SI 370 - Data Exploration

Fall 2017: MW 4:00-6:00pm 1255 NQ

Note: Some syllabus details may be subject to change.

Note: This syllabus draws heavily from earlier iterations of this and similar courses, particularly those of Eytan Adar and Kevyn Collins-Thompson.

Instructor: Chris Teplovs (cteplovs@umich.edu)
Office: 3389 NQ
Office hours: Tuesdays 3:00-5:00pm or by appointment, **3377 NQ (not my office)**

GSI: Kaifeng Chen (chenkf@umich.edu)
Office hours: Thursdays 4:00-6:00pm or by appointment, 1270 NQ

If you have questions about course material, homework, or labs, please feel free to come and talk with us during office hours. You can also contact us via email: please put "**370**" in the subject line so we can be sure to attend to it. (Please note that we may not be available on email over the weekend.)

We try to answer questions that are sent to me directly within about 24 hours. Responses on weekends and holidays may be slower. The course has a discussion board using Piazza set up on Canvas – if the answers to your questions might be of interest to any other students (in particular, programming or course logistics questions), we encourage you to post your questions to Piazza (and we may cross-post your question to Piazza myself if it makes sense).

TL;DR

This course is designed to provide an opportunity for you learn how to use Python to explore data. You will need to show up for lectures and labs. You'll have a mid-term exam, a bunch of lab and homework assignments, and a final project that you'll need to complete on your own. Help each other out but don't cheat. Have fun!

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 and use. Students in this course will learn basic concepts of information visualization and techniques of exploratory data analysis, using scripting, text parsing, structured query language, regular expressions, graphing, and clustering methods to explore data. Students will be able to make sense of and see patterns in otherwise intractable quantities of data. Though the focus of the class is a high-level understanding of EDA 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 that SI330 (Data Manipulation) useful to take in advance of this course (or at the same time) but this is not required this year. 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 power 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. Because of this, visualization will feature heavily in this class.

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 (Links to an external site.)Links to an external site.)
- 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 visualization 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.

Textbooks

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/., offcampus: http://proquest.safaribooksonline.com.proxy.lib.umich.edu

Schedule/Readings

Week 1	**September 4-8, 2017**	**Homeworks: none**
Sept. 6, 2017	Intro, welcome to SI370, basics of EDA	Lab: Tools of the trade (Jupyter notebooks)

Readings:

- Leek, Jeff, The Elements of Data Analytic Style, (2015). Read chapters 1-5, 13 and 14 (skim the rest if you want, it's a very short "book"). Available at datastyle.pdf .

Week 2	**September 11-15, 2017**	**HW1 out (9/15)**
Sept. 11, 2017	Python for Data I	Lab: Python for Data I

Readings:

- Grus, Joel. "A Crash Course in Python", Data Science from Scratch. Chapter 2. (2015). Available at grus-chapter2.pdf .
- McKinney, Wes. "IPython: An Interactive Computing and Development Environment". *Python for Data Analysis*. Chapter 3. (2012) - only pages 45-57 and page 72 required. Available at mckinney-chapter3.pdf .

Sept. 13, 2017 Pandas I Lab: Pandas I

Readings:

- R. Jordan Crouser, Data Wrangling with Python and Pandas (2015) Available at DataWrangling.pdf .
- McKinney, Wes. "Getting Started with pandas". *Python for Data Analysis*. Chapter 5. (2012) - Available at Python_for_data_analysis_567.pdf (as chapters 5-7).

Week 3 September 18-22, 2017 HW1 due, HW2 out (9/22)

Sept. 18, 2017 Pandas II Lab: Pandas II

Readings:

- McKinney, Wes. "Data Loading, Storage, and File Formats". *Python for Data Analysis*. Chapter 6. (2012) - Available at Python_for_data_analysis_567.pdf (as chapters 5-7).
-

Sept. 20, 2017 Pandas III Lab: Pandas III

Readings:

- McKinney, Wes. "Data Wrangling: Clean, Transform, Merge, Reshape". *Python for Data Analysis*. Chapter 7. (2012) - Available at Python_for_data_analysis_567.pdf (as chapters 5-7).

