Stat C131A: Statistical Methods for Data Science

Fall 2025

Lecture 1: The “first” lecture !

August 28 2025

Logistics

- **Instructor:** Naveen Ashish
- nashish@berkeley.edu
- **OH:**

- **GSI:** Dylan Webb
- dylancw@berkeley.edu
- **OH:** Tue, Thu 3-5PM; Evans 434

- **Resources***
  - Statistical Methods for Data Science, *Elizabeth Purdom*
  - R for Data Science (r4ds), *Hadley Wickham et.al.,*

Today

- Logistics
- Introductory …
  - Know a bit about you
  - My trajectory, through data science
- Syllabus*
  - This course, **Statistical Methods** *for* **Data Science**
- Short story
- Final project

you

You

- Background
- Goals
  - Especially this course
- Aspirations

my trajectory, through data science


NASA
Ames Research Center


it2 Calit2


UCI
School of
Medicine


Keck School
of Medicine
of USC


Fred Hutch
Cancer Center


UNIVERSITY of
WASHINGTON


BERKELEY


InferLink


abzooba
analytics simplified


dt dropthought


IntellikGraph


2005
2013
2016
2018
Now

- Data-driven solutions *for* Emergency Response
- Data … *for* Neuroscience
- Data …. *for* Clinical Translational Science

2005 2013 2016 2018 Now

- More Neuroscience, Data Science for Neuroscience

2005 2013 2016 2018 Now

- Data Science for Cancer Research
  - Bacterial growth analysis
  - Microbiome
  - Deep-learning based yet explainable (breast cancer) MRI image analysis
  - Women's Health Initiative (WHI) data mining


2005 2013 2016 2018 Now

- Mainly federal funded R&D innovation research contracts
  - ISR: Intelligence, Surveillance & Reconnaissance
  - Data-driven Emergency Response
    - Especially extreme events: natural, man-made
  - More generally, systems that help with
    - Situational awareness
      - Cognitive overload
    - Decision making (Course-of-Action)

