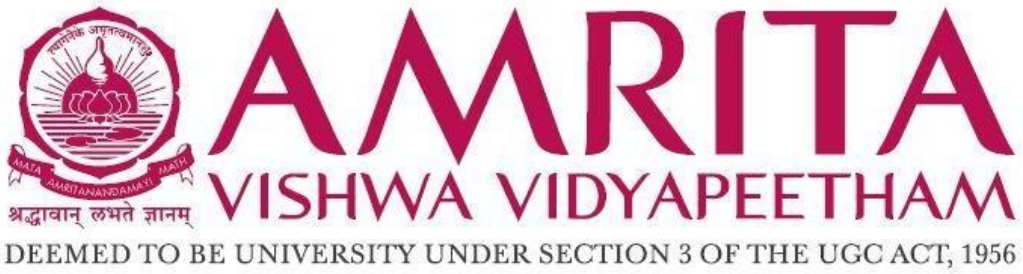


Overfitting versus Underfitting

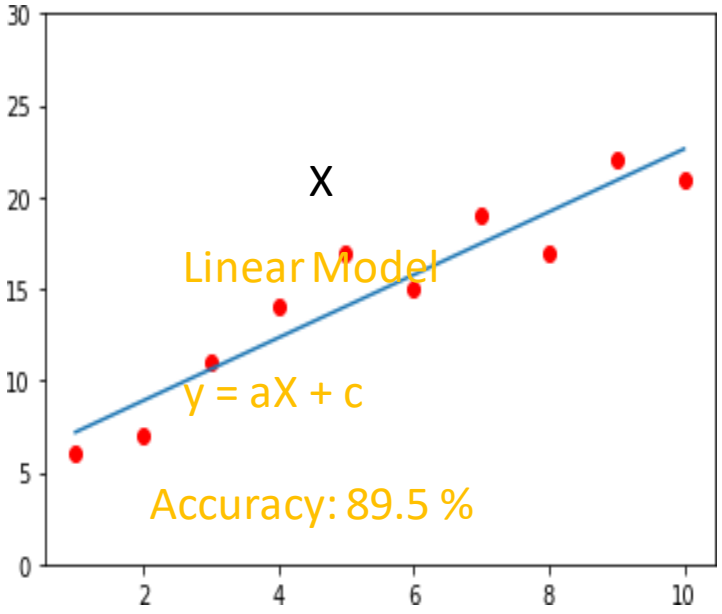
Supervised Machine Learning

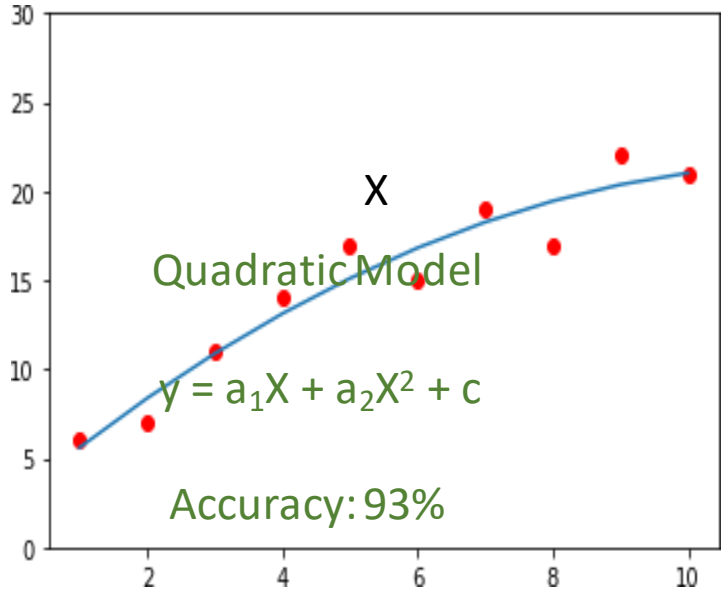
- Algorithm learns a model from training data
- Goal is to best estimate the mapping function (f)

