ORIE 5355: Applied Data Science - Decision-making beyond Prediction

Lecture 2: Common challenges in data collection

Nikhil Garg

Announcements

- Homework 1 posted by Tuesday, due 9/15
- Please fill out:
  - When2meet for office hours
  - Pre-course survey
  - Homework buddy form

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 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 the "**assignment** of numbers to a **variable** in which we are interested."

- Construct: what are we actually interested in?
- Representation/variable: ideal numerical representation
- Measurement/data: what we actually get in the dataset

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

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

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

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”

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

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
How healthy someone is	How much healthcare spending they have

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
How healthy someone is	How much healthcare spending they have

In all these cases, we might not even have good access to this numerical measurement!

- I mis-enter your grade
- A researcher doesn’t have access to data they need
- You don’t vote in the election

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
- Different courses have different grade distributions; students may expect “A”s for average work, but others might want to limit them to 20% of class

Why does this matter?

- You're Airbnb
  - Do you have the same threshold for badges and "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 analyzing a survey; when someone says they are "for environmental protection," does that mean they support raising taxes on fuel? Preventing housing to save the turtles? Government funding of clean energy?
- In a recommender system, does someone "liking" or "reposting" or "viewing" something mean they want to see more of that content?
- You collect reports on problems in a city (311). What does it mean when someone reports an "unacceptable" pothole to fix? Can you trust that different people are reporting problems at the same rates?

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")
- LLMs help in analyzing text, but you still need to summarize
- 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. It is better 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?

Basic setting: mean estimation

The task

- Each person j has some complex thoughts about the course, C_j
- This is reduced to a binary opinion, $Y_j \in \{0, 1\}$
- We want to measure $\mu = E[Y_j]$, the population mean opinion on some issue
- Each person also has covariates, X_j
- We also may care about *conditional* means

$$\mu_{\text{ORIE}} = E[Y_j \,|\, \text{ORIE program}]$$

Example:

"Do you like the class so far?"

Options: "yes" and "no"

What fraction of people like the class so far?

Degree program, whether you like afternoon classes, etc

Fraction of people in ORIE who like the class

Notation so far

- **Construct C_j:** Each person j's complex thoughts
- **Ideal measurement:** $Y_j \in \{0, 1\}$
- **Goal:** estimate $\mu = E[Y_j]$, the population mean
- **Covariates** about a person: X_j

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?

From ideal measurement to data

- If I run a survey of the class asking the binary question, in the best case, I get everyone's true $Y_j \in \{0, 1\}$
- Then, I can take the average and directly get $\mu = E[Y_j]$

Two potential issues:

- Not everyone responds: Let $O_j \in \{0, 1\}$ indicate that I *observe* a response from person j
- What if some people don't give me their true opinion? Let $Z_j \in \{0, 1\}$ be the response we actually get from person j

Thus, our ideal dataset is the set $\{Y_j \mid j \text{ is in the class}\}$

- Our actual dataset is $\{Z_j \mid j \text{ is in the class and responds}, O_j = 1\}$

Notation so far

- **Construct C_j:** Each person j's complex thoughts
- **Ideal measurement:** $Y_j \in \{0, 1\}$
- **Goal:** estimate $\mu = E[Y_j]$, the population mean
- **Covariates** about a person: X_j
- **Whether someone responds:** $O_j \in \{0, 1\}$
- **Someone's actual answer**: $Z_j \in \{0, 1\}$

Naïve method in the ideal case

- Get list of n students (from class roster)
- Call them; everyone answers and get $Z_j = Y_j$ from each
- We now have $\{Y_j\}_{j=1}^{n}$

Random sample of people in this class

- Simply do, $\hat{\mu} = \frac{1}{n}\sum_j Y_j$

Average opinion of the sample

- By law of large numbers, if Y_i is independent and identically distributed according to the true population's opinion, then

$$\hat{\mu} \to \mu \text{ as } n \to \infty$$

μ : Actual opinion of the class

Naïve method in the actual case

- Get list of n students (from class roster)
- Call them; n_{obs} people answer and we get Z_j from each
- We now have $\{Z_j\}_{j=1}^{n_{obs}}$
- And so our estimate is, $\hat{\mu} = \frac{1}{n_{obs}} \sum O_j Z_j$

