Quantitative Prediction of Electoral Vote for United States Presidential Election in 2016

Gang Xu
Senior Research Scientist in Machine Learning
Houston, Texas
(prepared on November 07, 2016)

Abstract

In this paper I am reporting the quantitative prediction of the electoral vote for United States presidential election in 2016. This quantitative prediction was based on the Google Trends (GT) data that is publicly available on the internet. A simple heuristic statistical model is applied to analyzing the GT data. This is intended to be an experiment for exploring the plausible dependency between the GT data and the electoral vote result of US presidential elections. The model's performance has also been tested by comparing the predicted results and the actual electoral votes in 2004, 2008 and 2012. For the year 2016, the Google Trends data projects that Mr. Trump will win the white house in landslide. This paper serves as a document to put this exploratory experiment in real test, since the actual election result can be compared to the prediction after tomorrow (November 8, 2016).

Introduction

Tomorrow, November 8 of 2016, shall be the election day for American people whose votes will determine the next US president. There have been several forecasting reports available on the internet. It is beyond the scope of this report to review all these work in details.

Lichtman had developed a pattern recognition method in early 1980s and has been able to correctly predict the past 30 years of presidential outcomes using this method [1]. Based on his method, Lichtman predicted that Trump is headed for a win in 2016 [2]. However his method cannot give the quantitative prediction of the electoral vote.

Silver has maintained a web site for the presidency forecasting [3]. As to the *ET 6:15 PM, November 7, 2016*, his model gave the results as shown in Fig 1, which predicted that Clinton would win the election with electoral vote of 299.

Pepper analyzed the Google Trends (GT) data for the keywords of candidate's last name with word “sign”. He discovered an interesting dependency pattern between the election result and the Google interest scores [3] since 2004. The work in this paper was inspired by the Pepper's analysis, since a curious question is raised as whether this dependency could be used to quantitatively predicting the electoral vote.

Who will win the presidency?

