Learning and Prediction over Massive Spatio-temporal Data

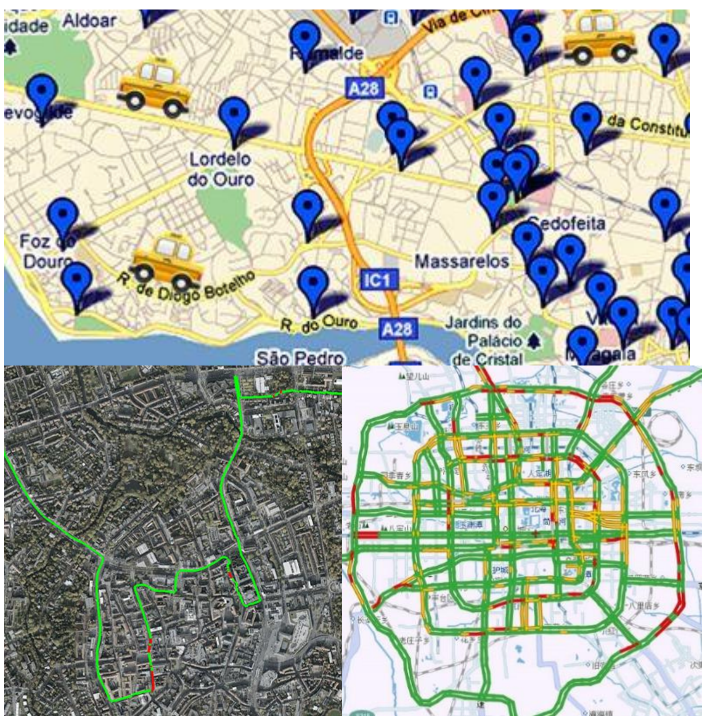
Jian Li
Institute for Interdisciplinary Information Sciences
Tsinghua University, China

清華大學
Tsinghua University

CONTENTS

Spatial Temporal data

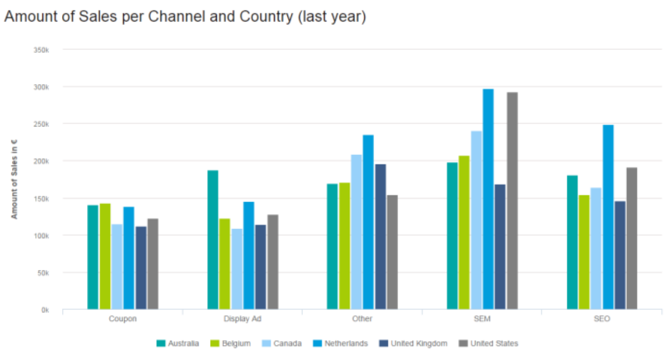
- GPS data, trajectory
- Online car-hiring data



Spatial Temporal data

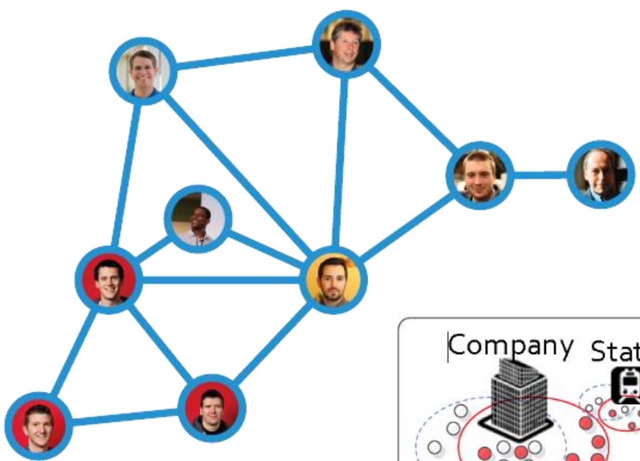
- Warehouse management

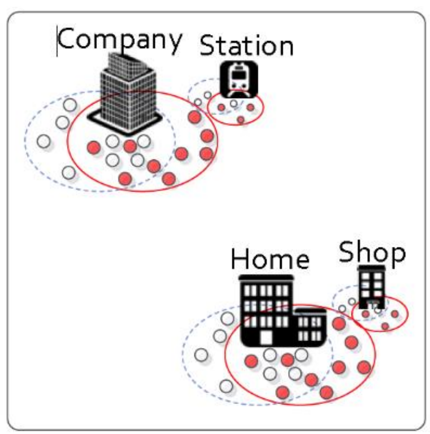
- Online retailers



Spatial Temporal data

Social relationship detection



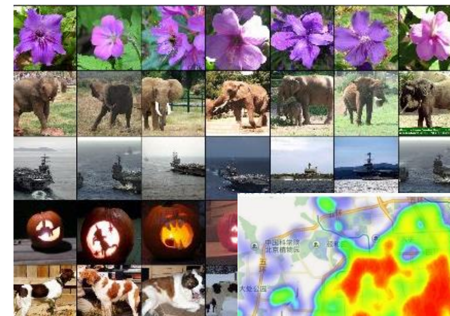


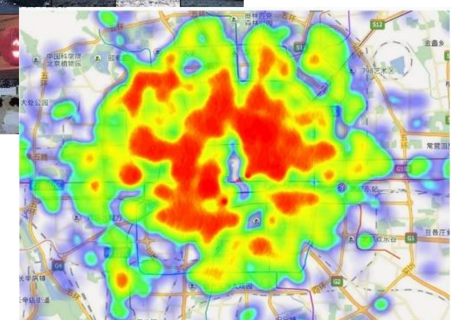
Financial data

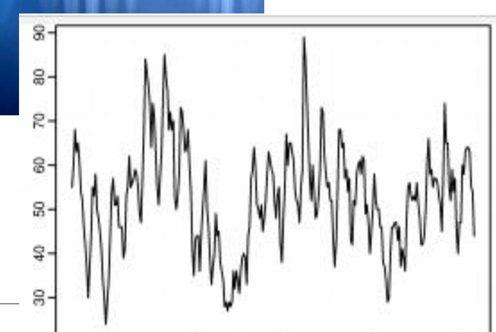
- Stock price prediction

Characteristic

- Spatial dependence
  - different locations interact on each other
  - compare with images:
    - city level scale, sensitive to the granularity
- Temporal dependence
  - past states affect the future
  - compare with texts/speech:
    - seasonality in multi-granularity
    - highly affected by sudden event (raining, traffic accident)

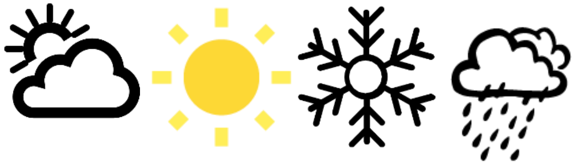






Characteristic

- Diverse data sources
  - GPS locations, orders, weather, POIs, etc.
- Massive, large volume, highly noisy

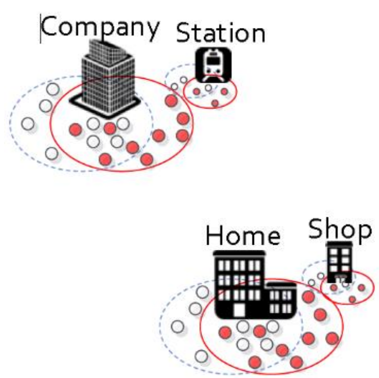


Spark

hadoop

Some Recent work of My Group

- **Supply-demand Prediction**
  - Online Car-hiring Services
- **When will you arrive?**
  - Estimating Travel Time Based on Recurrent Neural Networks
- **Where to build your store?**
  - Store location selection
- **User Identification**
  - Automatic User Identification across Heterogeneous Data Sources
- **Visitation Prediction**
  - Which POI you will visit next?
- **Traffic condition Prediction**
  - Traffic Condition Prediction System



清華大學
Tsinghua University

CONTENTS

Tsinghua University

Supply-Demand Prediction for Online Car-hiring Services using Deep Neural Network

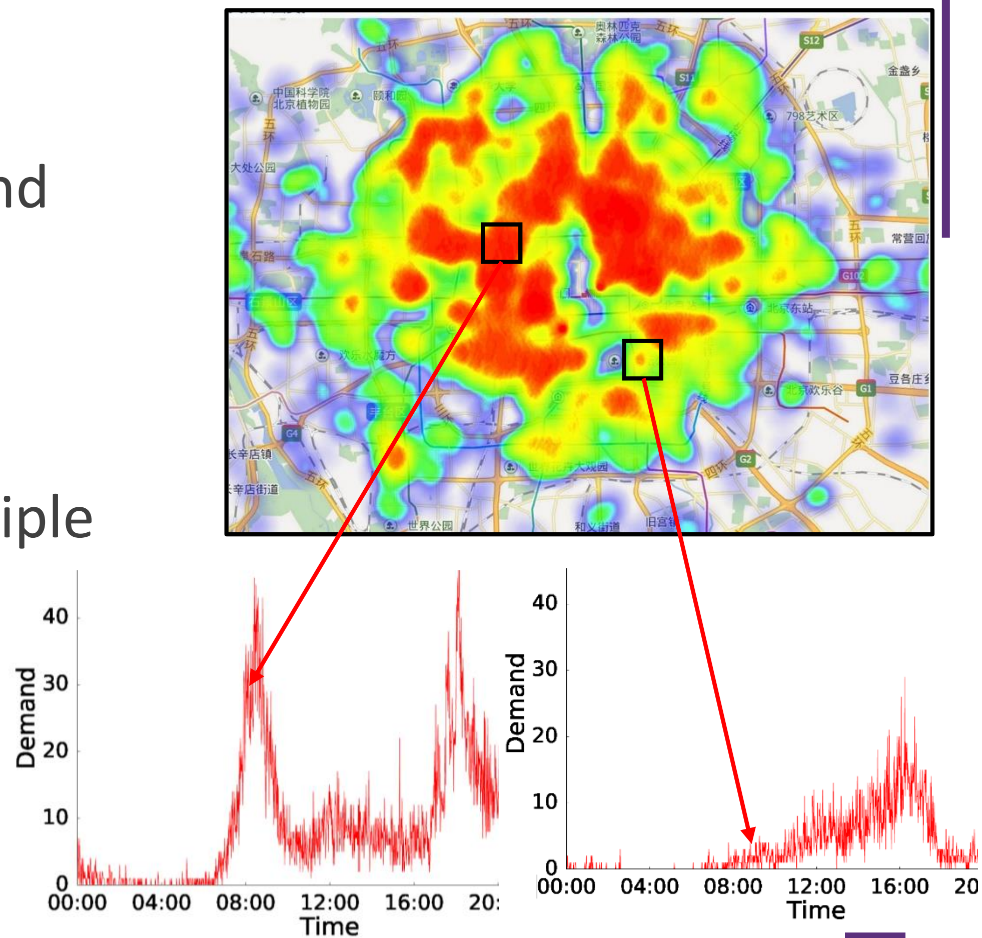
- **Objective**
  - Predict the gap between the car-hailing supply and demand in a certain area in the next few minutes.
- **Motivation**
  - Balance the supply-demand by scheduling the drivers in advance
  - Adjust the price dynamically
  - Recommend popular pick-up locations for drivers

