AI大模型在数字城市领域的探索和思考

演讲人：郑宇 教授、博士

IEEE Fellow，ACM杰出科学家、KDD China主席
京东集团副总裁、首席数据科学家
国家万人计划科技创新领军人才
http://urban-computing.com/yuzheng

AI大模型（AI模型、AI大模型、AI大语言模型）

人工智能是研究使用计算机来模拟人的某些思维过程和智能行为（如学习、推理、思考、规划等）的学科

机器学习是人工智能的重要基础和分支：符号学习 和 统计学习

		AI模型	AI大模型
统计机器学习		传统学习模型	深度学习模型
	判别式模型 P(Y|X)	逻辑回归（LR） 支持向量机（SVM） 决策树（DT）	深度神经网络（DNN） 深度卷积网络（CNN） 递归神经网络（RNN）
	生成式模型 P(X,Y)	朴素贝叶斯(Naïve Bayes) 高斯混合模型(GMMs) 隐马尔可夫模型(HMM) 线性判别分析 (LDA)	自编码器（AE） 生成式对抗网络（GAN） GPT（Generative Pre-trained Transformer）AI大语言模型(LLM)

AI大语言模型向数字城市领域的演进

自然语言处理（AI大语言模型）

场景驱动

自然语言的意图理解成为瓶颈；人机对话正好是一个文本的顺序生成过程；出错容忍度高；

模型架构

深度学习方法已经成熟，Transformer架构异军突起，连结和压缩更大的语义空间

强大算力

针对Transformer架构优化的GPU芯片，云计算提供了集群计算的基础设施

海量数据

除了互联网积累的大数据，大量书籍也完成了电子化（数据分布与语言对话基本相同）

数字城市领域

场景驱动

影响因素多、推理复杂度高；场景已经沉淀大量可用数据；影响因素时间、空间跨度大；对单条结果的可靠性要求较低；

模型架构

捕捉时空特性；多源、多模态数据融合；
行业知识注入；

强大算力

对海量时空数据的管理和分析能力尚有不足，数据进入模型前的预处理能力需要加强

海量数据

城市数据体量不小，但割裂孤立、且数据资源不一致，无法互认；数据分布和语义不同；

1. 构建海量城市数据（集）

数据孤岛突显

- 不仅缺少数据的联通和汇聚
- 更缺乏一致的数据资源体系

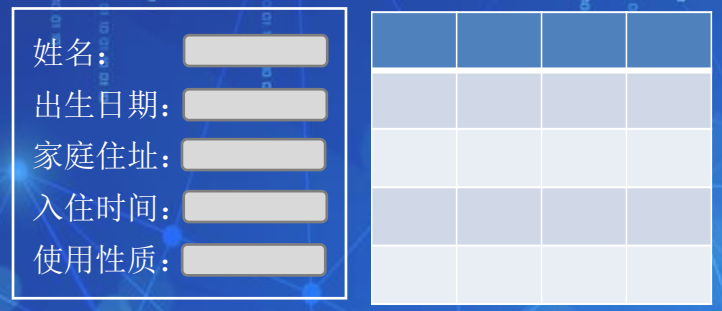
挑战

数据分布不同

- 时空数据的分布与互联网数据不同
- 政务数据表达方式与自然语言不同

城市数据本身极大丰富

结构化数据



以电子政务表单为代表

非结构化数据

以视频、语音和文本为代表

时空数据

以地理信息和物联网数据为代表

1. 构建海量城市数据：数据孤岛凸显

数字化控件

城市知识体系

数据孤岛问题

业界已形成了通用制式

可在不同系统间调用和共享

问题不突出

六大时空数据模型

城市知识体系

数据孤岛问题

结构化数据

姓名：
出生日期：
家庭住址：
入住时间：
使用性质：

非结构化数据

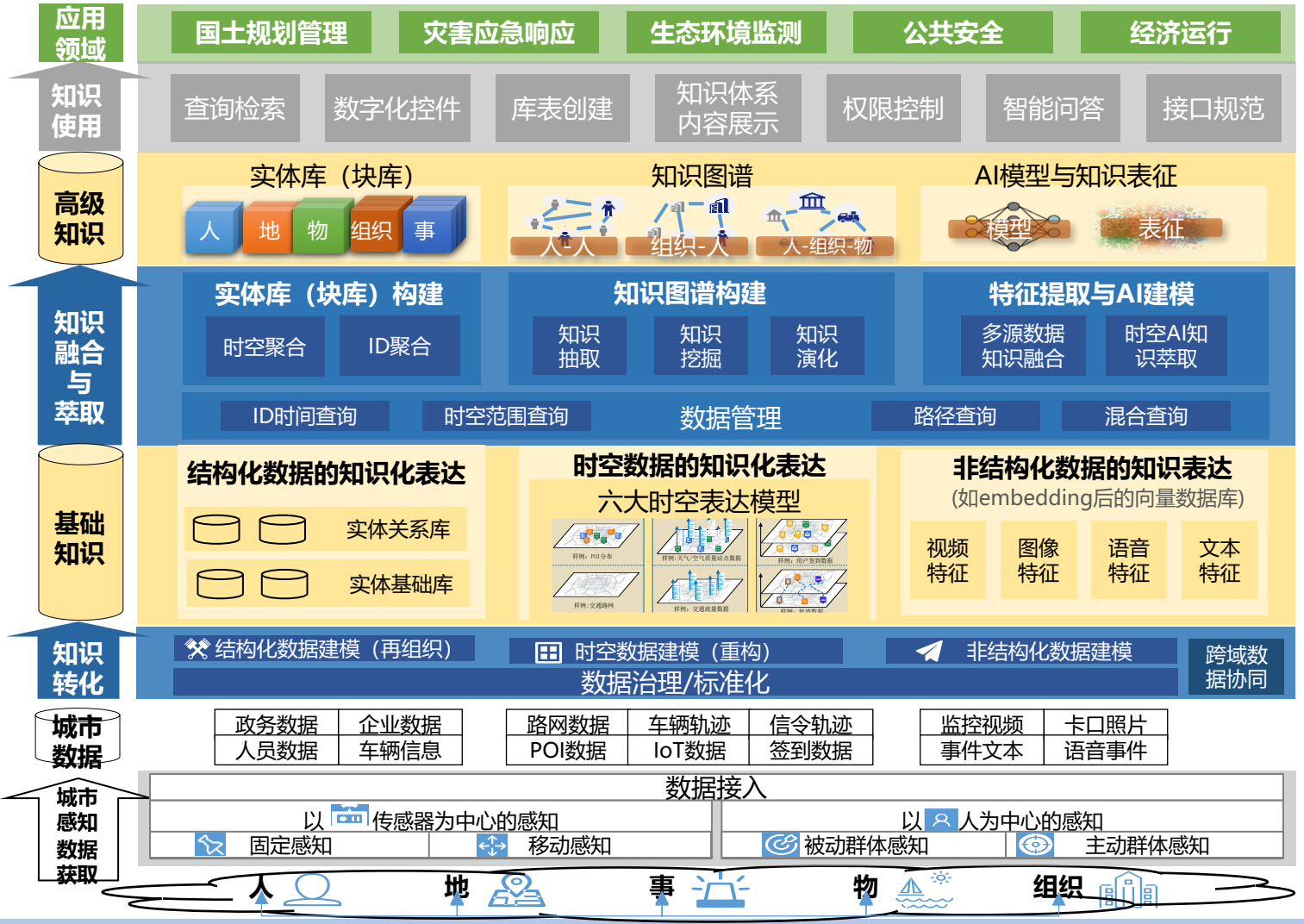
时空数据

郑宇. 基于城市知识体系的公共数据要素化方法. 大数据，2024

城市知识体系

京东智能城市
Powered by JD Cloud

将数据转化为知识，用知识来指引数据的对齐；用知识来解决更难、跟深层次的问题；



提供**对上层**国土规划、应急响应、生态环境等**城市智能应用的知识支撑**

高效**数据管理**手段为支撑，利用实体库构建、特征提取与AI建模、知识图谱等方法，将**初级知识**进一步加工成**高级知识**，让不同类型的数据可以做进一步的知识融合。

