Detecting Spatial Events in Geo-tagged Text

Presented by Sarah Davis, Hamed Mirzaei

CMPUT 692 – Fall 2020

Table of contents

- Problem description
- Introduction to project
- Related works
- Algorithm
- Datasets
- Evaluations
- Problems

Problem description

- Given a set of coordinates, how can we make sense of the spatial pattern that it shows? How do we describe "hot-spots" in the data?



Introduction to project

- Given a collection of geo-tagged events...
- Where certain events are taking place and how the information disseminate (centre(s) and spread/dispersion)?
- Focus on multiple centered events, based on Backstrom's single center method
- Tracking centers and spread of COVID-19 using Twitter's tweets is one example

Related Works
Backstrom et al. 2008

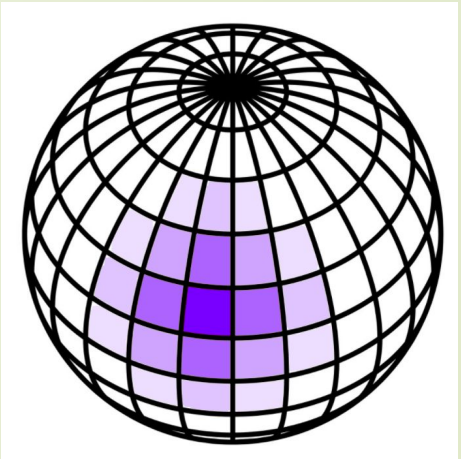
- Consider a "Term of Interest" or TOI one might search in a search engine
- It's possible this TOI is heavily influenced by geography
  - e.g., "Edmonton Oilers"
- The authors propose a probabilistic framework for identifying the center and dispersion of a TOI using users' query logs





Related Works
Backstrom et al. 2008

- 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
- Subdivide cell z with highest p until the true "center" is found



$$\sum_{i \in S} \log C d_i^{-\alpha} + \sum_{i \notin S} \log(1 - C d_i^{-\alpha})$$

Related Works
Backstrom et al. 2008

- This work can be extended to handle multiple centers

- Steps (based on k-means clustering):
  a. Place k centers at random on map
  b. Optimize each center as we did before
  c. Look at every geographic point (q), determine which of the centers gives it the highest probability
  d. Re-optimize each center, now only considering its subset of qs
  e. Repeat until convergence

