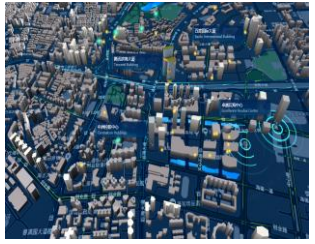


以人为中心的时空数据计算：从稀疏感知到不确定性

中国科学技术大学
计算机学院、软件学院、大数据学院
中国科大-数据智能实验室　汪　炀

2021.01.09

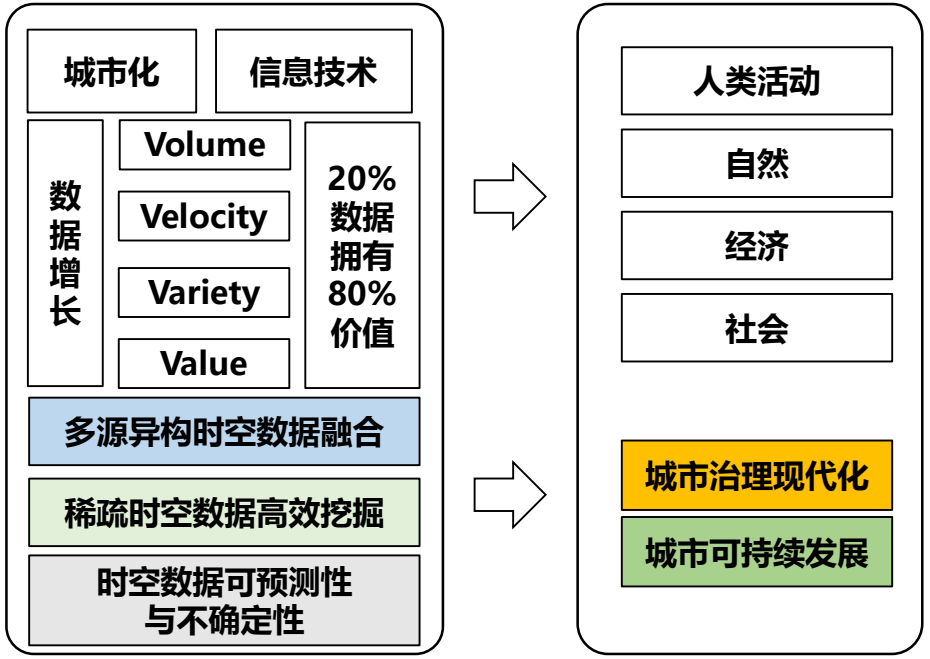
目 录

报告人介绍

报告人： 汪 炀，中国科学技术大学软件学院副院长，**安徽省杰青获得者。**主要研究方向包括城市计算与时空数据挖掘、分布式计算等，近年来作为负责人主持3项国家自然科学基金、1项国家自然科学基金青年基金等近30余项纵向课题。近5年来以第一/通信作者身份在顶级国际学术会议**Mobicom、Infocom、AAAI、IJCAI、Ubicomp、ICDM、ICDCS等**和**IEEE TVT、TKDE、ACM TIST**等期刊上发表城市计算和时空数据挖掘相关的CCF-B类以上、JCR一区论文30余篇。因在车联网与智能计算方面的贡献被授予IBM全球**杰出学者奖和国家留学基金委优秀教师奖。**

报告介绍

物联网技术和人工智能的快速发展，可获得的城市数据呈现指数级增长，而数据-价值往往遵循***关键少数法则***。因此，如何进行多源异构时空数据融合，高效利用已有的稀疏数据实现知识发现和信息挖掘成为城市计算的一大挑战。



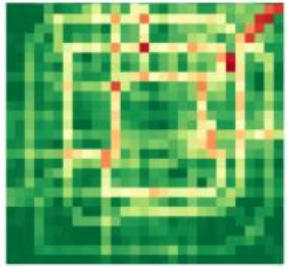
目 录

时空数据概述

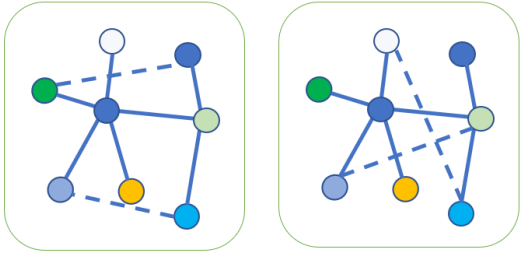
- **时空数据**是指具有空间分布且随着时间变化而变化的数据的集合。e.g. 路段（路口）交通流、人口密度、大气污染、网约车订单、国民经济发展、兴趣点（POI）签到。

- **来源：** 遥感（RS）、 城市监控（Monitoring）、 移动设备（Mobile device）

- **数据形态**



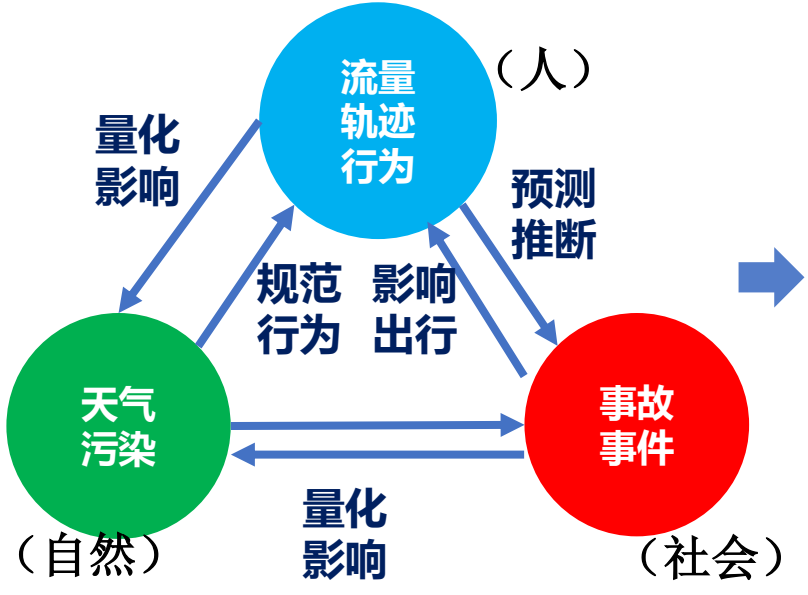
网格型（Grid）

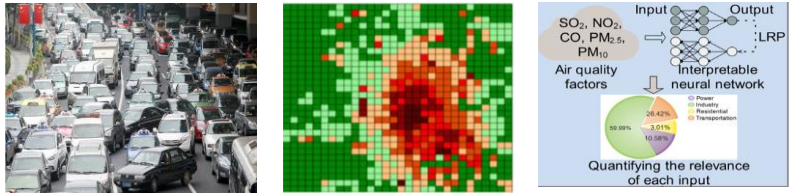


图型（Graph）

轨迹行为序列（Trajectory）

以人为中心的时空数据计算





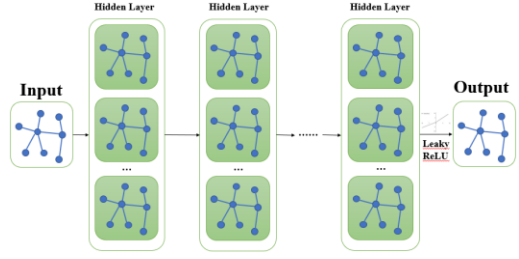
房价市场预测预警 | 兴趣点商家推荐配送优化 | 出租车供需优化 | 犯罪事件预测 | 交通事故预测 | 大气科学 | 水污染

