Data Science Lab. @ University of Piraeus

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
UNIVERSITY OF PIRAEUS

Mobility Data Analytics

Yannis Theodoridis
Data Science Lab.*, Univ. Piraeus

* **Credits**: Eva Chondrodima, Christos Doulkeridis, Harris Georgiou,
Yannis Kontoulis, Nikos Pelekis, Panagiotis Tampakis, George S. Theodoropoulos, Andreas Tritsarolis

MSc GeoInformatics @NTUA, May 2024

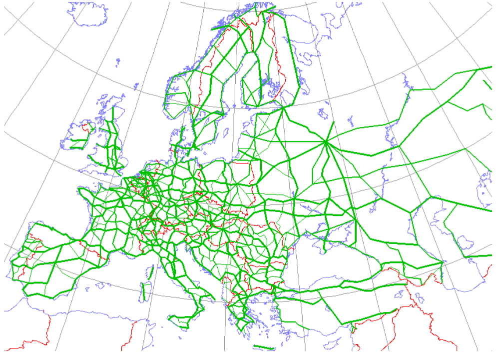
Outline

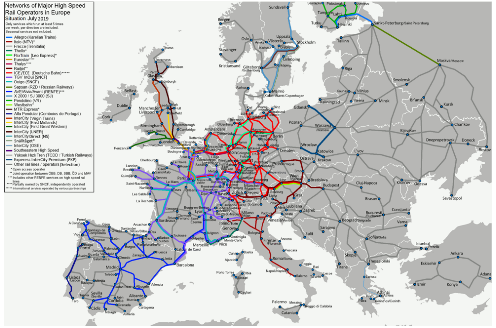
1. **Introduction - Getting familiar with mobility data**
2. **Pre-processing mobility data**
   - Cleansing, Simplification, Enrichment, Sampling, etc.
3. **Analyzing mobility data**
   - Cluster analysis (and collective movement behavior)
   - Future location & trajectory prediction
4. **Summary**

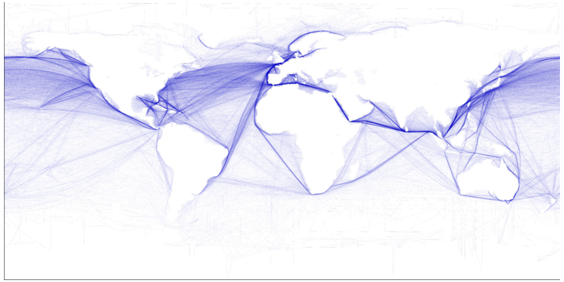
1.
Introduction –
Getting to know mobility data

Application domains

- **Urban**: movement of vehicles (private, taxis, buses), pedestrians, etc.
- **Maritime / Aviation**: movement of ships/aircrafts (also, challenges due to unmanned/autonomous objects)
- Examples:
  - Detect **typical vs. anomalous** movements, hot spots/paths, etc.
  - **Forecast** anticipated routes (or traffic), etc.







All images source: Wikipedia.org

Examples of datasets @ urban

- **GeoLife** (source: Microsoft Research Asia)
  - 182 user movements (under various transportation means) organized in 17,621 trajectories;
  - 68 Km in 2,7 hrs. per trajectory, avg.;
  - dense sampling (1 sample every ~5 sec)
- **T-Drive** (source: Microsoft Research Asia):
  - 2,357 taxis in Beijing for 1 week (15 million points, in total);
  - 869 Km per taxi, avg.;
  - sparse sampling (1 sample every ~3 min)

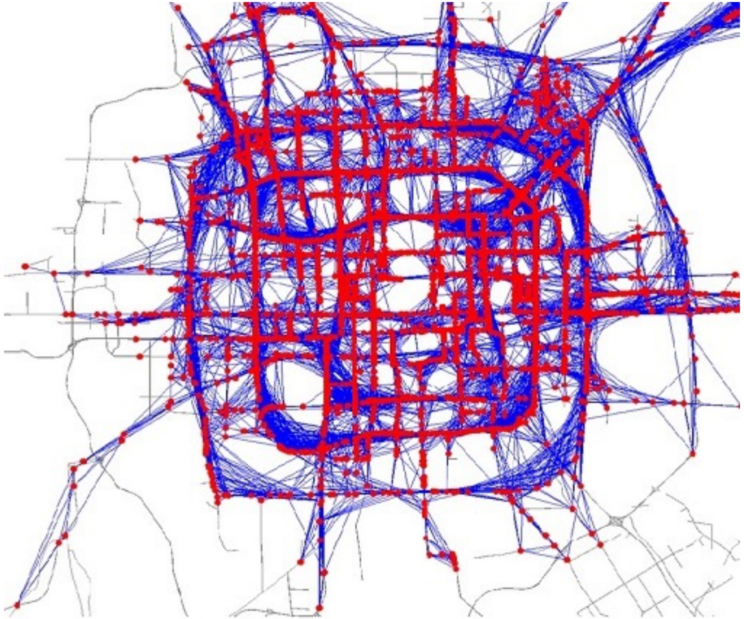
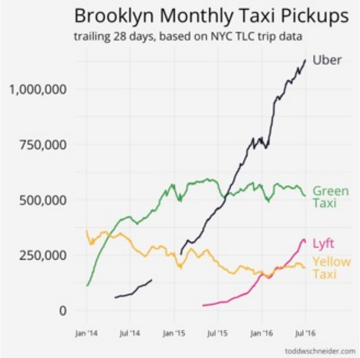


image source: research.microsoft.com

Examples of datasets @ urban (cont.)

- **NYC taxis** (source: NYC Taxi & Limousine Commission): 1.4 billion trips, Jan. 09 – Dec.17.
  - **Ride-hailing apps** data are also provided
  - Attention: pickup – drop-off locations are only available



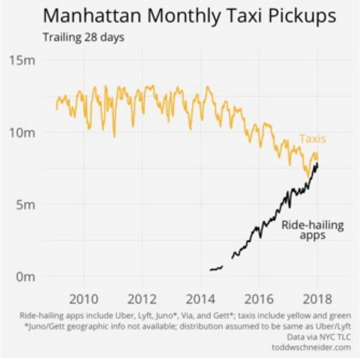
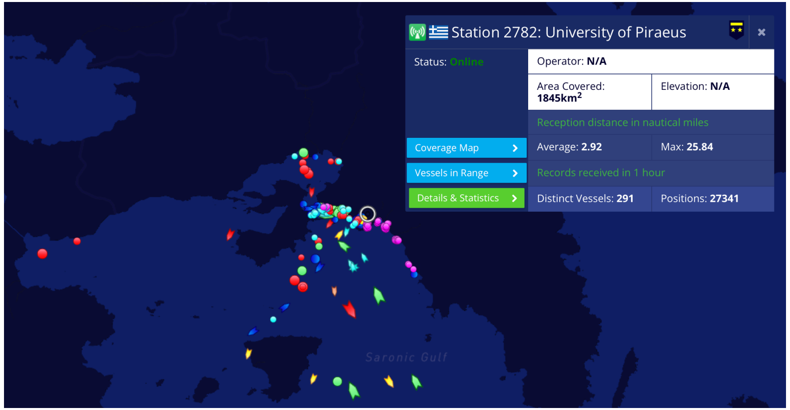


image source: toddwschneider.com

Examples of datasets @ maritime

- **AIS** (Automatic Identification System)
  - >250,000 vessels tracked daily (source: marinetraffic.com)
  - AIS signal transmitted: every 2 to 10 sec depending on speed while underway; every 3 min while at anchor



