Identifying Meaningful Locations of Social Media Users

Qingqing Chen, Research Associate
Ate Poorthuis, Assistant Professor
Singapore University of Technology and Design

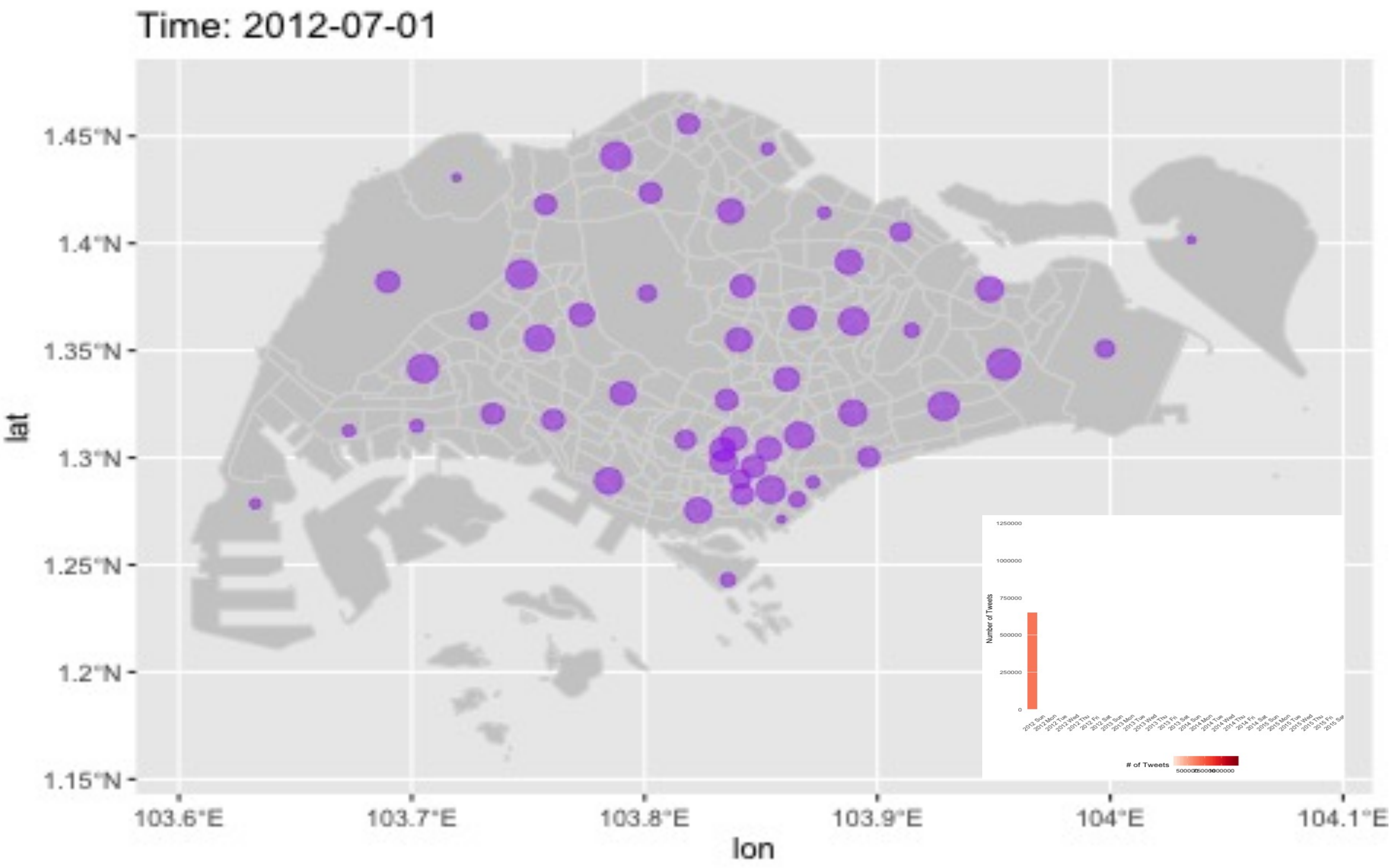
qingqing_chen@sutd.edu.sg
ate_poorthuis@sutd.edu.sg

Leaflet | Imagery provided by services from the Global Imagery Browse Services (GIBS), operated by the NASA/GSFC/Earth Science Data and Information System (ESDIS) with funding provided by NASA/

OUTLINES

- Meaningful Locations
- Challenges and Objectives
- Bottom-up Approach
- HOMELOCATOR Package
- Results and Conclusion

MEANINGFUL LOCATIONS



Time: 2012-07-01
1.45°N
1.4°N
1.35°N
1.3°N
1.25°N
1.2°N
1.15°N
lat
103.6°E
103.7°E
103.8°E
103.9°E
104°E
104.1°E
lon
Number of Tweets
# of Tweets

Users’ locations are extremely valuable for urban system study!

However

Locations are currently everywhere…

And

Users can go all over the city…

But

On the surface,

This tells us relatively little about the nature and meaningful of these locations…

So

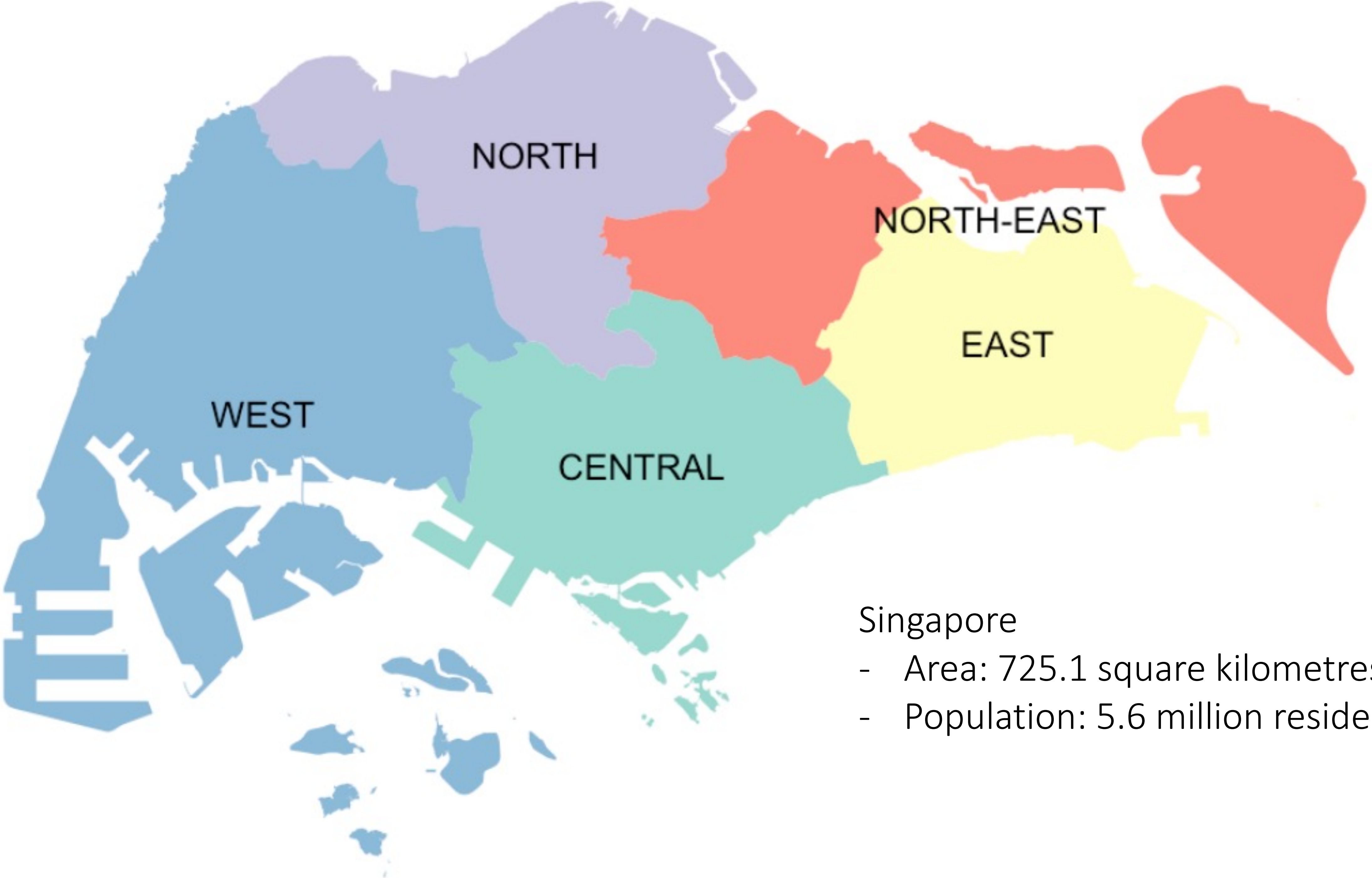
How can we extract those meaningful locations from the data?

