

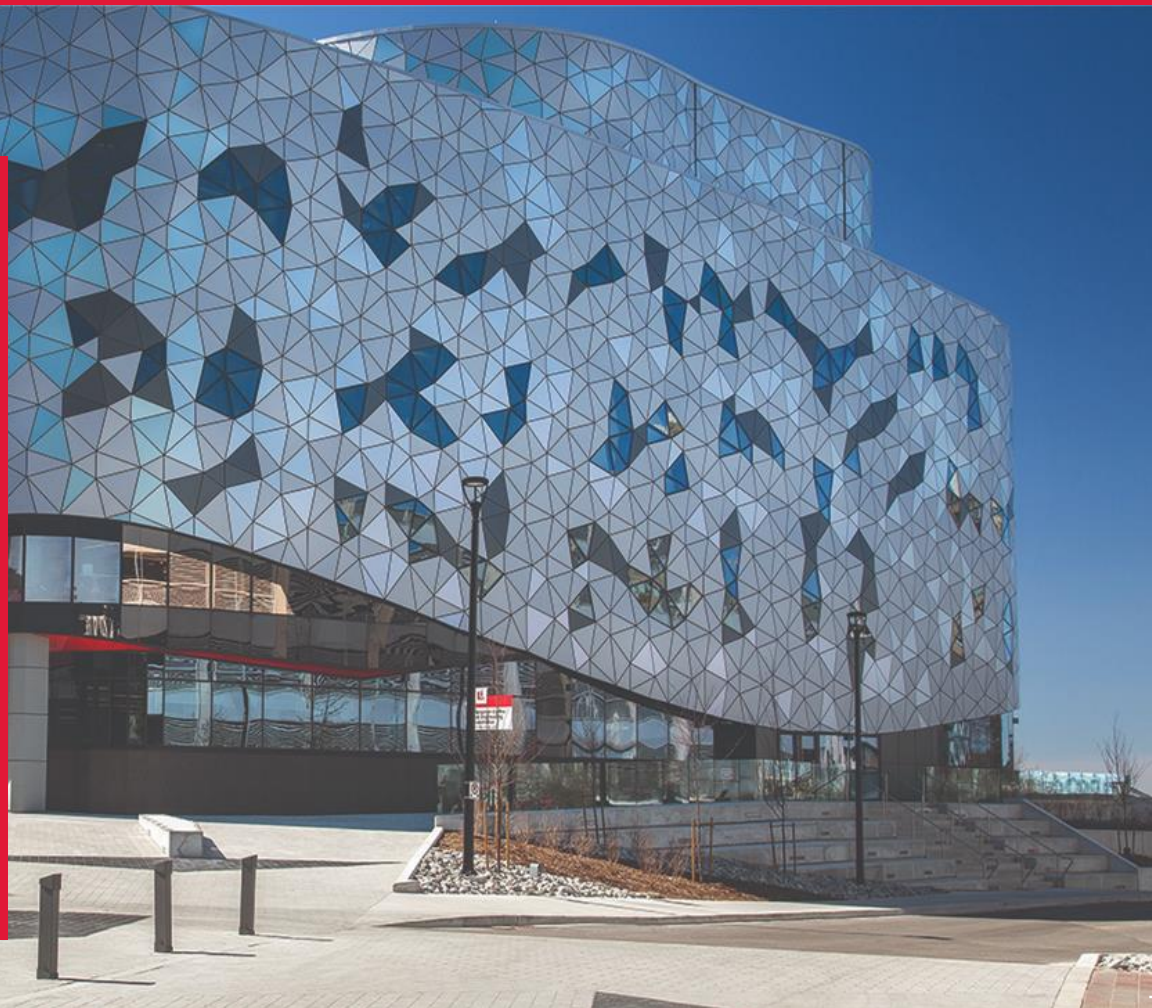
Trajectory Data Mining in the Age of Big Data and AI

Ontario Database Day
(OnDBD 2023)

Manos Papagelis

Wed, Dec 13, 2023
McMaster University

YORK 



Background & Motivation

Trajectory/Mobility Data

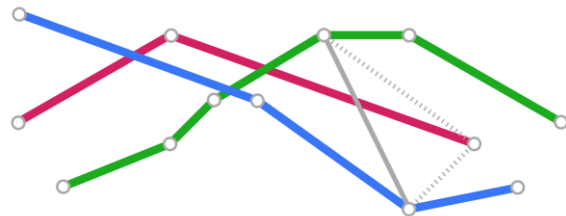
Trajectory: A Sequence of (Spatiotemporal) Points



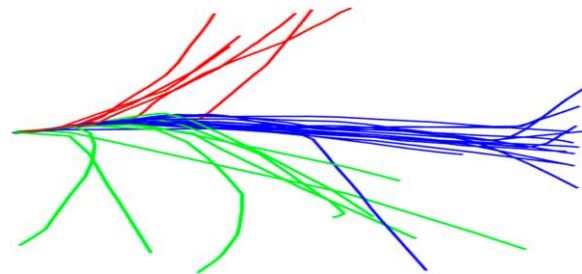
Vast Amounts of Trajectory/Mobility Data



Trajectory Data Mining



trajectory similarity



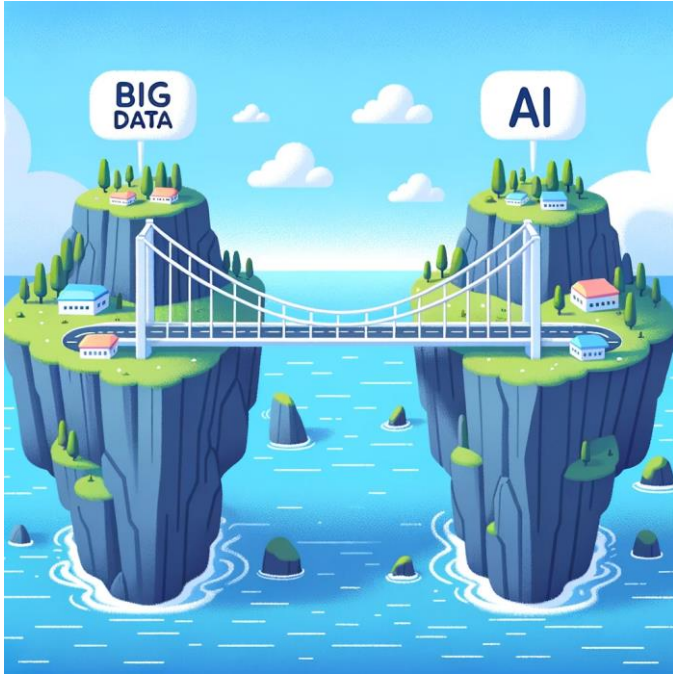
trajectory clustering

trajectory anomaly detection
trajectory network mining
trajectory classification

...

