ORIE 5355: People, Data, & Systems

Lecture 2: Common challenges in data collection

Nikhil Garg

Questions from last time?

Module overview

- What *is* data? Where does it come from? What does it *represent?*
- Common challenges in data collection

  Selection biases, censoring, and other challenges
- Polling/surveys as an extended example
  - What goes wrong in measuring opinions (mean estimation)
  - Some techniques that somewhat work
  - US 2016 election polls as a case study
- Other challenges and contexts: online ratings, privacy, etc.

What is data?

A quick primer on measurement theory

What is a quantitative data point?

A measurement is "**assignment** of numbers to a **variable** in which we are interested."

- Construct/variable**:** what are we actually interested in?
- measurement/datum: numerical representation

These are not the same thing, especially with complexities of people!

Examples of constructs and (often flawed) measurements

Construct	Measurement
How well you understand the course material	A 1-100 grade, or a coarser letter grade
Your opinion about a movie	1-5 star rating, or a paragraph text review
Your political views/ideal public policy	Reduced to binary choice in voting
Race + Ethnicity	“white,” “Black,” “Asian” “Hispanic” “Other”
Gender	Often reduced to binary in surveys/forms

People disagree on how measurements map to constructs

- Ratings in online marketplaces across countries
  - In the US, anything but 5 stars means “terrible.”
  - In other countries, 3 or 4 stars is the norm
  - Heterogeneity within a country/culture: some people rate everything a 5 and always tip, others never do
- What do political terms mean?
  - Hakeem Jefferson, “The Curious Case of Black Conservatives: Construct Validity and the 7-point Liberal-Conservative Scale.”

Why does this matter?

- You're AirBnB
  - Do you have the same threshold for badges/`high quality' across countries?
  - People travel across countries, how do you standardize their ratings?
  - How do you communicate ratings to people from different cultures?
- You're doing a regression and trying to predict political leaning
  - When someone says they are "for environmental protection," does that mean they support raising taxes on fuel?
  - Do you do something different for Black people who say they're conservative versus white people who do so?
- You collect reports on problems in a city (311). What does it mean when someone reports an "unacceptable" pothole to fix?

What to do about it?

When *collecting* data, you can opt for free form text to give flexibility

- Doesn't constrain people to your pre-determined categories
- Potentially allows people to add more detail to capture the "construct"

This makes *analyzing* the data harder; doesn't fully solve the problem

- Most machine learning methods take in numeric or categorical data
- Even most modern NLP techniques convert words to numbers ("embeddings")
- Doesn't solve the problem of people using the same words to mean different things

=> this is a fundamental issue with quantitative data analysis

Ok, so what *can* you do?

You're going to have to make measurement choices at some point. Best make them consciously than by default.

- What is the data going to be used for? Do you need to create categories if there isn't a downstream prediction task?
- Categories chosen should relate to downstream tasks

  "Hispanic/Latino" category:

  - To know what languages to support, need to separate "Brazilian"
  - To predict political lean, separate out "Cuban in Florida"
- Some measures are more consistent than others

  Ask about more "objective" traits such as responsiveness or cleanliness

Parting thoughts about constructs

- Quantitative data science is all about creating general beliefs about discrete categories

  Also known as “stereotyping,” and data science inherits all its problems

- Be thoughtful about whether the measurement you have is appropriate for the construct you care about
- Many of the challenges we’ll discuss in this class are just the measurement-construct dichotomy in disguise

  [You really care about X, but the data you have can only tell you Y]

Questions?

Mean estimation from surveys

The task

- Each person j has an opinion, $Y_j \in \{0, 1\}$
- We want to measure $\bar{y} = E[Y_j]$, the population mean opinion on some issue
- Each person also has covariates, x_j^k (e.g., where they live)
- Sometimes, we also care about *conditional* means

$$E[Y_j \mid \text{lives in Roosevelt Island}]$$

This problem is everywhere

- What fraction will vote for the Democrat in the next election
- What is the average rating of this product?
- Do people want the city to build a foot bridge to Manhattan?
- Are people happy with this new feature I just deployed?

