Welcome to

DS504/CS586: Big Data Analytics
Application I

Prof. Yanhua Li

Time: 6:00pm –8:50pm R
Location: AK 232
Fall 2016

- 18 critiques & Next Thur we have the last critique.
  - Already graded 8 of them.
  - Plan to grade 1-2 more.

- Grading
  - Projects (40%)
    - Project 1 (10%)
    - Project 2 (30%)
      - Final reports in the discussion forum (by 11:59pm 12/13);
      - Self-and-peer evaluation form for project 2 (by 11:59PM 12/13);
  - Written work (30%):
    - Critiques + Project reports (20%)
    - Quiz (10%, with 5% each)
  - Oral work (30%):
    - Presentation

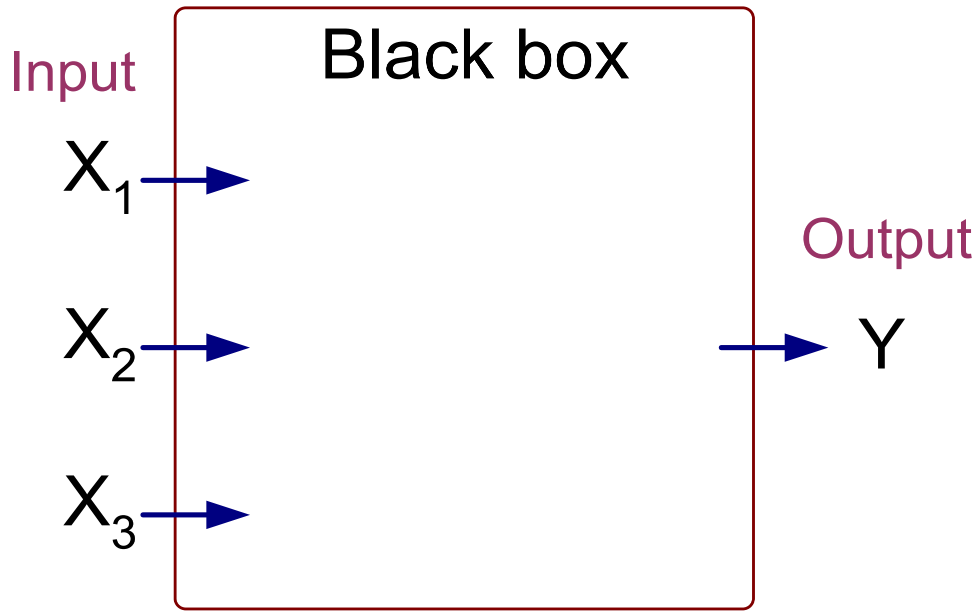
- Final Project Presentation
  - 22 minutes each group (including Q&A)
  - Schedule:
    - 12/15 Thu

Next Class: Summary and Discussion

- Review of the semester
- Plus the last critique/review

Artificial Neural Networks (ANN)

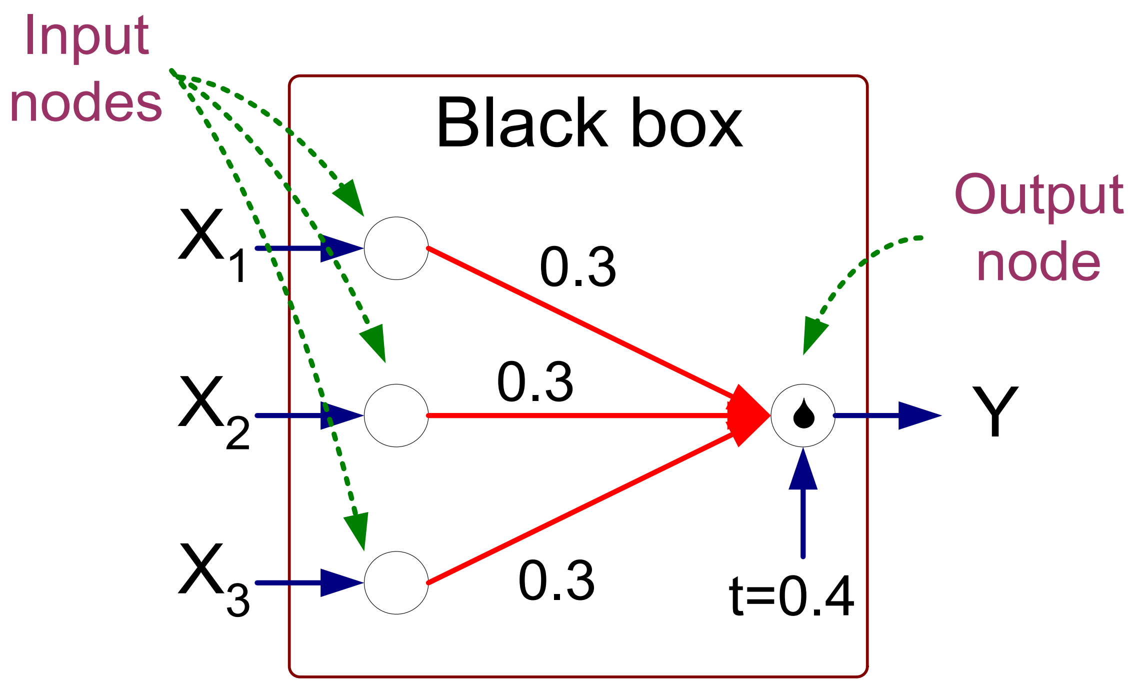
X_1	X_2	X_3	Y
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	1
0	0	1	0
0	1	0	0
0	1	1	1
0	0	0	0



Output Y is 1 if at least two of the three inputs are equal to 1.

Artificial Neural Networks (ANN)

X_1	X_2	X_3	Y
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	1
0	0	1	0
0	1	0	0
0	1	1	1
0	0	0	0

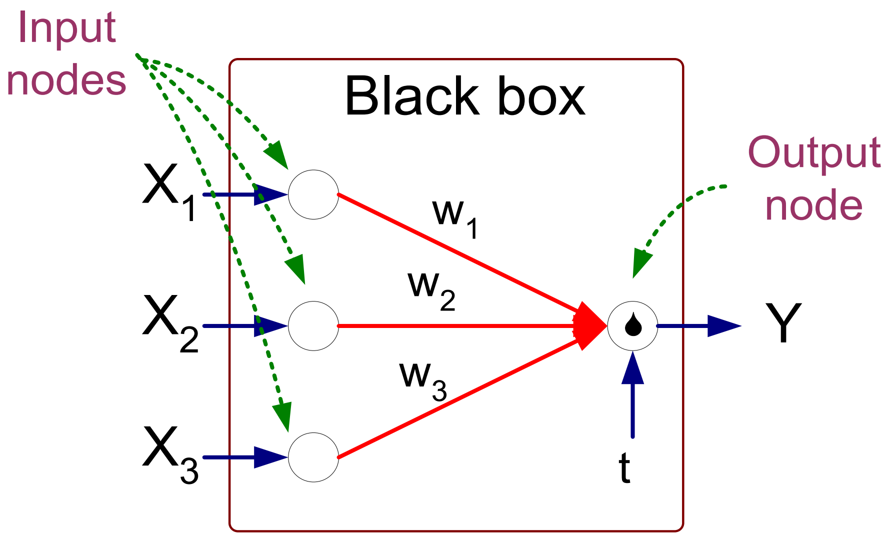


$$Y = I(0.3X_1 + 0.3X_2 + 0.3X_3 - 0.4 > 0)$$

$$\text{where } I(z) = \begin{cases} 1 & \text{if } z \text{ is true} \\ 0 & \text{otherwise} \end{cases}$$

Artificial Neural Networks (ANN)

- Model is an assembly of inter-connected nodes and weighted links
- Output node sums up each of its input value according to the weights of its links
- Compare output node against some threshold t

