ER 131: Data, Environment and Society

Instructor: Duncan Callaway, dcal@berkeley.edu
Graduate Student Instructor: Salma Elmallah, salmae@berkeley.edu

Fall, 2019
4.0 Units

Lecture time and location: Tu/Th 9:30-11:00am, 202 Wheeler

Lab time and location:

- Mo 10:00a-12:00p, 166 Barrows (#33106)
- Tu 2:00p-4:00p, 2032 VLSB (#33276)

Office hours:

- Wednesday 9:30-10:30a, Barrows 325, by appointment (Duncan)
- Thursday 11:00a-12:00p, Barrows 325, open door (Duncan)
- Wednesday 3:00-4:00p, Barrows 399, open door (Salma)

Course Description

This course will teach students to build, estimate and interpret models that describe phenomena in the broad area of energy and environmental decision-making. The effort will be divided between (i) learning a suite of data-driven modeling approaches, (ii) building the programming and computing tools to use those models and (iii) developing the expertise to formulate questions that are appropriate for available data and models. Our goal is that students will leave the course as both critical *consumers* and responsible *producers* of data driven analysis.

We will work in Python in this course, and students must have taken Data 8 before enrolling. The course is designed to complement and reinforce Berkeley's data science curriculum, in particular Data 100 (though D100 is not a prerequisite). Whereas Data 100 focuses on a very broad set of data science tools including modeling, web technologies, working with text, databases and statistical inference, this course focuses more on how to use prediction methods as decision-making tools in energy and environment contexts.

This is a four unit course, with three hours of lecture and two hours of lab section each week. Lectures will focus on theoretical and conceptual material but also introduce the programming structures required to use the material. Labs will be computer working sessions with a GSI and lab helpers available to work through weekly lab exercises.

Prerequisites

- (required) Foundations of Data Science (CS/ INFO/ STAT C8)
- (recommended) Computing: An introductory programming course (CS61A or CS88).
- Math:
  - (required) High school or college calculus.
  - (recommended) Linear Algebra (Math 54, EE 16a, or Stat89a).

Satisfaction of degree requirements

This course can be used to satisfy the following requirements.

- Upper division domain emphasis for Data Science major
- Engineering Elective for Energy Engineering
- Upper division requirement for Energy and Resources Group minor

Resources

- You will need your own computer, but virtually any operating system will do (OSX, Windows, Linux, Chromebook).
- We will draw some material from Berkeley's Data 100 course book, freely available here: https://www.textbook.ds100.org
- We will draw material from the excellent text book, *Introduction to Statistical Learning*, available in both print and pdf form.
- We will do a variety of readings from peer reviewed journals and popular press.
- Lectures, readings, and solutions will all be available on the course github site: `https://github.com/duncancallaway/ER131_2019`.
- You'll complete all HW and lab assignments using Python, within Jupyter notebooks hosted on `datahub.berkeley.edu`.
- Links for assignments will be posted on bCourses. You'll submit you work there, too.

Assessment

The course will have weekly labs and homework assignments, a mid-term and a final project. Grading will be as follows:

- Homework: 20%
  - There will be ten. We drop the lowest grade.

  - HW will be released on Thursdays and due the following Thursday.

- Lab assignments: 15%
  - There will be nine. We drop the lowest grade.
  - Released on Mondays and due the following Monday.
  - Attendance is 40% of lab grade, completing the lab is 60% of the grade.
  - Grading will focus on completeness rather than correctness.
- Mid-term: 25% (November 19, in class)
- Final poster: 10% (the poster session will be December 17, 3-6pm)
- Final project: 30% (due December 18 at 6am)

Late policy:

- You may request up to two extensions of two days over the course of the semester. You may distribute those extensions as you wish over homework and lab assignments. Otherwise, we will not accept late homeworks and labs. Coordinate extension requests with the GSI.
- The poster must be presented during the poster session to receive credit.
- For the final project, we drop 10 points out of 100 for each day late, or roughly a full letter grade. Projects submitted after 11:59am on December 20, 2019 will not receive credit.

Working in groups

Homework and labs

You are encouraged to learn from one another by brainstorming solution strategies. However the work you submit must clearly be your own. We will give zero credit for assignment submissions that are identical to one another. If you work with others, be sure to finish assignments on your own. Comments and markdown cells must clearly be your own.

