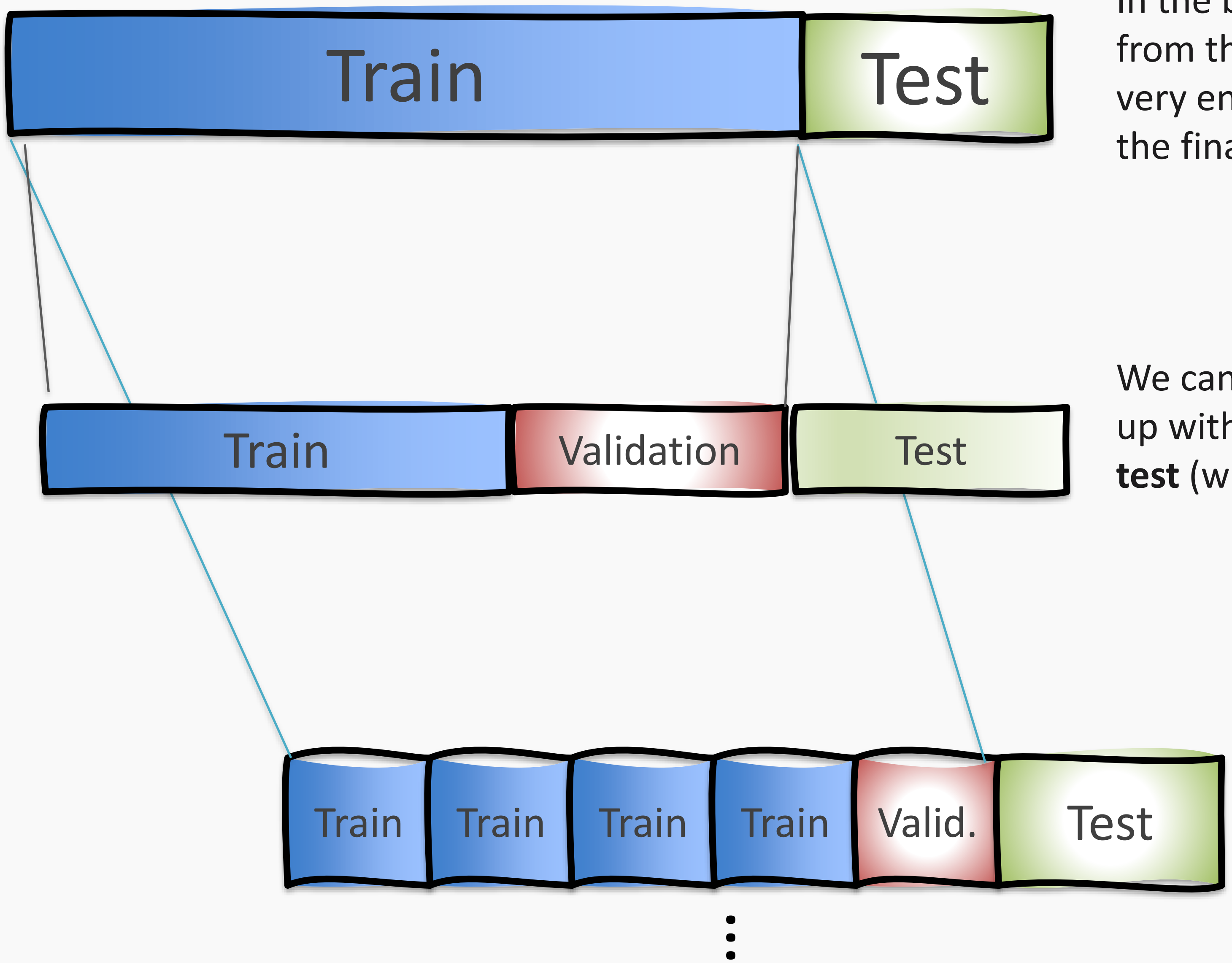
Generalization Error and Bias Variance Tradeoff

Pavlos Protopapas, Ignacio Becker

Outline

- Q&A
- Generalization Error, Bias Variance Tradeoff
- Regularization
    - Lasso and Ridge



In the beginning, we always separate a portion of the data from the main dataset, which we never touch until the very end when we want to evaluate the performance of the final model. Normally, this is called train + test split. *

We can split train data into train + validation, essentially ending up with train + validation (both used to find the best model) + **test** (which we still don’t use until the very end).

And then, we sometimes also use cross-validation, which has nothing to do with either test or validation splits? Because cross-validation uses the train data to split it into k buckets.

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

Top 10 questions about scaling

1. What if we scale X and Y?

$$Y' = \lambda Y$$
$$X' = \lambda X$$

$$\beta' = (X'^T X')^{-1} X'^T Y' \rightarrow \beta' = \left(X^T X \quad\right)^{-1} X^T Y \quad = \beta$$

2. What if we scale only X?

$$X' = \lambda X$$

$$\beta' = (X'^T X')^{-1} X'^T Y' \rightarrow \beta' = \left(X^T X \quad\right)^{-1} X^T Y \quad /\lambda = \beta/\lambda$$

Top 10 questions about scaling

3. What if we scale each predictor by different amount?

$$\beta_i{}' = \beta_i/\lambda_i$$

4. What are the different ways of scaling?

Normalization:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Standardization

$$x' = \frac{x - \bar{x}}{\sigma}$$

Top 10 questions about scaling

5. Will scaling change the performance of the model?

NO for linear models, e.g. linear regression, poly-regression

YES for non-linear models such as kNN or Ridge or Lasso

6. Will scaling change the interpretation of the coefficients?

YES. Since coefficients change, then the relative importance will change

Top 10 questions about scaling

7. Do we scale categorical variables?

Assuming all categorical variables are one-hot-encoded,

- normalization will not change anything – variable are already between 0 and 1
- Standardization makes no sense, no need to do it and will mess up the interpretation

Top 10 questions about scaling

8. How do we scale in sk-learn?

```
from sklearn.preprocessing import StandardScaler
x_transformed_train = StandardScaler().fit_transform(x_train)

from sklearn.preprocessing import MinMaxScaler
x_transformed_train = MinMaxScaler().fit_transform(x_train)
```

Top 10 questions about scaling

9. When to use normalization and when to use standardization?

normalization: when the scale of the variables matter and there are no outliers

standardization: when the scale is not important and when you suspect you have outliers

