



# Using Machine Learning Techniques to Interpret Open-ended Responses in Web Surveys

Laura Wronski, SurveyMonkey

Anna Boch, Stanford University

Reuben McCreanor, SurveyMonkey

Federal Committee on Statistical Methodology

March 7, 2018



# Open end potential in web surveys

# Pros and cons of using machine learning

Traditional Methods (for phone, in-person) | Machine Learning (for online surveys)

Pros

- Precodes simplify the task for interviewers

- Much faster than hand-coding for some types of analyses
- Preserves the individuality of responses
- Endless ways to analyze data

Cons

- Precodes/poor transcription may lose some of the nuance
- Hand-coding is onerous

- Difficult to combine analyses of short responses and long responses
- No one “correct” model to fit
- Time-consuming in its own way



# Experimental design

# Research Questions

- **Are some demographic groups more likely to respond to open-ended questions?**
  - Is this consistent for both political/apolitical questions?
- **What tools can be used to analyze open-ended data?**
  - How replicable are these methods?
  - Can we get to a point where the ease and consistency of analyzing open-ended questions matches that of closed-ended questions?

# Two-survey experiment

## Condition 1: Political open-end

- July 18-24, 2017
- 6,000+ respondents
- Trump approval:
  - *"Do you approve or disapprove of the way Donald Trump is handling his job as president?"*  
*Strongly approve*  
*Somewhat approve*  
*Somewhat disapprove*  
*Strongly disapprove*
- Immediately followed by open end:
  - *"Why?"*

## Condition 2: Apolitical open-end

- August 1-7, 2017
- 10,000+ respondents
- Trump approval:
  - *"Do you approve or disapprove of the way Donald Trump is handling his job as president?"*  
*Strongly approve*  
*Somewhat approve*  
*Somewhat disapprove*  
*Strongly disapprove*
- Standalone open end:
  - *"If you had \$1000 and you couldn't save it, invest it, or use it to pay off bills, what would you do with it?"*



Who answers open-  
ended questions?

# 1. Political open end is skipped more often

**Table 1. Item nonresponse for closed and open ended questions**

	Condition 1 – Political Open end		Condition 2 – Apolitical open end	
Total	99.4%	73.4%	97.8%	89.4%
Men	99.3%	74.3%	98.5%	88.2%
Women	99.5%	72.4%	97.1%	90.6%

## 2. Differences by demographics are consistent

	<i>Condition 1 – Political open end</i>		<i>Condition 2 – Apolitical open end</i>	
	<i>Trump approval</i>	<i>Open end</i>	<i>Trump approval</i>	<i>Open end</i>
<b>Age groups</b>				
18-34	99.4%	60.2%	97.4%	86.0%
35-49	98.9%	67.8%*	97.3%	89.4%***
50-64	99.6%	76.9%***	98.1%	91.6%*
65+	99.5%	83.4%***	98.4%	92.8%
<b>Race/Ethnicity</b>				
White	99.5%	75.2%	98.7%	90.8%
Black	98.7%	69.4%*	95.9%	88.4%
Hispanic	99.6%	65.9%*	96.6%	86.1%***
Asian	99.5%	54.8%***	92.8%	77.4%***
Other	98.3%	75.9%	96.3%	89.5%
<b>Education</b>				
HS or less	99.1%	64.4%	96.6%	87.1%
Some college	98.9%	71.3%	98.4%	91.5%***
College degree	99.6%	74.3%	98.8%	90.7%*
Graduate degree	99.7%	78.2%	98.6%	88.5%*

\*\*\* 0.001 \*\* 0.01 \* 0.05

### 3. Partisan effect is stronger for political open end

	Condition 1 – Political open end		Condition 2 – Apolitical open end	
	Trump approval	Open end	Trump approval	Open end
<b>Party</b>				
Republican	100%	69.3%	99.7%	89.8%
Lean Republican	99.6%	71.6%	99.0%	92.6%
Independent	99.2%	72.8%	97.7%	90.4%
Lean Democrat	99.8%	79.8%	98.9%	91.5%
Democrat	99.8%	77.7%	99.3%	90.4%
<b>Ideology</b>				
Very Conservative	100%	73.5%	98.6%	85.3%
Conservative	99.9%	68.9%**	99.2%	90.8%***
Moderate	99.6%	74.1%	98.5%	90.0%
Liberal	99.9%	79.4%	99.5%	92.5%
Very Liberal	99.4%	79.3%	99.9%	90.3%
<b>Trump approval</b>				
Strongly approve	100%	76.7%	100%	89.4%
Somewhat approve	100%	62.0%***	100%	88.4%
Somewhat disapprove	100%	64.9%**	100%	90.7%*
Strongly disapprove	100%	82.8%**	100%	91.2%

\*\*\* 0.001 \*\* 0.01 \* 0.05

# Summary

- Men and women are equally likely to respond
- Likelihood of response increases with age
  - Not just topic salience? Volunteerism bias?
- Whites are the most likely race to respond, Asians the least likely
  
- Political questions tend to magnify all of the above
- Magnitude of trump\_approval strong predictor of political open end response
  - Topic salience
- Political characteristics (party ID, ideology, trump\_approval) are irrelevant in predicting apolitical open end response



Who has more to say?