Final project

You must work in groups of 2-3 for the final project. The final project writeup must include a statement describing each team member's contributions and a statement that all team members agreed the division of labor was equitable.

Schedule

Session	Day / Week	Topic	Methods Reading	Domain Reading	Homework assigned
Lecture 1	8/29/19	Course introduction			
Lab 1	Week of 9/2/19	No lab – week of labor day			
Lecture 2	9/3/19	Data design		Blei and Smyth	
Lecture 3	9/5/19	Pandas, variable types and file types	DS100 Ch3		HW1: Getting started
Lab 2	Week of 9/9/19	Answer HW1 questions; Pandas			
Lecture 4	9/10/19	Pandas, ctd, and data for HW2 (PM2.5)		Kleinberg et al; Athey.	
Lecture 5	9/12/19	Merge, groupby, pivot	DS100 Ch4, 5		HW2: Pandas, PM2.5 and fires
Lab 3	Week of 9/16/19	Answer HW questions; Exploratory data analysis			
Lecture 6	9/17/19	Exploratory data analysis		Hino et al; Pelletier et al	
Lecture 7	9/19/19	Visualization	DS100 Ch6		HW3: EDA; Wildfire ignitions
Lab 4	Week of 9/23/19	Answer HW questions, visualization			
Lecture 8	9/24/19	Intro to modelling; KNN	DS100 Ch 10; ISLR Ch 2		
Lecture 9	9/26/19	Reviewing regression; Guest lecture from Dan Kammen		Sunter et al	HW4: Visualization; renewable energy data
Lab 5	Week of 9/30/19	Basic modeling, KNN			
Lecture 10	10/1/19	Regression	ISLR 3.1-3.3		
Lecture 11	10/3/19	Multiple Regression; Land Use Regression		Novotny et al	HW5: regression

Continued on next page

Schedule

Session	Day / Week	Topic	Methods Reading	Domain Reading	Homework assigned
Lab 6	Week of 10/7/19	Answer HW questions, regularization			
Lecture 12	10/8/19	Gradient Descent	DS100 Ch 11		
Lecture 13	10/10/19	Resampling	ISLR 5.1-5.2		HW 6: Gradient descent; a "theory" homework.
Lab 7	Week of 10/14/19	Ans HW questions; gradient descent			
Lecture 14	10/15/19	Resampling; Model selection and regularization	ISLR 6.1-6.2		
Lecture 15	10/17/19	Finish regularization		Metz	HW 7 - Resampling; A "theory" homework
Lab 8	Week of 10/21/19	Review resampling			
Lecture 16	10/22/19	Classification	ISLR 4.1-4.3		
Lecture 17	10/24/19	Classification and regression trees	ISLR 8.1-8.2		HW8 - Model selection applied to land use regression
Lab 9	Week of 10/28/19	Review classification			
Lecture 18	10/29/19	Boosting, bagging, random forests		Wattenberg et al	
Lecture 19	10/31/19	Environmental Justice		Pastor	HW9 - Classification and regression trees with Cal Enviroscreen Data
Lab 10	Week of 11/4/19	TBD			
Lecture 20	11/5/19	Support vector machines	ISLR 9.1-9.3		

Continued on next page

Schedule

Session	Day / Week	Topic	Methods Reading	Domain Reading	Homework assigned
Lecture 21	11/7/19	Wrap up support vector machines		Badger	HW10 – Support vector machines with Cal Enviroscreen Data
Lab 11	Week of 11/11/19	No lab – week of veteran's day			
Lecture 22	11/12/19	Career panel			
Lecture 23	11/14/19	Exam Review through HW10 / Lecture 21			
Lab 12	Week of 11/18/19	Exam Review through HW10 / Lecture 21			
Lecture 24	11/19/19	Exam			
Lecture 25	11/21/19	Guest lectures		Reading TBD	
Lab 13	Week of 11/25/19	Project check in			
Lecture 26	11/26/19	Neural Networks	Mah		
Lab 14	Week of 12/2/19	Project check in			
Lecture 27	12/3/19	Neural Networks	Gabrys et al		
Lecture 28	12/5/19	slack			