# Session 11: Bias/variance

wifi: GA-Guest, yellowpencil

GENERAL ASSEMBLY

```
cd ~/Documents/ga-ldn-ds37
git commit -am "your commit message here"
git pull
```

# Today's session plan

| | |
|---|---|
| **1800-1900** | Standup & review |
| **1900-1920** | Break |
| **1920-2000** | Exploring bias and variance in models |
| **2000-2100** | Training and testing splits |
| **US Presidential election exercise** | |

# At the end of the session, you will be able to ...

**Understand** bias and variance in models

**Use** k-fold cross validation to make more efficient use of data

**Describe** the meaning of underfitting and overfitting

Roland Berger x GA

# How good is your model?

# Bias and variance

A model's testing error consists of two main components; bias and variance.

A model with high **bias** doesn't describe the data very well.

A model with high **variance** is very sensitive to changes in the training data.

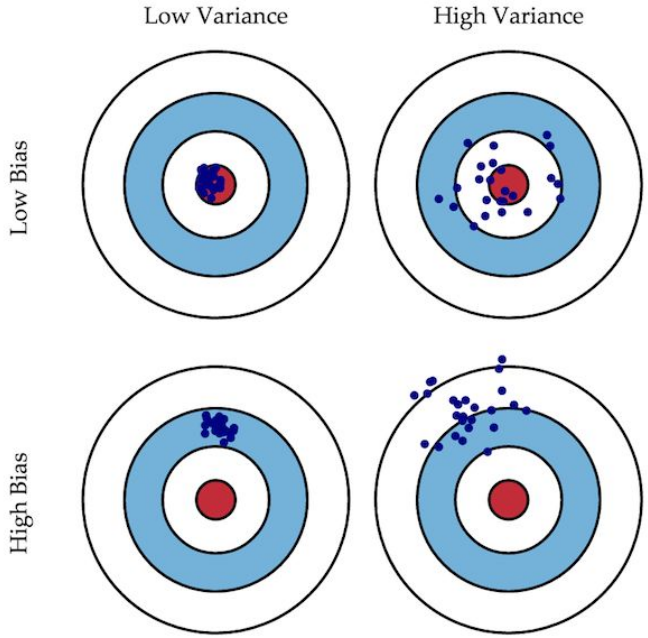



Fig. 1 Graphical illustration of bias and variance.

# Definitions

**Bias** Error that results from the correct value and the predicted value within our model.

Roughly, whether or not our model aims on target.

**Variance** Error that results from the variability of a model prediction for a given data point.

Roughly, whether or not our model is reliable.

**Group Exercise:**

# Are Complex Models Always Better?

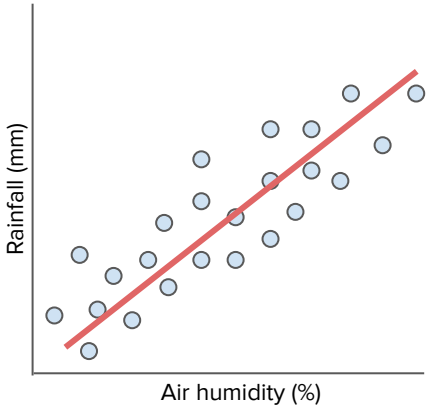
**10 minutes**

Imagine you're fitting a model to predict the rainfall on a given day (mm) using the air humidity (%). Your training set consists of 25 observations or data points. You try fitting two models to your dataset.

In your groups, discuss which model you would use and why. Which model would perform better on the testing data, and why?

