Welcome to

DS504/CS586: Big Data Analytics --Introduction & Logistics

Prof. Yanhua Li

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

What is *DS504/CS586* about?

- A second Level DS/CS course (primarily) for graduates
  - CS/DS Ph.D students in big data analytics and related areas;
  - then other Ph.D students or MS students with
    - Experience in databases and/or in data mining, or equivalent knowledge.
    - Sufficient programming experience is expected so that you are comfortable to undertake a course project.

Big Data Analytics

techniques and tools for managing, analyzing and extracting knowledge from “big data”

Introduction

What is "Big Data"?

Big Data – What is it?

- A “big” buzzword …
- No single standard definition…
- Talk to 1000 people, there will be 1000 “definitions” …

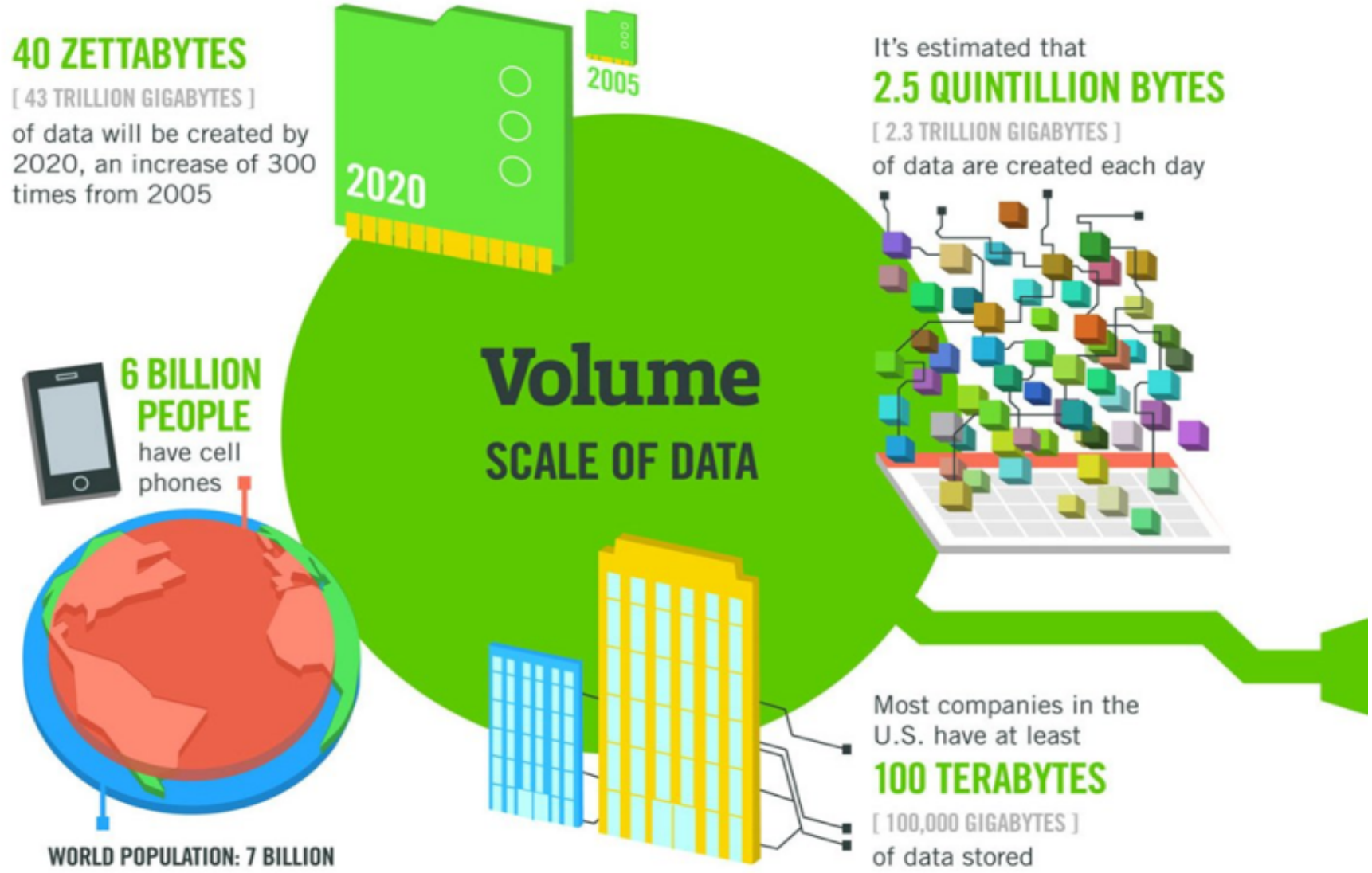
“***Big Data***” is data whose scale, diversity, complexity, and/or quality require new architectures, techniques, algorithms, analytics, and interfaces to manage it and extract value and hidden knowledge from it…

Why Now?

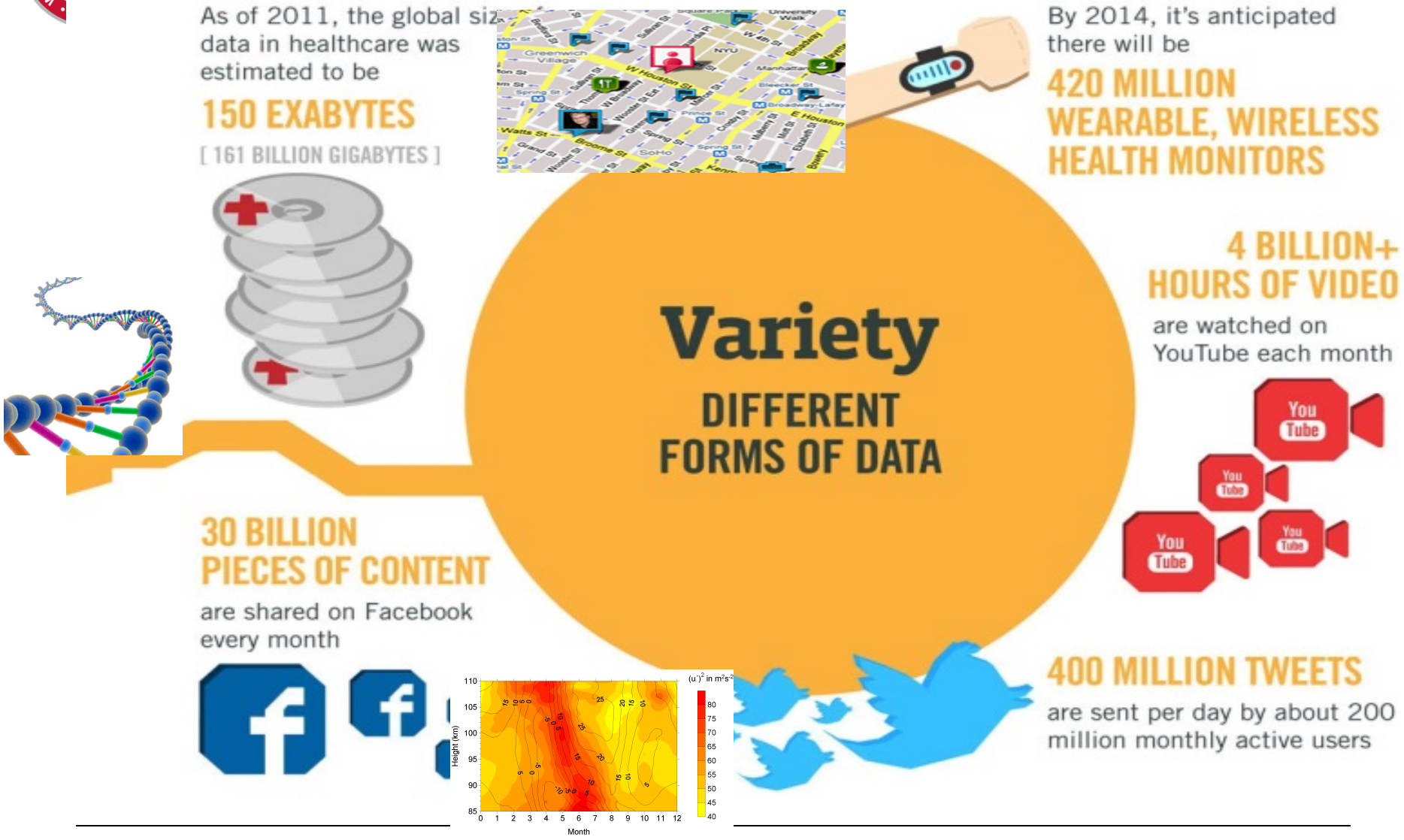
Big Data and Big Challenges

Big Data

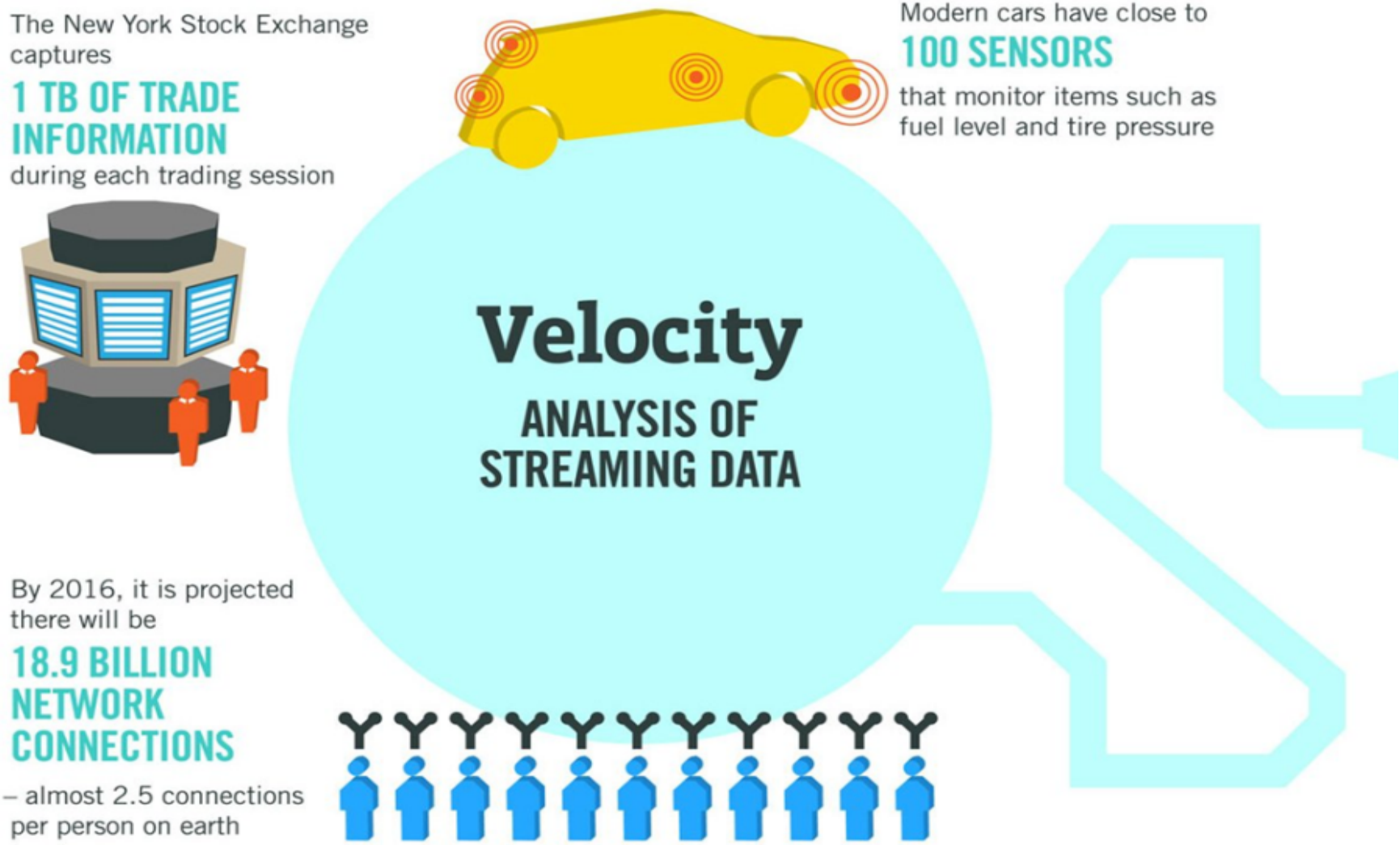
- Volume
- Variety
- Velocity
- Veracity



