Welcome to

DS3010: Computational Data Intelligence --Introduction & Logistics

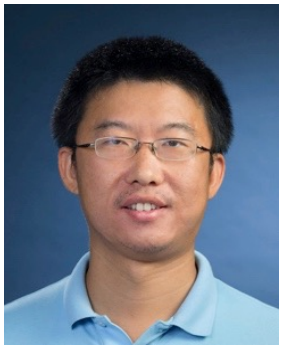
Prof. Yanhua Li

Time: 11:00pm – 12:50pm Mon & Thur

Location: Higgins Labs 114

Spring 2022

Who am I?



Yanhua Li, Pronounced as "yan-**wa** l**ee**"
Associate Professor
Computer Science & Data Science

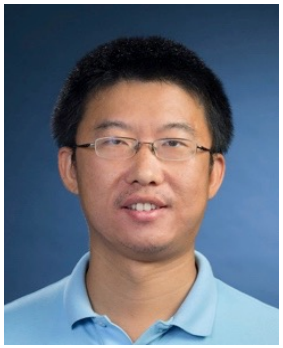
Remote Office hours: 10am-11am Friday Zoom (link in Canvas)

Other times are available by appointment.

Best ways to contact me:

- WPI email: yli15@wpi.edu

Who am I?



Yanhua Li, PhD
Associate Professor
Computer Science & Data Science

PhD, Computer Science, U of Minnesota, 2013
PhD, Electrical Engineering, BUPT, 2009

Research Interests: Big data analytics, Smart Cities, Measurement, Spatio-temporal Data Mining

Industrial Experience: Bell-Labs, Microsoft Research.

TA Hang Li

Office hour location:

Monday: 2-3pm

(Zoom at https://wpi.zoom.us/j/6644998206)

Friday: 1-2p

(In-person at Unity Hall (UH) 341),

WPI email: hyin@wpi.edu

Welcome!

- Basic course information
  - Course number- DS3010
  - Course name- Data Science III: Computational Data Intelligence
  - When/where - Mondays and Thursdays from 11am-12:50pm, HL114

Objectives for today

- Discuss the course mechanics (syllabus, grading, etc.)
- Start to get inspired Data Science!

High level course goals and learning objectives

This course builds on DS2010 and focuses on computational aspects of Data Science, covering a selection of key challenges in, and methodologies for, working with big data.

Topics to be covered include:

- **Data Collection**
- **The "Data Pipeline"**
- **Data Acquisition, and Cleaning**
- **Data Management**
- **Data Mining (e.g., mining graphs), and Machine Learning**
- **Deep Learning**

- **Applications (recommender system)**

High level course goals and learning objectives

We will also cover professional skills, such as

- **communication,**
- **presentation, and**
- **Application project.**

Students will acquire a working knowledge of data science through hands-on projects in a variety of Data Science domains.

“Recommended background” for the course

Recommended background:

.Data science basics equivalent to DS 1010

.Data analysis principles and modeling equivalent to DS 2010

.Knowledge of basic statistics equivalent to (MA2611 and MA2612),

.The ability to program equivalent to (CS 1004 or CS 1101 or CS 1102) and (CS 2102, CS2103 or CS 2119)

.An understanding of databases equivalent to (CS3431 or MIS3720) are assumed.

.Programming skill (in Python) are recommended.

Course Topics

- Large scale data sampling and estimation,
- Data Cleaning,
- Data management,
- Graph Data Mining,
- Data clustering,
- Machine Learning and Deep Learning
- Applications (Recommender system)

Just to get an idea of where we are...

- How many of you know Python?

Just to get an idea of where we are...

- How many of you know Python?
- How many of you have taken a database course?

Just to get an idea of where we are...

- How many of you know Python?
- How many of you have taken a database course?
- How many of you know what supervised/unsupervised machine learning is?

Just to get an idea of where we are…

- How many of you know Python?
- How many of you have taken a database course?
- How many of you know what supervised/unsupervised machine learning is?
- How many of you have done a project before that involves analyzing data?

Computing background for the course

- You will need to be able get your hands dirty playing with, processing, and plotting data using the ***Python*** computer language!
- That will be the officially supported language for the course and all lecture examples will be in Python.
- Now, with that being said, this is not intended to be a programming course (i.e., your code will not be graded), but actually working with data will be extremely important (i.e., the results of the code will be graded)!

Python

- Python itself can be found at:
  - http://www.python.org
- We will also be making use of iPython/Jupyter notebooks. They makes developing Python code much easier. They can be found at:
  - http://jupyter.org
- Good place to start:
  - Learning Python By Mark Lutz O'Reilly Media, September 2013.
    - http://shop.oreilly.com/product/0636920028154.do
    - Available for free from the library.
  - Python for Data Analysis by Wes McKinney, October 2012.
    - http://shop.oreilly.com/product/0636920023784.do

    - Available for free from the library.

Math background for the course

You will need to do some math for this course, especially using ideas in *supervised / unsupervised machine learning*.

In particular, you will need to be familiar with:

- **Basic statistics and probability – mean and variance, normal distributions, etc.**
- **Linear algebra – Matrices, vectors, eigenvalues, and decompositions such as the Singular Value Decomposition.**
- **Supervised machine learning – including the Bias-Variance trade-off, regression and classification techniques, error estimation**

Textbook

- **There is no single textbook for the course.**
- **However, there are four books are good references for this course:**
  - **Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Instagram, Github, and more,** Matthew Russell
    - https://www.amazon.com/Mining-Social-Web-Facebook-Instagram-dp-1491985046/dp/1491985046/ref=mt_paperback?_encoding=UTF8&me=&qid=1578332195
  - **Python for Data Analysis**, Wes McKinney
    - https://www.amazon.com/Python-Data-Analysis-Wrangling-IPython-ebook/dp/B075X4LT6K
  - **Python Data Science Handbook**, Jake VanderPlas
    - https://www.amazon.com/Python-Data-Science-Handbook-Essential-dp-1491912057/dp/1491912057/ref=mt_paperback?_encoding=UTF8&me=&qid=1577807451
  - **Python Data Science Essentials**, Alberto Boschetti and Luca Massaron
    - https://www.amazon.com/Python-Data-Science-Essentials-practitioners-dp-178953786X/dp/178953786X/ref=mt_paperback?_encoding=UTF8&me=&qid=1577807519
  - **Electronic versions of all four are available for free from the WPI library.**

Course activities

- **Lectures**: The lectures and *in class discussions* are an important part of the course.
  - The base lecture notes will be posted on the class web page before after class along with any annotations made during class.
- **Two Individual Projects**: Doing the case-studies is how you get experience solving interesting problems.
- **One Team Project**: This project is to be done in teams of 2-4 students, The team will work together to define, solve, summarize and present the outcomes.
- **Exams**: There will be a final exam, 2-3 in-class quizzes, and a final team work presentations.

Course requirements and grading standards

Individual Projects	35% (15% for Project 1 and 20% for Project 2)
Team Project	25%
Final exam	25%
Quizzes	15% (2-3 quizzes)

The final exam and quizzes will be in class, noncumulative, and open note, but **no collaboration will be allowed** and the exams be graded based upon demonstrated understanding of key concepts.

The team project will be performed in **groups of 2-4** and will be graded based upon the quality and completeness of presentations and the submitted reports.

I reserve the right to curve the final grades (either up or down) based upon the aggregate performance of the class.

Advice on doing the projects

- ***You will be expected to hand in an Jupyter notebook and a set of presentation slides for each project.***

  - Make sure that the presentation slides are stand alone. I.e., make sure that your slides tell a story that you don't need to read your iPython notebook to understand!

  - In addition, make sure that your iPython notebook fully documents your projects! Think of the iPython notebook as a written report that you are giving to the CEO of your company.

  - In other words, the iPython notebook and presentation are independent deliverable, and they should each be able to stand on their own.

- Make sure it is clear where **each part** of **each question** is answered.

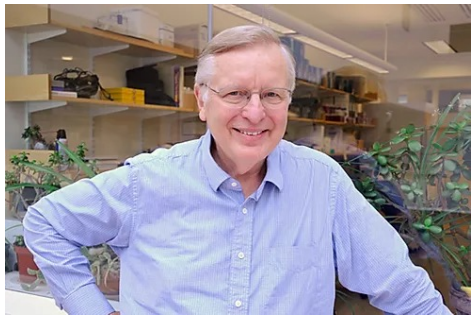
- The details are being finalized, but you will likely be grading each other on the team project presentations.

Success

Your success in life will be determined largely by your

- ability to speak,
- ability to write,
- and the quality of your ideas

in that order!



https://www.youtube.com/watch?v=Unzc731iCUY

Patrick Henry Winston,
Ford Professor of Artificial Intelligence and Computer Science
MIT

Course details available online, along with a lot of other stuff…

http://canvas.wpi.edu

Course Materials

- Textbooks
  - No Textbook.
- Slides
  - Will be posted on the class website after each class

Objectives for today

- Discuss the course mechanics (syllabus, grading, etc.)
- Start to get inspired Data Science!

