Generalization Error and Bias Variance Tradeoff

CS1090A Introduction to Data Science

Pavlos Protopapas, Kevin Rader, and Chris Gumb

Photo: Greg Mccutcheon
Bryce Canyon, Utah

Outline

- Recap – Model Selection
- Generalization Error, Bias Variance Tradeoff
- Regularization Techniques: Lasso, Ridge

Outline

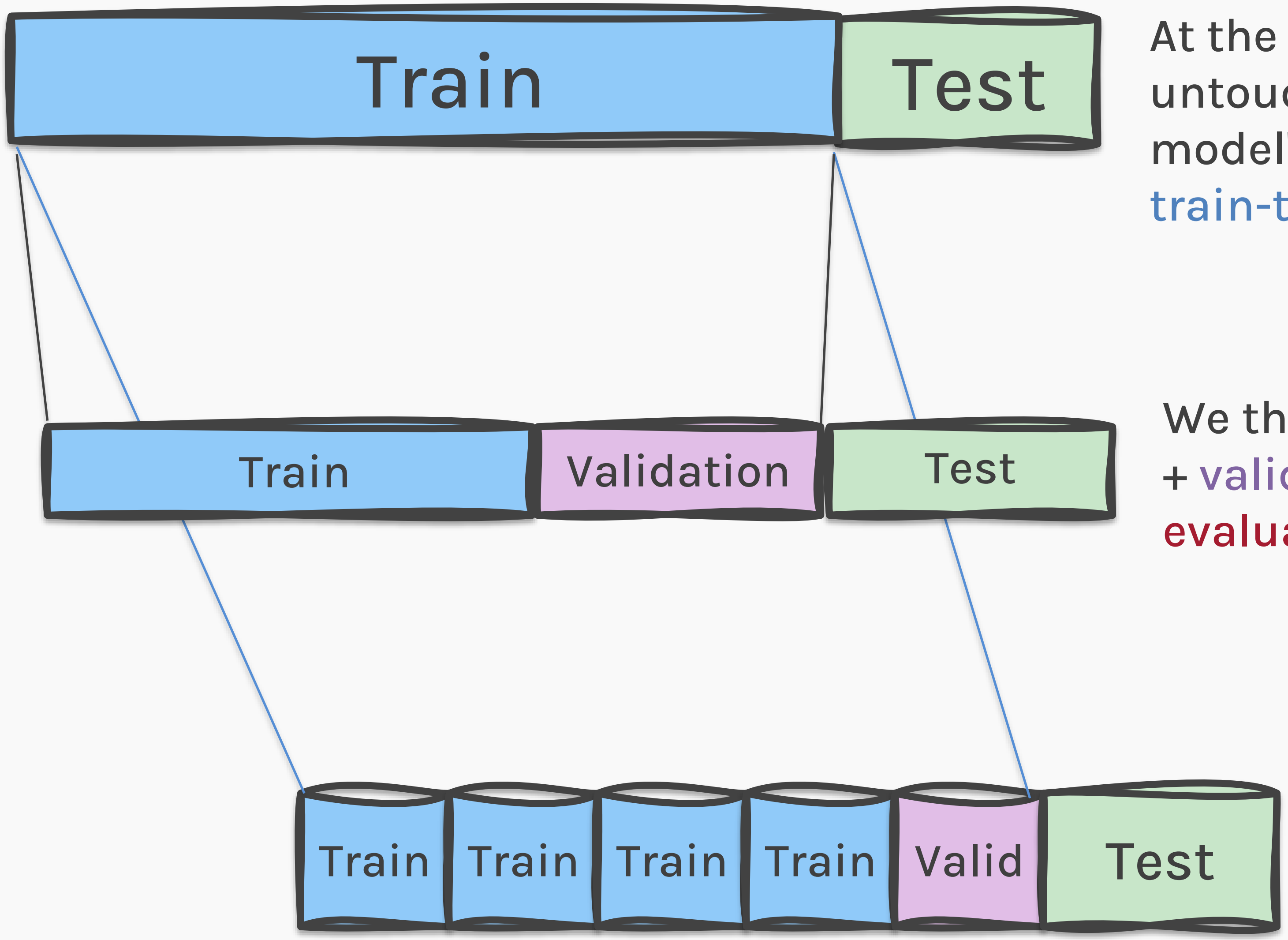
- **Recap – Model Selection**
- Generalization Error, Bias Variance Tradeoff
- Regularization Techniques: Lasso, Ridge

Recall - Model Selection

At the start, we set aside a portion of the data, untouched until the end, to evaluate the final model's performance. This is called the train-test split. *

* sometimes they (not us!) also call this train + validation split, while meaning train + test.

Recall - Model Selection



At the start, we set aside a portion of the data, untouched until the end, to evaluate the final model's performance. This is called the train-test split. *

We then saw we can split train data into train + validation (to find the best model) + test (to evaluate the performance of the model).

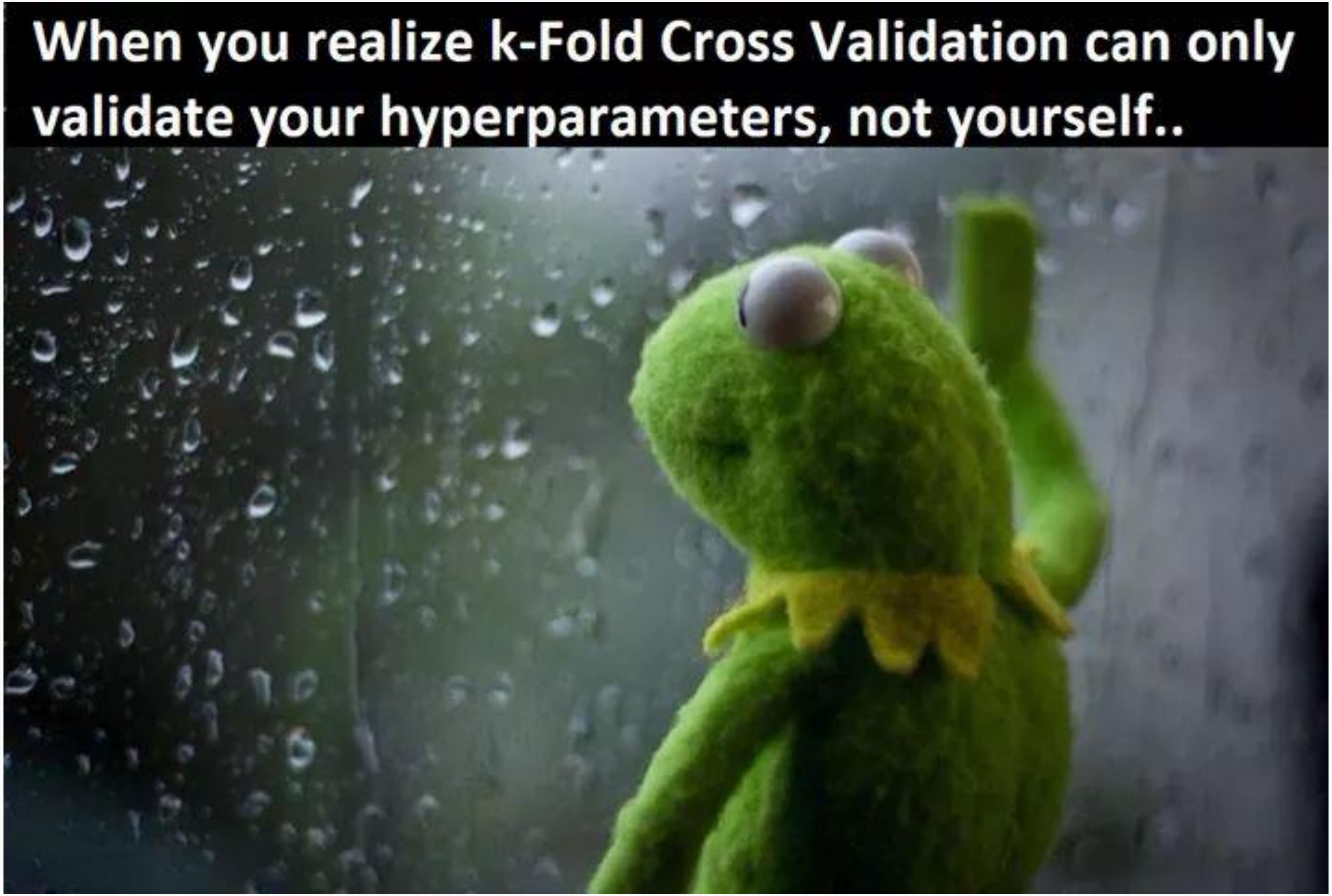
We then finally saw that we can use cross-validation. It splits the train data into k buckets and uses different chunks of data as the validation set.

