# Analysis of GSM Data in Conjunction with Twitter Data for Understanding Social Behaviours in Senegal

By
Leo Liu

Supervisor
Dr. Monica Wachowicz

# We are entering into a Big Data era

Every minute in 2014

- Google: 4,000,000 searches
- Email: 200,000,000 sent
- Twitter: 280,000 tweets sent

- Almost everything we do leaves a digital footprint (or data)

Important Big Data sources:

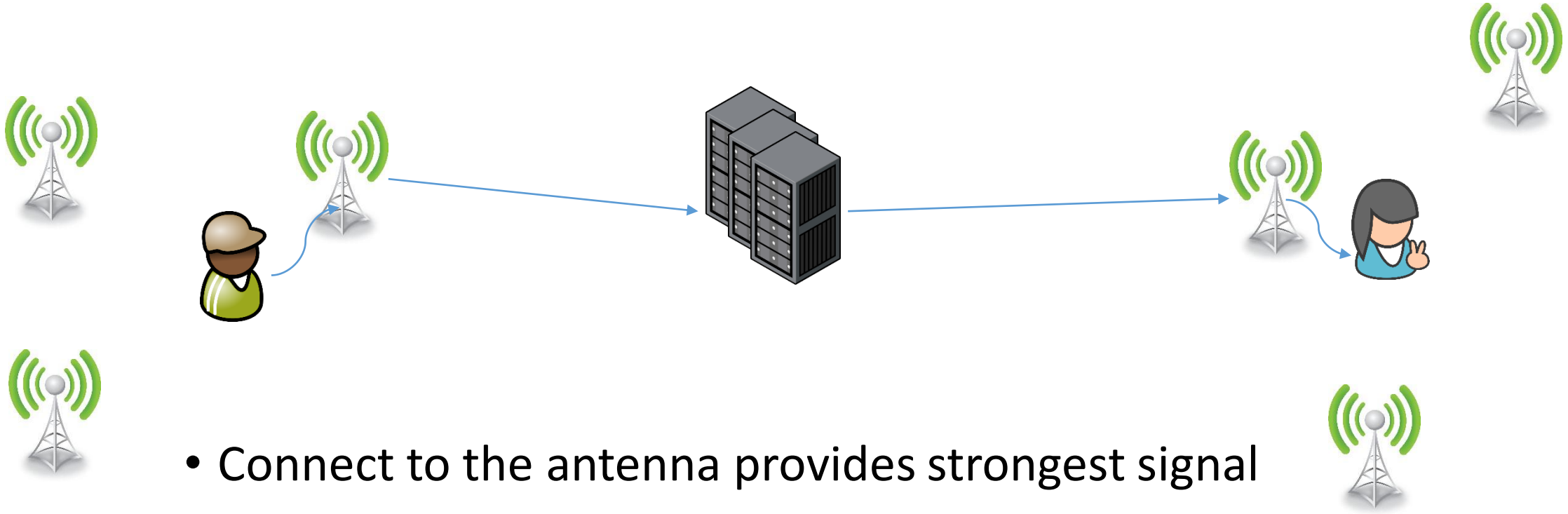
- Global System for Mobile Communications (GSM)
- Social media: Twitter

# Geo-Tweet Data:

- 32,536 geo-tagged tweets WERE collected
- In the region of Dakar, Senegal
- From February 4th to July 25th, 2013

| user | timestamp | lon | lat | text |
|---|---|---|---|---|
| souleey_man | 139147200 | -17.46445711 | 14.75180725 | fatou vient dans tg meu kaagne leu some hahaha |
| cheikhou_og | 139147404 | -17.47197470 | 14.75295200 | @iammaguee: julo lilagn nane mbiiww .. mdr |
| cheikhou_og | 139147425 | -17.47197470 | 14.75295200 | @iammaguee: cheikhou lilanla ni ? doul way lol |
| souleey_man | 139147510 | -17.46445711 | 14.75180725 | khodioo gaaw |
| souleey_man | 139147559 | -17.46445711 | 14.75180725 | signalez le compte de @cheikhou_og haahahahahhaaaa |
| momarginal | 139148704 | -17.46602556 | 14.71670566 | @senebelle haha i gotta get up for work in 2 hours, man |

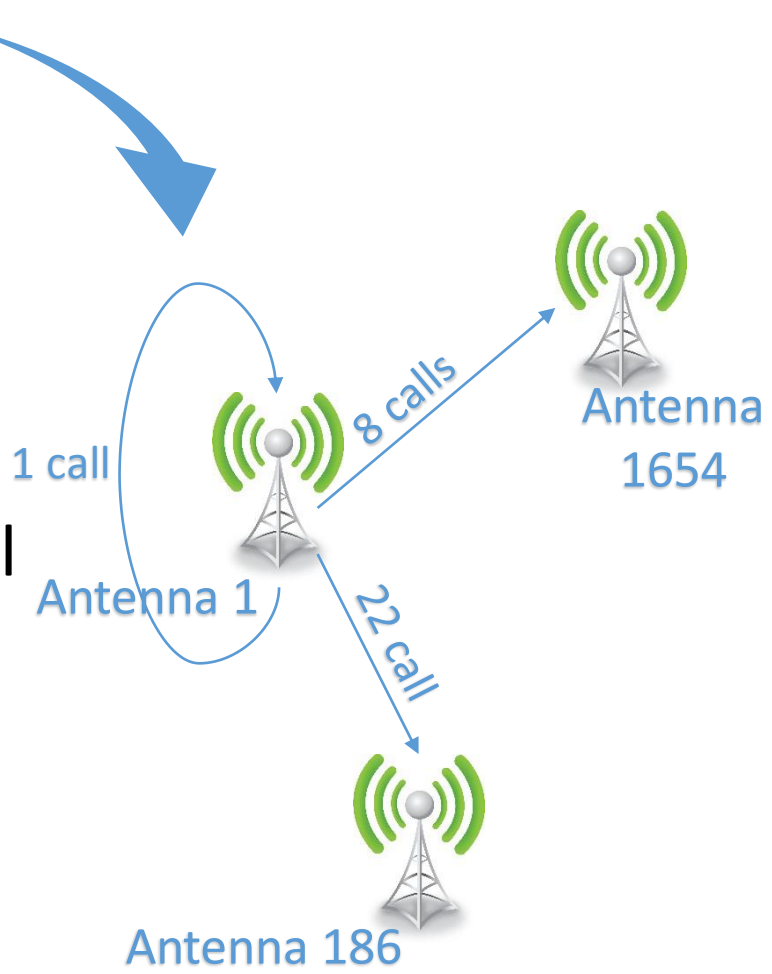
# GSM Data: When you are calling…



- Connect to the antenna provides strongest signal
- Signal strength ↓ as distance ↑
- Strongest ≈ Closest