Notation so far

- **Construct C_j:** Each person j's complex thoughts
- **Ideal measurement:** $Y_j \in \{0, 1\}$
- **Goal:** estimate $\mu = E[Y_j]$, the population mean

- **Covariates** about a person: X_j
- **Whether someone responds:** $O_j \in \{0, 1\}$
- **Someone's actual answer**: $Z_j \in \{0, 1\}$
- **Number of in the population we care about:** N
- **Number of people asked:** n
- **Number of people who actually responded:** $n_{obs} = \sum O_j$
- **And so our naïve estimator is:**

$$\hat{\mu} = \frac{1}{n_{obs}} \sum O_j Z_j$$

What goes wrong

What all can go wrong?

Why might $\hat{\mu} = \frac{1}{n_{obs}} \sum O_j Z_j$ be very far from $\mu = E[Y]$?

- What if I only asked 1 person? (Or more generally only got 1 response)

$$n_{obs} = 1$$

- What if some people systematically gave the wrong answer?

$$Z_j = 1 \text{ if } Y_j = 0$$

- What if everyone who hated the class refused to answer?

$$O_j = 0 \text{ if } Y_j = 0$$

You don't ask enough people (or don't get enough responses)

- Suppose the true opinion rate is μ
- Any single person I ask has an opinion $Y_j \in \{0, 1\}$

  $Y_j \sim \text{Bernoulli}\,(\mu)$ (opinion is 1 with probability μ)
- And so the mean opinion from n_{obs} is random as well

  $\sum O_j Y_j \sim \text{Binomial}(\mu, n_{obs})$

(If I only ask 2 people, their mean is 1 w.p. μ^2, 0 w.p $(1-\mu)^2$, and .5 otherwise)

This is called sampling *variance*

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 the other person is also going to rate you

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

People gave you Z_j, instead of Y_j

$$\hat{\mu} = \frac{1}{n_{obs}} \sum_j O_j Z_j$$

You lie because you want a better grade

$\hat{\mu}$ does not converge to μ, *unless errors cancel out*

This is an example of *bias*

People who appear in the observed sample differ from the target population you care about

- 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" (?)

Often called selection effects or differential non-response

Your sample does not represent your population, in math

- For each person j, let $O_j \in \{0,1\}$ be whether they answered
- You have data $\{Z_j \mid j \text{ responds}, O_j = 1\}$ Again, you do

$$\hat{\mu} = \frac{1}{n_{obs}} \sum_{j \in \{j \mid O_j = 1\}} Z_j$$

Uncorrelated: Whether you answered is unrelated to what your true opinion is

When does $\hat{\mu} \to \mu$? When Y_j and O_j are uncorrelated

- Suppose someone flipped a coin before deciding to answer
- Suppose someone only answered if their opinion was $Y_j = 1$

This is an example of *bias*

Putting things together

$\mu = E[Y]$ (true average opinion in population)

$\hat{\mu} = \frac{1}{n_{obs}} \sum_{j \in \{j \mid O_j = 1\}} Z_j$ (avg answer among responded sample)

$$\hat{\mu} - \mu = (\hat{\mu} - E[Z \mid O = 1]) + (E[Z \mid O = 1] - E[Y \mid O = 1]) + (E[Y \mid O = 1] - \mu)$$

Overall error

Finite sampling error: I got some small number of responses compared to if I had asked a billion people (source of variance)

Selection effect: true opinion among those who answer differs from population true opinion

