ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

UNIVERSITY OF PIRAEUS

DATA STORIES

Data Science Lab. @ University of Piraeus

18th Hellenic Data Management Symposium
Athens, Greece, July 1–2, 2024

Mobility Data Science: Perspectives and Challenges

Mokbel, M., Sakr, M., Xiong, L., Züfle, A., et al.

ACM Trans. Spatial Algorithms Syst. 10, 2, Article 10 (June 2024)

Presenter:
Yannis Theodoridis, Data Science Lab, Univ. Piraeus

Dagstuhl Seminar 22021
“Mobility Data Science”
Jan. 09-14, 2022

Mobility Data Science: Perspectives and Challenges

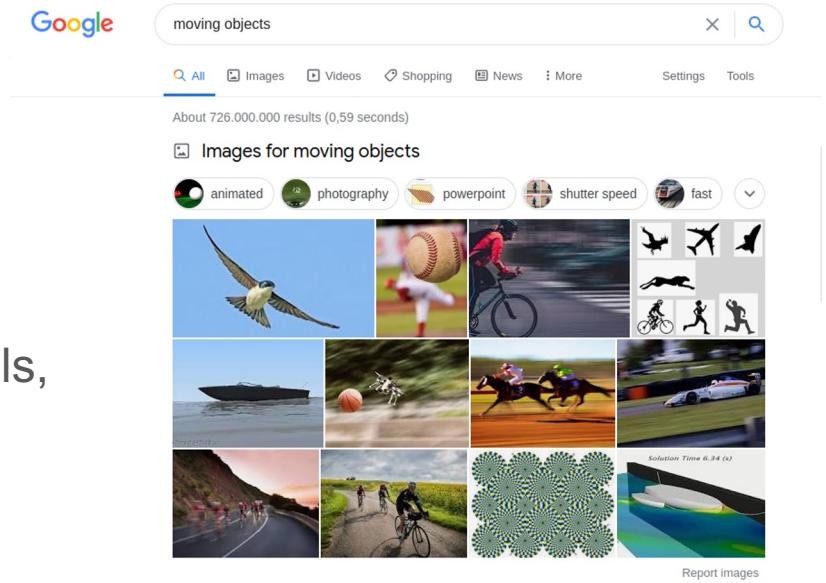
MOHAMED MOKBEL, Computer Science and Engineering, U
States
MAHMOUD SAKR, Université Libre de Bruxelles École polyt
and Ain Shams University, Cairo, Egypt
LI XIONG, Emory University, Atlanta, United States
ANDREAS ZÜFLE, Computer Science, Emory University, Atlan
JUSSARA ALMEIDA, Computer Science, Federal University of
TAYLOR ANDERSON, Department of Geography and Geoinfo
Fairfax, United States
WALID AREF, Computer Science, Purdue University, West Lafay
GENNADY ANDRIENKO, KD, Fraunhofer SIT Saint Augustin
NATALIA ANDRIENKO, KD, Fraunhofer SIT Saint Augustin B
YANG CAO, School of Computing, Kyoto University, Kyoto, Japa
SANJAY CHAWLA, Qatar Computing Research Institute, HBKU
REYNOLD CHENG, Computer Science, HKU, Hong Kong, Hor
PANOS CHRYSANTHIS, Computer Science, University of Pitt
XIQI FEI, Department of Geography and Geoinformation Science
States
GABRIEL GHINITA, Hamad Bin Khalifa University College of
ANITA GRASER, AIT Austrian Institute of Technology GmbH,
DIMITRIOS GUNOPULOS, National and Kapodistrian Unive
CHRISTIAN S. JENSEN, Department of Computer Science, Aa
JOON-SEOK KIM, Geospatial Science and Human Security Divi
Richland, United States
KYOUNG-SOOK KIM, Artificial Intelligence Research Cente
Science and Technology Tokyo Bay Area Center, Koto-ku, Japan
PEER KRÖGER, Kiel University, Kiel, Germany
JOHN KRUMM, Computer Science, University of Southern Cali
JOHANNES LAUER, HERE Technologies, Schwalbach am Taun
AMR MAGDY, Computer Science and Engineering, University of
MARIO NASCIMENTO, Khoury College of Computer Sciences,
SIVA RAVADA, Oracle, Nashua, United States

ACM 2374-0361/2024/5-ART
https://doi.org/10.1145/3652158

MATTHIAS RENZ, Computer Science, Kiel University, Kiel, Germany
DIMITRIS SACHARIDIS, Data Science Lab, Université Libre de Bruxe
Bruxelles, Belgium
FLORA SALIM, School of Computer Science and Engineering, Univ
Australia
MOHAMED SARWAT, Arizona State University, Tempe, United State
MAXIME SCHOEMANS, Data Science Lab, Université Libre de Bruxe
Bruxelles, Belgium
CYRUS SHAHABI, Computer Science, University of Southern Califor
BETTINA SPECKMANN, Technische Universiteit Eindhoven, Eindho
EGEMEN TANIN, University of Melbourne, Australia, Melbourne, Aus
XU TENG, Iowa State University, Ames, United States
YANNIS THEODORIDIS, University of Piraeus, Piraeus, Greece
KRISTIAN TORP, Aalborg Universitet, Aalborg, Denmark
GOCE TRAJCEVSKI, Iowa State University, Ames, United States
MARC VAN KREVELD, Information and Computing Sciences, Utrech
Netherlands
CAROLA WENK, Department of Computer Science, Tulane University
MARTIN WERNER, School of Engineering and Design, Techincal Univ
RAYMOND WONG, The Hong Kong University of Science and Techn
SONG WU, Data Science Lab, Université Libre de Bruxelles École po
Belgium
JIANQIU XU, College of Computer Science and Technology, Nanjing Un
tics, Nanjing, China
MOUSTAFA YOUSSEF, Computer Science and Engineering, The Ame
DEMETRIS ZEINALIPOUR, Department 5:Databases and Informati
Informatics, Saarbrücken, Germany and Computer Science, University of
MENGXUAN ZHANG, School of Computing, Australian National Un
ESTEBAN ZIMÁNYI, Data Science Lab, Université Libre de Bruxell
Bruxelles, Belgium