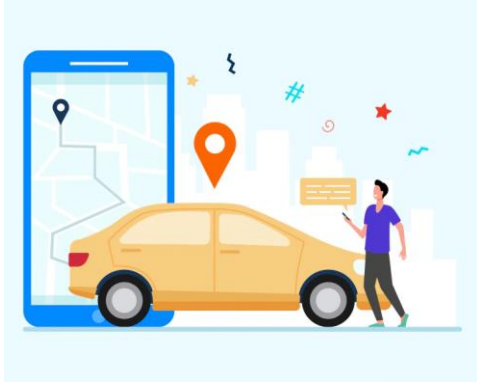
challenging computational problems

Trajectory Data Mining in the Age of Big Data and AI



a symbiotic relationship that presents a new strategy for addressing complex problems in trajectory data mining

Plethora of Applications



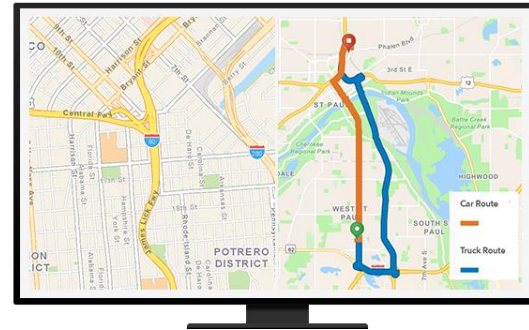
ridesharing



trip/POI (point-of-interest) recommendation



traffic analysis



route planning and optimization

Data Mining Lab @ YorkU's Journey on Trajectory Data Mining

- Trajectory simplification [ACM SIGSPATIAL '23]
- Trajectory similarity [Submitted]
- Trajectory dataset and resources [ACM SIGSPATIAL '23]
- Trajectory prediction [Submitted]
- Trajectory classification [IEEE MDM '23]
- Trajectory network analysis [Big Data Research, IEEE MDM '20, GeoInformatica, IEEE BigData '18, 2 x IEEE MDM '18]
- Mobility + epidemics [ACM SIGSPATIAL/SpatialEpi '24, ACM SIGSPATIAL/SpatialEpi '23, IEEE MDM '22]
- Transportation optimization [ACM SIGSPATIAL '22, ACM SIGSPATIAL '22]

Trajectory Simplification

The Trajectory Pathlet Dictionary Construction Problem

Trajectories on the Road Network

- Trajectory

- Denoted by τ
- Represented as:

$$\tau = \langle (x_1, y_1, t_1), \dots, (x_{|\tau|}, y_{|\tau|}, t_{|\tau|}) \rangle$$

object's geo-location specific time instance



- Road Network

Modelled as a **graph** $\mathcal{G}\langle \mathcal{V}, \mathcal{E} \rangle$

- $\mathcal{V}$: **Nodes** (set of road intersections)
- $\mathcal{E}$: **Edges** (set of road segments)

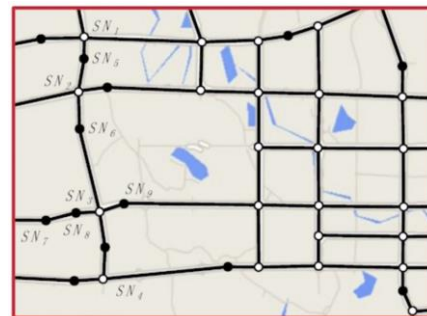
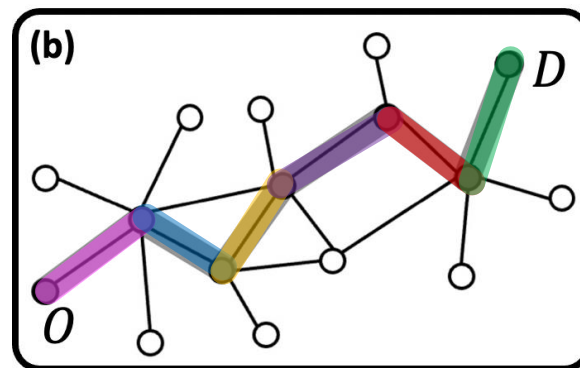
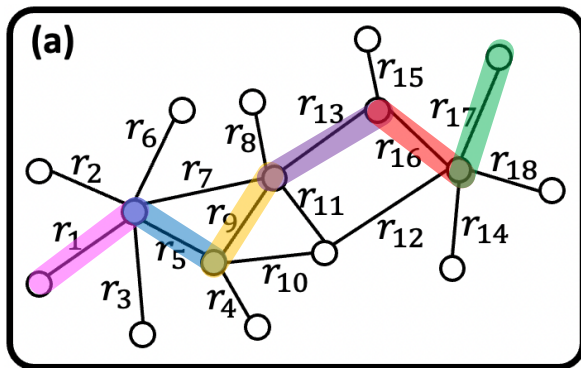


Image Source: "Updating Road Networks by Local Renewal from GPS Trajectories" [Wu et al, MDPI '16]

Road Segment-based Representation

- Each trajectory τ can be expressed as a set of road segments $R_s \subseteq R$
- This special representation is denoted by $\mathfrak{N}(\tau)$



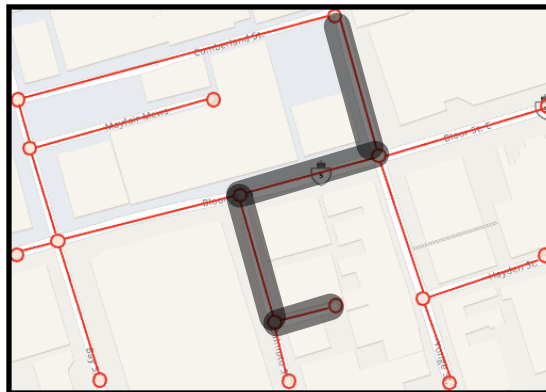
$$\mathfrak{N}(\tau) = \{r_1, r_5, r_9, r_{13}, r_{16}, r_{17}\}$$

Trajectory Pathlet Dictionary (PD) Construction

- **Pathlet Dictionary:** A small set of basic building blocks that can represent a wide range of trajectories
- Many names in the literature

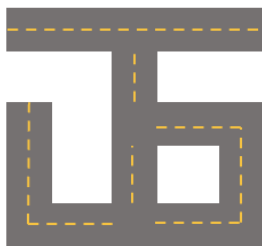
[Panagiotakis et al – TKDE '12, Chen et al – SIGSPATIAL '13, Sankararaman et al – SIGSPATIAL '13, Agarwal et al – PODS '18, Li et al – TSAS '18, Zhao et al – CIKM '18]

- Pathlet
- Subtrajectory
- Trajectory Segments
- Fragments
- ...