Recall - Model Selection

1. Model selection as a way to avoid overfitting
2. Validation set to select the best model
3. Cross validation to avoid overfitting to the validation set

Ways of model selection:

- Exhaustive search
- Greedy algorithms
- Fine tuning hyper-parameters
- **Regularization**


When you realize k-Fold Cross Validation can only validate your hyperparameters, not yourself..

Outline

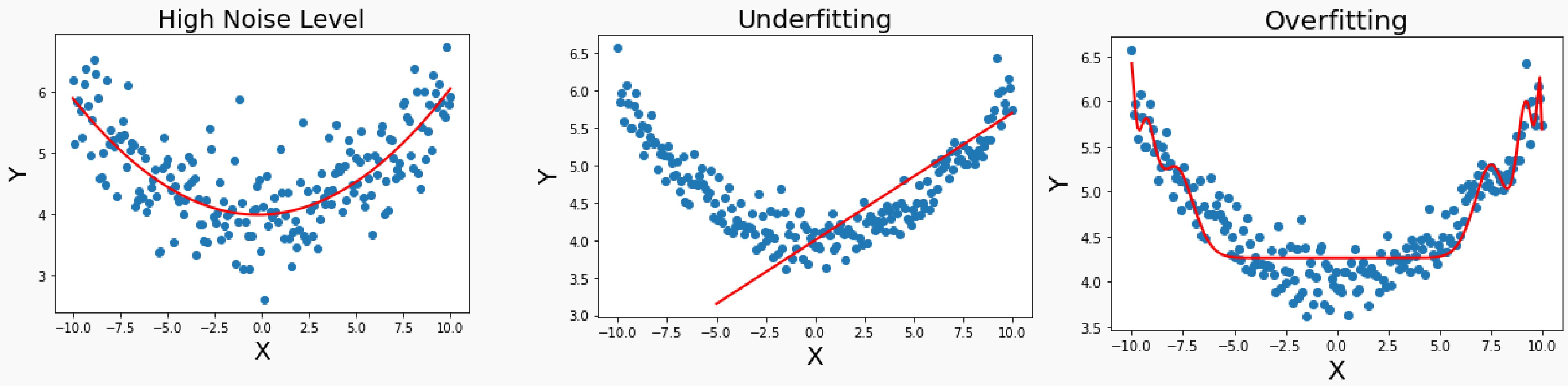
- Recap – Model Selection
- **Generalization Error, Bias Variance Tradeoff**
- Regularization Techniques: Lasso Ridge

Test Error and Generalization

We know to evaluate models on both train and test data because models can do well on train data but do poorly on new data.

When models do well on new data, it is called generalization.

There are at least three ways a model can have a high-test error.



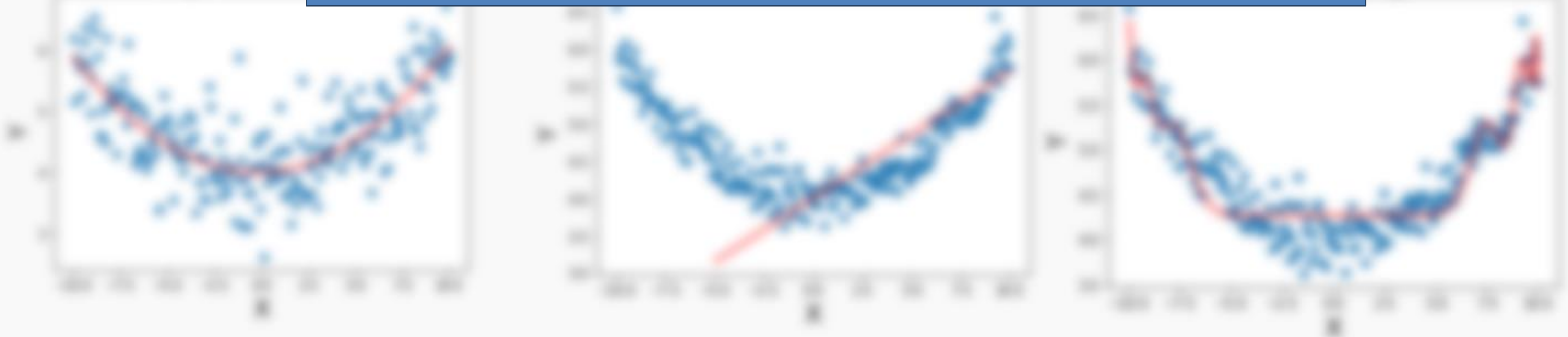
Test Error and Generalization

We know to evaluate models on both train and test data because models can do well on train data but do poorly on new data.

When models do well on new data, it is called generalization.

There are at

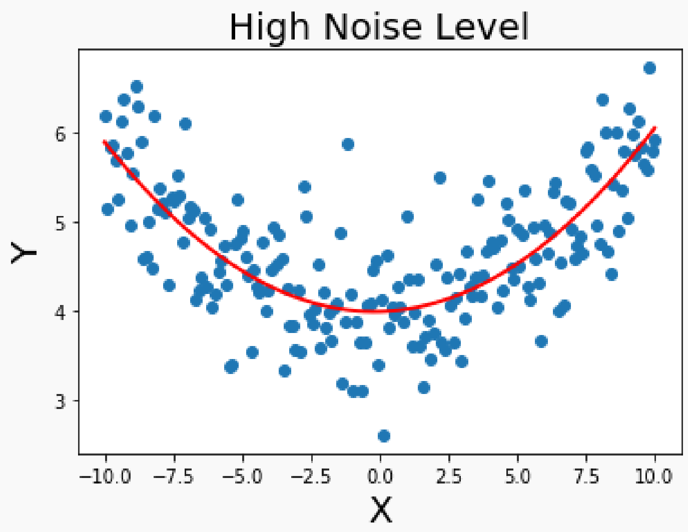
We can classify the test error into 2 types, **Irreducible** and **Reducible**



Irreducible and Reducible Errors

Irreducible error (or aleatoric error):

- This is error caused by random noise in the data.
- No matter how good the model is, this error cannot be reduced.



