Urban Computing

–Using Big Data to Solve Urban Challenges

Dr. Yu Zheng

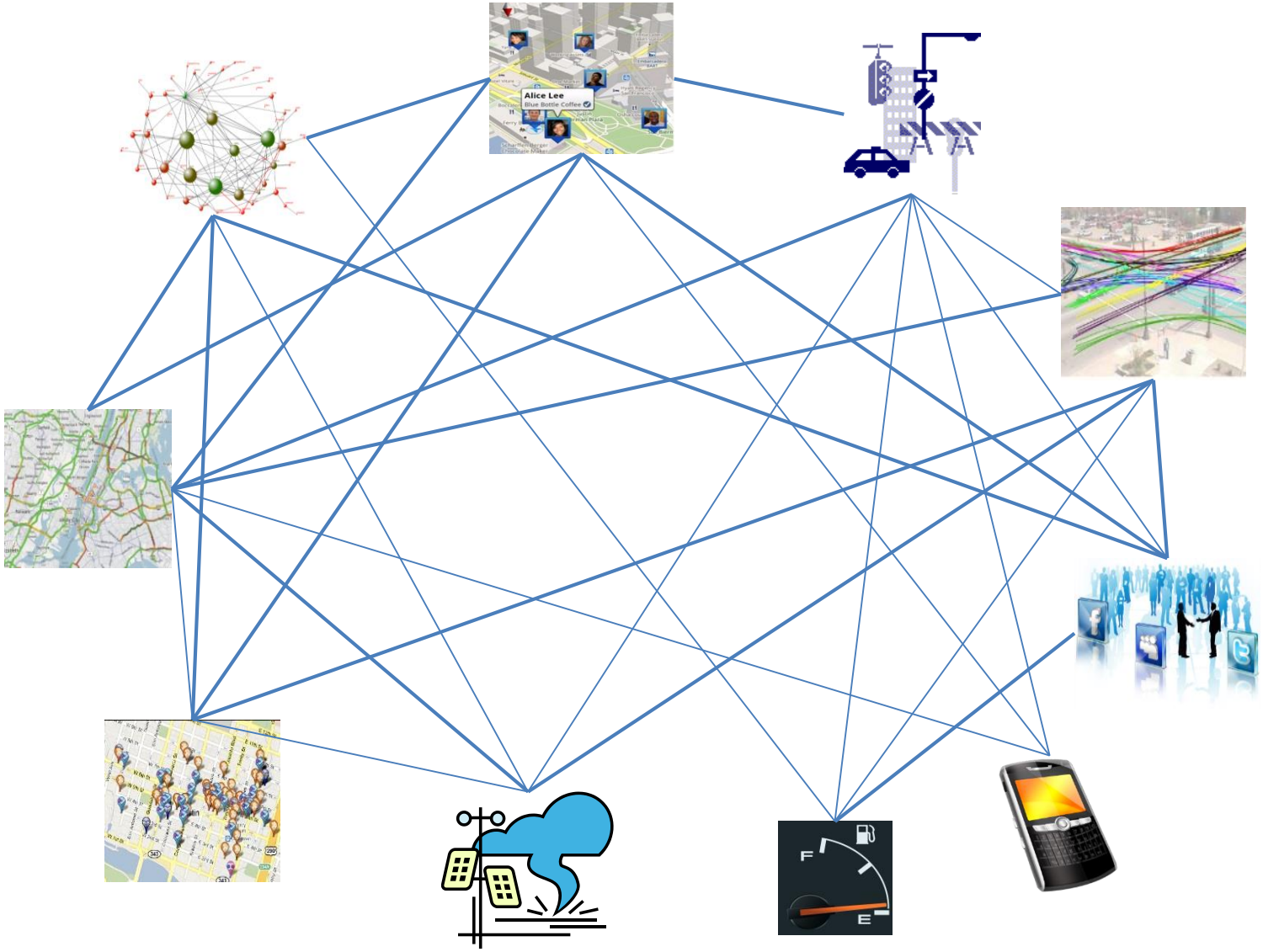
Lead Researcher, Microsoft Research Asia

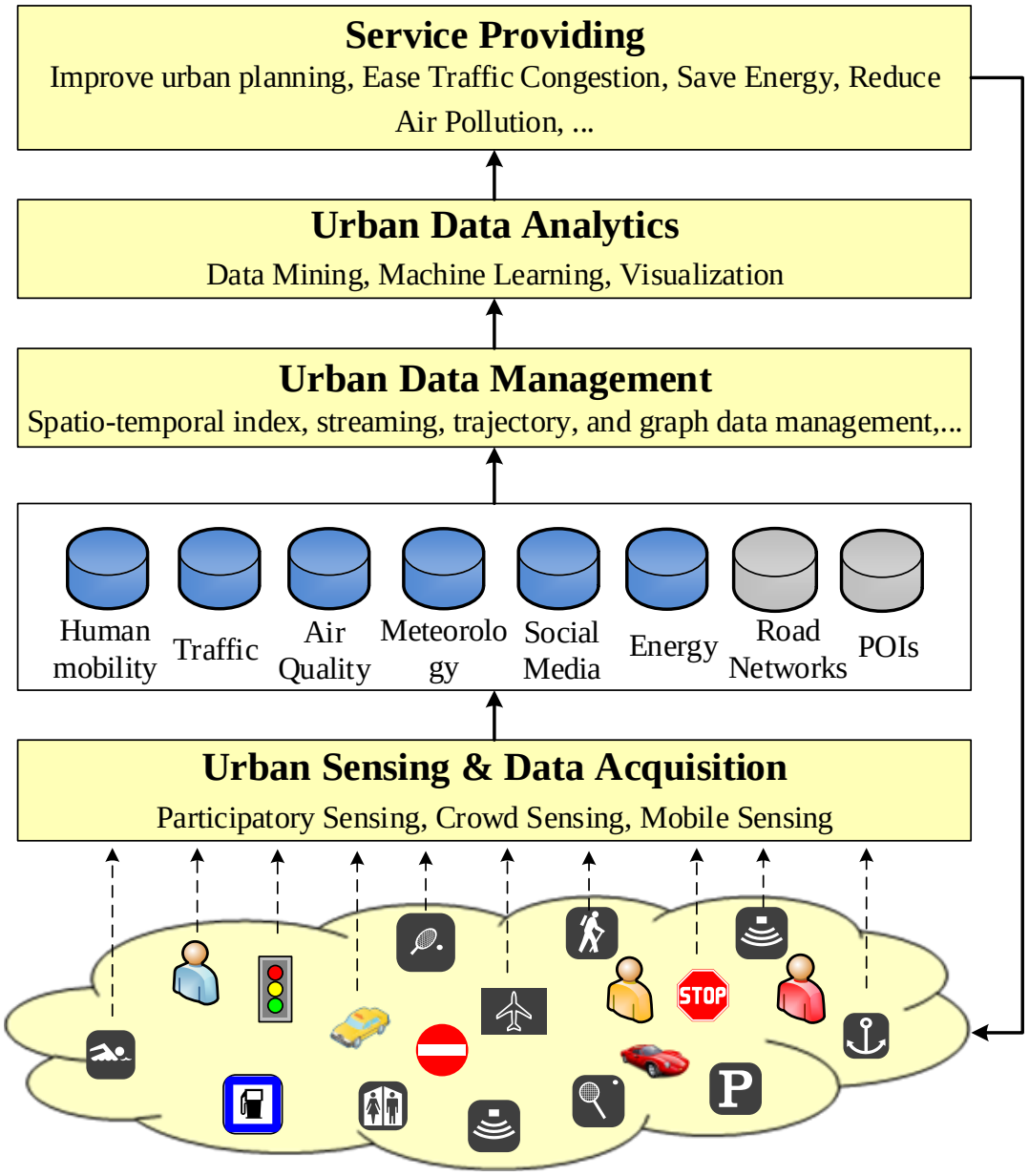
Chair Professor at Shanghai Jiaotong University

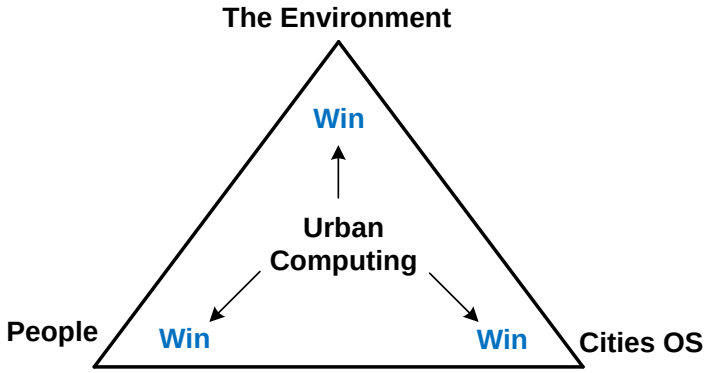
http://research.microsoft.com/en-us/projects/urbancomputing/default.aspx

Big Challenges in Big Cities

Big Data in Cities





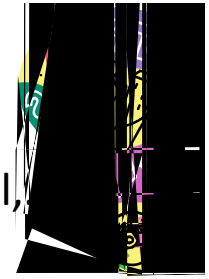


Tackle the Big challenges in Big cities using Big data!

Zheng, Y., et al. Urban Computing: concepts, methodologies, and applications. *ACM transactions on Intelligent Systems and Technology*.

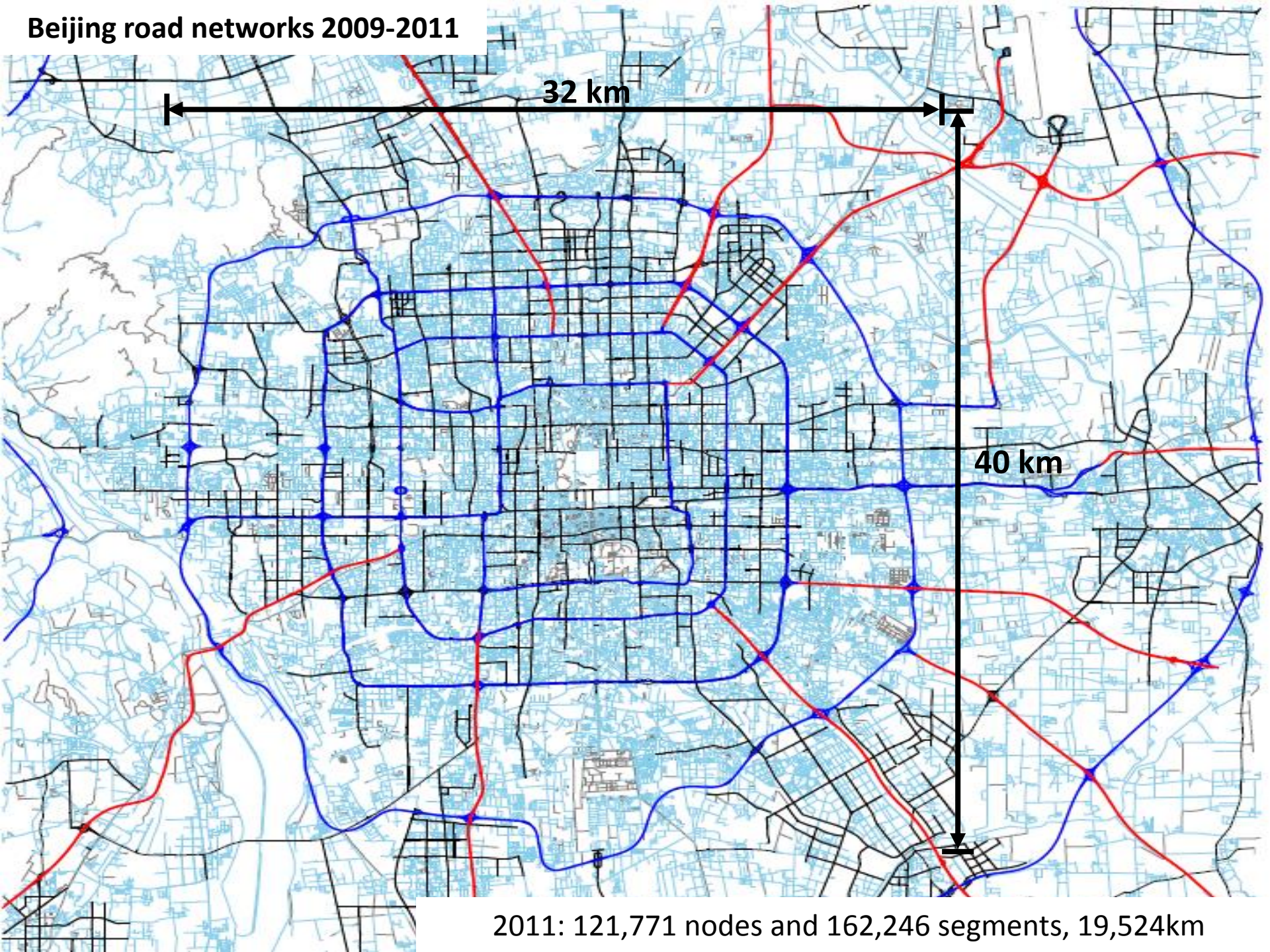
Key Focuses and Challenges

- 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

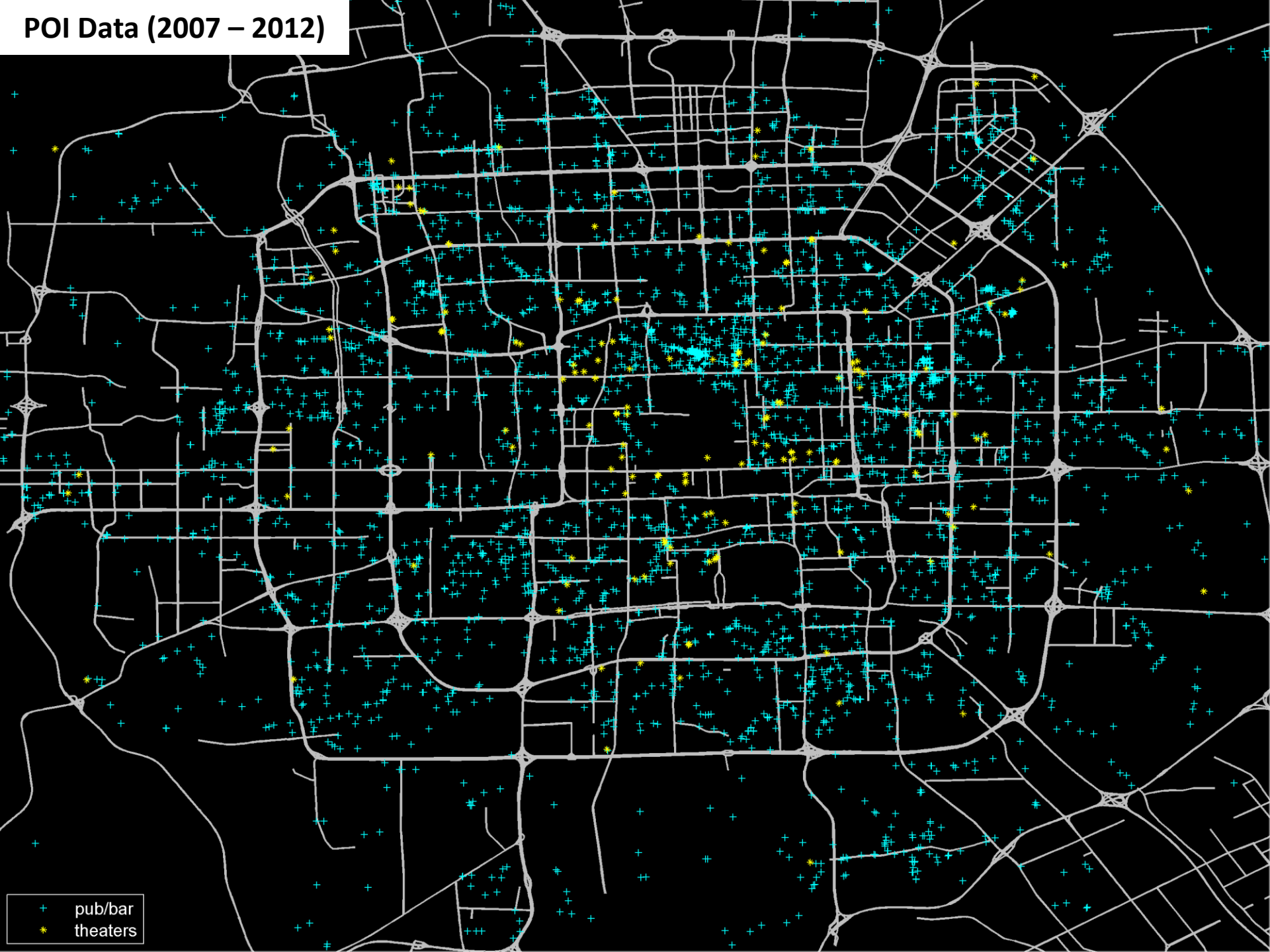


- Computing with heterogeneous data sources
  - Geospatial, temporal, social, text, images, economic, environmental,
  - Learn mutually reinforced knowledge across a diversity of data
  - Efficiency + Effectiveness: Data Management + Mining + Machine Learning

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

POI Data (2007 – 2012)




260
240
220
200
180
160
140
120
2007
2007.5
2008
2008.5
2009
2009.5
2010
2010.5
2011
2011.5
2012
theater

Air Quality Data

Forecast | Current AQI | More Maps

National Parks/Monuments Tribal Boundaries

The tribal boundaries shown here are provided by the Bureau of Indian Affairs and are intended to be used as a general spatial reference only. They are not a formal determination of tribal boundaries by the EPA.

