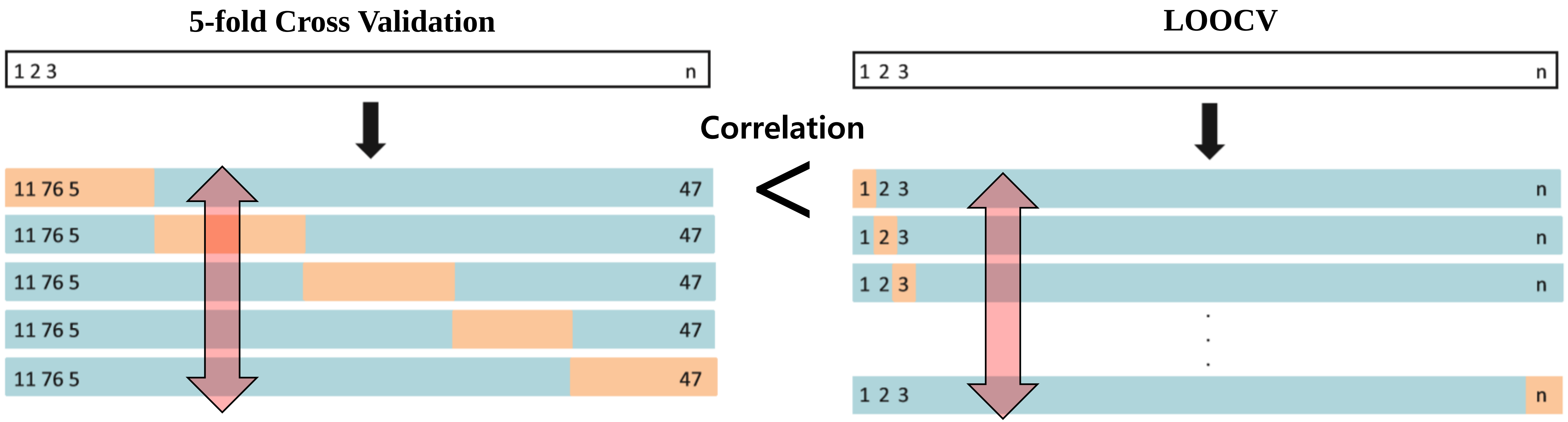
Training에 사용되는 데이터의 수가 많아지면서 Bias는 적어지게 된다.

Bias / Variance Trade-off



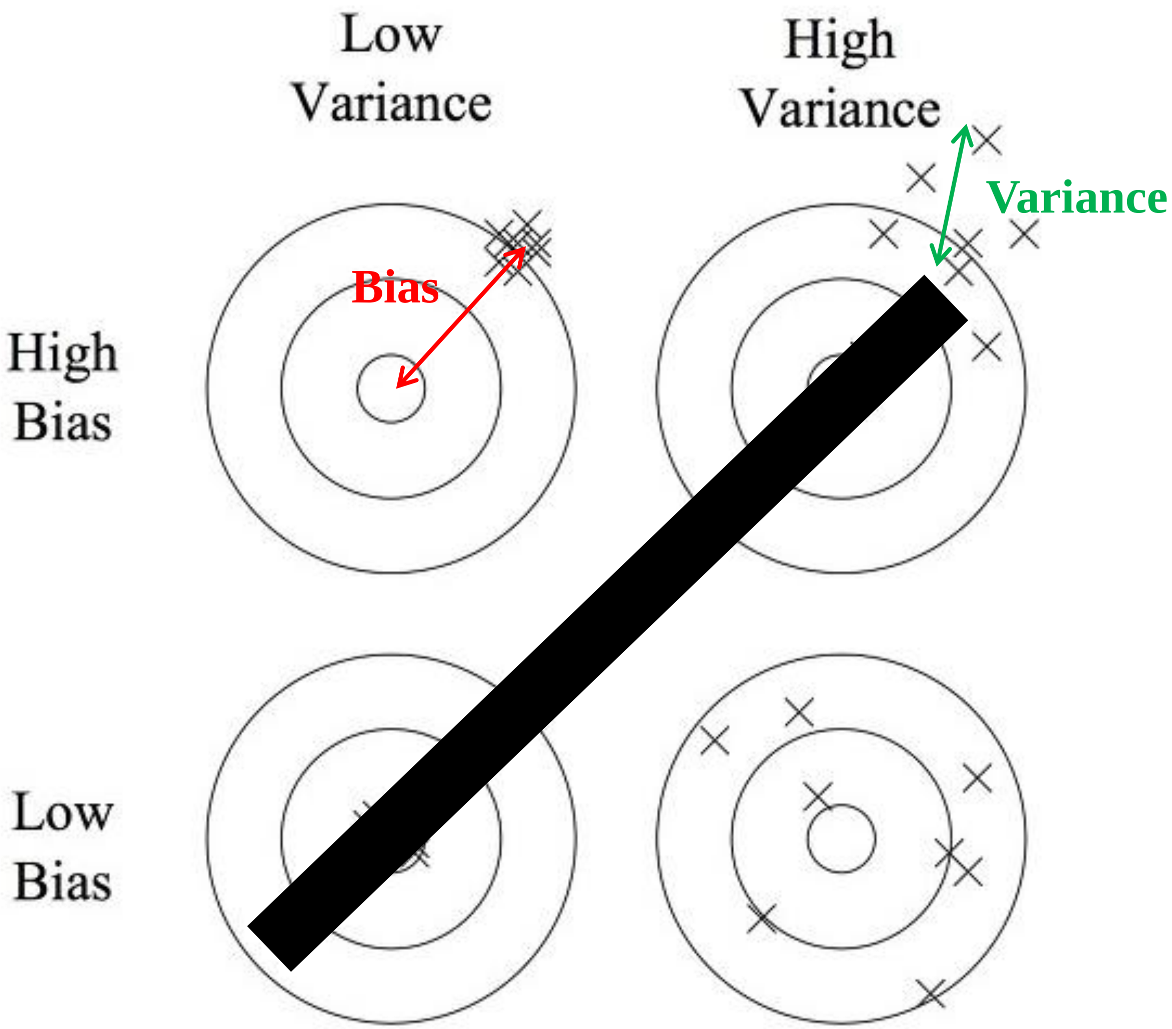
LOOCV가 Bias는 가장 적게 Fitting한다!

Bias / Variance Trade-off



LOOCV가 Variance는 CV 보다 크게 Fitting한다!

