DS 1000: How Data Shape Our World

Josh Clinton and Will Doyle

August, 2021

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.

This course will teach students how data science is used to make key decisions across multiple domains. Data science involves bringing together programming and high powered computing to analyze data using statistics. Data scientists provide insights that allow leaders in business, government and other institutions to make decisions. Students will engage directly with problems in data science, with hands-on work with real world data. Each topic area will also engage with leading intellectuals who are working in a given area.

We assume you know *nothing* about data, statistics, or programming. This is an introductory level class aimed at introducing and motivating the study of data in a way that focuses on both technical and critical thinking skills. Both are essential.

This class will focus on using various real-world examples to teach important topics in data science and we will attempt to motivate the tools we learn via motivating questions. Students will learn about the following key concepts in data science:

- Structuring and Manipulating Data
- Data Visualization
- Making Predictions from Data
- Understanding Error in Predictions
- Training and Testing Models

The examples are mostly drawn from the social sciences given our backgrounds and expertise, but hopefully of interest:

- Predicting elections using polls: become your own 538.com or The Upshot
- Understanding college admissions - become a Dean of Admissions?
- Textual forensics - using words to predict authorship?
- COVID-19 and public health.

NOTE: This is the first time this class been taught so things will assuredly change once we are underway as we gauge the pace and understanding level of class.

Evaluation

Students will complete ten assignments individually. These assignments will ask you to implement the skills covered in a given week.

In addition, students will work in groups on five guided exercises. These exercise will challenge you to implement the skills we discuss in class. The focus of these reports will be explaining an applied data analysis to an external audience, and will include both text and data visualizations.

Finally, students must participate in class. The final grade will also be based on student participation, which includes attending course meetings and completing in-class prompts.

Summary of Evaluation:

Assignments: 10% Participation: 10% Guided Exercises: 80%

Required Texts

There is not really a required book for this course as no book really combines the critical thinking and technical skills we aspire to teach. As a result we are not requiring you to purchase or read anything, but there are a few recommendations we could make

The closest book in terms of *technical content* is:

Wickham, Hadley and Garrett Grolemund. *R for Data Science*. It is freely available to Vanderbilt students here. The associated web page with the book is here. The book is terse – focusing on making R "work" rather than how to think about working with data (and R).

Critical thinking will also be important and there are many interesting books about how to think about data (and the analysis of data). An extremely small set includes:

O'Neil, Cathy. *Weapons of Math Destruction* Silver, Nate. *The Signal and The Noise*. Tufte, Edwards. *The Visual Display of Quantitative Information*

Also stay attuned to news about issues and controversies involving data science.

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.

Honor Code Statement

All assignments for this class, including weekly problem sets and the final project, are to be conducted under the obligations set out in Vanderbilt's Honor Code. Please click here to review the Honor Code.

If you have any questions at all about the Honor Code or how it will be applied, ask us right away.

Health and Safety

Vanderbilt's Guidelines for Covid Precautions are Here

TL;DR:

- Don't come to class sick– if you feel a tiny bit unwell, do not come to class. These absences will ALWAYS be excused.
- Proper masks must be worn at all times. "Proper" means multilayer masks held in place with elastic. Buffs and bandanas are not appropriate and are likely worse than nothing.

- Masks cannot be taken off in class for any reason. Pulling down one's mask to speak kind of defeats the whole purpose.
- No food or drink in class. Your laptop will thank you for this rule.

Communication

Email: please put the phrase "DS 1000" in any email you send to us. While we try to respond to emails in a timely manner, don't expect a response in less than 24 hours. Our office hours are below:

Office Hours: Clinton - Tuesday & Thursday 9AM -11AM Please Use My Booking Page »»»> c150591cd33448f714f5285d23da30c223610d7c

Office Hours: Doyle: Tuesday & Thursday 12PM -2PM Please Use My Booking Page

Office Hours: Melissa Meisels Mondays, 4-5 pm

Office Hours: Qi Xu Mondays, 3-4 pm

Schedule

Intro: The wonderful world of data science

Readings/Resources

Download R

Download RStudio You want the Desktop version, free license

RStudio Introduction and Resources

Lesson Notes

Lecture notes: PDF .Rmd

Intro: Hello, World: The basics of interacting with statistical programming languages

Readings/Resources

Basic Basics from R Ladies Sydney

Lesson Notes

Lecture notes: HTML PDF .Rmd

Intro: Data wrangling

Readings/Resources

"The Gender gap was expected to be historic."

"Exit polls, election surveys and more"

Wickham & Grolemund, Chapter 3

Lesson Notes

Univariate Data Analysis: Descriptives

Readings/Resources

Lesson Notes

Univariate Data Visualization Using `ggplot`

Thursday, September 9, 2021[1]

Readings/Resources

Wickham & Grolemund, Chapter 1

Lesson Notes

Bivariate Data: More `ggplot`, Smoothing, and Conditional Means

Readings/Resources

Latest Polls - 538

Forecasting the US Elections - *Economist*

Lesson Notes

Assignment 3 Due September 13, Midnight

Sampling, Resampling, and Measuring (Some) Uncertainty: Two Amazing Results

Readings/Resources

Lesson Notes

Application: Predicting Elections Using Conditional Means and Maps

Readings/Resources

Electoral College

You are not expected to understand all of what is being discussed, but the explanations raise questions about what to do.

538 Model

270 to Win Presidential Election Forecast

Lesson Notes

Description, Prediction, & Causality: What are we doing?

Readings/Resources

Lesson Notes

[1]Note, this lecture and the next may take 3 lectures.

Conditional Means: application 1

Readings/Resources

Lesson Notes

Conditional Means: application 2

Readings/Resources

Lesson Notes

Conditional Means: workshop day

Readings/Resources

Lesson Notes

Regression Linear regression: graphics

Readings/Resources

Lesson Notes

Regression Regression: intro

Readings/Resources

Lesson Notes

Assignments

Regression: multiple regression

Readings/Resources

Lesson Notes

Regression Regression application

Readings/Resources

Lesson Notes

Regression Regression: workshop day

Readings/Resources

Lesson Notes

Regression Regression: uncertainty

Readings/Resources

Lesson Notes

Binary Outcomes: intro

Readings/Resources

Lesson Notes

Binary Outcomes: Uncertainty

Readings/Resources

Lesson Notes

Binary Outcomes: cross validation

Readings/Resources

Lesson Notes

Binary Outcomes: Model Tuning and Comparison

Readings/Resources

Lesson Notes

Binary Outcomes: application

Readings/Resources

Lesson Notes

Binary Outcomes: workshop

Readings/Resources

Lesson Notes

Unsupervised Learning: introduction

Readings/Resources

Lesson Notes

Unsupervised Learning: uncertainty

Readings/Resources

Lesson Notes

Unsupervised Learning: models selection and comparison

Readings/Resources

Lesson Notes

Unsupervised Learning Unsupervised: application 1

Readings/Resources

Lesson Notes

Unsupervised Learning Unsupervised: application 2

Readings/Resources

Lesson Notes