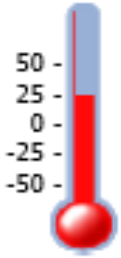
Good | Moderate | USG | Unhealthy | Very Unhealthy | Hazardous | ! Action Day

Click on the city name for more detailed information. printable summary	FORECAST Tue Aug 20	FORECAST Wed Aug 21	CURRENT AQI
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

Meteorological data

Weather Observation

Temp



26℃

Humidity: 73%

Wind direction

NE

Monthly Climate History

Accumulated Rainfall
159.7mm

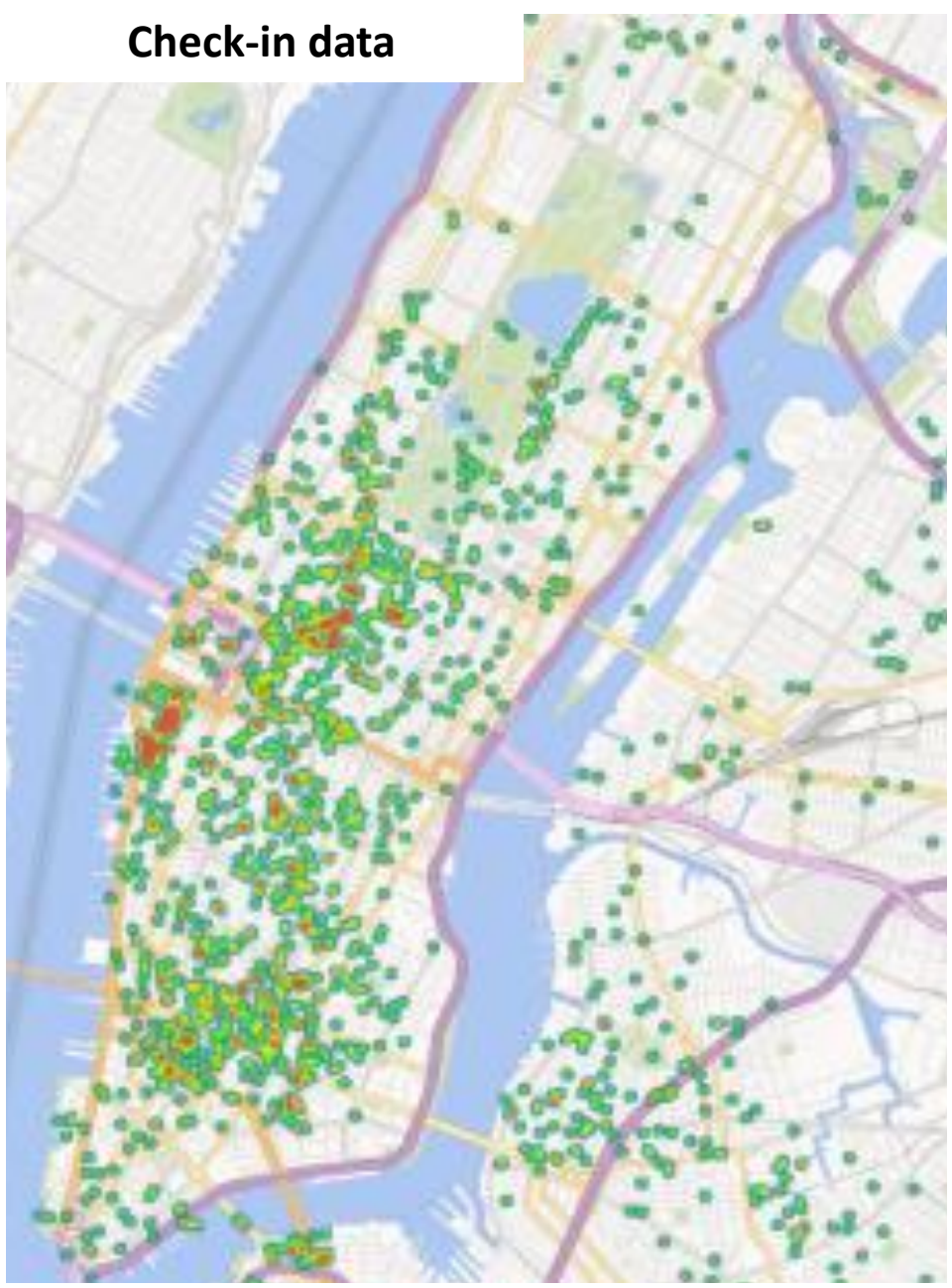
Average Temperature
24.9℃

High Temperature
36.1℃

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

4-7 Days Forecast

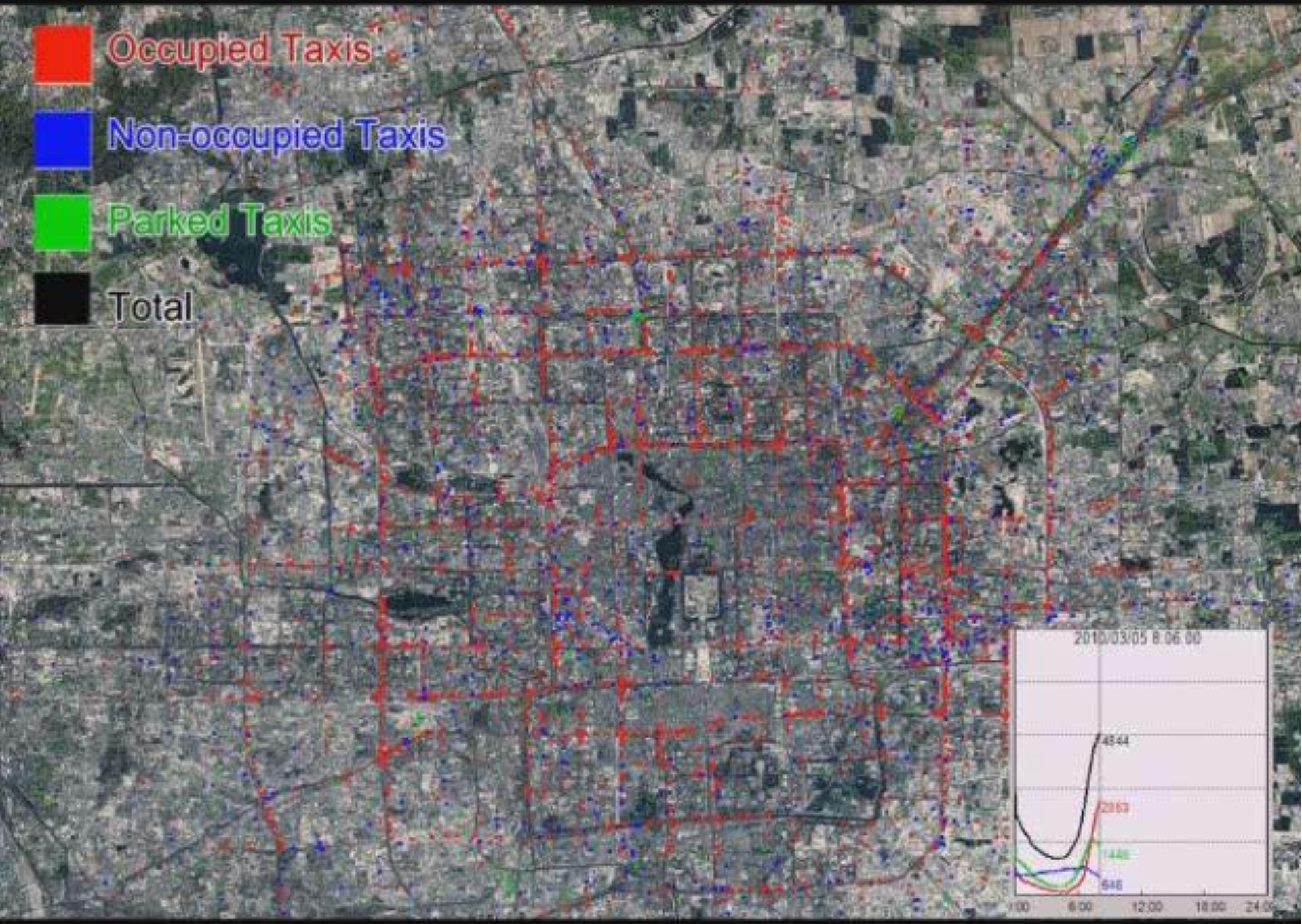
Date		weatherForecast		Temperature	wind
Tuesday Aug 20	night		Shower	Low: 23℃(73℉)	<12km/h
Wednesday Aug 21	day		Cloudy	High: 30℃(86℉)	<12km/h
	night		Cloudy	Low: 22℃(72℉)	<12km/h
Thursday Aug 22	day		Sunny	High: 29℃(84℉)	<12km/h
	night		Sunny	Low: 22℃(72℉)	<12km/h
Friday Aug 23	day		Sunny	High: 32℃(90℉)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



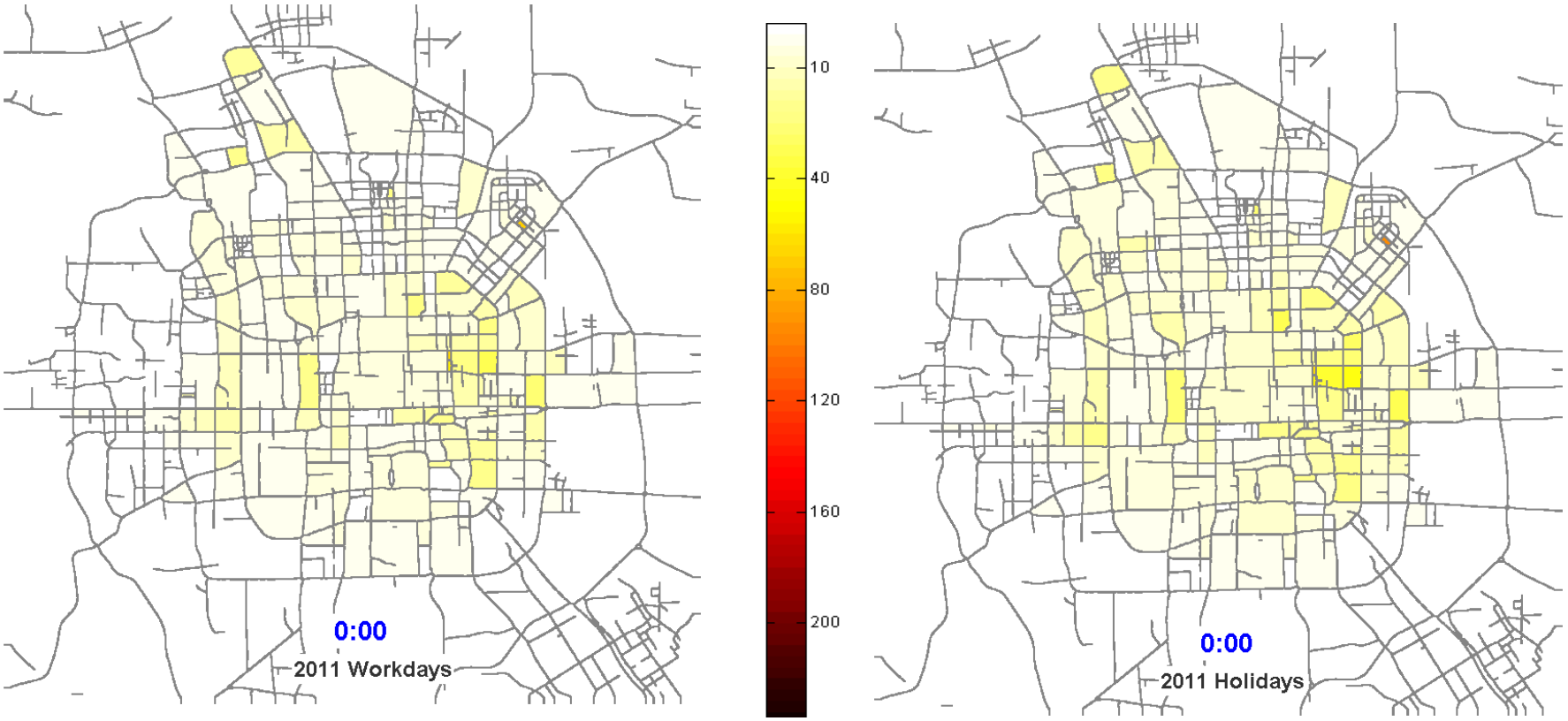
Check-in: Entertainment

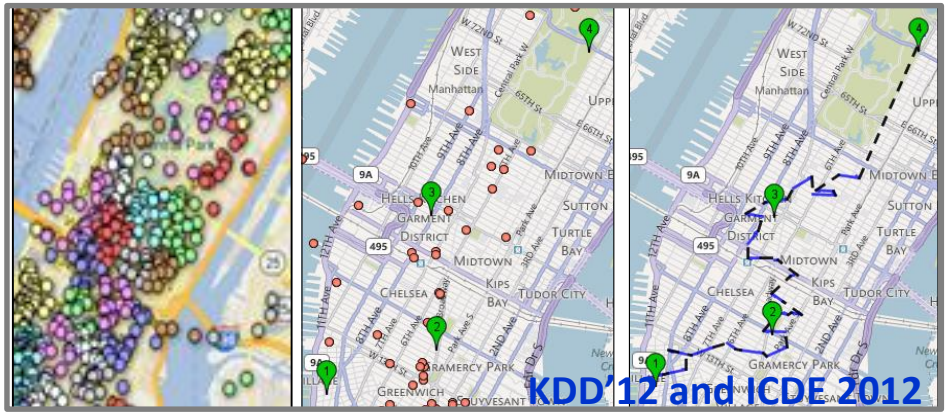
Check-ins: Nightlife Spot

GPS trajectories of 33,000 taxis from 2009 to 2013

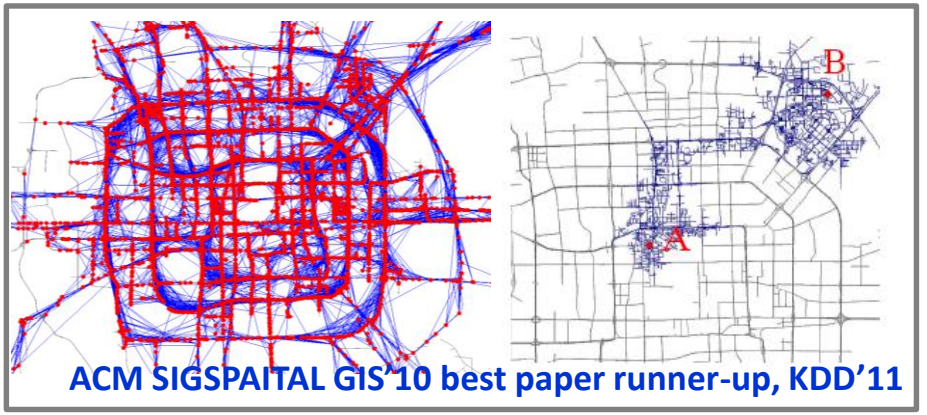

Occupied Taxis
Non-occupied Taxis
Parked Taxis
Total
2010/03/05 8:06:00
4844
2353
1446
646
1:00
6:00
12:00
18:00
24:00

Heat Maps of Beijing (2011)

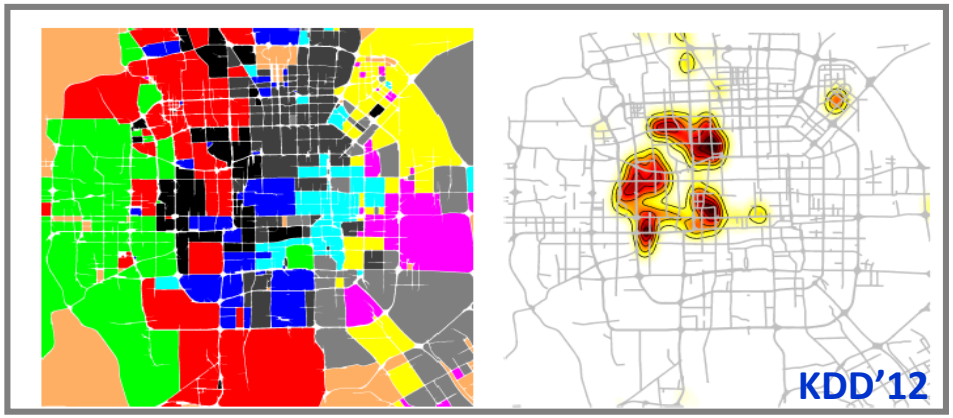




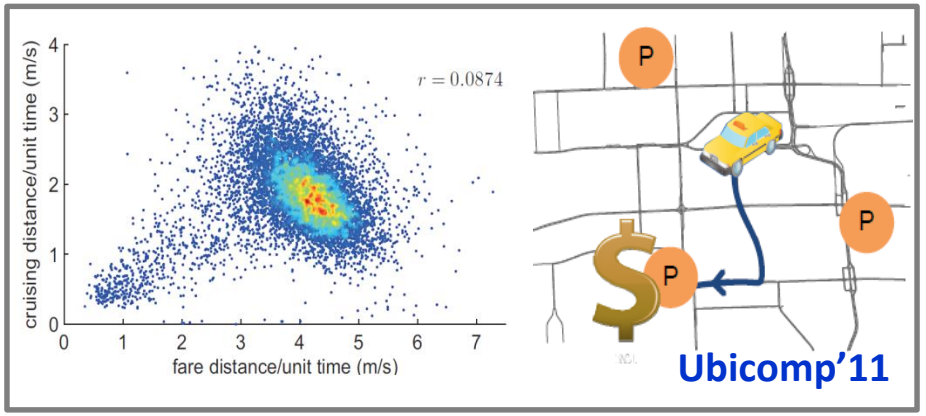
Route Construction from Uncertain Trajectories



