School of Computing
UNIVERSITY OF GEORGIA

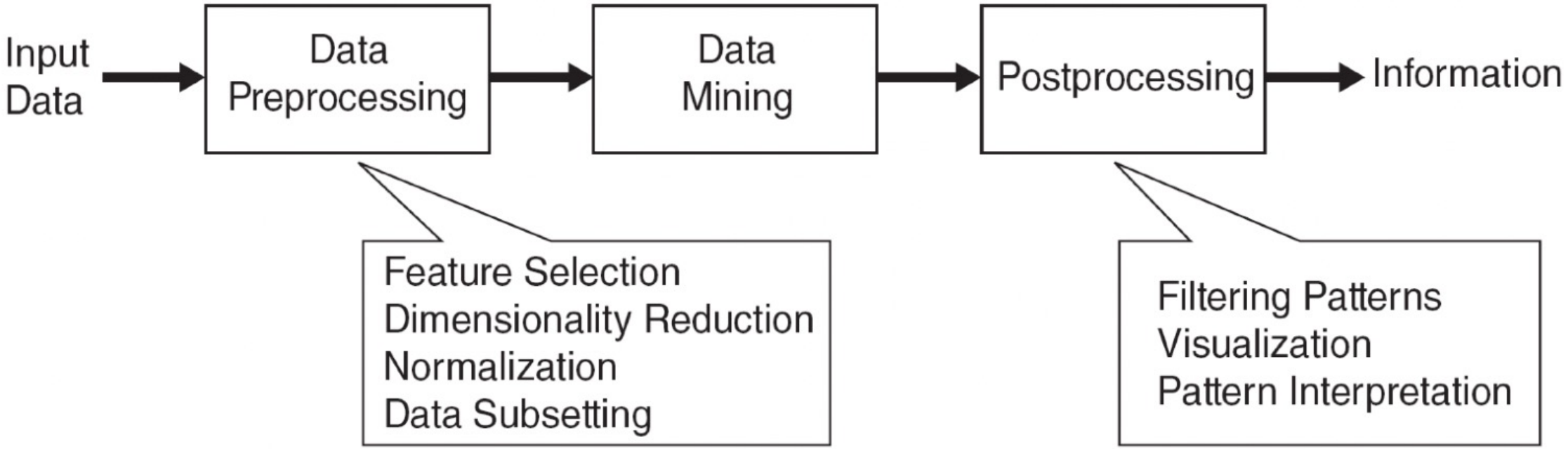
CSCI 4380/6380 DATA MINING

Fei Dou

Assistant Professor
School of Computing
University of Georgia

October 4, 5 2023

Recap: Data Mining Process



Evaluation Measures

Recap: Evaluation Measures

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a	b
	Class=No	c	d

- a: TP (true positive)
- b: FN (false negative)
- c: FP (false positive)
- d: TN (true negative)

Recap: Evaluation Measures

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

Most widely-used metric:

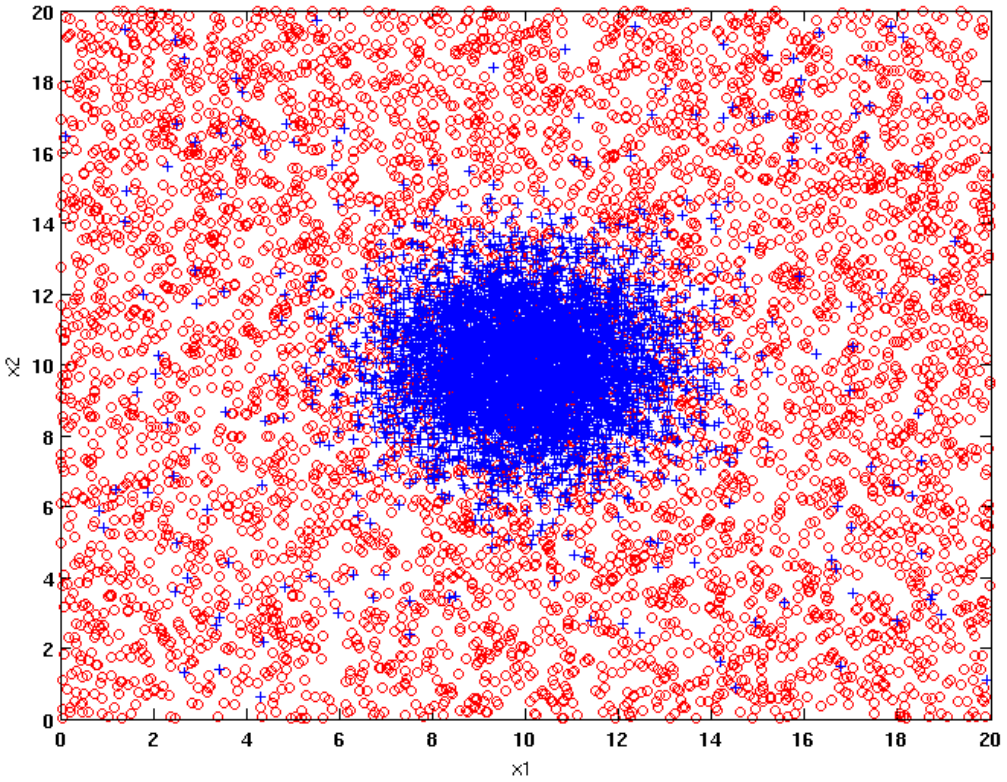
$$Accuracy = \frac{a + d}{a + b + c + d} = \frac{TP + TN}{TP + TN + FP + FN}$$

Recap: Evaluation Measures

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a	b
	Class=No	c	d

- Precision(p) = $\frac{a}{a+c}$
- Recall(r) = $\frac{a}{a+b}$
- F-measure(F) = $\frac{2rp}{r+p} = \frac{2a}{2a+b+c}$

Example Data Set



Two class problem:

+ : 5400 instances

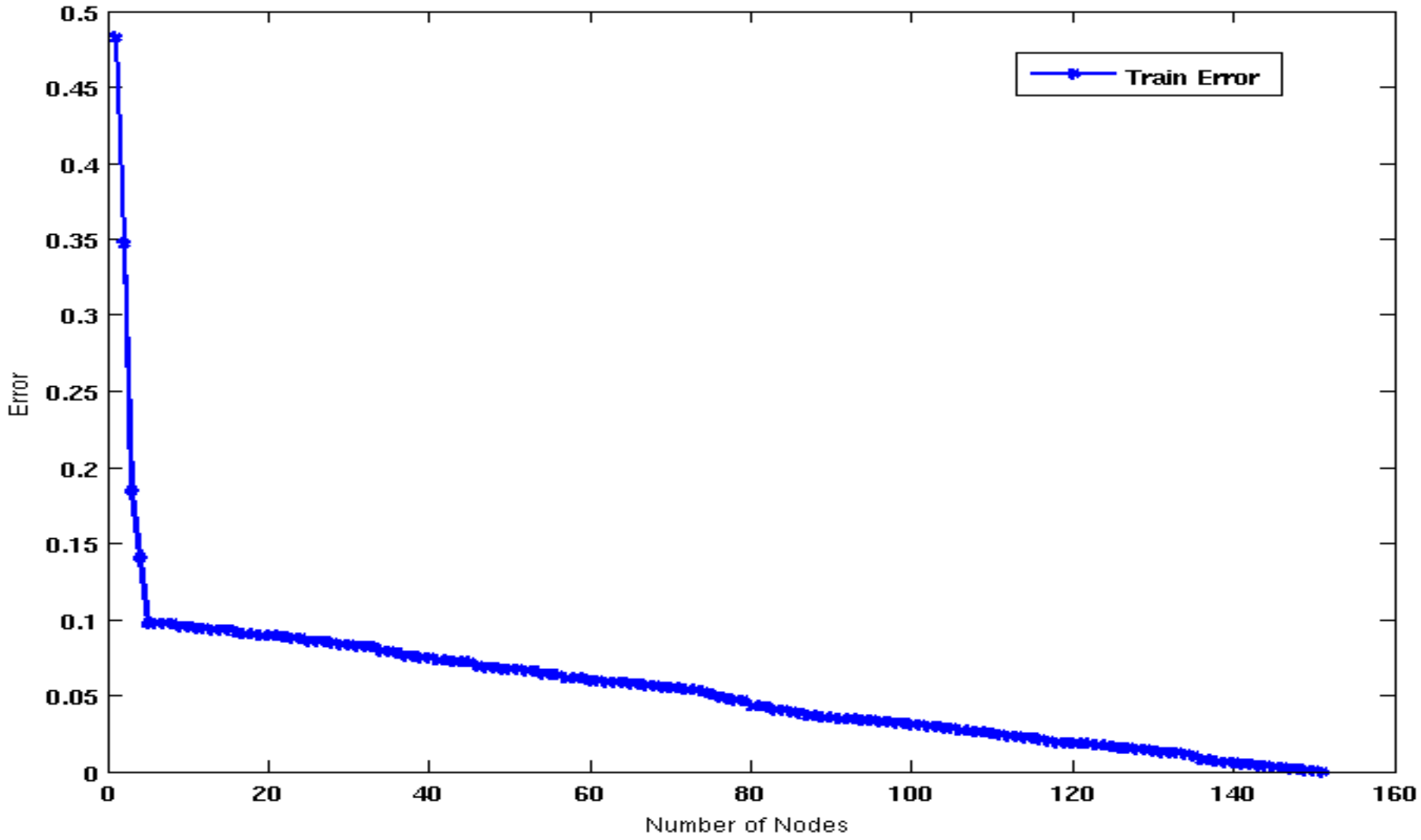
- 5000 instances generated from a Gaussian centered at (10,10)
- 400 noisy instances added