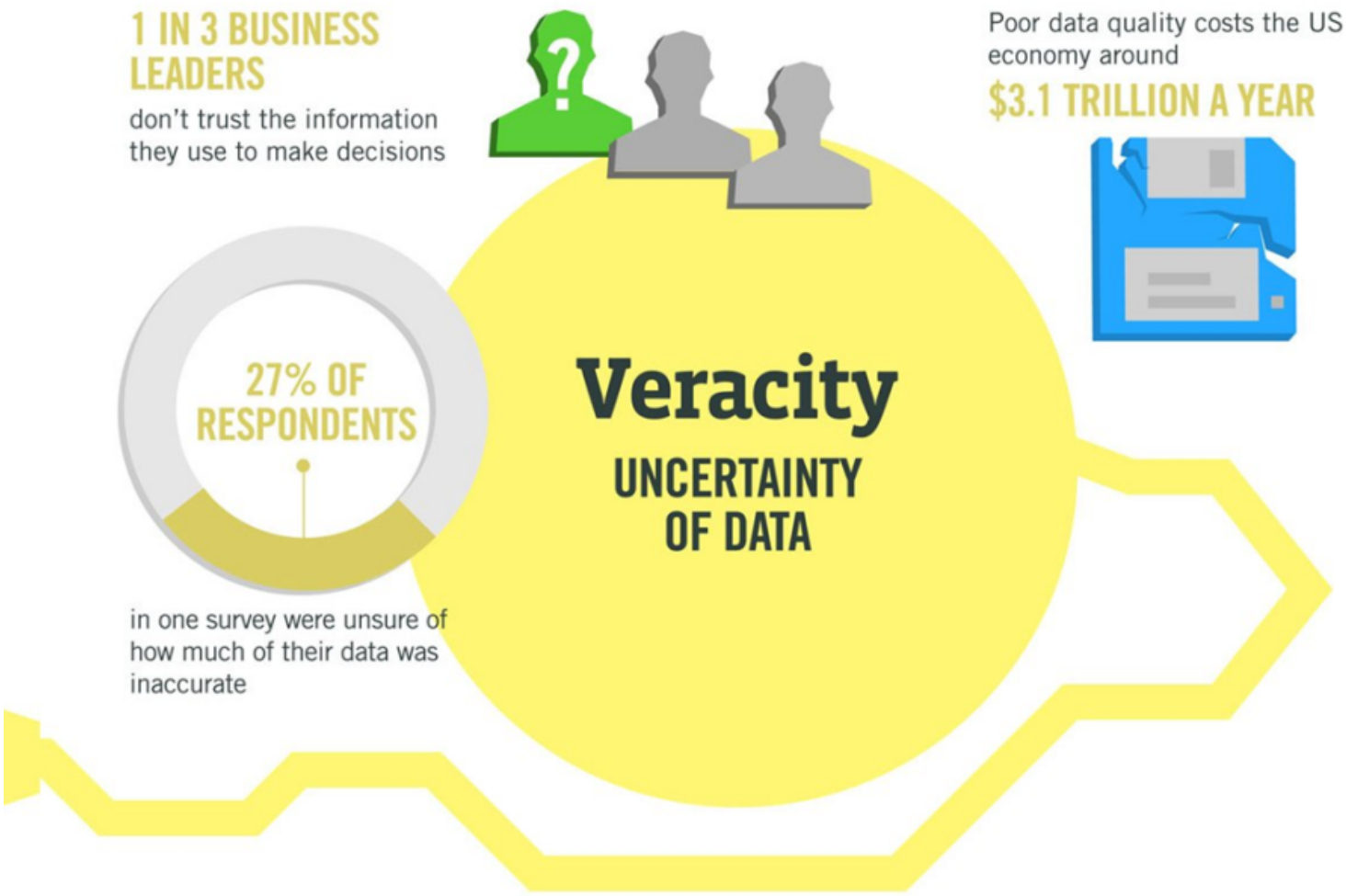
Thanks: http://www-01.ibm.com/software/data/bigdata/images/4-



Thanks: http://www-01.ibm.com/software/data/bigdata/images/4-Vs-of-big-data.jpg

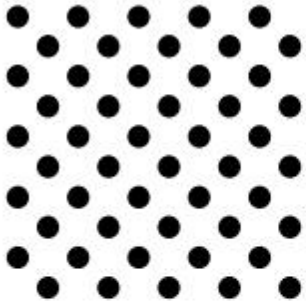
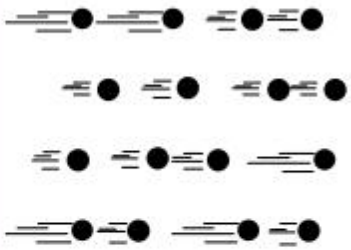
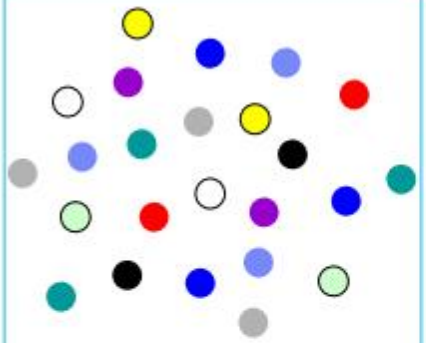
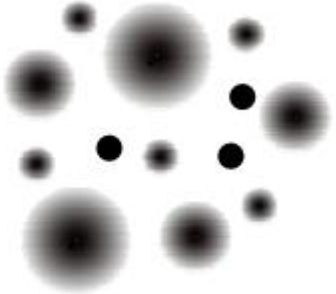


Thanks: http://www-01.ibm.com/software/data/bigdata/images/4-V



Thanks: http://www-01.ibm.com/software/data/bigdata/images/4-Vs-of-big-data.jpg

4Vs

Volume	Velocity	Variety	Veracity*
			
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**
Terabytes to exabytes of existing data to process	Streaming data, milliseconds to seconds to respond	Structured, unstructured, text, multimedia	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

The Model Has Changed…

Old Model of Generating/Consuming Data has Changed

Old Model:

Few privileged companies are generating and "owning" data,
all others are consuming data (in controlled packages)

The Model Has Changed…

- **New Model of Generating/Consuming Data has Changed**

Producers :

- **Everyone - Man, Woman and Child, and Devices**

Consumers:

- **Professionals**
- **Businesses**
- **Scientists**
- **And us**
- **Everyone wants a piece of this pie …**

What Sectors Can Benefit?

- **Businesses**
- **Transportation**
- **Science & Engineering**
- **Governments**
- **Energy**
- **Healthcare**
- **Education**
- **Entertainment**

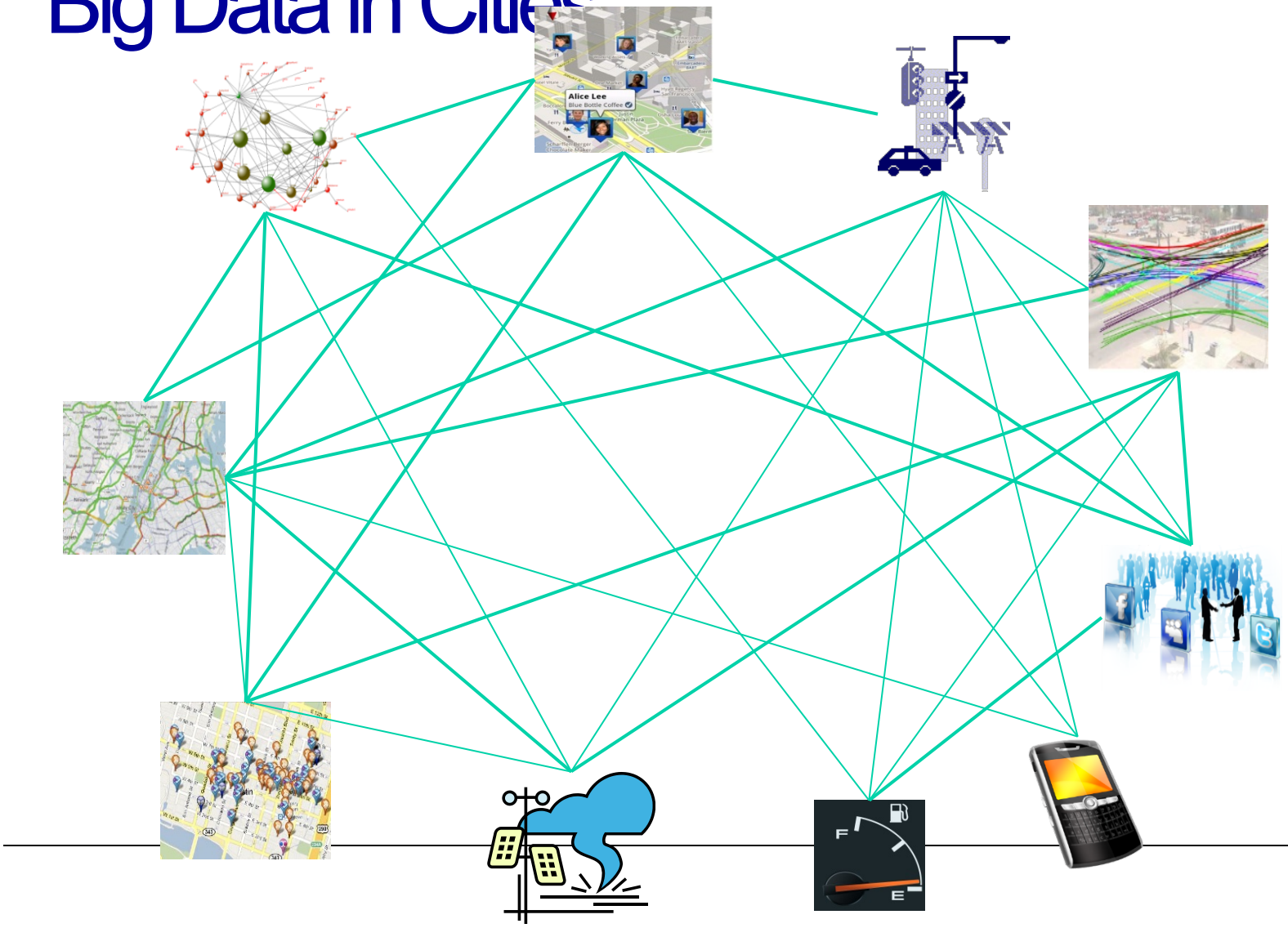
Utilize data to improve people's life quality

