Data Science – the case of Mobility Data

Yannis Theodoridis

InfoLab | University of Piraeus | Greece
infolab.cs.unipi.gr

Univ. Cyprus, Nov. 2012

What is "Data Science" from quora.com

- *"Looking at data"*
- *"Tools and methods used to analyze large amounts of data"*
- *"Anything you can do to get knowledge out of data"*
  - finding and gathering data, data mining and preprocessing, EDA, statistics, machine learning, natural language processing and data visualization

What is "Data Science" from quora.com

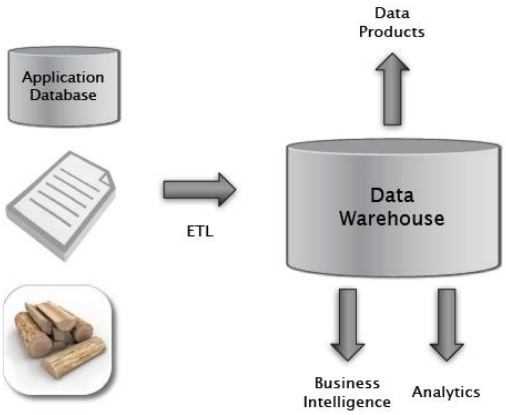
- *"The set of practices targeted at the storage, management and analysis of data sets large enough that require distributed computing and storage resources"*
  - Toolkit is a very broad one: distributed DBMS, noSQL, mapReduce, complexity of algorithms, visualization of large data sets, algorithms from Machine Learning, …

Historical note:

- the job title 'Data Scientist' came into its recent vogue when Jeff Hammerbacher (Berkeley) coined the term at Facebook in 2007

The processes involved in Data Science

- Data acquisition and cleaning
- Data integration, storage and processing
  - Flat files vs. relational vs. noSQL databases
  - Serial vs. parallel/distributed (e.g. Hadoop) processing
- Visualizing data for exploration and presentation
  - Graphs, plots, etc.
- Learning from / Building models upon data
  - ML/DM algorithms; Visual analytics
- Addressing data privacy issues

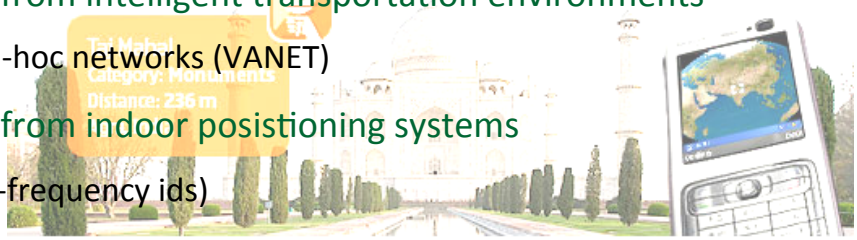


What is Mobility Data

- Large diffusion of mobile devices, mobile and location-based services
  ➔ **mobility data**

What is Mobility Data

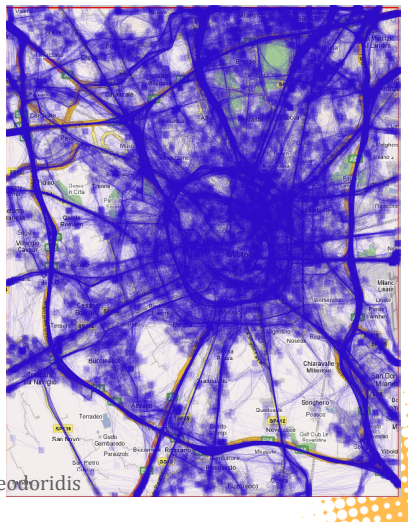
- Location data from GPS-equipped devices
  - Humans (pedestrians, drivers) with GPS-equipped smartphones
  - Vessels with AIS transmitters (due to maritime regulations)
- Location data from mobile phones
  - i.e., cell positions in the GSM/UMTS network
- Location data from intelligent transportation environments
  - Vehicular ad-hoc networks (VANET)
- Location data from indoor posistioning systems
  - RFIDs (radio-frequency ids)
  - Wi-Fi access points



Examples of Mobility Data

- Raw data: GPS recordings

```
N; Time; Lat; Lon; Height; Course; Speed; PDOP; State; NSat
…
8;22/03/07 08:51:52;50.777132;7.205580; 67.6;345.4;21.817;3.8;1808;4
9;22/03/07 08:51:56;50.777352;7.205435; 68.4;35.6;14.223;3.8;1808;4
10;22/03/07 08:51:59;50.777415;7.205543; 68.3;
11;22/03/07 08:52:03;50.777317;7.205877; 68.8;
12;22/03/07 08:52:06;50.777185;7.206202; 68.1;
13;22/03/07 08:52:09;50.777057;7.206522; 67.9;
14;22/03/07 08:52:12;50.776925;7.206858; 66.9;
15;22/03/07 08:52:15;50.776813;7.207263; 67.0;
16;22/03/07 08:52:18;50.776780;7.207745; 68.8;
17;22/03/07 08:52:21;50.776803;7.208262; 71.1;
18;22/03/07 08:52:24;50.776832;7.208682; 68.6;
…
```

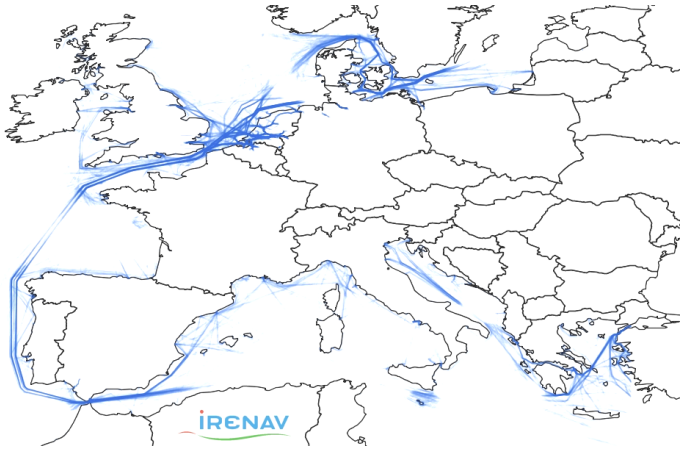


Examples of real trajectory datasets (1)

- GPS signals from vehicles driving in urban areas. Example:
  - "Milano dataset"; ~2M GPS recordings from 17241 distinct objects during 1 week

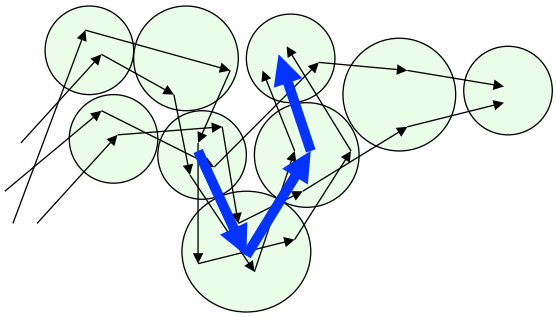
Examples of real trajectory datasets (2)