Top 10 questions about scaling

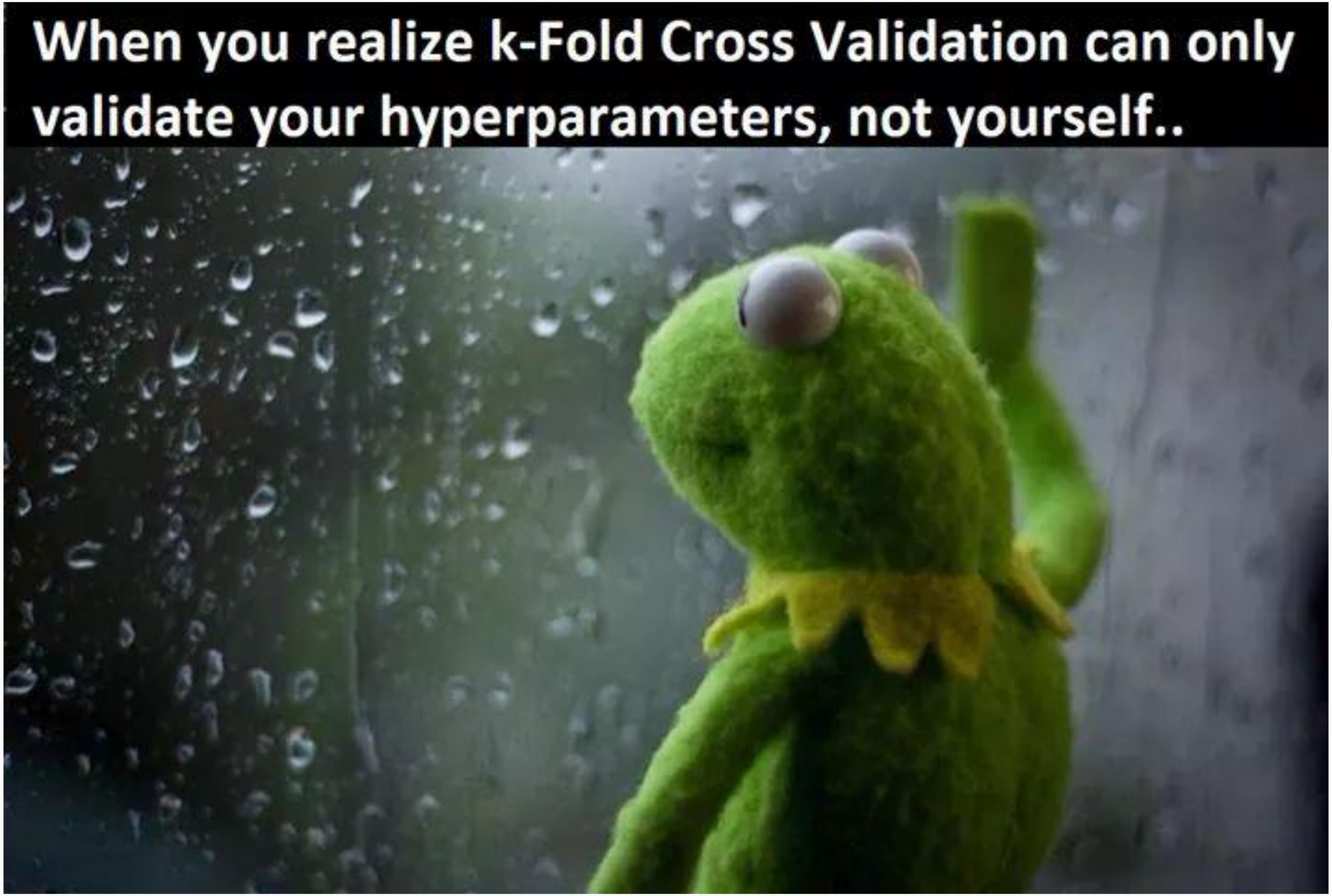
10. Shall I scale before or after creating features in poly regression.

It is all about the dynamic range. You need to make sure the max number and min number are within the range of your machine.

Top 10 questions about scaling

10. Shall I scale before or after creating features in poly regression.

It is all about the dynamic range. You need to make sure the max number and min number are within the range of your machine.


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

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**

Outline

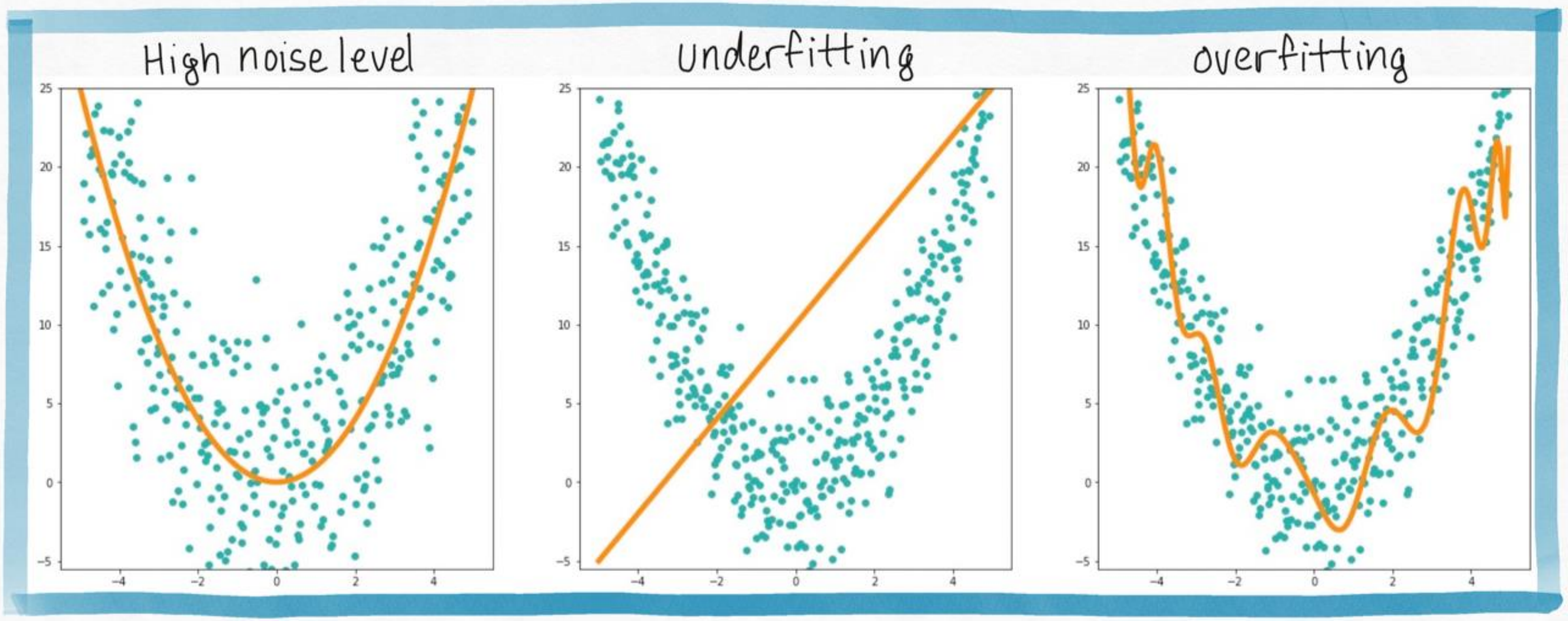
- Q&A
- **Generalization Error, Bias Variance Tradeoff**
- Regularization
    - Lasso and Ridge

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.



Irreducible and Reducible Errors

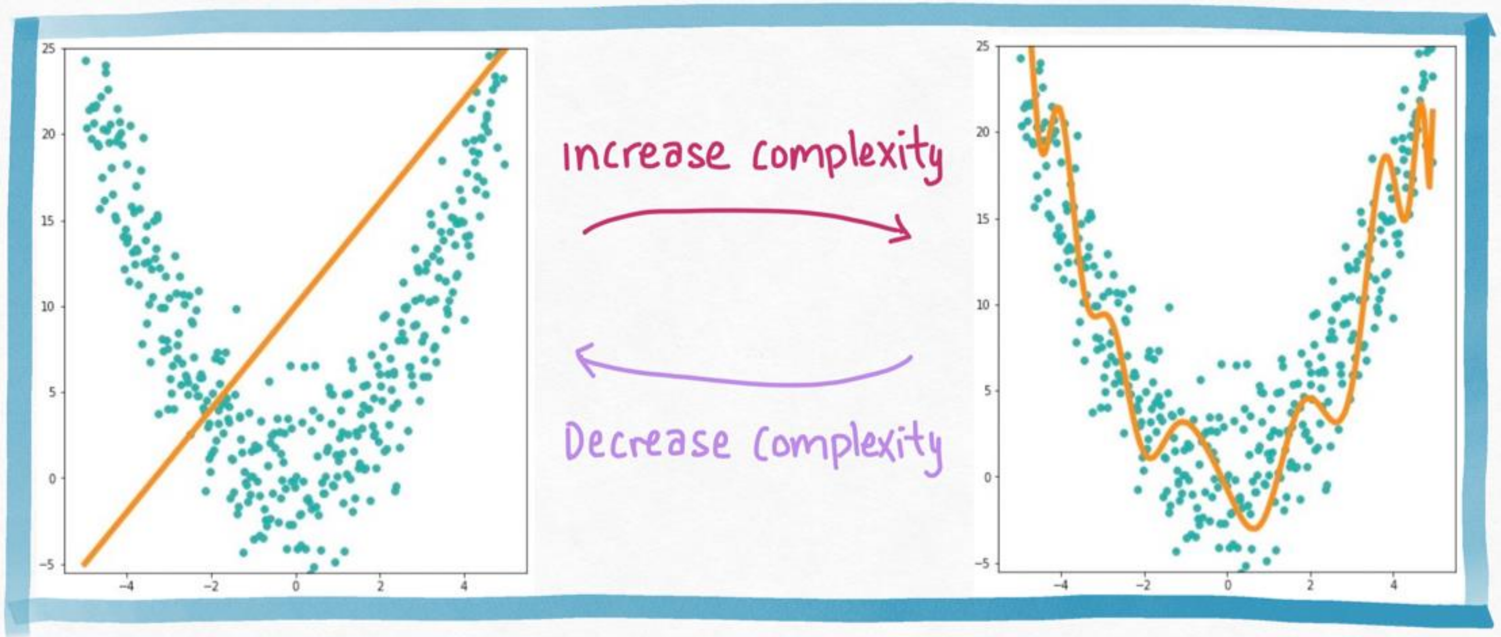
We distinguished the contributions of noise to the generalization error:

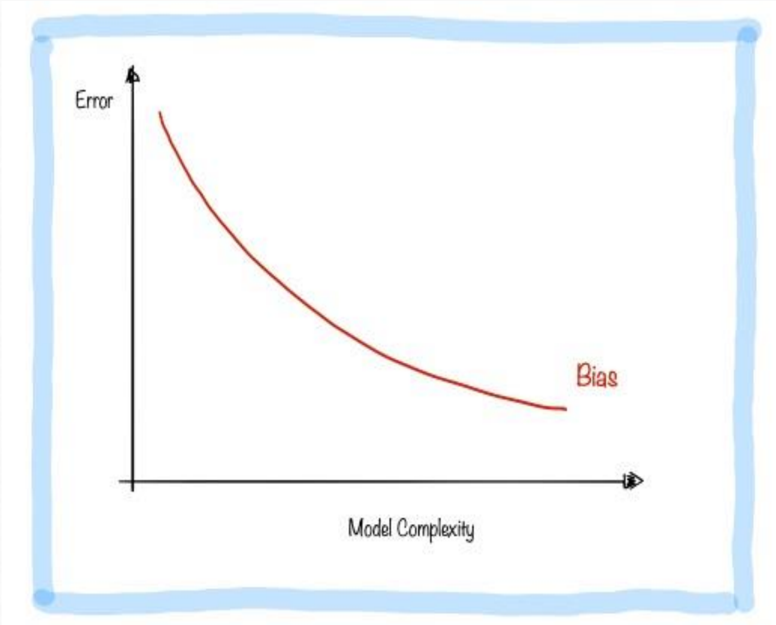
Irreducible error (or aleatoric error): we can’t do anything to decrease the error due to noise.

Reducible error (or epistemic error): we can decrease the error due to overfitting and underfitting by improving the model.

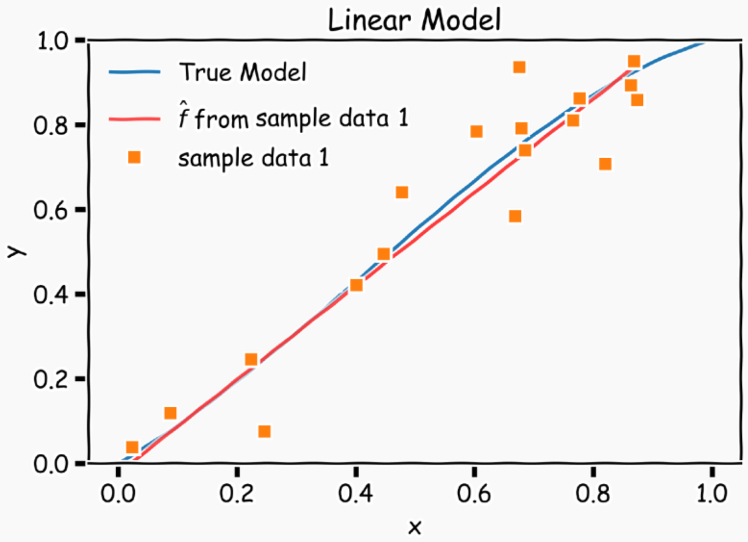
The Bias-Variance: Bias

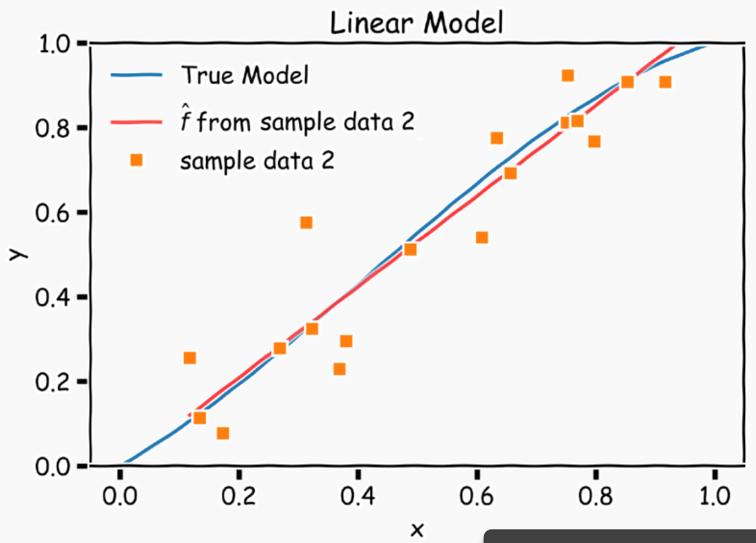
Reducible error comes from either underfitting or overfitting. There is a tradeoff between the two sources of errors:

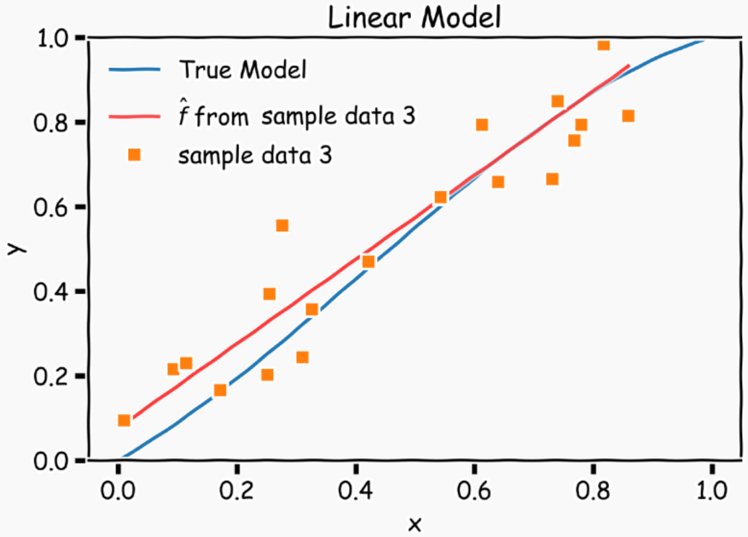


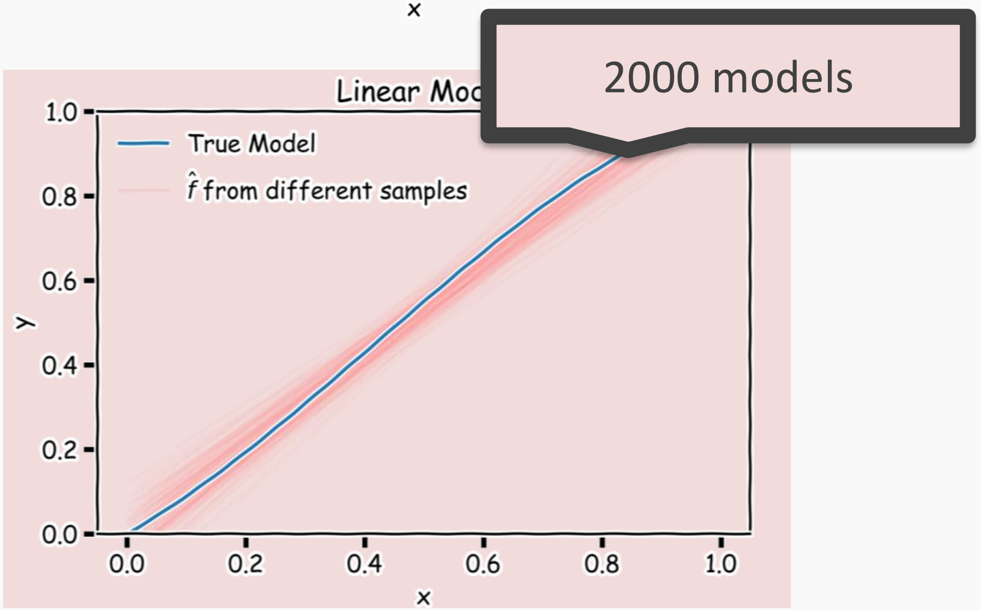


Bias vs Variance: Variance of a SIMPLE model



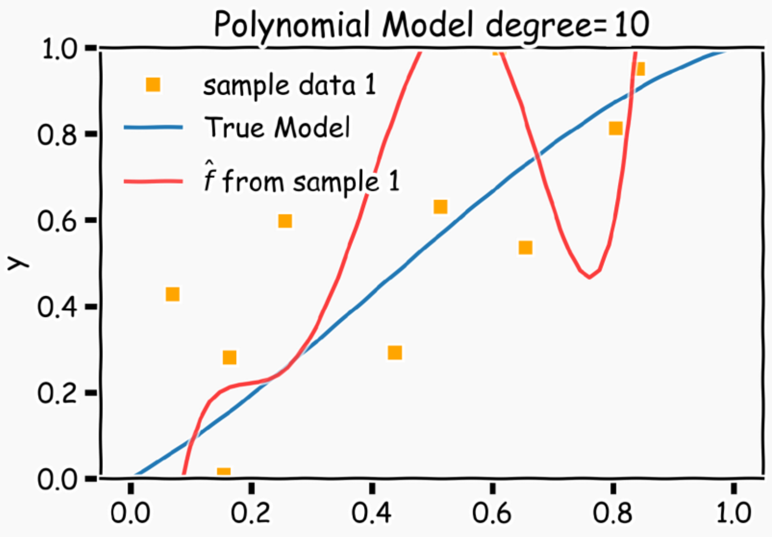


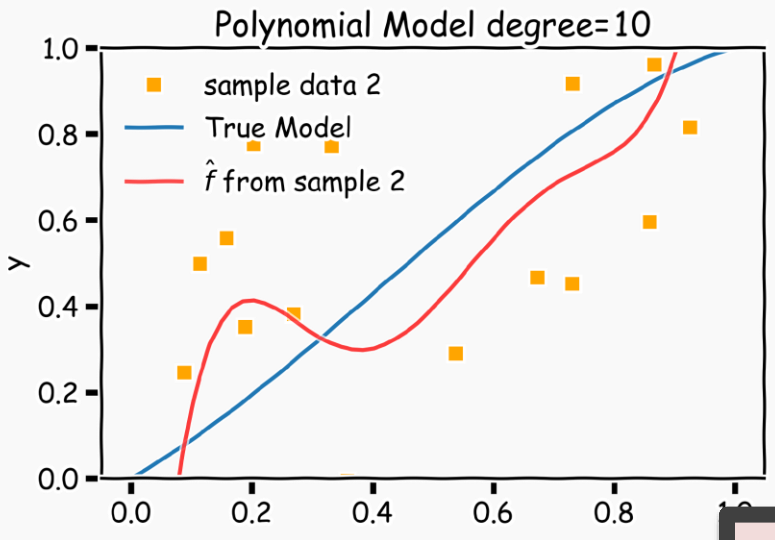


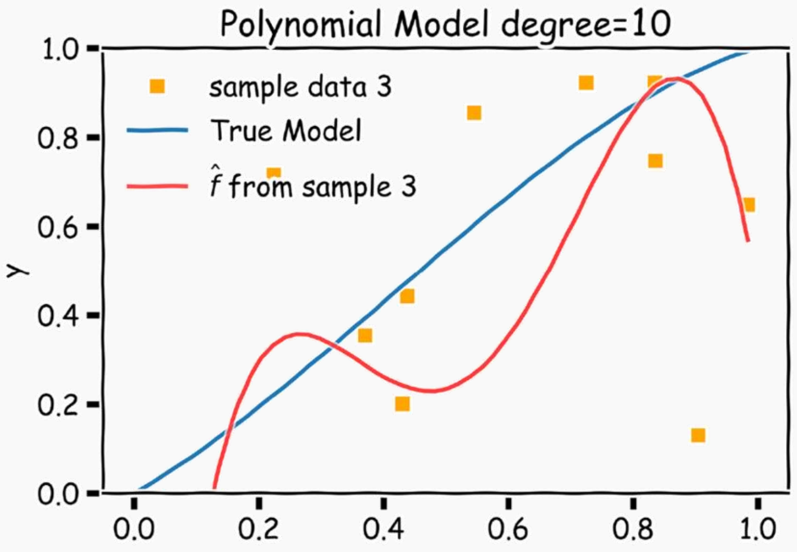


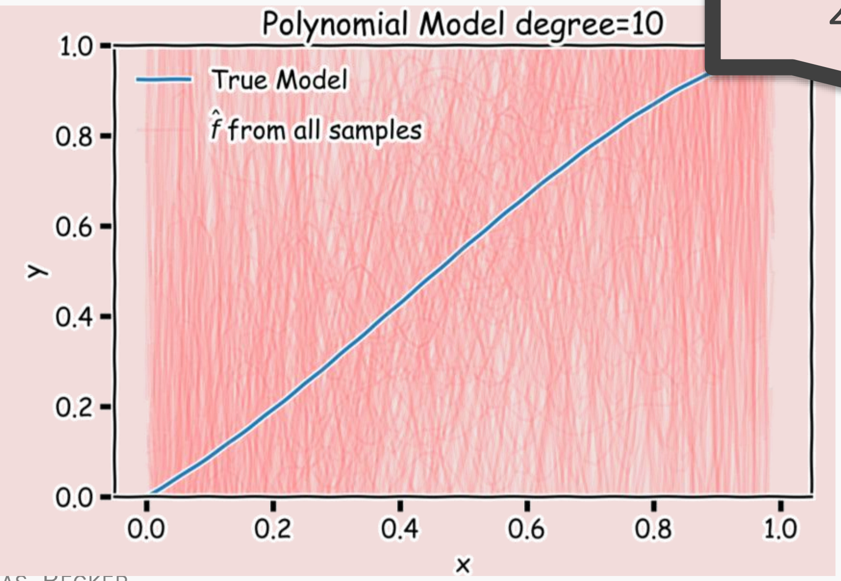
2000 models