o : 5400 instances

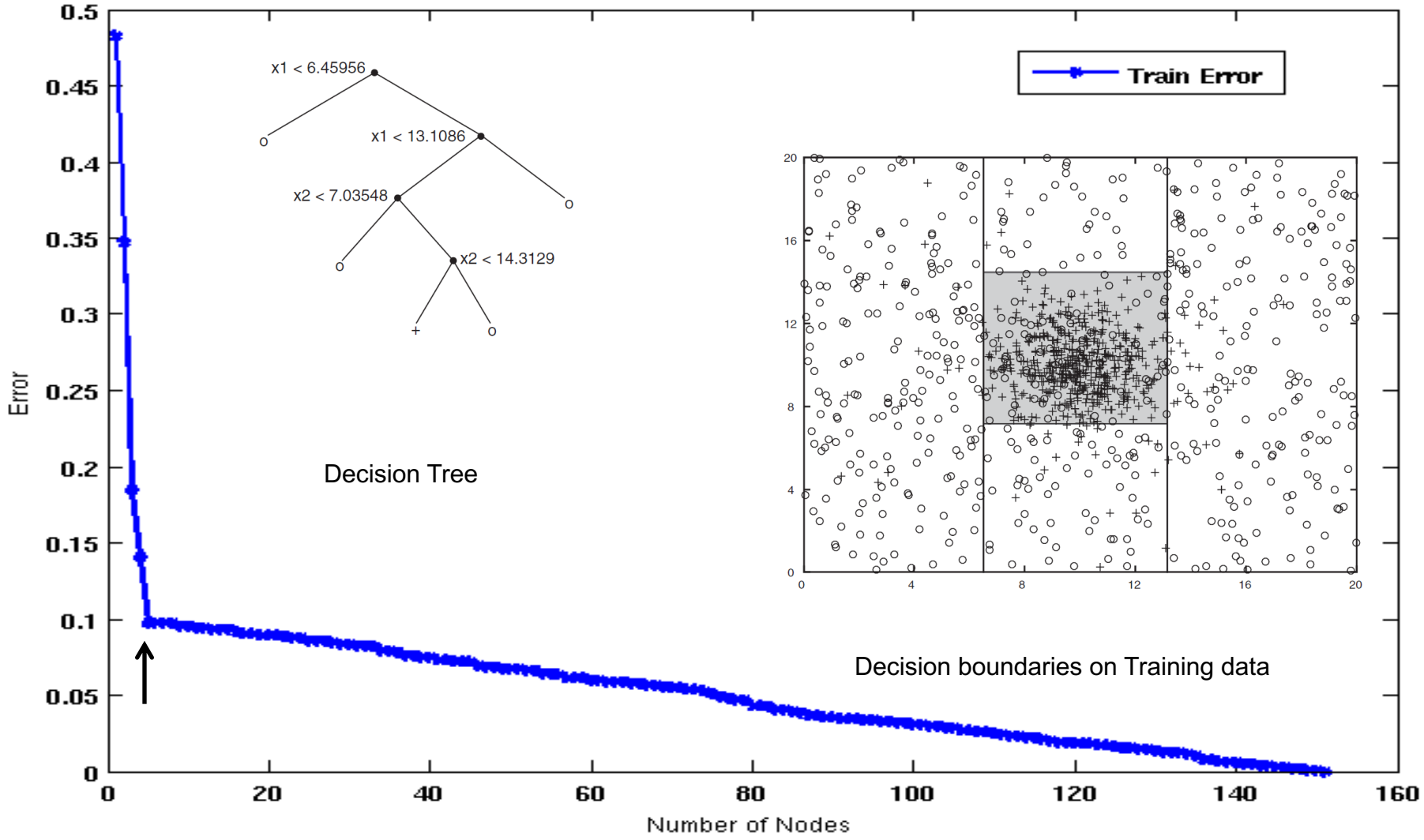
- Generated from a uniform distribution

10 % of the data used for training and 90% of the data used for testing

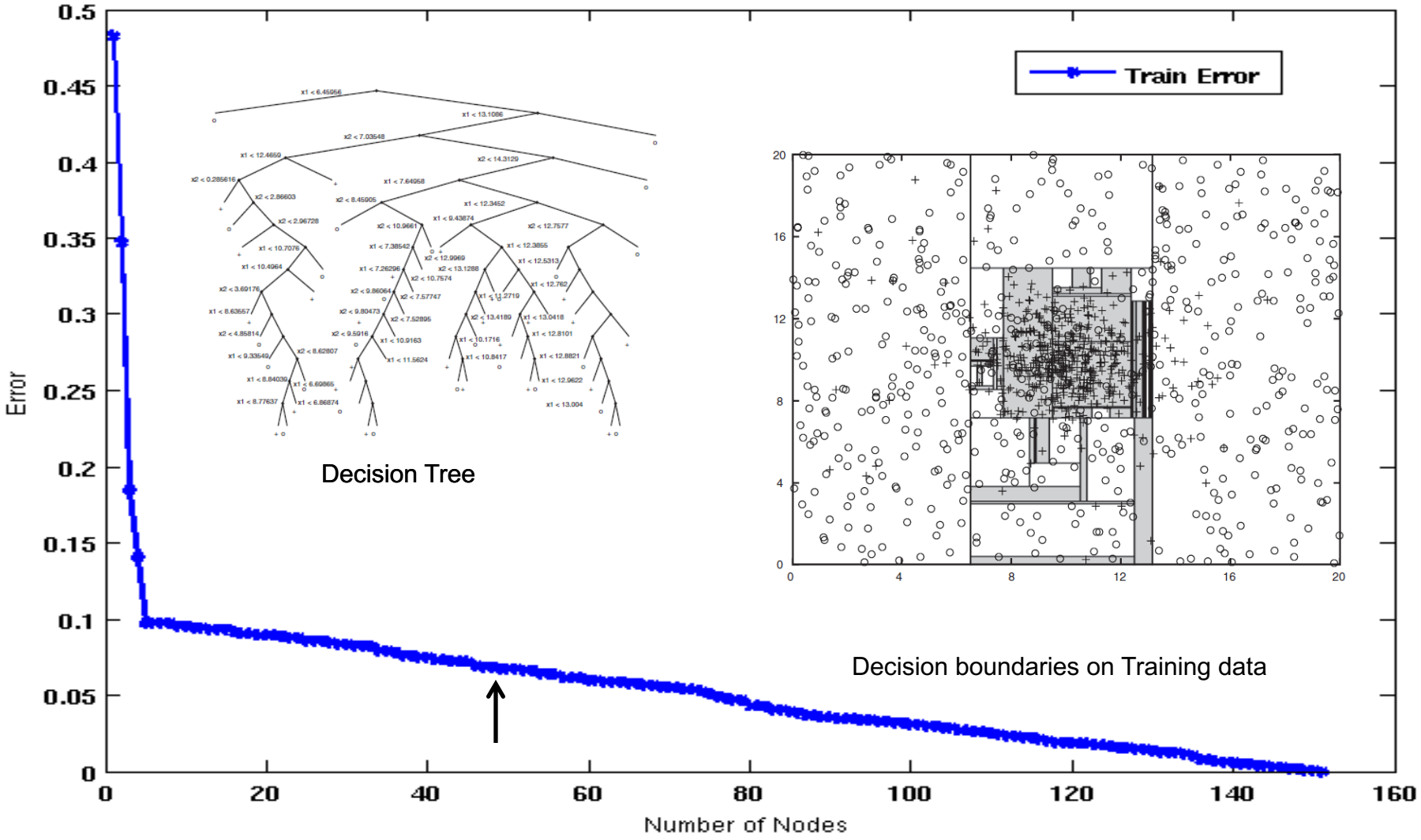
Increasing number of nodes in Decision Trees



