WHO WILL VOTE? ASK GOOGLE*

Seth Stephens-Davidowitz
sstephen@fas.harvard.edu

July 13, 2013

Abstract

Measuring intention to vote is complicated by social desirability bias: the majority of non-voters falsely tell surveys that they will vote. This paper argues that Google searches prior to an election can be used to proxy voting intention in different parts of the United States. I apply the data to forecasting elections. For the 2008 and 2010 elections, October search rates for "vote/voting," compared to four years earlier, explained 20-40 percent of state-level change in turnout rates. And out-of-sample predictions made prior to the 2012 election were strong. The data might prove useful in predicting candidate performance in midterm elections. If turnout is predicted to be high, the Democratic candidate can be expected to do better than the polls suggest. For presidential elections, the data can be useful in estimating the composition of the electorate, by comparing media market search rates to media market demographics. In the 2008 election, the Google data would have correctly predicted substantially increased African-American turnout. The out-of-sample 2012 demographics predictions using Google data were largely correct. It correctly forecast elevated Mormon turnout. It correctly forecast, contrary to some pollsters' predictions, that African-American turnout would remain elevated. Most important, perhaps, for political scientists, the Google data can be used as a dependent variable to find what determines interest in voting.

*I thank Alberto Alesina, Raj Chetty, David Cutler, Ed Glaeser, and Lawrence Katz for helpful conversations. I also received helpful comments at the Southern Political Science Association annual meeting.

I Introduction

People misreport likelihood of voting to polls. Rogers and Aida (2012) find that 67 % of individuals who will not vote tell pollsters they are almost certain to vote. Recent research casts doubt on the reliability of pollsters' tools to screen "likely voters." Ansolabehere and Hersh (2011) find that even the first screen used by polls, registration status, is subject to large deception. More than 60 % of non-registered voters falsely claim that they are registered. Rogers and Aida (2012) find that reported voting intention has little predictive power, controlling for previous voting behavior. Vote misreporting is correlated with other variables in important ways that may bias analysis (Ansolabehere and Hersh, 2011; Rogers and Aida, 2012; Vavreck, 2007).

Google data can measure socially sensitive behaviors (Stephens-Davidowitz, 2012). And the theory that Google can capture changes in turnout intention over time is compelling: the marginal voter, the individual who only votes in certain elections, is likely to need information prior to voting. Thus, he or she might search for reminders, Googling "how to vote" or "where to vote." The fact that habitual voters are unlikely to need to make these searches does not limit this exercise; these voters' intentions are easy to forecast.

In particular, using the 2008 and 2010 elections, I show that change in search rates for "vote" or "voting" in the October prior to an election, compared to October four years earlier, explains 20 to 40 percent of variation in changes in turnout rates compared to four years earlier. The predictive power is little affected by controlling for changes in registration rates, early voting rates, or a state's having a Senate race, three other sources of information available prior to an election that might be used to measure intention to vote. Prior to the 2012 election, I made out-of-sample predictions for turnout in the 2012 elections. The predictions were fairly strong and always strongly significant. There were, though, additional lessons that might improve predictions in the future, such as the importance of controlling for trends in voting searches and voting rates.

Can more be done with this data than projecting state-wide turnout? Might this data help political forecasters predict election results? And might this help political scientists better understand the causes of voting decisions? I examine two political forecasting applications and mention areas for possible future research.

Forecasters might use Google search data to help choose demographics weightings. In 2008, African-Americans turned out at elevated levels to support the first major-party general election black candidate. In 2012, there was great uncertainty as to whether the black share of the electorate would remain at its historic 2008 levels. Since African-Americans overwhelmingly support the Democratic candidate, a one percentage point increase in the black share of the electorate leads to nearly a one percentage point increase in the Democratic candidate's two-party vote share. The Google data would have correctly predicted a large increase in the black share of the electorate in 2008. The Google data showed large increases in the rates of voting-related searches in black communities prior to the 2008 compared to the 2004 election. And the Google data would have correctly predicted that the black share of the 2012 electorate would be similar to its historic 2008 levels. Many polls that underestimated Obama's support falsely assumed that the black share of the electorate would decline in 2012. Prior to the 2012 election, I predicted the composition of the 2008 electorate based on Google search data. I found that only one major demographic was a robust predictor of changing vote search rates, from October 2008 to October 2012: the share of the population that is Mormon. I thus predicted that the Mormon share of the electorate would increase. Other population groups would remain similar to their 2008 shares, with some adjustments for population increases. (For example, the Hispanic share of the electorate would increase, even though Hispanic turnout rates would not change.) The predictions were largely correct. This methodology would seem particularly promising for measuring a large change in turnout rates for a group whose population differs substantially in different parts of the country. Black turnout rates in 2008 and Mormon turnout rates in 2012 fit these criteria.

Forecasters might adjust their likely voter screen in a state based on Google search activity in that state. If a state is likely to have high turnout, pollsters can let more voters through the screen. Since most marginal voters are Democrats, an increase in the pool of likely voters will likely lead to a higher predicted vote total for the Democratic candidate (Gomez et al., 2008). An increase in Google-related voting activity did, on average, predict better Democratic performance than the polls predicted in the 2010 Midterm elections. However, it did not for the 2008 or 2012 presidential elections. This strategy, thus, appears to only work in Midterm elections.

The data source might prove useful in explaining past behavior. Two possibilities seem particularly promising. First, political scientists might study the determinants of actual voting, controlling for early intent to vote. Some variables that might be studied include negative advertising and voting laws. Second, the data source might be used as a high-frequency measure of intent to vote and can help measure the immediate effects of interventions. For example, a major aim of conventions is energizing the base to vote. Comparing vote-intention searches in the aftermath of a convention might proxy its success. By this metric, Romney's 2012 convention was the least successful. (Few individuals were looking up how to vote after the convention.) This fits some pundit commentary and polls-based statistical analysis.

This paper builds on a nascent literature, started by Varian and Choi (2010) and extended by Ginsberg et al. (2009) and Askitas and Zimmermann (2009) in using Google data to forecast the future. It follows Stephens-Davidowitz (2012) in using Google data to proxy a socially sensitive attitude.