[ICDE’17]

Tsinghua University

Challenges

- The car-hailing supply-demand varies dynamically
  - geographic locations
  - time intervals.
- The order data contains multiple attributes
- “hand-crafted” features are difficult to design



Definitions

Car-hailing order valid (invalid)

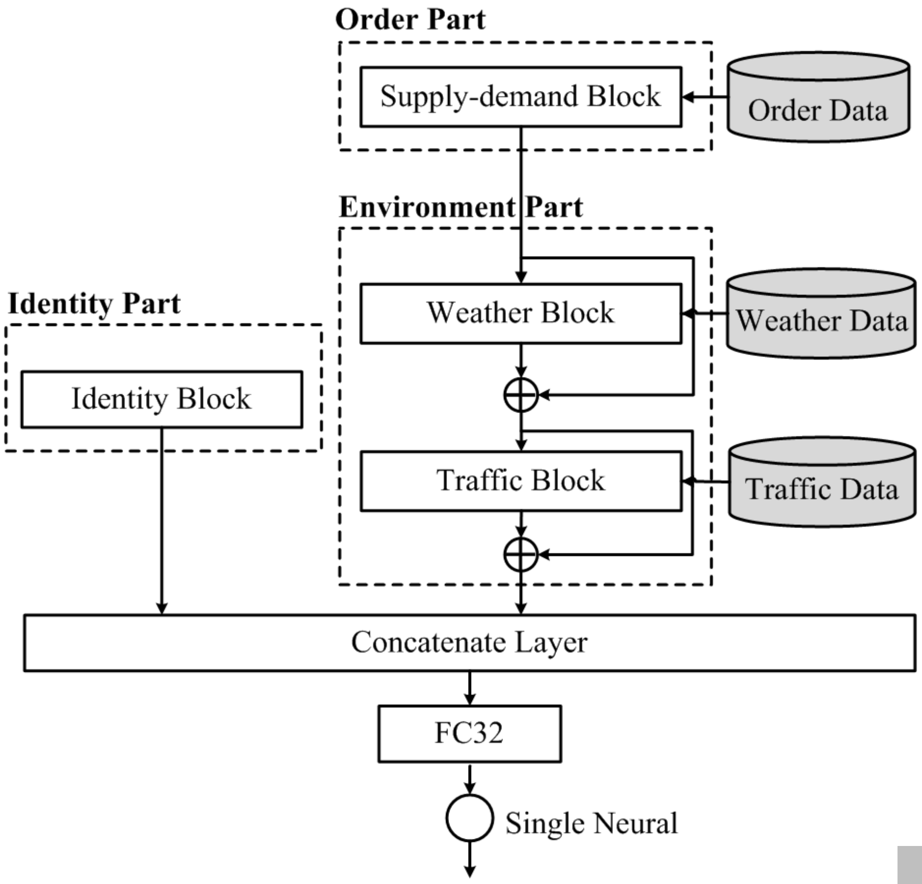
1. Date 2. Timeslot 3. Passenger ID
4. Star area ID 5. Destination area ID

Objective

Predict the supply-demand gap (eg. the number of invalid orders) of a certain area a, in 10 minutes from now.

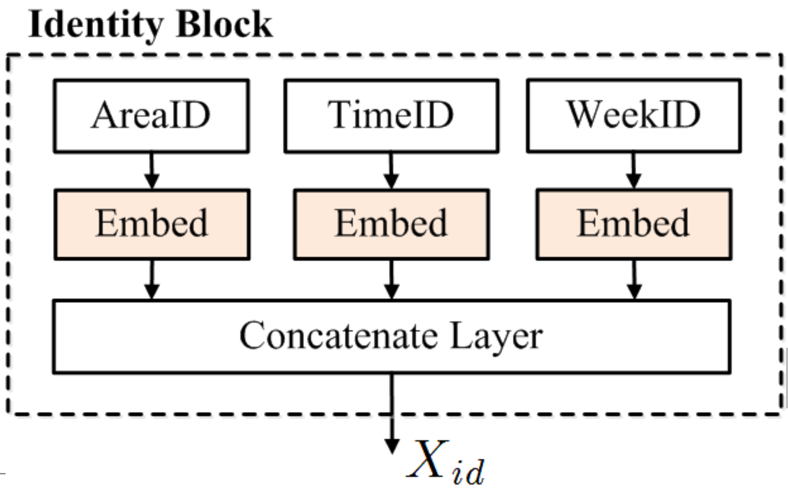
Framework

1. End-to-end model
2. Using embedding to "cluster" similar areas and timeslots
3. Learning the useful feature vector from the order data
4. Involve in the weather and traffic data through residual network

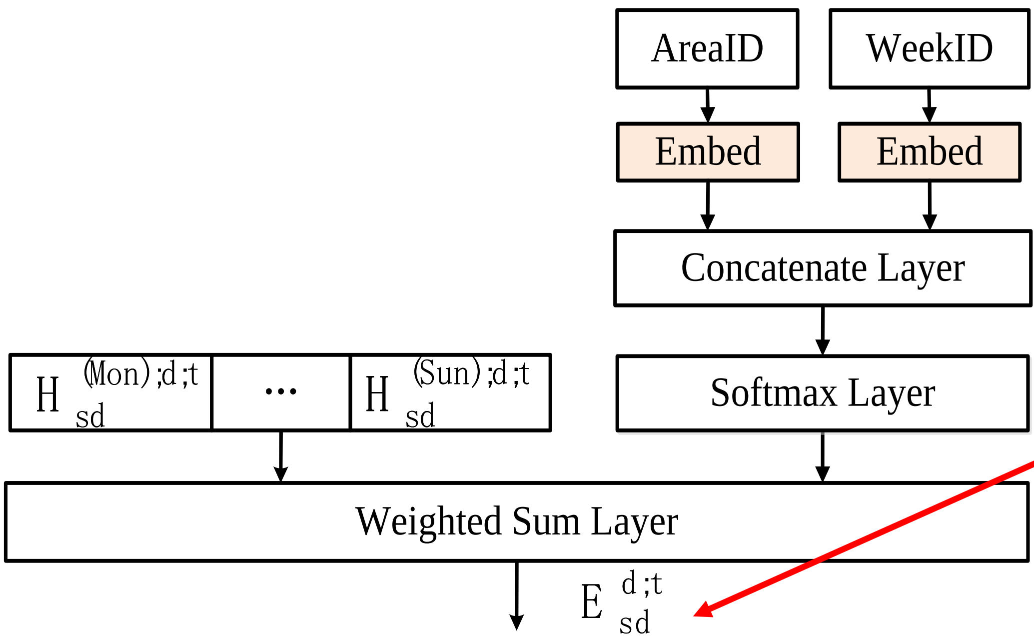


Identity Block

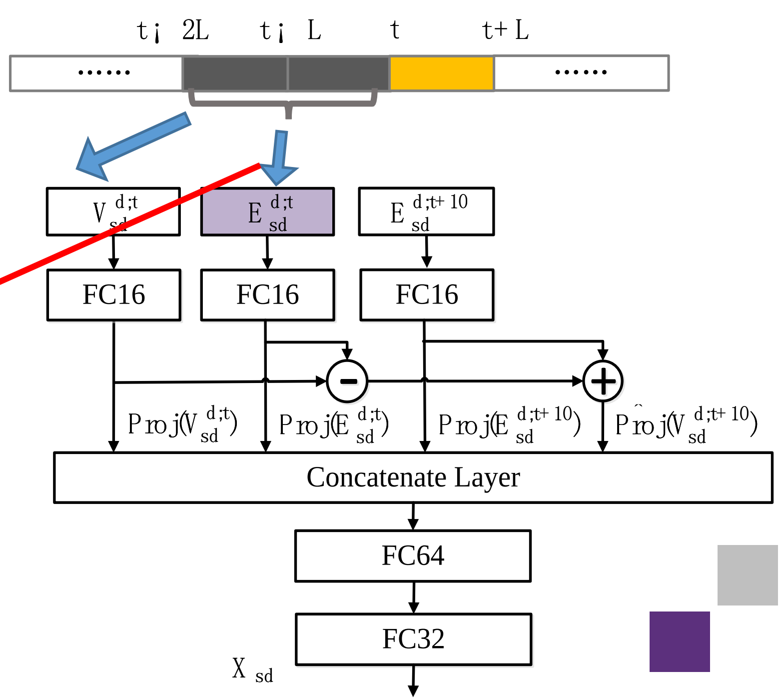
- Different areas at different time can share similar supply-demand patterns.
- Prior work clusters the similar data :
    - separate sub-task
    - manually design the distance measure