Bias vs Variance: Variance of a COMPLEX model







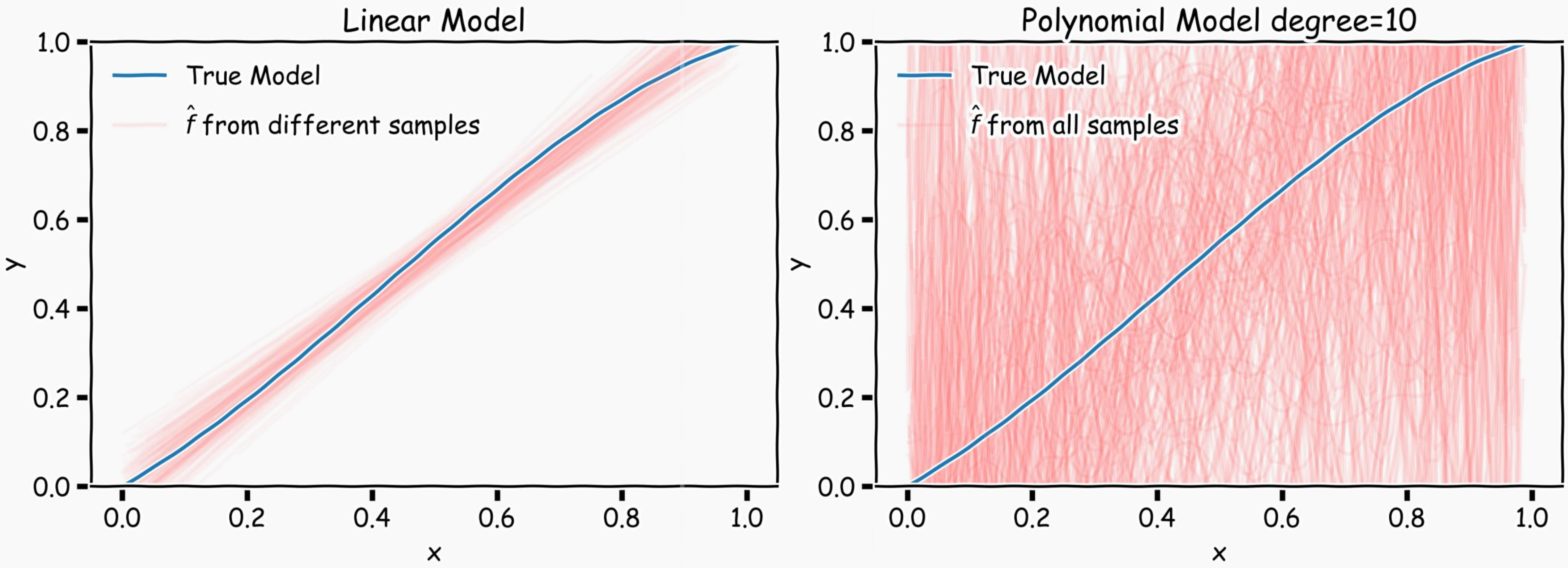


2000 models

Bias vs Variance

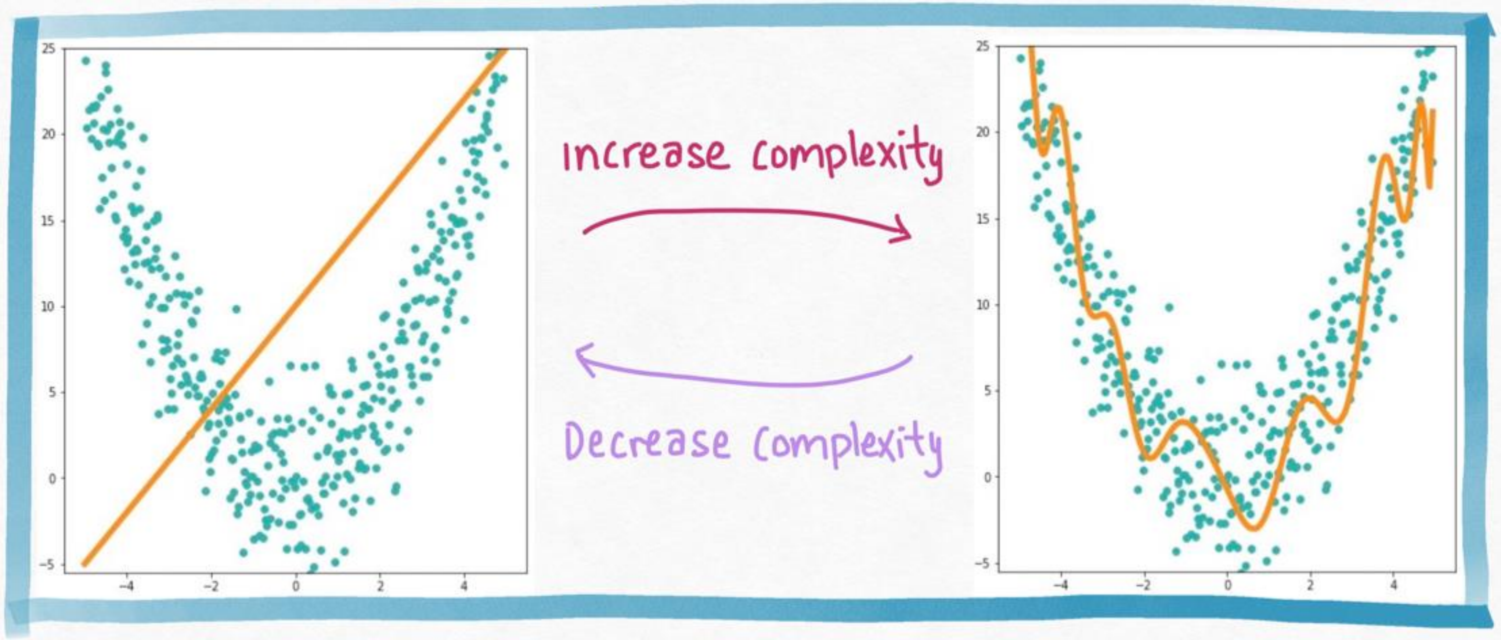
Left: 2000 best fit linear models, each fitted on a different 20-point training set.

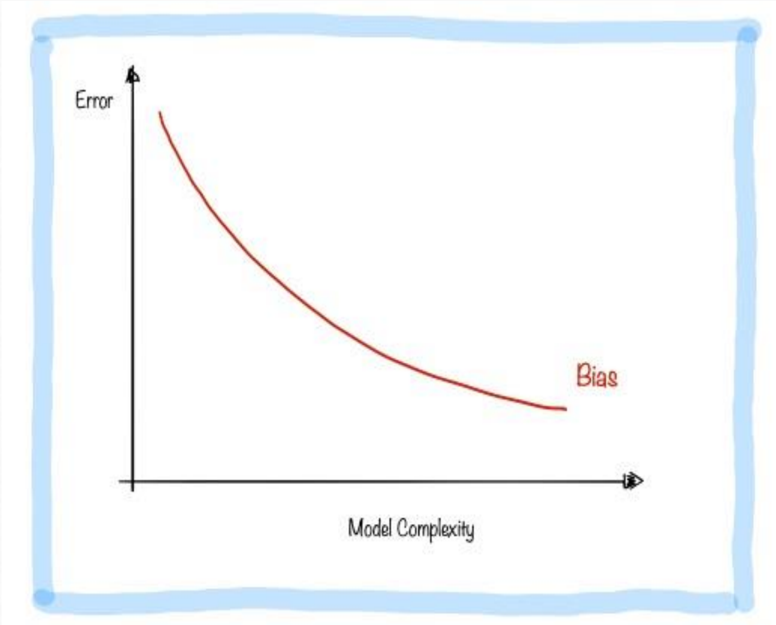
Right: 2000 best fit models using degree 10 polynomials.



