

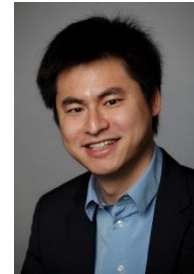
Urban Computing

–Using Big Data to Solve Urban Challenges

Dr. Yu Zheng

Lead Researcher, Microsoft Research

Chair Professor at Shanghai Jiao Tong University



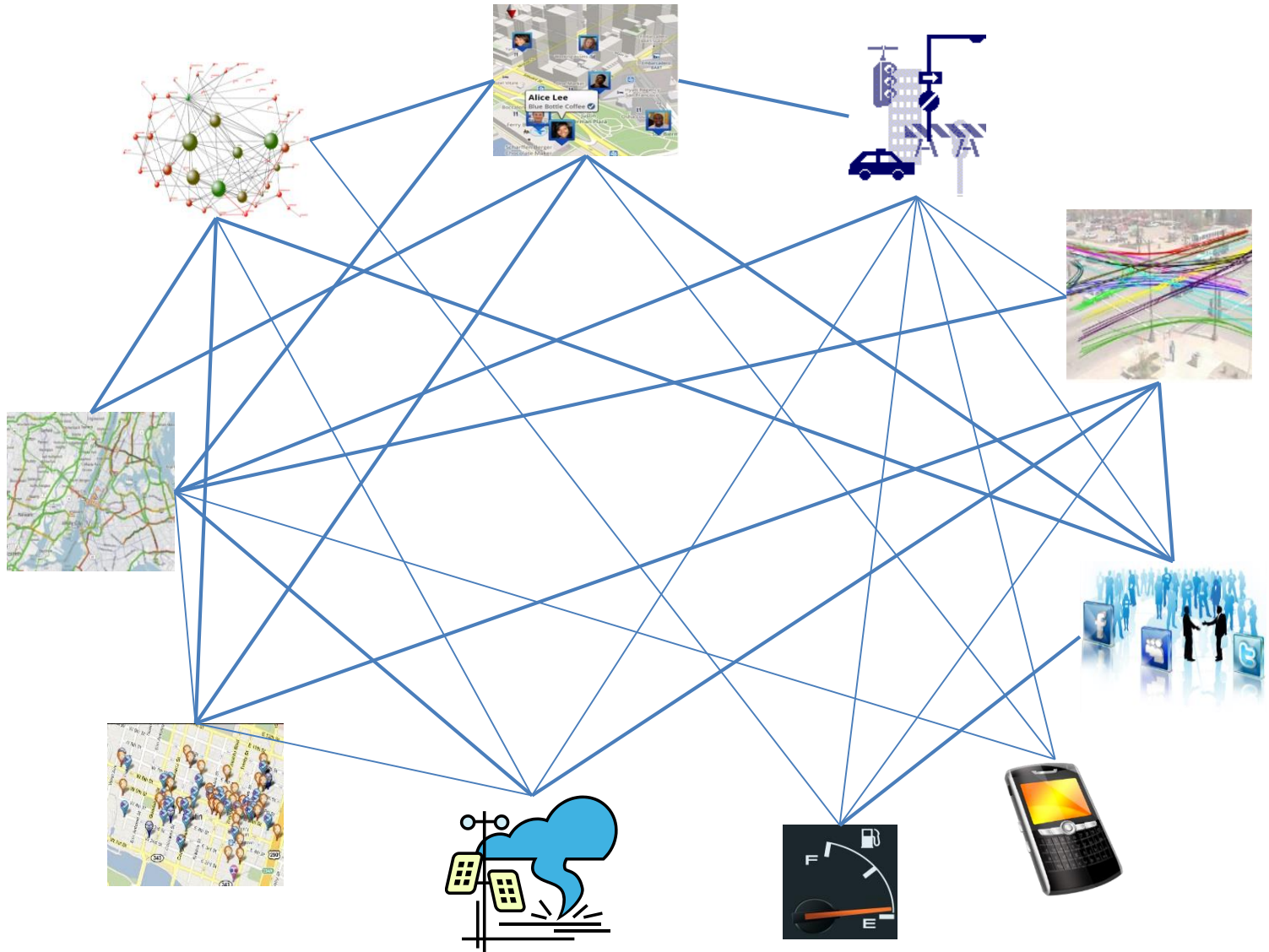
<http://research.microsoft.com/en-us/projects/urbancomputing/default.aspx>

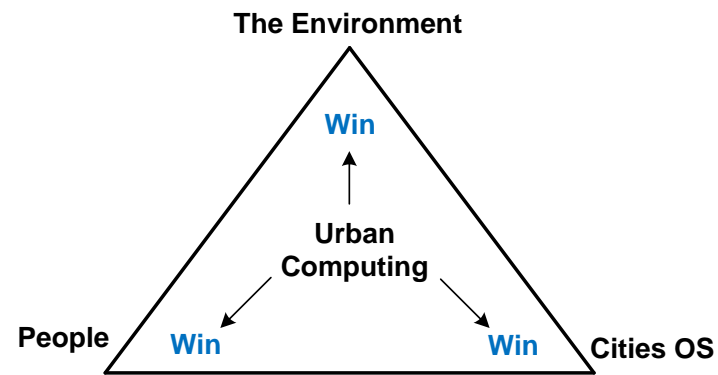
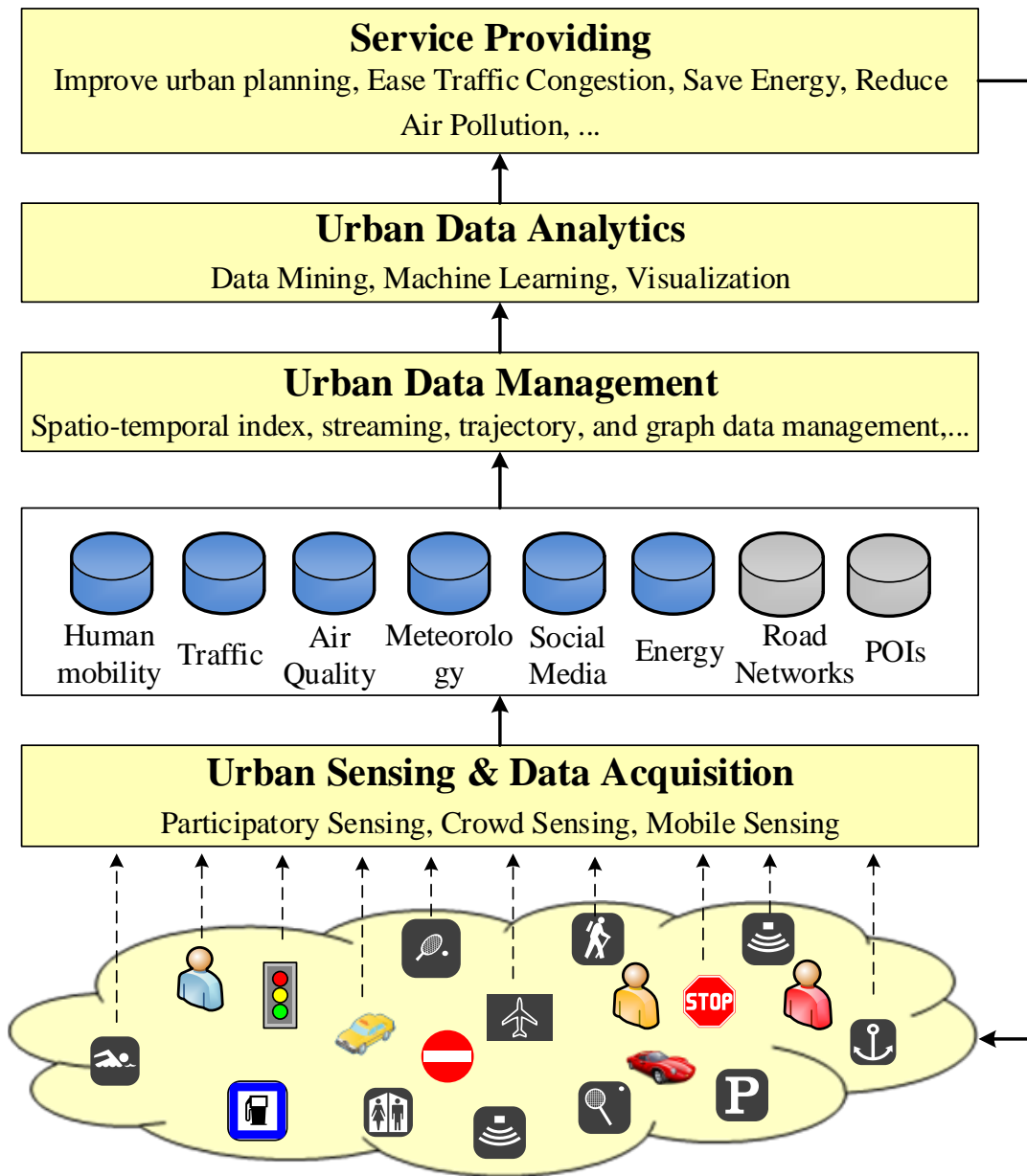


Big Challenges in Big Cities



Big Data in Cities





*Tackle the **Big** challenges
in **Big** cities
using **Big** data!*

Key Focuses and Challenges

- Sensing city dynamics

- Unobtrusively, automatically, and constantly
- A variety of sensors: Mobile phones, vehicles, cameras, loops,...
- **Human as a sensor:** User generated content (check in, photos, tweets)
 - Loose control and unreliable → data missing and skewed distribution
 - Unstructured, implicit, and noisy data
 - Trade off among energy, privacy and the utility of the data



- Computing with heterogeneous data sources

- Geospatial, temporal, social, text, images, economic, environmental,...
- Learn mutually reinforced knowledge across a diversity of data
- Efficiency + Effectiveness: Data Management + Mining + Machine Learning

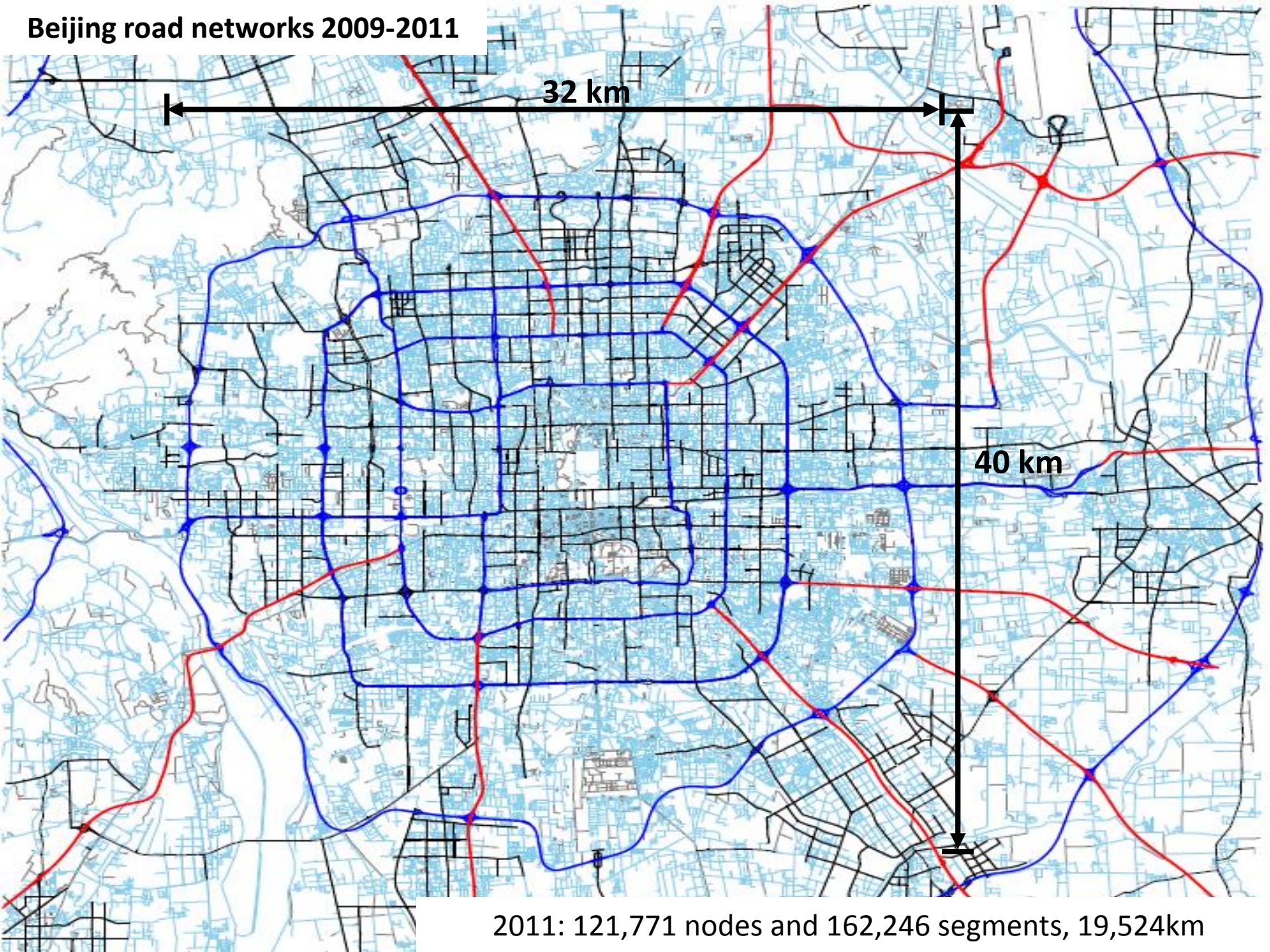


- Blending the physical and virtual worlds

- Serving both people and cities (virtually and physically)
- **Hybrid systems:** Mobile + Cloud, crowd sourcing, participatory sensing...

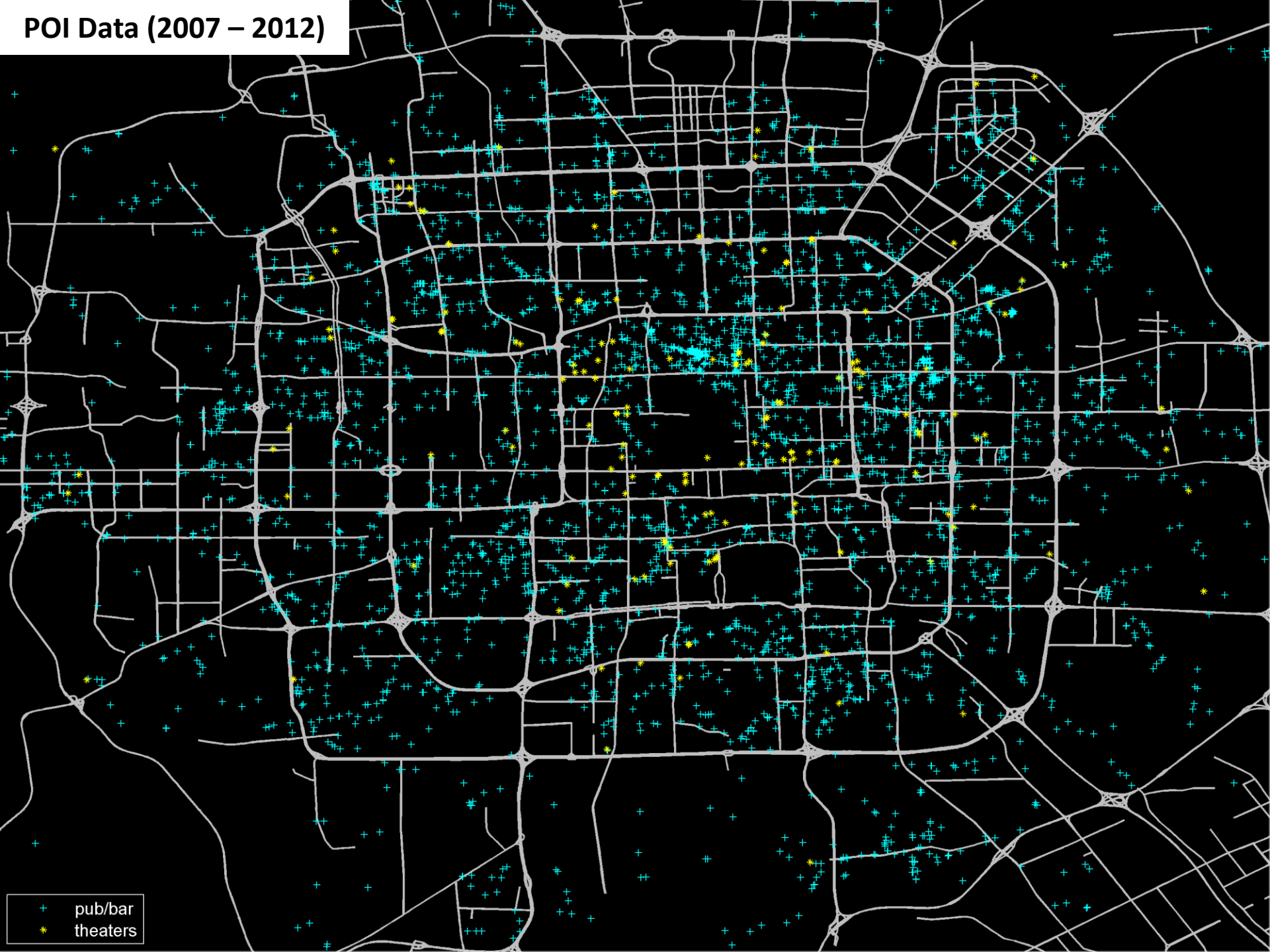


Beijing road networks 2009-2011

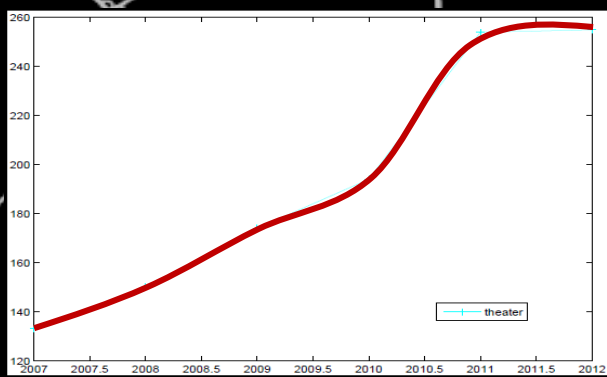


2011: 121,771 nodes and 162,246 segments, 19,524km

POI Data (2007 – 2012)



+ pub/bar
* theaters



Air Quality Data

Forecast | Current AQI | More Maps



| Click on the city name for more detailed information. printable summary | FORECAST | | CURRENT AQI |
|--|------------|------------|-------------|
| | Tue Aug 20 | Wed Aug 21 | |
| Ann Arbor | ! USG | n/a | 69 |
| Benton Harbor | Mod | n/a | 67 |
| Detroit | ! USG | n/a | 84 |
| Eastern U.P. | Mod | n/a | 62 |
| Flint | Mod | n/a | 66 |
| Grand Rapids | ! USG | n/a | 74 |
| Houghton Lake | Mod | n/a | 63 |
| Kalamazoo | Mod | n/a | 67 |
| Lansing | Mod | n/a | 72 |
| Ludington | ! USG | n/a | 63 |
| Saginaw | Mod | n/a | 73 |
| Traverse City | Mod | n/a | 63 |

