Understanding semantic similarity among subway stations using smart card data

Reporters: Zhuang Dingyi, Siyu Hao

Supervisors: Prof. Lee Der-Horng, Prof. Jiangang Jin

Department of Civil and Environmental Engineering

National University of Singapore

Contents

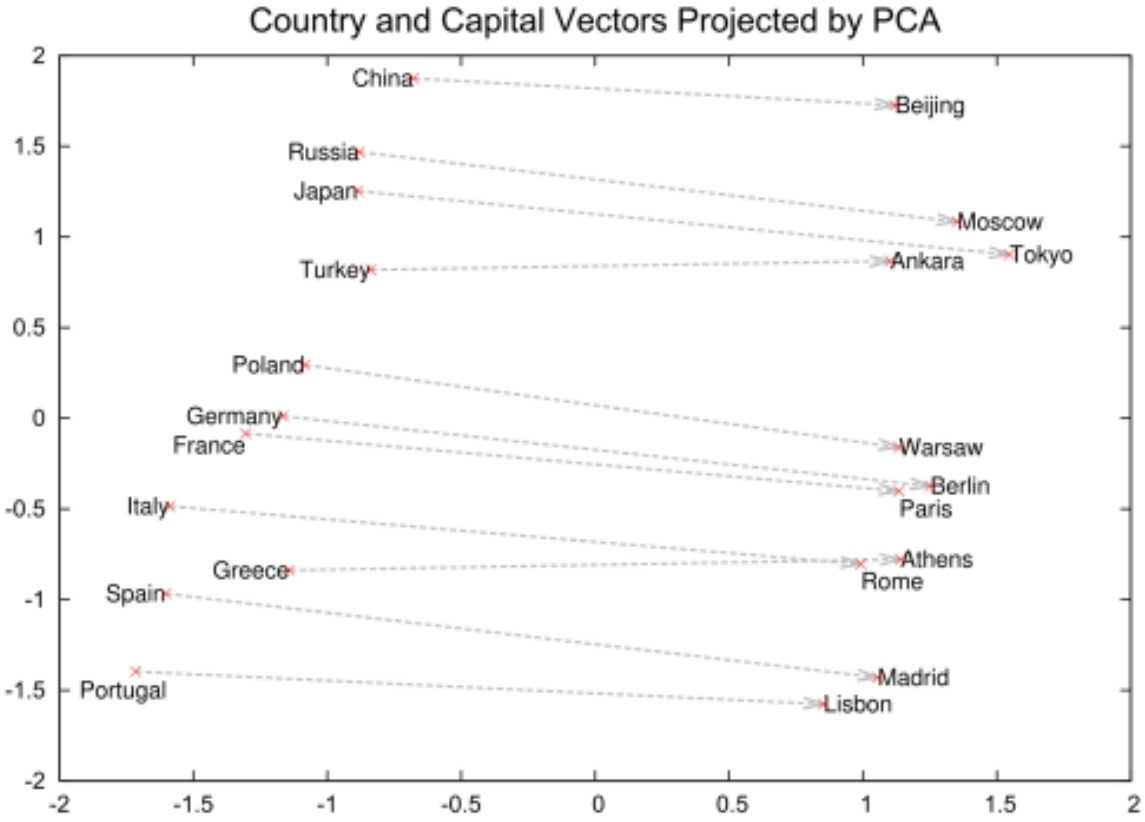
Motivation

Literature review

Contribution

1. INTRODUCTION

Motivation



Words → **Embedding** → Word vectors → **Clustering** → Semantic similarity → Stations?

Sequential features

Literature review

Previous station similarity analysis is based on shallow mobility features, such as aggregated passenger flow

Mohamed, K.et. al. in Clustering smart card data for urban mobility analysis.

Some transferred semantic models into urban computing, but regarding stations as documents and lack of further comprehensive analysis

Wang, J., Kong, X., Rahim, A., Xia, F., Tolba, A., & Al-Makhadmeh, Z. (2017). IS2Fun: Identification of Subway Station Functions Using Massive Urban Data. IEEE Access, 5, 27103-27113.

Semantic models are now widely applied in fields outside Natural Language Processing

Yuan, N. J., Zheng, Y., & Xie, X. (2018). Discovering Functional Zones in a City Using Human Movements and Points of Interest. In Spatial Analysis and Location Modeling in Urban and Regional Systems (pp. 33-62). Springer, Berlin, Heidelberg.

Contribution

Concept

Stations are like Chinese characters or compound words

Words (Stations)

- Meaning in sentence (Mobility pattern)
- Literal meaning, e.g. superman=super+man (Inherent features like POI)

Case studies

Analysis on similarity between MRT stations of Singapore in a planning perspective:

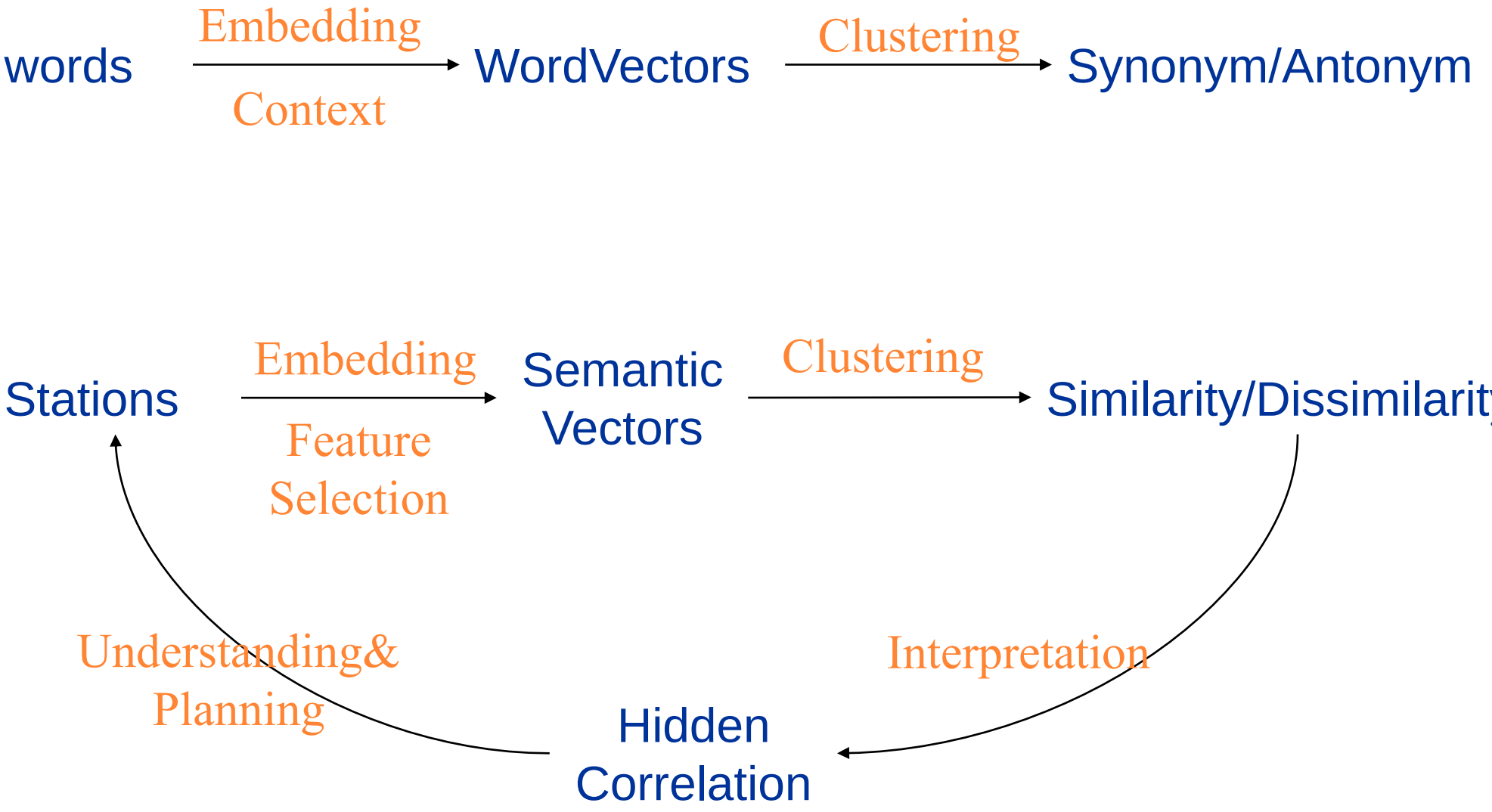
- 9 POI categories
- 5 case studies
- Planning suggestions

