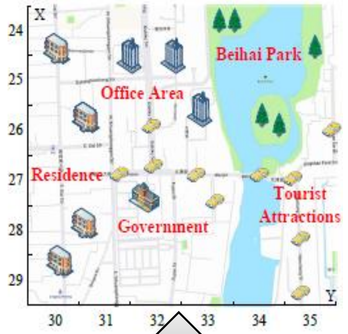
Big Data Intelligence: Challenges and Our Solutions

Tianrui LI
Southwest Jiaotong University, China
trli@swjtu.edu.cn

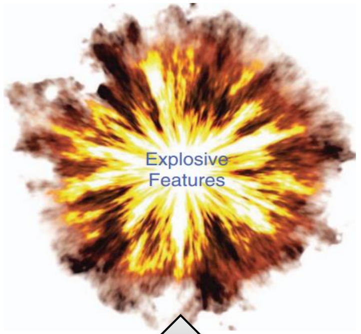
Big Data Intelligence: Challenges

Few Labeled Data



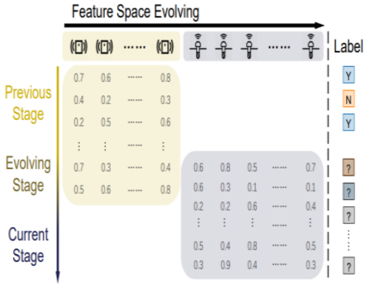
User Privacy

Explosive Features

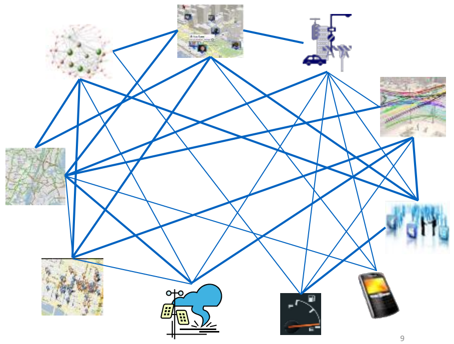


Big Data

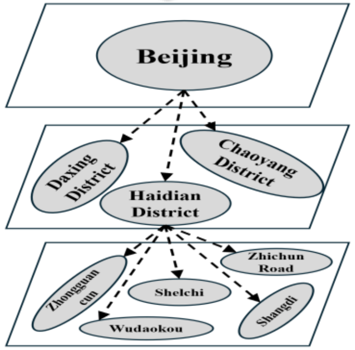
Open-world Dynamics



Multi-Source Heterogeneous



Multi-granularity



Problem 1: Few Labeled Data

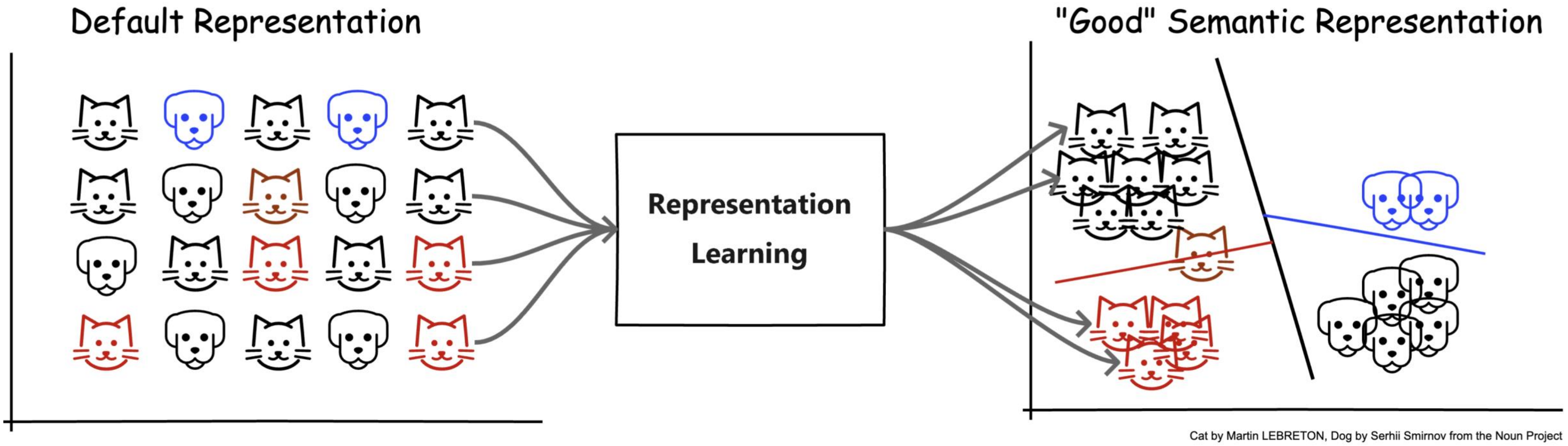
Micro-Supervised Disturbance Learning

Joint work with

Jielei Chu, Jing Liu, Hongjun Wang, Hua Meng, Zhiguo Gong, Tianrui Li. ***Micro-supervised disturbance learning: A perspective of representation probability distribution***. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.

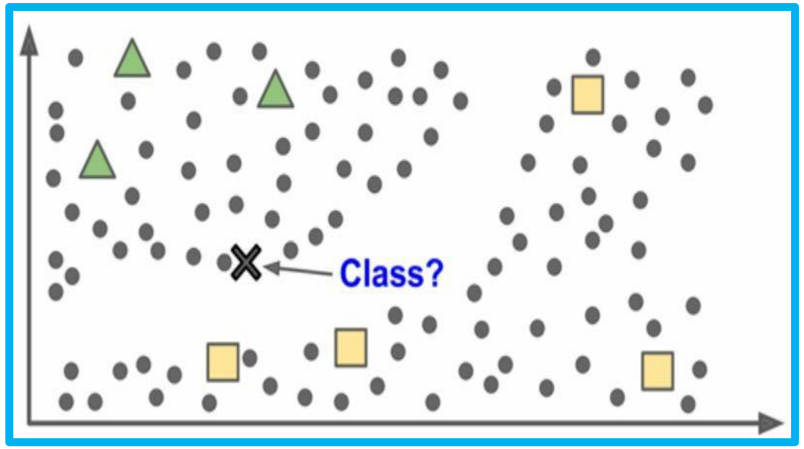
Learning Expressive Representations

- **Learning expressive representations** is a fundamental problem in machine learning area



Learning Expressive Representations

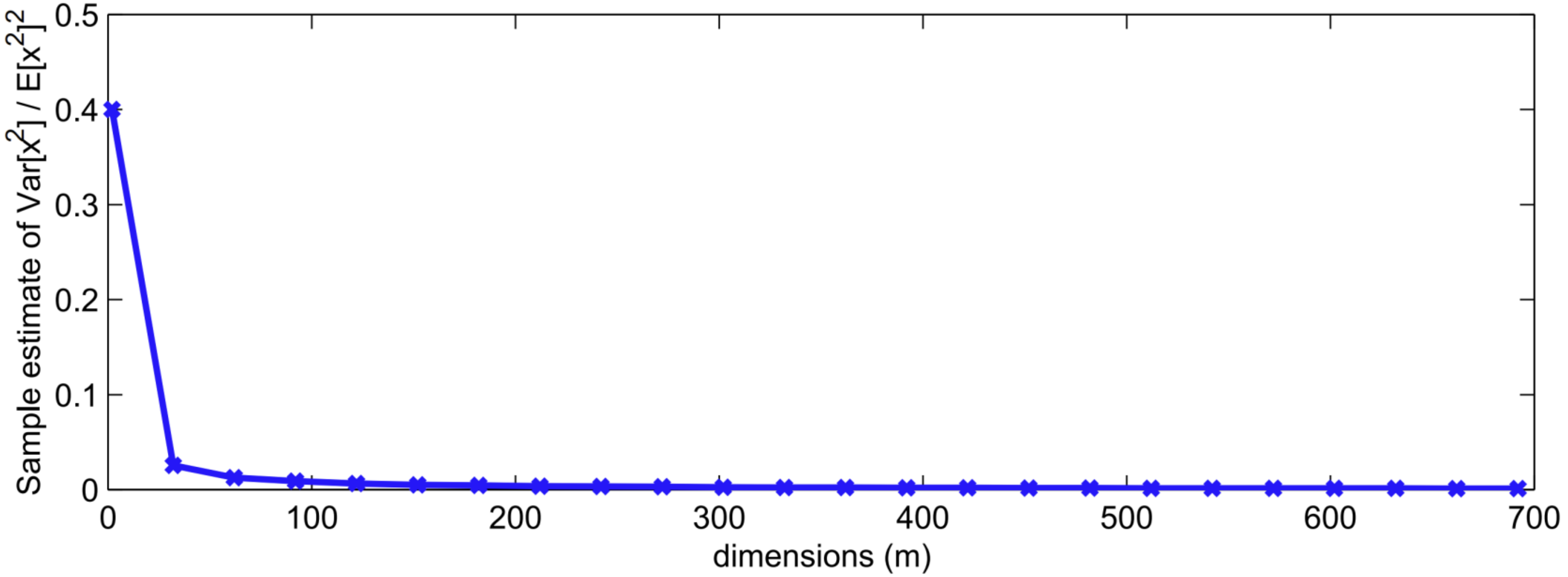
- **Semi-supervised learning** refers to a learning problem that involves a small portion of labeled examples and a large number of unlabeled examples from which a model must learn and make predictions on new examples.
- The **scarcity and high cost of labels** prompt us to explore more expressive representation learning methods which depends on **as few labels as possible**.



Semi-supervised learning

Learning Expressive Representations

- **Euclidean distance** is useful in semi-supervised models. However, these models show somewhat instabilities called **distance concentration phenomenon**.
  - As the data dimensionality increases, all the pairwise distances (dissimilarities) may converge to the same value.



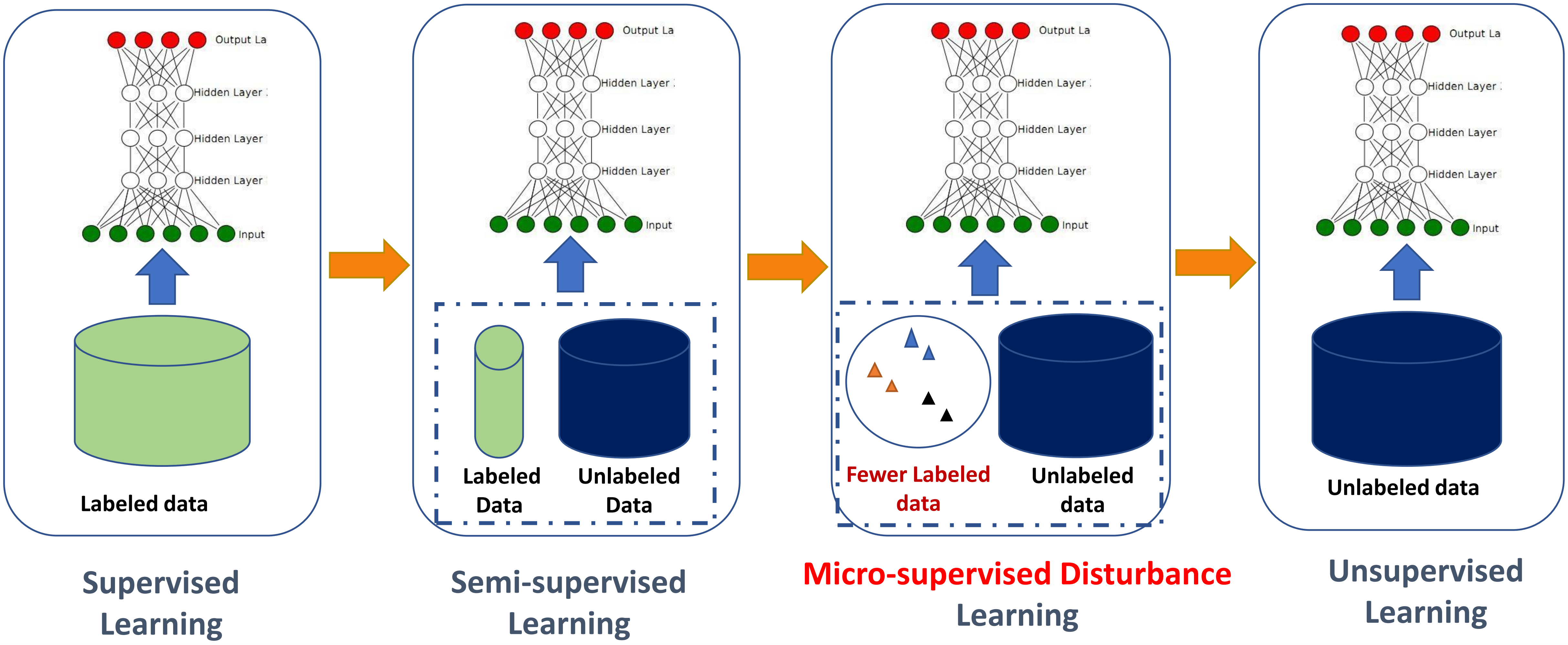
Micro-Supervised Disturbance Learning

- The motivation derives from the **small-perturbation ideology of physical systems.**
  - ➢ The original state of the physical system changes slightly under the stimulation of small disturbance.
- Two interesting problems:
  - ➢ Whether the **small disturbance** can be used to **stimulate** the representation learning model to fine-tune the expected representation probability distribution?
  - ➢ Whether the representation learning capability can significantly improve under the **continuous stimulation** of small disturbance?

Micro-Supervised Disturbance Learning

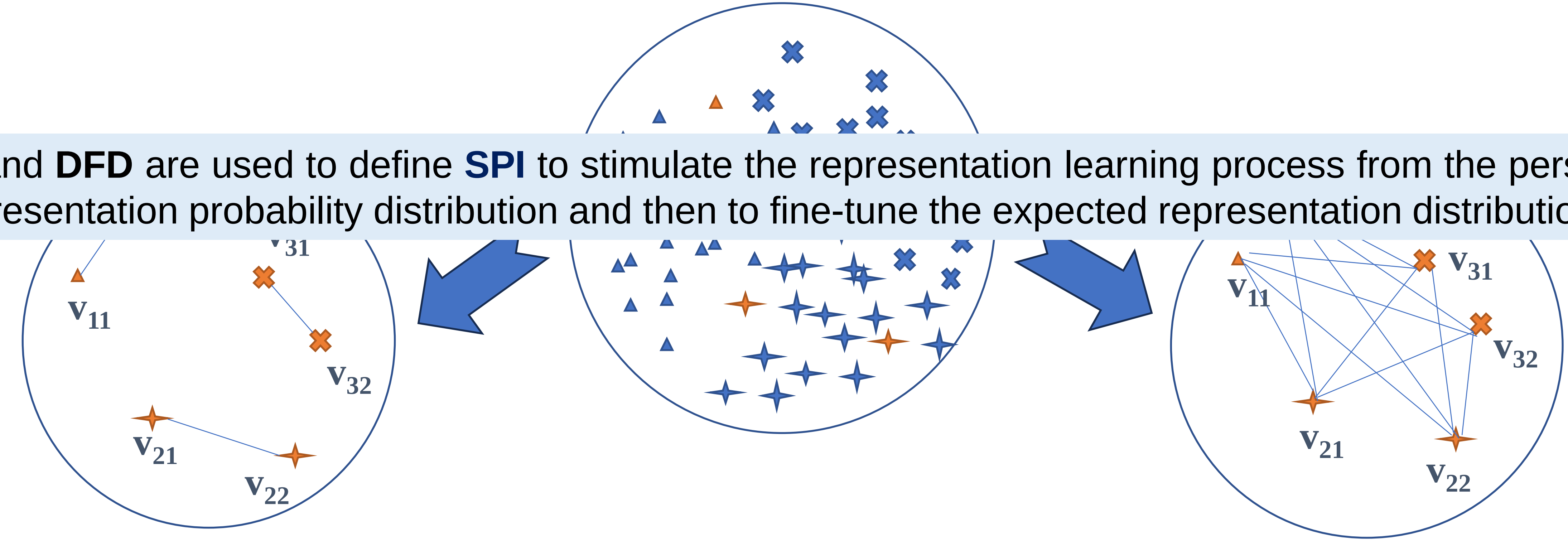
- To achieve these goals, the **small-perturbation information (SPI)** is used to stimulate the representation learning process from the perspective of representation probability distribution.
- Two variant models are proposed to fine-tune the expected representation distribution of RBM.
  - Micro-supervised Disturbance Gaussian-binary RBM (Micro-DGRBM).
  - Micro-supervised Disturbance RBM (Micro-DRBM) models.
- The **SPI** only depends on **two labels of each cluster**. Hence, we term this learning pattern as **Micro-supervised Disturbance Learning (Micro-DL)**.

Micro-Supervised Disturbance Learning



Micro-Supervised Disturbance Learning

SFD and **DFD** are used to define **SPI** to stimulate the representation learning process from the perspective of representation probability distribution and then to fine-tune the expected representation distribution.



Similar feature distribution (SFD) set

SFD=$\{(h_f, h_g) | p(h_f|v_f)$ and $p(h_g|v_g)$ are similar$\}$

v_f and v_g are in the same cluster.

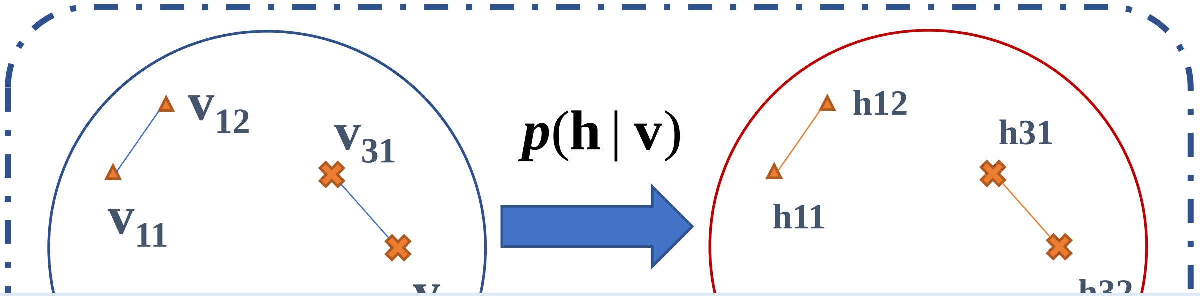
Dissimilar feature distributions (DFD) set

DFD=$\{(h_r, h_s) | p(h_r|v_r)$ and $p(h_s|v_s)$ are dissimilar$\}$

v_r and v_s are in the different cluster.

Micro-Supervised Disturbance Learning

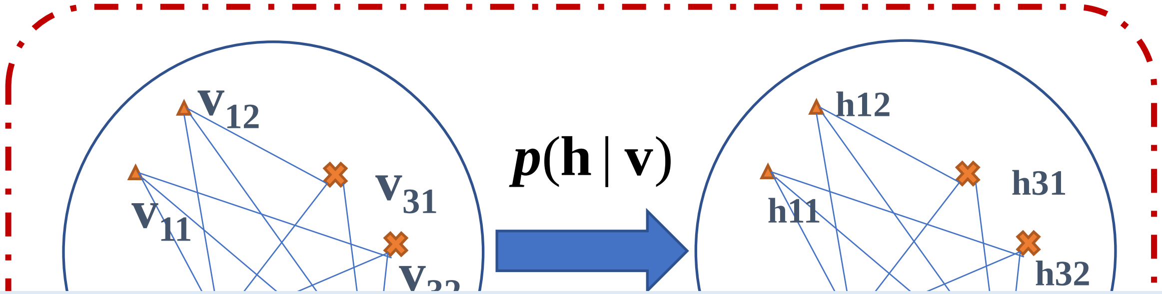
Similar feature distribution (SFD) set



The KL divergence of SPI is **minimized** in the **same** cluster to force the representation probability distributions to become more similar in Contrastive Divergence (CD) learning.

$$\min\{\sum_{\text{SFD}} \mathrm{KL}(p(\mathbf{h}_{ij} \mid \mathbf{v}_{ij}) \parallel p(\mathbf{h}_{i'j'} \mid \mathbf{v}_{i'j'}))\}$$

Dissimilar feature distributions (DFD) set

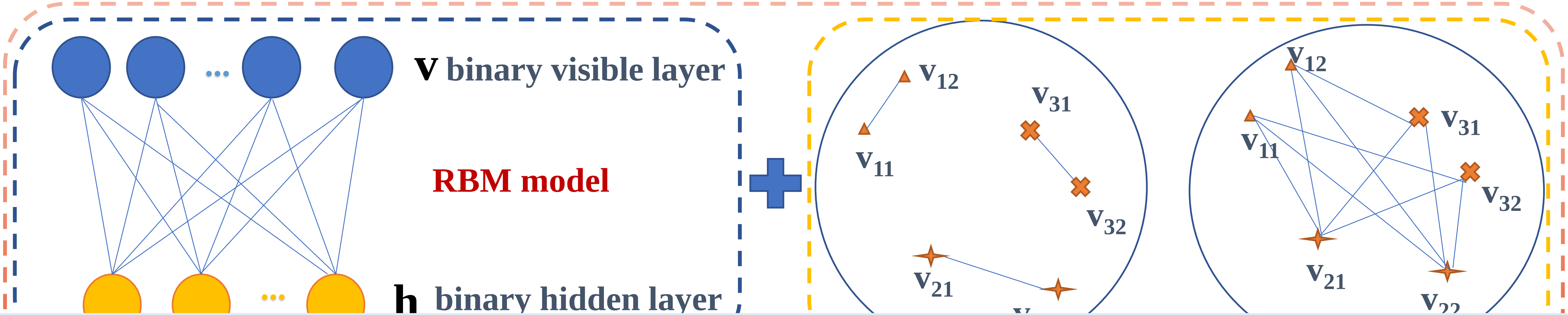


The KL divergence of SPI is **maximized** in the **different** clusters to force the representation probability distributions to become more dissimilar in CD learning.

$$\max\{\sum_{\text{DFD}} \mathrm{KL}(\boldsymbol{p}(\mathbf{h}_{\boldsymbol{pq}} \mid \mathbf{v}_{\boldsymbol{pq}}) \parallel \boldsymbol{p}(\mathbf{h}_{\boldsymbol{p'q'}} \mid \mathbf{v}_{\boldsymbol{p'q'}}))\}$$

The Kullback-Leibler (KL) divergence is used to measure the difference between two probability distributions.