# Mean and median response length - demographics

## Condition 1: Political open end

Men		Women	
83.5		102.4	
49		61	
18-34	35-49	50-64	65+
79.6	90.2	95.6	100.7
30.5	48	59	68.5
White	Black	Hispanic	Asian
97.4	79.2	68.2	62.1
61	31	41	18
HS or less	Some college	College grad	Graduate degree
66.4	83.0	96.8	109.4
30	49	58	68.5

# Mean and median response length - partisanship

## Condition 1: Political open end

Republican	Lean Rep	Independent	Lean Dem	Democrat
78.6	83.8	89.8	120.9	102.4
45	53	51	81	63

Very conservative	Conservative	Moderate	Liberal	Very liberal
86.6	77.8	93.5	113.6	110.0
52	44.5	55	72	72

Strongly approve	Somewhat approve	Somewhat disapprove	Strongly disapprove
86.1	59.9	64.6	114.6
51	32	26	75

# Mean and median response length - demographics

## Condition 2: Apolitical open end

Men		Women	
21.9		28.6	
14		18	
18-34	35-49	50-64	65+
28.0	23.9	23.4	25.9
18	14	15	17
White	Black	Hispanic	Asian
25.6	25.0	24.5	22.0
16	15	15	13
HS or less	Some college	College grad	Graduate degree
25.0	26.4	22.0	23.7
15	17	16	15

# Mean and median response length - partisanship

## Condition 2: Apolitical open end

Republican	Lean Rep	Independent	Lean Dem	Democrat
24.0	26.3	27.7	25.6	25.6
15	16	17	16	16

Very conservative	Conservative	Moderate	Liberal	Very liberal
22.2	24.6	26.2	27.1	28.8
14	15	16	17	17

Strongly approve	Somewhat approve	Somewhat disapprove	Strongly disapprove
23.8	23.7	25.4	25.4
15	15	17	17

# Summary

- Women consistently have longer responses than men – True for both political and apolitical open ends
- Those who express a “strong” opinion are more likely to respond to an open end follow-up, but not more likely to respond to another open end on a different topic later in the survey
- Everything else varies by the open end itself
  - Age
  - Race
  - Education
  - Party ID
  - Ideology



# Machine learning techniques

# Word counts & bigram counts

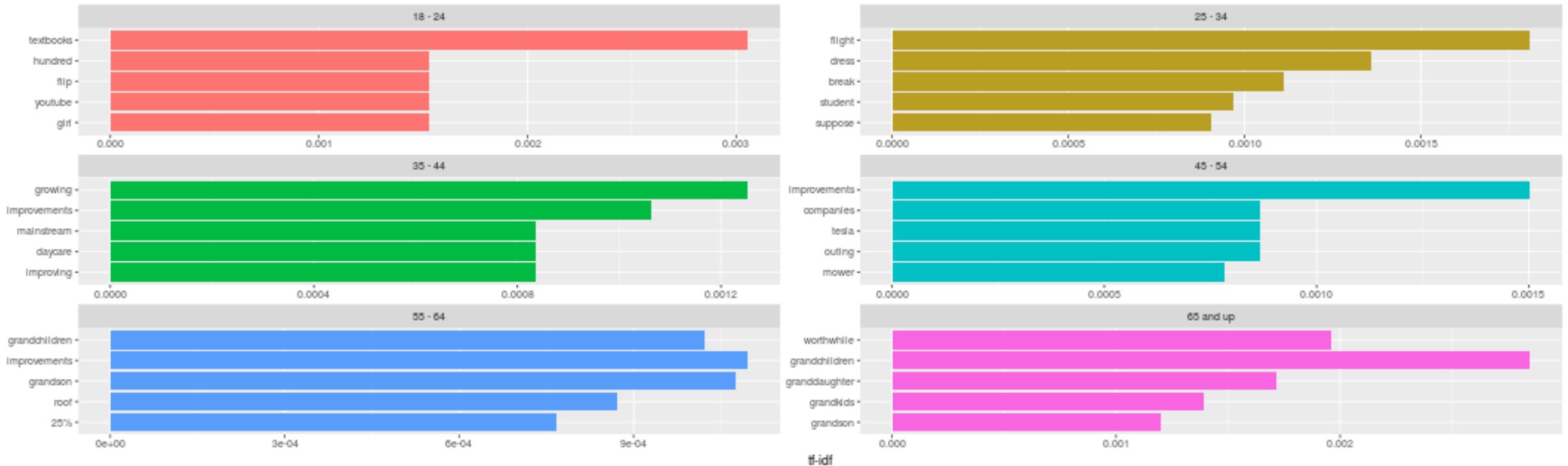
- Simple word or bigram frequencies
- Not a ML technique in itself, but correcting spelling, numeric characters, capitalization, etc. or simple grouping (e.g. invest|investing|investments) requires ML tools
- In R: *tm* package

