Ay128: Undergraduate 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: Prof. Dan Weisz

GSIs/uGSIs: TBD

Day, Time, Location: Monday 4-7pm in 131 and/or 541 Campbell, 3 hours of lecture / lab instruction per week

Email: dan.weisz@berkeley.edu

Office: Campbell 311

Office Hours: 2 hours per week, times TBD

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) 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%):

- Written up using Overleaf/Sharelatex using Astrophysical Journal publication templates
- due **before** specified class.
- -10% for each day late
- collaborate (talk, draw pictures, analyze data) with your lab mates, but 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 short journal style write up + a AAS style talk on the project to the class

Reading:

- lab instructions and topical handouts linked on the class webpage

Materials:

- you may use department computers; an account has been made for you.

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, Data storage, Collecting data & Querying databases, cleaning data / assessing fidelity, Pre-lab with Gaia Data Release 2 (construct H-R diagrams, compare with stellar models, and associated exercises), advice for lab write ups
    - Week 1
        - **Lecture Topics:** Introduction, Overview of Gaia Mission, The Integrity of Data
        - **Reading assignment:** Ch 1.1-1.5, Ch 2.1-2.3 of Ivezic, Ch 2 https://gea.esac.esa.int/archive/documentation/GDR2/index.html
        - **Coding Exercises:** Work through the instructor designed ADQL tutorial designed for our class
        - **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
    - Week 2
        - **Lecture Topics:** Cross-Matching Catalogs, Methods of 2D visualization (e.g., scatter points vs. density plots, color schemes (pros and cons), labeling plots
        - **Reading assignment:** Ch 1.6-1.7 Of Ivezic

        - **Coding Exercises:** Instructor designed coding problems
        - **Lab #0 assignment:** Use ADQL to query Gaia database, construct HRDs for various volumes and as a function of position on the sky, overplot predictions from stellar 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).
    - Week 3
        - **Lecture Topics:** Review of basic probability, Intro to Bayesian Analysis
        - **Reading assignment:** Ch 3,4 of Ivezic
        - **Coding Exercises:** Instructor designed coding problems
    - Week 4
        - **Lecture Topics:** Probabilistic Model Fitting: Principles & Practical Advice
        - **Reading assignment:** Ch 1, 2 of Hogg et al. 2010 (https://arxiv.org/pdf/1008.4686)
        - **Coding Exercises:** Exercises 1-4 from Hogg et al. 2010
    - Week 5
        - **Lecture Topics:** Sampling, Optimization, MCMC, Visualizing Posterior distributions
        - **Reading assignment:** Ch 1-5, 7 of Hogg & Foreman-Mackey 2018 (https://arxiv.org/abs/1710.06068)
        - **Coding Exercises:** Exercises 1-7 from Hogg & Foreman-Mackey 2018
    - Week 6
        - **Lecture Topics:** Intro to Machine Learning
        - **Reading assignment:** Ch 2. Of Ivezic
        - **Coding Exercises:** Instructor designed coding problems

- **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 & computational efficiency, and deep learning and neural networks.

- Week 7
  - **Lecture Topics:** Building data-driven models, linear models
  - **Reading assignment:** Ch 6 of Ivezic
  - **Coding Exercises:** Instructor designed coding problems
- Week 8
  - **Lecture Topics:** Interpolation, resampling
  - **Reading assignment:** Ch 7 of Ivezic
  - **Coding Exercises:** Instructor designed coding problems
- Week 9
  - **Lecture Topics:** Improving numerical and computational efficiency
  - **Reading assignment:** Ch 8 of Ivezic
  - **Coding Exercises:** Instructor designed coding problems
- Week 10
  - **Lecture Topics:** Deep Learning
  - **Reading assignment:** Ch 9 of Ivezic
  - **Coding Exercises:** Instructor designed coding problems

- **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.
  - Week 11
    - **Lecture Topics:** Approaches to Measuring Galaxy Distances, Time Series Analysis
    - **Reading assignment:** Ch 10 of Ivezic
    - **Coding Exercises:** Instructor designed coding problems

- Week 12
  - **Lecture Topics:** Building Hierarchical Models and using STAN to sample
  - **Reading assignment:** STAN tutorial (http://mc-stan.org/users/documentation/tutorials)
  - **Coding Exercises:** Instructor designed coding problems
- Week 13
  - **Lecture Topics:** Propagating Uncertainties
  - **Reading assignment:** Andre 2010 (https://arxiv.org/pdf/1009.2755)
  - **Coding Exercises:** Instructor designed coding problems
- Week 14
  - **Lecture Topics:** Advice on Professional Presentations
  - **Reading assignment:** http://www.astrobetter.com/wiki/Presentation+Skills, https://arxiv.org/pdf/1712.08088

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.