Week 4 **September 25-29, 2017** **HW2 due, HW3 out (9/29)**

Sept. 25, 2017 Basic Stats I (Univariate + Chi-squared) Lab: Python for Stats I

Readings:

1. Grus, Joel. "Statistics", Data Science from Scratch. Chapter 2. Pages 57-62, (2015). Available on proquest, but also: grus-chapter5-stats.pdf
2. Shasha and Wilson, "Statistics is Easy! 2nd ed" Sections 4.1, 4.2, 4.3 (2011). Available at Morgan Claypool, but also: statistics_is_easy_2.pdf

Sept. 27, 2017 Basic Stats II Lab: Multivariate Analysis I

Readings:

1. Grus, Joel. "Statistics", Data Science from Scratch. Chapter 2. Pages 57-62, (2015). Available on proquest, but also: grus-chapter5-stats.pdf
2. Shasha and Wilson, "Statistics is Easy! 2nd ed" Sections 4.1, 4.2, 4.3 (2011). Available at Morgan Claypool, but also: statistics_is_easy_2.pdf

Week 5 **October 2-6, 2017** **HW3 due, HW4 out (10/6)**

Oct. 2, 2017 Basic Stats III Lab: Multivariate Analysis II

Readings:

1. Shasha and Wilson, "Statistics is Easy! 2nd ed" Sections 4.1, 4.2, 4.3 (2011). Available at Morgan Claypool, but also: statistics_is_easy_2.pdf
2.

Oct. 4, 2017 Data Cleaning & Outliers Lab: Outliers

Readings:

1. David Lane, "Influential Observations" Available here: http://onlinestatbook.com/2/regression/influential.html

Week 6	**October 9-13, 2017**	**HW4 due, HW5 out (10/13)**
Oct. 9, 2017	Resampling	Lab: The Bootstrap

Readings:

I haven't found a great Python-focused chapter, but this one is decent (in R). Ignore the programming bits for now:

1. Karthik Ramasubramanian, Abhishek Singh, Chapter: "Sampling and Resampling Techniques." Available on Springer (Links to an external site.)Links to an external site. and also here: bootstrap_chapter.pdf

Oct. 11, 2017	Intro to Vis I (perception)	Lab: Tableau I

Readings:

1. Christopher G. Healey, "Perception in Visualization," available here.
2. We will also be starting Tableau. In addition to the documentation in Tableau, you can also find resources on proquest.

Week 7	**October 16-20, 2017**	**HW5 due (10/20)**
Oct. 16, 2017	Review	Lab: Review
Oct. 18, 2017	Midterm (in class)	

Readings: None

Week 8	**October 23-27, 2017**	**HW6 out, Project proposal due (10/27)**
Oct 23, 2017	Vis II (good design)	Lab: Tableau II

Readings:

1. Tufte, Edward, The Visual Display of Quantitative Information, Chapters 4-6, (2011). Available: Edward_tufte_the_visual_display_of_quantitative_informationChaps456.pdf

Oct. 25, 2017	Intro to Vis III (good design II)	Lab: Python Vis

Readings:

1. Segel and Heer, "Narrative Visualization: Telling Stories with Data," Available: 2010-Narrative-InfoVis.pdf .
2. Grus, Joel. "Visualizing Data", Data Science from Scratch. Chapter 3. (2015). Available on proquest (Links to an external site.)Links to an external site.site, but also: grus-chapter3-vis.pdf
3. McKinney, Wes, "Plotting and Visualization", Python for Data Analysis. Chapter 8. (2012). Available on proquest (Links to an external site.)Links to an external site. site, but also: mckinney-chapter8-plotting-vis.pdf

Week 9	**October 30-Nov. 3, 2017**	**HW6 due, HW7 out (11/3)**

Oct. 30, 2017	Many Dimensions I: General Techniques	Lab: Practical multidim/multivar

Readings:

1. Alfred Inselberg, "Multidimensional Detective" Inselberg1997.pdf

Nov. 1, 2017	Many Dimensions II: Dimensionality Reduction	Lab: Practical Dim Red

Readings:

1. Chapman et al. R for Marketing Research and Analytics, Chapter 8, available r_for_marketing_chap8.pdf (yes, it's about R, ignore that bit, we'll show you how to do similar things in Python)
2. Urdan, Statistics in Plain English, Chapter 15, available stats_in_plain_english_chap15.pdf

Week 10	**November 6-10, 2017**	**HW7 due, HW8 out (11/10)**
Nov. 6, 2017	Time Series I	Lab: Time Series I

Readings:

1. Jake VanderPlas, Data Science with Python, Chapter 3.11, "Working with Time Series," available on github: https://github.com/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/03.11-Working-with-Time-Series.ipynb

Nov. 8, 2017	Time Series II	Lab: Time Series II

Readings:

1. Wes McKinney, Python for Data Analysis, Chapters 10 and 11: python_for_data_analysis_time_series.pdf

Week 11	**November 13-17, 2017**	**HW8 due, HW9 out (11/17)**

Nov. 13, 2017	Clustering I	Lab: Clustering I

Readings:

1. Provost & Fawcett, Chapter 6, "Similarity, Neighbors, and Clusters". Data Science for Business (2013) Available via chapter-6-provost.pdf .
2. Grus. Chapter 19, "Clustering". Data Science from Scratch (2015), Available at chapter-19-grus.pdf

Nov. 15, 2017	Clustering II	Lab: Clustering II

Readings: See readings for Tuesday

Week 12	**November 20-24, 2017**	**HW9 due, HW10 out (11/24)**
Nov. 20, 2017	Classification I	Lab: Classification I

Readings:

1. Grus, Joel, *Data Science from Scratch*. Chapters 13-17. (2015). Available at: data_science_from_scratch_chap13%2B17.pdf

Nov. 22, 2017	Classification II	Lab: Classification II

Readings: See Tuesday reading

Week 13	**November 27-Dec. 1, 2017**	**HW10 due**
Nov. 27, 2017	Network I	Lab: Network I

1. Aggarwal, Charu, "Social Network Analysis," Data Mining, Chapter 19. Available here: Data+Mining+The+Textbook-sna.pdf
2. Munzner, Tamara, "Arrange Networks and Trees," Visualization Analysis & Design, Chapter 9. Available here: networks_and_trees.pdf

Nov. 29, 2017	Network II	Lab: Network II

Readings:

3. Gephi Tutorial, Available here: gephi-tutorial-quick_start.pdf

Week 14	**December 4-8, 2017**	
Dec. 4, 2017	Text	Lab: Text

Readings: TBA

Dec. 6, 2017	Presentations I	
Week 15	**December 11, 2017**	**Final Reports Due (12/11)**
	Last Class: Presentations	

Class format

Class is on Mondays and Wednesdays for 2 hours. We will spend the first hour discussing the high level theory behind our tools. The second hour will be an in-class lab. We will often give a 10-20 minute introduction to the lab and allow you to work in class for the remainder of the time. Labs will help you apply what you learned about in class and will get you going on homework.

Labs portion of class

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. Labs are a mix of short lectures and activities that you will do in pairs. 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).***

Attendance

Attendance in the class and participation in labs is mandatory. Repeatedly missing class (> 3 times, unexcused) or failing to participate will lead to a failing grade.

Readings

There are readings assigned almost every week. These are intended to supplement the lectures and labs. You will get much more out of class if you read these before you attend. These will be posted in advance of class and made available in the course canvas site. You may find it useful to print (at least the shorter ones) and bring them to class/lab.

Electronics Policy

Laptops/phones/tablets/etc. are not allowed while anyone (myself, guest lectures, or students) is up in front of the classroom presenting or leading discussions. All the research, and my personal experience, suggests that everyone learns better and enjoys the course more when the distraction of electronics is removed. In most cases the slides will be available before lecture if you need to print them. We will also give you plenty of breaks so you will have time to get online.

Giving and Receiving Assistance

SI106 has this policy and I think it's appropriate here as well (slightly modified from the SI106, W14 syllabus). Learning technical material is often challenging. We are going to cover a wide range of topics in the course and we will move quickly between topics. Because it is our goal for you to succeed in the course, we encourage you to get help from anyone you like.

All that said: *You* are responsible for learning the material, and you should make sure that all of the assistance you are getting is focused on gaining knowledge, not just on

getting through the assignments. If you receive too much help and/or fail to master the material, you will crash and burn later in the semester.

The final submission of each homework exercise must be 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).

If you are a more advanced student and are willing to help other students, please feel free to do so. Just remember that your goal is to help teach the material to the student receiving the help. It is acceptable for this class to ask for and provide help on an assignment via the Q&A site (Piazza or Slack), *including posting (short!) code fragments*. Just don't post complete answers. If it seems like you've posted too much, one of the instructional staff will contact you to let you know, so don't worry about it. When in doubt, err on the side of helping your fellow students. To reiterate, the collaboration policy is as follows. Collaboration in the class is allowed (and even encouraged) for assignments – you can get help from anyone as long as it is clearly acknowledged. Collaboration or outside help is not allowed on exams, though you will be allowed to use some materials that you bring with you. Use of solutions from previous semesters is not allowed. The authorship of any assignments must be in your own style.

Google/StackOverflow/etc.

Use these! Seriously… 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!**

Homework Assignments

Homeworks will be due on Friday at 11:59pm (see lateness policy below). We will also release new homeworks on Friday for the next week. You can expect that some of the material in the homeworks will be covered in the labs (in the week of the release) so it's in your interest to do as much of the labs as possible in class and make sure you understand the material. The goal of the assignments (and labs) is to make sure that

you understand how to use the material in class on practical/real-world problems. We will try to provide you with real or realistic data and problems as much as possible -- things you would actually encounter if you were doing data exploration in a professional context.

