Sweet Tweets! Exploring a New Method for Probability-based Twitter Sampling

Trent D. Buskirk, Novak Family Distinguished Professor of Data Science

Bowling Green State University

SODA, 5/18/2021

Acknowledgements…

- I’d like to acknowledge the members of the Buskirk BGSU Covid Team who have contributed TONS to this work!
  - Brian Blakely, BGSU
  - Ravinder Singh, BGSU
  - Youzhi Yu, BGSU
  - Dr. Herb McGrath, BGSU
  - Dr. Adam Eck, Oberlin College

The Rise of Social Media Data…

- Pew (2018) estimates that about 70% of US adults use some social media. And not all users "use" at the same rates with about 22% of the U.S. population using Twitter.

- A 2017 survey of 300 medium to large businesses conducted by Digital Marketing at CLUTCH found that (Anyan, 2017):

79% use Twitter for social listening activities

25% use social listening to improve products

21% use social listening to improve customer experience

Source: https://bit.ly/clutchsurvey2017

Sampling Twitter…

- There is limited work on how to create samples from Twitter that have desirable properties of representation and replication…
  - Hino and Fahey (2019) illustrate a population based method for for creating an archive of Twitter data that is highly representative of the USER population.
    - Sample is generated from user-based samples; May have issues temporally because of Twitters 3200 public tweet limits.
  - Berzofsky et al. (2018) also applied a similar approach taking a systematic sample of Twitter (User) IDs and built a sampling frame based on gathering Twitter User's with corresponding ID numbers.

Our Approach To Twitter Sampling

- In our method we focus on sampling Tweets from a particular day and geographic location, rather than users.

- In our preliminary work we discovered a connection between time and Tweet IDs that were previously discussed in by Kergl et al. (2014) and Pfeffer and colleagues (2018).

  - The 64-bit representation of a tweet id contains important time and processing information

  - Current publically available version of the Twitter search API (1.0) cannot sample at particular time points (only days) – the new beta version of the Search API (2.0) can incorporate a time window, but only up to the second, not millisecond and can’t filter by geography.

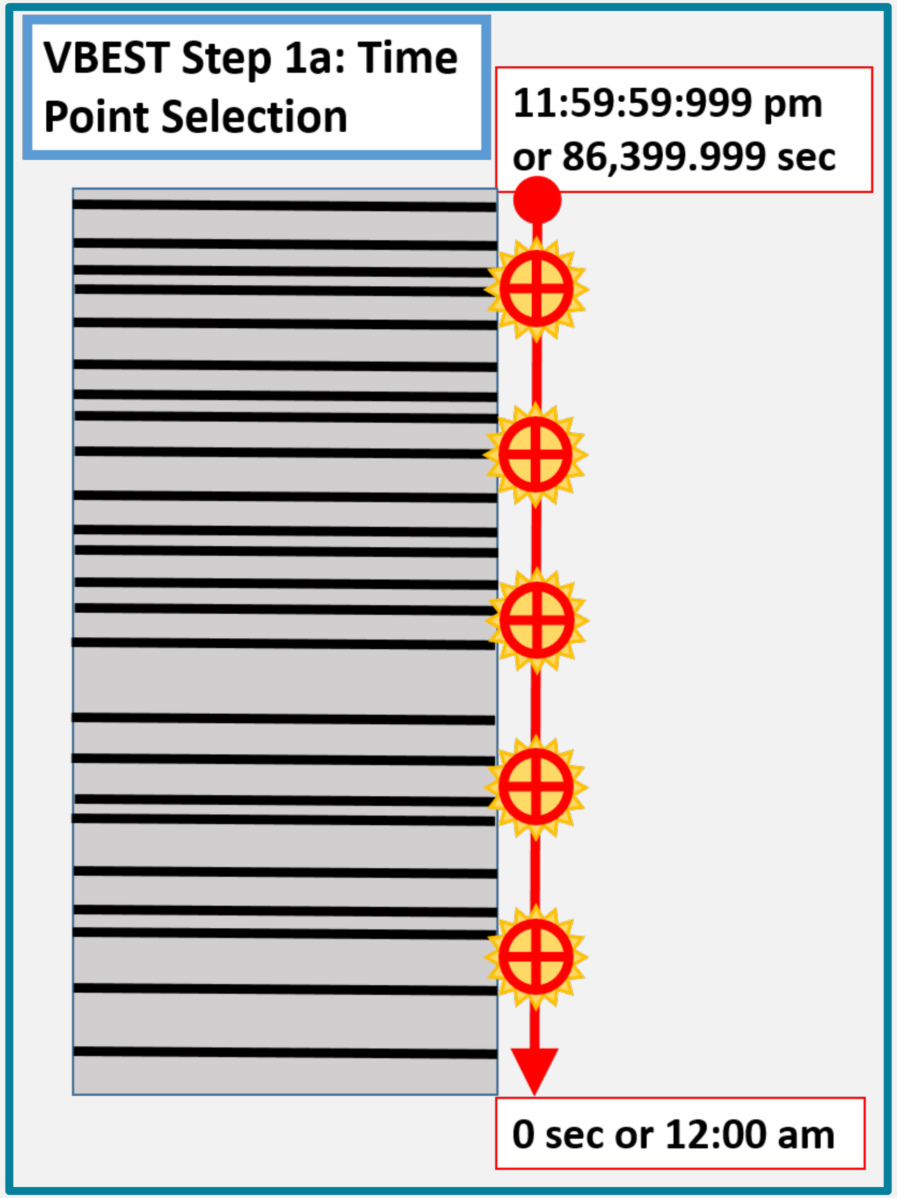
Our Approach to Twitter Sampling

- Our goal is to create a sample of tweets that reflects the volume of total tweets over a given day from a given region.
  - By creating synthetic tweet ids based on times of interest, we can restrict the twitter samples to particular time windows.
- By incorporating time into the sampling method we can be sure to traverse the entire twitter distribution for a given day.
- Twitter volumes (e.g. tweets per second over time) are converted to corresponding time intervals:
  - lower volume -> wider time intervals; Higher volume -> narrower intervals
- We will sample the time intervals (what we call Tweet PSUs).
  - Velocity Based Estimates for Sampling Tweets (VBEST)

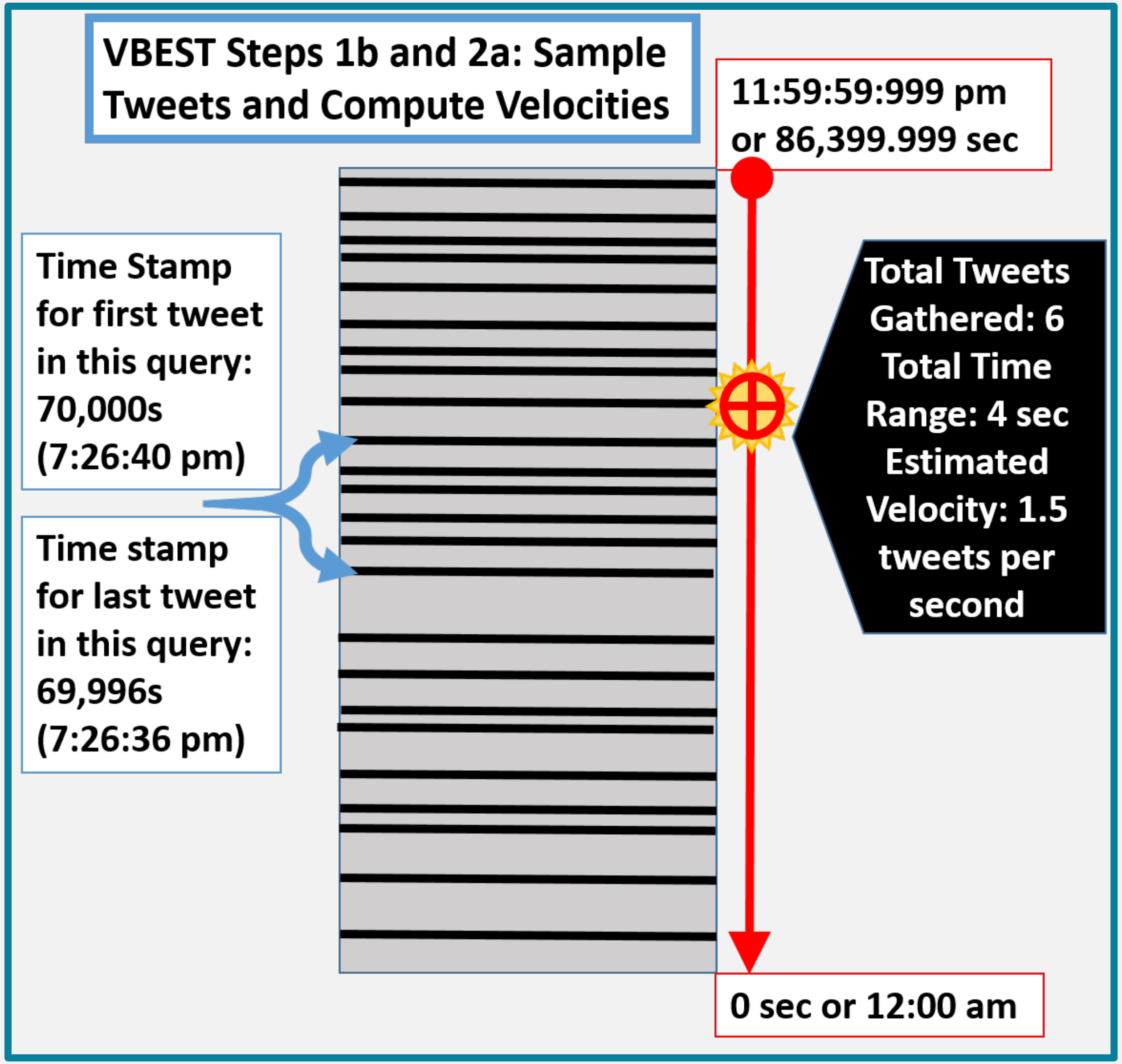
The VBEST Algorithm…

- Step 1: Select an initial tweets from uniform time points across a given day of interest
  - 1a: Time Point Selection: Select a systematic sample of 48 time points from across a given day/region and convert these time points into synthetic tweet ids to be used in Twitter Search API query calls
  - 1b: Initial Tweet Sample: Submit 2 queries (of 100 tweets per) for each of the sampled time points.
- Step 2: Estimate Twitter Volume over a given day/location
  - 2a: Compute Twitter velocities based on initial tweet samples
  - 2b: Estimate a Twitter Velocity Curve using locally estimated scatterplot smoothing (LOESS) with polynomial regression fitting.

