



UNIVERSITÀ
DEGLI STUDI
DI TRIESTE

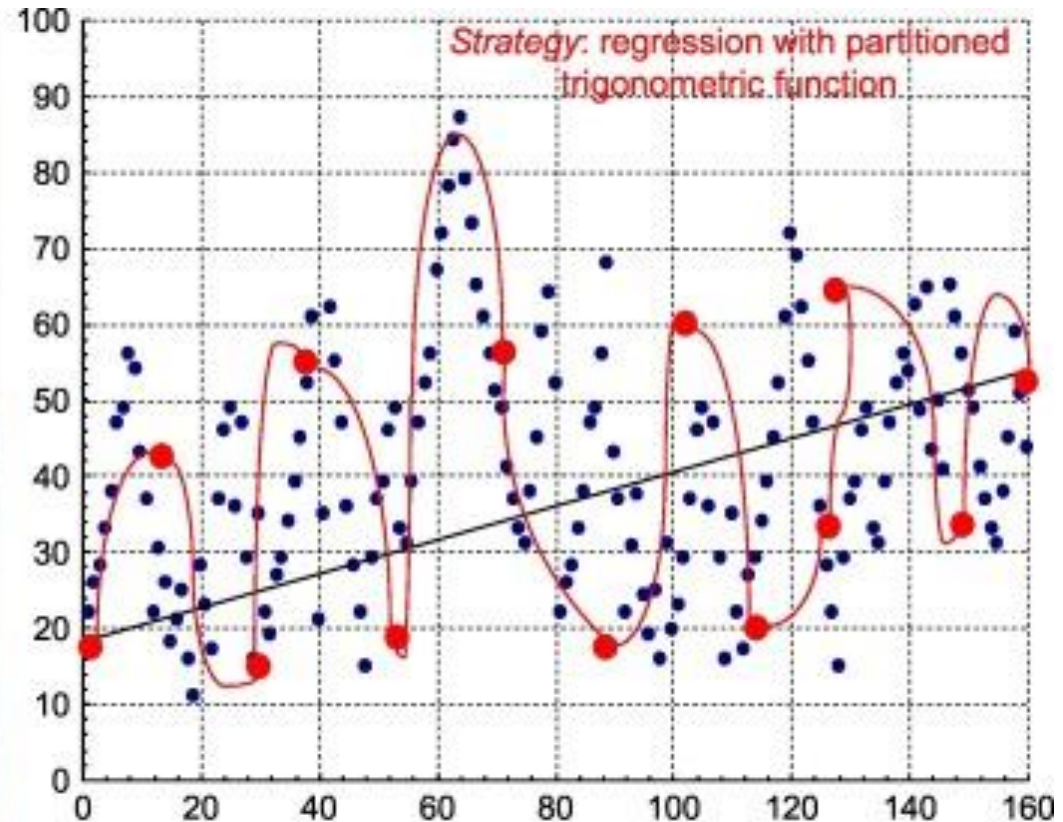
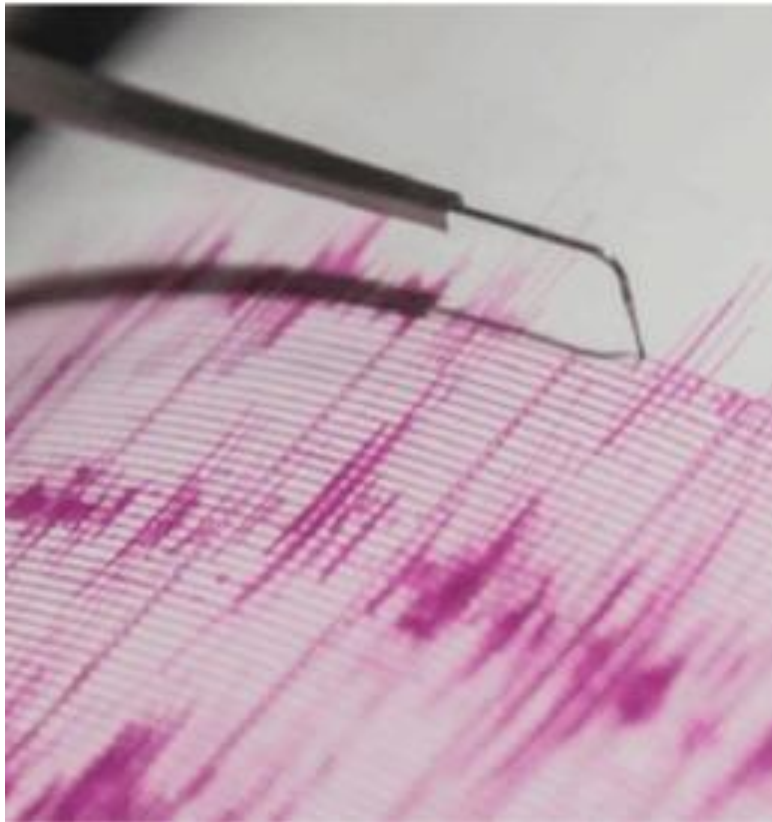
Statistics

Generalization and
bias-variance tradeoff

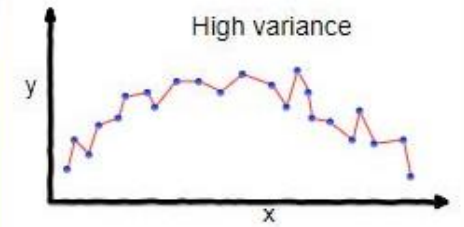
Luca Pennella
September 18th, 2024

Is a fitting line a good choice?

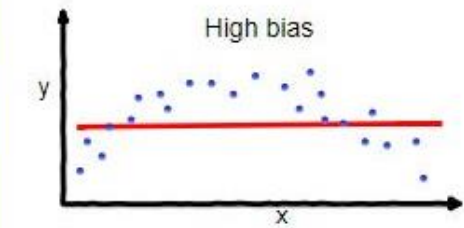
One needs to know priors on the phenomenon studied



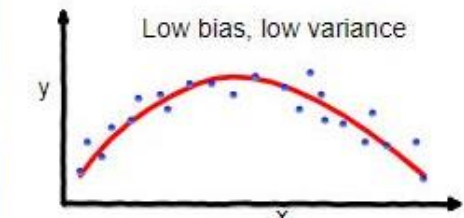
Bias-variance tradeoff and model complexity



overfitting



underfitting



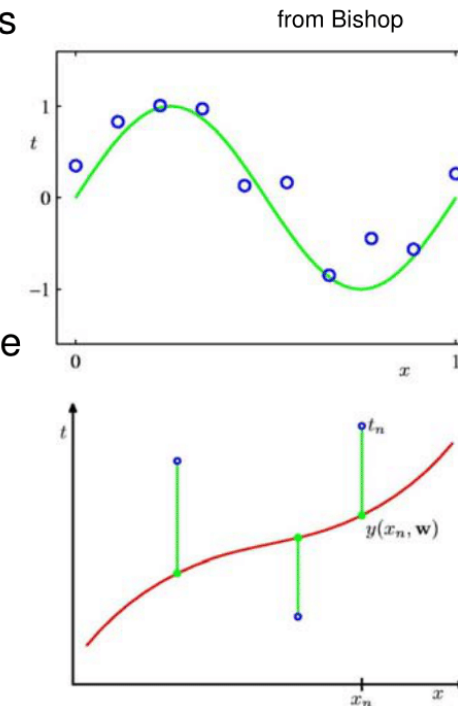
Polynomial curve fitting (Bishop)

Regression example: polynomial curve fitting

- The green curve is the true function (which is not a polynomial)
- The data points are uniform in x but have noise in y .
- We will use a **loss function** that measures the squared error in the prediction of $y(x)$ from x . The loss for the red polynomial is the sum of the squared vertical errors.

$$E(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^N \{y(x_i, \mathbf{w}) - t_i\}^2$$