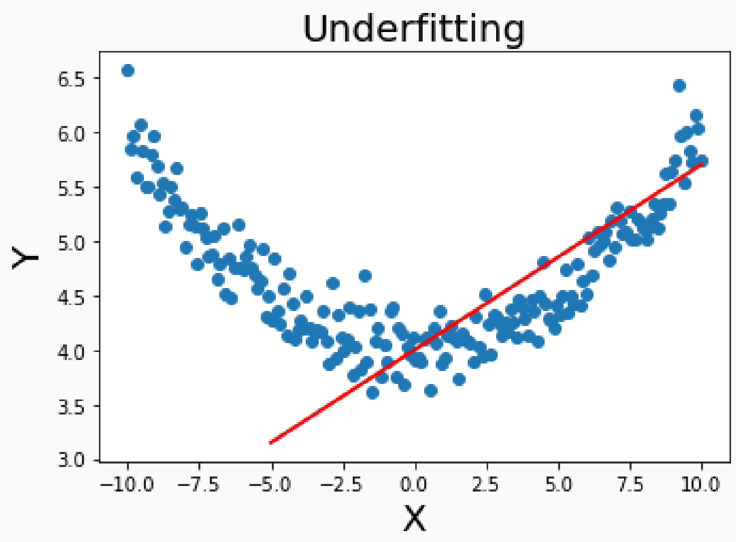
Irreducible error is always present and cannot be eliminated.

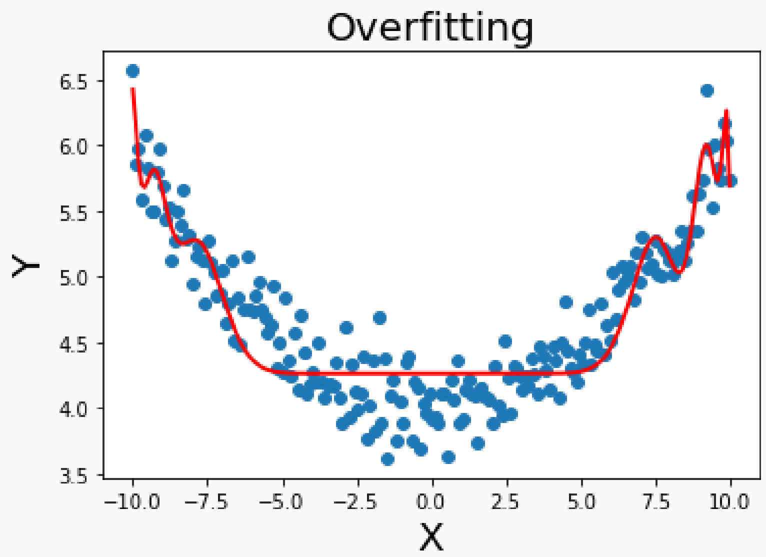
Irreducible and Reducible Errors

Reducible error (or aleatoric error):

- This comes from model limitations, such as overfitting or underfitting.

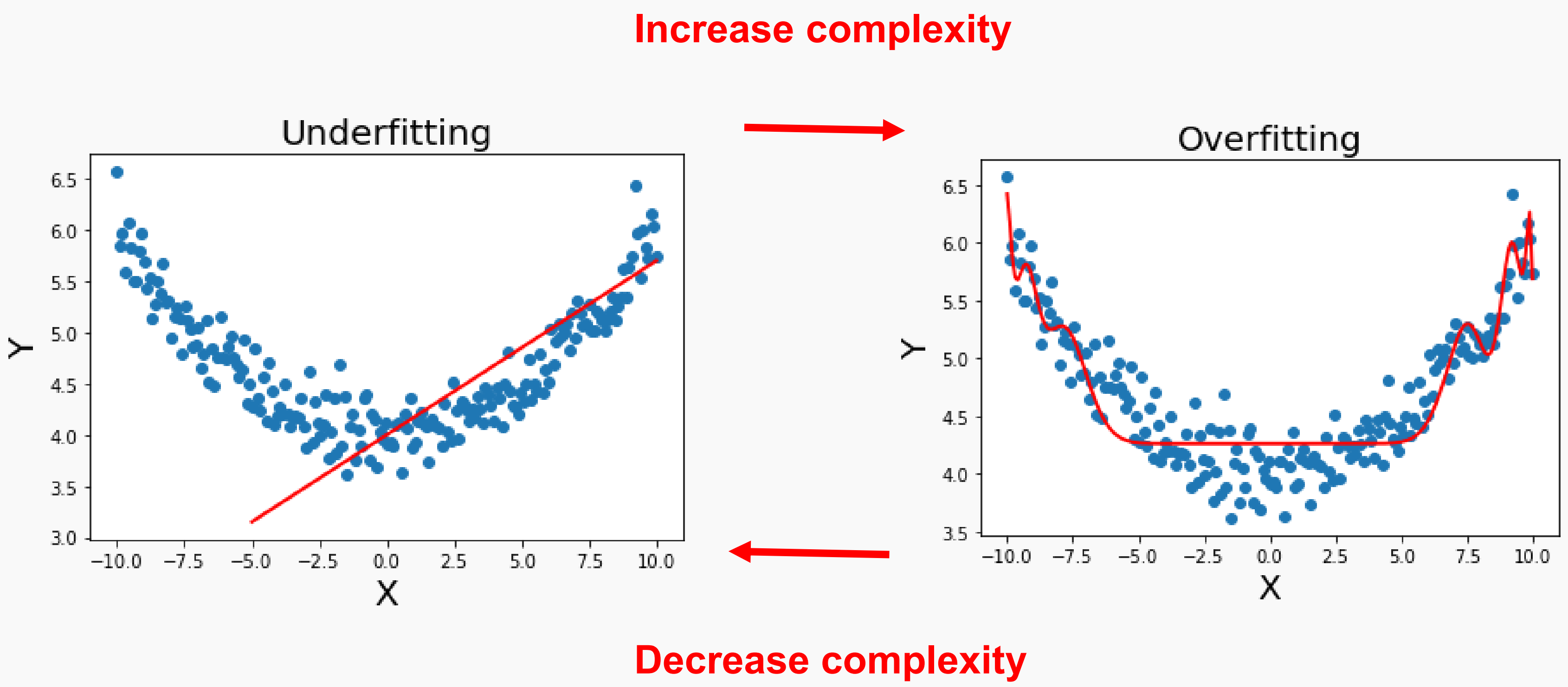
- We can reduce this error by improving the model, or using better data preprocessing.