2005 2013 2016 2018 Now

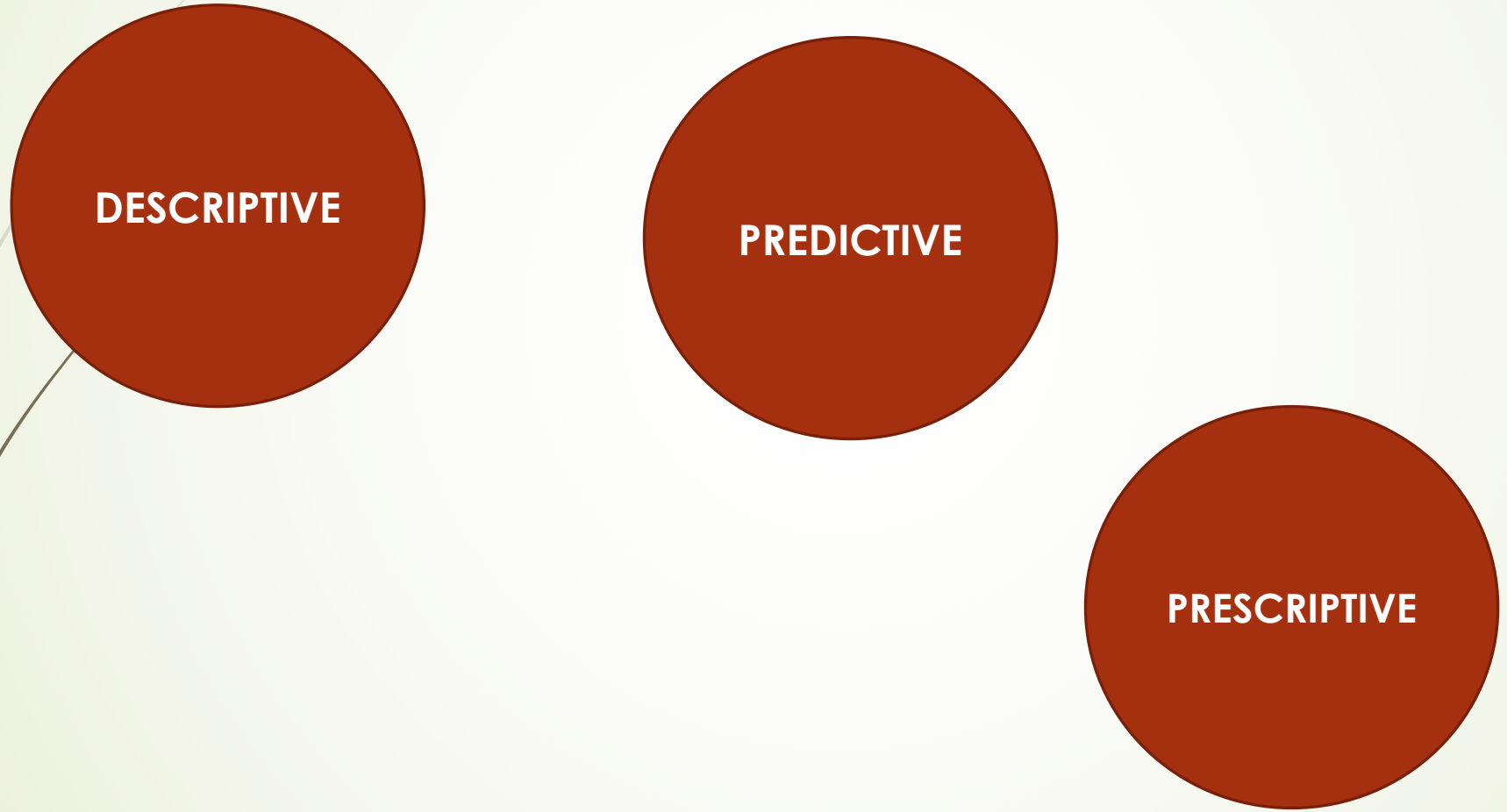
Disciplines, Problems, Customers

- **Data science for**
  - Neuroscience research
  - Clinical researcher
  - Emergency response
  - Cancer research
  - Materials science (new material design)
  - Climate science
  - Intelligence analysis (ISR)
  - …
- **Consulting for the private sector**
  - Market research - persona analytics
  - Predictive analytics for retail
  - Insurance: assessing particular chemicals for risk of human harm
    - Based on scientific literature
  - ….
- **Elements**
  - Decisions: what to do
  - Automated actions: what to ignore, cognitive overload
  - Deal with uncertainty
  - Establishing the limits of what we can do

In summary

- Encountered and employed statistical methods **through the lens of a** data scientist
- Help either
  - Humans
  - Automation
- Statistical methods have/continue to be relevant in about every problem !

Data Science **Pillars**



About

- **1) Summarization & Visualization (EDA)**
  - **Exploratory** Data Analysis **!**
  - Take a messy table of numbers and make it **legible**
    - what's typical, how spread out, any weird cases, how groups differ, …
  - Good pictures help you **think** ☺
    - and, often, change what model you choose later
- **2) Probability & Distributions**
  - **Distribution** tells you which values are common/rare
  - Lets you turn "there's a pattern here" into "how likely is that by chance?"
  - Landmark, and practically *very* applicable, discoveries: The **Central Limit Theorem**
  - Sampling, Errors

