Mobility, Data Mining and Privacy

Fosca Giannotti and Dino Pedreschi

Fosca.Giannotti@isti.cnr.it

Dino.Pedreschi@di.unipi.it

Pisa KDD Lab   www-kdd.isti.cnr.it

University of Pisa and ISTI-CNR, Italy

Tutorial @ ECML PKDD 2008

Antwerp, 15 September 2008

Mobile devices and services

- Large diffusion of mobile devices, mobile services and location-based services

Wireless networks as mobility data collectors

- Wireless networks infrastructures are the **nerves of our territory**
- besides offering their services, they gather highly informative **traces** about the human mobile activities
- UbiComp infrastructure will further push this phenomenon
- Miniaturization, wearability, pervasiveness will produce traces of increasing
  - positioning accuracy
  - semantic richness

Which mobility data?

- Location data from mobile phones, i.e. cell positions in the GSM/UMTS network.
- Location data from GPS-equipped devices – Galileo in the (near?) future
  - Next/current generation of Nokia mobile phones have on-board GPS receiver, and can transmit GPS tracks by SMS/MMS
- Location data from
  - peer-to-peer mobile networks
  - intelligent transportation environments – VANET
  - ad hoc sensor networks, RFIDs (radio-frequency ids)

Mobility, Data Mining and Privacy

- Towards an **archaeology of the present**?
- A scenario of great opportunities and risks:
  - mining mobility data can yield useful knowledge;
  - but, individual privacy is at risk.
- A new multidisciplinary research area is emerging at this crossroads, with potential for broad social and economic impact
  - F. Giannotti and D. Pedreschi (Eds.)
    Mobility, Data Mining and Privacy. Springer, 2008.

A paradigmatic project: ***GeoPKDD***

http://www.geopkdd.eu

A European FP6 project

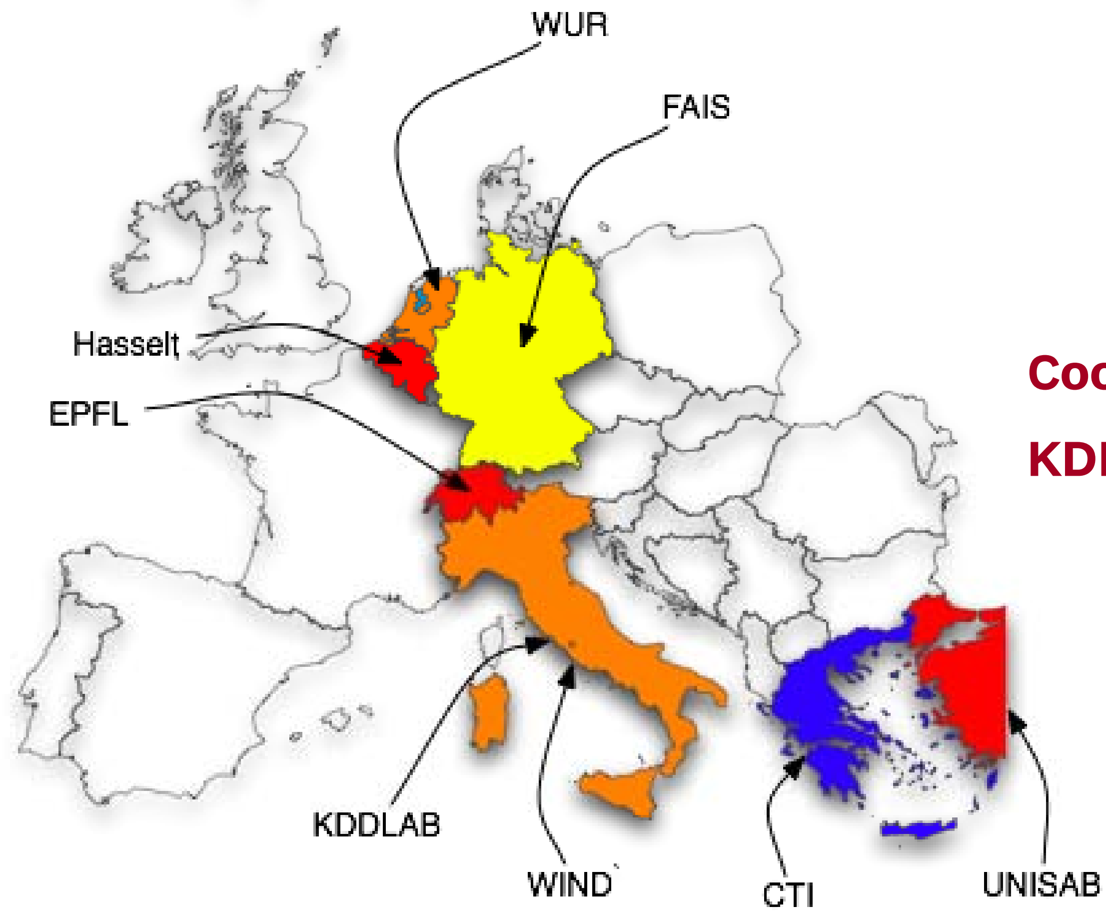
Geographic Privacy-aware

Knowledge Discovery and Delivery

www.geopkdd.eu

GeoPKDD

Geographic Privacy-aware Knowledge Discovery and Delivery



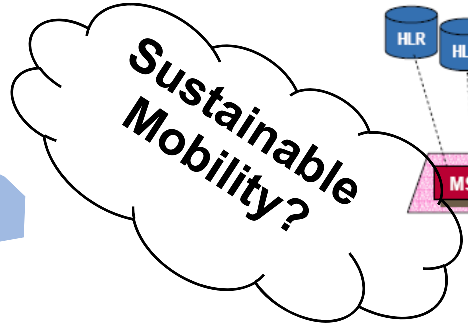
Coordinator:

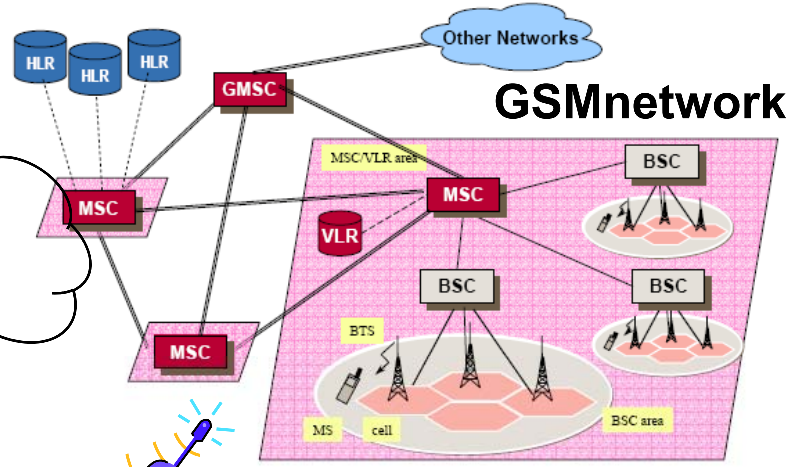
KDD-LAB Pisa, ISTI-CNR

The GeoPKDD scenario

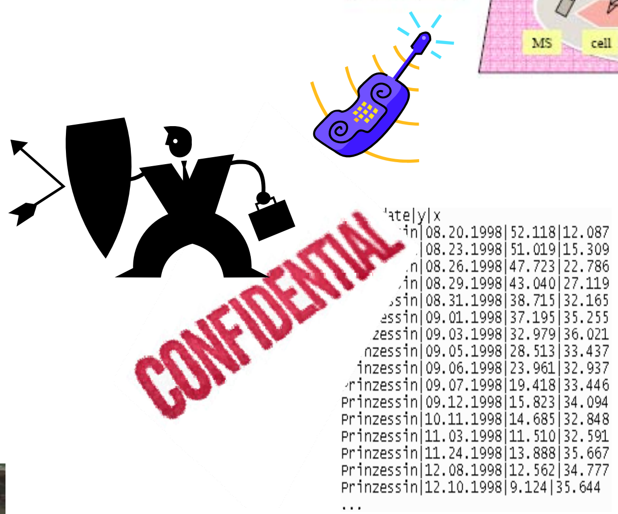
- **From the analysis of the traces of our mobile phones it is possible to reconstruct our mobile behaviour, the way we collectively move**
- **This knowledge may help us improving decision-making in many mobility-related issues:**
  - Planning traffic and public mobility systems in metropolitan areas;
  - Planning physical communication networks
  - Localizing new services in our towns
  - Forecasting traffic-related phenomena
  - Organizing logistics systems
  - Avoid repeating mistakes
  - Timely detecting changes.

