Prediction models

Seminar Data Science for Economics

Madina Kurmangaliyeva

m.kurmangaliyeva@uvt.nl

Spring 2021

Tilburg University

In general, all prediction tasks have the same basic steps:

1. Train a **less flexible model** on a **training sample**
2. Train a **more flexible model** on the training sample
3. Compare MSE of model 1 and 2 on the **validation sample** and choose the one with the smallest MSE
4. Calculate MSE of the chosen model on the **test sample**. $\sqrt{\text{MSE}}$ is the expected spread of the prediction errors of your best prediction model.

Having a validation set and test set?

Why don’t we use the MSE of the chosen model in the validation set?

Why do we need a test set?

Validation vs test MSE

Model 1

Model 2

Validation vs test MSE

Predictive power

Model 1 = **Model 2**

$\text{MSE} \sim N(5,1)$ $\text{MSE} \sim N(5,1)$

Validation vs test MSE

Predictive power

	Model 1	=	**Model 2**
	MSE ~ N(5,1)		MSE ~ N(5,1)
Validation 1	5.26		5.05

Validation vs test MSE

Predictive power

Model 1 = **Model 2**

$\text{MSE} \sim N(5,1)$ $\text{MSE} \sim N(5,1)$

Validation 1	5.26	5.05
Validation 2	5.90	6.56
Validation 3	4.63	4.86

Validation vs test MSE

Predictive power

Model 1 = **Model 2**

MSE ~ N(5,1) MSE ~ N(5,1)

	Model 1	Model 2
Validation 1	5.26	$\boxed{5.05}$
Validation 2	5.90	6.56
Validation 3	4.63	4.86

Validation vs test MSE

Predictive power

Model 1 = **Model 2**

	MSE ~ N(5,1)	MSE ~ N(5,1)
Validation 1	5.26	$\boxed{5.05}$
Validation 2	$\boxed{5.90}$	6.56
Validation 3	4.63	4.86

Validation vs test MSE

Predictive power

Model 1 = **Model 2**

	MSE ~ N(5,1)	MSE ~ N(5,1)
Validation 1	5.26	$\boxed{5.05}$
Validation 2	$\boxed{5.90}$	6.56
Validation 3	$\boxed{4.63}$	4.86

Validation vs test MSE

Predictive power

	Model 1	=	**Model 2**	**Best model's Validation MSE**
	MSE ~ N(5,1)		MSE ~ N(5,1)	
Validation 1	5.26		5.05	5.05
Validation 2	5.90		6.56	5.90
Validation 3	4.63		4.86	4.63

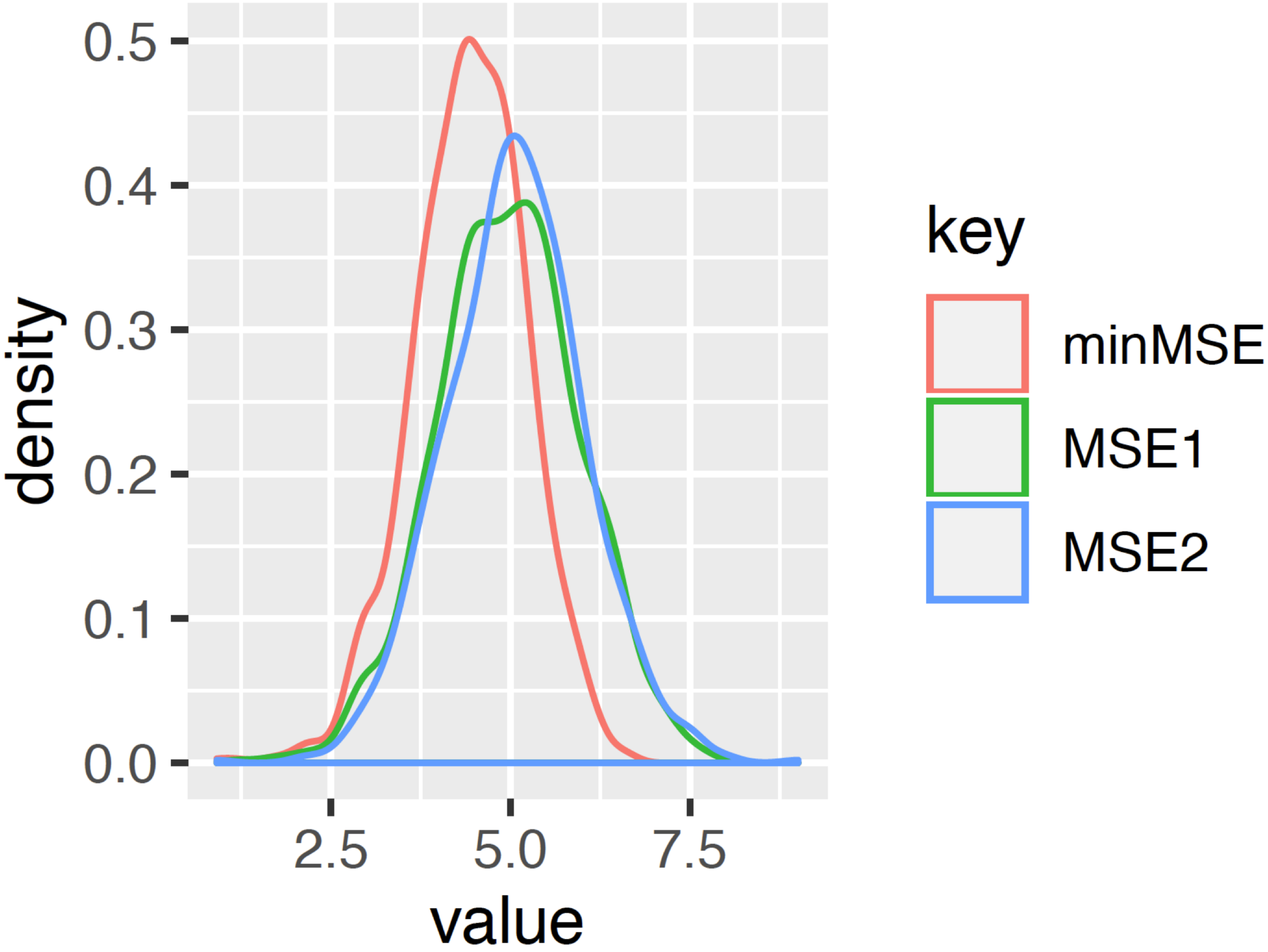
Validation vs test MSE

Predictive power

Model 1 = Model 2

	Model 1 MSE ~ N(5,1)	**Model 2** MSE ~ N(5,1)	**Best model's Validation MSE** MSE ~ min(N(5,1), N(5,1))
Validation 1	5.26	$\boxed{5.05}$	5.05
Validation 2	$\boxed{5.90}$	6.56	5.90
Validation 3	$\boxed{4.63}$	4.86	4.63