II Google Search Proxy

To predict change in area-level turnout, I examine changes in an area's percentage of searches that include "vote" or "voting" in the October prior to an election.

In particular, I am interested in Δln(Google Vote Search) =

$$\ln\left(\frac{\text{Oct. Searches w/ “Vote/Voting”}_{\text{i,t}}}{\text{Oct. Searches}_{\text{i,t}}}\right) - \ln\left(\frac{\text{Oct. Searches w/ “Vote/Voting”}_{\text{i,t}-4}}{\text{Oct. Searches}_{\text{i,t}-4}}\right)$$

Following Stephens-Davidowitz (2012), I choose the most salient words to constrain data-mining. I choose October data so that predictions can be meaningfully delivered prior to an election. One might suspect that likely voters would wait until the last minute to find information on voting. Indeed, search volume for "vote" and "voting" do increase significantly the day of and the day prior to the election. However, perhaps surprisingly, I did not find that including this November data improved the model. In results not shown, I find that the model is a bit worse including a longer range of data (August and September). It is a bit worse shortening the prediction by just including one week in October. While it is impossible to test this with data available, the evidence is consistent with likely voters searching for voting information many times in October and the first days of November.

The measure selected is an "omnibus" measure which captures many different ways that Google searching may predict turnout.

Figure I shows the words most often used in combination with the word "vote" or "voting" in October 2012. Notably, 25 percent of searches also include the word "early." An obvious concern is that changes in searches that include "vote" or "voting," from October 2008 to October 2012, picks up consistent voters switching to early voting. One possibility is taking out searches that include the word "early." However, I instead use all these searches because of the predictive power these searches have shown, presented shortly, the importance of capturing new voters who happen to choose early voting and would need to Google information, and the fact that I found little evidence for substantially different results excluding the term.

In addition, 20 percent of searches include the word “register” or “registration.” I include these in the measure for many of the same reasons I include searches for early voting.

The data are available at the state, media market, and city level from Google Trends. However, if absolute search volume is not high enough, Google Trends does not report data. Stephens-Davidowitz (2012) develops a methodology for obtaining this data. In principle, any data can be obtained with this – or a fairly similar – approach. However, the lower the absolute search volume, the more downloads necessary to obtain the data. I have weighed importance of the data and difficulty of obtaining it in choosing the geographic level of analysis.

There is one important issue with the data. Four states – Virginia, Vermont, California, and Delaware – and three media markets – San Francisco, Washington (DC), and Burlington, Vermont – consistently show up as outliers when measuring changes in search volume through time. While I am not entirely sure the reason for this, it seems to be related to incorrect geographic proxies. Google improved its geographic marks on January 2011, and notably these areas – and no others – saw sharp changes in search rates for most words from December 2010 to January 2011.

I do not include these four states and three media markets in the analysis and recommend omitting them from any analysis measuring changes through time prior to January 2011.

III Using Google to Predict Turnout

The previous section develops a proxy for changes in interest in voting information on Google between election cycles. I now compare this to changes in turnout rates. I measure turnout rate as total votes as a percentage of the voting eligible population. I download this data from Dr. Michael McDonald at `http://elections.gmu.edu`.

III.A. 2008 and 2010 Elections

Figure II shows the results for the 2008 presidential election and 2010 midterm election. Panel (a) compares changes in Google voting interest to changes in turnout rates, from 2004-2008. Panel (b) compares changes in Google voting interest to changes in turnout rates, from 2006-2010. The relationships are both highly statistically significant. If Google search activity for voting was higher in October 2008 (2010), than would be expected from October 2004 (2006), turnout was higher in 2008 (2010 than in 2004 (2006).

The relationship is stronger for the midterm elections than the presidential elections.

Table I examines the relationship between Google searches and turnout for presidential elections more systematically. Since the search proxy includes searches such as "register to vote" or "early voting," one might suspect that Google data is picking up other factors, and alternative publicly available data may be used instead to predict turnout. However, I show that the relationship between Google voting-related searches and turnout is little affected by including changes in registration rates and early voting rates.

In results not shown, I find that Google searches go beyond other factors that might predict midterm turnout. These include changes in registration rates and states that had a Senate race in one year considered but not the other.

III.B. 2012 Out-of-Sample Election

Section III.A. found that October search volume for "vote" or "voting," in October, compared to four years earlier, was a strong predictor of state-level turnout in both the 2008 presidential and 2010 midterm elections.

Since this analysis was all done in the run-up to the 2012 election, the 2012 election presented an excellent opportunity to give an out-of-sample test to this methodology.

Prior to the 2012 election, I calculated the change in Google search rates for "vote" or "voting," from October 2008 to October 2012. I then predicted change in state-level turnout rates based on this data. I made all predictions public on my website the morning of November 6, 2012.

Overall, there was a decrease in country-wide search volumes for "vote" or "voting." I assumed that this was not a meaningful change, as I have often found changes at the national level that did not correlate with national-level changes. I instead assumed that there would be no overall change in turnout but that states in which decreased search rates were more pronounced would see decreased turnout rates. States in which decreased search rates were least pronounced would see increased turnout rates. In fact, the decreased search volume at the country-level did prove meaningful. Overall turnout rates were lower in 2012 than in 2008.

Column (1) of Table II shows the predicted changes in turnout based on $\Delta\ln(\text{Google Vote Search})_{2012,2008}$ I use the coefficient from the midterm elections, and, as mentioned, assumed that overall turnout would be unchanged.

Column (3) shows the actual change in turnout rates, by state. Figure III compares the two. The relationship is statistically significant, though somewhat weak, in unweighted regressions. (Panel (a)). Panel (b) of Figure III weights regressions by 2000 population. The relationship is much stronger. Similar to the 2008 election, smaller states had significantly larger errors in 2012.

Table III further explores the relationship between a state's predicted change in turnout, from 2008 to 2012, and actual change in turnout, from 2008 to 2012. One interesting finding is that errors in predicting turnout, using Google, were highly serially correlated. States in which Google searches would have under-estimated turnout in both 2008 and 2010 also under-estimated turnout in 2012. This suggests that there are trends in voting-related searches that are not meaningful and should be controlled for in future predictions. This has the potential to greatly improve the methodology.