Mobility Manager



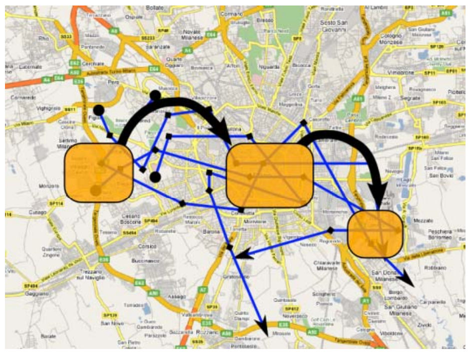


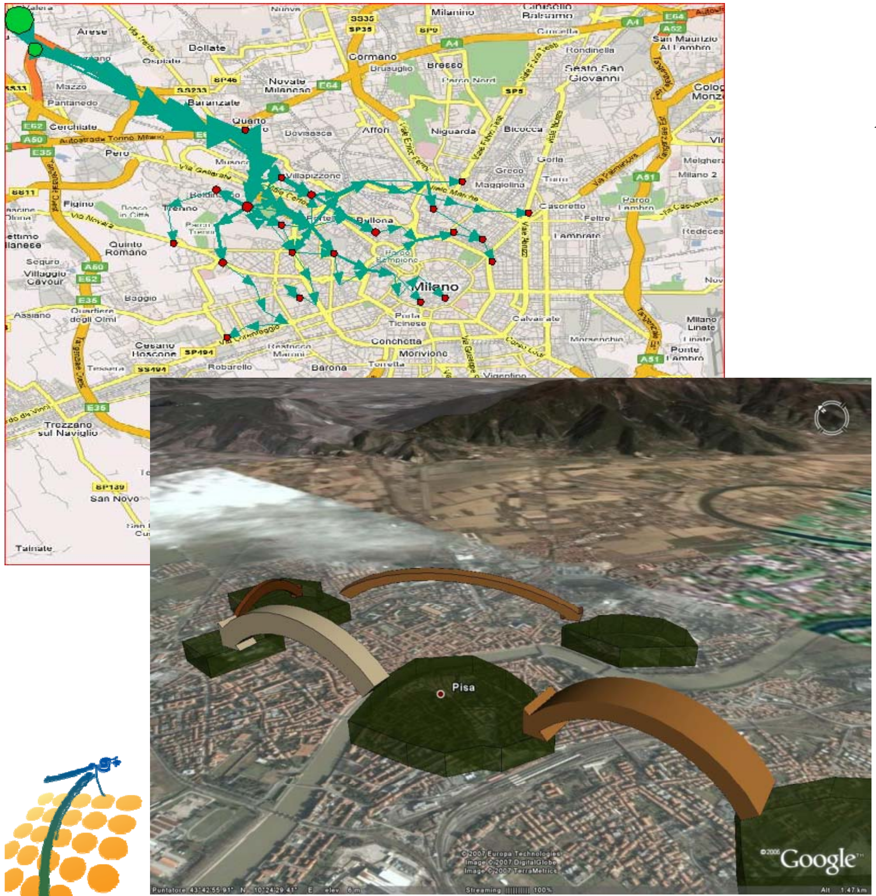
GSMnetwork



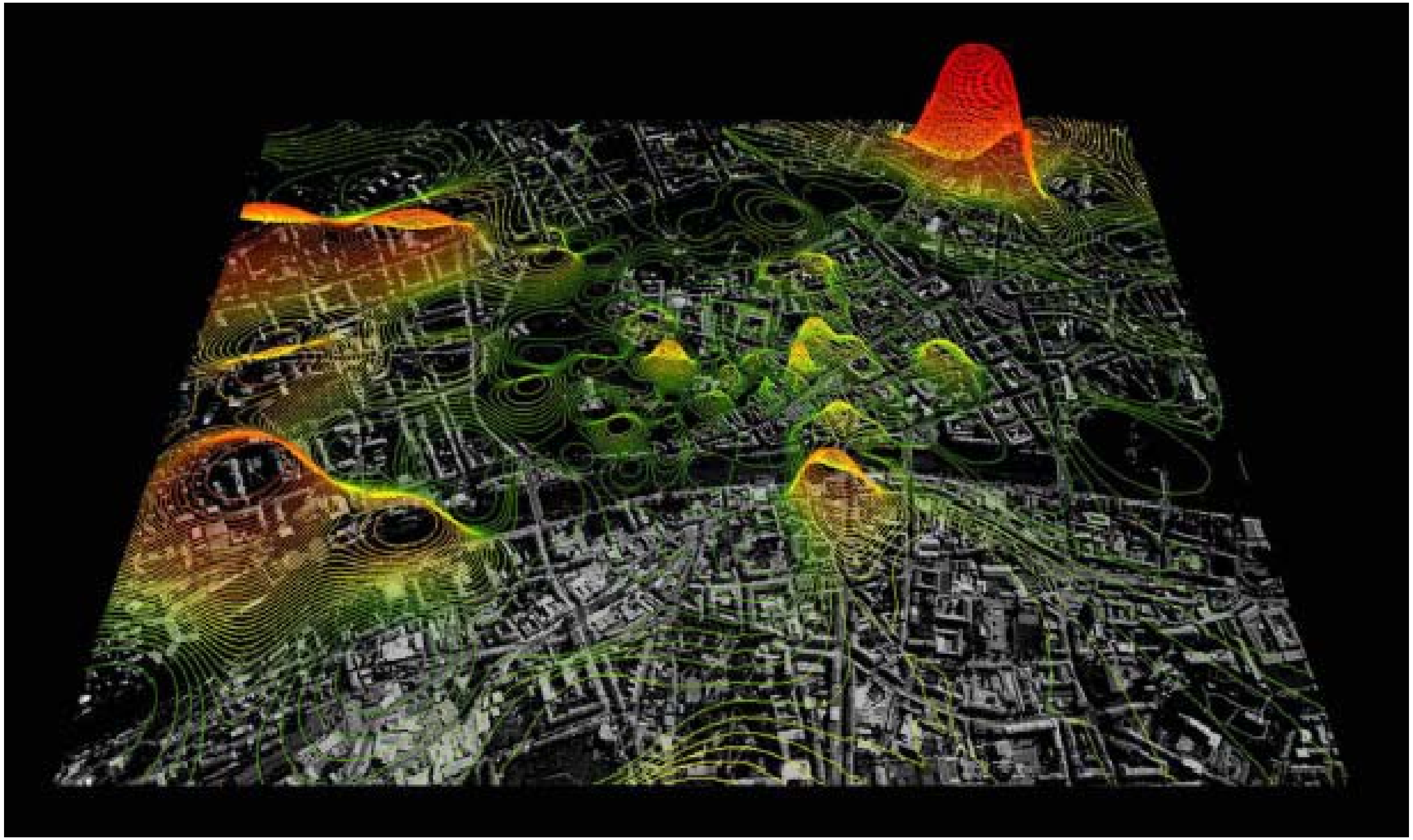
Location data

Mobility models

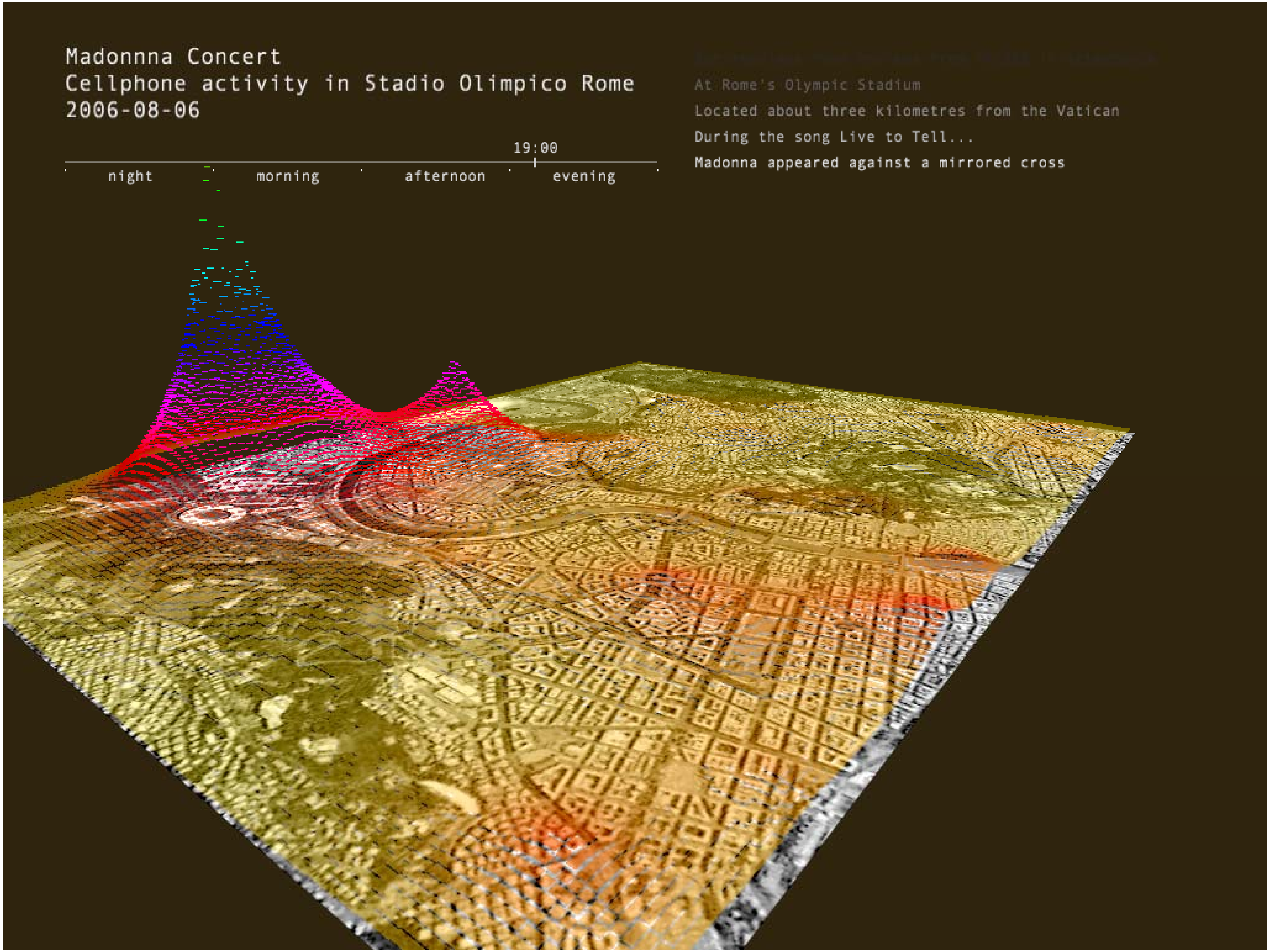




Real-time density estimation in urban areas



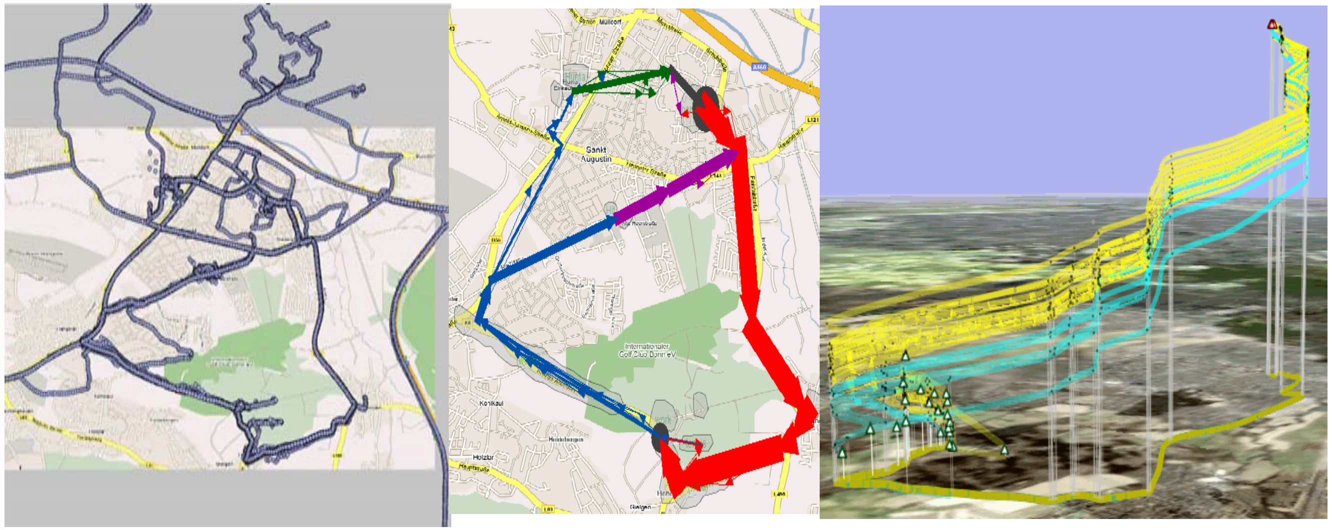
The senseable project: http://senseable.mit.edu/grazrealtime/


Madonnna Concert
Cellphone activity in Stadio Olimpico Rome
2006-08-06
19:00
night
morning
afternoon
evening
At Rome's Olympic Stadium
Located about three kilometres from the Vatican
During the song Live to Tell...
Madonna appeared against a mirrored cross

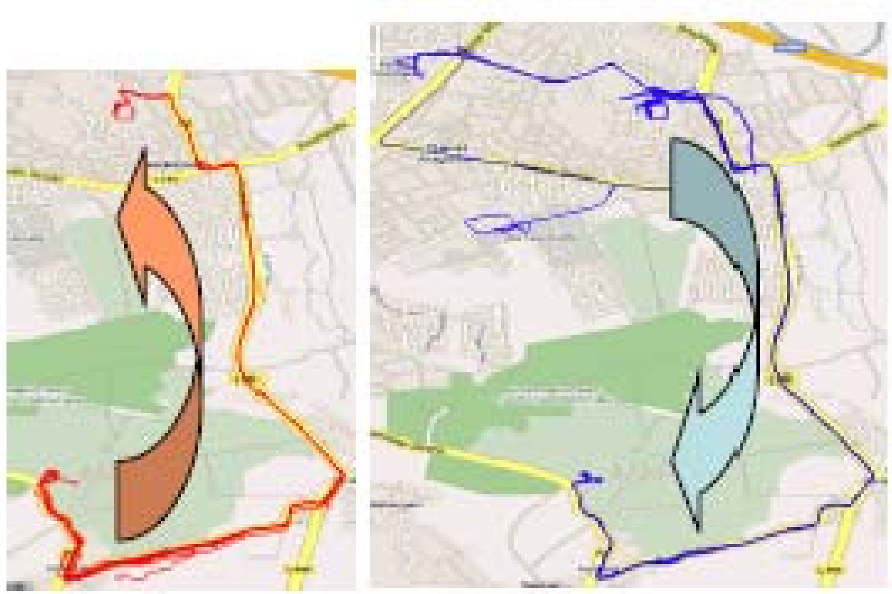
