Draft Syllabus Subject to Change

Advanced Topics in Data Science (Spring 2024)

CS 109b, AC 209b, Stat 109b, or CSCI E-109b

Instructors

Pavlos Protopapas (SEAS) & Alex Young (Statistics)

Lectures: Mon & Wed 9:45-11am SEC 1.321 **Labs:** Fri 9:45-11:30am SEC 1.321 **Advanced Sections:** Select Wednesdays 2:15-3:30pm Room & Dates TBD **Office Hours:** TBD

Prerequisites: CS 109a, AC 209a, Stat 109a, or CSCI E-109a or the equivalent.

Course Description

Advanced Topics in Data Science (CS109b) is the second half of a one-year introduction to data science. Building upon the material in Introduction to Data Science (CS109a), the course introduces advanced methods for data wrangling, data visualization, statistical modeling, and prediction. Topics include generative models, Bayesian modeling, sampling methods, unsupervised learning, data augmentation, and multiple deep learning architectures such as CNNs, RNNs, autoencoders, language models, transformers, and GANs.

The programming language used will be Python.

Tentative Course Topics

- Unsupervised Learning, Clustering
- Bayesian Inference
- Hierarchical Bayesian Modeling
- Fully Connected Neural Networks
- Convolutional Neural Networks
- Autoencoders
- Recurrent Neural Networks
- NLP / Text Analysis
- Transformers
- Generative Adversarial Networks

Course Goals

Upon successful completion of this course, you should feel comfortable with the material mentioned above, and you will have gained experience working with others on real-world problems. The content knowledge, project, and teamwork experience will prepare you for the professional world or further studies.

Differences Between CS109B & AC209B

- **Advanced Sections**: 209B students are required to either attend the advanced sections in person or watch the recordings. 109B students are welcome at advanced sections, but this material is not required for them.
- **Assignments**: 209B students will have a short, additional HW component for some assignments. This will cover material from the advanced sections.

- **Exams:** The quiz questions and coding portions of the two midterms will differ slighly between 109B and 209B students. 209B midterms will also cover material from advanced sections.
- **Projects:** Project groups which include one or more 209B students will be required to make use of some approach or method not explicitly covered in lecture or advanced section. The goal is that 209B students perform some additional, outside reading to inform their projects.

Course Format

Lectures, labs, and advanced sections will be live-stream for Extension School students and can be accessed through the 'Zoom' section on Canvas.

Recordings will be made available to all registered students with 24 hours and can be accessed through the 'Course Videos' section on Canvas.

Lectures The class meets for lectures twice a week (Mon & Wed). Attending and participating in lectures is a crucial component of learning the material presented in this course.

Students may be asked to complete short readings before certain lectures.

Attendance is required for on campus students.

Labs Lab will be held every Friday and will present deep-dives into the software libraries used to implement the methods described in lecture, preparing students to successfully complete their homework assignments.

Labs cover core, non-optional course material. Like lecture, attendance is required for on campus students.

Advanced Sections The course will include advanced sections for students enrolled in AC209b. These 75 min sessions will cover advanced topics like the mathematical underpinnings of the methods seen in the main course lectures and lab as well as extensions of those methods. The material covered in the advanced sections is required for AC209b students, but all are welcome. Tentative topics are:

- Gaussian mixture models
- Particle filters/sequential Monte Carlo
- Solvers
- Segmentation techniques, YOLO, Unet and M-RCNN
- Word2Vec
- Diffusion models

Note: Advanced Section are not held every week. Consult the course calendar for exact dates.

Assignments and Grading:

Midterms The course will feature two midterms, each with two components:

1. An in-class quiz (mostly multiple choice).
2. A set of take-home coding challenges.

The in-class quiz for the first midterm will be held from 9:45-11 AM on Friday, March 22nd, during our regular lab hours. After this quiz concludes, the associated coding challenges will be released, with submissions due by 11:59 PM on Sunday, March 24th.

The second midterm's in-class quiz is scheduled from 9:45-11 AM on Wednesday, April 24th. Similarly, upon the quiz's conclusion, the coding challenges will be distributed. The deadline for these challenges will be 11:59 PM on Friday, April 26th.

Further details and any additional resources for the midterms will be provided as their dates approach.

Projects Students will work in groups of 3-5 to complete a final project. Students will be encouraged to propose project topics early in the semester, provided they can supply relevant, public data. Course staff will select a subset of these proposals to become project options for the class. The final submission will consist of a written report, a Jupyter notebook with all relevant code, and a 6-minute, pre-recorded presentation video.