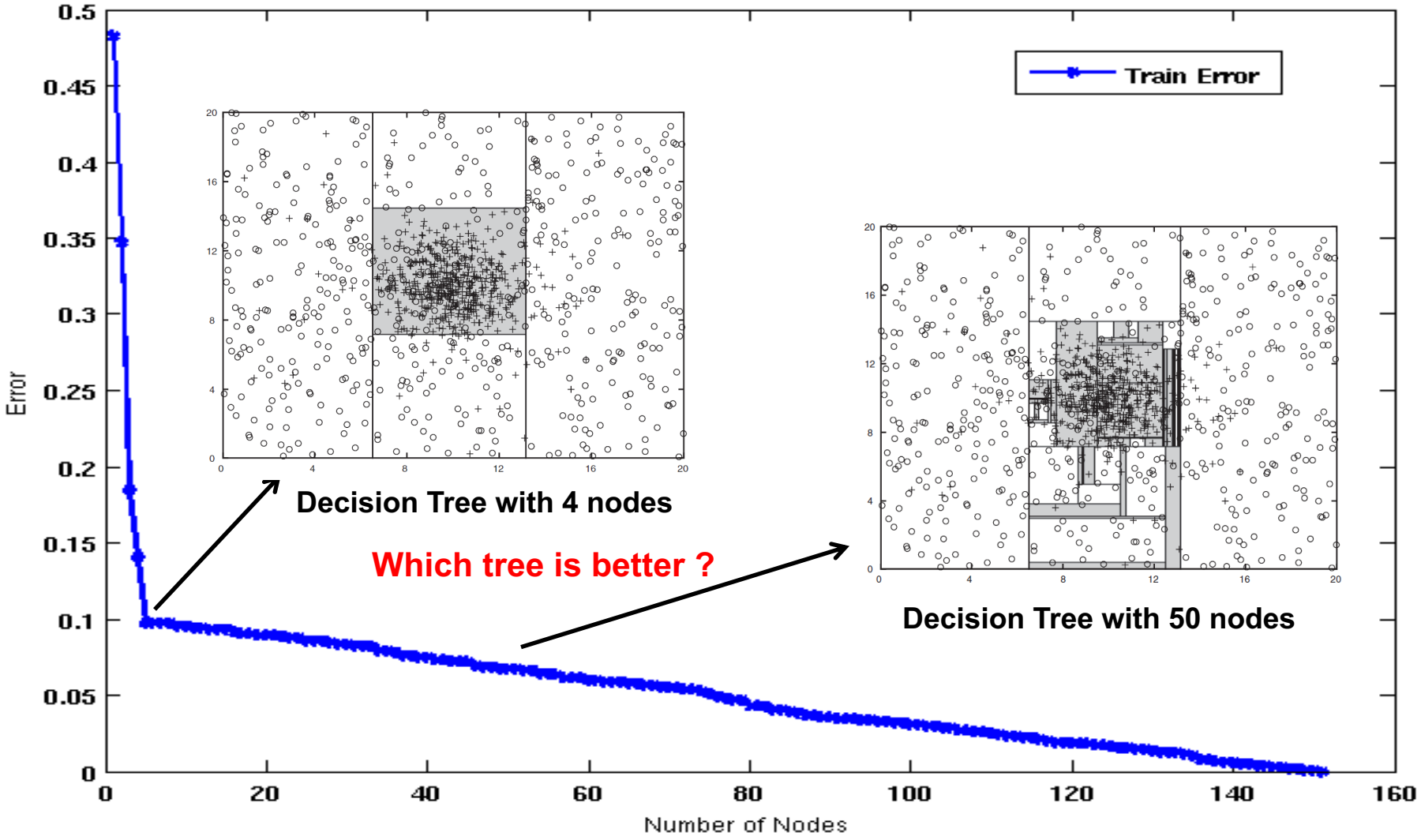
Decision Tree with 4 nodes



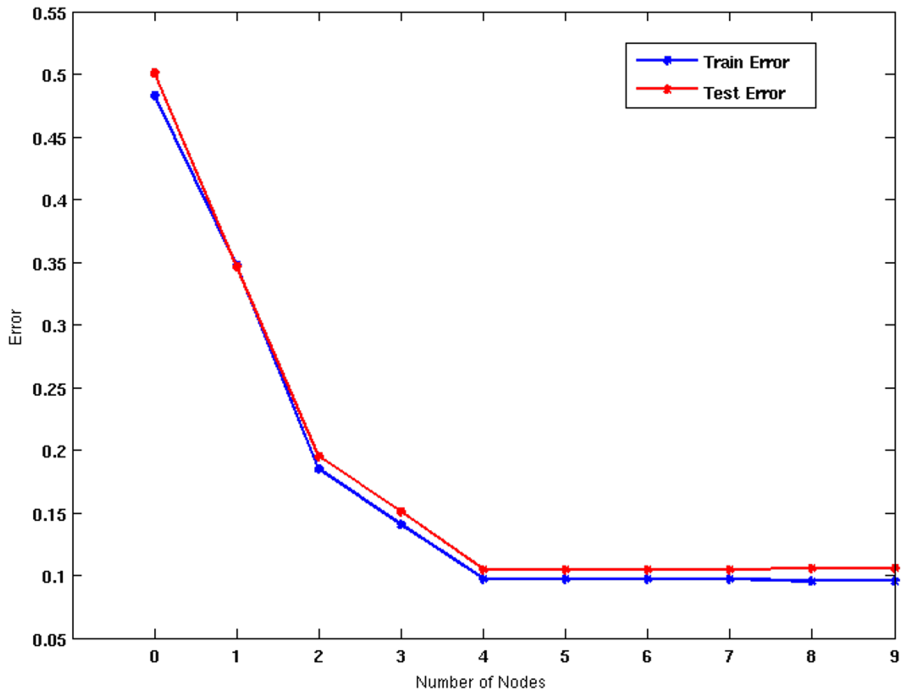
Decision Tree with 50 nodes

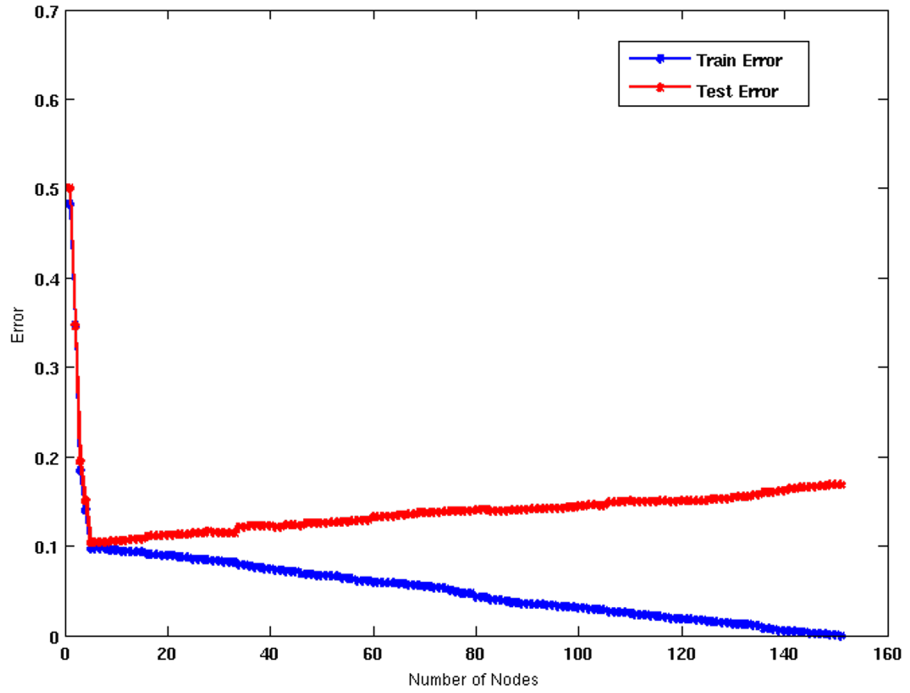


Which tree is better?



Model Underfitting and Overfitting



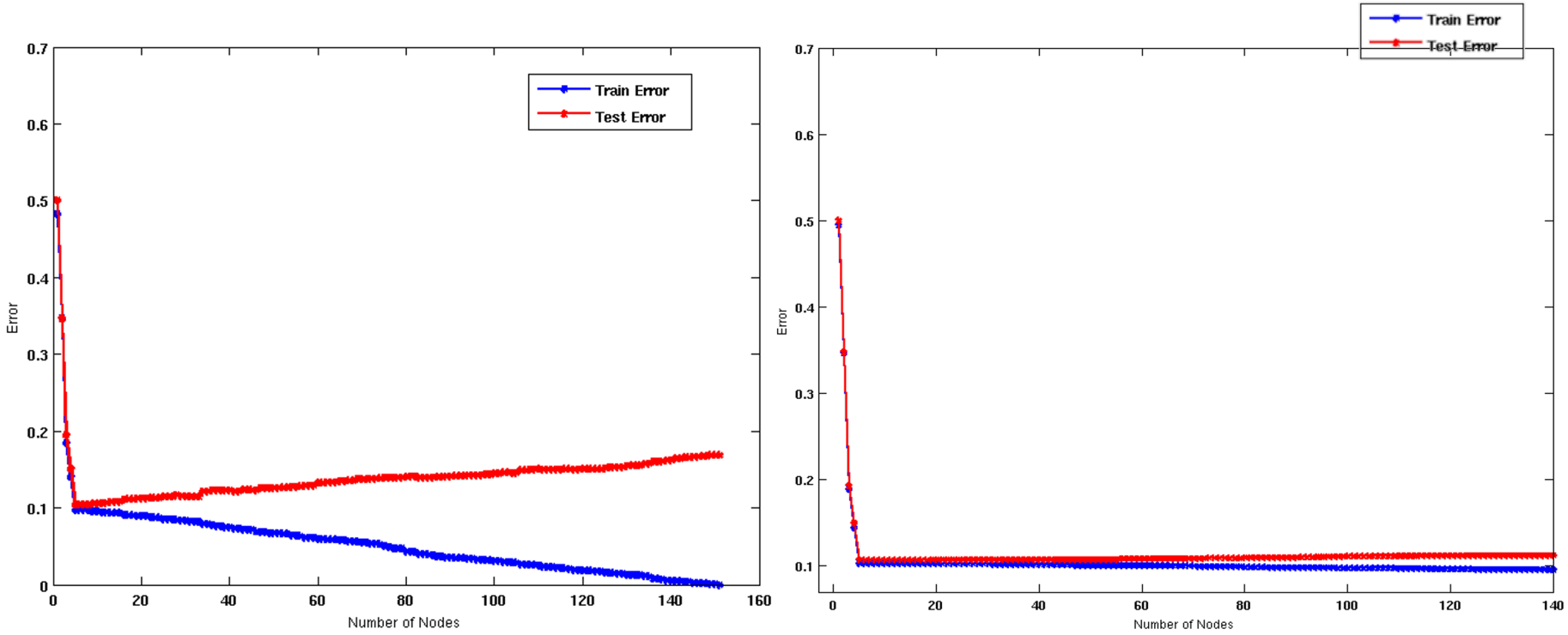


As the model becomes more and more complex, test errors can start increasing even though training error may be decreasing

