

# Statistics 154, Spring 2019

## Modern Statistical Prediction and Machine Learning

### Lecture 2: Problem formulation and experimental design

Instructor: Bin Yu

([binyu@berkeley.edu](mailto:binyu@berkeley.edu)); office hours: **Tu: 9:30-10:30 am; Wed: 9:00-10:00 am**  
**office: 409 Evans**

GSIs: Yuansi Chen (Mon: 10-12; 12-2); Raaz Dwivedi (Mon: 2-4; 4-6)

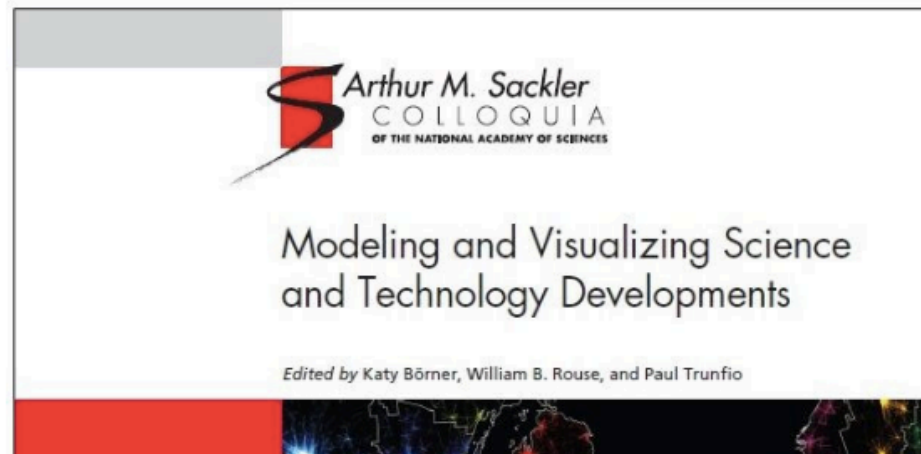
[yuansi.chen@berkeley.edu](mailto:yuansi.chen@berkeley.edu); [raaz.rsk@berkeley.edu](mailto:raaz.rsk@berkeley.edu)

(office hours to be announced)

# Proc. of National Academy of Sciences (PNAS)



*PNAS invites you to browse the articles from the Sackler Colloquium*  
**Modeling and Visualizing Science and Technology Developments**