Micro-supervised Disturbance RBM (Micro-DRBM)



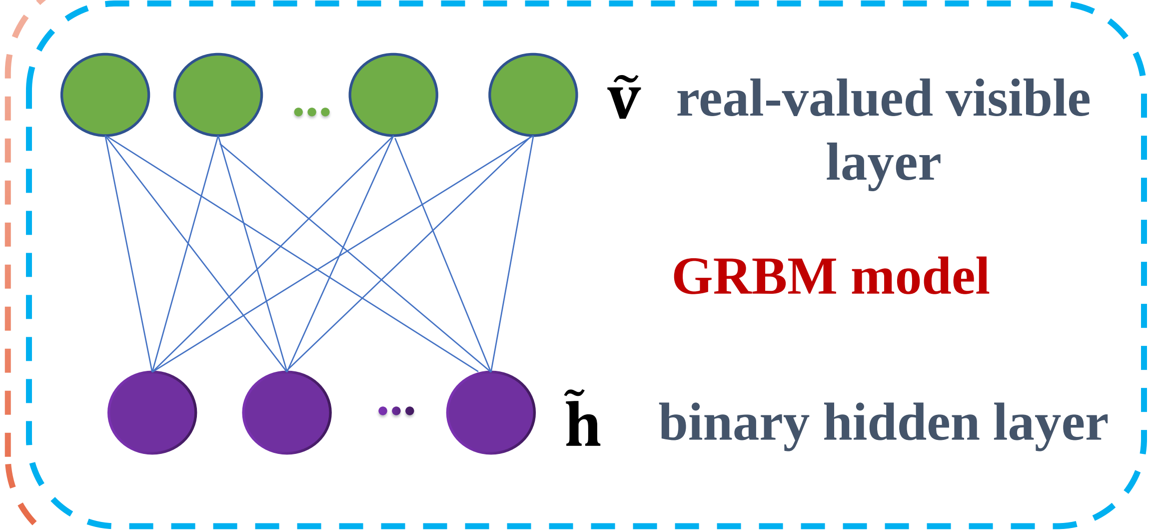
Under the stimulation of these Micro-supervised Disturbance, we expect that the representation probability distributions become more similar and dissimilar in the same and different clusters, respectively.

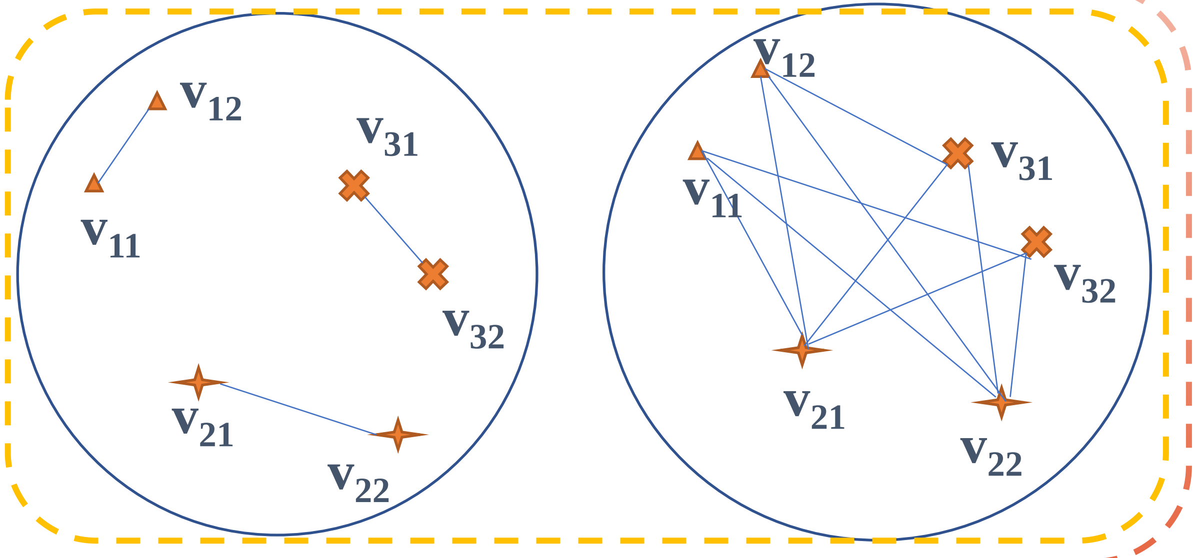
$$\min_{\theta}\Big\{-(1-\alpha)\Big(\mathbf{KL}(p_0 \,\|\, p_\infty)-\mathbf{KL}(p_1 \,\|\, p_\infty)\Big) + \alpha\Big[\frac{1}{K_S}\sum_{\mathcal{SFD}}\mathbf{KL}(P(\mathbf{h}_f|\mathbf{v}_f) \,\|\, P(\mathbf{h}_g|\mathbf{v}_g)) - \frac{1}{K_D}\sum_{\mathcal{DFD}}\mathbf{KL}(P(\mathbf{h}_r|\mathbf{v}_r) \,\|\, P(\mathbf{h}_s|\mathbf{v}_s))\Big]\Big\},$$

Contrastive Divergence (CD) Learning

Micro-supervised Disturbance

Micro-supervised Disturbance Gaussian-binary RBM (Micro-DGRBM)

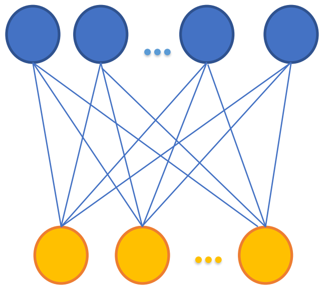




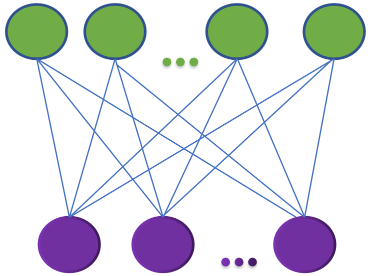
$$\min_{\tilde{\theta}}\Big\{-(1-\alpha)\Big(\mathbf{KL}(\widetilde{p_0}\,\|\,\widetilde{p_\infty})-\mathbf{KL}(\widetilde{p_1}\,\|\,\widetilde{p_\infty})\Big)+\alpha\Big[\frac{1}{\widetilde{K_S}}\sum_{\widetilde{\mathcal{SFD}}}\mathbf{KL}\big(P(\widetilde{\mathbf{h}_f}|\widetilde{\mathbf{v}_f})\,\|\,P(\widetilde{\mathbf{h}_g}|\widetilde{\mathbf{v}_g})\big)-\frac{1}{\widetilde{K_D}}\sum_{\widetilde{\mathcal{DFD}}}\mathbf{KL}\big(P(\widetilde{\mathbf{h}_r}|\widetilde{\mathbf{v}_r})\,\|\,P(\widetilde{\mathbf{h}_s}|\widetilde{\mathbf{v}_s})\big)\Big]\Big\}.$$

Contrastive Divergence (CD) Learning

Micro-supervised Disturbance

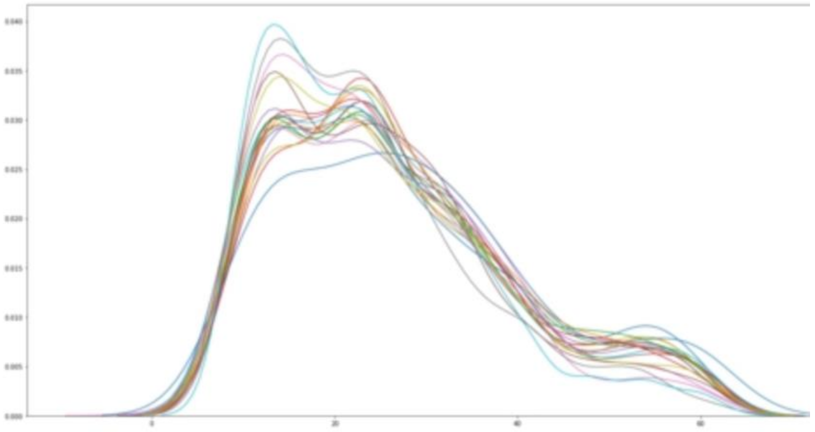


Micro-DRBM



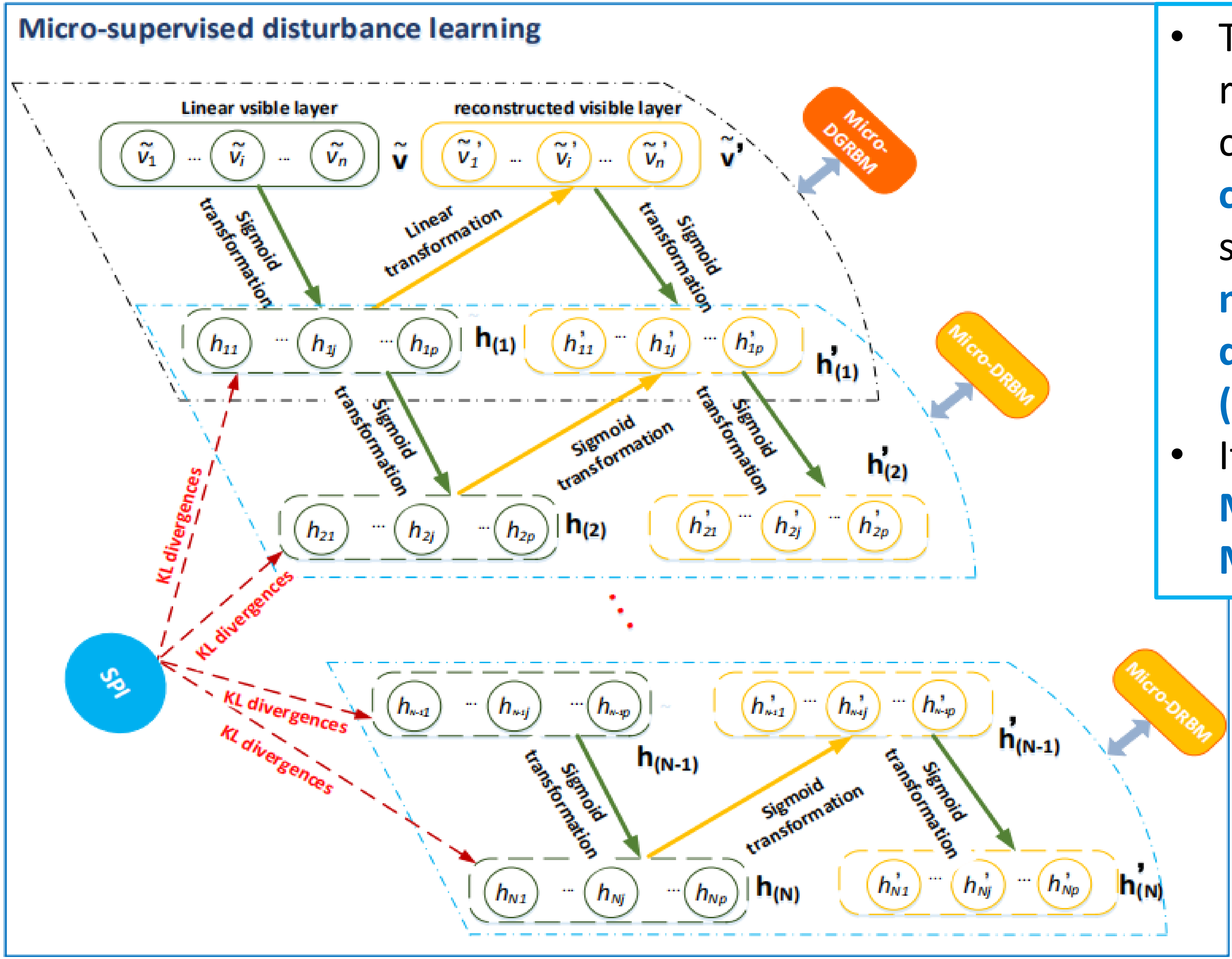
Micro-DGRBM

Representation Probability Distribution



Shallow models have litimted representation capability!

Whether **deep framework** has better representation capability under the **continuous stimulation** of small-perturbation information?



- To explore the representation learning capability under the **continuous stimulation** of small disturbance, a deep **micro-supervised disturbance learning (Micro-DL)** is developed.
- It consists of a stack of **one Micro-DGRBM** and ***N*** **Micro-DRBMs**.

Evaluation

- Our Micro-DL model outperforms state-of-the-art baseline models

Performance Comparisions (Rank) of Benchmarking Algorithms (Semi-SP), Shallow Models (pcGRBM and Semi-EAGR), and Deep Models (Semi-MG, VGAE, NMicro-DL and Micro-DL)

Dataset	Semi-SP	pcGRBM	Semi-EAGR	Semi-MG	VGAE	NMicro-DL	Micro-DL	Total
aquarium	-0.1006 (65)	-0.1218 (71)	-0.0827 (58)	0.0997 (25)	0.0453 (34)	-0.0091 (45)	0.1695 (6)	304
bathroom	-0.0787 (56)	-0.1708 (80)	-0.1214 (69)	0.1601 (8)	0.0142 (39)	-0.0341 (49)	0.2310 (2)	303
blog	-0.0788 (57)	-0.1433 (75)	-0.1142 (67)	0.1175 (18)	0.0274 (37)	-0.0040 (43)	0.1951 (3)	300
cactus	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
voituretuning	-0.0900 (60)	-0.1215 (70)	-0.0577 (53)	0.0819 (27)	0.0565 (31)	-0.0163 (47)	0.1469 (11)	299
car	-0.1607 (79)	0.1283 (16)	-0.0858 (59)	0.0786 (28)	-0.1407 (74)	0.0512 (32)	0.1290 (15)	303
KDD99	-0.4220 (84)	0.1244 (17)	0.0211 (38)	0.1045 (23)	-0.1508 (78)	0.1460 (12)	0.1766 (4)	256
segmentation	-0.2102 (83)	0.0507 (33)	0.0999 (24)	-0.1175 (68)	0.0694 (30)	-0.0046 (44)	0.1125 (20)	302
vowe	-0.0385 (50)	-0.0615 (54)	0.0043 (41)	0.0343 (35)	-0.0003 (42)	-0.0093 (46)	0.0710 (29)	297
Total	799	713	679	303	458	505	**113**	3570
Average Rank	66.5833	59.4167	56.5833	25.2500	38.1667	42.0833	**9.4167**	

It means the representation learning capability of our Micro-DL architecture has significantly enhanced under the continuous stimulation of small-perturbation information (SPI).

Problem 2:
Data Privacy Protection

Data Privacy Protection in Daily Schedule Recommendation

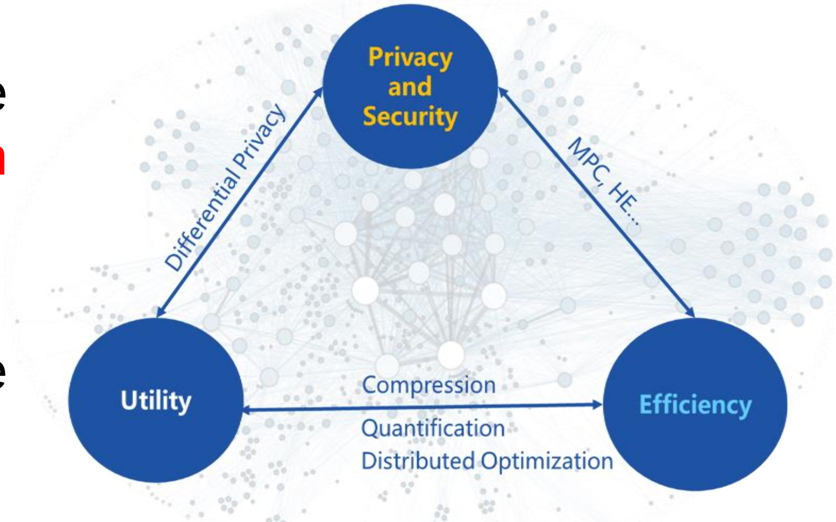
Joint work with Tencent 腾讯

Wei Huang, Jia Liu, Tianrui Li, Tianqiang Huang, Shenggong Ji, Jihong Wan. ***FedDSR: Daily schedule recommendation in a federated deep reinforcement learning framework***. IEEE Transactions on Knowledge and Data Engineering, 2023.

Data Privacy and Security

- Big Data inevitably involves the users' privacy
- Effective data privacy protection is very important for Big Data Intelligence

- In 2018, EU issued the General Data Protection Regulation (GDPR)
- Data privacy faces the Privacy-Efficiency-Utility Trilemma



Privacy-Efficiency-Utility Trilemma

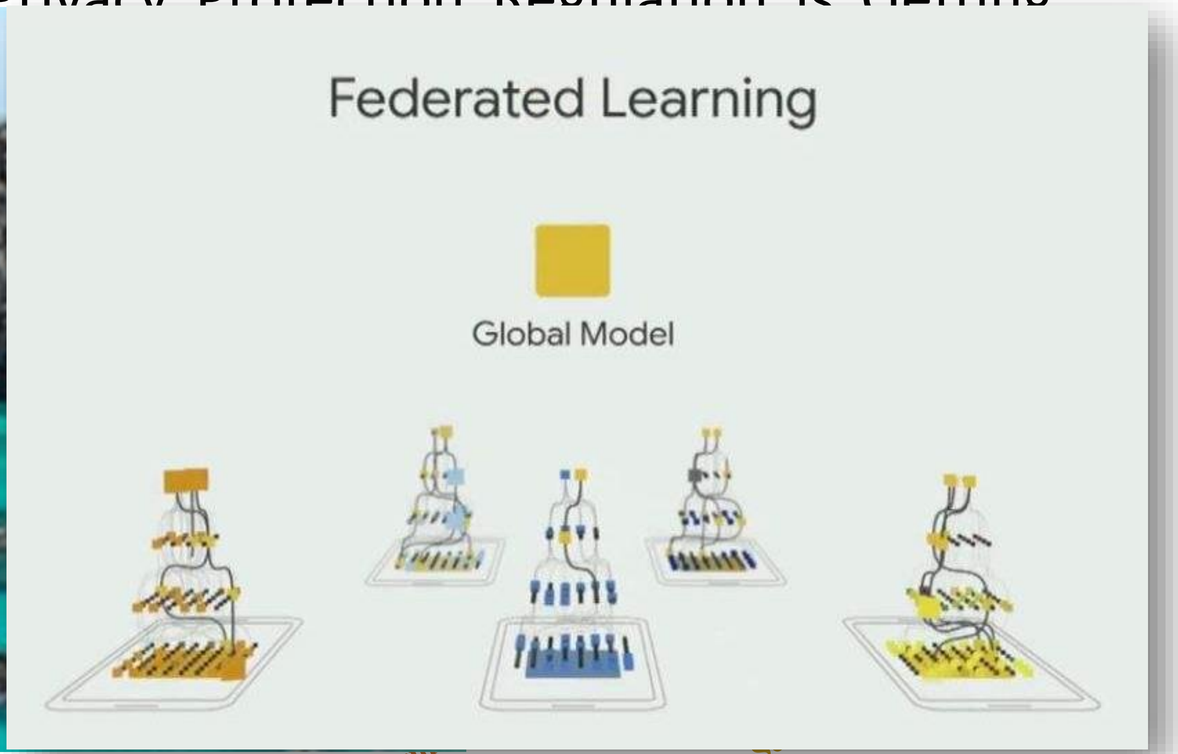
Data Privacy and Security

- Data Regulatory Legal System —— “Data Privacy Protection Regulation is Getting Tougher Around the


California Consumer Privacy Act (CCPA) Compliance Tips

California's Consumer Privacy Protection Act (CCPA) goes into effect on January 1, 2020. The CCPA is a broad-ranging legal framework regulating data and enhancing privacy rights for Californians. The law will change the way companies across the United States navigate data protection and privacy.

RICHT

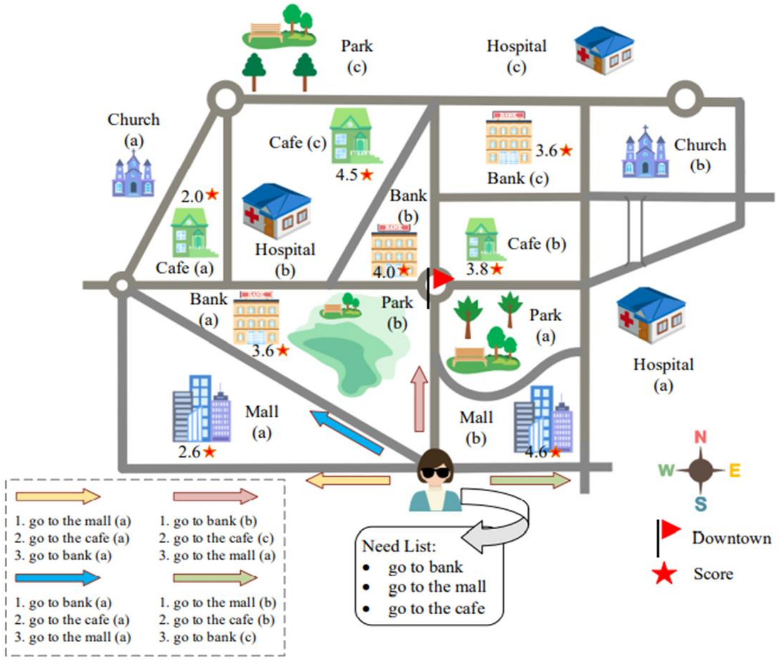


✓ CCPA in the United States California《California Consumer Privacy Act》

✓ China has successively introduced comprehensive and stringent regulations on data security protection.