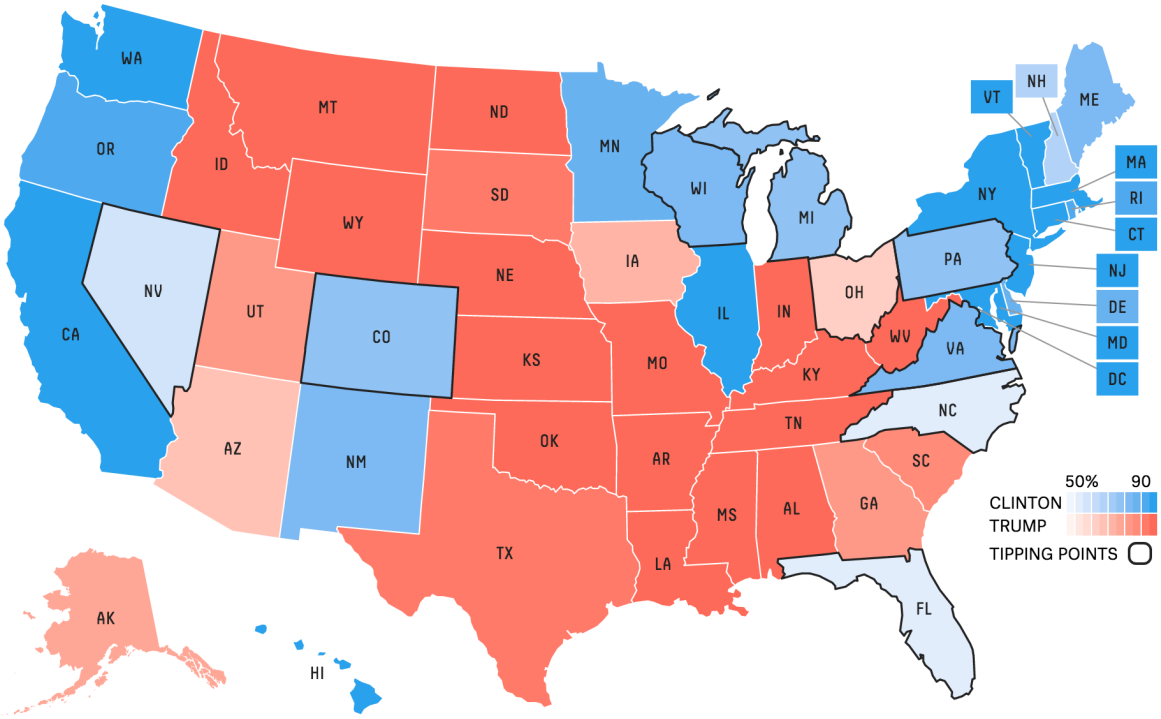
Chance of winning

Hillary Clinton

69.2%

Donald Trump

30.7%



Electoral votes

Hillary Clinton	299.1
Donald Trump	238.1
Evan McMullin	0.7
Gary Johnson	0.0

Popular vote

Hillary Clinton	48.6%
Donald Trump	45.1%
Gary Johnson	4.8%
Other	1.6%

at *ET 6:15 PM, November 7, 2016*

Data Description

The GT data with the same type of keywords as in Pepper's analysis were used in the current work. The keywords are the candidate's last name combined with the word “sign”. The GT data for the two candidates were downloaded (in CSV format). The original CSV data were slightly reformatted for the further processing and analyzing. The restricting conditions for the GT data search are:

- United States (region)
- 2004 – present (time)
- All categories (type)
- Web search (search)

Even though all the GT data from 2004 to present were downloaded, only the data of 3-year window up to the October of the election year were used for predicting the electoral vote.

As an illustrative example, the GT data for predicting the 2008 US presidential election are shown in Fig 2 and Tab 1.