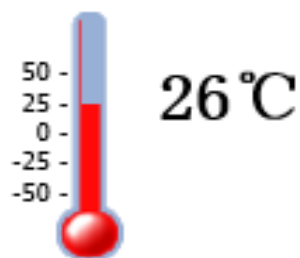
Good
Moderate
USG
Unhealthy
Very Unhealthy
Hazardous
! Action Day

Meteorological data



Weather Observation

Temp



Humidity: 73%

Wind direction



Monthly Climate History

Accumulated Rainfall

159.7mm

Average Temperature

24.9°C

High Temperature

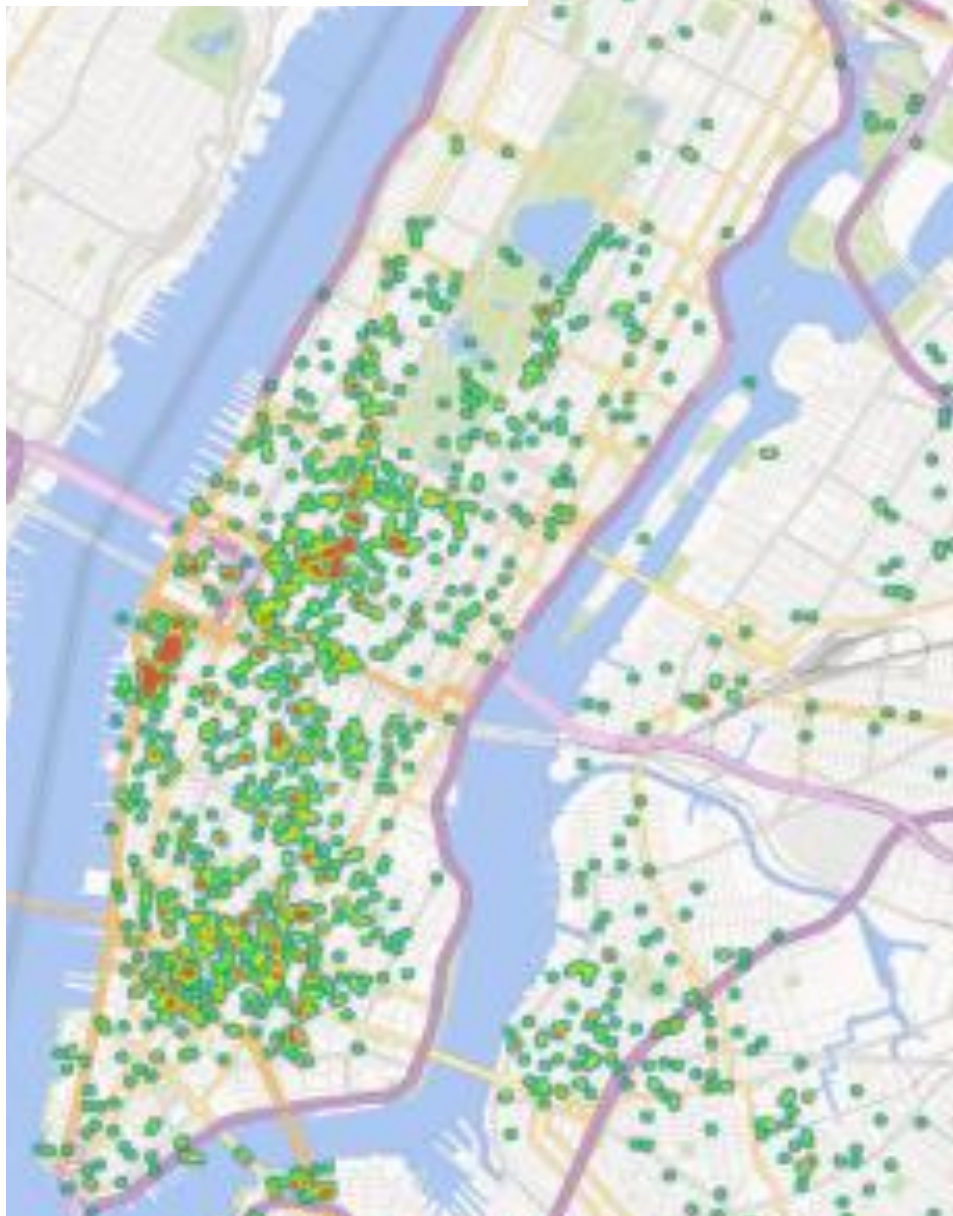
36.1°C

BeijingWeather Forecast (2013-08-20 18:00)

4-7 Days Forecast

| Date | | weatherForecast | | Temperature | wind |
|------------------|-------|-----------------|--------|-------------------|---------|
| Tuesday Aug 20 | night | | Shower | Low: 23°C (73°F) | <12km/h |
| Wednesday Aug 21 | day | | Cloudy | High: 30°C (86°F) | <12km/h |
| | night | | Cloudy | Low: 22°C (72°F) | <12km/h |
| Thursday Aug 22 | day | | Sunny | High: 29°C (84°F) | <12km/h |
| | night | | Sunny | Low: 22°C (72°F) | <12km/h |
| Friday Aug 23 | day | | Sunny | High: 32°C (90°F) | <12km/h |
| | night | | Sunny | Low: 22°C (72°F) | <12km/h |

Check-in data

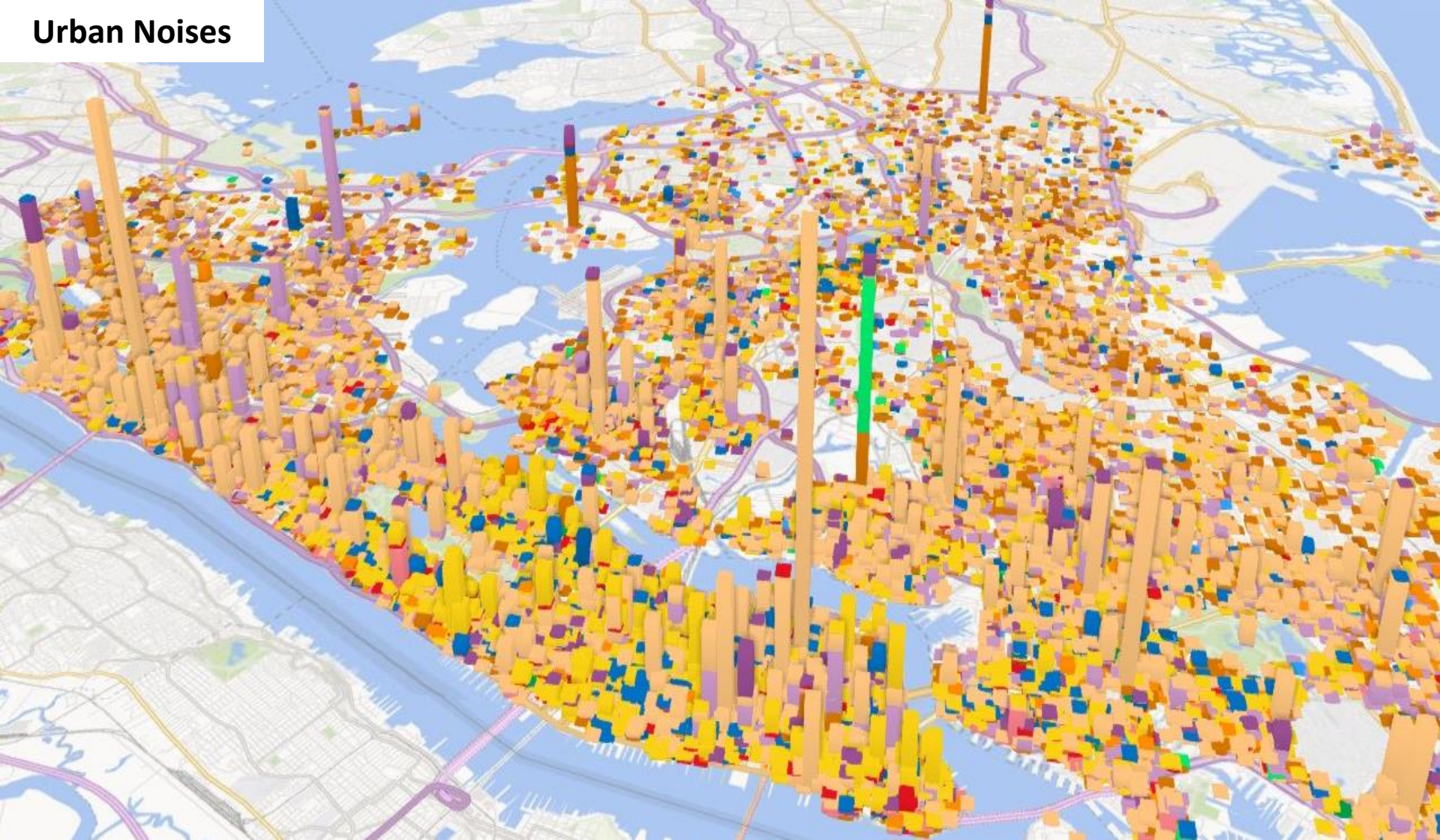


Check-in: Entertainment



Check-ins: Nightlife Spot

Urban Noises

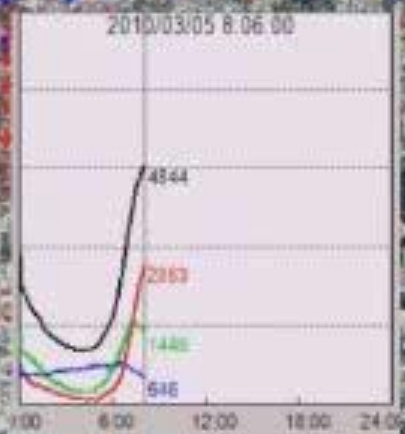
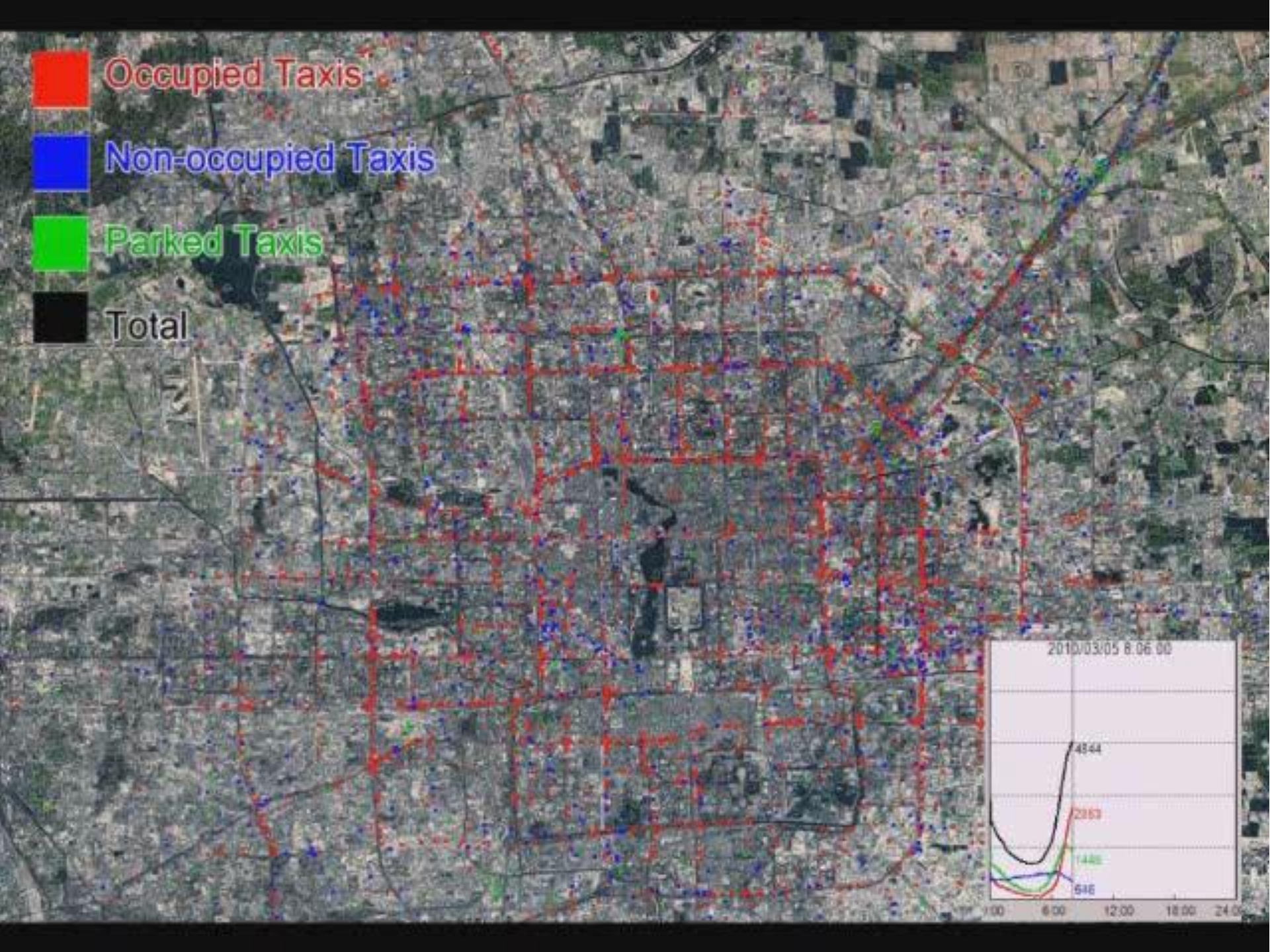


- | | | | |
|--|--|--|---|
|  Air condition/Ventilation equipment |  Loud Music/Party |  Horn Honking Sign Requested |  Others |
|  Alarms |  Loud Talking |  Jack Hammering |  Private carting noise |
|  Banging/Pounding |  Loud Television |  Lawn care equipment |  Vehicle |
|  Construction |  Manufacturing | | |

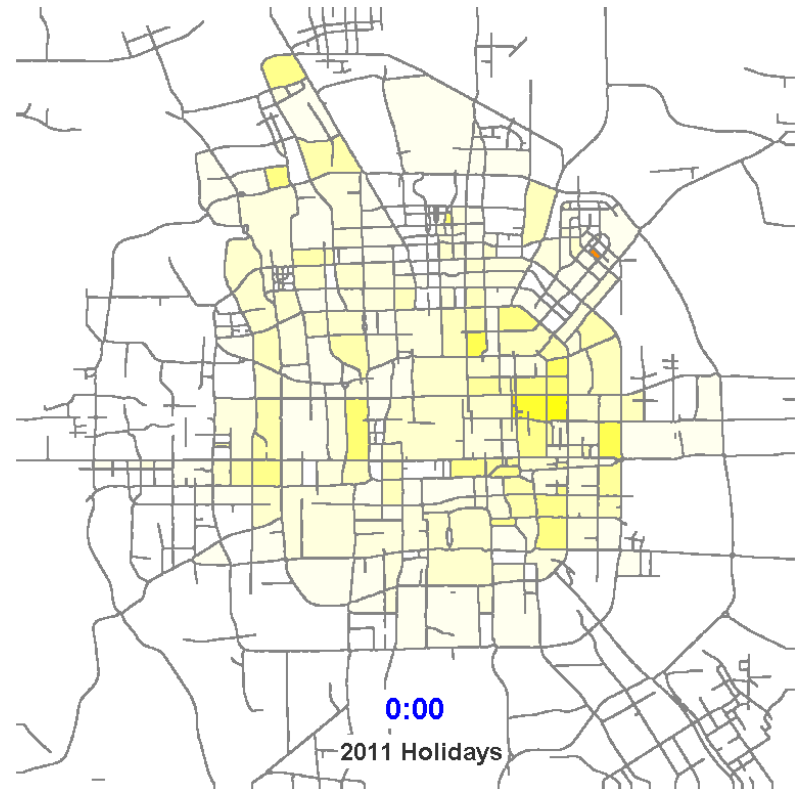
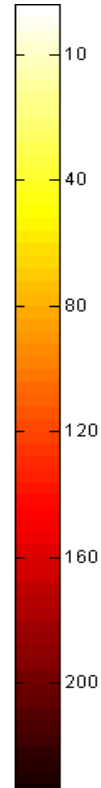
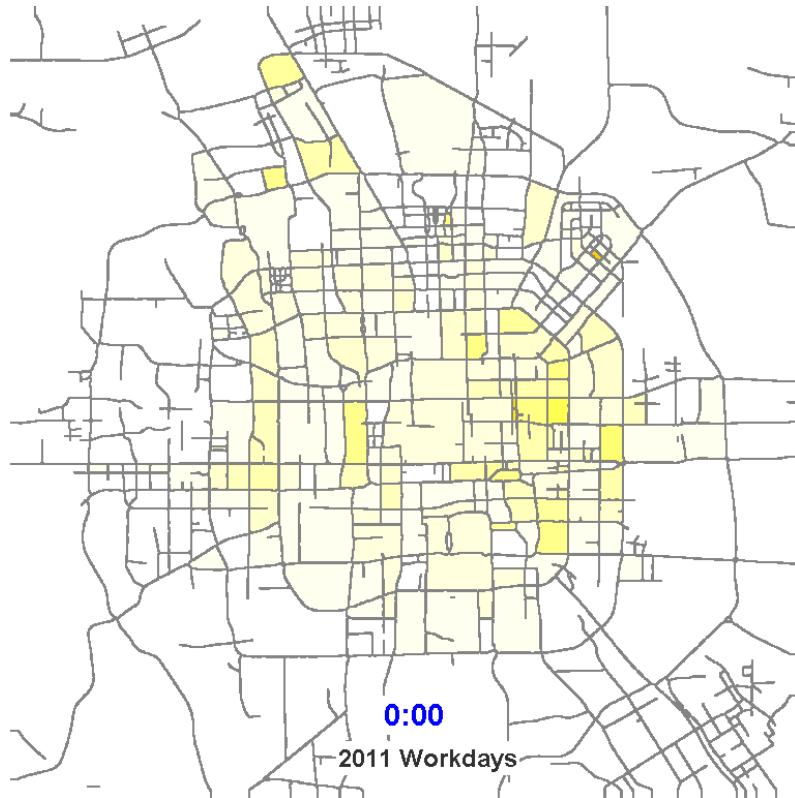