Daily Schedule Recommendation

- The daily schedule recommendation is to arrange a reasonable sequence of activities and the location of activities

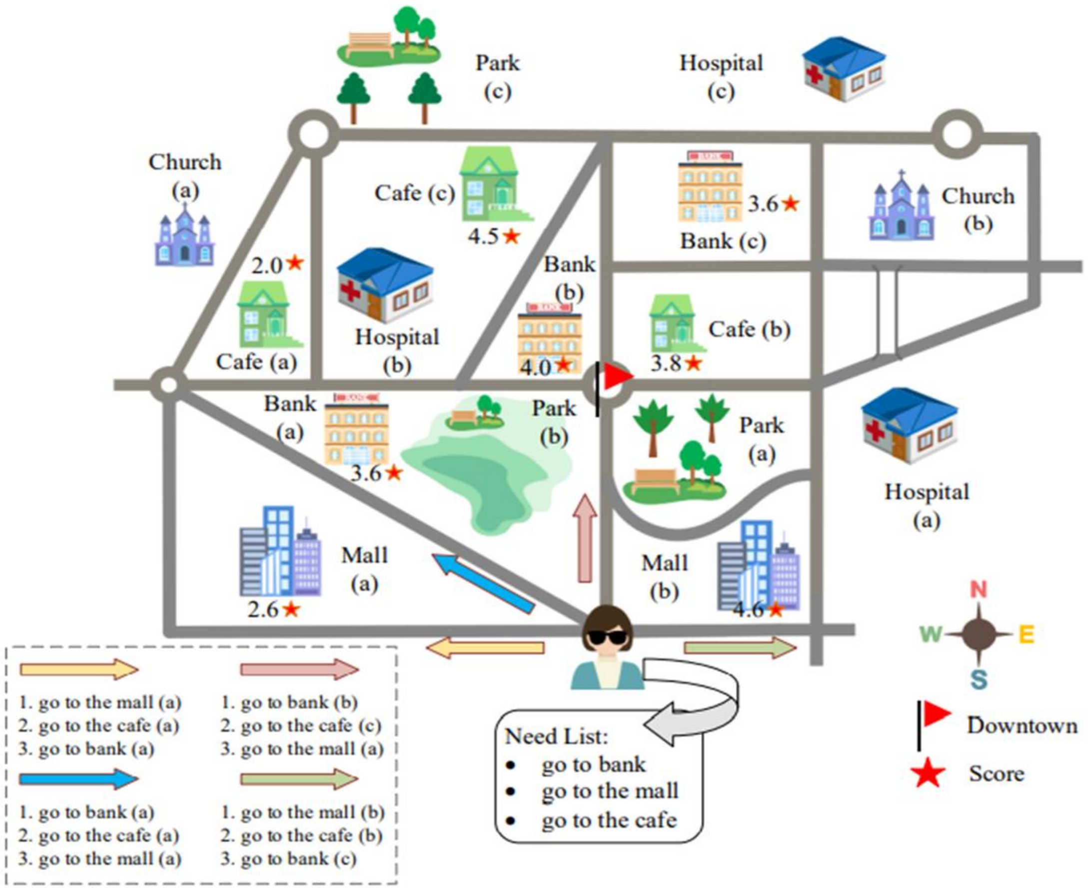


An example of user daily schedules

E.g., home address, behavioral habits

Daily Schedule Recommendation

- An example of user daily schedules



Problem

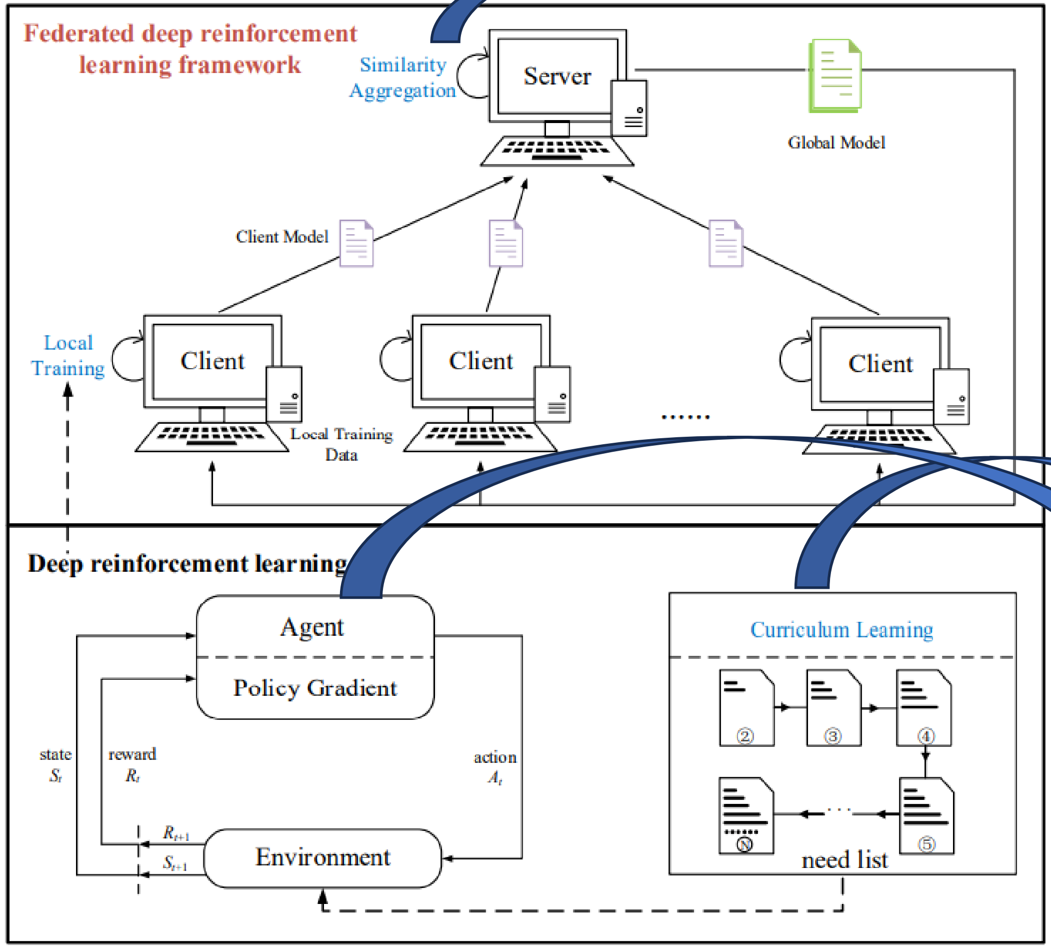
- How to recommend the order and location of activities that meet the user's needs while complying with the data security protection regulations?

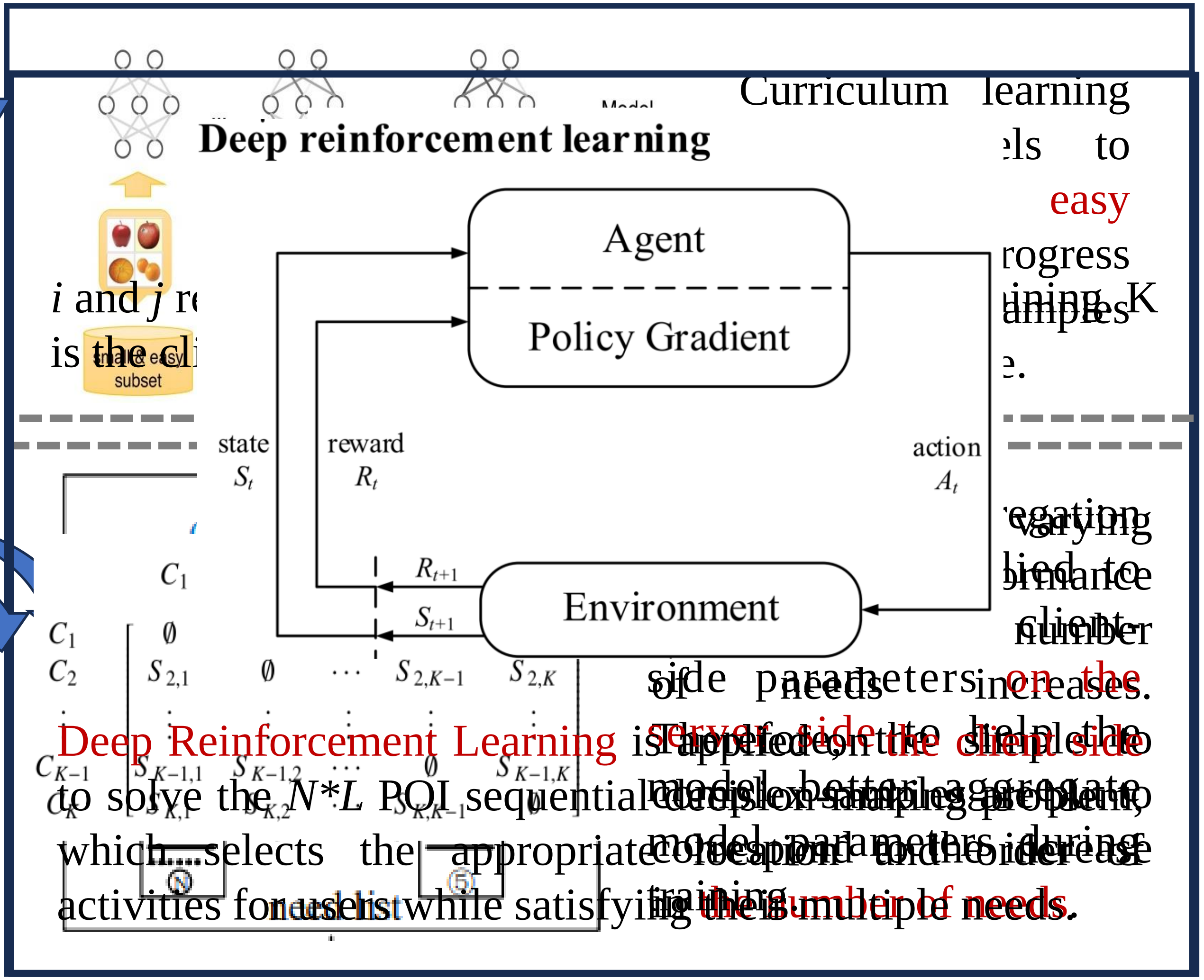
  Federated Learning + Deep Reinforcement Learning

analyze

- The longer the user's list of needs, the more complex the sequential decisions that deep reinforcement learning models need to be trained for. **Curriculum Learning**
- Users' location and their trajectory data are sensitive and private. **Federated Learning**

Architecture Design



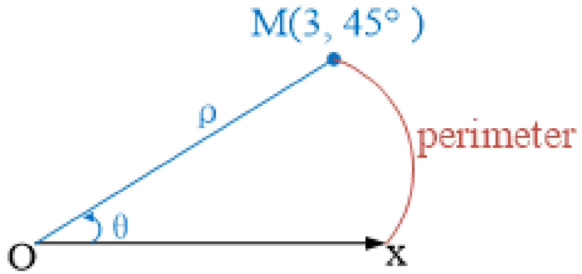


Evaluation

- The proposed FedDSR has better results in *distance*, *time*, and *score*, and the *perimeter* is lower than all the comparison algorithms

TABLE 3
Distribution of users' needs in the two datasets.

Dataset \ Need	2	3	4	5	6	7
Geolife	40.7%	26.8%	16.0%	9.8%	4.8%	1.9%
Chengdu	22.5%	26.6%	23.9%	18.8%	6.0%	2.2%



Method	*Chengdu*				*Geolife*			
	distance (km)	*time (min)*	*score (point)*	*perimeter*	*distance (km)*	*time (min)*	*score (point)*	*perimeter*
RS	4.12	35.51	2.37	117.36	1.81	20.83	2.21	73.03
DG	1.05	12.05	2.72	34.52	**0.78**	**8.02**	2.36	26.61
EG	1.14	12.06	2.94	31.22	1.44	13.90	2.79	38.60
SG	3.58	28.51	3.53	52.67	2.17	25.80	3.33	54.13
KNN-SD	2.97	24.90	2.56	76.35	1.39	13.22	3.20	29.90
MS	**1.02**	**11.28**	2.65	33.31	0.95	10.59	2.91	27.81
CDRL	1.87	15.53	3.77	24.00	1.26	13.19	3.43	26.02
FedDSR	2.06	16.02	**4.03**	**19.53**	0.82	9.58	**3.54**	**17.58**

Problem 3: Explosive Features

Scalable Feature Selection by Spark

Rough Hypercuboid Approach

Joint work with

Chuan Luo, Sizhao Wang, Tianrui Li, Hongmei Chen, Jiancheng Lv, Zhang Yi. ***Spark rough hypercuboid approach for scalable feature selection***. IEEE Transactions on Knowledge and Data Engineering, 2023.

The Curse of Dimensionality

- The explosive features brings new challenges to Big Data Intelligence

(c) ***LIBSVM DATABASE***

APPLICATION DOMAIN	DATA NAME	DIMENSION	YEAR
IMAGE	USPS	256	1994
	GISETTE	5,000	2003
LIFE SCIENCE	LEUKEMIA	7,129	1999
	COLON-CANCER	2,000	1999
	BREAST-CANCER	7,129	2001
TEXT	NEWS20	62,061	1995
	REAL-SIM	20,958	1998
	SECTOR	55,197	1998
	RCV1	47,236	2004
	NEWS20.BINARY	**1,355,191**	2005
	WEBSPAM	**16,609,143**	2006
	SIAM	30,438	2007
	LOG1P	**4,272,227**	2009
EDUCATION	KDD2010	**29,890,095**	2010

Data volume presents an immediate challenge pertaining to the ***scalability issue***.

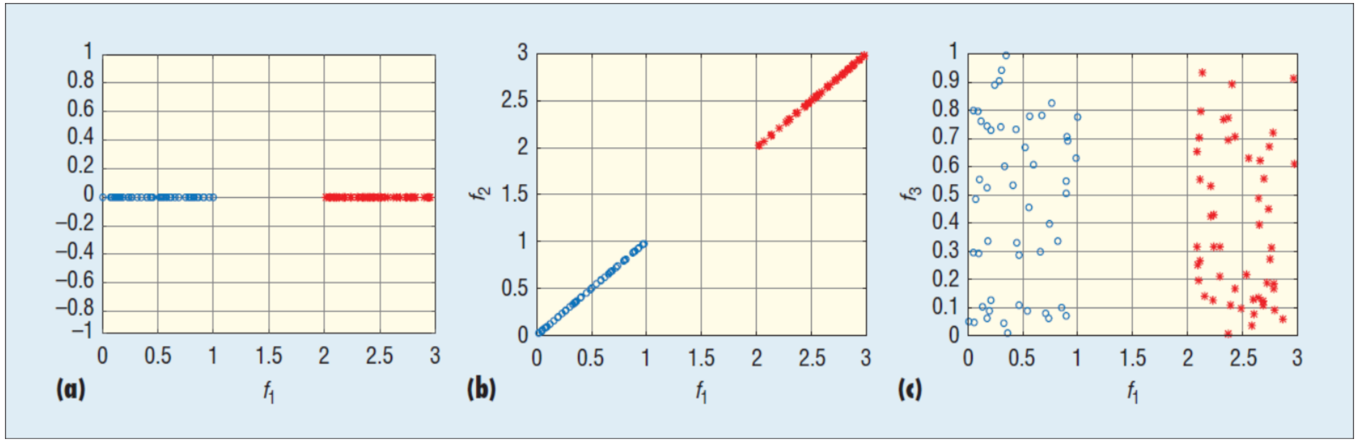
A single machine can no longer process or even store all the data. Only solution is to ***distribute data over large clusters***.

Most algorithms are **serial-computing** implementations and still struggle when processing large-scale datasets due to the **limited computational and storage resources**.

Feature Selection

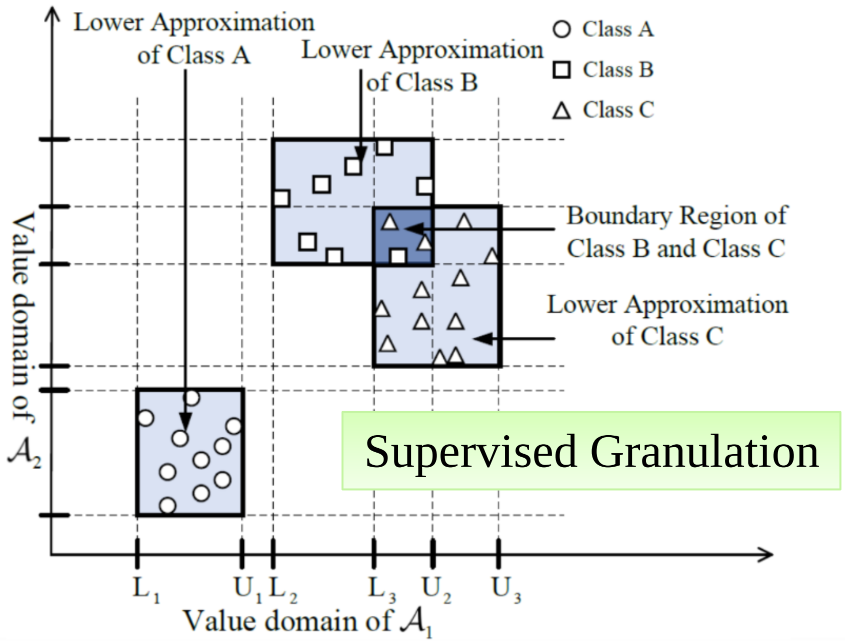
- Real-world data contains a lot of irrelevant, redundant and noisy features
- Feature selection
  - Preparing clean understandable data
  - Building more compact and efficient models
  - Improving data mining performance

Examples of (a) relevant, (b) redundant, and (c) irrelevant features



J.D. Li, et al, Feature Selection: A Data Perspective. ACM Comput. Surv. 50(6): 94:1-94:45 (2018)

- Rough hypercuboids approach
  - Integrating the merits of rough sets and hypercuboid learning
  - Hybrid objective function to measure discriminating ability of features



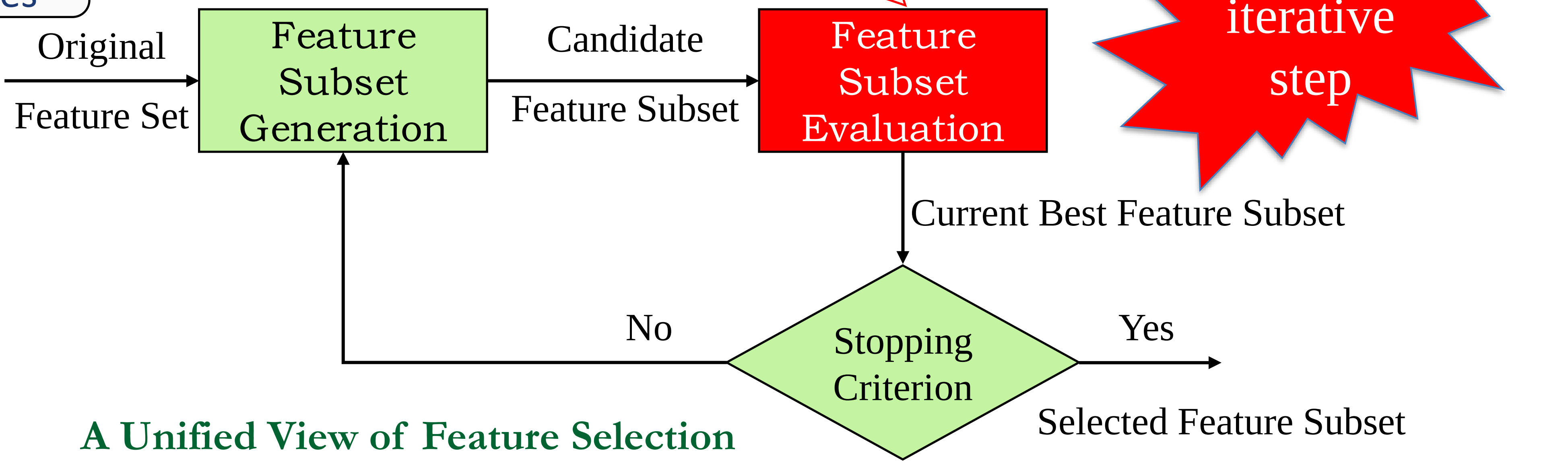
Parallel Optimization

Relevance between features and labels

Dependency of labels on the feature subset

$$J = \varpi \frac{1}{|\boldsymbol{S}|}\sum_{A_k\in \boldsymbol{S}}\gamma_{A_k}(\boldsymbol{D}) + \lambda(1-\varpi)\gamma_{\boldsymbol{S}}(\boldsymbol{D})$$
$$+(1-\lambda)(1-\varpi)\frac{\sum_{A_k\neq A_l\in \boldsymbol{S}}\{\sigma_{\{A_k,A_l\}}(\boldsymbol{D},A_k)+\sigma_{\{A_k,A_l\}}(\boldsymbol{D},A_l)\}}{|\boldsymbol{S}|(|\boldsymbol{S}|-1)}$$