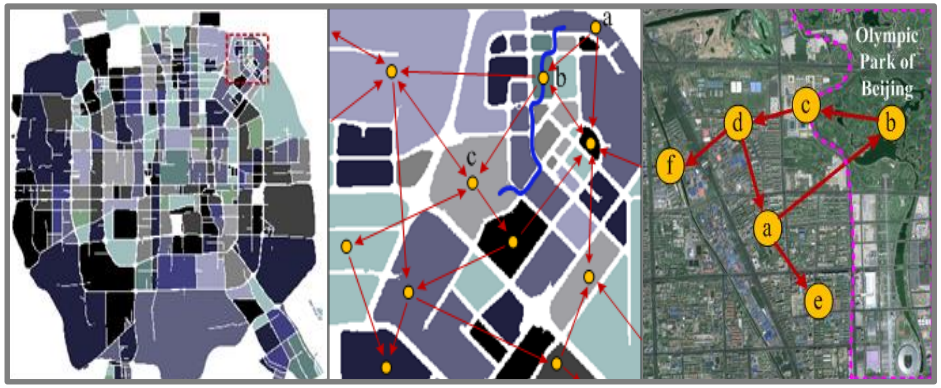
Finding Smart Driving Directions



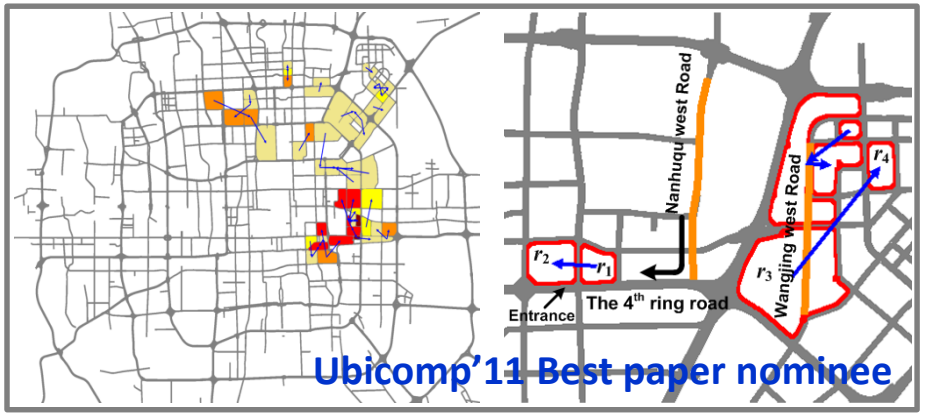
Discovery of Functional Regions



Passengers-Cabbie Recommender system

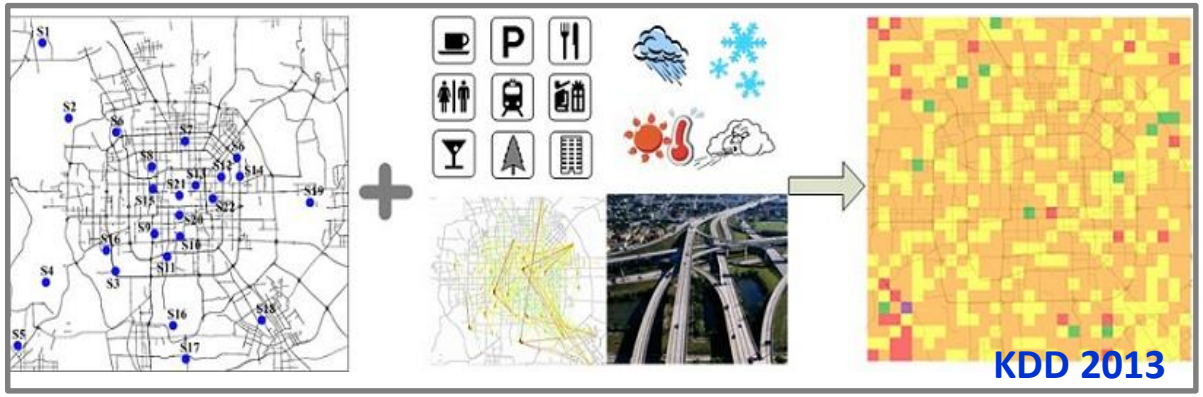


Anomalous Events Detection KDD'11 and ICDM 2012

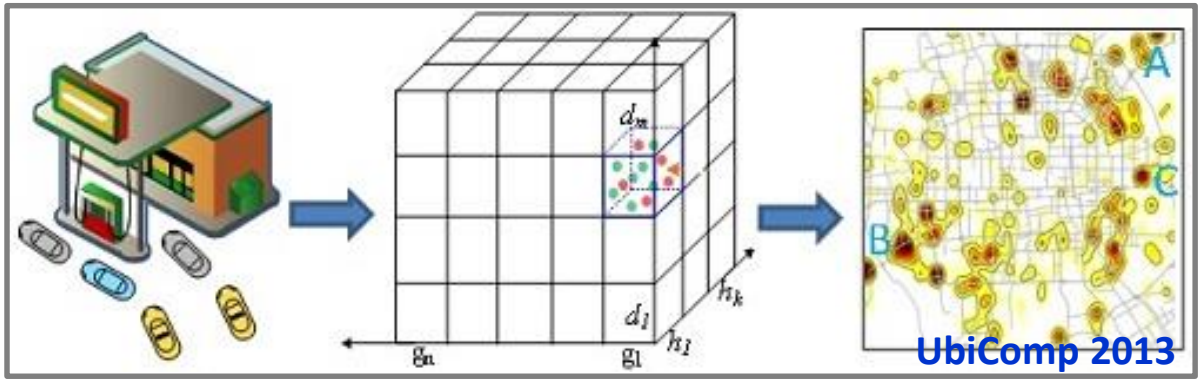


Urban Computing for Urban Planning

Real-time and large-scale dynamic ridesharing



Real-time and find-grained air quality inference using big data



Real-time city-scale gas consumption sensing

U-AIR

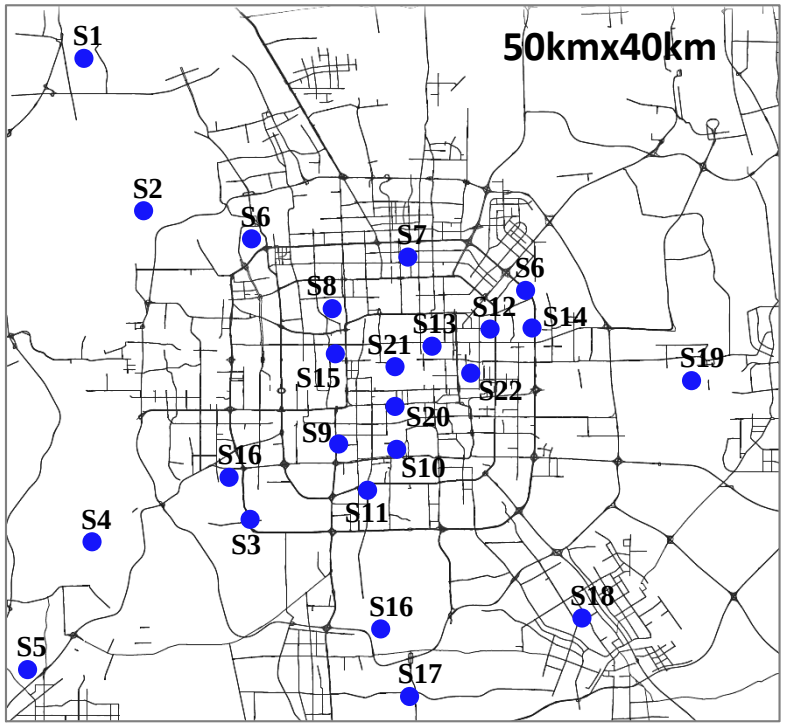
Real-Time and Fine-Grained Air Quality throughout a City

