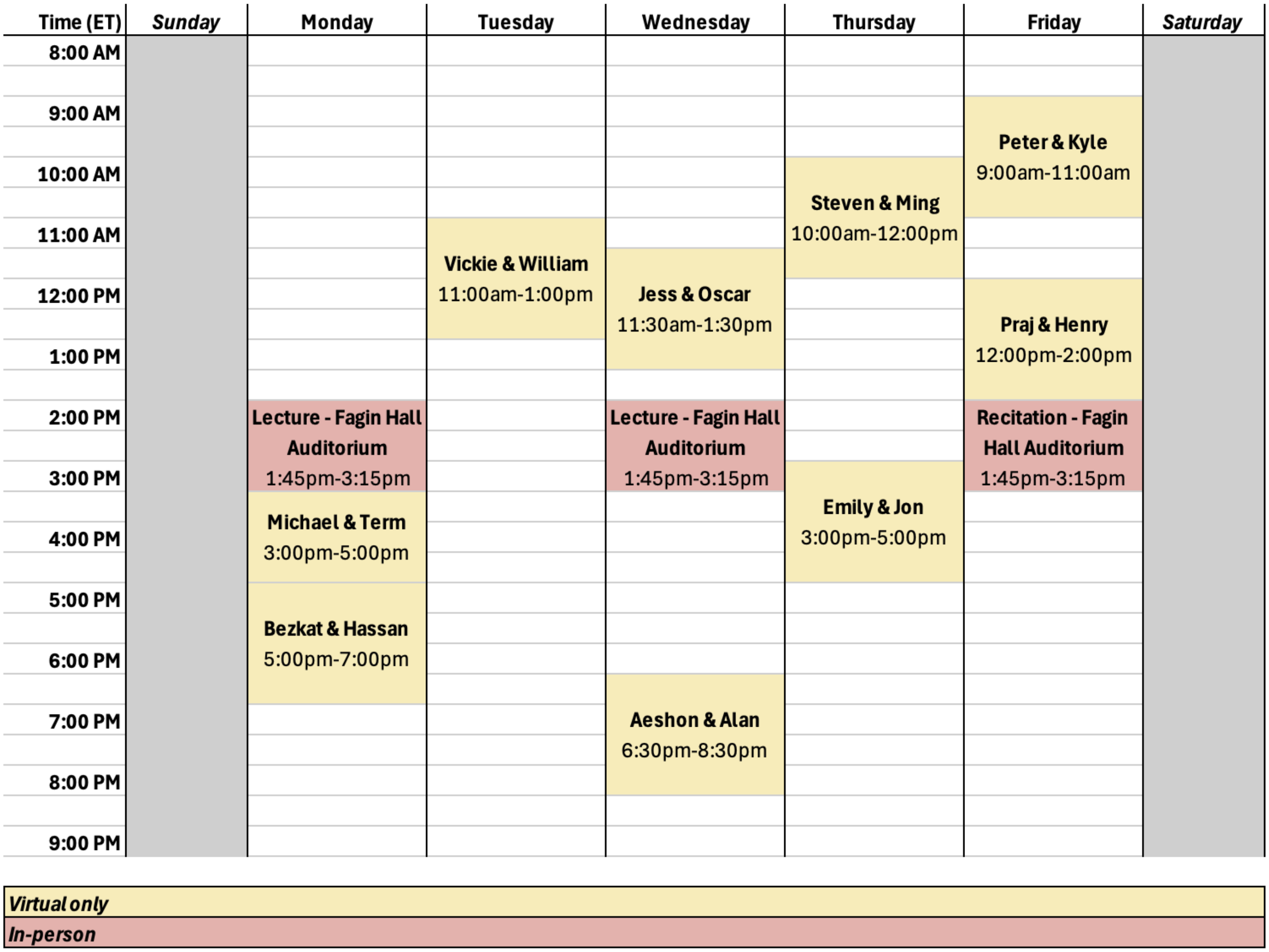
Logistics

Instructors: **[Ryan Marcus (https://rmarcus.info)](https://rmarcus.info)**

Zoom link:https://upenn.zoom.us/j/93428993003

Your fantastic TAs and their office hours:



Time (ET)	*Sunday*	Monday	Tuesday	Wednesday	Thursday	Friday	*Saturday*
8:00 AM							
9:00 AM						Peter & Kyle 9:00am-11:00am	
10:00 AM					Steven & Ming 10:00am-12:00pm		
11:00 AM			Vickie & William 11:00am-1:00pm				
				Jess & Oscar 11:30am-1:30pm			
12:00 PM						Praj & Henry 12:00pm-2:00pm	
1:00 PM							
2:00 PM		Lecture - Fagin Hall Auditorium 1:45pm-3:15pm		Lecture - Fagin Hall Auditorium 1:45pm-3:15pm		Recitation - Fagin Hall Auditorium 1:45pm-3:15pm	
3:00 PM					Emily & Jon 3:00pm-5:00pm		
		Michael & Term 3:00pm-5:00pm					
4:00 PM							
5:00 PM		Bezkat & Hassan 5:00pm-7:00pm					
6:00 PM							
				Aeshon & Alan 6:30pm-8:30pm			
7:00 PM							
8:00 PM							
9:00 PM							

Virtual only

In-person

Classroom: **[Fagin Hall Main Auditorium (https://facilities.upenn.edu/maps/locations/fagin-hall-claire-m)](https://facilities.upenn.edu/maps/locations/fagin-hall-claire-m)** (Ann L. Roy Auditorium). Lectures will be recorded and posted, but in-person attendance is highly encouraged.

- Lectures will be Monday and Wednesday, 1:45pm - 3:15pm in Fagin Hall. Prerecorded video lectures will also be made available, linked through Canvas.
- Recitations will be Fridays, from 1:45pm - 3:15pm in Fagin Hall (same as the lectures). Recitation will also be recorded and posted through Canvas. Recitation will primarily provide guidance on homework assignments.

- The midterm will take place on October 16th (**physical attendance required**). There will be a final exam during university final period (date TBD, **physical attendance required**).
  - The miderm will be a mix of multiple choice and free response questions, written "worksheet style" on an exam we provide.
  - The final exam will be free response questions only, written in a "blue book."
- December 9th, 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). Unless noted otherwise, a new homework assignment will be released when the previous homework assignment is due.
- 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.

(Previous iterations of the course: **Fall 2023 (https://sites.google.com/seas.upenn.edu/cis545/home)** , **Spring 2023 (https://sites.google.com/seas.upenn.edu/cis545-sp23)** , **Fall 2022 (https://sites.google.com/seas.upenn.edu/cis545-22f)** , **Spring 2022 (https://sites.google.com/seas.upenn.edu/cis545-sp22)**)

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 110, MCIT 590, 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 25%,
- midterm 15%,
- final exam 15%,
- participation (participating in person, posting to Ed) 5%

Late Policy

Homework assignments turned in late (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 Dec 9th at 10pm. Late homework assignments submitted after Dec 9th 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 perfect 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 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**)
- **OHQ** (https://ohq.io/courses/802) (for less frustrating office hour queuing)

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=AOvVaw337X8kb2yDyUPh5Qq5qZwO) . 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) .