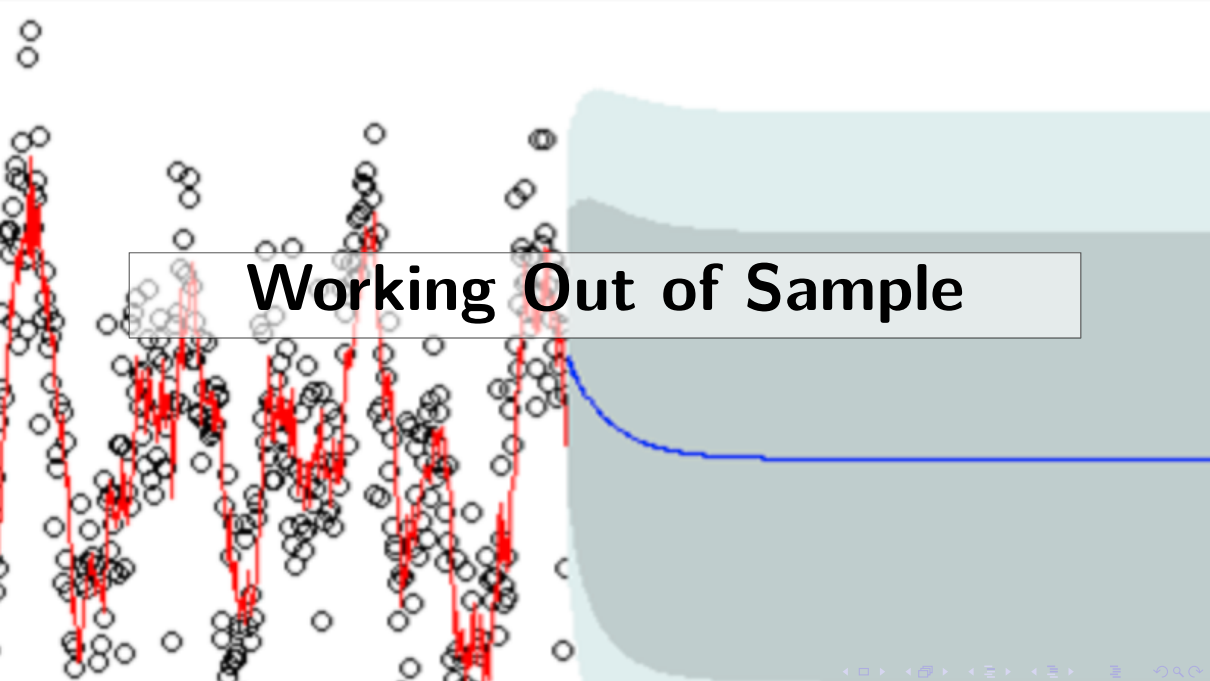
Working Out of Sample



Today:

- Introduce the ideas of political forecasting;
- Discuss these ideas in the context of the 2016 election;
- Talk about out of sample work in relation model assessment.

Working out of sample

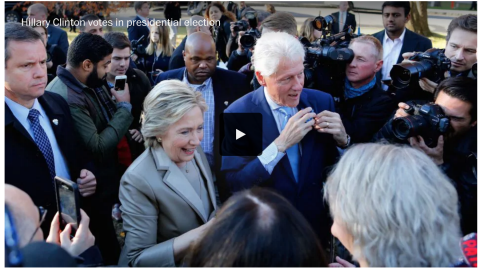
- Major question: what does your model say about data not used to create it?

- Two motivations for working out of sample:
  1. **Forecasting** – what does your model say about the future?
     - Used to drawn conclusions and make inferences;
     - Model is not generally under evaluation;
  2. **Assessment** – does your model predict well outside of the data on which it was made?
     - Used to evaluate the model and make sure it isn't 'overfit' to the data it was made with;
     - Model is very much under evaluation!

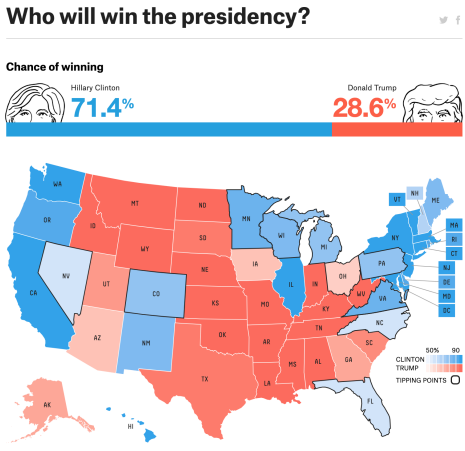
Forecasting – what does your model say about the future?

