CS 3319 Foundations of Data Science

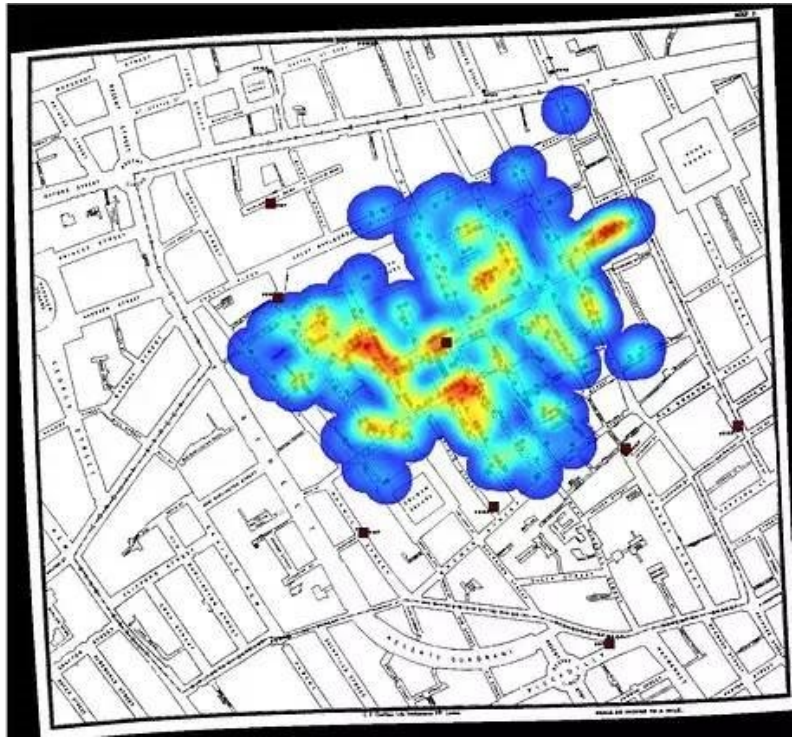
10. Spatio-Temporal Data

Jiaxin Ding

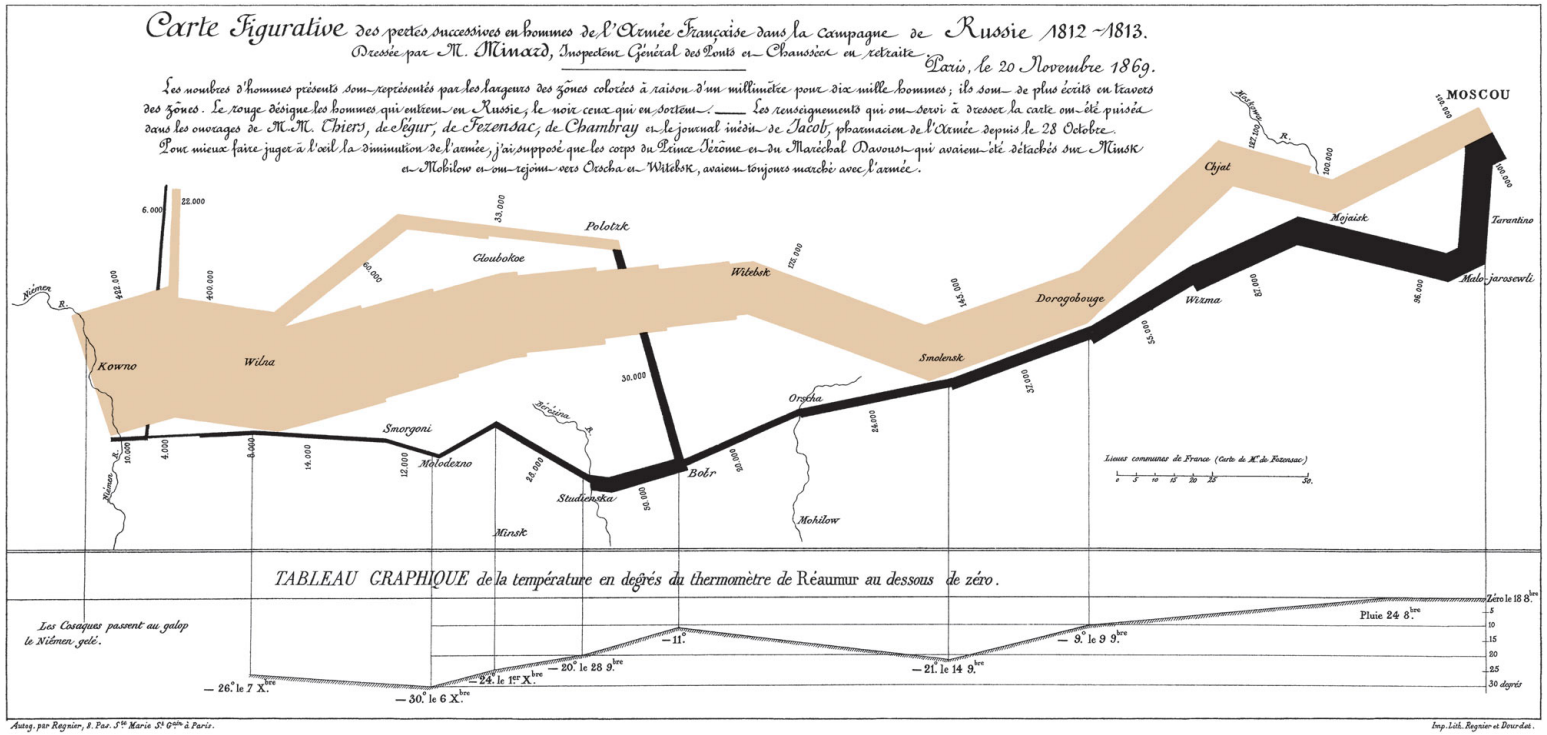
John Hopcroft Center

Spatio-Temporal Data

- Different data sources
  - Geographical map data



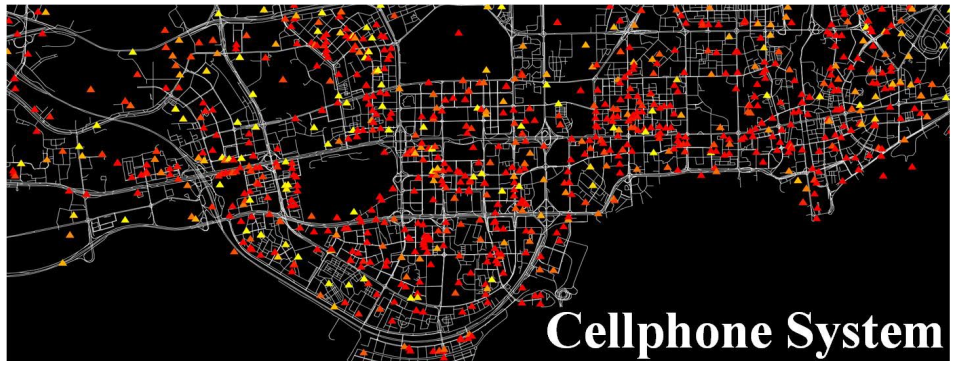
1854 Broad Street Cholera Outbreak

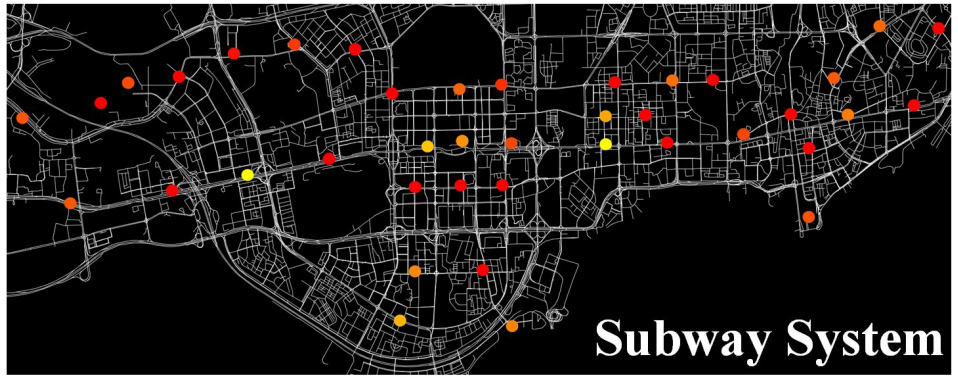


Charles Minard's map of Napoleon's disastrous Russian campaign of 1812

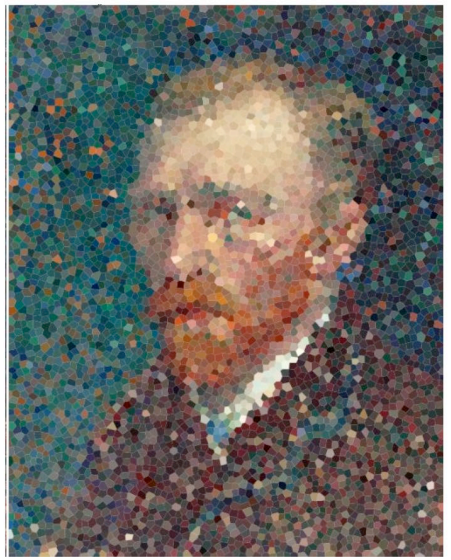
Spatio-Temporal Data

- Different data sources
  - Internet of Things data: static vs mobile

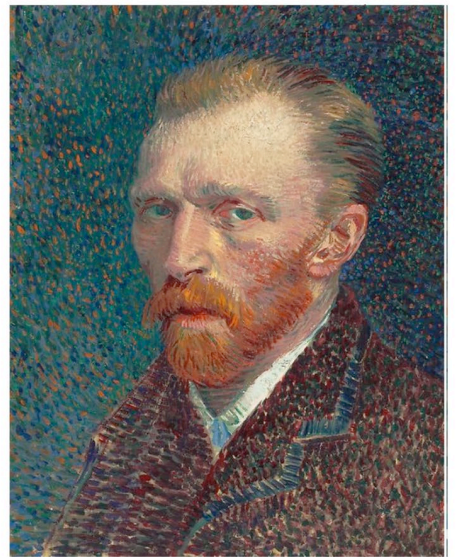




CDRs



GPS



From https://www.cs.rutgers.edu/~dz220/

Spatio-Temporal Data

- Different data sources
  - Social media, e.g. Foursquare Datasets

Spatio-Temporal Data

- Different data types
  - Structured data, semi-structured data, unstructured data
  - Values, semantics information

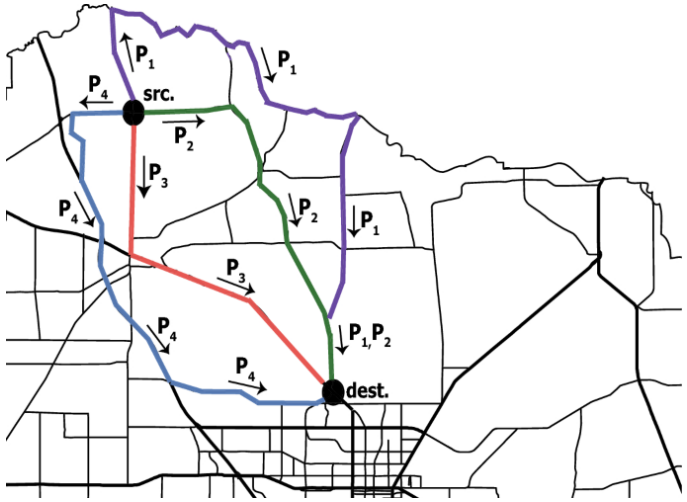
time	Longitude	Latitude
2020-12-14 00:09:01	121.43376767	31.0254039608
2020-12-14 12:00:30	121.43189012	31.0228388124
2020-12-14 14:13:23	…	…

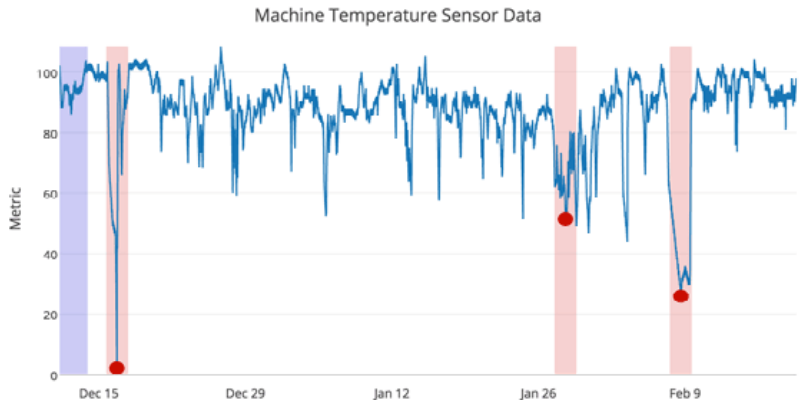
```
<Time_Metadata>
  <TimeCoverage>
    <StartTime>
        2020-12-14 00:09:01
    </StartTime>
    <EndTime>
        2020-12-14 12:00:30
    </EndTime>
    <Longitude>
        121.43189012
    <Longitude>
    ……
```

