Data, Environment and Society (ER131)

Lecture 1: Introduction

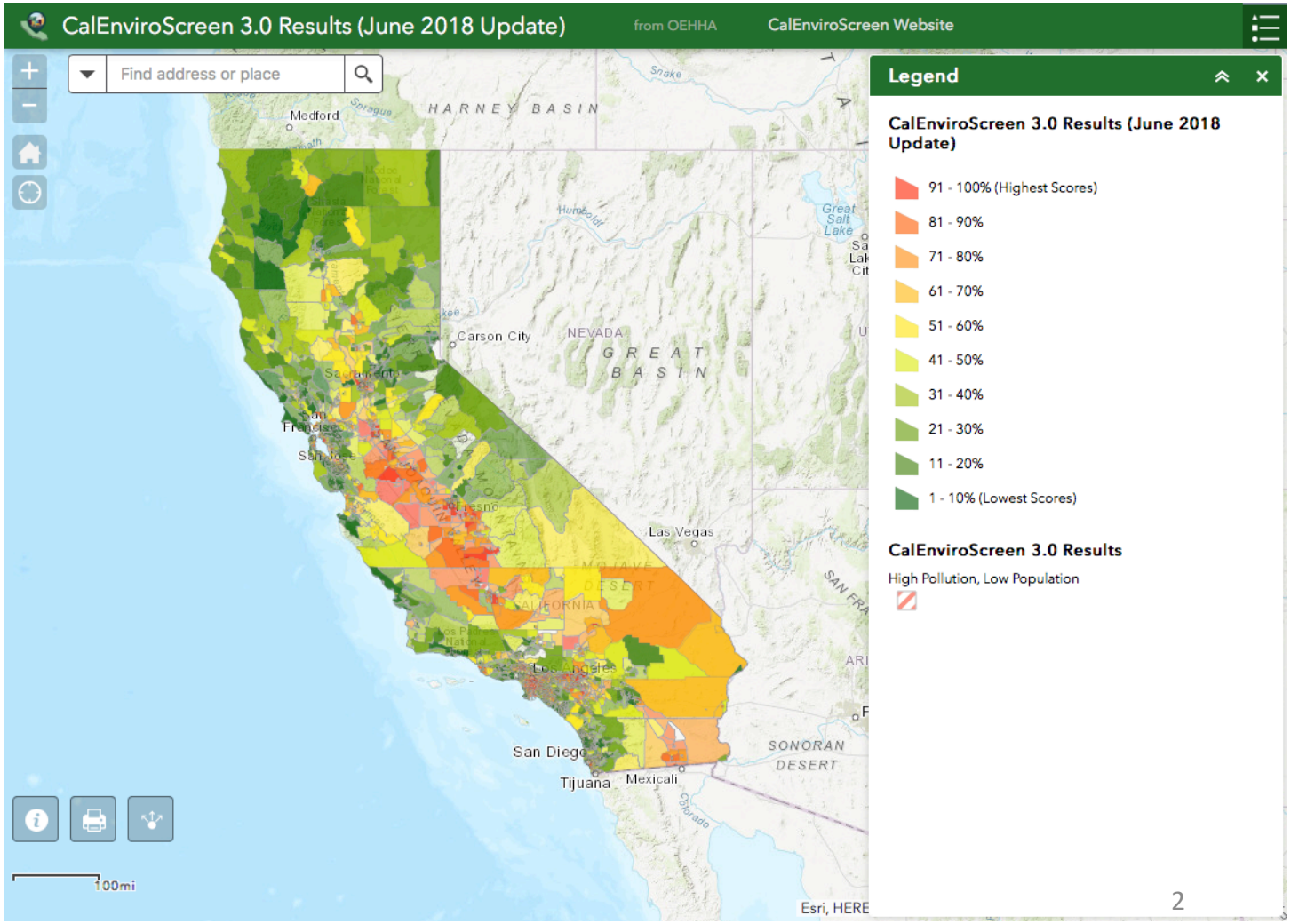
August 29, 2019

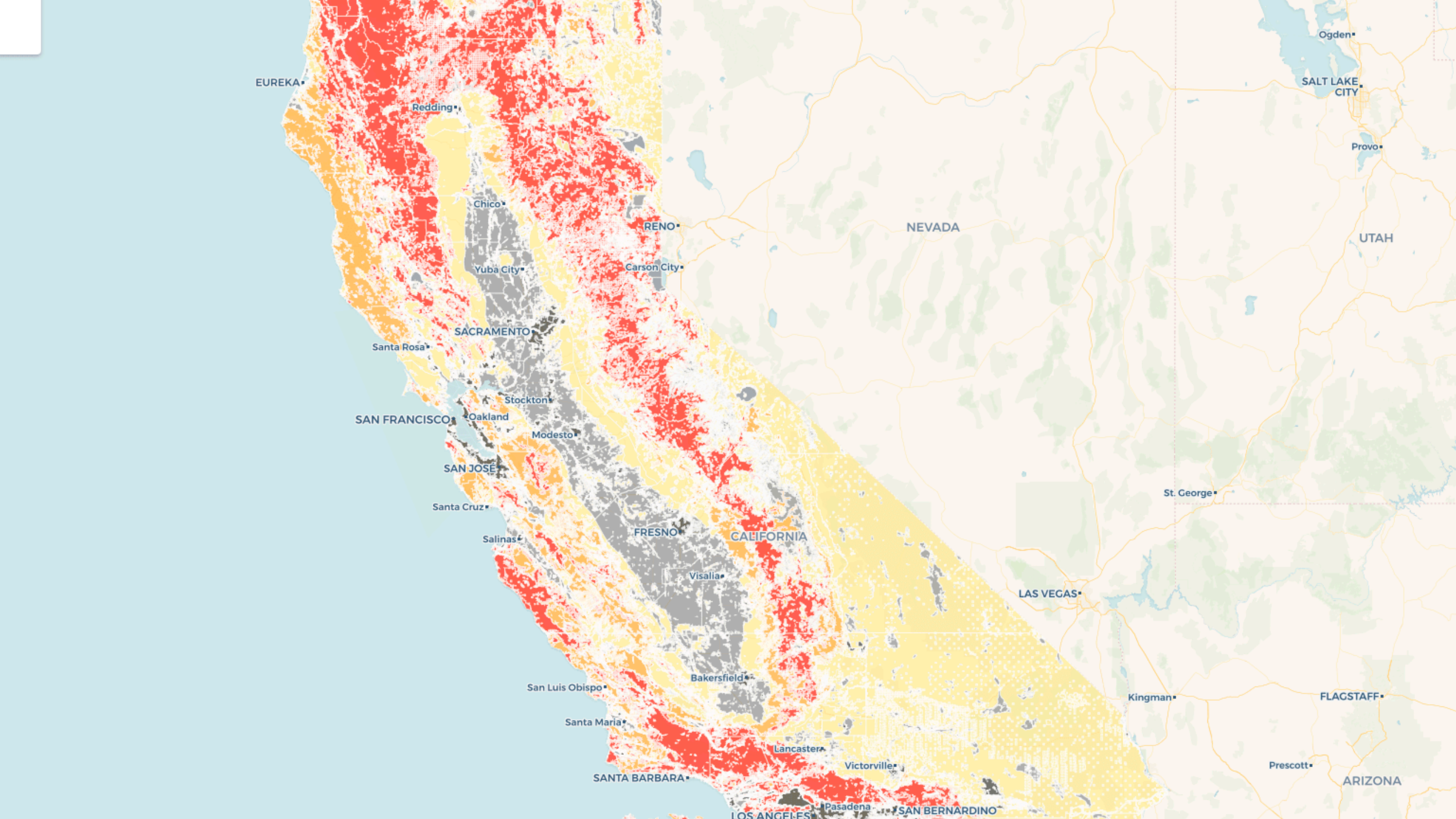
Instructor: Duncan Calllaway

GSI: Salma Elmallah

Why data, environment and society?

1. Energy consumption and other industrial activities generate pollution. Its distribution and impact are heterogeneously distributed.




EUREKA
Redding
Chico
RENO
Yuba City
Carson City
SACRAMENTO
Santa Rosa
Stockton
SAN FRANCISCO
Oakland
Modesto
SAN JOSE
Santa Cruz
