AIMSS Journal Club: October 27, 5:00 - 6:00 pm

Virtual on ZOOM - link @ linktr.ee/aimss



Topic:

Using Twitter for Modelling COVID-19 Hot-Spots

Presentation Overview

1. Introduction to the problem
2. Paper: *Spatial variation in search engine queries*
3. Application to COVID-19 / Experimental plans

Intro to Problem

- What are two things that both spread?
  - Information 📚🧠
  - Disease 🦟🦠
- Surely there is something to be learned in how information about disease spreads
- We are in a unique situation currently...



Data



- For the first time during a global pandemic, we have microblogging sites like Twitter
- Datasets are appearing that use the Twitter API to extract tweets with COVID-related phrases and hashtags
- Include tweet content, date/time, location***, etc...

— Active keywords and hashtags (archive: keywords.tsv) : "corona", "#corona", "coronavirus", "#coronavirus", "covid", "#covid", "covid19", "#covid19", "covid-19", "#covid-19", "sarscov2", "#sarscov2", "sars cov2", "sars cov 2", "covid_19", "#covid_19", "#ncov", "ncov", "#ncov2019", "ncov2019", "2019-ncov", "#2019-ncov", "pandemic", "#pandemic" "#2019ncov", "2019ncov", "quarantine", "#quarantine", "flatten the curve", "flattening the curve", "#flatteningthecurve", "#flattenthecurve", "hand sanitizer", "#handsanitizer", "#lockdown", "lockdown", "social distancing", "#socialdistancing", "work from home", "#workfromhome", "working from home", "#workingfromhome", "ppe", "n95", "#ppe", "#n95", "#covidiots", "covidiots", "herd immunity", "#herdimmunity", "pneumonia", "#pneumonia", "chinese virus", "#chinesevirus", "wuhan virus", "#wuhanvirus", "kung flu", "#kungflu", "wearamask", "#wearamask", "wear a mask", "vaccine", "vaccines", "#vaccine", "#vaccines", "corona vaccine", "corona vaccines", "#coronavaccine", "#coronavaccines", "face shield", "#faceshield", "face shields", "#faceshields", "health worker", "#health worker", "health workers", "#healthworkers", "#stayhomestaysafe", "#coronaupdate", "#frontlineheroes", "#coronawarriors", "#homeschool", "#homeschooling", "#hometasking", "#masks4all", "#wfh", "wash ur hands", "wash your hands", "#washurhands", "#washyourhands", "#stayathome", "#stayhome", "#selfisolating", "self isolating"

Plan

- (This project is in early stages — no experimental results yet)
- We want to use Twitter data to model, over time, the most active centers of COVID information exchange
  - We assume this is related to the spread
- To do this, we will be drawing inspiration the following paper…

Spatial variation in search engine queries

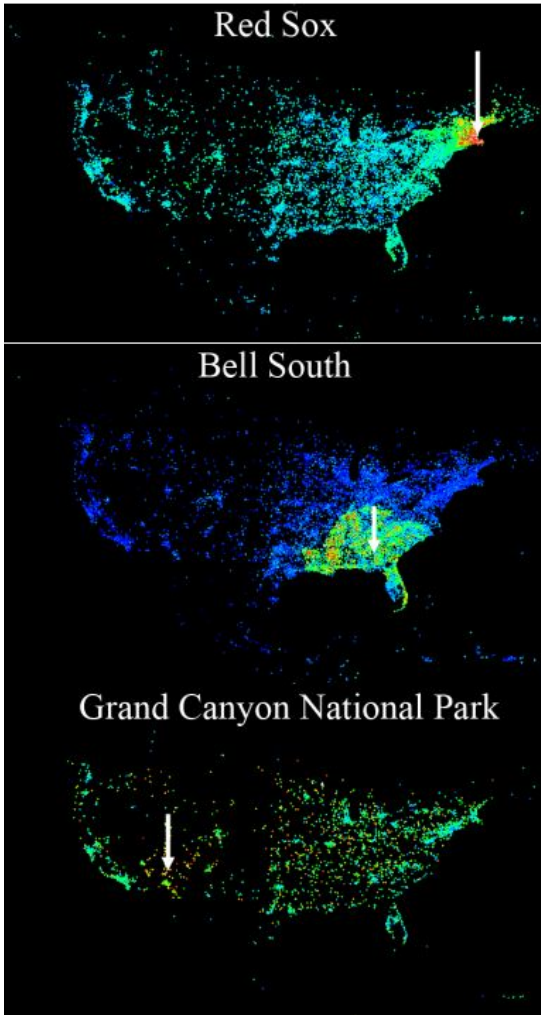
Backstrom, Klein, Kumar, and Novak, 2008

