GENERAL ASSEMBLY

DATA SCIENCE

11 WEEK PART TIME COURSE

Week 3 - Model Evaluation
Wednesday 6th January 2016

1. Motivation
2. How do we know if we have a good model?
3. What is the Bias-Variance tradeoff?
4. Lab
5. Homework Review

WHAT IS A GOOD MODEL?

Q: What's wrong with training error?

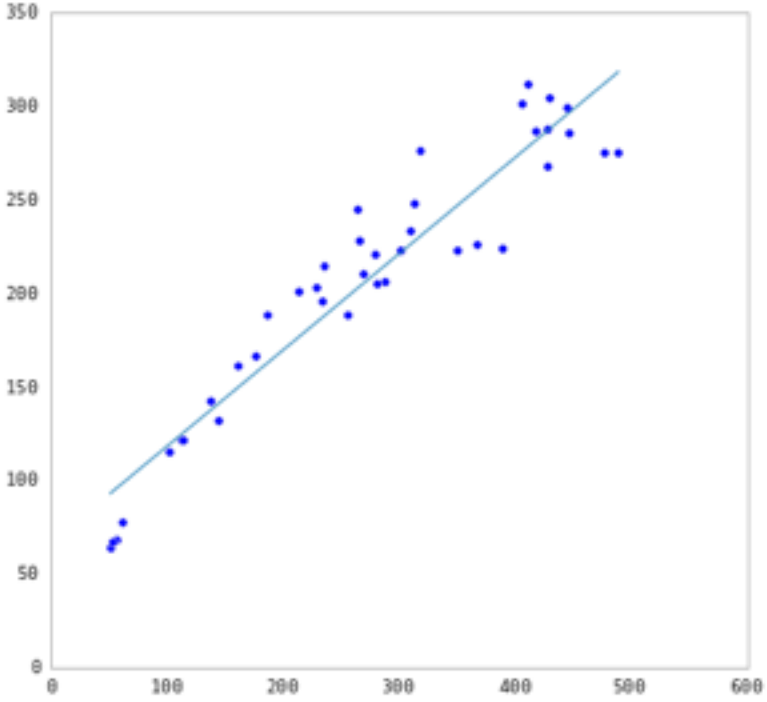
Thought experiment:

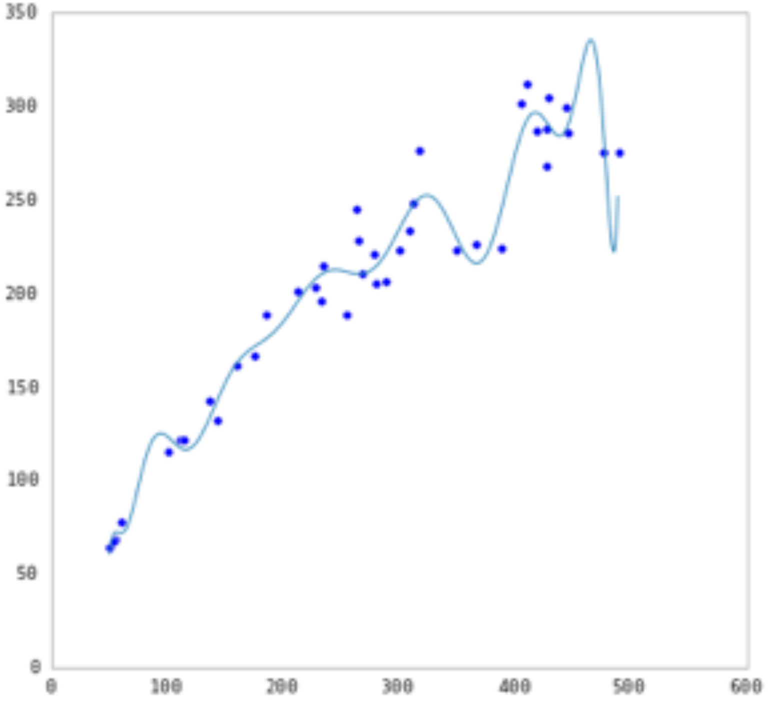
Suppose we train our model using the entire dataset.

Q: How low can we push the training error?
- We can make the model arbitrarily complex (effectively "memorizing" the entire training set).

A: Down to zero!

TRAINING ERROR





Q: What's wrong with training error?

Thought experiment:
Suppose we train our model using the entire dataset.
Q: How low can we push the training error?

- We can make the model arbitrarily complex (effectively "memorizing" the entire training set).

A: Down to zero!

A: Training error is not a good estimate of accuracy beyond training data.