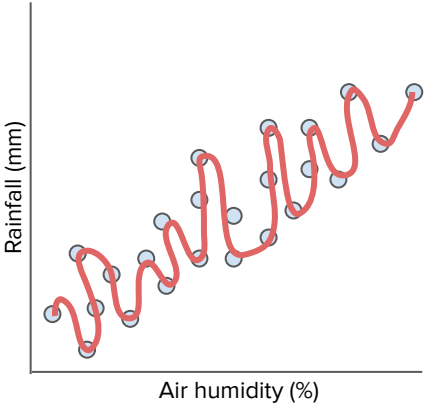
**Model 1**
Low training accuracy (doesn't pass through all the points in our training data)

Low complexity (it's just a straight line)




**Model 2**
High training accuracy (passes through all the points in our training data)

High complexity (it's a fancy squiggly line)



# The bias-variance tradeoff

**Bias: How close are predictions to the actual values?**

- Bias is error stemming from incorrect model assumptions
- If the model cannot represent the data's structure, our predictions could be consistent, but will not be accurate.
- **Example**: Assuming data is linear when it has a more complicated structure.

**Variance: How variable are our predictions?**

- Variance is error stemming from being overly sensitive to changes in the training data.
- A model with high variance will make very different predictions given slightly different training sets.
- **Example**: Using the training set exactly (e.g. 1-NN) for a model results in a completely different model -- even if the training set differs only slightly.

As model complexity **increases**:

- Bias **decreases**. (The model can more accurately model complex structure in data.)
- Variance **increases**. (The model identifies more complex structures, making it more sensitive to small changes in the training data.)

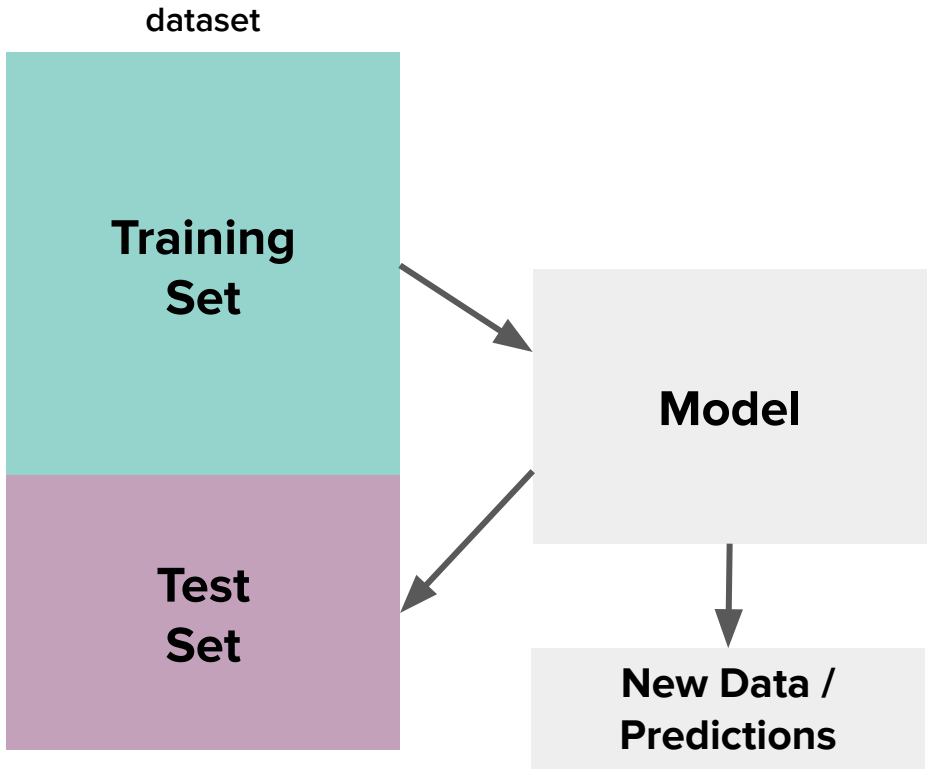
 

# Training vs testing accuracy

We will typically calculate the accuracy (or **error**) of a model using the sum of squared residuals, or the root mean squared error.

The **training** error doesn't tell us about how well our model will predict unseen data, only how well it fits our training data.