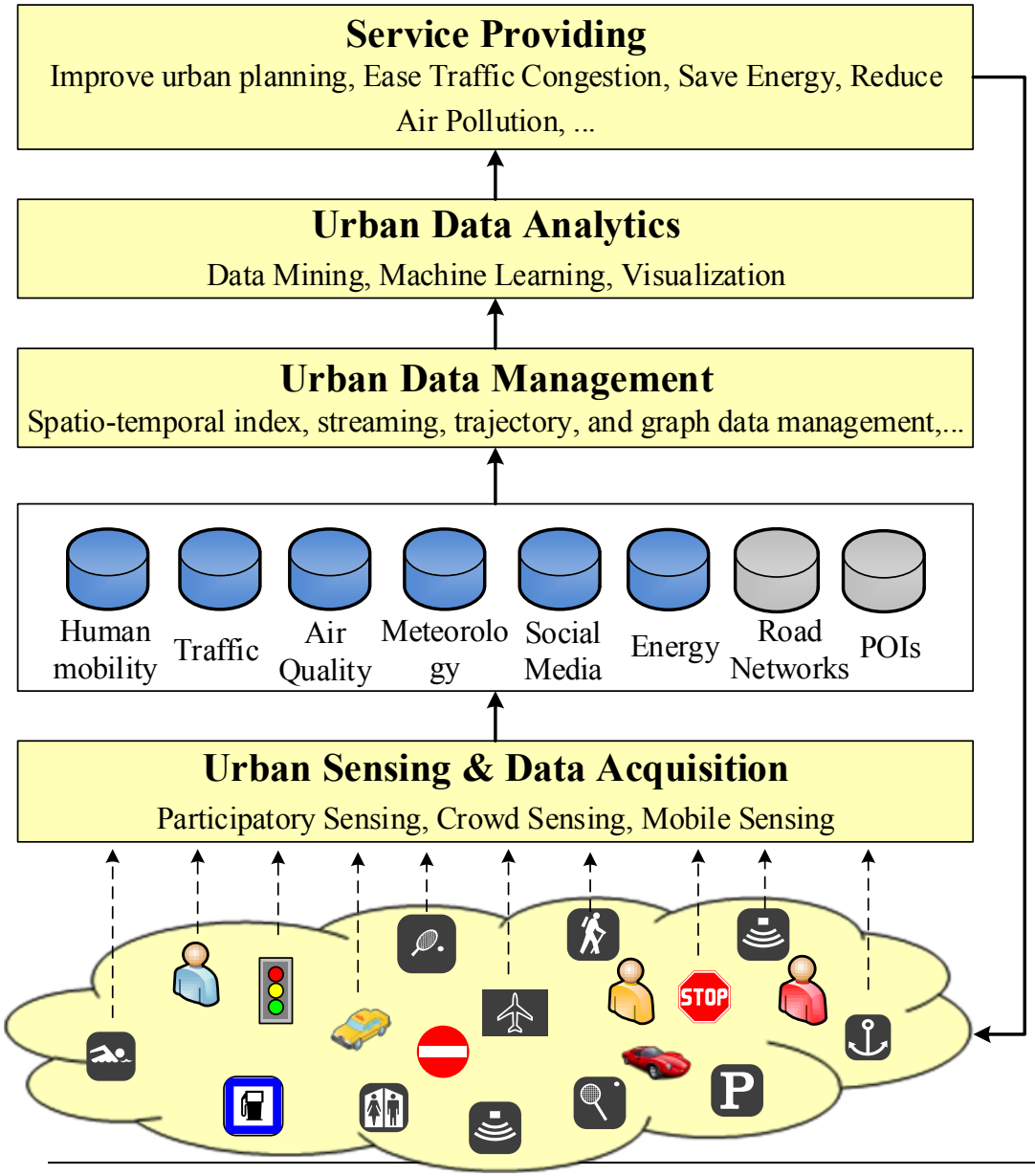
Done with the high level introduction

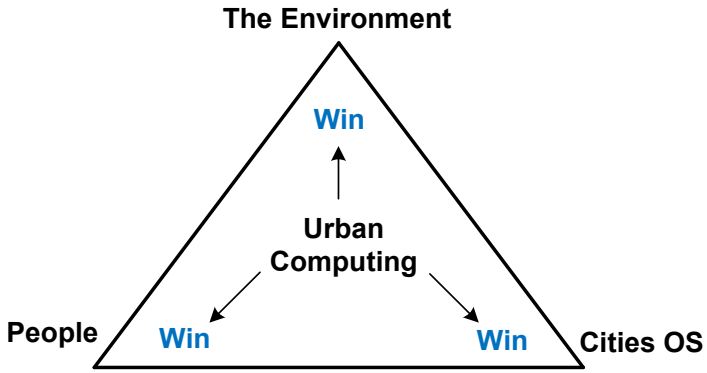
Begin with application stories

Big Challenges in Big Cities

Big Data in Cities

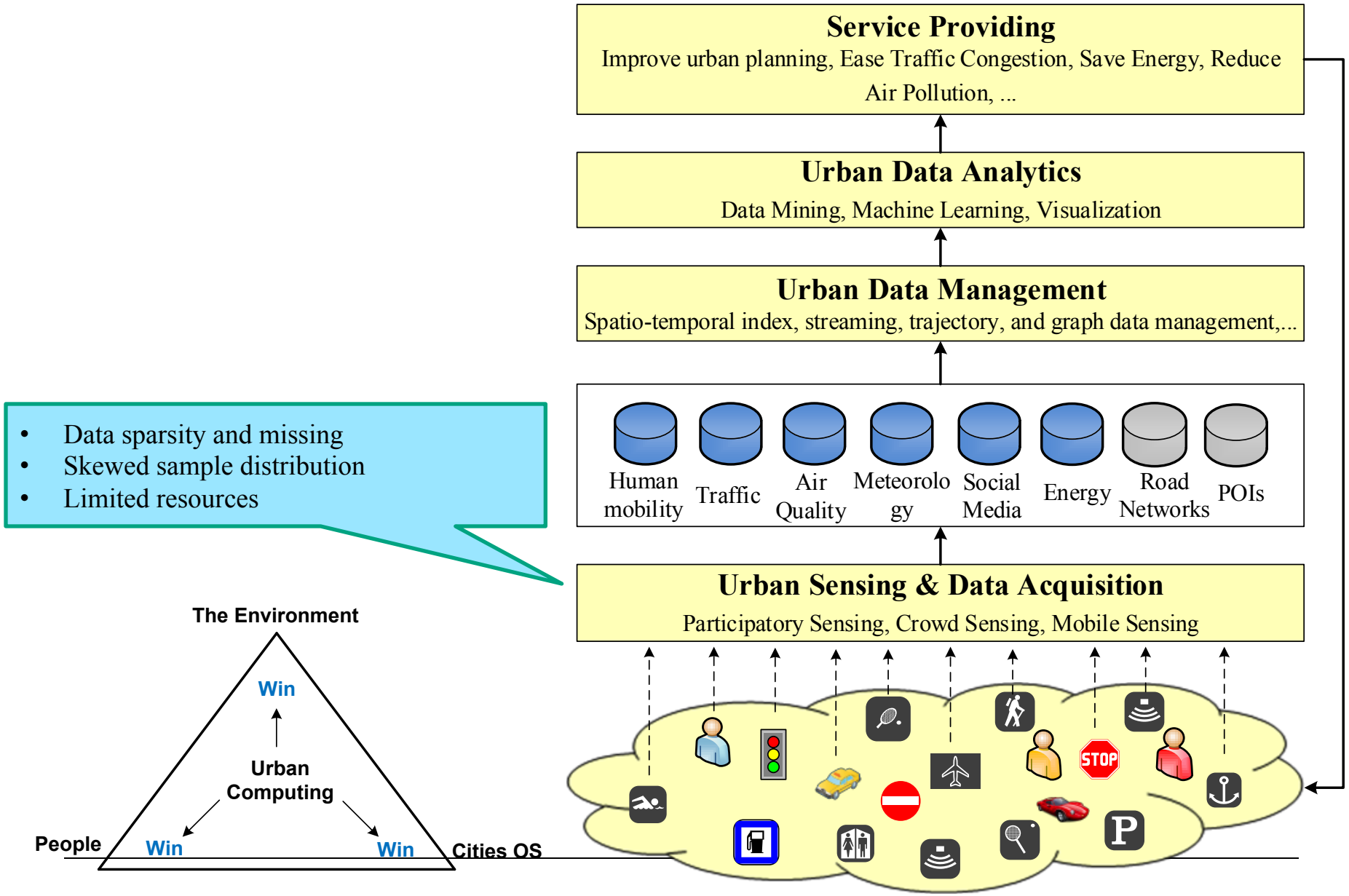






Tackle the Big challenges in Big cities using Big data!

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

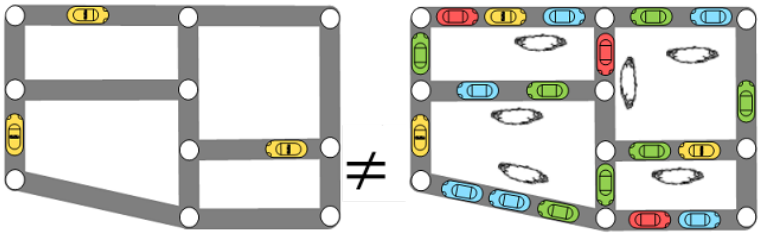


Zheng, Y., et al. Urban Computing: concepts, methodologies, and applications. *ACM transactions on Intelligent Systems and Technology*.

Urban Sensing

A sample of data → An entire dataset

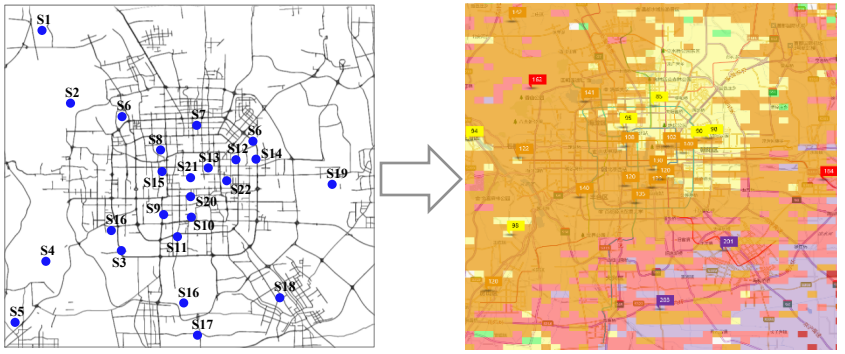
- Biased distribution



Taxi flow Entire traffic flow

Inferring Gas Consumption and Pollution Emission of Vehicles throughout a City. KDD 2014.

- Data sparsity and missing



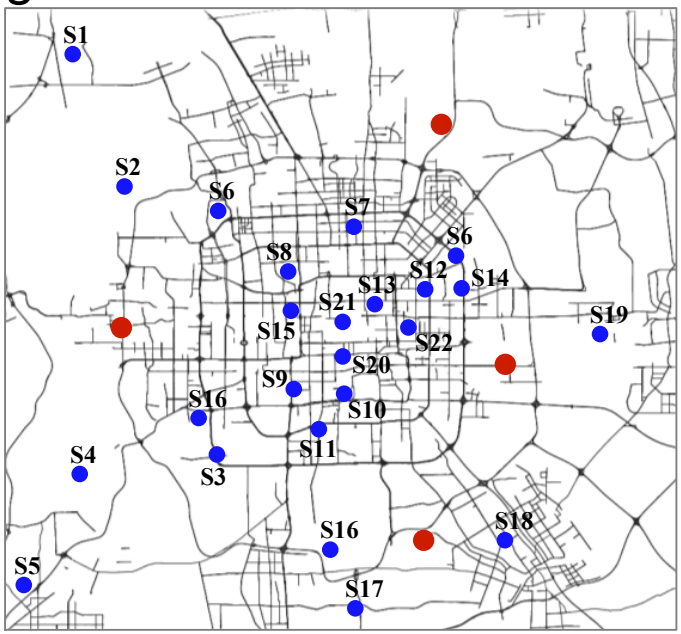
Air quality monitoring stations

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

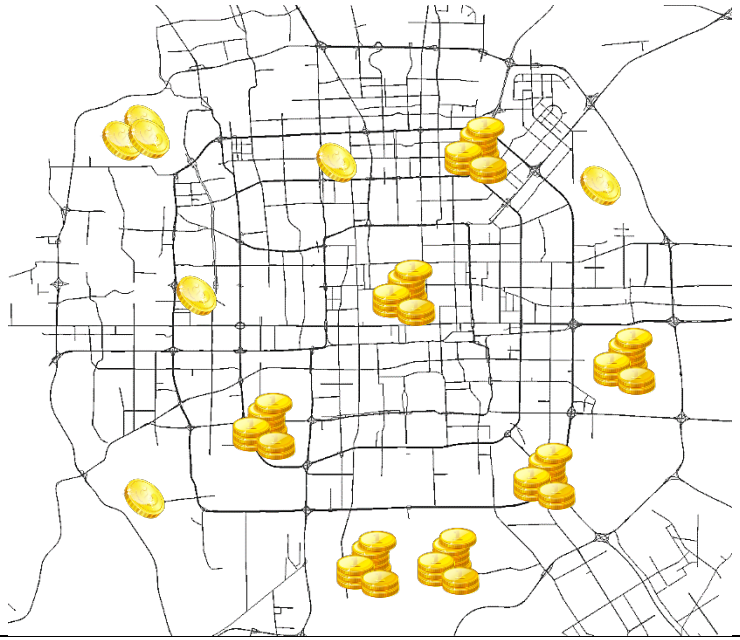
Urban Sensing

A limited resource (budget, labors, land…)

- Static sensing: Where to deploy sensor to maximize the gain?

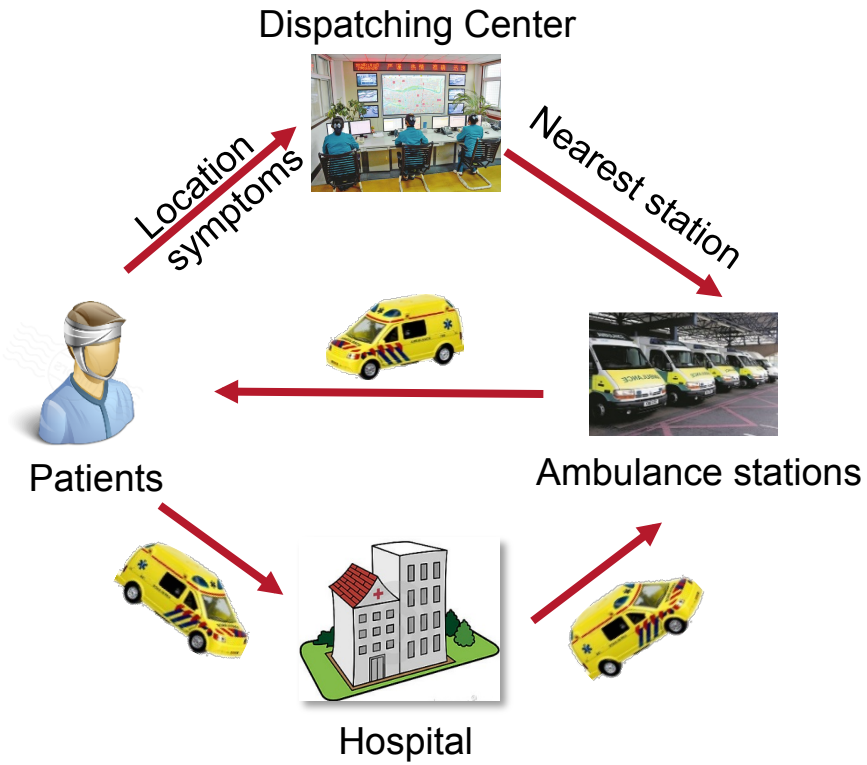


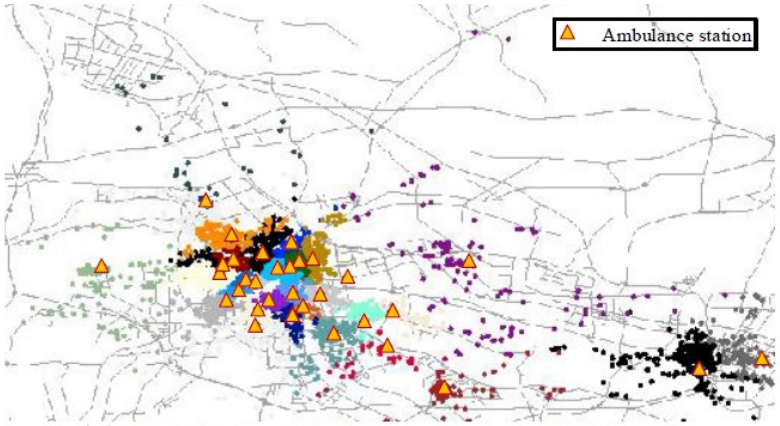
- Crowdsensing: How to arrange the incentives dynamically?