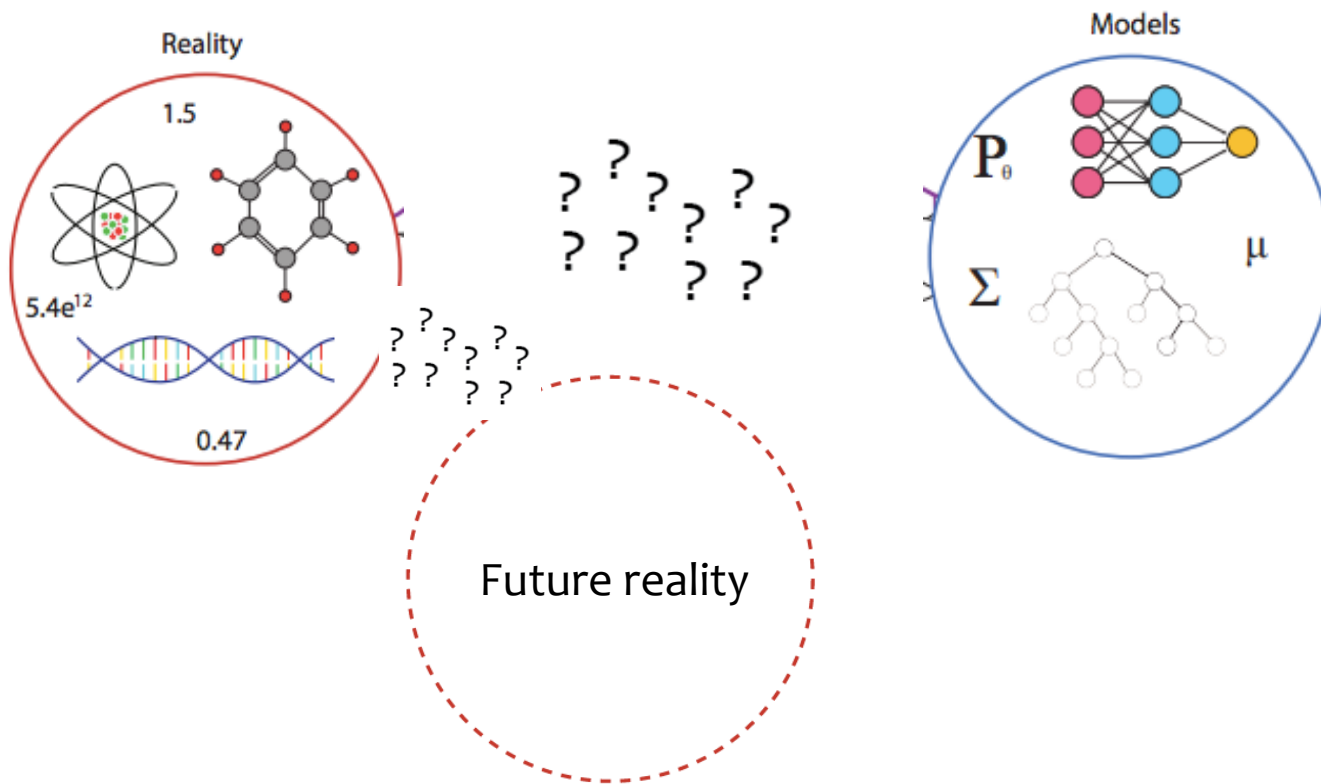




Human survival depends on the ability to predict future outcomes and make informed decisions. Such predictions increasingly rely on models of the structure and dynamics of natural, technological, and social systems. The 12 articles in the Sackler Colloquium on Modeling and Visualizing Science and Technology Developments examine the ways in which researchers analyze and present data, particularly in large datasets, to unravel the dynamics of a wide array of complex systems, as diverse as the career trajectories of scientific researchers and the flow of urban traffic.

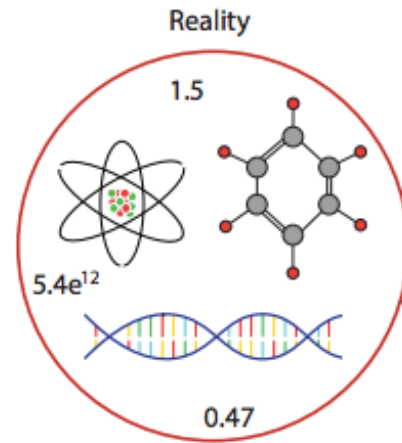
## Recap: Why are we here?

To solve prediction problems in real world  
by connecting the two solid circles below  
in a justifiable way to say things about the third circle

