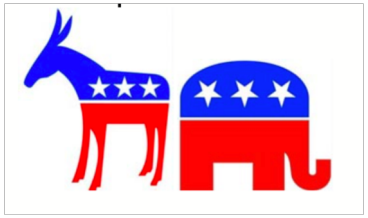


# Assessing Candidate Preference through Web Browsing History


Mark Crovella

with Giovanni Comarela, Ramakrishnan Duraraijan,

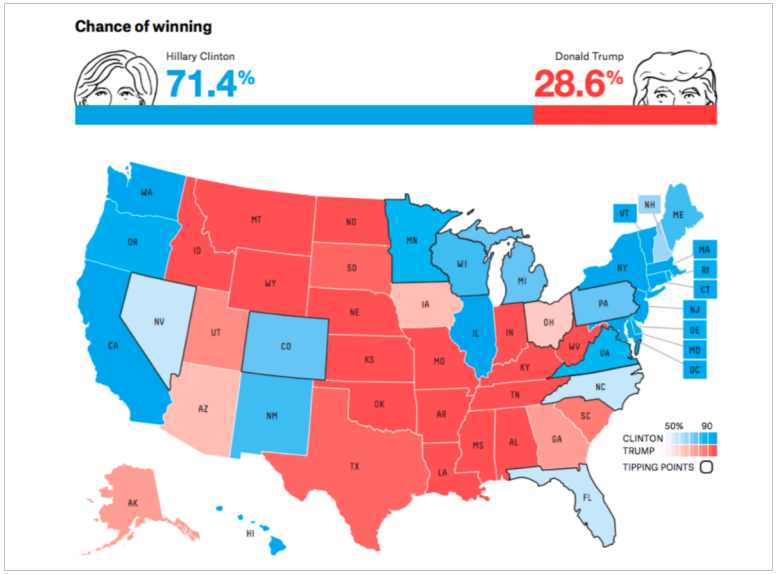
Paul Barford, and Dino Christenson







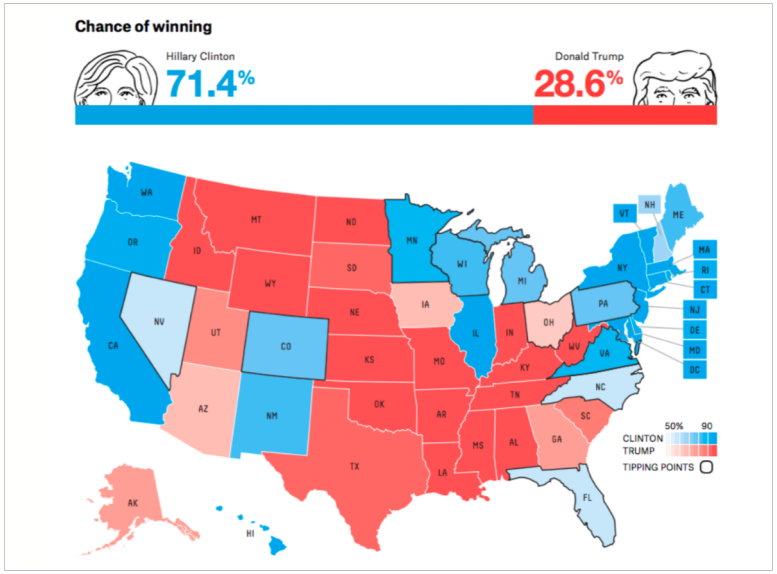

# Who cares about electorate opinion?




?


- Candidates
- Political Scientists
- Everyday People!

# Why not use polls?




?


- Require multiple days to complete
- Subject to interviewer effects, question wording, non-responsive, non-forthcoming subjects
- Random polling increasingly difficult

# What would be ideal?

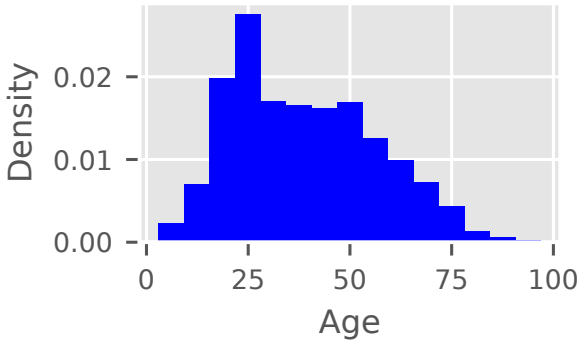
- Passive observation
  - Avoids respondents having to interpret questions or recall their own behavior
- Fine grained in space and time
  - US presidential elections are winner-take-all at the state level
  - requires predictions in 50 states + DC
  - Some questions require district level and even neighborhood level predictions
- Inference from behavior, not speech
  - Text analysis can be challenging
  - Statements do not always reflection opinions

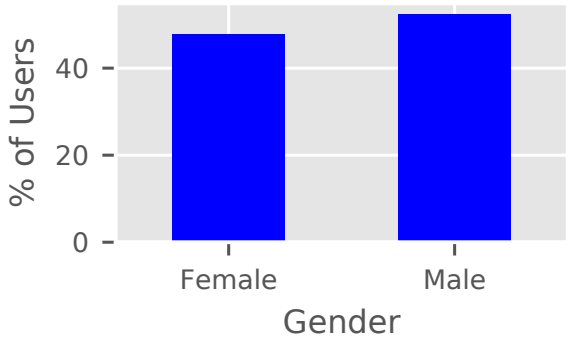
# A Potential Solution: Web Browsing Histories