I also add to the regressions dummy variables that take the value 1 for any state that had at least one fatality caused by Hurricane Sandy (Connecticut, Maryland, New Jersey, New York, North Carolina, and West Virginia). This variable is always negative but does not meaningfully affect the coefficient on the Google-predicted turnout.

As I only have 46 observations, I am limited in how many demographics variables I add. But adding percent black, percent Hispanic, and percent College Grad does not meaningfully change the coefficient on the Google-predicted turnout.

In sum, the out-of-sample predictions were pretty good. The predicted changes in turnout, from 2008 to 2012, correlated with actual changes in turnout, from 2008 to 2012. However, the analysis suggests two changes in future predictions. First, changes in national patterns may have more meaning than I realized. And data should be de-trended. The Google analysis led to consistent under or overestimates in certain states. This would not happen with a proper de-trending.

IV Using Google Data to Estimate Election Demographics

Section III shows that a state's Google searches prior to an election can predict an area's turnout.

Can this information be used to improve election projections?

In this section, I discuss how this information might be used to predict the composition of the electorate.

IV.A. 2008 Election

According to exit polls, roughly 88 % of African-American voters supported Democrat John Kerry in 2004. Roughly 95 % of African-American voters supported Obama in 2004.

Roughly 77 % of white evangelical voters supported George W. Bush in 2004. Roughly 74 % of white evangelical voters supported John McCain in 2008.

In addition, Hispanic, Jewish, and young voters lean Democratic. Non-Hispanic white, Mormon, and older voters lean Republican.

Since certain groups are so likely to support a particular party, composition of the electorate is crucial in determining the final vote totals. A one percentage point increase in black turnout will be expected to lead to a nearly one percentage point increase in the Democratic candidate's vote share.

By comparing areas' voting interest, as proxied by Google, to areas' demographics, we might predict whether any demographic can be expected to turnout in unusually high numbers.

This exercise can best be done using a smaller level of aggregation of the data than the state and thus allowing for bigger differences in demographics. I thus do this analysis using data from media markets in the United States.

Consider this exercise for the 2008 presidential election.

Column (1) of Table IV shows the correlation between the percent of the population that is black and change in October voting-related search rates from 2004 to 2008.

This relationship is highly statistically significant and robust. Column (2) adds a number of other demographics variables. The other variable that is consistently statistically significant is the Hispanic population. The Google data consistently predicted increases about 1/3 to 1/2 as large among the Hispanic population as among the African-American population. Column (3) reproduces the results of Column (1) but adds Census division fixed effects. Column (4) reproduces the results of Column (2) with Census division fixed effects.

Choosing different variables, in regressions not reported, does lead to some instability in some of the coefficients on the age, college, and percent Kerry variables. However, the percent black and percent Hispanic proved to be remarkably robust.

IV.B. 2012 Out-of-Sample Election

Prior to the election, I predicted the composition of the electorate using Google search data.

Table V shows results analyzing changes in Google search activity, from October 2008 to October 2012. There is a remarkably strong, robust, and stable relationship between changes in Google voting-related search rates and population size. About 50 percent of the variation can be explained by the log of the population rate. And I was unsuccessful in finding any correlate of population size that explained away this result. (There was no similar relationship between population size and changes in Google search rates, from 2004 to 2008, and the results in IV) are little affected by adding a control for the log of population.

I was unable to figure out the reason for this relationship. It does not seem to have to do with rural vs. urban areas, which might be affected by internet penetration rates. Nor does it seem to be a universal fact of Google data that population rates correlate with changed search rates over this time period. I assumed that population size would not meaningfully impact the voting composition of the electorate but I did control for it in all my regressions.

Column (1) of Table V, while including the population size control, adds the variable percent black. The relationship is not statistically significant. If we interpret these as actually capturing the likelihood of blacks' voting, as it seemed to in 2008, we can easily reject a drop in the black voting population as large as the gains in black voting population in 2012. Column (2) of Table V adds the same controls used in Column (2) of Table IV. None of the coefficients are statistically significant. at the 5 % level. Column (4) and (5) add division fixed effects to the regressions of Columns (1) and (2), respectively. Again, none of the demographics coefficients are statistically significant at the 10 % level.

Columns (3) and (6) add to the regressions of Columns (2) and (5) religion variables from the 2000 Congregations and Membership Study. Some hypothesized that conservative evangelicals would be less motivated to turn out due to Mitt Romney's Mormon faith. There is no statistically significant relationship between evangelical adherents per capita and changes in Google searches. Intriguingly, of all the religion variables, the only statistically significant coefficient is on Percent Other Religion. This consists mostly of Mormons. There is strong and robust evidence that Mormons will turn out at historic levels to support Romney. The coefficient on Percent Other Religion in Table V is similar in size and robustness to the coefficient on Percent Black in Table IV.

Table VI combines the results of Table V with changing demographics in the country to make out-of-sample predictions for the composition of the 2012 electorate. All predictions were made public, as shown, on my website the morning of the election.

The only robust predictor of changing Google search volumes was Other Religion. Based on the magnitude, I predicted that 'other' religion would rise from 6 to 7 percent of the voting population. I predicted that all other voting rates would stay the same. However, since the Hispanic share of the population grew substantially over the previous four years, I assumed that Hispanics would make up 10 percent of the electorate, compared to the 9 percent Hispanic share of the 2008 electorate.

Contrary to many, I did not foresee decreased share of black voters. As shown in Table VI, parts of the country with the highest black populations were Googling for voting information in 2012 at the elevated rates of 2008. And I predicted that white evangelicals would make up 26 percent of the electorate, just as they did in the 2008 electorate. I did not foresee major changes in the age composition of the electorate.

The predictions performed well. Hispanics, due to population growth, indeed made up 10 percent of the electorate. Blacks again made up 13 percent of the electorate. The white evangelical share of the electorate did not decrease despite Romney's presence on the ticket.

Percent declaring ‘other religion’ was indeed elevated, and this was almost certainly due to increased Mormon turnout. Utah was, as predicted in Table II, one of the states that ranked highest in the difference between 2012 and 2008 turnout rates.

The only number that was off was the share of the electorate age 18-29. I guessed that it would remain 18 percent of the electorate, wheres young voters rose to 19 percent of the electorate. My methodology may struggle to predict changes in voting rates by age, since media markets do not differ that much in age demographics.