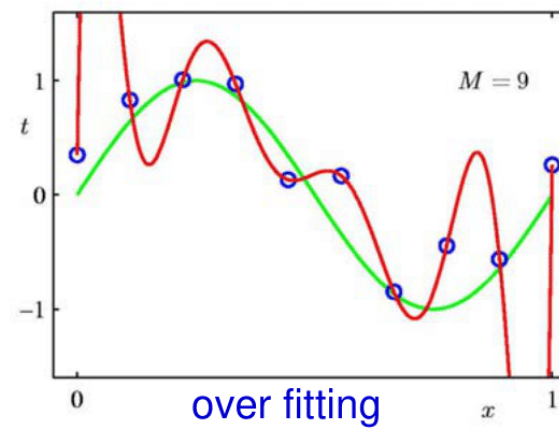
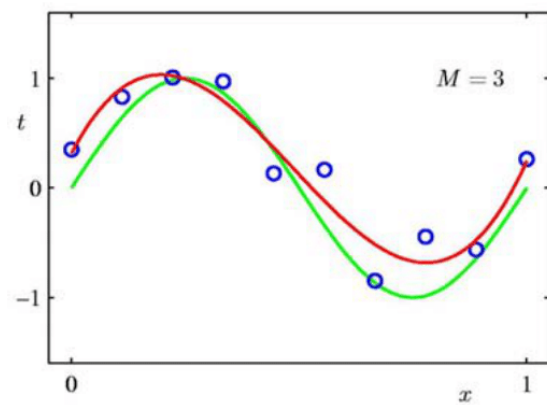
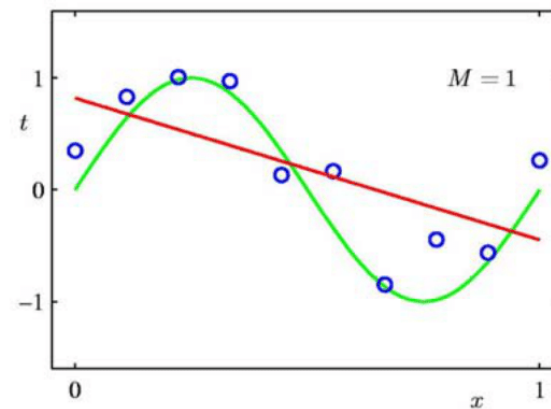
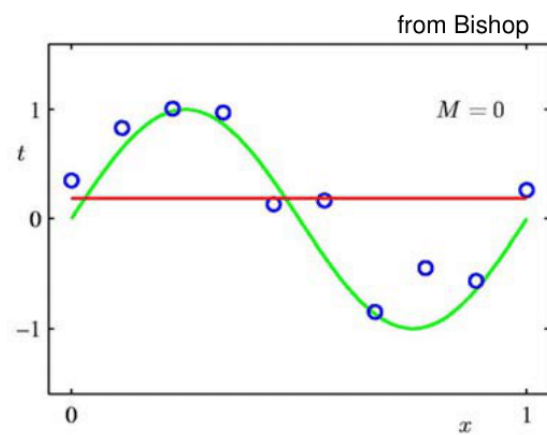
↑
target value



$$y(x, \mathbf{w}) = w_0 + w_1x + w_2x^2 + \dots + w_Mx^M = \sum_{j=0}^M w_jx^j$$

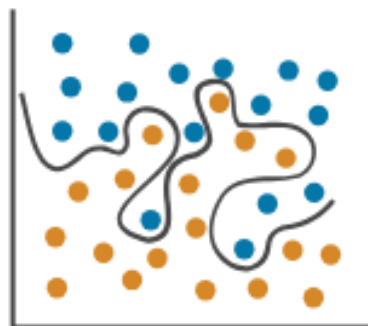
polynomial
regression

Some fits to the data: which is best?

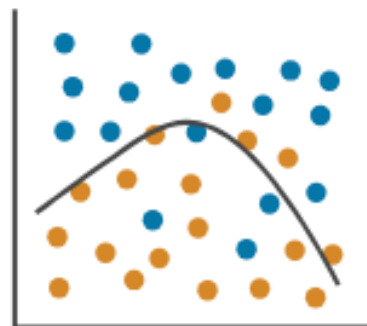


Classification

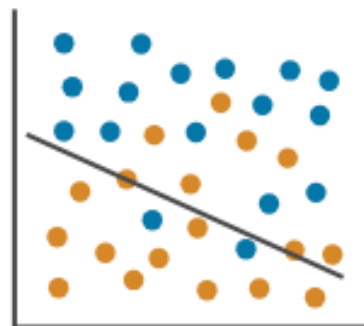
Overfitting



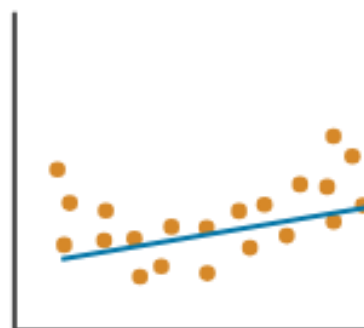
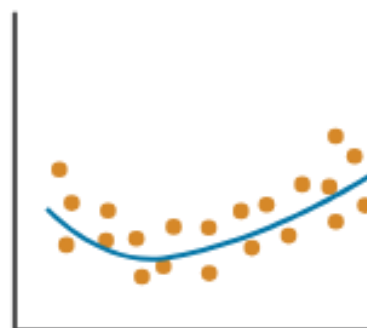
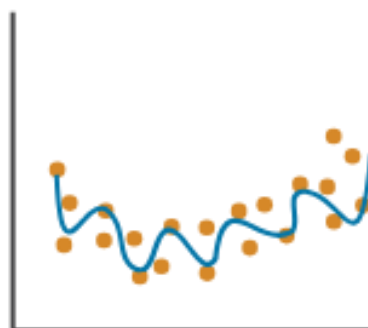
Right Fit



Underfitting



Regression

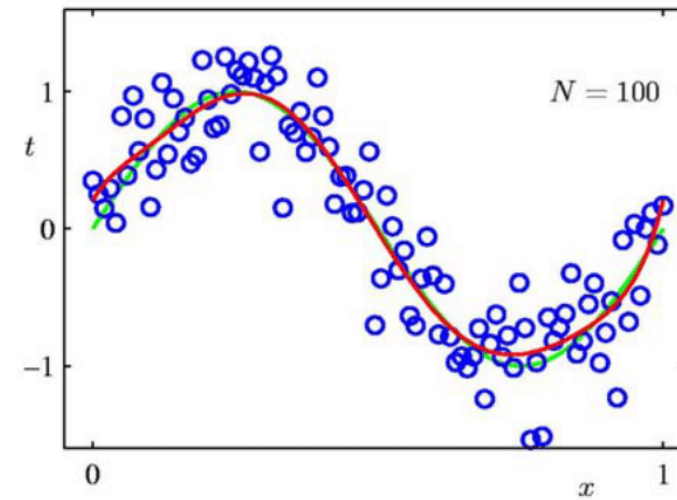
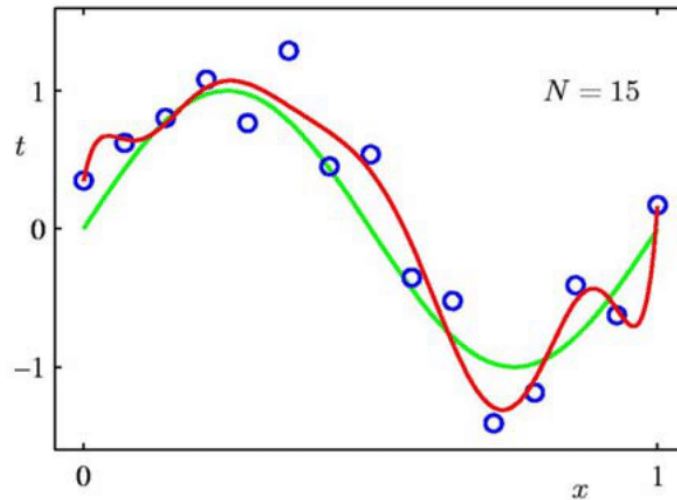


Trading off goodness of fit against model complexity

- If the model has as many degrees of freedom as the data, it can fit the training data perfectly
- But the objective in ML is generalization
- Can expect a model to generalize well if it explains the training data surprisingly well given the complexity of the model.