Spatio-Temporal Data

- Different ways of processing
  - Batch
  - Graph
  - Stream

time	x	y

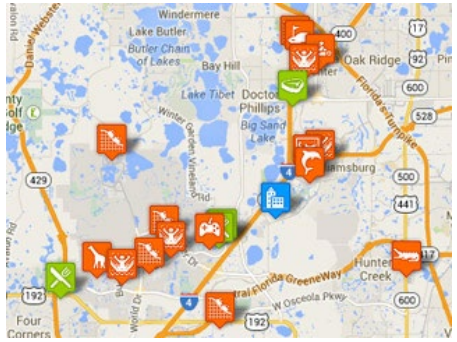


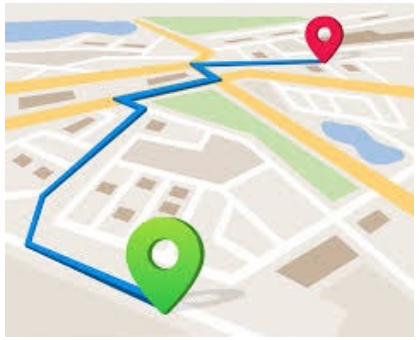


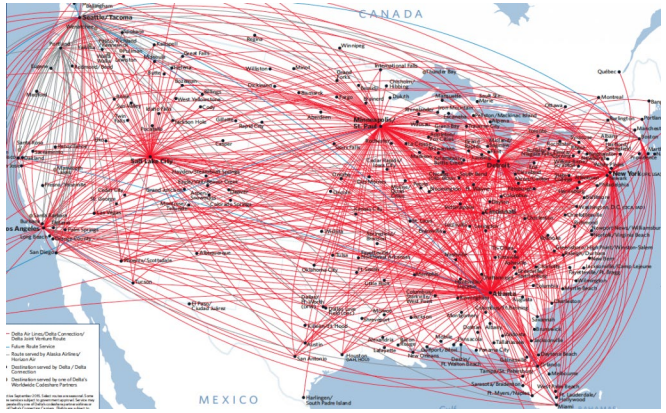
Spatio-Temporal Data

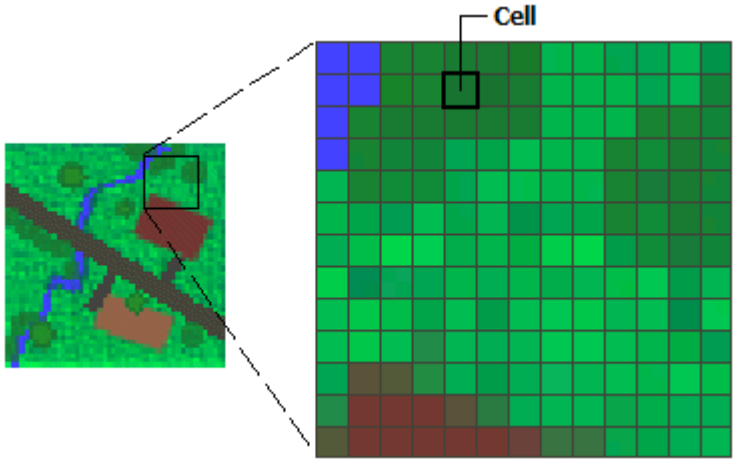
- Points
  - Point of Interests (POI)
- Lines
  - Route
- Graphs
  - Traffic networks,

    e.g. road networks, airlines
- Rasters
  - Partition into cells



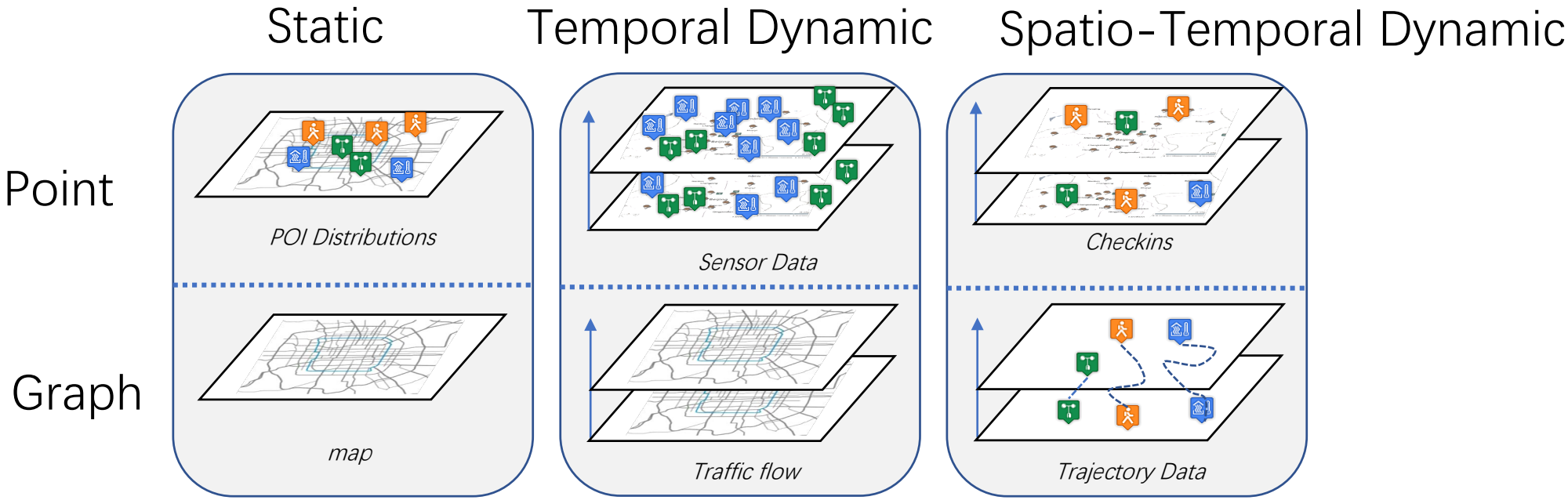






Spatio-Temporal Data

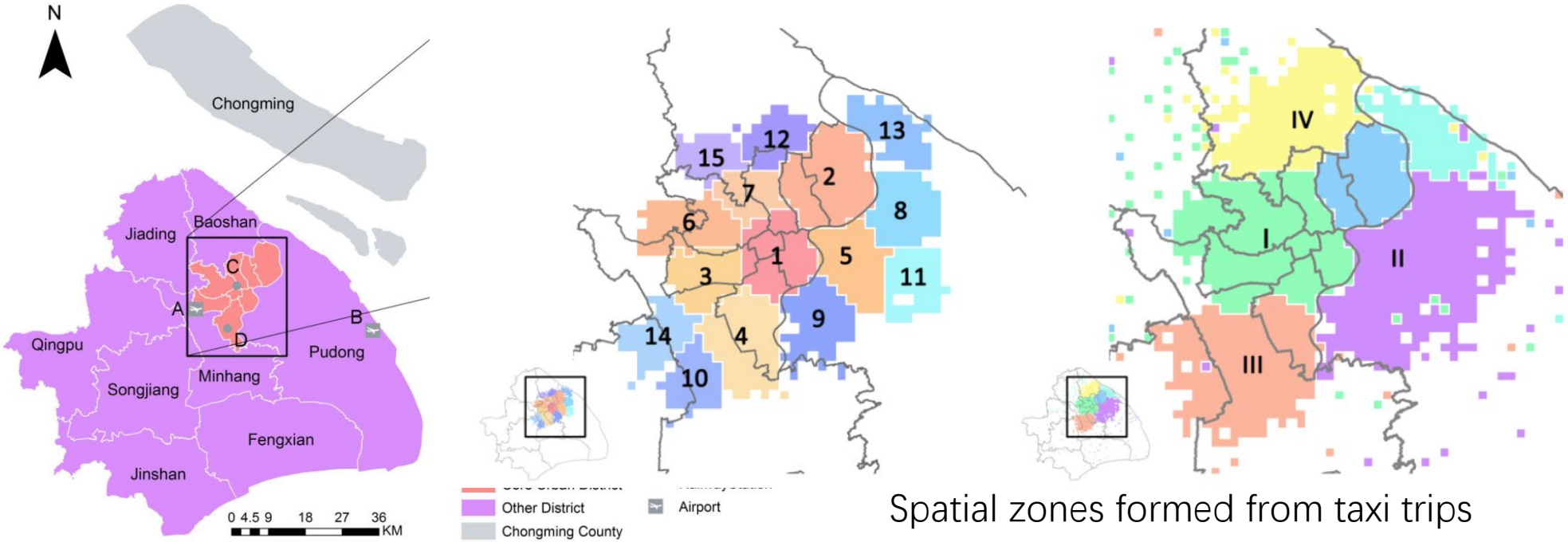
- Spatio-temporal: static/dynamic



Spatial and Temporal Data Characteristics

Spatial Characteristics

- Spatial hierarchy
  - Different spatial granularities
  - Hierarchical structures

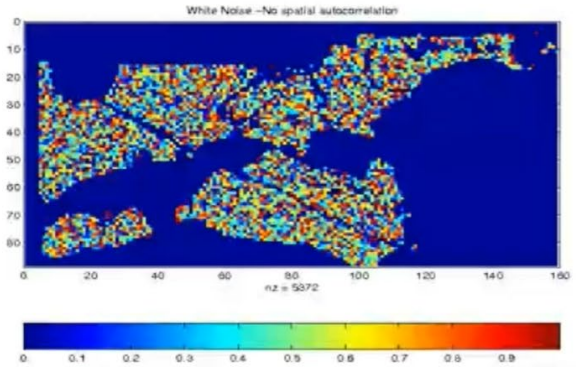


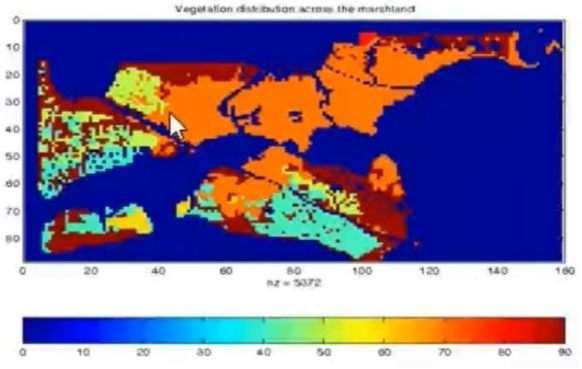
Spatial zones formed from taxi trips

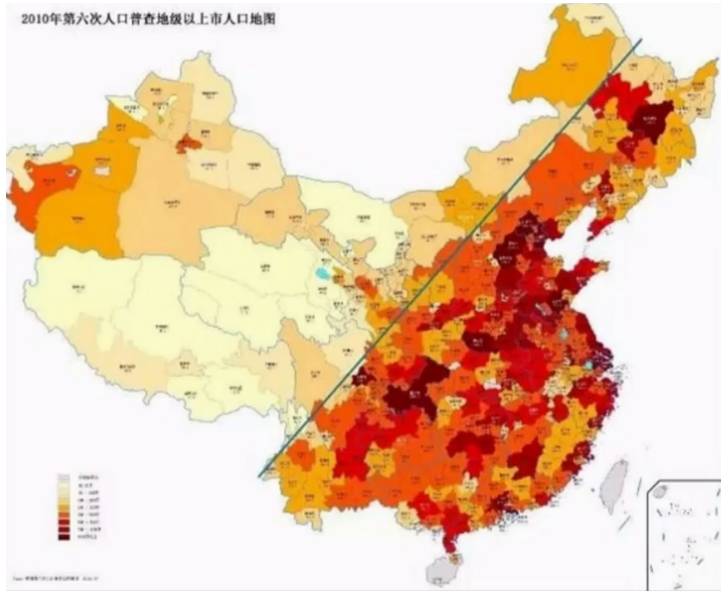
X Liu, L Gong, Y Gong, et al. Revealing travel patterns and city structure with taxi trip data[J]. Journal of Transport Geography, 2015, 43: 78-90.

Spatial Characteristics

- Spatial Auto-Correlation
  - First Law of Geography
    - All things are related, but **nearby things are more related** than distant things.
  - i.i.d assumption is typically not valid
    - i.i.d vs **auto-correlation**

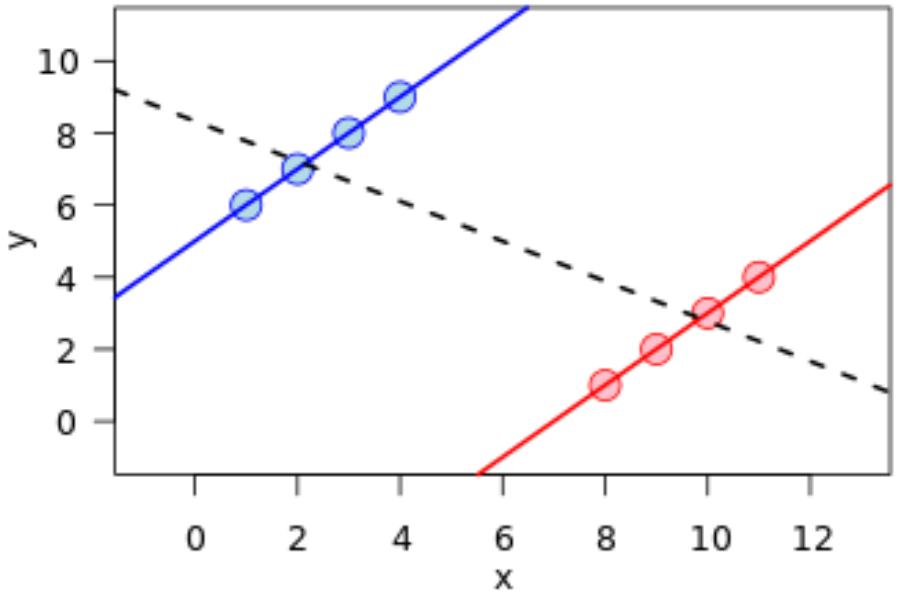


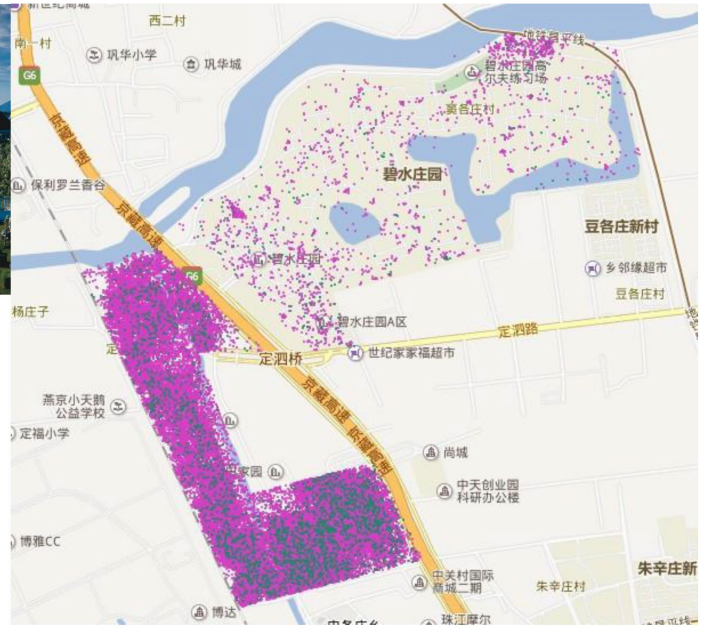




Spatial Characteristics

- Spatial **heterogeneity**
  - "Second law of geography"
  - Global model might be **inconsistent** with regional models
    - Spatial Simpson's Paradox