Underfitting: when model is too simple, both training and test errors are large

Overfitting: when model is too complex, training error is small but test error is large

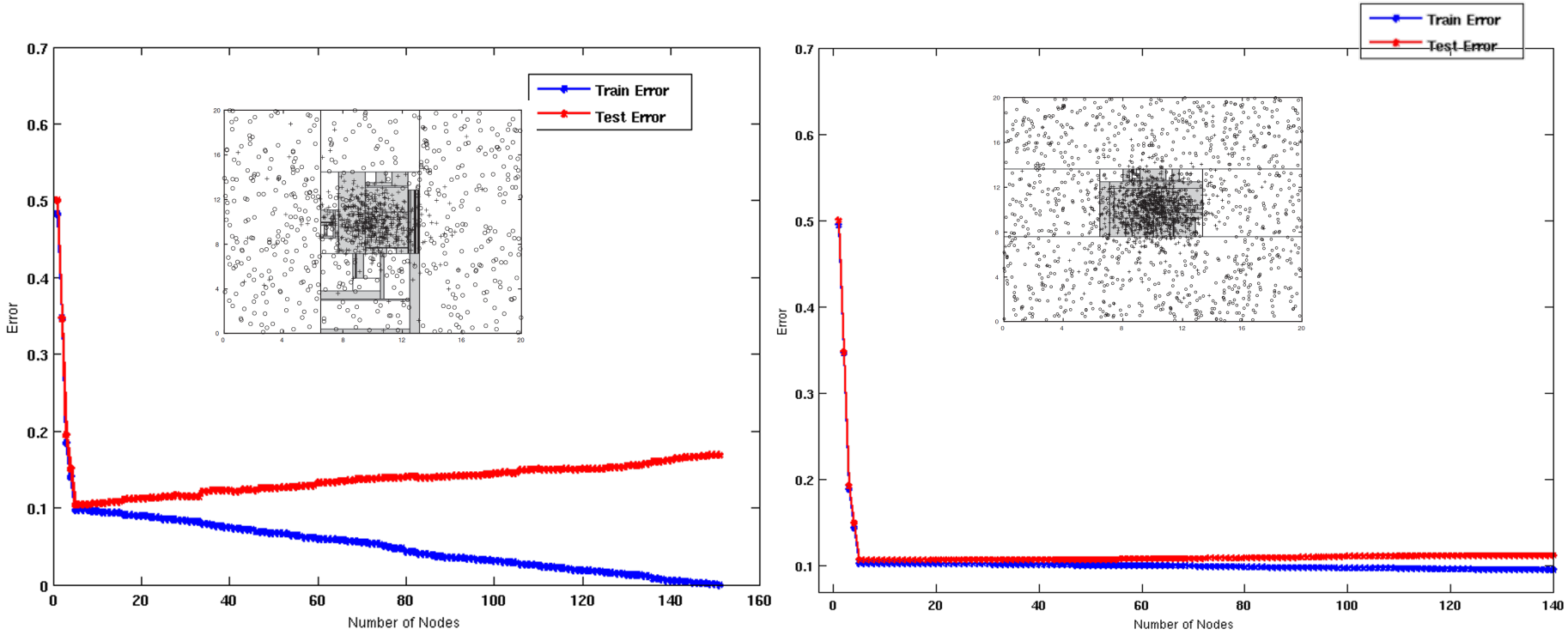
Model Overfitting – Impact of Training Data Size



Using twice the number of data instances

- Increasing the size of training data reduces the difference between training and testing errors at a given size of model

Model Overfitting – Impact of Training Data Size



Using twice the number of data instances

- Increasing the size of training data reduces the difference between training and testing errors at a given size of model

Reasons for Model Overfitting

- Not enough training data
- High model complexity
  - Multiple Comparison Procedure

Overfitting

- Overfitting results in decision trees that are more complex than necessary
- Training error does not provide a good estimate of how well the tree will perform on previously unseen records
- Need ways for estimating generalization errors

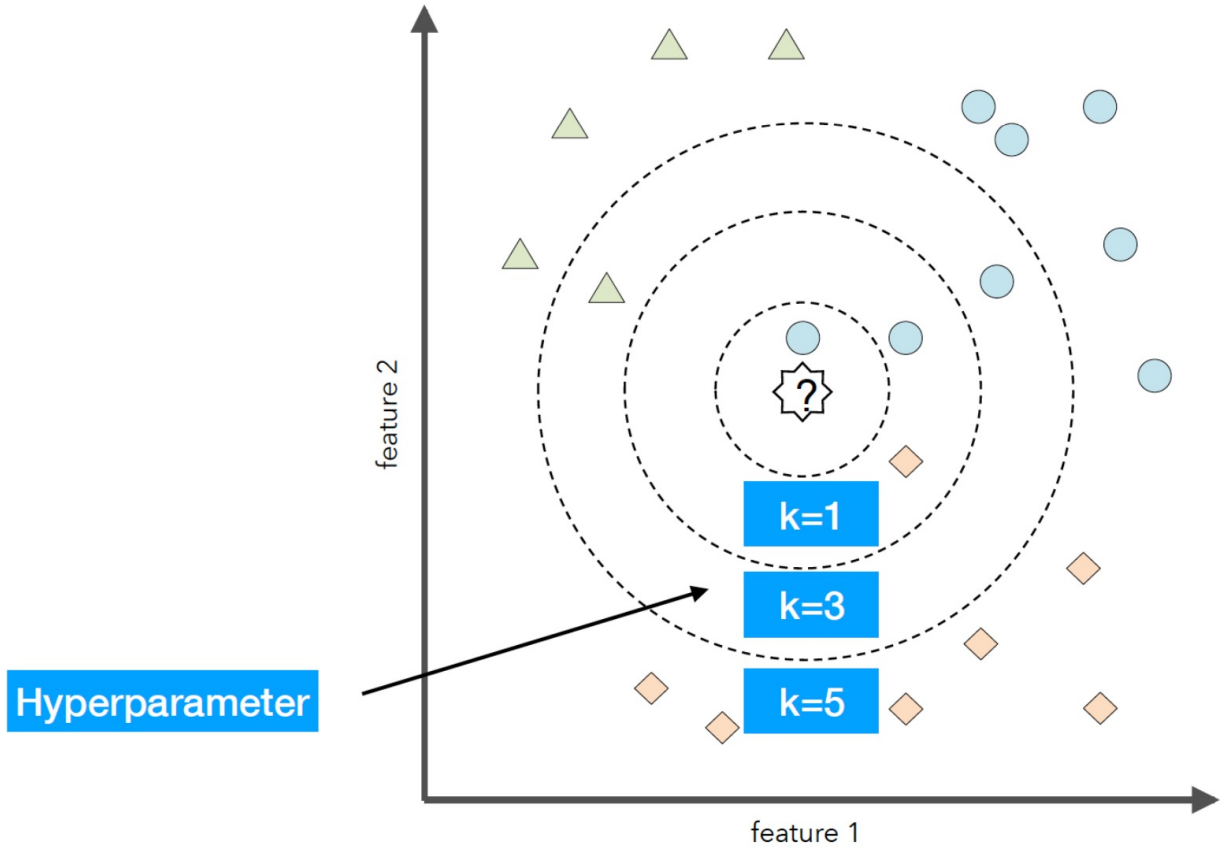
Model Selection

- Performed during model building

- Purpose is to ensure that model is not overly complex (to avoid overfitting)
  - Need to estimate generalization error
  - Using Validation Set Incorporating Model Complexity