*This is a community publication. The authors of this article met in Dagstuhl for Seminar #2
four authors co-organized the Dagstuhl seminar leading to this article and coordinated the
contributed equally to this research. The seminar was held in the week of January 9 - 14 , 202
topics: data management, mobility analysis, geography, privacy, urban computing, systems,
integration, and theory. Due to COVID-19, the seminar took place in hybrid mode, with 8
challenge of different time zones of the participants, all sessions were attended by at least 37

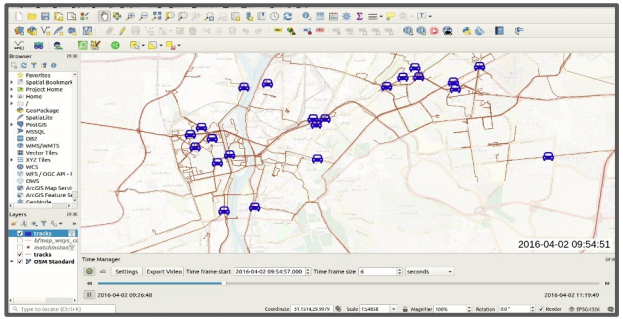
Moving Objects



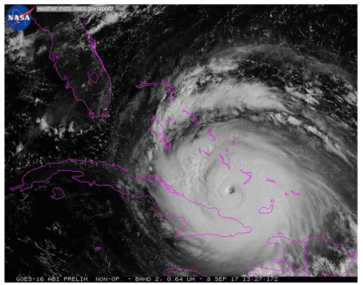
Informal definition: **Moving objects are objects that move in the space over time**

- **points** that change location (e.g., animals, sport players, vehicles, drones, …)
- **regions** that change location and, eventually, extent (hurricanes, etc.)

Most common sources: sensors (GPS), as well as radar, bluetooth, RFID, etc.



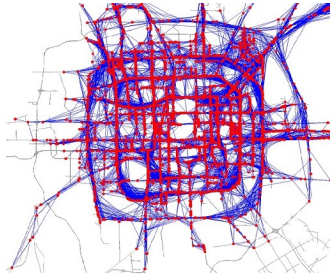
(source: mobilitydb.com)



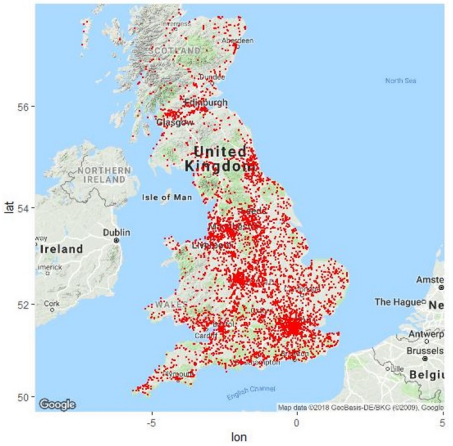
(source: giphy.com)

Examples of Mobility Data

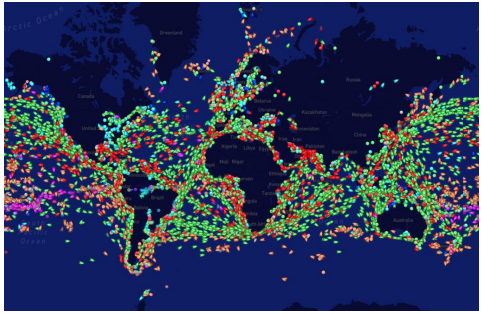
- ➔ Transportation (urban/maritime/aviation)
- ➔ Movement in indoor environments
- ➔ Location-Based Social Networking (LBSN), etc. etc.



T-Drive Urban traffic data
(source: research.microsoft.com)

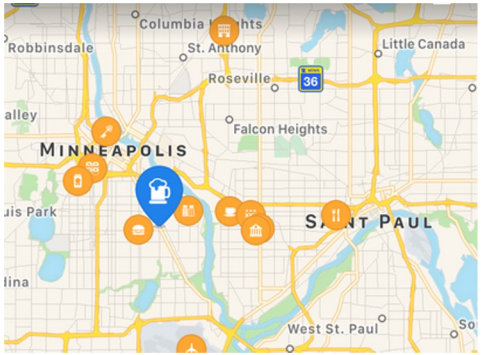


UK road accidents 2012-14
(source: kaggle.com)



Marine transportation data
(source: marinetraffic.com)

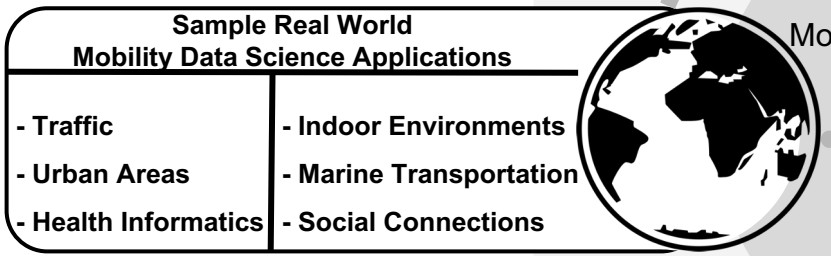
Aviation data
(source: flightradar24.com)

