数据驱动智能的挑战与解决策略

李天瑞

西南交通大学信息科学与技术学院

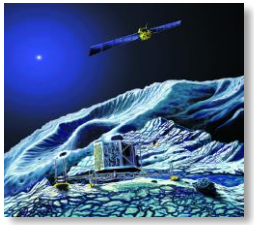
四川省云计算与智能技术高校重点实验室

trli@swjtu.edu.cn

数据驱动智能的挑战

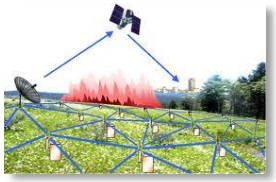
数据量大

科学仪器

社交网络

传感器




移动设备

数据驱动智能的挑战

实时处理

- 2013年3月4日7点20分，在北国长春就发生婴儿随车被盗事件。
- 一辆车牌号为吉AMM102的灰色RAV4车辆被盗，而当时，一个男婴就在车内。
- 2013年3月5日8点找到车。

数据驱动智能的挑战

不确定性

数据驱动智能的挑战

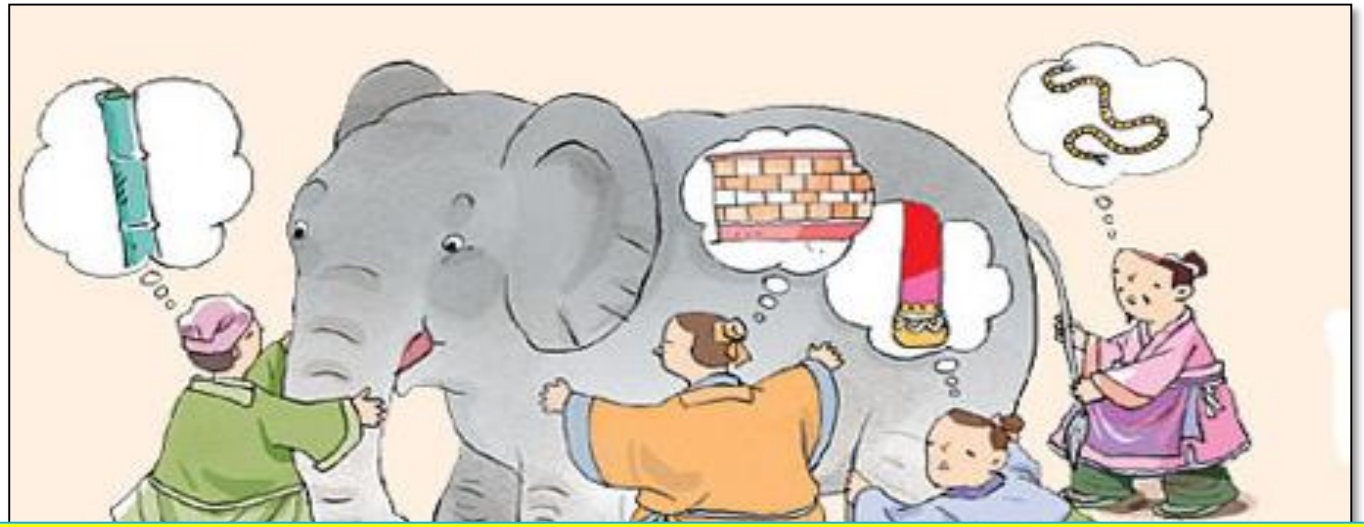
隐私问题

- 谷歌监视着我们的网页浏览习惯
- 淘宝监视着我们的购物习惯
- 微信似乎什么也都知道，不仅窃听到了我们的心思，还能描绘我们的社交关系网

技术层面

数据驱动智能的挑战

- 有效合理的数据采集

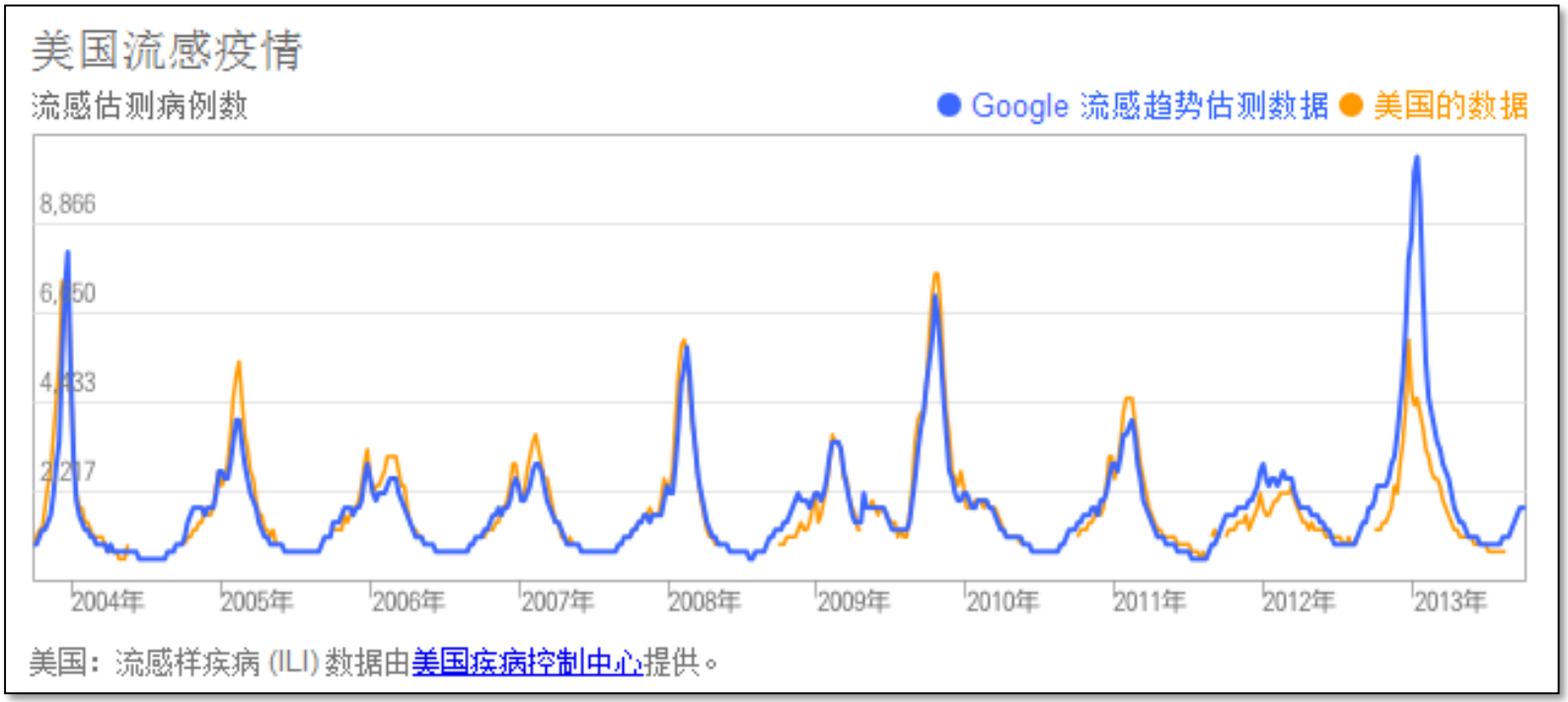


横看成岭侧成峰，远近高低各不同，
不识庐山真面目，只缘身在此山中。 ——苏轼

技术层面

数据驱动智能的挑战

- 处理技术还不成熟（数据的动态特性）



数据驱动智能的挑战

技术层面

数据的多源异构特性

数据驱动智能的核心

关键要素

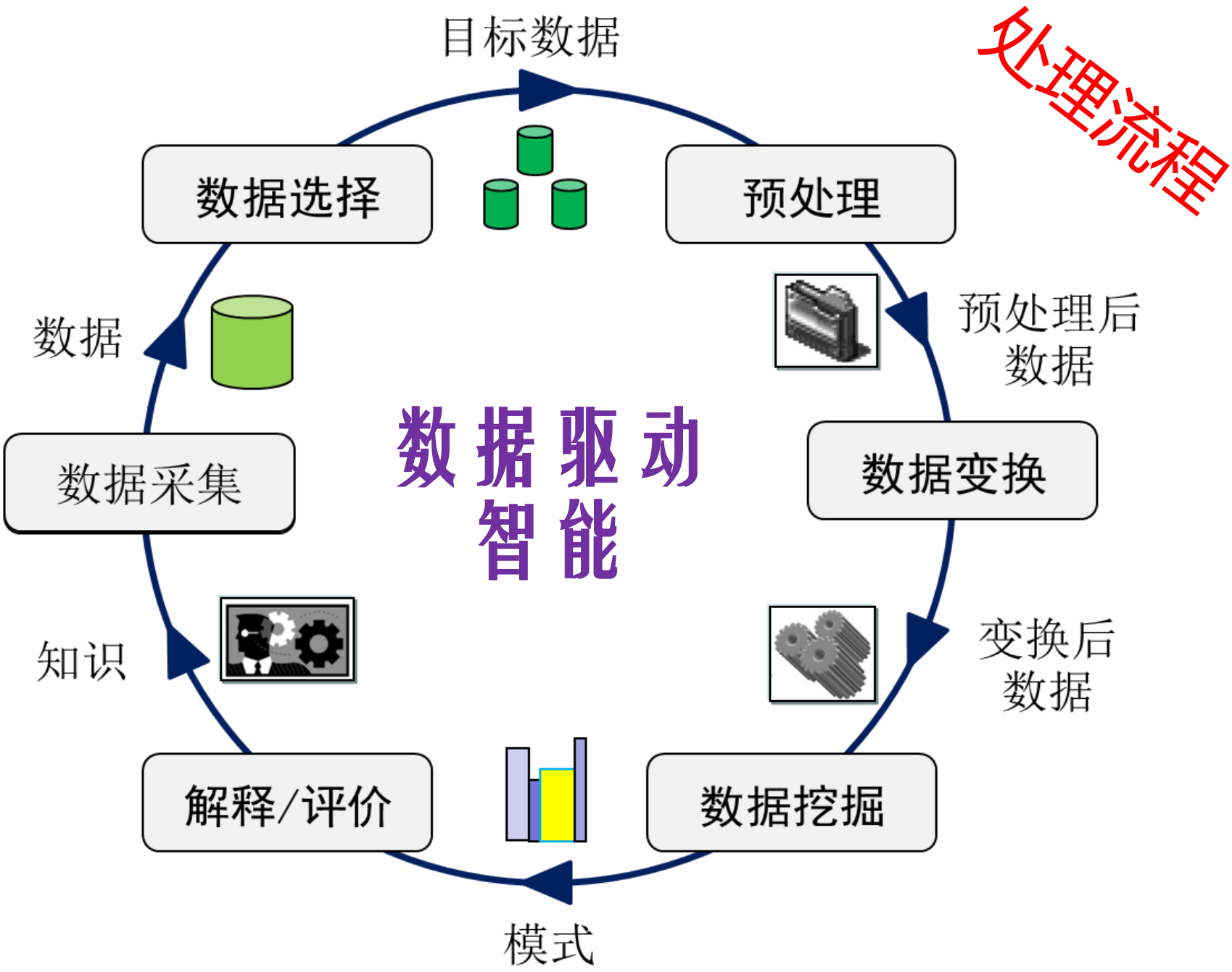
方向盘：行业应用

油：数据

引擎：算法

车轮：计算能力

参考芮勇PPT：AI系统的ABCD



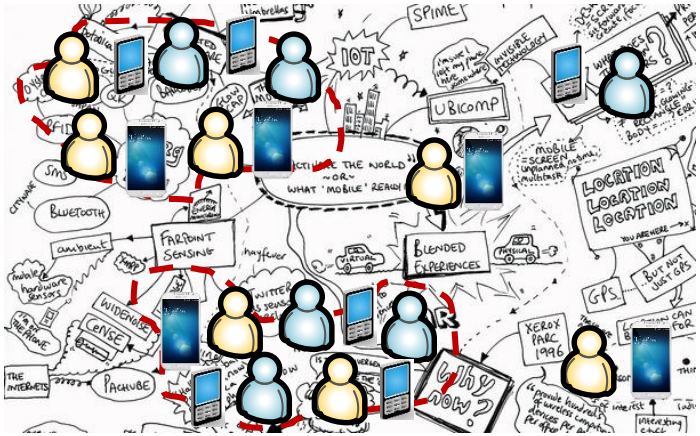
Tianrui Li, Da Ruan, ***An extended process model of knowledge discovery in database***, J. of Enter Inform Management, 2007