KDD 2013

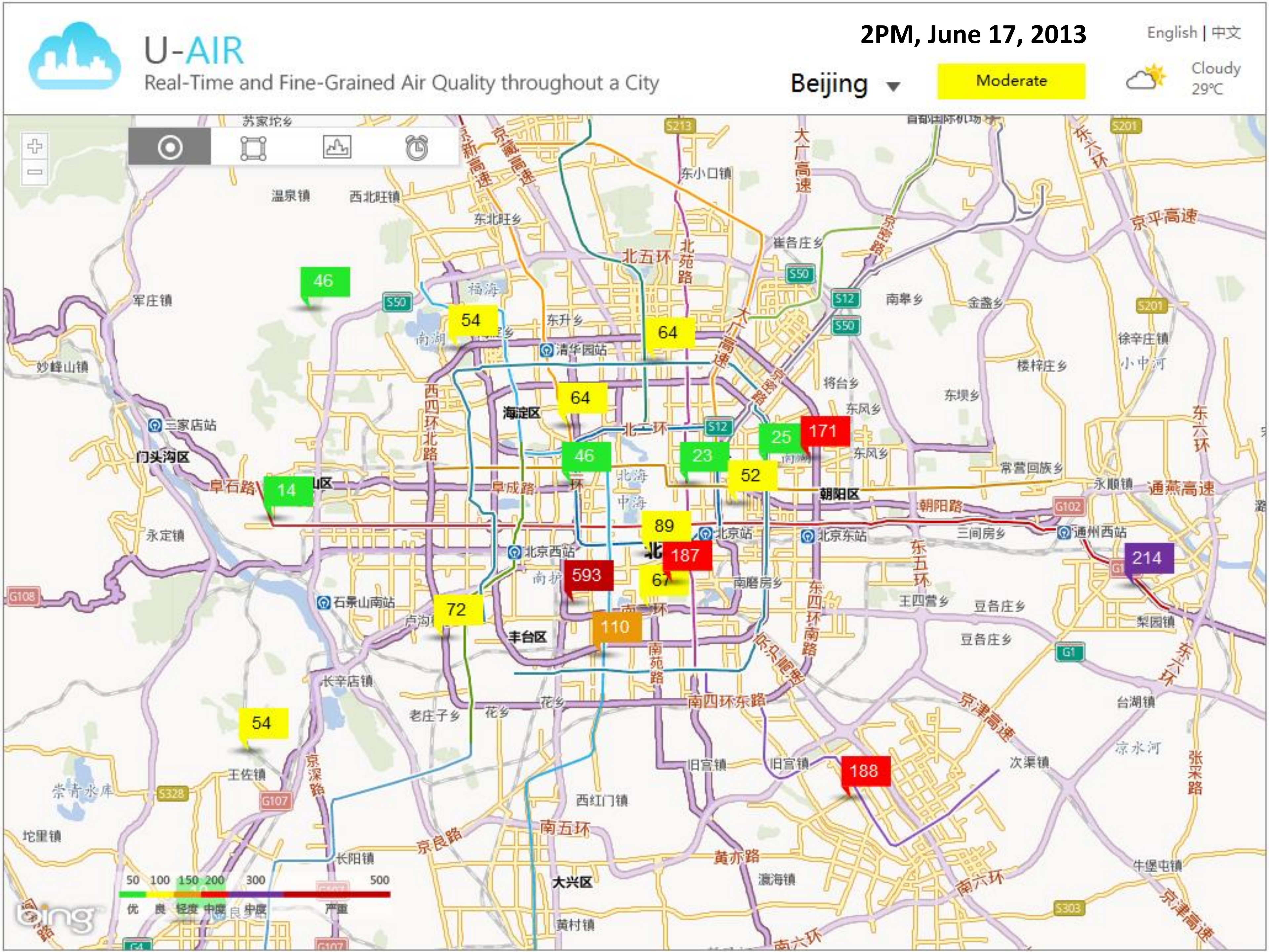
http://urbanair.msra.cn/

Background

- Air quality monitor station

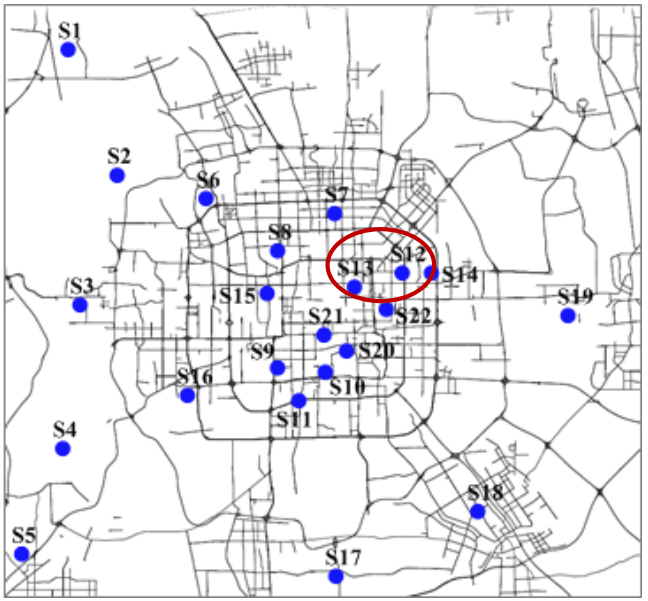


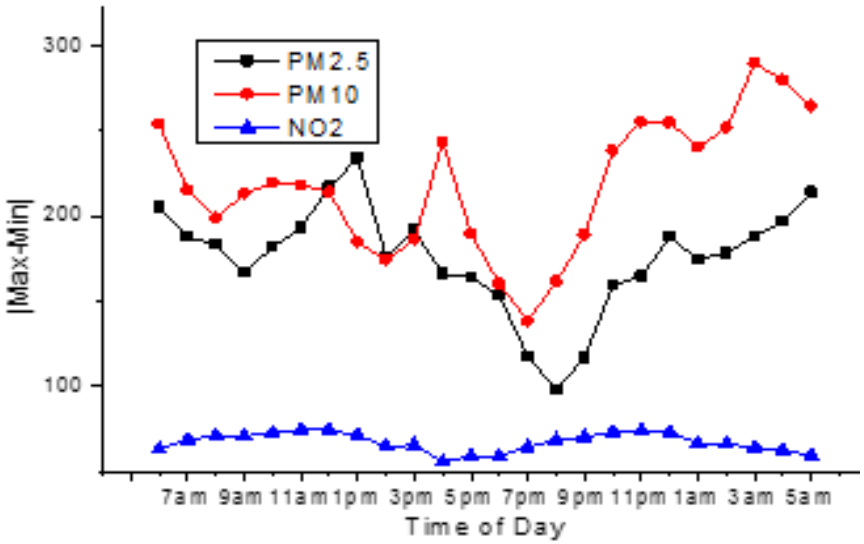
2PM, June 17, 2013



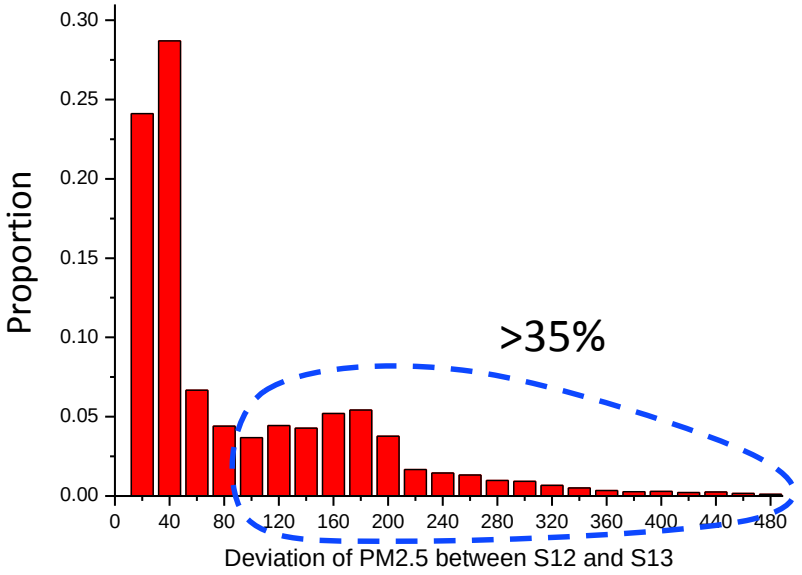
Challenges

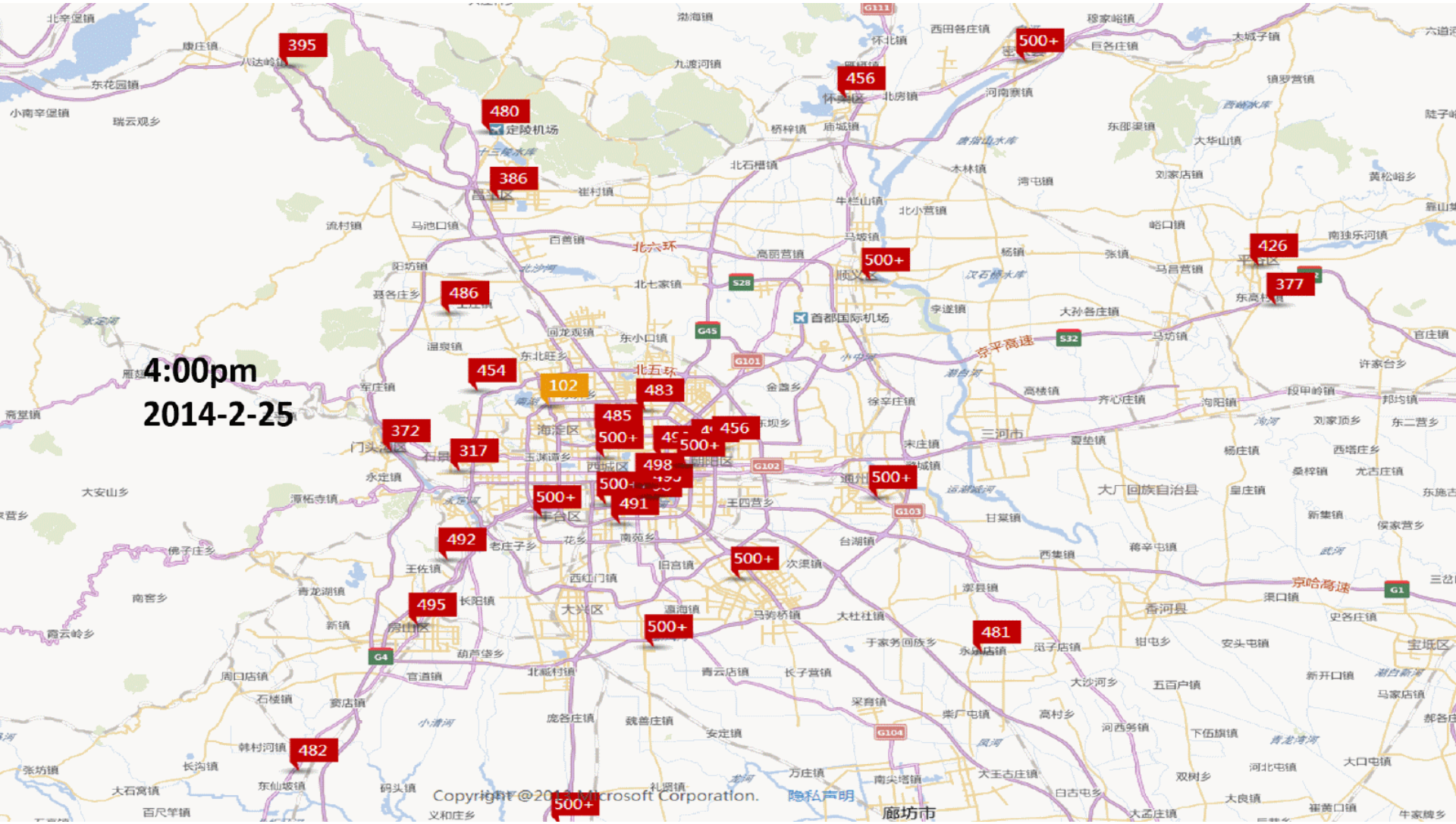
- Air quality varies by locations non-linearly
- Affected by many factors
  - Weathers, traffic, land use…
  - Subtle to model with a clear formula



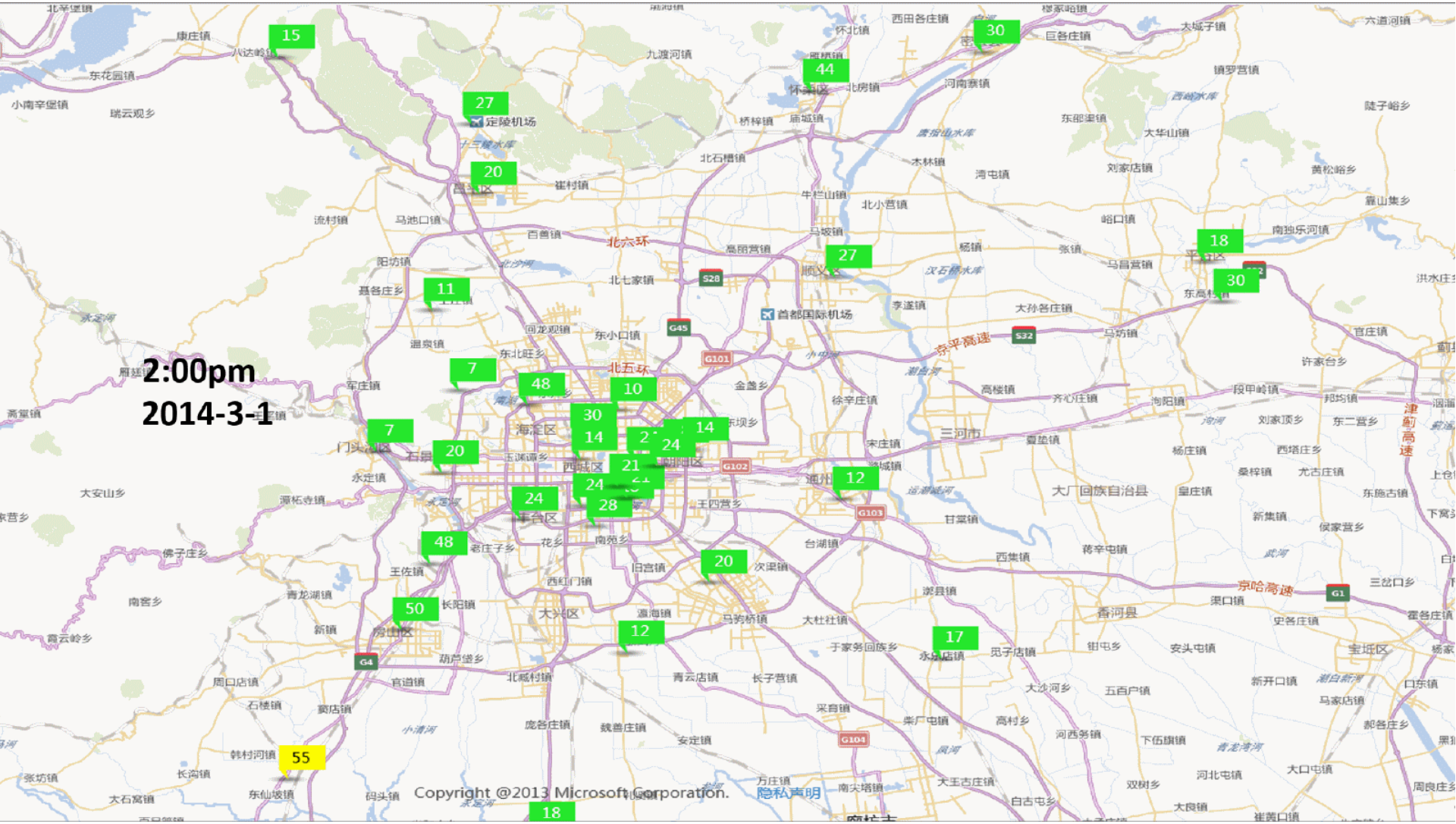


A) Beijing (8/24/2012 - 3/8/2013)

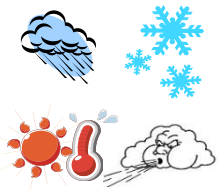



4:00pm
2014-2-25
395
480
386
456
500+
486
454
102
483
485
372
317
498
491
492
495
482
426
377
Copyright © 2014 Microsoft Corporation.
隐私声明
廊坊市

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



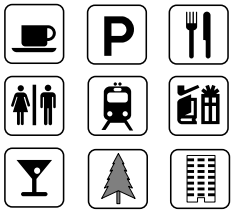
Inferring **Real-Time** and **Fine-Grained** air quality throughout a city using **Big Data**



