

# Predicting House Election and Voting Patterns with Polling Data

Group: Debbie Chung, Wilson Lui, Carolyn Morris, Kejia Shi\*  
Mentors: David Rothschild and Tobias Konitzer (Microsoft Research);  
Tian Zheng (Columbia University)

April 27, 2018

# Table of Contents

- Introduction
- Exploratory Polling Data Analysis
- Methodology: Why Use MRP Models?
- Model: District-level Election Prediction
- Extensions

\* The authors would like to be considered contributed equally.

# Introduction: 2016 Election



THE UPSHOT | 2016 Election Forecast: Who Will Be President?



## Hillary Clinton has an 85% chance to win.

Last updated Tuesday, November 8 at 10:20 PM ET

CHANCE OF WINNING



85%

Hillary Clinton

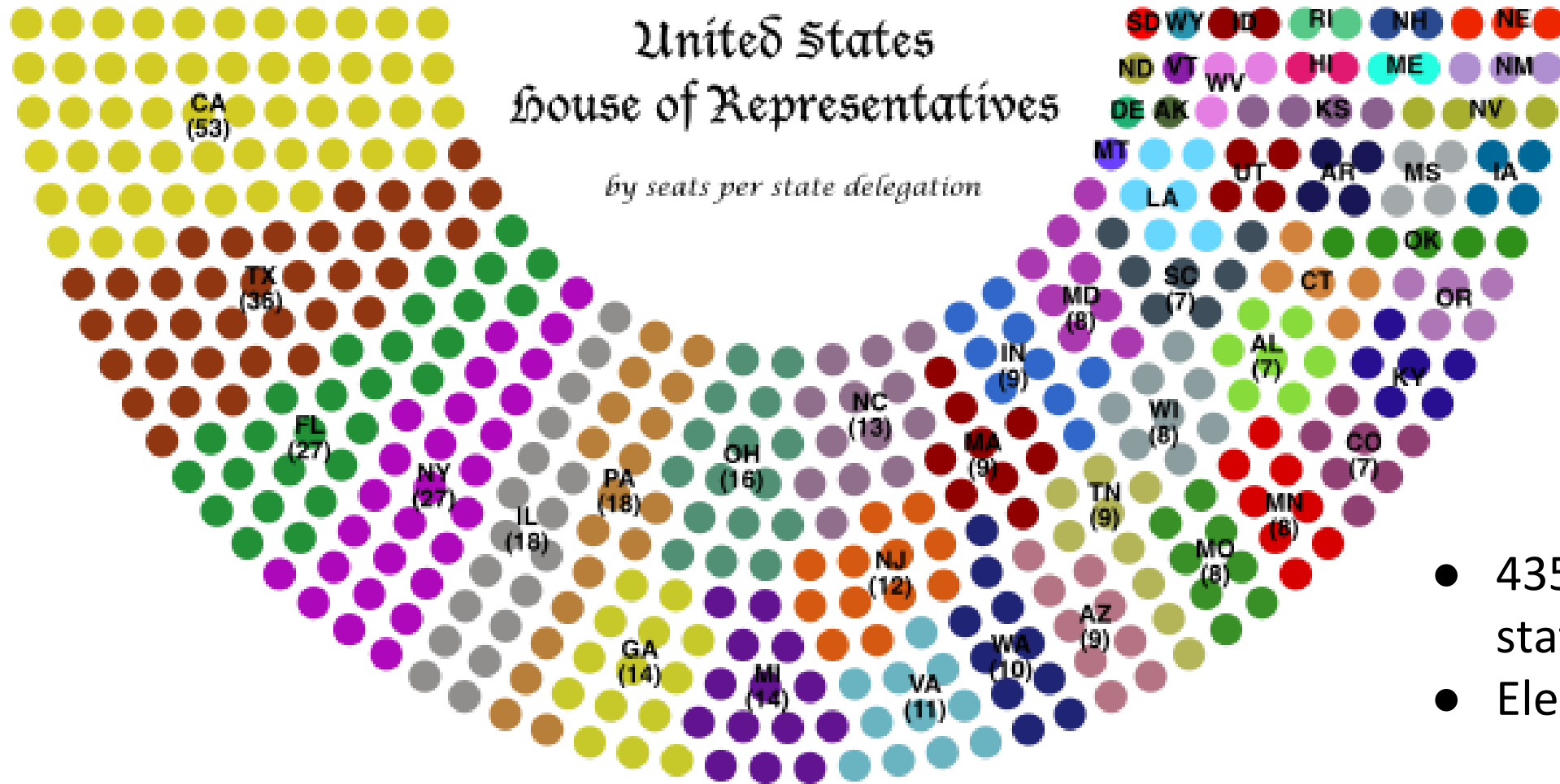


15%

Donald J. Trump



# Introduction: 2018 House of Rep. Election



- 435 Seats, allocated by state's population
- Elections every 2 years

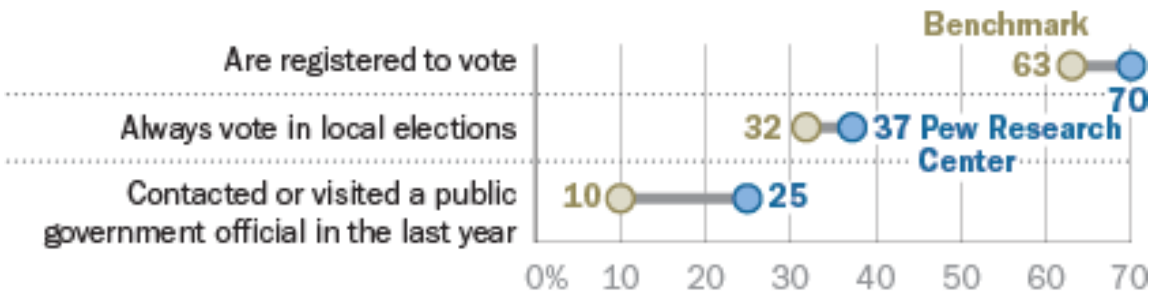
# Difficulties: Declining Response Rates

Response rate by year (%)



## Politically active adults overrepresented in Pew Research Center surveys

% who say they ...



# Difficulties: Selection of Survey Channels

A screenshot of a mobile app interface for a Pollfish survey. The question is "Compared to other recent presidents, is President Trump appropriate?". Below the question is a red button labeled "SELECT ONE". There are five radio button options: "Yes, very appropriate", "Yes, appropriate", "No, inappropriate", "No, very inappropriate", and "Don't know". A circular refresh icon is on the right side of the options. The Pollfish logo is at the bottom right of the app interface.

Q1 —

Compared to other recent presidents, is President Trump appropriate?