Order Part



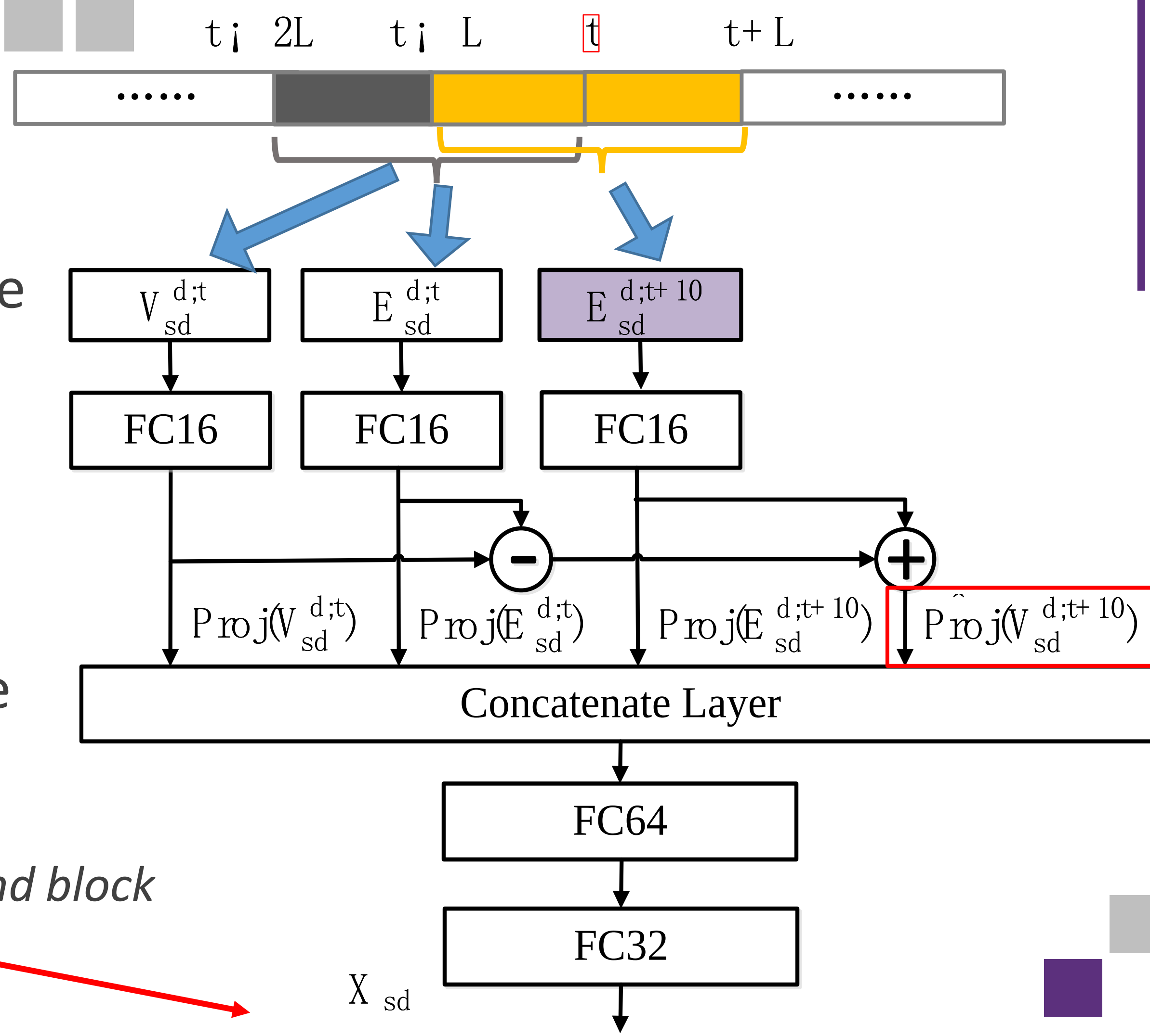
AreaID
WeekID
Embed
Embed
Concatenate Layer
Softmax Layer
···
Weighted Sum Layer



t+L
FC16
FC16
FC16
Concatenate Layer
FC64
FC32

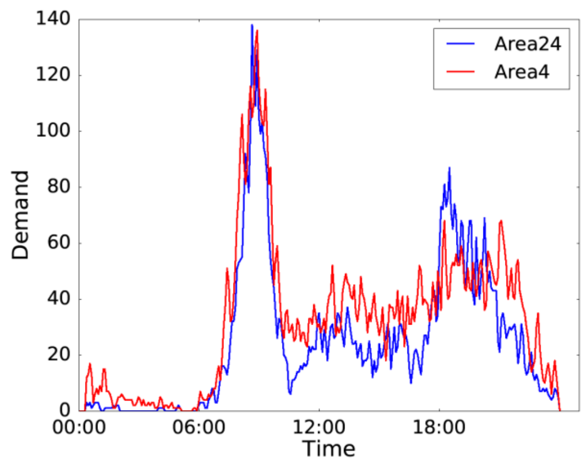
Order Part

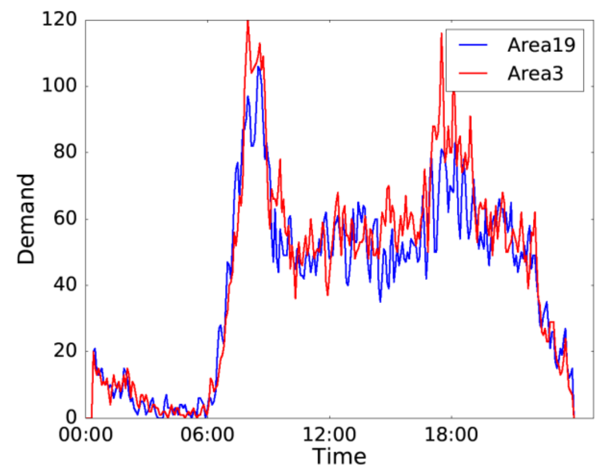
1. Project vectors into the same kernel space
2. Directly manipulate the vectors in the space
3. Stable training procedure & accurate result