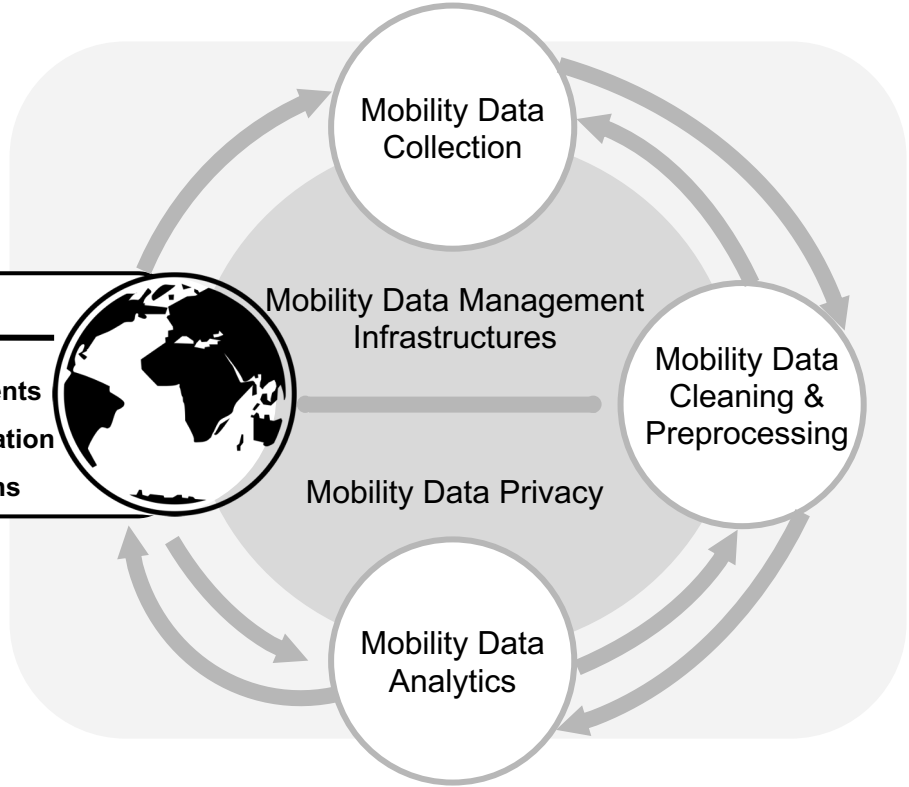


Foursquare check-ins
(source: foursquare.com)

“Mobility Data Science” Definition

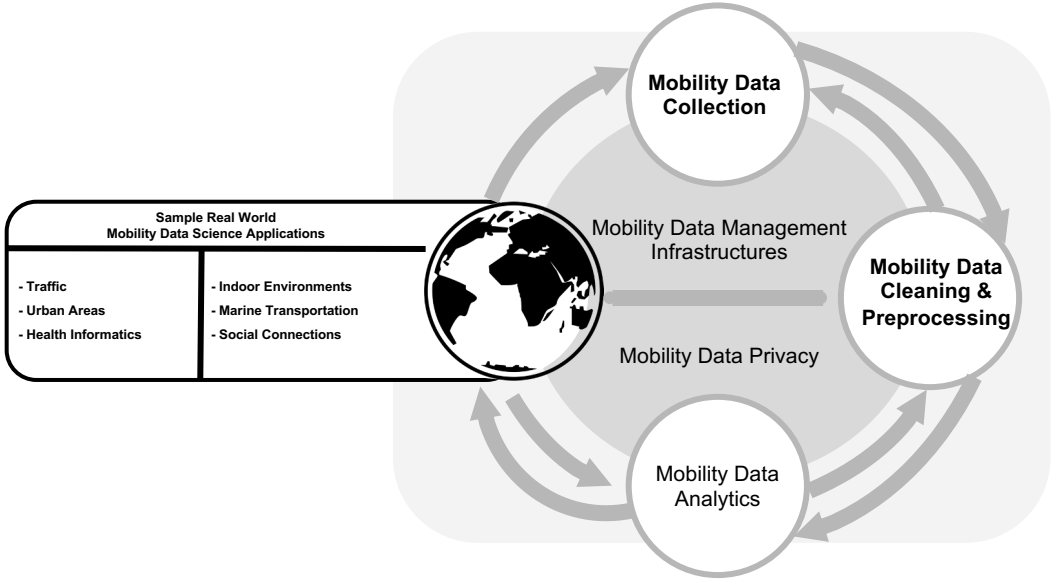
An **interdisciplinary** field that uses scientific **methods**, **processes**, **algorithms** and **systems** to extract or extrapolate **knowledge and insights** from potentially noisy, structured and unstructured **mobility data**, and apply knowledge from mobility data across a broad range of application domains





The Mobility Data Science pipeline

I. Mobility Data Collection, Cleaning & Preprocessing
II. Mobility Data Analytics and Data Management Infrastructures
III. Mobility Data Privacy



Open Mobility Data (1/2)

Trajectory datasets in the **urban domain** are limited to **small** size datasets, e.g.:

- **RioBuses** (Rio de Janeiro, Brazil): 12K trajectories (buses), 118M points, 1 month period, 1 min. sampling rate
- **Grab-Posisi** (Singapore): 84K trajectories (vehicles), 89M points, 1 month period, 1 sec. sampling rate
- **GeoLife** (Beijing, China): 17K trajectories (mixed), 26M points, 3 years period, 1-5 sec. sampling rate
- **RomaTaxi** (Rome, Italy): 320 trajectories (taxis), 21M points, 1 month period, 7 sec. sampling rate

Why do we call them “small”?

