The Art and Science of Election Analytics

Soubhik* Barari, Ph.D.

Research Scientist
SurveyMonkey (Momentive.ai)

June 6th, 2023
New York Open Statistical Programming Meet-Up

**Pronounced "Sho - Bik"*

@soubhikbarari

soubhikbarari.com

sbarari@g.harvard.edu

About me

I'm a Quantitative Social Scientist by training

- Finished my Ph.D. a week ago!!! (😎 but also 😭)
- My academic research uses data science tools (e.g. NLP, Bayesian models) to show how political media ecosystems in the U.S. work
- Side note: political science is the original data science

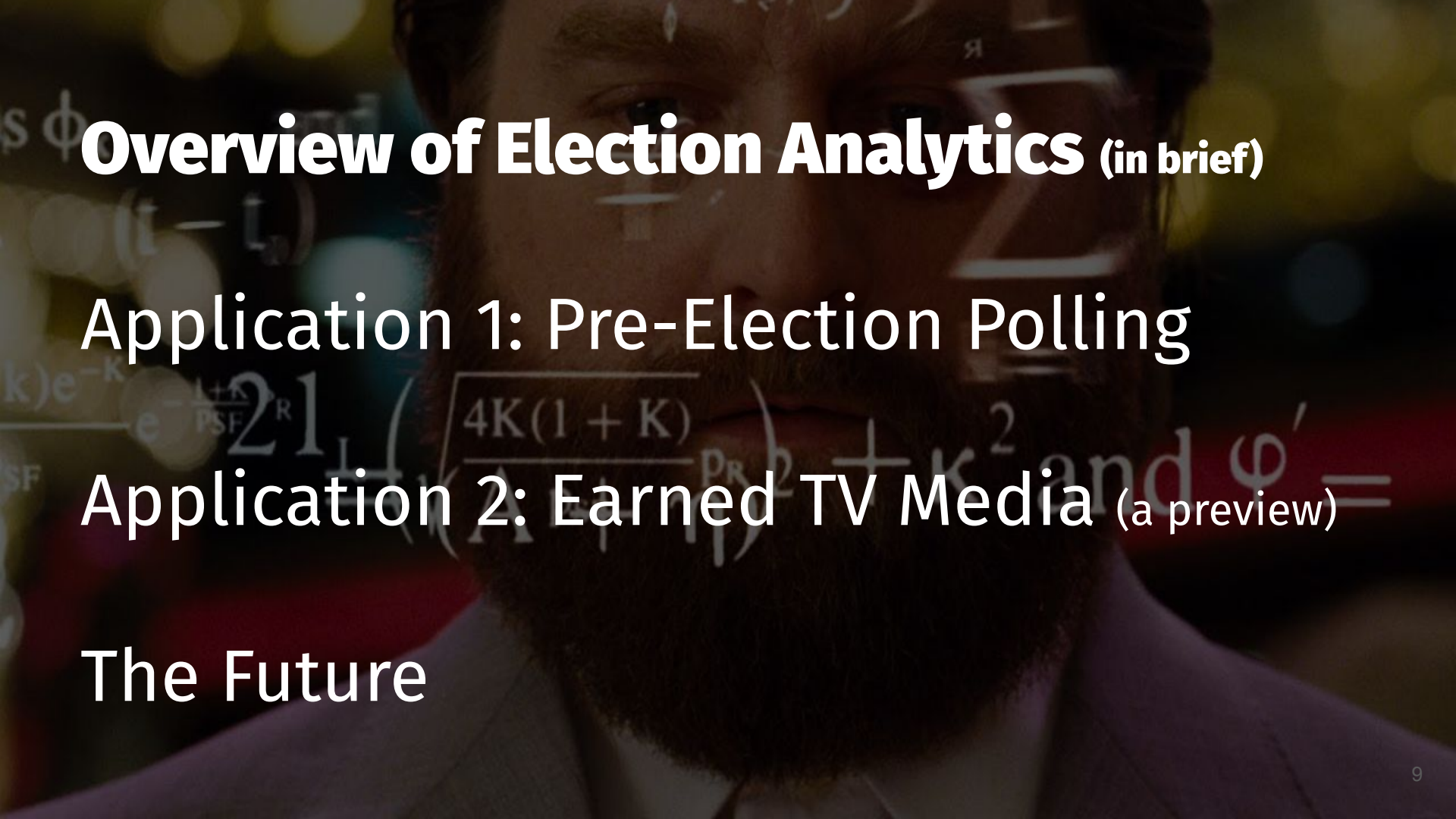
I'm a Research Scientist at SurveyMonkey

- I work on "full-stack" survey science (sampling, weighting, analysis)
- Partnerships with Harvard Medical School, AAPI Data, J. D. Power
- **Creator of R package `svmkR`**
- **Co-led 2022 pre-election polling with NBC News / UPenn PORES**

I'm a former (brief) Microsoftie*

- I interned at MSR in 2021 (Computational Social Science group)
- Using data science (e.g. SQL, causal inference) to understand the **coverage and effects of televised news on election campaigns**

*is that a thing?



Overview of Election Analytics (in brief)

Application 1: Pre-Election Polling

Application 2: Earned TV Media (a preview)

The Future

200+ years of innovation in election analytics

1788: The First Election

- Before parties…
- …but, *was* a huge party.
- 1% of population voted.

1788 1824 1860 1880 1924 1936 1952 1960 2006 2008 2012 2024

New Tools:

- booze
- bribes
- elite connections
- fun theme songs
- intimidation
- stump speeches

200+ years of innovation in election analytics

1824: The Birth of Polling

- First "poll" published by a newspaper.
- Straw polls were fun, but very biased and mostly benefited newspapers.


The Pennsylvanian.

HARRISBURG, PUBLISHED BY CHRISTIAN GLEIM & SAMUEL CONRAD.

"LIBERTY WITHOUT KNOWLEDGE IS BUT A PRECARIOUS BLESSING."

Vol. I. SATURDAY, AUGUST 21, 1824. No. 23.

TERMS OF PUBLICATION.

FLAX SEED.

A LIST OF CAUSES

the actual newspaper in question!

New Tools:

1788 1824 1860 1880 1924 1936 1952 1960 2006 2008 2012 2024

○ straw polls

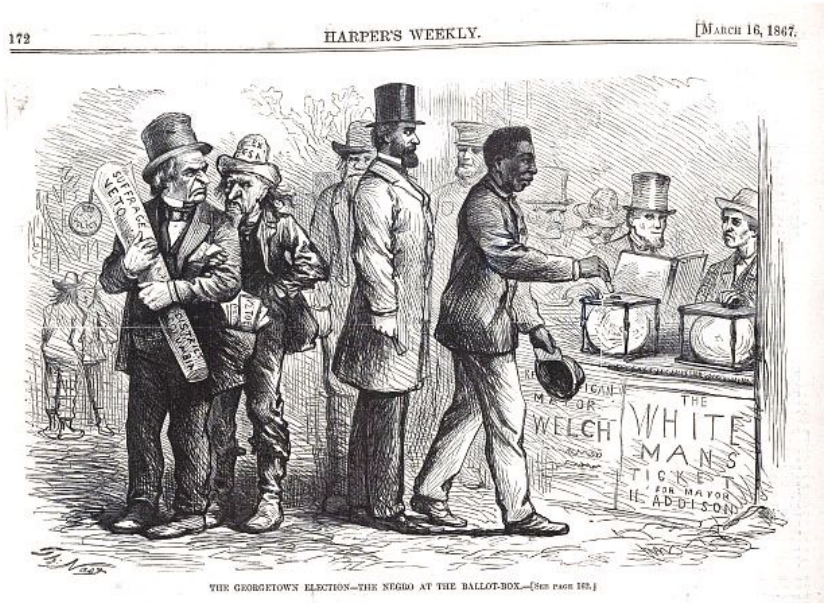
Recommended Reading:

"Strength in Numbers: How Polls Work and Why We Need Them" by G. Elliott Morris (2022)

200+ years of innovation in election analytics

1860-80: The First Voter Files

- Voting rights for Black Americans.
- Major voter registration reforms.
- Key opportunity for political parties to systematically track and target voters.



1788 1824 1860 1880 1924 1936 1952 1960 2006 2008 2012 2024

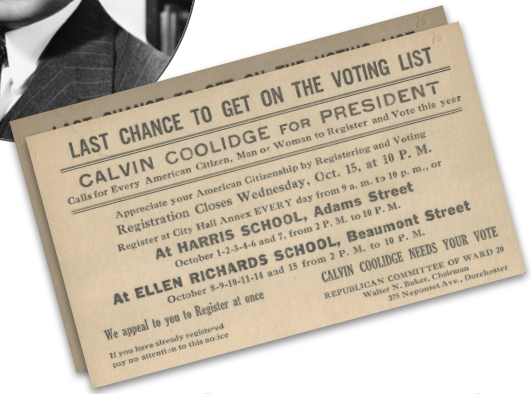
New Tools:

- crude **voter files**

Recommended Reading:
"Hacking the Electorate: How Campaigns Perceive Voters" by Eitan Hersh (2015)

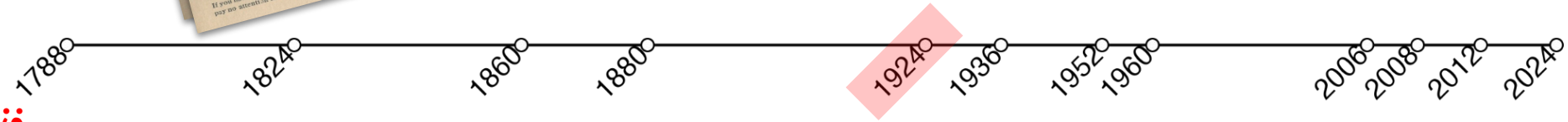
200+ years of innovation in election analytics

Harold Gosnell,
Professor at UChicago



1924: The First Election Field Experiment

- Two-stage randomized voter registration mailer in Chicago (*n* = 3,000).
- Treatment group +10% more likely to register for upcoming election.

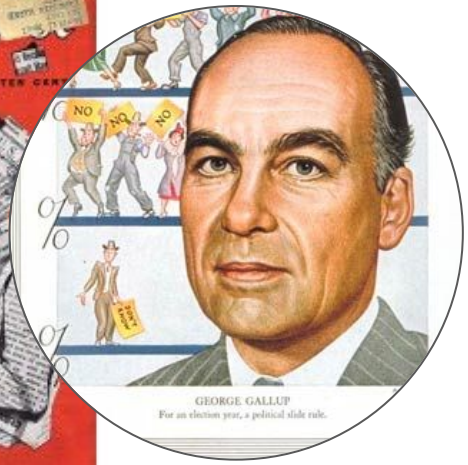


New Tools:

- **randomized control trials**
  (not particularly popular…)

Recommended Reading:
"The Victory Lab: The Secret Science of Winning Campaigns" by Sasha Issenberg (2012)

200+ years of innovation in election analytics



George Gallup,
Pollster (arguably the most famous)

1936: The Polling Revolution Begins

- Literary Digest's massive straw poll (*n* = 3 million) massively mis-predicts that FDR would lose massively in election.
- But one man saw it all coming and changed polling forever…

1788 1824 1860 1880 1924 **1936** 1952 1960 2006 2008 2012 2024

New Tools:

- large-sample straw polls,
- more scientific **"probability" polls**

Recommended Reading:
"Strength in Numbers: How Polls Work and Why We Need Them" by G. Elliott Morris (2022)

200+ years of innovation in election analytics

1952: The Rise of Television & The Behavioral Revolution

- First television ad (Eisenhower).
- UMich produces the academic foundations of behavioral political science.
- Birth of the longest running panel on U.S. political attitudes (ANES).

INSTITUTE FOR SOCIAL RESEARCH
UNIVERSITY OF MICHIGAN

ANES
American National Election Studies

1788 1824 1860 1880 1924 1936 1952 1960 2006 2008 2012 2024

New Tools:

- **media market** and audience data,
- field of **survey methodology**,
- field of **political behavior**
- large-scale survey panels

Recommended Reading:
"The Victory Lab: The Secret Science of Winning Campaigns" by Sasha Issenberg (2012)

200+ years of innovation in election analytics

Left: JFK and Nixon in the first televised presidential debate.

Right: "The People Machine" invented by Simulmatics Co.

1960: The First TV Debate & The Rise of Analytics Firms

- JFK vs. Nixon
- The Simulmatics Corporation

1788 1824 1860 1880 1924 1936 1952 1960 2006 2008 2012 2024

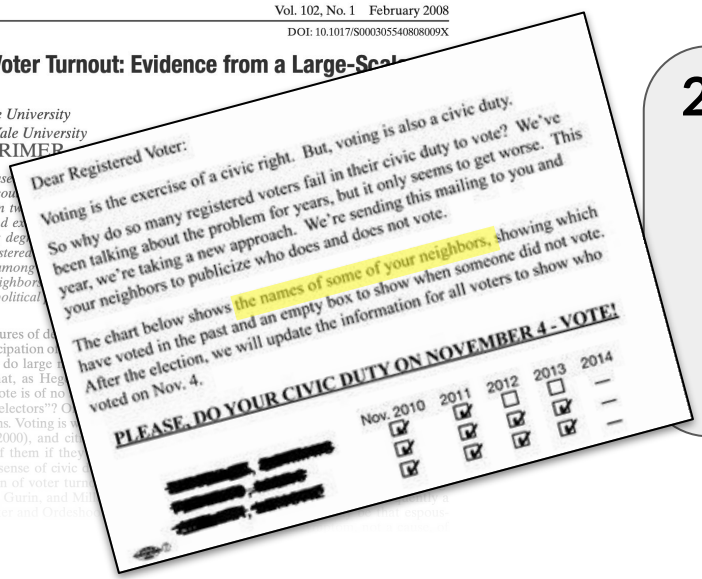
New Tools:

- crude **prediction** (e.g. county-level results)
- crude **simulation** (e.g. Electoral College)

Recommended Reading:

"If Then: How the Simulmatics Corporation Invented the Future" by Jill Lepore (2020)

200+ years of innovation in election analytics



American Political Science Review Vol. 102, No. 1 February 2008

DOI: 10.1017/S0003055408080009X

Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment

ALAN S. GERBER *Yale University*
DONALD P. GREEN *Yale University*
CHRISTOPHER W. LARIMER

Voter turnout theories base... turnout unless they accou... We distinguish between tw... in accordance with a norm and ex... motives and applying varying deg... several hundred thousand registered... higher turnout was observed among... to their household or their neighbor... pressure as an inducement to political...

Among the most striking features of d... political systems is the participation o... of voters in elections. Why do large ... of people vote, despite the fact that, as Heg... observed, "the casting of a single vote is of no... cance where there is a multitude of electors"? (... pothesis is adherence to social norms. Voting is w... regarded as a citizen duty (Blais 2000), and ci... worry that others will think less of them if they... to participate in elections. Voters' sense of civic ... has long been a leading explanation of voter turn... among both behavioral (Campbell, Gurin, and M... 1954) and formal (Downs 1957; Riker and Ordesh...

Dear Registered Voter:

Voting is the exercise of a civic right. But, voting is also a civic duty.

So why do so many registered voters fail in their civic duty to vote? We've been talking about the problem for years, but it only seems to get worse. This year, we're taking a new approach. We're sending this mailing to you and your neighbors to publicize who does and does not vote.

The chart below shows the names of some of your neighbors, showing which have voted in the past and an empty box to show when someone did not vote. After the election, we will update the information for all voters to show who voted on Nov. 4.