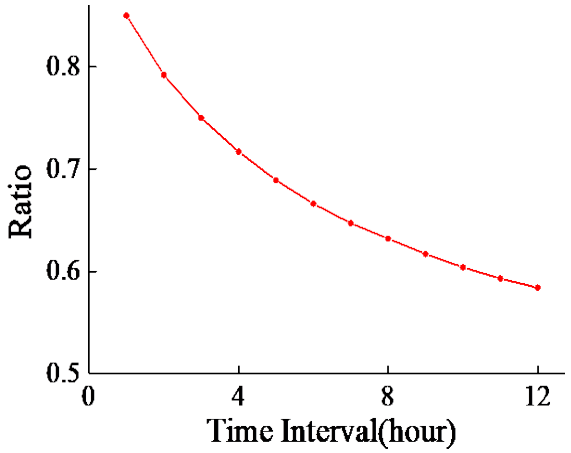




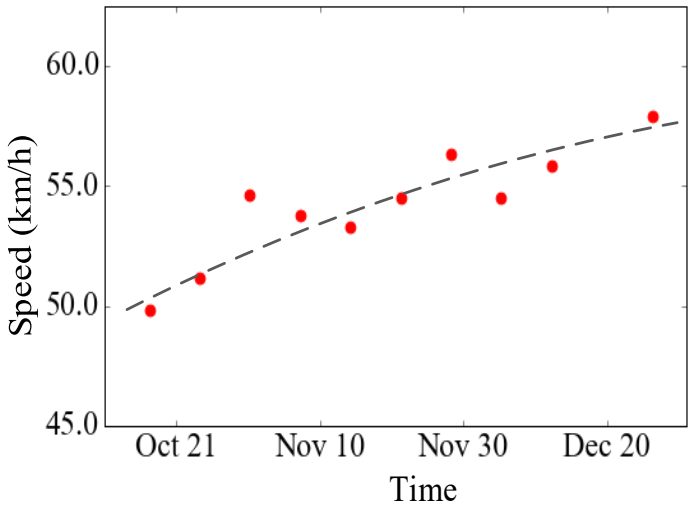
Temporal Characteristics

- Temporal
  - Closeness: correlation
  - Trend: long term

Closeness



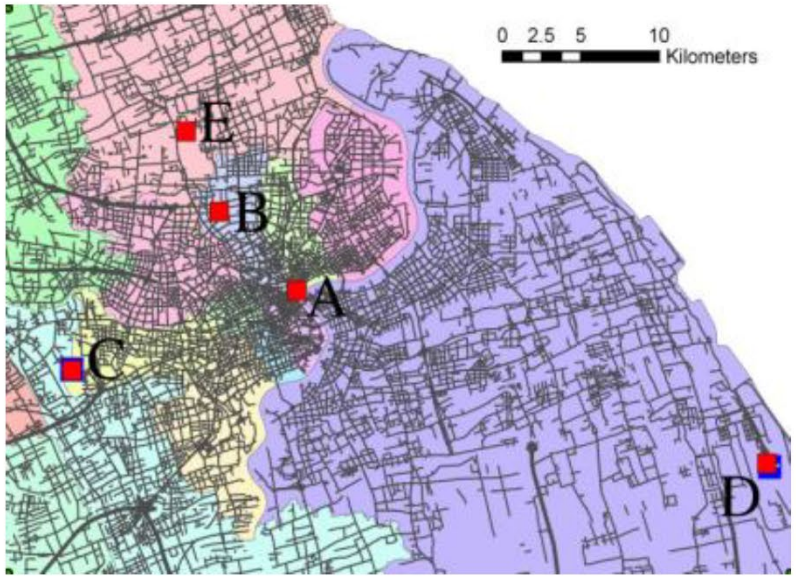
Trend

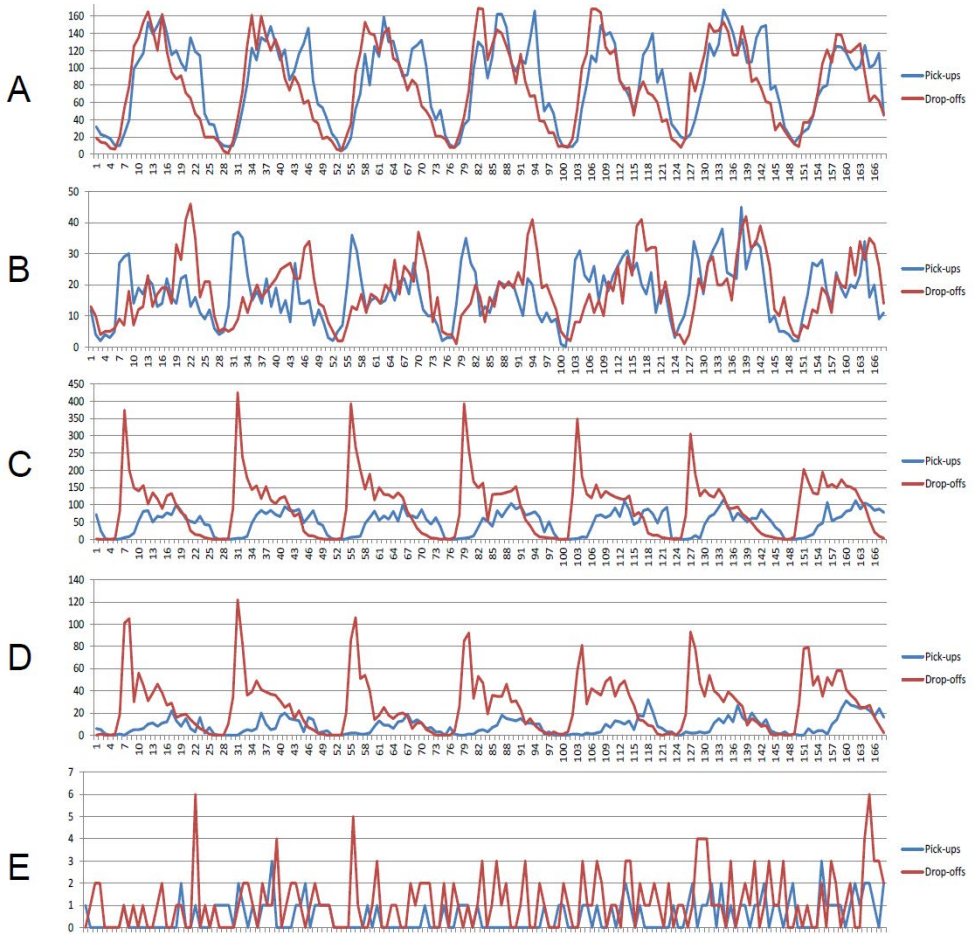


Traffic speed at 9-10am on consecutive weekends

Temporal Characteristics

Periodicity

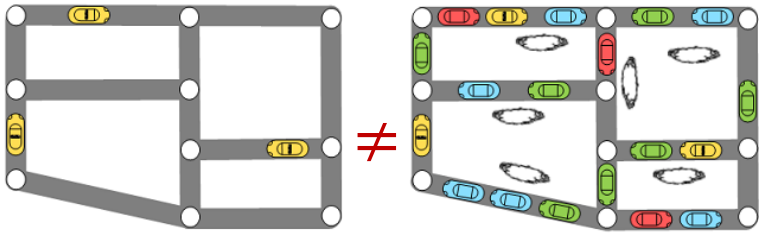




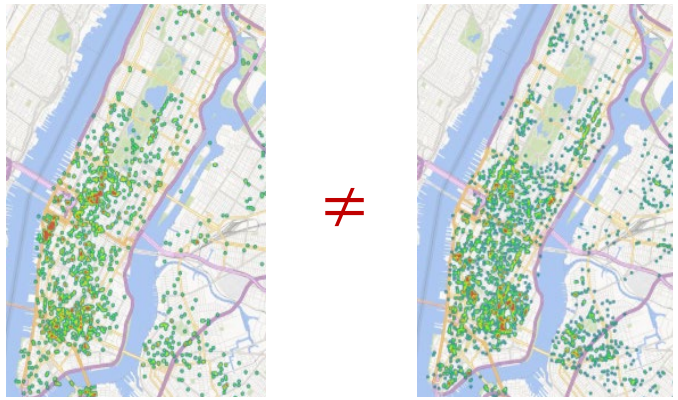
Spatio-Temporal Data Collection Characteristics

A sample of data → An entire dataset

- Biased distribution

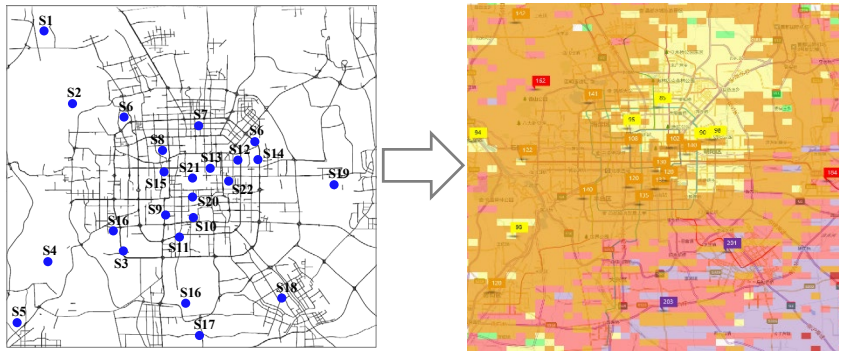


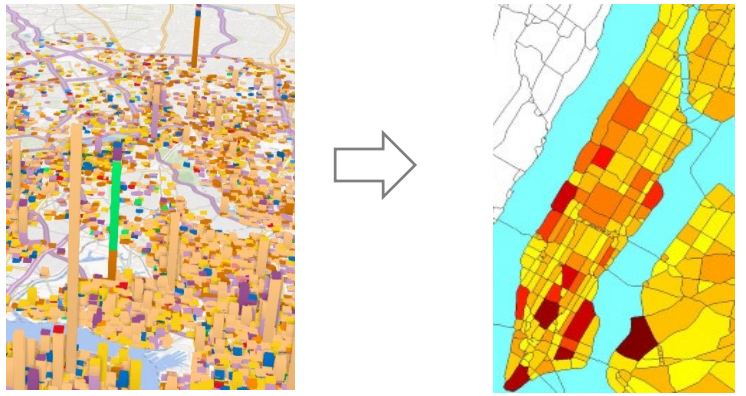
Taxi flow ≠ Entire traffic flow



Check in data ≠ Citywide human mobility

- Data missing and sparsity

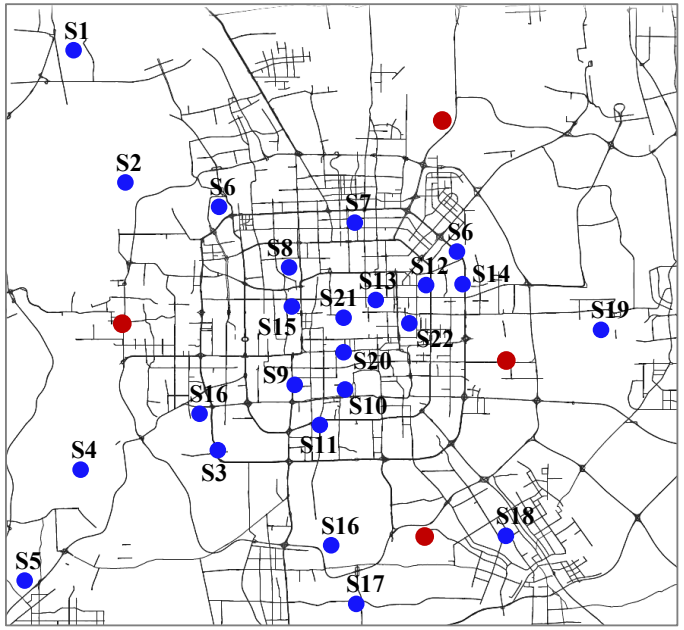


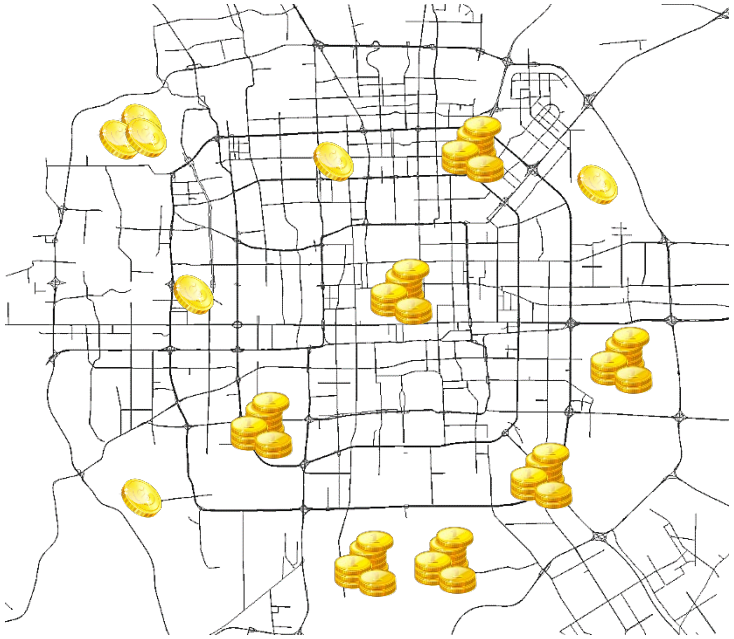


Spatio-Temporal Data Collection Characteristics

A limited resource (budget, labors, land…)