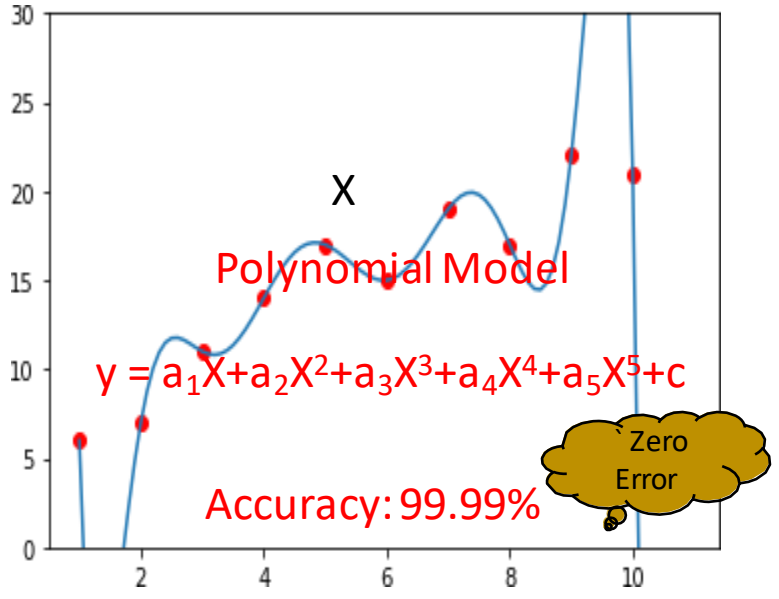
  f(X) -> Y
- Inductive Learning : Learning general concepts from specific examples

Underfitting and Overfitting

- How well a machine learning model learns and generalizes to new data
- Fit : Refers to how well you approximate a target function.







Simple : Underfitting **Complex : Overfitting**

Model Evaluation

- To better understand machine learning algorithms
- To get better performance on your data.

- Overfitting or underfitting the data - Cause of poor performance in machine learning
- Overfitting – Less Generalization
- Underfitting – More assumptions