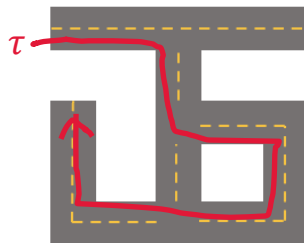


Pathlet-based Representation of a Trajectory

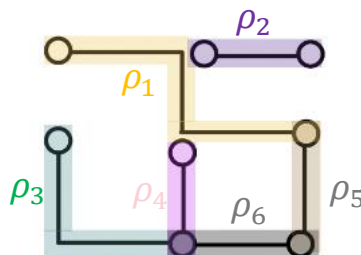
Denoted by $\Phi(\tau) = \{\rho^{(1)}, \rho^{(2)}, \dots, \rho^{(k)}\}$



(a)



(b)

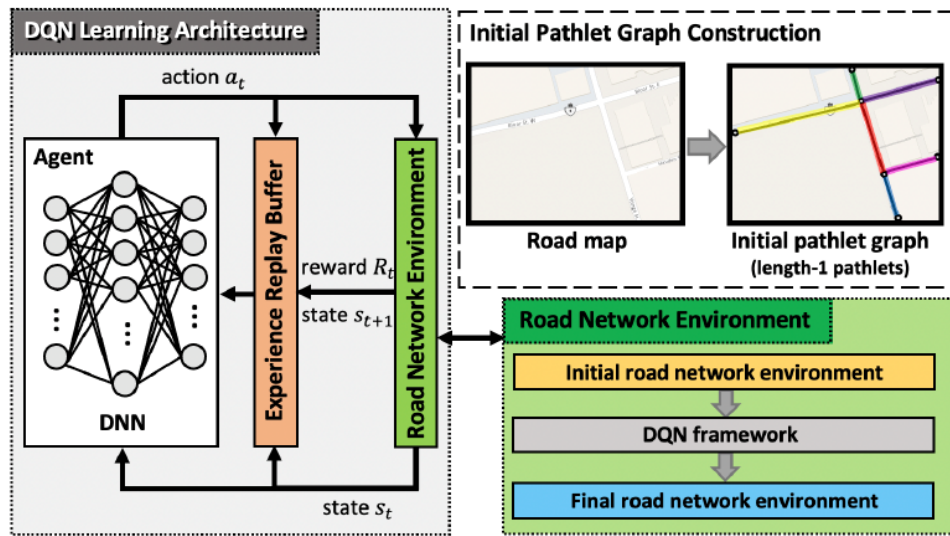


(c)

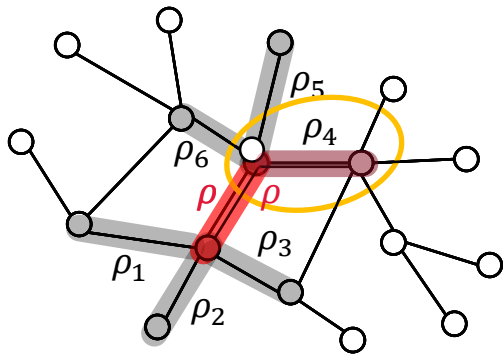
$$\Phi(\tau) = \{\rho_1, \rho_5, \rho_6, \rho_3\}$$

PathletRL - Overview

- Extracting candidate pathlets
- Deep Reinforcement Learning framework



Extracting Candidate Pathlets - Example



Choose pathlet ρ uniformly at random

Identify all neighbors of ρ

Compute utility of ρ_{merge}

Obtain ρ^* with the highest utility

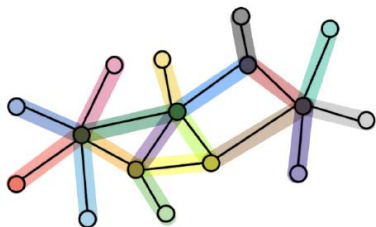
Merge ρ and ρ_4

New current pathlet ρ

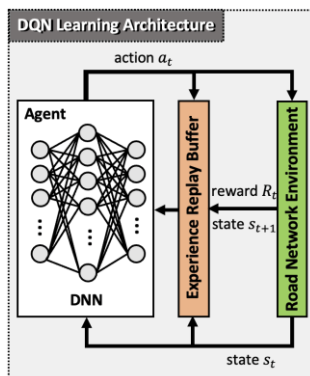


ρ_{merge}	Utility
MERGED(ρ, ρ_1)	+0.7
MERGED(ρ, ρ_2)	+1.8
MERGED(ρ, ρ_3)	-1.6
MERGED(ρ, ρ_4)	+5.5
MERGED(ρ, ρ_5)	-3.2
MERGED(ρ, ρ_6)	+2.9

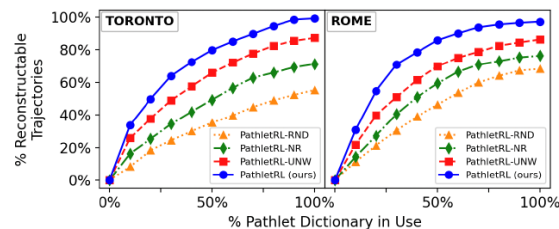
Approach & Contributions



Edge-disjoint pathlets



Deep Reinforcement Learning (DQN)



