Forecasting US Presidential Elections:

Using Mixed Effects Models

Agenda

1 **Overview of how the US elects presidents**

2 **Stylized facts about voting**

3 **Types of forecasting models**

4 **Building a fundamentals based forecasting model**

Agenda

1 **Overview of how the US elects presidents**

2 Stylized facts about voting

3 Types of forecasting models

4 Building a fundamentals based forecasting model

US Presidential Elections are Decided by the Electoral College

Each state* gets electoral votes based on:

(1) Number of house representatives the state has, which is determined by a state's population in the US census (see the Huntington-Hill method)

plus

(2) Number of senators the state has (always 2)

*Washington DC, from the 23rd amendment, gets electoral votes based on its population but no more than the least populous state, which is currently Wyoming with 3 electoral votes.

Current Number of Electoral Votes

435 *# of House of Representative Members*

+

100 *# of Senators*

+

3 *# of Electoral Votes for Washington DC*

---

538 *Electoral Votes in Electoral College*

51

50 states plus DC award electoral votes

270

electoral votes a candidate must get to win the election

2*

Maine and Nebraska award electoral votes differently than other states (based on congressional districts)

Candidates win electoral votes from a state by winning the majority of votes cast in the state*

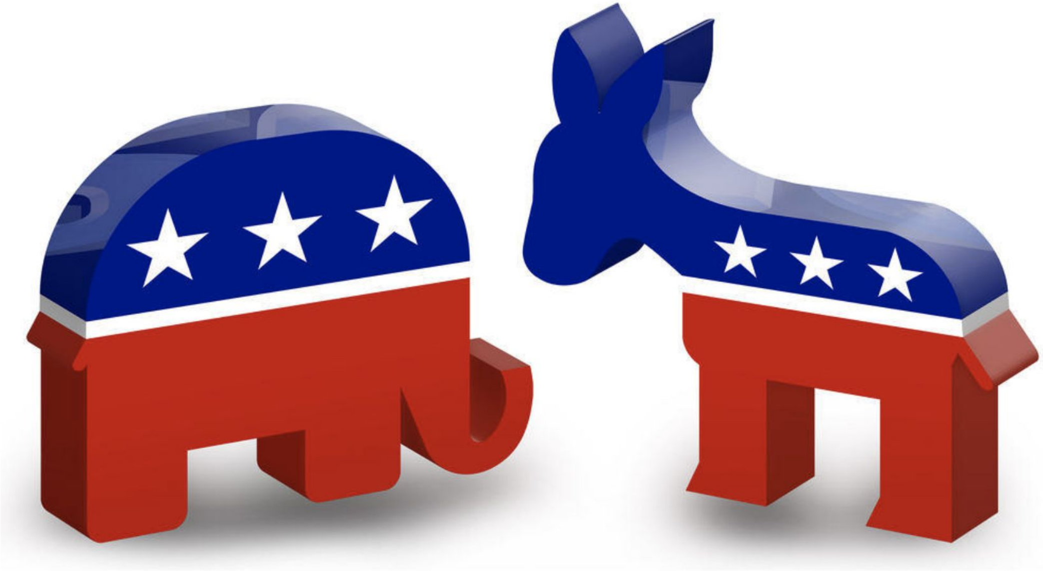
After voting is closed, we add up the votes each candidate gets per state, giving all electoral votes to the winner. The election winner is the candidate with the majority of electoral votes.

If no candidate gets a majority (270 or more electoral votes), the election results are decided by the house of representatives (with 1 vote per state).

The US is a Two Party System

The electoral college is basically **a winner take all system** as most of the time you get all of a state's electoral votes by winning it.

This makes it hard for third parties to matter in US Presidential elections, so we can ignore them for modeling purposes.



Side Note: Weird Things can Happen in the Electoral College

Because the electoral college awards electoral votes on a state by state basis, we can get weird outcomes.

It is possible for the candidate who gets the majority of votes overall i.e. wins the popular vote nationally to lose the electoral college and therefore the election.

This might sound like an edge case, but ***it has happened 5 times in 58 elections and 2 times in the last 5 elections (2000 and 2016)***!

Splits like these, where one party wins the popular vote and the other wins the electoral college, are especially likely to occur in close races. See here and here for more information.

Agenda

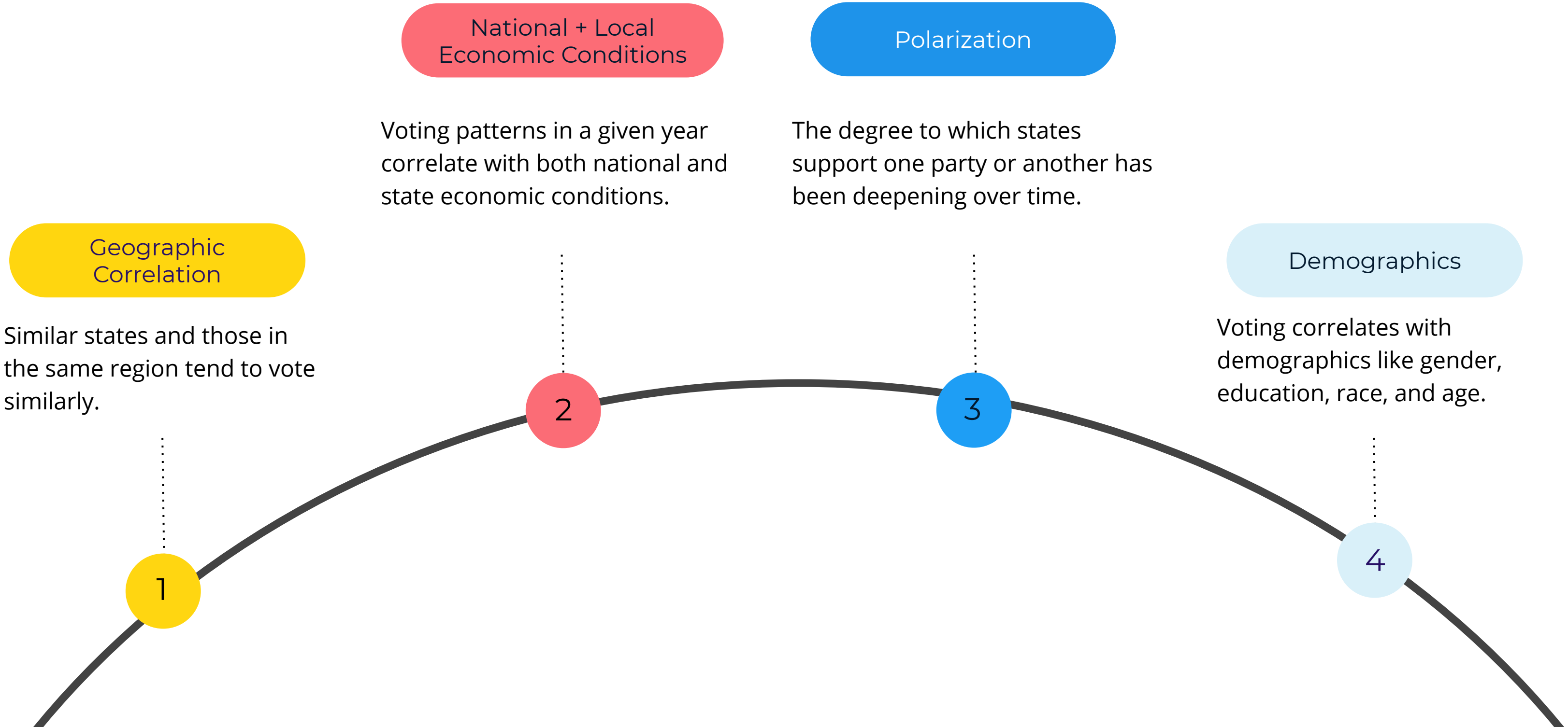
1 Overview of how the US elects presidents

2 **Stylized facts about voting**

3 Types of forecasting models

4 Building a fundamentals based forecasting model

Stylized Facts about Voting



1 Geographic Correlation

Similar states and those in the same region tend to vote similarly.

2 National + Local Economic Conditions

Voting patterns in a given year correlate with both national and state economic conditions.