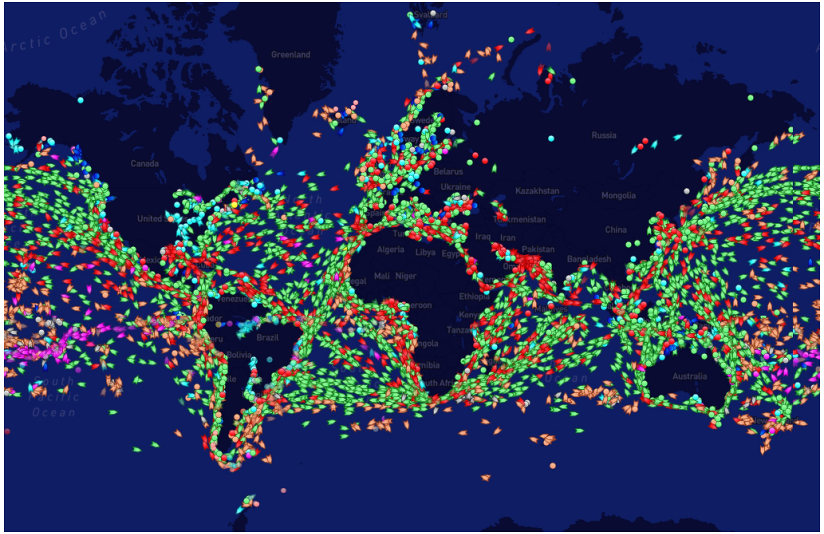


image source: marinetraffic.com

- top: global snapshot on May 26th, 2022; vessel colors correspond to different vessel types (e.g., cargo is green, tanker is red)
- left: vessels tracked by the Univ. Piraeus' AIS station

Examples of datasets @ aviation

- **ADS-B** (Automatic Detection System - Broadcast)
  - >15,000 aircrafts flying at the same time worldwide (source: flightradar24.com)
  - ADS-B signal transmitted: every 1 sec while on air; not transmitted while on the ground

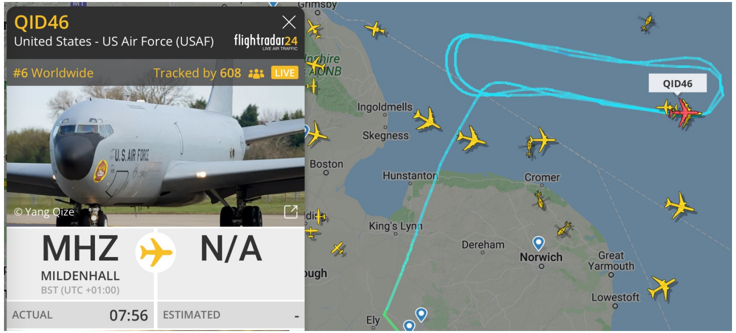


image source: flightradar24.com

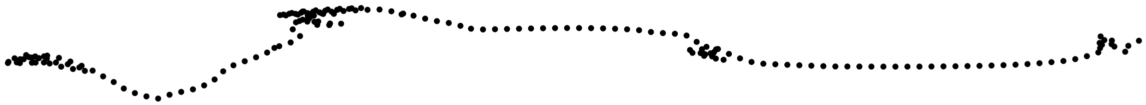
- top: global snapshot on May 25th, 2022; yellow vs. blue planes if located by terrestrial vs. satellite stations
- left: the route of a military aircraft

2.
Pre-processing mobility data

Data pre-processing

- Definition: **preparing data for analytics purposes**

$$T = \{ \langle p_1, t_1 \rangle, \langle p_2, t_2 \rangle, \ldots, \langle p_n, t_n \rangle \}$$

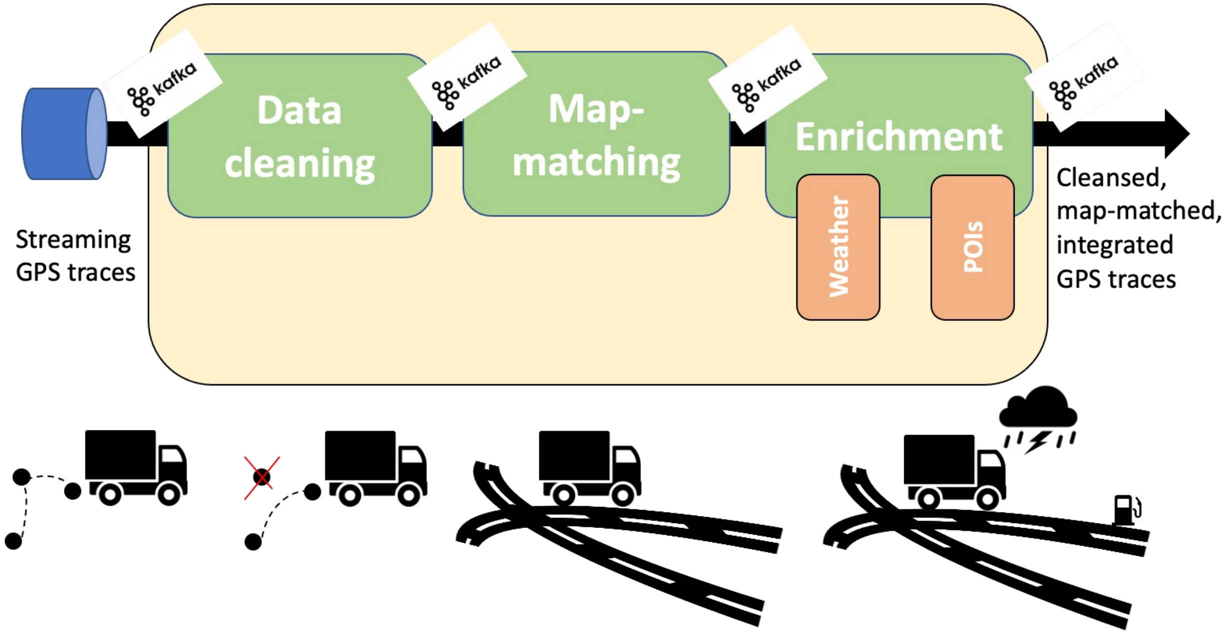


- Data pre-processing includes:
  - **Cleansing** (noise removal, smoothing, map matching, etc.)
  - **Transformation** (trajectory segmentation, simplification, etc.)
  - **Enrichment** (semantic annotation, data fusion, etc.)

  etc.

Data pre-processing (cont.)