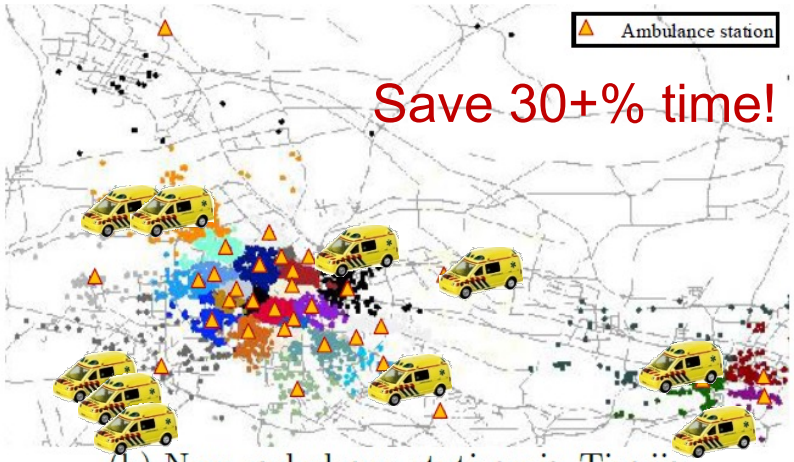
Suggesting locations for monitoring stations, KDD 2015

Improving Medical Emergency Services using Big Data





(a) Original ambulance stations in Tianjin

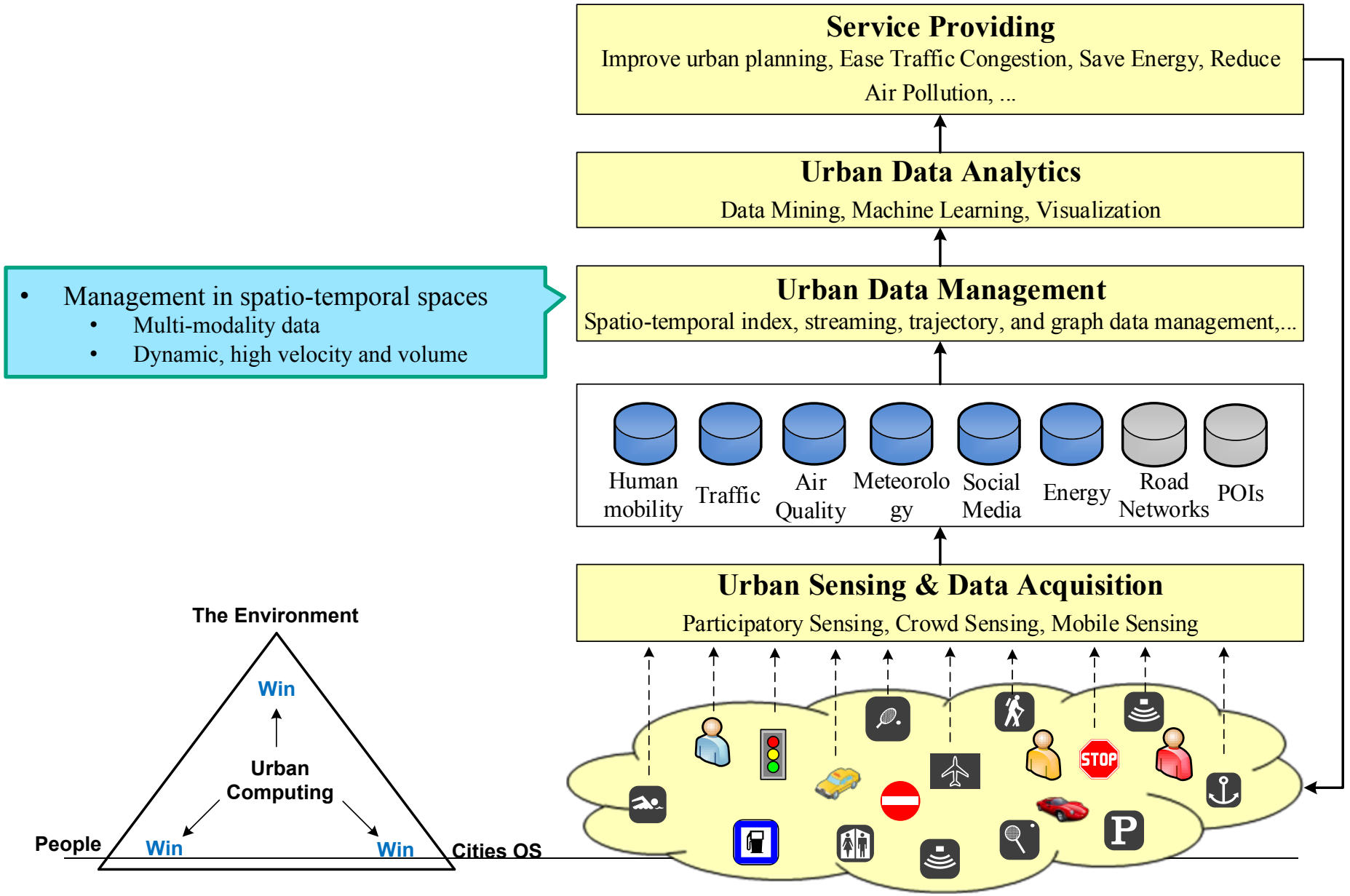


(b) New ambulance stations in Tianjin

- Select locations for Ambulance Stations
- Dynamic ambulance allocation

Yilun Wang, **Yu Zheng**, et al. Travel Time Estimation of a Path using Sparse Trajectories.. KDD 2014

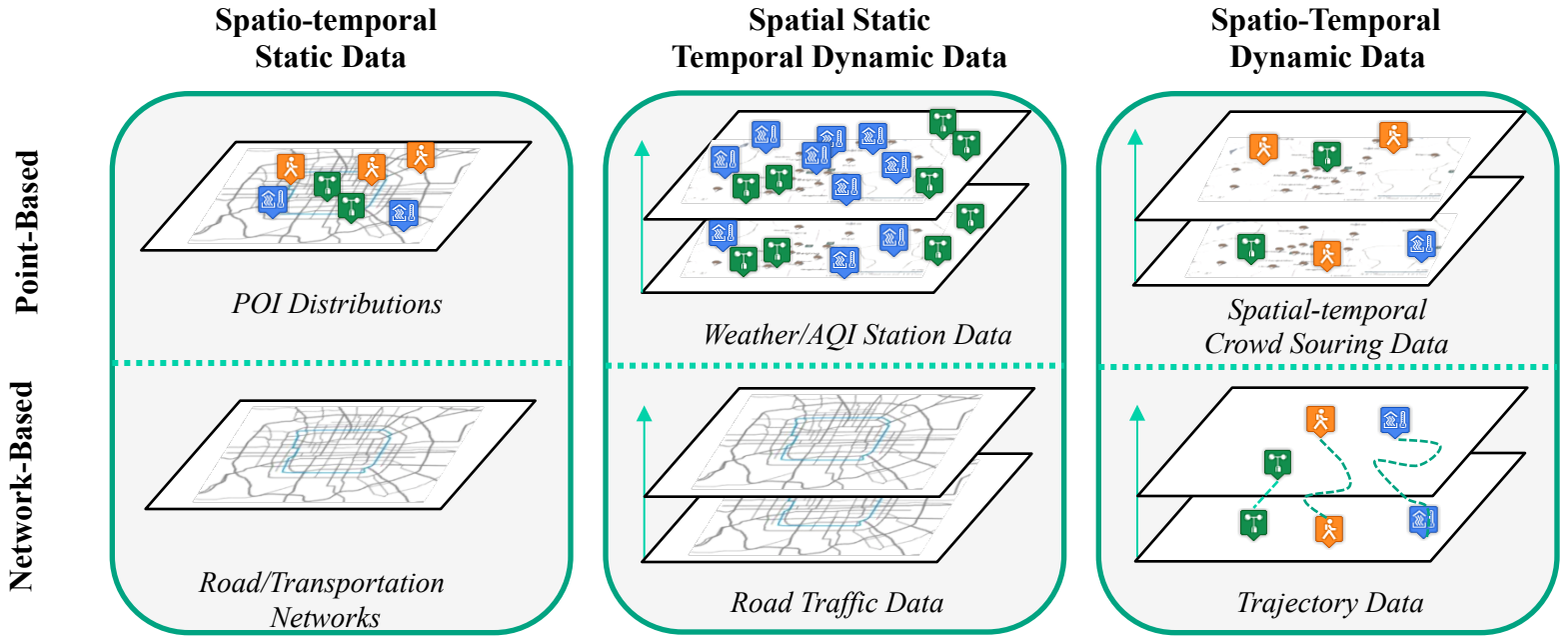
Location Selection for Ambulance Stations: A Data-Driven Approach, ACM SIGSPATIAL 2015



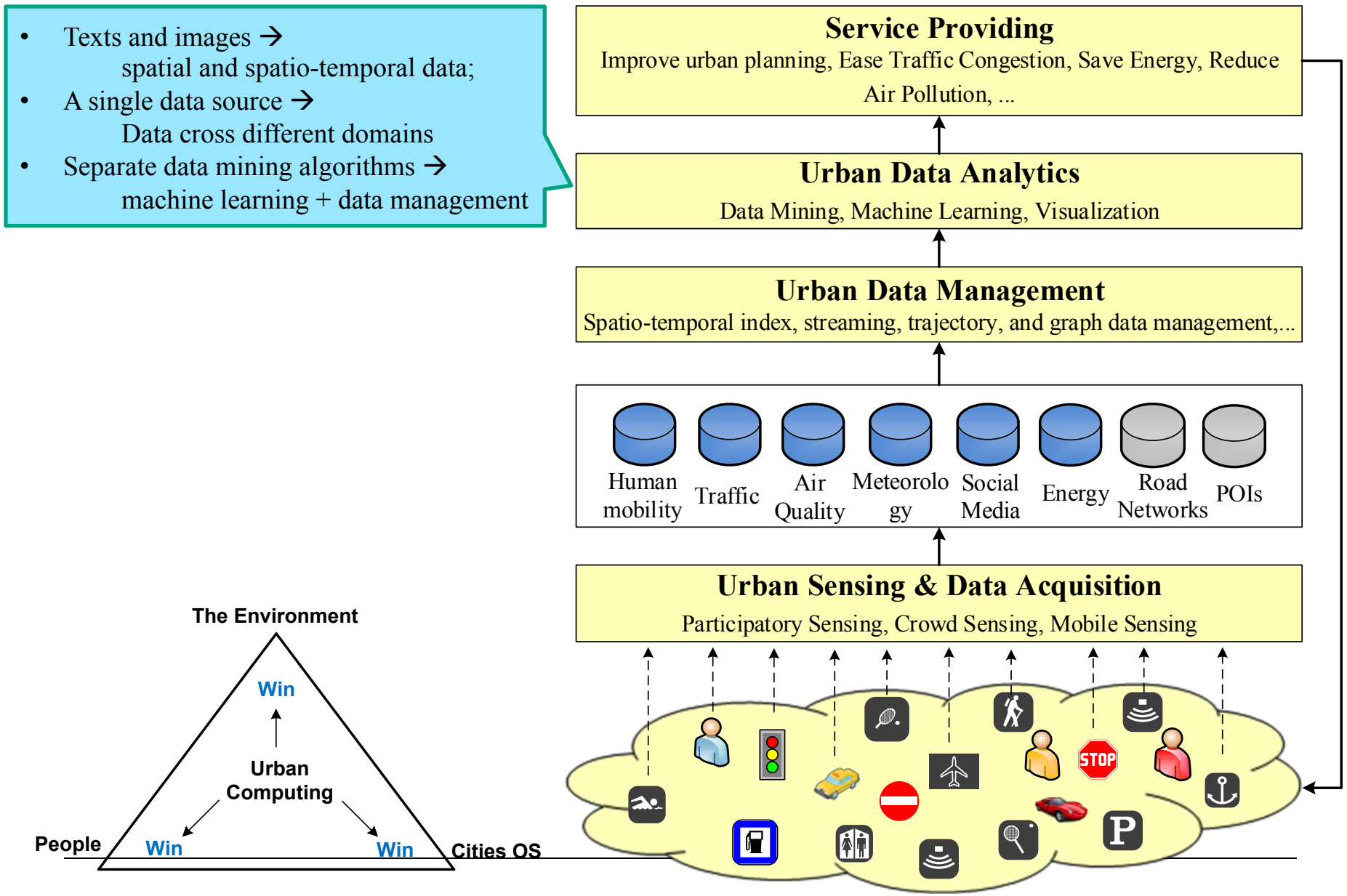
Zheng, Y., et al. Urban Computing: concepts, methodologies, and applications. *ACM transactions on Intelligent Systems and Technology*.