Meteorology

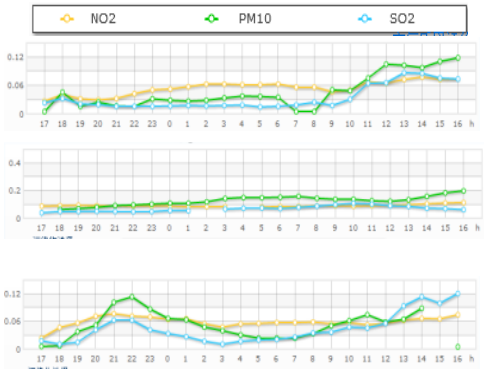
Traffic

Human Mobility



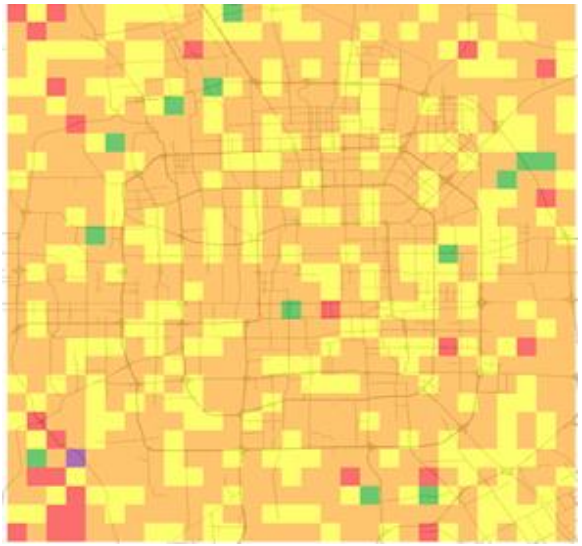
POIs

Road networks

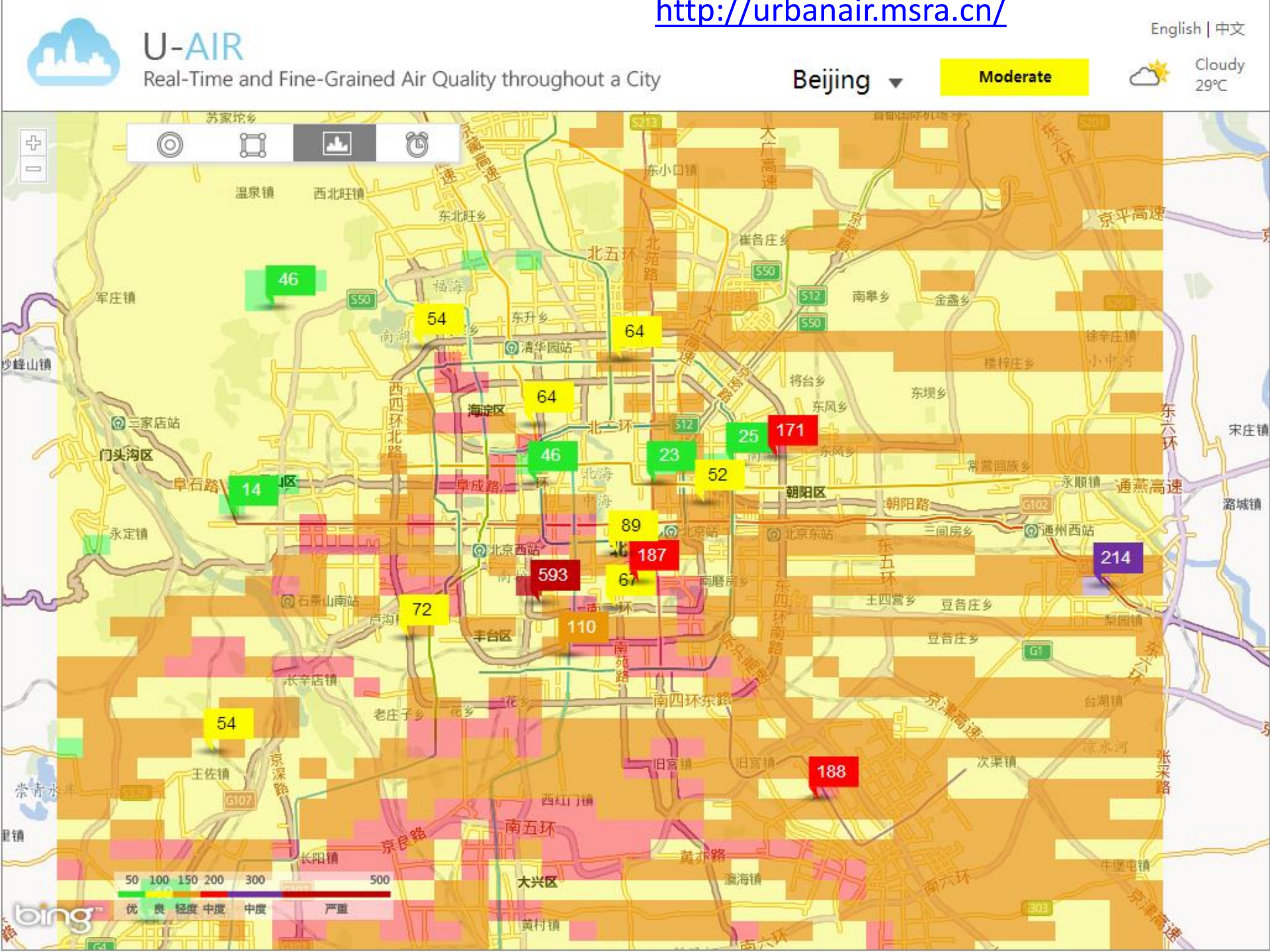


Historical air quality data

Real-time air quality reports



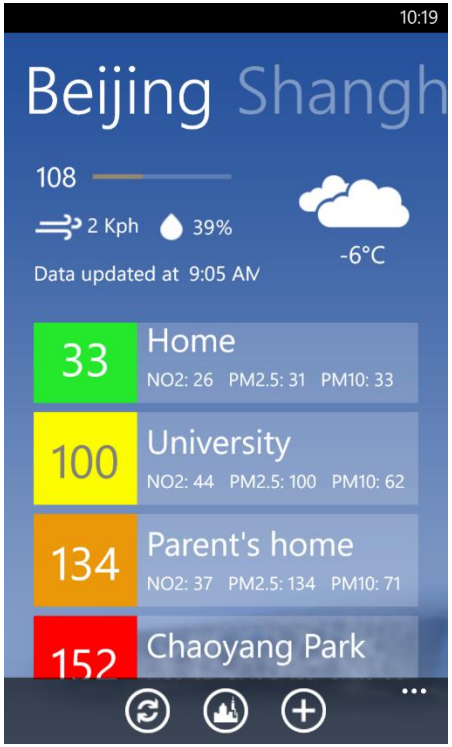
http://urbanair.msra.cn/



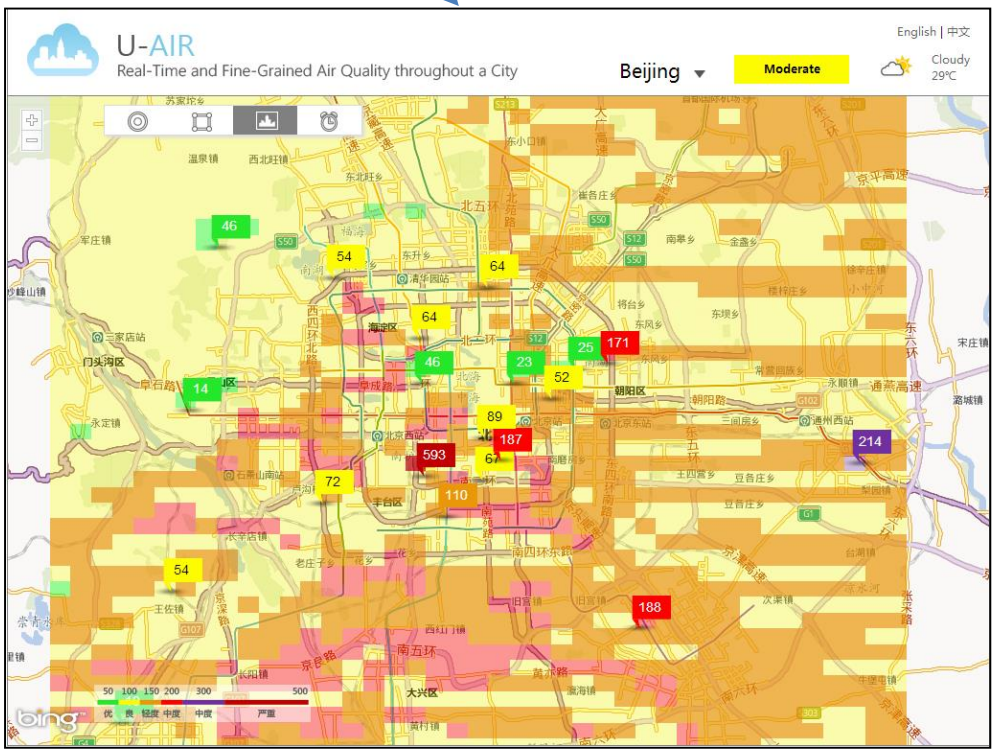
Cloud

MS Azure

Cloud + Client



Clients



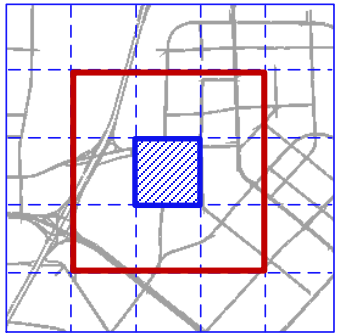
http://urbanair.msra.cn/

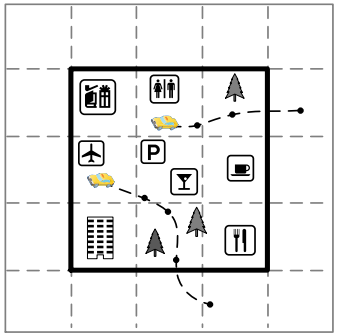
Difficulties

- 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



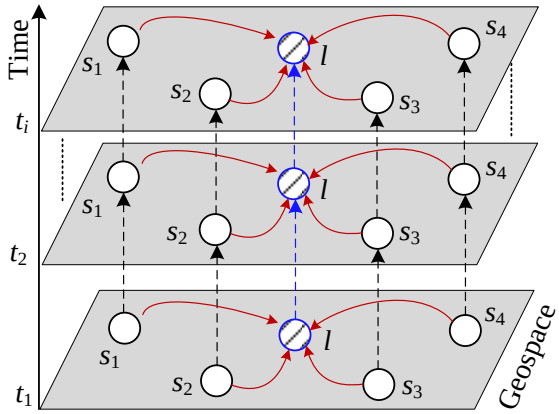


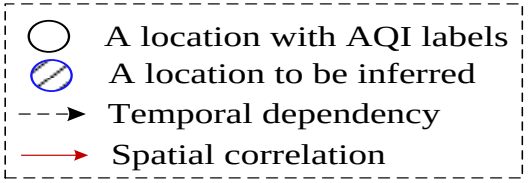
- Co-training-based semi-supervised learning model for each pollutant
  - Predict the AQI labels
  - Data sparsity
  - Two classifiers

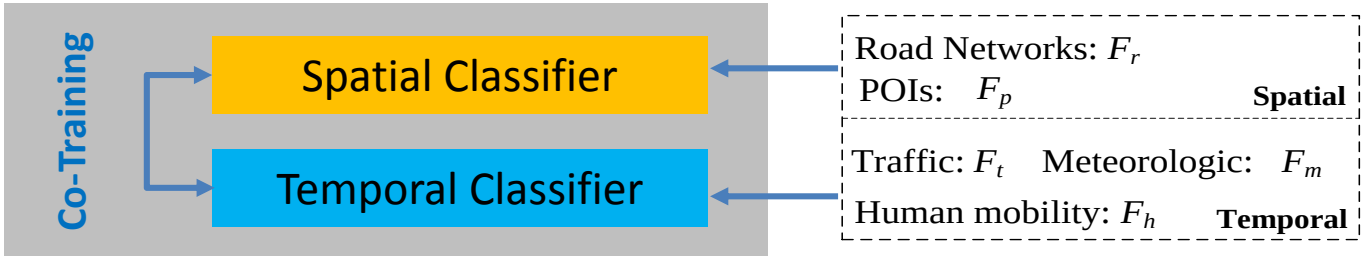
AQI	Values Levels of Health Concern	Colors
0-50	Good (G)	Green
51-100	Moderate (M)	Yellow
101-150	Unhealthy for sensitive groups (U-S)	Orange
151-200	Unhealthy (U)	Red
201-300	Very unhealthy (VU)	Purple
301-500	Hazardous (H)	Maroon

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

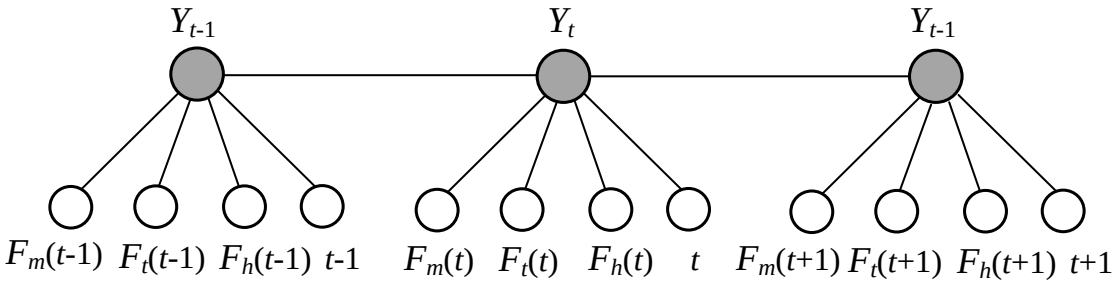






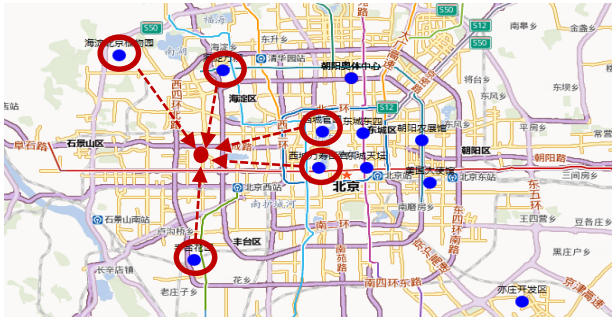
Semi-Supervised Learning Model

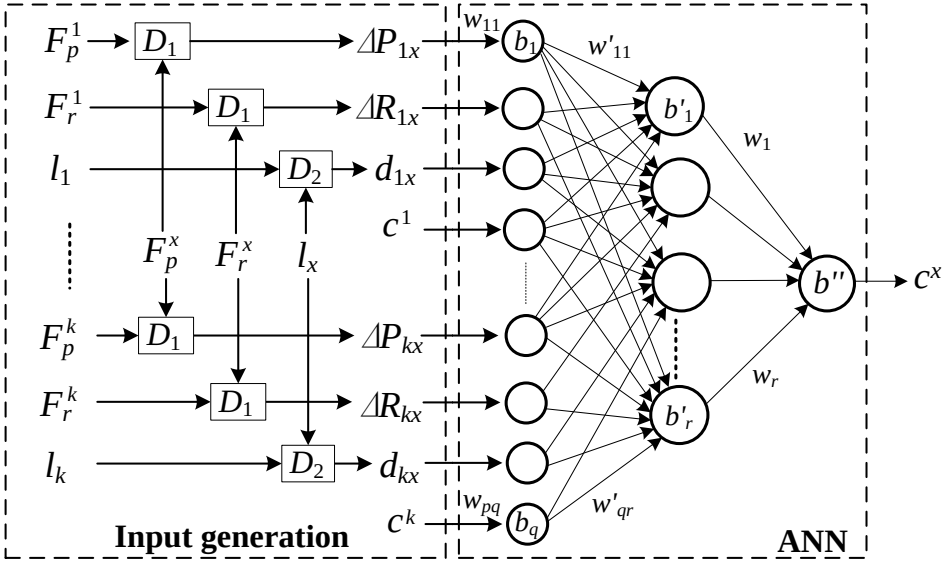
- Temporal classifier (TC)
  - Model the temporal dependency of the air quality in a location
  - Using temporally related features
  - Based on a Linear-Chain Conditional Random Field (CRF)



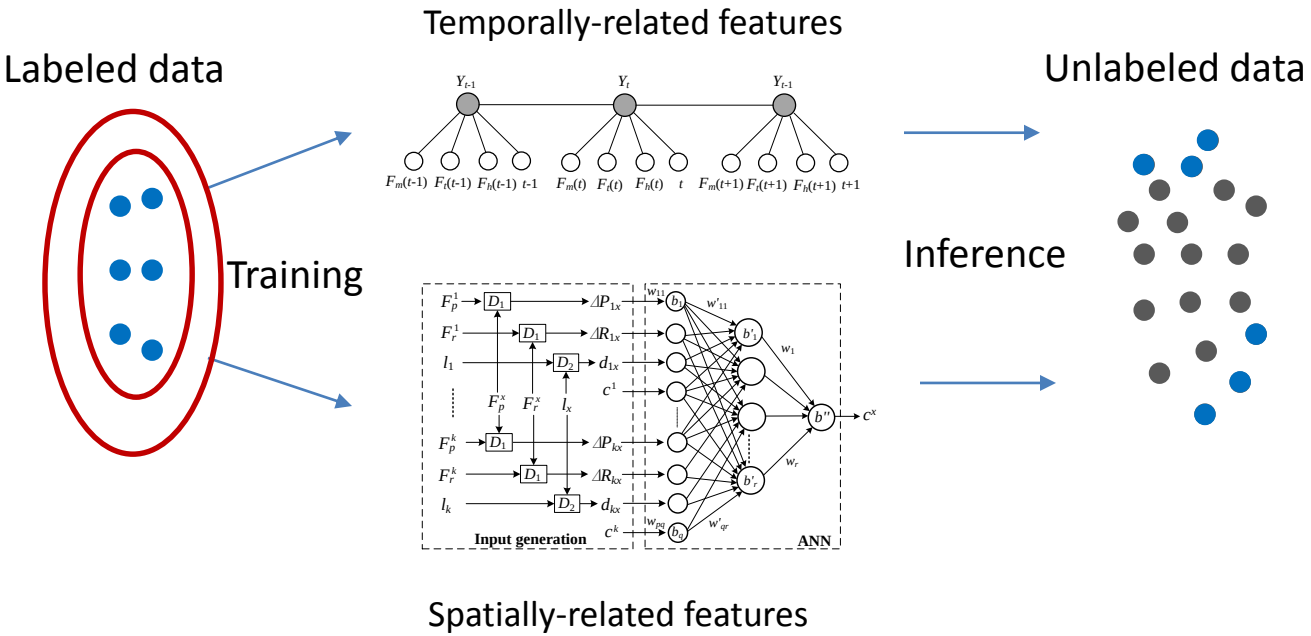
Semi-Supervised Learning Model

- 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

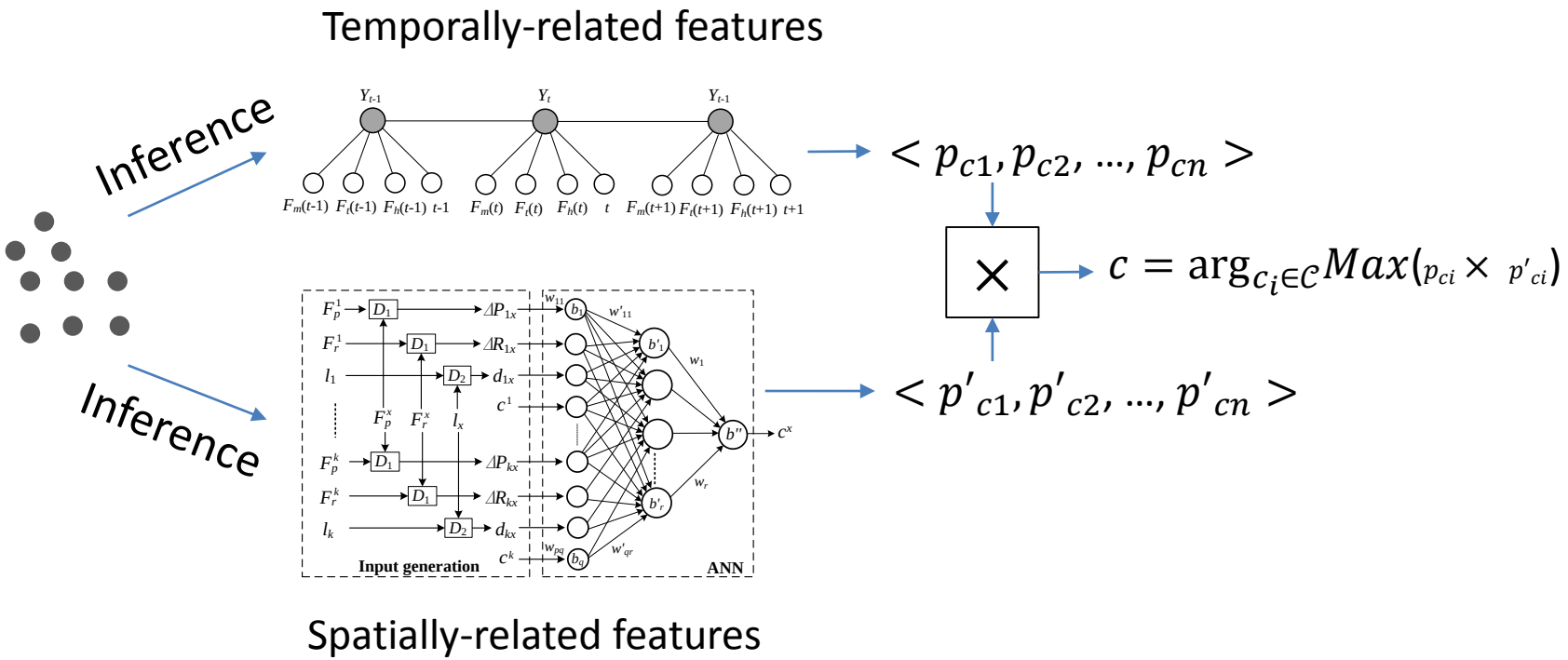




Learning Process of Our Model



Inference Process

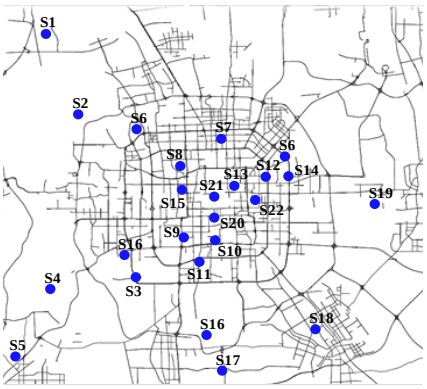


Evaluation

- Datasets

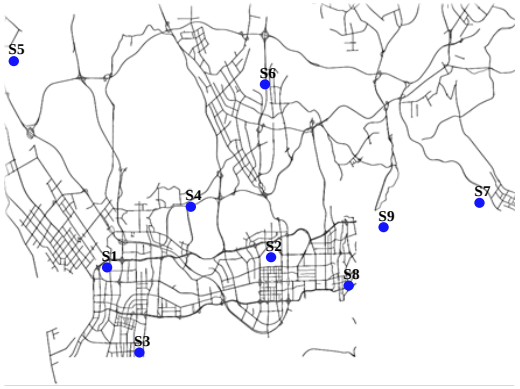
Data Released

Data sources		Beijing	Shanghai	Shenzhen	Wuhan
POI	2012 Q1	271,634	321,529	107,061	102,467
	2012 Q3	272,109	317,829	107,171	104,634
Road	#.Segments	162,246	171,191	45,231	38,477
	Highways	1,497km	1,963km	256km	1,193km
	Roads	18,525km	25,530km KM	6,100km	9,691km
	#. Intersec.	49,981	70,293	32,112	25,359
AQI	#. Station	22	10	9	10
	Hours	23,300	8,588	6,489	6,741
	Time spans	8/24/2012-3/8/2013	1/19/2013-3/8/2013	2/4/2013-3/8/2013	2/4/2013-3/8/2013
Urban Size (grids)		50×50km (2500)	50×50km (2500)	57×45km(2565)	45×25km (1165)