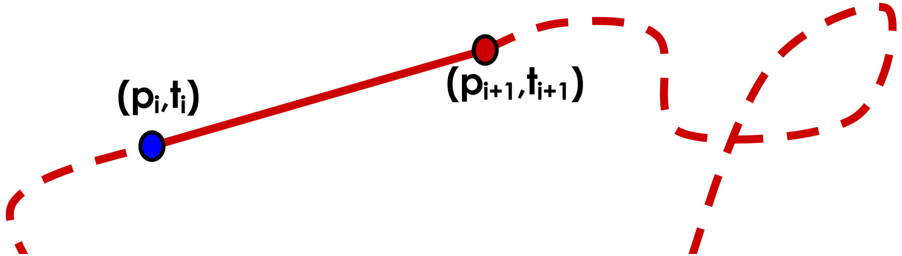
- An example: **data pre-processing pipeline (urban traffic)**



Source: Track & Know EU project

From GPS locations to trajectories

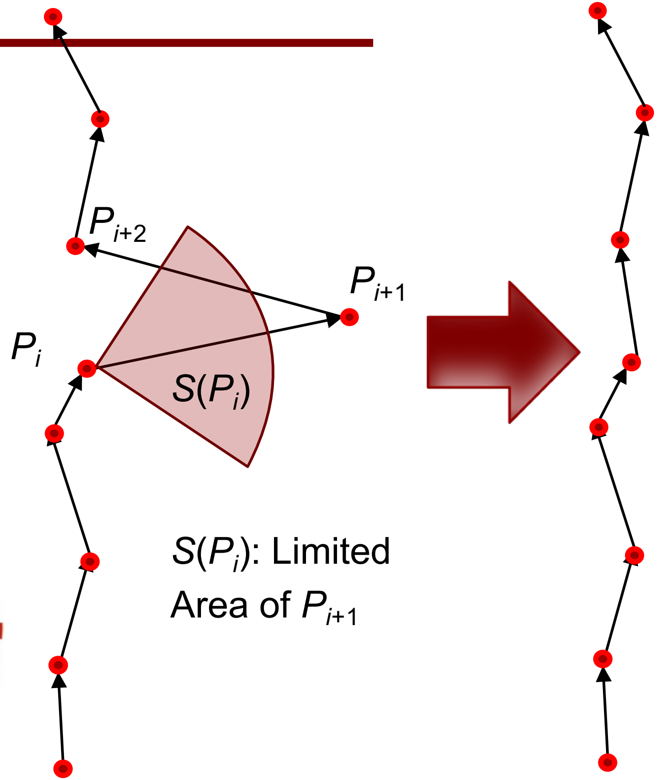
- GPS records correspond to **samples** (p_i, t_i) of our movement – inferring ‘continuous’ movement is not trivial.
- A typical representation of a moving object’s trajectory is a **polyline** (in 4D space; x-, y-, z-, t-) – vertices correspond to (p_i, t_i)
- Typically, **linear interpolation** is assumed between (p_i, t_i) and (p_{i+1}, t_{i+1})



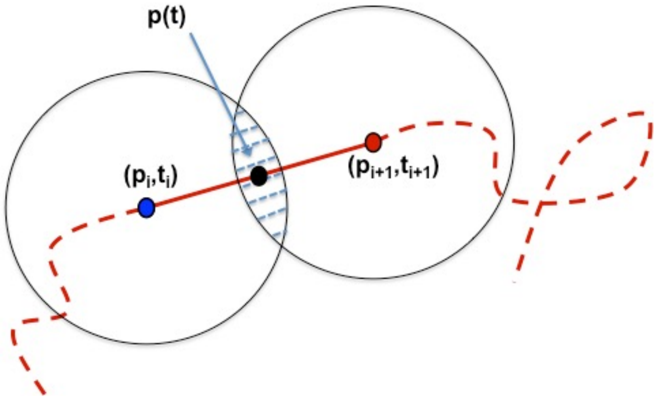
$$p(t) = \left(x_i + \frac{t - t_i}{t_{i+1} - t_i}(x_{i+1} - x_i), y_i + \frac{t - t_i}{t_{i+1} - t_i}(y_{i+1} - y_i) \right)$$

GPS Data Cleansing

- Erroneous recordings: noise vs. random errors
- **Noise** corresponds to values that are ‘impossible’ to appear
- Can be detected and removed using appropriate filters
  - e.g., maximum speed

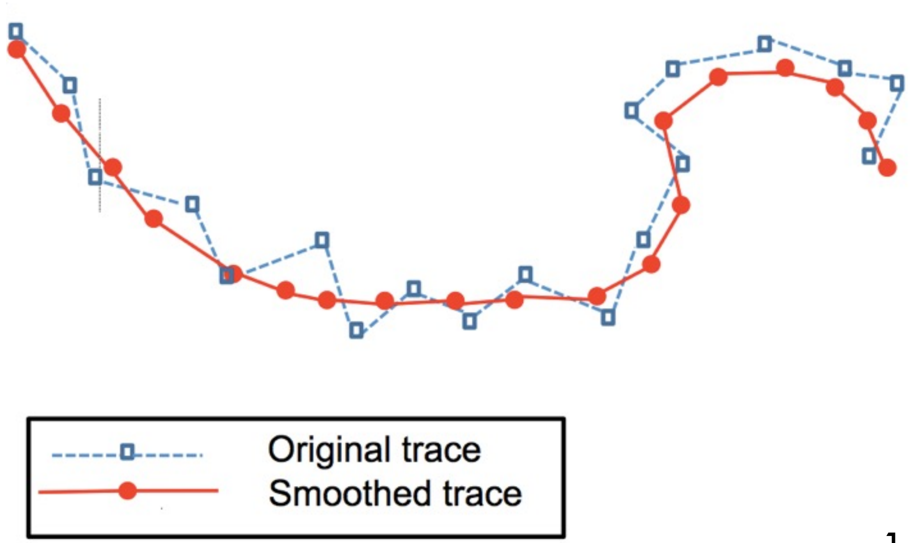


- **Potential Area of Activity** (PAA)



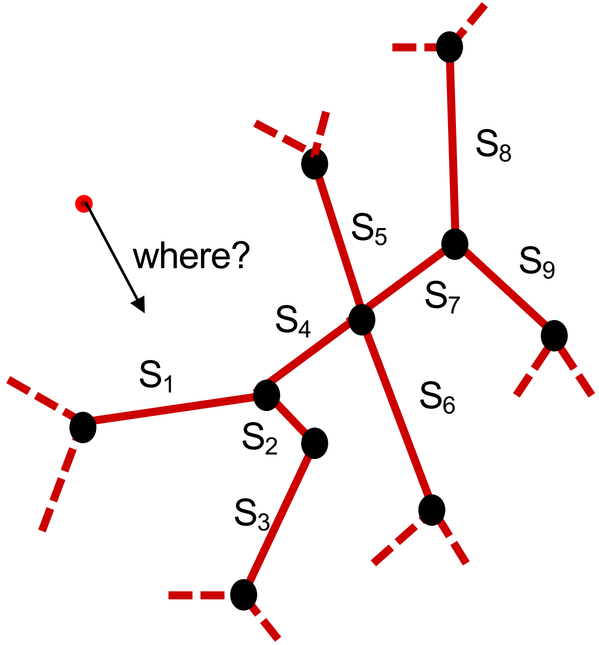
GPS Data Cleansing (cont.)

- Erroneous recordings: noise vs. random errors
- **Random errors** correspond to 'possible' values that appear to be small deviations from actual ones
- Can be smoothed using a plethora of statistical methods
  - e.g., least squares spline approximation (de Boor, 1978)



GPS Data Cleansing (cont.)