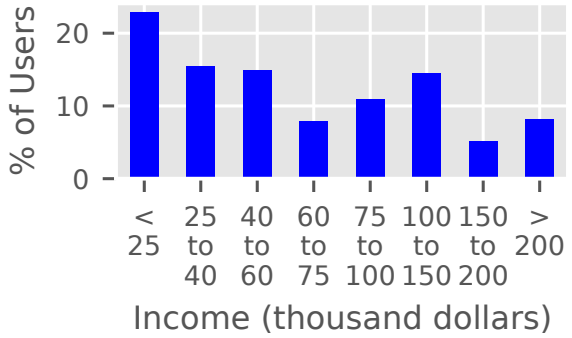
- Consider obtaining the Web browsing records of a large cohort of individuals
  - ✔ Passive ✔ Fine Grained ✔ Behavioral
  - ✘ But … how to obtain?
- In fact, (many) such cohorts exist:
  - comScore, hitWise, nielson NetRatings, …
  - Users are *compensated* for their data
  - Full informed-consent for data use and sharing
- Today, we will use one such cohort: comScore's US Web desktop panel

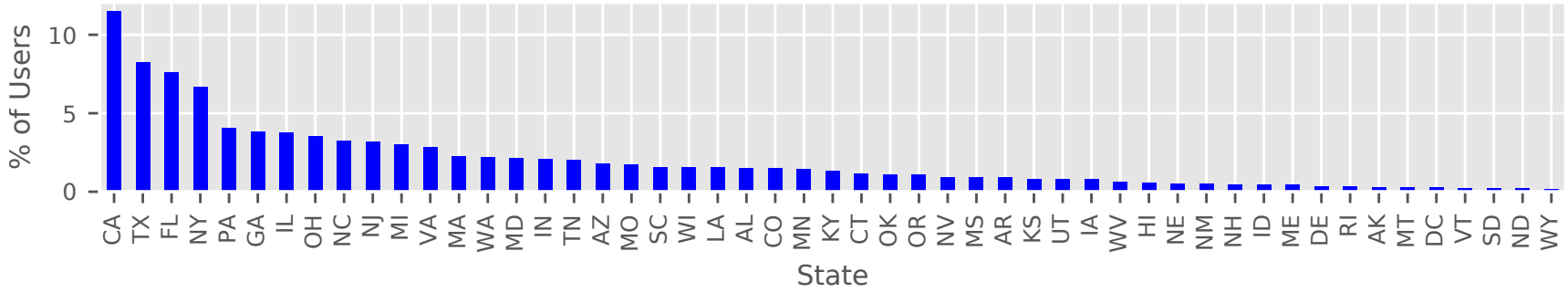
# The comScore Data

- About 120,000 panelists
- 56 days: from Sept 9, 2016 to November 3, 2016
- Complete browsing records including headers
- 70M unique URLs, 380K unique sites



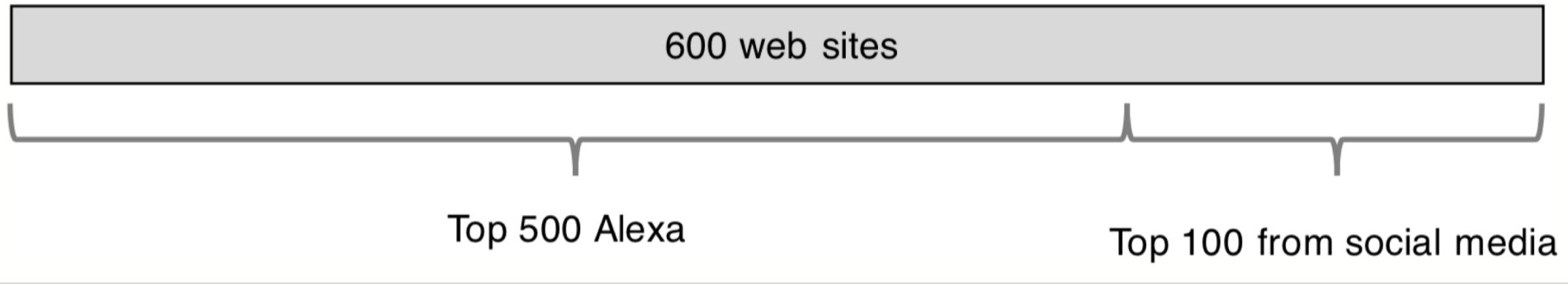



# Research Challenges

1. What features of Web browsing are most informative?
   - Detail vs data size tradeoff
2. How to apply ML?
   - Per-user labels are not available
3. Achieving Spatial Granularity
   - Need per-state predictions
4. Temporal Heterogeneity
   - User activity varies qualitatively over the week

# Feature Selection

- Sites? URLs? Headers? Query terms? Content?
- Using sites gives largest amount of data
- Using *referrals* increases signal
- Per user per day feature vector:



# Applying ML

One does not simply … train a classifier

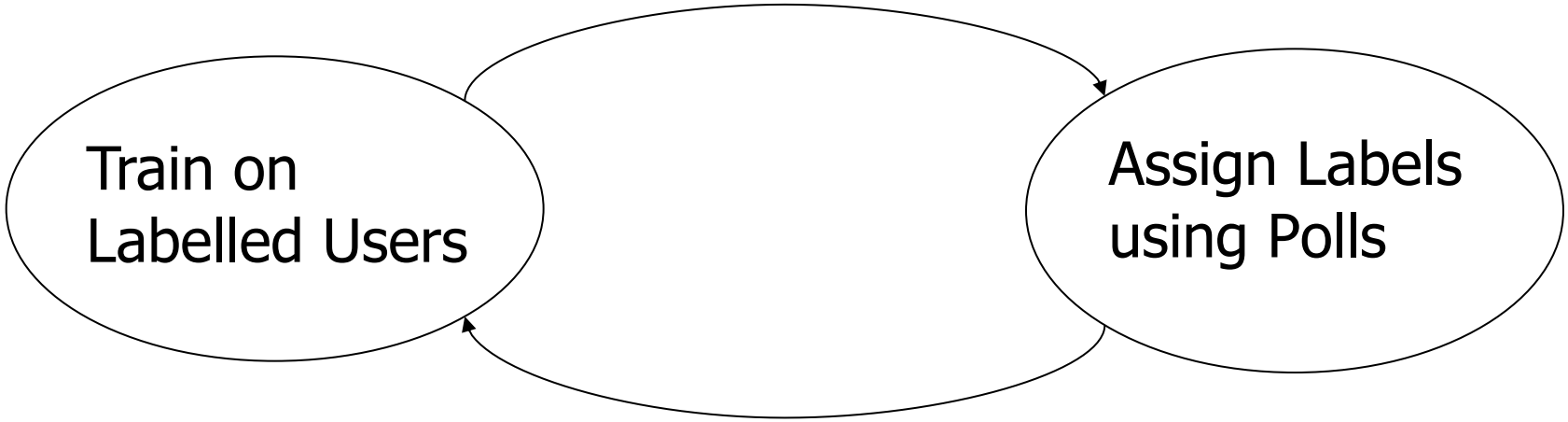
Our approach is to train using polling data, and then predict *forward in time* on a fine spatial and temporal scale

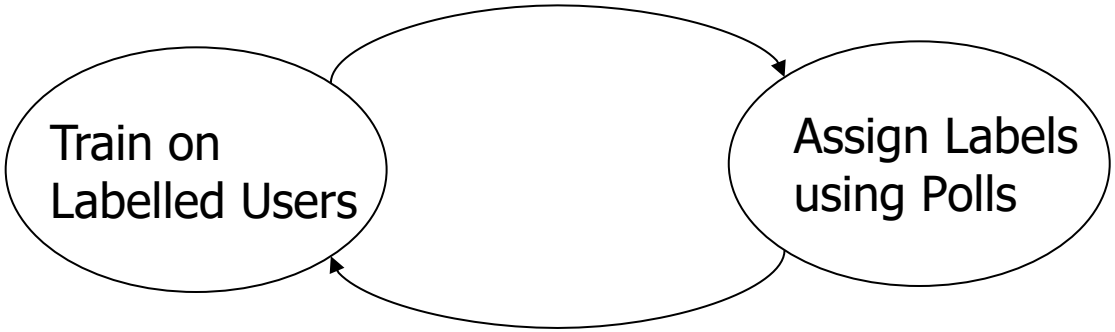
How to use polling data?

- Simple strategy:
  - Use per-state polls
  - Assign each user the majority opinion for the state
- Can we do better?

# An EM Algorithm

- Assign each user a label
- Train a classifier using those labels
  - logistic regression
- Now, assign new labels
- Repeat






Key step is label assignment –

Initially, use state majority labels

Subsequently, refine using per-state polls

For each state:

*rank* the users of that state using the classifier

for a state that polled *p* percent Republican,

assign the top *p* percent that label

**Algorithm 1:** `EM-training`

**Data:** $\mathbf{U}$, $R$, $B$, $\mathcal{A}$

$L' \leftarrow$ `InitLabels`($U$, $R$, $B$)

**repeat**

  $L'' \leftarrow L'$

  $\Theta \leftarrow$ train $\mathcal{A}$ on ($\mathbf{U}$, $L'$)

  **foreach** *region* $r$ **do**

    $t_r \leftarrow$ percentile $(1 - B(r))$ of $P(L(u) = 1|\Theta)$, $\forall u \in r$