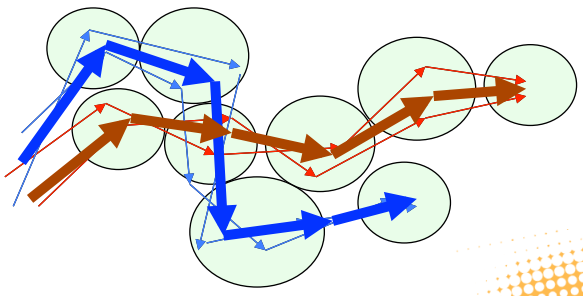
- AIS signals from vessels sailing at sea. Example:
  - "IMIS dataset – 3 days"; ~4.5M GPS recordings from 1753 distinct objects during 3 days ("in Eastern Mediterranean"
  - Full version: 3Tbytes data covering 3 years of activity



What is Data Science for Mobility Data

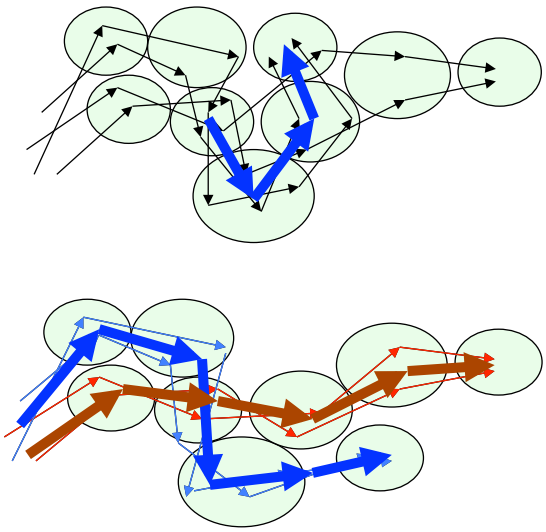
- *"Looking at **mobility** data"*
- *"Tools and methods used to analyze large amounts of **mobility** data"*
- *"Anything you can do to get knowledge out of **mobility** data"*
- *"The set of practices targeted at the storage, management and analysis of **mobility** data sets large enough that require distributed computing and storage resources"*





What is Data Science for Mobility Data

- Focusing on human activity …
  - *"the opportunity to discover, from the **digital traces** of human activity, the **knowledge** that makes us comprehend timely and precisely the way we live, the way we use our time and our land"* [1]



[1] Giannotti, F. and Pedreschi, D. 2008. Mobility, Data Mining and Privacy: Geographic Knowledge Discovery. Springer.

What can we learn from mobility data …

Querying vehicles datasets…

- (global) Traffic monitoring
  - How many cars are in the ring of the town?
  - Once an accident is discovered, immediately send alarm to the nearest police and ambulance cars

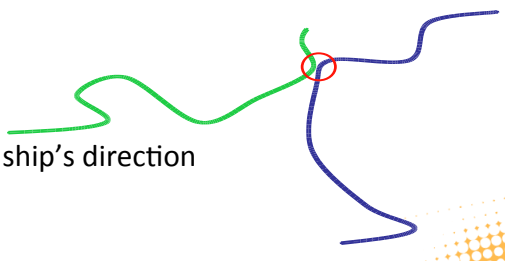
- (personalized) Location-aware queries
  - Where is my nearest Gas station?
  - What are the fast food restaurants within 3 miles from my location?
  - Let me know if I am near to a restaurant while any of my friends are there
  - Get me the list of all customers that I am considered their nearest restaurant

Querying vessels datasets…

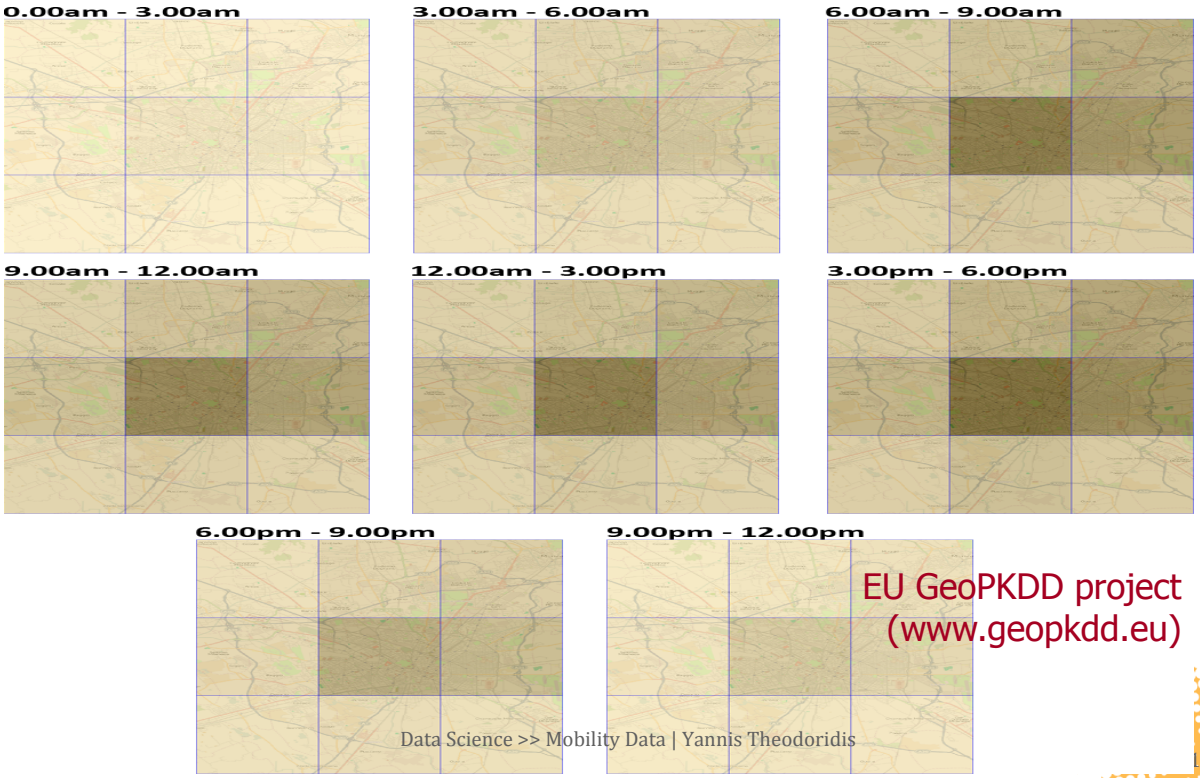
(requirements from Greek Maritime Conservation Agencies)

- Derived information
  - Extract / draw the ship tracks (detailed vs. simplified)
  - Calculate min distances between ships
  - Calculate max number of ships in the vicinity of another ship
  - Find whether (and how many times) a ship goes through specified areas (e.g. narrow passages, biodiversity boxes)

- Further analysis of trajectory data
  - Calculate the number of sharp changes in ship's direction
  - Find typical routes vs. outliers

