

# Exploring new ways of using Twitter content to augment survey data

Michael F. Schober



2022 FedCASIC Virtual Conference  
(Federal Computer Assisted Survey Information Collection)  
April, 2022



# Acknowledgments

- Collaborative agreement with Census Bureau (#CB20ADR0160002)
- Helpful conversations with Census colleagues from multiple directorates (Research and Methodology, Economic, Decennial, Communications)
- Team members on the collaborative agreement
  - Frederick Conrad, Johann Gagnon-Bartsch (faculty PI's, University of Michigan)
  - Mao Li, Juejue Wang (U of Michigan grad students)
  - Robyn Ferg (Westat)
  - Rebecca Dolgin, Unnati Shukla (The New School grad students)



How and in what ways can social media data best be used to connect with and enhance knowledge gained from traditional surveys?

- Early promising demonstrations that analyses of social media content can align with measurement from sample surveys
  - e.g., O’Connor et al. (2010), Tumasjan et al. (2011), Jensen & Anstead (2013), Ceron et al. (2014), Fu and Chan (2013), Sang and Bos (2012)
- Even though social media corpus is not designed to represent a population
  - or “designed” at all in Groves’ (2011) sense

**From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series**

**Brendan O’Connor<sup>†</sup>** **Ramnath Balasubramanyan<sup>†</sup>** **Bryan R. Routledge<sup>‡</sup>** **Noah A. Smith<sup>†</sup>**  
brenocon@cs.cmu.edu rbalasub@cs.cmu.edu routledge@cmu.edu nasmith@cs.cmu.edu

<sup>†</sup>School of Computer Science  
Carnegie Mellon University

<sup>‡</sup>Tepper School of Business  
Carnegie Mellon University

Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media

**Predicting Elections with Twitter:  
What 140 Characters Reveal about Political Sentiment**  
Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, Isabell M. Welpe

Technische Universität München  
Lehrstuhl für Betriebswirtschaftslehre Strategie und Organisation  
Leopoldstraße 139, 80804 Munich, Germany

# Excitement about using analyses of social media posts for social research

- At least two visions:
  1. Enhance survey data with social media content
    - e.g., include data derived from social media content in statistical models otherwise based on survey results
  2. Replace survey data with data derived from social media content
    - e.g., eliminate certain questions and variables they produce, or substitute social media-based measure in some waves

# Example early promising demonstration: O'Connor et al. (2010)

- Significant correlations over time between
  - sentiment ratios of tweets containing the word “jobs”
  - Gallup’s daily Economic Confidence tracking poll
  - U of Michigan’s monthly Index of Consumer Sentiment

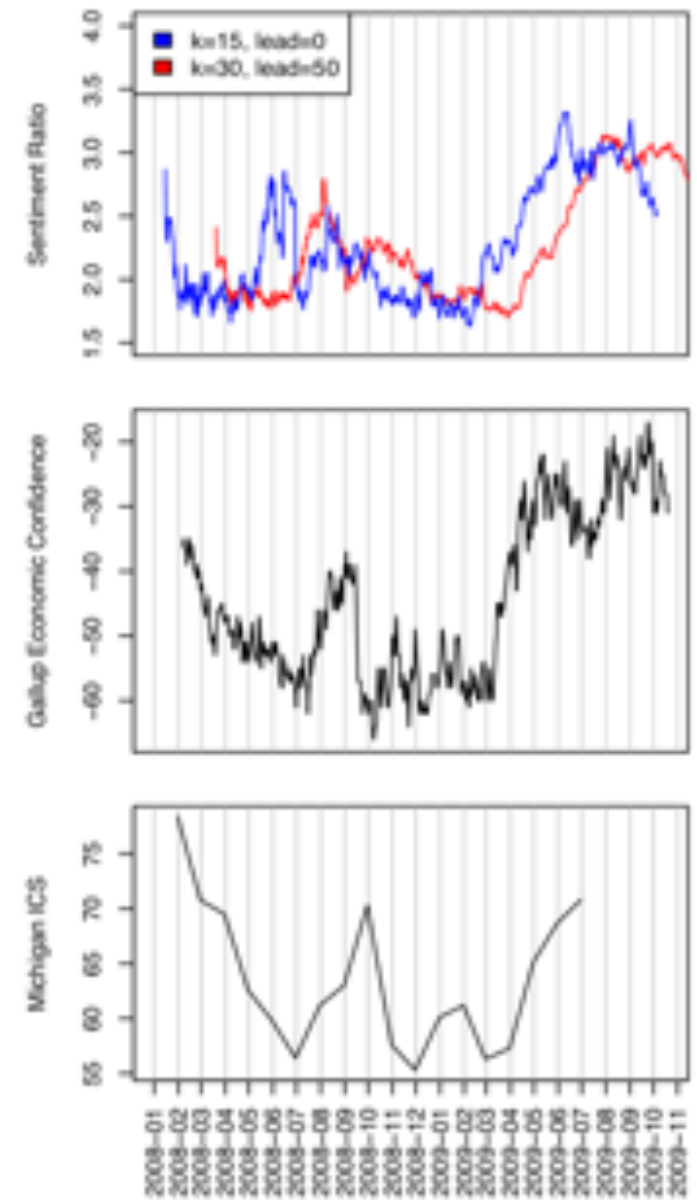


Figure 6: Sentiment ratio and consumer confidence surveys. Sentiment information captures broad trends in the survey data.

# But alignment may not persist over time

- Conrad et al. (2021) replicated correlation of “jobs” tweets and Index of Consumer Sentiment studied by O’Connor et al. (2010)
- But association did not extend into later time period and was not robust to changes in
  - Sentiment analysis method
  - Aggregation formula
  - Other data processing decisions (smoothing, lag)
- Removing clearly irrelevant tweets made association worse
- Original result was so fragile that authors concluded was likely spurious

# Intriguing demonstrations of possibilities continue to emerge

- Wikipedia Page Views
  - Smith & Gustafson (2017) modeled 100 US Senate races based on pre-election polls (surveys) with usual survey-based predictors (e.g., incumbent?) and either did or did not include number of views of candidates' Wikipedia pages
  - Models fit significantly better when number of pageviews was included
- Glassdoor
  - Symitsi et al. (2021) improved prediction of household expenditure in the US by combining survey-based consumer sentiment indices with employees' ratings of the business outlook for their companies on the Glassdoor platform
- (See other papers in *Public Opinion Quarterly* 2021 special issue “New Data in Social and Behavioral Research”)



# Jury is still out

- Need better theory about *when* social media data and survey data are most likely to tell us the same things about society
  - Survey and social media data are so different in nature that they could easily not align
  - But it's possible that under the right circumstances—when they tap into similar processes – they will align
- Perhaps more sophisticated analyses of textual material in social media posts can open new doors
- And there are likely other potential uses for social media data in social research beyond what has currently been investigated

# Two approaches our team has been taking

1. Understanding better the circumstances under which analyses of social media posts can plausibly align with survey findings
2. Exploring new potential uses of social media corpora to support researchers in generating qualitative insights

(I'll focus on each in turn)

# 1. Circumstances under which analyses of social media posts can plausibly align with survey findings

- Overview of our approach: Alignment should be more likely when
  1. survey responses have a relatively high signal-to-noise ratio over time (i.e., there is something for social media to align with)
    - Signal: Responses actually change over the time frame
    - Noise: Sampling variability can't be too high
  2. social media posts contain content semantically related to survey question content
    - NLP tools that select posts aligned with survey question content
  3. using available social media metrics that make sense for survey question content
    - Volume (frequency) of tweets when simply mentioning particular content is closely related to a survey response, e.g., Have you completed the Census?
    - Sentiment of tweets containing particular content when questions ask for valenced responses (e.g., positive or negative attitudes towards the topic)