How to prevent over fitting? I

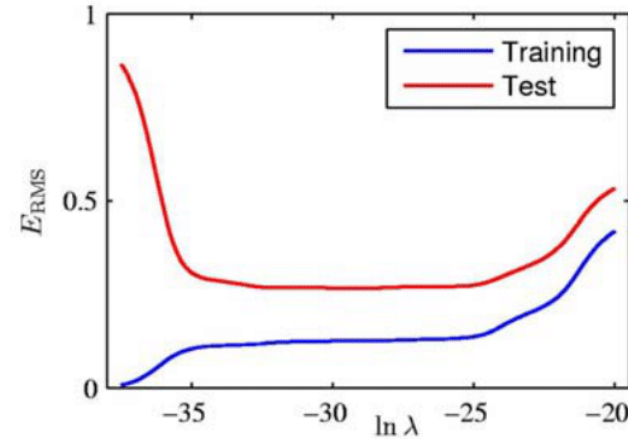
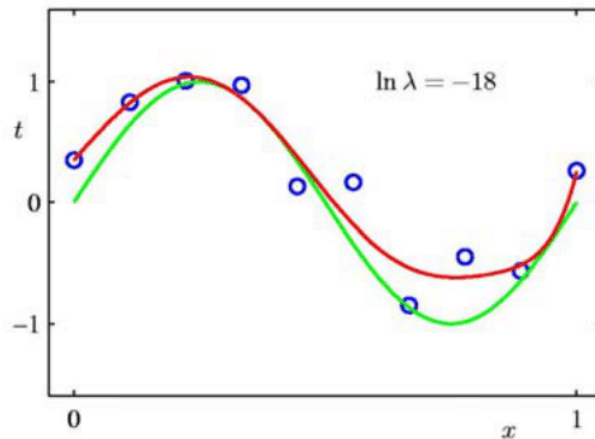
- Add more data than the model “complexity”
- For 9th order polynomial:



How to prevent over fitting? II

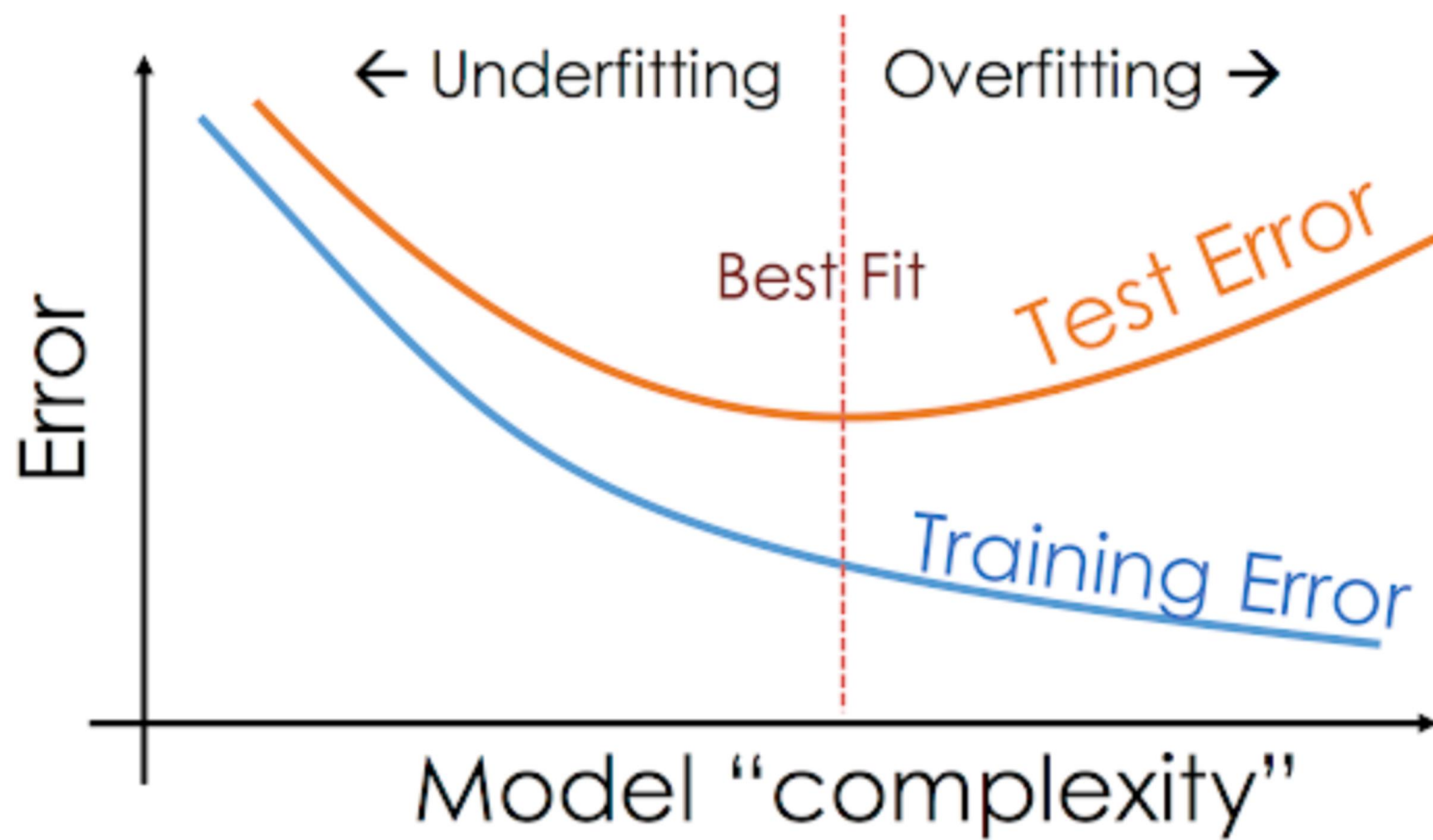
- Regularization: penalize large coefficient values

$$\tilde{E}(\mathbf{w}) = \underbrace{\frac{1}{2} \sum_{i=1}^N \{y(x_i, \mathbf{w}) - t_i\}^2}_{\text{loss function}} + \underbrace{\frac{\lambda}{2} \|\mathbf{w}\|^2}_{\text{regularization}} \quad \text{"ridge" regression}$$



In practice use **validation** data to choose λ (not test)