Urban Data Management

- Managing multi-modality data
  - Categorical and numeric data
  - Different scales, densities, updating frequency, and ST properties
- Dynamic and big volume
  - Group query strategy
  - Computing in parallel

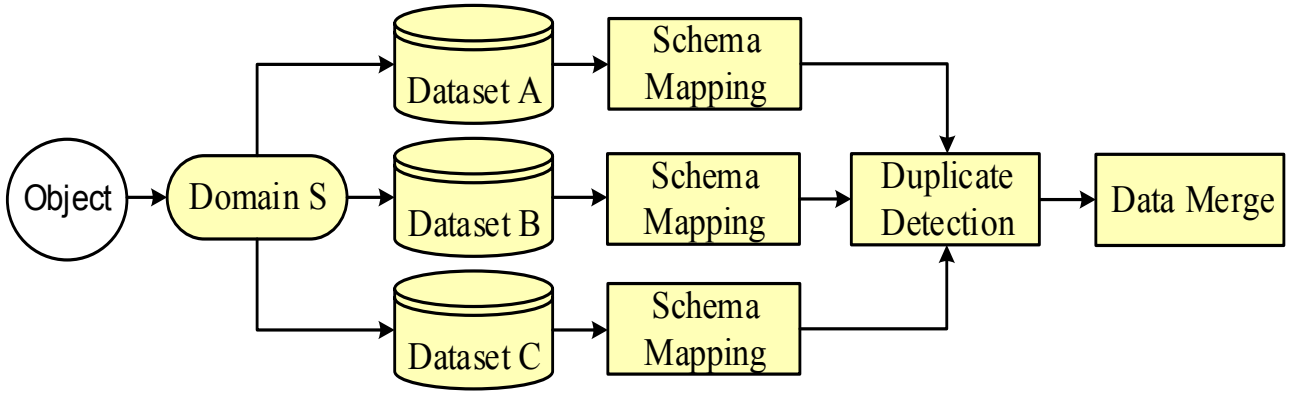


Yu Zheng. Trajectory Data Mining: An Overview. ACM Transactions on Intelligent Systems and Technology (**ACM TIST**). 2015

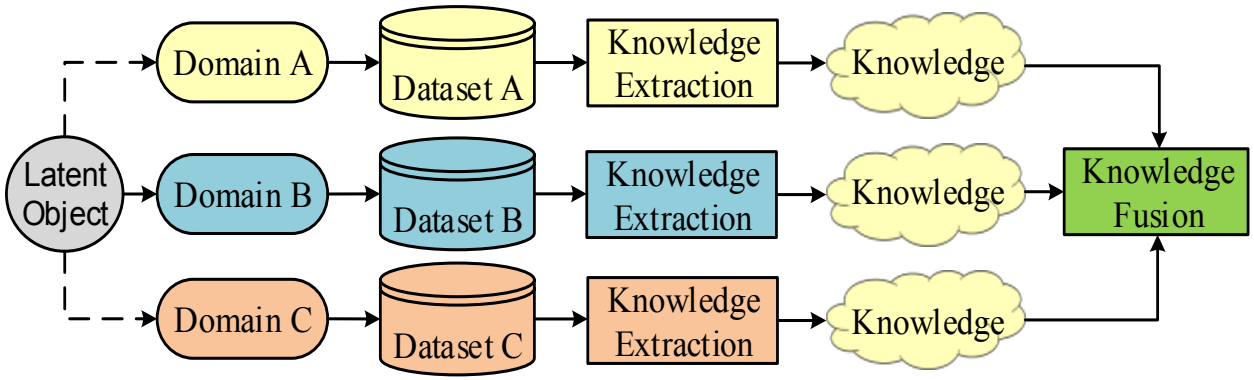


Zheng, Y., et al. Urban Computing: concepts, methodologies, and applications. *ACM transactions on Intelligent Systems and Technology*.

Data Integration vs Knowledge Fusion



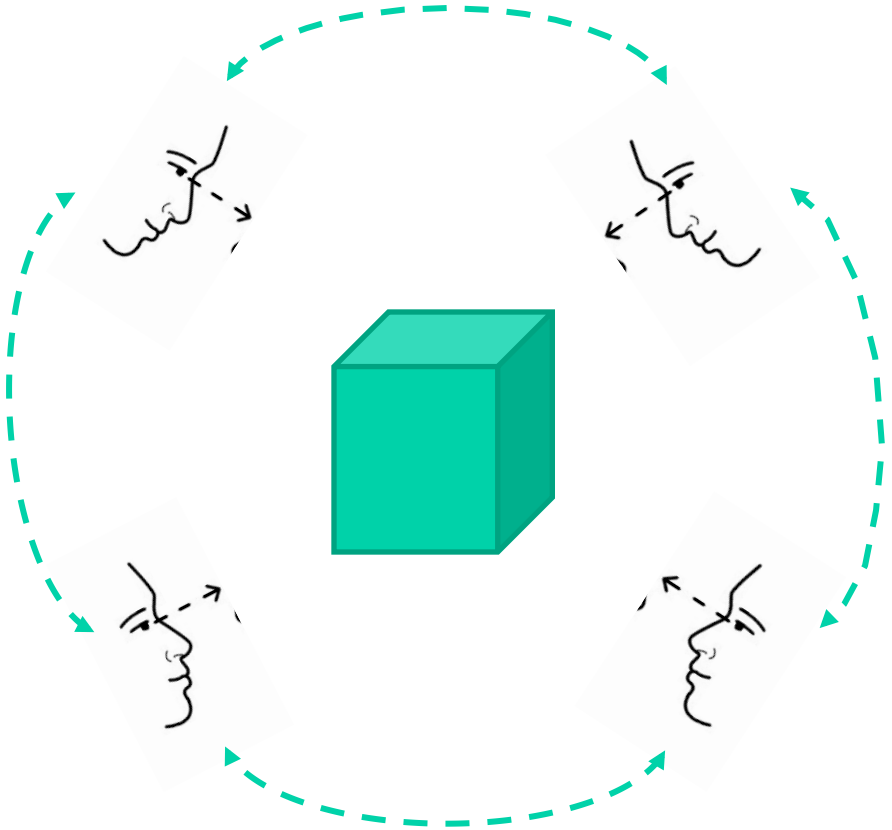
A) Paradigm of the conventional data fusion



Cross-Domain Data Fusion

Yu Zheng. Methodologies for Cross-Domain Data Fusion: An Overview. IEEE Transactions on Big Data, 1, 1, 2015.

Multi-View-Based Learning



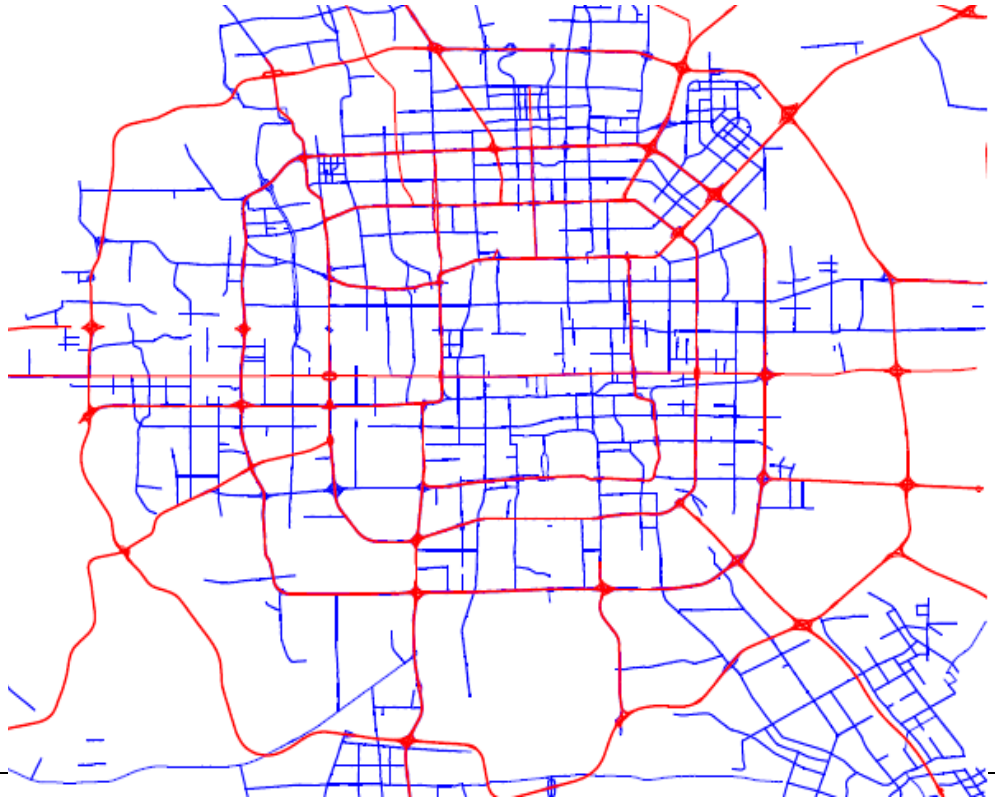
Urban Computing for Urban Planning

Best Paper Nominee Award at UbiComp 2011

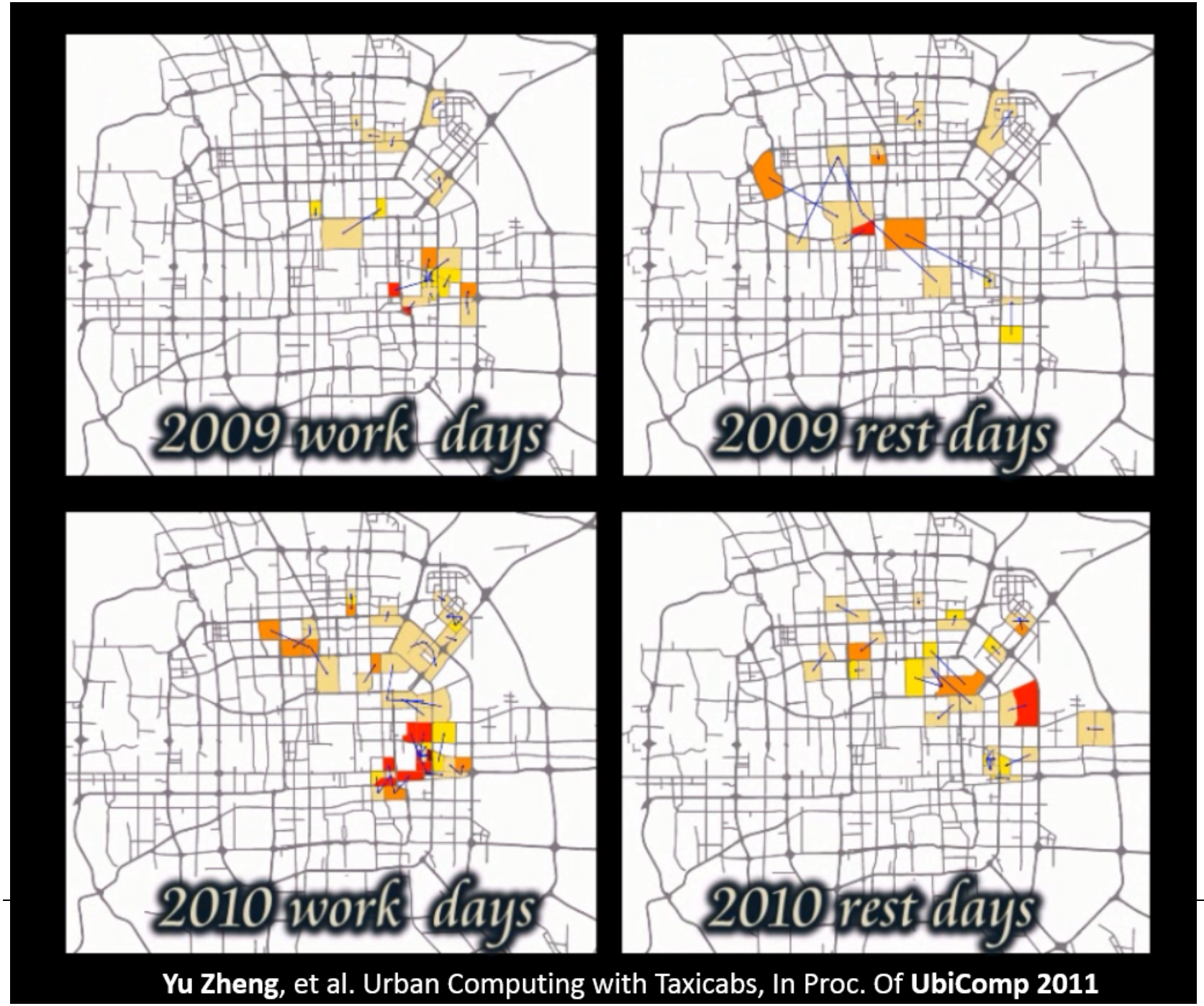
The Most Cited Paper

City-Wide Traffic Modeling

- Partition a city into regions with major roads
- Regions are root causes of the problem



Yu Zheng, et al. Urban Computing with Taxicabs, In Proc. Of **UbiComp 2011**



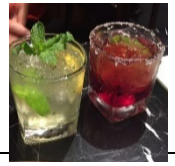
Yu Zheng, et al. Urban Computing with Taxicabs, In Proc. Of **UbiComp 2011**

