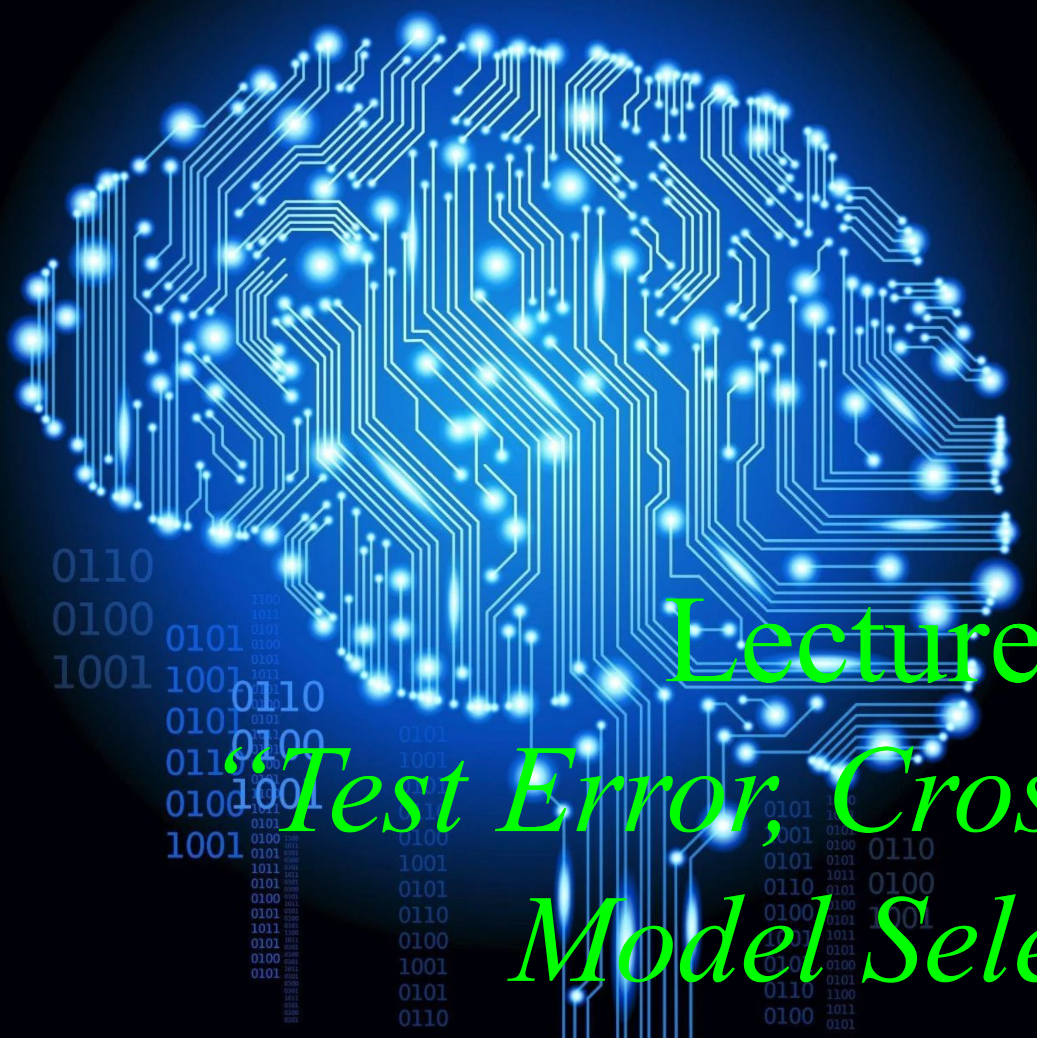
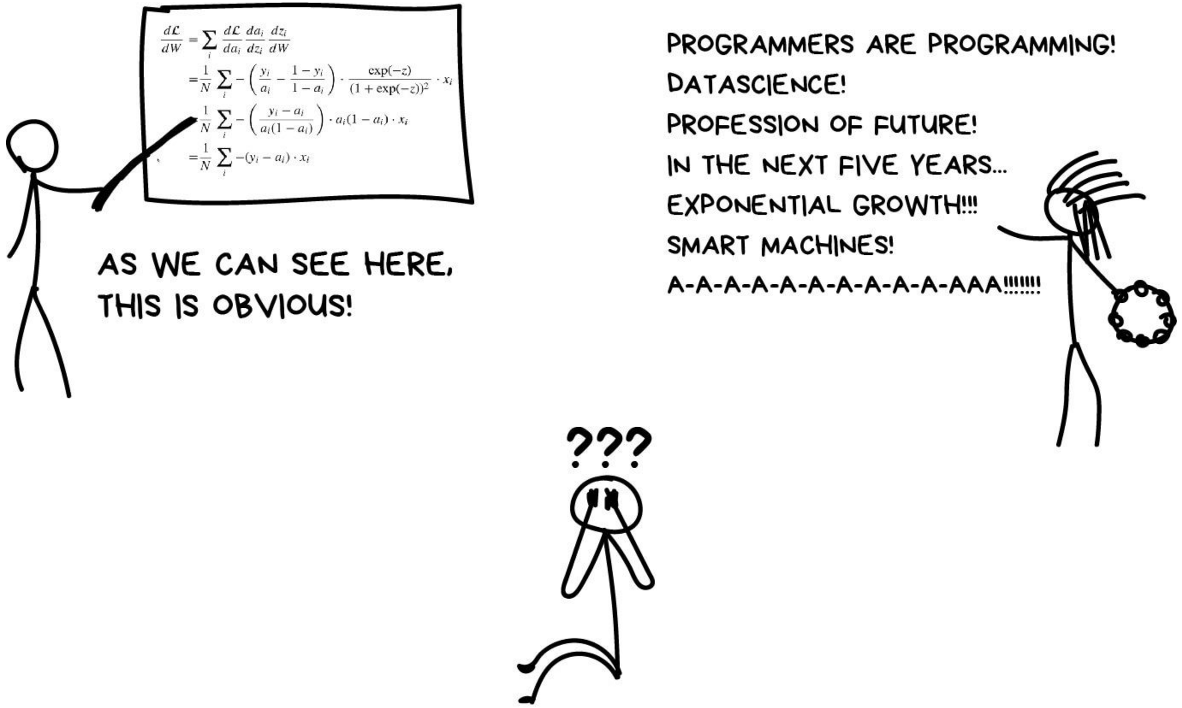


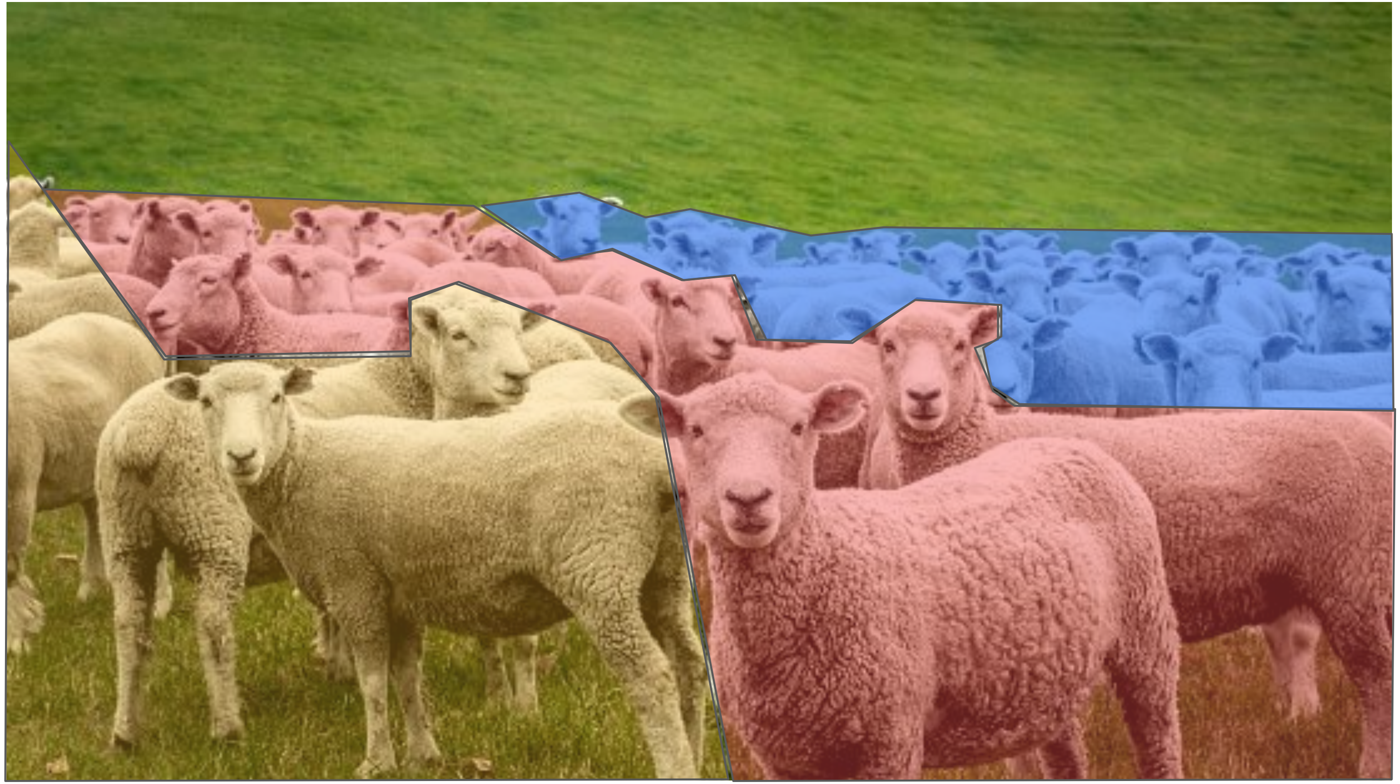
Lecture 05:

“Test Error, Cross-Validation, Model Selection”