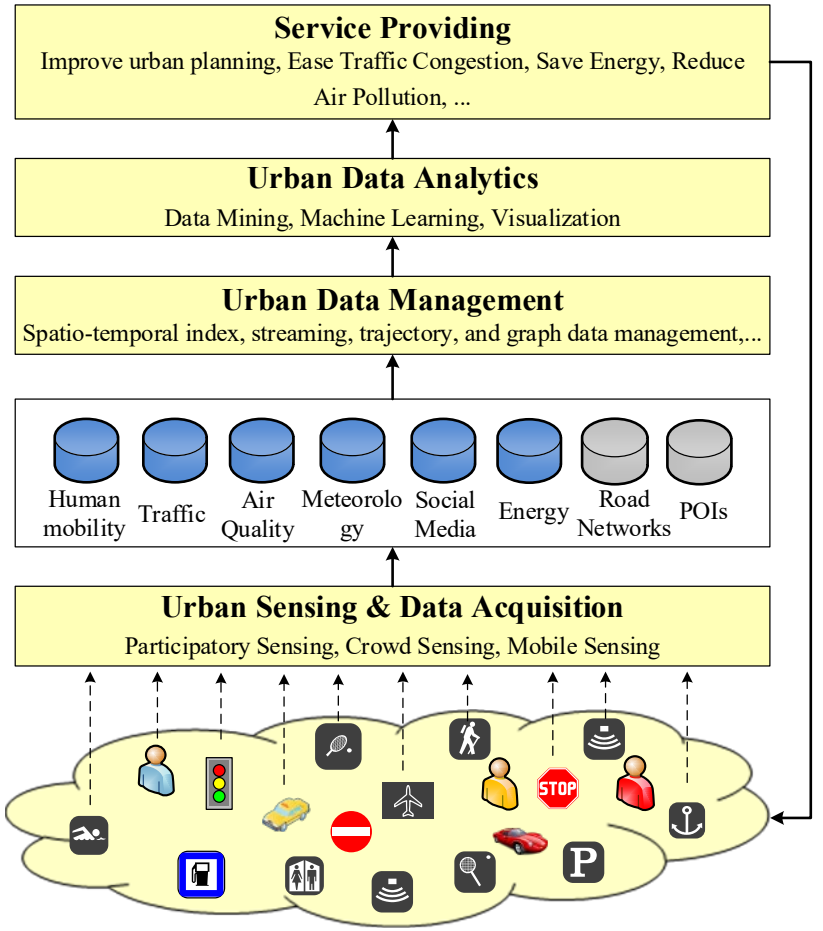
- **Static sensing**: Where to deploy sensor to maximize the gain?
- **Crowdsensing**: How to arrange the incentives dynamically?





Fundamental Techniques for Spatio-Temporal Data

- Spatio-temporal data acquisition
- Spatio-temporal data management
- Spatio-temporal data mining



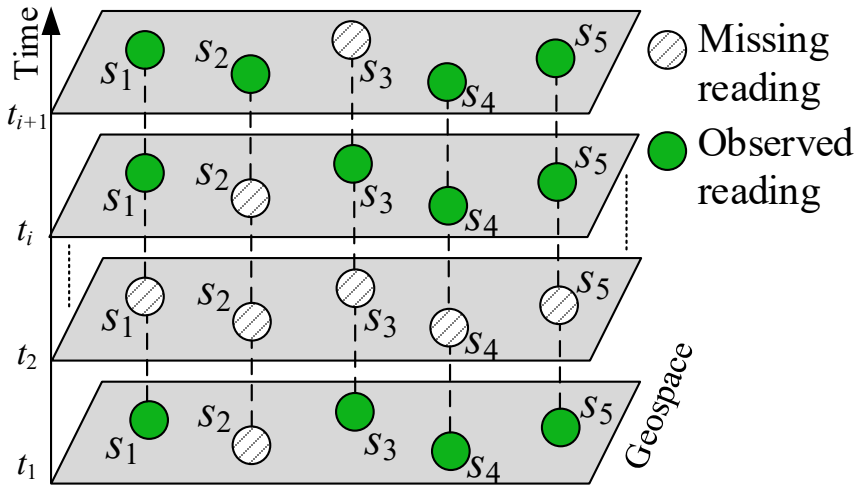
Zheng, Y., et al. Urban Computing: concepts, methodologies, and applications. *ACM transactions on Intelligent Systems and Technology*.

Spatio-Temporal Data Acquisition

Filling Missing Values in Spatio-Temporal Data

Filling Missing Values in Spatio-Temporal Data

- **Data missing** is a very common phenomenon in IOT data
- **Goal:** Inferring the values of those missing entries using collective information:



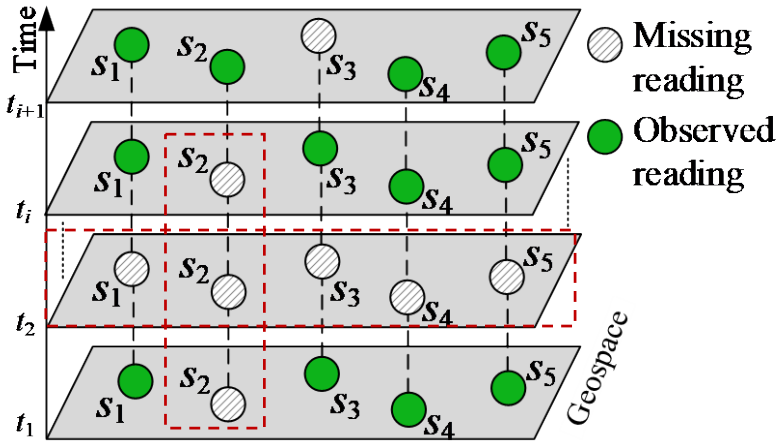
A) Missing situation

	PM2.5	NO_2	Humidity	Wind Speed
Missing rate	13.3%	16.0%	21.5%	30.3%

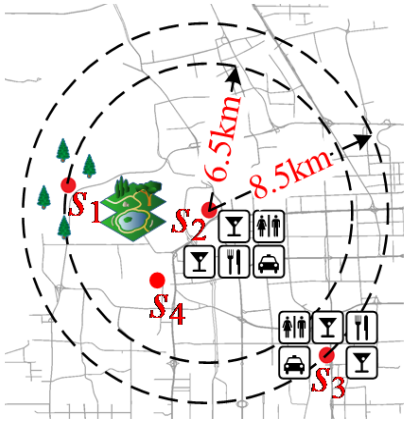
Xiuwen Yi, **Yu Zheng**, et al. ST-MVL: Filling Missing Values in Geo-sensory Time Series Data. IJCAI 2016

Filling Missing Values in Spatio-Temporal Data

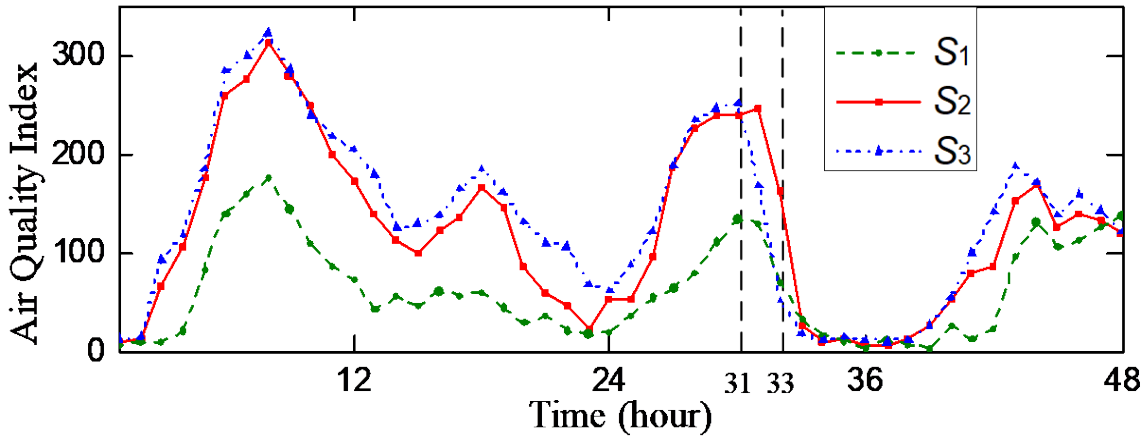
- **Difficulties**
  - Random missing and block missing
    - Not handled by fixed learning models
  - Readings changing over time and locating non-linearly
    - Not handled by simple interpolations



A) Missing situation



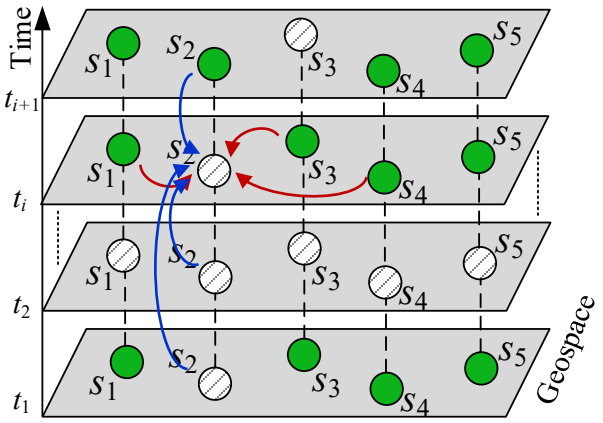
A) Geo-location of sensors

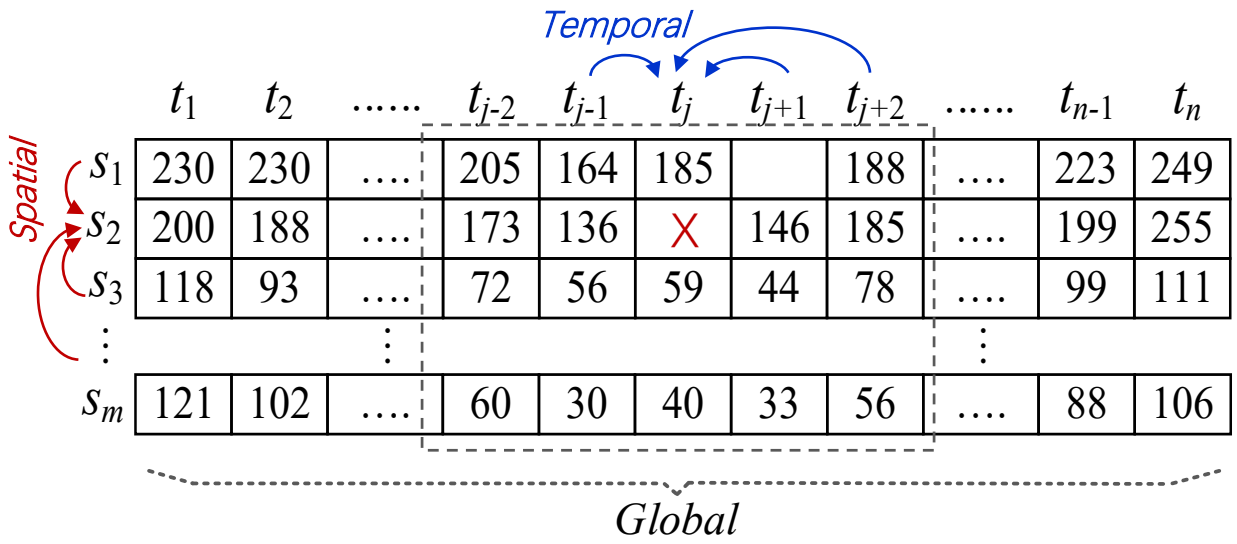


B) Air quality index over time

Fill Missing Values in Spatio-Temporal Datasets

- Achieve this goal from different perspectives
  - **Spatial and Temporal** perspectives
    - Spatial neighbors
    - Temporally adjacent time intervals
  - **Global and local** perspectives
    - Local: Recent context
    - Global: Long-term patterns

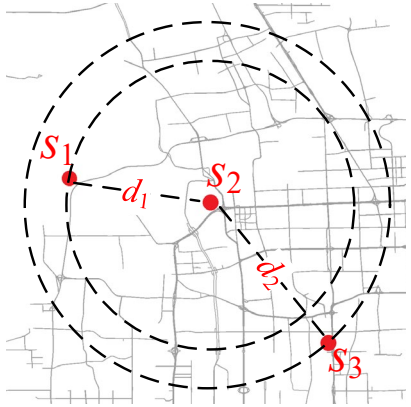




Global: long-term knowledge

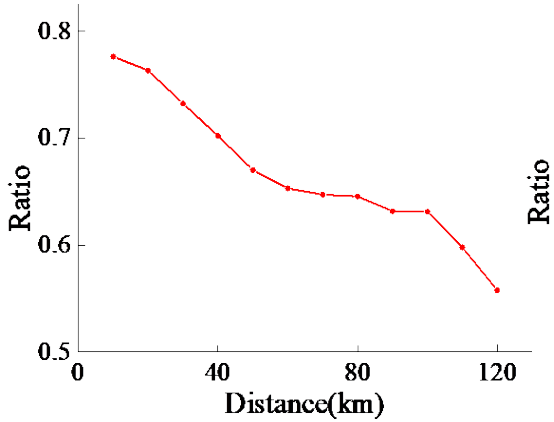
Spatial

Inverse Distance Weighting (IDW)

