ASTR 496: Foundations of Data Science in Astronomy

Prof. Gautham Narayan
✉: gsn@illinois.edu
☎: +1 (217) 300-7322

Lecture: Astronomy 134, Tue & Thur, 1230 – 1350
🌍: `https://github.com/gnarayan/ast496_2025_Fall`
Office Hours: Astronomy 129, Thu 1600 – 1700
Class Zoom: `https://go.illinois.edu/gsn`

COURSE DESCRIPTION & LEARNING GOALS

This 16 week course (3 contact hours) will cover a number of statistical techniques that are relevant to astrophysical studies. These include robust statistics, regression, model building and hypotheses testing, MCMC methods, parameter estimation, time series analysis, clustering and dimensionality reduction, and hierarchical modeling. We will also cover best practices for writing code and version control. These techniques are ubiquitous in science and industry. My goal is to provide a survey of these techniques, together with realistic problems, so that you see how they work and what their implicit assumptions are. The course TA is Joseph Weller, and he has office hours T 1130-1230 (before class) and R 1500-1600 (before my office hours).

By the end of this semester you should be able to:

- Apply a variety of existing models in astronomy, and interpret the results and compare to existing literature.
- Identify when existing models are inadequate descriptions of the astrophysical circumstances under study and develop new models using your knowledge of statistical and computational techniques.
- Report inferences from your models with appropriate uncertainties, develop visualizations that convey the robustness of the model you selected, and describe what you did in a manner appropriate for high-impact scientific journals.

To achieve these outcomes, students enrolled for 3 credit hours are expected to work 6 hours/week outside instruction time.

PREREQUISITES

Required: ASTR 210 & 310, to provide the necessary background in modern astronomy and Python programming. These have additional calculus & physics prerequisites.

Recommended: Prior coursework in undergraduate statistics and undergraduate linear algebra (e.g., MATH 227 or 257). You will also need a computer with a working `conda` and `git` installation for much of the coursework.

TEXTS & READINGS

- “Statistics, Data Mining, and Machine Learning in Astronomy”, Ž. Ivezić, A. Connolly, J. T. VanderPlas & A. Gray
- “Python Data Science Handbook”, J T. VanderPlas

OTHER RESOURCES

“Modern Statistical Methods for Astronomy”, E. Feigelson with J. Babu (**FB**) is detailed, though focuses on using R - available for free as a an e-book through NASA ADS. You may also find “Bayesian Models for Astrophysical Data”, J. M. Hilbe, R. S. de Souza, & E. E. O. Ishida helpful (1st edition., Cambridge University Press, ISBN 9781107133082).

The LSST Data Science Fellowship Program has a huge collection of worked notebooks and video lectures: `https://github.com/LSSTC-DSFP/LSSTC-DSFP-Sessions`.

GRADING

Your grade is determined from a combination of assignments, midterm and a final project. Policies for each are below. This course awards 3 credit hours. You will be required to deliver a 15-minute presentation in the last week of the semester covering one specialized analysis topic, chosen together with me. The presentation must survey how the analysis technique is used in the astrophysics and statistical literature at present, and provide a worked Jupyter notebook illustrating its application to a particular problem in astrophysics. Ideally, the technique is is directly relevant to the any ongoing research you are involved in. **All coursework uses Github, not a LM like Canvas, to replicate realistic research conditions.** Periodic overall grade updates are sent via email. You are welcome to discuss your grades and your work in the course with me during office hours.

Points:

- Weekly Assignments: 40% - excluding midterm/final week - i.e. 12 total, equally weighted
- Midterm: 25%
- Final + Presentation: 35%

This course uses a plus (+) and minus (−) grading scale for course grades. Grades are rounded up (i.e. ceil()) to the nearest integer.

97-100=A+; 93-96=A; 90-92=A-; 87-89=B+; 83-86=B; 80-82=B-; 77-79=C+; 73-76=C; 70-72=C-; 67-69=D+; 63-66=D; 60-62=D-; 0-59=F

COURSE POLICIES

I’ve outlined standards for this course below. Times listed in this syllabus are US/Central throughout. If something is not covered by my policies, please discuss it with me. My contact information is at the beginning of this syllabus.

Assignment & Exam Policies: Assignments, as well as midterm and final examinations are open book and take home. You may work in groups, and may discuss the assignments and ways to tackle it, but you must write/code your solution independently. This means you get to talk with each other, discuss how you’d solve a problem, but come up with your own solution, but not share your solutions. Over the course of the last three semesters, a total of 5 students thought I’d not catch that level of cheating. They all failed.

Assignments/exams will be posted to the course GitHub repo on Thursdays. Make a fork of the repo, create a folder with your name for your work, write/code up your solution as directed in the assignment, commit, and open a pull request when you are satisfied with your work before Noon the following Wednesday. You are allowed to drop ONE assignment from your total, for whatever reason, no questions asked (and if you don’t elect to, I’ll drop your lowest).

The midterm and final examinations will be posted online, and will be due by Noon six days later. If you have a conflict with the exam dates, please contact me as soon as possible. Make up examinations will have different questions. Exams include all material covered prior, and will require a more substantial time commitment that the weekly assignments.

Absences and Grades of Incomplete: As described in the Student Code, I will accomodate reasonable absences from class that result in missed work. Documentation must be provided to the instructor, and such documentation can be obtained from the Connie Frank CARE Center. Incomplete (I) grades are given by the college (LAS) only in situations where unexpected emergencies prevent you from completing the course and the remaining work can be completed the next semester. As described in the LAS Policies and Procedures, the college is the final authority on whether you qualify for an incomplete, not the instructor.