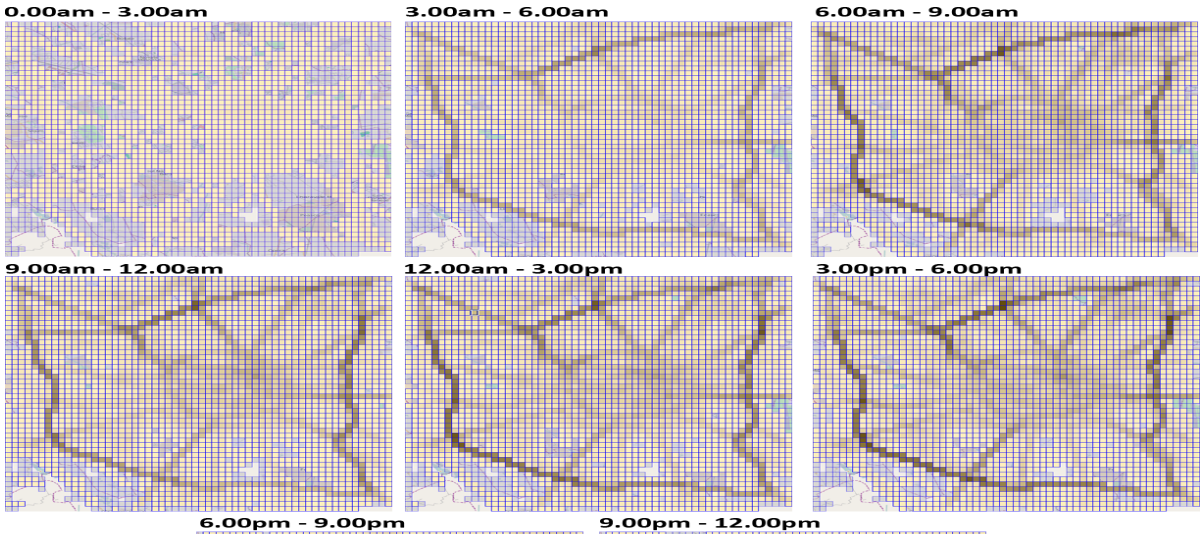


OLAP analysis (1. coarser detail)



EU GeoPKDD project
(www.geopkdd.eu)

OLAP Analysis (2. finer detail)



EU GeoPKDD project
(www.geopkdd.eu)

Key questions that arise

Q1. Which kind of **management** is appropriate for mobility data

- (batch vs. streaming) MOD engines
- Primitive vs. advanced queries on trajectory data

Q2. What kind of **analysis** fits to mobility data

- Analysis on current locations vs. trajectories
- Clusters, frequent patterns, anomalies / outliers

Q3. How to **preserve data privacy**

Outline of the (remaining) talk

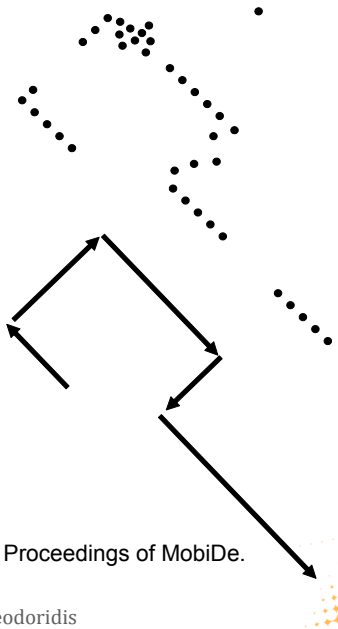
- (Preparatory step) **moving object database management**
  - Collecting mobility data - the trajectory reconstruction problem
  - Storing and querying trajectory databases
- **Trajectory OLAP analysis**
- **Trajectory data mining**
  - Frequent pattern mining; Trajectory clustering; etc.
- **Summary – the next step**
  - Semantic trajectories

Moving Object Database Management

- Collecting mobility data (& the trajectory reconstruction problem)
- Storing and querying trajectory databases

The trajectory reconstruction problem

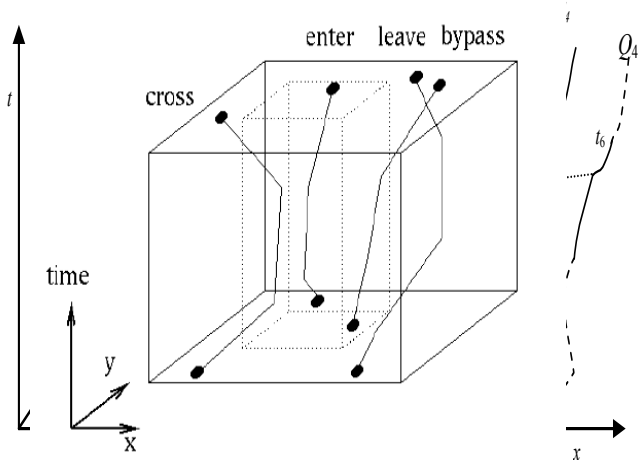
- From raw data, i.e., time-stamped locations
  - Raw data (3D points) arrive either one-by-one or in bulks
- … to trajectory data, i.e., continuous evolutions
  - Redundancy is reduced, noise is removed, etc.
- A solution [2]: filters / thresholds decide whether a new series of data is to
  - be appended to an existing trajectory, or
  - initiate a new trajectory, or
  - be considered as noise



[2] Marketos, G. et al. 2008. Building Real-world Trajectory Warehouses. In Proceedings of MobiDe.

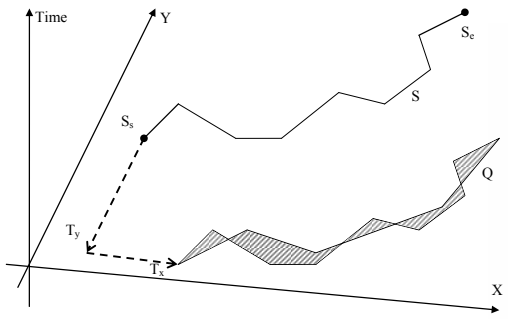
Querying trajectory data

- Primitive search operations:
  - Coordinate-based
    - Range, NN
  - Trajectory-based
    - Topological, Directional
- As well as more advanced…

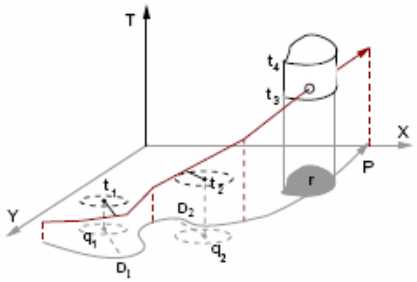


Querying trajectory data

- **Trajectory similarity** queries [3]
  - "Given a query trajectory *Q*, find the k- most similar trajectories to *Q* (perhaps, constrained is space and/or time)"



- **Spatio-temporal pattern** queries [4]
  - "Find objects that crossed through region *A* at time t_1, came as close as possible to point *B* at a later time t_2 then stopped inside circle *C* during interval (t_3, t_4)"



[3] Frentzos, E. et al. 2007. Index-based Most Similar Trajectory Search. In Proceedings of ICDE.
[4] Hadjieleftheriou, M. et al. 2005. Complex Spatio-temporal Pattern Queries. In Proceedings of VLDB.

