CAS CS 591
Computational Tools for Data Science
Fall 2016

Meeting Place: SCI 117

Meeting Time: TR 11-12:30

Instructor: Prof. Mark Crovella

- **Office:** MCS-140E
- **Office Hours:** M 2-3:30, R 3-4:30
- **Email:** crovella@bu.edu

Teaching Fellow: Ms. Katherine Missimer

- **Office Hours:** W 4-5:30, F 5-6:30
- **Office Hours Location:** Undergrad Lab, EMA 302
- **Lab Tutoring Hours:** F 3-5.
- **Email:** kzhao@bu.edu

Overview of the Course

This course is targeted at students who require a basic level of proficiency in working with and analyzing data. The course emphasizes practical skills in working with data, while introducing students to a wide range of techniques that are commonly used in the analysis of data, such as clustering, classification, regression, and network analysis. The goal of the class is to provide to students a hands-on understanding of classical data analysis techniques and to develop proficiency in applying these techniques in a modern programming language (Python).

Broadly speaking, the course breaks down into three main components, which we will take in order of increasing complication: (a) unsupervised methods; (b) supervised methods; and (c) methods for structured data.

Lectures will present the fundamentals of each technique; focus is not on the theoretical underpinnings of the methods, but rather on helping students understand the practical settings in which these methods are useful. Class discussion will study use cases and will go over relevant Python packages that will enable the students to perform hands-on experiments with their data.

Prerequisites: Students taking this class must have some prior familiarity with programming, at the level of CS 105, 108, or 111, or equivalent. CS 132 or equivalent (MA 242, MA 442) is required. CS 112 is also helpful.

Learning Outcomes

Students who successfully complete this course will be proficient in data acquisition, manipulation, and analysis. They will have good working knowledge of the most commonly used methods of clustering, classification, and regression. They will also understand the efficiency issues and systems issues related to working on very large datasets.

Readings

There is no text for the course. Lecture notes will be posted online.

Some of the lectures are based on *Introduction to Data Mining,* by Tan, Steinbach and Kumar. This is a good place to go for more detail if something is not clear.

Some other recommended texts are:

1. Python for Data Analysis (http://shop.oreilly.com/product/0636920023784.do). This is the definitive text for *Pandas* which we will use quite a bit.

2. Programming Collective Intelligence (http://shop.oreilly.com/product/9780596529321.do)

Web Resources

The slides I use are actually executable python scripts, using the `jupyter notebook`. You can download and execute the lectures on your own computer, and you can modify them any way you'd like, play around with them, experiment, etc.

The slides I use in lecture are published on `github`. The repository is `https://github.com/mcrovella/CS505-Data-Science-in-Python`. If you want to access the repository using `git`, please feel free. If you find a bug, feel free to submit a pull request.

Homeworks and Project

1. There will nine homework assignments. In a typical assignment you will analyze one or more datasets using the tools and techniques presented in class.

   Homeworks will be submitted via `github`. For this, we need your github account (create one if you don't already have it). After you have created it, fill out the form at `https://goo.gl/forms/8W0SOdvMn07UKdip2` to let us know what it is.

   You are expected to work individually on homeworks.

2. In addition, there will be a final project. For the project you will extract some knowledge or conclusions from the analysis of dataset of your choice. The analysis will be done using a subset of the methods we described in class. The final project will require a proposal, two progress reports, and a final presentation in poster form.

   The project will have three essential components: 1) a data collection piece (which may involve crawling or calls to an API, combining data from different sources etc), 2) a data analysis piece (which will involve applying different techniques we described in class for the analysis) and 3) a conclusion

component (where the results of the data analysis will be drawn). The students will submit a 5-page report explaining clearly all the three components of their project. Finally a poster presentation will be required where the students will be prepare to present their effort and results in front of their poster.

As an example, you may choose to collect data from Twitter related to a specific topic (e.g., Ebola virus) and then measure the intensity of posts about a topic in different areas of the world etc. Other examples of projects may include (but are not limited to): analysis of MBTA data, analysis of NYC data, crawling of YouTube (or other social media data) and analysis of social behavior like trolling, bullying etc.

The project is due by the last day of class (December 8). The project presentations will be given in the form of a final poster explaining components 1, 2 and 3 of the project.

You are expected to work in teams of two on the final project. I will leave it up to you to form teams on your own, but everyone must work in a team.