Late or Missed Assignments: All work is assigned on Thursday and due the following Thursday before class begins. If you know that you will be turning an assignment in late please notify me in advance. A full letter grade will be deducted for each day an assignment is late until a "F" grade is achieved, unless you have a documented medical excuse or you have notified me of other extenuating circumstances. Remember that you may drop ONE assignment from your total, for whatever reason, no questions asked.

Accessibility Accommodation: It is my goal that this class be an accessible and welcoming experience for all students, including those with disabilities that may impact learning in this class. If the design of this course poses barriers to you effectively participating and/or demonstrating learning in this course, please meet with me, with or without an Accessibility Services accommodation letter, to discuss reasonable options or adjustments. You are welcome to talk to me at any point in the semester about course design concerns, but it is always best if we can talk at least one week prior to the need for any modifications. During our discussion, I may suggest the possibility/necessity of your contacting the Office of Disability Resources and Educational Services (1207 S. Oak St., Champaign, IL 61820; 217-333-1970) disability@illinois.edu; `http://disability.illinois.edu/`) to talk about academic accommodations.

Plagiarism: **Don't.** You are going to be using GitHub for assignments, so there's a record of your commits, and it is trivial to check if chunks of your work match someone else. You may work in groups together, and may discuss the assignments and ways to tackle it, but you must write/code your solution independently. Read the University of Illinois' policy on plagiarism.

Plagiarism and cheating of any kind on an assignment or examination will result at least in an "F" for that work, and may also lead to an "F" for the entire course. Plagiarism and cheating subjects a student to referral to the Senate Committee for Student Discipline for further action.

I am confident in each of your ability to tackle the course work. My group work policy is designed to encourage you to learn how to collaborate, but the assignments are designed to test YOUR grasp of the material. If you feel you need help with material, come see me during office hours or any time my door is open.

Classroom Behavior: I expect you to live up to your roles as student-scholars. Students must follow the University of Illinois' standards for personal and academic conduct. Proper conduct entails creating a **positive** learning experience for all students, regardless of sex, race, religion, sexual orientation, social class, or any other feature of personal identification; therefore, **sexist,**

racist, prejudicial, homophobic, or other derogatory remarks will not be tolerated.

Syllabus Amendment: This syllabus may be amended or modified in any way upon notice, with the version on GitHub being authoritative. Most such changes will affect the tentatve schedule.

Important Dates:

- Aug. 25, 2025: First day of class
- Oct. 9, 2025: Midterm Exam Assigned (due Oct. 15 by Noon)
- Dec. 10, 2025: Last day of classes
- Dec. 11, 2025: Final Exam Assigned (due Dec. 17 by Noon)

CLASS SCHEDULE FALL 2025 (subject to revision)

- **Aug 28**
  First steps, crash course in python. **NO CLASS AUG 26**.
- **Sep 2, 4**
  Probability distributions, descriptive statistics, the Central Limit theorem and when it doesn't hold, robust statistics, and hypothesis testing (ICVG Ch. 3, FB Ch. 2). **CLASS OVER ZOOM THIS WEEK.**
- **Sep 9, 11**
  Statistical inference, frequentist properties such as unbiasedness & the Cramér–Rao bound, consistency, asymptotic limits, mean-squared errors (ICVG Ch. 4, FB Ch. 3)
- **Sep 16, 18**
  Maximum likelihood estimation and applications, ranting about minimizing χ^2 (ICVG Ch. 4). **CLASS OVER ZOOM ON SEP 18.**
- **Sep 23, 25**
  Regression & Inference: ordinary least squares, generalized least squares, orthogonal distance regression vs generative modeling of data (ICVG Ch. 8, FB Ch. 7)
- **Sep 30, Oct 2**
  Bayes in practice, sampling and Markov Chain Monte Carlo methods (ICVG Ch. 5)
- **Oct 7, 9**
  Building models, effective sampling techniques, estimating parameters & uncertainties, posterior predictive checks, other MCMC wizardry (ICVG Ch. 8). **Midterm exam posted.**
- **Oct 14, 16**
  Visualization (VdP Ch. 4), Midterm exam due. No homework assignment because it's spring break and I'm not that mean.
- **Oct 21, 23**
  Time-series analysis (ICVG Ch. 10, FB Ch. 11), Gaussian processes (ICVG Ch. 8.10, readings from Rasmussen & Williams)
- **Oct 28, 30**
  Probabilistic Graphical Models (PGMs) & hierarchical Bayes (Readings from Hilbe, de Souza & Ishida)
- **Nov 4, 6**
  The ABCs of not having a likelihood function (Readings from Hilbe, de Souza & Ishida)
- **Nov 11, 13**
  Intro to Machine learning, tree methods (ICVG Ch. 9, VdP Ch. 5)
- **Nov 18, 20**
  Gaussian mixture models, density estimation, unsupervised clustering techniques, and dimensionality reduction (ICVG Ch. 6, 7, FB Ch. 9 and bits of Ch. 6, VdP Ch. 5)
- **Dec 2, 4**
  Dealing with outliers, imbalanced, and missing data, supervised machine learning techniques
- **Dec 9**
  Student presentations. **Final exam posted on Dec 11th.**
- **Dec. 17th**
  Final exam due by Noon.