WARWICK
THE UNIVERSITY OF WARWICK

Polls, prediction, and the POTUS

Probabilistic forecasting of elections

15 October 2020

Fernando Antonio Zepeda Herrera MSc. Statistics

Polls, prediction, and the POTUS

- You've probably heard that Joe Biden will beat Donald Trump, but we all remember that the same was said of Hillary Clinton, so, how likely is it that Biden wins it all?

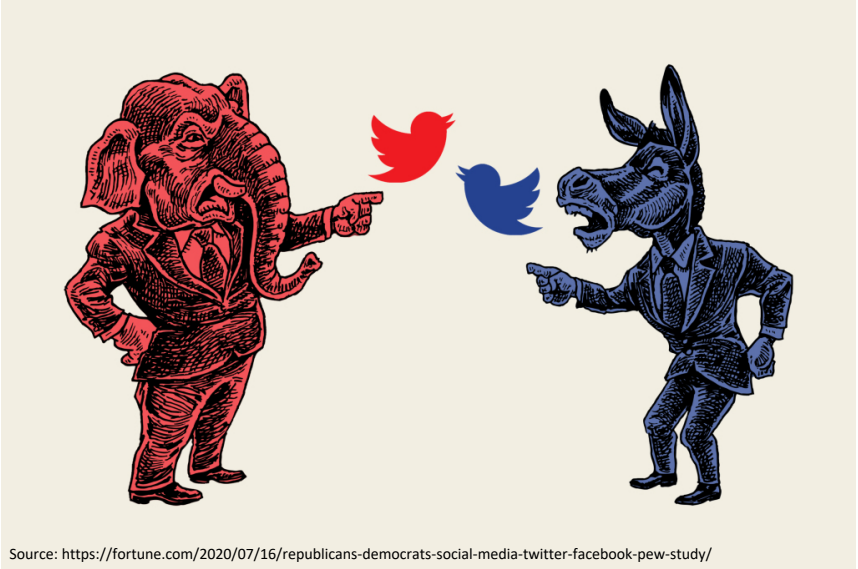
Source: https://tenor.com/view/biden-sunglasses-shades-gif-14042505

- In this talk we will explore as a case study how forecasting models for the POTUS election work.

Source: https://tenor.com/view/inauguration-cnn2017-donald-trump-ok-gif-7576940

Forecasting elections

- Interesting for several actors:
  - Candidates and campaigns
  - Media outlets
  - Political scientists
  - Twitter fighters
  - …

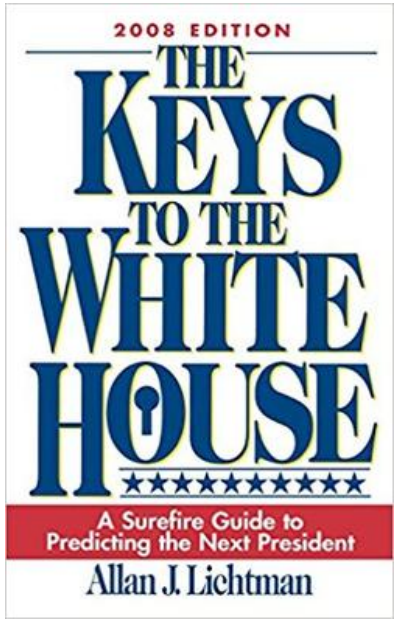


Source: https://fortune.com/2020/07/16/republicans-democrats-social-media-twitter-facebook-pew-study/

- In general, not an easy task:
  - Relatively short history
  - Political systems
  - Varying data sources and qualities

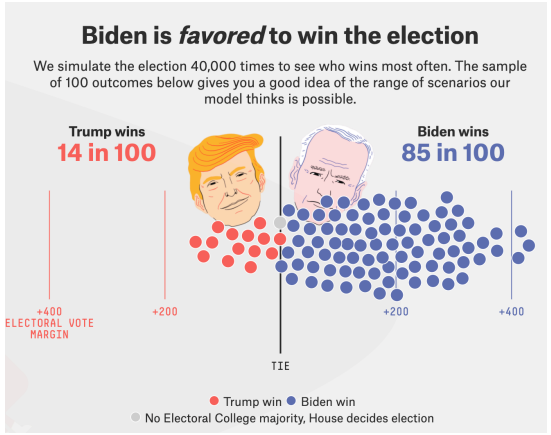
Different forecasts, the case with POTUS