# GSM Data:

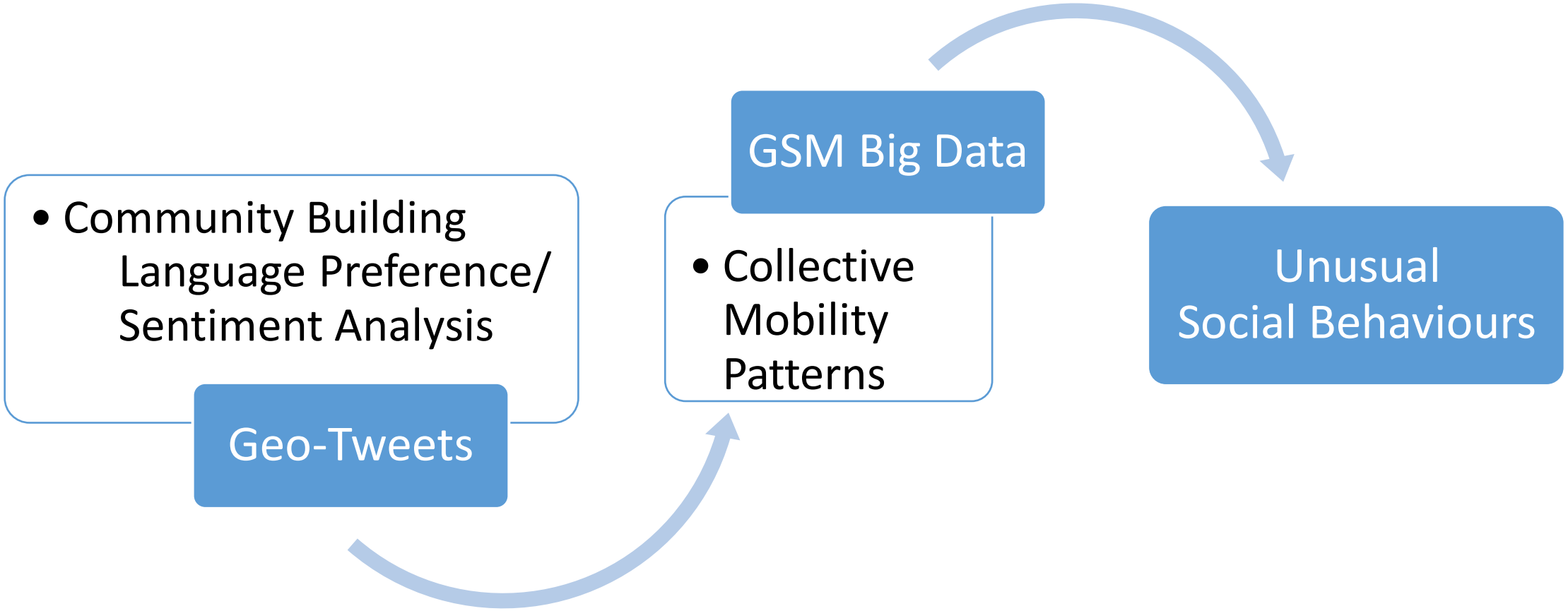
| Time | Out_Antenna | In_Antenna | Num_of_Calls |
|---|---|---|---|
| 2013-01-01 00 | 1 | 1 | 1 |
| 2013-01-01 00 | 1 | 1654 | 8 |
| 2013-01-01 00 | 1 | 186 | 22 |
| … … | … … | … … | … … |



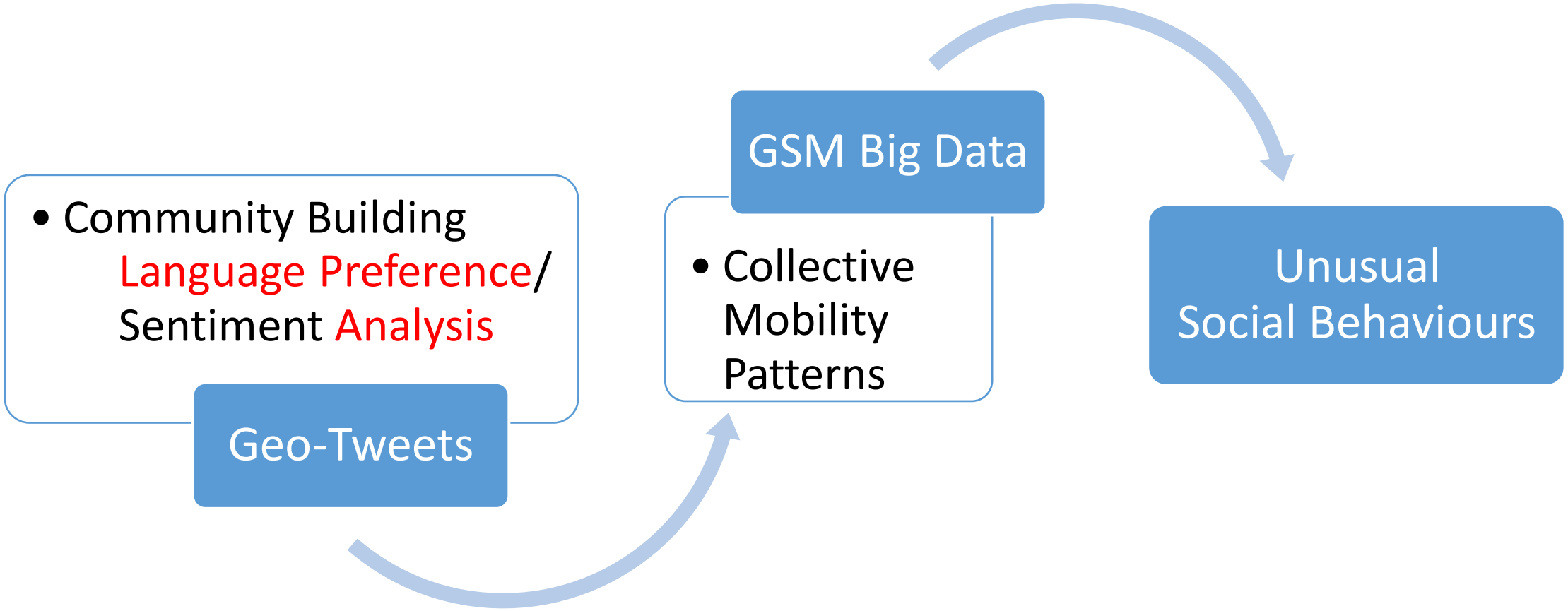

- Hourly antenna-to-antenna voice traffic in Senegal
- 1666 antennas in total, lat./ lon. given
- Data recorded for 1 Year (2013)