# Domain: Knowledge and public opinion about the US Decennial Census January-September 2020

- Unusual opportunity for exploring these propositions
  - Content of survey includes both behavioral and attitudinal questions on a range of topics—during a period of societal controversy about the Decennial Census
  - Daily measures over 9 months available from both survey and social media data sets
  - Ongoing monitoring of social media content within US Census Bureau provides excellent access

# Survey data set: 2020 Census Tracking Survey

- Asks questions about people's knowledge of Census, exposure to ads, likelihood of participating, trust in data privacy, beliefs about uses of the data
- Web survey data (n = 76,919) from January 1-September 13, 2020 producing daily estimates from about 300 households per day
  - Two opt-in non-probability panels (Dynata and ThinkNow)
  - Sample selected to match American Community Survey quotas for Age x Gender, Region, Race and Hispanic Origin, and Education
  - (Probability sample telephone data set also exists for shorter time period--not analyzed here)

# Social media data set: Daily tweets

- 3,499,628 tweets from January 1-September 14, 2020
- all tweets that mention US Census and related content each day
- Collected via complex query (designed by Census Bureau) through Sprinklr platform

## List of Keywords retrieved from Sprinklr for September 16<sup>th</sup>, 2020

```
((("Census" OR "American Community Survey" OR "American Communities Survey" OR "@uscensusbureau" OR "#Census" OR "#Census2020" OR "#Census2010" OR "censuses" OR "#2020census" OR "2020census" OR "UScensus" OR "citizenship question" OR "#CensusDay" OR "#bluecensusboycott" OR "#censusboycott" OR "2020census.gov" OR "my2020census.gov" OR "Household Pulse Survey" OR "becountednow.com") OR ("#census" OR "#2020census" OR "#census2020" OR "#citizenshipquestion" OR "#census2010" OR "#censusday" OR "#data" OR "#2010census" OR "#travel" OR "#photography" OR "#nature" OR "#vote2020" OR "#history" OR "#repost" OR "#tbt" OR "#instagood" OR "#jobs" OR "#hiring" OR "#love") AND ("Census" OR "American Community Survey" OR "American Communities Survey" OR "@uscensusbureau" OR "#Census" OR "#Census2020" OR "#Census2010" OR "censuses" OR "#2020census" OR "2020census" OR "UScensus" OR "citizenship question" OR "#CensusDay" OR "#bluecensusboycott" OR "#censusboycott" OR "2020census.gov" OR "my2020census.gov" OR "Household Pulse Survey" OR "becountednow.com")) AND NOT (("Azerbaijan" OR "Belarus" OR "Burkina Faso" OR "Burundi" OR "Cambodia" OR "Cameroon" OR "Chad" OR "Congo" OR "Cote d'Ivoire" OR "Democratic Republic of Congo" OR "Djibouti" OR "Ethiopia" OR "Guinea Bissau" OR "Haiti" OR "Kenya" OR "Liberia" OR "Maldives" OR "Mali" OR "New Caledonia" OR "Nicaragua" OR "Nigeria" OR "North Korea" OR "Solomon Islands" OR "Sudan" OR "Vatican City" OR "Vietnam" OR "Kenyan" OR "ethiopian" OR "nigerian" OR "India" OR "South Africa" OR "Statistics SA" OR "Pakistan" OR "kenyans" OR "#census2019" OR "#census19" OR "Kibera" OR "#censuskenya2019" OR "#intersexcensuske" OR "Maseno" OR "Kisumu" OR "Nairobi" OR "British" OR "Ksh" OR "Kshs" OR "#kenyacensus2019" OR "@KNBStats" OR "Kitengela" OR "co.ke" OR "Isiolo" OR "2011 census" OR "census 2011" OR "Kiharu" OR "Office of National Statistics" OR "Jammu City" OR
```

# Measuring alignment

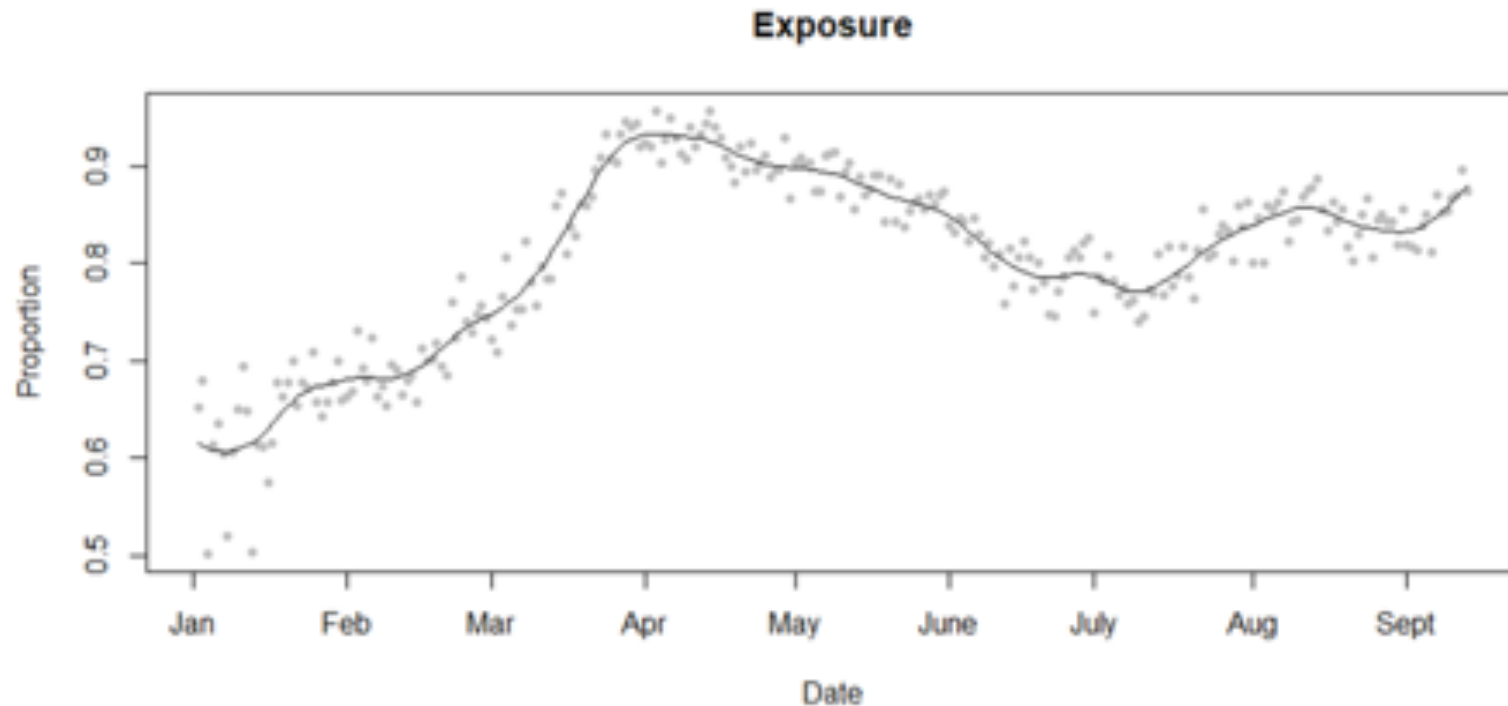
- *Comovement*: fraction of time that two time series move in the same direction (Fechner, 1897; Moore & Wallis, 1943; Goodman & Grunfeld, 1961)

$$\text{comovement}(x, y) = \frac{1}{T-1} \sum_{t=2}^T 1[\text{sgn}(x_t - x_{t-1}) = \text{sgn}(y_t - y_{t-1})]$$

- Ranges from 0 (perfect anti-association) to 1 (perfect association), with .5 being chance association
- Compared to Pearson correlation, less sensitive to outliers and less likely to detect spurious alignment
- Requires a specified time step (e.g., day to day vs. week to week)