Expected results

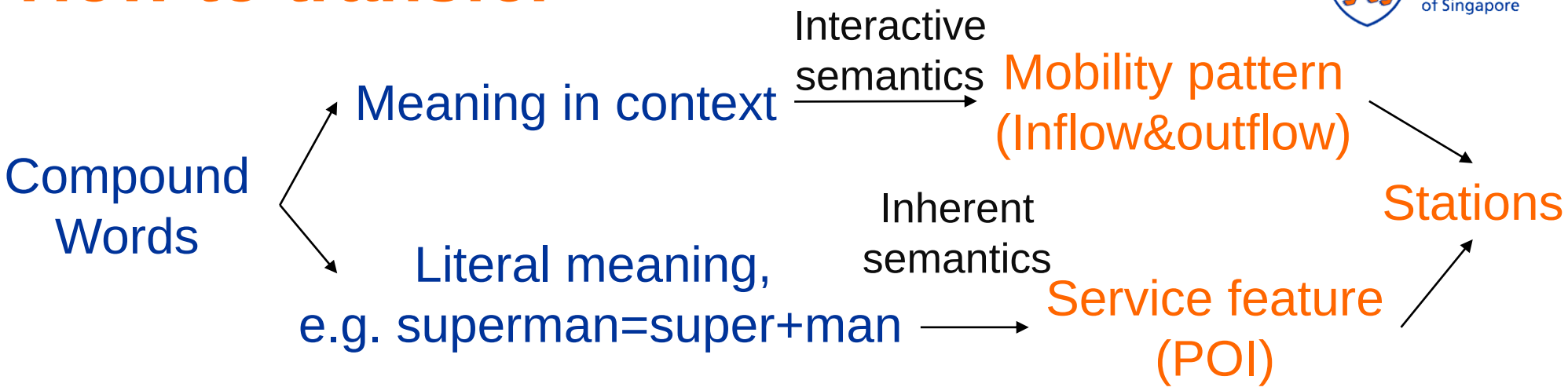
How to transfer

2. RESEARCH IDEAS

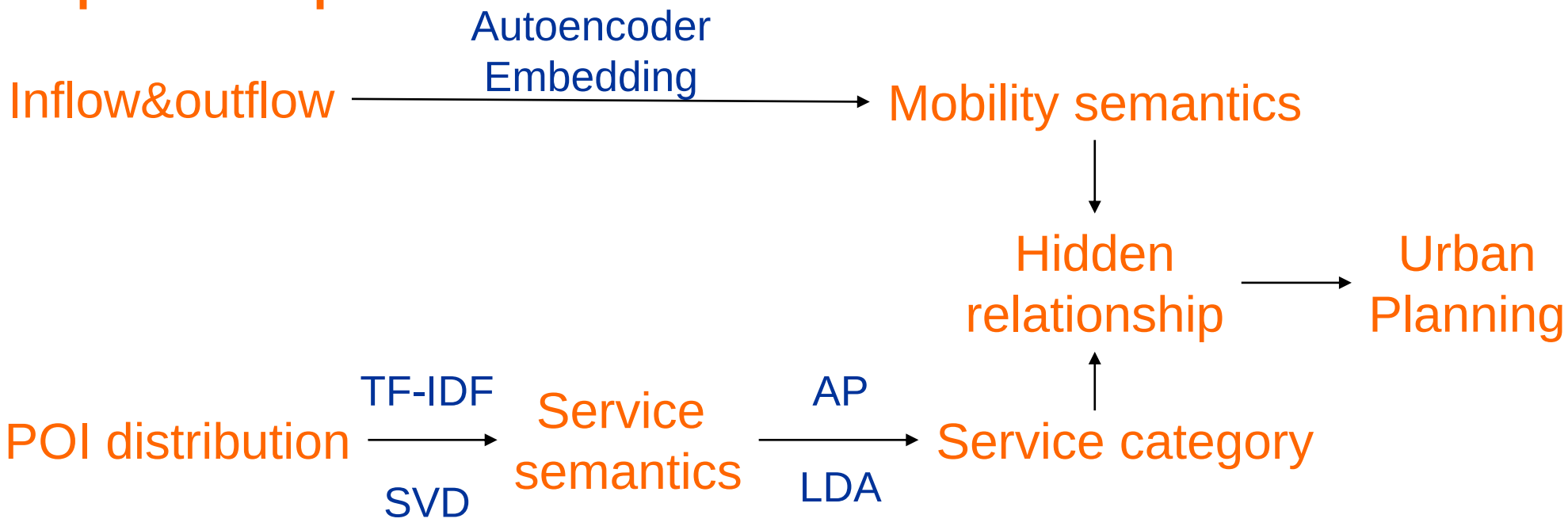
Expected results



How to transfer



Proposed steps:



Dataset

Stacked autoencoder

Mobility semantics

Service semantics

Case studies

3. RESEARCH RESULTS

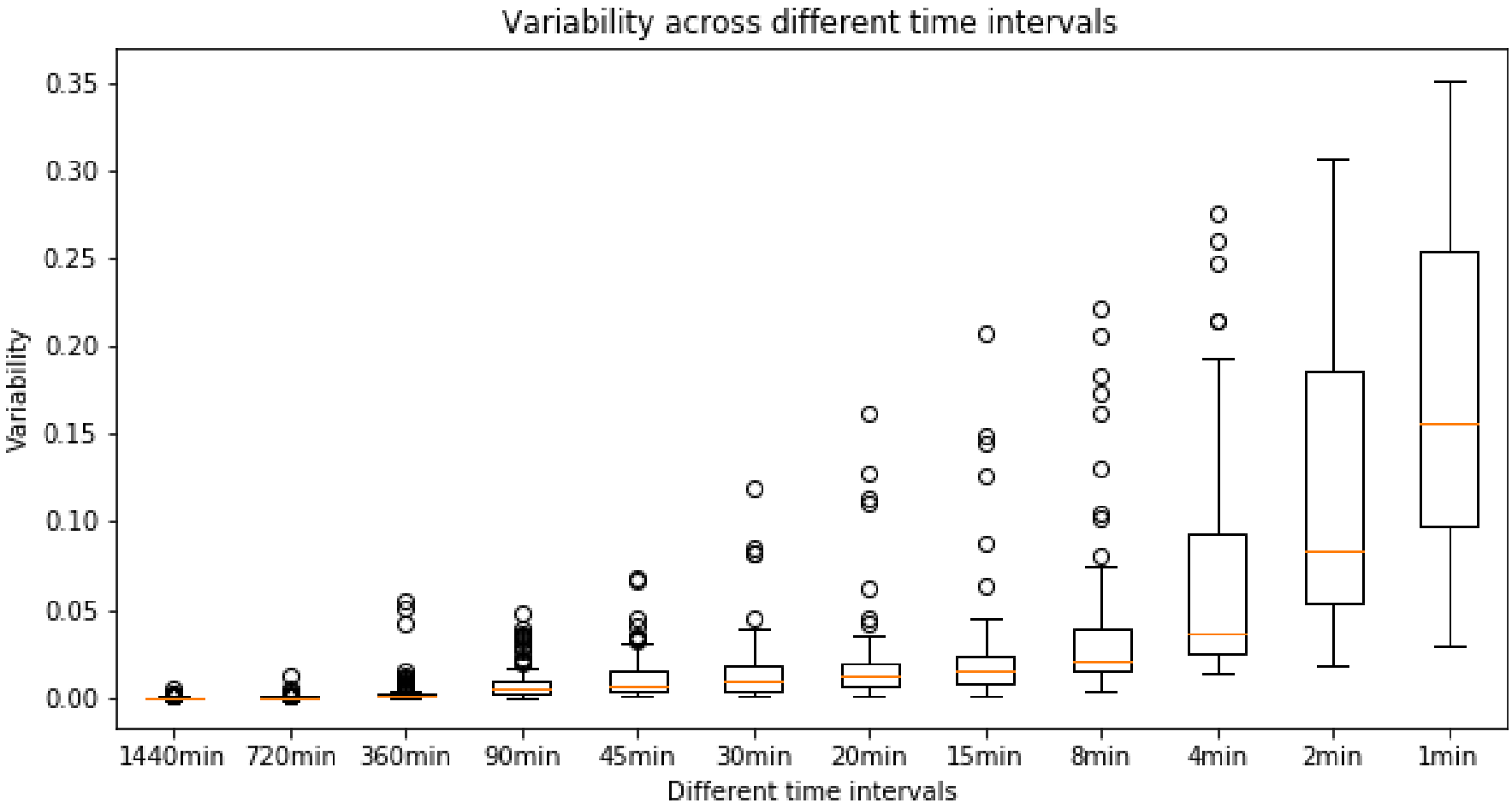
Dataset

Provided by Land Transport Authority (LTA), Singapore. Multi-model data (Bus&MRT), we only considered MRT.

Table for dataset description

Description	Value
Covered days	2012/3/19-2012/3/25 (Normal week)
Covered Stops	4702 (122 for MRT stations)
Average records number each day	>5,000,000
Data volume	4.1 GB
Average multi-model riding distance	7 km
Average multi-model riding time	20 min
Multi-model transferring percentage	30%
Average MRT riding distance	12 km
Average MRT riding time	27 min
MRT transferring percentage	23%

Dataset

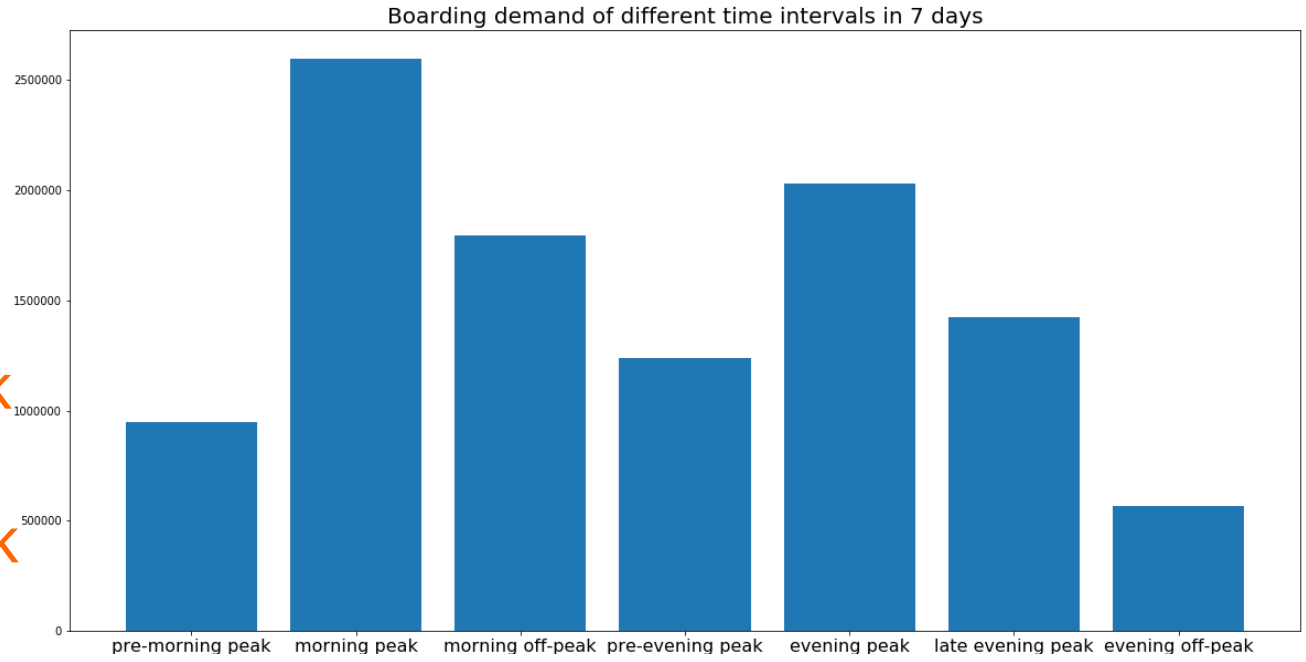


To better under stand temporal influence ,inspired by *Mohamed, K.et. al. in Clustering smart card data for urban mobility analysis*. We choose 1-3 hours as our time interval (LOW VARIABILITY).