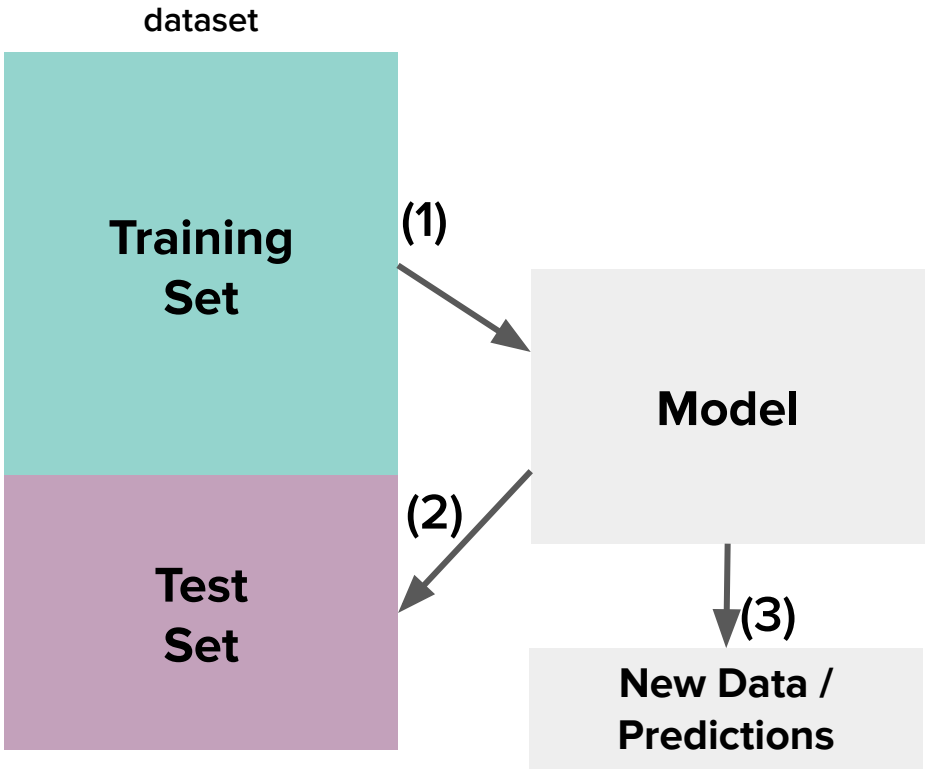
The **testing** error tells us how well our model generalises.



# Machine Learning Validation

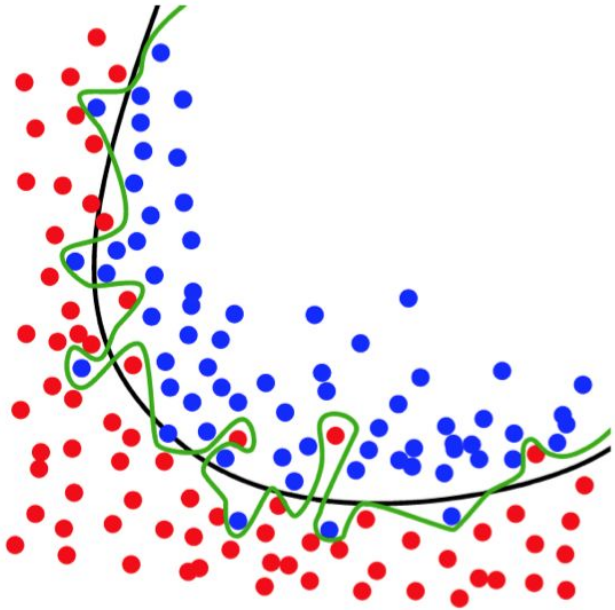
There are three sources of error:

1. Training error
2. Generalization error
3. Out-of-sample error



# Underfitting vs Overfitting



**Minimizing training error does not minimize generalization error!**

1. What is the training accuracy of the green line model to the left? The black line model?

2. Is the green line a better model or the black line?

**Solo Exercise:**

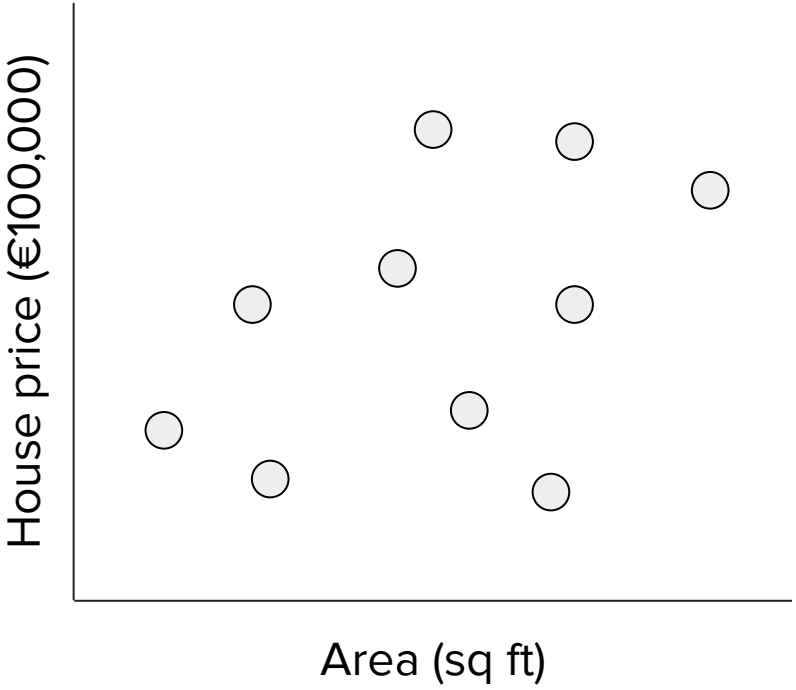
# Bias and variance

Think about the following models. Relative to each other, do they have:

High or low bias? Why?
High or low variance? Why?

**Linear regression**
**High-order polynomial**

Sketch out what the models might look like on paper, on a dataset like this:



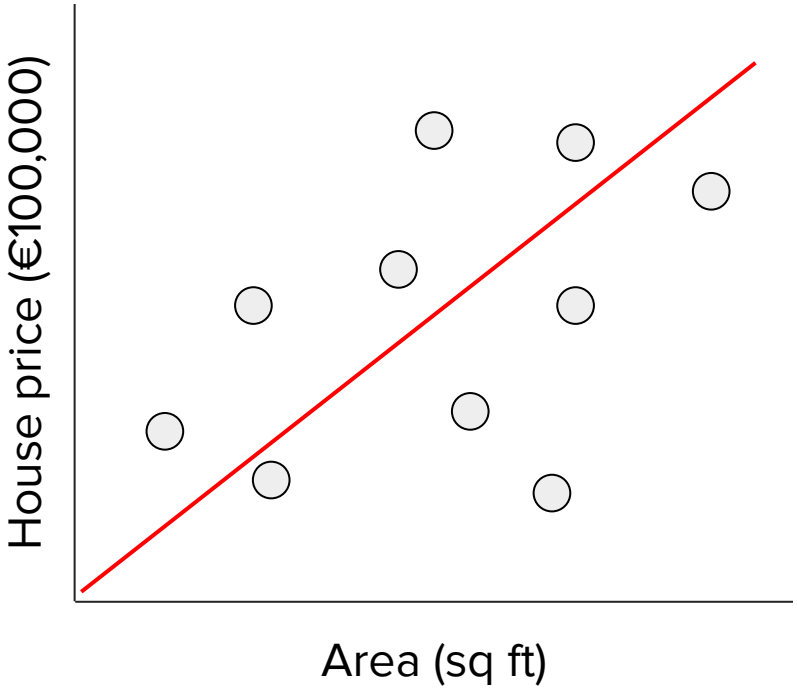
# Bias and variance

## Linear regression

**Low variance, High bias.**

If we train with a different subset of the training set, the model will be about the same. Hence, the model has low **variance**.

The resulting model will predict the training points incorrectly (unless they happen to be perfectly linear). Hence, it has high **bias**.