Computational **Data** Intelligence

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” …

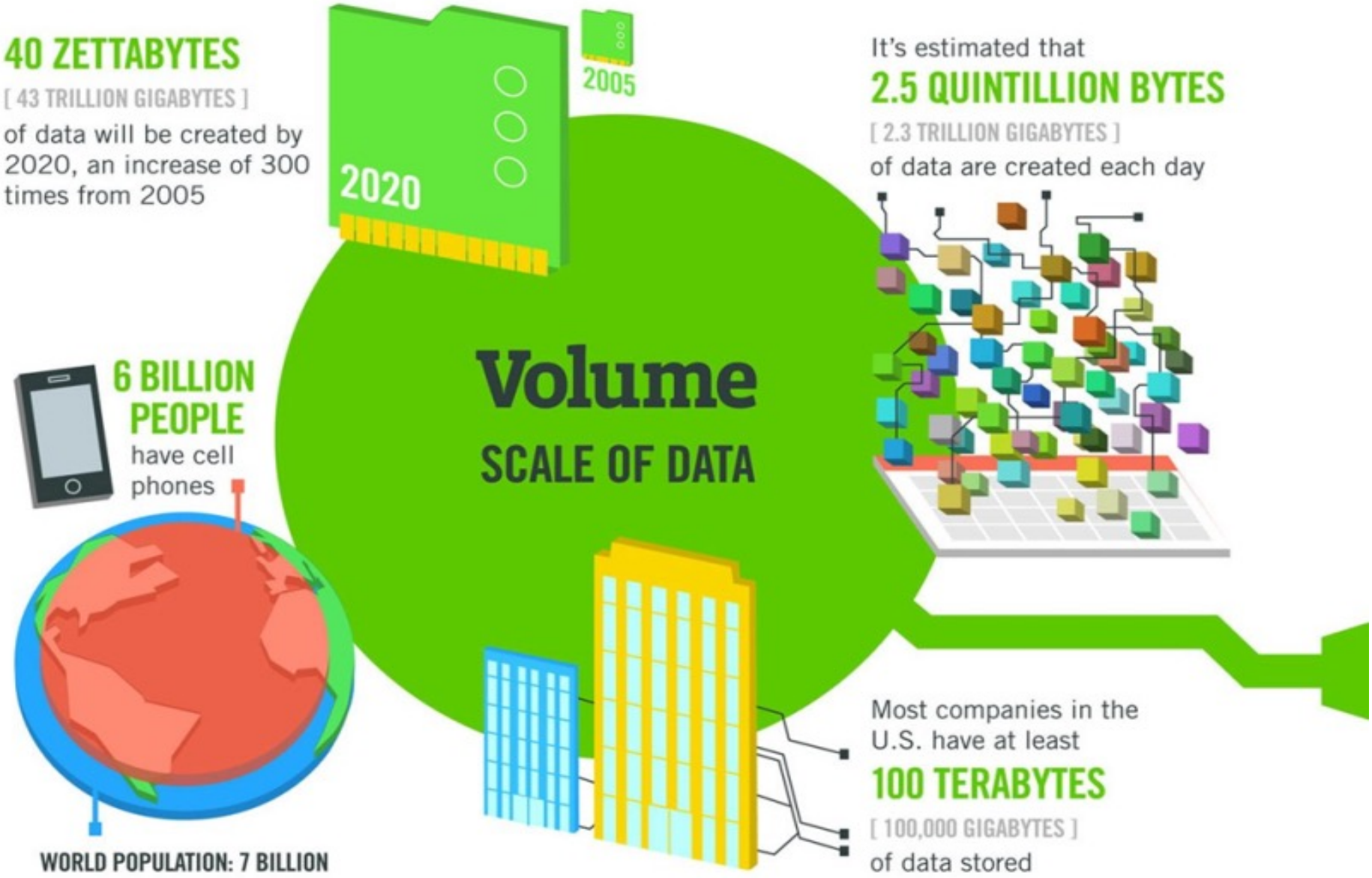
“***Big Data***” is data whose scale, diversity, complexity, and/or quality require new architectures, techniques, algorithms, analytics, and interfaces to manage it and extract value and hidden knowledge from it…

Why Now?