Let's forecast the 2016 Election – whoops!

A comprehensive average of election forecasts points to a decisive Clinton victory



Democratic presidential candidate Hillary Clinton voted in Chappaqua, N.Y., Nov. 8, accompanied by her husband. (Photo: EDUARDO MUNOZ ALVAREZ/The Washington Post)



FiveThirtyEight is a well known election forecasting model...

- How does it work? Some key facts that underpin how it uses data:
  - Electoral college votes are allocated at state level;
  - National vote $\neq$ state vote $\Rightarrow$ national vote $\neq$ electoral college vote;
  - Popular vote $\neq$ electoral college vote;
  - Probability of winning $\neq$ Popular vote $\Rightarrow$ probability of winning $\neq$ vote percentage;
  - Shouldn't give a "point prediction"—forecasts should incorporate **uncertainty**.

The Media Has A Probability Problem

The media's demand for certainty — and its lack of statistical rigor — is a bad match for our complex world.

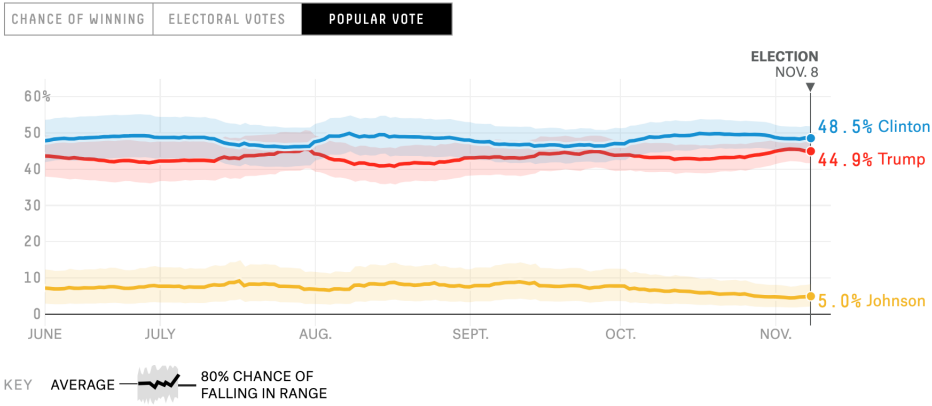
By Nate Silver

Filed under The Real Story Of 2016

Published Sep. 21, 2017

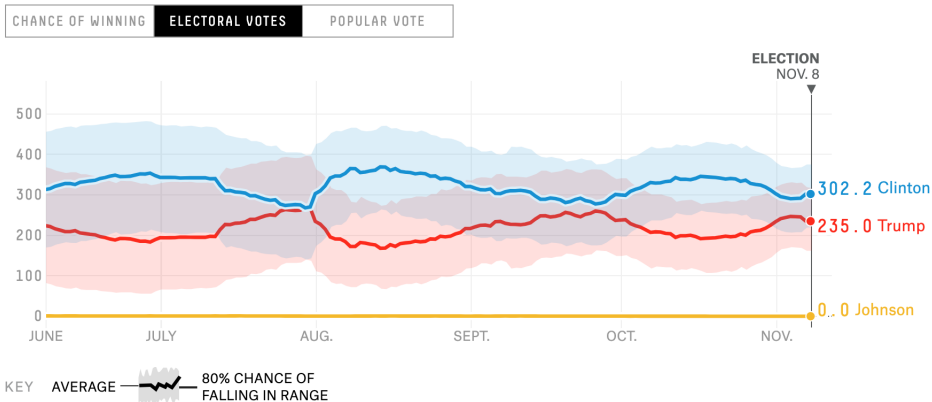
How the forecast has changed

We'll be updating our forecasts every time new data is available, every day through Nov. 8.



