Model Selection and Featurization

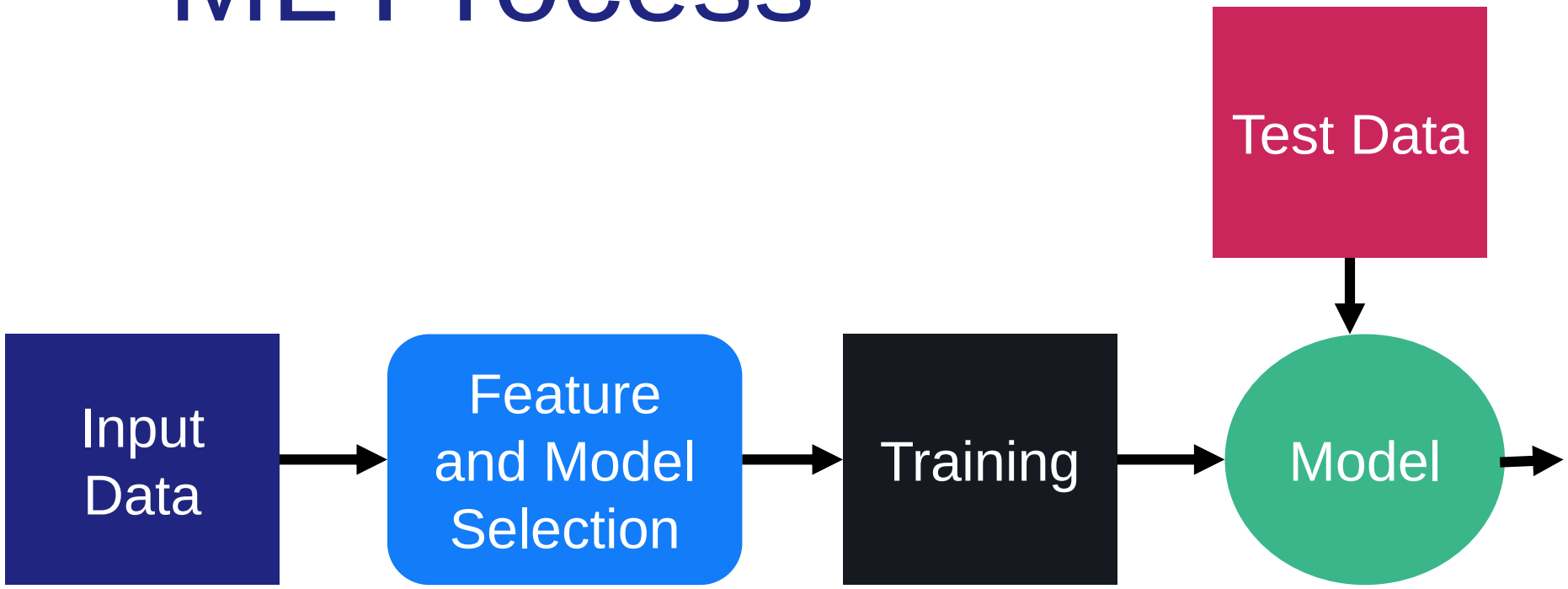
Outline

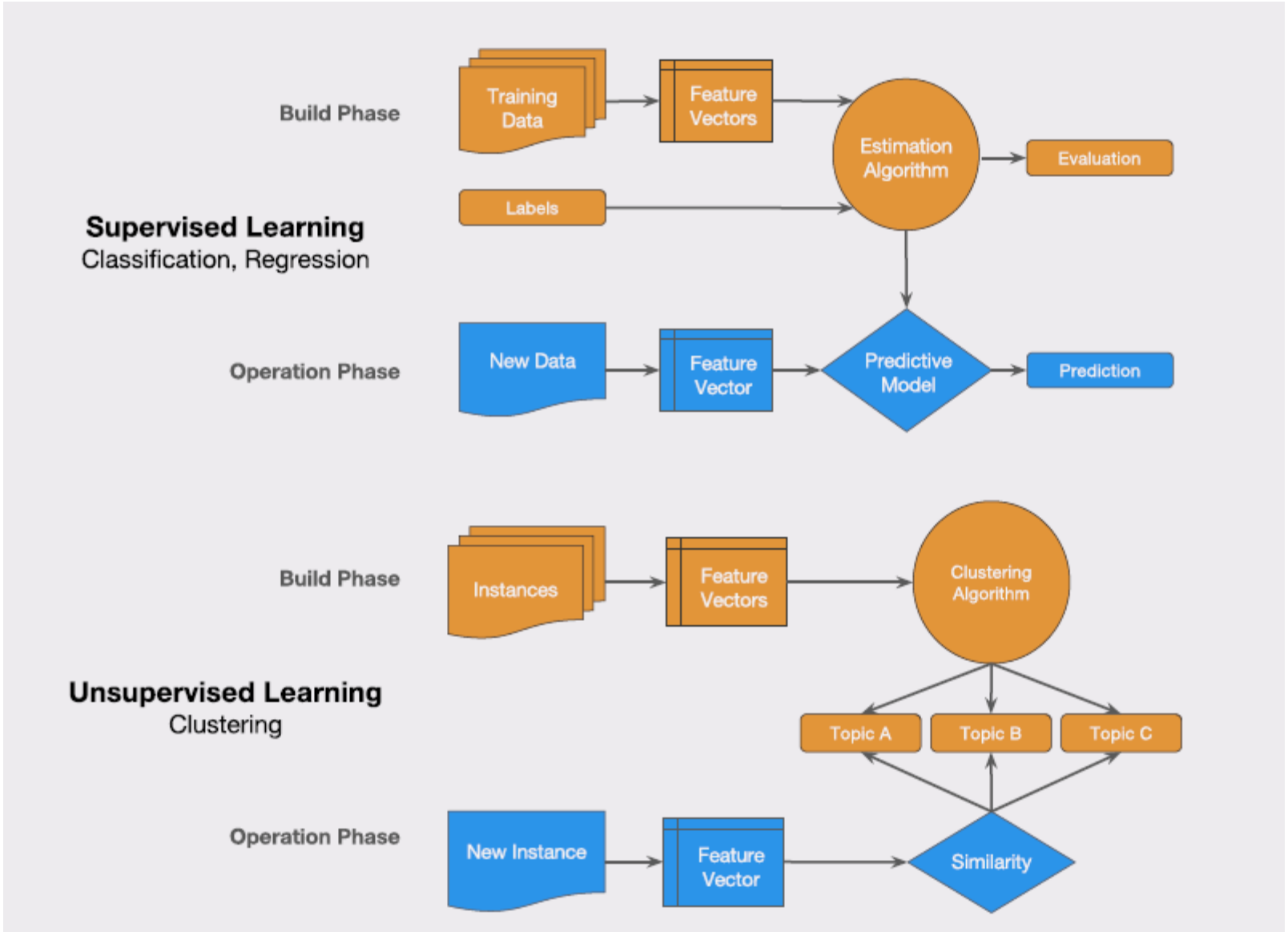
- Model Selection
- Training Set Error vs Test Set Error
- K-fold Cross Validation
- Feature Selection
- Possible Capstone Projects

What is Model Selection?