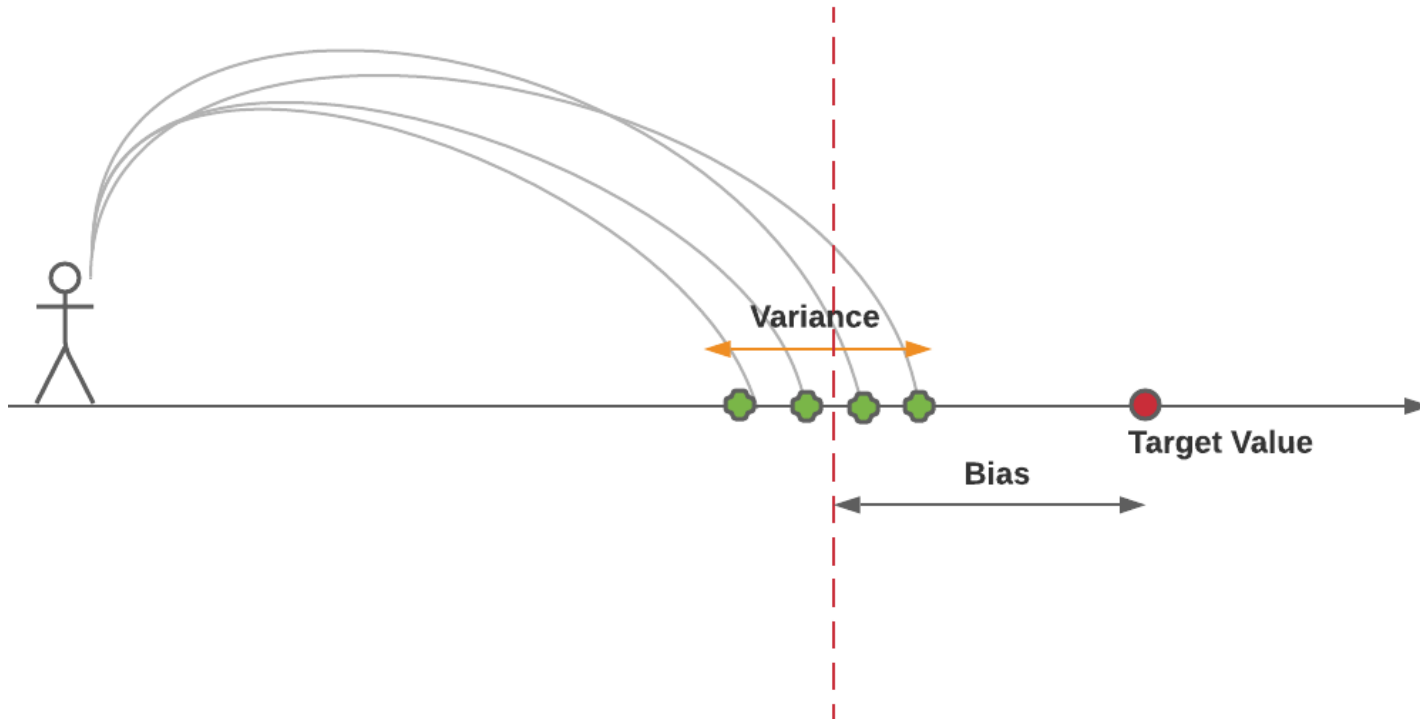
We will return to regularization later



Bias and variance

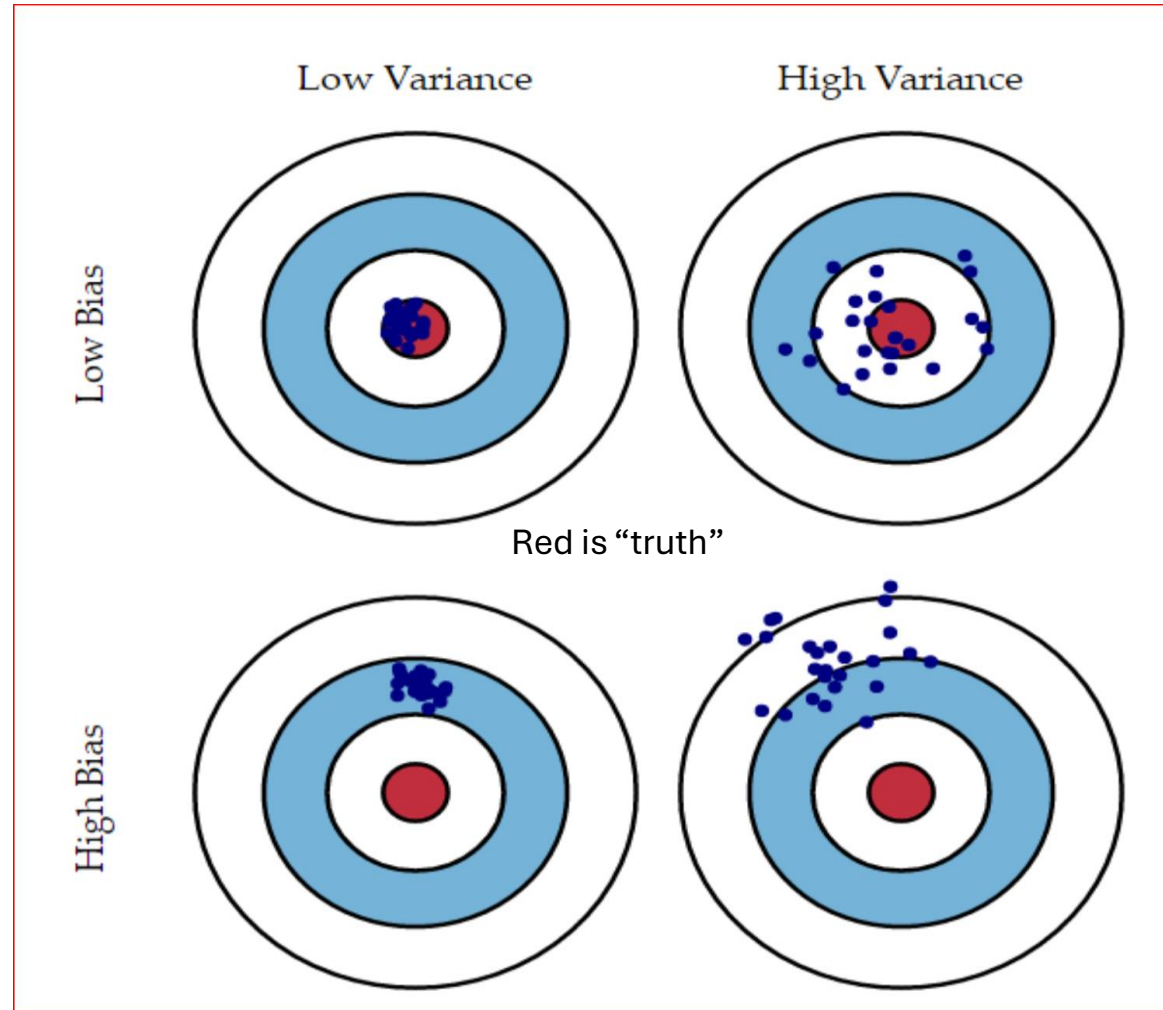
+ Bias: is measured by the difference between what the model gives and the true value