- Big Data!
- On average, for a month: 120,000,000 rows

# Research Objectives and Methods



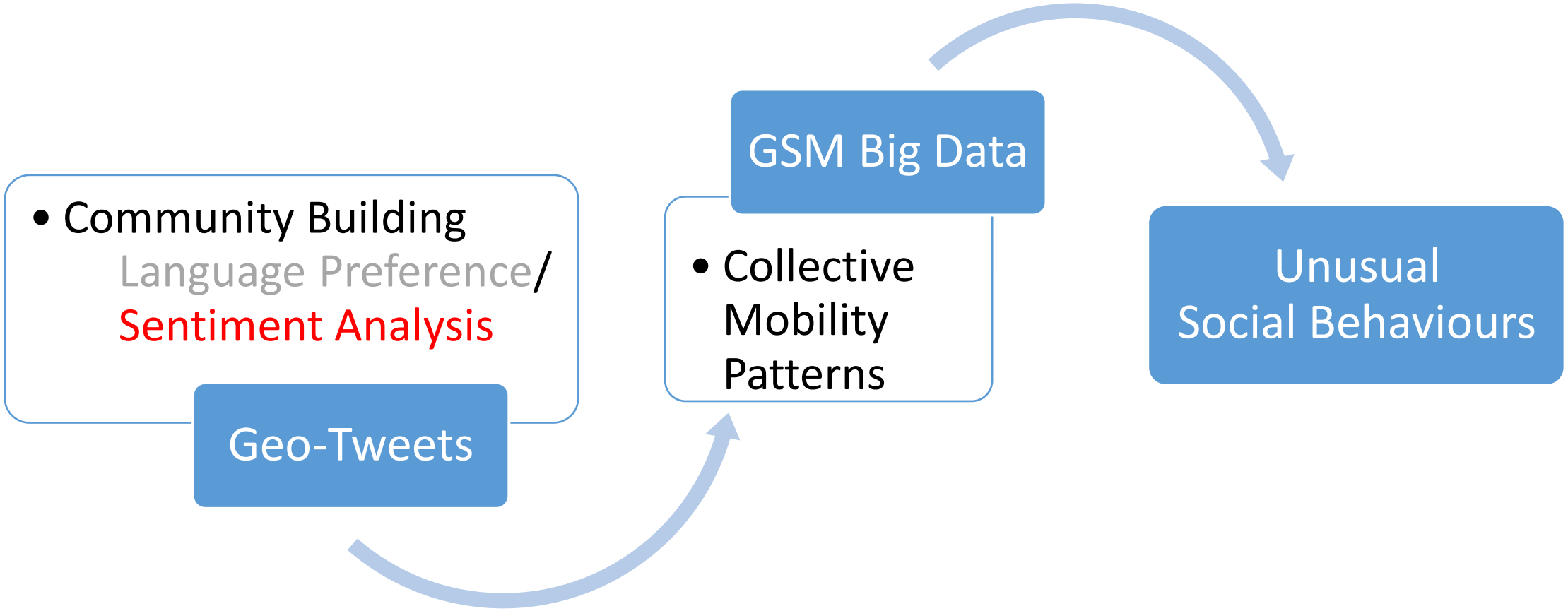
# Research Objectives and Methods



# Geo-Tweets: Language Preference Analysis

- Create 2 dictionary containing most frequently used English and French words
- Look for each word in each tweet in both dictionaries
- Clustering (English vs. French)
  Check clustering pattern: Globle Moran’s I
  Optimal distance: Incremental Spatial Autocorrelation
  Cluster building: Anselin Local Moran's I