Salinas
FRESNO
CALIFORNIA
Visalia
Bakersfield
San Luis Obispo
Santa Maria
Lancaster
Victorville
SANTA BARBARA
Pasadena
SAN BERNARDINO
LOS ANGELES
NEVADA
Ogden
SALT LAKE CITY
Provo
UTAH
St. George
LAS VEGAS
Kingman
FLAGSTAFF
Prescott
ARIZONA

Why data, environment and society?

2. Wildfire is caused by a range of factors and has a range of impacts.

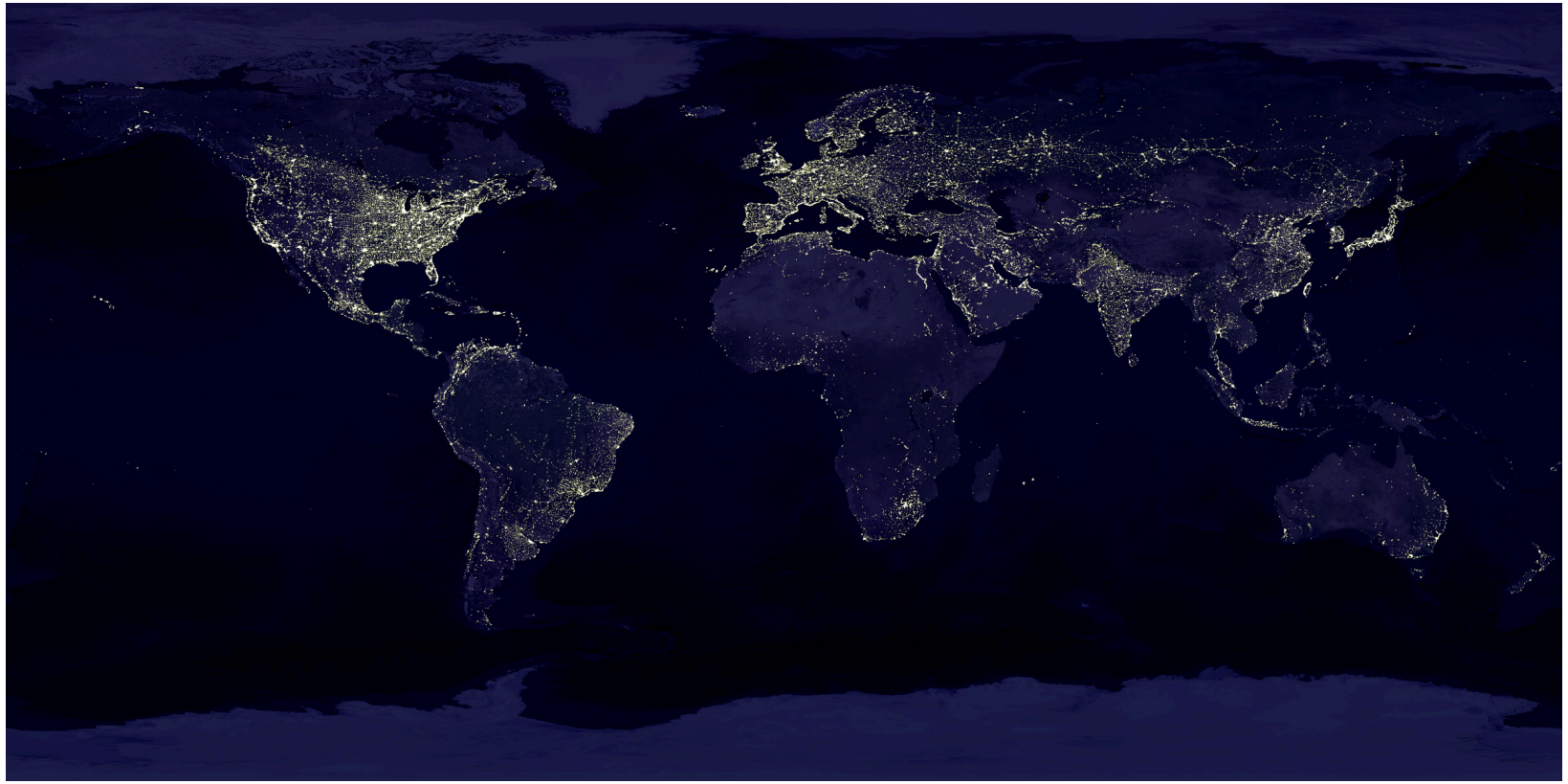
The temporality, scope and complexity of data available for analyzing historical fires and *predicting* future fires is enormous

