京东城市
JD iCity

数据驱动方法在城市中的应用

新问题、新思路与新方法

詹仙园 博士

京东智能城市研究院资深研究员

THU 2018/12/13

About Me

2018 - Present
JD iCity

2017 - 2018
Microsoft Research Asia

2013 - 2017
Ph.D.: Transportation Engineering at Purdue

2014 - 2016
M.S.: Computer Science at Purdue

2011 - 2012
M.S.: Transportation Engineering at Purdue

2007 - 2011
B.E.: Structure Engineering at THU

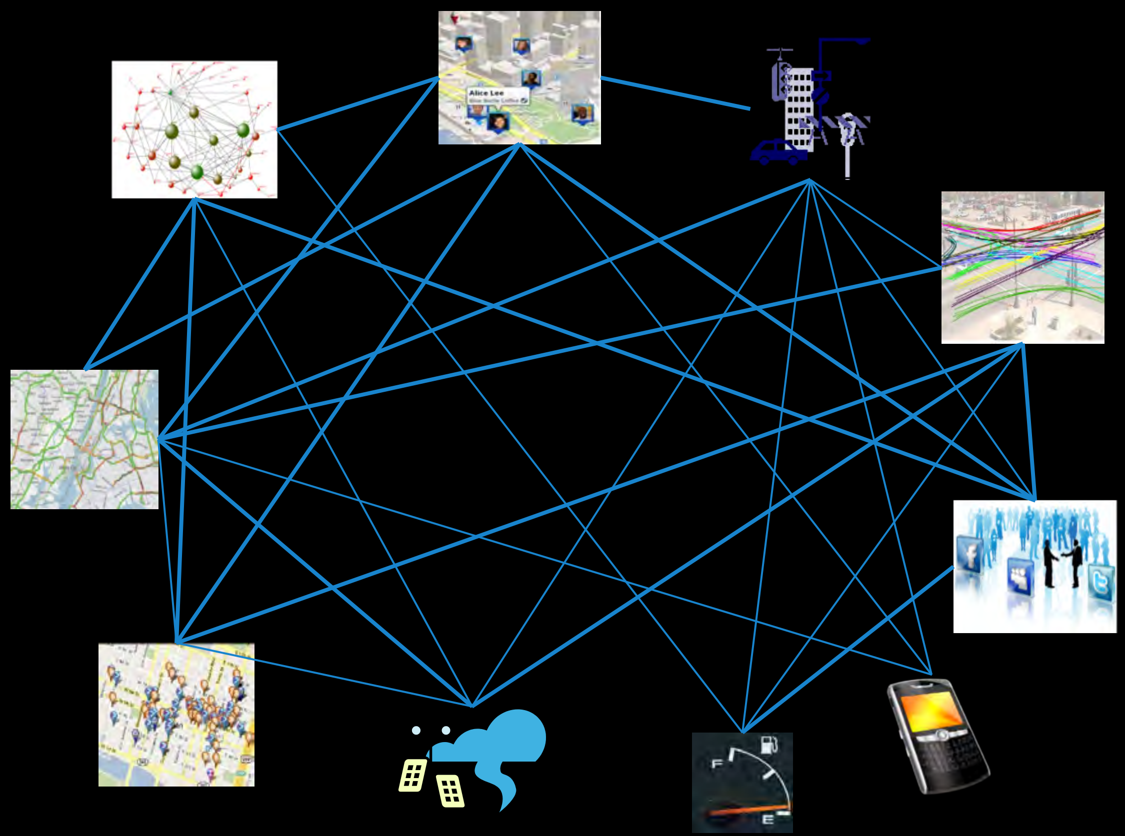
Agenda

- 新问题
  - 新数据 & 新问题
- 新思路
  - 数据驱动思维
- 新方法
  - 时空深度学习 & More

新问题

新数据 & 新问题

Massive Amount of Data in Urban Space



Massive self generated data from urban space

- Social media data
  - 3 million Foursquare check-ins per day (2011)
  - 500 million tweets per day on Twitter (2012)
- Mobile phone data
  - 30 million mobile phone records per day (Great Boston area, 2009)
- GPS data from taxis
  - 500,000 taxi trips from NYC (2013)
- Data from urban sensor networks
  - E.g. License-plate recognition camera data
    40,000 vehicle records per day for a single intersection (Langfang, 2015)

Tremendous Opportunities

- Game changer for urban systems modeling
- Allows us directly observe how system works
- Better solution for existing problems:
  - Traffic state monitoring
  - Inferring land use
- New possibilities for emerging problems:
  - Large-scale logistics/dispatching optimization
  - Water/air quality estimation/prediction
  - Detecting illegal parking
  - And many more

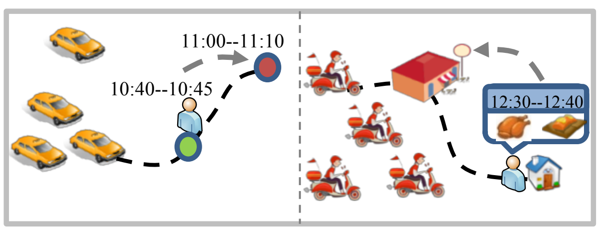
Why Data-driven Methods

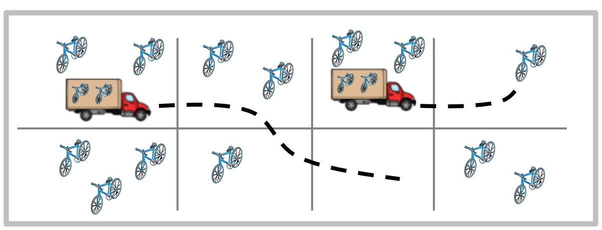
- Urban systems are highly complex
  - Millions of residents, large number of interacting sub-components
  - Simple analytical model simply will not work
- Requirement for efficiency and scalability
  - Solving analytical models are costly
  - Not suitable nor accurate for real-time applications
- Usability
  - Traditional analytic sometimes hardly useful for solving real world problems

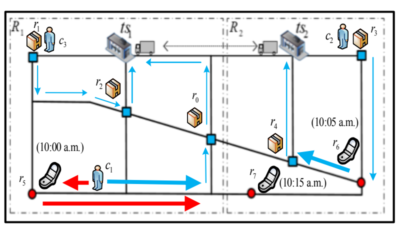
城市中的调度优化问题

- Urban services involve logistics/dispatching optimization :
  - Huge volumes of requests
  - Large amount of data
  - Real-time operation
  - Highly dynamic

Optimization matters!







数据驱动建模方法

- Highly dynamic
  - Requests changing over time and locations
  - States of resources: locations, load, tasks,…

→
- Use real-world data to mine patterns
- Use real-time data in models

- Large-scale
  - Thousands of candidates: locations, taxis, carriers…
  - Correlated (cannot be simply separated)
  - Most problems are NP-hard (scale is a disaster)

→
- New data-driven algorithms
- Problem/region partitioning
- Search space reduction

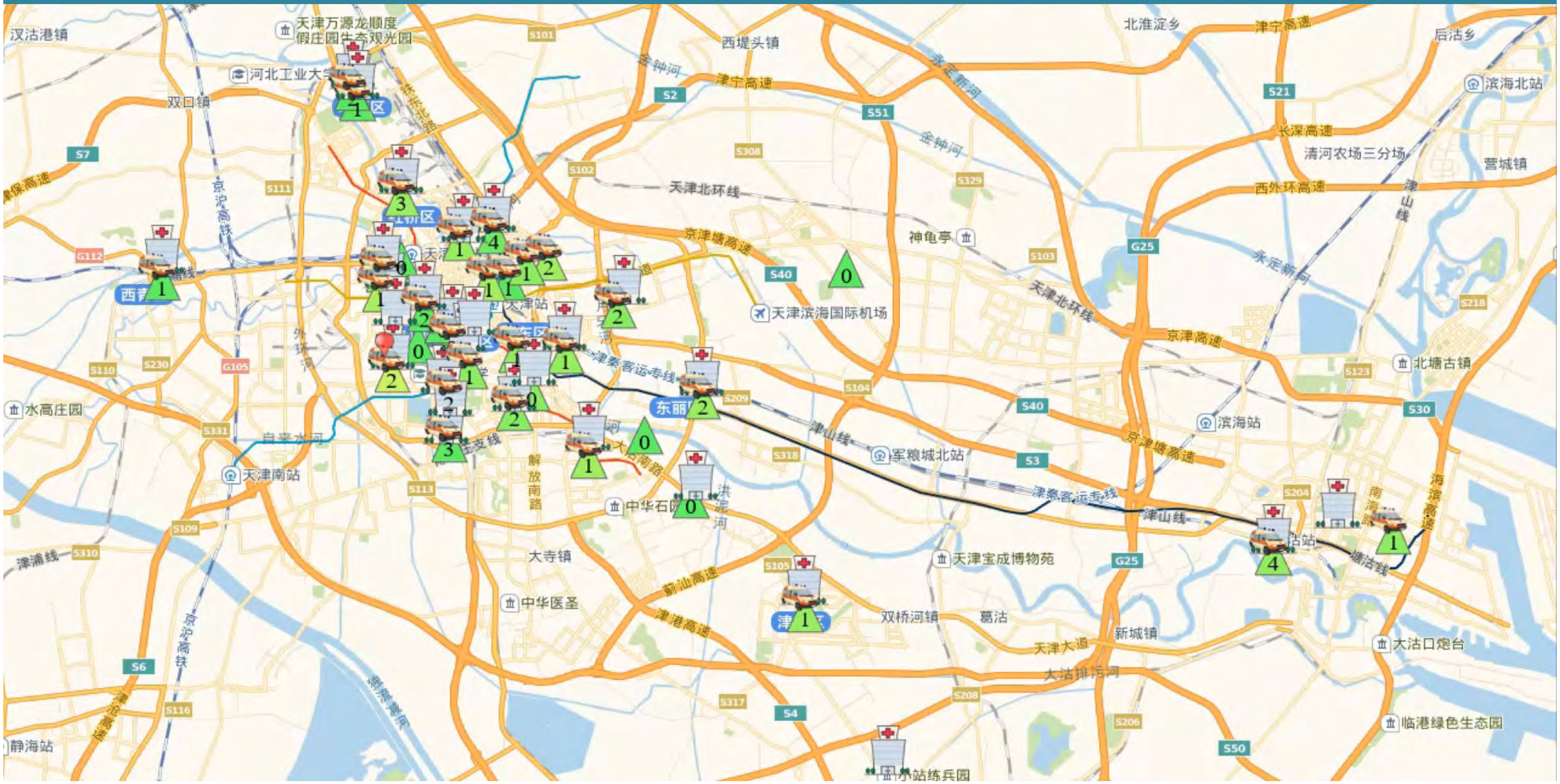
- Instantaneous answers

→
- Utilizing efficient data processing/management techniques, e.g. spatial-temporal indexing