数据采集

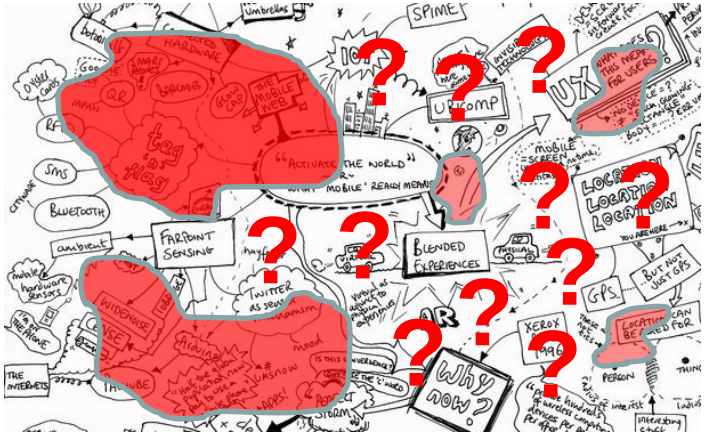
城市感知

- 收集高质量数据
  - 噪声、空气质量等
  - 把人作为传感器
- 研究目的
  - 实时监测与预测
  - 数据分析与应用

人类行为的不规则



数据的不平衡

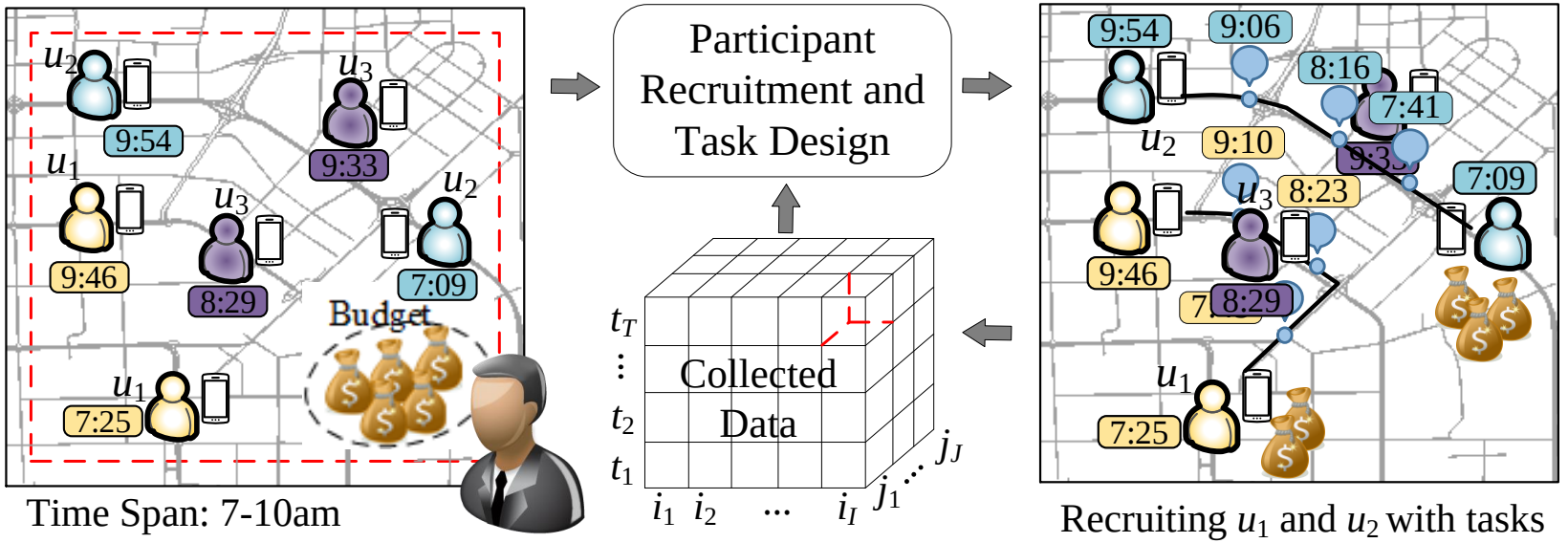


Shenggong Ji, Yu Zheng, Tianrui Li, ***Urban Sensing Based on Human Mobility***. UbiComp 2016 (CCF A)

城市感知

数据采集

- 给定一定的经费，设计一个移动群体感知框架，使得搜集到的数据质量（包括数量和平衡性）最大化

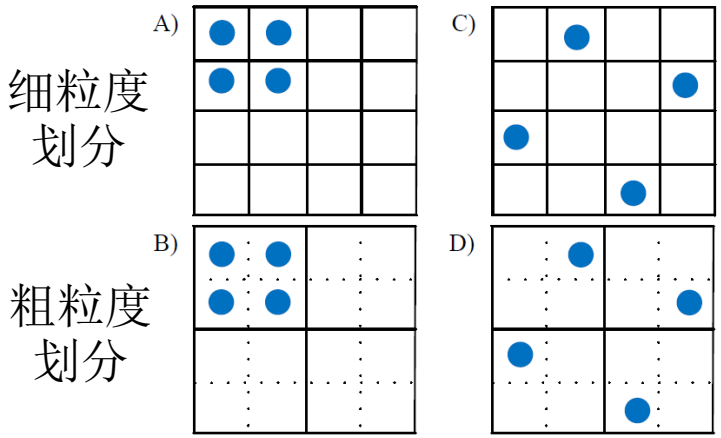


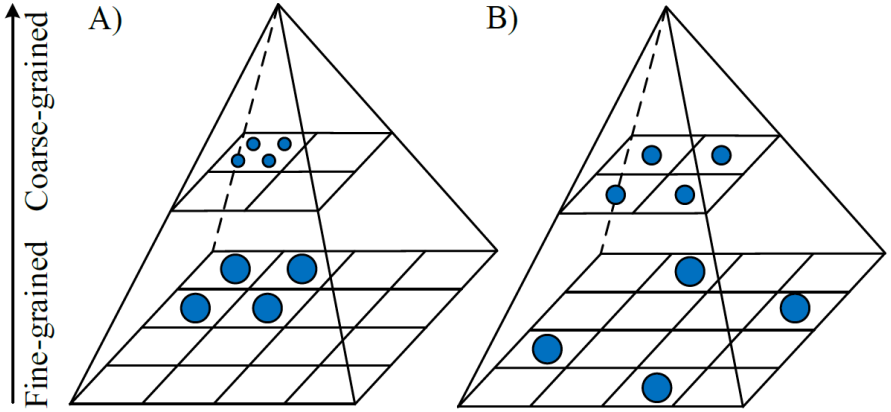
Time Span: 7-10am
Unit reward for each hour

Recruiting u_1 and u_2 with tasks

数据采集

城市感知





层次熵：可用来合理评价采集数据的质量（平衡性）

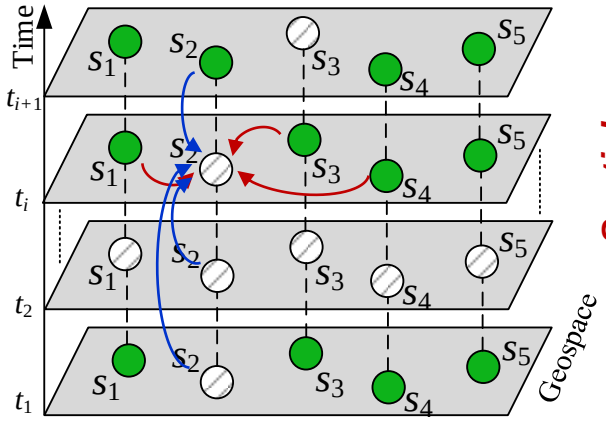
$$E(\mathcal{A}) = \sum_{k=1}^{k_{max}} \omega(k) E(\mathcal{A}(k)) / k_{max} \qquad E(\mathcal{A}(k)) = -\sum_{i,j,t} p(i,j,t|k) \log_2(p(i,j,t|k))$$

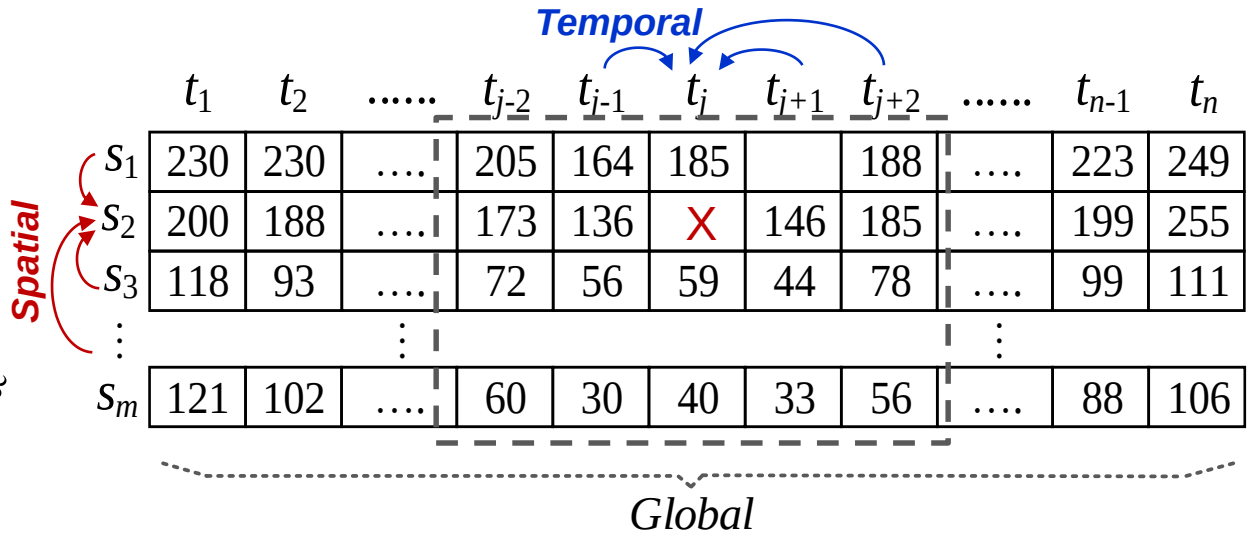
城市感知

预处理

- 时空数据库中的缺失值处理技术
  - 时间和空间角度
  - 全局和局部角度

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



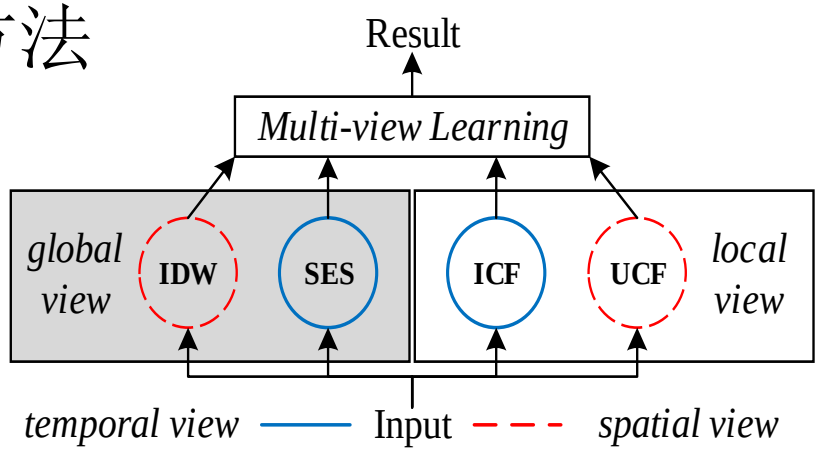


Xiuwen Yi, Yu Zheng, Junbo Zhang, Tianrui Li. ST-MVL: Filling Missing Values in Geo-sensory Time Series Data. IJCAI 2016 (CCF A)

城市感知

预处理

- 基于多视图的缺失值处理方法
  - IDW: Inverse Distance Weighting
  - SES: Simple Exponential Smoothing
  - UCF: User-based Collaborative filtering
  - ICF: Item-based Collaborative filtering
  - MVL: Linear least square