Heuristics/Rule of thumb

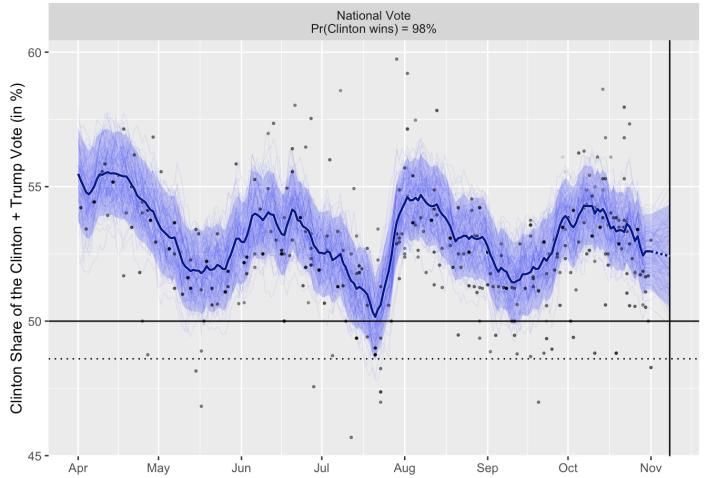


Source: https://en.wikipedia.org/wiki/The_Keys_to_the_White_House

Statistical modelling



Source: https://projects.fivethirtyeight.com/2020-election-forecast/



Source: http://www.slate.com/features/pkremp_forecast/report.html

Right now, our model thinks **Joe Biden is very likely to beat Donald Trump** in the electoral college.

	Chance of winning the electoral college	Chance of winning the most votes	Predicted range of electoral college votes (270 to win)
Joe Biden Democrat	around 9 in 10 or 92%	better than 19 in 20 or 99%	227-424
Donald Trump Republican	less than 1 in 10 or 8%	less than 1 in 20 or 1%	114-311

The probability of an electoral-college tie is <1%

Source: https://projects.economist.com/us-2020-forecast/president

No surprises here, we like statistical ones

- We like modelling
- Forecasting who will win matters,
- But… *how likely is it to win it all?*
- Managing uncertainty

WDSS
Warwick Data Science Society

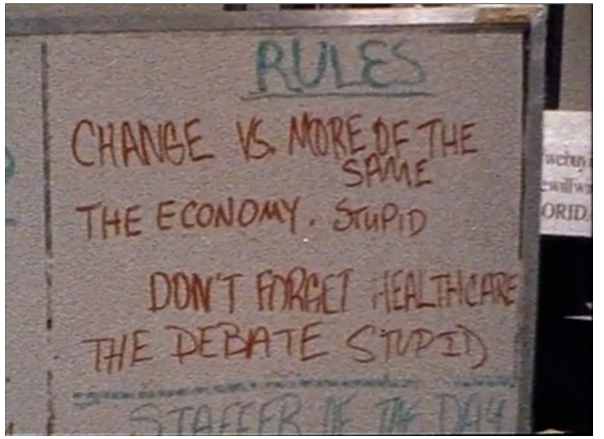
Source: https://tenor.com/view/b99-captainholt-holt-raymondholt-statistics-gif-8718500

Bayesian framework

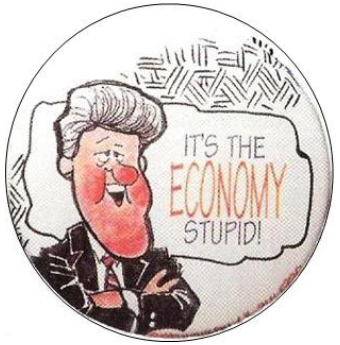
- Probability as a measure of uncertainty
- Joint data generating model with parameters, latent variables, etc.
- Alternatively, Bayes' Theorem:
  - Prior uncertainty on unknowns
  - Observed data gathered
  - Posterior, reassesed, uncertainty

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)}$$

Fundamentals



Source: https://www.business2community.com/marketing/3-economic-principles-will-transform-marketing-campaigns-01159950



Source: http://reconomy.org/the-economy-its-stupid/

Fundamentals

- Heuristics and rules of thumb forecasts suggest there are some structural or *fundamental* factors that would allow us to predict the election in advance reasonably well.

- Example: Abramowitz (2008)

  - Bad economy => TIME FOR CHANGE!

  - Bad incumbent => TIME FOR CHANGE!

  - Same party last 2+ terms => TIME FOR CHANGE!

Source: https://www.economist.com/united-states/2020/06/11/meet-our-us-2020-election-forecasting-model

Fundamentals

- We can use regression or any other tool to make a good prior probabilistic forecast. Examples,
  - The Economist selected their forecast and tried to avoid overfitting with elastic net regularisation and LOO-CV.
  - Nate Silver's 538 model seems to be more heuristical(?) this year, given COVID19, e.g. includes a NYT news component.
- This is important and non trivial! Domain knowledge matters.
- But it's the starting point.

Source: https://www.economist.com/united-states/2020/06/11/meet-our-us-2020-election-forecasting-model

Polls

Source: https://tenor.com/view/ijust-want-to-ask-afew-questions-asking-question-jack-ryan-season2-jack-ryan-gif-15467656

Let's start polling

- Electoral surveys are our main source of information about the state and evolution of a political campaign.

