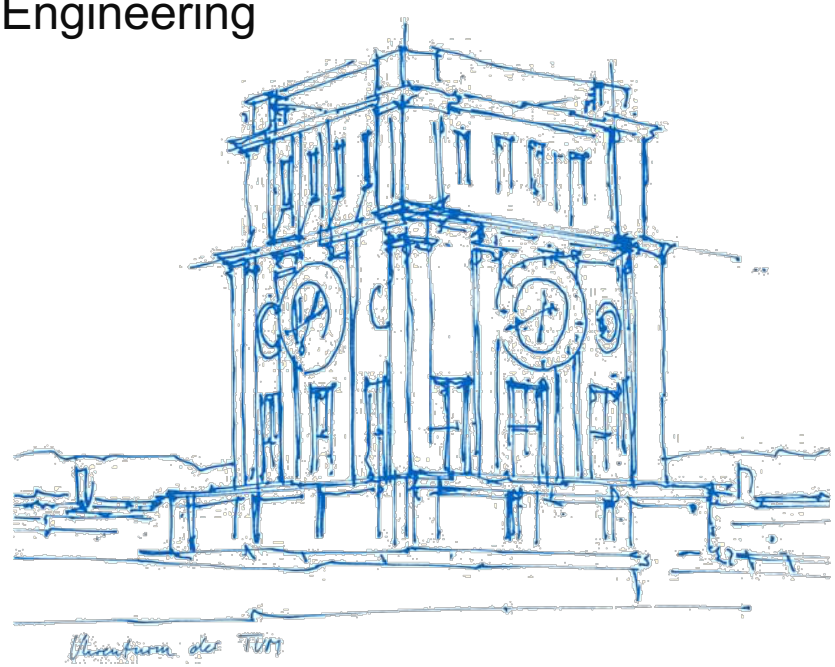
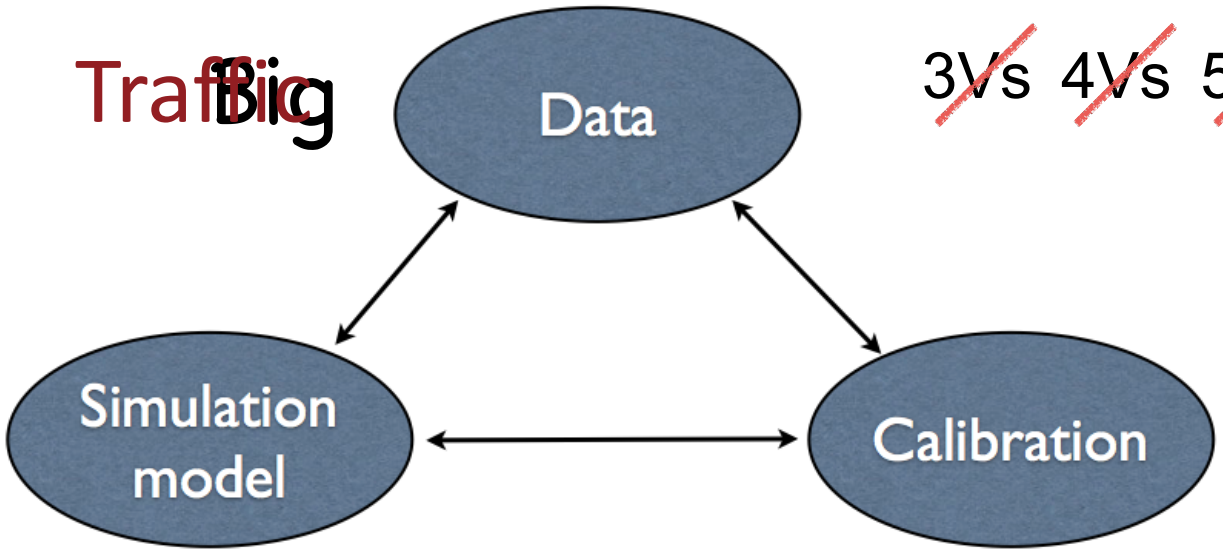
Mobility-related feature extraction from emerging data

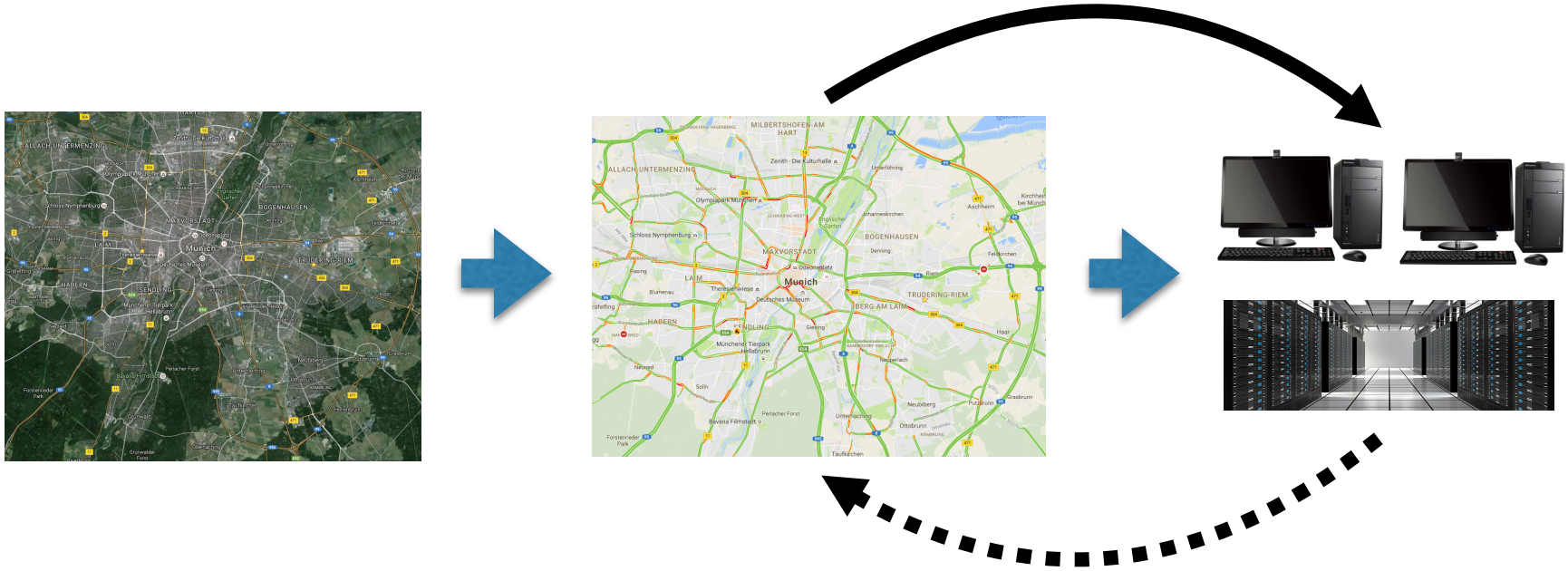
Emmanouil Chaniotakis and Constantinos Antoniou

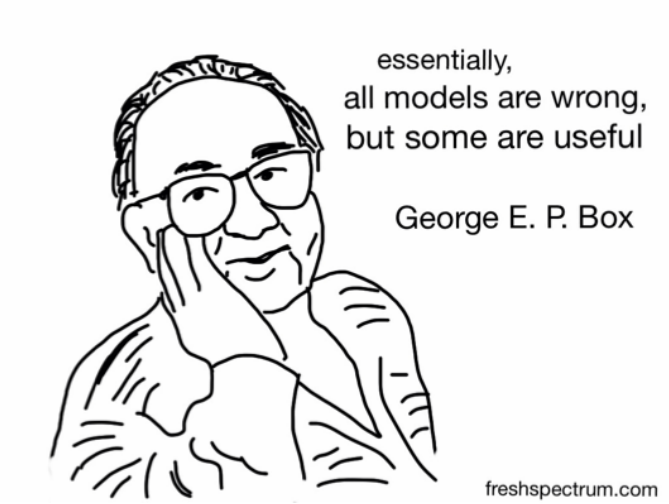
Chair of Transportation Systems Engineering
Department of Civil, Geo and Engironmental Engineering
Technical University of Munich (TUM)

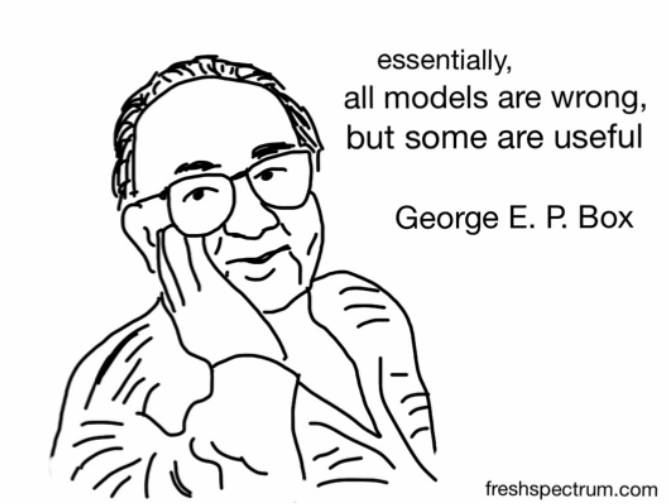



Traffic
Big
Data
3Vs 4Vs 5Vs 6Vs 7V
Simulation model
Calibration

Historical modeling approach

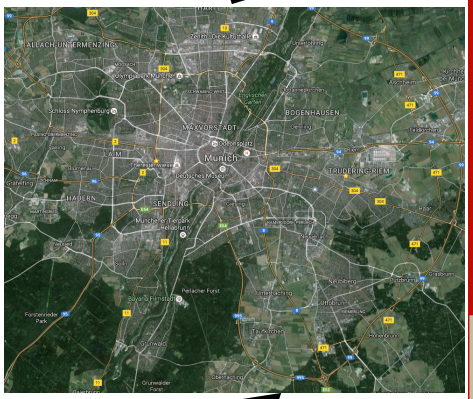



essentially,
all models are wrong,
but some are useful
George E. P. Box
freshspectrum.com

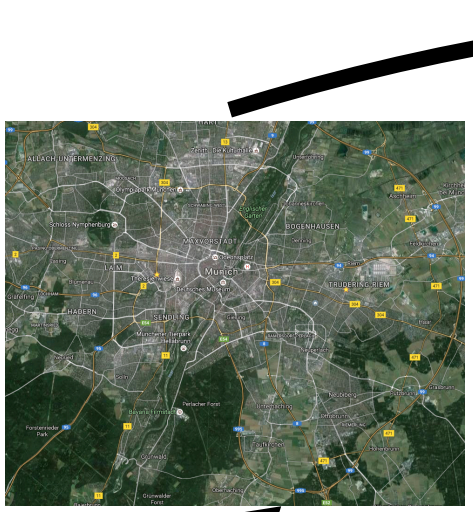


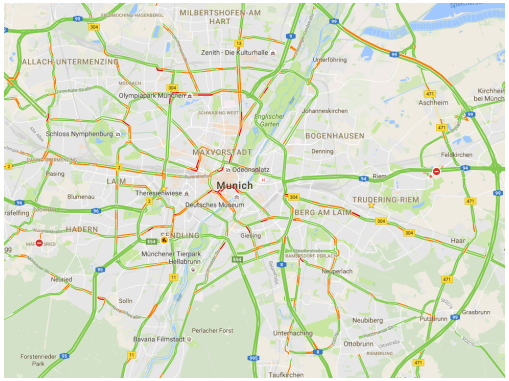
"All models are wrong,
and we can increasingly succeed without them"
C. Anderson (misquoting P. Norvig)
http://norvig.com/fact-check.html