- Multiple constraints: time, cost, capacity…

→
- Highly customized models
- Candidate pruning

AI改进救护车站点选址和调度优化

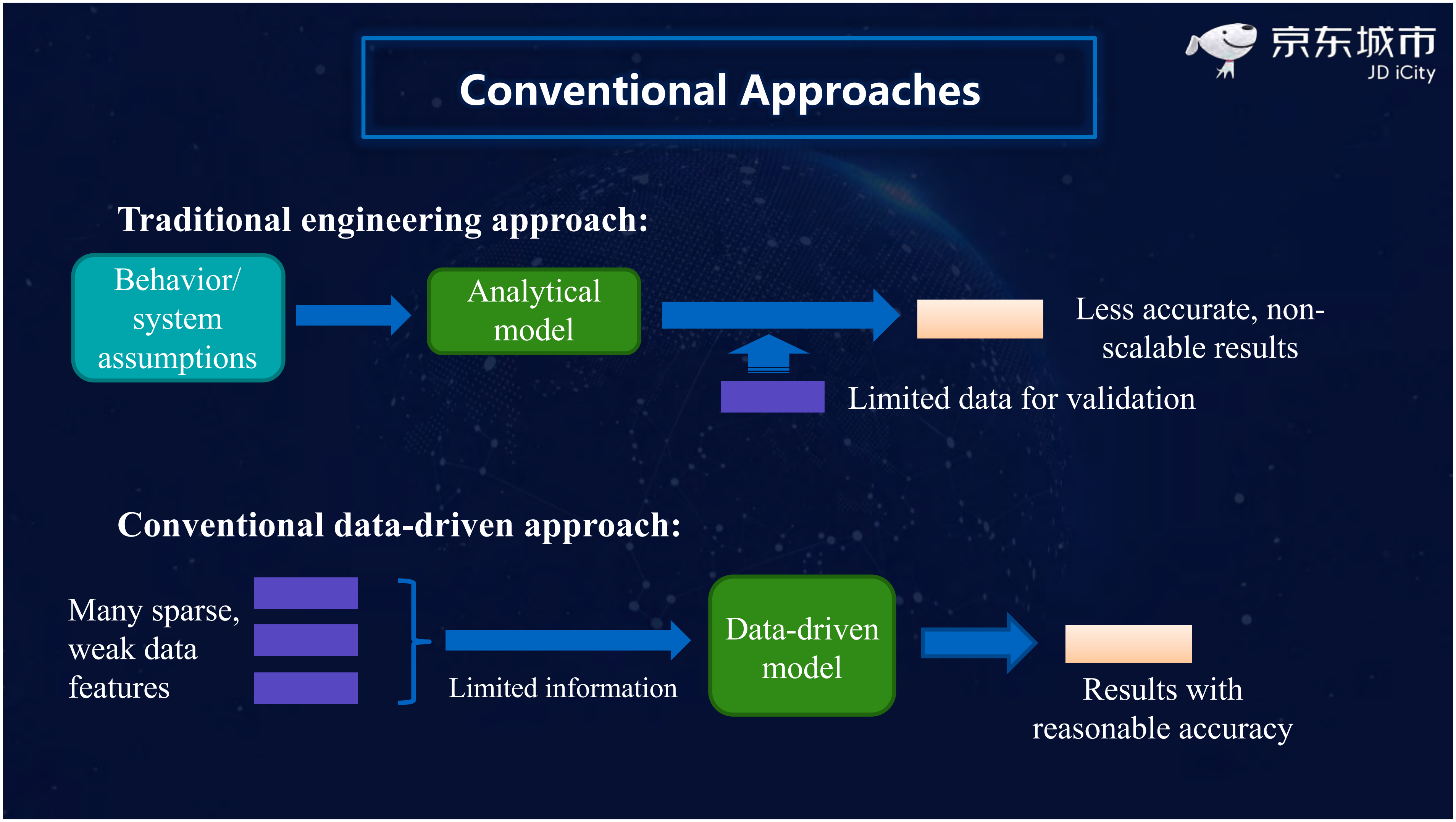


基于共享单车数据的城市违章停车智能监测

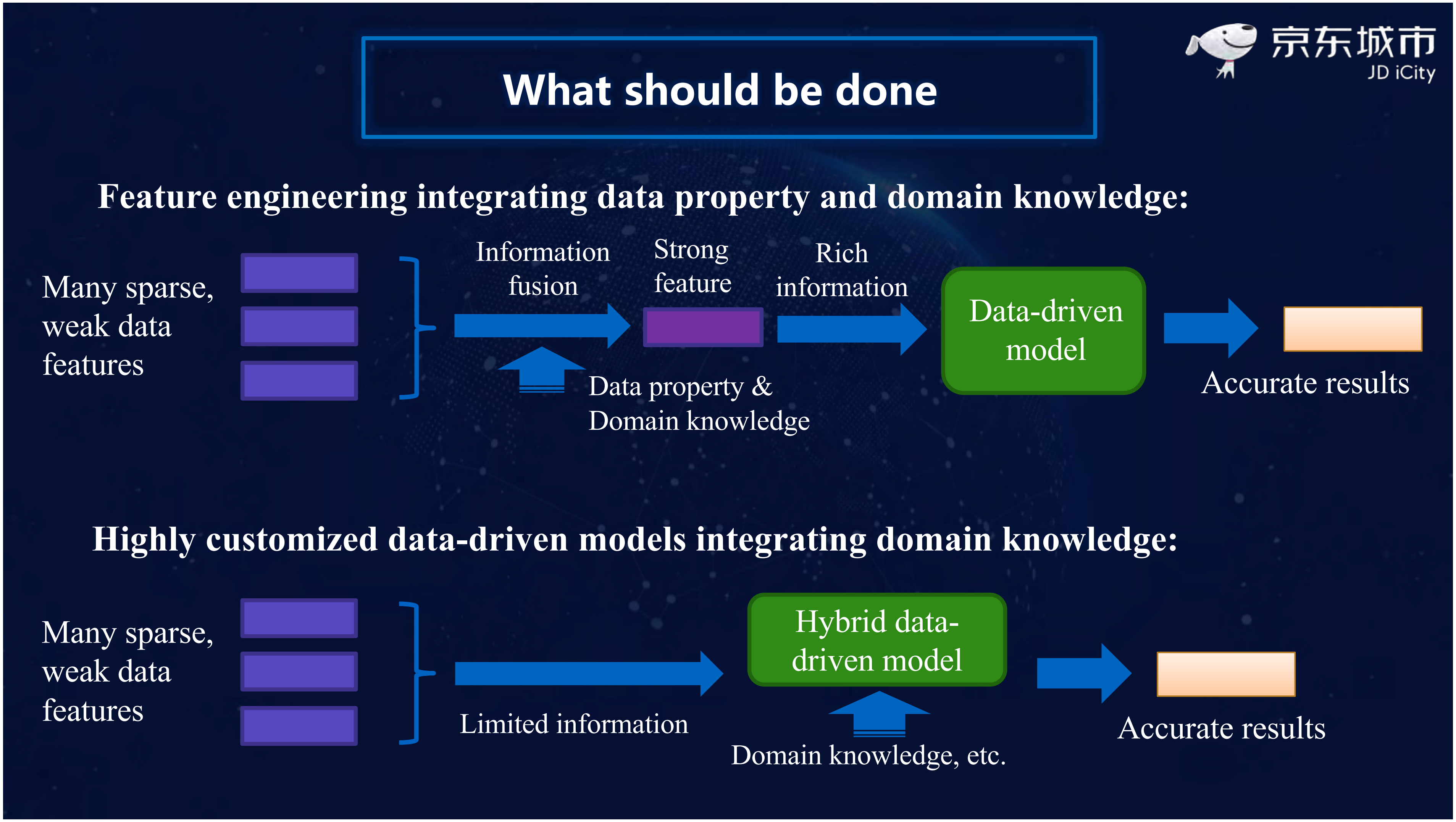
T. He, J. Bao, R. Li, S. R., Y. Li, C. Tian, Y. Zheng. Detecting Vehicle Illegal Parking Events using Sharing Bikes' Trajectories. KDD 2018

新思路

数据驱动思维



京东城市
JD iCity
Conventional Approaches
Traditional engineering approach:
Behavior/ system assumptions
Analytical model
Less accurate, non-scalable results
Limited data for validation
Conventional data-driven approach:
Many sparse, weak data features
Limited information
Data-driven model
Results with reasonable accuracy


What should be done
京东城市
JD iCity
Feature engineering integrating data property and domain knowledge:
Many sparse, weak data features
Information fusion
Strong feature
Rich information
Data-driven model
Data property & Domain knowledge
Accurate results
Highly customized data-driven models integrating domain knowledge:
Many sparse, weak data features
Limited information
Hybrid data-driven model
Domain knowledge, etc.
Accurate results

京东城市 JD iCity

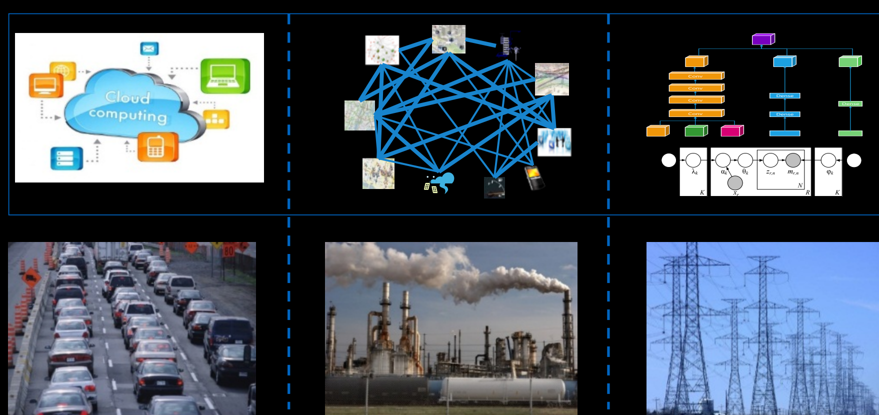
城市计算(Urban Computing)

城市数据的采集、管理、分析挖掘和服务提供

数据 + 计算

解决交通、规划、环境、能耗、公共安全、商业、医疗等痛点

云计算 + 大数据 + AI + 城市场景



服务提供

改进城市规划 缓解交通拥堵节约能耗 降低空气污染

城市数据分析

人工智能 模式识别 机器学习和可视化

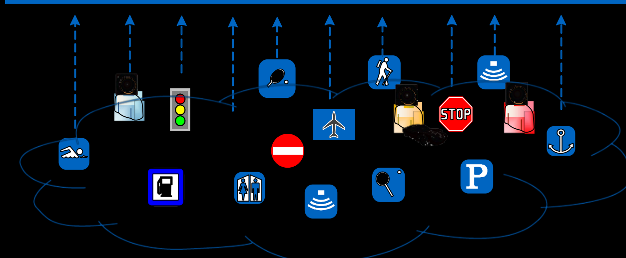
城市数据管理

时空索引 流数据 轨迹数据和图数据管理 异构数据索引

人的移动性数据　交通流　空气质量　气象数据　社交媒体　能耗数据　道路结构　兴趣点

城市感知和数据获取

参与感知，群体感知和移动感知计算

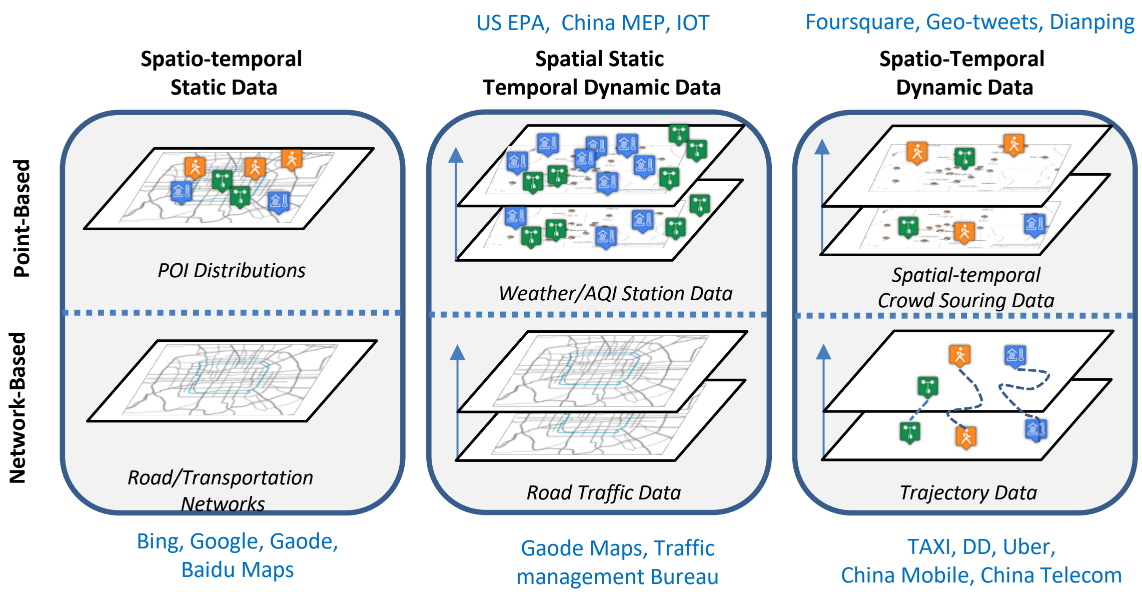


新方法

时空深度学习 & More

Taxonomy of Spatio-Temporal (ST) Data

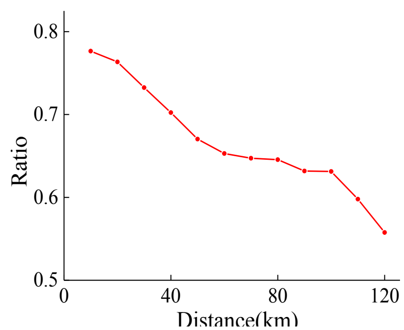
- Data Structures
- Spatio-temporal (ST) Properties

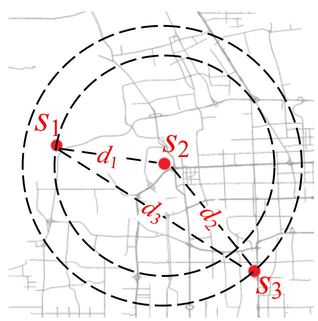


Why Spatio-Temporal Data Is Unique

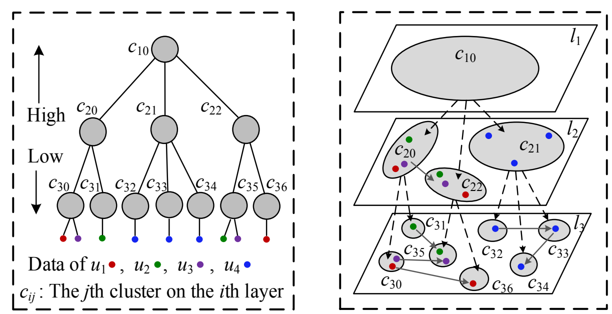
Spatial Properties

- Distance
  - Spatial closeness
  - Triangle inequality: $|d_1 - d_2| \leq d_3 \leq |d_1 + d_2|$
- Hierarchy
  - Different spatial granularities
  - City structures





$\boldsymbol{u_1}: u_2 > u_4$

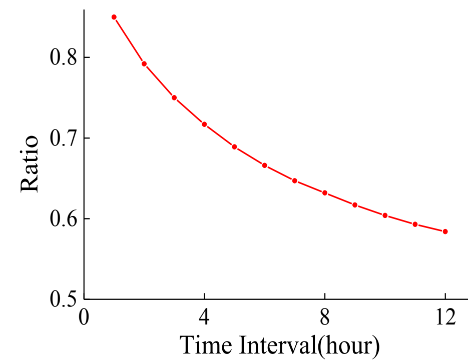


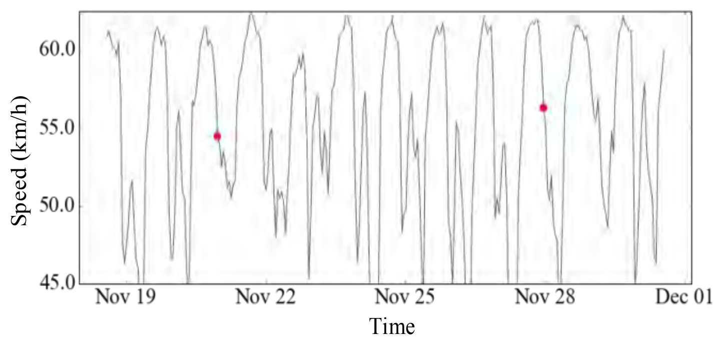
$\boldsymbol{u_1}: u_3 > u_2$

Junbo Zhang, Yu Zheng, et al. Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction, AAAI 2017