Training Error

A

Guessing: ~50%

Underfitting

B

Overfitting

Mr. know it all
~98%

Quiz based on
class work

Professor

C

Problem solving approach:
~92%

Good fit

Testing Error

Guessing: ~47%

A

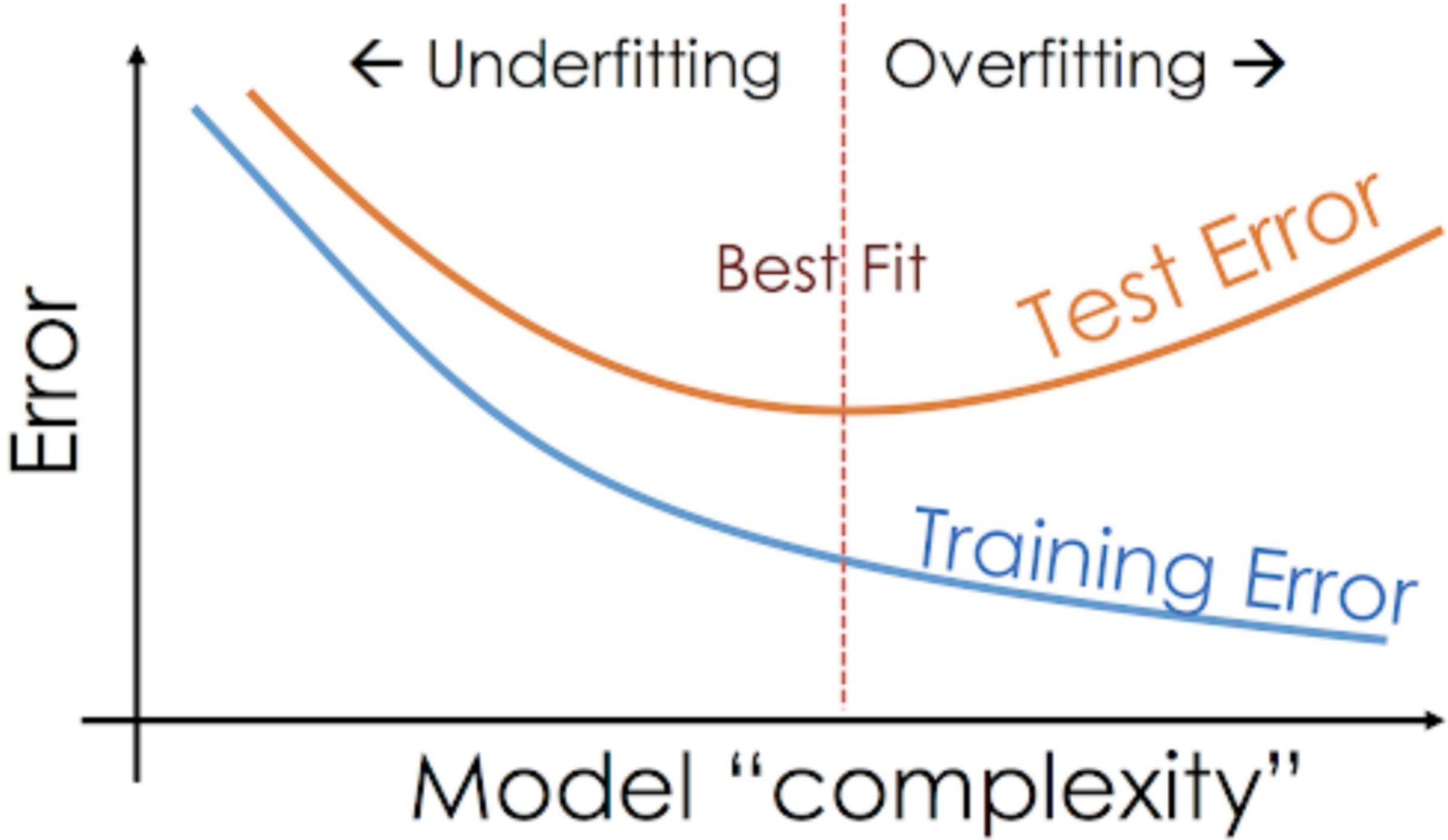
Mr. know it all
~69%

B

Problem solving approach:
~89%

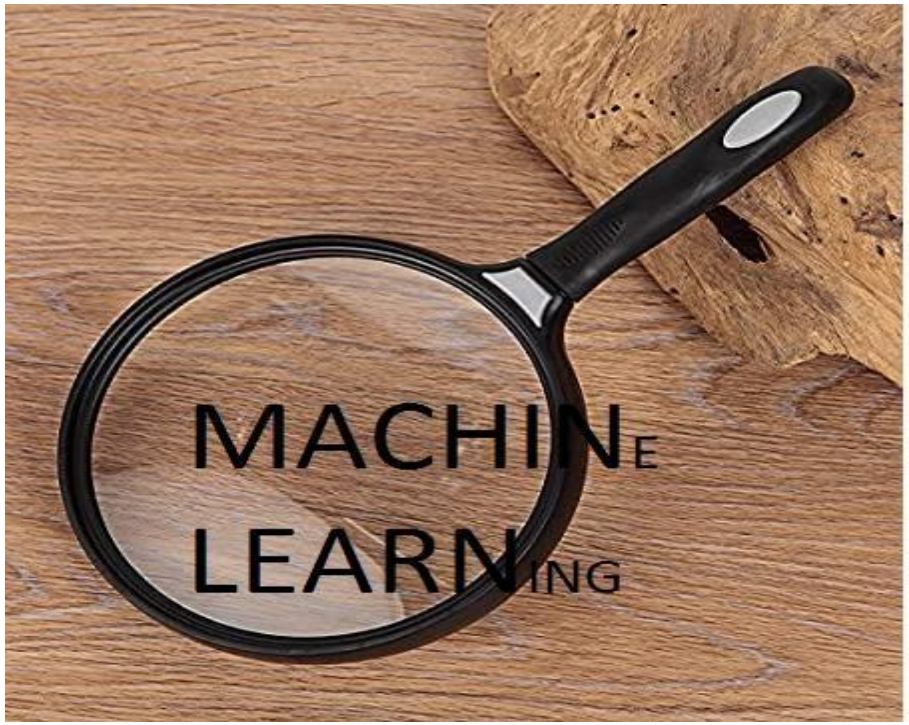
Semester Exam

Professor


← Underfitting
Overfitting →
Best Fit
Test Error
Training Error
Error
Model "complexity"

AMRITA VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY UNDER SECTION 3 OF THE UGC ACT, 1956

Bias-Variance Tradeoff



Prediction errors

- Influenced by ML algorithms
  - Bias Error
  - Variance Error
- Not Influenced by ML algorithms
  - Irreducible error

- Proper understanding of Bias and Variance errors help to build accurate models
- Avoid the mistake of overfitting and underfitting.