PLEASE, DO YOUR CIVIC DUTY ON NOVEMBER 4 - VOTE!

	Nov. 2010	2011	2012	2013	2014
[redacted]	☑	☑	☐	☑	—
	☑	☑	☑	☑	—
	☑				—

2006: The Rise of "Nudges"

- One of the most famous political science papers ever published: Gerber and Green (2008)
- Social pressure works in getting out the vote!

1788 1824 1860 1880 1924 1936 1952 1960 2006 2008 2012 2024

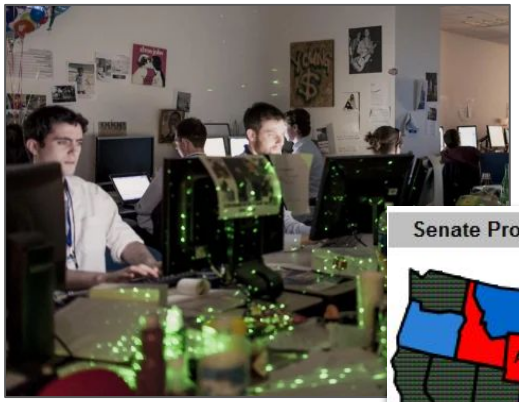
New Tools:

- field of **behavioral science**
- advances in **field experimentation**

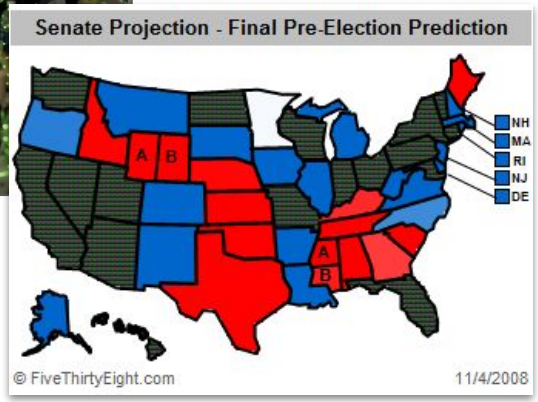
Recommended Reading:

"Get Out the Vote! How to Increase Voter Turnout" by Don Green and Alan Gerber (2015)

200+ years of innovation in election analytics



Left: The Obama campaign "Nerd Cave"



Right: FiveThirtyEight's 2008 (nearly perfect) Senate forecasts

2008: The "Big Data" Era Begins

- Obama '08 campaign modernized how campaign analytics is done.
- FiveThirtyEight / Nate Silver revolutionizes public election forecasts.

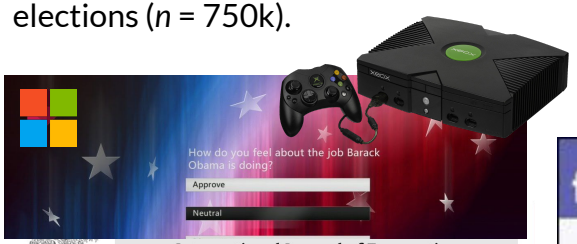
1788 1824 1860 1880 1924 1936 1952 1960 2006 2008 2012 2024

New Tools:

- **micro-targeting**
- **social media mobilization**
- **augmented voter files**
- **machine learning**
- complex **election forecasts**

200+ years of innovation in election analytics

Below: poll asked on XBox Live during 2012 elections (*n* = 750k).



Forecasting elections with non-representative polls

Wei Wang [a,*], David Rothschild [b], Sharad Goel [b], Andrew Gelman [a,c]
[a] Department of Statistics, Columbia University, New York, NY, USA
[b] Microsoft Research, New York, NY, USA
[c] Department of Political Science, Columbia University, New York, NY, USA

ARTICLE INFO

Keywords:
Non-representative polling
Multilevel regression and poststratification
Election forecasting

ABSTRACT

Election forecasts have traditionally been based on representative polls, in sampled individuals are asked who they intend to vote for. While represen historically proven to be quite effective, it comes at considerable costs of Moreover, as response rates have declined over the past several decad

nature

A 61-million-person experiment in social influence and political mobilization

Robert M. Bond, Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cam
Marlow, Jaime E. Settle & James H. Fowler

Abstract

Human behaviour is thought to spread through face-to-face social
ts in
online soci
ults from a
ages deliver
al elections
l self-expres
illions of pe
l elections
self-expres
illions of pe

2012-16: The "Tech Elections"

- Phone poll response rates below 10% for first time… 😱
- MSR paper shows how to accurately adjust highly unrepresentative polls
- FB experiment shows online social networks can influence elections.
- Nonresponse bias in 2016 polls correlates with education 💀

Left: randomized registration message on FB newsfeed (*n* = 61 million).

1788 1824 1860 1880 1924 1936 1952 1960 2006 2008 2012 2024

New Tools:

- **multilevel regression and poststratification (MRP)** (not really new)
- **social media experiments**
- **online polling methods**

200+ years of innovation in election analytics

Above: still from ad released by RNC created via generative AI.

Right: AI-generated image of Trump praying – directly shared by Donald Trump.

Below: AI-generated image of Trump's arrest created on Midjourney.

1788 1824 1860 1880 1924 1936 1952 1960 2006 2008 2012 2024

New Tools:

- **generative AI**
- **online / mixed polling methods**
- **???**

The current election analytics landscape

The Who

- Researchers
  (academia, think thanks)
- Practitioners
  (party organizations, campaigns, consulting firms, governments, polling firms, tech firms, advocacy groups, law firms)

The What

- Opinion polling
- Outcome forecasting
- Message testing
- Resource optimization
- Qualitative research
- Election auditing

The Why

- Improve campaign decisions
  (voter outreach, ad spend, earned media, messaging, fund-raising)
- Improve news coverage
- Improve advocacy
- Trust and safety

The How

- Data engineering
- Dashboard building
- SQL, Python, R, STATA/SPSS
- Applied regression modelling
- Statistical learning
- Causal inference
- Econometrics

5 big lessons we've learned about elections

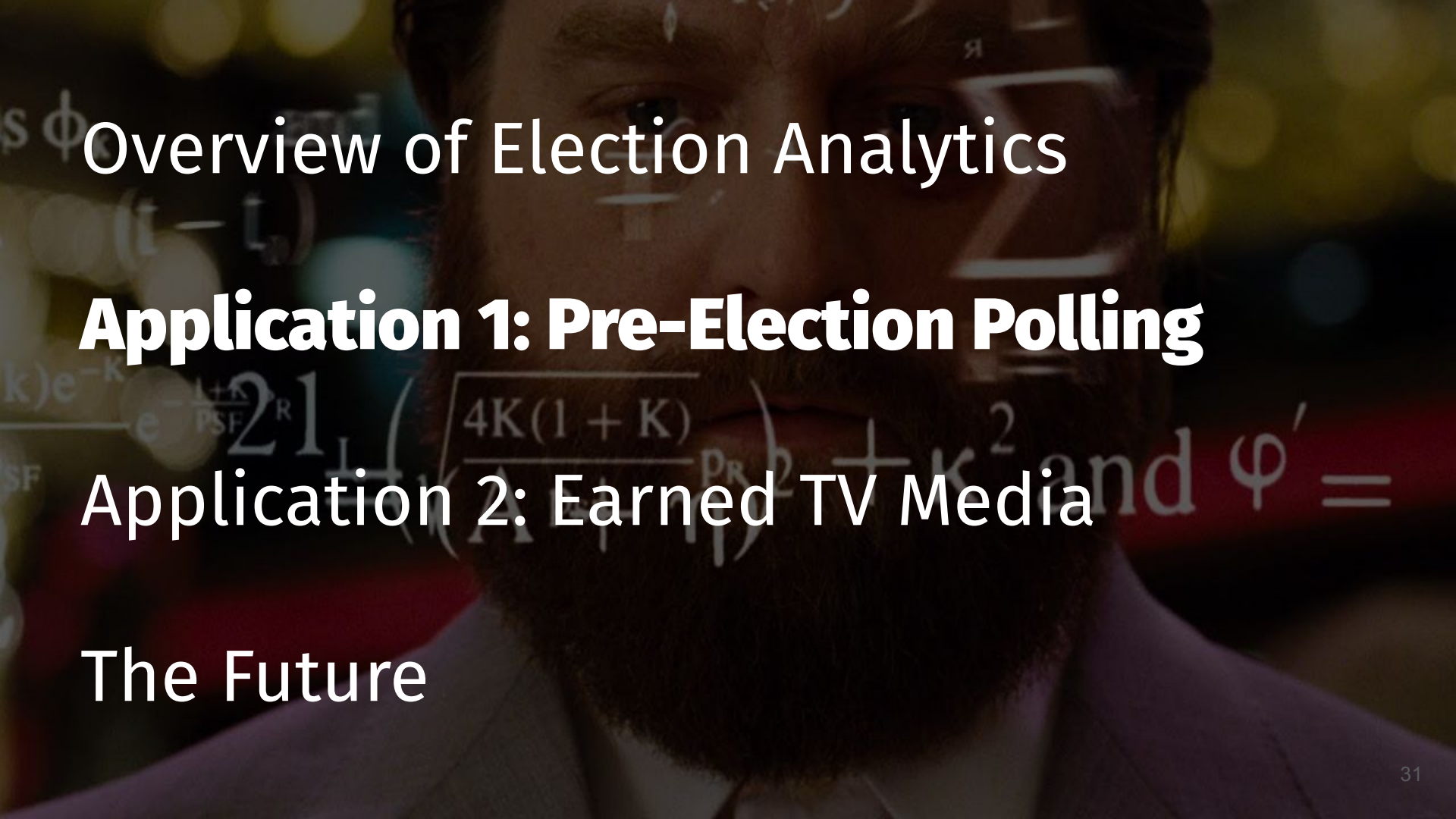
1. The **incumbency advantage** is still very large and explains most election outcomes.

2. Effects of media (e.g. ads, news coverage) are **minimal**.
   - ➢ **Agenda-setting**: Media "may not be successful much of the time in telling people what to think, but it is stunningly successful in telling its readers what to *think about*."

3. Votes are highly **nationalized** with very low levels of **split-ticketing**.
   - ➢ Party affiliation is still, mostly, strongest predictor of attitudes and behavior across offices.

4. But, **persuasion** does work *on the margins*.
   - ➢ Example: deep canvassing (Broockman and Kalla 2016).

5. In competitive elections, things like **candidate quality** can matter.



Overview of Election Analytics

Application 1: Pre-Election Polling

Application 2: Earned TV Media

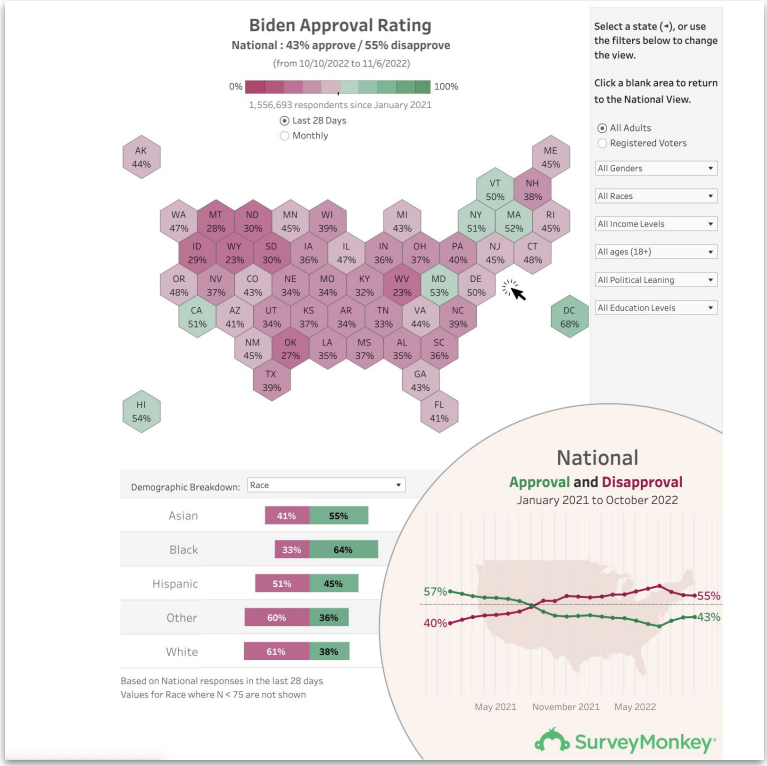
The Future

Pre-election polling in 2022 at SurveyMonkey

Partnership with UPenn Program on Opinion Research and Election Studies (PORES) to deliver weekly midterm pre-election polls to NBC News.

Our “Why”s:

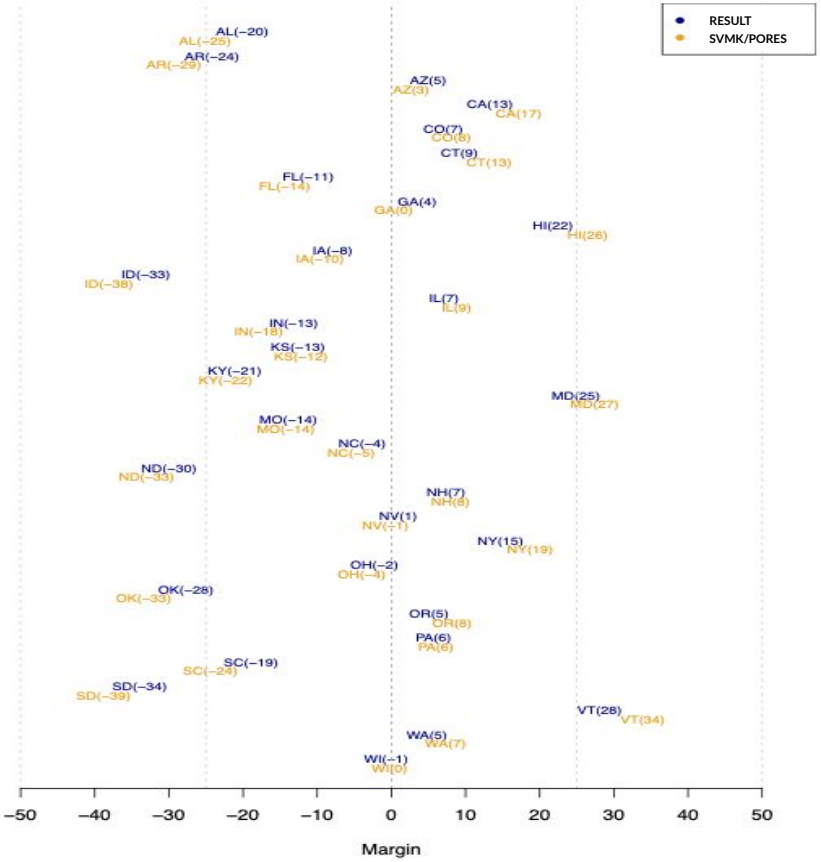
- Content for election news coverage.
- Data for decision desk analysts.
- Public dashboards and banners.
- Evaluate possibility of an online-only exit poll.



A few things
we learned from our polls

We characterized most battlegrounds correctly (hooray!)