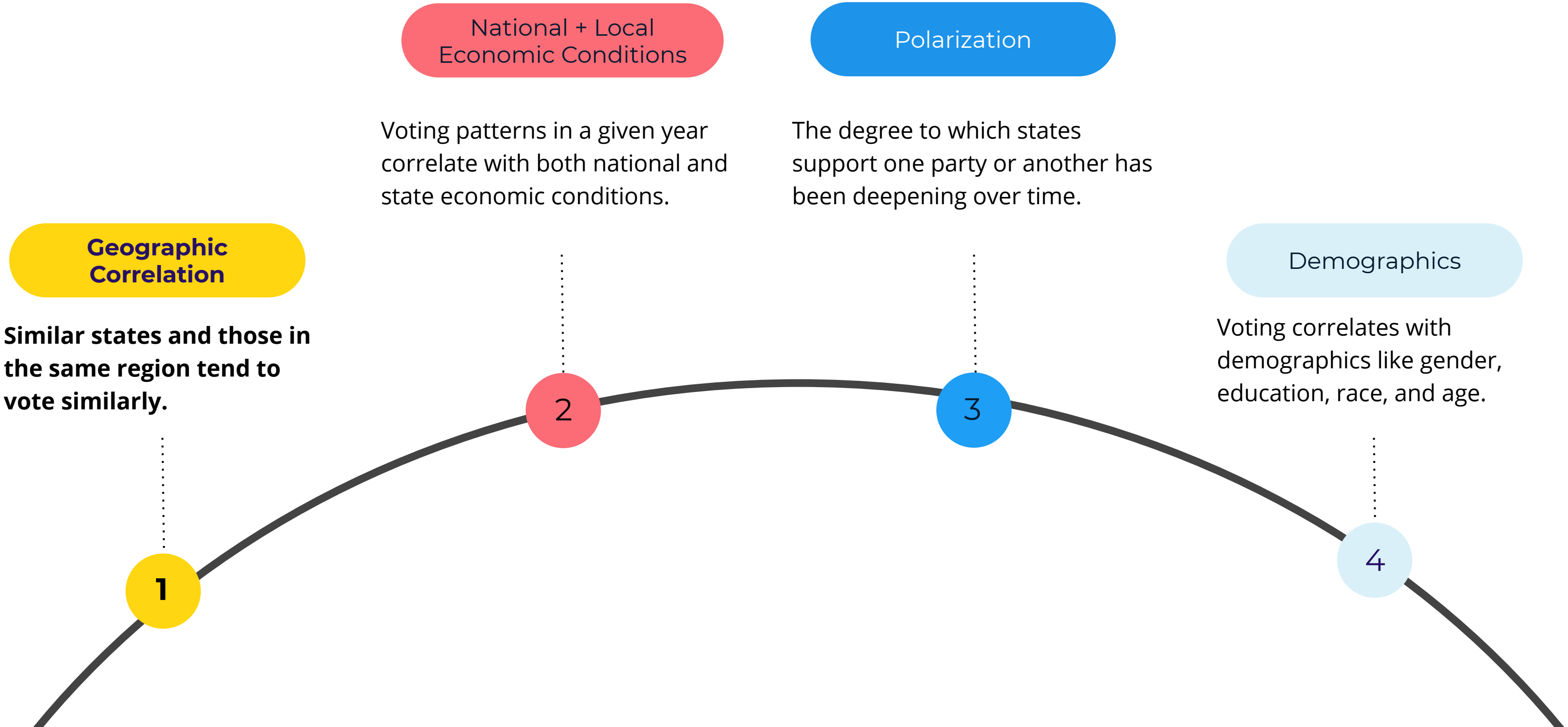
3 Polarization

The degree to which states support one party or another has been deepening over time.

4 Demographics

Voting correlates with demographics like gender, education, race, and age.

Stylized Facts about Voting



1 **Geographic Correlation**

Similar states and those in the same region tend to vote similarly.

2 National + Local Economic Conditions

Voting patterns in a given year correlate with both national and state economic conditions.

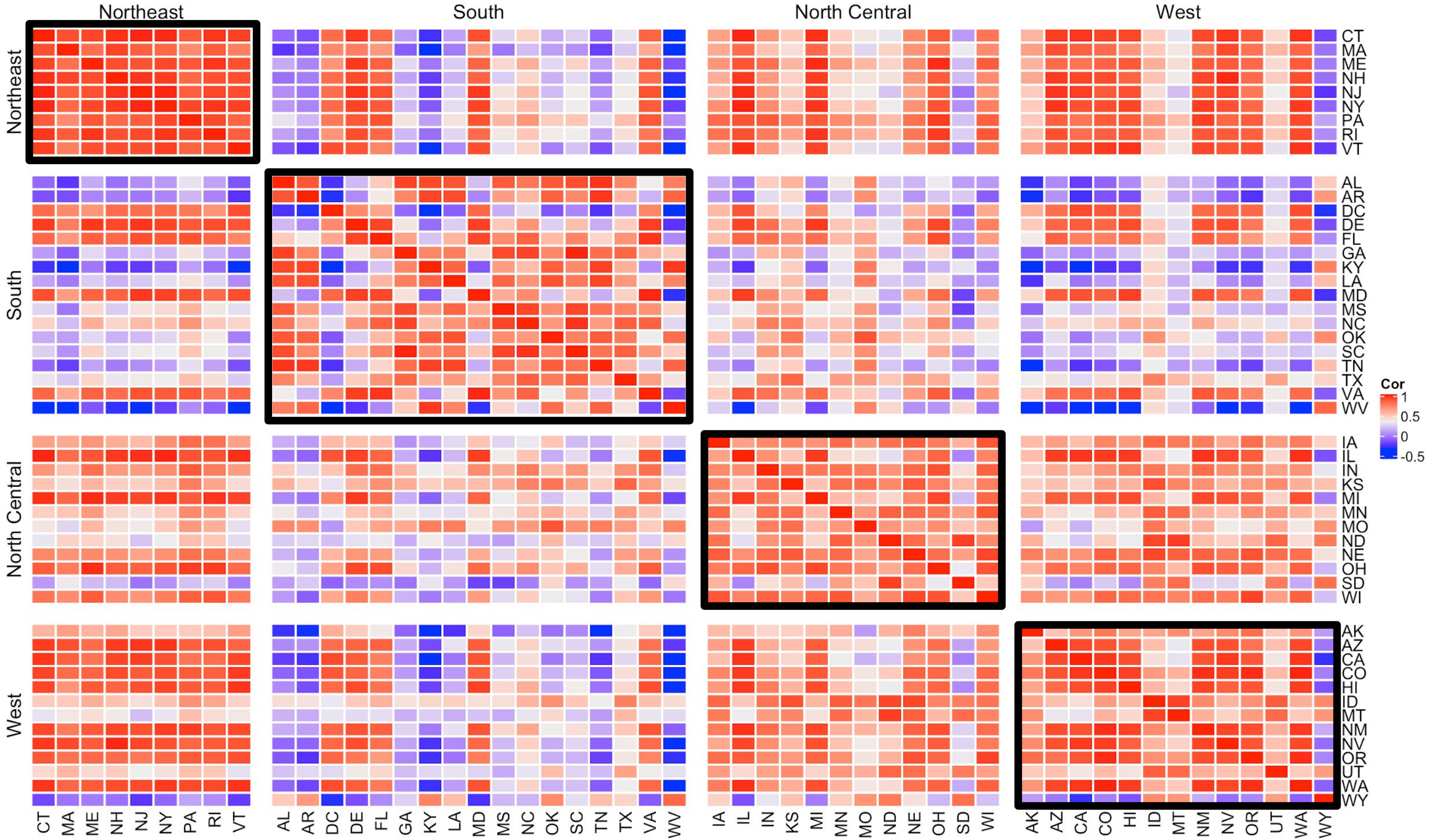
3 Polarization

The degree to which states support one party or another has been deepening over time.