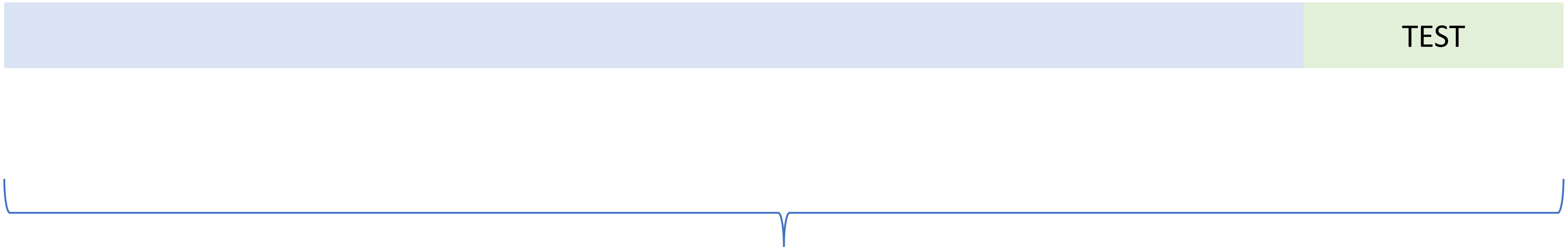
$E(\text{MSE}|\min \text{MSE}) \neq E(\text{MSE})$



https://madina-k.github.io/dse_mk2021

Hence, we need a yet untouched sample (test sample) to estimate the unbiased out-of-sample MSE

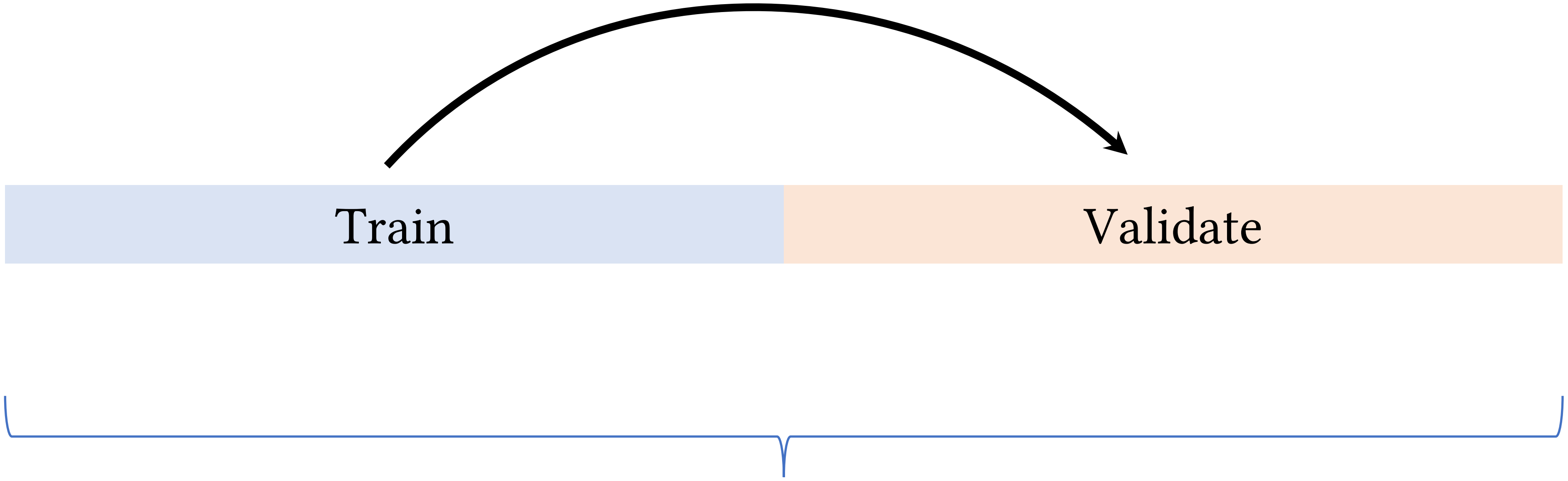
Isn’t it wasteful to split data in 3 equal parts (training, validation, and test)?