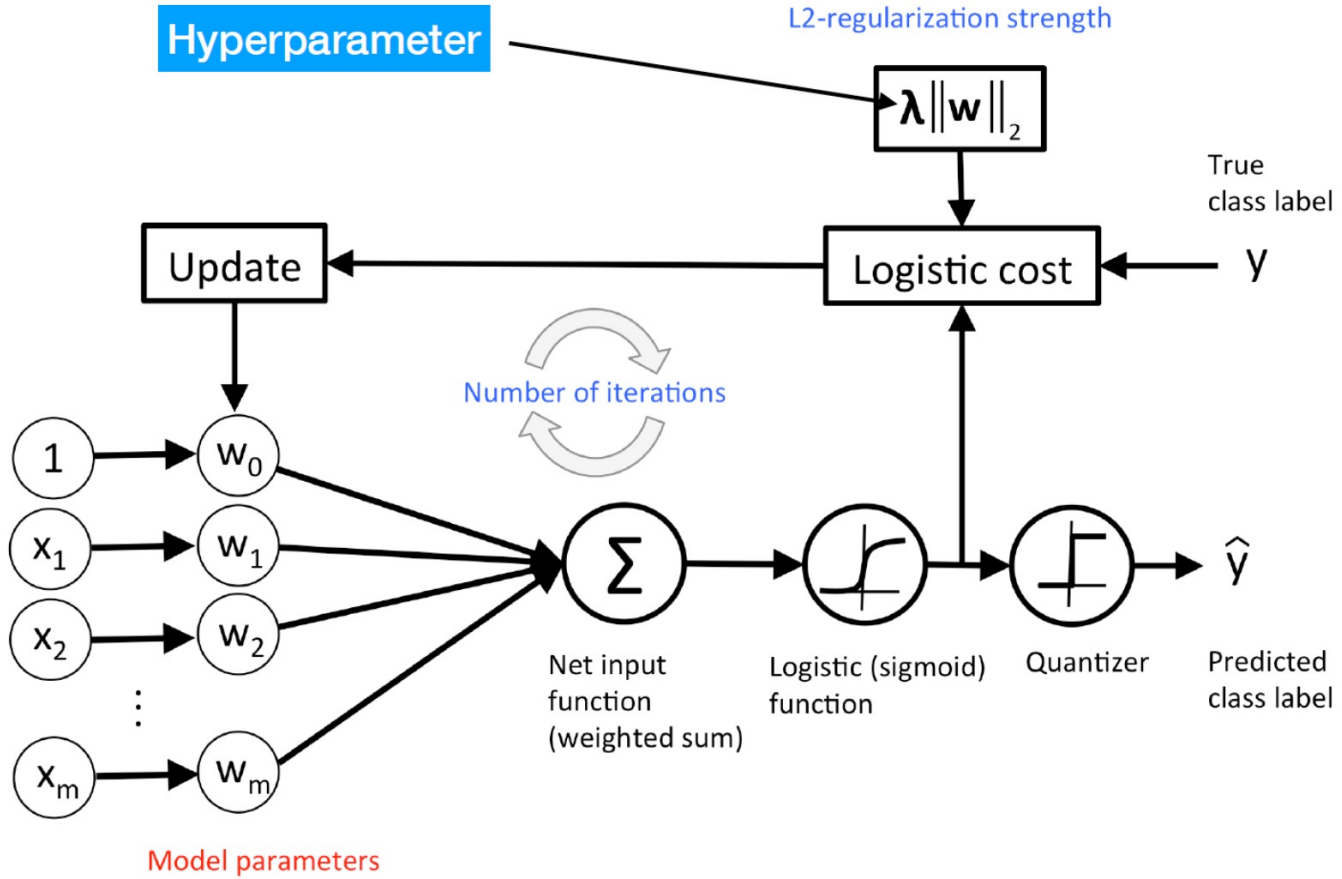
Cross-Validation: Motivation

- Nonparametric model: k-nearest neighbors
- Parametric model: logistic regression $\sigma(\mathbf{wx}) + \lambda\|\mathbf{w}\|_2$



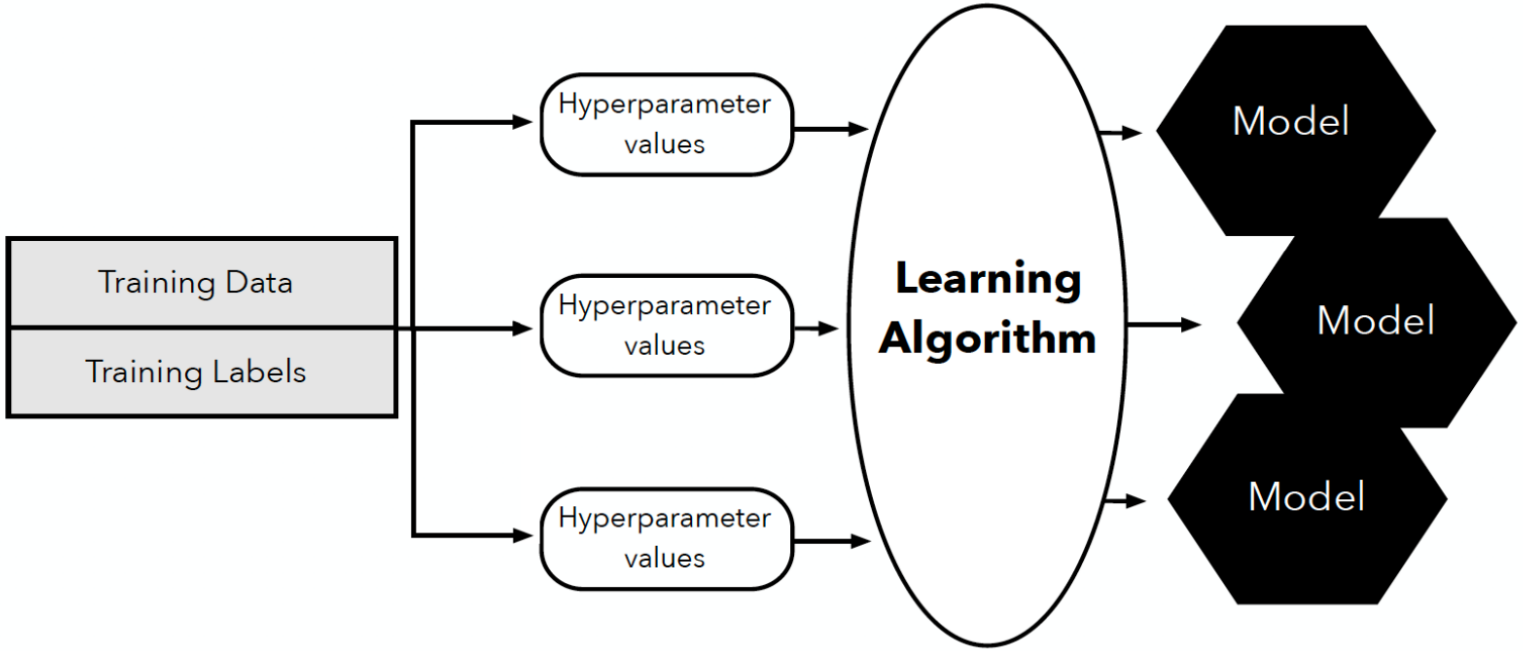
Cross-Validation: Motivation

- Nonparametric model: k-nearest neighbors
- Parametric model: logistic regression $\sigma(\mathbf{wx}) + \lambda\|\mathbf{w}\|_2$

