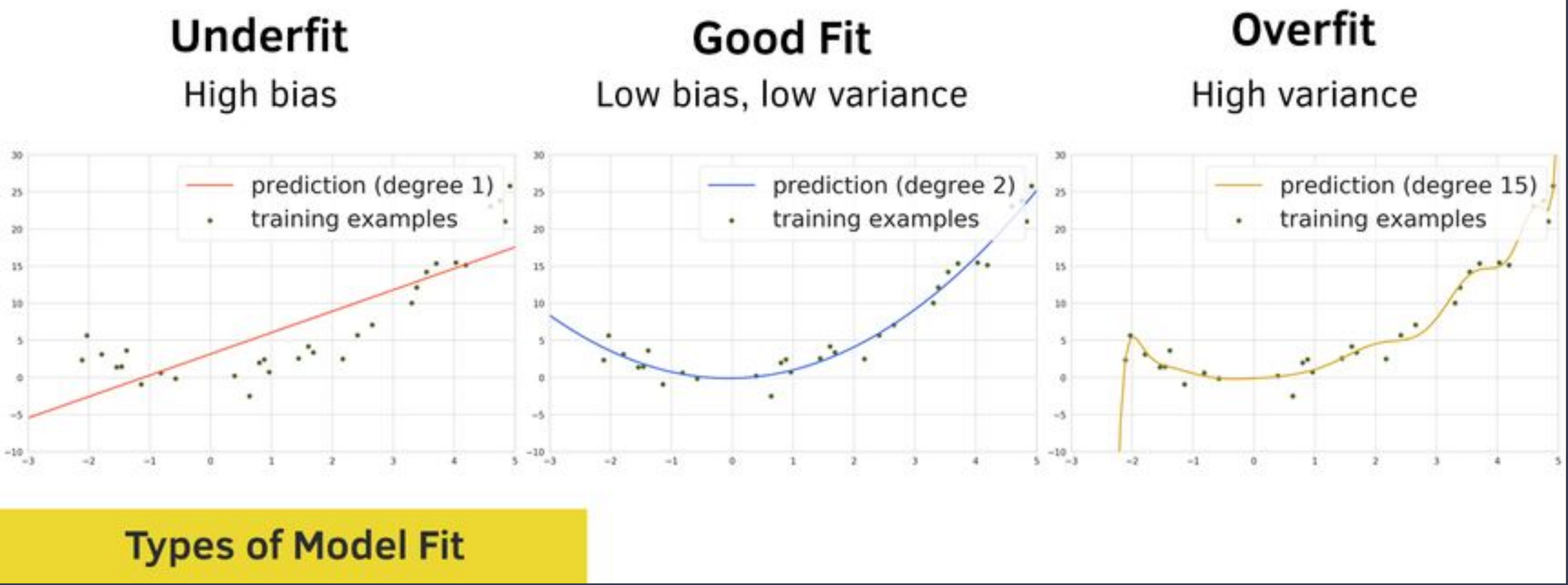
Index Class

1. Bias-Variance Dilemma
2. Train-test split

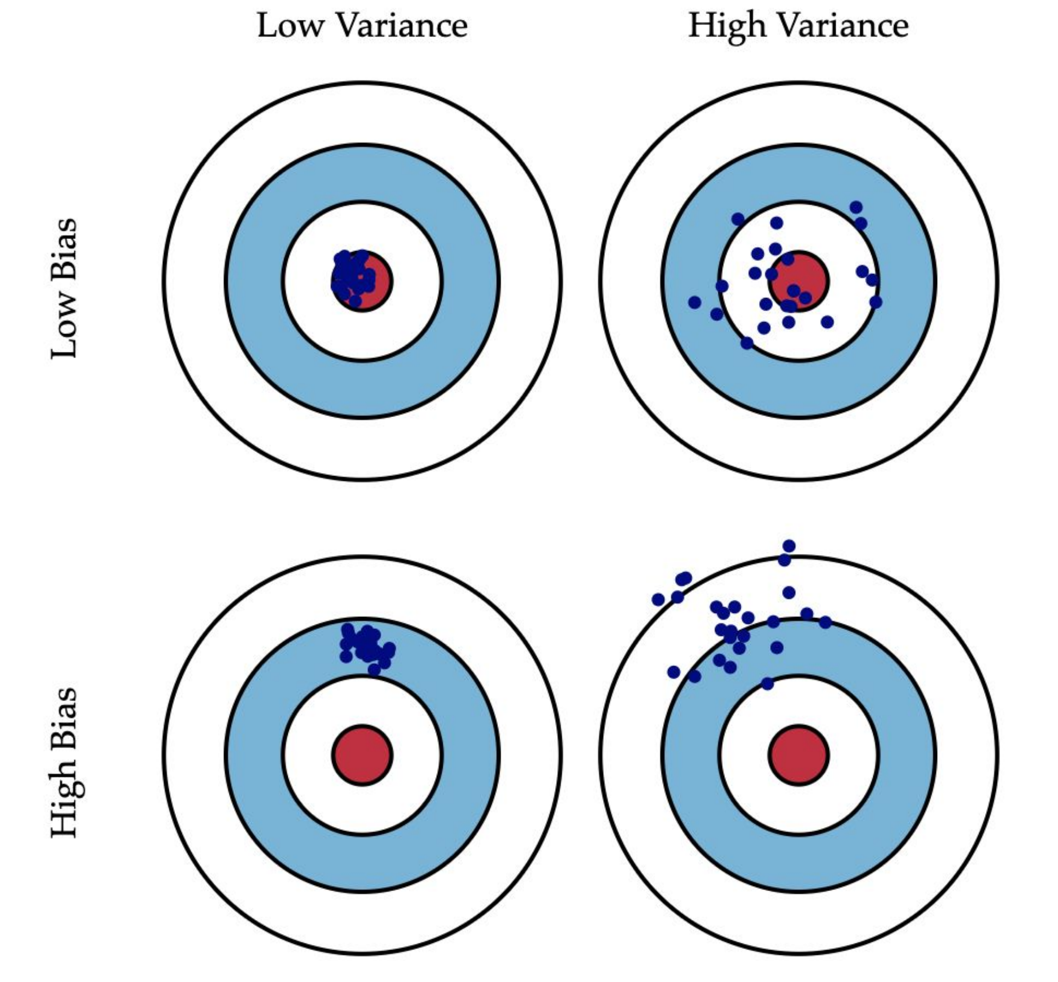
Types of Fit

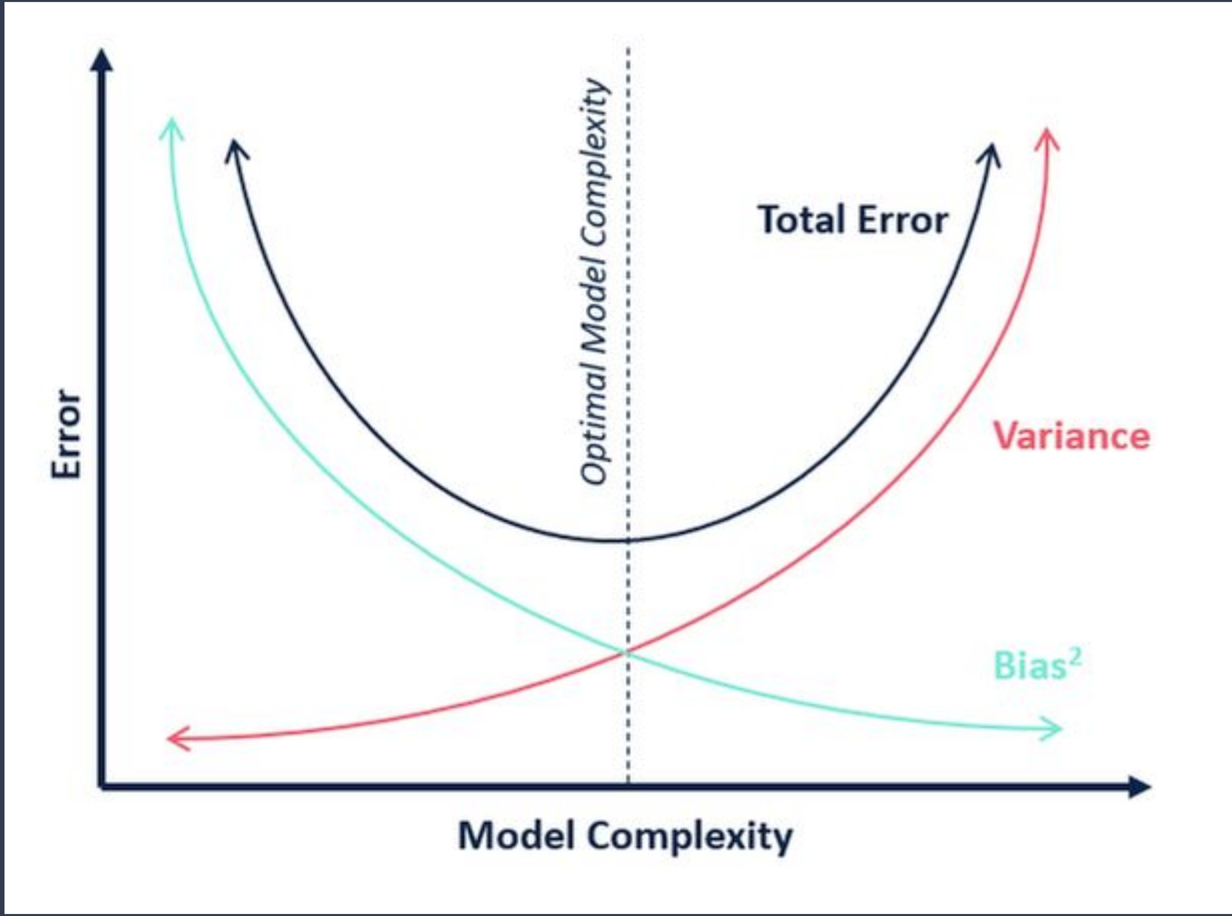

Underfit
High bias
prediction (degree 1)
training examples
Good Fit
Low bias, low variance
prediction (degree 2)
training examples
Overfit
High variance
prediction (degree 15)
training examples
Types of Model Fit

Bias-Variance dilemma

The **bias–variance dilemma** or **bias–variance problem** is the conflict in trying to simultaneously minimize these two sources of error that prevent supervised learning algorithms from generalizing beyond their training set:[1][2]

- The *bias* error is an error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting).