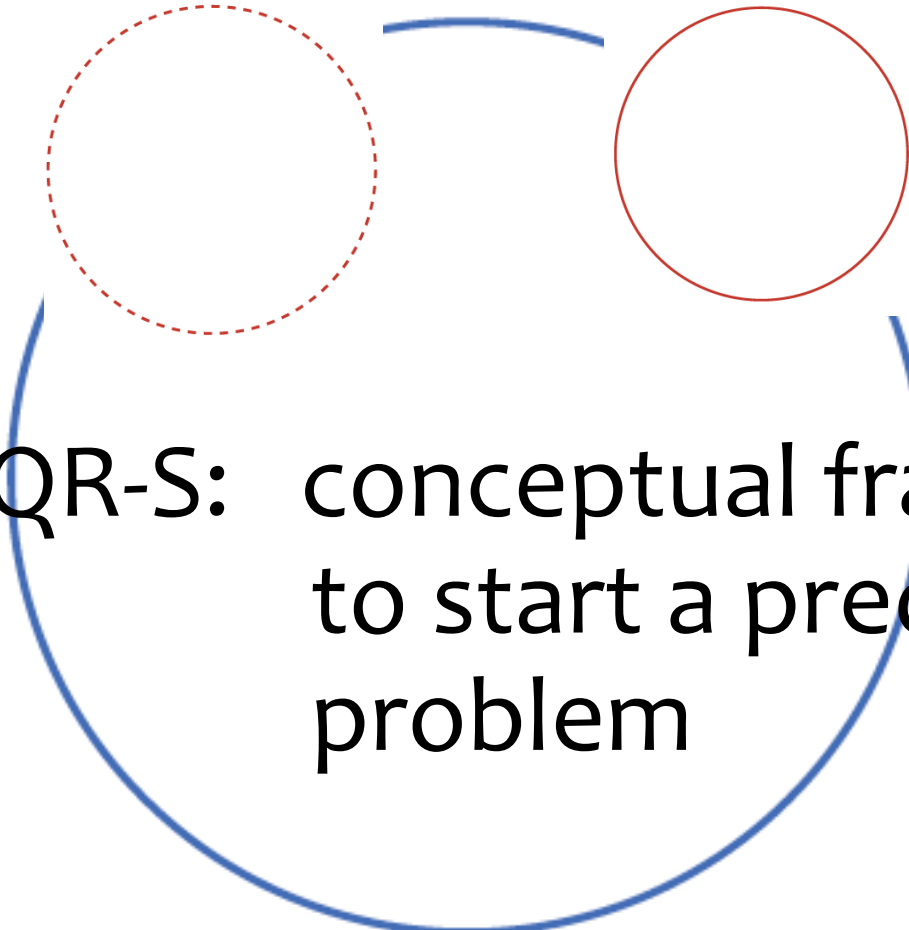


# Recap: What will you learn?

Future reality



- Problem formulation
- Data collection, data cleaning
- EDA (exploratory data analysis, visualization)
- Unsupervised learning (e.g. PCA, clustering)
- Supervised learning (LS, regularized LS, Kernel regression, logistic regression, Support Vector Machines (SVMs), Nearest Neighbor (NN), Decision trees, Random Forests, Deep Learning)
- Data results, validation, conclusions



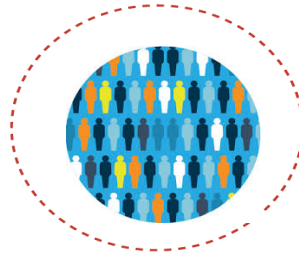
I.PQR-S: conceptual framing  
to start a prediction  
problem

or a check-list

# PQR-S helps you think straight!

- Population

- Question



- **Representative** data collection (data neutral, fairness)  
is  similar to  ?

- - (to be filled in throughout the course )

- Scrutinize or validate data results



P. “Population” of relevance -- this is a **generalization** of “population” in a traditional statistics class

- Population is the relevant group of people (objects, units) that the prediction will be applied to

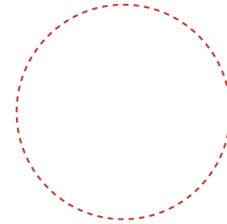


- Write the population down as a record for any ML project



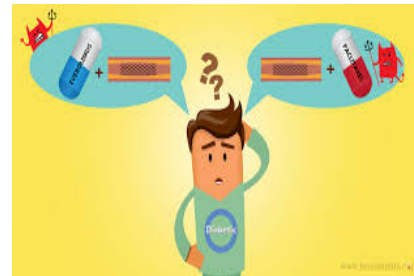
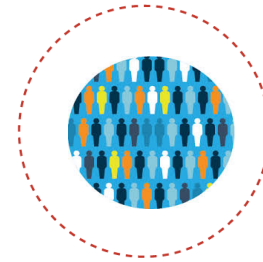
# “Population” for the 2016 election

- Oracle (future) population  
all the votes casted on election day
- Short of magic, we want to predict or guess these votes ahead of time
- How to go about this?



# Q ? Question, question, question

- Domain prediction question to answer
- Examples
  - Why didn't the polls predict well?
  - Is smoking predictive of lung cancer?
  - Does BrainPlus IQ change humanity?

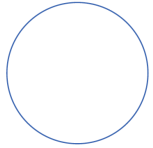


What would you do if you are the pollster?

# What would you do if you are the pollster?

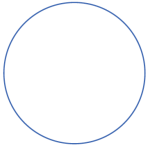
- What and how to collect data to answer the question?

# Survey design: what questions to ask?



- We need to translate the questions into a more precise question:
- Did the poll have the resources (energy, expertise, relevant data) to collect data and do the prediction?

# Survey design: what questions to ask?



- We need to translate the questions into a more precise question:

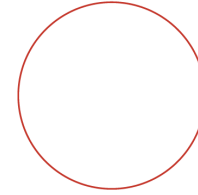
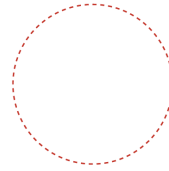
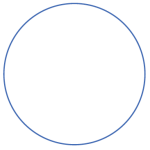
Are you a likely voter

Are you an undecided voter

Who would you vote if you vote today?

- Did Gallup Poll have the resources (energy, expertise, relevant data) to do the prediction?