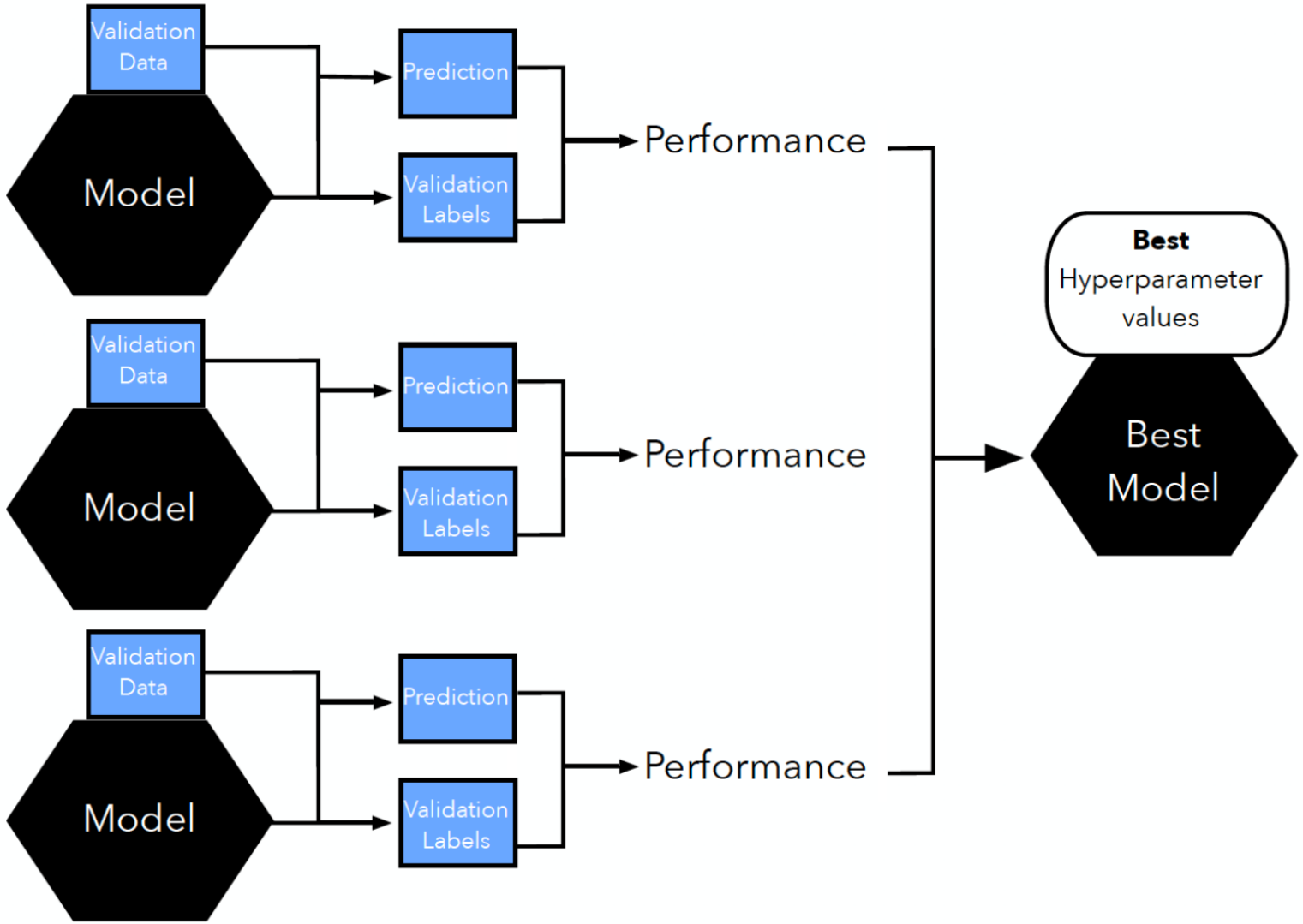


Cross-Validation: Three Way Split

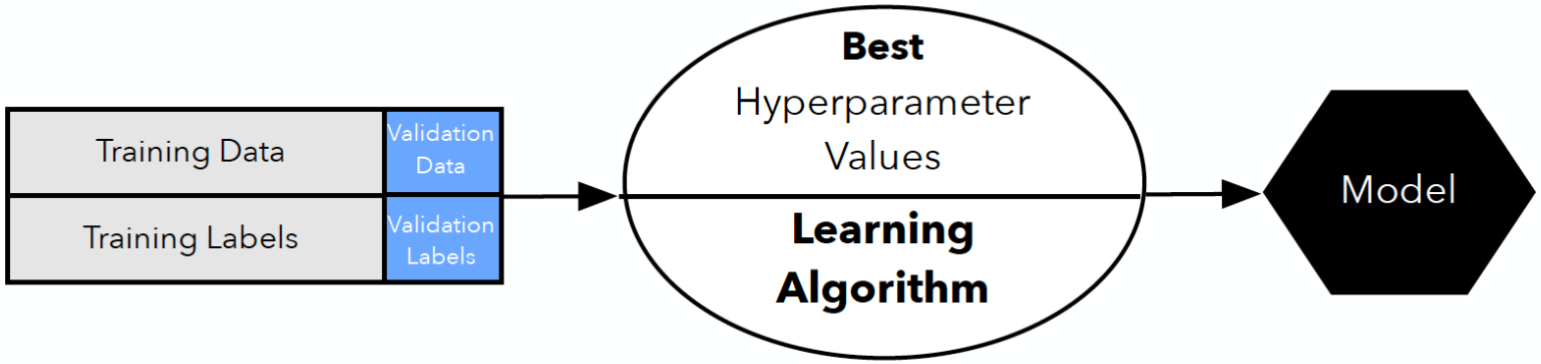
Cross-Validation: Three Way Split



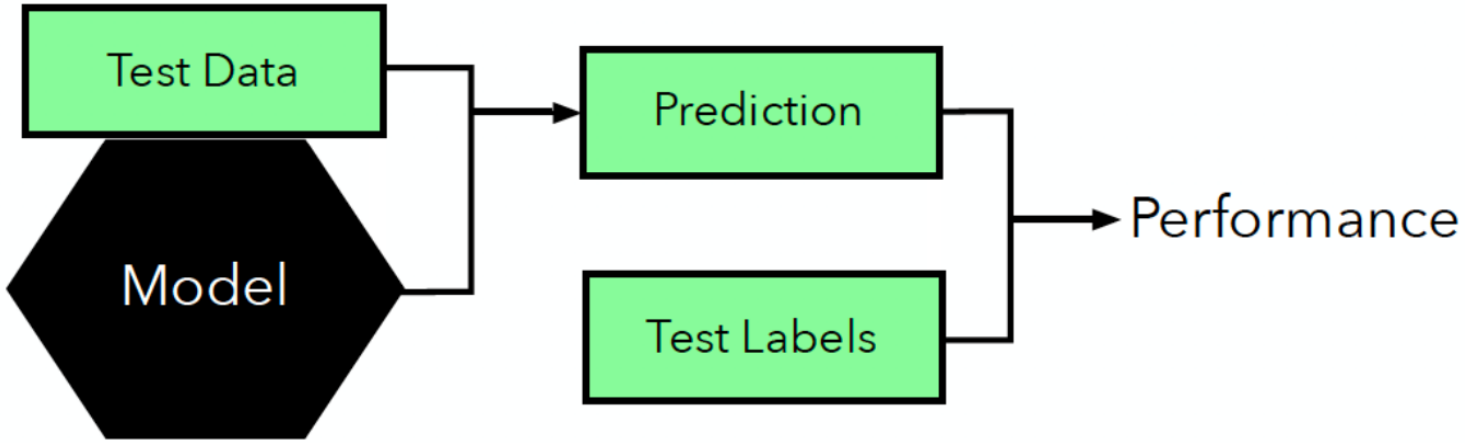
Cross-Validation: Three Way Split



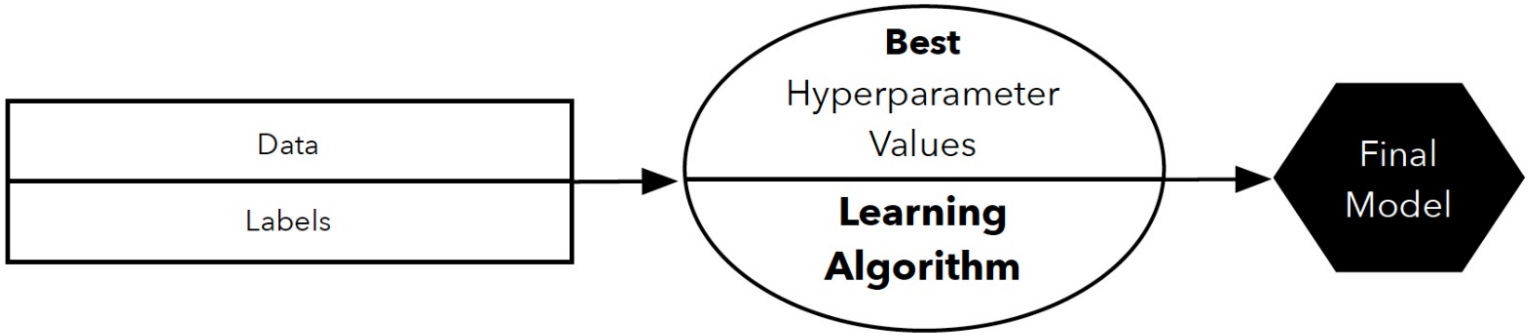