More ambitiously: mobility patterns

From mobility data to mobility patterns

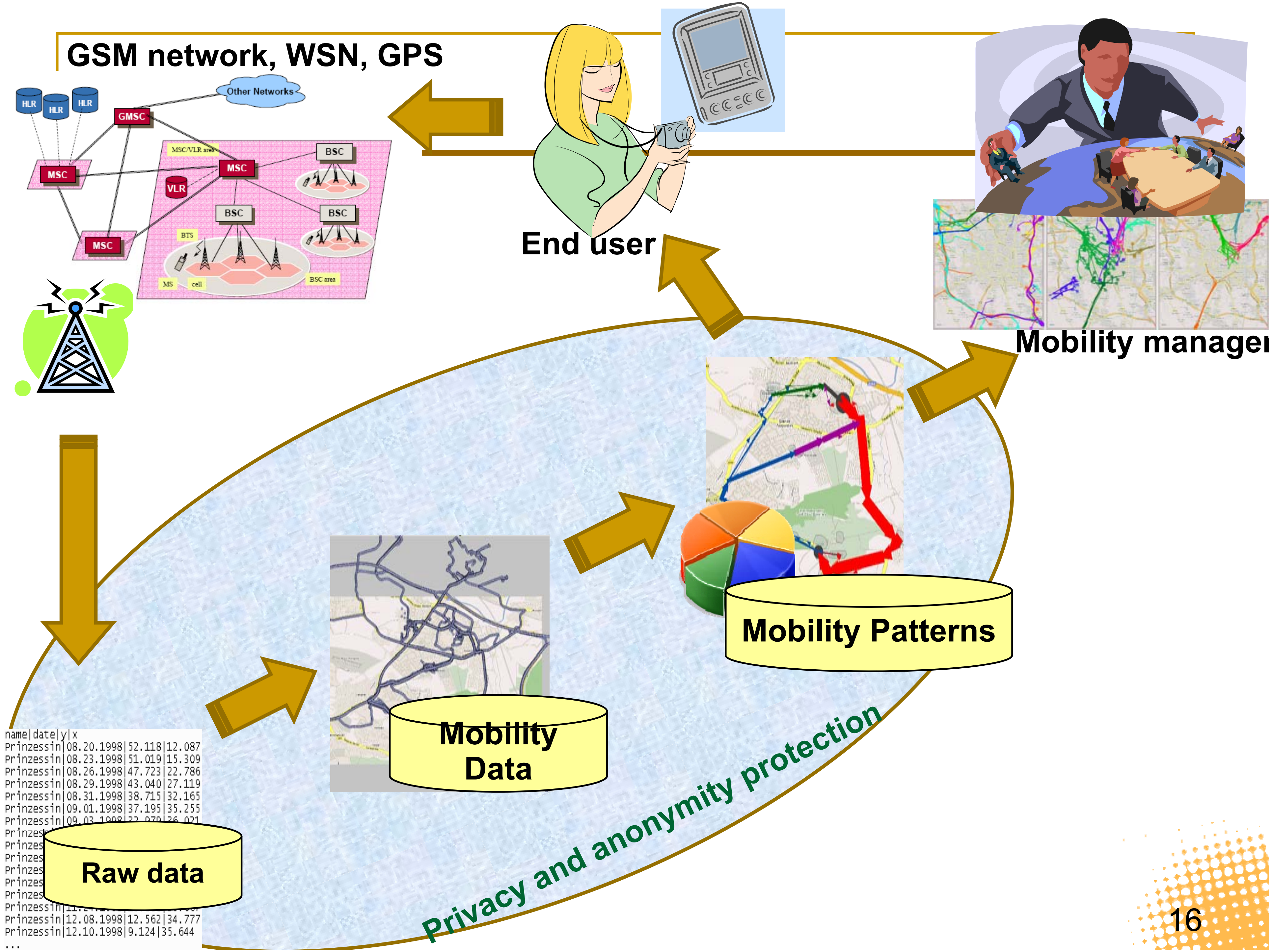
```
name|date|y|x
Prinzessin|08.20.1998|52.118|12.087
Prinzessin|08.23.1998|51.019|15.309
Prinzessin|08.26.1998|47.723|22.786
Prinzessin|08.29.1998|43.040|27.119
Prinzessin|08.31.1998|38.715|32.165
Prinzessin|09.01.1998|37.195|35.255
Prinzessin|09.03.1998|32.979|36.021
Prinzessin|09.05.1998|28.513|33.437
Prinzessin|09.06.1998|23.961|32.937
Prinzessin|09.07.1998|19.418|33.446
Prinzessin|09.12.1998|15.823|34.094
Prinzessin|10.11.1998|14.685|32.848
Prinzessin|11.03.1998|11.510|32.591
Prinzessin|11.24.1998|13.888|35.667
Prinzessin|12.08.1998|12.562|34.777
Prinzessin|12.10.1998|9.124|35.644
...
```