VBEST Algorithm Steps 1 and 2…



VBEST Step 1a: Time Point Selection



VBEST Steps 1b and 2a: Sample Tweets and Compute Velocities

Time Stamp for first tweet in this query: 70,000s (7:26:40 pm)

Time stamp for last tweet in this query: 69,996s (7:26:36 pm)

Total Tweets Gathered: 6 Total Time Range: 4 sec Estimated Velocity: 1.5 tweets per second

Our Approach to Twitter Sampling

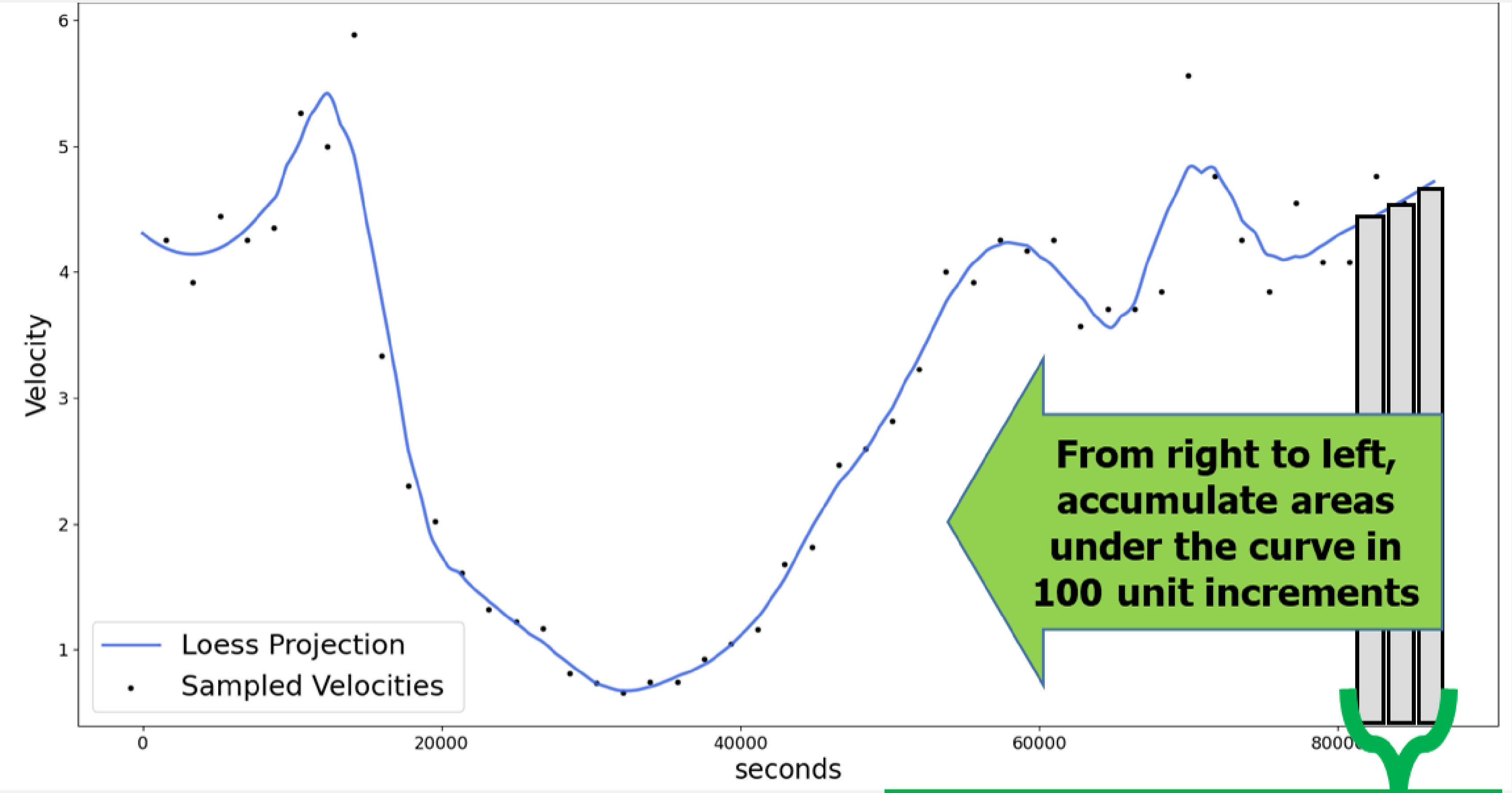
Step 3: Create Sampling Frame of Primary Sampling Units (Tweet PSUs)

- We estimate the area under the Twitter Velocity Curve using a 3-point trapezoidal rule over a grid of 86,400 one-second points.
- We estimate PSU boundaries to be those time points between which we estimate the twitter volume to be 100 (or just exceed it).

Step 4: Take VBEST Sample

- 4a. A 25-50% Systematic Sample of Tweet PSUs is taken.
- 4b. The upper endpoint of each selected Tweet PSU is then converted to Synthetic Tweet ID and submitted as a query for the Twitter Search API. The collection of resulting tweets from processing all of the queries forms the VBEST twitter sample.

