ENSF 611 – Machine Learning for Software Engineers

Week 3 – Model Selection

Lecture Goals

UNIVERSITY OF CALGARY

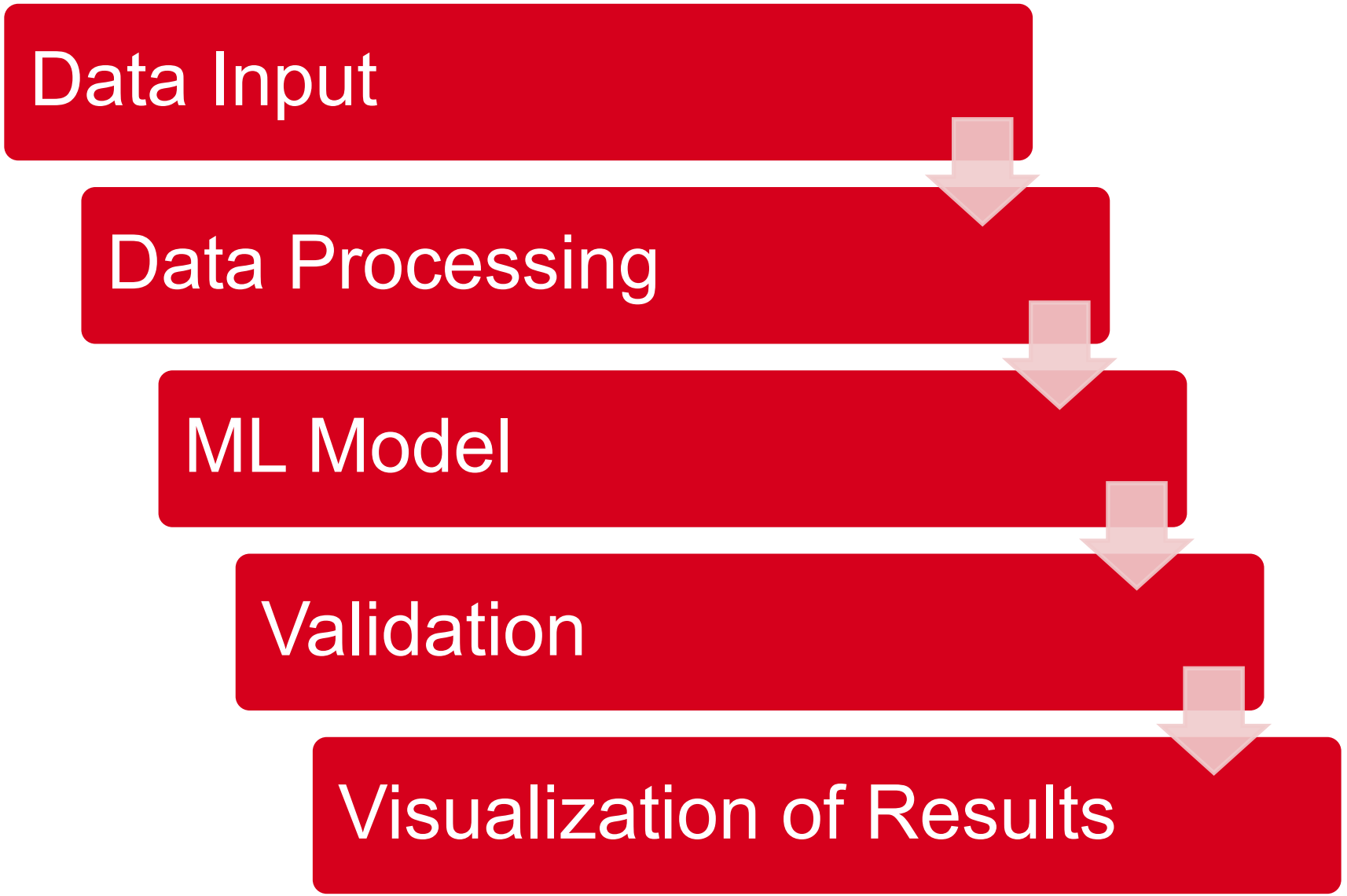
- Introduce the concept of cross-validation
- Discuss the bias-variance tradeoff