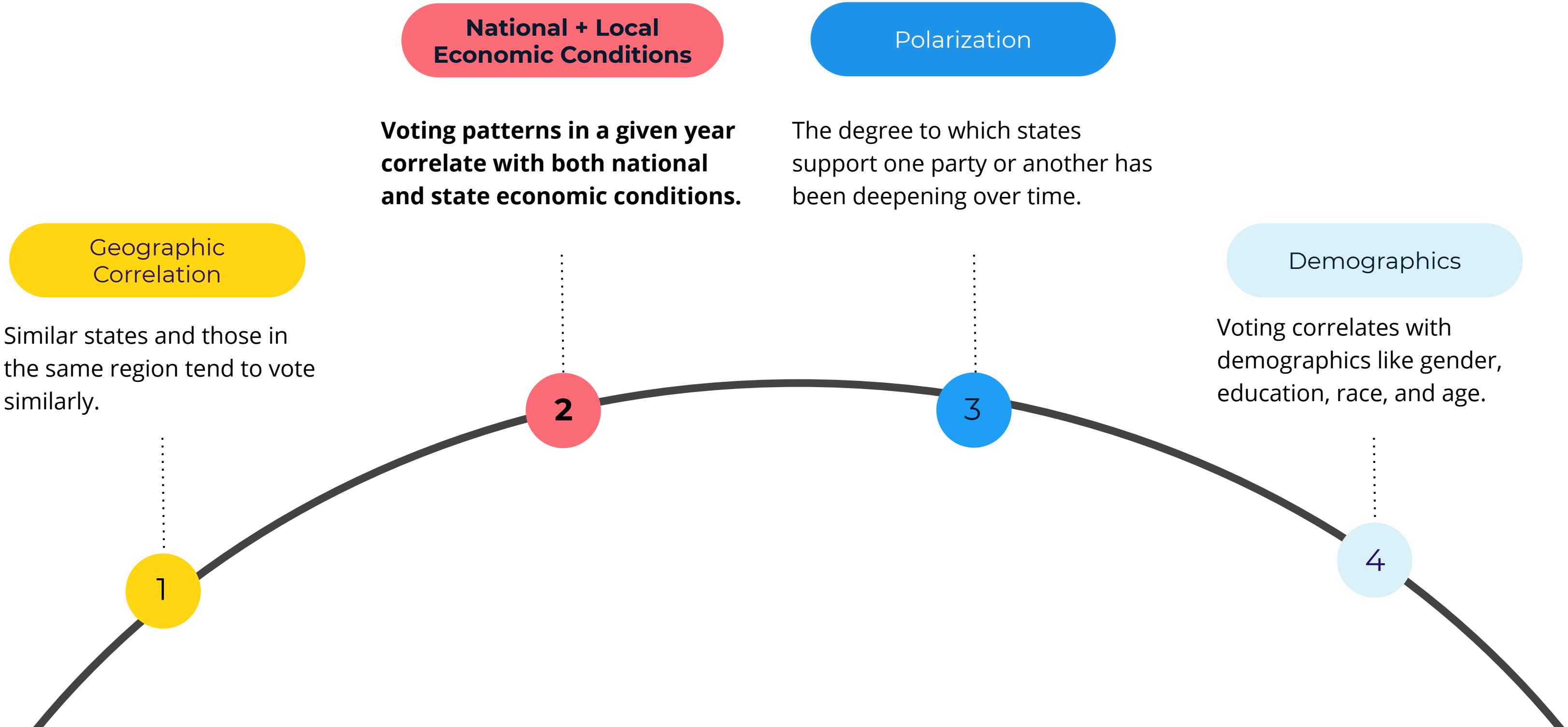
4 Demographics

Voting correlates with demographics like gender, education, race, and age.

High Correlation Across States in a Region



Stylized Facts about Voting



1

Geographic Correlation

Similar states and those in the same region tend to vote similarly.

2

National + Local Economic Conditions

Voting patterns in a given year correlate with both national and state economic conditions.

3

Polarization

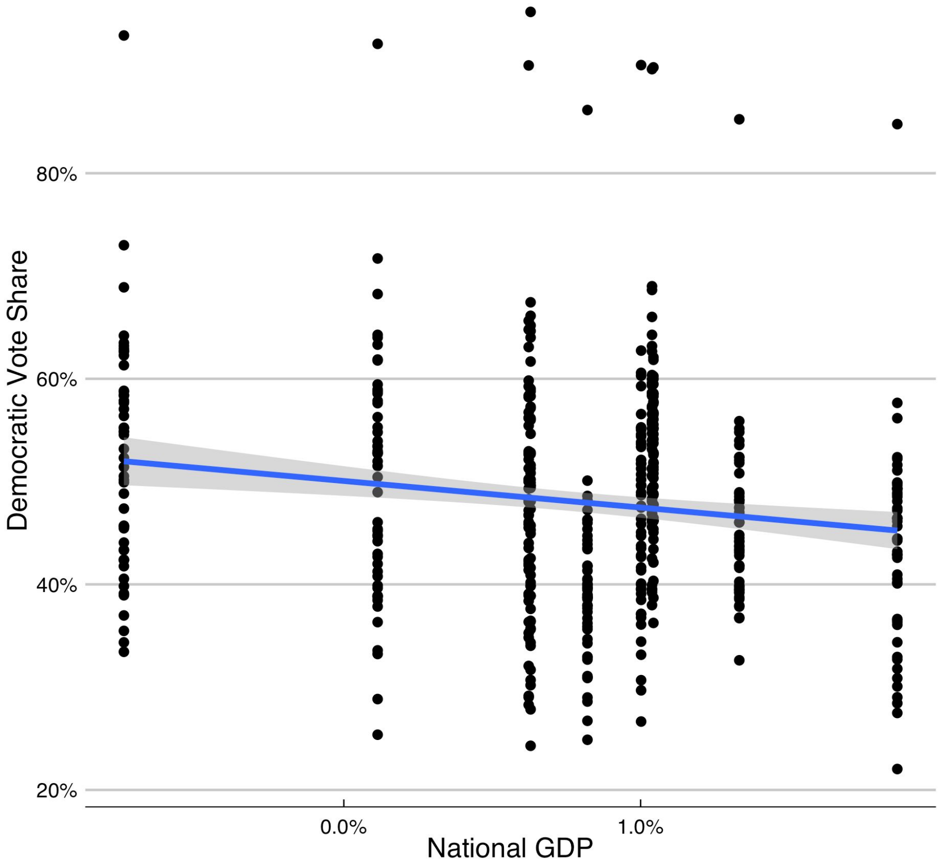
The degree to which states support one party or another has been deepening over time.