pay	2002
buy	1913
bills	1759
vacation	1042
donate	850
save	849
invest	757
home	704
spend	591
family	526

# Tf-idf (term frequency-inverse document frequency)

How would you spend \$1000?

- Singles out words that are used with greater frequency among subgroups
- In R: *tidytext*, *dplyr*, *ggplot2* packages



# Topic modeling

- In R: *stm* package

Topic 3 Top Words:

Highest Prob: obama, barack, campaign, biden, polit, will, debat

FREX: ayer, barack, obama, wright, biden, jeremiah, joe

Lift: oct, goolsbe, ayerss, ayr, bernadin, ayer, annenberg

Score: oct, obama, barack, ayer, wright, campaign, biden

Topic 7 Top Words:

Highest Prob: palin, governor, sarah, state, alaska, polit, senat

FREX: blagojevich, palin, sarah, rezko, alaska, governor, gov

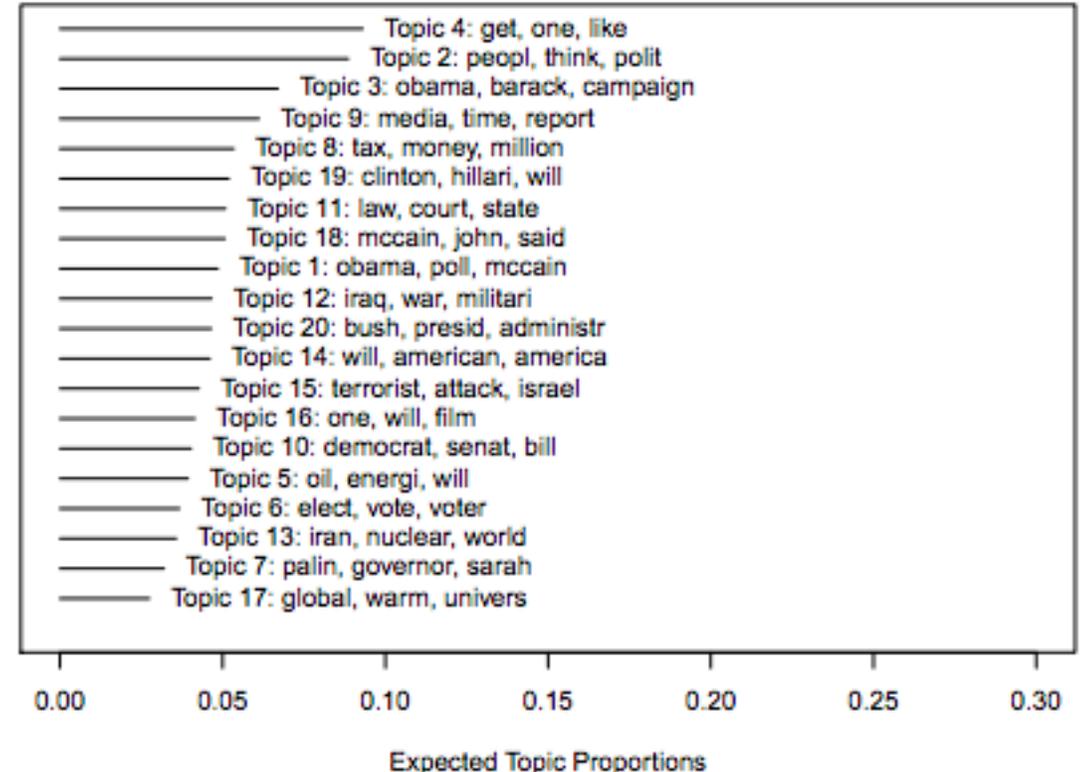
Topic 3

Here's video of the ad we reported on below that the Obama campaign is running in Ohio responding to the earlier Swift-Boating spot tying Obama to former Weatherman Bill Ayers... With all our pr