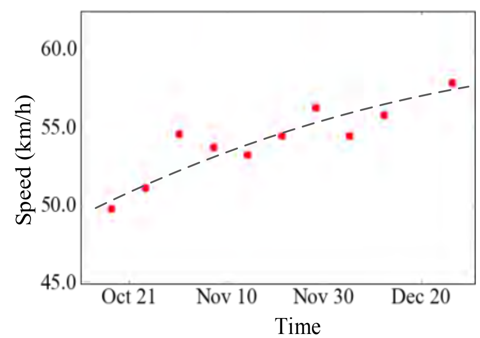
Why Spatio-Temporal Data Is Unique

- Temporal properties
  - Temporal closeness
  - Period
  - Trend





A) Hourly traffic speed on consecutive days



B) Traffic speed at 9-10am on consecutive weekends

Junbo Zhang, Yu Zheng, et al. [Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction](), AAAI 2017

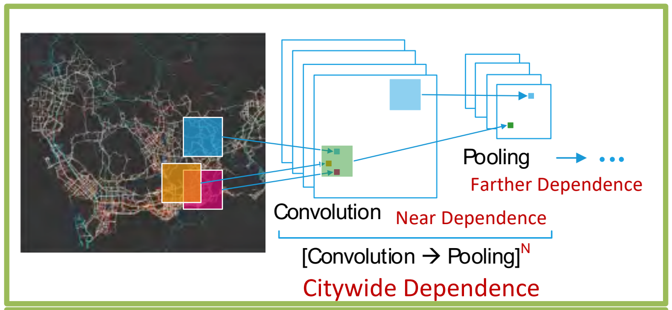
Deep Learning meets ST Data

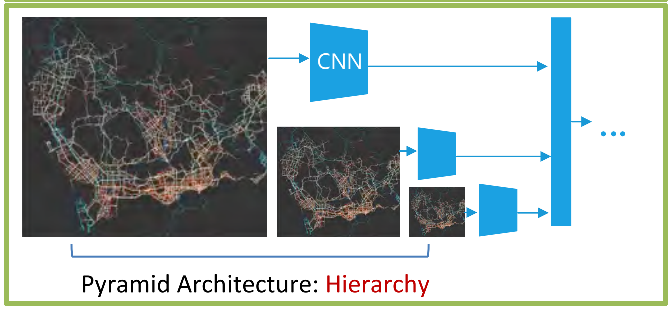
- What Deep Learning can do for ST Data
  - Encoding a (single) ST dataset
  - Fusing multiple ST datasets
- What ST data can provide to Deep Learning
  - Massive and diverse Data
  - Computing infrastructures are ready
  - Application scenarios requiring
    - Instantaneous responses at large spaces
    - Collective computing
    - (traditional machine learning models many not be able to handle)

Taxi Trajectory Data of Shenzhen

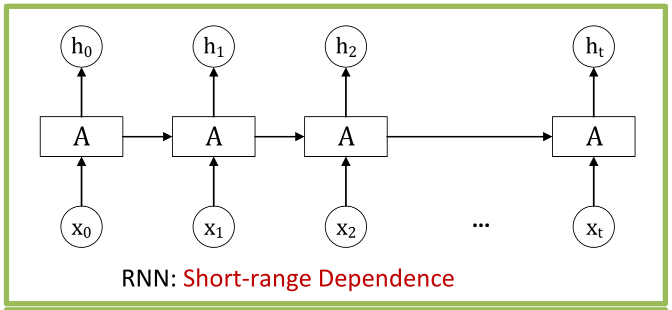
Encoding Spatio-Temporal Properties

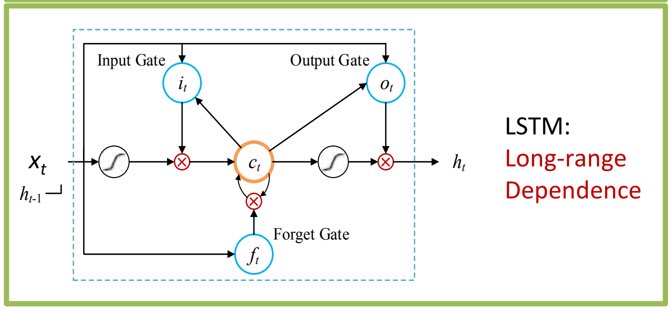
CNN is able to model spatial properties





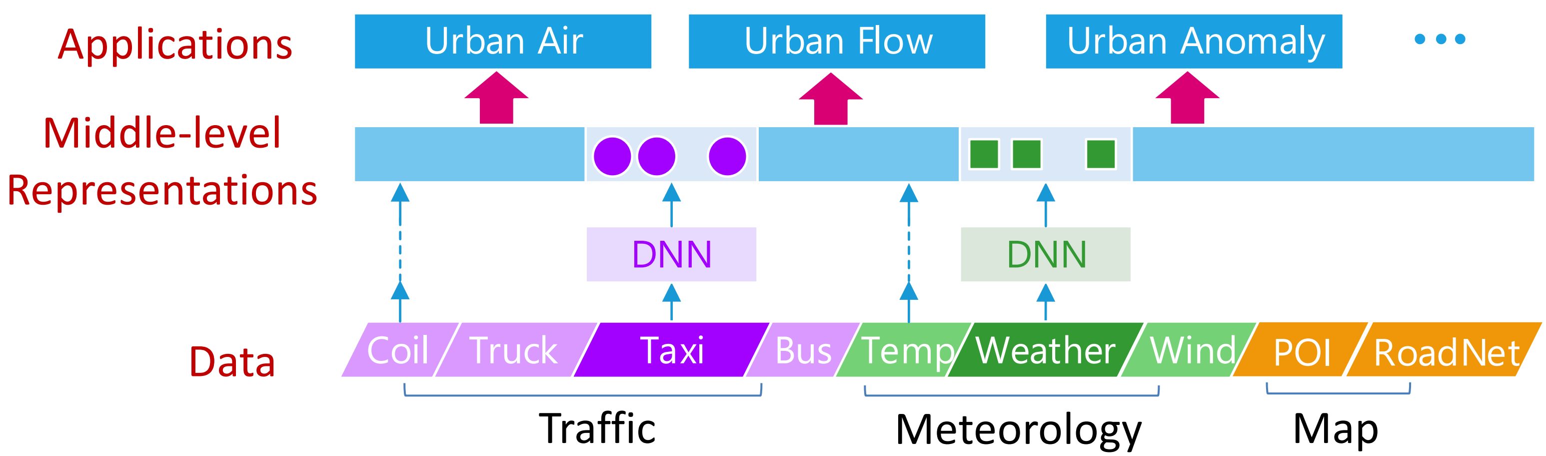
RNN/LSTM is able to model **temporal** properties





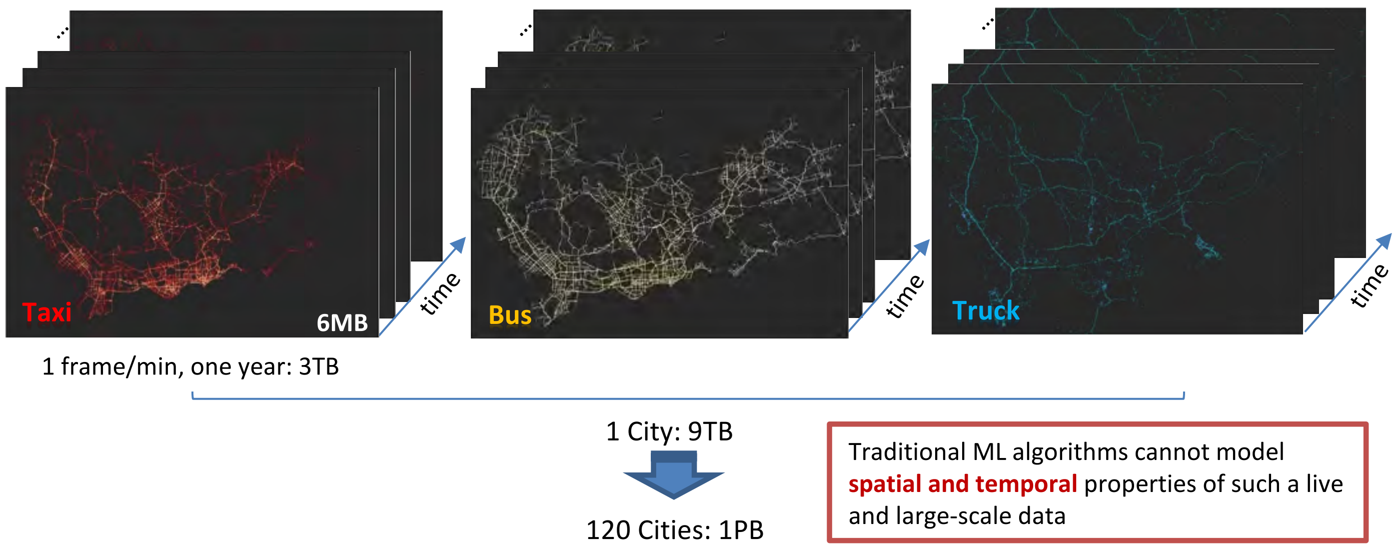
Trajectories of taxis, trucks and buses

Fusing Multiple ST-Datasets



Why Deep Learning for ST Data

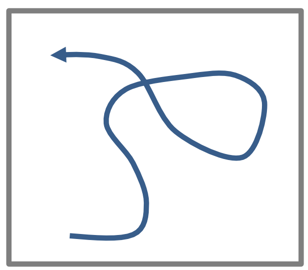
- Big ST-Data

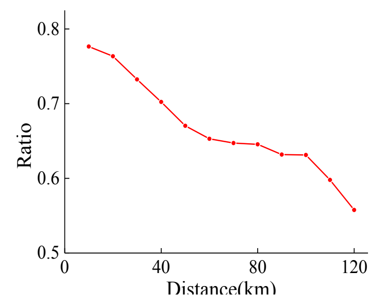


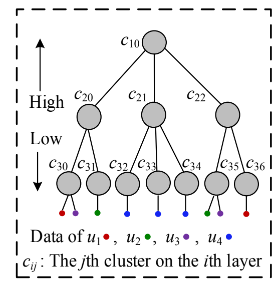
Traditional ML algorithms cannot model **spatial and temporal** properties of such a live and large-scale data

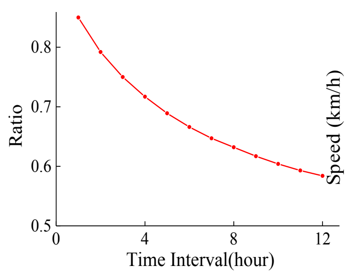
Challenges of DL for ST Data

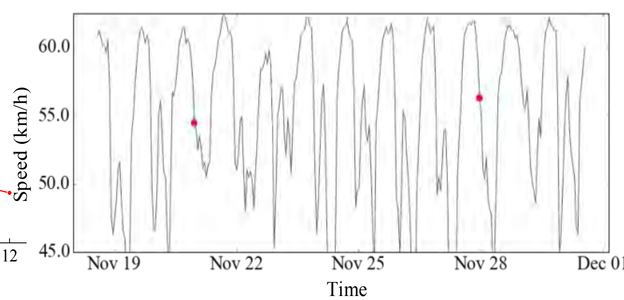
- Data transformation
- Encoding ST properties in DNNs