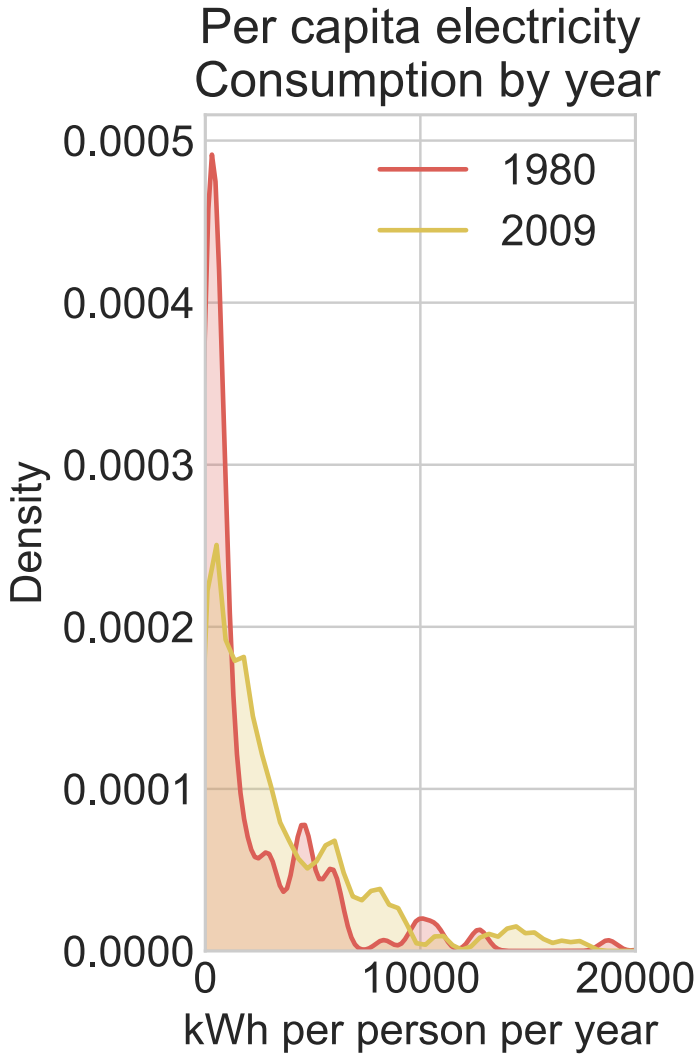
(sfgate)

(nasa)

Why data, environment and society?