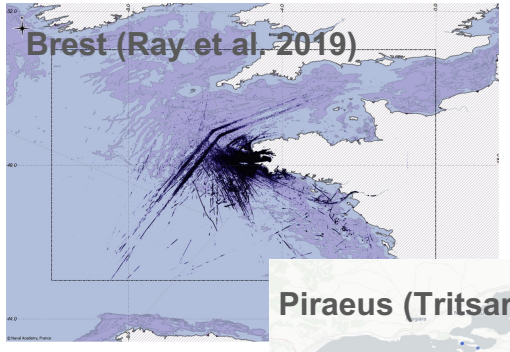
Other urban datasets only include **origin/destination** of each trajectory, e.g.,

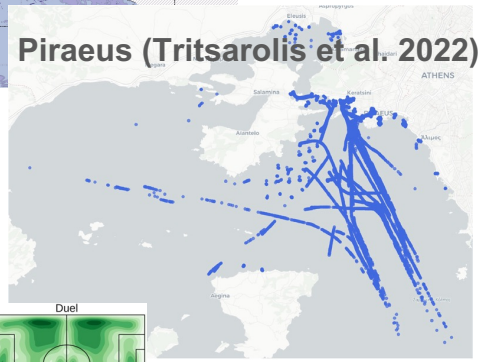
- **NYC Cabs** (New York City, NY, USA): 1.4 billion trips, 9 years period

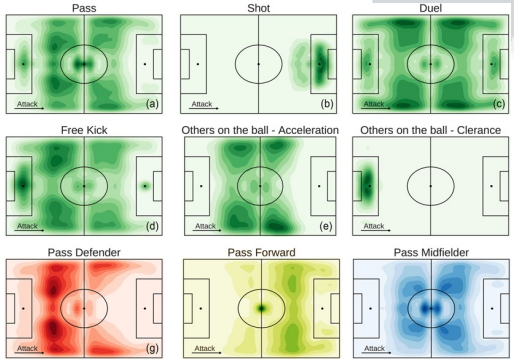
Open Mobility Data (2/2)

Other than road network- constrained datasets:

- **Maritime AIS data** (+metadata), e.g.:
  - Brest, France: 19M points (6 months)
  - Piraeus, Greece: 244M points (32 months)
- **Sport** (basketball, soccer, etc.) data
  - Wyscout soccer dataset: ~3M time-stamped and geo-positioned events during ~2K matches



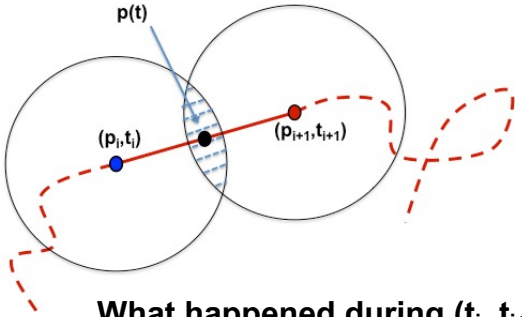




Soccer data (Pappalardo et al. 2019)

The Need for Mobility Data Preprocessing

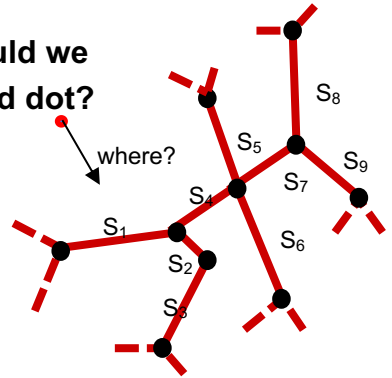
- GPS device inaccuracy; noise in data transmission; uncertainty of moving objects whereabouts between two recorded locations; etc.



What happened during (t_i, t_{i+1})?
Potential Area of Activity (PAA)

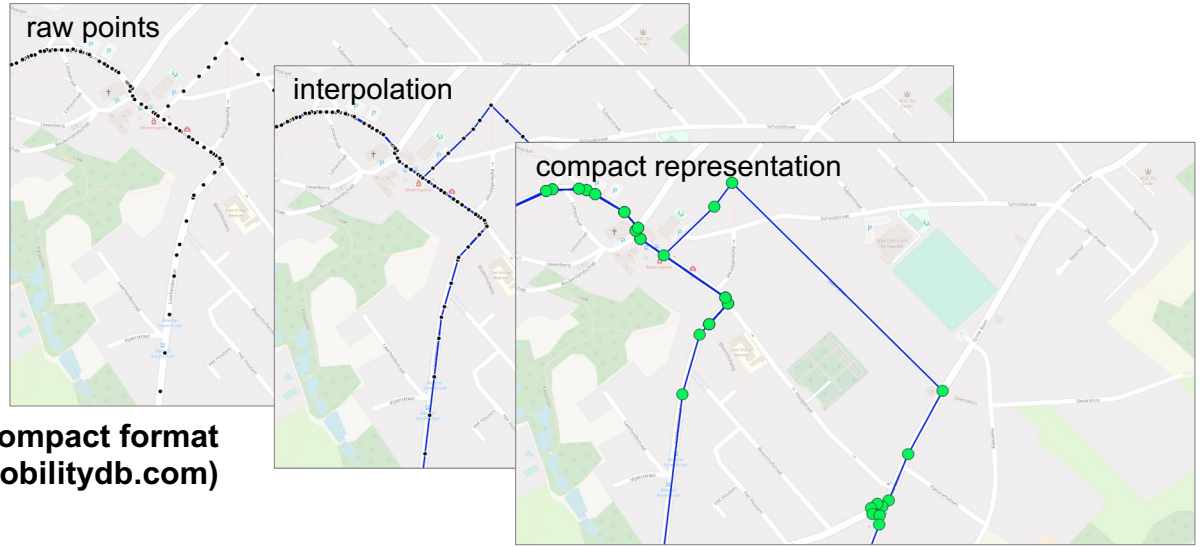
Noise in ADS-B Flight Aware positions during takeoff of an aircraft (examine the sequence of the timestamps)

Where should we place the red dot?