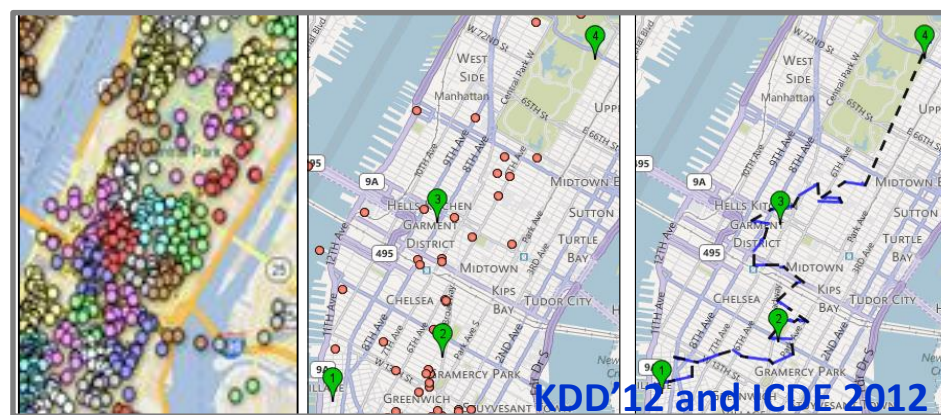
GPS trajectories of 33,000 taxis from 2009 to 2013

- Occupied Taxis
- Non-occupied Taxis
- Parked Taxis
- Total

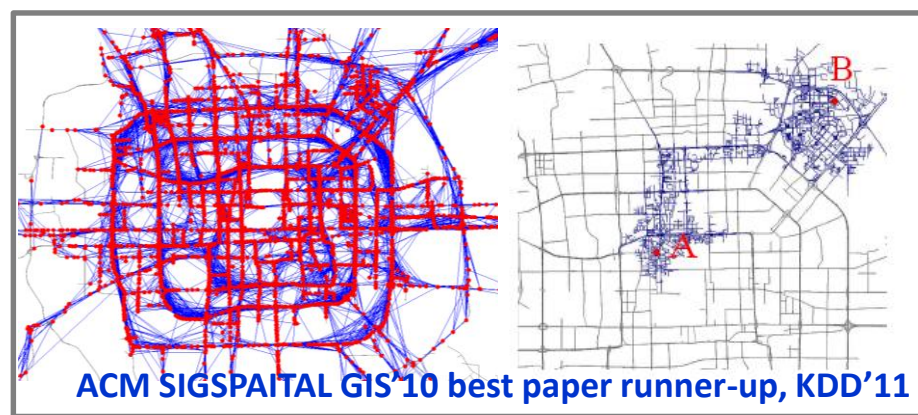


Heat Maps of Beijing (2011)

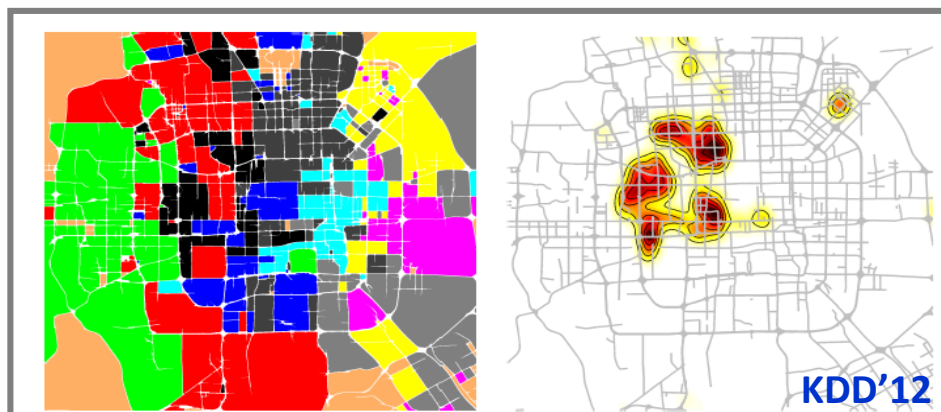




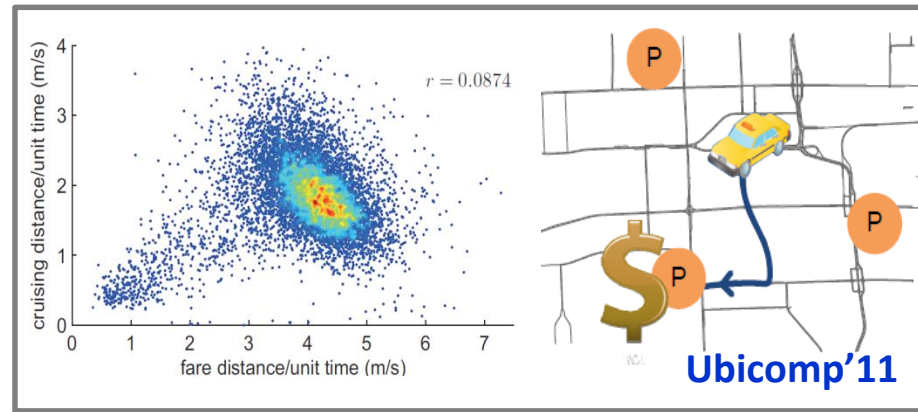
Route Construction from Uncertain Trajectories



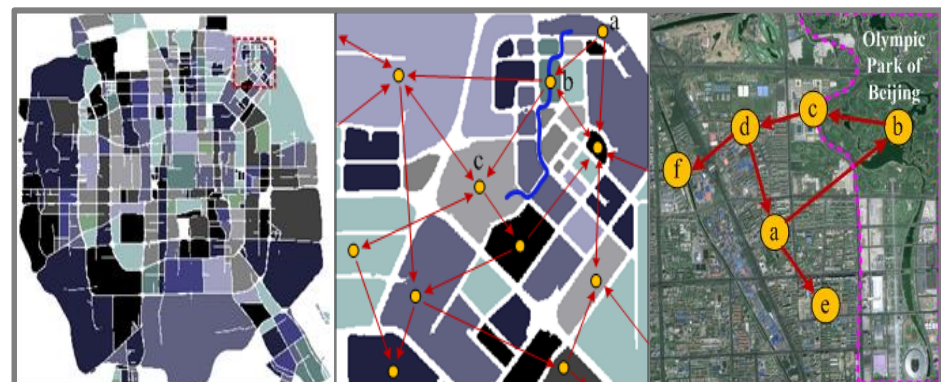
Finding Smart Driving Directions



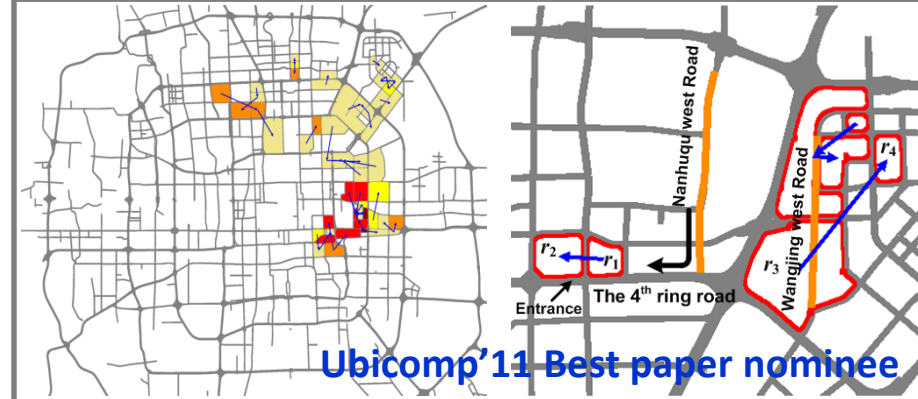
Discovery of Functional Regions



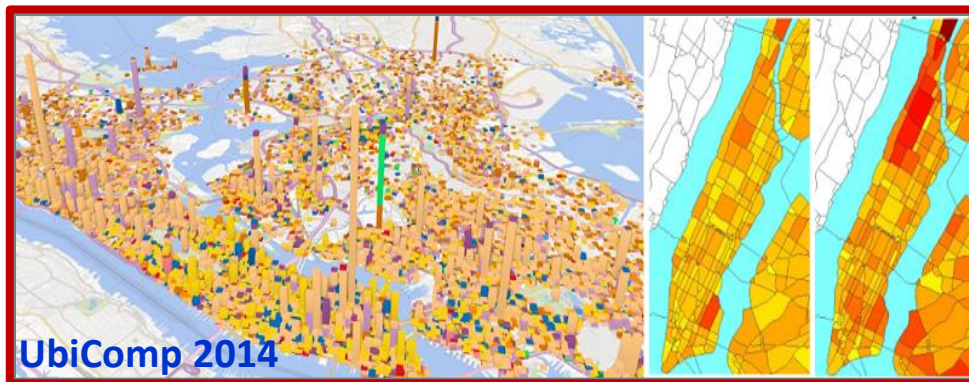
Passengers-Cabbie Recommender system



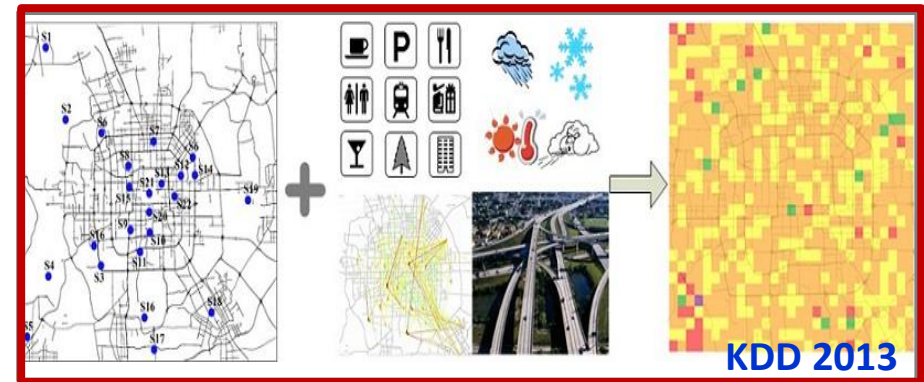
Anomalous Events Detection KDD'11 and ICDM 2012



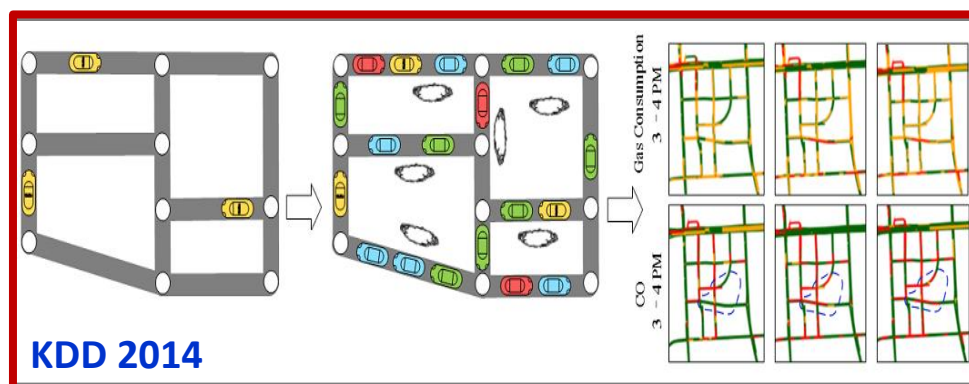
Urban Computing for Urban Planning



Diagnose urban noises using big data



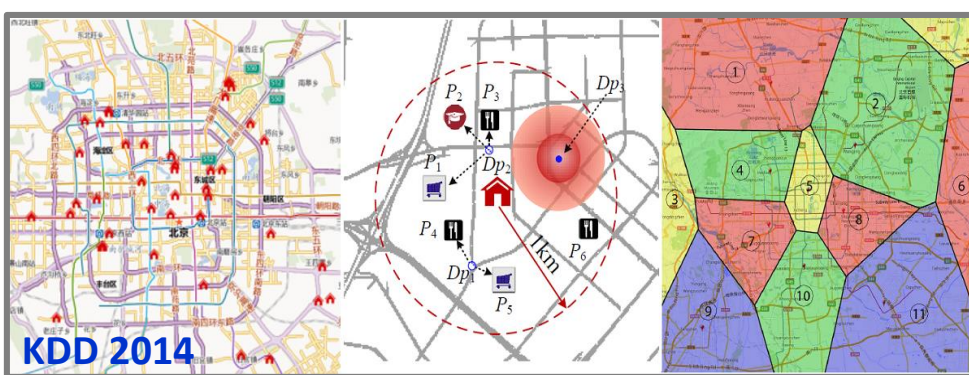
Infer air quality using big data



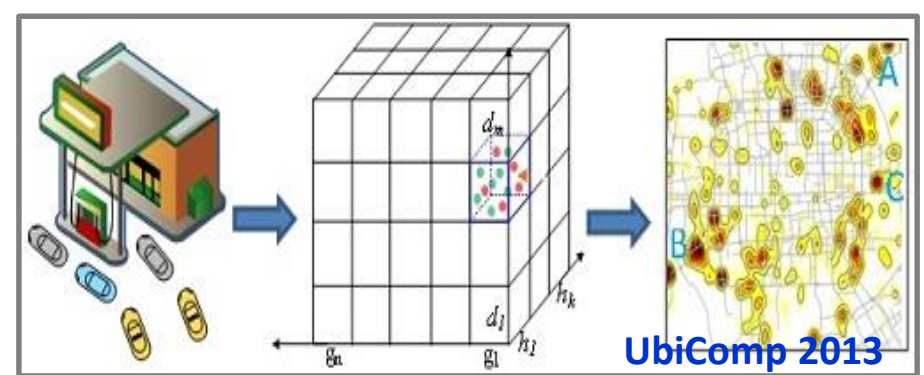
Real-time gas consumption and pollution emission



Real-time and large-scale dynamic ridesharing



Residential Real estate ranking and clustering

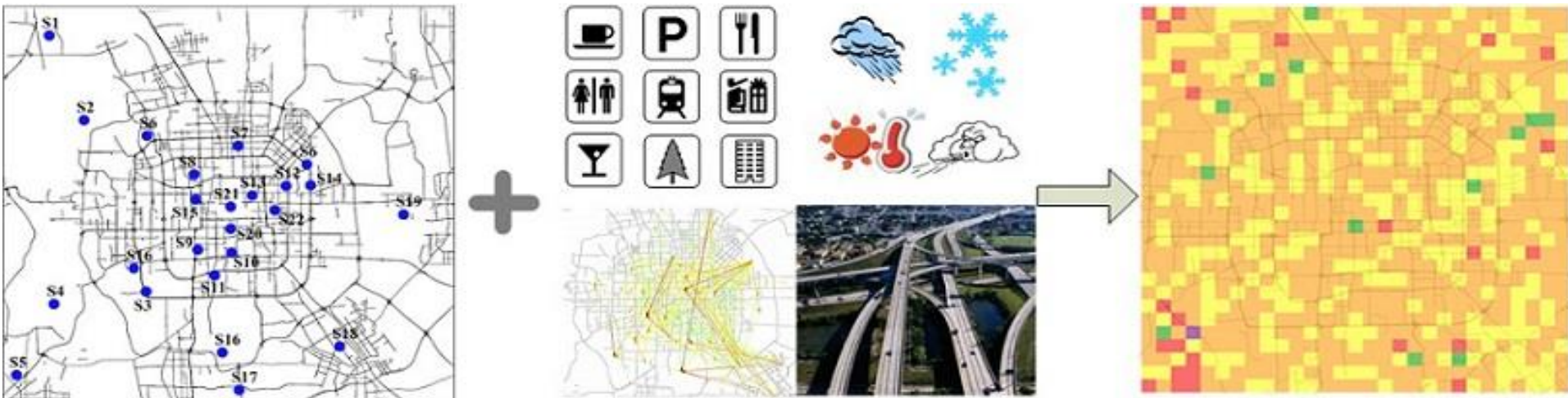


Real-time city-scale gas consumption sensing

When **Urban Air** Meets **Big Data**

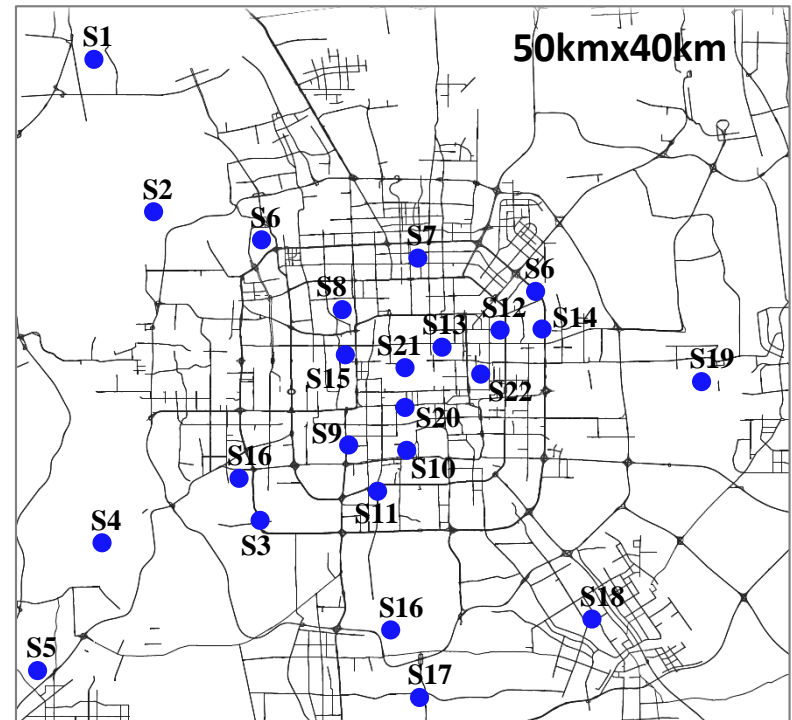
KDD 2013

<http://urbanair.msra.cn/>

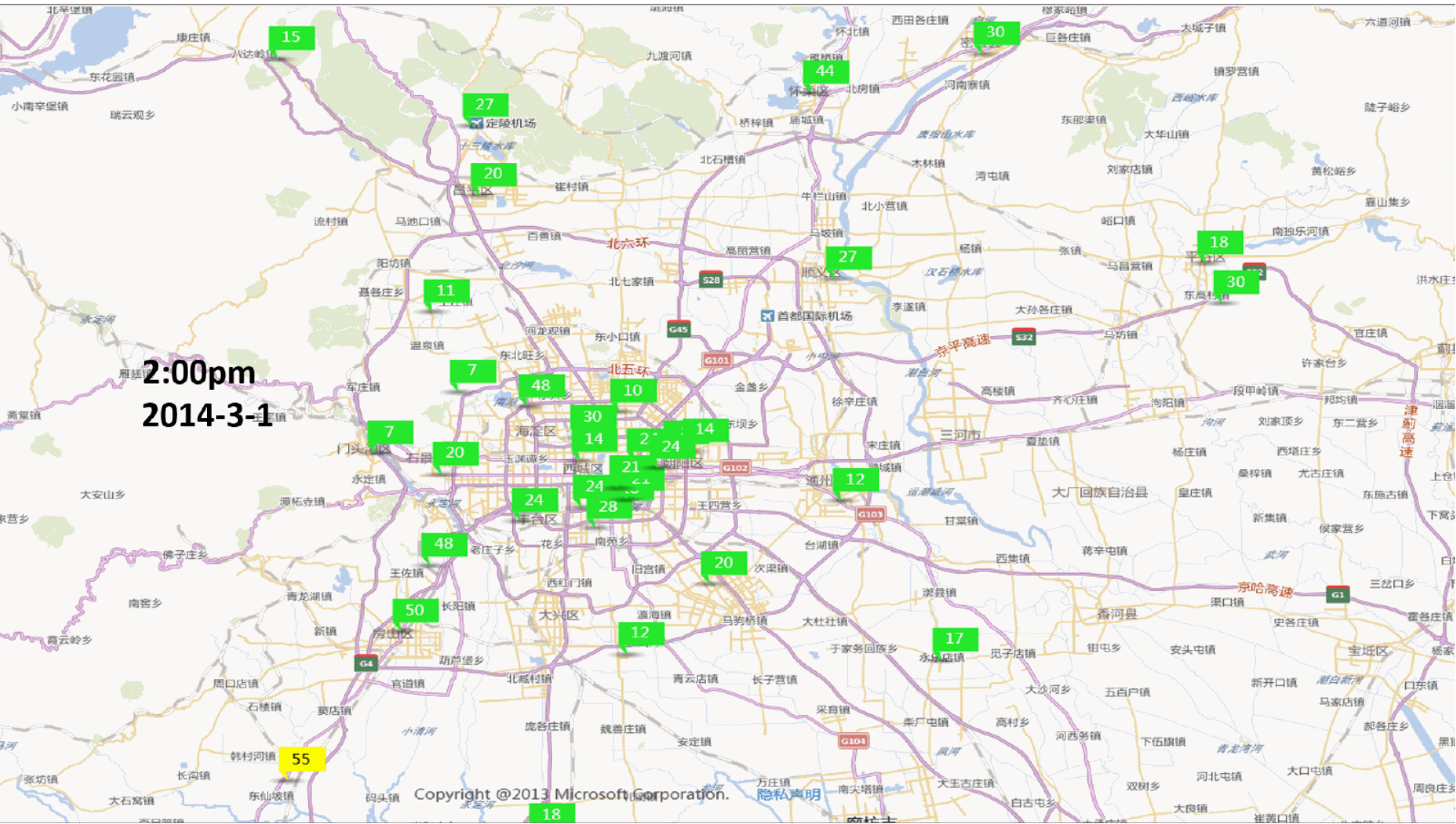


Background

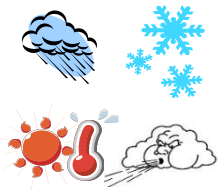
● Air quality monitor station



We do not really know the air quality of a location without a monitoring station!