Introduction





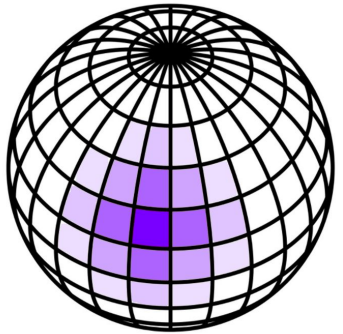
- Consider a “Term of Interest” or TOI one might search in a search engine
- It’s possible this TOI is heavily influenced by geography
  - e.g., “Edmonton Oilers”
- Determining the epicentre of search for a query is of interest (especially to advertisers)
- Determining the “dispersion” of a query can contain useful information about its “locality”



How do we model this?

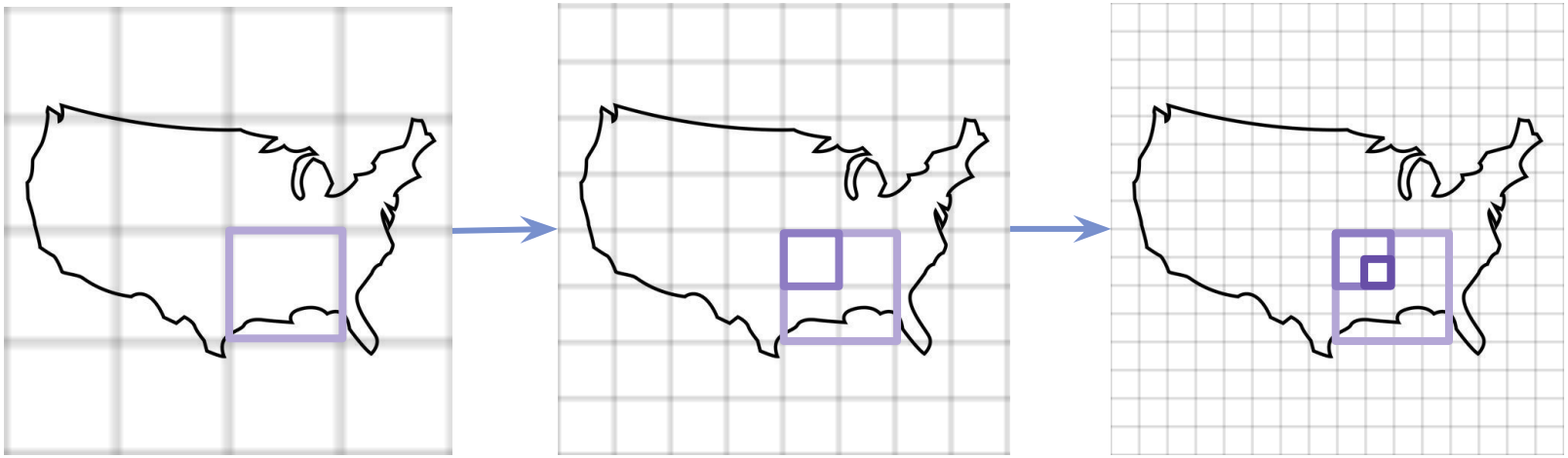
Methodology

- The authors propose a probabilistic framework for identifying the center and dispersion of a TOI using users' query logs
- First, divide the surface of the earth into a series of cells based on latitude and longitude
- Assume for each cell x & query q combination, there is a probability $p_{x,q}$ that a random search from x will be q
- From cell z with highest p (center), surrounding $p_{x,q}$ should taper off ...