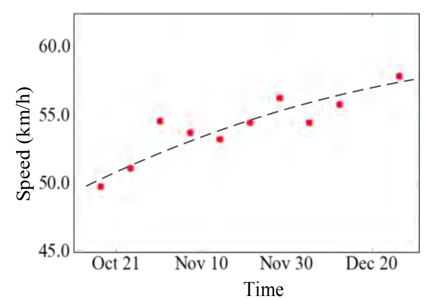






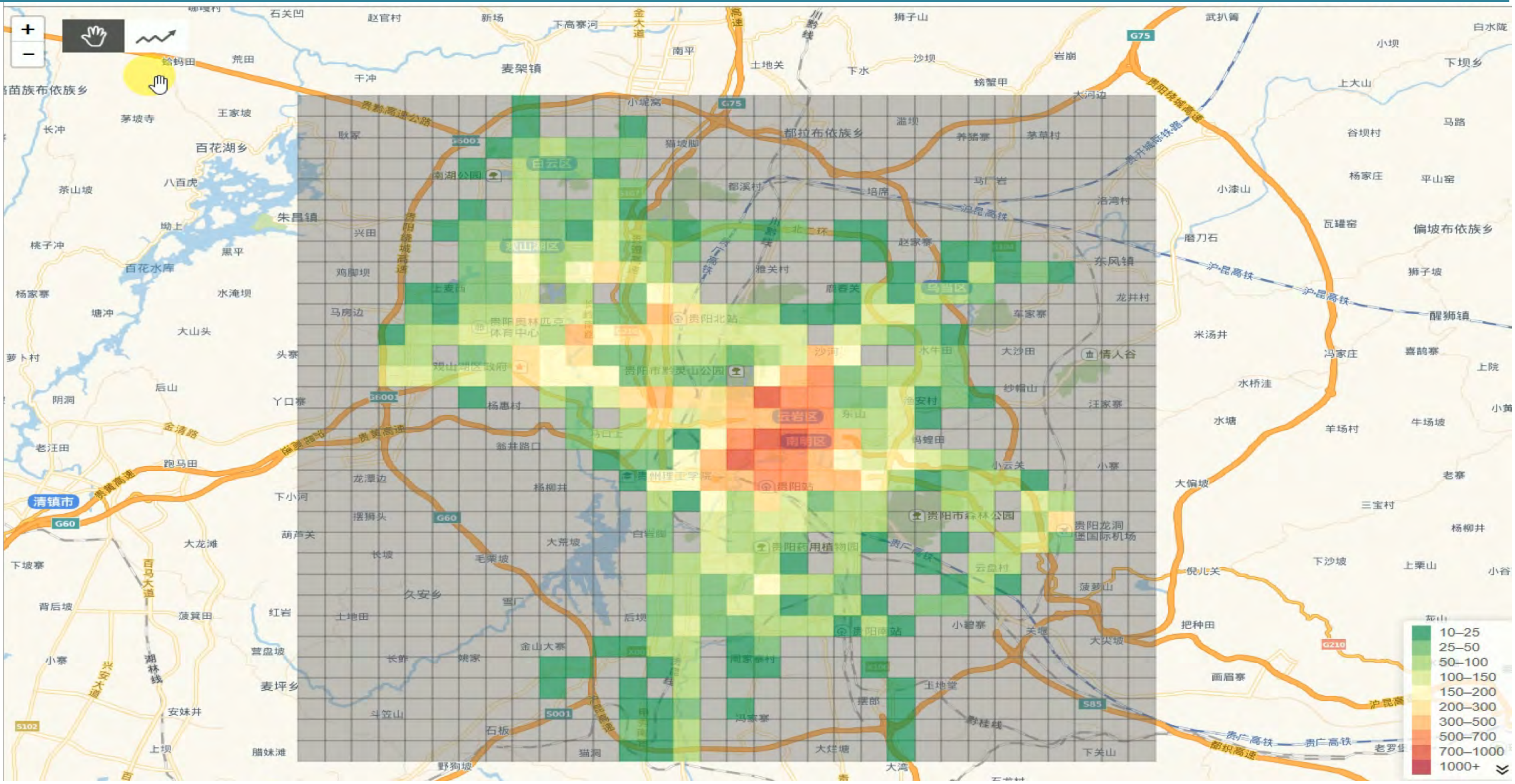


A) Hourly traffic speed on consecutive days



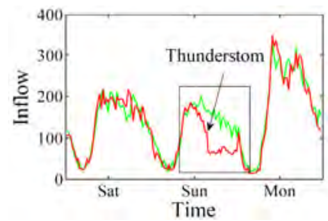
B) Traffic speed at 9-10am on consecutive weekends

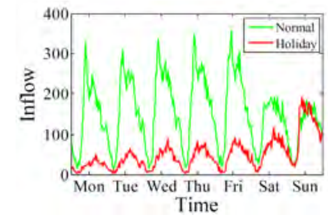
AI预测城市栅格区域人群流量

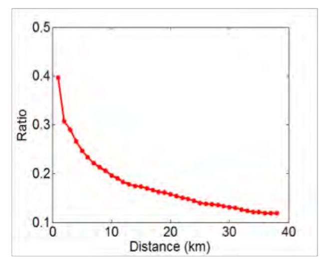


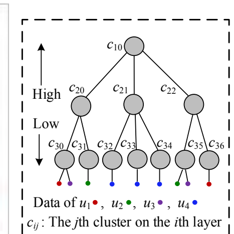
Challenges

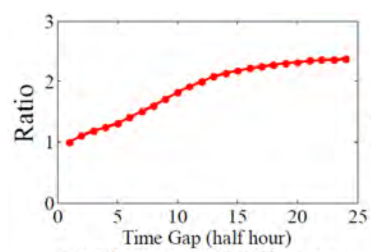
- Urban crowd flow depends on many factors
  - Flows of previous time interval
  - Flows of nearby regions and distant regions
  - Weather, traffic control and events
- Capturing spatial properties
  - Spatial distance and hierarchy
- Capturing temporal properties
  - Temporal closeness
  - Period and trend



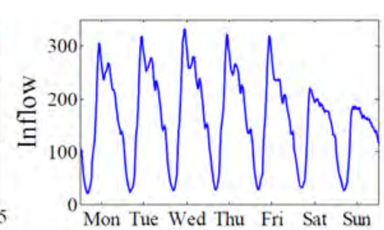




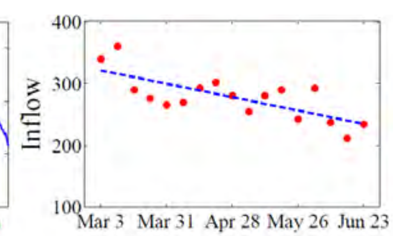




(a) Closeness of Office Area

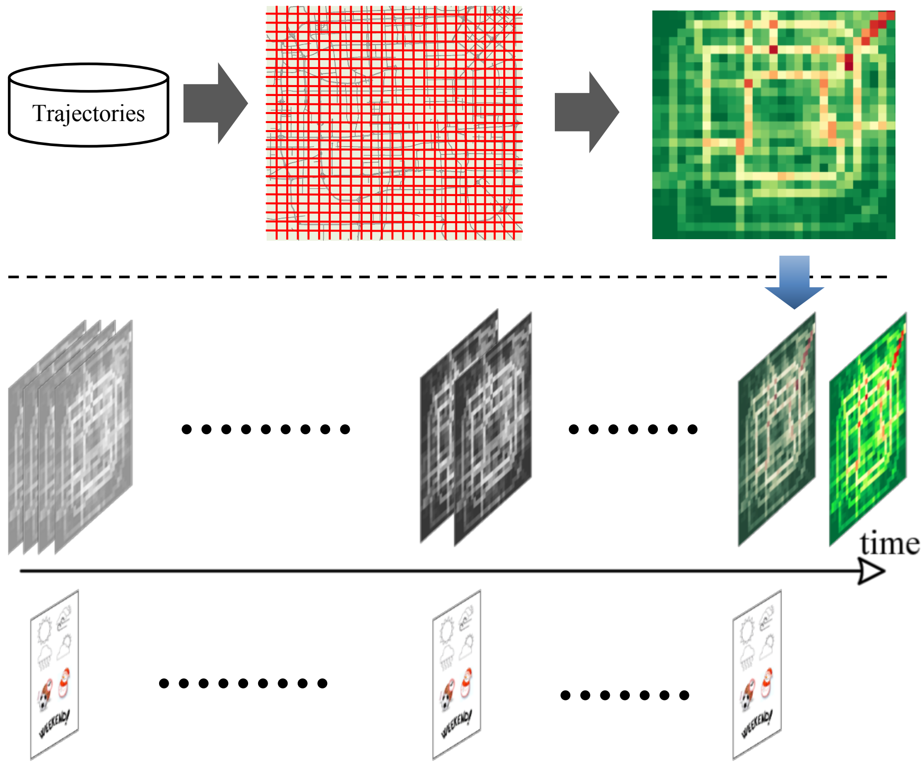


(b) Period of Office Area



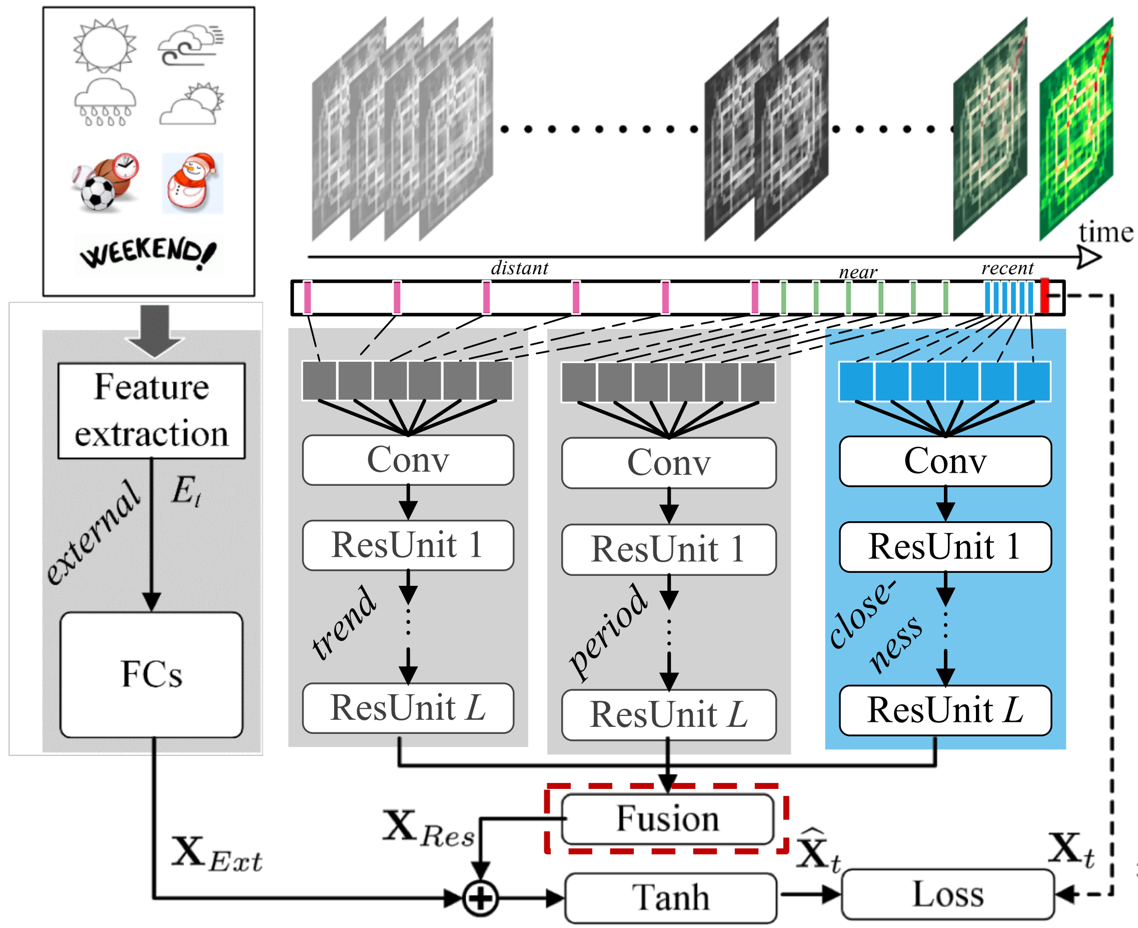
(c) Trend of Office Area

Converting Trajectories into Video-like Data



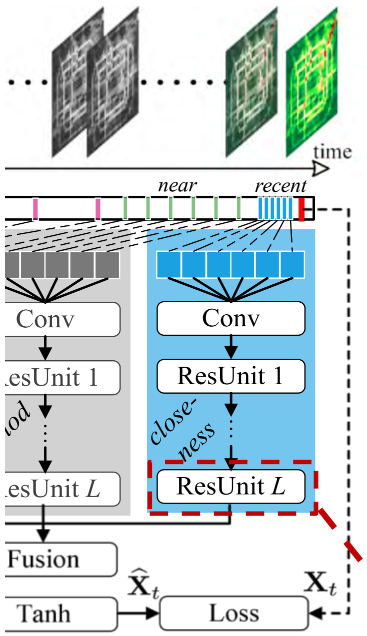
Junbo Zhang et al. Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction, AAAI 2017

ST-ResNet: A Collective Prediction

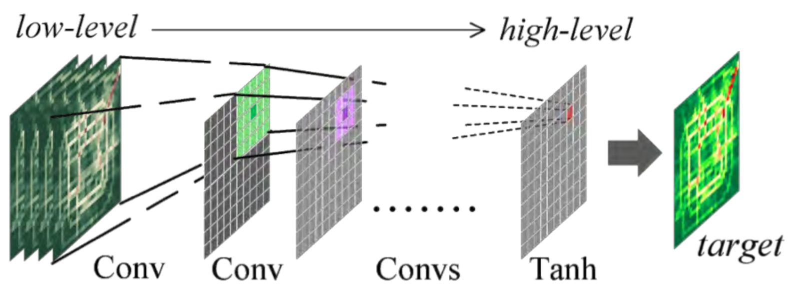


Junbo Zhang et al. [Predicting Citywide Crowd Flows Using Deep Spatio-Temporal Residual Networks](), AI Journal, 2018

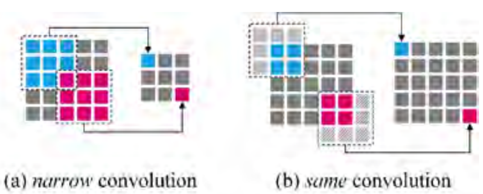
Residual Deep Convolutional Neural Network



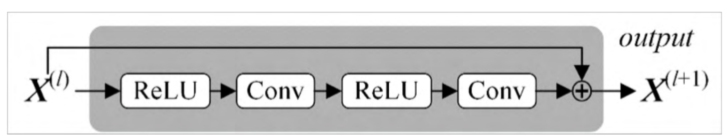
Capturing spatial correlation of both near and far



