The Effect of Population and “Structural” Biases on Social Media-based Algorithms – A Case Study in Geolocation Inference Across the Urban-Rural Spectrum

CS492 Paper presentation

April 8th, 2019

Jaeryoung Ka

Social Media Contributes to Making New Algorithms

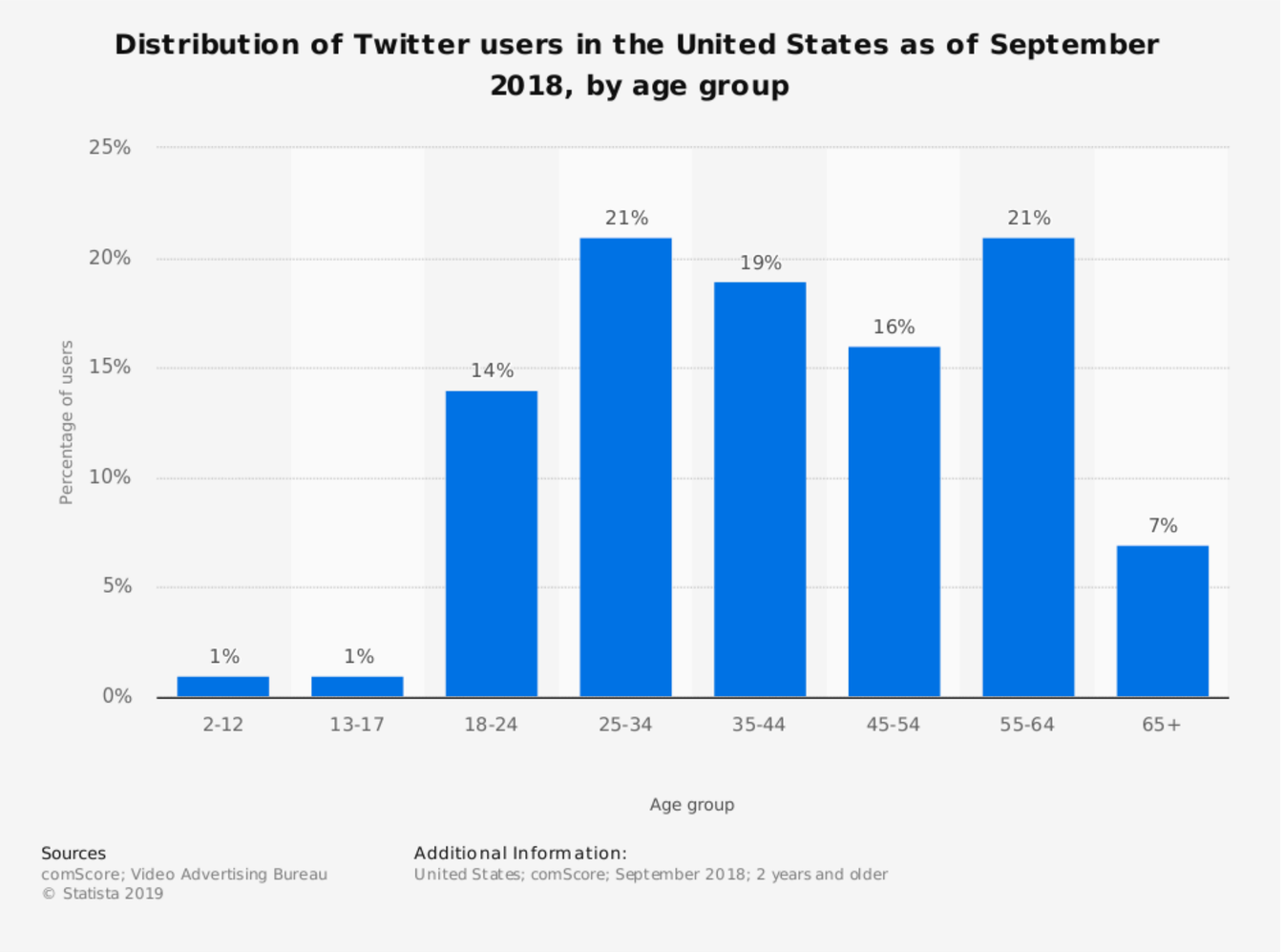
- Social Media plays a large role on developing new intelligent algorithms

- Recommender systems (Falher et al., 2015)
- Location inference (Jurgens et al., 2015)

Short Survey: the Age Group most likely to use Twitter is

- Below 17
- 17 - 24
- 25 - 34
- 35 - 44
- 45 - 54
- 55 - 64
- 65+

Short Survey: the Age Group most likely to use Twitter is



Social Media Data is prone to Suffer from Population Bias

- Population bias is endemic to most social media dataset
- Researchers recognized this and working on it
- Ruths, Derek, and Jürgen Pfeffer. "Social media for large studies of behavior.", *Nature*

SOCIAL SCIENCES

Social media for large studies of behavior

Large-scale studies of human behavior in social media need to be held to higher methodological standards

By **Derek Ruths**[1]* *and* **Jürgen Pfeffer**[2]

On 3 November 1948, the day after Harry Truman won the United States presidential elections, the *Chicago Tribune* published one of the most famous erroneous headlines in newspaper history: "Dewey Defeats Truman" (*1*, *2*). The headline was informed by telephone surveys, which had inadvertently undersampled Truman supporters (*1*). Rather than permanently discrediting the practice of polling, this event led to the

different social media platforms (*8*). For instance, Instagram is "especially appealing to adults aged 18 to 29, African-American, Latinos, women, urban residents" (*9*) whereas Pinterest is dominated by females, aged 25 to 34, with an average annual household income of $100,000 (*10*). These sampling biases are rarely corrected for (if even acknowledged).

Proprietary algorithms for public data. Platform-specific sampling problems, for example, the highest-volume source of public Twitter data, which are used by thousands of researchers worldwide, is not an accurate representation of the overall plat-

The rise of "embedded researchers" (researchers who have special relationships with providers that give them elevated access to platform-specific data, algorithms, and resources) is creating a divided social media research community. Such researchers, for example, can see a platform's inner workings and make accommodations, but may not be able to reveal their corrections or the data used to generate their findings.