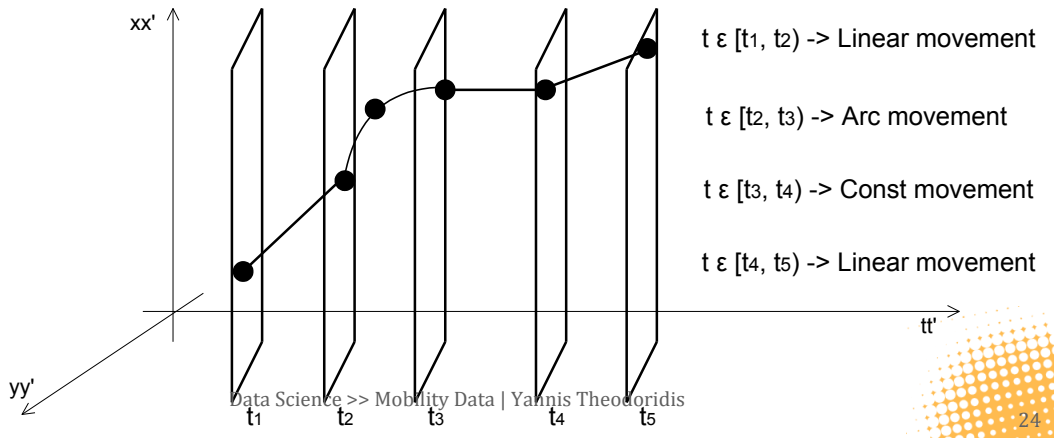
Moving Objects Database Systems

- From traditional (relational or spatial) DBMS to **Moving Object Database (MOD) engines**
  - Data types, indices, query processing & optimization strategies for trajectories
- Both spatial and temporal dimensions are considered as first-class citizens.
- State-of-the-art MOD engines
  - **HERMES** [5], **SECONDO** [6]

[5] Pelekis, N., et al. 2008. HERMES: Aggregative LBS via a trajectory DB engine. In Proceedings of SIGMOD.
[6] Güting, R. H. et al. 2010. SECONDO: A Platform for Moving Objects Database Research and for Publishing and Integrating Research Implementation. IEEE Data Engineering Bulletin, 33(2):56-63

http://infolab.cs.unipi.gr/hermes/

- A palette of Abstract Data Types on top of an extensible DBMS
  - Moving point, moving line, moving polygon, etc.
  - R-tree and TB-tree indexing support



Trajectory OLAP Analysis

- Data cubes (dimensions and measures)
- OLAP analysis demonstration

Trajectory OLAP analysis

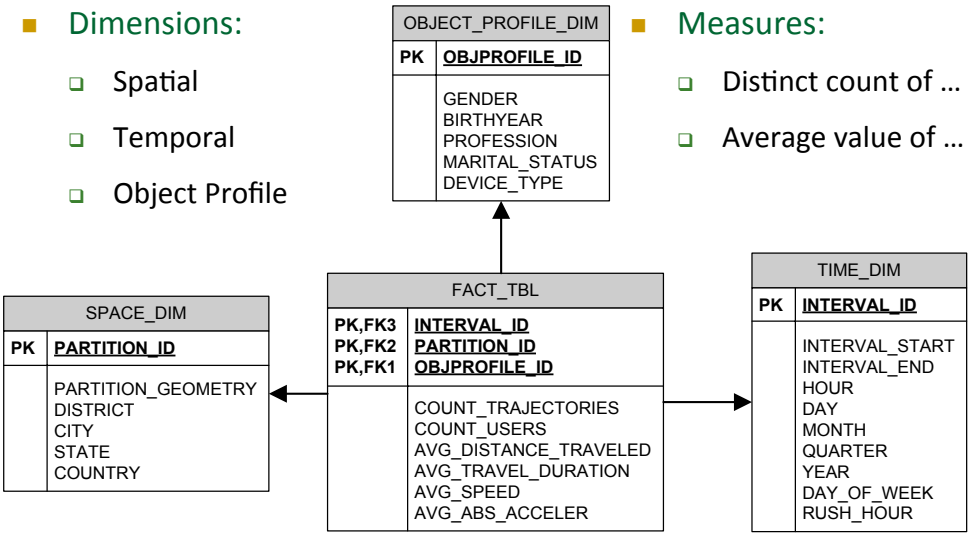
- Data warehouses:
  - "Subject-oriented, integrated, time-variable, non-volatile information systems aiming at decision making" [7]
- Data cubes:
  - Aggregates over the operational databases (using an Extract-Transform-Load (ETL) process)
  - Technically, a collection of relations (if relational model is adopted)
  - Typical structure: star schema
    - A fact table with measures + dimension tables with hierarchies

[7] Inmon, B. 1992. Building the Data Warehouse. 1st Edition. Wiley and Sons.

A data cube for trajectories

Our approach [8]:

- Dimensions:
  - Spatial
  - Temporal
  - Object Profile
- Measures:
  - Distinct count of …
  - Average value of …



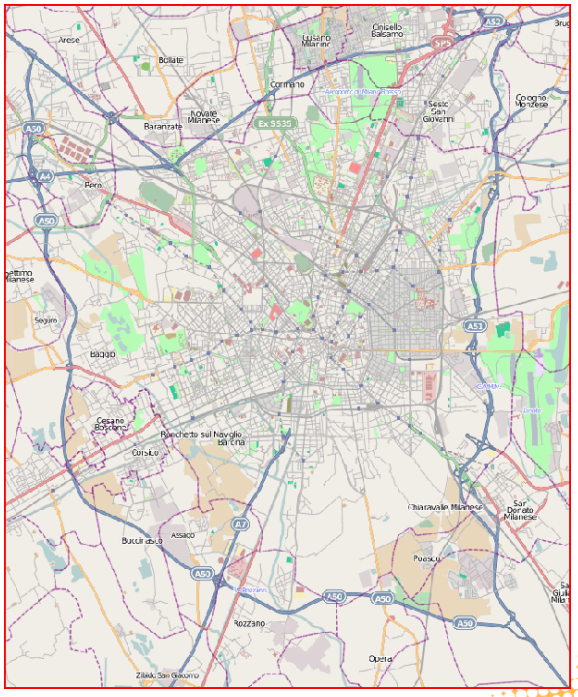
[8] Marketos, G. & Theodoridis, Y. 2010. Ad-hoc OLAP on Trajectory Data. In Proceedings of MDM.

Typical OLAP analysis (from end-users' point of view)