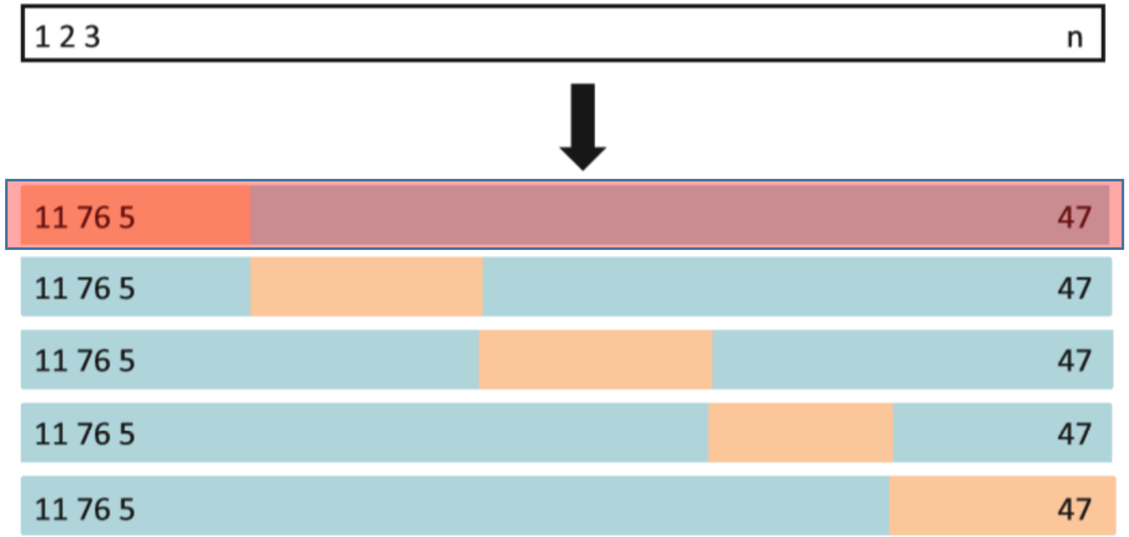
Bias / Variance Trade-off



Bias / Variance Trade-off

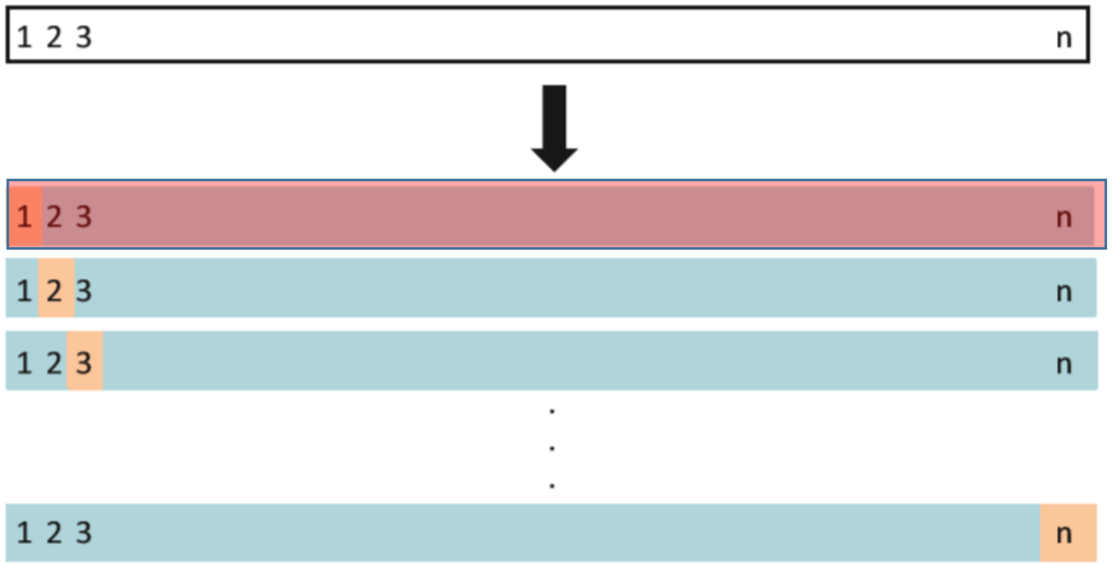
5-fold Cross Validation

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k} \boxed{\mathrm{MSE}_i}$$



LOOCV

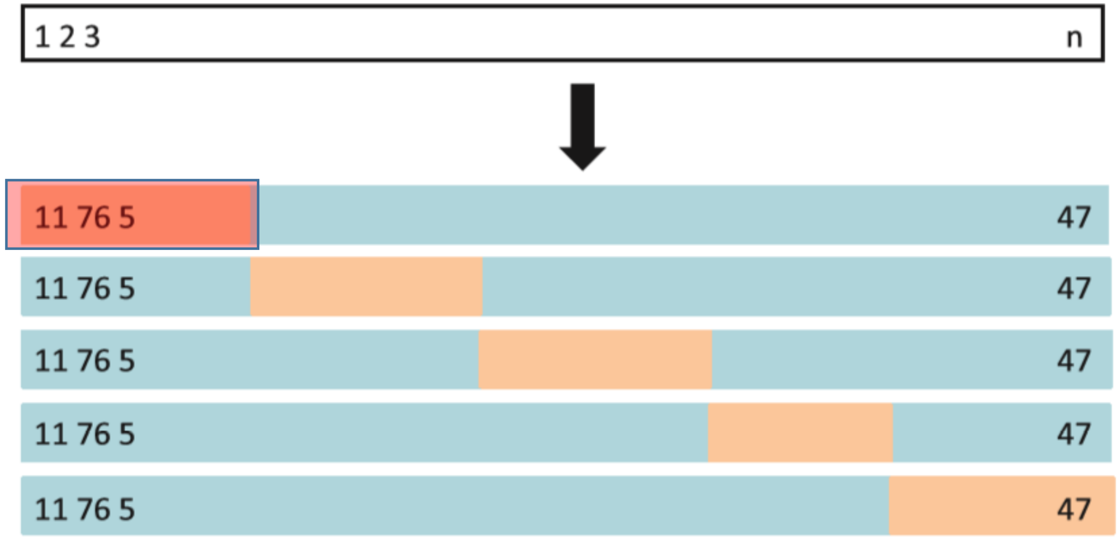
$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \boxed{\mathrm{MSE}_i}$$



Bias / Variance Trade-off

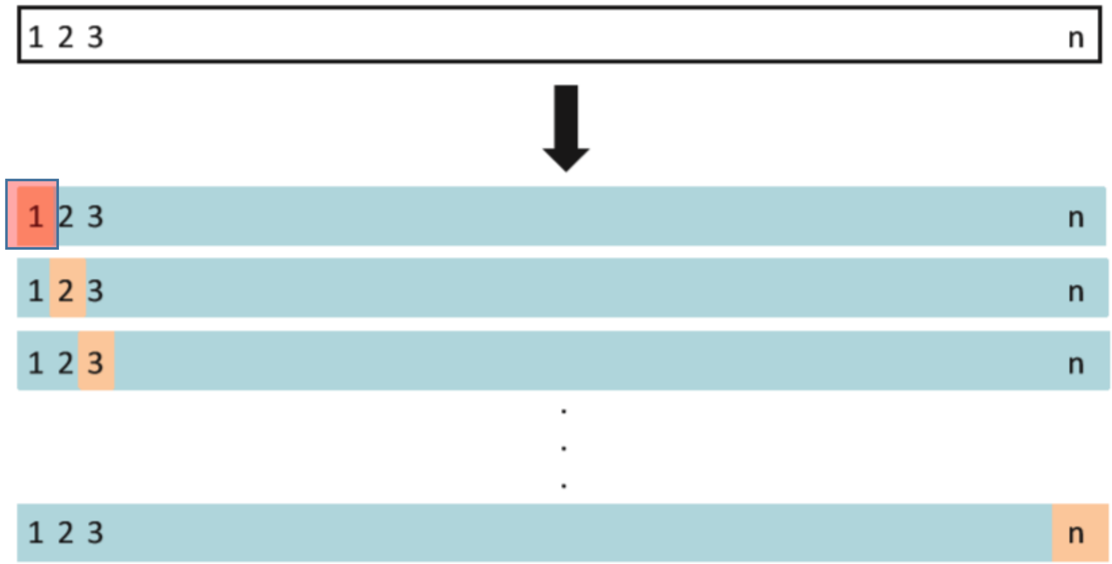
5-fold Cross Validation

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k} \boxed{\mathrm{MSE}_i}$$



LOOCV

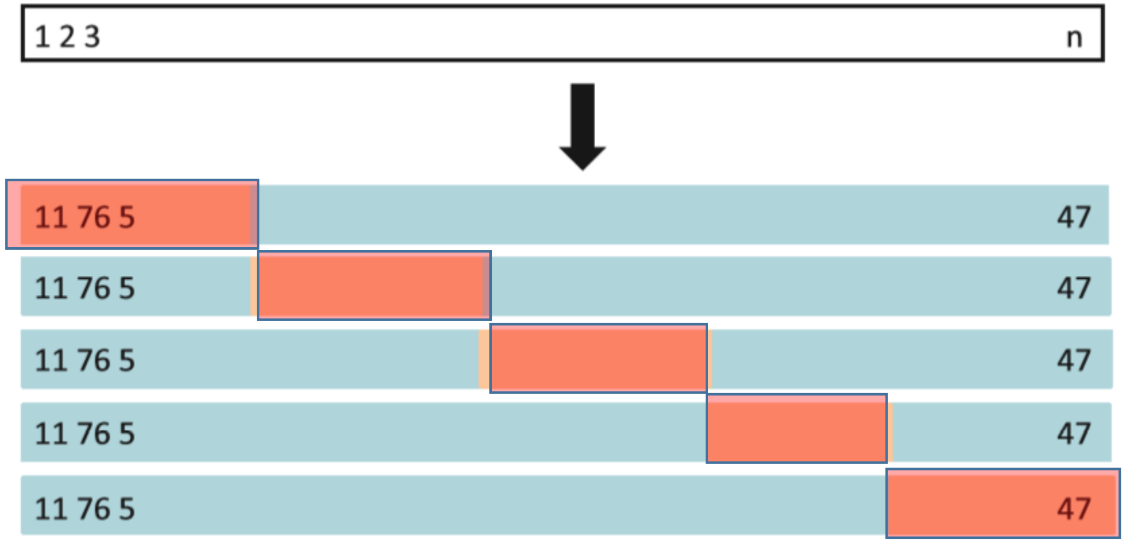
$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \boxed{\mathrm{MSE}_i}$$



Bias / Variance Trade-off

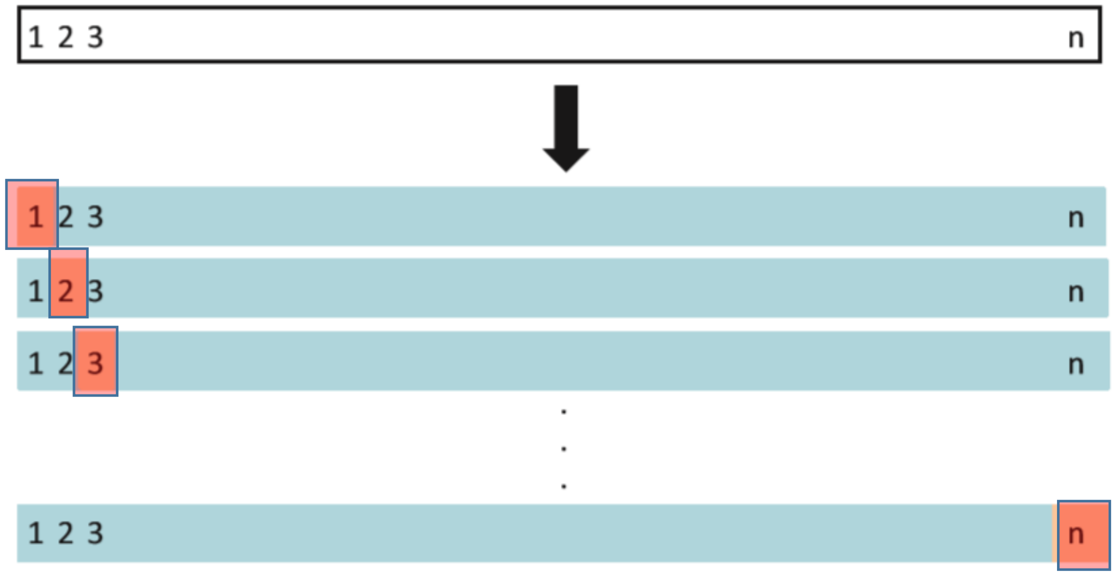
5-fold Cross Validation

$$\mathrm{CV}_{(k)} = \boxed{\frac{1}{k}\sum_{i=1}^{k}}\mathrm{MSE}_i$$



LOOCV

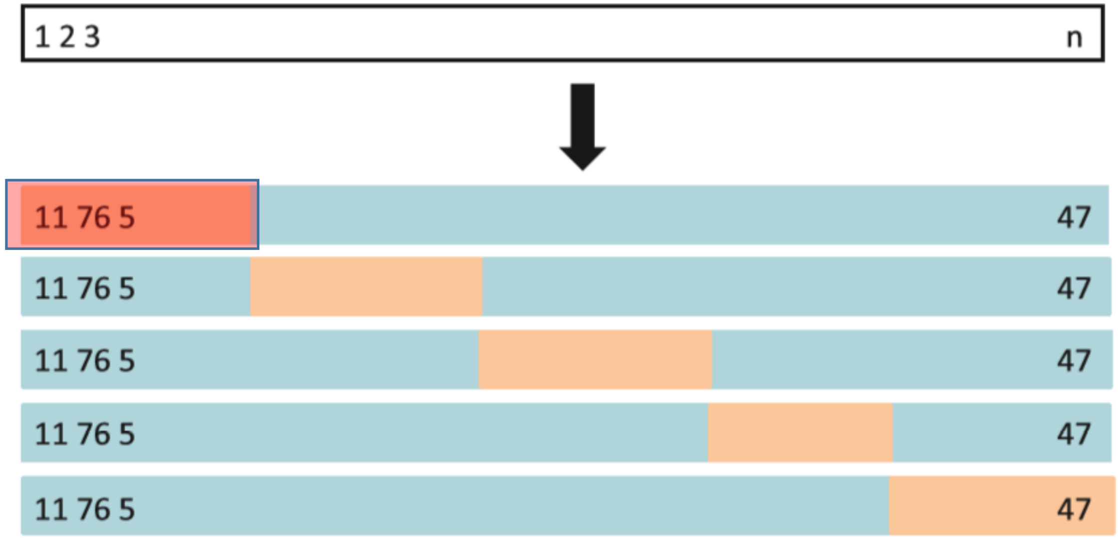
$$\mathrm{CV}_{(n)} = \boxed{\frac{1}{n}\sum_{i=1}^{n}}\mathrm{MSE}_i$$