Naïve method

- Get list of people (watched the movie; from phone book)
- Call them, suppose everyone answers and get Y_j from each
- We now have $\{Y_j\}_{j=1}^{N}$, if called N people
- Simply do, $\hat{y} = \frac{1}{N}\sum_j Y_j$
- By law of large numbers, if Y_i is independent and identically distributed according to the true population's opinion, then

$$\hat{y} \to \bar{y} \text{ as } \mathrm{N} \to \infty$$

What goes wrong

People don't give "true" opinion

Why?

- You're asking about something sensitive
- "social desirability" – people like making other people happy
- They're getting paid to answer the survey and just want to finish
- You know they other person is also going to rate you

Of course, then you're (likely) not going to succeed

People gave you $\widetilde{Y}_j$, instead of Y_j

$\hat{y} = \frac{1}{N}\sum_j \widetilde{Y}_j$

$\hat{y}$ does not converge to $\bar{y}$, *unless errors cancel out*

Your sample does not represent your population

- You just posted a poll on Facebook or Twitter, anyone could respond
- You called only landlines, and no one under 50 owns one anymore
- You only asked people to rate a movie after they've seen it
- You can only rate an item if you bought it *and didn't return it*
- Those with certain opinions are more likely not to answer
  - After bad experiences on online platforms
  - "Shy Trump voters" (?)

=> People who answer the poll are different than your population – "differential non-response"

Your sample does not represent your population, in math

- For each person j, let $A_j \in \{0,1\}$ be whether they answered
- You have $\boldsymbol{Y} = \{(A_j, Y_j)\}_{j=1}^{N}$, if called N people

  Where $Y_j = \emptyset$ if $A_j = 0$ (they did not answer)
- Again, you do

  $$\hat{y} = \frac{1}{|\{j \mid A_j=1\}|} \sum_{j \in \{j \mid A_j=1\}} Y_j$$

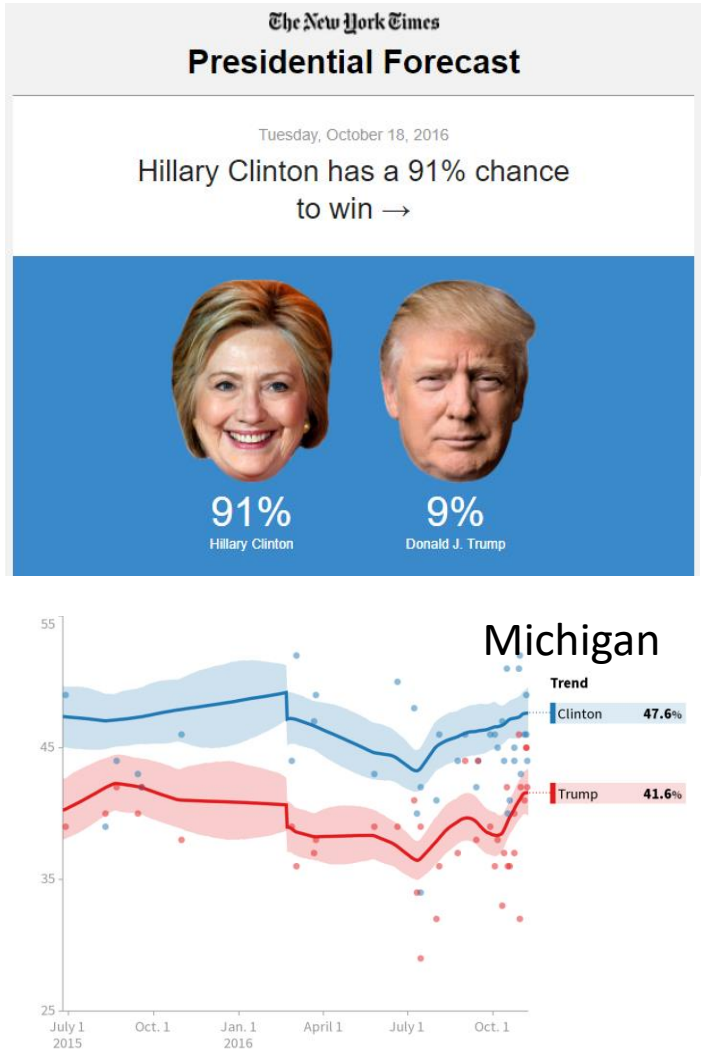
  where $\{j \mid A_j = 1\}$ denotes the set of people who answered