# Question with good (0.83) signal-to-noise ratio:

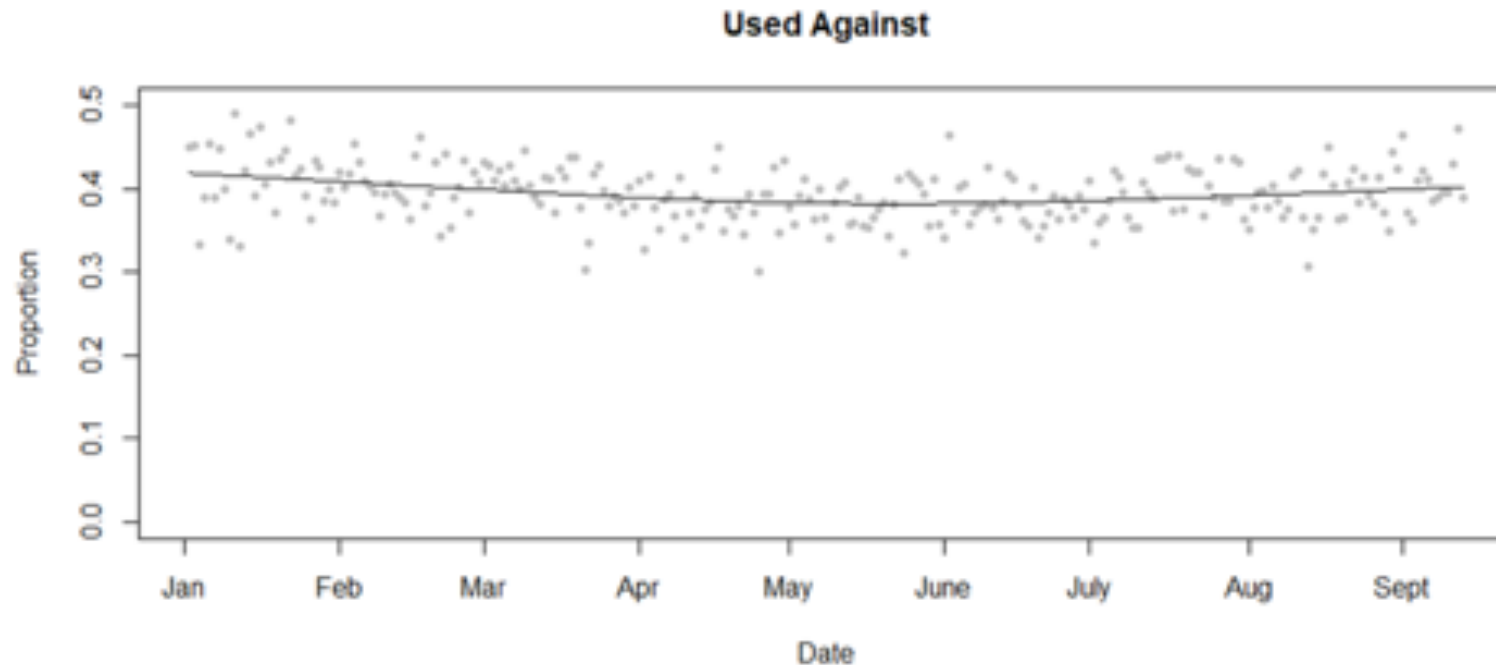
How much have you seen or heard recently – within the last week of so – about the 2020 Census?





# Question with poor (0.14) signal-to-noise ratio:

How concerned are you, if at all, that the answers you provide to the 2020 Census will be used against you?



# Efforts to find alignment:

## Selecting tweets that are relevant to the survey question

- Tested several NLP tools to automate selection of relevant tweets
- SBERT\* is best performing – because sensitive to meaning, not just word match
- Assigned score to each tweet indicating semantic distance from survey question and we found a cutoff using a systematic method involving some human judgment

\*Sentence-Bidirectional Encoder Representations from Transformers

See for one description: <https://medium.com/dair-ai/tl-dr-sentencebert-8dec326daf4e>



- (If interested: One useful overview of techniques, application domains, and particular challenges of topic-modeling short texts like social media posts)

<https://www.frontiersin.org/articles/10.3389/frai.2020.00042/full>

## Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis

Rania Albalawi<sup>1\*</sup>, Tet Hin Yeap<sup>1\*</sup> and Morad Benyoucef<sup>2\*</sup>

<sup>1</sup> School of Information Technology and Engineering, University of Ottawa, Ottawa, ON, Canada, <sup>2</sup> Telfer School of Management, University of Ottawa, Ottawa, ON, Canada

### OPEN ACCESS

#### Edited by:

Anis Yazid,  
OsloMet—Oslo Metropolitan  
University, Norway

#### Reviewed by:

Lei Jiao,  
University of Agder, Norway  
Ashish Rauniyar,  
University of Oslo, Norway  
in Collaboration With Reviewer LJ  
Imen Ben Sassi,  
Tallinn University of  
Technology, Estonia  
Desta Haileselassie Hagos,  
Oslo Metropolitan University, Norway  
in Collaboration With Reviewer IS

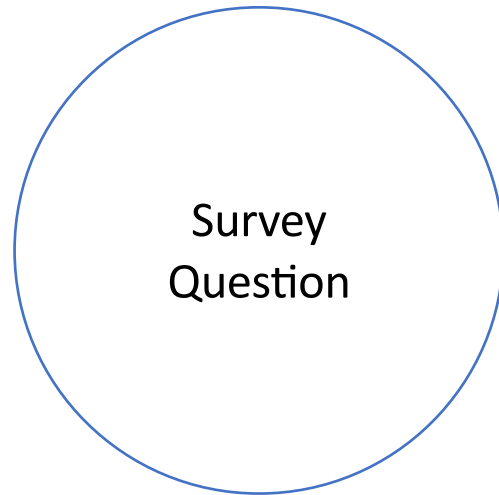
#### \*Correspondence:

Rania Albalawi  
raiba028@uottawa.ca  
Tet Hin Yeap  
tet@eecs.uottawa.ca  
Morad Benyoucef

With the growth of online social network platforms and applications, large amounts of textual user-generated content are created daily in the form of comments, reviews, and short-text messages. As a result, users often find it challenging to discover useful information or more on the topic being discussed from such content. Machine learning and natural language processing algorithms are used to analyze the massive amount of textual social media data available online, including topic modeling techniques that have gained popularity in recent years. This paper investigates the topic modeling subject and its common application areas, methods, and tools. Also, we examine and compare five frequently used topic modeling methods, as applied to short textual social data, to show their benefits practically in detecting important topics. These methods are latent semantic analysis, latent Dirichlet allocation, non-negative matrix factorization, random projection, and principal component analysis. Two textual datasets were selected to evaluate the performance of included topic modeling methods based on the topic quality and some standard statistical evaluation metrics, like recall, precision, *F*-score, and topic coherence. As a result, latent Dirichlet allocation and non-negative matrix factorization methods delivered more meaningful extracted topics and obtained good results. The paper sheds light on some common topic modeling methods in a short-text context and provides direction for researchers who seek to apply these methods.

# SBERT quantifies semantic distance between tweets and survey question, e.g.

“Do you think the 2020 Census questionnaire will or will not ask which people living in your household are U.S. citizens?”



Chose target survey question

“The 2020 Census questionnaire will NOT ask which people in their households are citizens.”

“#FAQ: Does the 2020 Census ask about citizenship status?  
No. The 2020 Census does not ask whether you or anyone in your home is a U.S. citizen.”

“RT @GUslavery Dr. Young was one of four doctors credited with founding the Medical Department at the college in 1850. Two other department founders, Flodoaro Howard and Johnson Elliot were also listed as slaveholders on the 1860 census.”