V Using Google Data to Predict Democratic Overperformance

The simplest use of the Google data is just adjusting predictions based on Google-predicted state-level turnout. Pollsters might weaken or strengthen their likely voter cutoff number based on Google search activity.

High turnout is usually a good sign for Democrats (Gomez et al., 2008). Table VII examine the relationship between changes in turnout, compared to four years earlier, and Democratic performance as compared to polls in the 2008 and 2012 presidential elections and 2010 Senate elections. As my measure of predicted performance based on the polls, I use the final predictions of Nate Silver.

V.A. 2008 and 2012 Presidential Elections

Columns (1) and (3) of Table VII show that the change in turnout rates, compared to four years earlier, did correlate with Obama’s over performance in both the 2008 and 2012 elections. However, of the two presidential elections, the relationship is only statistically significant with the 2008 election. Interestingly, there is positive serial correlation in Obama’s over performing the polls. If Obama did better than Nate Silver’s final predictions in a state in 2008, he was also expected to do better in 2012.

Columns (2) and (4) of Table VII show that changing Google search rates for voting, would not have successfully predicted Obama performance, relative to Nate Silver's predictions, in either the 2008 or 2012 presidential elections.

However, Columns (3) and (4) show that changes in October Google voting-related searches, while they did predict changes in turnout, did not predict Obama over performance.

V.B. 2010 Midterm Election

Column (5) of Table VII shows that change in turnout was a strong predictor of Democratic performance, compared to the polls, in 2010 Senate elections. If turnout was higher in 2010, relative to 2006, the Democratic Senate candidate could be expected to perform better than Nate Silver's final projections. Interestingly, once again Obama 2008 over-performance was positively correlated with Democratic over-performance in a different year. If Obama did better than Nate Silver predicted in 2008, Democratic Senate candidates would be expected to do better than Nate Silver predicted in 2010.

Column (6) shows that, for the 2010 Midterm election, elevated Google search rates were highly statistically significant predictors of Democratic performance compared to Nate Silver's final predictions.

Note, too, that the changing predictions are quantitatively significant. Figure II Panel (b), shows that $\Delta ln(TurnoutRate)$ varies between about 0.25 and -0.25. Column (5) of Table VII shows that changes in turnout of this magnitude would mean an increase in Democratic performance of about 2.5 percentage points or decrease in Democratic performance of about 2.5 percentage points, relative to Nate Silver's predictions. Column (6) shows that Google could have predicted much of this over or under performance.

Thus, Google search rates would not have predicted state-level over or under-performance in the 2008 or 2012 presidential elections. But it would have predicted state-level over or under-performance in the 2010 midterm Senate elections. The huge variance in turnout in Midterm elections may explain the difference. Google data for voting information in October should certainly be considered as a supplementary source in the 2014 midterm elections.

VI Using Google Searches as Dependent Variable

Earlier, I used Google searches for “vote” or “voting” as an independent variable to predict turnout levels. More interesting, perhaps, for political scientists is not predicting turnout a few weeks before an election. More interesting, perhaps, is finding the drivers of turnout. Traditionally, political scientists have used self-reported voting likelihood as the dependent variable. But Rogers and Aida (2012) questions the validity of this strategy.

If Google searches are a better predictor of voting intention, they can potentially be used as a dependent variable to understand the predictors of turnout. Figure IV shows a measure of convention effectiveness by how much interest in voting-related searches were made afterwards. The idea is that a successful convention will mobilize the base, and this will show up in interest in voting. In particular, I measure the percent of Google searches that included the word “vote” or “voting” on the Friday after every convention, starting in 2004. Note that this measure, as always, is normalized by total Google searches.

I add an additional normalize, dividing all variables by a common factor so that the top-scoring convention scores 100. Some interesting patterns emerge. The most successful convention, by this metric, over this time period, was the 2004 Republican convention. The least successful convention, by far, was the 2012 Republican convention. In fact, the 2012 Republican convention was followed by less than half the voting-related searches as any other convention. Media accounts generally agree that the 2012 Republican convention was a poor one: the conversation mostly focused on Clint Eastwood’s inappropriate speech, rather than the speeches of either Mitt Romney or Paul Ryan.

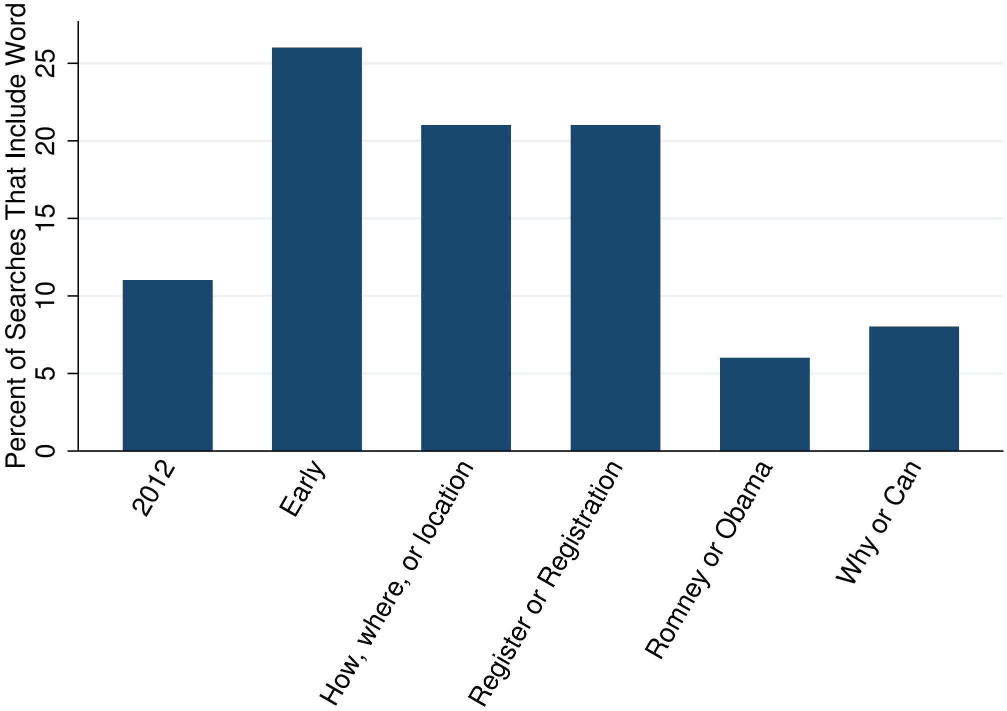
Figure IV is one application of using Google data to measure the effectiveness of various interventions in increasing turnout. More research, along these lines, might be done.

