

# Course Syllabus

[Jump to Today](#)

## Course Background

---

Machine learning has become increasingly accessible for anyone with a statistical bent and a penchant for writing code. Predictive analytics are being utilized by academia, government, and business to gain insights from their that may have previously been too large to model. Thanks to the rise in computational power, running complex algorithmic modeling techniques on these big data sets can be run from a personal laptop.

R has become a go-to programming language for data scientist packed with helpful libraries and functions to shape and transform data. One of the many reasons why data scientists love using the R language is that it's **open sourced**, where you are free to use it and it comes with an active community continually adding resources.

## Course Objectives

---

This class will be an introduction to machine learning techniques. The course will require a basic understanding of statistical concepts and coding skills. The course will not go into every detail of each technique but will provide further reading and there will be class discussions to stimulate ideas.

The first portion of the class will be dedicated to learning the basics of R especially around shaping data for machine learning. The second portion of class will delve into strategies dealing with larger data sets; topics will include parallel and distributed computing. The third and last portion of the class will be working with algorithms. Some of the machine learning tasks the class will explore are classification, regression, and clustering algorithms. The techniques we will cover include k-nearest neighbors, random forests, gradient boosting machines, k-means clustering and linear regression.

The class will work on tuning parameters and evaluating performance of models. A diverse selection of urban data sets will be used in the class; some of them include 311 data, PLUTO, US census, and Citi bike data.

## Text Book

---

R for Everyone: Advanced Analytics and Graphics by Jared Lander ([amazon](https://www.amazon.com/Everyone-Advanced-Analytics-Graphics-Addison-Wesley/dp/0321888030)  [\(https://www.amazon.com/Everyone-Advanced-Analytics-Graphics-Addison-Wesley/dp/0321888030\)](https://www.amazon.com/Everyone-Advanced-Analytics-Graphics-Addison-Wesley/dp/0321888030) )

Machine Learning with R – second edition by Brett Lantz (Safari Online)

## Online Resources

---

[Online Stat Book](http://onlinestatbook.com/2/introduction/introduction.html)  [\(http://onlinestatbook.com/2/introduction/introduction.html\)](http://onlinestatbook.com/2/introduction/introduction.html) -Basic stats knowledge

## Assignments

---

There will be weekly readings and assignments for the first third of the class, reinforcing and pushing topics covered in class.

# Grading

---

- Assignments 50%
- Reading Discussion 10%
- Final Project 40%

# Syllabus

---

## Part I - R Basics

- Getting started with R
- R Environment - intro to RStudio
- Essential R Programming
  - Data Types
  - Data Structure
  - Functions
  - Logical Operators
  - Group Manipulation

## Part II - Machine Learning

- Data Munging
  - Grep (regular expressions)
  - Dplyr
  - Reshape
  - Missing Data
- Exploratory Data Analysis
  - Getting started with ggplot2
  - Histograms and distributions
- Introduction to ML
  - What is machine learning?
  - Machine learning problems
  - Supervised vs. Unsupervised Learning
  - Regression, Classification, Clustering
- Clustering Algorithms
  - K-means
  - Clustering exercise (PLUTO data)
- Classification
  - KNN K nearest neighbors (lazy learners)
  - KNN exercise (311 data)
  - Naive Bayes (probabilistic learning)
  - Naive Bayes exercise
  - Decision Trees
  - Decision tree exercise
- Regression
  - Generalized linear models
  - GLM exercise (Citi bike data)
  - Regression with trees
  - Regression trees exercise
- Model Evaluation
  - Classification

- Confusion matrix
- Sensitivity, specificity, precision, recall
- ROC curves
- Regression
- Residuals
- P-value
- RMSE

## Part III - Final Projects

Bring a classification or regression problem to life with machine learning. Choose a data set of your liking and use one of the tools learned in class to see how well you can model the topic.

## Calendar

### Week 1 - 3/7

- Course Intro
- R Basics

### Week 2 - 3/14

- SPRING BREAK

### Week 3 - 3/21

- Munging
- EDA
- Reading (Critical Theory of Technology)

### Week 4 - 3/28

- Intro to machine learning
- Clustering
- Reading (Technological Utopianism in American Culture)

### Week 5 - 4/4

- Classification
- Regression
- Reading(Rethinking Objectivity)

### Week 6 - 4/11

- Model Evaluation
- Reading(Trust in Numbers, Weapons of Math Destruction)






### Week 7 - 4/18

- Work day
- Interactive Data Viz with Shiny
- Probabilistic Learning

### Week 8 - 4/25

- Final Presentations

## Course Summary:

Date	Details	
Tue Mar 28, 2017	 Assignment 1	due by 11:59pm
Tue Apr 4, 2017	 Assignment 2	due by 7pm
Tue Apr 11, 2017	 Assignment 3	due by 11:59pm
Tue Apr 25, 2017	 Final Project	due by 11:59pm
	 Roll Call Attendance	