3. Energy conversion and use is at the core of core of nearly all societies

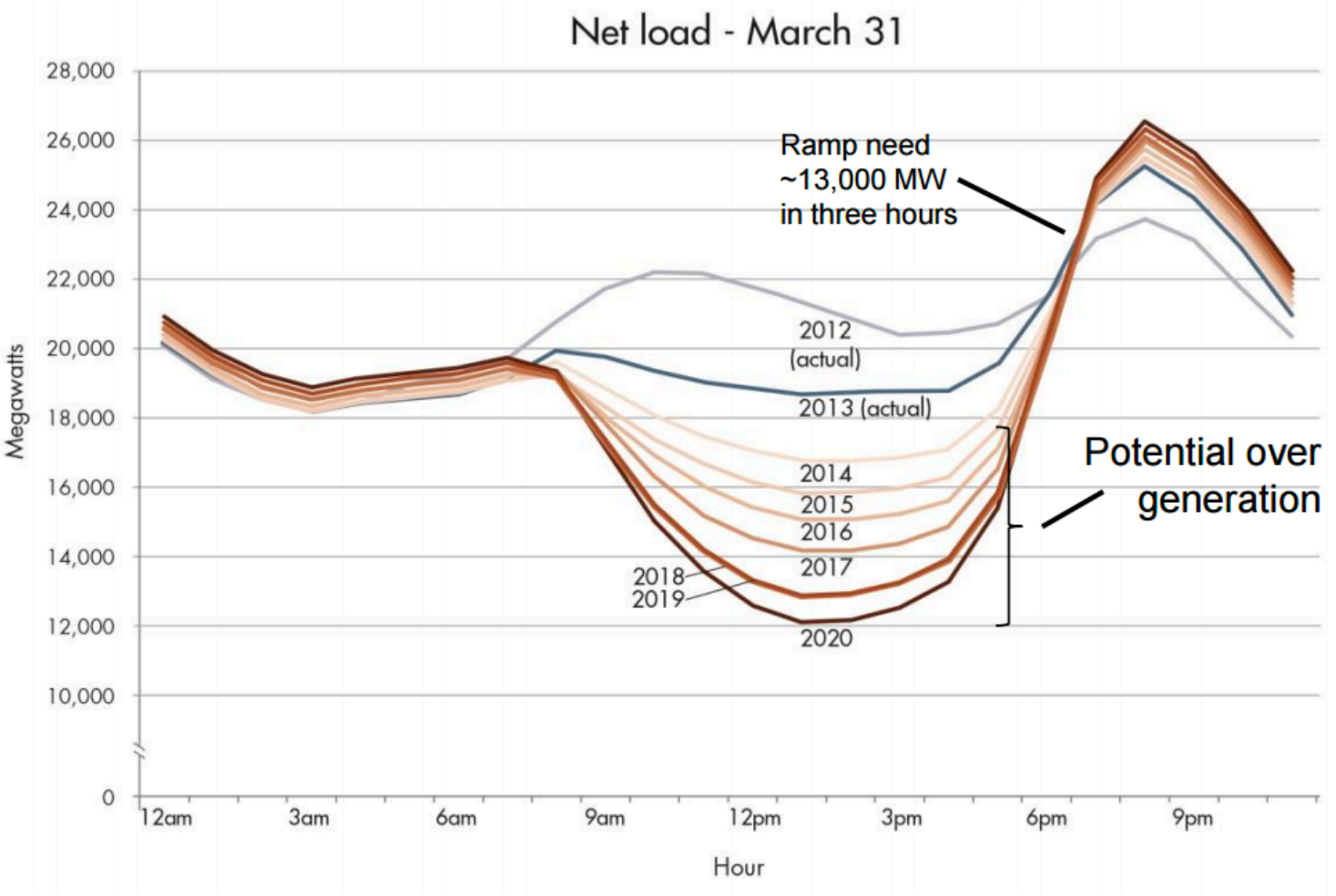




Why data, environment and society?

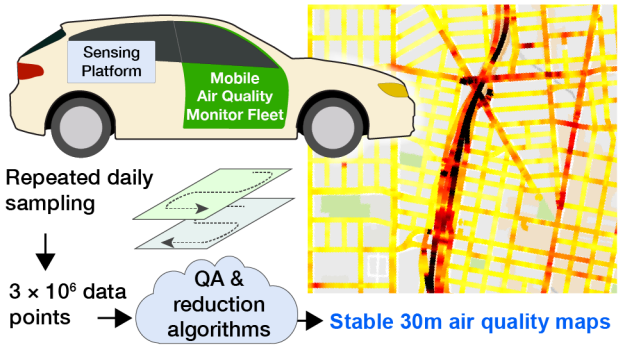
4. As renewable electricity proliferates, electric power systems must transform their operations in response

Analyzing and *predicting* energy production patterns is crucial to economical and reliable future operations



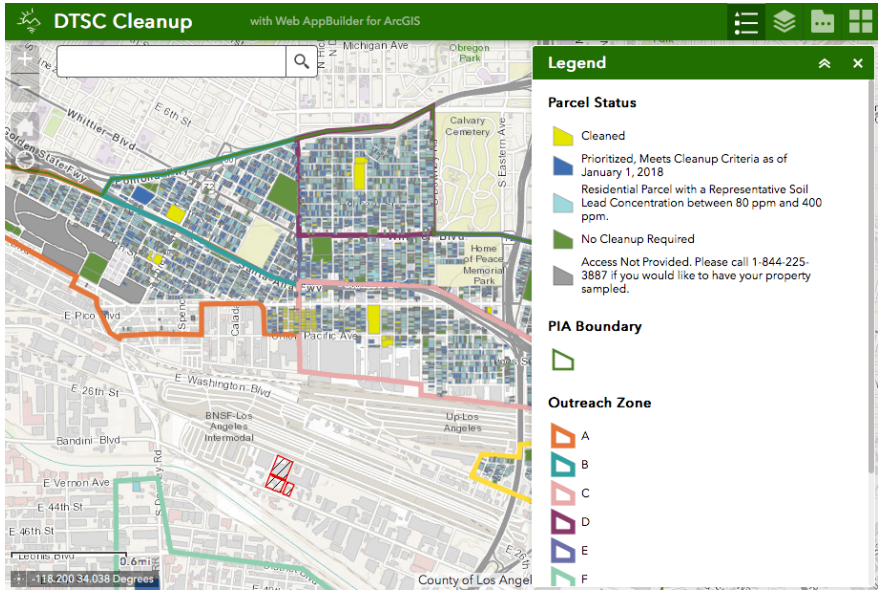
Why data, environment and society?

5. Data (and access to the data) to study these issues is growing



West Oakland Environmental Indicators Project
know which way the wind blows

Environmental Data Commons



DATAPORT
PECAN STREET

PurpleAir

OPEN DATA

CalEnviro Screen

Why data, environment and society?