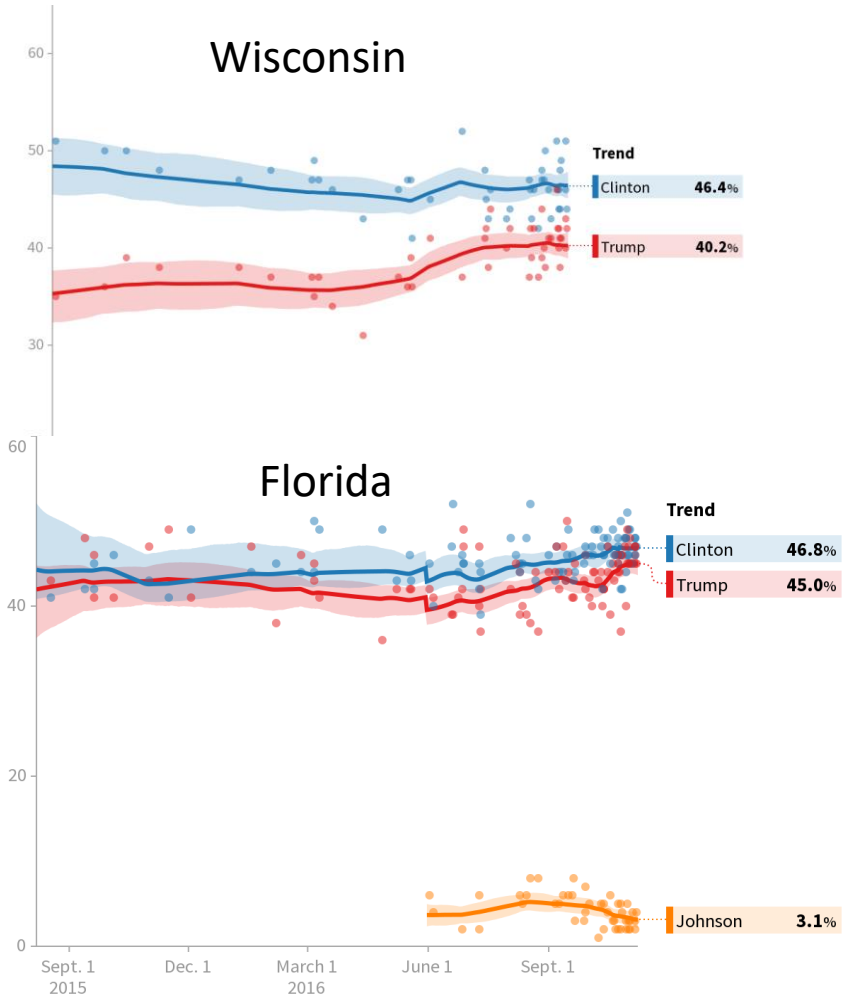
  and so $|\{j \mid A_j = 1\}|$ is the number of people who answered