Cross-Validation: Three Way Split



Cross-Validation: Three Way Split



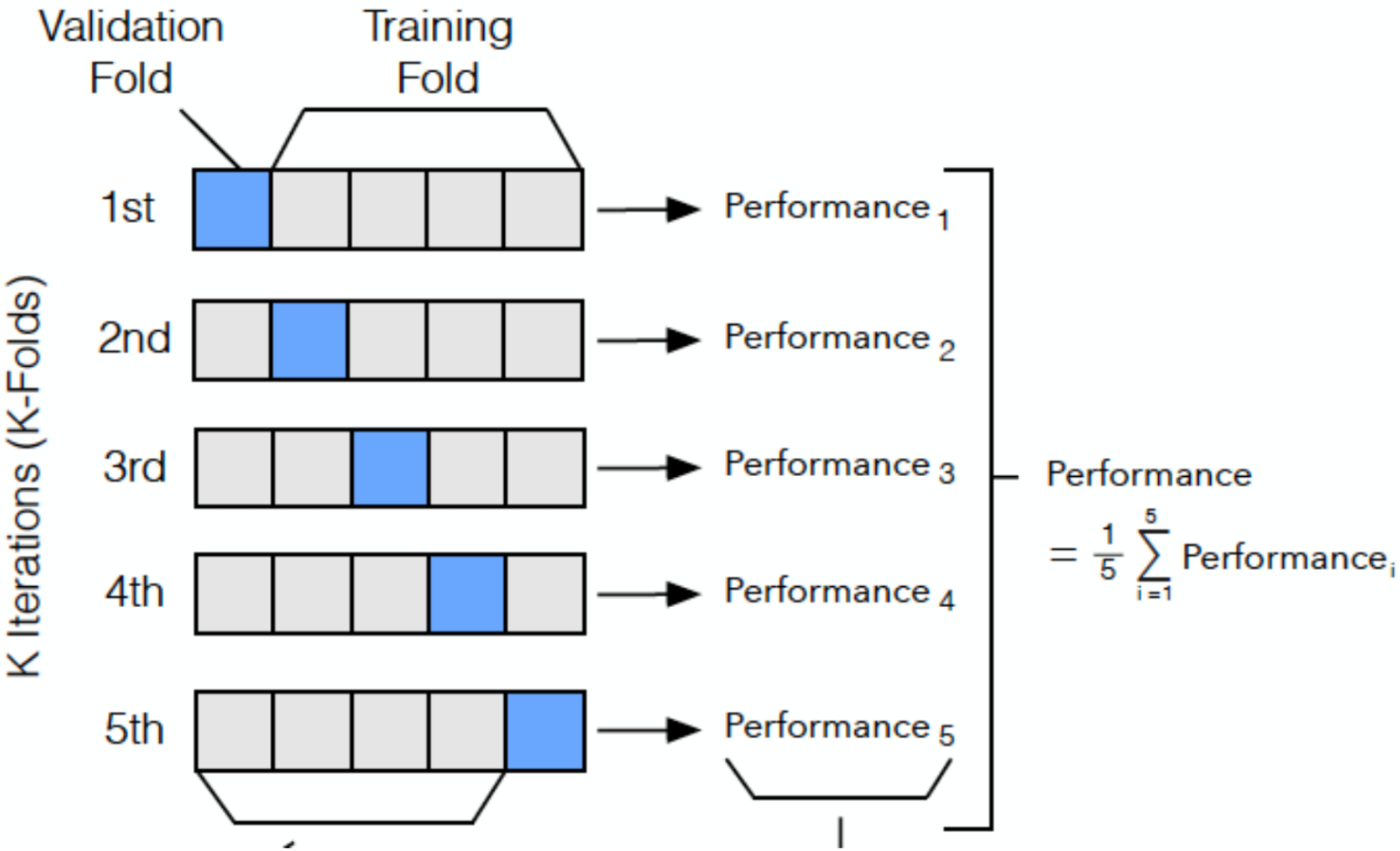
Cross-Validation: Three Way Split



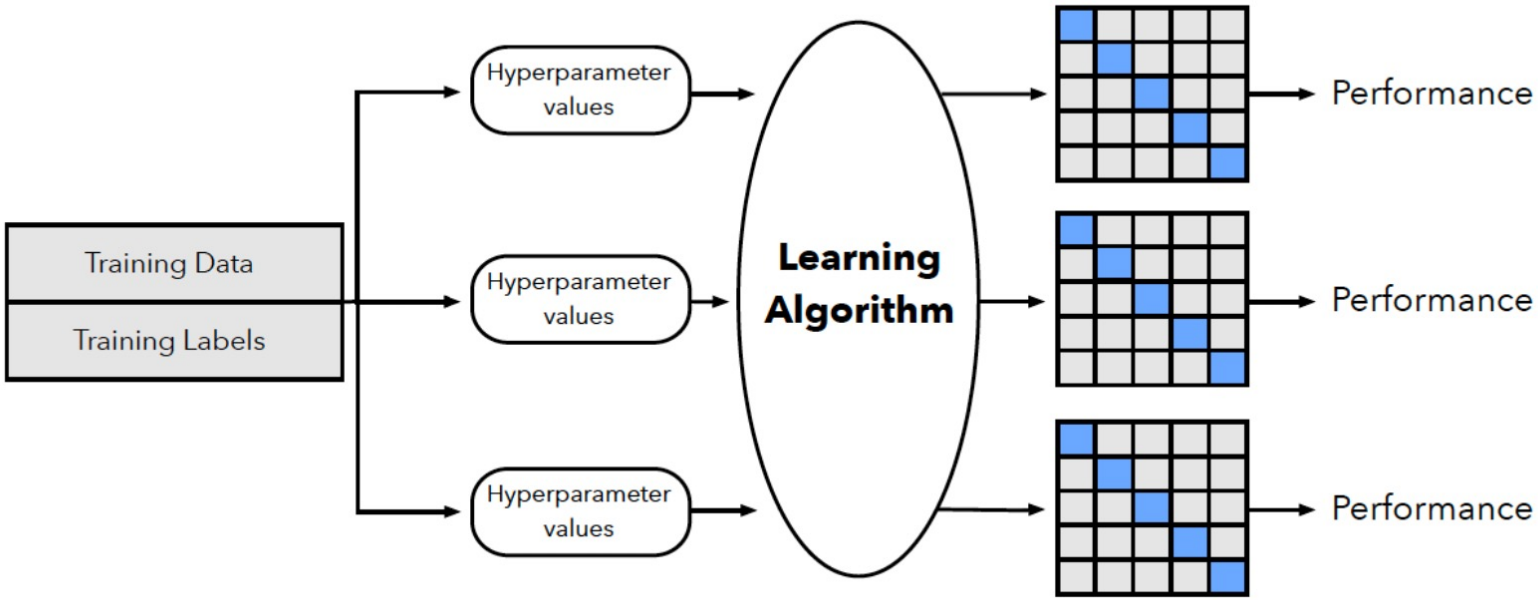
- Pro: fast and simple
- Con: high variance, bad use of data

