Syllabus for CIS 5450

Logistics

Instructors: **Harry Smith** (http://www.ourpropeller.com/), **Ryan Marcus** (https://rmarcus.info)

Office hours: **Schedule** (https://docs.google.com/spreadsheets/d/1hIq4XzgC3KKulB9rmVq0e4YjvhShby8NLJo658PqYTE/edit?usp=sharing) (subject to change) and **OHQ** (https://ohq.io/courses/925)

Classroom: **Meyerson** (https://facilities.upenn.edu/maps/locations/meyerson-hall) B1. Lectures will be recorded and posted a few days after class, but in-person attendance is highly encouraged.

- Lectures will be Tuesday and Thursday, 1:45pm - 3:15pm in Meyerson B1.
- Recitations will be Fridays, at either 10:45am or noon, in Towne 100. Recitation will also be recorded and posted through Canvas. Recitation will bridge the lecture and homework content.
- The midterm will take place on March 3rd (**physical attendance required**).
  - The midterm will be a mix of multiple choice and free response questions, written "worksheet style" on an exam we provide.
- There will be a final exam (date TBD), during university final period, **physical attendance required**).
  - The final exam will be free response questions only, written in a "blue book."
  - You can see the university policies on rescheduling a final exam here: **https://srfs.upenn.edu/registration-catalog-calendar/final-exams** (https://srfs.upenn.edu/registration-catalog-calendar/final-exams)
- April 28th, the last day of class, will feature an in-class activity and a fun (in my opinion) game (**physical attendance required**).
- Homework assignments will be released via announcements on Ed Discussion (see link in the left bar).
- Grades will be posted through GradeScope.
- If you have questions, feel free to search and post on Ed Discussion to get help from TAs and your fellow classmates.

Course Description

In the era of big data, we are increasingly faced with the challenges of converting massive amounts of data to actionable *knowledge*. Given the limits of individual machines (compute power, memory, bandwidth), increasingly the solution is to clean, integrate, and process the data using

statistical machine learning techniques, in parallel on many machines. This course focuses on the fundamentals of scaling computation to handle common data analytics tasks. You will learn about basic tasks in collecting, wrangling, and structuring data; programming models for performing certain kinds of computation in a scalable way across many compute nodes; common approaches to converting algorithms to such programming models; standard toolkits for data analysis consisting of a wide variety of primitives; and popular distributed frameworks for analytics tasks such as filtering, graph analysis, clustering, and classification.

Prerequisites

This course expects broad familiarity with probability and statistics, as well as programming in Python. CIS 1100, MCIT 5900, or the equivalent are *required*. Additional background in statistics, data analysis (e.g., in Matlab or R) is *helpful* but not required.

Grading

The grade breakdown will be as follows:

- homeworks (5-6 expected) 40%,
- term project 30%,
- midterm 15%,
- final exam 15%,

Late Policy

Each homework assignment in this course is due 2 weeks after the assignment is released. Homework assignments turned more than two weeks past their release date (even by one second) will receive a 10 percentage point penalty. For example, if an assignment submitted late would have gotten a score of 95%, the late assignment will instead get a score of 85%. An assignment that would have gotten a score of 100%, but is submitted late will get a score of 90%. Homework assignments can be turned in late until April 29th at 10pm. Late homework assignments submitted after April 29th at 10pm will get a score of 0.

Resubmission: you can submit your work for grading after the due date multiple times, but your resubmitted work will get the 10 percentage point penalty. For example, if you submit your work on time and get a grade of 70%, you could fix your homework (correct all errors) and resubmit late for a score of 90%. Resubmission can never raise your grade over 90%.

Collaboration Policy

You are responsible for knowing **Penn's Code of Academic Integrity (https://catalog.upenn.edu/**

pennbook/code-of-academic-integrity) . In particular, copying solutions from other students or other resources (e.g. the Web or from students who have taken the class in previous years) is NOT allowed. While you can verbally discuss high level ideas and discuss concepts, you are NOT allowed to share code with each other. Needless to say, making answers to homework assignments or exams available to others either directly or by posting on the web is NOT allowed.

We will not have a sense of humor about violations of this policy!

AI/Large Language Model (LLM) Policy

Modern AI tools can be of great help in understanding concepts, and we have no concerns about you using ChatGPT, Gemini, etc. to get alternative explanations for topics or syntax. Such tools are a great addition to a data scientist's toolbox! You should feel free to use LLMs as much as you'd like. Keep in mind that copying and pasting a solution from ChatGPT might score you some points on the homework, but **you'll be doing the midterm and final exam with pen and paper.** Therefore, we highly recommend using the homework assignments as an opportunity to practice your skills.

Resources:

- **Colab** (http://colab.research.google.com/) (for homeworks; you'll need a Google@SEAS or personal GMail ID)
- Ed Discussion, see link in sidebar (questions, discussion)
- Canvas, you are here (for access to the lecture recordings, which will be linked below)
- Gradescope, see link in sidebar (for homework submission and exams; you'll be **auto-added to this via Canvas**)

Readings:

We recommend several books for students of different skill levels..

For students with a limited CS background: You should get the book ***Data Science from Scratch*** (https://www.google.com/url?q=https%3A%2F%2Fwww.oreilly.com%2Flibrary%2Fview%2Fdata-science-from%2F9781492041122%2F&sa=D&sntz=1&usg=AOvVaw25djjmSGt5qQeKprNSCzOE) , (https://www.google.com/url?q=https%3A%2F%2Fwww.oreilly.com%2Flibrary%2Fview%2Fdata-science-from%2F9781492041122%2F&sa=D&sntz=1&usg=AOvVaw25djjmSGt5qQeKprNSCzOE) 2nd ed, by Grus, from O'Reilly. This book provides a quick refresher in Python, probability, statistics, and linear algebra. An online version can be accessed through the **Penn libraries** (https://www.google.com/url?q=https%3A%2F%2Fwww.library.upenn.edu%2F&sa=D&sntz=1&usg=AOvVaw337X8kb2yDyUPh5Qq5

qZwO) . You may also find the UC Berkeley free book **The Foundations of Data Science** (https://www.google.com/url?q=https%3A%2F%2Fwww.inferentialthinking.com%2Fchapters%2Fintro&sa=D&sntz=1&usg=AOvVaw3mKk0r9jrWLBa9svAg1ko4) useful.

For all students: *Python for Data Analysis* (https://www.google.com/url?q=https%3A%2F%2Fwesmckinney.com%2Fpages%2Fbook.html&sa=D&sntz=1&usg=AOvVaw2istGsI3y8YFvpBc9QVwrc) **, by McKinney, from O'Reilly. Again, an online version is accessible via the Penn libraries** (https://www.google.com/url?q=https%3A%2F%2Fwww.library.upenn.edu%2F&sa=D&sntz=1&usg=AOvVaw337X8kb2yDyUPh5Qq5qZwO) **.**

For advanced students: ***Python Machine Learning*** (https://www.google.com/url?q=https%3A%2F%2Fsebastianraschka.com%2Fbooks.html&sa=D&sntz=1&usg=AOvVaw1Gg_uanN_pmZK4rxphi7AR) , 3rd edition by Raschka, from Packt. And indeed, an online version of this book is also accessible via the **Penn libraries** (https://www.google.com/url?q=https%3A%2F%2Fwww.library.upenn.edu%2F&sa=D&sntz=1&usg=AOvVaw337X8kb2yDyUPh5Qq5qZwO) .