接入数据，对数据进行**治理、组织**，形成**结构化、时空类型、非结构化3大类基础知识，解决同类数据共享和知识融合问题。**

通过**城市应用**提炼**5大类实体**，基于**城市感知**方法针对抽象出的本体**获取数据**

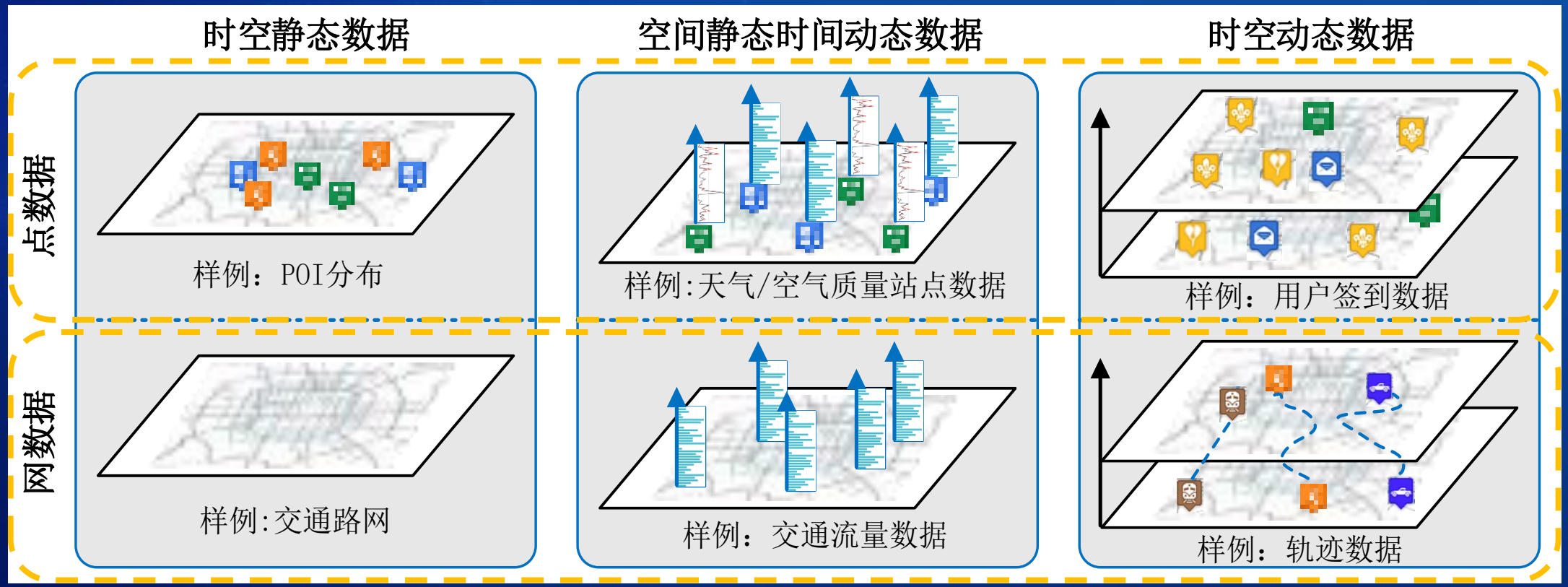
郑宇.城市知识体系[J/OL].武汉大学学报(信息科学版):1-20[2022-07-05].

1.1 构建海量城市数据：数据孤岛凸显

挑战：时空数据种类繁杂、形态各异，缺乏数据标准

方法：城市知识体系-六大时空数据模型

贡献：构建时空数据资源体系，确保数据的一致性和系统的可扩展性



郑宇. 智能城市操作系统. 计算机学会通讯，2020，16（12）：39-44

1.2 构建海量城市数据：数据孤岛凸显

城市知识体系-实现结构化数据的要素化

智慧城市应用

医疗物资保障

医院的氧气罐

教育登记簿

在校学生

学生导师

...

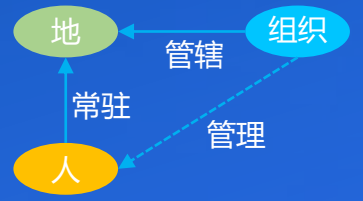
安全隐患排查

城管和井盖

井盖维修

重点人群信息采集

社区居民管理

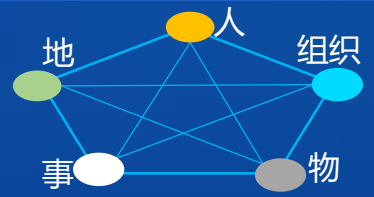


城市知识体系

抽象与凝练

5大类实体

25种 同类实体间关系



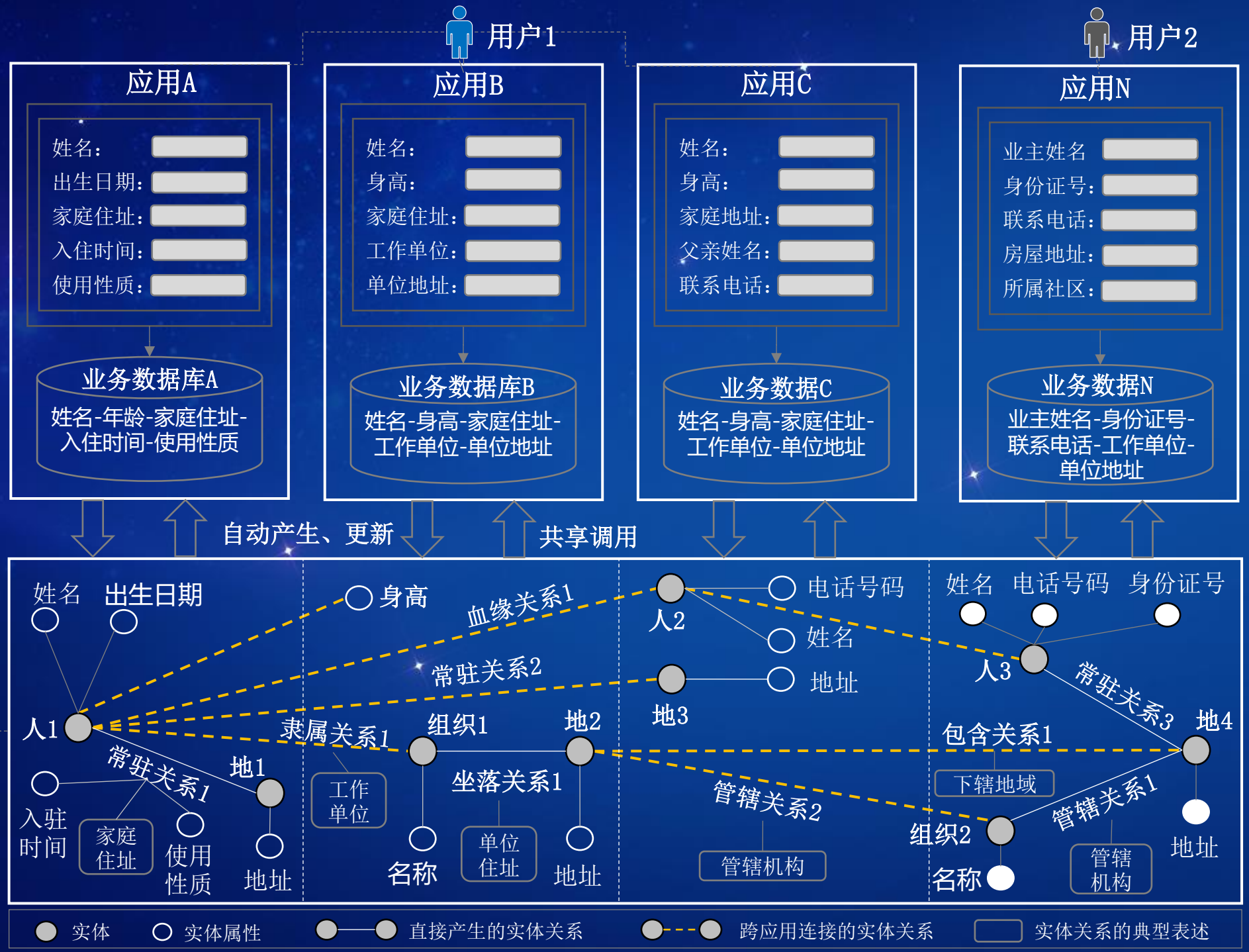
52种 实体间相互关系

600+ 实体和关系属性

底层数据要素

数据标准与规范

郑宇. 城市知识体系[J]. 武汉大学学报（信息科学版), 2023, 48(1): 1-16.