Inferring **Real-Time** and **Fine-Grained** air quality throughout a city using **Big Data**



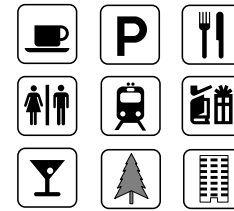
Meteorology



Traffic



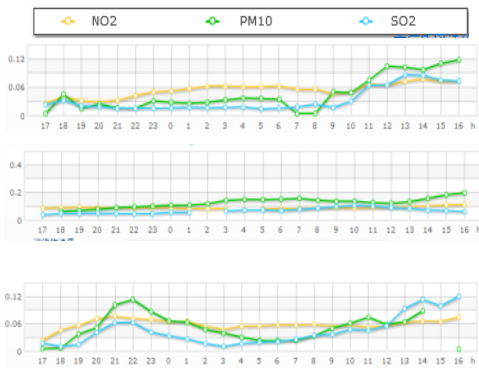
Human Mobility



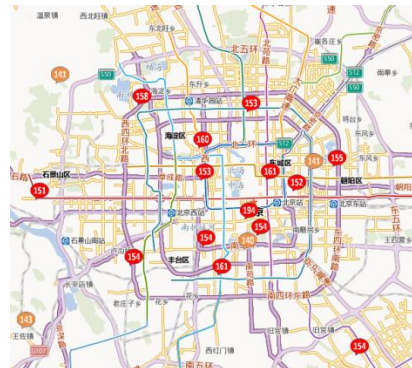
POIs



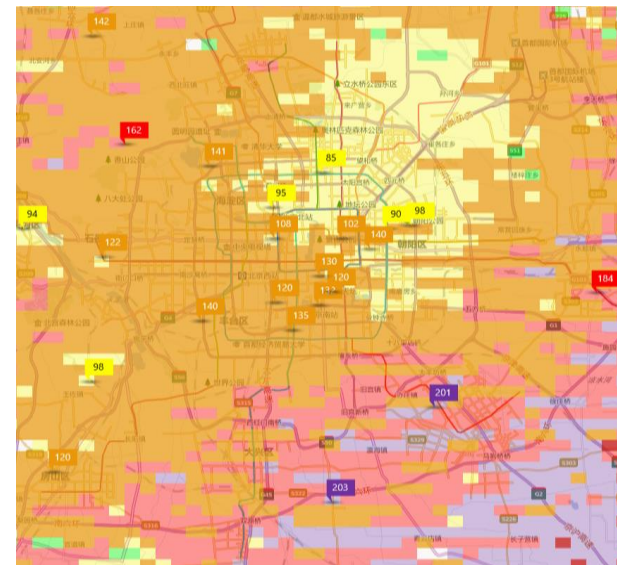
Road networks



Historical air quality data



Real-time air quality reports



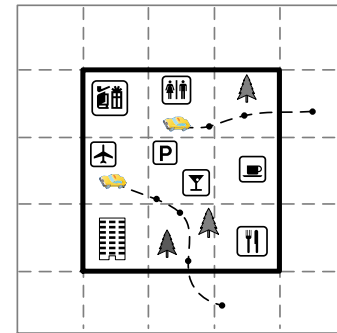
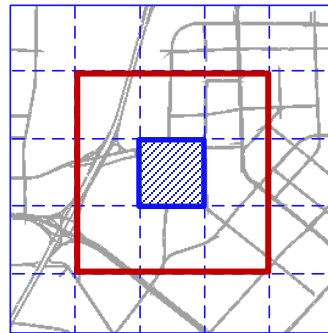
Difficulties

- Incorporate multiple heterogeneous data sources into a learning model
 - Spatially-related data: POIs, road networks
 - Temporally-related data: traffic, meteorology, human mobility
- Data sparseness (little training data)
 - Limited number of stations
 - Many places to infer
- Efficiency request
 - Massive data
 - Answer instant queries

Methodology Overview

- Partition a city into disjoint grids
- Extract features for each grid from its impacting region

- Meteorological features
- Traffic features
- Human mobility features
- POI features
- Road network features



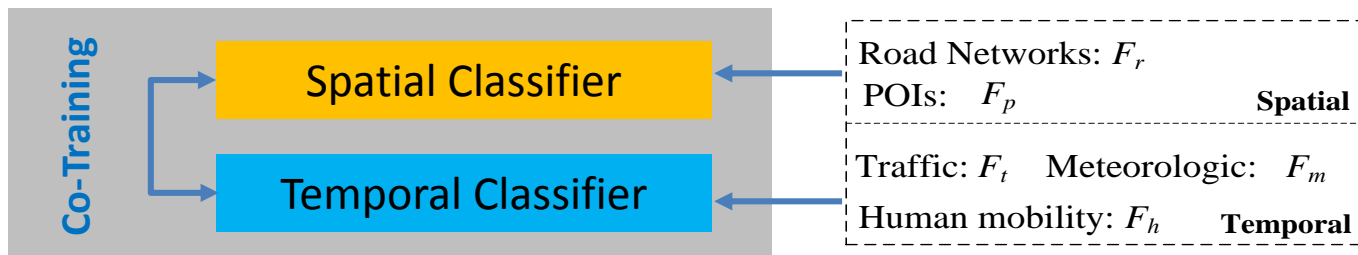
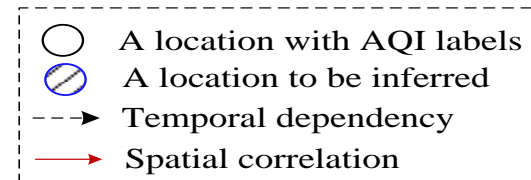
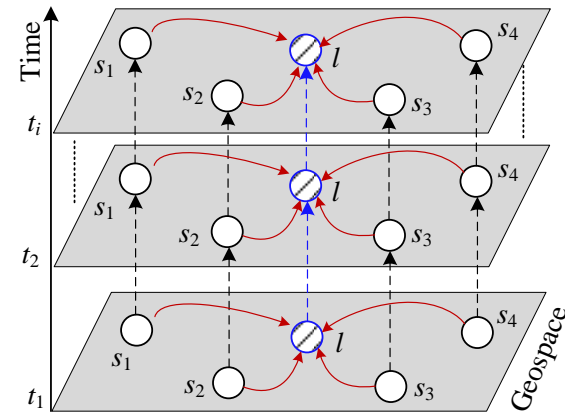
- Co-training-based semi-supervised learning model for each pollutant

- Predict the AQI labels
- Data sparsity
- Two classifiers

| AQI | Values Levels of Health Concern | Colors |
|---------|--------------------------------------|--------|
| 0-50 | Good (G) | Green |
| 51-100 | Moderate (M) | Yellow |
| 101-150 | Unhealthy for sensitive groups (U-S) | Orange |
| 151-200 | Unhealthy (U) | Red |
| 201-300 | Very unhealthy (VU) | Purple |
| 301-500 | Hazardous (H) | Maroon |

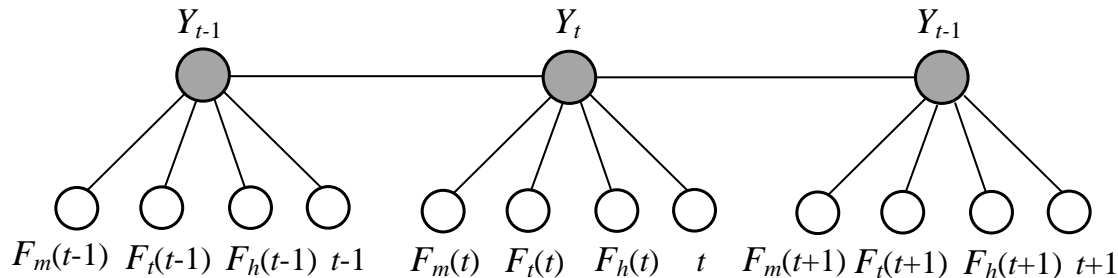
Semi-Supervised Learning Model

- Philosophy of the model
 - States of air quality
 - Temporal dependency in a location
 - Geo-correlation between locations
 - Generation of air pollutants
 - Emission from a location
 - Propagation among locations
 - Two sets of features
 - Spatially-related
 - Temporally-related



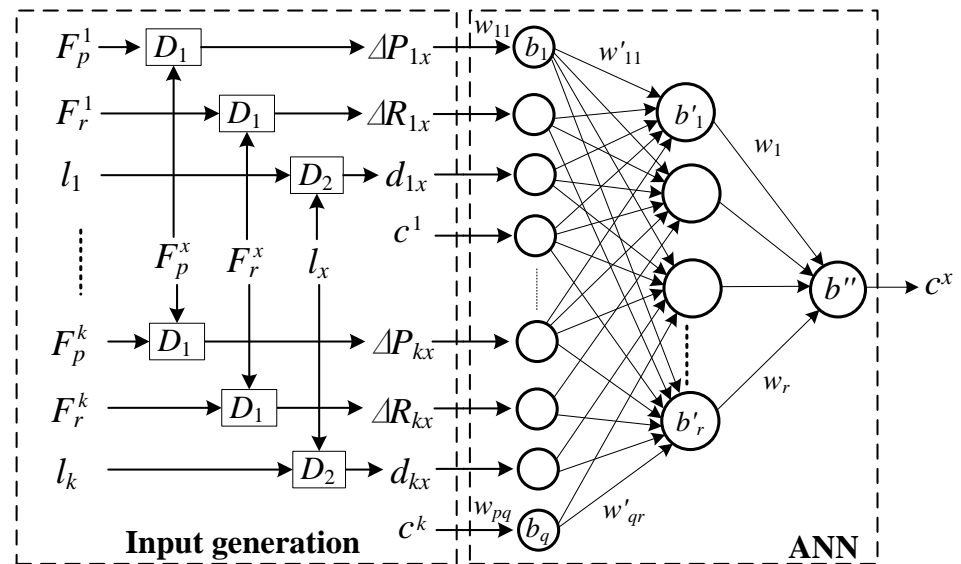
Semi-Supervised Learning Model

- Temporal classifier (TC)
 - Model the temporal dependency of the air quality in a location
 - Using temporally related features
 - Based on a Linear-Chain Conditional Random Field (CRF)

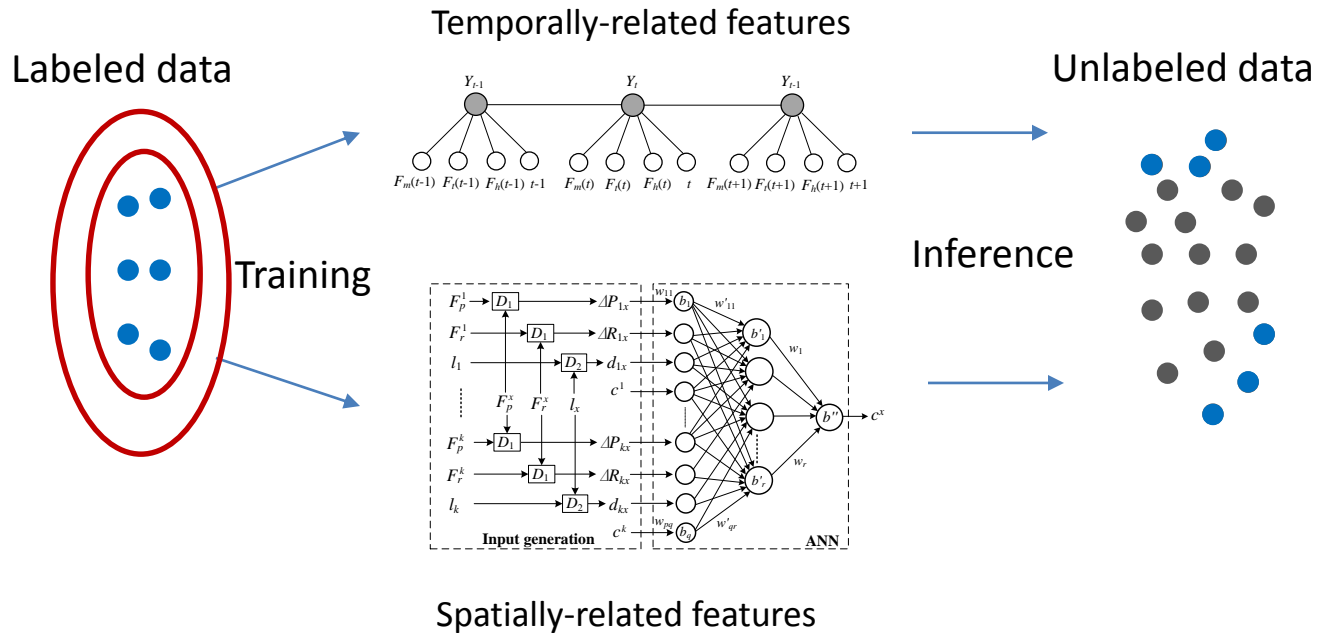


Semi-Supervised Learning Model

- Spatial classifier (SC)
 - Model the spatial correlation between AQI of different locations
 - Using spatially-related features
 - Based on a BP neural network
- Input generation
 - Select n stations to pair with
 - Perform m rounds

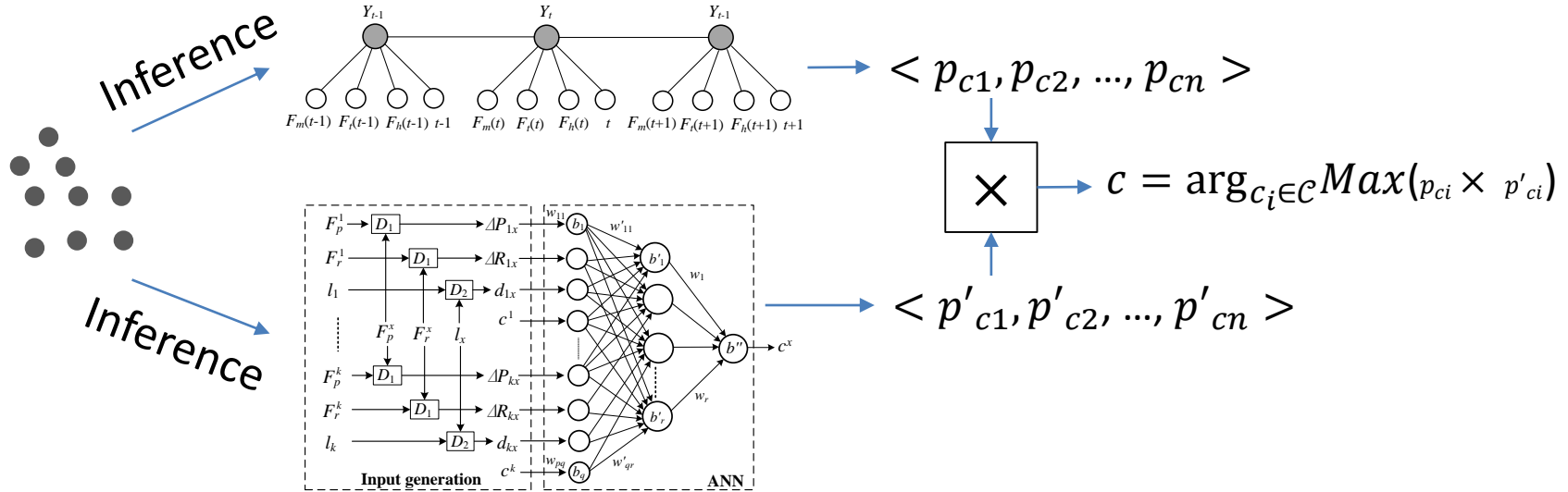


Learning Process of Our Model



Inference Process

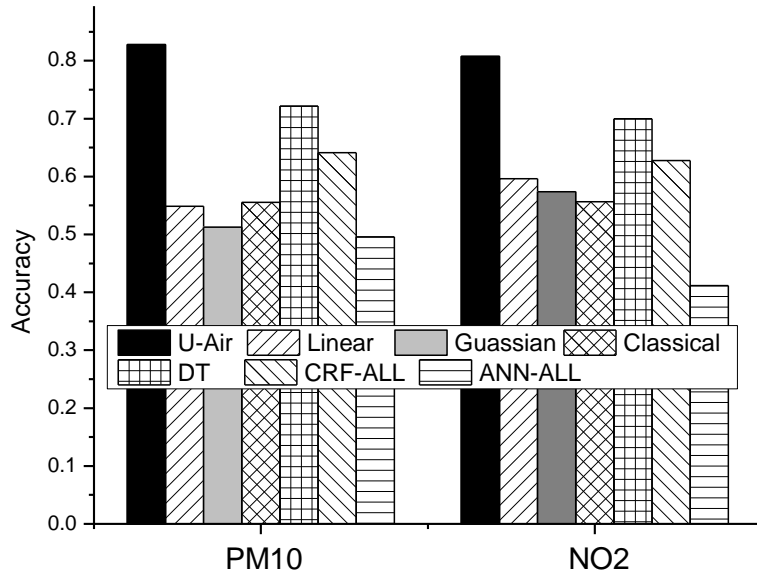
Temporally-related features



Spatially-related features

Evaluation

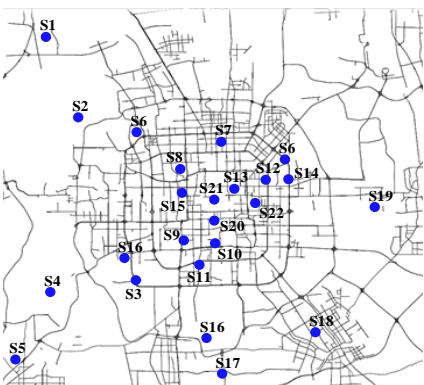
- Overall performance



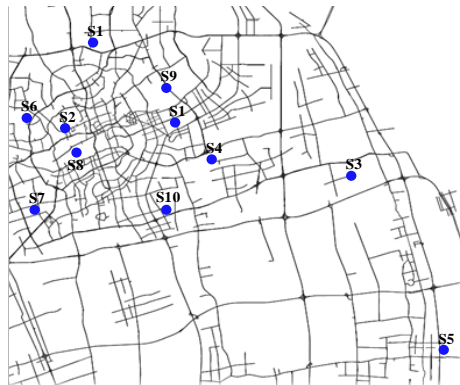
Yu Zheng, et al.