Typical Preprocessing Pipeline

1. Noise elimination
2. (if network-constrained data) Map matching
3. (if trajectory-oriented analysis) Gap filling, Trajectory segmentation
4. Data enrichment



From raw points to a compact format (source: mobilitydb.com)

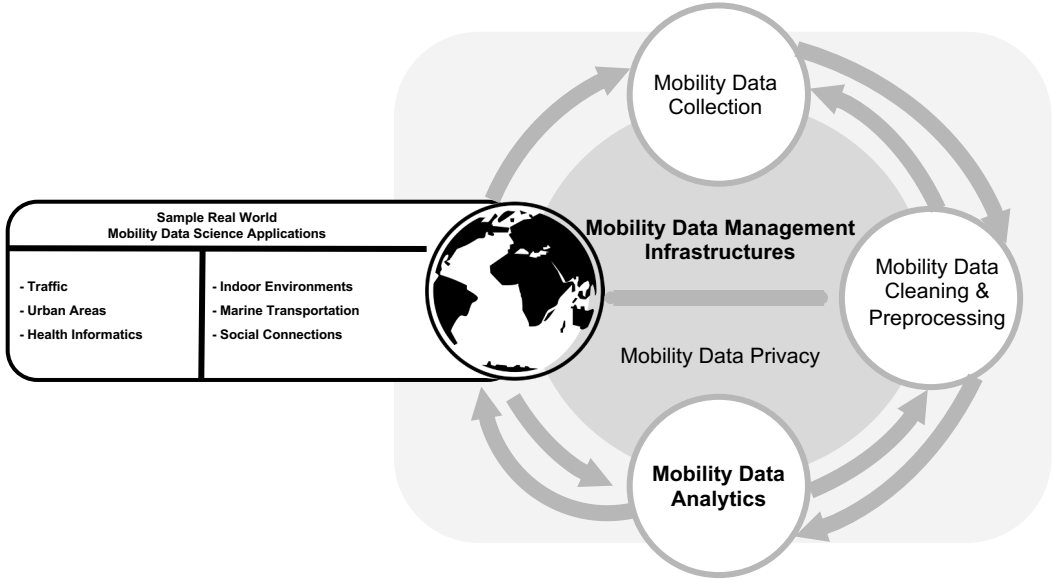
Challenges in Mobility Data Collection & Prep.

1. **Mobility data privacy**
    - *tradeoff between fine granularity detailed mobility data and privacy*
2. **Mobility data bias**
    - *equitable, fair actions and policies based on mobility data science results*
3. **Incentives for data sharing**
    - *incentives to drivers to share their mobility traces, even for sporadic trips*
4. **Simulated mobility data**
    - *work with social scientists to create realistic individual-level human mobility data*
5. **Inaccuracy in the movement space infrastructure**
    - *map inference algorithms that go beyond inferring the map topology to inferring map metadata*
6. **Filling in temporal mobility gaps**
    - *scalable, fine-grained imputations that mimic a continuous trajectory data stream*

I. Mobility Data Collection, Cleaning & Preprocessing

II. Mobility Data Analytics and Data Management Infrastructures

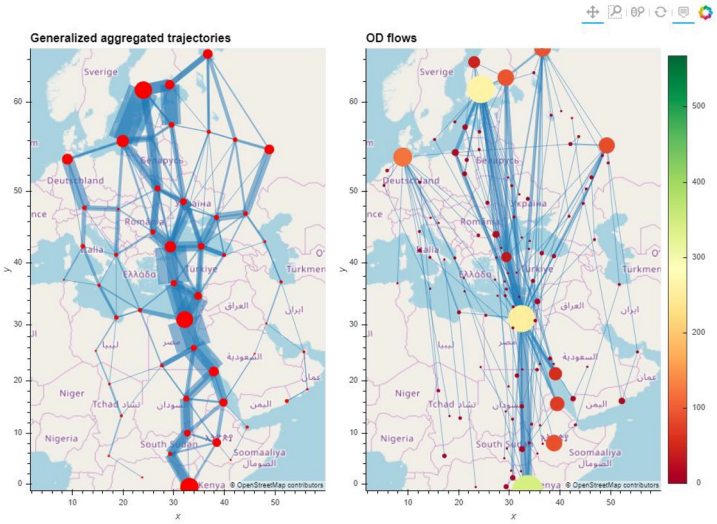
III. Mobility Data Privacy



Goals of Mobility Data Analytics

A wide palette of analytics themes. To name but a few:

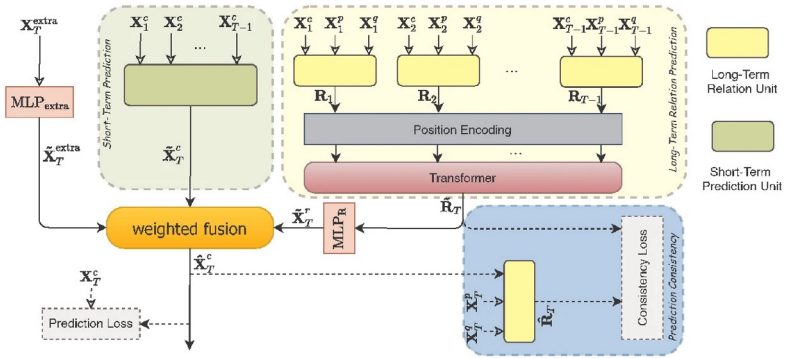
- **Urban mobility (traffic data)**: green routing, traffic anomaly detection, hot spot / path analysis, road traffic prediction, travel time estimation, update of road network
- **Public transportation (ticketing data)**: understanding passenger demand and movement patterns, strategic long-term planning of the network
- **Personal mobility of individuals (GPS and other sensor data)**: activity recognition, personalized routing, matching with ride-sharing services, crowd-sourcing