社会经济 | 自然

- 城市可持续发展
- 城市治理能力现代化

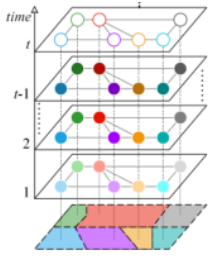
目 录

时空数据计算的一般方法

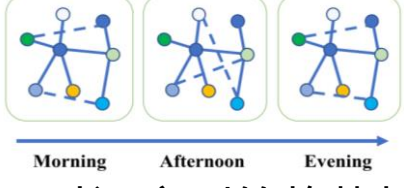


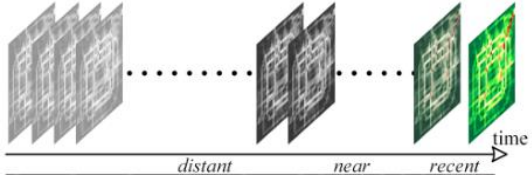
时空数据特点

1. 路网传播特性



2. 区域之间的动态空间关联

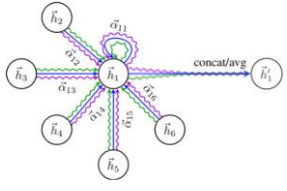


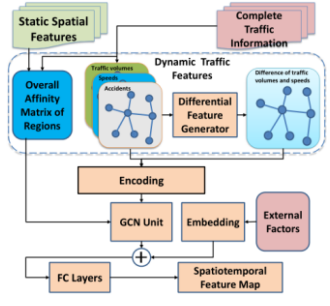


3. 时间序列的趋势性、连续型、周期性

对应技术

网格数据CNN 、图网络建模GCN

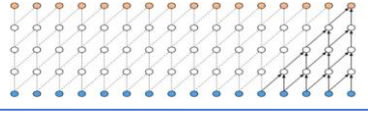




注意力机制、GAT、

时变模式抽取

RNN变体时间序列建模、TCN



时空数据计算相关工作回顾

传统机器学习

事件预测（Classification）
Xgboost, Random Forest

流量速度预测（Regression）
非负矩阵分解NMF
核密度估计 KDE
滑动平均自回归ARIMA

深度学习

Graph-based

Grid-based

ST-ResNet AAAI17

DCRNN (ICLR18)

Hetero-ConvLSTM (KDD18)

ASTGCN(AAAI19)
STG2Seq(IJCAI19)
ST-MetaNet(KDD19)
MDL(TKDE19)

3D-CNN TITS'19
UrbanFM (KDD19)

……

AGCRN(NIPS'20)
RiskOracle (AAAI20)

Curb-GAN (KDD20)

RiskSeq (TKDE21)

……

时空数据挖掘的两大挑战

以人为中心的时空数据计算挑战：从稀疏性到不确定性

以人为中心的时空数据

多样性　稀疏性　规模性

- 设备部署成本
- 设备存储能力
- 偶发事件记录本质稀少

城市物联网系统

- 时空关联特性
- 多源信息互补
- 价值性

时空数据挖掘的两大挑战

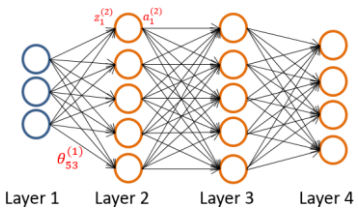
以人为中心的时空数据计算挑战：从稀疏性到不确定性

以人为中心的时空数据

多样性　稀疏性　规模性

关键少数法则
价值性（价值密度）

数据驱动的预测模型



稀疏　噪声

数据不确定性

模型不确定性

预测结果不确定性

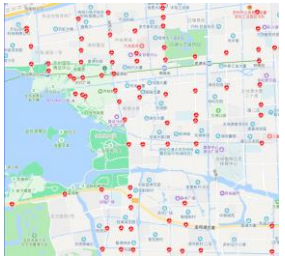
从稀疏产生源头划分两类稀疏场景

本质稀疏： 数据记录本身偶发，即本身产生很少，即使全部获取也有限，如：犯罪事件、交通拥堵（事故）事件、偶发疫情。无法补全。

伪稀疏： 指数据随时随处产生，但是由于采集设备能力有限，无法全部获取。如：路段通行速度、卡口流量、断续房价记录、大气环境数据等。 可填充。

- 训练过程的零膨胀问题（zero-inflated issue）
- 有效数据覆盖面小，难以支撑训练

（a）本质稀疏


（b）伪稀疏

时空数据挖掘稀疏挑战的主要解决策略

伪稀疏：基于多方数据集的协同学习

- 基于生成式学习的方法
- 基于数据集交叉时空域的协同推断学习
- 基于多源数据和多任务的预测

本质稀疏：数据与损失函数变换策略

- 数据增强与样本生成
- 基于样本不平衡问题的缓解策略
- 基于先验信息的数据间隔最大化

目 录

时空数据挖掘稀疏挑战的主要解决策略

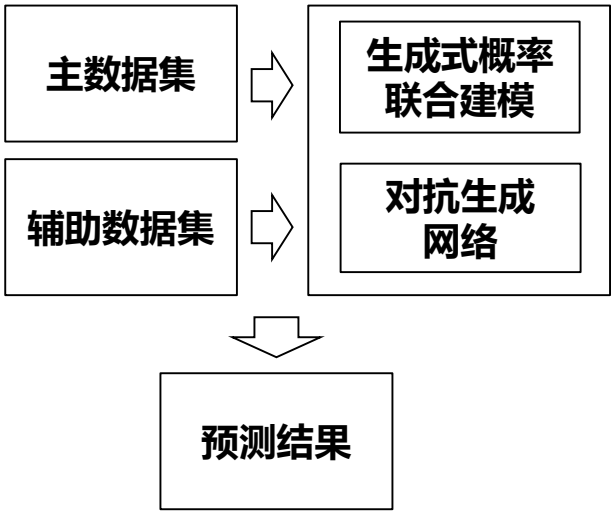
伪稀疏：基于多方数据集的协同学习

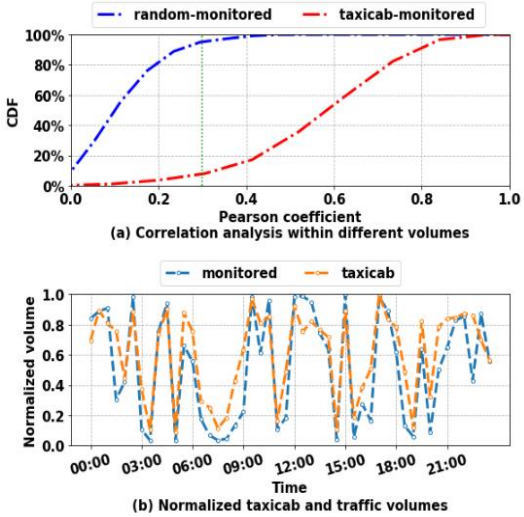
- **基于生成式学习的方法**

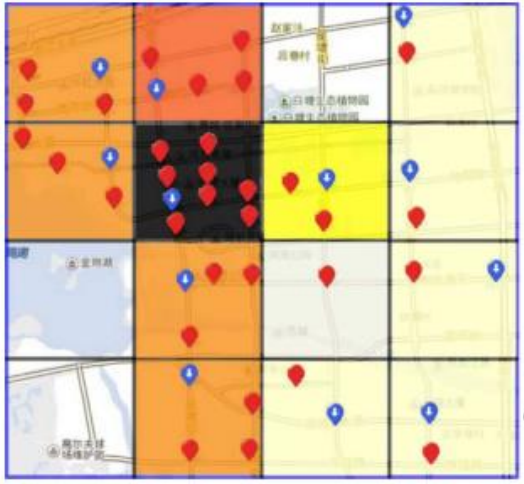
- 基于数据集交叉时空域的协同推断学习

- 基于多源数据和多任务的预测

基于生成式学习的方法


主数据集
辅助数据集
生成式概率联合建模
对抗生成网络
预测结果


random-monitored
taxicab-monitored
CDF
100%
80%
60%
40%
20%
0%
0.0
0.2
0.4
0.6
0.8
1.0
Pearson coefficient
(a) Correlation analysis within different volumes
monitored
taxicab
Normalized volume
1.0
0.8
0.6
0.4
0.2
0.0
00:00
03:00
06:00
09:00
12:00
15:00
18:00
21:00
Time
(b) Normalized taxicab and traffic volumes