SELECT ONE


☐ Yes, very appropriate

☐ Yes, appropriate

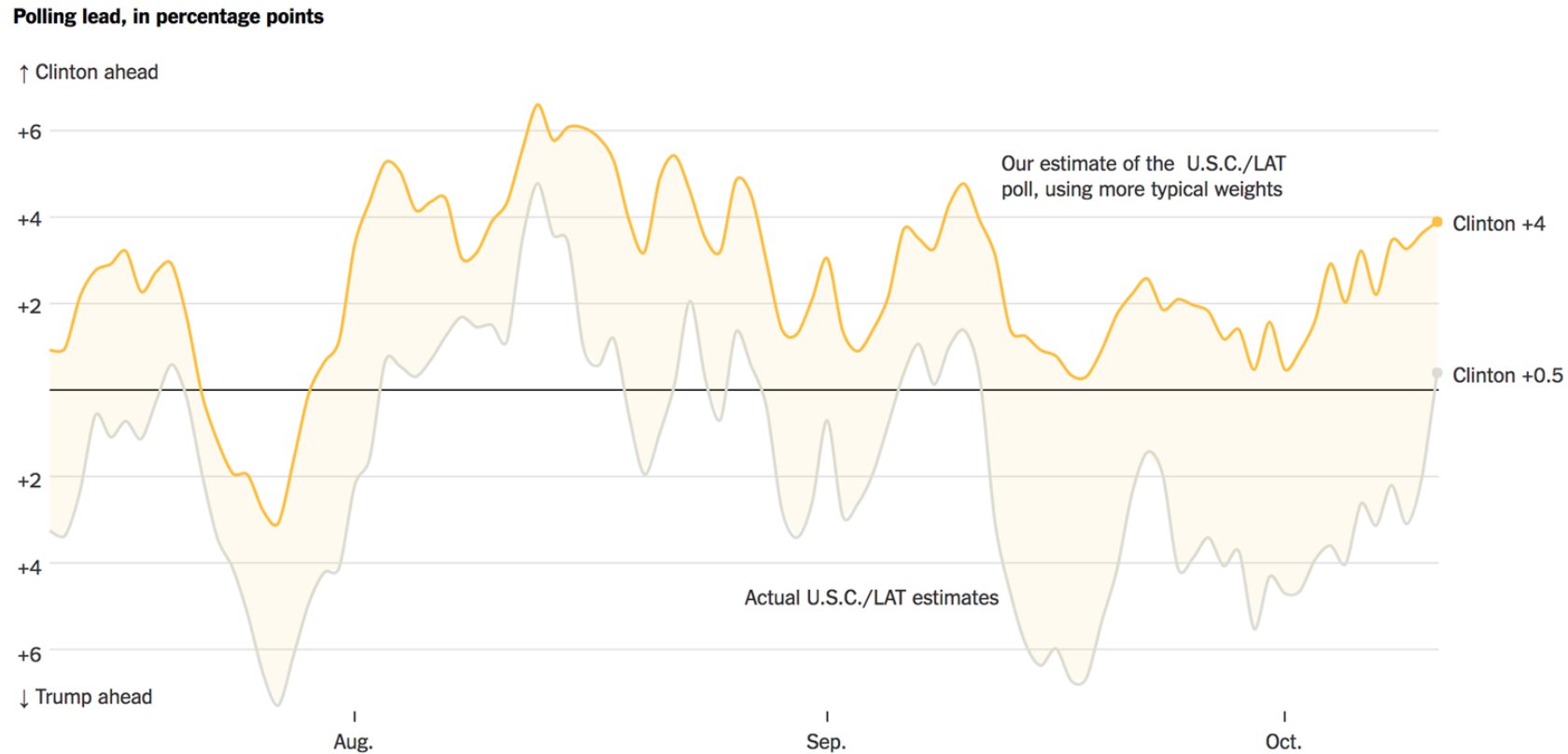
☐ No, inappropriate

☐ No, very inappropriate

☐ Don't know



# Difficulties: Weight Adjustment for Nonresponse



# Methodology: Dealing with Nonresponse

- Ignore the questionnaire-unconscious people completely (unweighted mean)
  - **DANGEROUS!**
  - Ideal nonresponse pattern: similar to population in every dimension
  - Bias: smaller if mean of the nonrespondents is close to mean of the respondents, but we have no idea.
    - Increasing sample size won't reduce nonresponse bias!



# Methodology: Dealing with Nonresponse

- Weighting methods - to reduce such bias  $\sum_{i \in S} w_i y_i$ 
  - Sampling weights: reciprocals of inclusion probabilities  $w_i = 1/\pi_i$
  - Stratification weights: if unit  $i$  belongs to stratum  $h$   $w_i = N_h/n_h$
  - Adjustment for nonresponse:

$$P(\text{respond } i \text{ selected}) = P(\text{unit } i \text{ in selected sample})P(\text{unit } i \text{ respond})$$

- Poststratification (Little 1993, Lohr 2010)
  - Modify the weights so that the sample is calibrated to population counts in the poststrata
  - Raking
    - Repeatedly estimating weights across each variable in turn until the weights converge
    - Force the survey totals to match the known population totals by assigning weights
    - Additional assumption: **the response probabilities depend only on the row and column and not on the particular cell**

# Methodology: Dealing with Nonresponse

- Raking Difficulties (continued)
  - **The algorithm may not converge if some of the cell estimates are zero.**
  - **“overadjustment” Danger** — if there is little relation between the extra dimension in raking and the cell means, raking can **increase the variance** rather than decrease it
- Other techniques
  - Imputation
  - Multilevel Regression and Poststratification (selected method)