VBEST Step 3: Create Tweet PSUs

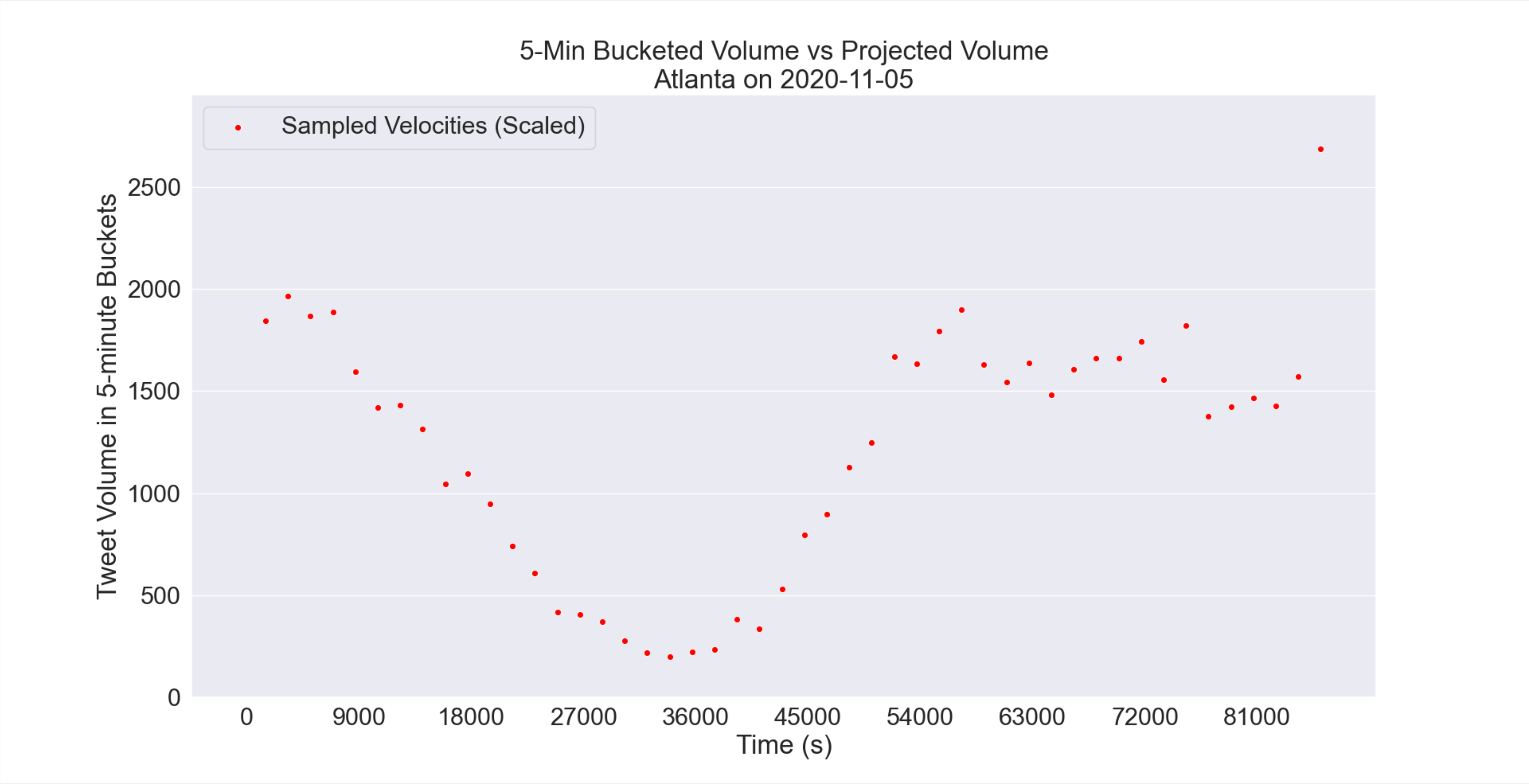


From right to left, accumulate areas under the curve in 100 unit increments

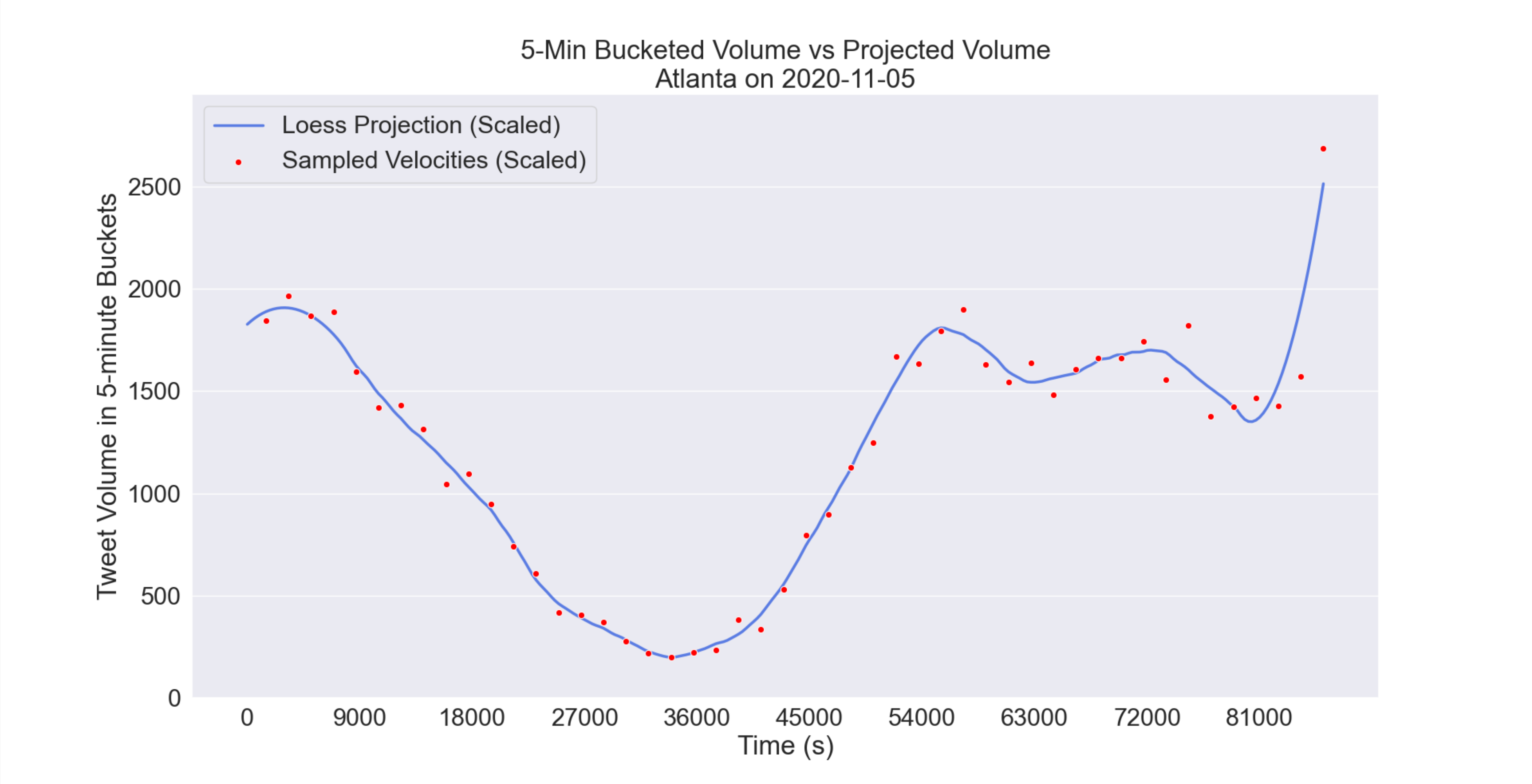
Expected number of tweets here is 100 so this interval forms first PSU (e.g. from 84000 to 86400 seconds).

VBEST Step 3

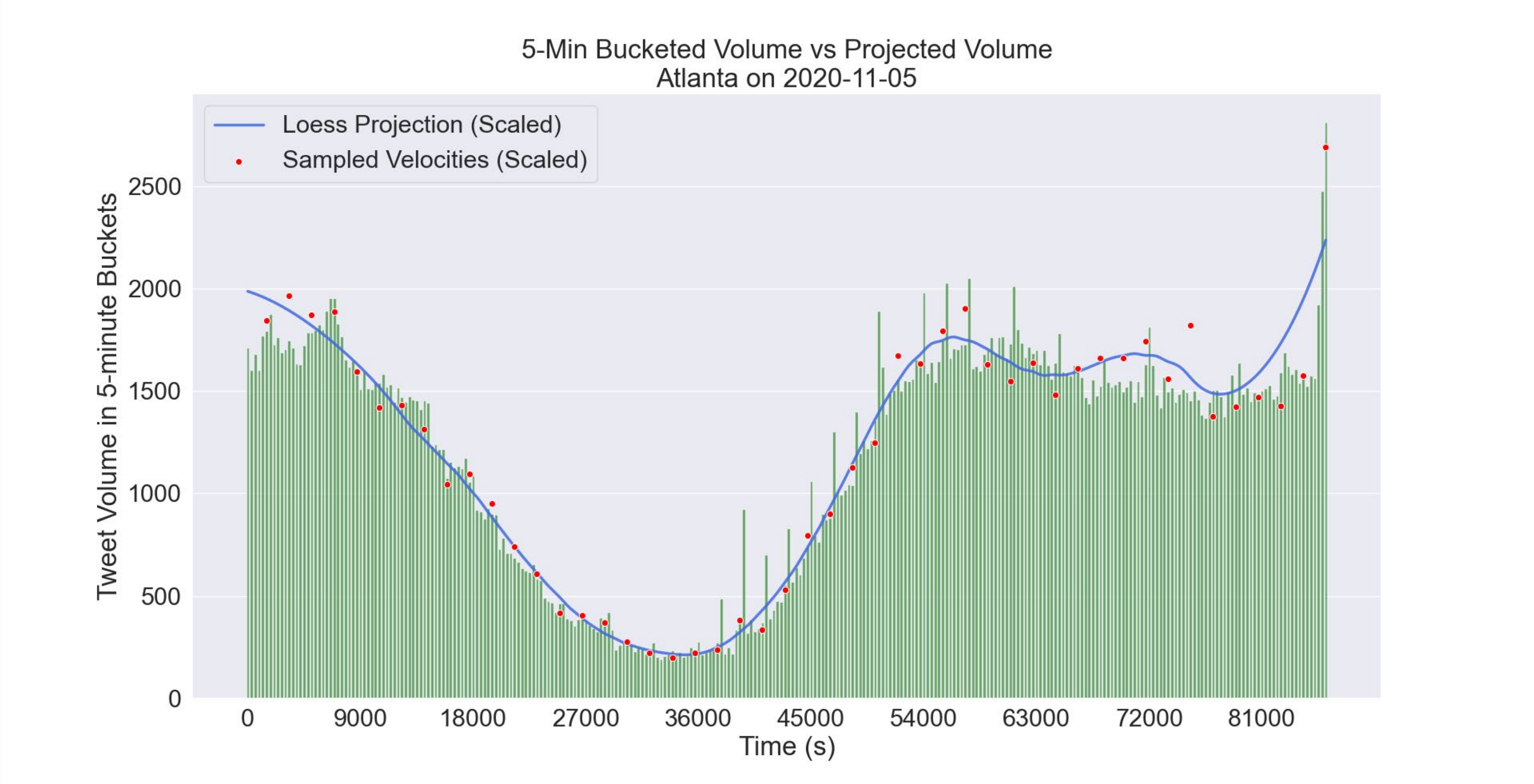
VBEST Steps 2 and 3 Illustration: Atlanta, Nov 5, 2020



VBEST Illustration: Atlanta, Nov 5, 2020



VBEST Illustration: Atlanta, Nov 5, 2020



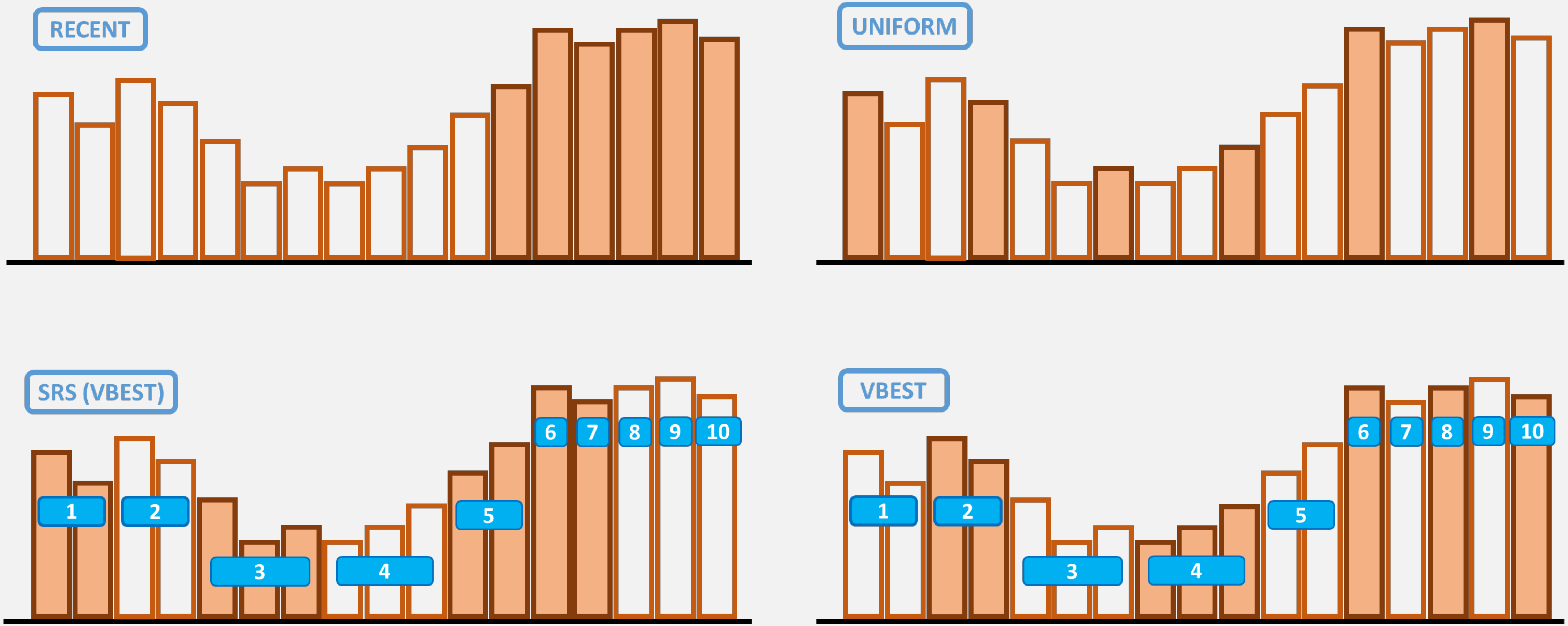
Experiment to Test the VBEST Algorithm

- Our experiment crossed two factors: Query Method and Sample Size (e.g. number of queries) for each of 38 days within 5 regions.