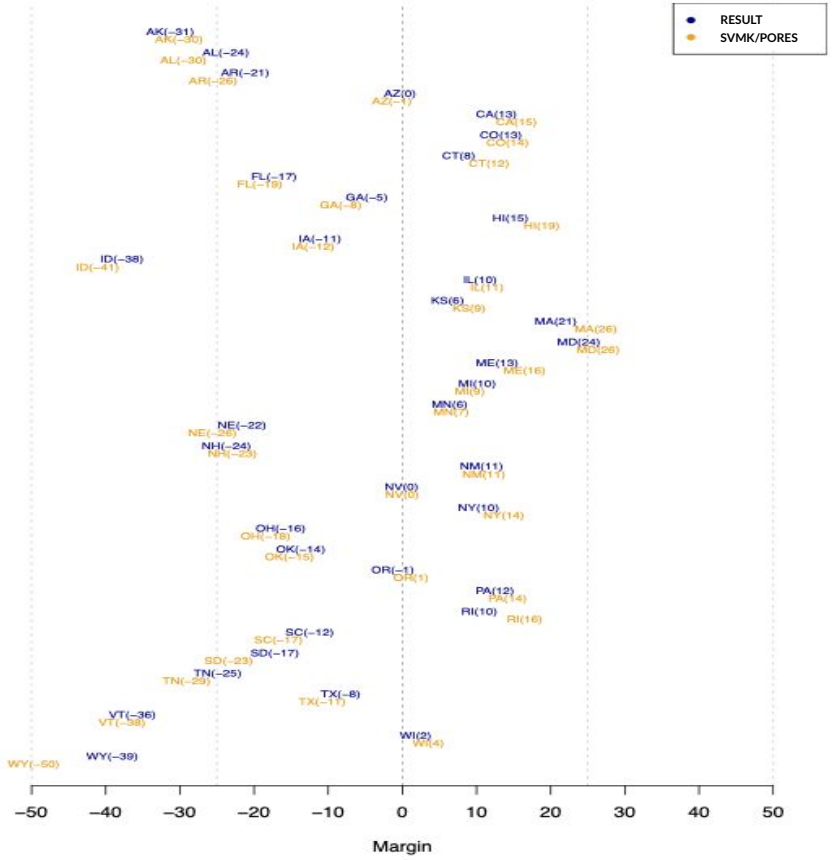
Dem-Rep Margin in Senate Races



Weighted to national likely voters
Responses from 10/26-11/8
Margin of error: 4%

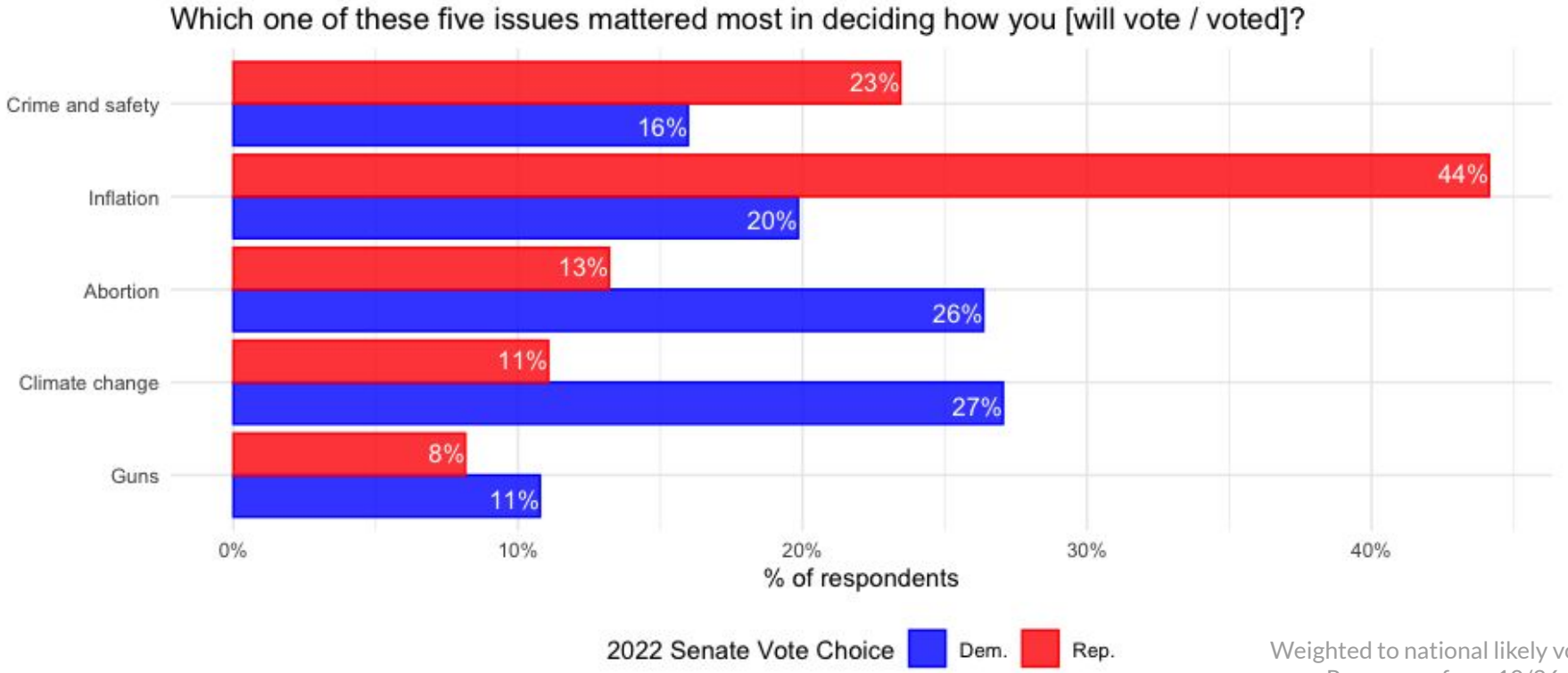
We characterized most battlegrounds correctly (hooray!)

Dem-Rep Margin in Gubernatorial Races



Weighted to national likely voters
Responses from 10/26-11/8
Margin of error: 4%

Climate change, abortion, and inflation were the dominant issues



Weighted to national likely voters
Responses from 10/26-11/8
Margin of error: 4%

A year of sophisticated policy preferences and meaningful split-ticketing … this was not.

SurveyMonkey

Weighted to national likely voters
Responses from 10/26-11/8
Margin of error: 4%

			Senate candiate choice (conslidated)			
		Total	Democratic Party candidate	Republican Party candidate	Did/would not vote	Other candidate
House candidate choice (consolidated)	*Unweighted N*	*43173*	*18926*	*18491*	*1232*	*320*
	Democratic Party candidate	47.7%	95.1%	3.3%	36.4%	53.3%
	Republican Party candidate	52.3%	4.9%	96.7%	63.6%	46.7%
	Did/would not vote	0.0%	0.0%	0.0%	0.0%	0.0%
	Other candidate	0.0%	0.0%	0.0%	0.0%	0.0%
	No answer	0.0%	0.0%	0.0%	0.0%	0.0%

Small, but meaningful split-ticketing against "low-quality" Rep candidates in some races (Georgia)

SurveyMonkey

Weighted to likely voters in **Georgia**
Responses from 10/26-11/8
Margin of error: 4%

7.2% of **Kemp (R)** voters voted for **Warnock (D)**

			Gubernatorial candidate choice (consolidated)				
		Total	Stacey Abrams (Dem)	Brian Kemp (Rep)	Did/would not vote	Other candidate	No answer
Senate candiate choice (conslidated)	*Unweighted N*	*2327*	*919*	*1114*	*91*	*41*	*162*
	Raphael Warnock (Dem)	43.5%	93.3%	7.2%	2.6%	9.0%	5.9%
	Herschel Walker (Rep)	37.6%	1.9%	82.4%	2.1%	17.1%	5.8%
	Did/would not vote	6.4%	1.3%	2.9%	91.9%	6.3%	3.4%
	Other candidate	3.8%	1.1%	4.7%	0.3%	62.5%	2.2%
	No answer	8.7%	2.4%	2.8%	3.0%	5.0%	82.6%

...but split-ticketing trends in Pennsylvania were surprising

SurveyMonkey

Weighted to likely voters in **Pennsylvania**
Responses from 10/26-11/8
Margin of error: 4%

Fewer Mastriano voters voted for Fetterman than did Shapiro voters for Oz

			Gubernatorial candidate choice (consolidated)			
		Total	Josh Shapiro (Dem)	Doug Mastriano (Rep)	Did/would not vote	No answer
Senate candiate choice (conslidated)	*Unweighted N*	*3022*	*1628*	*962*	*170*	*262*
	John Fetterman (Dem)	44.2%	86.0%	1.9%	10.7%	16.2%
	Mehmet Oz (Rep)	38.7%	9.1%	91.9%	7.2%	15.0%
	Did/would not vote	7.1%	1.4%	1.7%	79.5%	1.2%
	Other candidate	0.0%	0.0%	0.0%	0.0%	0.0%
	No answer	10.1%	3.6%	4.5%	2.6%	67.6%

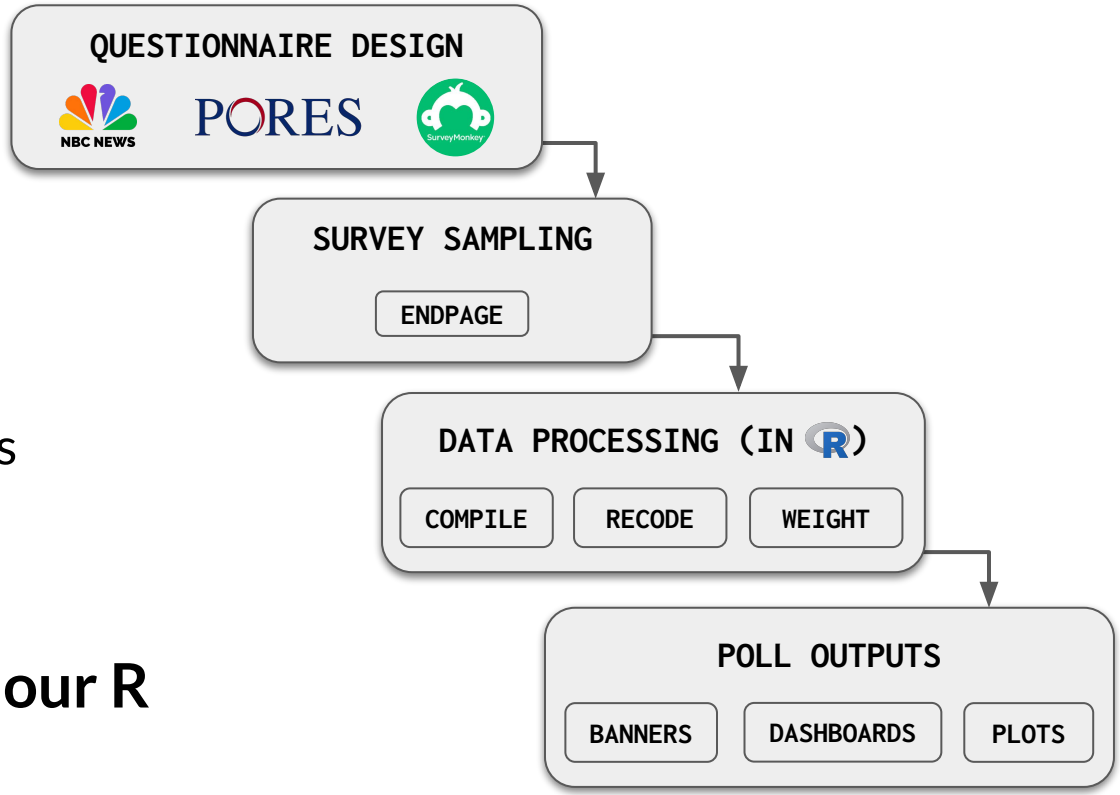
How we do pre-election polling

Our polling workflow at a high level

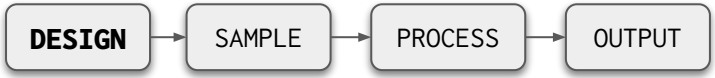
Daily surveys fielded between Jan 20 2021 and Nov 14 2022.

- **>200** total questions
- **1.56 million** total responses
- **2.3k** average daily responses
- **29k** average weekly responses

Many steps automated in our R package `svmkR`.



Tracking midterm opinions



Attitudinal Questions

- **Biden / Trump approval**
- **Issues** (covid, climate, abortion, racism, gun control, inflation, immigration, crime, elections)
  - Issue importance
  - Policy positions
  - Pulse questions (e.g. after Dobbs decision)
- **Battleground Races** (e.g. AZ Gov, PA Sen, GA Sec State, FL Gov)
  - Candidate perceptions (e.g. *do you think Ron DeSantis is too extreme?*)
  - Candidate approval

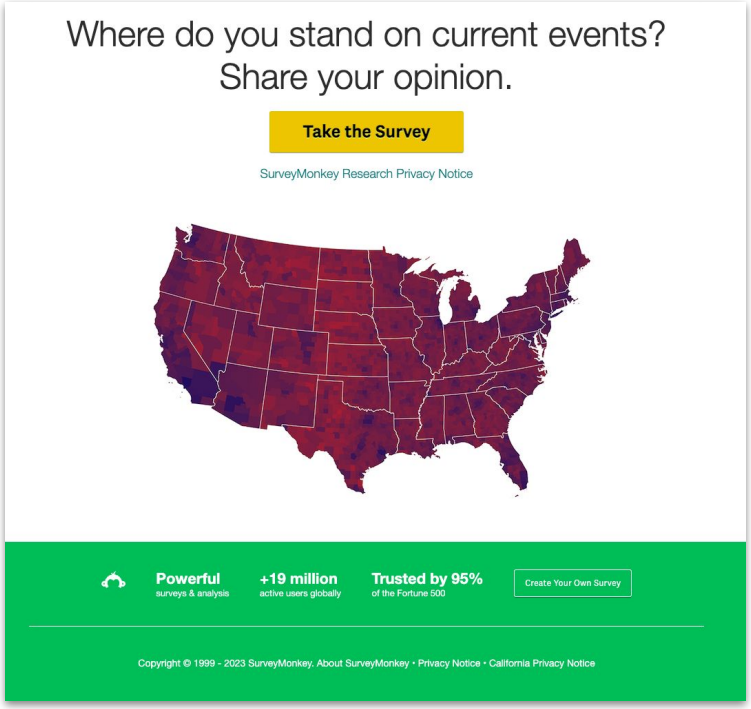
Respondent Identity Questions

- **Standard demographics** (e.g. race, education, gender, sex orientation, income)
- **"Professional" survey-taking** (e.g. *how many surveys did you take today?*)
- **Turnout profile**
  - Self-reported voter registration
  - Self-reported vote mode
  - Self-reported past turnout history
  - Vote intention

SurveyMonkey's End Page

(a.k.a. the special sauce)

DESIGN → **SAMPLE** → PROCESS → OUTPUT



Randomized research invitation on end page for every SurveyMonkey survey.

- 20% of daily end pages directed to our PORES/SVMK poll from Jan 22–Nov 23.

But is this (an online opt-in river sample) any better than a straw poll???

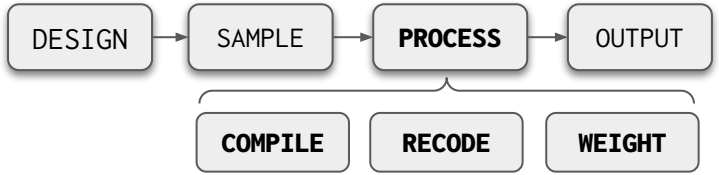
How to evaluate a poll with population benchmarks

Error (Source % - Benchmark %)

	Benchmark	Endpage (n=9,310)	Cint Gen Pop (n=2,003)
Unemployment (CPS)	4%	**+7%**	+12%
Licensed Driver (DOT)	90%	**+3%**	-3%
SSI Recipient (SSA)	3%	**+8%**	+19%
Social Security Recipient (SSA)	21%	**≈0%**	+10%
Medicare Recipient (CMS)	20%	**+15%**	+25%
Cigarette Smoker (NHIS)	12%	**-3%**	+12%
Marital Rate (CPS)	49%	+4%	**-2%**
Divorce Rate (CPS)	10%	+5%	**+4%**
Citizen Rate (CPS)	92%	**+4%**	+5%
Biden Approval (538)	37%	**≈0%**	+5%

Evaluation surveys fielded on Cint and SurveyMonkey Endpage in July 2022. Totals shown here are unweighted.