First-genera[keep calm and pick the low hanging fruit]pach



Second-generation big-data approach





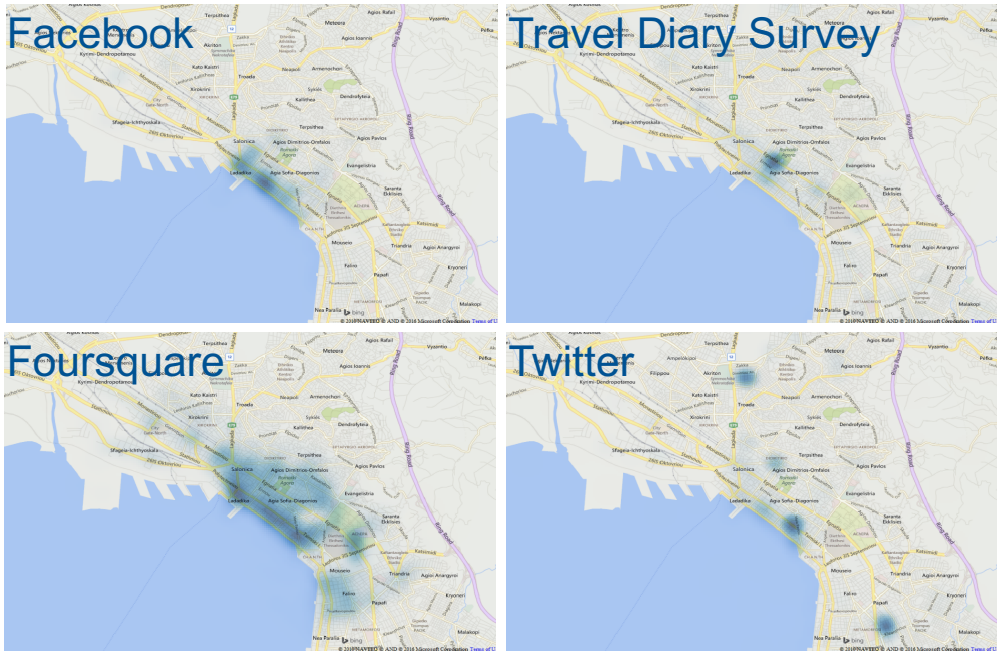
Motivation

Facebook the 3rd most visited website

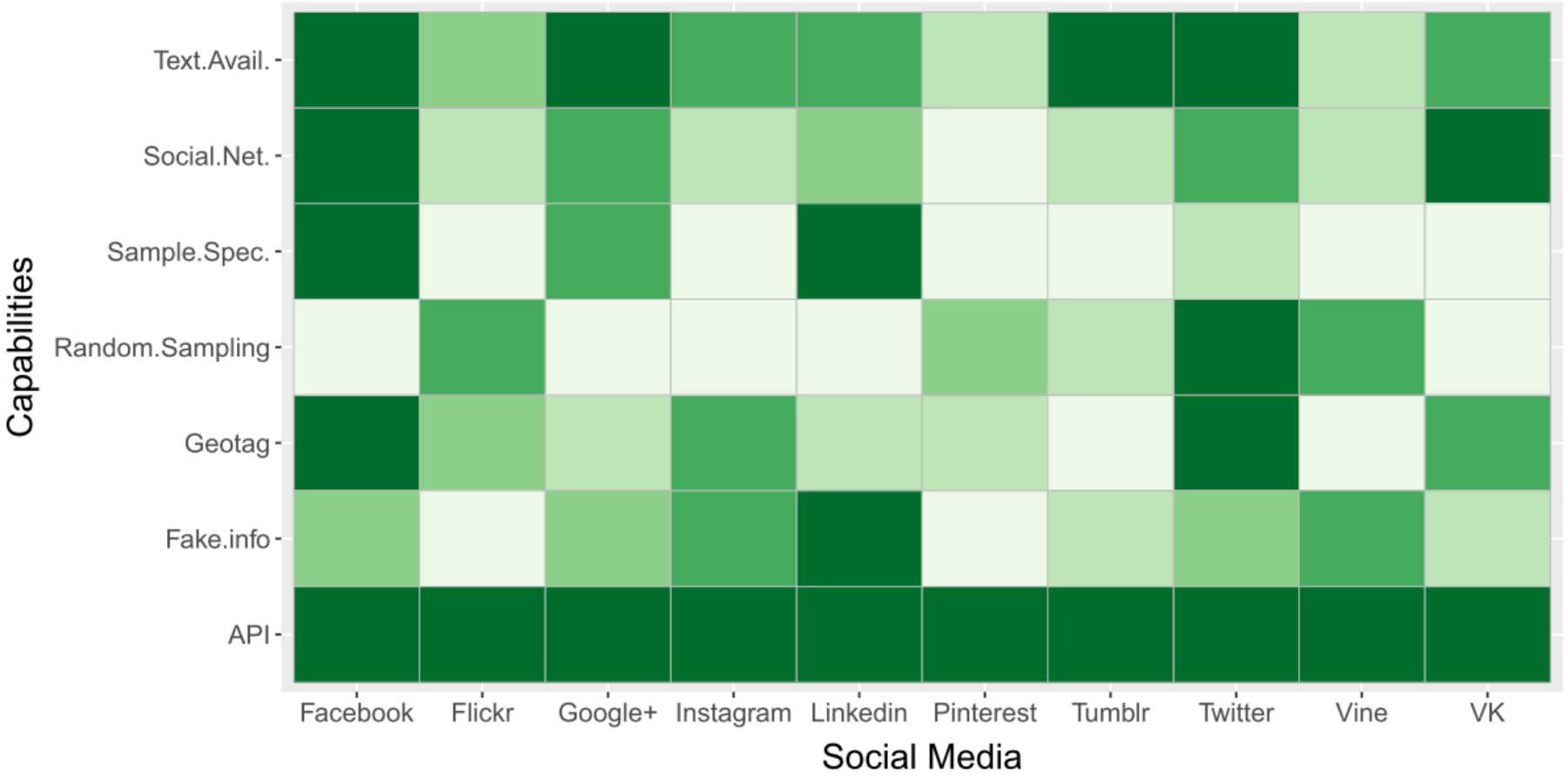
Twitter the 13th most visited website

Facebook 1.09 billion daily active users

Twitter 100 million daily active users

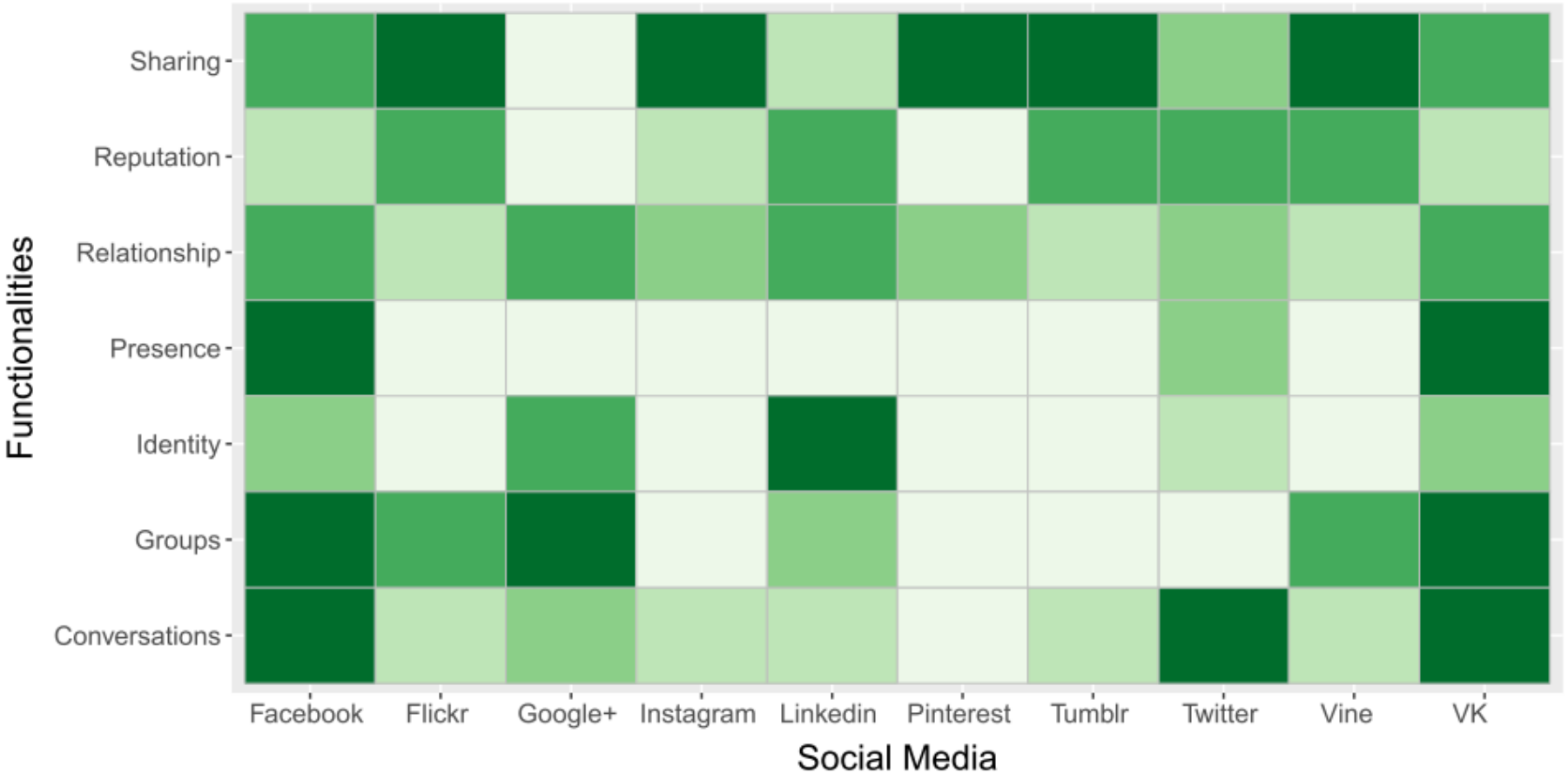


Social Media Functionalities



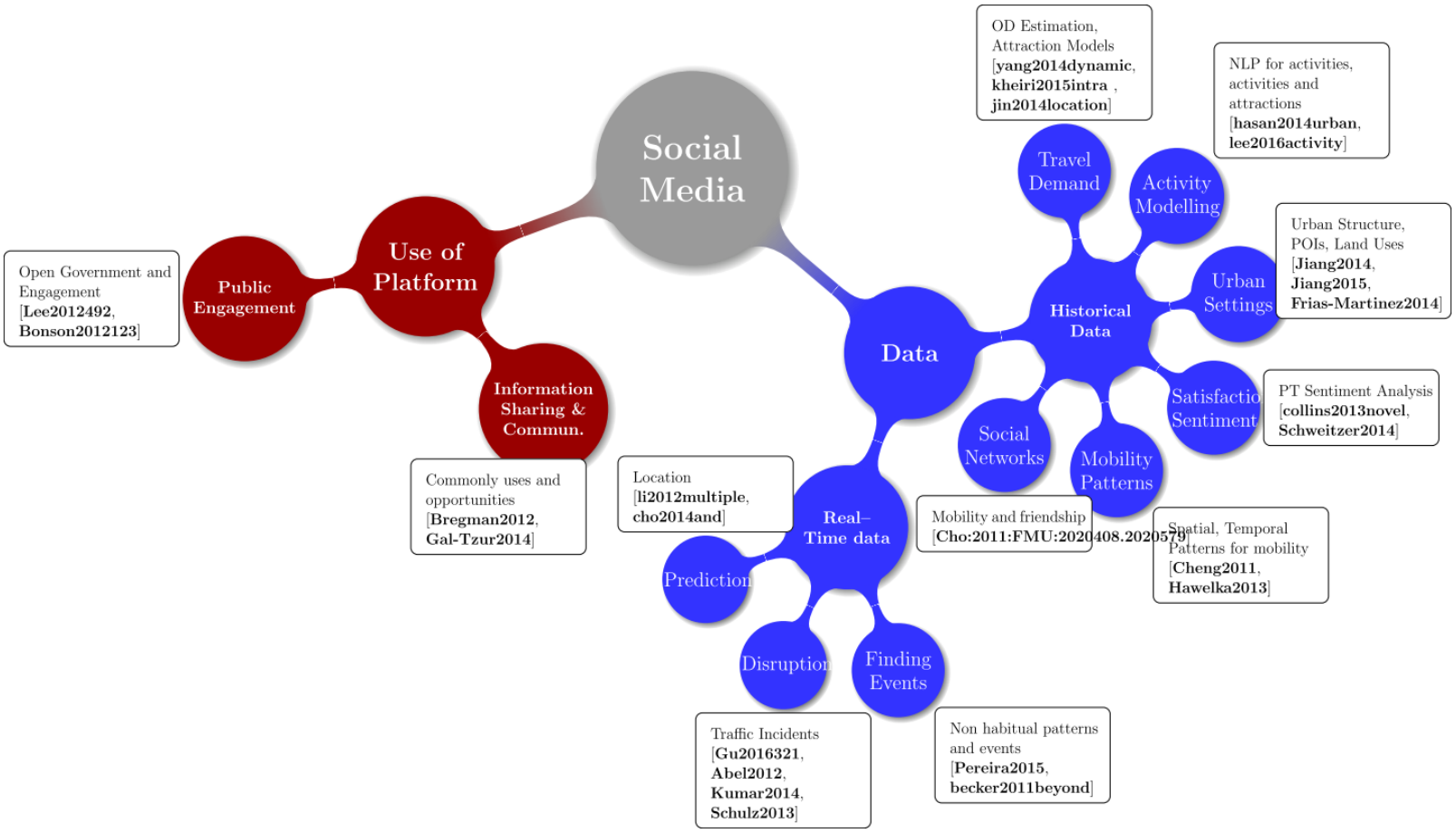
Chaniotakis, Antoniou, Pereira. "Mapping Social Media for Transportation Studies." *IEEE Intelligent Systems* 31.6 (2016): 64-70.

Social Media Data Availability



Chaniotakis, Antoniou, Pereira. "Mapping Social Media for Transportation Studies." *IEEE Intelligent Systems* 31.6 (2016): 64-70.

Social Media in Transportation



Chaniotakis, Antoniou, Pereira. "Mapping Social Media for Transportation Studies." *IEEE Intelligent Systems* 31.6 (2016): 64-70.

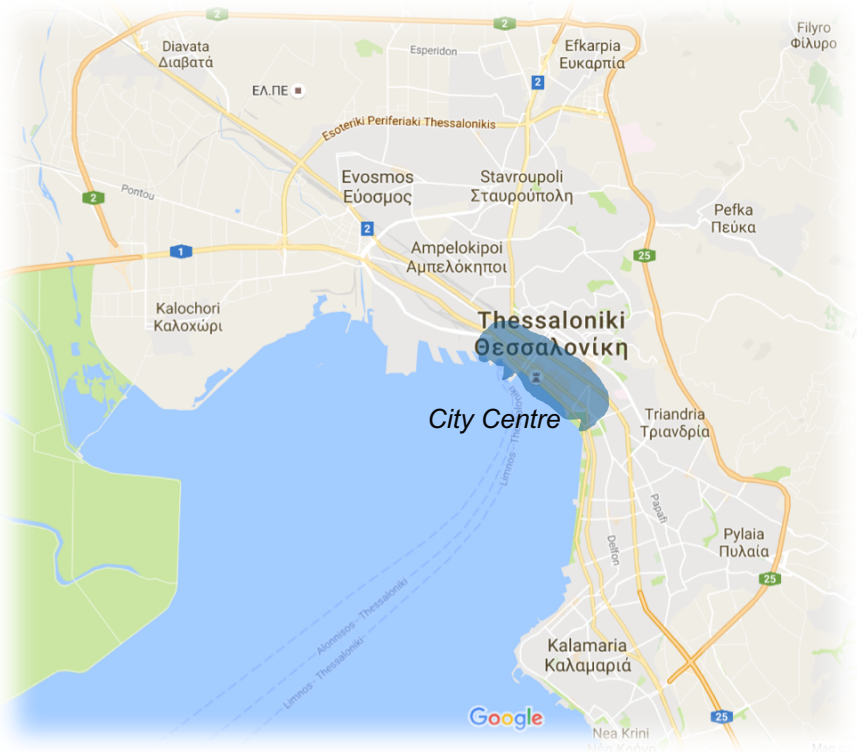
1st Case Study: SM vs. Surveys

Datasets

- 3 Social Media
  - Facebook
  - Twitter
  - Foursquare
- Recent travel diary survey

Study Area

- Thessaloniki, Greece
- 2nd largest city in Greece
- 1m inhabitants (metropolitan area)
- Moderate Social Media use



