STAT 4710/5710: Modern Data Mining
Fall 2025

Lectures, Section 401: Tue & Thu every week, 3:30-5:00pm, F85 JMHH
Lectures, Section 402: Tue & Thu every week, 5:15-6:45pm, F85 JMHH

Teaching Team & Communication

Instructor: Ying Jin

Email: yjinstat@wharton.upenn.edu
Office hour: Wed 4-5:30pm, 411 ARB

PhD head TA: Fred Zhang

Email: fwzhang@wharton.upenn.edu
Office hour: Thu 1:45-3:15pm, F92 JMHH

Undergraduate TA

Nhat-Ha Pham, nhathapt@sas.upenn.edu
James Ward, jrward@wharton.upenn.edu

(ARB = Academic Research Building, JMHH = Jon M. Huntsman Hall)

The instructor and teaching assistant will hold office hours every week. Outside of office hours, students can ask the teaching staff questions on the Ed Discussion platform. Students should only email the instructor in exceptional circumstances.

Course Description

With the advent of the internet age, data are being collected at unprecedented scale in almost all realms of life, including business, science, politics, and healthcare. Data mining—the automated extraction of actionable insights from data—has revolutionized each of these realms in the 21st century. The objective of the course is to teach students the core data mining skills of exploratory data analysis, selecting an appropriate statistical methodology, applying the methodology to the data, and interpreting the results. The course will cover a variety of data mining methods including linear and logistic regression, penalized regression, tree-based methods, and deep learning. Students will learn the conceptual basis of these methods as well as how to apply them to real data using the programming language R.

Prerequisites

Students are required to have taken two semesters of statistics courses, including the equivalent of STAT 4310. In particular, students must be comfortable with multiple linear regression. Students are also required to have programming experience in at least one language (R is preferred, but experience in another language is sufficient). Students are expected to have access to a computer on which they have installed the R Studio IDE, and any necessary support packages.

Course Outline

The course is tentatively structured into five units.

Unit 1: R for Data Mining (2 weeks)

- Course introduction, data visualization, data pre-processing

Unit 2: Prediction fundamentals (2 weeks)
- Model evaluation, complexity, bias-variance trade-off, cross-validation, classification

Unit 3: Regression-based methods (3.5 weeks)
- Logistic regression, regression in high dimensions, ridge regression, lasso regression

Unit 4: Tree-based methods (2.5 weeks)
- Growing decision trees, tree pruning, bagging and random forests, boosting, feature importance

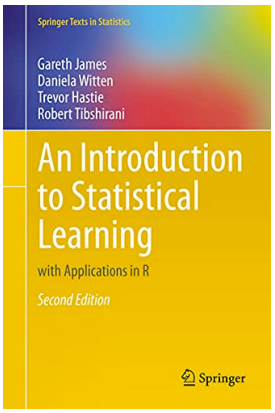
Unit 5: Deep learning (3 weeks)
- Neural networks, optimization, deep learning for image and text processing, AI frontiers

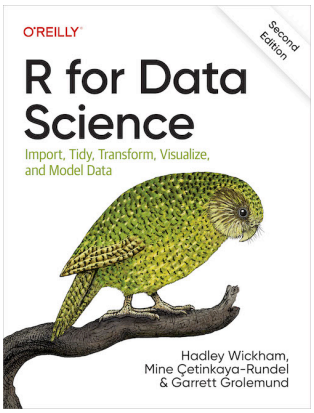
Course Materials

Please download R and Rstudio (we need both) prior to the first class. The R statistical software program, which is available here. RStudio is an Integrated Development Environment (IDE) for R, available here. Coding sessions in class will be in R, yet we allow students to use Python for homework if they prefer.

Our primary textbook (required) is
- Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. *An Introduction to Statistical Learning.* Second edition. 2021.

This textbook is available for purchase at the Penn Bookstore and freely available online.





We will use following textbook for R programming:
- Hadley Wickham, Mine Çetinkaya-Rundel, and Garrett Grolemund. *R for Data Science.* Second Edition. 2023.

This textbook is freely available online.

Assignments and assessments

	Homework	Quizzes	Exams
Format	Take home	In class	In class
Submission	Electronic	Paper	Paper
Duration	1-2 weeks	30 minutes	90 minutes
Number of assessments	5 (one per unit)	4 (except unit 1)	2
Percentage of grade	40% = 5 x 8%	20% = 4 x 5%	40% = 2 x 20%
Collaboration allowed*	✓	✗	✗
Course materials allowed	✓	✗	✗
Internet & AI tools allowed*	✓	✗	✗

*See course policies below for details.

Homework

There will be 5 homework assignments (roughly 1 per unit). These homeworks will be prescriptive and involve developing Data Mining (or ML) models. Homeworks will be uploaded in the "Homework" Folder on Canvas. They will tentatively be assigned and be due **11:59am ET** on the following dates:

- Hw 1: Assigned on 8/28, due on 9/16
- Hw 2: Assigned on 9/11, due on 9/30
- Hw 3: Assigned on 10/2, due on 10/23
- Hw 4: Assigned on 10/28, due on 11/11
- Hw 5: Assigned on 11/13, due on 12/8

Submission. Homeworks are submitted by students as PDFs (containing both code chunks and required outputs) and graded through Gradescope.

Late submission. Homeworks are penalized by 25% up to 1 day late. Homeworks more than 1 day late will receive a 0 (since solutions will have been posted on Canvas by then). Submit early to avoid last-minute technical failures.

