The Advantages (and Limitations) of Using Individual-Level Web Search Data to Study Political Behavior

Masha Krupenkin

Stanford University

October 19, 2018
Search engines are popular
Search engines are popular

1,200,000,000,000 Google searches per year (worldwide)
Search is an powerful data source to study political behavior
Overview

1. Why use search data?
Overview

1. Why use search data?
2. How to use and validate Bing search data
Overview

1. Why use search data?
2. How to use and validate Bing search data
3. Search data as a solution to survey response bias (Study 1)
Overview

1. Why use search data?
2. How to use and validate Bing search data
3. Search data as a solution to survey response bias (Study 1)
4. Using search data for causal inference (Study 2)
1. Why use search data?
## Benefits of Search Data

**Survey**

1. Potential for non-attitudes
2. Social desirability bias
3. Small sample
4. Temporally static
5. Connection to behavior contested

**Search**

1. Searching is motivated behavior
2. Avoids social desirability bias
3. Millions of users
4. High temporal granularity
5. Well documented connection to behavior
Search data predicts a variety of behaviors

- **Economic**: Purchasing and consumption behavior (Goel et al 2010; Bordino et al 2012; Choi and Varian 2012; Kinski 2016)
- **Health**: Drug adverse effects (Yom-Tov and E Gabrilovich 2013), cancer (Ofran et al 2012), and suicide (Recupero, Harms, and Noble 2008)
- **Political**: Voter registration (Street et al 2015) and turnout (Stephens-Davidowitz 2013)
- **Other**: Tourism (Choi and Varian 2009a) and unemployment (Choi and Varian 2009b)
## Limitations of Search Data

<table>
<thead>
<tr>
<th>Survey</th>
<th>Search</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Nationally Representative</td>
<td>Non-Representative</td>
</tr>
<tr>
<td>2. Long-lived data</td>
<td>Short-lived data</td>
</tr>
<tr>
<td>3. Experimentation possible</td>
<td>No experimentation</td>
</tr>
<tr>
<td>4. Some high profile failures</td>
<td>Some high profile failures</td>
</tr>
<tr>
<td>(1936 Literary Digest Poll)</td>
<td>(Google Flu Trends)</td>
</tr>
</tbody>
</table>
Google Flu Trends

- GFT was an attempt to predict flu incidence from search terms
Google Flu Trends

- GFT was an attempt to predict flu incidence from search terms
- Outcome variable = CDC data for flu-related doctor visits
Google Flu Trends

- GFT was an attempt to predict flu incidence from search terms (2003-2008)
- Outcome variable = CDC data for flu flu-related doctor visits
- Early results promising, serious issues starting in 2011
What Can We Learn From Google Flu Trends?

- User interaction with technologies (search engines) change over time
What Can We Learn From Google Flu Trends?

- User interaction with technologies (search engines) change over time
- Careful validation is key
What Can We Learn From Google Flu Trends?

- User interaction with technologies (search engines) change over time
- Careful validation is key
- Models can be improved (Yang, Santillana, and Kou 2015)
2. How to use and validate Bing search data
Who uses Bing?

### Worldwide
- Market share: 9%
- Monthly searches: 12B

### United States
- Market share: 33%
- Monthly searches: 5B

### Brazil
- Market share: 4%
- Monthly searches: 288M

### Canada
- Market share: 17%
- Monthly searches: 405M

### Latin America
- Market share: 5%
- Monthly searches: 895M
  - Argentina
  - Brazil
  - Chile
  - Colombia
  - Mexico
  - Peru
  - Venezuela
Linking to individual level covariates
Validating search data
Validating search data

1. Demographics
Validating search data

1. Demographics
2. Seasonality
Validating search data

1. Demographics
2. Seasonality
3. Survey Responses
Validating with demographics

Using established relationships between variables of interest and age/gender/ethnicity to link search and real world behavior
Validating with demographics

Mental health searches by gender
Validating with demographics

Voting searches by age (2016)
Validating with seasonality

Using established relationships between variables of interest and time to link search and real world behavior
Validating with seasonality

House purchase searches

Jan Apr Jul Oct Jan
Jan Apr Jul Oct Jan
Validating with seasonality

Searches for depression

% Searchers who search 'depression' (log scale)

Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec Jan
Validating with survey data

Explicitly asking survey respondents the question of interest and observing their searches
How we survey

Trump's voter fraud plan is a magnet for controversy

Which party do you most closely identify with?