基于生成式学习的方法

多元正态分布
密集信息转移矩阵
} 时空建模

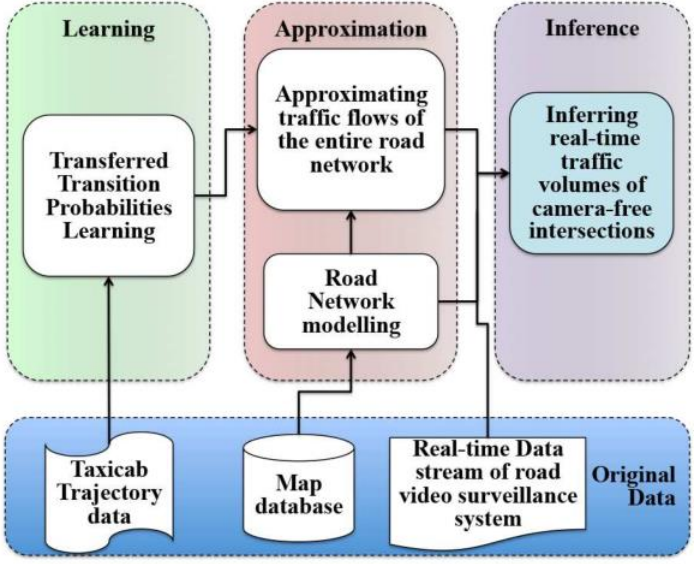
MND条件期望

Case1: 基于稀疏卡口流量的实时交通模式分析与推断

稀疏卡口流量（全类型车辆）+ 密集的出租车轨迹

基于转移矩阵 - 捕获车辆的转移模式

将交通元素建模为**多元正态分布（MND）**



Wang et,al. Real-time Traffic Pattern Analysis and Inference with Sparse Video Surveillance Information (IJCAI 2018).

基于生成式学习的方法

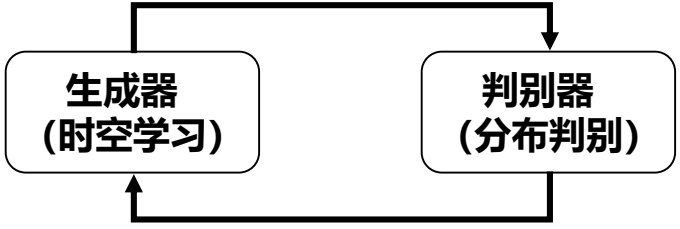
Case2: 基于生成对抗网络的稀疏卡口流量实时推断

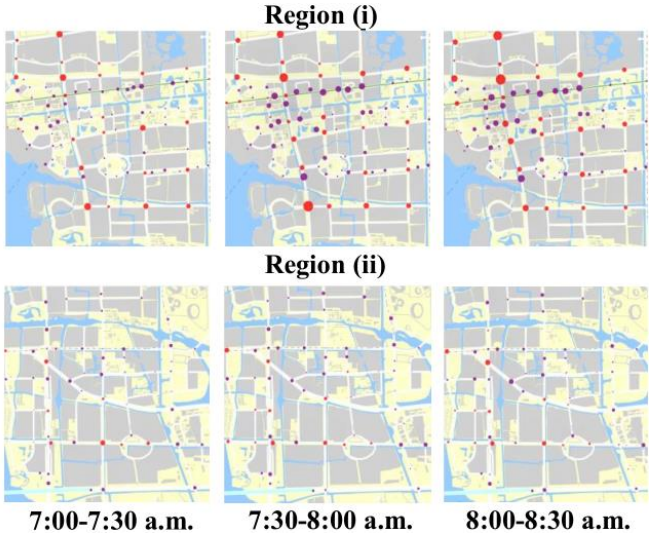
稀疏卡口流量 + 密集的出租车轨迹

依据出租车车流分布与总体车流分布的相似性

基于生成-对抗式网络学习分布

基于出租车密集数据训练->基于稀疏卡口流量实时推断





Zhu et, al. Understanding Road Network Statuses in Traffic Services with Sparse Traffic Sensors.(Under review)

时空数据挖掘稀疏挑战的主要解决策略

伪稀疏：基于多方数据集的协同学习

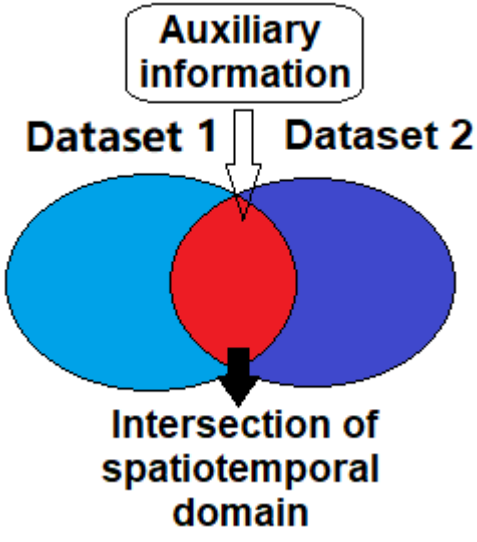
- 基于生成式学习的方法

- **基于交叉时空域的协同推断学习**

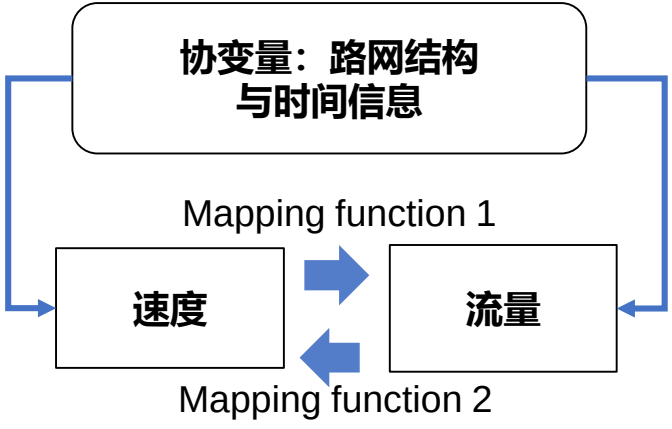
- 基于多源数据和多任务的预测

基于交叉时空域的协同推断学习

时空域交集



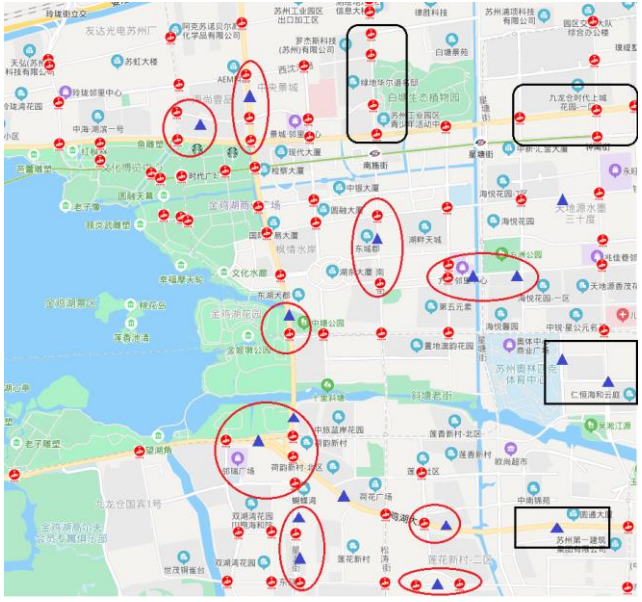
基于时空域交集和相关静态协变量实现相互推断（Mutual Inference）



基于交叉时空域的协同推断学习

Case: 交通事故预测中路网速度、交通流量覆盖率低伪稀疏问题，最大化推断速度与流量

- ✓ 多源时空特征存在高阶交互影响，
  e.g. 路网结构与天气形成“共振”；
- ✓ 车流与车速等动态特征存在非线性关联；
- ✓ 深度因子分解机可捕获不同数据域之间的交互关联，即产生vector-level的高阶项x1x2。