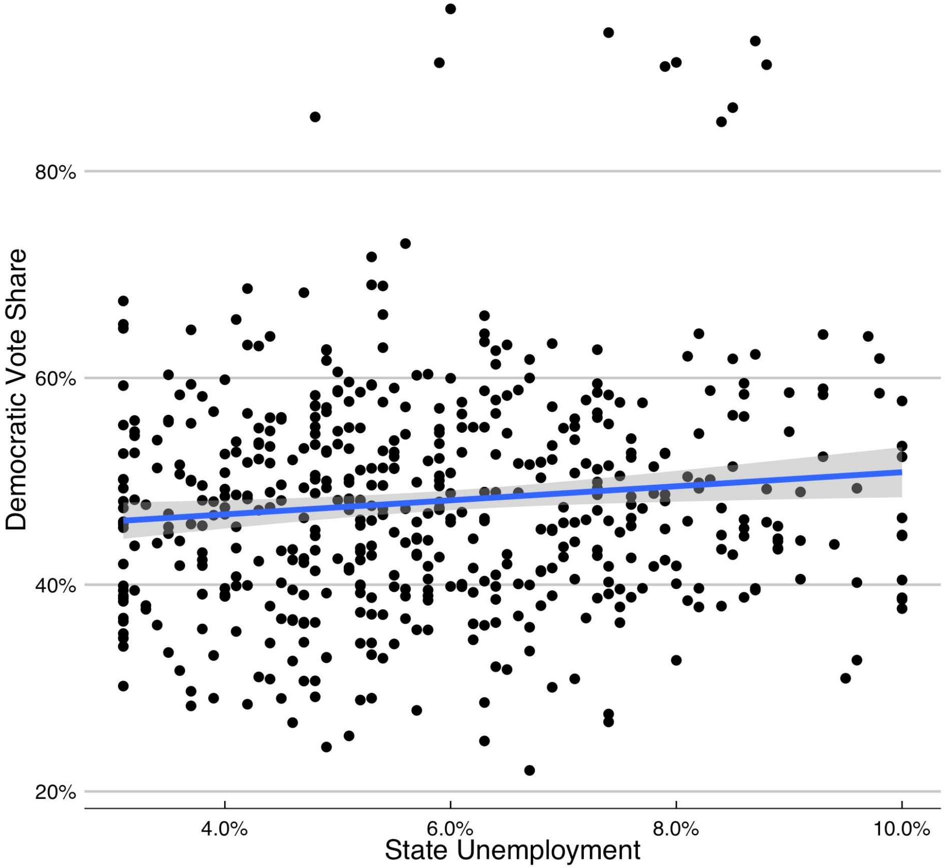
4

Demographics

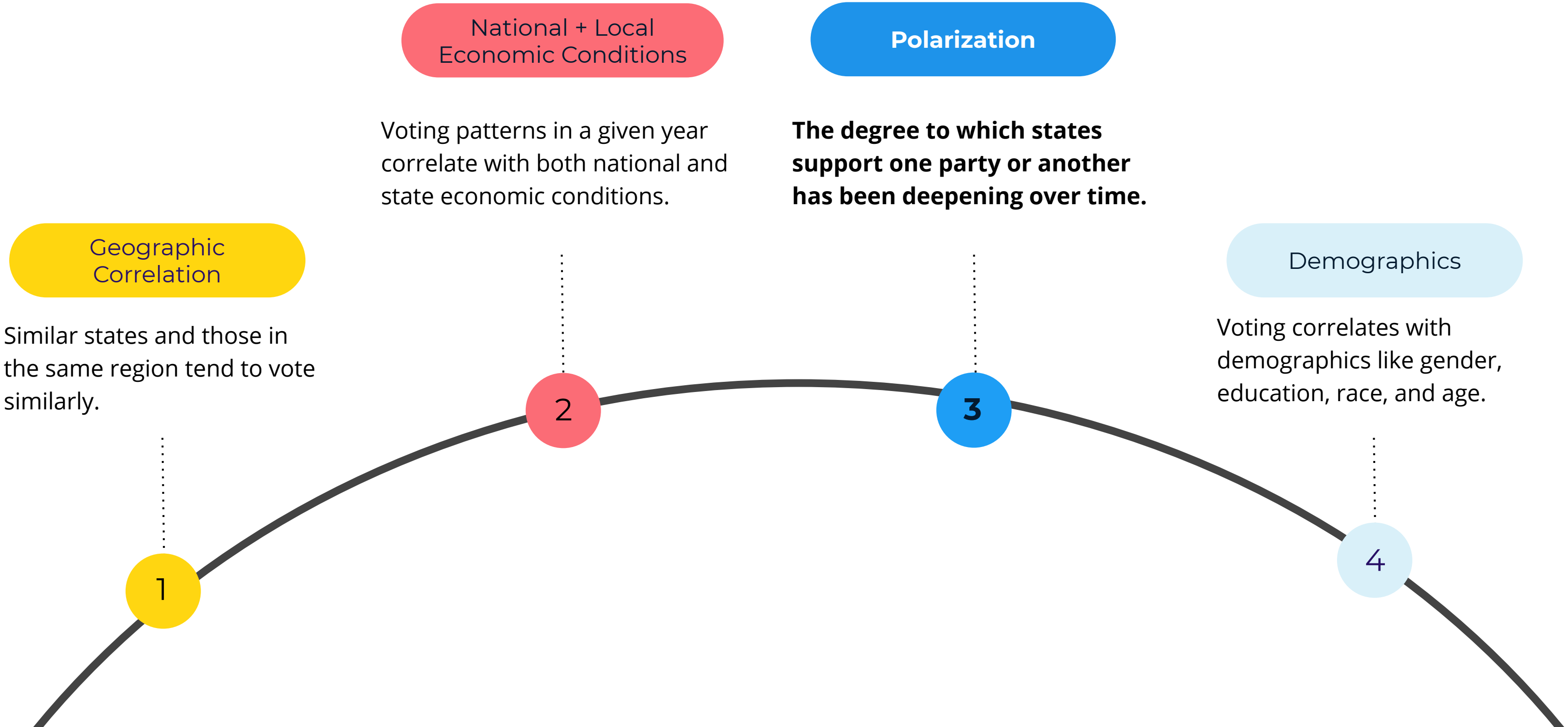
Voting correlates with demographics like gender, education, race, and age.

National & State Economic Conditions Matter





Stylized Facts about Voting



Geographic Correlation

Similar states and those in the same region tend to vote similarly.

National + Local Economic Conditions

Voting patterns in a given year correlate with both national and state economic conditions.

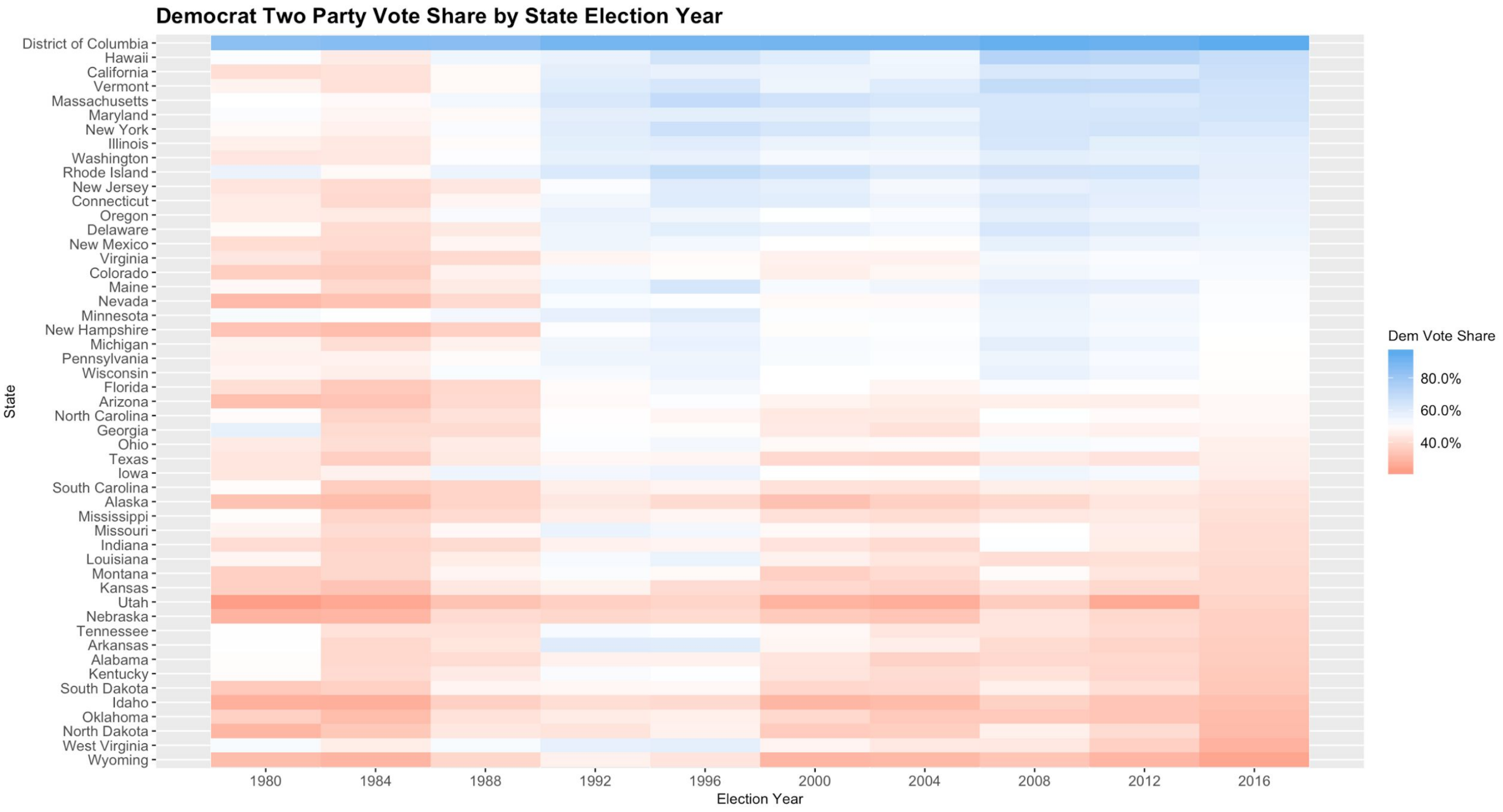
Polarization

The degree to which states support one party or another has been deepening over time.

Demographics

Voting correlates with demographics like gender, education, race, and age.