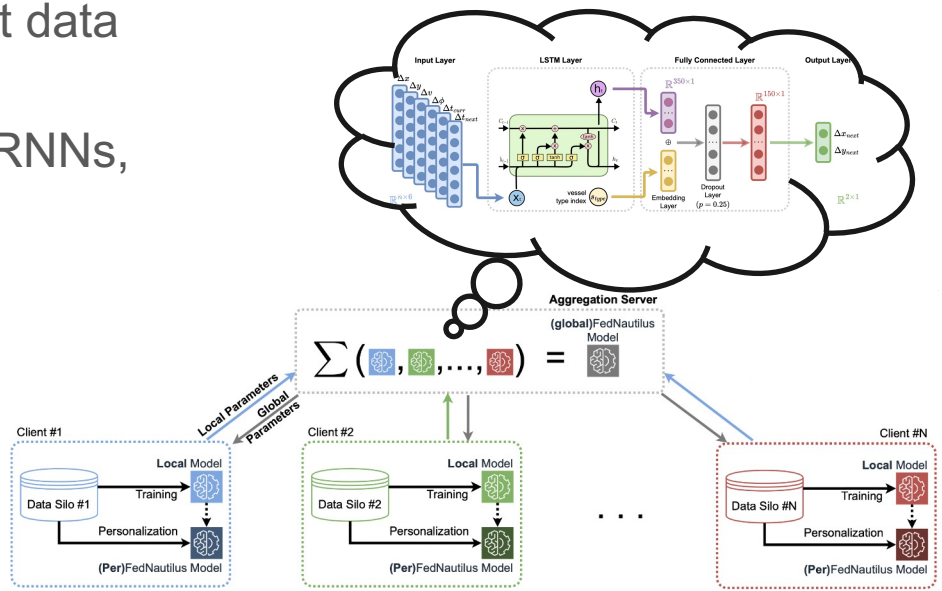
Analyzing aggregated trajectories and OD flows with MovingPandas (source: movingpandas.org)

Mobility Data Analytics Architectures

- from Python libraries for movement data analysis (e.g., MovingPandas)
- … to deep learning architectures (RNNs, GANs, Transformers, etc.)



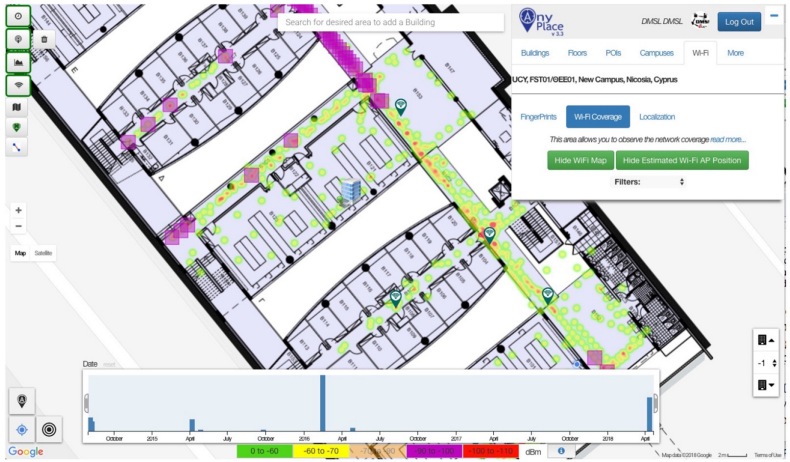
The TermCAST architecture for urban flow forecasting (Xue & Salim, 2021)



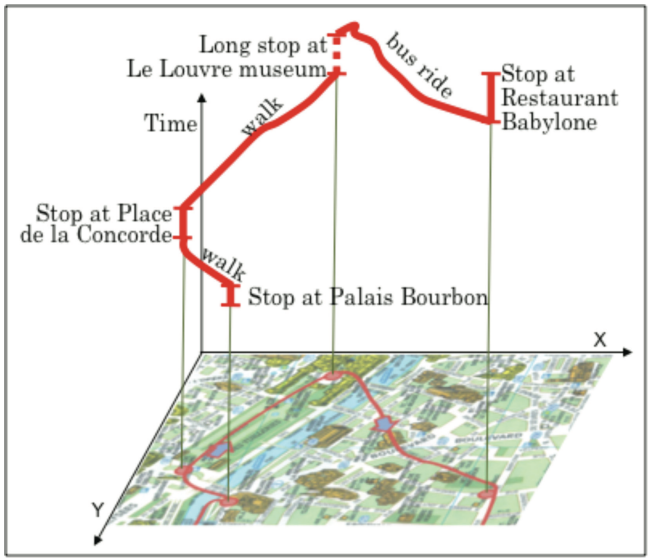
The FedNautilus architecture for (federated) maritime traffic forecasting (Tritsarolis et al. 2024)

Advances in Mobility Data Mgmt Infra. (1/2)

From models and complex data types to capture, e.g. **mobility semantics** or **indoor movement** …



Fingerprint management in indoor environments (Laoudias et al. 2021)

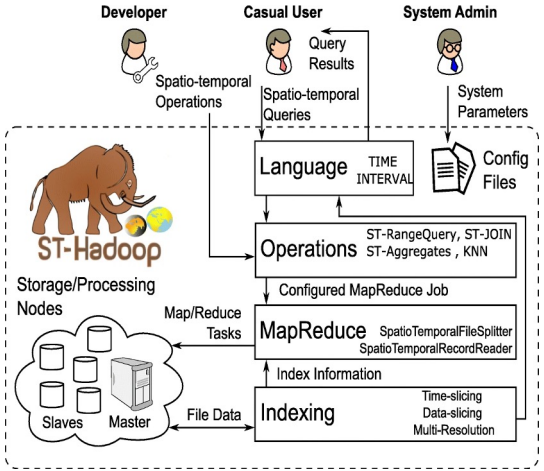


Semantic Trajectories (Parent et al. 2013)