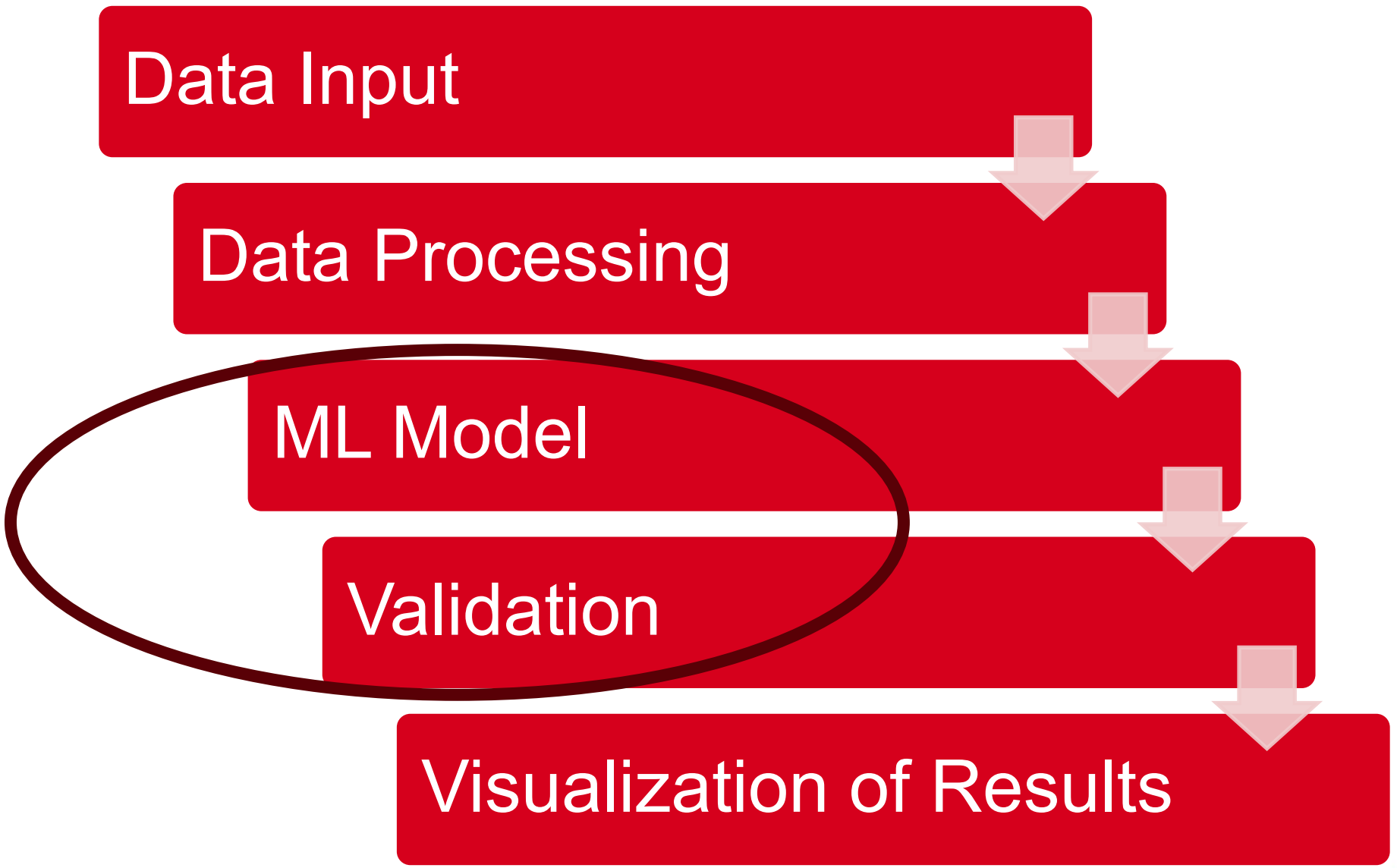
- Chapter 5 of Python Data Science Handbook

Review

- Last week we…
  - Introduced the machine learning workflow
  - Demonstrated how to read in data and clean it before using in a machine learning model (fill in missing/erroneous values)
  - Discussed the need for splitting the data into training and testing sets
  - Demonstrated how to implement a basic machine learning model
  - Introduced basic accuracy metrics

ML Workflow


Data Input
Data Processing
ML Model
Validation
Visualization of Results


Data Input
Data Processing
ML Model
Validation
Visualization of Results

Model Validation

Validating your model

- Your model performance depends on how you select your training and testing sets
- In the simple machine learning example that we did last week, we only calculated the accuracy score for one split of the data (using train_test_split)
  - Note that the default split is 75% training, 25% testing
- What if the way the data was split positively or negatively affected the accuracy?