+ Variance: variability of predictions from different rounds of runs



$$\sigma^2 = \frac{\sum (xi - \bar{x})^2}{N}$$

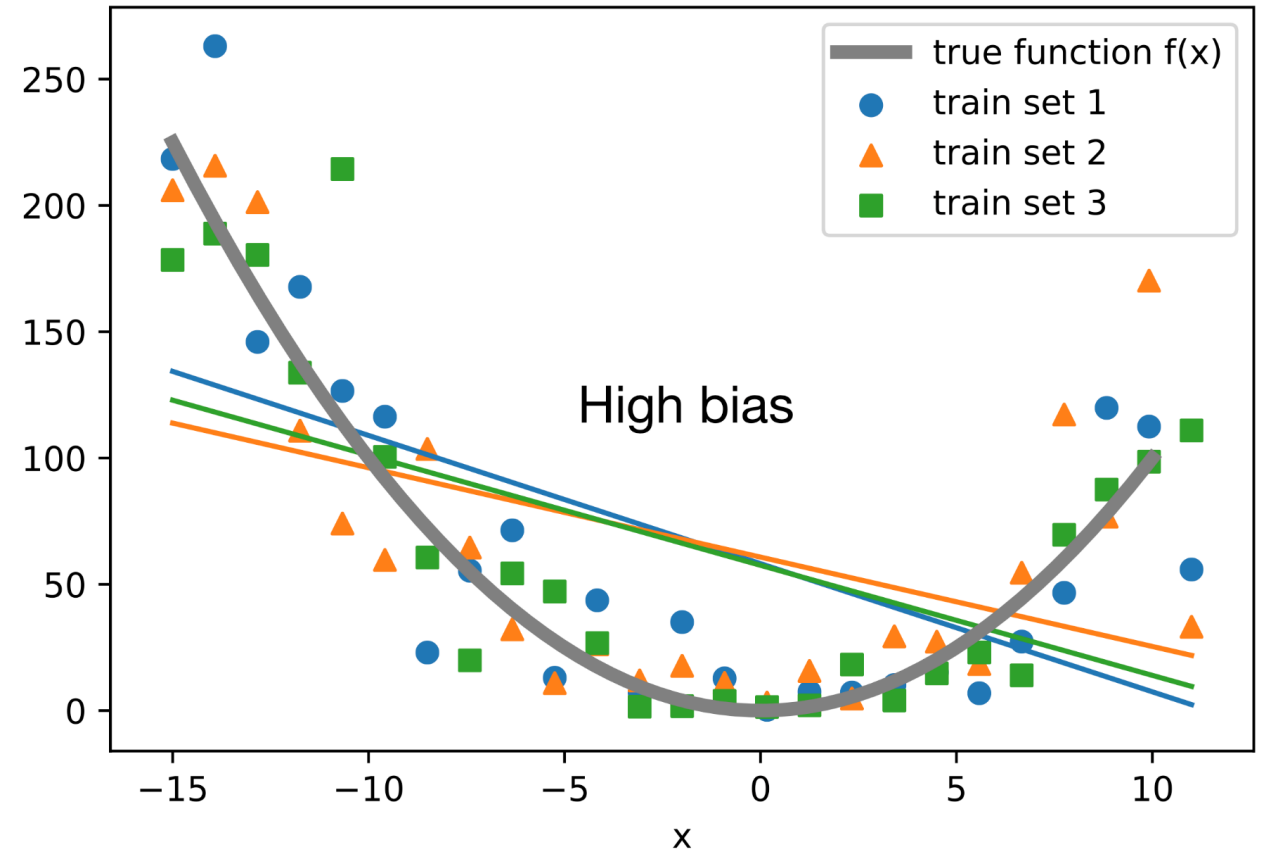
Bias and variance



Bias

+ The difference between the true value and the model prediction is large : high bias

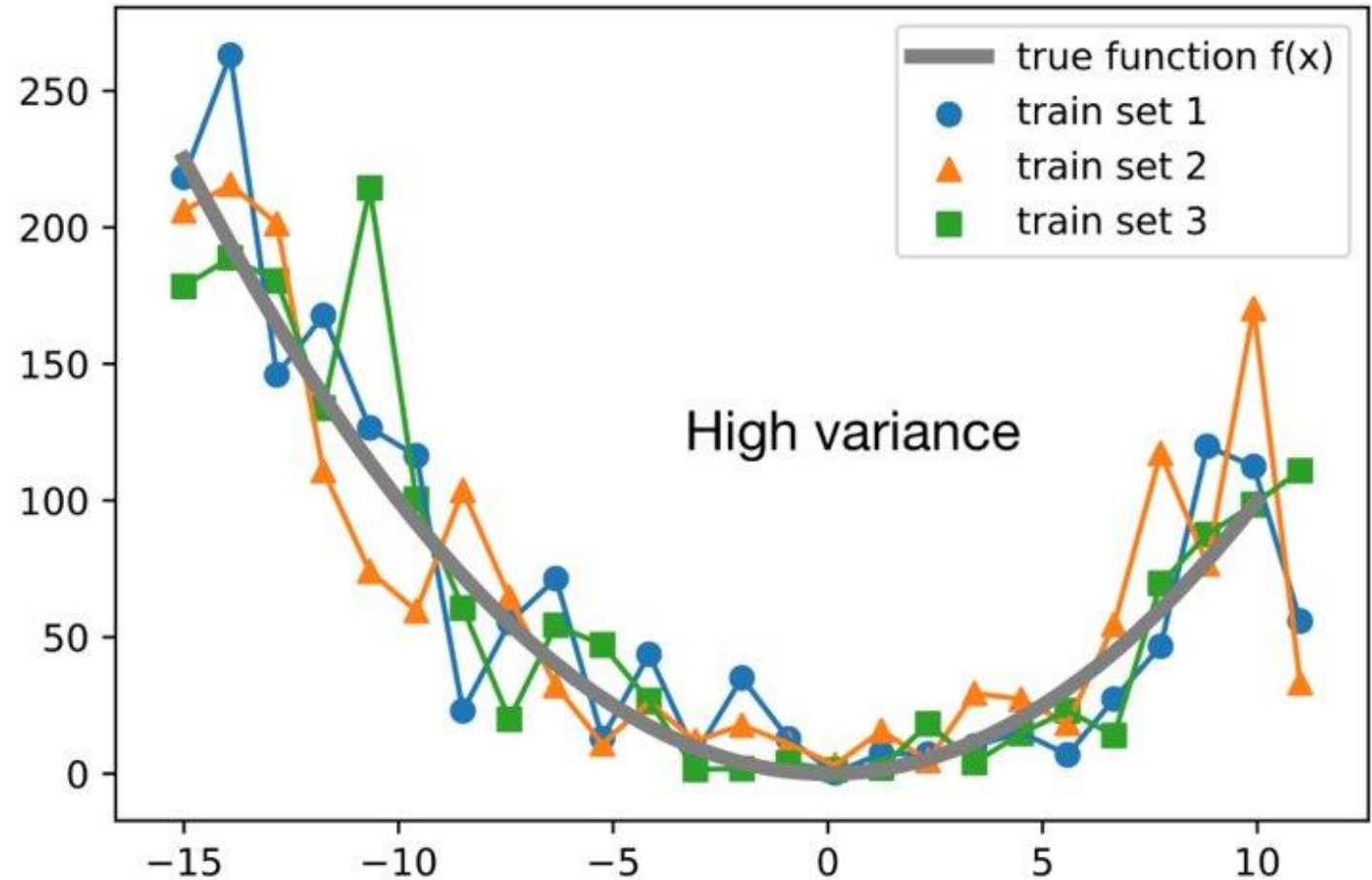
+ What's the variance across the three models prediction?



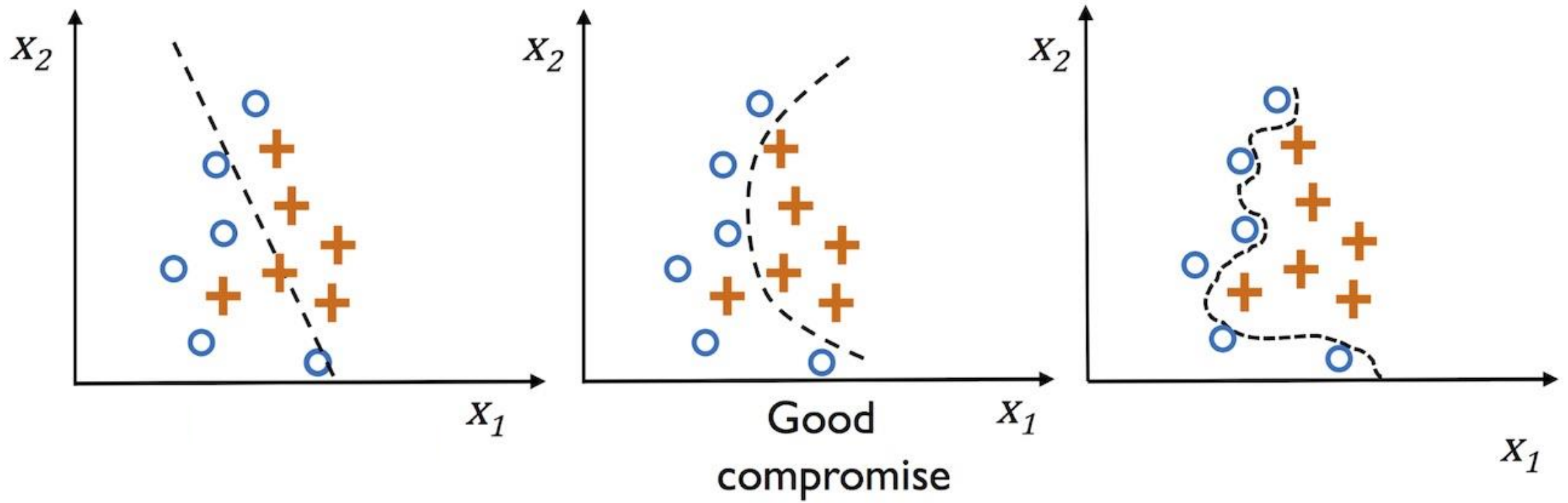
Variance

+ For different datasets (in the same distribution) the y predicted by the model is very different