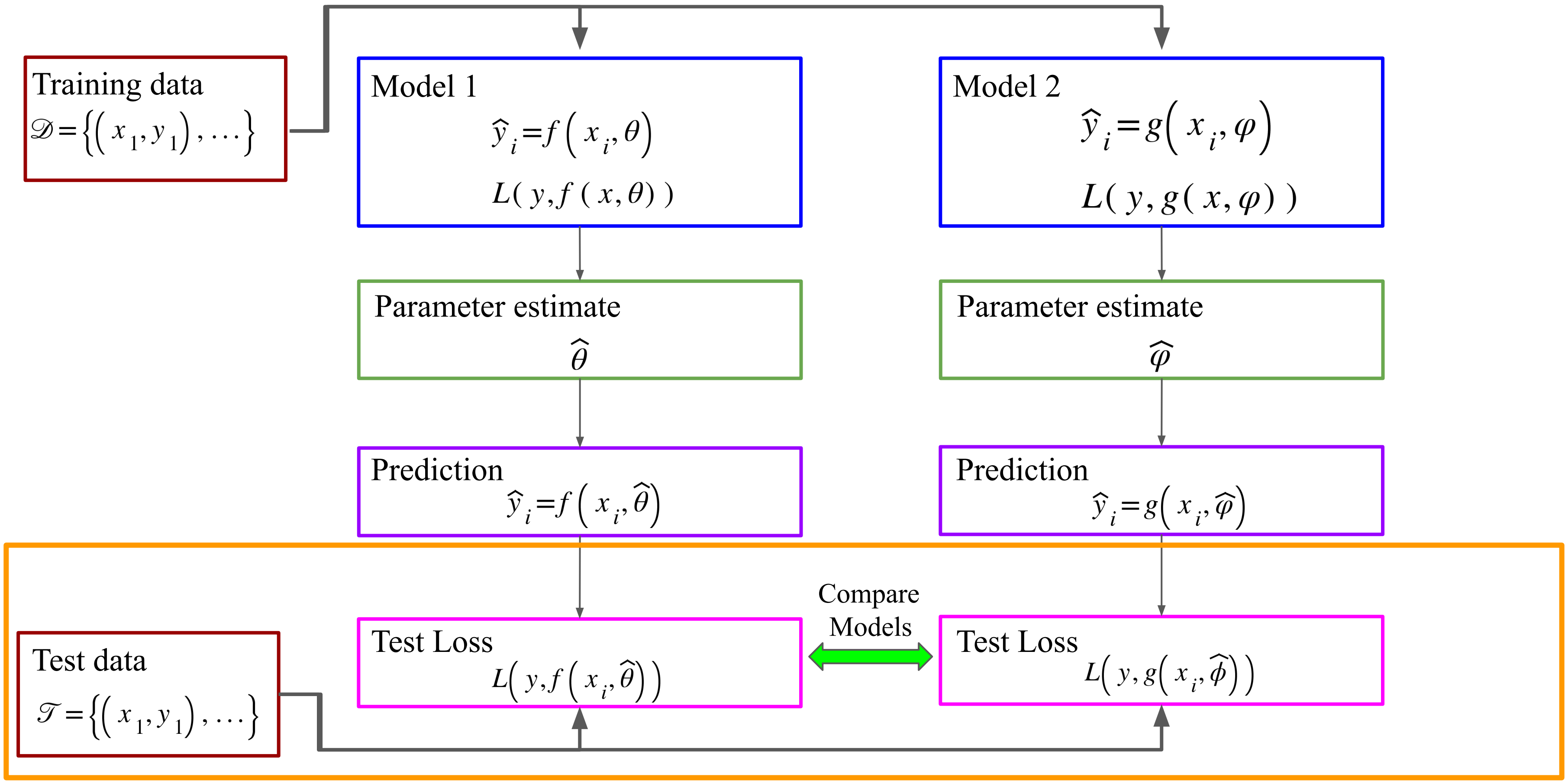
Practical Matters



How to split our data?



Supervised Learning (Recall)



Test Error

Given a dataset (collection of realizations) $\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$ of (X,Y) which were **not** used to train the model, we define the **test error** as the following:

$$\frac{1}{n}\sum_{i=1}^{n} L\left(y_i, f\left(x_i, \hat{\theta}\right)\right)$$

Generalization error (*Conditional test error*) = the expectation of test error over different test sets, given a particular training set.

$$E_{\mathscr{T}}\left(L\left(y_i, f\left(x_i, \hat{\theta}\right)\right) \middle| \mathscr{D}\right)$$

Prediction error (*Expected test error*) = the expectation of test error over different training and test sets.

$$E_{\mathscr{D},\mathscr{T}}\left(L\left(y_i, f\left(x_i, \hat{\theta}\right)\right)\right)$$

Bias-Variance Decomposition

What influences our expected test error? There are **3 factors**:

1. **Bias**: Systematic difference of the best fitted model from the true relationship

$$E\left(\widehat{f}\left(x_i\right)\right) - f\left(x_i\right)$$

2. **Variance** of the fit around the average fit.

$$E\left(\widehat{f}\left(x_i\right) - E\left(\widehat{f}\left(x_i\right)\right)\right)^2$$

3. **Irreducible error:** Variability in data around the true relationship between x and y.

$$y_i = f\left(x_i\right) + \boxed{\varepsilon} \leftarrow \sigma^2_{\varepsilon}$$

Bias-Variance Decomposition

What influences our expected test error? There are **3 factors**:

1 **Bias**: Systematic difference of the best fitted model from the true relationship

$$E\left(\hat{f}(x_i)\right) - f(x_i)$$

2 **Variance** of the fit around the average fit.

$$E\left(\hat{f}(x_i)\right) - E\left(\hat{f}(x_i)\right)^2$$

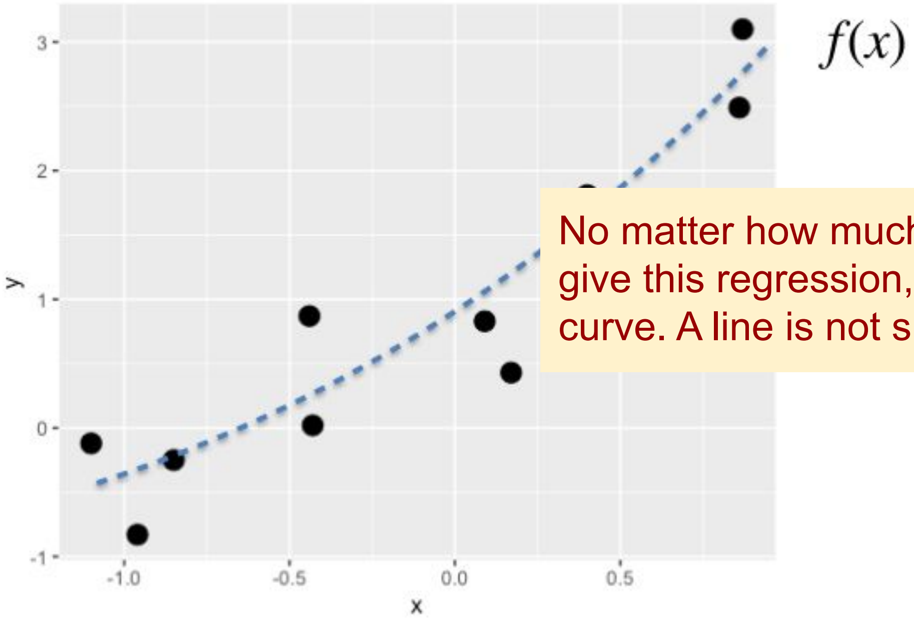
3 **Irreducible error:** Variability in data around the true relationship between x and y.

$$y_i = f(x_i) + \varepsilon \leftarrow \sigma^2_{\varepsilon}$$

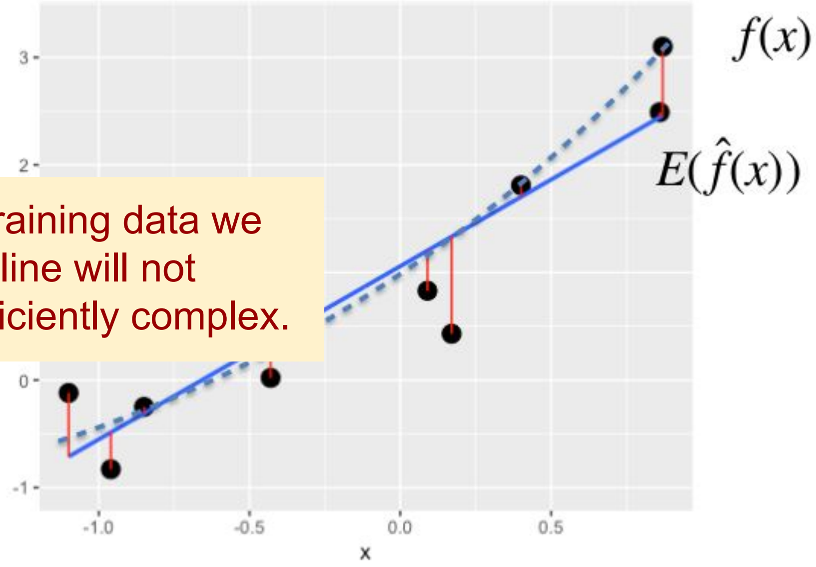
Bias-Variance Decomposition

1 **Bias**: Systematic difference of the best fitted model from the true relationship

$$E\left(\widehat{f}\left(x_i\right)\right) - f\left(x_i\right)$$



No matter how much training data we give this regression, a line will not curve. A line is not sufficiently complex.

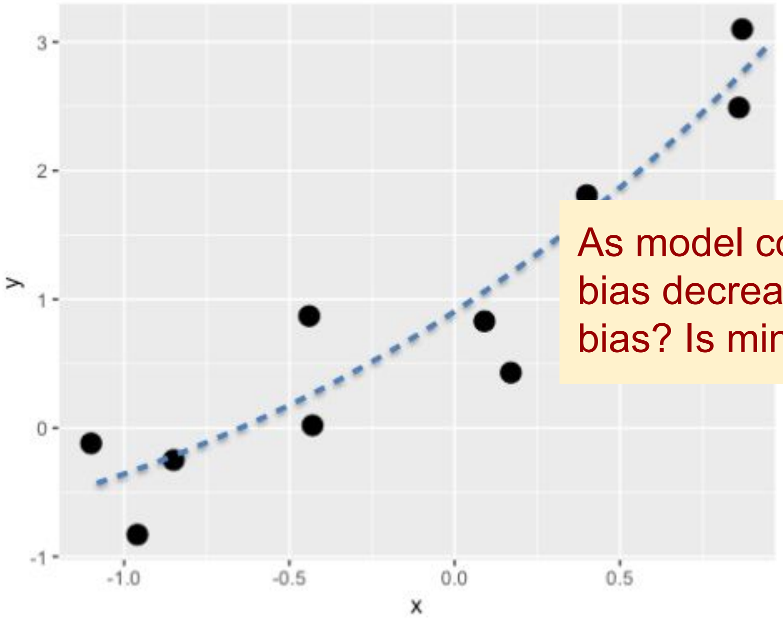


$\hat{y} = 1.1 + 1.6x$

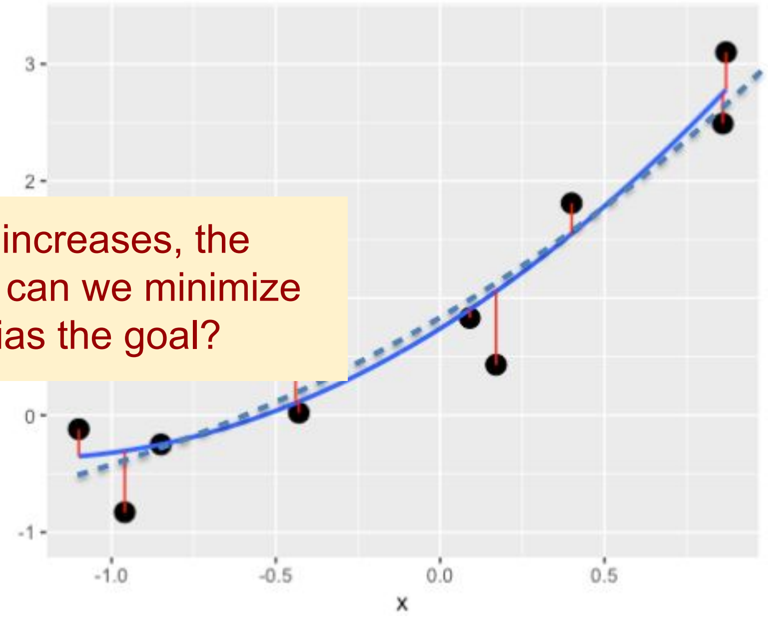
Bias-Variance Decomposition

1 **Bias**: Systematic difference of the best fitted model from the true relationship