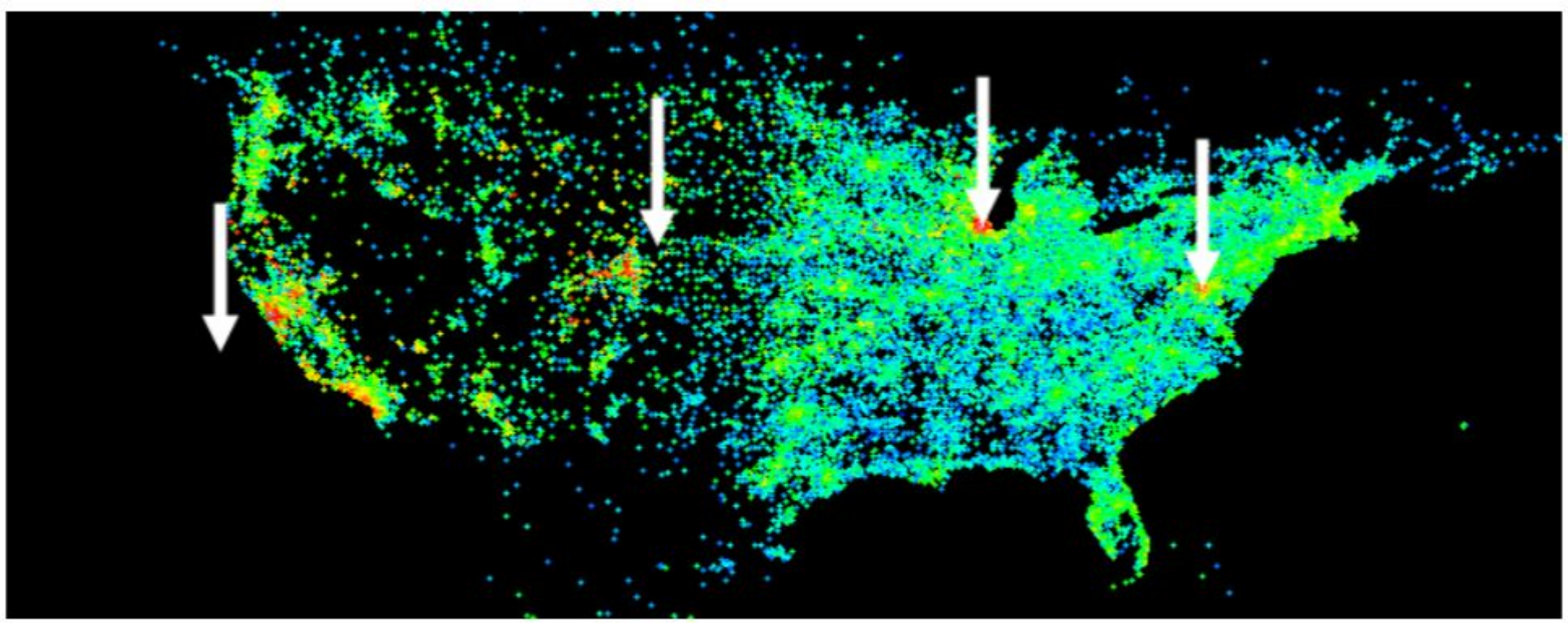
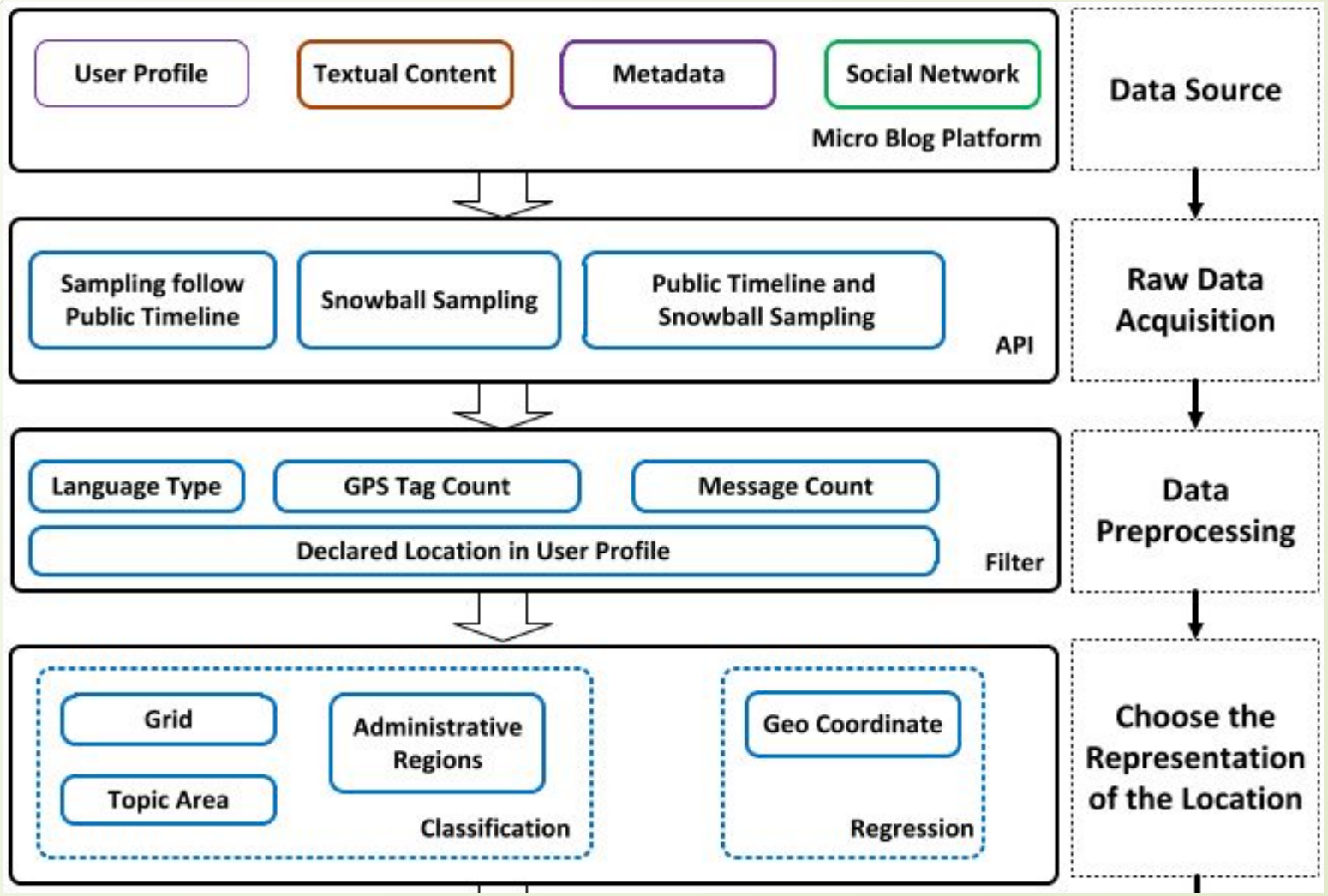