Topic 20

Waxman calls for release of FBI interviews with Bush and Cheney. In a letter to Attorney General Michael Mukasey today, Rep. Henry Waxman (D-CA), the Chairman of the Hous e Committee on Oversight

Top Topics





# Findings & recommendations

# Open end best practices

- **Open end responses aren't as representative as closed ended responses**
  - Skipped more frequently
  - Answer quality varies
- **But, open ended questions can be good complements to closed ended questions, especially when it comes to understanding differences between groups**
  - Word counts
  - Tf-idf
  - Topics
- **Still very difficult for the analysis to be standardized**

Thank you



[lauraw@surveymonkey.com](mailto:lauraw@surveymonkey.com)  
[research@surveymonkey.com](mailto:research@surveymonkey.com)



# Appendix



# SurveyMonkey's sampling methodology

# SurveyMonkey's unique recruitment flow

**500K surveys sent every month  
on 1000s of topics...**



# SurveyMonkey's unique recruitment flow

**500K surveys sent every month  
on 1000s of topics...**



**...receive 90M responses from a diverse  
cross-section of the population...**

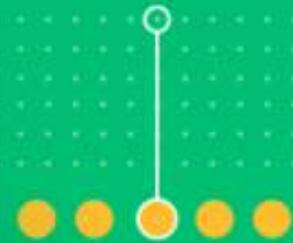
St

# What do you want to know?

Put your curiosity to work and get answers with the world's #1 online survey software.

PRO SIGN UP

SIGN UP FREE



**5.7**  
billion  
questions  
asked every  
year

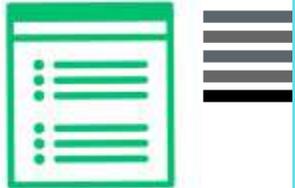
500K su  
on 1000



tion  
orm

# Survey

500K surveys sent  
on 1000s of topics

The main landing page features a light blue background. At the top center is a white icon of a dog's head. Below it, the title "Pause to help a puppy" is displayed in a large, bold, dark font. Underneath the title is a sub-headline: "Take a survey to raise \$0.50 for a participating charity of your choice. You could even win \$100!". On the right side of the page is a large, high-quality photograph of a brown and white puppy sitting and looking towards the camera. On the left side, there is a form with the label "Email Address:" above a white input field containing the text "you@example.com". Below the input field is a prominent green button with the text "GET STARTED" in white. At the bottom of the form area, there is a line of text: "By creating an account, you agree to our [Terms of Service](#) and [Privacy Statement](#)."

The creation platform interface has a green background. It features the text "What do you want to know?" in white. Below this is a smaller line of text: "The user-friendly way to learn and get answers. Ask the world &amp; get answers, without." At the bottom, there are two buttons: a yellow one labeled "ASK ABOUT IT" and a white one labeled "ASK ABOUT IT". On the right side, a woman is partially visible, holding a small object.

Creation  
Platform

The contribute panel interface is a smaller version of the main landing page, set against a light blue background. It includes the same puppy image on the right, the "Pause to help a puppy" title, the sub-headline, the "Email Address:" label, the input field with "you@example.com", and the green "GET STARTED" button. The text at the bottom is also present: "By creating an account, you agree to our [Terms of Service](#) and [Privacy Statement](#)."

Contribute  
Panel

# SurveyMonkey

Thank you for taking this survey.  
Powered by SurveyMonkey

# HOW

Where do you stand on current events? Share your opinion.



Take the Survey

Your responses will remain confidential and are for research purposes only.

Privacy Notice

Image courtesy of Robert Vanderbei

SurveyMonkey

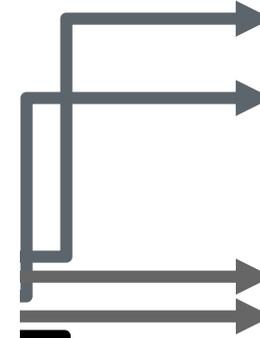
Create Your Own Survey

500K surveys sent every month on 1000s of topics...



...receive cross

ectly to a survey



with charity

Thank you for taking this survey.  
Powered by SurveyMonkey

Where do you stand on current events? Share your opinion.



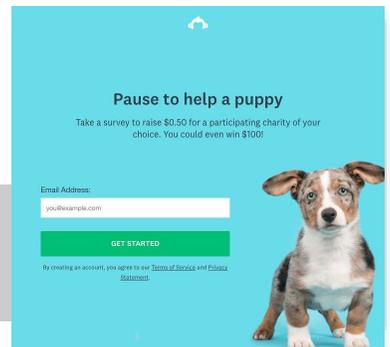
Take the Survey

Your responses will remain confidential and are for research purposes only.

## Election Tracking



## Creation Platform



## Contribute Panel