Course Syllabus

Jump to Today ✎ Edit

SI 370 – Data Exploration

Winter 2018

Instructor: Daniel M Romero <drom@umich.edu>

Instructional Aide: Ruihan Wang <ruihwang@umich.edu>

Lecture/Lab: 1255 NQ, Tuesday/Thursday 2:30 - 4:30PM

Office hours

Romero: Mondays 3-5pm (3340 NQ)

Wang: Wednesdays 1-3pm (1277 NQ)

Or by appointment.

Course Description

SI370 aims to help students get started with their own data acquisition and exploratory data analysis (EDA). EDA is crucial in evaluating and designing solutions and applications, as well as understanding information needs. Students in this course will learn basic concepts and techniques of exploratory data analysis, using scripting, text parsing, structured query language, regular expressions, graphing, visualization, and clustering methods to explore data. Students will be able to make sense of and see patterns in otherwise intractable quantities of data. Much of the class will be using Python to implement solutions.

Where this course fits in the curriculum

This course offers advanced material for those in the BSI beyond the 106/206 sequence. You will find SI330 (Data Manipulation) useful to take in advance of this course. In SI330 you'll learn more about the data structures for analysis or how to obtain data. We will do our best to provide you with "clean" data so that manipulation is not strictly necessary (but in real world settings, both obtaining and manipulating as well as analysis skills are necessary).

The focus of the class is on data exploration--how we can understand new datasets (with a focus on applications for people, by people, and about people) and ask high level questions. Exploratory Data Analysis is a critical piece of the "sensemaking" that goes into any analytical workflow: we often need to get a sense of what is going on with the data in order to better approach Confirmatory Data Analysis. That is, we want to understand what we have before we decide on the statistical analysis or other experiments that we might want to do. We apply ideas from statistics in this process, but often what we glean from our data

will guide the statistical analysis we do after. Visualization plays a key role in this process as it provides a visual summary of the data in a way that we can understand it.

Though we largely focus on the exploration process of the analyst, where one or a small group are looking at the data, the tools you will learn in the class apply more broadly to the communication of information. For example, how you might visualize the data and convey it to others. You'll find this useful in many of the other classes (as well as in "real-world" settings) where you have to present information or results to others.

Learning Objectives

Competency:

- Apply tools of EDA in new situations (specifically techniques/methods/workflows to: (a) maximize insight into a data set; (b) uncover underlying structure; (c) extract important variables; (d) detect outliers and anomalies; (e) test underlying assumptions; (f) develop parsimonious models; and (g) determine optimal factor settings. http://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm
- Compute and visualize summary statistics of datasets (specific implementation in Python, but general understanding of how to use other languages/systems)
- Combine the use of graphical aesthetics with data manipulation to visualize relationships between variables
- Use factors to analyze categorical data and exploratory clustering analysis for unlabeled data.
- Produce polished information graphics for publication.

Literacy:

- Be familiar with basic concepts and design principles of information visualization
- Be familiar with basic analysis and techniques for time series data, multidimensional data, network data, and text data

Awareness:

- Be aware of Python modules to process complex types of data such as networks, time series, and images.
- Be aware of advanced data visualization techniques.

Text Books

There will be no required text for this class, we will be providing PDFs as necessary. That said, there are a number of books that expand on the topics we will be talking about and can be useful for you as you make progress through your homework.

Recommended:

- Foster Provost and Tom Fawcett, Data Science for Business (2013)
- Joel Grus, Data Science from Scratch (2015)
- Wes McKinney, Python for Data Analysis (2012)

- Tamara Munzner, Visualization Analysis and Design (2014)

All are available from the Proquest site for free:

on campus: **http://proquest.safaribooksonline.com/ (http://proquest.safaribooksonline.com/)**,
off campus: **http://proquest.safaribooksonline.com.proxy.lib.umich.edu (http://proquest.safaribooksonline.com.proxy.lib.umich.edu/)**