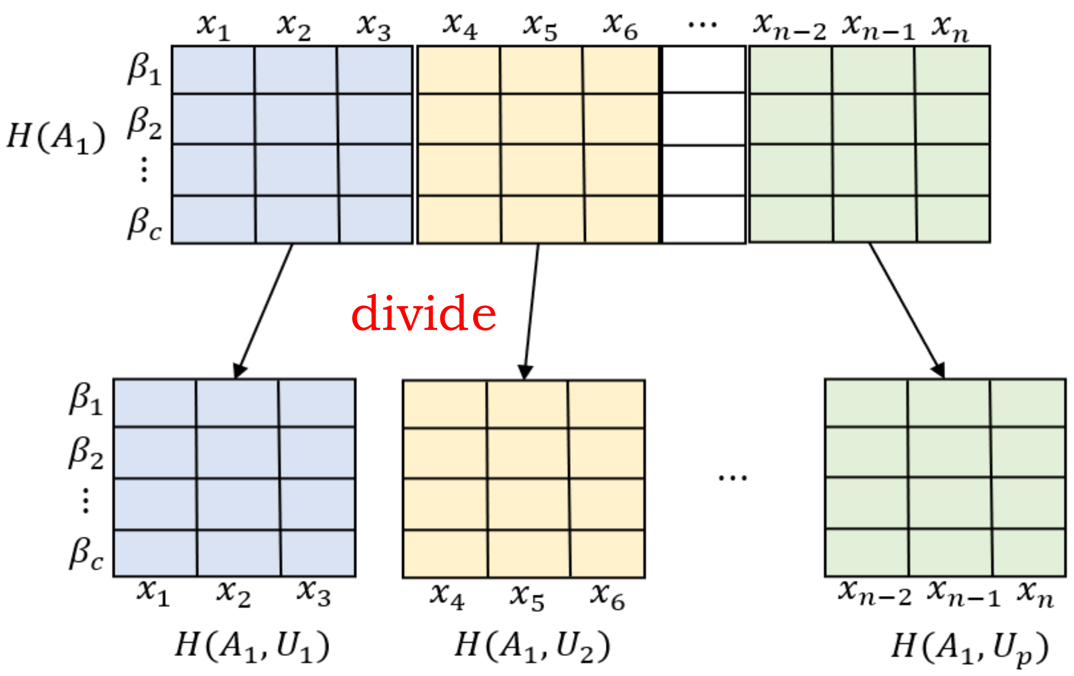
Significance among selected features



A Unified View of Feature Selection

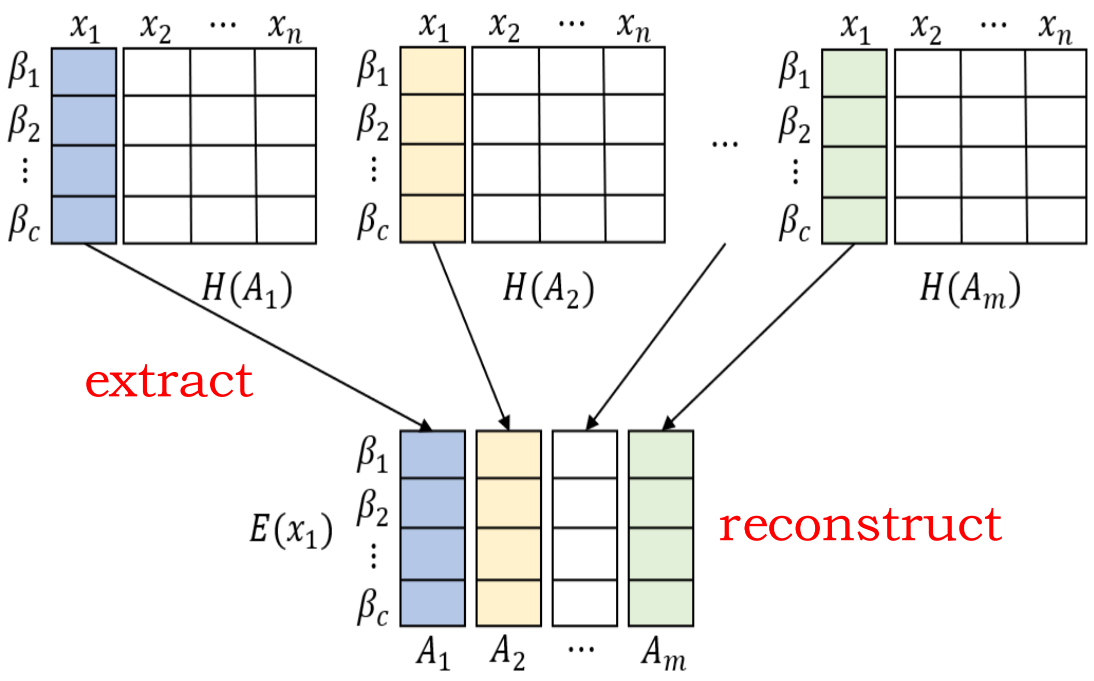
Data Parallelism Strategies

- Vertical partitioning
  - Partitions data along feature space



Hypercuboid matrix of feature

- Horizontal partitioning
  - Partitions data along sample space



Hypercuboid matrix of object

Horizontal Partitioning Oriented Parallelizations

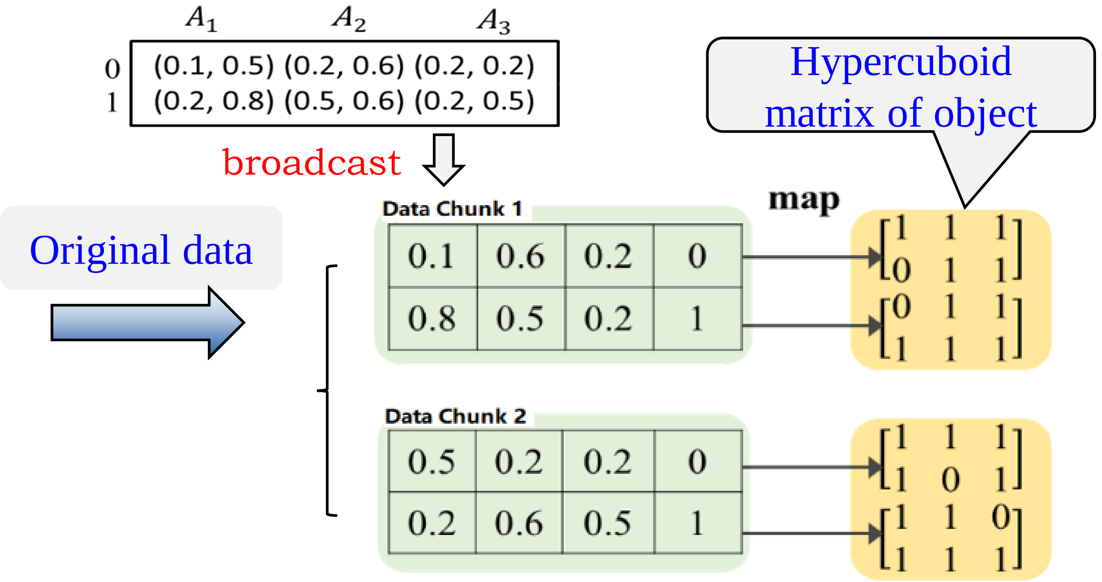
- Data parallelism

Feature

$$E(x_k) = \begin{bmatrix} e_{11}(x_k) & e_{12}(x_k) & \cdots & e_{1m}(x_k) \\ e_{21}(x_k) & e_{22}(x_k) & \cdots & e_{2m}(x_k) \\ \cdots & \cdots & \cdots & \cdots \\ e_{c1}(x_k) & e_{c2}(x_k) & \cdots & e_{cm}(x_k) \end{bmatrix}$$

Object

$$e_{ij}(x_k) = \begin{cases} 1, & \mathrm{L}_i \le f(x_k, A_j) \le \mathrm{U}_i, \\ 0, & \text{otherwise.} \end{cases}$$



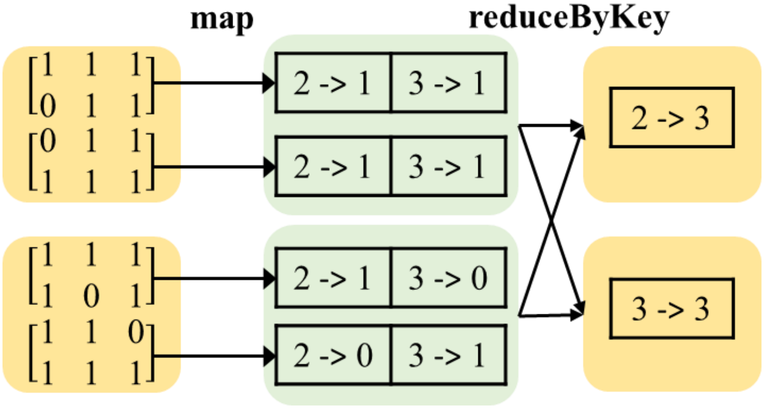
- Task parallelism

Dependency of feature

$$\gamma_{S\cup\{A_j\}}(D) = \frac{1}{n}\sum_{k=1}^{n}\left\{1 - \min\{1, \sum_{i=1}^{c} \cap_{A_l \in S\cup\{A_j\}} e_{il}(x_k) - 1\}\right\}$$

reduceByKey
Cumulative Sum

map
Membership value of an object



Vertical Partitioning Oriented Parallelizations

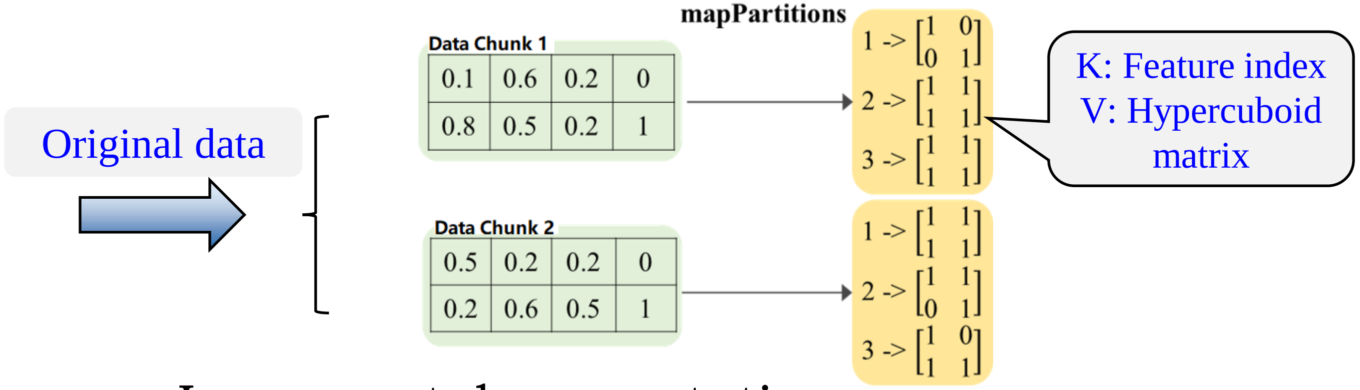
- Data parallelism

Feature

$$H(A_k, U_p) = \begin{bmatrix} h_{11}(A_k, U_p) & h_{12}(A_k, U_p) & \cdots & h_{1|U_p|}(A_k, U_p) \\ h_{21}(A_k, U_p) & h_{22}(A_k, U_p) & \cdots & h_{2|U_p|}(A_k, U_p) \\ \cdots & \cdots & \cdots & \cdots \\ h_{c1}(A_k, U_p) & h_{c2}(A_k, U_p) & \cdots & h_{c|U_p|}(A_k, U_p) \end{bmatrix}$$

Partition

$$h_{ij}(A_k, U_p) = \begin{cases} 1, & \mathrm{L}_i \le f(x_j, A_k) \le \mathrm{U}_i \\ 0, & \text{otherwise.} \end{cases}$$



- Task parallelism

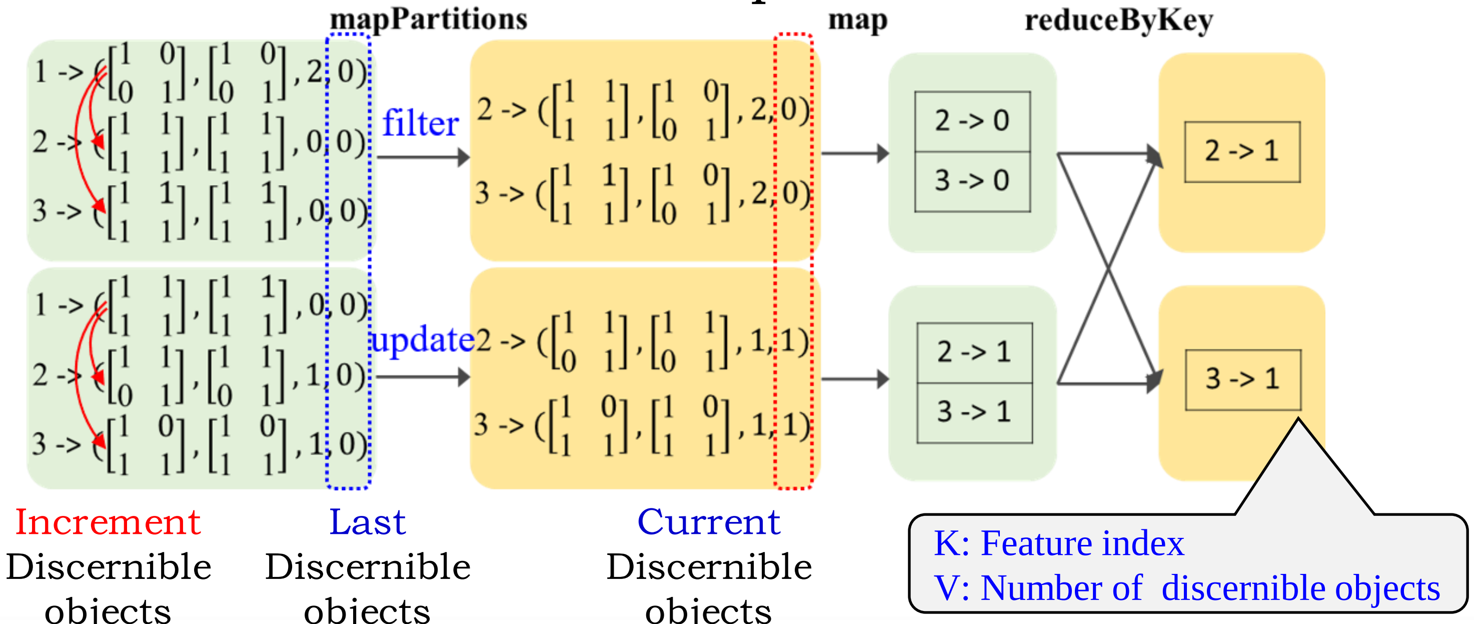
Current iteration value

$$J_{avgsig}(A_k, S) = \frac{1}{|S|}\{(|S|-1)J_{avgsig}(A_k, S-\{A_s\}) + \sigma_{\{A_k, A_s\}}(D, A_k)\}$$

Last iteration value

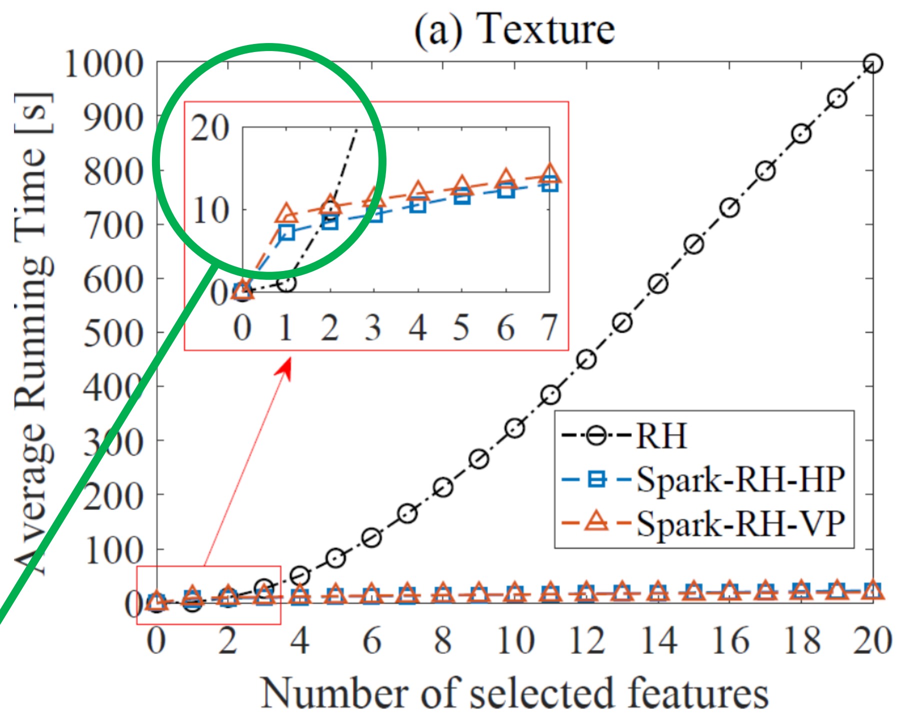
Significance of the new feature

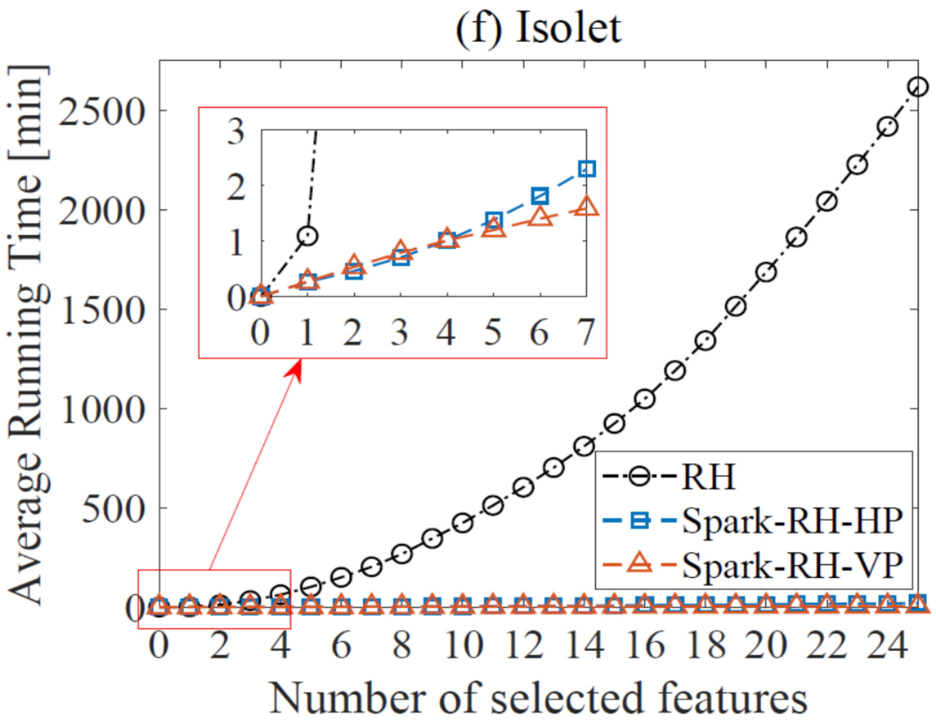
Incremental computation



Evaluation

- Both horizontal and vertical parallelizations can produce selected features in very less time compared to the standard sequential algorithms





A fixed initialization time is required in the beginning of distributed program deployment in Spark cluster.

Problem 4: Open-world Dynamics

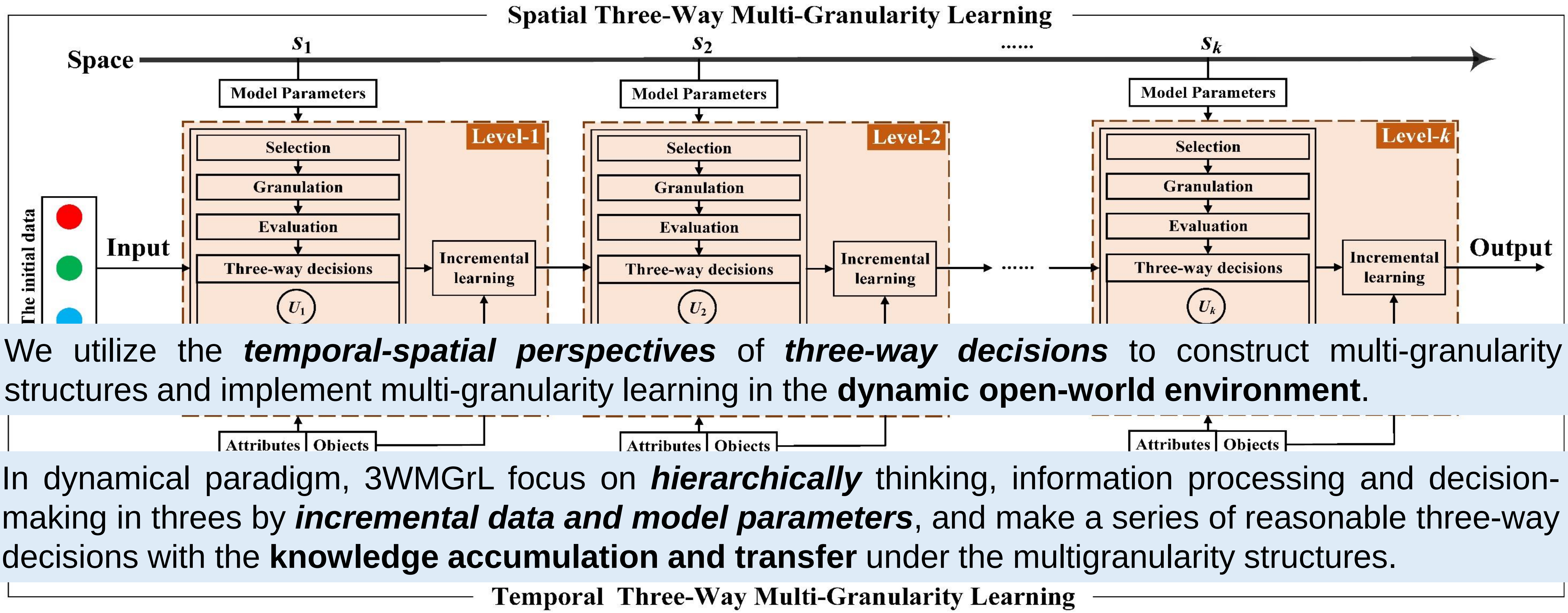
Three-way multi-granularity learning (3WMGrL)

for Dynamic Fuzzy Environment

Joint work with

Xin Yang, Yujie Li, Dun Liu, Tianrui Li. ***Hierarchical fuzzy rough approximations with three-way multigranularity learning***. IEEE Transactions on Fuzzy Systems, 2022.

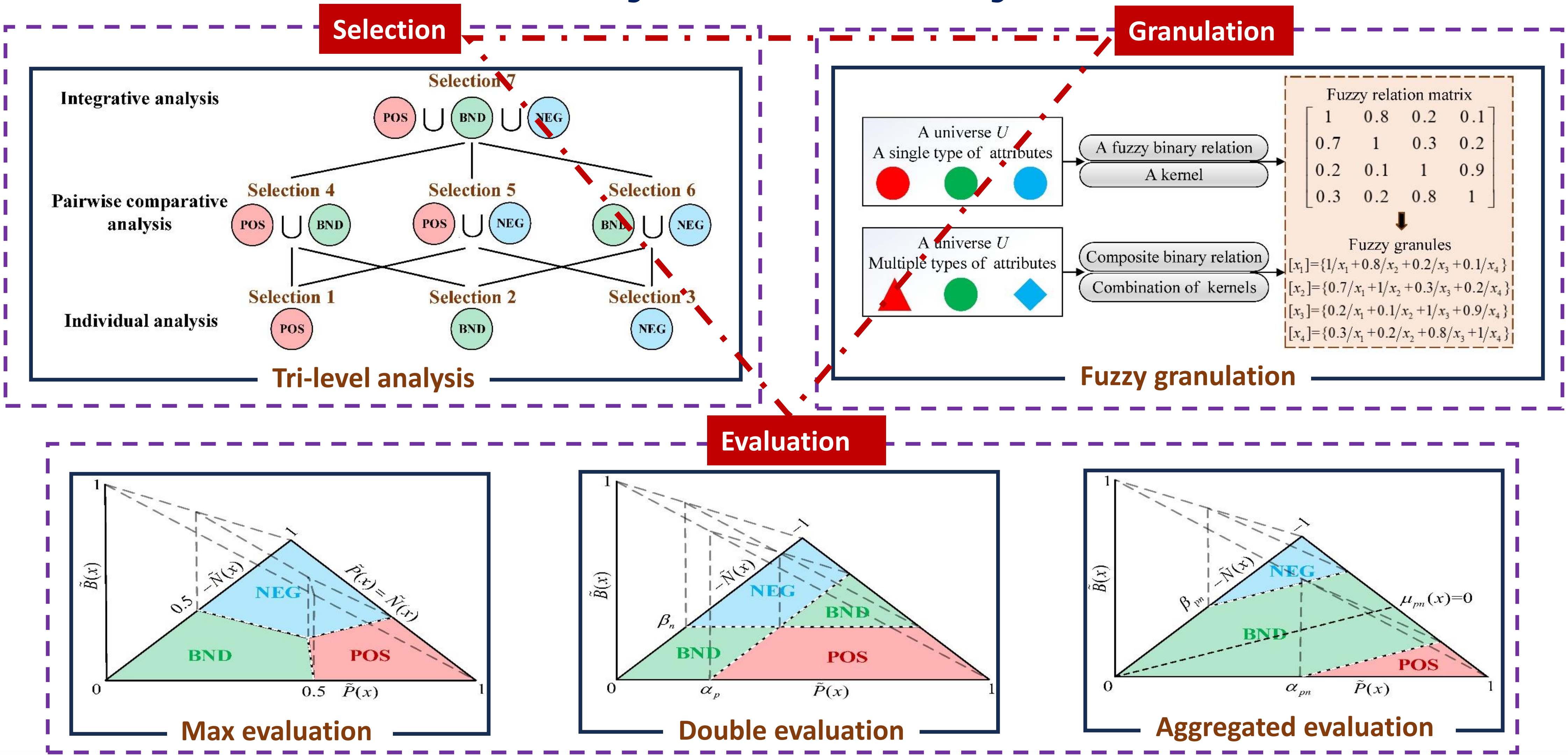
The framework of 3WMGrL



We utilize the ***temporal-spatial perspectives*** of ***three-way decisions*** to construct multi-granularity structures and implement multi-granularity learning in the **dynamic open-world environment**.