Dataset

We divide the time into 7 time intervals:

- 5-7 pre-morning peak
- 7-10 morning peak
- 10-16 morning off-peak
- 16-17 pre-evening peak
- 17-19 evening peak
- 19-22 late evening peak
- 22-24 evening off-peak



Mobility vectors of same time intervals: m*n dimension, where

m: 122*7= 854 (122 stations * 7 days)
n: 122+122+7= 251 (inflow&outflow from&to all stations + one-hot code for day)

Dataset

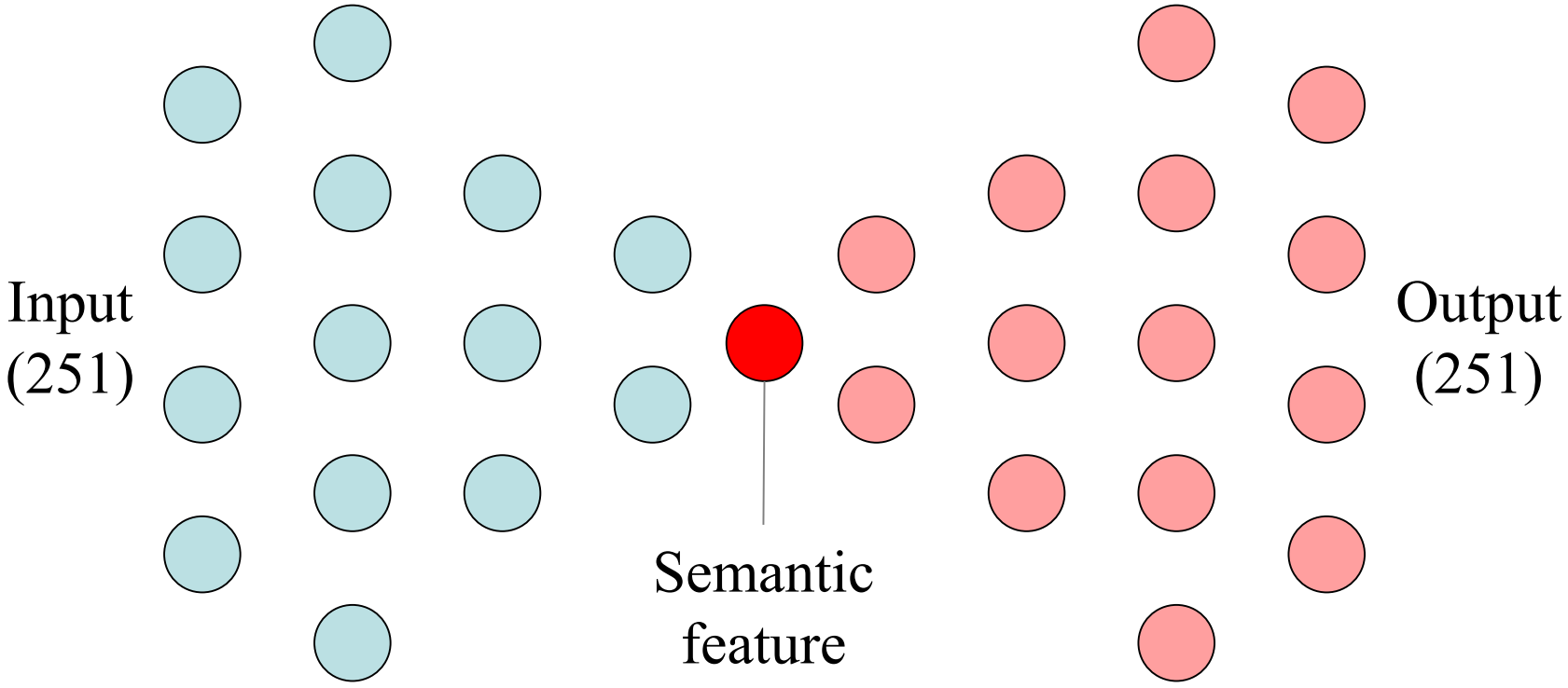
POI dataset is powered by Google Maps, contains:

- 22 categories
  - 'atm','bank','bus_station', 'transit_station', 'place_of_worship' 'supermarket', 'shopping_mall', 'education' , 'parking', 'park', 'political', 'storage', 'intsec','lodging','hospital','car_rental', 'car_dealer','car_repair','bar','cafe','local_government_office','bicycle_store'

- 10 MRT lines
  - 'NS','EW','NE','CC','CE','BP','CG','PE','SW','SE'

Stacked autoencoder

Reduce the dimension of flow vectors from 251 into 16. Train 7 models for 7 time intervals respectively. Train data use Min-Max normalization.

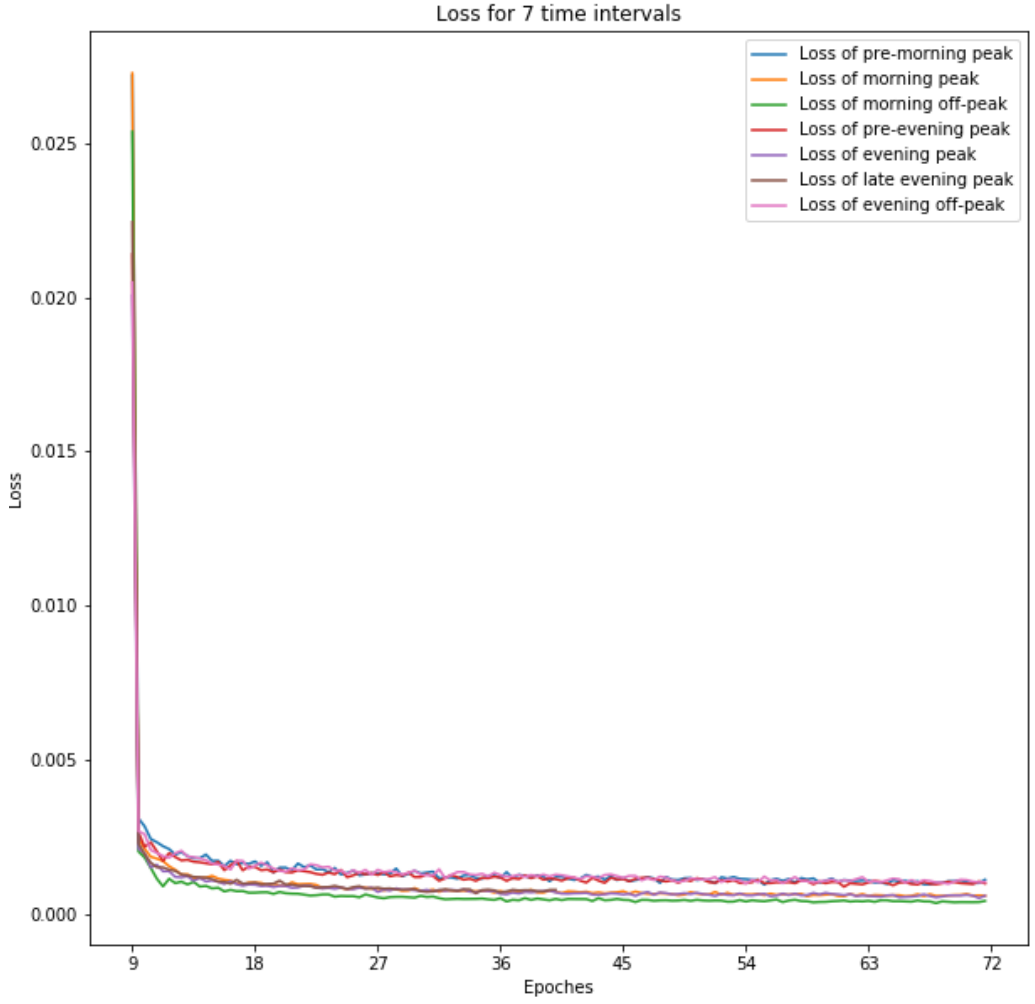


$$Loss = \frac{(output - input)^2}{sampleSize}$$

Stacked autoencoder

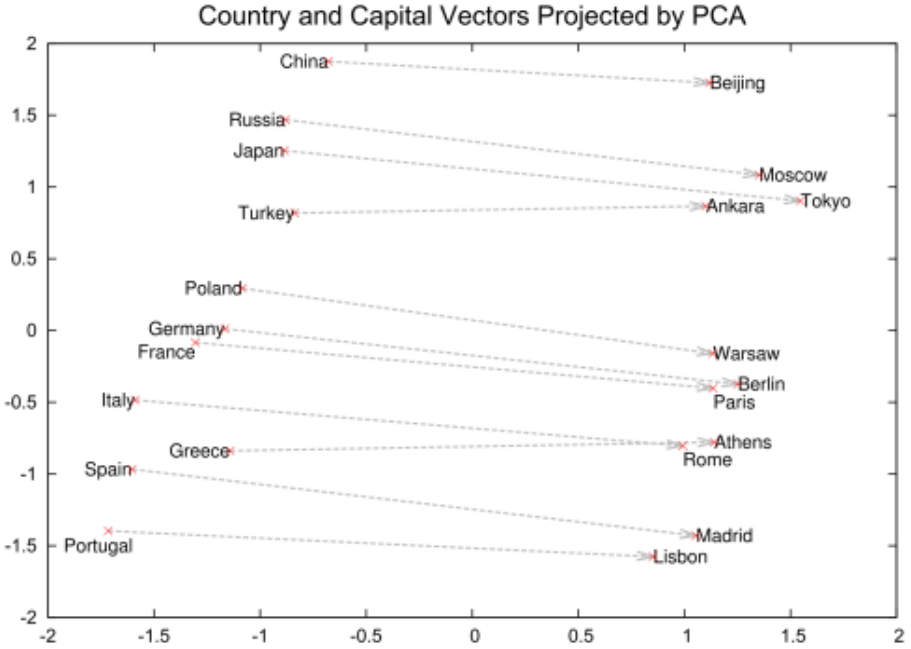
Platforms and training parameters:

- Epoch:200000, batch size: 128, optimizer: adaGradient (LR:0.01)
- 8 E5 cores, 16GB RAM, 1060 3GB, take 7hours to train one model



