COMP 598: Introduction to Data Science Syllabus

Overview

Data science is a cornerstone for our modern age. In the right hands, the tools of data science can transform large, unstructured piles of data into key insights that inform business decisions, automate industrial processes, or deliver nuanced understanding of opportunities or challenges. Data science is quickly becoming one of the major tools that governments, companies, and even individuals use to make important decisions.

That such important and wide-ranging decisions are based on the analysis of large and unwieldy data sets highlights to importance of doing it right. This is, at its heart, the purpose of this class: to teach you to the fundamentals of how to use the powerful tools of data science responsibly.

As a result, this class will take a holistic perspective on the practice of data science. The class will be technical – we will learn a diverse array of techniques ranging from data scraping to statistical modeling to visualization. The class will also be reflexive – we will develop an awareness of how even very well-intentioned analyses can completely misrepresent the real-world and lead to wrong insights. We will learn how to avoid this.

My goal is for you to leave this class both capable of applying data science in the real-world and cautious to ensure that you do so responsibly.

Class Schedule

ALL lectures will be pre-recorded and posted on MyCourses.

Videos will be posted on Sunday night (EST). The expectation is that the student will view them by that Thursday, when the homework for that week is posted.

Contact Information

Instructor: Professor Derek Ruths

Office Hours: See office hour information on the homepage of MyCourses

Mode of contact: Please, please **use my office hours as your first line for getting in touch with me**. For emergencies or other personal or private matters, please use my email, derek.ruths@mcgill.ca, and start your email title with "[COMP 598]". If you don't use this title opener, I can't guarantee that I will respond to your email quickly (I get a LOT of emails everyday, so if you don't flag it, I likely won't see it).

Teaching Assistants:

TBD

For office hours, consult the course homepage on MyCourses.

Homework Submission:

Through MyCourses, there will be an Assignment area for each assignment due.

Class Structure

Resources

There is no textbook for this class.

As part of this course, you will setup and run a cloud server – we will cover how to do this in Amazon Web Services, but you are also welcome to use Google Cloud, Digital Ocean, or any other cloud server system. We will cover all this in detail during lecture.

The cost for this server will be approximately $120 CAD. If you apply for the Github Developer Education Pack, you can get an AWS voucher for $100 USD, which will cover most/all of the cost for the semester. I encourage you to do this.

Grading

Each student's final grade in this course will be determined by approximately 12 homework assignments and a project report. There are no exams.

Grade calculation. Each student's grade will be determined be:

- 75% assignments: the total number of points a student receives over the semester will be divided by the total number of points they could have received on all assignments
- 25% final project report: this is a team assignment. the grading rubric will be circulated when the project is first assigned.

Assignments. There will be roughly one assignment each week.

- Timing: it will be assigned on Thursday morning, due the following Friday night (at 11:59 PM).

Project. Starting week 9, students will work in teams of three on a final project. They will do so independently (though with some structure provided each week). They are encouraged to check in with TAs and the instructor, but the only assessment on this project will be the final report, submitted on the last day of the semester.

Late Assignments

If submitting an assignment by the due date presents a problem, contact me as soon as possible to determine whether a late submission can be accommodated. If a later due date has not been arranged, then the late assignment's final grade will be penalized at 10% per day.

Extenuating Circumstances
I want every student in this course to succeed. If unforeseen situations arise that interfere with your ability to complete coursework or devote adequate time to this course, *please contact me as soon as you suspect there could be a problem.* While I cannot guarantee that I will oblige every request and situation, the sooner you notify me of the situation, the sooner we can work to find a way to accommodate any issues you may be dealing with. Please bear in mind that **requests that have waited till the last minute will not be accommodated.**

Academic Integrity
Except where specifically noted, homework may be discussed with other students and I encourage group work. However, all work (code, writing, and answers) must be the student's own. Copying another student's work, in any form, constitutes an act of cheating.

McGill University values academic integrity. Therefore, all students must understand the meaning and consequences of cheating, plagiarism and other academic offences under the Code of Student Conduct and Disciplinary Procedures (see www.mcgill.ca/integrity for more information).

Right to Submit Work in English or French
In accord with McGill University's Charter of Students' Rights, students in this course have the right to submit in English or in French any written work that is to be graded.

Schedule

Week	Units	Assignments
Week 1	(1) Welcome (2) What is data science?	HW 1: data science projects by hand
Week 2	(3) Your new home: Unix	HW 2: AWS exercises, CLI analysis of dataset
Week 3	(4) Core data science tools (python, git, jupyter, bokeh)	HW 3: basic analysis, run a live dashboard the TAs can grade.

Week 4	(5) Question formulation	HW 4: Question formulation
Week 5	(6) Data collection – scraping & organization	HW 5: Scraping and API collection for US election
Week 6	(6) Data annotation with keywords & manual coding	HW 6: Keyword annotation of text
Week 7	(7) Data annotation with crowd sourcing	HW 7: Build simple coding interface in flask
Week 8	(8) Responsible data annotation and collection	HW 8: Crowd source to another group PROJ Step 1: Evaluating different data collection strategies
Week 9	(9) Modeling – bias and statistical significance	PROJ Step 2: Data annotation
Week 10	(10) Analysis – visualization	PROJ Step 3: Sentiment analysis HW 10: Advanced visualization design
Week 11	(11) Analysis – characterizing error	PROJ Step 4: Analysis HW 11: Mixed method error analysis
Week 12	(12) Communication (presentation & writing)	PROJ Step 5: Report HW 12: Presenting error analysis