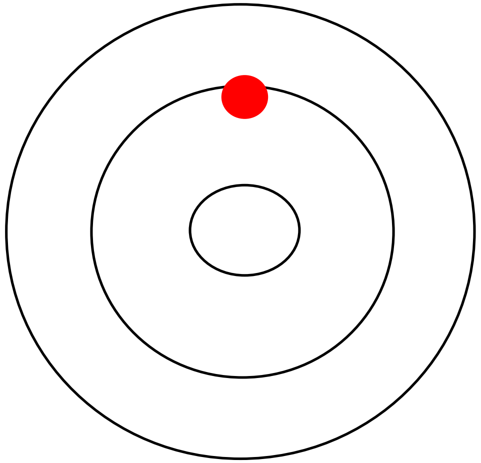


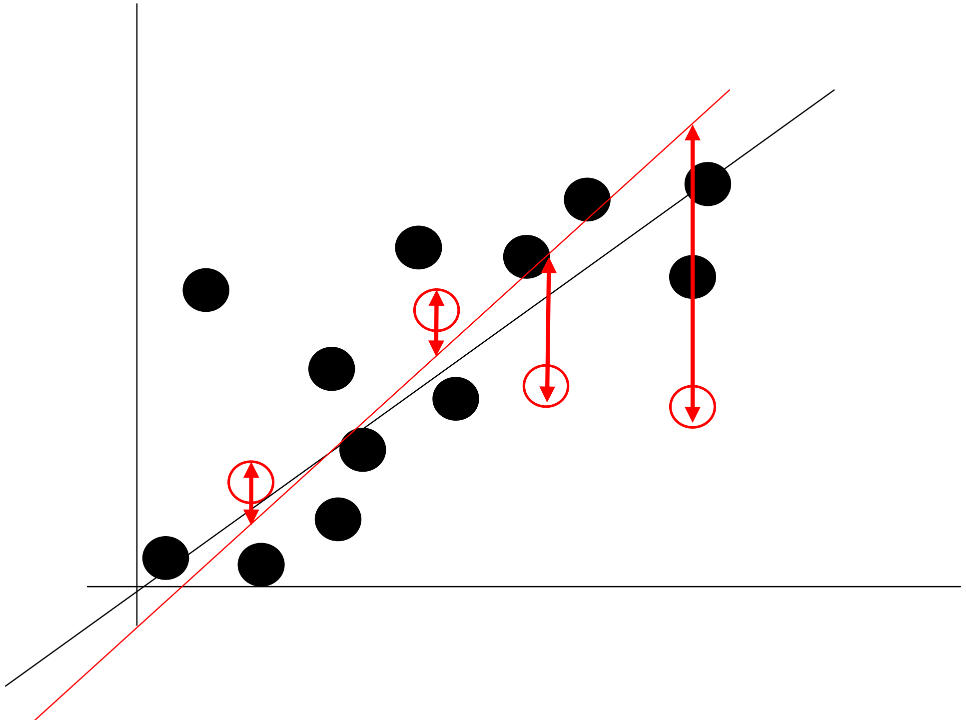
Bias / Variance Trade-off

5-fold Cross Validation

$$\mathrm{CV}_{(k)} = \boxed{\frac{1}{k}\sum_{i=1}^{k}}\mathrm{MSE}_i$$



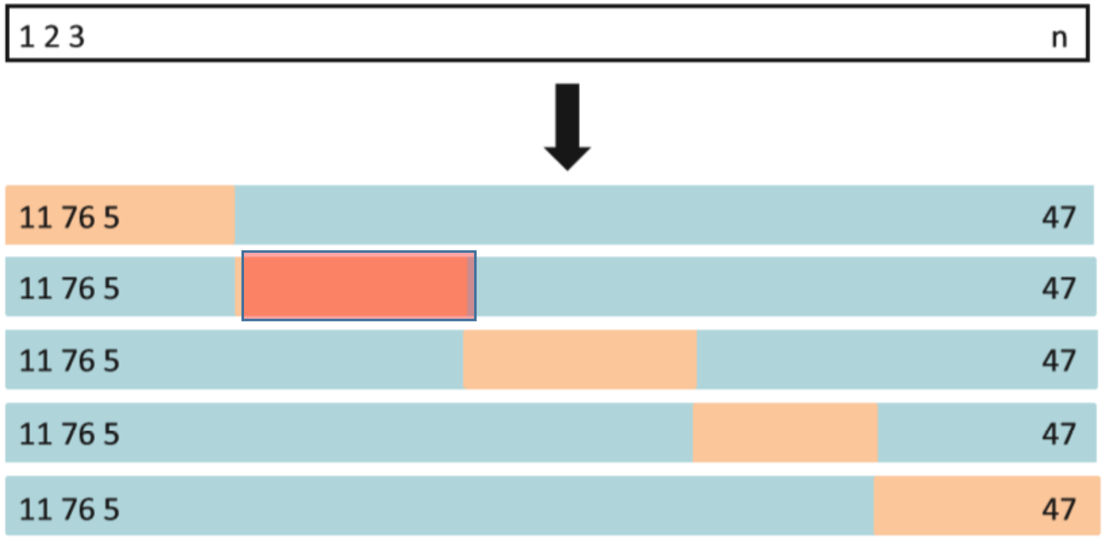


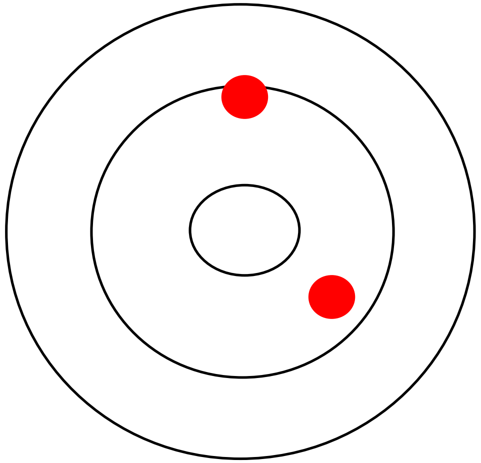


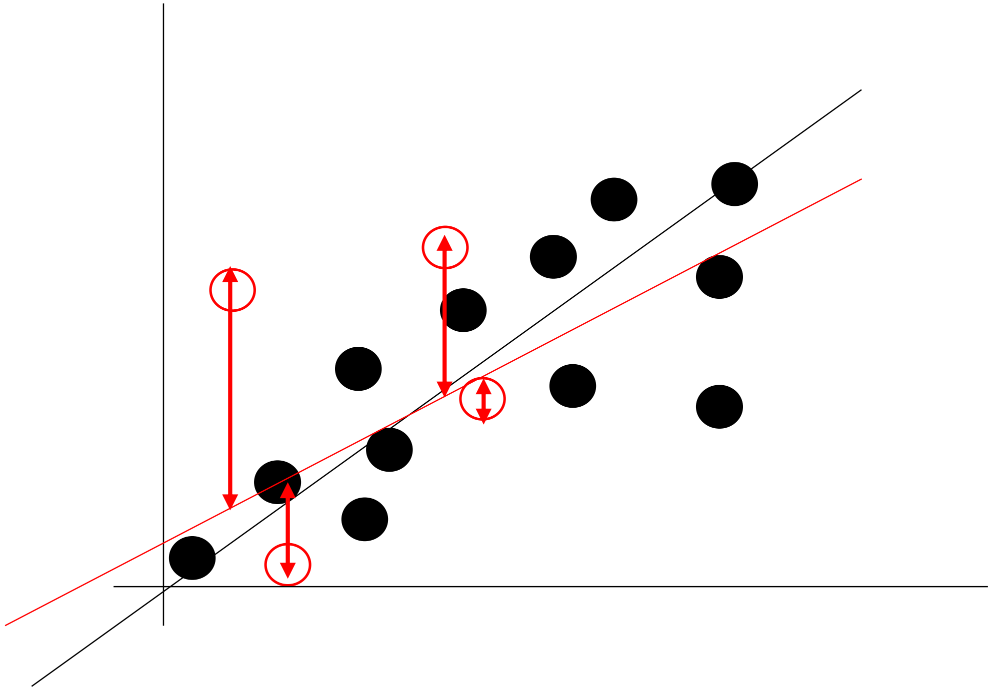
Bias / Variance Trade-off

5-fold Cross Validation

$$\mathrm{CV}_{(k)} = \boxed{\frac{1}{k}\sum_{i=1}^{k}}\mathrm{MSE}_i$$