Project groups with AC209b students will be held to a higher standard, and are expected to explore and incorporate methods that are not explicitly covered in class.

More information on projects and the proposal process to come.

Homework Assignments There will be 6 graded homework assignments, each due either one or two weeks after being assigned (see schedule for details). For these assignments, students have the option to collaborate and submit their work in pairs. Please consult the collaboration policy below.

AC209B students will have a short, additional HW component for some assignments. This will cover material from the advanced section.

Quizzes

During lectures, students will take a brief quiz focusing on both pre-class and in-class materials. Please note, the quizzes will not include content from the advanced sections.

The quizzes are short enough to be completed right after the lecture but they will remain available until the beginning of the following lecture. The lowest third of your quiz scores will be discarded.

Exercises Some lectures may feature coding exercises centered on the recently introduced topics. However, no exercises will be conducted during the AC209B advanced sections.

The exercises are short enough to be completed during the time allotted in lecture but they will remain available until the beginning of the following lecture.

Your final course grade will be calculated twice: once including exercise grades and once without. Your final course grade will be determined based on the higher of the two. In this way, exercises can only improve your grade.

Note: no exercises will be dropped in the grade calculation which includes them.

Course Resources

Online Materials Instructional course materials, including lecture slides, lab & section notebooks, and coding exercises will be published on Ed.

Assignments will only be posted on Canvas.

Working Environment You will be working in Jupyter Notebooks which you can run in your own machine or in the SEAS JupyterHub which will be available through Canvas.

Recommended Textbooks

- ISLR: An Introduction to Statistical Learning, 2nd ed. by James, Witten, Hastie, Tibshirani (Springer: New York, 2021)
- BDA3: Bayesian Data Analysis by Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, Donald B. Rubin (CRC Press: New York, 2013)
- Goodfellow: Deep Learning by Goodfellow, Bengio and Courville. (The MIT Press: Cambridge, 2016)
- Glassner: Deep Learning: A Visual Approach by Andrew Glassner (No Starch Press, 2021)
- SLP Speech and Language Processing by Jurafsky and Martin (3rd Edition Draft)
- INLP Introduction to Natural Language Processing by Jacob Eisenstein (The MIT Press: Cambridge, 2019) Free electronic versions are available (ISLR, Goodfellow, SLP, INLP) or hard copy through Amazon (ISLR, Goodfellow, Glassner, SLP, INLP).

Articles & Excerpts

- Unsupervised learning:
  - **Basics**: James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning (2nd ed.). New York: Springer. Chapter 12 https://hastie.su.domains/ISLR2/ISLRv2_website.pdf
  - **Silhouette plots**: https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html
  - **Gap statistic**: Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(2), 411-423. https://hastie.su.domains/Papers/gap.pdf
  - **DBSCAN**: Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. ACM Transactions on Database Systems (TODS), 42(3), 1-21. https://dl.acm.org/doi/pdf/10.1145/3068335?casa_token=_P479lYnlpsAAAAA:PckzU6ZiTt3yMNzFrXyzESZ3N_pp904kN0N2QEwIoq6CxtfPCxnL9bNTGtjhuiNtzSfKyXQI

- Bayesian material
  - **Basics**: Glickman, Mark E. and Van Dyk, David A. (2007) "Basic Bayesian Methods" In Topics in Biostatistics (Methods in Molecular Biology). Edited by Walter Ambrosius. The Humana Press Inc., Totowa, NJ. ISBN 1-58829-531-1. pp 319-338. Chapter accessible from http://www.glicko.net/research/glickman-vandyk.pdf
  - **Importance sampling, rejection sampling, MCMC, Metropolis, Gibbs sampler**: Andrieu, C., De Freitas, N., Doucet, A., & Jordan, M. I. (2003). An introduction to MCMC for machine learning. Machine learning, 50(1), 5-43. Article accessible from: https://www.cs.ubc.ca/~arnaud/andrieu_defreitas_doucet_jordan_intromontecarlomachinelearning.pdf
  - **Bayesian examples - regression, hierarchical modeling**: Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., & Rubin, D.B. (2013). Bayesian Data Analysis (3rd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b16018, http://www.stat.columbia.edu/~gelman/book/BDA3.pdf
    - Chapter 14: Introduction to regression models
    - Chapter 15: Hierarchical linear models
    - Chapter 16: Generalized linear models (includes logistic regression)

Getting Help For questions about homework, course content, package installation, etc., the process is:

- try to troubleshoot yourself by reading the lecture, lab, and section notes, and looking up online resources.