$$E\left(\widehat{f}\left(x_i\right)\right) - f\left(x_i\right)$$



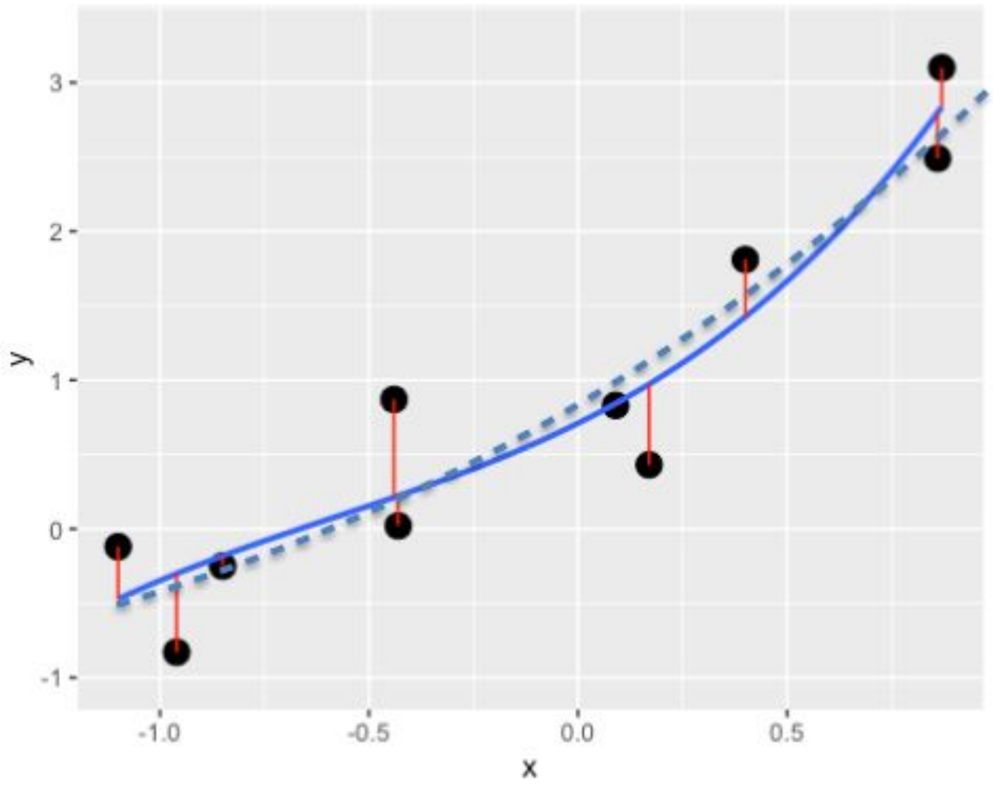
$f(x)$



$\hat{y} = 0.74 + 1.75x + 0.69x^2$

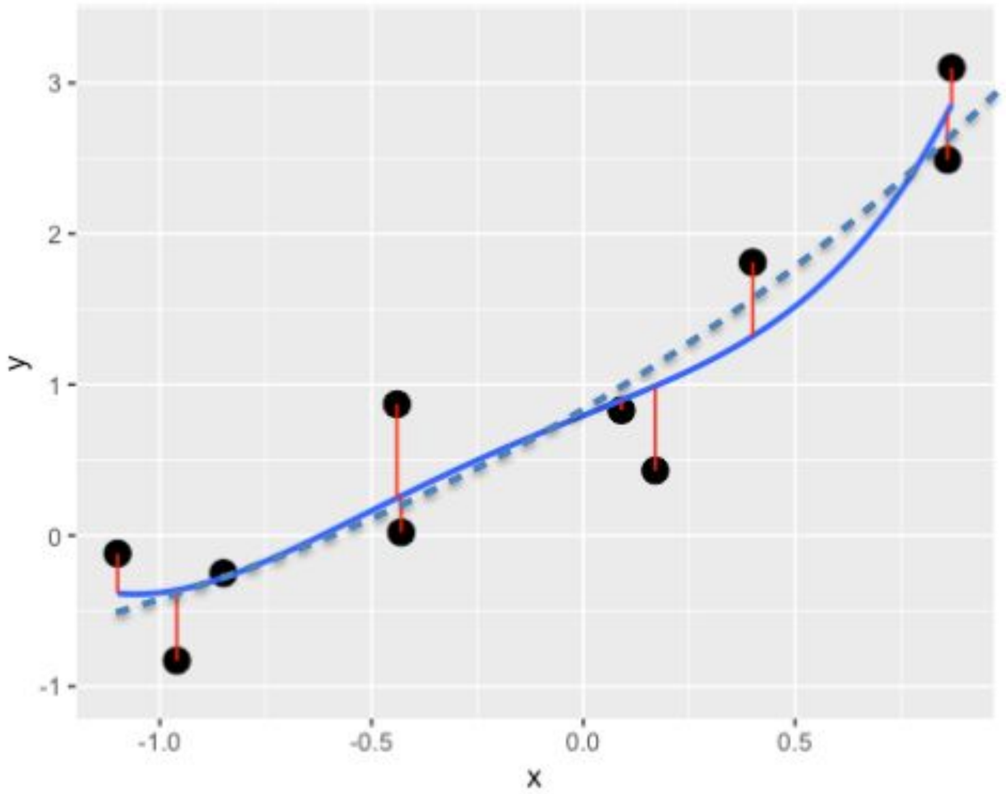
As model complexity increases, the bias decreases. How can we minimize bias? Is minimizing bias the goal?

Order 3 polynomial



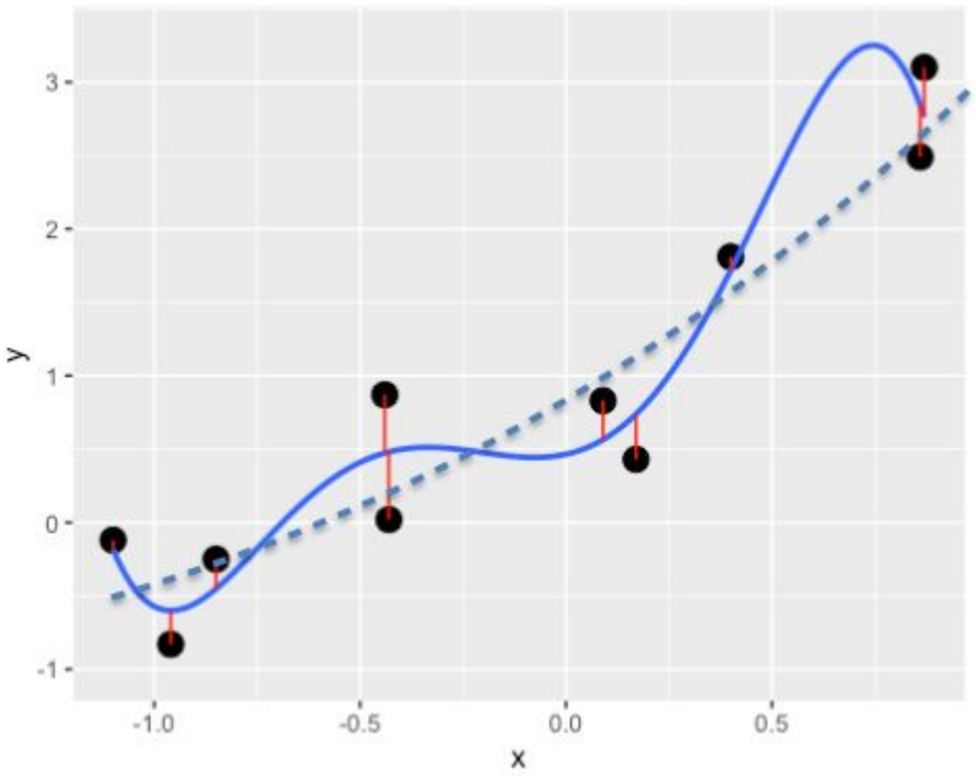
$$\hat{y} = 0.71 + 1.39x + 0.8x^2 + 0.46x^3$$

Order 4 polynomial



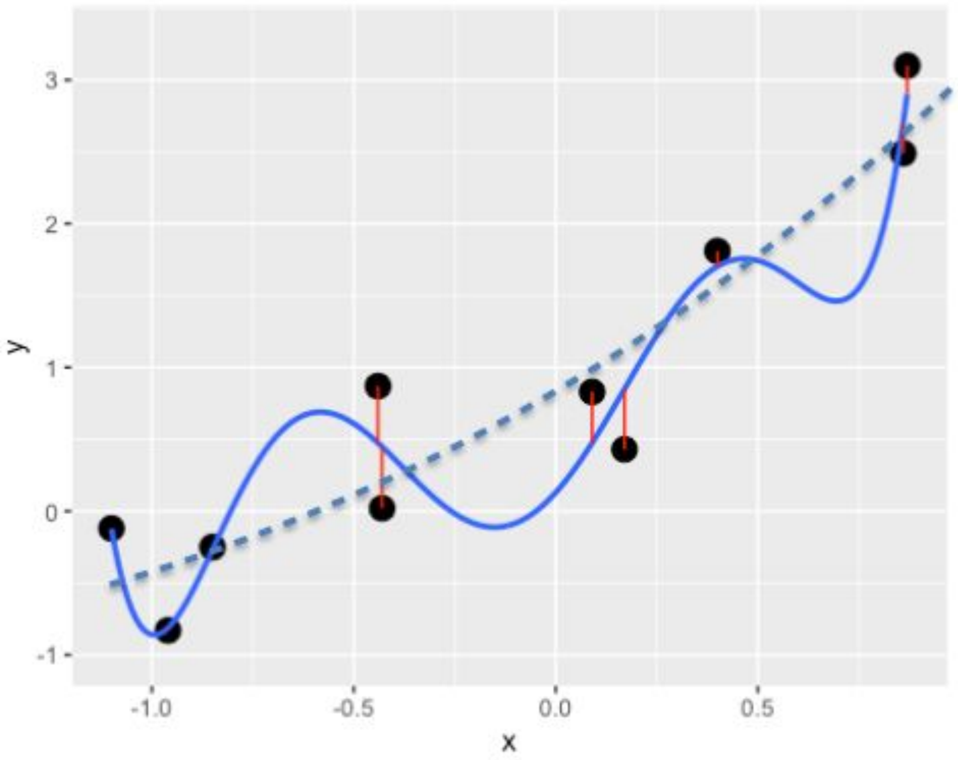
$$\hat{y} = 0.795 + 1.128x - 0.039x^2 + 0.905x^3 + 0.898x^4$$