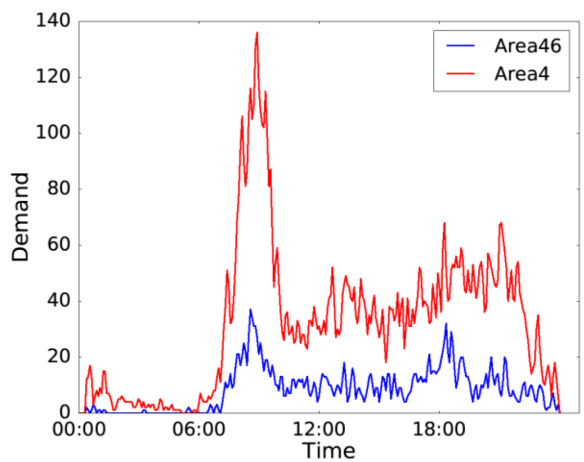


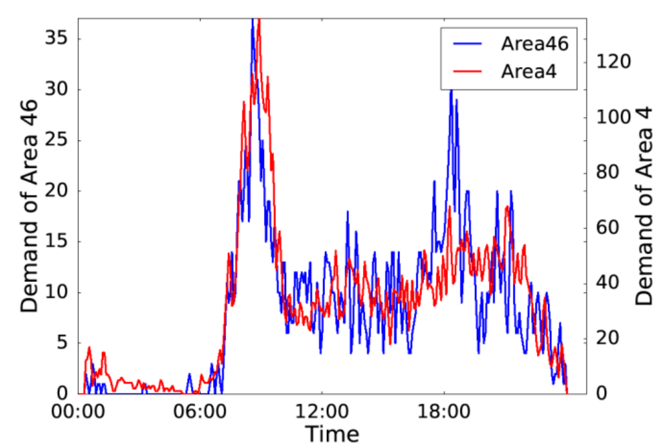
Output of the supply-demand block

Experiment – Effects of Embedding

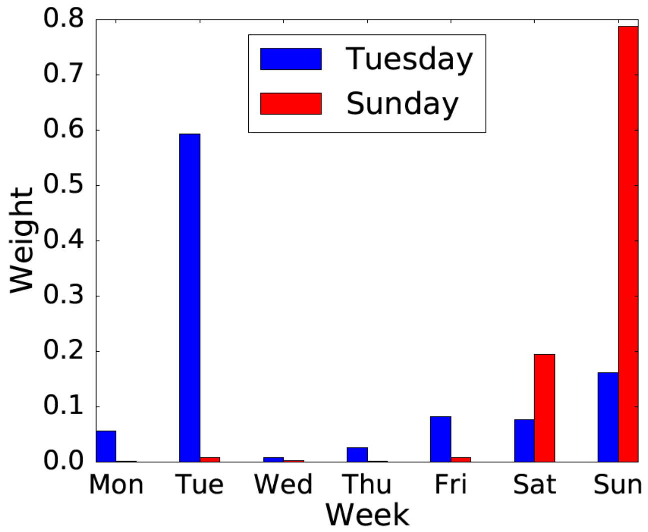




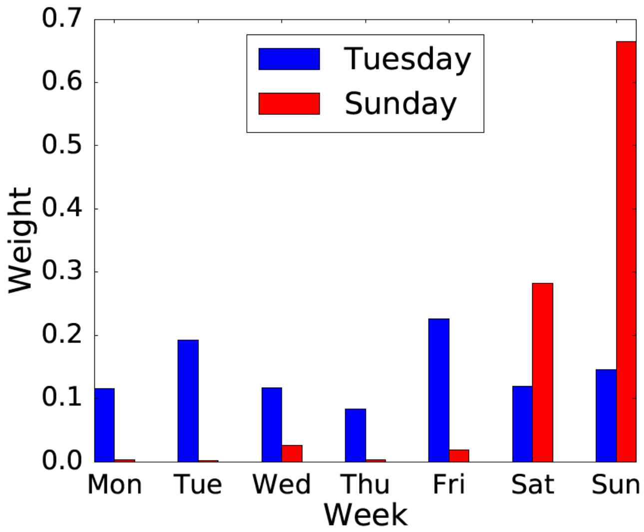




Experiment – Effects of Embedding



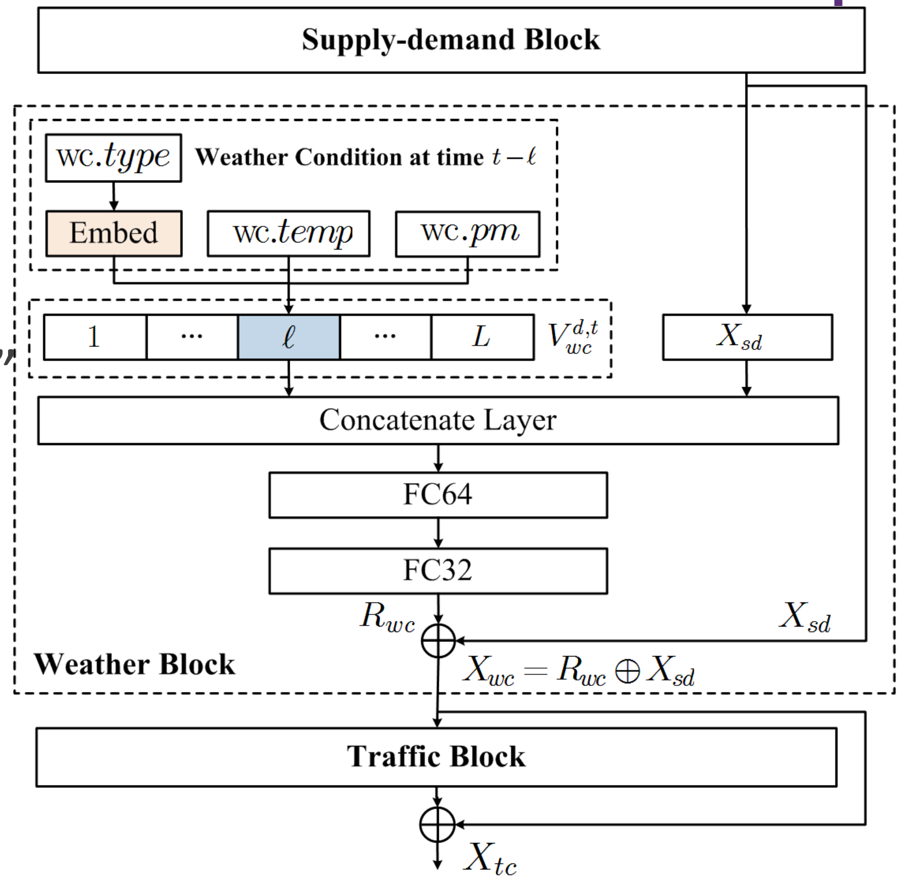
Area 1



Area 26

Environment Part

- Weather Block
- Residual Connection
  - Take the output as the “residual”
  - Makes the model more flexible

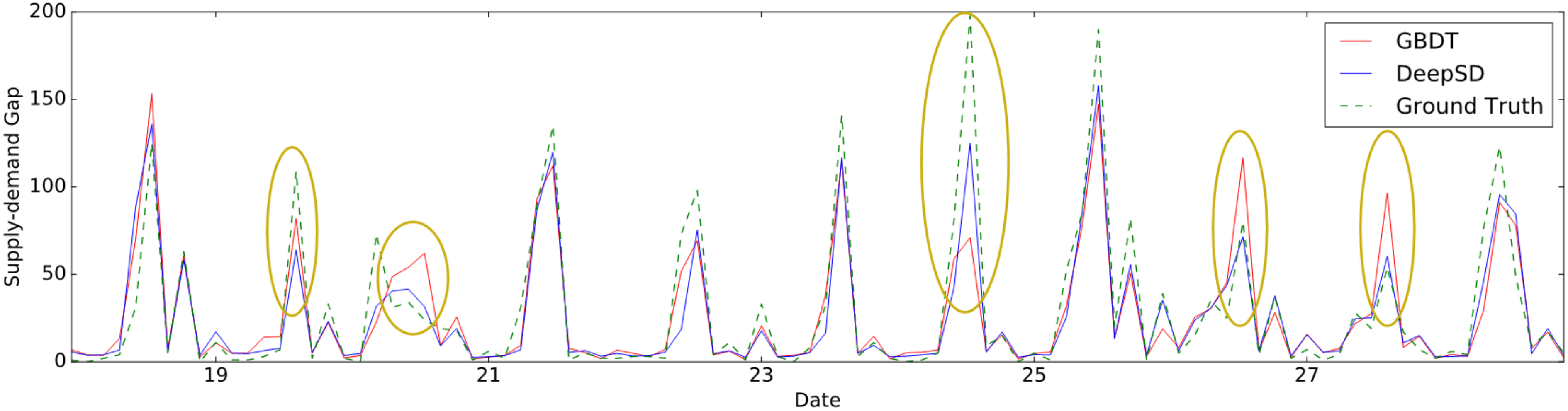


Experiment

Table: Performance Comparison

Model	Error Metrics	
	MAE	RMSE
Average	14.58	52.94
LASSO	3.82	16.29
GBDT	3.72	15.88
RF	3.92	17.18
Basic DeepSD	3.56	15.57
Advanced DeepSD	**3.30**	**13.99**

Experiment


200
150
100
50
0
Supply-demand Gap
19
21
23
25
27
Date
GBDT
DeepSD
Ground Truth

清華大學
Tsinghua University

CONTENTS

Estimating Travel Time Based on Recurrent Neural Networks

When will you arrive?

Motivation

- Routes planning, Navigation
- Traffic dispatching

Previous work

- Estimate for each individual road
- Road intersections and traffic lights
- No driving habits

清華大學
Tsinghua University

Definitions

Objective

Given:

1. path, 2. driver, 3. start time

Estimate:

the travel time for the given path.

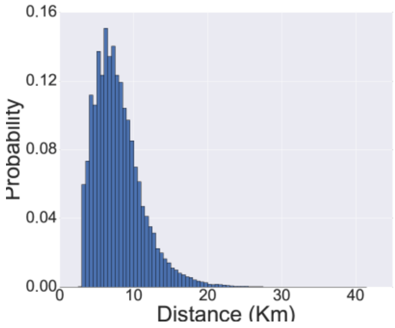
⁎ *We assume that the travel path S is specified by the user or generated by the route planning apps.*

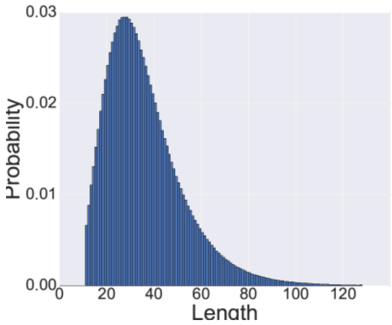
清華大學
Tsinghua University

Challenges

- The travel time of a specific path can be very different
  - ✓ Peak/Non-peak hour
  - ✓ The day of the week

- Diverse values of trajectory length/distance.