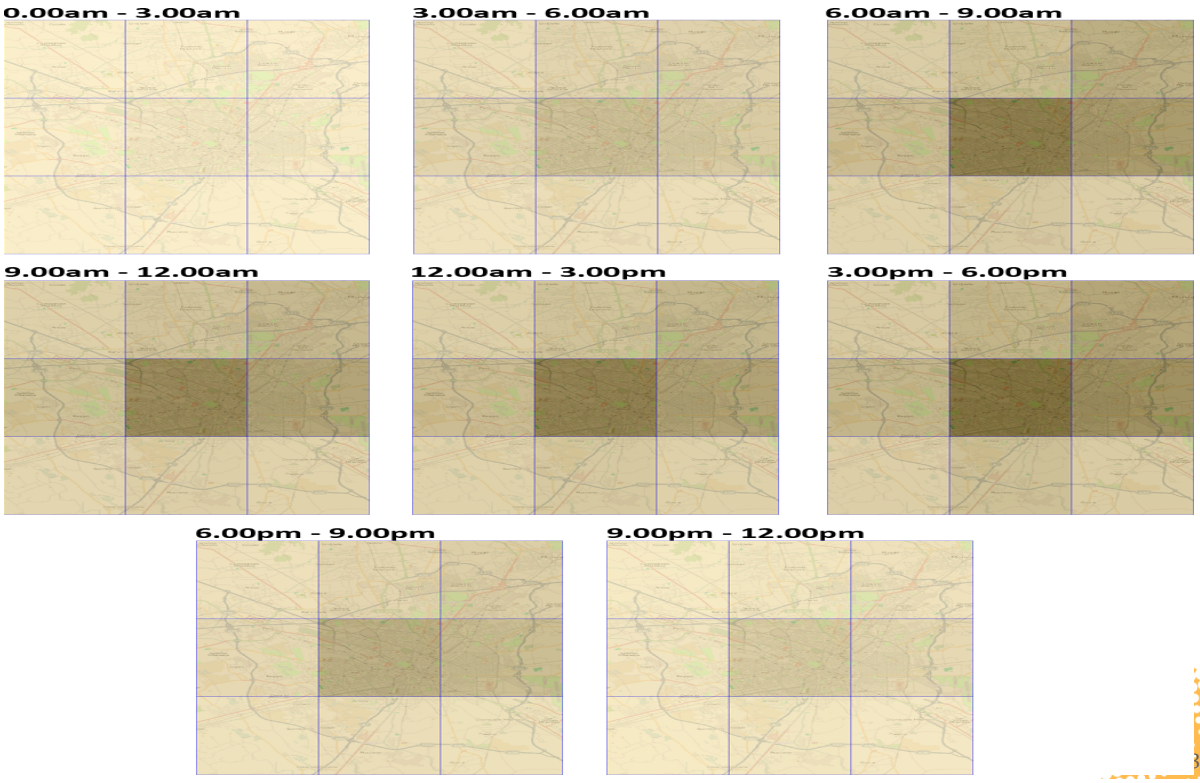
- How does traffic flow and speed change along the week?
  - Q1: Where does the highest traffic appear? at what hour?
  - **A1: unclassified choropleth map (for a specific period of time)**
  - Q2: What exactly happens at the road network level?
  - **A2: drill-downs in space and/or time**
  - Q3: How does movement propagate from place to place?
  - **A3: data cube measures' correlation (speed vs. presence)**

Milano dataset

- What?
  - 2M observations (GPS recordings)
    - for 7 days (Sun. 1 - Sat. 7 April ’07)
  - 200K trajectories (after reconstruction)
- How?
  - Stored in Hermes MOD engine
  - Feeding a trajectory data cube

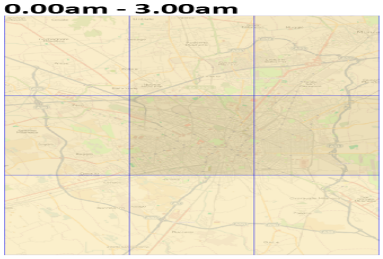


Presence on Tuesday
(coarser level)

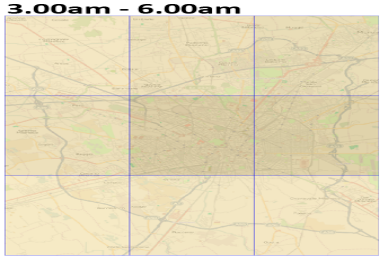


Presence on Saturday
(coarser level)

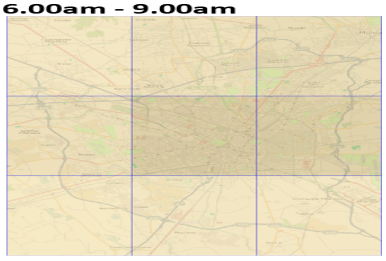
0.00am - 3.00am



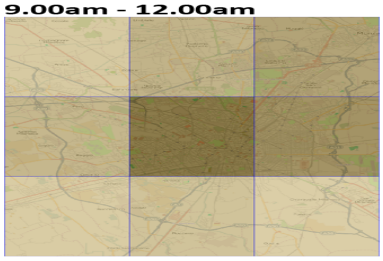
3.00am - 6.00am



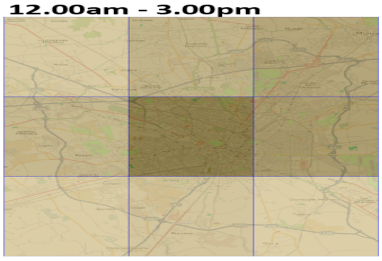
6.00am - 9.00am



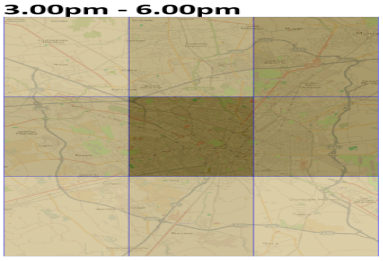
9.00am - 12.00am



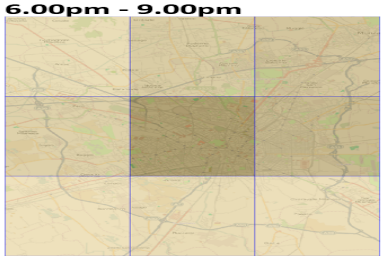
12.00am - 3.00pm



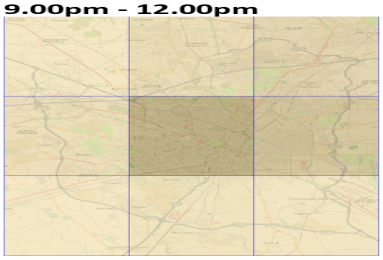
3.00pm - 6.00pm



6.00pm - 9.00pm

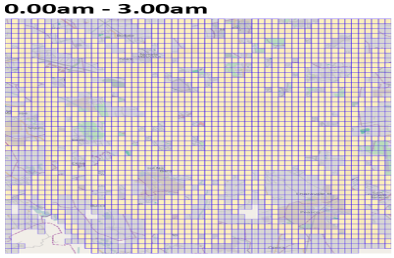


9.00pm - 12.00pm

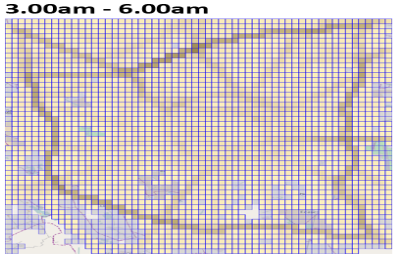


Presence on Tuesday
(finer - road network - level)