Piazza

We will be using Piazza for class discussion. The system is really well tuned to getting you help fast and efficiently from classmates, Ms. Missimer, and myself. Rather than emailing questions to the teaching staff, I encourage you to post your questions on Piazza. Our class Piazza page is at: `https://piazza.com/bu/fall2016/cs505/home`. We will also use Piazza for distributing materials such as homeworks and solutions.

When someone posts a question on Piazza, if you know the answer, please go ahead and post it. However pleased *don't* provide answers to homework questions on Piazza. It's OK to tell people *where to look* to get answers, or to correct mistakes; just don't provide actual solutions to homeworks.

Programming Environment

We will use `python` as the language for teaching and for assignments that require coding. Instructions for installing and using Python are on Piazza.

Course and Grading Administration

Homeworks are due at 7pm on Fridays. Assignments will be submitted using `github`. Ms. Missimer will explain how to submit assignments.

NOTE: IMPORTANT: Late assignments **WILL NOT** be accepted. However, you may submit **one** homework up to 3 days late. You **must** email Ms. Missimer before the deadline if you intend to submit a homework late.

Final grades will be computed based on the following:

50% Homework assignments.

50% Final Project

The exact cutoffs for final grades will be determined after the class is complete.

Academic Honesty

You may discuss homework assignments with classmates, but you are solely responsible for what you turn in. Collaboration in the form of discussion is allowed, but all forms of cheating (copying parts of a classmate's assignment, plagiarism from books or old posted solutions) are NOT allowed. We – both teaching staff and students – are expected to abide by the guidelines and rules of the Academic Code of Conduct (which is at `http://www.bu.edu/academics/policies/academic-conduct-code/`).

Graduate students must also be aware of and abide by the GRS Academic Conduct code at `http://www.bu.edu/cas/students/graduate/forms-policies-procedures/academic-discipline-procedures/`.

You can probably, if you try hard enough, find solutions for homework problems online. Given the nature of the Internet, this is inevitable. Let me make a couple of comments about that:

1. If you are looking online for an answer because you don't know how to start thinking about a problem, talk to Ms. Missimer or myself, who may be able to give you pointers to get you started. Piazza is great for this – you can usually get an answer in an hour if not a few minutes.

2. If you are looking online for an answer because you want to see if your solution is correct, ask yourself if there is some way to verify the solution yourself. Usually, there is. You will understand what you have done *much* better if you do that. So ... it would be better to simply submit what you have at the deadline (without going online to cheat) and plan to allocate more time for homeworks in the future.

Course Schedule

Date	Topics	Reading	Assigned	Due
9/6	Introduction to Python		HW 0	
9/8	Essential Tools (Git, Jupyter Notebook, Pandas)			
9/13	Probability and Statistics Refresher			HW 0
9/15	Linear Algebra Refresher			
9/20	Numpy, Scikit-learn, Distance and Similarity Functions			
9/22	Intro to Timeseries		HW 1.1	
9/27	Clustering I: k-means			
9/29	Clustering II: In practice		HW 1.2	
9/30				HW 1.1
10/4	Clustering III: Hierarchical Clustering			
10/6	Clustering IV: GMM and Expectation Maximization		HW 2.1, 2.2	
10/7				HW 1.2
10/11	NO CLASS; Monday Schedule			
10/13	Singular Value Decomposition I : Low Rank Approximation			
10/7				HW 2.1
10/18	SVD II: Dimensionality Reduction			
10/20	SVD III: Anomaly Detection		HW 3.1	
10/21				HW 2.2
10/25	Web Scraping			
10/27	Classification I: Decision Trees			
10/28				
11/1	Classification II: k-Nearest Neighbors			
11/3	Classification III: Naive Bayes, SVM			
11/4				HW 3.1, Proj Proposal
11/8	Regression I: Linear Regression		HW 3.2	
11/10	Regression II: Logistic Regression			
11/11				Prog Report 1
11/15	Regression III: More Linear Regression			
11/17	Recommender Systems		HW 4	
11/18				HW 3.2
11/22	Map Reduce			Prog report 2
11/24	NO CLASS; Thanksgiving Break			
11/29	Network Analysis I			
12/1	Network Analysis II		HW 5	
12/2				HW 4
12/6	Graph Clustering, Text Analysis			
12/8	Poster Session			
12/12				HW 5