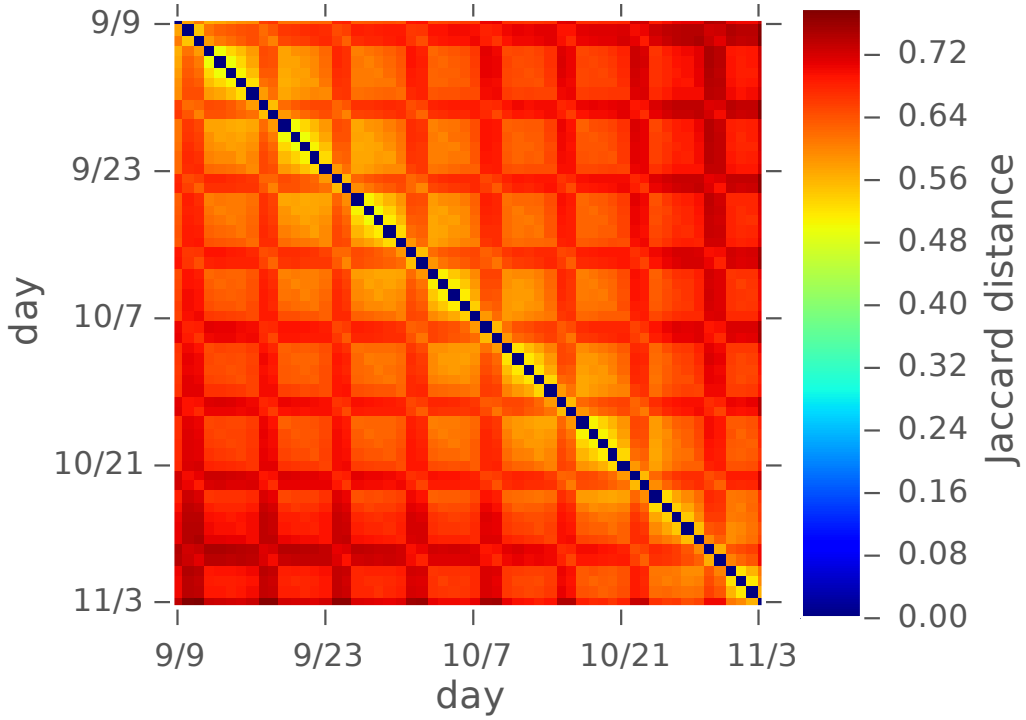
  **foreach** *row* $u$ **do**

$$L'(u) = \begin{cases} 1, & \text{If } P(L(u) = 1|\Theta) \geq t_{R(u)} \\ 0, & \text{otherwise} \end{cases}$$

**until** $L' \approx L''$;

**return** $\Theta$

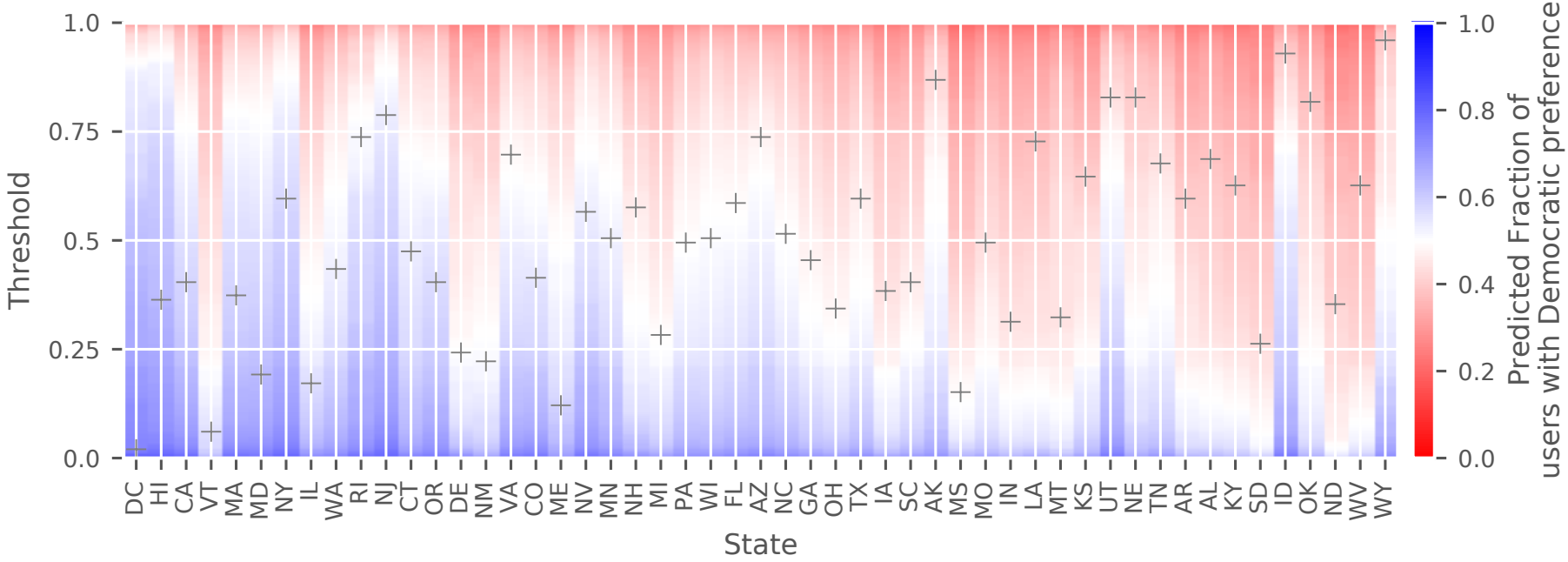
# Temporal Heterogeneity




Jaccard Distance of Daily User Sets

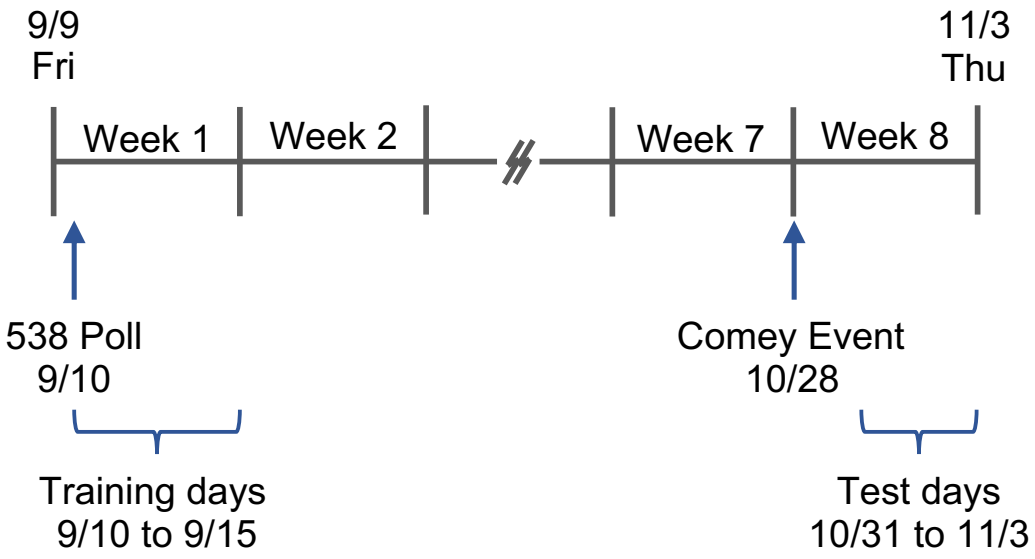
- Models trained on weekday users tend to predict that weekend users are more Democratic than in reality
- Need to train weekdays and weekends separately