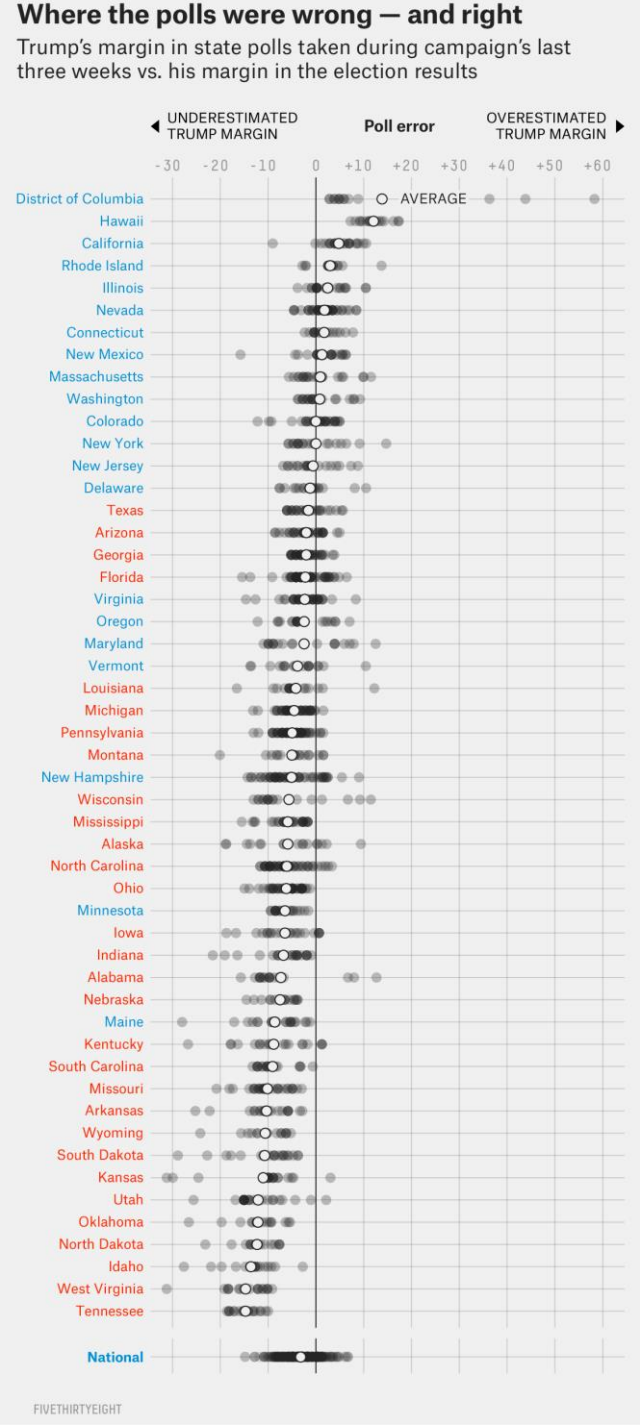
$\hat{y}$ does not converge to $\bar{y}$ *unless* Y_j and A_j are uncorrelated