Irreducible Errors

- Cannot be reduced by creating good models.
- Erupts due to inconsistent data/noisy data.
- Problem framing strategy
- Avoidance or neglecting variables that influence the target function

Bias

- Simplifying assumptions made by a model to make the target function easier to learn.

$$y = \theta_0 + \theta_1 x_1$$

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_d x_d = \sum_{j=0}^{d} \theta_j x_j$$

- Less flexible and Lower predictive
- **Low Bias**: Low assumptions about the form of the target function.
- **High-Bias**: High assumptions about the form of the target function.

Assumptions lead to Bias error!!!

Variance

- Variance is the change in the estimate of the target function with change in the training data.
- The algorithm should be good at picking out the hidden underlying mapping between the inputs and the output variables.
- **Low Variance**: A small change in the data sample leads to a small change to the estimate of the target function.

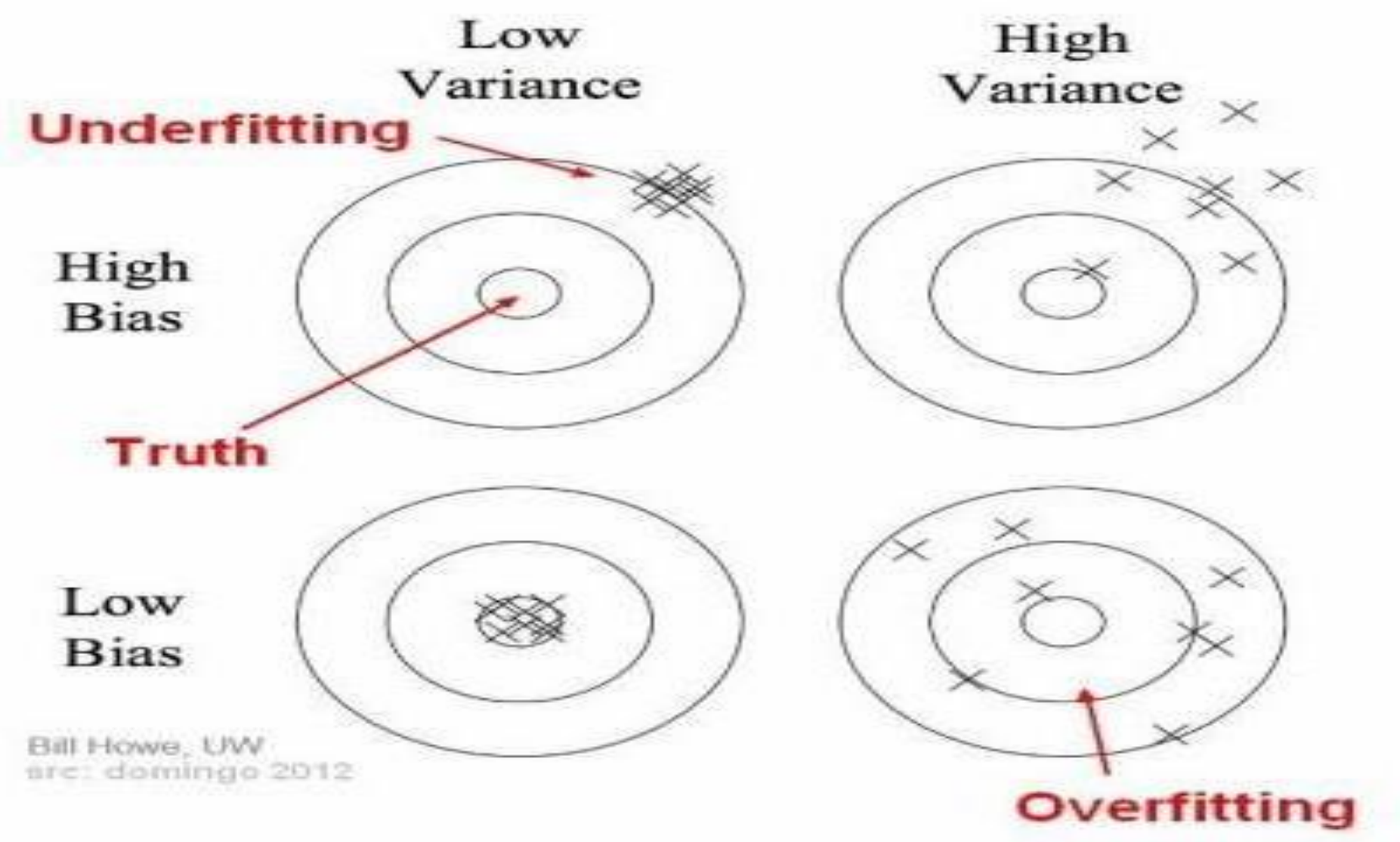
$$y = 2x + 2$$

- **High Variance**: A small change in the training data leads to very big change to the estimate of the target function.

$$y = 3x^3 + 2x^2 + 5x + 2$$

Complex functions lead to Variance error!!!

Graphical illustration of bias and variance



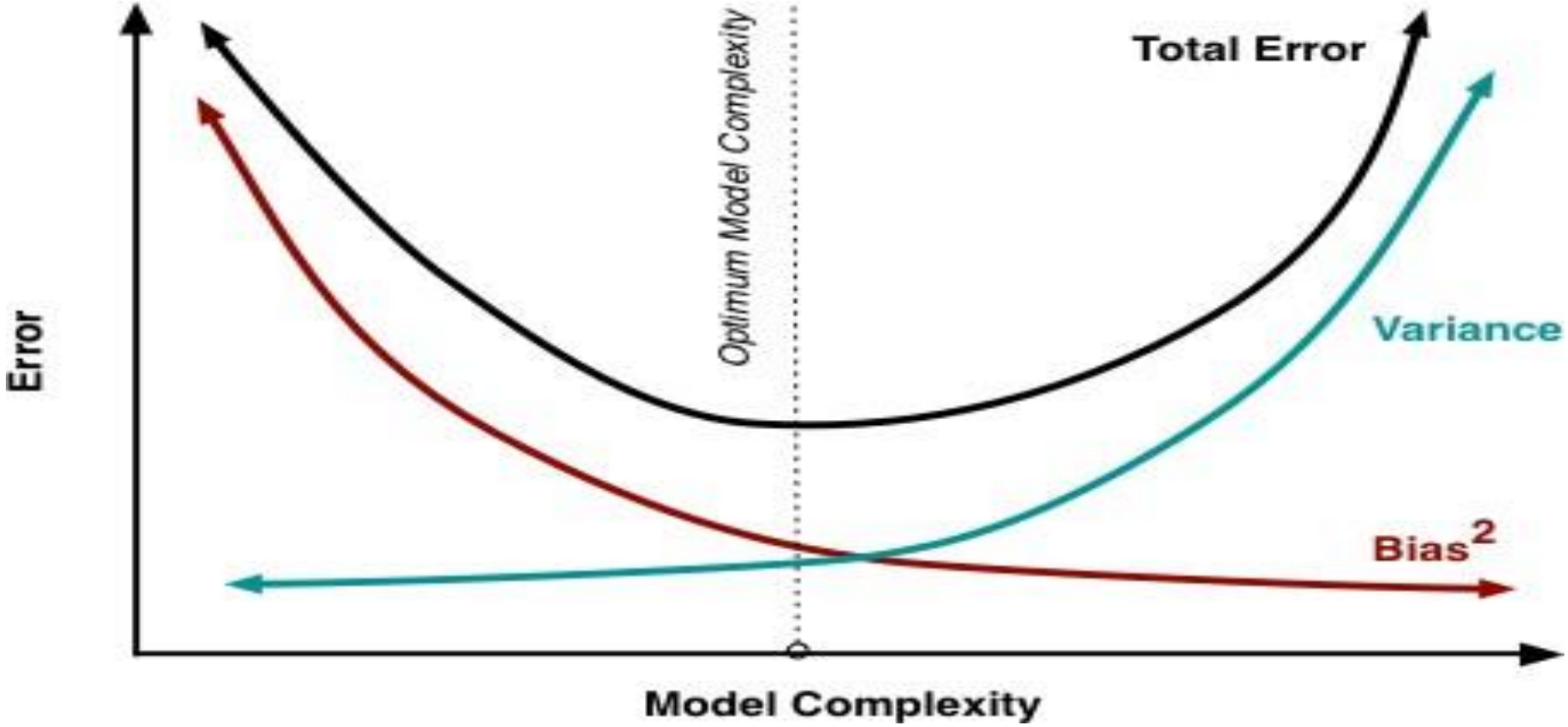
Bias – Variance Trade-Off

- Achieve low bias and low variance.
- Achieve good prediction performance.