Cross-validation

UNIVERSITY OF CALGARY

- Can use cross-validation to validate the model performance with the full dataset
  - Repeat validation process multiple times to validate the model and the selected hyperparameters
- To test the model five times, use the following code:

```
from sklearn.cross_validation import cross_val_score
cross_val_score(model, X, y, cv=5)
```

Cross-validation example

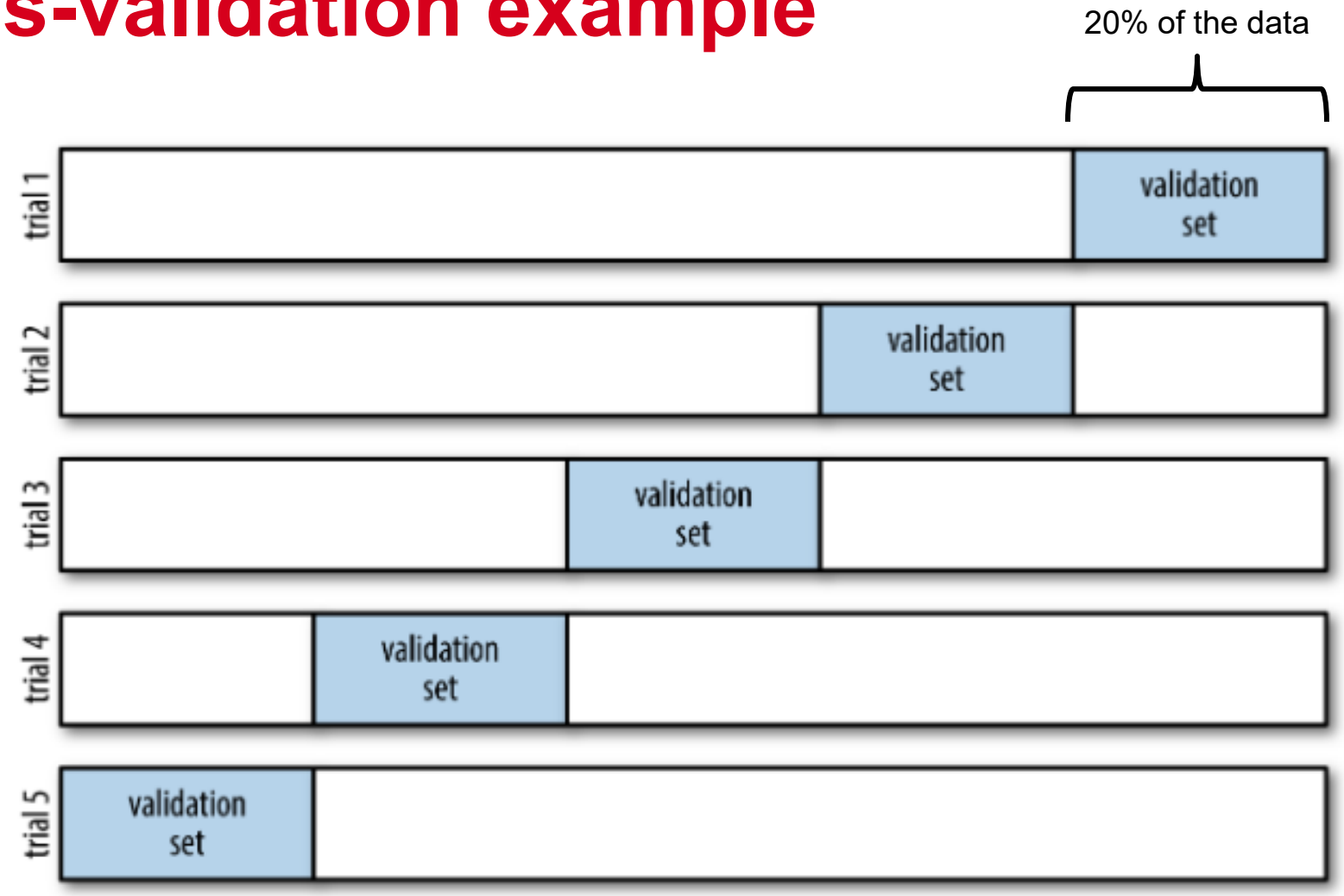


Figure 5-23. Visualization of five-fold cross-validation

Validation set vs. testing set

- In the previous slide, we reference the training set and the validation set
- Typically, we split the data into three sets:
  - Training set – used to train the model (estimate the relationship between the input and output)
  - Validation set – used to validate the hyperparameters selected