TEST

Your data

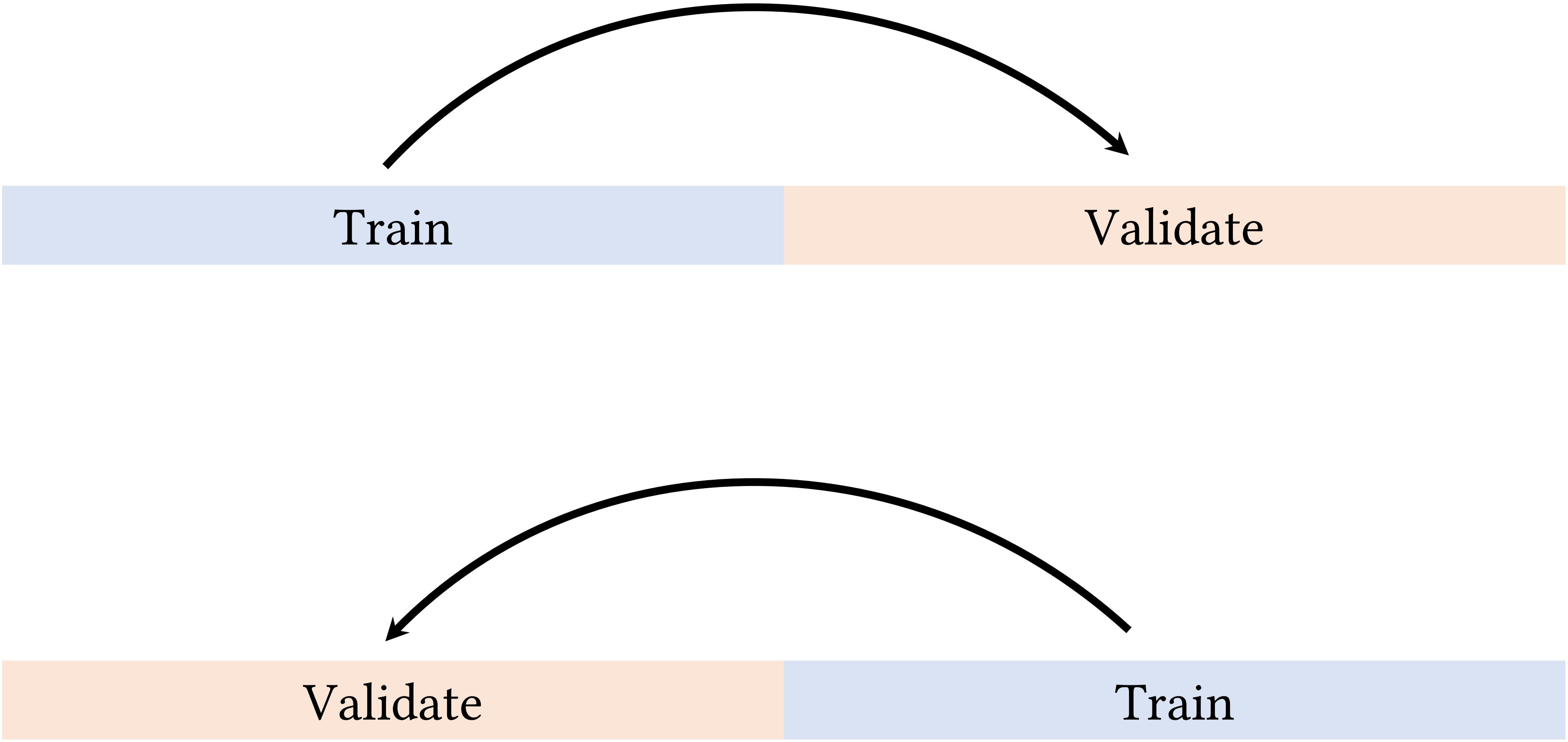
Your training + validation data

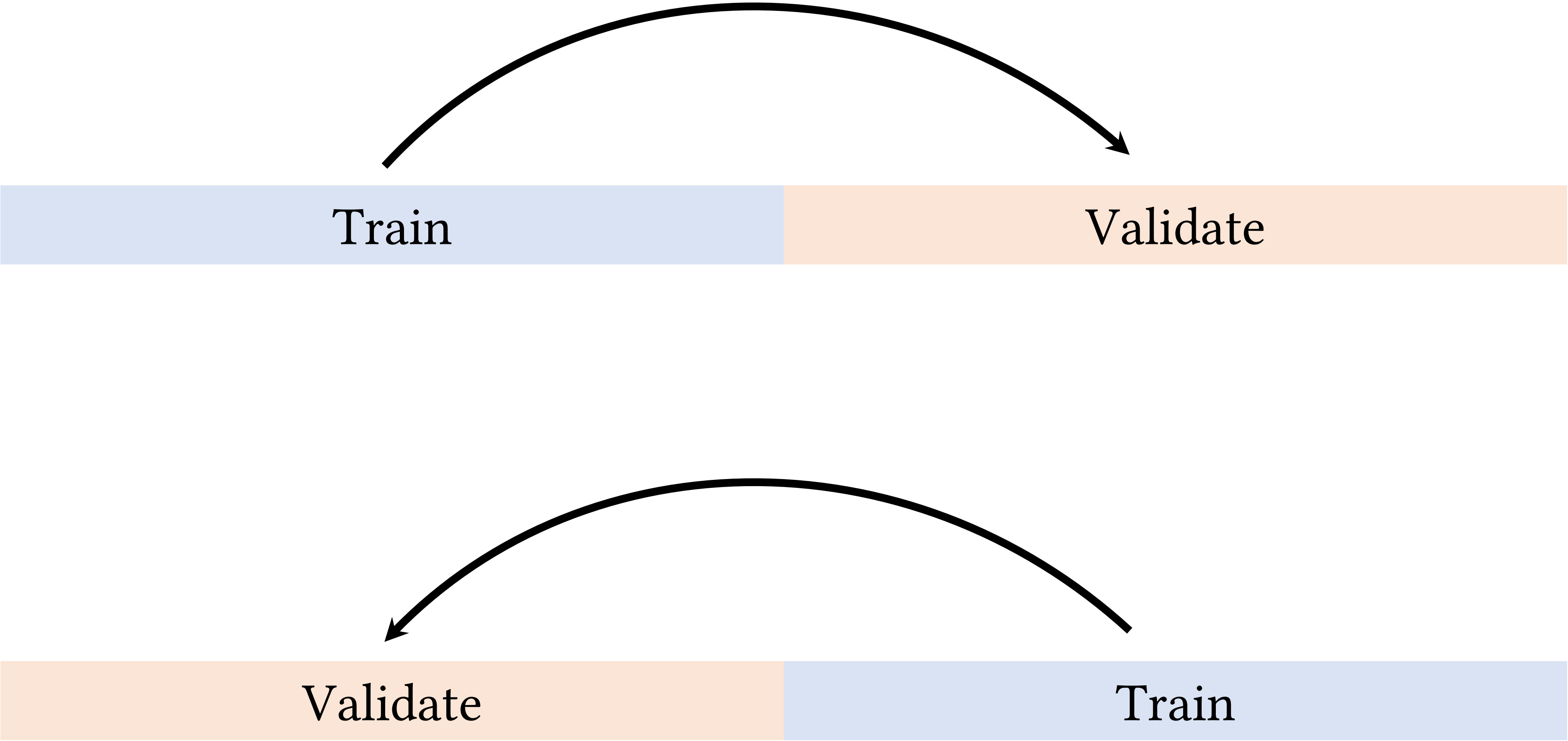


Train

Validate

Your training + validation data


Train
Validate
Validate
Train



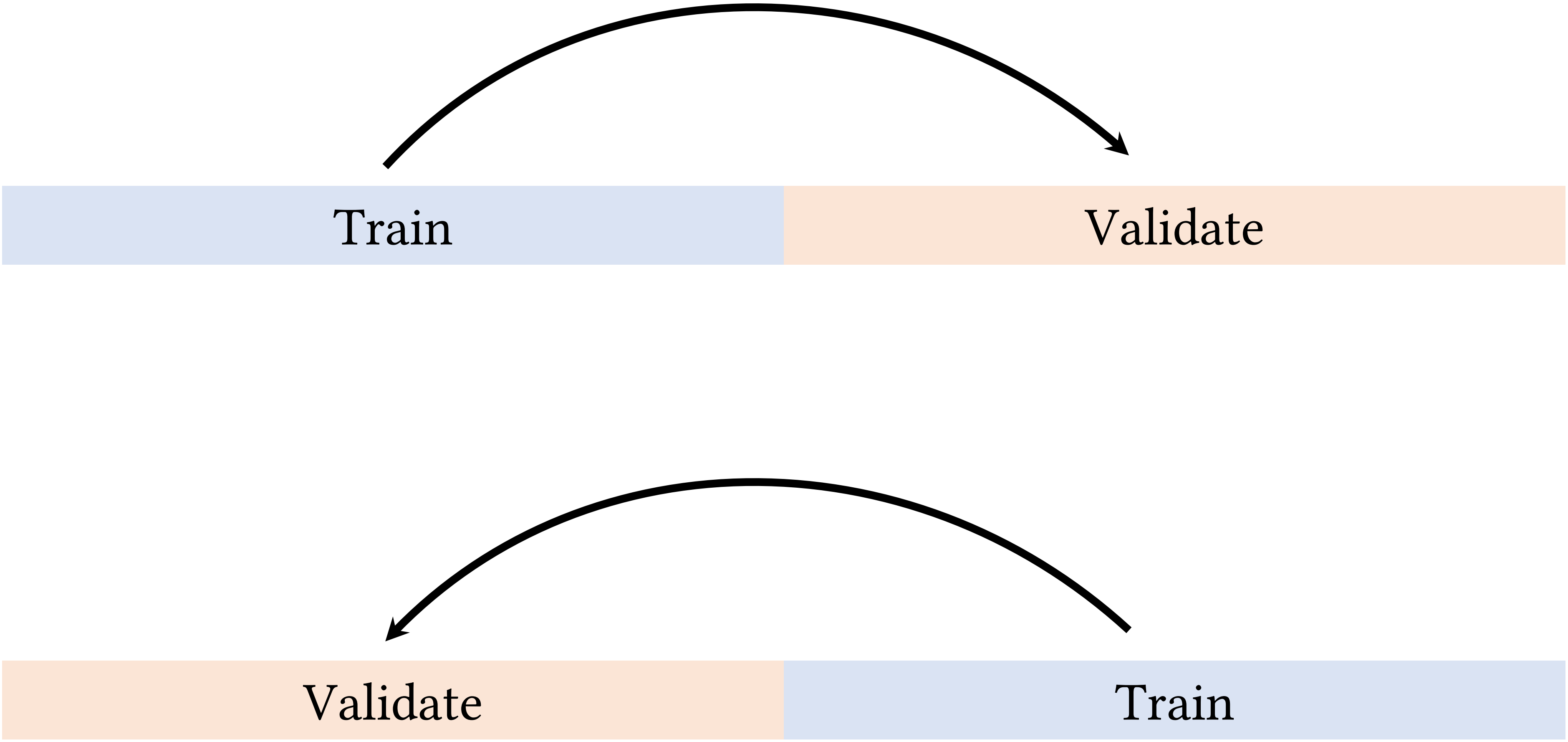