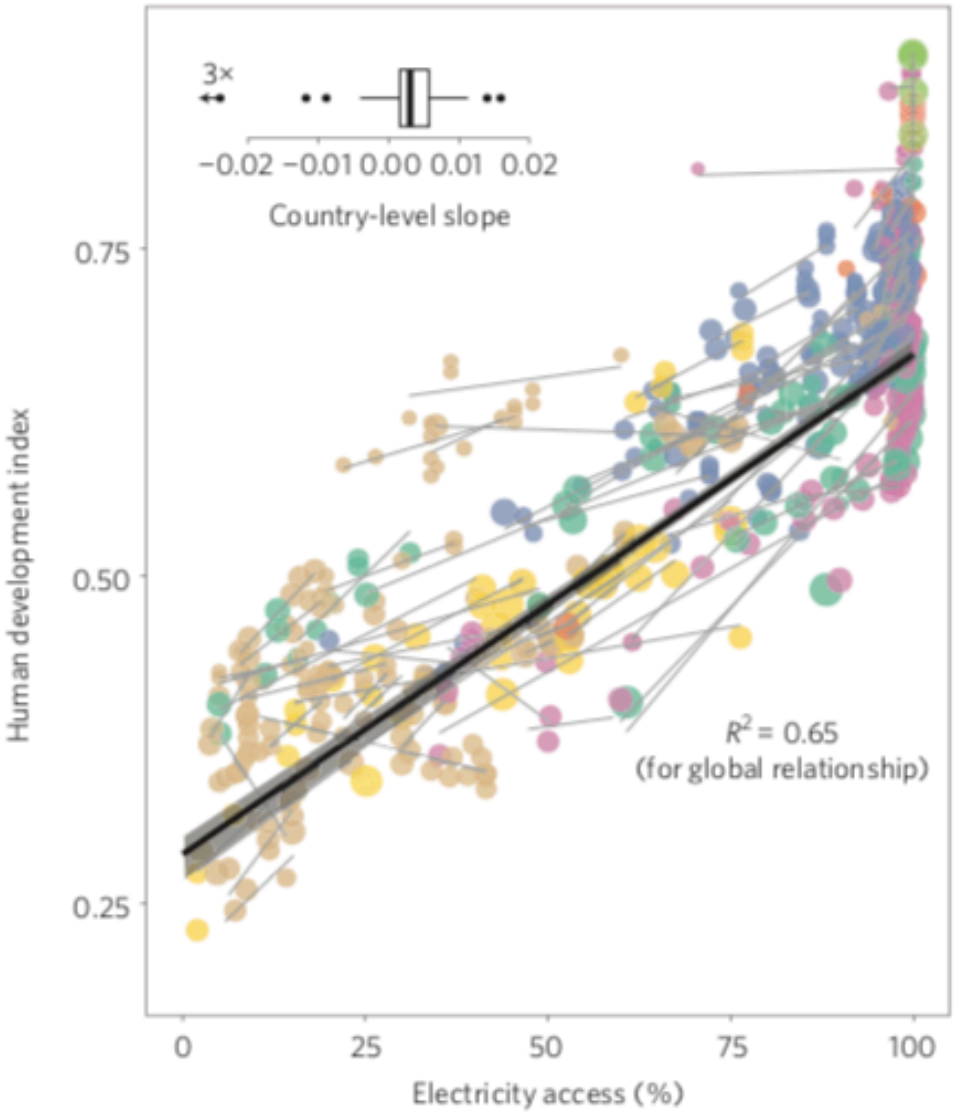
6. There is an incredible growing toolkit available to anyone with a computer…

a. New Algorithms
b. Improving computing infrastructure
c. New analysis tools: Jupyter, Python libraries

In sum: *Problems and data are abundant, the algorithms are maturing, and everything is accessible to someone with relatively little training*

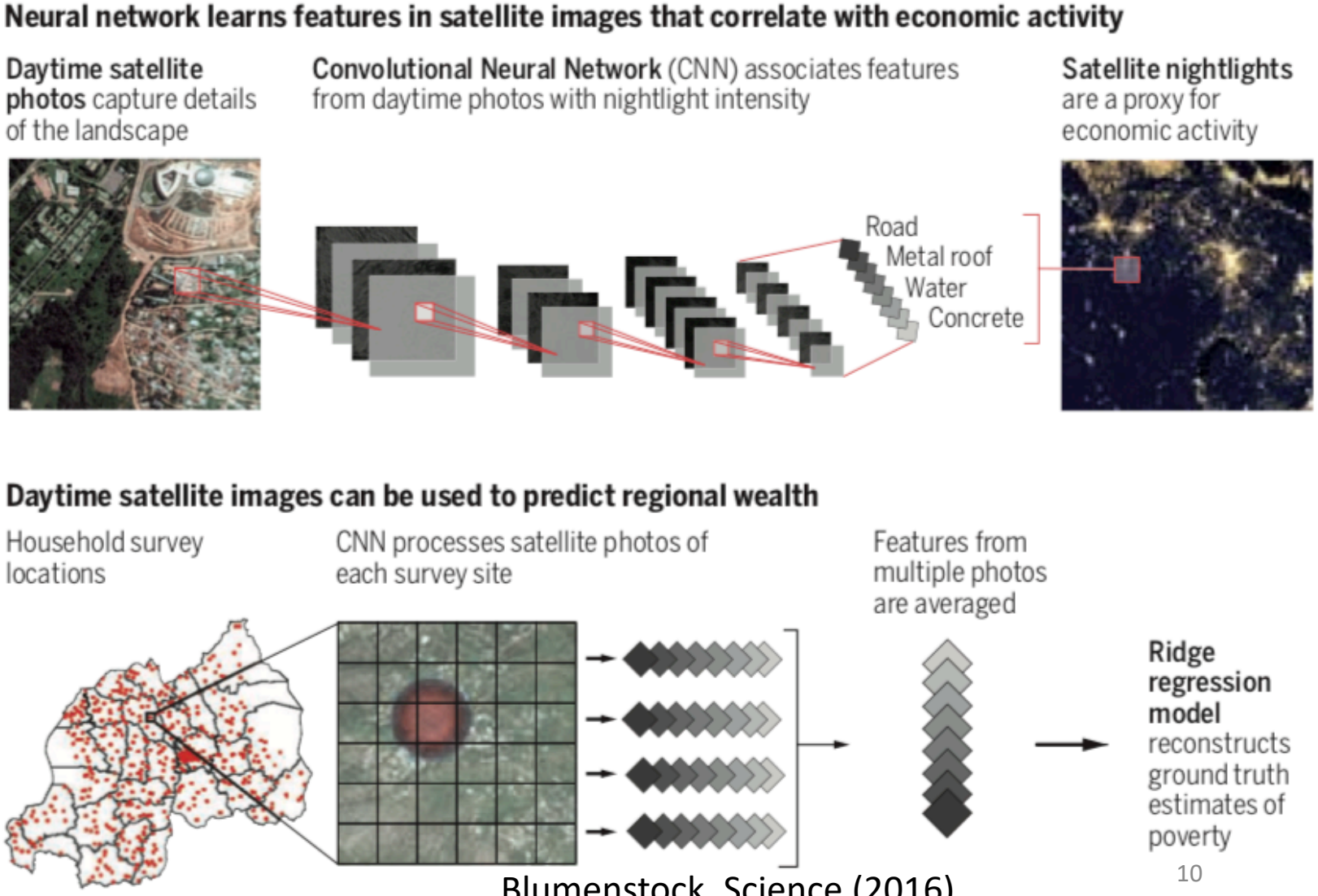
Example: Human development and energy

- From Alstone, Gershenson and Kammen, Nature Climate Change (2015)



Example: Predicting poverty

Using day and night satellite data, survey data to predict poverty in locations without surveys (Jean *et al* Science 2018)



Blumenstock, Science (2016)

What is data science?