- Most basic *model* for a given poll:

Respondents prefering Biden

Total respondents revealing preferences

$$y|\pi \sim Binom(n, \pi)$$

Latent preference for Biden in population

Key assumptions

- Notion of random sample
  - Independent Bernoulli trials given a shared probability of voting for a candidate
  - Sampling error is the only quantified survey error
- Notion of snapshot
  - The population parameter of interest is interpreted as latent preferences or the result of a hypothetical election *held at the time of the survey*

Poll of polls

- We want to aggregate more than one poll throughout the campaign

$$\{(y_k, n_k)\} \text{ for } k = 1, 2, \ldots, K$$

- We will allow latent preferences to evolve in time from the start of the analysis $t = 1$ until the election at time $t = T$.

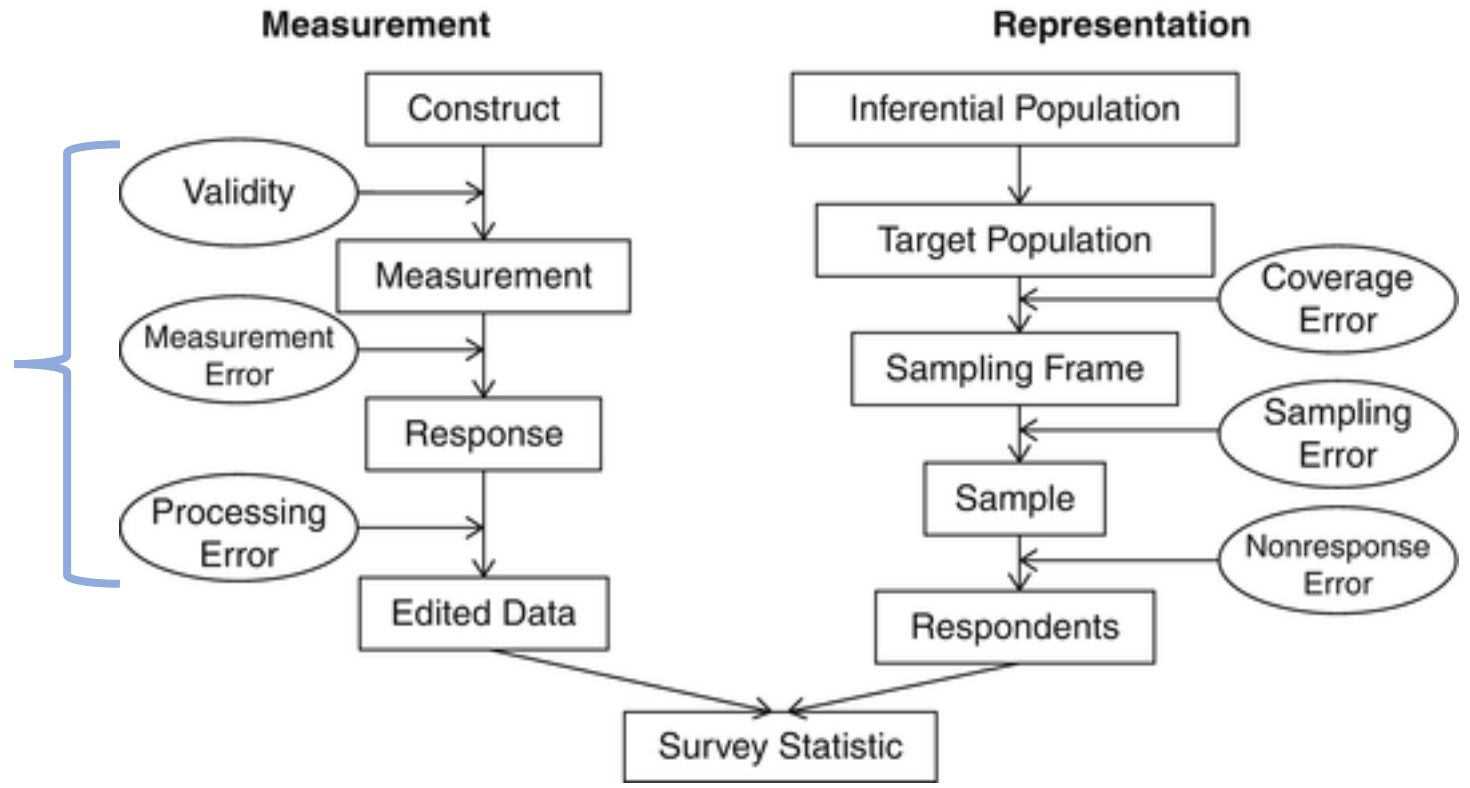
$$\underline{\pi} = (\pi_1, \pi_2, \ldots, \pi_T)$$

$$y_k | \underline{\pi} \sim Binom(n_k, \pi_{t[k]})$$

Total Survey Error

- But we can't forget about Total Survey Error!