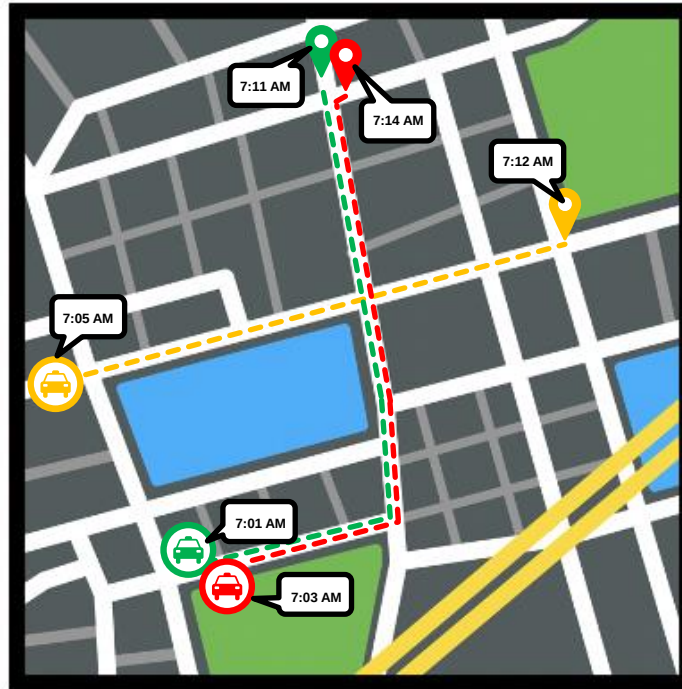
Partial trajectory reconstruction
~85%

Trajectory Similarity

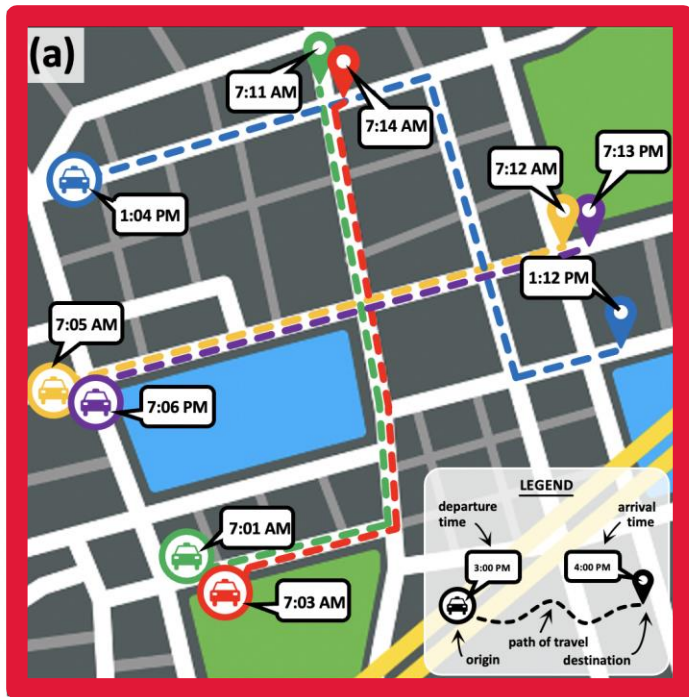
The Top-k Trajectory Similarity Search Problem

Trajectory Similarity

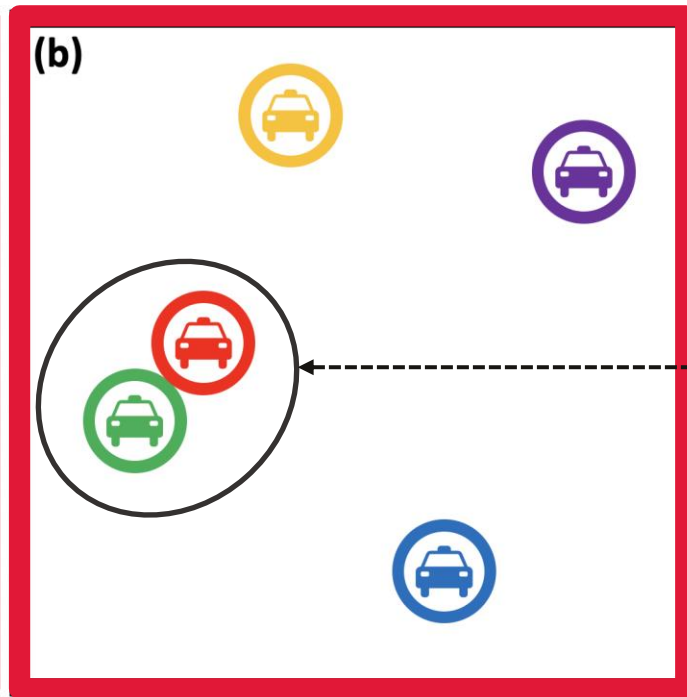
- How similar two trajectories are
- Several ways to define



Spatiotemporal Similarity – Example



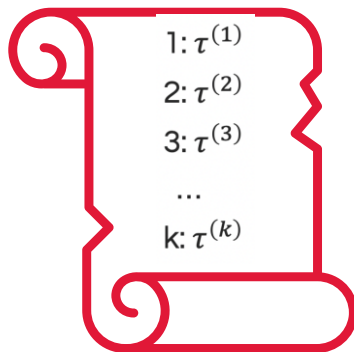
Taxi Trajectories



Embedding Space

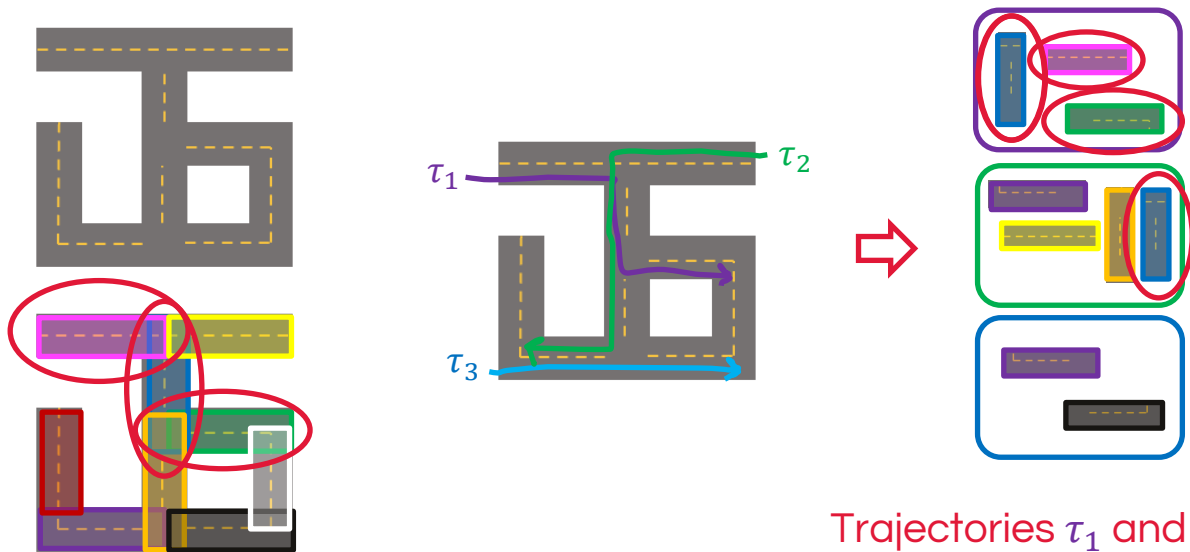
Problem Statement

- Top- k Trajectory Similarity Search Task
 - **Given:** Trajectory set $\mathcal{T}$
Query trajectory τ_q
Positive integer $k \geq 1$
 - Find the (ranked) list of top k trajectories in $\mathcal{T}$:
 - **Criterion:** Similarity with τ_q