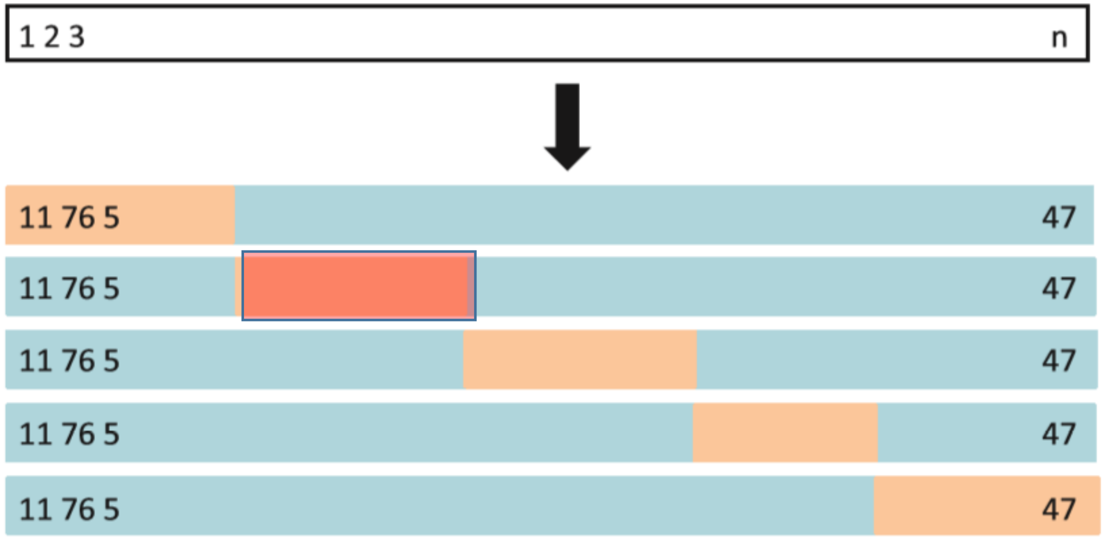


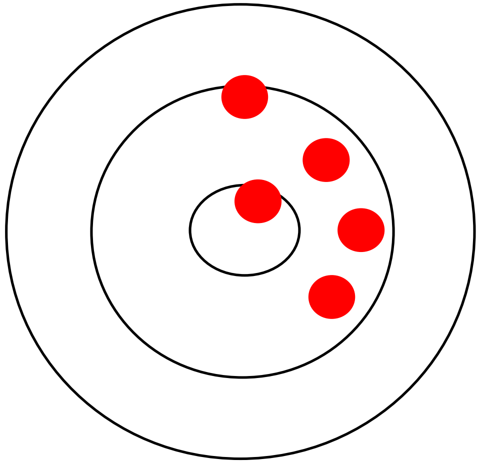
SW

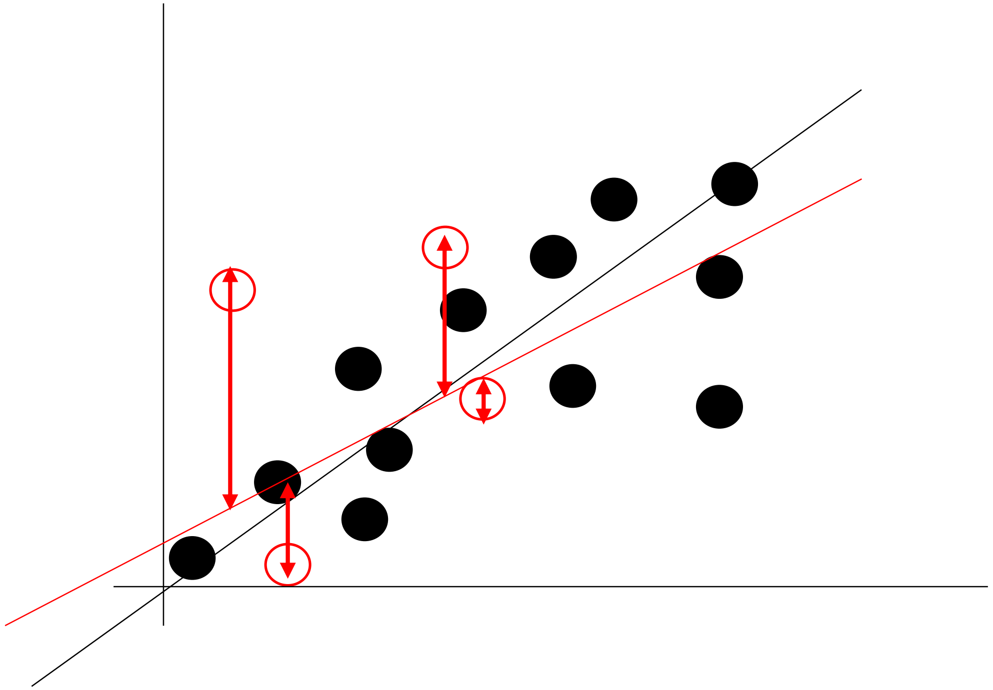
Bias / Variance Trade-off

5-fold Cross Validation

$$\mathrm{CV}_{(k)} = \boxed{\frac{1}{k}\sum_{i=1}^{k}}\mathrm{MSE}_i$$

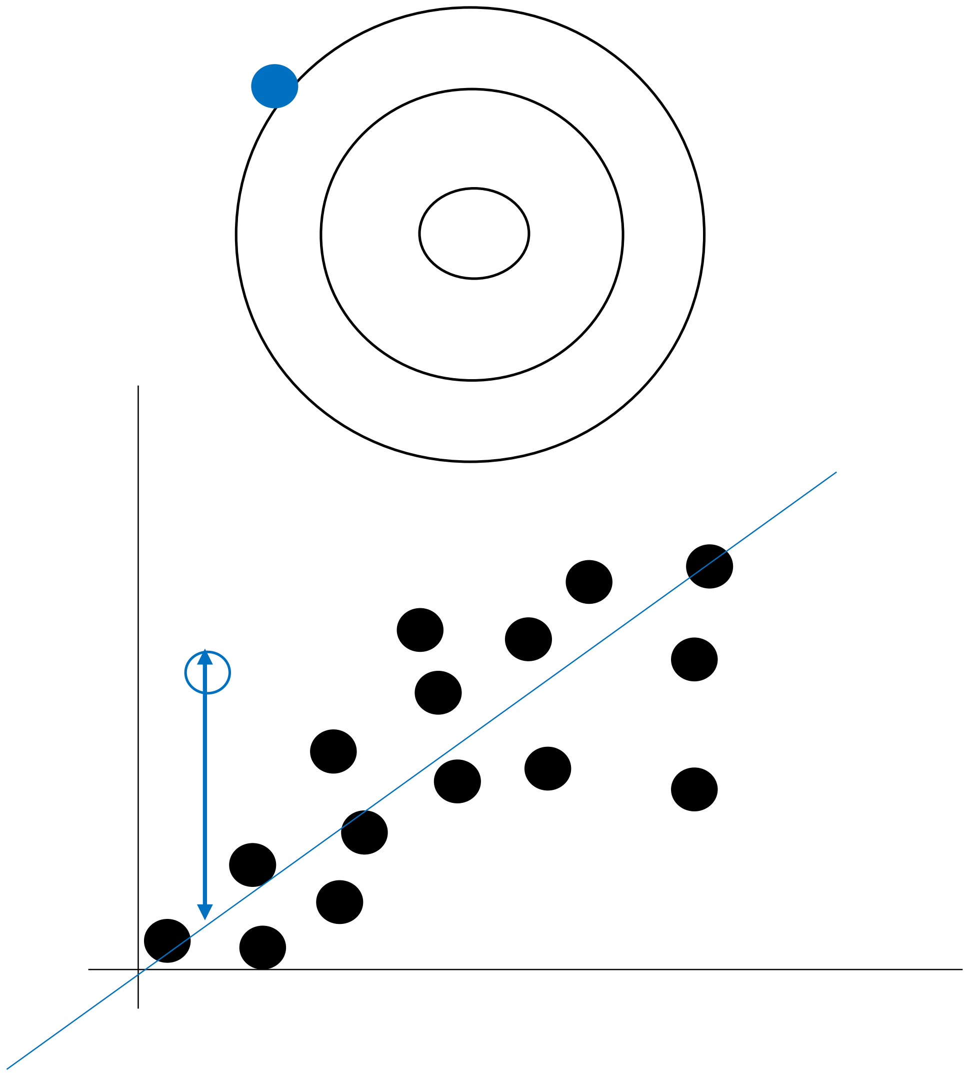






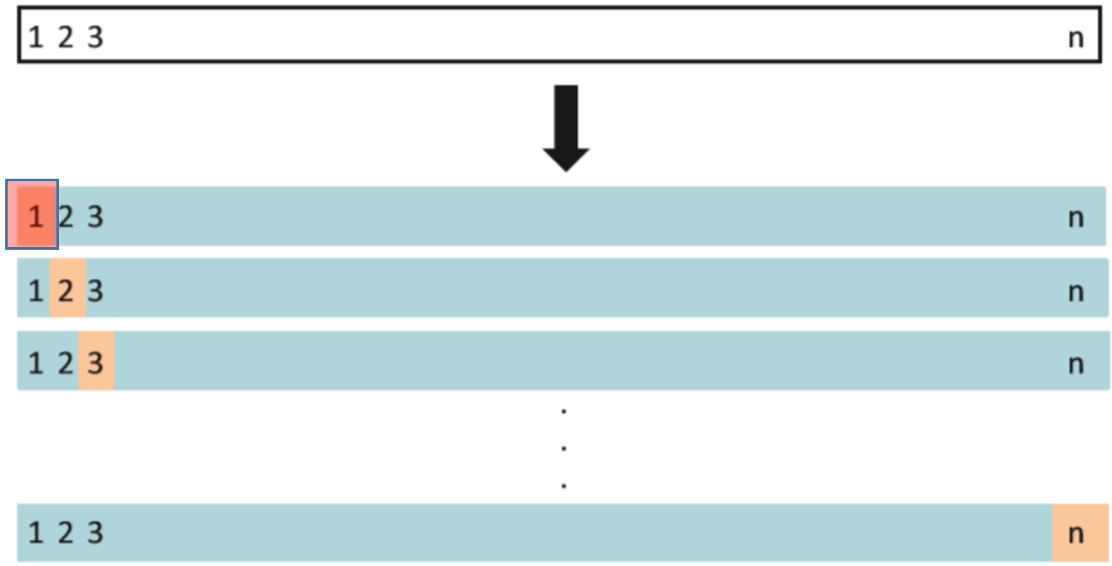
SW

Bias / Variance Trade-off

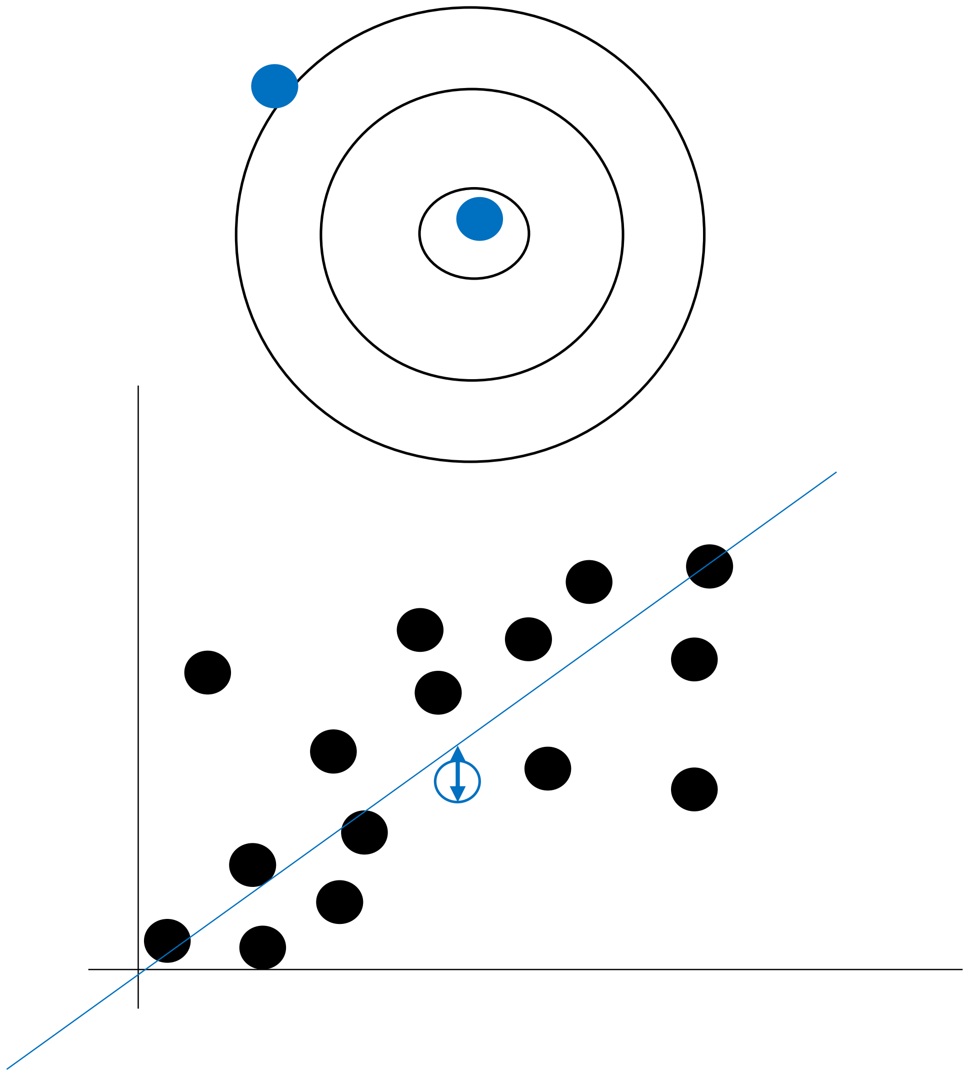


LOOCV