Con
v

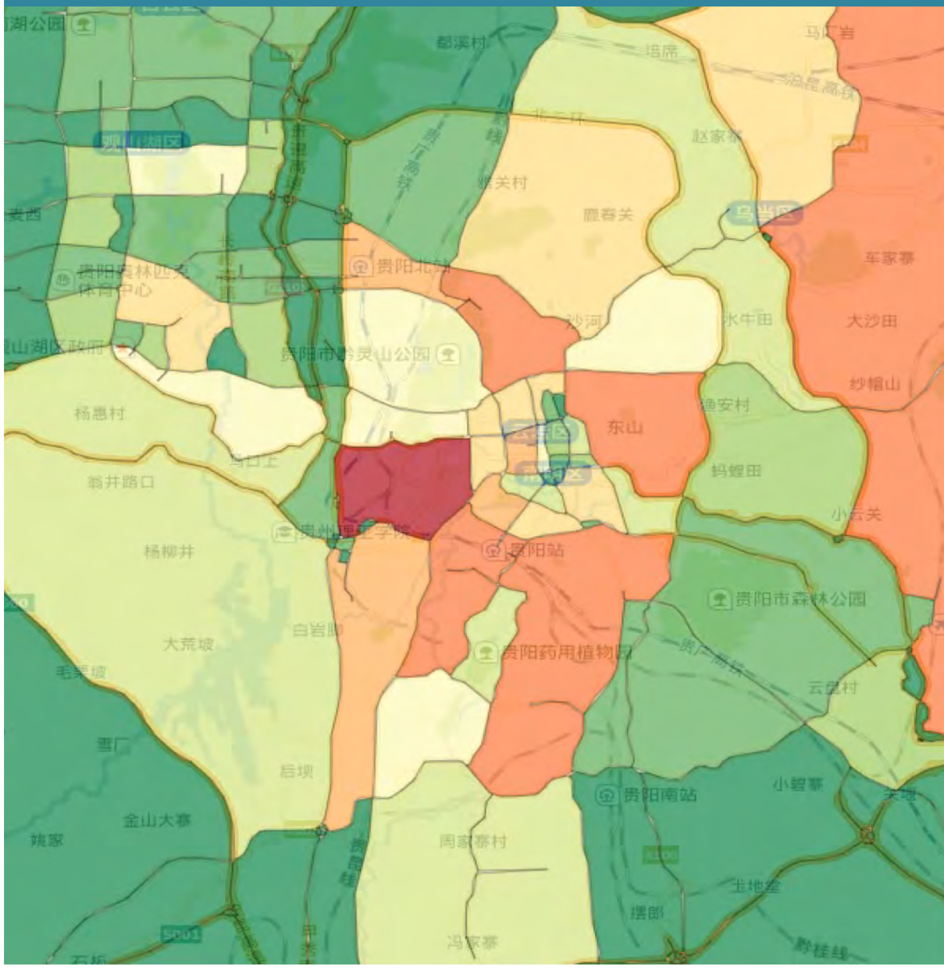


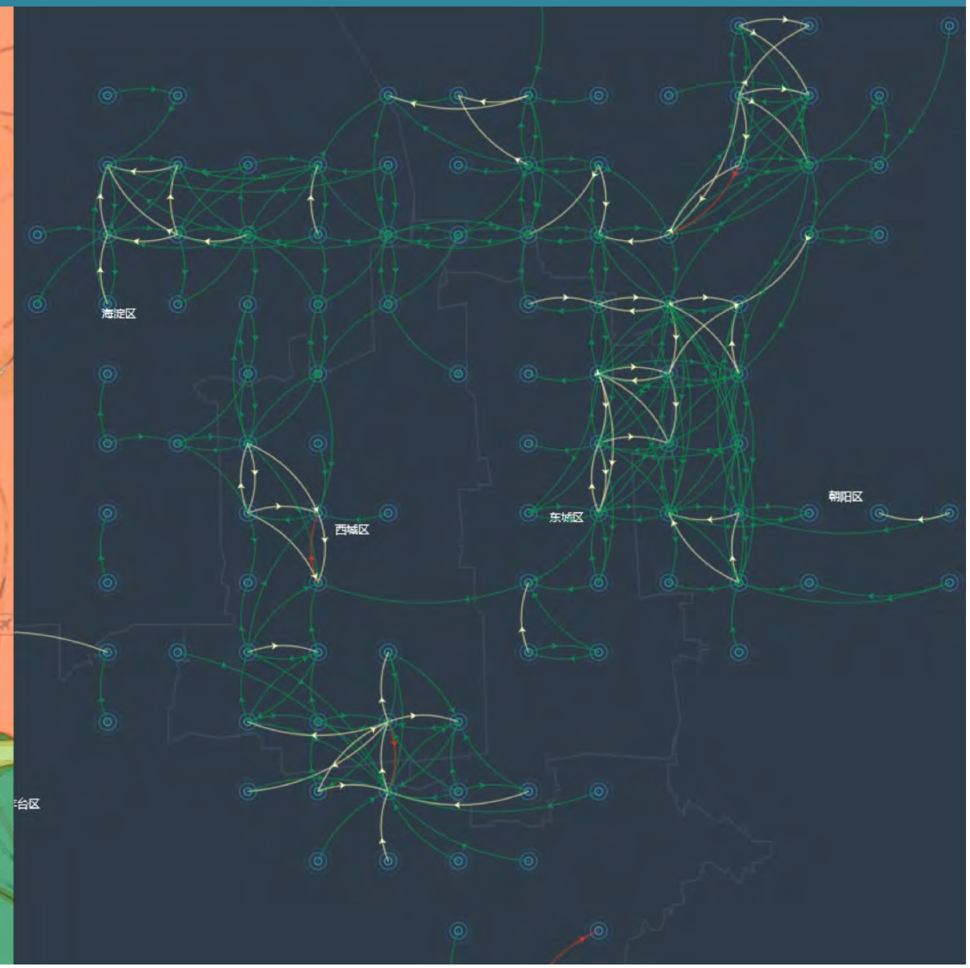
Using residual network framework to help training



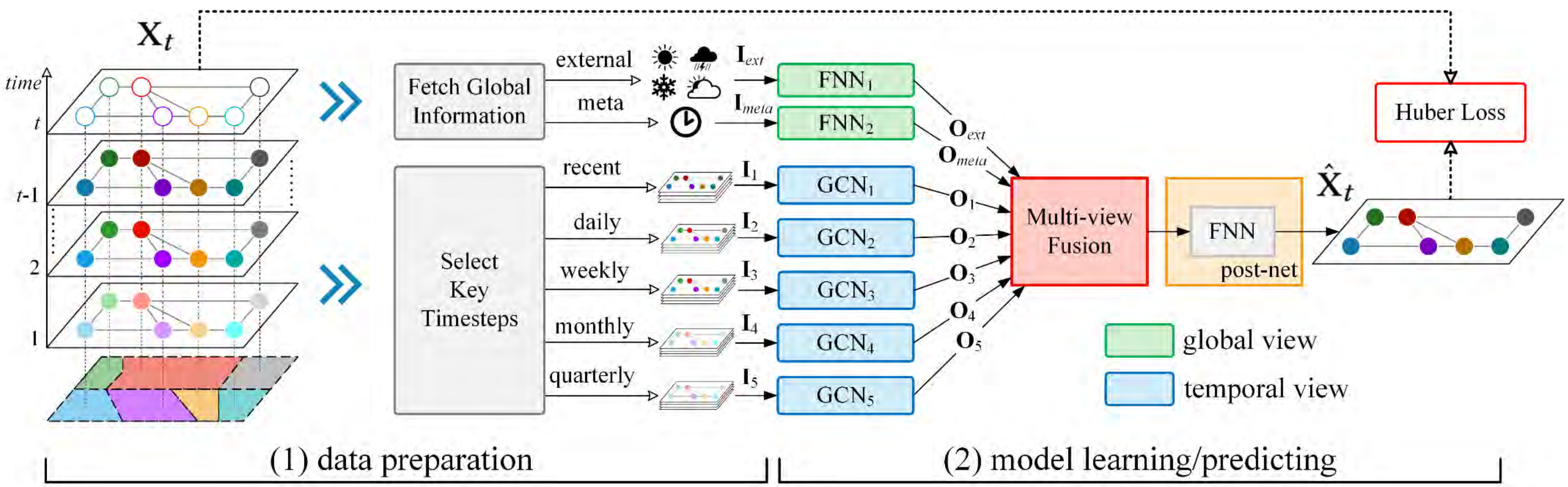
Junbo Zhang et al. Predicting Citywide Crowd Flows Using Deep Spatio-Temporal Residual Networks, AI Journal, 2018

AI预测城市区域人流量及流转

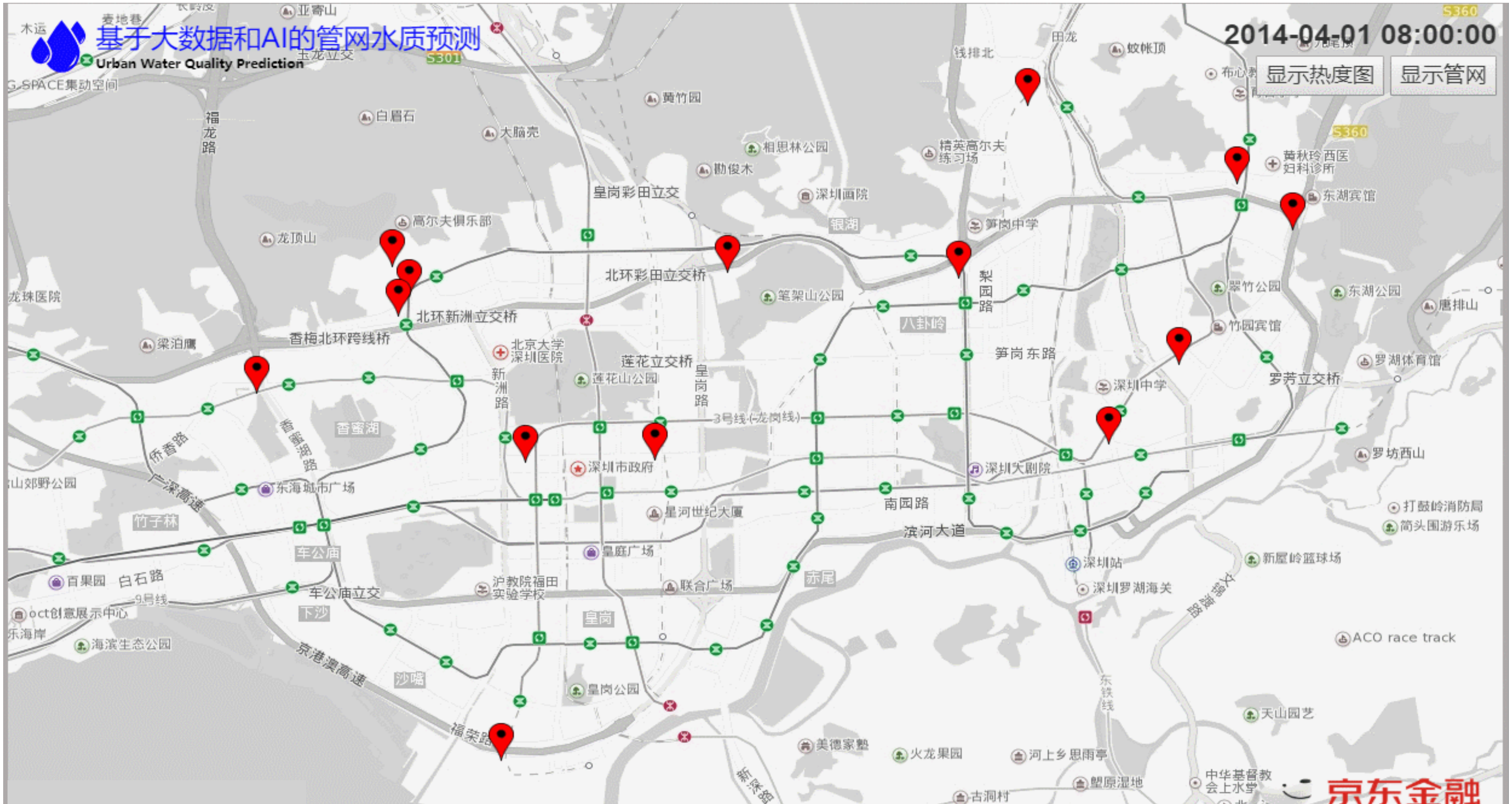




Multi-view Graph Convolutional Networks



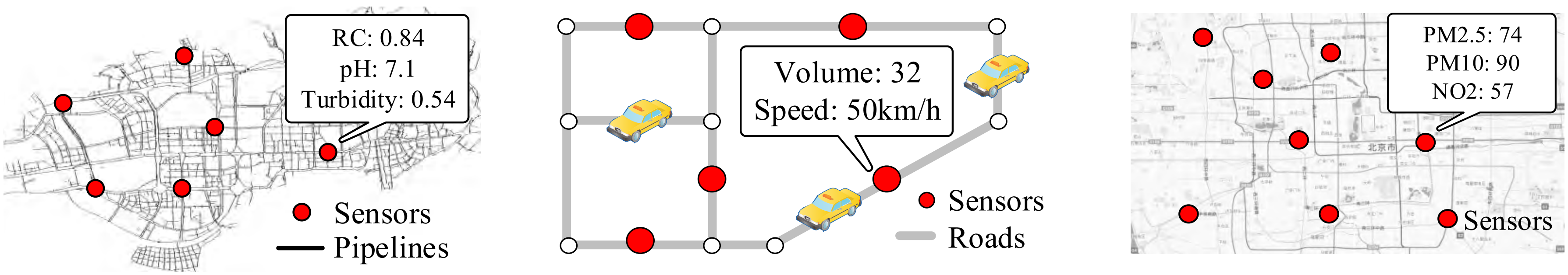
Junbo Zhang et al. Predicting Citywide Crowd Flows in Irregular Regions Using Multi-View Graph Convolutional Networks. Submitted