Let's think about House Effects



Binomial likelihood only addresses this

Figure from Groves, R., Fowler, F., Couper, M., Singer, E., & Tourangeau, R. (2004). Survey methodology

Poll of polls

- We can then consider house effects via a GLM changing our parameter of interest decomposing it into “signal and noise”

- Survey estimates

  - measure latent preferences
  - disturbed by sampling error and non-sampling error attributed to house effects

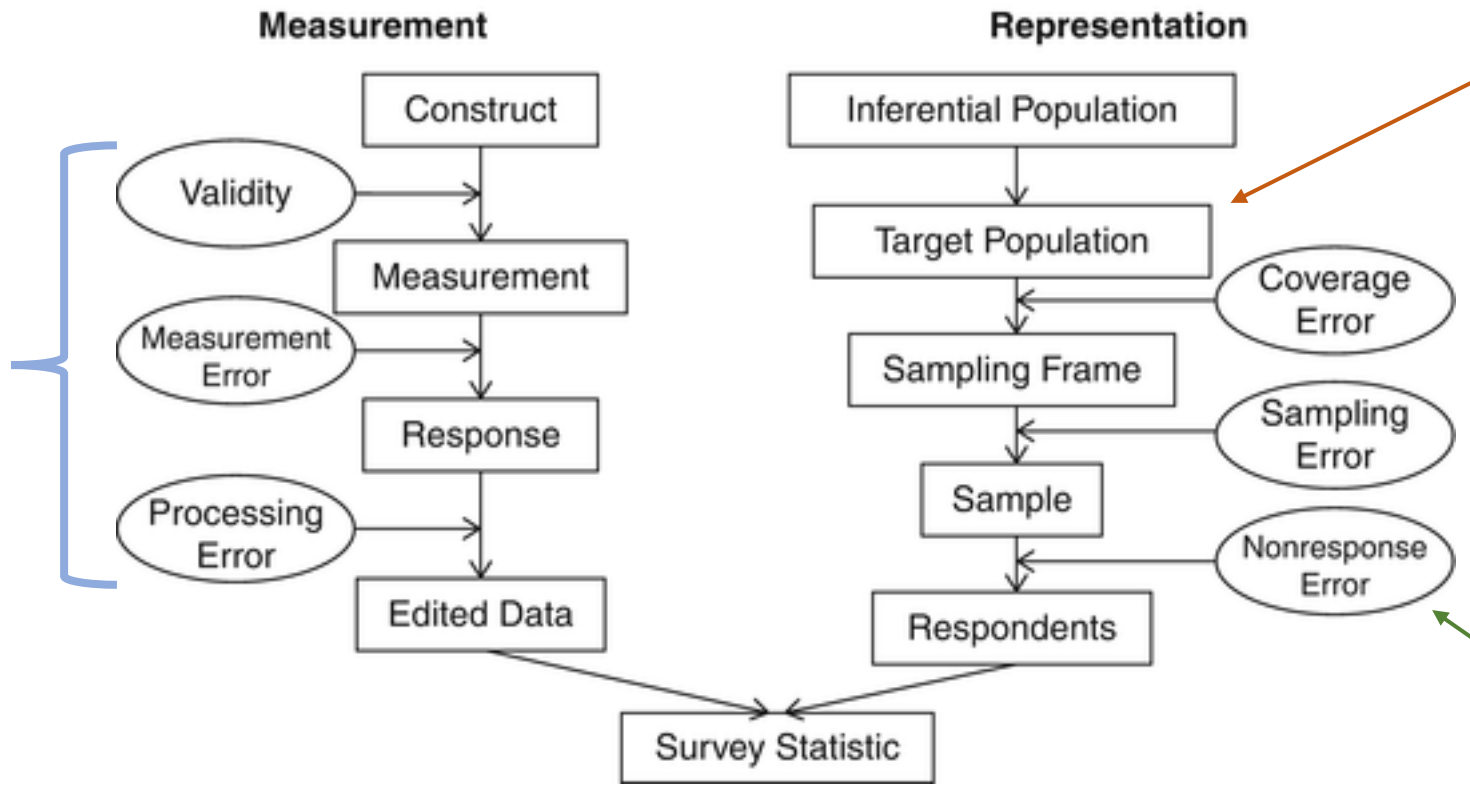
$$y_k|\underline{\pi} \sim Binom(n_k, \pi_k)$$

$$logit(\pi_k) = \mu_{t[k]} + \delta_{h[k]}$$

Total Survey Error

- But we can't forget about Total Survey Error!

Let's think about House Effects



Registered vs Likely voters
F2F vs Cell vs Web

Binomial likelihood only addresses this

Partisan bias

Figure from Groves, R., Fowler, F., Couper, M., Singer, E., & Tourangeau, R. (2004). Survey methodology

Poll of polls

- Other sources of error could be incorporated

- Survey estimates
  - measure latent preferences
  - disturbed by sampling error and
  - considering non-sampling error attributed to house effects population surveyed, partisan non-response bias, etc. (invariant in time)

$$y_k|\underline{\pi} \sim Binom(n_k, \pi_k)$$

$$logit(\pi_k) = \mu_{t[k]} + \underbrace{\delta_{h[k]} + \delta_{p[k]} + \delta_{r[k]}}_{\delta_k}$$

How do things evolve?

Back to the Future!