REPRESENTATION OF HUMAN BEHAVIOR. *Human behavior and online platform design.* Many social forces that drive the

There is Another Bias: Structural Bias, comes with Design Choices

- What contribute to the bias of an algorithm?
- Of course, Population bias
- Also structural bias, which arises from algorithmic design choices

- Algorithmic Bias = Population Bias + Structural Bias + ε

Sadly, Little Investigation is Done on Structural Bias

- Unlike many works which dealt with Population bias, there is little investigation on the Structural bias

- Goal of the paper: Helping people address this gap; taking a detailed look at structural bias

Population bias	Structural bias

Case study: Geolocation inference using Tweets

Models and Datasets

Geolocation inference using Tweets

- Goal: predicting the location of a twitter user or tweets
- Typically done by analyzing the content of tweet and/or geographical configuration of explicitly encoded social ties
- Decided to perform a case study since social science is too broad

Two kinds of Geolocation Inference Algorithms

- Priedhorsky et al
  - Text-based approach, geolocate a tweet

- Jurgens et al
  - Network-based approach, geolocate a twitter user

- Paper utilized the all two approaches to get the result

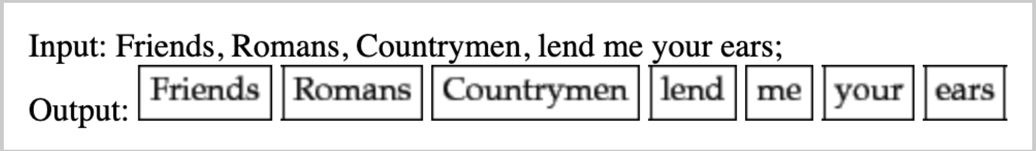
Text-based Algorithm from Priedhorsky et al.

- Trained on a set number of tweets with known locations

- They analyzed: tokens of the tweet, Timezone of the user, Self-reported location field, and specified language

What is a token?

Tokenization is the task of chopping a character sequence up into pieces, called *tokens*.



Some characters are thrown away, such as punctuations.

More detail: https://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html

The token analyzed in this paper is an NLP term; not the tokens that applications use to access Twitter user's data.

Text-based Algorithm: GMMs

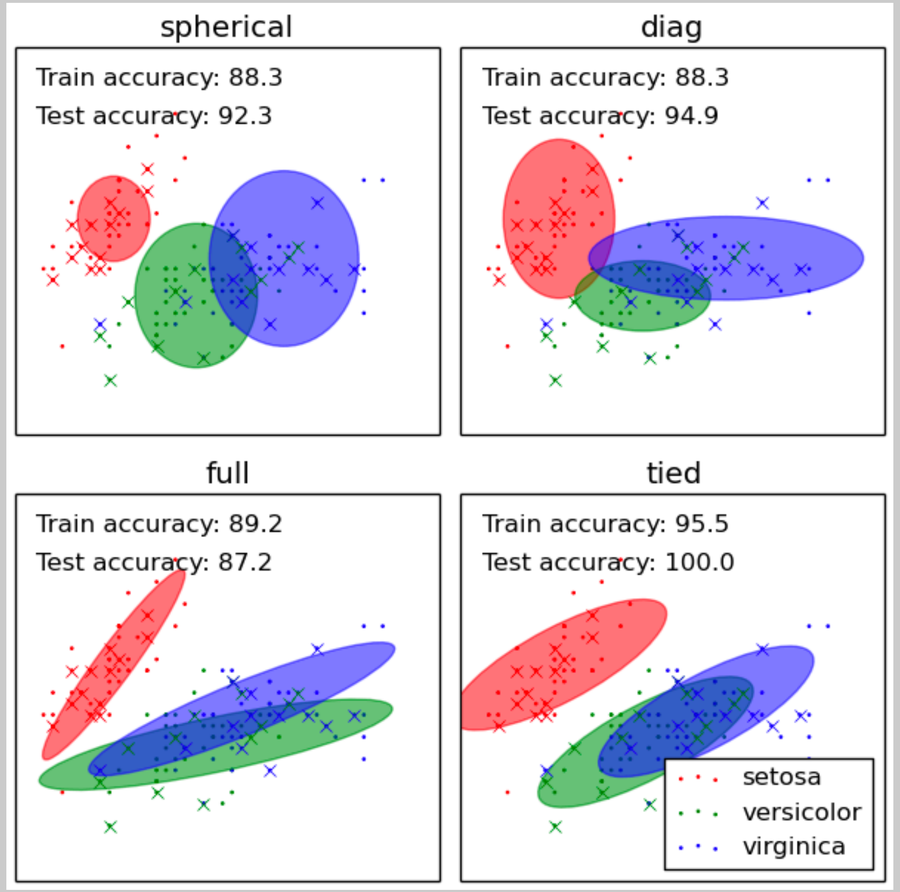
- Used Gaussian mixture models (GMMs) to deal with the data

- GMM captures the probability that a given token originated from an area based on the training data

- 1. Tokenize the tweet
  2. Weigh, combine individual GMM for each token in the tweet
  3. Find the highest probable area in the result

What is Gaussian mixture models (GMM) ?

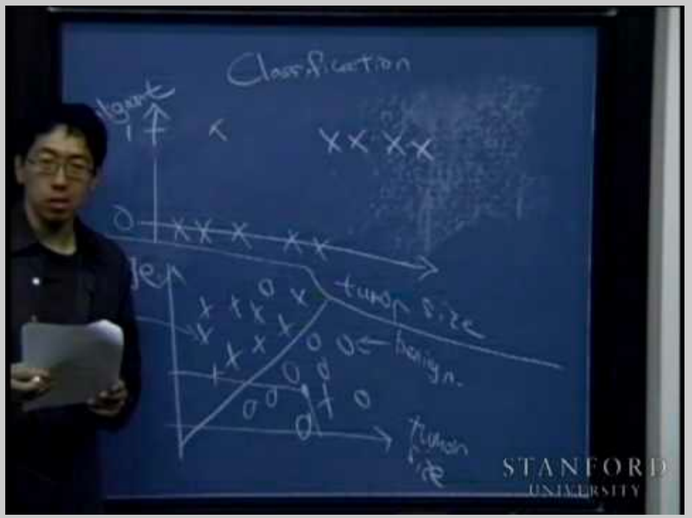
- A probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters.
- Commonly used for clustering purpose;
- Good to model differences that come from analyzed tokens
- Figure: GMM on Iris flower data set



What is Gaussian mixture models (GMM) ?