Processing workflow



Cleaning stack:

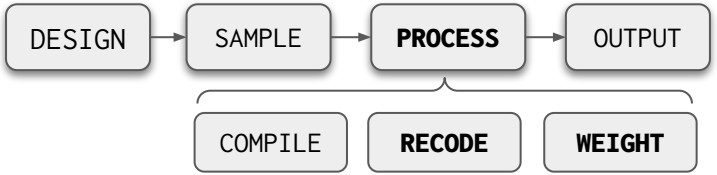
- Large `tidyverse` pipeline to clean and prep data for weighting

How we handle item non-response:

- For *substantive questions* → listwise delete for item-specific outputs.
  - Assumes MCAR, but standard (perhaps archaic) practice in polling.

- For *weighting variables* → re-code as explicit `NA`.
  - Other methods (mean, multiple imputation) are slow or presumptive.
  - Weights and outputs not hugely sensitive to choice of imputation.

How we weight our polls



Target populations are combinations of:

Individual	Geography	Period
○ Adult ○ Likely Voter	○ U.S. National ○ Battleground States	○ Monthly ○ Rolling 28 Day ○ Weekly ○ Daily

Adult target population distributions:

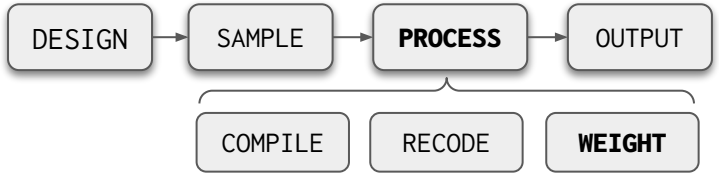
- *ACS marginals*: `Age` (6 pt), `Race` (5 pt), `Gender` (3 pt), `Educ` (6 pt), `Division`
- *ACS joints*: `Gender*Race`, `White*Educ`, `White*Gender*Educ`, `Region*Educ`, `Region*Gender*Age`, `Region*White*Educ`
- *2020 vote*: `Candidate` (2 pt), `Turnout`

Adult weights:

- Single-stage raking (`survey` package), trimmed at 99% percentile

… So you didn't use MRP??

How we weight likely voters



$$\text{Weight}_{LV,g,t} = \text{Weight}_{Adult,g,t} \times \left|\Pr(\text{Vote in 2022} \mid \mathbf{X}) > 0.5\right|$$

geography, period, survey covariates

1. Fit mixed effects models of turnout on validated CES surveys
2. Select model with highest cross-validated accuracy (93%):

```
Vote ~ Female + log(Age) + Educ6 + Employed2 + Income6 + Race5 +
       PartyID3 + Ideology5 + VoteIntent6 + VoteLastElection2 +
       Gender2*log(Age) + RaceWhite*log(Age) + Educ6*Age +
       Gender2*Educ + RaceWhite*Educ6 + PartyID3*Female +
       PartyID3*log(Age) + PartyID3*Educ6 + PartyID3*RaceWhite +
       (1|County:State) + (1|State)
```

Fixed effects

Random effects

3. Predict Pr(Vote in 2022 | **X**) for SVMK/PORES poll respondents.
4. Adjust Pr(Vote in 2022 | **X**) (e.g. state registration deadlines).
5. Manually validate (“eyeball test”) of the 10 most/least likely voters.

Our data products

DESIGN → SAMPLE → PROCESS → **OUTPUT**

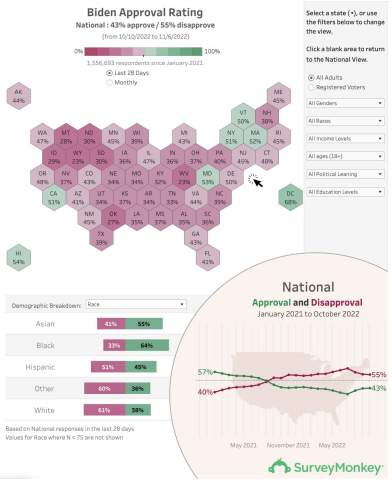
Data

Automated dumps

googledrive

Dashboards (public)

Made in Tableau



Banners

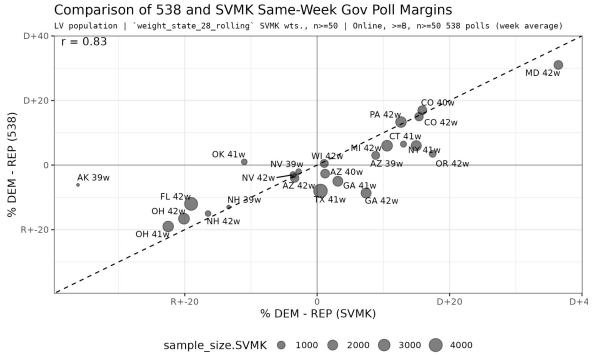
- Built with `expss` and `openxlsx`
- Bootstrapped margins of error

SurveyMonkey

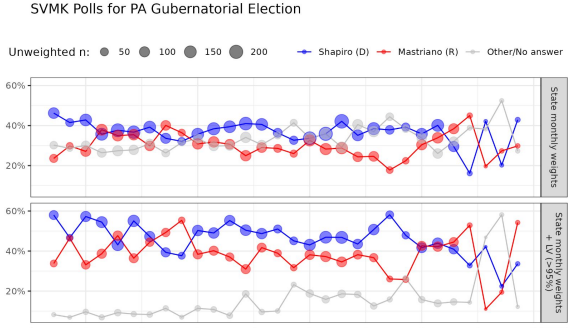
			Vote method (cast / intent)				
		Total	Election-day in-person	Early in-person	Mail/absentee	Will not vote	No answer
Senate candidate choice (consolidated)	*Unweighted N*	*58413*	*19741*	*12520*	*15208*	*2823*	*328*
	Democratic Party candidate	45.8%	38.7%	49.3%	62.0%	16.8%	44.3%
	Republican Party candidate	43.2%	58.6%	43.4%	32.3%	16.1%	35.4%
	Did/would not vote	10.2%	2.5%	6.8%	5.3%	66.9%	20.0%
	Other candidate	0.9%	0.2%	0.5%	0.4%	0.1%	0.2%
House candidate choice (consolidated)	*Unweighted N*	*45503*	*12920*	*8296*	*11117*	*1943*	*243*
	Democratic Party candidate	47.1%	39.4%	51.0%	63.2%	15.4%	47.4%
	Republican Party candidate	45.5%	58.6%	47.3%	35.2%	17.8%	38.0%
	Did/would not vote	7.4%	2.0%	1.6%	1.6%	66.8%	14.5%
	Other candidate	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Gubernatorial candidate choice (consolidated)	*Unweighted N*	*63355*	*21369*	*13938*	*16346*	*3012*	*351*
	Democratic Party candidate	45.3%	38.4%	48.0%	60.2%	16.6%	40.3%
	Republican Party candidate	44.1%	59.1%	45.4%	32.9%	18.7%	39.7%

Visualizations

- Poll comparisons

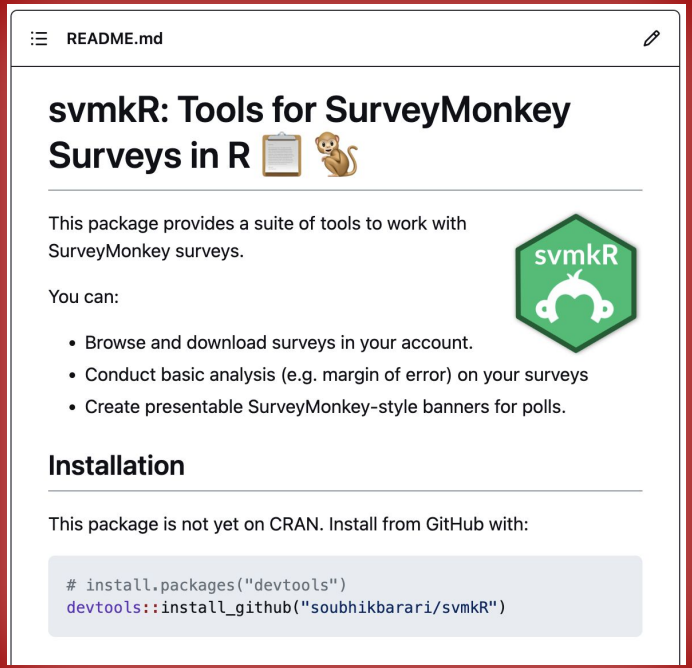


- Question trends



We can automate (most of) our workflow with `svmkR`

github.com/soubhikbarari/svmkR



README.md

svmkR: Tools for SurveyMonkey Surveys in R 📋🐒

This package provides a suite of tools to work with SurveyMonkey surveys.

svmkR

You can:

- Browse and download surveys in your account.
- Conduct basic analysis (e.g. margin of error) on your surveys
- Create presentable SurveyMonkey-style banners for polls.

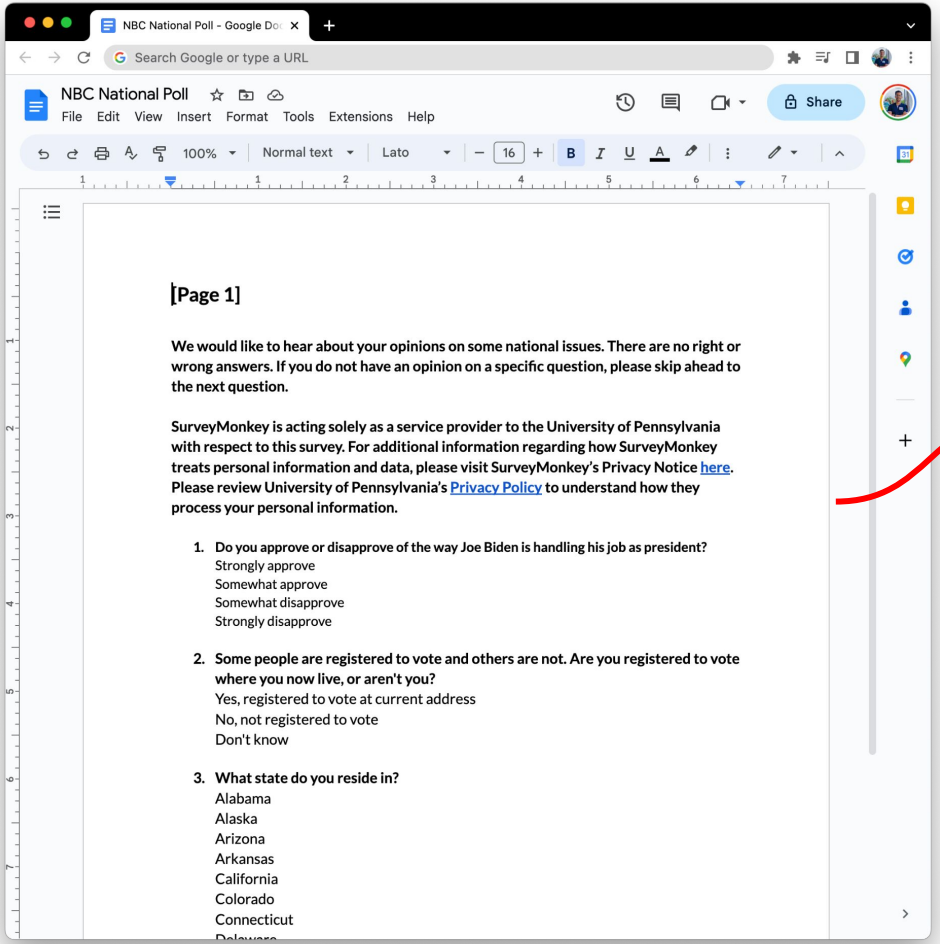
Installation

This package is not yet on CRAN. Install from GitHub with:

```
# install.packages("devtools")
devtools::install_github("soubhikbarari/svmkR")
```

QDOC: A questionnaire markup syntax

DESIGN → SAMPLE → PROCESS → OUTPUT



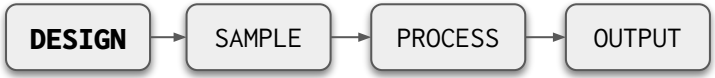

```
[[PageBreak]]

[[Question:Presentation:DescriptiveText]]
We would like to hear about your opinions on some national issues. There are no right or wrong answers. If you do not have an opinion on a specific question, please skip ahead to the next question.

SurveyMonkey is acting solely as a service provider to the University of Pennsylvania with respect to this survey. For additional information regarding how SurveyMonkey treats personal information and data, please visit SurveyMonkey's Privacy Notice here. Please review University of Pennsylvania's Privacy Policy to understand how they process your personal information.

[[Question:SingleAnswer]]
Do you approve or disapprove of the way Joe Biden is handling his job as president?
[[Choices]]
Strongly approve
Somewhat approve
Somewhat disapprove
Strongly disapprove

[[Question:SingleAnswer]]
Some people are registered to vote and others are not. Are you registered to vote where you now live, or aren't you?
[[Choices]]
Yes, registered to vote at current address
No, not registered to vote
Don't know

[[Question:SingleAnswer:Dropdown]]
What state do you reside in?
[[Choices]]
```

nyhackr_demo.R

Source on Save | Run | Source

```
## set up
library(svmkR)
svmkR::set_token("XXX")
```

31:16 (Top Level) R Script

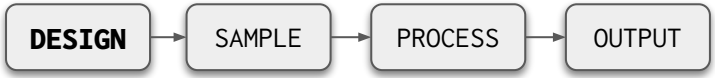


nyhackr_demo.R

Source on Save | Run | Source

```r
## set up
library(svmkR)
svmkR::set_token("XXX")

## read marked-up Google Doc questionnaire
qdoc <- svmkR::read_qdoc(gdoc="docs.google.com/document/XXX")
```

31:16 (Top Level) R Script



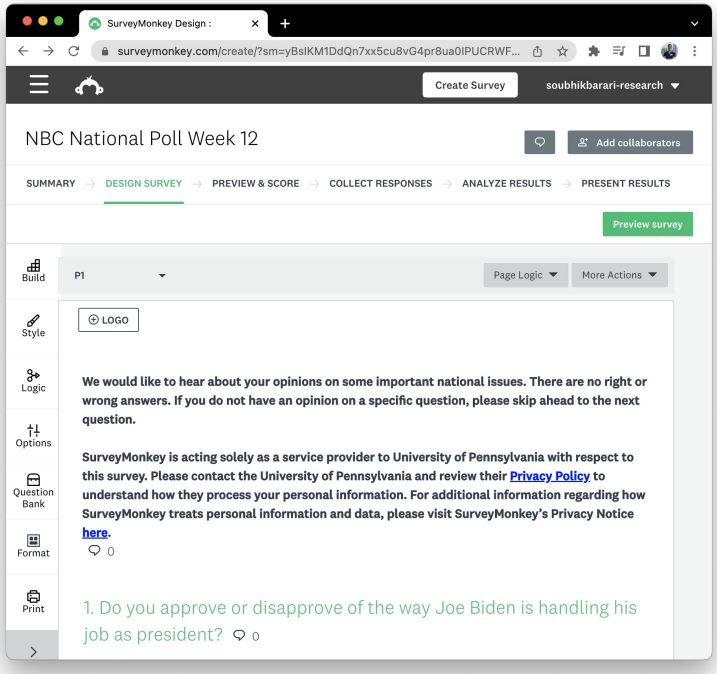
nyhackr_demo.R

Source on Save | Run | Source

```
1  ## set up
2  library(svmkR)
3  svmkR::set_token("XXX")
4
5  ## read marked-up Google Doc questionnaire
6  qdoc <- svmkR::read_qdoc(gdoc="docs.google.com/document/XXX")
7
8  ## upload to account
9  svmkR::upload_qdoc(qdoc, title=nbc_poll_title)
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
```