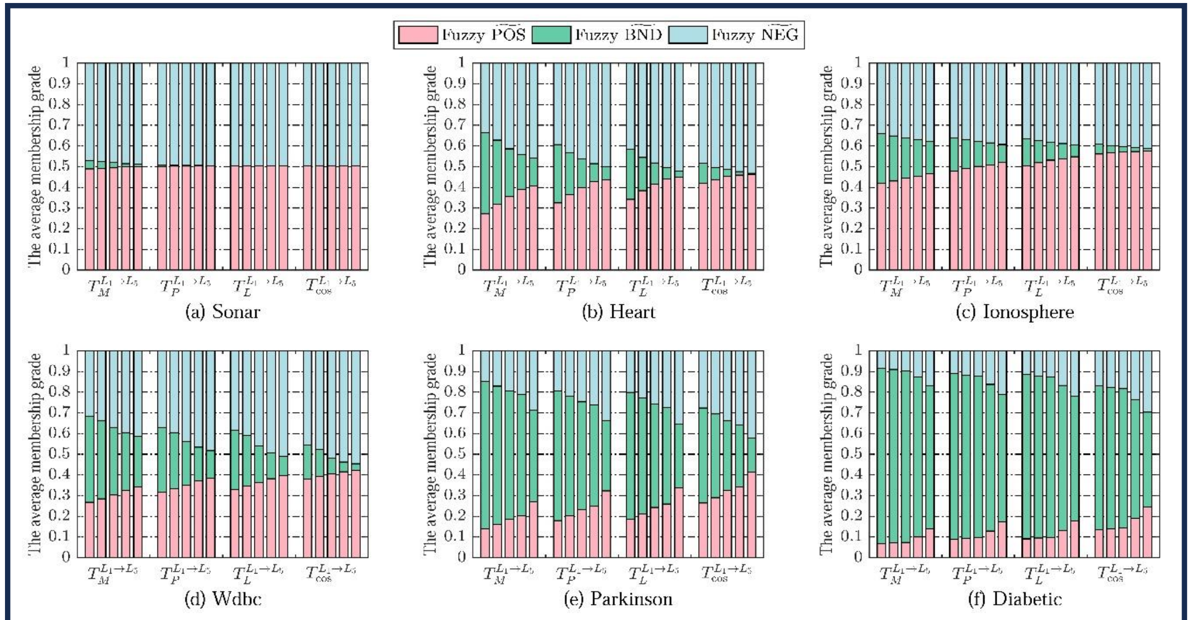
In dynamical paradigm, 3WMGrL focus on ***hierarchically*** thinking, information processing and decision-making in threes by ***incremental data and model parameters***, and make a series of reasonable three-way decisions with the **knowledge accumulation and transfer** under the multigranularity structures.

3WMGrL for Dynamic Fuzzy Environment

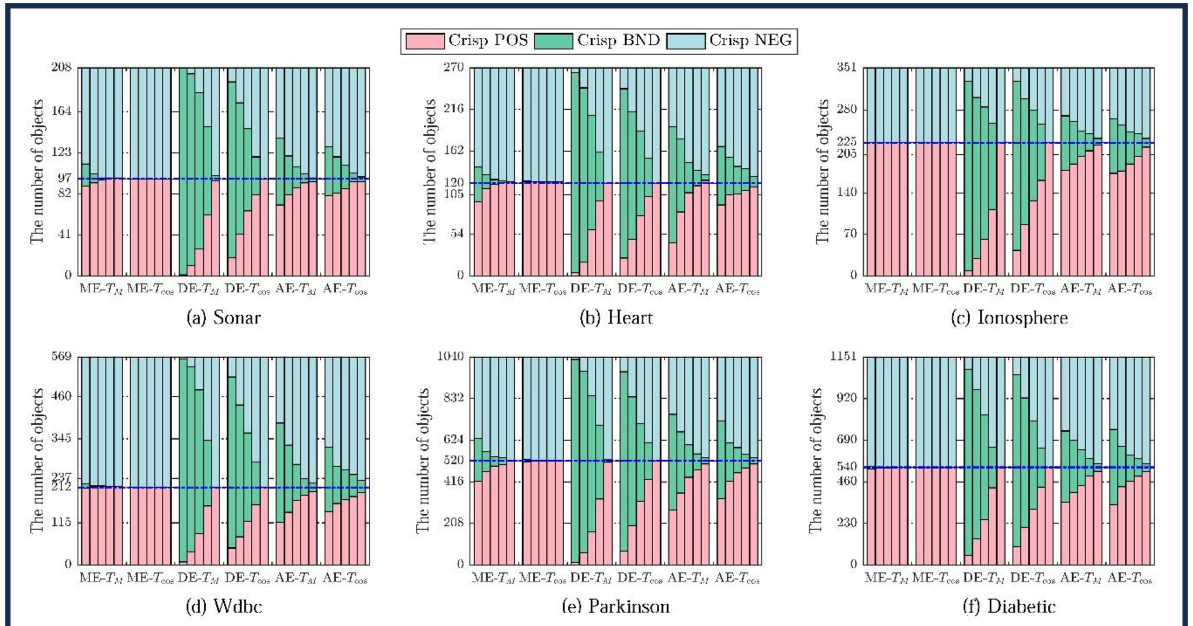


Evaluation

- The uncertainty with the boundary regions is reduced incrementally



The changes of fuzzy three-way regions



The changes of crisp three-way regions

Problem 5: Multi-source Heterogeneity

Predicting Citywide Crowd Flows Using Deep Spatio-Temporal Residual Networks

Joint work with Microsoft JD.COM 京东

Junbo Zhang, Yu Zheng, Dekang Qi, Ruiyuan Li, Xiuwen Yi, Tianrui Li, ***Predicting Citywide Crowd Flows Using Deep Spatio-Temporal Residual Networks***, Artificial Intelligence, 2018

Background

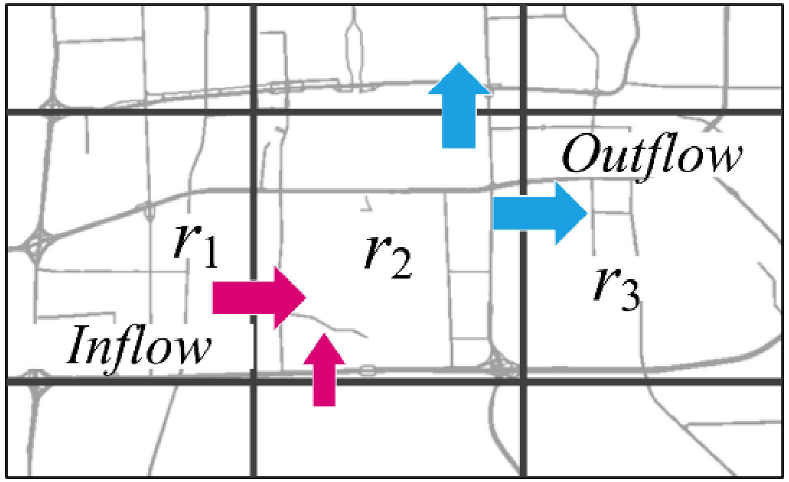
- Predicting crowd flows in a city is of great importance to traffic management, risk assessment, and public safety.
  - At least 146 dead after stampede during Halloween festivities in Itaewon, South Korea.

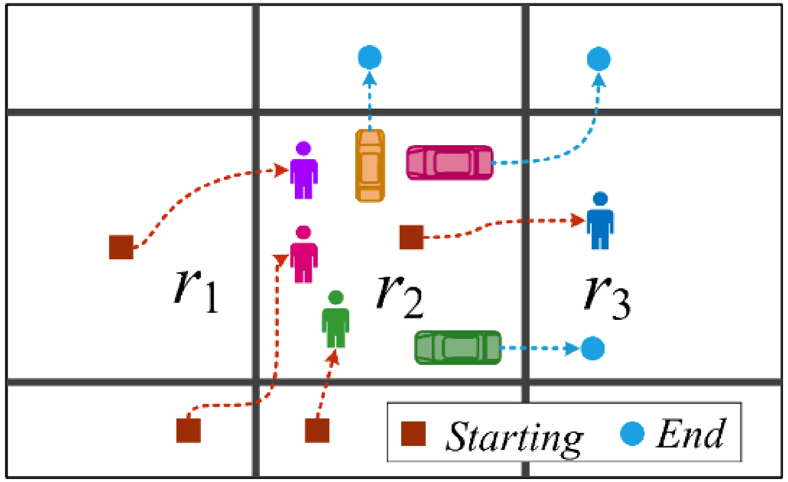
If one can predict the crowd flow in a region, such tragedies can be mitigated or prevented by utilizing emergency mechanisms, e.g., conducting traffic control, sending out warnings, or evacuating people, in advance.

Predicting Citywide Crowd Flows Using Deep Spatio-Temporal Residual Networks

- Inflow is the total traffic of crowds entering a region from other places during a given time interval.
- Outflow denotes the total traffic of crowds leaving a region for other places during a given time interval.

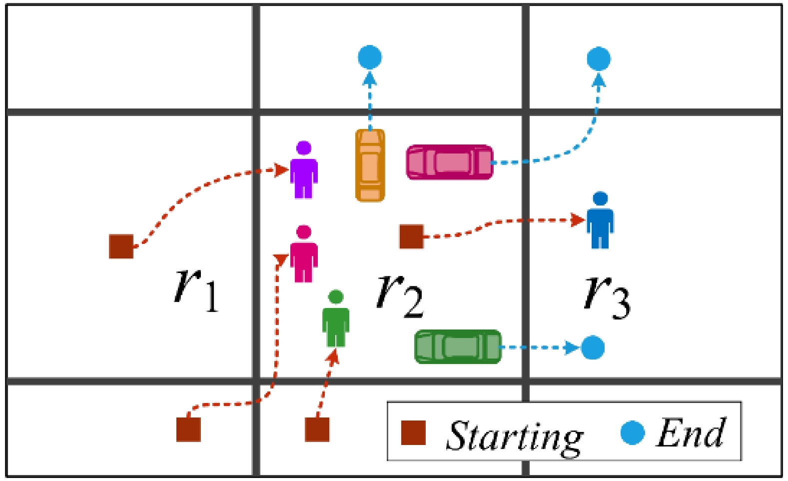
Task: To predict two types of crowd flows: inflow and outflow.





Predicting Citywide Crowd Flows Using Deep Spatio-Temporal Residual Networks

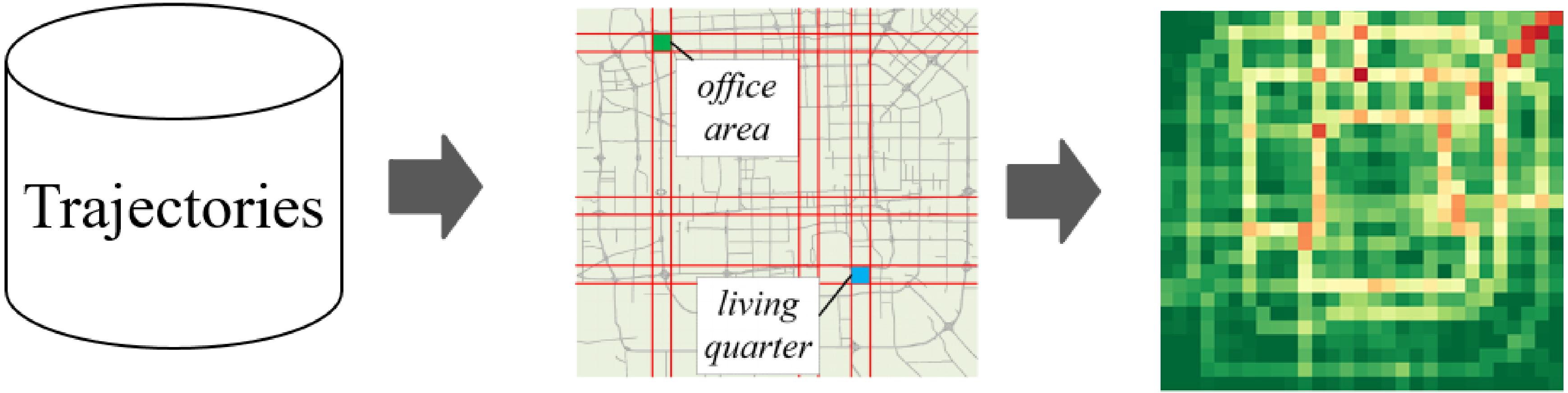
- Inflow/outflow can be measured by
  - the number of pedestrians
  - the number of cars driven nearby roads
  - the number of people traveling on public transportation systems (e.g. metro, bus)
  - or all of them together if data is available.



We can use mobile phone signals to measure the number of pedestrians, showing that the inflow and outflow of r2 are (3,1), respectively.
Similarly, using the GPS trajectories of vehicles, two types of flows are (0,3), respectively.
Therefore, the total inflow and outflow of r2 are (3,4), respectively.

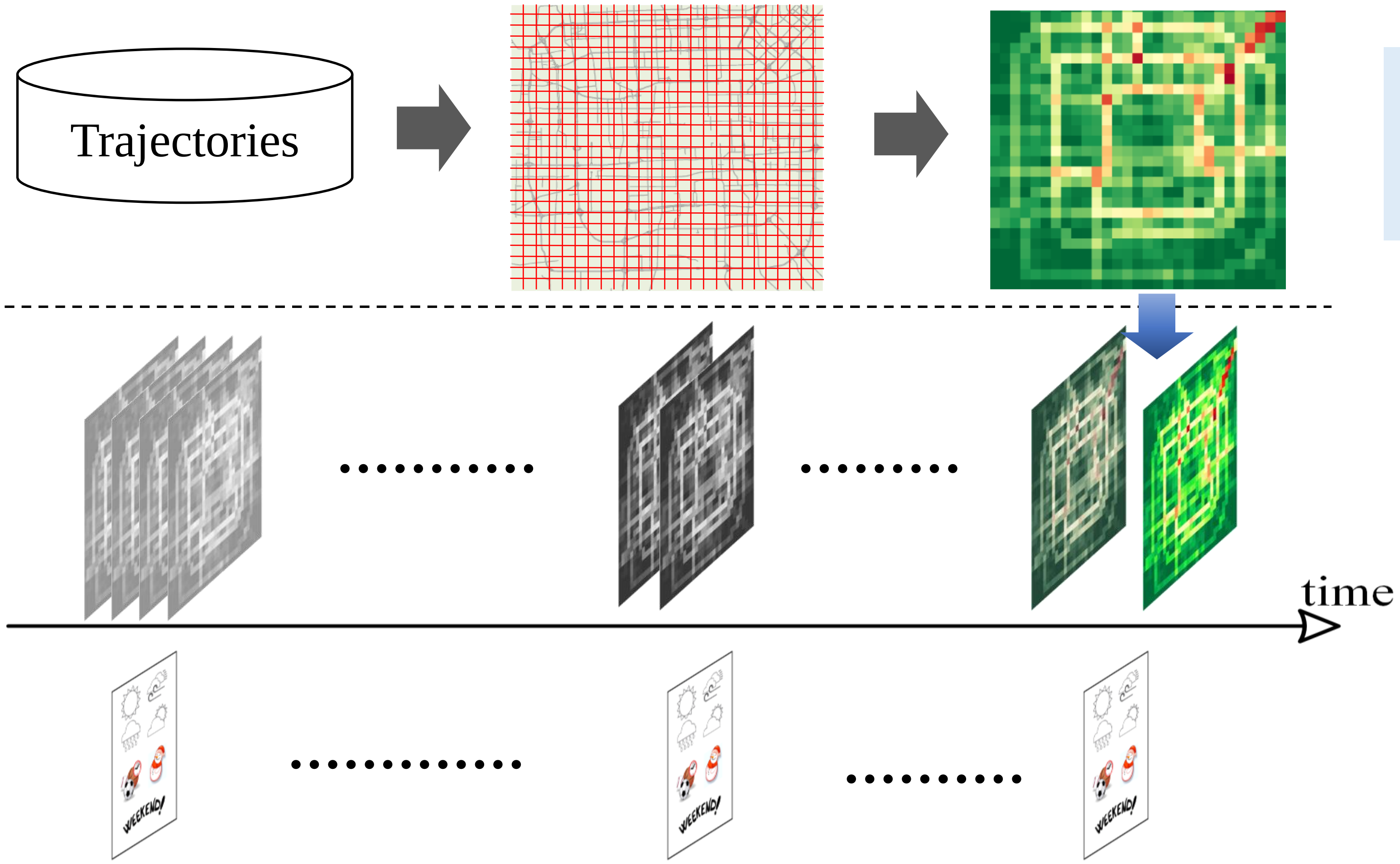
Predicting Citywide Crowd Flows Using Deep Spatio-Temporal Residual Networks