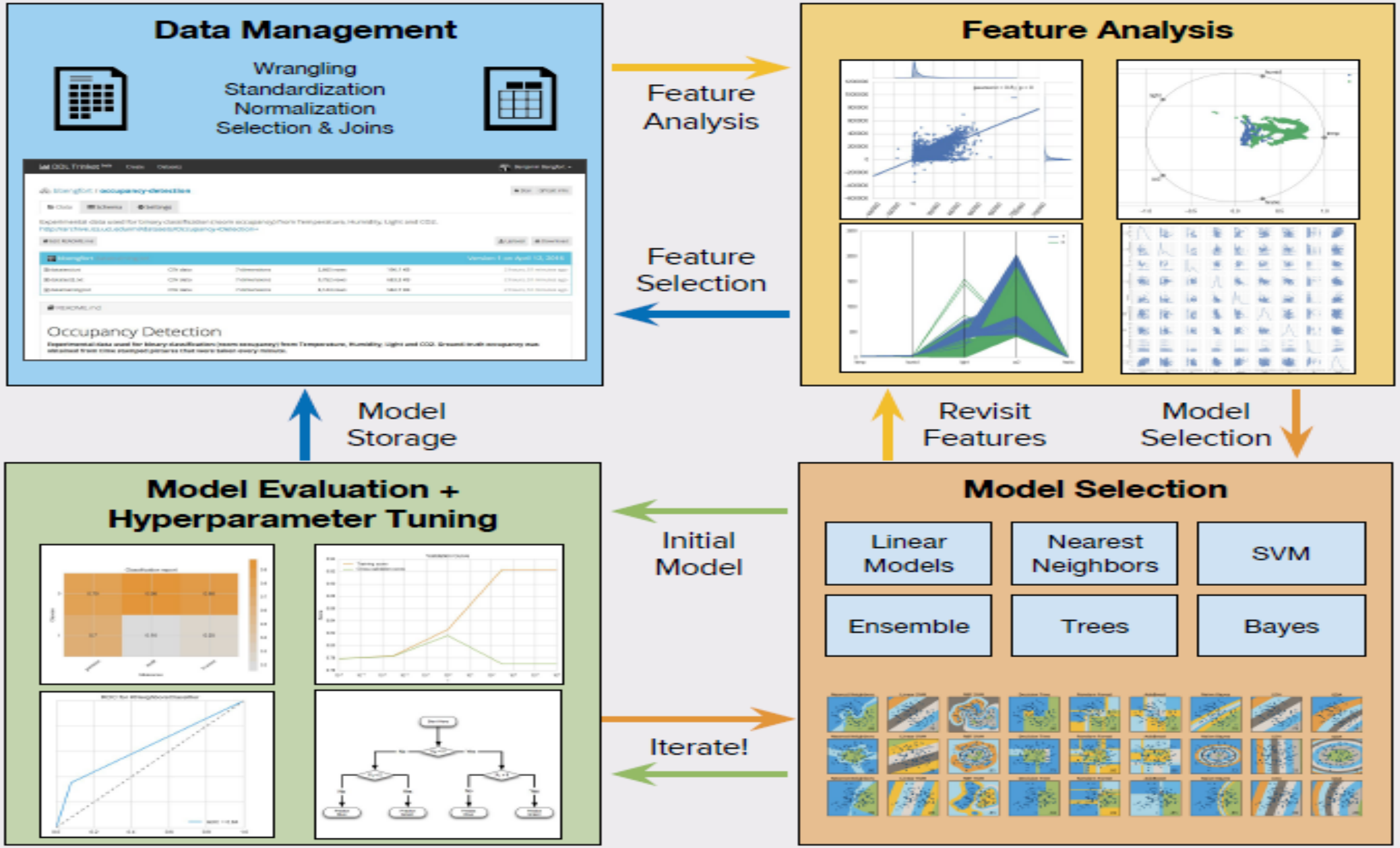
- Given a set of models $M = \{M_1, M_2, \cdots, M_R\}$, choose the model that is expected to do the best on the test data.
- M may consist of
  - Same learning model with different complexities or hyper parameters
    - Linear Regression
    - Nonlinear Regression: Polynomials with different degrees
    - K-Nearest Neighbors: Different choices of K
    - Decision Trees: Different choices of the number of levels / leaves
    - … and almost any learning model

ML Process


Test Data
Input Data
Feature and Model Selection
Training
Model


Build Phase
Training Data
Feature Vectors
Estimation Algorithm
Evaluation
Labels
Supervised Learning
Classification, Regression
Operation Phase
New Data
Feature Vector
Predictive Model
Prediction
Build Phase
Instances
Feature Vectors
Clustering Algorithm
Topic A
Topic B
Topic C
Unsupervised Learning
Clustering
Operation Phase
New Instance
Feature Vector
Similarity


Without Feedback
Models are Disconnected
They cannot adapt, tune, or react.


Data Management
Wrangling
Standardization
Normalization
Selection & Joins
Occupancy Detection
Feature Analysis
Feature Analysis
Feature Selection
Model Storage
Revisit Features
Model Selection
Model Evaluation + Hyperparameter Tuning
Model Selection
Initial Model
Linear Models
Nearest Neighbors
SVM
Ensemble
Trees
Bayes
Iterate!

Why Optimization?