Chaniotakis, Antoniou, Salanova, Dimitriou, "Can Social Media data augment travel demand survey data?." *Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on*. IEEE, 2016.

Twitter Data

Data Collection of both Real-time and Historical (per user) data:

- 7,856 Twitter users using Real-time data collection (~1 year)
- User's timeline data collection (tweets not restricted to Thessaloniki)
- 49,169 geotagged tweets (only in Thessaloniki)

Type of information:

- Unique user ID
- unique tweet ID
- tweet text
- time of tweet
- (lon, lat)

Chaniotakis, Antoniou, Salanova, Dimitriou, "Can Social Media data augment travel demand survey data?." *Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on*. IEEE, 2016.

Facebook & Foursquare Data

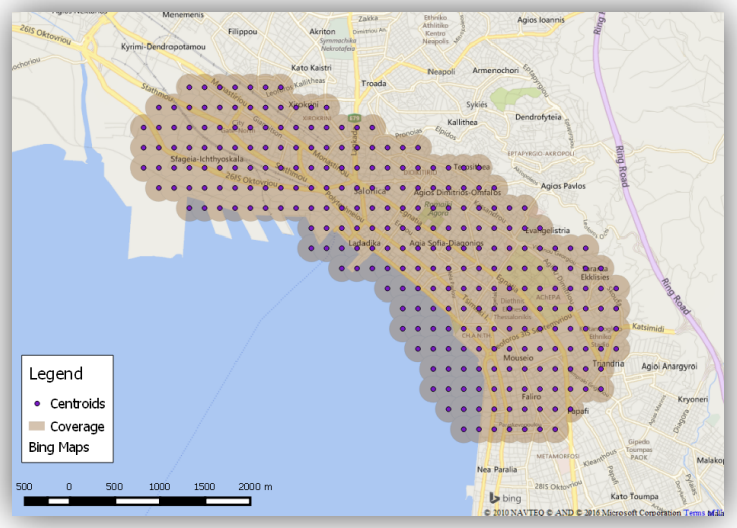
Data Collection of **venues** check-ins (public data)

- 30 min query interval
- Area based search → venues & characteristics
- Centre (lat,lon) & radius (150m)
- Only in the city centre of Thessaloniki
  (due to API request number limitations)

7,511 Venues from Facebook
9,135 Venues from Foursquare

Type of Information

- venue ID
- the venue name
- the venue category
- the total number of visitors

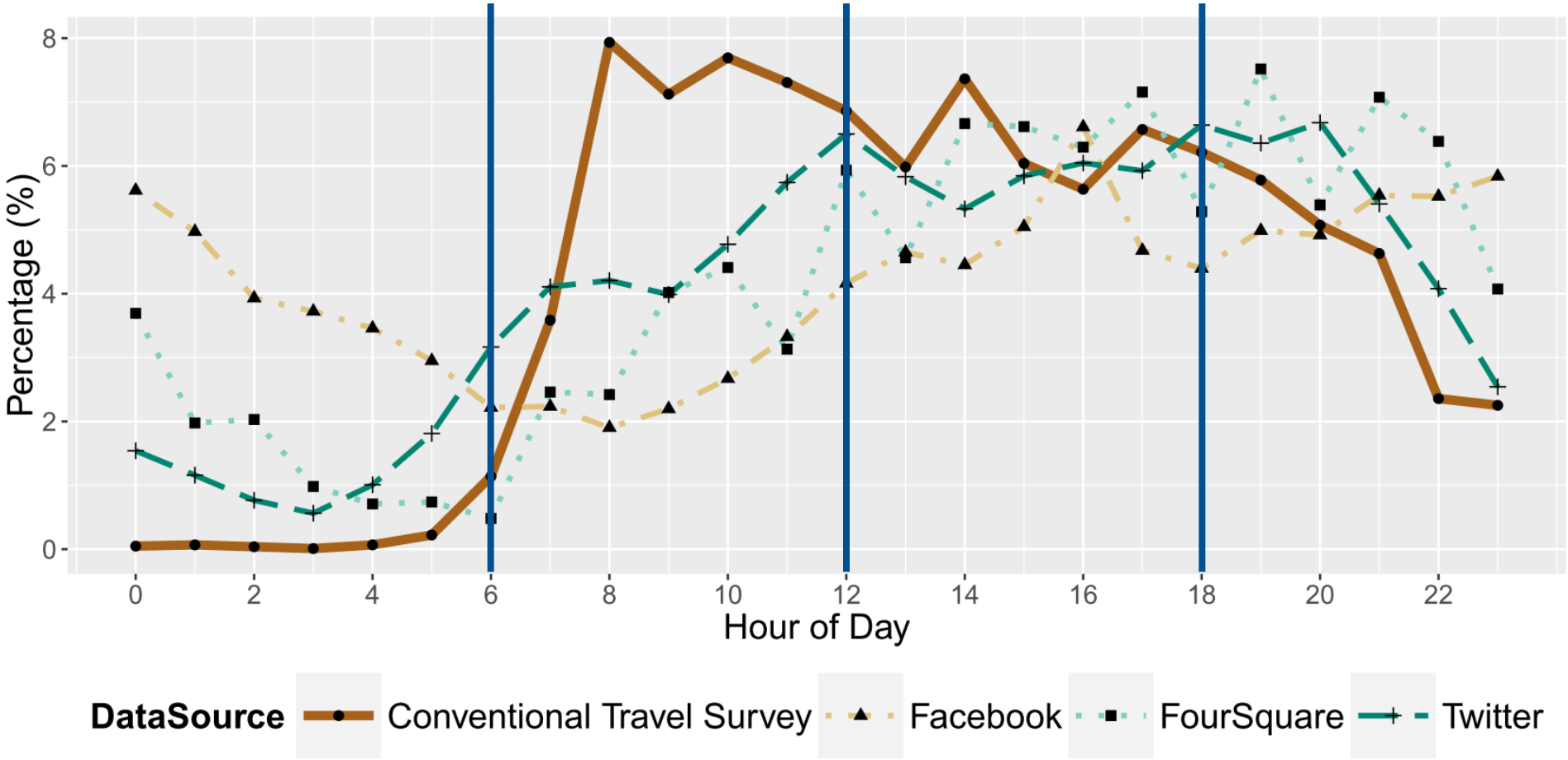


Chaniotakis, Antoniou, Salanova, Dimitriou, "Can Social Media data augment travel demand survey data?."
Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on. IEEE, 2016.

