Ay128/256: Astronomy Data Science Lab

UC Berkeley, Spring 2019

This course consists of three data-centric laboratory experiments that draw on a variety of tools used by professional astronomers. Students will learn to procure and clean data (drawn from a variety of world-class astronomical facilities), assess the fidelity/quality of data, build and apply models to describe data, learn statistical and computational techniques to analyze data (e.g., Bayesian inference, machine learning, parallel computing), and effectively communicate data and associated scientific results. This class will make use of data from facilities such as Gaia, the Sloan Digital Sky Survey, and the Hubble Space Telescope to explore the structure and composition of the Milky Way, stars, and galaxies throughout the local and distant Universe. There is a heavy emphasis software development in the Python language, statistical techniques, and high-quality communication (e.g., written reports, oral presentations, and data visualization).

**Instructor:** Profs. Dan Weisz (dan.weisz@berkeley.edu) and Josh Bloom (joshbloom@berkeley.edu)

**GSIs/uGSIs:** Kareem El-Badry (kelbadry@berkeley.edu)

**Day, Time, Location:** Monday 4-7pm in 131 Campbell

**Email:** datalab@astro.berkeley.edu

**Website:** https://ucb-datalab.github.io

**Offices:** 311 Campbell (Dan), 203 Campbell (Josh), 407 Campbell (Kareem)

**Office Hours:**

Kareem: Tu 5-6pm and Fri 9-10am in 419 Campbell

Dan: Weds 4-5pm in 355 Campbell

Josh: Thurs: 2:30-3:30 in 203 Campbell

**Textbook:** "Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data" by Željko Ivezić et al. **Link:** http://a.co/i3fnkw4

and select readings to be provided.

**Prerequisites:**

- This class assumes that you have completed introductory astrophysical instruction (Astro 7A and 7B, or higher such as Astro 160/161) as well as knowledge of calculus (including Math 53) and linear algebra (Math 54 or Physics 89)
- You should have proficiency or fluency in the Python programming language. This class heavily emphasizes software development, and is **not** the place to learn Python for the first time.

**Class Participation (25% of grade):**

- Active engagement in class discussion and lecture
- Presenting work during weekly "show and tell"
- "Show and Tell": During an ongoing lab, we will start class with show and tell so that everyone knows the status. This is your opportunity to solve your problems and see how others are approaching the task in hand. Come to class prepared to describe what you have done in the previous week. Ask questions and interrogate the instructors and your fellow students.

- Coding exercises: discussion of results / challenges during weekly "show and tell", timely posting of code to github

**Lab Reports (60%):**

- Submitted as annotated Jupyter notebooks via Piazza
- due **before** specified class.
- -10% for each day late
- collaborate (talk, draw pictures, analyze data) with your lab mates, but you **MUST** implement separately (your own equations, code, plots, writing)

**Final Project & Presentation (15%):**

- Students pitch a project to the instructors and will have 2-3 weeks at the end of the semester to complete it. A complete final project includes a Jupyter notebook and a AAS style talk on the project to the class. We will have a presentation day on the last day of regular classes (4/29).

**Reading:**

- lab instructions and topical handouts linked on the class webpage

**Materials:**

- We can set up an account on a department computer, if you wish
- Datalab instance

**Schedule:**

- Format: 1 weekly 3 hour meeting: First 1.5 hours consist of "show and tell" progress reports from all students, oral presentations on lab findings (when labs are due). The second 1.5 hours is lecture from the instructor on a new topic related to the ongoing or upcoming lab.

- **Weeks 1-2**: Logistics, Introduction to data science in astronomy.
    - **Lab #0 assignment:** Use ADQL to query Gaia database, construct HRDs for various volumes of the Milky Way, and as a function of position on the sky, over-plot predictions from stellar evolution models

- **Weeks 3-6:** Lab Assignment # 1 -- **Gaia, RR Lyrae, and Galactic Dust** — The aim of this lab is use the sample of RR Lyrae in the Gaia catalog to build a 2D dust map of the Milky Way. We make use of the fact that RR Lyrae are standard candles to probe the dust content along the line of sight. This lab includes several ADQL queries, how to identify "bad" data, fitting models to time series data, modeling dust along the line of sight. Technical skills include Bayesian model fitting vs. optimization, sampling, posterior and convergence checks, visualization (posteriors, 2D dust maps).

- **Weeks 7-10:** Lab Assignment #2 -- **Data Driven Modeling of Stellar Spectra** — The goal of this lab is to (a) build a model to predict what the spectrum of a star should look like for a given set of stellar parameter or "labels" (e.g., Teff, logg, abundances) and (b) use this model to infer properties of stars observed by APOGEE by fitting their spectra. Technical topics include: interpolation and resampling techniques, linear models, techniques for numerical

- **Weeks: 11-14:** Lab Assignment # 3 -- **The Hubble Constant** — The aim of this lab is measure the local Hubble Constant, H_0, by building a hierarchical model for the distance ladder and ultimately using SNe Ia as standard candles. We'll use the data and general method outlined in Riess et al. (2016). Though we'll use the same data as Riess et al., we'll develop a hierarchical

Bayesian model instead of the maximum likelihood approach they use. Technical skills hierarchical modeling, efficiently sampling high dimensional models, STAN, more sophisticated time series fitting, systematic vs. random errors, appropriately propagating errors.

- **Important Note:** You will also be working on your final presentation/project during part of Lab 3. This will be an extremely busy and intense period of the class, so please plan accordingly.

**Class Philosophy:**
This class is effectively an introduction to astronomical research from a data-centric perspective. The labs are designed to be challenging, yet manageable mini-research projects. They will be time consuming.

Often when doing research, you will not have all the necessary technical skills to complete a project. It is normal to have to learn entirely new skills along the way. A main aim of this class is to provide exposure to those skills, in a setting that requires you to think and act like a researcher, but with extra guidance and resources. You should also get used to asking questions that seem basic (sometimes embarrassingly so) in hindsight, and working in open and collaborative environments.

The weekly lectures will provide you with broad background, but are not intended to teach all the nitty-gritty technical skills you need. Much like real research, these are skills you learn on your own, in groups, and/or in consultation with an advisor. The equivalent with this class are questions via Piazza and office hours, particularly with Kareem.

Research is fun, but challenging. It usually does not progress linearly, and you often have to take a few steps backward before ultimately moving forward.

**Class Conduct:**
This is a work-intensive class. You are going to spend significant time on your own in the lab with minimal supervision. At all times, you are expected to abide by the UC Berkeley Code of Conduct (http://sa.berkeley.edu/code-of-conduct), acting with respect to your peers, GSIs, and instructor. Should you experience any form of harassment or discrimination, we maintain a list of resources that can help you decide how to respond. (https://astro.berkeley.edu/department-resources/reporting-harassment). GSIs and instructors are non-confidential reporters; we have a legal obligation to act on any reports of harassment. Please know that we take our responsibility seriously.