COSC/MATH 4931: Introduction to Data Science

Spring 2017

Instructor:	Shion Guha	**Time:**	TuTh 11:00am - 12:15pm
Email:	shion.guha@marquette.edu	**Place:**	412 Cudahy Hall

Description: This course is a first introduction to the new and rapidly evolving field of data science. We will understand the philosophy and basic concepts of doing data science by applying and combining state-of-the- art concepts from mathematics, statistics and computer science. The student will learn how to use popular programming and data analysis platforms to clean, organize and analyze data from a modern machine learning perspective. Emphasis will be placed on interpretation and visualization.

Prerequisites: MATH 1451 (Calculus 2) and COSC 1010 (Intro Programming) I will assume that everyone knows how to search for information online and has access to a computer. If you don't please email me right away. Our classroom, 412 Cudahy has computers and I can arrange for students to have computer access setup if required. You can obviously, use your personal computing devices for this course. You will use a computer everyday in class so please plan accordingly.

Course Page: `https://www.shionguha.net/teaching/cosc4931sp17`

Office Hours: TuTh 12:30pm - 2:30pm at 318 Cudahy.

Books: There are no **required** textbooks for this class. Let's all save some money! I will post all recommended books, readings, datasets and code fragments on the course website as we progress through the class. All software used will be free and open source. You will be required to search for appropriate information online to complete the assignments.

Objectives: This course is a first introduction to modern data science for undergraduates. We will chiefly focus on an expansive, outward looking understanding of modern data science specifically featuring a survey of different statistical and numerical methods of analyzing data computationally. Thus, this course will draw from machine learning, statistics, mathematics and numerical analysis At the end of the course, a successful student should be able to:

- collect, process and clean data using the R-Python stack,
- analyze data, using a suite of modern analytic approaches,
- make rigorous inferences from data and diagnose problems in data analysis,
- visualize, write and explain inferences obtained from data in a time-bound deliverable form.

Timeline:

- Jan 17: introduction to the course; policies and expectations
- Jan 19: the data science stack: jupyter, publishing to github; basic python tomfoolery
- Jan 24: the data science stack: R + Stan, publishing via Shiny; basic R tomfoolery
- Jan 26: data cleaning; tidy data concepts
- Jan 31: data cleaning; tidy data problems

- Feb 02: guest lecture from a professional data scientist
- Feb 07: guest lecture from a professional data scientist
- Feb 09: clustering: hierarchical and centroid based models
- Feb 14: clustering: distribution and density based models
- Feb 16: project topic presentations
- Feb 21: project topic presentations
- Feb 23: classification: naive bayes approach
- Feb 28: code for mid-term project in class **(I am away for a conference)**
- Mar 02: code for mid-term project in class **(I am away for a conference)**
- Mar 07: classification: binary logistic models
- Mar 09: data science, validity and hypothesis testing; mid term project reports submission
- Mar 14: **spring break! no class!**
- Mar 16: **spring break! no class!**
- Mar 21: regression: linear models for continuous data
- Mar 23: regression: poisson models for count data
- Mar 28: decision trees
- Mar 30: random forests
- Apr 04: graphical models: hierarchical bayes
- Apr 06: better predictions: model regularization
- Apr 11: visualization; project topic paper submission
- Apr 13: **easter break! no class!**
- Apr 18: networks: sampling and structural properties
- Apr 20: networks: modeling tie strength
- Apr 25: interpretability in data science
- Apr 27: dimension reduction: principle component analysis
- May 02: dimension reduction: latent variable models
- May 04: ethics, fairness and accountability in data science
- May 13: **final project due!**

Grading Policy: Project Topic Presentation (20%), Mid Term Project Submission (30%), Project Topic Paper (20%), Final Project (30%). This course will **not** be graded on a curve. The final grades will depend on the following scores:

A: 93 - 100; AB: 85 - 92; B: 77 - 84; BC: 69 - 76; C: 61 - 68; CD: 53 - 60; D: 45 - 52; F: 0 - 44

All regrade requests will be limited to computational errors. I will provide partial credit for particular questions. Please refer to the course policy section for more details.

Important Dates:

February 16, 21: Project Topic Presentations
March 09: Mid Term Project Report Submission
April 11: Project Topic Paper Submission
May 13: Final Project Submission

Course Policy:

- You are responsible for your own progress. Please check your marquette email, the course website and d2l about course announcements regularly.
- Please bring a computing device everyday to class. We will be writing code together everyday. This is necessary. If you are unable to get a computing device, we will setup an account for you to use using the computers in our classroom.
- Each of you will create your own free github accounts to maintain your project and any code you write in class. This is part of a data science project's lifecycle and is expected to be shown to employers by data science job applicants.
- All submission of papers/project will be online via d2l. Any submission after the deadline will be considered late. I will post an exemplar and demonstrate in class, the right way to submit your mid term project.
- You may discuss your project and your paper with other students, but you must write up your papers and code independently in your own words.
- You must cite every reference in your papers and every library (if used) in your project in ACM style. Failure to do so will be regarded as a violation of academic integrity. Please refer to the section on academic integrity for more details.
- The project topic paper should be around 4 pages, ACM extended abstract style adhering strictly to the current CSCW conference norms. It will be due at 11:59 PM local time. I will inform you about the expected content at an appropriate time.
- Regular attendance is essential and expected. A student who incurs an excessive number of absences may be withdrawn from the class at the instructor's discretion.
- I will take in class attendance on the first day of class and again, just after withdrawl date to get an accurate count of class registration. Otherwise, I will not take attendance regularly.
- Please make sure your cell phone is turned fully off, or silent. No texting, reading emails, playing games, or anything else. I will reserve the right to ask you to leave the class if you are being distracting or disruptive.

Ott Memorial Writing Center: The Ott Memorial Writing Center offers free one-on-one consultations for all writers, working on any project, at any stage of the writing process. Marquette's writing center is a place for all writers who care about their writing, because every writer can benefit from conversation with an interested, knowledgeable peer. Writing center tutors can help you brainstorm ideas, revise a rough draft, or fine-tune a final draft. You can schedule a 30- or 60-minute appointment in advance (288-5542 or www.marquette.edu/writing-center), but walk-ins (in 240 Raynor or our other satellite locations) are also welcome. The Ott Memorial Writing Center also offers free workshops and hosts writing retreats.

Academic Integrity: Marquette University takes academic integrity very seriously. This is a core part of who we are as reflected by our Jesuit values. All students are required to take the Academic Integrity Tutorial. If you haven't, please go take it right now. All students are required to adhere to the Honor Pledge and follow the Honor Code. Please familiar yourself with the Academic Integrity website. There is a lot of useful information there.

I take a zero tolerance policy with violations of academic integrity. All papers and projects are run through a plagiarism detection software. If you are flagged, be assured that we will have a conversation. Let's not have that conversation shall we? If I determine that you have indeed violated academic integrity, you will receive a failing grade for that component of the course or for the entire course depending on the nature and severity of the violation. Please help me to make sure that there are no such incidents.

Accessibility Policy: If you have any accessibility needs, please contact Office of Disability Services (ODS) to register them as soon as possible. ODS works with students with documented disabilities to provide accommodations for their educational needs. The course has been designed for multiple different styles of learning. However, if you have any specific learning styles that you want me to know about which would not be addressed by AES, please reach out to me within the first week of class so I can try to accommodate.