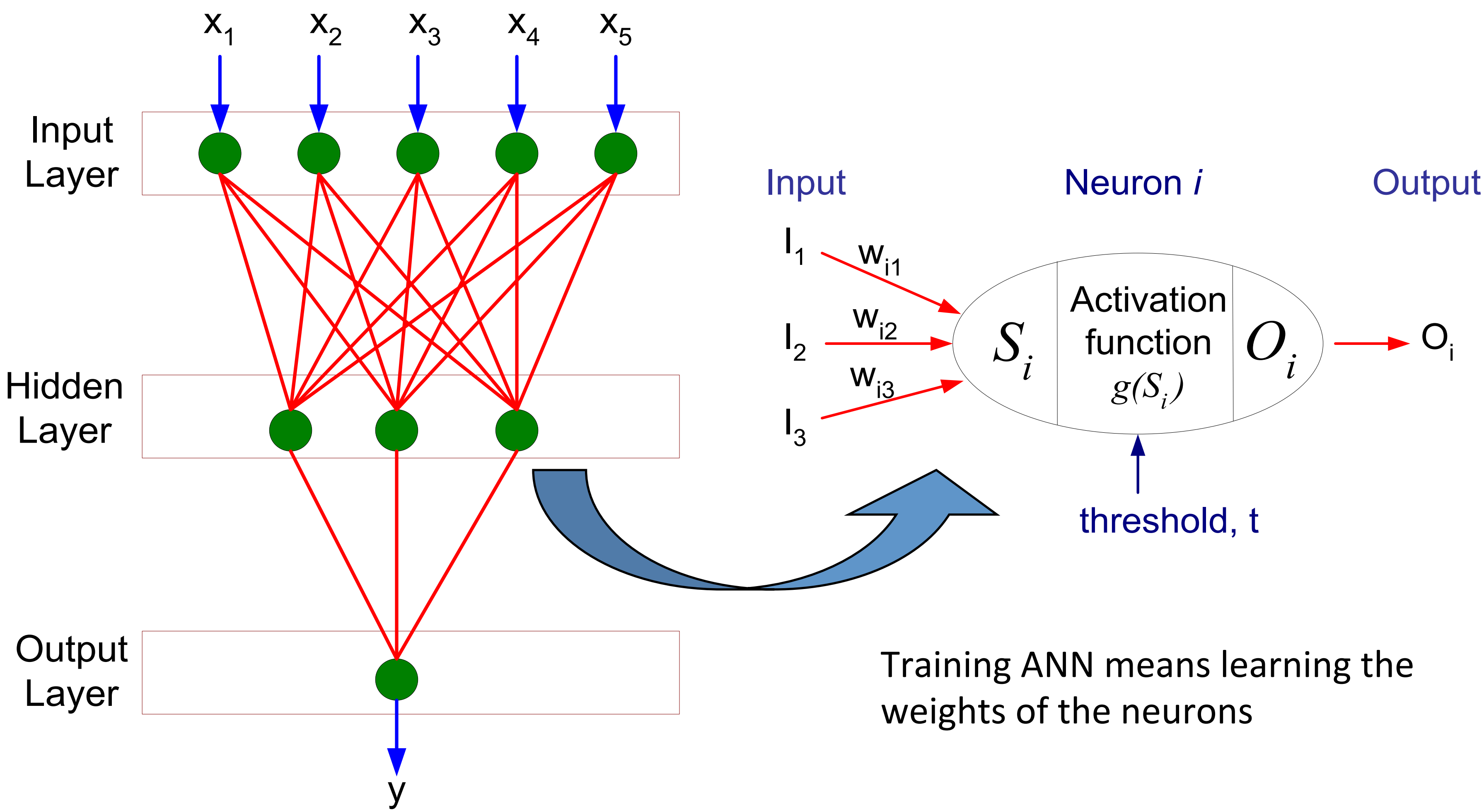


Perceptron Model

$$Y = I\left(\sum_i w_i X_i - t\right) \quad \text{or}$$

$$Y = sign\left(\sum_i w_i X_i - t\right)$$

General Structure of ANN



Training ANN means learning the weights of the neurons

Real-world problems are always messy

- Multiple models
- Key features
- Data Sparsity

- What do we do to solve a classification/inference/ prediction problem?
  - Data Cleaning
  - Feature selection
  - Inference model
  - Evaluation
- An example of how to solve real world application problem

U-Air: When Urban Air Quality Meets Big Data

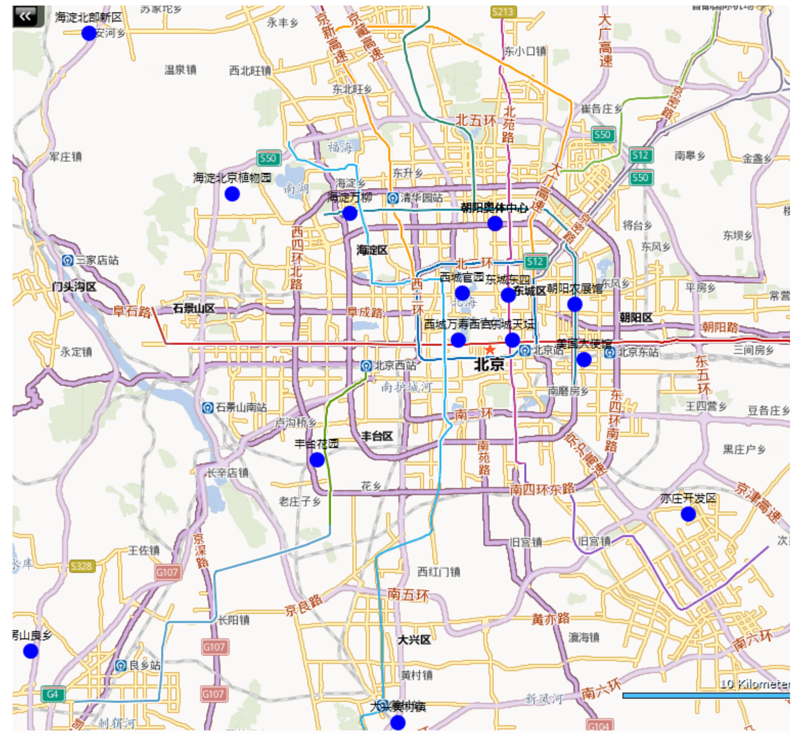
Authors: Yu Zheng, Microsoft Research Asia

U-AIR

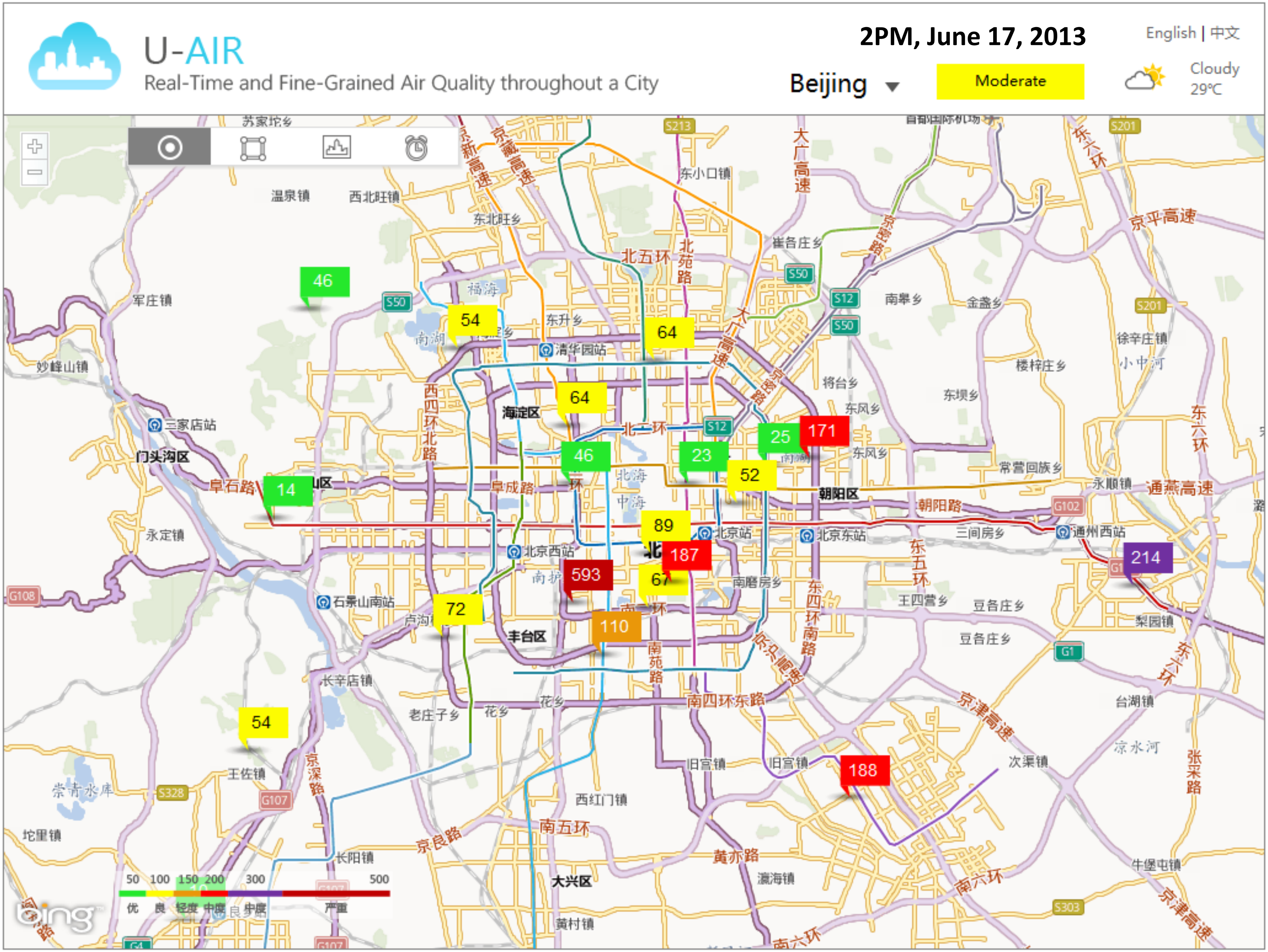
Real-Time and Fine-Grained Air Quality throughout a City