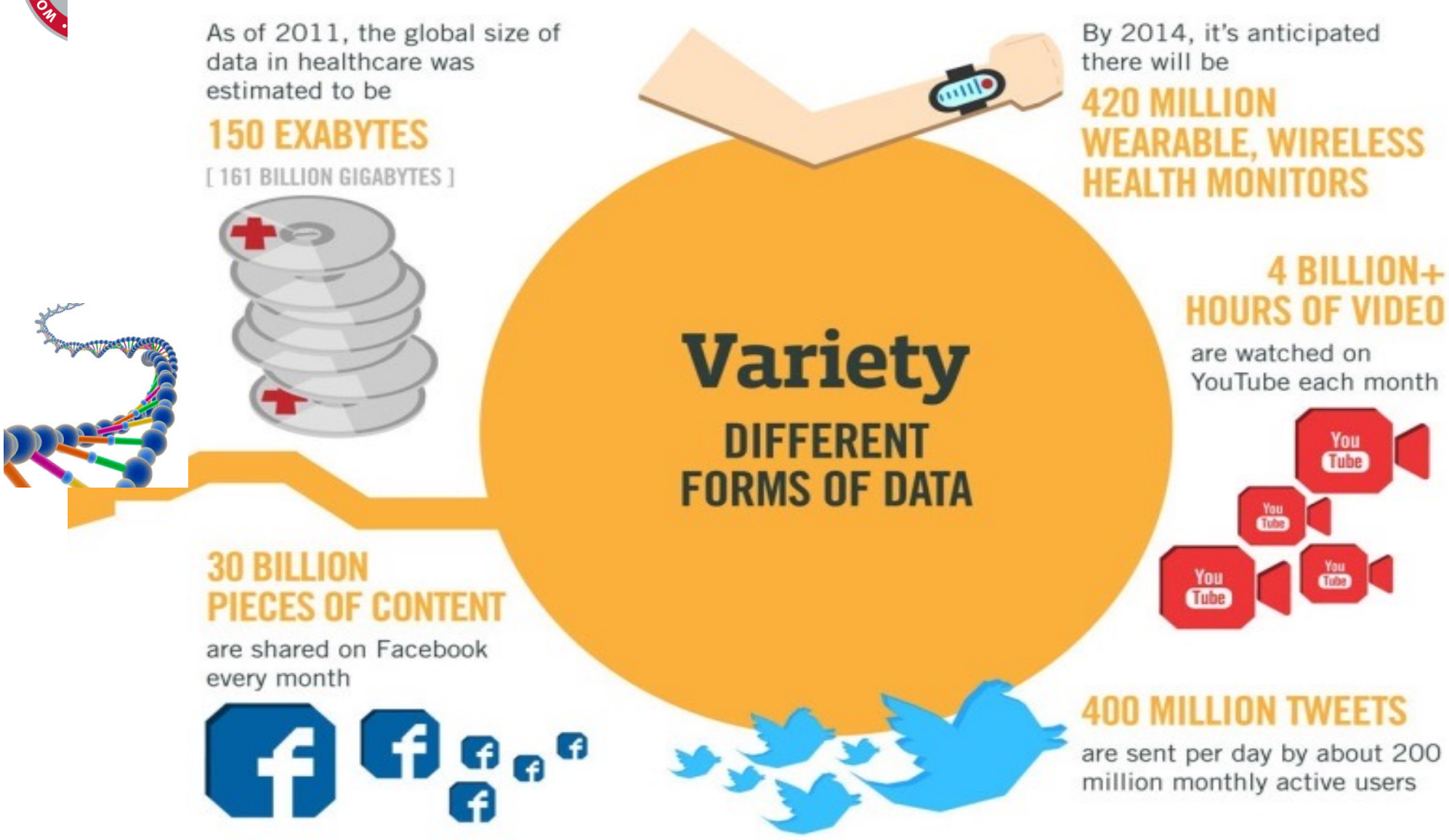
Big Data and Big Challenges

Big Data

- Volume
- Variety
- Velocity
- Veracity



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



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

The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

during each trading session

Modern cars have close to

100 SENSORS

that monitor items such as fuel level and tire pressure

Velocity

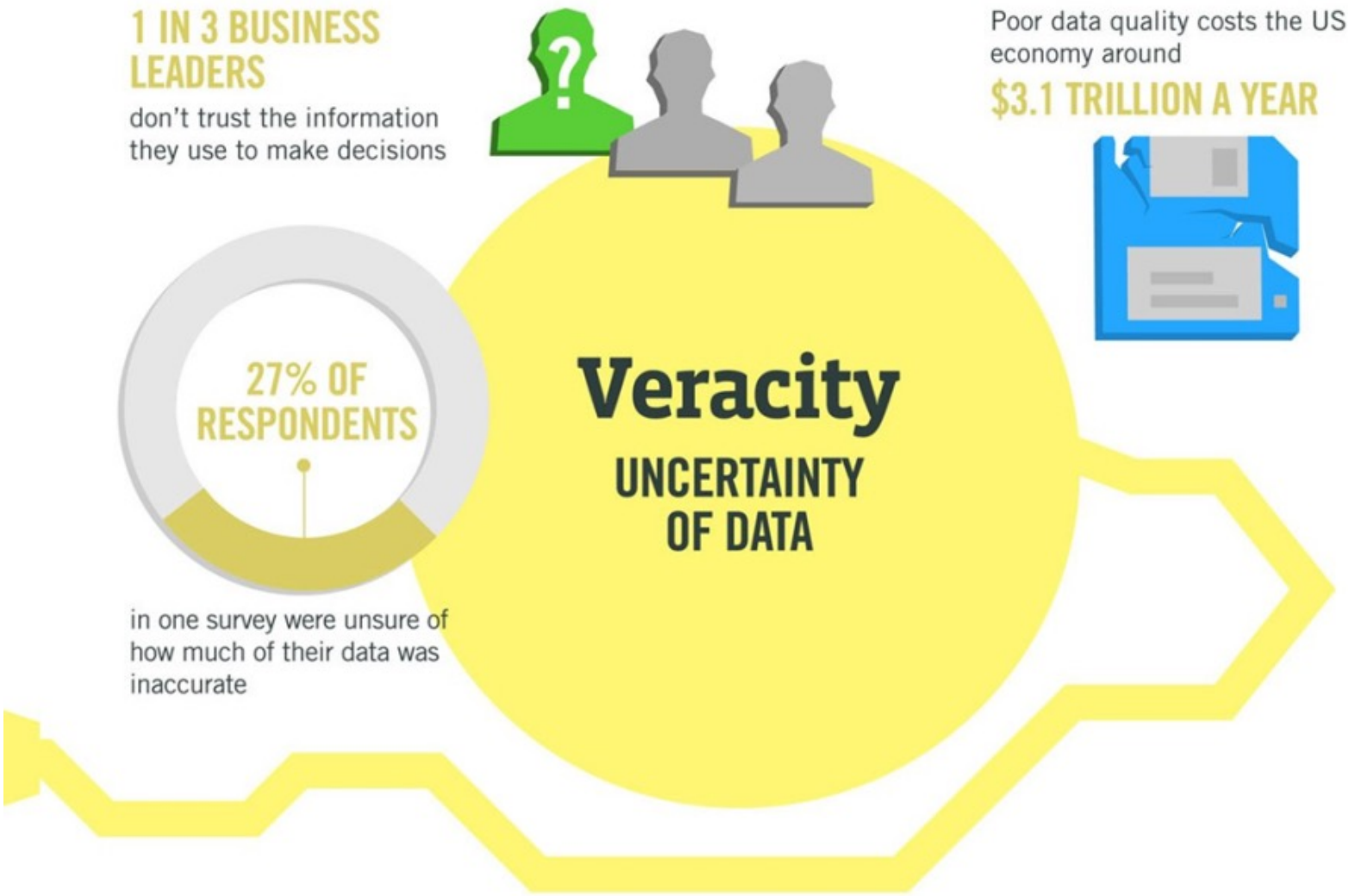
ANALYSIS OF STREAMING DATA

By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

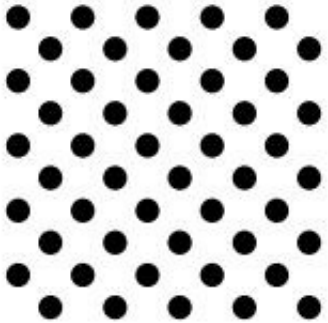
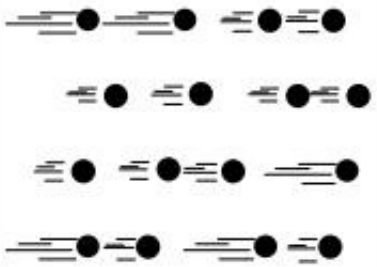
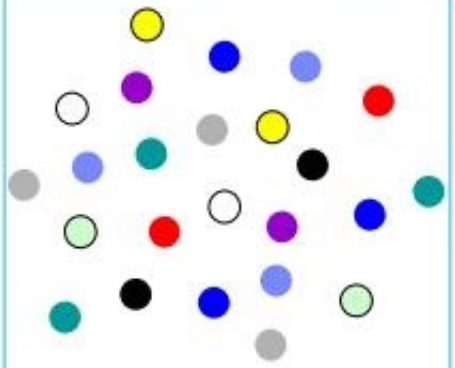
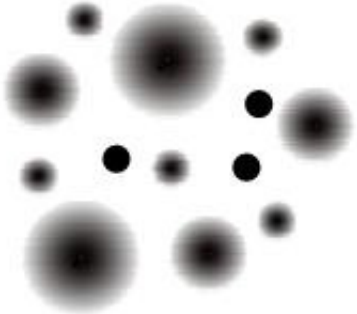
– almost 2.5 connections per person on earth

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-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

- **E.g., YouTubers, Twitter account**

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

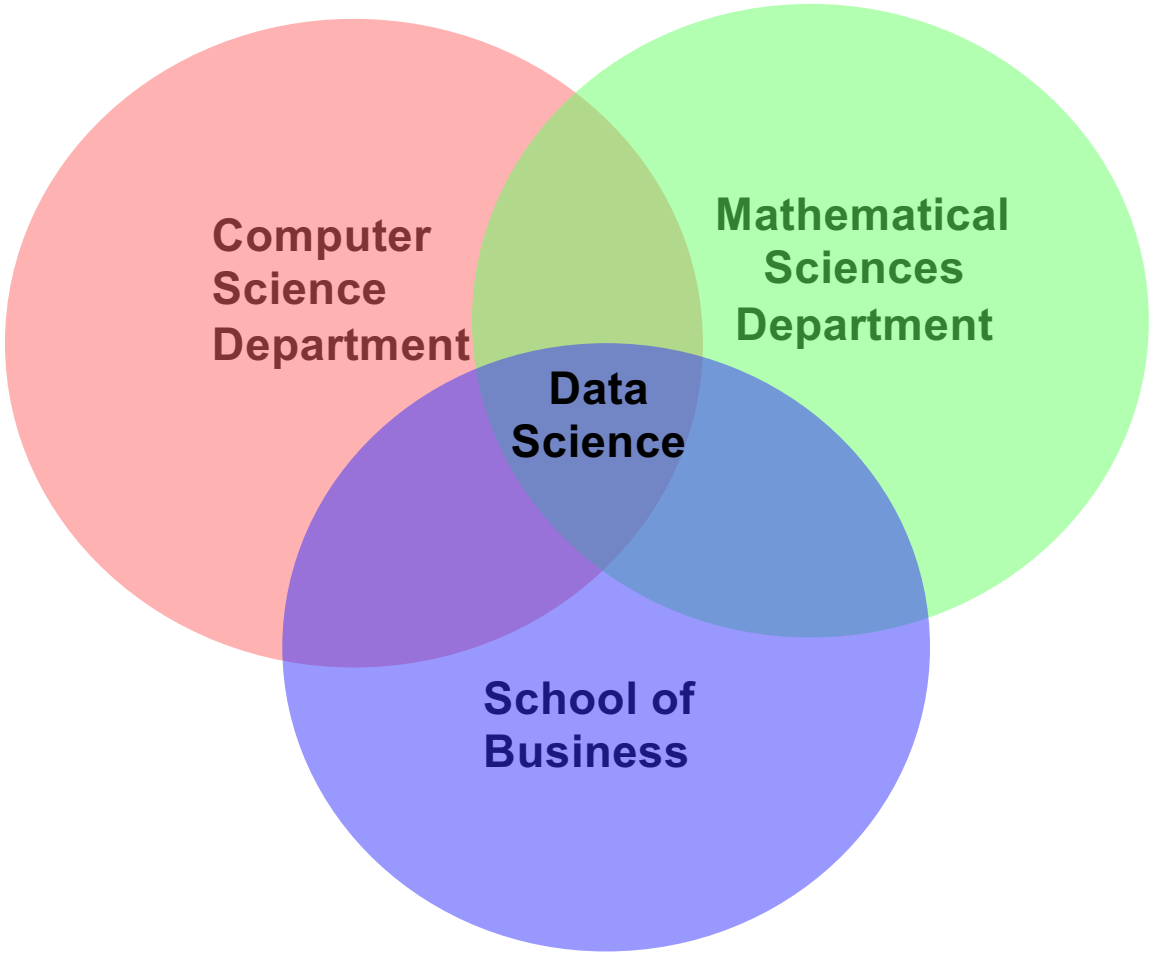
Data Science?

Data Science

Data science combines multiple fields, including statistics, scientific methods, artificial intelligence (AI), and data analysis, to extract value from data, and use the value for applications.

From Oracle: https://www.oracle.com/data-science/what-is-data-science/

What is Data Science? One view...



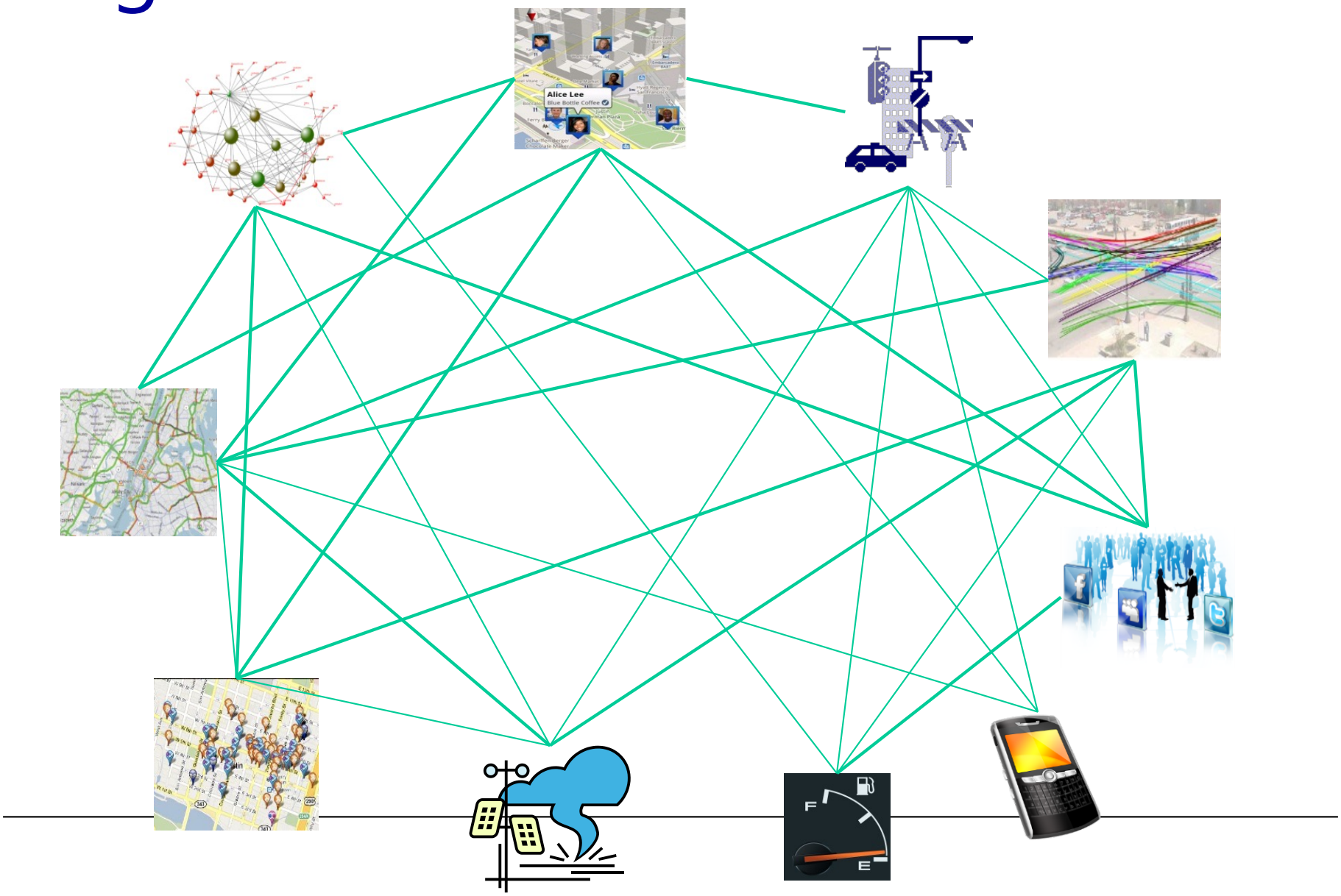
- Based upon Drew Conway's Data Science Venn Diagram
- http://en.wikipedia.org/wiki/Data_science