+ What's the bias: think about the models average

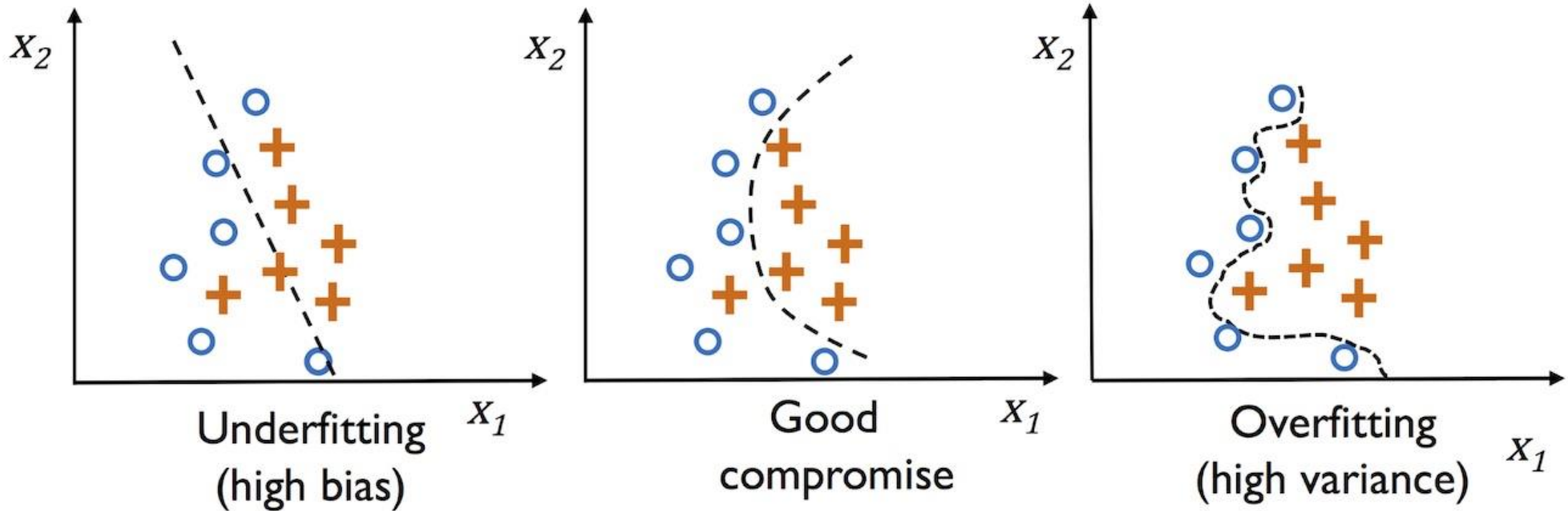


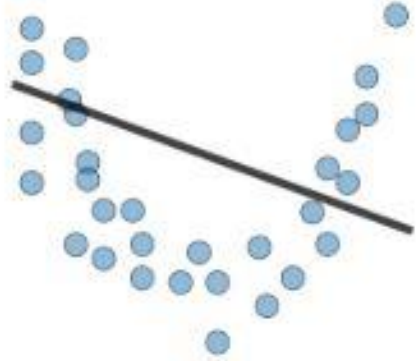


Bias and variance tradeoff



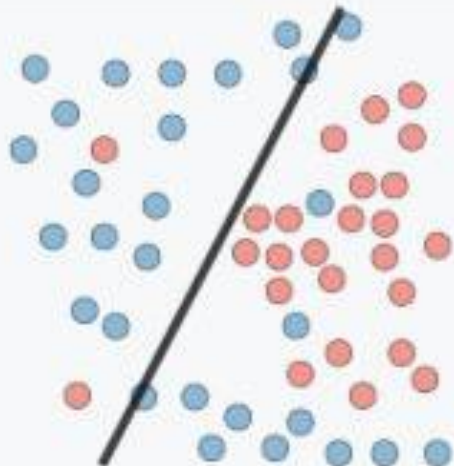
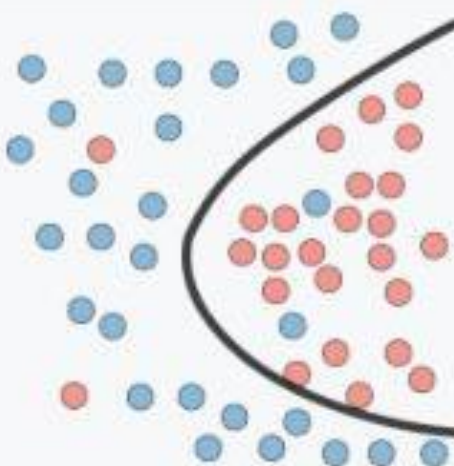
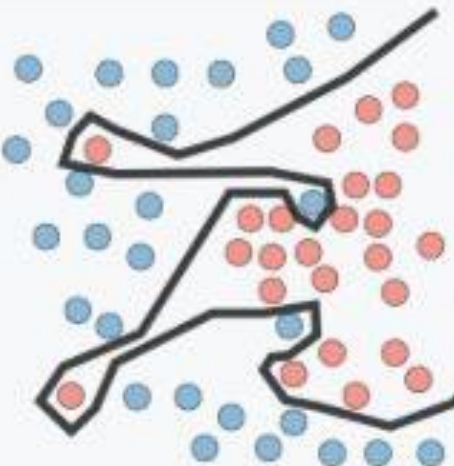
Bias and variance tradeoff


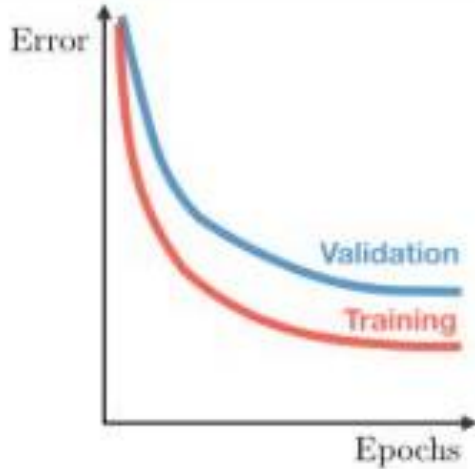
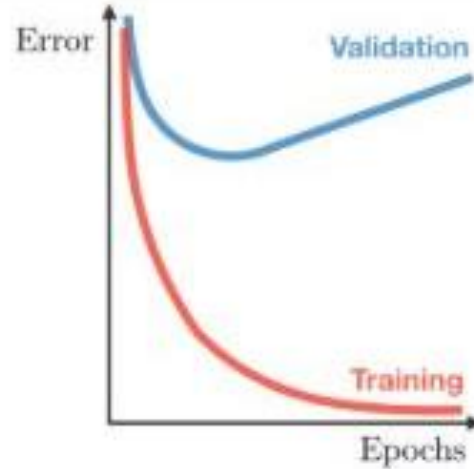
+ Suppose you have another dataset (from the same distribution): how would the separators look like in 1 and 3 in terms of difference from the ones above? How much is the classifier varying for different training data?