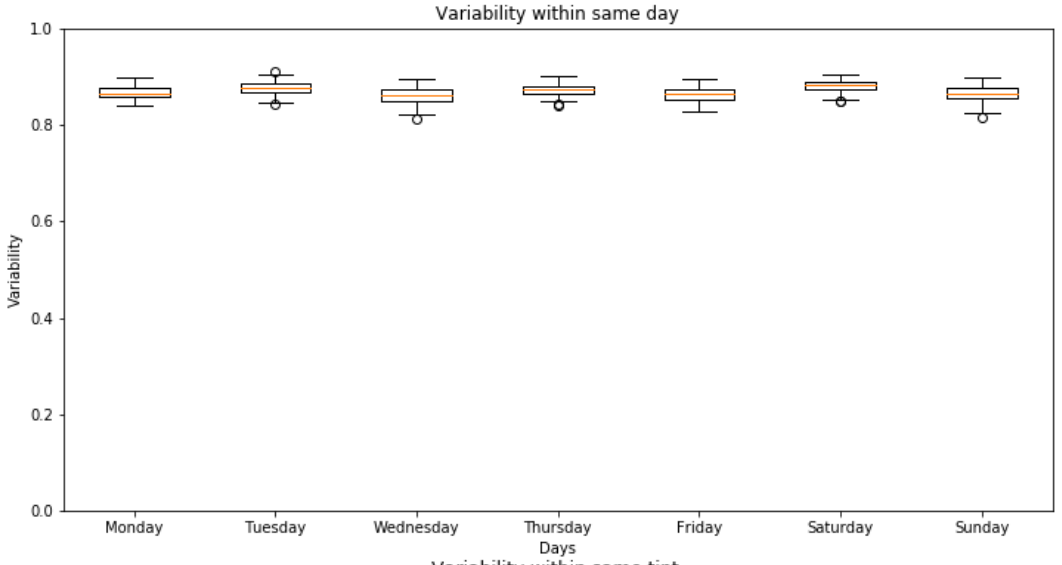
Time interval	R-squared value
pre-morning peak	0.881
morning peak	0.951
morning off-peak	0.959
pre-evening peak	0.882
evening peak	0.948
late evening peak	0.947
evening off-peak	0.865
Mean	0.919

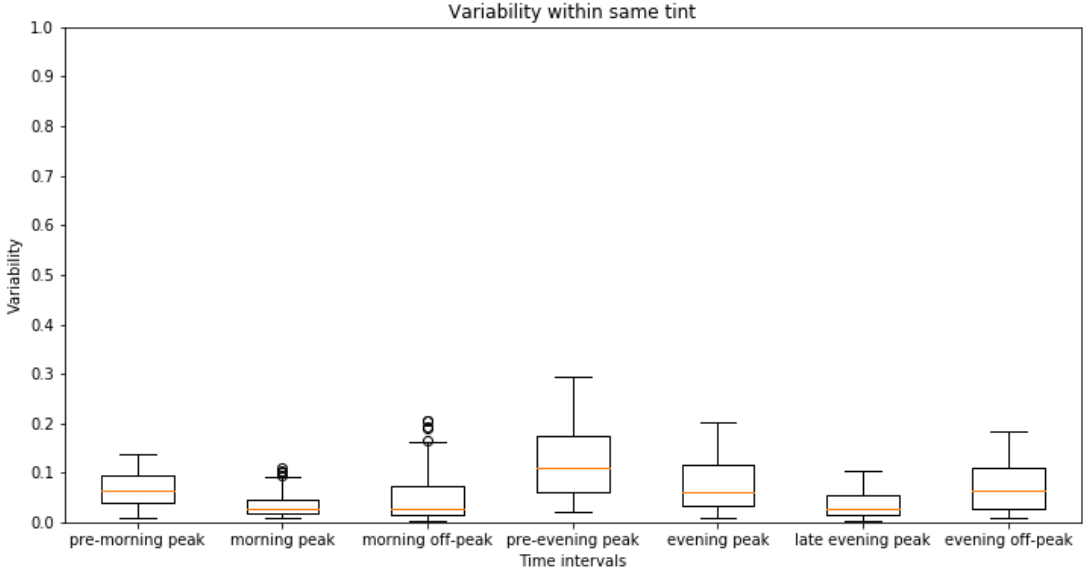
Mobility semantics



Mobility semantic vector decomposition

“Capital”&”Country”~ Mobility semantics in Different time intervals



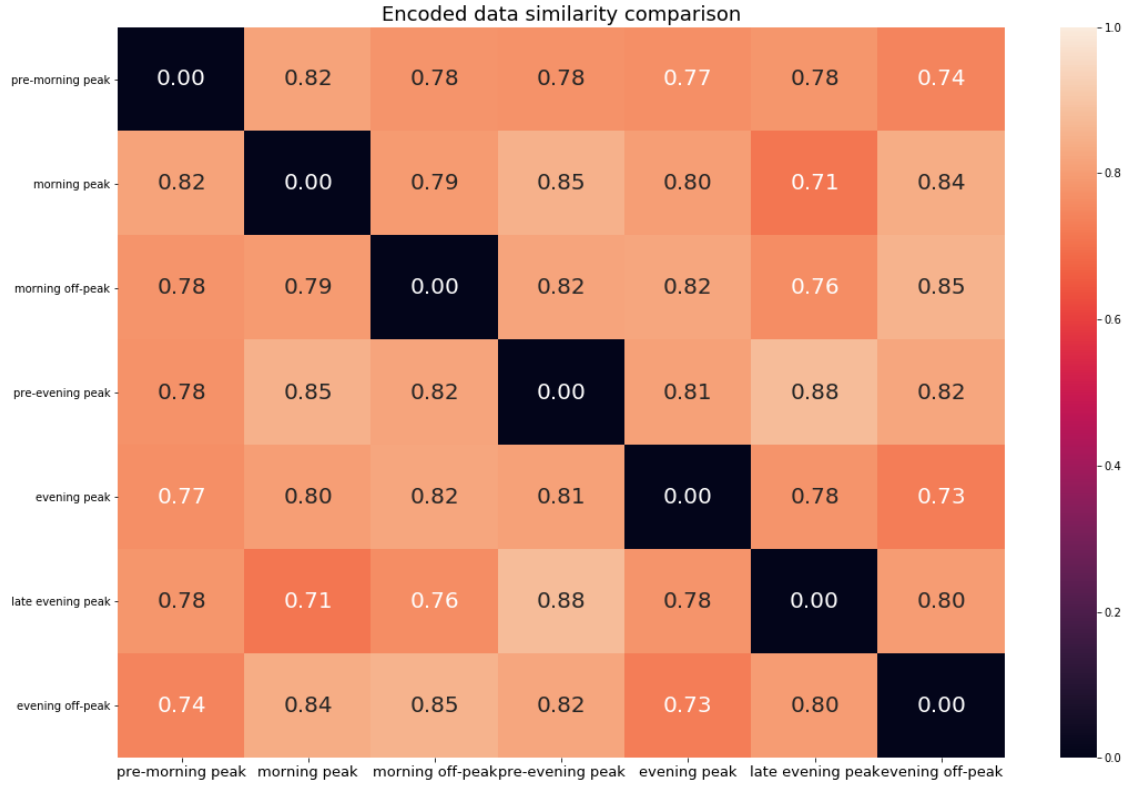


Mobility semantics

For semantic vectors Beijing-China≈Tokyo-Japan

↓ Cosine similarity

Stn1_Monday_MorningPeak-Stn1_Monday_EveningPeak
≈ Stn2_Friday_MorningPeak-Stn2_Friday_EveningPeak

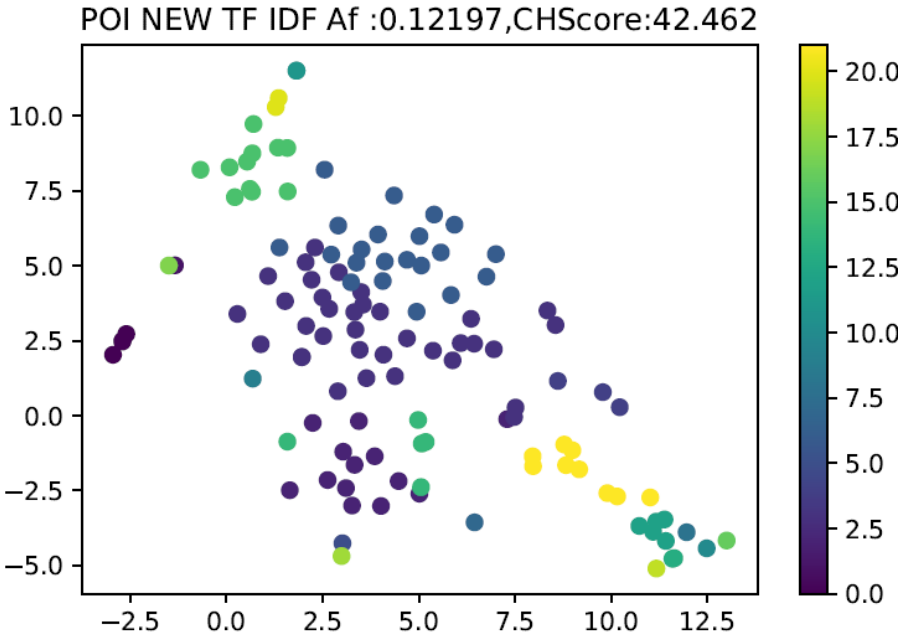


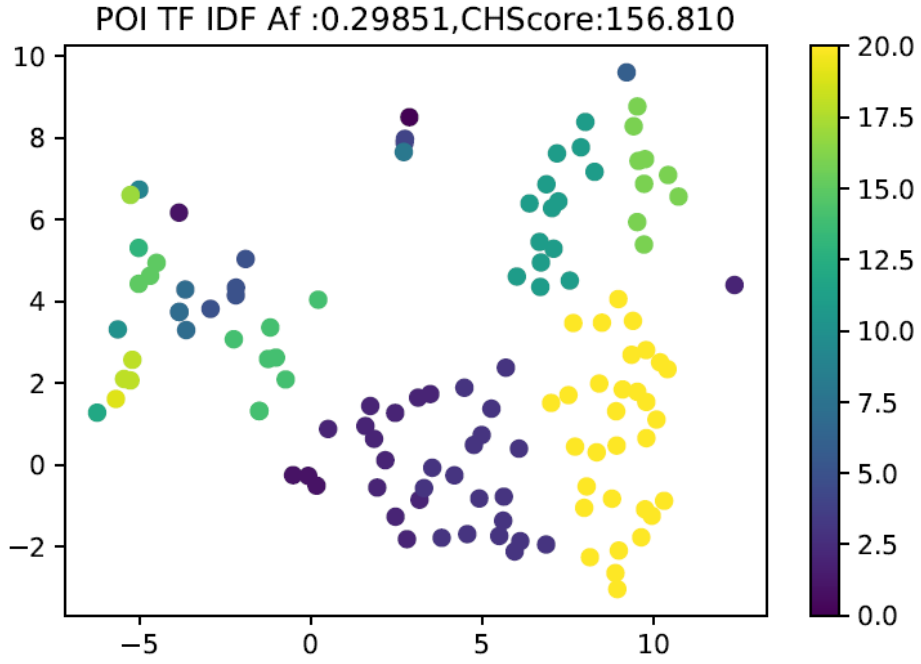
Elements similarity of each two time interval group's subtraction vector