Shanghai Big Data Hotpot Restaurant











When Urban Air Meets Big Data

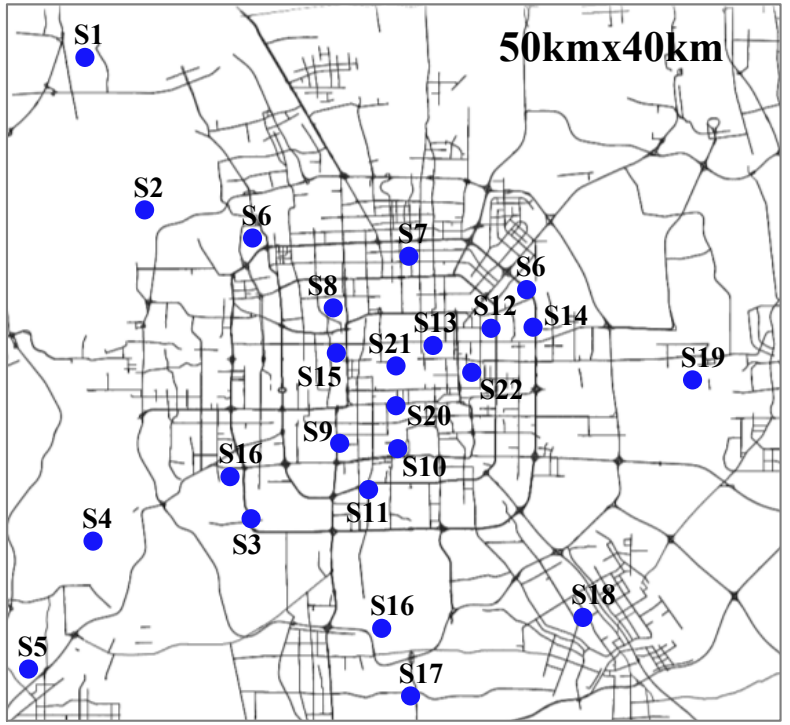
KDD 2013

http://urbanair.msra.cn/

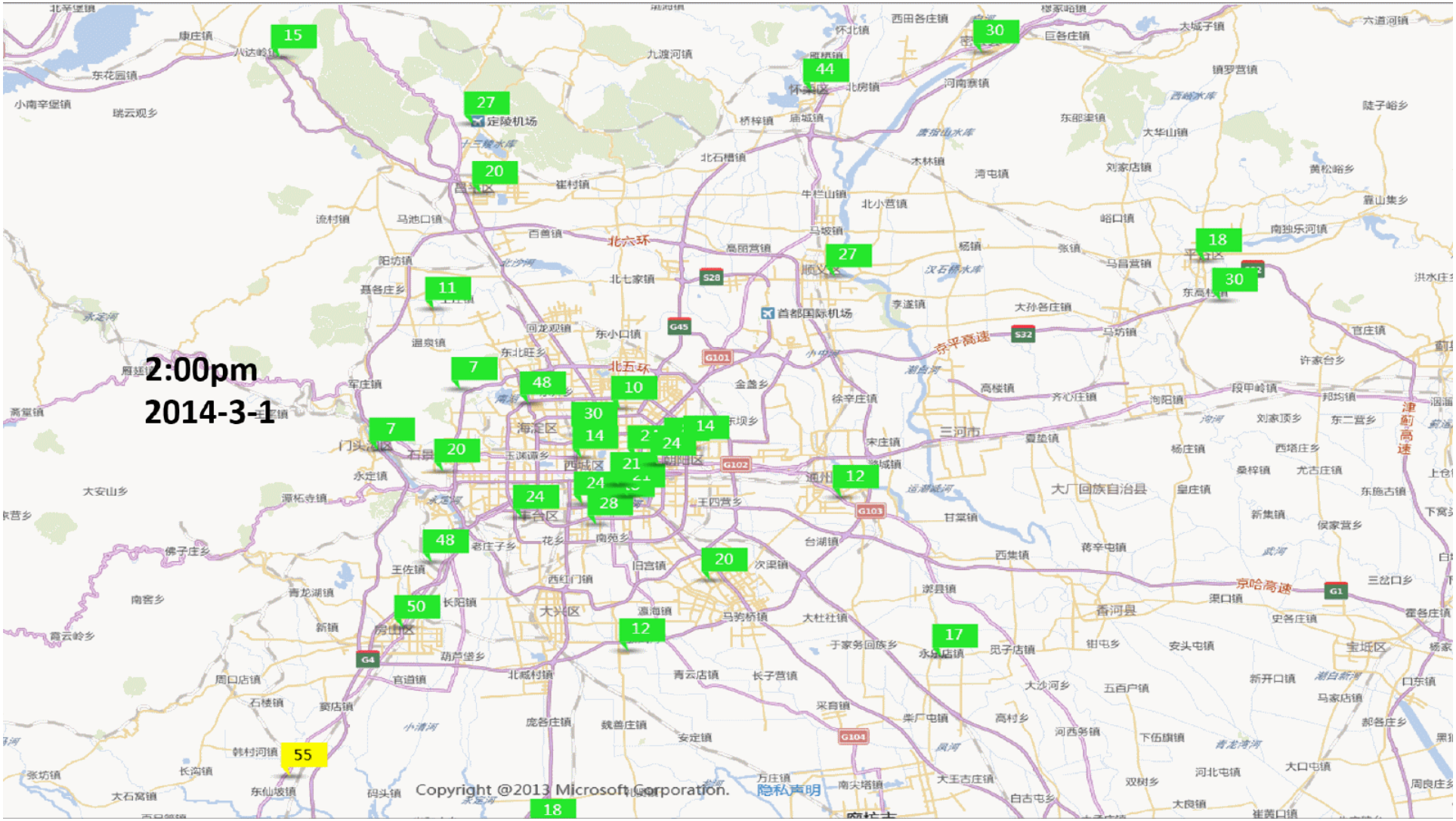
Air Pollution: A Global Concern !

PM2.5, PM10, NO_2, SO_2, CO, O_3

● Air quality monitor station



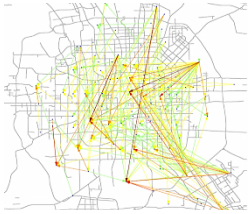
We do not really know the air quality of a location without a monitoring station!



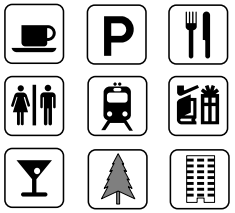
Inferring **Real-Time** and **Fine-Grained** air quality throughout a city using **Big Data**

Meteorology

Traffic

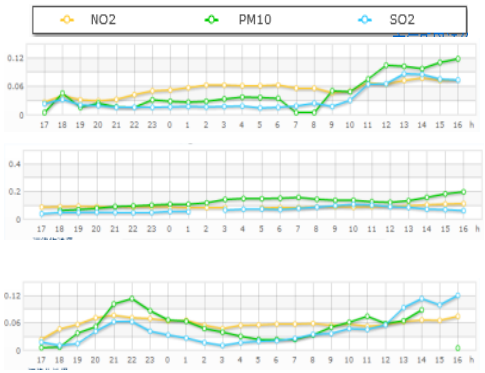


Human Mobility



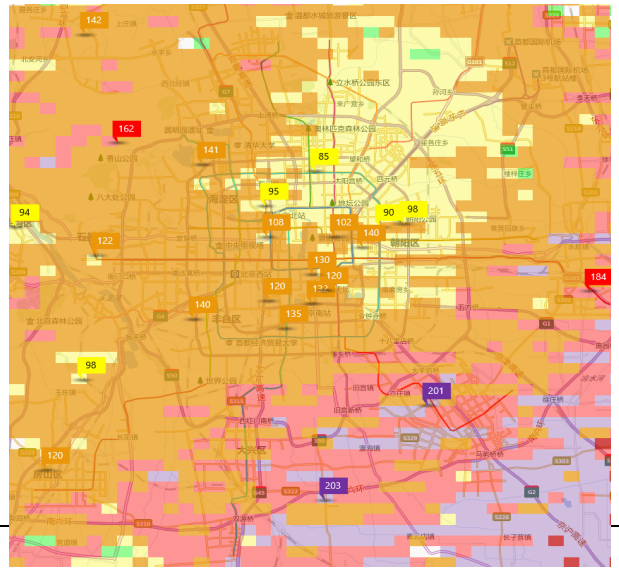
POIs

Road networks



Historical air quality data

Real-time air quality reports



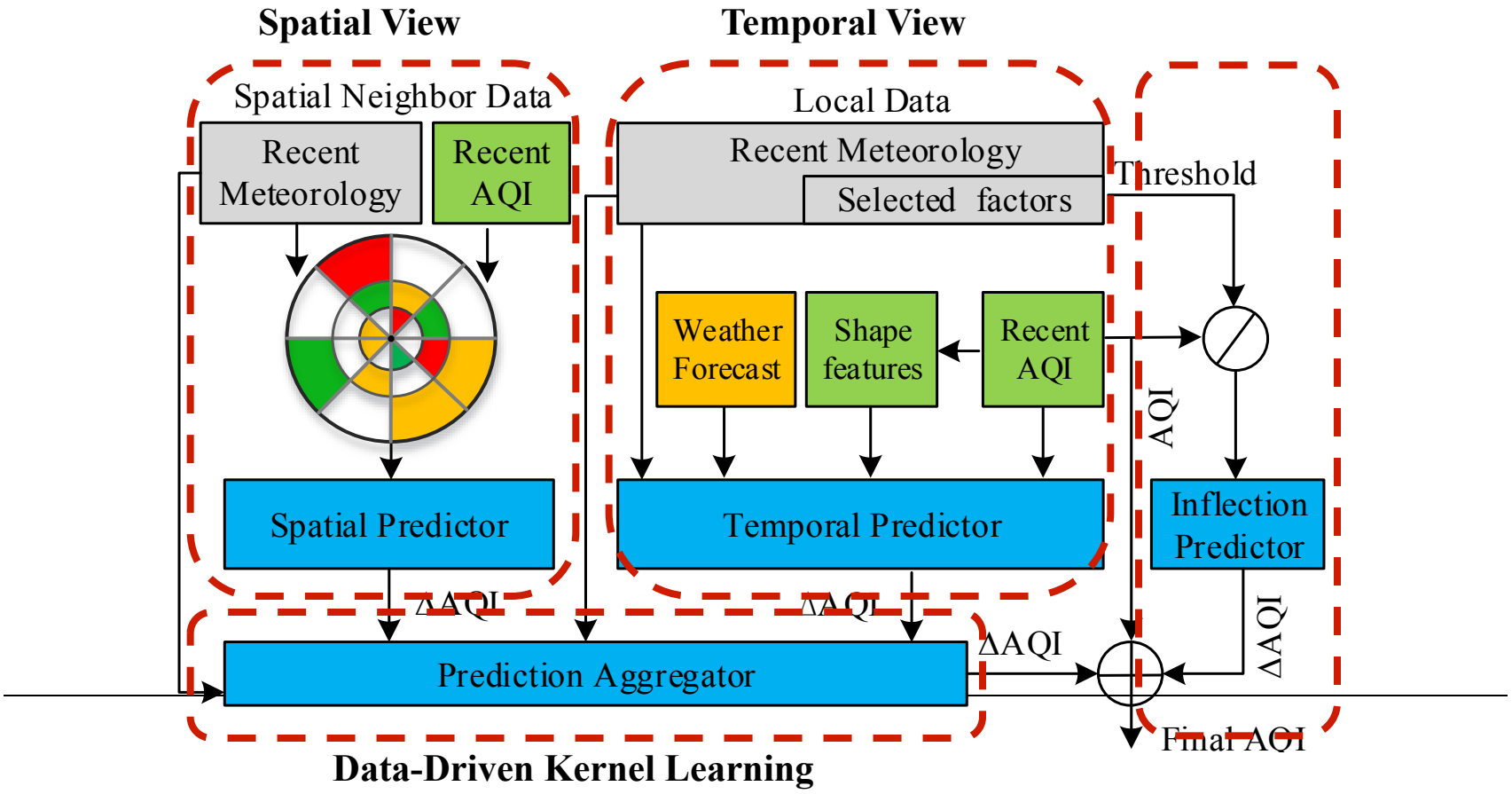
Zheng, Y., et al. U-Air: when urban air quality inference meets big data. KDD 2013

Urban Air System

Zheng, Y., et al.
U-Air: When Urban Air Quality Inference Meets Big Data. KDD

Multi-View Learning Framework

- Features: Non-overlapping features providing different views
- Models: Model extrapolation and trend regression respectively
- Training: Combination of small models vs. a big model