VII Conclusion

I show that search volume for "vote" or "voting," in October, strongly correlates with final turnout, in United States elections. I show two ways that this can help political forecasters.

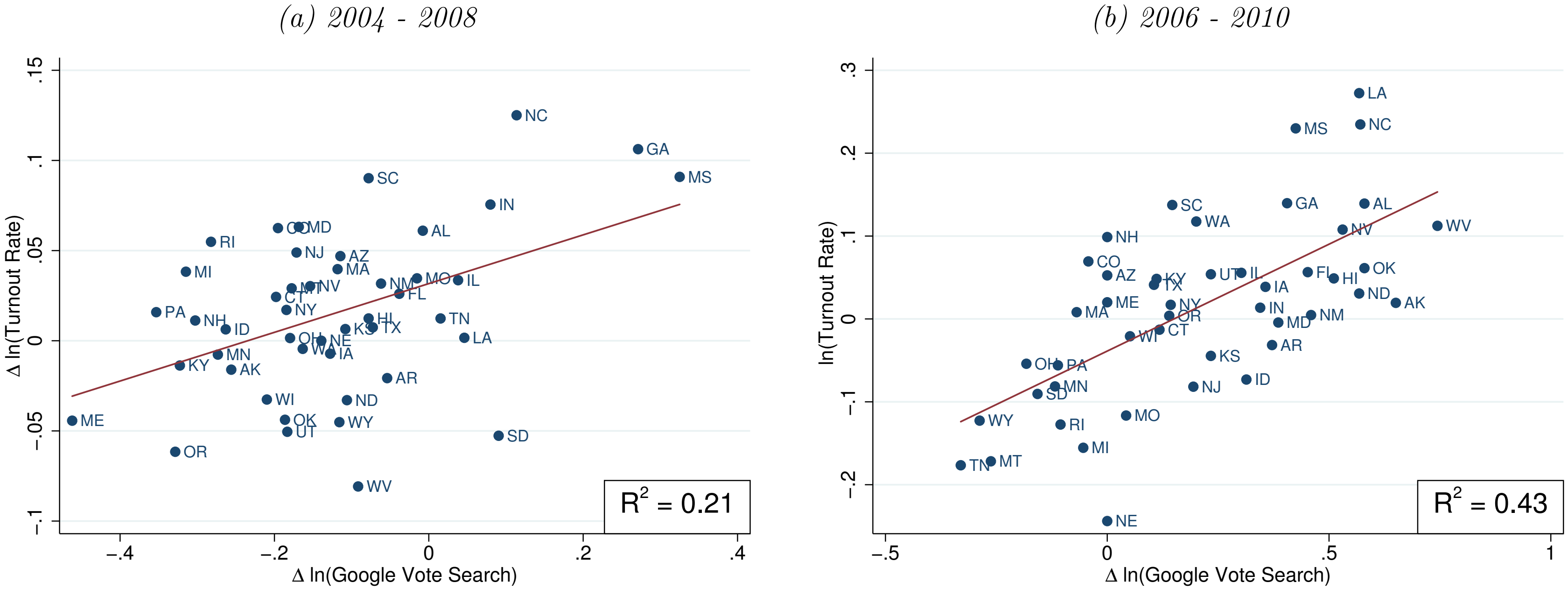
The data might also prove useful in understanding the determinants of voting decisions. High-frequency search data can be used to study the effectiveness of various interventions in increasing voting intentions. In addition, comparing voting intention – as proxies on Google – to actual voting behavior can be used to better studying how last-minute laws or advertising campaigns influence turnout.

Figure I
Searches for “Vote” or “Voting,” October 2012



Notes This shows the percentage of searches that include “vote” or “voting” in October 2012 that also included the word(s) shown. This was done by dividing the rate of searches that include either “vote” or “voting” and the word(s) shown by the rate of searches that include “vote” or “voting.” All Google data are downloaded from Google Trends.

Figure II
Change in Turnout Rates and Change in October Google Voting-Related Search Rate