# Spatial Heterogeneity



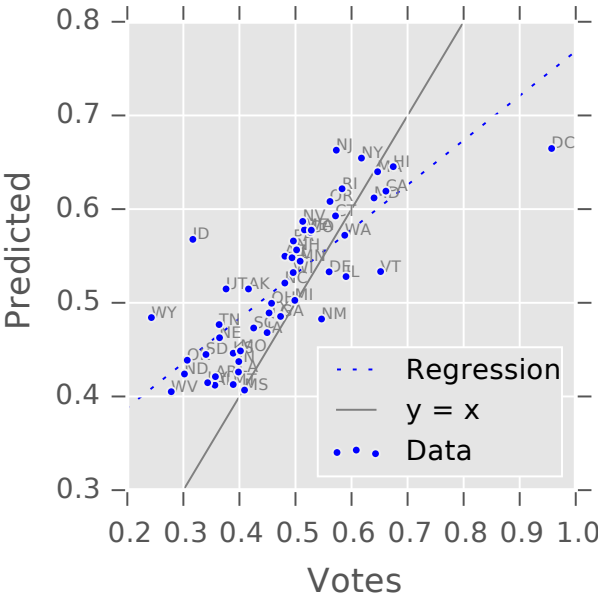

- Browsing behavior considered "x% Democratic" could be considered much more (or less) Democratic in another state

# Putting it All Together - Results

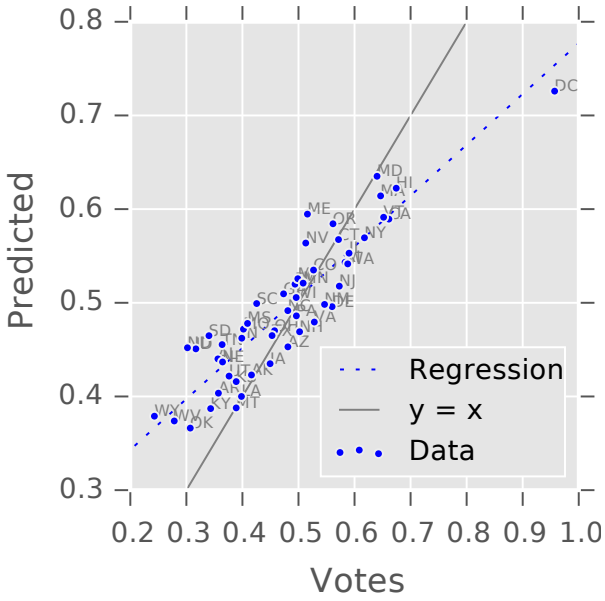



- We trained using 51 per-state polls as of September 10, 2016
- We made predictions ~8 weeks later, on daily and per-state basis