$$CV_{(n)} = \boxed{\frac{1}{n}\sum_{i=1}^{n}} MSE_i$$

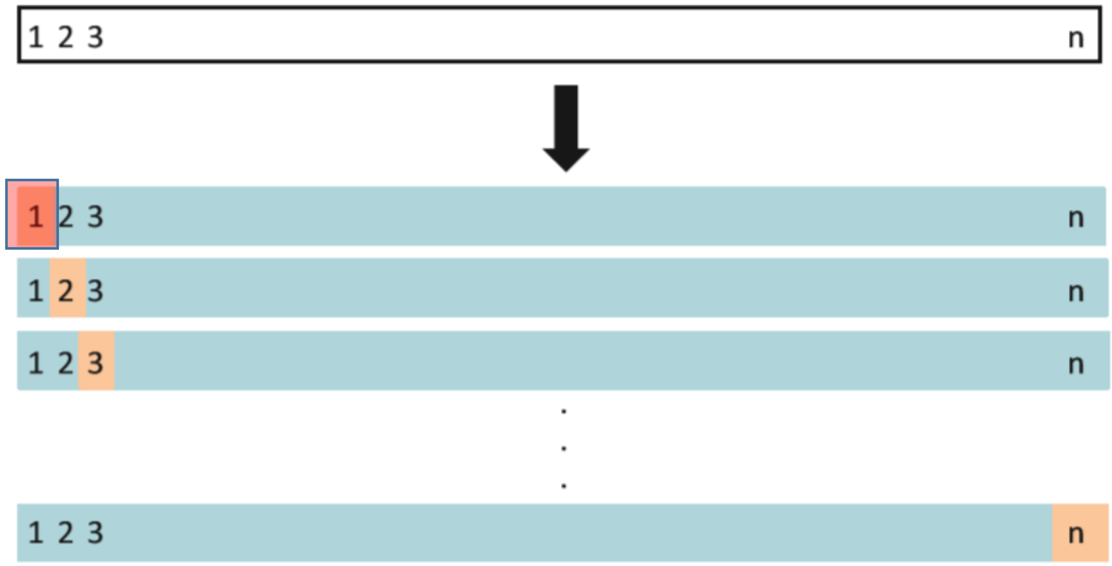


Bias / Variance Trade-off



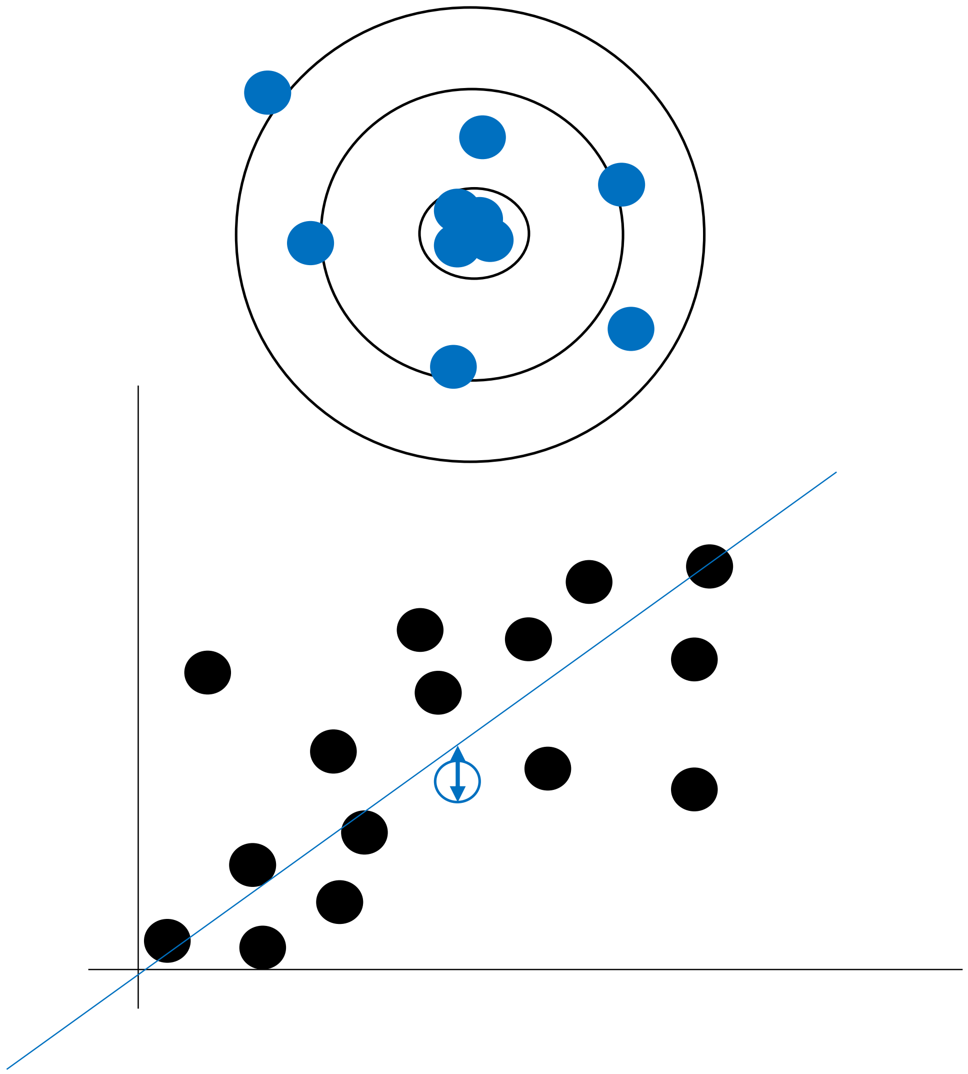
LOOCV

$$CV_{(n)} = \boxed{\frac{1}{n}\sum_{i=1}^{n}} \mathrm{MSE}_i$$



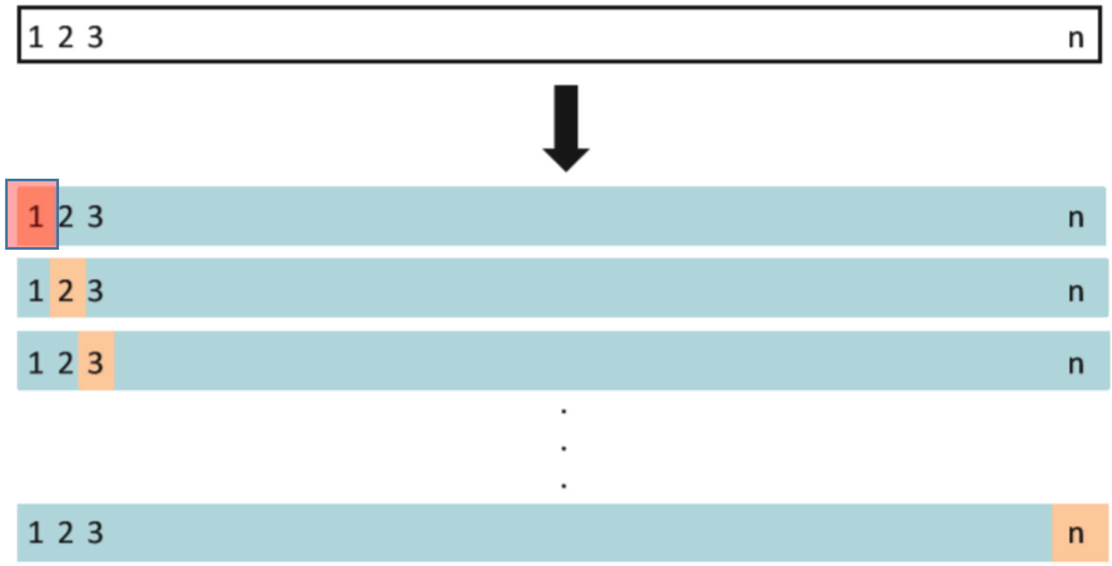
SW

Bias / Variance Trade-off



LOOCV

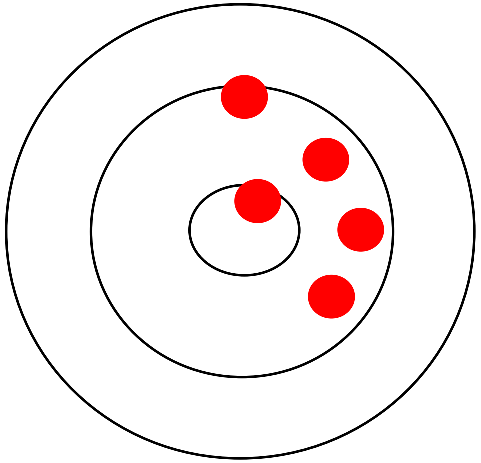
$$\mathrm{CV}_{(n)} = \boxed{\frac{1}{n}\sum_{i=1}^{n}}\mathrm{MSE}_i$$