31:16 (Top Level) R Script

DESIGN → SAMPLE → PROCESS → OUTPUT



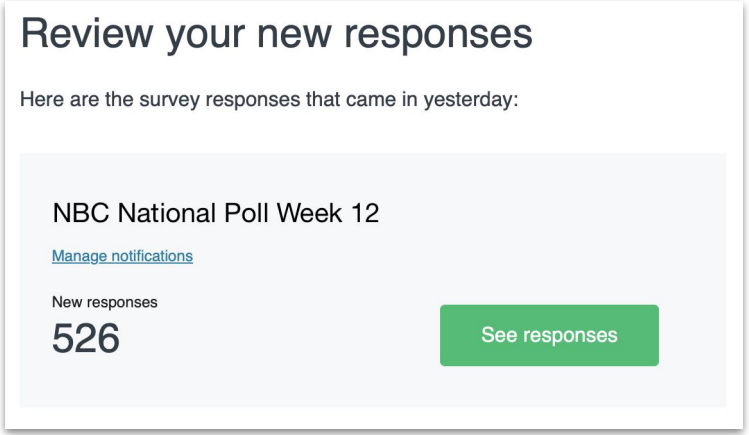
nyhackr_demo.R

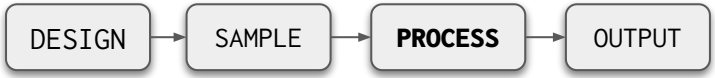
Source on Save | Run | Source

```r
## set up
library(svmkR)
svmkR::set_token("XXX")

## read marked-up Google Doc questionnaire
qdoc <- svmkR::read_qdoc(gdoc="docs.google.com/document/XXX")

## upload to account
svmkR::upload_qdoc(qdoc, title=nbc_poll_title)

## after collecting responses...
```

31:16 (Top Level) R Script

DESIGN → **SAMPLE** → PROCESS → OUTPUT





nyhackr_demo.R

Source on Save | Run | Source

```r
## set up
library(svmkR)
svmkR::set_token("XXX")

## read marked-up Google Doc questionnaire
qdoc <- svmkR::read_qdoc(gdoc="docs.google.com/document/XXX")

## upload to account
svmkR::upload_qdoc(qdoc, title=nbc_poll_title)

## after collecting responses...

## download responses
my_surveys <- svmkR::browse_surveys()
```

31:16 (Top Level) R Script

Console Terminal

R 4.2.0 · ~/Desktop/nyhackr-talk/

```
> my_surveys
# A tibble: 5 × 3
  title                          url                                  id
  <chr>                          <chr>                                <chr>
1 NBC National Poll Week 12      https://api.surveymonkey.com/v3/survey… 4028…
2 SurveyMonkey News Survey       https://api.surveymonkey.com/v3/survey… 5125…
3 SurveyMonkey News Survey       https://api.surveymonkey.com/v3/survey… 4075…
4 Untitled Survey                https://api.surveymonkey.com/v3/survey… 5127…
5 Brand + Methodology Test Survey https://api.surveymonkey.com/v3/survey… 4077…
>
```

DESIGN → SAMPLE → **PROCESS** → OUTPUT

nyhackr_demo.R

Source on Save | Run | Source

```r
## set up
library(svmkR)
svmkR::set_token("XXX")

## read marked-up Google Doc questionnaire
qdoc <- svmkR::read_qdoc(gdoc="docs.google.com/document/XXX")

## upload to account
svmkR::upload_qdoc(qdoc, title=nbc_poll_title)

## after collecting responses...

## download responses
my_surveys <- svmkR::browse_surveys()
nbc_poll_id <- subset(my_surveys, title=="NBC National Poll Week 12")$id
```

31:16 (Top Level) R Script

Console Terminal

R 4.2.0 · ~/Desktop/nyhackr-talk/

```
> my_surveys
# A tibble: 5 × 3
  title                          url                                    id
  <chr>                          <chr>                                  <chr>
1 NBC National Poll Week 12      https://api.surveymonkey.com/v3/survey… 4028…
2 SurveyMonkey News Survey       https://api.surveymonkey.com/v3/survey… 5125…
3 SurveyMonkey News Survey       https://api.surveymonkey.com/v3/survey… 4075…
4 Untitled Survey                https://api.surveymonkey.com/v3/survey… 5127…
5 Brand + Methodology Test Survey https://api.surveymonkey.com/v3/survey… 4077…
>
```

DESIGN → SAMPLE → **PROCESS** → OUTPUT

nyhackr_demo.R

Source on Save | Run | Source

```r
## set up
library(svmkR)
svmkR::set_token("XXX")

## read marked-up Google Doc questionnaire
qdoc <- svmkR::read_qdoc(gdoc="docs.google.com/document/XXX")

## upload to account
svmkR::upload_qdoc(qdoc, title=nbc_poll_title)

## after collecting responses...

## download responses
my_surveys <- svmkR::browse_surveys()
nbc_poll_id <- subset(my_surveys, title=="NBC National Poll Week 12")$id
out <- svmkR::fetch_survey_obj(nbc_poll_id)
```

31:16 (Top Level) R Script

Console Terminal

R 4.2.0 · ~/Desktop/nyhackr-talk/

```
You have 49 requests left today before you hit the limit
You have 48 requests left today before you hit the limit
You have 47 requests left today before you hit the limit
You have 46 requests left today before you hit the limit
You have 45 requests left today before you hit the limit
You have 44 requests left today before you hit the limit
You have 43 requests left today before you hit the limit
You have 42 requests left today before you hit the limit
Time difference of 4.233146 mins
+ Parsing responses ↖
  |++++++++++++++++++++++++++++++++++++++++++++++++++| 100% elapsed=27s
+ Merging responses 🤝
+ Levelling columns 🎚
+ Labelling columns 🏷

DONE 🍾

19750 rows x 56 columns
>
```

DESIGN → SAMPLE → **PROCESS** → OUTPUT

nyhackr_demo.R

Source on Save | Run | Source

```r
## set up
library(svmkR)
svmkR::set_token("XXX")

## read marked-up Google Doc questionnaire
qdoc <- svmkR::read_qdoc(gdoc="docs.google.com/document/XXX")

## upload to account
svmkR::upload_qdoc(qdoc, title=nbc_poll_title)

## after collecting responses...

## download responses
my_surveys <- svmkR::browse_surveys()
nbc_poll_id <- subset(my_surveys, title=="NBC National Poll Week 12")$id
out <- svmkR::fetch_survey_obj(nbc_poll_id)

## select weighting target
svmkR::get_target("us_genpop_acs19")
```

31:16 (Top Level) R Script

Console Terminal

R 4.2.0 · ~/Desktop/nyhackr-talk/

```
> get_target("us_genpop_acs19")
```

DESIGN → SAMPLE → **PROCESS** → OUTPUT

nyhackr_demo.R

Source on Save | Run | Source

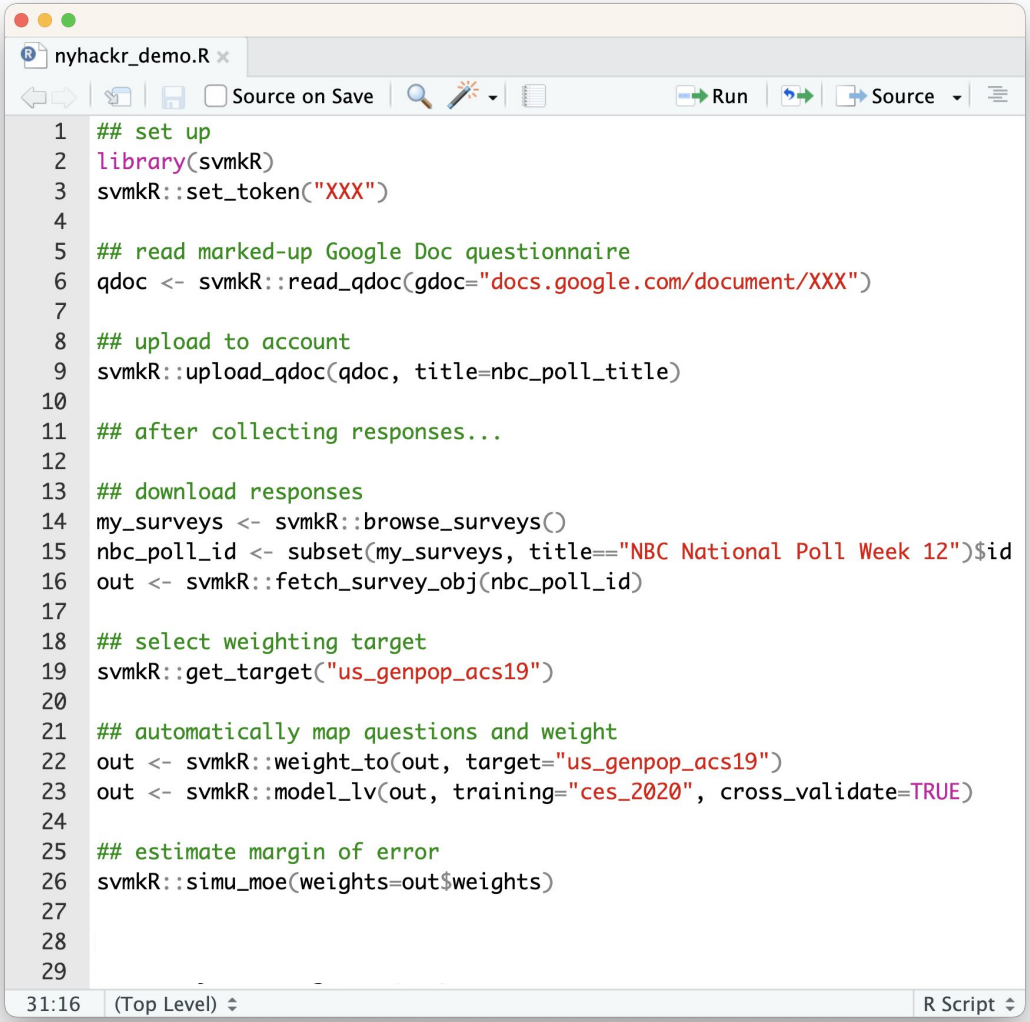
```r
## set up
library(svmkR)
svmkR::set_token("XXX")

## read marked-up Google Doc questionnaire
qdoc <- svmkR::read_qdoc(gdoc="docs.google.com/document/XXX")

## upload to account
svmkR::upload_qdoc(qdoc, title=nbc_poll_title)

## after collecting responses...

## download responses
my_surveys <- svmkR::browse_surveys()
nbc_poll_id <- subset(my_surveys, title=="NBC National Poll Week 12")$id
out <- svmkR::fetch_survey_obj(nbc_poll_id)

## select weighting target
svmkR::get_target("us_genpop_acs19")

## automatically map questions and weight
out <- svmkR::weight_to(out, target="us_genpop_acs19")
out <- svmkR::model_lv(out, training="ces_2020", cross_validate=TRUE)
```

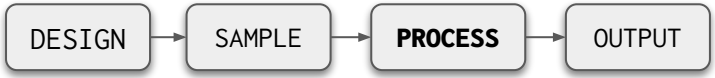
31:16 (Top Level) R Script

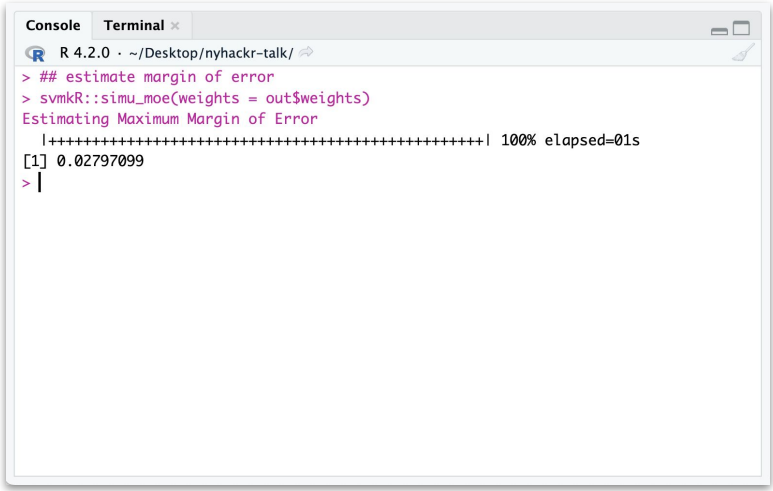
Console Terminal

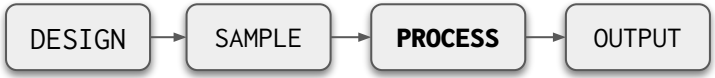
R 4.2.0 · ~/Desktop/nyhackr-talk/

```
> out <- svmkR::weight_to(out, target="us_genpop_acs19")
```


nyhackr_demo.R
Source on Save
Run
Source
set up
library(svmkR)
svmkR::set_token("XXX")
read marked-up Google Doc questionnaire
qdoc <- svmkR::read_qdoc(gdoc="docs.google.com/document/XXX")
upload to account
svmkR::upload_qdoc(qdoc, title=nbc_poll_title)
after collecting responses...
download responses
my_surveys <- svmkR::browse_surveys()
nbc_poll_id <- subset(my_surveys, title=="NBC National Poll Week 12")$id
out <- svmkR::fetch_survey_obj(nbc_poll_id)
select weighting target
svmkR::get_target("us_genpop_acs19")
automatically map questions and weight
out <- svmkR::weight_to(out, target="us_genpop_acs19")
out <- svmkR::model_lv(out, training="ces_2020", cross_validate=TRUE)
estimate margin of error
svmkR::simu_moe(weights=out$weights)
31:16
(Top Level)
R Script


DESIGN
SAMPLE
PROCESS
OUTPUT


Console
Terminal
R 4.2.0 · ~/Desktop/nyhackr-talk/
> ## estimate margin of error
> svmkR::simu_moe(weights = out$weights)
Estimating Maximum Margin of Error
|++| 100% elapsed=01s
[1] 0.02797099
>



nyhackr_demo.R

Source on Save | Run | Source

```r
## set up
library(svmkR)
svmkR::set_token("XXX")

## read marked-up Google Doc questionnaire
qdoc <- svmkR::read_qdoc(gdoc="docs.google.com/document/XXX")

## upload to account
svmkR::upload_qdoc(qdoc, title=nbc_poll_title)

## after collecting responses...

## download responses
my_surveys <- svmkR::browse_surveys()
nbc_poll_id <- subset(my_surveys, title=="NBC National Poll Week 12")$id
out <- svmkR::fetch_survey_obj(nbc_poll_id)

## select weighting target
svmkR::get_target("us_genpop_acs19")

## automatically map questions and weight
out <- svmkR::weight_to(out, target="us_genpop_acs19")
out <- svmkR::model_lv(out, training="ces_2020", cross_validate=TRUE)

## estimate margin of error
svmkR::simu_moe(weights=out$weights)

## clean/make relevant variables
out <- my_cleaning_func(out)
```