Train

Validate

Validate

Train

$$\text{CV\_MSE(model } i \text{)} = \frac{1}{2}\left(MSE_1^i + MSE_2^i\right)$$



Train

Validate

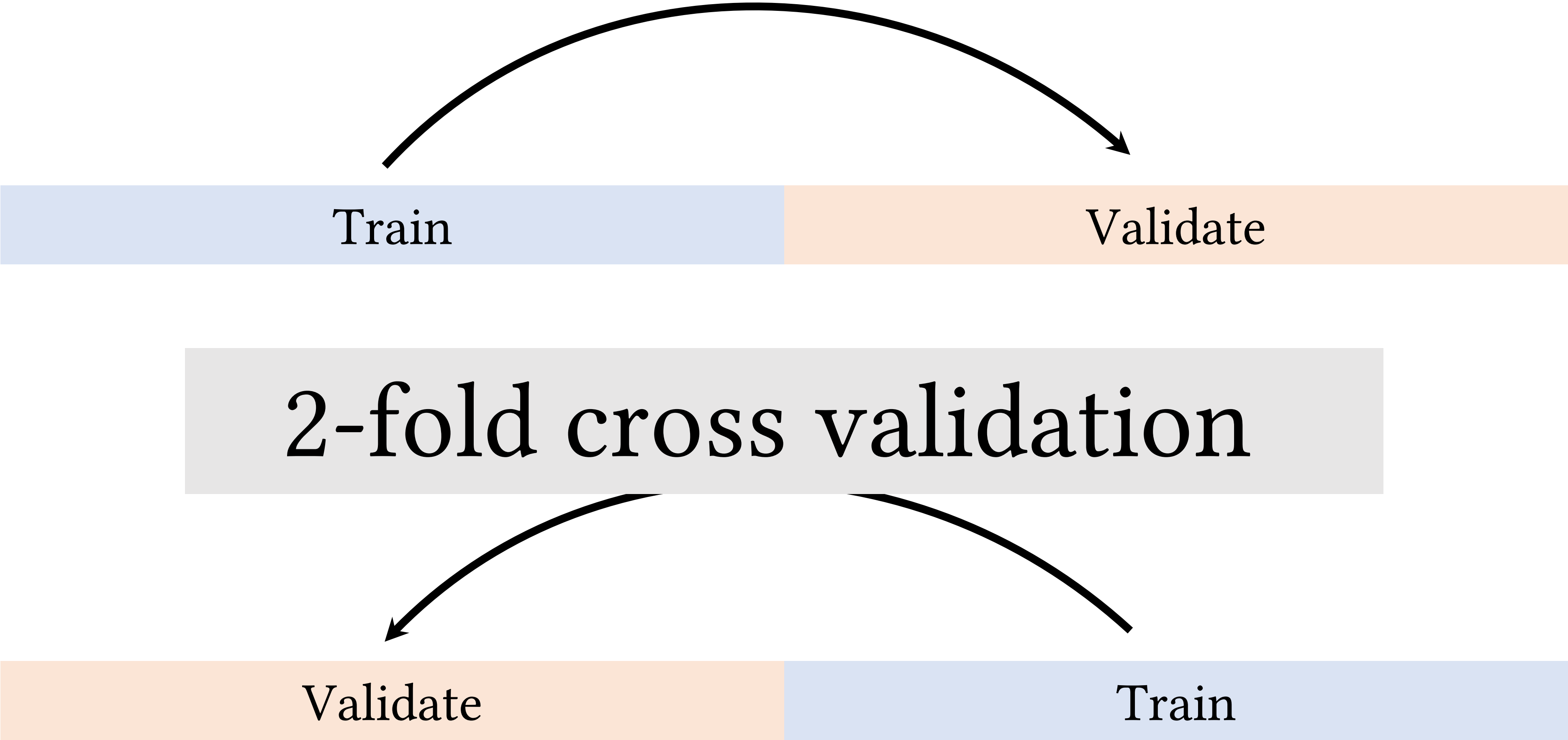
Validate

Train