First Objective

Bottom-up approach:

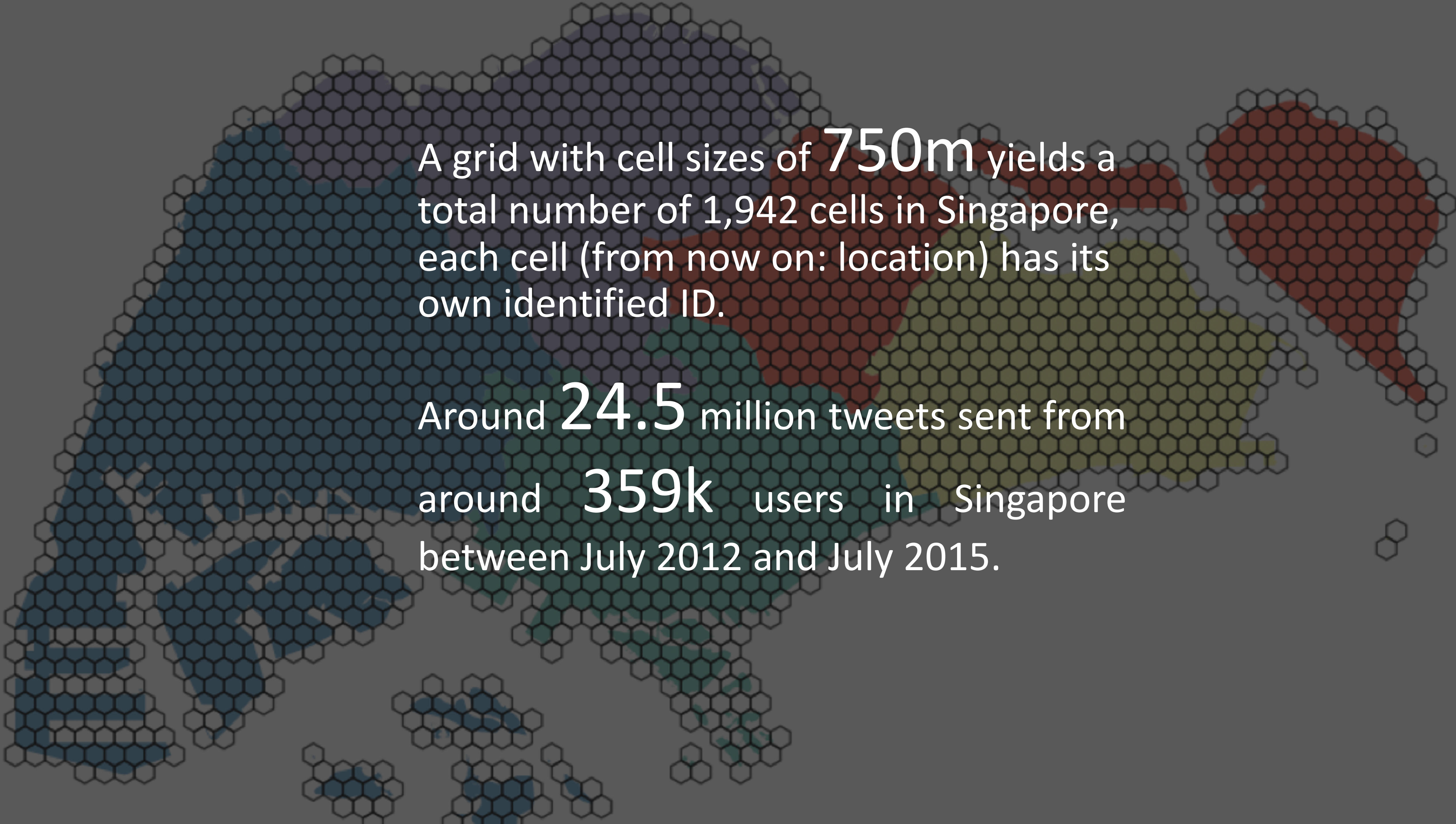
Identify meaningful locations from mobile technology, or more specifically speaking, from Social Media Data, based on its temporal and spatial features.