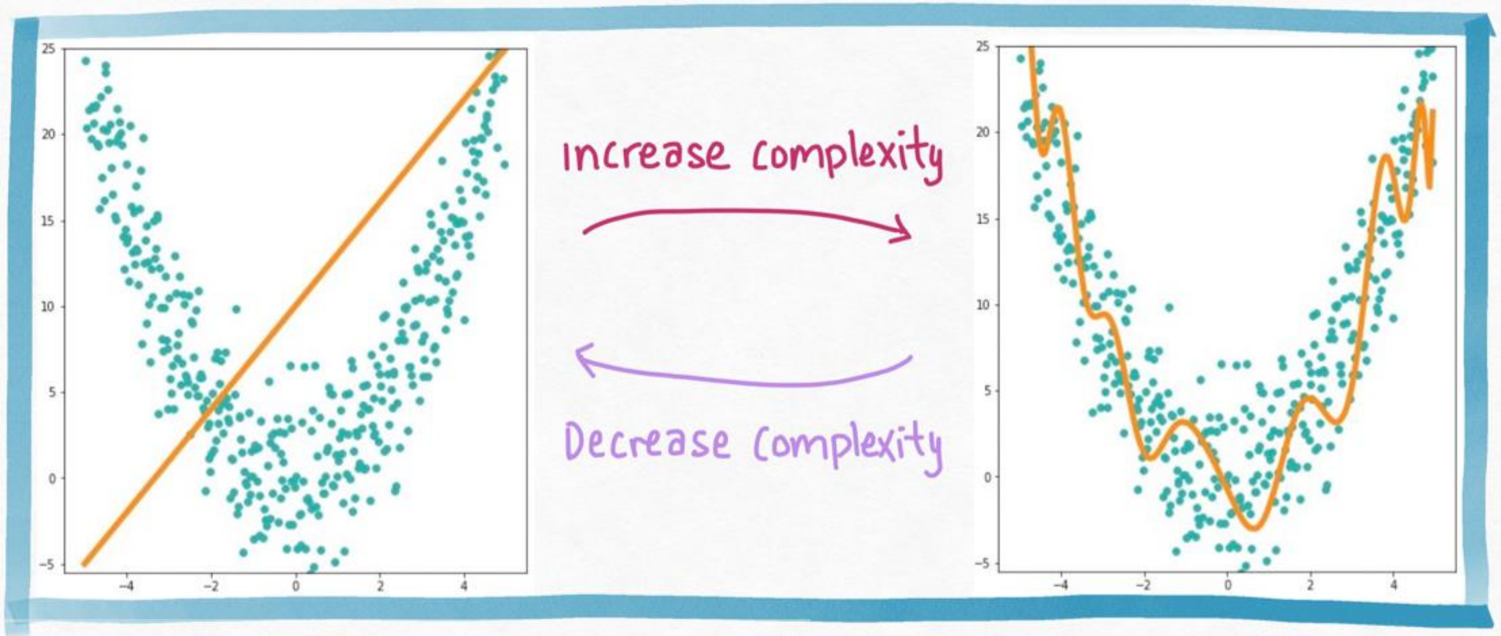
The Bias-Variance: Bias

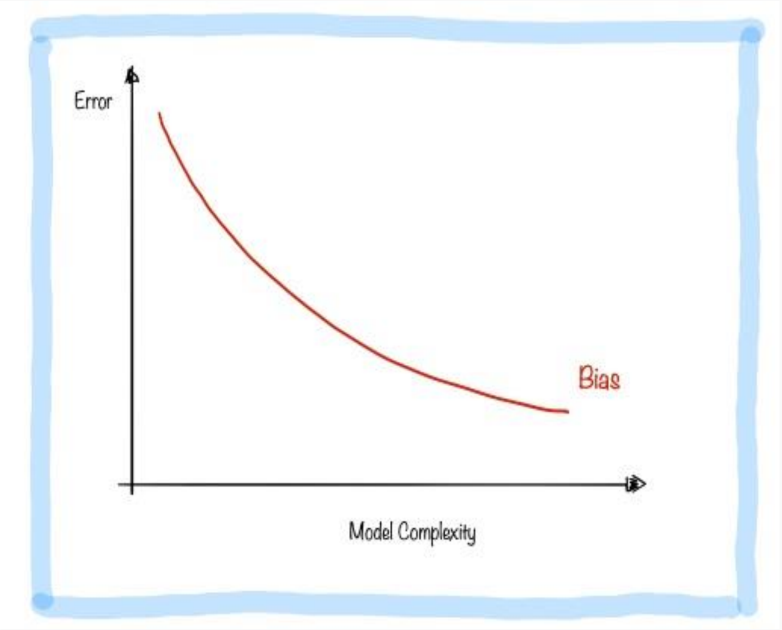
Reducible error comes from either underfitting or overfitting. There is a tradeoff between the two sources of errors:

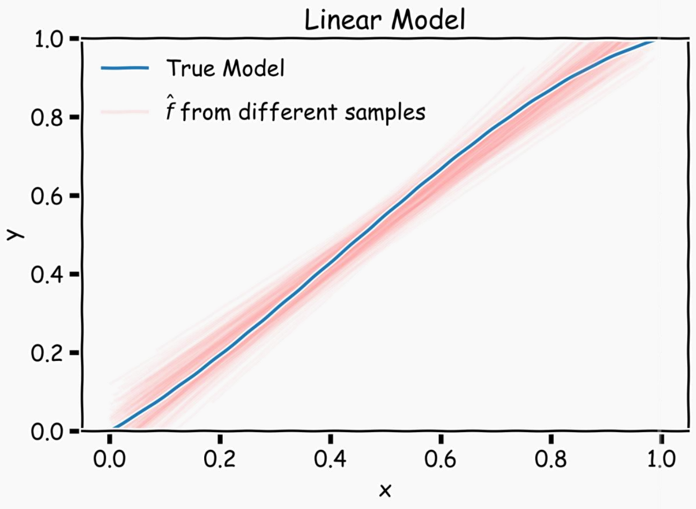


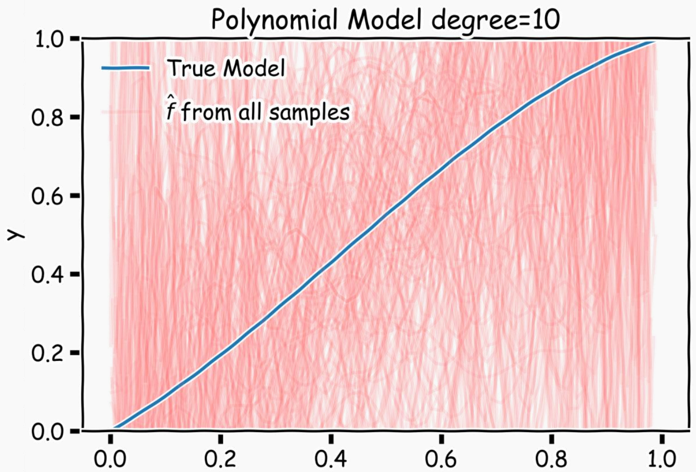


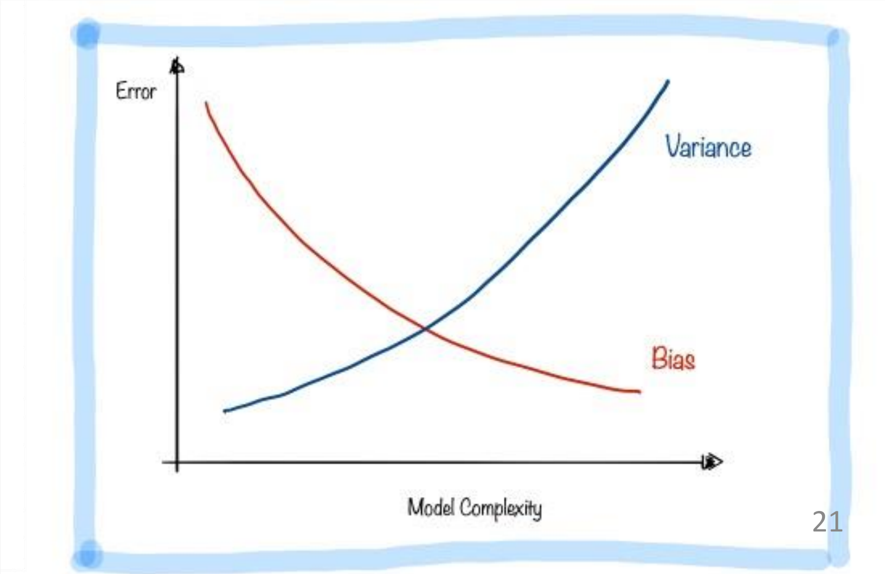
The Bias-Variance Trade Off





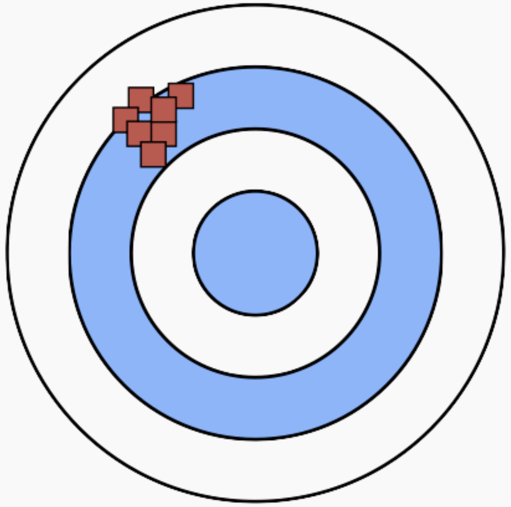


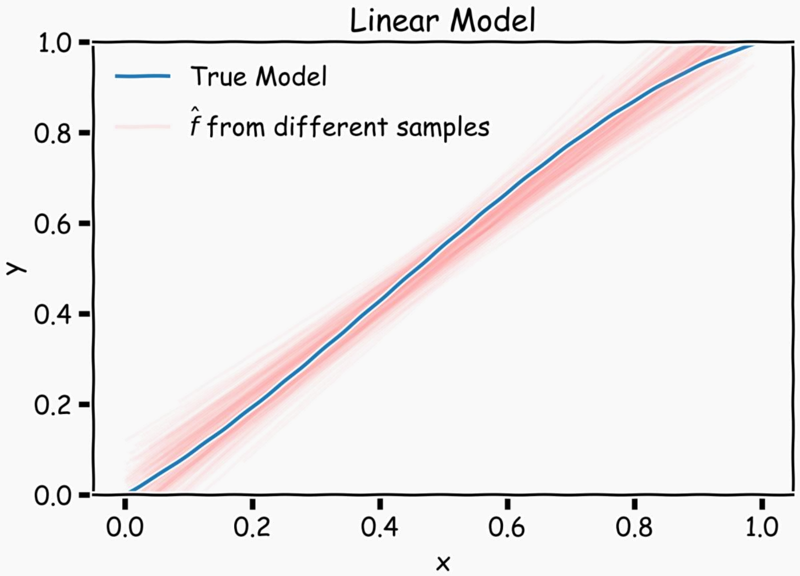




Low Variance
(Precise)

High Bias
(Not Accurate)

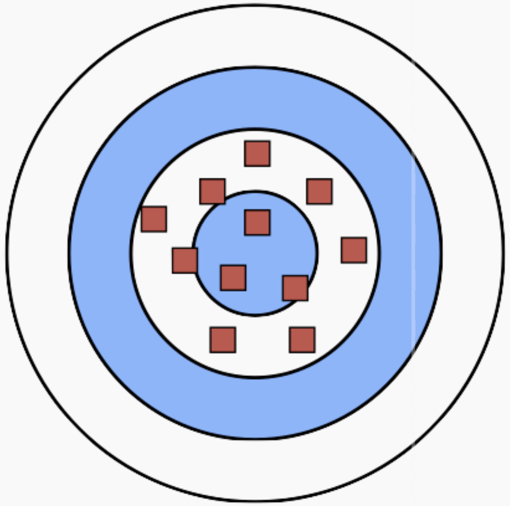


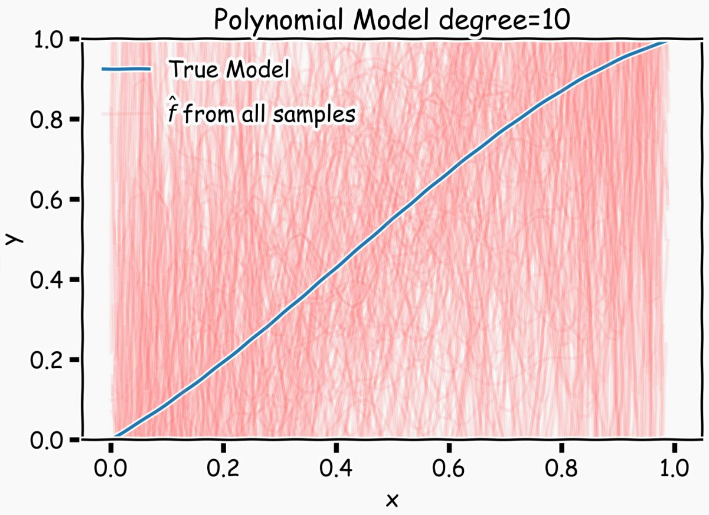


Low Variance
(Precise)

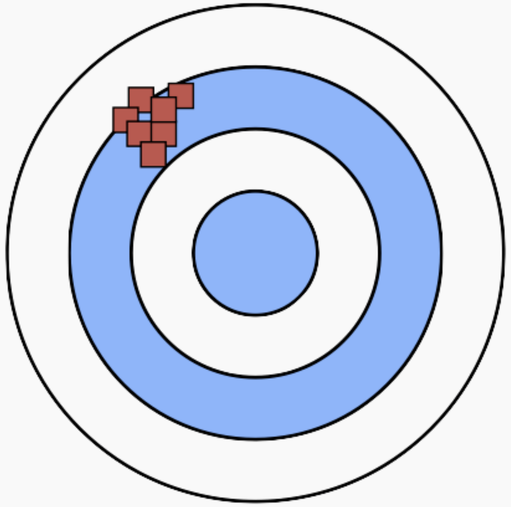
High Variance
(Not Precise)

Low Bias
(Accurate)