	Underfitting	Just right	Overfitting
Symptoms	<ul style="list-style-type: none"> • High training error • Training error close to test error • High bias 	<ul style="list-style-type: none"> • Training error slightly lower than test error 	<ul style="list-style-type: none"> • Very low training error • Training error much lower than test error • High variance
Regression illustration			
Possible remedies	<ul style="list-style-type: none"> • Complexify model • Add more features • Train longer 		<ul style="list-style-type: none"> • Perform regularization • Get more data

Why getting more data is a remedy for Overfitting?

	Underfitting	Just right	Overfitting
Symptoms	<ul style="list-style-type: none"> • High training error • Training error close to test error • High bias 	<ul style="list-style-type: none"> • Training error slightly lower than test error 	<ul style="list-style-type: none"> • Very low training error • Training error much lower than test error • High variance
Classification illustration			
Possible remedies	<ul style="list-style-type: none"> • Complexify model • Add more features • Train longer 		<ul style="list-style-type: none"> • Perform regularization • Get more data

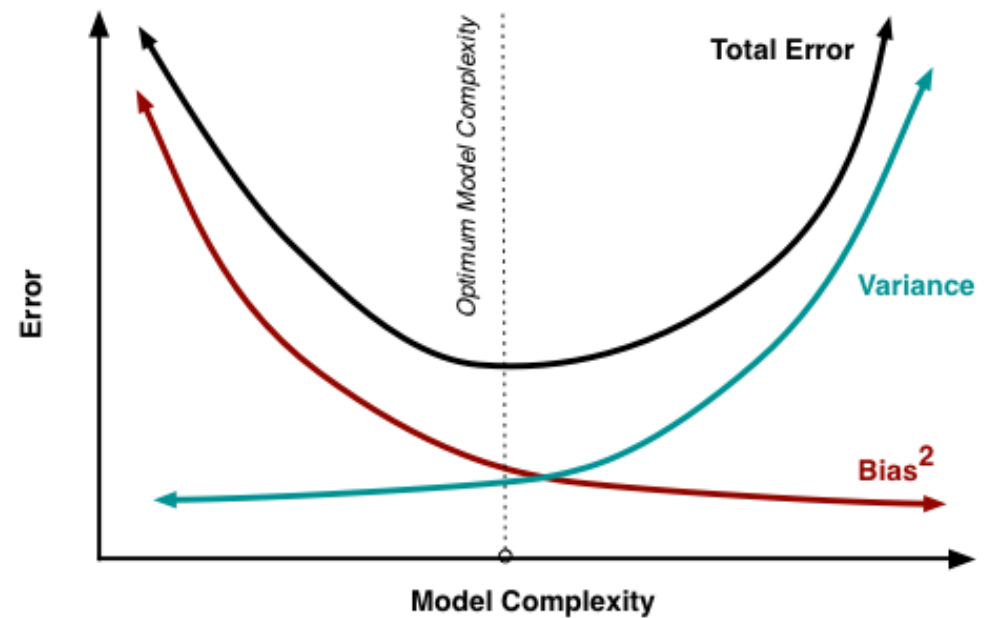
	Underfitting	Just right	Overfitting
Symptoms	<ul style="list-style-type: none"> • High training error • Training error close to test error • High bias 	<ul style="list-style-type: none"> • Training error slightly lower than test error 	<ul style="list-style-type: none"> • Very low training error • Training error much lower than test error • High variance
Deep learning illustration			
Possible remedies	<ul style="list-style-type: none"> • Complexify model • Add more features • Train longer 		<ul style="list-style-type: none"> • Perform regularization • Get more data

Variance: Captures how much your classifier changes if you train on a different training set. How "over-specialized" is your classifier to a particular training set (overfitting)? If we have the best possible model for our training data, how far off are we from the average classifier?

Bias: What is the inherent error that you obtain from your classifier even with infinite training data? This is due to your classifier being "biased" to a particular kind of solution (e.g. linear classifier). In other words, bias is inherent to your model.

SO...

- + We have the data and we split in train and test (validation)
- + We choose a model and train on train data and test on test data
- + We fix an acceptable error: if the training or test error is too high we have a poor model





UNIVERSITÀ
DEGLI STUDI
DI TRIESTE

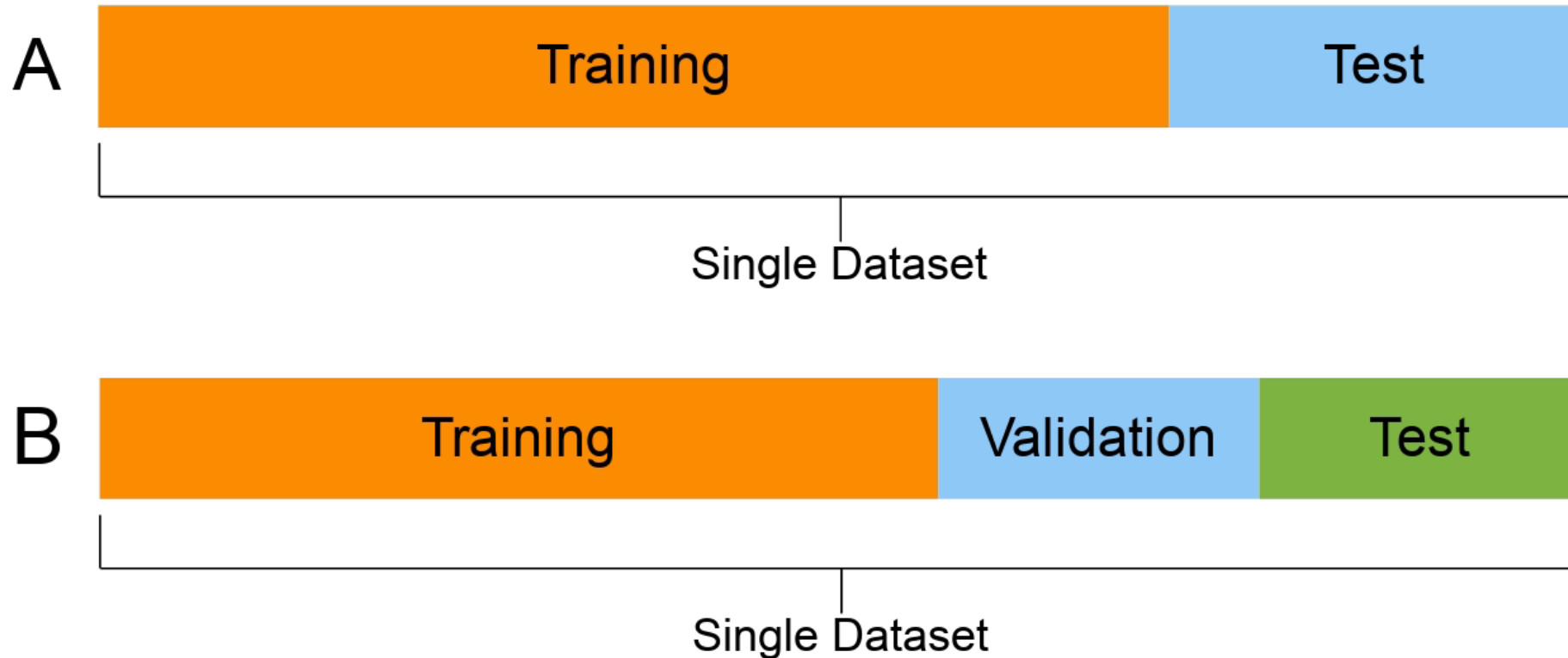
SUPERVISED LEARNING MODEL SELECTION & ASSESSMENT

Testing for good models: validation e cross-validation

- Testing if the model has extracted the correct information from the data.
- A validation dataset is a sample of data held back from training your model that is used to give an estimate of model skill while tuning model's hyperparameters.
- The validation dataset is different from the test dataset that is also held back from the training of the model, but is instead used to give an unbiased estimate of the skill of the final tuned model when comparing or selecting between final models.