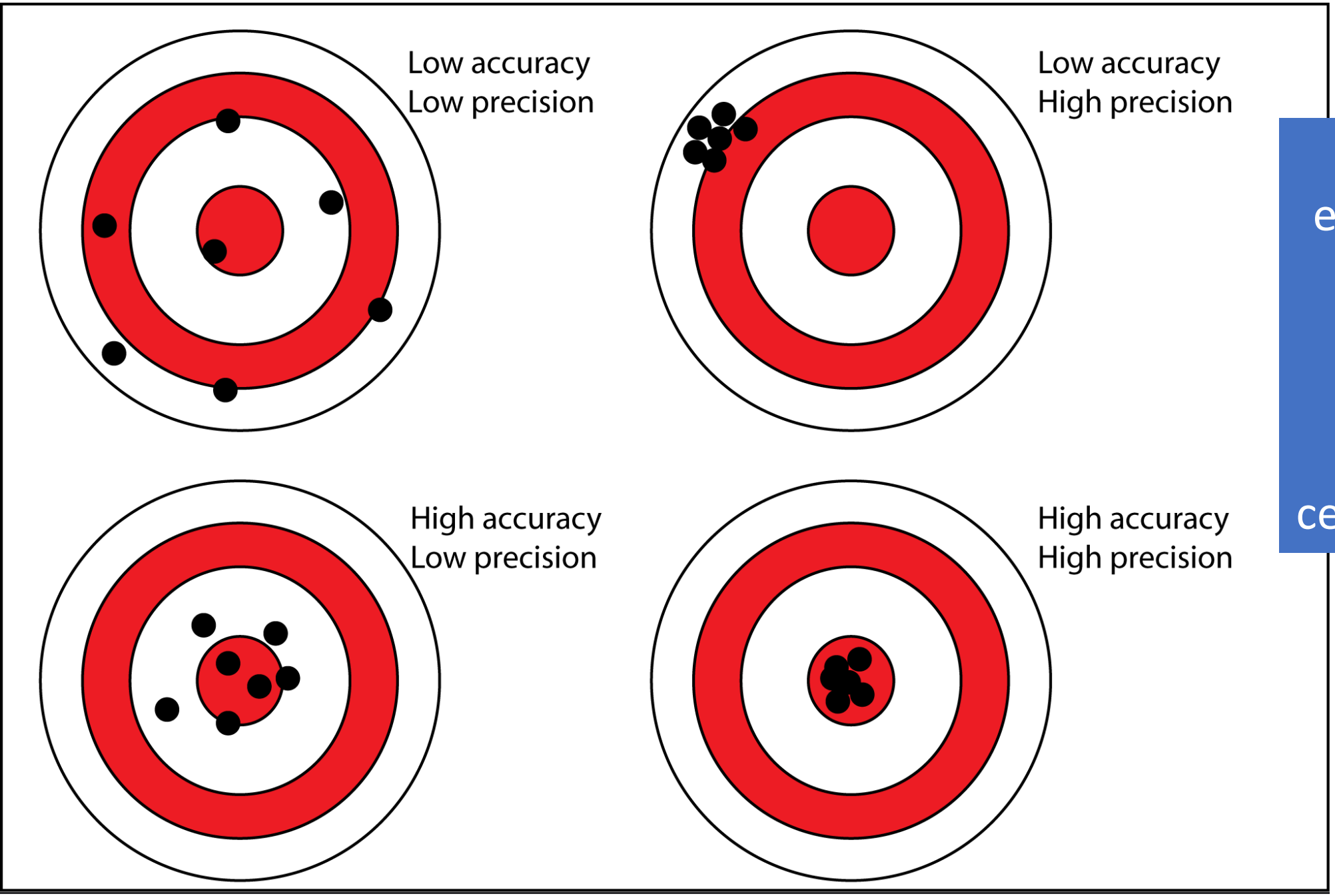
Measurement error: Given someone responded, their response doesn't equal their actual opinion

Not included:
Construct mismatch: Y isn't actually what I wanted!

Possible Errors

- Finite sampling error: I got some small number of responses compared to if I had asked a billion people (source of variance)
- Measurement error: Given someone responded, their response doesn't equal their actual opinion
- Selection effect: true opinion among those who respond differs from population true opinion
- Construct mismatch: Y isn't actually what I wanted!

All matter in every data science task!