Service semantics

Term Frequency–Inverse Document Frequency (TF-IDF)

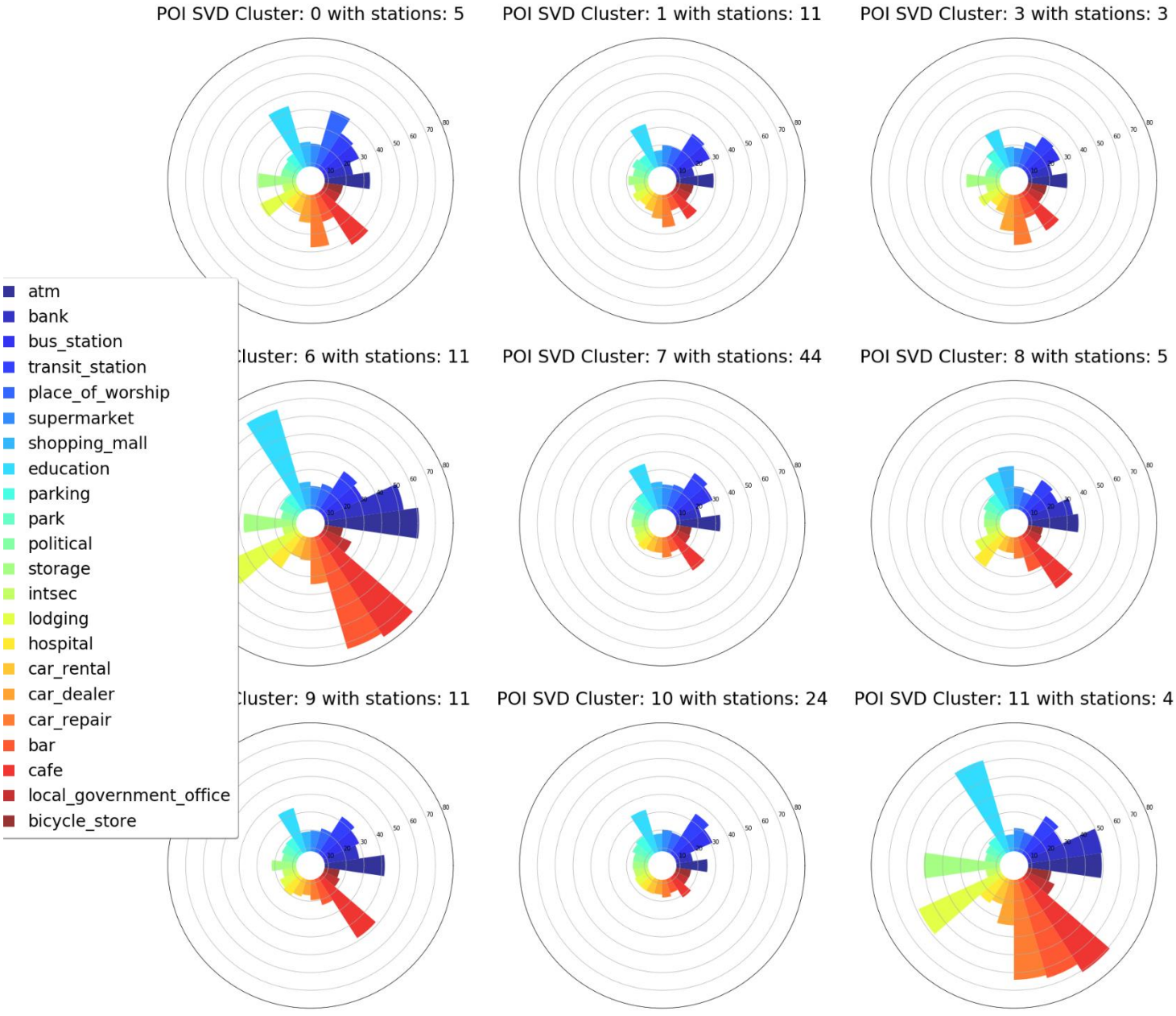
$$\mathrm{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$





SVD to find semantics
(refer to literatures)

Service semantics



Categories are hard to classify, use topic modeling to help us better find the division

Service semantics

Words

Topics

Words

Document

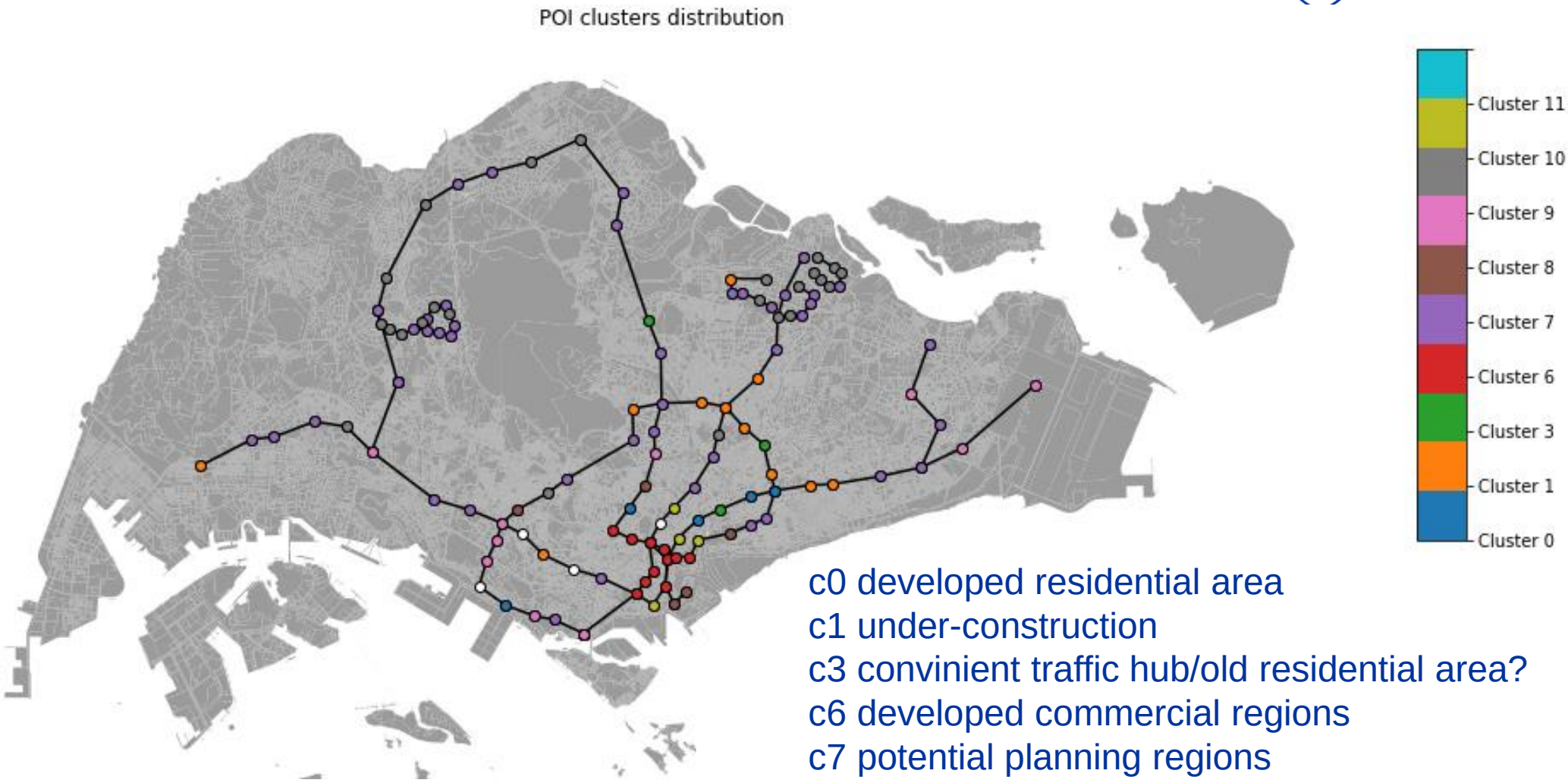
	Cluster 0	Cluster 1
atm	1.3696	1.8781
bank	1.1391	1.1029
bus_station	1.3124	2.2211
transit_station	1.3558	2.453
place_of_worship	1.3757	1.2871
supermarket	1.0753	1.3188
shopping_mall	1.0883	1.0688
education	1.5323	2.2542
parking	1.0407	1.1003
park	0.9873	1.2058
political	1.0004	1.0501
storage	1.3591	1.2316
intsec	1.0233	1.0223
lodging	1.2389	1.1629
hospital	1.0146	1.0558
car_rental	1.0232	1.1717
car_dealer	1.148	1.4002
car_repair	1.3397	1.7815
bar	1.1428	1.0909
cafe	1.3791	1.6682
local_government_office	1.0347	1.1135
bicycle_store	1.0768	1.1253

Cluster 1
Cluster 3
Cluster 5
Cluster 6
Cluster 7
Cluster 8
Cluster 9
Cluster 10
Cluster 12

s_station']
ar_repair']
bar' 'cafe']
education']
tm' 'cafe']
ion' 'cafe']
it_station']
pair' 'cafe']
it_station']