Bias / Variance Trade-off

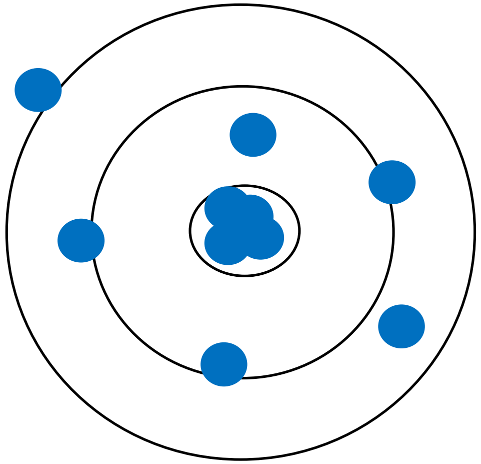
5-fold Cross Validation

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{MSE}_i$$



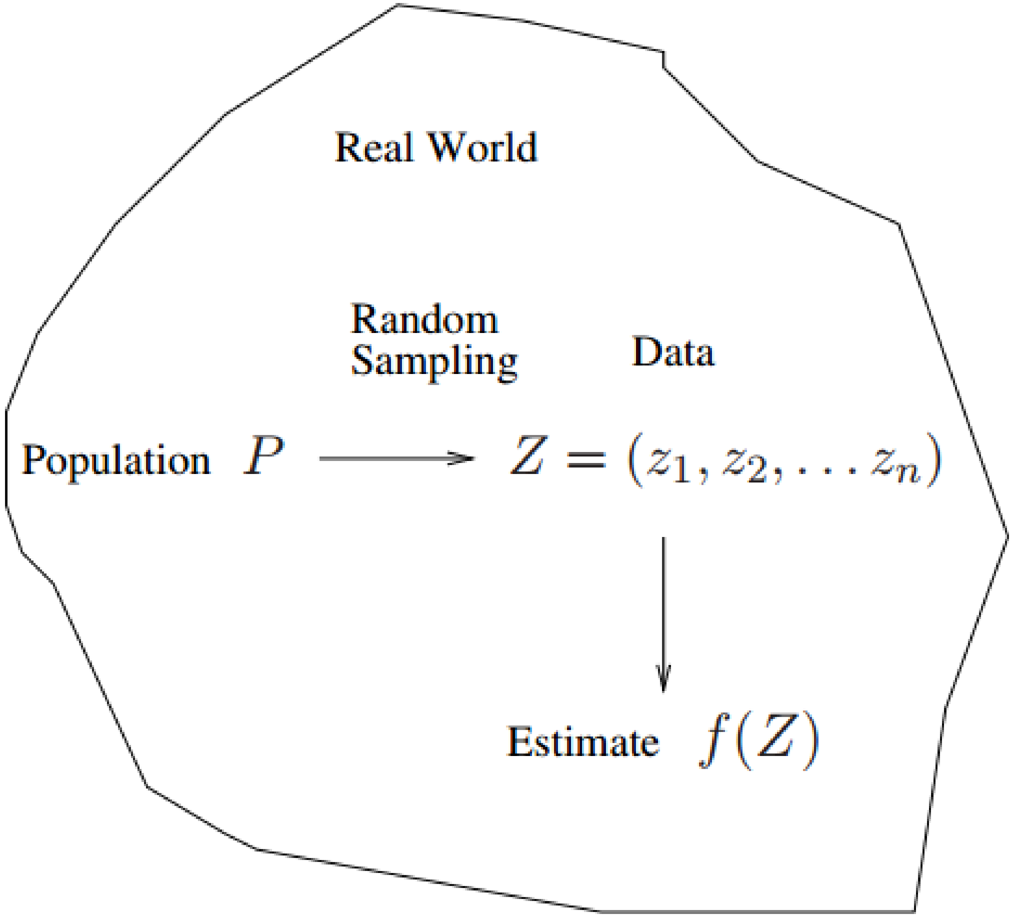
LOOCV

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{MSE}_i$$

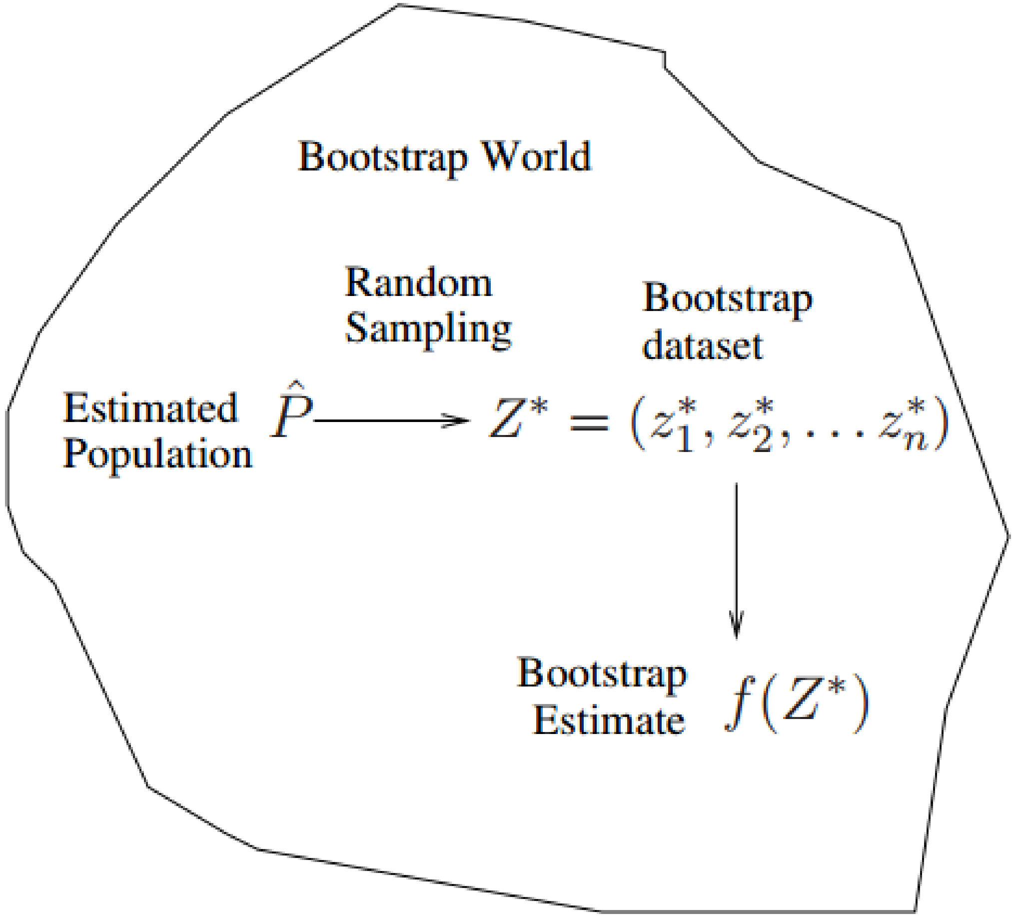


SW

Bootstrap



Real World
Random Sampling
Data
Population P
$Z = (z_1, z_2, \dots z_n)$
Estimate $f(Z)$