速度和流量探测覆盖示意图（SIP）

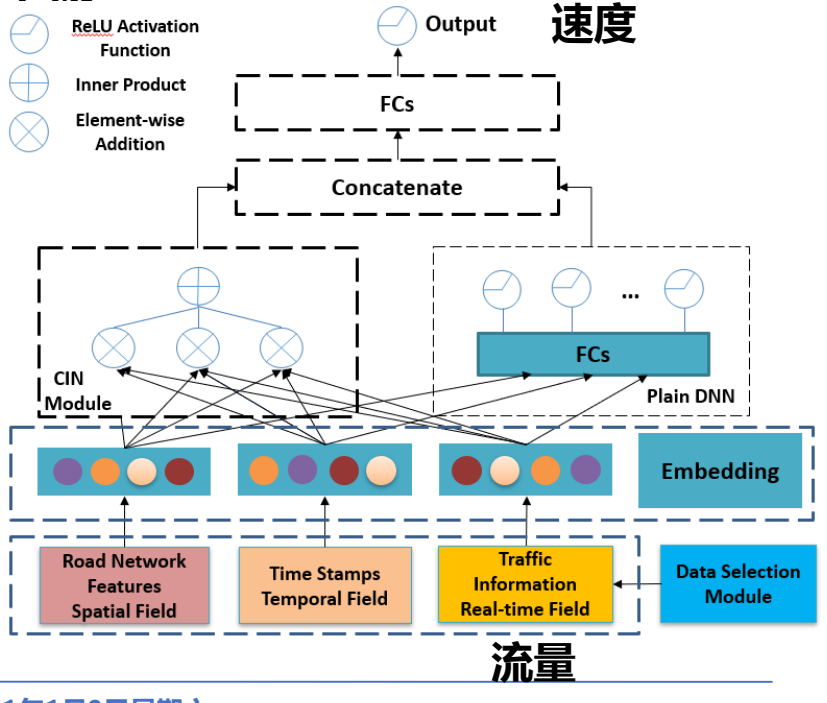
基于交叉时空域的协同推断学习

Case: 交通事故预测中路网速度、交通流量覆盖率低

基于ST-DeepFM的动态交通信息互推断模型

- 对每个区域而言，设计压缩交互网络（CIN）和深度特征提取模块（DNN）
- 筛选对应的静态路网特征，时间戳分别放入对应field
- 自适应地选择邻近的区域动态信息放入Real-time field