Figure 6: Multiple centers for the query “United Airlines.”

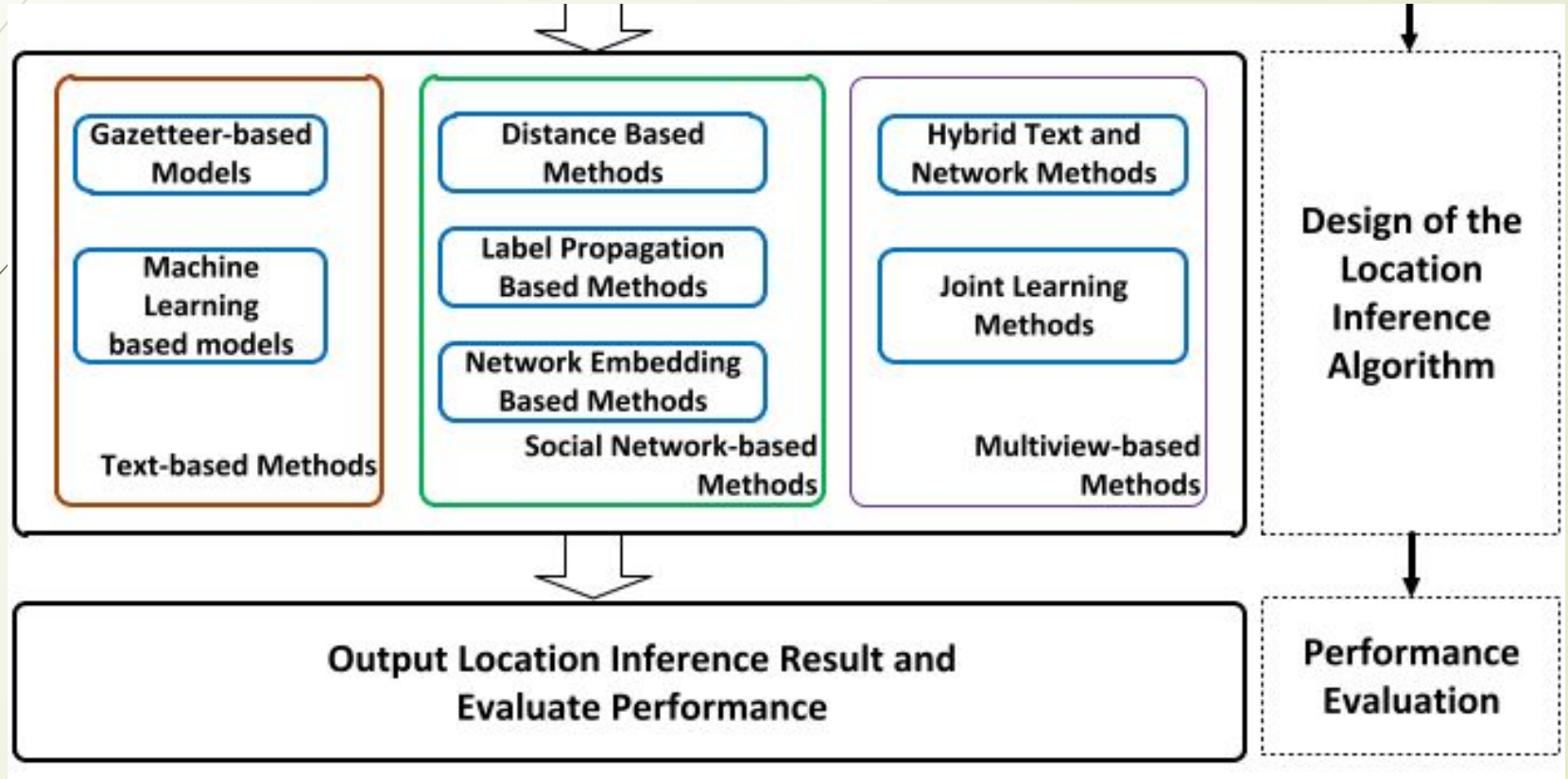
Related Works
Others

- Frequency-based methods
- Density-based methods
- Probabilistic-based methods
- Hybrid methods
- All these methods follow the same framework

Related Works
Framework



Related Works
Framework - cont.