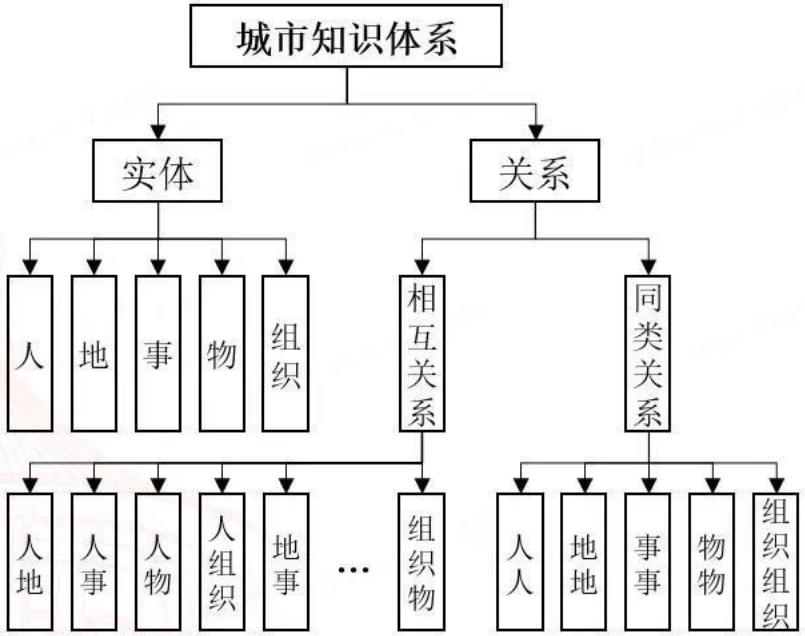
数据要素化三大目标

1. 数据跟应用分离，以便在不同应用间精准共享。

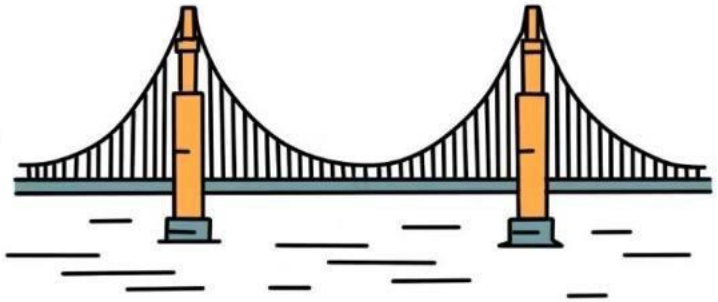
2. 数据要素能自动产生和更新，以便形成规模效应。

3. 不同数据要素能自动连接，以便充分发掘数据价值。

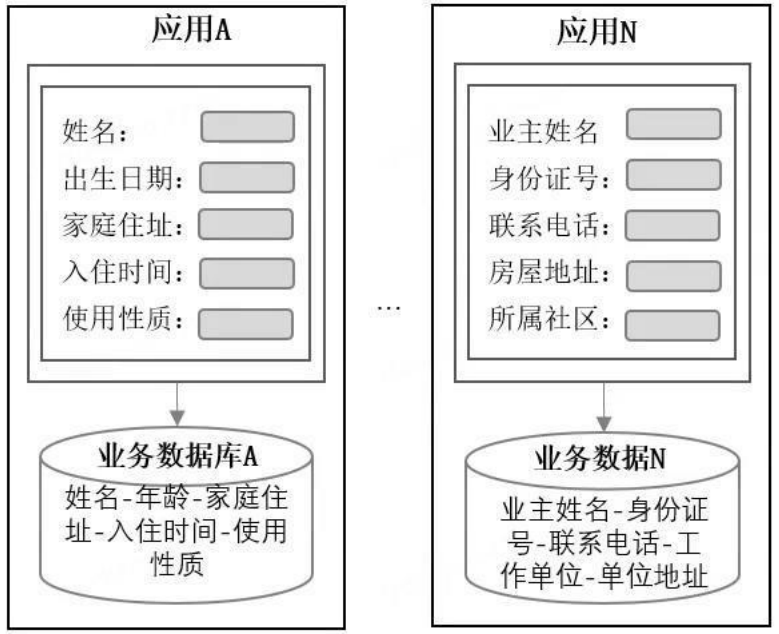
郑宇. 基于城市知识体系的公共数据要素化方法. 大数据，2024，10(4)



数据要素理论



数字化控件



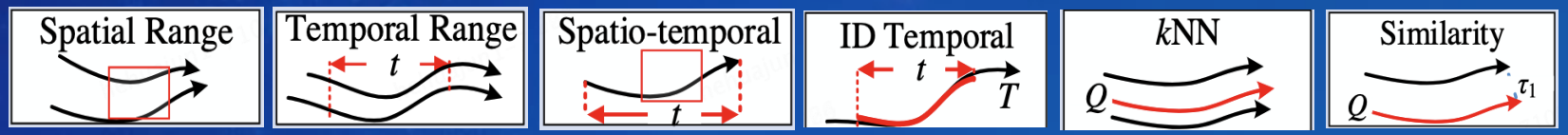
城市业务应用

搭建了数据要素理论到应用的桥梁

2. 提供强大算力：时空数据管理

挑战：时空数据体量大、更新频、查询方式不同、响应要求快，现有数据库系统不能满足即时处理的需求；

时空查询方式



SQL语句引擎

支持六种查询方式，满足业务需求，符合使用习惯

时空数据搜索

时空剪枝优化

分布并行加速

时空查询效率提升10倍

时空数据存储

数据无损压缩

存储空间减少85%，效率提升7倍

数据编码索引

减少精确匹配样本数80%

He H, et al. TMan: A High-Performance Trajectory Data Management System Based on Key-value Stores, ICDE 2024

He H,, et al.
Trass: Efficient trajectory similarity search based on key-value data stores.
ICDE 2024

Li R.. et al.
Elf: Erasing-Based Lossless Floating-Point Compression, VLDB 2023

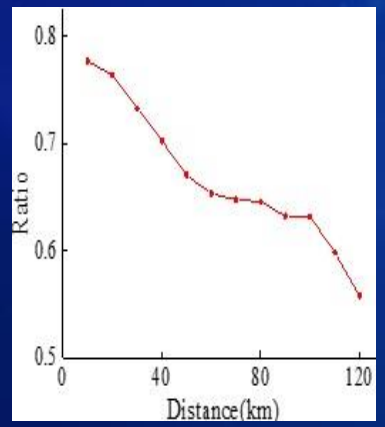
3. 设计城市AI大模型架构

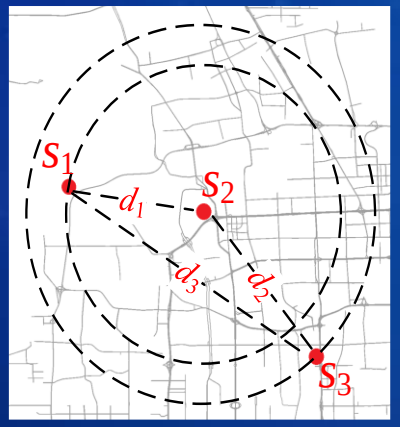
挑战：通用人工智能技术对时空属性的理解和支撑不足

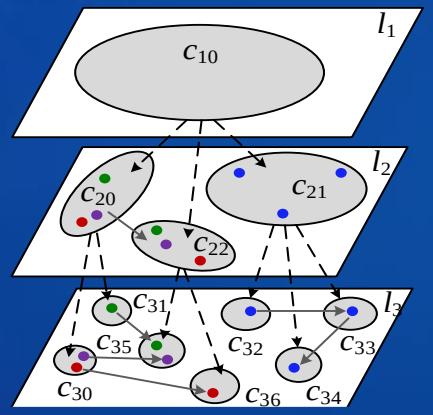
3.1 时空属性建模

降低模型复杂度、捕获时空数据特性

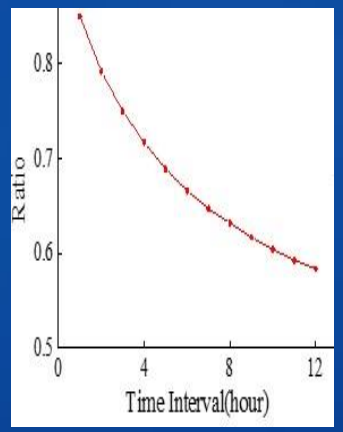
空间临近性、空间层次与距离

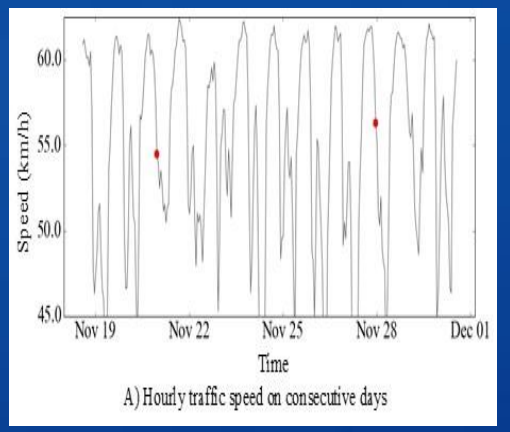


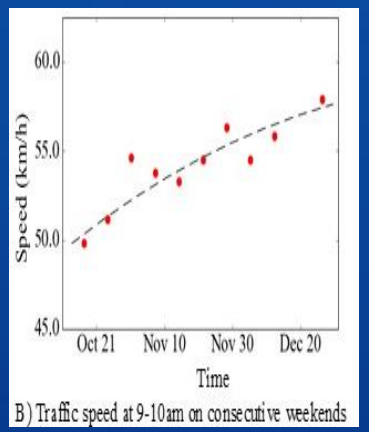




时间临近性、周期性、趋势性







在人流量、交通流和环境监测等预测任务上，将业界的预测准确率提升了10%、计算资源减少70%。

Zhang J., Zheng Y. Qi D., Deep spatio-temporal residual networks for citywide crowd flows prediction, AAAI 2017