METHOD for accessing Tweets	Description
1. Popular	One of three methods for Twitter Search API that returns the most popular results in the query.
2. Mixed	the current default method for Twitter Search API that includes both popular and real time results.
3. Recent	an alternate option for Twitter Search API (tweets pulled from end of day towards beginning of day)
4. Uniform	Queries taken at uniformly position time points throughout the day
5. SRS (VBEST)	Simple random sample (without replacement) of Tweet PSUs
6. VBEST	Systematic random sample (without replacement, circular) of Tweet PSUs

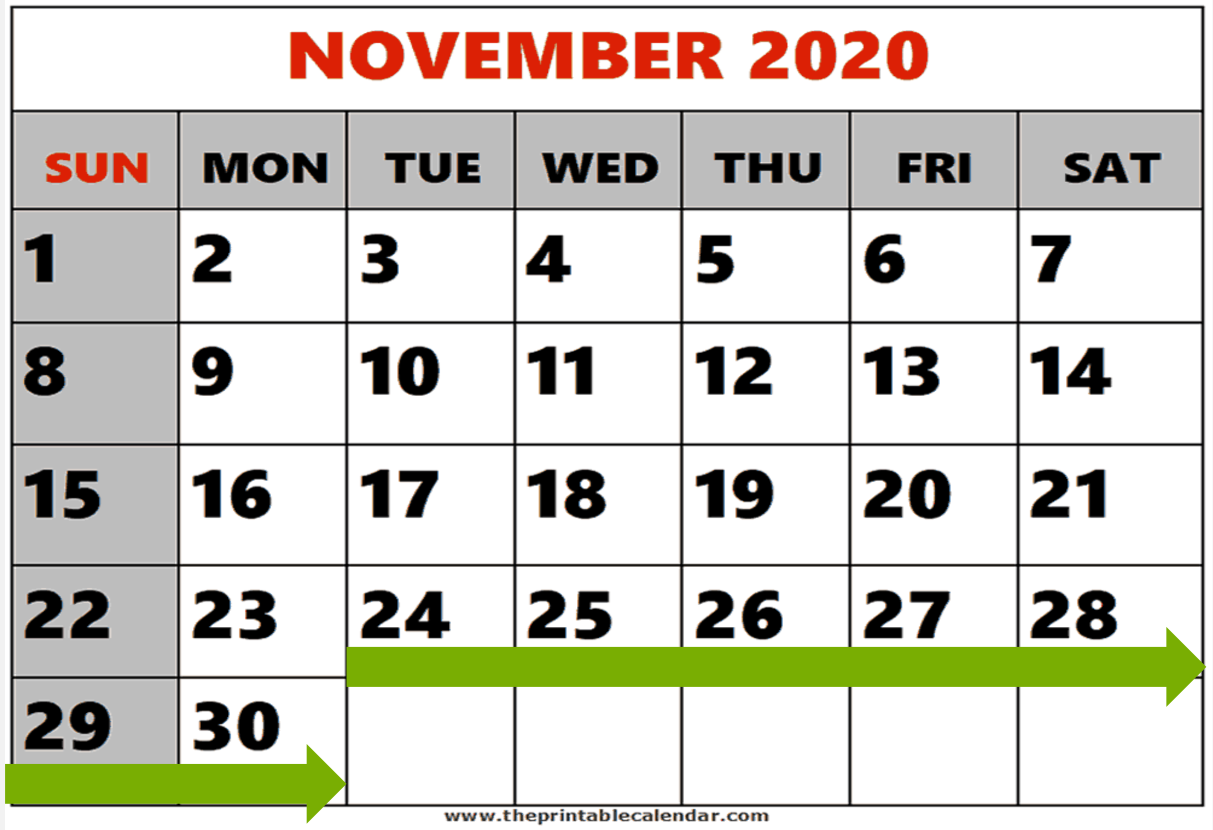
Sample Size
720 total queries for 1-4 96 + 624 for 5-6
540 total queries for 1-4 96+444 for methods 5-6
360 total queries for 1-4 96+264 for methods 5-6

Visual Depiction of Data Gathering Algorithms



Blue Box labels indicate different “Tweet PSUs”

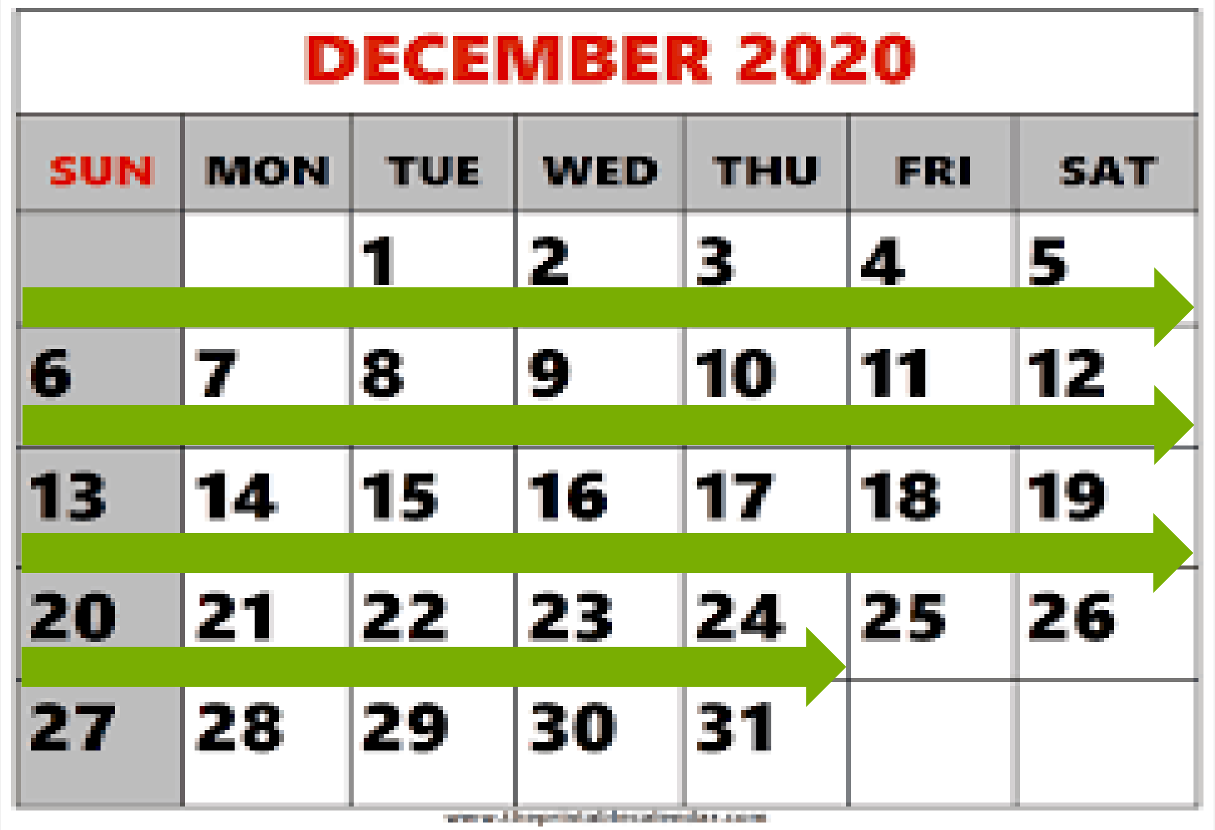
Field Test Period…



NOVEMBER 2020

SUN	MON	TUE	WED	THU	FRI	SAT
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30					

www.theprintablecalendar.com



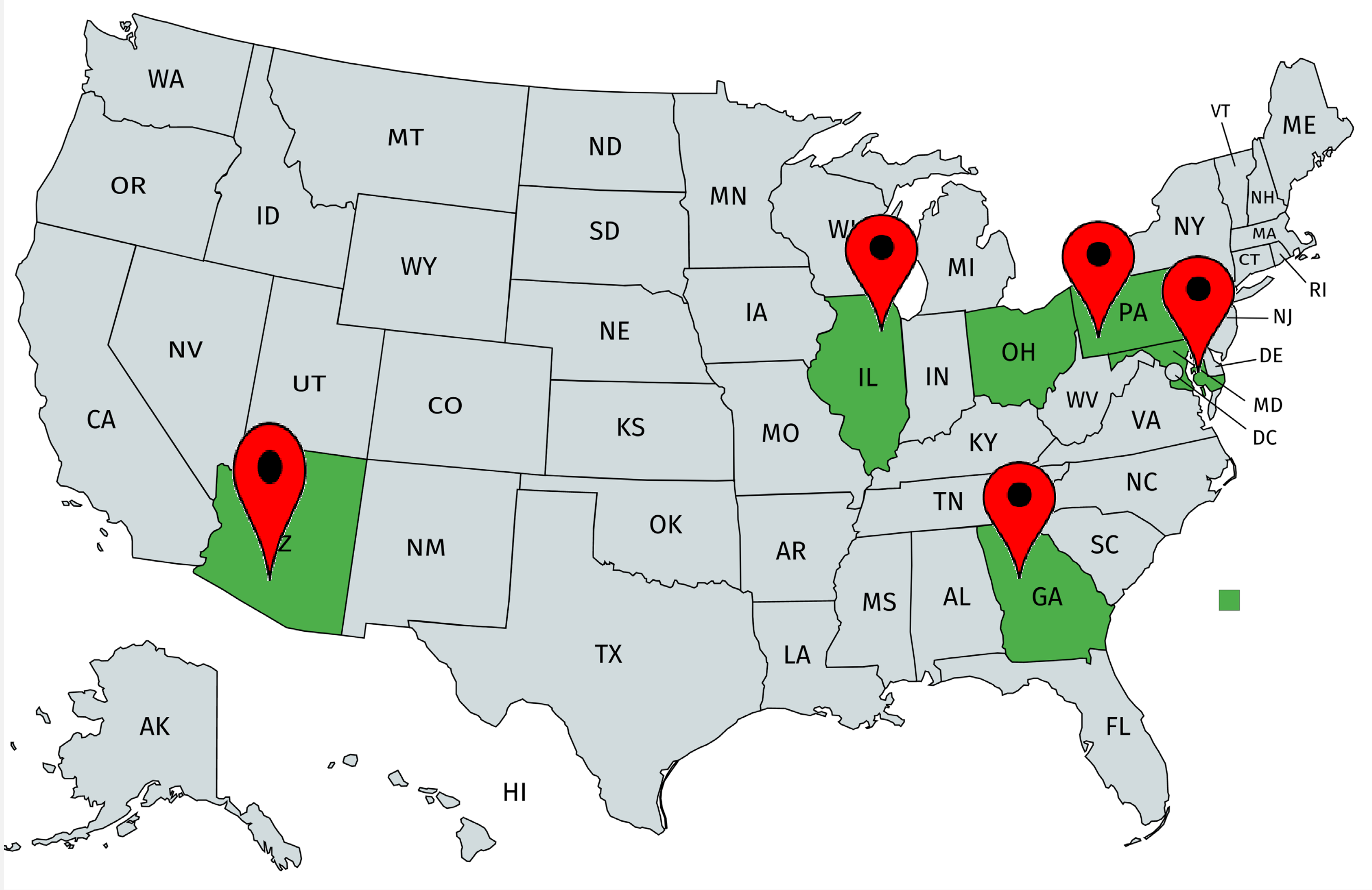
DECEMBER 2020

SUN	MON	TUE	WED	THU	FRI	SAT
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