Yuxuan Liang, Songyu Ke **Junbo Zhang et al**. GeoMAN: Multi-level Attention Networks for Geo-sensory Time Series Prediction. IJCAI 2018

Geo-sensory Time Series

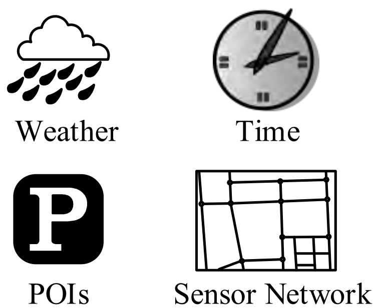
- There are massive sensors deployed in physical world

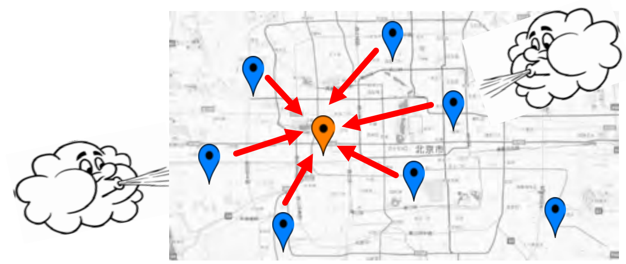


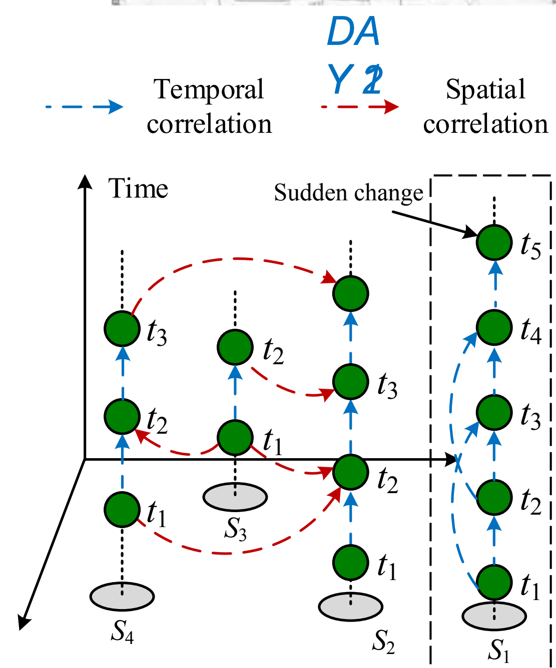
- Properties
  - Each sensor has a unique geospatial location
  - Constantly reporting time series readings about different measurements
  - With geospatial correlation between their readings

Challenges

- Dynamic inter-sensor correlations
- Dynamic temporal correlations
- Affected by many factors
  - Readings of previous time interval
  - Readings of other sensors in nearby regions
  - External factors: weather, time and land use

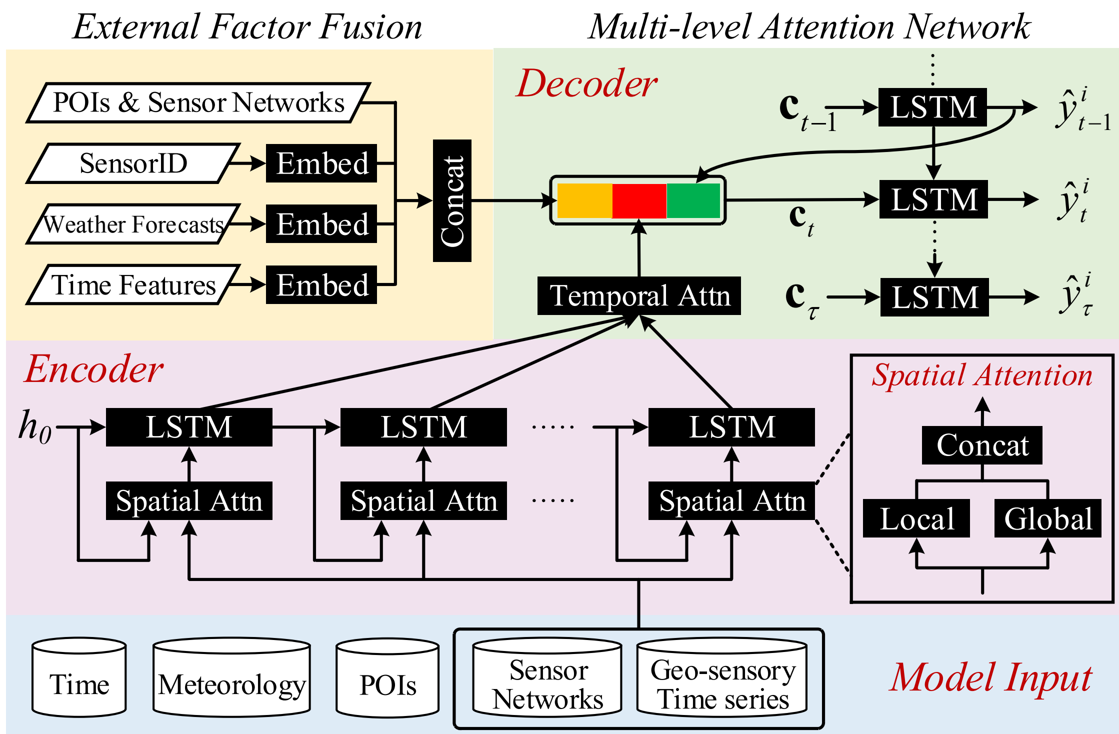






GeoMAN: Multi-level Attention Networks

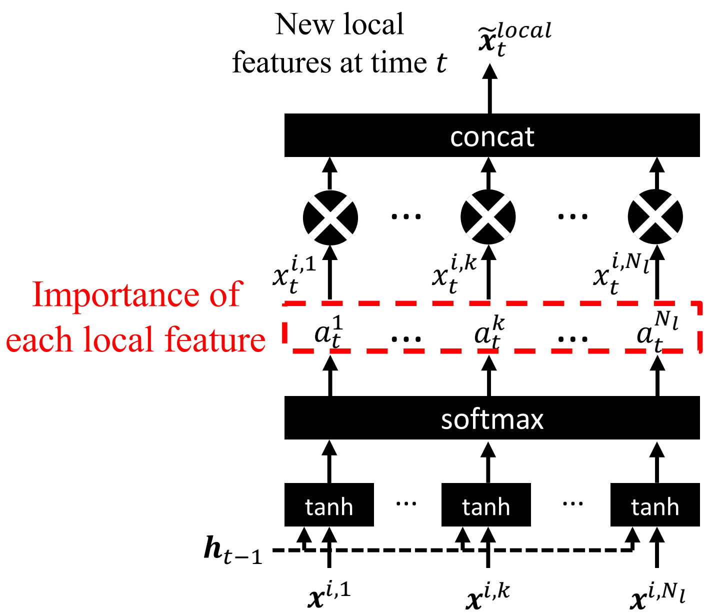
- Spatial attention to capture complex spatial correlations
- Temporal attention to model dynamic temporal correlations
- Fusion module to incorporate the external factors



Yuxuan Liang, Songyu Ke, Junbo Zhang, et al. GeoMAN: Multi-level Attention Networks for Geo-sensory Time Series Prediction. IJCAI 2018

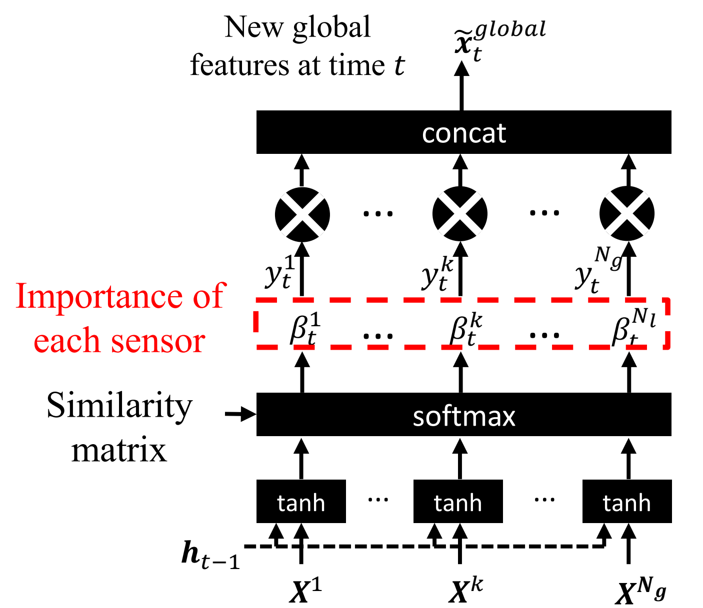
Spatial Attention

- Local spatial attention



Local features of a given sensor

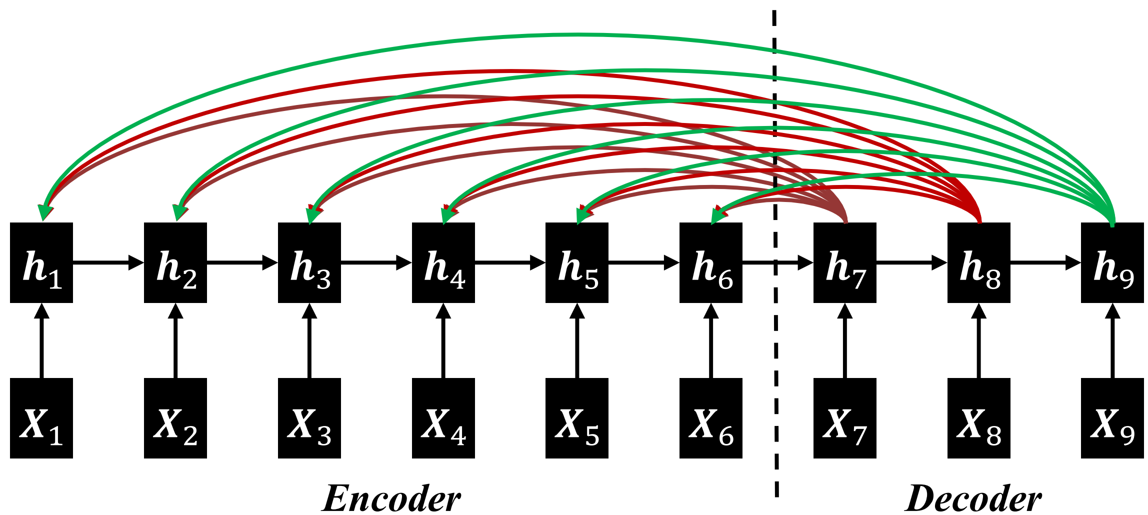
- Global spatial attention



Historical readings of each sensor

Temporal Attention

- Sequence-to-sequence learning architecture
- Select relevant previous time slots to make predictions



Evaluation

- Task 1 - water quality prediction
  - Water quality data
    - Residual chlorine
    - 10 kinds of time series
    - From 14 sensors in Shenzhen
    - Update each 5 minutes
  - Meteorology data
  - POIs data

- Task 2 - air quality prediction
  - Air quality data
    - PM2.5
    - 19 kinds of time series
    - From 35 sensors in Beijing
    - Hourly updates
  - Meteorology data
  - POIs data

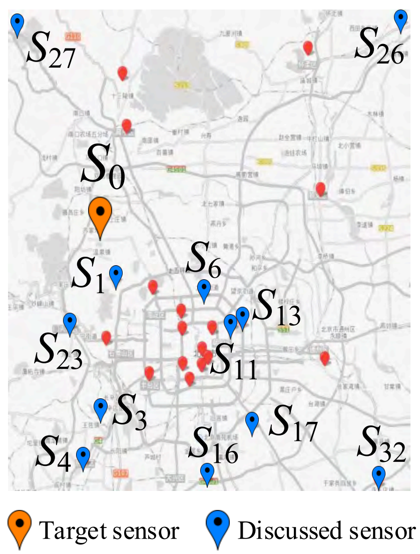
Results

Method	Water Quality		Air Quality	
	RMSE	MAE	RMSE	MAE
ARIMA	8.61E-02	7.97E-02	31.07	20.58
VAR	5.02E-02	4.42E-02	24.60	16.17
GBRT	5.17E-02	3.30E-02	24.00	15.03
FFA	6.04E-02	4.10E-02	23.83	15.75
stMTMVL	6.07E-02	4.16E-02	29.72	19.26
stDNN	5.77E-02	3.99E-02	25.64	16.49
LSTM	6.89E-02	5.04E-02	24.62	16.70
Seq2seq	5.80E-02	4.03E-02	24.55	15.09
DA-RNN	5.02E-02	3.52E-02	24.25	15.17
GeoMAN	**4.34E-02**	**3.02E-02**	**22.86**	**14.08**

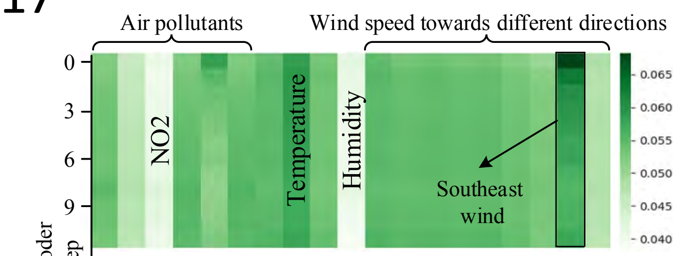
Yuxuan Liang, Songyu Ke, Junbo Zhang, et al. GeoMAN: Multi-level Attention Networks for Geo-sensory Time Series Prediction. IJCAI 2018