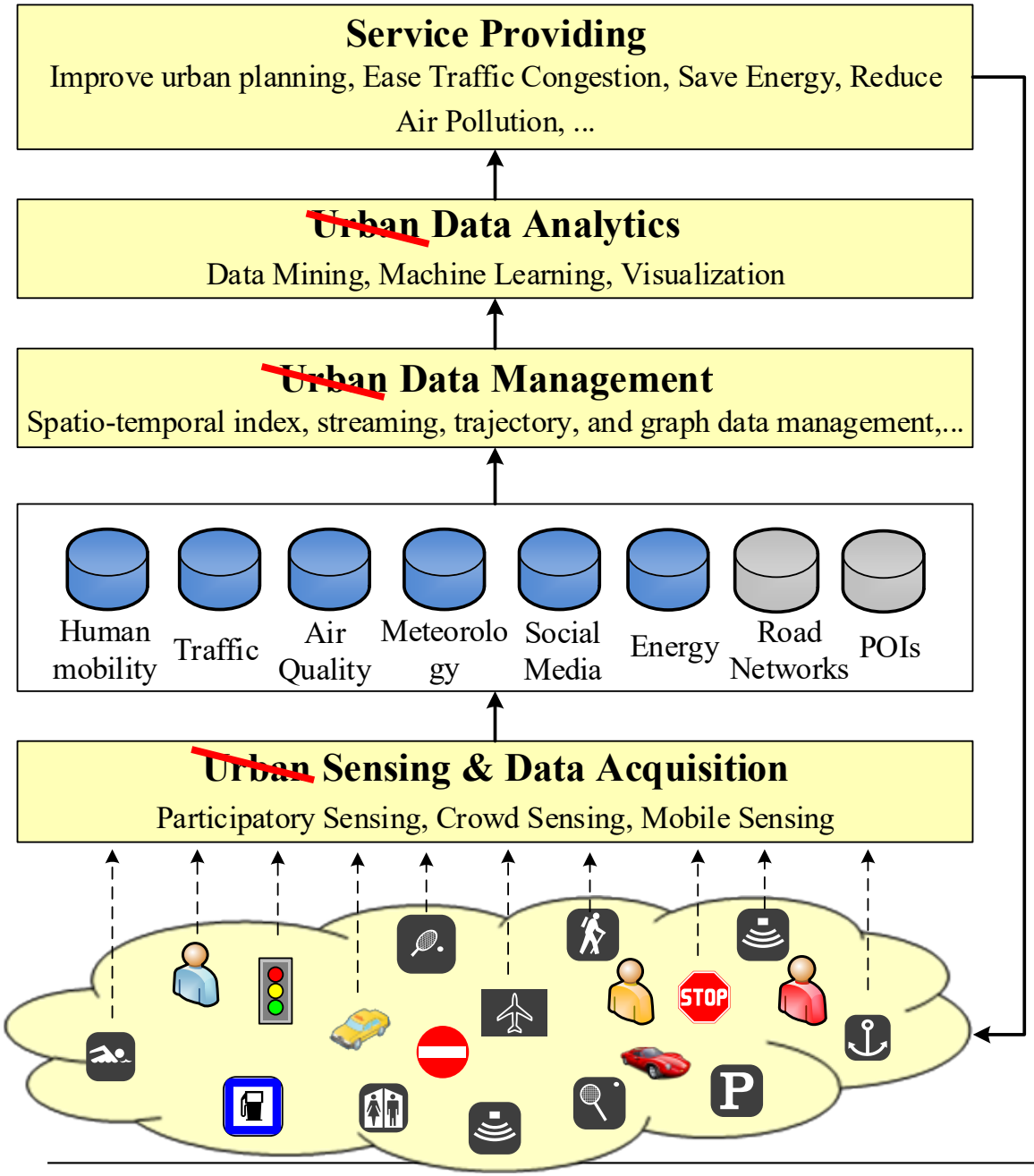
- http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

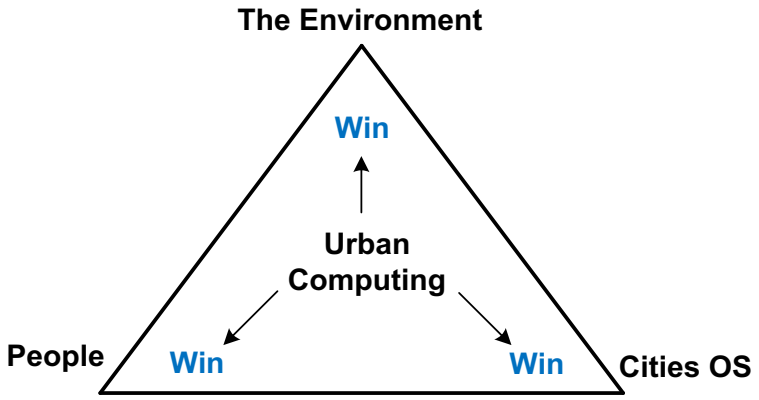
Some application stories

Big Challenges in Big Cities

Big Data in Cities

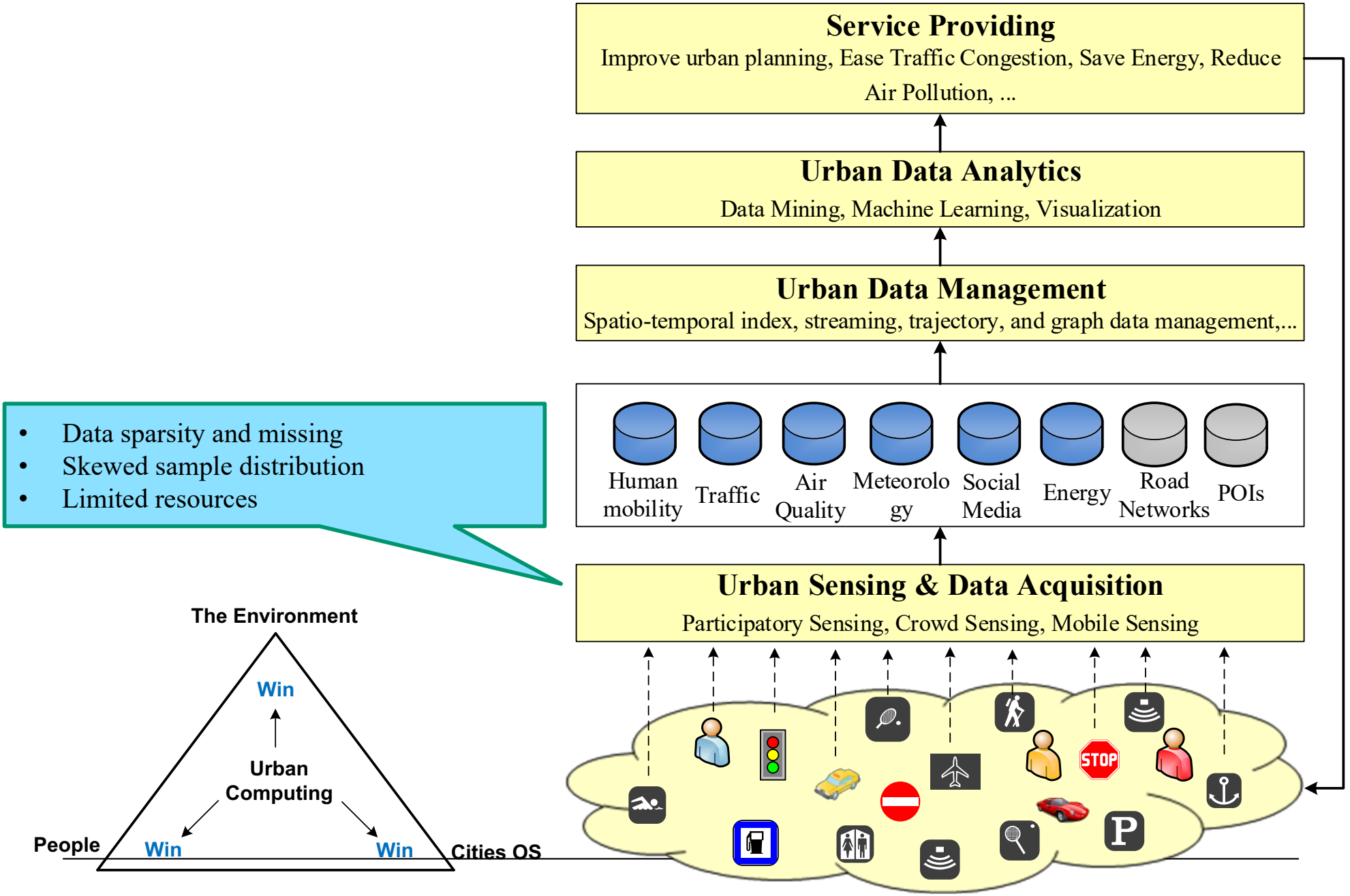






Tackle the Big challenges in Big cities using Big data!

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

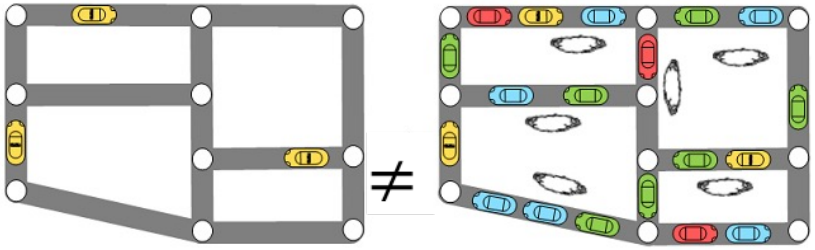


Zheng, Y., et al. Urban Computing: concepts, methodologies, and applications. *ACM transactions on Intelligent Systems and Technology*.