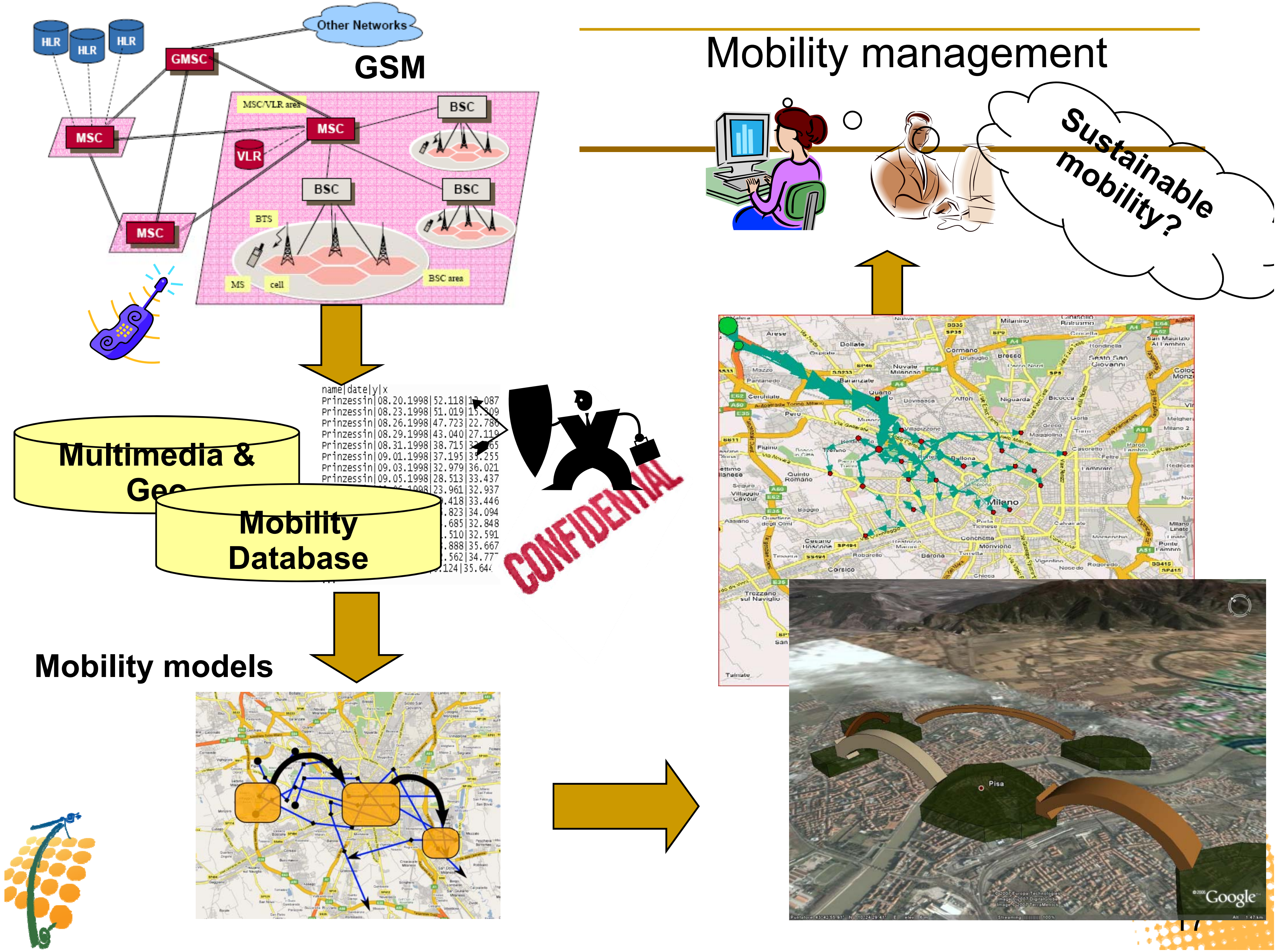
From mobility data to mobility patterns

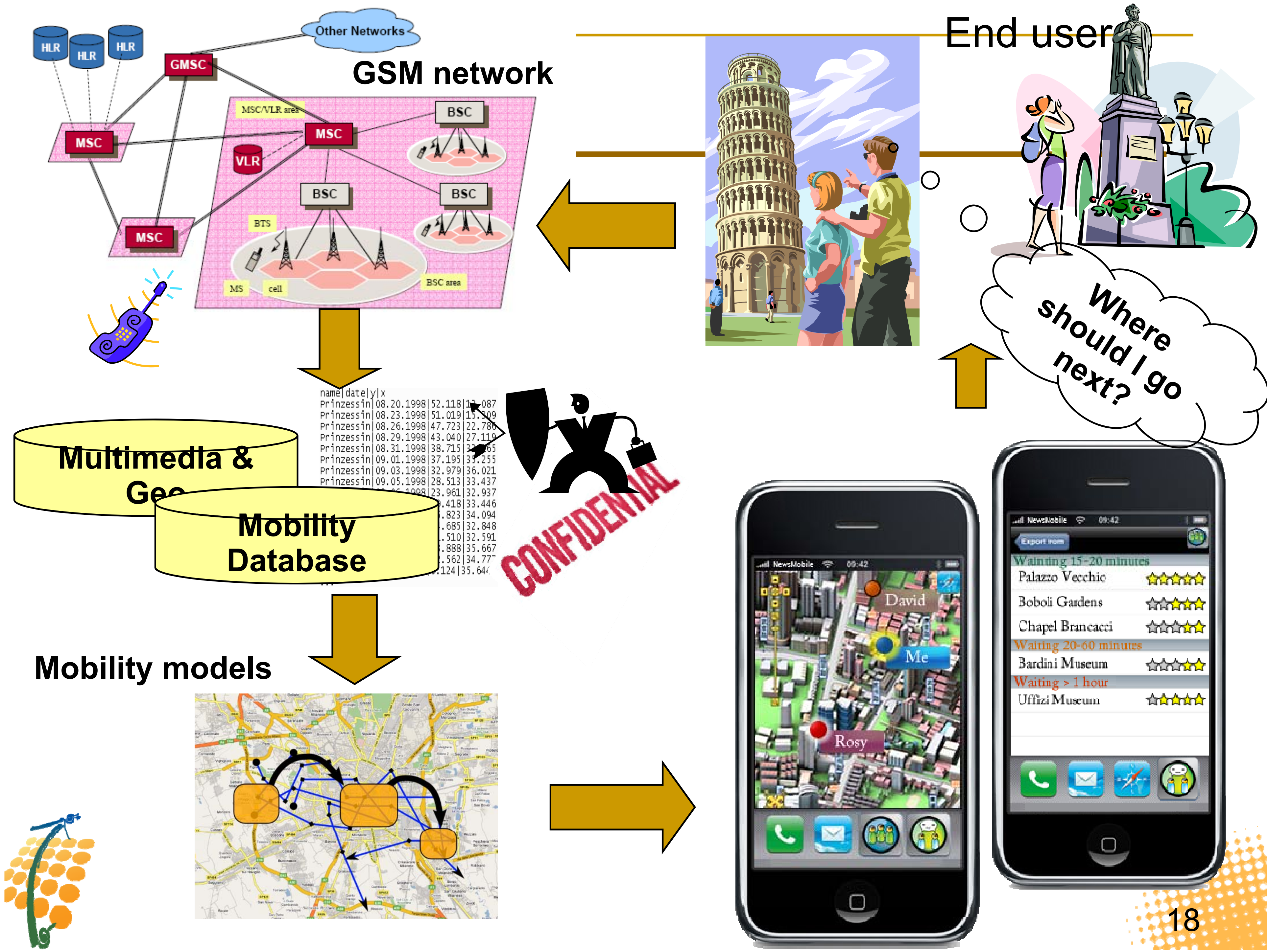


Mobility data mining and the Geographic Knowledge Discovery process


GSM network, WSN, GPS
HLR
GMSC
Other Networks
MSC
MSC/VLR area
BSC
VLR
BTS
MS
cell
BSC area
End user
Mobility manage
Mobility Patterns
Mobility Data
Raw data
name|date|y|x
Prinzessin|08.20.1998|52.118|12.087
Prinzessin|08.23.1998|51.019|15.309
Prinzessin|08.26.1998|47.723|22.786
Prinzessin|08.29.1998|43.040|27.119
Prinzessin|08.31.1998|38.715|32.165
Prinzessin|09.01.1998|37.195|35.255
Prinzessin|12.08.1998|12.562|34.777
Prinzessin|12.10.1998|9.124|35.644
...
Privacy and anonymity protection

Mobility management




GSM network
Other Networks
HLR
HLR
HLR
GMSC
MSC/VLR area
MSC
MSC
MSC
VLR
BSC
BSC
BSC
BTS
MS
cell
BSC area
End user
Where should I go next?
name|date|y|x
Prinzessin|08.20.1998|52.118
Prinzessin|08.23.1998|51.019
Prinzessin|08.26.1998|47.723|22.786
Prinzessin|08.29.1998|43.040|27.119
Prinzessin|08.31.1998|38.715
Prinzessin|09.01.1998|37.195
Prinzessin|09.03.1998|32.979|36.021
Prinzessin|09.05.1998|28.513|33.437
CONFIDENTIAL
Multimedia & Geo
Mobility Database
Mobility models
David
Me
Rosy
Waiting 15-20 minutes
Palazzo Vecchio
Boboli Gardens
Chapel Brancacci
Waiting 20-60 minutes
Bardini Museum
Waiting > 1 hour
Uffizi Museum

Key questions

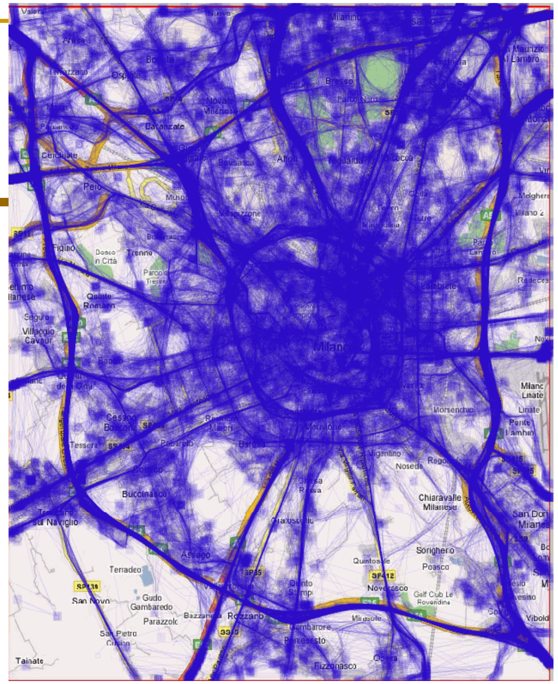
- How to reconstruct a trajectory from raw logs, how to store and query trajectory data?
- How to classify trajectories according to means of transportation (pedestrian, private vehicle, public transportation vehicle, …)?
- Which spatio-temporal pattern and /models are useful abstractions of mobility data?
  - How to compute such patterns and models efficiently?
- Privacy protection and anonymity – how to make such concepts formally precise and measurable?
  - How to find an optimal trade-off between privacy protection and quality of the analysis?