Order 5 polynomial



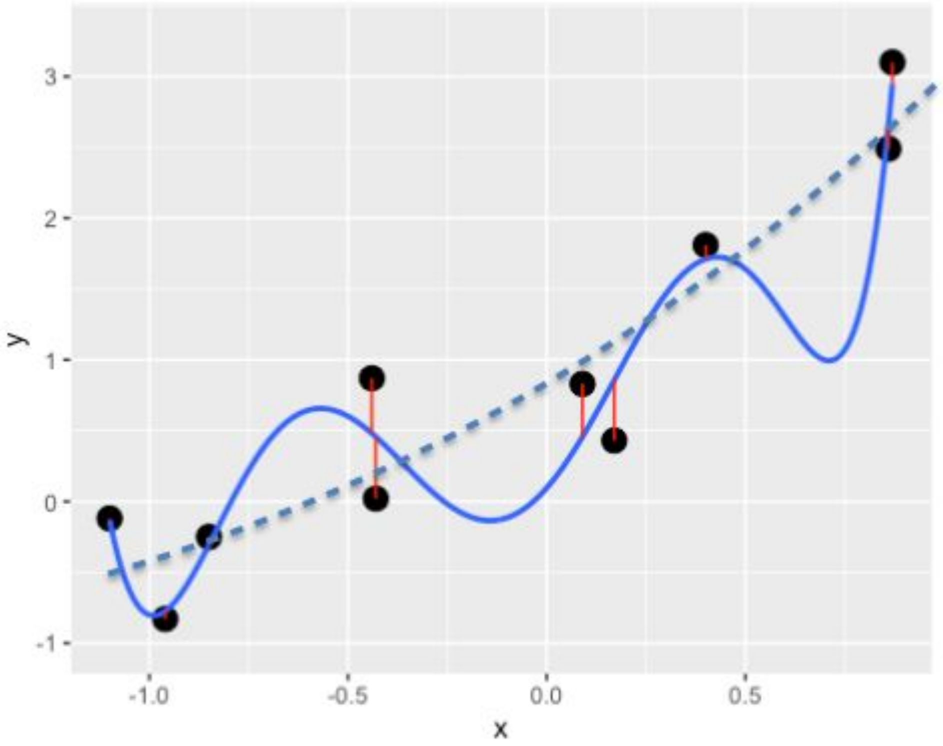
$$\hat{y} = 0.47 + 0.62x + 4.86x^2 + 6.75x^3 - 5.25x^4 - 6.72x^5$$

Order 6 polynomial



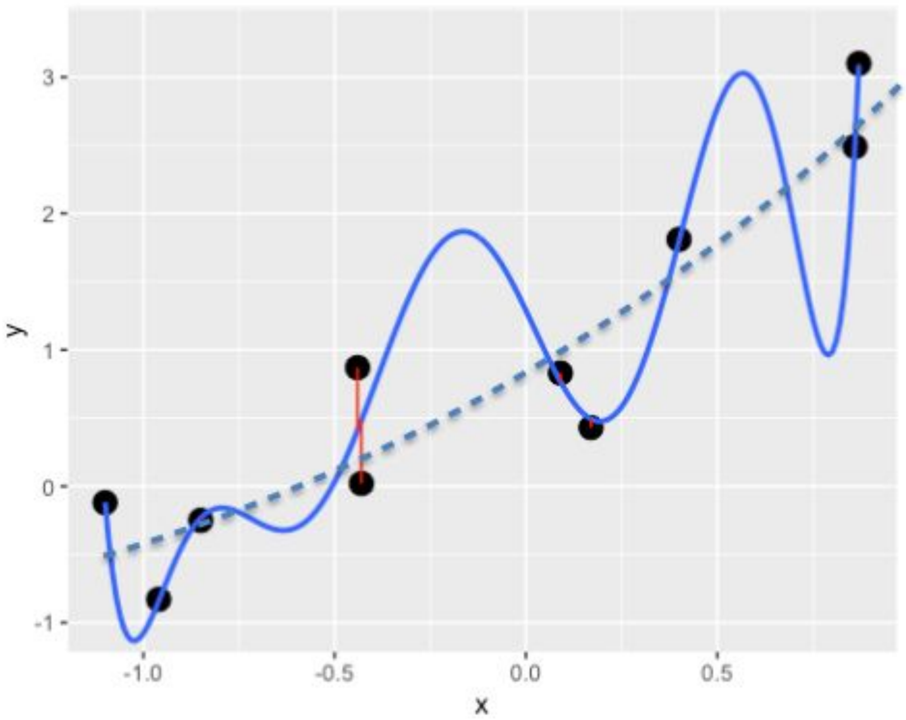
$$\hat{y} = 0.13 + 3.13x + 8.99x^2 - 11.11x^3 - 23.83x^4 + 12.52x^5 + 18.38x^6$$

Order 7 polynomial



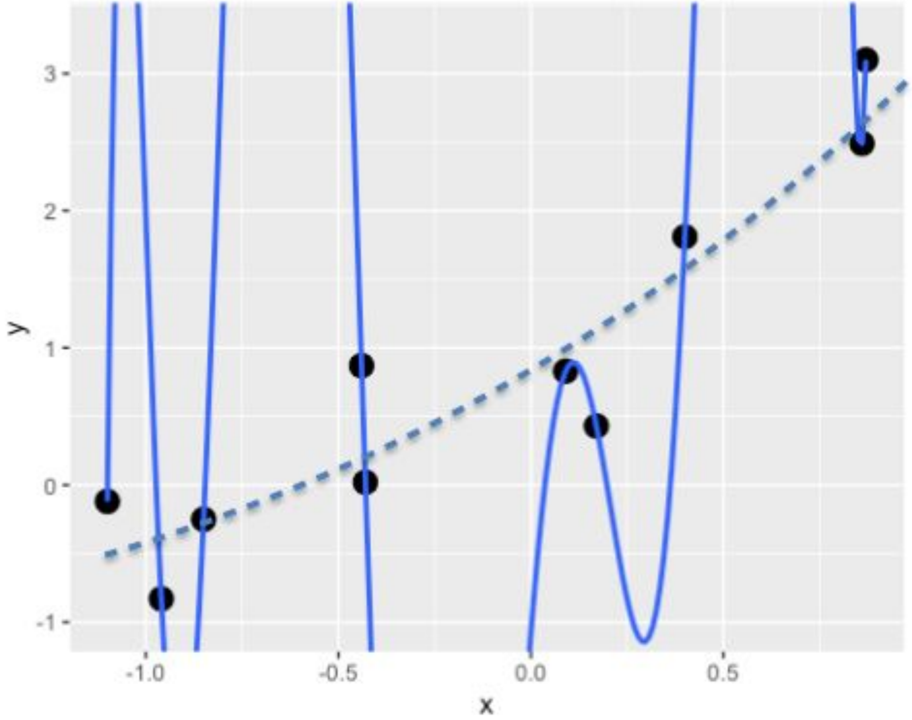
$$\hat{y} = 0.096 + 3.207x + 10.193x^2 - 11.078x^3 - 30.742x^4 + 8.263x^5 + 25.527x^6 + 5.483x^7$$

Order 8 polynomial



$$\hat{y} = 1.3 - 5.9x - 5.1x^2 + 69.9x^3 + 48.8x^4 - 172x^5 - 131.9x^6 + 123.3x^7 + 101.2x^8$$

Order 9 polynomial



$$\hat{y} = -1.1 + 34.8x - 127.9x^2 - 379.9x^3 + 1186.9x^4 + 1604.8x^5 - 2475.4x^6 - 2627.6x^7 + 1499.6x^8 + 1448.1x^9$$

Bias-Variance Decomposition

What influences our expected test error? There are **3 factors**:

1 **Bias**: Systematic difference of the best fitted model from the true relationship

$$E\left(\hat{f}\left(x_i\right)\right) - f\left(x_i\right)$$

2 **Variance** of the fit around the average fit.

$$E\left(\hat{f}\left(x_i\right)\right) - E\left(\hat{f}\left(x_i\right)\right)^2$$

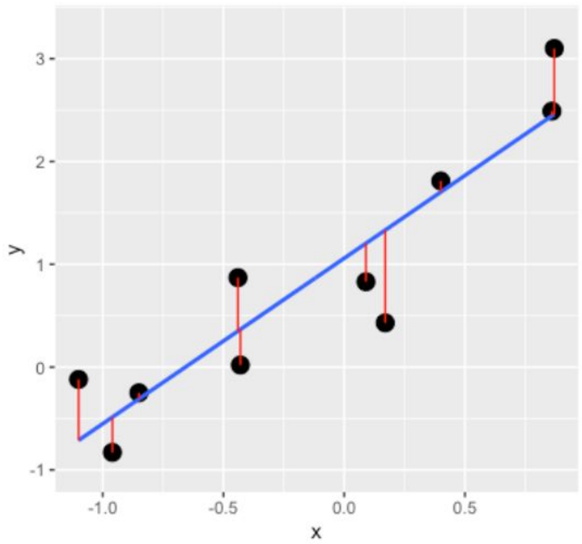
3 **Irreducible error:** Variability in data around the true relationship between x and y.

$$y_i = f\left(x_i\right) + \varepsilon \leftarrow \sigma^2_{\varepsilon}$$

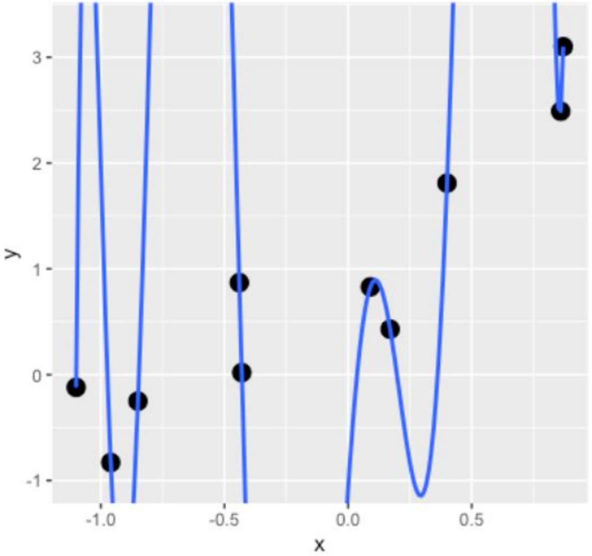
Bias-Variance Decomposition

2 **Variance** of the fit around the average fit.