$$\text{CV\_MSE(model } i \text{)} = \frac{1}{2}\left(MSE_1^i + MSE_2^i\right)$$

Choose model that has the lowest CV_MSE

2-fold cross validation



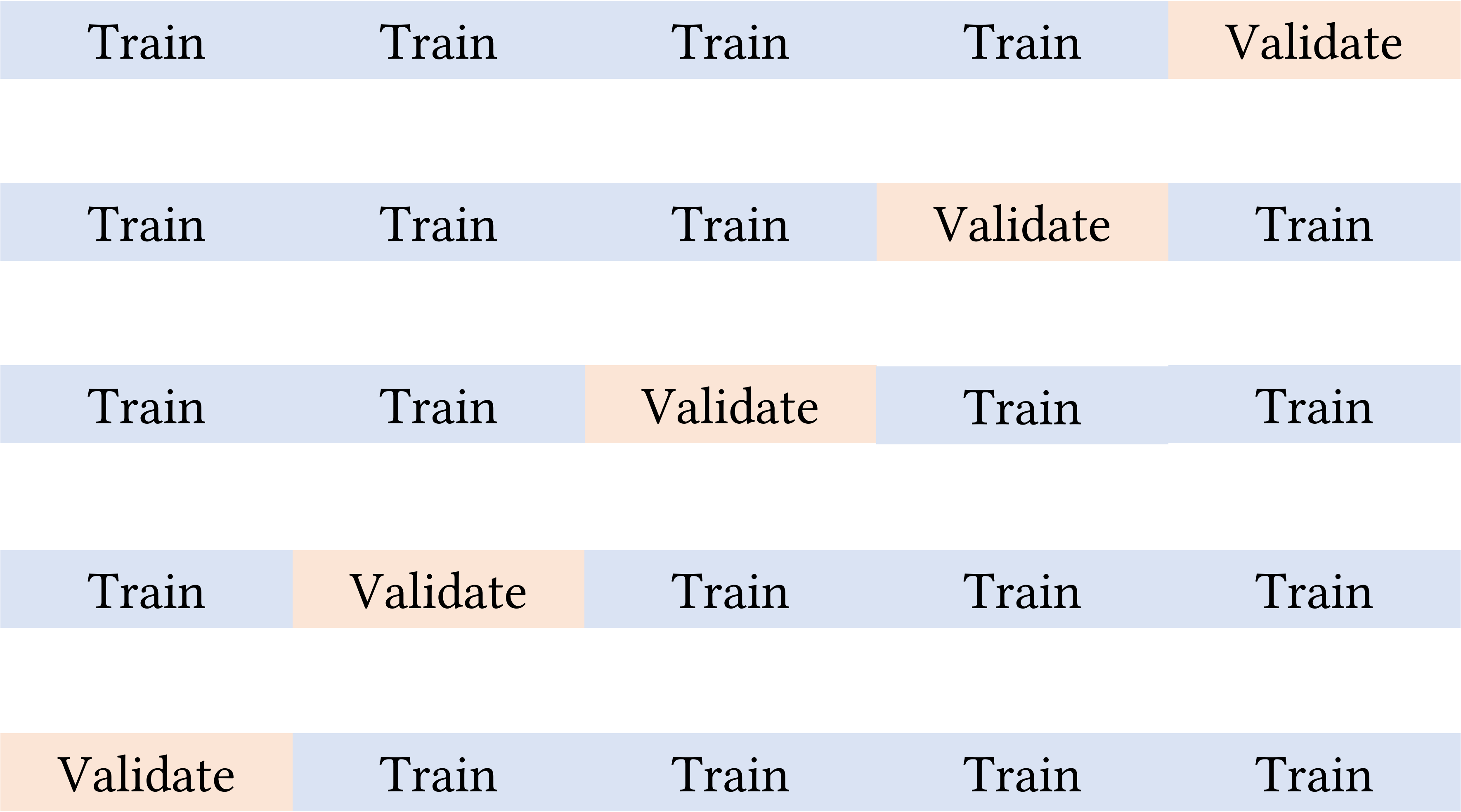
$$\text{CV\_MSE(model } i \text{)} = \frac{1}{2}\left(MSE_1^i + MSE_2^i\right)$$