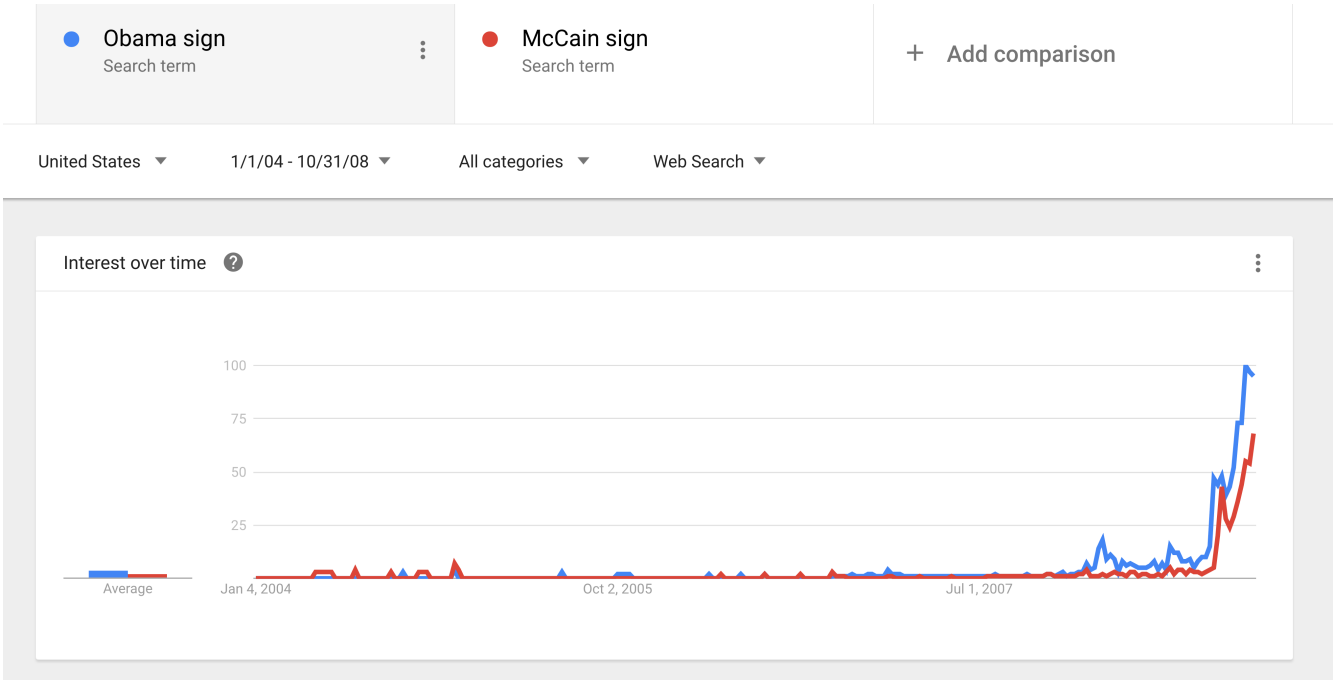


Figure 2. Google Trend Data for Predicting the US Electoral Vote in 2008

Table 1. Google Trend Data for Predicting the US Electoral Vote in 2008

Year	Month	Obama-sign	McCain-sign
2005	10	1	1
2005	11	1	1
2005	12	0	0
2006	1	0	0
2006	2	0	0
2006	3	0	0
2006	4	0	1
2006	5	0	1
2006	6	0	0
2006	7	0	0
2006	8	0	0
2006	9	0	1
2006	10	0	0
2006	11	0	0
2006	12	0	0
2007	1	1	0
2007	2	2	1
2007	3	1	0
2007	4	0	0
2007	5	0	0
2007	6	0	0
2007	7	2	0
2007	8	0	0
2007	9	1	0
2007	10	1	0
2007	11	1	0
2007	12	1	0
2008	1	5	1
2008	2	11	2
2008	3	7	2
2008	4	6	2
2008	5	8	2
2008	6	13	4
2008	7	9	3
2008	8	32	10
2008	9	51	36
2008	10	100	58

A Heuristic Theory and Statistical Model

To explore the quantitative relation between the above GT data and the actual electoral vote, a heuristic model is applied. The essential idea can be shown by a simple estimation given as follows.

Taking the data in Tab 1, summing the interest scores in columns of *Obama-sign* and *McCain-sign* gives overall scores *254* and *126*, respectively. The fractional ratio of the overall interest score for the *Obama-sign* is calculated as *254/(254+126)* ~ ***66.8%***. The fractional ratio of the overall interest score for the *McCain-sign* is calculated as *126/(254+126)* ~ ***33.2%***.

These two results can be compared to the actual electoral vote in 2008:

- Obama: *365 / 538* ~ ***67.8%***
- McCain: *173 / 538* ~ ***32.2%***

With **the theoretical assumption** *that the above numerical matching is not by coincidence, instead the electoral vote be statistically correlated with the GT interest scores for the presidential candidates,* we may therefore develop a statistical model to quantitatively predict the electoral vote.

A few technical details for the heuristic statistical model is briefly summarize as follows:

- The ensemble of models, based on the bootstrapping approach, is adopted to account for the statistical uncertainty.
- A deterministic bootstrapping procedure is applied. The ensemble set of bootstrap samples is given by $\{X_t \mid t_{min} \le t \le t_{max}, X_t = \{d_t, d_{t+1}\ldots, d_{max}\}\}$, where d_t represents a single data "point" of the GT data at time t (a certain month).
- The bootstrapping procedure is designed to give the higher weight to the GT data sample that are closer to the election month (November of the election year).
- The histogram of bootstrap samples is smoothed by a radial-basis kernel density model, so the *MAP*-like (*maximum a posteriori probability*) estimate, as well as the mean estimate, can be obtained

Comparison of Predictions with Electoral Vote in 2004, 2008 and 2012

The prediction performance has been evaluated using the historical ballot results in 2004 (Tab 2 and Fig 3), 2008 (Tab 3 and Fig 4) and 2012 (Tab 4 and Fig 5).

