Introduction to Data Science

# Introduction

We have entered a time in which vast amounts of data are more widely available than ever before. At the same time, a new set of tools has been developed to analyze this data and provide decision makers with information to help them accomplish their goals. Those who engage with data and interpret it for organizational leaders have taken to calling themselves data scientists, and their craft data science. Other terms that have come into vogue are "Big Data," "Predictive Analytics" and "Data Mining." These can seem to be mysterious domains. The point of this class is to demystify much of this endeavor for individuals who will be organizational leaders.

The class is structured around developing students' skills in three areas: getting data, analyzing data to make predictions, and presenting the results of analysis. For each area, the subtopics are as follows:

## Getting Data Topics

* Tools of the Trade: R and Rstudio
* Working with pre-processed data and flat files
* Getting data from the web: webscraping, using forms, using Application Programming Interfaces
* Using databases

## Analyzing Data Topics

* Descriptives and conditional means
* Regression
* Supervised learning: classification
* Unsupervised learning: K-means and nearest neighbors clustering
* Cross Validation

## Presenting Data Analysis Topics

* Descriptives: histograms, density plots, bar plots, dot plots
* Scatterplots
* Lattice graphics and small multiples
* Interactive graphics
* Communicating results effectively

# Evaluation

Students will be evaluated based in two areas: weekly assignments and the final project.

* Problem sets: 65% Each week I will assign a problem set for students to complete. These problem sets will be assigned on Monday, and will be due the next Sunday night at 11:59:59 pm. No late assignments will be accepted. Each assignment will be graded on a 100 point scale. Your lowest grade will be dropped.
* Final Project 35%: During the course of the semester you will work on a final assignment utilizing your skills as a data analyst. There will be four progress reports due for the final project, each of which will be worth 12.5% of the final grade for the project. No late progress reports will be accepted. The final product will account for the remaining 50%. No late final projects will be accepted.

# Texts

*Required Texts*

We will have two texts for the course. The first is Hadley Wickham's book, [R for Data Science](http://r4ds.had.co.nz/).

Wickham, H., & Grolemund, G. (2016). R for data science: import, tidy, transform, visualize, and model data. San Francisco: O'Reilly Media, Inc.

The other text is Nate Silver's *Signal and the Noise*.

Silver,N. (2012). *The signal and the noise: Why so many predictions fail-but some don’t*. New York: Penguin.

# Software

We will use only free, [open source](https://opensource.org/osd-annotated) software in this course.

We will use [R](https://cran.r-project.org/), an open-source data analytic platform for all analysis. R appears to be the most widely used data analysis software in data science. We will utilize [Rstudio](https://www.rstudio.com/home/) as our integrated development environment (IDE) for R.

# Honor Code Statement

All assignments for this class, including weekly assignments and the final project, are to be conducted under the obligations set out in Vanderbilt's Honor Code. Please click [here](http://www.vanderbilt.edu/student_handbook/the-honor-system/)to review the honor code.

There will be two quite different standards for completing the assignments and the final project.

*Assignments* You may collaborate with anyone and you may utilize any resource you wish to complete these assignments.

*Final Project* All of the work on the final assignment must be your own. Anyone's work that you reference should be cited, as usual. All data that you do not personally collect must be cited, as with any other resource.

If you have any questions at all about the honor code or how it will be applied, ask me right away.

# Schedule

## Week 1: Topic for the Week: Getting Data 1-- Tools of the Trade

*Resources*

Wickham: [Introduction](http://r4ds.had.co.nz/introduction.html), [Explore: Introduction](http://r4ds.had.co.nz/explore-intro.html), [Workflow: basics](http://r4ds.had.co.nz/workflow-basics.html), [Workflow: projects](http://r4ds.had.co.nz/workflow-projects.html)

Silver, Chapters 1-4

R Intro and Resources

[Download R](https://cran.r-project.org/)

[R Basics](https://www.datacamp.com/home)

[Download Rstudio](https://www.rstudio.com/products/rstudio/download3/) You want the "Desktop" version, free license

[Rstudio Intro and Resources](https://www.rstudio.com/online-learning/)

*Lesson Notes*

Chapter 1, Introduction: 01-intro.Rmd.

*Synchronous session* R Basics, "verbs" of data wrangling

## Week 2 Topic for the Week: Analyzing Data: Conditional Means

*Resources*

Wickham: [Data transformation](http://r4ds.had.co.nz/transform.html)  
Silver, Chapters 5-9, 12-13

*Lecture Notes*

Chapter 2, Condtional Means: 02-conditional\_means.Rmd.

*Assignments*

Assignment 1 Due Midnight Sunday

*Synchronous Section: Conditional Means*

## Week 3 Topic for the Week: Presenting Data-- Descriptives

*Resources*

Wickham: [Data visualization](http://r4ds.had.co.nz/data-visualisation.html)

[Data transformation](http://r4ds.had.co.nz/exploratory-data-analysis.html)

[Cookbook for R:Bar and Line Graphs](http://www.cookbook-r.com/Graphs/Bar_and_line_graphs_%28ggplot2%29/)

[Cookbook for R: Plotting Distributions](http://www.cookbook-r.com/Graphs/Plotting_distributions_%28ggplot2%29/)

*Lecture Notes*

Chapter 3, Plotting Distributions and Condtional Means: 03-plot\_means.Rmd.