www.theprintablecalendar.com

Collected 5 weeks' worth (38 days) of Twitter samples for our Experiment

Data Collection from 6 Medium to Large MSAs



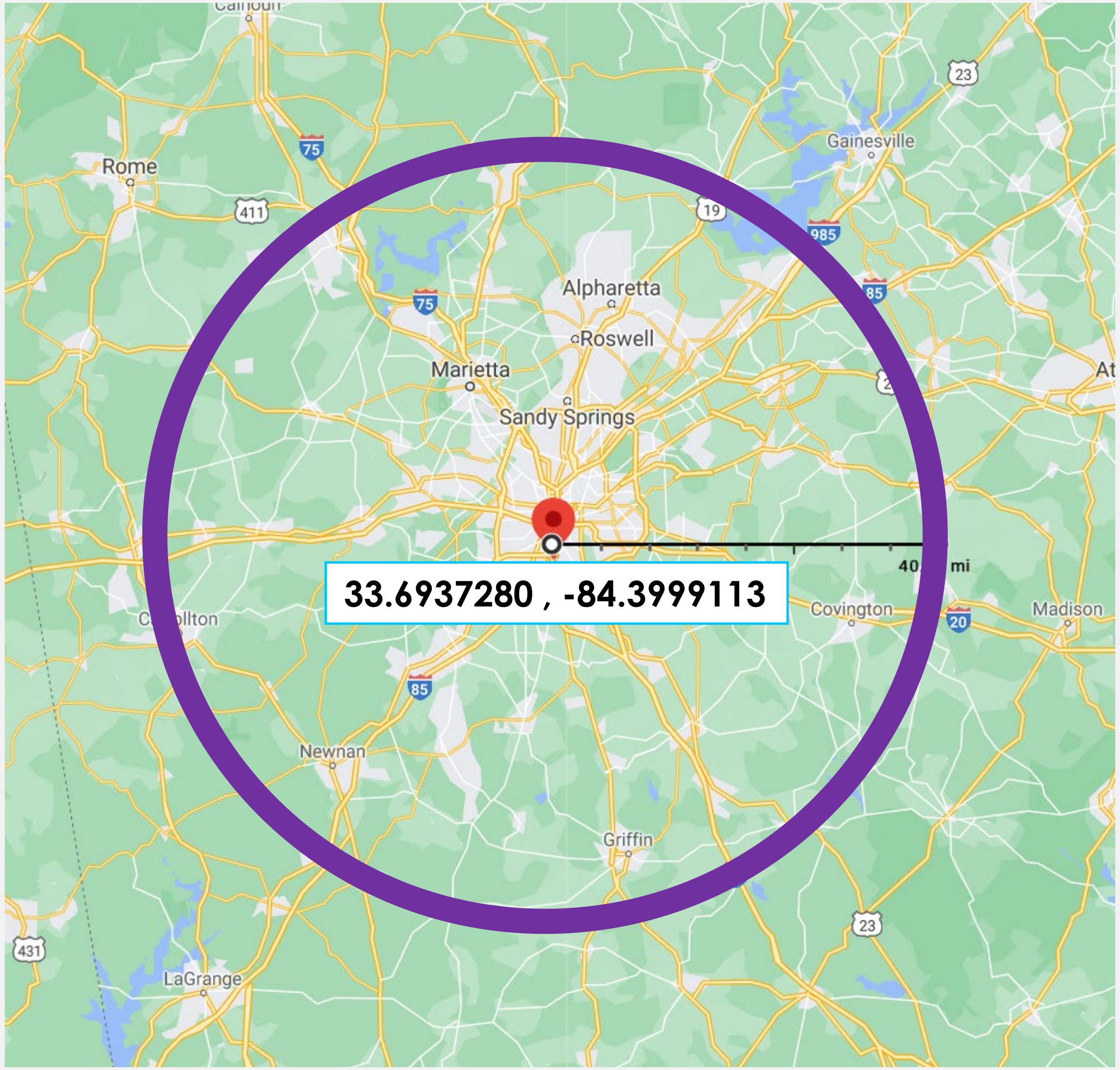
MSA REGIONS
Phoenix-Mesa-Chandler
Chicago-Naperville-Elgin
Columbus, OH
Pittsburgh, PA
Baltimore-Columbia-Towson
Atlanta-Sandy Springs-Alpharetta

Example of Geographical Search Parameters in Twitter Search API

- For Atlanta, we specify Lat/Long GPS coordinates (33.6937280 , -84.3999113) as the center and a radius of 40 miles.

- This area will allow us to capture the major cities that comprise the Atlanta-Sandy Springs-Alpharetta MSA

- Our radii for the 6 MSAs ranged from 16 miles for Baltimore to 53 miles for Phoenix



Primary Keyword Groupings Used in Filtering Twitter Sample

Keyword Group							
Covid	Social Distancing	Working	Masks	Sanitizing	General Virus	Symptoms	Treatment
covid	social distance	wfh	face mask (s)	hand sanitizer	virus	can't smell	ventilator
covid-19	social distancing	working from home	mask (s, ed)	disinfect	flu	no Smell	remdesivir
covid19	six feet apart	work from home	PPE	disinfectant	pandemic	can't taste	vaccine
covid test (ing)	6 ft apart	not working now	N95	lysol	sars	no taste	contact tracing
covid cases	6 feet apart	furlough	face cover (ing)	sanitize	pneumonia	cough	
coronavirus	hunker down	reopen	face shield	sanitizing	Fauci	fever	
rona	lockdown	reopening		sanitizer		chills	
cv19	quarantine	stimulus checks		hand wash		sore throat	
	quarantining	remote work		hand washing		asymptomatic	
		working remotely		bleach			
		unemployed		washing hands			

Evaluation Criteria for Our Method – Stage 2

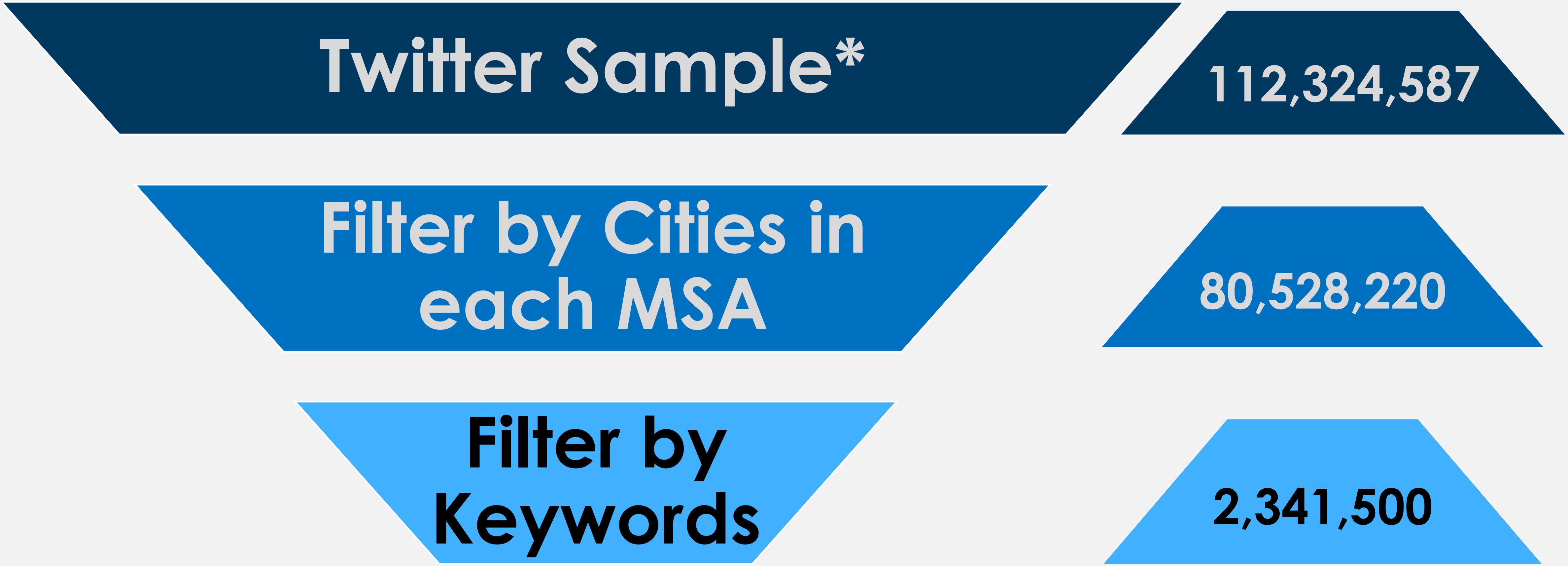
- We also evaluate our method in two ways based on the resulting samples selected

- For a given Sampling Method, S and Twitter Keyword Category, C, we estimate the performance of our sample using the Percent Relative Absolute Bias (PRAB) of the estimate ($\widehat{\theta}_{SC}$) of the true prevalence of that keyword within the reference Firehose Sample (θ_C) from our vendor.

$$\boldsymbol{PRAB}\left[\hat{\theta}_{SC}\right] = \mathbf{100} * \frac{\left|\widehat{\boldsymbol{\theta}}_{\boldsymbol{SC}} - \boldsymbol{\theta}_{\boldsymbol{C}}\right|}{\boldsymbol{\theta}_{\boldsymbol{C}}}$$

$$\boldsymbol{MeanPRAB}[\boldsymbol{S}] = \frac{\mathbf{1}}{\mathbf{8}} \sum_{\boldsymbol{C}} \boldsymbol{PRAB}\left[\hat{\theta}_{SC}\right]$$