Method	General Missing		Spatial Block Missing		Temporal Block Missing		Sudden Change		Overall	
	MAE	MRE	MAE	MRE	MAE	MRE	MAE	MRE	MAE	MRE
NMF	11.21	0.163	18.98	0.239	12.73	0.217	34.37	0.381	13.08	0.188
NMF-MVL	11.16	0.162	18.97	0.238	12.66	0.217	34.33	0.380	13.06	0.187
ST-MVL	10.81	0.158	17.85	0.217	11.71	0.208	33.15	0.368	12.12	0.174
ST-MVL*							28.98	0.322	10.42	0.149

Xiuwen Yi, Yu Zheng, Junbo Zhang, Tianrui Li. ST-MVL: Filling Missing Values in Geo-sensory Time Series Data. IJCAI 2016 (CCF A)

数据驱动智能的核心

关键要素

方向盘：行业应用

油：数据

引擎：算法

车轮：计算能力

增量学习

iLgC: Incremental Learning Based on Granular Computing

The past

The present

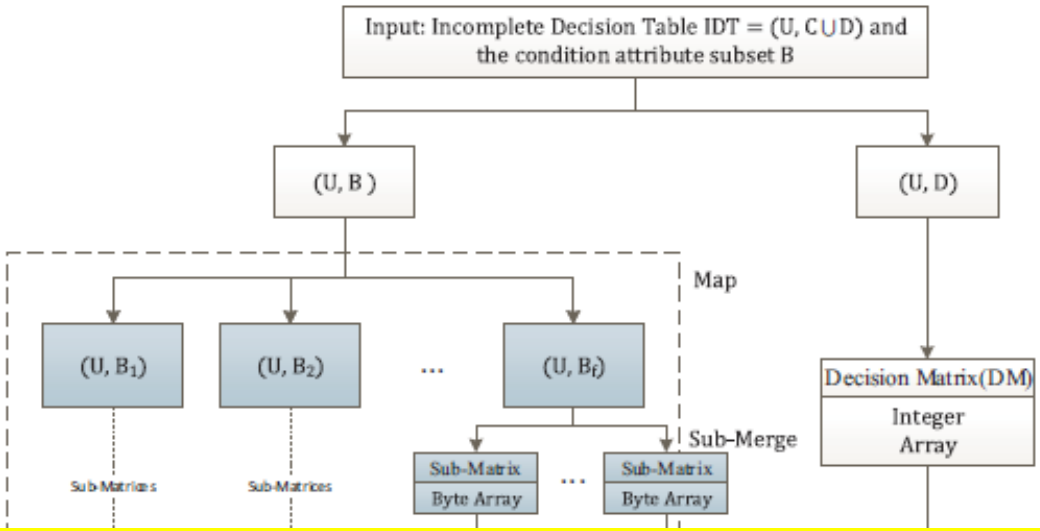
The future

主要贡献

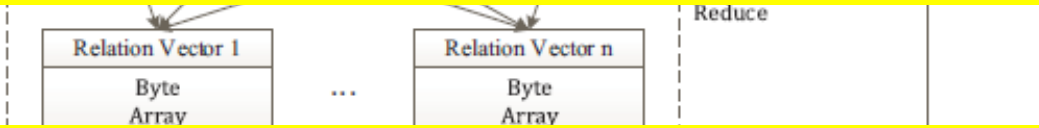
- Dynamic maintenance of approximations, the mining task in Rough Set like frequent pattern in ARM
  - Variation of the object set
    - New patients' records are added
  - Variation of the attribute set
    - New disease features become available
  - Variation of attribute values
    - The feature values may be revised

- A parallel matrix-based method for computing approximations in incomplete information systems (IIS)
  - S1: A MapReduce-based parallel method to construct the relation matrix is designed for fast computing approximations
    - A Sub-Merge operation is used
  - S2: An incremental method is applied to the process of merging the relation matrices.
    - The relation matrix is updated in parallel and incrementally to efficiently accelerate the computational process.
  - S3: A sparse matrix method is employed to optimize the proposed matrix-based method
    - To further improve the performance of the algorithm.

Junbo Zhang, Jian-Syuan Wong, Yi Pan, Tianrui Li, ***A parallel matrix-based method for computing approximations in incomplete information systems***. IEEE Transaction on Knowledge and Data Engineering, 2015

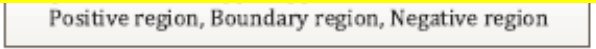


S1: A parallel strategy based on MapReduce.
To reduce space complexity, we use byte arrays to storage the sub-relation matrices.



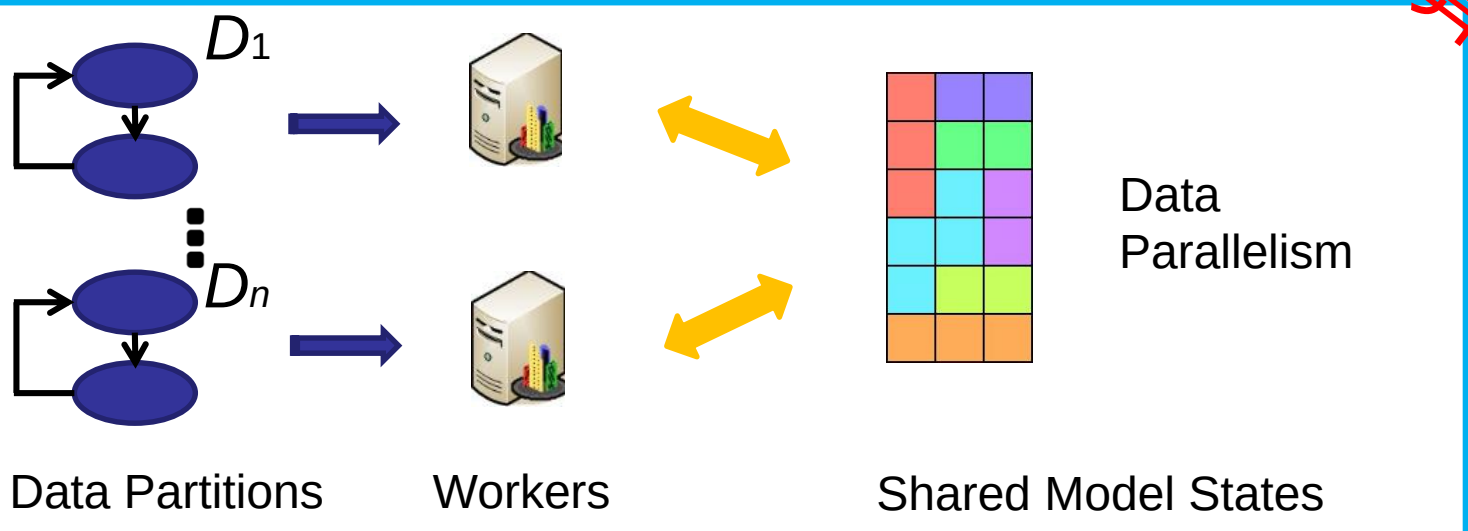
S2: The process of merging can be viewed as a process of adding attributes one by one (A typical incremental process).

S3: As the number of condition attributes increases, there are more and more zero entries in the relation matrix.

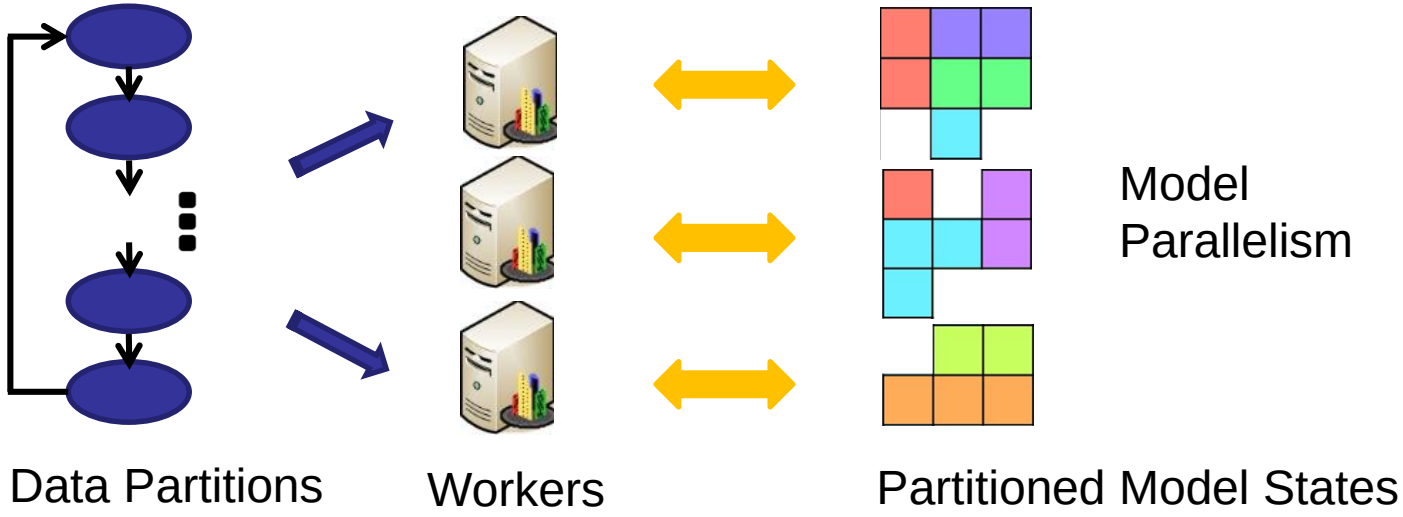


并行处理

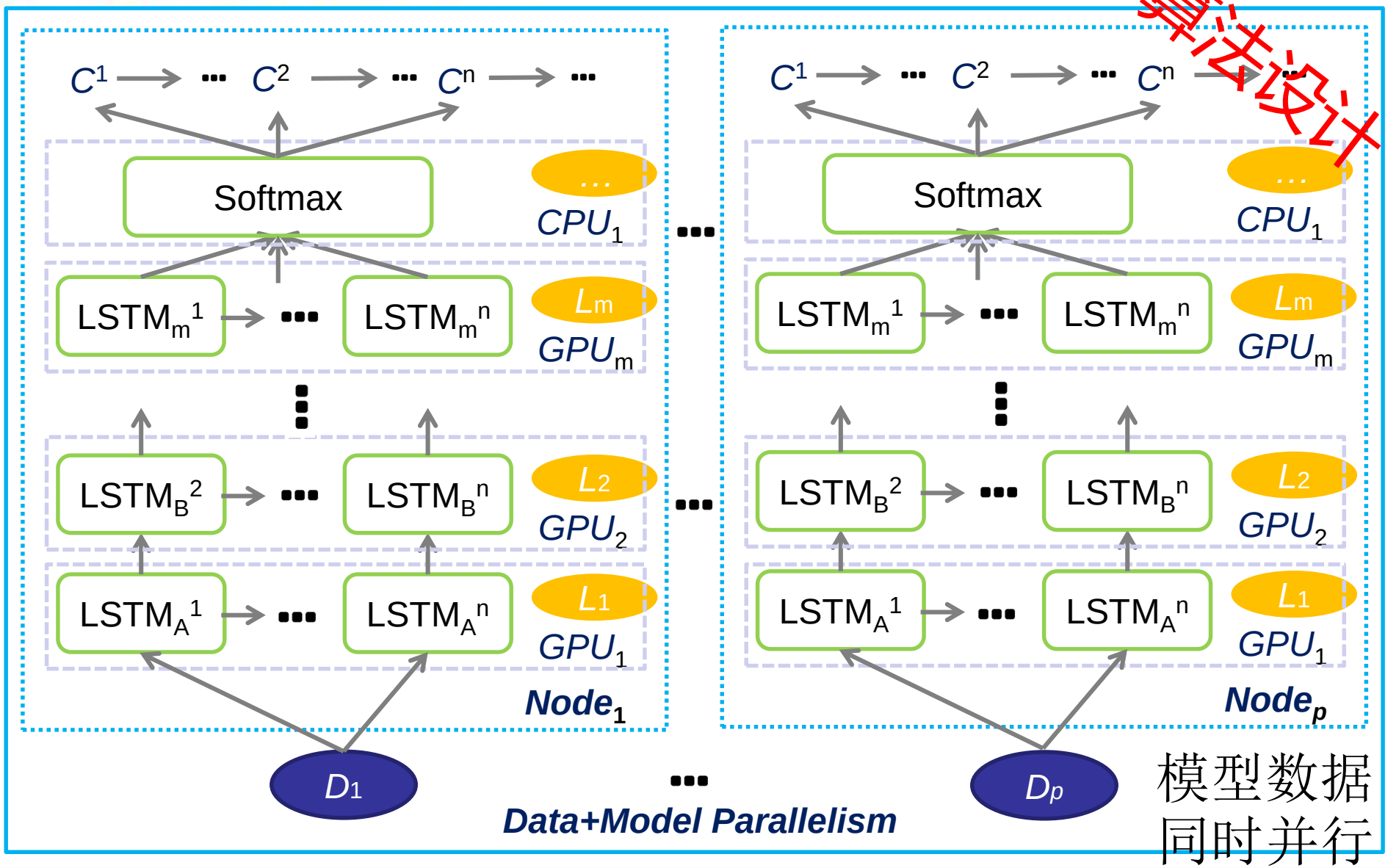
算法设计



数据并行



模型并行


算法设计
C1
C2
Cn
Softmax
CPU1
LSTMm1
LSTMmn
Lm
GPUm
LSTMB2
LSTMBn
L2
GPU2
LSTMA1
LSTMAn
L1
GPU1
Node1
Nodep
D1
Dp
Data+Model Parallelism
模型数据
同时并行