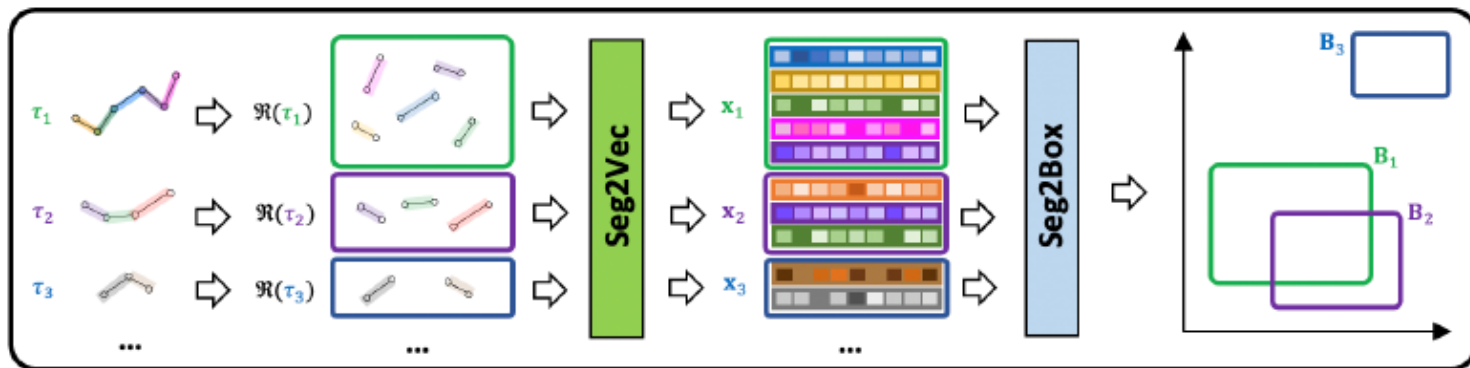
Approach: Reducing Trajectory Similarity to Set Similarity Problem

- Treat **each trajectory** as a **set**; its **elements** are the **road segments** it has traversed (road-based representation $\mathfrak{R}(\tau)$)
- Similar (Dissimilar) trajectories map to similar (dissimilar) sets



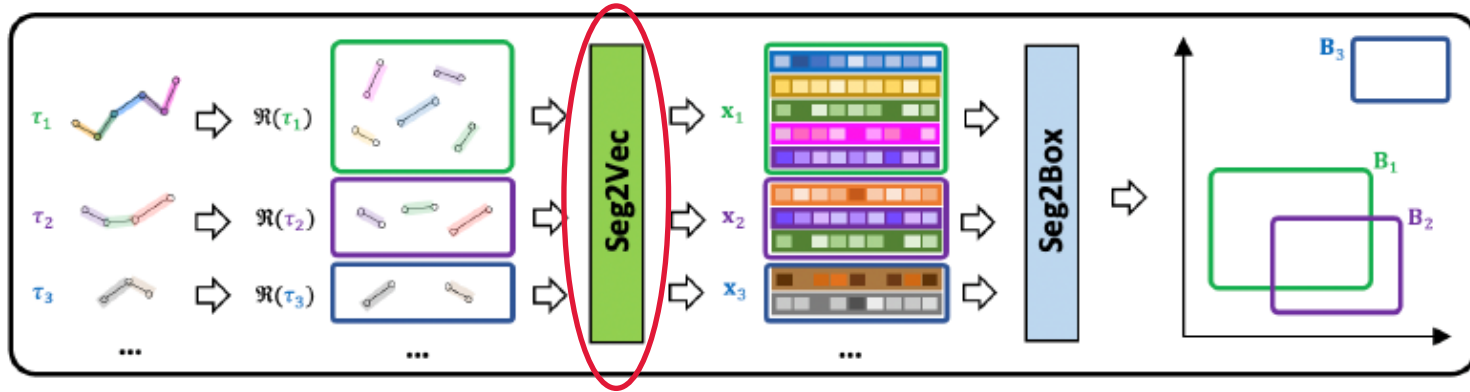
ST2Box Overview

- Spatiotemporal Trajectories to Box Embeddings for Similarity Learning



ST2Box Overview

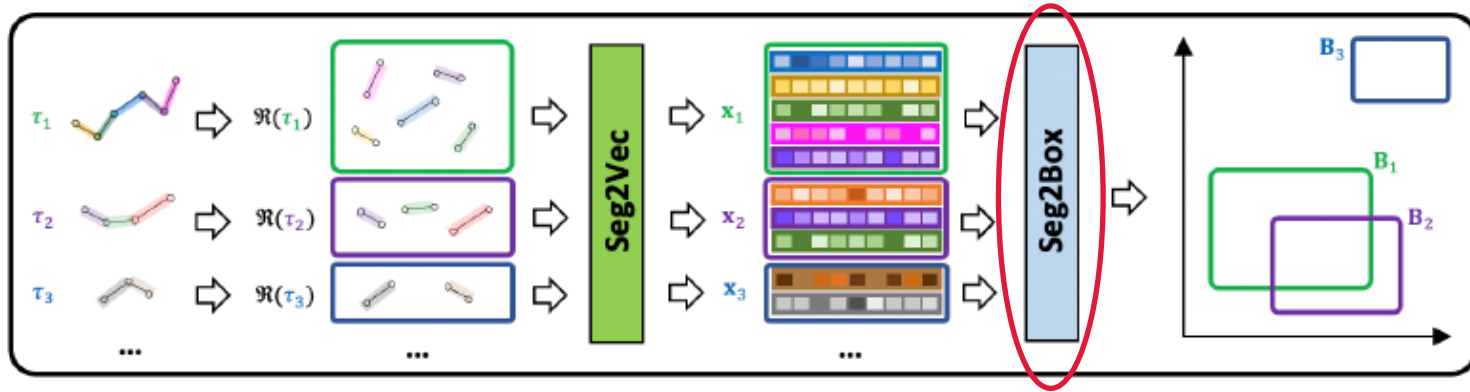
- Spatiotemporal Trajectories to Box Embeddings for Similarity Learning



- Seg2Vec** – spatiotemporal vector representations of road segments

ST2Box Overview

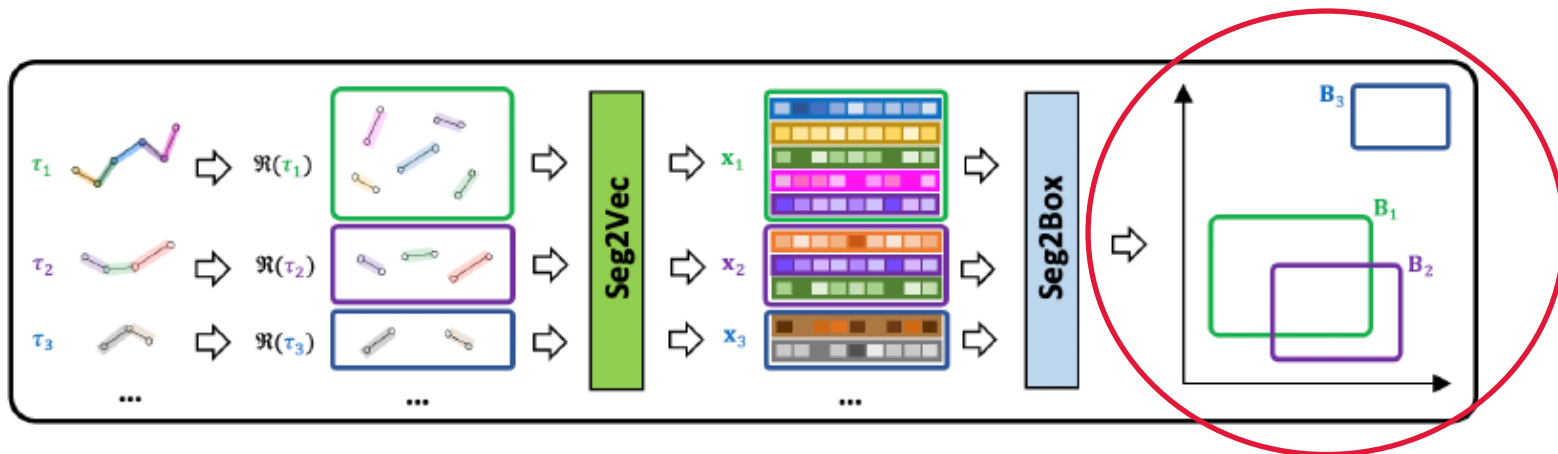
- Spatiotemporal Trajectories to Box Embeddings for Similarity Learning