A guided tour on mobility data mining technologies

- Trajectory databases
- Trajectory warehouses and OLAP
- Mobility data mining
- Privacy-preserving mobility data mining
- Visual analytics for mobility data

Acquiring, Storing and Querying trajectories

Data: typical structure and typical size

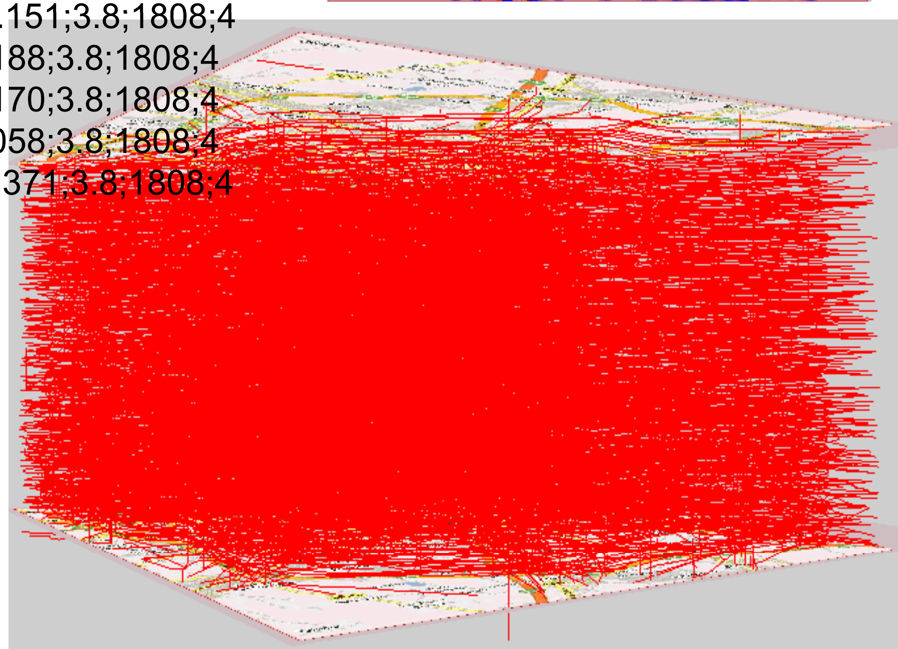


N;Time;Lat;Long;Height;Course;Speed;PDOP;State;NSat

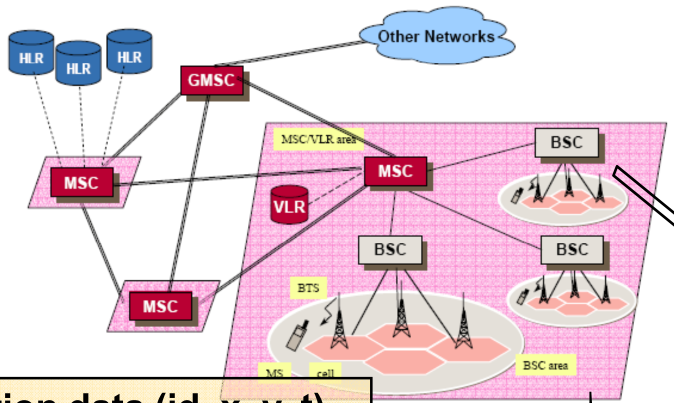
…

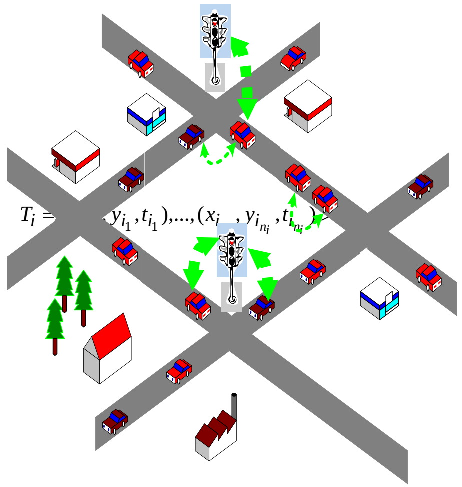
8;22/03/07 08:51:52;50.777132;7.205580; 67.6;345.4;21.817;3.8;1808;4
9;22/03/07 08:51:56;50.777352;7.205435; 68.4;35.6;14.223;3.8;1808;4
10;22/03/07 08:51:59;50.777415;7.205543; 68.3;112.7;25.298;3.8;1808;4
11;22/03/07 08:52:03;50.777317;7.205877; 68.8;119.8;32.447;3.8;1808;4
12;22/03/07 08:52:06;50.777185;7.206202; 68.1;124.1;30.058;3.8;1808;4
13;22/03/07 08:52:09;50.777057;7.206522; 67.9;117.7;34.003;3.8;1808;4
14;22/03/07 08:52:12;50.776925;7.206858; 66.9;117.5;37.151;3.8;1808;4
15;22/03/07 08:52:15;50.776813;7.207263; 67.0;99.2;39.188;3.8;1808;4
16;22/03/07 08:52:18;50.776780;7.207745; 68.8;90.6;41.170;3.8;1808;4
17;22/03/07 08:52:21;50.776803;7.208262; 71.1;82.0;35.058;3.8;1808;4
18;22/03/07 08:52:24;50.776832;7.208682; 68.6;117.1;11.371;3.8;1808;4

…



Location data producers:GSM, GPS, WiFI





Location data (id, x, y, t)
are generated

Trajectory stream manager +
Trajectory reconstruction

$$T_i = < (x_{i_1}, y_{i_1}, t_{i_1}), \ldots, (x_{i_{n_i}}, y_{i_{n_i}}, t_{i_{n_i}}) >$$

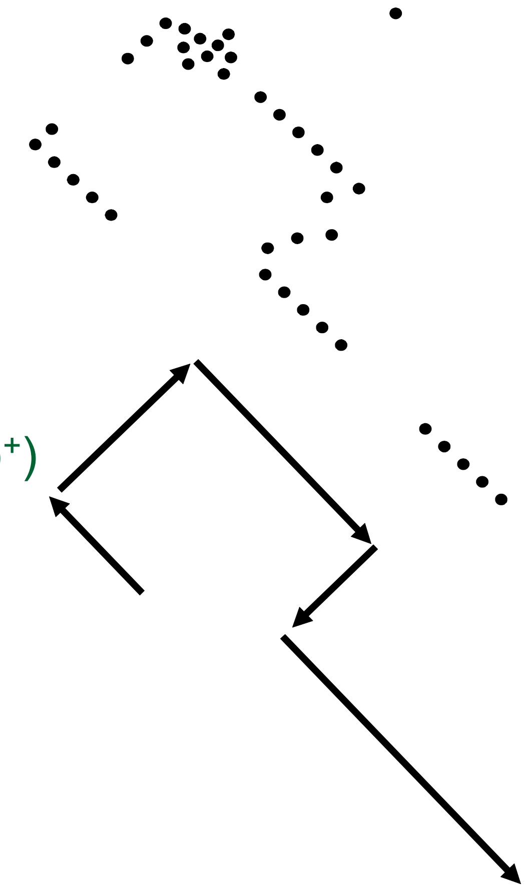
trajectory data
(obj-id, traj-id, (x, y, t)*)
are reconstructed

Moving
Object
Database

The trajectory reconstruction problem

- From raw location data (obj-id, x, y, t)

a sample of a user's movement (GPS recordings)

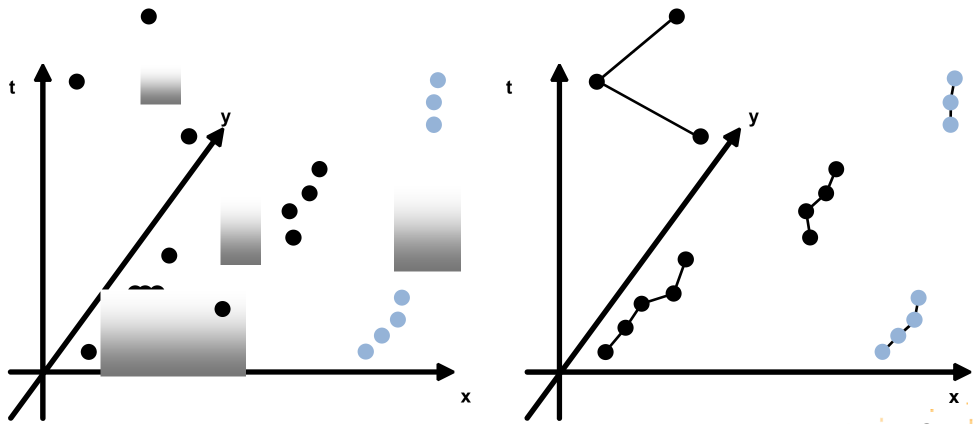


- To trajectory data (obj-id, traj-id, $(x, y, t)^+$)

a sample of reconstructed trajectories

Reconstructing trajectories

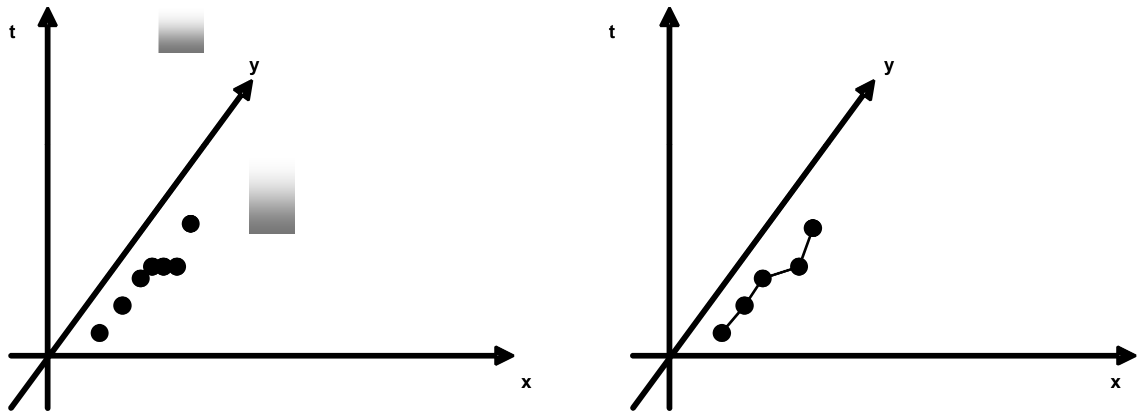
- Collected raw data represent time-stamped geographical locations
  - Raw points arrive in bulk sets
  - We need a filter that decides if the new series of data is to be appended to an existing trajectory or not:
    - Tolerance distance
    - Temporal gap
    - Spatial gap
    - Maximum speed
    - Maximum noise duration