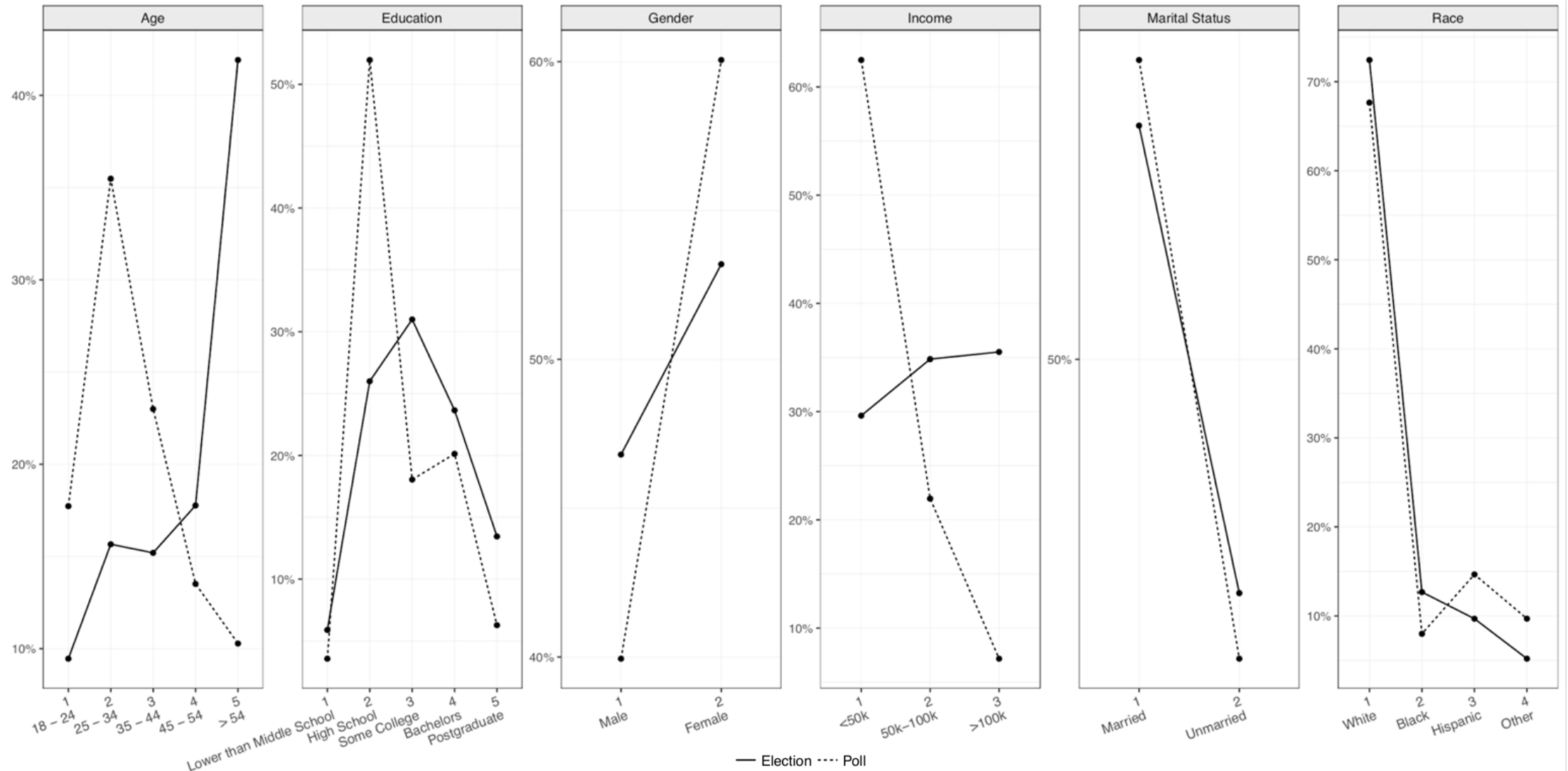
# Dataset

## About Our Data

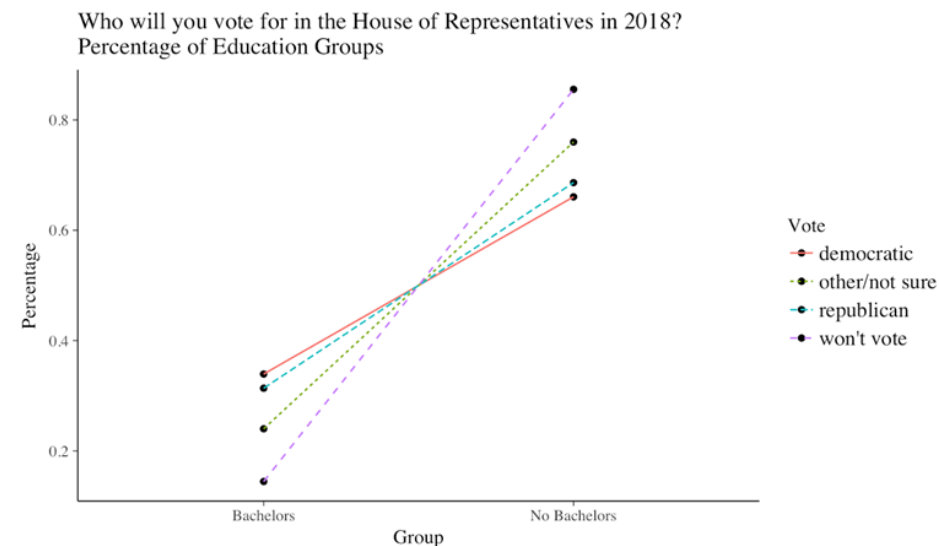
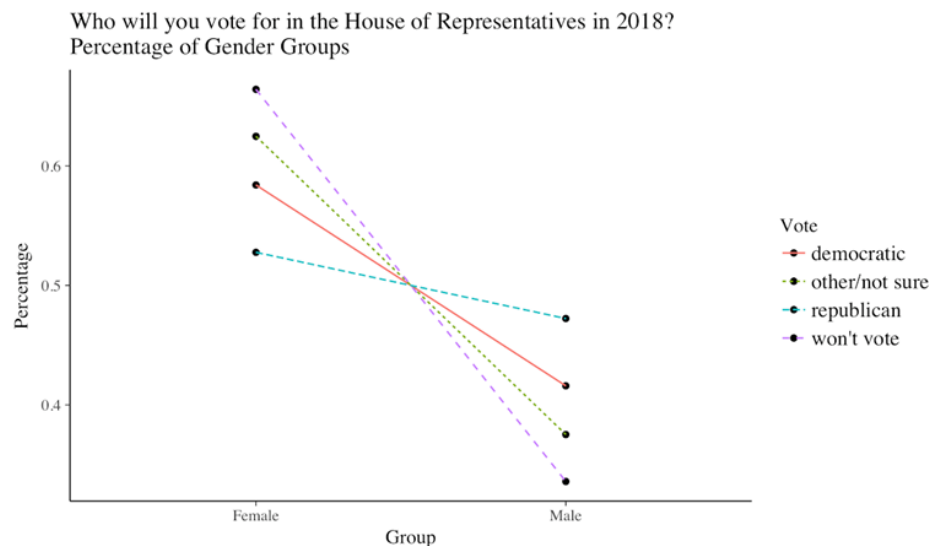
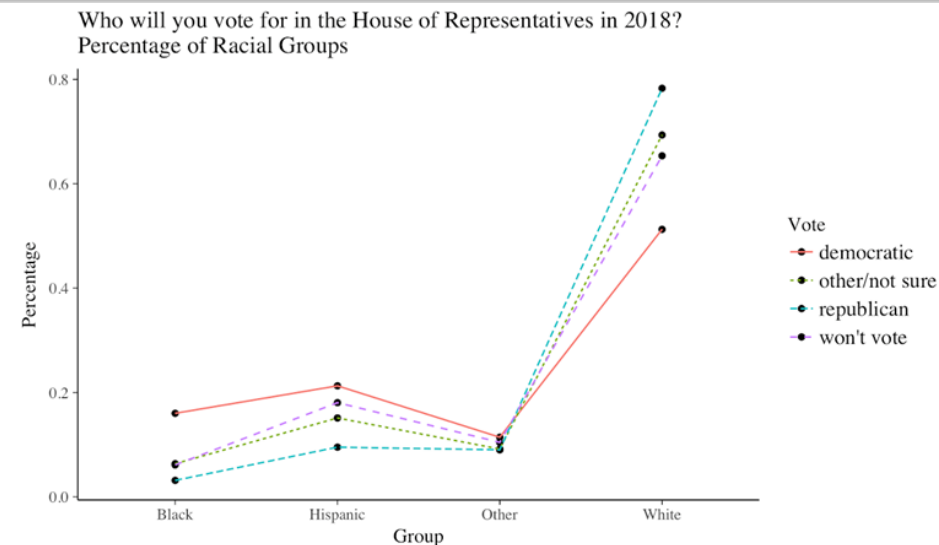
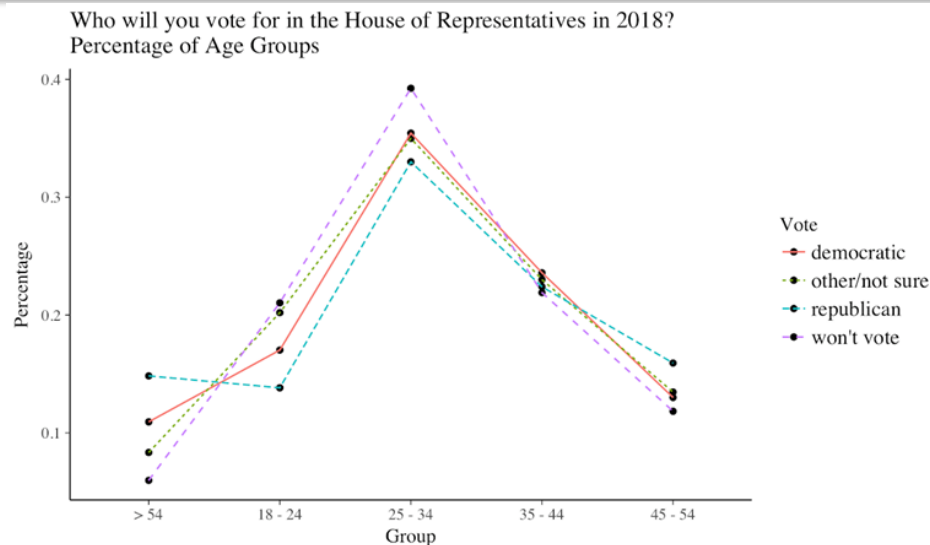
- A single month survey = 1,000 observations (total 6 months)
- Data table = Demographics + General questions  
+ Psychometric questions
  - Demographics
    - Age, gender, educational level, ...
  - General questions
    - Approve/disapprove Trump, *Which party you'll vote in 2018...*
  - 9 psychometric topics/surveys
    - Populism, Racial Resentment, Traditionalism, Compassion, Globalism, Economic Populism, Authoritarianism, Trust in Institutions, and Climate Change
- Non-representative sample