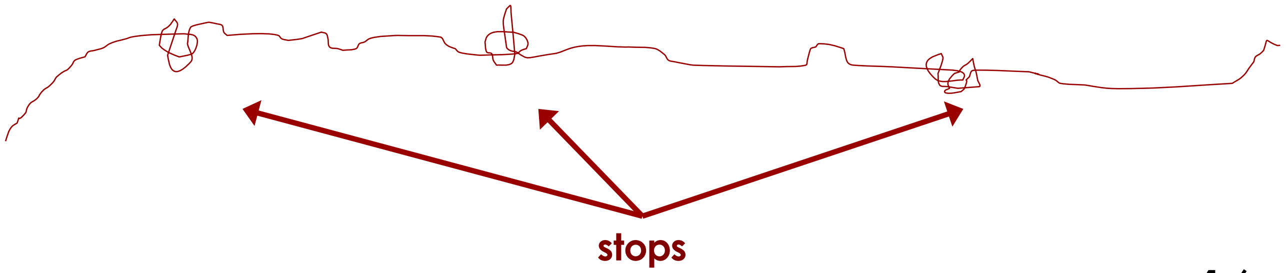
- Special case: network-constrained movement
- Requires an additional step: **map-matching**
- Several techniques (Quddus et al. 2003; 2007):
  - Geometric map-matching
  - Topological map-matching
  - Probabilistic map-matching
  - Hybrid map-matching



Trajectory segmentation

- Goal: **Segment sequences of points** in homogeneous sub-sequences (hereafter, called **trajectories** or **routes**)

- Various approaches:
  - Segmentation via raw (spatial / temporal) gap or via stop discovery
  - Segmentation via prior knowledge (e.g., office / sleeping hours, arrival at ports)



Trajectory simplification

- The need for simplification: efficiency in storage, processing time, etc.
  - Simplification is a form of data compression
- Goal: maintain the original 'signature' as much as possible by only keeping the set of **critical points**
- Approaches
  - Offline (i.e., multi-pass), vs.
  - Online (i.e., single-pass)

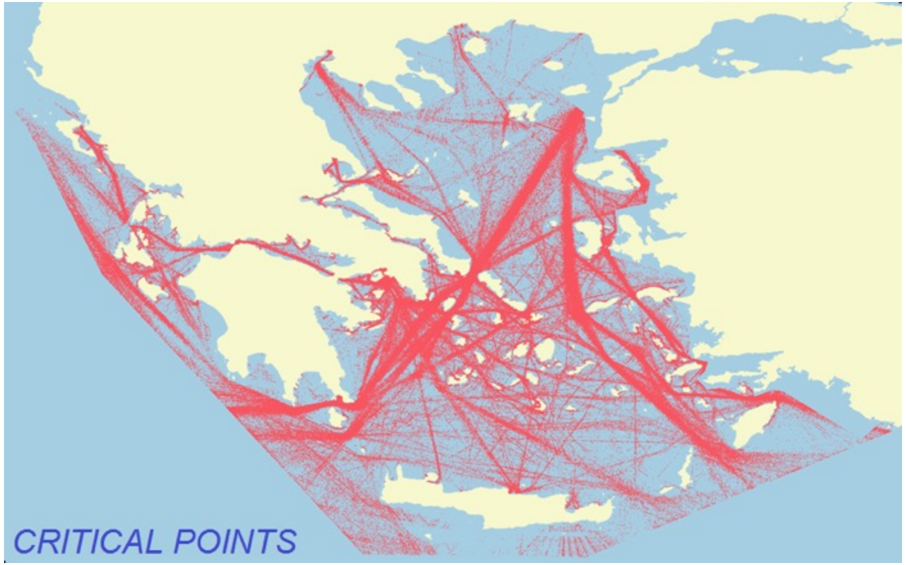


image source: aminess.eu

Trajectory simplification (cont.)

- Offline approaches:
  - top-down vs. bottom-up vs. sliding window vs. opening window
- e.g., **Synchronous Euclidean Distance – SED** (Meratnia & de By, 2004)
  - Adapts the popular Douglas & Peucker polyline simplification (1973) to the mobility domain

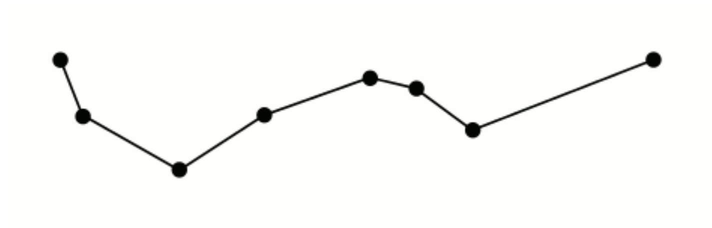
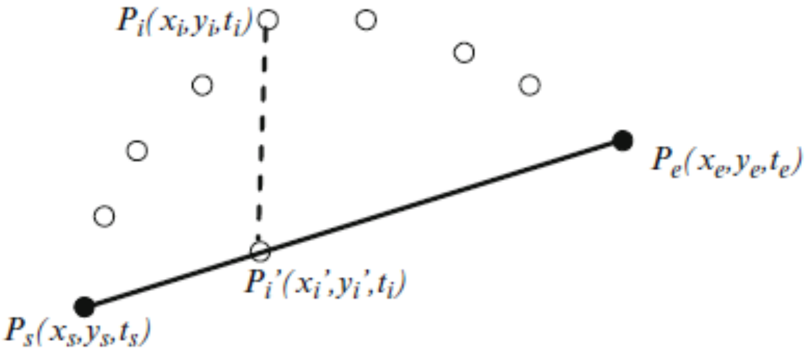
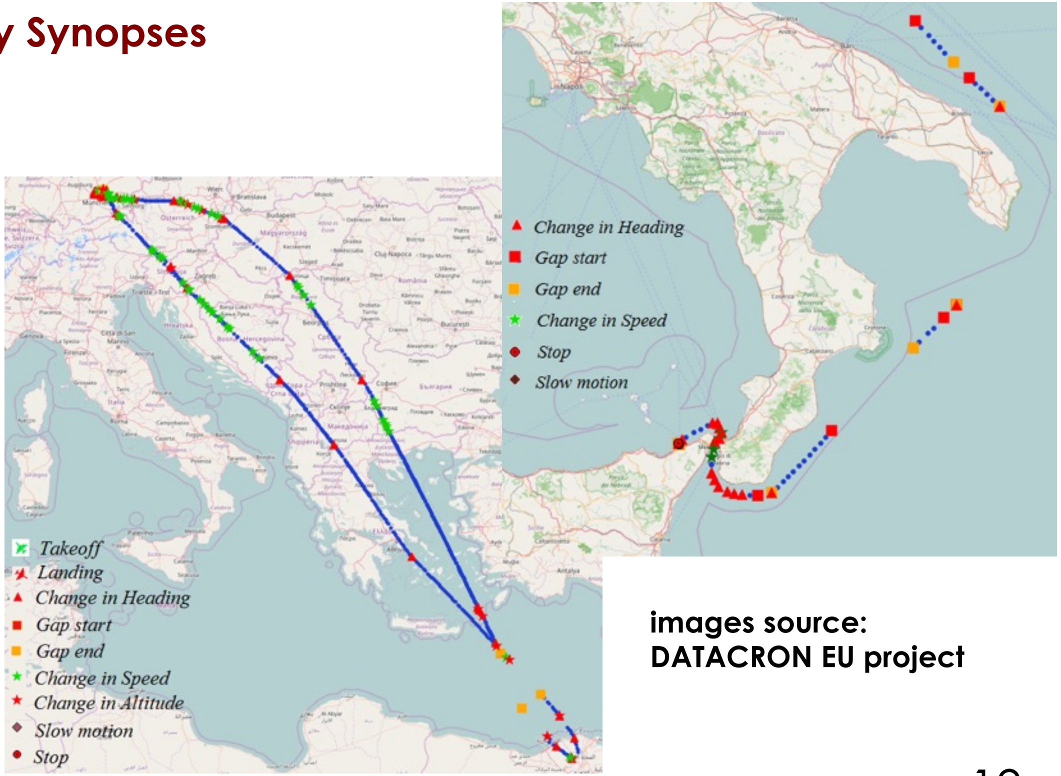


image source: https://commons.wikimedia.org/wiki/File:Douglas-Peucker_animated.gif



Trajectory simplification (cont.)