# Research Objectives and Methods

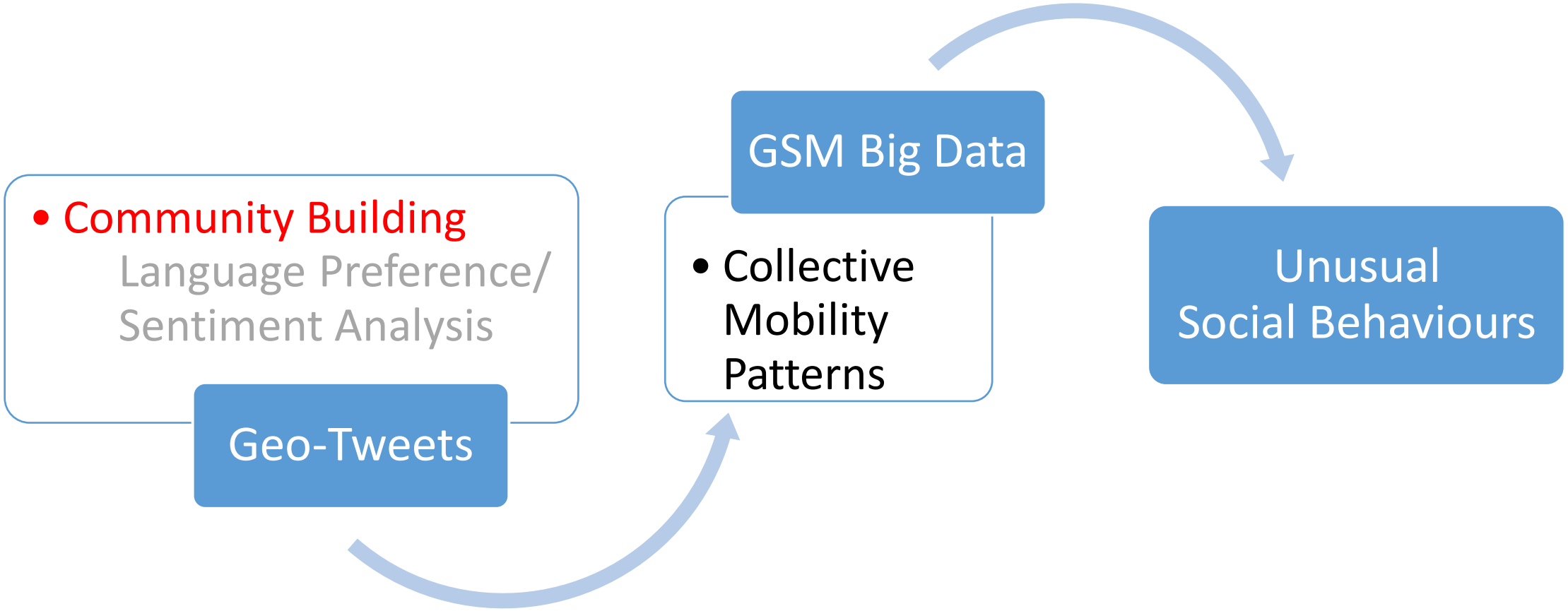


# Geo-Tweets: Sentiment Analysis

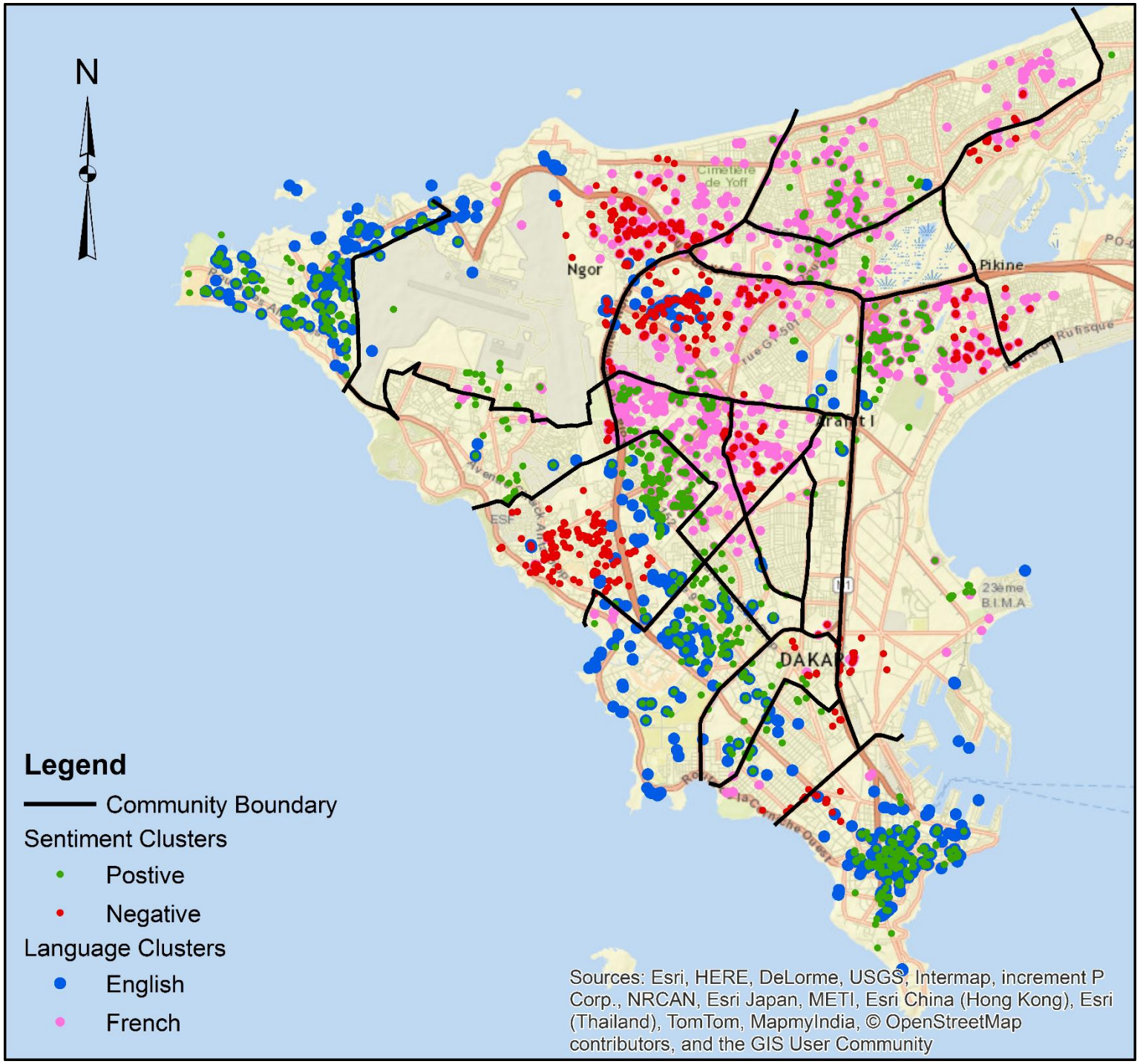
- Translate all the tweets into English
- Use SentiWordNet 3.0.0 dictionary to evaluate each tweet's sentiment word by word
- Clustering
  (positive vs. negative sentiments)
  Similar to language preference analysis

| ID | PosScore | NegScore | SynsetTerms |
|---|---|---|---|
| 1740 | 0.125 | 0 | able#1 |
| 2098 | 0 | 0.75 | unable#1 |
| 3700 | 0.25 | 0 | dissilient#1 |
| 3829 | 0.25 | 0 | parturient#2 |
| 5107 | 0.5 | 0 | uncut#7 full-length#2 |
| 5205 | 0.5 | 0 | absolute#1 |
| 5473 | 0.75 | 0 | direct#10 |
| 5599 | 0.5 | 0.5 | unquestioning#2 implicit#2 |
| 5718 | 0.125 | 0 | infinite#4 |
| 5839 | 0.5 | 0.125 | living#3 |
| 6032 | 0.25 | 0.5 | relative#1 comparative#2 |
| 6245 | 0 | 0 | relational#1 |
| 6336 | 0 | 0 | absorptive#1 absorbent#1 |
| 6777 | 0.375 | 0 | sorbefacient#1 absorbefacient#1 |