# Visualization: Population (Election VS Polling)



# Exploratory Polling Data Analysis



# Methodology: MRP

## Multilevel Regression and Poststratification (MRP)

### ■ History

- Adopted widely for static evaluation of polling data to recover state-level estimates of public opinion
  - Especially useful when having small number of observation in any states
- First developed by Gelman and Little (1997)
- Takes into account geographics (Park et al., 2006)

MRP = Bayesian Hierarchical Model + Poststratification

- Bayesian hierarchical modeling (improves over classical regression)
  - Random effects VS fixed effects
- Post-stratification

# Methodology: MRP

- Why MRP?
  - No issues like raking or simple weight adjustment
  - Sparsity in demographical cells
  - Small data to predict nationwide population
  - Explanatory power on demographics

# *Model: Categorical Variables of Demographics*

**Survey respondents are categorized based on 5 demographic variables**

<b>Demographic Variable</b>	<b>Possible Categories</b>
Gender	Male, Female
Age	18-24, 25-34, 35-44, 45-54, over 54
Race	White, Black, Hispanic, Other
Education	No Bachelor's Degree, Bachelors Degree
Party Affiliation	Democrat, Independent, Republican
Congressional District	One of 436 districts



# Model: Bayesian Hierarchical Model

$$\begin{aligned}\eta = & \beta_0 + \alpha_{\text{district}} + \alpha_{\text{age}} + \alpha_{\text{gender}} + \alpha_{\text{race}} + \alpha_{\text{education}} + \alpha_{\text{party}} \\ & + \alpha_{\text{age} \times \text{gender}} + \alpha_{\text{age} \times \text{education}} + \alpha_{\text{age} \times \text{party}} + \alpha_{\text{age} \times \text{race}} \\ & + \alpha_{\text{gender} \times \text{education}} + \alpha_{\text{gender} \times \text{party}} + \alpha_{\text{gender} \times \text{race}} \\ & + \alpha_{\text{education} \times \text{party}} + \alpha_{\text{education} \times \text{race}} + \alpha_{\text{party} \times \text{race}} \\ \mu_{\text{district}} = & \alpha_{\text{division}} + \beta_{\text{trump}} \text{logit}(\text{vote}_{2016}) + \beta_{\text{cook}} \text{cook}_{\text{district}}\end{aligned}$$

$$\begin{aligned}\alpha_{\text{district}} & \sim \text{N}(\mu_{\text{district}}, 1), \alpha_{\text{division}} \sim \text{N}(0, 1), \\ \beta_{\text{trump}} & \sim \text{N}(0, 1), \beta_{\text{cook}} \sim \text{N}(0, 1)\end{aligned}$$

# Model: Poststratification Space

## The Poststratification Space of Likely Voters

	age	gender	state	race	marstat	edu	party	vote	district	N
1	18 – 24	male	AK	white	married	bachelors	democratic	democratic	1	0.374011473
2	18 – 24	male	AK	white	married	bachelors	democratic	other/not sure	1	0.097090531
3	18 – 24	male	AK	white	married	bachelors	democratic	republican	1	0.037771121
4	18 – 24	male	AK	white	married	bachelors	independent	democratic	1	0.051371248
5	18 – 24	male	AK	white	married	bachelors	independent	other/not sure	1	0.311413933
6	18 – 24	male	AK	white	married	bachelors	independent	republican	1	0.046290979
7	18 – 24	male	AK	white	married	bachelors	republican	democratic	1	0.105836126
8	18 – 24	male	AK	white	married	bachelors	republican	other/not sure	1	0.412813559
9	18 – 24	male	AK	white	married	bachelors	republican	republican	1	1.436392861
10	18 – 24	male	AK	white	married	no bachelors	democratic	democratic	1	0.388705554
11	18 – 24	male	AK	white	married	no bachelors	democratic	other/not sure	1	0.142020455
12	18 – 24	male	AK	white	married	no bachelors	democratic	republican	1	0.054881938
13	18 – 24	male	AK	white	married	no bachelors	independent	democratic	1	0.038749441
14	18 – 24	male	AK	white	married	no bachelors	independent	other/not sure	1	0.381645357
15	18 – 24	male	AK	white	married	no bachelors	independent	republican	1	0.049214879
16	18 – 24	male	AK	white	married	no bachelors	republican	democratic	1	0.085634773
17	18 – 24	male	AK	white	married	no bachelors	republican	other/not sure	1	0.635887197
18	18 – 24	male	AK	white	married	no bachelors	republican	republican	1	1.659618607
19	18 – 24	male	AK	white	unmarried	bachelors	democratic	democratic	1	0.730733320

# *Model: District-level 2018 Election Prediction*

- Our model predicts the two-party vote shares in the upcoming 2018 Midterm Elections for each of the 435 congressional districts, with Washington D.C. considered as a 436<sup>th</sup> district
- Using a hierarchical model is a compromise between assuming that every district is different (one model per district) and that every district is the same (one model for all districts)
- The model learns the voting behavior of each demographic and district
- We augment our model with data from external sources to help it learn the differences between districts

# Model: Cook Score

## Cook Partisan Voting Index (CPVI)

- A measure of how much more a district leans towards a particular party compared to the national average
- For example, D+26 is the CPVI for New York's 10th Congressional District

# Model: District-level 2018 Election Prediction

## Major Party Voter Turnout Model

This model predicts if a respondent will choose one of these answers:

Who will you vote for in the House of Representatives in 2018?	Will vote Democratic	Will vote Republican	Will vote other/not sure	Won't Vote
--	----------------------	----------------------	--------------------------	------------

For each subpopulation, the model estimates the probability that a person belonging to that group will vote for a major party candidate:

US Congressional District	Gender	Race	Age	Education	What is your political party affiliation?	
Alabama-1	female	black	18 - 24	College	Democrat	0.785794
					Independent	0.264017
					Republican	0.649078
				No college	Democrat	0.718228
					Independent	0.175414
					Republican	0.596680

# Model: District-level 2018 Election Prediction

## Democratic Party Vote Turnout Model

This model is trained and tested on a reduced dataset containing only the respondents who will vote for a major party.