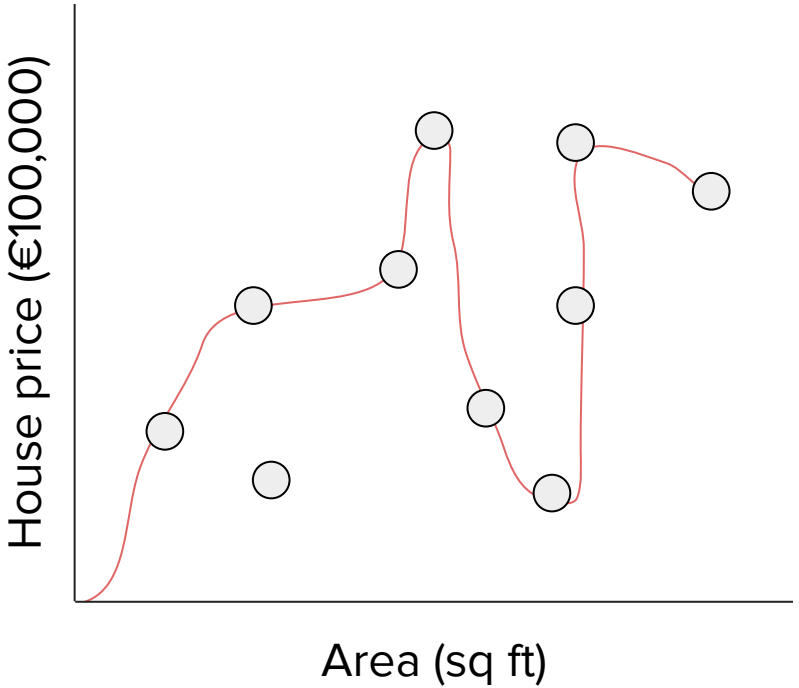
# Bias and variance

## High order polynomial

**High variance, low bias.**

A more complex polynominal will represent the underlying structure of the data slightly better than linear regression (assuming the underlying relationship is not linear) so will have low **bias**.

If the degree of the polynomial is very high, it will be extremely sensitive to changes in the training data/will be more prone to fitting to outliers and noise, and will therefore have a high **variance**.



**Computers Out:**

# Exercises

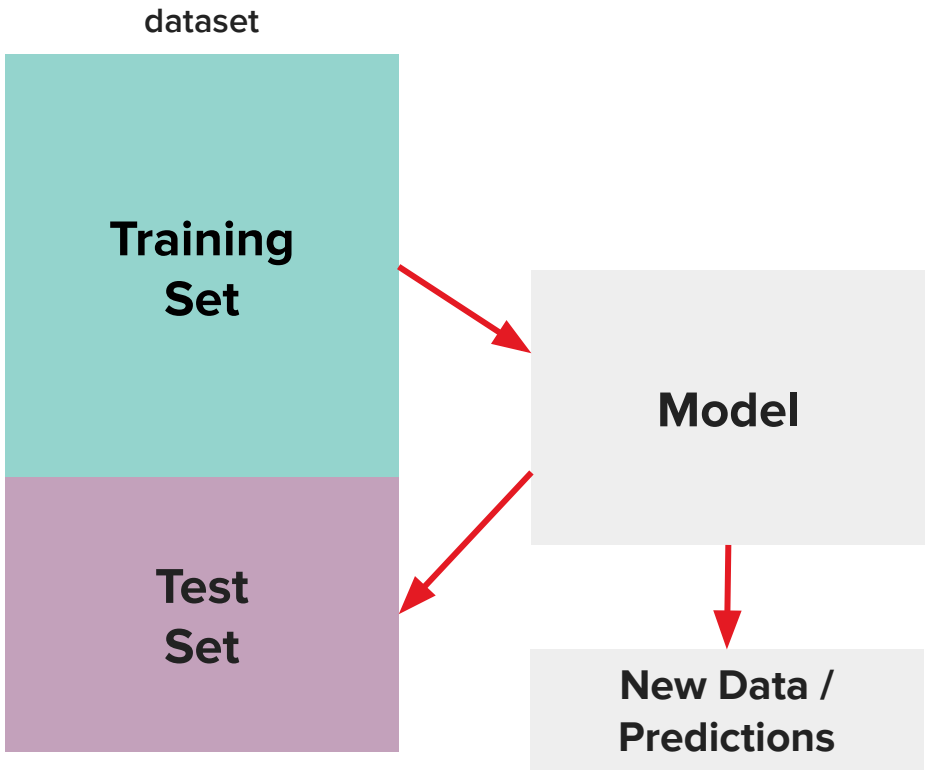
Open ds37-11-01.ipynb and let's start working through some examples