Activities

Chaniotakis, Antoniou, Salanova, Dimitriou, "Can Social Media data augment travel demand survey data?." *Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on*. IEEE, 2016.

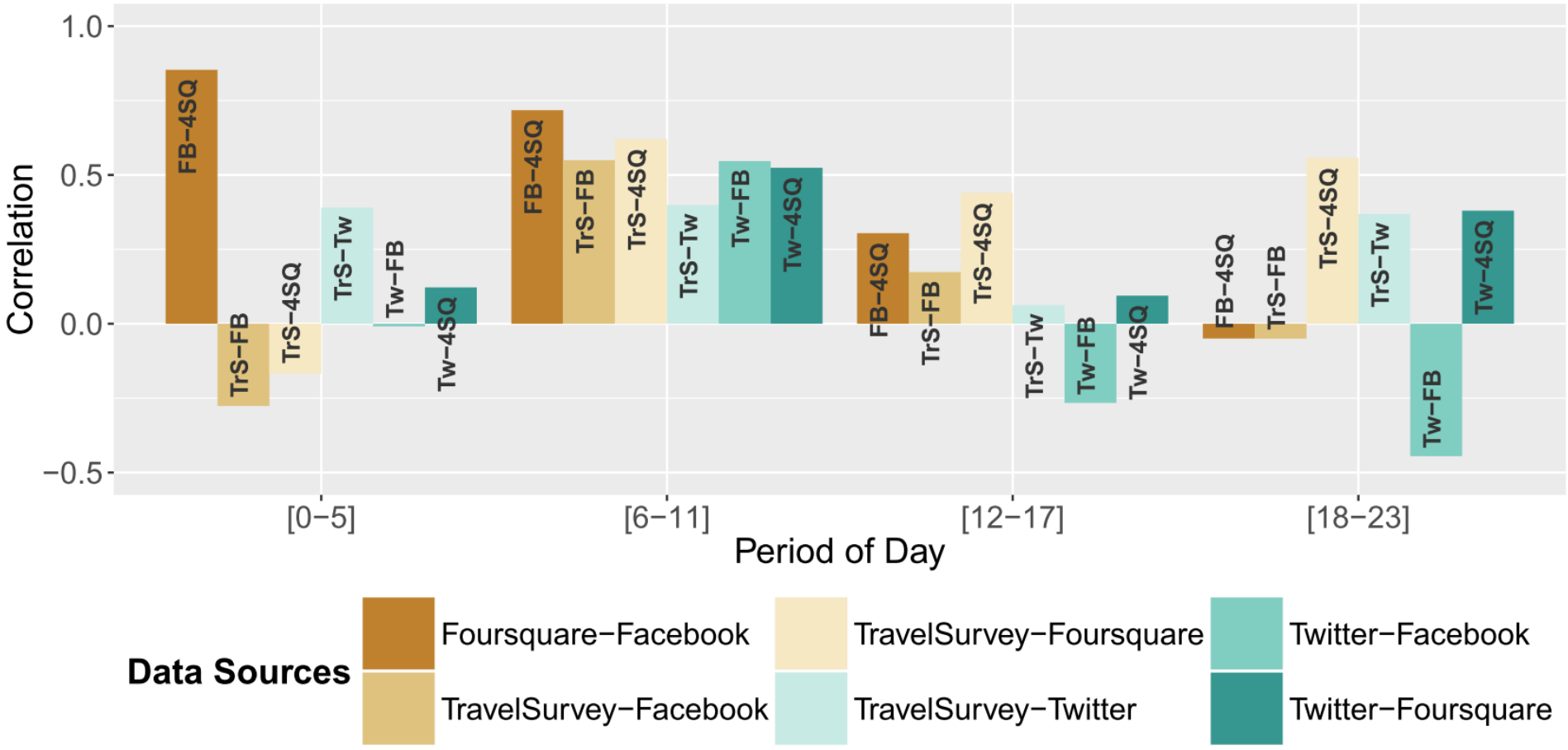
Temporal Analysis (1)



Chaniotakis, Antoniou, Salanova, Dimitriou, "Can Social Media data augment travel demand survey data?." *Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on*. IEEE, 2016.

Chair of Transportation Systems Engineering
Department of Civil, Geo and Environmental Engineering
Technical University of Munich

Temporal Analysis (2)



Chaniotakis, Antoniou, Salanova, Dimitriou, "Can Social Media data augment travel demand survey data?." *Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on*. IEEE, 2016.

2nd Case study: Inferring Activities from Social Media

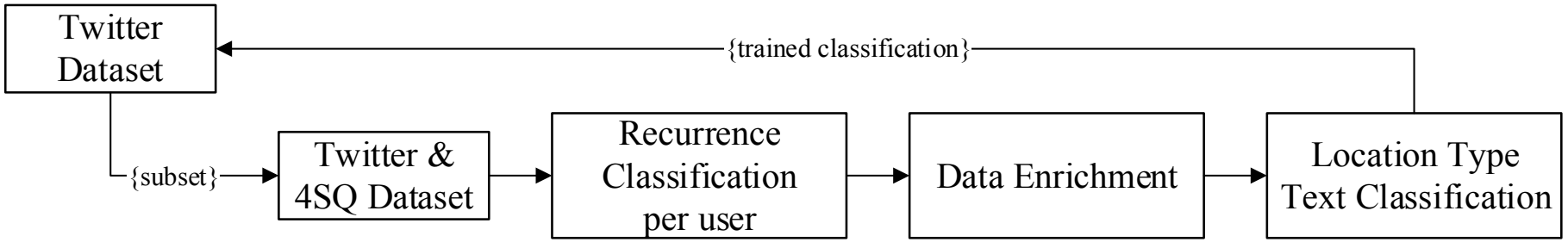
Text and geolocation offered by one SM platform

Location characterization is offered by another

The connection between them may be available (URL)

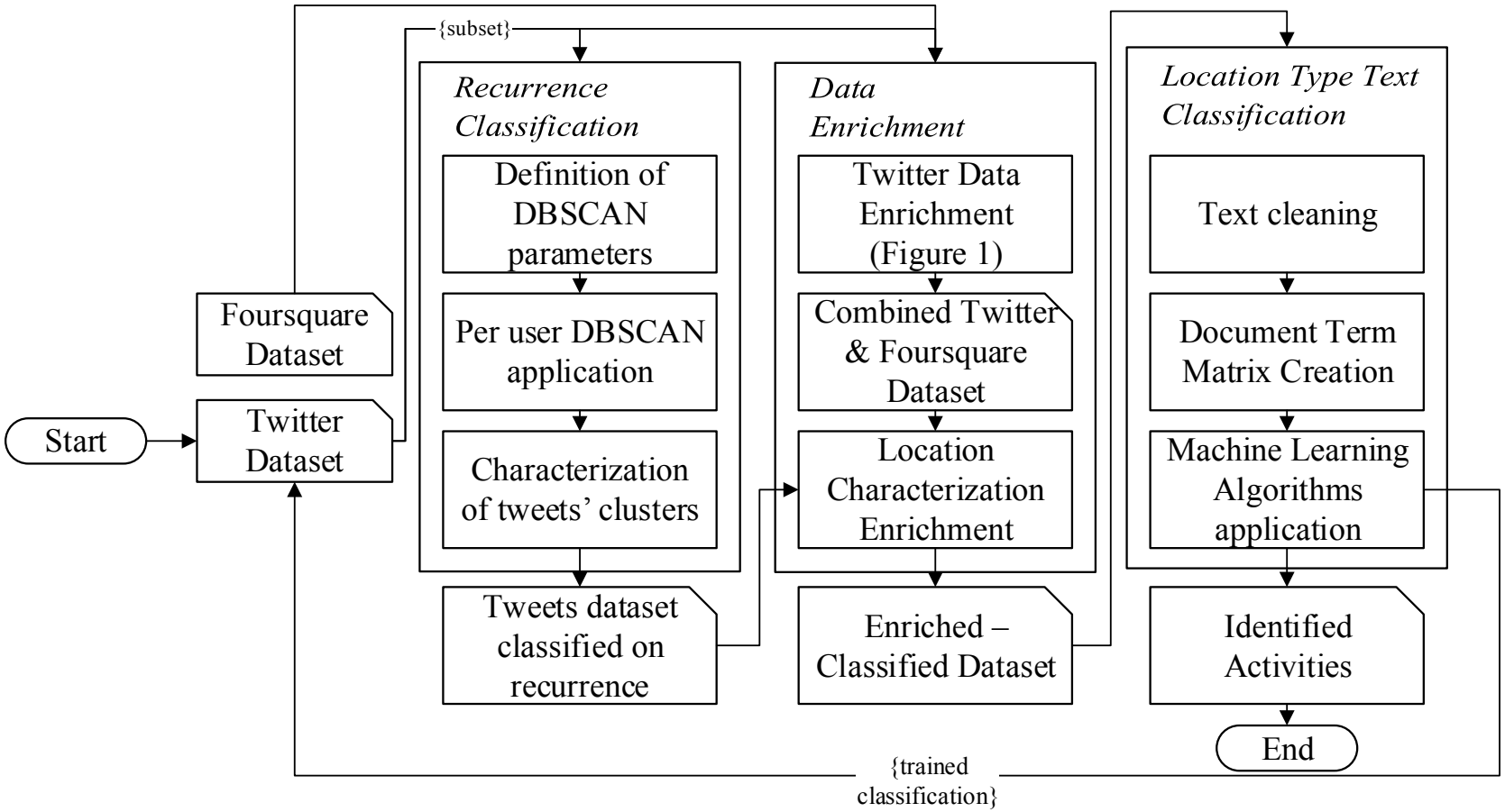
User-centric approach is possible.