Who will you vote for in the House of Representatives in 2018?	Will vote Democratic	Will vote Republican
--	----------------------	----------------------

For each subpopulation, the model estimates the probability that a person belonging to that group will vote for the Democratic Party:

US Congressional District	Gender	Race	Age	Education	What is your political party affiliation?	
Alabama-1	female	black	18 - 24	College	Democrat	0.964243
					Independent	0.722461
					Republican	0.325900
				No college	Democrat	0.945041
					Independent	0.624978
					Republican	0.303258

# Model: District-level 2018 Election Prediction

## Poststratification

### Two Party Vote Share Conditioned on Major Party Vote Probability

Using the probabilities generated by the previous two models, the probability of a respondent from state  $s$  voting for a democratic candidate in the 2018 House of Representatives elections given that he or she will be voting for a major party candidate can be calculated using the following formula:

$$p_s = \frac{\sum_{j \in s} N_j \pi_j^{major} \pi_j^{dem}}{\sum_{j \in s} N_j \pi_j^{major}}$$

Each  $\pi$  in this formula represents a vector where each value represents a probability associated with a specific demographic cell.  $N_j$  represents a vector containing the population counts for each demographic cell in state  $s$ .

District	Democratic Party Vote Share	Republican Party Vote Share
Alabama-1	0.347612	0.652388
Alabama-2	0.393858	0.606142
Alabama-3	0.367572	0.632428
Alabama-4	0.180493	0.819507
Alabama-5	0.330053	0.669947
Alabama-6	0.258670	0.741330
Alabama-7	0.763045	0.236955

# *Model: District-level 2018 Election Prediction*

## **Model Validation Results**

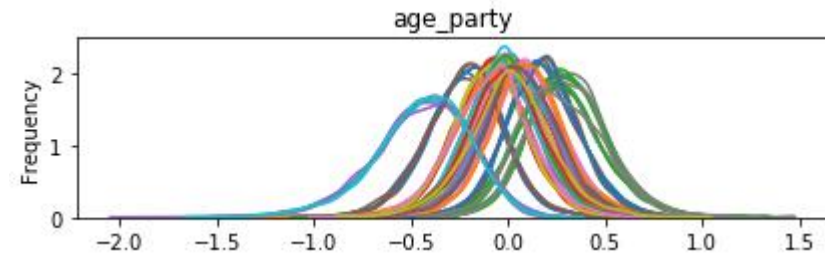
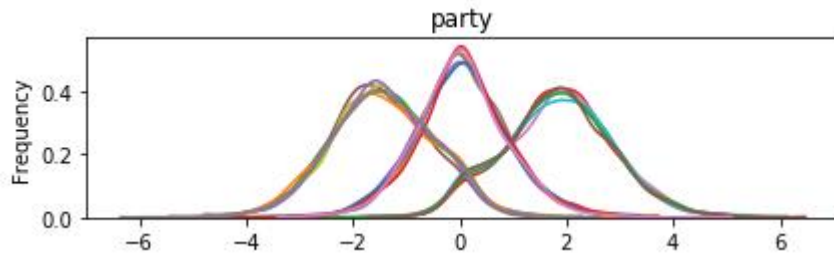
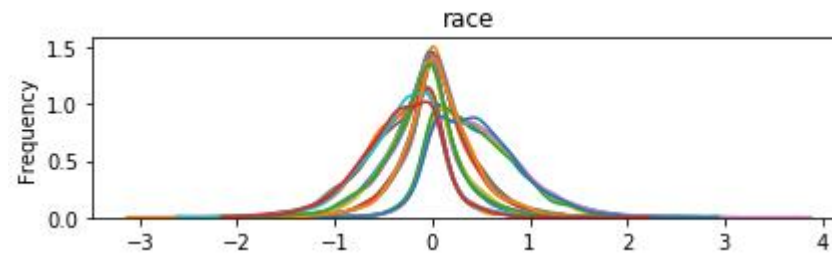
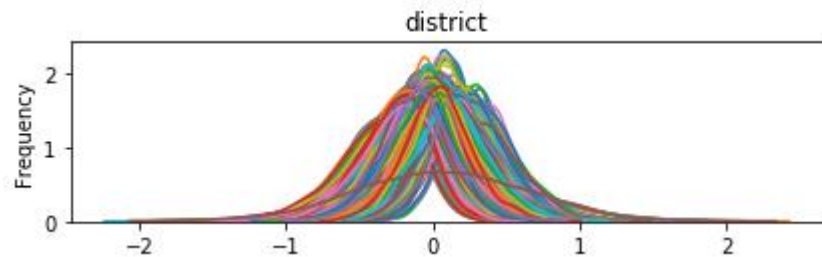
<b>Model</b>	<b>Test Set Prediction Accuracy</b>
Major Party Voter Turnout	75.54%
Democratic Party Preference	86.69%

We performed cross-validation, using 80% of the survey data as the training set and 20% as the test set.