Case study: Polling in US 2016 presidential election

Polls were off (a bit) in the 2016 e

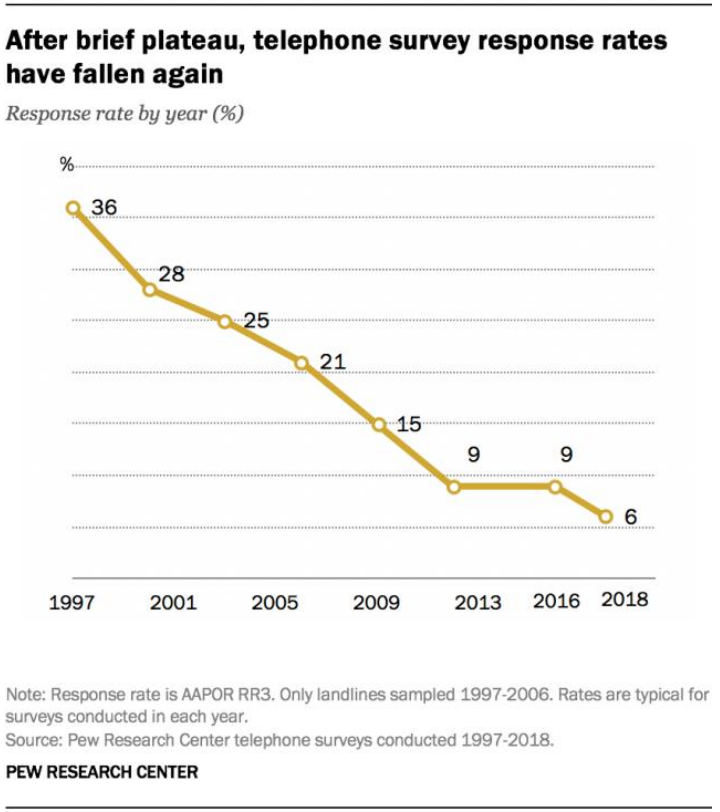






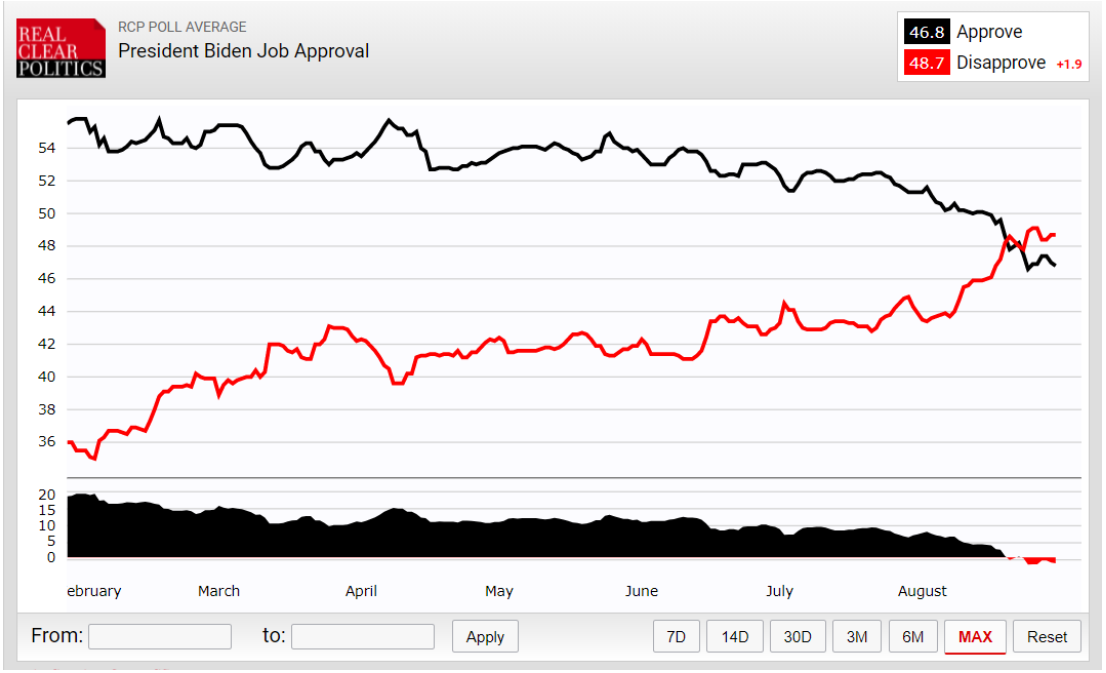
What happened?