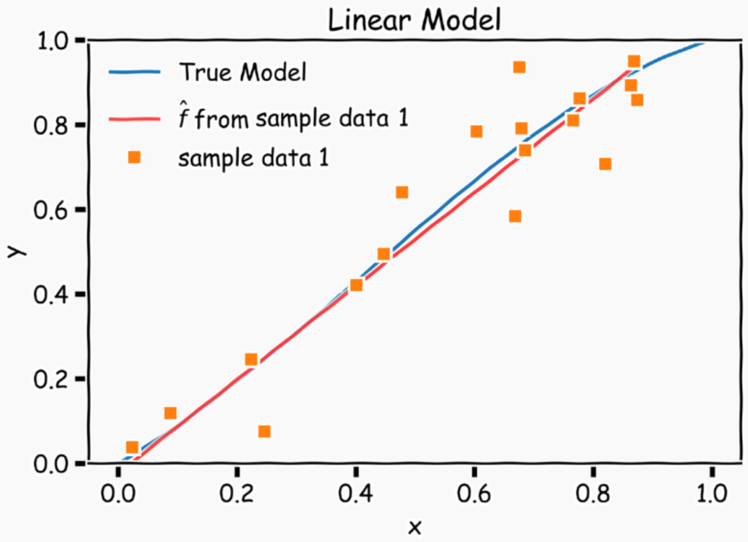
The Bias-Variance: Bias

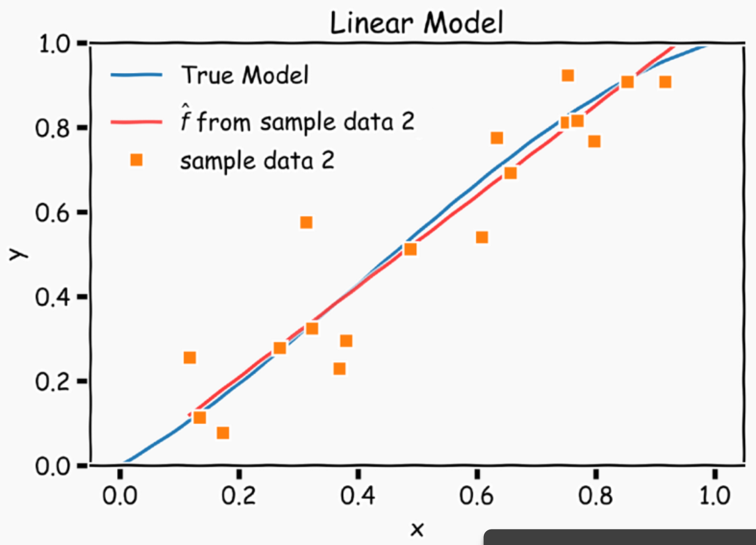
Reducible error comes from either underfitting or overfitting. There is a tradeoff between the two sources of errors:

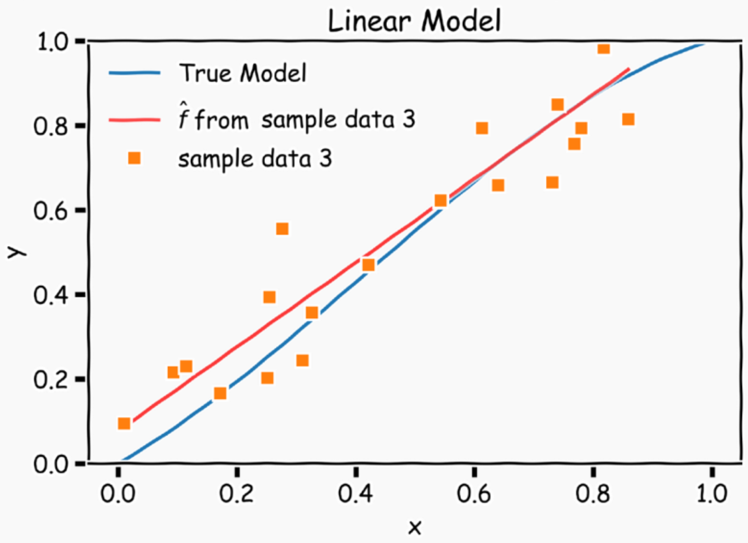


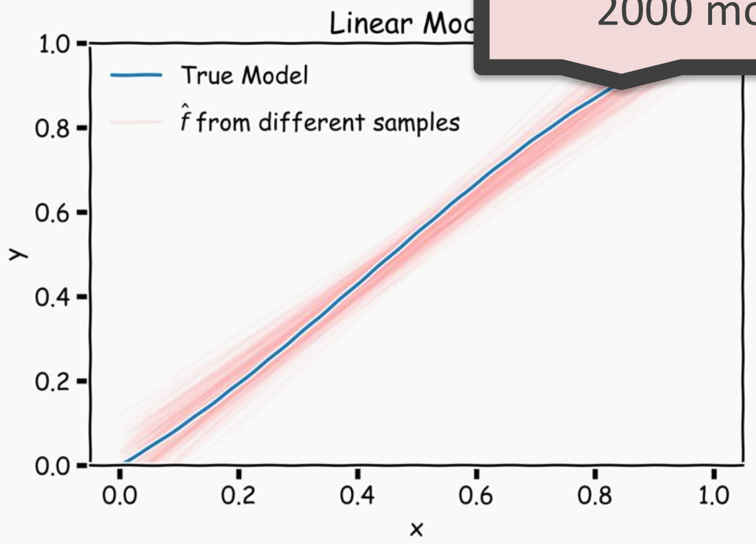

NYPD
WITH GREAT
MODEL COMPLEXITY
COMES GREAT
RISK OF OVERFITTING
imgflip.com

Bias vs Variance: Variance of a SIMPLE model



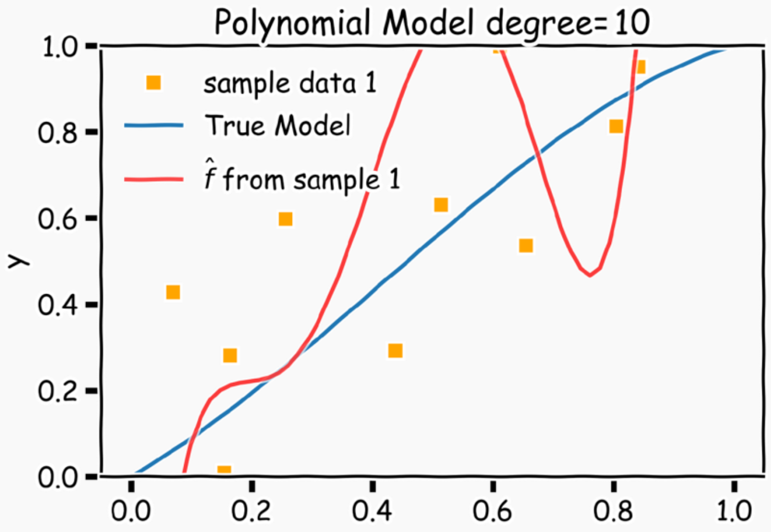


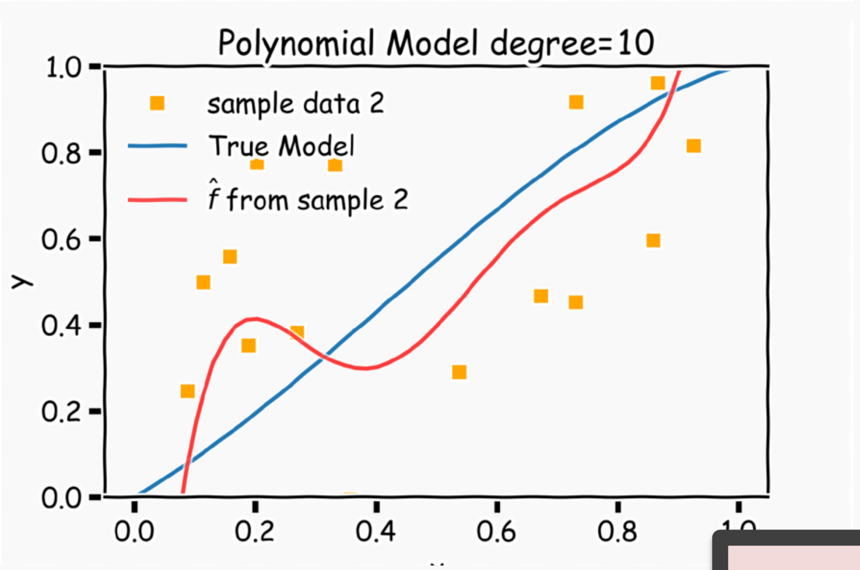


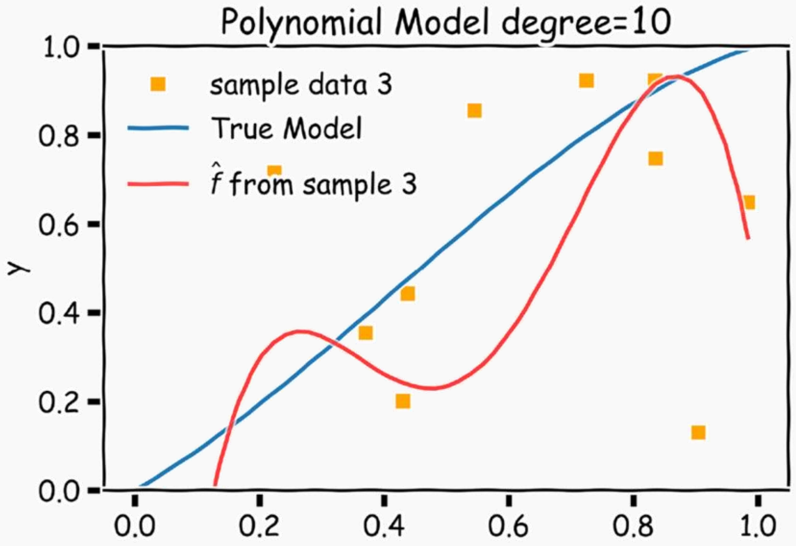


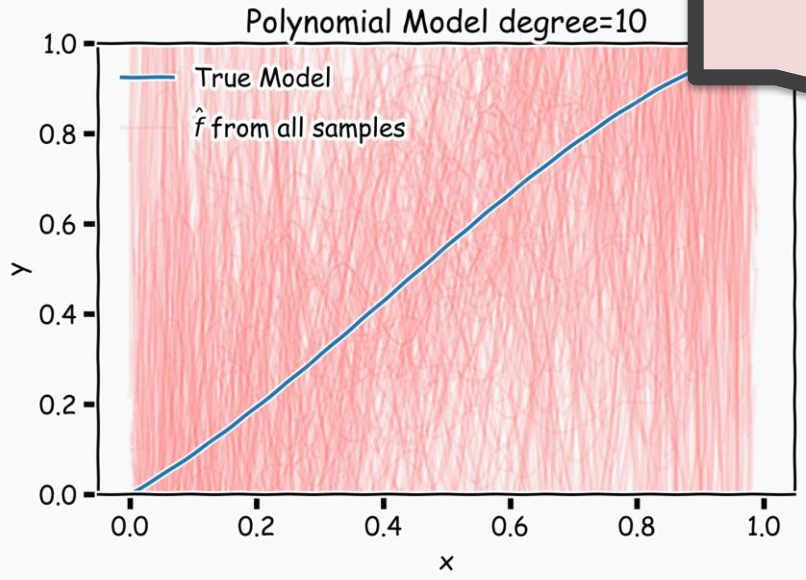
2000 models

Bias vs Variance: Variance of a COMPLEX model