Methodology (Step 1)

- To find the *center* for a query, $p_{x,q}$ is calculated for a coarse mesh, then a finer mesh, then a finer mesh, … until we find z



Methodology (Step 2)

- For each q, we also define C and α:
  - C is the frequency of the search at its center
  - α indicates how quickly the $freq(q)$ tapers as we move away from z

- That means, for each q, the probability that user with distance d from the query's center issued q is:

$$Cd^{-\alpha}$$

(large α = local,
small α = national/global)

Methodology (Step 2)

- C and a must be tuned for each query such that Cd^{-a} is actually the probability that a query-issuing user of distance d from the center issued query q
- The authors determine that the log-likelihood as a function of C and a is unimodal, meaning they can both be optimized fairly easily at the same time
  - Authors' (unoptimized) implementation converges to maximum in under 30 seconds per q

Newspaper	α
The Wall Street Journal	0.111327
USA Today	0.263173
The New York Times	0.304889
New York Post	0.459145
The Daily News	0.601810
Washington Post	0.719161
Los Angeles Times	0.782538
The Star Ledger	0.998462
Detroit Freepress	1.068055
San Francisco Chronicle	1.091030
Chicago Tribue	1.102554
Philadelphia Inquirer	1.140618
Chicago Sun Times	1.165482
The Boston Globe	1.171179
The Arizona Republic	1.284957
Dallas Morning News	1.286526
Houston Chronicle	1.289576
Star Tribune	1.337356

Evaluation

- The authors tested their localization model using queries with ground truth locations (sports teams, U.S. senators, etc.)
- Used function $d(a_q, b_q)$, the distance between the actual center of q and the computed center of q
- Localized almost all baseball team names to within 60 miles of their respective home cities