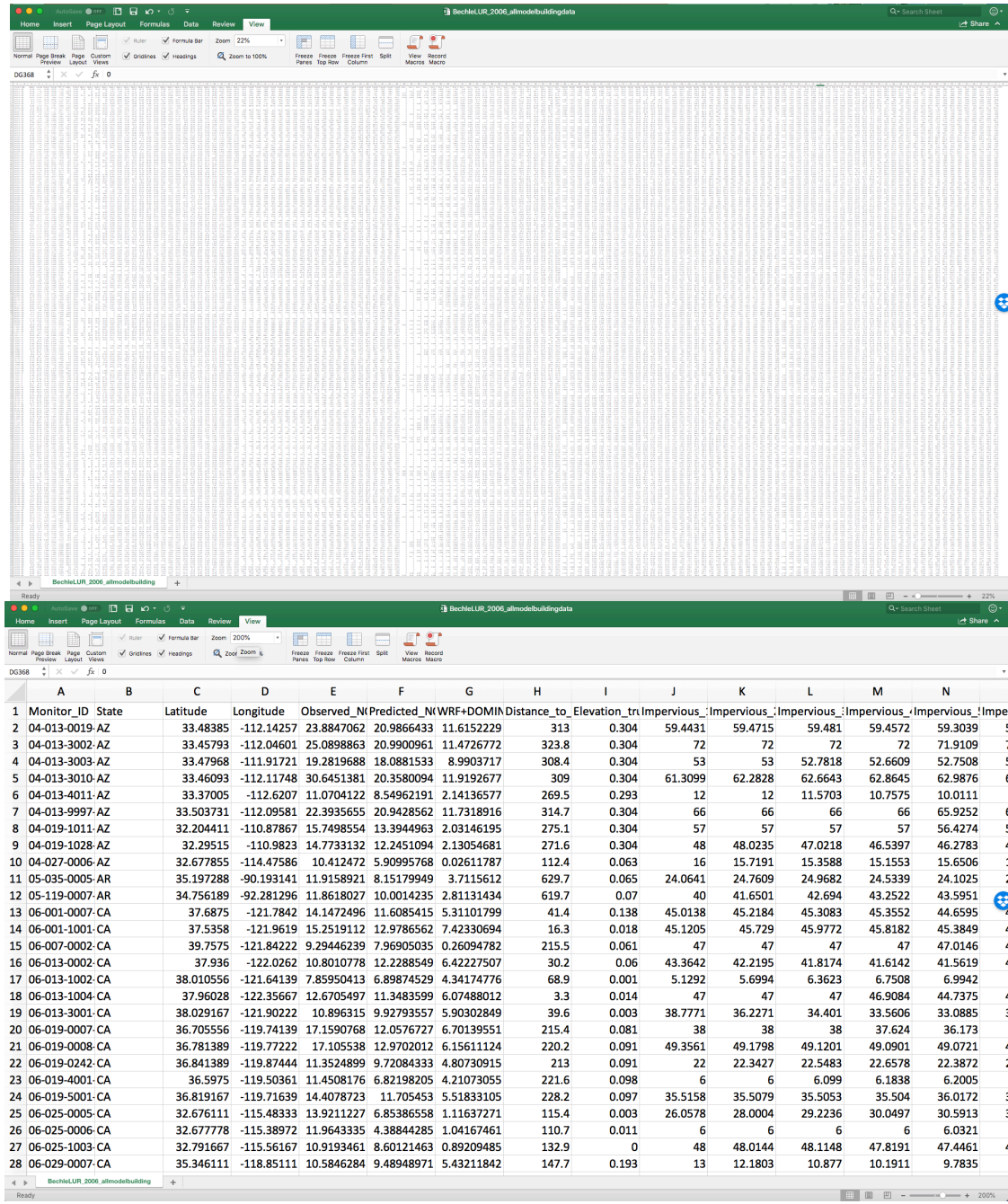
"Data science focuses on exploiting the modern deluge of data for prediction, exploration, understanding, and intervention"

--Blei and Smith (2017)

- It is a mix of data, computing and statistics.
- It is not **only** statistics: the algorithms have deep origins in computing and computer science
- It is not **only** computing: one must understand the fundamentals of the origins and characteristics of data

What does “big data” mean?

- Two characteristics
  - Data sets are large (many data points in total, or cells in a spreadsheet)
  - Data sets are *wide* (many different categories of observations, or columns in a spreadsheet)
- Many of the “data science” things we do in Python are things one could do in excel with small data sets.



What are those other buzzwords?

Artificial intelligence

Machine Learning

Statistical Learning

These refer to processes to build models that use data to learn from and make decisions about the world based on available observations.

We'll learn from a statistical learning textbook, and many of the tools we use are also known as machine learning.

Central theme of the course: using *prediction* methods in support of *resource allocation* decisions

Let’s unpack this a little further…

What can we do with data science? Rain dances vs. Umbrellas

(adapted from Kleinberg AER 2017)

(amazon.com)

- Consider two policy makers studying rain
  - One seeks to understand if a rain dance will make it rain
  - One seeks to decide whether or not to invest in umbrellas
- Both require models and data, the analyses are qualitatively different. Why?
  - "Rain dance" problems are *impact analyses.* The seek to identify a *causal* effect.
  - We call "umbrella" problems *resource allocation* problems. They seek to *predict* the future so people can decide where to allocate resources and effort.