# Feasibility of question

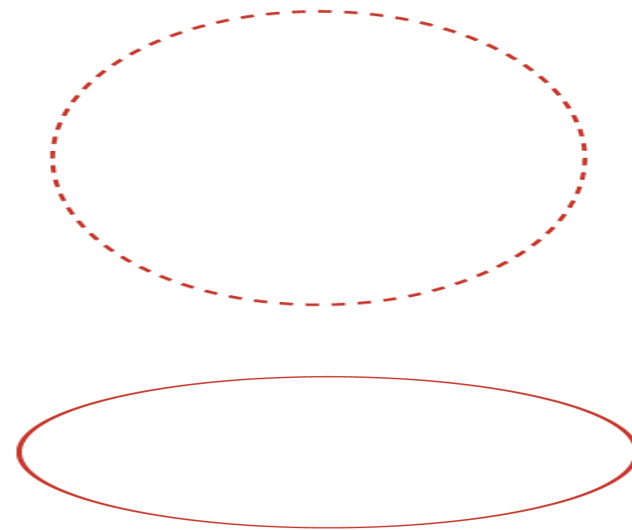
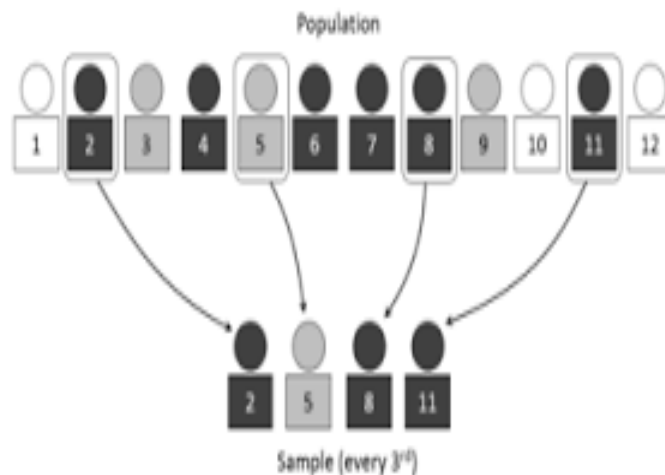


- Is the question feasible to answer using given data?

## R. Representative data collection for 2016 election?

- Ideally, we want the poll data on Oct. 30 to be “representative” of the voter population on Nov. 8, or **data neutral**

Problem: Population does not exist on Oct. 30, so can not be randomly sampled with all the money in the world



Domain knowledge can come in to help, but we can never be 100% sure about something of the future.



# An approach used in the past

- Convenience sample

It is a judgment call whether the poll data is useful for predicting election results, and the following helps:

- There is always prior knowledge or data
- Qualitative: literature, human co-workers
- Quantitative: previous data, new pilot study

## Difficulties due to two sources of uncertainty on Nov. 8 that can't be helped by careful design

- Who are going to vote on election day?
- How is a voter going to vote?
- We can't ask all voters, even if we could, people change their minds – for 2016, about 12% undecided voters, compared with 4-5% in the past

# Assumptions in the poll data collection to ensure R

1. Votes on Oct. 30 are the same as on the election day: People are asked what they would vote if they are voting today –
2. People are telling the truth
3. Undecided voters vote similarly as decided voters
4. The polled group is representative of the voter population

## How to make the sample representative or data neutral?

- A whole field of statistics, **Survey Sampling**, has been devoted to this. It recommends random sampling of different kinds. (Take **Stat 152**)
- Simplest: **simple random sampling (SRS)** without (or with) replacement – putting all entities of the population in a hat and randomly drawing one by one

# One particular poll: Gallup Poll

- Simplest form: 1000 SRS samples from numbers of phones (landline and cell phones) (one landline can map to many voters).
- What is the population? Is it the same as the voter population?



- What if people without phones voted very differently from people with phones?

# Gallup Poll: some quick calculations

- Suppose Gallup poll is an SRS of the voter population on election day:



=



, and previous assumptions

139 million votes, 68 for Trump, 71 for Clinton – less than 2% diff.

- Margin of error is about 3%, one Gallup Poll is not enough to predict such a close race **even if all the assumptions hold and they do not.**

# Gallup Poll: some quick calculations

- Suppose Gallup poll is an SRS of the voter population on election day:



=



, and previous assumptions

139 million votes, 68 for Trump, 71 for Clinton – less than 2% diff.