- go to office hours this is the best way to get help.
- post on the class Ed forum; we want you and your peers to engage in helping each other. TFs and the instructors also monitor Ed and will respond within 24 hours. Note that Ed questions are visible to everyone. If you are citing your homework solutions (code or other aspects of your work) you must post privately so that only the staff sees your message.
- watch for official announcements via Ed. These announcements will also be sent to the email address associated with your Canvas account so make sure you have it set appropriately.
- send an email to the Helpline cs109b2024@gmail.com for administrative issues and non-content specific questions.
- for personal matters that you do not feel comfortable sharing with the TFs, you may send an email to either or both of the instructors.

Course Policies and Expectations

Grading Your final grade for the course will be computed using the following weights:

Assignment	Final Grade Weight
Homework 1-6	50%
Midterms	20%
Quizzes	5%
Exercises	5%
Project	20%
Total	**100%**

This course uses the grading system outlined in the FAS Student Handbook:

A, A– Earned by work whose excellent quality indicates a full mastery of the subject and, in the case of the grade of A, is of extraordinary distinction.

B+, B, B– Earned by work that indicates a good comprehension of the course material, a good command of the skills needed to work with the course material, and the student's full engagement with the course requirements and activities.

C+, C, C– Earned by work that indicates an adequate and satisfactory comprehension of the course material and the skills needed to work with the course material and that indicates the student has met the basic requirements for completing assigned work and participating in class activities.

D+, D, D– Earned by work that is unsatisfactory but that indicates some minimal command of the course materials and some minimal participation in class activities that is worthy of course credit toward the degree.

E Earned by work that is unsatisfactory and unworthy of course credit toward the degree.

Numerical scores in the class will be converted to letter grades at the end of the course by the instructors.

Late Work Policy

Extension School Late Days **Extension School** students are allocated a total **4 late days** with **at most 2 days applied to any single homework**.

On-Campus Students Late Days **On-campus students** students are initally allocated a total of **3 late days**. with the possibility of acquiring more through attendance (see attendance policy below). **At most 2 late days can applied to any single homework**.

General Late Day Policies If a student has exhausted all their late days, late homework will not be accepted unless there is a medical (if accompanied by a doctor's note) or other official, University-excused reasons. There is no need to ask before using one of your late days.

Late days cannot be applied to quizzes, exercises, midterm components, or project milestones.

Attendance Policy

Attendance at lectures and labs is required for all on-campus students**. The teaching staff will record on-campus attendance. For every 8 sessions attended (i.e., lecture or lab), on-campus students will earn 1 additional late day.** Any effort to misrepresent attendance will be considered a violation of the honor code and be dealt with accordingly.**

Academic Integrity

We expect you to adhere to the Harvard Honor Code at all times. Failure to adhere to the honor code and our policies may result in serious penalties, up to and including automatic failure in the course and reference to the ad board.

DCE Academic Integrity Policy

If you are an Extension School student, you are responsible for understanding Harvard Extension School policies on academic integrity (https://extension.harvard.edu/for-students/student-policies-conduct/academic-integrity/) and how to use sources responsibly. Stated most broadly, academic integrity means that all course work submitted, whether a draft or a final version of a paper, project, take-home exam, online exam, computer program, oral presentation, or lab report, must be your own words and ideas, or the sources must be clearly acknowledged. The potential outcomes for violations of academic integrity are serious and ordinarily include all of the following: required withdrawal (RQ), which means a failing grade in the course (with no refund), the suspension of registration privileges, and a notation on your transcript.

Using sources responsibly (https://extension.harvard.edu/for-students/support-and-services/using-sources-effectively-and-responsibly/) is an essential part of your Harvard education. We provide additional information about our expectations regarding academic integrity on our website. We invite you to review that information and to check your understanding of academic citation rules by completing two free online 15-minute tutorials that are also available on our site. (The tutorials are anonymous open-learning tools.)

Student Collaboration

If you work with a partner on an assignment make sure both parties solve all the problems. Do not divide and conquer. You are expected to be intellectually honest and give credit where credit is due. In particular:

- if you work with a fellow student and want to submit the same notebook you need to form a group prior to the submission. Details in the assignment. Not all assignments will permit group submissions.
- you need to write your solutions entirely on your own or with your collaborator
- if you worked with a fellow student on a paired assignment but decide in the end to submit different notebooks individually, include the name of the other student as a comment at the top of your notebook.
- you are welcome to take ideas from code presented in labs, lecture, or sections but you will need to change it, adapt it to your style, and ultimately write your own. Simply copying verbatim will rarely be successfully.
- if you use code found on the internet, books, or other sources you need to cite those sources.
- you should not view any written materials or code created by other students for the same assignment.