Back to the future

- Remember the fundamentals forecast? That's our prior guess as to how things will go *on Election day*
- Our poll of polls tells us how things are going on *at each point in time.*
- What if we travel backwards, from Election day until the first day of the analysis?
- How much things change between periods is controlled by another parameter.

$$\mu_1|\mu_2 \sim N(\mu_2, \sigma^2) \Longleftarrow \mu_t|\mu_{t+1} \sim N(\mu_{t+1}, \sigma^2) \Longleftarrow \mu_T \sim N(\mu_f, \sigma_f^2)$$

As Election Day approaches, polls dominate the forecast.

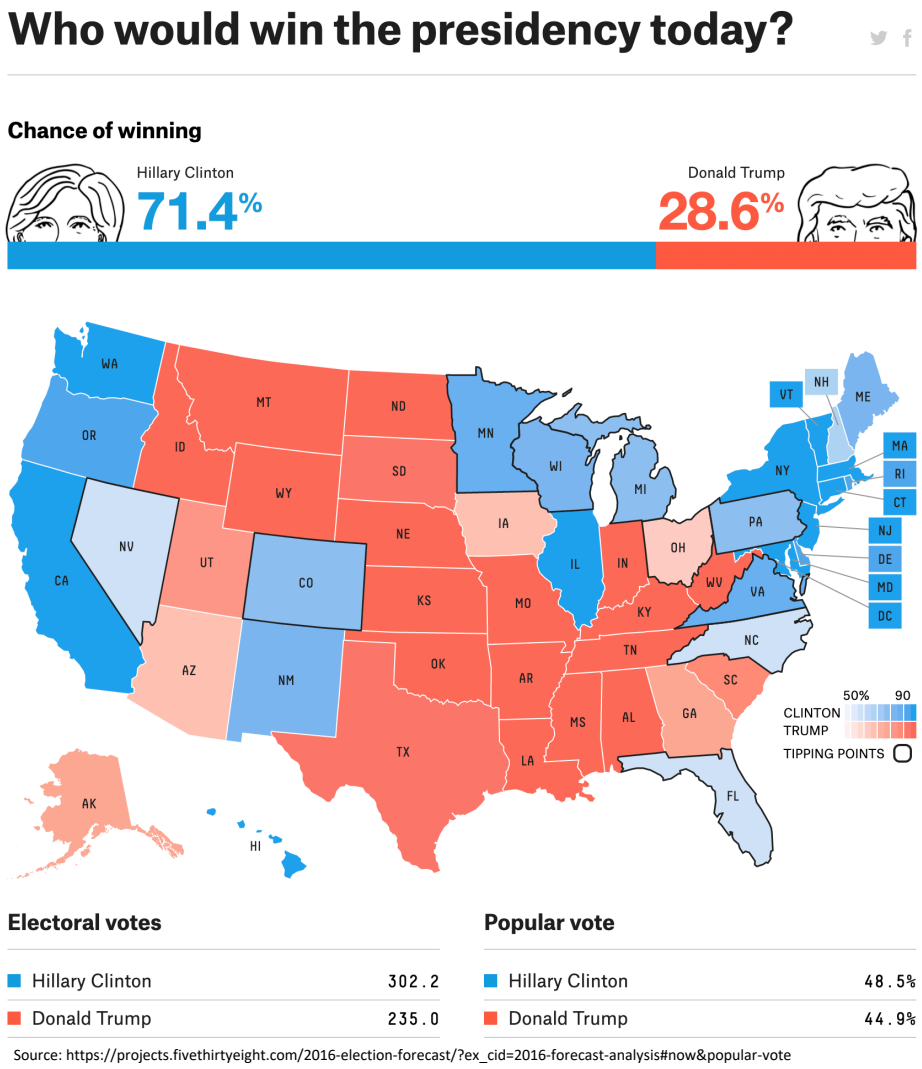
Source: hhttps://tenor.com/view/thats-it-igot-it-iknow-idea-dr-emmett-brown-gif-16500866

Well, actually not, but that's the general idea!

So… what about 2016?

https://tenor.com/view/but-the-logistics-were-off-wrong-statistics-mrs-maisel-gif-15627343

What about 2016?