3.2 设计城市AI大模型架构-多源数据融合

单一数据源、多模态数据

第74分钟，场上僵局终于被打破，恩多耶禁区右侧获队友直塞后横敲中路，恩博洛门前近距离捅射入网，瑞士领先，1-0！

第80分钟，赖斯分球，萨卡大禁区前沿右侧连续内切后怒
左侧死角，英格兰扳平比分，1-1！

Latent topic

第88分钟，埃泽带球抹入禁区左侧，闪开角度右脚爆射打偏。第90+2'沙尔右路传中，阿坎吉后点头球攻门未能顶正部位，祖贝尔禁区左侧跟上得球右脚凌空爆射高出横梁，瑞士错失绝杀机会。90分钟常规时间结束，双方1-1战平，比赛被迫入加时。

IEEE TBD的创刊文章

IEEE Transaction on Big Data, vol.1, no.1 1

Methodologies for Cross-Domain Data Fusion: An Overview

Yu Zheng, *Senior Member*

Abstract— Traditional data mining usually deals with data from a single domain. In the big data era, we face a diversity of datasets from different sources in different domains. These datasets consist of multiple modalities, each of which has a different representation, distribution, scale, and density. How to unlock the power of knowledge from multiple disparate (but potentially connected) datasets is paramount in big data research, essentially distinguishing big data from traditional data mining tasks. This calls for advanced techniques that can fuse the knowledge from various datasets organically in a machine learning and data mining task. This paper summarizes the data fusion methodologies, classifying them into three categories: stage-based, feature level-based, and semantic meaning-based data fusion methods. The last category of data fusion methods is further divided into four groups: multi-view learning-based, similarity-based, probabilistic dependency-based, and transfer learning-based methods. These methods focus on knowledge fusion rather than schema mapping and data merging, significantly distinguishing between cross-domain data fusion and traditional data fusion studied in the database community. This paper does not only introduce high-level principles of each category of methods, but also give examples in which these techniques are used to handle real big data problems. In addition, this paper positions existing works in a framework, exploring the relationship and difference between different data fusion methods. This paper will help a wide range of communities find a solution for data fusion in big data projects.

Index Terms— Big Data, cross-domain data mining, data fusion, multi-modality data representation, deep neural networks, multi-view learning, matrix factorization, probabilistic graphical models, transfer learning, urban computing.

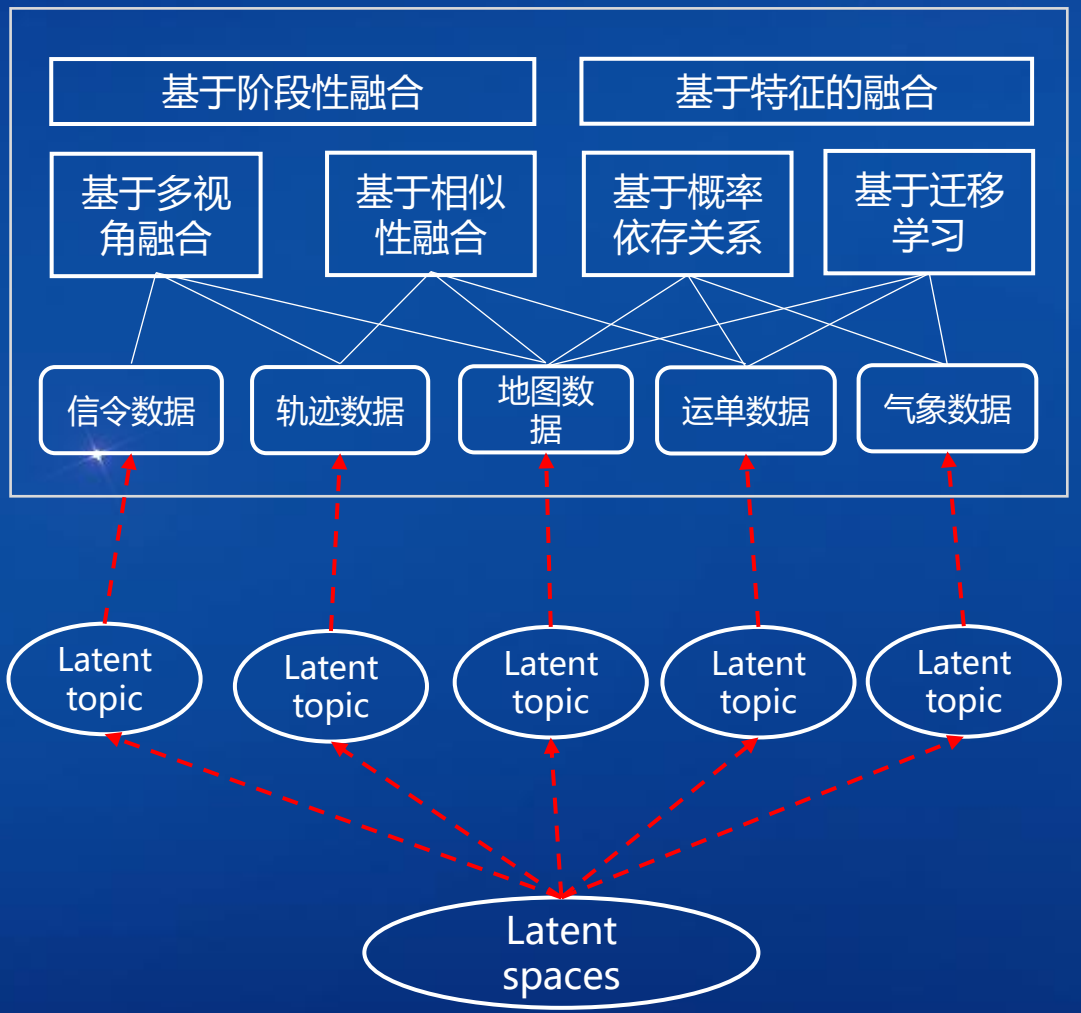
1 Introduction

In the big data era, a wide array of data have been generated in different domains, from social media to transportation, from health care to wireless communication networks. When addressing a problem, we usually need to harness multiple disparate datasets [84]. For example, to improve urban planning, we need to consider the structure of a road network, traffic volume, points of interests (POIs) and populations in a city. To tackle air pollution, we need to explore air quality data together with meteorological data, emissions from vehicles and factories, as well as the dispersion condition of a place. To generate a more accurate travel recommendation for users, we shall consider the user's behavior on the Internet and in the physical world. To better understand an image's semantic meanings, we can use its surrounding text and the features derived from its pixels. So, how to unlock the power of knowledge from multiple datasets across different domains is paramount in big data research, essentially distinguishing big data from tradition data mining tasks.

However, the data from different domains consists of multiple modalities, each of which has a different representation, distribution, scale and density. For example, text denoted as a spatial graph. Treating different datasets equally or simply concatenating the features from disparate datasets cannot achieve a good performance in data mining tasks [8][46][56]. As a result, fusing data across modalities becomes a new challenge in big data research, calling for advanced data fusion technology.

This paper summarizes three categories of methods that can fuse multiple datasets. The *first* category of data fusion methods use different datasets at different stages of a data mining task. We call them stage-based fusion methods. For example, Zheng et al. [86] first partition a city into disjoint regions by road network data, and then detect the pairs of regions that are not well connected based on human mobility data. These region pairs could denote the design that is out of date in a city's transportation network. The *second* category of methods learns a new representation of the original features extracted from different datasets by using deep neural networks (DNN). The new feature representation will then be fed into a model for classification or prediction. The *third* category blends data based on their semantic meanings, which can be further classified into four

不同数据源、多模态数据



Yu Zheng, Methodologies for cross-domain data fusion: An overview, IEEE transactions on big data 1 (1), 16-34