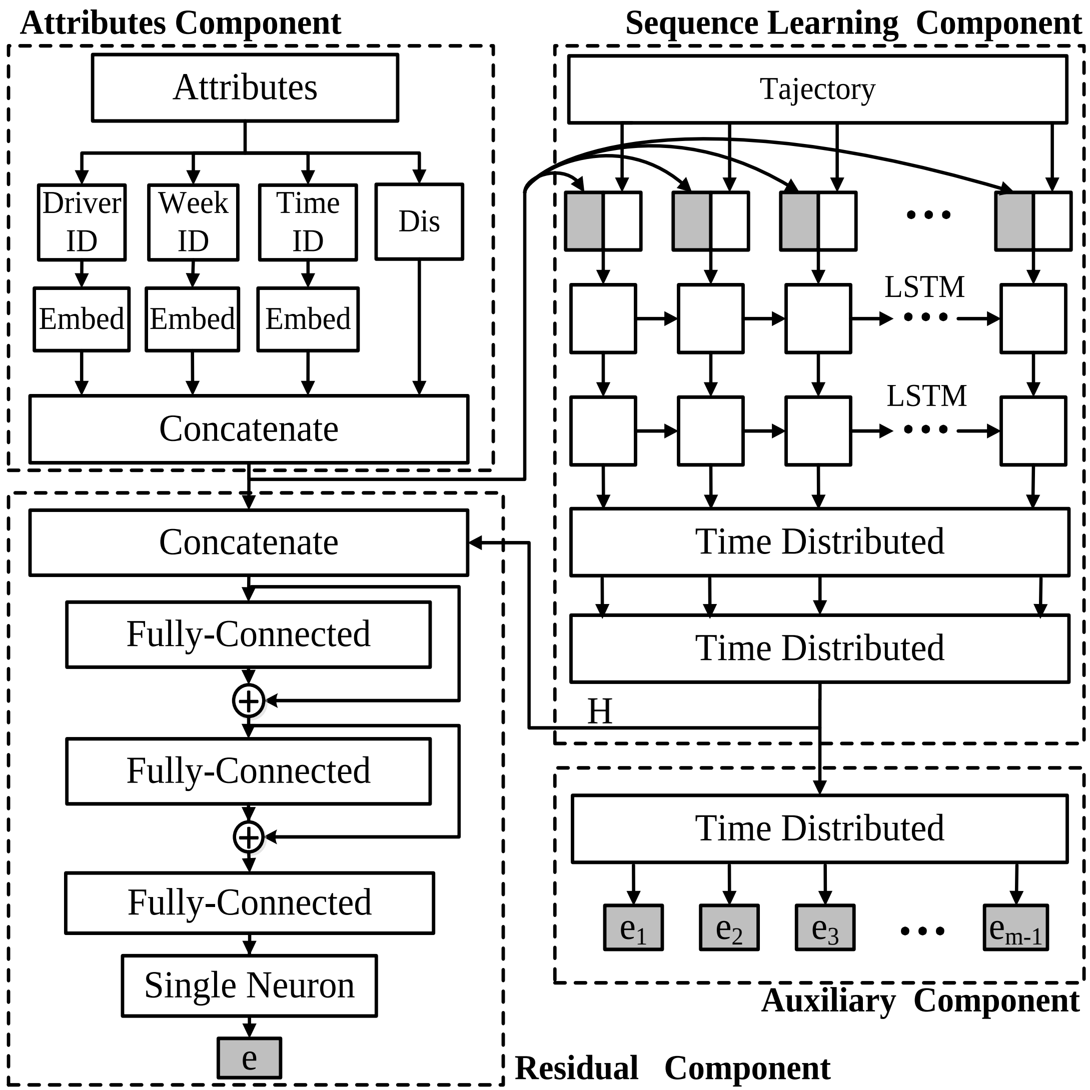


- Different driving habits

Tsinghua University

Architecture

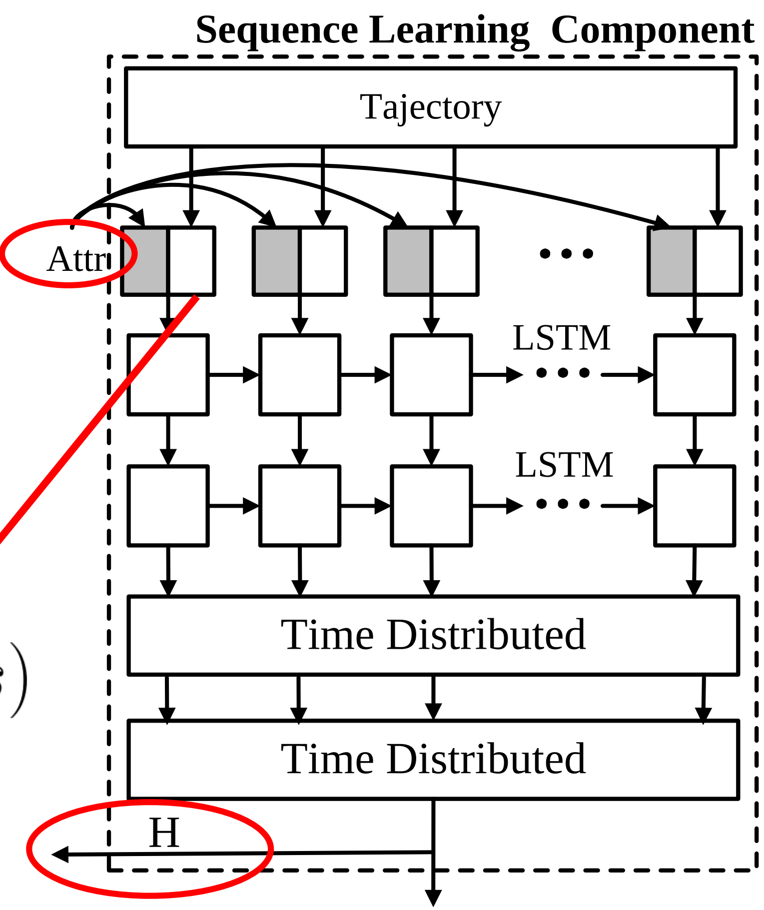
1. Incorporate various factors
2. Using LSTM to capture the temporal relationships
3. Using ResNet to predict the travel time
4. Extend to multi-task learning by introducing an auxiliary component



Sequence Learning Component

- Using LSTM to capture temporal dependency
- Handling different trajectory length
  - Mean Pooling Trick
  - Sampling Trick

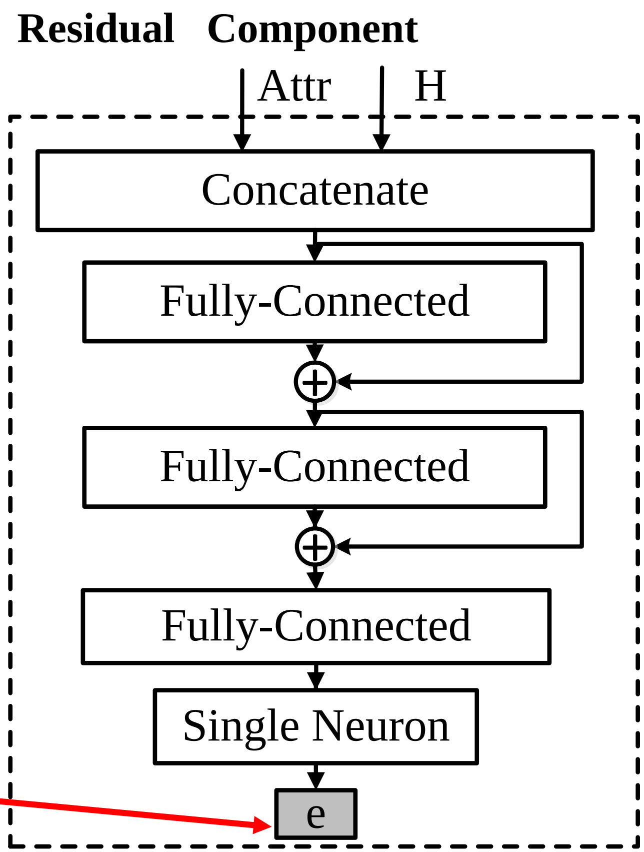
$$(lng_i, lat_i, lng_{i+1}, lat_{i+1}, dis)$$



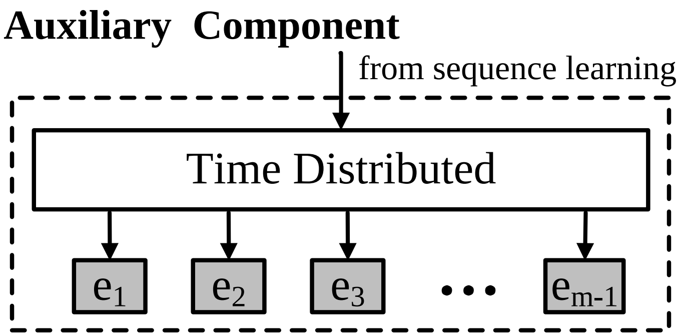
Residual Component

- Concatenate the output of the **sequence learning component** and the **attribute component**.
- Connect three fully-connected layers by residual connnection

The estimated travel time of the given path



Auxiliary Component



To utilize the "local information"

- extend to a multi-task model
- estimate the travel time of GPS point pairs
- used as the auxiliary output

Model Training

- Evaluate: mean absolute percentage error (MAPE)
  - Residual Component

$$\text{loss}_{seq} = \left| e - \Delta t_{p_1 \to p_{L_m}} \right| / \Delta t_{p_1 \to p_{L_m}}.$$

  - Auxiliary Component

$$\text{loss}_{aux} = \frac{1}{m-1} \sum_{i=1}^{m-1} \frac{\left| e_i - \Delta t_{p_{L_i} \to p_{L_{i+1}}} \right|}{\Delta t_{p_{L_i} \to p_{L_{i+1}}} + \epsilon}.$$

  - Final loss:

$$\text{loss} = \text{loss}_{seq} + \alpha \cdot \text{loss}_{aux}$$

Experiment

Data Description

- 1.4 billion GPS records of 14,864 taxis in Oct. 2014 in Chengdu.
- Total number of trajectories: 9,653,822. (60GB)
- Use the last 7 days (from 24th to 30th) as the test set and the remaining ones as the training set.

Experiment

Table: Performance Comparison

Model	MAPE
Gradient Boosting	20.32%
MLP-3 layers	16.17%
MLP-5 layers	15.75%
Vanilla RNN	18.85%
DeepTTE	**13.14%**

CONTENTS

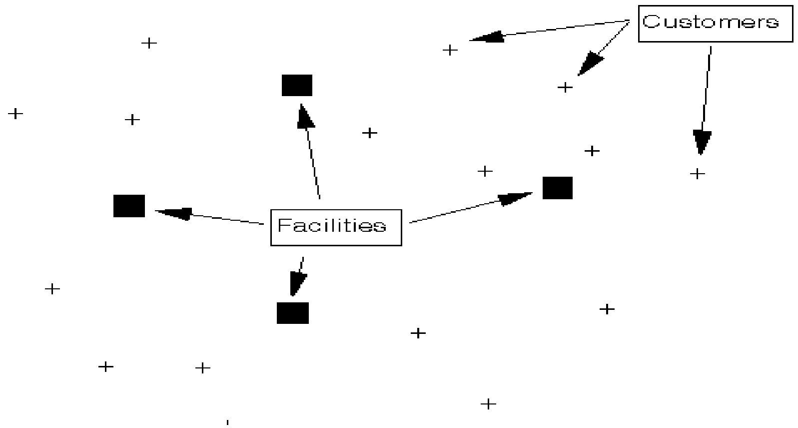
Store Site Selection

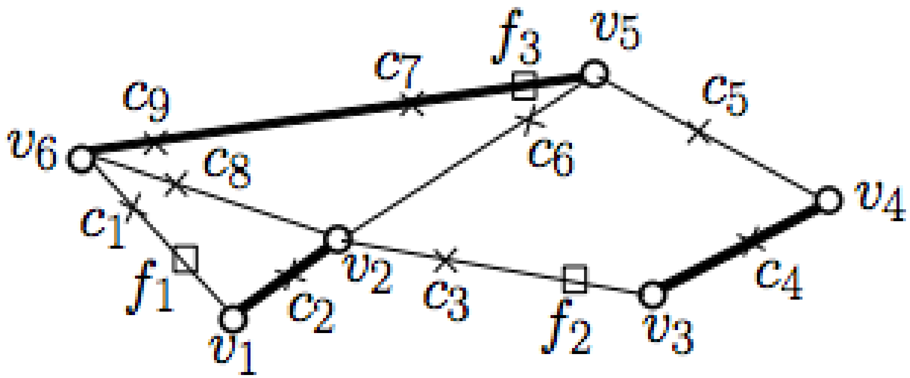
Where to open a new store (optimal facility location problem)?

- demand prediction
- existing competition
- crowd profile (we are planning to use Topic models)
  - A store: a word
  - A sequence of stores a user visited: a sentence
- location selection (a geometric optimization problem)

Store Site Selection

Introduction





Joint work with Baidu Big Data lab

Store Site Selection

Existing work

- Previous work
  - Make decision in given locations.
  - Check-in data
  - Linear supervised learning model.