Our Contributions

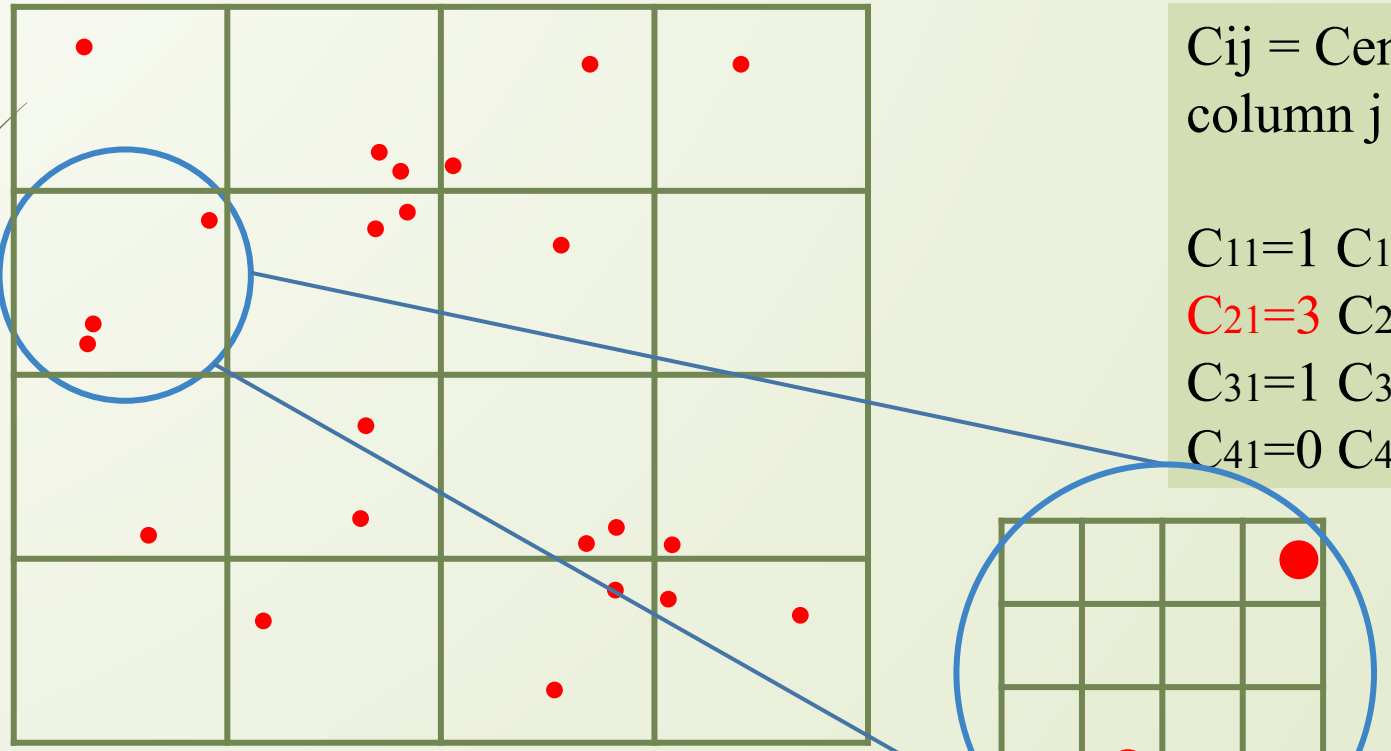
1. Alternative way of thinking about multiple centers
   - Instead of picking number of centers, pick a frequency threshold
   - Find all possible centers based on the activity

2. Flexible cells algorithm
   - Early termination of bigger cells
   - Detects centers at borders of cells
   - Pre-calculated distances