模型并行的特征选择

算法设计

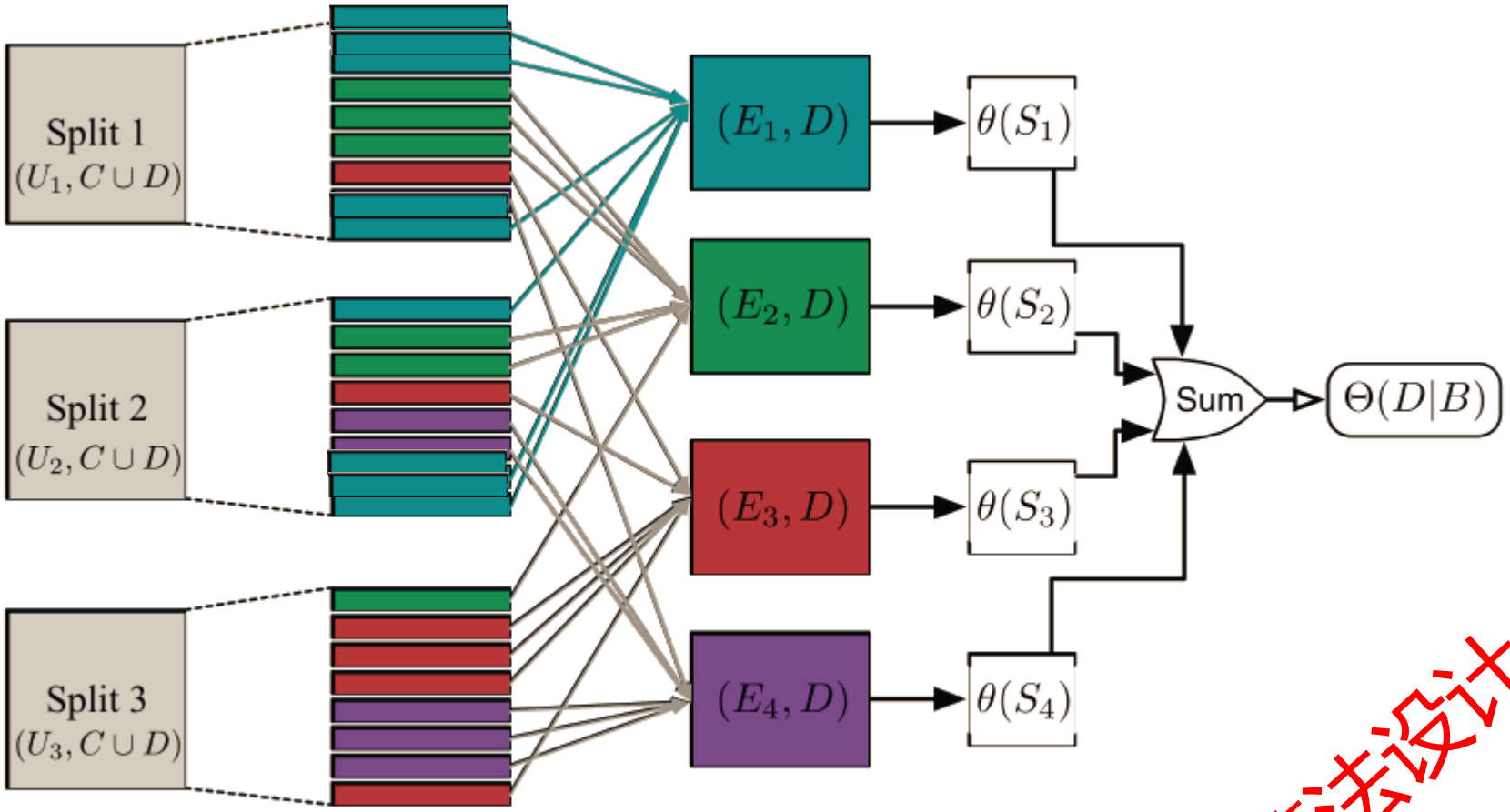
Method	$\Theta(D\|B)$
PR	$\gamma(D\|B) := -\gamma_B(D) = -\frac{\|POS_B(D)\|}{\|U\|}$
SCE	$\mathcal{H}(D\|B) = -\sum_{i=1}^{e} p(E_i) \sum_{j=1}^{m} p(D_j\|E_i) \log(p(D_j\|E_i))$
LCE	$\mathcal{H}_L(D\|B) = \sum_{i=1}^{e} \sum_{j=1}^{m} \frac{\|D_j \cap E_i\|}{\|U\|} \frac{\|D_j^c - E_i^c\|}{\|U\|}$
CCE	$\mathcal{H}_Q(D\|B) = \sum_{i=1}^{e} \left(\frac{\|E_i\|}{\|U\|} \frac{C_{\|E_i\|}^2}{C_{\|U\|}^2} - \sum_{j=1}^{m} \frac{\|E_i \cap D_j$

A unified representation

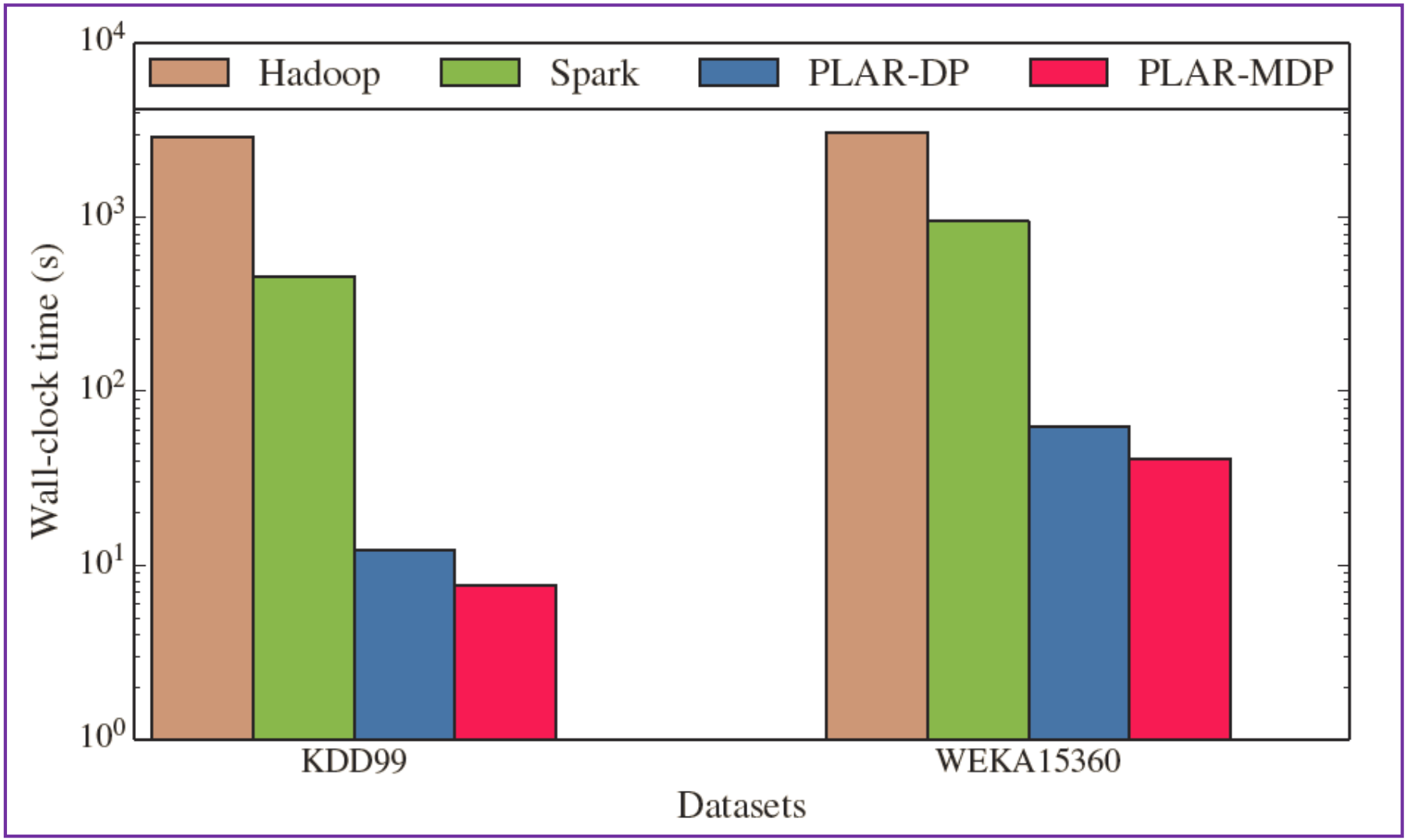
$$\Theta(D|B) = \sum_{i=1}^{e} \theta(S_i)$$

Method	$\theta(S_i)$
PR	$-\frac{\|E_i\| sgn_{PR}(E_i)}{\|U\|}$
SCE	$-\frac{1}{\|U\|} \sum_{j=1}^{m} \|D_{ij}\| \log \frac{\|D_{ij}\|}{\|E_i\|}$
LCE	$\sum_{j=1}^{m} \frac{\|D_{ij}\|(\|E_i\| - \|D_{ij}\|)}{\|U\|^2}$
CCE	$\frac{\|E_i\|^2 \times (\|E_i\| - 1)}{\|U\| C_{\|U\|}^2} - \sum_{j=1}^{m} \frac{\|D_{ij}\|^2 \times (\|D_{ij}\| - 1)}{\|U\| C_{\|U\|}^2}$

模型数据并行的特征选择



算法设计



Parallel Large-scale Attribute Reduction based on Hadoop, Spark and MDP model

特征选择实验结果

算法设计

高维数据上的表现

数据集	Gisette
样本数	6000
特征数	5000

第 i 次迭代	PLAR-DP	PLAR-MDP：模型并行度				
		2	4	8	16	32
1	6262	3080	1570	885	472	350
2	5975	2982	1480	873	465	343
3	6261	3059	1497	869	470	344
4	6115	3017	1484	877	468	344
5	6194	3155	1512	885	465	348
总耗时	30806	15293	7543	4389	2340	1730