Our Solution: DST-ResNet

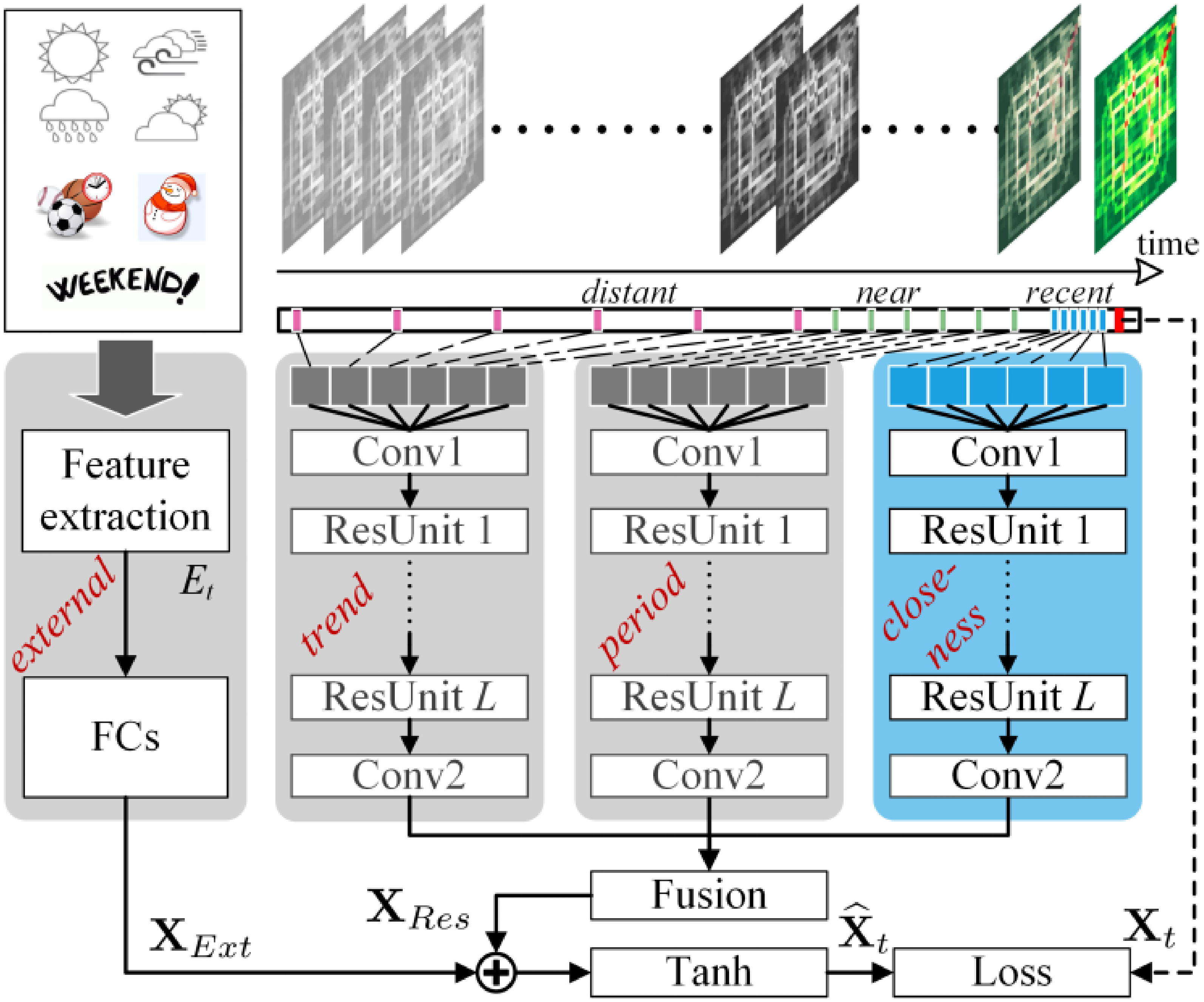


Converting trajectories into video-like data

Predicting Citywide Crowd Flows Using Deep Spatio-Temporal Residual Networks



The brighter the area, the more people there are



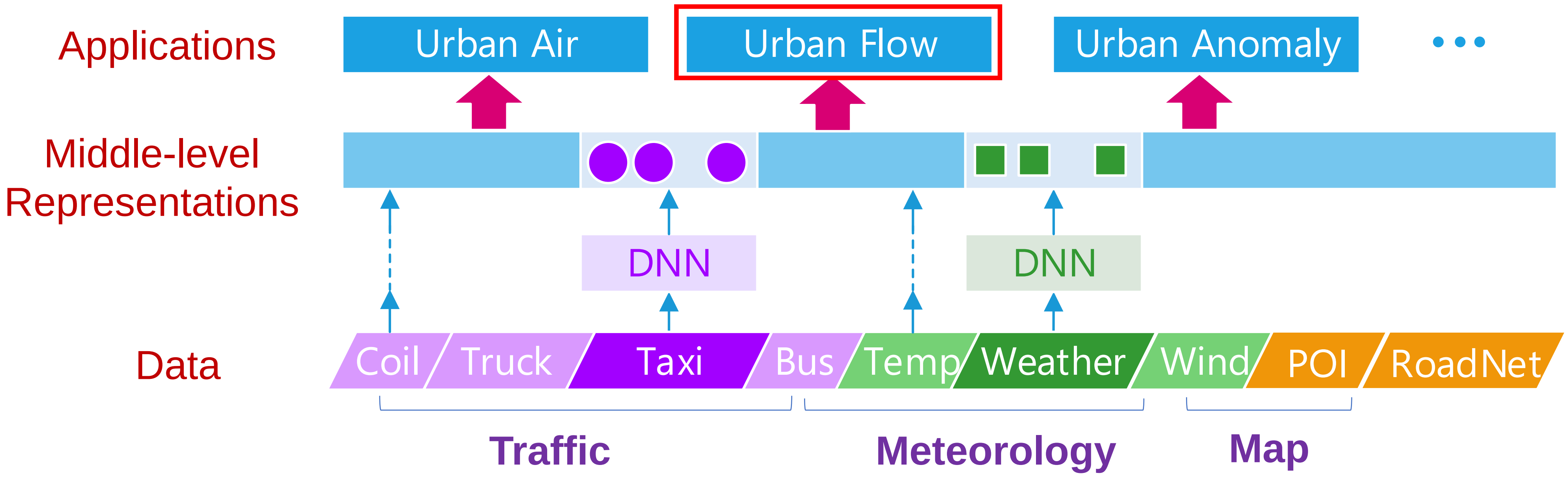
Multi-source data fusion

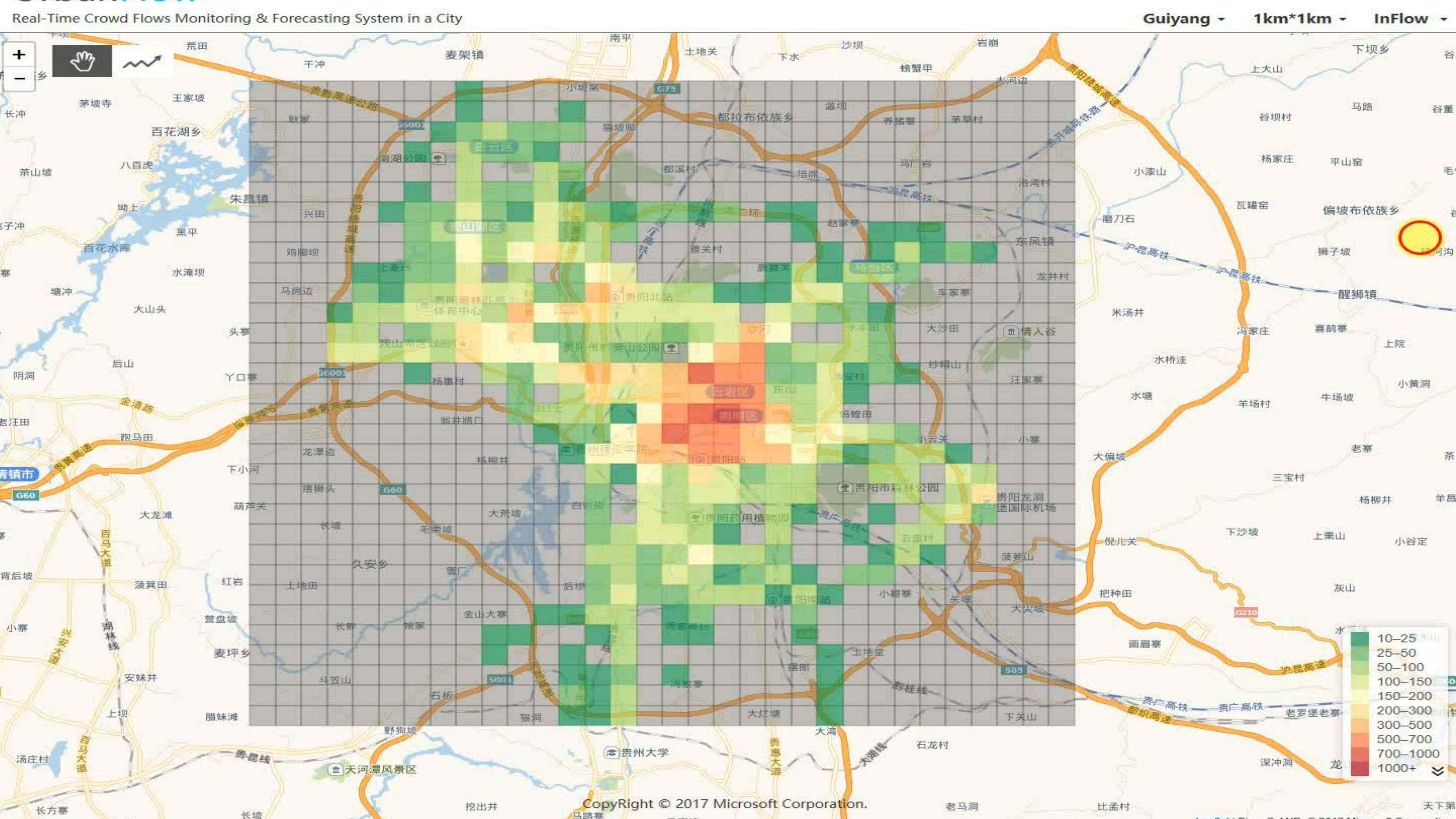
- The architecture of ST-ResNet.
- It comprised of four major components modeling temporal ***closeness, period, tre nd***, and ***external influence***, respectively.

Conv: Convolution;
ResUnit: Residual Unit;
FC: Fully-connected.

Predicting Citywide Crowd Flows Using Deep Spatio-Temporal Residual Networks

- Fusing multiple datasets

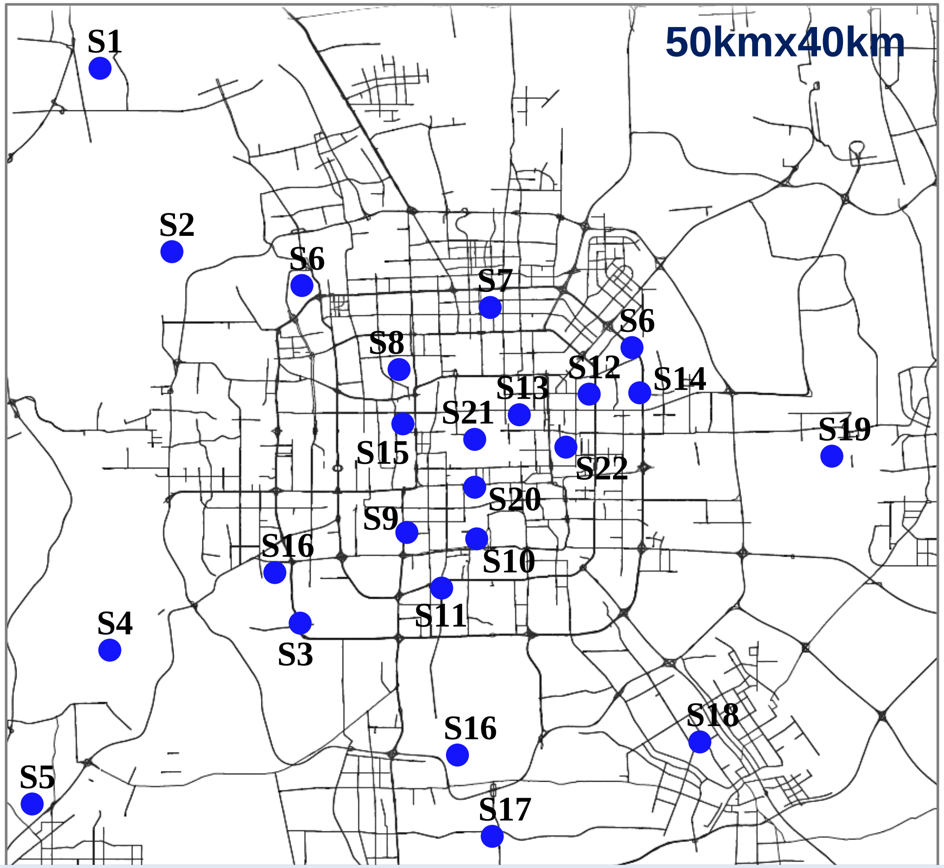



Real-Time Crowd Flows Monitoring & Forecasting System in a City
Guiyang
1km*1km
InFlow
10–25
25–50
50–100
100–150
150–200
200–300
300–500
500–700
700–1000
1000+
CopyRight © 2017 Microsoft Corporation.

Deep Distributed Fusion Network
for Air Quality Prediction

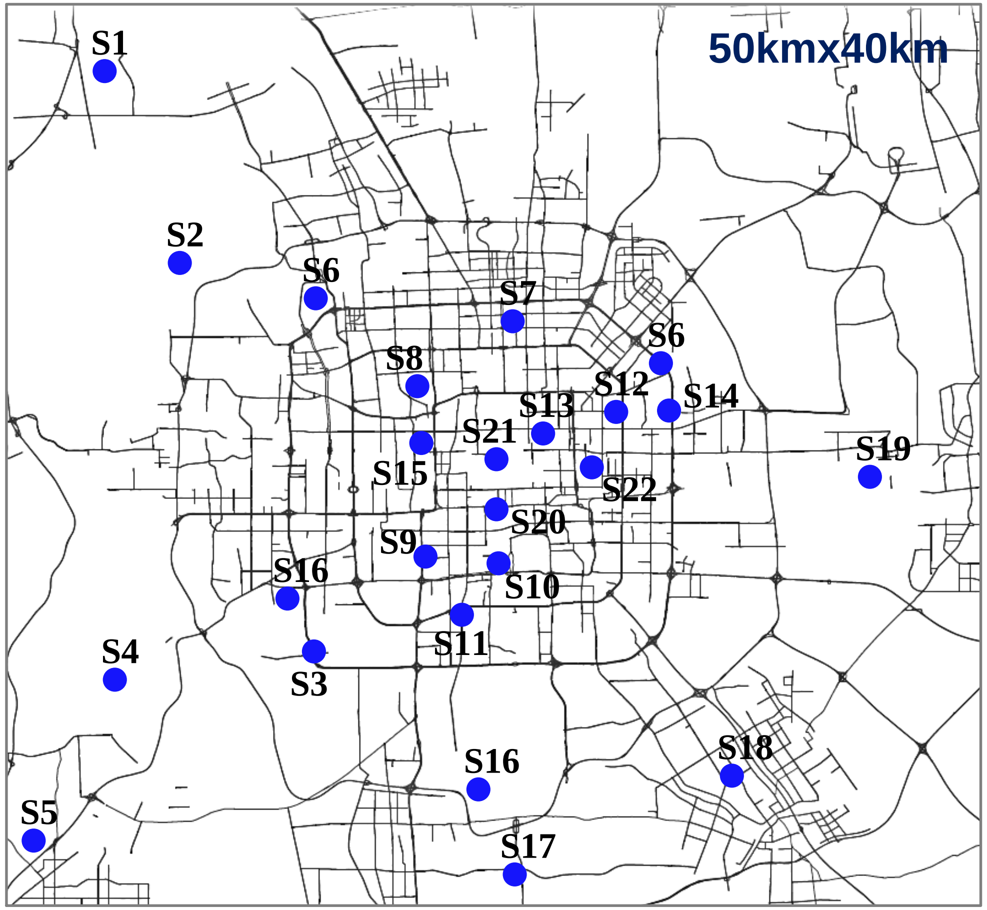
Joint work with Microsoft JD.COM 京东

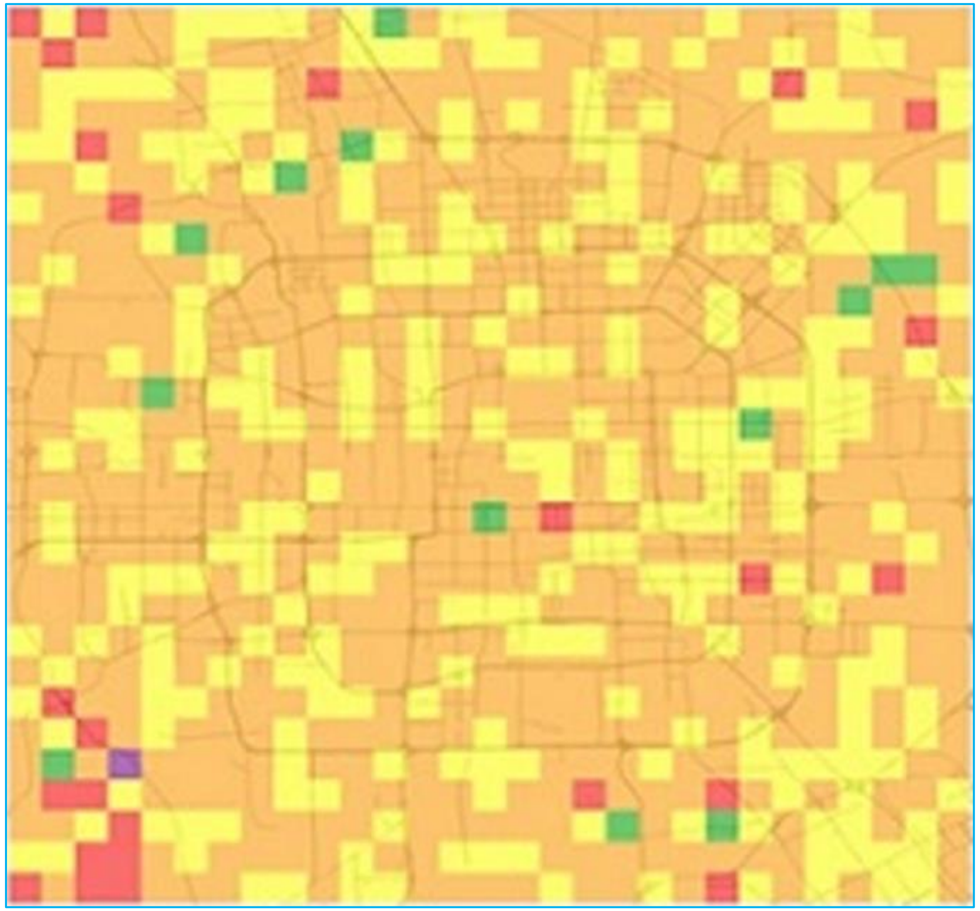
Background



With the rapid development of urbanization, air pollution is becoming a severe environmental and societal issue for all developing countries around the world.

Background

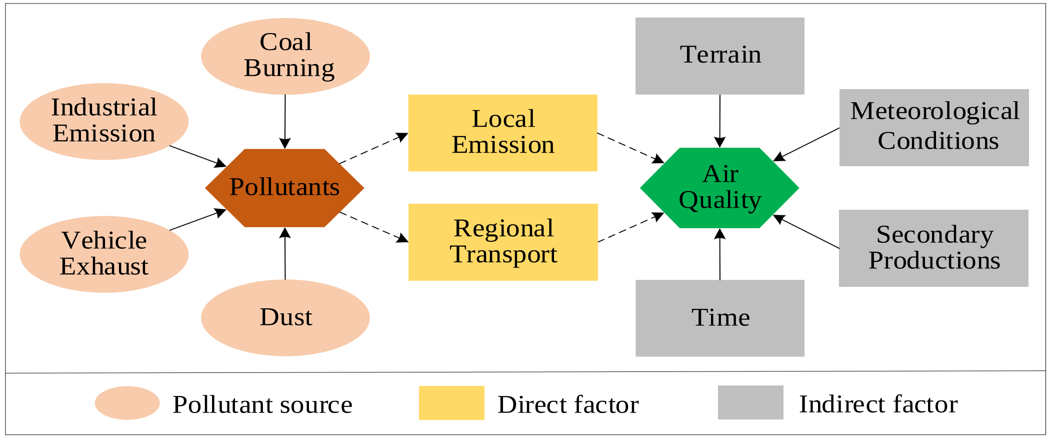




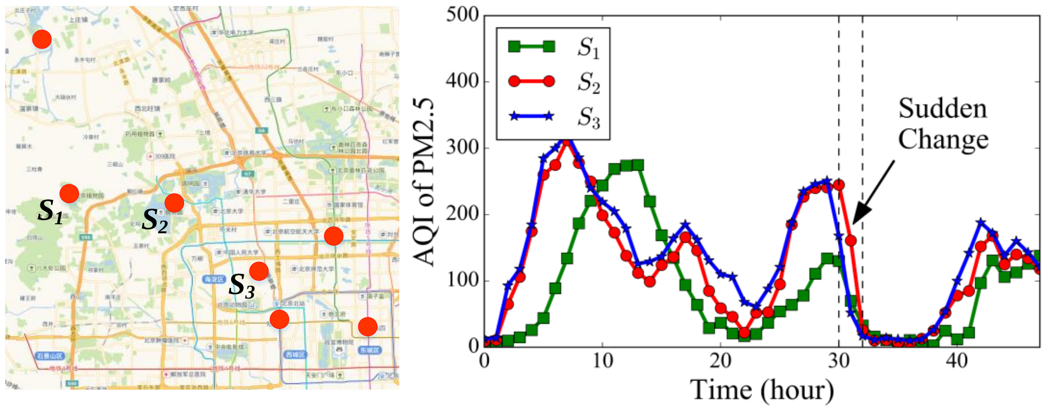
Air quality monitoring stations are limited. How can we infer the air quality at any location?

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

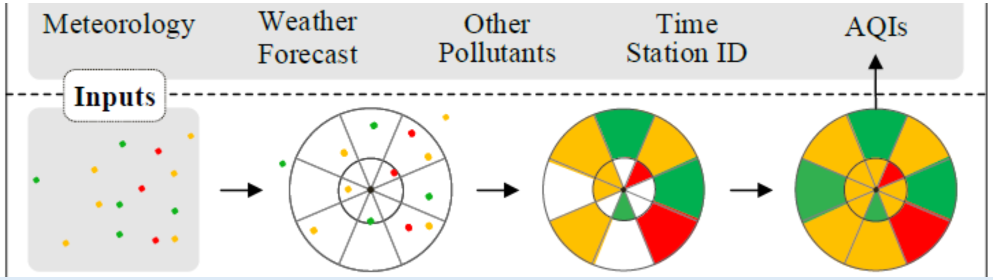


A) Monitoring stations B) AQI change over time

Deep Distributed Fusion Network

Multi-source data fusion

- **Spatial Transformation**
  - Air pollution dispersion
  - Spatial correlation
  - Scalability



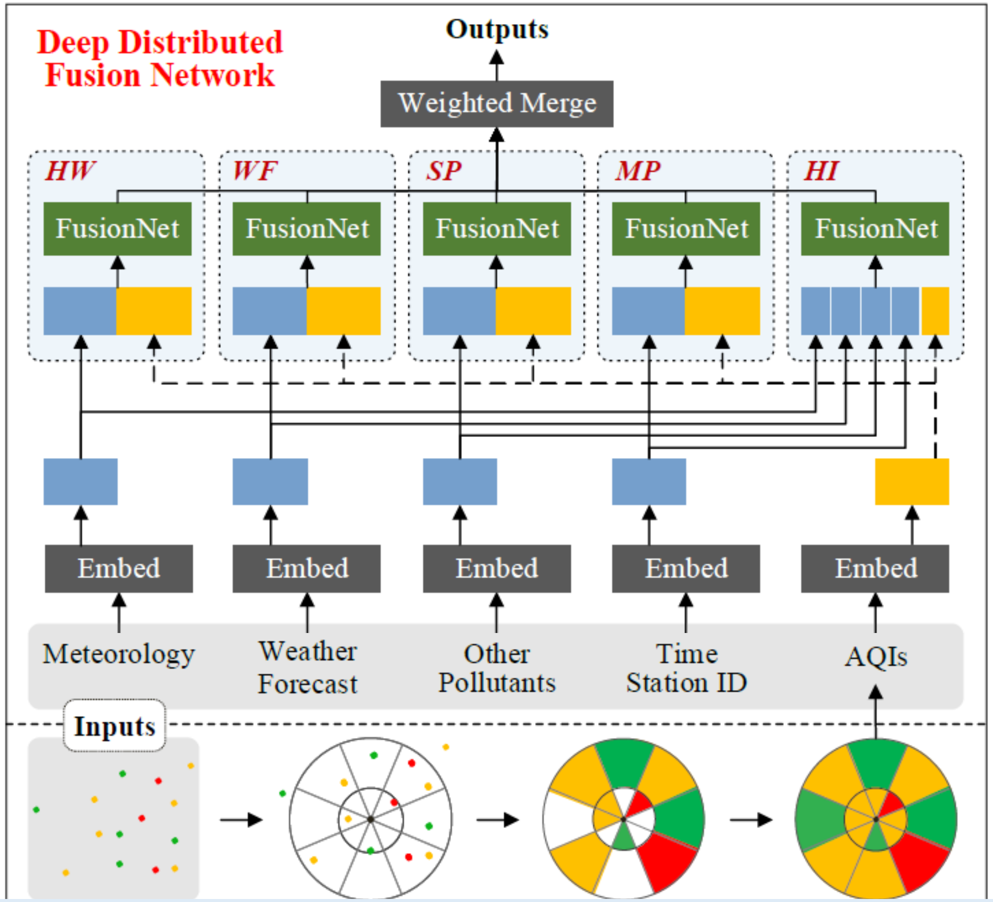
Considering air pollutants’ spatial correlations, the former component converts the spatial sparse air quality data into a consistent input to simulate the pollutant sources.