- Online approaches, e.g., **Trajectory Synopses** (Patroumpas et al. 2015; 2017)
- Maintains a **velocity vector** per moving object in order to detect **instantaneous events**
  - stop; change in velocity vector; etc.
- Tradeoff: degree of compression vs. quality of approximation



images source: DATACRON EU project

Trajectory enrichment

- From "raw" sequences (p,t) of time-stamped locations
- … to meaningful mobility tuples <where, when, what>
- **Semantic trajectory** (Parent et al. 2015)
  - semantically-annotated representation of the motion path of a moving object
  - **sequence of episodes** (**stop/move** segments of routes) along with appropriate **tags**

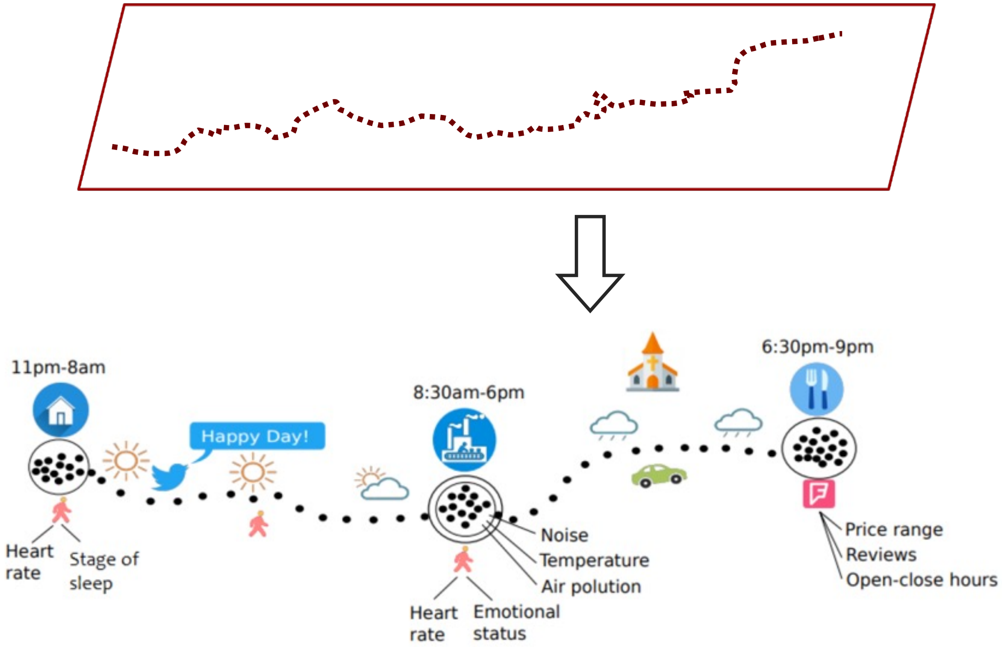


Image source: MASTER EU project

3.
Analyzing mobility data

Types of mobility data analytics

- Discovering **groups** and **outliers**
- Discovering **frequent routes** (hot paths) and **frequent locations** (hot spots)
- **Prediction/forecasting** tasks

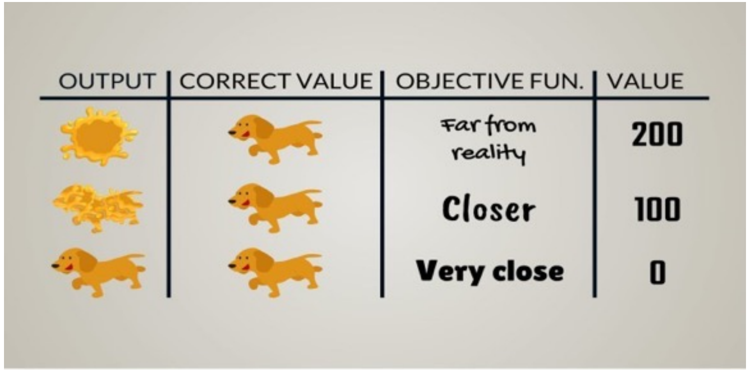
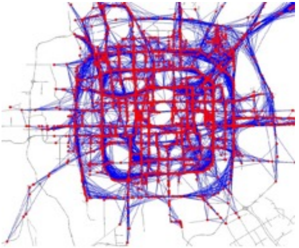
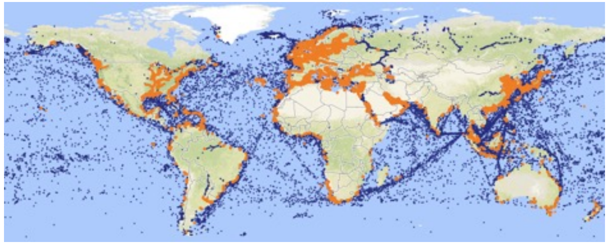


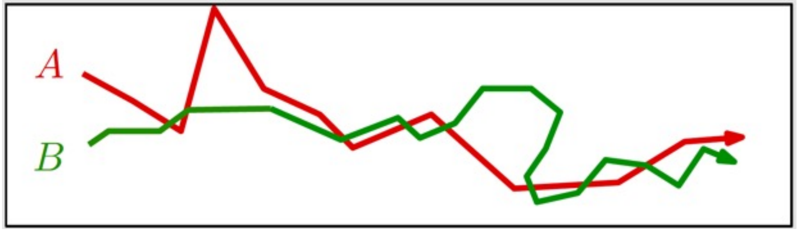
image source: kdnuggets.com





Orthogonal issue: Trajectory similarity

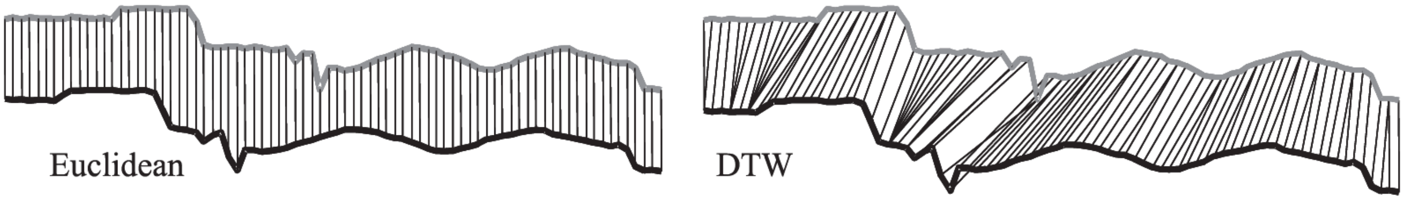
- How do we measure **similarity** between two trajectories A, B?
  - not so trivial as it sounds



- Alternative approaches:
  - Trajectory as a 2D time-series
  - Trajectory as a 2D polyline
  - Trajectory as a movement function

Trajectory as a time series

- Time series similarity has been studied extensively (e.g., Vlachos et al. 2002; Chen et al. 2005). Examples:
  - Euclidean distance, Chebyshev distance, Dynamic Time Warping (DTW),
  - Longest Common SubSequence (LCSS),
  - Edit Distance on Real sequences (EDR),
  - Edit distance with Real Penalty (ERP), etc.