Cross-Validation: k-Fold CV

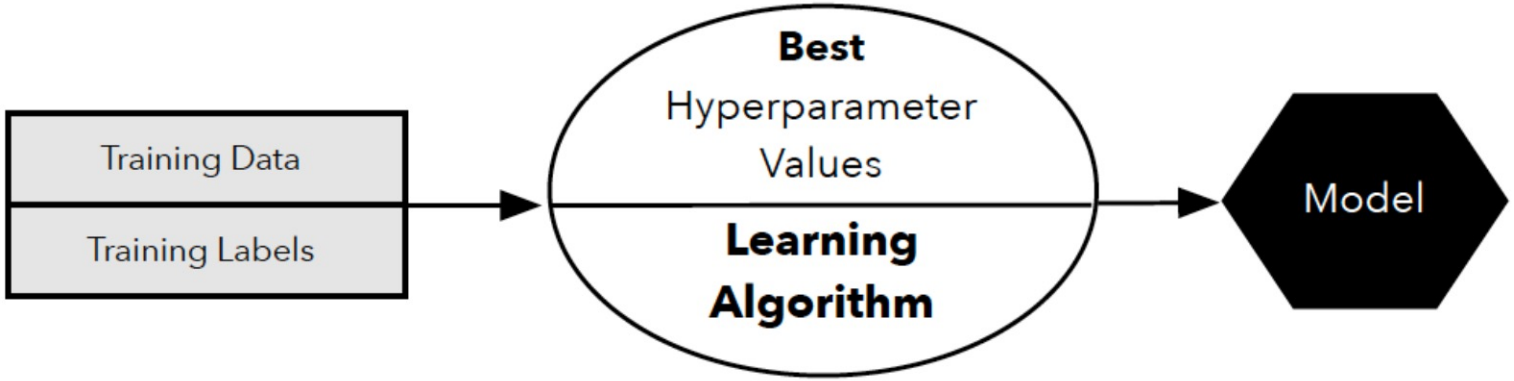
Cross-Validation: k-Fold CV



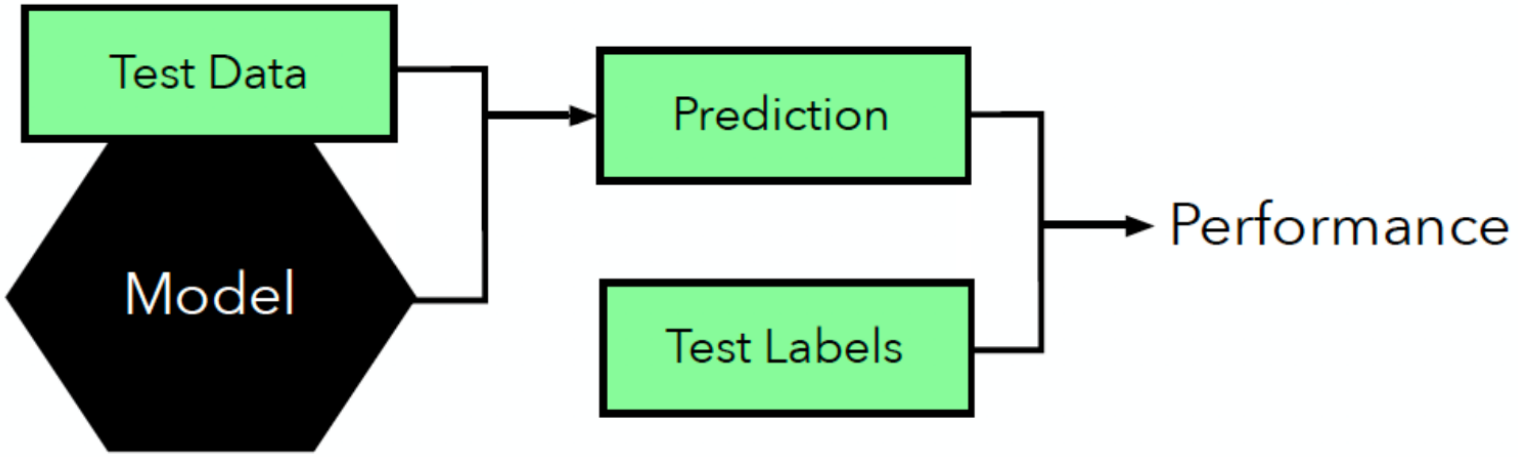
Cross-Validation: k-Fold CV



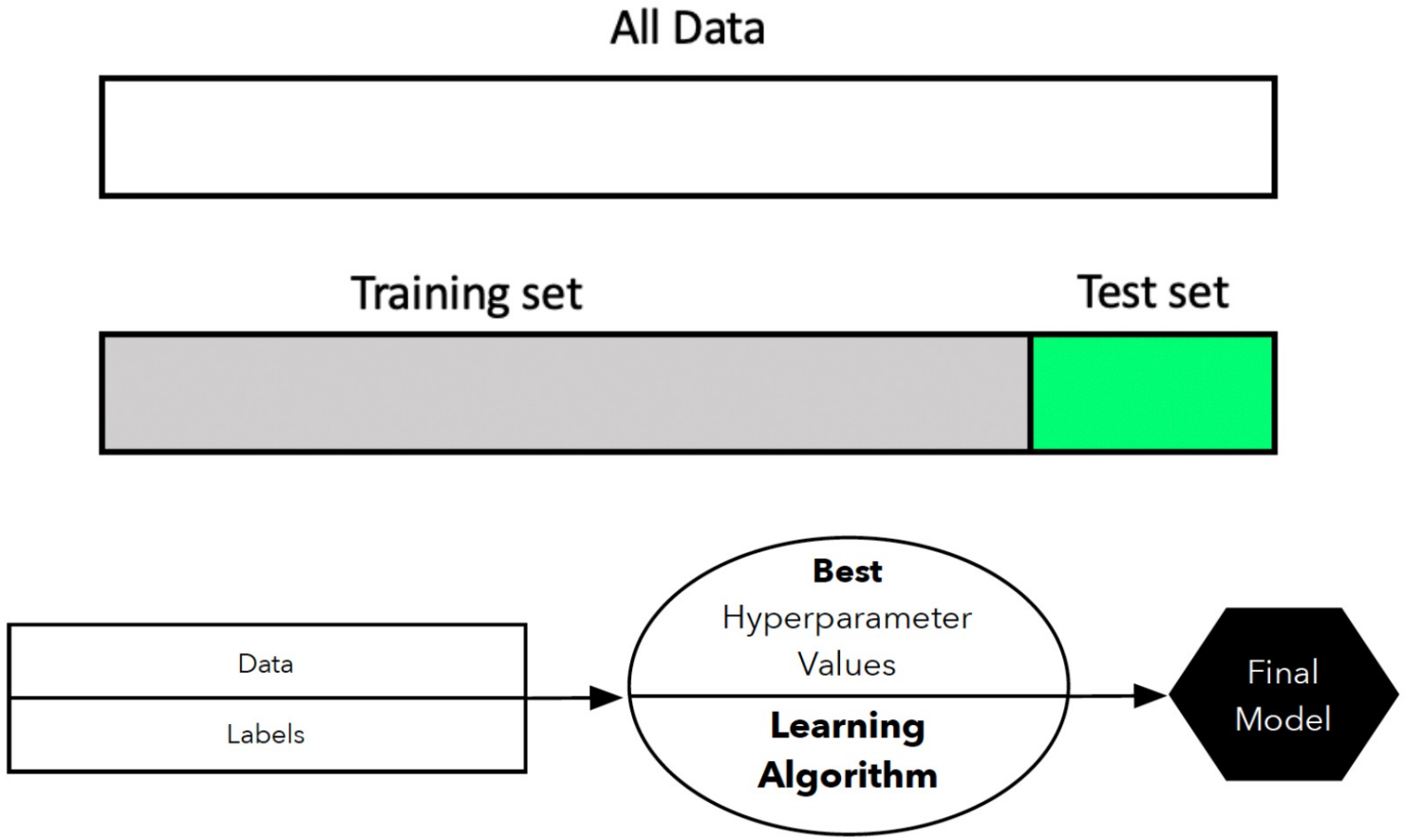
Cross-Validation: k-Fold CV



Cross-Validation: k-Fold CV

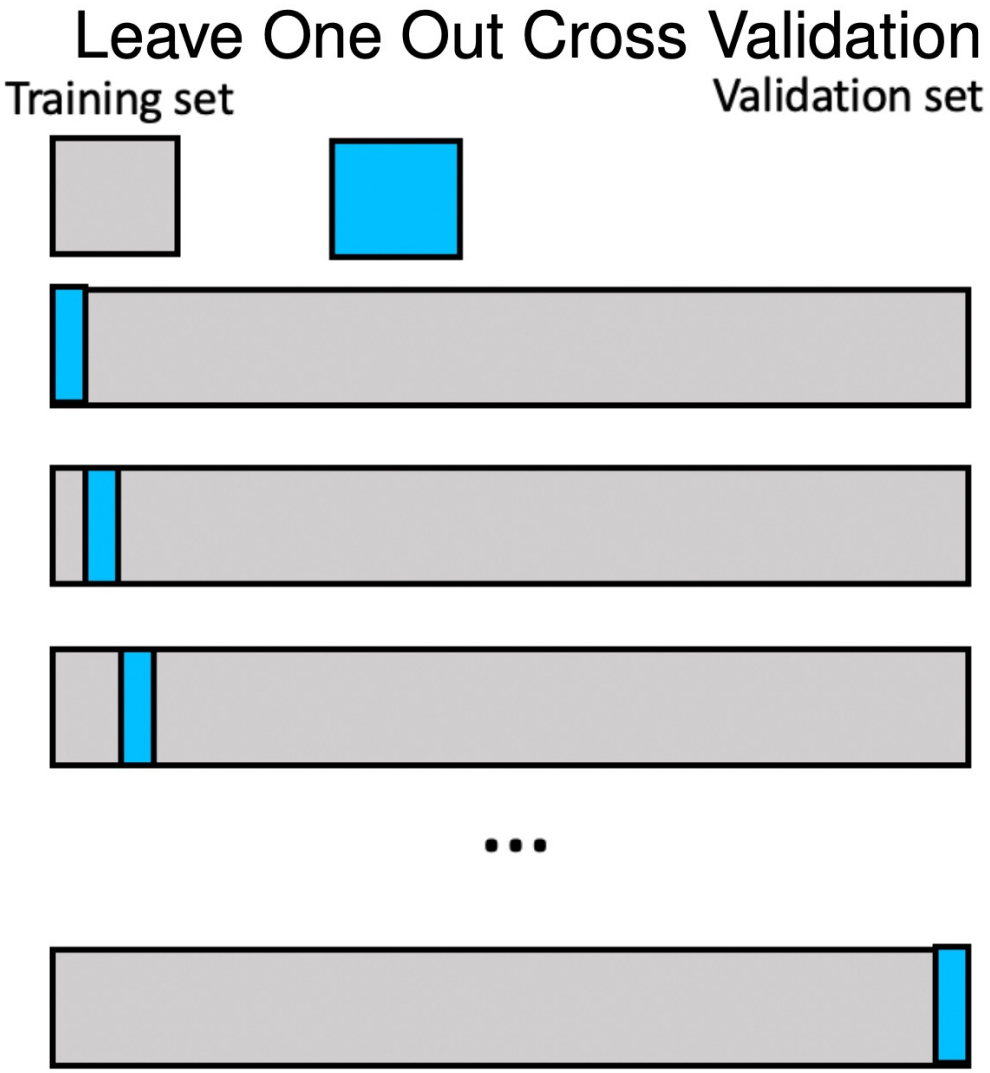


Cross-Validation: k-Fold CV



- Pro: more stable, more data
- Con: slower

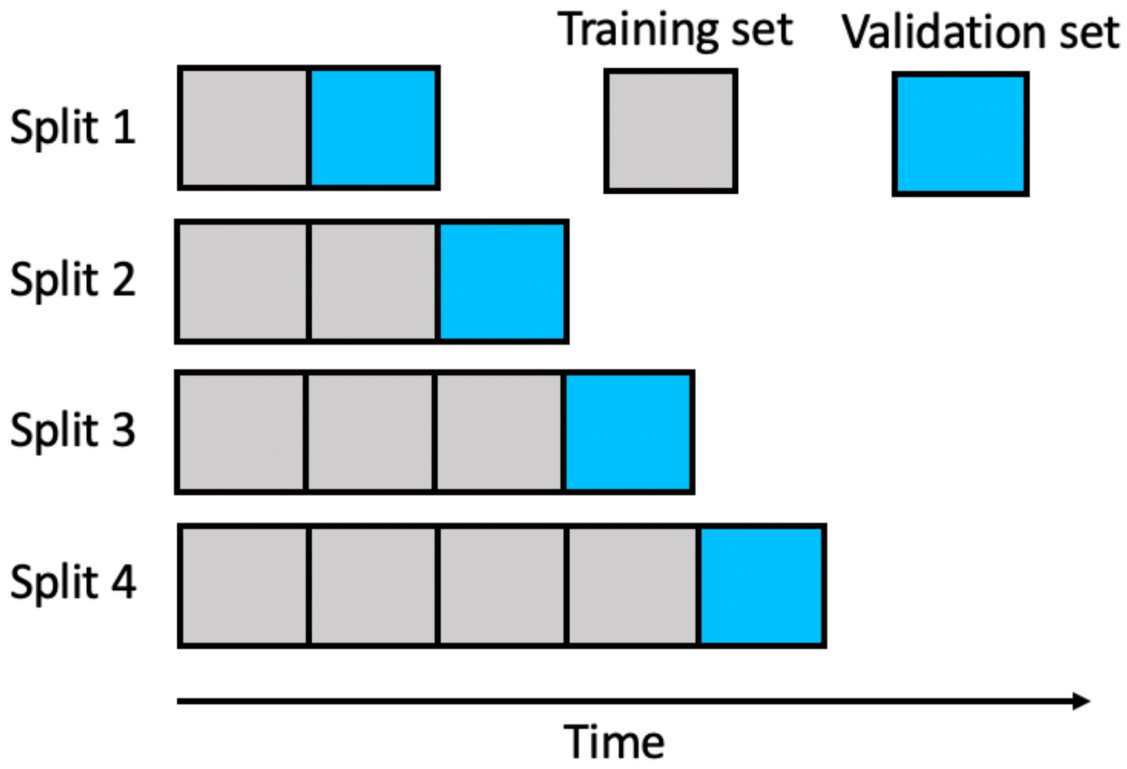
Cross-Validation: LOOCV



Cross-Validation: Summary

	Downside	Upside
Validation-set	may give unreliable estimate of future performance	cheap
Leave-one-out	expensive	doesn't waste data
10-fold	wastes 10% of the data,10 times more expensive than validation set	only wastes 10%, only 10 times more expensive instead of **n** times
3-fold	wastes more data than 10-fold, more expensive than validation set	slightly better than validation-set

Cross-Validation: Sequence Data Split



Cross-Validation: Stratified Split