  - Testing set – used to test if the model generalizes well with the selected hyperparameters
- Will discuss this in more detail later in the course

Model Selection

How to select the best model?

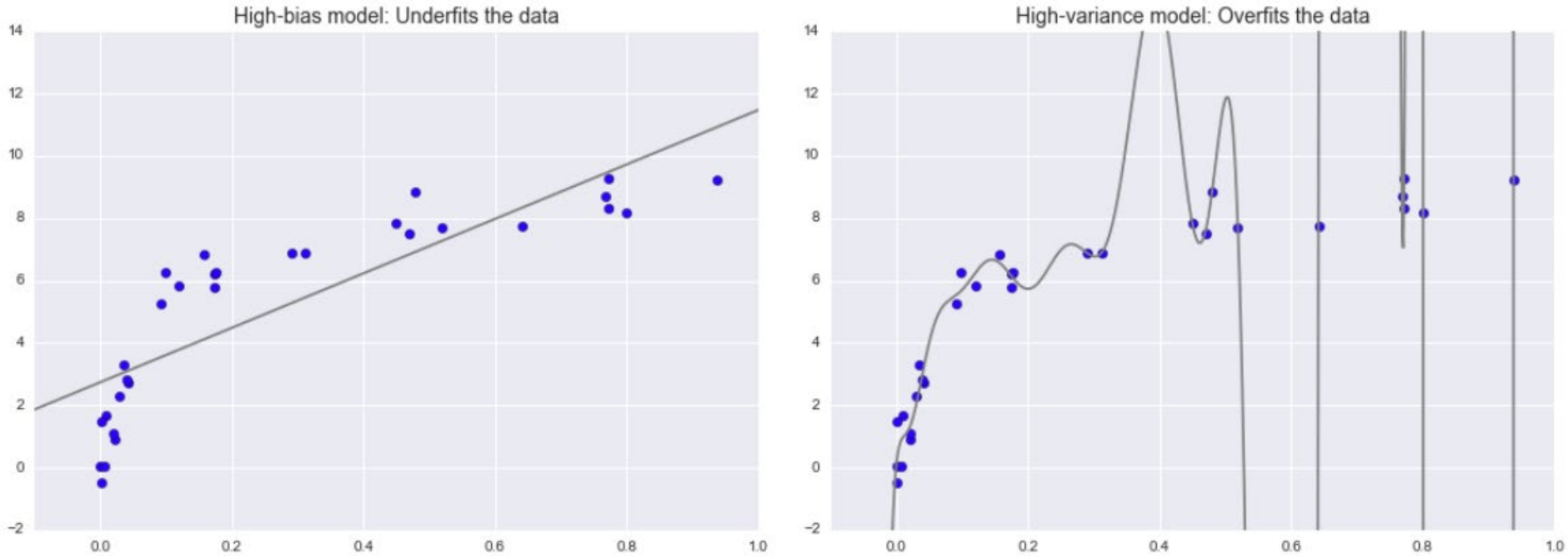


Figure 5-24. A high-bias and high-variance regression model

The bias-variance trade-off

UNIVERSITY OF CALGARY

- Neither of these models is a good fit to the data, but they fail in different ways
- The model on the left attempts to find a straight-line fit through the data
- Because the data is more complicated than a straight line, the straight-line model will never be able to describe this dataset well
- Such a model is said to **underfit** the data - it does not have enough model flexibility to suitably account for all the features in the data
- Another way of saying this is that the model has **high bias**