U-Air: when urban air quality inference meets big data.

KDD 2013



A) Beijing



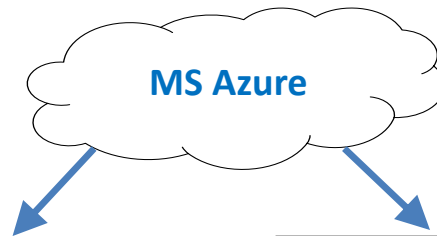
B) Shanghai



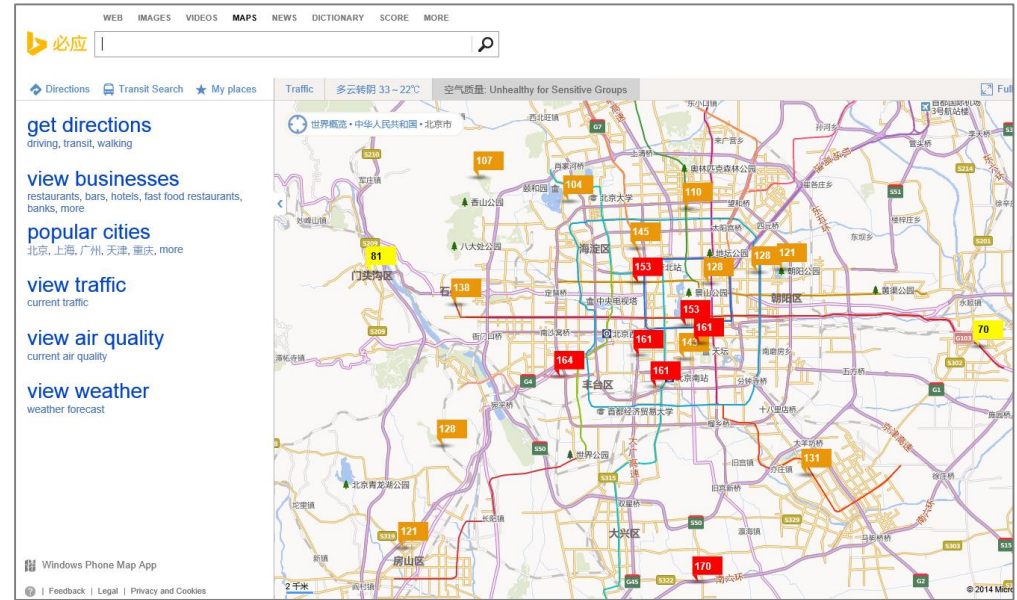
C) Shenzhen



D) Wuhan



Cloud + Client



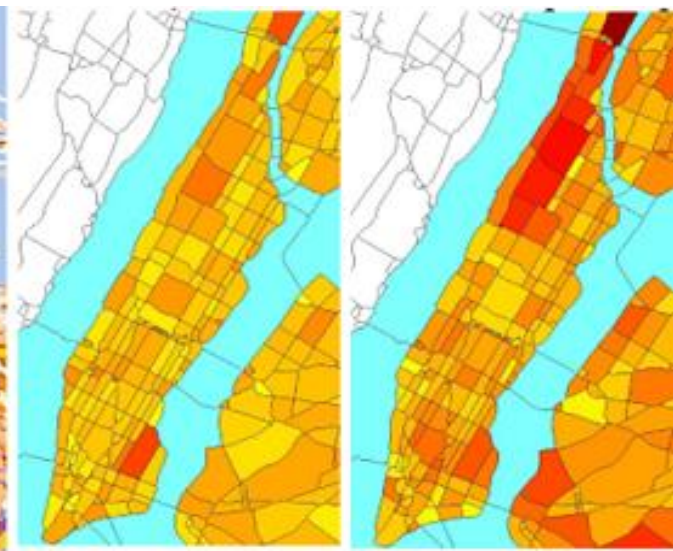
Urban Air

Bing Map

- Transferred to CityNext and Bing Map China
- Working with Chinese Ministry of Environmental Protection
- Forecasting air quality in the near future
- To identify the root cause of the air pollution

Diagnosing Urban Noises using Big Data

UbiComp 2014



Background

- Many cities suffer from noise pollutions
 - Traffic, loud music, construction, AC...
 - Compromise working efficiency
 - Reduce sleep quality
 - Impair both physical and mental health
 - ...
- Urban noise is difficult to model
 - Change over time very quickly
 - Vary by location significantly
 - Depends on sound levels and people's tolerance
 - The composition of noises is hard to analyze



Loud Music/Party 25.5
Loud Talking 13
Vehicle 12
Construction 10.7
AC/Ventilation 5.7



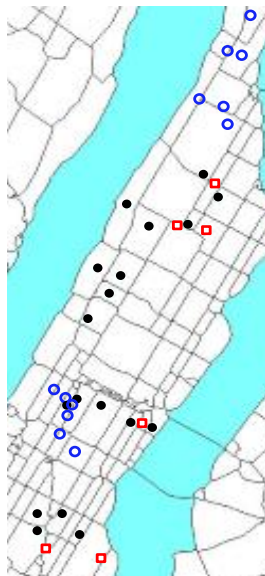
311 in NYC

- 311 Data
 - A platform for citizen's non-emergent complaints
 - Associated with a location, timestamp, and a category
 - Human as a sensor → crowd sensing
 - Implies people's reaction and tolerance to noises

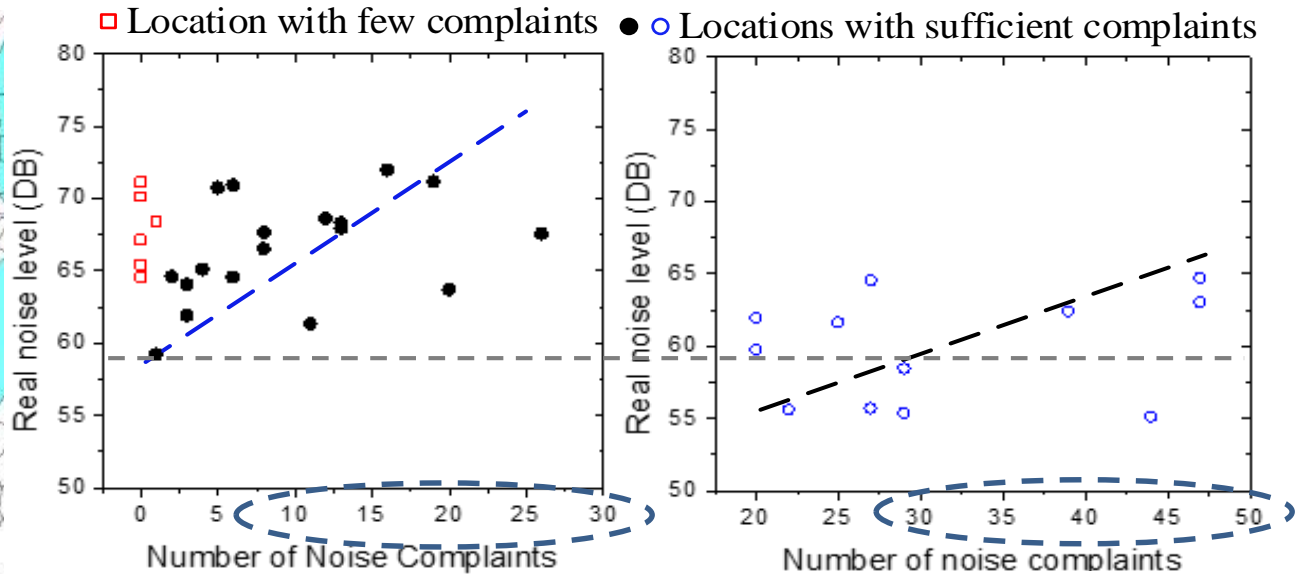


311 in NYC

- Correlation between 311 complaints and real noise levels
 - Measured the real noise levels of 36 locations in Manhattan
 - People's tolerances vary in time of day



A) Locations



B) Correlation in 6am-6pm

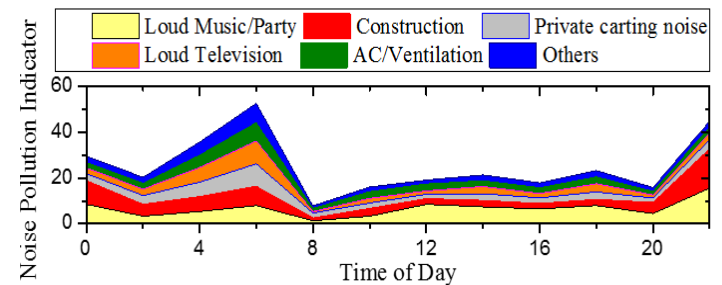
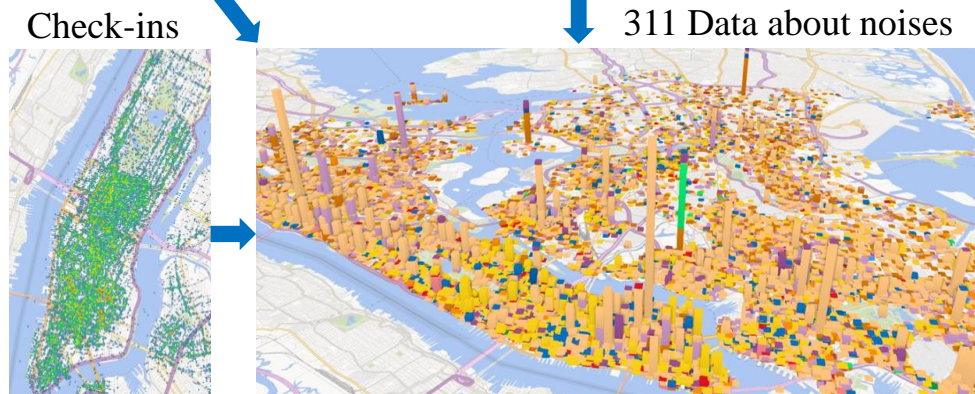
24 locations

C) Correlation in 7pm-11pm

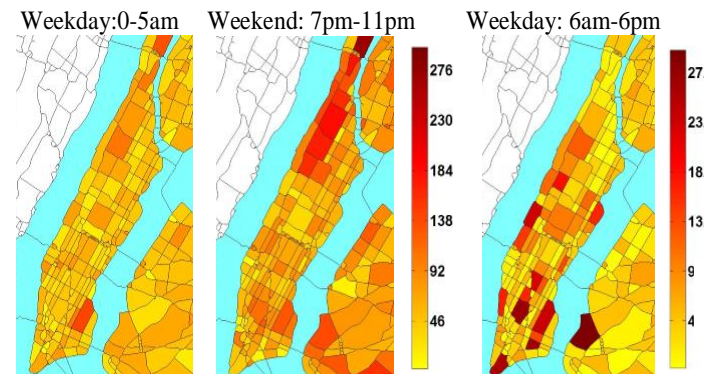
12 locations

Goal

- Reveal the noise situation of each region in each hour
 - A noise indicator denoting the noisy level
 - Composition of noises in each location



C) Noise of different categories in Time Square

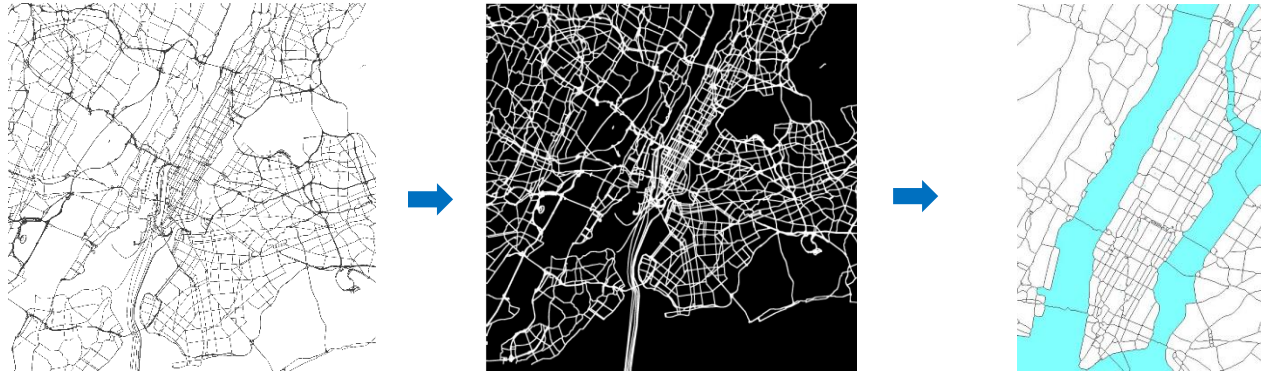


A) Overall noises

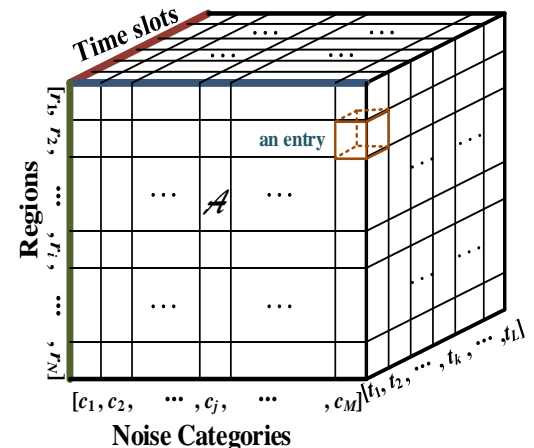
B) Construction

Methodology

- Partition NYC into regions by major roads

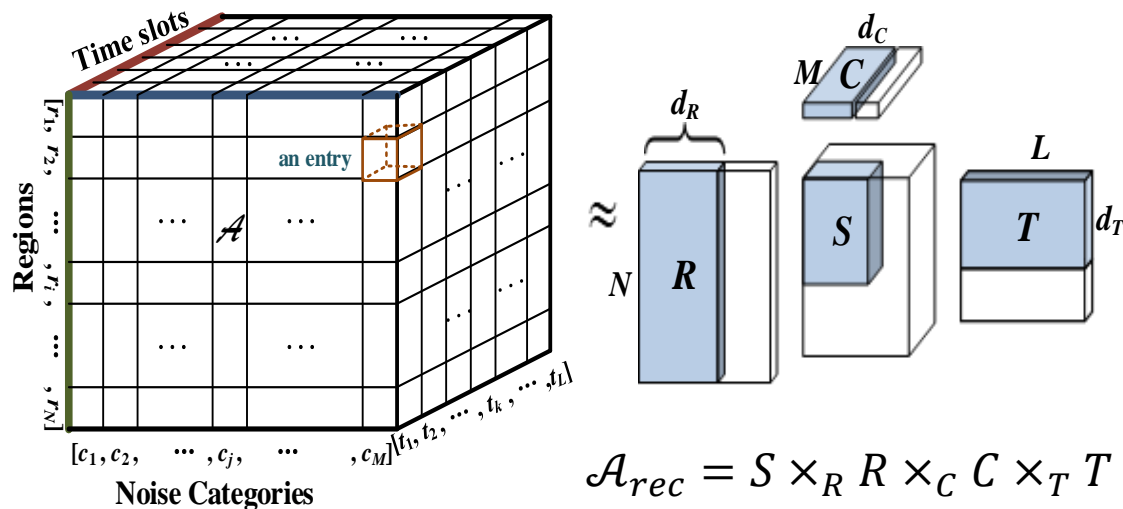


- Build a 3D tensor to model the noises
 - Region
 - Time slot
 - Noise categories
- Supplement the missing entries through
 - A context-aware tensor decomposition
 - In collaborative filtering



Methodology

- Simple idea: Tensor Decomposition
- Not very accurate
- Need more inputs from other sources



$$\mathcal{L}(S, R, C, T) = \frac{1}{2} \|\mathcal{A} - S \times_R R \times_C C \times_T T\|^2 + \frac{\lambda}{2} (\|S\|^2 + \|R\|^2 + \|C\|^2 + \|T\|^2)$$

Methodology

- Geographical Features

- Road networks

- Number of intersections f_s
- Length of road segments in different levels f_r

- POIs

- Total number of POIs f_n and density of POIs f_d
- Distribution over different categories f_c

$$X = \begin{matrix} & f_s & f_r & f_n & f_d & f_c \\ \begin{matrix} r_1 \\ r_2 \\ \vdots \\ r_i \\ \vdots \\ r_N \end{matrix} & & \mathbf{F}_r & & & \\ & & & \mathbf{F}_p & & \end{matrix}$$