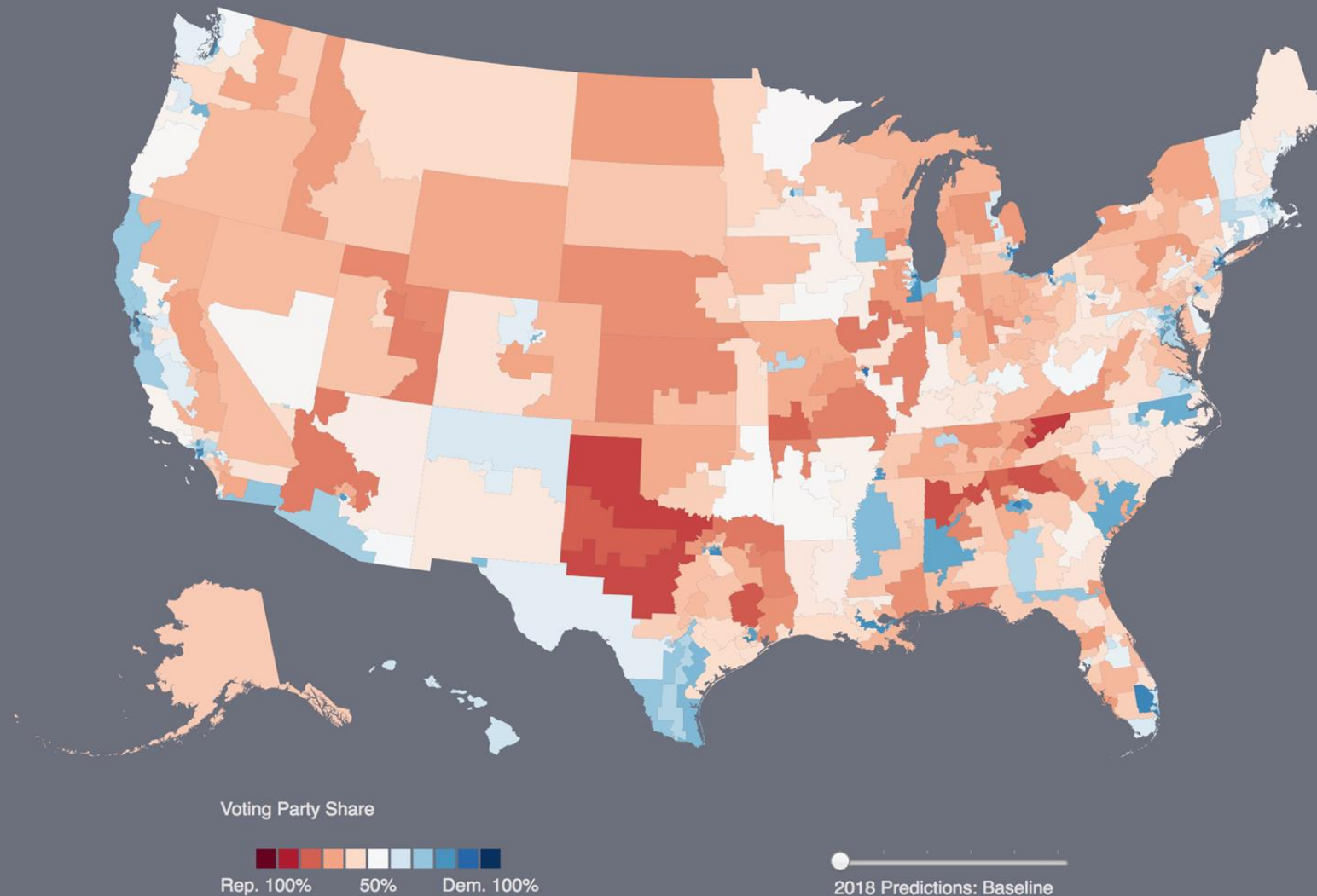


# Model: District-level 2018 Election Prediction

## Informative Variables



# *Model: District-level 2018 Election Prediction*



# *Extensions: 1 Adjusting Voter Turnout*

- What changes in voter turnout could tip the balance of power?
  - For Democrats to win exactly 218 seats:
    - White Democrats need to have 40% higher turnout.
    - Black Democrats need to have 266% higher turnout.
    - Hispanic Democrats need to have 586% higher turnout.
    - Other race Democrats need to have 890% higher turnout.

# *Extensions: 2 Incorporating Psychometric Variables*

## **Measuring Psychometric Variables Using Survey Questions**

- The surveys indirectly measure the presence of certain personality traits or beliefs by asking questions about the respondent's views or asking if they agree with certain statements related to one of nine topics
- Each group of questions was answered by a subset of the respondents (6,000 respondents for each group)
- For each topic, we incorporate answers to the corresponding questions into our model and restrict the training and test sets to the respondents who answered those questions

# Extensions: 2 Incorporating Psychometric Variables

Topic	Example Question
Populism	"The system is stacked against people like me."
Racial Resentment	"Racial minorities can overcome prejudice without any special favors."
Traditionalism	"More influence of churches in daily life would make society better."
Compassion	"Homelessness is sometimes the right price to pay for lacking work ethic."
Economic Populism	"Too many millionaires and billionaires lead to inequality hurting people like me."
Authoritarianism	Which quality is more important for children: independence vs. respect for elders?
Trust in Institutions	Do you trust what the FBI and CIA reports?
Presidential	Compared to other recent presidents, is President Trump a moral leader?
Climate	How do you feel about government spending and regulations to address climate change?

# Extensions: 2 Incorporating Psychometric Variables

## Test Set Prediction Accuracy with Psychometric Variables

Topic	Major Party Turnout Model				Democratic Party Vote Model			
	Baseline	Scheme 1	Scheme 2	Scheme 3	Baseline	Scheme 1	Scheme 2	Scheme 3
Racial Resentment	74.05%	74.05%	74.05%	73.95%	87.43%	89.49%	89.87%	88.74%
Traditionalism	75.24%	75.52%	76.18%	73.25%	86.14%	87.99%	87.62%	85.95%
Authoritarianism	72.75%	72.75%	72.75%	72.84%	85.32%	86.62%	86.43%	87.17%

# Extensions: 2 Incorporating Psychometric Variables

## POLITICS

### *Trump Voters Driven by Fear of Losing Status, Not Economic Anxiety, Study Finds*

By NIRAJ CHOKSHI APRIL 24, 2018



A Trump supporter at a campaign rally in Sacramento in June 2016. A new study found that many Trump voters were driven by fear of losing their status in society. Damon Winter/The New York Times

#### RECENT COMMENTS

**JS** 2 days ago

"Cultural anxiety" is such a nice euphemism for racism isn't it? And ya know what? Being a white christian male in America is STILL a pretty...

**Martha Barron** 2 days ago

I dont think this is a brand new theory. Read "Strangers in Their Own Land" by Arlie Russell Hochschild, published shortly before the 2016...

**Clarence Dupiton** 2 days ago

dear white men, welcome to the club. sincerely, everyone else.

[SEE ALL COMMENTS](#)



# Extensions: 3 Dynamic MRP

## How do we predict the opinion change when new data keep coming?

Ways to do it:

Time series models (but only on fixed effect...)

Fit a MRP at every time point (but not very systematic...)

We include time feature and interactions in random effects!

$$\begin{aligned}\eta = & \beta_0 + \mu_{state} + \alpha_{month} + \alpha_{age} + \alpha_{gender} + \alpha_{race} \\ & + \alpha_{education} + \alpha_{marstat} \\ & + \alpha_{age \times gender} + \alpha_{age \times education} + \alpha_{age \times marstat} + \alpha_{age \times race} \\ & + \alpha_{gender \times education} + \alpha_{gender \times marstat} + \alpha_{gender \times race} \\ & + \alpha_{education \times marstat} + \alpha_{education \times race} + \alpha_{marstat \times race} + \text{SCORE}_{month} \\ \mu_{state} = & s_{state} + (s_{state,2} + \beta_{month} + \beta_{month^2})\alpha_{state}\end{aligned}$$

(Gelman et al. 2018)