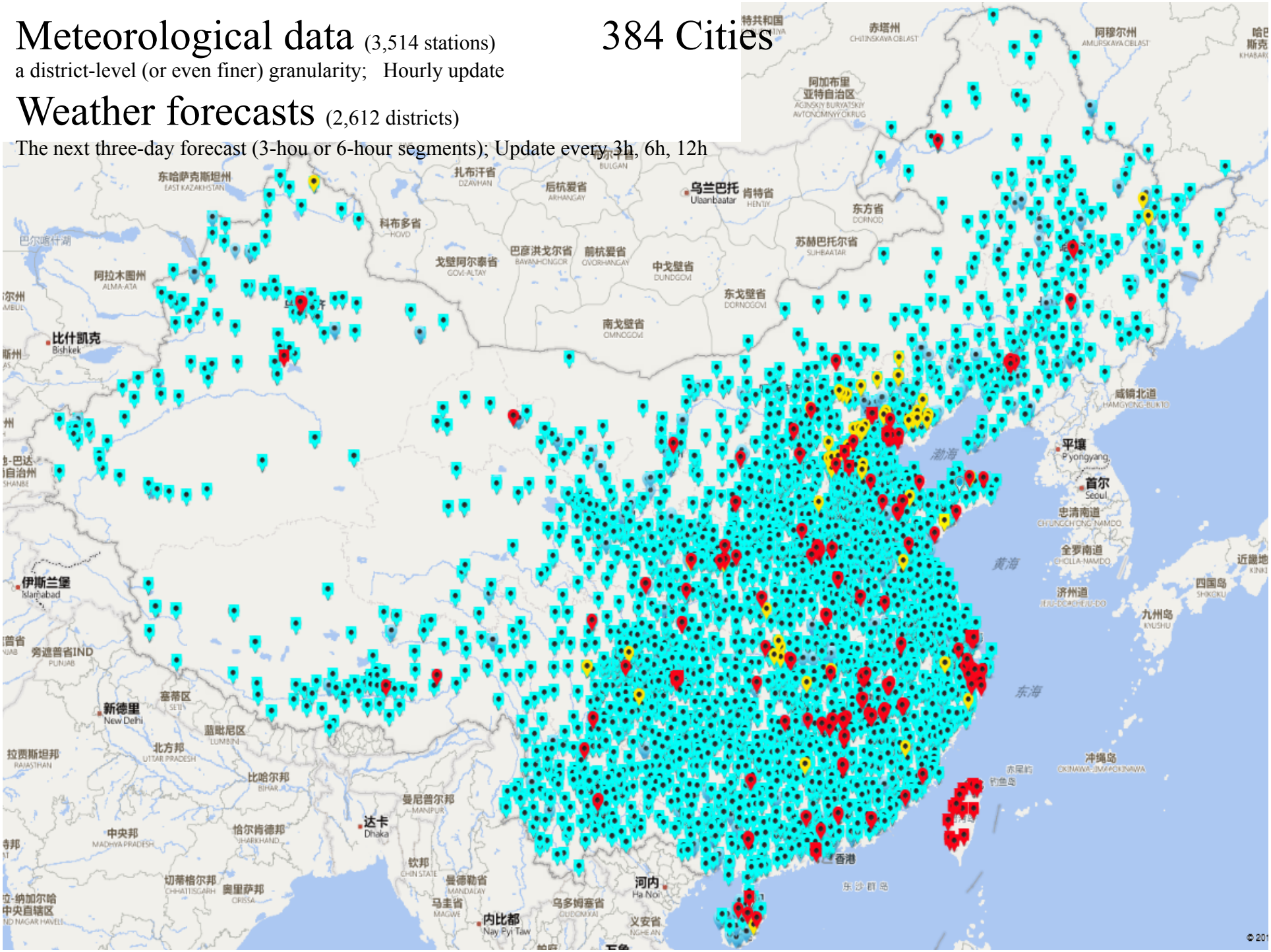
Meteorological data (3,514 stations)

a district-level (or even finer) granularity; Hourly update

Weather forecasts (2,612 districts)

The next three-day forecast (3-hou or 6-hour segments); Update every 3h, 6h, 12h

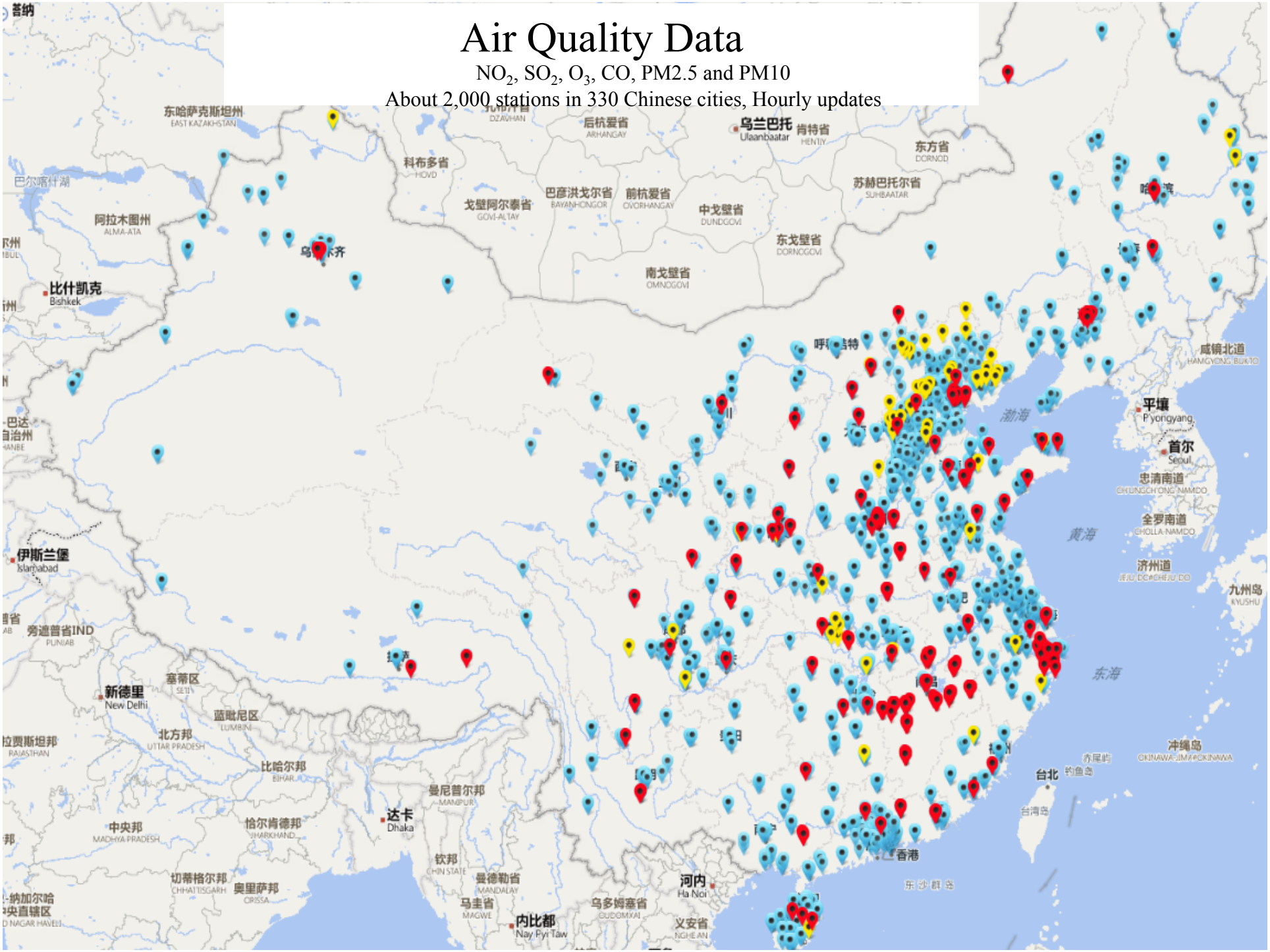
384 Cities



Air Quality Data

NO_2, SO_2, O_3, CO, PM2.5 and PM10

About 2,000 stations in 330 Chinese cities, Hourly updates



Revisit Big Data

- **NOT** a single data source which is very big
- **NOT** mean full data
- **NOT** mean very dense data
- May need **less** domain knowledge
- Tools are ready

- Data across different domains
- Sample of (label) data
- Data sparsity always exists
- More understanding of data itself and data science
- Many unsolved problems

Big Data ≠ Mining Single Dataset ≠ Simple Statistics

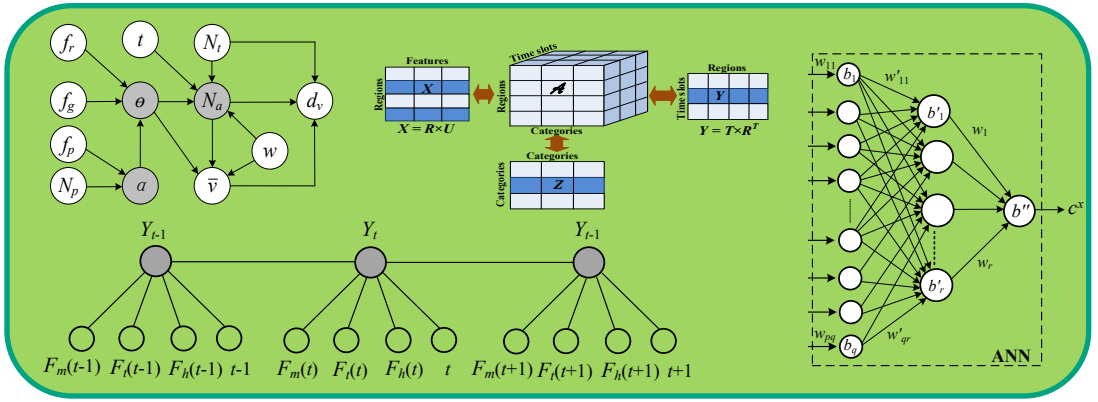
Big Data ≠ Machine learning ≠ Deep Learning

Big Data ≠ Cloud Computing ≠ Hadoop

Big Data needs comprehensive capabilities to deliver end-to-end services!

Data Scientist

Models and Algorithms



Problems

Data

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

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

Yu Zheng. Trajectory Data Mining: An Overview. *ACM Transactions on Intelligent Systems and Technology*. 2015

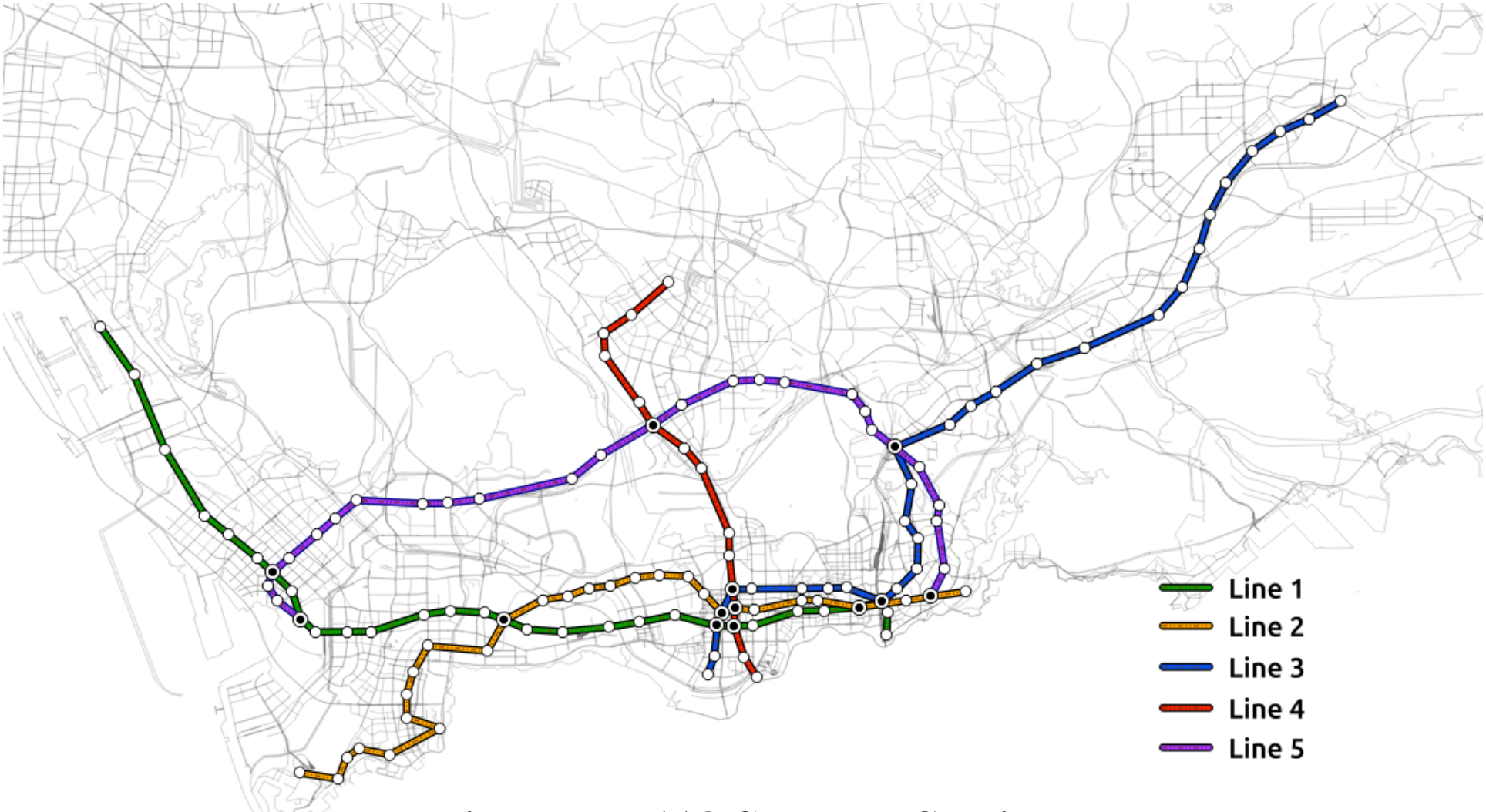
~~**Yu Zheng**. Methodologies for Cross-Domain Data Fusion: An Overview. *IEEE Transactions on Big Data*, 1, 1,~~ 2015.

What data available in our course?