Trajectory as a polyline

- **DISSIM** (Nanni & Pedreschi, 2006; Frentzos et al. 2007)
  - Extension of Euclidean distance:

$$DISSIM(R,S) = \int_{t_1}^{t_n} L_2\big(R(t), S(t)\big)dt$$

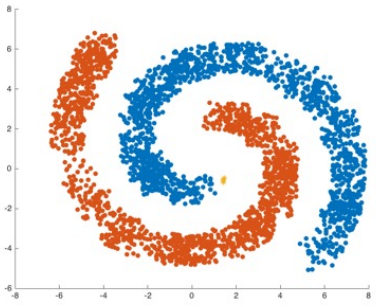
$$DISSIM(R,S) \approx \frac{1}{2}\sum_{k=1}^{n-1}\Big(\Big(L_2\big(R(t_k), S(t_k)\big) + L_2\big(R(t_{k+1}), S(t_{k+1})\big)\Big) \cdot (t_{k+1} - t_k)\Big)$$

  - DISSIM function is a metric
    - Conditions: (1) non-negativity; (2) identity of indiscernibles; (3) symmetry; (4) triangle inequality

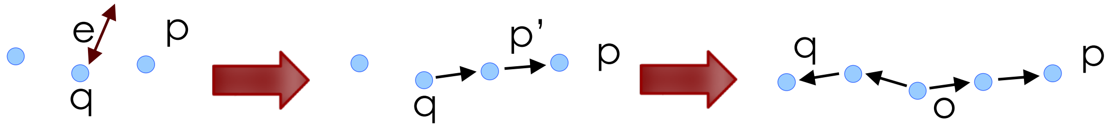
1. $d(x,y) \geq 0$
2. $d(x,y) = 0 \Leftrightarrow x = y$
3. $d(x,y) = d(y,x)$
4. $d(x,z) \leq d(x,y) + d(y,z)$

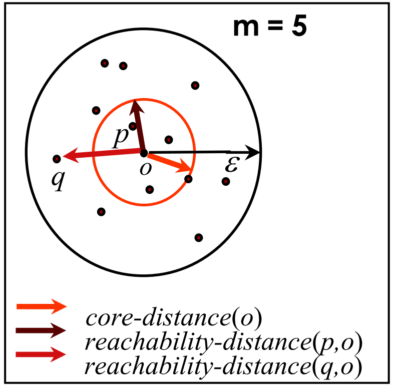
From point clustering …

- **DBSCAN** (Ester et al. 1996), **OPTICS** (Ankerst et al. 1996), etc.: A family of density-based point clustering methods
  - Key parameters (recall that we talk about density-based methods):
    - **radius** of an object's neighborhood (e)
    - minimum **population** within an object's neighborhood (m)
  - Classification of points: **core points** vs. **borders** vs. **noise**
  - Clusters are built around core points wrt. **density reachability**



m = 3





… to Trajectory clustering

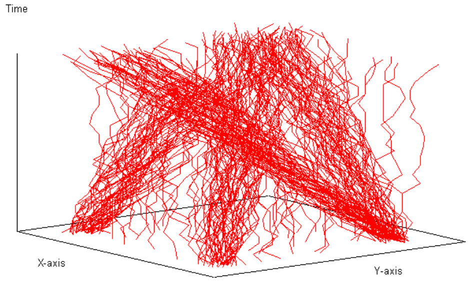
- Objectives:
  - Cluster trajectories w.r.t. similarity
  - Eventually, detect outliers
- Issues:
  - Which similarity function?
  - Upon the entire trajectories or portions (sub-trajectories?

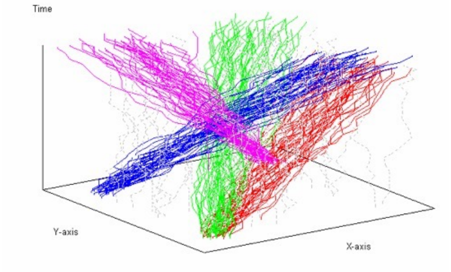
- State-of-the-art:
  - Clustering on the entire trajectories: **T-OPTICS** (Nanni & Pedreschi, 2006)
  - Clustering on sub-trajectories: **TraClus** (Lee et al. 2007); **S²T-Clustering** (Pelekis et al. 2017a, 2017b), **DSC** (Tampakis et al. 2019)

… to Trajectory clustering (cont.)

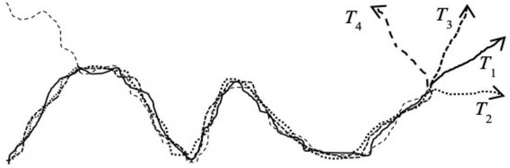
- Clustering at entire trajectory level, e.g. **T-OPTICS**
  - Builds upon OPTICS and DISSIM distance function

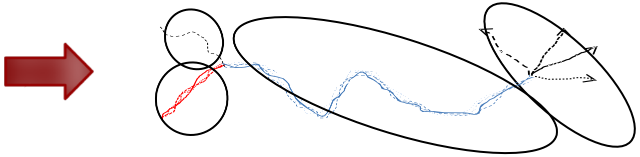
$$DISSIM(R,S) = \int_{t_1}^{t_n} L_2(R(t),S(t))dt$$





- Clustering at sub-trajectory level, e.g. **S^{2}T-Clustering**
  - Finds the most 'popular' sub-trajectories and builds clusters around them

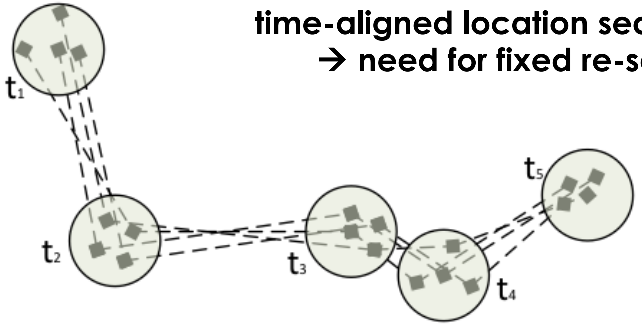


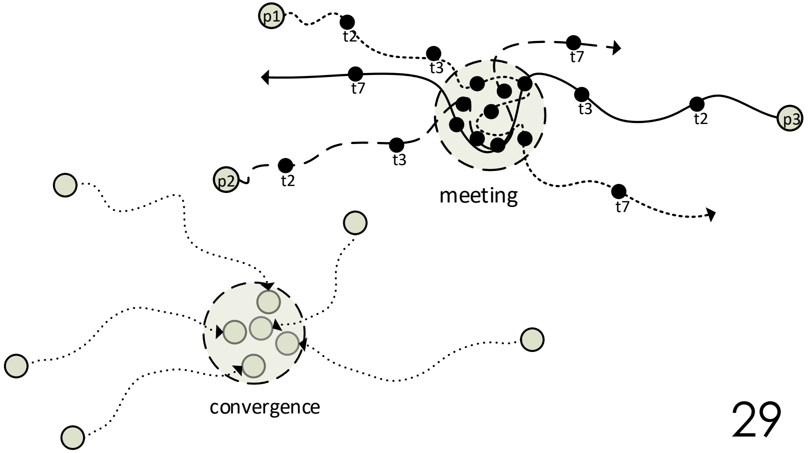


Location-based clustering

- Detecting a large enough subset of objects moving along paths close to each other for a certain time
  - Spherical-like clustering: **Flocks** (Laube et al. 2005; Gudmundsson & van Kreveld, 2006) vs.
  - Density-based clustering: **Convoys** (Jeung et al. 2008); **Swarms** (Li et al. 2010), etc.
- Interesting variants of the flock/convoy methods:
  - **meeting/convergence points**, **leaders and followers**, **evolving clusters** (Tritsarolis et al. 2021), etc.