- The *variance* is an error from sensitivity to small fluctuations in the training set. High variance may result from an algorithm modeling the random noise in the training data (overfitting).


Low Variance
High Variance
Low Bias
High Bias



Error
Optimal Model Complexity
Total Error
Variance
Bias²
Model Complexity

Regression	
'explained_variance'	`metrics.explained_variance_score`
'max_error'	`metrics.max_error`
'neg_mean_absolute_error'	`metrics.mean_absolute_error`
'neg_mean_squared_error'	`metrics.mean_squared_error`
'neg_root_mean_squared_error'	`metrics.root_mean_squared_error`
'neg_mean_squared_log_error'	`metrics.mean_squared_log_error`
'neg_root_mean_squared_log_error'	`metrics.root_mean_squared_log_error`
'neg_median_absolute_error'	`metrics.median_absolute_error`
'r2'	`metrics.r2_score`
'neg_mean_poisson_deviance'	`metrics.mean_poisson_deviance`
'neg_mean_gamma_deviance'	`metrics.mean_gamma_deviance`
'neg_mean_absolute_percentage_error'	`metrics.mean_absolute_percentage_error`
'd2_absolute_error_score'	`metrics.d2_absolute_error_score`
'd2_pinball_score'	`metrics.d2_pinball_score`
'd2_tweedie_score'	`metrics.d2_tweedie_score`

Cheat Sheet – Bias-Variance Tradeoff

What is Bias? $bias = \mathbb{E}[f'(x)] - f(x)$

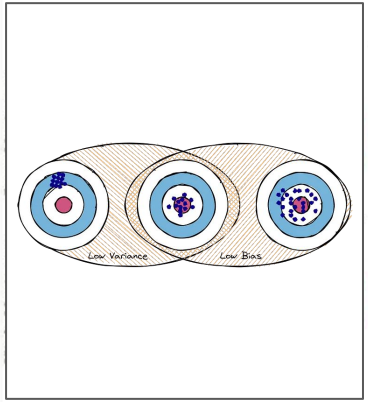
- Error between average model prediction and ground truth
- The bias of the estimated function tells us the capacity of the underlying model to predict the values

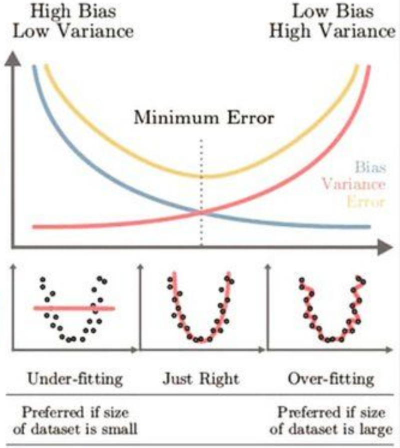
What is Variance? $variance = \mathbb{E}\left[\left(f'(x) - \mathbb{E}[f'(x)]\right)^2\right]$