Background

- Air quality
  - NO2, SO2
  - Aerosols: PM2.5, PM10
- Why it matters
  - Healthcare
  - Pollution control and dispersal
- Reality
  - Building a measurement station is not easy
  - A limited number of stations (poor coverage)

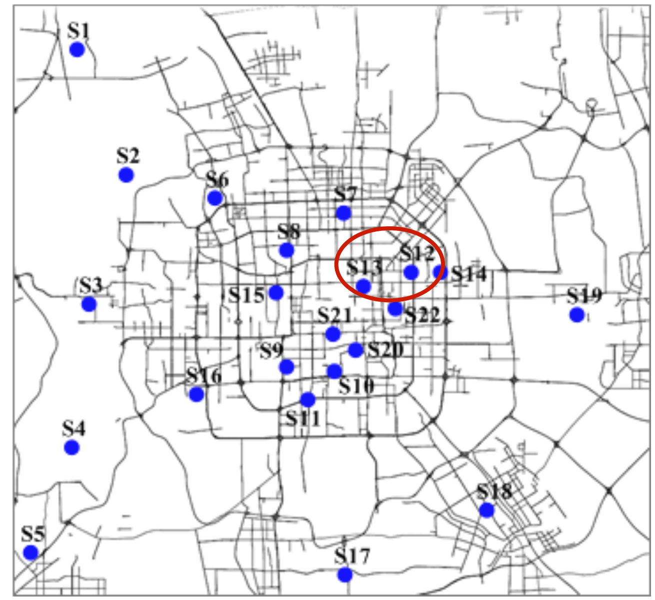


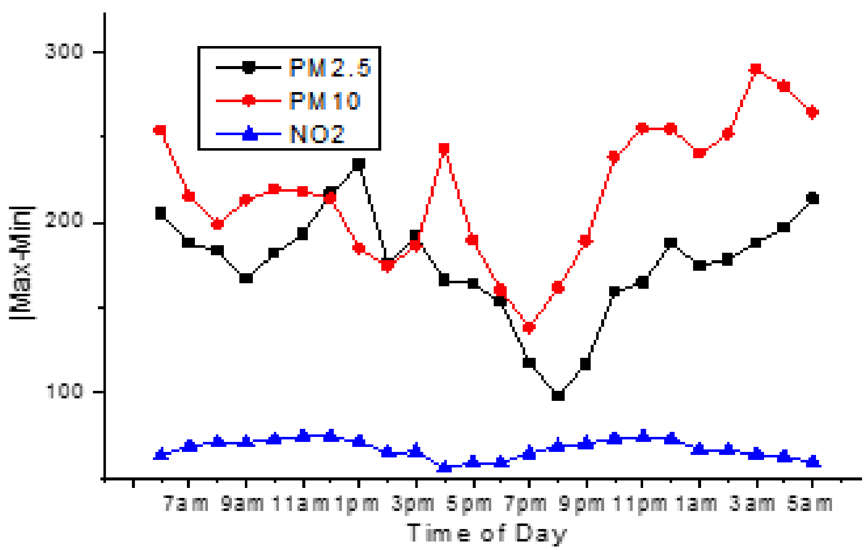
Beijing only has 22 air quality monitor stations in its urban areas (50kmx40km)


U-AIR
Real-Time and Fine-Grained Air Quality throughout a City
2PM, June 17, 2013
English | 中文
Beijing
Moderate
Cloudy
29°C
46
54
64
64
46
14
23
25
171
52
89
187
593
67
72
110
214
54
188
50 100 150 200 300 500
优 良 轻度 中度 重度 严重
海淀区
丰台区
朝阳区
门头沟区
大兴区
北五环
北京西站
北京站
北京东站
清华园站
石景山南站
通州西站
京平高速
通燕高速
京津高速
南四环东路
南五环
南六环

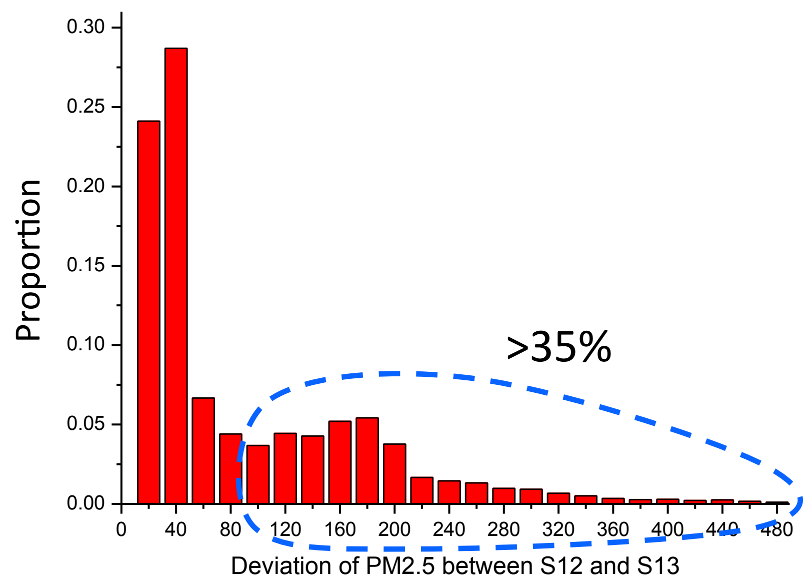
Challenges

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





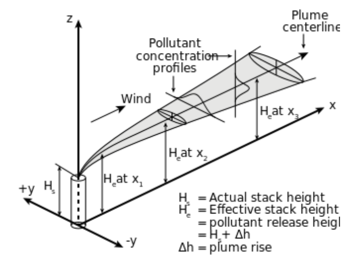
A) Beijing (8/24/2012 - 3/8/2013)



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

Challenges

- Existing methods do not work well
  - Linear interpolation
  - Classical dispersion models
    - Gaussian Plume models and Operational Street Canyon models
    - Many parameters difficult to obtain: Vehicle emission rates, street geometry, the roughness coefficient of the urban surface…

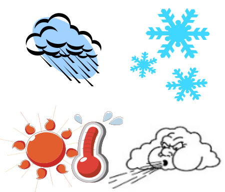


  - Satellite remote sensing
    - Suffer from clouds
    - Does not reflect ground air quality
    - Vary in humidity, temperature, location, and seasons

  - Outsourced crowd sensing using portable devices
    - Limited to a few gasses: CO2 and CO
    - Sensors for detecting aerosol are not portable: PM10, PM2.5
    - A long period of sensing process, 1-2 hours

30,000 + USD, 10ug/m^3
202×85×168（mm）

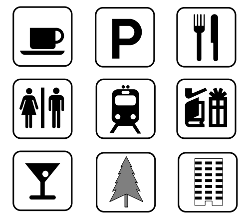
Inferring **Real-Time** and **Fine-Grained** air quality throughout a city using **Big Data**