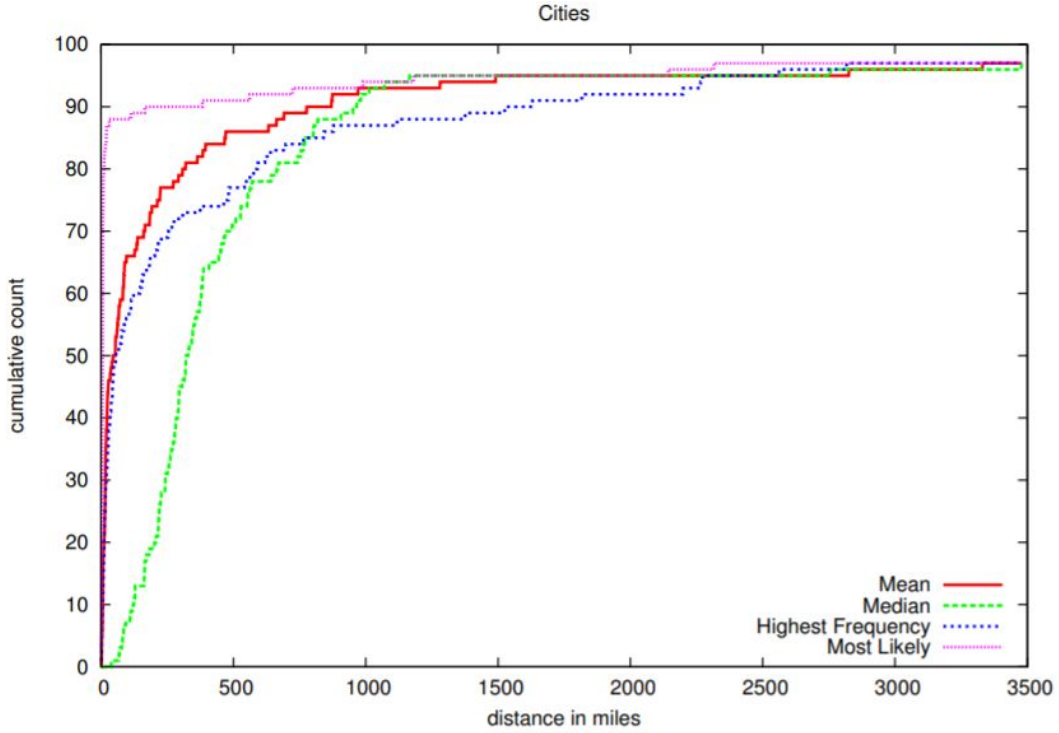


Figure 3: Performance of algorithms on high-population U.S. cities. (cumulative distribution of error distances)

Mean = weighted center of gravity over all instances of *q*

Median = median latitude and longitude of all *q*

Highest frequency = highest local density of instances of *q*

Most likely = this paper's work

Extensions to Algorithm

- It is easy to find queries that vary over time, e.g.:
  - The *center* stays the same, but C or α changes
  - The *center* itself changes