Possibly the quietest Wagamama I've ever been in - and in the middle of London!! https://t.co/pY3UVIV4iw



Chaniotakis, E., Antoniou, C., Aifadopoulou, G., & Dimitriou, L., *2017* "*Inferring Activities from Social Media Data.*" In *Transportation Research Record* (accepted).

Inferring Activities from Social Media (2)



Chaniotakis, E., Antoniou, C., Aifadopoulou, G., & Dimitriou, L., *2017* "*Inferring Activities from Social Media Data.*" In *Transportation Research Record* (accepted).

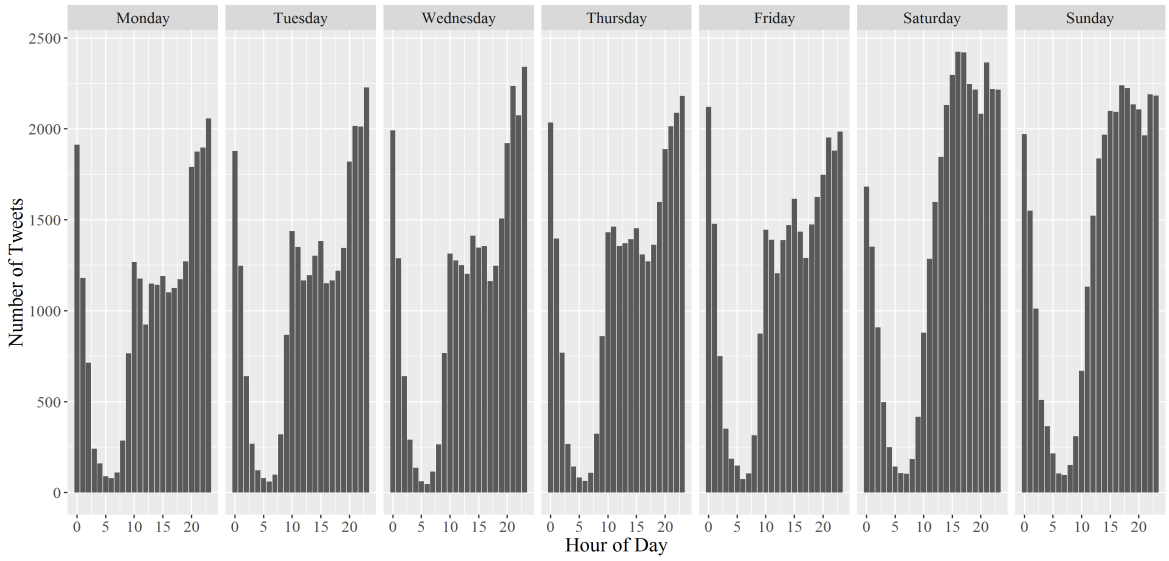
Datasets (1)

2 years' data collected from London (Twitter API)

482,883 unique users

Collected timeline (for a random sample of 90,000 users)

11,060,814 tweets in total

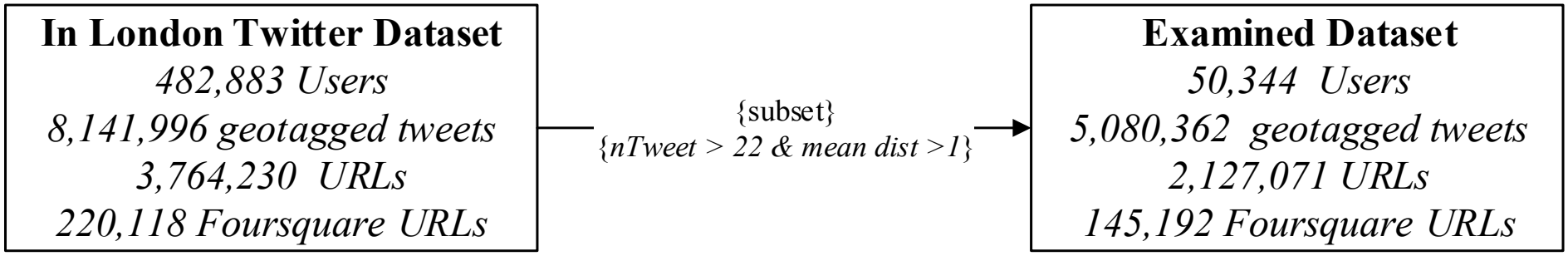


Chaniotakis, E., Antoniou, C., Aifadopoulou, G., & Dimitriou, L., *2017* "*Inferring Activities from Social Media Data.*" In *Transportation Research Record* (accepted).

Datasets (2)

Subset Formation

- 101.8 tweets with a standard deviation of 311.4
- 0.8% standard deviation of percentage posted per day (close to uniform)



Chaniotakis, E., Antoniou, C., Aifadopoulou, G., & Dimitriou, L., *2017 "Inferring Activities from Social Media Data.*"
In *Transportation Research Record* (accepted).

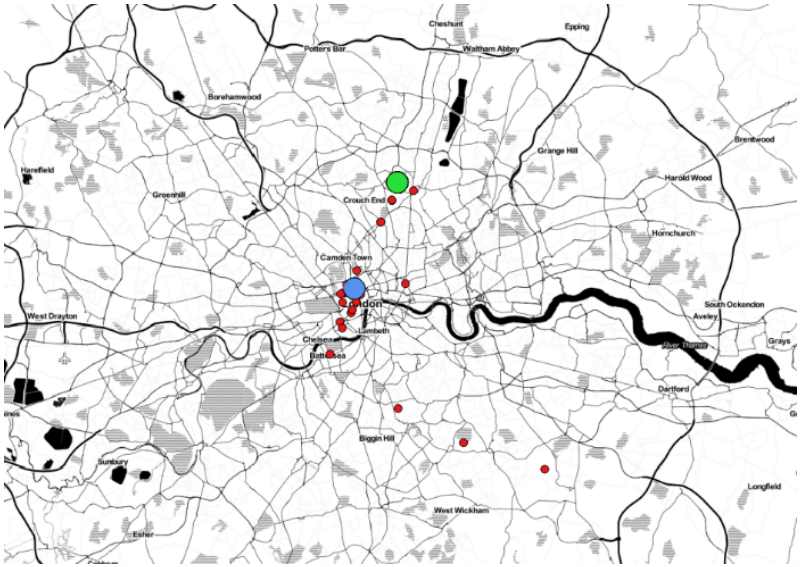
Recurrence Classification (1)

8.3% of subset have posted at least one 4SQ link

39% of Twitter-4SQ users' posts include 4SQ link

Density–based Spatial Clustering of Applications with Noise

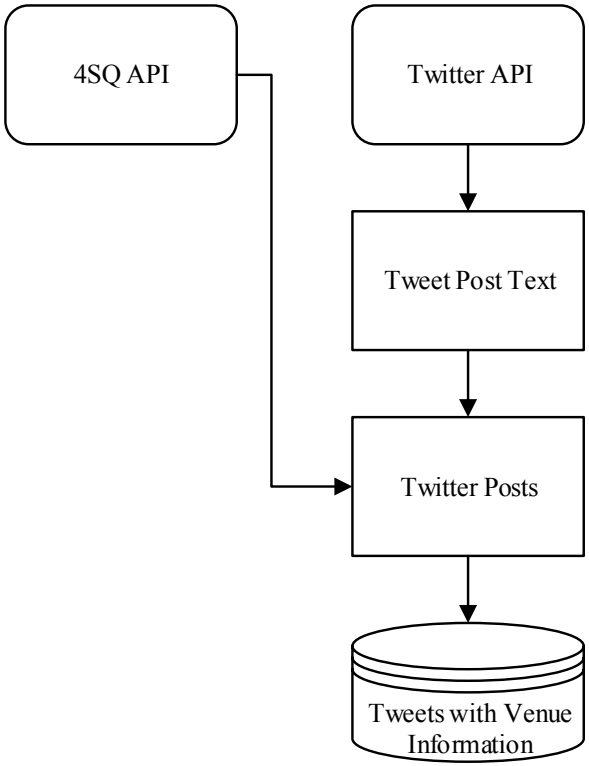
- Point parameter (*Eps*) = 0.002 (GPS accuracy)
- Clustering > 5 posts per location

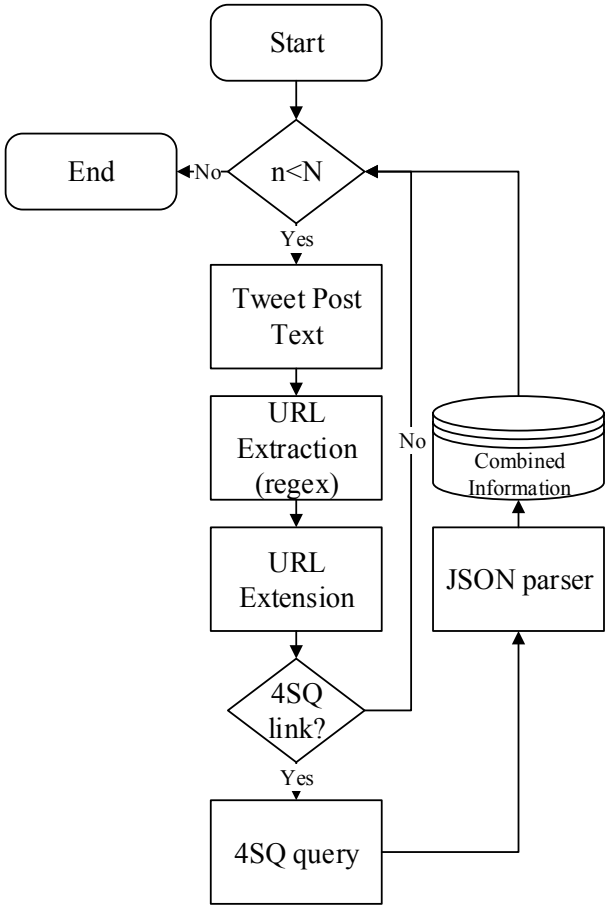


Chaniotakis, E., Antoniou, C., Aifadopoulou, G., & Dimitriou, L., *2017* "*Inferring Activities from Social Media Data.*" In *Transportation Research Record* (accepted).