The results might change occasionally since samples are small

Service semantics

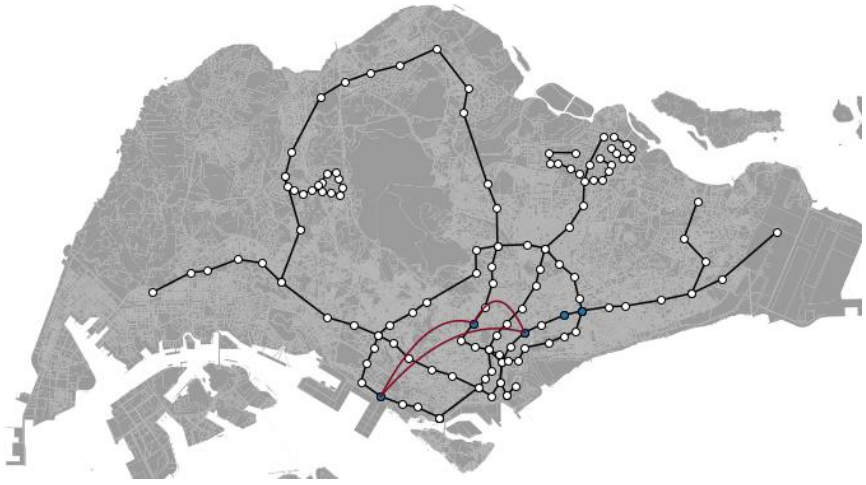


c0 developed residential area
c1 under-construction
c3 convinient traffic hub/old residential area?
c6 developed commercial regions
c7 potential planning regions
c8 entertainment
c9 scientific&educational
c10 emerging residential area
c11 emerging commercial regions

Case studies

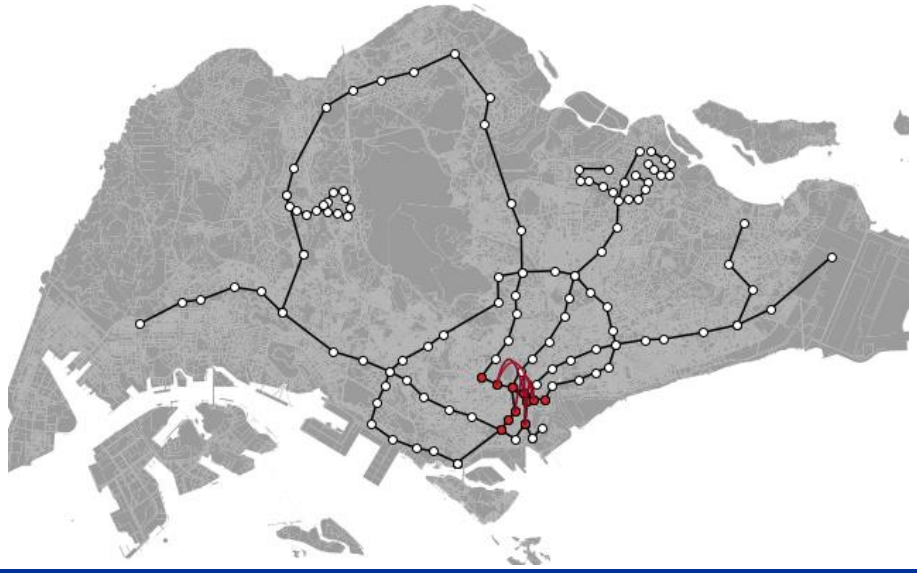
1. Different lines, same POI semantics, same flow semantics

Same flow features in different lines of Cluster 0

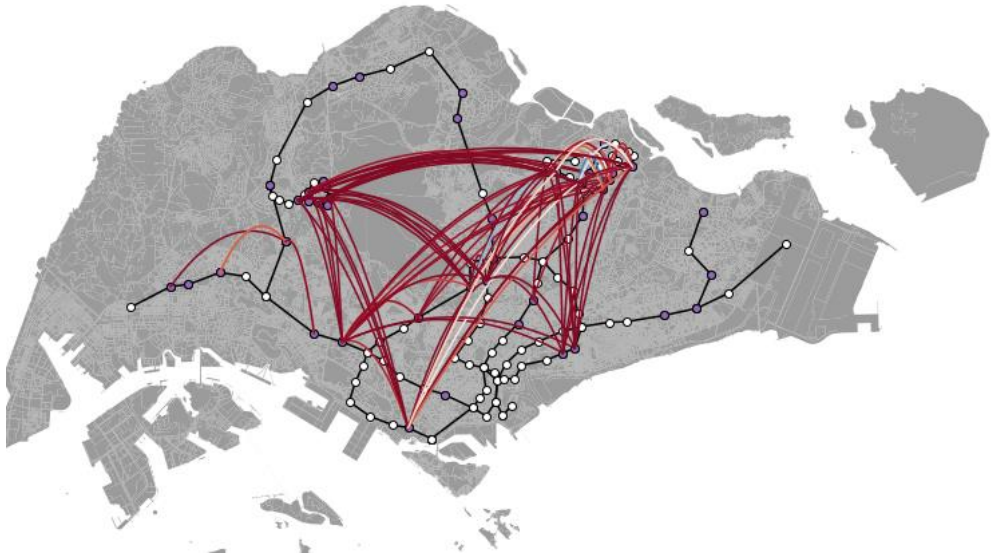


Same flow features in different lines of Cluster 1

Same flow features in different lines of Cluster 6

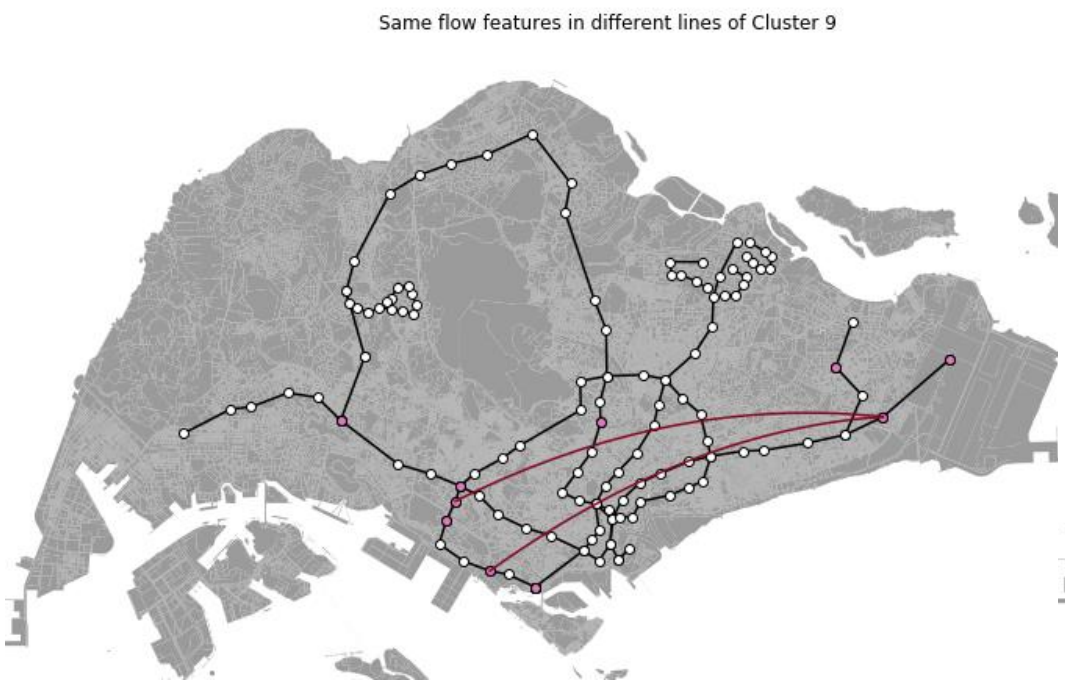


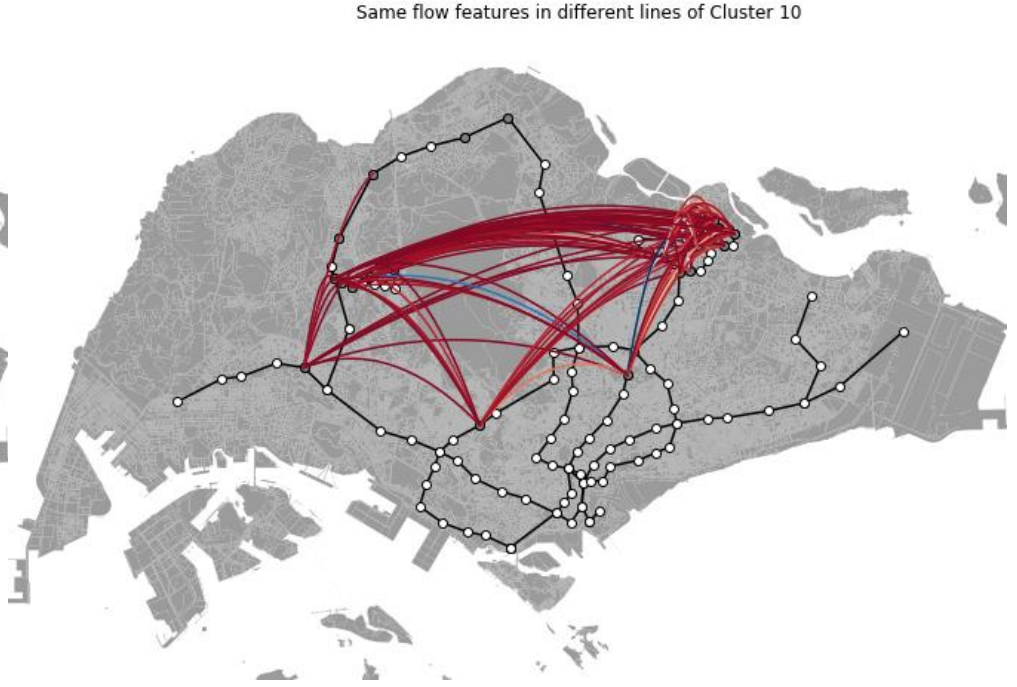
Same flow features in different lines of Cluster 7