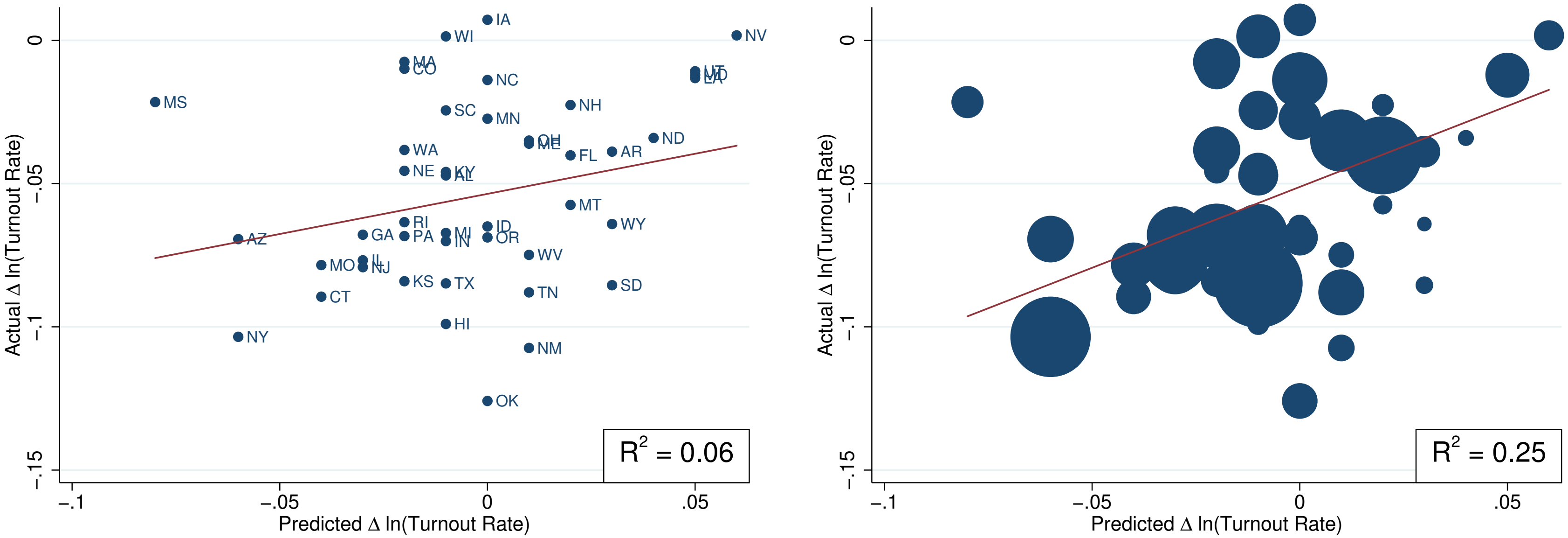


(a) 2004 - 2008

(b) 2006 - 2010

Notes: On the x-axis is the change in the natural log of October search rates for "vote" or "voting." On the y-axis is the change in the natural log of the turnout rate. All Google data are downloaded from Google Trends. Turnout Rate is total presidential votes divided by total eligible voters. These data were downloaded from The United States Elections Project, at `http://elections.gmu.edu/bio.html`.

Figure III
Change in Turnout Rates and Change in October Google Voting-Related Search Rate, 2008-2012

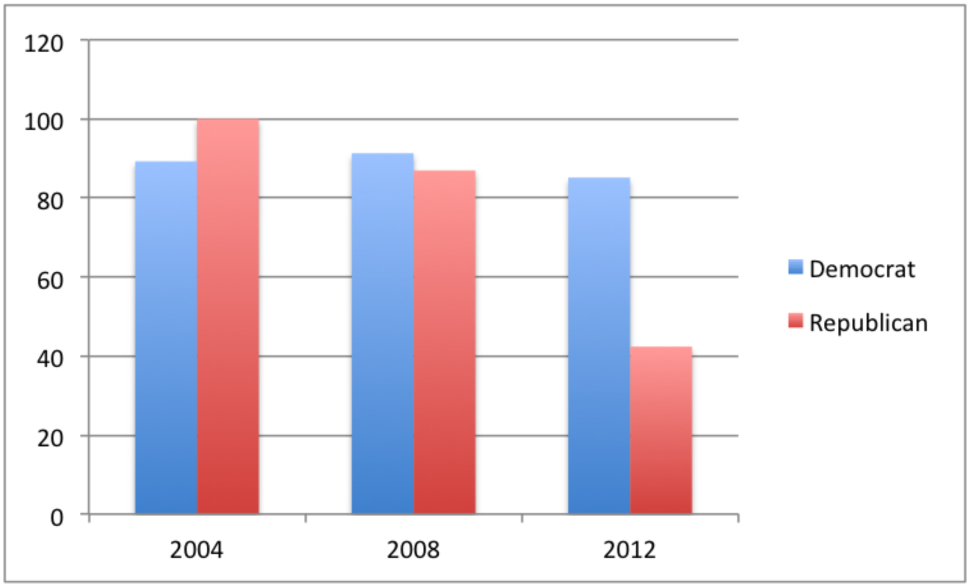


(a) Unweighted

(b) Weighted

Notes: On the x-axis is the change in the natural log of October search rates for “vote” or “voting.” On the y-axis is the change in the natural log of the turnout rate. All Google data are downloaded from Google Trends. Turnout Rate is total presidential votes divided by total eligible voters. These data were downloaded from The United States Elections Project, at `http://elections.gmu.edu/bio.html`. Weighted regressions are weighted by 2000 population, from the Census.

Figure IV
Google Voting-Related Search Rate, Friday After Convention



Notes: This figure shows the percent of Google searches that included the word "vote" or "voting" the Friday after the convention. Results are normalized so that the highest-scoring convention scores 100.

Table I
Turnout, 2004-2008

		Δ ln(Turnout Rate)		
	(1)	(2)	(3)	(4)
Δ ln(Google Vote Search)	0.135***	0.134***	0.151***	0.144***
	(0.038)	(0.037)	(0.039)	(0.031)
Δ ln(Registration Rate)		0.162		0.312***
		(0.109)		(0.090)
Δ ln(Early Vote Rate)		-0.004		-0.004
		(0.015)		(0.011)
Adjusted R-squared	0.19	0.19	0.31	0.41
Observations	46	46	46	46
Weighting	Unweighted	Unweighted	Weighted	Weighted

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Notes: The dependent variable is the change in the natural log of the turnout rate from 2004 to 2008. Turnout Rate is total presidential votes divided by total eligible voters. These data were downloaded from The United States Elections Project, at `http://elections.gmu.edu/bio.html`. $\Delta ln(GoogleVoteSearchRate)$ is the difference between the natural log of Google search rates that include "vote" or "voting," from October 2004 to October 2008. All Google data are downloaded from Google Trends. $\Delta ln(RegistrationRate)$ is change in the change in natural log of registered voters on election day per eligible voters, from 2004 to 2008. $\Delta ln(EarlyVoteRate)$ is change in early votes as a proportion of eligible voting population, from 2004 to 2008. These data were also downloaded from the United States Elections Project, at `http://elections.gmu.edu/bio.html`. California, Vermont, Virginia, Delaware, and Washington D.C. are not included for reasons discussed in the text. Weighted regressions are weighted by 2000 population size, according to the Census.