A) Loud talking



B) POI: Food



C) Loud Music

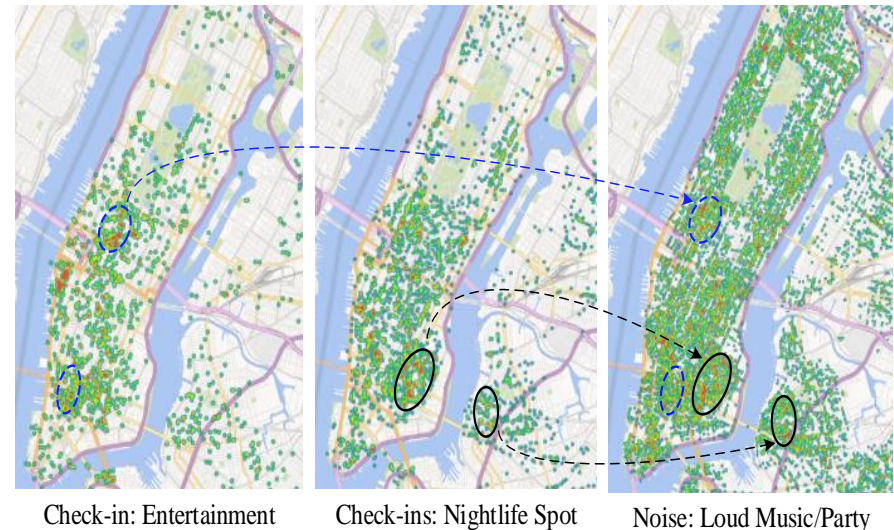
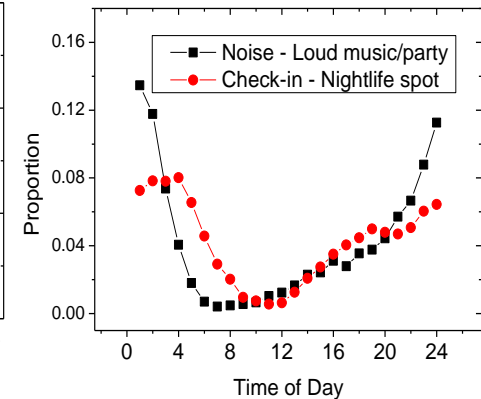
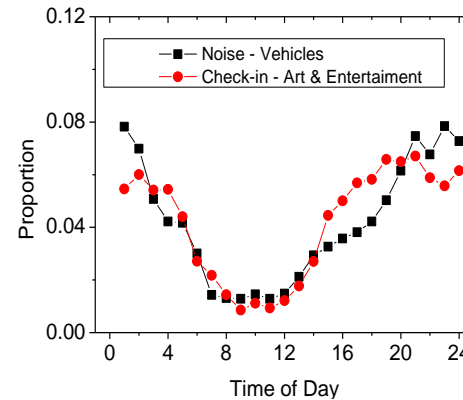


D) POI: Entertainment

Methodology

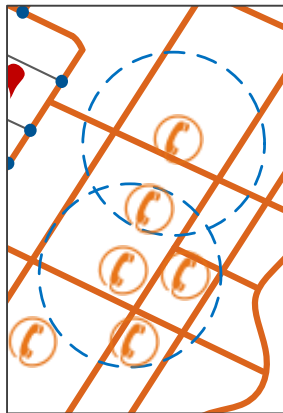
- Check in data in NYC
 - Gowalla: 127,558 check-ins (4/24/2009 to 10/13/2013)
 - Foursquare: 173,275 check-ins (5/5/2008 to 7/23/2011)
- Correlation with noises
- Y implies
 - correlation between different time slots
 - correlation between different regions

$$Y = \begin{matrix} t_1 \\ t_2 \\ \vdots \\ t_k \\ \vdots \\ t_L \end{matrix} \begin{bmatrix} r_1 & r_2 & \dots & r_i & \dots & r_N \\ d_{11} & \dots & \dots & \dots & \dots & \dots \\ \vdots & \ddots & & d_{ki} & & \vdots \\ \vdots & & & \vdots & \ddots & \vdots \\ \vdots & & & \vdots & & d_{LN} \end{bmatrix}$$

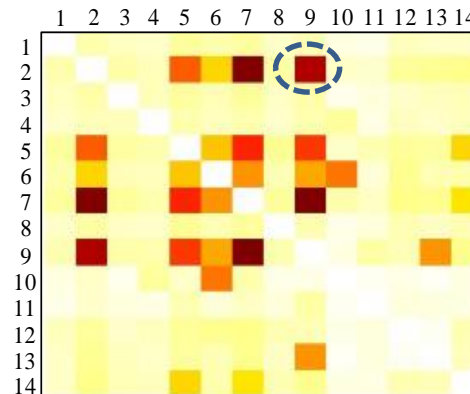
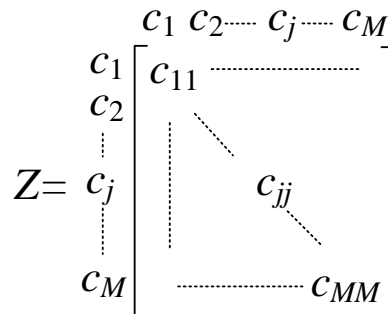


Methodology

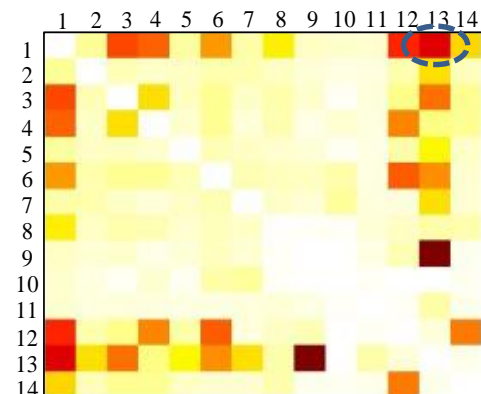
- Correlation between different noise categories



| Categories | % | Categories | % |
|--------------------------|------|--------------------------------|-----|
| c_1 . Loud Music/Party | 42.2 | c_8 . Alarms | 1.7 |
| c_2 . Construction | 17.2 | c_9 . Private carting noise | 0.8 |
| c_3 . Loud Talking | 14.6 | c_{10} . Manufacturing | 0.3 |
| c_4 . Vehicle | 13.7 | c_{11} . Lawn care equipment | 0.3 |
| c_5 . AC/Ventilation | 3.9 | c_{12} . Horn Honking | 0.2 |
| c_6 . Banging/Pounding | 2.1 | c_{13} . Loud Television | 0.1 |
| c_7 . Jack Hammering | 2.1 | c_{14} . Others | 0.8 |



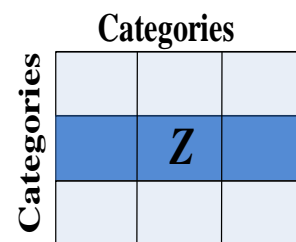
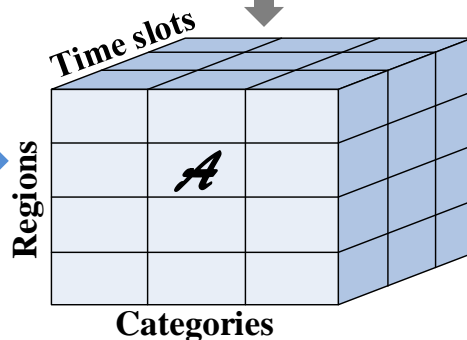
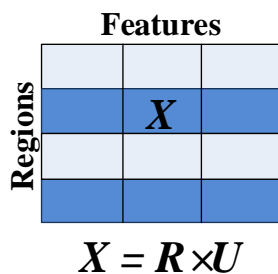
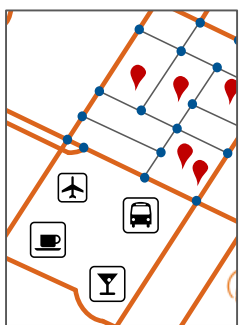
A) weekday



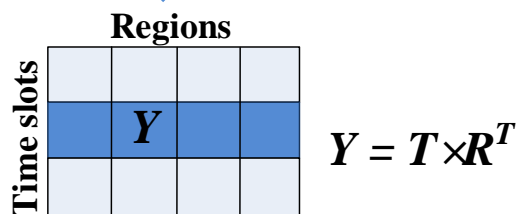
B) weekend

Methodology

POIs and Road Networks



Check-ins

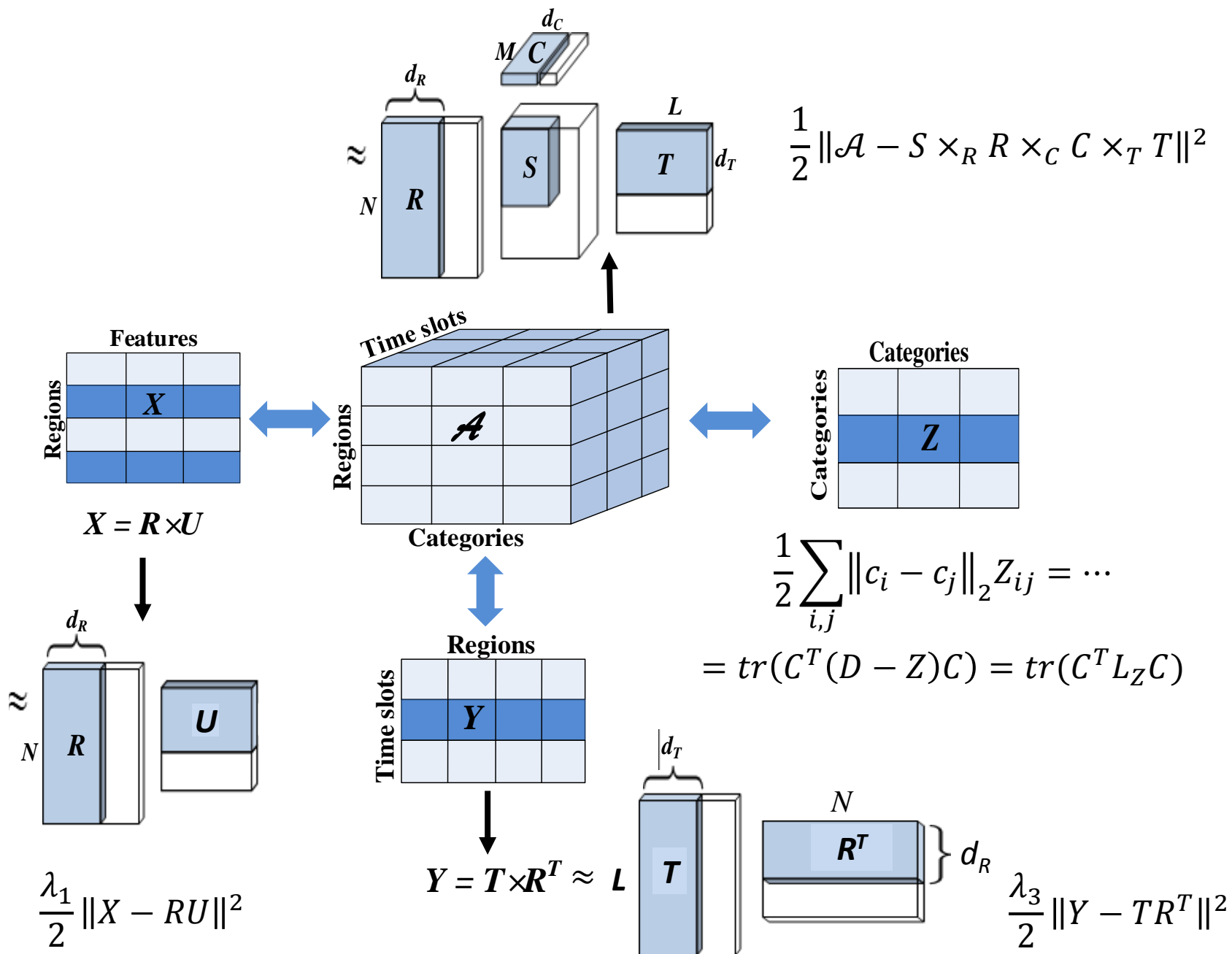


311 complaints about noises



$$\mathcal{L}(S, R, C, T, U) = \frac{1}{2} \|\mathcal{A} - S \times_R R \times_C C \times_T T\|^2 + \frac{\lambda_1}{2} \|X - RU\|^2 + \frac{\lambda_2}{2} \text{tr}(C^T L_Z C) + \frac{\lambda_3}{2} \|Y - TR^T\|^2 + \frac{\lambda_4}{2} (\|S\|^2 + \|R\|^2 + \|C\|^2 + \|T\|^2 + \|U\|^2)$$

$$\mathcal{L}(S, R, C, T, U) = \frac{1}{2} \|\mathcal{A} - S \times_R R \times_C C \times_T T\|^2 + \frac{\lambda_1}{2} \|X - RU\|^2 + \frac{\lambda_2}{2} \text{tr}(C^T L_Z C) + \frac{\lambda_3}{2} \|Y - TR^T\|^2 + \frac{\lambda_4}{2} (\|S\|^2 + \|R\|^2 + \|C\|^2 + \|T\|^2 + \|U\|^2)$$



<http://citynoise.azurewebsites.net/>

New York City's Noise

Data Panel

Weekday ▾

All Categories ▾

Noise Layer

Time Panel

0:00 ~ 0:00

Overall Time

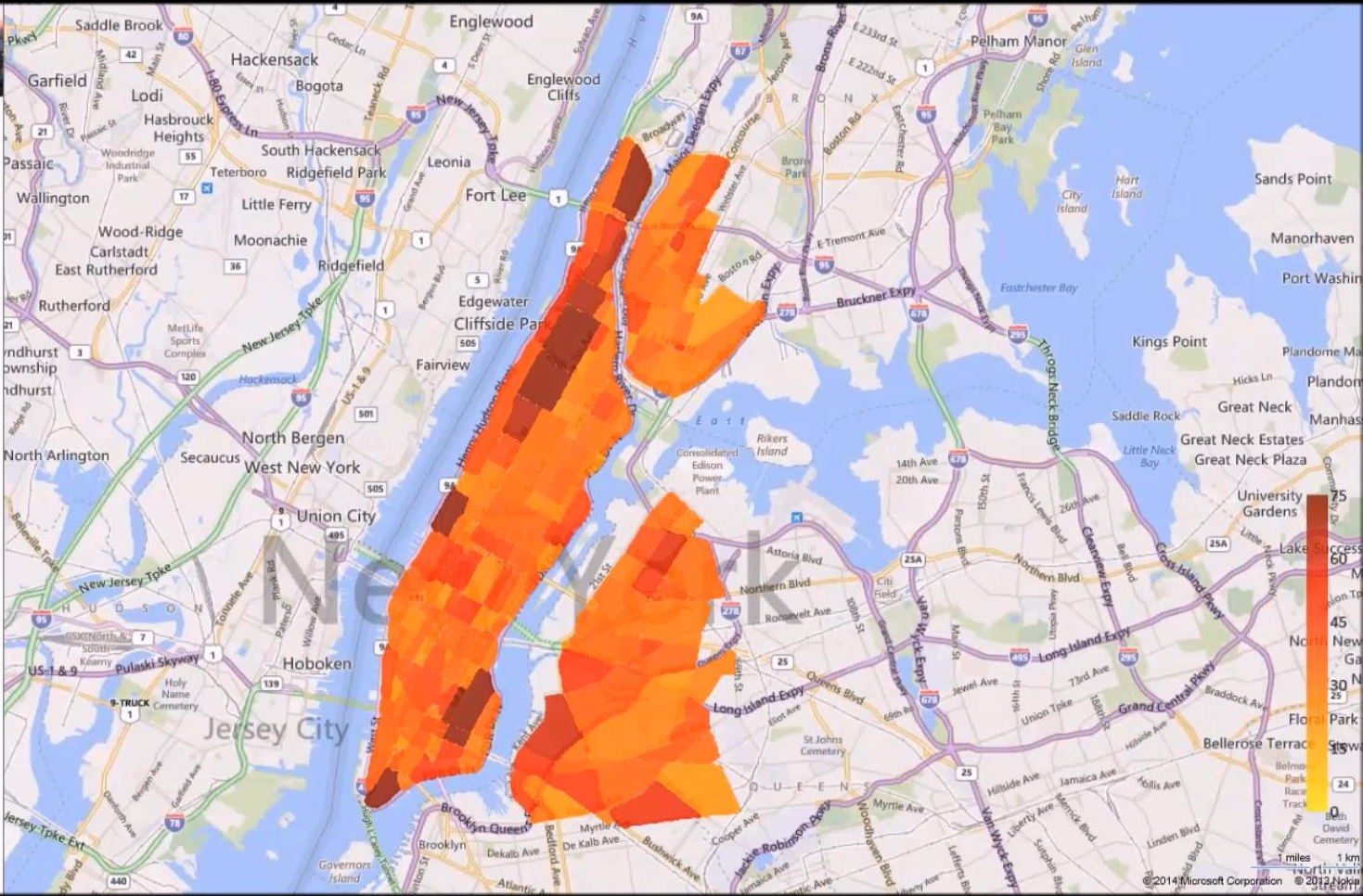
Top 5 Noisiest Regions

| | |
|---|---|
| #1 Audubon Ave: 135.01 | → |
| #2 East Village & Stuyvesant Town: 105.07 | → |
| #3 Columbia University: 98.16 | → |
| #4 Wall Street: 97.63 | → |
| #5 E Houston St & 1st Ave: 89.06 | → |

Region Noise Analysis

0 2 4 6 8 10 12 14 16 18 20 22

This demo is based on 311 data from May 23, 2013 to Jan. 31, 2014.