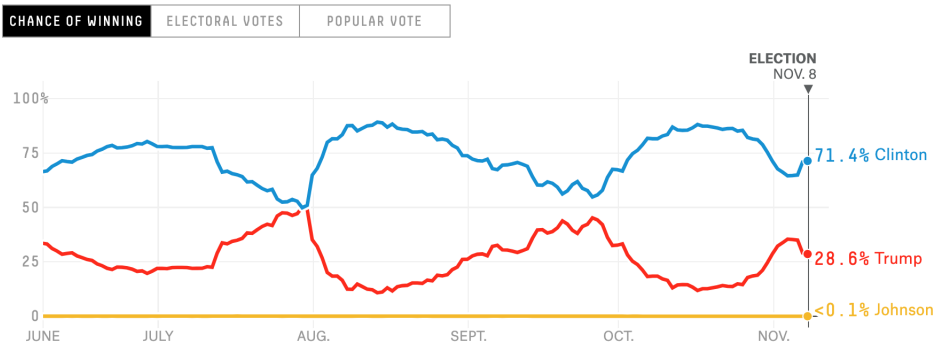
How the forecast has changed

We'll be updating our forecasts every time new data is available, every day through Nov. 8.



How the forecast has changed

We'll be updating our forecasts every time new data is available, every day through Nov. 8.



What does this 71/28 probability mean?

- Out of all the possible scenarios...
  - In each state, either Trump or Clinton will win – 2^{50} possibilities;
  - How many ways can you configure the electoral college map (2^{50})?
- ...that are consistent with our data...
  - Not all scenarios are equally likely
  - We learn about the plausibility of scenarios from history and polls
- ...how often is Clinton the winner?
  - A poll has Clinton 1 pt. below Trump
  - What is the range over what Clinton's "true vote" could be?
  - How many of the "true votes" in that range are greater than Trump's?

Building blocks

- FiveThirtyEight uses fundamentals (and historical data) to build "prior expectations" for Election Day;
  - Anchors each state as the model projects to Election Day;
  - Explores how each state differs from the average;
  - When state predictions are wrong, allows assessment of which states are "wrong together;"
- New polls "update our prior expectations" about Election Day.

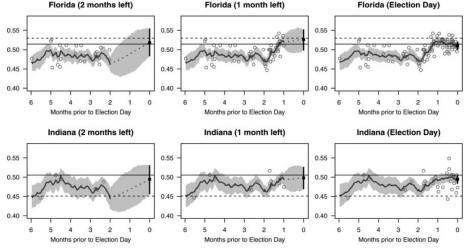


Figure 3. Forecasting the 2008 presidential election in real time. Results are shown for Florida and Indiana. The vertical axis is the percentage

Three types of error

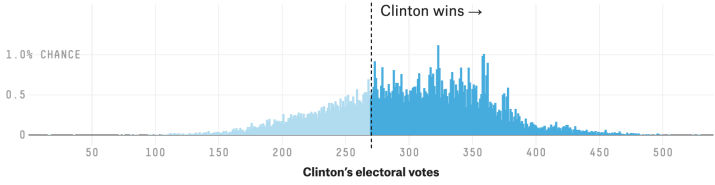
Each simulation accounts for three potential types of error and uncertainty:

- National error. The polls are systematically off throughout the country.
- Demographic and regional error. The polls are off in states that have demographic or geographic factors in common.
- State-specific error. The polls are off in a particular state, with no effect on other states.

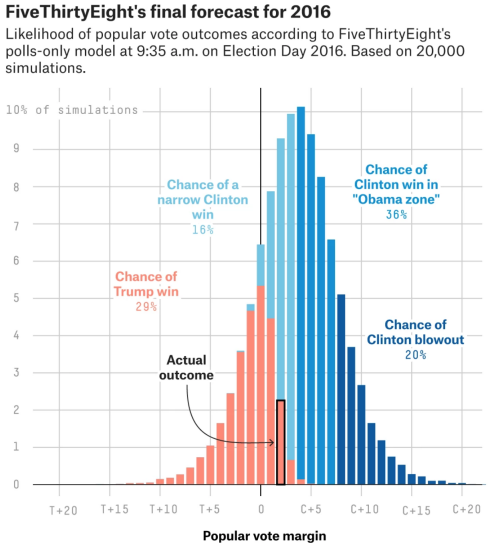
What scenarios were compatible with the data?

What to expect from the Electoral College

In each of our simulations, we forecast the states and note the number of electoral votes each candidate wins. That gives us a distribution for each candidate, where the tallest bar is the outcome that occurred most frequently.



What scenarios were compatible with the data?