Similarity score between survey question and tweets using SBERT

Closer = more related to question

Farther = less related to question

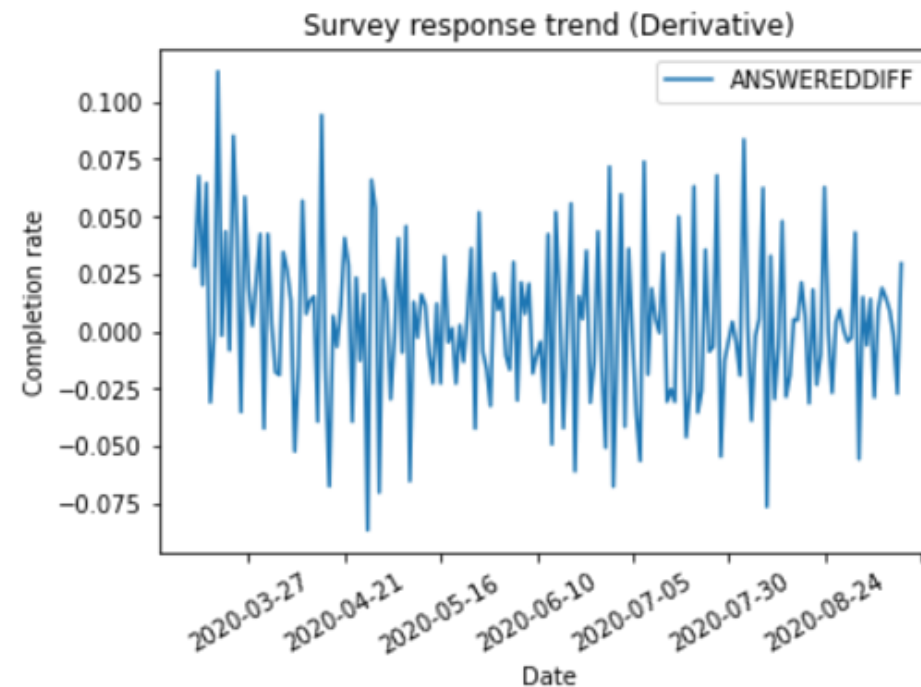
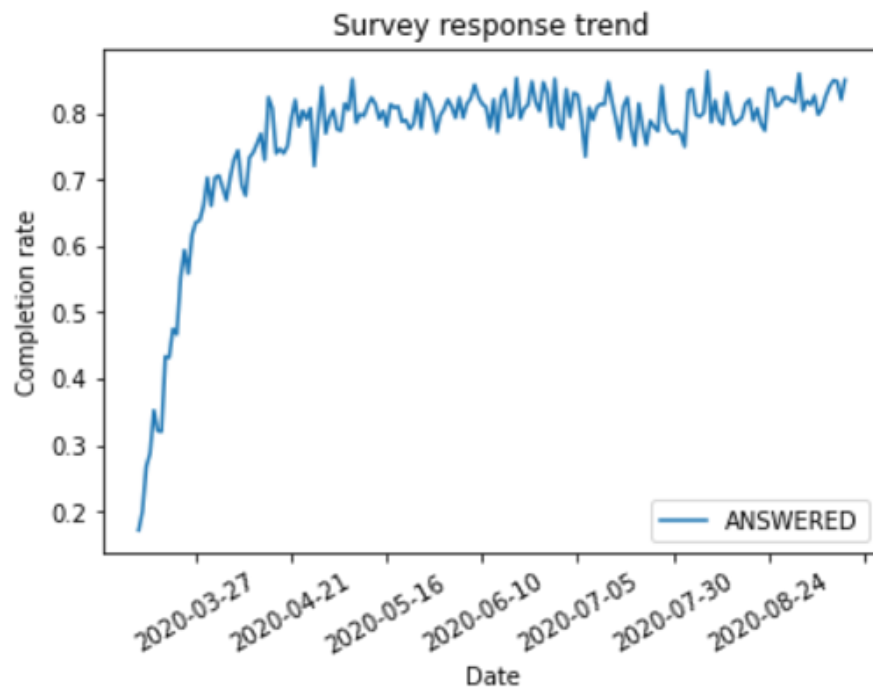
Cut-off point

# Efforts to find alignment: Accounting for cumulative character of responses to some survey questions

Have you or someone in your household answered the 2020 Census questions, or has your household not answered them yet? (starting March 12)

“Yes” responses will only increase over time

By taking the derivative, the response data become daily change data which is more comparable to variation in daily tweets



We observe co-movement!\*

Volume, full corpus: 9/24 questions

Volume, relevant tweets: 8/24

Sentiment, full corpus: 8/24

Sentiment, relevant tweets: 6/24

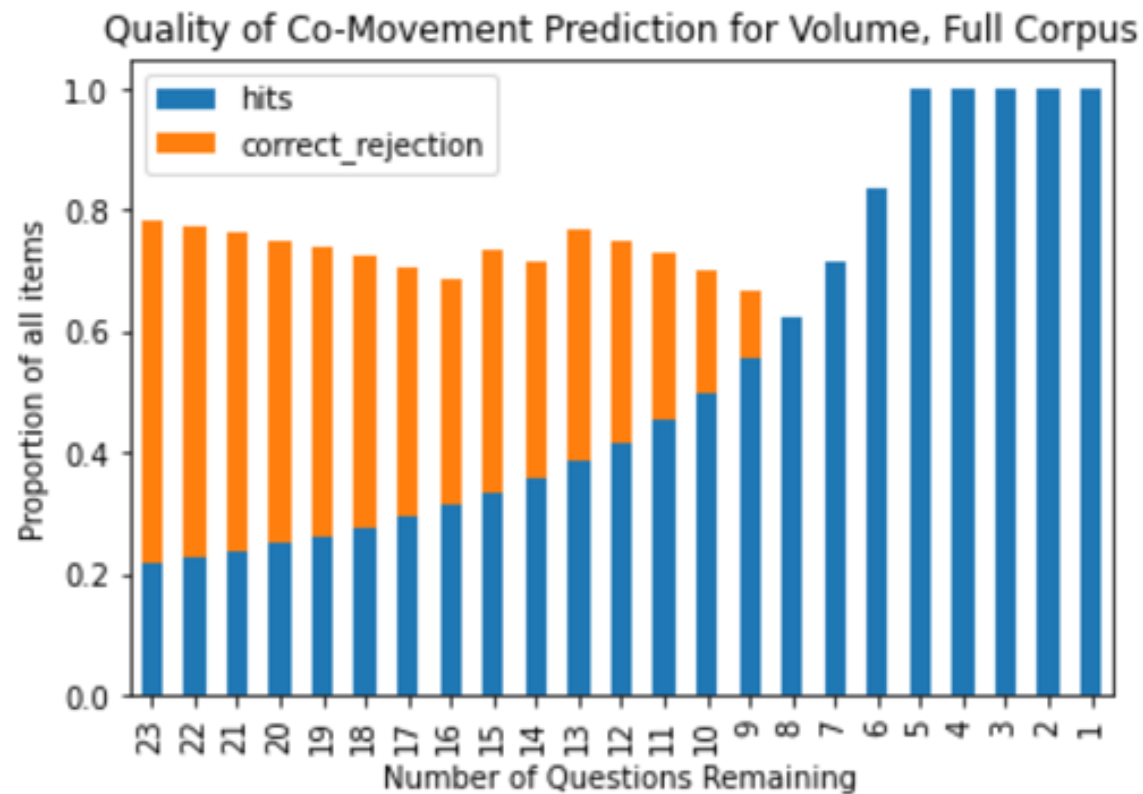
Are these what we would predict, given question content for each combination of measure and corpus?

How does signal-to-noise ratio affect prediction of co-movement?