Roland Berger x GA

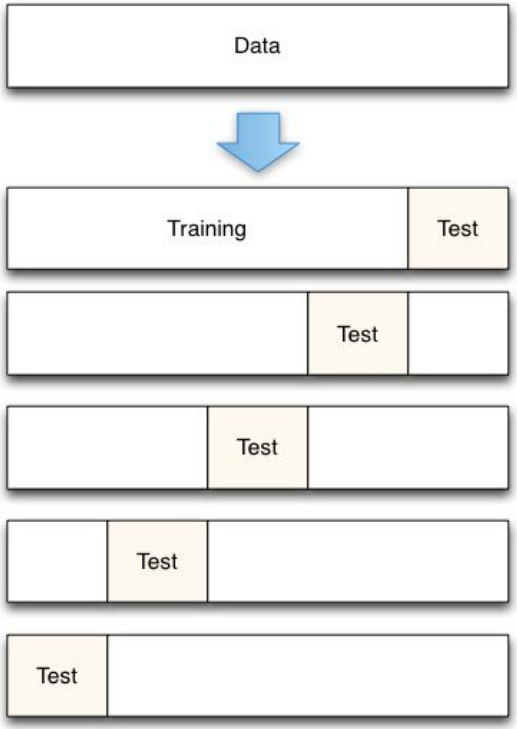
# Training and testing

# Training and testing split

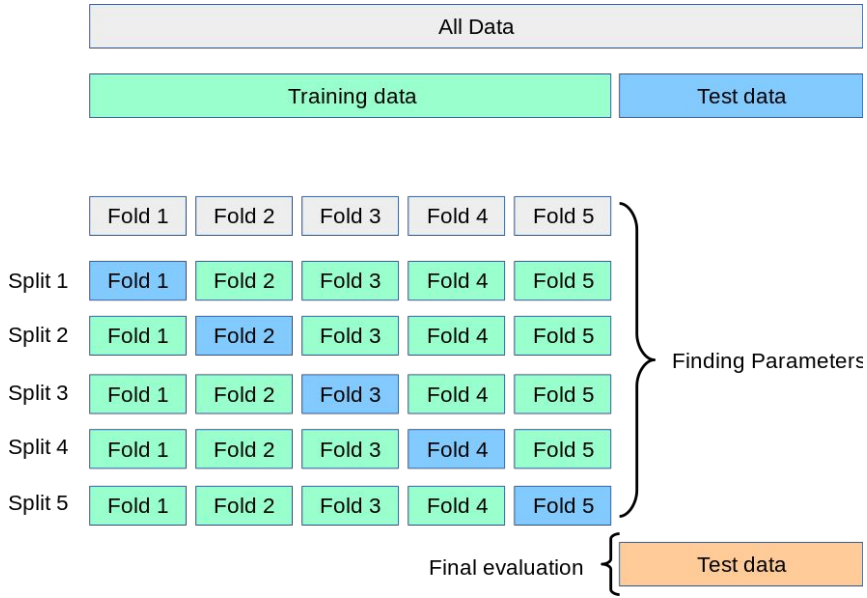


# K-fold cross validation



# Leave one out cross validation



Intro to Python

# Let's Review

# At the end of the session, you will be able to ...

**Understand** bias and variance in models

**Use** k-fold cross validation to make more efficient use of data

**Describe** the meaning of underfitting and overfitting

# Coming up next time...

- KNN classifiers


GA