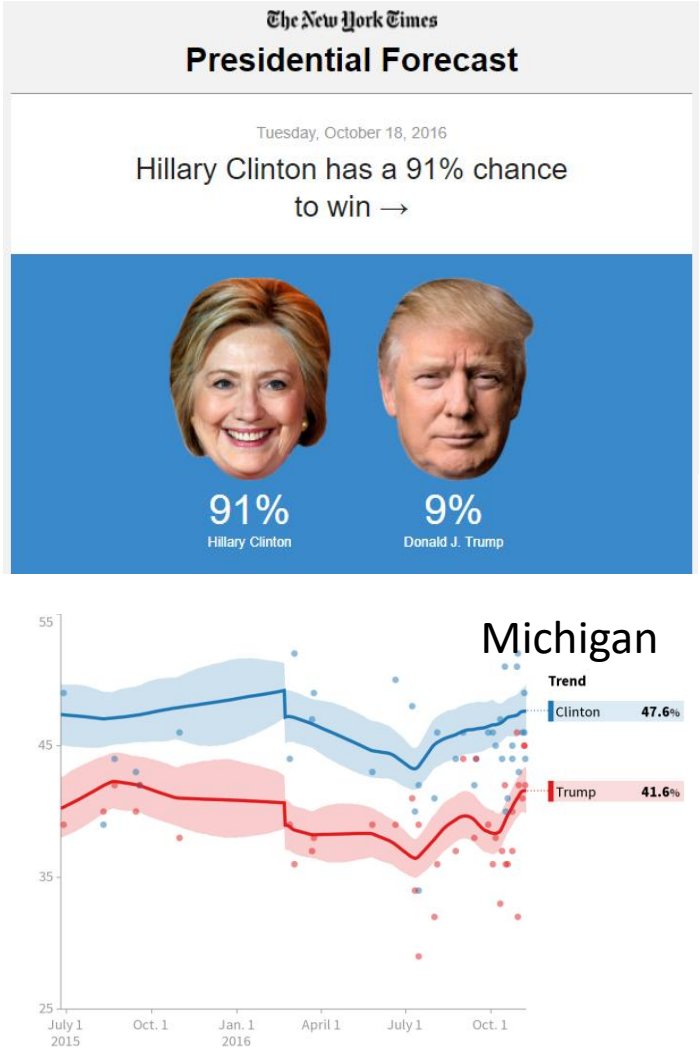


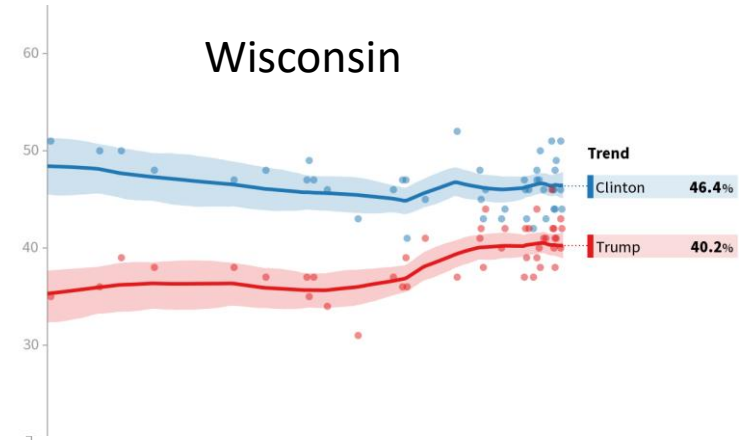
Low variance: estimates cluster together