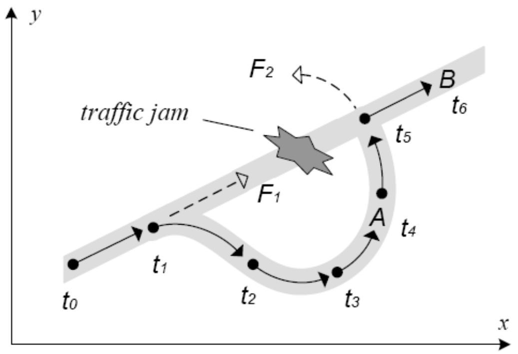
Note: these methods work on time-aligned location sequences → need for fixed re-sampling

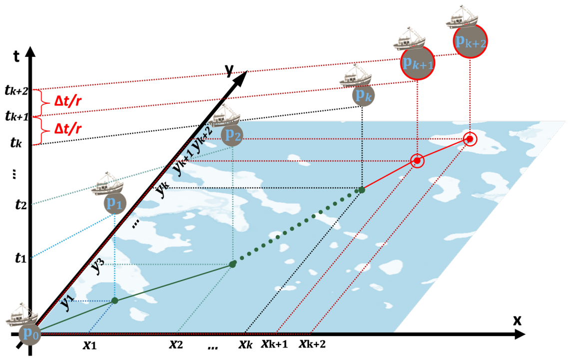




Location / Trajectory prediction

- **Future location / trajectory prediction (FLP/TP)** aims to predict the future location(s) of a moving object within a time horizon.
- Main approach: mathematical formulae- (Tao et al. 2004) vs. **Pattern-based**, i.e., patterns are built upon the objects' history
  - urban (Trasarti et al. 2017);
  - maritime (Chondrodima et al. 2022, 2023; Tritsarolis et al. 2024);
  - aviation (Georgiou et al. 2018, 2020)
- Interesting variants: traffic flow forecasting, collision risk assessment, estimated time of arrival (ETA) prediction, etc.





Location / Trajectory prediction (cont.)

- **MyWay** (Trasarti et al. 2017) maintains a Personal Mobility Data Store (PMDS) per participating person
  - How is a person moving?
    - According to his/her past movement patterns
  - What if the personal datastore is not adequate?
    - Look into the collective knowledge base
- 3 predictors: personal (red), collective (blue), hybrid (green)

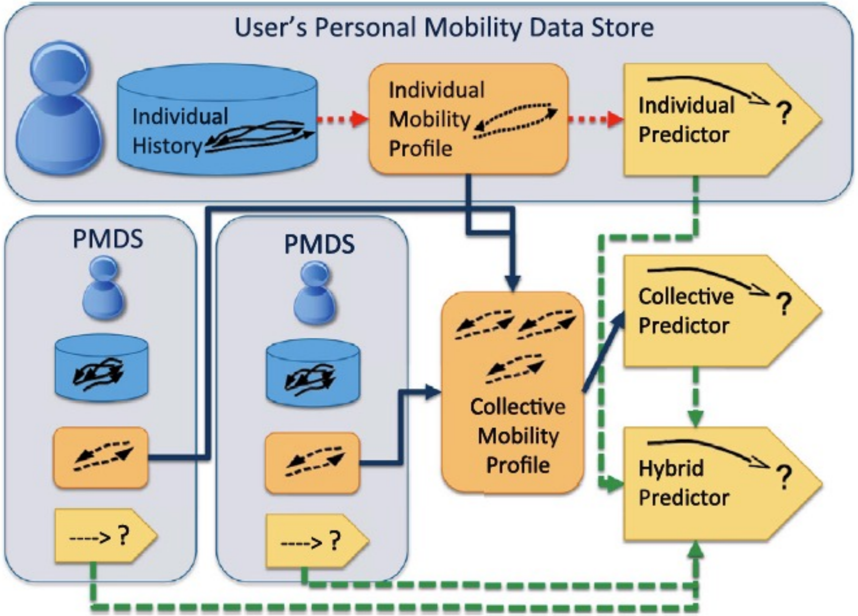
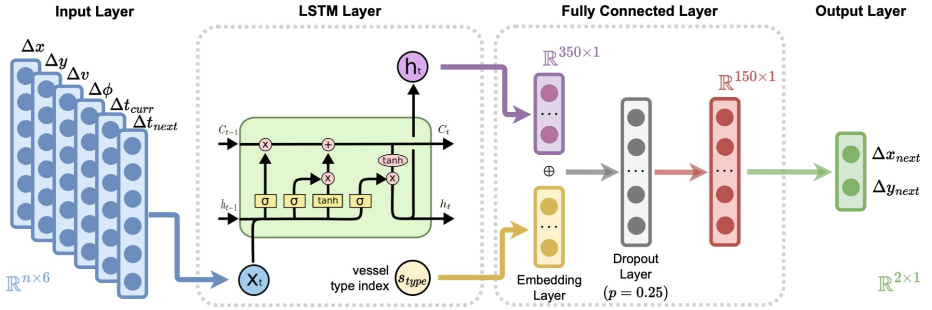


image source: kdd.isti.cnr.it

Location / Trajectory prediction (cont.)

- **(Fed)Nautilus** (Tritsarolis et al. 2024) trains an LSTM neural network with past trajectories of vessels
  - Two variants: centralized (Nautilus) vs. Federated learning- based (FedNautilus) architecture
  - The FL approach achieves ~90% savings in communication cost
    - only model parameters are exchanged between data silos and aggregation server



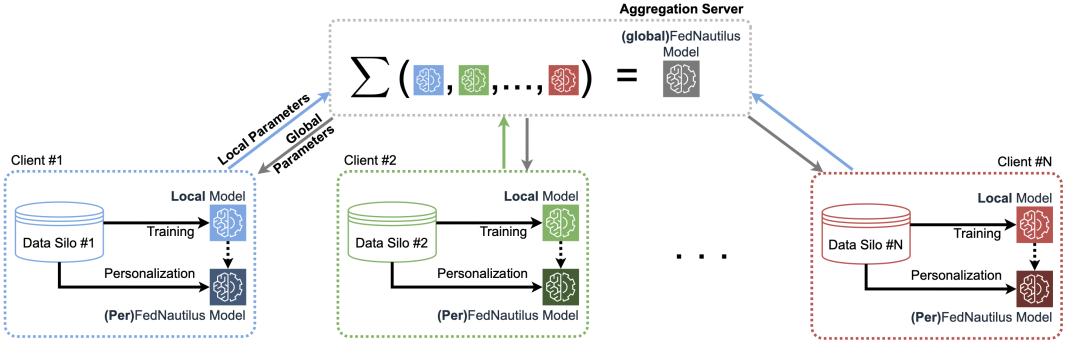
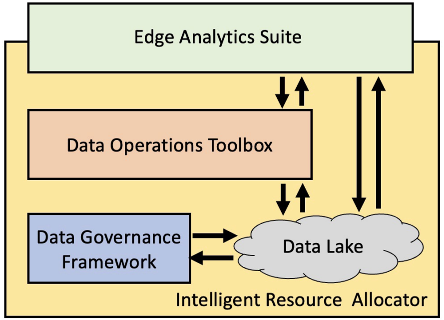


image source: datastories.org/maritime/

5.
Summary

Summary

- The **Mobility Data Analytics** field (Pelekis & Theodoridis 2014) includes many success stories on:
  - **Data management** - access methods & query processing techniques, DBMS extensions (the so-called, Moving Object Databases), etc.
  - **Data mining** – clusters, flocks, convoys, hot spots, etc.
- Current research trends revolve around:
  - **Semantically-enriched trajectory management and analytics** (Parent et al. 2013): information about when / where / what
  - **Extreme-scale mobility data processing** (Vouros et al. 2018): voluminous, streaming, disperse information about objects' movement
  - **Mobility data spaces** (Doulkeridis et al. 2023): exchanging data and models among actors (producers/consumers) – the MobiSpaces.eu project