Road Map in Shenzhen

20,656 Road Segments

Subway Lines

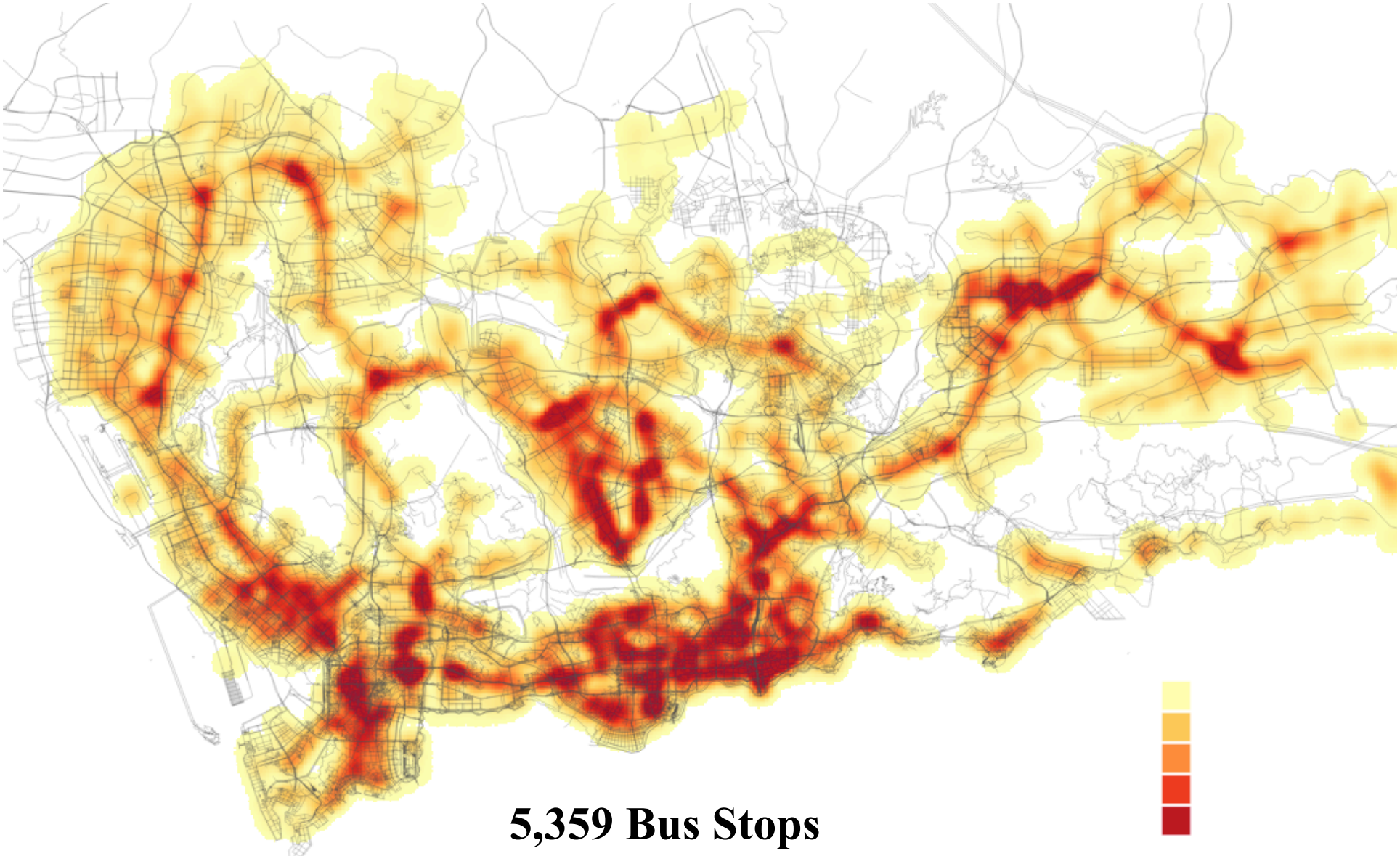


5 subway lines and 118 Subway Stations

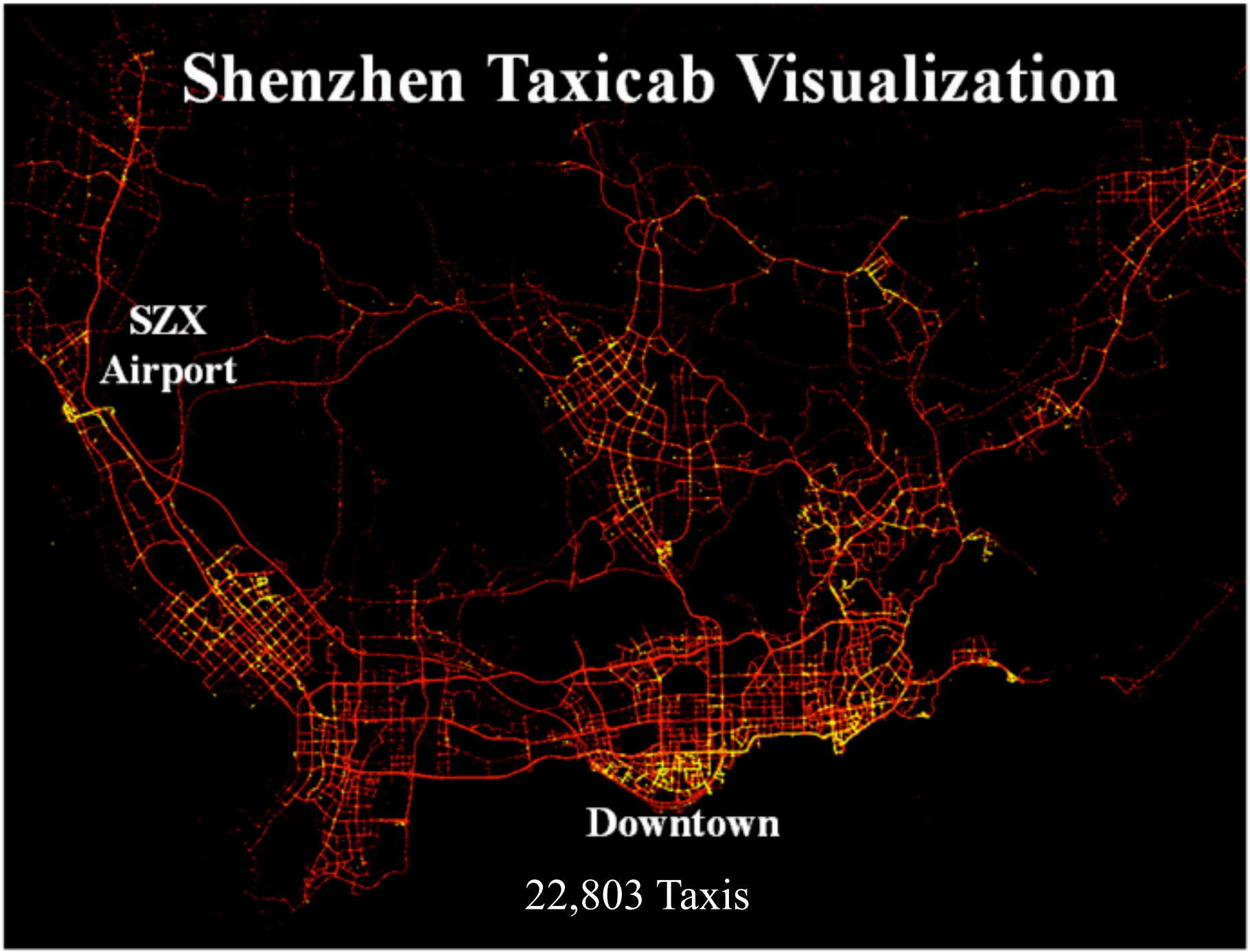
Bus Routes

8,875 Buses serve 814 Bus Routes

Bus Stop Distribution



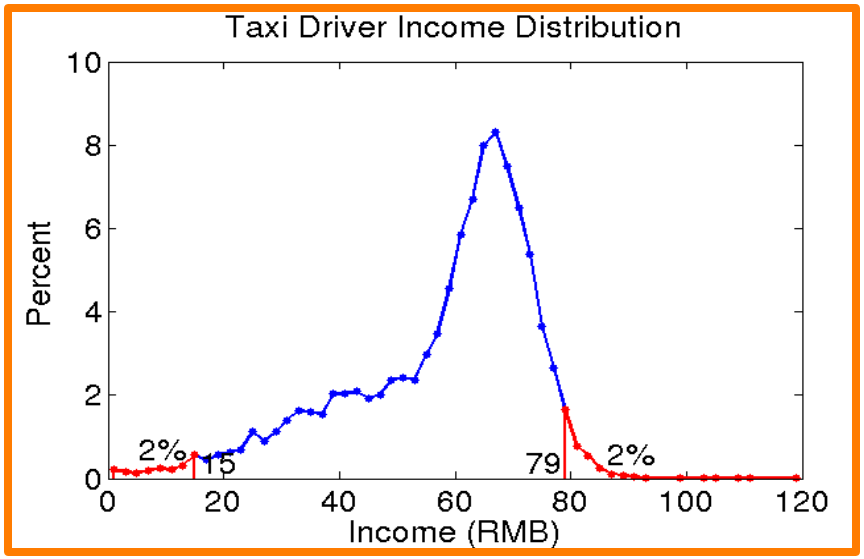
5,359 Bus Stops


Shenzhen Taxicab Visualization
SZX
Airport
Downtown
22,803 Taxis

Transportation Billing Data

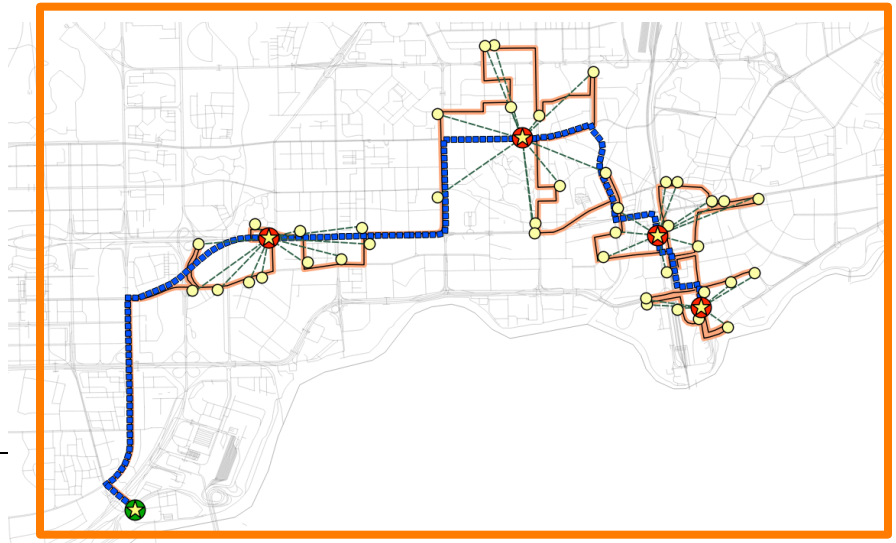
Urban Issues

Regional Weather-Traffic Sensitivity



Smart Taxi

Traffic Estimation & Prediction

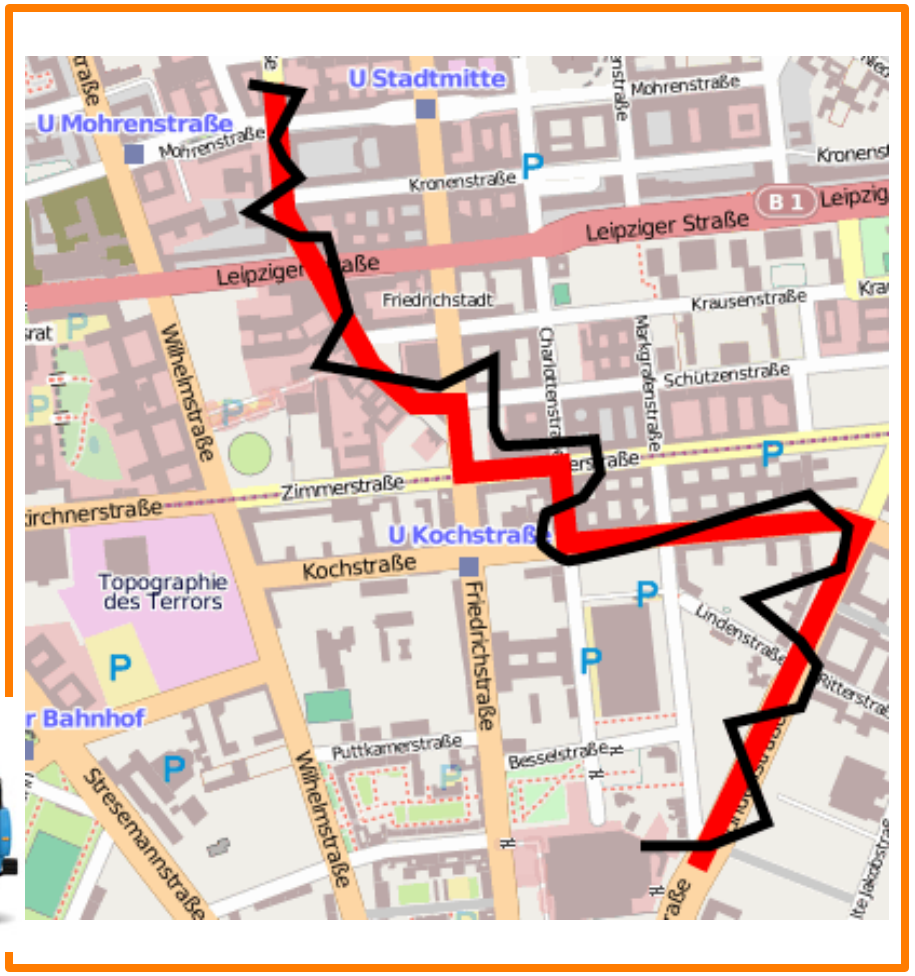


Smart shuttle service

Urban Issues (cont.)

Population Census

Logistic Planning



Low Sample Rate
Map Matching

Questions?