- When a center changes, we can observe this change if we consider the data in *time-slices* (e.g., every 24 hours offset by 1 hour)
  - We can often see α getting slowly smaller...

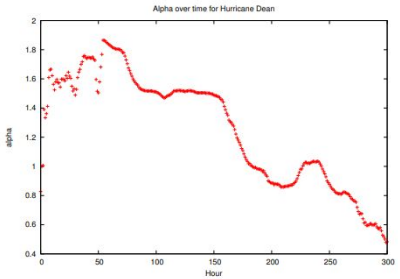




Figure 4: The path of Hurricane Dean's storm center, moving west through the Caribbean, alongside the smoothed path of query centers for "Hurricane Dean."

Extensions to Algorithm

- Sometimes, having one center isn't enough

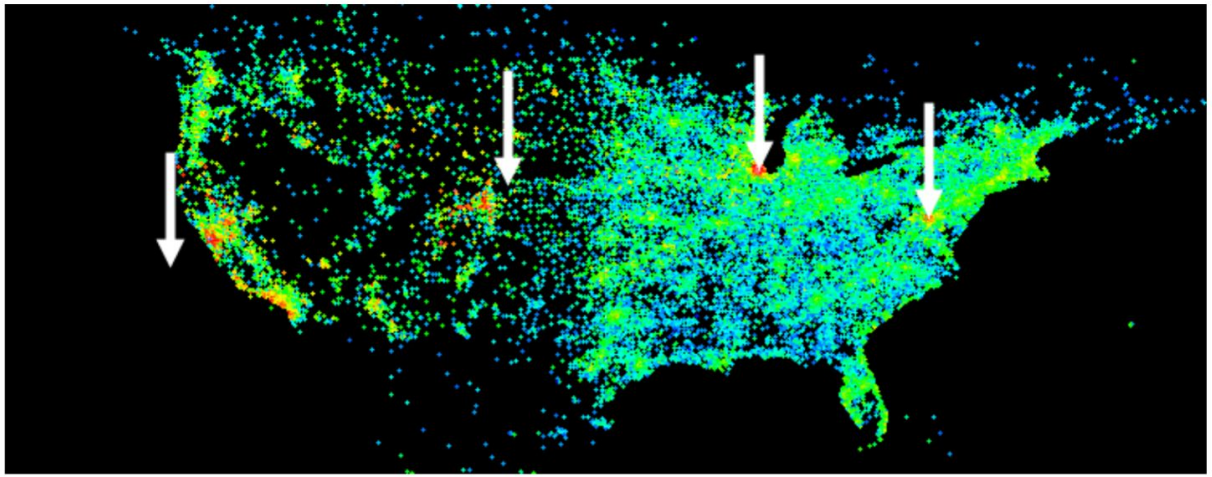
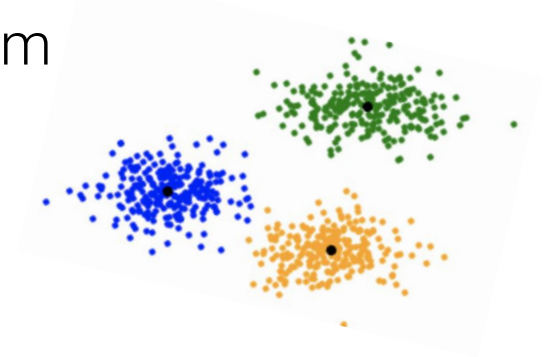