About

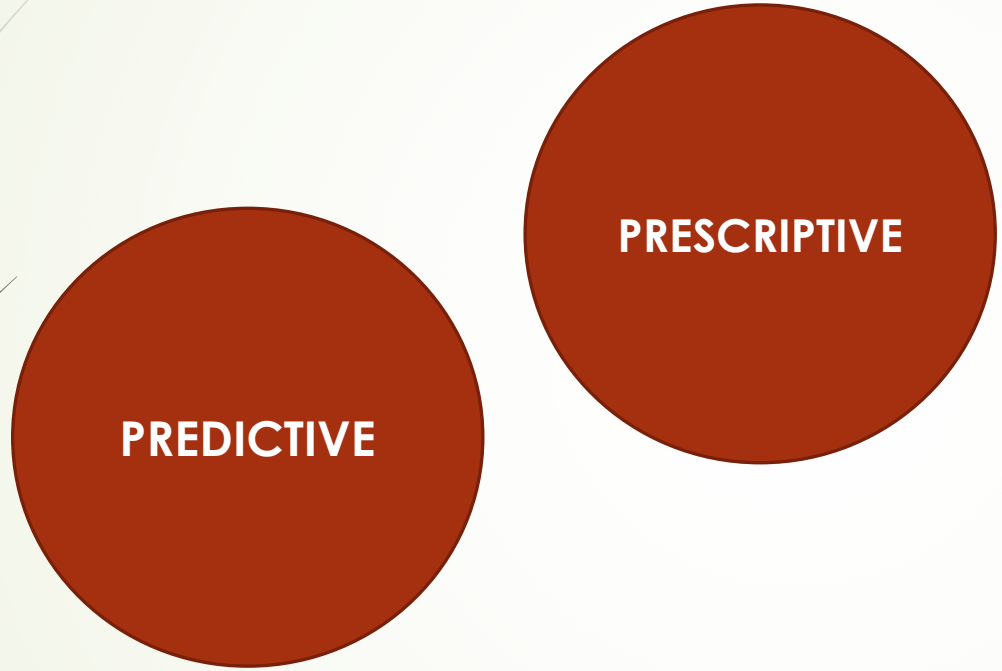
- **4) Estimation, Confidence Intervals**
  - Give a **range** that's honest about uncertainty
  - Confidence intervals: Size, Uncertainty
- **5) Hypothesis Testing**
  - "If there were **no real difference**, how surprising is the gap we see?"
    - null/alternative, p-value, Type I/II error, t-test, ANOVA (many groups), multiple comparisons, ANOVA
- **6) Multiple Linear Regression (explanation & adjustment)**
  - Predict a numeric outcome from several inputs and read off **how much each input moves the outcome**, holding others steady.
  - Separates **signal from confounding** and gives **interpretable** effects

About

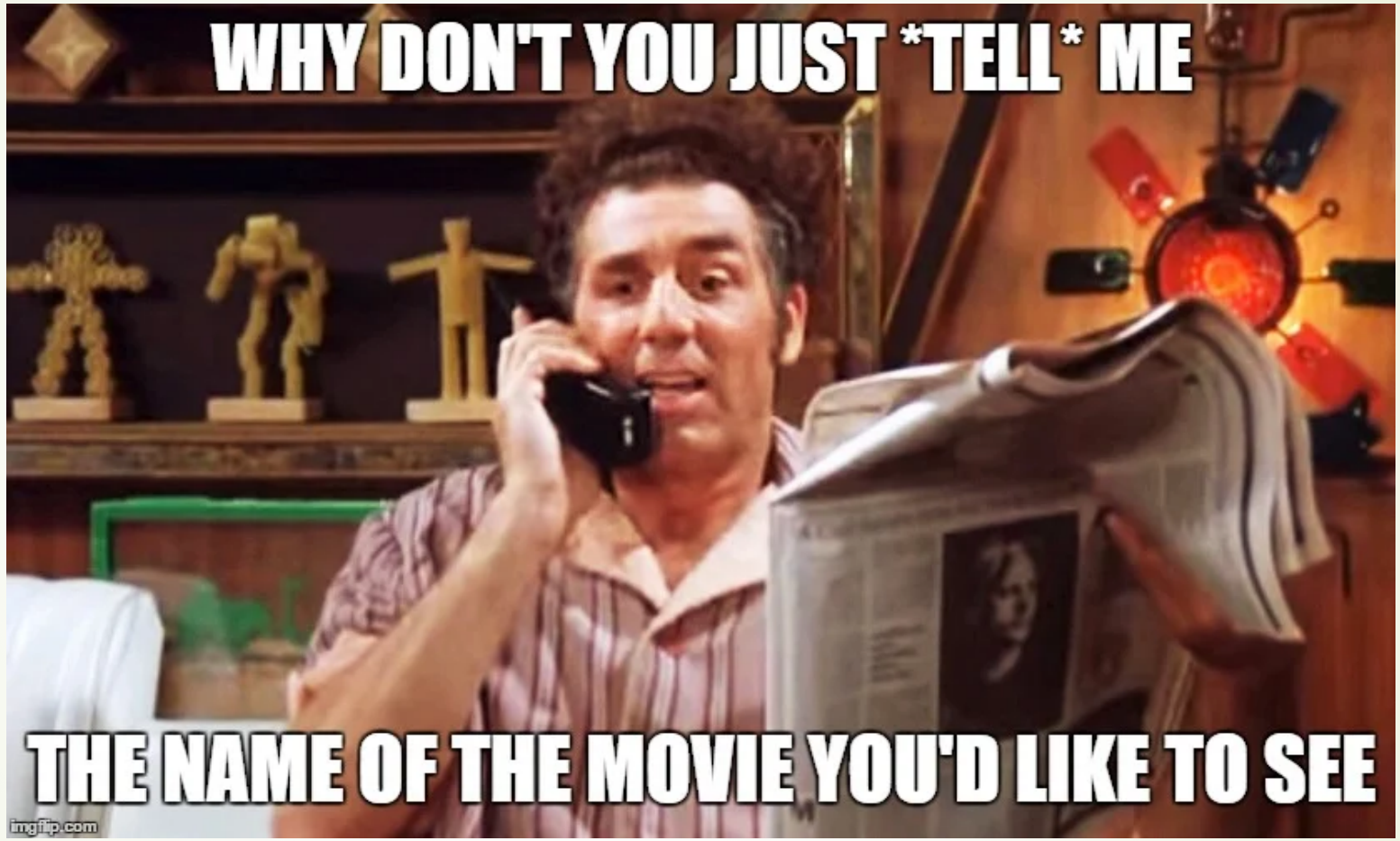
- **7) Logistic Regression (classification)**
  - When the outcome is **yes/no**, estimate the **probability** of "yes" from inputs.
  - Turns probabilities into decisions while keeping interpretation.
- **8) Trees & Random Forests (nonlinear patterns)**
  - Let the data **split** itself into meaningful regions ("if BMI>28 and age<40, predict higher risk")
    - Combine many trees (**random forest**) for stability.
  - Captures **thresholds and interactions** you might miss with a straight line.
- **9) Model Selection & Validation**
  - Don't trust a model because it fits the data you used to train it
    - **Hold out** some data or use **cross-validation** to check how it does on new cases (train/validation/test split, k-fold CV, overfitting, AIC/BIC (fit vs simplicity).
  - Ensures your results **generalize**
    - Beyond your sample