- **Linear** machine learning algorithms - High bias ; Low variance.
- **Nonlinear** machine learning algorithms - Low bias ; High variance.

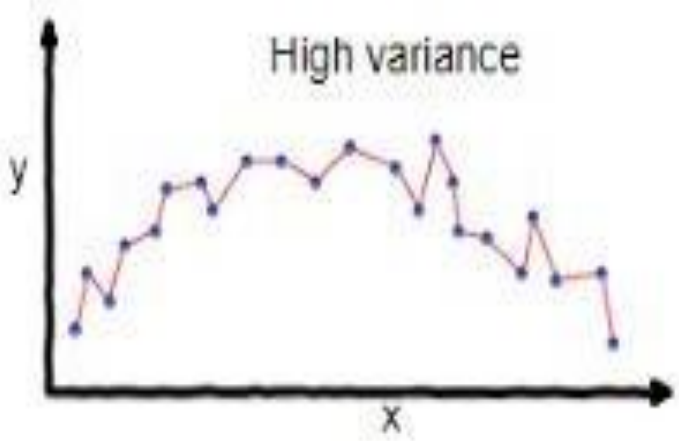
- Parameterization of machine learning algorithms is often a battle to balance out bias and variance.

Contribution to total error

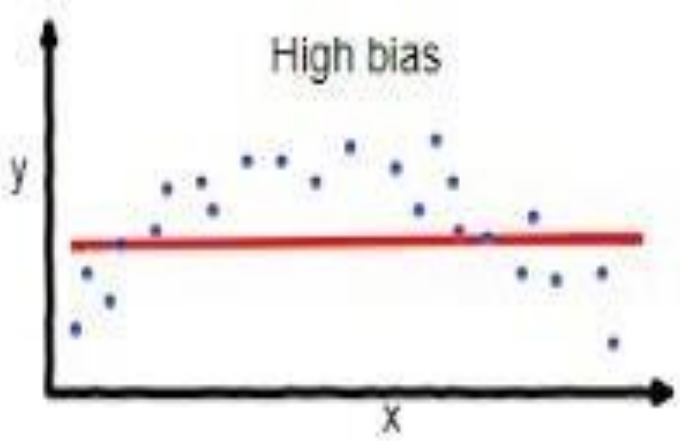


Relationship between Bias and Variance

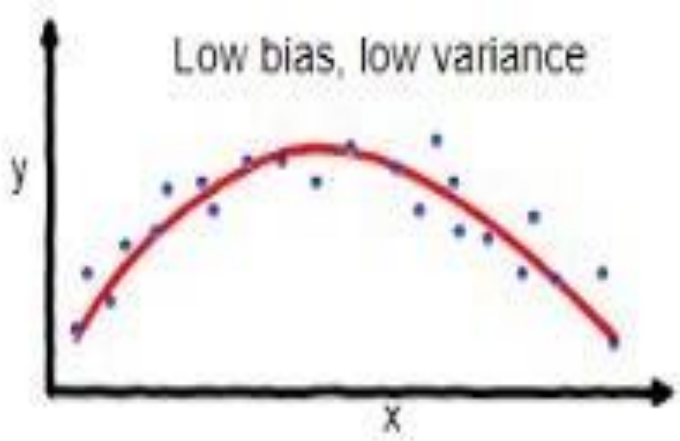
- Increasing the bias will decrease the variance.
- Increasing the variance will decrease the bias.