- The probability of belonging to each cluster is calculated, and then a classification is achieved by assigning each observation to the most likely cluster.

- More materials: Stanford CS229 Lecture Note http://cs229.stanford.edu/notes/cs229-notes7b.pdf



Text-based Algorithm: Datasets

- Twitter Streaming API
  - with bounding box which confines data to U.S. (unlike original algorithm)
  - only tweets in contiguous United States are used

- 51.2M tweets from 1.6M unique Twitter users

- Contiguous United States excludes the non-contiguous states of Alaska and Hawaii, and all other off-shore insular areas.

Network-based Algorithm from Jurgen et al.

- Building a bidirectional network by making an edge between two users who mentioned each other in Twitter

- Iteratively propagates the location of the known users to any of their neighbors with unknown-location, as the median of previously located neighbors

- This is repeated for 5 times

Network-based Algorithm: Datasets

- Building a mention network from dataset from randomly collected tweets
  99M tweets from 26M Twitter users from August 2015 to September 2015

- Restricted dataset to U.S. only for consistency, yielding
  3.2M tweets from 1.2M Twitter users from same date range

- 113K comprised the ground truth of final network

Ground Truth and Fun Facts

- U.S. National Center for Health Statistics classified each county in the U.S. from 1 to 6 scale: 1 being most urban, 6 being most rural
- Home location of a user is determined by geocoding location field using the Wikipedia-based geocoder (Brent Hecht and Darren Gergle, 2010) to reduce the noise
- The most urban users (class 1) is highly overrepresented (130% to 210%)
- The most rural users (class 6) is highly underrepresented (24% to 45%)

Geolocation inference using Tweets

Evaluations and Results

Evaluation Framework: Definitions

- Definition of algorithm precision: % of predictions within 100km range of the actual location

- Definition of true positive: Prediction lying on same country as the ground truth

- In the range from 1 to 6, defined by NCHS,
  - Urban refers to one in code 1 and 2
  - Rural refers to one in code 5 and 6

Evaluation Framework on Text-based Model

- Tested 5 models for each condition
- Constrained data to ensure no overlap in training phases and testing phases
- Training set size is 30K, which is equal to the size of reference paper

Evaluation Framework on Network-based Model

- 5-fold cross validation for each network-based model
- Training dataset contains 24K users; smaller than that used in the reference paper, Due to the limited number of tweets in selected data source (streaming API)

What is K-fold cross Validation? Why do it?

- Randomly partition the original sample into k partitions
- 1 Subsample used as a validation data, k-1 subsamples as a training data
- This repeated for k times, each subsample being a validation data once
- Generally results in a less biased result for the model
- For more information, refer to Stanford CS229 Lecture Notes (http://cs229.stanford.edu/notes/cs229-notes5.pdf)

Result: Geolocation Inference Algorithms are Worse for rural users

- They found that geolocation algorithms have lower performance for rural users regardless of the methodology or definition of problem.

	% of Training Data		Precision (Recall)		
Text-Based Models	**Urban**	**Rural**	**Urban**	**Rural**	**Overall**
Typical (Baseline)	63.8%	9.0%	23.0 ± 3.0%	9.9 ± 0.4%	20.6 ± 1.9%
Population Bias Balanced	55.3%	15.0%	22.1 ± 1.5%	10.5 ± 0.6%	20.4 ± 1.1%
Urban Boosted	100%	0%	27.6 ± 3.0%	4.9 ± 0.3%	19.5 ± 2.0%
Rural Boosted	0%	100%	5.0 ± 0.5%	17.9 ± 1.5%	6.8 ± 0.6%
Network-Based Models	**Urban**	**Rural**	**Urban**	**Rural**	**Overall**
Typical (Baseline)	75.0%	5.2%	25.7 ± 0.4% (13.1%)	20.6 ± 1.7% (8.1%)	25.0 ± 0.4% (12.3%)
Population Bias Balanced	55.3%	15.0%	20.6 ± 0.8% (12.4%)	39.2 ± 5.2% (9.5%)	22.2 ± 0.8% (11.9%)
Urban Boosted	100%	0%	27.0 ± 3.9% (13.3%)	5.0 ± 1.9% (9.0%)	22.3 ± 3.2% (11.6%)
Rural Boosted*	0%	100%	1.0 ± 0.3% (4.6%)	59.2 ± 5.3% (8.4%)	3.8 ± 0.4% (4.5%)

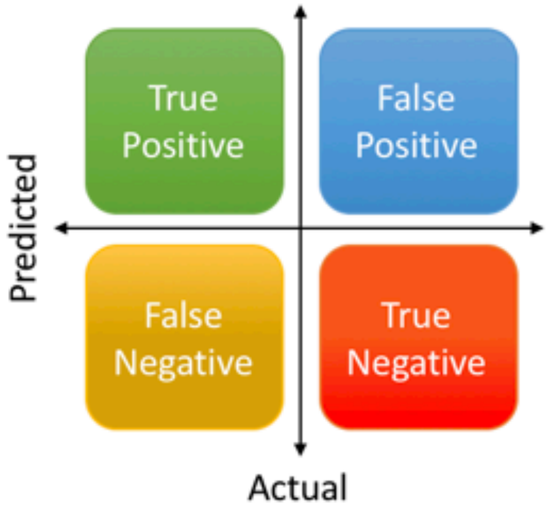
Table 1. Urban-rural results for typical training data as well as various population bias manipulations.
Precision: confidence intervals for percentage of predictions within 100 km of true location.
Recall: values in parentheses, except the text-based models which had recall always around 100%.
***The network-based rural-boosted model was trained on 6000 tweets due to lack of data.**

Wait.. what is Precision and Recall?

$$\textbf{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

$$\textbf{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

$$\textbf{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{Total}}$$



Now let’s back to the original biases problem

Research Questions and findings

Research Questions

- RQ1:
  Is there algorithmic bias?