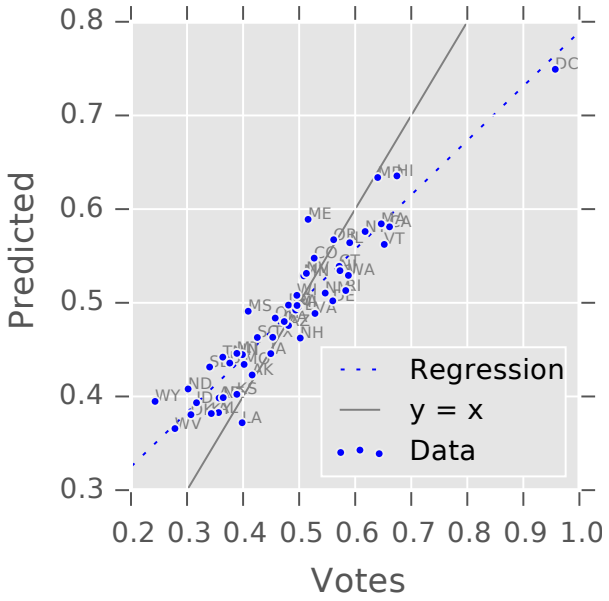
# Impact of Model Refinements




Baseline
ρ = 0.80



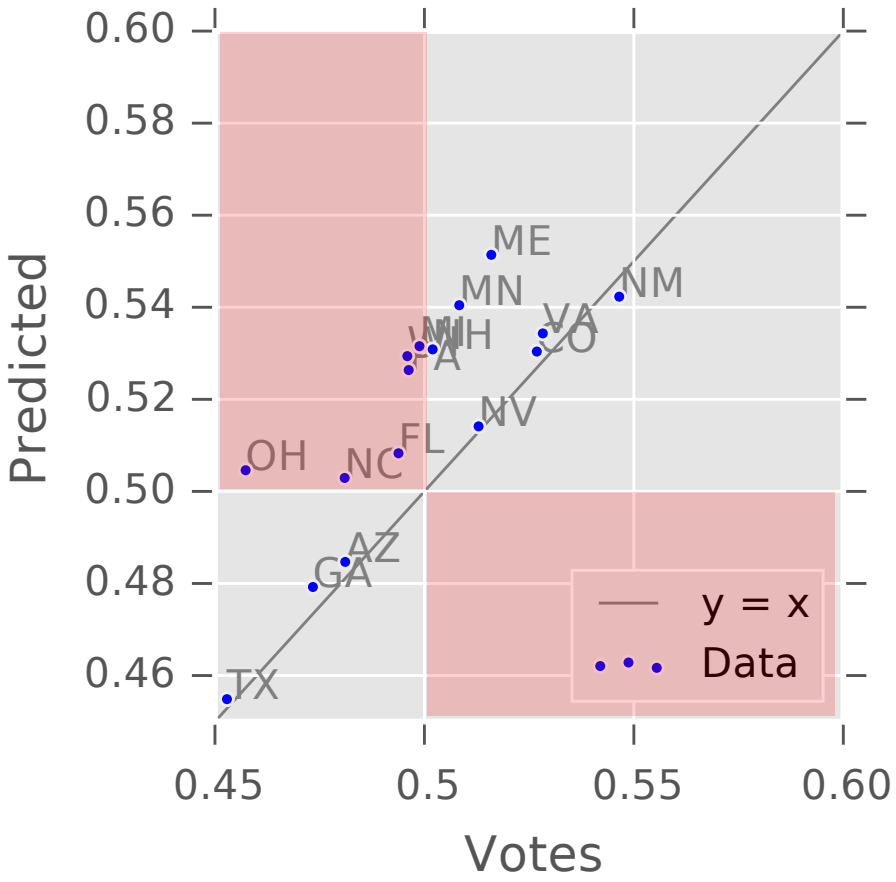

Per State Model
ρ = 0.91




Plus EM
ρ = 0.94

# Battleground States

- Prediction in close states matters most
- US presidential elections hinge on "battleground" states.



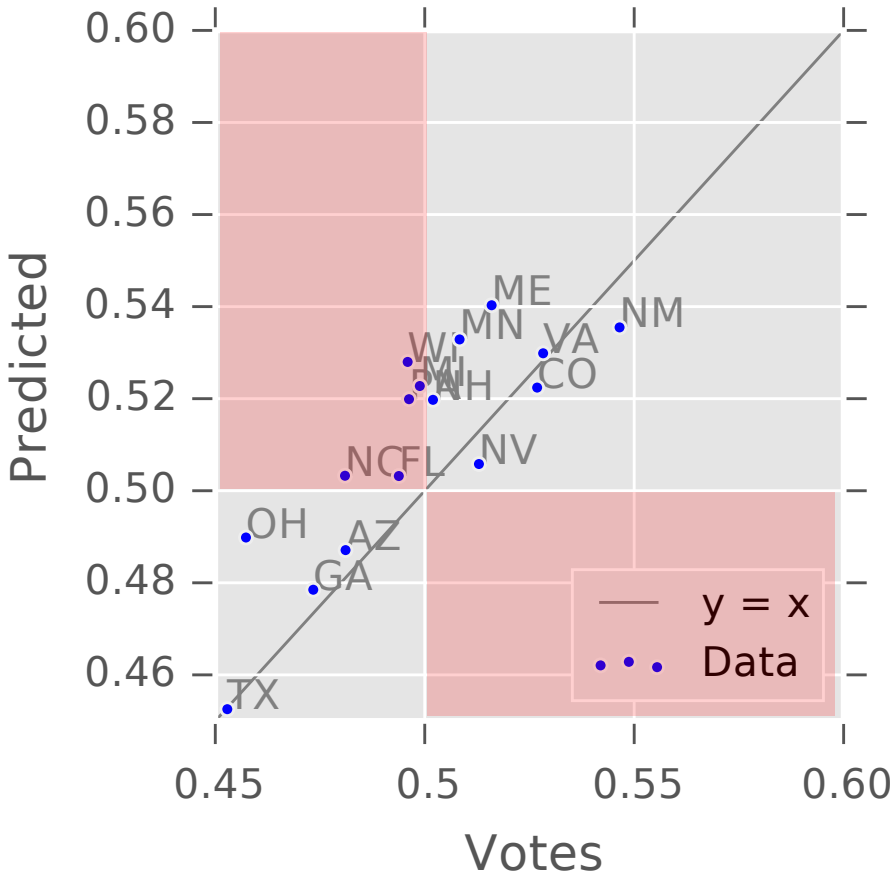

What if we had done nothing …

ie, used the polls from Sept 10 to predict the vote in battleground states

7 states mispredicted (MI, PA, WI, FL, NC, OH, and IA)

# Polls ”got it wrong”

- What if we had used the last polls pre-election to predict the vote?