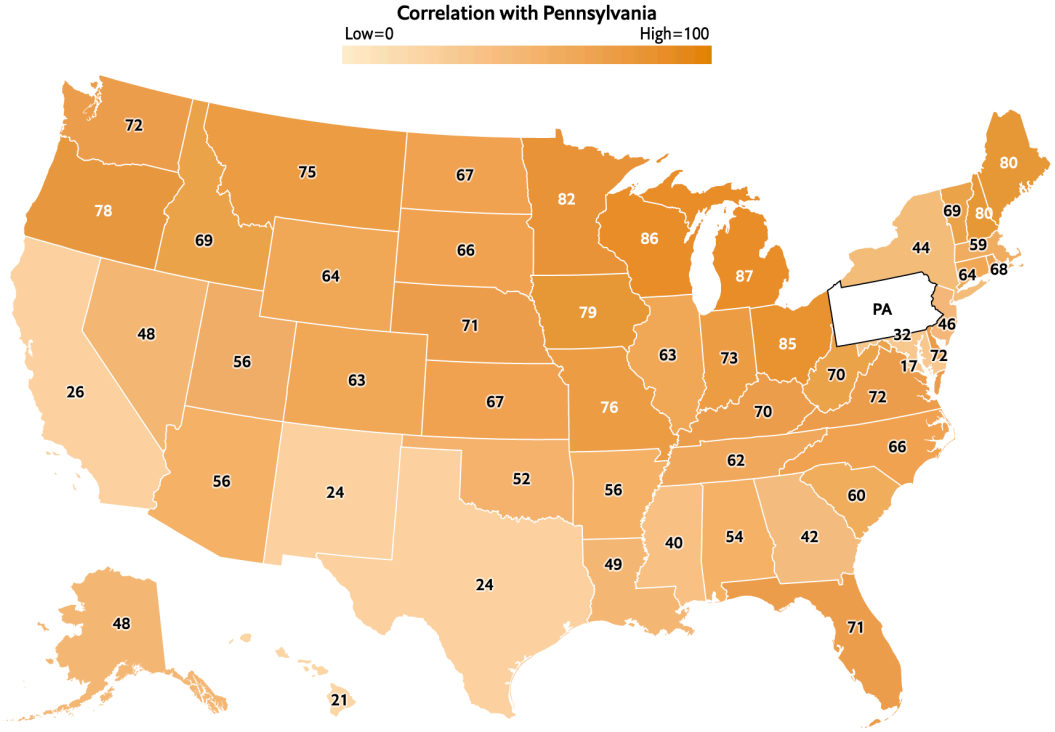


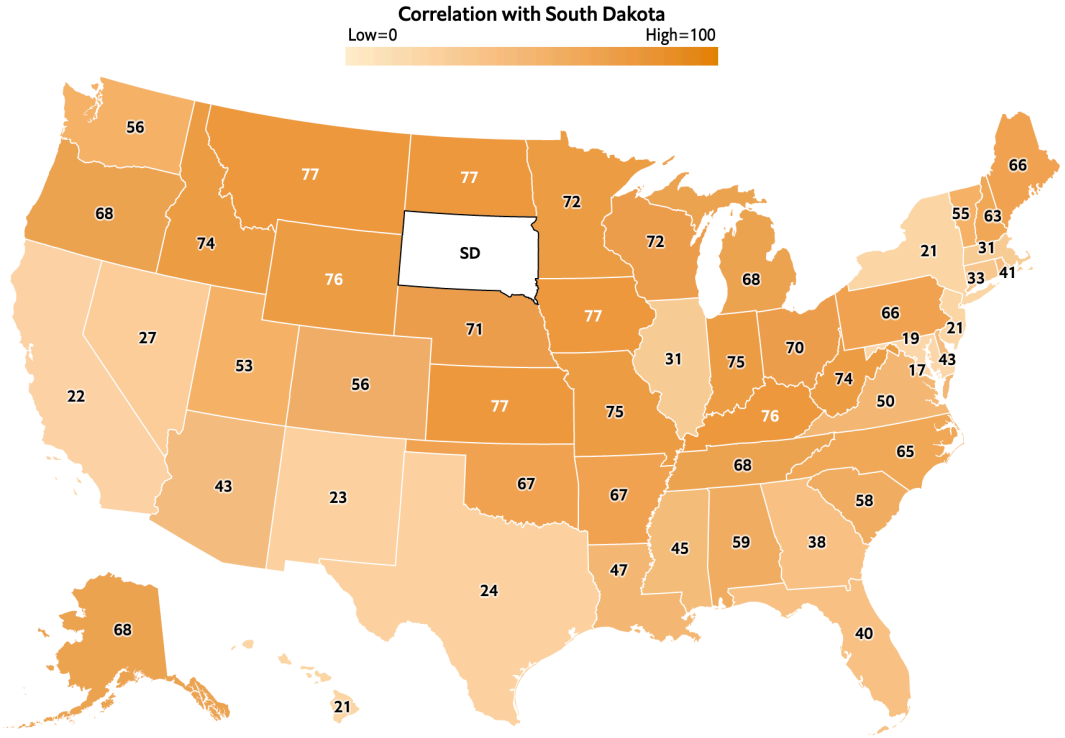
- Actual popular vote:
  - Clinton 48.2% vs Trump 46.1%
- We should model electoral votes!
- Relative side note
  - 25% probability events are “1 in 4”
  - An NFL kicker missing a field goal from about 40-45 yards
  - A 1 die attacker wins a unit against 2 dice deffenders in Risk

Correlation, correlation!

How states move together

Our model also simulates what would happen if the race moves, or the polls are biased, in similar amounts in like states. We calculate similarity between states by comparing their demographic and political profiles, such as the share of white voters who live there, how religious they are and how urban or rural the state is.