Deep Distributed Fusion Network

Multi-source data fusion

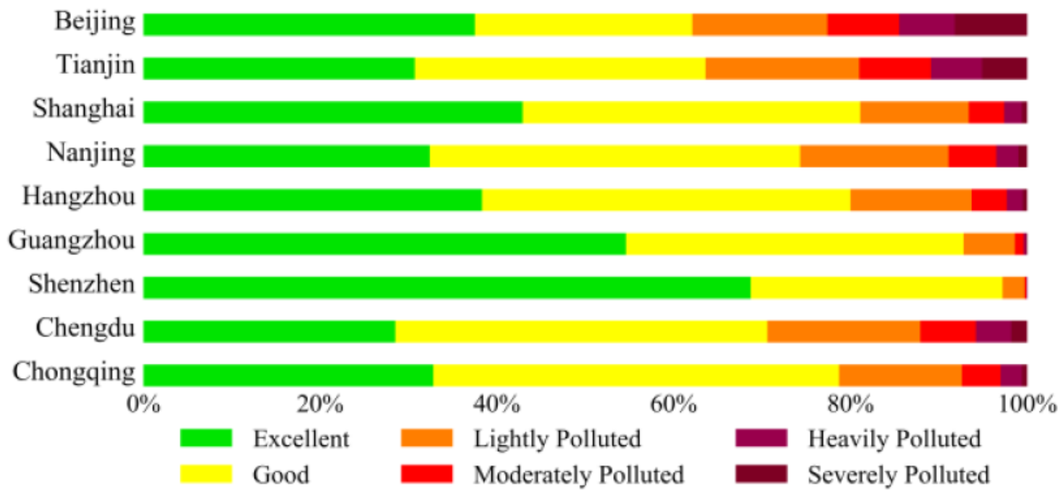
- **Spatial Transformation**
  - Air pollution dispersion
  - Spatial correlation
  - Scalability
- **Distributed FusionNet**
  - HW/WF/SP/MP nets to capture different individual influences
  - Capture holistic influence (HI)
- **Weighted Merge**

$\hat{y} = Sigmoid(\, y_{hw} \circ w_{hw} + y_{wf} \circ w_{wf} +$



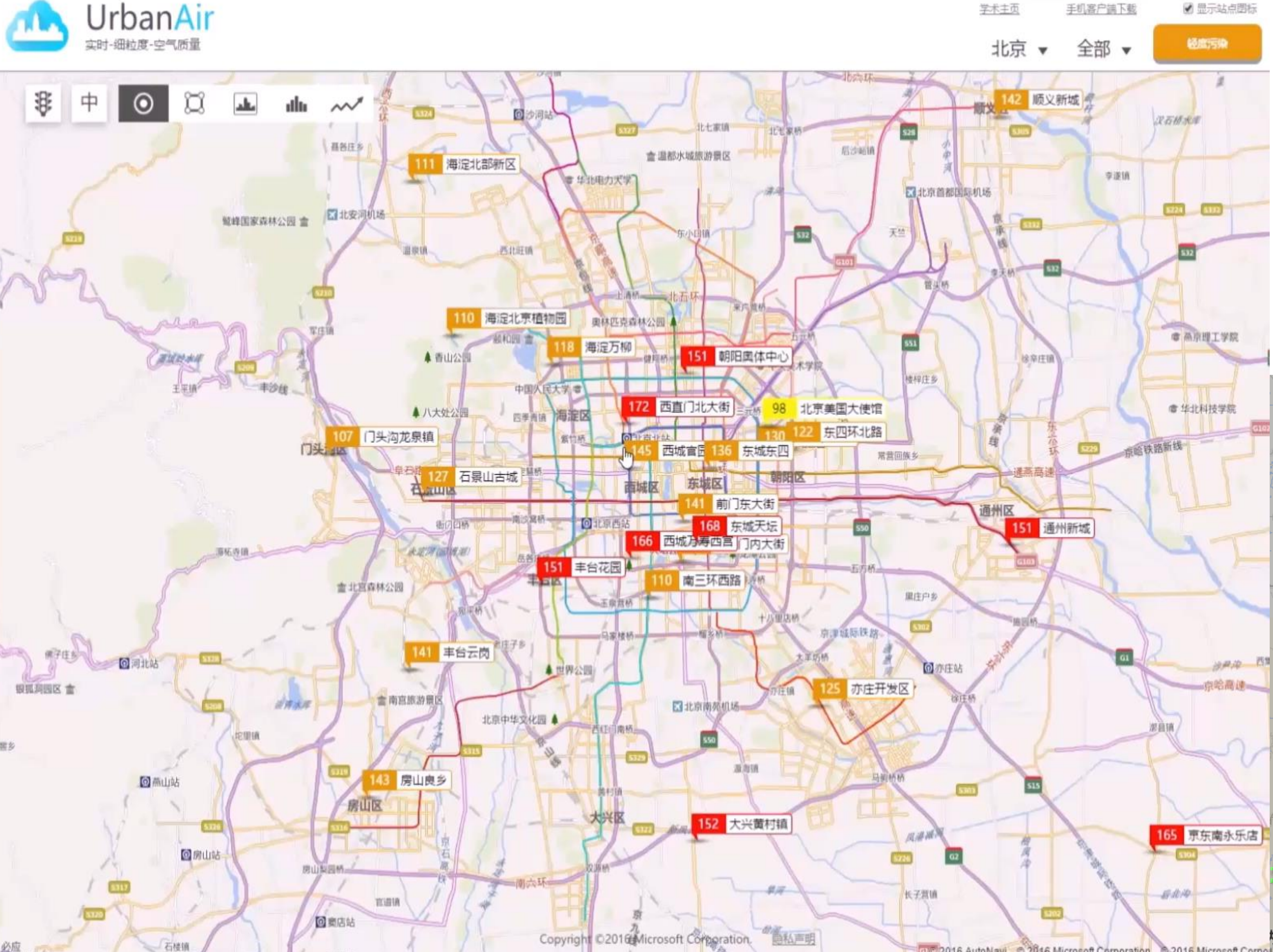
The latter network adopts a neural distributed architecture to fuse heterogeneous urban data for simultaneously capturing the factors affecting air quality, e.g. meteorological conditions.

Evaluation



Air Quality	In-city stations	36
	Instances	875,394
	Sudden changes	20,540
	Average $PM_{2.5}$	118.2
	Neighbor stations	74
Meteorology	Sources	17
	Instances	327,514
Weather Forecast	Sources	17
	Instances	298,790

Method	1-6h		7-12h		13-24h		24-48h		Sudden Change	
	acc	*mae*	*acc*	*mae*	*acc*	*mae*	*acc*	*mae*	*acc*	*mae*
ARIMA	0.751	28.3	0.576	52.1	0.458	65.4	0.307	74.6	0.066	112.9
LASSO	0.790	21.9	0.620	39.7	0.534	48.9	0.452	57.1	0.273	87.2
GBDT	0.792	21.8	0.629	38.8	0.540	48.0	0.458	56.5	0.321	21.8
LSTM	0.780	23.1±0.1	0.606	41.2±0.1	0.491	53.2±0.1	0.380	64.8±0.1	0.240	90.1±1.1
LSTM-STC	0.794	21.6±0.2	0.622	39.6±0.2	0.508	51.4±0.1	0.396	63.0±0.3	0.314	82.5±1.6
DeepST	0.806	20.4±0.1	0.633	38.1±0.2	0.545	47.5±0.2	0.466	55.7±0.7	0.380	74.5±2.9
DMVST-Net	0.806	20.4±0.1	0.638	37.8±0.3	0.550	47.4±0.5	0.481	53.9±0.7	0.419	70.4±2.0
DeepFM	0.808	20.1±0.1	0.643	37.3±0.2	0.549	47.2±0.6	0.474	54.9±0.6	0.396	72.3±1.9
DeepSD	0.811	19.7±0.1	0.645	37.1±0.2	0.551	46.8±0.8	0.479	54.3±0.7	0.428	69.5±3.3
DeepAir	**0.812**	**19.5±0.2**	**0.656**	**36.1±0.2**	**0.569**	**45.1±0.1**	**0.500**	**52.1±0.3**	**0.471**	**63.8±2.8**



We have deployed DeepAir in AirPollutionPrediction system, providing fine-grained air quality forecasts for **300+** Chinese cities every hour.

Metro Train Scheduling Optimization to Shorten Passengers' Travel Time

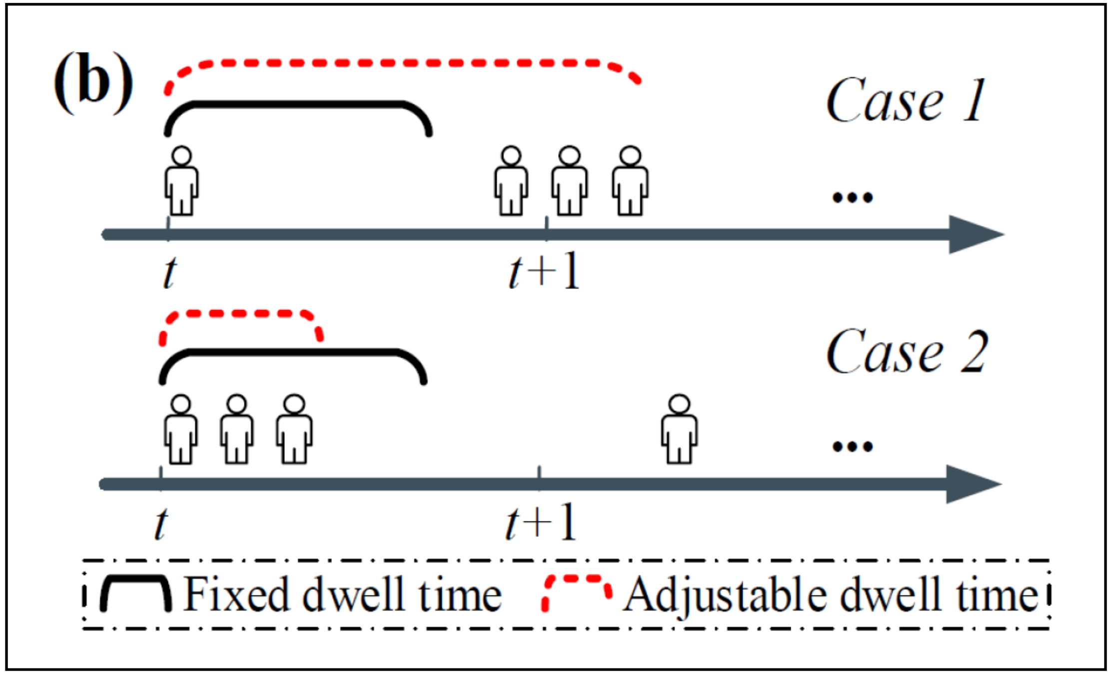
Joint work with Tencent 腾讯 JD.COM 京东

Zhaoyuan Wang, Zheyi Pan, Shun Chen, Shenggong Ji, Xiuwen Yi, Junbo Zhang, Jingyuan Wang, Zhiguo Gong, Tianrui Li, Yu Zheng. ***Shortening passengers' travel time: A dynamic metro train scheduling approach using deep reinforcement learning***. IEEE Transactions on Knowledge and Data Engineering, 2023.

Background

- **Shorten passenger travel time by adjusting train dwell time within a reasonable range**
  - Shorten passenger travel time -> Improve work efficiency

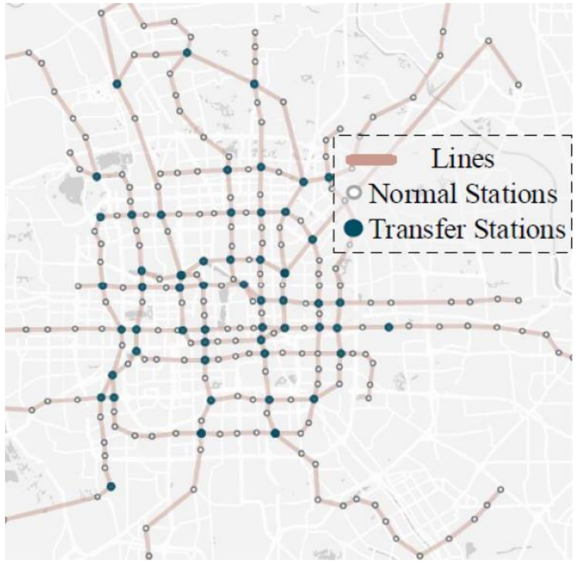
  - **Traditional method:** increasing the number of trains/ accelerating train speed



Metro Train Scheduling Optimization

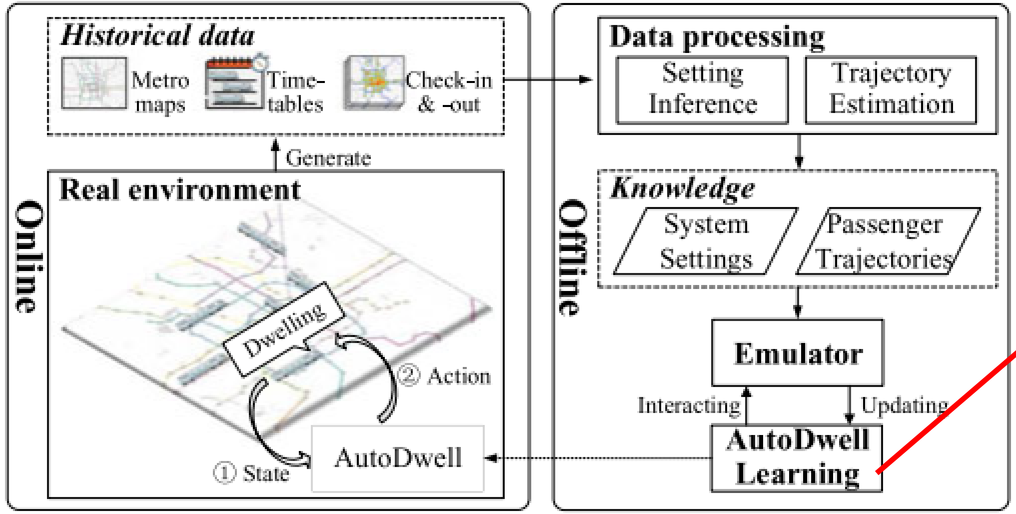
Challenges and Solutions

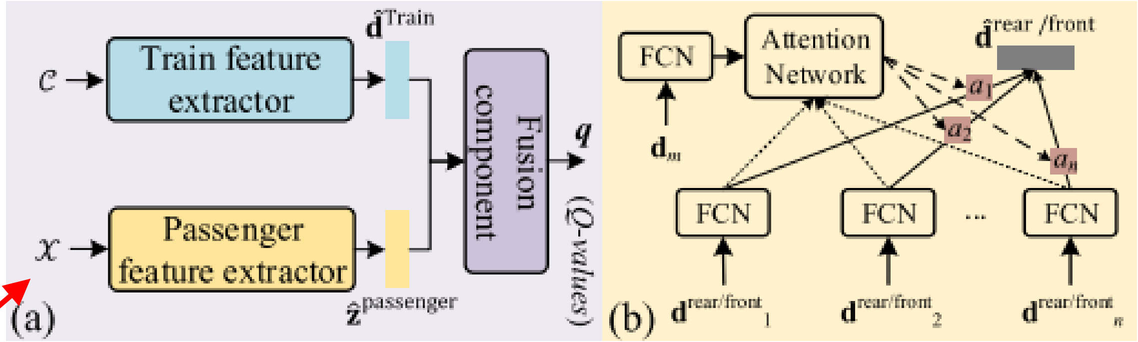
- Adjusting the dwell time will have a long-term impact
  - ✓ Use deep reinforcement learning models to capture this impact
- The complex spatio-temporal relationship affects distribution of passengers
  - ✓ Design a deep learning module composed of networks, e.g. graph attention mechanism to characterize it
- Complex interactions between trains are generated
  - ✓ Develop a deep learning module composed of attention networks and other components to model it



Scheduling System Framework

- A deep neural network named ***AutoDwell*** is proposed as the scheduling policy to boost passengers’ experience



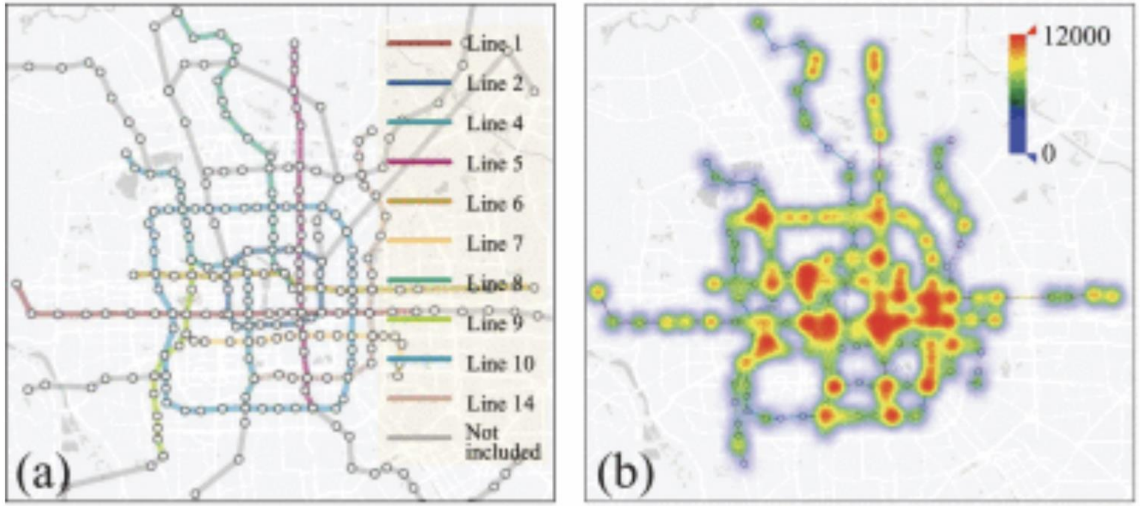


- System framework: (a) Overview of AutoDwell; (b) Structure of the train feature extractor.

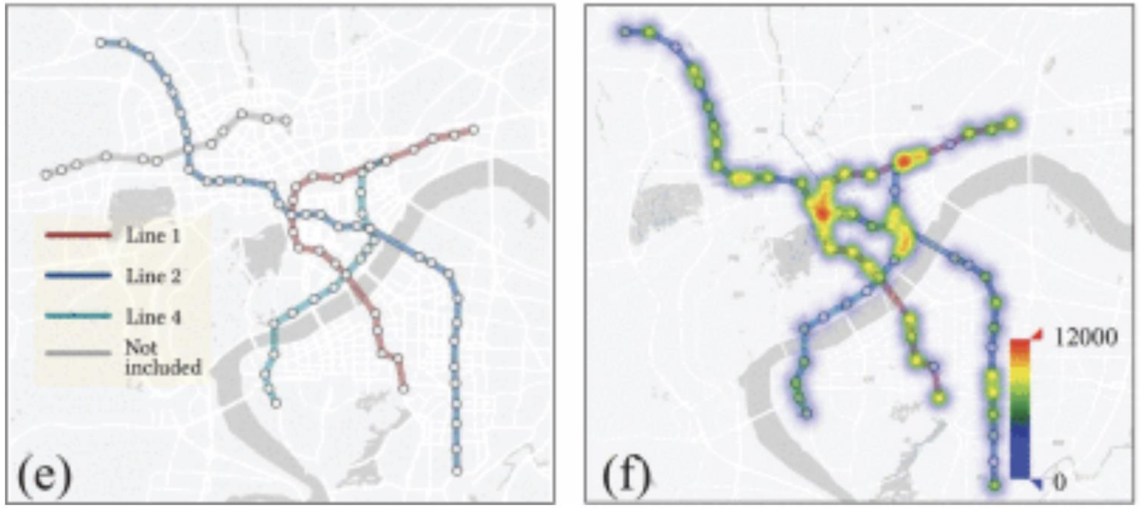
AutoDwell unlocks long-term rewards of actions according to the observed state by the guidance of the immediate reward, consisting of three components: train feature extractor, passenger feature extractor, and fusion network.

Data Sets

	Statistical levels & indicators		Beijing	Hangzhou
System	# lines		10	3
	# stations		182	66
	# transfer stations		36	5
	Transfer ratio		0.59	0.31
	Total length		274.34km	89.99km
	# daily records	aver.	4,215,786.91	1,154,317
		SD	163,251.11	88,848.90
	# stations for a trip	aver.	9.84	7.54
		SD	5.47	4.80
	Trip time	aver.	1,923.71s	1,515.50s
		SD	905.76	840.72
	# lines for a transfer trip	aver.	1.38	1.06
		SD	0.52	0.24
Line	# stations of one line	max	45	32
		min	13	18
		aver.	21.80	23.67
		SD	8.42	6.02
	# of trips per day & line	max	757,302	594,262
		min	170,982	155,484
		aver.	351,336.70	350,789.30
		SD	167,671.38	182,339.6
Station	# of check-in records per day & station	max	84,561	93,372
		min	2,051	3,418
		aver.	19,304.21	16,677.12
		SD	12,176.76	13,279.19
	Distance between two neighbors	max	3.00km	3.32km
		min	0.42km	0.60km
		aver.	1.31km	1.32km
		SD	0.43	0.48



(a) Metro map of Beijing; (b) Daily spatial distribution of check-in records in Beijing



(e) Metro map of Hangzhou; (f) Daily spatial distribution of check-in records in Hangzhou

Experiments on Saving Travel Time

Methods			FM *Min*	FM *Max*	FM *Aver*	HM *Day*	HM *Hour*	PM *ARIMA*	PM *RNN*	PMM *ARIMA*	PMM *RNN*	autoDwell *w/o-T*	autoDwell *w/o-P*	autoDwell *w-T&P*
Beijing	$\mathcal{P}_1$	$\bar{\delta}$	1896.391	1938.902	1904.756	1874.787	1868.257	1883.834	1885.194	1865.512	1864.206	1868.823	1849.393	**1842.855**
		$\bar{\xi}$	0.232	0.233	0.228	0.227	0.225	0.230	0.231	0.225	0.223	0.225	0.216	**0.212**
	$\mathcal{P}_2$	$\bar{\delta}$	1893.647	1936.336	1902.402	1872.604	1865.263	1881.762	1881.268	1863.428	1862.060	1864.347	1844.769	**1838.188**
		$\bar{\xi}$	0.231	0.230	0.229	0.225	0.223	0.228	0.228	0.224	0.222	0.224	0.215	**0.210**
Hangzhou	$\mathcal{P}_1$	$\bar{\delta}$	1503.352	1519.288	1513.241	1486.735	1485.984	1498.058	1501.693	1482.276	1479.821	1493.215	1463.429	**1455.967**
		$\bar{\xi}$	0.293	0.298	0.299	0.287	0.286	0.294	0.295	0.288	0.287	0.290	0.279	**0.275**
	$\mathcal{P}_2$	$\bar{\delta}$	1499.453	1503.363	1501.068	1484.080	1482.817	1485.427	1483.420	1480.180	1476.190	1487.137	1460.343	**1452.075**
		$\bar{\xi}$	0.285	0.284	0.285	0.278	0.277	0.285	0.284	0.278	0.276	0.286	0.277	**0.271**

Experiments on Saving Train Resources

Methods	Beijing $\mathcal{P}_1$ Ti-	Beijing $\mathcal{P}_1$ Tr+	Beijing $\mathcal{P}_2$ Ti-	Beijing $\mathcal{P}_2$ Tr+	Hangzhou $\mathcal{P}_1$ Ti-	Hangzhou $\mathcal{P}_1$ Tr+	Hangzhou $\mathcal{P}_2$ Ti-	Hangzhou $\mathcal{P}_2$ Tr+
FM-*Max*	4	11	4	11	16	14	15	13
HM-*Hour*	2	5	1	3	10	7	9	6
PM-*RNN*	2	5	2	5	13	10	14	12
PMM-*RNN*	1	3	1	3	8	5	8	5

- It can save an average of **20 seconds** per trip and millions of minutes of commuting time per day;
- It can guide the daily savings of **dozens of train resources** while maintaining the current level of commuting.