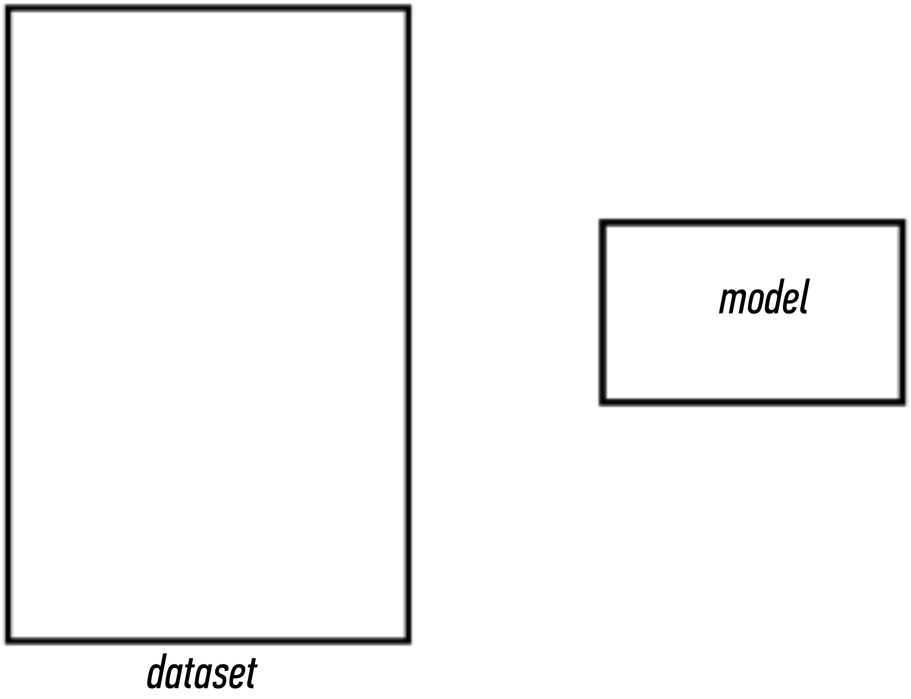
WHY THIS MATTERS

The data that we are given for prediction won't always be the end of the data we are interested in! We may not have access to all the data of interest

We will gather data and build and iterate over models however a main reason for building the model was to predict unseen test cases.

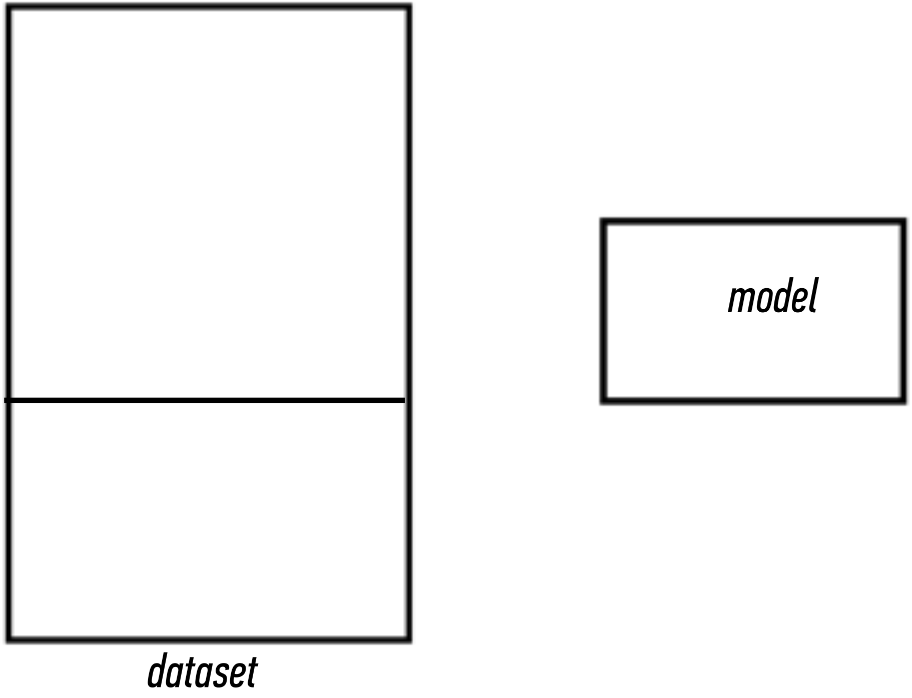
TEST SET APPROACH

Q: How can we make a model that generalizes well?



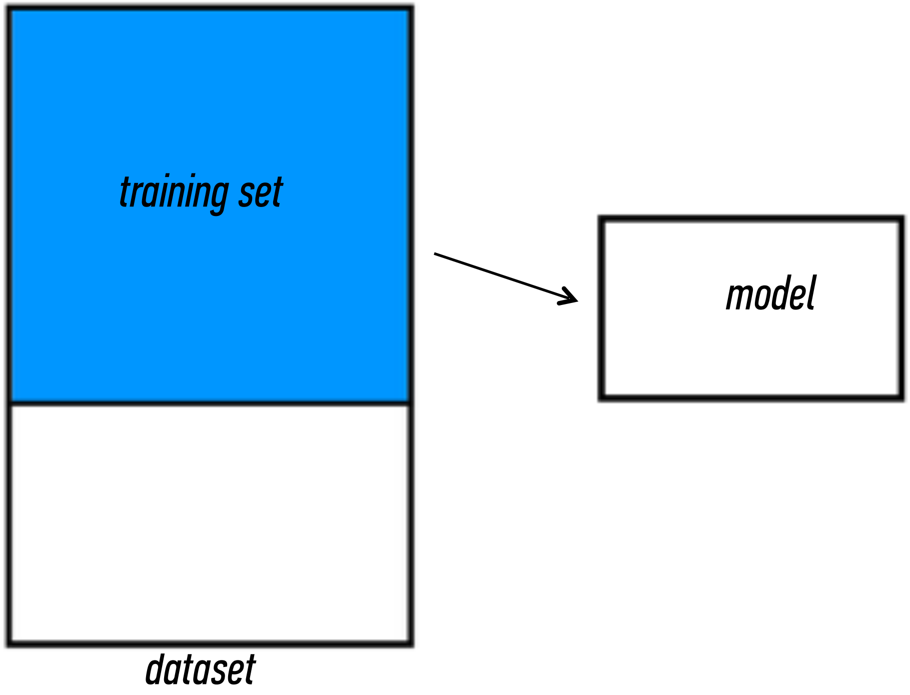
Q: How can we make a model that generalizes well?