Choose model that has the lowest CV_MSE

Train	Train	Train	Train	Validate

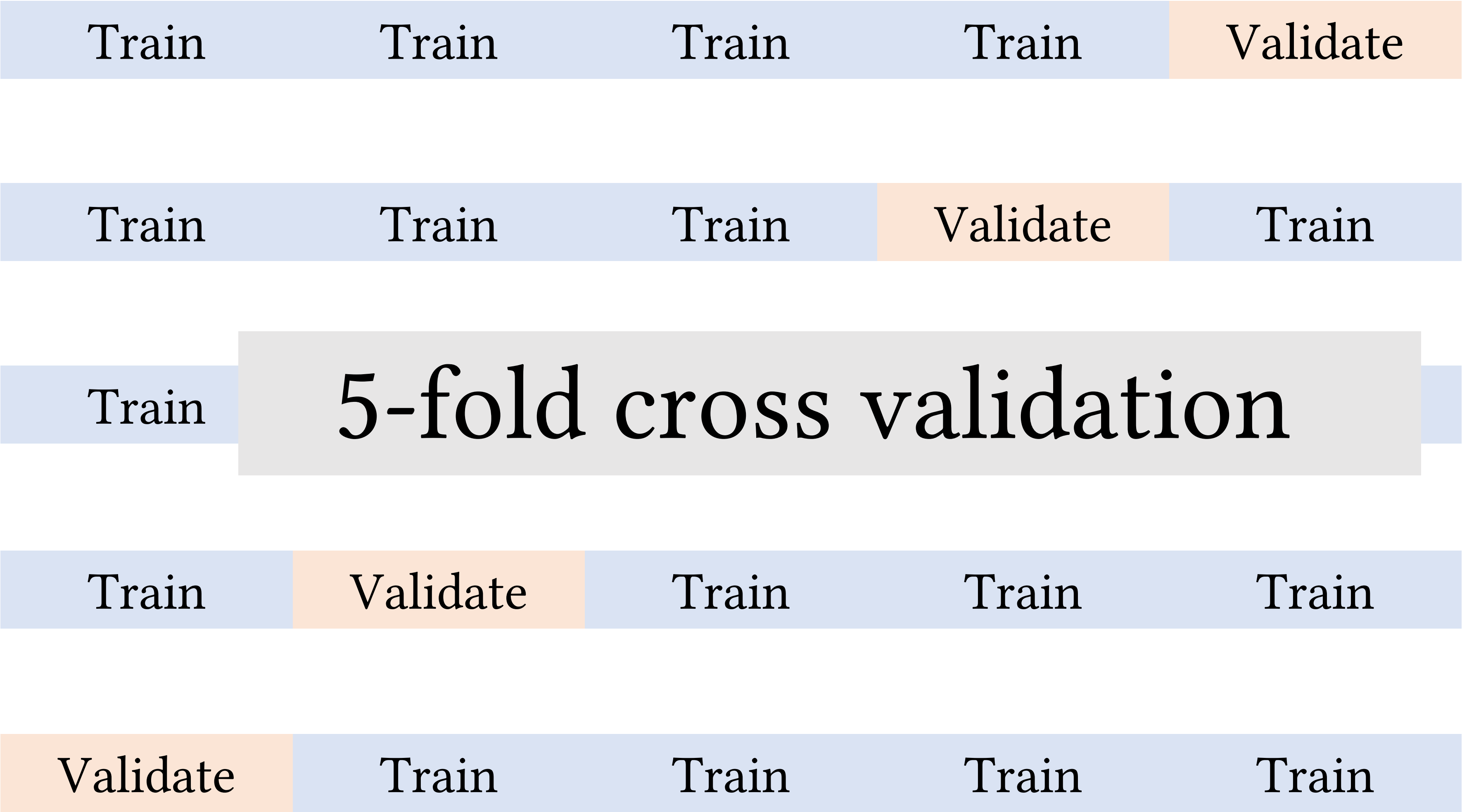
Train	Train	Train	Train	Validate
Train	Train	Train	Validate	Train

Train	Train	Train	Train	Validate
Train	Train	Train	Validate	Train
Train	Train	Validate	Train	Train
Train	Validate	Train	Train	Train
Validate	Train	Train	Train	Train



Train	Train	Train	Train	Validate
Train	Train	Train	Validate	Train
Train	Train	Validate	Train	Train
Train	Validate	Train	Train	Train
Validate	Train	Train	Train	Train

$$\text{CV\_MSE(model } i \text{)} = \frac{1}{5}\sum_{j=1}^{5} MSE_j^i$$



Train	Train	Train	Train	Validate
Train	Train	Train	Validate	Train
Train				
Train	Validate	Train	Train	Train
Validate	Train	Train	Train	Train