$$E\left(\hat{f}(x_i) - E\left(\hat{f}(x_i)\right)\right)^2$$



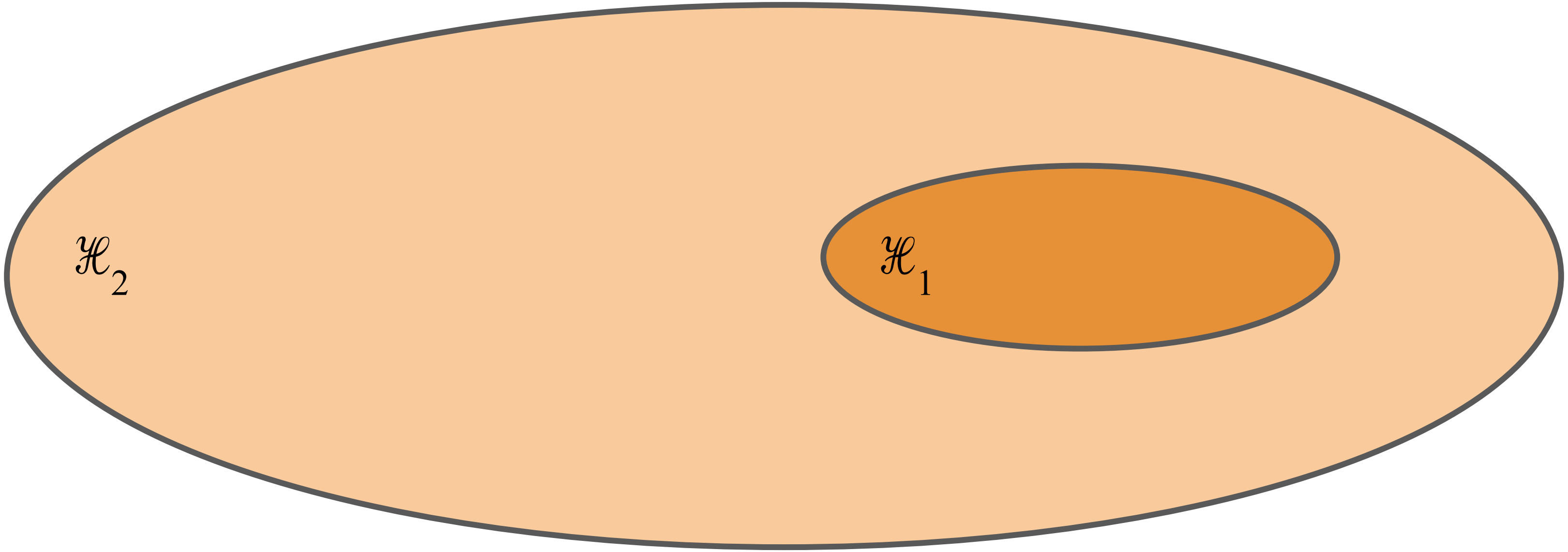
MSE training loss: 0.22



MSE training loss: 0

Model Complexity (Training loss)

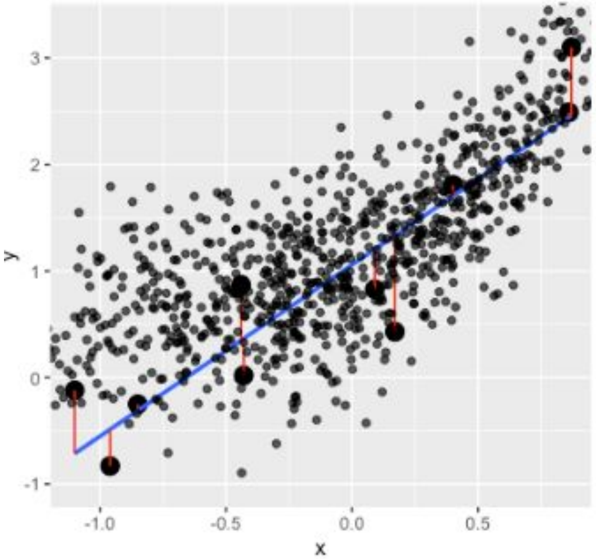
- Larger model spaces lead to lower training loss
- Consider $\mathscr{H}_1$ as the set of all linear functions; consider $\mathscr{H}_2$ as the set of all quadratic functions. We note that $\mathscr{H}_1 \subset \mathscr{H}_2$



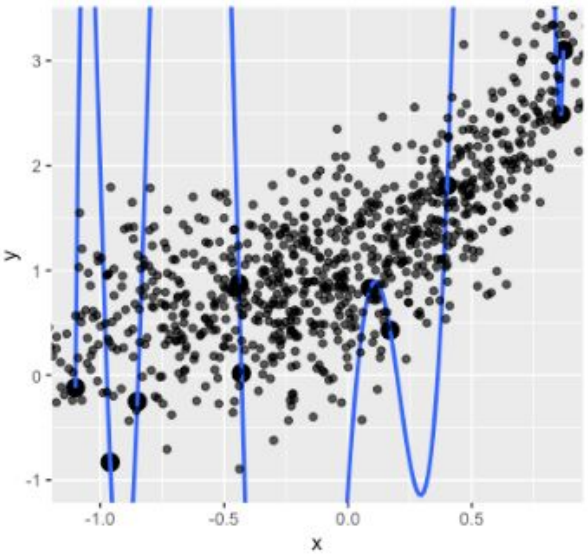
Bias-Variance Decomposition

2 **Variance** of the fit around the average fit.

$$E\left(\widehat{f}(x_i) - E\left(\widehat{f}(x_i)\right)\right)^2$$

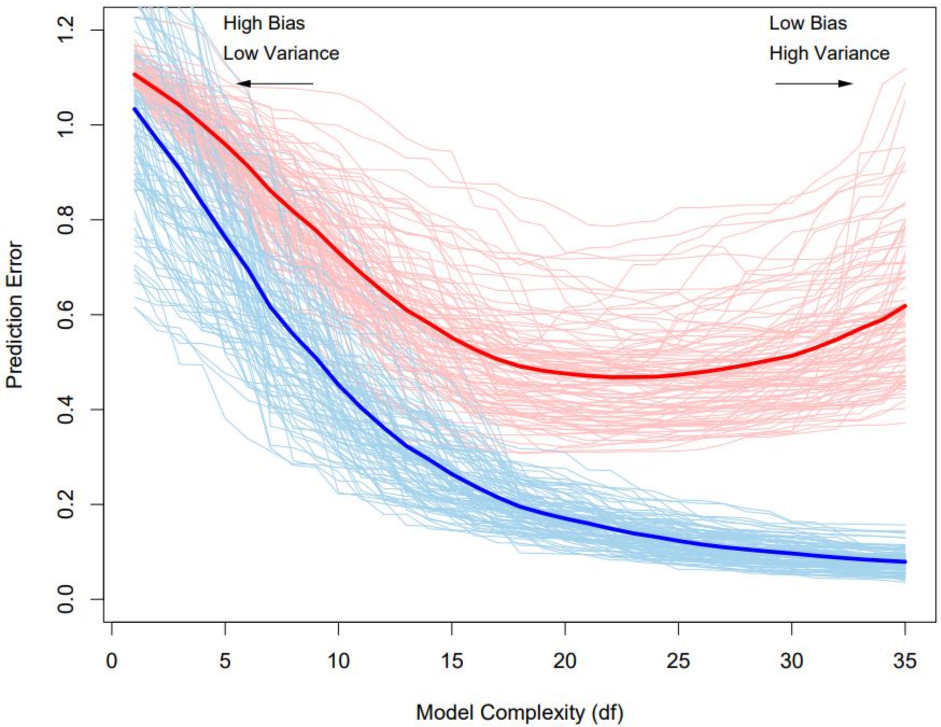


MSE training loss: 0.22
MSE test error: 1.36



MSE training loss: 0
MSE test error: 87422896497

Bias-Variance Tradeoff



- **<u>Overfitting</u>** → the *training error* is **decreasing**, but the *test error* is **increasing**

Bias-Variance Decomposition

What influences our expected test error? There are **3 factors**:

1 **Bias**: Systematic difference of the best fitted model from the true relationship

$$E\left(\widehat{f}\left(x_i\right)\right) - f\left(x_i\right)$$

2 **Variance** of the fit around the average fit.

$$E\left(\widehat{f}\left(x_i\right)\right) - E\left(\widehat{f}\left(x_i\right)\right)^2$$

3 **Irreducible error:** Variability in data around the true relationship between x and y.

$$y_i = f\left(x_i\right) + \varepsilon \leftarrow \sigma^2_{\varepsilon}$$

Bias-Variance Decomposition

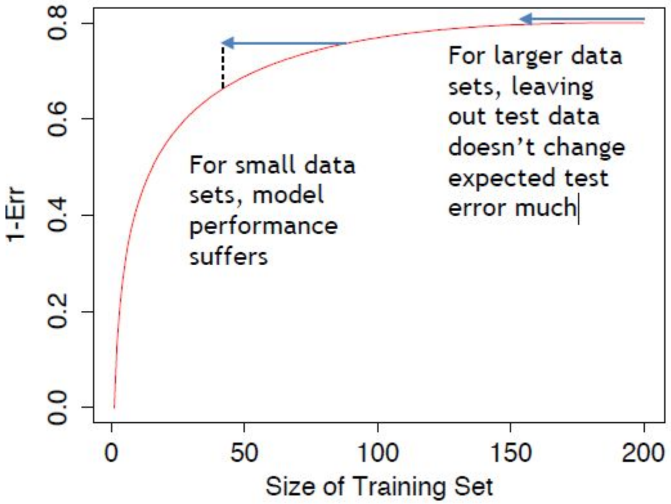
3 **Irreducible error:** Variability in data around the true relationship between x and y.

$$y_i = f(x_i) + \boxed{\varepsilon} \leftarrow \sigma^2_{\varepsilon}$$

- Assume that we have the correct model class (i.e. say that it is a linear model)
- Then we can correctly predict outputs from inputs.. right?
- Actually, there are variables outside of X which can have some effect on Y (i.e. noise). In other words, there will be a part of Y which is determined by unobserved phenomena. Even if we had infinite (X,Y) data, we still could not completely determine Y from X.
- Often when we overfit, we are actually fitting our model to noise.

How big should the test set be?

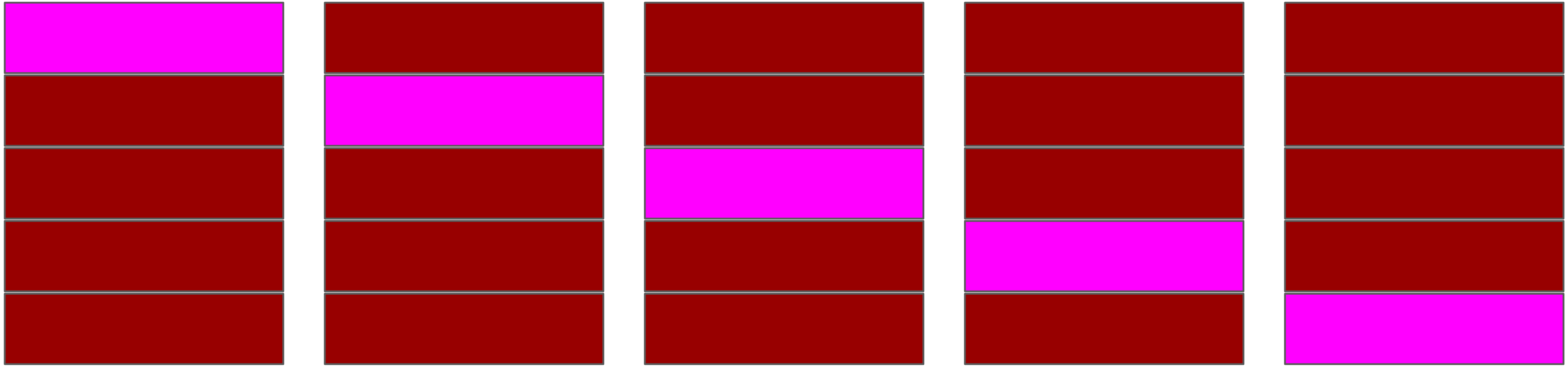
- The test set should be large enough to detect differences in test errors
- The test set should be small enough such that data is left for training (model fitting)



Our Approach So Far...