1) split dataset



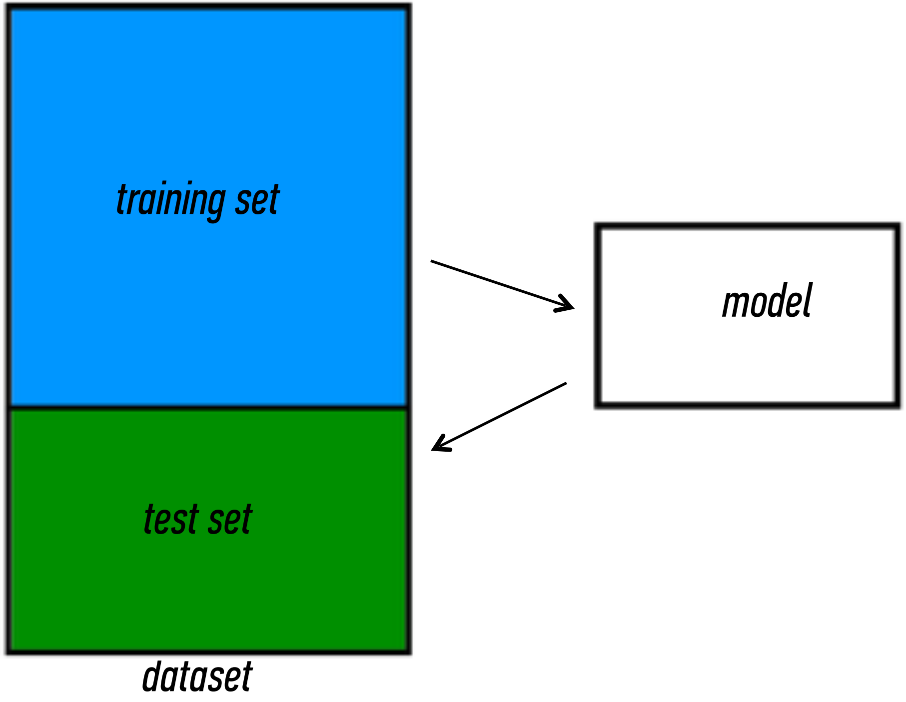
Q: How can we make a model that generalizes well?

1) split dataset
2) train model



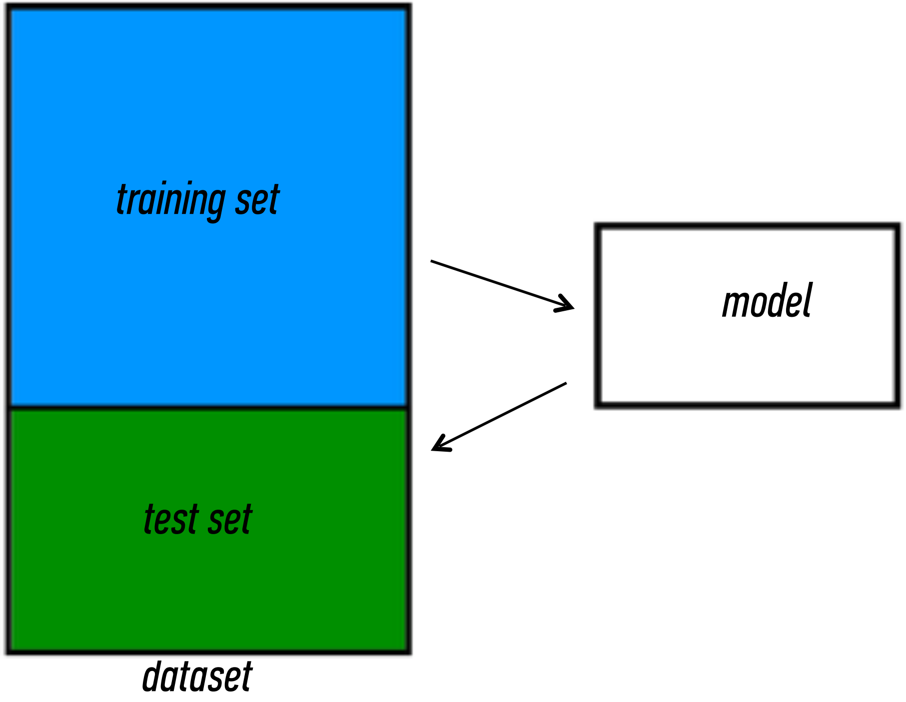
Q: How can we make a model that generalizes well?