Case studies

1. Different lines, same POI semantics, same flow semantics (dL_sP_sF).

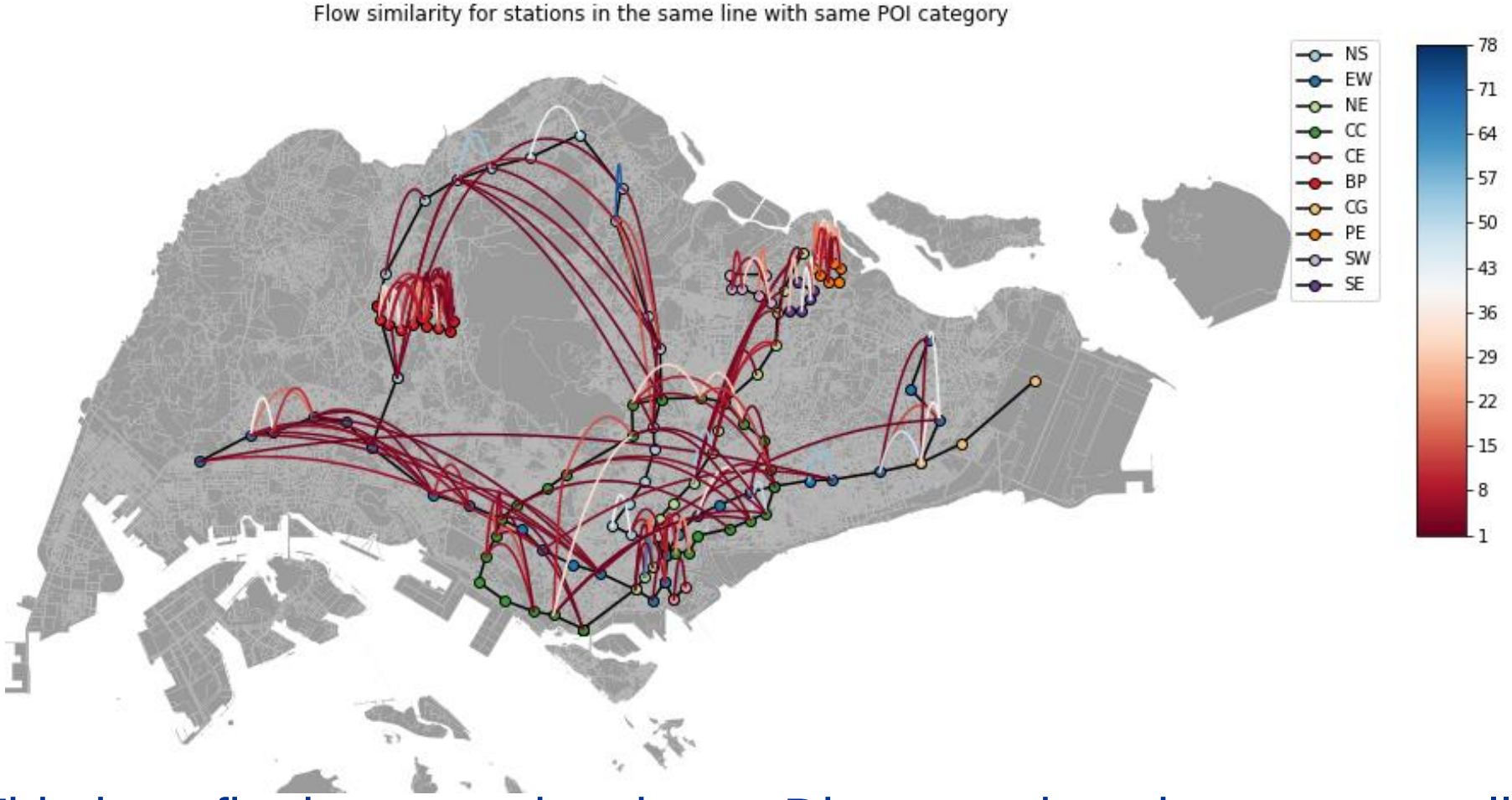




Discovered stations are usually LRT or other remote stations, because they same interaction station. Like Farmway and Woodleigh (C10 emerging residential area), might both share similar flow patterns from Sengkang.

Case studies

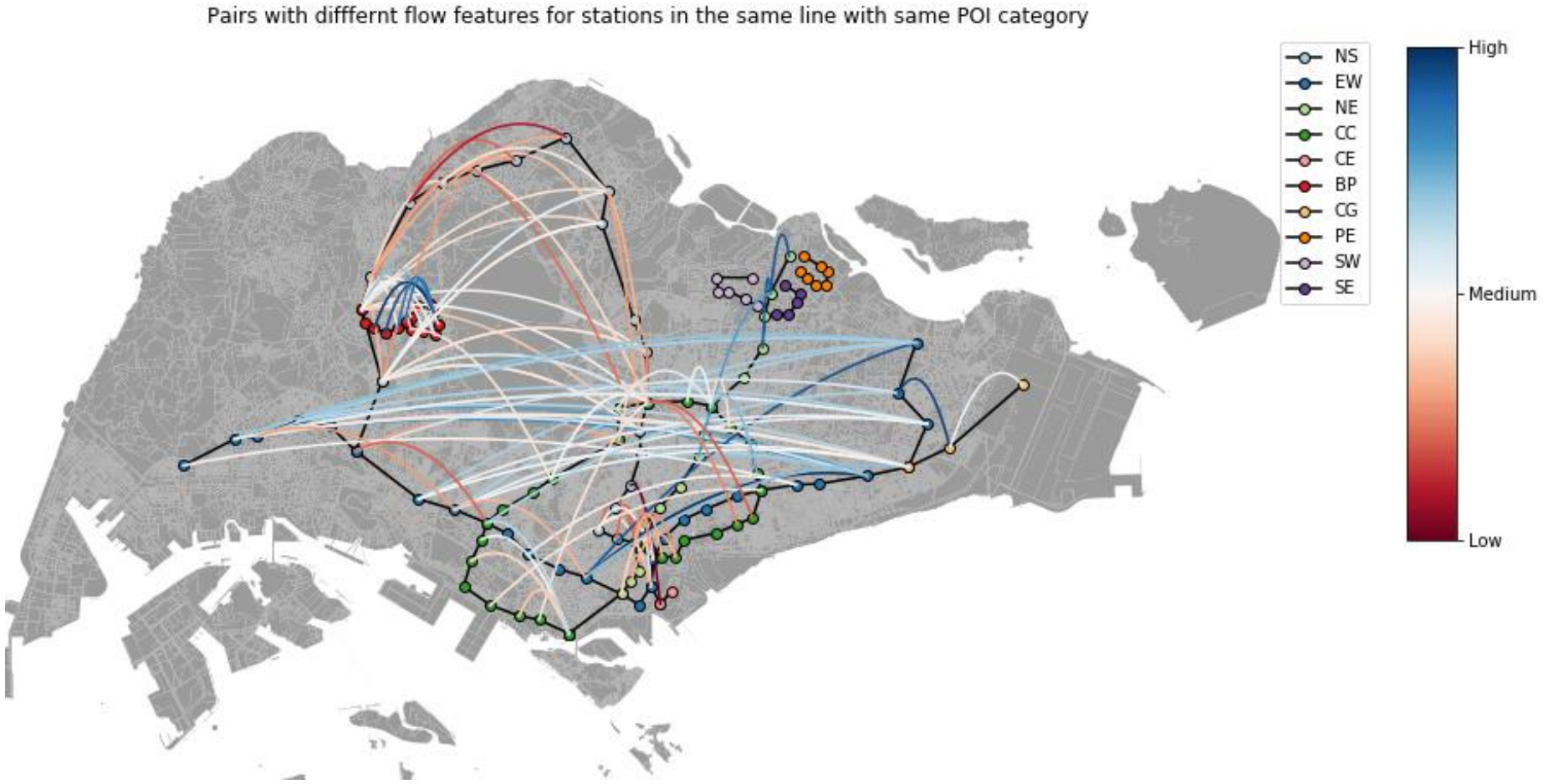
2. Same line, same POI semantics, same flow semantics(sL_sP_sF)



This benefits best to advertisers. Discovered stations are usually the adjacent stations in the same line, such as Somerset and Orchard or Pioneer and Bonn Lay.

Case studies

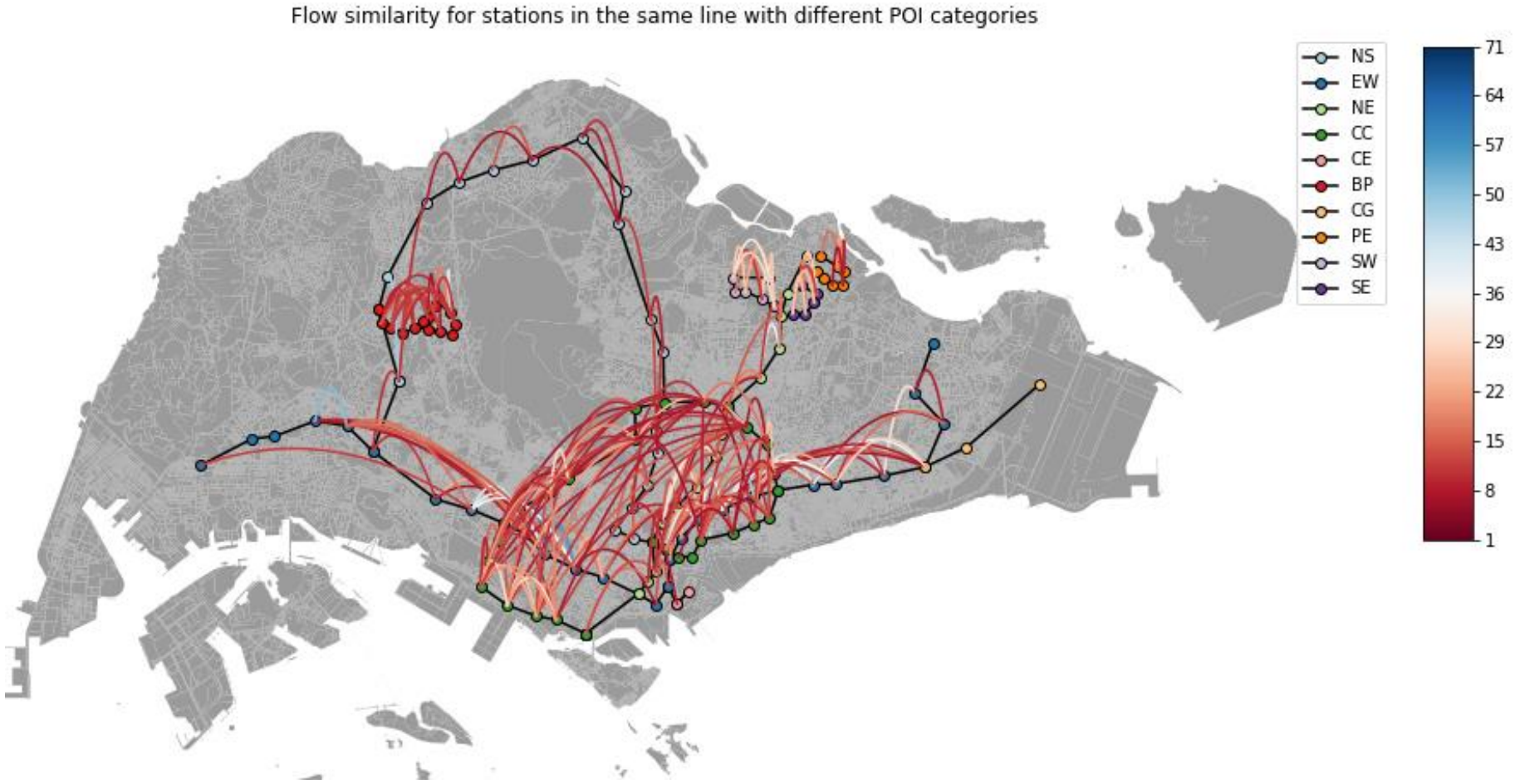
3. Same line, same POI semantics , different flow semantics (sL_sP_sF)



Remote stations in the same line, like Pasir Ris and Dover. While stations in residential region like Jurong East and Buona Vista are intersections to connect flow demand from different places.