- This approach is called "Hold-out". We "hold-out" the test set.
- Why it's good:
  - It measures what we want (performance of learned model)
  - It's simple
- Why it's bad:
  - Smaller training sets can lead to variable performance and performance estimates; they can also lead to favoring simpler models
  - Smaller test sets can give poor estimates of performance

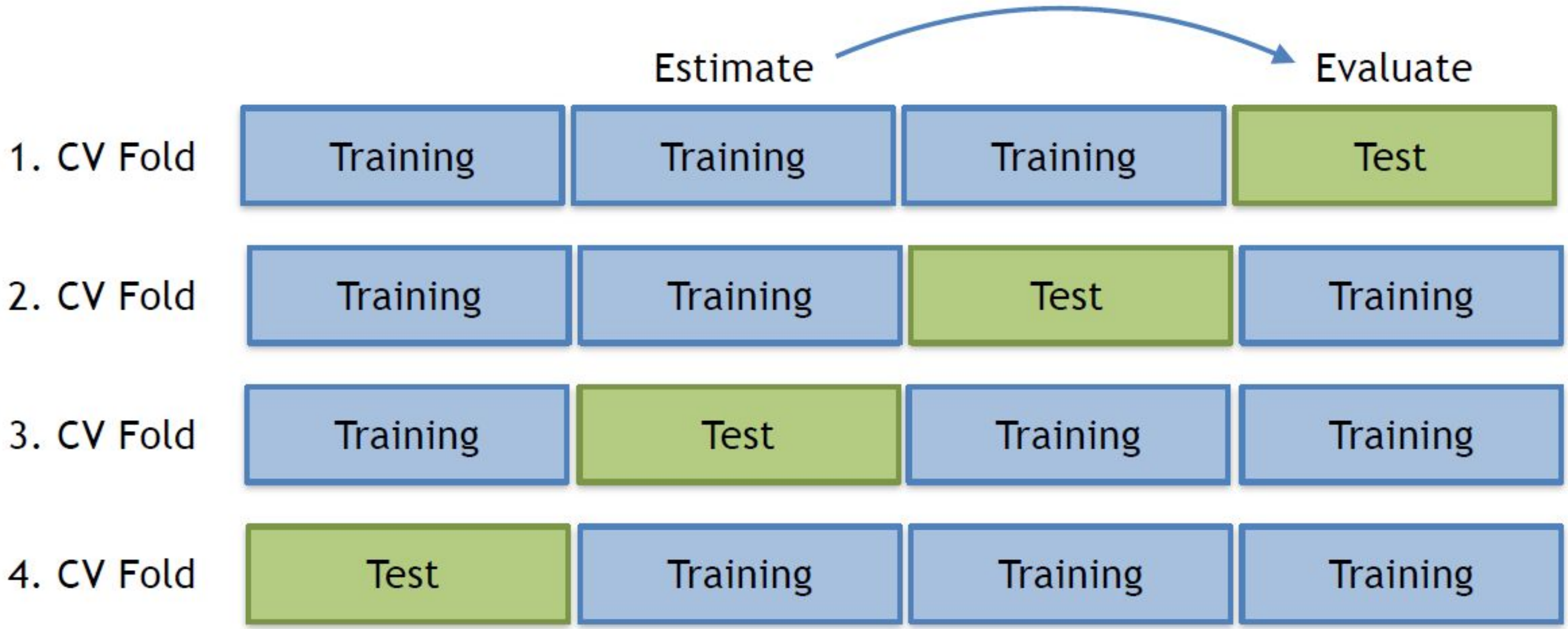
k-fold Cross-Validation



The idea:

1. Split the data into k disjoint partitions (or “folds” of size n/k)
2. For i in range(1,k), train/test with *partition i* as the test set and with the remaining data as the training set
3. Compute the average test error across all test results = **Cross-validation error**. This error has lower variance than error on one partition.

4-fold CV



- In the end, we average the test error across all folds. Each error will be slightly different.

Common Regression Errors

Mean-squared error (MSE)

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

Root-mean squared error (RMSE)

$$RMSE = \sqrt{\left(\frac{1}{n}\right)\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2}$$

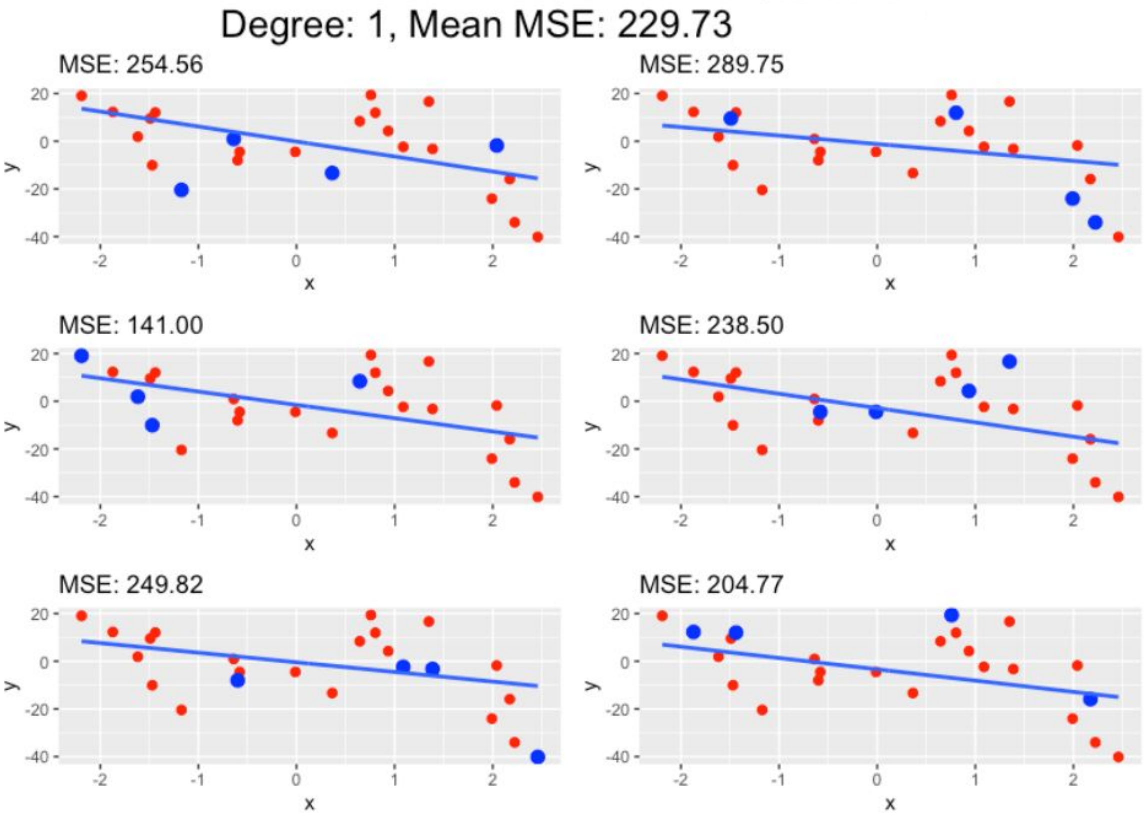
Mean absolute error (MAE)

$$MAE = \left(\frac{1}{n}\right)\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right|$$

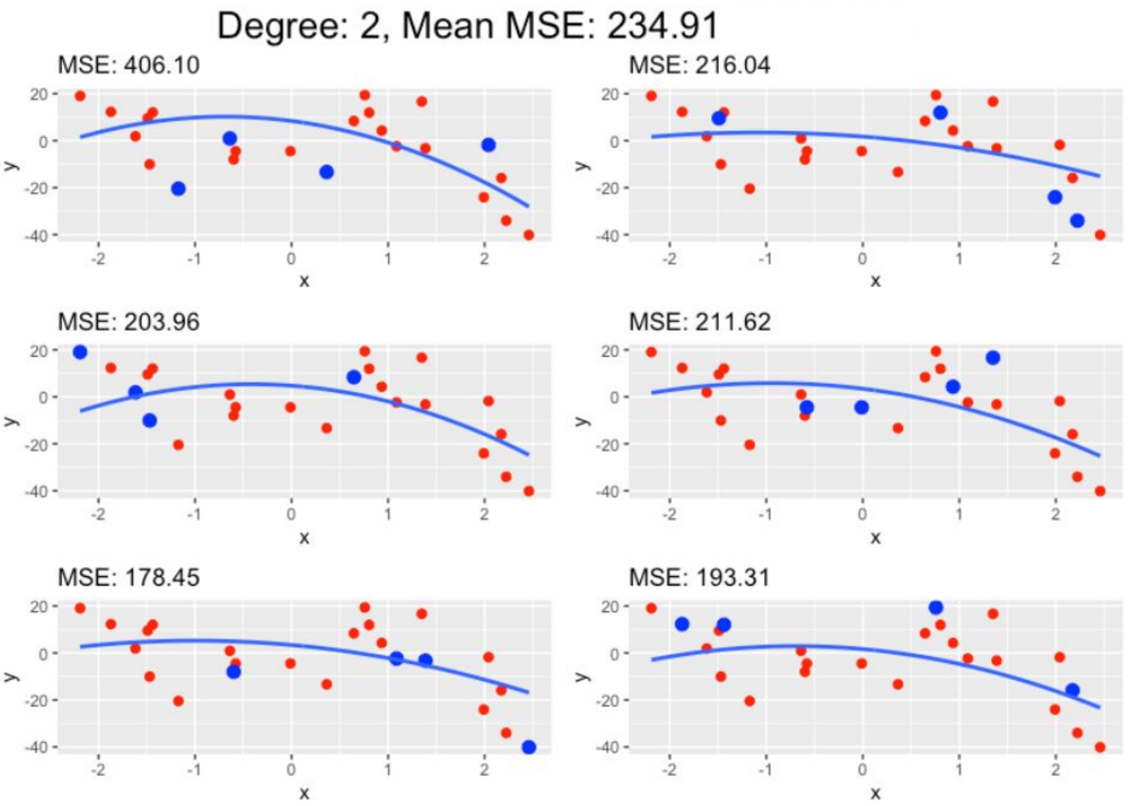
Mean relative error (MRE)

$$MRE = \left(\frac{1}{n}\right)\sum_{i=1}^{n}\frac{\left|\hat{y}_i - y_i\right|}{\left|y_i\right|}$$