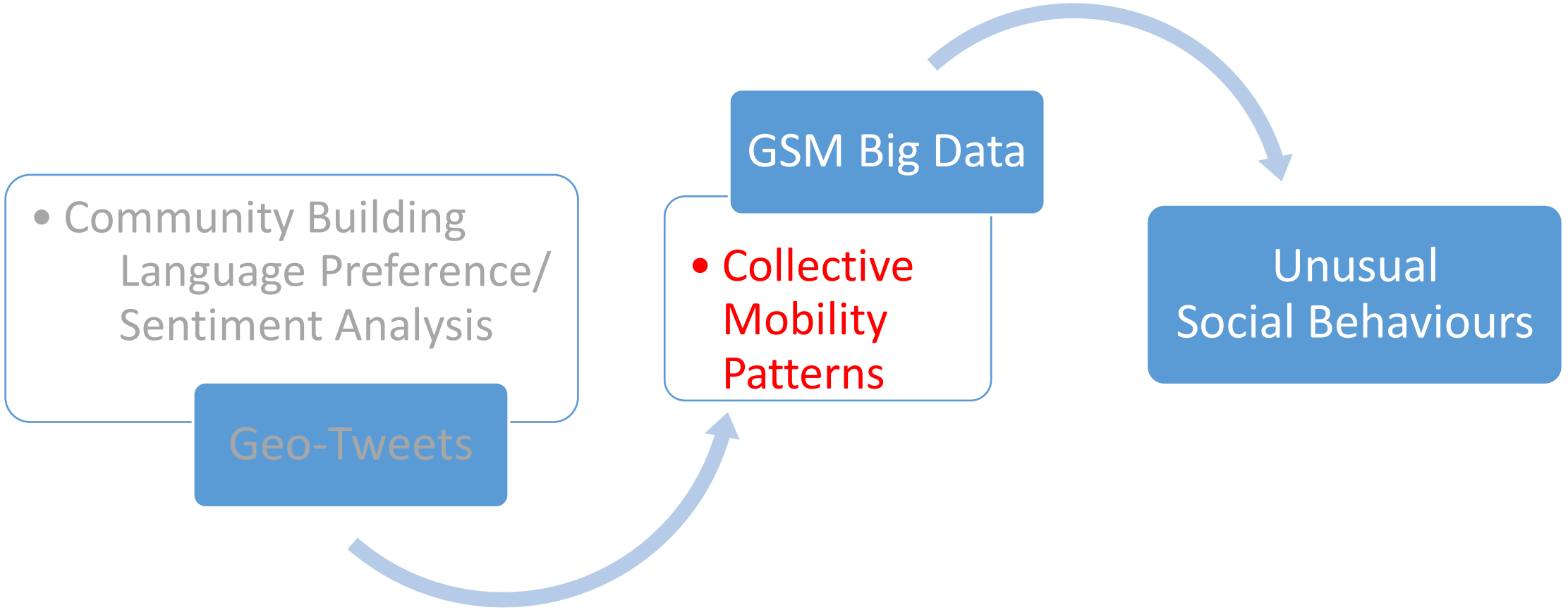
# Research Objectives and Methods



# Community Building

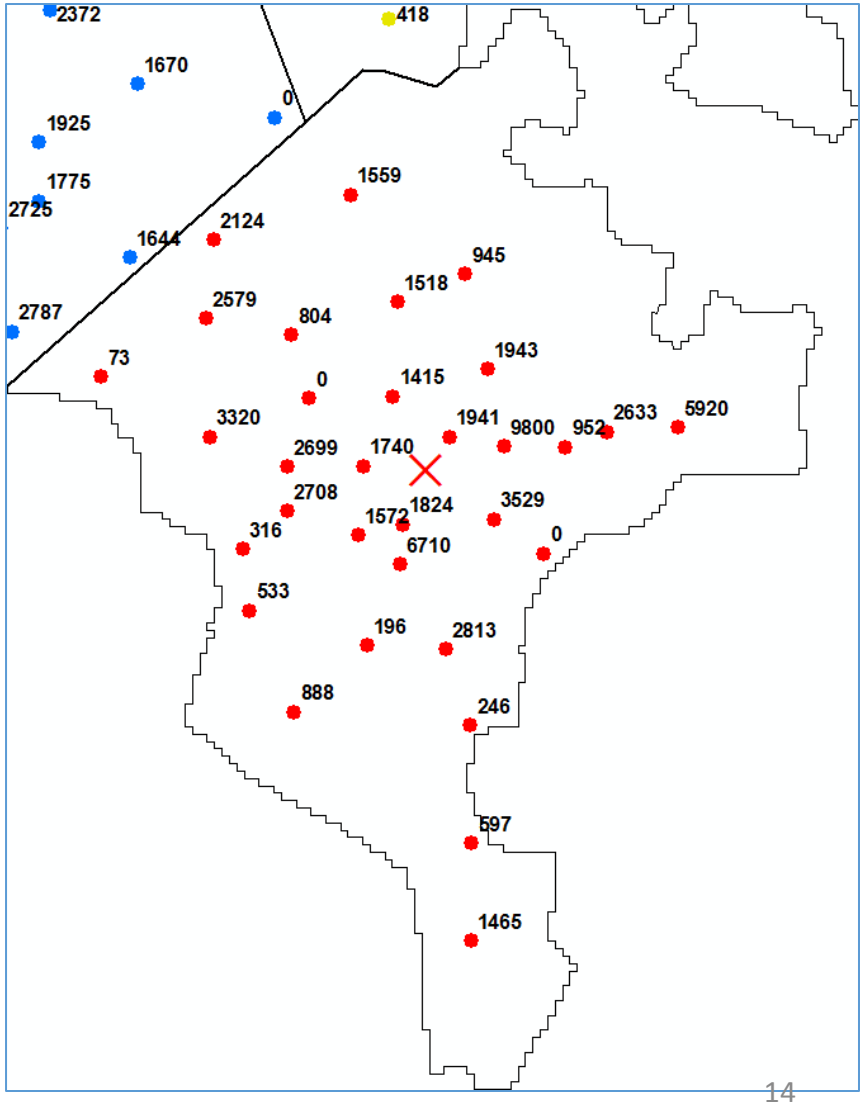


# Research Objectives and Methods

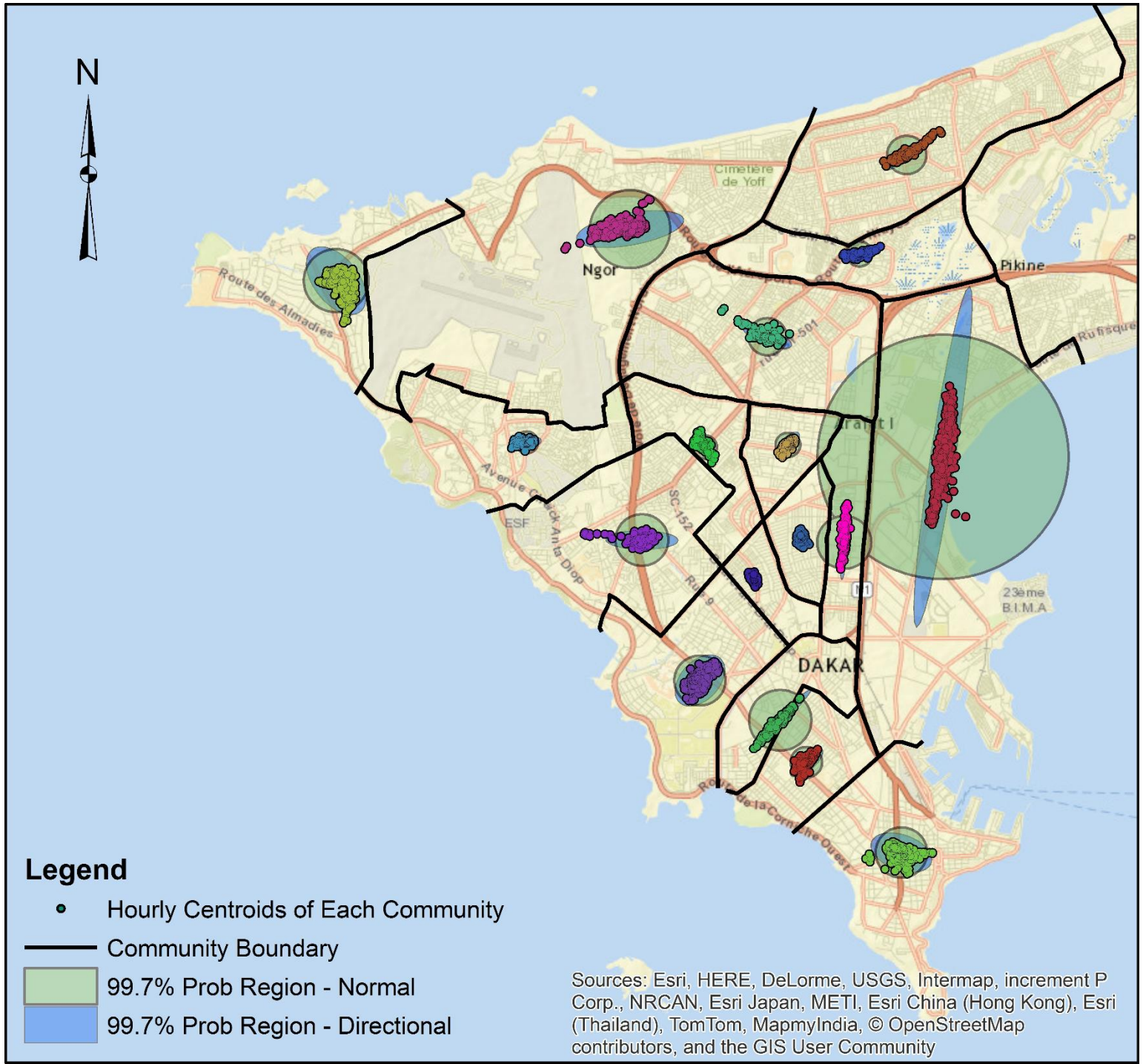


# Collective Mobility Patterns

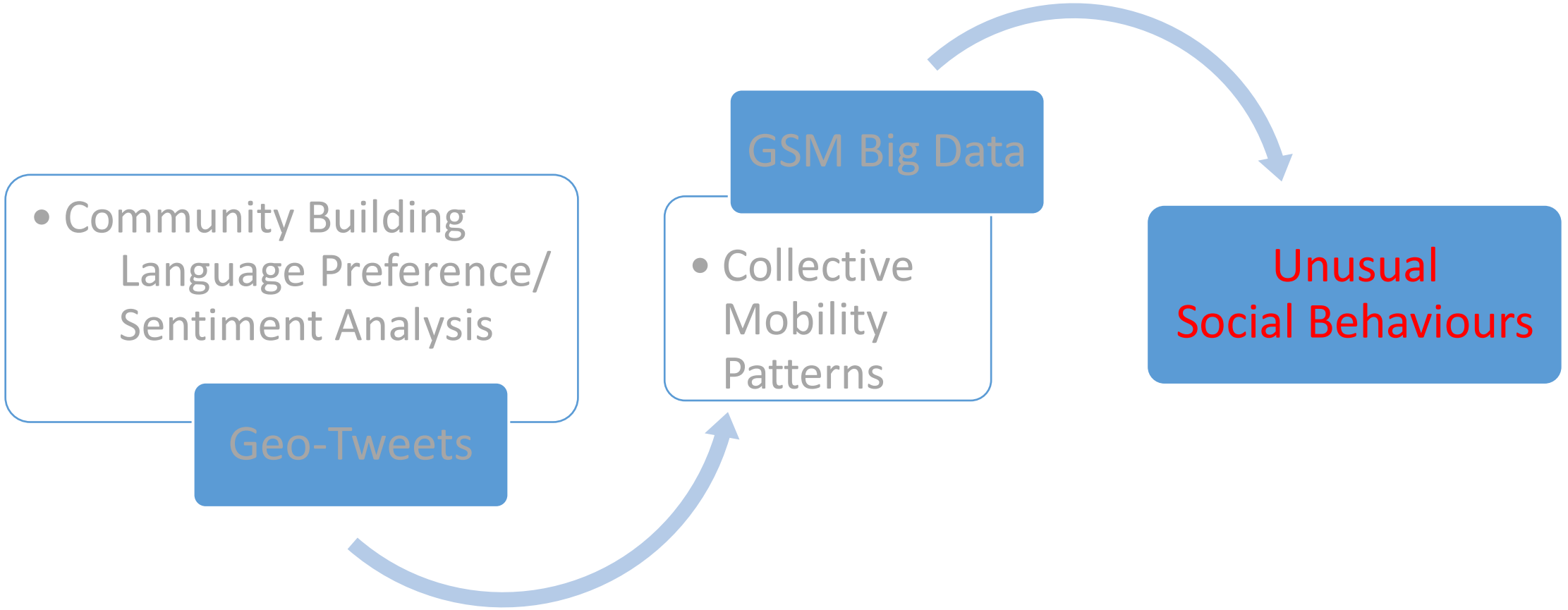
- Compute amount of hourly traffic for each antenna
- Calculate hourly weighted mean (centroid) of traffic amount of antennas which belong to a same community
- # of calls as weight
- In a month with 31 days
  31*24 = 744 centroids will be calculated for each community



# Collective Mobility Patterns



# Research Objectives and Methods



# Uncovering Unusual Social Behaviours

- An hourly centroid of a community falls outside from the community's 99.7% probability region:

  May be an unusual social behaviour

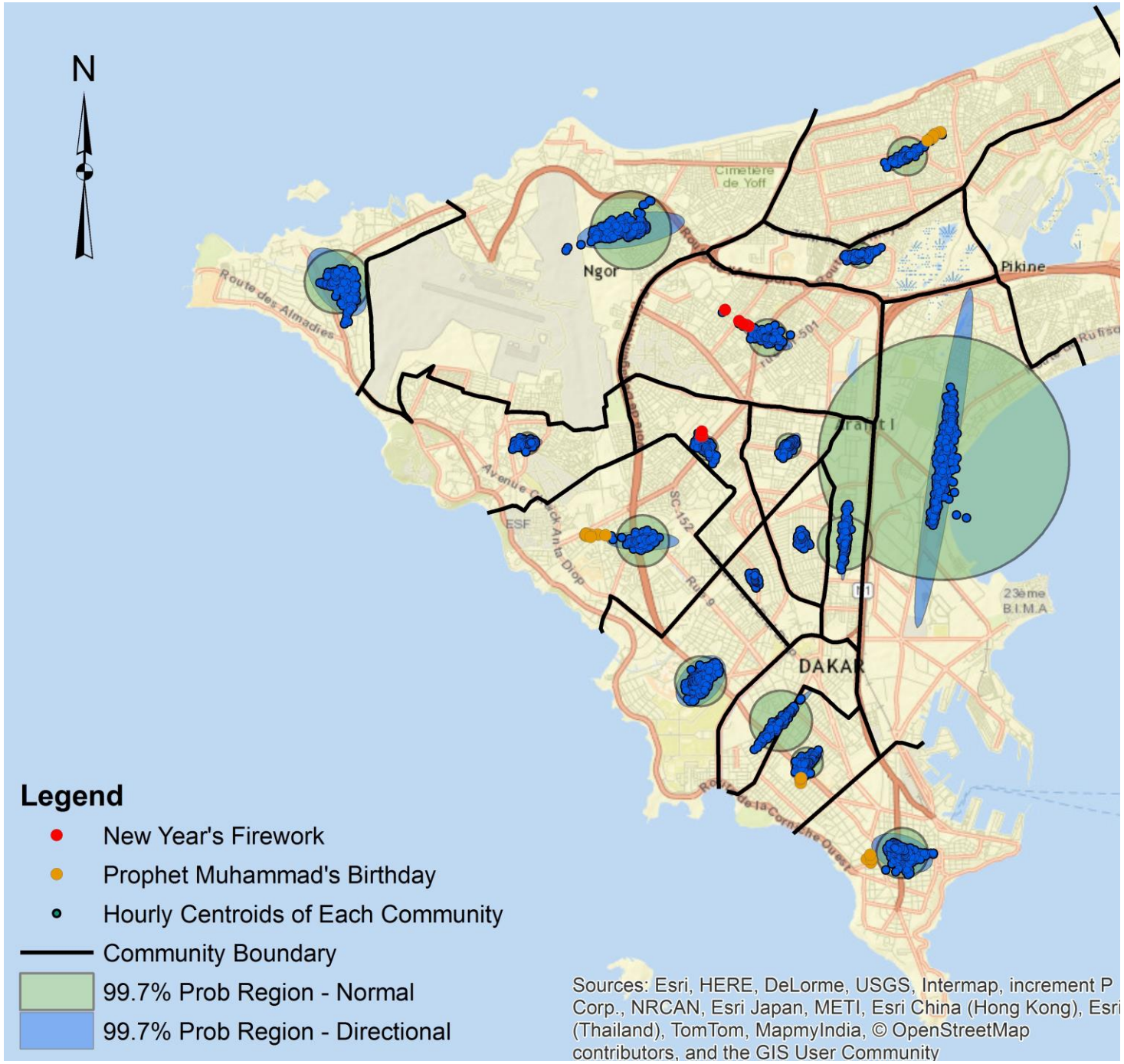
- If the following two hours' centroids continuously fall outside from the 99.7% probability region

  The unusual social behaviour is confirmed

# Result

Jan 1, 2013 - New Year Eve Firework

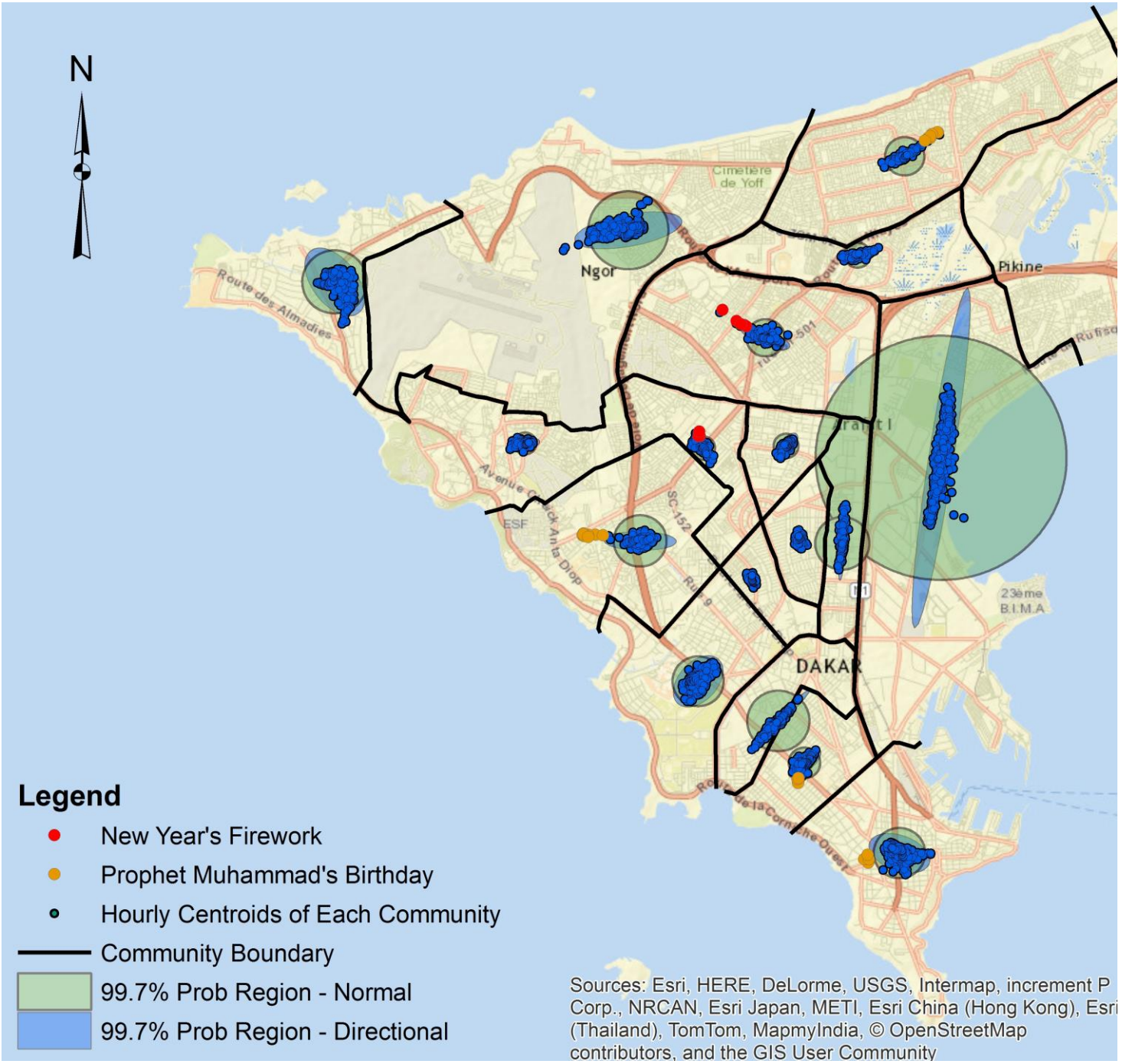
Jan 24, 2013 - Prophet Muhammad's Birthday