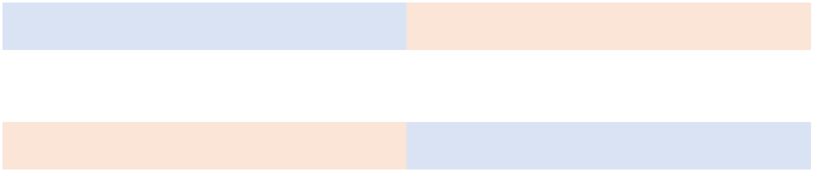
5-fold cross validation

$$\text{CV\_MSE(model } i \text{)} = \frac{1}{5} \sum_{j=1}^{5} MSE_j^i$$

In general, can generalize to a k-fold CV procedure

k = 2

Split in half



k = n-1

Leave-one-out CV

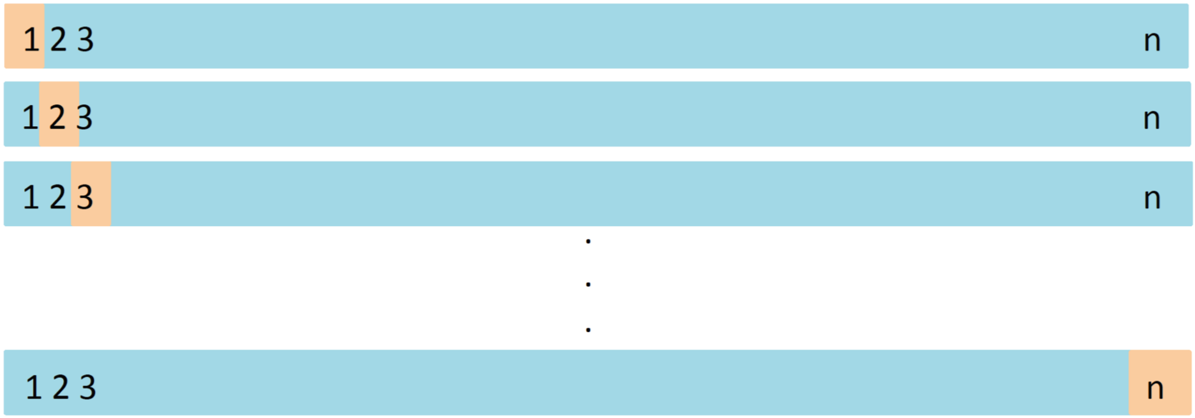
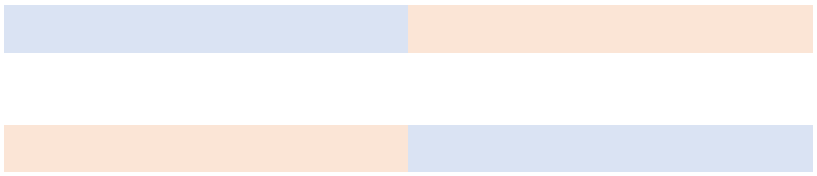


fig 5.3 from ISLR

In general, can generalize to a k-fold CV procedure

k = 2

Split in half



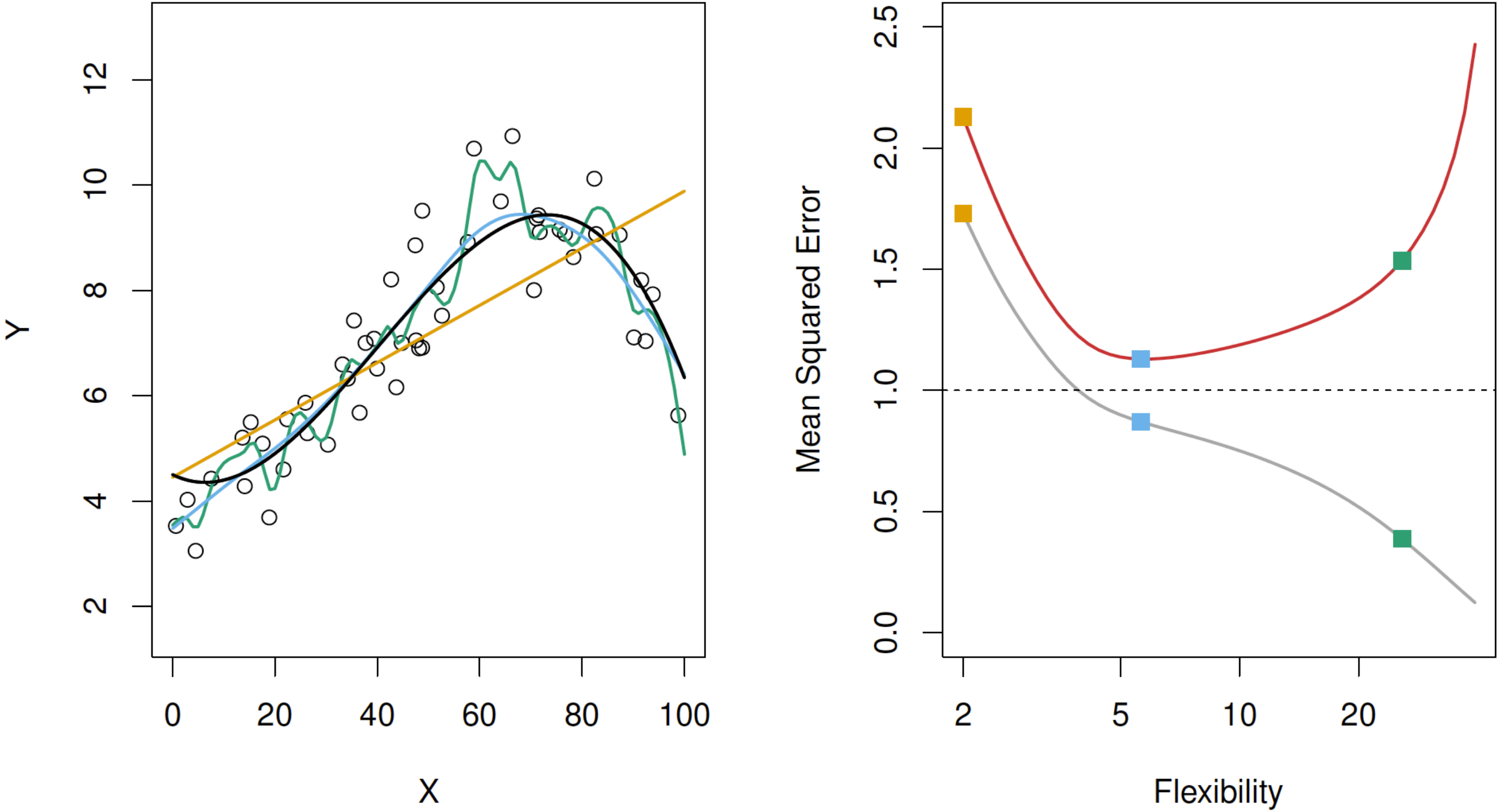
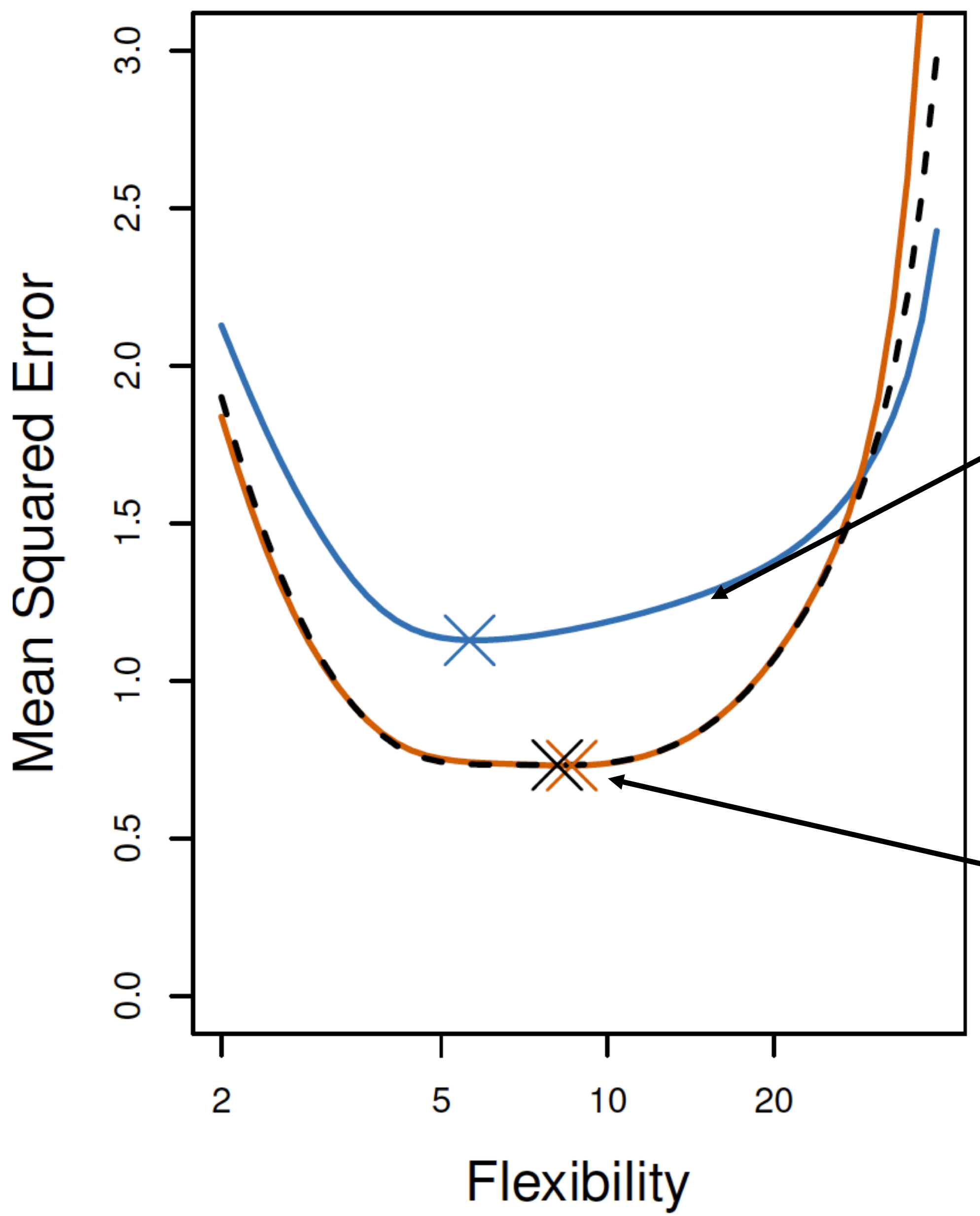