- RQ2:
  Is the observed bias due to population bias?

- RQ3:
  Can any remaining underperformance for a specific population be fixed by training solely on data from that population?

RQ1: Is there algorithmic bias?

- Yes. Typical (Baseline) shows the low performance of algorithm for rural users.
- Note that the higher the value is, the more data is actually put in the map.

	% of Training Data		Precision (Recall)		
Text-Based Models	**Urban**	**Rural**	**Urban**	**Rural**	**Overall**
Typical (Baseline)	63.8%	9.0%	23.0 ± 3.0%	9.9 ± 0.4%	20.6 ± 1.9%
Population Bias Balanced	55.3%	15.0%	22.1 ± 1.5%	10.5 ± 0.6%	20.4 ± 1.1%
Urban Boosted	100%	0%	27.6 ± 3.0%	4.9 ± 0.3%	19.5 ± 2.0%
Rural Boosted	0%	100%	5.0 ± 0.5%	17.9 ± 1.5%	6.8 ± 0.6%
Network-Based Models	**Urban**	**Rural**	**Urban**	**Rural**	**Overall**
Typical (Baseline)	75.0%	5.2%	25.7 ± 0.4% (13.1%)	20.6 ± 1.7% (8.1%)	25.0 ± 0.4% (12.3%)
Population Bias Balanced	55.3%	15.0%	20.6 ± 0.8% (12.4%)	39.2 ± 5.2% (9.5%)	22.2 ± 0.8% (11.9%)
Urban Boosted	100%	0%	27.0 ± 3.9% (13.3%)	5.0 ± 1.9% (9.0%)	22.3 ± 3.2% (11.6%)
Rural Boosted*	0%	100%	1.0 ± 0.3% (4.6%)	59.2 ± 5.3% (8.4%)	3.8 ± 0.4% (4.5%)

Table 1. Urban-rural results for typical training data as well as various population bias manipulations.
Precision: confidence intervals for percentage of predictions within 100 km of true location.
Recall: values in parentheses, except the text-based models which had recall always around 100%.
***The network-based rural-boosted model was trained on 6000 tweets due to lack of data.**

RQ2: Is the observed bias due to population bias?

- First, we can check that there is obvious population bias by comparing the ratio of the classification result.

	% of Training Data		Precision (Recall)		
Text-Based Models	**Urban**	**Rural**	**Urban**	**Rural**	**Overall**
Typical (Baseline)	63.8%	9.0%	23.0 ± 3.0%	9.9 ± 0.4%	20.6 ± 1.9%
Population Bias Balanced	55.3%	15.0%	22.1 ± 1.5%	10.5 ± 0.6%	20.4 ± 1.1%
Urban Boosted	100%	0%	27.6 ± 3.0%	4.9 ± 0.3%	19.5 ± 2.0%
Rural Boosted	0%	100%	5.0 ± 0.5%	17.9 ± 1.5%	6.8 ± 0.6%
Network-Based Models	**Urban**	**Rural**	**Urban**	**Rural**	**Overall**
Typical (Baseline)	75.0%	5.2%	25.7 ± 0.4% (13.1%)	20.6 ± 1.7% (8.1%)	25.0 ± 0.4% (12.3%)
Population Bias Balanced	55.3%	15.0%	20.6 ± 0.8% (12.4%)	39.2 ± 5.2% (9.5%)	22.2 ± 0.8% (11.9%)
Urban Boosted	100%	0%	27.0 ± 3.9% (13.3%)	5.0 ± 1.9% (9.0%)	22.3 ± 3.2% (11.6%)
Rural Boosted*	0%	100%	1.0 ± 0.3% (4.6%)	59.2 ± 5.3% (8.4%)	3.8 ± 0.4% (4.5%)

Table 1. Urban-rural results for typical training data as well as various population bias manipulations.
Precision: confidence intervals for percentage of predictions within 100 km of true location.
Recall: values in parentheses, except the text-based models which had recall always around 100%.
***The network-based rural-boosted model was trained on 6000 tweets due to lack of data.**

RQ2: Is the observed bias due to population bias?

- Then, the paper balanced out the population bias to 55.3% and 15.0%, which is the result of the actual census.

	% of Training Data		Precision (Recall)		
Text-Based Models	**Urban**	**Rural**	**Urban**	**Rural**	**Overall**
Typical (Baseline)	63.8%	9.0%	23.0 ± 3.0%	9.9 ± 0.4%	20.6 ± 1.9%
Population Bias Balanced	55.3%	15.0%	22.1 ± 1.5%	10.5 ± 0.6%	20.4 ± 1.1%
Urban Boosted	100%	0%	27.6 ± 3.0%	4.9 ± 0.3%	19.5 ± 2.0%
Rural Boosted	0%	100%	5.0 ± 0.5%	17.9 ± 1.5%	6.8 ± 0.6%
Network-Based Models	**Urban**	**Rural**	**Urban**	**Rural**	**Overall**
Typical (Baseline)	75.0%	5.2%	25.7 ± 0.4% (13.1%)	20.6 ± 1.7% (8.1%)	25.0 ± 0.4% (12.3%)
Population Bias Balanced	55.3%	15.0%	20.6 ± 0.8% (12.4%)	39.2 ± 5.2% (9.5%)	22.2 ± 0.8% (11.9%)
Urban Boosted	100%	0%	27.0 ± 3.9% (13.3%)	5.0 ± 1.9% (9.0%)	22.3 ± 3.2% (11.6%)
Rural Boosted*	0%	100%	1.0 ± 0.3% (4.6%)	59.2 ± 5.3% (8.4%)	3.8 ± 0.4% (4.5%)

Table 1. Urban-rural results for typical training data as well as various population bias manipulations.
Precision: confidence intervals for percentage of predictions within 100 km of true location.
Recall: values in parentheses, except the text-based models which had recall always around 100%.
***The network-based rural-boosted model was trained on 6000 tweets due to lack of data.**

RQ2: Is the observed bias due to population bias?

- Text-based models remain highly urban-biased, network-based models become less urban-biased. In other words, we cannot conclude that population bias is the sole contributor to this.