A) Beijing

B) Shanghai

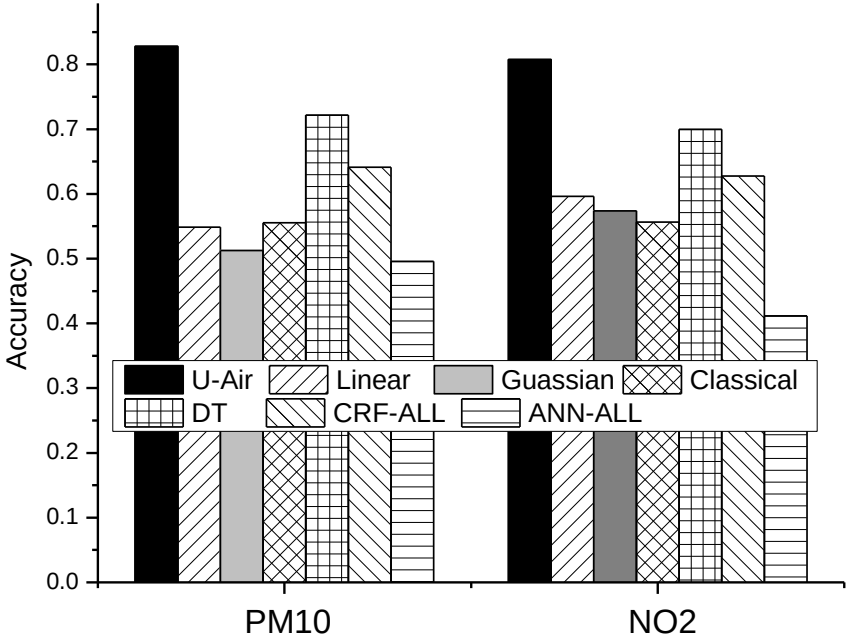


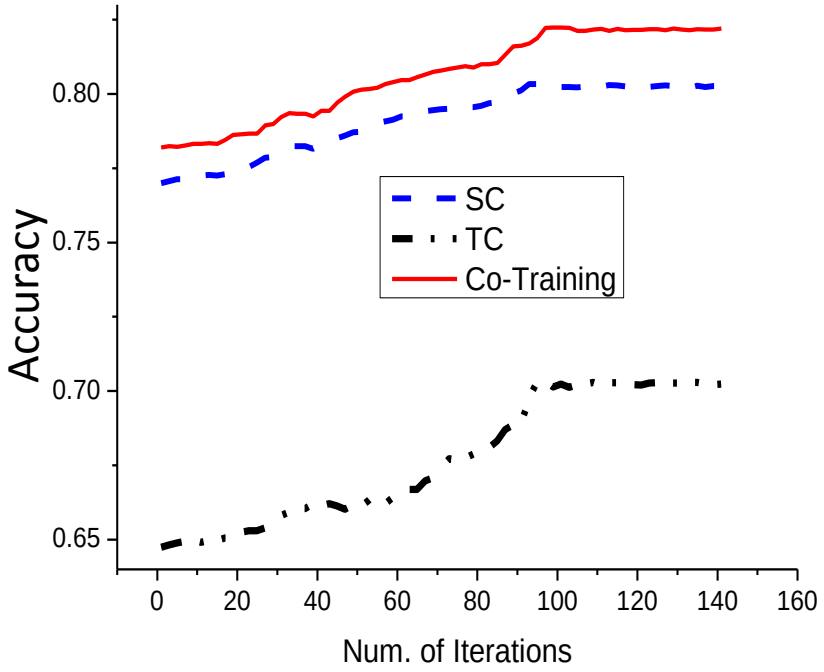
C) Shenzhen

D) Wuhan

Evaluation

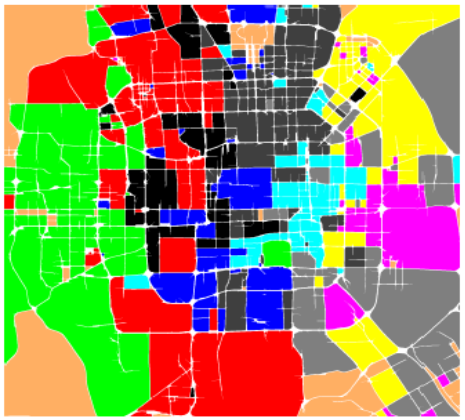
- Overall performance of the co-training





Zheng, Y., et al. U-Air: when urban air quality inference meets big data. KDD 2013

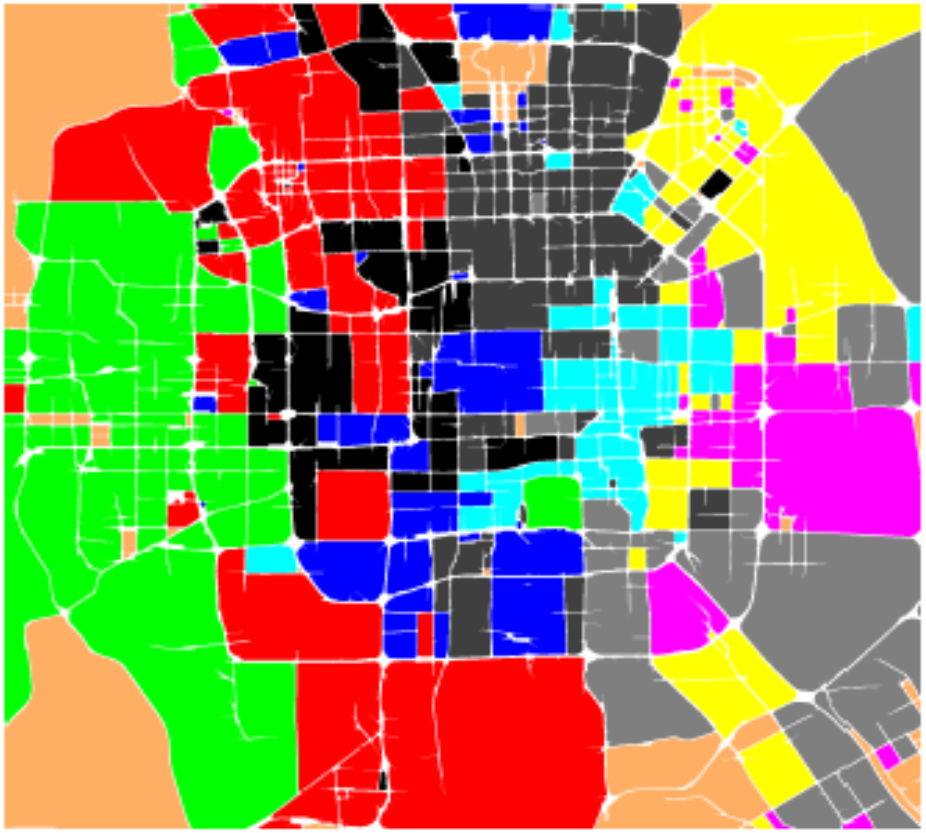
Discover Regions of Different Functions using Human Mobility and POIs



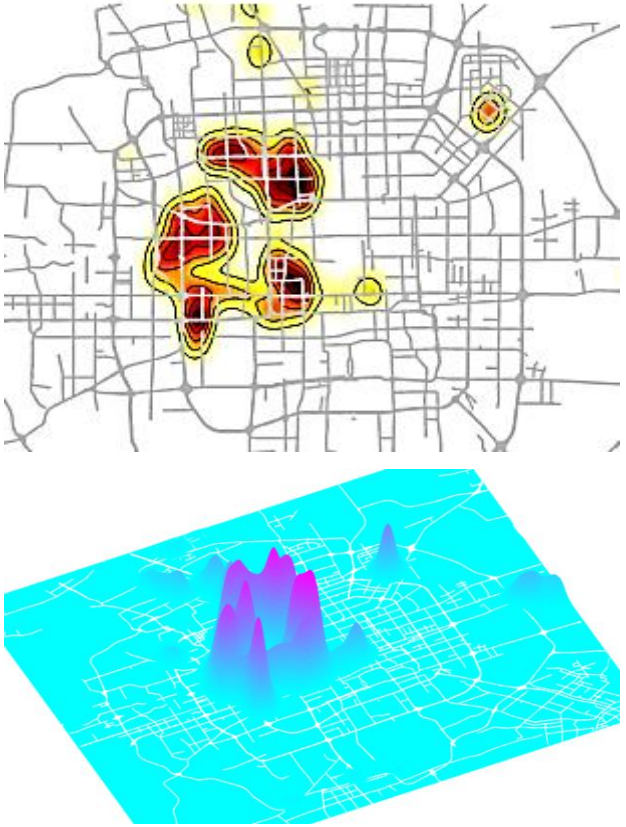
In KDD 2012

Goals

- Discover regions of different functions in urban areas
- Identify the kernel density of a functionality



Functional Regions



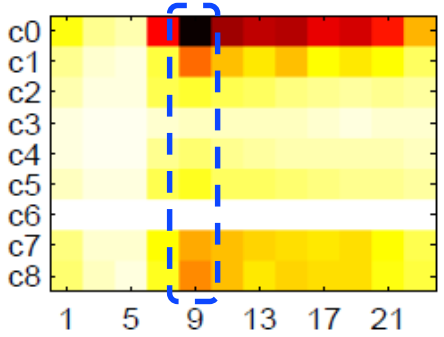
Functionality Density

Motivation and Challenges

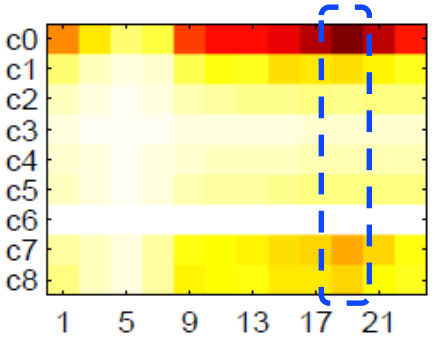
- POIs indicate the function

- But not enough
  - Compound
  - Quality

- Human mobility
  - Differentiate between POIs of the same category
  - Indicate the function of a region



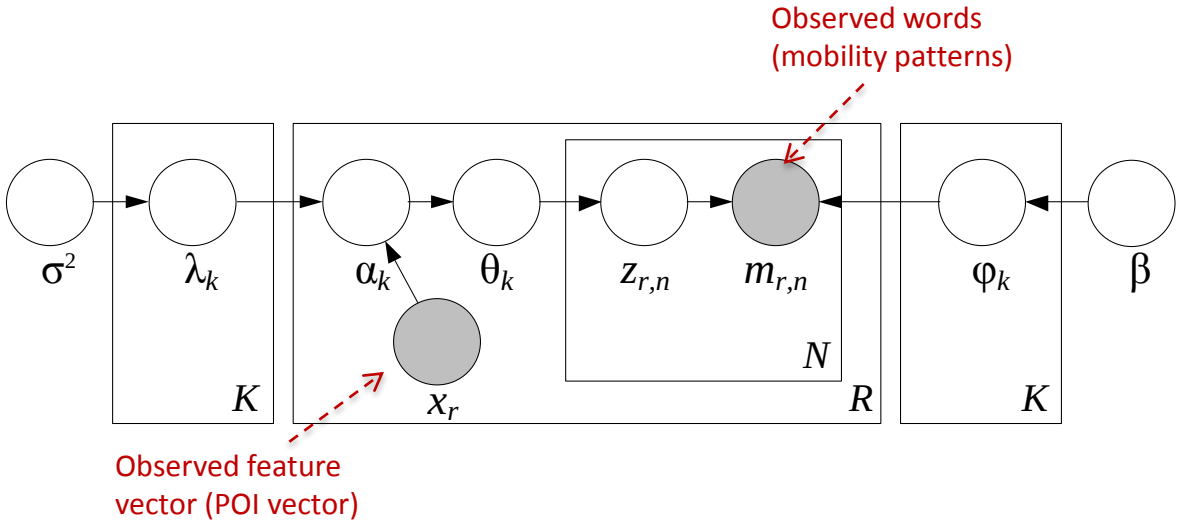
Leaving



Arrival

Methodology Overview

- **Mapping from regions to documents**
  - Regions → Documents (*R*)
  - Functions → Topics (*K*)
  - Mobility patterns → Words (*N*)
  - POIs → meta data like Key words and authors

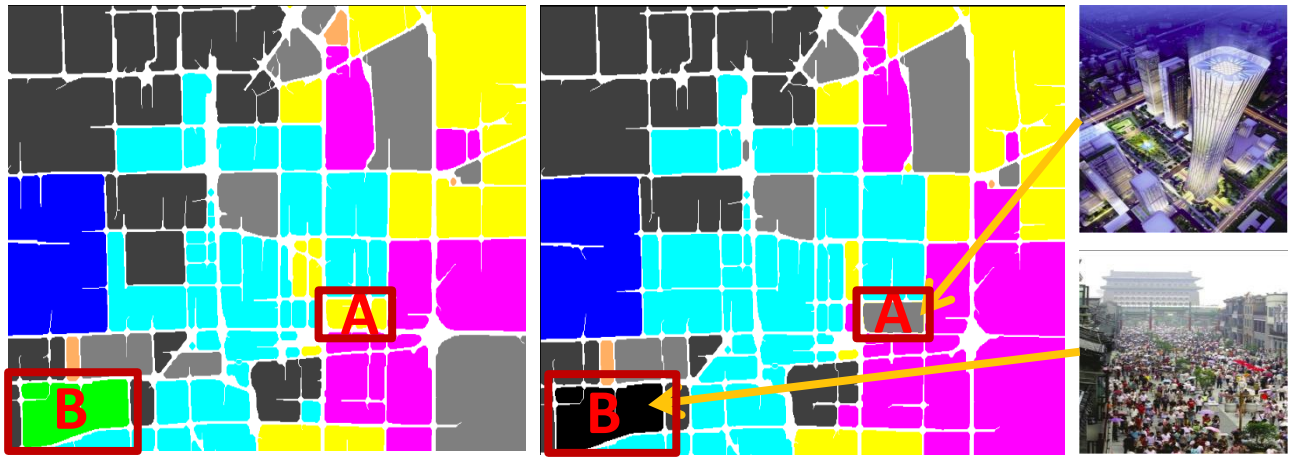


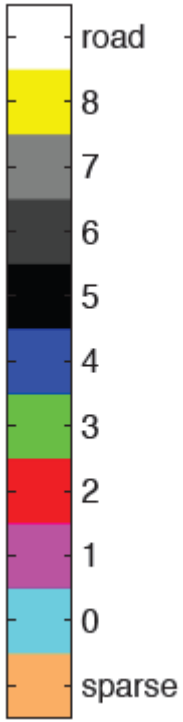
Infer the topic distribution using a LDA(Latent Dirichlet allocation)-variant topic model

Results

2010

2011



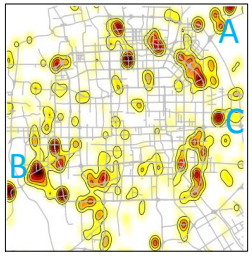


Land use planning (2002-2010)

Results of 2011

Sensing the Pulse of Urban Refueling Behavior

UbiComp 2013



Questions

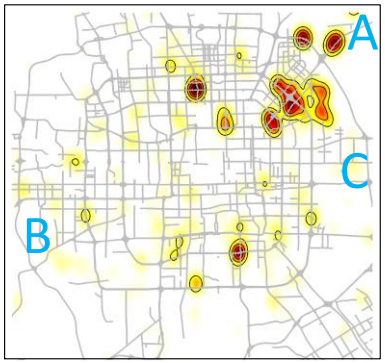
How many liters of gas have been consumed in the past 1 hour in NYC?

Which gas station in 3 miles has the shortest queue?