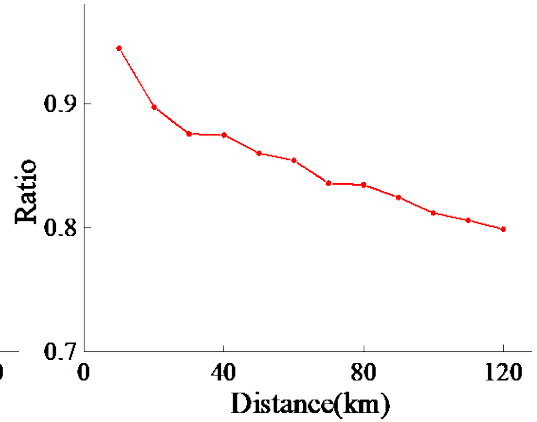


$$\hat{v}_{gs} = \frac{\sum_{i=1}^{m} v_i * d_i^{-\alpha}}{\sum_{i=1}^{m} d_i^{-\alpha}}$$

Beijing from May 2014 to May 2015



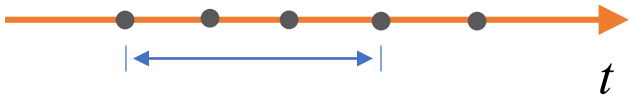
A) Air quality data



B) Humidity

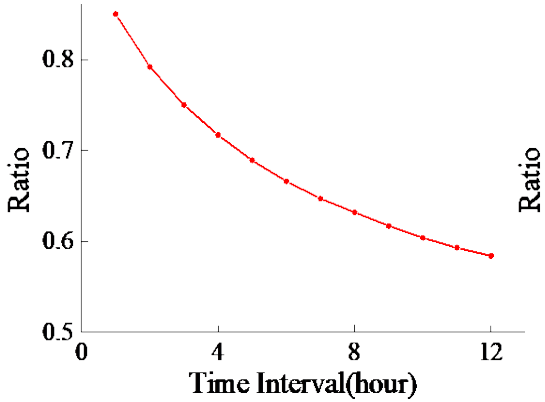
Temporal

Simple Exponential Smoothing (SES)

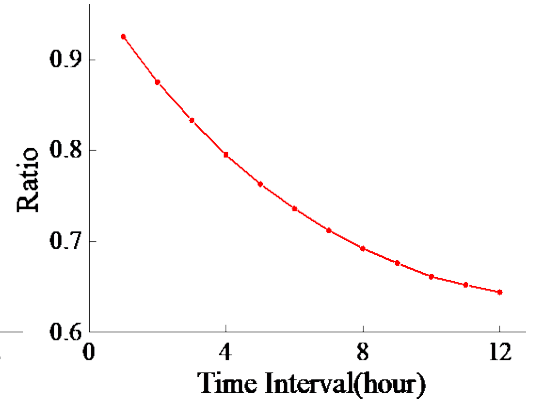


Time interval

$$\hat{v}_{gt} = \beta v_j + \beta(1-\beta)v_{j-1} + \cdots + \beta(1-\beta)^{t_j-1}v_1$$



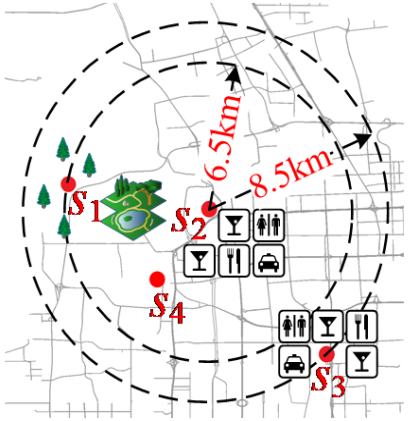
A) Air quality data



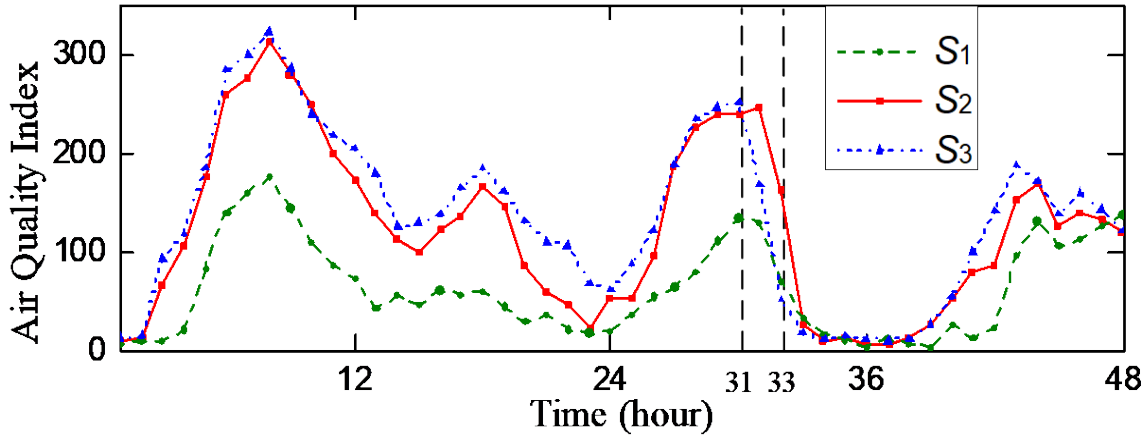
B) Humidity

Local: Recent Context

- Some situations break long-term patterns



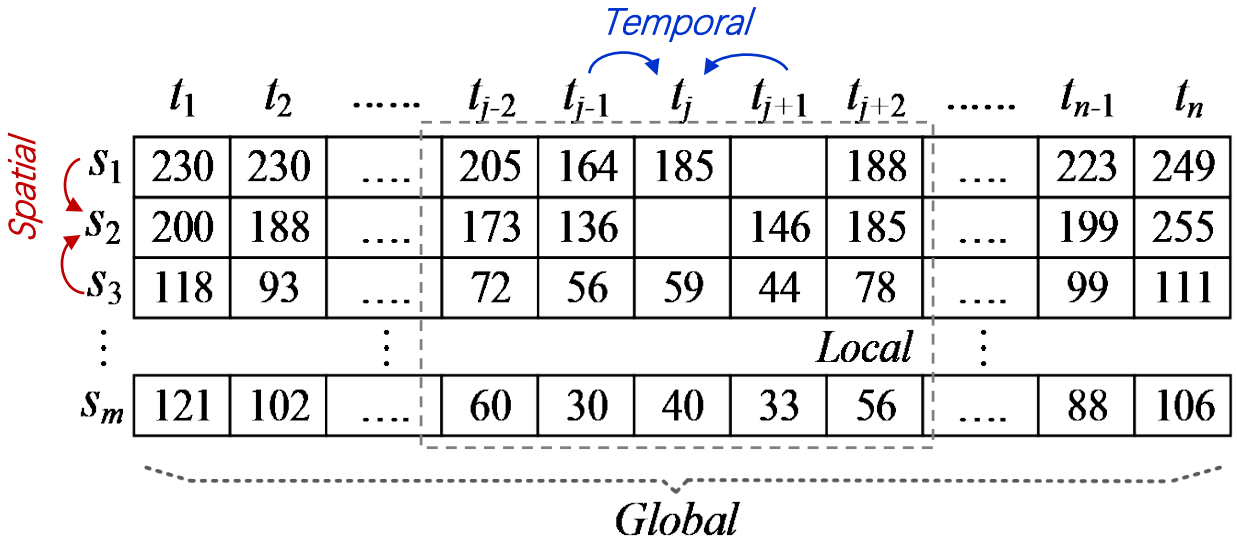
A) Geo-location of sensors



B) Air quality index over time

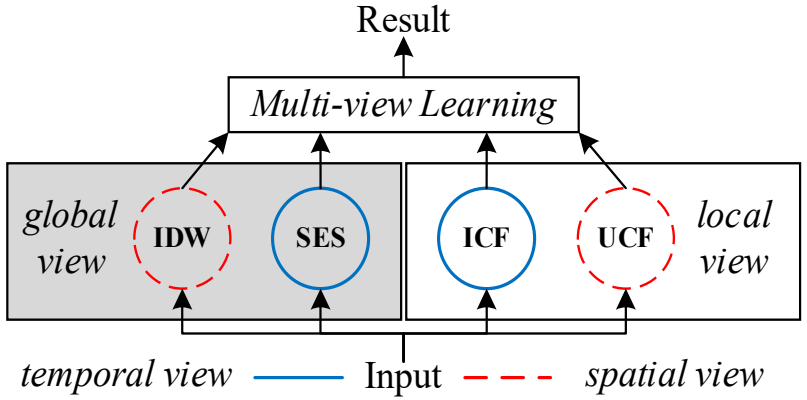
Collaborative Filtering

- Sensors → users
- Time intervals → items

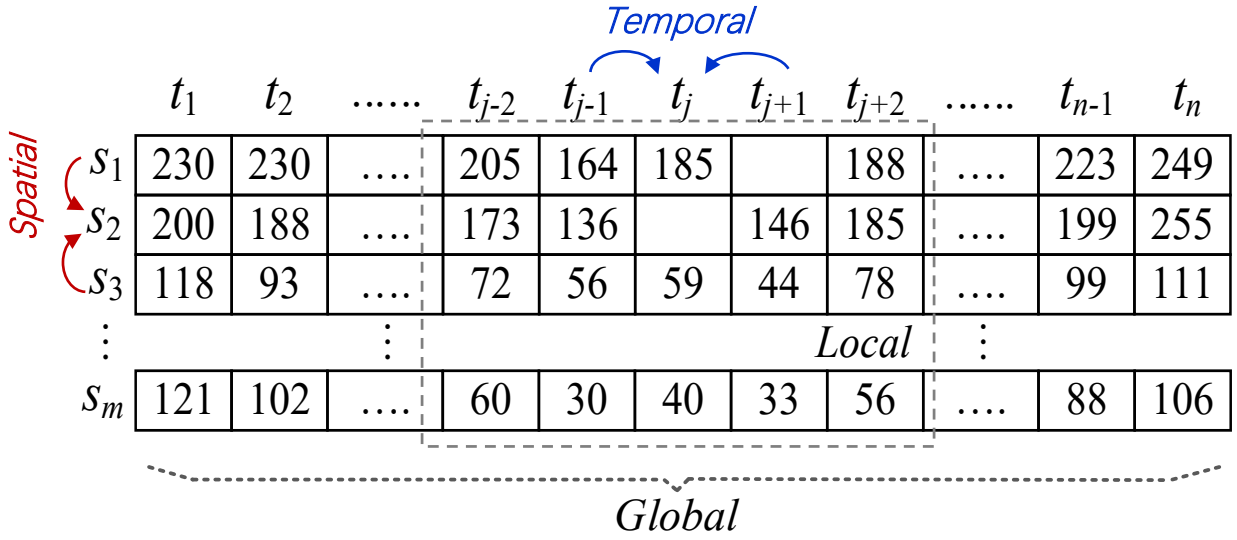


Fill Missing Values in Spatio-Temporal Datasets

- A multi-view-based method
  - IDW: Inverse Distance Weighting
  - SES: Simple Exponential Smoothing
  - UCF: User-based Collaborative filtering
  - ICF: Item-based Collaborative filtering



$$\hat{v}_{mvl} = w_1 * \hat{v}_{gs} + w_2 * \hat{v}_{gt} + w_3 * \hat{v}_{ls} + w_4 * \hat{v}_{lt} + b$$



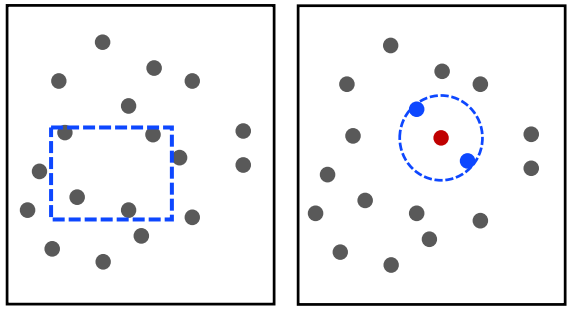
Spatio-Temporal Data Management

Managing Spatio-Temporal Big Data

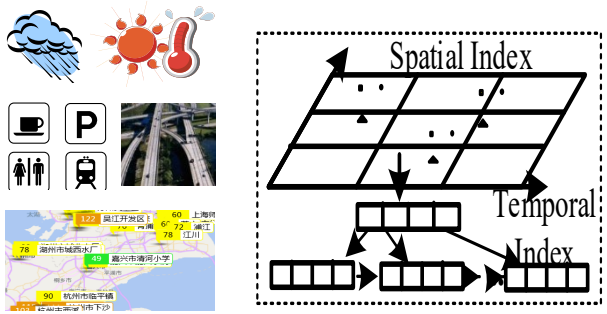
- **Difficulties**
  - Large-scale and highly dynamic
  - Cloud computing platforms do not support ST Data well
    - Unique ST data structures: trajectories (the most complex ST Data)
    - Unique queries: ST-Range queries and KNN queries rather than key words
    - Data across different domains: Hybrid indexing for managing multi-modality data