1. Synthetic data generator

Algorithm
Backstrom's algorithm

- Full map is divided into cells of a specific size (starts with 2 and ends in 0.1)

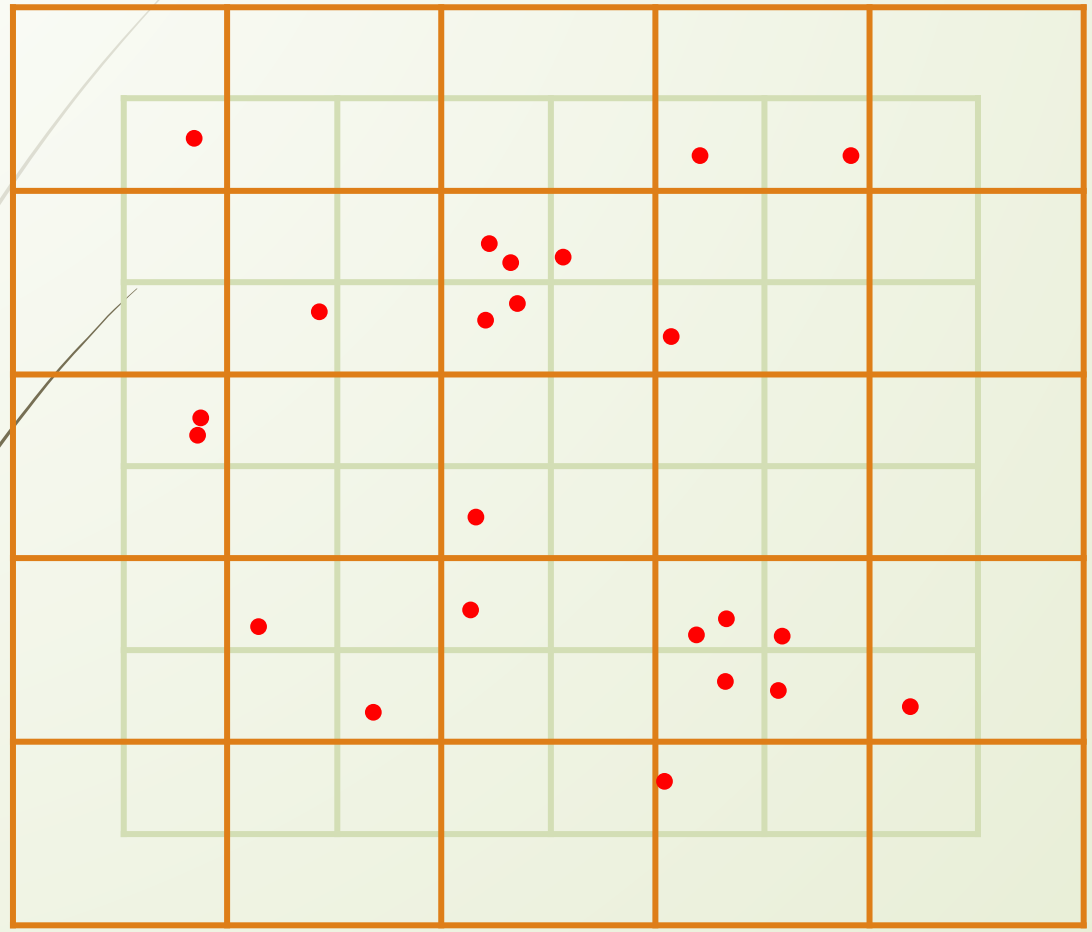


Detecting Spatial Events in Geo-tagged Text

Cij = Center of cell at row i and column j

$C_{11}=1$ $C_{12}=2$ $C_{13}=2$ $C_{14}=1$
$C_{21}=3$ $C_{22}=2$ $C_{23}=1$ $C_{24}=0$
$C_{31}=1$ $C_{32}=2$ $C_{33}=2$ $C_{34}=1$
$C_{41}=0$ $C_{42}=1$ $C_{43}=2$ $C_{44}=2$