*Assignments*

Assignment 2 Due Midnight Sunday

*Final Project* First progress reports due, Sunday midnight

*Synchronous session: Presenting results in graphical format: barplots, density plots, dot plots, histograms*

## Week 4 Topic for the Week: Getting Data --pre-processed data, flat files, basic concepts of "tidy data"

*Resources*

Wickham: [Data import](http://r4ds.had.co.nz/data-import.html), [Tidy data](http://r4ds.had.co.nz/tidy-data.html)

*Lecture Notes*

Chapter 4, Flat Data 04-flat\_data.Rmd

*Assignments*

Assignment 3 Due Midnight Sunday

*Synchronous session: working with various data formats*

## Week 5 Topic for the Week: Analyzing Data--Linear Regression

*Resources*

Wickham: [Model: Introduction](http://r4ds.had.co.nz/model-intro.html), [Model Basics](http://r4ds.had.co.nz/model-basics.html), [Model Building](http://r4ds.had.co.nz/model-building.html)

*Lecture Notes*

Chapter 5, Linear Regression 05-regression.Rmd

*Assignments*

Assignment 4 Due Midnight Sunday

*Synchronous session: using linear regression, training and testing Models*

## Week 6 Topic for the Week: Presenting data--scatterplots

*Resources*

Wickham: [Data Visualization](http://r4ds.had.co.nz/data-visualisation.html), [Graphics for Communication](http://r4ds.had.co.nz/graphics-for-communication.html)

Tufte, Visual Display chapters 4 and 5.

Tufte, Envisioning Information, chapter 2

*Lecture Notes*

Chapter 6, Scatterplots 06-scatterplots.Rmd

*Assignments*

Assignment 5 Due Midnight Sunday

*Final Project*

Second Progress Report for Final Project Due

*Synchronous session: Presenting Data via Scatterplots*

## Week 7 Topic for the week: Getting Data-- Getting data from the web

*Resources*

Rvest Vignette: <https://cran.r-project.org/web/packages/rvest/vignettes/selectorgadget.html>

[Reed College rvest introduction](http://www.reed.edu/data-at-reed/resources/R/rvest.html)

[rvest tutorial](https://stat4701.github.io/edav/2015/04/02/rvest_tutorial/)

*Lecture Notes*

Chapter 7, Web Scraping and APIs, 07-webscrape.Rmd

*Assignments*

Assignment 6 Due Midnight Sunday

*Synchronous session: Accessing data from the web*

## Week 8 Topic for the Week: Analyzing Data--Classification

*Resources*

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 6). New York: Springer. [Chapter 4](http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Sixth%20Printing.pdf) , [Chapter 4 Lab R Code](http://www-bcf.usc.edu/~gareth/ISL/Chapter%204%20Lab.txt)

Althoff, T., Danescu-Niculescu-Mizil, C., & Jurafsky, D. (2014, May). How to Ask for a Favor: A Case Study on the Success of Altruistic Requests. In ICWSM. (Available Here)[<http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/download/8106/8101>]

*Lecture Notes*

Chapter 8, Classification, 08-classification.Rmd

*Assignments*

Assignment 7 Due Midnight Sunday

*Synchronous Session: Classification*

Classifying behavior via text analysis: random acts of pizza.

## Week 9 Topic for the Week: Presenting Data-- Plots and Tables for Classification

*Resources*

*Assignments*

Assignment 8 Due Midnight Sunday

*Final Project*

Third Progress Report Due

*Lecture Notes*

Chapter 9, plots and tables for classification 09-plotting\_classification.Rmd

*Synchronous Session: Presenting Data, Plots and Tables for classification*

## Week 10 Topic for the Week: Analyzing Data-- Cross Validation

*Resources*

Wickham [Many Models](http://r4ds.had.co.nz/many-models.html)

Lecture Notes

## Week 11 Topic for the Week: Getting Data--databases and relational data

*Resources* Wickham [Relational Data](http://r4ds.had.co.nz/relational-data.html)

Working with Databases in R, available: <https://cran.r-project.org/web/packages/dplyr/vignettes/databases.html>

*Lecture Notes*

Chapter 11, Databases 11-databases.Rmd

*Assignments*

Assignment 10 Due Midnight Sunday

*Synchronous session* Databases and relational data

## Week 12 Topic for the Week: Unsupervised learning

Subtopics: k-means, nearest neighbor clustering

*Resources*

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 6). New York: Springer. [Chapter 10](http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Sixth%20Printing.pdf) , [Chapter 10 Lab R Code](http://www-bcf.usc.edu/~gareth/ISL/Chapter%2010%20Labs.txt)

*Lecture Notes*

Chapter 12, Unsupervised Learning 12-unsupervised.Rmd

*Assignments*

Assignment 11 Due Midnight Sunday

*Final Project*

Fourth Progress Report Due

\*Synchronous [session:\*](session:*) K-means clustering, nearest neighbor classification

## Week 13: Topic for the Week: Presenting Data-- Interactive Graphics

*Resources*

*Lecture Notes*

Chapter 13, Interactive Graphics 13-interactive.Rmd

*Assignments*

Assignment 12 Due Midnight Sunday

*Synchronous Session* Interactive Graphics

## Week 14: Topic for the Week: Communicating Results

Assignment 13 Due Midnight Sunday

*Synchronous Sessions* Class Presentations

## Final Projects Due Two Weeks After Last Class