Table 2. Comparison of Prediction with Electoral Vote in 2004

	MAP Est.	Mean Est.	Actual Electoral Vote
Kerry	245	246	251
Bush	292	291	286

Table 3. Comparison of Prediction with Electoral Vote in 2008

	MAP Est.	Mean Est.	Actual Electoral Vote
Obama	359	353	365
McCain	179	185	173

Table 4. Comparison of Prediction with Electoral Vote in 2012

	MAP Est.	Mean Est.	Actual Electoral Vote
Obama	343	341	332
Romney	195	197	206

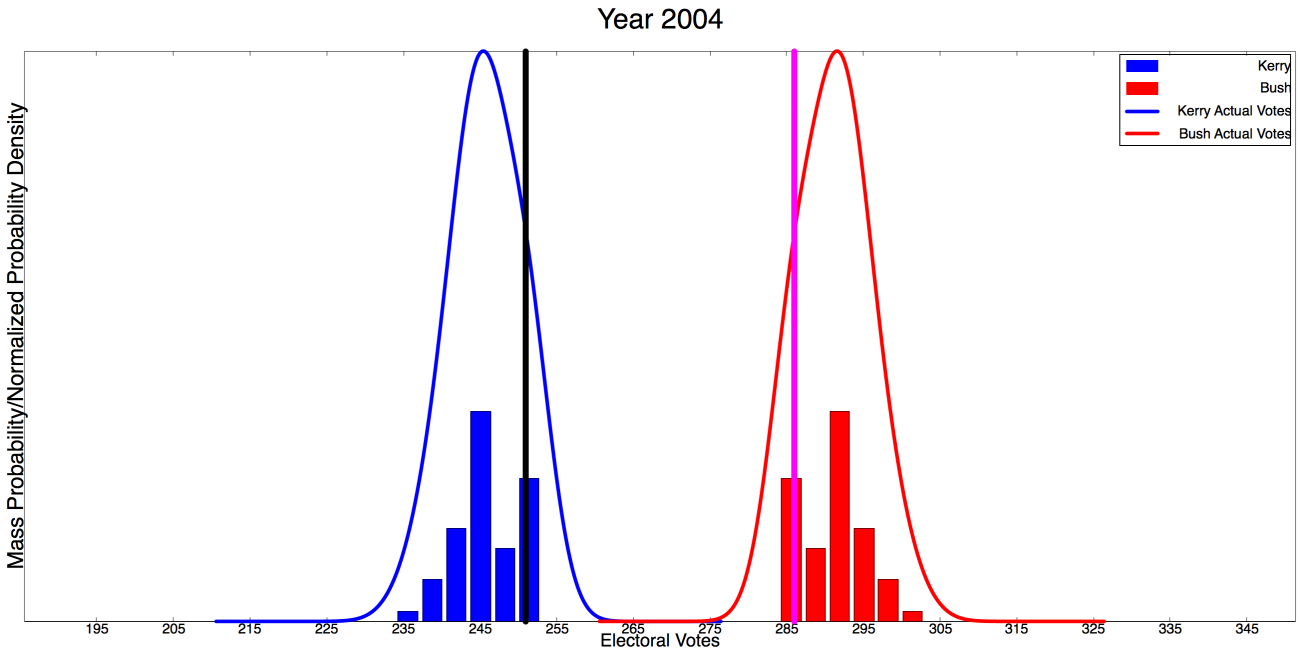


Figure 3. The Predicted Distribution of Electoral Vote (solid curves and histograms) versus The Actual Electoral Vote (vertical lines) in 2004