1) split dataset
2) train model
3) test model



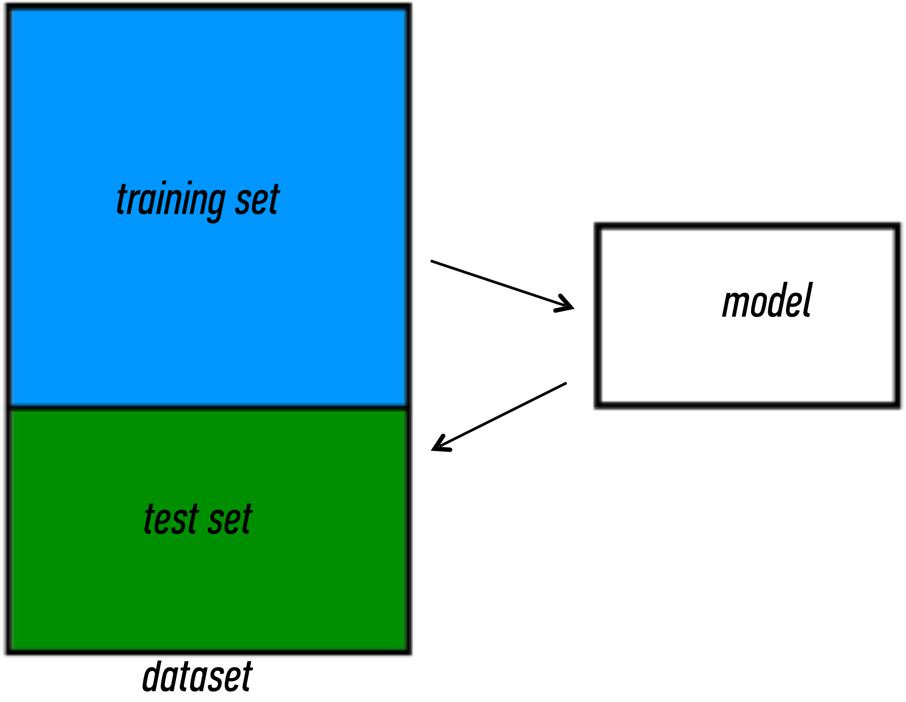
Q: How can we make a model that generalizes well?

1) split dataset
2) train model
3) test model
4) parameter tuning



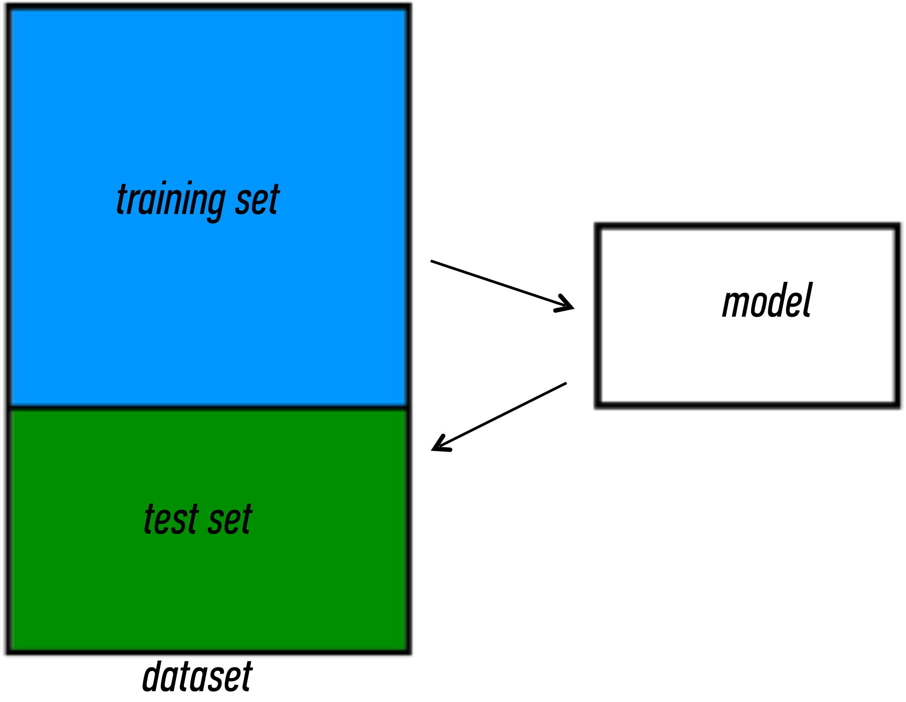
Q: How can we make a model that generalizes well?

1) split dataset
2) train model
3) test model
4) parameter tuning
5) choose best model



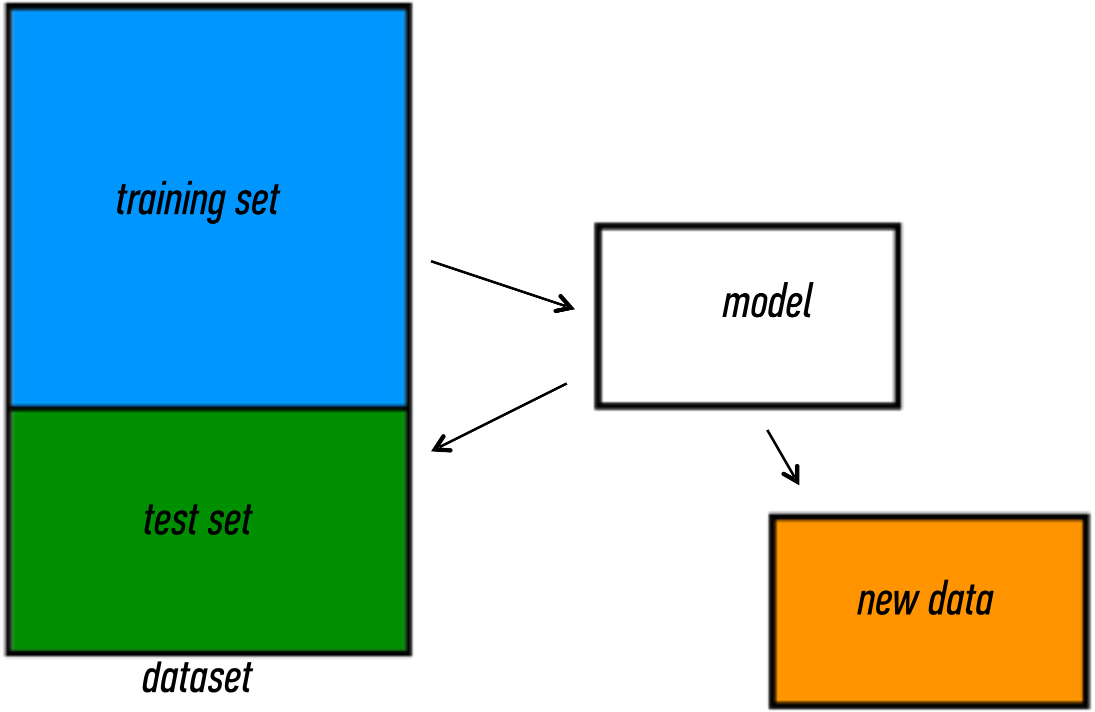
Q: How can we make a model that generalizes well?

