Election Forecasting

Predicting the Winner Before any Votes are Cast

15.071 – The Analytics Edge

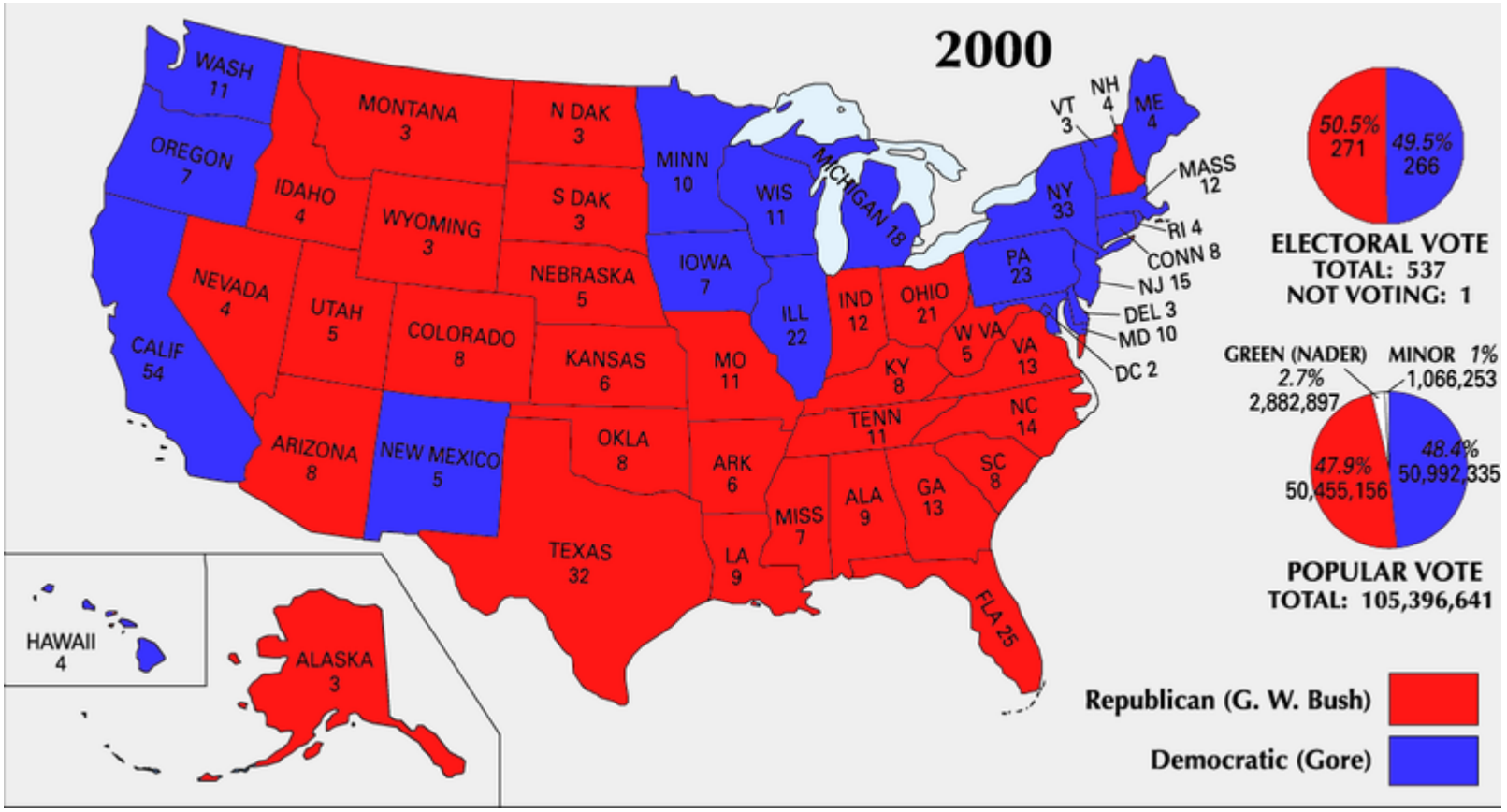
United States Presidential Elections

- A president is elected every four years
- Generally, only two competitive candidates
  - Republican
  - Democratic

The Electoral College

- The United States have 50 states
- Each assigned a number of *electoral votes* based on population
  - Most votes: 55 (California)
  - Least votes: 3 (multiple states)
  - Reassigned periodically based on population change
- Winner takes all: candidate with the most votes in a state gets all its electoral votes
- Candidate with most electoral votes wins election

2000 Election: Bush vs. Gore



Election Prediction

- Goal: Use polling data to predict state winners

- Then-*New York Times* columnist Nate Silver famously took on this task for the 2012 election

The Dataset

- Data from RealClearPolitics.com
- Instances represent a state in a given election
  - *State*: Name of state
  - *Year*: Election year (2004, 2008, 2012)
- Dependent variable
  - *Republican*: 1 if Republican won state, 0 if Democrat won
- Independent variables
  - *Rasmussen, SurveyUSA*: Polled R% - Polled D%
  - *DiffCount*: Polls with R winner – Polls with D winner
  - *PropR*: Polls with R winner / # polls

Election Forecasting

Predicting the Winner Before any Votes are Cast

15.071 – The Analytics Edge

Simple Approaches to Missing Data

- Delete the missing observations
  - We would be throwing away more than 50% of the data
  - We want to predict for all states
- Delete variables with missing values
  - We want to retain data from Rasmussen/SurveyUSA
- Fill missing data points with average values
  - The average value for a poll will be close to 0 (tie between Democrat and Republican)
  - If other polls in a state favor one candidate, the missing one probably would have, too

Multiple Imputation

- Fill in missing values based on non-missing values
  - If Rasmussen is very negative, then a missing SurveyUSA value will likely be negative
  - Just like *sample.split*, results will differ between runs unless you fix the random seed
- Although the method is complicated, we can use it easily through R's libraries
- We will use Multiple Imputation by Chained Equations (mice) package