25

Reconstructing trajectories: parameters

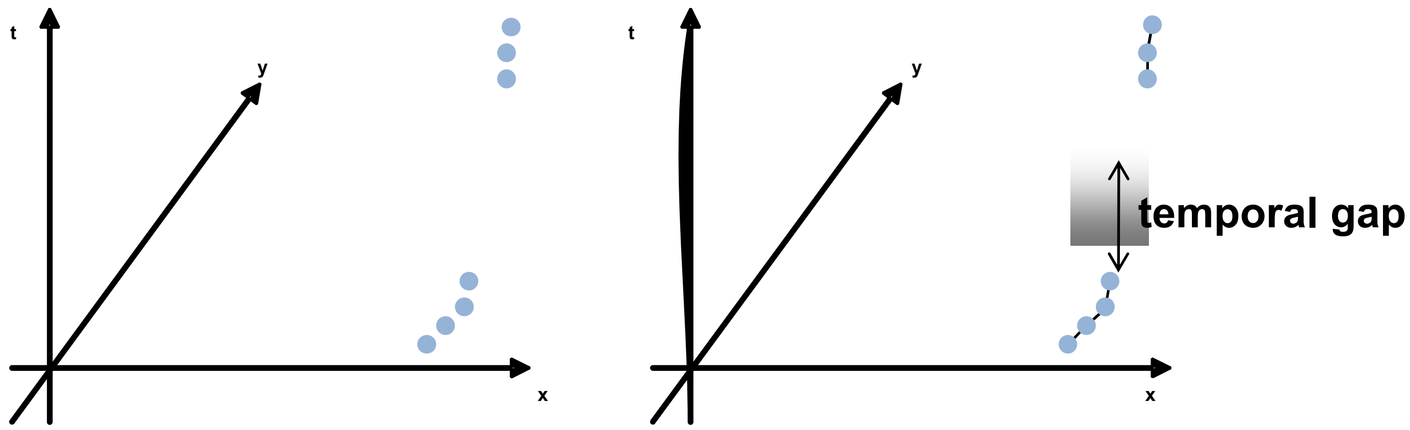
- Tolerance distance
  - The tolerance of the transmitted time-stamped positions. In other words, it is the **maximum distance between two consecutive time-stamped positions** of the same object in order for the object to be **considered** as **stationary**



26

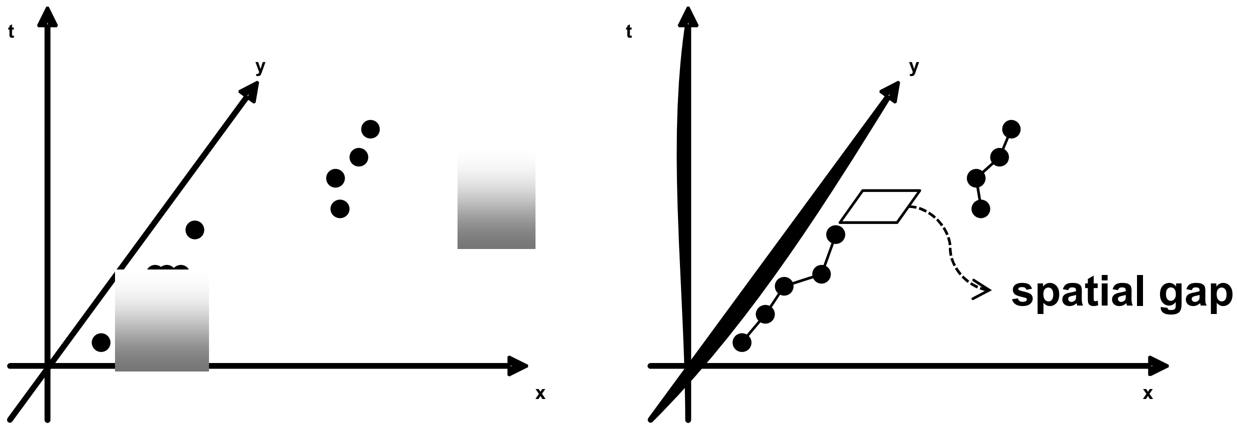
Reconstructing trajectories: parameters

- Tolerance distance
- Temporal gap between trajectories
  - The **maximum allowed time interval** between two consecutive time-stamped positions of the same trajectory for a single moving object



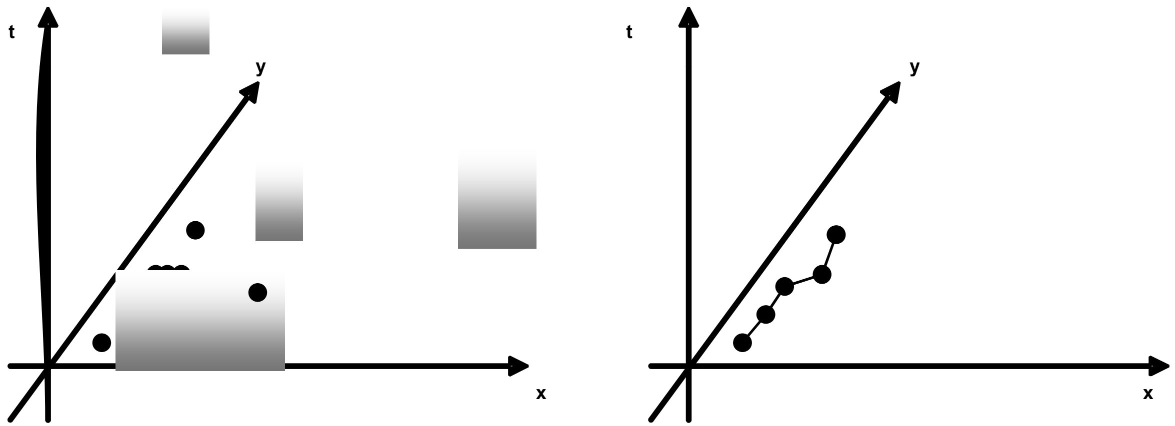
Reconstructing trajectories: parameters

- Tolerance distance
- Temporal gap between trajectories
- Spatial gap between trajectories
  - The **maximum allowed distance** in 2D plane between two consecutive time-stamped positions of the same trajectory



Reconstructing trajectories: parameters

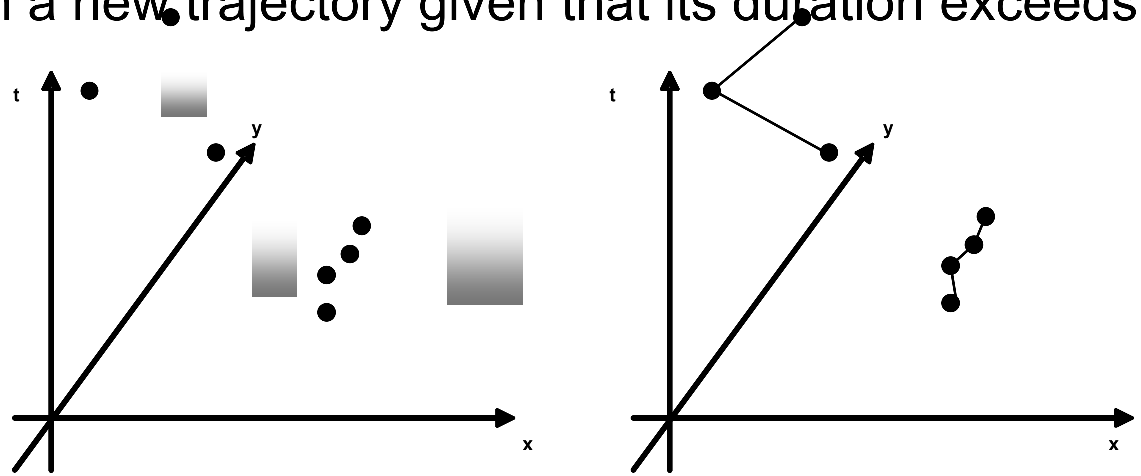
- Tolerance distance
- Temporal gap between trajectories
- Spatial gap between trajectories
- Maximum speed
  - It is used in order to determine whether a reported time-stamped position must be considered as **noise** and consequently discarded from the output trajectory



Reconstructing trajectories: parameters

- Tolerance distance
- Temporal gap between trajectories
- Spatial gap between trajectories
- Maximum speed

- Maximum noise duration
  - The **maximum duration of a noisy part** of a trajectory. Any sequence of noisy time-stamped positions of the same object will result in a new trajectory given that its duration exceeds *noise*$_{max}$