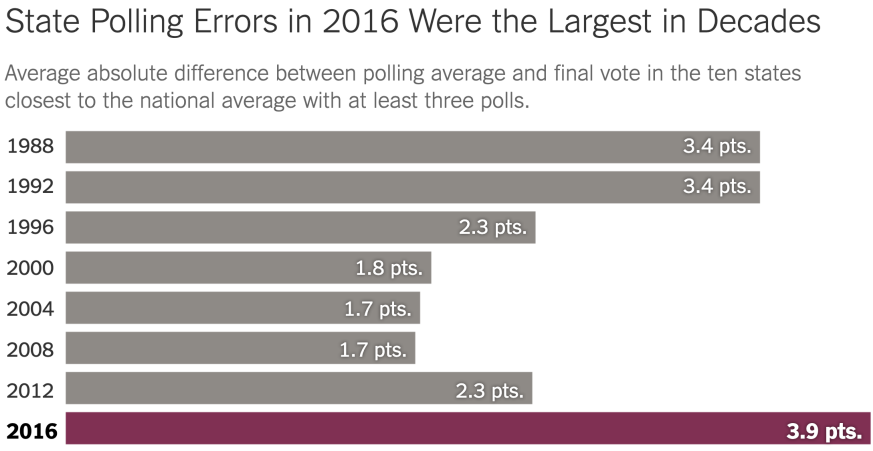
What scenarios were compatible with the data?

Crazy and not-so-crazy scenarios

Here are the chances we'll see these election outcomes.

Electoral College deadlock no candidate gets 270 electoral votes	1.0%
Electoral College 269-269 tie	0.5%
Recount at least one decisive state within 0.5 ppt	8.3%
Clinton wins popular vote	81.4%
Trump wins popular vote	18.6%
Clinton wins popular vote but loses Electoral College	10.5%
Trump wins popular vote but loses Electoral College	0.5%
Johnson wins at least one electoral vote	0.3%
McMullin wins at least one electoral vote	13.5%
Clinton majority wins at least 50 percent of the vote	28.7%
Trump majority wins at least 50 percent of the vote	2.3%
Clinton landslide double-digit popular vote margin	6.1%
Trump landslide double-digit popular vote margin	0.3%
Map exactly the same as in 2012	0.2%
Clinton wins at least one state Mitt Romney won in 2012	71.6%
Trump wins at least one state President Obama won in 2012	85.0%

What went wrong?!



Projecting confidence: How the probabilistic horse race confuses and demobilizes the public

Sean J. Westwood[1], Solomon Messing[2], and Yphtach Lelkes *[3]

[1]Program in Quantitative Social Science, Dartmouth College
[2]Data Labs, Pew Research Center
[3]Annenberg School of Communication, University of Pennsylvania

February 6, 2018

Assessment – does your model predict well outside of the data on which it was made?

How do we vet models?

- Much of class has focused on **inference** – our ability to reason about a **population** from a **sample**;
  - Average causal effect = estimating general effects of a treatment;
  - Hypothesis testing = inference about population parameters, e.g. the number of campaign supporters in the voting population;
  - Linear/logistic regression = measuring the effect of an independent variable on a dependent variable;

How do we vet models?

- Much of class has focused on **inference** – our ability to reason about a **population** from a **sample**;
  - Average causal effect = estimating general effects of a treatment;
  - Hypothesis testing = inference about population parameters, e.g. the number of campaign supporters in the voting population;
  - Linear/logistic regression = measuring the effect of an independent variable on a dependent variable;

- Something monumental happened last week – as soon as we started to talk about 'good' models we also started to talk about **prediction** rather than inference;
  - RMSE – typical model **prediction** error;
  - R^2 – typical model **prediction** error compared with a naive model;
  - F-test – hypothesis test of fit (measured by **prediction** error);
  - Confusion matrix – binary **prediction** error aggregated across the data;
  - Accuracy – binary percent correctly **predicted**;
  - Precision/recall – measurement of **prediction** quality;

How do we vet models using out of sample work?

1. Divide the data into a **training** set and a **test** set;
   - No observations in common;
   - Test set is usually smaller;
   - Same dependent variable distribution across the two;

2. Create the model (e.g. linear regression) using the training data;

3. Use the model created with the training data to predict the test data – compute metrics!

Why should we care?

Out of sample work is where the model becomes useful for helping us to control our world.