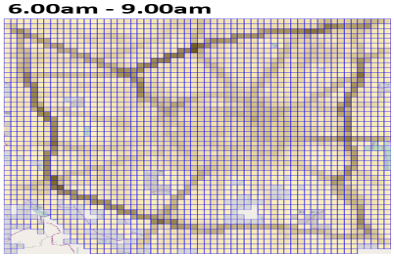
0.00am - 3.00am



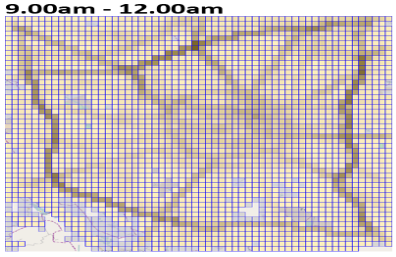
3.00am - 6.00am



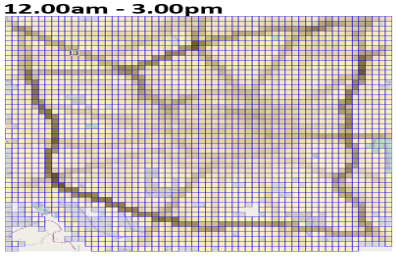
6.00am - 9.00am



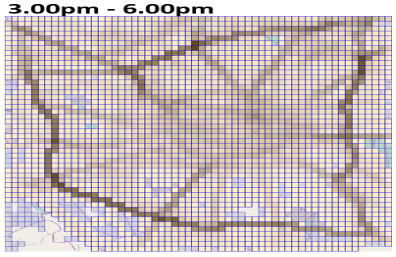
9.00am - 12.00am



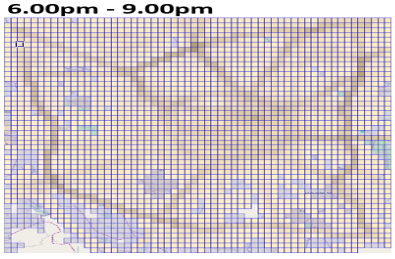
12.00am - 3.00pm



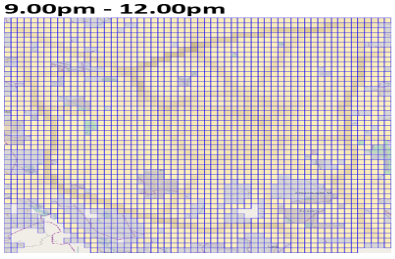
3.00pm - 6.00pm



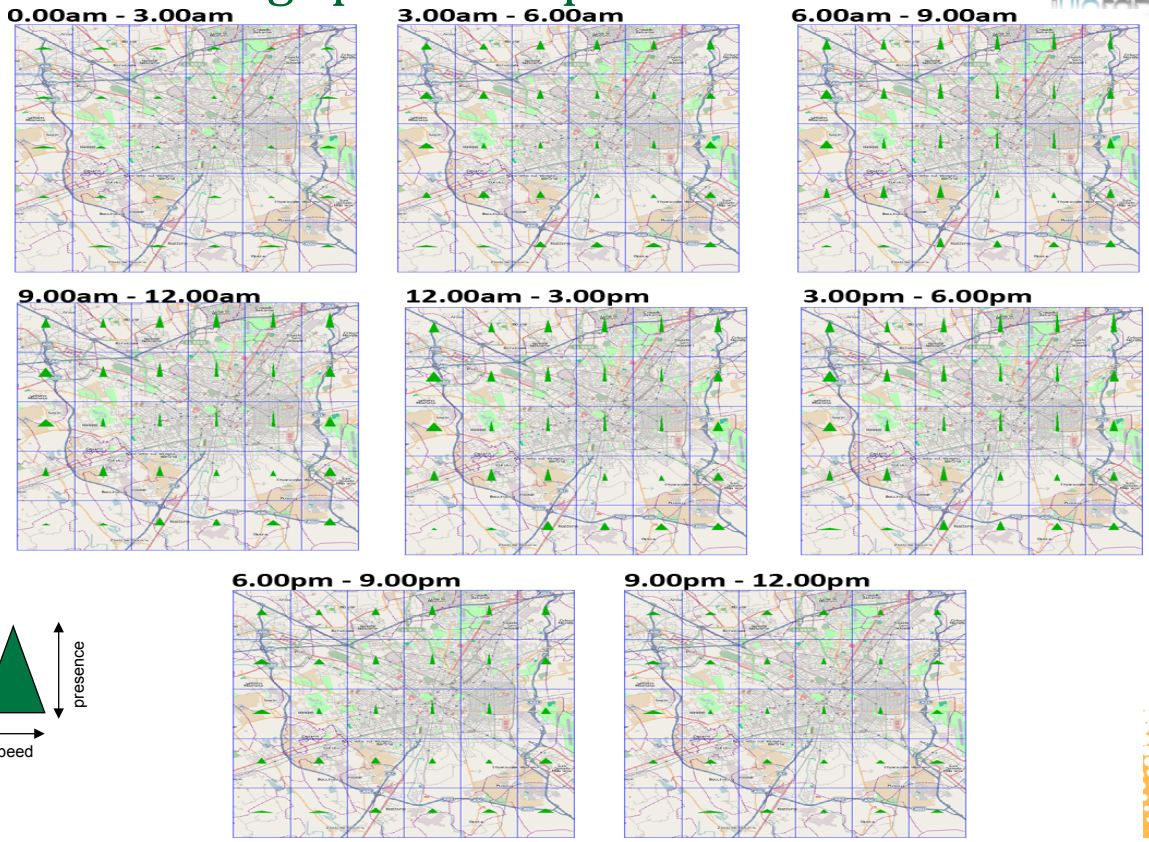
6.00pm - 9.00pm



9.00pm - 12.00pm



Correlating speed and presence

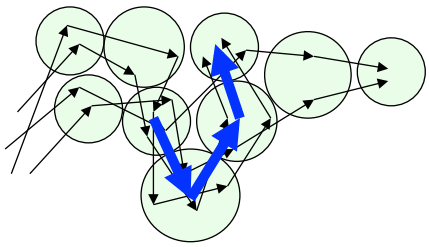


Trajectory Data Mining

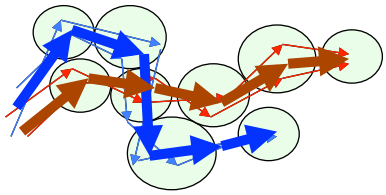
- Frequent pattern mining
- Trajectory clustering

Examples of mobility data mining

- **Frequent pattern mining**
  - Identify 'frequent' or 'popular' patterns
  - Discover hot spots, hot paths, etc.



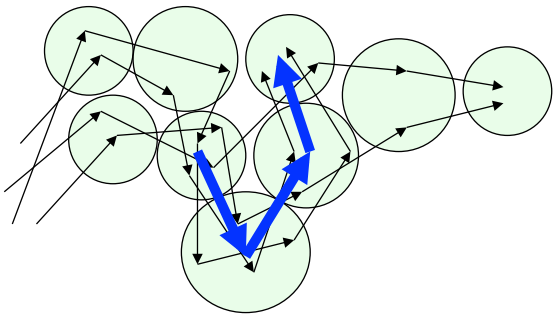
- **Trajectory clustering**
  - Cluster trajectories w.r.t. similarity
    - For each cluster, find its 'centroid' or 'representative'
  - Discover moving clusters (flocks), outliers, etc.