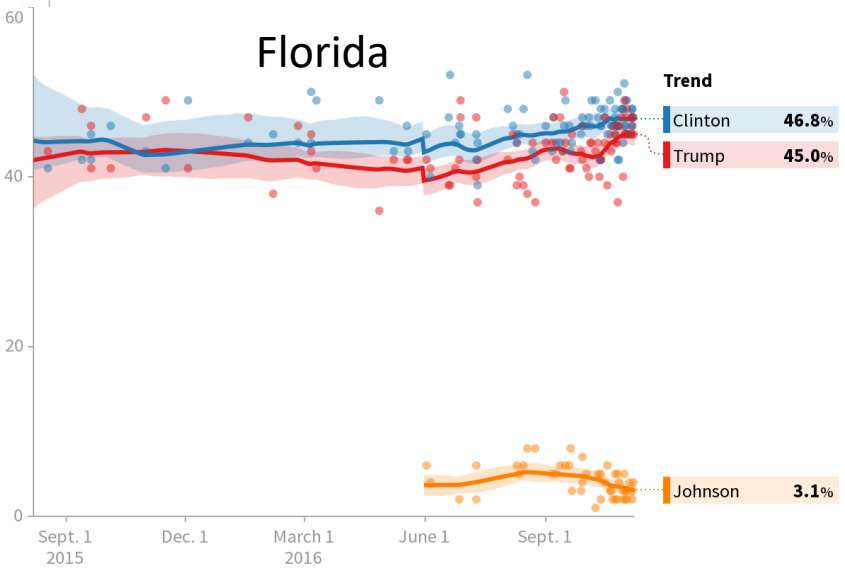
Low bias: estimates centered on target

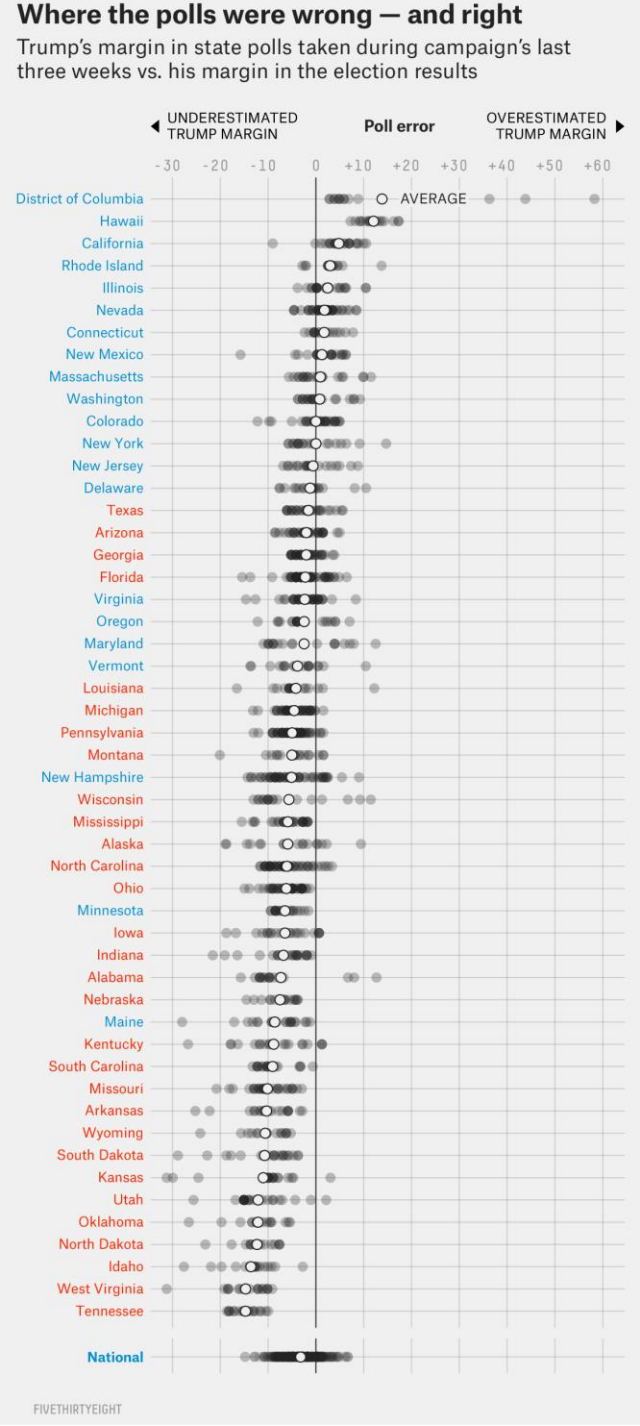
Case study: Polling in 2016 U.S. presidential election

