COMP 370
Introduction to Data Science
Course Overview

Lesson 1

Derek Ruths

Lesson overview

Objectives

- Welcome
- Become familiar with course format & structure

Outline

- What is this course about?
- Lecture & assignment structure
- Grading
- Resources
- About me…

What is this course about?

This class is all about “good data science” –

- Collect, curate, and analyze data as a means of understanding phenomena in the world
- Combine knowledge about the problem and technical expertise & experience

By the end of class, students will

- Be able to design and execute a “good data science project”
- Have done a “good data science project”

Lecture structure

Course organized into *units*

- Each unit lasts approximately 1 week

We have two lectures a week: Monday & Wednesday

I don’t know if they’ll be reliably recorded…

(just assume they won’t be)

Assessments

- One assignment each week
  - Assigned on Monday, due the following Monday
- 4 in-class exams (make-up exam the following Monday from 11 AM – 12:30 PM)
- One team project
  - Begins 2/3rds of the way through the semester
  - Execute a full, guided data science project
  - Random team assignments (respecting grad/undergrad and credit/pass-fail)
  - Due last day of classes

Grading

15% completion credit for assignments

65% in-class exams

20% final project report

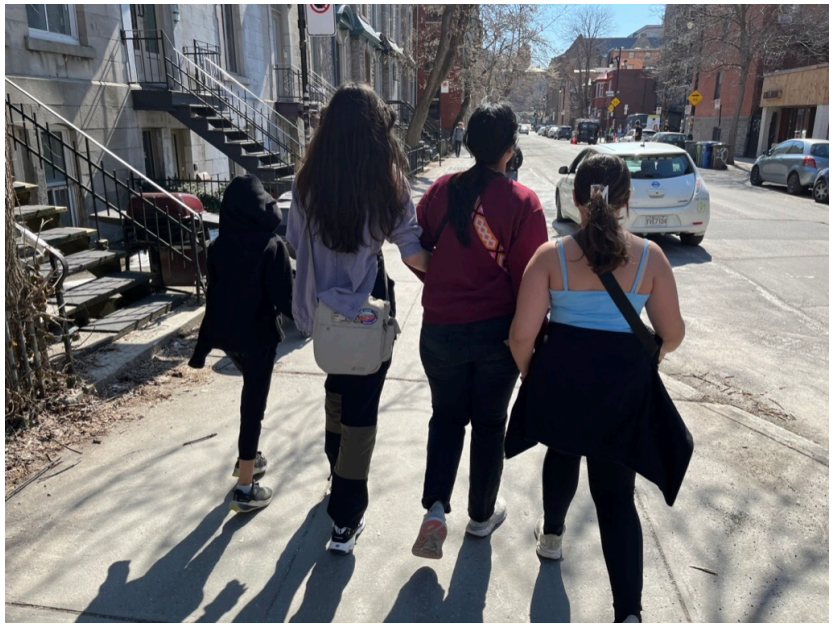
Resources

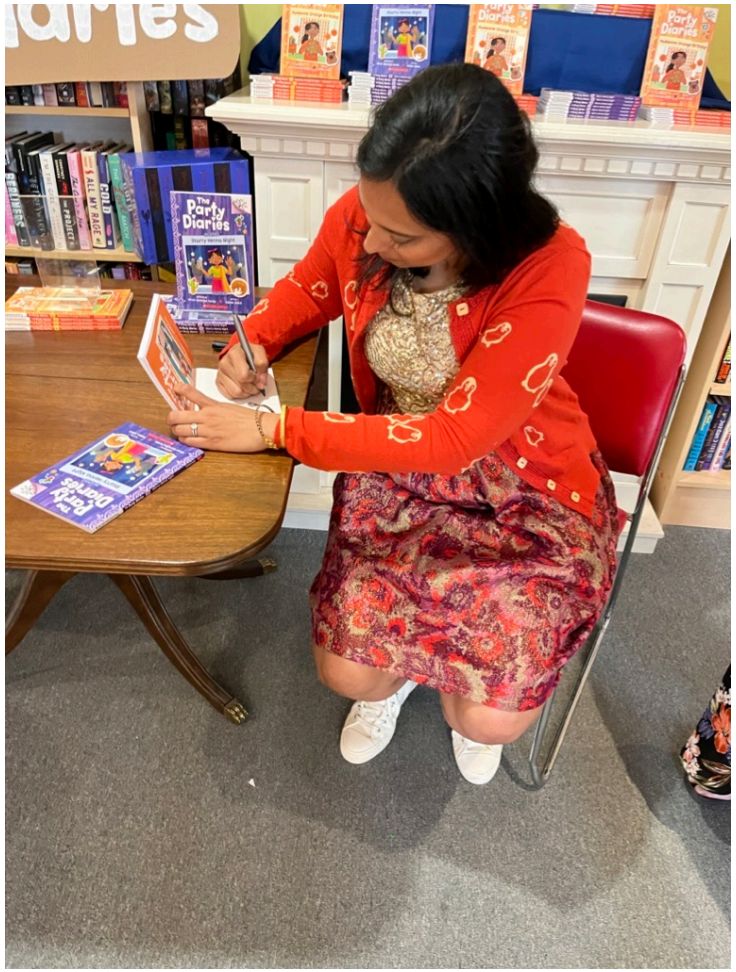
No textbook for the class

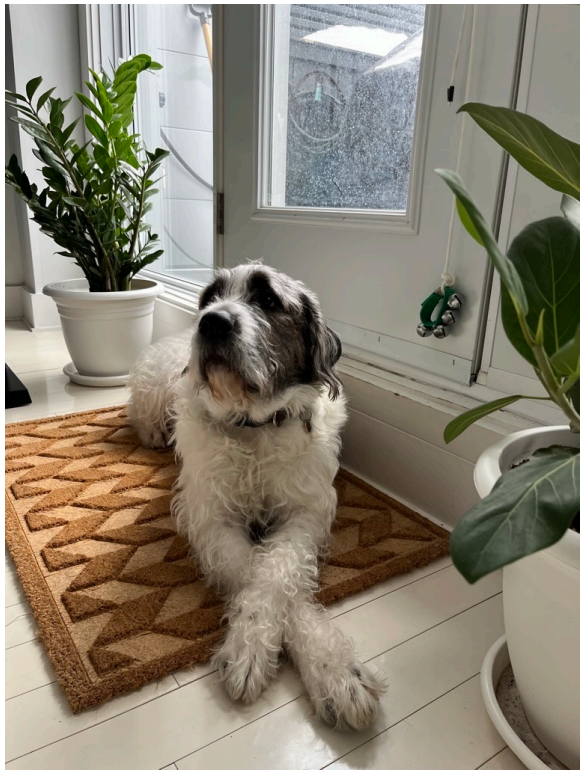