- Seg2Box** – box representations of sets of road segments

ST2Box Overview

- Spatiotemporal Trajectories to Box Embeddings for Similarity Learning



- Overlapping boxes $\Rightarrow$ Similar sets $\Rightarrow$ Similar trajectories

ST2Box Properties

Accurate, Versatile, Generalizable, Robust, Fast, Scalable

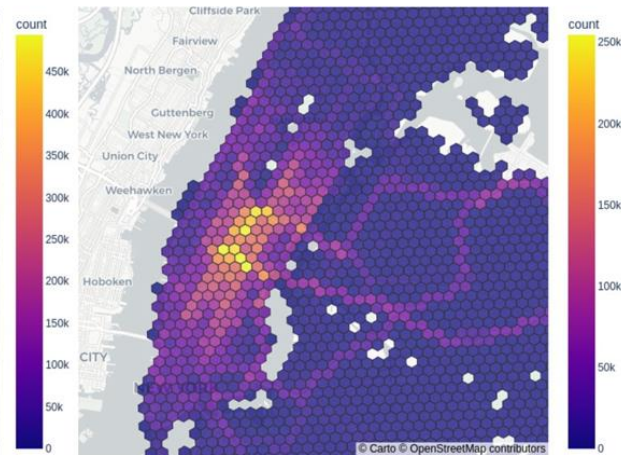
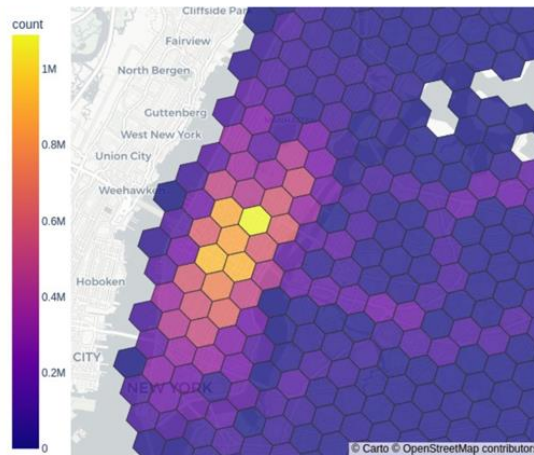
Up to ~30% Performance Gain

Higher-order Mobility Flow Data

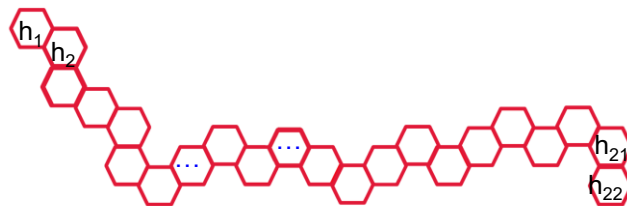
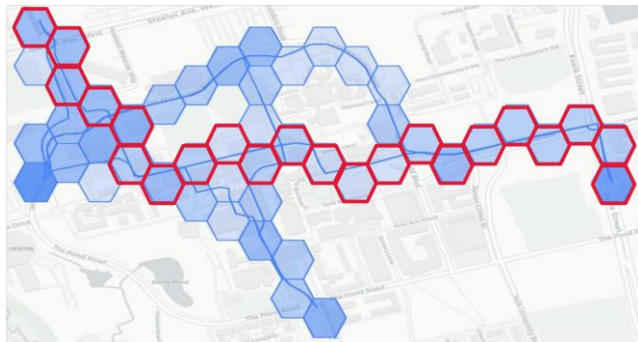
Map Tessellation

lower
resolution

higher
resolution



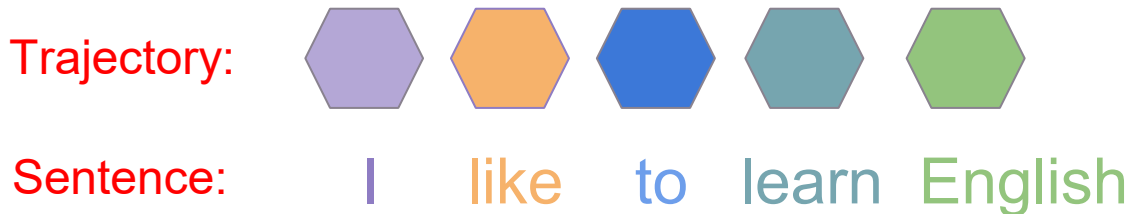
Trajectories: Sequences of Hexagons



Trajectory: $h_1, h_2, h_3 \dots h_{20}, h_{21}, h_{22}$

Treat Trajectories as Language Statements

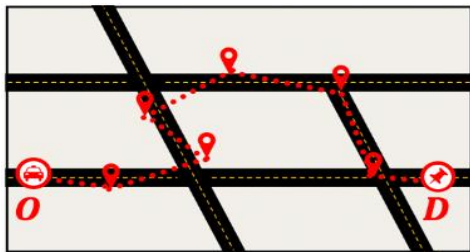
Hexagons represent 'tokens' & trajectories represent 'sentences'



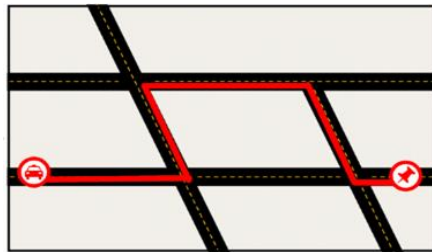
Advantages:

- Reduced data sparsity
- More compatible with well-known ML models (e.g., sequence models, LLMs)

Point2Hex: Overview of the Pipeline



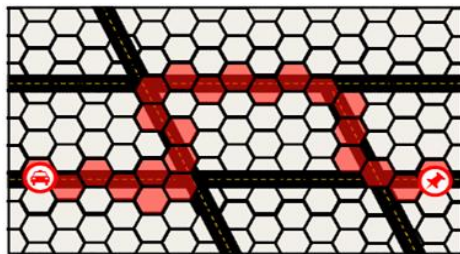
GPS Traces or POI
Check-Ins
(input)