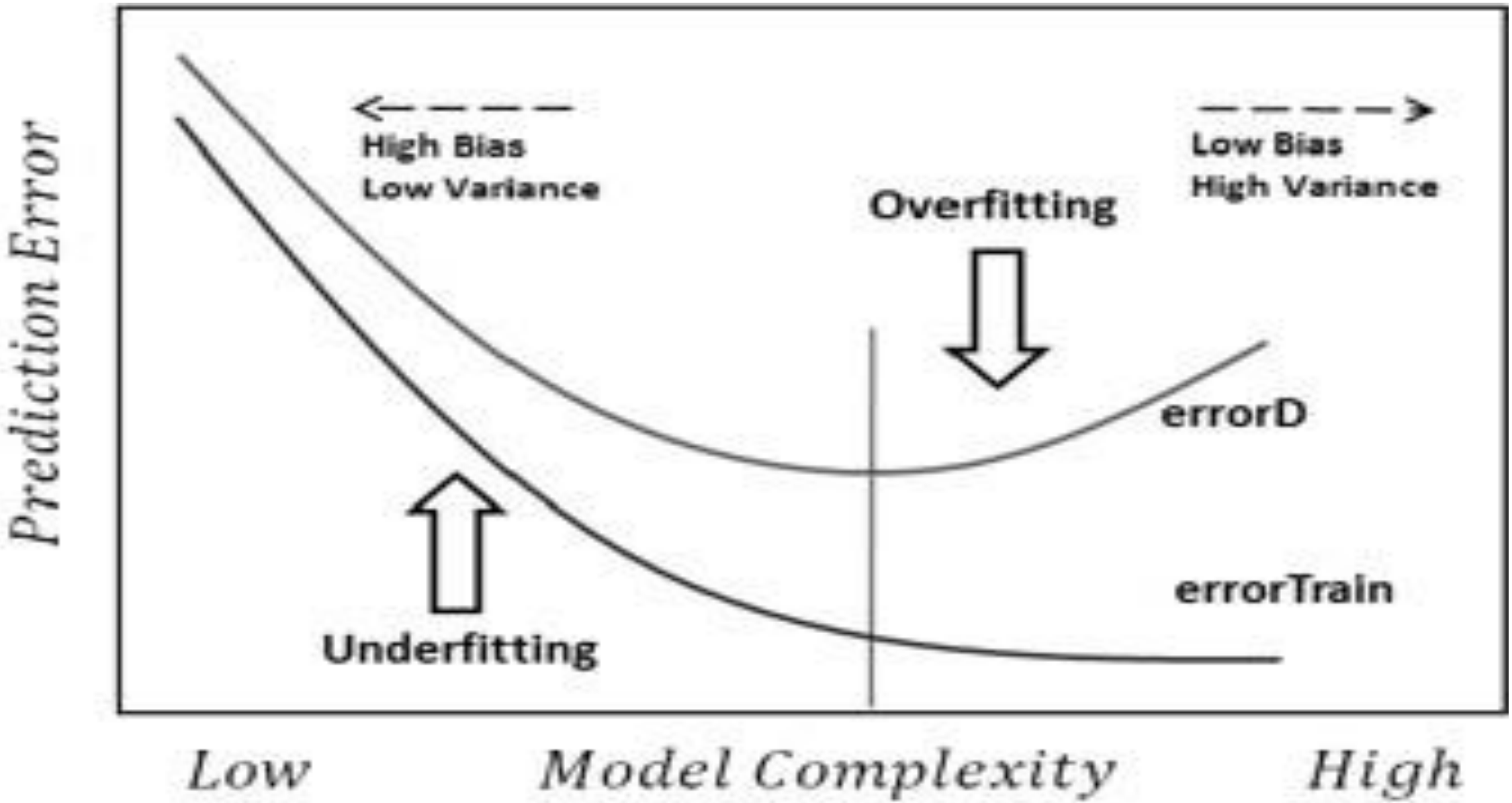
overfitting



underfitting



Good balance


Prediction Error
High Bias
Low Variance
Low Bias
High Variance
Overfitting
errorD
errorTrain
Underfitting
Low
Model Complexity
High

Curb Overfitting/Underfitting

Overfitting (High Variance)

- Cross validation
- Increase number of samples
- Reduce number of features
- Reduce the significance of the features(Regularization)

Underfitting(High Bias)

- Increase number of features
- Decrease number of samples
- Try adding polynomial features