About

- **11) Reproducible Statistical Computing (in R)**
  - Analyses should be **scripted** so anyone can re-run them and get the same results.
  - Science (and your grade!) needs **repeatable** results.
- **12) Communication for Decisions**
  - Tell a clear story: **What changed? By how much? How sure are we? What should we do?**
  - **Persuasive:** The goal isn't just to compute, it's to **inform a choice**.



the problem is…



what IS the problem !

another story

Operation Neptune Spear: Decision

- May 2 2011 Raid: GO decision not a unanimous one, but taken by President Obama
  - Post 40 intelligence reviews and multiple meetings
- March 2011 meeting convened by the President to *lay out the **(un)certainty** concerning bin Laden's location as **clearly** as possible*
  - CIA Team Leader: as high as 95%
  - Deputy Director of Intelligence Michael Morell offered a figure of 60%
  - Most advisors seemed to place their confidence level at about 80%, though some as low as 30%
  - Red-Team: 40-60%
- The President's predicament: *"(the discussion) provided not more certainty, but more confusion"*, and *"the advisors were offering probabilities that disguised uncertainty as opposed to actually providing (you) with more useful information"*

Friedman, J. A., & Zeckhauser, R. (2015). Handling and mishandling estimative probability: Likelihood, confidence, and the search for Bin Laden. *Intelligence and National Security*, *30*(1), 77-99.

Peter Bergen. Manhunt: The Ten-Year Search for Bin Laden--from 9/11 to Abbottabad

Conditional Probabilities !