Polls were off (a bit) in the 2016 e



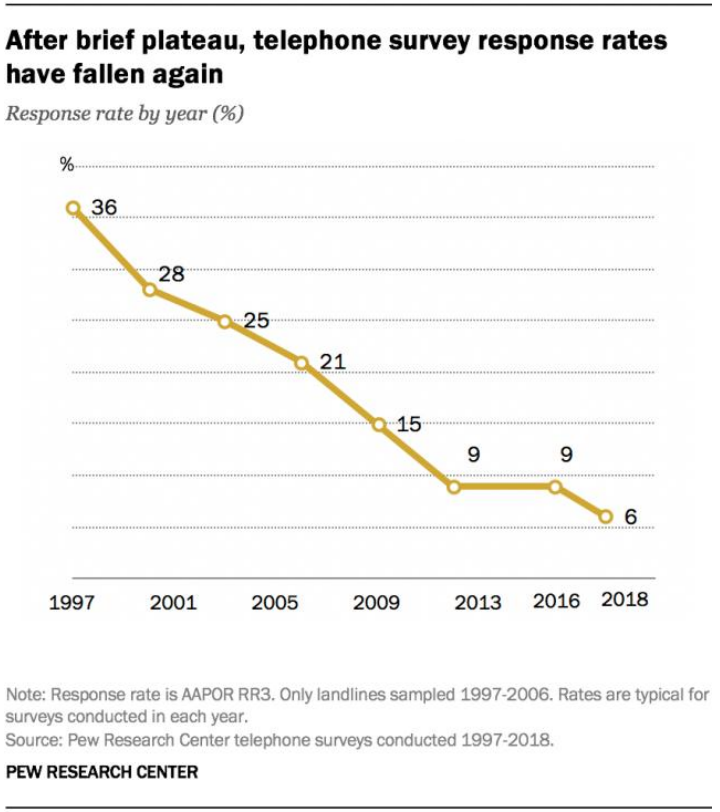






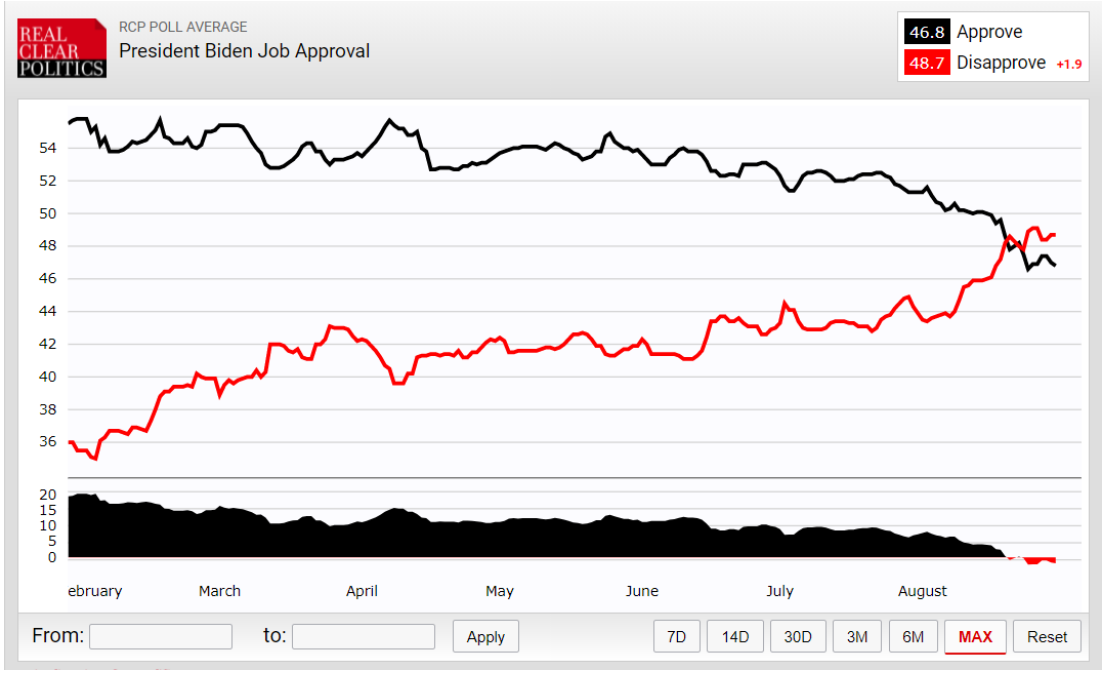
What happened?

