Stat 214, Spring 2026, UC Berkeley

Data Analysis and Machine Learning for Real-World Decision Making

Instructor: Professor Bin Yu, binyu@berkeley.edu

Lectures: T/Th: 11-12:30 (Social Science Building 20); Office hours: TBA (Evans 409)

Two discussion sections: Friday 1-2, 2-3 in Evans 334.

GSIs: Zach Rewolinski zachrewolinski@berkeley.edu, Anqi Wang aqwang@berkeley.edu, Sequoia Andrade srandrade@berkeley.edu, Sean Richardson seanrichardson@berkeley.edu; Office Hours TBA.

GSIs will be in charge of the discussion sessions, Ed Discussions on bcourses, and the labs/homework (reading summaries of and selected problems from the VDS book by Yu and Barter) .

Textbooks:

- "Veridical data science: the practice of responsible data analysis and decision-making" by Bin Yu and Rebecca Barter (MIT Press, in-press) (free online version at vdsbook.com) (required)
- Statistical models, David Freedman (Cambridge Press, 2009, 2nd Ed.) (required). (open-source pdf)
- The elements of statistical learning, Trevor Hastie, Rob Tibshirani, Jerome Friedman (Springer, 2016, 2nd Ed.) (recommended). (open-source pdf)

  Location of course materials: internet and bcourses

Prerequisites: stat 134 and stat 135 (or data 100 and data 140) or equivalents.

Computing prerequisites: stat 243 or equivalent.

This is an MA class in statistics. Students will be engaged in open-ended data projects for decision making to solve domain problems. It mirrors the entire data science life cycle in practice, including problem formulation, data cleaning, exploratory data analysis, statistical and machine learning modeling and computational techniques, and interpretation of results in context. It is guided by the Predictability-Computability-Stability (PCS) framework for veridical data science and emphasizes critical thinking and documenting human judgment calls and code. It coaches not only the technical but also communication and teamwork skills in order to obtain responsible and reliable data-driven conclusions for solving complex real world problems.

Grading:

- 45% lab assignments (three projects: first a single-person project (20%), and second a team project (25%))
- 10% reading assignments and selected problems from VDS book
- 2.5% class participation (lectures and discussions)
- 2.5% peer lab review performance
- 5% paper presentations
- 35% final project (team project)

Attendance policy: email notices to GSIs are required for missing lectures or discussion sessions. Attendances will be taken at lectures and discussion sessions. No exams.

Assignment descriptions: reading and selected problems from VDS book, and all three projects or data labs are based on using data sets with background domain information given (including possible guest lectures from domain experts) to arrive at data-driven conclusions or decisions. The projects or labs mimic data science practice and guide students through the whole data science life cycle (DSLC).

Assignment policies: no late lab reports in general, except under special circumstances.

Student conduct (academic integrity): class discussion on academic integrity and professional conduct in the beginning of the semester. Every lab report will be turned in with statements from students on their contributions and about whether and how they used AI tools such as chatGPT.

Comments, Suggestions, Gripes: Before or after the lectures, email, or talk to the instructor and the GSIs.

Ed Discussion: Questions and discussion about course material, assignments, and labs can be posted on the Ed Discussion page (accessed on bCourses). The GSIs will regularly monitor this to ensure all questions are answered in a timely manner, but students are encouraged to help their classmates as well. Please think carefully before asking questions specifically about the projects. For example, questions concerning how to do something specific in Python are fine, but questions asking what other people did for their analysis are not. Questions asking about clarifications are fine.

bCourses: https://bcourses.berkeley.edu/courses/1551682
Course website: https://stat214.berkeley.edu

Tentative lecture schedule

Date(s)	Topic	Notes
Week 1 (1/20, 1/22)	Overview (VDS, PCS, DSLC), Collaborative culture. Problem formulation, Data Cleaning, EDA	Lab 0 as a self-study exercise assigned on Friday 1/24
Week 2 (1/27, 1/29)	Guest lecture by Aaron Kornblith from UCSF PCA, Clustering	Lab 0 turned in (no grading); Lab 1 assigned (single person lab) (3 weeks)
Week 3 (2/3, 2/5)	Clustering. Basic deep learning. autoencoder.	
Week 4 (2/10, 2/12)	LS. Logistic regression. Guest lecture by Eileen Long from Nvidia	
Week 5 (2/17, 2/19)	Decision tree. Introduction of random variables and probabilistic generative modelsGLMs. Fitting GLMs through (IRWLS)	Lab 1 due, Lab 2 assigned with checkpoint schedule (team project, 4 weeks)
Week 6 (2/24, 2/26)	Class presentation on papers. Penalized GLMs including Lasso, Ridge and ElasticNet.	
Week 7 (3/3, 3/ 5)	PCS UQ and conformal inference. Boosting, SVM MA Panel?	

Week 8 (3/10, 3/12)	Class presentation on papers. Kernel ridge, supervised DL, CNN	
Week 9 (3/17, 3/19)	Transformer, generative LLM, Guest lecture from Google	Lab 2 due on 3/21 Lab 3 (final project) assigned (7 weeks, due May 9) sub-lab due dates?
Week 10 (3/24, 3/26)	SPRING BREAK	
Week 11 (3/31, 4/2)	Class presentation Importance index, Interpretable ML (LIME, SHAP, CD)	
Week 12 (4/7, 4/9)	Class presentation Diffusion models	Final project: first check-point
Week 13 (4/14, 4/16)	Neyman-Rubin ATE adjustments, Guest Lecture from Biotech?	
Week 14 (4/21, 4/23)	Cate, RF, iRF, RF+	Final project: second check-point
Week 15 (4/28, 4/30)	RL basics; Revisit PCS.	

Week 16 (5/8) **Final project due**

—--------

LAB SCHEDULE: we don’t have to send this out in the syllabus.

Date(s)	Topic	Notes
Week 1 (1/23)	GitHub Setup & Lab 0	Zach
Week 2 (1/30)	Lab #1 Introduction	Zach
Week 3 (2/6)	Data Visualizations + ACCESS Setup	Zach
Week 4 (2/13)	Check ACCESS Works, Dimension Reduction	Zach or ???
Week 5 (2/20)		Lab #1 Due, Lab #2 Assigned; Anqi
Week 6 (2/27)		Anqi
Week 7 (3/6)		Anqi
Week 8 (3/13)		Anqi
Week 9 (3/20)		Lab #2 Due, Lab #3 Assigned
Week 10 (3/27)	SPRING BREAK	
Week 11 (4/3)		
Week 12 (4/10)		
Week 13 (4/17)		
Week 14 (4/24)		
Week 15 (5/1)		