Now Red States are Redder & Blue States are Bluer

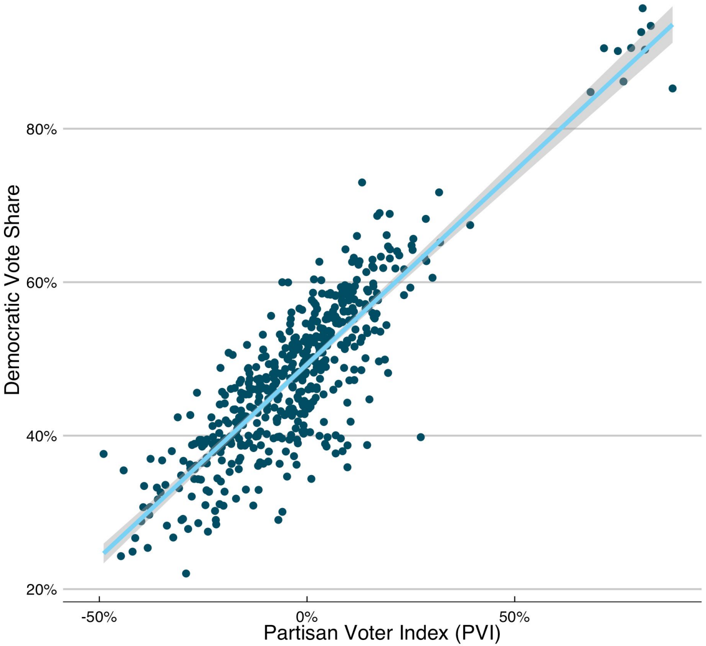


Key Variable: Partisan Voter Index (PVI)

$\text{PVI}_{\text{state, year}}$ = (Dem Vote $\text{Share}_{\text{state, year}}$ - Rep Vote $\text{Share}_{\text{state, year}}$) - (Dem Vote $\text{Share}_{\text{national, year}}$ - Rep Vote $\text{Share}_{\text{national, year}}$)

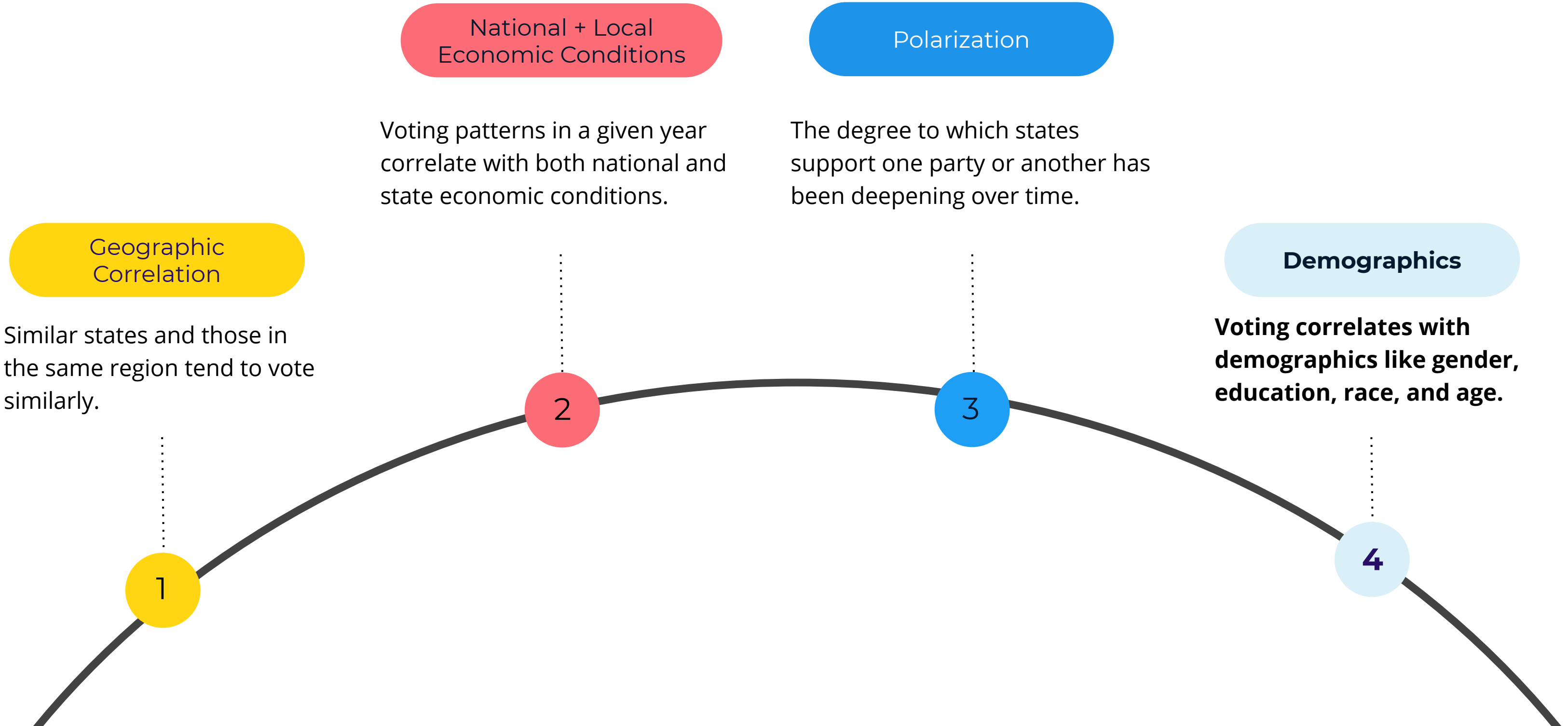
Calculate predictor using a mix of PVI in the past two elections before a given election year

Ex: PVI predictor for 2016 is a 75%, 25% mix of PVI in 2012 and 2008; this reduces overfitting to last election



This metric intuitively captures how much more Democratic or Republican leaning a state has been recently than the nation as a whole and is a strong predictor of future voting patterns.

Stylized Facts about Voting



1

Geographic Correlation

Similar states and those in the same region tend to vote similarly.

2

National + Local Economic Conditions

Voting patterns in a given year correlate with both national and state economic conditions.

3

Polarization

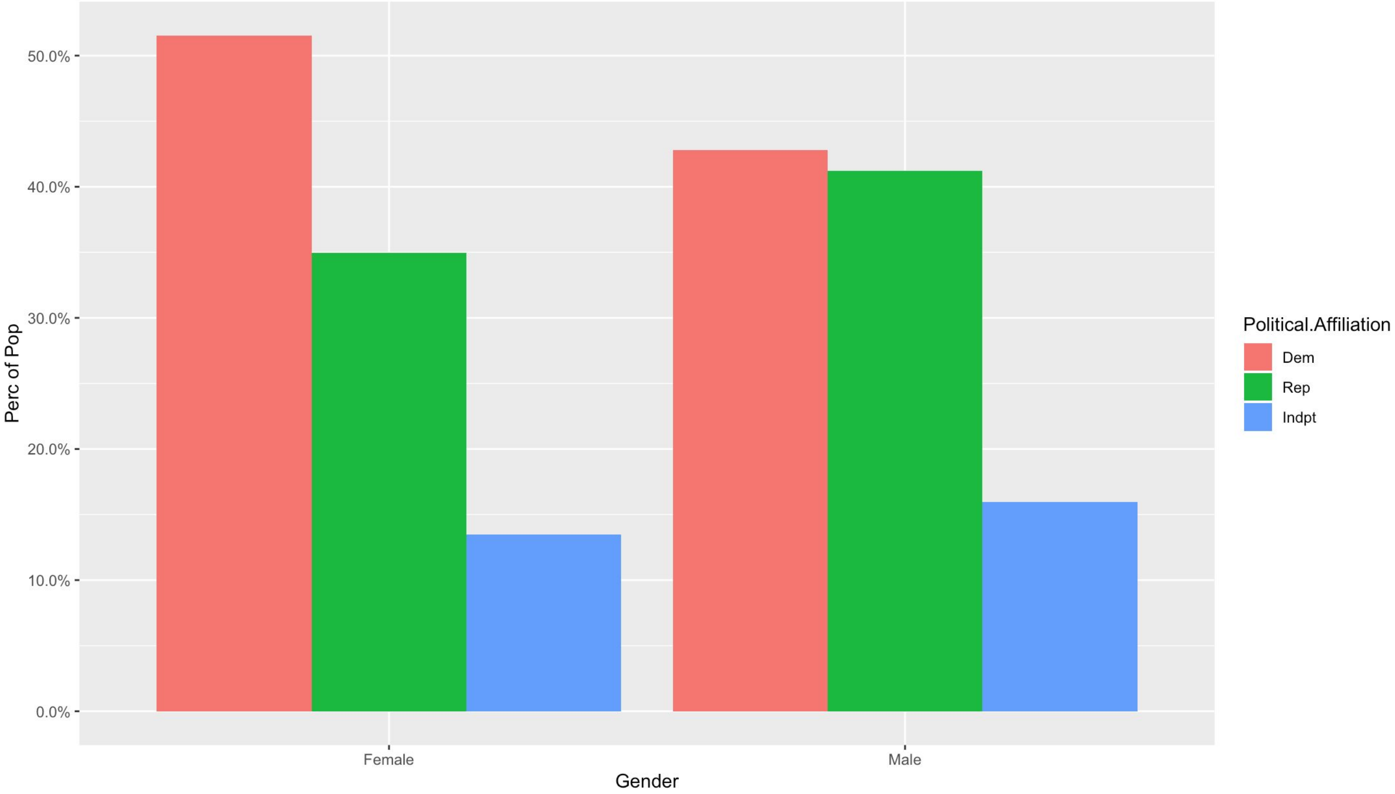
The degree to which states support one party or another has been deepening over time.