Meteorology

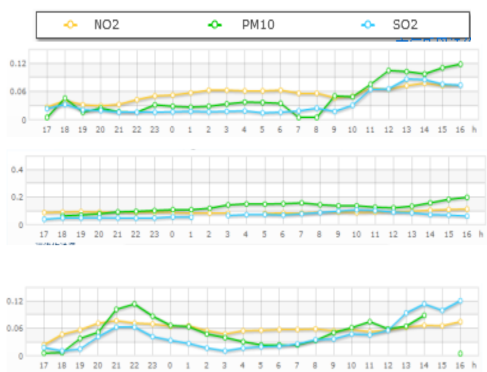
Traffic

Human Mobility

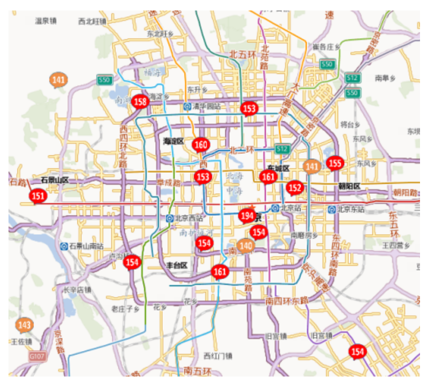


POIs

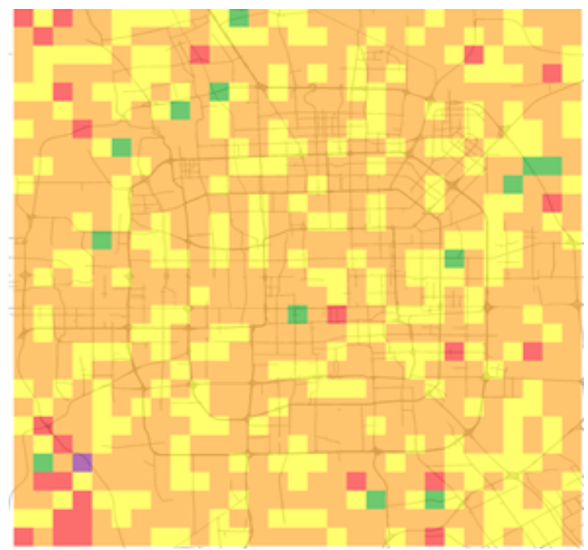
Road networks

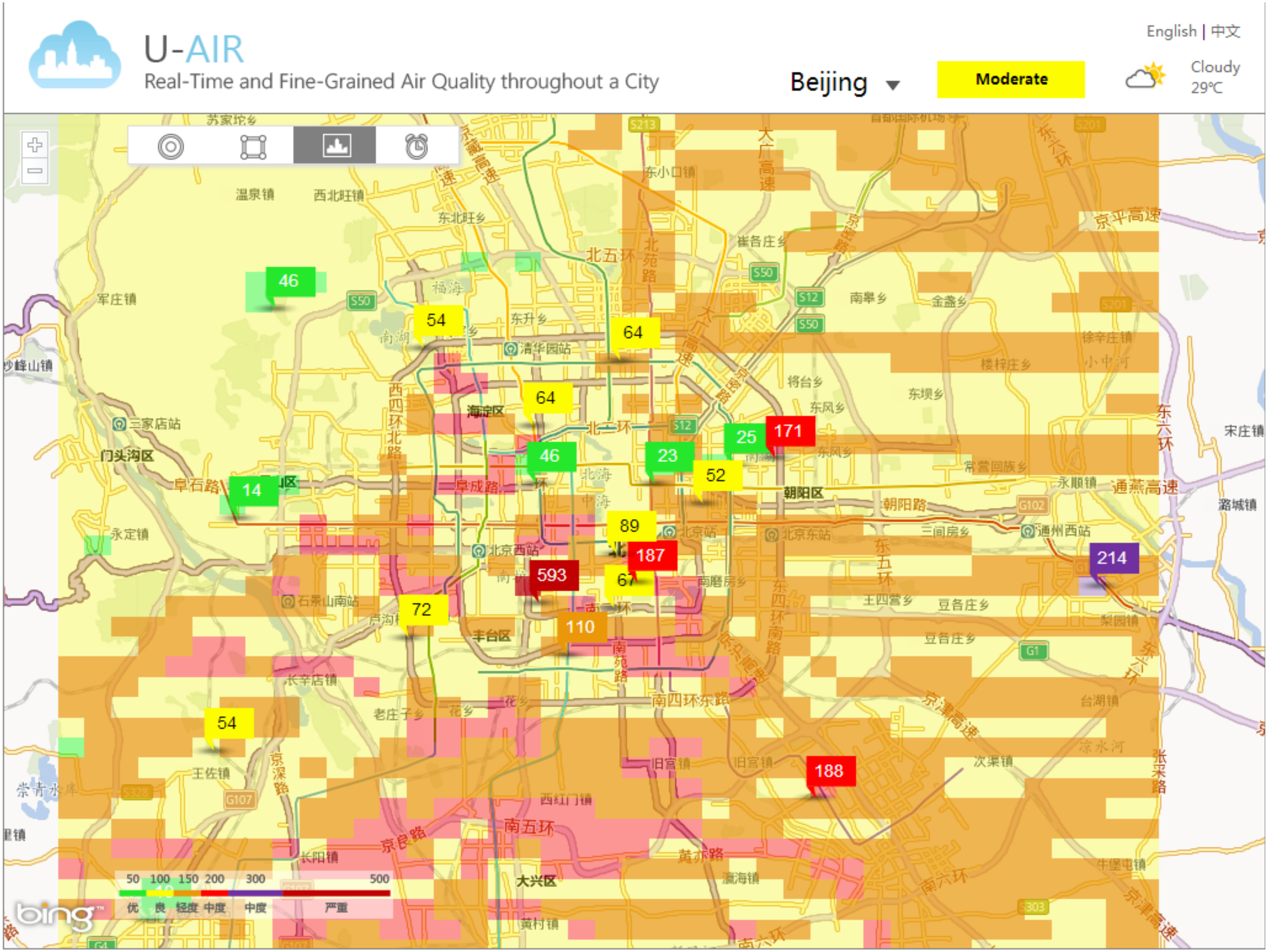


Historical air quality data



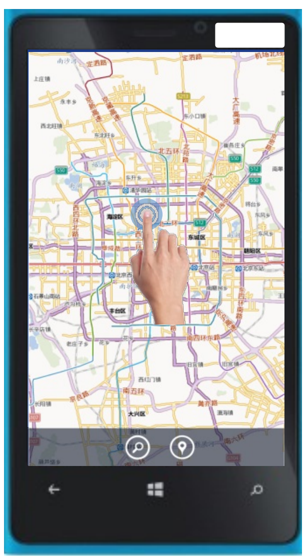
Real-time air quality reports

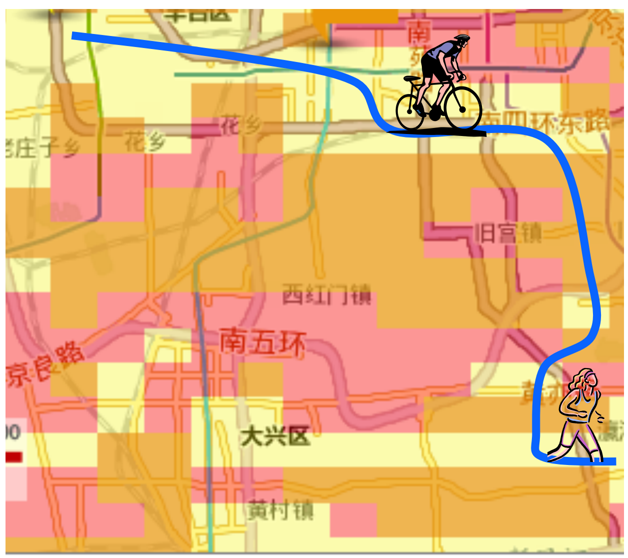


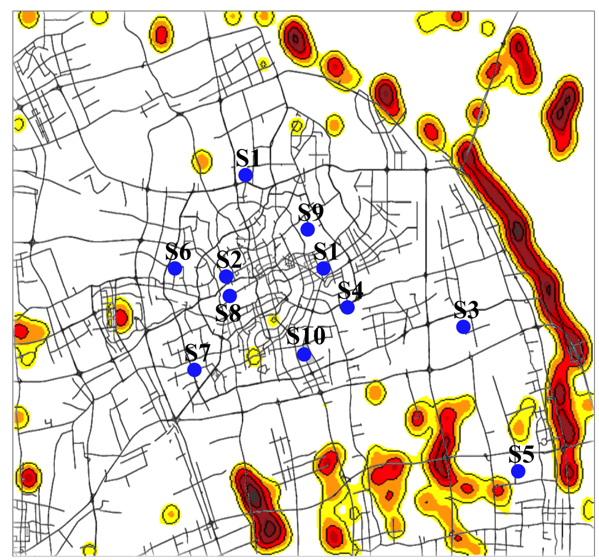

U-AIR
Real-Time and Fine-Grained Air Quality throughout a City
English | 中文
Beijing
Moderate
Cloudy
29℃
46
54
64
64
46
14
23
25
171
52
89
187
593
67
214
72
110
54
188
50 100 150 200 300 500
优 良 轻度 中度 中度 严重
海淀区
朝阳区
丰台区
大兴区
门头沟区
北京西站
北京站
北京东站
清华园站
三家店站
石景山南站
通州西站
北五环
京平高速
通燕高速
京津高速
南五环
南六环
南四环东路
阜成路
朝阳路
京良路
温泉镇
西北旺镇
军庄镇
永定镇
长辛店镇
王佐镇
长阳镇
西红门镇
黄村镇
旧宫镇
次渠镇
台湖镇
宋庄镇
潞城镇
永顺镇
梨园镇
豆各庄乡
南皋乡
金盏乡
崔各庄乡
东小口镇
bing

Applications

- Location-based air quality awareness
  - Fine-grained pollution alert
  - Routing based on air quality
- Deploying new monitoring stations
- A step towards identifying the root cause of air pollution





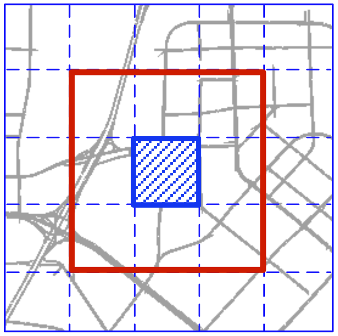


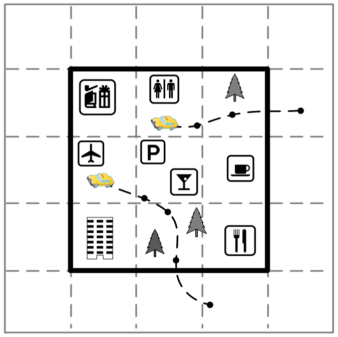
Difficulties

- 1. How to identify features from each kind of data source
- 2. Incorporate multiple heterogeneous data sources into a learning model
  - Spatially-related data: POIs, road networks
  - Temporally-related data: traffic, meteorology, human mobility
- 3. Data sparseness (little training data)
  - Limited number of stations
  - Many places to infer