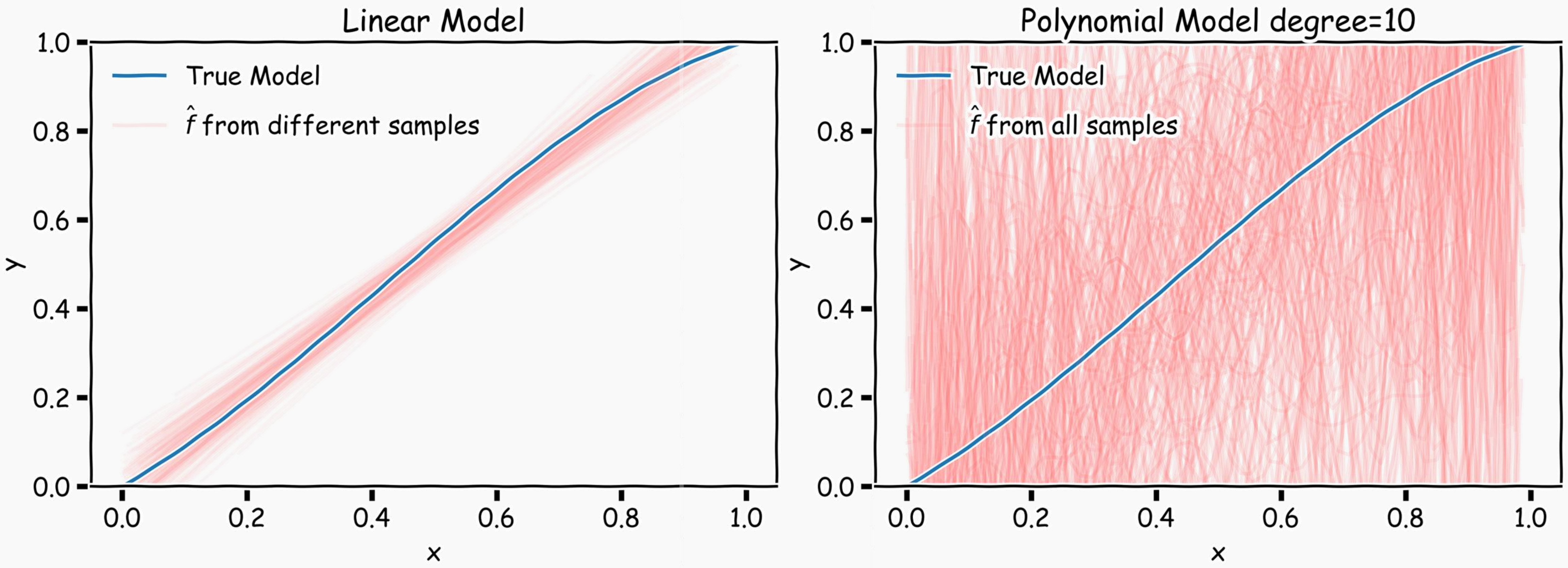




2000 models

Bias vs Variance

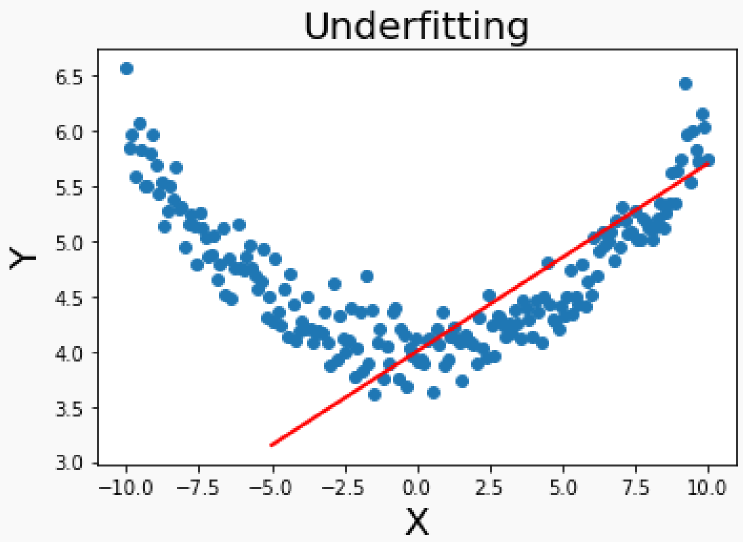
Left: 2,000 best-fit linear models, each fitted to a different 20-point training set.
Right: 2,000 best-fit models using degree-10 polynomials.

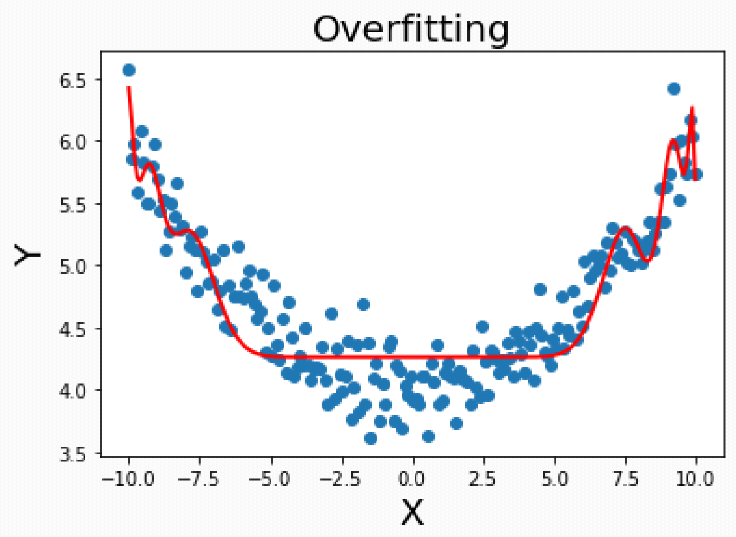


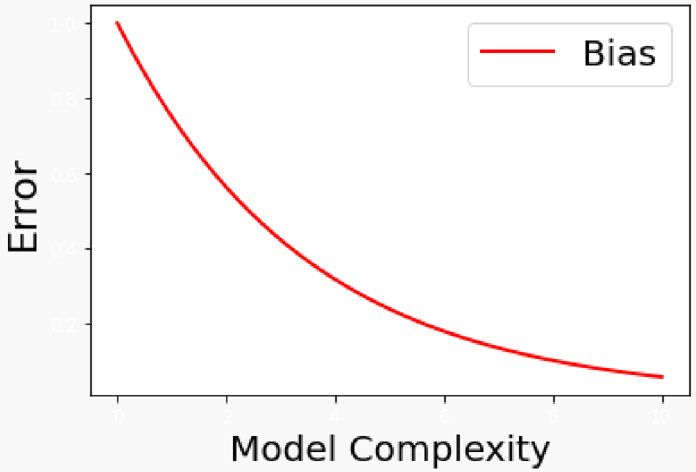
The Bias-Variance: Bias

Bias refers to how far off a model's predictions are from the actual truth.