H: OBL is in target house H

$\neg H$: OBL is not in target house H

U_1: Complete absence of electronic communication
U_2: Simultaneous trusted courier awareness denials

E_1: AAAK traced to house H

E_2: House H has multiple privacy indicators

E_3: The "pacer" does not leave the house

E_4: 911 detainees interrogated are known to have lied

E_5: Biogeographic theory based places an 84.5% probability of OBL residing in simila town in the region

V: Target is as high value as OBL
T: AAAK is the trusted courier
R: "Pacer" is OBL

$$\boldsymbol{P(H \mid E_1, E_2, E_3, E_4, E_5, V, T, R)}$$

project proposal

The **DATA**

- National Health and Nutrition Examination Survey (NHANES), run by CDC/NCHS
  - Nationally representative of the civilian, non-institutionalized U.S. population (all ages)
  - Continuous program released in 2-year cycles (e.g., 2017–2018), with minor changes across cycles
  - Two parts: household interview (health, diet, behaviors) + standardized **MEC** exam with physical measures and laboratory tests
- Person key **SEQN**
  - Most components are one row per person; some are multi-row per person
- Public-use data are free (SAS XPT + codebooks); in R
  - Many pull tables with the **nhanesA** package
- Capabilities: national estimates (means/prevalence), subgroup comparisons, associations between diet, measures, and labs

NHANES

- National Health and Nutrition Examination Survey (NHANES), run by CDC/NCHS
- Person-level microdata with a unique participant ID (**SEQN**)
  - interview, exam, diet, and lab files by SEQN.
- Anonymized public files (no names/addresses), detailed geography suppressed; some variables top-coded to protect privacy.
- Age coverage from infants to older adults; very old ages reported as **80+** in public data.
- About **10,000 participants per 2-year cycle** (≈5,000 examined per year), sampled to be nationally representative with oversampling of selected groups.
- Components include: household interview (demographics, health, behaviors), **24-hour dietary recalls** (Day 1 for all; Day 2 for a subset), mobile exam center measurements (e.g., height/weight/BMI, blood pressure), and laboratory tests (e.g., lipids, glucose/HbA1c, hs-CRP, vitamins, metals).
- Most component tables are **one row per person**
- A typical merged analytic file spans **hundreds** of variables per person; pulling many questionnaires, labs, and diet totals can yield **1,000+** columns.
- Public releases are **SAS XPT** files with codebooks; easy to load in R (commonly via the nhanesA package).

Problem

- Each group investigates how a **dietary exposure** (you choose) relates to a **health outcome** (you choose) in NHANES adults
- Using
  - design-based analysis
  - rich EDA
  - a defensible model with diagnostics.

For instance

- **A) Exposure (choose one)**
  - **Fiber density** (g / 1000 kcal)
  - **Added/total sugar density** (g / 1000 kcal, or % energy)
  - **Saturated fat density** (g / 1000 kcal, or % energy)
  - **Sodium density** (mg / 1000 kcal)
  - **Caffeine intake** (mg/day), optionally as density
- **B) Outcome (choose one)**
  - **Continuous lab/measure:** log hs-CRP (inflammation), SBP/DBP mean (blood pressure), HDL/total cholesterol, BMI.
  - **Binary clinical cut:** high hs-CRP (e.g., ≥3 mg/L), hypertension (measured or self-report), obesity (BMI≥30).

For instance

- **Model family (choose one primary; you may show a secondary for comparison)**
  - **GLM (design-based)**
    - Continuous outcome: linear model (often on log scale).
    - Binary outcome: logistic model.
    - Allow **nonlinearity** in diet via **splines/GAM** or **piecewise**.
  - **Quantile regression (design-aware)** for median effects when outcomes are skewed.
  - **Tree/forest** (predictive comparison) with honest discussion of interpretability vs fit.
- All teams must define the **NHANES survey design** and justify **covariates** (e.g., age, sex, race/ethnicity, BMI, smoking, *total energy* if not using densities). Use the weight tied to your *most restrictive* component.

Evaluation

- **A) Study Design & Data Foundations**
  - Problem framing & design
  - Data handling & reproducibility
- **B) Exploratory Insight & Inference**
  - Visualization for thinking (EDA)
  - Estimation & testing (effect size + CI + p-value)
- **C) Modeling Rigor**
  - Modeling & diagnostics (form justified, checks shown)
  - Validation & robustness (holdout/CV, sensitivity)
- **D) Communication & Impact**
  - Interpretation & decision relevance
  - Presentation
  - Written report

Class

- Attendance ?
  - Discussion based
- Project

???
questions