单条数据量小，信息精准度高

单条数据量大，信息精准度低

结构化数据

姓名：

出生日期：

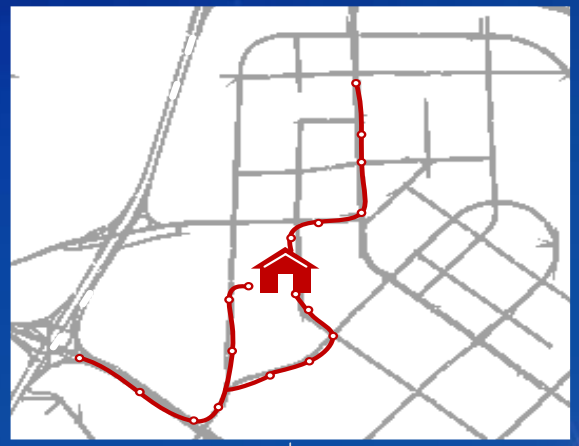
家庭住址：

入住时间：

使用性质：

政务数据库

时空数据



轨迹数据

路网数据

兴趣点数据

非结构化数据

夕阳的余晖洒落在回家的小路上
人们在楼下的小酒馆悠闲的交流。

转发 203　评论 175　赞

转发　评论　赞

图片数据

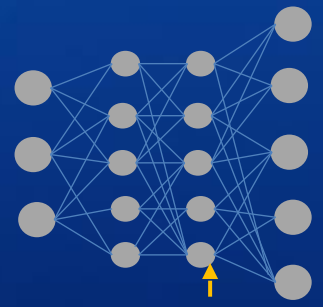
文本数据

视频数据

高精度融合

人 常驻 地

低精度融合



相互结合

3.3 时空AI-模型构建框架

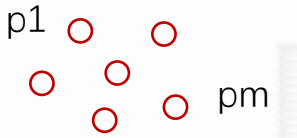
Raw ST Data → Data Transformation → Applicable Data Modes → Model Selection → ST-AI models → Deployment → Applications


POI


Road Networks

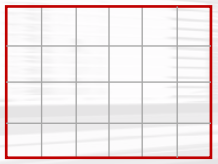

trajectories

ST Point-Based Data

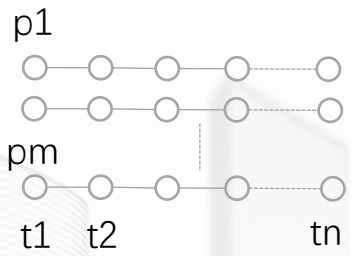


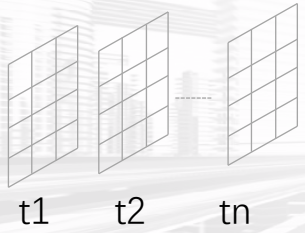
ST Sequential Data

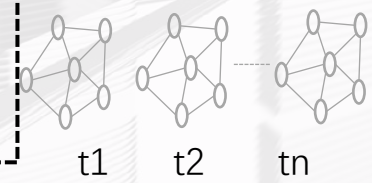
ST-Grid Based Data



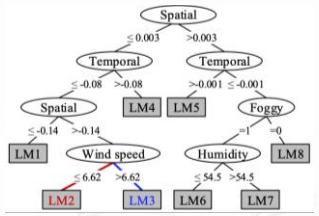
ST Network-Based Data

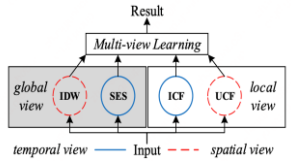




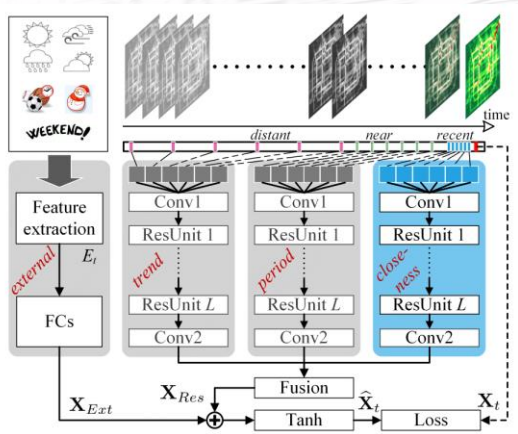


ST Feature Engineering+ small machine learning models

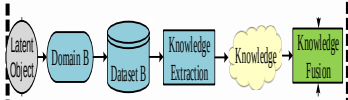




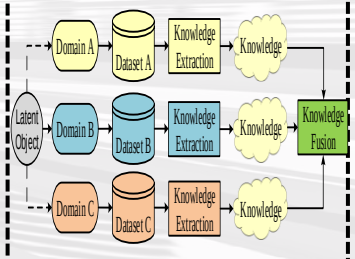
ST Deep Neural Networks



Single domain



Cross-domain

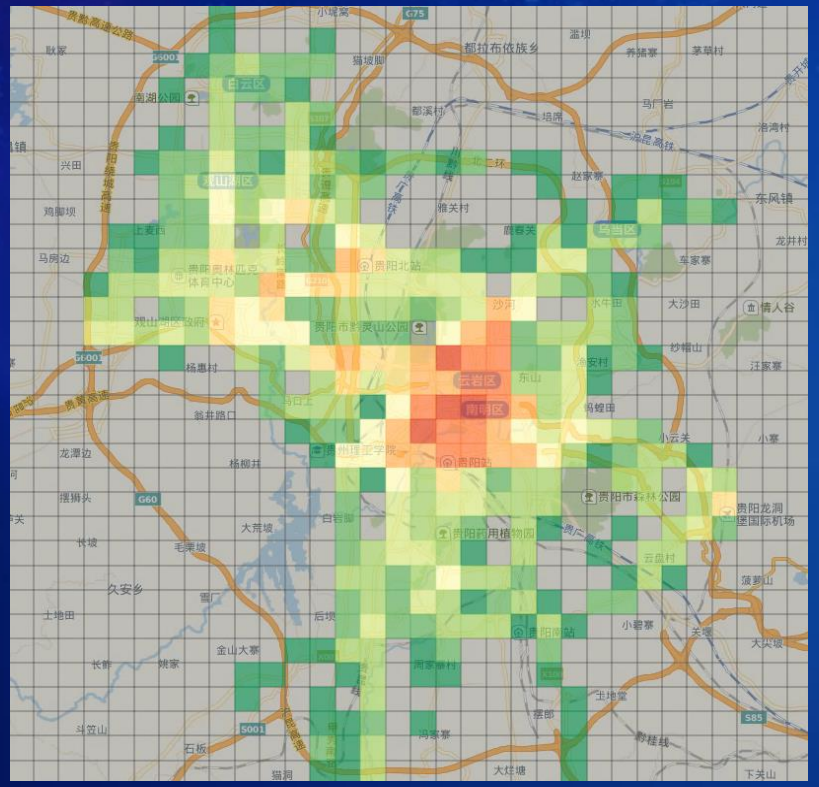


Tasks

- Forecast
- Anomaly detections
- imputations
- interpolations
- Recommendations
- Ranking
- Scheduling
- ⋮

