Logistics

Instructors: Jacob Gardner (https://jacobrgardner.github.io/) and Ryan Marcus (https://rmarcus.info)

Your fantastic TAs:

S24 Big Data Contact Info : Contact Info

Name	Email	Office Hours
Instructors		
Jake Gardner	jacobrg@seas.upenn.edu	TBD
Ryan Marcus	rcmarcus@seas.upenn.edu	TBD
TAs		
(Head TA) Arnav Jhaveri	arnavjha@seas.upenn.edu	Wednesdays 3:15 PM – 5:15 F
(Head TA) Emily Liu	liuemily@wharton.upenn.edu	Thursdays 11:45 AM – 1:45 PM
(Head TA) Jeffrey Li	lijeff@seas.upenn.edu	Wednesdays 3:15 PM – 5:15 F
Aashvi Manakiwala	aashvi@seas.upenn.edu	Saturdays 12:00 PM – 2:00 PM
Akanksha Tripathy	atripath@seas.upenn.edu	Thursdays 11:45 AM – 1:45 PM
Faaiz Quaisar	faaizq@sas.upenn.edu	Tuesdays 11:30 AM – 1:30 PM
Federico Cimini	fcimini@seas.upenn.edu	Mondays 10:00 AM – 12:00 PM
Karan Sampath	ksampath@seas.upenn.edu	Wednesdays 5:15 PM – 7:15 F
Kyle Liao	kyleliao@sas.upenn.edu	Wednesdays 9:30 AM – 11:30
Liang-Yun Cheng	lycheng@seas.upenn.edu	Mondays 10:00 AM – 12:00 PM
Matthew Sonder	msonder@sas.upenn.edu	Wednesdays 9:30 AM – 11:30

Contact Info

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

- Lectures will be Tuesday and Thursday, 1:45pm - 3:15pm in Meyerson Hall B1. Prerecorded video lectures will also be made available, linked through the syllabus.
- Recitations will be Fridays, at either 10:15-11:45am or noon-1:30p in TOWNE 100.
- Materials will be posted weekly in Ed Discussion. Recitations are optional, but highly recommended.

(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 20%,
- midterm 15%,
- final exam (2nd midterm) 15%,
- **quizzes (1 week deadline from lecture!) 7%,**
- participation (participating in person, posting to Ed) combine to make 3%.

Late Days

Students can submit homework assignments (including HW0 but not the term project), up to 48 hours late cumulatively, with no penalty. In other words, split across all homeworks, you have 48 late hours. If you exceed those 48 hours, a penalty of -1% per hour will be applied to that HW. Please note that we will be rounding up (e.g. Gradescope will count a submission that is even 1 minute late as using up 1 late hour)

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, Bard, etc. to get alternative explanations for topics. However -- given that we are trying to teach general, reusable skills -- we expect you to write your code without help from an LLM or from a classmate. Please note that the exams will be tailored with this in mind (not focused on syntactic details, but on the ability to tackle problems) so you should make sure you can solve problems on your own!

Readings and 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://www.google.com/url?q=https%3A%2F%2Fohq.io&sa=D&sntz=1&usg=AOvVaw0PZMnniUn4liLNPwRJmpzE) (for less frustrating office hour queuing)

Readings:

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

For students who do not have at least 2 years of a CS degree: 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=AOvVaw2istGsl3y8YFvpBc9QVwrc) **, 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) .