Fig. 2.9 from ISLR



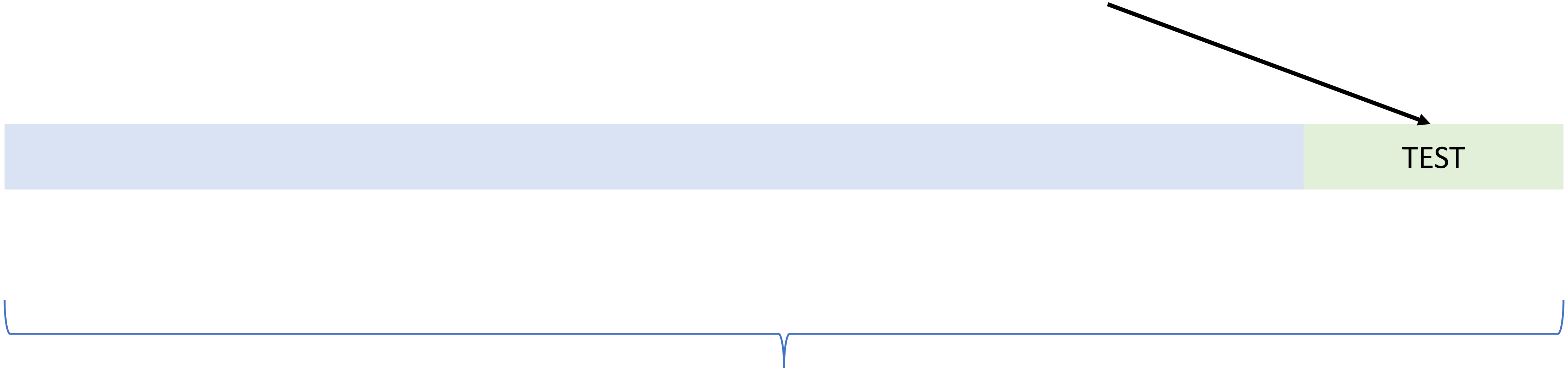
True expected MSE

CV errors:
Orange is k = 10
Dashed is k = n-1

Fig. 5.6 from ISLR.

Bottomline

To have or not to have the test set?



Your data