Methodology Overview

- Partition a city into disjoint grids
- Extract features for each grid from its affecting region
  - Meteorological features
  - Traffic features
  - Human mobility features
  - POI features
  - Road network features



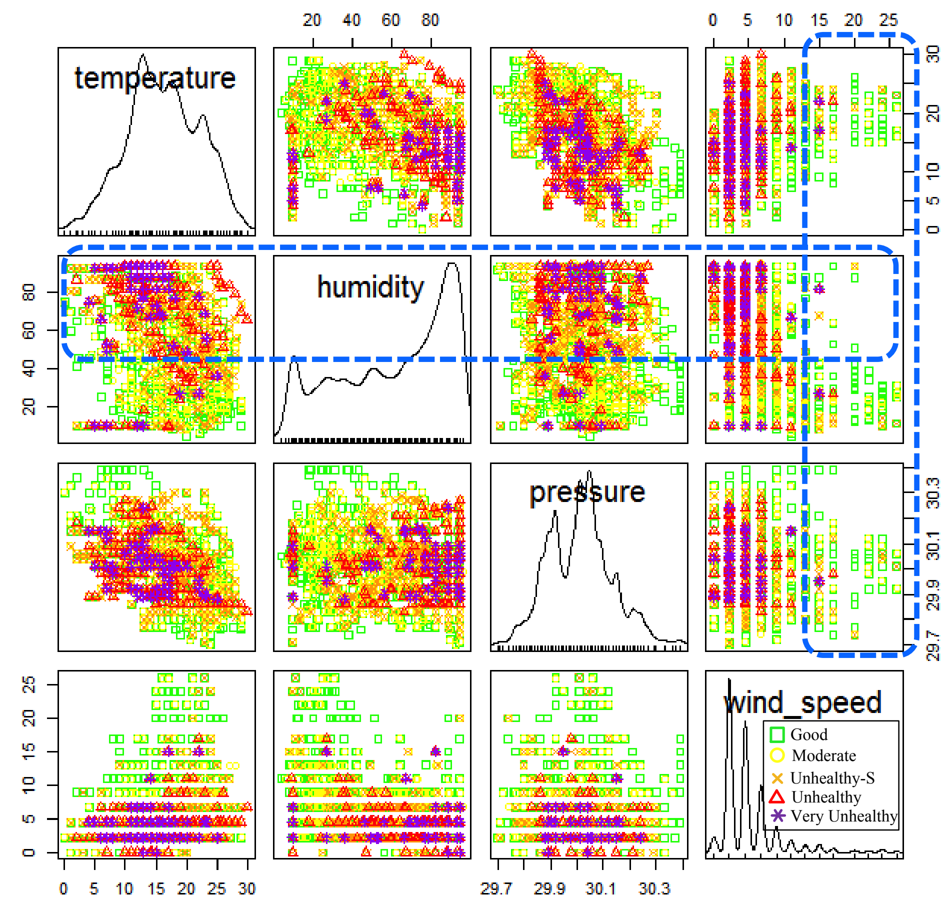


- Co-training-based semi-supervised learning model for each pollutant
  - Predict the AQI labels
  - Data sparsity
  - Two classifiers

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

Meteorological Features: F_m

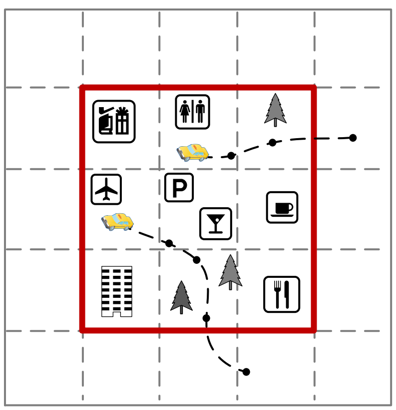
- Rainy, Sunny, Cloudy, Foggy
- Wind speed
- Temperature
- Humidity
- Barometer pressure



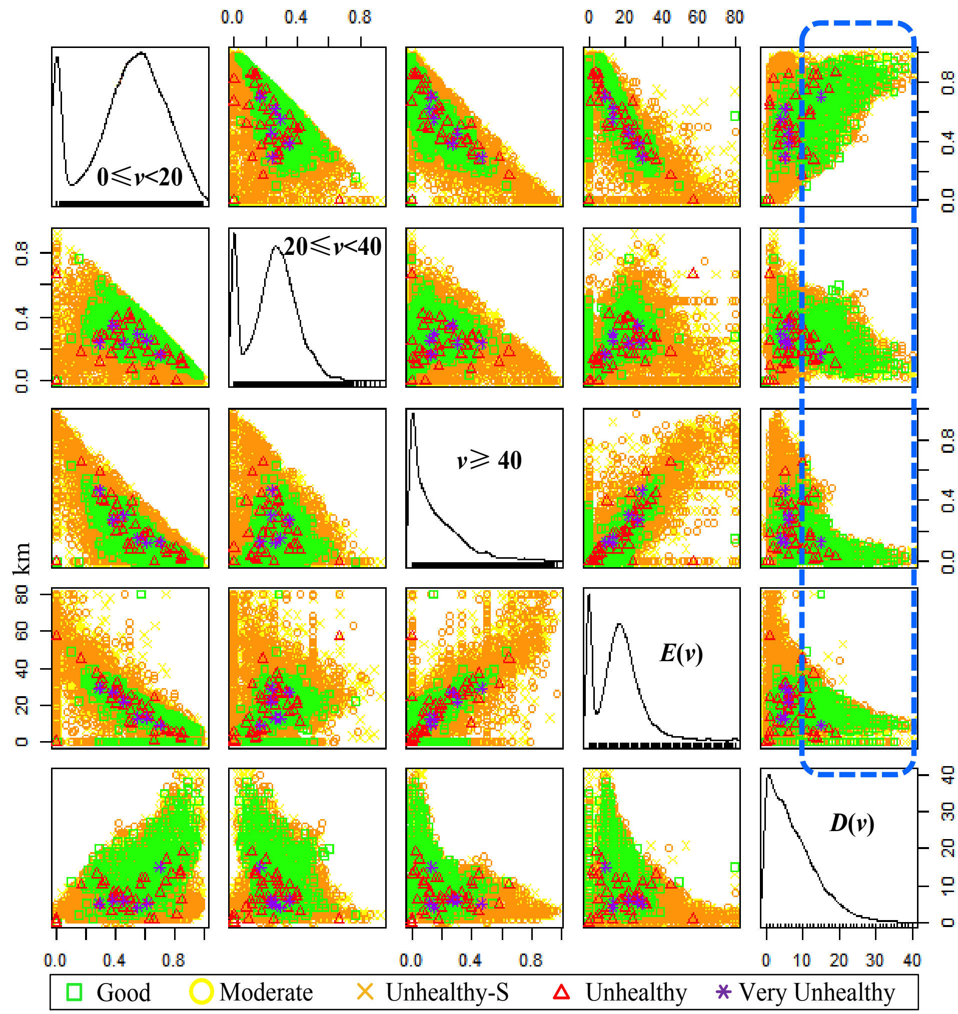
AQI of PM_{10}
August to Dec. 2012 in Beijing

Traffic Features: F_t

- Distribution of speed by time: F(v)
- Expectation of speed: E(V)
- Standard deviation of Speed: D

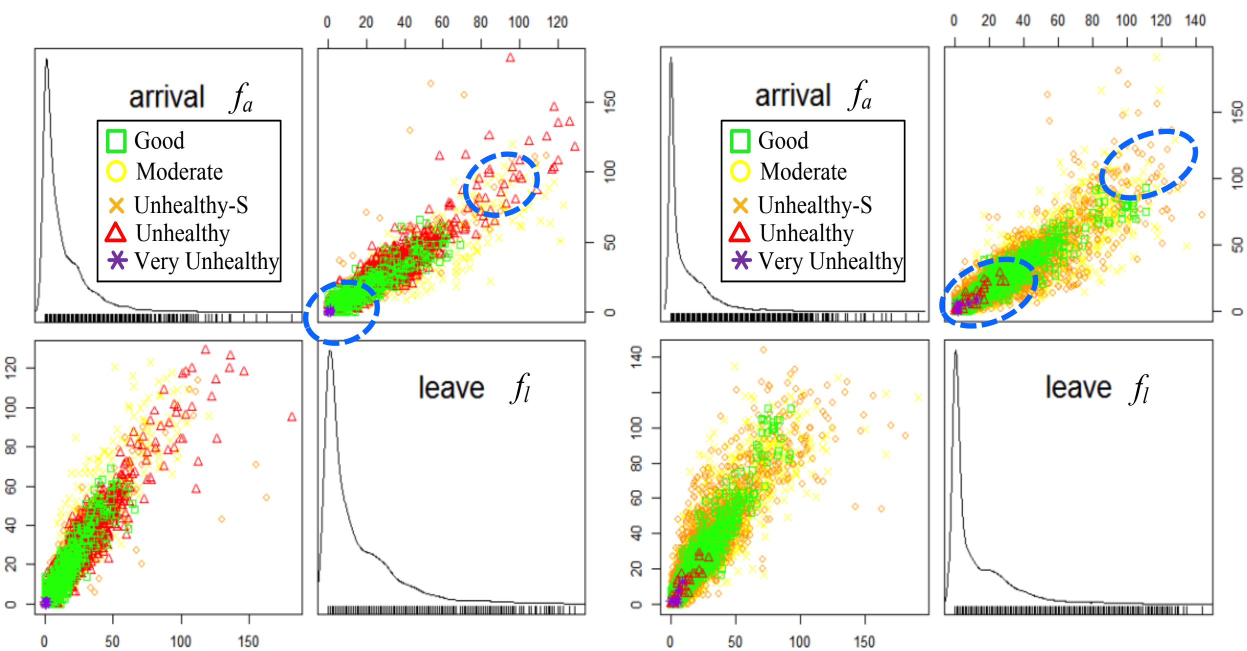


GPS trajectories generated by over 30,000 taxis
From August to Dec. 2012 in Beijing