Is prediction enough for resource allocation?

(Adapted from Susan Athey's 2017 Science article.)

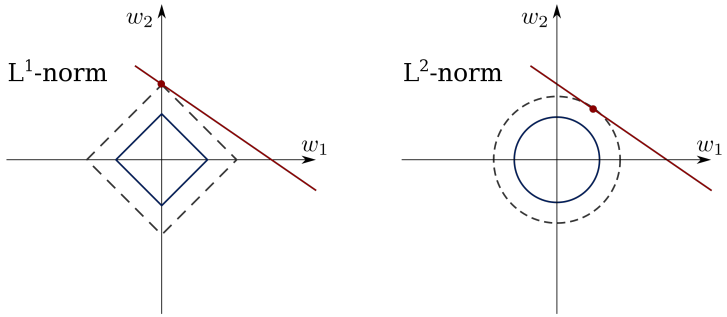
- Consider a scarce resource: Environmental health and safety inspectors at CalEPA.
- If one could predict which chemical facilities are *likely* to have an unplanned toxic release, one could dispatch inspectors to those locations
- Why might this be a bad strategy?
  - The behavior of the facilities being inspected might change during an inspection.
  - The cost to remedy identified problems might be different at different locations, which might make some problems less likely to be solved.

(mssu.edu)

The models we study in this course are good for *prediction*

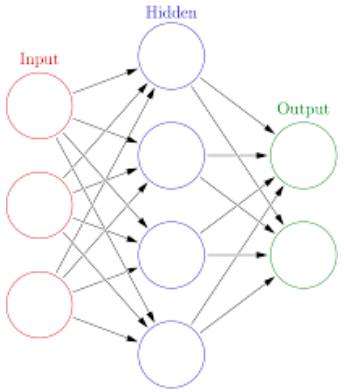
We'll focus on using them to answer resource allocation problems

Some of the specific tools we'll learn



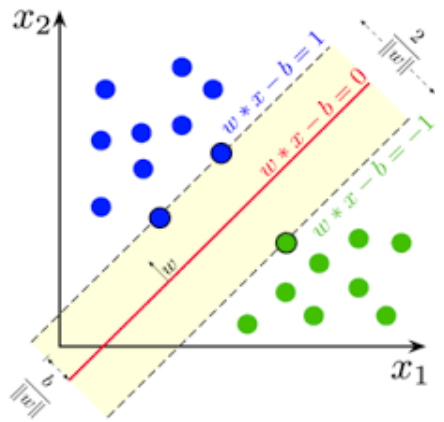
Shrinkage methods

(wikimedia)



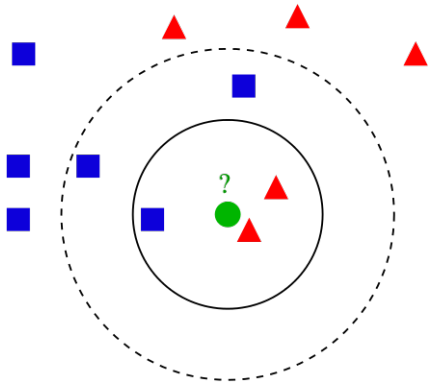
Neural networks

(wikimedia)



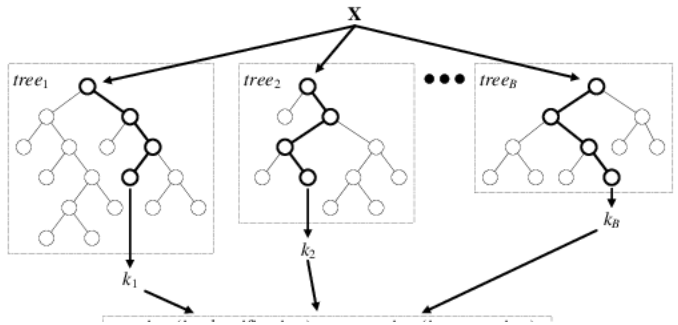
Support vector machines

(Wikimedia)



K nearest neighbors

(wikimedia)



Classification and regression trees

(researchgate.net)

- Visualization
- Exploratory data analysis
- Gradient descent
- Pandas and scikit-learn

The importance of context and narratives

From Pastor's 2007 article in Reclaiming Nature on Environmental Justice (EJ):

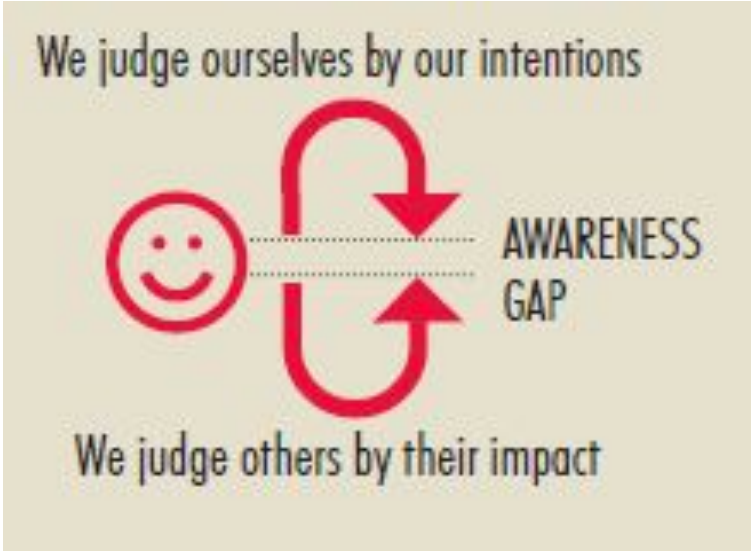
"Traditional environmentalists tend to favor 'rational' processes of debate, objective scientific research on hazards and their risks…such frameworks tend to produce negotiation between businesses and their hired experts…"

"EJ activists favor 'democratic' epistemologies in which community participation facilitates story-telling about lived experiences…in the minds of many EJ advocates, this allows for community empowerment."