\*significant and marginally significant

# Accuracy of co-movement predictions for different levels of signal-to-noise ratio

- For each measure-corpus pair, predicted co-movement based on question content
- Calculated hits and Correct Rejections
- Successively removed (left-to-right) questions with lowest signal-to-noise ratio



This work leads us to think that

Predicting when co-movement is and is not likely seems to be possible

At least for this questionnaire and corpus of Census-related tweets

But it takes a lot of work, computing skill, and judgment about how responses and tweets should move together given content of question

Might be a different story for a different survey questionnaire; worth exploring

(To be presented at AAPOR in Chicago)



## 2. Exploring new potential uses of social media corpora to support researchers in generating qualitative insights

- Rather than starting with the metaphor of social media posts on a topic as being like responses to survey questions (which leads down the road of exploring alignment) what other analogies might be fruitful?
- *Focus group metaphor*: social media data may be regarded as similar to data gathered from a (very large!) focus group
  - Comments generated by a broad range of stakeholders
  - who have self-selected into discussing the topic
  - and may display a broader range of opinions than a small focus group could ever hope to target (see Chen & Tomblin, *POQ* 2021)
- (Of course, different in many ways than a focus group that is moderated, topic-focused, and synchronously copresent)
- Just as researchers generate useful insights about public opinion from analysis of focus group discussion (and even coding of transcripts), might researchers generate useful insights from Twitter corpora?

# Challenge: how to select samples of tweets from a large corpus for analyst to review?

- How many?
  - Could someone look at 1,000? 300? There's some limit - how much can people tolerate, and when do we hit diminishing returns?
  - How long does it take?
  - How painful is it?

→ automated support is needed

- How to sample?
  - All (including retweets) vs. unique
  - Weighted (more weight to more retweeted?) vs. unweighted (random)
  - Filtered for content first? (narrow down by keyword or topic cluster)

# Our ambition: Develop a “Tweet Browser” tool that provides automated support for unearthing insights from social media corpus

Requires better understanding how analysts developing insights from careful reading of social media content

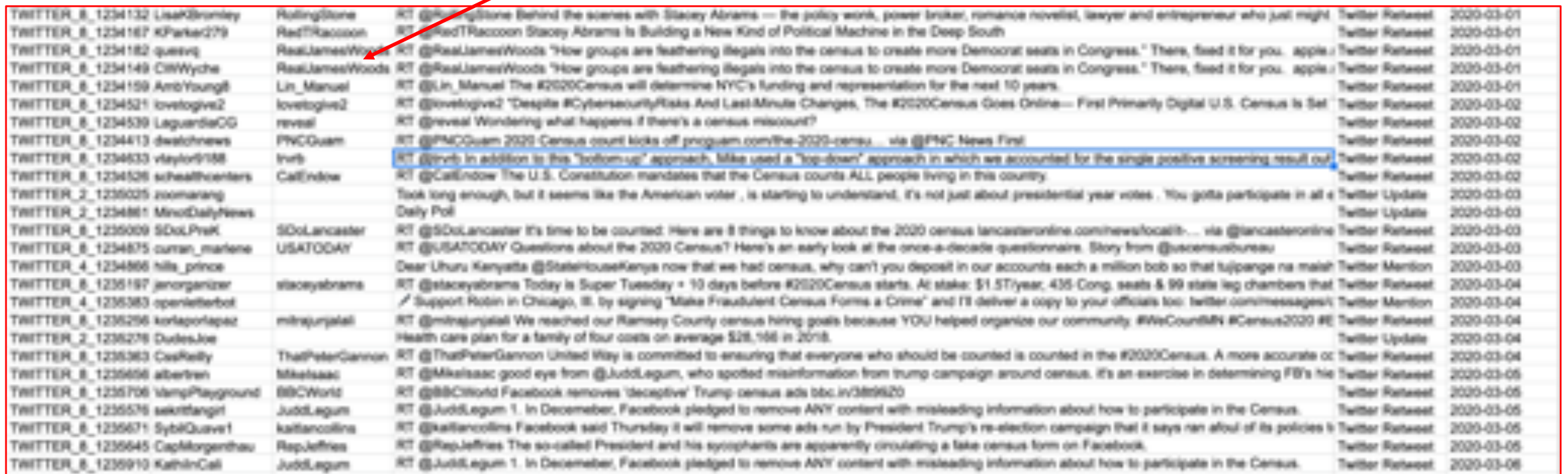
Our approach thus far has been to iteratively carry out, amongst ourselves and with Census colleagues, a series of exercises in which we

- Generated and read several samples of 150-300 tweets, and evaluated our ability to generate insights
- Explored:
  - How to define a corpus (e.g., keyword)?
  - How coherent / narrow to make the corpus
  - How to sample tweets
  - How many tweets
  - How to format (order by date, frequency? additional information like number of retweet,? etc.)
  - How to focus analyst’s goal (what themes to look for – e.g., barriers to participation? Citizenship question?)
  - How to record analyst’s insights (format for reporting)



# How to format: Displaying tweets: **retweets** vs unique tweets

@RealJamesWoods retweets, which accounted for 5469 tweets in the corpus, came up 9 times in a random sampling that included retweets. This led to the question: Are multiple repeats helpful to see like this?