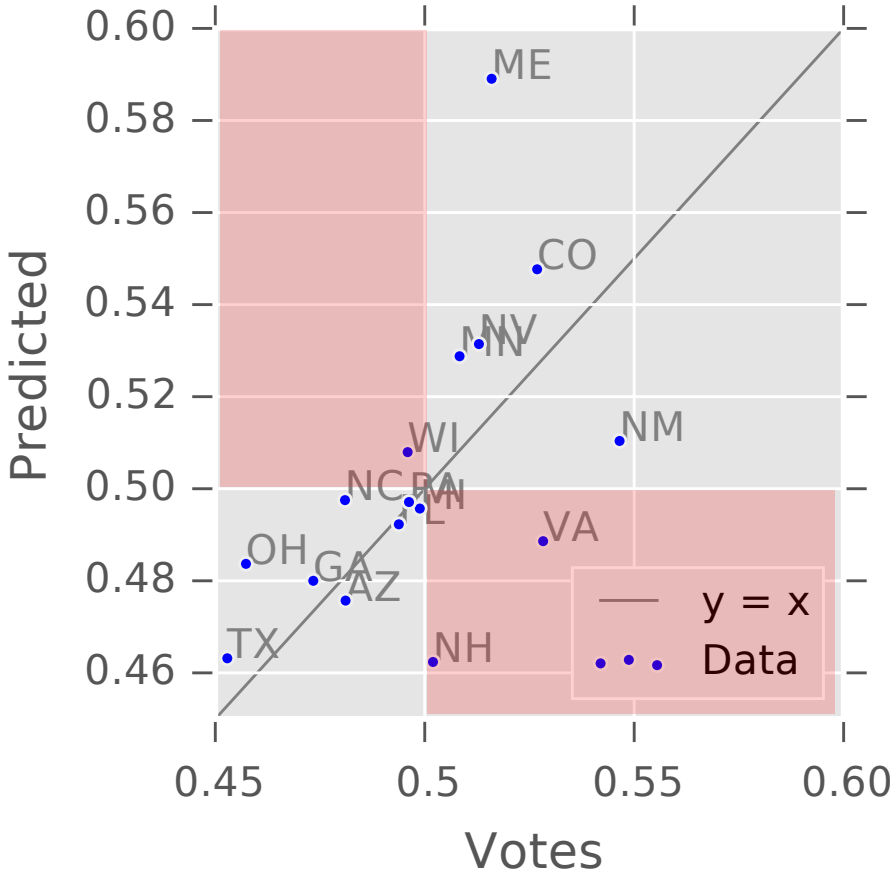



Polls on Nov 7 versus final vote

5 states mispredicted (MI, PA, WI, FL, NC)

# Model Predictions

- More absolute error than polls
- But less biased!



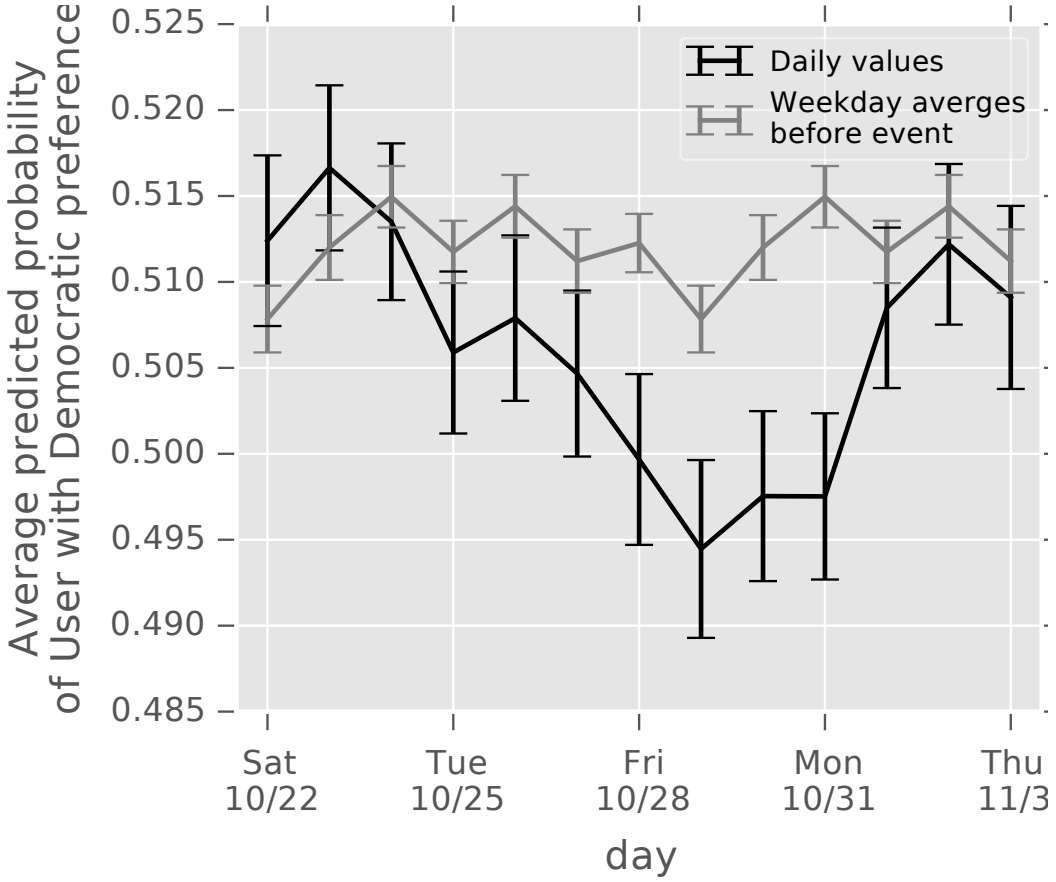

Model predictions versus final vote

3 states mispredicted (VA, NH, WI)

# Assessing an Event

- On Oct 28, FBI director James Comey announced re-opening of Clinton investigation
- Immense controversy, many said that this was decisive factor in election outcome
- Our methods allow fine-grained examination of public opinion around this event