实际大数据中的表现

数据集	天文大数据 SDSS
样本数	320000
特征数	5201

算法	128 核	32 核
PR	7432	24274
SCE	7312	24181
LCE	7207	24372
CCE	7383	24295

关键要素

数据驱动智能的核心

方向盘：行业应用

油：数据

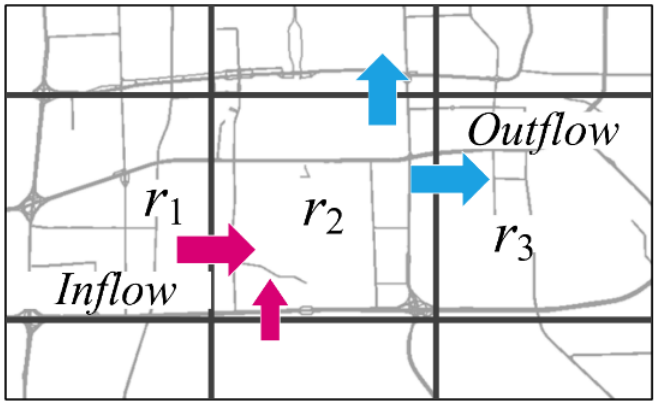
引擎：算法

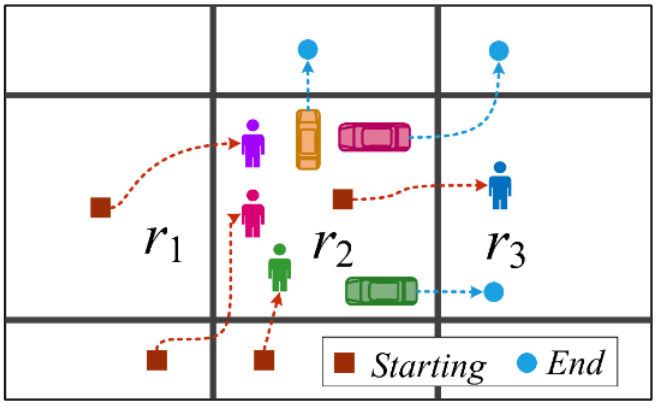
车轮：计算能力

行业应用

城市人流预测

- 目标是预测整个城市里每个区域在未来时刻有多少人进、有多少人出。
- 目的是使管理者能迅速了解每个区域公共安全状况，及时采取预警措施。



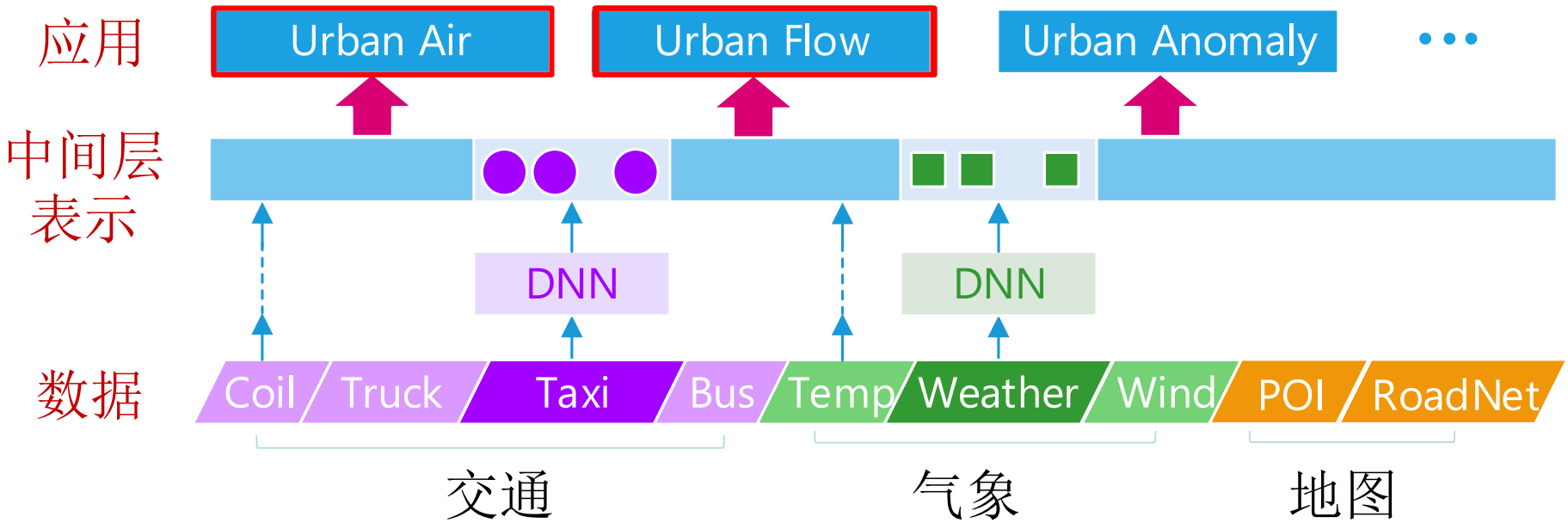


Junbo Zhang, Yu Zheng, Dekang Qi, Ruiyuan Li, Xiuwen Yi, Tianrui Li, Predicting Citywide Crowd Flows Using Deep Spatio-Temporal Residual Networks, Artificial Intelligence, 2018 (CCF A)

城市人流预测

行业应用

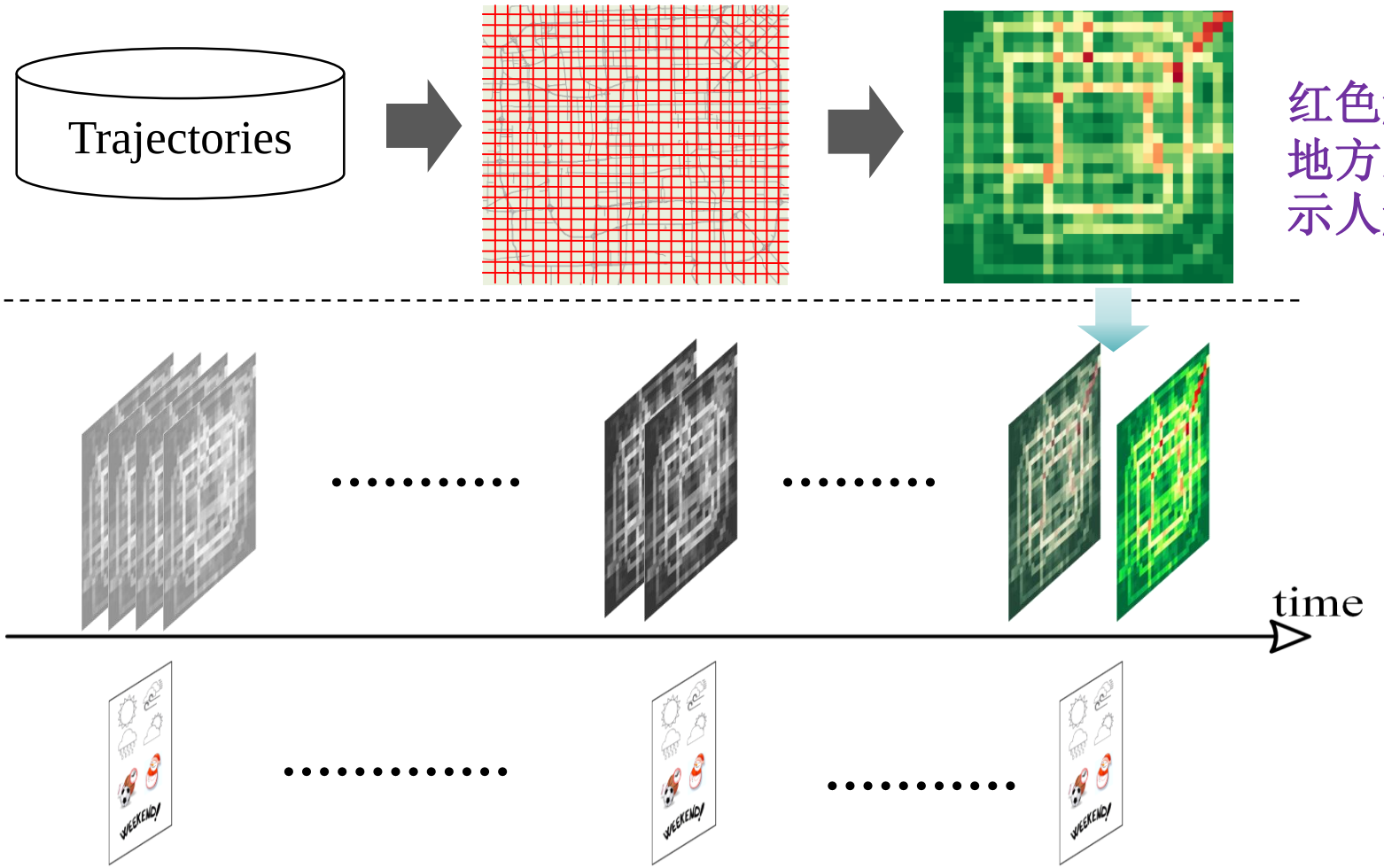
- 不同数据源的融合

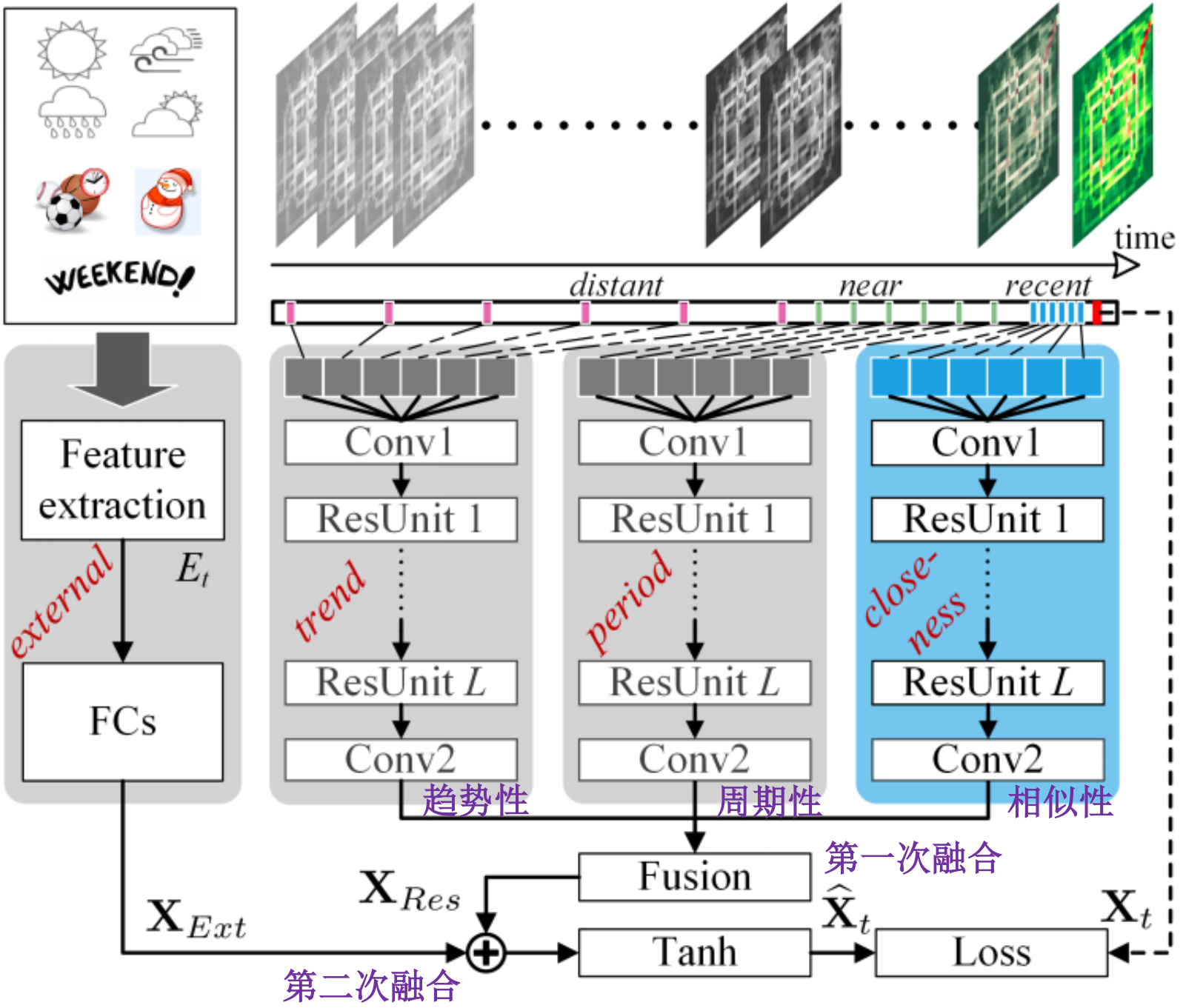


Junbo Zhang, Yu Zheng, Dekang Qi, Ruiyuan Li, Xiuwen Yi, Tianrui Li, Predicting Citywide Crowd Flows Using Deep Spatio-Temporal Residual Networks, Artificial Intelligence, 2018 (CCF A)