- **Trajectory classification**
  - Assign trajectories to predefined classes
  - Find rules that may predict future behavior of moving objects

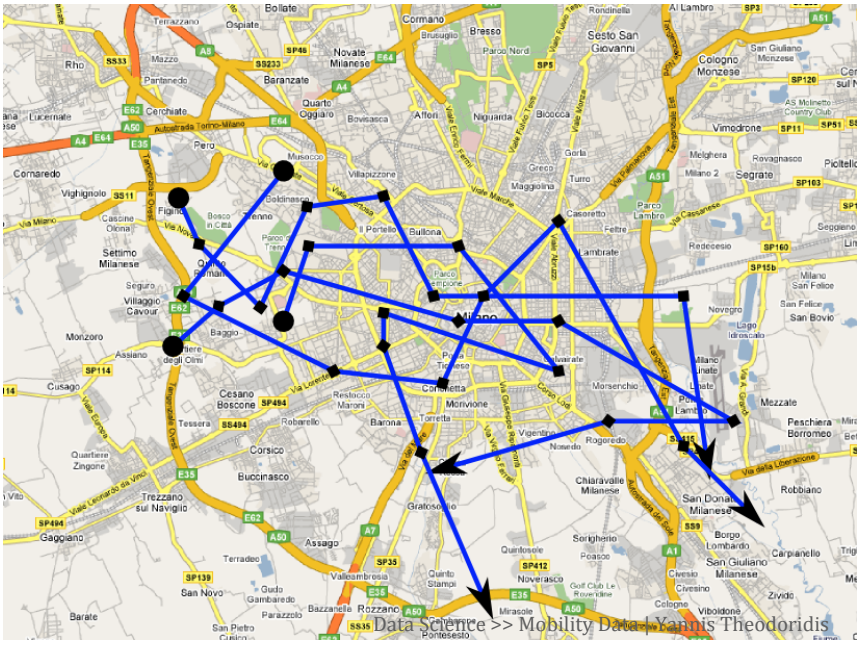
"Frequent pattern mining" techniques

- Technical objectives:
  - Identify 'frequent' or 'popular' patterns
  - Discover hot spots, hot paths, etc.
- State-of-the-art:
  - T-Patterns [9]



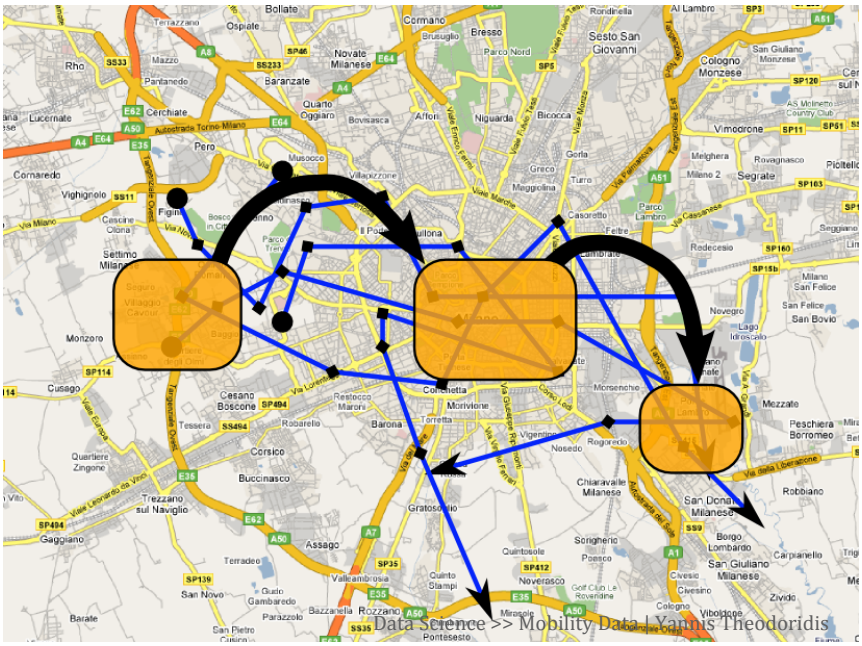
[9] Giannotti, F. et al. 2007. Trajectory Pattern Mining. In Proceedings of KDD.

What is a “frequent pattern” for trajectories?



T-patterns

- **T-pattern** is a sequence of visited regions, **frequently** visited in the **specified order** with **similar transition times**



T-Pattern discovery

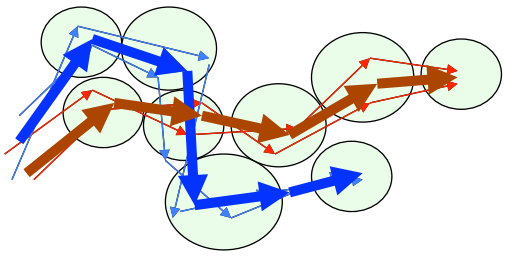
Input:
Trajectory
Dataset

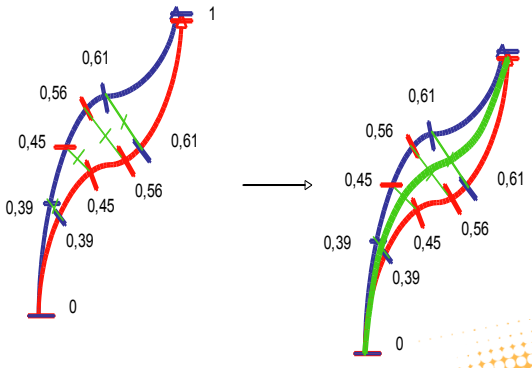
Output: T-Patterns

Intermediate result:
Regions of Interest

Clustering mobility data

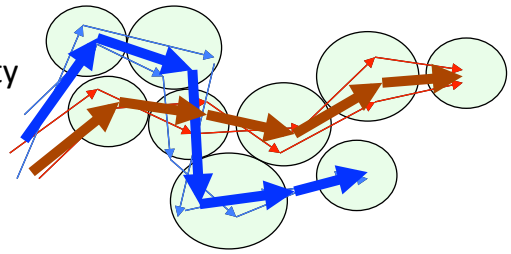
- Questions arising:
  - Which distance function between trajectories?
  - How do we define intra- and inter-cluster distances?
  - Which kind of clustering? (e.g. Partitioning vs. Density-based)
  - How does a 'centroid' look like?
    - A "trajectory" representing the members of a cluster, as better as possible





“Trajectory clustering” techniques

- Technical objectives:
  - Cluster trajectories w.r.t. similarity
    - For each cluster, find its ‘centroid’ or ‘representative’
  - Discover moving clusters (flocks), outliers, etc.

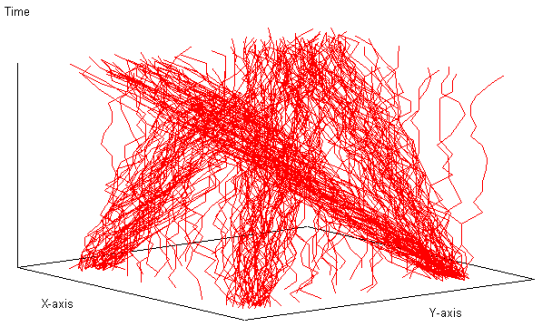


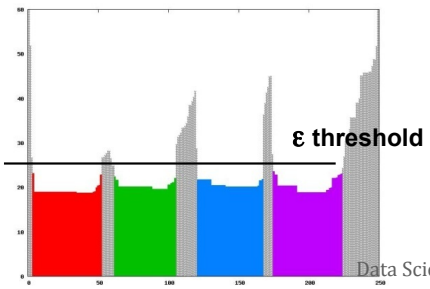
- State-of-the-art
  - Density-based clustering: **T-OPTICS** [10]
  - Partition-based clustering: **CenTR-I-FCM** [11]

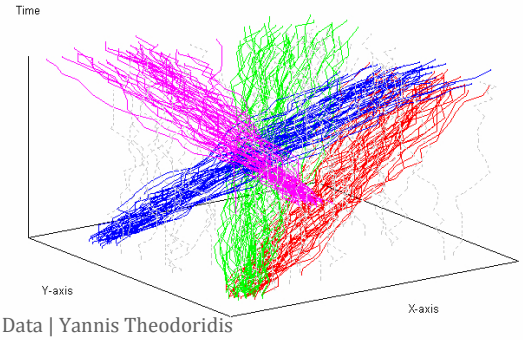
[10] Nanni, M. & Pedreschi, D. 2006. Time-focused clustering of trajectories of moving objects. J. of Intelligent Information Systems, 27: 267-289.
[11] Pelekis, N. et al. 2009. Clustering Trajectories of Moving Objects in an Uncertain World. In Proceedings of ICDM.