Regrading. Regrade requests for these assignments can also be made through Gradescope. Please be aware of the following regrade rules:

- Regrade requests will be considered only when there is a clear discrepancy between the rubric and the grade. Regrade requests for homework will be considered only if submitted within a week of the date the grade was posted. Regrade requests for exams will be considered only if submitted within 3 days of the date the grade was posted.
- Any request will result in regrading the whole assignment.
- Quizzes are automatically graded; any disputes of the intended answers should be submitted via private message on Ed Discussion.

Collaborations and AI for homework. Students are permitted to work together on homework assignments, but must write up and submit solutions individually. In particular, students may not copy each others' solutions. Students may consult all course materials, textbooks, the internet, or AI tools (e.g. ChatGPT or GitHub Copilot) to complete their homework. Students may not use solutions to problems that may be available online and/or from past iterations of the course. For each homework, students must disclose all classmates with whom they collaborated, which AI tools they used, and how they used them. Failure to do so will result in a 10-point penalty. Plagiarized code will receive a 0 for all parties.

Quizzes and exams

Quiz dates. There will be 4 in-class quizzes (one per unit except unit 1), 30 mins each:

- Quiz 1: Tue, 9/23
- Quiz 2: Tue, 10/14
- Quiz 3: Tue, 11/6
- Quiz 4: Tue, 11/25

Exam dates. In addition, there will be 2 in-class exams, 90 mins each.

- Exam 1: Thu, 10/16
- Exam 2: Thu, 12/4

Exam 2 will focus on post-Exam-1 materials.

IMPORTANT: Please consider your continued enrollment in this course based on your ability to not miss any of these quizzes or exams.

Materials. Students may not consult any materials for quizzes and exams except for a calculator and both sides of one sheet of 8.5x11-inch paper with 1-inch margins and the equivalent of 10-point font.

Makeup and absence. No makeup quizzes and exams will be offered. However, each student's lowest quiz grade will be dropped.

Missing an exam. In exceptional circumstances* that a student must miss an exam, the student must explain the absence with a letter from an academic advisor or departmental representative at least one week in advance (except in emergencies). At the instructor's discretion, the grade for the missed exam may be replaced by the average of the student's other exam grade and average quiz grade.

*Exceptional circumstances include serious illness, family emergencies, or university-sanctioned events (e.g., varsity athletics, academic conferences). Vacations, travel issues, or oversleeping are not acceptable reasons.

Letter grades

An overall numeric grade will be computed for each student at the end of the semester by weighting the homework, quizzes, and exams according to the above percentages. **Final letter grades will then be assigned based on numeric grade thresholds; a tentative grading scheme is below.**

- A+/A/A-: $90 \leq \text{score} \leq 100$
- B+/B/B-: $80 \leq \text{score} < 90$
- C+/C/C-: $70 \leq \text{score} < 80$
- D: $60 \leq \text{score} < 70$

These thresholds represent the **lowest numeric grade (not rounded)** that will guarantee the corresponding letter grade; thresholds may be lowered at the instructor's discretion (e.g., to account for overall class performance). For example, a final numeric grade of 89.0 guarantees at least a B+, while a 90.0 guarantees at least an A-.

By the grade type change deadline (Nov 3), students will have grades for two homework assignments, two quizzes, and the first exam.

Extra credit

Extra credit will be awarded at the discretion of the instructor for **participation in class and on Ed Discussion**. On Ed Discussion, answering questions will be weighted more heavily than asking questions, with greatest weight given to instructor-endorsed answers. Apart from this, there are no other extra credit opportunities beyond **possible extra credit questions** on the homeworks, quizzes, and/or exams.

Other Policies & Information

Key dates

- Tuesday, August 26: First day of class
- Thursday, October 9: Fall break, no class
- Monday, October 6: Drop period ends
- Friday, October 24: Grade mode change deadline
- Monday, November 3: Last day to withdraw from a course
- Thursday, November 27: Thanksgiving break, no class
- Thursday, December 4: Last lecture

Classroom expectations

Attendance. Attendance is not required, however missing classes will likely affect learning, and missing in-class quizzes will affect the final grade. Students are encouraged to ask questions in class.

Electronics. It is recommended that students bring their laptops to class for programming exercises. Outside of this context, students are encouraged not to have their laptops out. Electronic devices should be silenced and are not to be used for activities that distract fellow students or the instructor. No electronic devices, except calculators, are allowed in quizzes.

Class materials. Class recordings will be provided on Canvas and all in-class lecture materials will also be posted to the course webpage.

Academic integrity

In accordance with Penn's Code of Academic Integrity, students must comply with the course collaboration policies described in this syllabus and in the assignment instructions. **All suspected academic integrity violations will be reported to the Office of Student Conduct and all assignments where violations occurred will receive grades of zero.** If you have any questions about collaboration policies, please do not hesitate to contact the instructor.

Accessibility for students with disabilities

The instructor is committed to creating a learning experience that is as accessible as possible. Students with disabilities should reach out to the Office of Student Disabilities Services (SDS) by calling 215-573-9235 (services are confidential) and email the instructor. The instructor will then work with the student and SDS to provide reasonable accommodations. For more on academic accommodations, please see the Weingarten Center.

Accommodation requests

Please submit all accommodation requests on the Ed Discussion platform under the "accommodations" category.