Twitter



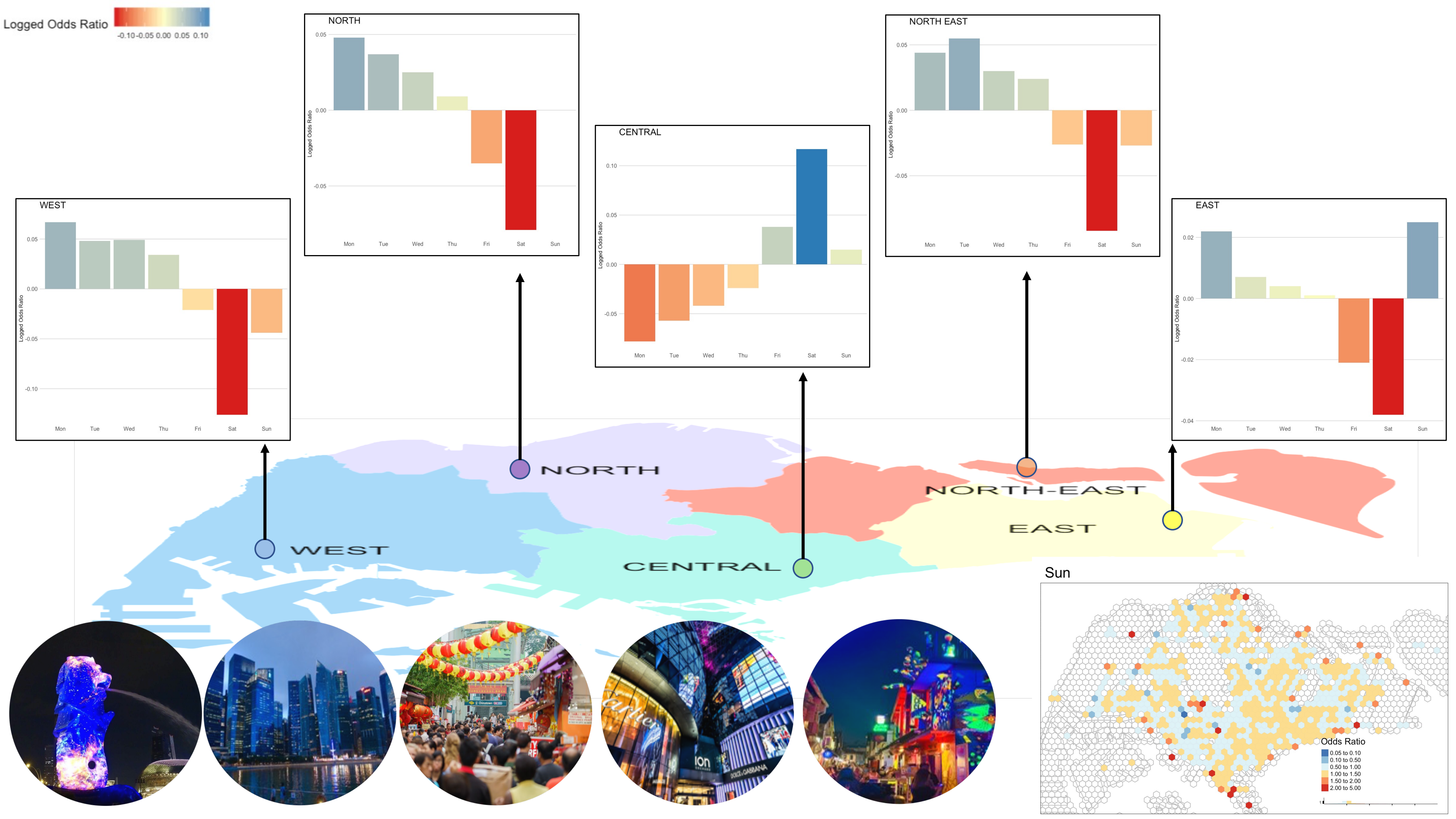
Singapore

- Area: 725.1 square kilometres
- Population: 5.6 million residents

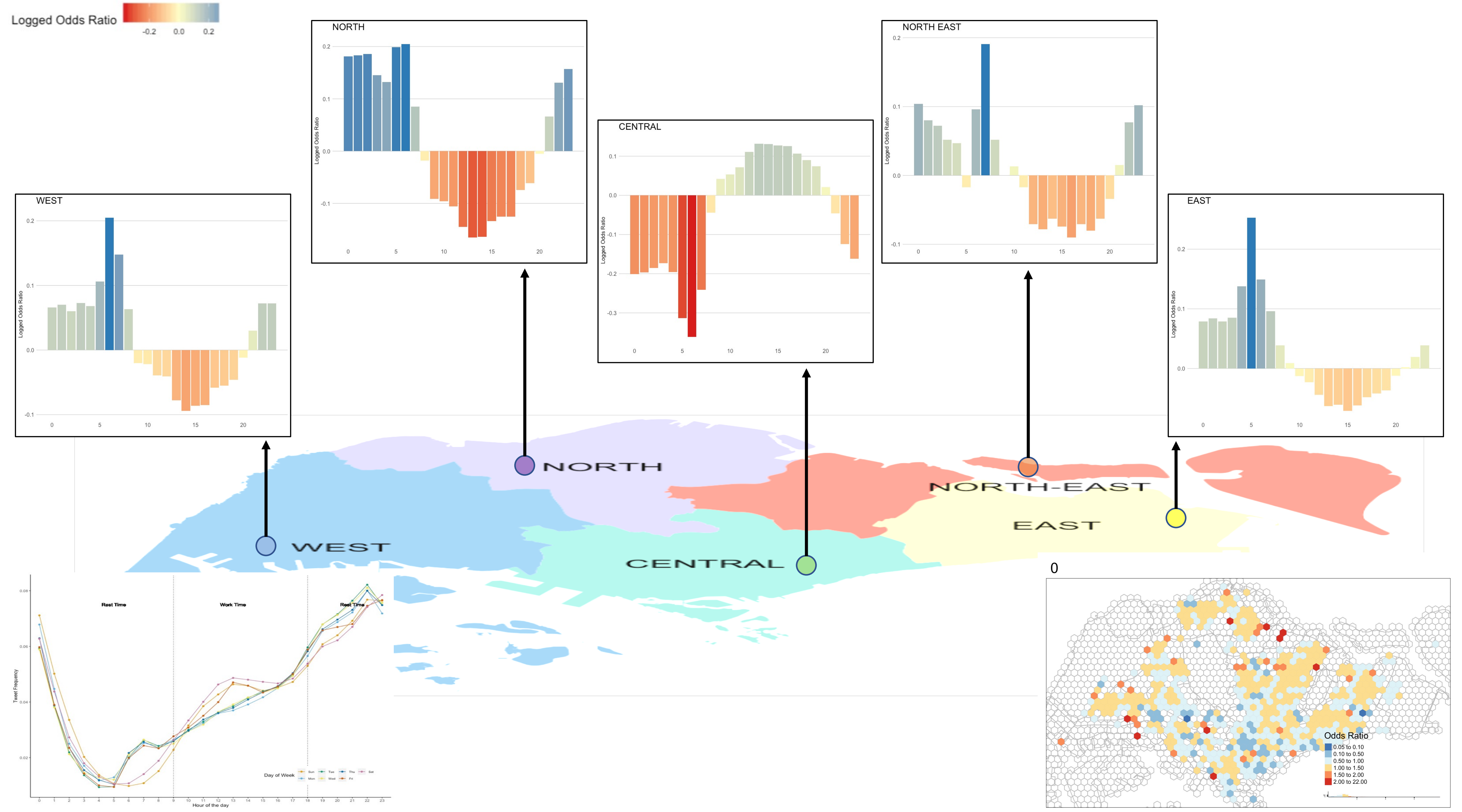


A grid with cell sizes of 750m yields a total number of 1,942 cells in Singapore, each cell (from now on: location) has its own identified ID.

Around 24.5 million tweets sent from around 359k users in Singapore between July 2012 and July 2015.



Logged Odds Ratio
-0.10 -0.05 0.00 0.05 0.10
NORTH
CENTRAL
NORTH EAST
WEST
EAST
Logged Odds Ratio
Mon
Tue
Wed
Thu
Fri
Sat
Sun
NORTH
NORTH-EAST
EAST
WEST
CENTRAL
Sun
Odds Ratio
0.05 to 0.10
0.10 to 0.50
0.50 to 1.00
1.00 to 1.50
1.50 to 2.00
2.00 to 5.00



Logged Odds Ratio
-0.2
0.0
0.2
NORTH
CENTRAL
NORTH EAST
WEST
EAST
Logged Odds Ratio
NORTH
NORTH-EAST
EAST
WEST
CENTRAL
Rest Time
Work Time
Rest Time
Tweet Frequency
Hour of the day
Day of Week
Sun
Mon
Tue
Wed
Thu
Fri
Sat
0
Odds Ratio
0.05 to 0.10
0.10 to 0.50
0.50 to 1.00
1.00 to 1.50
1.50 to 2.00
2.00 to 22.00

Custom Model

Deriving time related variables from Timestamp

Such as year, month, day, day of the week, hour of the day, etc.