Urban Sensing

A sample of data → An entire dataset

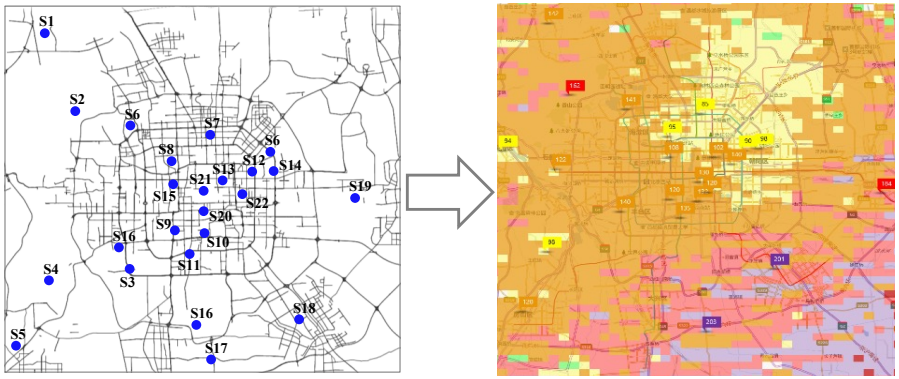
- Biased distribution



Taxi flow Entire traffic flow

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

- Data sparsity and missing



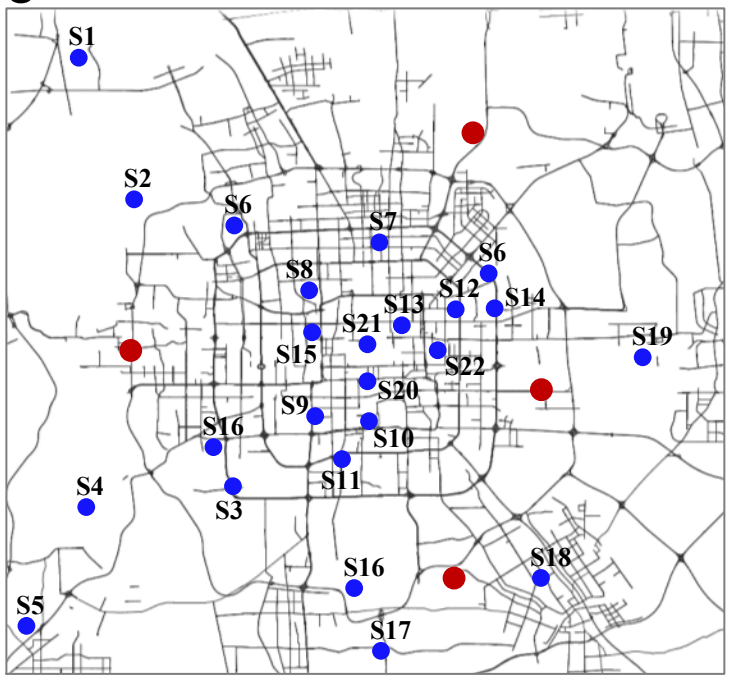
Air quality monitoring stations

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

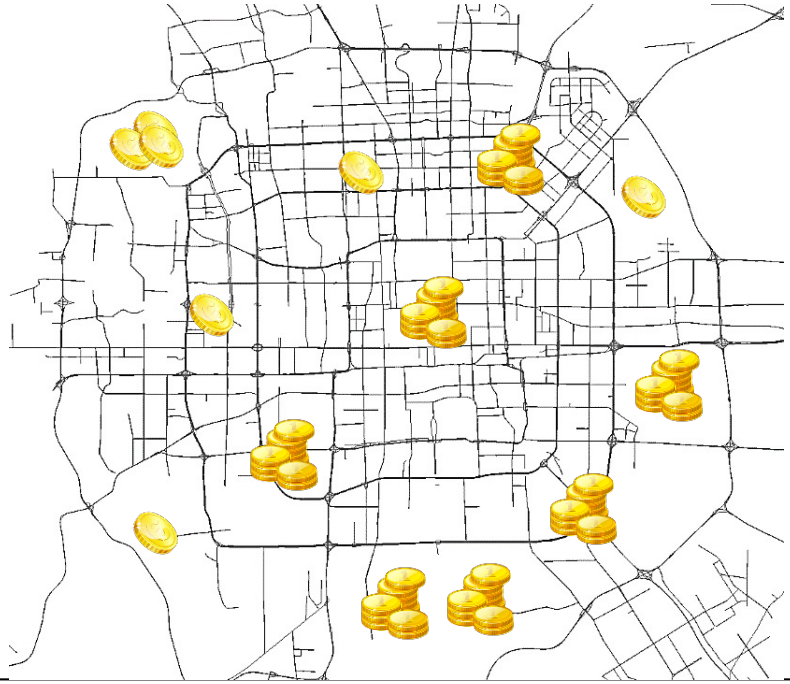
Urban Sensing

A limited resource (budget, labors, land…)

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

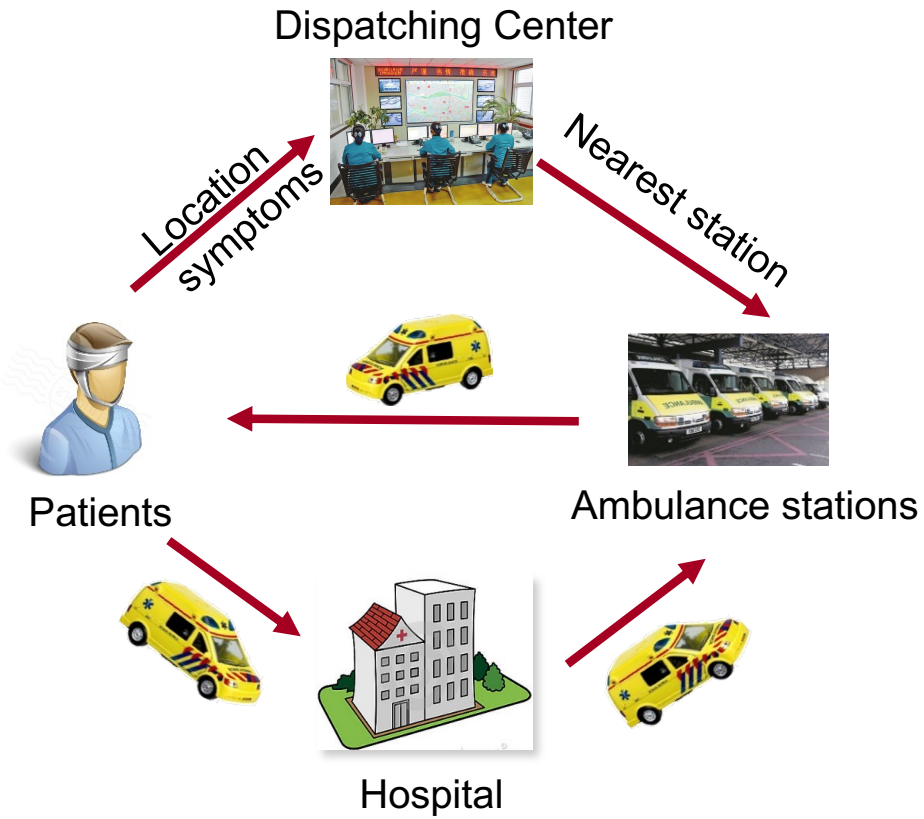


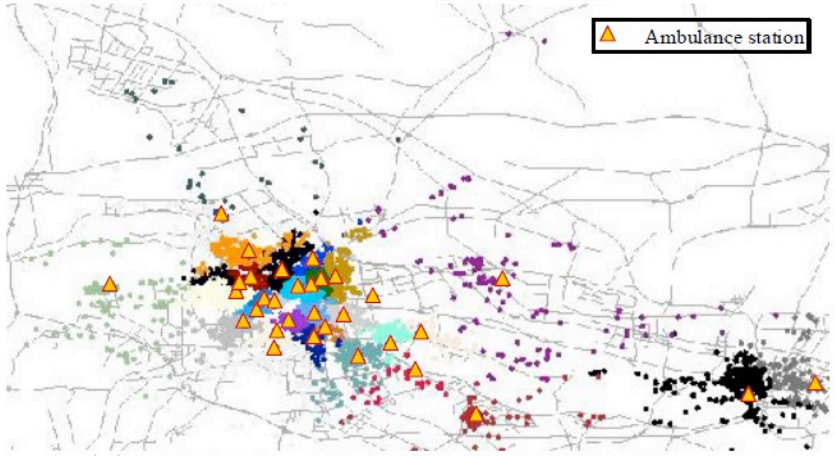
- Crowdsensing: How to arrange the incentives dynamically?



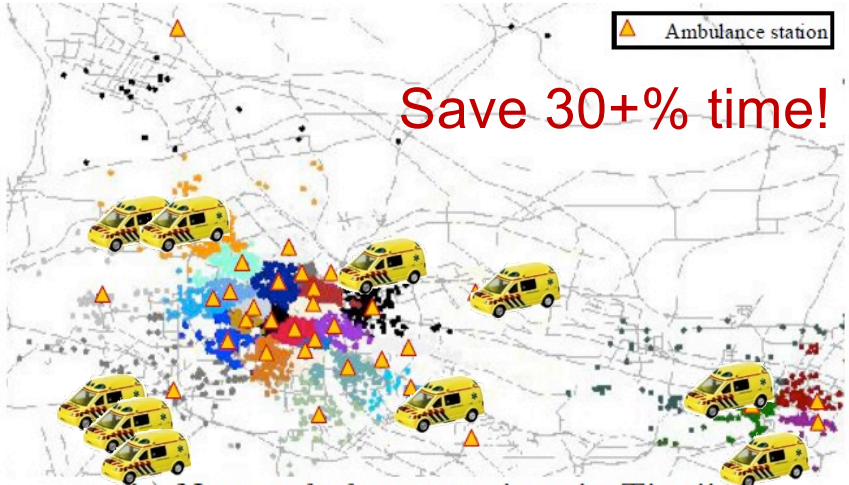
Suggesting locations for monitoring stations, KDD 2015