# Questions?



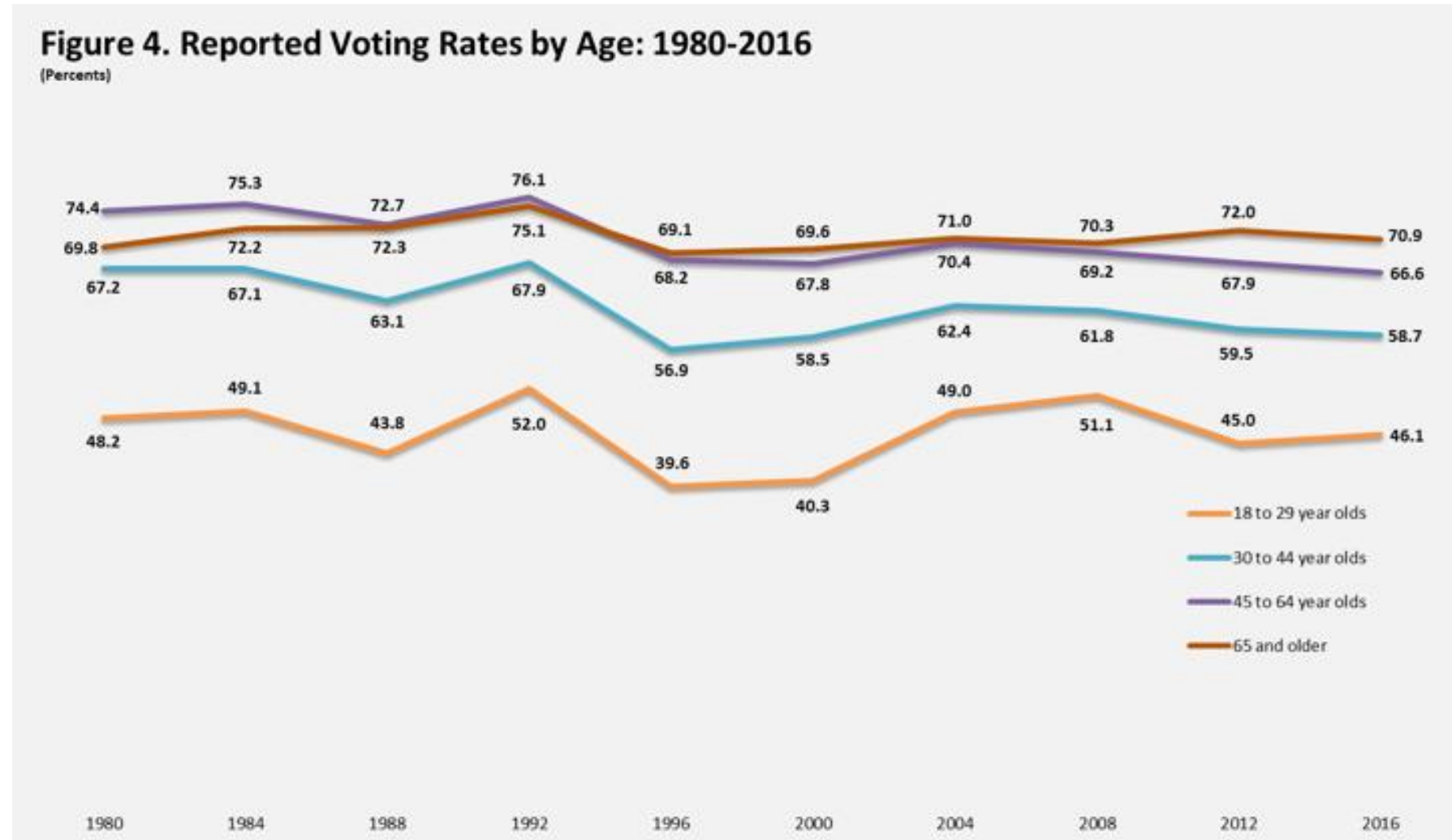
# Appendix

# *Introduction: 2018 House of Rep. Election*

In the 2016 Presidential election, **60%** of Americans voted.

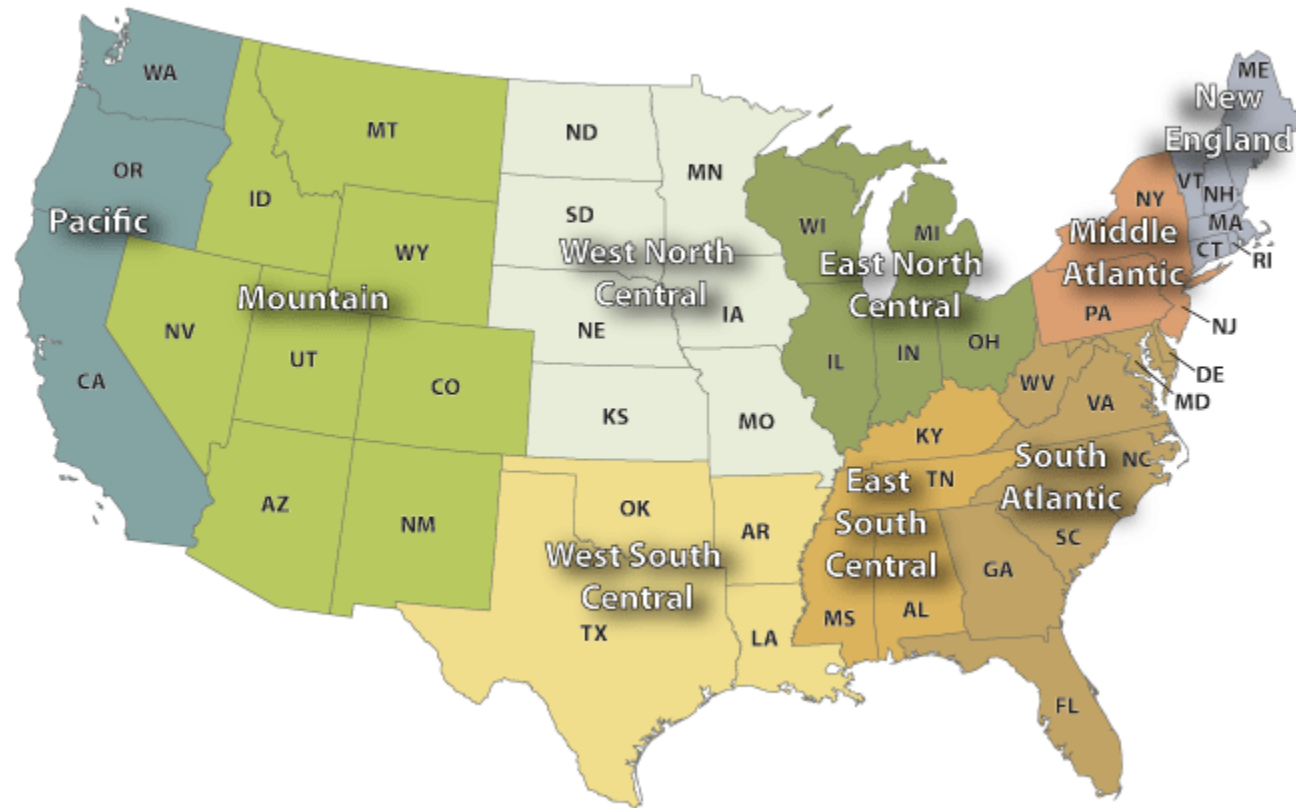
In a typical midterm election (non-Presidential election), **40%** of Americans will vote.