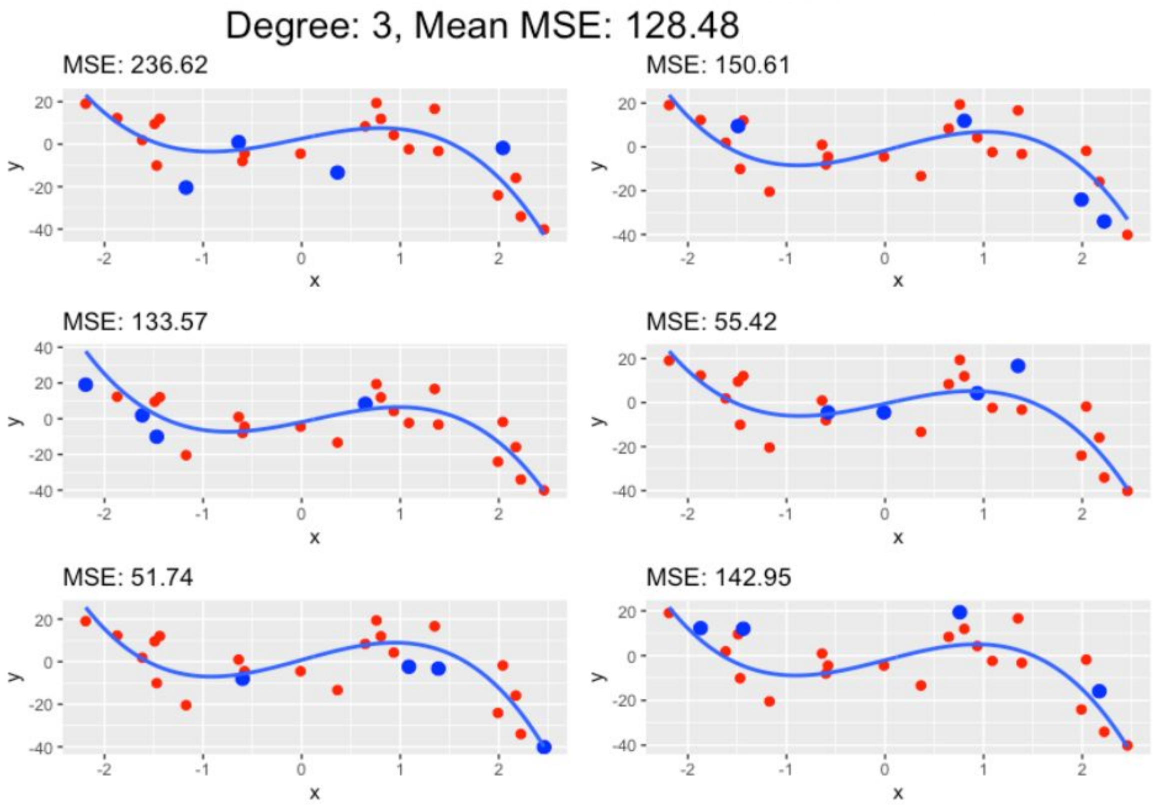
CV for different models



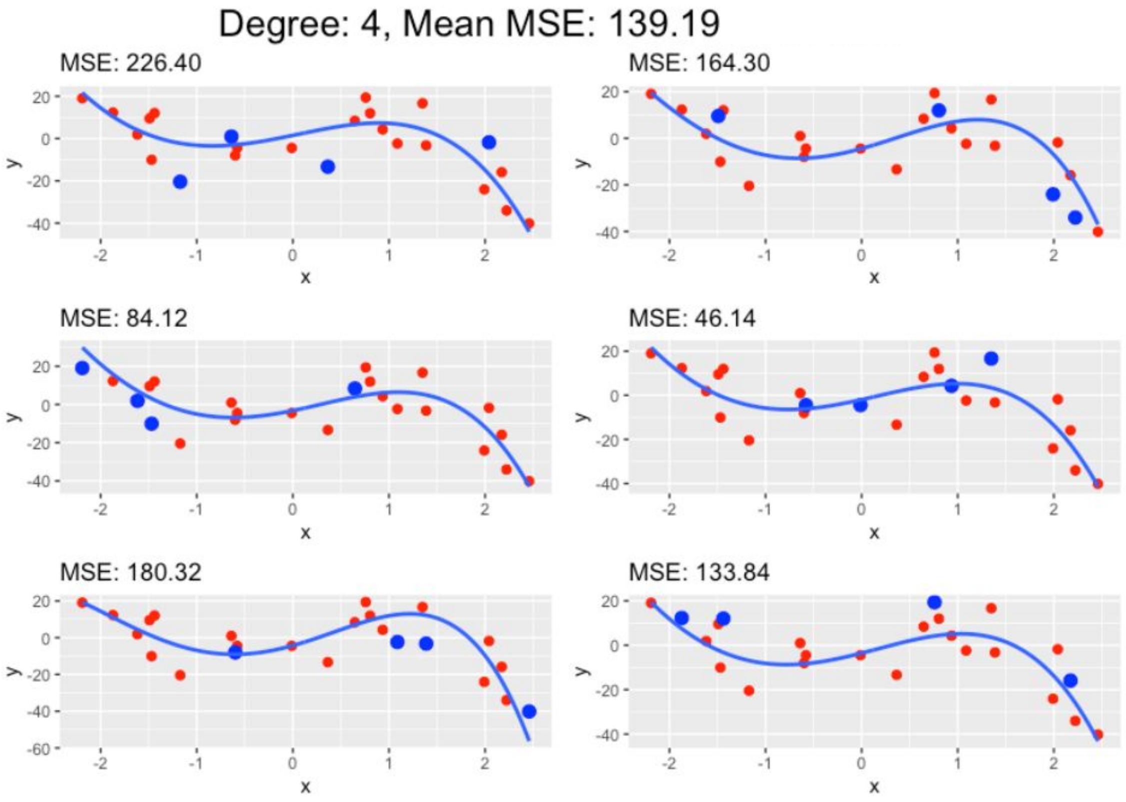
CV for different models



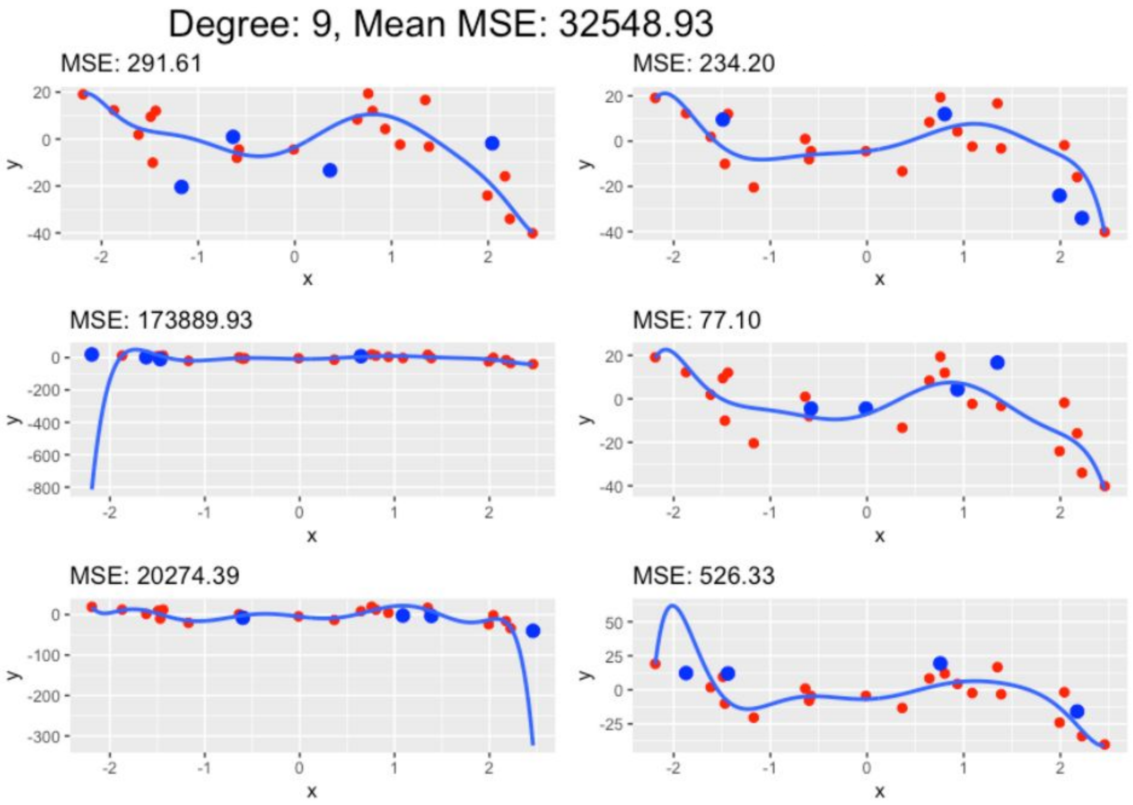
CV for different models



CV for different models



CV for different models



How many CV folds?

	2-Fold CV (split-half)	K-Fold CV (often K=5-10)	N-Fold CV Leave-One-out
Overestimation bias of prediction error:	bad	present	nearly unbiased
Computational cost:	low	favourable	high
Variance of estimate:	low	low	high
Training sets are:	independent	similar	nearly identical

How many CV folds?

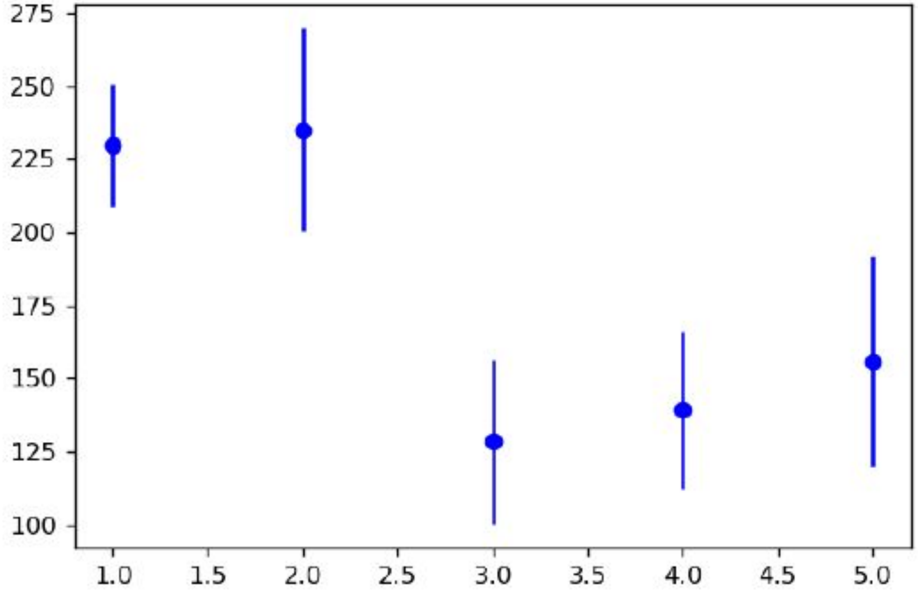
	2-Fold CV (split-half)	K-Fold CV (often K=5-10)	N-Fold CV Leave-One-out
Overestimation bias of prediction error:	bad	present	nearly unbiased
Variance of estimate:	low	low	high

https://jmlr.org/papers/volume11/cawley10a/cawley10a.pdf

https://stats.stackexchange.com/questions/61783/bias-and-variance-in-leave-one-out-vs-k-fold-cross-validation

Model Selection

<u>Which model is the best?</u>

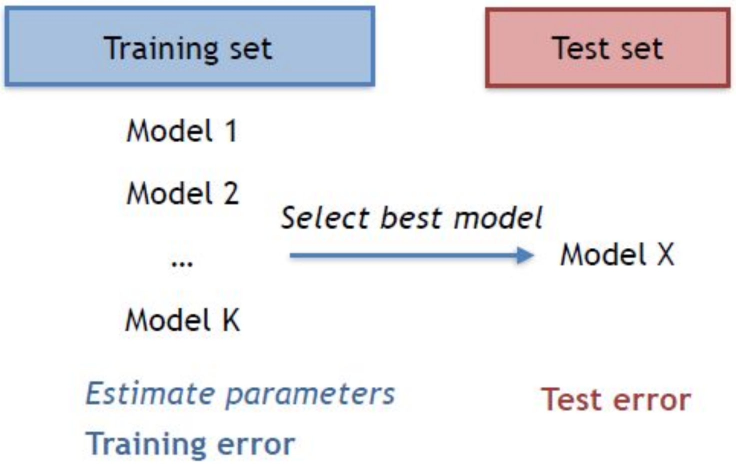


Convention: Select the simplest model with error no more than one stderr. from the best model