Case studies

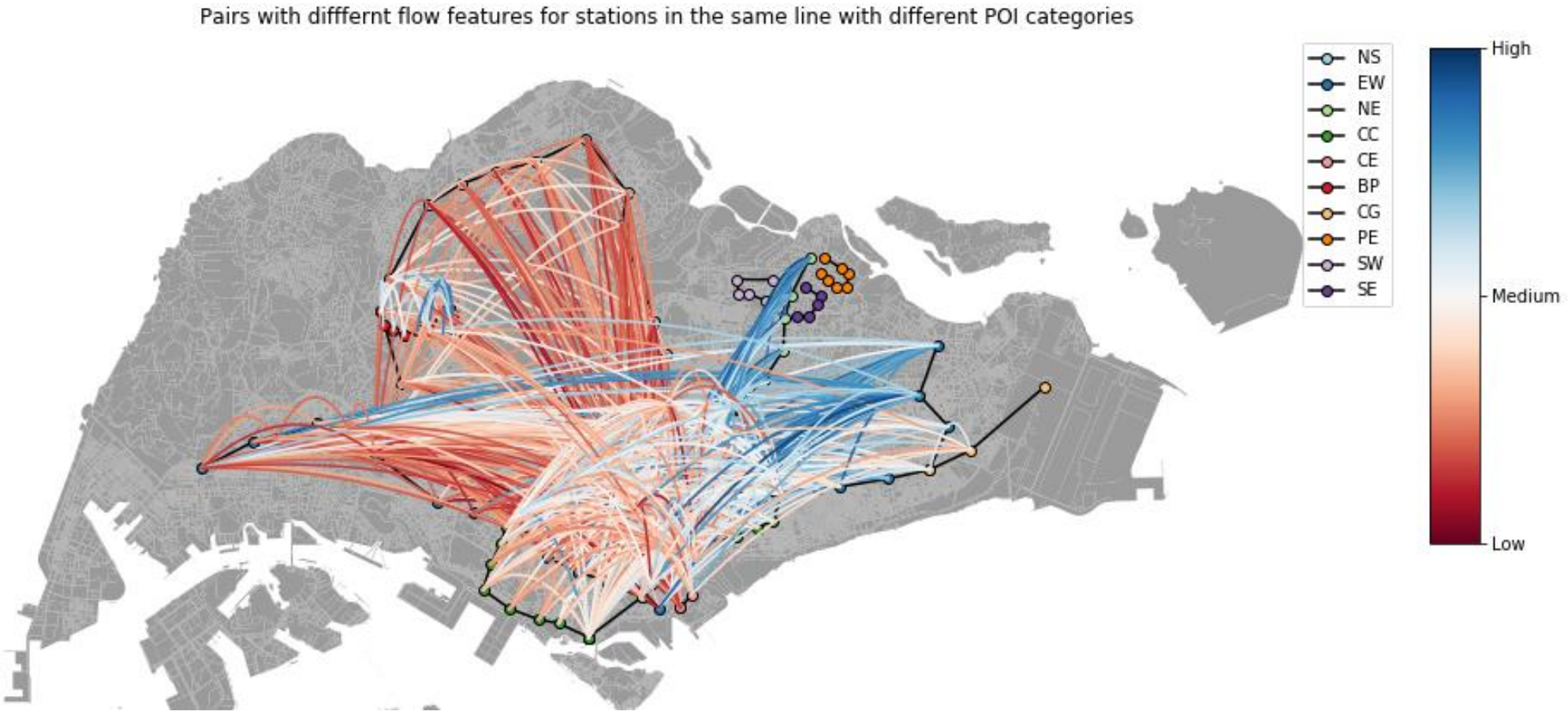
4. Same line, different POI semantics, same flow semantics (sL_dP_sF)



Circle line and LRT lines are the most typical since they serve only particular regions. POI are quite different in the opposite sides but customer flow remains similar.

Case studies

5. Same line, different POI semantics, different flow semantics (sL_dP_dF)



This result satisfies our knowledge, since distant stations in the same line serve different needs and located in various circumstances

Commercial interests

Urban planning

Further work

4. DISCUSSION AND ANALYSIS

Commercial interests

Advertisement. Advertisers can focuses on stations with same POI& flow feature and avoid targeting stations with different POI& flow feature. In general, advertising among adjacent stations in the same line.

Site selection. For small and medium-size enterprises targeting at regular or similar customers, like cheap clothing stores, snack bars or barber shops can refer to stations with same flow features to develop core customers.

Urban planning

Infrastructure. Lanes, bus stops, etc. can be constructed according to same flow features or same POI, like Tampines and Jurong East (highest overlapping in sL_sP_sF).

Traffic monitoring. Crowd with similar boarding or alighting patterns can provide insight to understand customers mobility for emergent evacuation, especially for circle line.

Land use. Flow and POI relationship, no matter similar or not, could provide comprehension of urban land use. Low utilized stations, like Ten Mile Junction, Farmway and Woodleigh can be abolished for better land use.

Further work

POI category division. Our service semantics only gives a roughly divided POI categories, but sophisticated division might be further analyzed.

Bus stops consideration. We only focused on MRT stations, which, however, is only part of the public transportation system.

Highlights

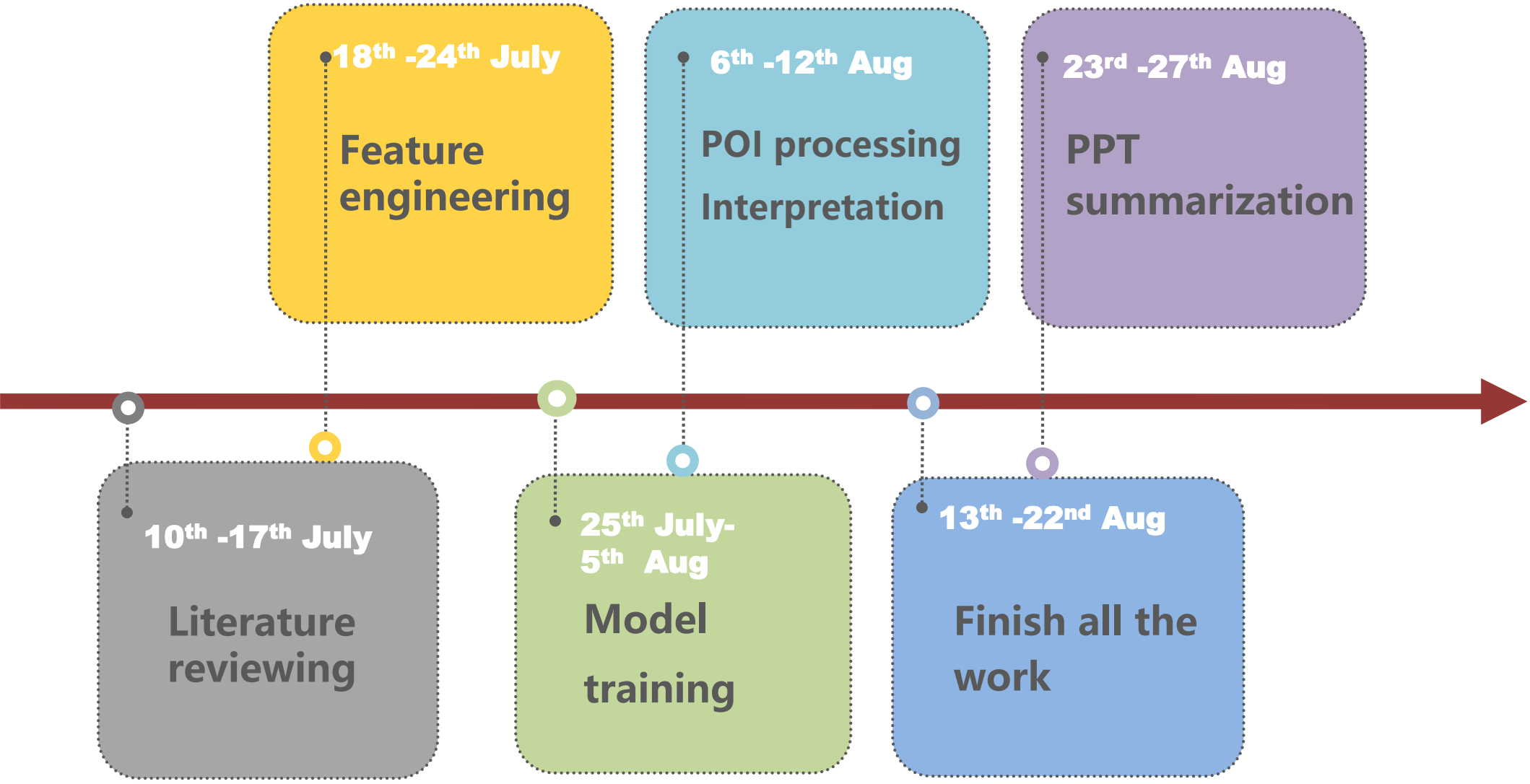
Timeline

5. CONCLUSION

Highlights

- Transplant semantic models on urban mobility discovery
- Proposed a new comprehension of semantic model
- Discovering specific relationship between MRT stations
- Give solid urban planning analysis and suggestions

Timeline


18th -24th July
Feature engineering
6th -12th Aug
POI processing
Interpretation
23rd -27th Aug
PPT summarization
10th -17th July
Literature reviewing
25th July-5th Aug
Model training
13th -22nd Aug
Finish all the work

NUS
National University
of Singapore

THANK YOU