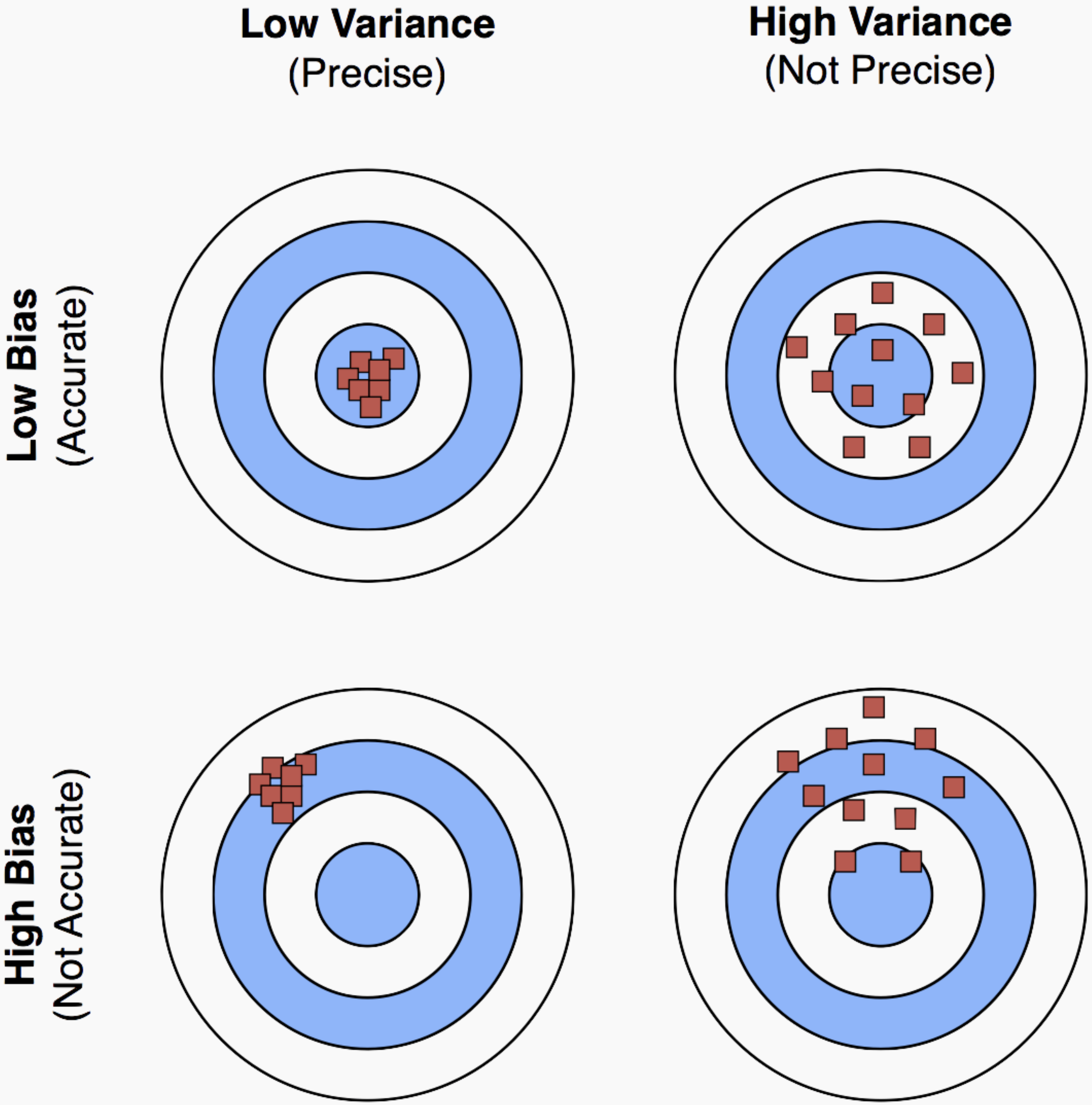




High Bias
(Not Accurate)



WE WANT THIS



WE WANT TO AVOID THIS

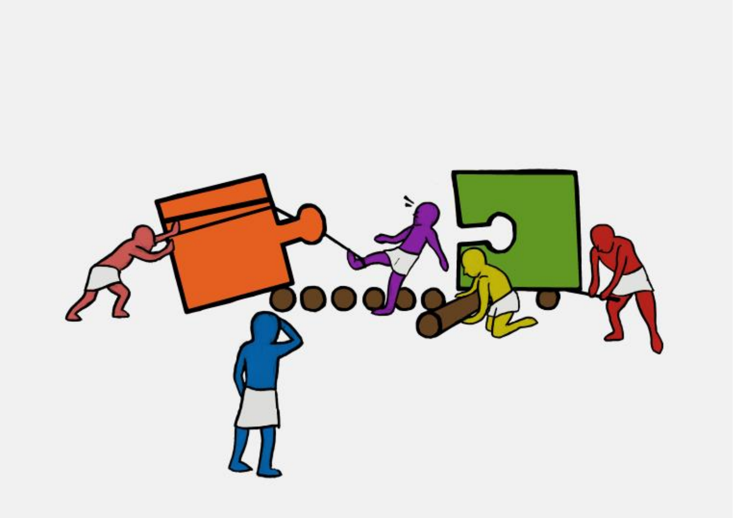
Overfitting

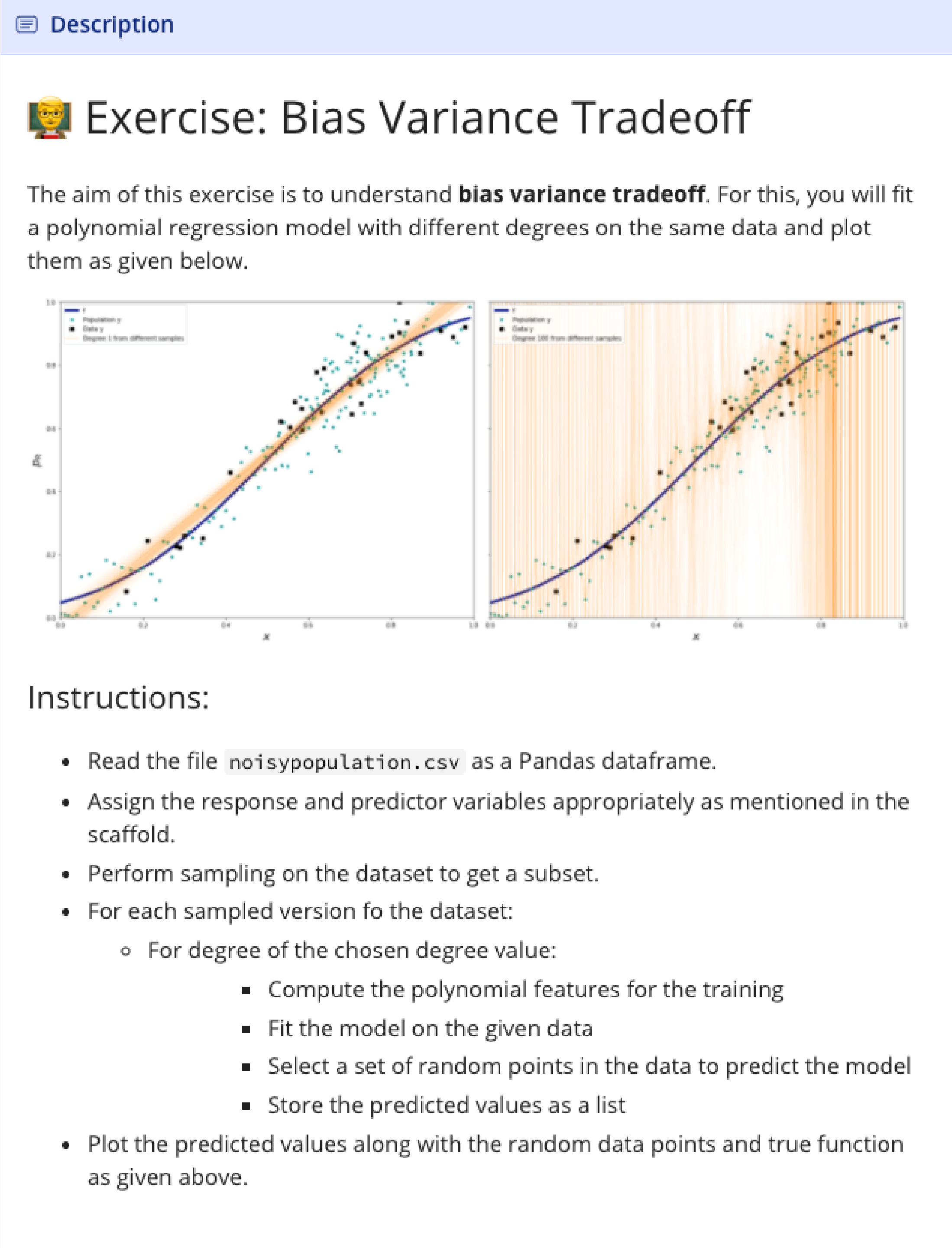
Overfitting occurs when a model corresponds too closely to the training set, and as a result, the model fails to fit additional data.

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.





Description

Exercise: Bias Variance Tradeoff

The aim of this exercise is to understand **bias variance tradeoff**. For this, you will fit a polynomial regression model with different degrees on the same data and plot them as given below.

f
Population y
Data y
Degree 1 from different samples
f
Population y
Data y
Degree 100 from different samples
x
x

Instructions:

- Read the file `noisypopulation.csv` as a Pandas dataframe.
- Assign the response and predictor variables appropriately as mentioned in the scaffold.
- Perform sampling on the dataset to get a subset.
- For each sampled version fo the dataset:
    - For degree of the chosen degree value:
        - Compute the polynomial features for the training
        - Fit the model on the given data
        - Select a set of random points in the data to predict the model
        - Store the predicted values as a list
- Plot the predicted values along with the random data points and true function as given above.