最终获得尽可能多的路网动态信息

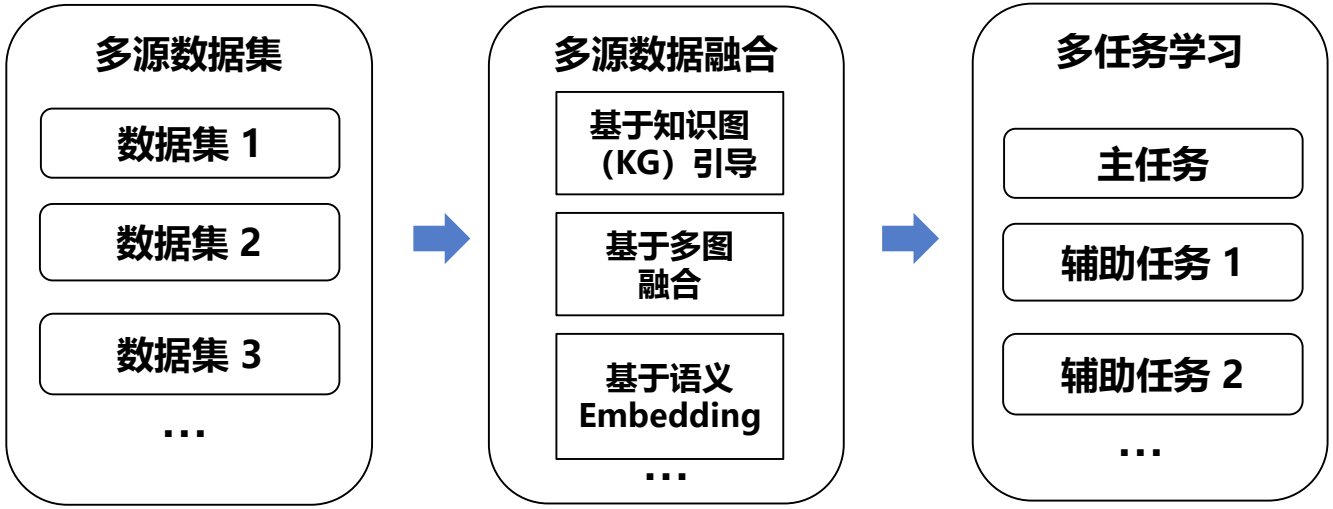


时空数据挖掘稀疏挑战的主要解决策略

伪稀疏：基于多方数据集的协同学习

- 基于生成式学习的方法
- 基于数据集交叉时空域的协同推断学习
- **基于多源数据和多任务的预测**

基于多源数据和多任务的预测



- 任务正则化
- 信息互补
- 辅助数据集利用
- 增强表征能力

基于多源数据和多任务的预测

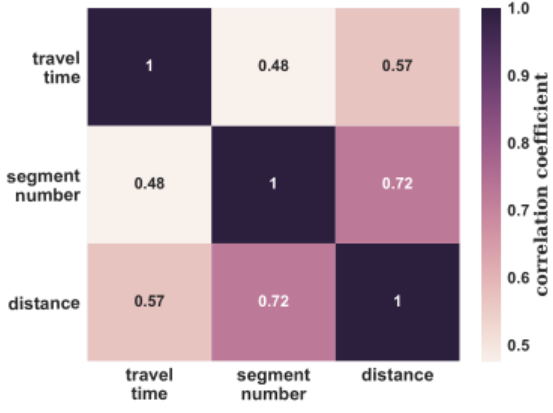
Case1: 基于稀疏交通流量的实时全城通行时间估计

密集的OD和历史轨迹序列，缺少实时交通状态

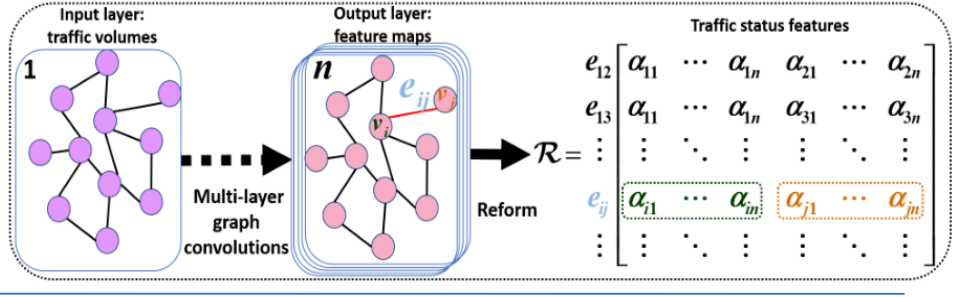
- 引入稀疏卡口的交通监控流量，反映实时路况

针对交通流量的稀疏性，为更好支持OD通行时间估计：

- 基于**路口转移信息**实现动态GCN路网状态学习
- 基于Bi-LSTM的轨迹序列**自编码**
- 构造路网状态、通行时间、路段数等**多任务学习**



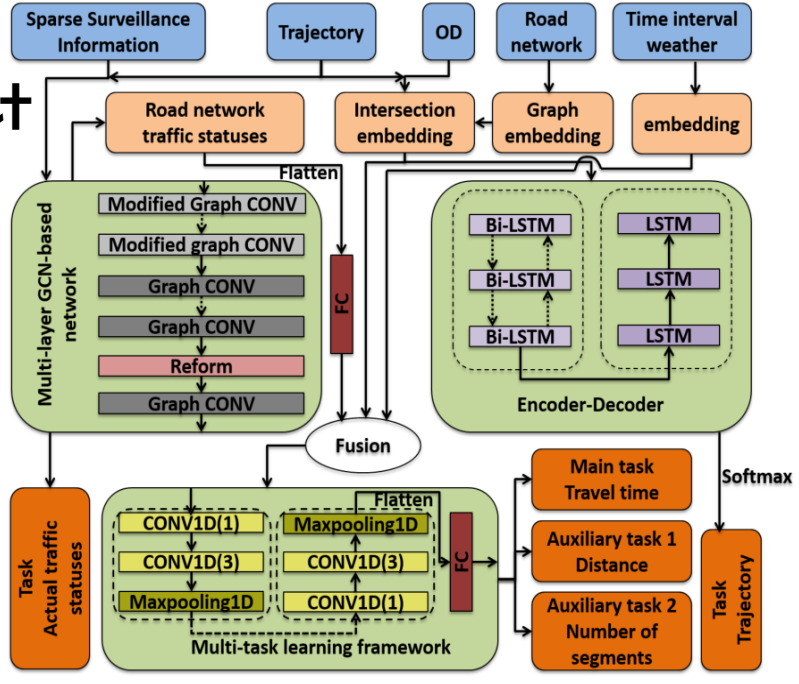
多任务间相关性矩阵



基于多源数据和多任务的预测

Case1: 基于稀疏实时交通流量的全城通行时间估计

- 对主任务进行正则化，提升稀疏特征表示能力
- 多项任务相互促进，实现信息互通与互补
- 基于轨迹在路口的转移信息，基于GCN实现稀疏信息的空间传递



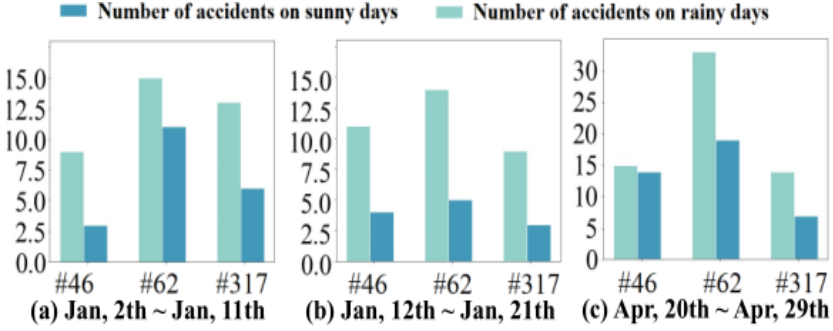
Zhang et, al. Real-time Travel Time Estimation with Sparse Reliable Surveillance Information (Ubicomp 2020).

基于多源数据和多任务的预测

Case2: 基于多源历史数据预测未来多粒度事故分布

Challenges:

- ✓ 多源动态交通信息稀疏；
- ✓ 事故事件本质稀疏，多步时序依赖性不强；
- ✓ 信息多源异构，相关性复杂；
- ✓ 城市各区域对天气的敏感程度不同。



季节性影响下，不同区域对雨水天气敏感性不同（空间异质特性）

基于多源数据和多任务的预测

Case2: 基于多源历史数据预测未来多粒度事故分布

数据预处理: PKDE& ST-DFM

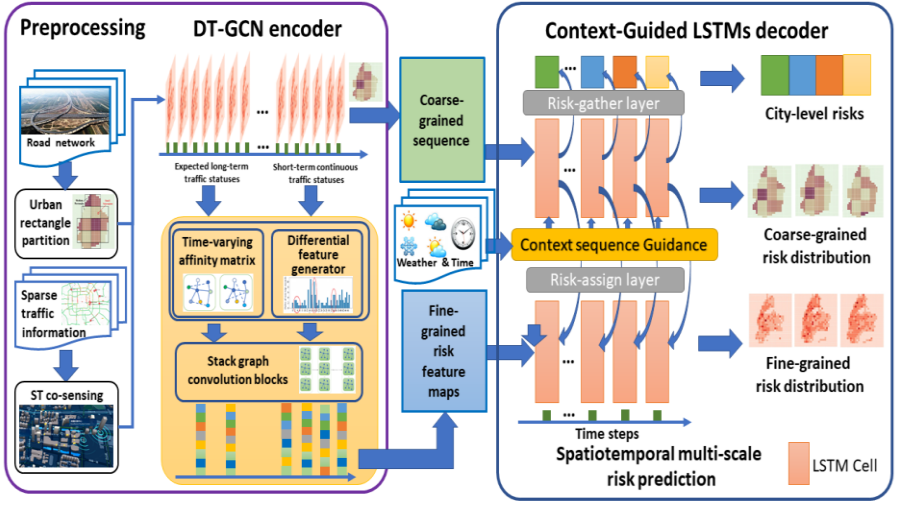
- 空间多粒度网格划分
- 稀疏信息填充与变换

时空建模: DTGCN & CG-LSTM

- 上下文引导的多步预测
- 空间多尺度依赖建模

后处理-事故筛选

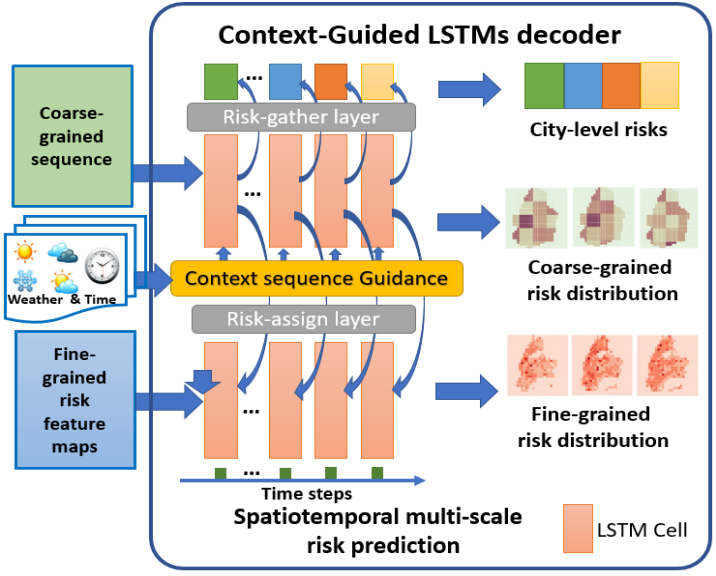
自适应事故区域Ranking机制



TKDE'21: Foresee Urban Sparse Traffic Accidents: A Spatiotemporal Multi-Granularity Perspective