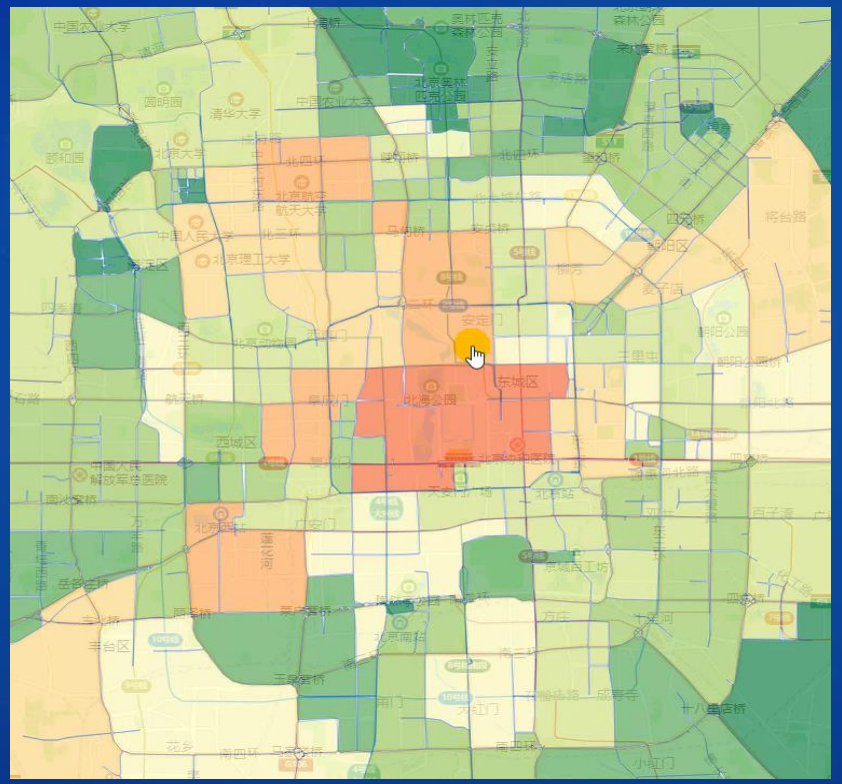
设计城市AI大模型架构：时空深度学习

Grid-based Applications



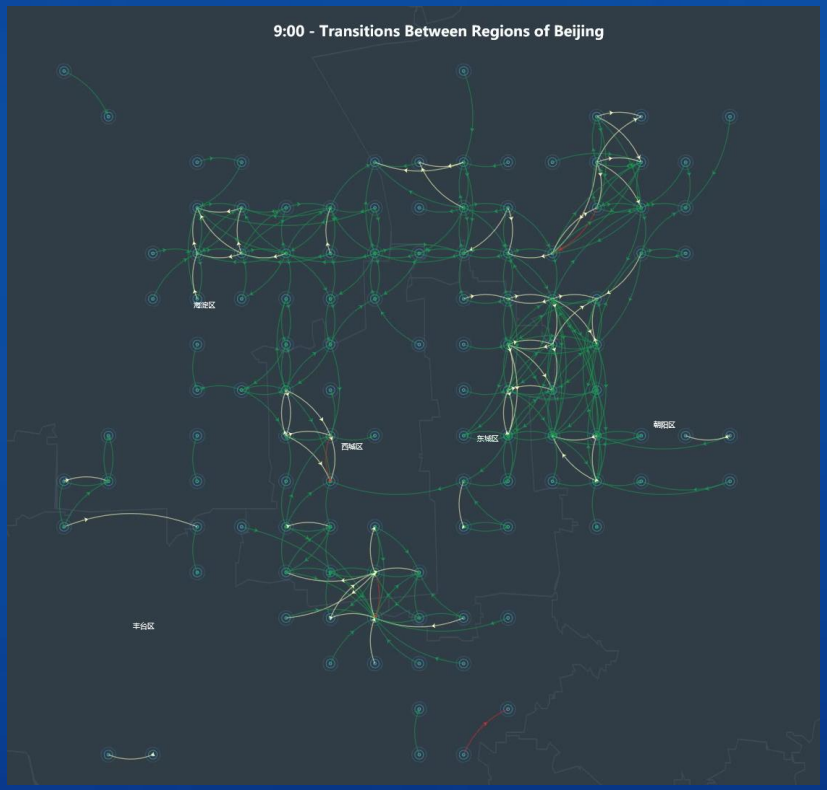
Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction. AAAI 2017

Irregular-region Based Applications



Predicting Citywide Crowd Flows in Irregular Regions Using Multi-View Graph Convolutional Networks. IEEE TKDE, 2020

Graph-based Applications



Flow Prediction in Spatio-Temporal Networks Based on Multitask Deep Learning. IEEE TKDE, 2020

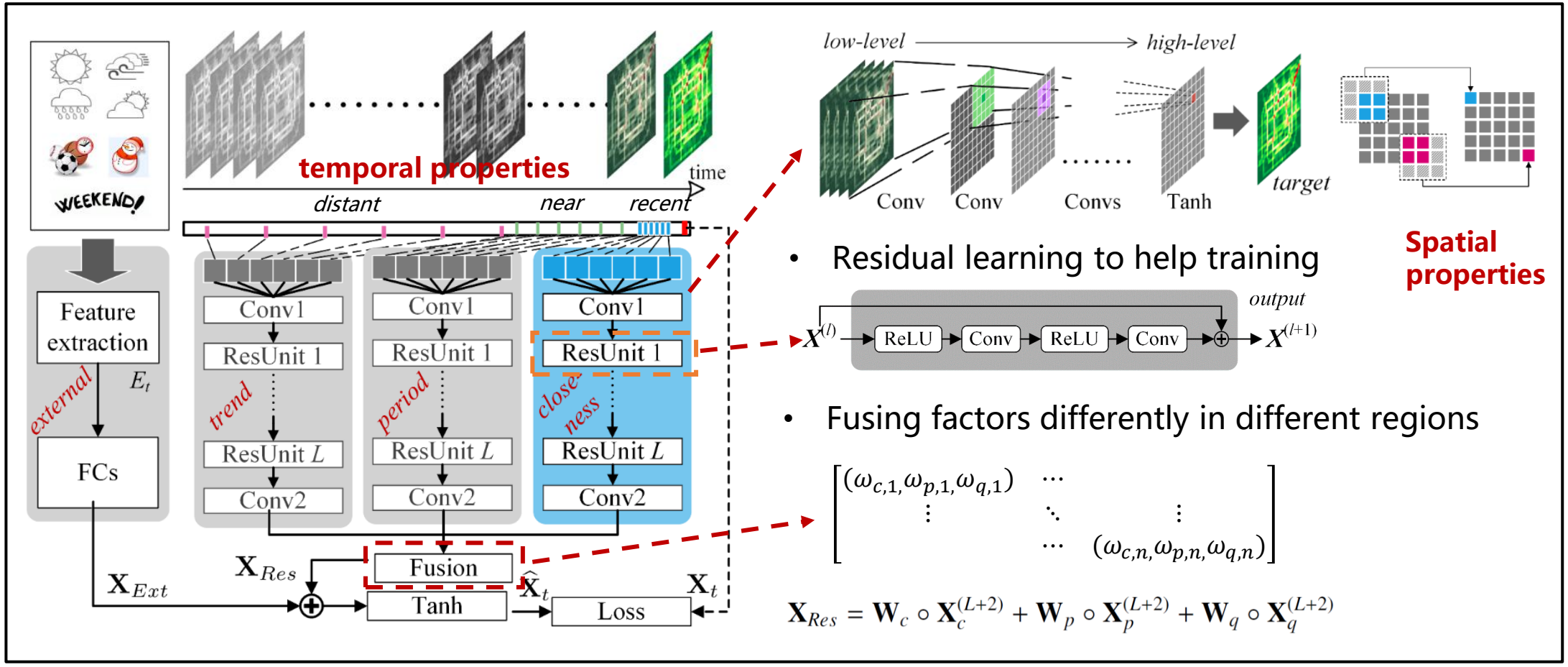
Grid-based Applications

JD iCity
Powered by JD Cloud

- Capture spatial correlation of both near and far distances
- Capture temporal closeness, period, and trend
- Capture external factors

The earliest self attention!(April 2017)

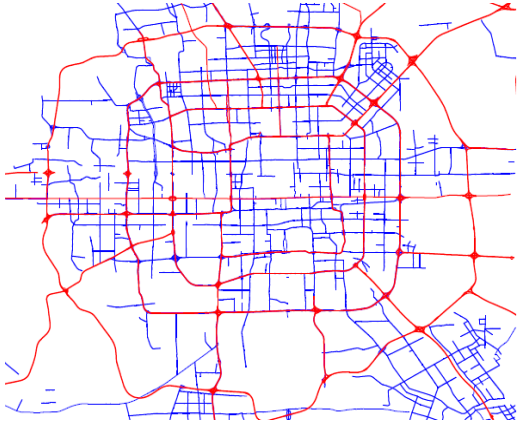
ST-ResNet Architecture: A Collective Prediction

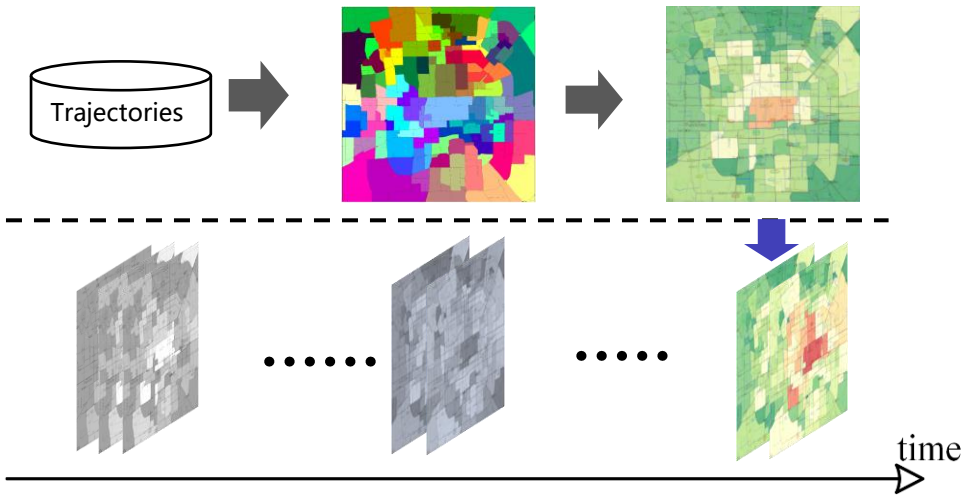


Junbo Zhang, Yu Zheng, et al. Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction, AAAI 2017