Algorithm
Backstrom's drawback



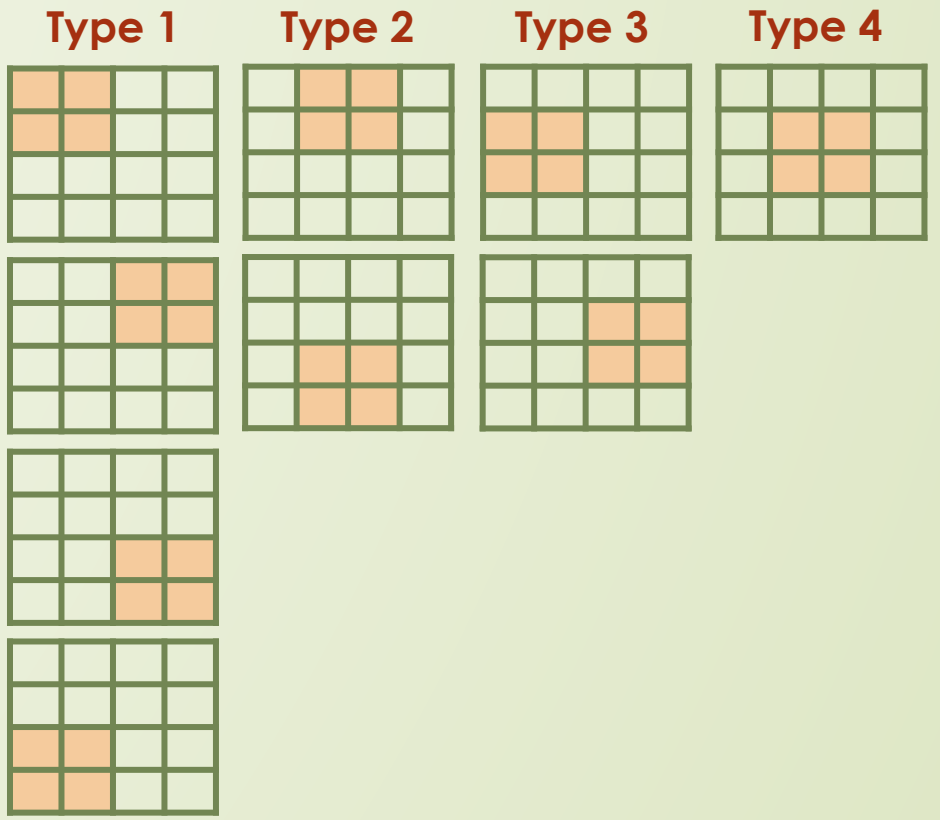
C_{ij} = Center of cell at row i and column

C_{11}=1 C_{12}=0 C_{13}=0 C_{14}=2 C_{15}=0
C_{21}=0 C_{22}=1 C_{23}=5 C_{24}=1 C_{25}=0
C_{31}=2 C_{32}=0 C_{33}=1 C_{34}=0 C_{35}=0
C_{41}=0 C_{42}=2 C_{43}=1 C_{44}=5 C_{45}=1
C_{51}=0 C_{52}=0 C_{53}=0 C_{54}=1 C_{55}=0

Flexible Cells Algorithm
Solution - Part1: narrow it down recursively

- Start with a cell of size full map
- Divide it into 9 possible cells of size half
- Do it for each new cell
- Termination: when, by halving, new cells won't cover the SIZE_THR

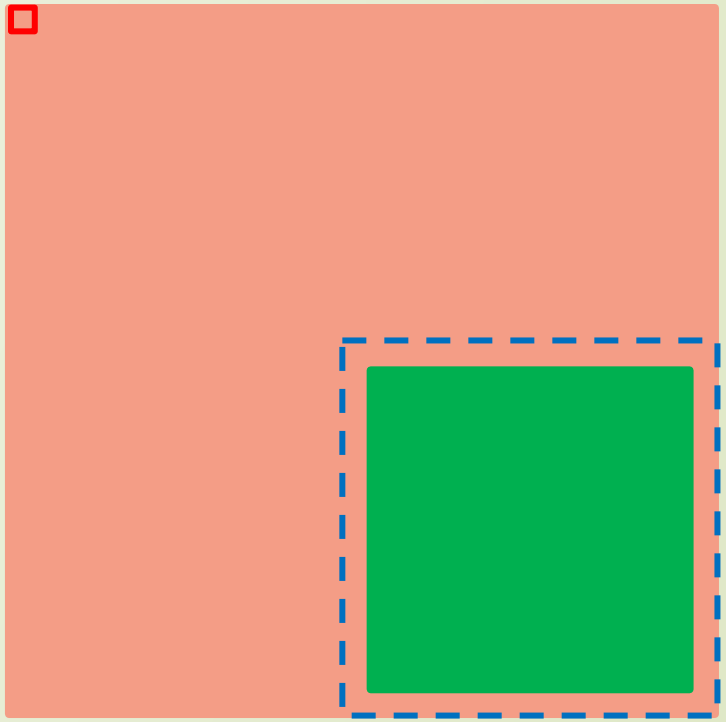
- In order to remove duplicate cells
  - Type1: check all 9 possible cells
  - Type2: check 3 vertically middle cells
  - Type3: check 3 horizontally middle cells
  - Type4: check only the middle cell