- Strong Democrat
- Lean Democrat
- Independent/other
- Lean Republican
- Strong Republican

VOTE

Politics

Facebook gave more to Mueller on Russian buys than to Congress

The Hill
Validating with survey data

Is now a good or a bad time to buy a car?
Validating with survey data

Is now a good or a bad time to buy a house?
3. Search data as a solution to survey response bias (Study 1)
Flood of post-election mental health stories

In a divided U.S., therapists treating anxiety are hearing the same name over and over: Donald Trump

'Trump Anxiety Disorder' may not be an official diagnosis, but therapists know the symptoms

Matt Kwong - CBC

July 28, 2018

Dana Milbank - *President Trump Actually is Making Us Crazy* (Sep 2017)
... and Some Survey Support

### U.S. Worry Rises Significantly More After Election 2016 Than Election 2008

Did you experience the following feelings during a lot of the day yesterday? How about worry?

<table>
<thead>
<tr>
<th></th>
<th>U.S. adults</th>
<th>Republicans</th>
<th>Independents</th>
<th>Democrats</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
</tr>
<tr>
<td><strong>2008-2009</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-Election</td>
<td>34.7</td>
<td>28.3</td>
<td>36.1</td>
<td>37.9</td>
</tr>
<tr>
<td>Post-Inauguration</td>
<td>35.6</td>
<td>29.6</td>
<td>36.8</td>
<td>38.3</td>
</tr>
<tr>
<td>Change (pct. pts.)</td>
<td>+0.9</td>
<td>+1.3</td>
<td>+0.7</td>
<td>+0.4</td>
</tr>
<tr>
<td><strong>2016-2017</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-Election</td>
<td>29.2</td>
<td>25.7</td>
<td>30.7</td>
<td>30.8</td>
</tr>
<tr>
<td>Post-Inauguration</td>
<td>33.3</td>
<td>26.5</td>
<td>33.6</td>
<td>39.3</td>
</tr>
<tr>
<td>Change (pct. pts.)</td>
<td>+4.1</td>
<td>+0.8</td>
<td>+2.9</td>
<td>+8.5</td>
</tr>
</tbody>
</table>


GALLUP-HEALTHWAYS WELL-BEING INDEX
Did Trump’s election cause a mental health crises among Democrats?
Serious Social Desirability Problems
Serious Social Desirability Problems

1. Mental health is a stigmatized topic
Serious Social Desirability Problems

1. Mental health is a stigmatized topic
2. Potential for expressive reporting or virtue signalling
Research approach

- Look at searches for "depression", "anxiety", "stress", "suicide", "therapy" and antidepressant drugs
Research approach

- Look at searches for "depression", "anxiety", "stress", "suicide", "therapy" and antidepressant drugs
- Republican vs Democrat (300K searchers)
Research approach

- Look at searches for "depression", "anxiety", "stress", "suicide", "therapy" and antidepressant drugs
- Republican vs Democrat (300K searchers)
- Spanish-speaking (proxy for Latino) searchers vs English-only searchers (1 Million searchers)
Research approach

- Look at searches for "depression", "anxiety", "stress", "suicide", "therapy" and antidepressant drugs
- Republican vs Democrat (300K searchers)
- Spanish-speaking (proxy for Latino) searchers vs English-only searchers (1 Million searchers)
- Spanish-speaking = used Bing.es
Research approach

- Look at searches for "depression", "anxiety", "stress", "suicide", "therapy" and antidepressant drugs
- Republican vs Democrat (300K searchers)
- Spanish-speaking (proxy for Latino) searchers vs English-only searchers (1 Million searchers)
- Spanish-speaking = used Bing.es
- Searchers who used Bing before AND after election
Results
Latinos experienced a large rise in mental health searches...
... and English-only searchers did not
... Nor did Democrats
Conclusion

- Trump election was associated with a large rise in mental health-related searches among Latinos
- No similar response among Republicans, Democrats (or English-only searchers)
- Social desirability bias among Democrats
4. Using search data for causal inference (Study 2)
Do negative ads increase voter engagement?

- Highly-debated question in political science (Lau, Sigelman, and Rovner 2007)
Do negative ads increase voter engagement?

- Highly-debated question in political science (Lau, Sigelman, and Rovner 2007)
- Serious causal inference challenges:
Do negative ads increase voter engagement?

- Highly-debated question in political science (Lau, Sigelman, and Rovner 2007)
- Serious causal inference challenges:
  - Internal validity: confounders in observational studies
Do negative ads decrease voter engagement?