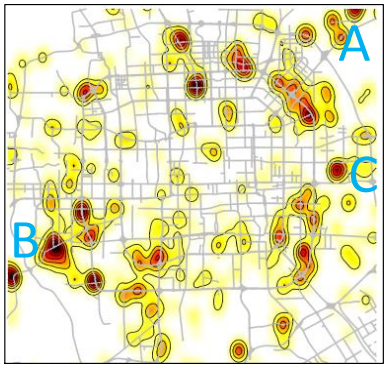
Goal

- Use GPS-equipped taxicabs as a sensor to capture both
  - Waiting time at a gas station
  - City-wide petrol consumption

Waiting time of taxis in a gas station

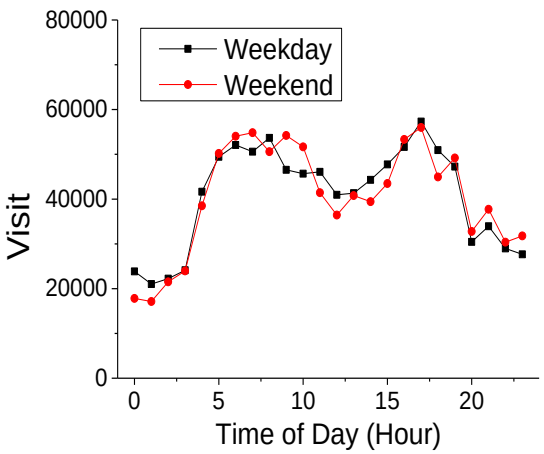


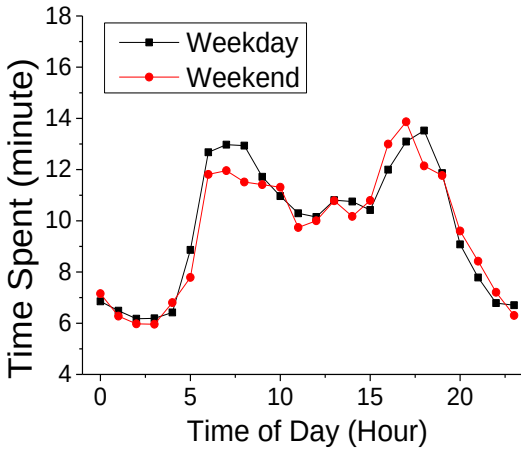
City-scale Gas consumption



Motivation

- Gas stations are owned by competing organizations
  - Do not want to make data available to competitors
  - There is a cost but no benefit for them
  - No time information
- Benefits
  - Gas station recommendation
  - Support the planning and operation of gas stations
  - Monitoring real-time city-scale energy consumption



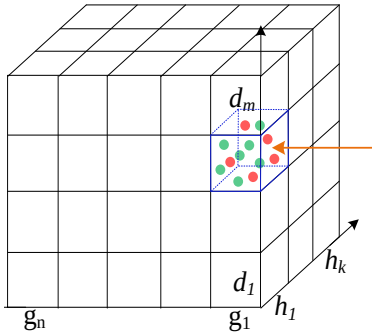


Methodology Overview

Spatio-temporal clustering and classification

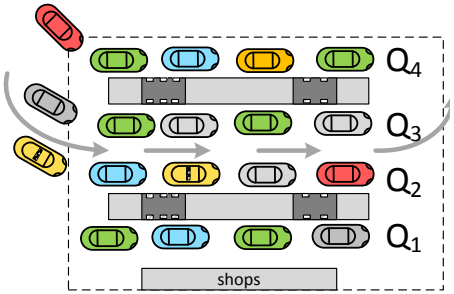
1. Refueling event detection in a gas station

Tensor Decomposition



2. Waiting time inference across different stations

Queue theory

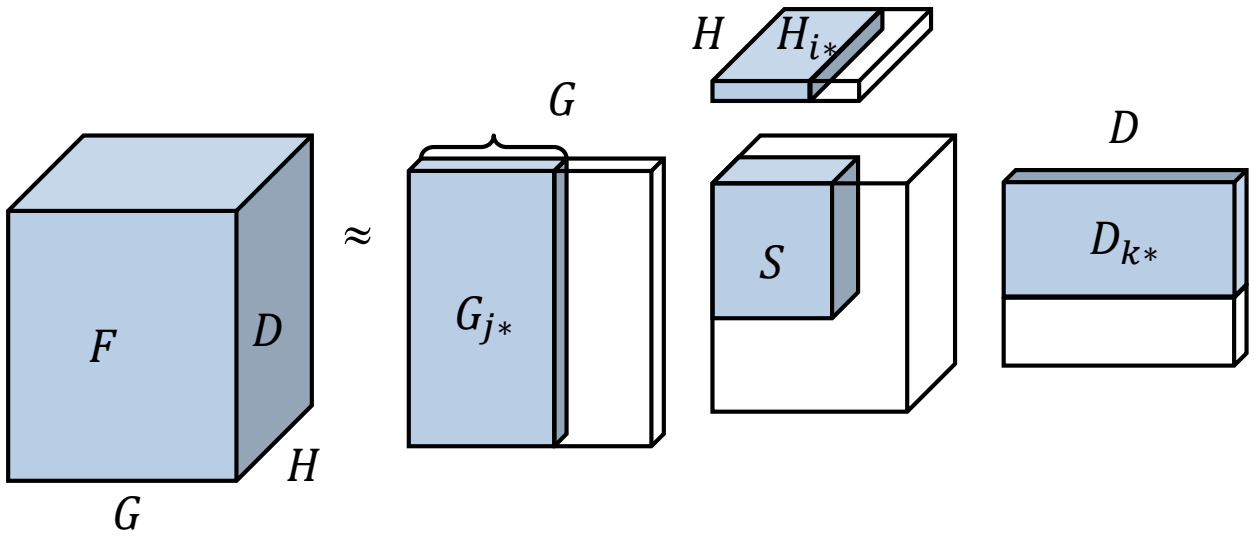


3. Estimation number of vehicles in a station

Expected Duration Learning

- Tensor decomposition
  - Approximate a tensor with the multiplication of three (low-rank) matrices and a core tensor
  - High order singular value decomposition (HOSVD)

$$F_{ijk} = S \times_H H \quad \times_G G \quad \times_D D \quad \approx S \times_H H_{i*} \times_G G_{j*} \times_D D_{k*}$$

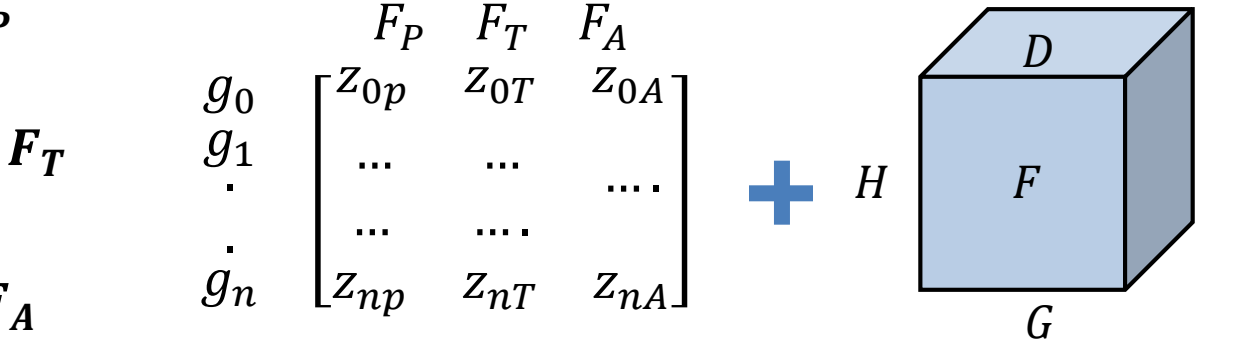


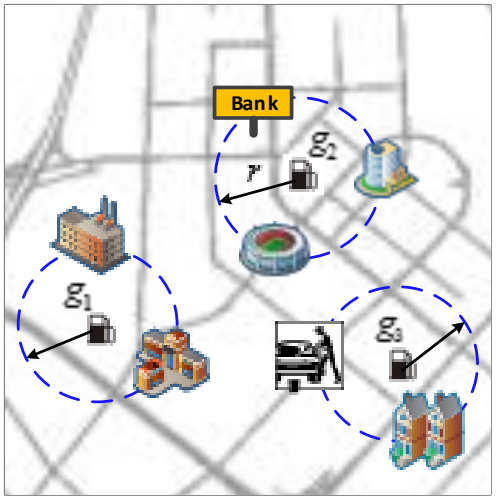
Expected Duration Learning

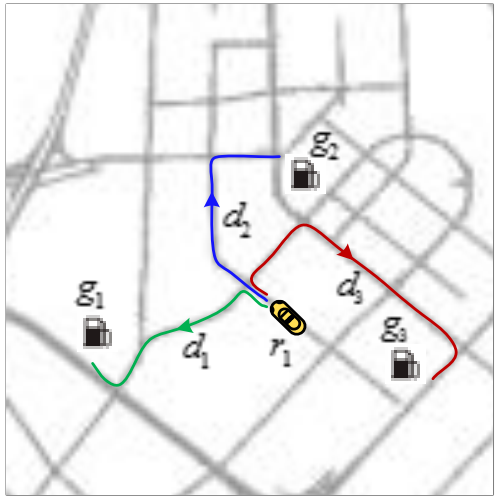
- The context of a station
  - POI feature $\boldsymbol{F_P}$
  - Traffic feature $\boldsymbol{F_T}$
  - Area feature $\boldsymbol{F_A}$

Stations with similar contextual features tend to have a similar duration

$$\begin{array}{c} \\ g_0 \\ g_1 \\ \cdot \\ \cdot \\ g_n \end{array} \begin{array}{c} \begin{array}{ccc} F_P & F_T & F_A \end{array} \\ \begin{bmatrix} z_{0p} & z_{0T} & z_{0A} \\ \cdots & \cdots & \cdots \\ \cdots & \cdots & \\ z_{np} & z_{nT} & z_{nA} \end{bmatrix} \end{array} + H$$





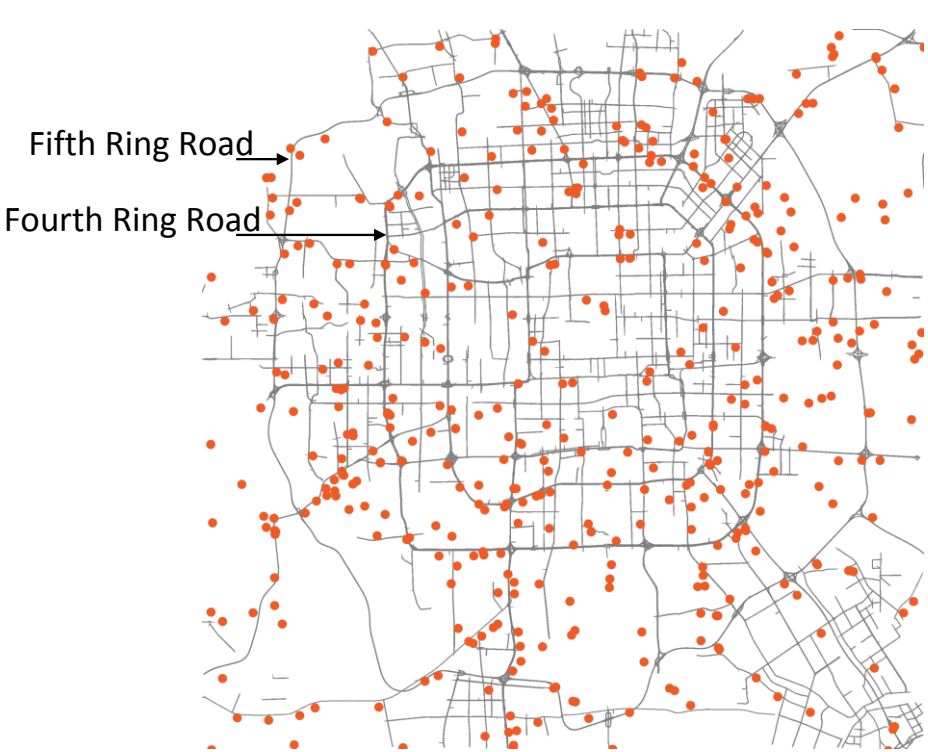


Evaluation

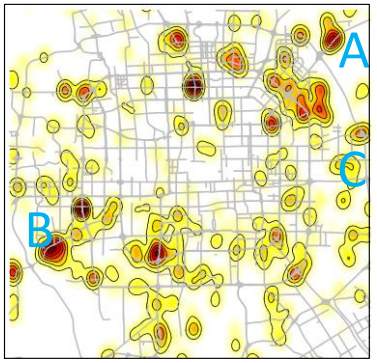
- In the field study
  - Sent students to two stations to observe the queues
  - Oct.17 to Nov.15 in 2012, 5-6pm
  - Recorded the number of vehicles and the waiting time of some
  - Evaluate waiting time and number of vehicles

Visualization

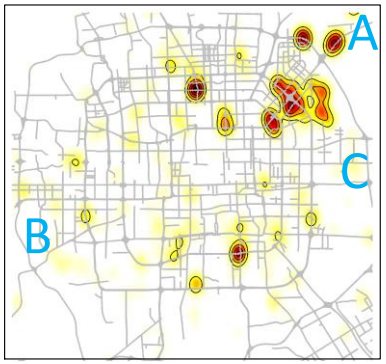
- Geographic View



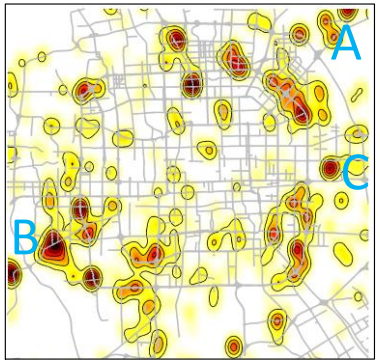
(a) stations' distribution



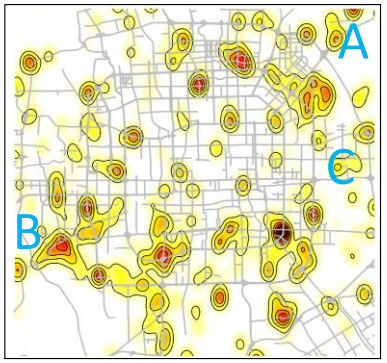
(b) taxis' time spent



(c) taxis' visits

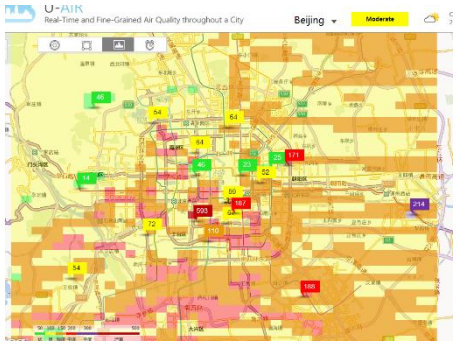


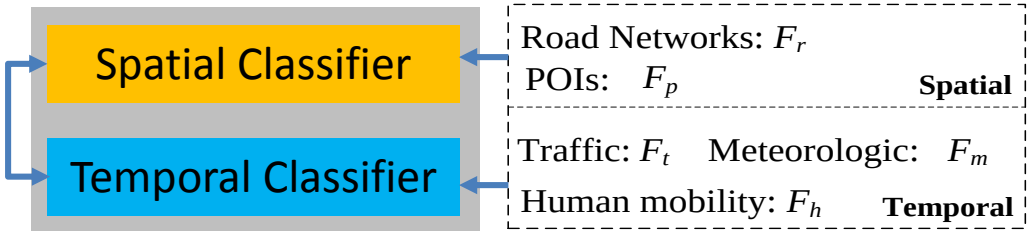
(d) urban's time spent



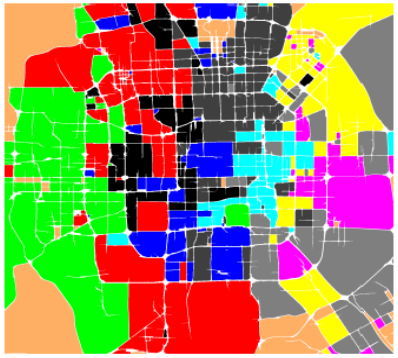
(e) urban's visits

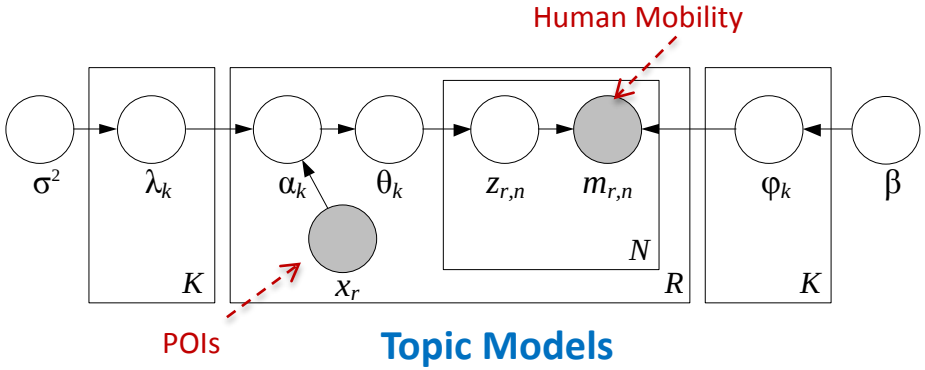
Computing with Multiple Heterogeneous Data Sources



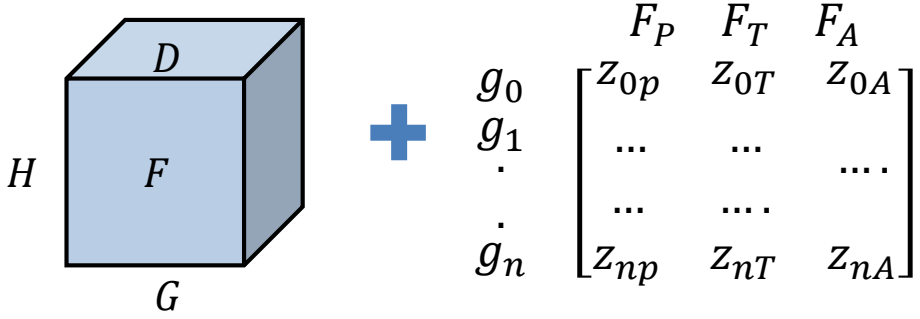


Co-Training-based Semi-supervised learning





Topic Models



Context-aware Tensor Decomposition

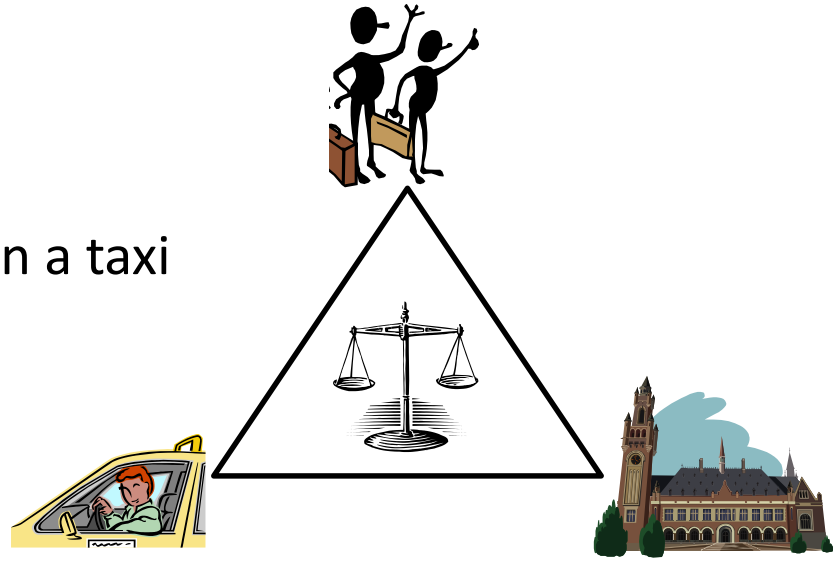
T-Share: A Large-Scale Dynamic Taxi Ridesharing Service