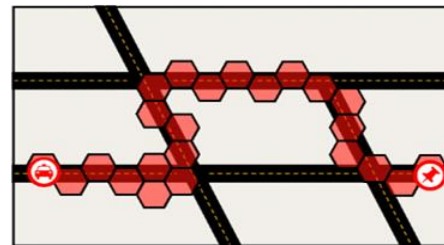
Linestring of
Trajectories
(Map-matching)



Map Tessellation with
Trajectories
(Hexagon-shaped cells)



Intersection of Linestrings and Polygons
(Computational Geometry)



Higher-order Mobility Flow
(Output)

Higher-order Mobility Flow: Datasets and Data Generator

Dataset	Trajectories	Time Period	Resolutions
HO-T-Drive	65,117	02/02/08 - 02/08/08	{6,...10}
HO-Porto	1,668,859	07/01/13 - 06/30/14	{6,...10}
HO-Rome	5,873	02/01/14 - 03/02/14	{6,...10}
HO-GeoLife	2,100	04/01/07 - 10/31/11	{6,...10}
HO-FourSquare-NYC	49,983	04/12/12 - 02/16/13	{6,...10}
HO-FourSquare-TKY	117,593	04/12/12 - 02/16/13	{6,...10}
HO-NYC-Taxi	2,062,554	01/01/16 - 06/30/16	{6,...10}



Datasets @ [Zenodo](#)

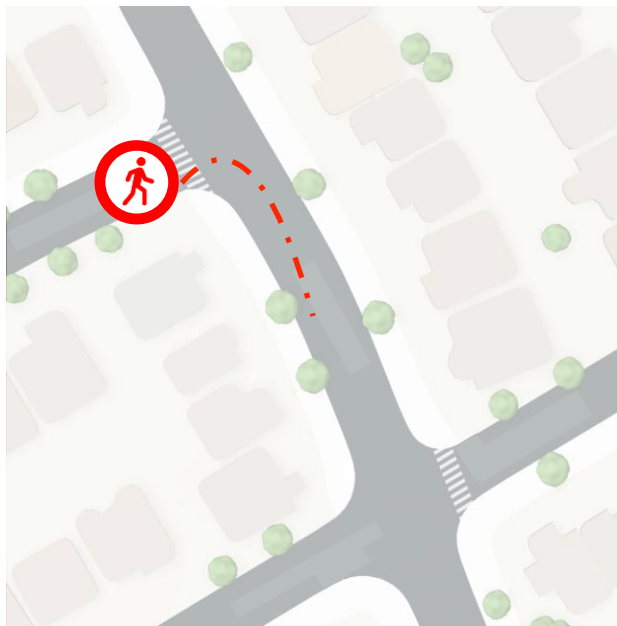


Data Generator @ [GitHub](#)

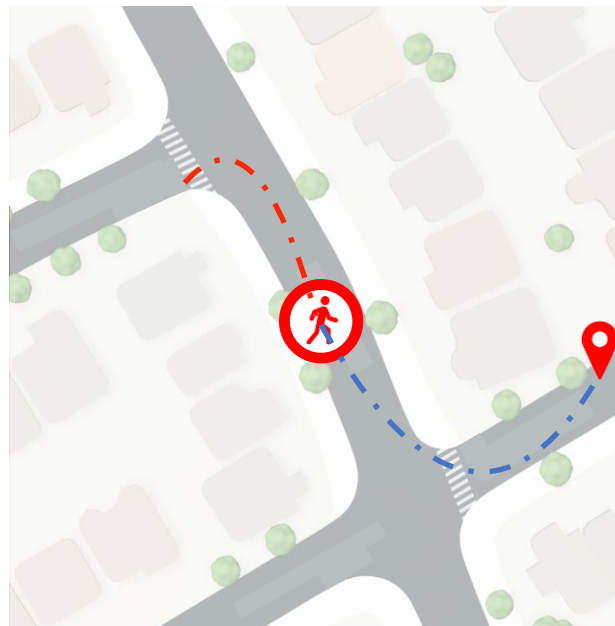
Trajectory Prediction

Predict the Next-k Trajectory Steps Problem

Problem of Interest: Trajectory Prediction



History trajectory



Predict future trajectory

Trajectory Prediction (Revisited)

Let

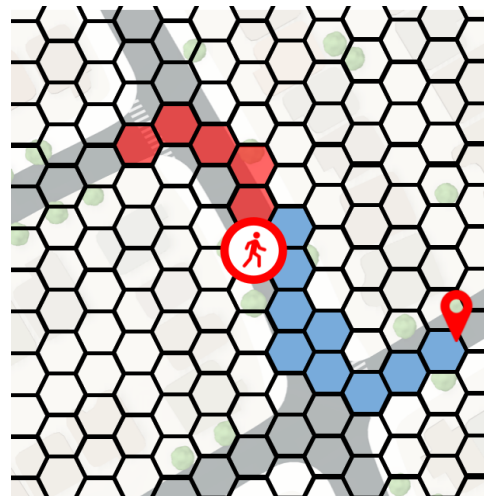
- an observation area
- a set of objects and their history trajectories
- an observation period

Input: Given

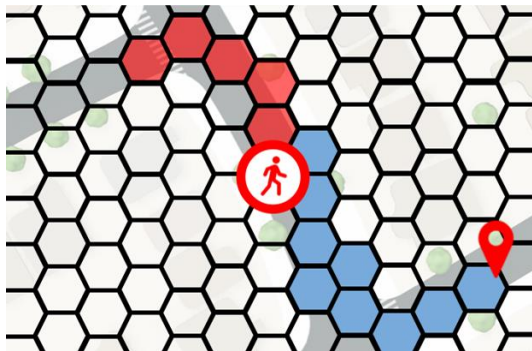
- a moving object n
- a partial trajectory = $\langle p_1, p_2, \dots, p_t \rangle$
- a prediction horizon $k > 0$

Output: We want to

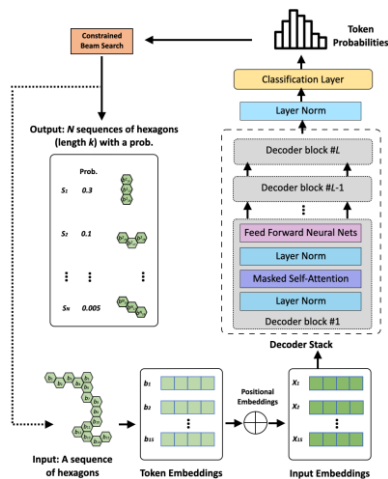
predict the **next k hexagons** of the input partial trajectory