Set pre-conditions to keep meaningful users

Condition	Min. Requirement
Data points sent by the user	> 10 tweets
Unique locations the user was active	> 10 locations
Data points sent at a location by the user	> 10 tweets
Unique hours the user was active at a location	> 10 hours
Unique days the user was active at a location	> 10 days
Time period the user was active at a	
Potential Twitter bots	

Split activity into different time frames

Time frame	Time period
Weekend	Saturday & Sunday
Weekday	Monday to Friday
Rest Time	1:00 - 8:00 & 19:00 - 24:00
Work Time	9:00 - 18:00
Early Morning	6:00 - 12:00
Non-early Morning	1:00 - 5:00 & 13:00 - 24:00

Weight and score the variables

Activity	Expression	Weight	Score
percentage of tweets on weekends	$\frac{T_{wk}}{T_{wk}+T_{wd}}$	0.2	$0.2 \times \frac{T_{wk}}{T_{wk}+T_{wd}}$
percentage of tweets on rest times	$\frac{T_{rt}}{T_{rt}+T_{wt}}$	0.2	$0.2 \times \frac{T_{rt}}{T_{rt}+T_{wt}}$
Normalized Tweets at a location	$\frac{T_p}{max(T_p), p \in P_u}$	0.1	$0.1 \times \frac{T_p}{max(T_p), p \in P_u}$
Normalized unique days at a location	$\frac{N_{d,p}}{max(N_{d,p})}, \forall p \in P_u$	0.1	$0.1 \times \frac{N_{d,p}}{max(N_{d,p})}$
Normalized time period at a location	$\frac{TP_p}{max(TP_p)}, \forall p \in P_u$	0.1	$0.1 \times \frac{TP_p}{max(TP_p)}$
Normalized unique months	$\frac{N_{m,u}}{12}$	0.1	$0.1 \times \frac{N_{m,u}}{12}$
Normalized unique day of week	$\frac{N_{dow,u}}{7}$	0.1	$0.1 \times \frac{N_{dow,u}}{7}$
Normalized unique hours at a location	$\frac{N_{h,p}, \forall p \in P_u}{24}$	0.1	$0.1 \times \frac{N_{h,p}, \forall p \in P_u}{24}$

Extract the ‘home’ location based on the score

Current State-of-the-Art

- Based on spatial-temporal features
  - Calculate the weight of posted tweets or tract activity across different time frames (e.g., Ahas et al. 2010, Efstathiades et al. 2015; Lin et al. 2018, etc.)

- Based on often-available social network
  - Use the user's friend network and tie strength (e.g., Jurgens 2013; McGee et al. 2013; Hironaka et al. 2016; Chen et al. 2016, etc.)

- Based on actual-content estimation
  - Detect the actual content of social media messages or user profiles (Hecht et al. 2011; Chandar et al. 2011; Chang et al. 2012, etc.)