TWITTER_8_1234132	LisaKBrumley	RollingStone	RT @RollingStone Behind the scenes with Stacey Abrams — the policy wonk, power broker, romance novelist, lawyer and entrepreneur who just might	Twitter Retweet	2020-03-01
TWITTER_8_1234167	KParker279	RealTRaccoon	RT @RealTRaccoon Stacey Abrams is Building a New Kind of Political Machine in the Deep South	Twitter Retweet	2020-03-01
TWITTER_8_1234182	queevq	RealJamesWoods	RT @RealJamesWoods "How groups are feathering illegals into the census to create more Democrat seats in Congress." There, fixed it for you. apple	Twitter Retweet	2020-03-01
TWITTER_8_1234149	CIRWyche	RealJamesWoods	RT @RealJamesWoods "How groups are feathering illegals into the census to create more Democrat seats in Congress." There, fixed it for you. apple	Twitter Retweet	2020-03-01
TWITTER_8_1234159	AmbYoung8	Lin_Manuel	RT @Lin_Manuel The #2020Census will determine NYC's funding and representation for the next 10 years.	Twitter Retweet	2020-03-01
TWITTER_8_1234521	lovetogive2	lovetogive2	RT @lovetogive2 "Despite #CybersecurityRisks And Last-Minute Changes, The #2020Census Goes Online— First Primarily Digital U.S. Census Is Set	Twitter Retweet	2020-03-02
TWITTER_8_1234539	LeguandaCG	reveal	RT @reveal Wondering what happens if there's a census miscount?	Twitter Retweet	2020-03-02
TWITTER_8_1234413	dwatchnews	PWCGuam	RT @PWCGuam 2020 Census count kicks off program.com/the-2020-census... via @PWC News First	Twitter Retweet	2020-03-02
TWITTER_8_1234633	vtaylor9188	ivrb	RT @ivrb In addition to his "bottom-up" approach, Mike used a "top-down" approach in which we accounted for the single positive screening result out	Twitter Retweet	2020-03-02
TWITTER_8_1234526	ishealthcenters	CallEndow	RT @CallEndow The U.S. Constitution mandates that the Census counts ALL people living in this country.	Twitter Retweet	2020-03-02
TWITTER_2_1235025	zoomarang		Took long enough, but it seems like the American voter , is starting to understand, it's not just about presidential year votes . You gotta participate in all e	Twitter Update	2020-03-03
TWITTER_2_1234861	MindDailyNews		Daily Pol	Twitter Update	2020-03-03
TWITTER_8_1235009	SDoLancaster	SDoLancaster	RT @SDoLancaster It's time to be counted. Here are 8 things to know about the 2020 census lancasteronline.com/news/localit... via @lancasteronline	Twitter Retweet	2020-03-03
TWITTER_8_1234875	curran_marlene	USATODAY	RT @USATODAY Questions about the 2020 Census? Here's an early look at the once-a-decade questionnaire. Story from @uscensusbureau	Twitter Retweet	2020-03-03
TWITTER_4_1234866	hills_prince		Dear Uhuru Kenyatta @StateHouseKenya now that we had census, why can't you deposit in our accounts each a million bob so that kujipange na maish	Twitter Mention	2020-03-03
TWITTER_8_1235187	janorganizer	staceyabrams	RT @staceyabrams Today is Super Tuesday + 10 days before #2020Census starts. At stake: \$1.5T/year, 435 Cong. seats & 99 state leg chambers that	Twitter Retweet	2020-03-04
TWITTER_4_1235383	openletterbot		/ Support Robin in Chicago, IL, by signing "Make Fraudulent Census Forms a Crime" and I'll deliver a copy to your officials too: twitter.com/messages/h	Twitter Mention	2020-03-04
TWITTER_8_1235296	korlaportopez	mitrajpratal	RT @mitrajpratal We reached our Ramsey County census hiring goals because YOU helped organize our community. #WeCountMN #Census2020 #E	Twitter Retweet	2020-03-04
TWITTER_2_1235278	DudesJoe		Health care plan for a family of four costs on average \$28,166 in 2018.	Twitter Update	2020-03-04
TWITTER_8_1235363	CwiReilly	ThatPeterGannon	RT @ThatPeterGannon United Way is committed to ensuring that everyone who should be counted is counted in the #2020Census. A more accurate oc	Twitter Retweet	2020-03-04
TWITTER_8_1235656	alberten	Mikaelisac	RT @Mikaelisac good eye from @JuddLegum, who spotted misinformation from trump campaign around census. it's an exercise in determining FB's lie	Twitter Retweet	2020-03-05
TWITTER_8_1235706	vangPlayground	BBCWorld	RT @BBCWorld Facebook removes 'deceptive' Trump census ads bbc.in/38t9620	Twitter Retweet	2020-03-05
TWITTER_8_1235576	sekitfangri	JuddLegum	RT @JuddLegum 1. In December, Facebook pledged to remove ANY content with misleading information about how to participate in the Census.	Twitter Retweet	2020-03-05
TWITTER_8_1235671	SybilQuave1	kaitlancollins	RT @kaitlancollins Facebook said Thursday it will remove some ads run by President Trump's re-election campaign that it says ran afoul of its policies b	Twitter Retweet	2020-03-05
TWITTER_8_1235645	CapMorgenthau	Rep.Jeffries	RT @Rep.Jeffries The so-called President and his sycophants are apparently circulating a fake census form on Facebook.	Twitter Retweet	2020-03-05
TWITTER_8_1235910	KathInCal	JuddLegum	RT @JuddLegum 1. In December, Facebook pledged to remove ANY content with misleading information about how to participate in the Census.	Twitter Retweet	2020-03-06

Tweet sample showing retweets

"How groups are feathering illegals into the census to create more Democrat seats in Congress." There, fixed it for you. apple.news/AL2pHcFrMR6Kwz...

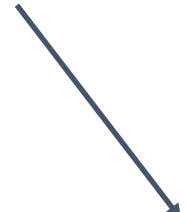
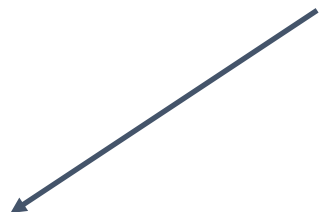
# How to format: Retweets vs **unique** tweets

In this alternate approach, @RealJamesWoods retweet shows up only once, but with the number of total retweets

Number of retweet	Retweet content	Original user	Day
1	Texas town's homeless population count misses "hidden" victims A yearly census-like count of the nation's nrv.ly/FUyQZD	stand_4_america	2020-03-01
1	Thank you to @ClownMarkets on 12th St/Ave C for letting our community know about the census and thanks to @GOLESNYC MoRUSNYC	MoRUSNYC	2020-03-01
5499	How groups are feathering illegals into the census to create more Democratic seats in Congress. Here, feed it for you. apple.news/AL2pHoFtMR6Kwz	RealJamesWoods	2020-03-01
79	The #2020Census will determine funding for:	HouseDemocrat	2020-03-02
1	We are partnering with DeKalb County Census team to host a Carnival/Human Services Expo at South DeKalb Mall on Saturday, March 7th from 11 a.m.	DSTDAC	2020-03-02
2	@SFbrigade @OpenOakland @codeforanajose RSVP here: sf.gov/events/march-7...	DataSF	2020-03-02
9	NEW: We don't know yet if @uscensusbureau has met its national recruiting goal of 2.7 million applicants for #2020Census jobs, but number of paid ten	hanslowang	2020-03-03
1	How does this keep happening? What the fuck is the point of the census if you people cant hire the correct amount of workers to make the voting proces	ACEXINFINITE	2020-03-03
2	How can Census data be used to increase economic opportunities in communities of color throughout the state? @blackfutureslab #askusaboutthecens	StepUpLA	2020-03-03
7	What's the 2020 Census and how does it affect Indian Country?	IndianCountry	2020-03-04
3	Make sure to participate in the 2020 census and to accurately report all children. All information provided to the U.S. Census Bureau is protected by law	NYBCF	2020-03-04
241	You know who doesn't want you to complete the census?	CarolynBMalone	2020-03-04
2	Taking the Census supports our HEB ISD community. How? Census data affects \$675 billion of funding for school lunches, mental health services, high	hebid	2020-03-05
2	You may notice census takers in our community this year.	upperwohlan	2020-03-05
1731	#MarkZuckerberg—you said Facebook would ban lies about the census. But now you're running +1000 Trump ads tricking people about the census!	SachaBaronCoh	2020-03-05
1283	There is simply no excuse for @Facebook approving thousands of Trump campaign ads that misled users and spread misinformation related to the Cen	SenKamalahari	2020-03-06
25	Advice that David Card gave me when I was a graduate student. "Your comparative advantage as a grad student is finding previously unused datasets"	Devin_G_Pope	2020-03-06
672	Shape your future. Like this tweet to receive a reminder when the #2020Census is live. https://t.co/MIT1T15xIW	uscensusbureau	2020-03-06
9312	"The census is a once-a-decade population count that determines how \$1.5 trillion in federal spending is allocated, and also how many congressional se	RealJamesWoods	2020-03-07
61	THIS ONE OF MANY REASONS WHY SHE NEEDS TO BE REMOVED FROM OFFICE: Open Borders Activist Ocasio Cortez Urges Illegal Aliens to Take Part in 2020 US	gatorfun1	2020-03-07
429	Open Borders Activist Ocasio Cortez Urges Illegal Aliens to Take Part in 2020 US Census (VIDEO) thegatewaypundit.com/2020/03/open-b... via @gate	gatewaypundit	2020-03-07
1	"Clarion County is falling pretty far behind as far as needing enumerators, to have Census workers. There are only a few more days to apply." Whiting s	exploreciarion	2020-03-08
27	How do you feel about the US Census...??	TheyCallMeTom	2020-03-08
27	@jd155551 @SenSchumer @SpeakerPelosi Average Income of Families and Persons in the United States in 1969 was \$5,400 -U.S. Census Bureau	FIGHT_2_KAG	2020-03-08
5	We are a proud partner of the @uscensusbureau!	SAPHAnfo	2020-03-09
42	The @CountVonCount loves #Census2020 because he knows #EveryChildCounts! Pediatricians can help children in their communities receive critical s	AmerAcadPeds	2020-03-09
1	I am being actively recruited to be a census taker on Instagram and I just want to know what makes me such a desirable candidate for this job?	Spirballn_Here	2020-03-09
5	@KristenClarkeJD and @LawyersComm have set up a hotline for the #2020Census.	NatUrbanLeague	2020-03-10
1	@RuedGestures, thank you for requesting to receive a reminder to complete the 2020 Census.	uscensusbureau	2020-03-10
1	Census Bureau site goes live as counting begins in earnest	NewsandRecord	2020-03-10
46	ICE Chief Shuts Down Idea That Enforcement Would Stop During Census Counting dailycaller.com/2020/03/11/ice...	DailyCaller	2020-03-11
1	Respond your way to the 2020 Census and other neighborhood news!	BDOCenter	2020-03-11
2	Census safety: What to know about the process, your privacy and what residents are required to do ketv.com/article/census...	KETV	2020-03-11