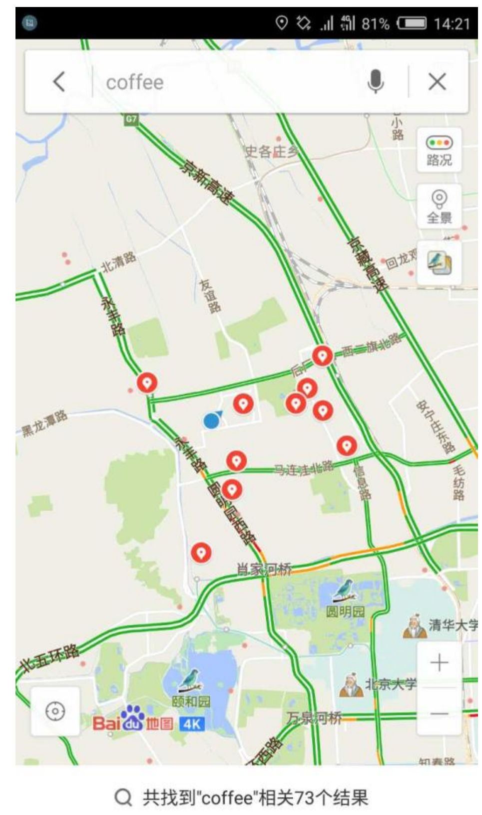
- Our work
  - Demand based
  - Mining features and target from multiple spatial-temporal data sources

Store Site Selection

User Demand Analysis

Specific('Starbucks')

General('Coffee shops')



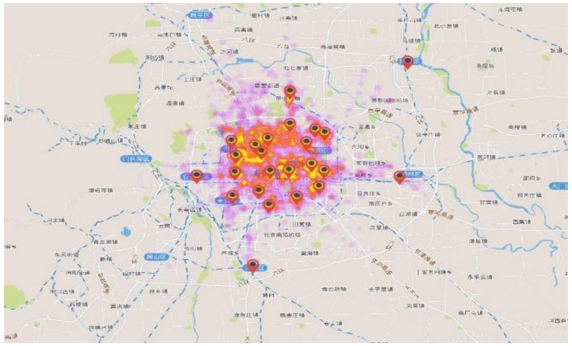
- [u, 2015-08-08, (116.34, 40.02), 08:42:28,"Starbucks"]
- User demand: D = (lat, lng, t)
- Two types of demands

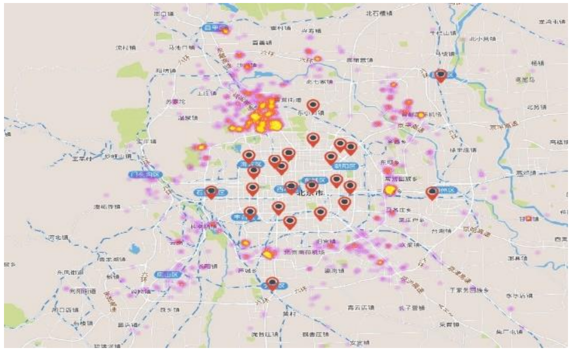
Store Site Selection

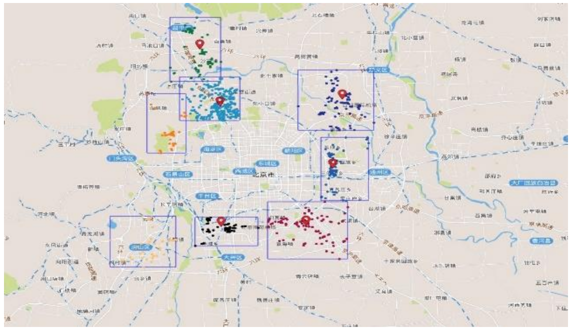
清華大學 Tsinghua University

Finding Demand Centers

- Identify demand points
- Exclude supplies
  - Specific demands
  - General demands
- Clustering demands







Store Site Selection

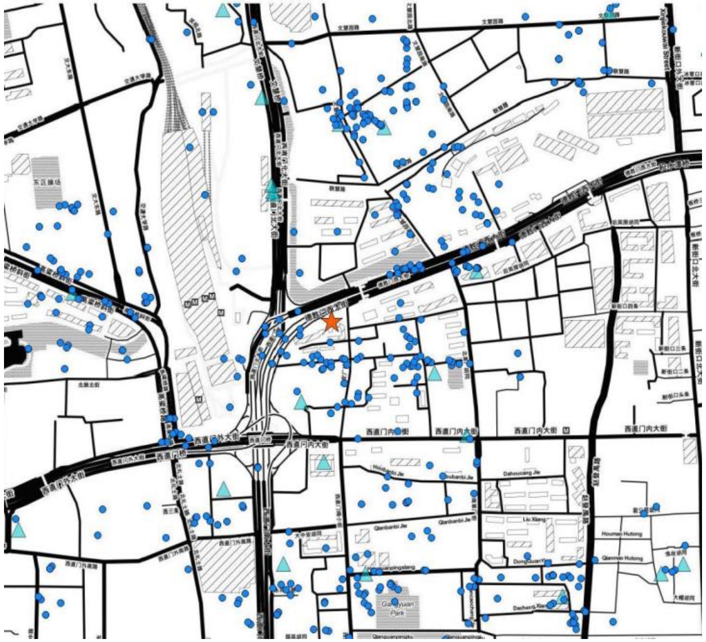
Exclude supplies: General

Exclude supplies with some probabilities

Distance score $S_d=1-e^{-d(lu-ld)^2/\sigma^2})$

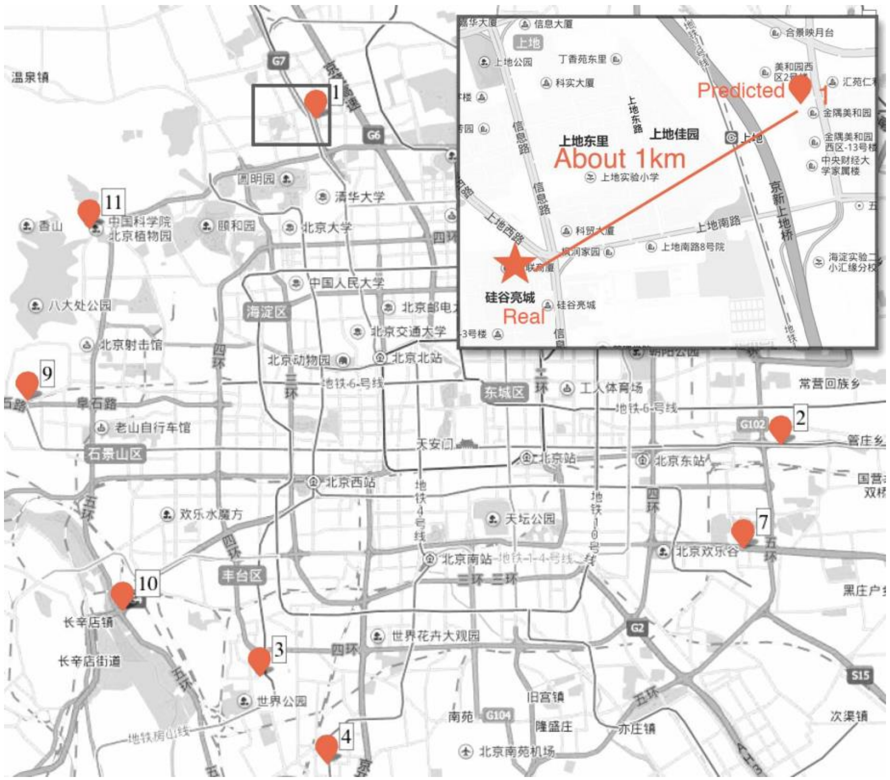
Supply Score $S_s = e^{-\varepsilon N}$

Remaining score $S_r=\alpha S_d+(1-\alpha)S_s$



Store Site Selection

Real Cases



chain hotpot restaurant “HaiDiLao” opened at September, 2015

“Starbucks” opened at January, 2016

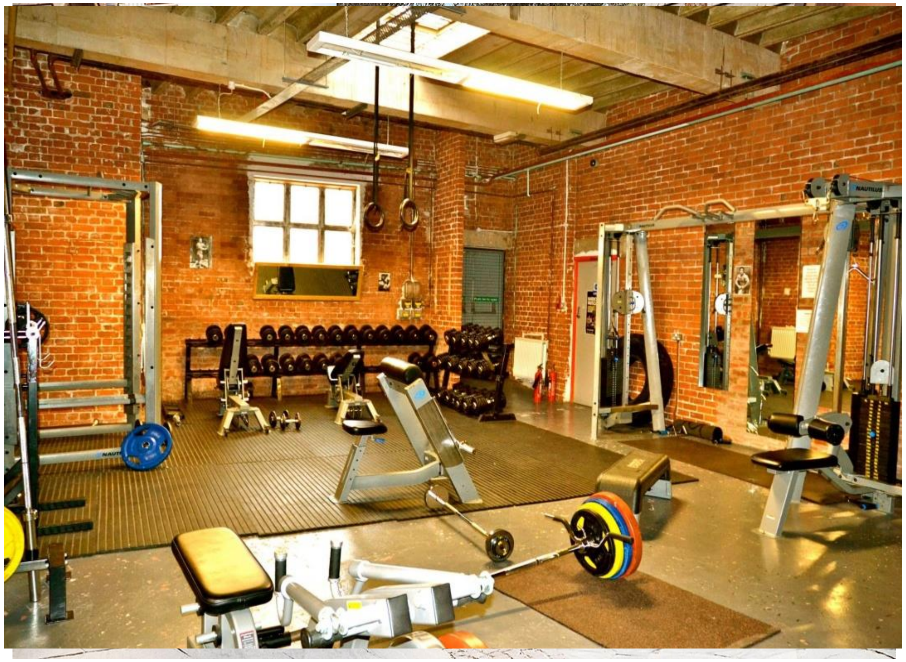
CONTENTS

Inferring POI Visitation

清華大學 Tsinghua University

Introduction

Given a user, the location and the corresponding timestamp, we want to figure out the actual POI that she or he most likely has visited.



✓Understanding the characteristics of users.

✓Useful for recommendation systems, advertisements, check-in systems.