Increase complexity



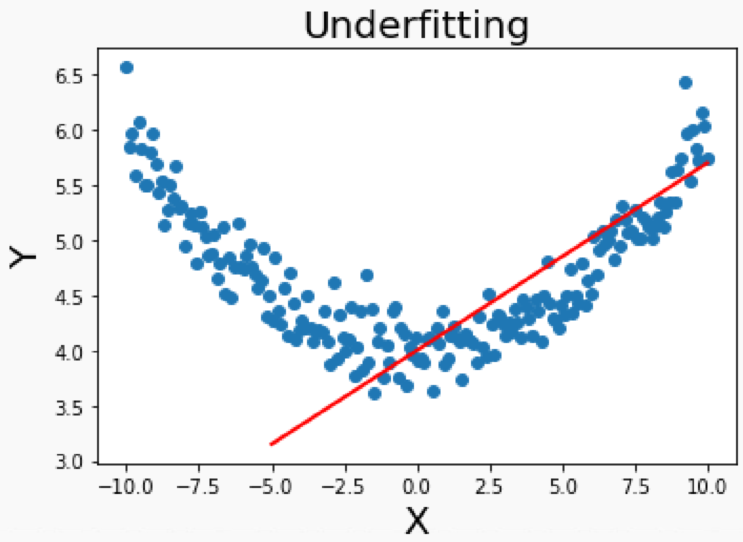


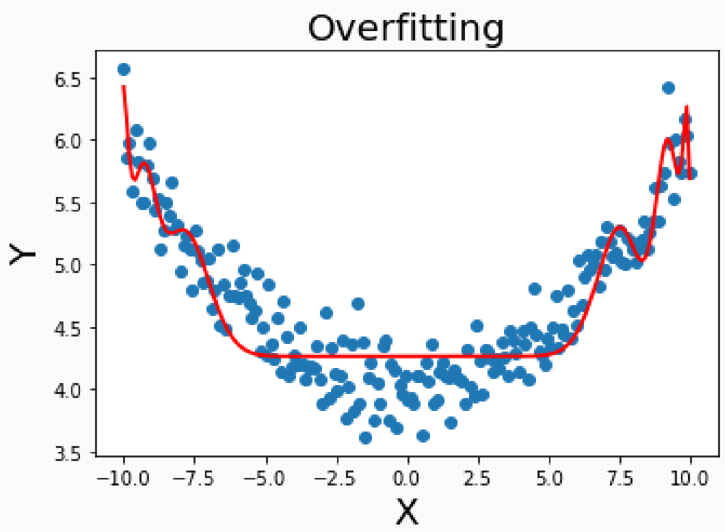


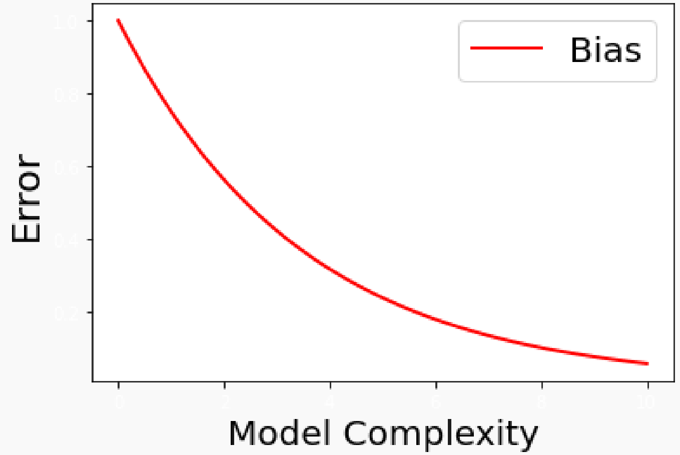
Decrease complexity

The Bias-Variance Trade Off

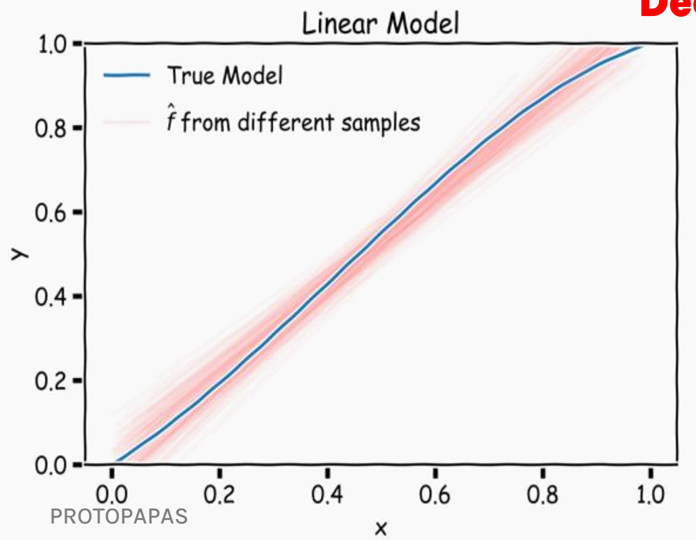
Increase complexity

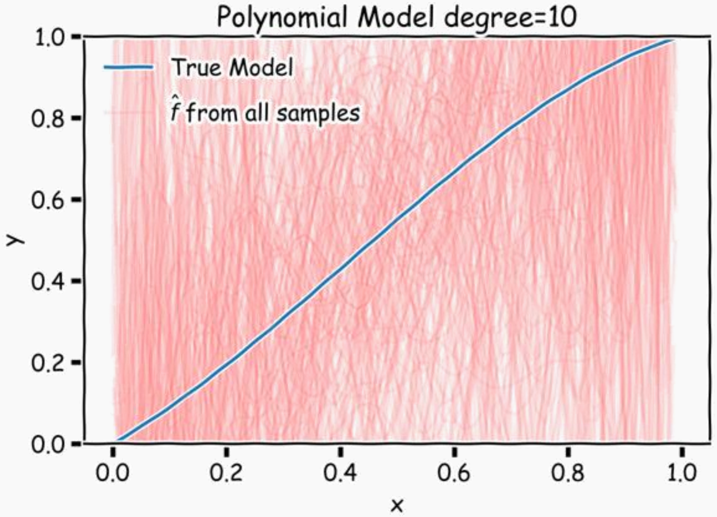


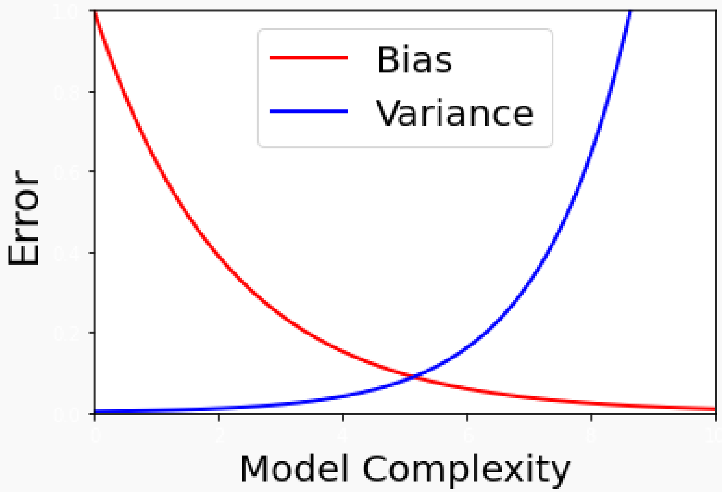




Decrease complexity

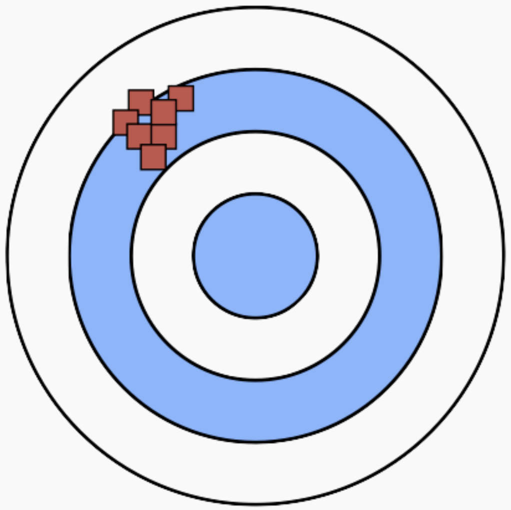


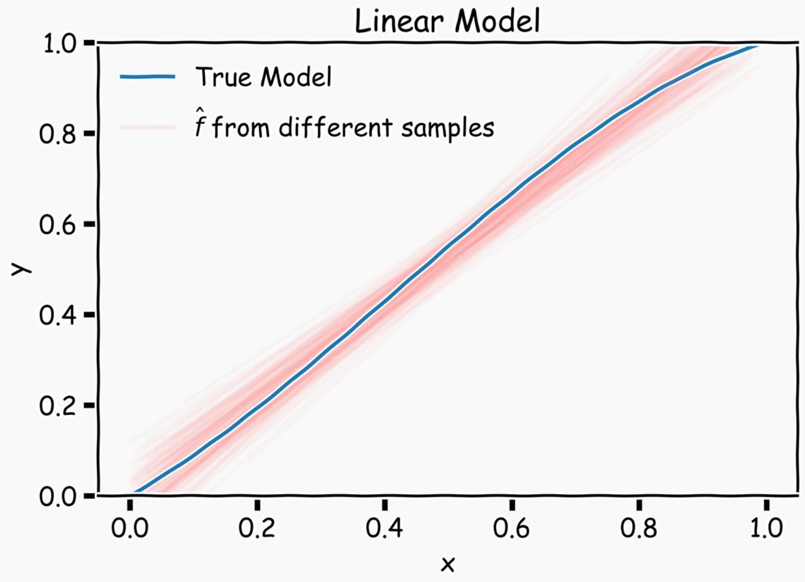




Low Variance
(Precise)

High Bias
(Not Accurate)