Source: https://projects.economist.com/us-2020-forecast/president

State polls

- We should model correlated state preferences
- We need state level polls!
- We can still allow for national trends

$$y_k|\underline{\pi} \sim Binom(n_k, \pi_k)$$

$$logit(\pi_k) = \mu_{s[k],t[k]} + \gamma_{t[k]} + \delta_k$$

State preference at the time of polling

Temporal national shocks

Placeholder for all of non-sampling error terms

Back to the future II

- Now we have vectors of preference at each point in time, but the idea remains very similar.

$$\underline{\mu}_{s,t} = \begin{pmatrix} \mu_{1,t} \\ \vdots \\ \mu_{50,t} \end{pmatrix}$$

KEY: Correlations between states via covariance matrix

$$\underline{\mu}_{s,t} | \underline{\mu}_{s,t+1} \sim N(\underline{\mu}_{s,t+1}, \Sigma) \Longleftarrow \underline{\mu}_{s,T} \sim N(\underline{\mu}_{s,f}, \Sigma_f)$$

Backwards evolution of both states preferences and national shocks

$$\gamma_t | \gamma_{t+1} \sim N(\gamma_{t+1}, \sigma_\gamma^2)$$

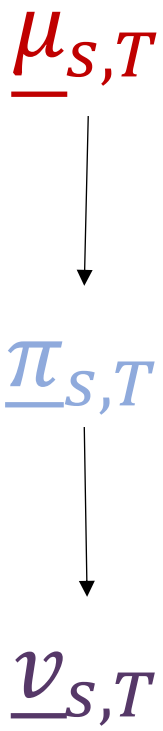
Fundamentals forecast, including no shock on election day (priced in)

$$\gamma_T = 0$$

Correlated states allow to "fill" the polling gaps in less surveyed states, there's shared information.

Modelled Electoral College

- Finally, we have latent preferences at Election Day.
- We can convert them back into vote shares for each state.
- The winner takes it all*: the winning candidate in each state gets assigned its electoral votes.

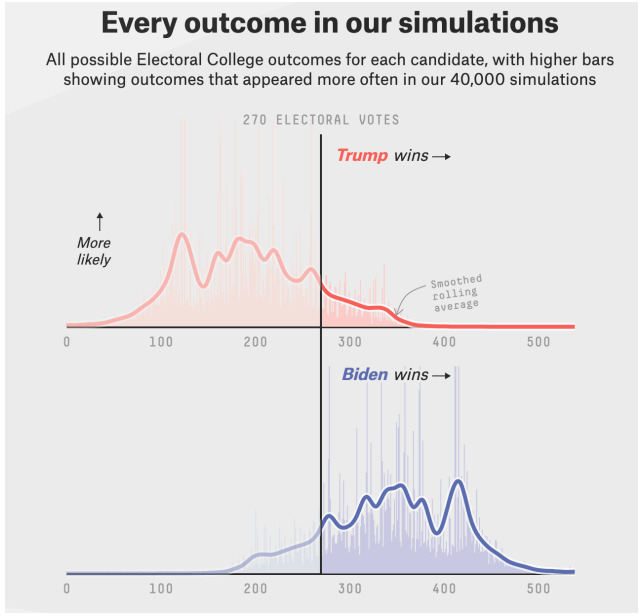


* Terms and conditions may apply… ME, NE

How likely is it to win it all?

- Why bother with all the notation? It means a probability model, and as such, we can simulate from it!
- In short, we simulate a lot of plausible trajectories, which gives us different simulated elections and winners.

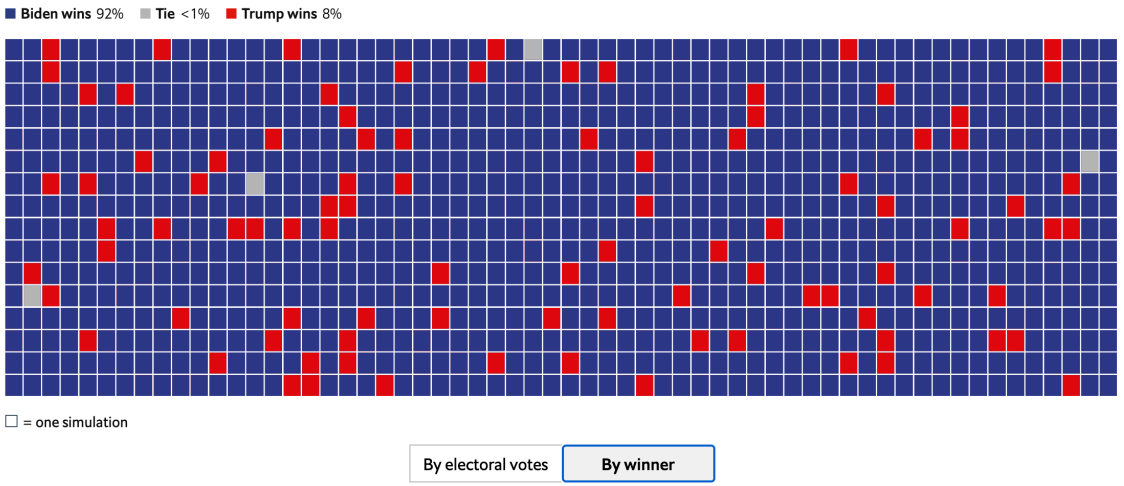
$$\begin{array}{ccccc} \underline{\mu}_{s,T} & \cdots & \underline{\mu}_{s,T} & \cdots & \underline{\mu}_{s,T} \\ \downarrow & & \downarrow & & \downarrow \\ \underline{\pi}_{s,T} & \cdots & \underline{\pi}_{s,T} & \cdots & \underline{\pi}_{s,T} \\ \downarrow & & \downarrow & & \downarrow \\ \underline{v}_{s,T} & \cdots & \underline{v}_{s,T} & \cdots & \underline{v}_{s,T} \end{array}$$