Joint work with Baidu Big data Lab

Inferring POI Visitation

Tsinghua University

Limitation of Existing work

- Distance based neighborhood models:
- Supervised learning-to-rank algorithms

Inferring POI Visitation

Tsinghua University

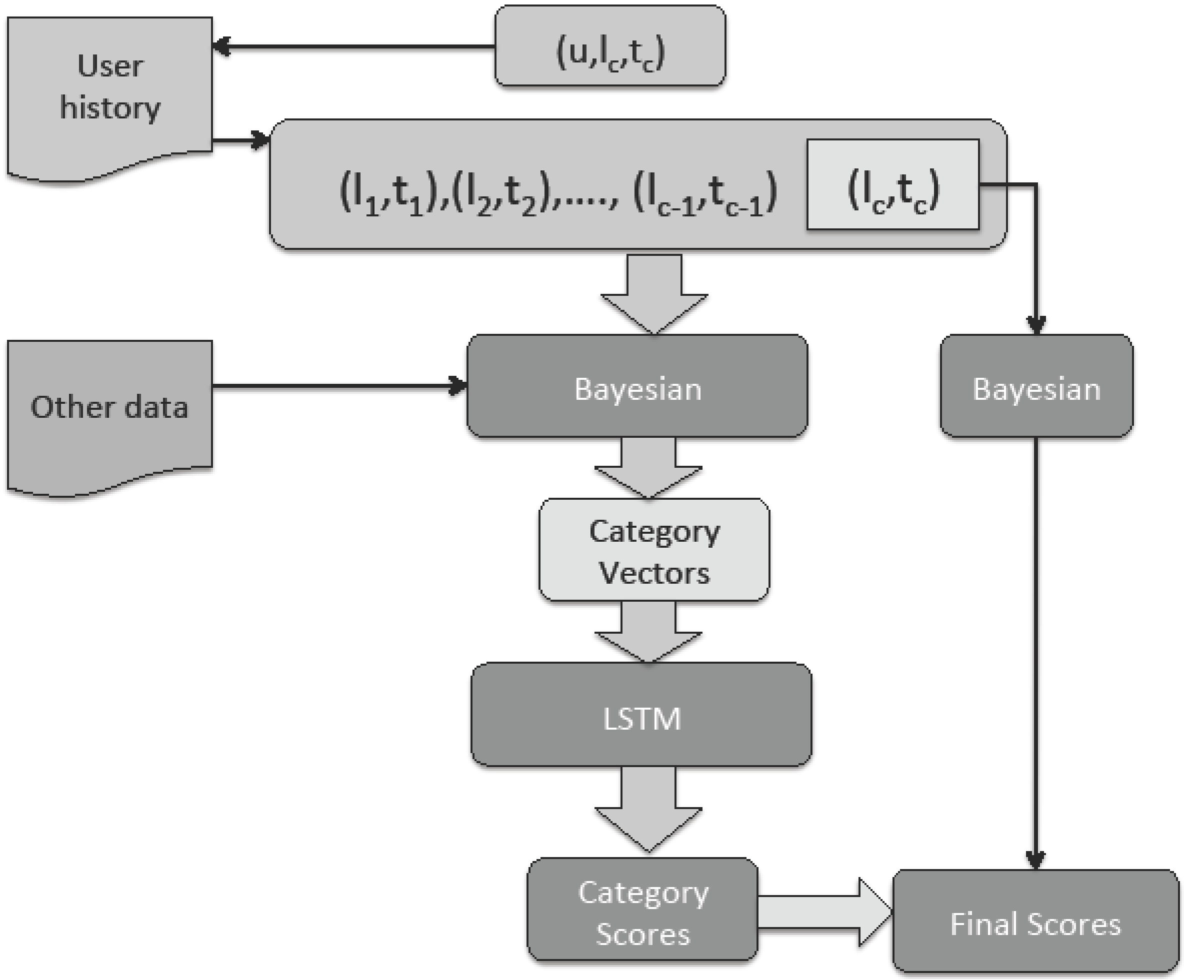
Framework

- Three steps:

✓ Bayesian inference.

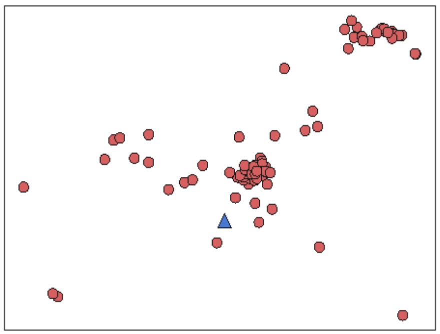
✓ LSTM-based inference.

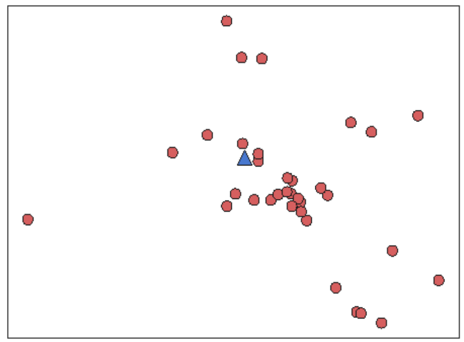
✓ Model fusion.

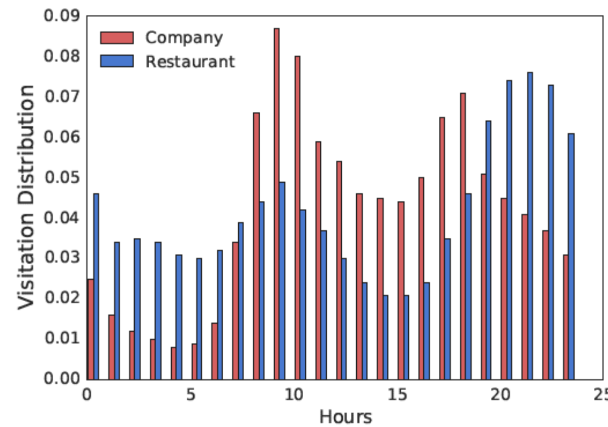


Inferring POI Visitation

Bayesian Inference







$$P(l_u|poi = p) = \sum_{\mu_k} w_p(\mu_k)P(l_u|\mu_k)$$

$$P(l_u|\mu_k) = N(l_u|l_{\mu_k}, \sigma^2)$$

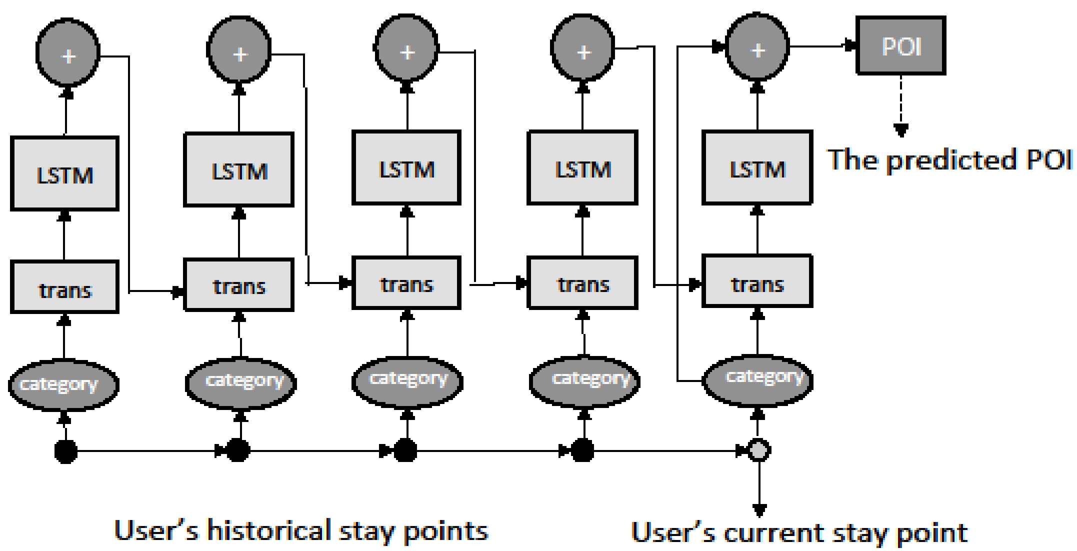
$$w_p(\mu_k) = \frac{{}^{1}/_{d(\mu_k, p) + 1}}{\Sigma_j {}^{1}/_{d(\mu_j, p) + 1}}$$

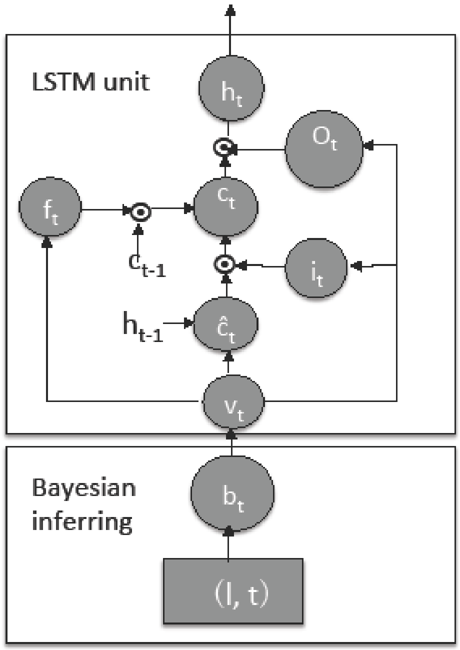
$$P(t|p) = P(hour|p)P(week|p)$$

$$P(p) \propto P(wifi_p) * P(query_p)$$

Inferring POI Visitation

LSTM-based model





$$S_{p_i}(hs_c, l_c, t_c) = P(poi = p_i | l_c, t_c) * h_c$$

Inferring POI Visitation

Experiments

- Data1: used to generate the POI features for Bayesian model
  Map query data, GPS data, and WiFi data from Jun. 2015 to Dec. 2015.

- Data2: groudtruth
  Check-in data of Nuomi in China from Jan. 1st, 2016 to Mar. 31th, 2016

- 23 categories in POIs of Nuomi.
  Accuracy for predicting POI's category with the LSTM based model is about 0.40.