城市人流预测

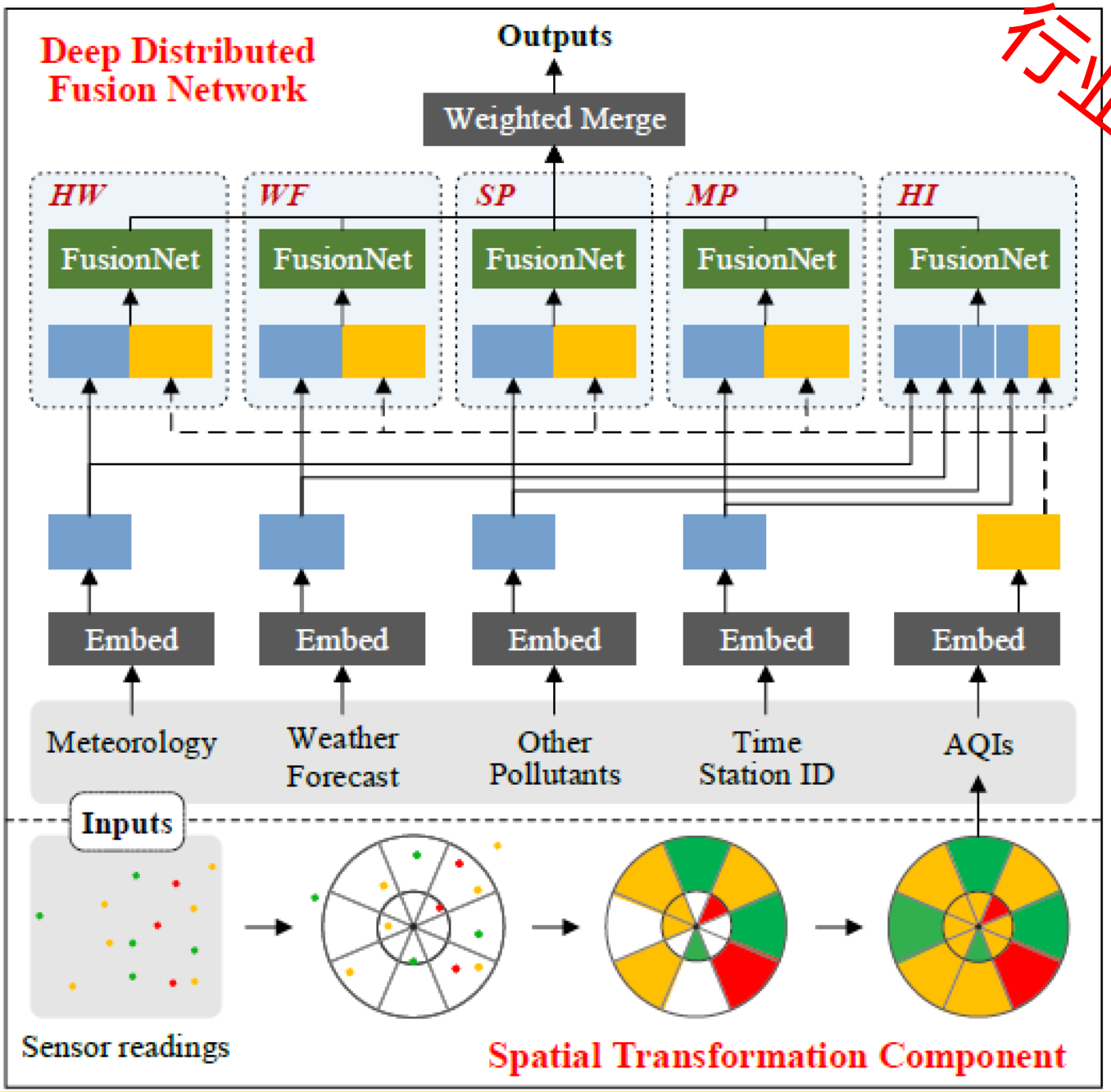
行业应用




WEEKEND!
time
distant
near
recent
Feature extraction
external
E_t
FCs
Conv1
ResUnit 1
trend
ResUnit L
Conv2
Conv1
ResUnit 1
period
ResUnit L
Conv2
Conv1
ResUnit 1
close-ness
ResUnit L
Conv2
趋势性
周期性
相似性
Fusion
第一次融合
$\mathbf{X}_{Res}$
$\mathbf{X}_{Ext}$
$\hat{\mathbf{X}}_t$
$\mathbf{X}_t$
Tanh
Loss
第二次融合
深度时空残差网络

空气质量预测

行业应用



Xiuwen Yi, Junbo Zhang, Zhaoyuan Wang, Tianrui Li and Yu Zheng, Deep Distributed Fusion Network for Air Quality Prediction, KDD

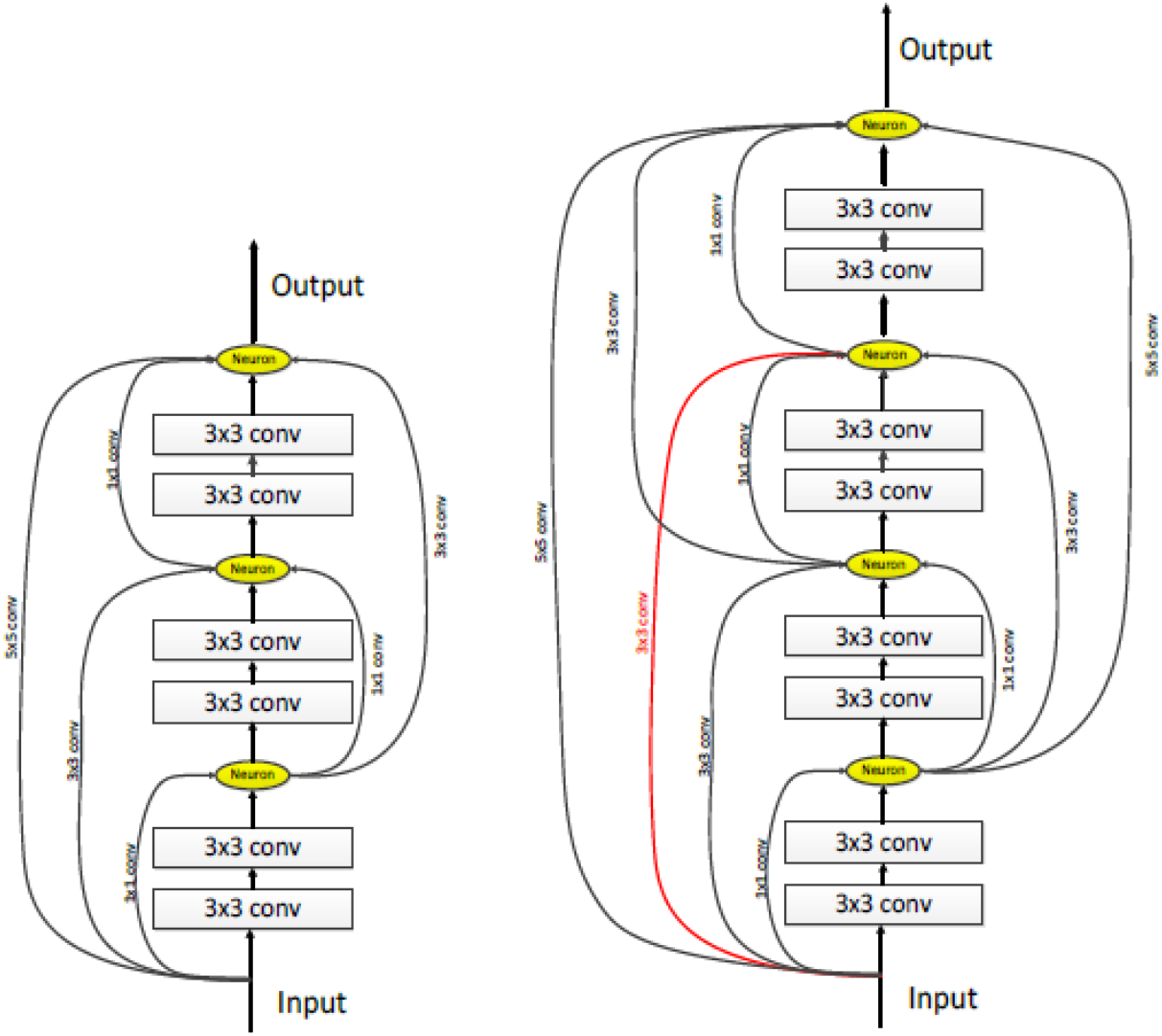


Figure 1: Cross-layer architectures.

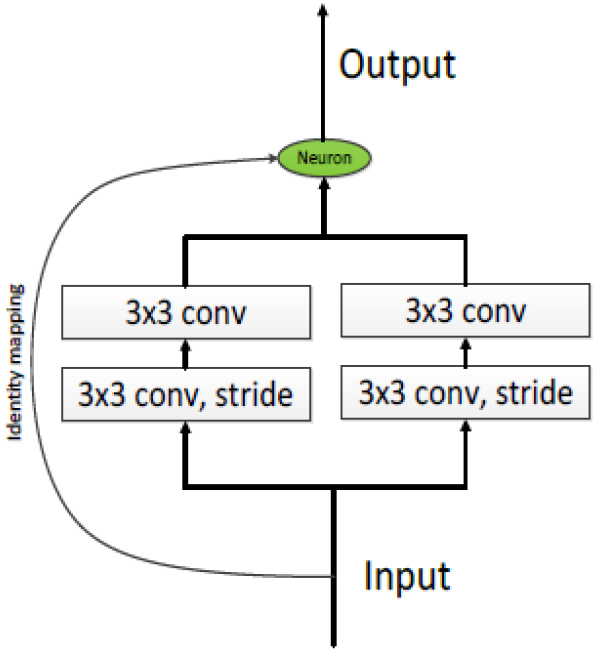
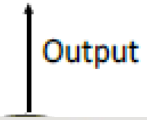


Figure 2: Connection-layer architecture

Zeng Yu, Ning Yu, Yi Pan, **Tianrui Li**: A Novel Deep Learning Network Architecture with Cross-Layer Neurons. BDCloud-SocialCom-SustainCom 2016: 111-117, Atlanta, USA, October 8-10, 2016

Zeng Yu, **Tianrui Li**, et al., Convolutional networks with cross-layer neurons for image recognition, Information Sciences, 2018



Paper 48	
Title:	A Novel Deep Learning Network Architecture with Cross-Layer Neurons
Paper:	(Jun 01, 00:18 GMT)
Author keywords:	deep learning convolutional networks cross-layer architecture cross-layer neurons
EasyChair keyphrases:	cross layer (535), cross layer neuron (507), identity mapping (307), cross layer architecture (158), lower level (150), layer neuron network (142), level layer (140), layer neuron (135), connection layer (130), mapping identity mapping (110), identity mapping identity (110), deep residual network (110), convolution filter (100), x1 conv neuron (79), classification error rate (79), identity mapping neuron (79), deep convolutional network (79), upper layer (70), convolutional layer (60), conv neuron (60), deeper convnet (50), computer vision (50), deep network (50), neural network (50), cross layer increase (47), mini batch size (47), x8 avg pool (47), imagenet large scale visual recognition (46), large scale visual recognition challenge (46), neural information processing system (40)
Abstract:	Very deep convolutional networks have recently demonstrated impressive classification performance on competitive benchmarks such as the ImageNet or COCO tasks. However, training such deep convolutional networks becomes more difficult. In this paper, we propose a novel deep network structure called cross-layer architecture to make the best use of information learned from all the lower-level layers. It utilizes cross-layer neurons to control all the lower-level layers to gather and send information to the upper layers. It shows that our new architecture can be applied in training deeper networks. The classification performance on CIFAR-10 benchmark dataset demonstrates that our new architecture can effectively improve the performance by using cross-layer neurons. Particularly, our new architecture can achieve up to 16.17% relative accuracy improvement compared to the state-of-the-art methods.
Submitted:	Jun 01, 00:18 GMT
Last update:	Jun 01, 00:18 GMT