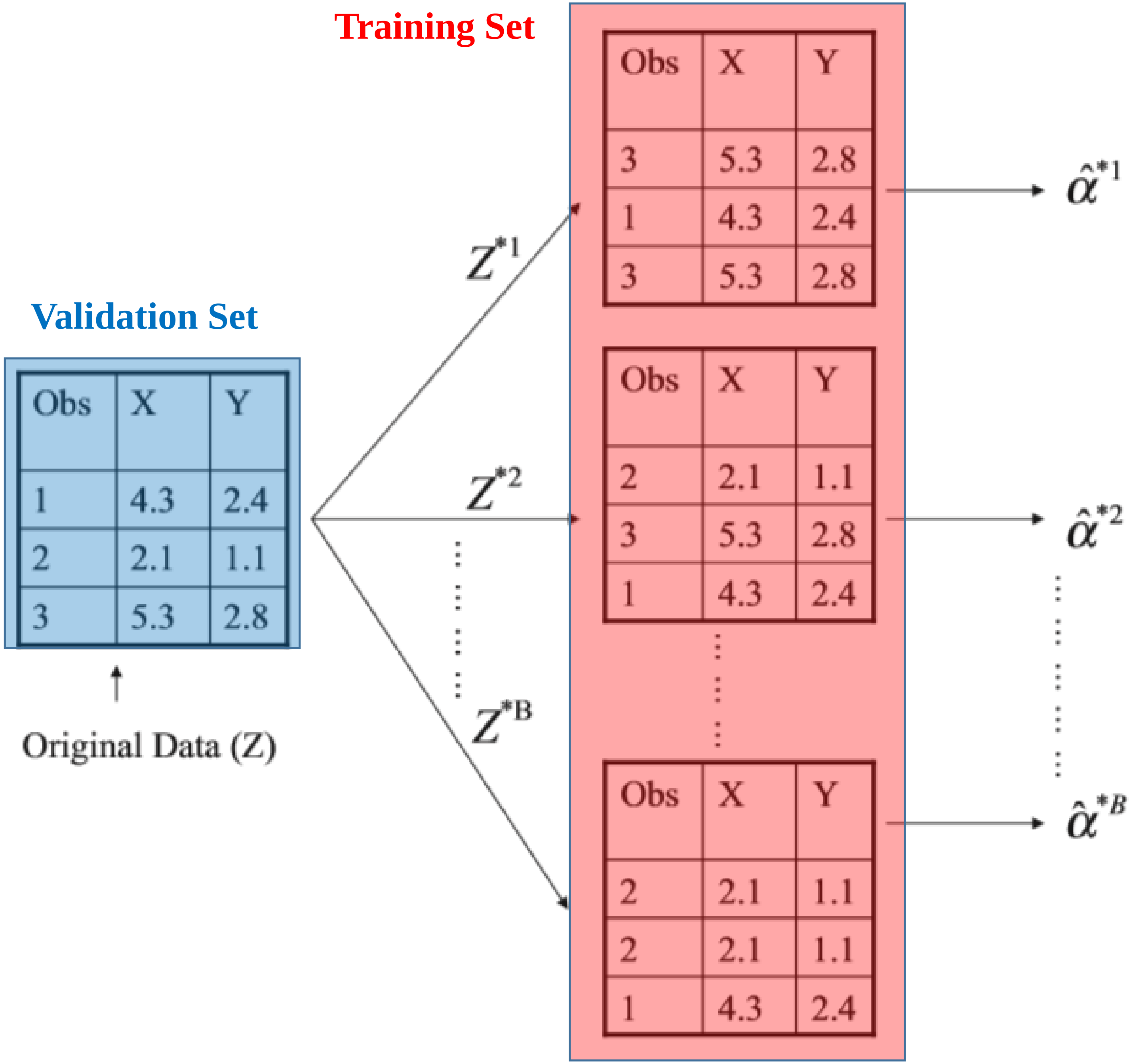


Bootstrap World
Random Sampling
Bootstrap dataset
Estimated Population $\hat{P}$
$Z^* = (z_1^*, z_2^*, \dots z_n^*)$
Bootstrap Estimate $f(Z^*)$

Bootstrap

반복을 허용하고,

관찰 값에서 **Resampling** 한다!