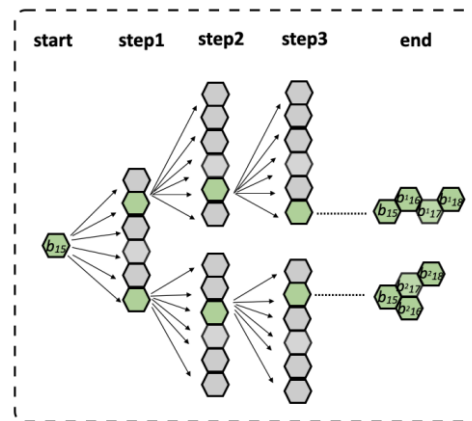
Approach & Contributions



Trajectory Prediction
(Revisited)



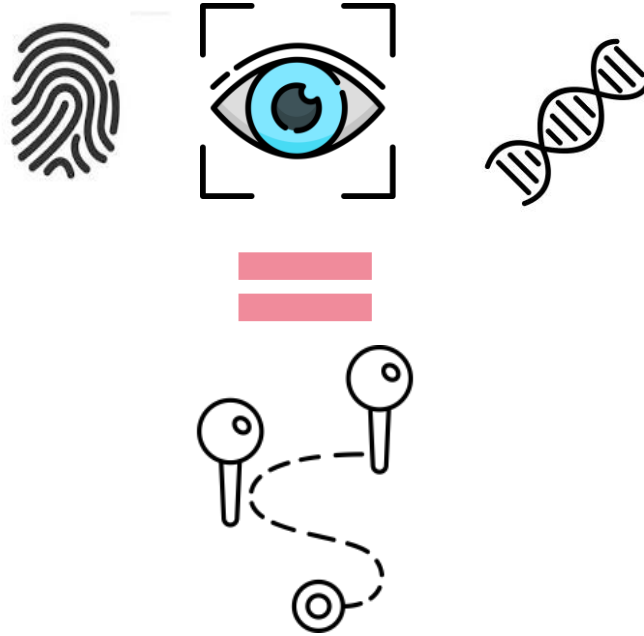
TrajLearn: Trajectory
Deep Generative Model



Beam search

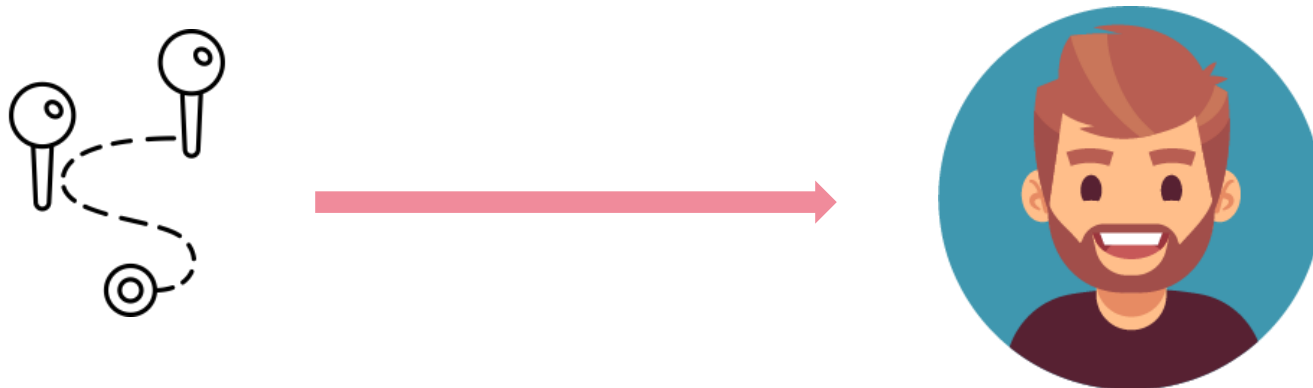
Trajectory Classification

The Trajectory-User Linking Problem



can trajectories
help to **identify** a person?

Trajectory-user Linking (TUL)



trajectory-user linking **aims at linking** anonymous trajectories to users who generate them

Problem Definition

Trajectory-user linking aims at linking anonymous trajectories to users

Given:

$\mathcal{U} = \{u_1, u_2, u_3, \dots, u_c\}$ – users

$\mathcal{T} = \{Tr_1, Tr_2, \dots, Tr_n\}$ – unlinked trajectories

TUL is defined as **a multiclass classification problem**

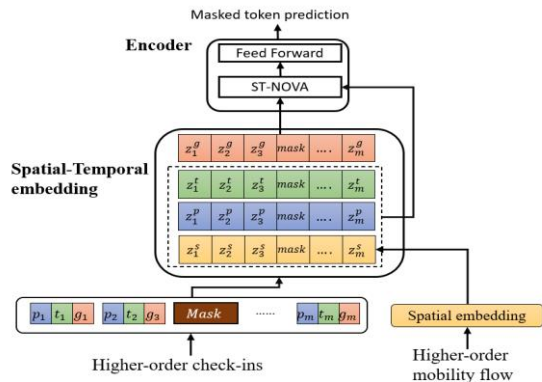
$$\min_{f \in \mathcal{F}} \mathbb{E}[\mathcal{L}(f(Tr_i), u_i)] \text{ over } \mathcal{F}$$

*where $\mathcal{F}$ is the set of all classifiers in the hypothesis space
 $\mathcal{L}(\cdot)$ is the loss between the predicted label $f(Tr_i) \in \mathcal{U}$ and the true label $u_i \in \mathcal{U}$*

Approach & Contributions



Higher-order mobility flow data generation



TULHOR: A spatiotemporal model that deals with sparsity and low data quality of the TUL problem

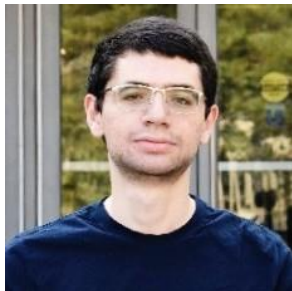
TULHOR **outperforms** baselines by up to 8%

Questions

Credits



Gian Alix



Mahmoud Alsaeed



Ali Faraji



Jing Li



Nina Yanin



Amirhossein Nadiri

PathletRL: Trajectory Pathlet Dictionary Construction using Reinforcement Learning. G. Alix, M. Papagelis. **ACM SIGSPATIAL 2023.**

Trajectory-User Linking using Higher-order Mobility Flow Representations. M. Alsaeed, A. Agrawal, M. Papagelis. **IEEE MDM 2023.**

Point2Hex: Higher-order Mobility Flow Data and Resources. A. Faraji, J. Ling, G. Alix, M. Alsaeed, N. Yanin, A. Nadiri, M. Papagelis. **ACM SIGSPATIAL 2023.**

St2Box: Trajectory Similarity Learning using Set to Box Representations. G. Alix, M. Papagelis. **Submitted.**

TrajLearn: Leveraging Generative Models for Trajectory Prediction Learning. A. Nadiri, A. Faraji, J. Ling, M. Papagelis. **Submitted.**

Thank you!