# Result

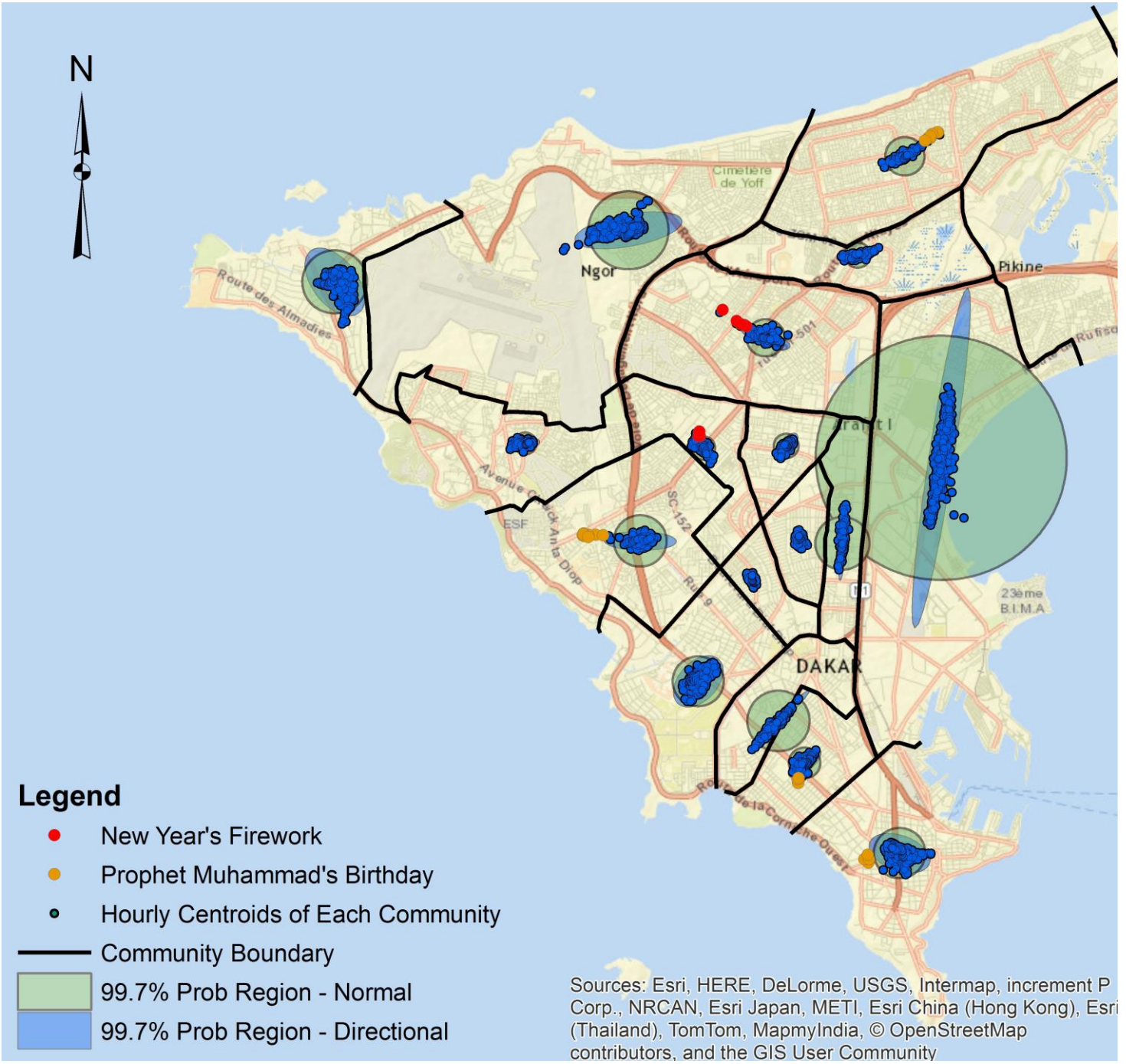
Jan 1, 2013 - New Year Eve Firework

| Centroid | XCoord | YCoord | Community | Hour | Day |
|---|---|---|---|---|---|
| 57 | -1943298.453 | 1648243.086 | 6 | 3 | 1 |
| 74 | -1943272.931 | 1648225.662 | 6 | 4 | 1 |
| 91 | -1943415.873 | 1648301.804 | 6 | 5 | 1 |
| 108 | -1943348.489 | 1648253.616 | 6 | 6 | 1 |
| 125 | -1943618.741 | 1648468.230 | 6 | 7 | 1 |
| 76 | -1943942.523 | 1646592.451 | 8 | 4 | 1 |
| 93 | -1943963.149 | 1646592.233 | 8 | 5 | 1 |
| 110 | -1943953.637 | 1646659.011 | 8 | 6 | 1 |




## Jan 24, 2013 - Prophet Muhammad's Birthday

| Centroid | XCoord | YCoord | Community | Hour | Day |
|---|---|---|---|---|---|
| 9388 | -1940613.685 | 1651019.540 | 4 | 0 | 24 |
| 9405 | -1940502.412 | 1651102.546 | 4 | 1 | 24 |
| 9422 | -1940508.567 | 1651107.019 | 4 | 2 | 24 |
| 9439 | -1940592.240 | 1651077.932 | 4 | 3 | 24 |
| 9456 | -1940650.085 | 1651015.970 | 4 | 4 | 24 |
| 9473 | -1940617.702 | 1651045.504 | 4 | 5 | 24 |
| 9490 | -1940694.016 | 1650981.382 | 4 | 6 | 24 |
| 9391 | -1945453.157 | 1645126.536 | 7 | 0 | 24 |
| 9408 | -1945562.481 | 1645129.617 | 7 | 1 | 24 |
| 9425 | -1945626.517 | 1645146.684 | 7 | 2 | 24 |
| 9442 | -1945637.427 | 1645118.731 | 7 | 3 | 24 |
| 9459 | -1945617.967 | 1645106.984 | 7 | 4 | 24 |
| 9476 | -1945624.254 | 1645132.200 | 7 | 5 | 24 |
| 9493 | -1945559.513 | 1645091.225 | 7 | 6 | 24 |
| 9510 | -1945340.277 | 1645119.125 | 7 | 7 | 24 |
| 9417 | -1942515.231 | 1641481.918 | 16 | 1 | 24 |
| 9434 | -1942521.903 | 1641440.494 | 16 | 2 | 24 |
| 9451 | -1942532.122 | 1641497.490 | 16 | 3 | 24 |
| 9468 | -1942518.896 | 1641514.979 | 16 | 4 | 24 |
| 9452 | -1941511.973 | 1640256.701 | 17 | 3 | 24 |
| 9469 | -1941569.272 | 1640306.845 | 17 | 4 | 24 |
| 9486 | -1941503.265 | 1640310.359 | 17 | 5 | 24 |
| 9503 | -1941519.272 | 1640375.256 | 17 | 6 | 24 |



# Conclusions

- Unusual social behaviour can be discovered by analyzing GSM and Geo-Tweet data
- Timely and efficiently
- Benefit many industries such as marketing, public health and urban planning

- Python with ArcPy toolbox can be used to process and analyze big data only by using computing power of average personal computers

# Acknowlegements


- My supervisor: Dr. Monica Wachowicz
- Dr. Emmanuel Stefanakis
- Lola Arteaga
- David Fraser
- Orange France