Human Mobility Features: F_h

- Human mobility implies
  - Traffic flow
  - Land use of a location
  - Function of a region (like residential or business areas)
- Features: Number of arrivals f_a and leavings (departures) f_l

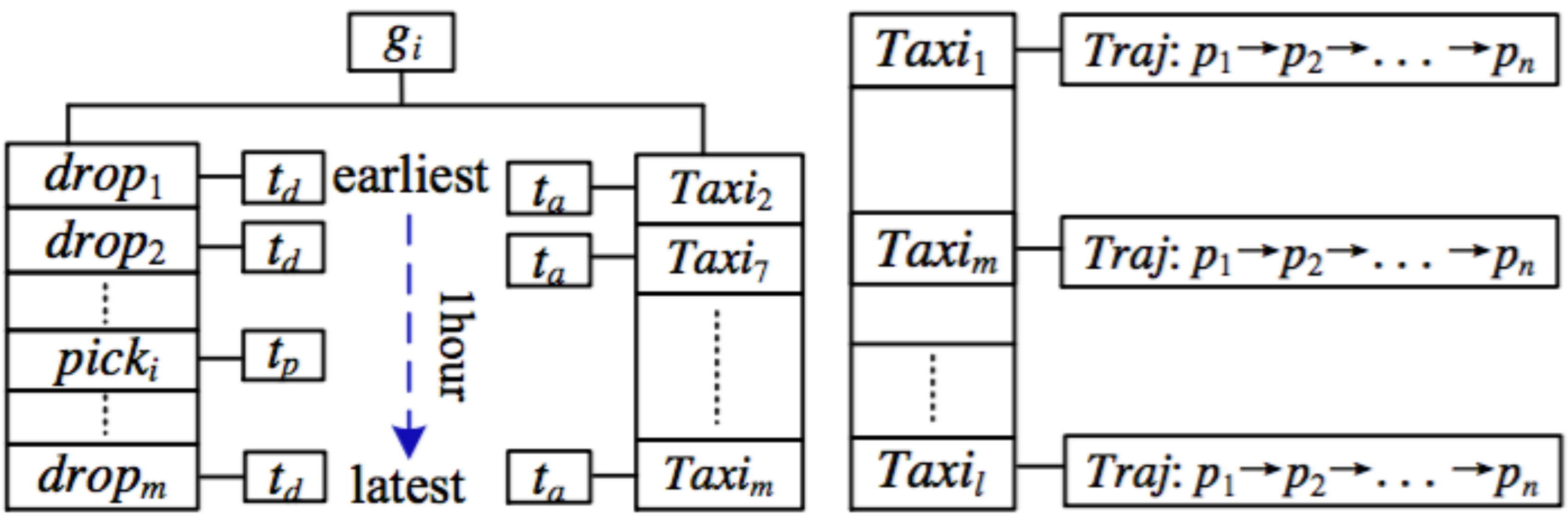


A) AQI of PM_{10} B) AQI of NO_2

Parks vs factories

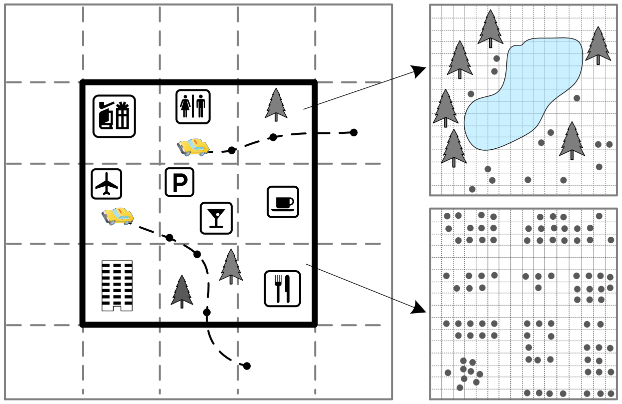
Extracting Traffic/Human Mobility Features

- Offline spatio-temporal indexing
- t_a: arrival time
- *Traj*: traj ID
- I_i: the index of the first GPS point (in the trajectory) entering a grid
- I_o: the index of the last GPS point (in the trajectory) entering the grid



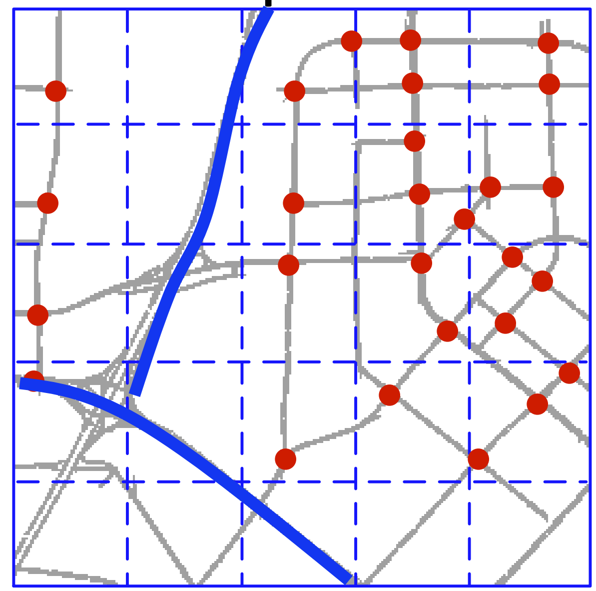
POI Features: F_p

- Why POI
  - Indicate the land use and the function of the region
  - the traffic patterns in the region
- Features
  - Numbers of POIs over categories
  - Portion of vacant places
  - The changes in the number of POIs
    - Factories, shopping malls,
    - hotel and real estates
    - Parks, decoration and furniture markets



C_1: Vehicle Services (gas stations, repair)	C_7: Sports
C_2: Transportation spots	C_8: Parks
C_3: Factories	C_9: Culture & education
C_4: Decoration and furniture markets	C_{10}: Entertainment
C_5: Food and beverage	C_{11}:Companies
C_6: Shopping malls and Supermarkets	C_{12}:Hotels and real estates

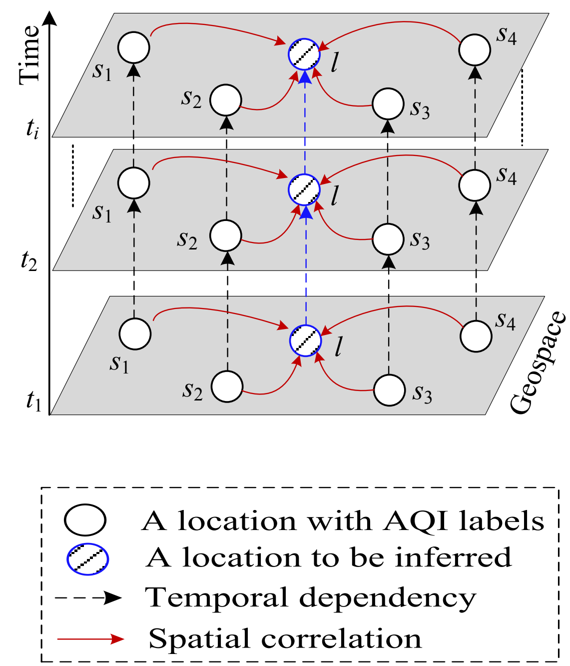
Road Network Features: F_r

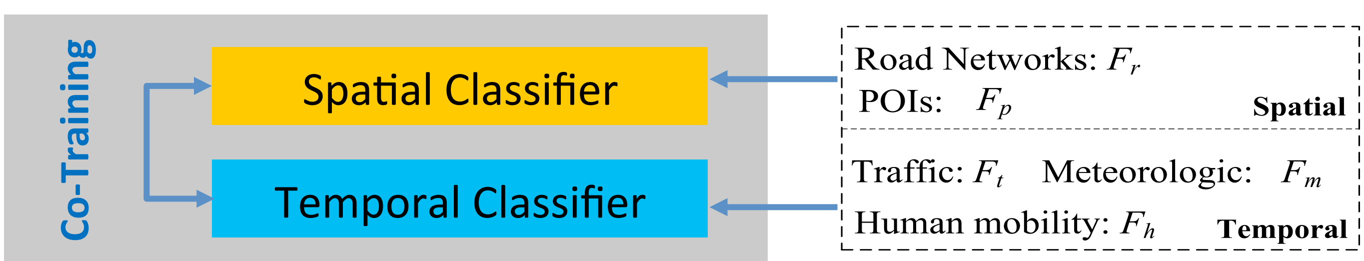


- Why road networks
  - Have a strong correlation with traffic flows
  - A good complementary of traffic modeling
- Features:
  - Total length of highways f_h
  - Total length of other (low-level) road segments f_r
  - The number of intersections f_s in the grid's affecting region