Figure 6: Multiple centers for the query "United Airlines."

Extensions to Algorithm

- Sometimes, having one center isn't enough
- This generative model can be extended such that multiple centers are placed, each with their own C and α
- This makes optimization harder — authors suggest a heuristic based on the *K*-means clustering algorithm



Extensions to Algorithm

- Steps:
  a. Place K centers at random on map
  b. Optimize each center based on $p_{x,q}$ as we did before
  c. Look at every geographic point (q), determine which of the centers gives it the highest probability according to Cd^{-a}
  d. Re-optimize each center, now only considering its subset of qs
  e. Repeat steps b-d until convergence

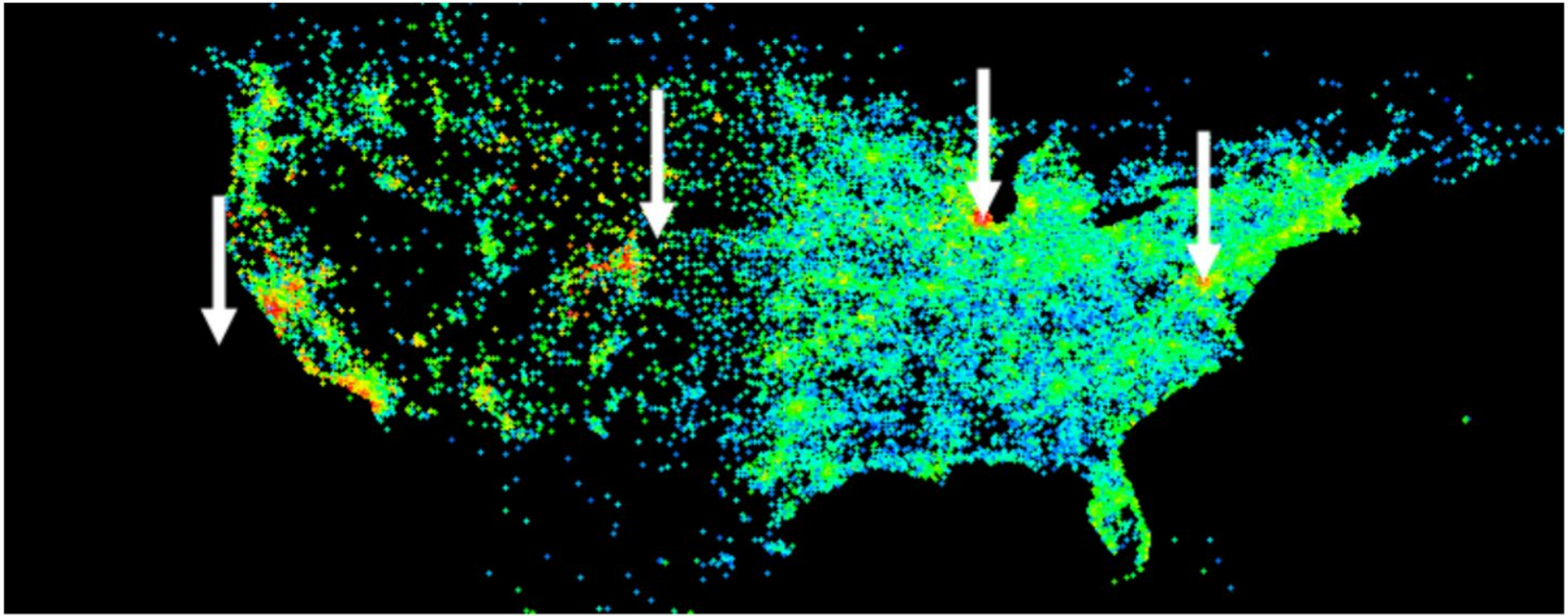


Figure 6: Multiple centers for the query "United Airlines."

Application to COVID-19

Data

- Lots of people are tweeting about COVID
- There is a small subset of Twitter users that use the “Tweet Geo-Location” setting — meaning we have the exact latitude and longitude of the place they tweeted
- The CORONAVIRUS (COVID-19) GEO-TAGGED TWEETS DATASET contains 246,699 located tweets

— Active keywords and hashtags (archive: keywords.tsv) : "corona", "#corona", "coronavirus", "#coronavirus", "covid", "#covid", "covid19", "#covid19", "covid-19", "#covid-19", "sarscov2", "#sarscov2", "sars cov2", "sars cov 2", "covid_19", "#covid_19", "#ncov", "ncov", "#ncov2019", "ncov2019", "2019-ncov", "#2019-ncov", "pandemic" "#pandemic" "#2019ncov", "2019ncov", "quarantine", "#quarantine", "flatten the curve", "flattening the curve", "#flatteningthecurve", "#flattenthecurve", "hand sanitizer", "#handsanitizer", "#lockdown", "lockdown", "social distancing", "#socialdistancing", "work from home", "#workfromhome", "working from home", "#workingfromhome", "ppe", "n95", "#ppe", "#n95", "#covidiots", "covidiots", "herd immunity", "#herdimmunity", "pneumonia", "#pneumonia", "chinese virus", "#chinesevirus", "wuhan virus", "#wuhanvirus", "kung flu", "#kungflu", "wearamask", "#wearamask", "wear a mask", "vaccine", "vaccines", "#vaccine", "#vaccines", "corona vaccine", "corona vaccines", "#coronavaccine", "#coronavaccines", "face shield", "#faceshield", "face shields", "#faceshields", "health worker", "#health worker", "health workers", "#healthworkers", "#stayhomestaysafe", "#coronaupdate", "#frontlineheroes", "#coronawarriors", "#homeschool", "#homeschooling", "#hometasking", "#masks4all", "#wfh", "wash ur hands", "wash your hands", "#washurhands", "#washyourhands", "#stayathome", "#stayhome", "#selfisolating", "self isolating"