Visualization: Dynamic Correlation

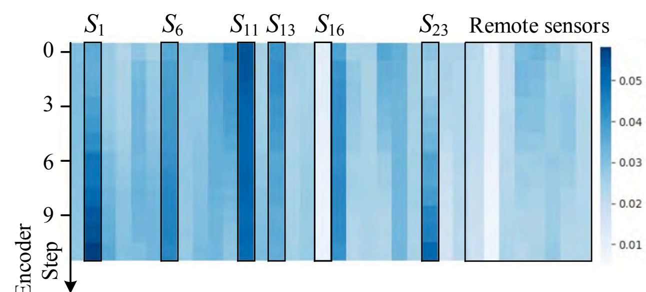
- Case study over air quality dataset
  - Discuss on sensor S_0
  - 4:00 to 16:00 on Feb. 28, 2017



(a) Air quality stations in Beijing

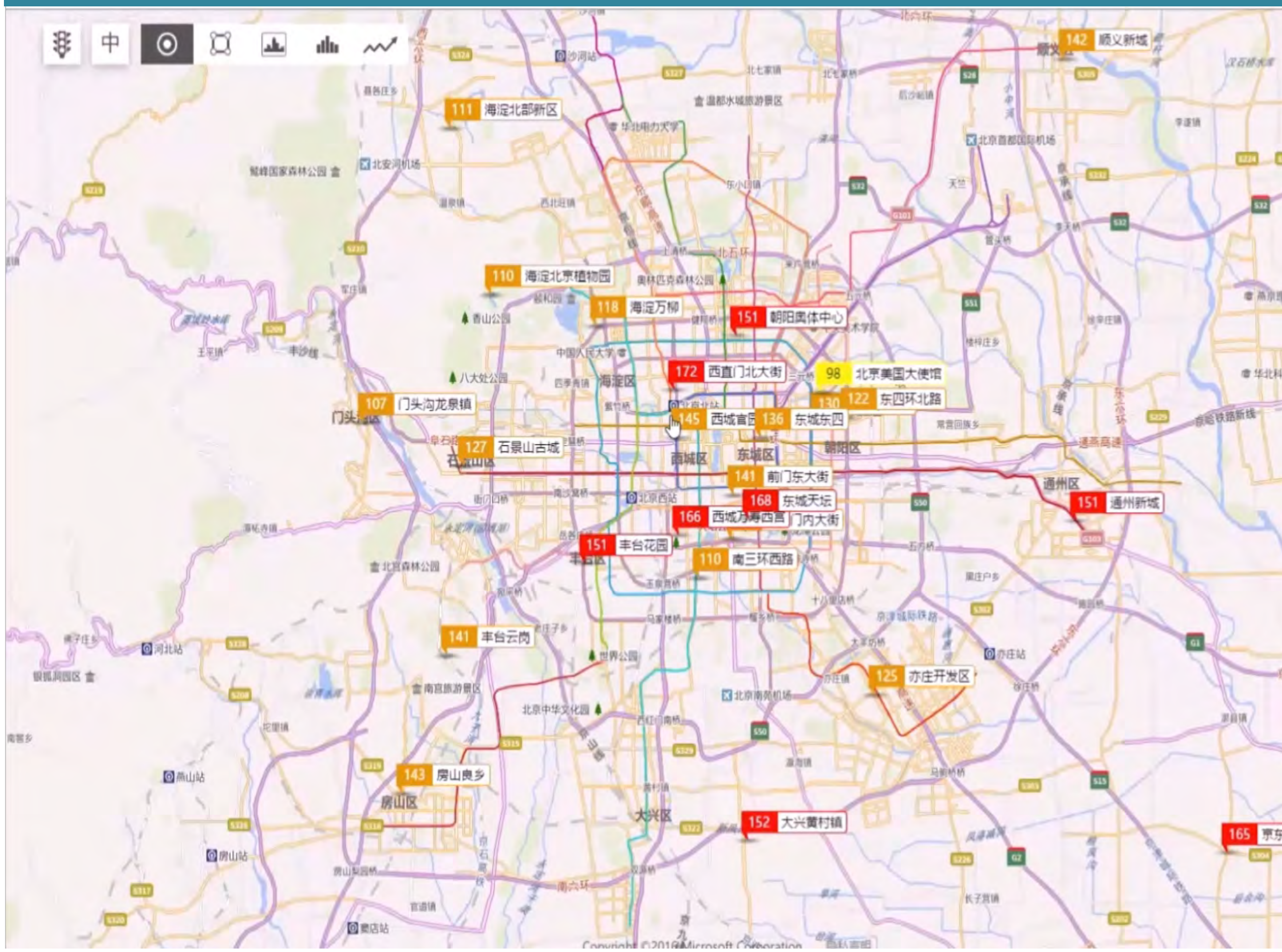


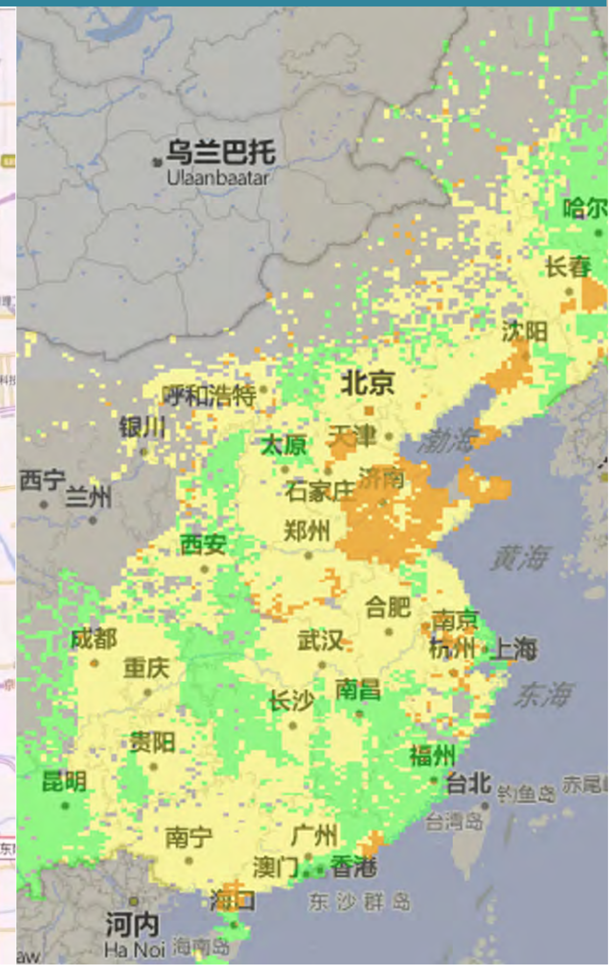
(b) Plot of local spatial attention weights



(c) Plot of global spatial attention weights

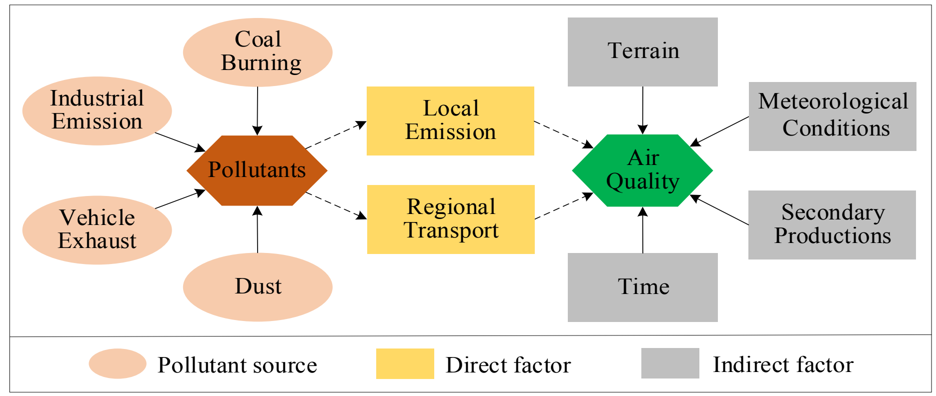
基于大数据和AI的空气质量预测



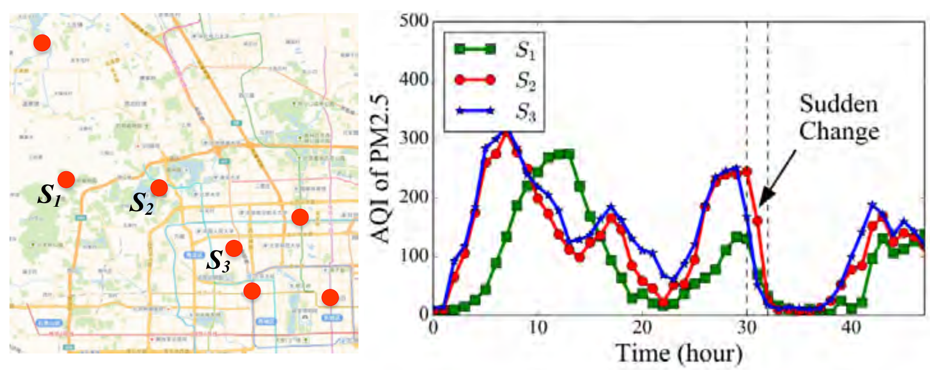


Challenges

- Multiple influential factors with complex interactions
  - Pollution sources, direct factors and indirect factors
  - Affected by multiply factors simultaneously

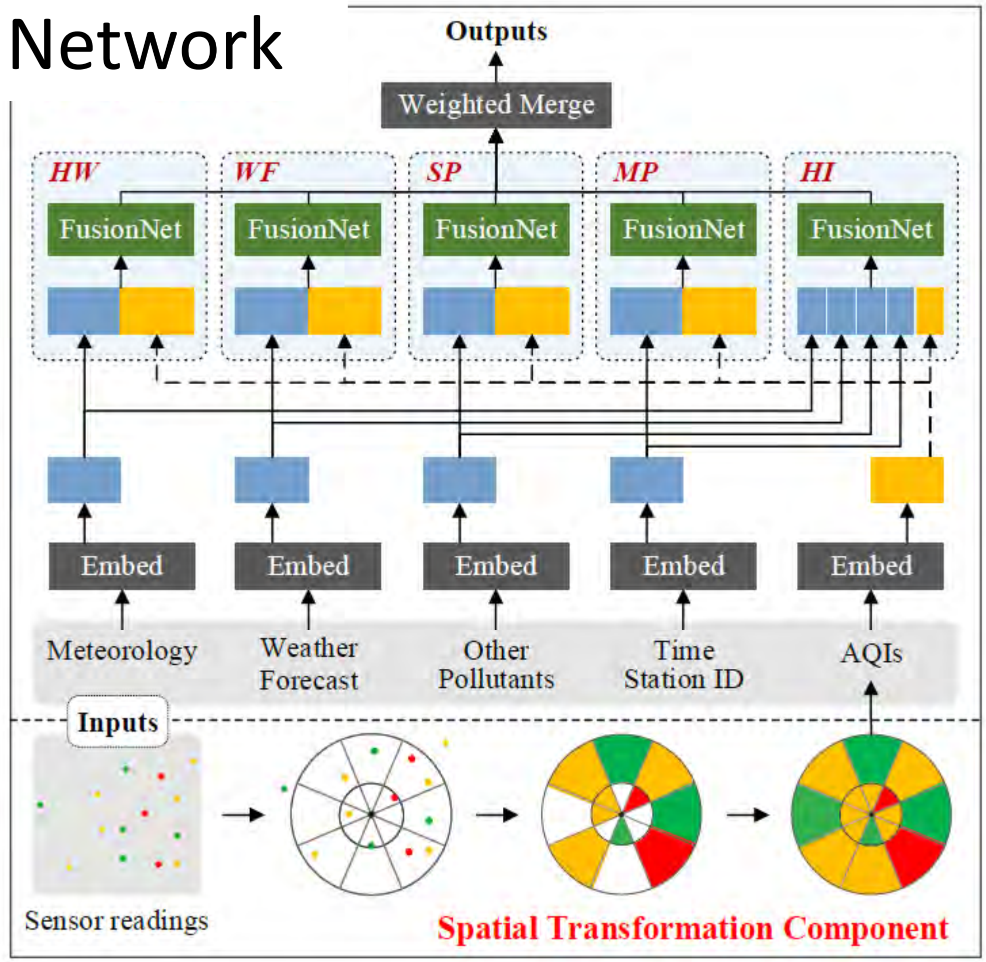


- Dynamic spatio-temporal correlation and sudden changes
  - Urban air changes over location and time significantly
  - AQI drops very sharply in a very short time span



A) Monitoring stations B) AQI change over time

Deep Distributed Fusion Network



Spatial Transformation

- Air pollution dispersion
- Spatial correlation
- Scalability

Distributed FusionNet

- HW/WF/SP/MP nets to capture different individual influences
- Capture holistic influence (HI)

Weighted Merge

$$\hat{\boldsymbol{y}} = Sigmoid(\boldsymbol{y}_{hw} \circ \boldsymbol{w}_{hw} + \boldsymbol{y}_{wf} \circ \boldsymbol{w}_{wf} + \boldsymbol{y}_{sp} \circ \boldsymbol{w}_{sp} + \boldsymbol{y}_{mp} \circ \boldsymbol{w}_{mp} + \boldsymbol{y}_{hi} \circ \boldsymbol{w}_{hi})$$

Xiuwen Yi, Junbo Zhang, et al. Deep Distributed Fusion Network for Air Quality Prediction. KDD

Official Prediction

- Advantages beyond Weather-Forecast-Based Method (WFM)
  - Spatial granularity: station vs district
  - Farther predictive capability: 48 vs 12 hours
  - Updating frequency: 1 hour vs 12 hours
  - Need less data sources
  - More accurate, 22% improvement

10/1/2014 to 12/30/2016.