Semi-Supervised Learning Model

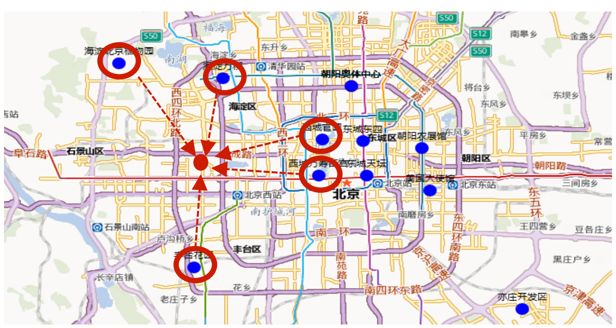
- Philosophy of the model
  - States of air quality
    - Temporal dependency in a location
    - Geo-correlation between locations
  - Generation of air pollutants
    - Emission from a location
    - Propagation among locations
  - Two sets of features
    - Spatially-related
    - Temporally-related

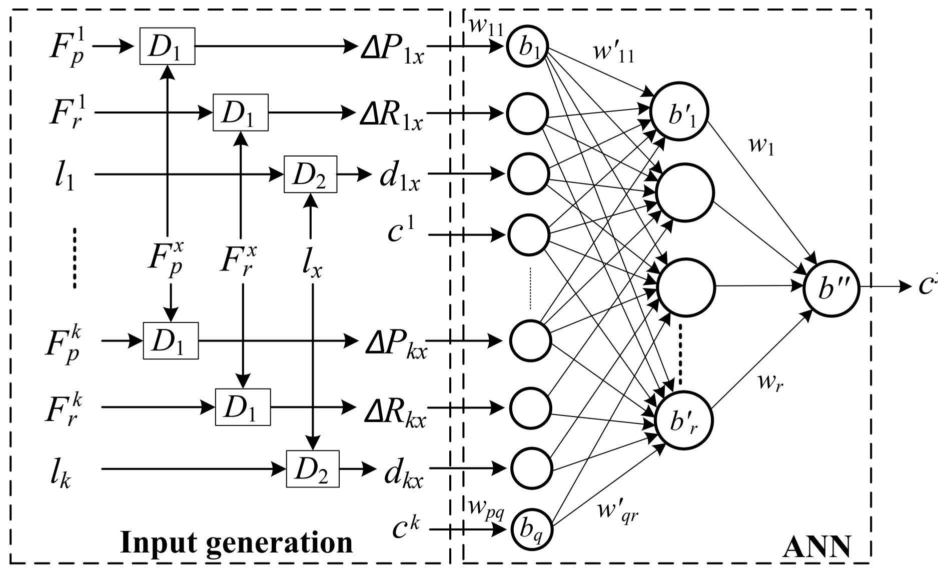




Co-Training-Based Learning Model

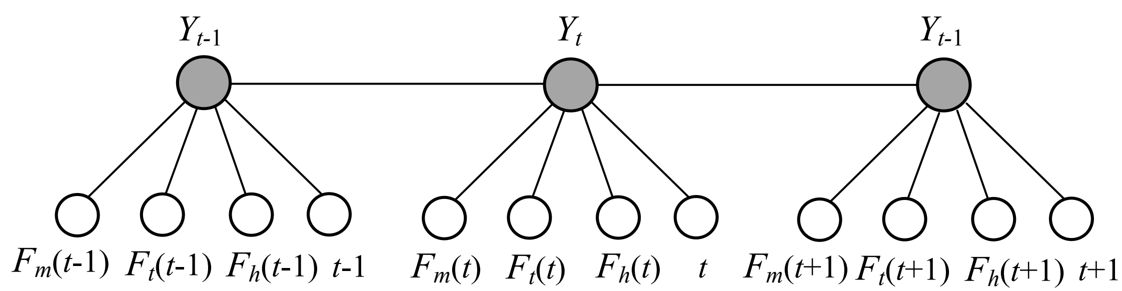
- Spatial classifier
  - Model the spatial correlation between AQI of different locations
  - Using spatially-related features
  - Based on a BP neural network
- Input generation
  - Select *n* stations to pair with
  - Perform *m* rounds