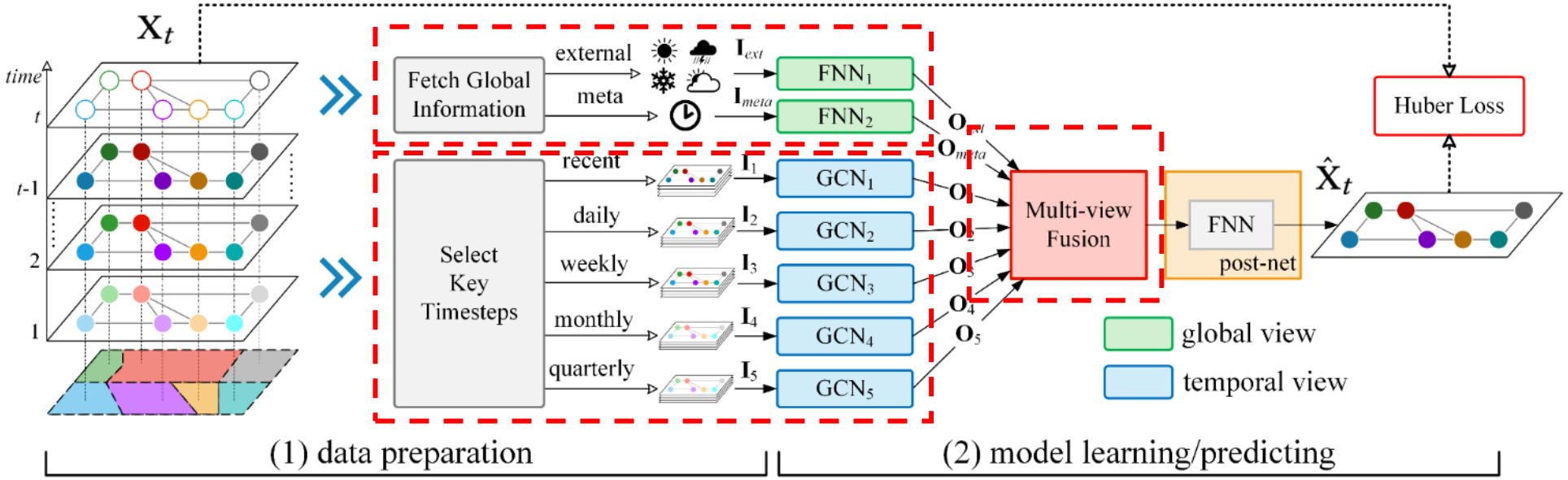
Irregular-region Based Applications

JD iCity
Powered by JD Cloud





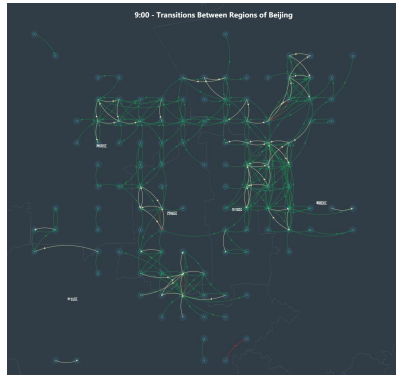
MVGCN Architecture: A Multi-View Framework

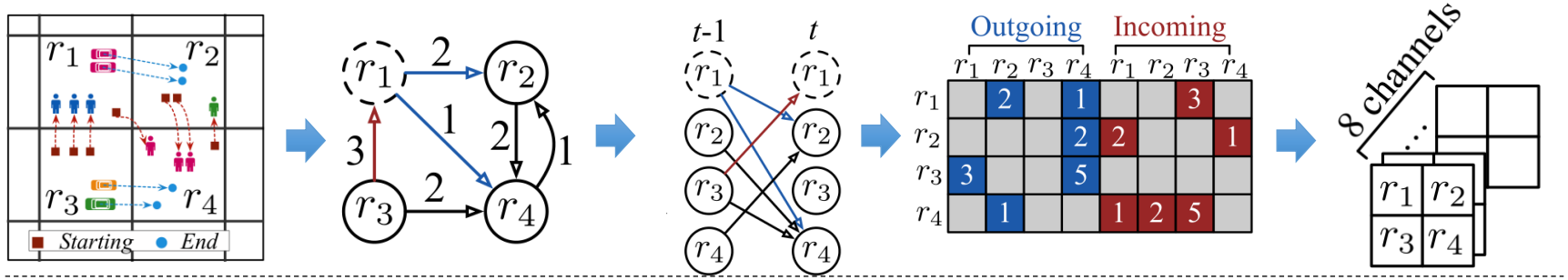


J. Sun, J. Zhang, et al. Predicting Citywide Crowd Flows in Irregular Regions Using Multi-View Graph Convolutional Networks. *IEEE TKDE, 2020*

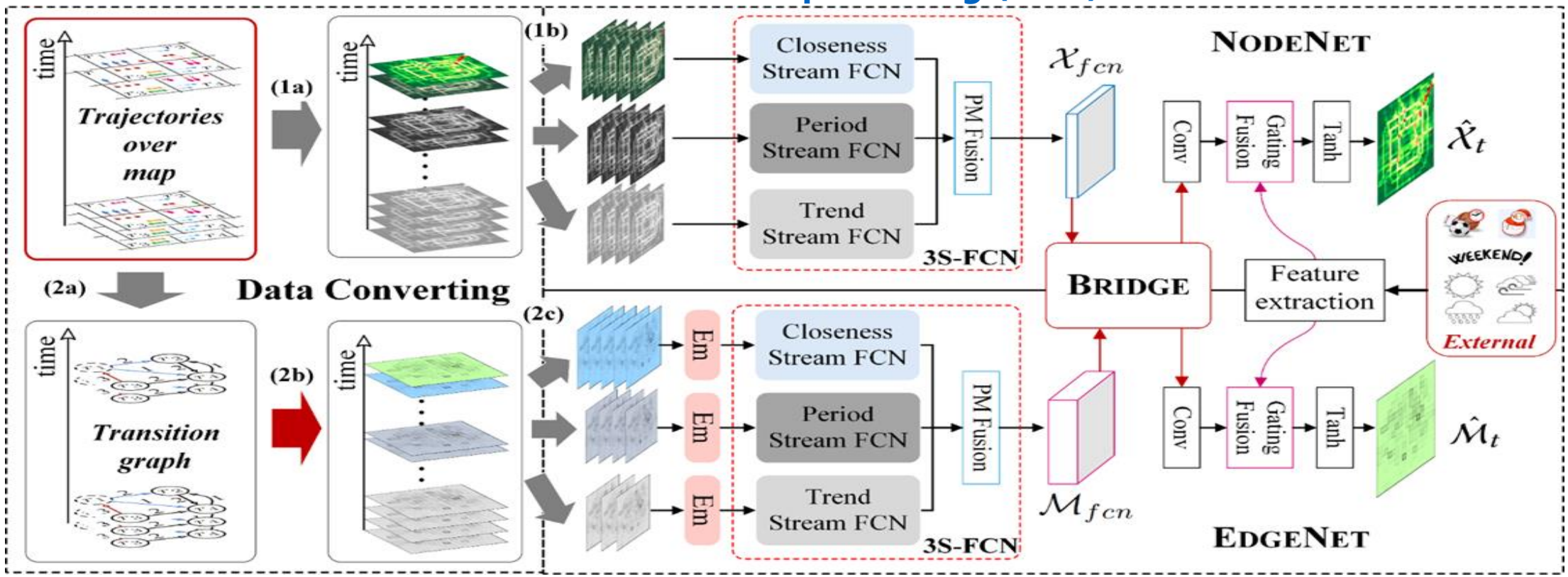
Graph-Based Applications

JD iCity
Powered by JD Cloud





Multitask Deep Learning (MDL) Framework



Junbo Zhang, **Yu Zheng**, et al. Flow Prediction in Spatio-Temporal Networks Based on Multitask Deep Learning. IEEE TKDE, 2020

4. 选择合适的城市应用场景

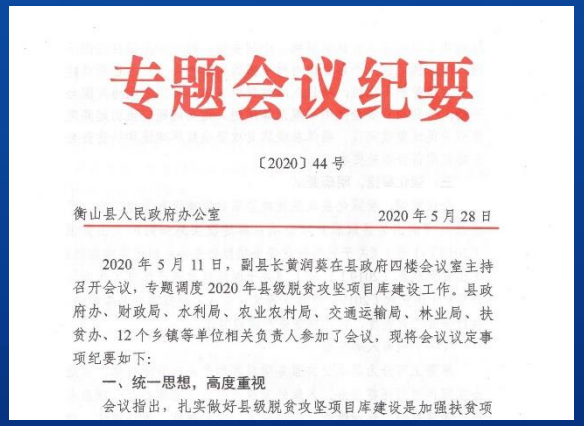
- 影响因素多、推理复杂度高；
- 场景已经沉淀大量可用数据；
- **影响因素在时间、空间上跨度大；**
- **对单条结果的可靠性要求较低；能反映大致的趋势和统计意义即可；**

AI大语言模型

以下场景均不理想！！！

12345智能政务热线

政务服务咨询机器人


专题会议纪要

〔2020〕44号

衡山县人民政府办公室　　2020年5月28日

2020年5月11日，副县长黄润葵在县政府四楼会议室主持召开会议，专题调度2020年县级脱贫攻坚项目库建设工作。县政府办、财政局、水利局、农业农村局、交通运输局、林业局、扶贫办、12个乡镇等单位相关负责人参加了会议，现将会议议定事项纪要如下：