Beijing Municipal Environmental Monitoring Center (using *WFM*)

Method	Station Level		District Level		Update	Grained
	acc	*mae*	*acc*	*mae*	Hours	level
WFM	0.54	54.5	0.64	46.1	12	District
DeepAir	0.77	26.7	0.86	17.9	1	Station

火力发电行业背景

假如发电效率从90%提到到90.5%

- 发电形式
  - 水电，火电，核电，风电和太阳能
  - 火力发电约占总发电量70%
- 火电装机容量
  - 全国总10.5亿千瓦
  - 约2000台机组（锅炉）

每年为国家节约100个亿!

			价值
一台60万千万火电机组一年可节省煤	0.55万吨	=	500万元
全国一年可节省煤	962.5万吨	=	87.5亿元

经济效益 ≈

全北京四个月的总用煤量

环境效益

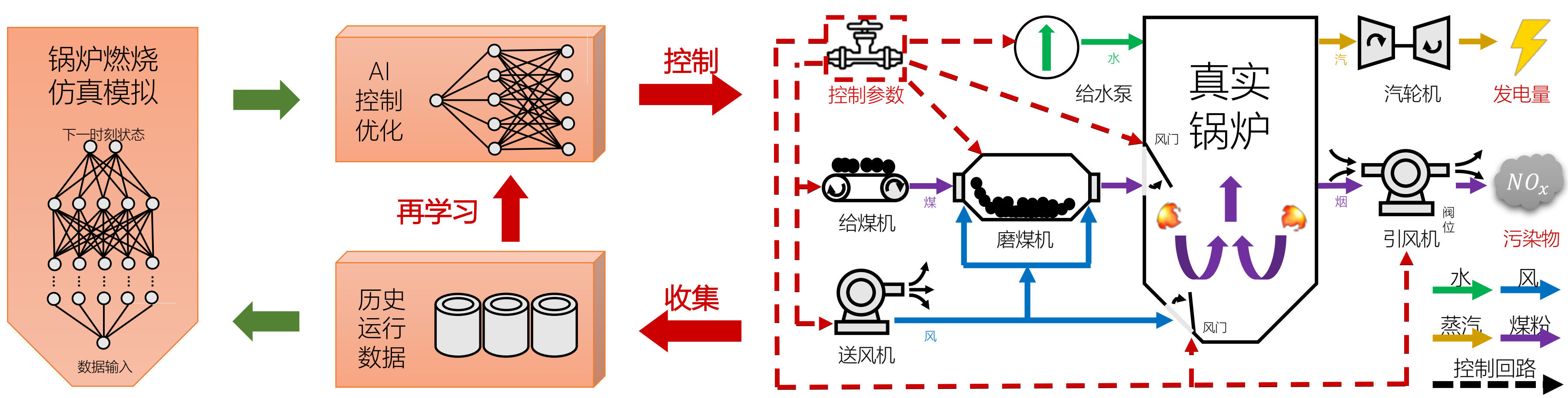
减少污染物排放

		总治理成本
二氧化硫	38.5万吨	23.1亿元
氮氧化物	48.1万吨	

AI + 火力发电

用更少的煤　　发更多的电　　更少的污染

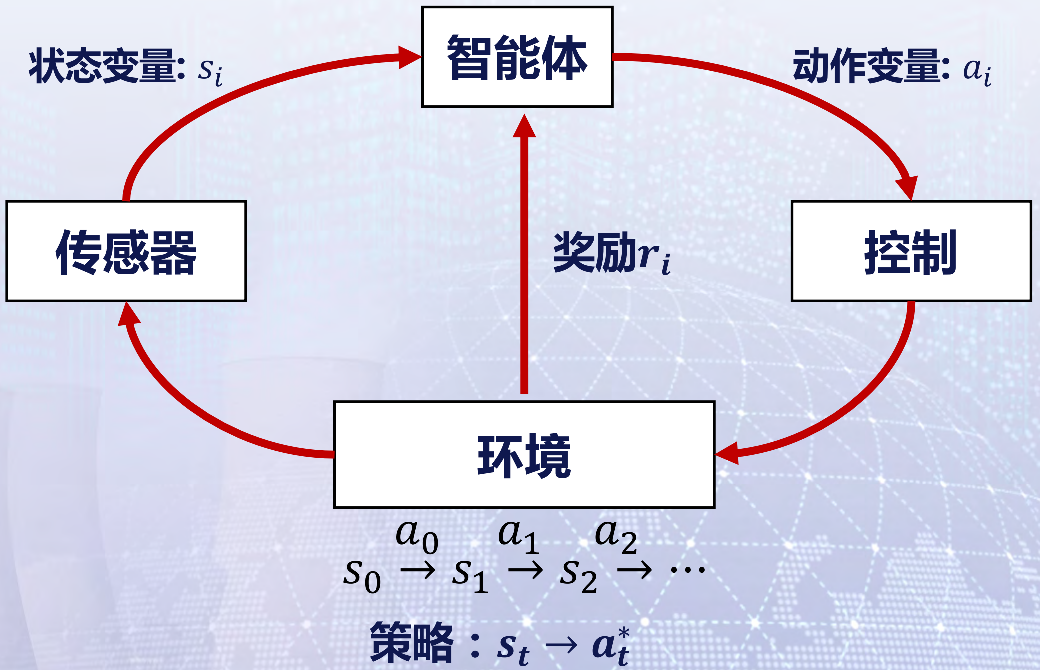
离线学习　　在线运行



深度强化学习优化

京东城市 JD iCity

- 磨煤机出口压力
- 燃烧器风粉温度
- 主蒸汽压力
主蒸汽温度
- 炉膛负压
- 冷/热一次风量
- 给水流量
给水温度
- 过量空气系数
- ⋮

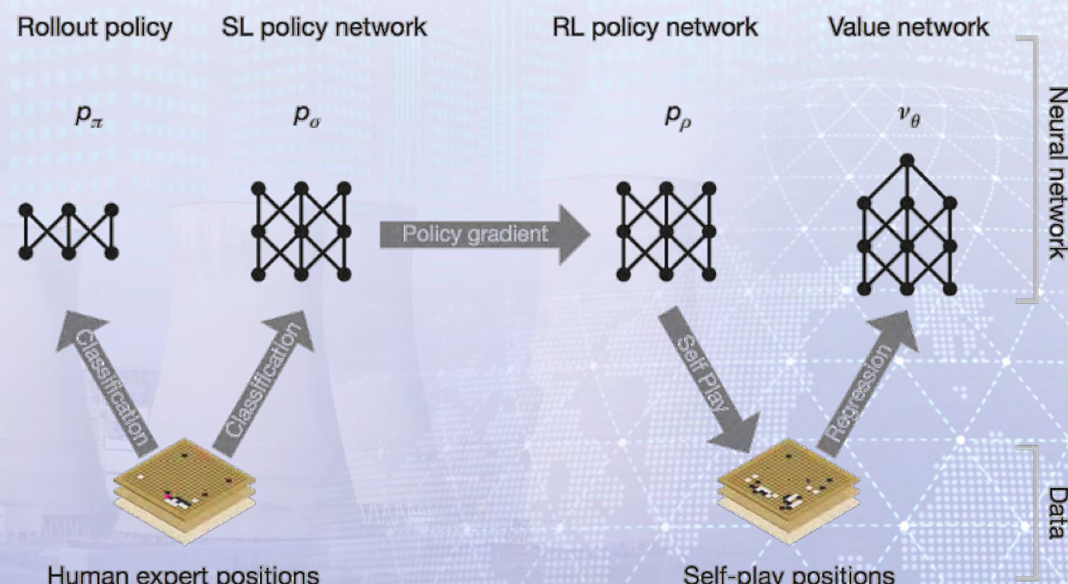


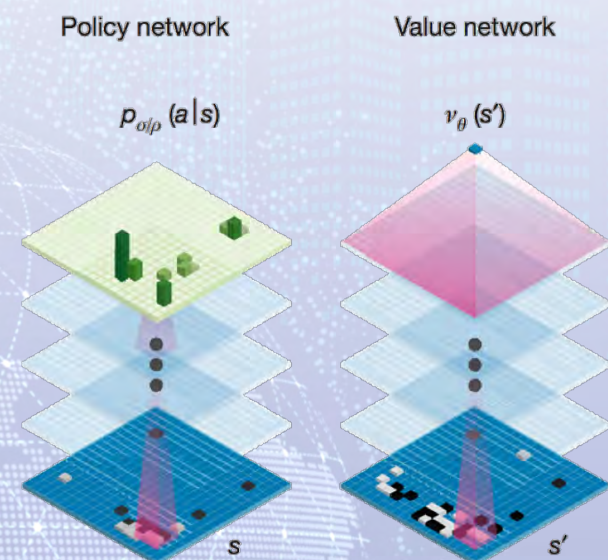
- 给煤机给煤量
- 冷/热风阀门
混合风阀门
- 减温水调节阀
- 二次风C，F挡板开度
- 燃尽风箱风门开度
- 一次风机导叶位置
送风机导叶位置
引风机导叶位置
- 过热器烟气挡板
- ⋮

深度强化学习优化

京东城市
JD iCity

AlphaGo





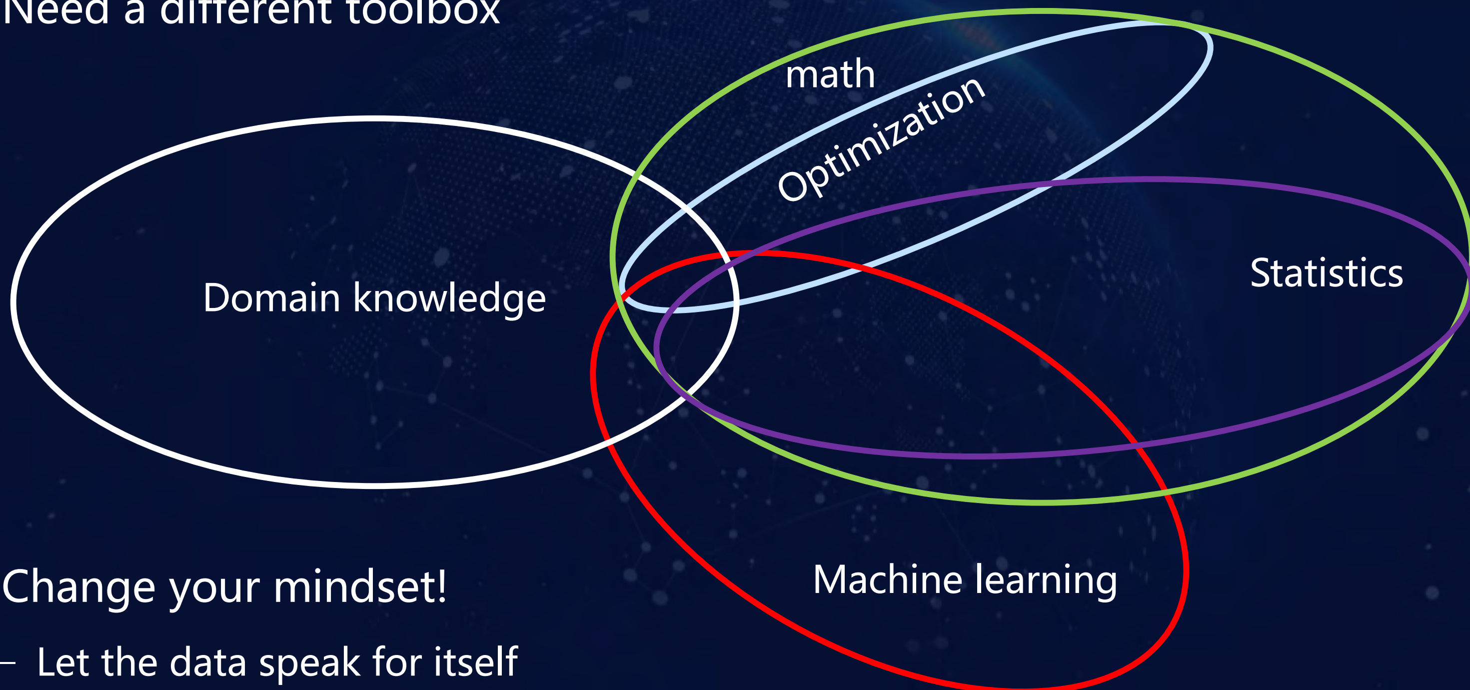

No.319

Take away message

- No universal approach
  - Highly dependent on data / problem / model
- Know your data
  - Develop highly customized models based on the property of the data
  - Important in model development!
- Incorporate domain knowledge if possible
  - Exploit structural properties in the data
  - Helpful especially when data is limited
- Get rid of assumptions!!!!!
  - Assumptions are bad!
  - If any, all assumptions should come from data

Take away message

- Need a different toolbox



- Change your mindset!
  - Let the data speak for itself
  - Stay away from traditional way of thinking

What to expect

- New and hot area, lots of opportunities
  - Many open questions
  - Solve problems that are not solvable before!
  - Many things to be done, especially for traditional engineering fields
- Conduct high impact research
  - Work on really interesting problems
- Develop something really useful
  - Build practical, deployable real-time applications
  - Papers should not be the final outcome

京东城市
JD iCity

Q/A

詹仙园 博士
zhanxianyuan@jd.com

京东城市
http://icity.jd.com

We Are Hiring!