Data Enrichment

Combine 4SQ & Twitter Data





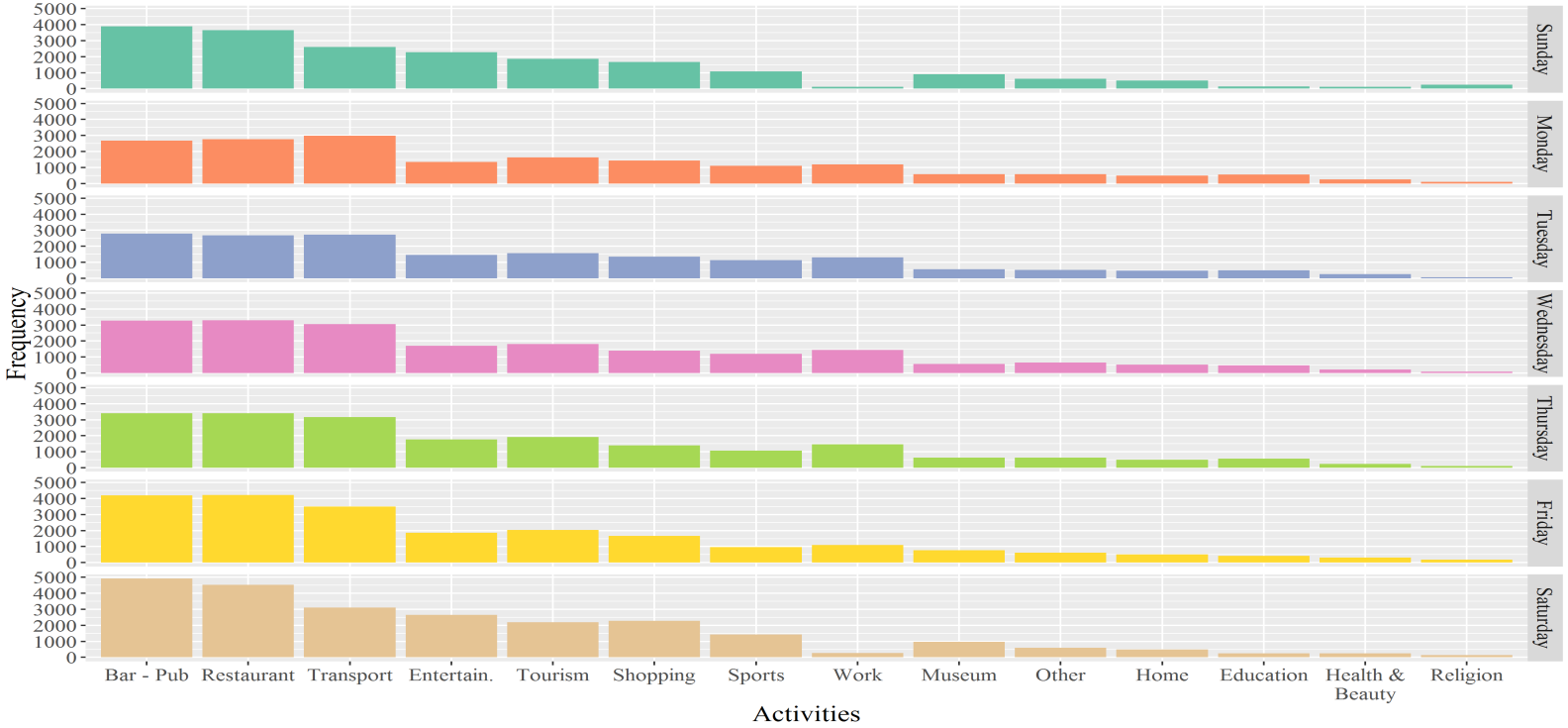
Chaniotakis, E., Antoniou, C., Aifadopoulou, G., & Dimitriou, L., *2017 "Inferring Activities from Social Media Data.*"
In *Transportation Research Record* (accepted).

4SQ Activities Distribution

Manually aggregated in 14 categories

Tendency towards leisure activities

Education and work higher represented during week days



Chaniotakis, E., Antoniou, C., Aifadopoulou, G., & Dimitriou, L., *2017* "*Inferring Activities from Social Media Data.*" In *Transportation Research Record* (accepted).

Location Type Classification

Linked tweet text with activities

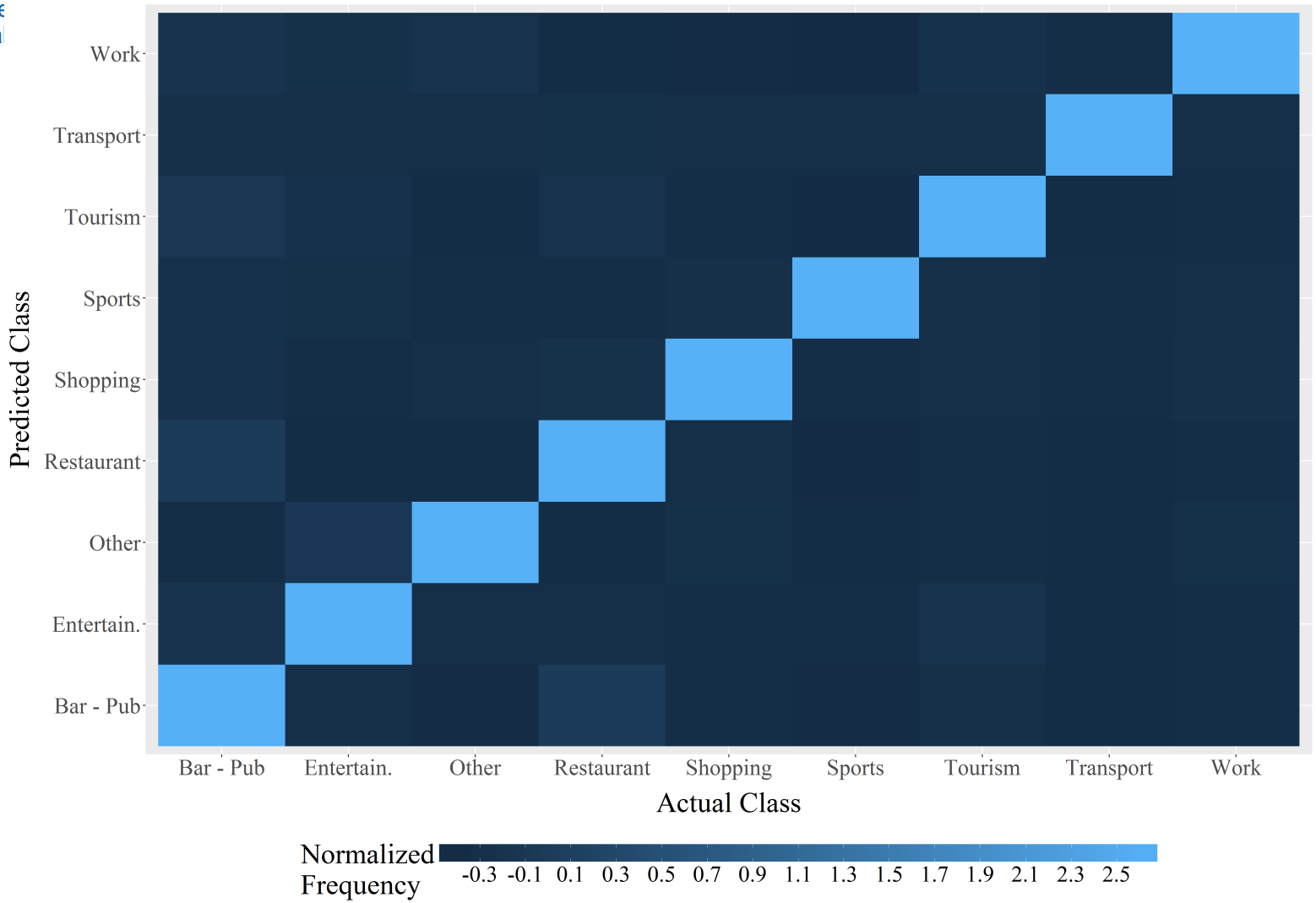
Removed stop words, punctuation

Automatic Text Classification:

- Support Vector Machine (SVM)
- Generalized Linear Model via Penalized Maximum Likelihood (GLMPML)
- Maximum Entropy (MAXENT)

- RTextTools library
  - Repeated 10 times
  - 20,000 randomly selected cases
  - 85% – (17,000 entries) training set
  - 15% – (3000 remaining entries) testing set

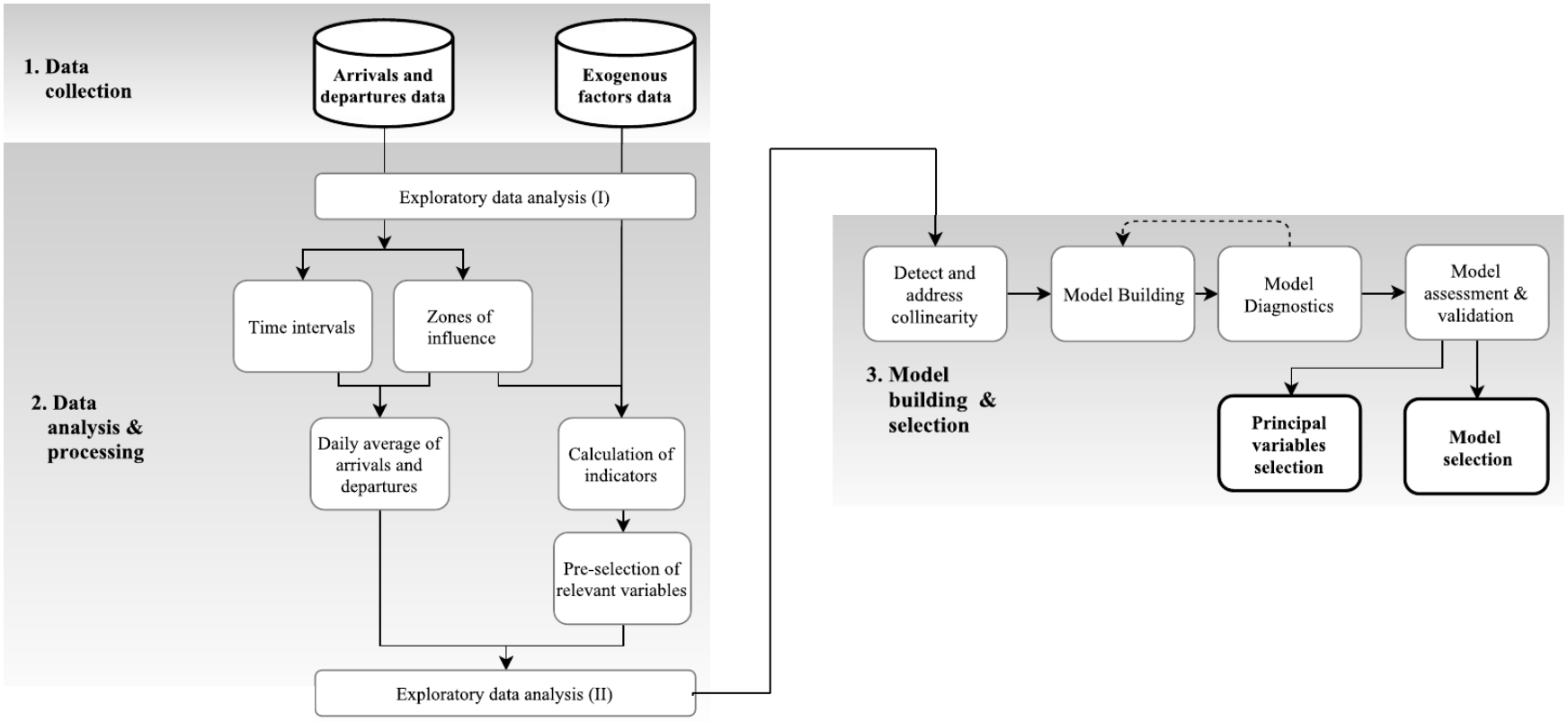
Chaniotakis, E., Antoniou, C., Aifadopoulou, G., & Dimitriou, L., *2017* "*Inferring Activities from Social Media Data.*" In *Transportation Research Record* (accepted).