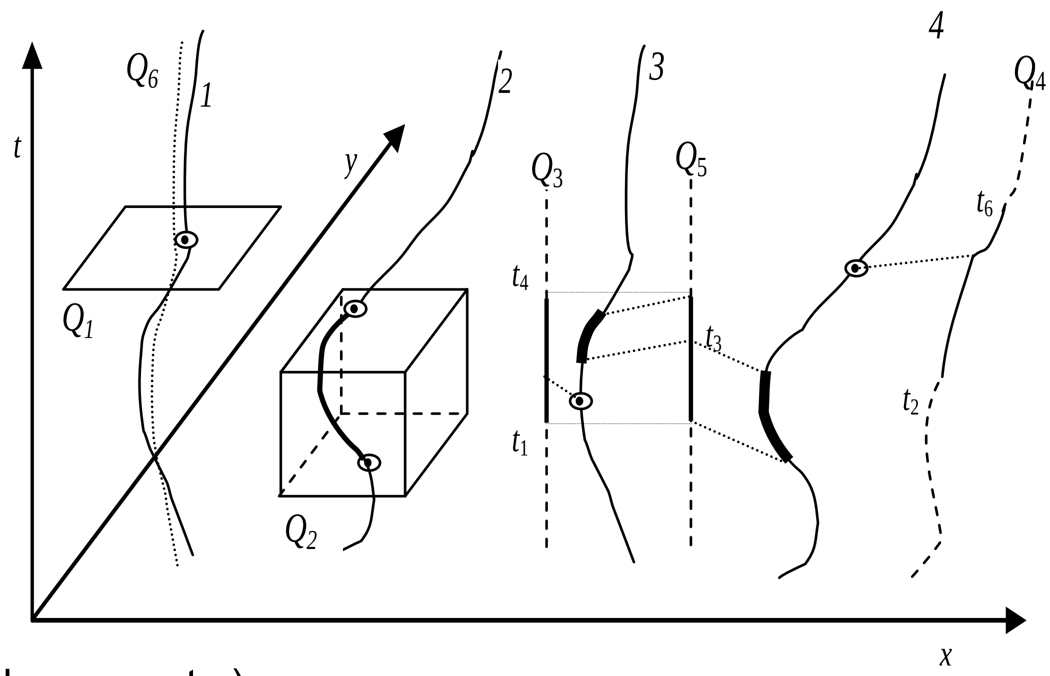


Moving Objects Databases

- The traditional database technology has been extended into **Moving Object Databases** (MODs) that handle modeling, indexing and query processing issues for trajectories
- Spatial and temporal dimensions are considered as first-class citizens.
- Both past and current (as well as anticipated future) positions of moving objects are of interest.
  - SECONDO: Ralf Hartmut Guting, et. al. SECONDO: An Extensible DBMS Platform for Research Prototyping and Teaching. In Proceeding of the International Conference on Data Engineering, ICDE, pages 1115{1116, Tokyo, Japan, April 2005.
  - PLACE: Mohamed F. Mokbel, et al. PLACE: A Query Processor for Handling Real-time Spatio-temporal Data Streams (Demo). In Proceeding of the International Conference on Very Large Data Bases, VLDB, pages 1377{1380, Toronto, Canada, August 2004.
  - DOMINO: Ouri Wolfson, et al.. Management of Dynamic Location Information in DOMINO (Demo). In Proceeding of the International Conference on Extending Database Technology, EDBT, pages 769{771, Prague, Czech Republic, March 2002.
  - Location-aware Query Processing and Optimization: A Tutorial by Mohamed F. Mokbel, MDM07

Querying the Moving Object Database

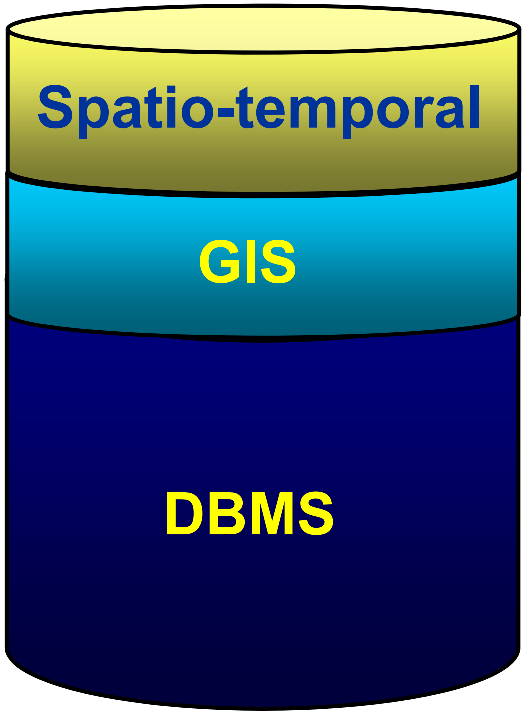
- Traditional spatial search
  - Range / distance-based / NN queries
- Trajectory-sub-sequence search
  - Spatial / temporal intersections of trajectories
- Topological / directional search



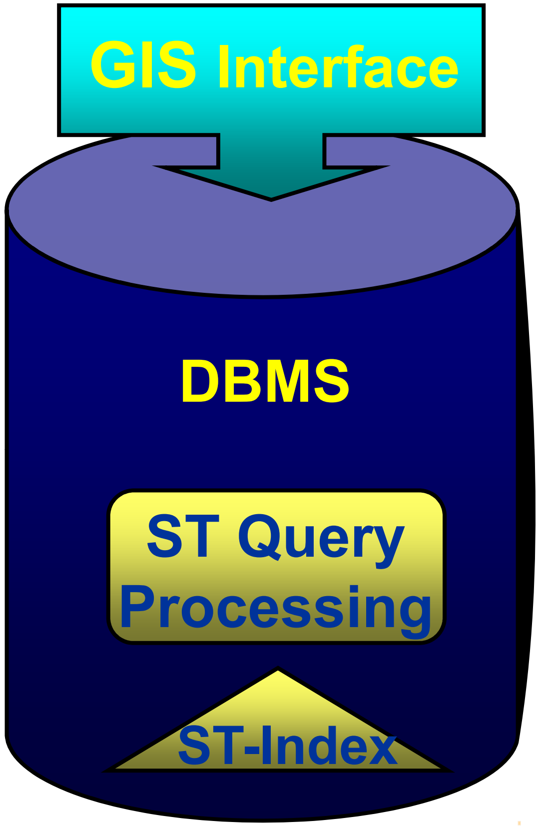
  - enter (cross, leave, bypass, etc.) an area
  - located west (south, etc.) of a (static) area
  - located left of (right of, in front of, etc.) a (moving) object

Location-based Database Servers

Layered Approach



Built-in Approach



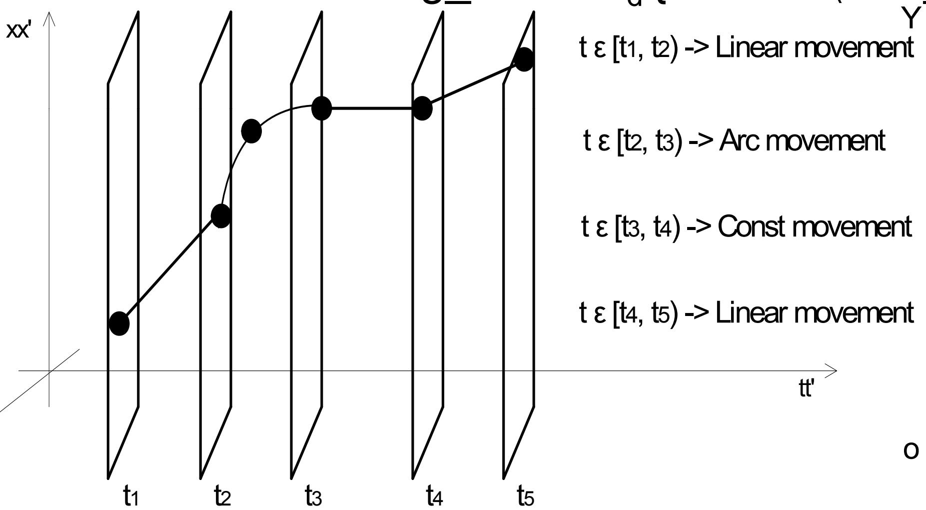
HERMES: A Database Engine for Moving Objects

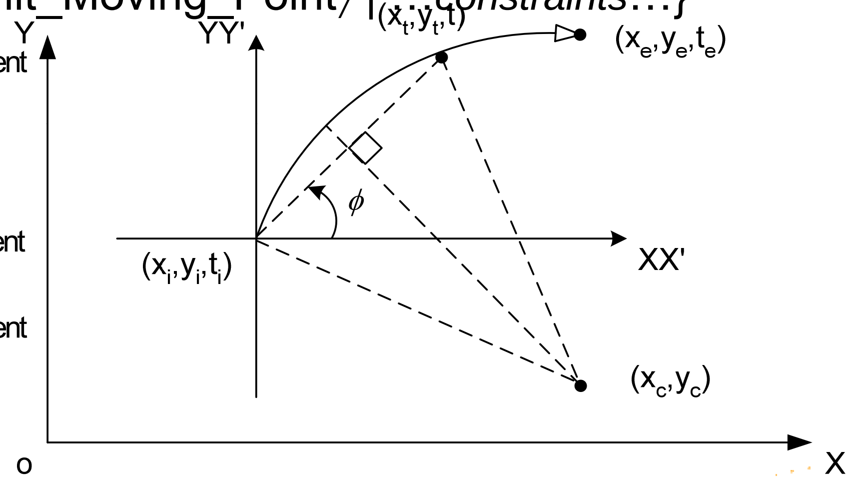
- **Built on top of ORACLE 10**
- **Data model: absolute vs. relative location coordinates**
  - **Current location as a function in time over the starting location**
  - **linear and arc movement functions**
- **Trajectory management**
  - **Insert/Update/Delete a moving object or a segment of its trajectory**
  - **Functions over trajectories or sets of trajectories**
- **Data management**
  - **Supported indices: R-tree (for stationary data)**
  - **Development of a specialized index (TB-tree)**
- Nikos Pelekis, Yannis Theodoridis: **Boosting location-based services with a moving object database engine**. MobiDE 2006: 3-10
- Nikos Pelekis, Yannis Theodoridis, Spyros Vosinakis, Themis Panayiotopoulos: **Hermes - A Framework for Location-Based Data Management**. EDBT 2006: 1130-1134