1) split dataset
2) train model
3) test model
4) parameter tuning
5) choose best model
6) train on all data



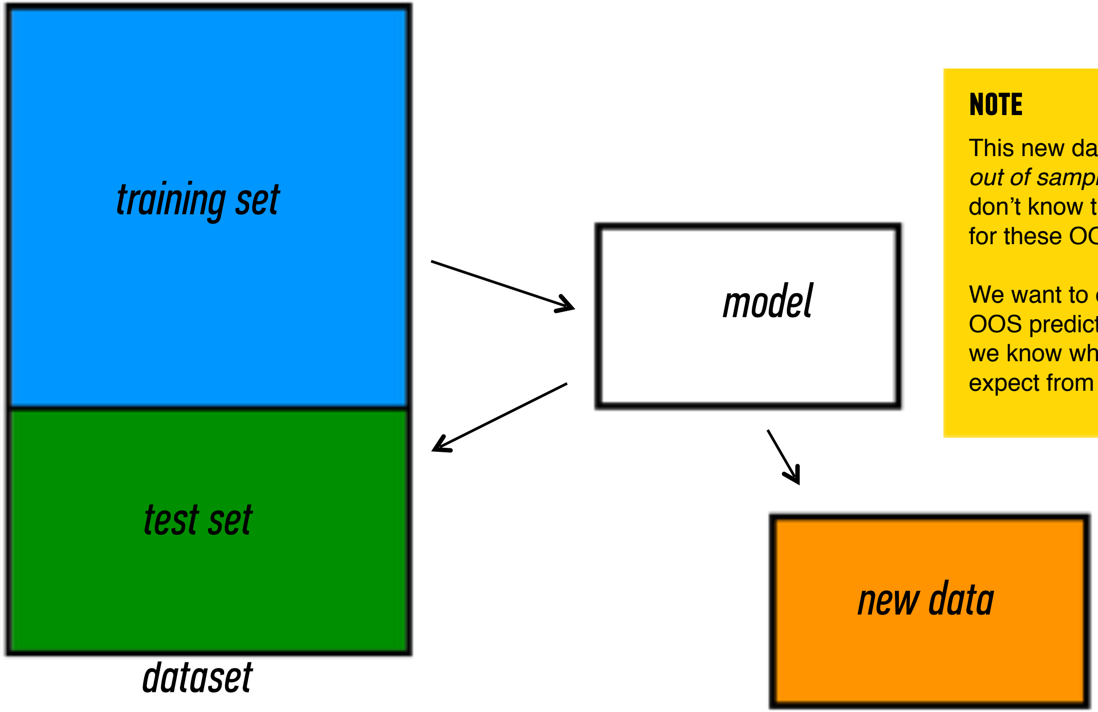
Q: How can we make a model that generalizes well?

1) split dataset
2) train model
3) test model
4) parameter tuning
5) choose best model
6) train on all data
7) make predictions on new data



Q: How can we make a model that generalizes well?

1) split dataset
2) train model
3) test model
4) parameter tuning
5) choose best model
6) train on all data
7) make predictions on new data



NOTE

This new data is called *out of sample* data. We don't know the labels for these OOS records!

We want to estimate OOS prediction error so we know what to expect from our model.

Suppose we do the train/test split.
Q: How well does test set error predict OOS?

Thought experiment:
Suppose we had done a different train/test split.
Q: Would the test set error remain the same?
A: Of course not!

A: On its own, not very well.

Suppose we do the train/test split.
Q: How well does test set error predict OOS?

Thought experiment:
Suppose we had done a different train/test split.
Q: Would the test set error remain the same?
A: Of course not!

A: On its own, not very well.

NOTE

The test set error gives a *high-variance estimate* of OOS accuracy.