Trajectories

Range Queries KNN Queries

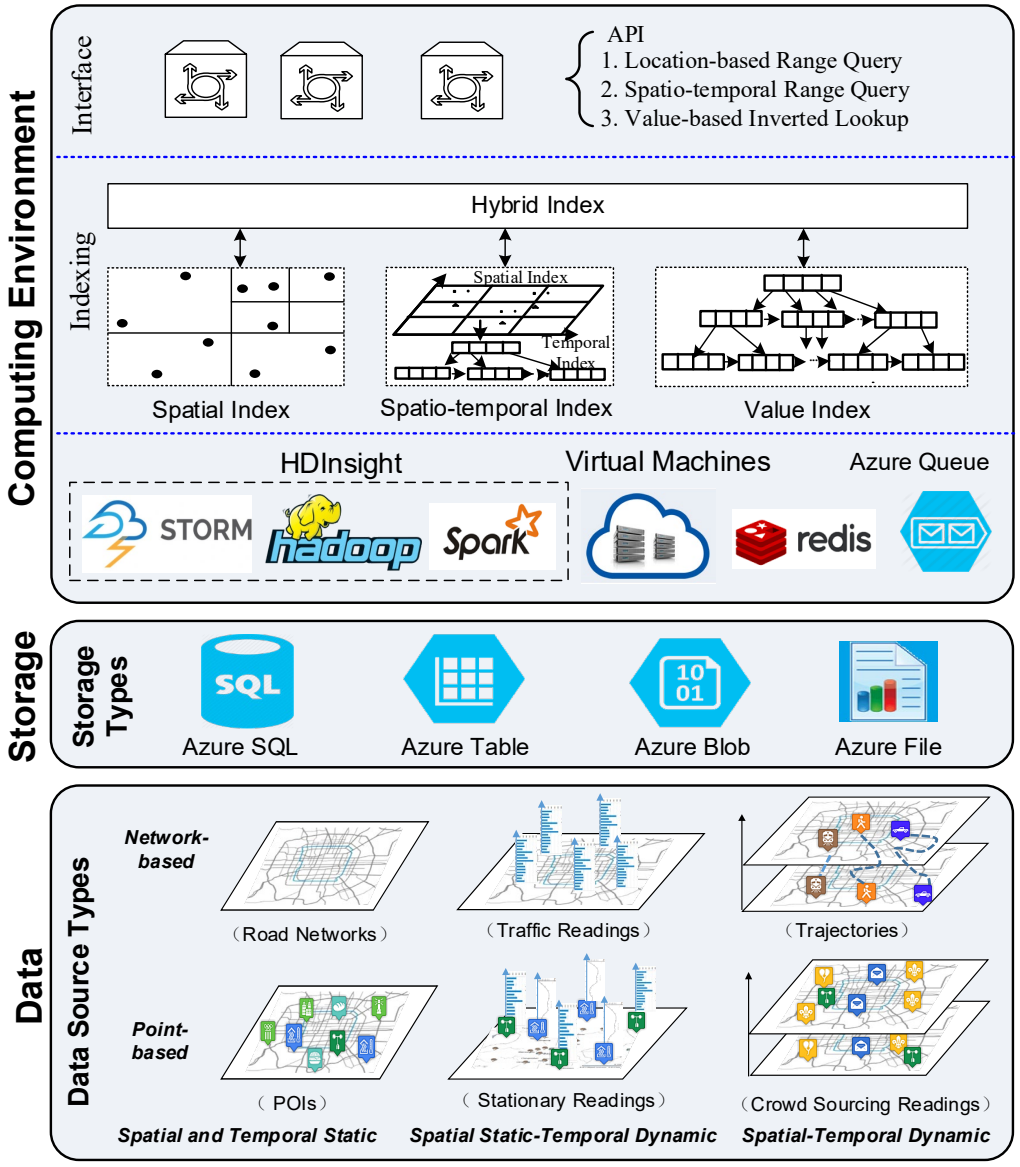


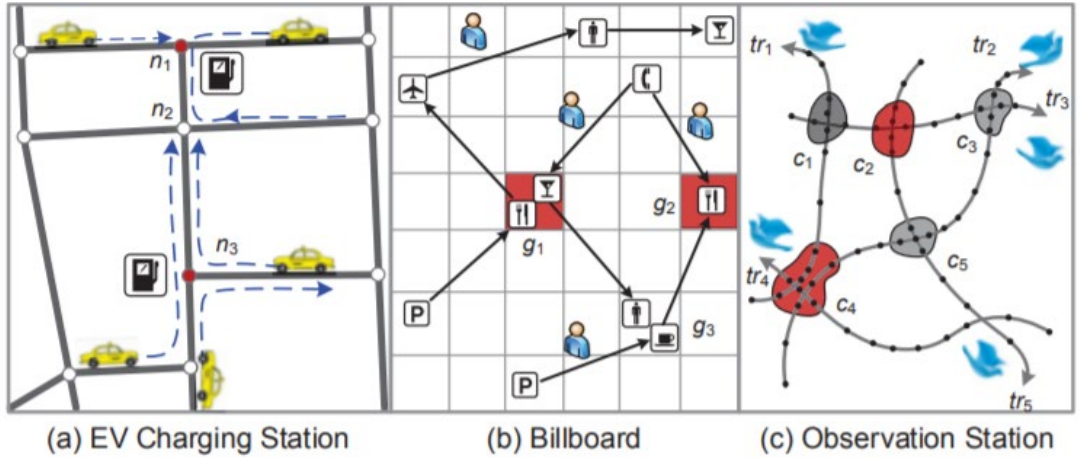
Hybrid indexing



Jie Bao, et al. Managing Massive Trajectories on the Cloud, ACM SIGSPATIAL 2016

Cloud Computing for ST Data





[Mining the Most Influential k-Location Set from Massive Trajectories](), ACM SIGSPATIAL 2016

Jie Bao, Ruiyuan Li, Xiuwen Yi, Yu Zheng.
[Managing Massive Trajectories on the Cloud]().

Finding Top-k Most Influential Location Set

"SmartAdP: Visual Analytics of Large-scale Taxi Trajectories for Selecting Billboard Locations", VAST 2016

Spatio-Temporal Data Mining

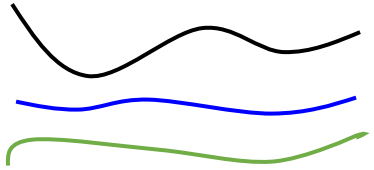
Application

Spatio-Temporal Data Mining

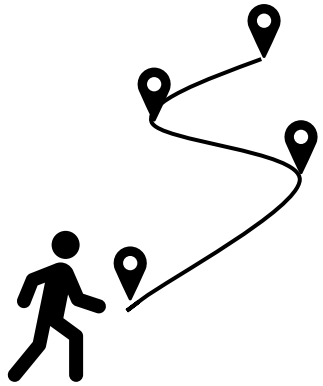
Smart phones and Intelligence Vehicles

Social Media

Call Detail Records



Spatio-Temporal Data

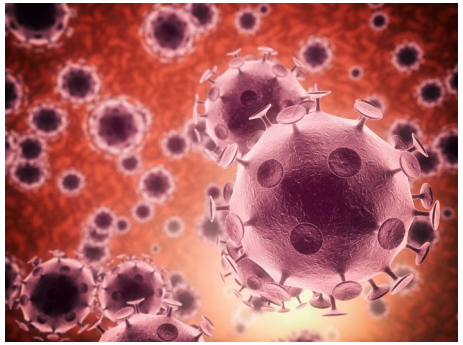


Mobility Patterns and Daily Habits

Various Downstream Tasks

Traffic Monitoring

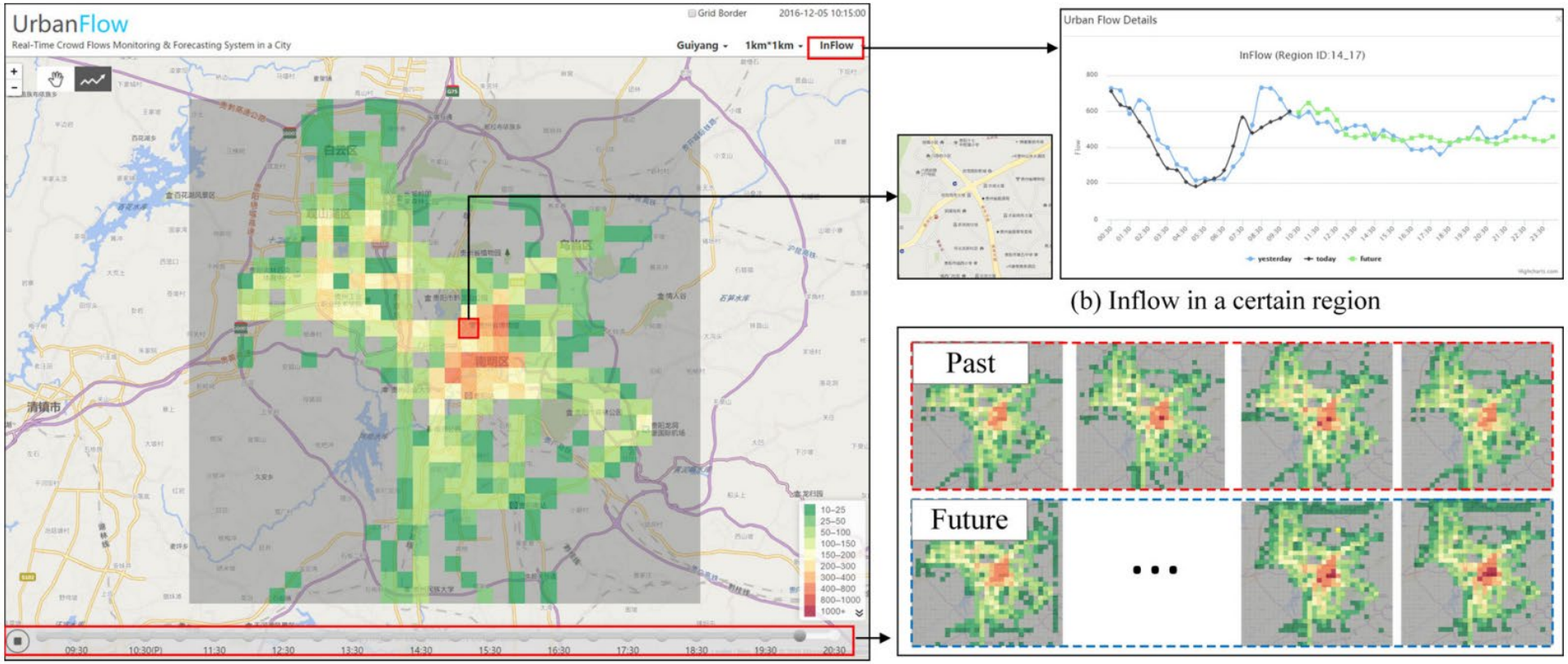
Route Planning



Pandemic Spread Prediction

Application

- Crowd flow prediction



(a) Web user interface of UrbanFlow

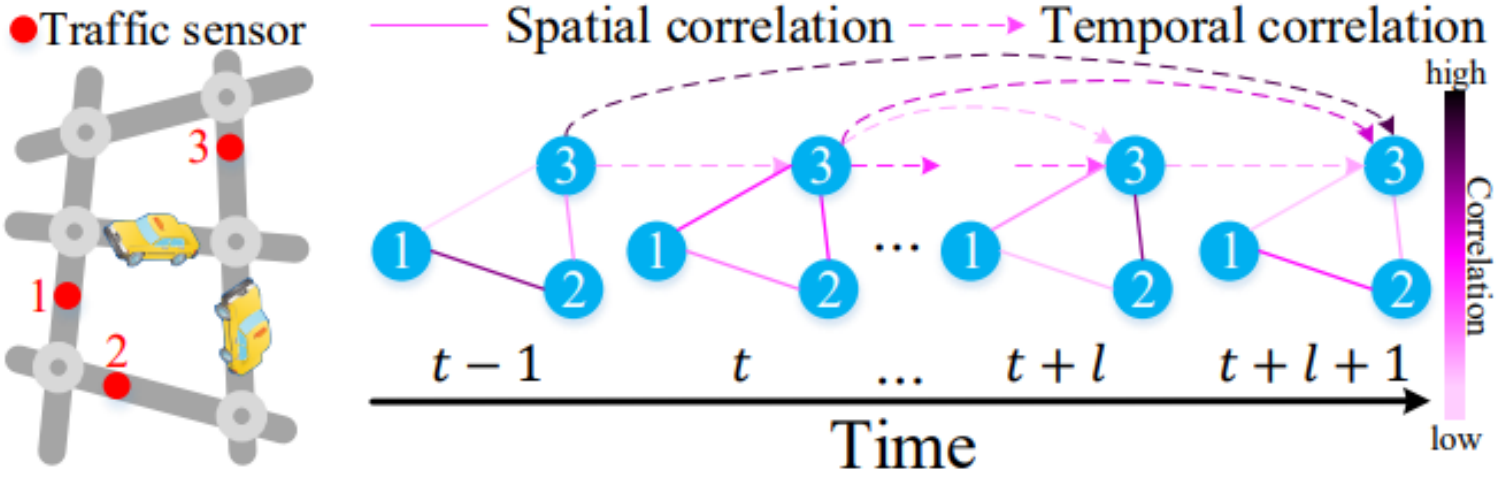
(b) Inflow in a certain region

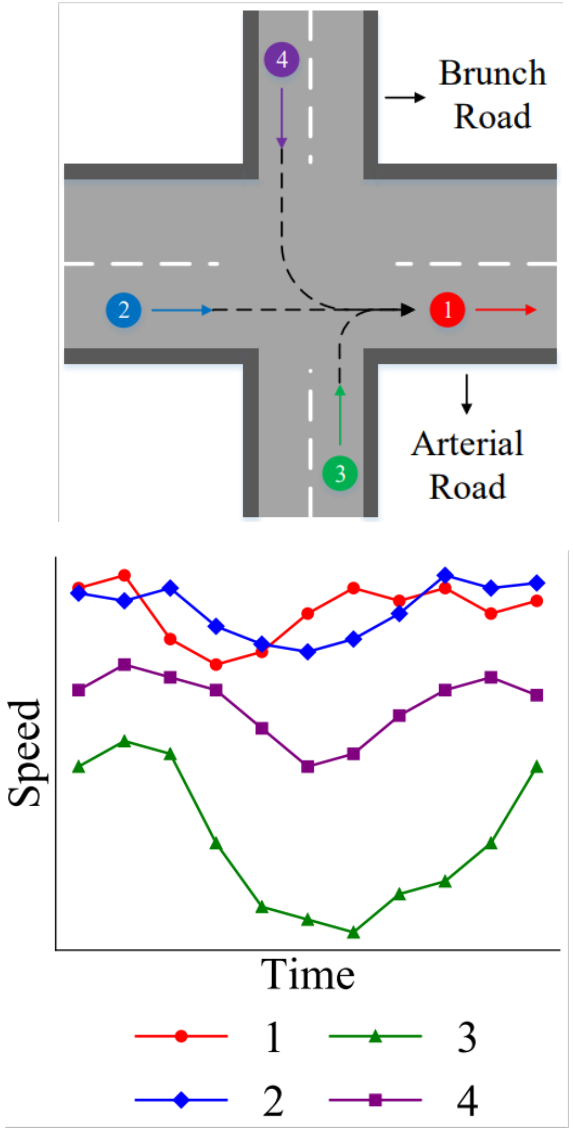
(c) Heatmaps in past & future

Zhang J, Zheng Y, Qi D, et al. Predicting citywide crowd flows using deep spatio-temporal residual networks[J]. Artificial Intelligence, 2018, 259: 147-166.