# Introduction: 2018 House of Rep. Election



# Model: District-level 2018 Election Prediction

U.S. Census Divisions

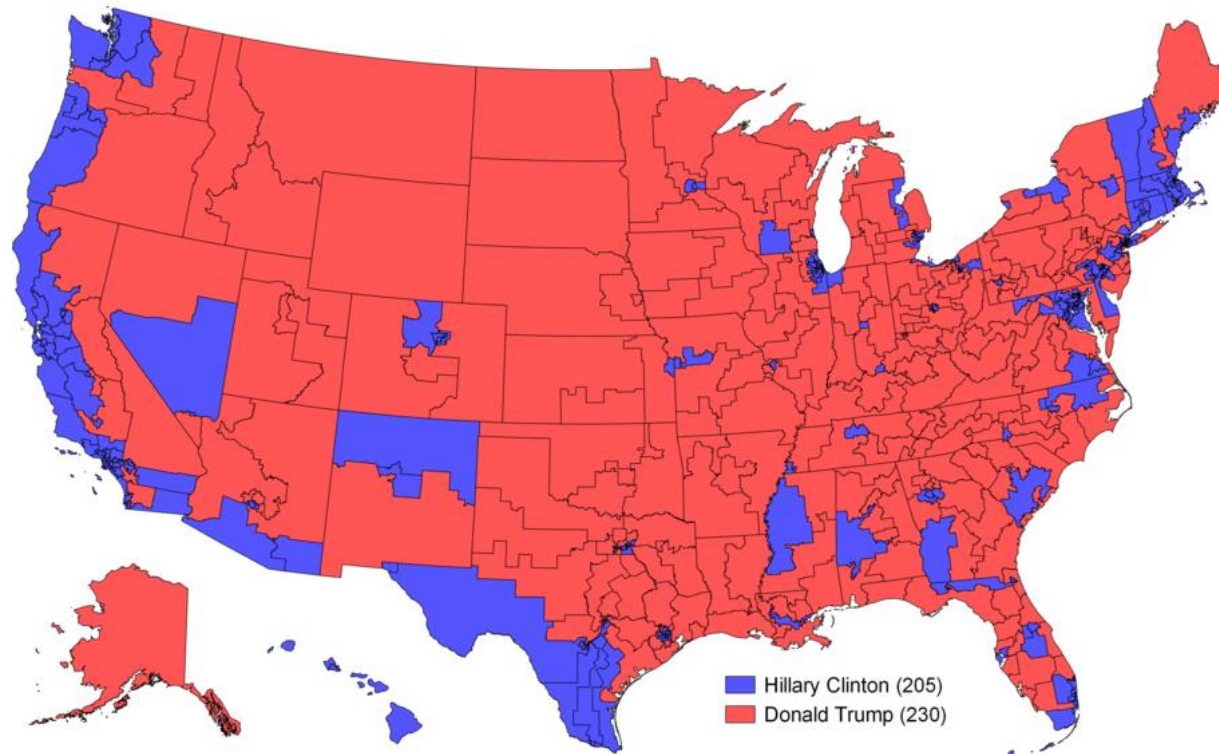


<https://www.ncdc.noaa.gov/monitoring-references/maps/us-census-divisions.php>



# Model: District-level 2018 Election Prediction

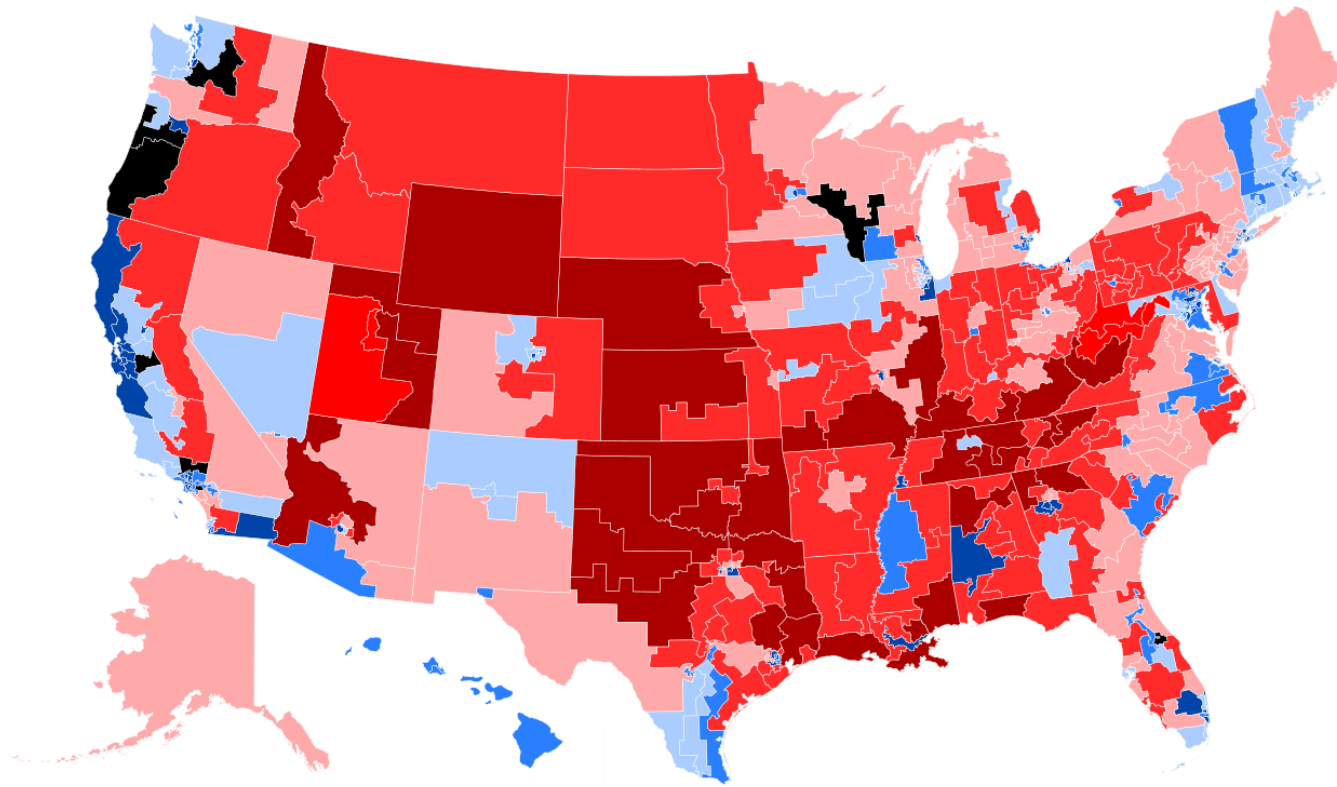
Donald Trump's two-party vote share in each district in the 2016 Election



<https://www.dailykos.com/stories/2017/1/30/1627319/-Daily-Kos-Elections-presents-the-2016-presidential-election-results-by-congressional-district>

# *Model: District-level 2018 Election Prediction*

## **District CPVI Map after the 2016 Election**



[https://en.wikipedia.org/wiki/Cook\\_Partisan\\_Voting\\_Index](https://en.wikipedia.org/wiki/Cook_Partisan_Voting_Index)

# *Model: District-level 2018 Election Prediction*

## **How the CPVI is Calculated**

This 2017 release has updated our PVI scores to incorporate the results of the November 2016 presidential election. A Partisan Voting Index score of D+2, for example, means that in the 2012 and 2016 presidential elections, that district performed an average of two points more Democratic than the nation did as a whole, while an R+4 means the district performed four points more Republican than the national average. If a district performed within half a point of the national average in either direction, we assign it a score of EVEN.

<https://www.cookpolitical.com/introducing-2017-cook-political-report-partisan-voter-index>



# *Extensions: Incorporating Psychometric Variables*

## **Answer Choice Weighting Schemes**

Example: Which quality is more important for children: independence vs. respect for elders?

**Scheme #1:**            1 - Respect for elders, 0 - Independence/Don't know

We consider the respondent to have the trait/attitude of interest if he or she has received 1 point for a majority of the questions.

**Scheme #2:**

Same as #1, except we use the respondent's point total in the model instead of labeling the respondents with "have" or "don't have".

**Scheme #3:**            2 - Respect for elders, 1 - Don't know, 0 - Independence

Same as #2, except we assign different point values to each choice depending on how much it agrees with the trait/attitude of interest.

# References

1. Gelman, A., & Little, T. C. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology*, 23(2), 127-35.
2. Little, R. J. (1993). Post-stratification: a modeler's perspective. *Journal of the American Statistical Association*, 88(423), 1001-1012.
3. Park, D. K., Gelman, A., & Bafumi, J. (2006). State-level opinions from national surveys: Poststratification using multilevel logistic regression. *Public opinion in state politics*, 209-28.