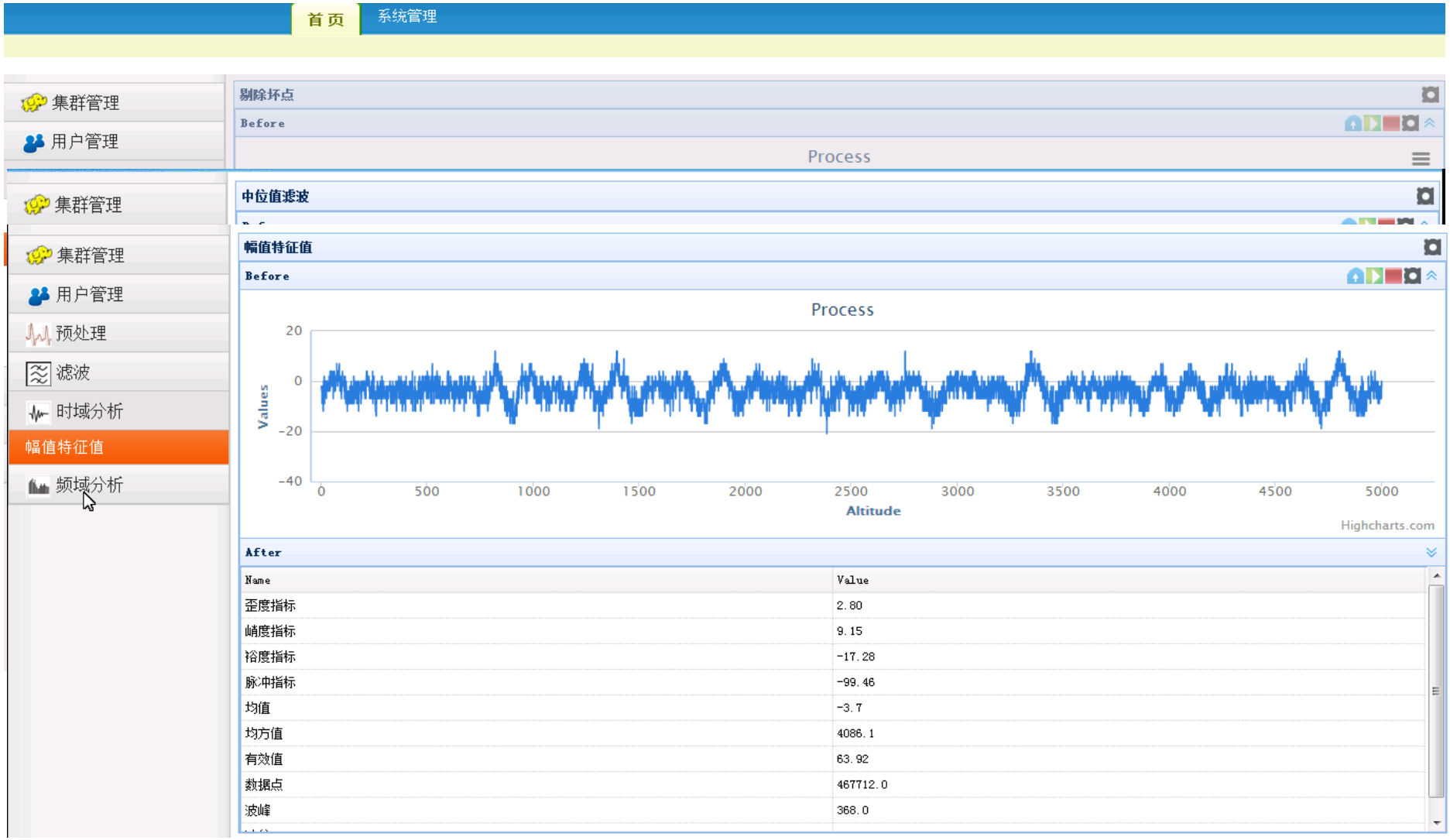
Figure 1: Cross-layer architectures.

Zeng Yu, Ning Yu, Yi Pan, **Tianrui Li**: A Novel Deep Learning Network Architecture with Cross-Layer Neurons. BDCloud-SocialCom-SustainCom 2016: 111-117, Atlanta, USA, October 8-10, 2016

Zeng Yu, **Tianrui Li**, et al., Convolutional networks with cross-layer neurons for image recognition, Information Sciences, 2018

高铁交通大数据+云计算+人工智能

高铁大数据云处理系统



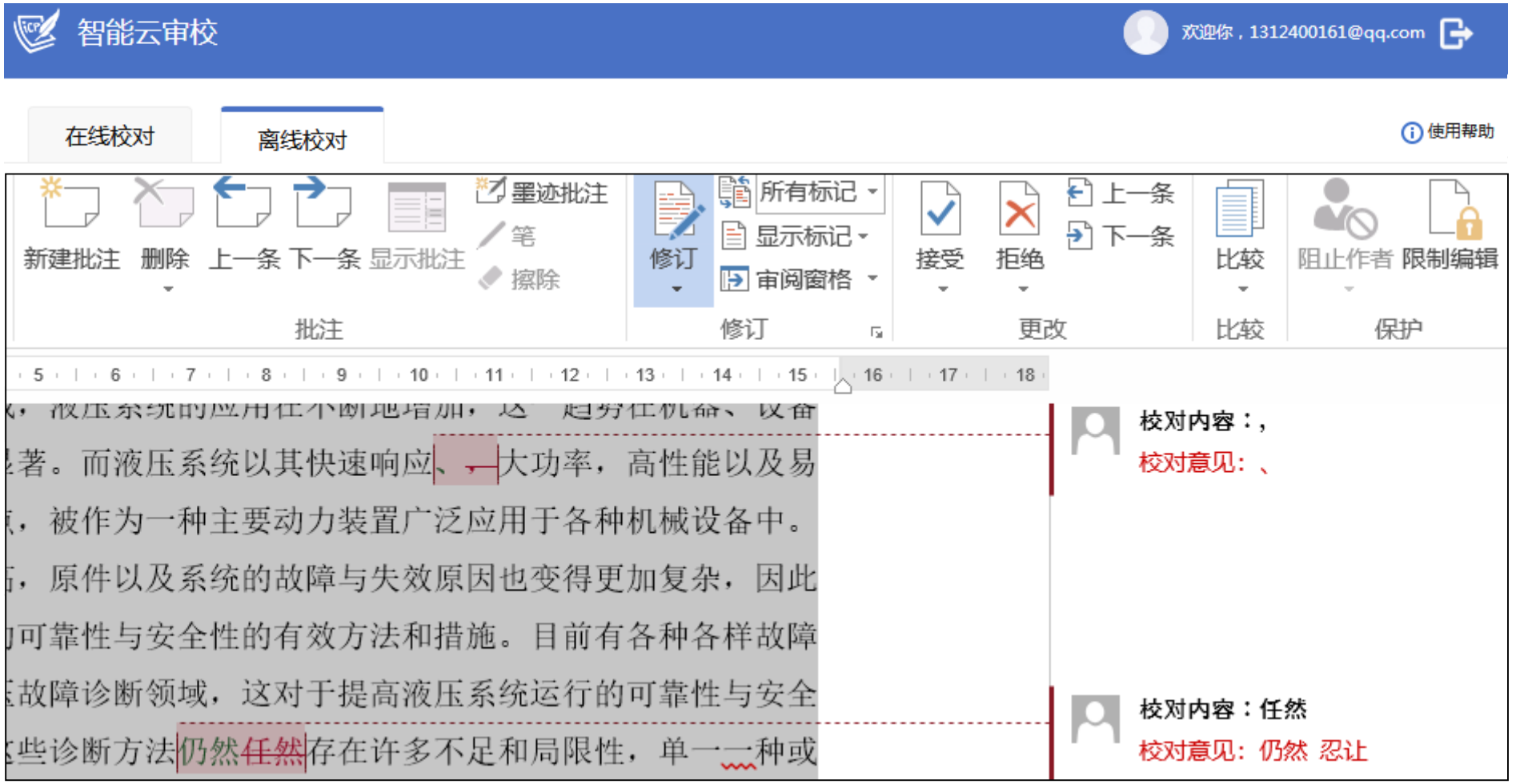
综合交通大数据+人工智能+可视化

空铁联运

智慧出行可视化决策支持平台

大数据+云计算+人工智能+自然语言处理

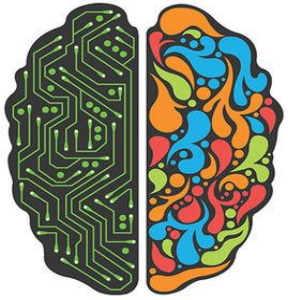
智能云审校（iproofread.cn）



数据驱动智能的变革

四维一体

形象思维



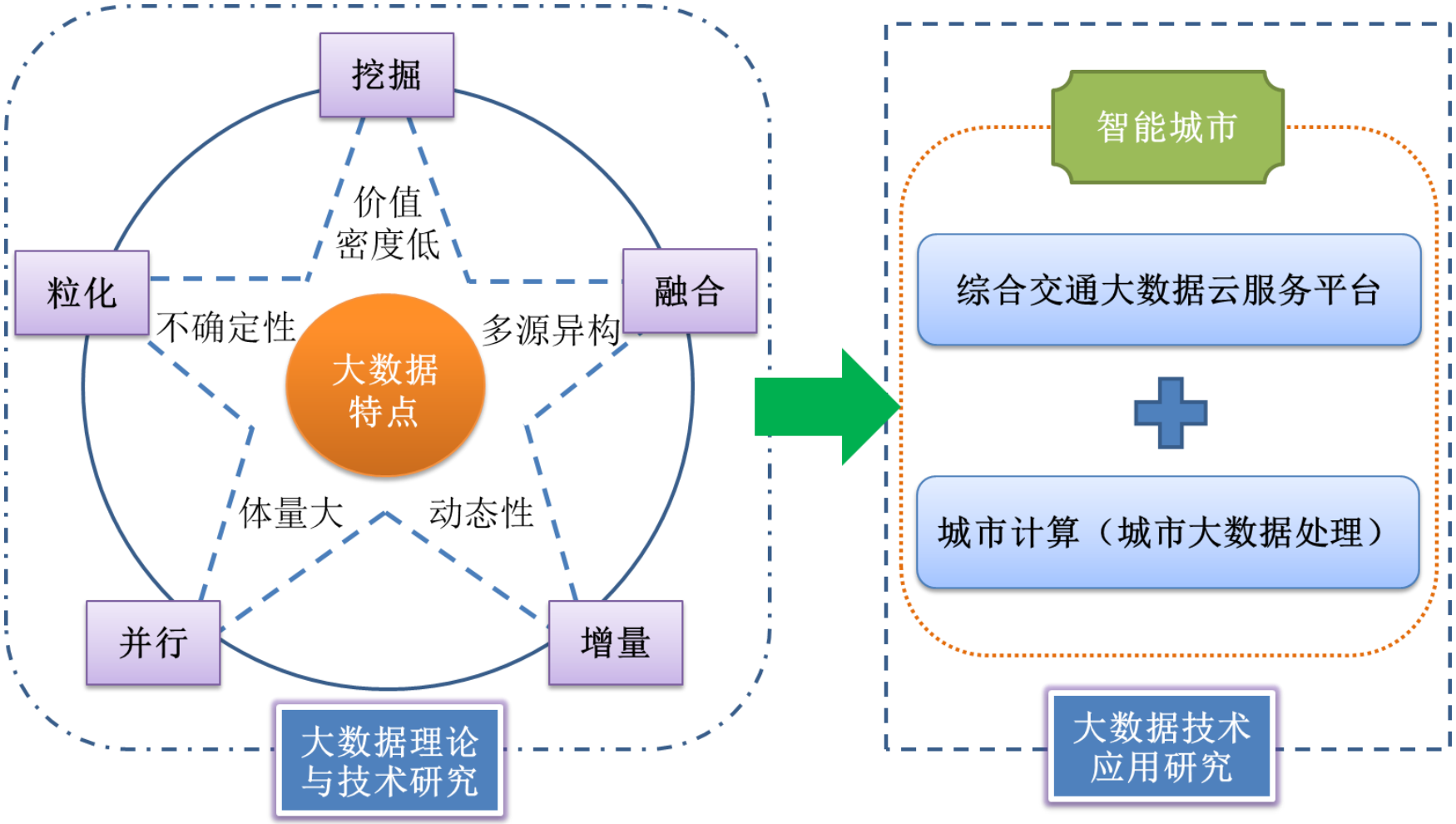
抽象思维

计算思维

数据思维

运用数据科学的概念进行问题求解、决策分析等涵盖数据科学之广度的一系列思维活动。

数据驱动智能的工作进展



研究成果

数据驱动智能的工作进展

国际刊物

- IEEE Trans. on PAMI
- IEEE Trans. on Know. & Data Eng.
- IEEE Trans. on Evolutionary Comp.
- IEEE Trans. on Image Processing
- IEEE Trans. on Cybernetics
- IEEE Trans. on Inf. Foren. and Sec.
- IEEE Trans. on Fuzzy Systems
- ……