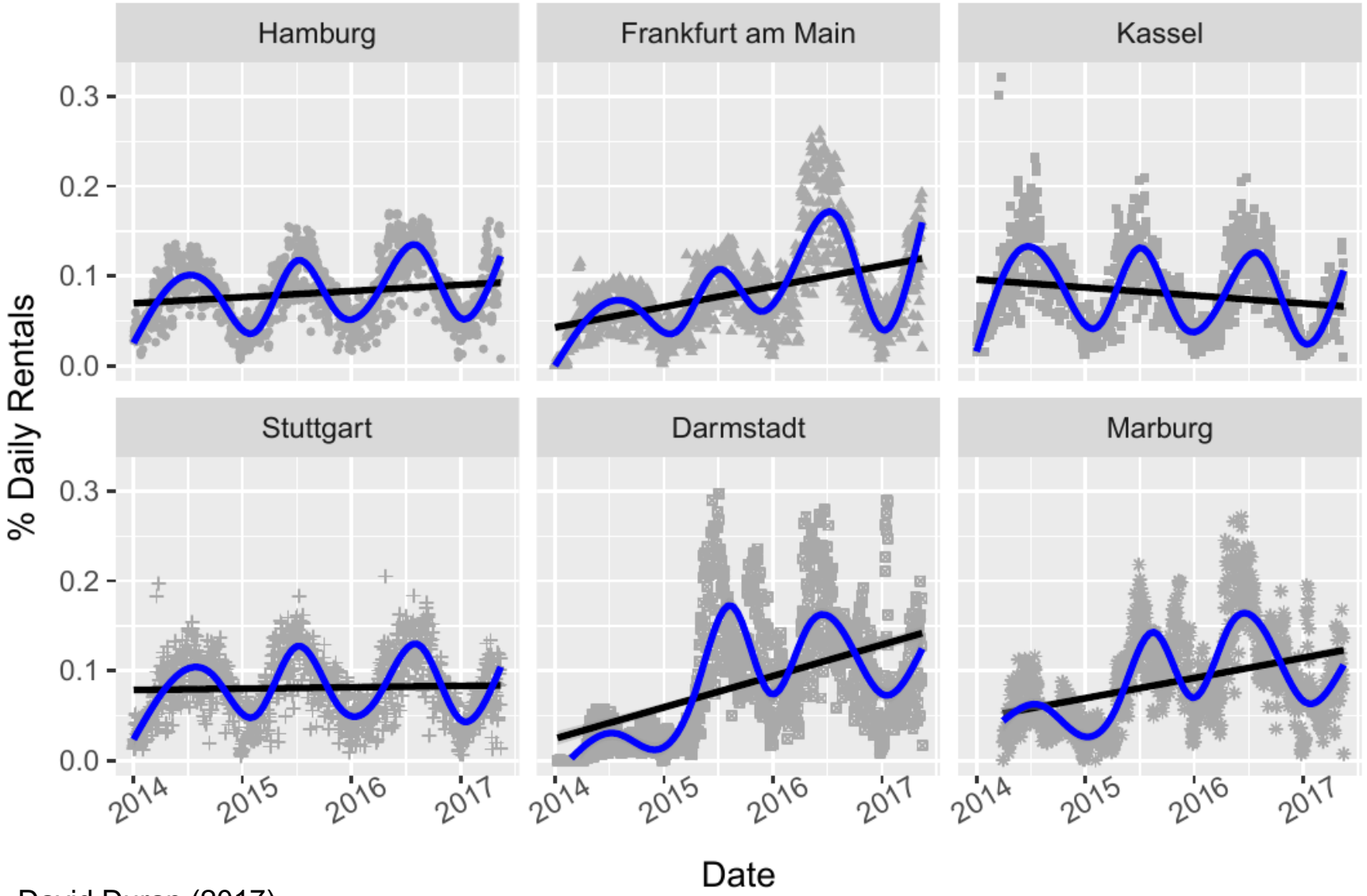
Max Entropy

Chaniotakis, E., Antoniou, C., Aifadopoulou, G., & Dimitriou, L., *2017* "*Inferring Activities from Social Media Data.*" In *Transportation Research Record* (accepted).

Factors affecting bike-sharing – Methodological framework

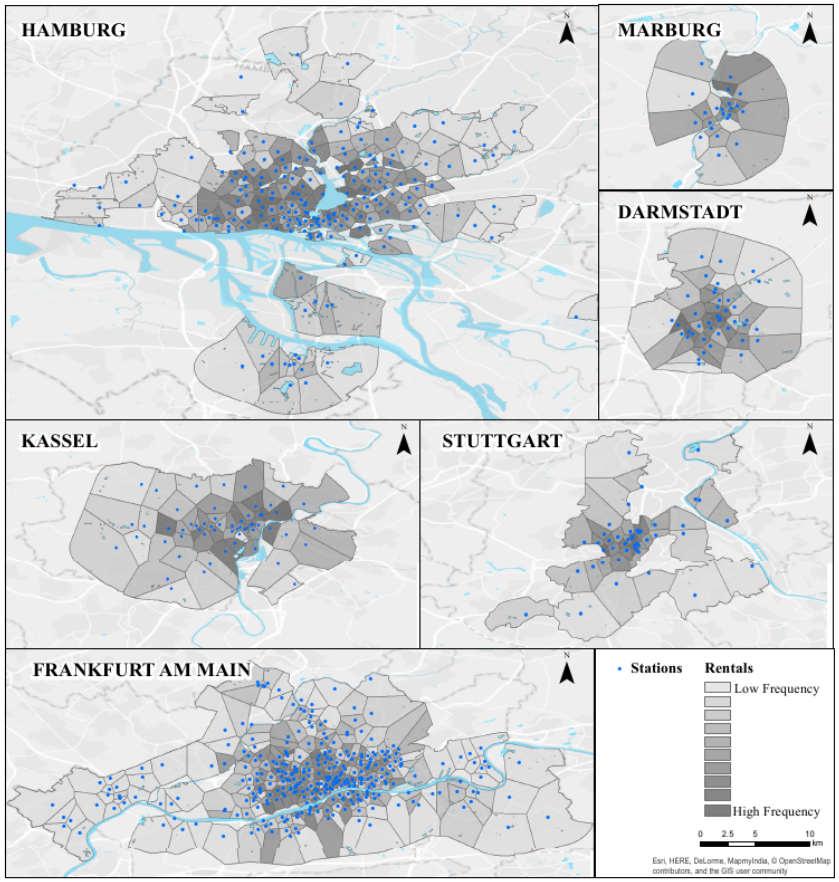


David Duran (2017)



David Duran (2017)

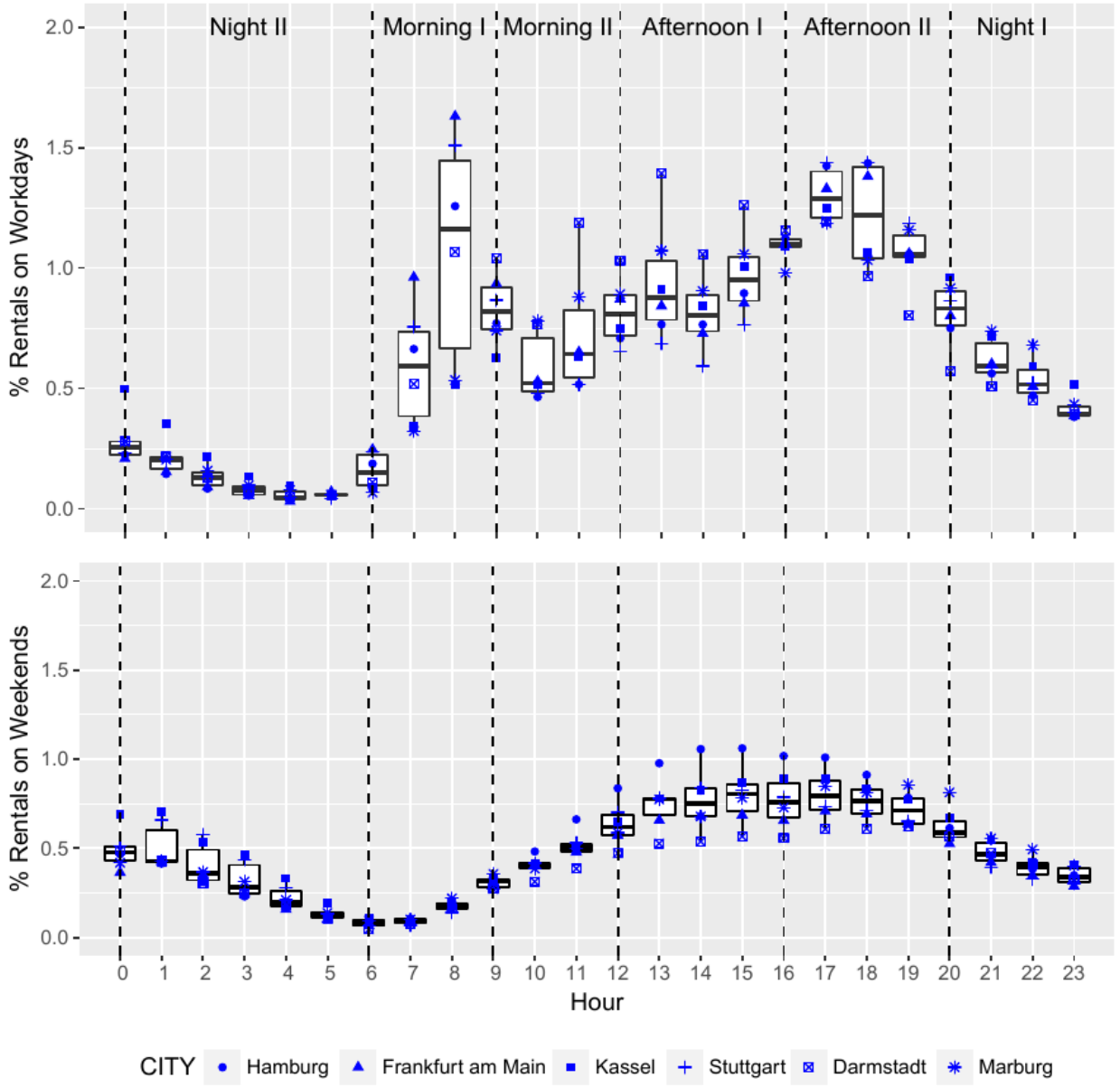
Dependent variables



David Duran (2017)