Improving Medical Emergency Services using Big Data





(a) Original ambulance stations in Tianjin

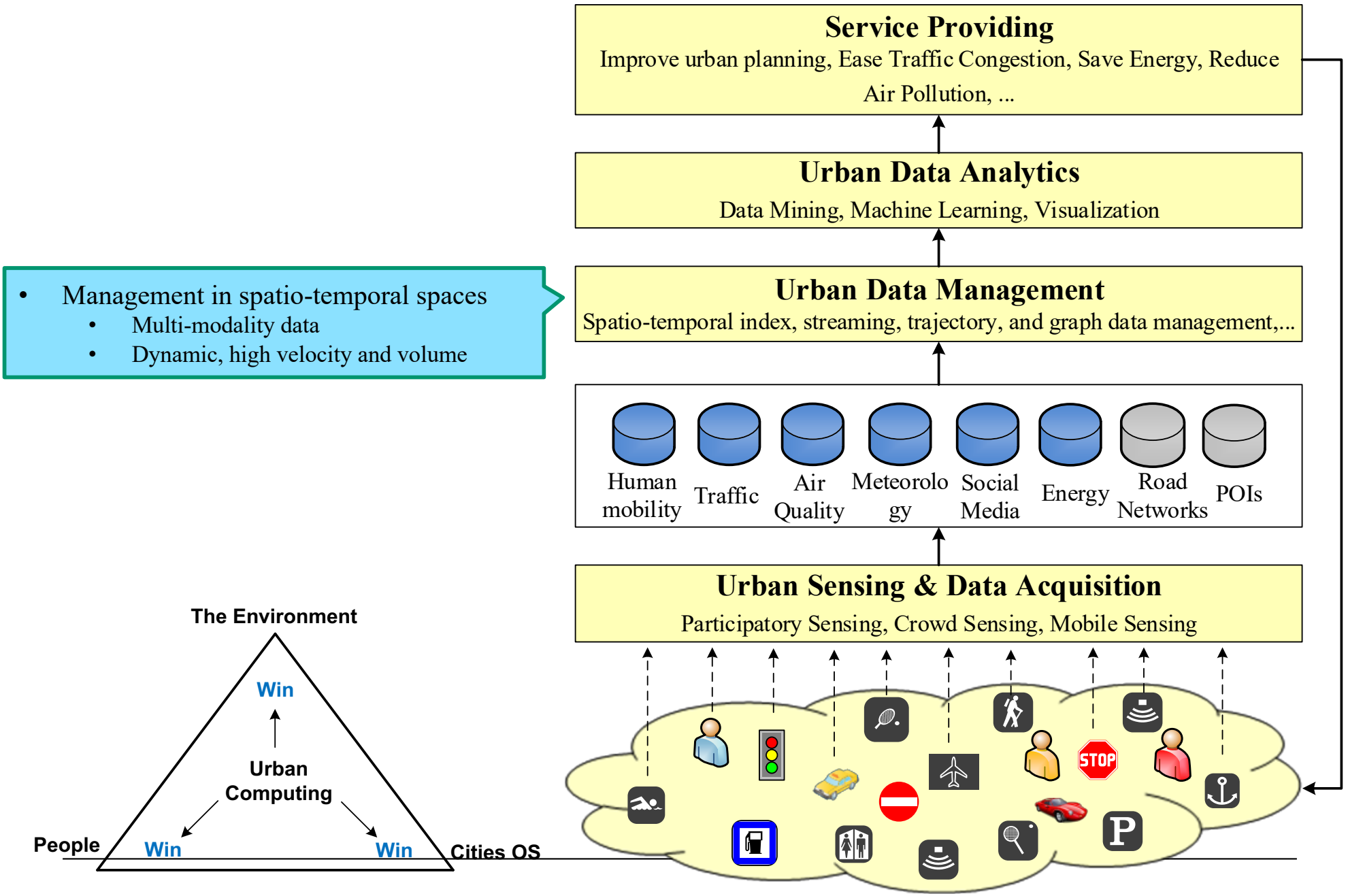


(b) New ambulance stations in Tianjin

- 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

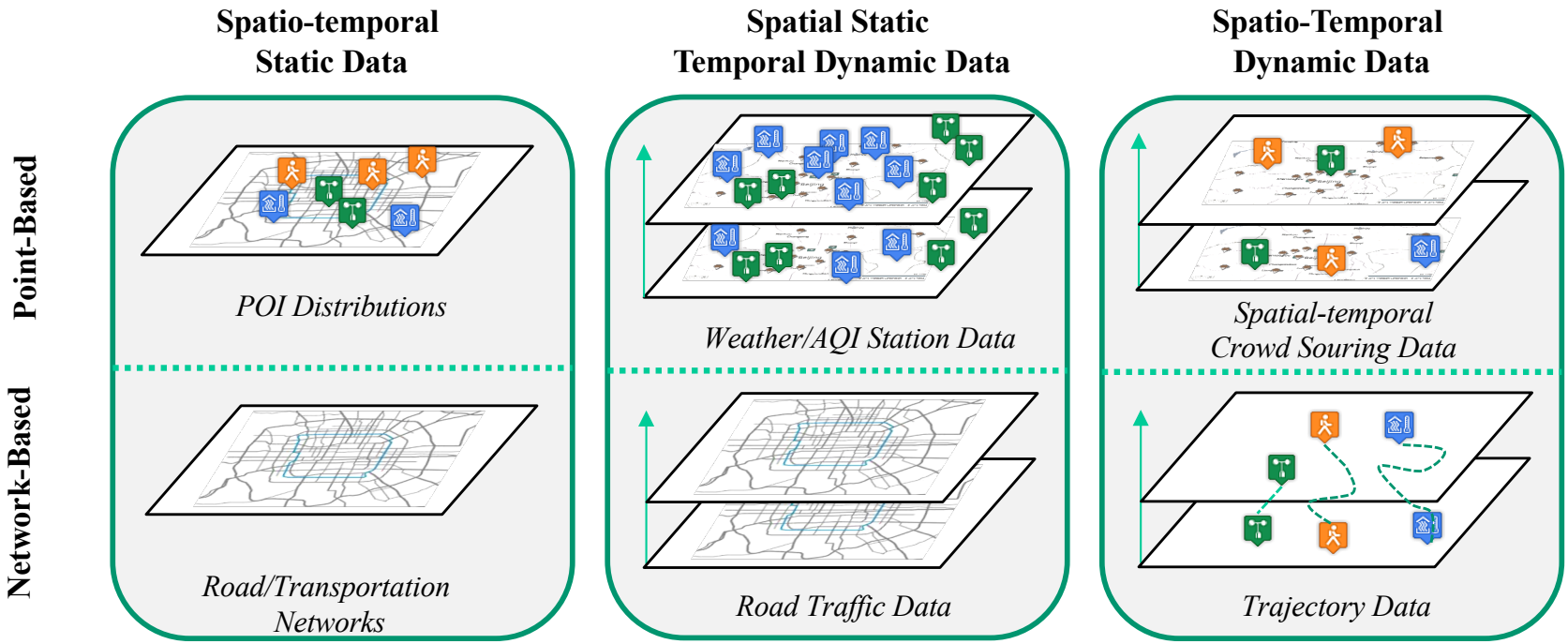
Location Selection for Ambulance Stations: A Data-Driven Approach, ACM SIGSPATIAL 2015



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

- Texts and images → spatial and spatio-temporal data;
- A single data source → Data cross different domains
- Separate data mining algorithms → machine learning + data management

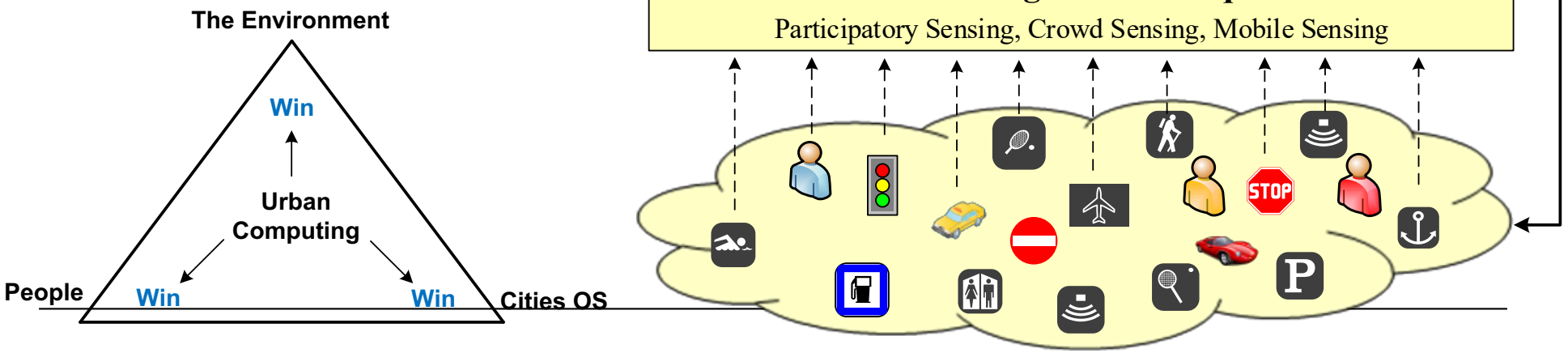
Service Providing
Improve urban planning, Ease Traffic Congestion, Save Energy, Reduce Air Pollution, ...

Urban Data Analytics
Data Mining, Machine Learning, Visualization

Urban Data Management
Spatio-temporal index, streaming, trajectory, and graph data management,...