You will need AWS t3.small EC2 instance (or equivalent)

- This will cost ~$120 for the semester

About me







About me















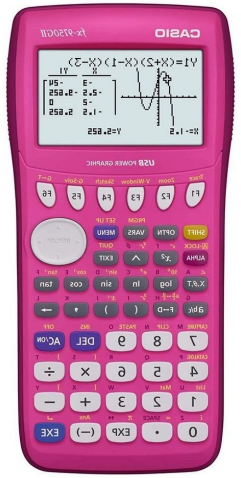


Getting good at data science is about experience

- Will this question produce a useful answer?
- What obstacles will I encounter with this data?
- What's the best way to organize this data?
- How much compute do I need?
- Should this be going faster? Should I spend time speeding it up?
- Have I answered the original question?
- What results can I show that prove it works?
- How do I deliver my solution?

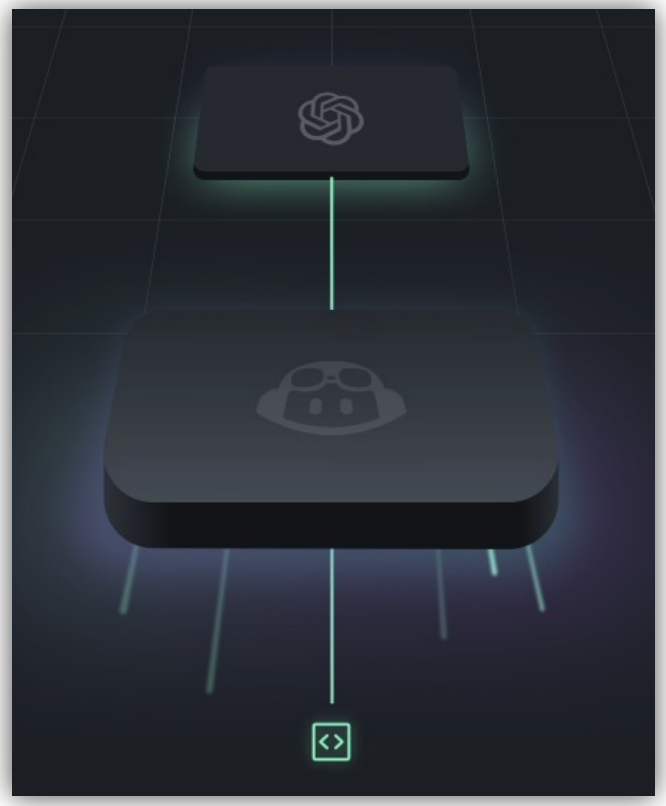
Do. the. homework.

Some thoughts on ChatGPT (LLMs)



The paradox:

- ChatGPT **can** solve many homework-scale problems
- ChatGPT **can't** solve many real-world-scale problems (by itself)



Avoid leaning on it while learning.

Rule of thumb: if you couldn't produce what it's writing, study it, turn it off, delete it, then write it yourself.

ChatGPT & this class

I'm going to do my best to prepare you to be a data scientist in a ChatGPT-powered world.

Observations:

- LLMs don't help (much) with problem/solution design
- LLMs make a lot of mistakes when writing code

Exams will focus on what you'll definitely have to do in real-world situations:

- Design
- Debugging

Do. the. homework.

Lesson wrap-up

Takeaways

- This course is going to be awesome
- Invest time and effort in the course - it’s worth it!

Up next

- Course topics
- How to get support