Tentative weekly topics

Week of	Topics
Jan 1	Intro, welcome to SI370, basics of EDA
Jan 8	Python for Data I and Intro to Pandas
Jan 15	More Pandas
Jan 22	Basic Statistics
Jan 29	More Statistics and Data Cleaning
Feb 5	Resampling and Intro to Visualization
Feb 12	More Visualization
Feb 19 **Midterm**	Review and midterm
Feb 26 **Spring Break**	
March 5	Multidimensional Data
March 12	Time Series

March 19	Clustering
March 26	Classification
April 2	Networks
April 9	Text
April 16	Project Presentations

Accommodations for students with disabilities

If you think you need an accommodation for a disability, please let me know at your earliest convenience. Some aspects of this course, the assignments, the in-class activities, and the way we teach may be modified to facilitate your participation and progress. As soon as you make me aware of your needs, we can work with the Office of Services for Students with Disabilities (SSD) to help us determine appropriate accommodations. SSD (734-763-3000; http://www.umich.edu/sswd/) typically recommends accommodations through a Verified Individualized Services and Accommodations (VISA) form. I will treat any information that you provide in as confidential a manner as possible.

Class format

We will spend the first hour discussing the high level theory behind our tools. The second hour will be an in-class lab. Labs will help you apply what you learned about in class and will get you going on homework.

Labs are intended to help you learn how to practically build what you learn about in lecture. They also give you the chance to ask questions about things you don't understand. You can assume that some of what you learn in lab will show up in your homework assignments. **Please bring your laptop (if this is a problem for you, please let us know ASAP).**

Grading

Grades breakdown:

Attendance: 5%
Labs: 15%

Weekly assignments: 30%

Midterm (In class on Feb 22): 20%

Project: 30%

Letter grade assignment:

[98,100] = A+	[87, 90) = B+	[77, 80) = C+	[67, 70) = D+
[93, 98) = A	[83, 87) = B	[73, 77) = C	[63, 67) = D
[90, 93) = A-	[80, 83) = B-	[70, 73) = C-	[60, 63) = D-
			[0, 60) = E

Collaborating on Assignments

We strongly encourage collaboration while working on some assignments. Active learning is effective. Collaboration with other students in the course will be especially valuable in summarizing material and picking out the key concepts. **You must, however, write your assignment submission on your own, in your own words**. If you receive assistance on an assignment, please indicate the nature and the amount of assistance you received. If the assignment is computer code, add a comment indicating who helped you and how. Any excerpts from the work of others must be clearly identified as a quotation, and a proper citation provided (e.g., in the comments of the code if it is a code fragment you have borrowed).

Google/StackOverflow/etc.

Use these! Make an effort to discover the answers on your own by using the Web. You'll be surprised how often you can find related code to help you. Give us the URLs of where you found help and even if you don't totally solve the problem, if it looks like you were heading the right direction in finding the solution we'll give you some credit.

Plagiarism

Unless otherwise specified in an assignment all submitted work must be your own, original work. Any excerpts from the work of others must be clearly identified as a quotation, and a proper citation provided. Any violation of the School's policy on Academic and Professional Integrity will result in serious penalties, which might range from failing an assignment, to failing a course, to being expelled from the program. Violations of academic and professional integrity will be reported to UMSI Student Affairs. Consequences impacting assignment or course grades are determined by the faculty instructor; additional sanctions may be imposed by the Senior Associate Dean for Academic Affairs.

Free Insurance Policy

We understand that unexpected events out of your control such as illness, job interviews, and family emergencies happen. That's why we will: (i) drop two assignments with the lowest grade and (ii) allow two missed lectures/labs without penalty. No explanations are required to use these "freebies". At the same time, late assignments will not be accepted and no additional make-up opportunities will be granted. Please do not ask for extensions or exceptions.

Extra Credit

From time to time, we will offer the class extra credit opportunities. Extra credit points will be applied to the current week's assignment and scores are capped at 100%. Extra credit points are not meant to increase your grade significantly and only a small number of them will be offered. Please do not ask for additional extra credit points.

This syllabus, lectures, labs, and other class material draw from earlier versions of this course, particularly those of Eytan Adar and Chris Teplovs.

Course Summary:

Date	Details