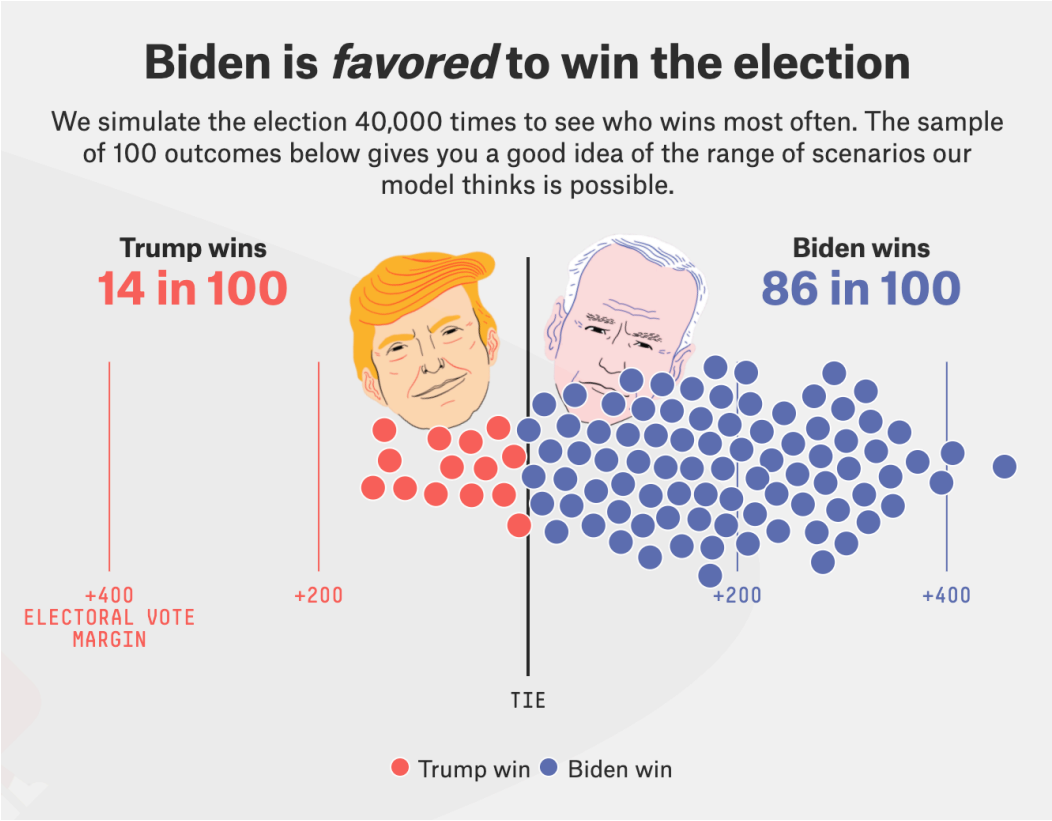


How likely is it to win it all?

Electoral-college simulations

Our model works by simulating 20,000 paths for the election, each time varying candidates' vote shares to account for polling error, changes in turnout or the political environment and the effects of campaigning. The bars below represent the predicted likelihood of every plausible electoral-vote outcome.





Thank you!

Fernando Antonio Zepeda Herrera MSc. Statistics

Some references

- The Economist
  - https://projects.economist.com/us-2020-forecast/president
  - https://github.com/TheEconomist/us-potus-model
  - https://projects.economist.com/us-2020-forecast/president/how-this-works
- Fivethirtyeight
  - https://projects.fivethirtyeight.com/2020-election-forecast/
  - https://fivethirtyeight.com/features/how-fivethirtyeights-2020-presidential-forecast-works-and-whats-different-because-of-covid-19/
- Abramowitz' Time for Change
  - https://www.jstor.org/stable/20452296?refreqid=excelsior%3A2396684a98b138d72acca751e6032897&seq=1
- Drew Linzer
  - https://votamatic.org/wp-content/uploads/2013/07/Linzer-JASA13.pdf
- Pierre-Antoine Kremp
  - http://www.slate.com/features/pkremp_forecast/report.html
- My own Poll of Polls for Mexico's 2018 Presidential Election (no fundamentals, in Spanish)
  - https://www.fazepher.me/project/01_poll-of-polls-2018/