Processing the Twitter Samples



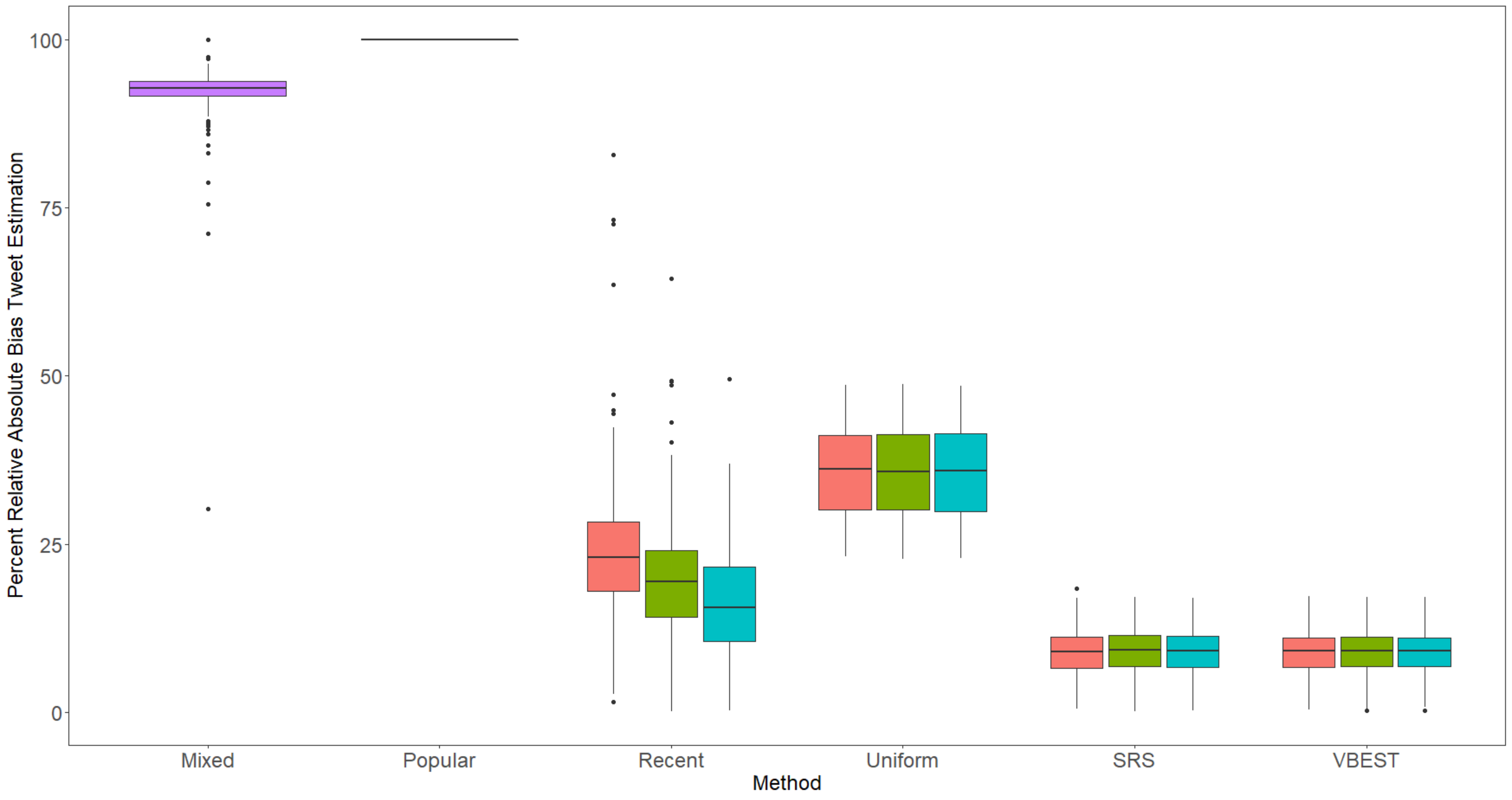
Twitter Sample* — 112,324,587

Filter by Cities in each MSA — 80,528,220

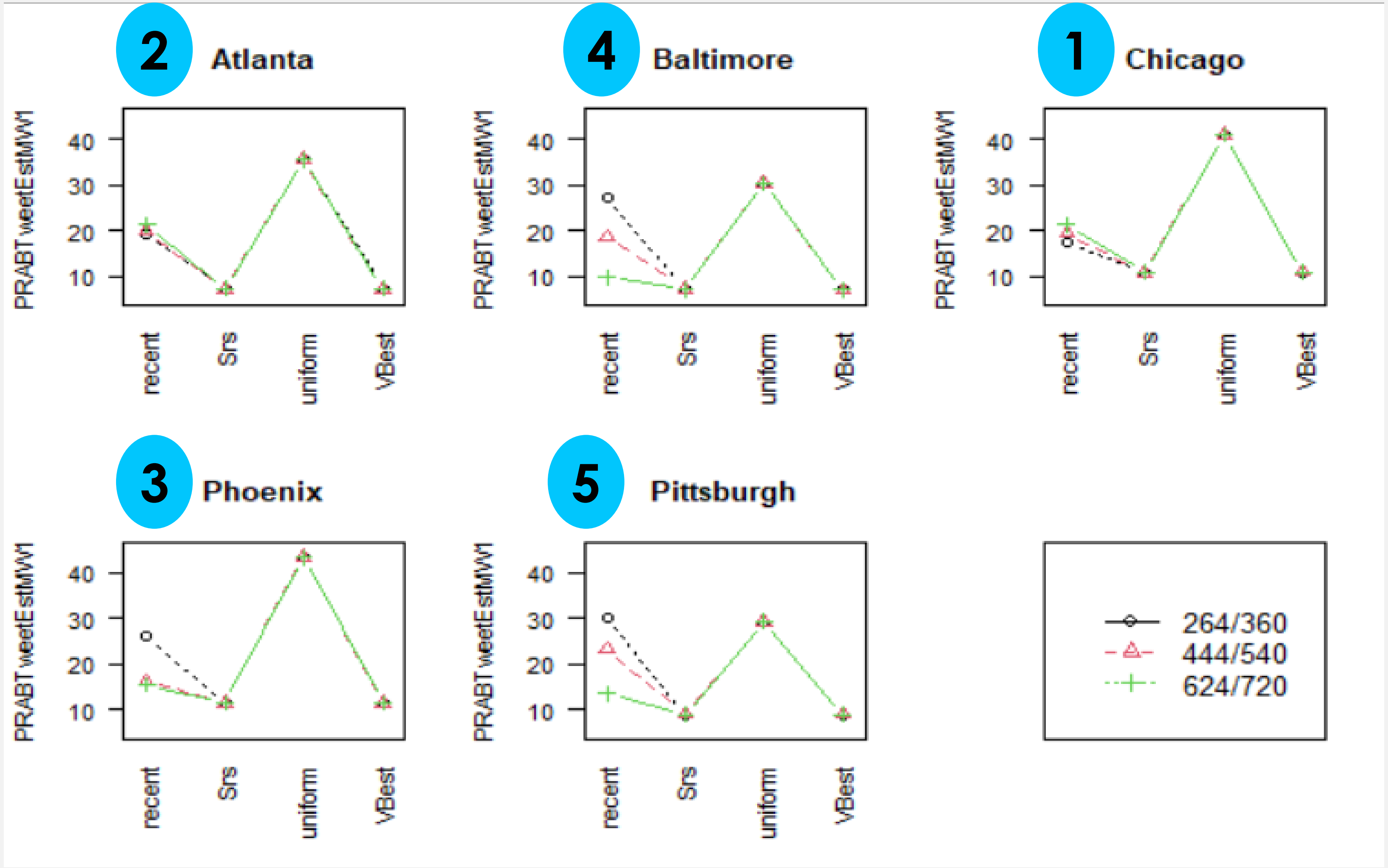
Filter by Keywords — 2,341,500

*__From MSAs:__ Atlanta, Baltimore, Chicago, Phoenix and Pittsburgh

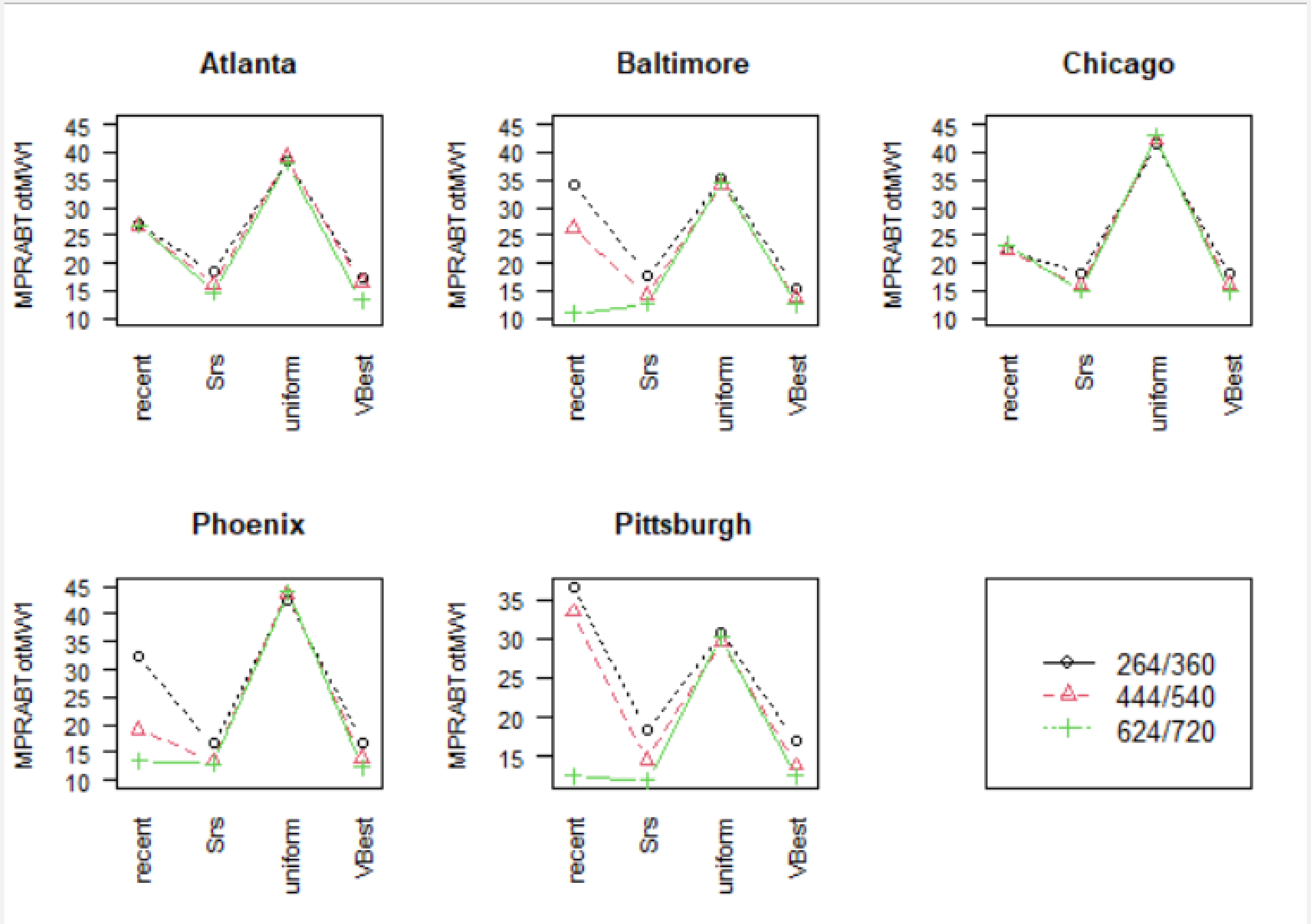
Preliminary Results: PRAB for Estimates of Size of the Tweet Population by Method and Sample Size