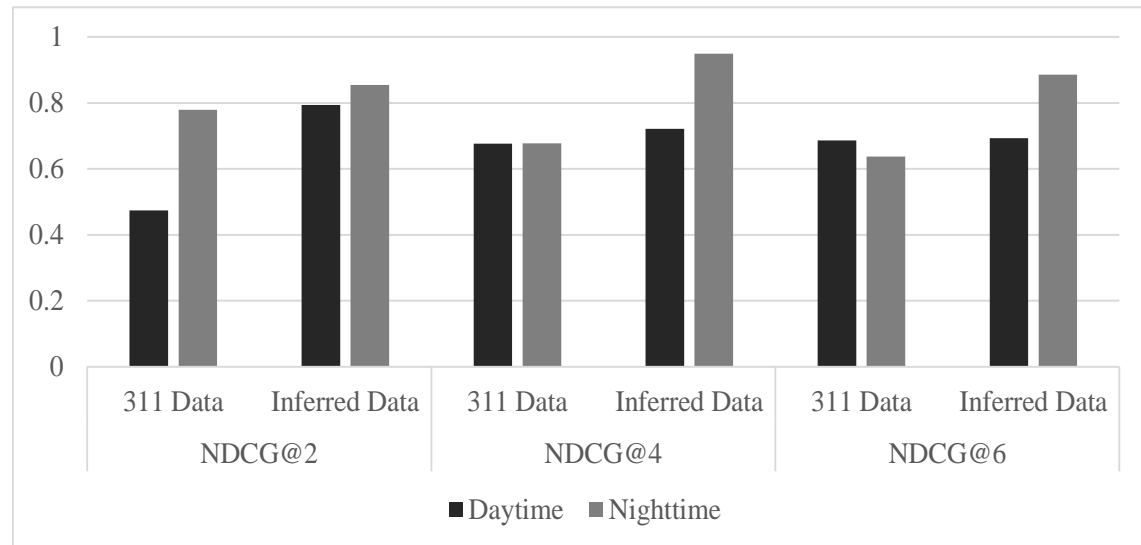
Experiments

- Accuracy of the inferences
 - Remove 30% non-zero entries
 - Metrics: RMSE & MAE
 - Compared with six Baseline methods

| Methods | Weekdays | | Weekends | |
|-----------------------------------|--------------|--------------|--------------|--------------|
| | RMSE | MAE | RMSE | MAE |
| AWR | 4.736 | 2.582 | 4.446 | 2.599 |
| AWH | 4.631 | 2.461 | 4.42 | 2.522 |
| MF | 4.600 | 2.474 | 4.393 | 2.516 |
| Kriging | 4.59 | 2.424 | 4.253 | 2.495 |
| TD | 4.391 | 2.381 | 4.141 | 2.393 |
| TD+ X | 4.285 | 2.279 | 4.155 | 2.326 |
| TD+ $X + Y$ | 4.160 | 2.110 | 4.003 | 2.198 |
| TD+ $X + Y + Z$ | 4.010 | 2.013 | 3.930 | 2.072 |

Experiments

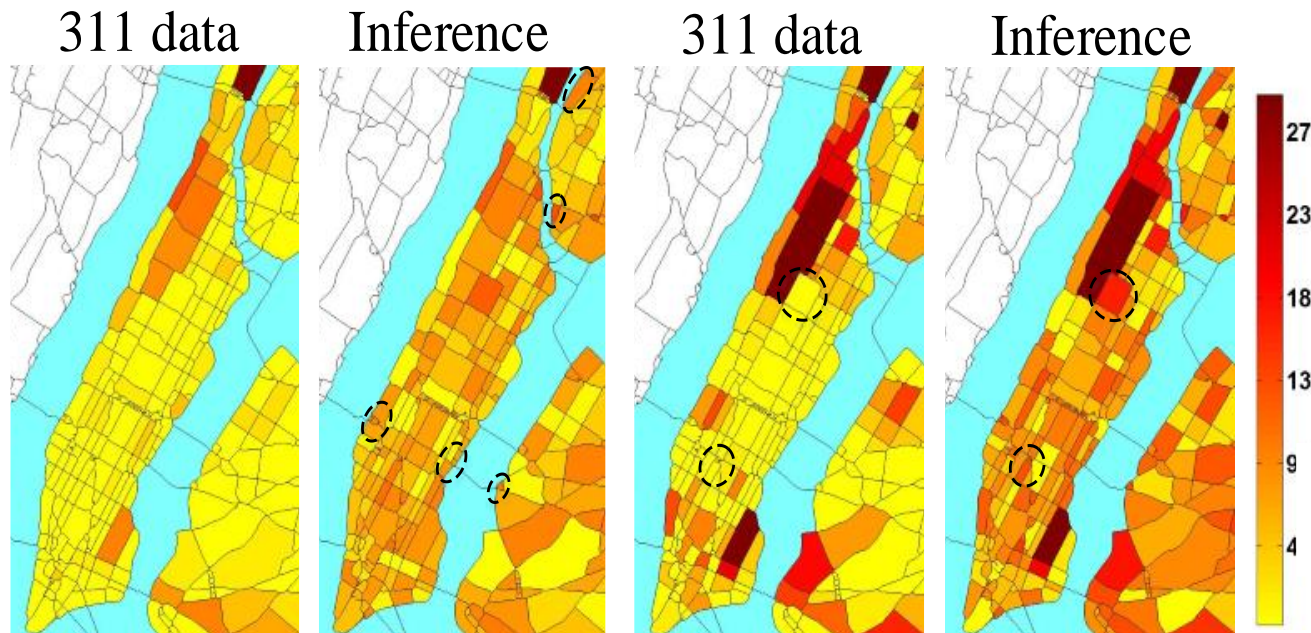
- Relative ranking performance
 - Ranked by the inferred noise indicators
 - Ground truth: measured by a mobile phone running a client program
 - Metric: NDCG



Experiments

- 311 vs. Inferences

- 2.75% of entries on weekdays are from 311 data (97.25% by inference)
- 1.83% of entries on weekends are from 311 data (98.17% by inference)

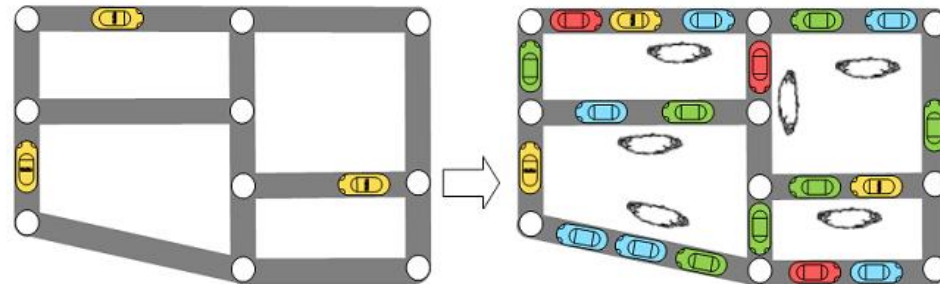


A) Vehicles (6am-6pm)

B) Loud Talking (0-5am)

Inferring Gas Consumption and Pollution Emission of Vehicles throughout a City

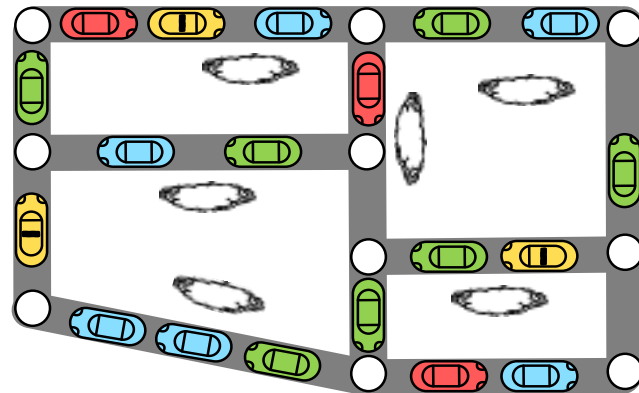
KDD 2014



Questions

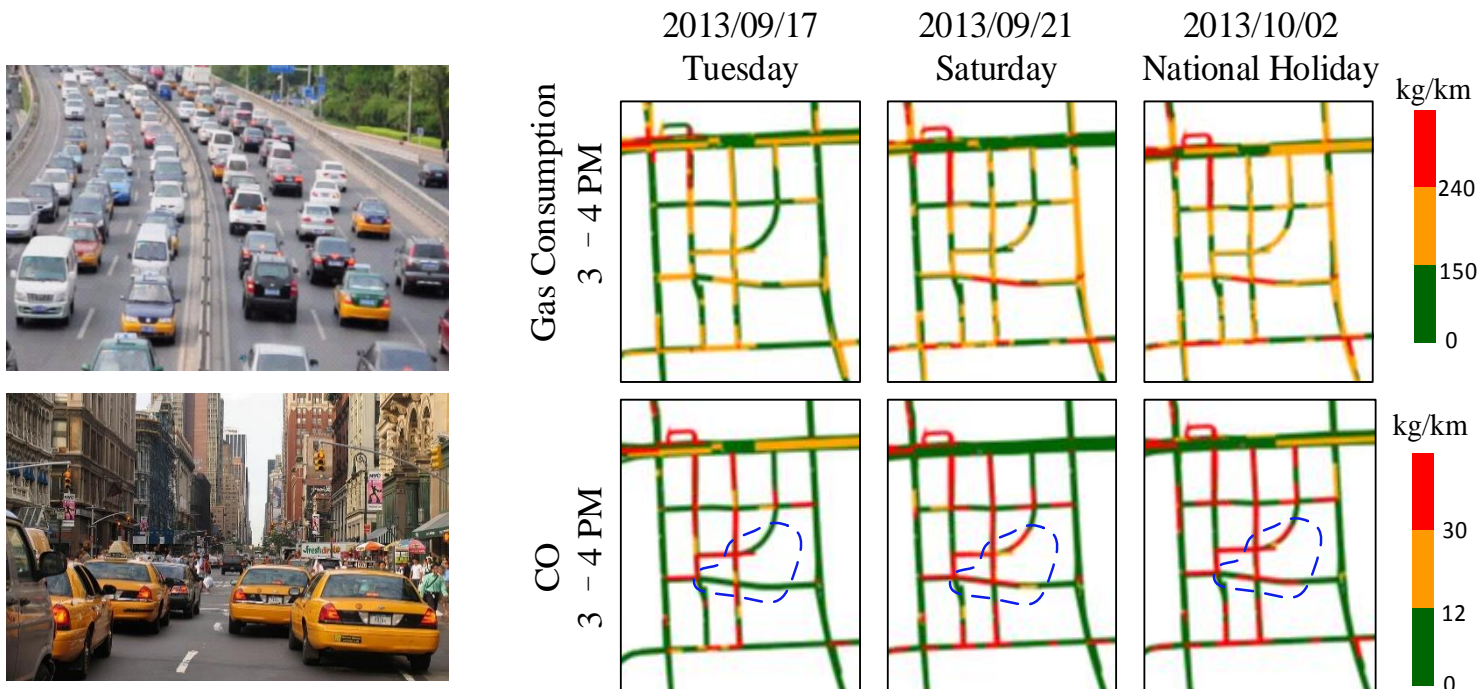
How many liters of gas have been consumed by the vehicles, in the entire city, in the past one hour?

What is the volume of PM_{2.5} that has been generated accordingly?



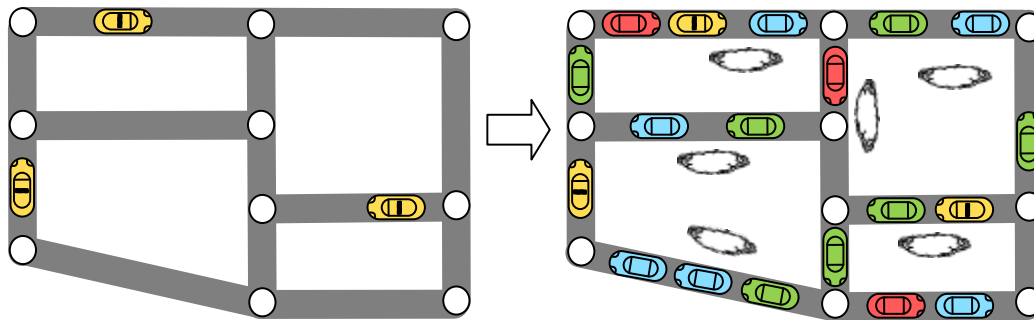
Goals

- Estimate the gas consumption and vehicle emissions
 - on arbitrary road segment
 - at any time intervals
 - using GPS trajectories of a sample of vehicles



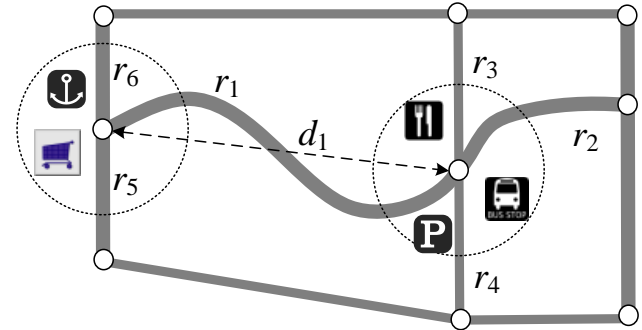
Our Approach

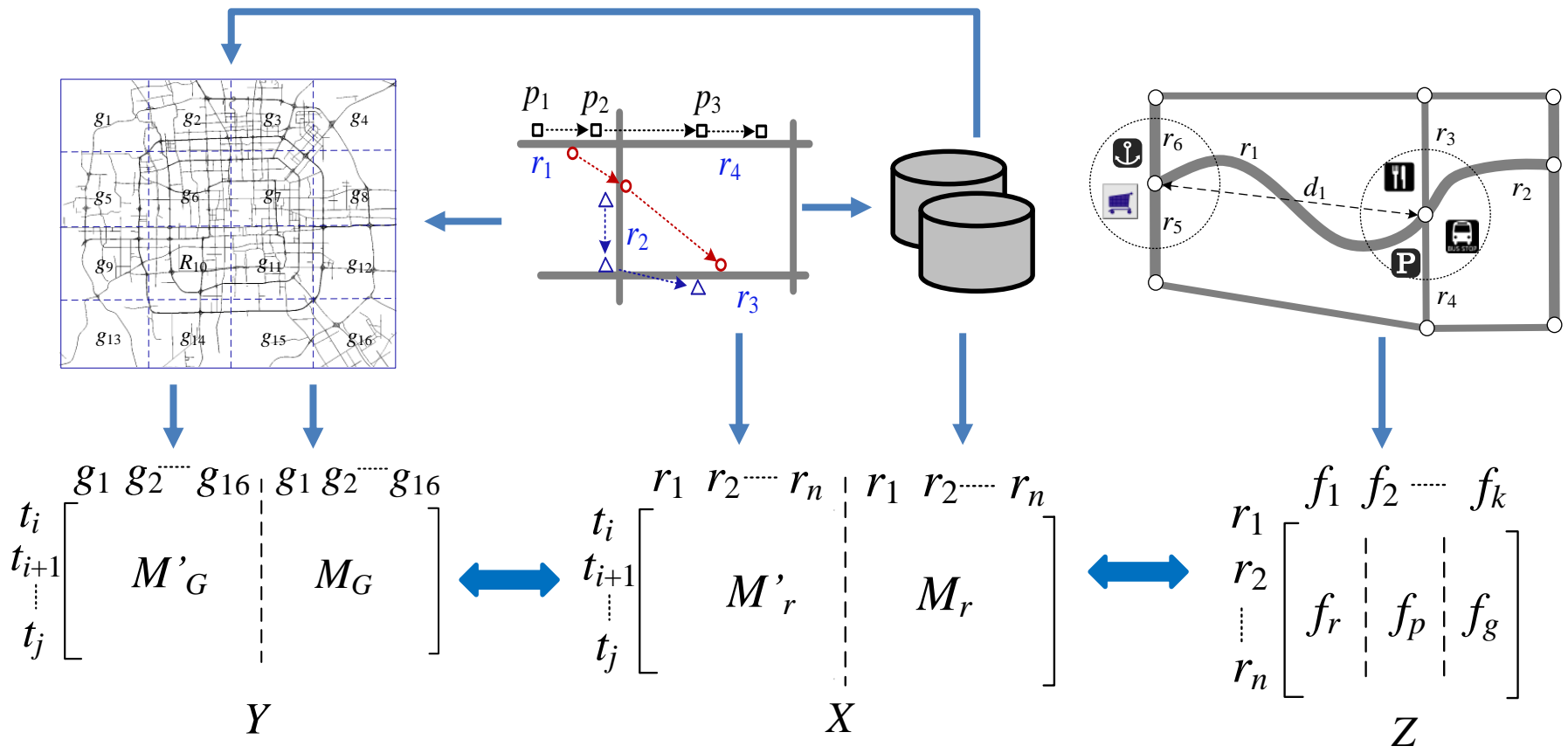
- Using the GPS trajectories of a sample of vehicles
 - Estimate travel speed on each road segment
 - Infer the traffic volume of each road segment
 - Calculate the gas consumption and emission of vehicles



Difficulties

- Data sparsity
- From speed to volume
 - Depends on multiple factors, such as
 - the current travel speeds and density of vehicles
 - the length, shape, and capacity of a road
 - weather conditions
 - Insufficient training data
 - Biased distribution of the samples
- Real-time and citywide
 - Over 100,000 road segments to infer
 - Need to finish it in a few minutes





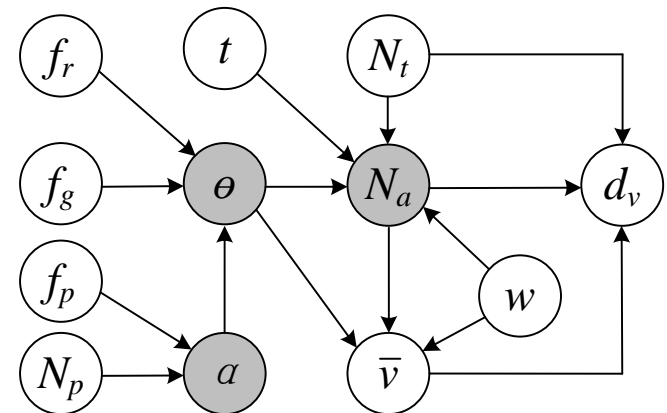
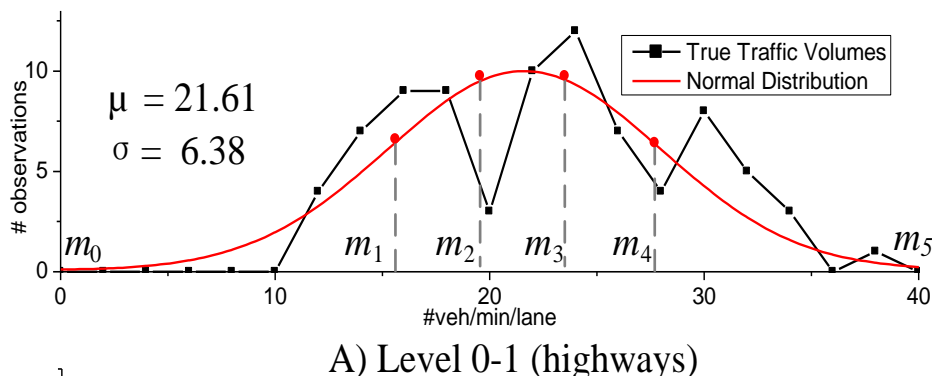
$$Y \approx T \times (G; G)^T; X \approx T \times (R; R)^T; Z \approx R \times F^T$$

- X : denotes fine-grained traffic conditions
- Y : denotes coarse-grained traffic conditions
- Z : geographical contexts of road segments

$$L(T, R, G, F) = \frac{1}{2} \|Y - T(G; G)^T\|^2 + \frac{\lambda_1}{2} \|X - T(R; R)^T\|^2 + \frac{\lambda_2}{2} \|Z - RF^T\|^2 + \frac{\lambda_3}{2} (\|T\|^2 + \|R\|^2 + \|G\|^2 + \|F\|^2),$$

Traffic Volume Inference (TVI)

- Objective: From travel speed to traffic volume
- Traffic volume depends on
 - the current travel speeds and density of vehicles
 - the length, shape, capacity of a road, and weather conditions
- Biased distribution between taxis and other vehicles
- Unsupervised learning approach



Energy and Emission Calculation

$$EF = (a + cv + ev^2)/(1 + bv + dv^2).$$

| | a | b | c | d | e |
|------------------|-----------------------|------------------------|------------------------|------------------------|-----------------------|
| CO | 71.7 | 35.4 | 11.4 | -0.248 | 0 |
| Hydrocarbon | 5.57×10^{-2} | 3.65×10^{-2} | -1.1×10^{-3} | -1.88×10^{-4} | 1.25×10^{-5} |
| Nox | 9.29×10^{-2} | -1.22×10^{-2} | -1.49×10^{-3} | 3.97×10^{-5} | 6.53×10^{-6} |
| Fuel Consumption | 217 | 9.6×10^{-2} | 0.253 | -4.21×10^{-4} | 9.65×10^{-3} |

$$E = EF \times r.N_a \times r.n \times r.len$$

Experiments

- Evaluation on TSE

| Methods | RMSE of \bar{v} | RMSE of dv | Time (sec) |
|-------------------------|-------------------|--------------|-------------|
| MF(M'_r) | 2.172 | 1.833 | 2.2 |
| MF($M'_r + Z$) | 1.939 | 1.385 | 18.2 |
| MF($M'_r + M'_g + Z$) | 1.908 | 1.314 | 20.2 |
| TSE | 1.369 | 1.035 | 22.2 |
| Kriging | 2.340 | 1.300 | 1,000 |
| KNN | 3.360 | 1.590 | 0.14 |