Real-time Ambulance Redeployment

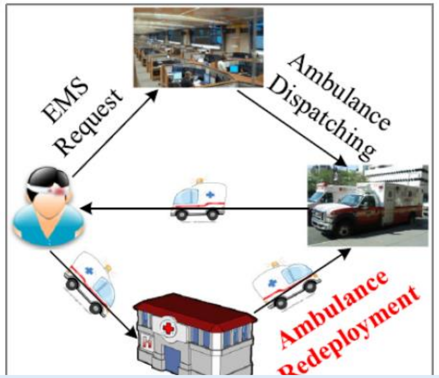
Joint work with JD.COM 京东

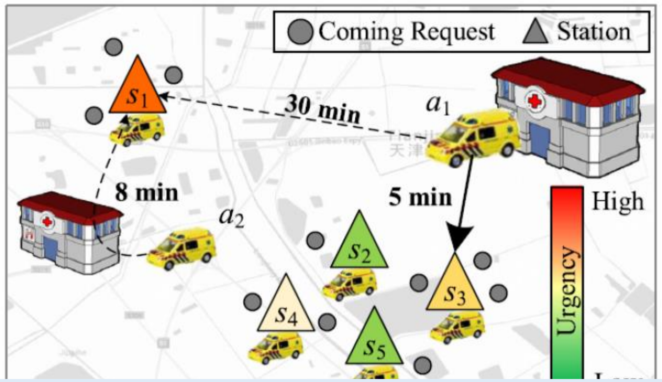
Shenggong Ji, Yu Zheng, Wenjun Wang, Tianrui Li. ***Real-time ambulance redeployment: A data-driven approach***. IEEE Transactions on Knowledge and Data Engineering, 2020.

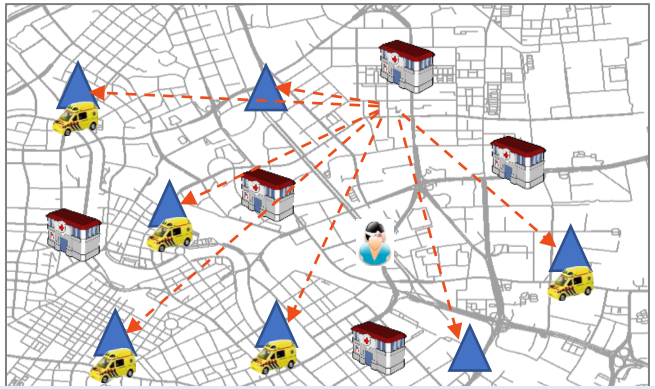
Q:

Emergency Medical Services (EMS)

- EMS are of great importance to saving people's lives
  - Traffic accident, emergent disease
- EMS significantly depends on the real-time redeployment strategy of ambulances
  - Select one station for redeployment
- Goal: minimize the waiting time of patients
  - Make a call → be picked up



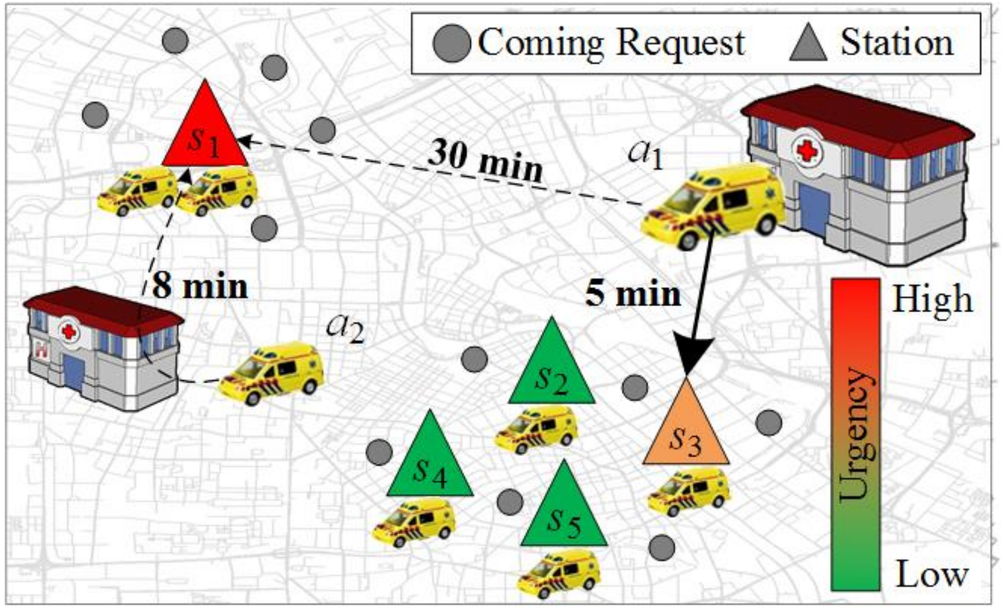




Q: Which station should an ambulance be redeployed to, after it becomes available (after it transports a patient to a hospital or after it finishes the in-site treatment for a patient)?

Main Challenges

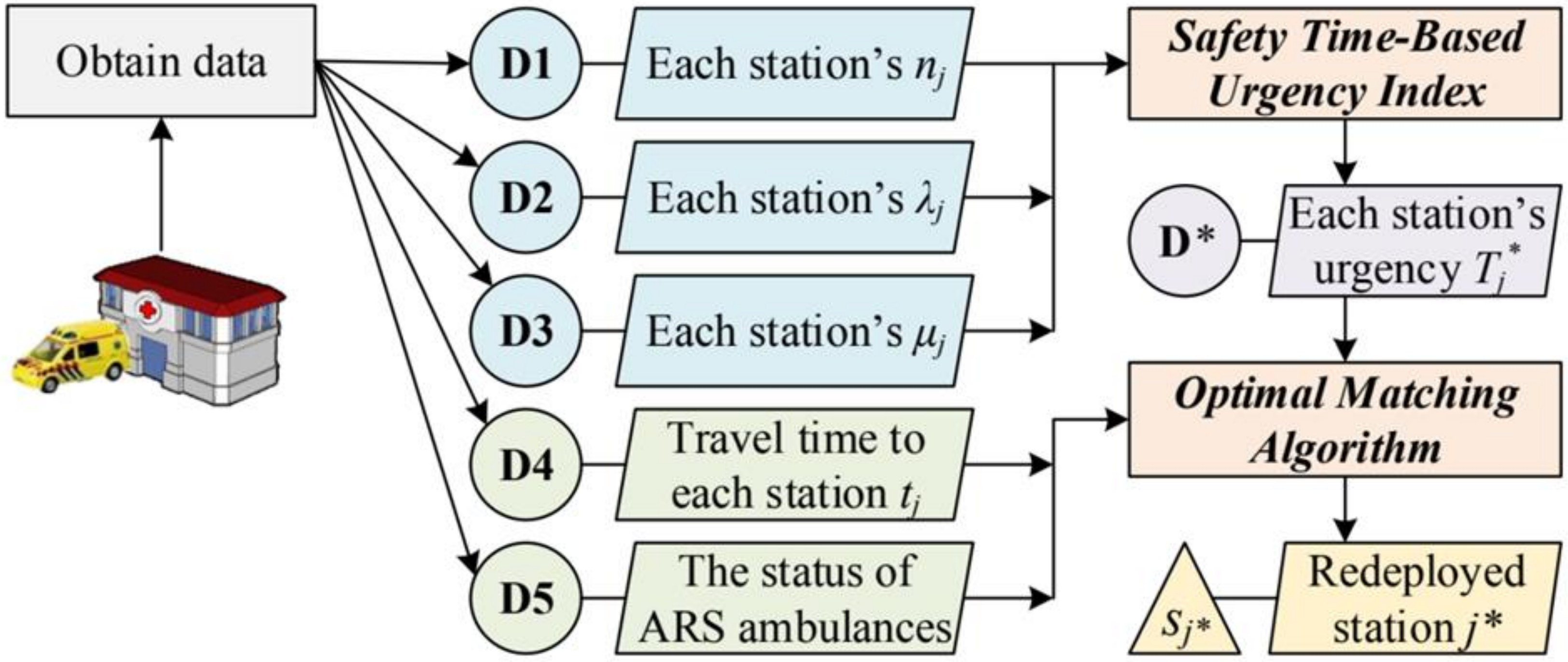
- When redeploying an ambulance, the following five factors need to be considered for each station:
  - D1: The number of available ambulances at each station.
  - D2: The number of EMS requests nearby each station in the future.
  - D3: The geographical location of each ambulance station.
  - D4: The travel time for the current available ambulance to reach each ambulance station.
  - D5: The status of other occupied ambulances.



How to properly redeploy an available ambulance requires a careful consideration of all these factors is an open challenge for redeployment strategy of ambulances.

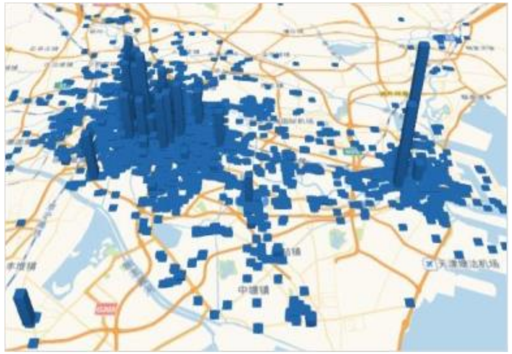
Data-driven Ambulance Redeployment

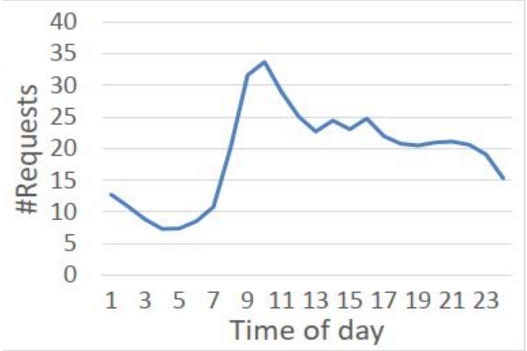
- Safety time-based urgency index: D1、D2、D3 → D*
- Spatial-temporal optimal matching algorithm: D*、D4、D5 → Redeployment result

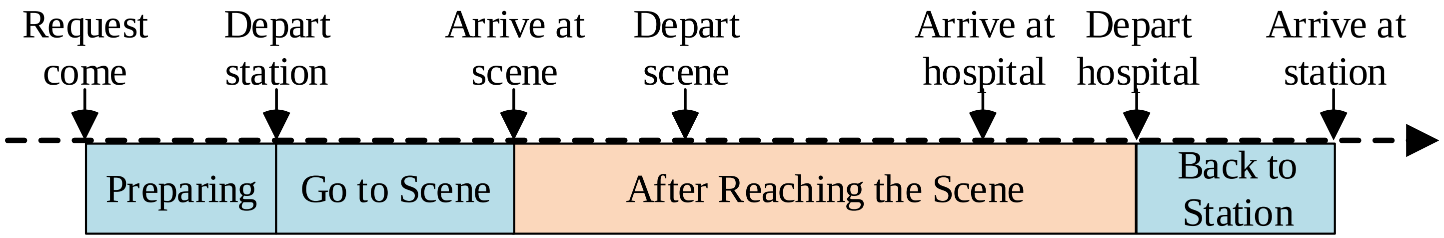


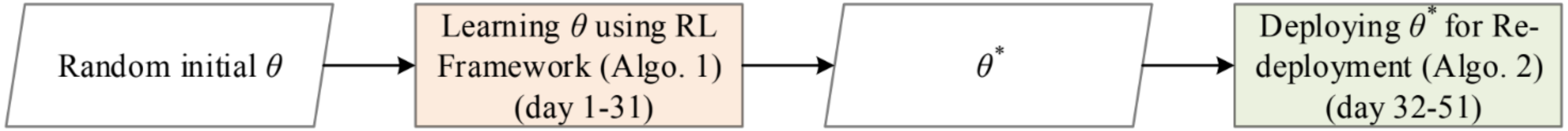
Evaluation

- Simulation based on real datasets
  - Patients in history
    - Oct. 1 to Nov. 21, 2014
    - 23,549 EMS requests
  - Ambulance stations (34)
  - Hospitals (41)
  - Road networks









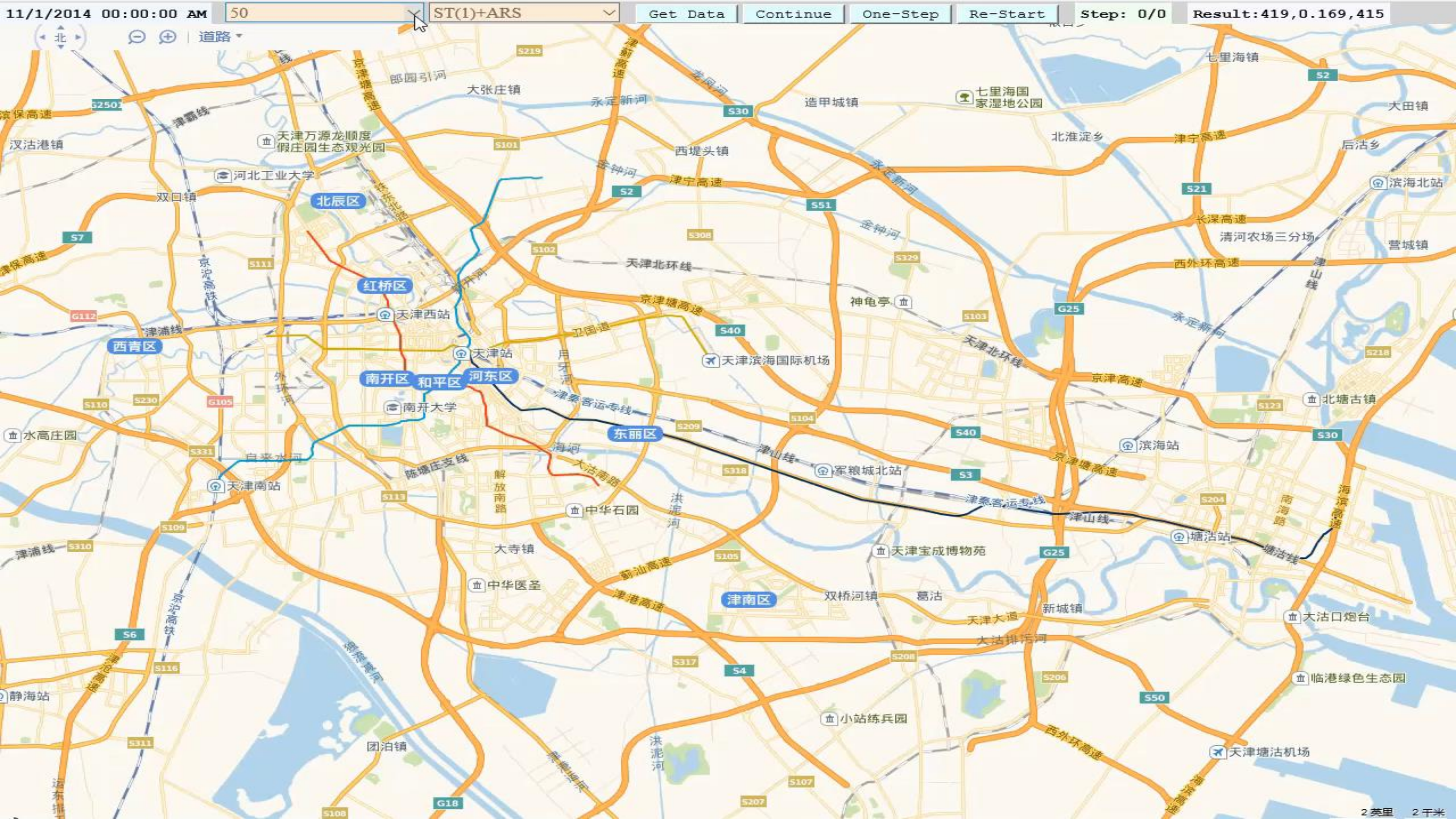
Evaluation

- Effectiveness

	#ambu = 50		#ambu = 60		#ambu = 70		#ambu = 80		#ambu = 90	
	AvePT	RelaPT	AvePT	RelaPT	AvePT	RelaPT	AvePT	RelaPT	AvePT	RelaPT
RS	856.5	0.531	778.7	0.569	759.6	0.583	747.9	0.591	741.4	0.596
NS	773.1	0.585	767.3	0.589	753.2	0.602	747.6	0.607	736.2	0.616
LS	603.8	0.745	531.2	0.785	480.1	0.808	444.8	0.823	423.9	0.831
ERTM	505.2	0.786	432.9	0.830	398.8	0.844	389.5	0.848	384.1	0.850
MEXCLP	502.4	0.774	461.8	0.822	409.1	0.852	392.9	0.862	376.9	0.872
DMEXCLP	516.9	0.773	447.3	0.826	408.7	0.852	387.1	0.865	375.2	0.872
DRLSN (ours)	*402.8*	*0.838*	*367.0*	*0.864*	*351.2*	*0.874*	*342.0*	*0.879*	*338.0*	*0.880*

- **AvePT**: Average pickup/waiting time
- **RelaPT**: Ratio of patients picked up within 10 minutes

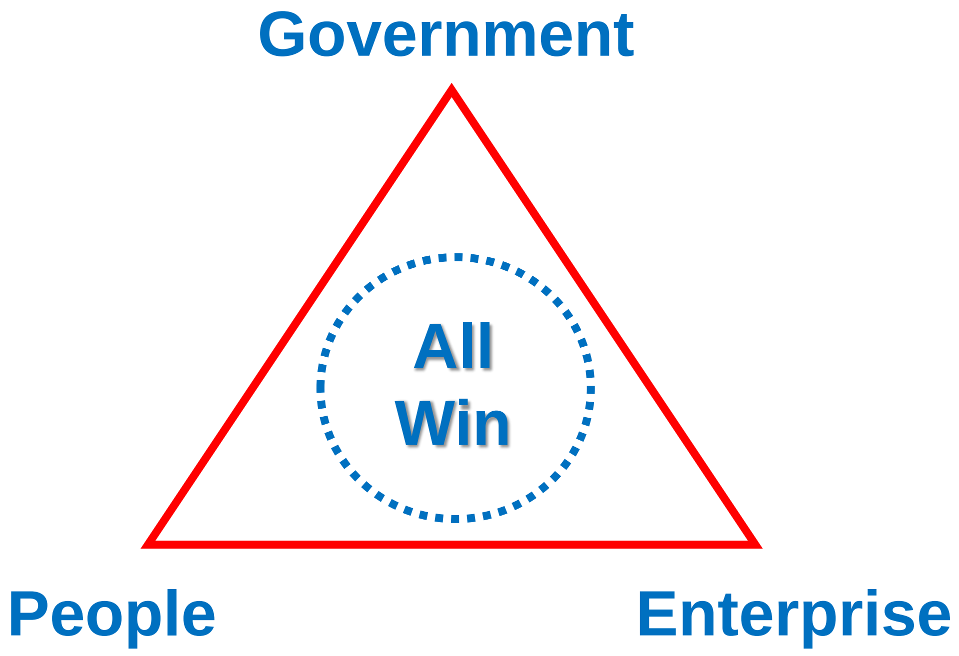
(#ambu=50) AvePT: save **~99 seconds (~20%)**
(#ambu=50) RelaPT: from **0.786** to **0.838**



11/1/2014 00:00:00 AM
50
ST(1)+ARS
Get Data
Continue
One-Step
Re-Start
Step: 0/0
Result:419,0.169,415
北
道路
北辰区
红桥区
天津西站
西青区
天津站
南开区
和平区
河东区
南开大学
东丽区
天津南站
天津滨海国际机场
中华石园
大寺镇
中华医圣
津南区
双桥河镇
葛沽
新城镇
大沽口炮台
临港绿色生态园
天津塘沽机场
塘沽站
滨海站
北塘古镇
滨海北站
军粮城北站
天津宝成博物苑
小站练兵园
团泊镇
静海站
水高庄园
汊沽港镇
双口镇
河北工业大学
天津万源龙顺度假庄园生态观光园
大张庄镇
西堤头镇
造甲城镇
七里海国家湿地公园
七里海镇
大田镇
后沽乡
北淮淀乡
清河农场三分场
营城镇
神龟亭
天津北环线
京津塘高速
津宁高速
京津高速
津山线
津秦客运专线
陈塘庄支线
海河
永定新河
2英里
2千米

Conclusions

- **Big Data Intelligence Challenges**
  - Small number of labeled samples
  - Privacy protection issues
  - High-dimensional data
  - Evolving data
  - Multi-source heterogeneous data
- **Application Case Study**
- **Future Work**
  - Data + Knowledge
  - Interpretability (Rough Set, Three-Way Decision, etc.)



Thanks！

South Gate of SWJTU

Library of SWJTU

School of Computing and Artificial Intelligence