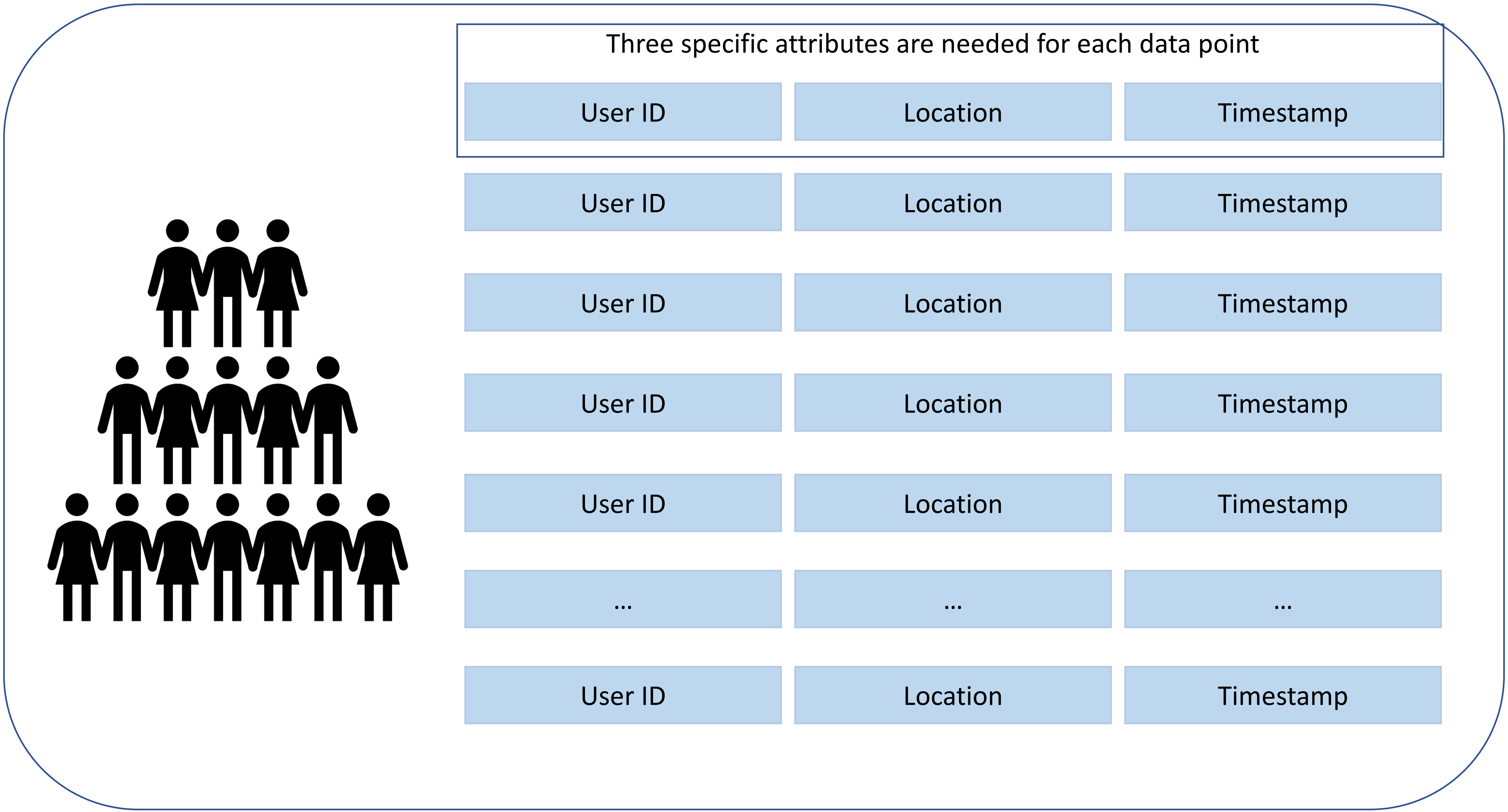
Problems Faced

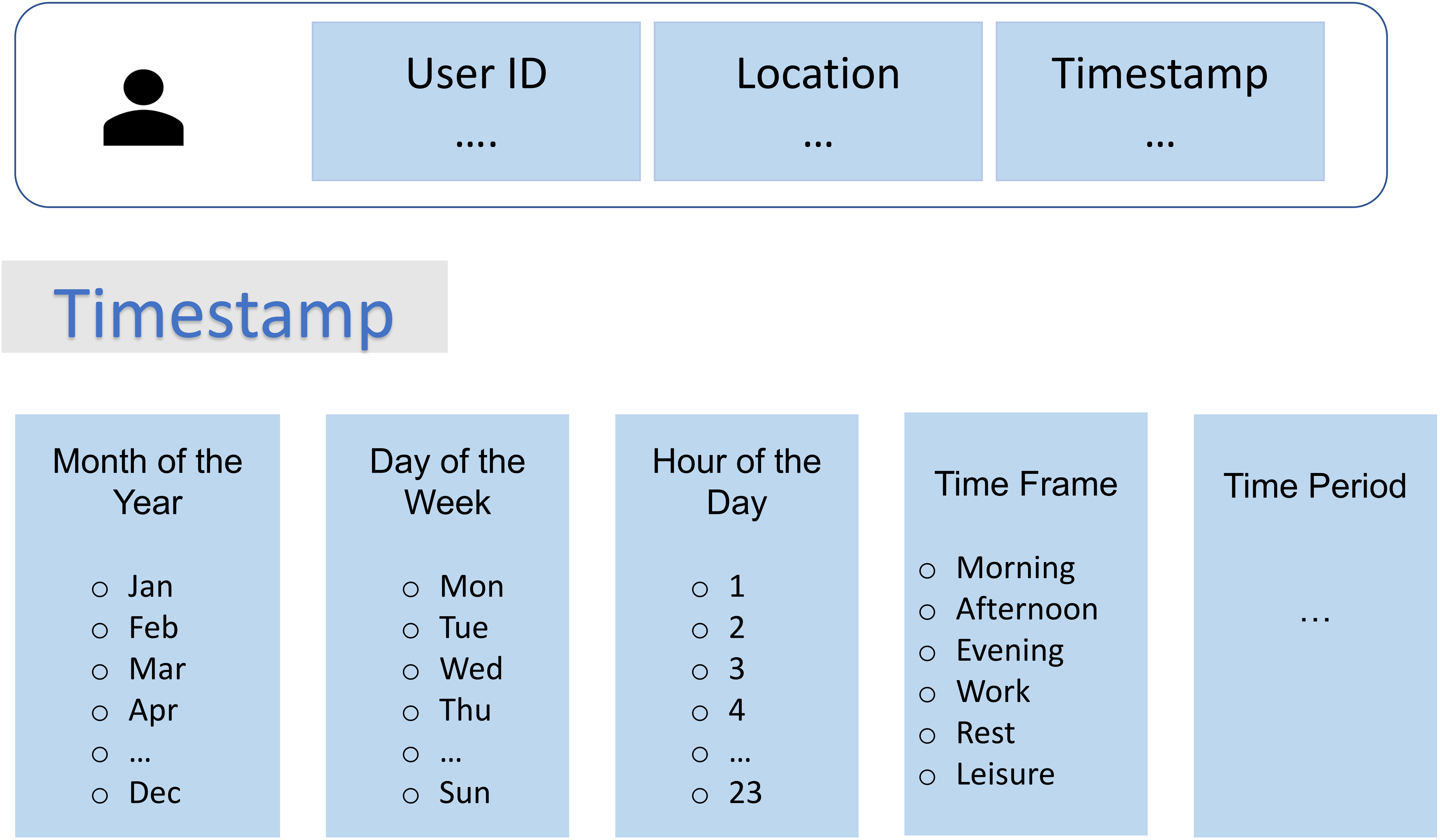
- Wide variety of different approaches without a clear ‘best-practice’.
- Difficult to evaluate the effectiveness of each approach as a ‘ground-truth’ most often doesn’t exist.
- Algorithms are not always discussed in detail in publications, which makes comparing algorithms as well as reproducing work difficult.
- Difficult to evaluate the robustness of subsequent findings due to the terms and conditions that often prevent the sharing of social media data.

