Vanderbilt University
Human and Organizational Development
Course Number HOD 2790
Fall 2014

Introduction to Data Science

William R. Doyle
Office: 207E Payne
Office Hours: Mondays and Wednesdays, 11:30-12:30 or by appointment
Email: w.doyle@vanderbilt.edu
Twitter: @wdoyle42
Phone: (615) 322-2904

Benjamin Skinner
Office: 207F Payne
Office Hours: by appointment
Email: b.skinner@vanderbilt.edu

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:

1. Tools of the Trade: R and Rstudio, Python and Canopy
2. Working with pre-processed data and flat files
3. Getting data from the web
4. Using databases

Analyzing Data Topics:

1. Conditional means
2. Regression

3. Classification
4. K-means and nearest neighbors clustering

Presenting Data Analysis Topics:

1. Descriptives: histograms, density plots, bar plots, dot plots
2. Scatterplots
3. Lattice graphics and small multiples
4. Maps

Evaluation

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

Problem sets: 65% Each week I will assign a problem set for students to complete. These probelm 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.

Texts, Software, and Resources

Texts

There are three required texts for the course:

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

Referred to in this syllabus as *Silver* .This book will be used to give you a sense of what you should be trying to accomplish with your final projects. It won't give you any sense of HOW to do your final projects.

O'Neil, C. and Schutt, R. (2013). *Doing Data Science: Straight Talk from the Frontline.* " O'Reilly Media, Inc."

Referred to in this syllabus as *O'Neil and Schutt.* This is a book written by two authors after teaching a class much like this one. It has a great deal of useful information on thinking about and completing projects in data science.

Lander, J. P. (2013). *R for Everyone: Advanced Analytics and Graphics*. Addison Wesley Data & Analytics Series. Addison-Wesley Professional

Your go-to resource for help with R.

McKinney, W. (2012). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. " O'Reilly Media, Inc."

Your go-to resource for doing data analysis in Python.

I've placed three books on reserve for the class:

Tufte, E. R. (2001). *The visual display of quantitative information*. Graphics Press, Cheshire, Conn., 2nd ed edition

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

Tufte, E. R. (1997). *Visual explanations: images and quantities, evidence and narrative*. Graphics Press, Cheshire, Conn

You should take a look at these for ideas and insipration–I've noted the sections that are most helpful in various parts of the syllabus.

Software

We will use two tools for data acquisition, analysis and presentation. The first will be R, an open-source data analytics platform. R appears to be the most widely used data analysis software in data science. The second will be Python, an open-source general programming tool. Python is widely used by data scientists, particularly for more open-ended programming tasks. Last, we will use a tool for version control and sharing code called git. All of these tools are completely free. I will help you get set up and running with the software for the course in the first few weeks.

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.

Communication

My office is in 207E Payne, and my phone number is (615) 322-2904. Please always feel free to stop by during office hours (Mondays and Wednesdays 11:30-12:30) 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 2790" in your subject line. I will search for these messages every time I log on. You can also use OAK'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.

Honor Code Statement

All assignments for this class, including quizzes, policy memos, midterm, and final are to be conducted under the obligations set out in Vanderbilt's Honor Code. Please click here 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 utlize 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.

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

Schedule of Meetings

August 20

Topics

Class introductions

August 25

Topic for the Week: Getting Data 1—Tools of the Trade

Resources:

Silver, Chapters 1-4

Lander, Chapters 1-4

O"Neil and Schutt Chapter 1

RStudio Intro and Resources:

http://www.rstudio.com/resources/training/online-learning/

Enthought Canopy:

https://www.enthought.com/products/canopy/

Github:

https://github.com/

https://mac.github.com/

https://windows.github.com/

August 27

Topic for the Week: Getting Data 1—Tools of the Trade
Standing meetings
Practical: Using R and Rstudio, Python and Canopy
Resources
Lander, Chapters 1-4 (again)
McKinney, Chapters 1-3

September 1

Topic for the Week: Analyzing Data 1—Conditional means
Resources:
Silver, Chapter 5-9, 12-13
Lander, Chapters 11 and 15
O'Neil and Schutt Chapter 2
McKinney Chapter 9 (optional)
Assignment 1 due August 31

September 3

Standing Meetings
Lab Practical: Conditional Means

September 8

Topic for the Week: Presenting Data 1—Descriptives
Subtopics: histograms, barplots, dot plots
Resources:
Lander, Chapter 7
McKinney Chapter 8 (optional)
Assignment 2 due September 7

September 10

Standing meetings
Lab Practical: Presenting results in graphical format: histograms, barplots, dot plots

September 15

Topic for the Week: Getting Data 2—pre-processed data, flat files
Resources
Lander, Chapters 5 and 6
McKinney Chapter 5 (optional)
Assignment 3 due September 14

September 17

Standing meetings
Lab Practical: working with various data formats

September 22

Topic for the Week: Analyzing Data 2—Linear Regression
Resources
Lander, Chapter 16
ONeil Pages 51-71
Assignment 4 due September 21

September 24

Standing meetings
Lab Practical: Using linear regression for prediction

September 29

Topic for the Week: Presenting Data 2—Scatterplots
Resources
Tufte, 2011, chapter 4 (On Reserve)
Assignment 5 due September 28

October 1

Standing Meetings
Practical: presenting data using scatterplots

October 6

Topic for the Week: Getting Data 3—Getting data from the web
Resources
Lander, Chapter 6.7.
McKinney, Chapter 6. (Optional)
Assignment 6 due October 5

October 8

Standing meetings
Practical: Web scraping, HTML,JSON, XML formats

October 13

Topic for the Week: Analyzing Data 3—Classification
Resources
O'Neil and Schutt, Chapter 5, p. 184-198
Lander, Chapter 17.1, 20.4, and 20.5
Assignment 7 due October 12

October 15

Standing meetings
Practical: logistic regression, classification and regression trees

October 20

Topic for the Week: Presenting Data 3—Lattice Graphics and Small Multiples
Resources
Tufte, 1990, p. 67-80
Tufte, 1997, p. 79-105
Assignment 8 due October 19

October 22

Standing Meetings
Practical: Lattice graphics

October 27

Topic for the Week: Getting Data 4: Databases
Resources
Ripley, "ODBC Connectivity in R"
http://cran.r-project.org/web/packages/RODBC/vignettes/RODBC.pdf
Assignment 9 due October 26

October 29

Standing Meetings
Practical: SQL, MySQL, MS Access databases

November 3

Topic for the Week: Analyzing Data 4: K-Means Clustering and Nearest Neighbors
Resources
O'Neil and Schutt Chapter 8
Lander Chapter 22
Assignment 10 due November 2

November 5

Standing meetings
Practical: algorithms for matching

November 10

Topic for the Week: Presenting Data 4: Maps
Resources
Loecher "Plotting on Google Static Maps in R"
http://www.icesi.edu.co/CRAN/web/packages/RgoogleMaps/vignettes/RgoogleMaps-intro.pdf
Assignment 11 due October November 9

November 12

Standing Meetings
Lab Practical: Mapping

November 17

Topic: Open Week
Assignment 12 due October November 16

November 19

Practical: Providing students help with ongoing projects

November 24

No Class, Thanksgiving Break

November 26

No Class, Thanksgiving Break

December 1

Final Presentations

December 3

Final Presentations
Final Assignment Due December 10