COMP 370: 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 and 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 the importance of doing it right. This is, at its heart, the purpose of this class: to teach you the fundamentals of how to use the powerful tools of data science responsibly and effectively.

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

Unless otherwise noted, all lectures will be in person. Recordings will be provided to the extent the classroom supports them and the hardware works. But they are not guaranteed.

Time: Monday/Wednesday from 11:35 AM – 12:55 PM.
Location: McConnell 204

Contact Information

For any class, homework, course-related questions, please use the course Discord server (link will be posted in MyCourses).

Instructor: Professor Derek Ruths

Office Hours: See office hour information in the #office-hours-schedule channel of the class Discord.

Mode of contact: **Use Discord and my office hours as your first line for getting in touch with me**. For emergencies or other personal or private

matters, email, derek.ruths@mcgill.ca, and begin the title with "[COMP 370]". 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 #office-hour-schedule channel in the class Discord server.

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

Class Structure

Resources

There is no textbook for this class.

Cloud server. 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.

If you turn the server off when you're not using it, the cost for this server will be no more than $120 CAD for the entire semester (probably a lot less).

ChatGPT. You are welcome to use GPT-based technologies for the course - in fact, I encourage responsible use. That said, you will be unable to use these technologies for any of the evaluations in this course. So it is vital that you use them as learning aids - not as a means of skipping "hard stuff". As a general rule, if you can't produce what GPT is providing you with, then you should spend time figuring that out.

Assessments

Each student's final grade in this course will be determined by in-class participation, homework submissions (checked only for submission, not correctness), 3 in-class exams, and a final project report.

Grade calculation. Each student's grade will be determined by a combination of graded assessments and class participation. The total available points a student will receive is calculated using the following components:

- 15% assignments (completion credit): any on-time homework assignment in which *some* attempt was made for each problem will receive full marks. Late submissions are not accepted.
- 40% in-class exams: this is across 3 in-class exams (each contributing approx. ~13% weight). 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 three exams.
- 25% final exam: a cumulative final exam during exam period
- 20% final project report: this is a team assignment. The grading rubric will be circulated when the projemycct is first assigned.

In-class participation via the Vevox system will determine the number of points (scored above) that are available to the student's final grade, according to this framework:

Minimum % of polls completed	Maximum grade
< 50	74
< 65	84
>= 75	100

Assignments. There will be one assignment each week.
- Timing: assignments will be posted on Wednesday, due the following Friday night (at 11:59 PM).
- Solutions: homework will not be graded or corrected. Solution guides will be posted to MyCourses immediately following the due date.

In-class exams. Exams will be given in-class as shown in the schedule. Exams will be hand-written ONLY in pen (if in pencil, regrades are not possible). Exams will start at 11:35 and end promptly at 12:55. Students registered with the SAA will be able to take their exam at SAA facilities.

Project report. Starting the week of November 3th, students will work in teams of 3-4 on a final project. Teams will be assigned randomly. Teams will work on their project 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.

Final exam. A 3-hour final exam will be administered during final exam period. It will be cumulative, covering all material in the course.

Regrade Deadline
Exams may be submitted for a regrade within 1 week of when exams were returned. The format for submitting a regrade request is to staple a printed page explaining the concern to the front of the exam booklet and hand it to the instructor at the beginning/end of a class.

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

The timing of each unit is approximate – though the homework assignment and exam timings are exact unless noted during classes.

Week	Units	Assessments
Aug 27	(1) Welcome (2) What is data science?	HW 1: data science projects by hand
Sept 1	(3) Your new home: Unix	HW 2: AWS exercises, CLI analysis of dataset
Sept 8	(4) Core data science tools (python, git, jupyter, bokeh)	HW 3: basic analysis, run a live dashboard.
Sept 15	(5) Question formulation	HW 4: Question formulation
Sept 22	(6) Data collection – scraping & organization	HW 5: Scraping and API collection **Exam 1 on Units 1-5 (Wednesday)**

Sept 29	(6) Data annotation with keywords & manual coding	HW 6: Keyword annotation of text
Oct 6	(7) Data annotation with crowd sourcing	HW 7: Build simple coding interface in flask
Oct 13	– Break –	– Break –
Oct 20	(8) Responsible data annotation and collection	HW 8: Crowd sourcing **Exam 2 on Units 6-7 (Wednesday)**
Oct 27	(9) Modeling – selecting methods	HW 9: Modeling methods
Nov 3	(9) Modeling – bias and statistical significance	PROJ Step 1: Evaluating different data collection strategies HW 10: Bias
Nov 10	(10) Analysis – visualization	PROJ Step 2: Data annotation HW 11: Advanced visualization design **Exam 3 on Units 7-9 (Wednesday)**
Nov 17	(11) Analysis – characterizing error	PROJ Step 3: Sentiment analysis HW 12: Mixed method error analysis
Nov 24	(12) Communication (presentation & writing)	PROJ Step 4: Analysis HW 13: Presenting error analysis
Dec 1	(13) Careers in Data Science	PROJ Step 5: Report HW 14: Job Search