国内刊物

- 中国科学
- 软件学报
- 计算机学报
- 计算机研究与发展
- ……

入选ESI热点论文3篇和高被引论文12篇，ESI扩展版论文21篇（占全校1/3），承担国际合作项目2项，国家科技支撑计划、国家自然科学基金等国家级项目26项，省部级项目9项

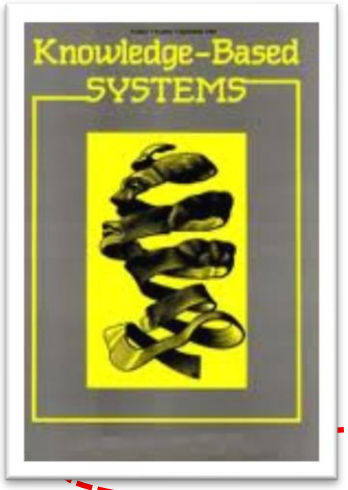
研究成果

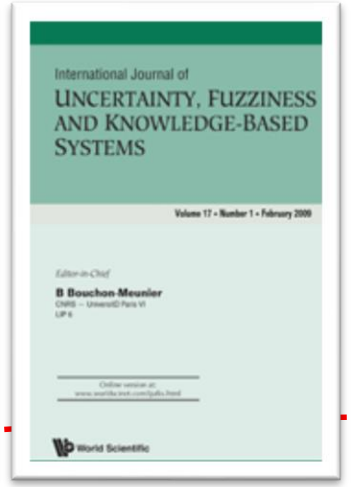
出版的专著

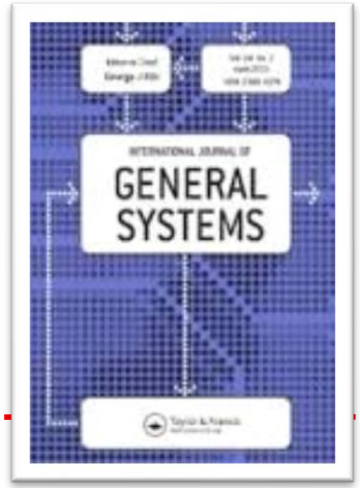

Chapman & Hall/CRC
Numerical Analysis and Scientific Computing
Cloud-Computing
Data-Intensive Computing and Scheduling
Frédéric Magoulès, Jie Pan and Fei Teng
CRC Press
粒计算研究丛书
三支决策与粒计算
粒计算研究丛书
动态知识发现与三支决策
——基于优势粗糙集视角
粒计算研究丛书
大数据挖掘的原理与方法
——基于粒计算与粗糙集的视角
李天瑞 罗川 陈红梅 张钧波 著
三支决策理论与应用
南京大学出版社
决策粗糙集理论及其研究进展
Human Face Perception: from 2D to 3D
从二维到三维

研究成果

SCI期刊客座编辑







ELSEVIER

World Scientific
www.worldscientific.com

Taylor & Francis Online

ATLANTIS PRESS

IOS Press

标志性成果

人才培养

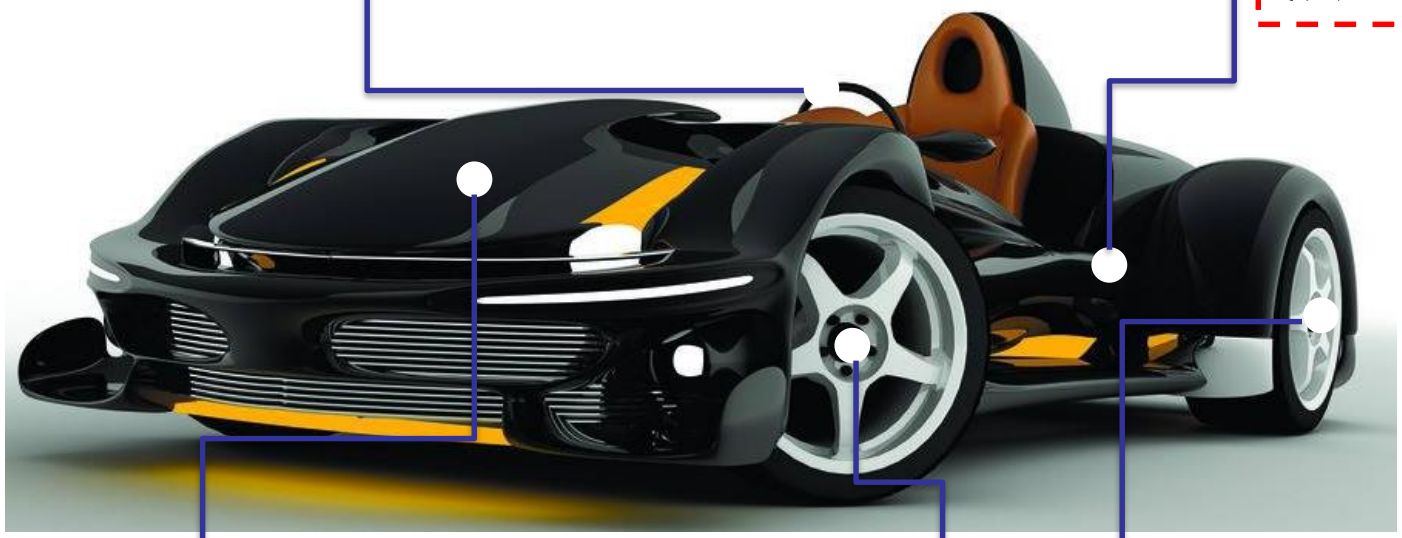
数据驱动智能的工作进展

- 天池大数据**新浪微博互动预测大赛**冠军（奖金20万元）
  - 全球40个国家和地区、1541所学校、29212名选手报名参赛
  - 算法已经在新浪微博部署上线运营
- 国际人工智能联合会IJCAI2016**社会影响力大数据分析竞赛**亚军

数据驱动智能的核心

方向盘：行业应用

油：数据



引擎：算法

车轮：计算能力

Reference

- Predicting Citywide Crowd Flows Using Deep Spatio-Temporal Residual Networks, **Artificial Intelligence**. 2018
- An Iterative Locally Auto-Weighted Least Squares Method for Microarray Missing Value Estimation, **IEEE Transactions on Nanobioscience**, 2017
- A Modified Ant Colony Optimization Algorithm for Network Coding Resource Minimization, **IEEE Transactions on Evolutionary Computation**, 2016
- A parallel matrix-based method for computing approximations in incomplete information systems, **IEEE Transactions on Knowledge and Data Engineering**, 2015
- A rough set-based method for updating decision rules on attribute values' coarsening and refining, **IEEE Transactions on Knowledge and Data Engineering**, 2014
- A rough-set based incremental approach for updating approximations under dynamic maintenance environments. **IEEE Transactions on Knowledge and Data Engineering**, 2013

Reference

- Efficient Parallel Boolean Matrix Based Algorithms for Computing Composite Rough Set Approximations, **Information Sciences**, 2016
- Fast algorithms for computing rough approximations in set-valued decision systems while updating criteria values, **Information Sciences**, 2015 (ESI高被引论文)
- Incremental Update of Approximations in Dominance-based Rough Sets Approach under the Variation of Attribute Values, **Information Sciences**, 2015 (ESI高被引论文)
- Dynamic Maintenance of Approximations in Set-valued Ordered Decision Systems under the Attribute Generalization. **Information Sciences**, 2014 (ESI高被引论文)
- Composite Rough Sets for Dynamic Data Mining. **Information Sciences**, 2014 (ESI热点论文，ESI高被引论文)
- A Parallel Method for Computing Rough Set Approximations. **Information Sciences**, 2012
- Probabilistic model criteria with decision-theoretic rough sets. **Information Sciences**, 2011 (ESI高被引论文)

Reference

- Hierarchical cluster ensemble model based on knowledge granulation, **Knowledge-based Systems**, 2016 (ESI高被引论文)
- Matrix approach to decision-theoretic rough sets for evolving data, **Knowledge-based Systems**, 2016 (ESI高被引论文)
- Update of Approximations in Composite Information Systems, **Knowledge-based Systems**, 2015
- Incremental Updating Approximations in Probabilistic Rough Sets under the Variation of Attributes, **Knowledge-based Systems**, 2015 (ESI高被引论文)
- Incremental Updating Approximations in Dominance-based Rough Sets Approach under the Variation of the Attribute Set, **Knowledge-based Systems**, 2013 (ESI高被引论文)
- Incremental Updating Approximations in Dominance-based Rough Sets Approach under the Variation of the Attribute Set, **Knowledge-based Systems**, 2013

Reference

- Incorporating logistic regression to decision-theoretic rough sets for classifications, **International Journal of Approximate Reasoning**, 2014 (ESI高被引论文)
- A Comparison of Parallel Large-scale Knowledge Acquisition using Rough Set Theory on Different MapReduce Runtime Systems. **International Journal of Approximate Reasoning**, 2014 (ESI高被引论文)
- Dynamic Maintenance of Approximations in Dominance-based Rough Set Approach under the Variation of the Object Set, **International Journal of Intelligent Systems**, 2013
- Rough Sets Based Matrix Approaches with Dynamic Attribute Variation In Set-valued Information Systems, **International Journal of Approximate Reasoning**, 2012
- Neighborhood rough sets for dynamic data mining. **International Journal of Intelligent Systems**, 2012

Reference

- Deep Distributed Fusion Network for Air Quality Prediction, **KDD 2015**
- ST-MVL: Filling Missing Values in Geo-sensory Time Series Data, **IJCAI 2016**
- Urban Sensing Based on Human Mobility, **UbiComp 2016**
- Forecasting Fine-Grained Air Quality Based on Big Data, **KDD 2015**
- Supervised Deep Learning with Auxiliary Networks, **KDD 2014**

- 基于矩阵运算的复杂网络构建方法研究, **中国科学**, 2016
- 云平台下基于粗糙集的并行增量知识更新算法, **软件学报**, 2015
- 大数据环境下移动对象自适应轨迹预测模型. **软件学报**, 2015

Reference

- **Special Issue**: Three-way Decisions and Granular Computing, **KBS**, 2016
- **Special Issue**: Computational Technique in Data Science, **IJIS**, 2015
- **Special Issue**: Advances on Rough Sets and Knowledge Technology, **FI**, 2014
- **Special Issue**: Fuzziness in Systems Modelling, **IJGS**, 2013
- **Special Issue**: Computational Intelligence for Policy Making and Risk Governance, **IJUFKBS**, 2012
- **Special Issue**: Computational Intelligence in Data Analysis, **IJCIS**, 2012
- **Special Issue**: New Trends on Intelligent Decision Support Systems, **KBS**, 2012
- **Special Issue**: Computational Intelligence in Decision Making, **IJCIS**, 2011

Special Issues

Reference

- 陈红梅, 李少勇, 罗川, **李天瑞**, 动态知识发现与三支决策: 基于优势粗糙集视角, 科学出版社, 2017
- **李天瑞**, 罗川, 陈红梅, 张钧波, 大数据挖掘的原理与方法——基于粒计算与粗糙集的视角, 科学出版社, 2016
- 于洪, 王国胤, **李天瑞**等, 三支决策: 复杂问题求解方法与实践, 科学出版社, 2015
- 刘盾, **李天瑞**, 苗夺谦等, 三支决策与粒计算, 科学出版社, 2013
- 贾修一, 商琳, 周献中, 梁吉业, 苗夺谦, 王国胤, **李天瑞**等. 三支决策理论与应用, 南京大学出版社, 2012
- 李华雄, 周献中, **李天瑞**, 王国胤, 苗夺谦, 决策粗糙集理论及其研究进展, 科学出版社, 2011

出版的编者