Inferring POI Visitation

Tsinghua University

Experiments

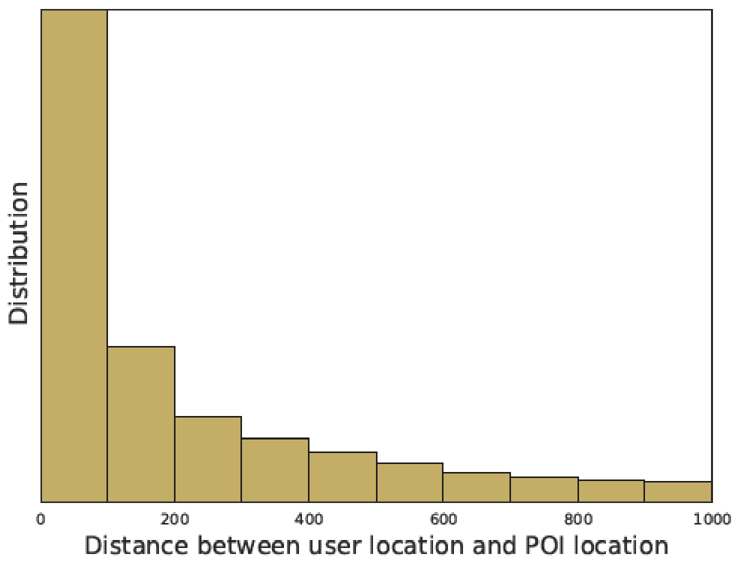
- The check-in data is very sparse, hard to extract sequential information from RNN by only using the check-in data.

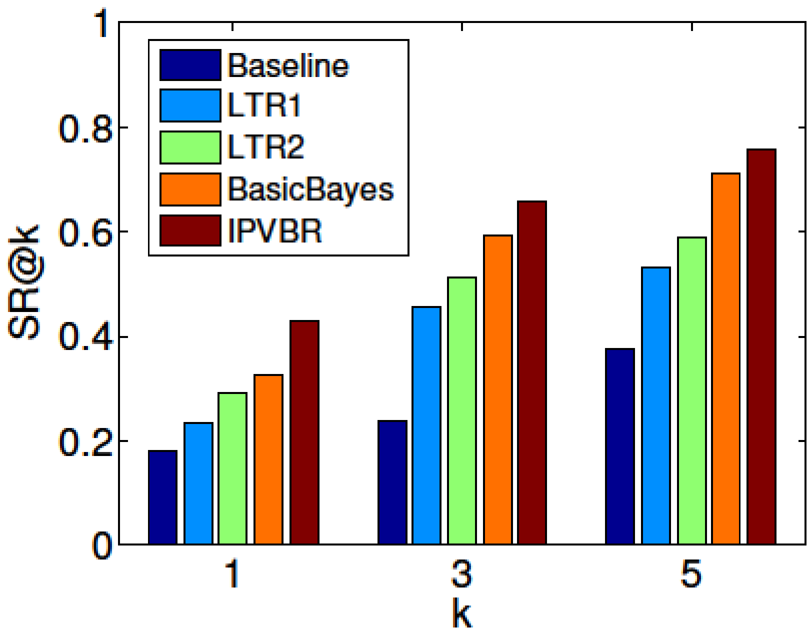
- However, in our framework IPVBR, we use historical stay points as the input to avoid such problem.

Time duration	Average number of check-ins for users
3 months	2.7
6 months	9.1
12 months	21.4

Inferring POI Visitation

Experiments





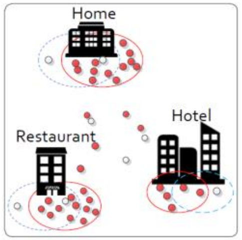
Other work: Automatic User Identification across Heterogeneous Data Sources

Goal: Identify the same user from the historical trajectory data set.

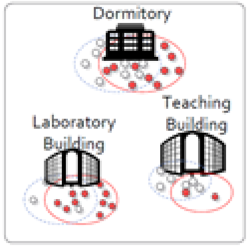
Motivation: human mobility, data integration, improve data quality

Challenges:

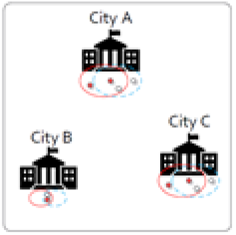
- Very different sampling rates
- Information loss in sparse trajectories
- Temporally disjoint
- Distinguish the overlaps



Same person, different sampling rates



School mates, significant overlap

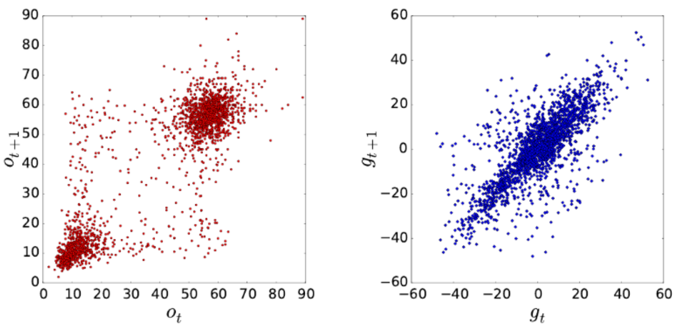


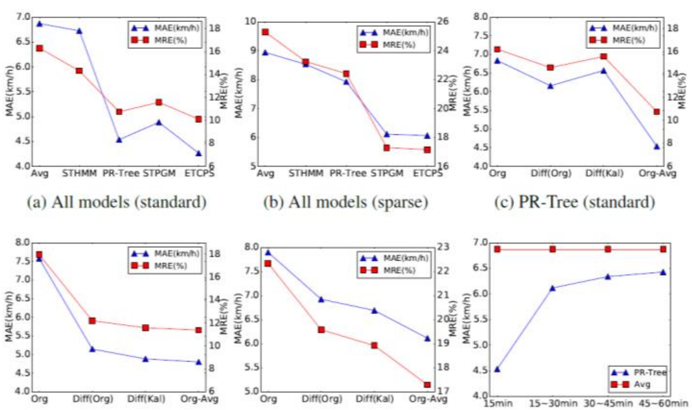
Same person, sparse rate, occurred in several places

[in ICDE16]

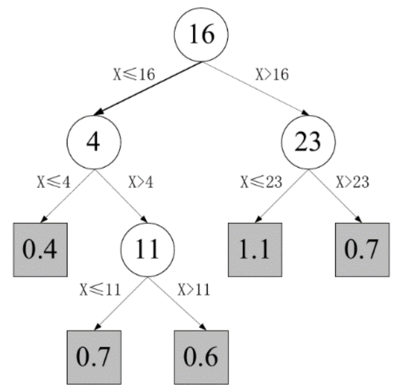
Other work: Traffic Condition Prediction

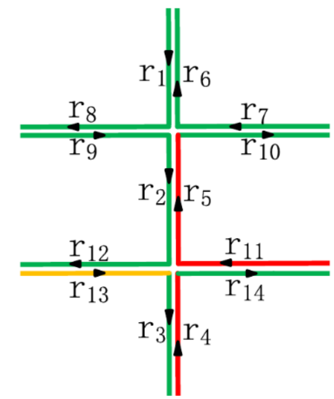
Relationships observation




(a) All models (standard) (b) All models (sparse) (c) PR-Tree (standard)
(d) STPGM (standard) (e) STPGM (sparse) (f) Predict longer

- PR-Tree models the traffic condition time series of each individual roads





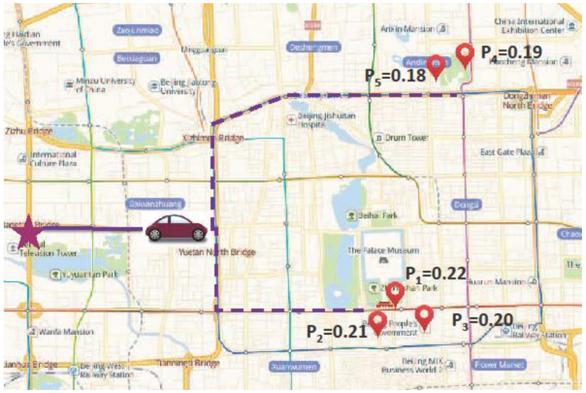
- STPGM models the relationship between different roads
- Our best quality prediction is achieved by a careful ensemble of the two models.

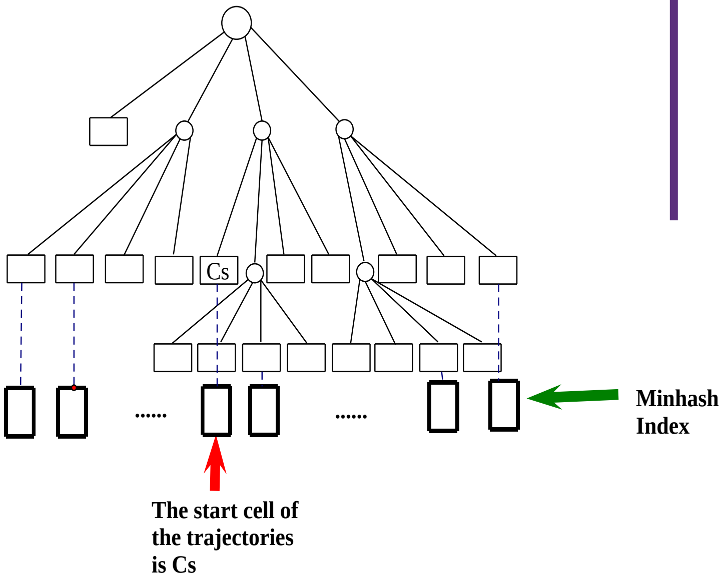
[DASFAA’16]

Other work: Destination Prediction

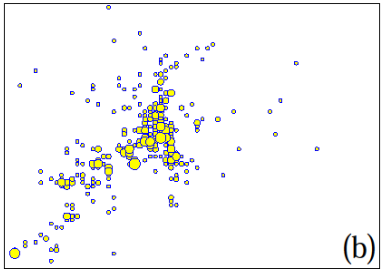
Problem Definition

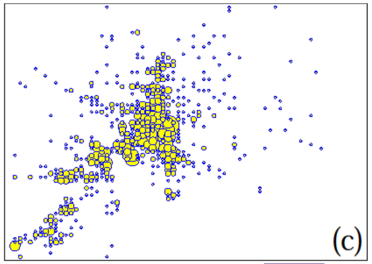
Destination prediction is to predict the destination of a trip given a partial passed trajectory.





Applications





[UbiComp'16]

Thank you