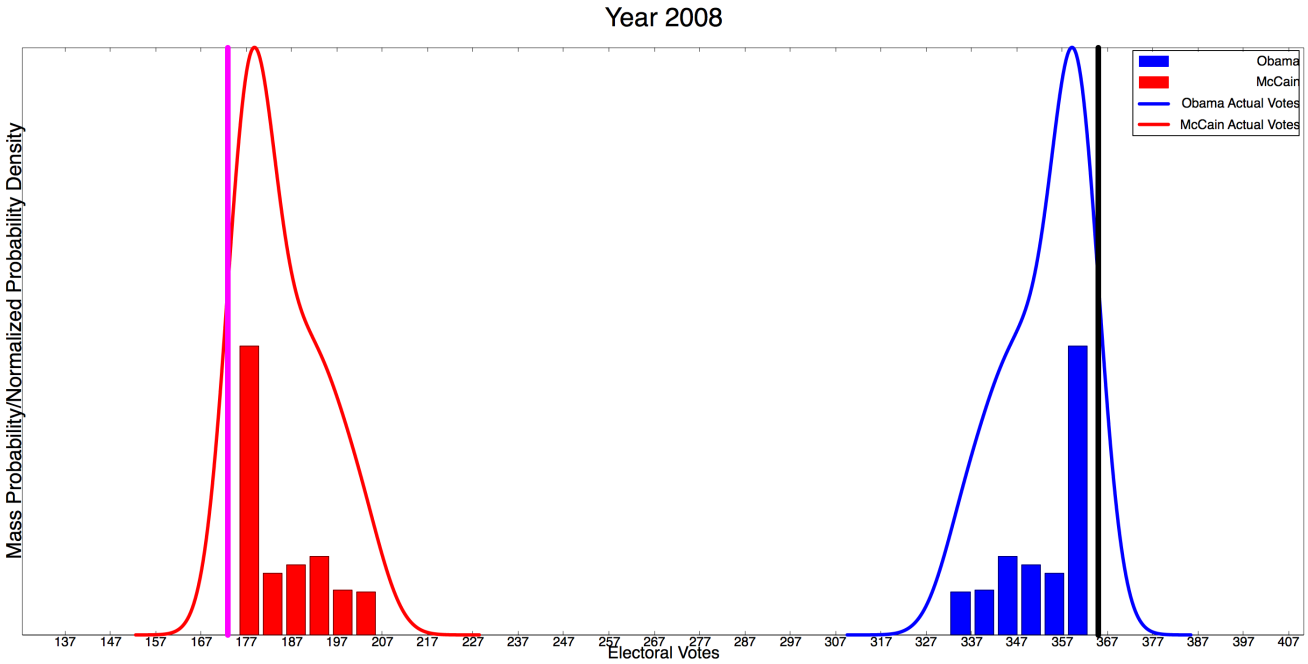


Figure 4. The Predicted Distribution of Electoral Vote (solid curves and histograms) versus The Actual Electoral Vote (vertical lines) in 2008

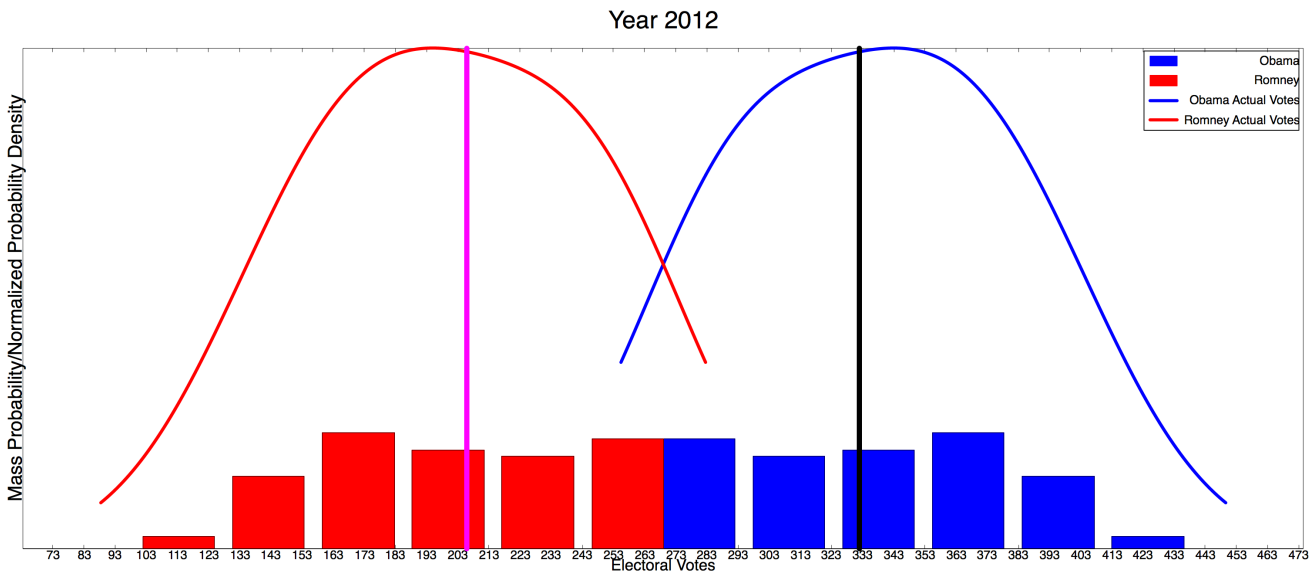


Figure 5. The Predicted Distribution of Electoral Vote (solid curves and histograms) versus The Actual Electoral Vote (vertical lines) in 2012

Prediction of Presidential Electoral Vote in 2016

The quantitative prediction and estimation of presidential electoral vote in 2016 have been shown in Tab 5 and Fig 6. It is seen that the distributions of electoral vote for Clinton and Trump are well separated in two different regions (124 – 163 for Clinton, 375 – 414 for Trump). It therefore indicates Trump will win the 2016 US presidential election in landslide (with *70% - 77%* of total electoral vote). This model prediction is subject to the falsification by the actual ballot result.

Discussions

Using the Google Trends data for quantitatively predicting the electoral vote appears to be an appropriate approach. It seems to well correlate the voter's sentiment and interest over the US nation with the actual electoral vote, based on 3 historical ballot results for the US presidential election.