Second Objective

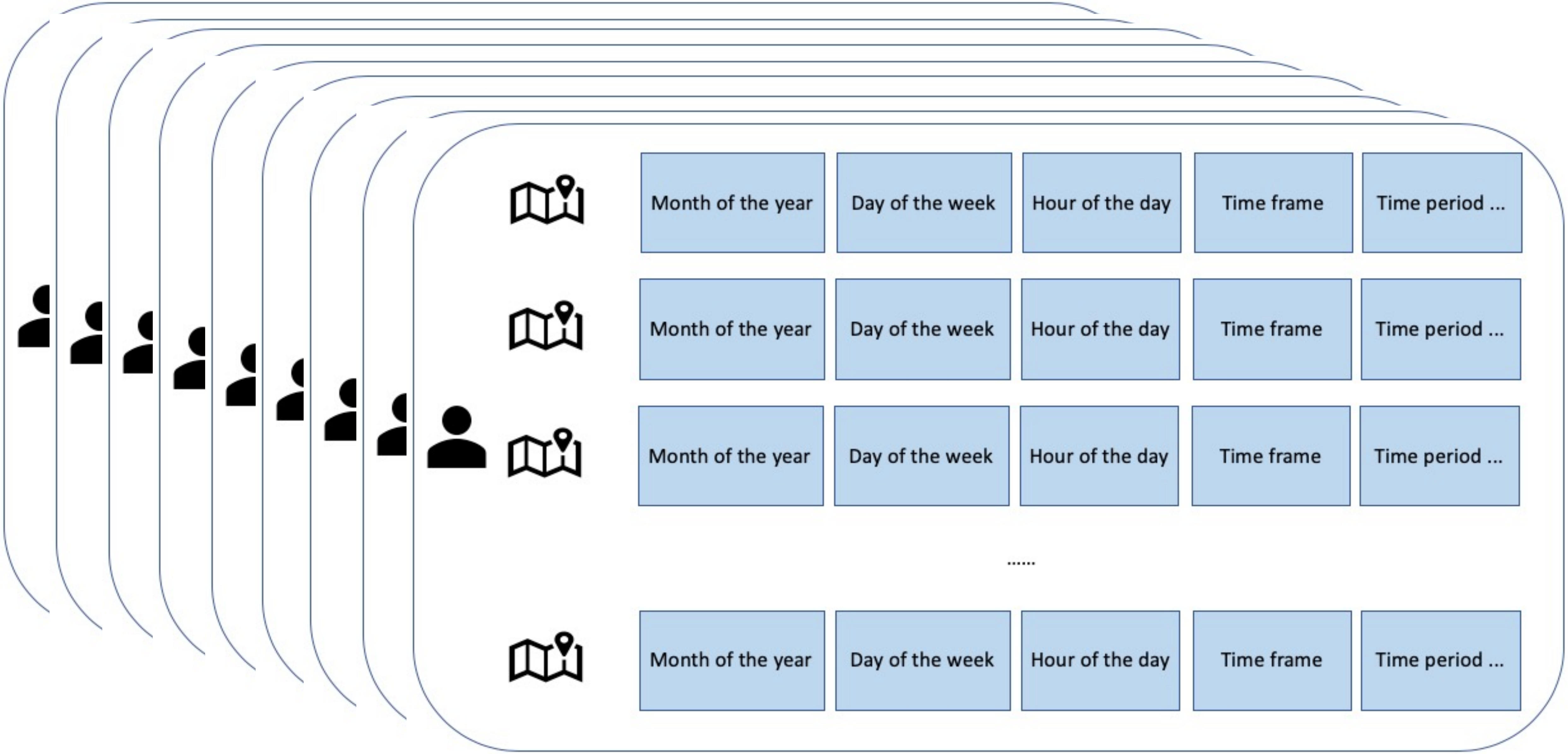
R Package: HOMELOCATOR

- Provide a consistent framework and interface for the adoption of different approaches to meaningful location identification
  - Approach can be written as a ‘recipe’, which make it easy to be used
  - Make comparison across different algorithms become possible
  - Functions of the package are flexible enough for people to create new variables or tune the existing variables’ thresholds of the recipes


Three specific attributes are needed for each data point
User ID
Location
Timestamp
User ID
Location
Timestamp
User ID
Location
Timestamp
User ID
Location
Timestamp
User ID
Location
Timestamp
…
…
…
User ID
Location
Timestamp

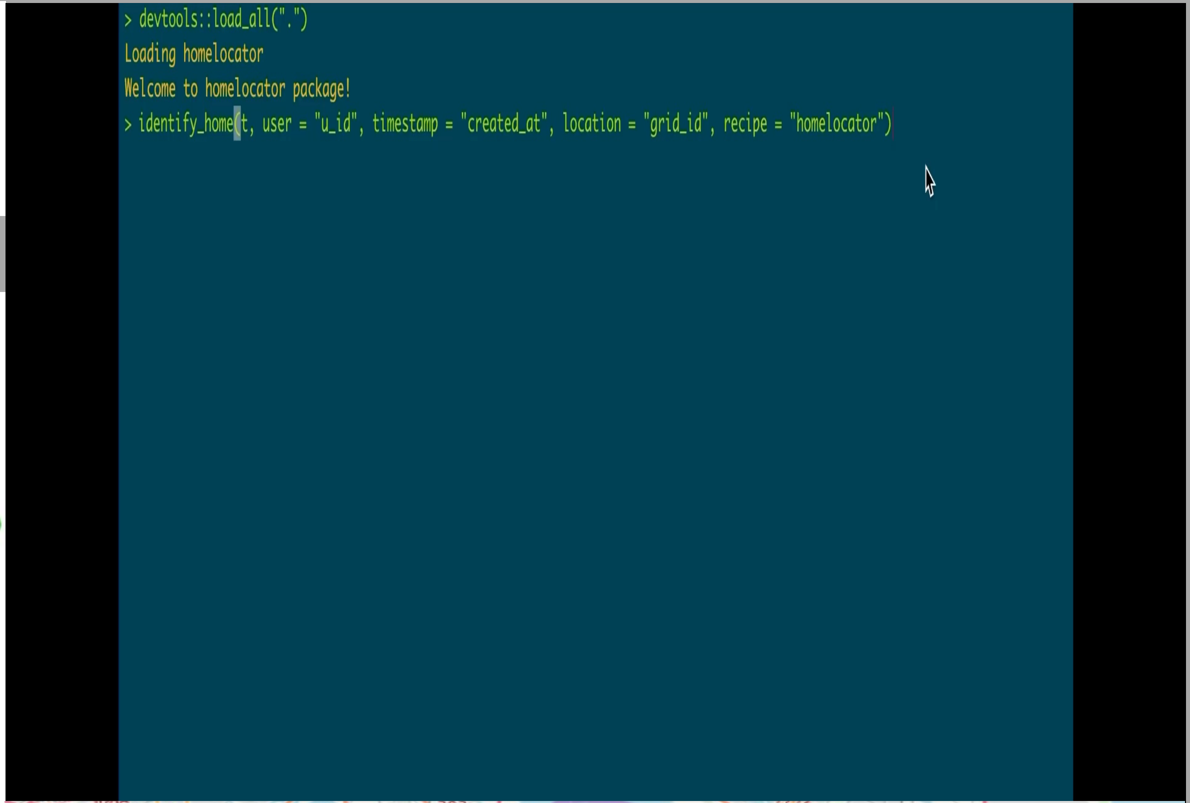

User ID
....
Location
...
Timestamp
...
Timestamp
Month of the Year
Jan
Feb
Mar
Apr
...
Dec
Day of the Week
Mon
Tue
Wed
Thu
...
Sun
Hour of the Day
1
2
3
4
...
23
Time Frame
Morning
Afternoon
Evening
Work
Rest
Leisure
Time Period
...

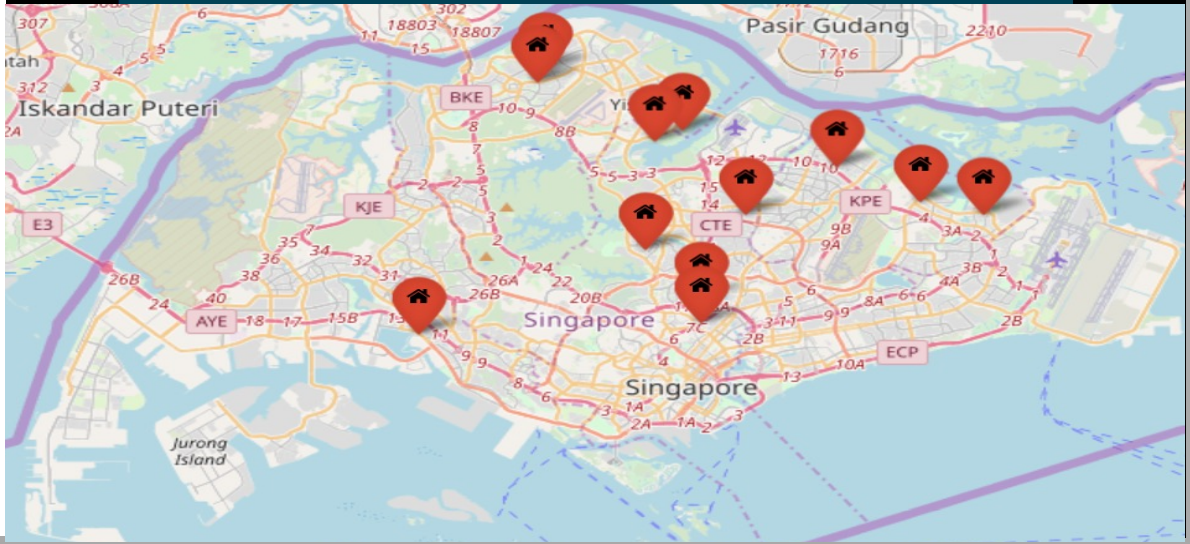
Parallel Computing



```r
# Use of algorithm with available recipes (OSN, MPD, HLC, FREQ)
homes <- identify_home(df, ..., recipe = "OSN")
# df is an input table with one row per data point and variables for
    'user', 'location' and 'timestamp'
```