基于多源数据和多任务的预测

Case2: 基于多源历史数据预测未来多粒度事故分布



事件稀疏与零膨胀

空间异质性、多步时序依赖性低

上下文引导的LSTM
Context-guided LSTM

- 空间多粒度的多任务预测
- 粗粒度事故分布视为中间学习信息

- 引入天气时间上下文进行逐步引导
- 中间粗粒度风险信息传递至细粒度并融合预测

基于多源数据和多任务的预测

Case3: 基于大规模稀疏轨迹的细粒度轨迹预测

稀疏位置信息+有限细粒度轨迹

➡ 未来细粒度轨迹预测

稀疏轨迹（粗粒度）：车辆过卡口记录

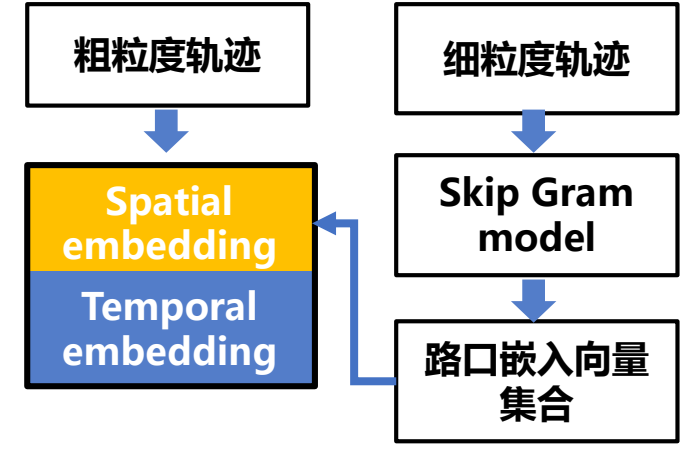
（具有摄像头才记录，时空均稀疏采样）

细粒度轨迹：出租车GPS记录

（秒级上传，密集采样）

TrajForesee: How limited detailed trajectories enhance large-scale sparse information to predict vehicle trajectories? ICDE 2021

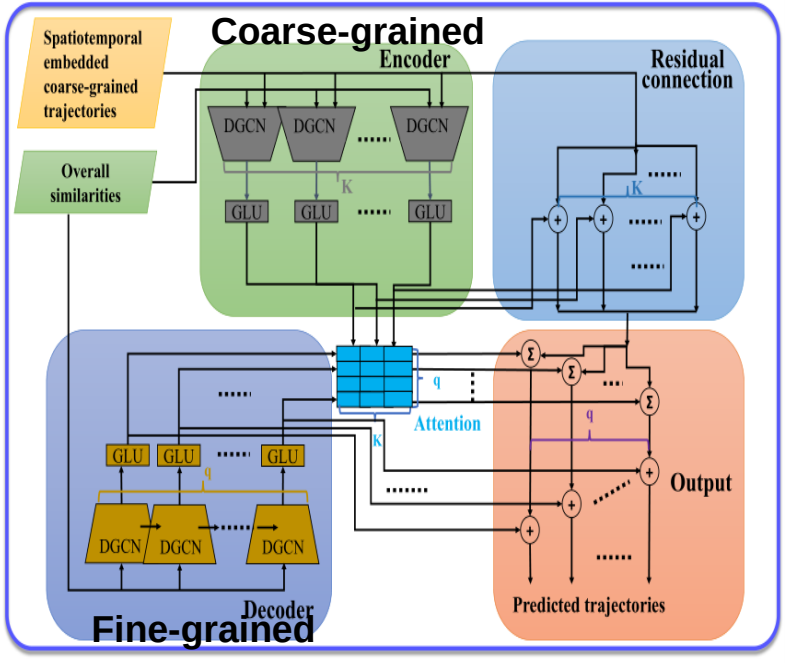
Motivation: 车辆细粒度轨迹在每个路口出现的频率分布和单词在文本中出现频率分布相似。



基于多源数据和多任务的预测

Case3: 基于大规模稀疏轨迹的细粒度轨迹预测

- 基于Skip-gram的**路口**时空语义嵌入
- 两种轨迹相似性
  - 车辆历史行为相似性（最常出现的路口-历史）
  - 车辆单个轨迹相似性（单轨迹路口序列-实时）
- 基于**DGCN2Seq**的编码-解码，映射粗粒度-细粒度轨迹序列
- 轨迹、到达时间、相似性度量多任务预测



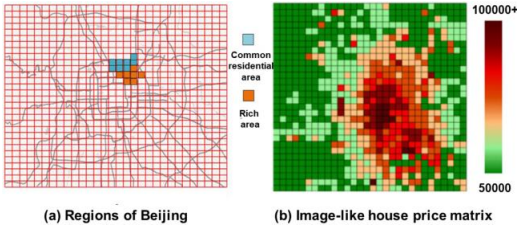
TrajForesee: How limited detailed trajectories enhance large-scale sparse information to predict vehicle trajectories? ICDE 2021

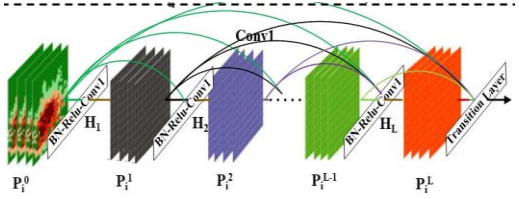
基于多源数据和多任务的预测

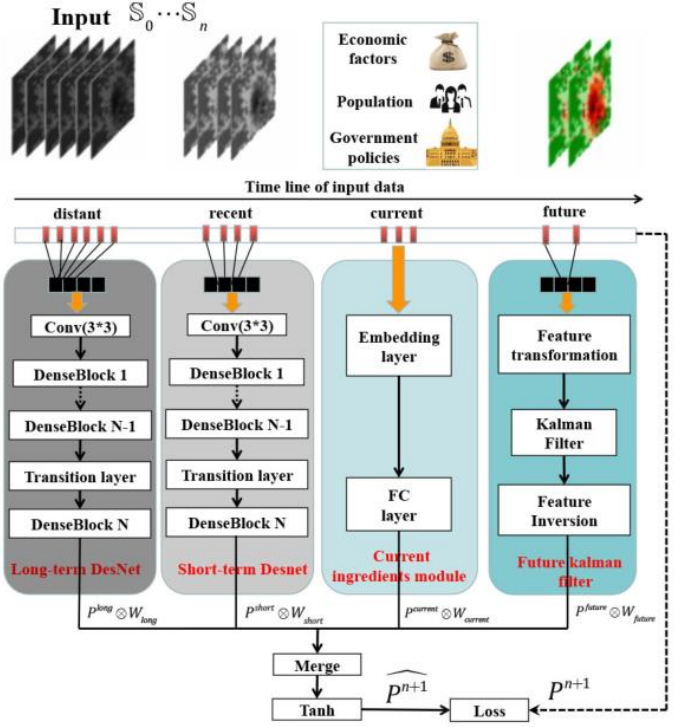
Case4: 具有断续记录的房价预测

- 将房价预测问题的空间粒度提升至英里级别区域
- 基于DenseNet价格趋势拟合-特征重用

  密集连接-元素级相加，缓解时空数据断续稀疏
- 基于卡尔曼滤波的趋势预测
- 经济、社会数据多源数据融合







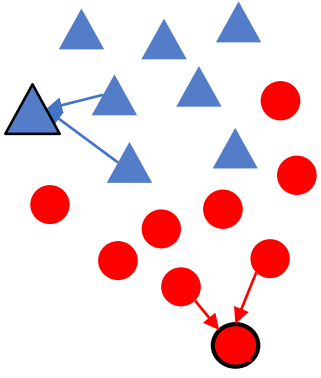
Ge et,al. FTD DesNet: An Integrated Model for Urban Subregion Housing Price Forecasting (ICDM 19)

时空数据挖掘稀疏挑战的主要解决策略

本质稀疏：数据与损失函数变换策略

- 基于先验信息的数据间隔最大化
- 数据增强与样本生成
- 基于样本不平衡问题的缓解策略

本质稀疏：数据与损失函数变换策略



数据增强与新样本生成
类内样本加权->新样本

样本重采样
over-sampling
under-sampling

权值重分配
基于样本数
基于学习分类难度
Focal loss

基于样本不平衡问题的缓解策略

Q Zhong, et, al. Towards good practices for recognition & detection (CVPR workshops 2016)
Tsung-Yi Lin, et, al. Focal loss for dense object detection. (IEEE ICCV 2017)

基于先验信息的数据间隔最大化

Zhou Z, et, al. Foresee Urban Sparse Traffic Accidents: A Spatiotemporal Multi-Granularity Perspective(TKDE 2021).