	% of Training Data		Precision (Recall)		
Text-Based Models	**Urban**	**Rural**	**Urban**	**Rural**	**Overall**
Typical (Baseline)	63.8%	9.0%	23.0 ± 3.0%	9.9 ± 0.4%	20.6 ± 1.9%
Population Bias Balanced	55.3%	15.0%	22.1 ± 1.5%	10.5 ± 0.6%	20.4 ± 1.1%
Urban Boosted	100%	0%	27.6 ± 3.0%	4.9 ± 0.3%	19.5 ± 2.0%
Rural Boosted	0%	100%	5.0 ± 0.5%	17.9 ± 1.5%	6.8 ± 0.6%
Network-Based Models	**Urban**	**Rural**	**Urban**	**Rural**	**Overall**
Typical (Baseline)	75.0%	5.2%	25.7 ± 0.4% (13.1%)	20.6 ± 1.7% (8.1%)	25.0 ± 0.4% (12.3%)
Population Bias Balanced	55.3%	15.0%	20.6 ± 0.8% (12.4%)	39.2 ± 5.2% (9.5%)	22.2 ± 0.8% (11.9%)
Urban Boosted	100%	0%	27.0 ± 3.9% (13.3%)	5.0 ± 1.9% (9.0%)	22.3 ± 3.2% (11.6%)
Rural Boosted*	0%	100%	1.0 ± 0.3% (4.6%)	59.2 ± 5.3% (8.4%)	3.8 ± 0.4% (4.5%)

Table 1. Urban-rural results for typical training data as well as various population bias manipulations.
Precision: confidence intervals for percentage of predictions within 100 km of true location.
Recall: values in parentheses, except the text-based models which had recall always around 100%.
***The network-based rural-boosted model was trained on 6000 tweets due to lack of data.**

RQ3: Can we fix biases through oversampling?

- To answer this question, the paper trained and tested each group separately, and then compared the result. (i.e. used separate models for each group)

	% of Training Data		Precision (Recall)		
Text-Based Models	**Urban**	**Rural**	**Urban**	**Rural**	**Overall**
Typical (Baseline)	63.8%	9.0%	23.0 ± 3.0%	9.9 ± 0.4%	20.6 ± 1.9%
Population Bias Balanced	55.3%	15.0%	22.1 ± 1.5%	10.5 ± 0.6%	20.4 ± 1.1%
Urban Boosted	100%	0%	27.6 ± 3.0%	4.9 ± 0.3%	19.5 ± 2.0%
Rural Boosted	0%	100%	5.0 ± 0.5%	17.9 ± 1.5%	6.8 ± 0.6%
Network-Based Models	**Urban**	**Rural**	**Urban**	**Rural**	**Overall**
Typical (Baseline)	75.0%	5.2%	25.7 ± 0.4% (13.1%)	20.6 ± 1.7% (8.1%)	25.0 ± 0.4% (12.3%)
Population Bias Balanced	55.3%	15.0%	20.6 ± 0.8% (12.4%)	39.2 ± 5.2% (9.5%)	22.2 ± 0.8% (11.9%)
Urban Boosted	100%	0%	27.0 ± 3.9% (13.3%)	5.0 ± 1.9% (9.0%)	22.3 ± 3.2% (11.6%)
Rural Boosted*	0%	100%	1.0 ± 0.3% (4.6%)	59.2 ± 5.3% (8.4%)	3.8 ± 0.4% (4.5%)

Table 1. Urban-rural results for typical training data as well as various population bias manipulations.
Precision: confidence intervals for percentage of predictions within 100 km of true location.
Recall: values in parentheses, except the text-based models which had recall always around 100%.
***The network-based rural-boosted model was trained on 6000 tweets due to lack of data.**

RQ3: Can we fix biases through oversampling?

- Bias still exists, so the reply to the question is No. However, the degree differs.
- More bias lingered on text-based models compared to the network-based models.

	% of Training Data		Precision (Recall)		
Text-Based Models	**Urban**	**Rural**	**Urban**	**Rural**	**Overall**
Typical (Baseline)	63.8%	9.0%	23.0 ± 3.0%	9.9 ± 0.4%	20.6 ± 1.9%
Population Bias Balanced	55.3%	15.0%	22.1 ± 1.5%	10.5 ± 0.6%	20.4 ± 1.1%
Urban Boosted	100%	0%	27.6 ± 3.0%	4.9 ± 0.3%	19.5 ± 2.0%
Rural Boosted	0%	100%	5.0 ± 0.5%	17.9 ± 1.5%	6.8 ± 0.6%
Network-Based Models	**Urban**	**Rural**	**Urban**	**Rural**	**Overall**
Typical (Baseline)	75.0%	5.2%	25.7 ± 0.4% (13.1%)	20.6 ± 1.7% (8.1%)	25.0 ± 0.4% (12.3%)
Population Bias Balanced	55.3%	15.0%	20.6 ± 0.8% (12.4%)	39.2 ± 5.2% (9.5%)	22.2 ± 0.8% (11.9%)
Urban Boosted	100%	0%	27.0 ± 3.9% (13.3%)	5.0 ± 1.9% (9.0%)	22.3 ± 3.2% (11.6%)
Rural Boosted*	0%	100%	1.0 ± 0.3% (4.6%)	59.2 ± 5.3% (8.4%)	3.8 ± 0.4% (4.5%)

Table 1. Urban-rural results for typical training data as well as various population bias manipulations.
Precision: confidence intervals for percentage of predictions within 100 km of true location.
Recall: values in parentheses, except the text-based models which had recall always around 100%.
***The network-based rural-boosted model was trained on 6000 tweets due to lack of data.**

Algorithmic Bias = Population Bias + Structural Bias + ε

More things on Structural Bias

Text-based case: 25% more *wikifiable* words in Urban area

- Examined average number of words that can be identified as geographic Wikipedia concepts (e.g. words tied directly to place)