How 'good' is the model? Test set

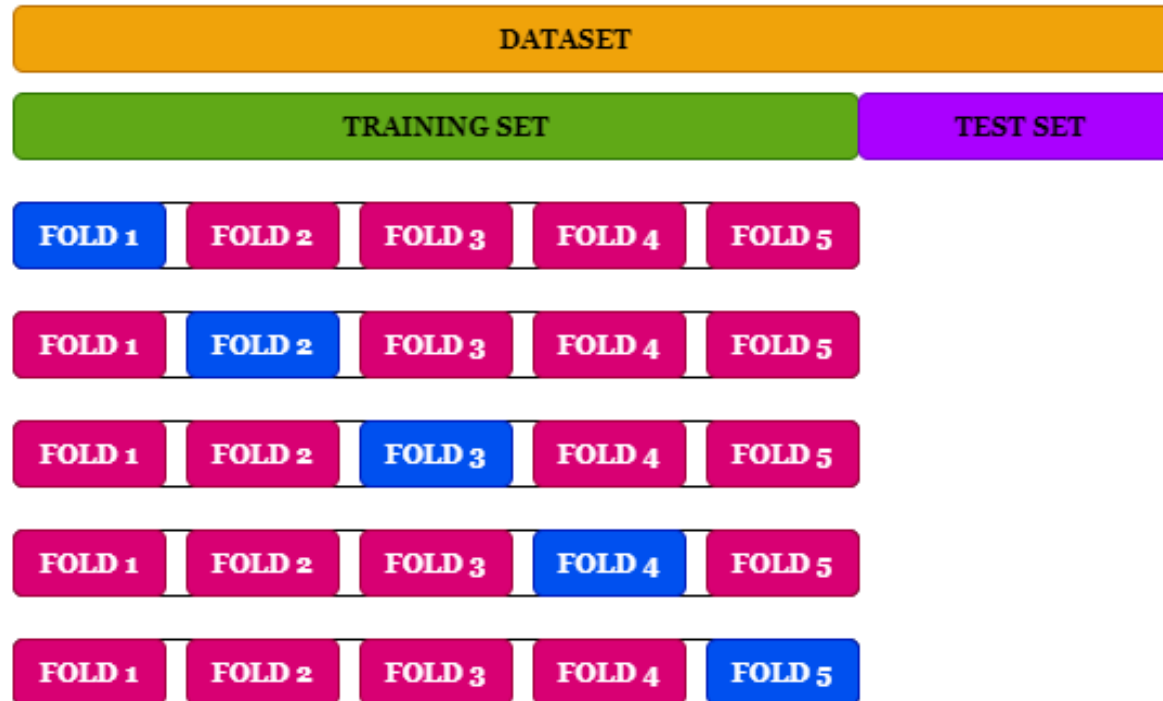
+ The ability to predict well is called
GENERALIZATION



Validation

- The scores for evaluating the model are highly variable depending on the observations that make up in train set and validation set from a single train validation split.
- Reserving much of the data for a single validation set reduces the number of observations we can use to train the model.

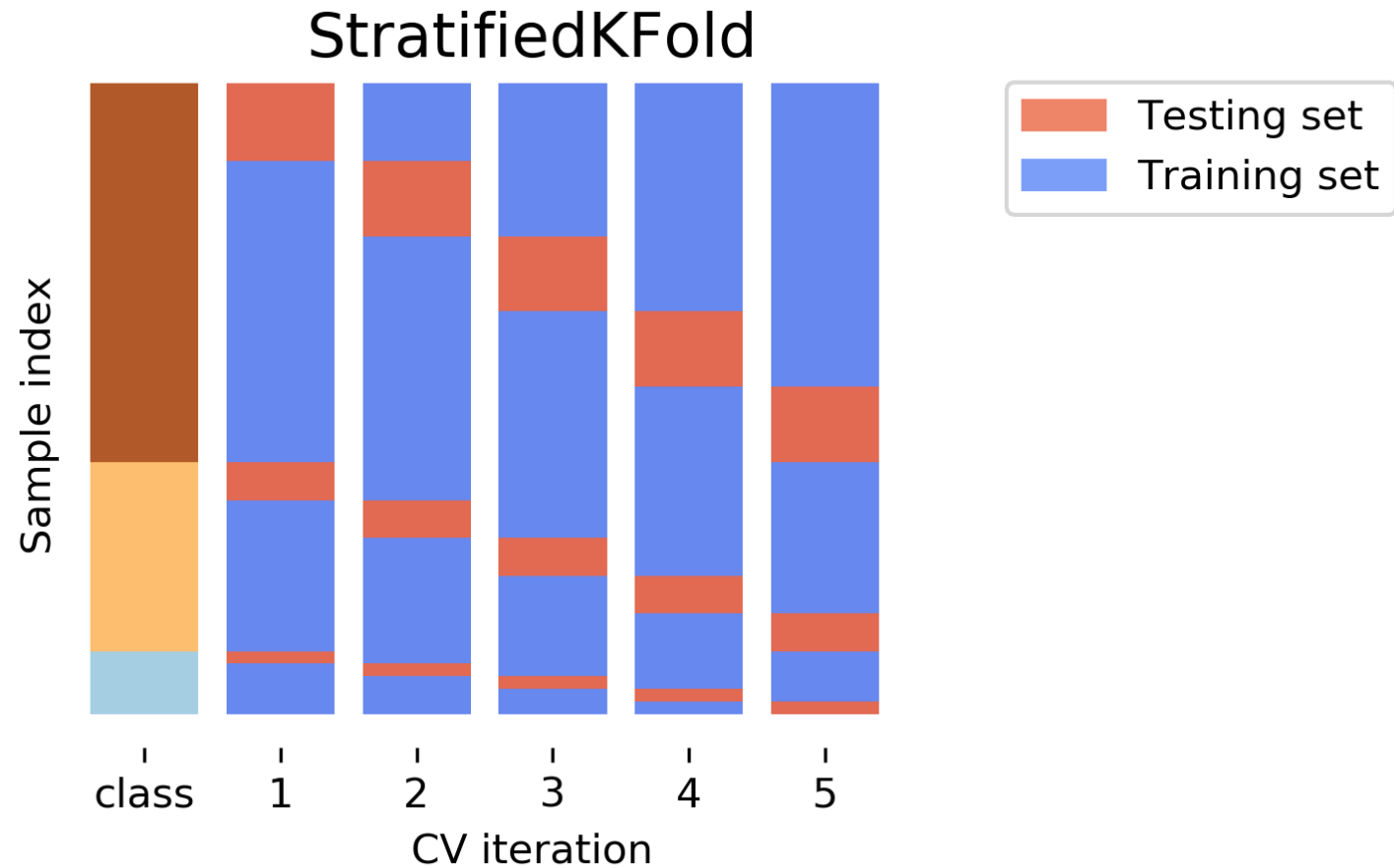
K-fold cross-validation



1. Divide the sample data into k parts.
2. Use k-1 of the parts for training, and 1 for testing.
3. Repeat the procedure k times, rotating the test set.
4. Determine an expected performance metric (mean square error, misclassification error rate, confidence interval, or other appropriate metric) based on the results across the iterations

Stratified K-fold cross-validation

- Needed when the classes are unbalanced
- Aims to maintain data set distribution in each fold.



Choosing K

- Two extreme cases:
 - $K = 2$
 - $K = N$ (Leave-one-out)
- With $K = 2$ there's no correlation between the learned models, while with $K > 2$ there is some overlap between the data used for each model.
- Could you foresee a problem with $K = 2$?

Choosing K

- In terms of accuracy, LOO often results in high variance as an estimator for the test error. Intuitively, since $N - 1$ of the N samples are used to build each model, models constructed from folds are virtually identical to each other and to the model built from the entire training set.
- However, if the learning curve is steep for the training size in question, then 5- or 10- fold cross validation can overestimate the generalization error.
- As a general rule, most authors, and empirical evidence, suggest that 5- or 10- fold cross validation should be preferred to LOO.