The MobiSpaces Ref. Architecture

Bibliographical references (1/4)

- Alvares LO, et al. (2007) A model for enriching trajectories with semantic geographical information. In Proceedings of GIS.
- Ankerst M, et al. (1999) OPTICS: Ordering points to identify the clustering structure. In Proceedings of SIGMOD.
- de Boor C (1978) A practical guide to splines. Springer-Verlag.
- Buchin K, et al. (2009) Finding long and similar parts of trajectories. In Proceedings of SIGSPATIAL-GIS.
- Cao H, et al. (2007) Discovery of periodic patterns in spatiotemporal sequences. IEEE Transactions on Knowledge and Data Engineering, 19(4).
- Chen L, et al. (2005) Robust and fast similarity search for moving object trajectories. In Proceedings of SIGMOD.
- Chondrodima E., et al. (2022) Machine Learning Models for Vessel Route Forecasting: An Experimental Comparison. In Proceedings of MDM.
- Chondrodima E., et al. (2023) An Efficient LSTM Neural Network-Based Framework for Vessel Location Forecasting. IEEE Transactions on Intelligent Transportation Systems, 24(5).
- Claramunt C, et al. (2017) Maritime data integration and analysis: recent progress and research challenges. In Proceedings of EDBT.
- Douglas D, Peucker T (1973) Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. The Canadian Cartographer, 10(2).
- Doulkeridis C, et al. (2023) MobiSpaces: An Architecture for Energy-Efficient Data Spaces for Mobility Data. In Proceedings of IEEE Big Data.

Bibliographical references (2/4)

- Ester M, et al. (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of KDD.
- Frentzos E, et al. (2007) Index-based most similar trajectory search. In Proceedings of ICDE.
- Georgiou H, et al. (2018) Moving objects analytics: survey on future location & trajectory prediction methods. Technical Report. arXiv:1807.04639.
- Georgiou H, et al. (2019) Semantic-aware aircraft trajectory prediction using flight plans. Int. J. Data Sci. and Analytics.
- Giannotti F, et al. (2007) Trajectory pattern mining. In Proceedings of KDD.
- Gudmundsson J, van Kreveld MJ (2006) Computing longest duration flocks in trajectory data. In Proceedings of GIS.
- Jeung H, et al. (2008) Discovery of convoys in trajectory databases. In Proceedings of VLDB.
- Laube P, et al. (2005) Discovering relative motion patterns in groups of moving point objects. Int. J. Geo, Info. Sci., 19(6).
- Lee JG, et al. (2008) Trajectory outlier detection: A partition-and-detect framework. In Proceedings of ICDE.
- Lee JG, et al. (2007) Trajectory clustering: a partition-and-group framework. In Proceedings of SIGMOD.
- Li Z, et al. (2010) Swarm: Mining relaxed temporal moving object clusters. Proceedings of VLDB, 3(1).
- Lin N, et al. (2014) An overview on study of identification of driver behavior characteristics for automotive control. Math. Probl. in Eng.

Bibliographical references (3/4)

- Meratnia N, de By RA (2004) Spatiotemporal compression techniques for moving point objects. In Proceedings of EDBT.
- Monreale A, et al. (2009) WhereNext: a location predictor on trajectory pattern mining. In Proceedings of KDD.
- Nanni M, Pedreschi D (2006) Time-focused clustering of trajectories of moving objects. J. Intelli. Info. Sys., 27(3).
- Palma AT, et al. (2008) A clustering-based approach for discovering interesting places in trajectories. In Proceedings of ACM-SAC.
- Parent C, et al. (2013) Semantic trajectories modeling and analysis. ACM Computing Surveys, 45(4), Article no. 42.
- Patroumpas K, et al. (2017) Online event recognition from moving vessel trajectories. GeoInformatica, 21(2).
- Patroumpas K, et al. (2015): Event Recognition for Maritime Surveillance. In Proceedings of EDBT.
- Pelekis N, et al. (2017a) In-DBMS sampling-based sub-trajectory clustering. In Proceedings of EDBT.
- Pelekis N, et al. (2017b) On temporal-constrained sub-trajectory cluster analysis. Data Mining and Knowl. Disc., 31(5).
- Pelekis N, Theodoridis Y (2014) Mobility data management and exploration. Springer.
- Quddus MA, et al. (2007) Current map-matching algorithms for transport applications: state-of-the-art and future research directions. Transp. Res. Part C: Emerging Technologies, 15(5).
- Quddus MA, et al. (2003) A general map matching algorithm for transport telematics applications. GPS Solutions, 7(3).

Bibliographical references (4/4)

- Tampakis P, et al. (2019) Scalable distributed sub-trajectory clustering. In Proceedings of IEEE Big Data.
- Tampakis P, et al. (2020) Distributed subtrajectory join on massive datasets. ACM Trans. Spatial Algorithms & Systems, 6(2), article no. 8.
- Tao Y, et al. (2004) Prediction and indexing of moving objects with unknown motion patterns. In Proceedings of SIGMOD.
- Trasarti R, et al. (2017) MyWay: location prediction via mobility profiling. Inf. Syst. 64, pp. 350-367.
- Tritsarolis A, et al. (2021) Online discovery of co-movement patterns in mobility data. Int. J. Geogr. Inf. Sci. 35(4).
- Tritsarolis A., et al. (2024) On Vessel Location Forecasting and the Effect of Federated Learning. In Proceedings of MDM.
- Vlachos M, et al. (2002) Discovering similar multidimensional trajectories. In Proceedings of ICDE.
- Vouros GA, et al. (2018) Big data analytics for time critical mobility forecasting: recent progress and research challenges. In Proceedings of EDBT.
- Wang W, et al. (2019) Driving style analysis using primitive driving patterns with Bayesian nonparametric approaches. IEEE Trans Int. Transp. Sys. 20(8).
- Yan Z, et al. (2011) SeMiTri: A Framework for Semantic Annotation of Heterogeneous Trajectories. In Proceedings of EDBT.
- Yan Z, et al. (2012) Semantic trajectories: Mobility data computation and annotation. ACM Trans. Intelligent Systems and Technology, 9(4), Article no. 49.

Acknowledgments

Research supported by EU grants:

- **EMERALDS** – Extreme-scale Urban Mobility Data Analytics as a Service. 2023-25 [emeralds-horizon.eu]
- **Green.Dat.AI** – Energy-efficient AI-ready Data Spaces. 2023-25 [greendatai.eu]
- **MobiSpaces** – New data spaces for green mobility. 2022-25 [mobispaces.eu]
- **VesselAI** – Enabling Maritime Digitalization by Extreme-scale Analytics, AI and Digital Twins. 2021-23 [vessel-ai.eu]
- **Track & Know** – Big Data for Mobility Tracking Knowledge Extraction in Urban Areas. 2018-20 [trackandknowproject.eu]
- **MASTER** – Multiple Aspect Trajectory Management and Analysis, 2018-22 [master-project-h2020.eu]
- **datAcron** – Big Data Analytics for Time Critical Mobility Forecasting, 2016-18 [datacron-project.eu]

The Data Science Lab @ UniPi.GR

Our research agenda:

- **Extreme-scale mobility data processing**
- **Mobility data analytics at the edge**
- **Time series analytics & forecasting**
- **Data fusion & semantic integration**
- etc.

https://www.datastories.org