Table II
State Turnout Predictions for 2012 Election

	Predicted 2012 Turnout **FILLED IN PRIOR TO 2012 ELECTION**		*Actual 2012 Turnout*	
	(1)	(2)	(3)	(4)
State	Δ *ln(Turnout)*	*Turnout Rate*	Δ *ln(Turnout)*	*Turnout Rate*
NV	.06	60.2	0	57.1
UT	.05	58.8	-.01	55.4
MD	.05	70.3	-.01	66.2
LA	.05	64.1	-.01	60.4
ND	.04	65.3	-.03	60.6
AR	.03	53.8	-.04	50.5
SD	.03	66.6	-.09	59.4
WY	.03	65	-.06	58.9
AK	.03	69.7	-.14	58.9
MT	.02	67.9	-.06	62.6
FL	.02	67.5	-.04	63.5
NH	.02	73	-.02	70.1
TN	.01	57.6	-.09	52.2
NM	.01	61.6	-.11	54.7
OH	.01	67.6	-.03	64.6
ME	.01	71.6	-.04	68.1
WV	.01	50.5	-.07	46.3
ID	0	63.9	-.06	59.6
MN	0	78.1	-.03	75.7
OR	0	67.4	-.07	63.2
IA	0	69.2	.01	69.9
OK	0	55.6	-.13	49.2
NC	0	65.8	-.01	64.6
HI	-.01	48.5	-.1	44.2
SC	-.01	57.4	-.02	56.6
KY	-.01	57.4	-.05	55.3
IN	-.01	58.6	-.07	55.1
AL	-.01	60.4	-.05	58
WI	-.01	71.7	0	72.5
TX	-.01	53.8	-.08	49.7

	Predicted 2012 Turnout **FILLED IN PRIOR TO 2012 ELECTION**		*Actual 2012 Turnout*	
	(1)	(2)	(3)	(4)
State	Δ *ln(Turnout)*	*Turnout Rate*	Δ *ln(Turnout)*	*Turnout Rate*
MI	-.01	68.3	-.07	64.7
PA	-.02	62.2	-.07	59.4
MA	-.02	65.6	-.01	66.3
WA	-.02	65.1	-.04	64.1
KS	-.02	61.1	-.08	57
NE	-.02	61.5	-.05	60.1
CO	-.02	69.7	-.01	70.3
RI	-.02	60.7	-.06	58
GA	-.03	60.8	-.07	58.4
NJ	-.03	65	-.08	61.9
IL	-.03	61.6	-.08	58.9
CT	-.04	63.9	-.09	60.9
MO	-.04	64.9	-.08	62.5
AZ	-.06	53.1	-.07	52.9
NY	-.06	55.6	-.1	53.2
MS	-.08	55.9	-.02	59.7

Notes: This Table compares state-level turnout predictions, based on Google search data, made prior to the 2012 election, to actual state-level turnout rates. Column (1) shows predicted change in the natural log of the turnout rate, from 2008 to 2012. Column (2) shows the predicted 2012 turnout rate. Column (3) shows actual change in the natural log of the turnout rate, from 2008 to 2012. Column (4) shows actual 2012 turnout rate. Turnout Rate is total presidential votes divided by total eligible voters. These data were downloaded from The United States Elections Project, at `http://elections.gmu.edu/bio.html`. California, Vermont, Virginia, Delaware, and Washington D.C. are not included for reasons discussed in the text.

Table III
Actual Turnout and Predicted Turnout, 2012

			Δ ln(Turnout Rate)			
	(1)	(2)	(3)	(4)	(5)	(6)
Δ ln(Google Vote Search)	0.293*	0.482***	0.470***	0.571***	0.688***	0.694***
	(0.173)	(0.161)	(0.162)	(0.149)	(0.135)	(0.128)
Residual, 2008		0.205	0.227		0.263**	0.232*
		(0.122)	(0.146)		(0.118)	(0.120)
Residual, 2010		0.205***	0.205***		0.182***	0.217***
		(0.056)	(0.060)		(0.052)	(0.051)
Affected by Hurricane Sandy		-0.003	-0.005		-0.004	-0.006
		(0.013)	(0.015)		(0.010)	(0.010)
Percent Black			-0.000			-0.001
			(0.001)			(0.001)
Percent Hispanic			-0.001*			-0.001***
			(0.001)			(0.000)
Percent College Grad			0.000			0.001
			(0.001)			(0.001)
Adjusted R-squared	0.04	0.29	0.29	0.23	0.44	0.53
Observations	46	46	46	46	46	46
Weighting	Unweighted	Unweighted	Unweighted	Weighted	Weighted	Weighted

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Notes: The dependent variable is the change in the natural log of the turnout rate from 2008 to 2012. Turnout Rate is total presidential votes divided by total eligible voters. These data were downloaded from The United States Elections Project, at `http://elections.gmu.edu/bio.html`. Predicted Δ ln(Turnout Rate) is the Google-based predictions, in Column (3) of Table II. All Google data are downloaded from Google Trends. Residual, 2008 is the residual from a regression of $\Delta ln(TurnoutRate), 2004 - 2008$ on $\Delta ln(GoogleVoteSearch), 2004 - 2008$. Residual, 2010 is the residual from a regression of $\Delta ln(TurnoutRate), 2006 - 2010$ on $\Delta ln(GoogleVoteSearch), 2006 - 2010$). Affected by Hurricane Sandy is a dummy variable that takes the value 1 for states Connecticut, Maryland, New Jersey, New York, North Carolina, and West Virginia. The demographics variables are from the 2000 Census. Weighted regressions are weighted by 2000 population, also from the Census. California, Vermont, Virginia, Delaware, and Washington D.C. are not included for reasons discussed in the text.

Table IV
Change in Google Search Rate and Demographics, 2004-2008

	Δ ln(Google Voting Search)			
	(1)	(2)	(3)	(4)
Percent Black	0.722***	0.864***	0.615**	1.028***
	(0.230)	(0.189)	(0.274)	(0.266)
Percent Hispanic		0.301***		0.533***
		(0.108)		(0.147)
Percent 18-34		0.156		0.249
		(0.863)		(0.854)
Percent 65-Plus		-0.085		0.281
		(0.873)		(0.981)
Percent College Grad		-0.488		-0.230
		(0.469)		(0.495)
Percent Kerry		-0.576		-0.865**
		(0.437)		(0.402)
Adjusted R-squared	0.09	0.21	0.16	0.25
Observations	176	176	176	176
Division Fixed Effects	No	No	Yes	Yes

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Notes: The dependent variable is the change in the natural log of Google search rates for the words "vote" or "voting," from October 2004 to October 2008. Regressions are for 176 media markets for which data was available. Percent Kerry is Kerry's share of the two-party vote. The demographics variables are from the 2000 Census. All regressions are weighted by 2000 population, also from the Census.