The bias-variance trade-off

UNIVERSITY OF CALGARY

- The model on the right attempts to fit a high-order polynomial through the data
- This model has enough flexibility to account for the fine features in the data, but its precise form seems to be more reflective of the noise properties of the data rather than the intrinsic properties of whatever process generated that data
- Such a model is said to **overfit** the data - it has so much model flexibility that the model ends up accounting for random errors as well as the underlying data distribution
- Another way of saying this is that the model has **high variance**

Comparing testing and validation

Figure 5-25. Training and validation scores in high-bias and high-variance models

Comparing testing and validation

- The R^2 score (coefficient of determination) measures how well a model performs compared to taking the mean of the data
- The goal is to have a R^2 score close to one – negative values indicate a bad model fit
- For high-bias models, the performance of the model on the validation set is similar to the performance on the training set
- For high-variance models, the performance of the model on the validation set is far worse than the performance on the training set

Validation curve

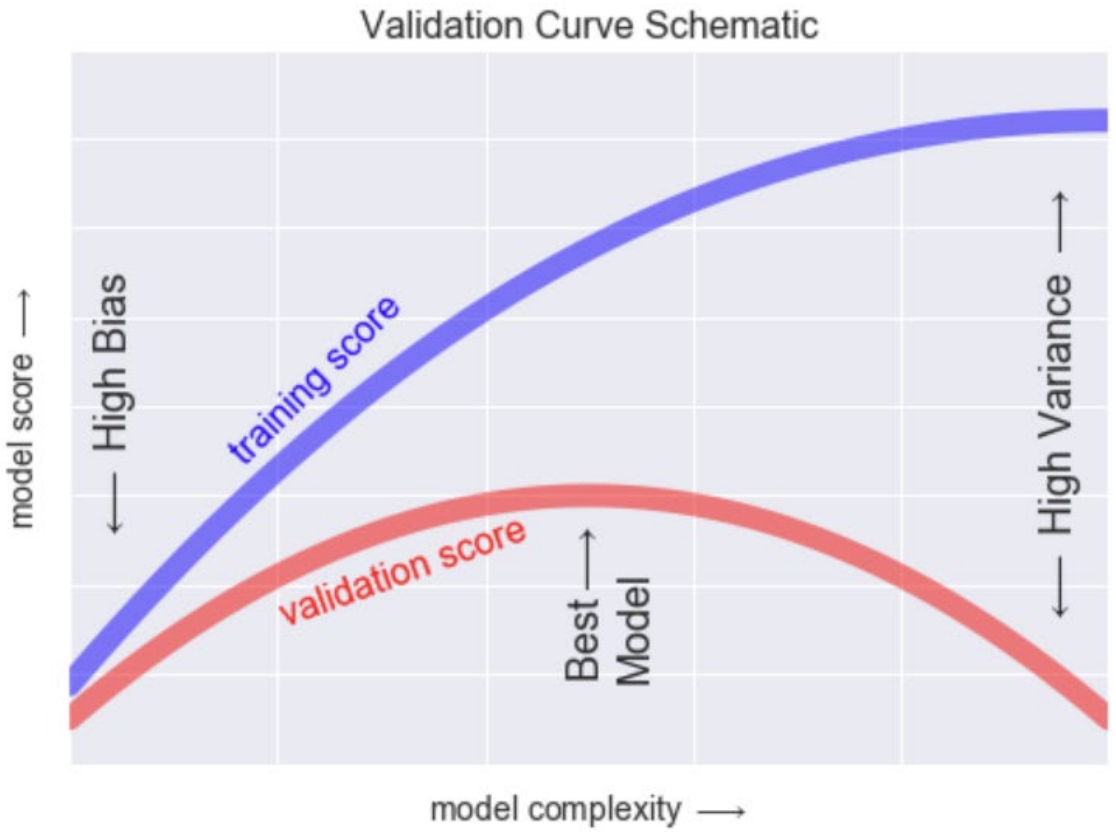


Figure 5-26. A schematic of the relationship between model complexity, training score, and validation score

Validation curve

- The training score will almost always be above the validation score
- The goal is to find the highest validation score, indicating a good trade-off between bias and variance
- May need to tune the hyperparameters to find the best model

What's next?

- We will be covering different supervised learning methods
  - Linear methods
  - Tree-based methods
  - SVM
  - K-nearest neighbours
  - Etc.