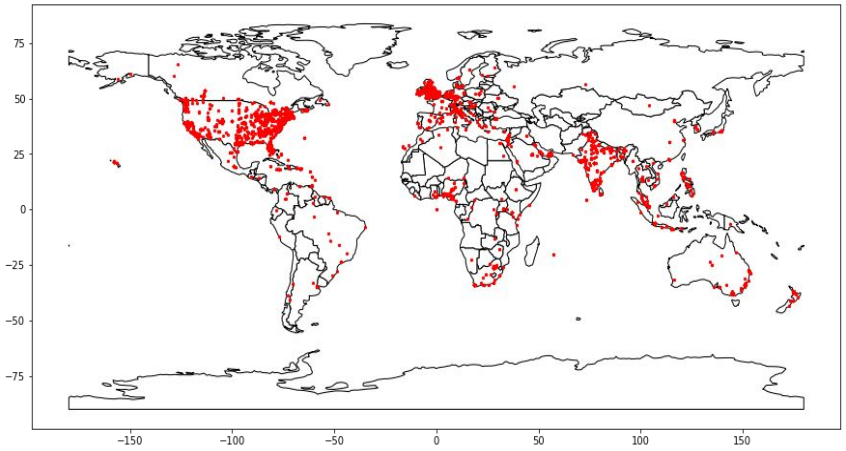






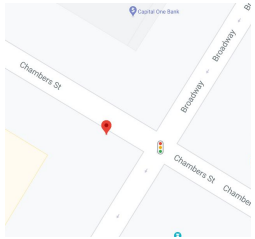
Data

- To replicate the paper, we will be restricting our analysis to the continental United States




```
[23]  1 locator = Nominatim(user_agent="myGeocoder")
      2 coordinates = "40.7142,-74.0064"
      3 location = locator.reverse(coordinates)
      4 location.raw
```

```
{'address': {'city': 'New York',
  'city_block': 'Manhattan Community Board 1',
  'country': 'United States of America',
  'country_code': 'us',
  'county': 'New York County',
  'house_number': '80',
  'neighbourhood': 'Civic Center',
  'postcode': '10007',
  'road': 'Chambers Street',
  'state': 'New York',
  'suburb': 'Manhattan'}
```



Plan

- (Authors did not provide implementations as this was a partially commercial project…)
- All tweets made that contain COVID keywords/phrases can be seen as analogous to a COVID “query” in a search engine
  - Have date/time, latitude and longitude for each tweet
- First, generate some synthetic datasets with predefined center(s)
  - Used to test how good our implementation for single and multiple centers is, get $d(a_q, b_q)$

Plan

- With the COVID data, we will first use the “single center” algorithm (how does a change?), then extend to the multiple center algorithm
- Generate multiple center models with many different values of k
  - How do we choose k?
- Once we have the multiple centers, we will evaluate how well the frequencies C correspond with the cumulative number of cases in the area

Other things to explore

- COVID likely contains multiple centers AND changes over time
  - Attempt to combine their temporal smoothing and *k*-means procedures
  - PROBLEM: *k*-means clustering for multiple centers sensitive to initial random placements

- Each COVID tweet doesn't necessarily contain the same information (#COVID19 vs. #wearamask vs. #vaccines)
  - Model each as a separate query or categorize
  - Local vs. national!

— Active keywords and hashtags (archive: keywords.tsv) : "corona", "#corona", "coronavirus", "#coronavirus", "covid", "#covid", "covid19", "#covid19", "covid-19", "#covid-19", "sarscov2", "#sarscov2", "sars cov2", "sars cov 2", "covid_19", "#covid_19", "#ncov", "ncov", "#ncov2019", "ncov2019", "2019-ncov", "#2019-ncov", "pandemic", "#pandemic" "#2019ncov", "2019ncov", "quarantine", "#quarantine", "flatten the curve", "flattening the curve", "#flatteningthecurve", "#flattenthecurve", "hand sanitizer", "#handsanitizer", "#lockdown", "lockdown", "social distancing", "#socialdistancing", "work from home", "#workfromhome", "working from home", "#workingfromhome", "ppe", "n95", "#ppe", "#n95", "#covidiots", "covidiots", "herd immunity", "#herdimmunity", "pneumonia", "#pneumonia", "chinese virus", "#chinesevirus", "wuhan virus", "#wuhanvirus", "kung flu", "#kungflu", "wearamask", "#wearamask", "wear a mask", "vaccine", "vaccines", "#vaccine", "#vaccines", "corona vaccine", "corona vaccines", "#coronavaccine", "#coronavaccines", "face shield", "#faceshield", "face shields", "#faceshields", "health worker", "#health worker", "health workers", "#healthworkers", "#stayhomestaysafe", "#coronaupdate", "#frontlineheroes", "#coronawarriors", "#homeschool", "#homeschooling", "#hometasking", "#masks4all", "#wfh", "wash ur hands", "wash your hands", "#washurhands", "#washyourhands", "#stayathome", "#stayhome", "#selfisolating", "self isolating"

Thank you!

Questions?