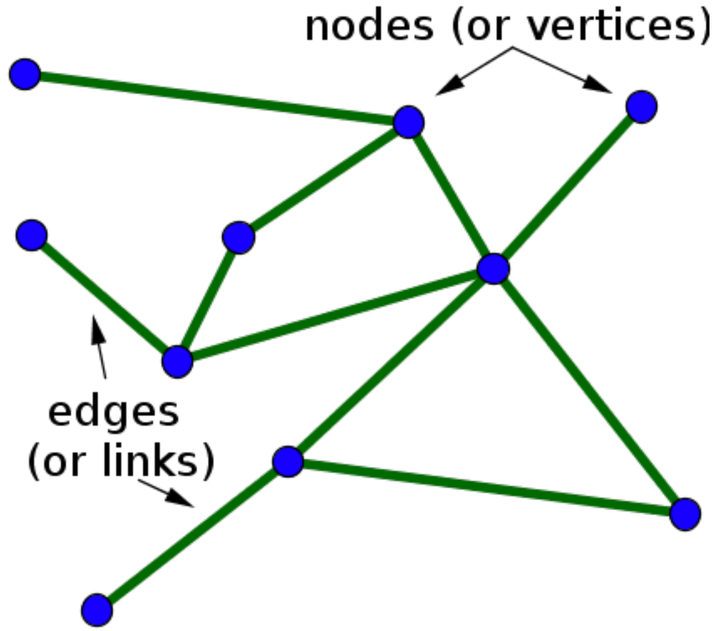
- Used Wikification algorithms already implemented in Sen et al. (2014)

- Most urban tweets (class 1) have 25% more *wikifiable* words than most rural tweets (class 6)

Network-based case: Birds of a Feather Flock Together

In Network-based algorithm, very high homophily is observed.

Consider edges between users whose location is known. Then 90% of the edges are one between a same rural-class group.



Network-based case: Birds of a Feather Flock Together

For example, suppose that A and B mentioned each other in Twitter.

If we know the location of A and B and A is classified in class 6 (most rural), then B has a 90% chance to be classified as a person in class 6.

Pitfall of fixed-distance parameter

- Consider a situation where you tweet an elementary school in both urban and rural area. How will the tweet in the two situations contribute to locating you?

- Note that rural area has bigger school district by area, due to low population density.

Pitfall of fixed-distance parameter

- Fixed-distance parameter assumes a fixed range of spatial autocorrelation for tweet usage, which isn't true.

- This would be the one cause of structural bias. Fixed distance parameters brought about failing to capture the full predictive power of each rural tweet in the text-based algorithm.

More discussions on the outcomes
(Yay!)

Seeking for Parity

- Suppose you are a researcher who want to utilize people's tweets for your project. Which you think is the better, text-based one or network-based one?

	% of Training Data		Precision (Recall)		
Text-Based Models	**Urban**	**Rural**	**Urban**	**Rural**	**Overall**
Typical (Baseline)	63.8%	9.0%	23.0 ± 3.0%	9.9 ± 0.4%	20.6 ± 1.9%
Population Bias Balanced	55.3%	15.0%	22.1 ± 1.5%	10.5 ± 0.6%	20.4 ± 1.1%
Urban Boosted	100%	0%	27.6 ± 3.0%	4.9 ± 0.3%	19.5 ± 2.0%
Rural Boosted	0%	100%	5.0 ± 0.5%	17.9 ± 1.5%	6.8 ± 0.6%
Network-Based Models	**Urban**	**Rural**	**Urban**	**Rural**	**Overall**
Typical (Baseline)	75.0%	5.2%	25.7 ± 0.4% (13.1%)	20.6 ± 1.7% (8.1%)	25.0 ± 0.4% (12.3%)
Population Bias Balanced	55.3%	15.0%	20.6 ± 0.8% (12.4%)	39.2 ± 5.2% (9.5%)	22.2 ± 0.8% (11.9%)
Urban Boosted	100%	0%	27.0 ± 3.9% (13.3%)	5.0 ± 1.9% (9.0%)	22.3 ± 3.2% (11.6%)
Rural Boosted*	0%	100%	1.0 ± 0.3% (4.6%)	59.2 ± 5.3% (8.4%)	3.8 ± 0.4% (4.5%)

Table 1. Urban-rural results for typical training data as well as various population bias manipulations.
Precision: confidence intervals for percentage of predictions within 100 km of true location.
Recall: values in parentheses, except the text-based models which had recall always around 100%.
***The network-based rural-boosted model was trained on 6000 tweets due to lack of data.**

Seeking for Parity

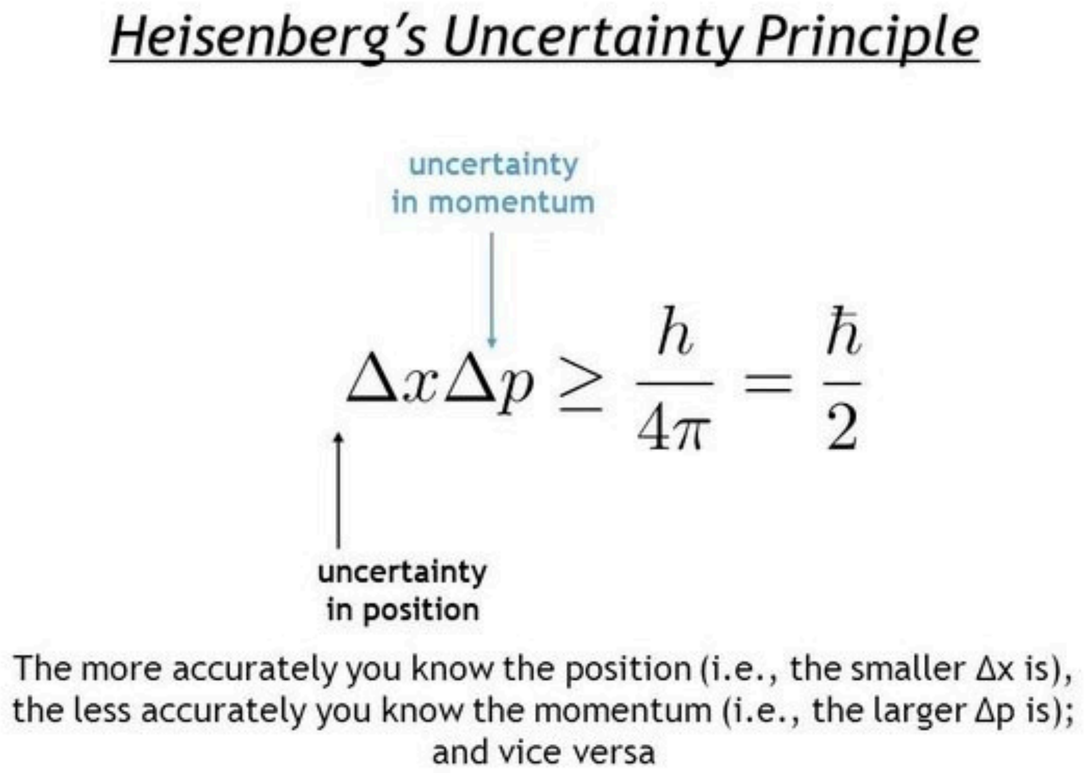
- In the network-based algorithm, it can result in more parity to investigate more efforts on data collection over underrepresented populations.