# Comey Letter Effect




Week of Comey Letter vs 6 previous weeks average

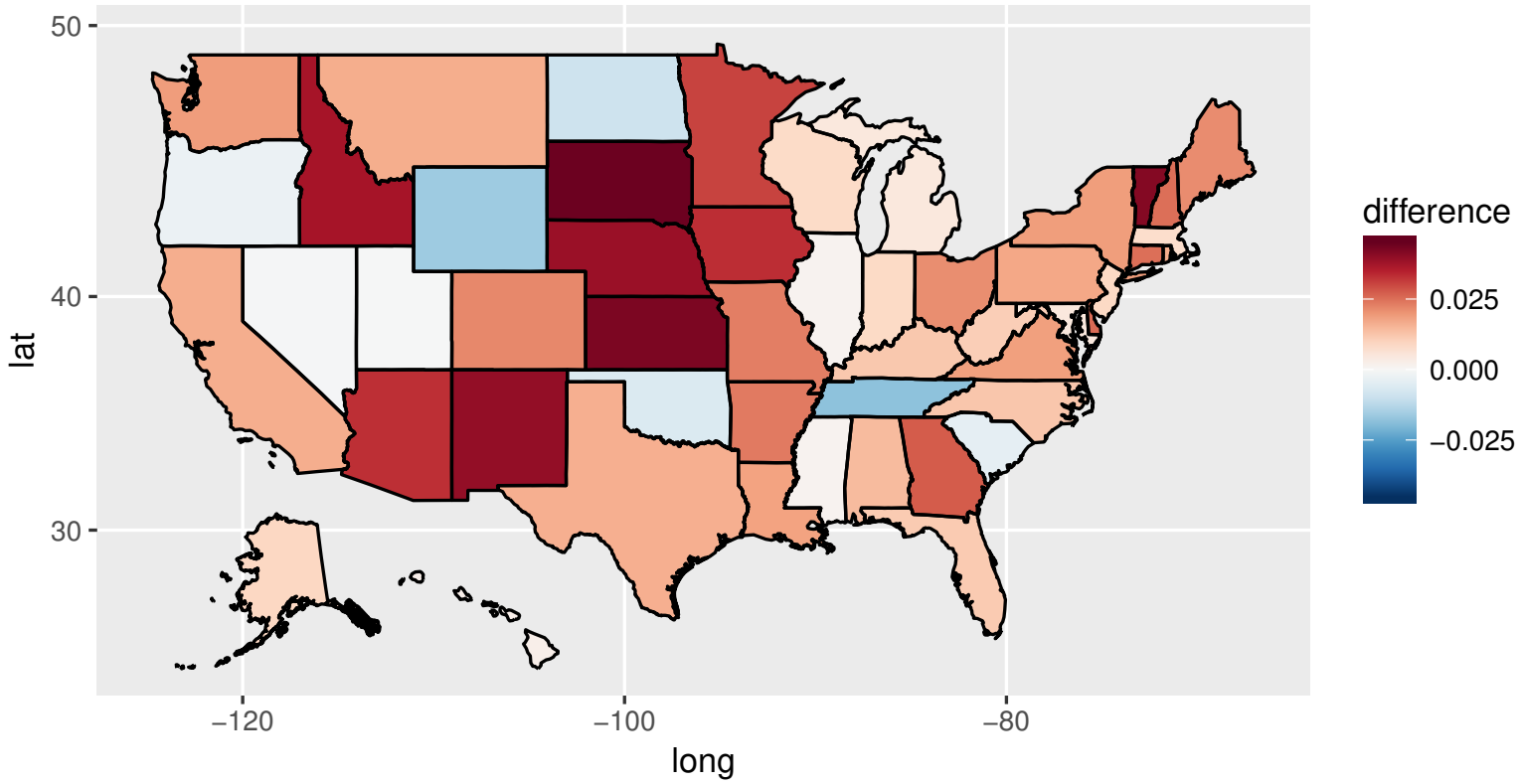
Statistically significant drop in support for Clinton

Starting around October 25

"The evidence for a meaningful effect on the election from the FBI letter is mixed at best. Based on Figure 6, it appears that Clinton's support started to drop on October 24th or 25th. October 28th falls at roughly the midpoint (not the start) of the slide in Clinton's support."

*An Evaluation of the 2016 Election Polls in the United States. Public Opinion Quarterly (February 3 2018)*

# Comey Letter Effect, Spatially




- Strongest effect in west and Midwest
- Moves away from Clinton in states with very close margins (NH, MN, AZ, PA)

# Is this approach sustainable?

- We believe yes
- Panels are commonly deployed by industry
  - Compensation on the order of < $25 /year /person
  - Web tracking software unintrusive
- Data could be collected in a secure fashion
  - All data could be handled only in encrypted form
  - Secure Multi-Party Computation
- Statistical correction methods for sampling issues are well understood

# Open Questions

- When (ie, why) does the EM algorithm work?
- What range of personal opinion questions can be explored using this approach?
  - Issue preference, attitudes, identification
- Can new political science questions be answered?
  - What is the effect on opinion when candidate X gives speech Y on date Z in city W?
  - How long does it last?

# Conclusions

- First method to use history of visits to web sites to assess preference for political candidates
- We clarify and address the challenges
  - New method for training a classifier using only label proportions
  - Feature selection, spatial and temporal heterogeneity
- We illustrate the power of this data + these methods for answering questions of interest to political scientists (and society)