4

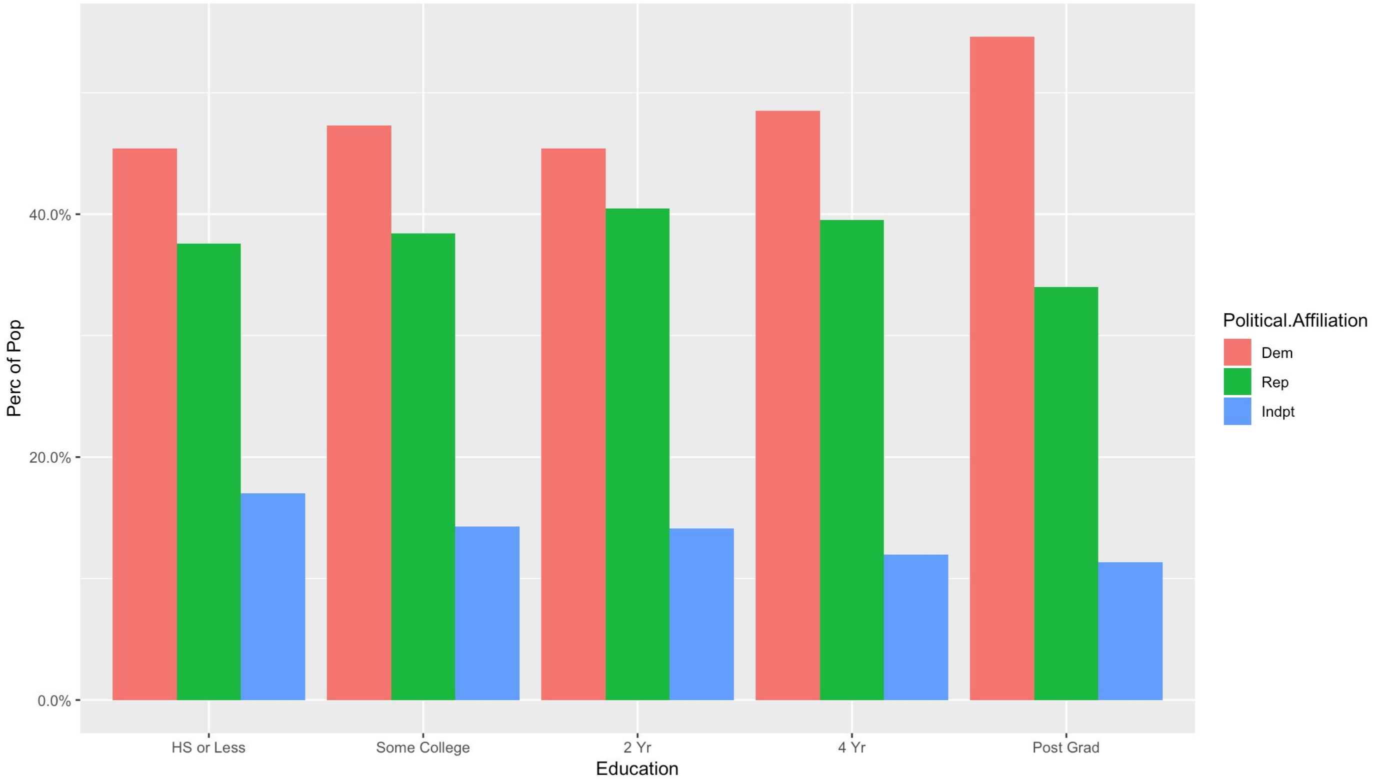
Demographics

Voting correlates with demographics like gender, education, race, and age.

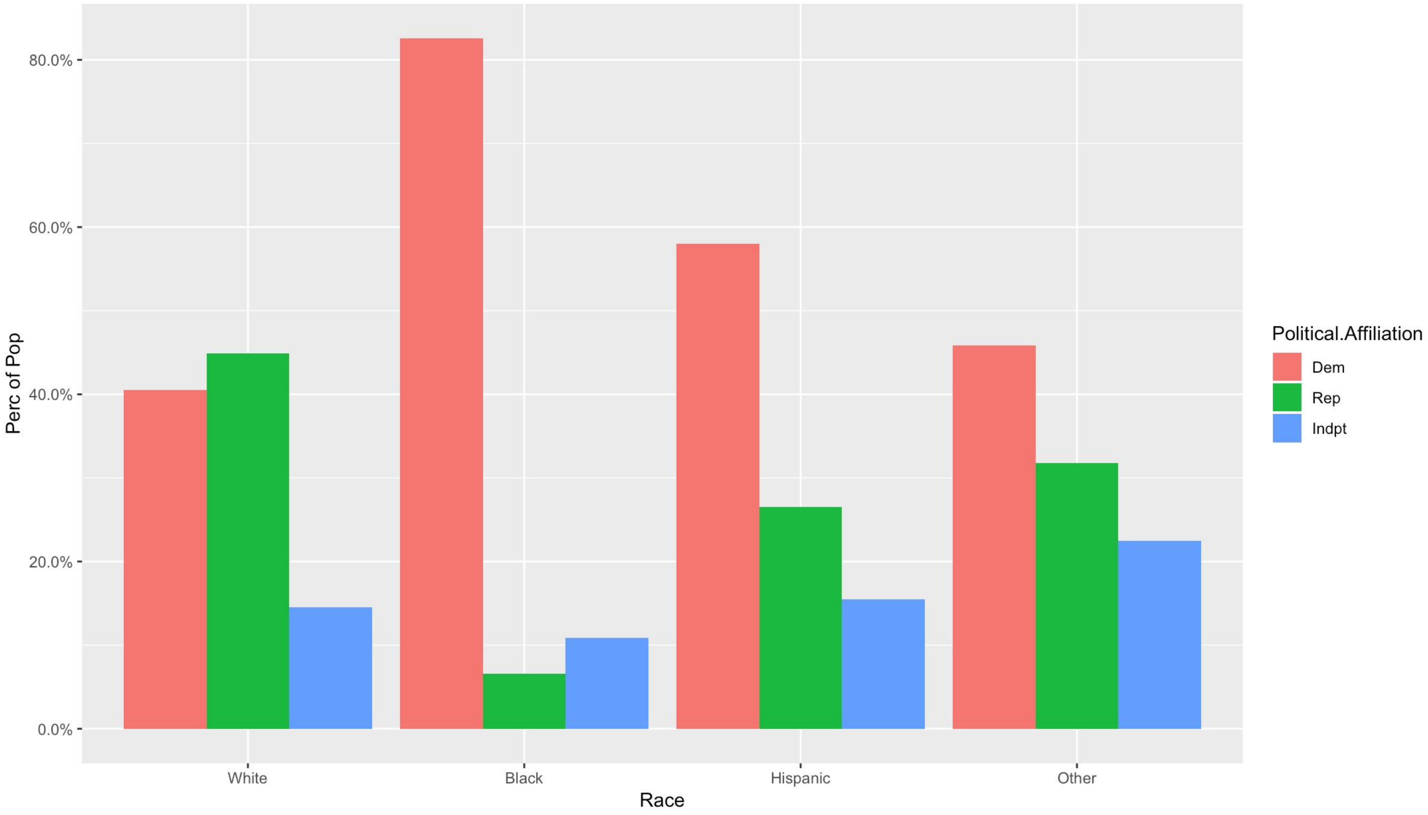
Women are More Likely to Vote Democratic