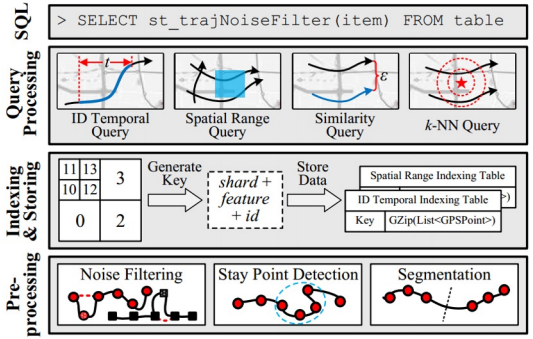
Advances in Mobility Data Mgmt Infra. (2/2)

…to **indexing** and **query processing** techniques, aiming to boost performance

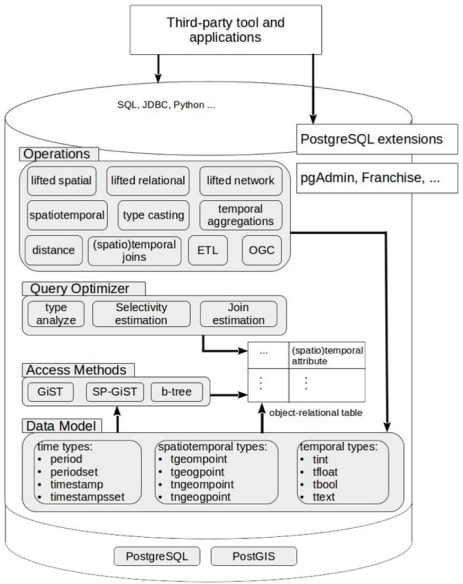
… to **systems** (all-in-one solutions)



ST-Hadoop (Alarabi et al. 2018)



TrajMesa (Li et al. 2020)

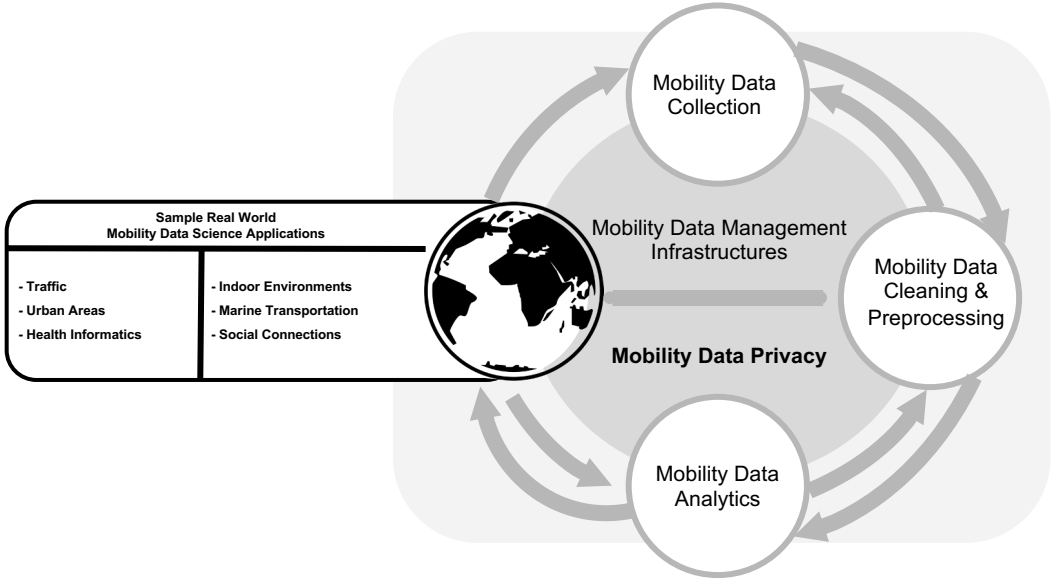


MobilityDB (Zimányi et al. 2020)

Challenges in Mobility Data Analytics & Data Mgmt Infra.

7. **ML for mobility data**
   - *build analysis primitives and common building blocks shaping a framework of ML-based mobility data analysis*
8. **Movement behavior understanding**
   - *modelling and understanding mobility behavior (using e.g., XAI), robust to changes due to societal events*
9. **Visual Analytics**
   - *develop VA techniques facilitating visual discovery of behavioral patterns*
10. **Building mobility-aware systems**
    - *native support for mobility data (from spatially- and temporally-aware DBMSs to scalable big data and NoSQL systems)*
11. **Location data as first-class citizens**
    - *support multi-models in one seamlessly integrated location+X system ("X" being keywords, graphs, relational data, click streams, document data, etc.)*
12. **Hybrid (streaming, batch) workloads**
    - *adopt the concepts behind HTAP (= OLTP + OLAP) systems to support the nature of mobility data*

I. Mobility Data Collection, Cleaning & Preprocessing
II. Mobility Data Analytics and Data Management Infrastructures
III. Mobility Data Privacy