	% of Training Data		Precision (Recall)		
Text-Based Models	**Urban**	**Rural**	**Urban**	**Rural**	**Overall**
Typical (Baseline)	63.8%	9.0%	23.0 ± 3.0%	9.9 ± 0.4%	20.6 ± 1.9%
Population Bias Balanced	55.3%	15.0%	22.1 ± 1.5%	10.5 ± 0.6%	20.4 ± 1.1%
Urban Boosted	100%	0%	27.6 ± 3.0%	4.9 ± 0.3%	19.5 ± 2.0%
Rural Boosted	0%	100%	5.0 ± 0.5%	17.9 ± 1.5%	6.8 ± 0.6%
Network-Based Models	**Urban**	**Rural**	**Urban**	**Rural**	**Overall**
Typical (Baseline)	75.0%	5.2%	25.7 ± 0.4% (13.1%)	20.6 ± 1.7% (8.1%)	25.0 ± 0.4% (12.3%)
Population Bias Balanced	55.3%	15.0%	20.6 ± 0.8% (12.4%)	39.2 ± 5.2% (9.5%)	22.2 ± 0.8% (11.9%)
Urban Boosted	100%	0%	27.0 ± 3.9% (13.3%)	5.0 ± 1.9% (9.0%)	22.3 ± 3.2% (11.6%)
Rural Boosted*	0%	100%	1.0 ± 0.3% (4.6%)	59.2 ± 5.3% (8.4%)	3.8 ± 0.4% (4.5%)

Table 1. Urban-rural results for typical training data as well as various population bias manipulations.
Precision: confidence intervals for percentage of predictions within 100 km of true location.
Recall: values in parentheses, except the text-based models which had recall always around 100%.
***The network-based rural-boosted model was trained on 6000 tweets due to lack of data.**

Trade-off: we can’t always seek for two rabbits

- What is a trade-off?



Trade-off: we can’t always seek for two rabbits

- What is a trade-off?

- Space–time tradeoff

FSU DEPARTMENT OF COMPUTER SCIENCE

Loop Unrolling

- Loop Unrolling means that multiple copies of the loop body are made.
- Original Loop:

```
for (i = 0; i < n; i++)
   a[i] = b[i] + c[i];
```

- Unrolled Loop:

```
for (i = 0; i < n % 4; i++)
   a[i] = b[i] + c[i];
for (; i < n; i += 4) {
   a[i] = b[i] + c[i];
   a[i+1] = b[i+1] + c[i+1];
   a[i+2] = b[i+2] + c[i+2];
   a[i+3] = b[i+3] + c[i+3];
   }
```

- Reduces loop overhead and provides more opportunities for scheduling.

Trade-off: we can’t always seek for two rabbits

- What is a trade-off?
- Space–time tradeoff: you need to ‘trade’ decreasing elapsed time with increased space usage

기억이 나지 않습니다…
(I don’t remember…)

겨울연가 12화
From Drama ‘Winter Sonata’ E12

Trade-off between Equity and Effectiveness

- Overall column shows performance on randomized population
- The better the rural performance is, the worse the overall performance is.

	% of Training Data		Precision (Recall)		
Text-Based Models	**Urban**	**Rural**	**Urban**	**Rural**	**Overall**
Typical (Baseline)	63.8%	9.0%	23.0 ± 3.0%	9.9 ± 0.4%	20.6 ± 1.9%
Population Bias Balanced	55.3%	15.0%	22.1 ± 1.5%	10.5 ± 0.6%	20.4 ± 1.1%
Urban Boosted	100%	0%	27.6 ± 3.0%	4.9 ± 0.3%	19.5 ± 2.0%
Rural Boosted	0%	100%	5.0 ± 0.5%	17.9 ± 1.5%	6.8 ± 0.6%
Network-Based Models	**Urban**	**Rural**	**Urban**	**Rural**	**Overall**
Typical (Baseline)	75.0%	5.2%	25.7 ± 0.4% (13.1%)	20.6 ± 1.7% (8.1%)	25.0 ± 0.4% (12.3%)
Population Bias Balanced	55.3%	15.0%	20.6 ± 0.8% (12.4%)	39.2 ± 5.2% (9.5%)	22.2 ± 0.8% (11.9%)
Urban Boosted	100%	0%	27.0 ± 3.9% (13.3%)	5.0 ± 1.9% (9.0%)	22.3 ± 3.2% (11.6%)
Rural Boosted*	0%	100%	1.0 ± 0.3% (4.6%)	59.2 ± 5.3% (8.4%)	3.8 ± 0.4% (4.5%)

Table 1. Urban-rural results for typical training data as well as various population bias manipulations.
Precision: confidence intervals for percentage of predictions within 100 km of true location.
Recall: values in parentheses, except the text-based models which had recall always around 100%.
***The network-based rural-boosted model was trained on 6000 tweets due to lack of data.**

Trade-off between Equity and Effectiveness

- There is a clear trade-off between equity and effectiveness, at least in the case of Twitter geolocation problem.

- In other words, you need to sacrifice some effectiveness to seek for more equity and vice versa.