- Average variability in the model prediction for the given dataset
- The variance of the estimated function tells you how much the function can adjust to the change in the dataset

High Bias → Overly-simplified Model
→ Under-fitting
→ High error on both test and train data

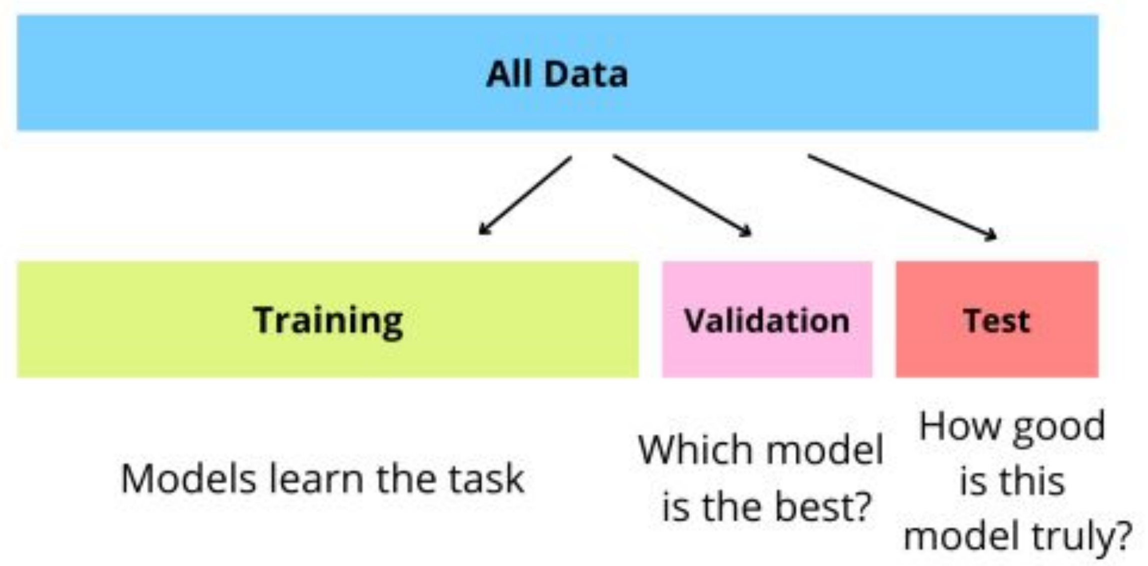
High Variance → Overly-complex Model
→ Over-fitting
→ Low error on train data and high on test
→ Starts modelling the noise in the input





Bias variance Trade-off

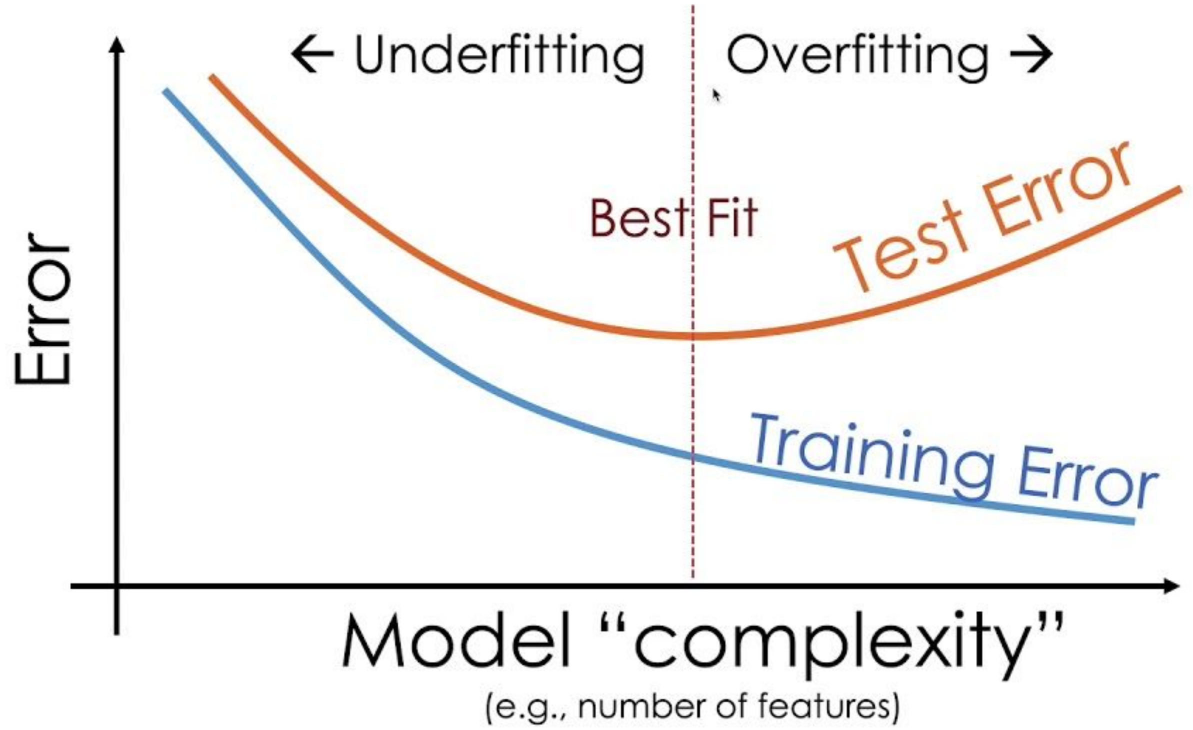
- Increasing bias reduces variance and vice-versa
- Error = bias^2 + variance +irreducible error
- The best model is where the error is reduced.
- Compromise between bias and variance