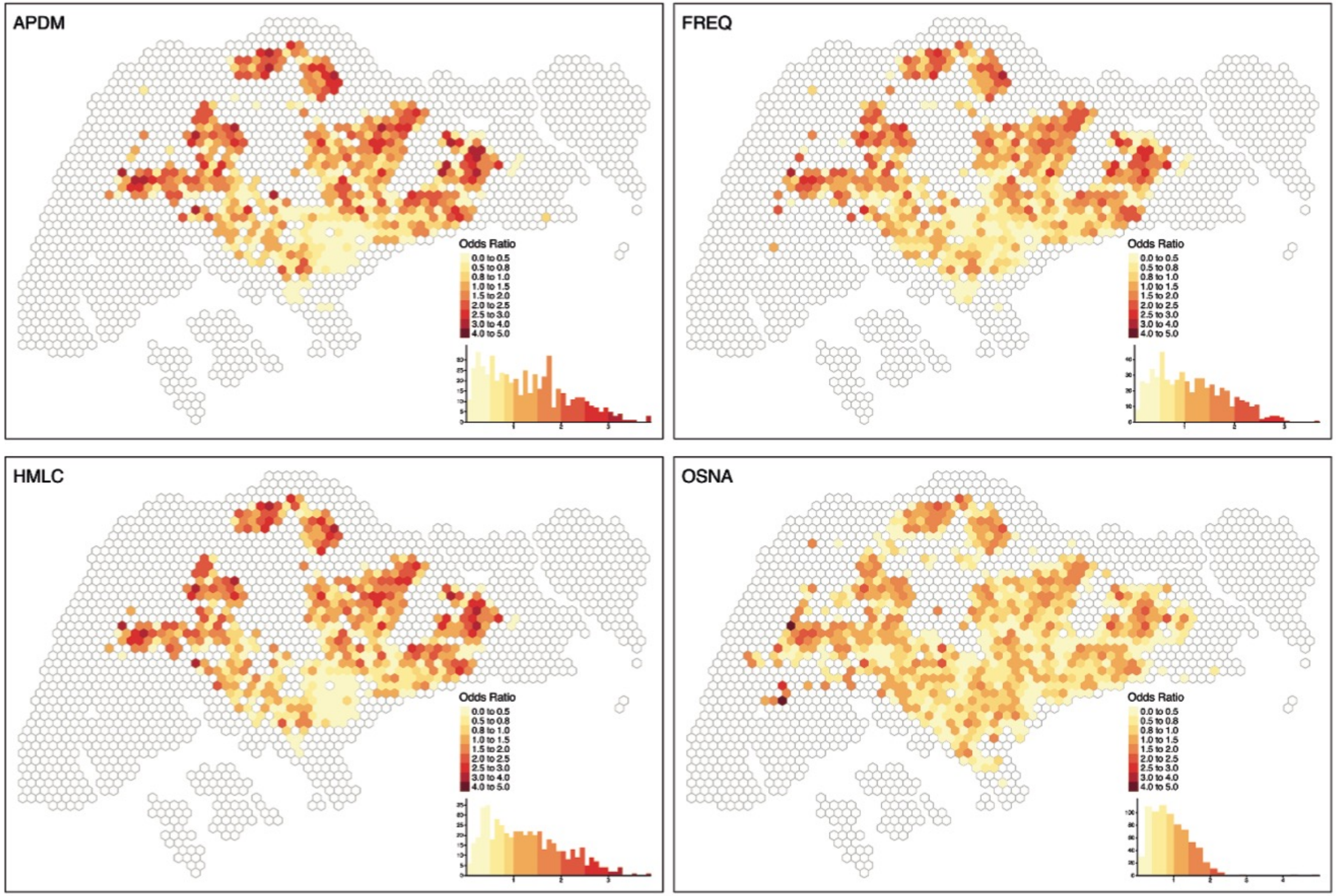
```r
# Validate input dataset, ensuring necessary columns are present
validate_dataset(df, user, timestamp, location)

# derive additional temporal variables (e.g. day of week)
enrich_timestamp(df, timestamp) # Derive temporal features from timestamp

# nest by user (creates a nested tibble for each user, allowing
    subsequent parallel processing per user)
nest_by_user(df_valid, group_var = user)

# add additional column for each data point within nested tibble
add_col_in_nest(wd_or_wk = if_else(wday %in% c(1,7), "weekend",
    "weekday")) # create weekday/weekend column

# filter based on column within nest
filter_in_nest(n_points_per_location > 10) # only keep locations with
    more than 10 data points

# extract most likely residential location after previous filter and
    weighting steps
extract_home(...)
```





Results

- Four approaches focused on residential (home) locations
- 22174 (6.18%) users have been assigned a location by all four algorithms (common users)
- 13233 (59.68%) users get assigned the same home location even though the approaches are quite distinct
- The package opens the door for comparing across different algorithms .