Is being underrepresented always bad?: Crypsis

Why have animals evolved so that they visually resemble the surroundings?

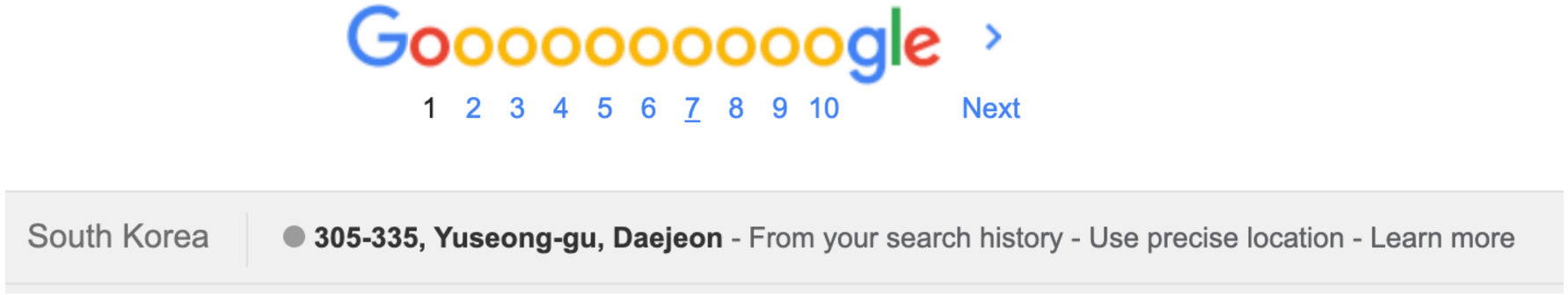
Inference attacks

Being easily spotted isn’t good for survival: they need to avoid being located

Inference attacks

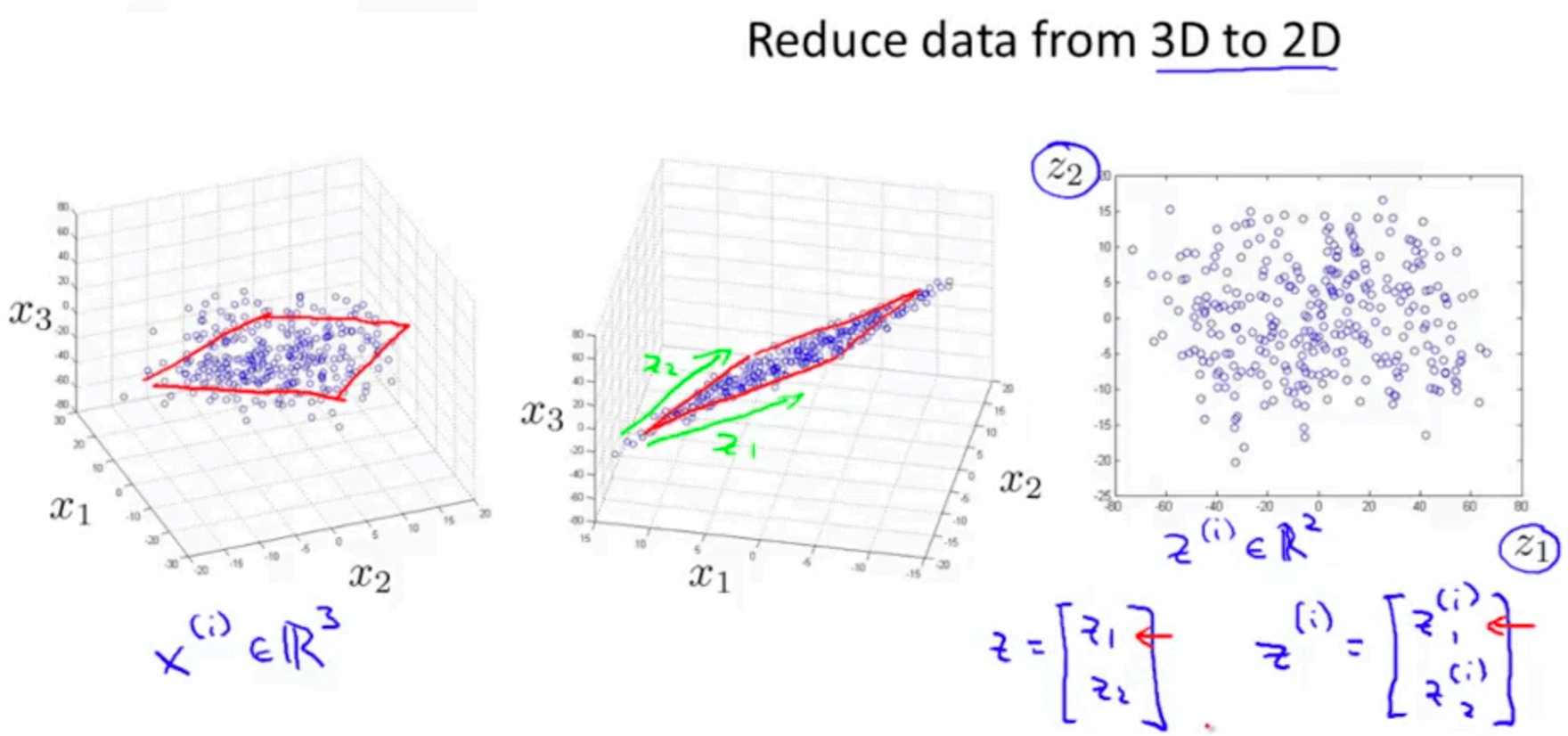
Being susceptible to getting located by others can bring about another problems

What if you need to worry about being tracked by everyone whenever you do something one the Internet?



Limitations: Reduced to 1-Dimension Spectrum

- Data categorized diverse people in 1-Dimension urban-rural spectrum
  - Doing so may result in information loss



Limitations

- Only geographically tagged tweets were used for network-based algorithm
  - Can be another source of bias: who will be more likely to use geotag?

- Confined data to the tweets in contiguous United States only
  - other regions might have different cultures, meaning of mentions, and so on
  - they can result in somewhat different outcomes

- How to mitigate the structural bias is left as a future researchers' mission

Questions and Answers