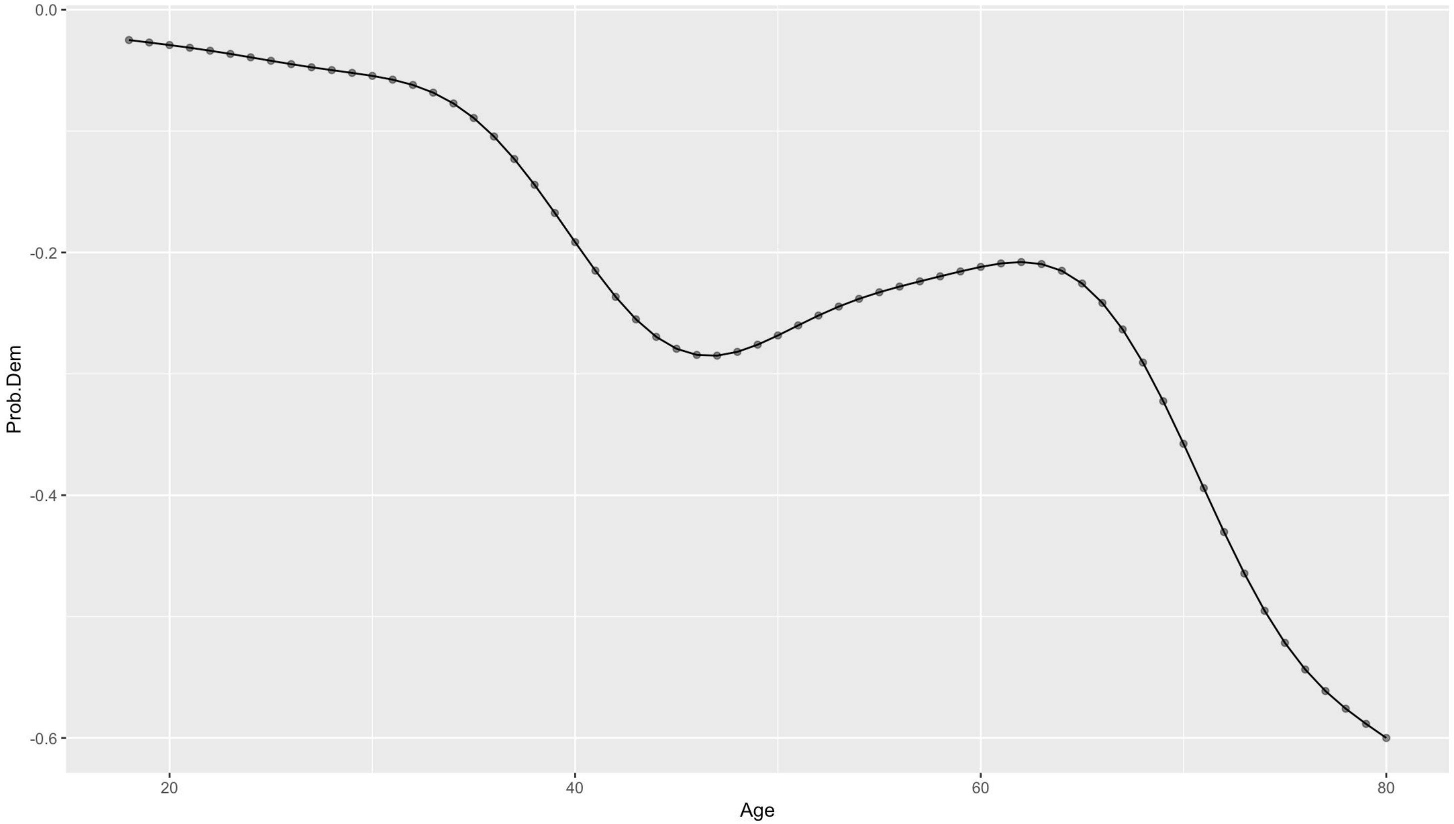
Voting Democratic is Correlated with Education



White Voters are Less Likely to Vote Democratic



Younger Voters Tend to Vote Democratic



Agenda

1 Overview of how the US elects presidents

2 Stylized facts about voting

3 **Types of forecasting models**

4 Building a fundamentals based forecasting model

Types of Forecasting Models

	Fundamentals Based	Polls Based	Fundamentals + Polls Hybrid
What They Are	Assume elections are based on things like **the economy, past voting history, incumbency, approval, etc**.	Use **national and state polls** prior to the election to reveal what voter preferences are.	Idea is to **combine fundamental predictors and polls** together to get the merits of both.
Pros / Cons	Historically there is not that much data since there have been few elections. As a result fundamentals based models tend to overfit.	Generally performs better than fundamentals based models. But polls have issues: (1) **Herding**: polls can try to copy one another (2) **House effects**: polls can be biased towards one party (3) **Sparsity**: certain states might see few (or zero) polls	Empirically this method seems to perform the best.

Types of Forecasting Models

	Fundamentals Based	Polls Based	Fundamentals + Polls Hybrid
What They Are	Assume elections are based on things like **the economy, past voting history, incumbency, approval, etc**.	Use **national and state polls** prior to the election to reveal what voters preferences are.	Idea is to **combine fundamental predictors and polls** together to get the merits of both.
Pros / Cons	Historically there is not that much data since there have been few elections. As a result fundamentals based models tend to overfit.	Generally performs better than fundamentals based models. But polls have issues: (1) **Herding**: polls can try to copy one another (2) **House effects**: polls can be biased towards a party (3) **Sparsity**: certain states might see few (or zero) polls	Empirically this method seems to perform the best.

Fundamentals Based Models

We will build a fundamentals based model together because:

- Polling data is hard to find
- Fundamental models are simpler
- They provide a good benchmark
- We can construct one that is historically quite accurate

We will start with a cleaned data set that puts several sources together for simplicity.

Data Sources

MIT Election Lab	By State Election Results
US Elections Project	Voter Turnout
UCSB President Approval Data	Approval Data
IPUMs CPS Survey Data	Demographic Data
FRED Economic Database	National + State Economic Data

Agenda

1 Overview of how the US elects presidents

2 Stylized facts about voting

3 Types of forecasting models

4 **Building a fundamentals based forecasting model**

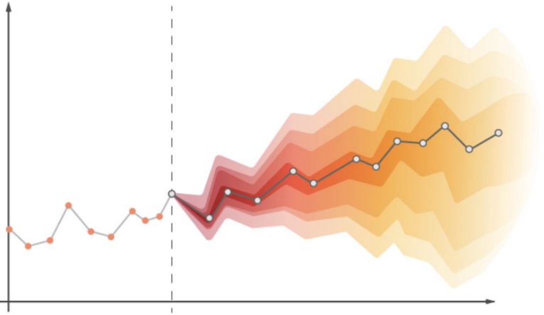
Building the Model

Key Principles

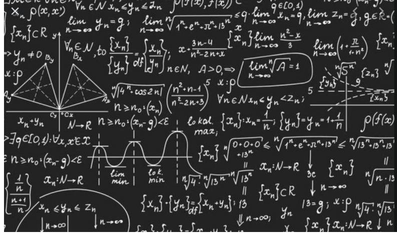
(1) Ignore Candidate Specifics (mostly irrelevant and hard to quantify) ✅
(2) Because the US is a two party system, we will ignore third parties ✅
(3) Make probabilistic forecasts to capture uncertainty in model predictions ✅
(4) We will focus on predicting Republican vote shares without loss of generality ✅

High Level Steps

(1) Build a model to predict vote shares in each state to determine which party will win (Republican or Democrat)
(2) Build a model to estimate errors / random noise in voting to account for uncertainty in forecasts
(3) Based on the model and error estimates, decide the winner of each state and therefore the number of electoral votes each party gets to find election winner
(4) Repeat steps (1) through (3) many times to simulate the probability of each party winning



The (High Level) Math



For each election and state, calculate the Republican two party vote share

$$\textbf{RepublicanVoteShare}_{\textbf{state, year}} =$$

$$\textbf{RepublicanVotes}_{\textbf{state, year}} \,/\, (\textbf{DemocraticVotes}_{\textbf{state, year}} + \textbf{RepublicanVotes}_{\textbf{state, year}})$$

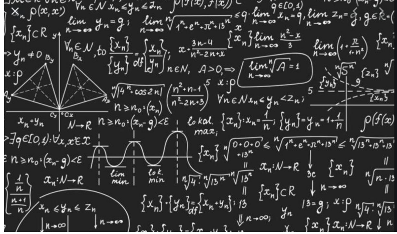
This is our **y variable of interest**.

If we predict it is above 0.5 for a state, then the Republicans are expected to win that state in a given election.

Calculating this for each state allows us to know who will win its electoral votes.*

*Note we ignore the complications of Maine and Nebraska for simplicity here.

The (High Level) Math



We will then build a model using a mixed effects regression of the form:

$$\mathbf{RepublicanVoteShare_{state,\ year} = X_{state,\ year} * B + NationalError + RegionalError + StateError}$$

B consists of (fixed effects) predictors like economic, approval, and demographic data by state (national level data like GDP is repeated for each state).

NationalError, RegionalError, and StateError are random effects (intercepts) and will be simulated from a distribution.

RegionalError is based on the census region a state is in. We can use this to *create correlation among the states.*

A single simulation draw from the model involves **drawing the three errors from their distributions and then adding them to the model prediction X*B**. Repeating this many times allows us to make a probabilistic forecast.