Model Selection (Strategy 1)

There are 3 strategies for model selection. In the next few slides we will consider each:

Strategy 1: Choose the model which **fits best to the training data.**



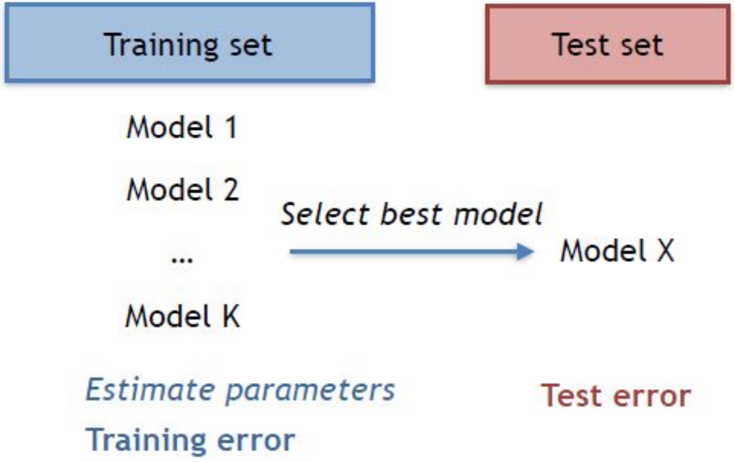
Can you think of an issue with this strategy?

One issue: If we pick the model which best fits to training data only, we will select the most complex model (recall bias-variance tradeoff)... this will lead to **overfitting**.

Model Selection (Strategy 1)

There are 3 strategies for model selection. In the next few slides we will consider each:

Strategy 1: Choose the model which **fits best to the training data.**



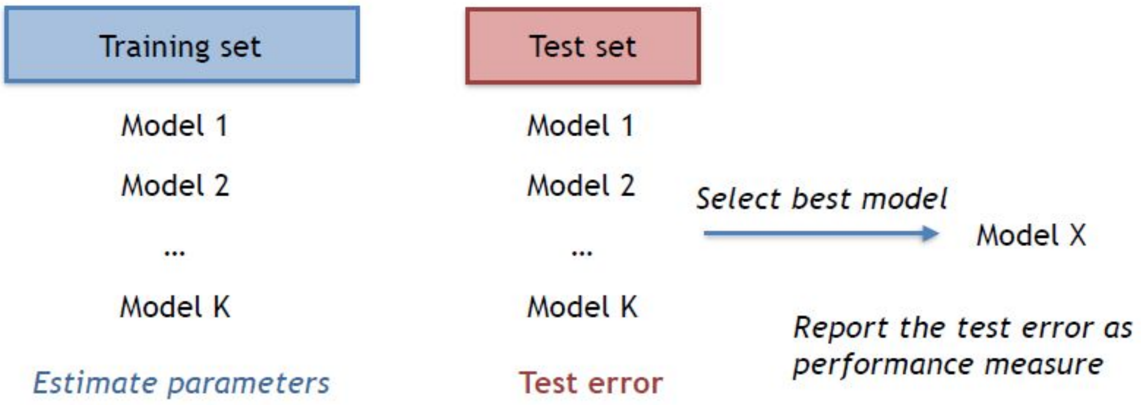
For simple models, if we adjust the training error upwards we can get a less-biased generalization error estimate: AIC, BIC, etc.

When there is limited data, this method may also be preferable.

Model Selection (Strategy 2)

There are 3 strategies for model selection. In the next few slides we will consider each:

Strategy 2: Choose the model which **has the lowest test error.**



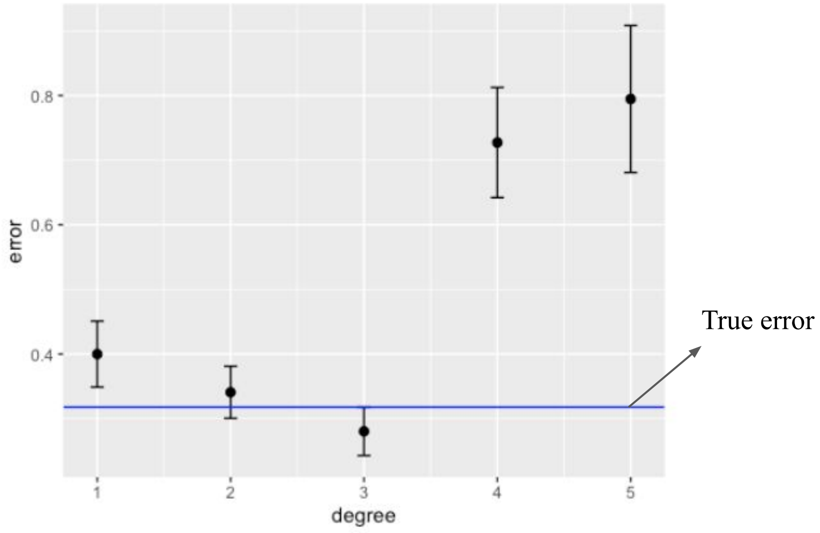
Can you think of an issue with this strategy?

One issue: We underestimate the test error. This leads to another form of “overfitting”

Model Selection (Strategy 2)

Strategy 2: Choose the model which **has the lowest test error.**

Let's say that we generate a dataset, split it into train and test sets + compute test performance OR use cross-validation. We then select the "best" model which has the lowest expected test error.



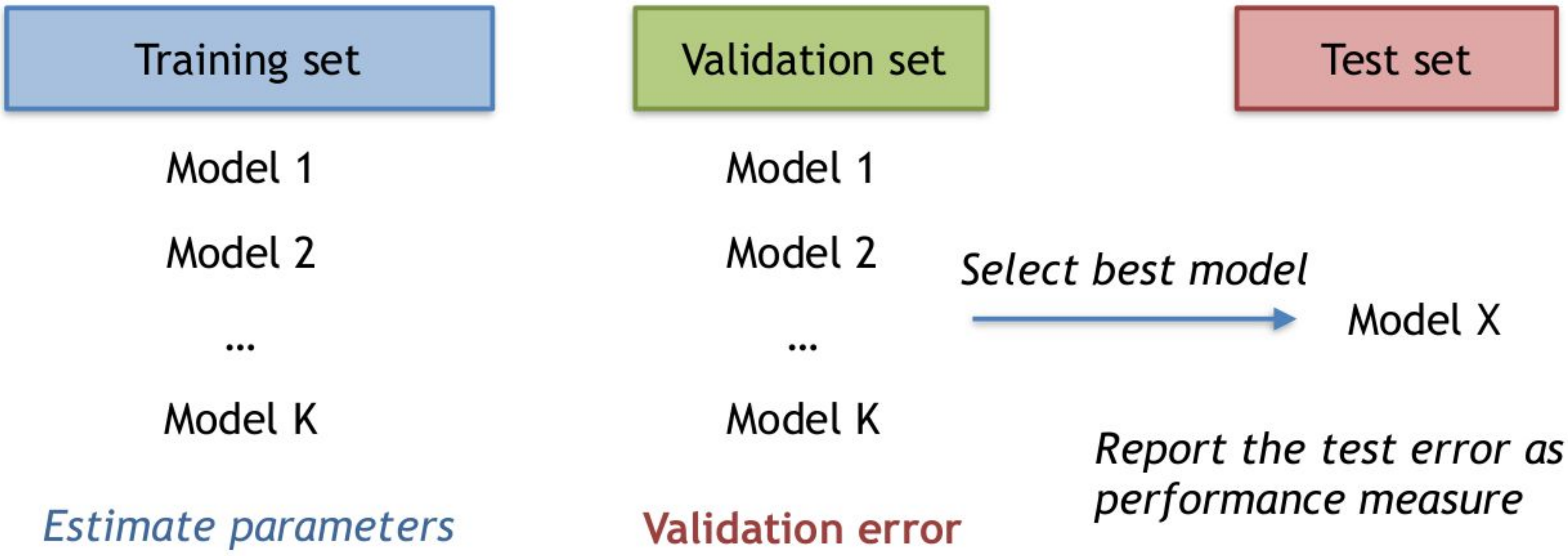
Let's say that we then perform feature selection (i.e. find best combo of features):

- (ESLII pp. 245) 100-feature dataset
- One possible pitfall: Select best features on all data, calculate CV/test error for each model
- Another possible pitfall: Select best features on CV/test error
- Both can lead to dramatic underestimated prediction error.

Model Selection (Strategy 3)

There are 3 strategies for model selection. In the next few slides we will consider each:

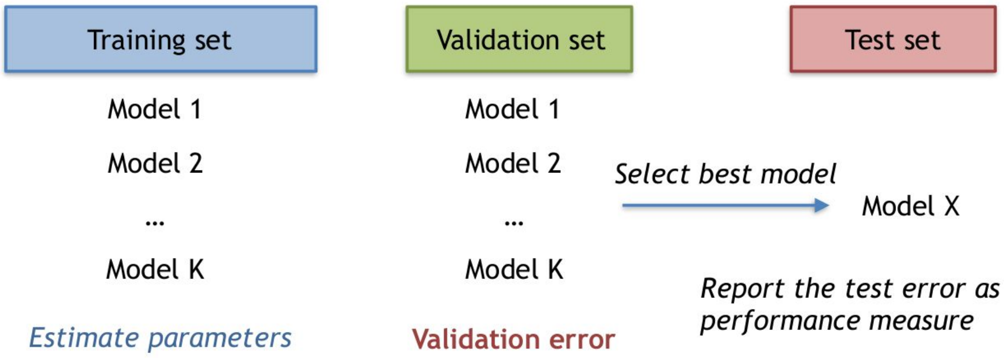
Strategy 3: Choose the model which **has the lowest validation error.**



Model Selection (Strategy 3)

There are 3 strategies for model selection. In the next few slides we will consider each:

Strategy 3: Choose the model which **has the lowest validation error.**

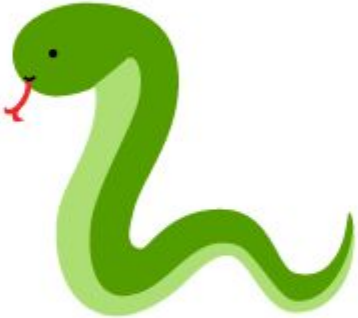


We partition the data into three disjoint subsets:

1. **Training set** to find parameters (θ)
2. **Validation set** to find right model space (i.e. degree of polynomial) → can call this decision another parameter (η)
3. **Test set** to estimate generalization error of a model $M(\eta,\theta)$

- In practice, we often use CV to select the best model → we use all possible validation sets for each model

Let’sss try it in Python...



Summary

- Test Error
- Bias and Variance
  - Bias-variance tradeoff
  - Underfitting, Overfitting
- Choosing a test size
- Cross-Validation
  - Choosing number of folds
- Model Selection
  - 3 Strategies for selecting best model