- Professional pollsters spend a lot of time on getting opinions right
  - [We'll cover some of their techniques next time]
- But, polling is an increasingly challenging business
  - Basically no one answers a phone poll
  - Modeling opinions/turnout in diverse democracy is hard
  - "social desirability" → "shy Trump voters" (?)
- In 2016, turns out that less educated voters both:
  - Were less likely to answer polls
  - Were more likely to vote Trump



Differential non-response is everything

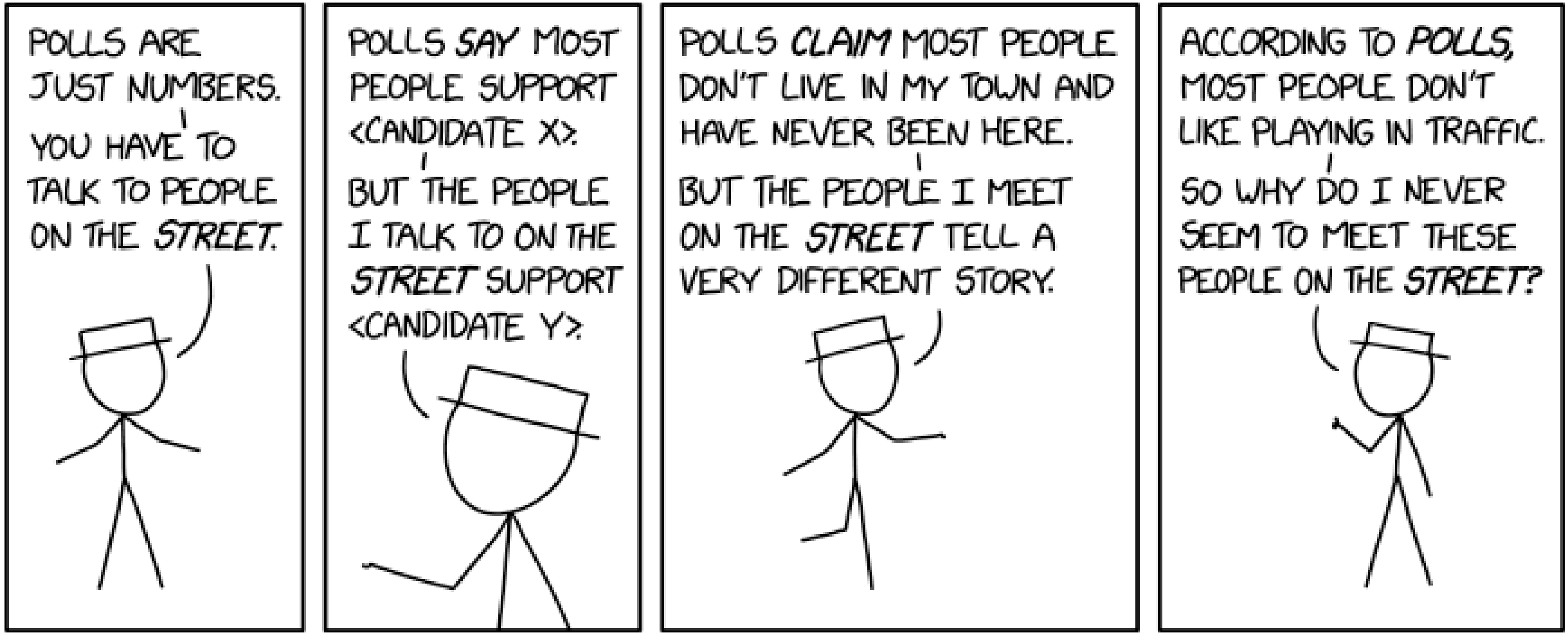
- Differential non-response shows up everywhere you're gathering opinions
- Your training data for whatever model you train faces the same issue!
- Standard "margin of error" calculations do not take this into account
- Differential non-response *over time* often explains "swings" in polls!



Parting thoughts

Be purposeful! Does your numeric data capture what you want it to?

Be skeptical! Just because a poll was “random” doesn’t make it good



Other pollsters complain about declining response rates, but our poll showed that 96% of respondents would be 'somewhat likely' or 'very likely' to agree to answer a series of questions for a survey.

xkcd: Polls vs the Street

Announcements

- Homework 1 being posted on Friday
- TA office hours start ~~next week~~ Friday On zoom (access via canvas)
- My office hours today, 4:30p, ~~outside café~~ On zoom (access via canvas)

Questions?