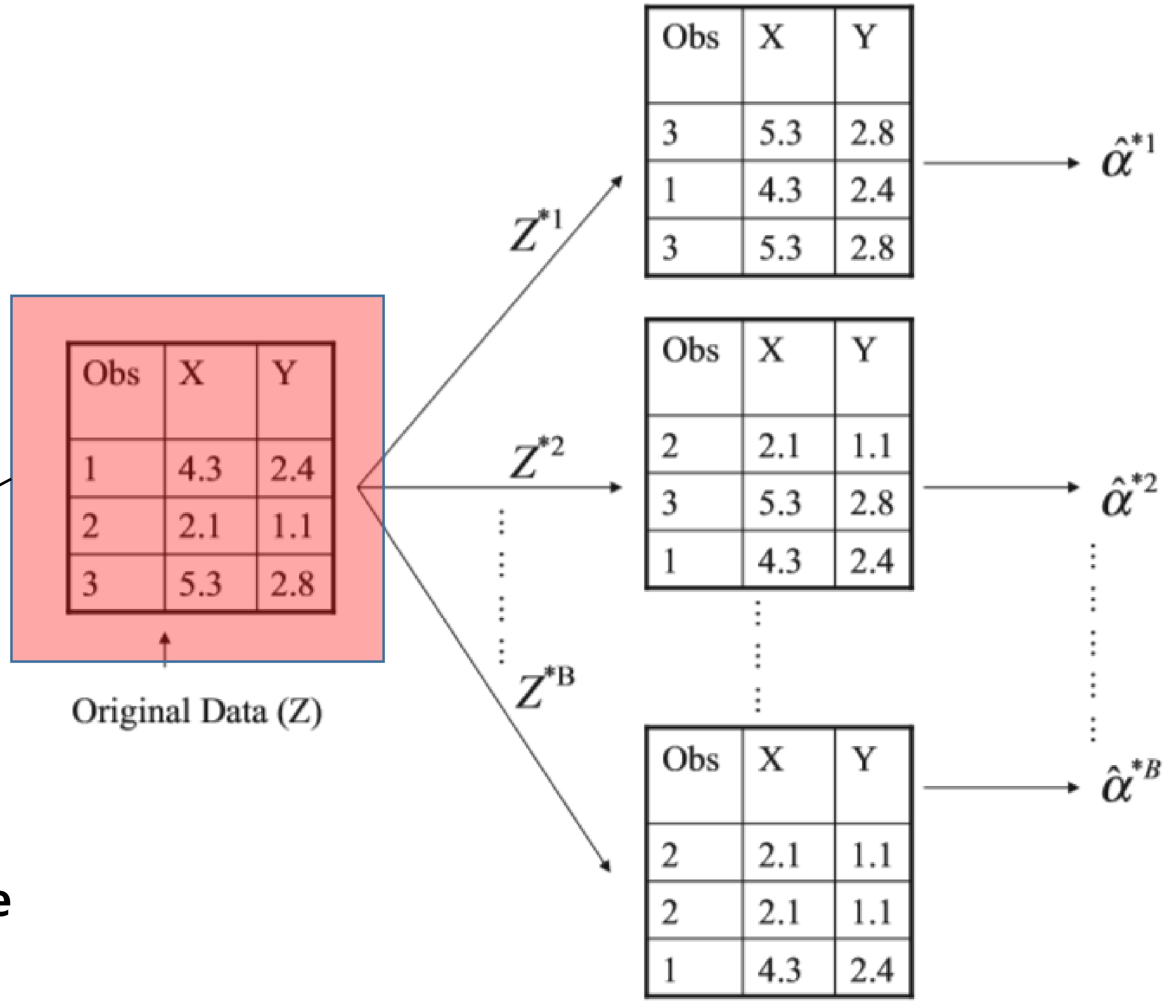
Obs	X	Y
1	4.3	2.4
2	2.1	1.1
3	5.3	2.8

Obs	X	Y
3	5.3	2.8
1	4.3	2.4
3	5.3	2.8

Obs	X	Y
2	2.1	1.1
3	5.3	2.8
1	4.3	2.4

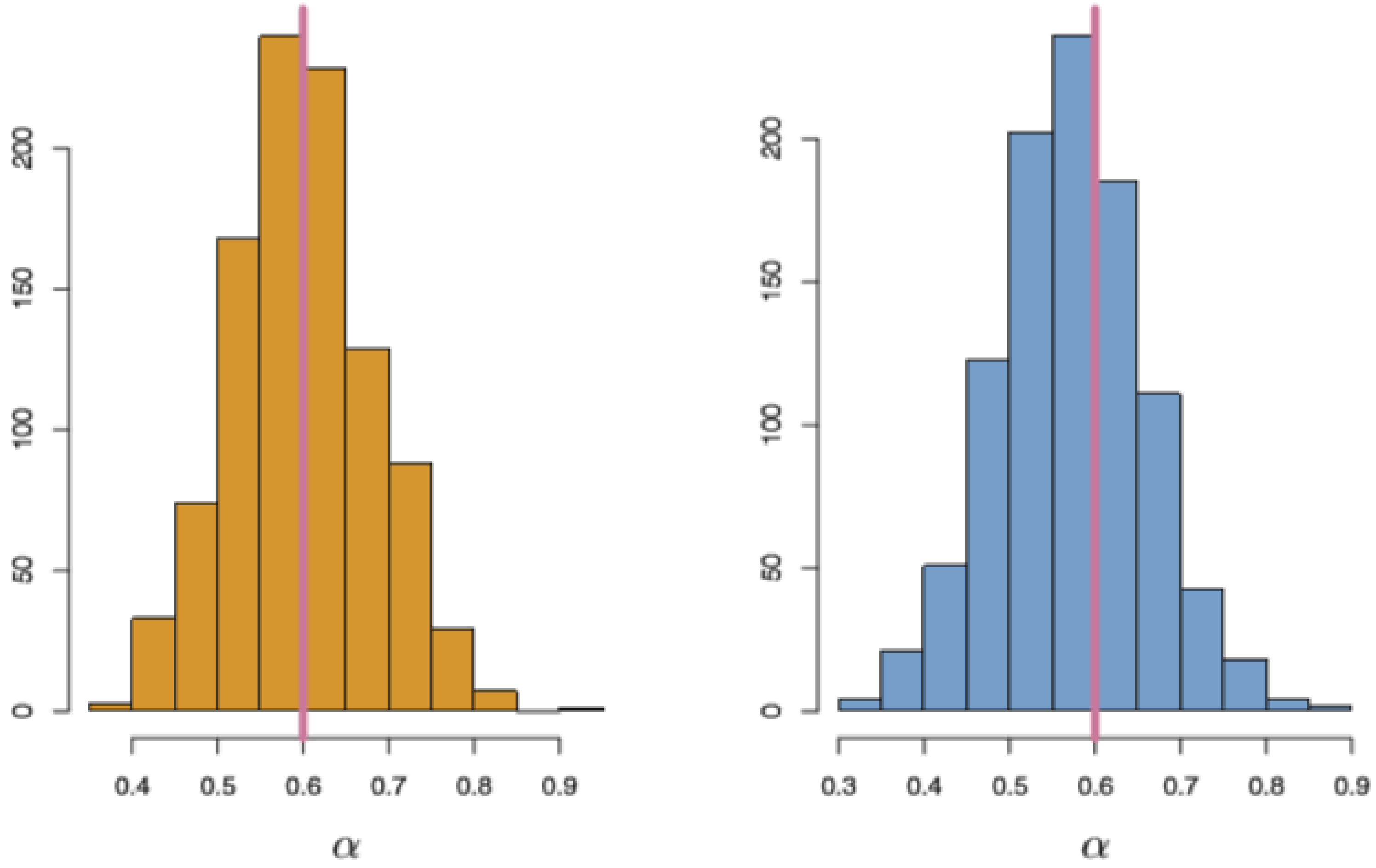
Obs	X	Y
2	2.1	1.1
2	2.1	1.1
1	4.3	2.4

Bootstrap – test error 추정이 가능? X



Original Sample = Training Sample = Validation Sample

Bootstrap – Variability with parameter

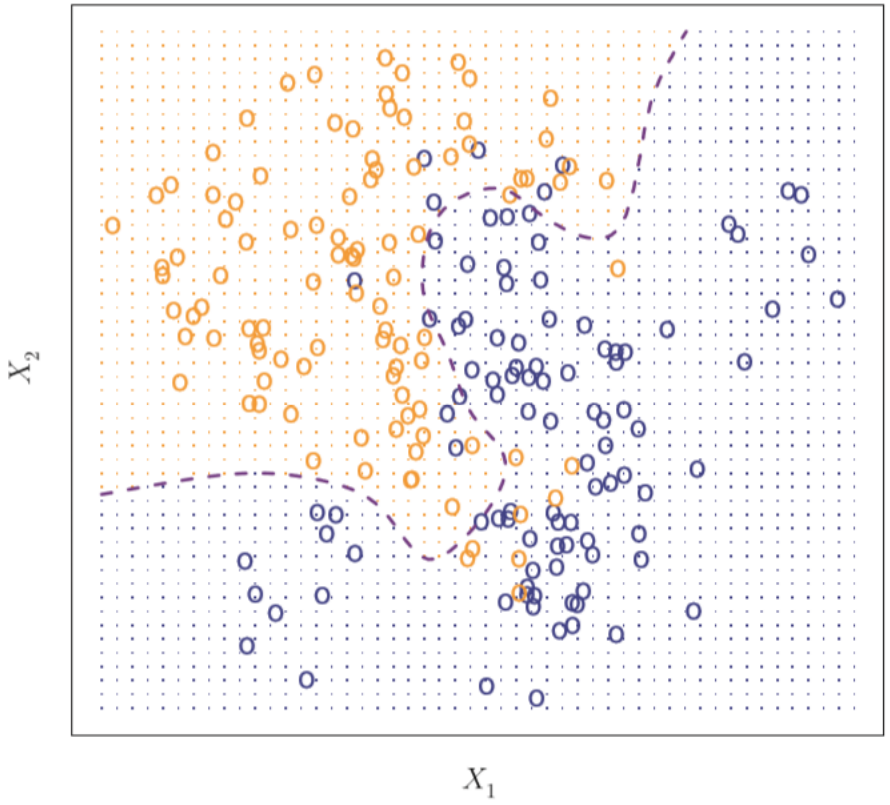


SW

Bayes VS. Bayes Naïve 링크:
https://www.quora.com/What-is-the-difference-between-a-Bayes-classifier-and-a-naive-Bayes-one

지난 주 이슈 1 – Bayes Classifier

$$P(Y=k\mid X=x)=\frac{f_k(x)\times\pi_k}{\sum_{l=1}^{K}\pi_l f_l(x)}$$

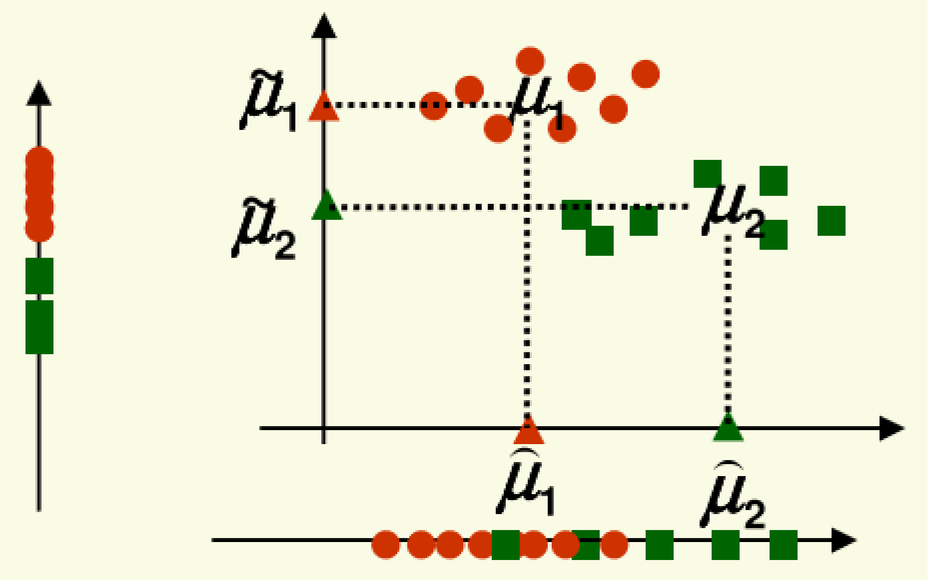


Naive Bayes Classifier

$$f_k(x)=P(X=x\mid Y=k)$$

$$P(X=x\mid Y=k)\approx\prod_{i=1}^{m}P(X=x_i\mid Y=k)$$

지난 주 이슈 2 – LDA for dimension reduction

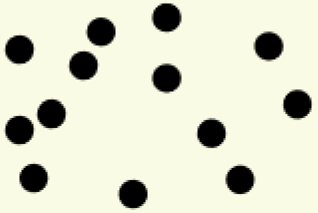


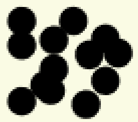
Define their ***scatter*** as

$$s = \sum_{i=1}^{n} (z_i - \mu_z)^2$$

samples $z_1, \ldots, z_n$

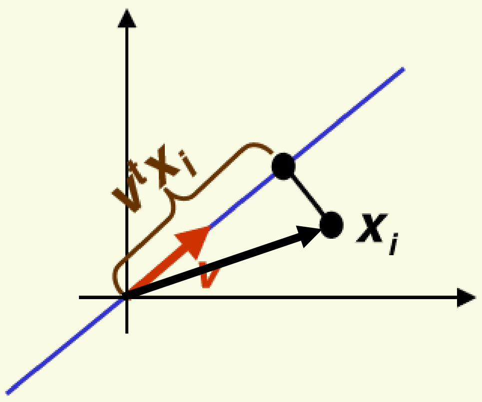
Sample mean is $\mu_z = \frac{1}{n} \sum_{i=1}^{n} z_i$

larger scatter: 

smaller scatter: 

SW

지난 주 이슈 2 – LDA for dimension reduction

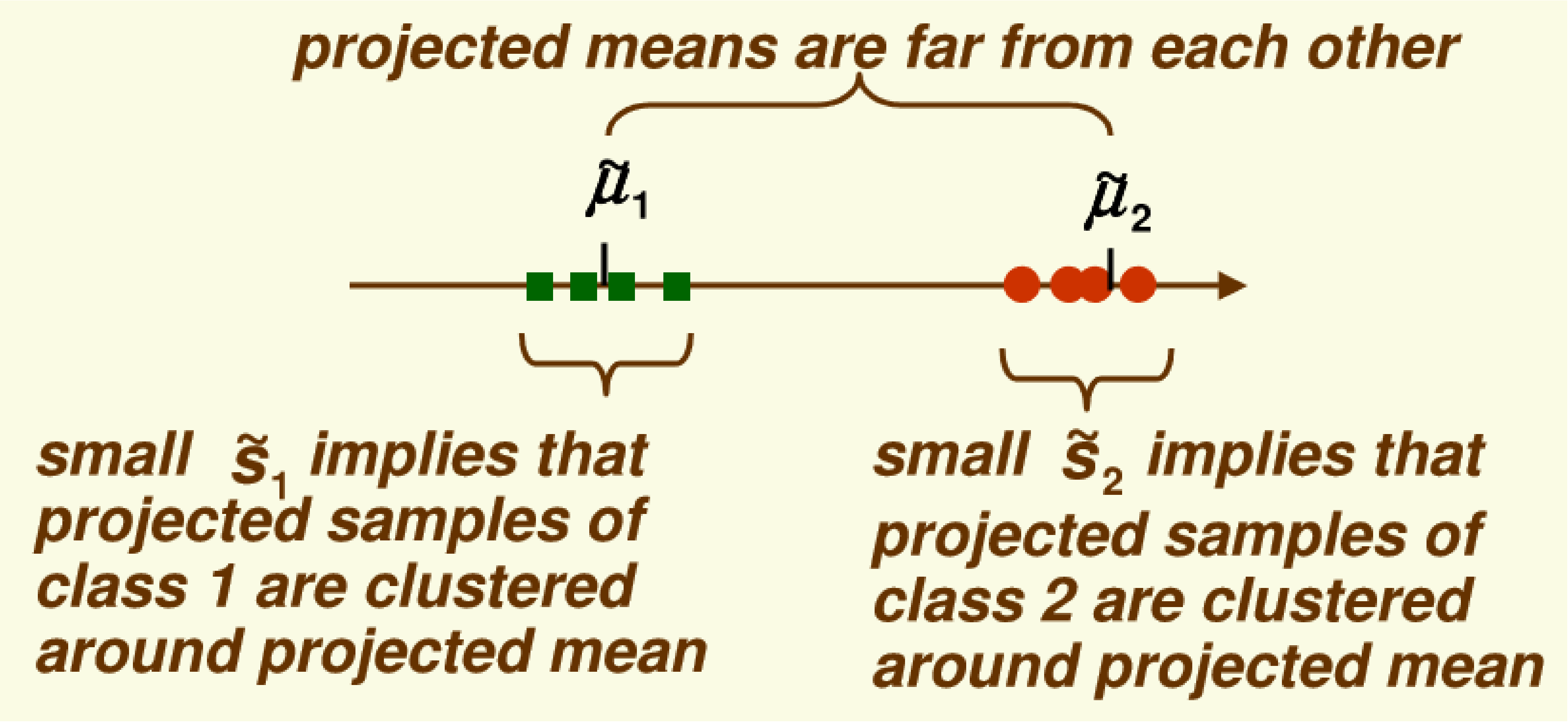


$$J(v) = \frac{\overbrace{(\tilde{\mu}_1 - \tilde{\mu}_2)^2}^{\text{want projected means are far from each other}}}{\tilde{s}_1^2 + \tilde{s}_2^2}$$

want scatter in class 1 is as small as possible, i.e. samples of class 1 cluster around the projected mean $\tilde{\mu}_1$

want scatter in class 2 is as small as possible, i.e. samples of class 2 cluster around the projected mean $\tilde{\mu}_2$

지난 주 이슈 2 – LDA for dimension reduction



지난 주 이슈 2 – LDA for dimension reduction

$$J(v) = \frac{(\tilde{\mu}_1 - \tilde{\mu}_2)^2}{\tilde{s}_1^2 + \tilde{s}_2^2} = \frac{v^t S_B v}{v^t S_W v}$$

J(v)를 v에 대해 미분 = 0 이 후, 수식 정리했을 때,

아래와 같이 고유 값 문제가 된다.!

$$\Rightarrow \underbrace{S_B v = \lambda S_W v}_{\textit{generalized eigenvalue problem}}$$

Thank you!