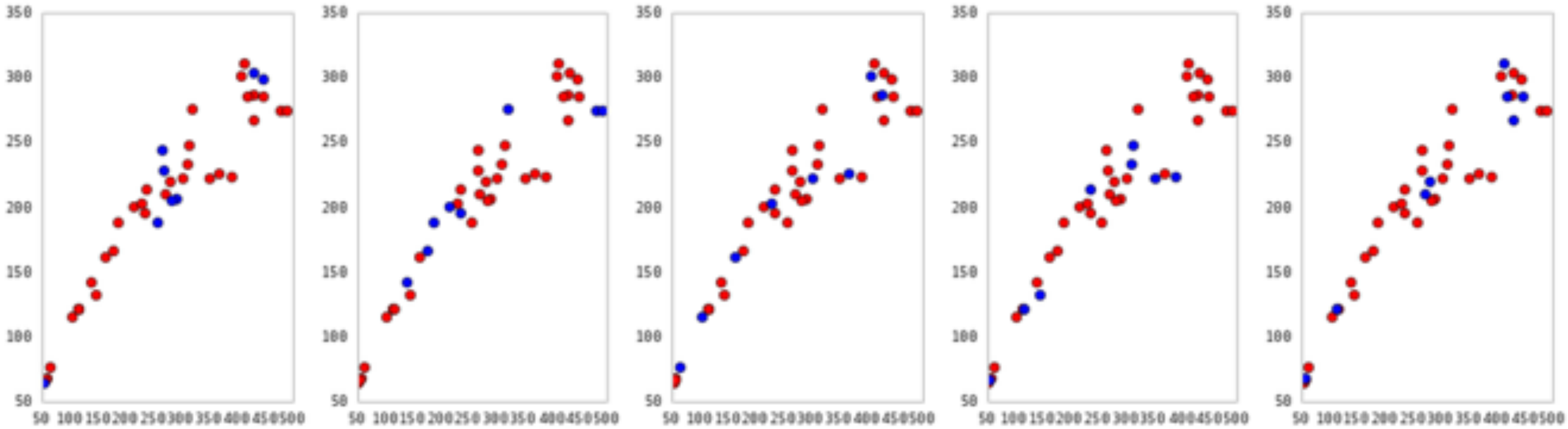
Something is still missing!
Q: How can we do better?

Thought experiment:
Different train/test splits will give us different test set errors.
Q: What if we did a bunch of these and took the average?
A: Now you're talking!

A: Cross-validation.

Steps for K-fold cross-validation:

1) Randomly split the dataset into K equal partitions.
2) Use partition 1 as test set & union of other partitions as training set.
3) Calculate test set error.
4) Repeat steps 2-3 using a different partition as the test set at each iteration.
5) Take the average test set error as the estimate of OOS accuracy.



5-fold cross-validation: red = training folds, blue = test fold

Source: http://nbviewer.ipython.org/github/fonnesbeck/Bios366/blob/master/notebooks/Section6_3-Model-Selection-and-Validation.ipynb

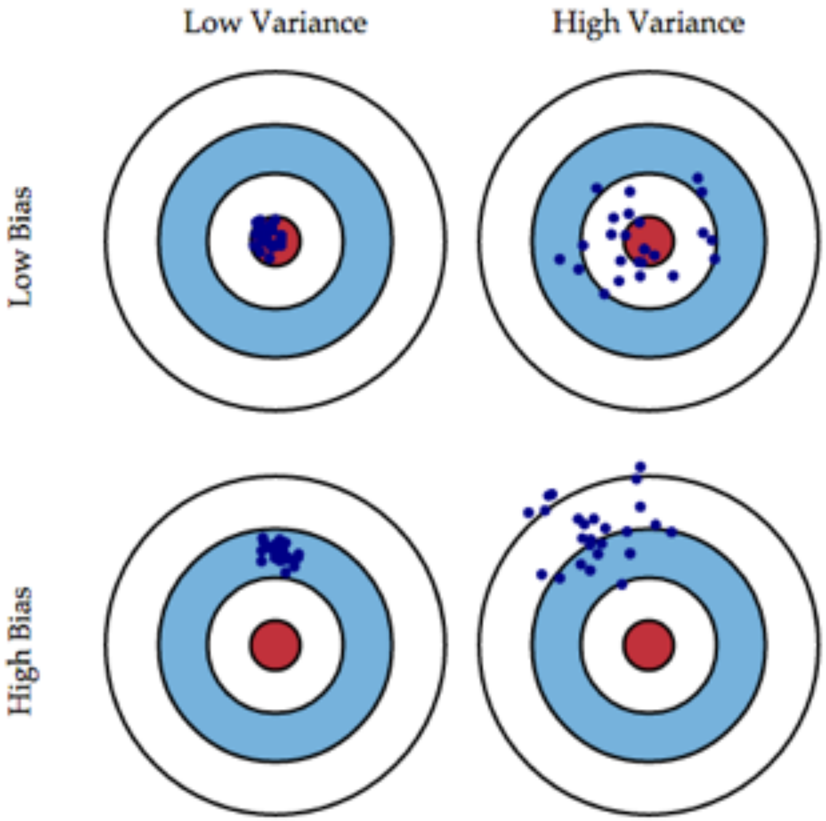
Features of K-fold cross-validation:

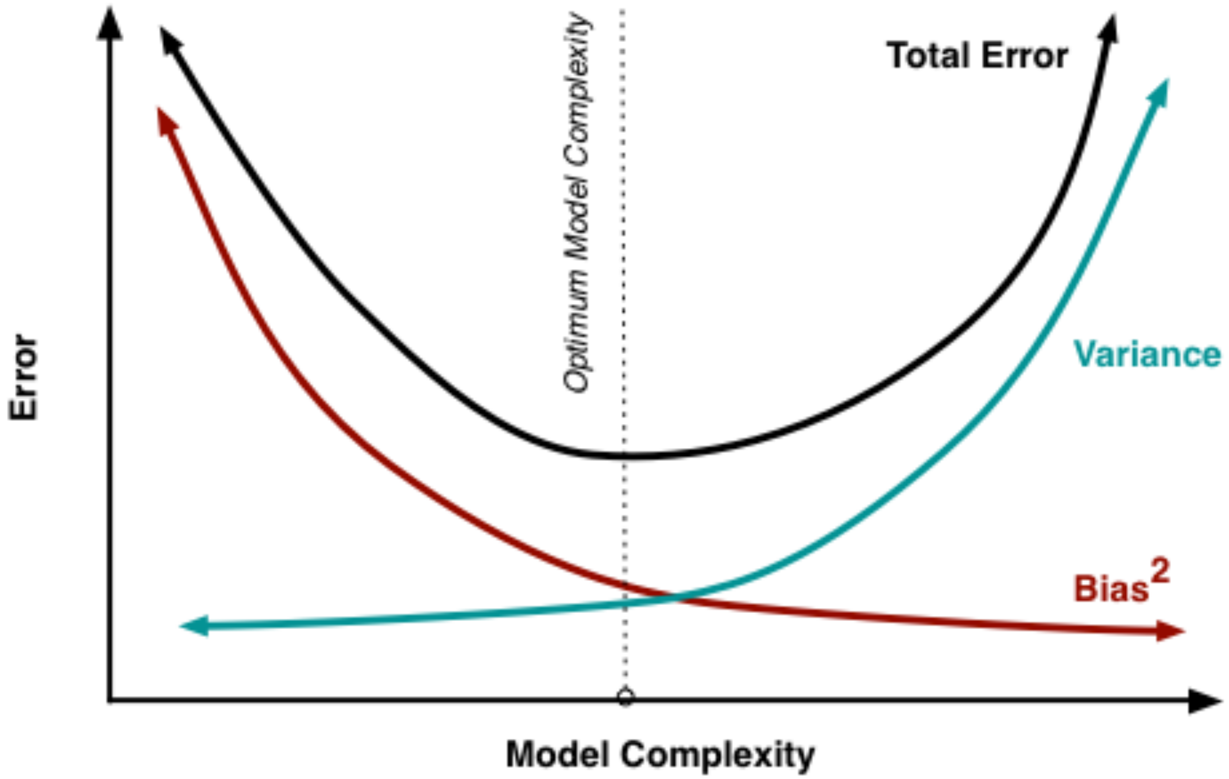
- More accurate estimate of OOS prediction error.
- More efficient use of data than single train/test split.
  - Each record in our dataset is used for both training and testing.
- Presents tradeoff between efficiency an computational expense.
  - 10-fold CV is 10x more expensive than a single train/test split
- Can be used for parameter tuning and model selection.

Errors due to Bias
When we are training over multiple data sets we will have different errors. Bias measures how far off we are in general the predictions are from the actual values

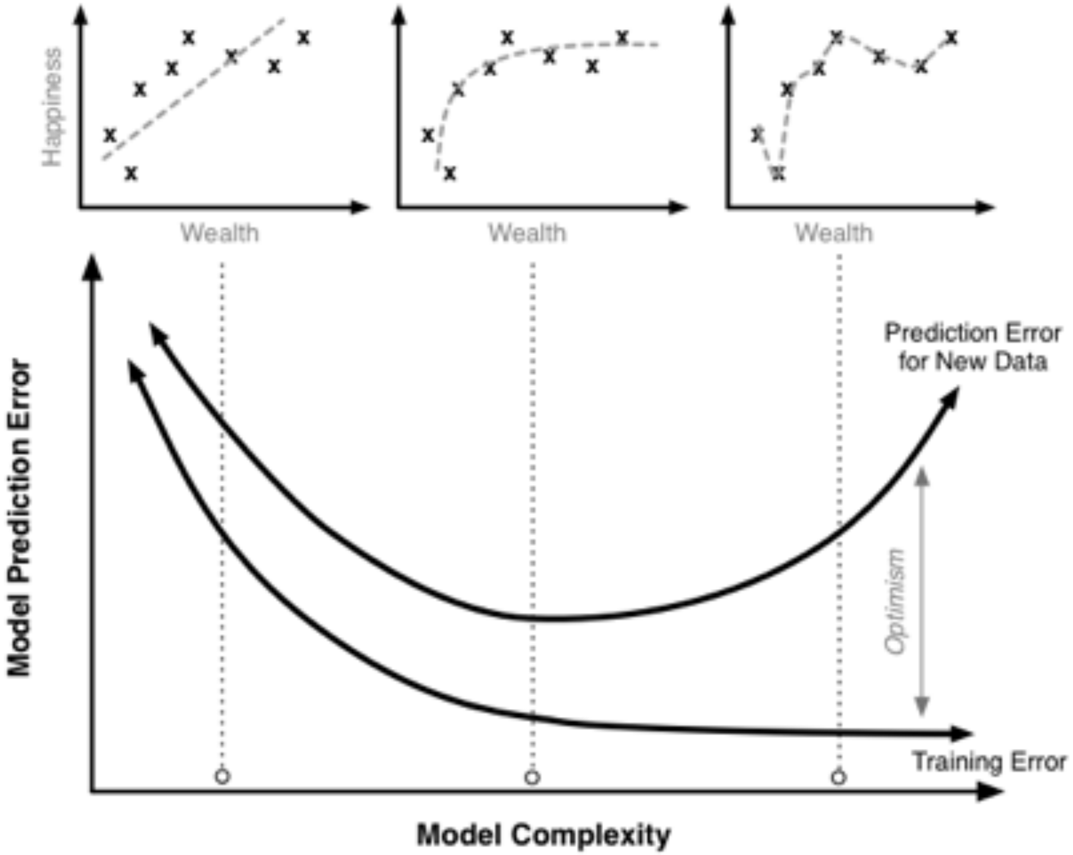
Errors due to Variance
This is how variable our model is for a given data point. The variance calculates how much the predicted are from the actual values