T-OPTICS

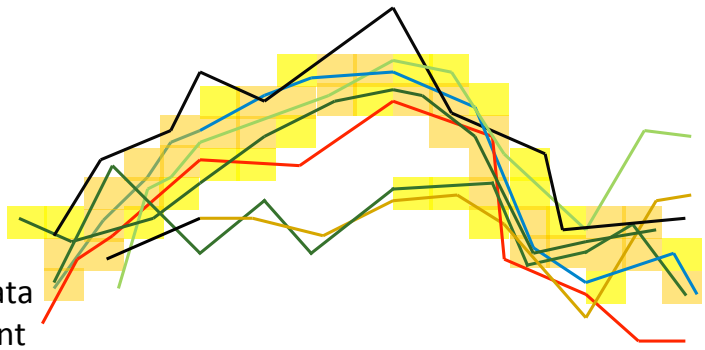
- Builds upon OPTICS
- Keywords:
  - distance, core trajectories, reachability
- Reachability plot (valleys and hills)
  - Valleys → clusters !!







CenTR-I-FCM: Clustering under uncertainty



- Builds upon Fuzzy-C-Means
  - a variation of K-means for uncertain data
- Motivation:
  - uncertainty of trajectory data should be taken into account
- Three phases:
  - Step 1: **mapping** of trajectories in an intuitionistic fuzzy vector space
  - Step 2: **discovering the centroid** of a bundle of trajectories (algorithm CenTra)
  - Step 3: **clustering** trajectories under uncertainty (algorithm CenTR-I-FCM)

Summarizing…

Summary

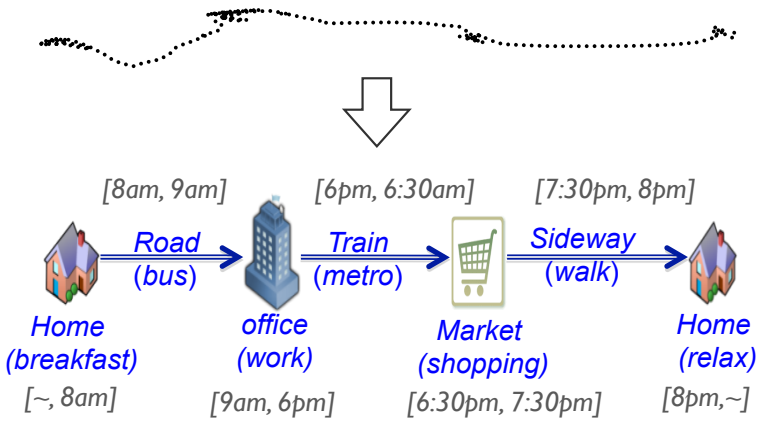
- Data Science and the case of Mobility Data
  - "*making data tell its story*"
  - A big challenge – asking for efficient, scalable algorithms
    - Be reminded of the "BIG (mobility) DATA" story

- Next steps:
  - The "**semantic trajectories**" era
    - Adding context → better mobility understanding
  - From **Location-based (LBSN)** to **Mobile Social Networking (MSN)**
    - Towards complex social networks of moving interacting objects.

From 'raw' to semantic trajectories

(raw) trajectories
e.g., GPS feeds

meaningful tuples
(aka, episodes) [12]



• Semantic Trajectory: a sequence of episodes $T=\{e_{first}, \ldots, e_{last}\}$

• Episode: $e_i=(t_{from}, t_{to}, place, tag)$

[12] Parent, C. et al. 2013. Semantic Trajectories Modeling and Analysis. ACM Computing Surveys, to appear.

Why semantic trajectories?

- Trajectory = **a sequence of episodes (stops/moves)**
  - E.g., home, shopping, move with bus, in train …
  - Better mobility understanding

Raw GPS Points

(traditional) trajectory

Semantic trajectory

	Road name	Start time
Walk	Ch. veilloud	08:50:26
	Rt. du Boi	08:54:46
	Rt. de Villar	08:57:24
	Tir Fédéra	08:58:41
Metro	M1	08:59:24
Walk	Rt. de la Sorg	09:03:57
	Ch. du Barrag	09:04:42
	La Diagonal	09:05:24

Mobile social networks

- **Facebook places**
  - Tag yourselves and find tagged friends
- **Foursquare, Gowalla, etc.**
  - Tell your friends where you are, suggest places, etc.

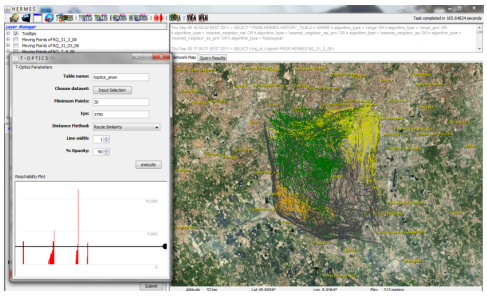
Mobile social networks

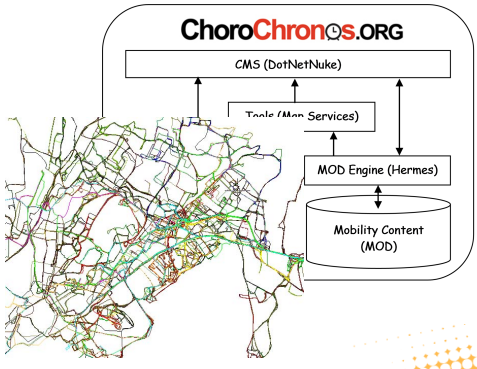
- Objective: **discovering interaction patterns**

Movement behaviour is a complex system of interactions.

InfoLab research results

- S/w tools and prototypes
  - **HERMES MOD engine**
    - 2 implementations: Oracle Spatial vs. PostgreSQL
    - plus its **privacy-preserving skin**, **HERMES++**
  - **ChoroChronos.org repository / CMS** of real mobility datasets
  - **GSTD* generator of synthetic datasets** simulating various movement distributions
- Algorithms
  - **Cen-TR-I-FCM** (for trajectory clustering)
  - **T-Sampling** (for trajectory sampling)
  - **S²T-Clustering** (for sub-trajectory clustering), etc.





Acknowledgments

- Joint work with Dr. Nikos Pelekis and our students @ InfoLab
- Many thanks to the partners of the following projects …
  - FP6 / GeoPKDD (www.geopkdd.eu), 2005-09
  - FP7 / MODAP (www.modap.org), 2009-13
  - ESF / COST-MOVE (move-cost.info), 2009-13
  - FP7 / DATASIM (www.datasim-fp7.eu), 2011-14
  - Marie Curie / SEEK (www.seek-project.eu), 2012-15
- … for the longtime collaboration and brainstorming on this research agenda

Thank you !!