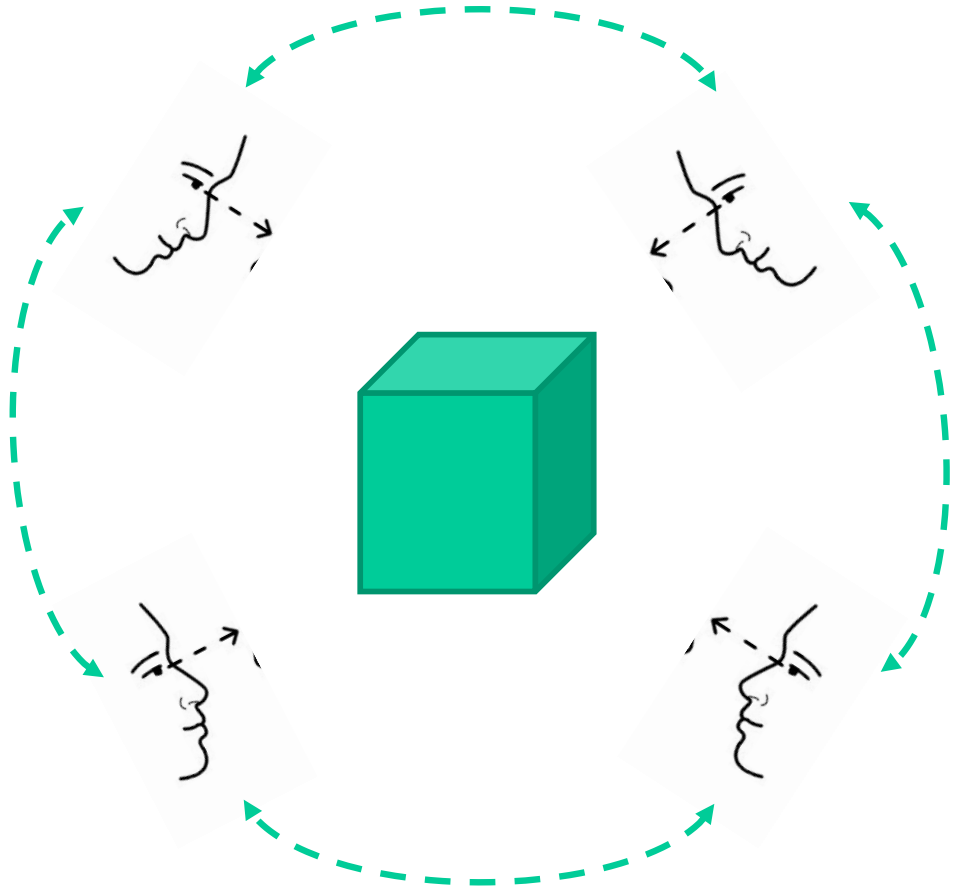
Human mobility | Traffic | Air Quality | Meteorology | Social Media | Energy | Road Networks | POIs

Urban Sensing & Data Acquisition
Participatory Sensing, Crowd Sensing, Mobile Sensing



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

Multi-View-Based Learning

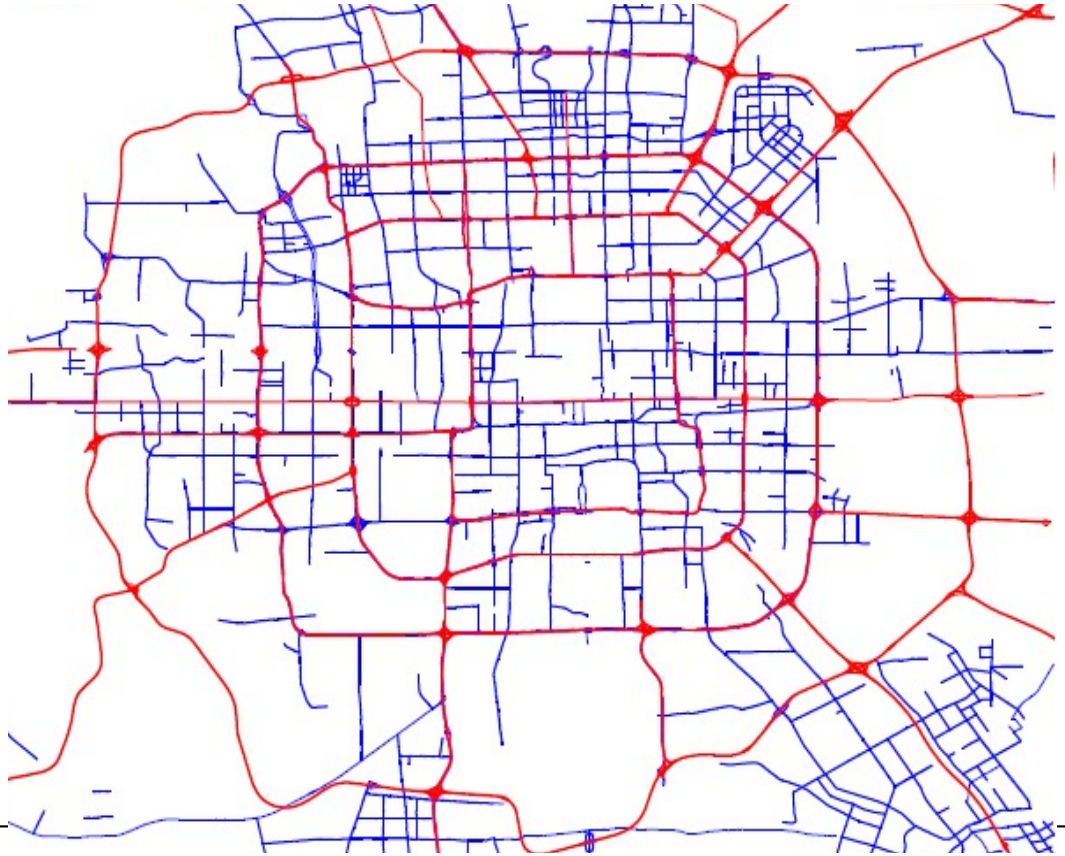


Urban Computing for Urban Planning

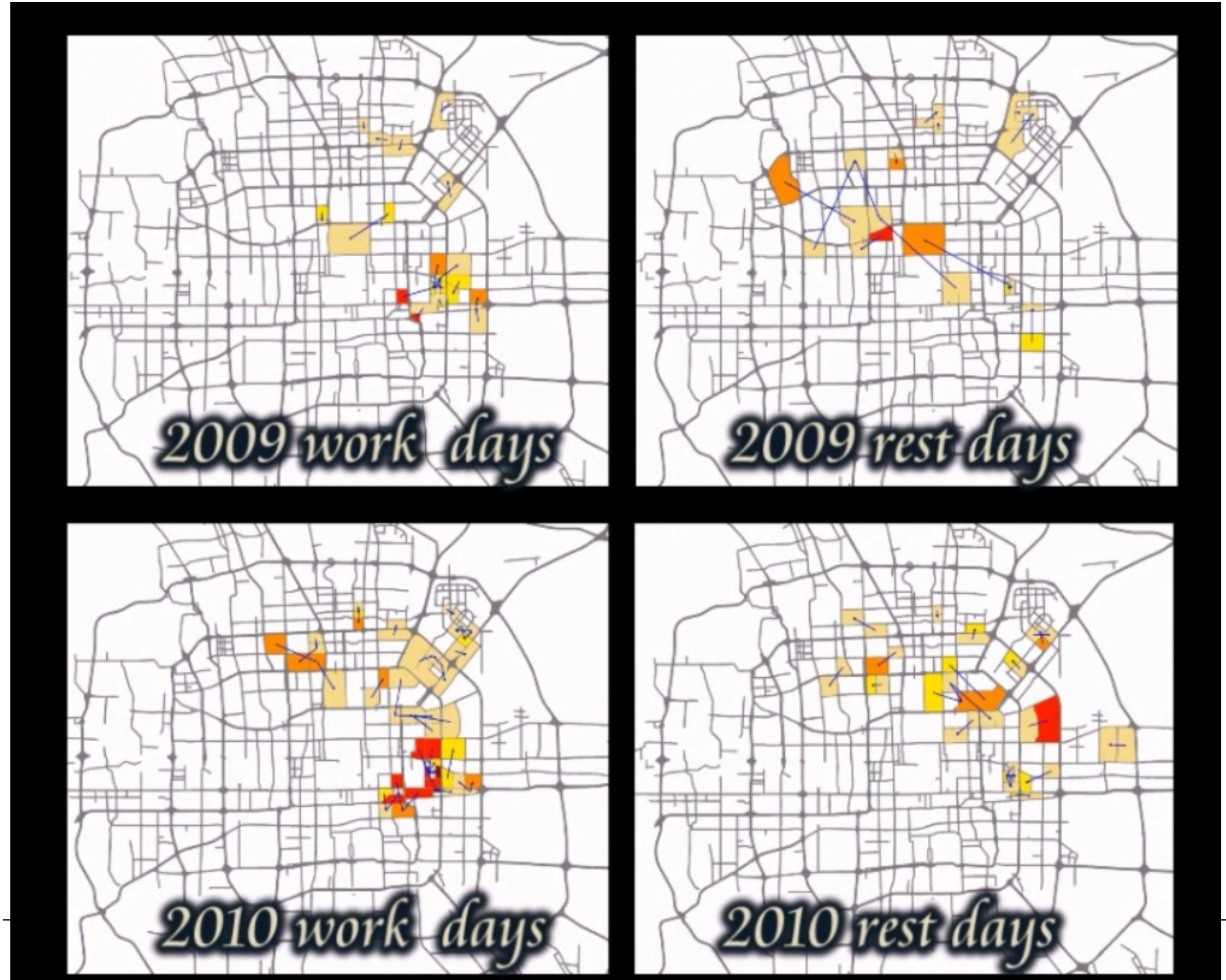


City-Wide Traffic Modeling

- 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**



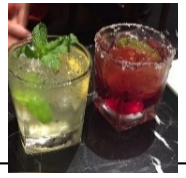
Yu Zheng, et al. Urban Computing with Taxicabs, In Proc. Of **UbiComp 2011**