Tweet sample showing unique tweets

# Selecting samples: whole corpus or keyword (citizen)



Number of retweet	message_cleaned	OriginalAuthor	Day
520	"How groups are feathering illegals into the census to create more Democra	RealJamesWoo	2020-03-01
11	Caribbean-Americans and other census respondents will now be able to rec	NYDailyNews	2020-03-01
1	census 조사, 통계	vocaforme	2020-03-01
8	According to the National Crime Survey administered by the Bureau of the C	2AWisdom	2020-03-02
9	It's #2020Census Statistics in Schools Week: Everyone Counts! APS stude	APSVirginia	2020-03-02
111	Hi, ionelts! We conduct a census, which taken for the 1st place. The purpos	NuSpiaggiaMu	2020-03-02
11	It's true! I'm doing a webchat with mumsnet tomorrow - focusing on the Fem	K_IngalaSmith	2020-03-03
1	Read this joyful geek's commitment to making sense of data legitimacy by fc	crislopezg	2020-03-03
1	Why are there census ads I thought we didn't have a choice with that shit	CampCollins8	2020-03-03
6	Do you know why it's important for people with disabilities and their families	TheArcUS	2020-03-04
2	The 2020 Census is right around the corner! It is critical that we make every	NorthHempstea	2020-03-04
2	<a href="#">Join us THIS SATURDAY @ Villa-Parke Community Center for JUMP INTO</a>	PasadenaGov	2020-03-04
45	As the 2020 census ramps up, you'll likely see misinformation — and some	NPR	2020-03-05
394	Facebook is taking down Trump ads that link to a fake "census" form becau	B52Malmet	2020-03-05
1	@JBJ_Marketing is conducting a media round table with @ABlinfluence & N	johnsonwillisj	2020-03-05
92	That won't pan out well for them	Inevitable_ET	2020-03-06
7	Facebook will take down Trump campaign posts that look like official census	verge	2020-03-06
1	The #2020Census is more than a population count. It's an opportunity to shi	fanwoodboroug	2020-03-06
75	Alexandria Ocasio-Cortez calls on illegal immigrants to fill out 2020 census	MariaBonanno9	2020-03-07
1	Aeon's Resident Connections team organized a presentation with Frank Sar	AeonMN	2020-03-07
98	AOC Encourages Illegals To Participate In 2020 Census   Zero Hedge	HyltonRobin	2020-03-07
1298	"The census is a once-a-decade population count that determines how \$1.5	RealJamesWoo	2020-03-08
46	Trump's latest corruption of democracy: The census will be made bogus for	psychdr100	2020-03-08
103	More than 1.5 million people of color, including black, Hispanic, and Native	ComcastNewsm	2020-03-08
10	What you need to know about the 2020 census buff.ly/2Q1ifWO with @NP hari		2020-03-09
2	Support Karen in Kensington, Calif. by signing "We need Plan B for votin	openletterbot	2020-03-09
2	Today's Weekly includes information about Kindergarten registration, Middle	Skokie735	2020-03-09
1	Census Bureau site goes live as counting begins in earnest trib.al/vRc4P7n	eurekaTS	2020-03-10
2	It's time for #Census2020! The census determines political representation. F	FIAnational	2020-03-10
15	Still think data can't be biased?	TalkPowerTv	2020-03-10

Number of retweet	message_cleaned	OriginalAuthor	Day
1	WHY DO DEMOCRATS SUPPORT SANCTUARY CITIES ? ITS	TomWestlock	2020-03-01
2	According to U.S. census data, there were 18,871,831 black American	JimW_in_NM	2020-03-01
1	@FutureSpecOps @RealJamesWoods @lesagre66751588 @FoxNews	adalifts	2020-03-01
5	Did you know the Census determines how much funding and services our	okcareertech	2020-03-02
23	NEW: The #2020Census will NOT include the now-blocked citizenship	hansilowang	2020-03-02
1	i want to work as a census taker rly bad but im not a citizen and theres	mutedbeatings	2020-03-02
1	@cavaticat Normalized to citizen population in each category (Census	JLShannonhous	2020-03-03
20	Still have questions about the #2020Census? We have answers.	NYCImmigrants	2020-03-03
6	Worth noting AG Holder was also one of the people behind the effort to	bhweingarten	2020-03-03
4	Did you know this week is @uscensusbureau's "Statistics in Schools	lpornoy	2020-03-04
4	"This report is the first to analyze Census Bureau data documenting recent	autselfadvocacy	2020-03-04
1	I have always thought this would be a great idea. I also supported a census	2Real2Day	2020-03-04
461	<a href="#">Today, people spoke and Facebook backed down!</a>	SachaBaronCoh	2020-03-05
233	Here are some facts to help debunk common misconceptions about the	NPR	2020-03-05
176	Hmm, I'm starting to suspect that maybe the Trump administration wasn't	Susan_Henness	2020-03-05
75	Today, people spoke and Facebook backed down!	SachaBaronCoh	2020-03-06
1	@Grace4NY I filled out a census form last fall that I received in the mail.	JohnBar866082	2020-03-06
14	"The Trump administration has already sowed significant confusion about	BrennanCenter	2020-03-06
1	@helloitsthao This is good news! Other states must take notice now &	Erimiana	2020-03-07
879	AOC Urges Illegal Immigrants to Fill Out Census Form🙏	no_silenced	2020-03-07
1	ILLEGAL ALIENS SHOULD NOT be counted in the Census. It's original	jock34_us2	2020-03-07
7	While you are all running around looking for hand sanitizer, remember that	Dagny_Galt	2020-03-08
103	For crying out loud. @AOC is supposed to be a #Lawmaker. All she's	cindievaccaro	2020-03-08
2	@uscensusbureau Can I protest as a citizen which you don't care about	HermitPJ	2020-03-08
86	The Trump administration attempted to put a citizenship question on the	mgrant76308	2020-03-09
18	Every service member deserves to be counted, and the fact that they WILL	RandPaul	2020-03-09
17	The Trump administration attempted to put a citizenship question on the	mgrant76308	2020-03-09
3	The Trump administration attempted to put a citizenship question on the	mgrant76308	2020-03-10
4	Everyone counts in #Census2020! No matter where you live, how old you	GaWomenStand	2020-03-10
1	What is the #2020census?	SI_Census	2020-03-10

# Overlap in insights from tweets vs focus groups?

A case study: We compared data on barriers to response from 2020 CBAMS Focus Groups groups to a sample of 300 tweets from March 1, 2020 to April 20, 2020.