31:16 (Top Level) R Script

DESIGN → SAMPLE → PROCESS → **OUTPUT**

nyhackr_demo.R

Source on Save | Run | Source

```
31  ## make banners
32  banners <- out %>%
33    make_banners(row.vars            = list("biden_approval",...),
34                 col.vars            = list("response_date",...),
35                 weight.var          = "weights",
36                 date.var            = "response_date",
37                 total.row.position = "below") %>%
38    write_banners(file.path = "nbc_week_12.xlsx",
39                  title     = "NBC/SVMK Week 12 Midterm Poll") %>%
40    upload_banners(drive.file.name    = "nbc_week_12.xlsx",
41                   drive.folder.path = "drive.google.com/XXX")
42
```

32:19 (Top Level) R Script

SurveyMonkey

			Vote method (cast / intent)				
		Total	Election-day in-person	Early in-person	Mail/absentee	Will not vote	No answer
Senate candidate choice (consolidated)	*Unweighted N*	*58413*	*19741*	*12520*	*15208*	*2823*	*328*
	Democratic Party candidate	45.8%	38.7%	49.3%	62.0%	16.8%	44.3%
	Republican Party candidate	43.2%	58.6%	43.4%	32.3%	16.1%	35.4%
	Did/would not vote	10.2%	2.5%	6.8%	5.3%	66.9%	20.0%
	Other candidate	0.9%	0.2%	0.5%	0.4%	0.1%	0.2%
House candidate choice (consolidated)	*Unweighted N*	*45503*	*12920*	*8296*	*11117*	*1943*	*243*
	Democratic Party candidate	47.1%	39.4%	51.0%	63.2%	15.4%	47.4%
	Republican Party candidate	45.5%	58.6%	47.3%	35.2%	17.8%	38.0%
	Did/would not vote	7.4%	2.0%	1.6%	1.6%	66.8%	14.5%
	Other candidate	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Gubernatorial candidate choice (consolidated)	*Unweighted N*	*63355*	*21369*	*13938*	*16346*	*3012*	*351*
	Democratic Party candidate	45.3%	38.4%	48.0%	60.2%	16.6%	40.3%
	Republican Party candidate	44.1%	59.1%	45.4%	32.9%	18.7%	39.7%

Summary

Challenges for polling at SurveyMonkey and beyond

Did we (all) get lucky in 2022?

- Typically less non-response bias in midterms

FiveThirtyEight

Polls were historically accurate in 2021-22

Weighted-average error of polls in the final 21 days* before presidential primary and presidential, Senate, House and gubernatorial general elections since 1998

	PRESIDENT					
CYCLE	PRIMARY	GENERAL	SENATE	HOUSE	GOV.	COMBINED
1998	–	–	7.5	7.1	8.1	7.7
1999-2000	7.9	4.4	6.0	4.3	4.9	5.5
2001-02	–	–	5.5	5.6	5.2	5.4
2003-04	7.0	3.3	5.3	5.8	5.5	4.8
2005-06	–	–	5.2	6.5	5.1	5.7
2007-08	7.7	3.5	4.7	5.9	4.4	5.5
2009-10	–	–	4.9	7.0	4.7	5.8
2011-12	8.9	3.7	4.7	5.5	4.9	5.3
2013-14	–	–	5.3	6.8	4.5	5.3
2015-16	10.2	4.9	5.0	5.8	5.4	6.8
2017-18	–	–	4.2	4.9	5.2	4.9
2019-20	10.2	5.0	5.8	6.5	6.4	6.3
2021-22	**–**	**–**	**4.8**	**4.0**	**5.1**	**4.8**
All years	9.2	4.3	5.4	6.1	5.4	6.0

Includes polls of special elections and runoffs. Excludes polls from pollsters that are banned by FiveThirtyEight, New Hampshire primary polls taken before the Iowa caucuses and other states' primary polls taken before the New Hampshire primary. Also excludes presidential primary polls if their leader or runner-up dropped out before that primary was held, if any candidate receiving at least 15 percent in the poll dropped out or if any combination of candidates receiving at least 25 percent in the poll dropped out.

Polls are weighted by one over the square root of the number of polls that their pollster conducted for that particular type of election in that particular cycle.

*Based on the poll's median field date.

SOURCES: POLLS, STATE ELECTION OFFICIALS

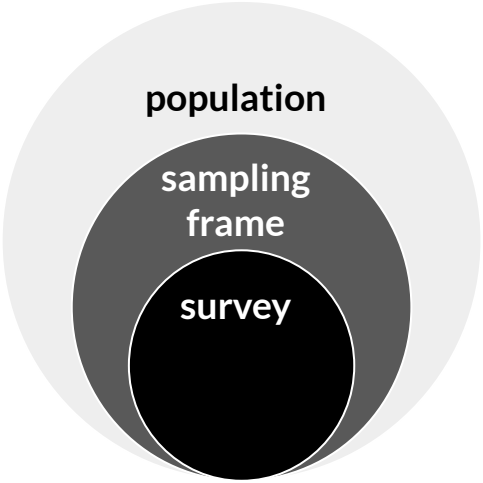
Challenges for polling at SurveyMonkey and beyond

Did we (all) get lucky in 2022?

- Typically less non-response bias in midterms

What is our End Page's sampling frame?

- Need a better understanding, but issues with privacy…



Challenges for polling at SurveyMonkey and beyond

Did we (all) get lucky in 2022?

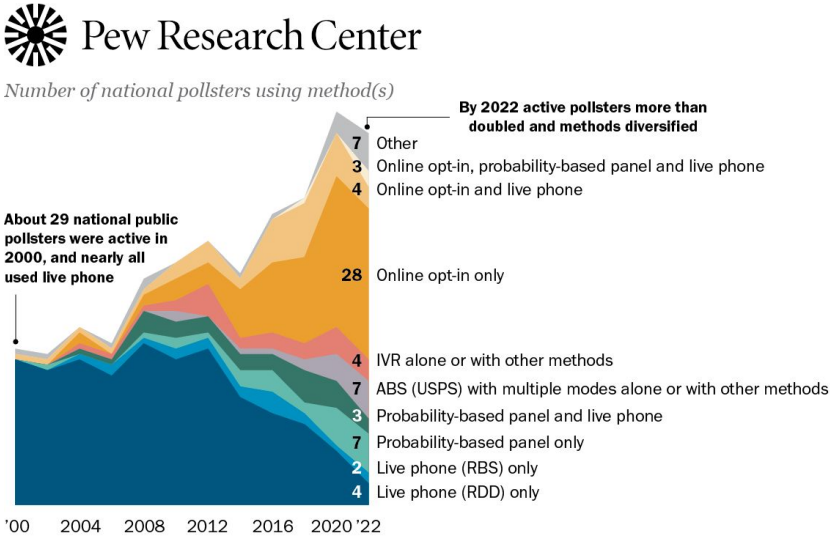
- Typically less non-response bias in midterms

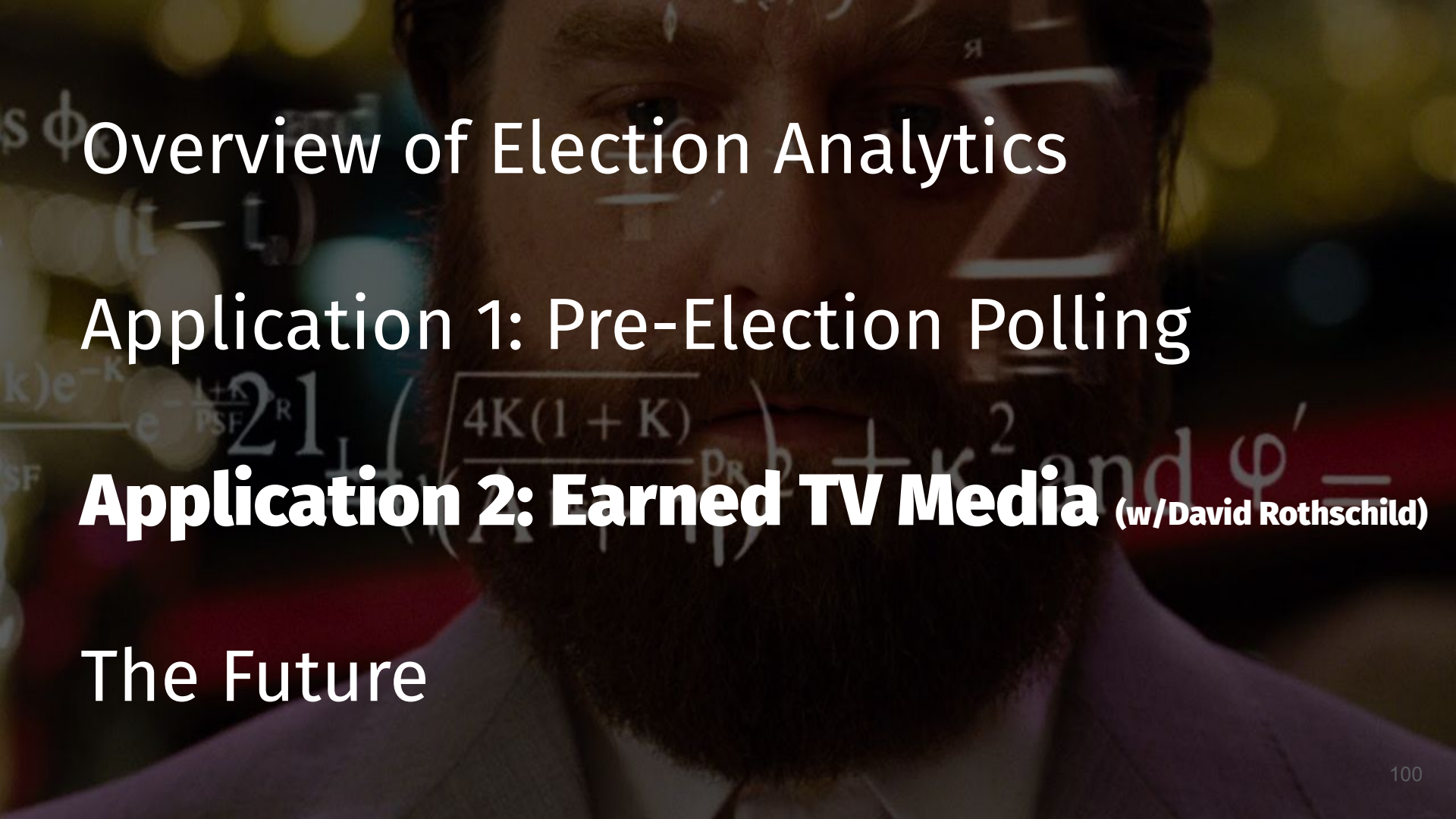
What is our End Page's sampling frame?

- Need a better understanding, but issues with privacy…

Can we blend End Page with other panels/modes?

- So far, relatively few gains ↛ best sticking with one or the other
- Opportunities to (i) dynamically blend or (ii) anchor to a probability sample





Overview of Election Analytics

Application 1: Pre-Election Polling

Application 2: Earned TV Media **(w/David Rothschild)**

The Future

The 2016 election: an enduring puzzle

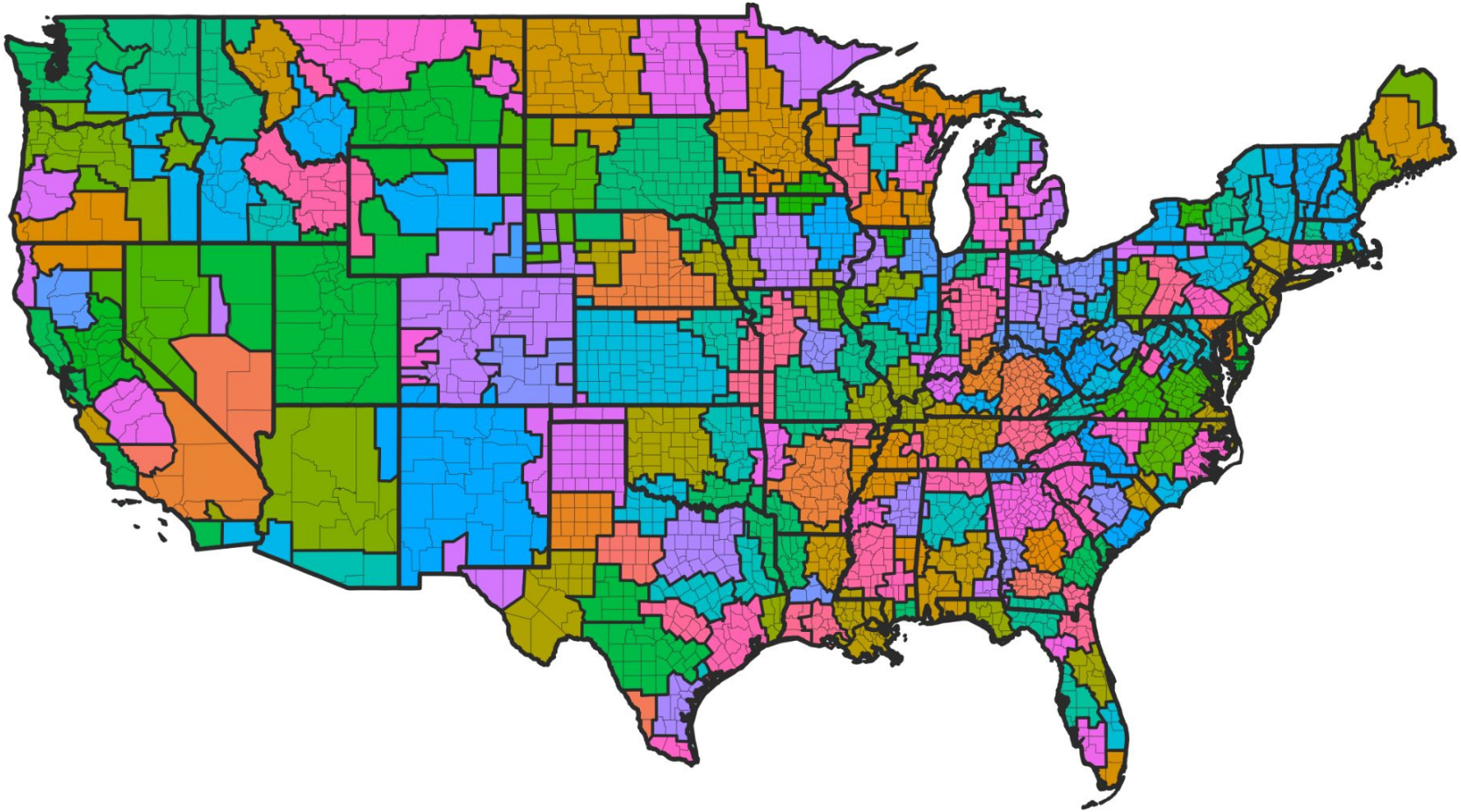
Dominant narrative that free media helped Trump win

- Lots of evidence on “The Fox News Effect”
  (DellaVigna and Kaplan 2007; Hopkins and Ladd 2013; Martin and Yurukoglu 2017; D. Broockman and J. Kalla 2022)
- Yet, little on the campaign effects of broadcast news
  (>10x the nightly viewership of prime-time Fox News)
- Compared to **paid media** (e.g. ads), we know much less about **earned media** (e.g. news coverage) in elections.