Table V
Change in Google Search Rate and Demographics, 2008-2012

			Δ ln(Google Voting Search)			
	(1)	(2)	(3)	(4)	(5)	(6)
ln(Total Population)	-0.152***	-0.124***	-0.148***	-0.160***	-0.128***	-0.160***
	(0.013)	(0.017)	(0.024)	(0.017)	(0.023)	(0.027)
Percent Black	-0.127	-0.108	-0.121	-0.184	-0.098	-0.197
	(0.155)	(0.163)	(0.183)	(0.198)	(0.215)	(0.211)
Percent Hispanic		0.029	0.032		-0.093	-0.168
		(0.082)	(0.117)		(0.099)	(0.147)
Percent 18-34		0.796	0.055		0.686	-0.352
		(0.836)	(0.669)		(0.903)	(0.745)
Percent 65-Plus		1.219*	0.658		1.230	0.290
		(0.707)	(0.971)		(0.870)	(0.993)
Percent College Grad		-0.554*	-0.672**		-0.565	-0.712*
		(0.281)	(0.328)		(0.359)	(0.371)
Percent Kerry		-0.009	0.061		0.016	0.107
		(0.129)	(0.172)		(0.154)	(0.135)
Percent Evangelical			0.121			0.203
			(0.205)			(0.265)
Percent Mainline Prot			-0.007			0.362
			(0.273)			(0.361)
Percent Catholic			0.063			0.224
			(0.201)			(0.193)
Percent Jewish			0.197			0.841
			(0.918)			(1.389)
Percent Islam			2.550			3.491
			(4.450)			(5.519)
Percent Other Religion			0.480***			0.684***
			(0.129)			(0.240)
Adjusted R-squared	0.67	0.69	0.70	0.69	0.70	0.72
Observations	194	194	194	194	194	194
Division Fixed Effects	No	No	No	Yes	Yes	Yes

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Notes: The dependent variable is the change in the natural log of Google search rates for the words "vote" or "voting," from October 2008 to October 2012. Regressions are for 194 media markets for which data were available. The demographics variables in Columns (2) and (5) are the same ones used in Table IV. Religion variables are from 2000 Congregations and Membership Study. All regressions are weighted by 2000 population, from the Census.

Table VI
Predicted 2012 Demographics

Demographic	*2004 Share*	*2008 Share*	*Predicted 2012 Share*	*Actual 2012 Share*
			FILLED IN PRIOR TO 2012 ELECTION	
	(1)	(2)	(3)	(4)
Black	11	13	13	13
Hispanic	8	9	10	10
White evangelical	23	26	26	26
Other religion	7	6	7	7
Age 18-29	17	18	18	19
Age 65+	16	16	16	16

Notes: Columns (1), (2), and (4) are demographics estimates from CNN's analysis of the 2004, 2008, and 2012 exit polls. Column (3) are predictions made prior to the 2012 election, based on the regressions in Table V.

Table VII
Democratic Overperformance and Change in Turnout

	Actual Democrat-Predicted Democrat					
	2008		2012		2010	
	(1)	(2)	(3)	(4)	(5)	(6)
Δ ln(Turnout Rate)	0.138**		0.018		0.099**	
	(0.056)		(0.099)		(0.047)	
Δ ln(Google Vote Search)		-0.014		-0.028		0.053***
		(0.015)		(0.021)		(0.018)
Actual Democrat-Predicted Democrat, 2008			0.310*	0.280*	0.918***	0.996***
			(0.157)	(0.163)	(0.275)	(0.204)
Adjusted R-squared	0.05	-0.01	0.09	0.13	0.46	0.54
Observations	46	46	46	46	31	31

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Notes: The dependent variable is the difference between the Democrat's share of the two-party vote and Nate Silver's final prediction for the Democrat's share of the two-party vote. In 2008 and 2012, this is Obama's vote share. In 2010, this is the Senate Democratic candidate's vote share. Δ ln(Turnout Rate) is the change in turnout rate compared to four years earlier. Δ ln(Google Vote Search) is the change in Google search rates for "vote" or "voting," in October, compared to the October four years earlier. Actual Democrat-Predicted Democrat, 2008 is the difference between Obama's 2008 vote share and Obama's 2008 vote share, as predicted by Nate Silver. Thus, it is also the dependent variable in Columns (1) and (2). Turnout Rate is total presidential votes divided by total eligible voters. These data were downloaded from The United States Elections Project, at `http://elections.gmu.edu/bio.html`. All Google data are downloaded from Google Trends. Nate Silver's predictions come from `http://trumantolong.blogspot.com/2008/11/grading-nate-silver-and-538.html`, `http://elections.nytimes.com/2010/forecasts/senate`, and `http://fivethirtyeight.blogs.nytimes.com/`. California, Vermont, Virginia, Delaware, and Washington D.C. are not included for reasons discussed in the text.

References

Ansolabehere, Stephen and Eitan Hersh, "Pants on Fire: Misreporting, Sample Selection, and Participation," *Working Paper*, 2011.

Askitas, Nikolaos and Klaus F. Zimmermann, "Google Econometrics and Unemployment Forecasting," *Applied Economics Quarterly*, 2009, *55* (2), 107 – 120.

Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant, "Detecting Influenza Epidemics Using Search Engine Query Data.," *Nature*, February 2009, *457* (7232), 1012–4.

Gomez, Brad T., Thomas G. Hansford, and George A. Krause, "The Republicans Should Pray for Rain: Weather, Turnout, and Voting in U.S. Presidential Elections," *The Journal of Politics*, July 2008, *69* (03).

Rogers, Todd and Masa Aida, "Why Bother Asking? The Limited Value of Self-Reported Vote Intention," *HKS Faculty Research Working Paper Series*, 2012.

Stephens-Davidowitz, Seth, "The Cost of Racial Animus on a Black Presidential Candidate: Evidence Using Google Search Data," *Working Paper*, 2012.

Varian, Hal R. and Hyunyoung Choi, "Predicting the Present with Google Trends," *SSRN Electronic Journal*, August 2010.

Vavreck, Lynn, "The Exaggerated Effects of Advertising on Turnout: The Dangers of Self-Reports," *Quarterly Journal of Political Science*, 2007, *2* (4), 325–343.