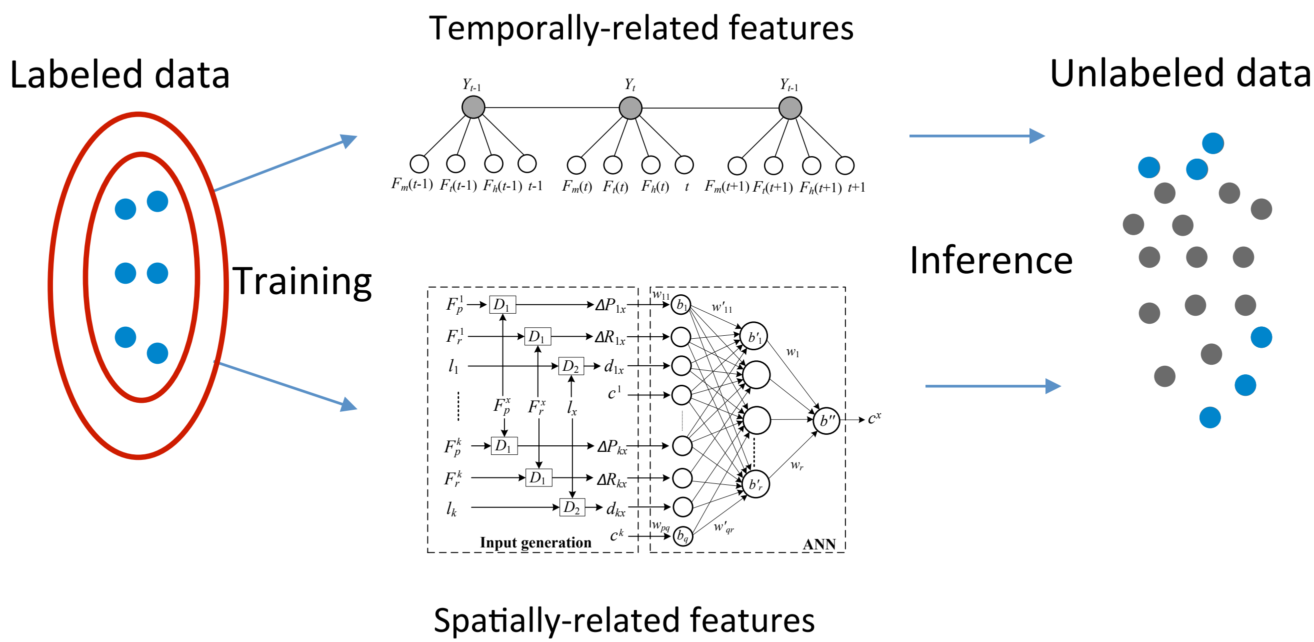


Co-Training-Based Learning Model

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

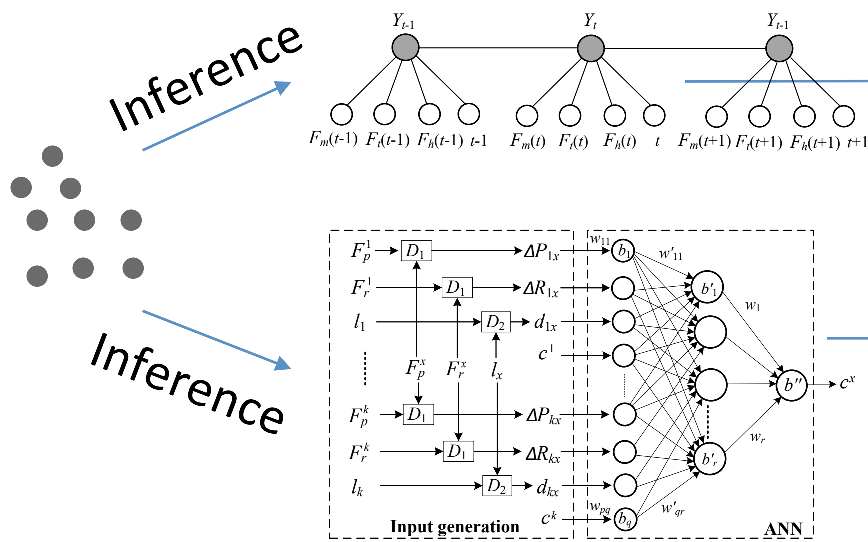


Learning Process



Inference Process

Temporally-related features



$$< p_{c_1}, \cdots, p_{c_n} >$$

$$\times$$

$$c = arg_{c_i \in C} Max(P_{SC}^{c_i} \times P_{TC}^{c_i})$$

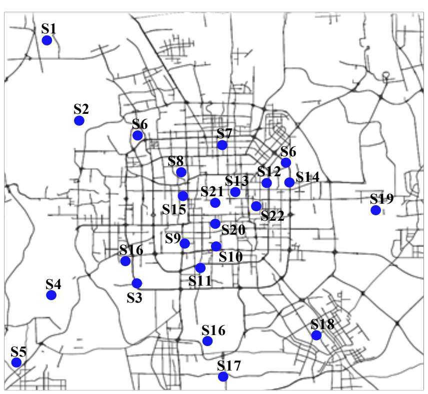
$$< p'_{c_1}, \cdots, p'_{c_n} >$$

Spatially-related features

Evaluation

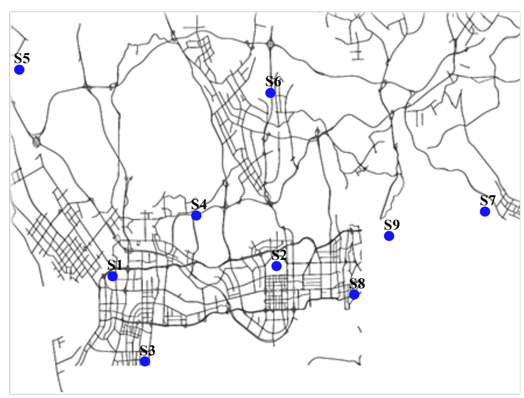
- Datasets

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)		5050km (2500)	5050km (2500)	5745km(2565)	4525km (1165)

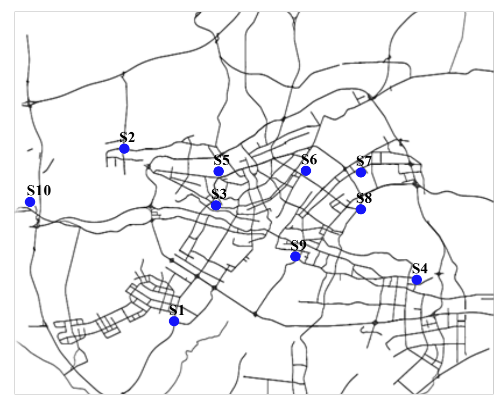


A) Beijing

B) Shanghai



C) Shenzhen



D) Wuhan

Evaluation

- Ground Truth
  - Remove a station
  - Cross cities
- Baselines
  - Linear and Gaussian Interpolations
  - Classical Dispersion Model
  - Decision Tree (DT):
  - CRF-ALL
  - ANN-ALL

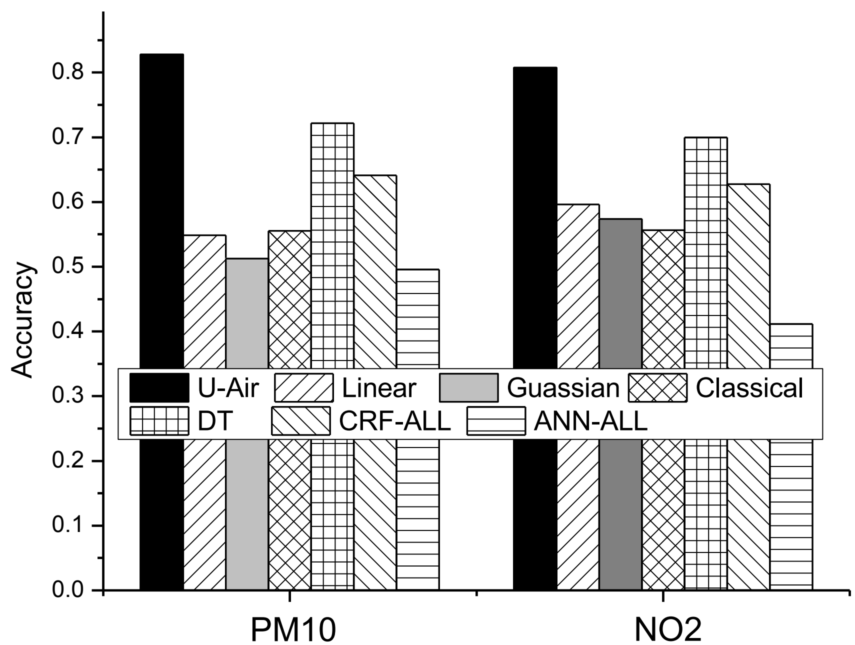
Evaluation

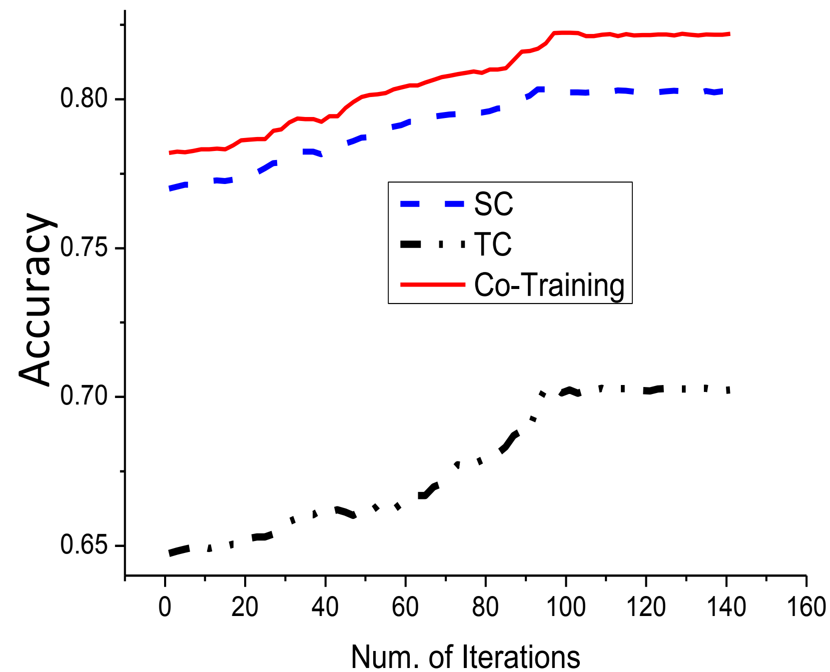
- Does every kind of feature count?

	PM10		NO2	
Features	Precision	Recall	Precision	Recall
Fm	0.572	0.514	0.477	0.454
Ft	0.341	0.36	0.371	0.35
Fh	0.327	0.364	0.411	0.483
Fp+Fr	0.441	0.443	0.307	0.354
Fm+Ft	0.664	0.675	0.634	0.635
Fm+Ft+Fp+Fr	0.731	0.734	0.701	0.691
Fm+Ft+Fp+Fr+Fh	0.773	0.754	0.723	0.704

Evaluation

- Overall performance of the co-training





Evaluation

- Confusion matrix of Co-Training on PM_{10}

Ground Truth	Predictions					
	G	M	S	U		
G	3789	402	102	0	0.883	Recall
M	602	3614	204	0	0.818	
S	41	200	532	50	0.646	
U	0	22	70	219	0.704	
	0.855	0.853	0.586	0.814	0.828	
	Precision					

Evaluation

- Performance of Spatial classifier

Cities	PM2.5		PM10		NO2	
	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
Beijing	0.764	0.763	0.762	0.745	0.730	0.749
Shanghai	0.705	0.725	0.702	0.718	0.715	0.706
Shenzhen	0.740	0.737	0.710	0.742	0.732	0.722
Wuhan	0.727	0.723	0.731	0.739	0.744	0.719

Evaluation

- Efficiency study
- Single grid 131s
- Inferring the AQIs for entire Beijing in 5 minutes

<table>
<tr><th colspan="2">Procedures</th><th>Time(ms)</th><th colspan="2">Procedures</th><th>Time(ms)</th></tr>
<tr><td rowspan="3">Feature extraction (per grid)</td><td>Ft&Fh</td><td>53.2</td><td rowspan="2">Inference (per grid)</td><td>SC</td><td>21.5</td></tr>
<tr><td>F_p</td><td>28.8</td><td>TC</td><td>13.1</td></tr>
<tr><td>F_r</td><td>14.4</td><td colspan="2">Total</td><td>131</td></tr>
</table>

Conclusion

- Infer fine-grained air quality with
  - Real-time and historical air quality readings from existing stations
  - Other data sources: meteorology, POIs, road network, human mobility, and traffic condition
- Co-Training-based semi-supervised learning approach
  - Deal with data sparsity by learning from unlabeled data
  - Model the spatial correlation among the air quality of different locations
  - Model the temporal dependency of the air quality in a location
- Results
  - 0.82 with traffic data (co-training)
  - 0.76 if only using spatial classifier

Questions?