Mixed Effects Models Overview

Combine “fixed” and “random” effects into the same model

$$Y = X*B + Z*U + e$$

Here X is a matrix of fixed effects and Z is a matrix of random effects.

We use maximum likelihood to estimate:

(1) the parameters B (fixed effect coefficients); analogous to regular linear regression

(2) Variance parameters i.e. variance covariance matrix of U ~ MultivariateNormal(0,$\boldsymbol{\Sigma}$)

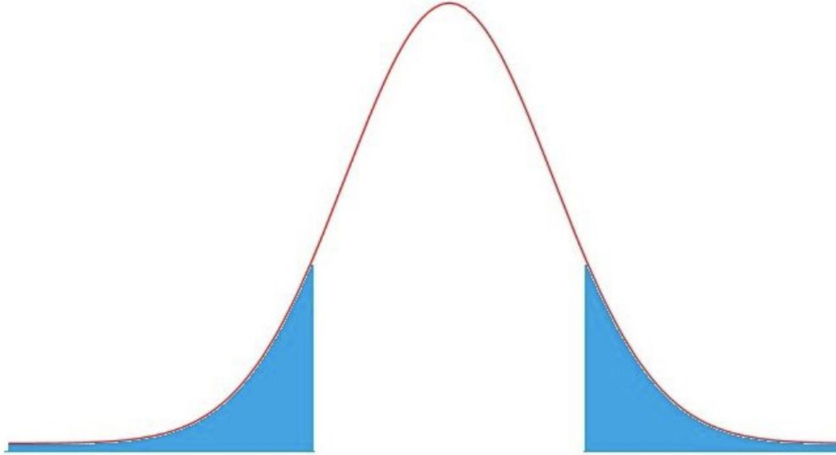
(3) Residual variance e ~ Normal (0,σ^2)

Random effects are random deviations from the overall average (note mean of zero) with an estimated variance. There are two basic types: (1) random intercepts and (2) random slopes, and they are made to vary across “groups” in a dataset.

Why Random Effects?

Across elections we see **national swings in voter sentiment** towards a particular party, and we also know that **certain states tend to vote similarly** to one another.

Therefore, to make probabilistic forecasts, we **need a way to account for potential national swings in voter sentiment and also account for the correlation between states**.



Why Random Effects? (Cont.)

Random effects (in this case random intercepts) allow us to **decompose the error of the model into specific pieces** that match what we expect intuitively:

Observation Variance = NationalError + RegionalError + StateError

We also get other benefits such as shrinkage to reduce overfitting and easily enabling simulations by drawing directly from the distributions of the random effects.

For more on mixed effects models and their benefits **see *Data Analysis Using Regression and Multilevel/Hierarchical Models by Gelman and Hill***.

Random Intercepts

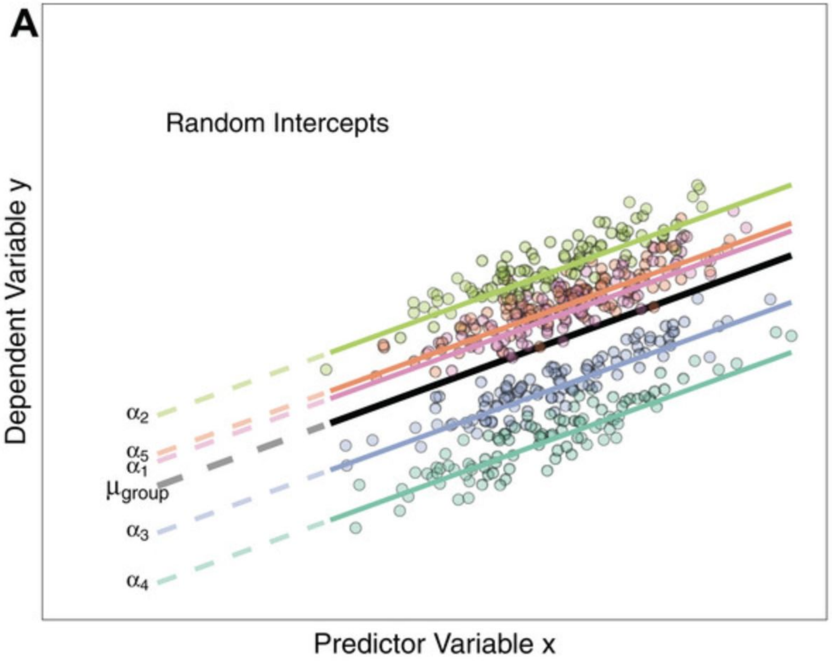
Random intercepts are deviations from the population intercept.

The model would look like $y_i = \alpha + \alpha_j$ where $\alpha_j \sim \text{Normal}(0,\sigma^2)$.

So the random intercepts follow the same normal distribution with mean zero and a fixed variance.

For example, suppose we create a random intercept in our model for each election year.

This would measure the deviation from the overall average across elections, and can then be thought of as the typical national swing we might expect.



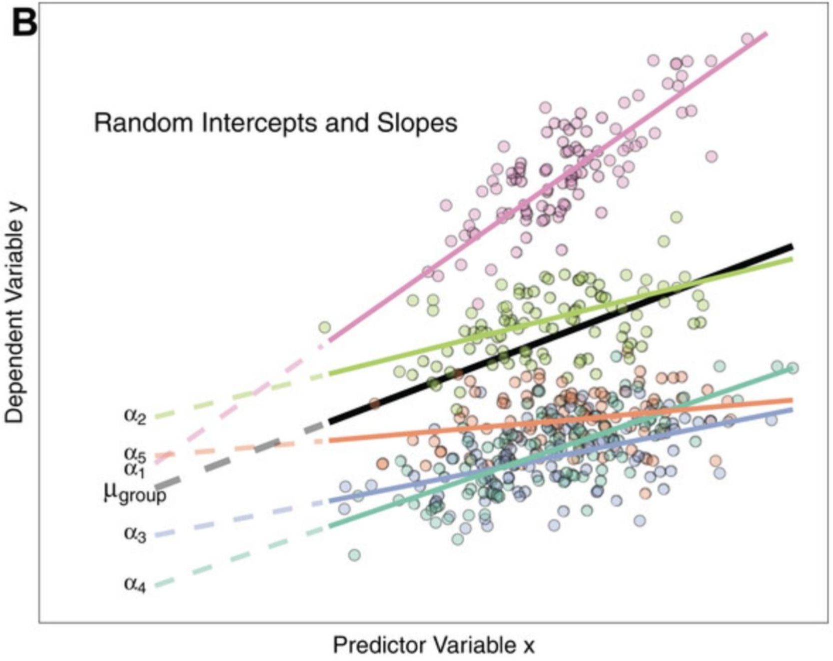
Random Slopes

Random slopes are deviations from the population slope.

The model would look like $y_i = \alpha + \beta x_i + \beta_j x_i$ where $\beta_j \sim \text{Normal}(0,\sigma^2)$.

So the random slopes follow the same normal distribution with mean zero and a fixed variance.

While we won't use random slopes in our forecasting model, they can help if we think the influence of a predictor differs across groups in a population (such as the impact of approval ratings by state).



Model Schematic

(1) Fit Model where $X_{\text{state, year}}$ are fixed effects and error terms are random intercepts:

$$\text{RepublicanVoteShare}_{\text{state, year}} = X_{\text{state, year}} * B + \text{NationalError} + \text{RegionalError} + \text{StateError}$$

(2) Calculate $E(\text{RepublicanVoteShare}_{\text{state, year}}) = X_{\text{state, year}} * B$ → This is our expected prediction

(3) Simulate errors

Draw from random intercepts distributions*

$$\text{NationalError}_{\text{Draw1}} \sim \text{StudentT}(0,\sigma_{NE}{}^2,df_{NE})$$
$$\text{RegionalError}_{\text{Draw1}} \sim \text{StudentT}(0,\sigma_{RE}{}^2,df_{RE})$$
$$\text{StateError}_{\text{Draw1}} \sim \text{StudentT}(0,\sigma_{SE}{}^2,df_{SE})$$

(4) Add errors to expected prediction:

$$X_{\text{state, year}} * B + \text{NationalError}_{\text{Draw1}} + \text{RegionalError}_{\text{Draw1}} + \text{StateError}_{\text{Draw1}}$$

(5) Repeat (3)-(4) many times to calculate our probabilistic forecast

*We use a StudentT distribution instead of a normal distribution because it has heavier tails and is more conservative. This requires us to specify an additional parameter for the degrees of freedom.

To R We Go!

Now let's go to R to build a fundamentals based forecasting model for the 2020 election.