Application

- Urban traffic prediction



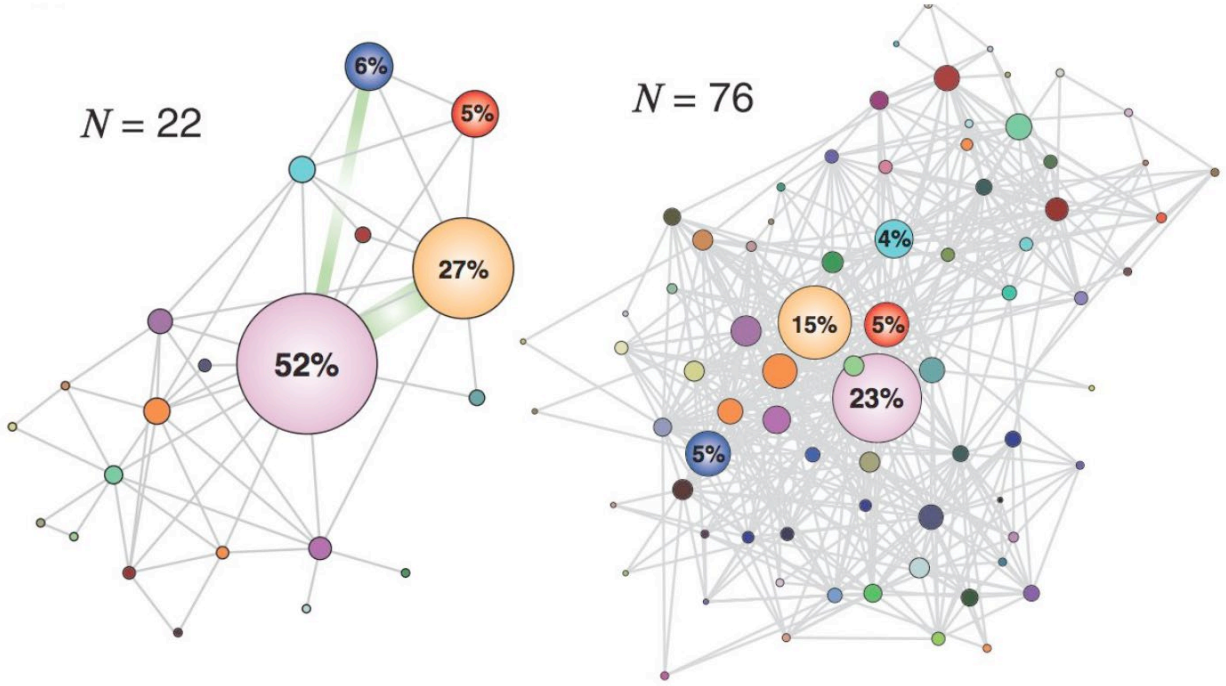


Zheng C, Fan X, Wang C, et al. Gman: A graph multi-attention network for traffic prediction[C] Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(01): 1234-1241.

Application

- Study human mobility

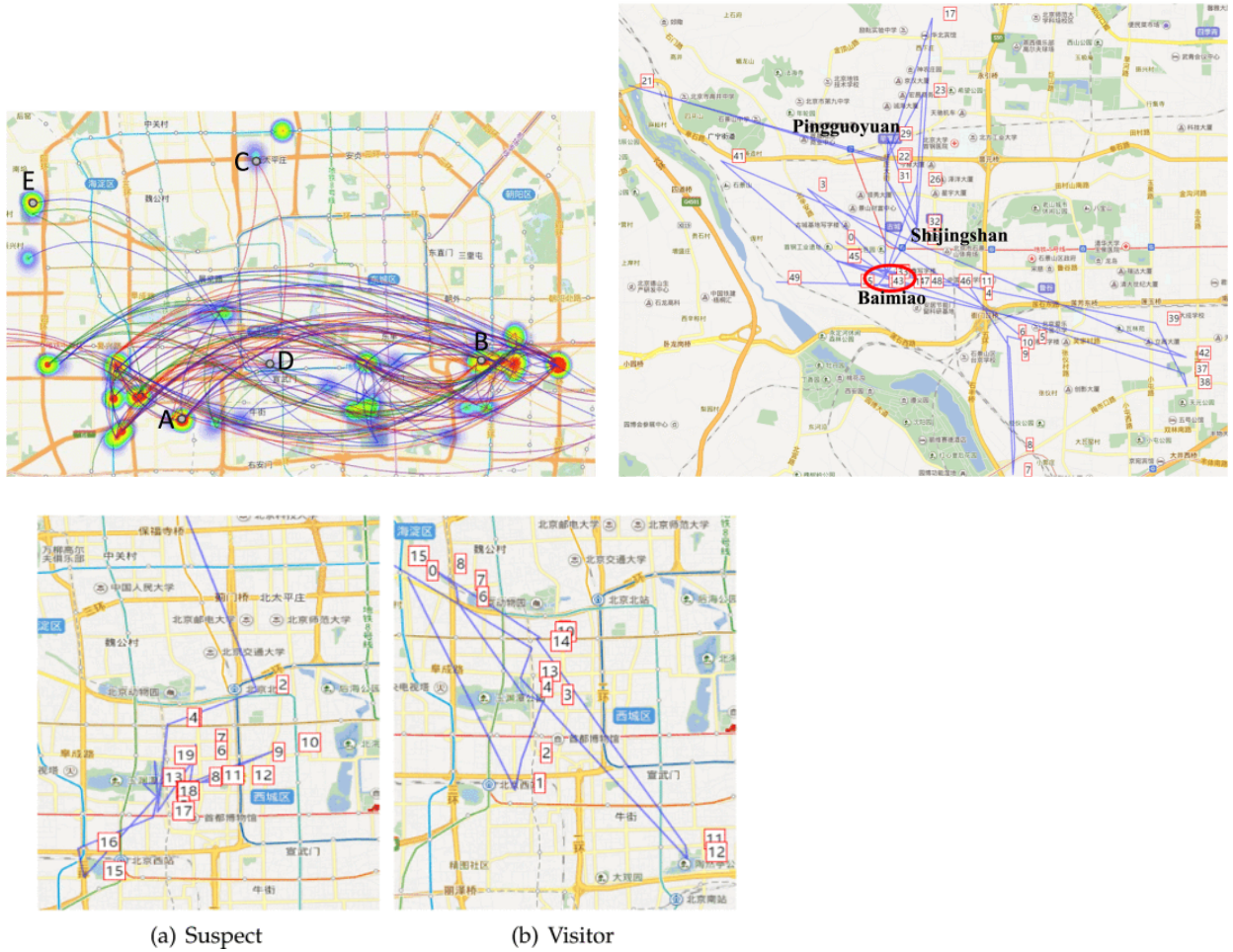
IMN: a network describing the typical movements of an individual



Song et al., Limits of predictability in human mobility, Science, 2010.

Application

- Public safety and security
  - Detecting mobility anomalies
  - Predicting the flow of the crowds

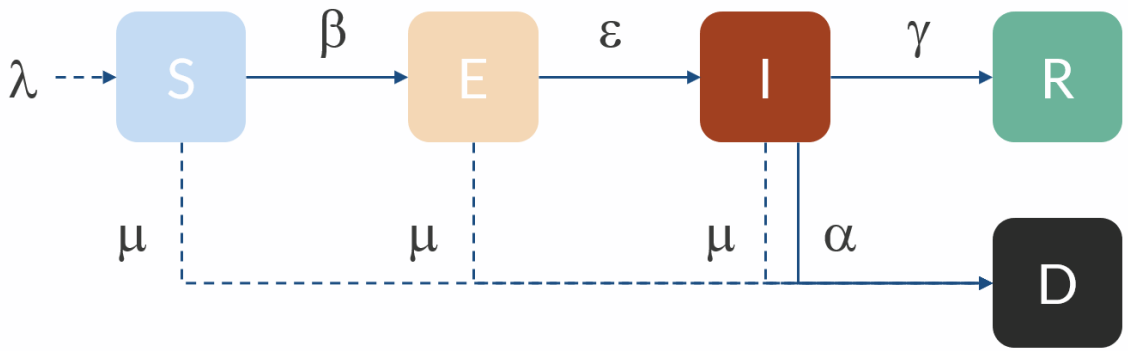


(a) Suspect (b) Visitor

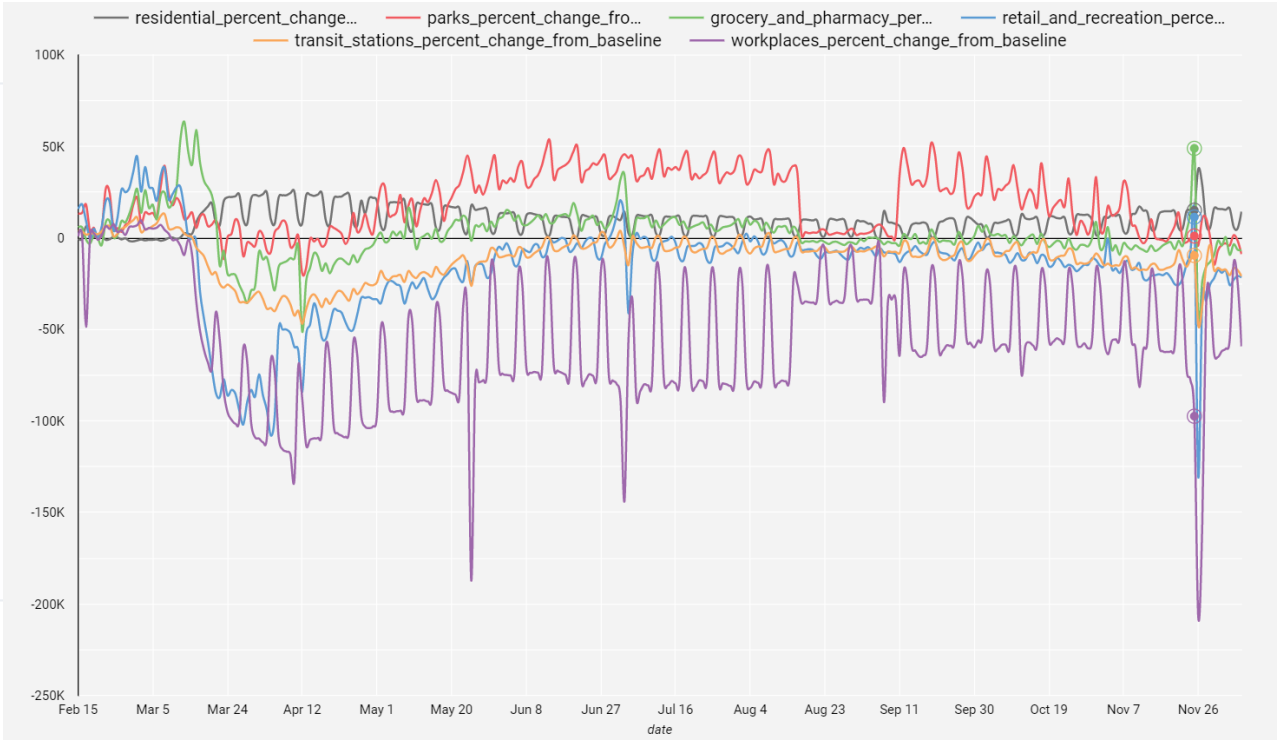
Du, Bowen, et al. "Detecting pickpocket suspects from large-scale public transit records." *IEEE Transactions on Knowledge and Data Engineering* 31.3 (2018): 465-478.