Shanghai Big Data Hotpot Restaurant









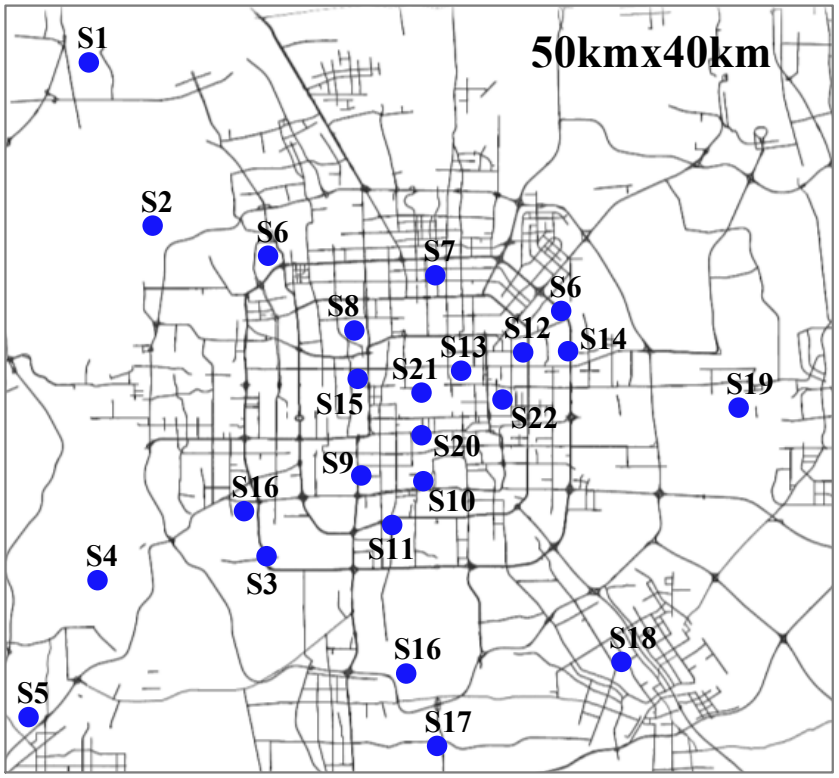


When Urban Air Meets Big Data

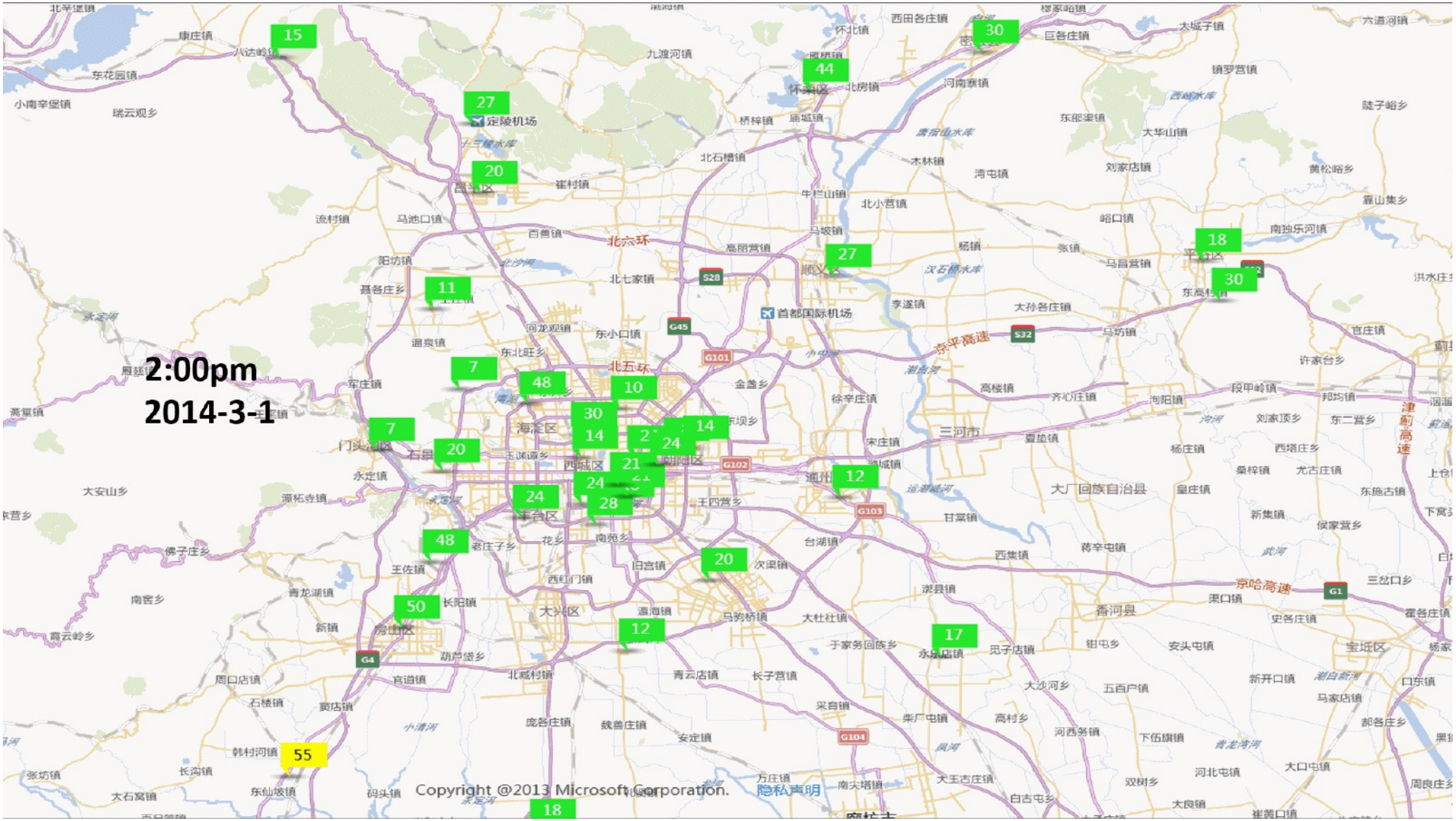
Air Pollution: A Global Concern !

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

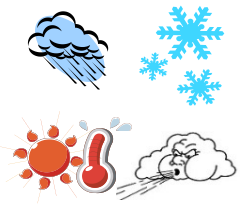
● Air quality monitor station



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



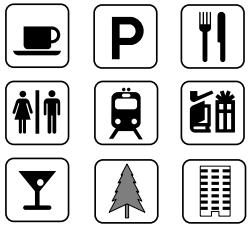
Inferring **Real-Time** and **Fine-Grained** air quality throughout a city using **Big Data**



Meteorology

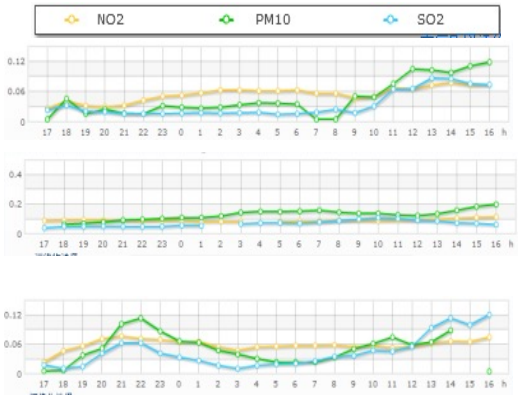
Traffic

Human Mobility



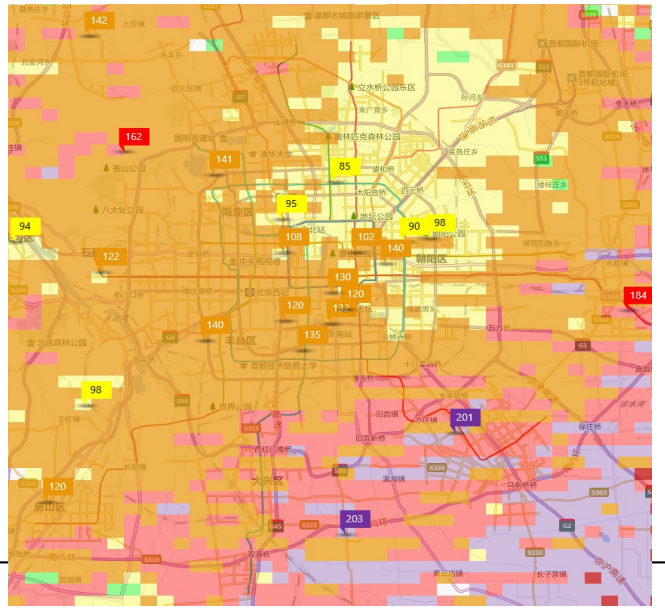
POIs

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

Revisit Data Science

- **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 itselfMany unsolved problems

Big Data ≠ Mining Single Dataset ≠ Simple Statistics

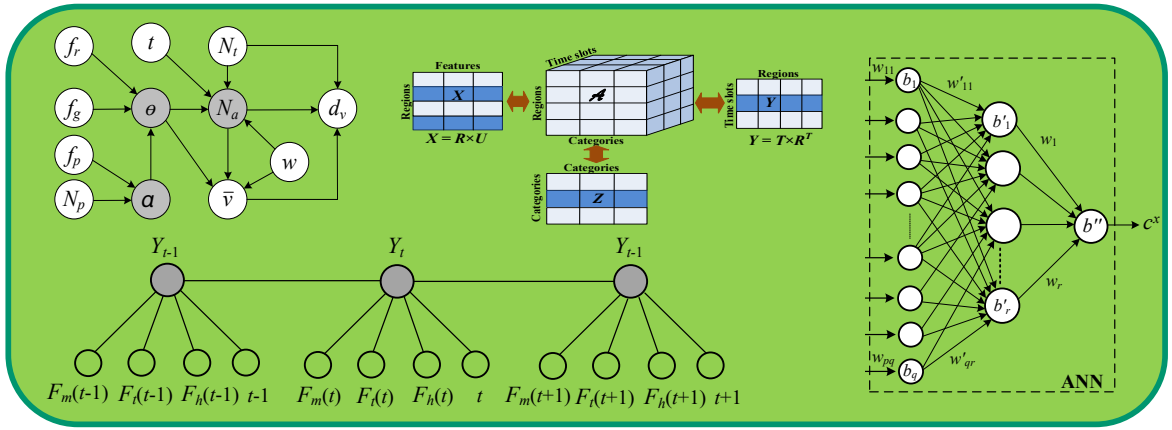
Big Data ≠ Machine learning ≠ Deep Learning

Big Data ≠ Cloud Computing ≠ Hadoop

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

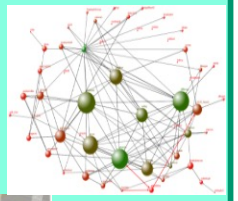
Data Scientist

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.~~

Questions?