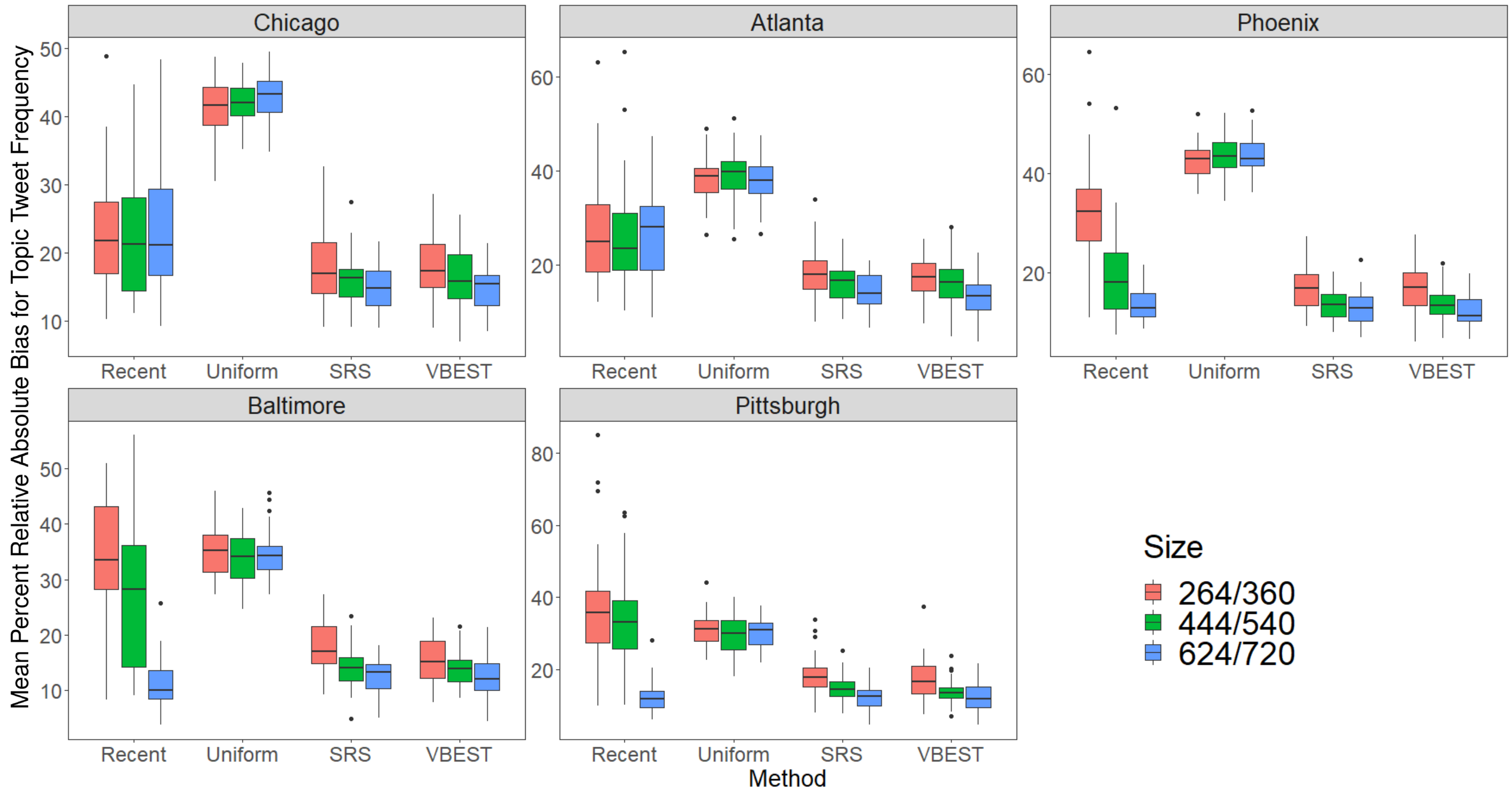
Preliminary Results: Overall PRAB for Estimates of Size of the Tweet Population for Each Region by Method and Sample Size



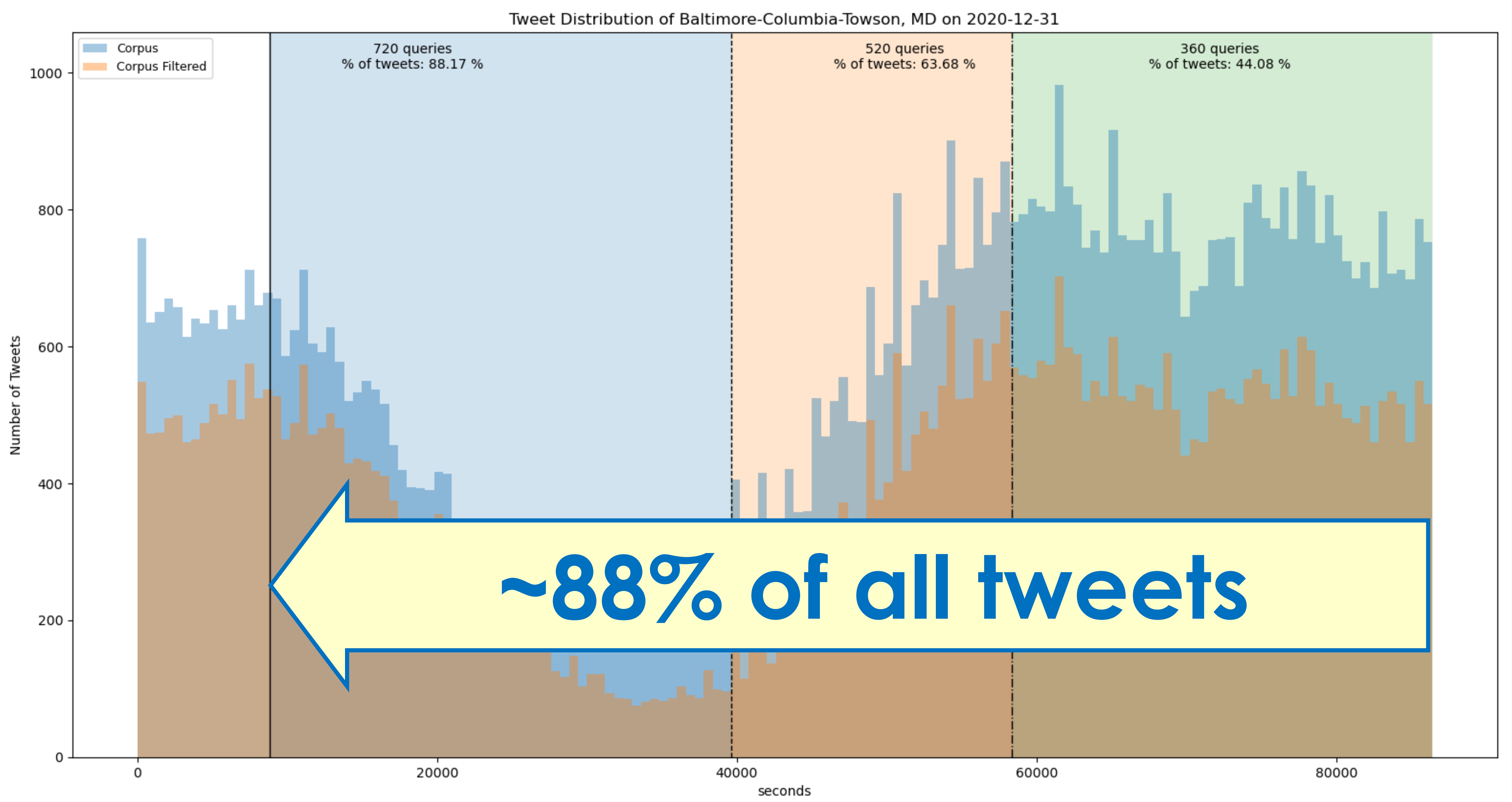
Preliminary Results: Overall Mean PRAB for Estimates of the Keyword Category Frequencies for Each Region by Method and Sample Size



Preliminary Results: Overall Mean PRAB for Estimates of the Topic Frequencies for Each Region by Method and Sample Size

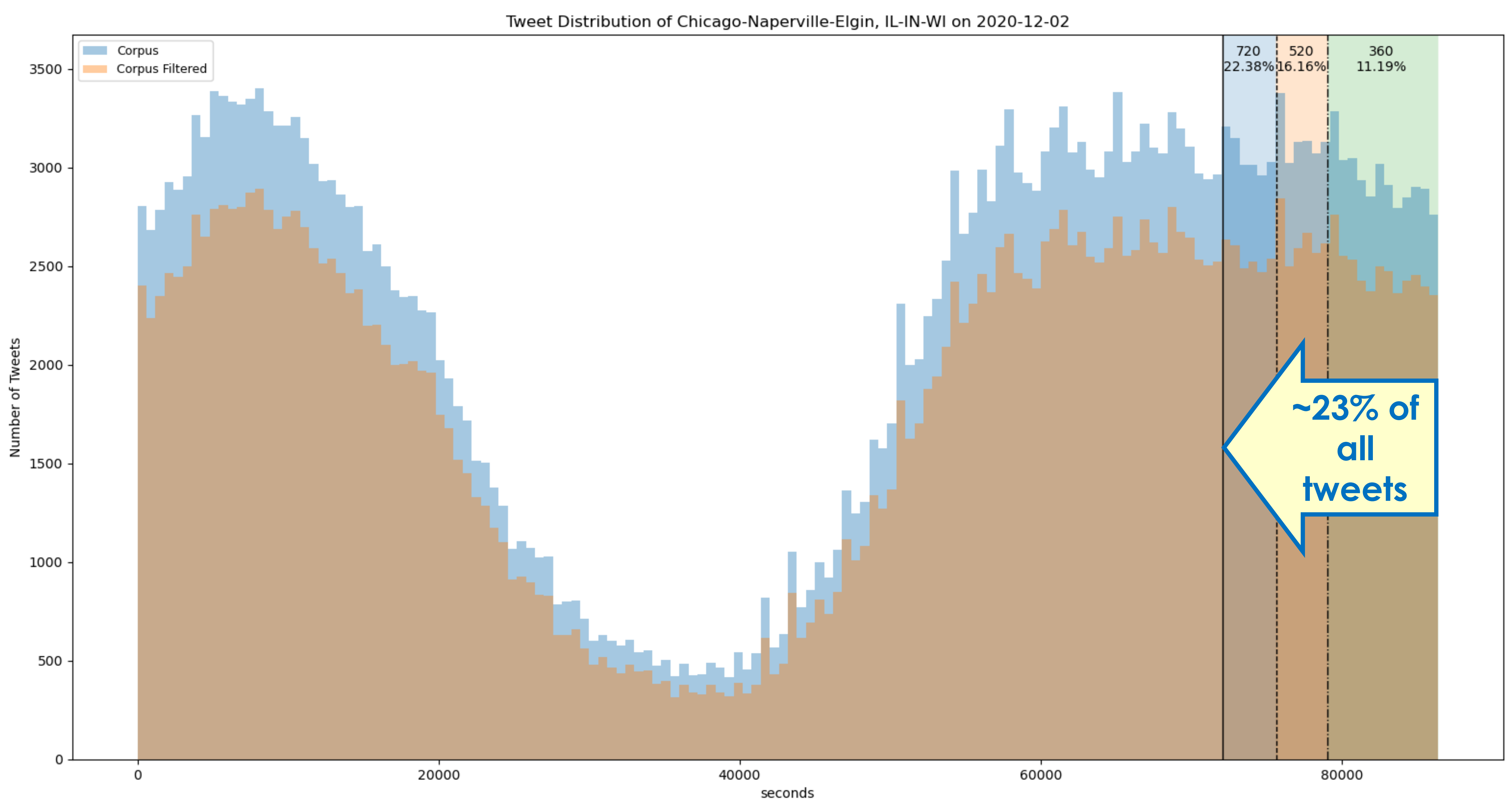


Why the Region by Method by Size Interaction?
Baltimore… (second smallest MSA by Tweet Volume)