Application

- Epidemiologic prediction



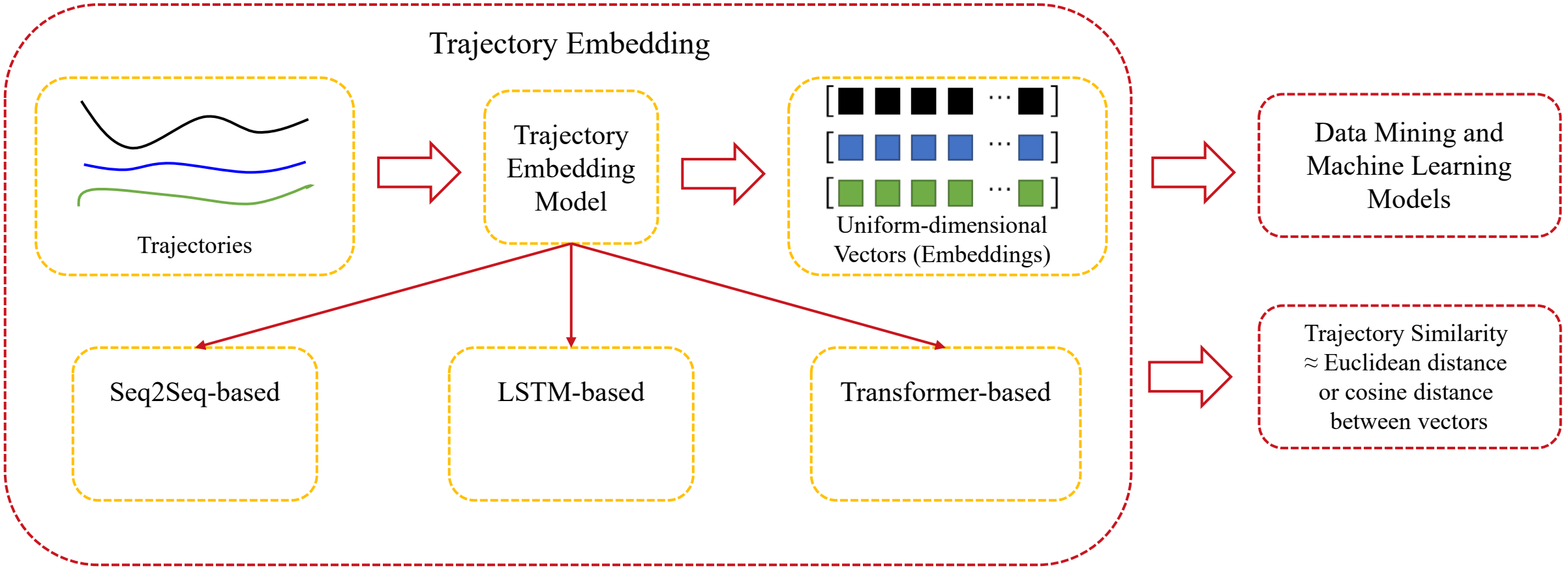
SEIR: susceptible, exposed, infectious, recovered



Google Mobility

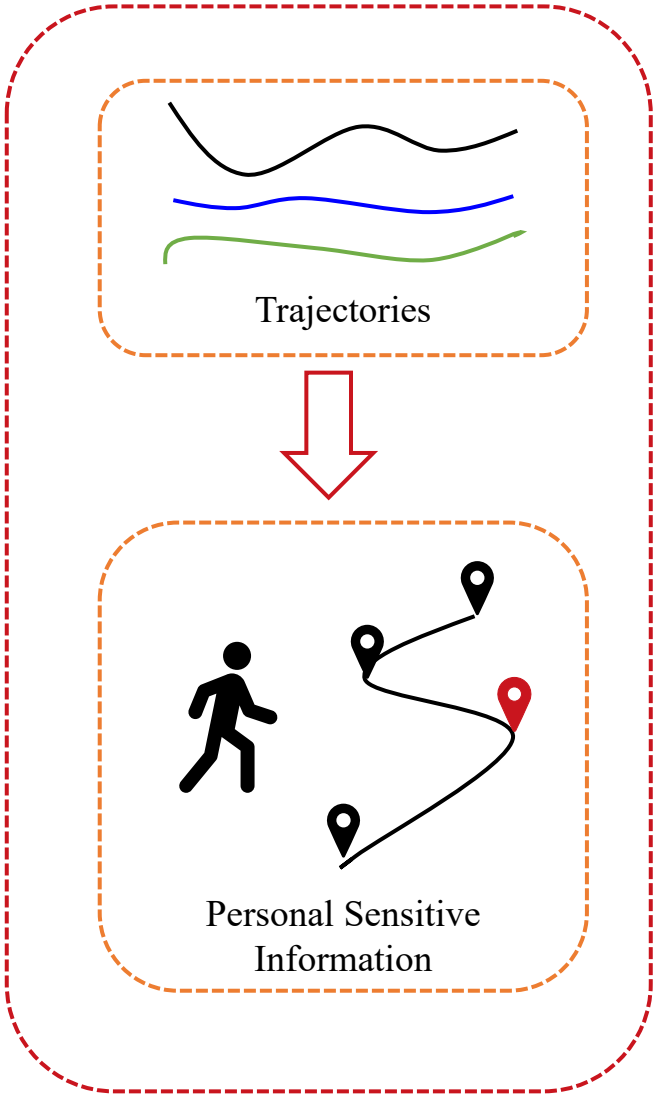
Application

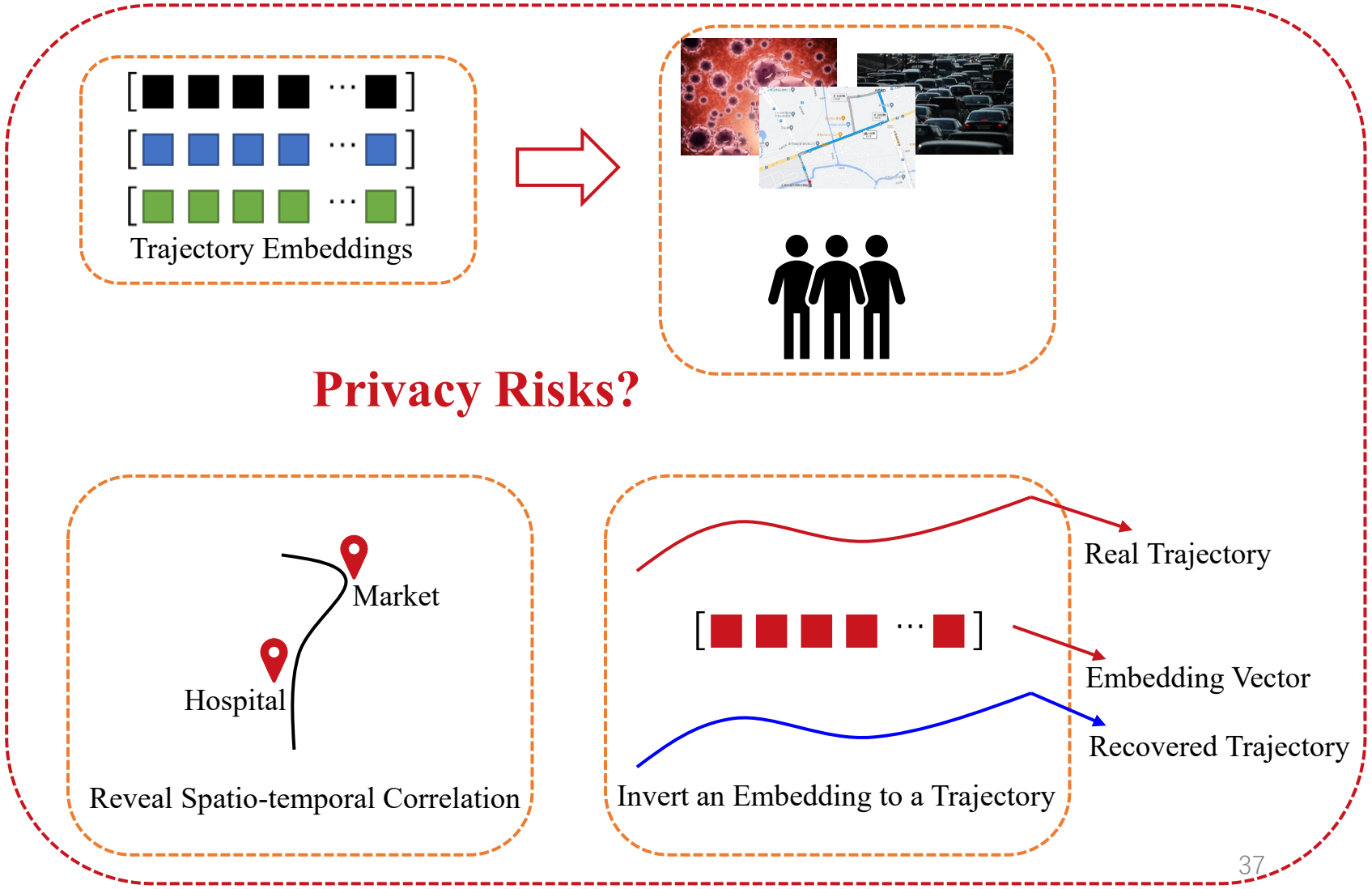
- Trajectory representation learning



- Xiucheng Li, Kaiqi Zhao, Gao Cong, Christian S Jensen, and Wei Wei. 2018. Deep representation learning for trajectory similarity computation. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*. IEEE, 617–628.
- Di Yao, Gao Cong, Chao Zhang, and Jingping Bi. 2019. Computing Trajectory Similarity in Linear Time: A Generic Seed-Guided Neural Metric Learning Approach. In 2019 IEEE 35th International Conference on Data Engineering (ICDE). 1358–1369. https://doi.org/10.1109/ICDE.2019.00123
- Yile Chen, Xiucheng Li, Gao Cong, Zhifeng Bao, Cheng Long, Yiding Liu, Arun Kumar Chandran, and Richard Ellison. 2021. Robust Road Network Representation Learning: When Traffic Patterns Meet Traveling Semantics. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 211–220.
- Jiaxin Ding, Bowen Zhang, Xinbing Wang, Chenghu Zhou. "TSNE: trajectory similarity network embedding." Proceedings of the 30th International Conference on Advances in Geographic Information Systems. 2022.

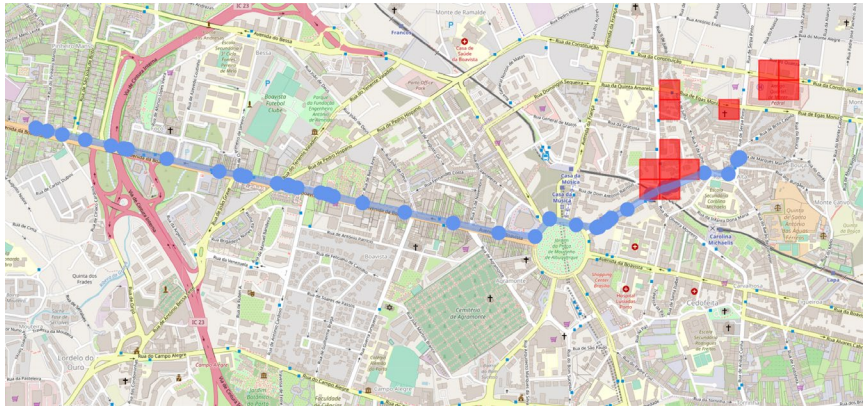
Application

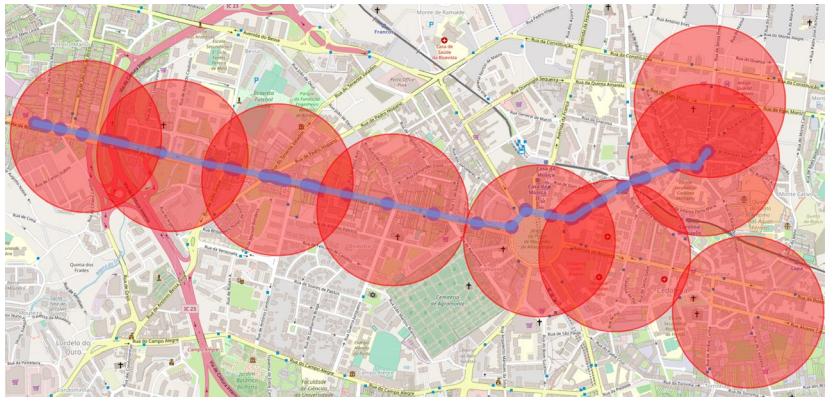




Application

- Trajectory representation attacks and privacy protection

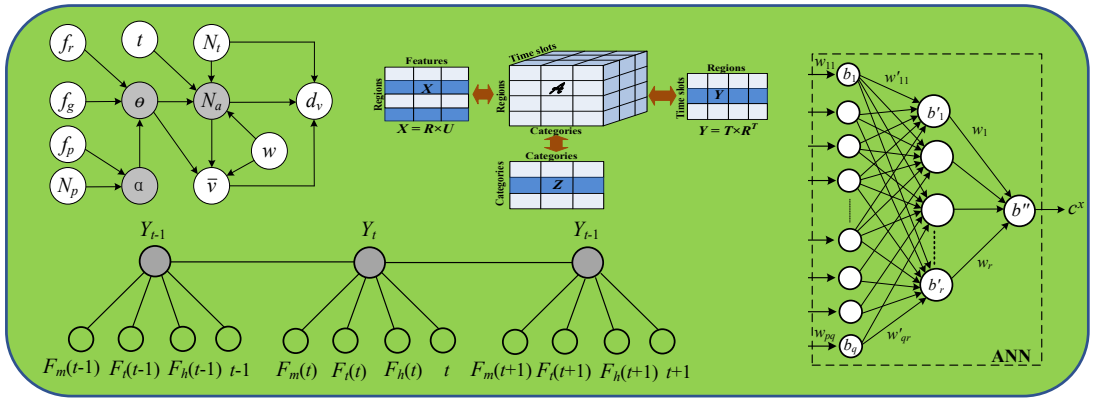




Ding J, Xi S, Wu K, et al. Analyzing sensitive information leakage in trajectory embedding models[C]//Proceedings of the 30th International Conference on Advances in Geographic Information Systems. 2022: 1-10.

- Similarity in high dim.
- Ranking the nodes
- Feature extracting
- Social networks
- Streaming queries
- Distributed computing
- Privacy

Models and Algorithms



- Statistics
- Locality Sensitive Hashing
- PageRank
- Graph representation
- SVD
- MapReduce

Problems

Data

- High Dimension batch data
- Graph
- Streaming
- Spatio-temporal data