See above for information on Giving/Receiving help. Copying wholesale (and/or failing to acknowledge your source) will be considered cheating and will be turned over to the academic advising office. Note that I am extremely good at catching people doing this. If you're good enough to hide the fact that you did it, you probably are better off just doing the assignment.

Midterm

There will be a 90-minute written test in late October (by default during the class period, but depending on class size we may need to switch to a bigger room: specific date and room to follow shortly), covering all previous material in the course to that date. Except in cases of serious illness or other emergencies, you are expected to take the midterm on the scheduled date. We'll hold a review session prior to the midterm. You will be allowed to bring certain materials into the exam with you (no electronics), but it will not be open book.

Final Project

We will provide additional information on the final project closer to that time. The goal is for you to try things that you learned in class on real datasets. We will provide some ideas and some datasets for you to think about and give you some examples as we get closer to the proposal date. Whereas the proposal contributes only 5% to this component of the grade (1.5% of the overall grade), you cannot submit a final project without having completed the proposal. This will be an individual project.

Grading

The graded work in the course will be weighted roughly as follows to determine a final percentage grade:

Course and Lab Participation: 20%

Homework Assignments:	30%
Midterm:	20%
Project:	30%

Late Policy

I realize that the occasional crisis might screw up your schedule enough to require a bit of extra time in completing a course assignment. Thus, I have instituted the following late policy that gives you a limited number of flexible “late day” credits.

You have **three (3)** free late days to use during SI 370. One late day equals exactly one 24-hour period after the due date of the assignment (including weekends). No fractional late days: they are all or nothing. Once you have used up your late days, **25% penalty** for each subsequent 24h period after the deadline that an assignment is late For example, if due date is 5pm Friday, with no late days left, penalties would be:

Before 5pm Saturday: 25% deduction

Before 5pm Sunday: 50% deduction

Before 5pm Monday: 75% deduction

After 5pm Monday: 100% deduction

You don't need to explain or get permission to use late days, and we will track them for you. In cases where late days can be assigned in multiple ways (e.g. you have only one late day left but hand in two late assignments) we will always allocate late days in a way that maximizes your grade. Note that resubmissions after the deadline will be counted as late submissions. **Also, late days may not be applied to the final project.**

There will be multiple opportunities for extra credit during the course, although not enough to make up for missing multiple assignments.

Bonus Credit Opportunities

- Active involvement in online discussion forum (piazza), including both raising meaningful questions and providing help to others, as determined by the instructors
- Active involvement in discussions in lectures/labs, as determined by the instructors.

- Completing bonus tasks and challenges in homeworks and lab sessions.
- Other achievement or contribution as determined by the instructors to be worthy of bonus credits.

E-mail policy

We will do our best to answer questions by e-mail within ~24 hours (just don't count on a response an hour before a deadline though responses on weekends may be slower). A discussion board will be set up on Piazza/Slack. You might be better off posting your questions there (especially programming questions).

Original Work

See above for information on Giving/Receiving help (which applies exclusively to homeworks). Unless otherwise specified in an assignment, all submitted work must be your own, original work. You may discuss general approaches with others on individual assignments, but may not copy code or answers wholesale and must indicate on your turned-in assignment who you worked with and how. 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 severe penalties, which might range from failing an assignment, to failing a course, to being expelled from the program, at the discretion of the instructor and the Associate Dean for Academic Affairs.

One caveat in regards to SI370 (Data Exploration) or any other related class. You may not submit the same work to both classes for your final project. We are aware of the students in both and while it is ok to broadly tackle the same problem, the work you turn in for both classes must be significantly different. If you're not sure about what that means, come talk to us.

Accommodations for Students with Disabilities

If you think you need an accommodation for a disability, please let us 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 us 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. We will treat any information you provide as private and confidential.

Student Mental Health and Wellbeing

The University of Michigan is committed to advancing the mental health and wellbeing of its students, while acknowledging that a variety of issues, such as strained relationships, increased anxiety, alcohol/drug problems, and depression, directly impacts students' academic performance.

If you or someone you know is feeling overwhelmed, depressed, and/or in need of support, services are available. For help, contact Counseling and Psychological Services (CAPS) at (734) 764-8312 and https://caps.umich.edu/ during and after hours, on weekends and holidays or through its counselors physically located in schools on both North and Central Campus. You may also consult University Health Service (UHS) at (732) 764-8320 and https://www.uhs.umich.edu/mentalhealthsvcs, or for alcohol or drug concerns, see www.uhs.umich.edu/aodresources.

For a more comprehensive listing of the broad range of mental health services available on campus, please visit: http://umich.edu/~mhealth/