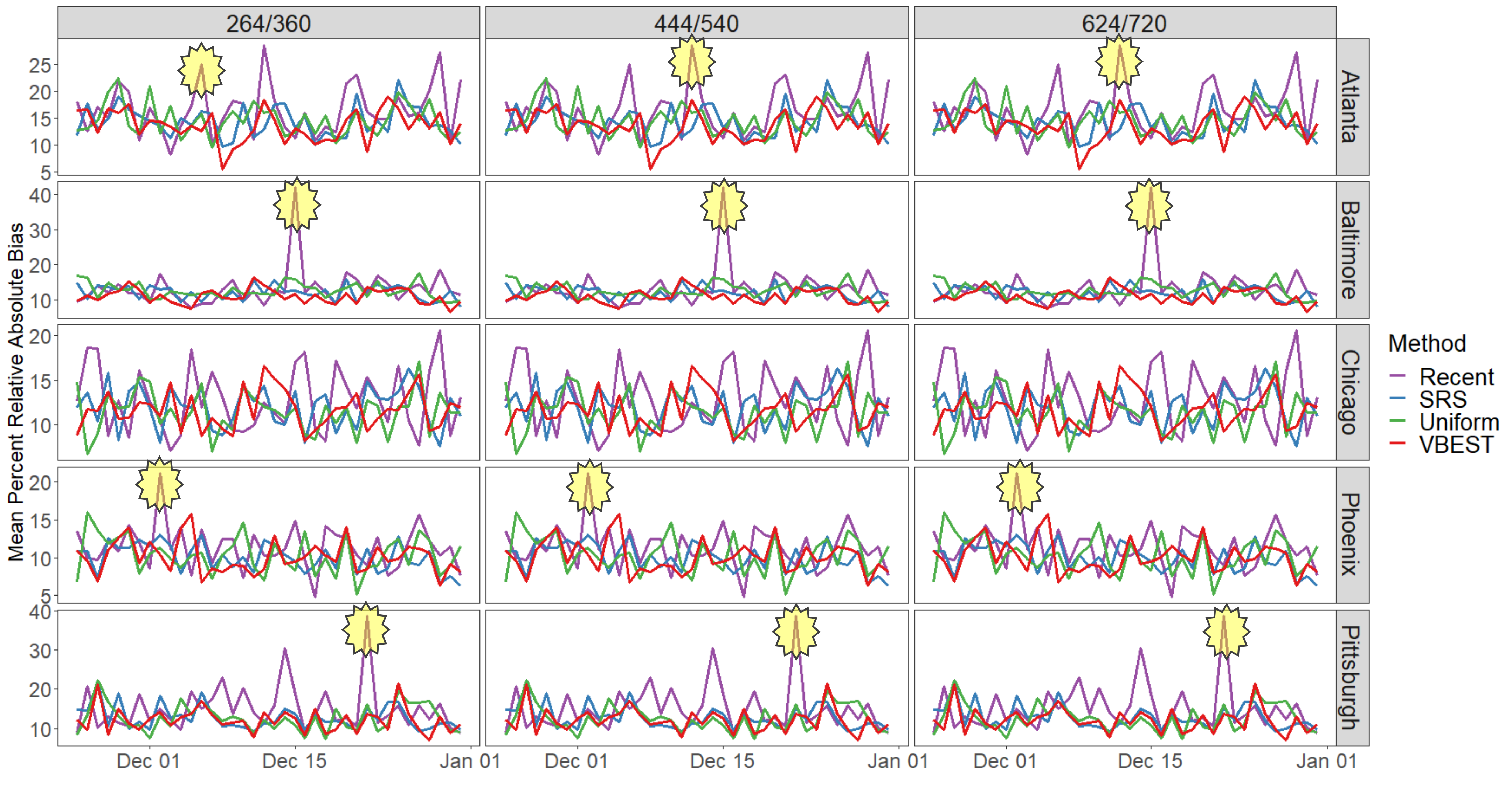
~88% of all tweets

Why the Region by Method by Size Interaction?
Chicago… (largest MSA by Tweet Volume)

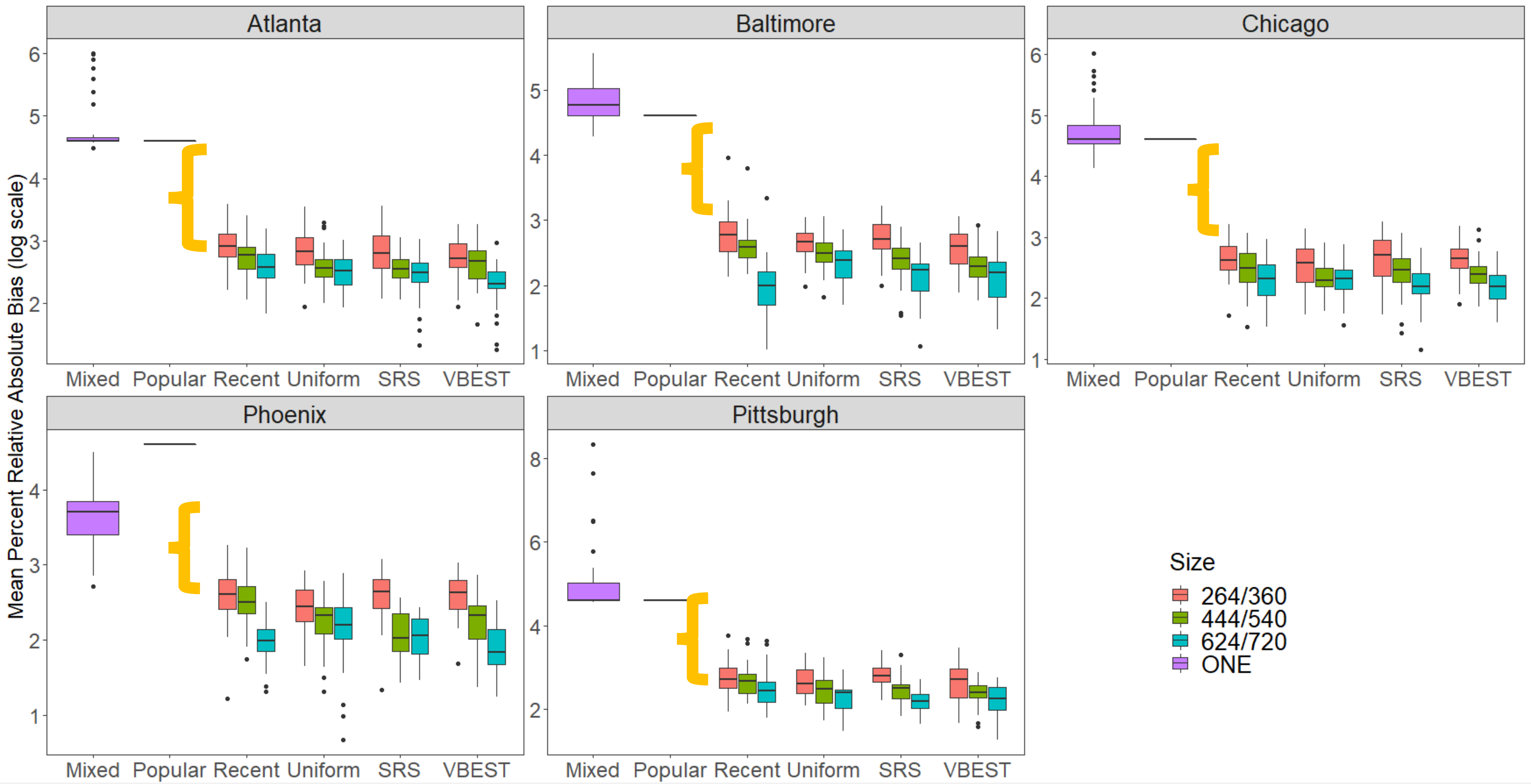


~23% of all tweets

Preliminary Findings: Mean PRAB (across all keyword categories) by Region, Method and Size across 38 Days



Preliminary Results: Mean PRAB by Region, Method and Size



Key Takeaways and Next Steps

- **We analyzed two main outcomes: Tweet Population for each Region and overall Incidence of Covid Related Keywords.**
  - Generally, mixed and popular should not be used to derive estimates from Twitter samples – their bias is orders of magnitude worse than any of the other methods.
  - The recent method performs reasonable well and in many cases is similar to Uniform, SRS or VBEST – however, it can have more volatility from day to day.
  - The uniform method, while spacing out tweet samples across the day, suffers from duplication issues and potentially over-estimates activity in the "trough period" in the middle of the night.
  - VBEST and SRS have the most predictable and stable behavior – larger sample sizes produce more accurate measures, generally using 90 fewer overall queries for the point estimates. The VBEST methodology also produces an estimate of twitter volume that can be used outright or for planning purposes. The other methods do not produce such an estimate.

Next Steps

- We are continuing to analyze the individual keyword incidences to understand the relationship between bias, overall popularity of a keyword, method and sample size.

- We are also working on an extension of the VBEST algorithm at Step 1 to produce estimates of keyword incidence based on modeling the distribution throughout the day.

- We are fielding a second experiment to look at different ways to compute initial velocities and velocity curves to potentially improve the construction of Tweet PSUs.

THANK YOU!

- Live Q & A, Thursday May 13 @ 2:30 pm EDT
- @trentbuskirk
- buskirk@bgsu.edu