Flexible Cells Algorithm
Solution - Part2: find best marginal cover

- Suppose we are looking for the green square and we come up with the light red square as one of our found cells

- For each found cell:
  - Divide it into a grid of smaller cells, each of size 0.1*SIZE_THR (red cell)
  - Find the square of size 1.2*SIZE_THR (i.e. 12 smaller cell) in the grid with the highest frequency (the blue dashed square)

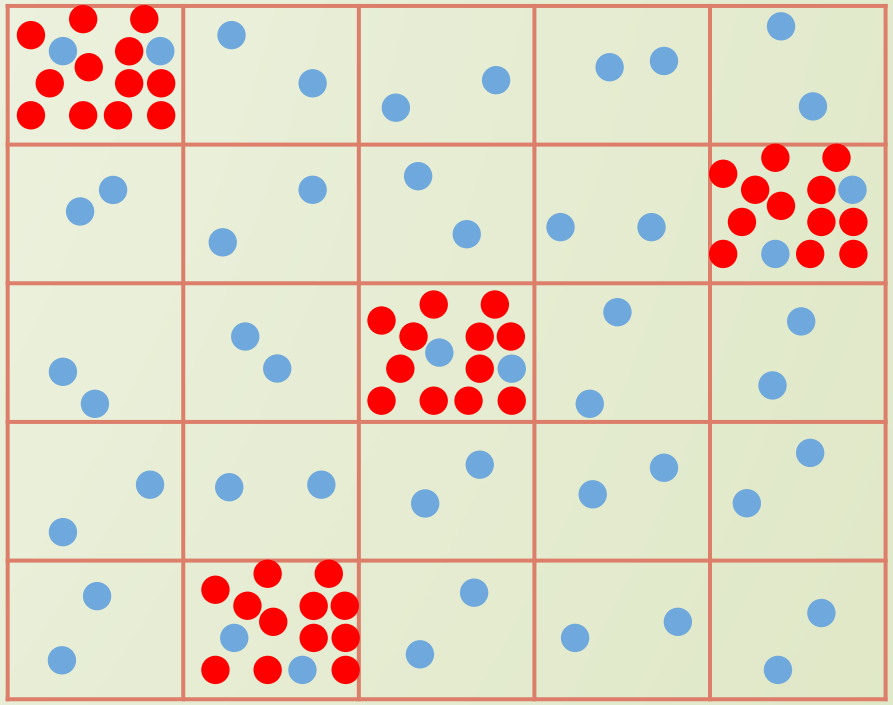


Datasets
Types

- Two types of datasets
  - User-dependent (one per user per location):
    - Covid-19
  - User-independent
    - US Accidents, UK Accidents, World Earthquakes
- An algorithm for generating synthetic data is implemented

Datasets
Synthetic data generation algorithm

- Parameters:
  - NOE (number of events)
  - NOC (number of centers)
  - MCF (min center frequency)
  - CS (cell size)
- Random generation of distinct centers
- Uniform distribution



Datasets
Covid-19



- 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 227,443 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"

Datasets
Covid-19 - cont.

- Final curation of Covid-19 dataset after removing noises:
  - 227443 -> 102056 (restrict to continental USA)
  - 102056 -> 44862 (one tweet per person per location)

- We were going to include Canada to have more data, but doing so only added ~4000 tweets

Evaluation

- In our synthetic data, for various cell sizes (2 -> 0.5 degrees), our centers was placed in the "ground truth" cells every single time
- We have run this algorithm on our COVID data as well
- Demonstration

Problems

1. This is a very computationally heavy task - running experiments can take days on big datasets and small cells of interest
2. The multiple center algorithm was merely suggested by Backstrom, they didn't provide any implementation - we were going off their vague instructions and had to do a lot of problem solving
3. We have our centers for COVID data too, how do we evaluate it? Something to do with COVID case data?