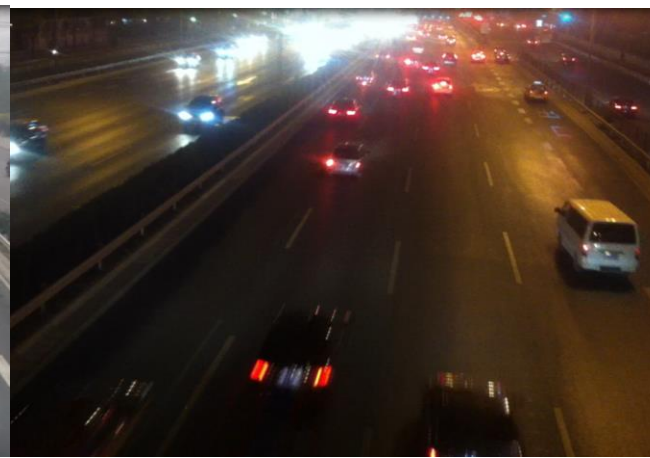
| RMSE | level 0-1 | level 2 | level 3 | level ≥ 4 |
|----------|-----------|---------|---------|----------------|
| Speed | 2.575 | 1.189 | 0.977 | 1.612 |
| Variance | 1.526 | 1.128 | 0.628 | 1.234 |

Experiments

- Evaluation on TVI

| Methods | MAE | MRE | Inference time (us/road) |
|--------------|------|------|--------------------------|
| TVI | 3.01 | 29% | 7.27 |
| TVI w/o dv | 3.19 | 31% | 7.18 |
| TVI w/o w | 3.15 | 29% | 7.10 |
| LR | 3.06 | 27% | 0.15 |
| FD | 2.66 | 16% | 0.13 |
| FD-SC | 3.9 | 42% | 0.13 |
| FD-DC | 6.7 | 137% | 0.13 |

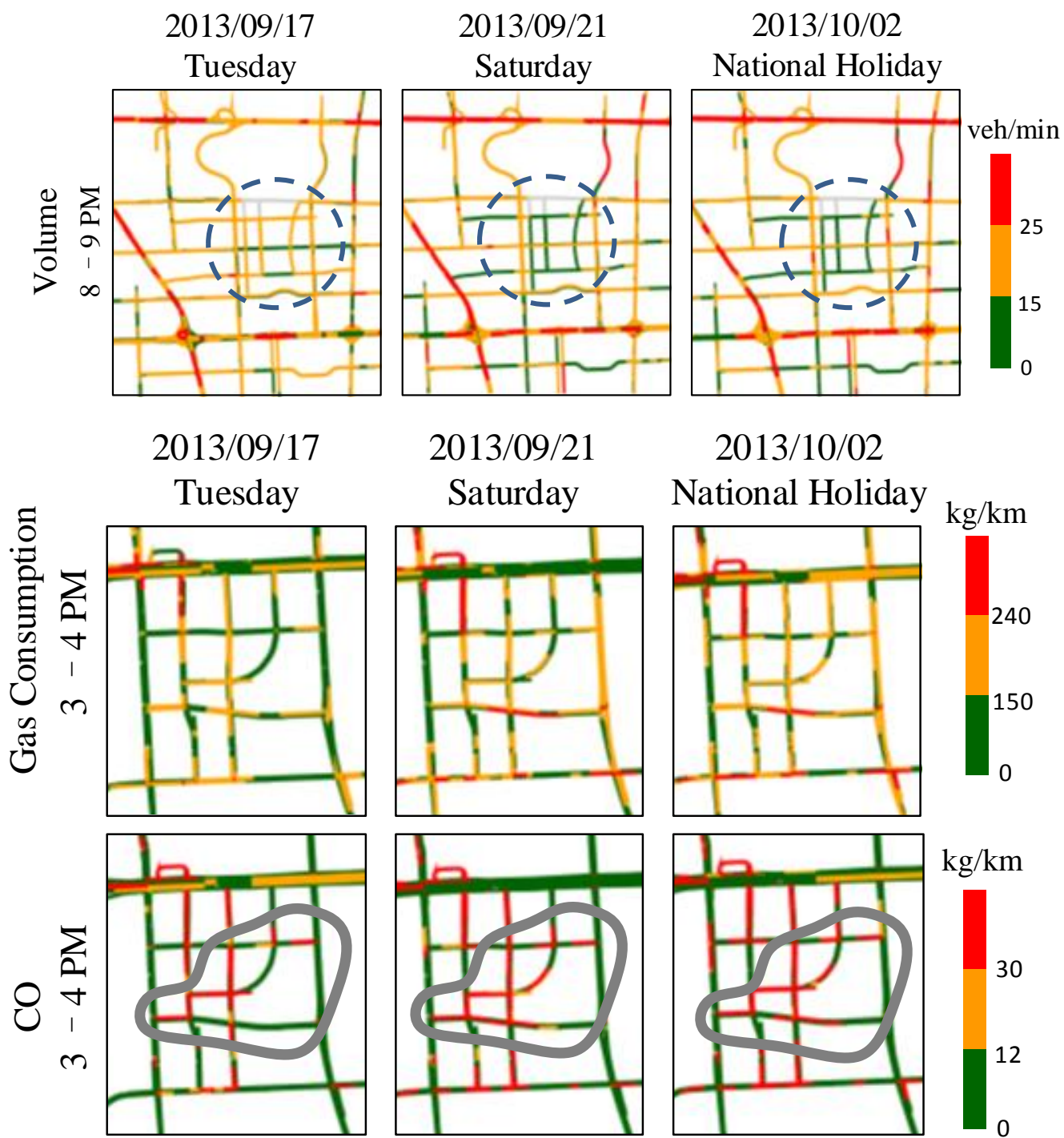
| | level 0-1 | level 2 | Weekday | Weekend |
|-----|-----------|---------|---------|---------|
| MAE | 5.55 | 2.23 | 2.97 | 3.28 |
| MRE | 22% | 41% | 29% | 30% |

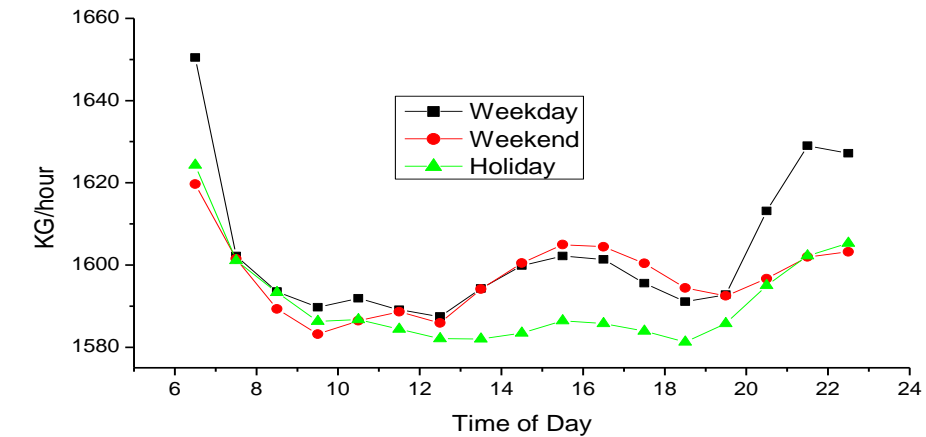
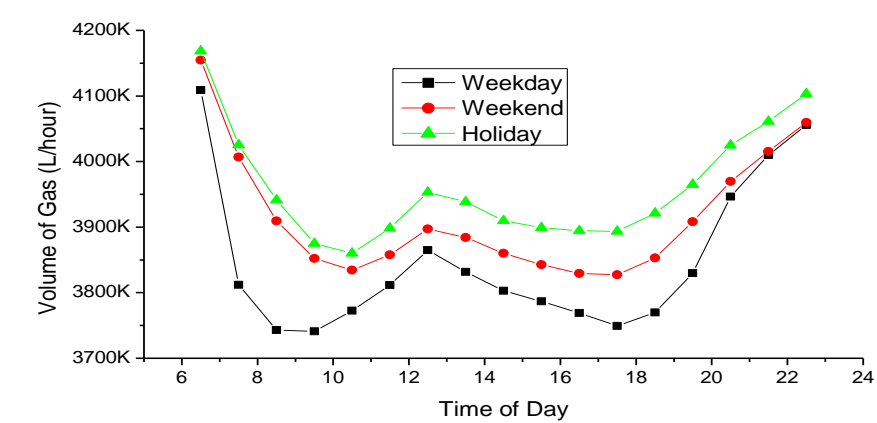
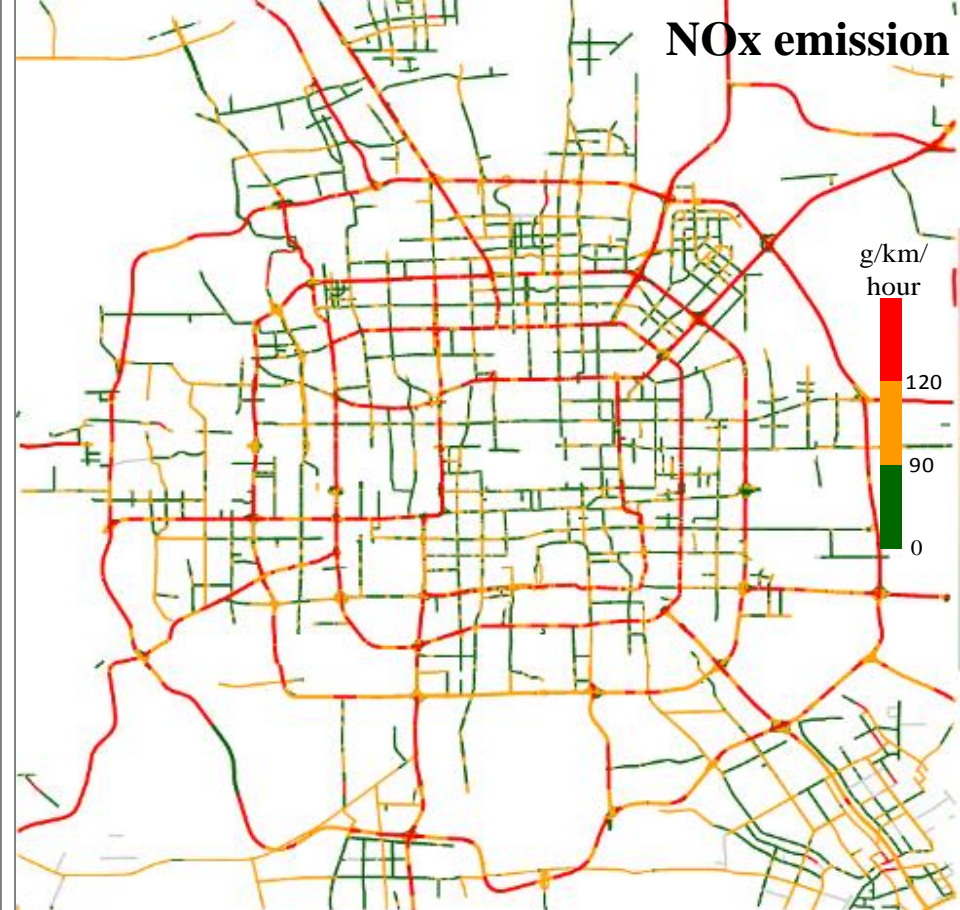
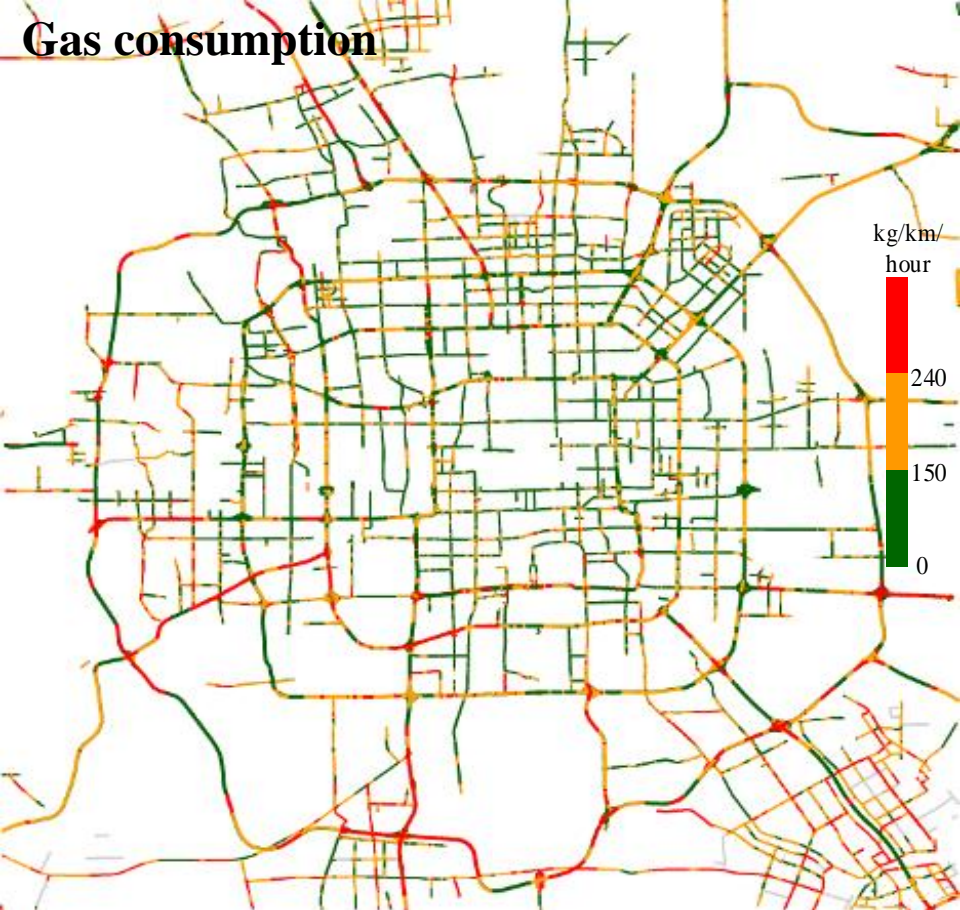


| Time | 7:00 ~ 10:00 | | | 10:00~16:00 | | | 16:00~20:00 | | | after 20:00 | | | total |
|---------|--------------|----|---|-------------|----|---|-------------|----|---|-------------|----|---|-------|
| Level | 0,1 | 2 | 3 | 0,1 | 2 | 3 | 0,1 | 2 | 3 | 0,1 | 2 | 3 | |
| Holiday | 0 | 0 | 0 | 6 | 14 | 4 | 6 | 8 | 1 | 4 | 6 | 0 | 49 |
| Workday | 7 | 28 | 8 | 29 | 74 | 9 | 28 | 92 | 7 | 6 | 17 | 4 | 309 |
| Total | 43 | | | 136 | | | 142 | | | 37 | | | 358 |

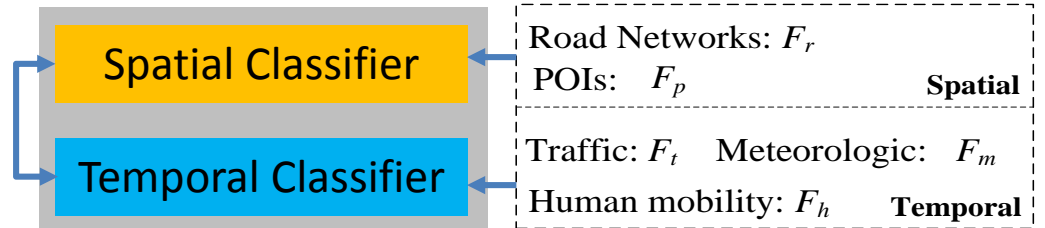
- Efficiency

| Online components | Time | Offline components | Time |
|-------------------|----------------|-------------------------------|-------------|
| Map-matching | 4.94min | Geo-feature extraction | 149s |
| TSE | 22.2s | Historical pattern extraction | 240s |
| TVI (inference) | 0.84s | TVI learning | 89s |
| Total | 5.32min | Total | 478s |

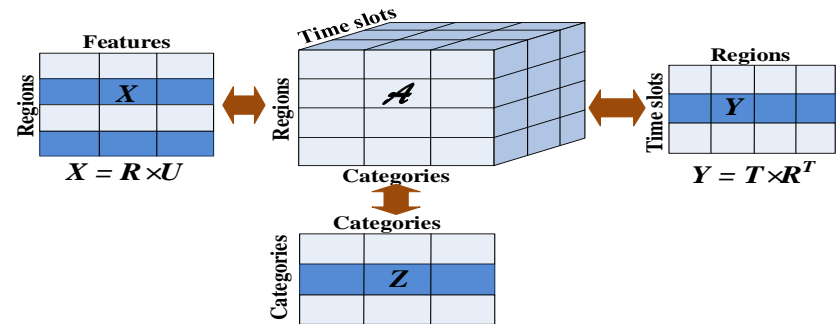




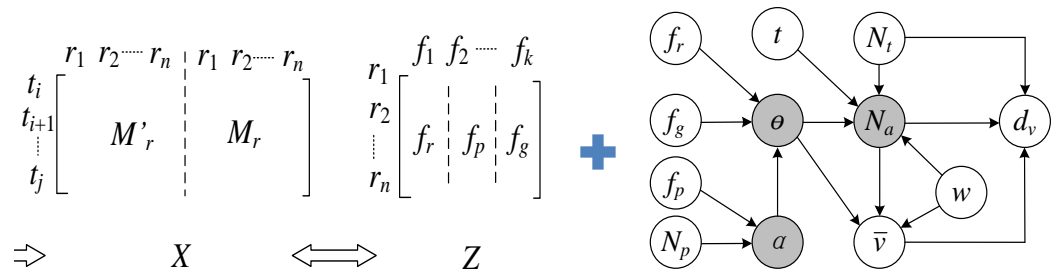
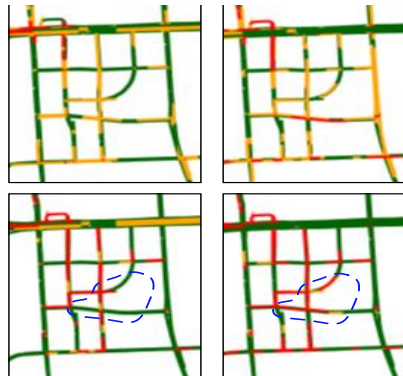
Computing with Multiple Heterogeneous Data Sources



Co-Training-based Semi-supervised learning



Context-aware Tensor Decomposition



Matrix Factorization + Graphical Models

Take Away Messages

- **3B**: *B*ig city, *B*ig challenges, *B*ig data
- **3M**: Data *M*anagement, *M*ining and *M*achine learning
- **3W**: *W*in-*W*in-*W*in: people, city, and the environment

3·BMW

Search for “[Urban Computing](#)”



[Download
Urban Air App](#)

Thanks!

Yu Zheng

yuzheng@microsoft.com



[Homepage](#)