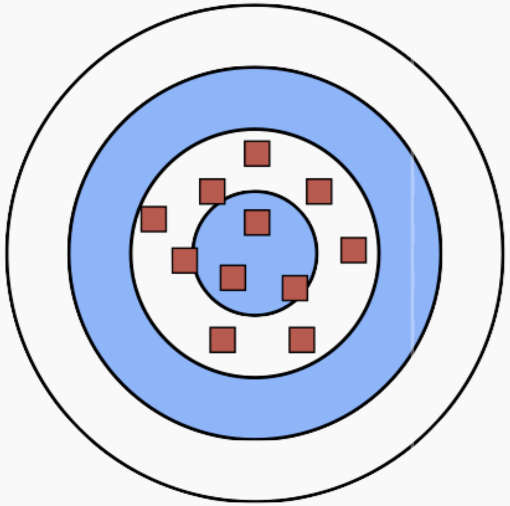


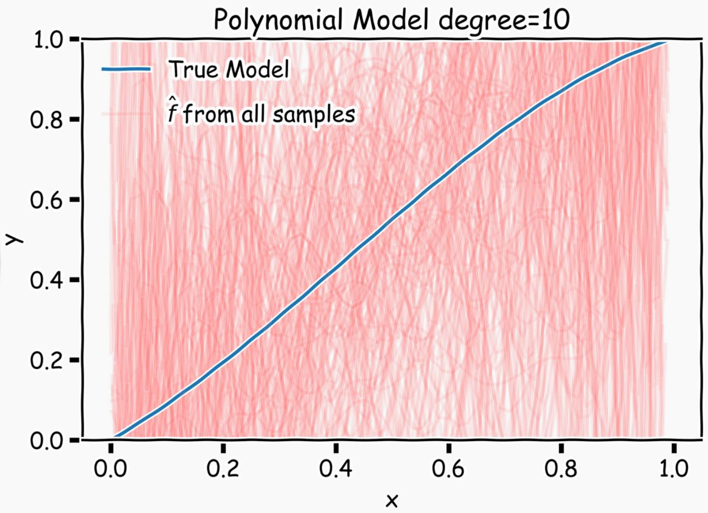


Low Variance
(Precise)

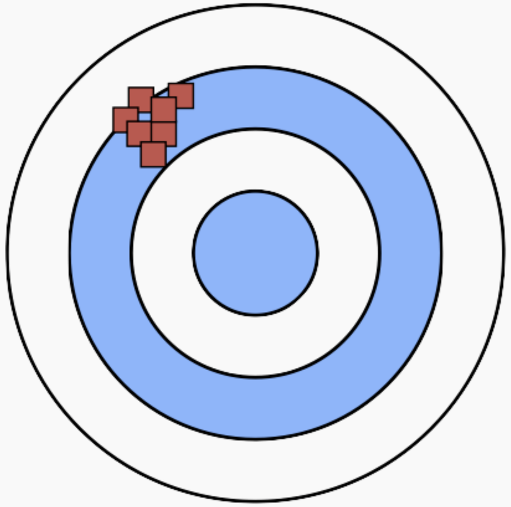
High Variance
(Not Precise)

Low Bias
(Accurate)

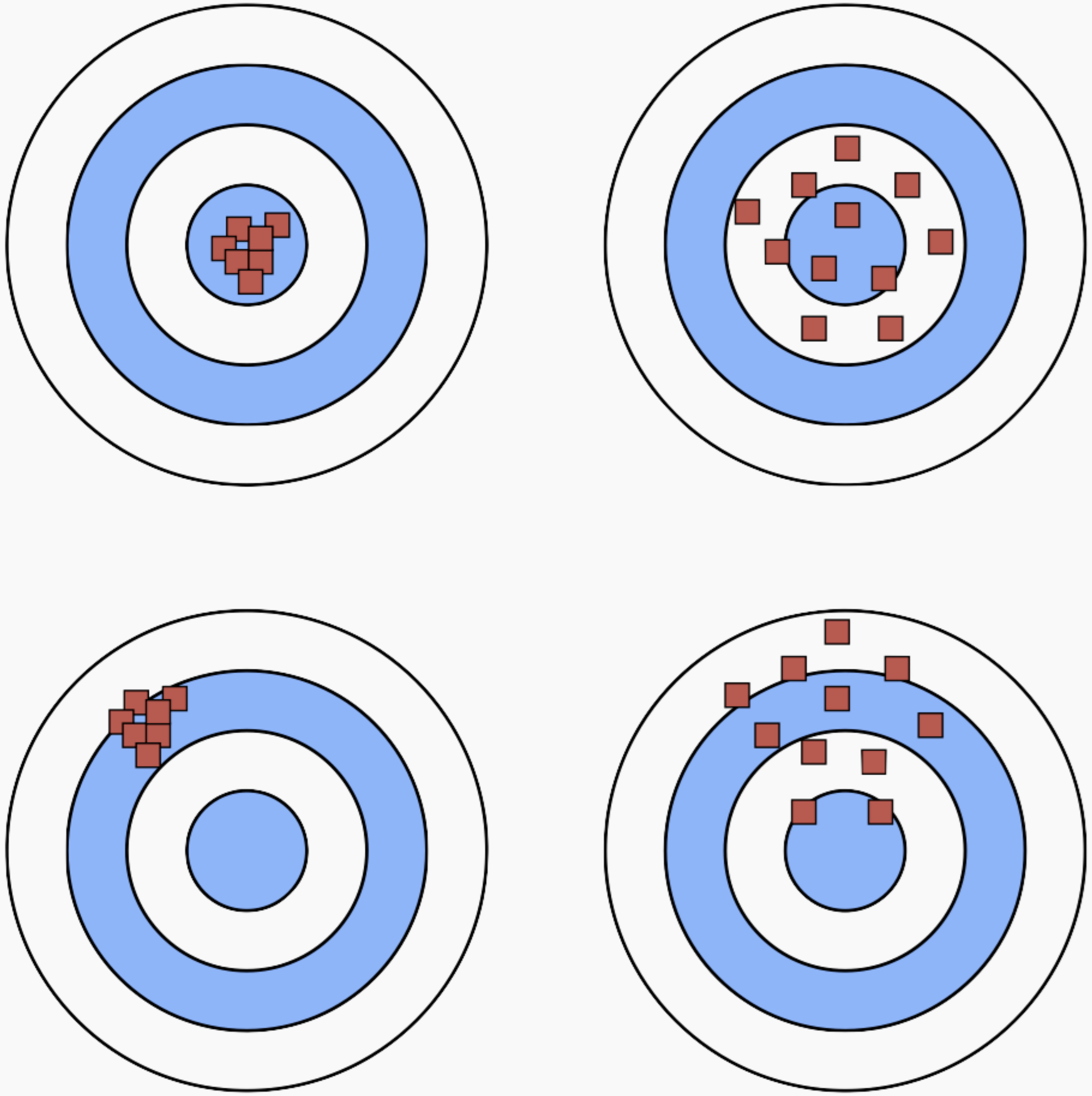




High Bias
(Not Accurate)



WE WANT
THIS !!!



WE WANT TO
AVOID THIS !!

Recall - Overfitting

Overfitting happens when a model learns the training data too well, making it perform poorly on new data.

So far, we have seen that overfitting can happen when:

- too many parameters
- the degree of the polynomial is too large
- too many interaction terms

Soon, we will see other evidence of overfitting, which will point to a way of avoiding overfitting: **Ridge and Lasso regressions**.