本质稀疏：基于先验信息的数据变换策略

Motivation

- ✓ 扩大正负样本距离，显著区分潜在风险不同的区域
- ✓ loss与label的一致性

Case: 基于多源历史数据预测未来多粒度事故分布

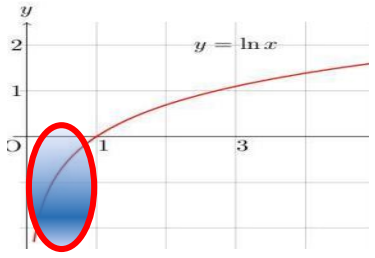
基于数据集先验信息的本质稀疏缓解方法

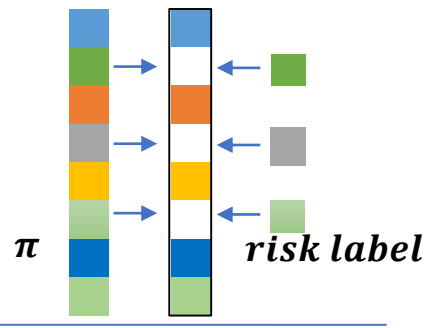
Step1: 统计每个区域在这个数据集上的事故总数，并归一化为0~1之间的概率值ε_{v_i}；

Step2: 将ε_{v_i}利用对数log 转化成一个负数，并且用参数b1、b2来使其和正的risk风险值相一致，如正的风险值在0-5之间，那么负数值也在-5~0之间。

$$\varepsilon_{v_i} = \frac{1}{N_{week}} \sum_{j=1}^{N_{week}} \frac{r_{v_i}(j)}{\sum_{k=1}^{m} r_{v_k}(j)}$$

$$\pi_{v_i} = b_1 \log_2 \varepsilon_{v_i} + b_2$$





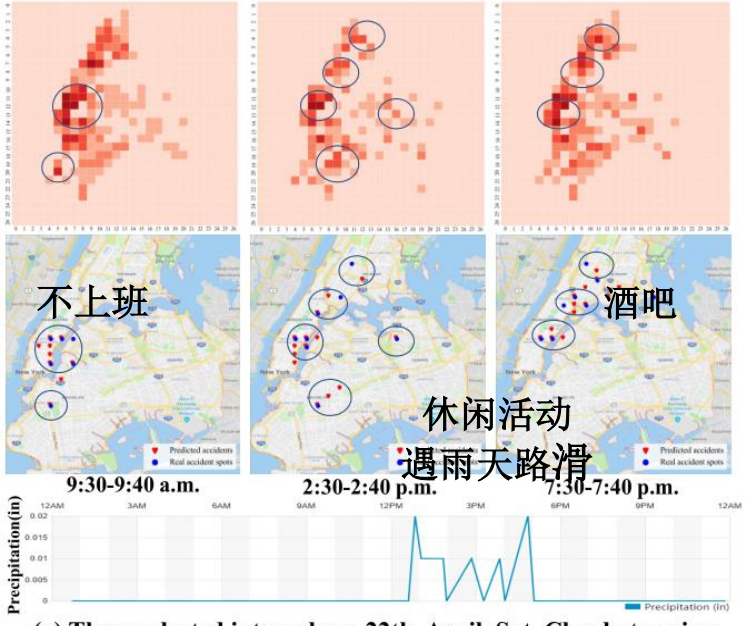
Case study: 时空交通事故预测

案例分析：

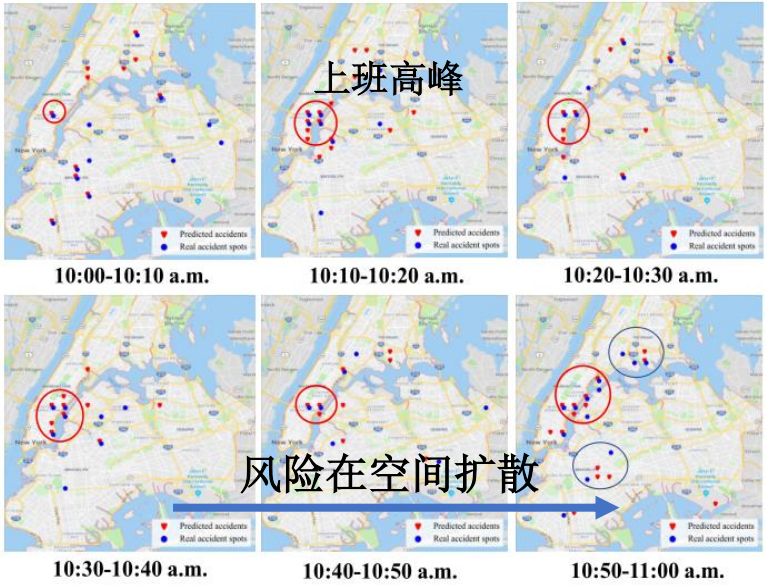
√ 预测事故风险分布大致吻合，筛选结果较好

√ 高风险区域呈现明显的时变特性

√ 预测结果能够跟随上下文变化



(a) Three selected intervals on 22th, April, Sat, Cloudy to rainy



(b) Sequential results on 3rd, April, Mon, Rainy

预测准确率瓶颈与时空不确定性

事故时空预测准确率瓶颈

约为55%（击中率=Top-k区域准确预测/所有事故区域）

因：事件“多因一果” 特性，突发性与偶然性

可预测性与不确定性

猜想： 连续时空元素的规律性强？可预测性大？e.g., 速度，流量

离散事件受多种不可控因素影响，可预测性较小？ e.g., 事故，犯罪事件

基于稀疏数据源的预测不确定性大（本质稀疏，伪稀疏，两者并存）？

哪些数据是可预测的，给定的这些多源信息，他的可预测度是多少?

不确定性来源与稀疏下的不确定性

- **人类活动等时空数据蕴藏高度的不确定**
  - 个体活动的不确定：情绪影响、突发事件
    新地点探索、出行取消
  - 环境因素的不确定：自然环境、社会环境
    跟风效应、管制措施
  - 数据采集不确定：采集噪声、数据稀疏
- **数据愈稀疏，可获得信息量愈小，模型难以捕获内在规律，学习过程不确定越大**

Captured information and traffic volume analysis of a camera

不确定性分类与量化方法

[1] Yongqi Liu, et, al. Probabilistic spatiotemporal wind speed forecasting based on a variational Bayesian deep learning model (Applied Energy 2020).
[2] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? (NIPS 2017)
[3] Balaji Lakshminarayanan, et, al. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles (NIPS 2017).
[4] Lingkai Kong, et, al. SDE-Net: Equipping Deep Neural Networks with Uncertainty Estimates (ICML 2020).