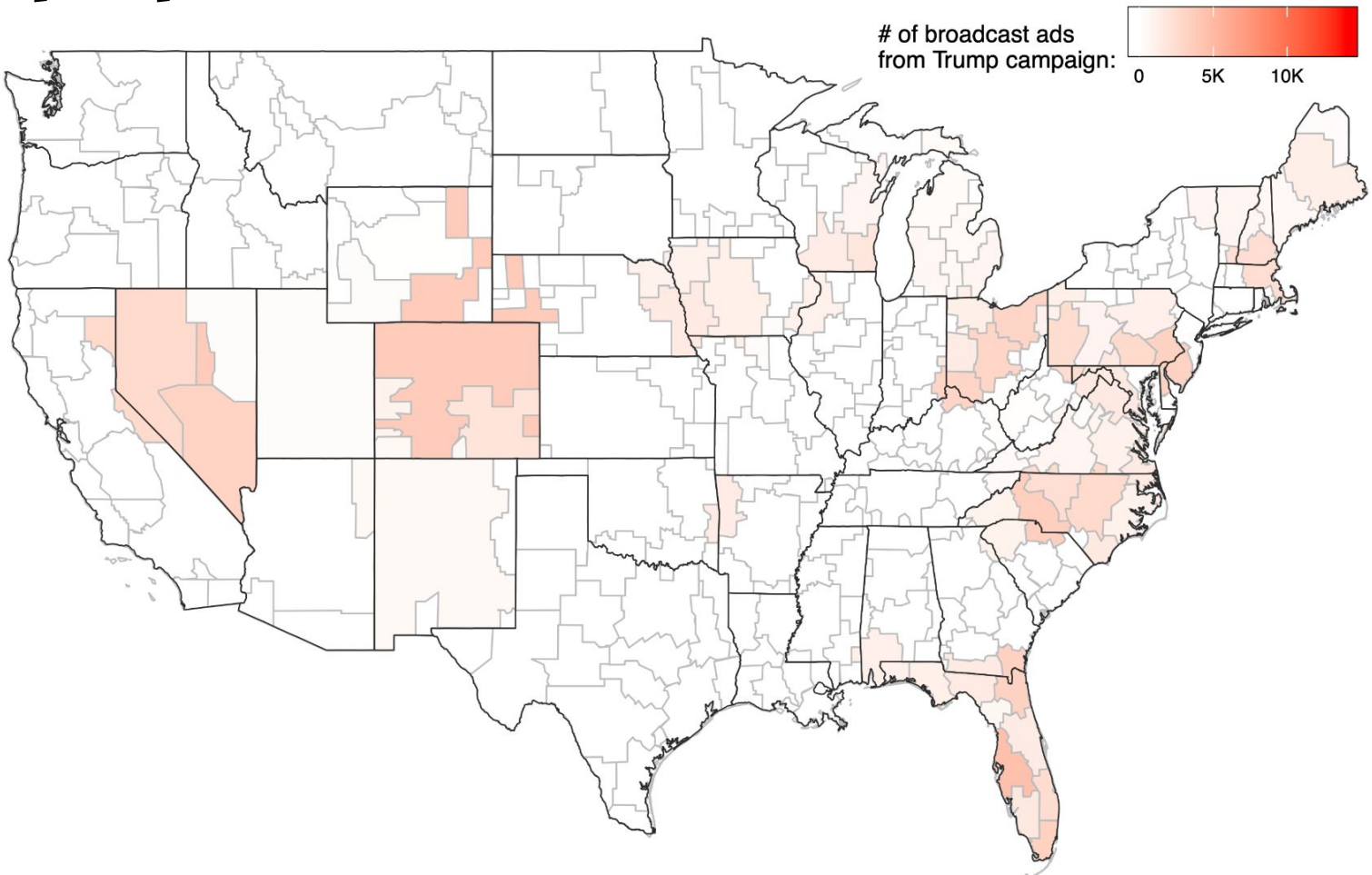
The structure of broadcast media markets

The structure of broadcast media markets

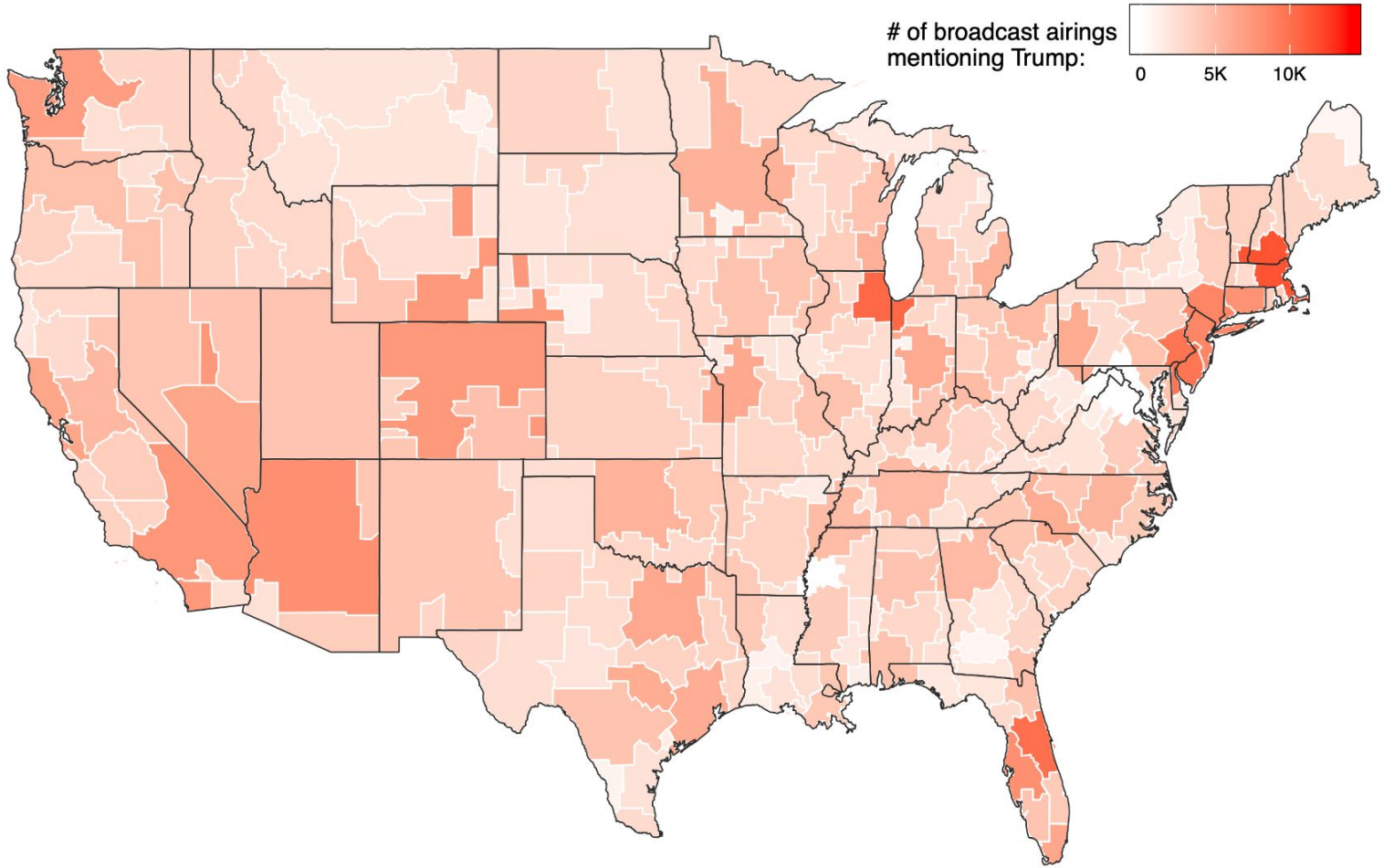
The structure of broadcast media markets



Trump's paid broadcast ads (e.g. Local ABC Affiliates)



Trump's earned broadcast media (e.g. Local ABC Affiliates)



Our research questions

Descriptive:

- Do Republicans have an ad advantage?
- Do candidates named Donald Trump have an advantage?
- Do incumbents have an advantage?

Causal:

- What are the returns from *earned media* relative to *paid media*?
- What are the returns from different kinds of coverage?

Data

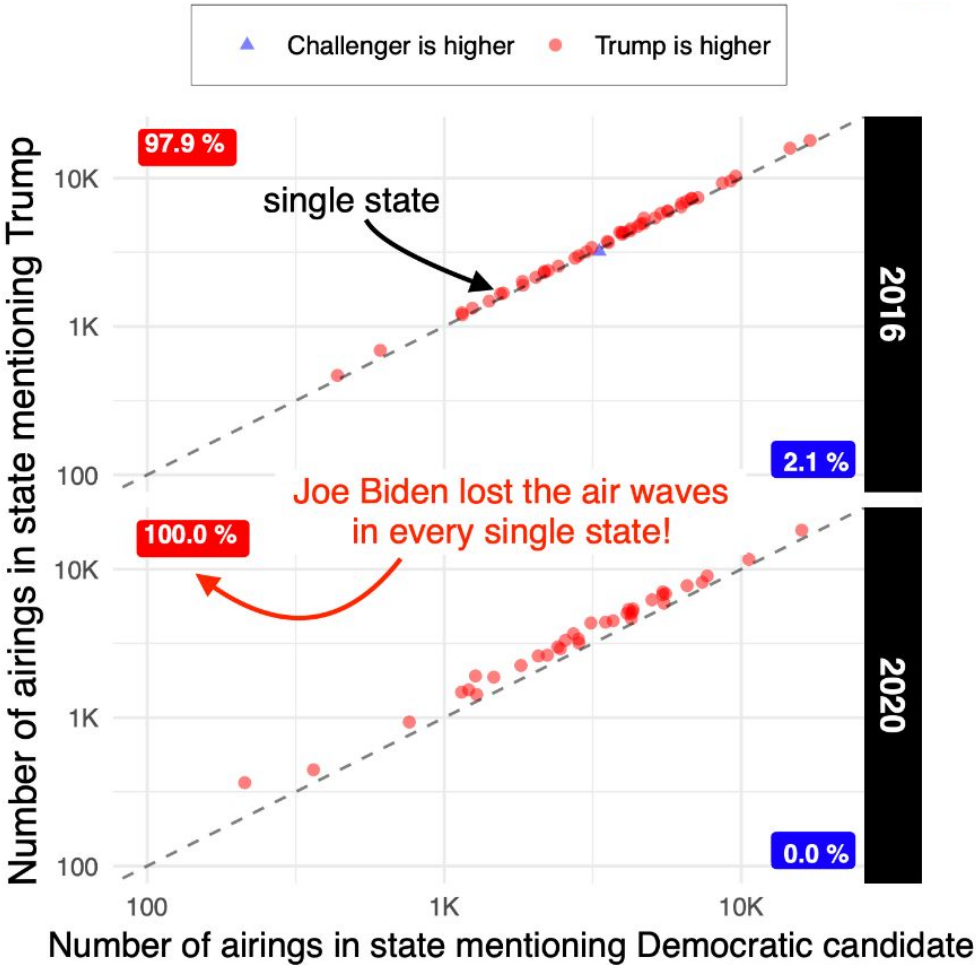
- Comprehensive transcripts of media market-level **broadcast news mentions:**

	House	Senate	President
Cycles	2014, 2016 2018, 2020	2014, 2016 2018, 2020	2016 2020
Candidates	1,214	227	3
Competitive Races	241	53	21 (states)
Total Unique Observations (Last Two Months):			
Media Markets	209	209	209
News Programs	16,081	18,648	28,330
News Airings	841,669	1,138,784	3,287,131
Ad Airings	1,701,568	947,151	263,443
Voting Counties	3,075	3,107	3,114
Border Voting Counties	1,988	2,006	2,007
Mean Per Candidate (Last Two Months):			
Media Markets	≈2	≈7	206
News Programs	105	1,117	212,350
News Airings	113	1,231	236,862
Ad Airings	2,193	5,176	131,154
Impressions Per News Airing	329,267	686,173	2,771,000
Voting Counties	≈9	≈66	3,023
Border Voting Counties	≈7	≈44	1,962

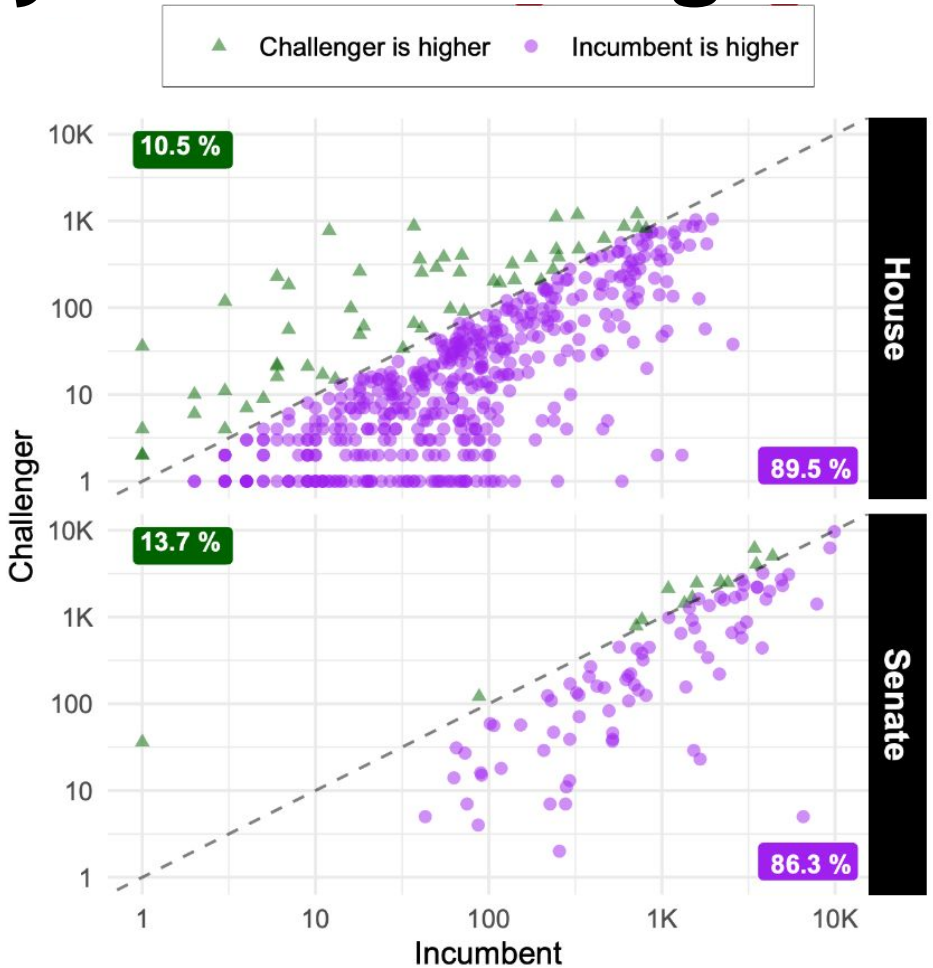
- **Tone** coded using Lexicoder Sentiment Dictionary.
- **Categories** (e.g. position-taking) hand-coded and validated.
- **Vote** returns at county level from `ourcampaigns.com`.

Descriptive results:
Who gets media coverage?

