

# Introduction to Data Science

**Vanderbilt University**

**Human and Organizational Development**

**Course Number HOD 3200**

**Fall 2017**

William R. Doyle

Office: 207D Payne

Office Hours: Mondays and Wednesdays, 1-3 or by appointment

Email: [w.doyle@vanderbilt.edu](mailto:w.doyle@vanderbilt.edu)

Twitter: @wdoyle42

Phone: (615) 322-2904

## 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, git and Github
- 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
- Evaluating multiple models

## Presenting Data Analysis Topics

- Descriptives: histograms, density plots, bar plots, dot plots
- Scatterplots
- Lattice graphics and small multiples
- Maps
- Interactive graphics

## 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. We will discuss this assignment and my expectations in detail during the course of the semester. 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. The final product will account for the remaining 50%.

## Texts

### *Required Texts*

We will have two texts for the course. The first is Hadley Wickham's book, R for Data Science. Wickham is generously making this book available for free. However, I strongly encourage you to buy this book from O'Reilly.

Amazon

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.

Amazon

Half.com

Your local bookseller

### *Reserve*

I've placed three books by Edward Tufte on reserve for you. These are masterpieces in the area of visualizing quantitative information. You should take a look at these for ideas and inspiration—I've noted the sections that are most helpful in various parts of the syllabus.

Tufte, E. R. (1990). *Envisioning information*. Cheshire, CT: Graphics Press.

Tufte, E. R. (1997) *Visual explanations*. Cheshire, CT: Graphics press.

Tufte, E. R. (2001). *The visual display of quantitative information* (2nd Edition). Cheshire, CT: Graphics press.

## Lecture Notes

My lecture notes are available at: [willdoyle.us/data\\_step](http://willdoyle.us/data_step). These include both code and notes for the week. They will also be available in your private github repository.

## Web Resources

When appropriate for each week, web resources are linked directly from the syllabus. You will find a wealth of resources online, including other versions of this class offered as Massive Online Open Courses. I encourage you to take full advantage of the wealth of online materials that are available. Stack Overflow is your friend, but search carefully for your question. It is VERY likely that your question has already been asked.

## Software

We will use only free, open source software in this course.

We will use R, 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 as our integrated development environment (IDE) for R.

We will also use git, a distributed version control program, and Github, an online hosting platform. Github Desktop will serve as our Graphical User Interface to git and GitHub: RStudio is fully integrated with git and Github, making it an ideal IDE for these purposes. Class assignments will be distributed through GitHub and will be collected and graded through GitHub as well.

## Communication

My office is in 207D Payne, and my phone number is (615) 322-2904. Please always feel free to stop by during office hours (Mondays and Wednesdays 12:30-3, right after class) or to call. If my office hours don't work for you, please make an appointment. Student communications, including emails are my priority. However, due to the volume of email I receive, I may miss your message. To help with this problem, please place the phrase "HOD 3200" in your subject line. I will search for these messages every time I access my email. You can also use Brightspace's email function, which will automatically do this for you. If you have a general question that I can answer for the whole class, send me a message on twitter at @wdoyle42, tagged #hoddatasci, or you can send a direct message.

## 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 hereto 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

### Wednesday, August 23 Topic for the Week: Getting Data 1– Tools of the Trade

#### *Resources*

Wickham: Introduction, Explore: Introduction, Workflow: basics, Workflow: projects

Silver, Chapters 1-4

R Intro and Resources

Download R

R Basics

Download Rstudio You want the “Desktop” version, free license

Rstudio Intro and Resources

Download git

Download GitHub Desktop

Github Intro and Resources

### Monday, August 28 Getting Data 1: Tools of the Trade

Subtopics: “verbs” of data wrangling, file types, working with git and GitHub.

#### *Lesson Notes*

Chapter 1, Introduction: 01-intro.Rmd.

### Wednesday, August 30 Getting Data 1: Tools of the Trade

#### *Standing meeting*

*Lab Practical* R Basics, “verbs” of data wrangling

### Monday, September 4 Topic for the Week: Analyzing Data 1– Conditional Means

#### *Resources*

Wickham: Data transformation

Silver, Chapters 5-9, 12-13

#### *Lecture Notes*

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

#### *Assignments*

Assignment 1 Due Midnight, Sunday, August 27

### Wednesday, September 6 Conditional Means, continued

#### *Standing Meetings*

*Lab Practical: Conditional Means*

## **Monday, September 11 Topic for the Week: Presenting Data 1– Descriptives**

Subtopics: bar plot, density plot, dot plots, histograms

### *Resources*

Wickham: Data visualization

Data transformation

Cookbook for R: Bar and Line Graphs

Cookbook for R: Plotting Distributions

### *Lecture Notes*

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

### *Assignments*

Assignment 2 Due Midnight Sunday, September 3

## **Wednesday, September 13 Descriptive Graphics, Continued**

*Standing Meeting* Progress reports due for section 1.