In the current exploratory experiment, I have applied a simple statistical model to some well-selected Google Trends data. In future this work could be extended to several directions.

(1) to rigorously assess whether the speculated correlations exist or not;

(2) to understand why and how such quantitative correlations could be established, if ever existed;

(3) to improve the prediction accuracy by developing the better models, or the better selected data (e.g., Google Trends combined with other data sources)

Table 5. MAP Estimates, Mean Estimates, and Estimated Vote Ranges for Clinton and Trump (2016)

	MAP Est.	Mean Est.	Estimated Vote Range
Clinton	147	143	124 – 163
Trump	391	395	375 – 414

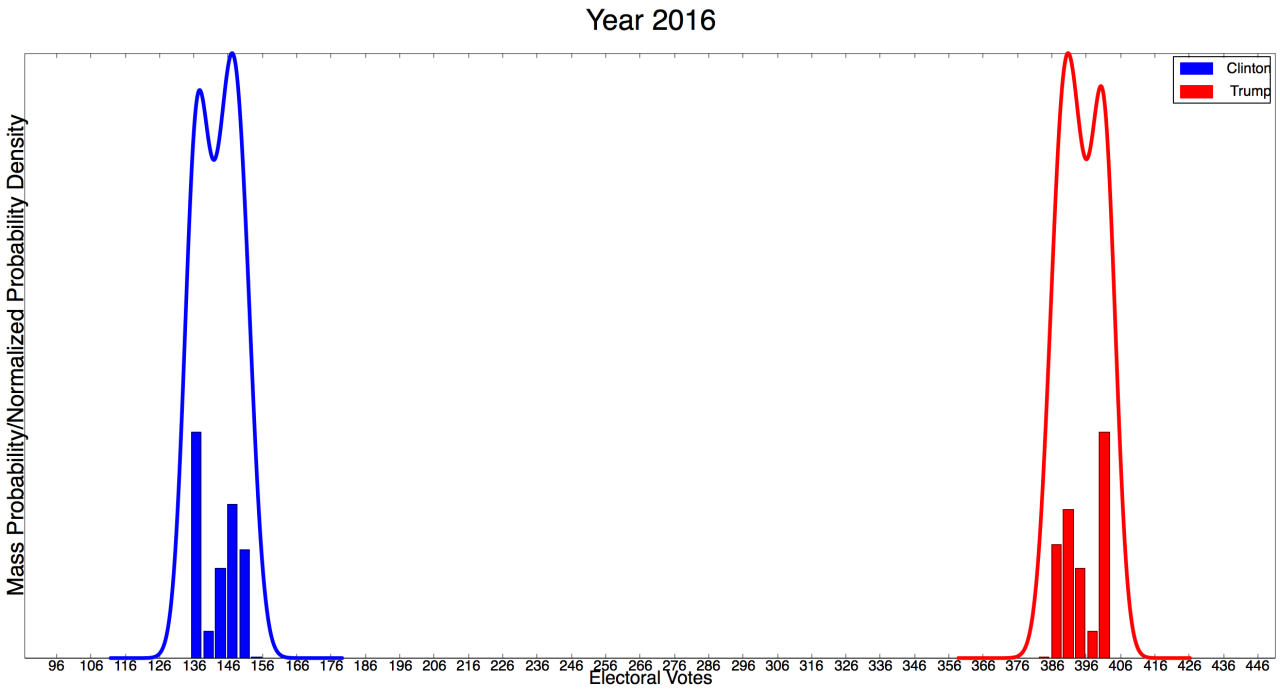


Figure 6. The Predicted Distribution of Electoral Vote (solid curves and histograms) for Clinton and Trump in 2016

References

1. A. J. Lichtman and V. I. Keilis-Borok, Pattern recognition applied to presidential elections in the United States, 1860-1980: Role of integral social, economic, and political traits, *Proc. Natl. Acad. Sci. USA*, Vol. 78, No. 11, pp. 7230-7234 (1981)
2. https://www.washingtonpost.com/news/the-fix/wp/2016/09/23/trump-is-headed-for-a-win-says-professor-whos-predicted-30-years-of-presidential-outcomes-correctly/
3. http://projects.fivethirtyeight.com/2016-election-forecast/?ex_cid=rrpromo
4. Ethan Pepper, Google Trends Indicate Trump Landslide, August 30, 2016, http://regated.com/2016/08/googles-trends-indicate-trump-landslide/