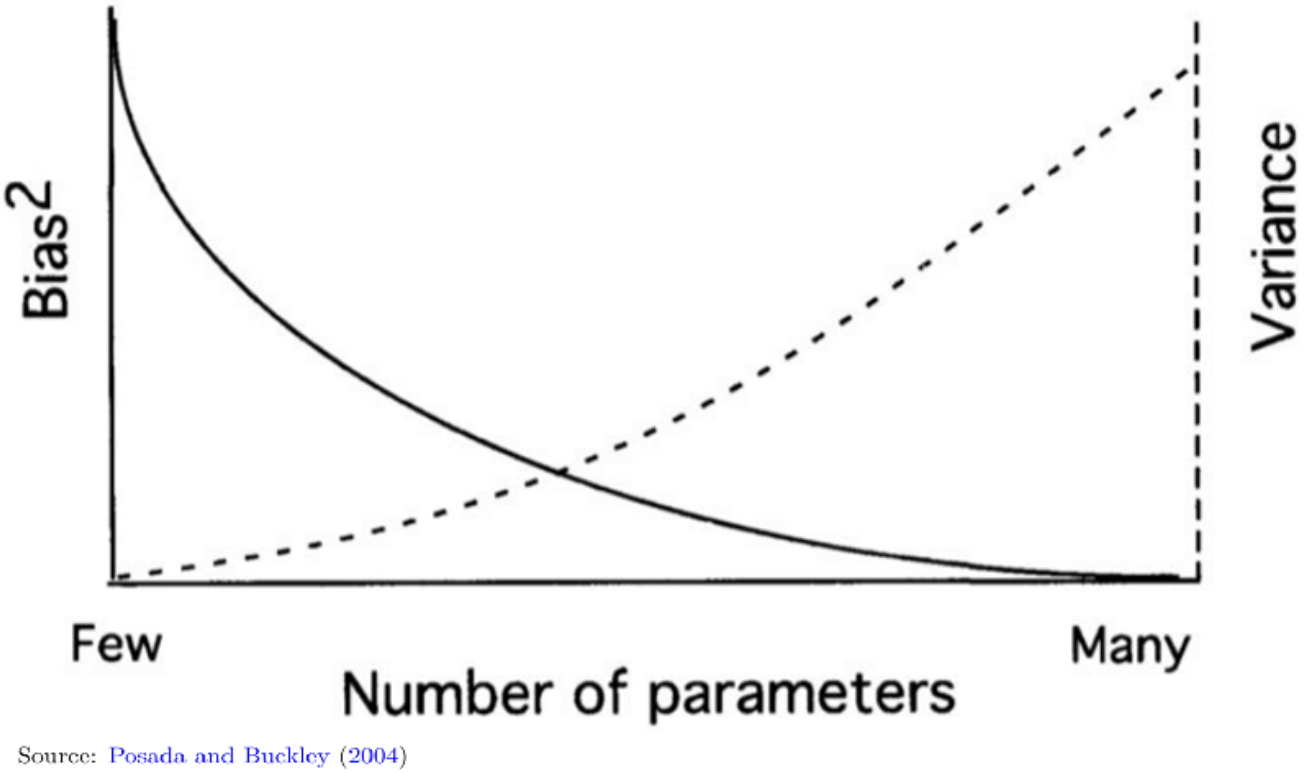
Independent variables

Exogenous Factors
City Size
Climate
Mobility behavior
Population density
Demographic factors
Economic factors
Geographic factors (hilliness)
Existing Infrastructure
Financial Situation
Political Situation



David Duran (2017)

Model selection criteria

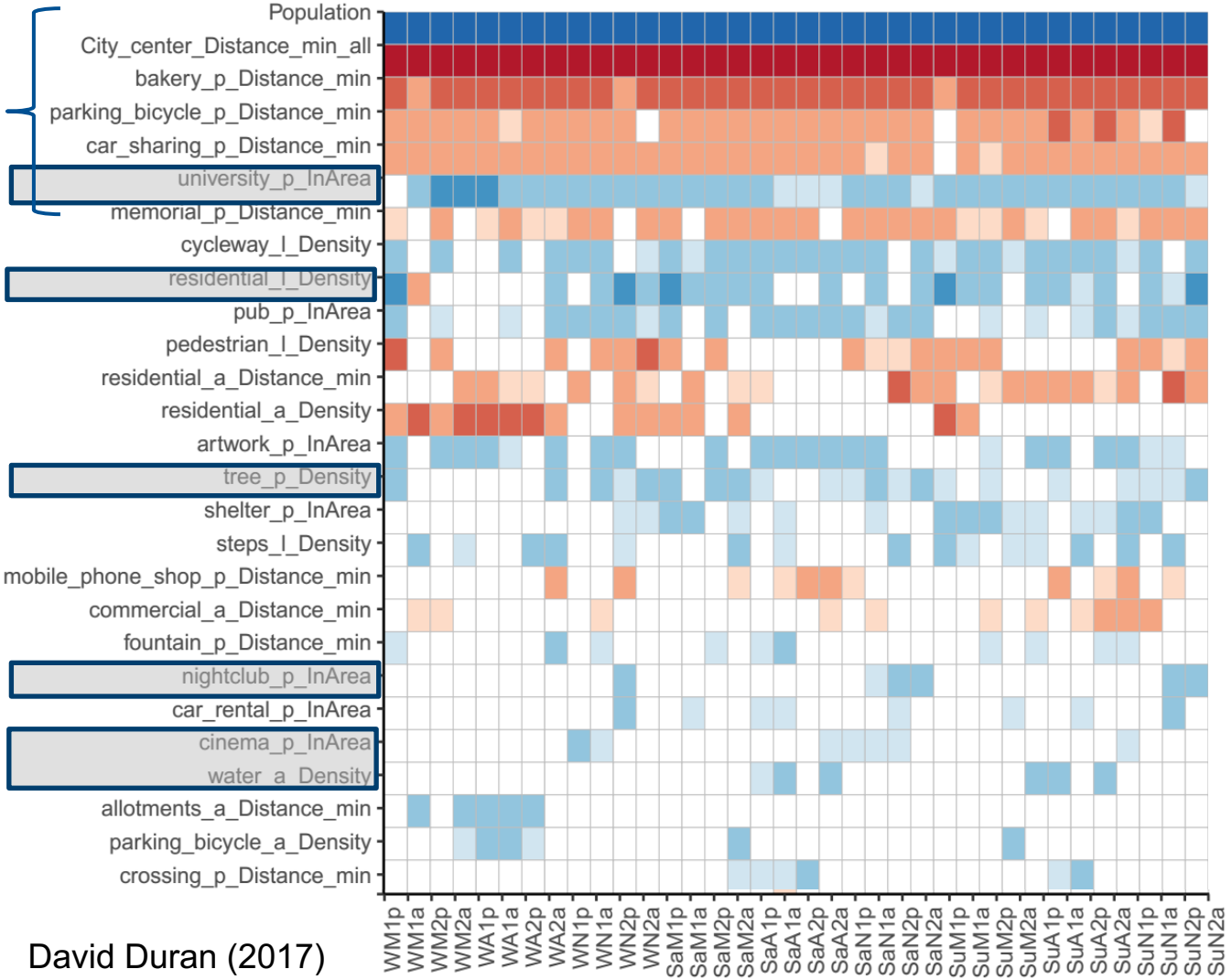


Source: Posada and Buckley (2004)

David Duran (2017)

Most influential variables selection

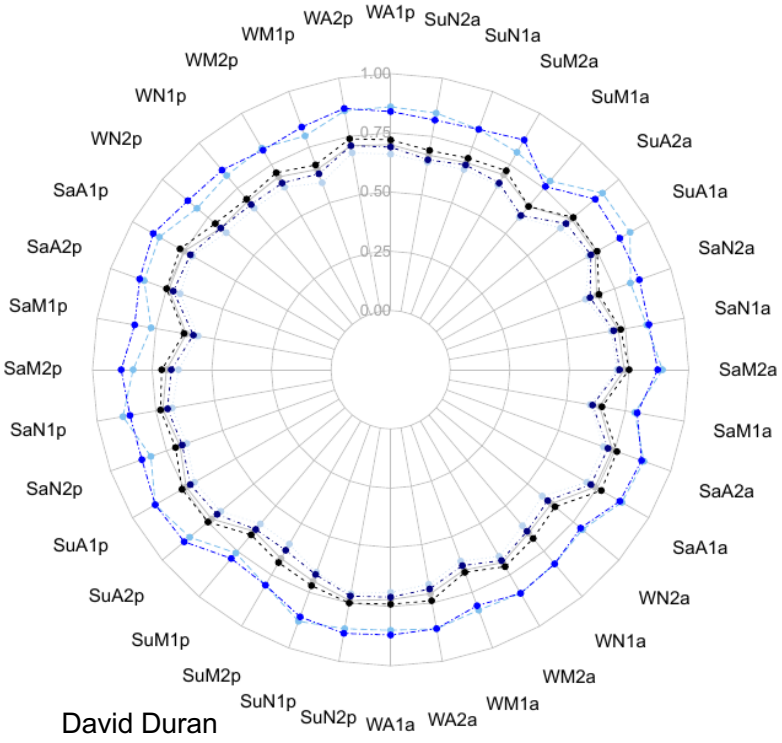
Stepwise+log



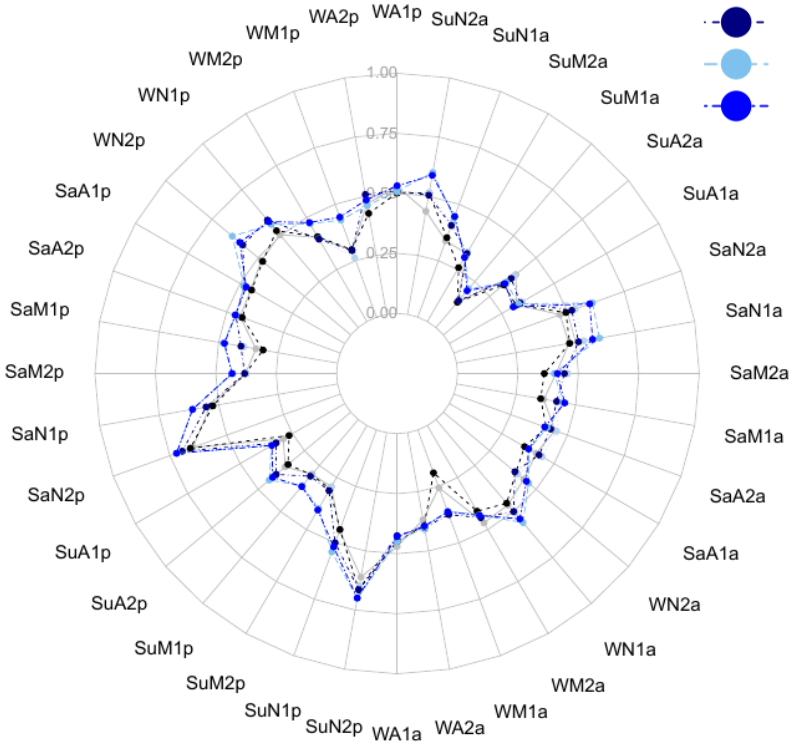
p	point
l	line
a	area
InArea	Presence in zone of influence
Density	Density in zone of influence
Distance_min	Distance to closest feature in zone of influence
Distance_min_all	Distance to closest feature in the city

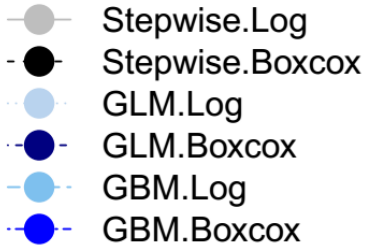
David Duran (2017)

R^2 adjusted



R^2 from validation





David Duran

Mobility-related feature extraction from emerging data

Prof. Constantinos Antoniou

Chair of Transportation Systems Engineering

Department of Civil, Geo and Environmental Engineering

Technical University of Munich

c.antoniou@tum.de