- Margin of error is about 3%, one Gallup Poll is not enough to predict such a close race **even if all the assumptions hold and they do not. (Actually, Gallup poll sat out for 2016 election)**



# Scrutinizing Gallup polls by comparing with actual election results in the past



## S: 2012 Gallup poll

- Final prediction: 49% Romney vs. 48% Obama
- Actual: 48% Romney vs. 51% Obama

# Gallup's review on the factors below

<https://news.gallup.com/poll/162887/gallup-2012-presidential-election-polling-review.aspx>

- A. Survey and Sample Design Factors ..
  - 1. Tracking Design
  - 2. RDD List-Assisted Landline vs. Listed Landline Samples
  
- B. Survey Field Management
  - 3. The Gallup Name
  - 4. Race of Interviewer
  - 5. Gender of Interviewer
  - 6. Neutral Probing of “Don’t Know” and “Refused” Responses
  - 7. Geographic Distribution of Interviews
  - 8. Interview Completion Time
  - 9. Cellphone and Landline Phone Distribution
  
- C. Data Handling
  - 10. Measuring and Weighting Race
  - 11. Handling of Third-Party Candidates
  - 12. Candidate Name Order in Question
  - 13. Likely Voter Estimating

# Scrutinizing data results in general

- Prediction on test data
- Stability analysis
- Post-hoc EDA or visualization
- Domain knowledge verification
- ...

## Longer time scale

- Down-stream consequences
- Further studies
- ...



## Recall “danger zone”

- Danger zone: algorithms applied to domain problems without understanding of statistical concepts/issues such as population, representative data collection, and uncertainty, ...
- For election and personalized medicine, there is NO averaging effect to aid algorithms as in IT applications where ML algorithms have been traditionally successful.

One has to get representative samples AND try to use other information to reduce variability for a one-time shot!

# Clinton Campaign fell into danger zone?

 POLITICO

Magazine ▾ Trump Presidency Policy ▾ PRO 🔍 👤 🌐 U.S. Edition ▾

2016

## How Clinton lost Michigan — and blew the election

Across battlegrounds, Democrats blame HQ's stubborn commitment to a one-size-fits-all strategy.

By EDWARD-ISAAC DOVERE | 12/14/16 05:08 AM EST

<http://www.politico.com/story/2016/12/michigan-hillary-clinton-trump-232547>

“In results that narrow, Clinton’s loss could be attributed to any number of factors — FBI Director Jim Comey’s letter shifting late deciders, the lack of a compelling economic message, the apparent Russian hacking. But heartbroken and frustrated in-state battleground operatives worry that a lesson being missed is a simple one: Get the basics of campaigning right.”

“Clinton never even stopped by a United Auto Workers union hall in Michigan, though a person involved with the campaign noted bitterly that the UAW flaked on GOTV commitments in the final days, and that AFSCME never even made any, despite months of appeals.”

Thanks to J. Sekhon

2016

## How Clinton lost Michigan — and blew the election

Across battlegrounds, Democrats blame HQ's stubborn commitment to a one-size-fits-all strategy.

By EDWARD-ISAAC DOVERE | 12/14/16 05:08 AM EST

“Brooklyn mandated that they not worry about data entry. Operatives watched packets of real-time voter information piled up in bins at the coordinated campaign headquarters. The sheets were updated only when they got ripped, or soaked with coffee. Existing packets with notes from the volunteers, including highlighting how much Trump inclination there was among some of the white male union members the Clinton campaign was sure would be with her, were tossed in the garbage.”

“The Brooklyn command believed that television and limited direct mail and digital efforts were the only way to win over voters, people familiar with the thinking at headquarters said. Guided by polls that showed the Midwestern states safer, the campaign spent, according to one internal estimate, about 3 percent as much in Michigan and Wisconsin as it spent in Florida, Ohio and North Carolina. Most voters in Michigan didn’t see a television ad until the final week.”

“Most importantly, multiple operatives said, the Clinton campaign dismissed what’s known as in-person “persuasion” — no one was knocking on doors trying to drum up support for the Democratic nominee, which also meant no one was hearing directly from voters aside from voters they’d already assumed were likely Clinton voters, no one tracking how feelings about the race and the candidates were evolving. This left no information to check the polling models against — which might have, for example, showed the campaign that some of the white male union members they had expected to be likely Clinton voters actually veering toward Trump — and no early warning system that the race was turning against them in ways that their daily tracking polls weren’t picking up.”

# ML lessons

- Analytical algorithms CAN NOT automatically detect non-representative or biased samples



- Shoe leather work needs to be taken seriously and with analytical algorithms:

Information about undecided voters and people who do not respond to polls can be obtained only through ground operatives in their talking and interactions with such people.



# Reading assignments

- Review of pre-requisites
- Reading of James et al book chapter on cross validation