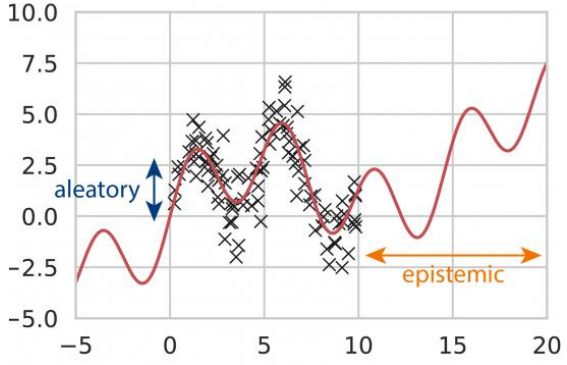
两种不确定

认知不确定 Epistemic uncertainty

模型尚未学到，可增加学习样本减小

偶然不确定 Aleatoric uncertainty

输入信息的噪声，可建模，不可约减



不确定性量化方法

贝叶斯深度学习(BNN)：把模型参数视作一个分布，Dropout/变分[1-2]；

深度集成学习(DeepEnsemble)[3]：不同的初始化模型，并进行集成；

非贝叶斯学习

基于神经随机微分方程的不确定性估计[4]：扰动学习（布朗运动）。

稀疏感知与不确定性思考

- **基于可预测性和规律性，在不确定性建模中引入时间维度和空间依赖，探索在时空上下文中，个体与集体活动随时空演变的不确定性。**
  - **一个新视角：内部不确定与外部不确定**

    内部不确定：Task-specified数据本身观测产生，噪声评估与数据质量估计

    外部不确定：上下文等多元外部因素产生的交互影响

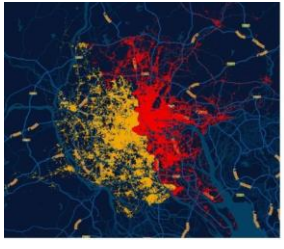
    不确定的时序依赖性：不确定性在时序上的演化和空间上的传播特性
  - **探索稀疏性、不确定性关联与模型稳定性关联**

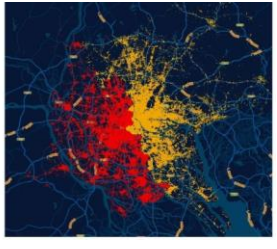
目 录

时空数据计算前沿研究

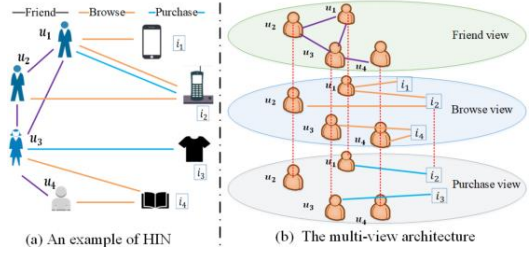
时空数据的安全性

个人轨迹、行程、POI签到的私有性（**差分隐私**）





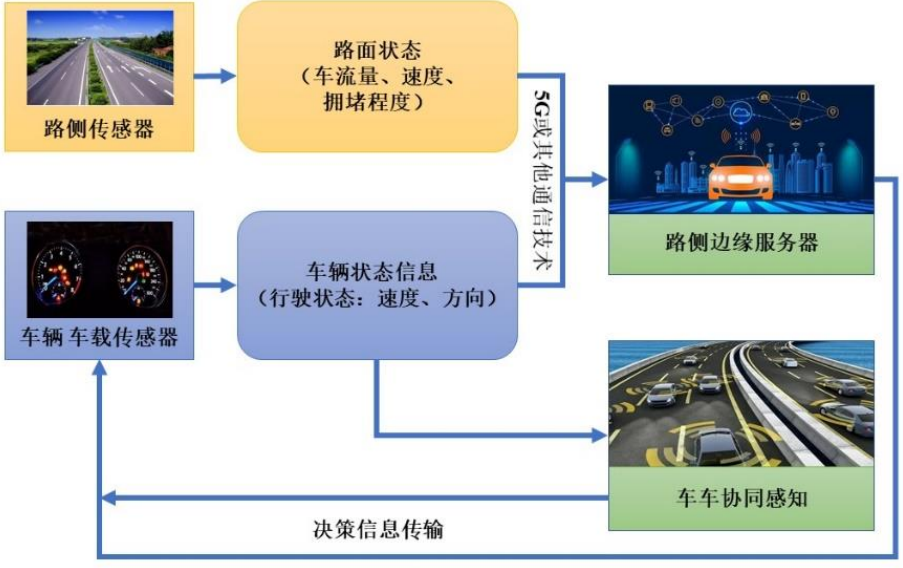
异质图信息网络、知识图引导、元学习



时空数据计算前沿研究

云计算与边缘计算

减轻终端计算负载，将部分共性的识别、发现和决策的计算负载迁移至边缘服务器，由其完成智能辅助驾驶的相关计算。



典型应用：车路协同

- 本地计算
- 任务调度
- 协同计算
- 目标：通信时间总和最短

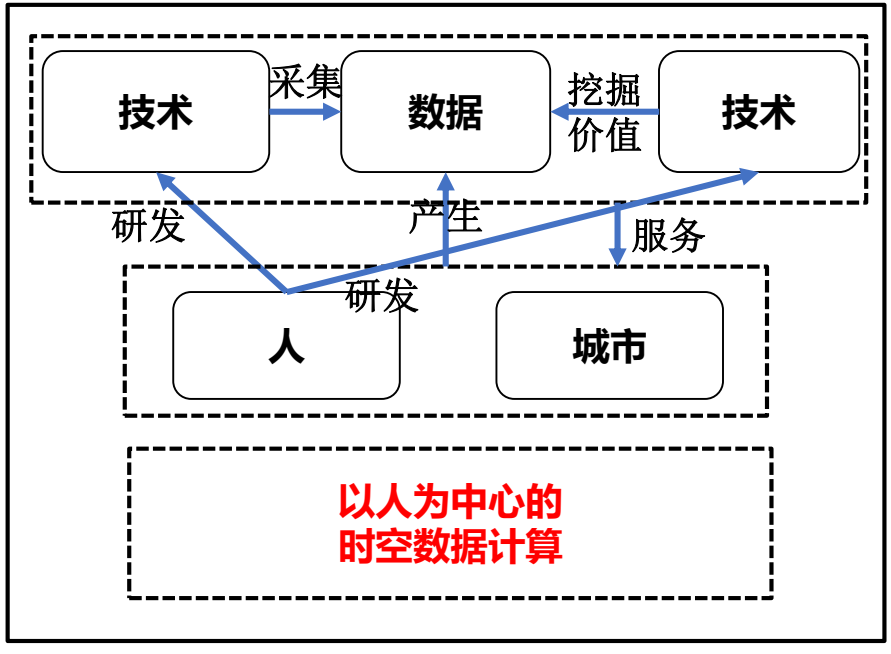
Take-away message

技术，人与城市

- 算法、数据服务于人类幸福与城市智慧化
- 技术推动城市治理精细化、科学化、现代化

经济价值
社会价值

事故预测 环境估计
轨迹预测 流量推断
订单匹配 不确定性估计
房价预测 可预测性分析



THANKS!

Q&A

祝大家新年快乐，万事胜意

Paper多多！

学术主页：

http://staff.ustc.edu.cn/~angyan/ 汪 炀

http://home.ustc.edu.cn/~zzy0929/Home/ 周正阳

联系方式：

Email: angyan@ustc.edu.cn 汪 炀

zzy0929@mail.ustc.edu.cn 周正阳