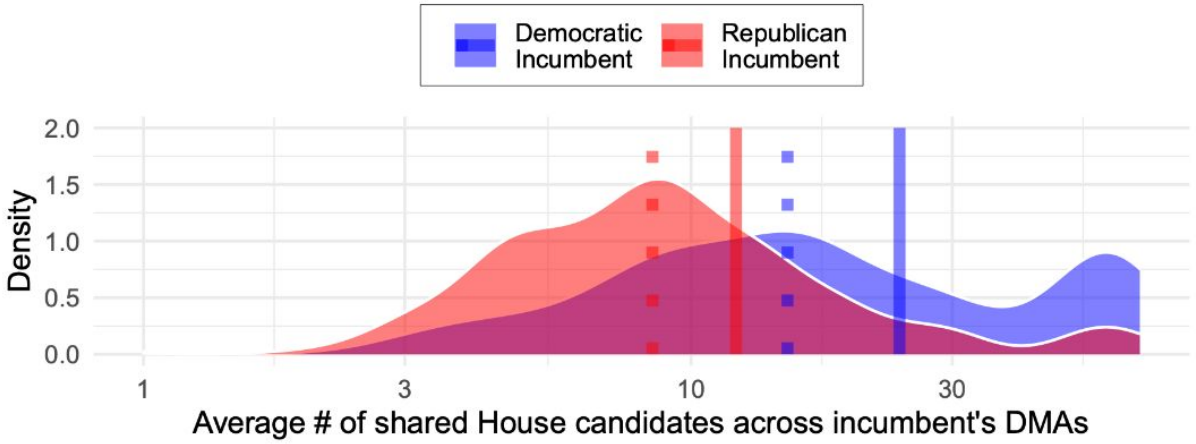
Trump's earned broadcast media advantage



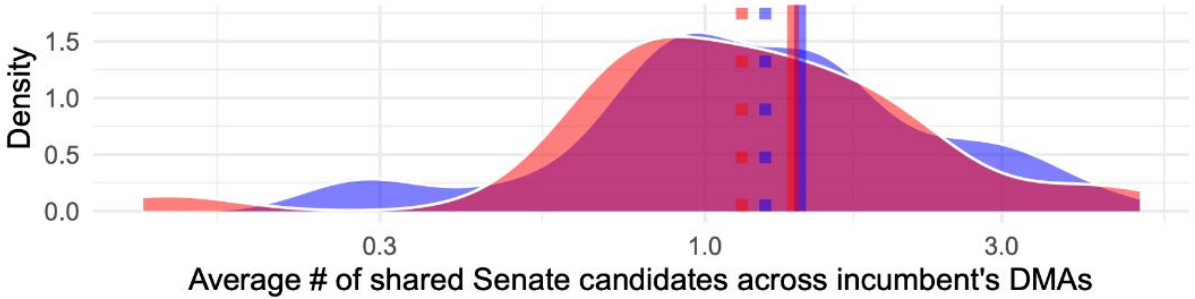
Incumbency media advantage in Congress



GOP incumbents are in less saturated markets

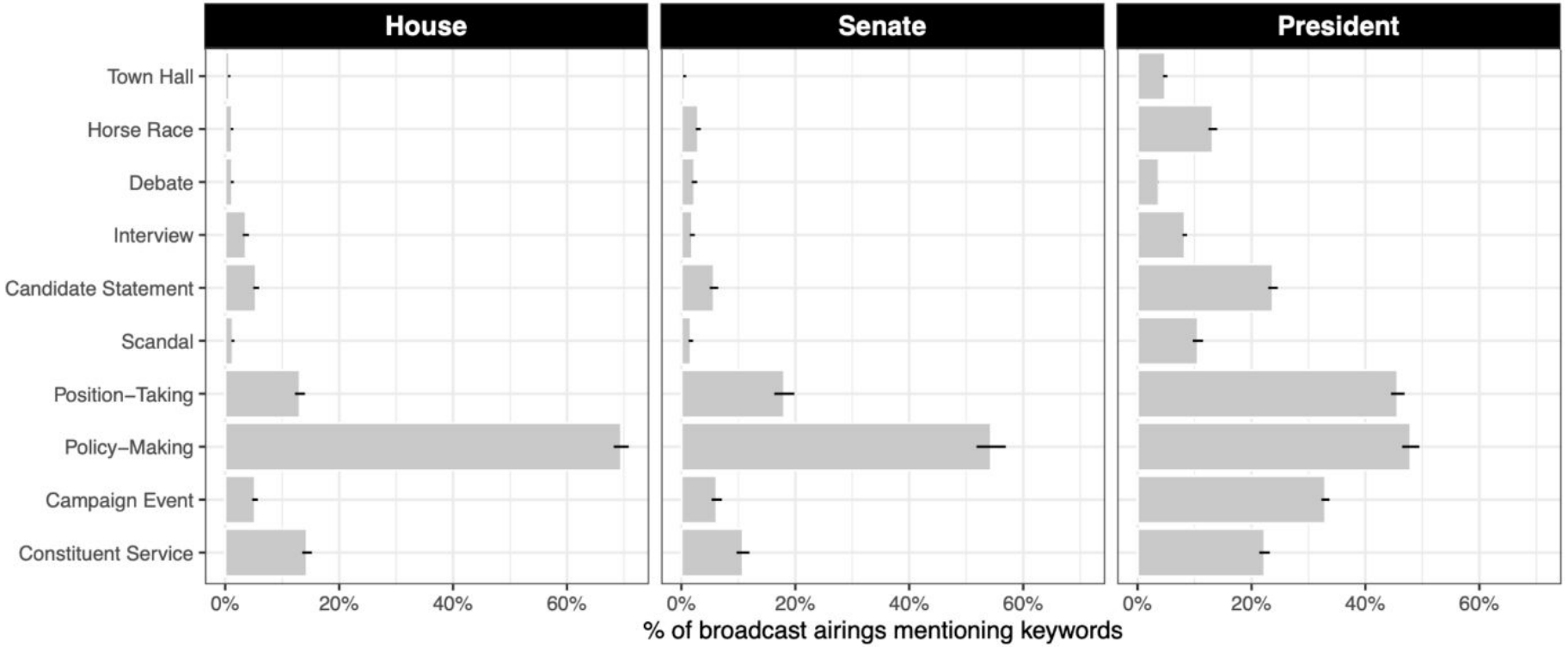


(a) **House Incumbents**



(b) **Senate Incumbents**

Categories of media coverage by office



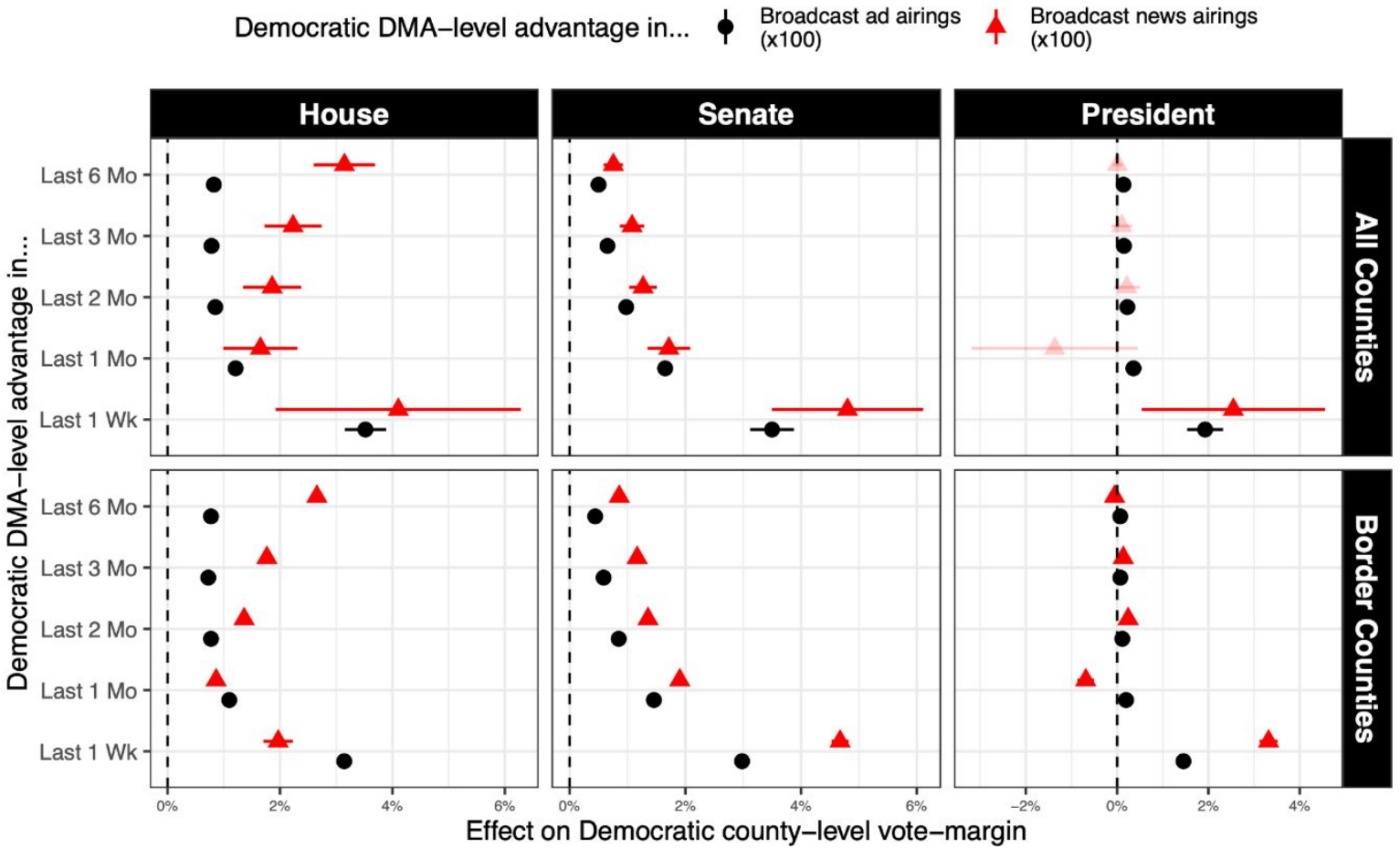
Causal results:
Does media coverage matter?

Research design (simplified)

We regress county-level Democratic vote margin on market-level earned media margin (following from Sides, Vavreck, and Warshaw 2022):

- To deal with:
  - Time-invariant county-level confounders
    - ⇝ county fixed effects.
  - Time-varying state-/national-level confounders
    - ⇝ state-year fixed effects.
  - Unobserved time-varying county-level confounders
    - ⇝ separate estimates for counties on market borders.
  - Cluster standard errors at different levels (market-year, border pair-year, states).
- Additional robustness checks for lag/lead effects.
- Control for level of ad spending in that media market.

Does earned media matter? A little, but often more than ads.



Summary

What we've learned so far

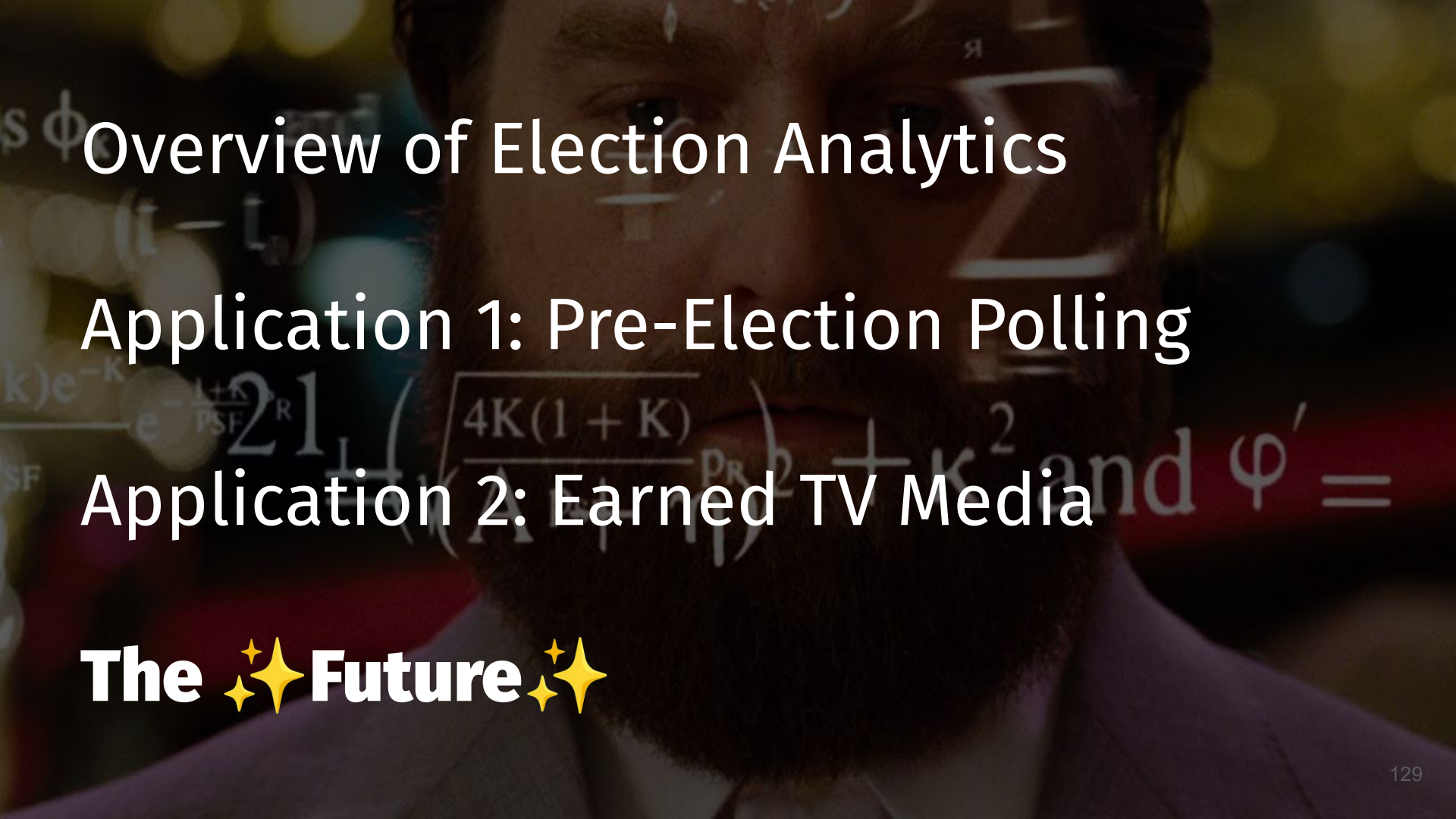
Large detectable advantages for Republicans and incumbents in earned media.

- Some of this comes from geographic advantages in district-market overlaps.
- Competitive races have much more earned media overall.*

Small effects of earned media on vote share.

- But often bigger than ads.
- Effects are largely zero at Presidential level.
- Returns are lower in the races with lots of earned media (competitive races).*
- Returns are higher from substantive coverage about constituent service, policy-making.*

*Not covered today, but covered in forthcoming paper.



Overview of Election Analytics

Application 1: Pre-Election Polling

Application 2: Earned TV Media

The ✨Future✨

3 emerging trends in election analytics

1. **The generative AI revolution(?)**
   - ➢ Already used to generate fund-raising emails (NYT 2023)
   - ➢ Potential for chatbots to simulate "deep canvassing" (Velez and Liu 2023; Velez 2023)
2. **Expanding scope (issues and groups) of analytics**
   - ➢ Example: AAPI voters (turnout up by 20% in 2022 from 2020)
3. **More analytics at the local and state levels**
   - ➢ Less down-ballot "partisan calcification" = more opportunity to shape issues.
   - ➢ Republicans have a concerted strategy up and down the ballot … Democrats?
   - ➢ More data about **you** at the local level (e.g. `LocalView`)

Let's keep the conversation going!

@soubhikbarari www. soubhikbarari.com sbarari@g.harvard.edu