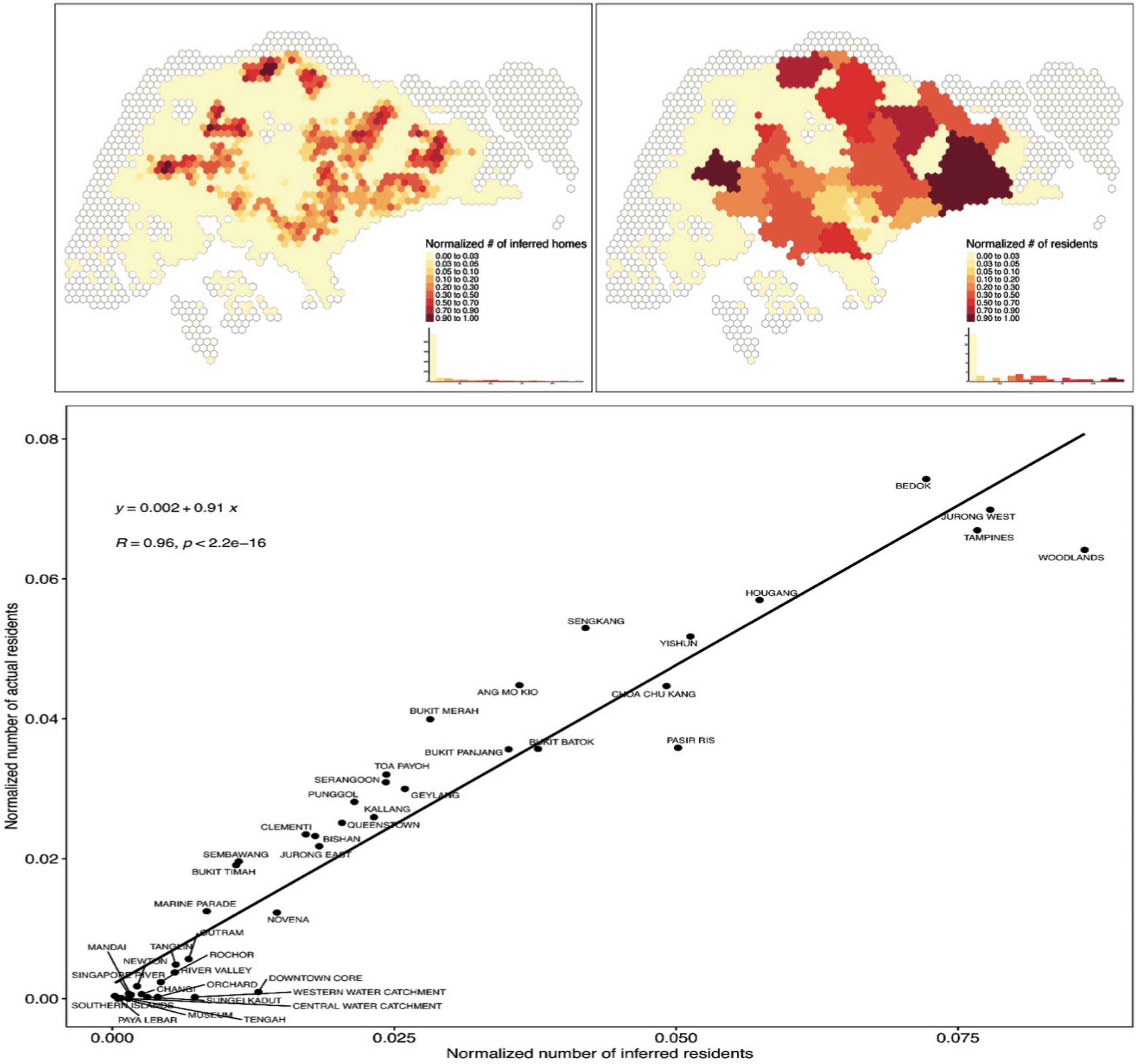


Approach	Results
APDM (Ahas *et al.* 2010)	Identified 40,374 (31.0%) users' homes of 130,311 users
FREQ	Identified 47,263 (36.3%) users' homes of 130,311 users
HMLC	Identified 33,488 (25.7%) users' homes of 130,311 users
OSNA (Efstathiades *et al.* 2015)	Identified 116,104 (89.1%) users' homes of 130,311 users
Ensemble	Identified 21,863 (78.1%) users' homes of 28,007 shared users

Results

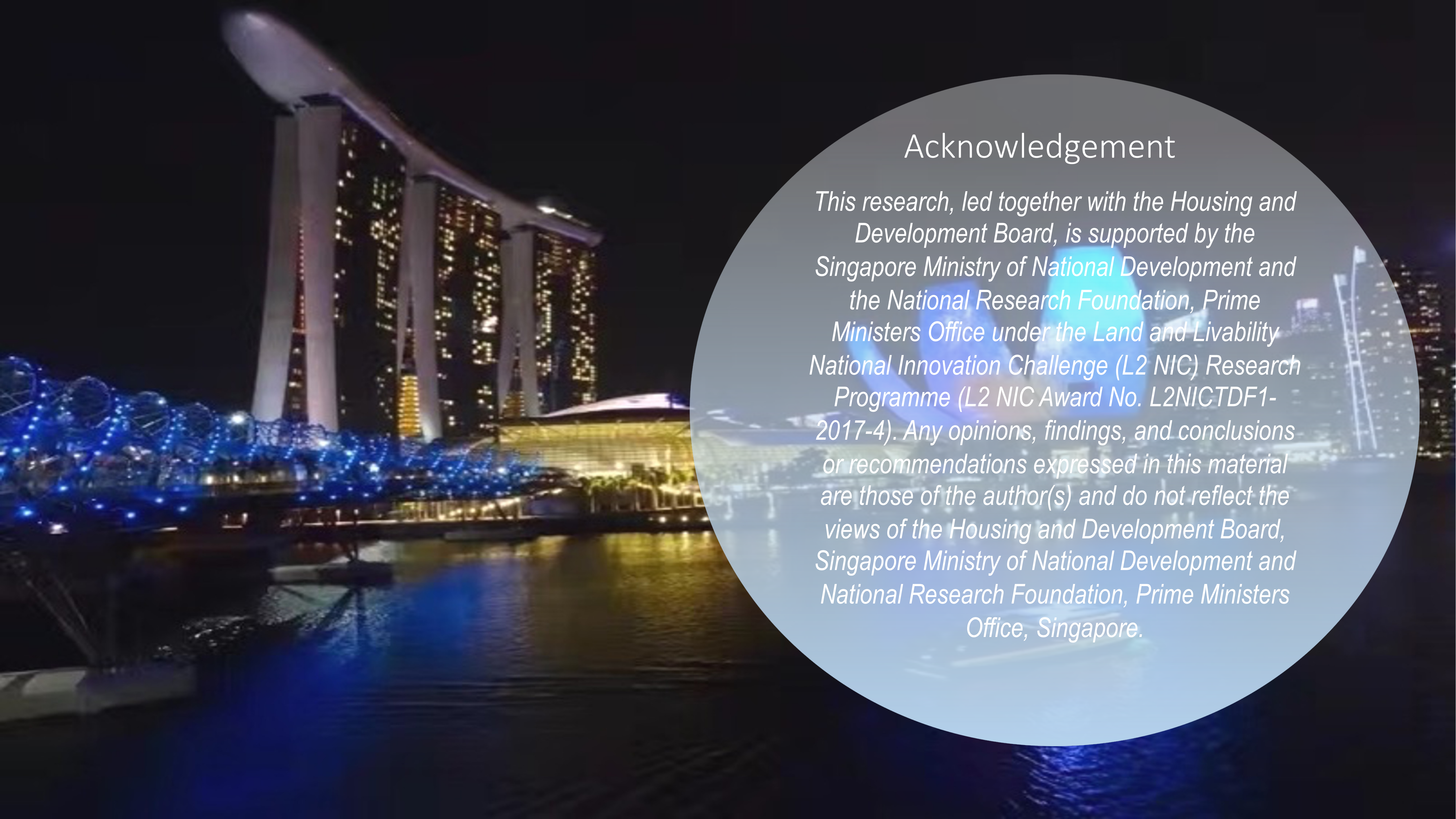
- Strong linear correlation (R = 0.96) between the normalized number of inferred residents and the normalized number of actual residents in Singapore

- Combination of algorithm is a fruitful way to infer the underlying the geography of a population and thus further be used in analysis of urban processes



Conclusion

Outline	Outline an ensemble approach to inferring meaningful locations from geotagged social media content through a 'homelocator' R package
Evaluate	The resulting spatial patterns from the ensemble approach closely correlated to the actual population distribution
Make	Make comparison across different approaches and algorithms • Easier • More accessible • More customizable
Increase	Increase transparency and reproducibility of work that relies on the inference of meaningful locations
Future work	- Apply the methods to the current social media climate like cell phone mobility data. - Extracted meaningful locations will be used to analyse and predict urban mobility patterns, for different groups of people and for different neighbourhoods.



Acknowledgement

This research, led together with the Housing and Development Board, is supported by the Singapore Ministry of National Development and the National Research Foundation, Prime Ministers Office under the Land and Livability National Innovation Challenge (L2 NIC) Research Programme (L2 NIC Award No. L2NICTDF1-2017-4). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the Housing and Development Board, Singapore Ministry of National Development and National Research Foundation, Prime Ministers Office, Singapore.