OPT at the heart of ML

$$\min_{\mathbf{w}\in\mathcal{W}} \sum_{i=1}^{n} \ell_i(\mathbf{w}) + \lambda \cdot \mathcal{R}(\mathbf{w})$$

Measures model fit for data point i *(avoids under-fitting)*

Measures model "complexity" *(avoids over-fitting)*

Train and Test errors

- Test error rate
  - The average error that results from using a machine learning method to predict the response on a new observation, i.e., a measurement that was not used in training the method
  - Given a data set, the use of a particular machine learning method is warranted if it results in a low test error
- Training error
  - Calculated by applying the machine learning method to the observations used in its training
  - Training error rate often is quite different from the test error rate, and in particular the former can dramatically underestimate the latter

The Test Set Method: *car-mpg* data set

- Analyse car-mpg data set
  - There appears to be a non-linear relationship between ***mpg*** and ***horsepower***
  - A model that predicts ***mpg*** using ***horsepower*** and ***horsepower*^{2}** gives better results than a model that uses only a linear term
  - Any observations?

The Test Set Method

- Good News
  - Very Simple
  - Can we then choose the method with the best test score?
- Is there a downside?
  - Yes. It wastes 30% of data which can be critical when there is sparsity of data.
- What are the alternatives?
  - K-fold cross validation

Resampling methods

- Cross –Validation
  - Used to estimate the test error associated with a given machine learning method in order to evaluate its **performance**, or to select the appropriate level of **flexibility**
  - Model assessment: The process of evaluating a model's performance
  - Model selection: The process of selecting the proper level of flexibility for a model

Cross-Validation

- In the absence of a very large designated test set that can be used to directly estimate the test error rate, a number of techniques can be used to estimate this quantity using the available training data
- We study a class of methods that estimate the test error rate by *holding out* a subset of the training observations from the fitting process, and then applying the machine learning method to those held out observations

K-Fold Cross Validation

- Create K equal sized partitions of the training data
- Each partition has $\frac{N}{K}$ examples
- Train using $K-1$ partitions, validate on the remaining partition
- Repeat the same K times, each with a different validation portion

- Finally choose the model with smallest average validation error
- Usually K is chosen as 10

K-fold: ***Auto*** data set

- Calculate K-fold metric

Cross-validation for classification

- Instead of computing the sum squared errors on a test set, you should compute…

  The total number of misclassifications on a test set.

Very serious remark

- Intensive use of cross validation can over fit.
- What can be done about it?
    - Hold out an additional test set before doing any model selection. Check the best model performs well even on the additional test set.
    - Or: Randomization Testing

Feature Selection

- Suppose you have a learning algorithm LA and a set of input attributes { X1 , X2 .. Xm }
- You expect that LA will only find some subset of the attributes useful.
- Question: How can we use cross-validation to find a useful subset?
- Two ideas:
  - Forward selection
  - Backward elimination

Forward Selection

- Begin with null model - a model that contains an intercept but no predictors
- Then fit p simple linear regressions and add to the null model the variable that results in the lowest RSS(or highest R^2)
- Then add to that model the variable that results in the lowest RSS(or highest R^2) for the new two-variable model
- Continue this approach until some stopping rule is satisfied

Backward Elimination

- Start with all variables in the model
- Remove a variable from the above model and check the increment in RSS (decrement in R^2) and remove the variable which has least influence, i.e., the variable that is least significant
- The new (p-1) variable model is fit and the variable with the least significance is removed.
- Continue this procedure until a stopping rule is reached

Mixed Selection

- This is a combination of forward and backward selection
- We start with no variables in the model and as in forward selection, we add the variable that provides the best fit
- At times, the significance of variables can become low as new predictors are added to the model
- Thus, if at any point, the significance for one of the variables in the model falls below a certain threshold, then we remove that variable from the model
- We continue to perform these forward and backward steps until all variables in the model have a sufficiently high significance and all the variables outside the model would have a low significance if added to the model

Features Selection: *Auto* data set

- Select features based on
  - Forward selection
  - Backward selection

Possible Capstone Projects

- Consumer complaints data analysis and case outcome prediction
  - Collect case status data from district, state and NCDRC
  - Extract important basic info about each case (using NLP techniques) and add to the features
  - Build predictive models for the outcome of the complaints
- Consumer complaints judgments summarization
  - Real estate
  - Healthcare
- Similar consumer complaints judgements identification
- NLP based features engineering for consumer complaints judgements(may need to use sequential models also)

Summary

- Model Selection
- Training Set Error vs Test Set Error
- K-fold Cross Validation
- Feature Selection
- Possible Capstone Projects

Questions?