Which do you prefer?

What's the awareness gap?

- For many in the room, talking about these issues is abstract, but for others it may not be
- We will discuss ideas about intent vs impact, for example in identifying environmental racism
- It is also important to think about intent vs impact in the words we use.
- …it matters little if you have the best of intentions if your actions still have a negative or unfair impact on others



calliopelearning.com

There is a human side to the algorithms themselves…

TheUpshot

Who's to Blame When Algorithms Discriminate?

A proposed rule from HUD would make it harder to hold people accountable for subtler forms of discrimination.

DEPT. OF TECHNOLOGY NOVEMBER 12, 2018 ISSUE

IN THE AGE OF A.I., IS SEEING STILL BELIEVING?

Advances in digital imagery could deepen the fake-news crisis—or help us get out of it.

By Joshua Rothman November 5, 2018

What are the course objectives?

- Teach students to build, estimate and interpret models that describe phenomena in the areas of energy & environmental decision-making.
- We will
  - Focus on analysis and prediction
  - Learn a suite of data-driven modeling approaches
  - Build the programming and computing skills to use those models (in Python and Jupyter notebooks)
  - Develop the expertise to formulate questions that are appropriate for available data and models.
- Students will leave the course as both critical consumers and responsible producers of data driven analysis.

Now let’s go into the nuts and bolts of the course

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).

How does this course fit into Berkeley's curriculum?

- The course is designed to fit into Berkeley's emerging data science curriculum. It is similar to Data 100.
  - But we will place a stronger emphasis on how to use *prediction* methods as decision-making tools in energy and environment contexts
  - It has less emphasis on web technologies, working with text, databases and statistical *inference*.
- The course satisfies:
  - Upper division domain emphasis for Data Science major
  - Engineering Elective for Energy Engineering
  - Upper division requirement for Energy and Resources Group minor

What resources are required?

- Hardware: You'll need a laptop.
  - Mac, Windows, Linux, Chromebook all ok.
  - If you don't have a laptop please see me or Salma after class.
- Software:
  - All work can be done in the cloud, on Berkeley's datahub
  - If you wish, you can install Python locally with the Anaconda distribution.
  - We'll use Python 3.x
- Internet connection: you'll be computing the cloud.
- Read-ware:
  - Introduction to Statistical Learning, available as PDF or for sale from online retailers.
  - The DS100 online textbook
  - Other reading will be distributed through github.

Note taking in lecture

- I'll post slides ahead of time in github. They'll be numbered.
- Prior students report that taking notes on paper during lecture works best
  - …then review slides after.
- If you wish you can take notes on your computer
- There will be a few lectures where you'll need your computer
- (…and you'll always need your computer in lab sections)

Github and bCourses

- bCourses
  - We'll use it for distributing homework (mainly hyperlinks to jupyter notebooks).
  - You'll upload your homework there too.
- Github: https://github.com/duncancallaway/ER131_2019
  - Readings posted here
  - Lecture notes here
  - Homework and lab files
  - All other materials

Datahub

- Berkeley's super cool jupyter notebook cloud server.
- You'll get there by using links in posted labs and homeworks.
- For example, here is hw01
  - You can do all your work here.
  - Your personal copy will get saved here.
  - When you're done, create a pdf, save it and upload it to bCourses for grading.

Datahub vs Anaconda

- If you wish, you can install a local version of Python and a starter package of data science libraries with the “Anaconda” distribution
  - See http://www.ds100.org/sp18/setup for instructions on how to do this.
- If you encounter problems running a notebook, the first thing we’ll ask is “is it working on datahub?”
  - We can’t provide support if the answer is yes…
- In other words: You don’t *have* to use Datahub. But we encourage it.

How are students evaluated?

1. Homework (20%). Submission by submitting pdf of Jupyter notebook. 10 assignments, we drop the lowest grade.
2. Labs (15%). Submission by submitting pdf of Jupyter notebook. 9 assignments, we drop the lowest grade.
3. Mid term (25%). In class on November 19
4. Final project:
   a. Poster session presentation (10%). December 17
   b. Jupyter notebook (30%). Due morning of December 18

Late policy

- You may request up to two extensions of two days over the course of the semester.
- 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
  - Brainstorm together!
  - However the work you submit must clearly be your own. 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

Final project

- We'll get started in late September
- We'll ask you to do small steps toward the final goal on your homework assigments
- The final project will be structured to include
  - A clear motivating resource allocation question
  - Data cleaning
  - Exploratory data analysis
  - A deep exploration of several prediction methods
- We'll give a few project ideas, but the goal is for you to work independently

Who are the instructors?

Duncan Callaway

Salma Elmallah

When and where?

- Lectures: Tu/Th 9:30-11, Wheeler 202
- Labs:
  - Mo 10-12, Barrows 166
  - Tu 2-4, VLSB 2032
- Office hours
  - Duncan (Barrows 325): We 9:30-10:30, by appointment; Th 11:15-12:15, open door.
  - Salma (Barrows 399): We 3-4, open door
- Labs due Mondays, pdf upload to bCourses
- HWs due Thursdays before class, pdf upload to bCourses

Next week

- No lab meetings (due to Labor Day)
  - Salma will hold extended office hours during Tuesday lab time, in VLSB 2032
- Lab 1 notebook will be distributed for you to complete on Monday
- Tuesday 9/3:
  - Reading: Science and Data Science, available on Github
- Thursday 9/5:
  - Reading: Chapter 3 from the DS100 textbook – see link in Github
- HW 1 will be released Thursday 9/5.