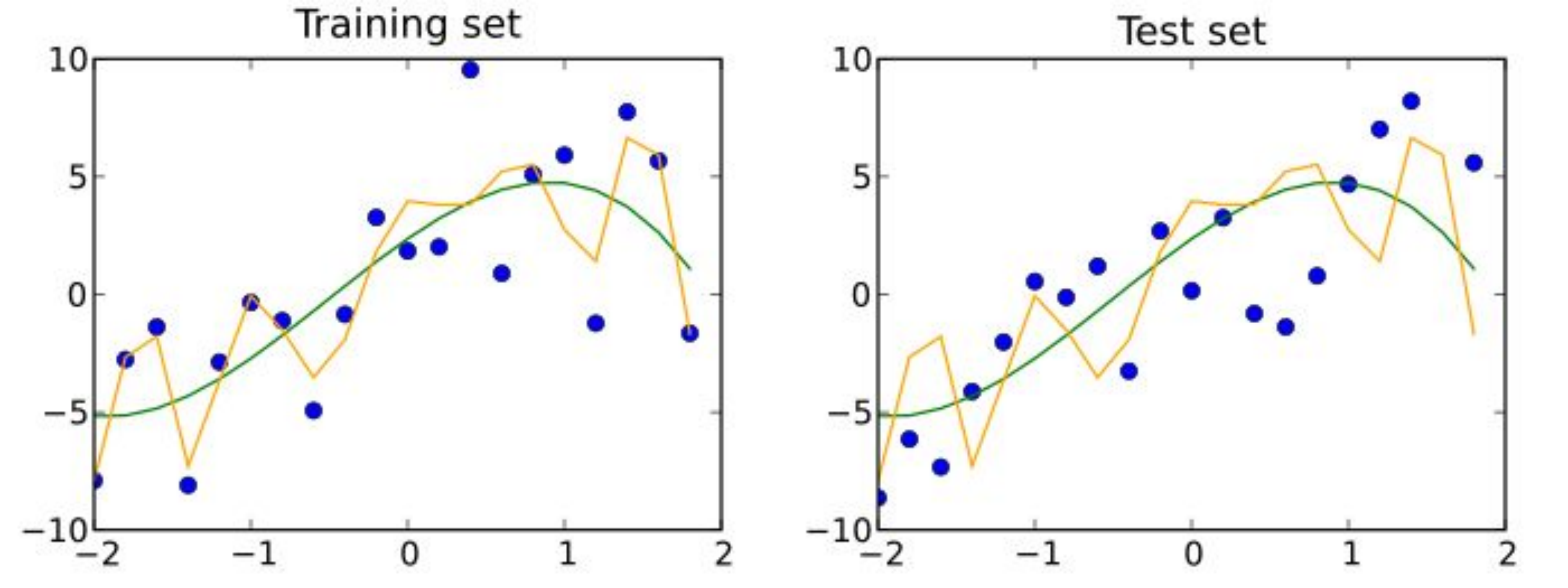

All Data
Training
Validation
Test
Models learn the task
Which model is the best?
How good is this model truly?


A
Training
Test
Single Dataset
B
Training
Validation
Test
Single Dataset

Training vs **Test** Error

Training error typically *under estimates* **test error**.





Training set
Test set
10
5
0
−5
−10
−2
−1
0
1
2