Unit 3 Schedule, And Beyond

**Unit 3 Overview**: This portion of the class is devoted towards model building, and the implementation of various types of statistical methods to make forecasts of various sorts. It’ll move beyond merely data cleaning and getting descriptive statistics and providing deeper statistical insights about what might happen *in the future.*

# Class 1: OLS

**Overview:** Objective Least Squares. Using linear regression to fit predictive models. Intuitions, graphing, basic principles.

**Suggested Reading:** ISL, Chapter 3. Pgs. 59-82 especially.

# Class 2 & 3: Cross Validation & Regularization

**Overview:** Training and test sets. KFold Cross Validation. Overfitting, bias and variance.

**Suggested Reading:** ISL, Chapter 5, Pgs. 175-190 especially.

# Class 4: Ensemble Models

**Overview:** Decision trees, boosting, bagging, ensembles, model interpretation.

**Suggested Reading:** ISL, Chapter 8, Pgs. 303-321

# Class 5: Classification

**Overview:** Predicting binary outcomes w/ logistic regression. Sigmoid, l1+l2 norms for classification.

**Suggested Reading:** ISL, Chapter 4, Pgs. 127-137

-------------------------------------------------------------------------------

**Other Topics Covered:** variable handling, categorical variables, ordinal variables, feature engineering

# Homework: Presentations on August 14th

# Unit 4

**Overview:** This unit is a bit more open-ended, and is built to add different layers of differentiation to the skills developed so far in this course. The emphasis will not necessarily be so statistical in nature, but rather emphasize various types of professional skills that compliment what you already know how to do.

To this end, the schedule is a bit open ended. There is one required class, but after that we can select from a variety of topics.

# Class 1: Connecting to API’s, Web Scraping

**Overview:** How to harness API data to connect to external web services. Basics of web scraping, including using the requests, json, and beautiful soup library. Reading html with pandas.

# Potential Topics:

**Ensemble Models, Continued:** They’re a very important technique, and xgboost is perhaps the most accurate learning method on structured data, so if we want to continue to refine our chops for developing accurate predictions, this is a good way to go.

**Cloud Computing:** Walk through of how to deploy models over AWS, deploy them as end points to make predictions, automatic scheduling of training and predictions, basic overview of cloud infrastructure.

**Using models as endpoints in software programs.** Go beyond merely making a model, and learn how to serialize it, and deploy it in a software application. We’ll cover pickling, and the basics of the flask API for using ML models in web applications. Will discuss techniques such as stochastic gradient descent for real-time optimization.

**Time Series Analysis:** Developing and interpreting ARIMA(p, d, q) models to make predictions based off of the passage of time.

**Deep Learning.** Basics of neural network construction. PyTorch API, overview of neural network architecture: CNNs, RNNs, and their common uses. Overview of unstructured data manipulation.

**Natural Language Processing.** Word embeddings, count vectorizers, stopwords, etc.

**Bonus Class:** Class will not have enough time to go over every single topic mentioned here. If availability permits, we can see about scheduling one additional bonus class for a topic that doesn’t get covered in class limits.