

Syllabus for CIS 5450

Logistics

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

Office hours: [OH Schedule](https://docs.google.com/spreadsheets/d/17_BJtStASqszRioWmFbQkWEUFTkTe1tv/edit?usp=sharing&ouid=111795220873593385887&rtpof=true&sd=true)  [_ \(https://docs.google.com/spreadsheets/d/17_BJtStASqszRioWmFbQkWEUFTkTe1tv/edit?usp=sharing&ouid=111795220873593385887&rtpof=true&sd=true\)](https://docs.google.com/spreadsheets/d/17_BJtStASqszRioWmFbQkWEUFTkTe1tv/edit?usp=sharing&ouid=111795220873593385887&rtpof=true&sd=true)  [_ \(https://ohq.io/courses/925\)](https://ohq.io/courses/925)

Classroom: [Meyerson](https://facilities.upenn.edu/maps/locations/meyerson-hall) [_ \(https://facilities.upenn.edu/maps/locations/meyerson-hall\)](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 Monday and Wednesday, 1:45pm - 3:15pm in Meyerson B1.
- Recitations will be Fridays, at either 10:15am or 1:45pm, in Towne 100. Recitation content will vary week to week, and will sometimes be used for oral exams.
- The midterm will take place on Oct 7th (**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 "blue book" style.
 - You can see the university policies on rescheduling a final exam here: <https://srfs.upenn.edu/registration-catalog-calendar/final-exams> [_ \(https://srfs.upenn.edu/registration-catalog-calendar/final-exams\)](https://srfs.upenn.edu/registration-catalog-calendar/final-exams)
- Dec 7th, 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) 20%,
- oral exams (2-3 expected) 10%,
- term project 30%,
- midterm 20%,
- final exam 20%,

Cutoffs for letter grades will be no higher than: A (95%), A- (90%), B+ (87%), B (83%), B- (80%), C+ (77%), C (73%), C- (70%), D+ (67%), D (63%), D- (60%). We may curve the class up, but we will not curve the class down. These cutoffs are from [David Ludden \(https://www.sas.upenn.edu/~dludden/GRADING%20CALCULATIONS.htm\)](https://www.sas.upenn.edu/~dludden/GRADING%20CALCULATIONS.htm).

Homework grades are determined by the score shown on GradeScope after you submit.

Homework Late Policy

Each homework assignment in this course is due 1 or 2 weeks after the assignment is released, at 10pm. Homework assignments turned in past their due 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 Dec 7th at 10pm. Late homework assignments submitted after Dec 7th 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%. Students who are found to have violated Penn's code of academic integrity are not eligible to resubmit homework assignments.

Oral exams

This course will require oral examinations. The purpose of the oral exams is to assess your ability to reason through and explain course concepts, rather than simply reproduce memorized answers. Near the start of the semester, you will be randomly assigned 2-3 oral exam topics. You will sign up ahead of time for a scheduled slot.

Oral exam topics (you will be randomly assigned 2 or 3 of these):

- Regular expressions (between 9/2 and 9/21)
- Relational schema (between 9/21 and 10/7)
- Hypothesis testing (between 10/7 and 10/26)
- Ensemble models and model evaluation (between 10/26 and 11/18)
- Distributed computing: scaling, faults, and data distribution (between 11/18 and 12/2)

Each oral exam will be about 15 minutes. The oral exams are designed to follow a technical interview setup: the examiner will ask you a question, and you will then walk the examiner through the process of developing a solution. Oral exams will be held in person, done on either a whiteboard or scratch paper. Students may not use any electronic devices. The examiner will record the oral exam on their laptop in order to help Ryan adjudicate grade disputes -- you may opt out of this recording, but note that an oral exam grade can only be corrected if we have a recording. All recordings will be deleted after course grades are finalized.

You will be graded on the correctness of your solution, your ability to communicate that solution and any assumptions it makes to the examiner, and your responses to the examiner's follow up questions. Failing to appear for your scheduled oral exam will result in a grade of 0%.

Collaboration Policy

You are responsible for knowing [Penn's Code of Academic Integrity \(https://catalog.upenn.edu/pennbook/code-of-academic-integrity\)](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.

The use of generative AI tools does not change these rules. If two submissions are substantially

similar, "we both got the same answer from an LLM" is not an acceptable explanation: you are responsible for everything you submit, including any text or code produced with AI assistance. Submitting AI-generated work that is substantially similar to another student's submission, or to publicly available material, will be treated the same way as any other impermissible copying or collaboration.

When we suspect that a violation of Penn's code has occurred, we submit a case to [Penn CSA \(https://csa.upenn.edu/\)](https://csa.upenn.edu/). The CSA's verdict will be final, and honored by the instructor. If the CSA finds that a violation occurred, the instructor's discretion will determine a resolution: students will *at least* receive a 0 on that assignment, and will *at most* be given a failing grade in the class.

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, or for writing example code. 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/)  (<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>),  (<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>)

[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 \(https://find.library.upenn.edu/catalog/9977533143503681?hld_id=53589066410003681\)](#). 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 \(https://find.library.upenn.edu/catalog/9977534934703681?hld_id=53589640820003681\)](#).

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 ↗\(https://www.google.com/url?q=https%3A%2F%2Fwww.library.upenn.edu%2F&sa=D&sntz=1&usg=AOvVaw337X8kb2yDyUPh5Qq5qZwO\)](#).

Accommodation Procedures

The University of Pennsylvania provides reasonable accommodations to students with disabilities who have self-identified and received approval from Disability Services. If Disability Services has approved your request for accommodations, please make an appointment to meet with me as soon as possible in order to discuss the arrangements for your accommodations.

If you have not yet contacted Disability Services and would like to request accommodations or have questions, you can make an appointment by calling (215) 573-9235. The office is located in the Weingarten Center at Stouffer Commons, 3702 Spruce Street, Suite 300. Please visit the Disability Services website at <https://weingartencenter.universitylife.upenn.edu/> (<https://weingartencenter.universitylife.upenn.edu/>).

Disability Services are free and confidential.