一、统一思想，高度重视

会议指出，扎实做好县级脱贫攻坚项目库建设是加强扶贫项

政务写作和会议纪要

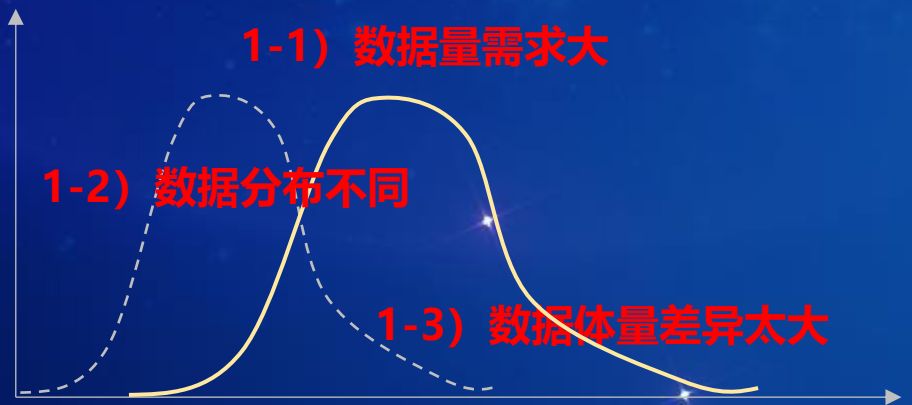
常见问题

正确的废话： 笼统含糊、不解决问题

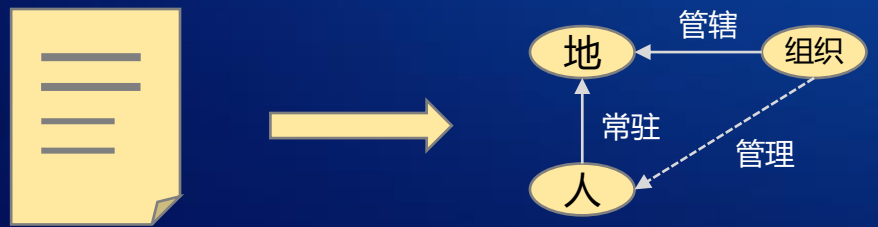
知识不全面： 关键信息缺失

错误的幻觉： 答案不准确

区分度不足： 答案同质化严重

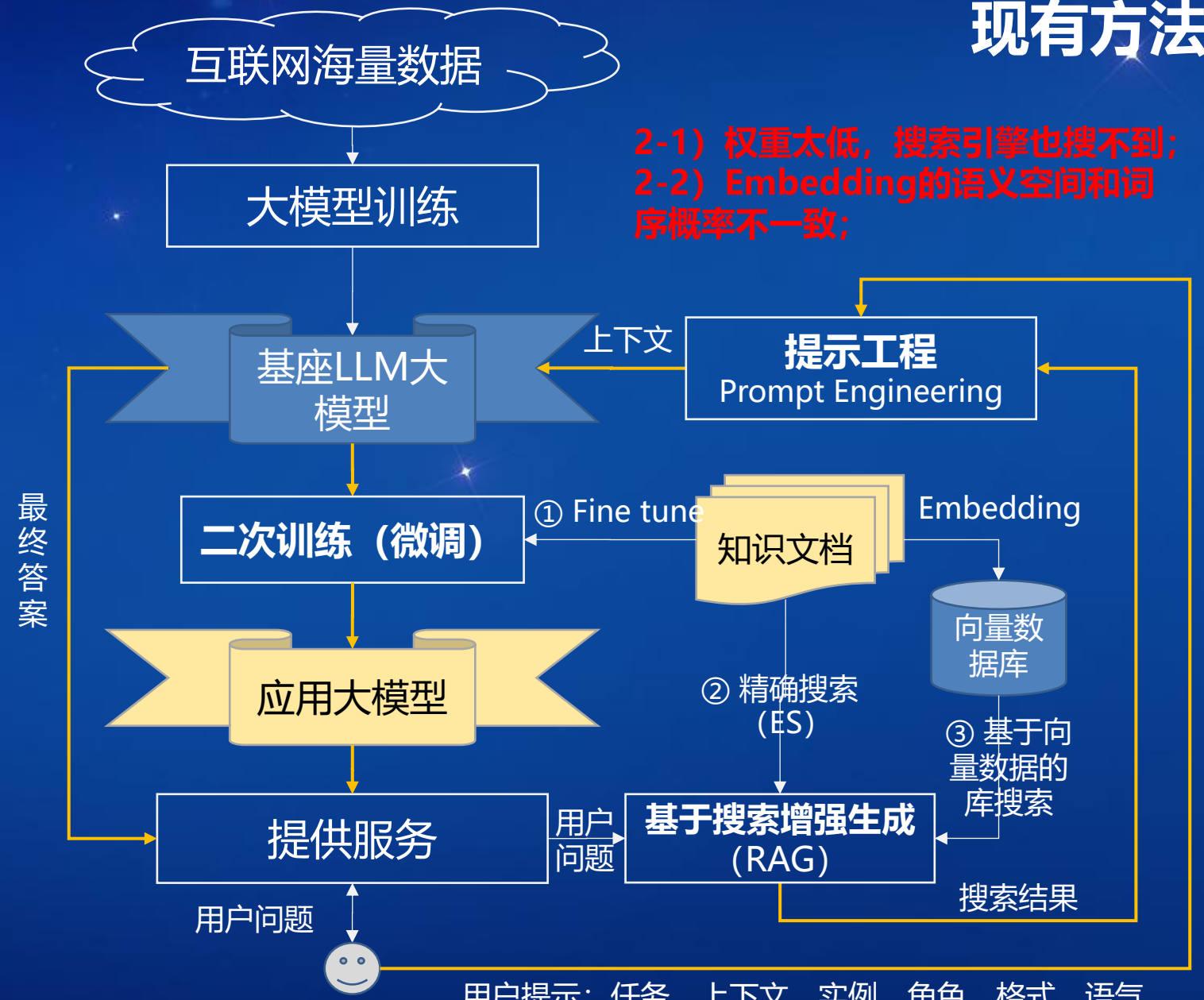


3）含复杂结构的问题处理不好



Q: A街道去年在老人关怀方面做了哪些工作?

现有方法



2-1）权重太低，搜索引擎也搜不到；
2-2）Embedding的语义空间和词序概率不一致；

用户提示：任务、上下文、实例、角色、格式、语气

正确方法

- 借鉴大模型的思想而非照搬模型结构
- 利用AI大模型而非必须用LLM解决问题
- 构建高质量、大规模的城市大数据集
- 基于行业知识设计AI模型结构
- 把知识图谱和大模型相结合

可能的方向

- 交通人流量预测
- 高精度短期降雨预测
- 城市热点事件态势推演
- 宏观经济分析和预判
- 城市规划和演进的仿真
- 长跨度粗颗粒度天气预报

AI大模型

AI大语言模型

基于城市知识体系
构建海量城市数据

海量可互认的
城市数据

基于时空数据管理技术
提供强大算力

AI大模型设计
捕获时空特性、融合多源数据

AI大语言模型设计及训练

场景大模型

基础大模型

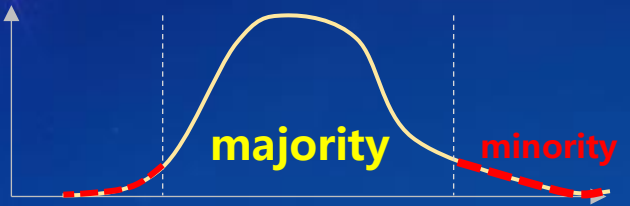
提供服务

agents

若干端到端的中小模型

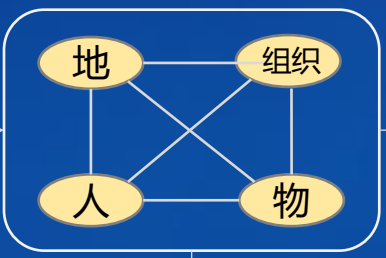
例如：人流预测、短期天气预报、空气质量预测

行业知识



购物、综艺、财经、教育、热点新闻、畅销书籍、热点影视题材等

高精度知识图谱



majority

Graph RAG

minority

例如：自行车道、公交线路、地铁规划等

Memories

AI大有可为，但仍处于婴儿阶段，未来的关键是跟行业深度融合

大语言模型是AI发展道路上的重要里程碑，但不是最终形态！

学习大模型的思想，不被模型架构禁锢，审慎判断、合理使用！

郑宇 博士

京东集团副总裁、IEEE Fellow

国家万人计划科技领军人才

谢谢！