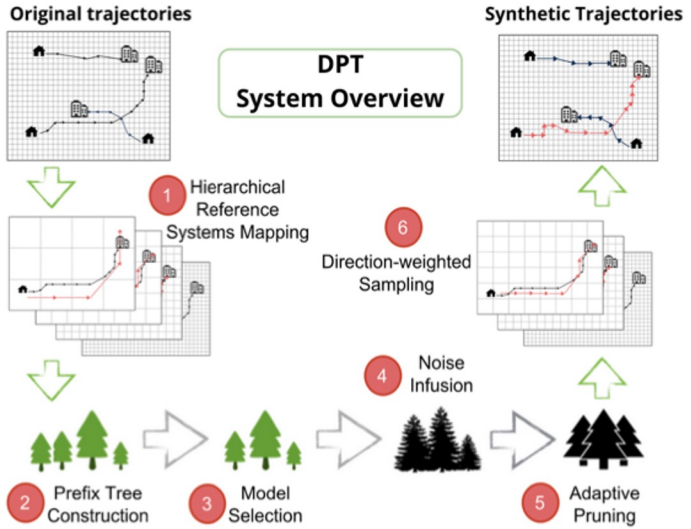
Efforts in Mobility Data Privacy

@ **data collection stage** (local setting):

- Applying Local Differential Privacy (LDP) schemes to location data
- Perturbation methods (e.g., GeoInd)

@ **data analysis stage** (central setting):

- Differential Privacy techniques for training ML models, e.g., Differential Private Trajectories (DPT)
- ML models for the assessment of privacy attacks (e.g., home-work attack) over raw mobility data



DPT overview (He et al. 2015)

Challenges in Mobility Data Privacy

13. **Threat models and privacy definitions**
    - *DP-relaxed versions may be needed given specific threat models to enhance the privacy and utility tradeoff*
    - *Better understanding on what sensitive information may be revealed and reconstructed from mobility data based models*
14. **Privacy and utility tradeoff and other factors**
    - *Designing realistic synthetic data generation methods for optimal privacy utility tradeoff (also, ensuring the fairness)*
15. **Explainability and societal education**
    - *Principles, design guidelines, and tools for explaining DP's protection and limitation to the society / stakeholders*

Geo-Indistinguishability (Andrés et al. 2013)

Conclusions

- **Mobility Data Science** is a distinct branch of (generic) Data Science
  - Space-time dimensions call for different methods of data acquisition, management, analysis, and privacy preservation

- We surveyed recent advances, described motivating applications, and identified major research questions ⇒ a list of **15 challenges** on
  - Mobility data collection & data preprocessing (6)
  - Mobility data analytics & data mgmt infra. (6)
  - Mobility data privacy (3)

1. **Mobility data privacy**
2. **Mobility data bias**
3. **Incentives for data sharing**
4. **Simulated mobility data**
5. **Inaccuracy in the movement space infrastructure**
6. **Filling in temporal mobility gaps**
7. **ML for mobility data**
8. **Movement behavior understanding**
9. **Visual Analytics**
10. **Building mobility-aware systems**
11. **Location data as first-class citizens**
12. **Hybrid (streaming, batch) workloads**
13. **Threat models and privacy definitions**
14. **Privacy and utility tradeoff**
15. **Explainability and societal education**

Mokbel, M., Sakr, M., Xiong, L., Züfle, A., et al.
Mobility Data Science: Perspectives and Challenges.
ACM Trans. Spatial Algorithms Syst. 10, 2, Article 10 (June 2024)

Yannis Theodoridis acknowledges the support of the EU's Horizon Europe research and innovation program under grant agreements No. 101070279 (MobiSpaces) and No. 101093051 (EMERALDS).

MobiSpaces
new data spaces for green mobility

EMERALDS
Extreme-scale Urban Mobility Data Analytics as a Service

Thank you for your attention !!

References (1/3)

Datasets:

- **Rio**: https://ieee-dataport.org/open-access/crawdad-coppe-ufrjriobuses
- **Grab-Posisi**: https://doi.org/10.1145/3356995.3364536
- **GeoLife**: https://www.microsoft.com/en-us/research/publication/geolife-gps-trajectory-dataset-user-guide/
- **RomaTaxi**: https://ieee-dataport.org/open-access/crawdad-romataxi
- **NYC cabs**: https://www.kaggle.com/c/nyc-taxi-trip-duration/data

Papers …

References (2/3)

- Alarabi, L., et al. (2018) **ST-Hadoop: a MapReduce Framework for Spatio-temporal Data**. GeoInformatica, 22, 4.
- Andrés, M.E., et al. (2013) **Geo-indistinguishability: Differential privacy for location-based systems**. Proceedings of SIGSAC.
- He, X., et al. (2015) **DPT: differentially private trajectory synthesis using hierarchical reference systems**. Proceedings of the VLDB Endowment, 8, 11.
- Laoudias, C., et al. (2021) **Indoor Quality-of-position Visual Assessment Using Crowdsourced Fingerprint Maps**. ACM Trans. Spatial Algorithms Syst. 7, 2, Article 10.
- Li, R., et al. (2020) **TrajMesa: A distributed NoSQL storage engine for big trajectory data**. Proceedings of ICDE.
- Pappalardo, L., et al. (2019) **A public data set of spatio-temporal match events in soccer competitions**. Sci Data 6, 236

References (3/3)

- Parent, C., et al. (2013) **Semantic trajectories modeling and analysis**. ACM Comput. Surv. 45, 4, Article 42.
- Ray, C., et al. (2019) **Heterogeneous integrated dataset for Maritime Intelligence, surveillance, and reconnaissance**. Data in Brief, 25, 104141
- Tritsarolis, A., et al. (2022) **The Piraeus AIS dataset for large-scale maritime data analytics**. Data in Brief, 40, 107782
- Tritsarolis, A., et al. (2024) **On Vessel Location Forecasting and the Effect of Federated Learning**. Proceedings of MDM.
- Xue, H., & Salim, F.D. (2021) **TERMCast: Temporal relation modeling for effective urban low forecasting**. Proceedings of PAKDD.
- Zimányi, E., et al. (2020) **MobilityDB: A Mobility Database Based on PostgreSQL and PostGIS**. ACM Trans. Database Syst. 45, 4.