*Lab Practical: Presenting results in graphical format: barplots, density plots, dot plots, histograms*

## **Monday, September 18 Topic for the Week: Getting Data 2–pre-processed data, flat files, basic concepts of “tidy data”**

### *Resources*

Wickham: Data import, Tidy data

### *Lecture Notes*

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

### *Assignments*

Assignment 3 Due Midnight Sunday, September 10

## **Wednesday, September 20 Pre-processed data, continued**

*Standing Meeting* Progress reports due for section 2.

*Lab Practical: working with various data formats*

## **Monday, September 25 Topic for the Week: Analyzing Data 2–linear regression**

### *Resources*

Wickham: Model: Introduction, Model Basics, Model Building

### *Lecture Notes*

Chapter 5, Linear Regression 05-regression.Rmd

### *Assignments*

Assignment 4 Due Midnight Sunday, September 17

## **September Wednesday, September 27 Linear Regression, continued**

*Standing Meetings*

*Lab Practical: linear regression*

## **Monday, October 2 Topic for the Week: Presenting data 2–scatterplots**

*Resources*

Wickham: Data Visualization, Graphics for Communication

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, September 24

## **Wednesday, October 4 Scatterplots, continued**

*Standing Meetings* Progress reports due for section 1.

*Lab Practical: Presenting Data via Scatterplots*

## **Monday, October 9 Topic for the week: Getting Data 3– Getting data from the web**

*Resources*

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

Reed College rvest introduction

rvest tutorial

*Lecture Notes*

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

*Assignments*

Assignment 6 Due Midnight Sunday, October 1

## **Wednesday, October 11 Web data, continued**

*Standing Meetings* Progress reports due for section 2.

*Lab Practical*

## Monday, October 16 Topic for the Week: Analyzing Data 3–Classification

### *Resources*

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 6). New York: Springer. Chapter 4 , Chapter 4 Lab R Code

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, October 8

## Wednesday, October 18 Classification, continued

### *Standing Meetings*

### *Lab Practical*

Classifying behavior via text analysis: random acts of pizza.

## Monday, October 23 Topic for the Week: Presenting Data 3– Lattice plots and small multiples

### *Resources*

Tufte, Envisioning Information, Chapter 4

### *Lecture Notes*

Chapter 9, Small Multiples 09-small\_multiples.Rmd

### *Assignments*

Assignment 8 Due Midnight Sunday, October 15

## Wednesday, October 25 Lattice Plots and Small Multiples, continued.

### *Standing Meetings*

Progress reports due for section 1.

### *Lab Practical*

## Monday, October 30 Topic for the Week: Getting Data 4–databases and relational data

### *Resources* Wickham Relational Data

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

### *Lecture Notes*

Chapter 10, Databases `databases.Rmd`

*Assignments*

Assignment 9 Due Midnight Sunday, October 22

## **Wednesday, November 1 Databases, continued**

*Standing Meetings*

*Lab Practical* Databases and relational data, collaborating via github

## **Monday, November 6 Topic for the Week: Training and testing multiple models**

*Resources*

Wickham Many Models

*Lecture Notes*

Chapter 13, Multiple Models `multiple_models.Rmd`

*Assignments*

Assignment 10 Due Midnight Sunday, October 29

## **Wednesday, November 8 Training and testing, continued**

*Standing Meeting*

*Lab Practical*

## **Monday, November 13 Topic for the Week: Presenting Data 4–Maps**

*Resources*

Lovelace, Cheshire et al (2015) Introduction to visualising spatial data in R. Online

*Lecture Notes*

Chapter 12, Mapping: `12-mapping.Rmd`

*Assignments*

Assignment 11 Due Midnight Sunday, November 5

## **Wednesday, November 15 Mapping, continued**

*Standing Meetings*

Progress reports due for section 1 and 2.



## **Monday, November 20 Wednesday, November 22**

Thanksgiving Break: No Class, but lots of data analysis to do!

### *Lecture Notes*

Take a look at Chapter 14, Interactive Graphics: `14-interactive_graphics.Rmd` This will be very helpful for creating nifty graphics for your final assignment.

## **Monday, November 27 Topic for the Week: Analyzing Data 4–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 , Chapter 10 Lab R Code

### *Lecture Notes*

Chapter 11, Unsupervised Learning `11-unsupervised.Rmd`

### *Assignments*

Assignment 12 Due Midnight Sunday, November 26

## **Wednesday, November 29 Unsupervised Learning, continued**

### *Standing Meetings*

*Lab Practical:* K-means clustering, nearest neighbor classification

## **Monday, December 4 Class Presentations**

Section 1

### *Assignments*

Assignment 13 Due Midnight Sunday, December 3

## **Wednesday, December 6 Class Presentations**

Section 2

**Final Projects Due Monday, December 11, midnight**