BIAS – VARIANCE TRADEOFF

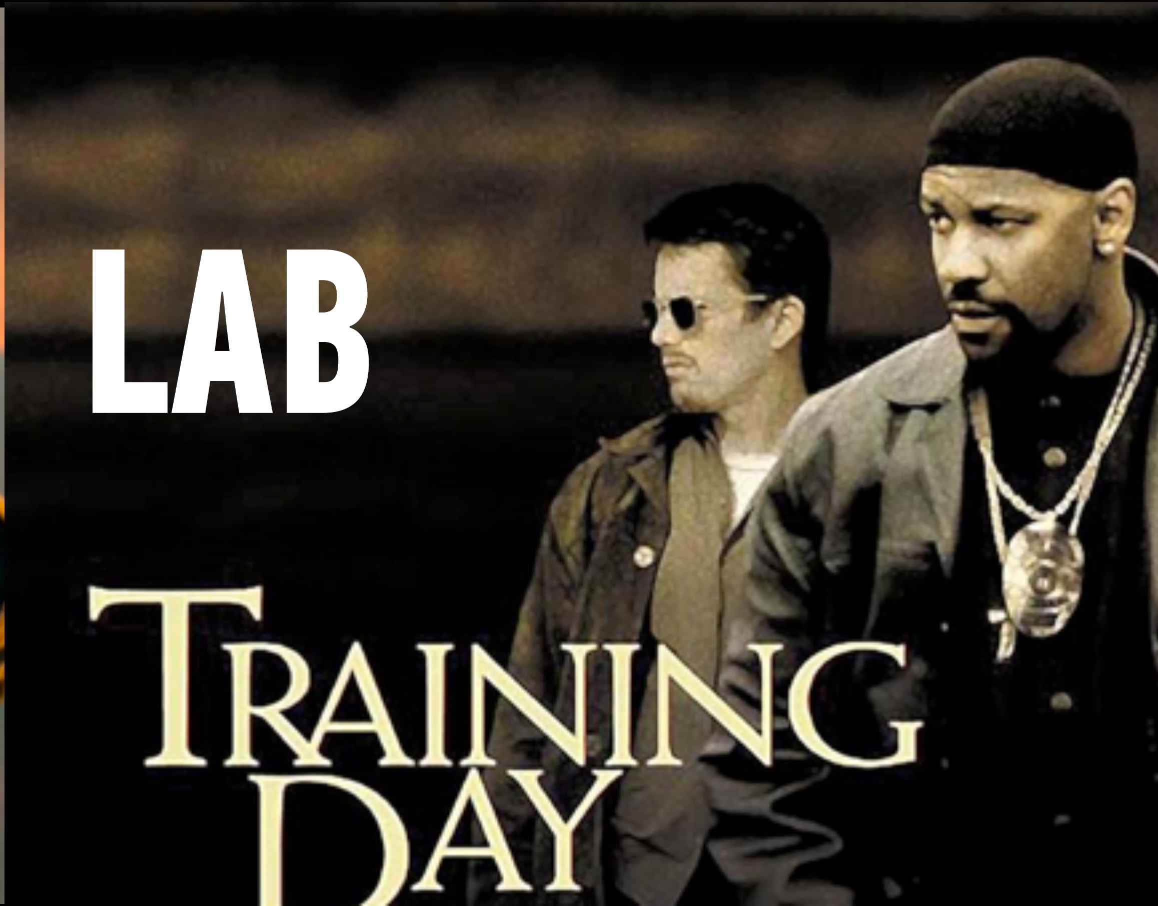



Optimum Model Complexity
Total Error
Variance
Bias^2
Error
Model Complexity

BIAS – VARIANCE TRADEOFF



LAB



DISCUSSION TIME

- Questions from previous lesson?
- Further Reading for Model Evaluation
- Check in with homework
- Tasks for next week

WEEK 3 Monday

QUESTIONS

- What are we trying to do when we use Logistic Regression?
- Why use it instead of Linear Regression for classification?
- Evaluating a logistic Regression model

DATA SCIENCE – Week 3 Day 2

DISCUSSION TIME

An Introduction to Statistical Learning

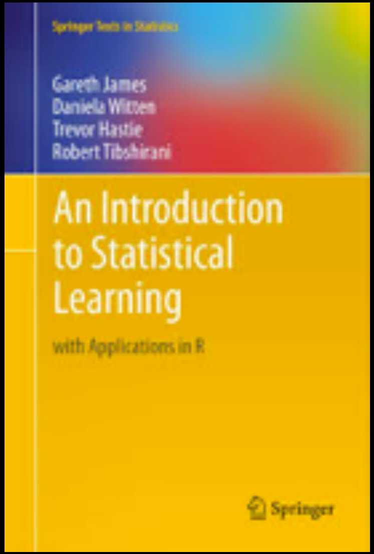
- Chapter 5 –Resampling Techniques

Logistic Regression applied to loan applications

- https://github.com/nborwankar/LearnDataScience

Odds Ratio in Logistic Regression

- http://www.ats.ucla.edu/stat/mult_pkg/faq/general/odds_ratio.htm



Tasks for Monday 11th January (< 30 mins)

- ☐ Have a git Repository setup with a Project Outline in the README.md
- ☐ Read and work though the code on this website on your own machine (copy and run)
  http://www.datarobot.com/blog/regularized-linear-regression-with-scikit-learn/
- ☐ Write out your computer specs (RAM, Hard Drive and CPU)
- ☐ If you are interested in doing kaggle competition send me your username so I can add you
- ☐ Fill in a Class Survey / Exit Ticket

SEND ME VIA SLACK PROOF OF THE ABOVE