- you may not provide or make available solutions to individuals who take or may take this course in the future. If you are using a remote git repository such as GitHub to work on your assignments **you must make it private.**

Use of AI Models

Purpose of Policy: This policy outlines the acceptable use of AI models, including but not limited to ChatGPT, in completing assignments for this course.

Policy Guidelines:

1. **Original Work:** Students are expected to complete assignments using their original thoughts and interpretations. AI models can be used to help understand concepts, generate ideas, or learn about different perspectives, but they should not write or complete assignments for students.
2. **Collaboration with AI:** Students may use AI models for brainstorming or generating preliminary ideas, but the final work submitted must be substantially their own. Students should be able to explain their reasoning, logic, and conclusions without relying on the model's output.
3. **Restrictions for Specific Assignments:** There may be specific assignments (e.g. quiz part of the midterms) or parts of the course where the use of AI models is entirely prohibited. These restrictions will be clearly stated in the assignment guidelines.
4. **Ethical Considerations:** Students are encouraged to approach the use of AI with ethical considerations in mind, including issues related to privacy, bias, and authenticity.

Consequences for Non-Compliance: Failure to adhere to this policy may result in academic penalties as outlined in the course's academic integrity policy.

Questions and Clarifications: If students have questions about the appropriate use of AI models in an assignment, they should consult the course instructor or teaching assistants before proceeding.

Please refer to the University's policy for further information.

Accommodations for Students with Disabilities

Harvard students needing academic adjustments or accommodations because of a documented disability must present their Faculty Letter from the Accessible Education Office (AEO) and speak with the professor by the end of the second week of the term, (fill in specific date). Failure to do so may result in the Course Head's inability to respond in a timely manner. All discussions will remain confidential, although Faculty are invited to contact AEO to discuss appropriate implementation.

Harvard Extension School is committed to providing an inclusive, accessible academic community for students with disabilities and chronic health conditions. The Accessibility Services Office (ASO) https://www.extension.harvard.edu/resources-policies/accessibility-services-office-aso offers accommodations and supports to students with documented disabilities. If you have a need for accommodations or adjustments in your course, please contact the Accessibility Services Office by email at accessibility@extension.harvard.edu or by phone at 617-998-9640.

Diversity and Inclusion Statement

As educators, we aim to build a diverse, inclusive, and representative community offering opportunities in data science to everyone. We will encourage learning that advances ethical data science, exposes bias in the ways data & data science can be (and all too frequently is) used, and advances research into fair and responsible data science.

We need your help to create a learning environment that supports a diversity of thoughts, perspectives, and experiences, and honors your identities (including but not limited to race, gender, class, sexuality, religion, ability, etc.) To help accomplish this:

- If you have a name and/or set of pronouns that differ from those in your official Harvard records, please let us know!
- If you feel like your performance in the class is being impacted by your experiences outside of class, please do not hesitate to come and talk with us. We want to be a resource for you. Remember that you can also submit anonymous feedback (which will lead to us making a general announcement to the class, if necessary, to address your concerns). If you prefer to speak with someone outside of the course, you may find helpful resources at the Harvard Office of Diversity and Inclusion.
- We (like many people) are still learning about diverse perspectives and identities. If something was said in class (by anyone) that made you feel uncomfortable, please talk to us about it.
- As a participant in course discussions, you are expected to respect your classmates' diverse backgrounds and perspectives.

Our course will discuss diversity, inclusion, and ethics in data science. Please contact us (in person or electronically) or submit anonymous feedback if you have any suggestions for how we can improve.

For additional resources, guidance, and support related to diversity and inclusion, please refer to the Harvard Office for Equity, Diversity, Inclusion, & Belonging.

Auditing

To request to audit the course, send an email to cs109b2024@gmail.com with your HUID (required) and a statement of agreement to the terms below. **Note:** Please make sure you are not currently enrolled in the course when you send your request. You can't be added as an auditor in Canvas if you are currently listed there as an enrollee.

All auditors must agree to abide by the following rules:

- Auditors must attend class in person. This is a Harvard policy. Auditors who do not confirm their presence during the first week of in-class instruction will lose course access.
- Auditors are held to the same standard of academic honesty as our registered students. Please do not share homeworks or solutions with anyone. Violations will be reported to the Harvard Administrative Board.
- Auditors are not permitted to take the course for credit in the future.
- Auditors should **not** submit HWs or midterms, or participate in projects.
- Auditors should refrain from using any course and TF resources that are designed for our registered students like Ed, FASOnDemand, and office hours.