- 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 a diverse democracy is hard
  - "social desirability" → "shy Trump voters" (?)
- In 2016, it turned out that voters without college degrees both:
  - Were less likely to answer polls
  - Were more likely to vote Trump



Differential non-response is everything

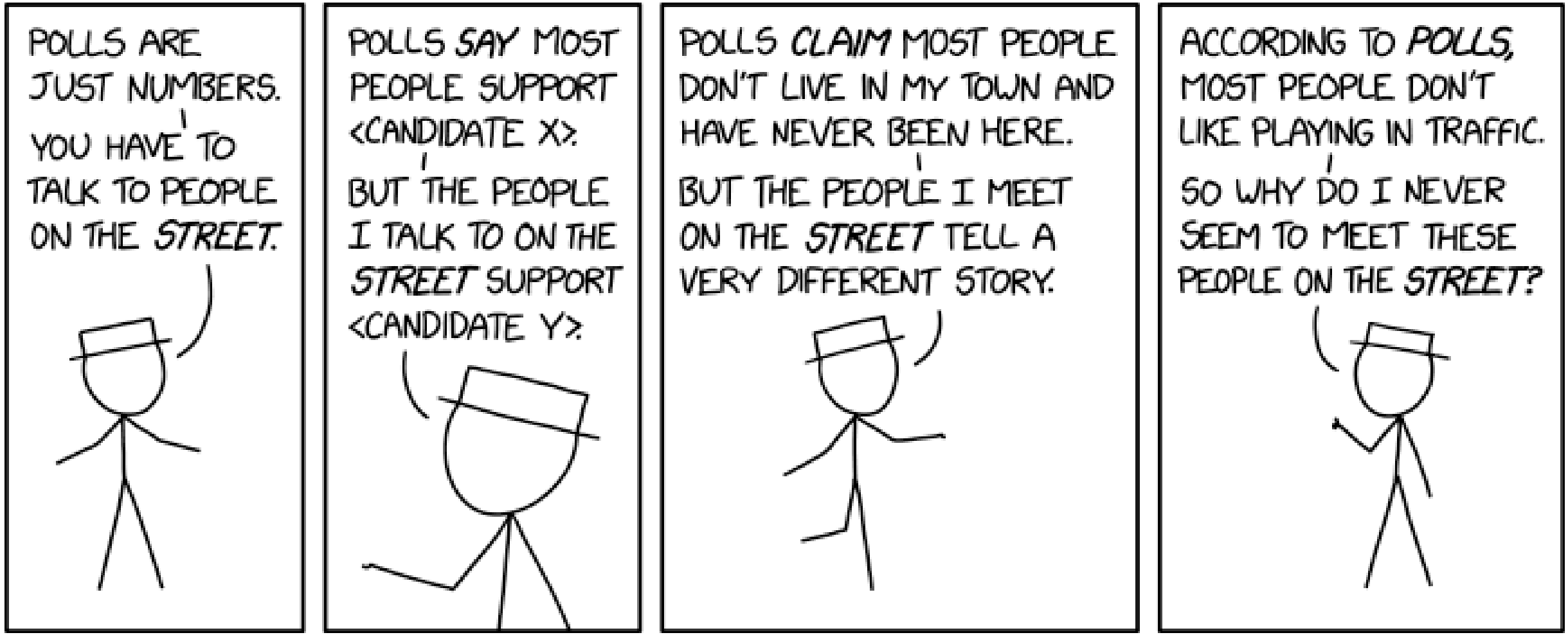
- Differential non-response shows up everywhere you're gathering opinions
- Your training data for any model you build faces the same issue!
- Standard "margin of error" calculations do not take this into account
- Differential non-response *over time* often explains "swings" in polls!



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

Questions?