2020 CBAMS Focus Groups ten barrier-to-response themes based on focus group transcripts

1. Confidentiality and privacy concerns\*
2. Lack of knowledge or understanding of purpose
3. Apathy toward the census and lack of efficacy
4. Inclusion of citizenship question
5. Fear of repercussions\*
6. Online data security concerns\*
7. Distrust of government
8. Displacement\*
9. Frauds and scams
10. Language barriers

*\*barriers that were hypothesized prior to the beginning of the 2020 CBAMS Focus Group research*



# Example: topic overlap between focus group statements and tweets

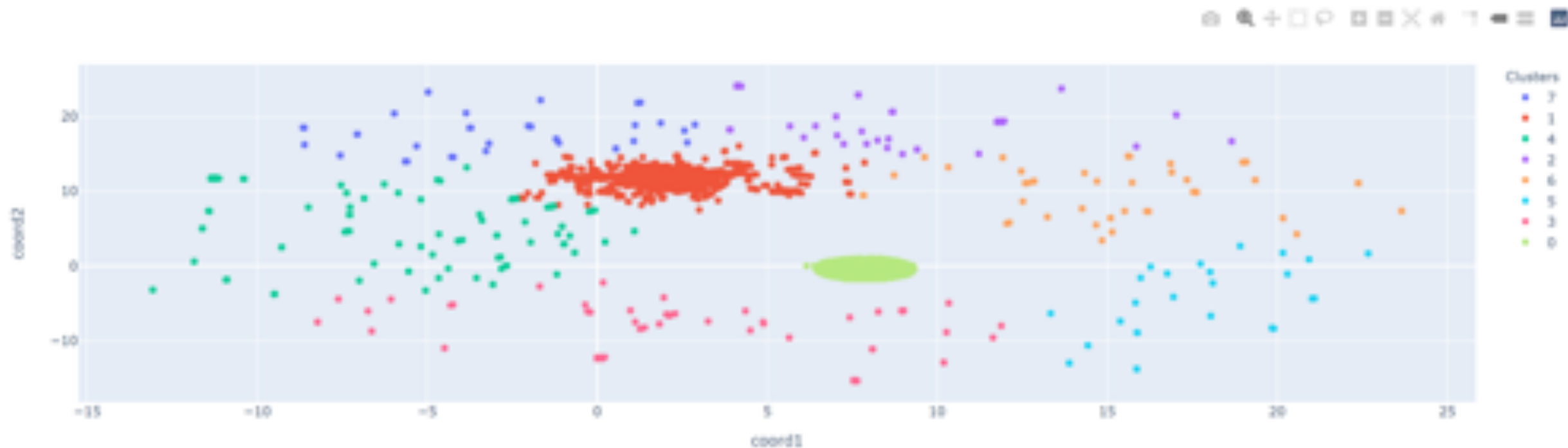
- Focus group: “[Census information is shared] with the entire government. With everyone in the government...police, immigration, hospitals, everything, everything, everything. Everything is connected.”
- Tweets:
  - "I might be on r/conspiracy too much but I'm not trying to give my info to the census. They magically have all my info if I owe \$ though."
  - "The Census is for everyone living in the U.S. It won't ask if you're a citizen, and your information will not be shared with other government agencies. Find out more at 2020Census.gov. #2020Census #CountAllKids <https://t.co/IYnoqtp5Ph>"
  - 'RT @GovPritzker I want to remind everyone that there is NO citizenship question on the Census. \r\n\r\nYou will not be asked your immigration status and your information will not be shared with anyone.'
  - '@ezduzit63 @oracle\_ed @BIZPACReview And on the 2020 U.S. Census none of the info can be given to @ICEgov . Intellectual dishonesty! Census info is used for services which lawful citizens depend upon.'
- Note: topic overlap doesn't mean stance or intention are the same...

# Example: topics in tweets that were not prominent in focus groups

- Concerns about Census data uses for congressional apportionment
  - RT @RealJamesWoods “How groups are feathering illegals into the census to create more Democrat seats in Congress.” There, fixed it for you.  
[apple.news/AL2pHcFrMR6Kwz...](https://apple.news/AL2pHcFrMR6Kwz...)
    - (retweeted 5000+ times)
- → Potential for social media posts to uncover opinions from people who might never participate in a focus group or respond to a Census survey

# Explore Tweet Clusters

The plot below contains 26311 total tweets.



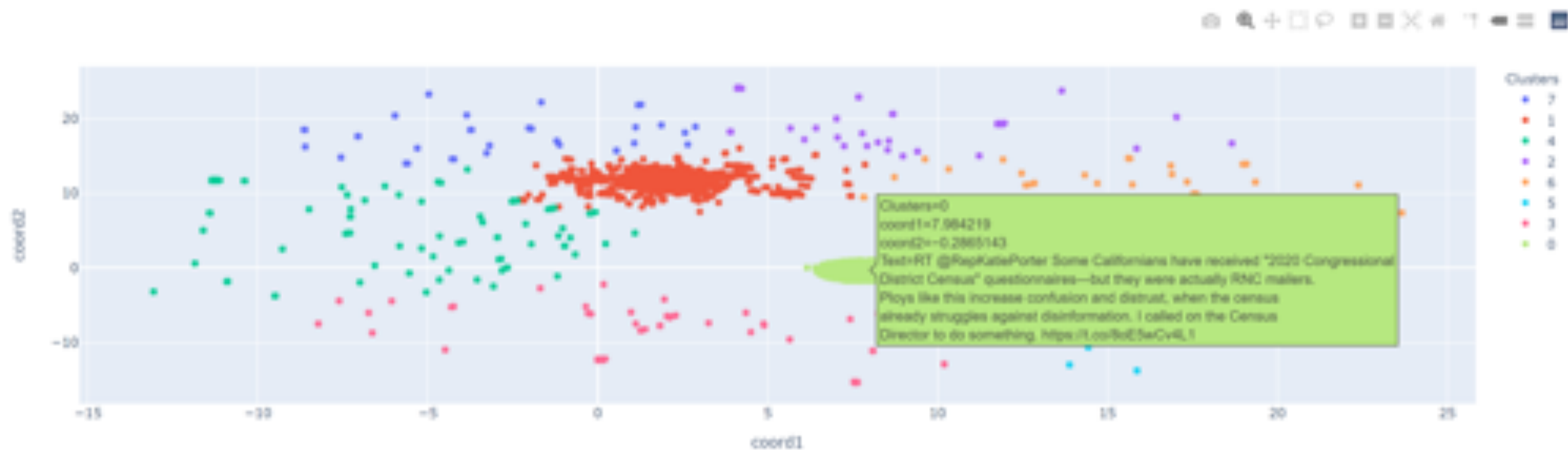
Most common words in each cluster:

Cluster	Proportion of Tweets	Number of Tweets	Top 5 Stemmed Words
2	0.007	983	cell good mail sun trust
0	0.557	14644	but californian questionnaires reputationstruggl
3	0.040	1124	case julg mail probe trust



## Explore Tweet Clusters

The plot below contains 26311 total tweets.



# What have we learned and hypothesized so far?

- There is some overlap between what tweets can show and what focus group results using same research question can show, but there can also be unique insights from social media
- There are certainly limits to how many tweets an analyst can handle at a time, and automated support will be useful
- How tweets are sampled and displayed seems certain to affect analyst experience, and likely insights generated
- Analysts may vary in their strategies, e.g. how much they are influenced by frequent tweets vs. “outlier” tweets
- Topical focus of sample of tweets makes a big difference
  - hard to keep track of what’s happening when the topics range too widely
- Insights are likely to differ depending on (1) whether there is a predefined (as opposed to emergent) research question and (2) what the question is

# Concluding proposals

- Using social media data to improve social research in connection with survey data continues to be promising
- But the best uses are likely to look different and be more complicated than early promising findings suggest
  - Being more thoughtful about content of and processes that generate both social media posts and survey responses will be needed
  - More sophisticated analysis tools—which require substantial infrastructure and advanced researcher skills—will certainly be needed
    - And even the most advanced text-analytic tools aren't perfect (and may never be)
- Alternate ways of using social media data, e.g. for research purposes like those of focus groups, are worth exploring for finding out what members of the public are thinking
  - perhaps especially members of “publics” whose opinions aren't always represented in representative sample surveys