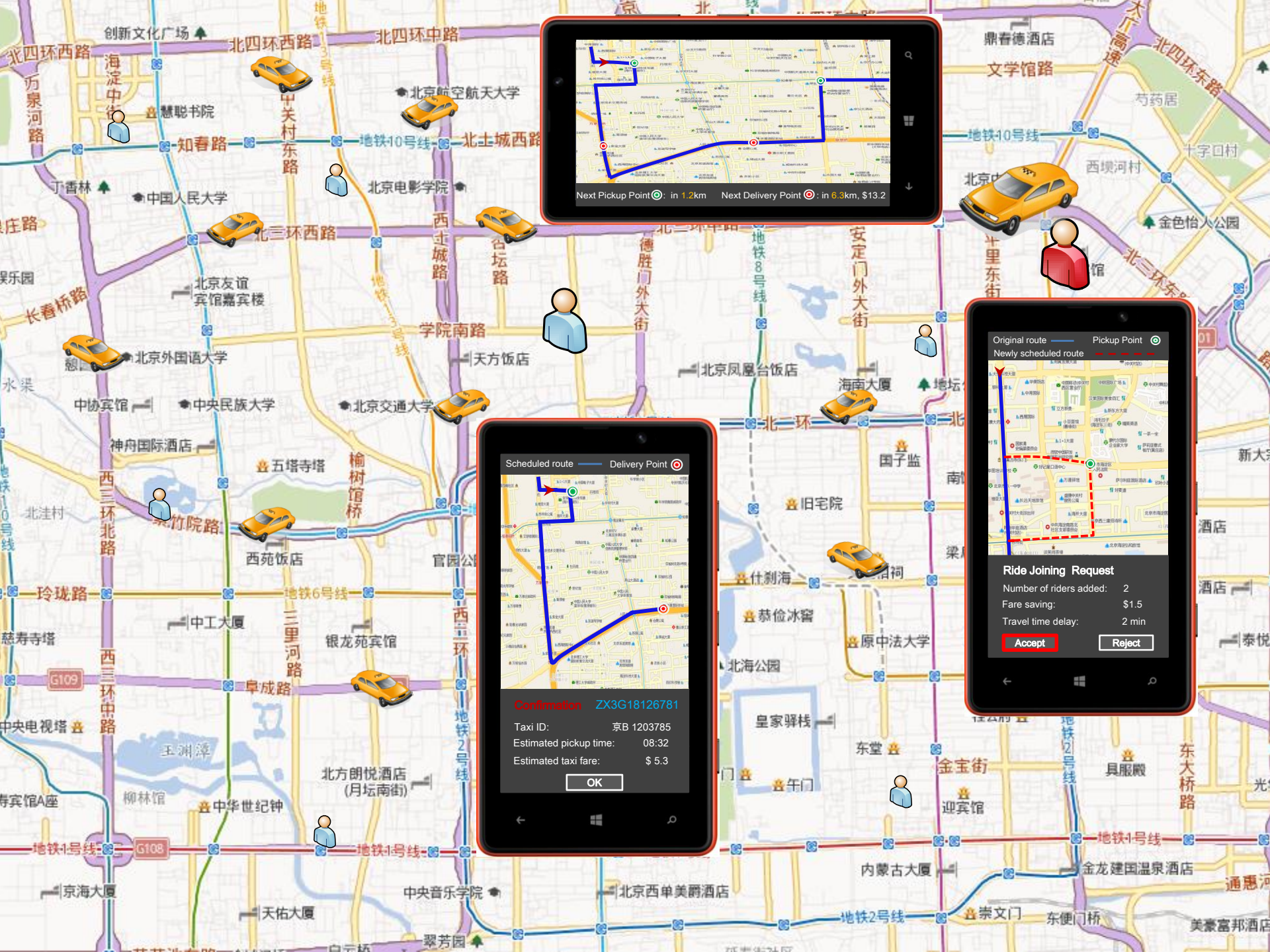
Best Paper Runner up Award at ICDE 2013

Difficult to take a taxi!

- A problem among passengers, taxi drivers and government
- Possible Solutions
  - Increasing taxis？
  - Taxi dispatching？ ？
- There are quite a few seats left in a taxi

Taxi Ridesharing




Next Pickup Point: in 1.2km
Next Delivery Point: in 6.3km, $13.2
Scheduled route
Delivery Point
Confirmation
ZX3G18126781
Taxi ID:
京B 1203785
Estimated pickup time:
08:32
Estimated taxi fare:
$ 5.3
OK
Original route
Pickup Point
Newly scheduled route
Ride Joining Request
Number of riders added:
2
Fare saving:
$1.5
Travel time delay:
2 min
Accept
Reject

Problem Definition

- Query Q=< $Q.o$, $Q.d$, $Q.wp$, $Q.wd$, n >
  - Origin and destination: $Q.o$ and $Q.d$
  - Pickup time: $Q.wp$
  - Delivery time: $Q.wd$

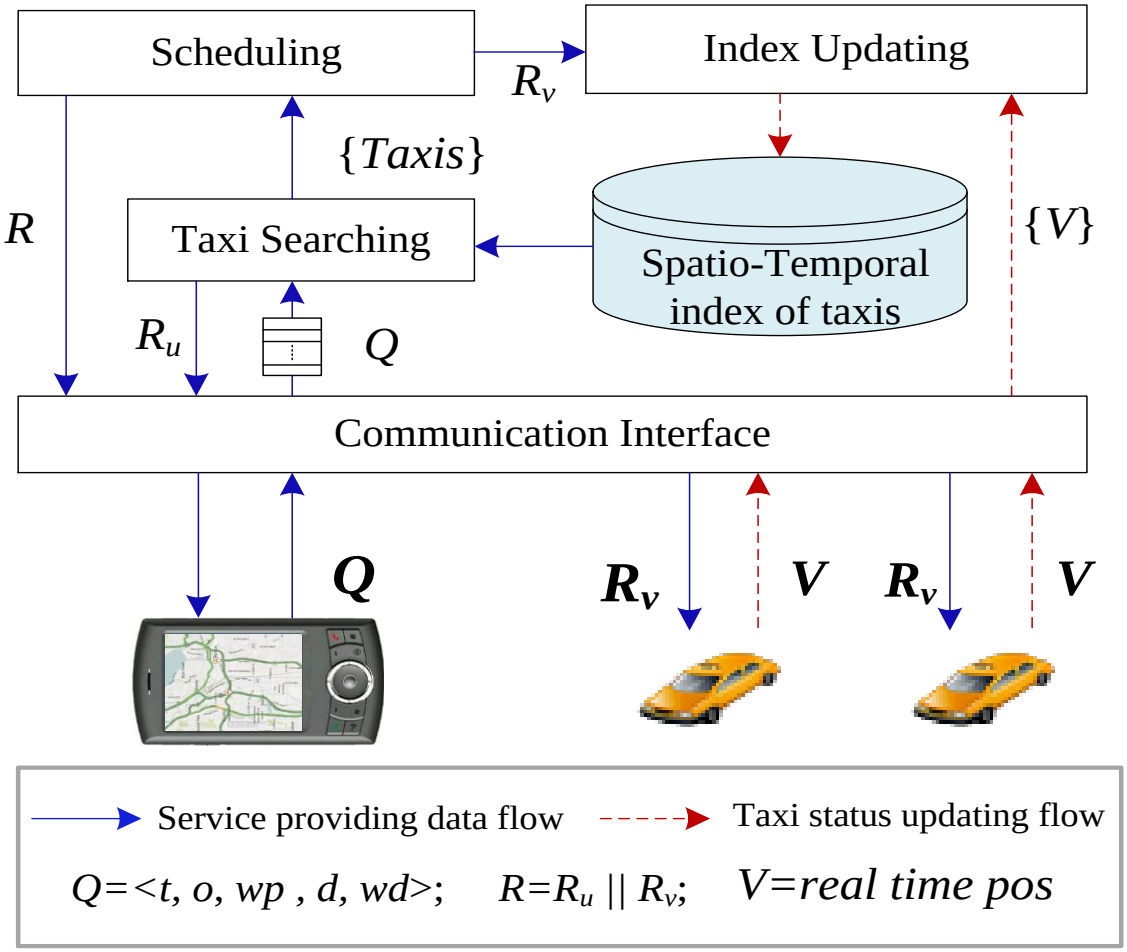
Given a fixed number of taxis traveling on a road network and a stream of queries, we aim to serve each query Q in the stream by dispatching the taxi which

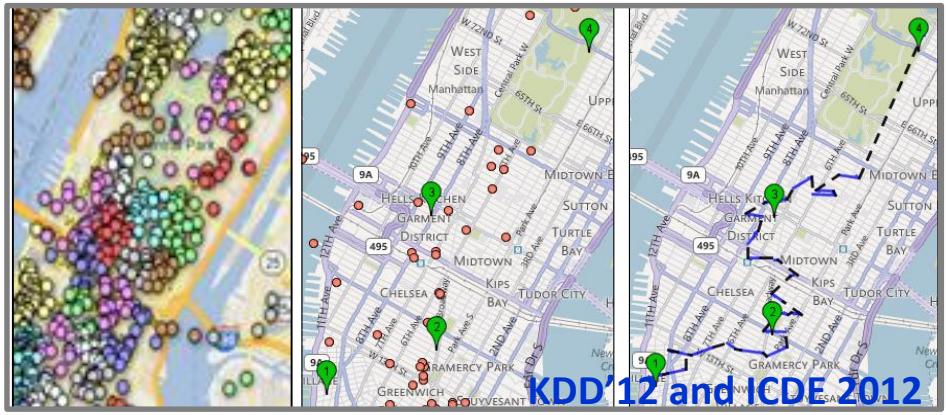
- *satisfies schedule constraint, capacity constraint of a taxi, and monetary constraint*
- *with the minimum increase in travel distance*.

Value

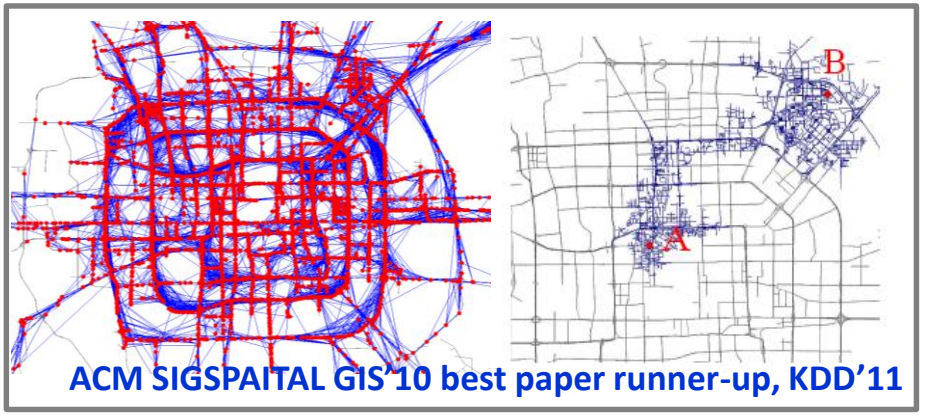
- **Government**
  - Save 120 million liter gasoline per year
    - Supporting 1M cars for 1.5 months
    - Worth about 150 million USD
    - 246 million KG CO2 emission
- **Passengers**
  - Serving rate increased 300%
  - Save 7% expense on average
- **Taxi drivers** increase profit 10% on average

Architecture

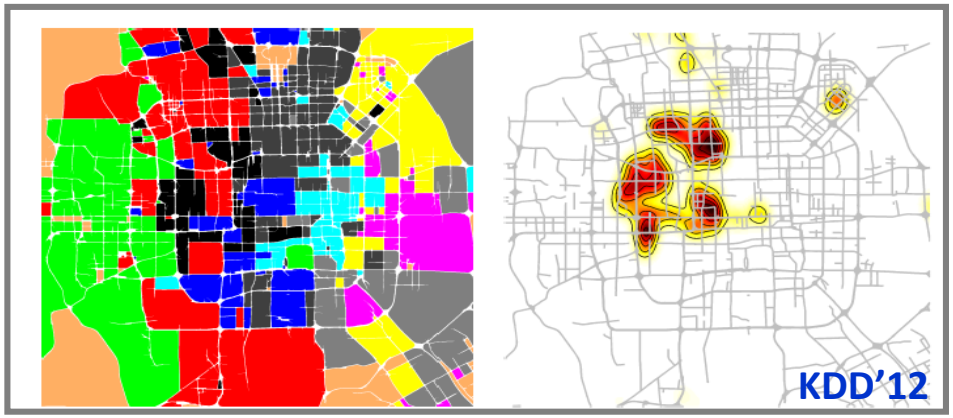




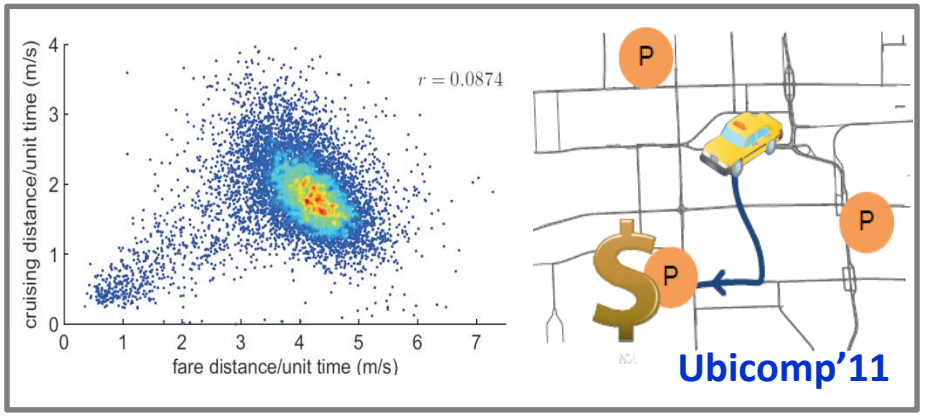
Route Construction from Uncertain Trajectories



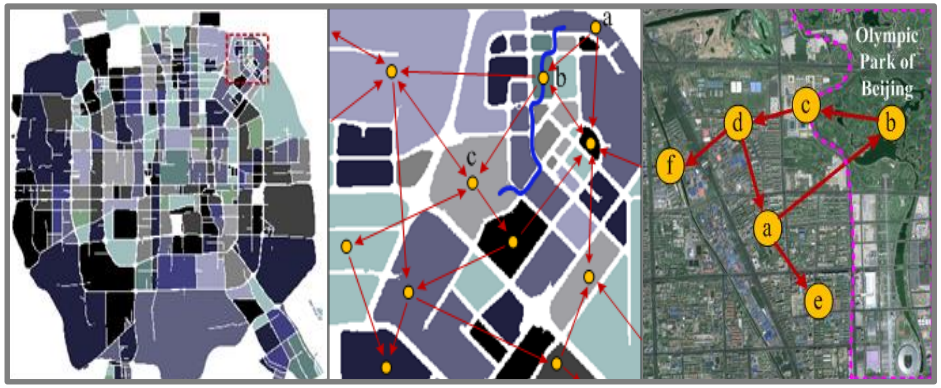
Finding Smart Driving Directions



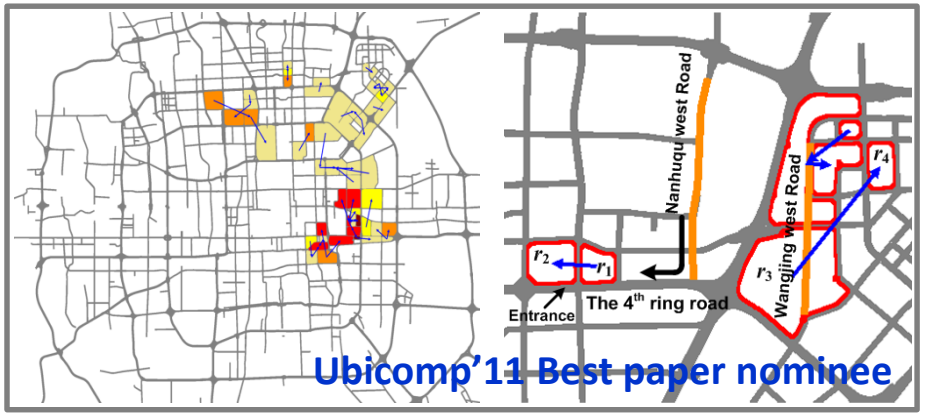
Discovery of Functional Regions



Passengers-Cabbie Recommender system



Anomalous Events Detection KDD'11 and ICDM 2012



Urban Computing for Urban Planning

Take Away Messages

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