- Highly-debated question in political science (Lau, Sigelman, and Rovner 2007)
- Serious causal inference challenges:
  - Internal validity: confounders in observational studies
  - External validity: lab experiments do not measure real-world engagement
Approach

- Leverage temporal granularity to study causal effect of campaign ads
Approach

- Leverage temporal granularity to study causal effect of campaign ads
- Look at searches for candidates and voter registration 2 minutes before and 2 minutes after ad
Methods

- 3,939 national ads in 2016 that were at least 2 mins away from another presidential ad and above the adspend threshold (median adspend)
Methods

- 3,939 national ads in 2016 that were at least 2 mins away from another presidential ad and above the adspend threshold (median adspend)
- Ads weighted by adspend, which is a measure of audience. Higher adspend = more people watching the ad
Methods

- 3,939 national ads in 2016 that were at least 2 mins away from another presidential ad and above the adspend threshold (median adspend)
- Ads weighted by adspend, which is a measure of audience. Higher adspend = more people watching the ad
- Ads coded as positive or negative based on buyer (eg ad about Clinton from Trump campaign is negative)
Results
Ads increase engagement (candidate searches)

![Graphs showing the relationship between minutes since ad and searches weighted by adspend for pro-Clinton, anti-Clinton, pro-Trump, and anti-Trump categories.](image-url)
Ads increase engagement (voter registration searches)
No difference between positive and negative ads

<table>
<thead>
<tr>
<th></th>
<th>Candidates</th>
<th>Registration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trump</td>
<td>0.537***</td>
<td>-0.025***</td>
</tr>
<tr>
<td></td>
<td>(0.021)</td>
<td>(0.0004)</td>
</tr>
<tr>
<td>Days to Election</td>
<td>-0.007***</td>
<td>-0.025***</td>
</tr>
<tr>
<td></td>
<td>(0.0004)</td>
<td>(0.0004)</td>
</tr>
<tr>
<td>Post Ad</td>
<td>0.002</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td>(0.005)</td>
<td>(0.012)</td>
</tr>
<tr>
<td>Negative Ad</td>
<td>-0.098***</td>
<td>-0.437***</td>
</tr>
<tr>
<td></td>
<td>(0.025)</td>
<td>(0.048)</td>
</tr>
<tr>
<td>Adspend</td>
<td>-0.00005</td>
<td>-0.0005</td>
</tr>
<tr>
<td></td>
<td>(0.0001)</td>
<td>(0.0001)</td>
</tr>
<tr>
<td>Days to Election x Post Ad</td>
<td>-0.00003 (0.0001)</td>
<td>-0.0003* (0.0002)</td>
</tr>
<tr>
<td>Post Ad x Negative</td>
<td>-0.001 (0.005)</td>
<td>-0.007 (0.014)</td>
</tr>
<tr>
<td>Negative Ad x Adspend</td>
<td>-0.0001 (0.0002)</td>
<td>0.001 (0.001)</td>
</tr>
<tr>
<td>Post Ad x Adspend</td>
<td>0.0001***</td>
<td>0.0002**</td>
</tr>
<tr>
<td></td>
<td>(0.00002)</td>
<td>(0.0001)</td>
</tr>
<tr>
<td>Post x Negative x Adspend</td>
<td>-0.00002 (0.0001)</td>
<td>-0.0001 (0.0002)</td>
</tr>
<tr>
<td>Constant</td>
<td>3.176***</td>
<td>1.917***</td>
</tr>
<tr>
<td></td>
<td>(0.059)</td>
<td>(0.173)</td>
</tr>
<tr>
<td>Day of Week FE</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Hour FE</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

Note: *p < 0.1; **p < 0.05; ***p < 0.01
SE Clustered by Ad
Conclusion

- Political ads increase both information-seeking behavior about candidates and searches for voter registration
Conclusion

- Political ads increase both information-seeking behavior about candidates and searches for voter registration
- Positive and negative ads have similar effects
Final Thoughts
Final Thoughts

- Search data can solve common problems with survey research
Final Thoughts

- Search data can solve common problems with survey research
- Applications for a wide variety of topics: Political participation, perceptions of the economy, online news consumption, emotional responses to news, racist attitudes, advertising effects
Final Thoughts

- Search data can solve common problems with survey research
- Applications for a wide variety of topics: Political participation, perceptions of the economy, online news consumption, emotional responses to news, racist attitudes, advertising effects
- Useful properties for causal inference