34

Hermes: trajectory data type

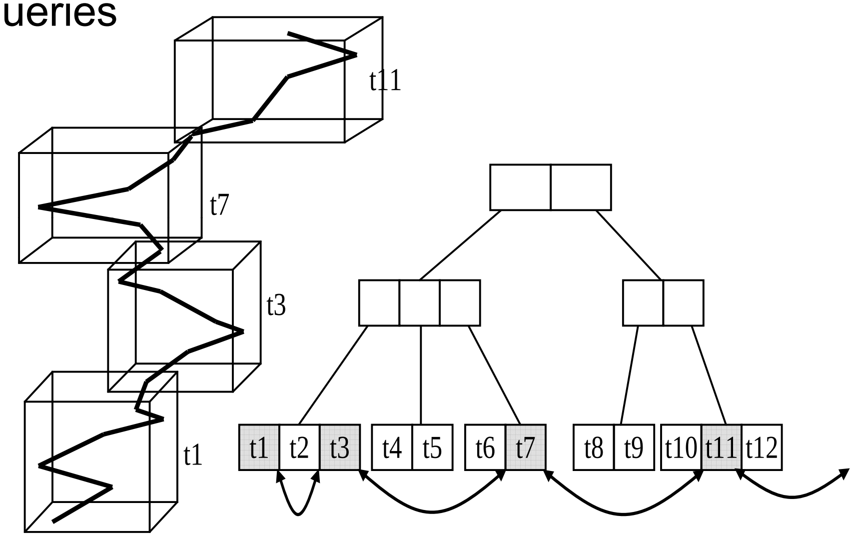
- Primitive definition:
  - Unit_Function = $_d$ ⟨ x_i:double, y_i:double, x_e:double, y_e:double, x_c:double, y_c:double, v:double, a:double, flag:TypeOfFunction ⟩ , where
  - TypeOfFunction={ CONST, PLNML_1, ARC_<1..8> }
  - Unit_Moving_Point = $_d$ ⟨ p: Period⟨SEC⟩, m: Unit_Function⟩
  - Moving_Point = $_d$ { tab: set⟨Unit_Moving_Point⟩ |*…constraints…*}





TB-Tree support in Hermes MOD engine

- TB-Tree Index
  - Maintains the ‘trajectory’ concept
    - Each node consists of segments of a single trajectory
    - Nodes are linked together in a chain
  - Effective for trajectory-oriented queries
  - Implemented in Hermes using Oracle’s indexing extensibility



HERMES includes

- Spatial entities:
  - Road Network Data (Nodes, Links)
  - Landmarks (ID, geometry, address, area, type)
  - Regions (ID, name, geometry)
- “Moving” entities:
  - Vehicles (object_id, traj_id, route)

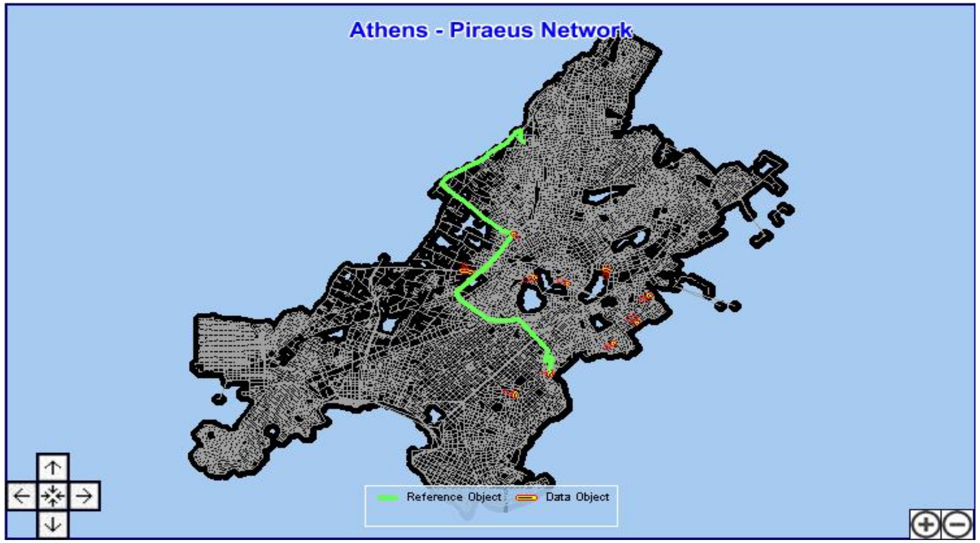
Query Operations

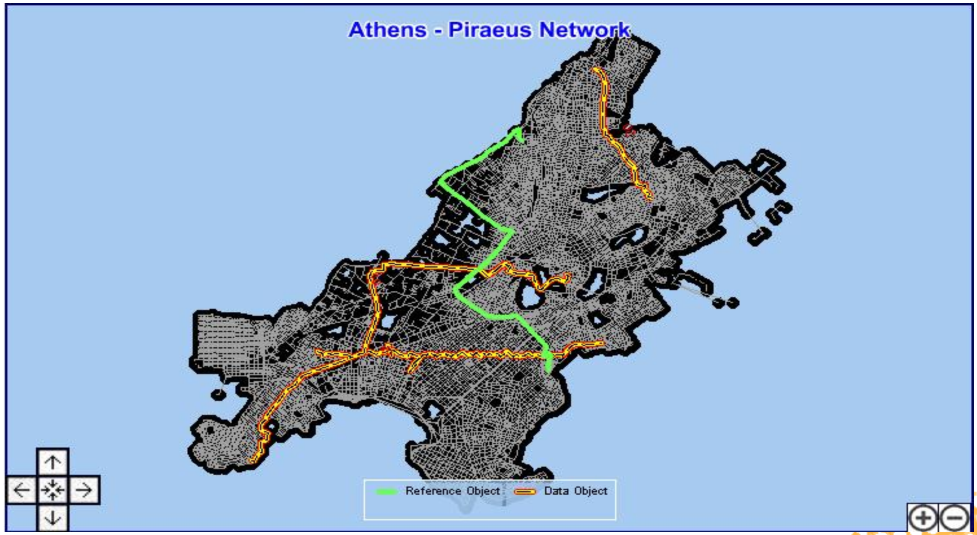
- Entities involved in a query
  - Reference Object: the type (trajectory or spatial entity) of the object based on which query answers are retrieved
  - Data Object: the type (trajectory or spatial entity) of the objects participating in the posed query answer

- Query classification
  - Moving Point – Moving Point
  - Moving Point – Static Spatial
  - Static Spatial – Moving Point

Moving Point – Moving Point

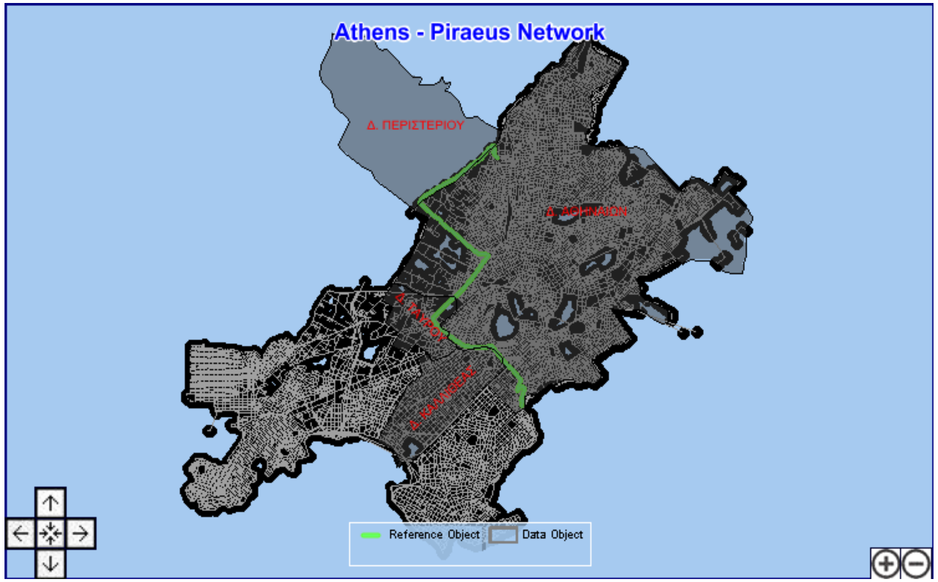
- Nearest Neighbor queries
  - Given a trajectory T, find the K nearest (during T's lifetime) parts of other trajectories
- Similarity queries
  - Spatial similarity
  - Spatiotemporal similarity
  - Speed-pattern similarity
  - Direction-pattern similarity

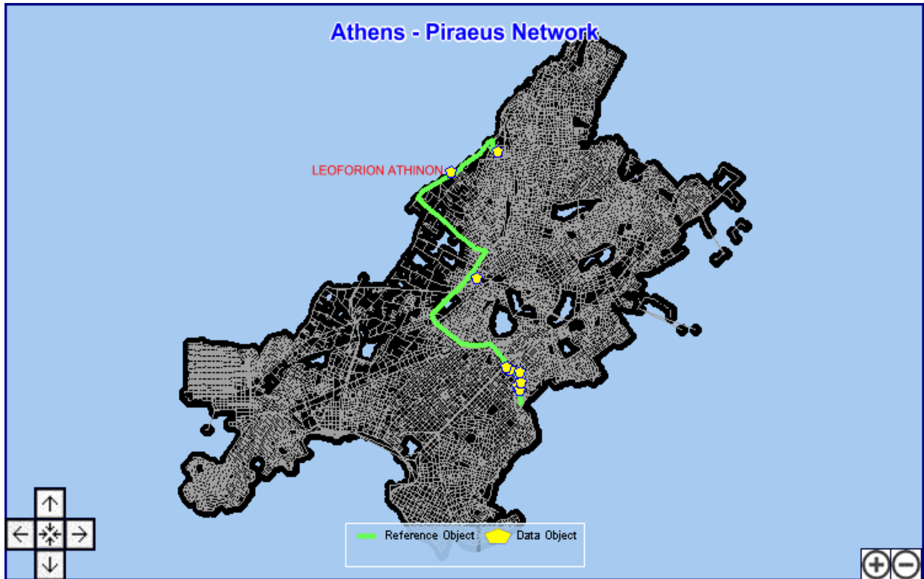




Moving Point – Static Spatial

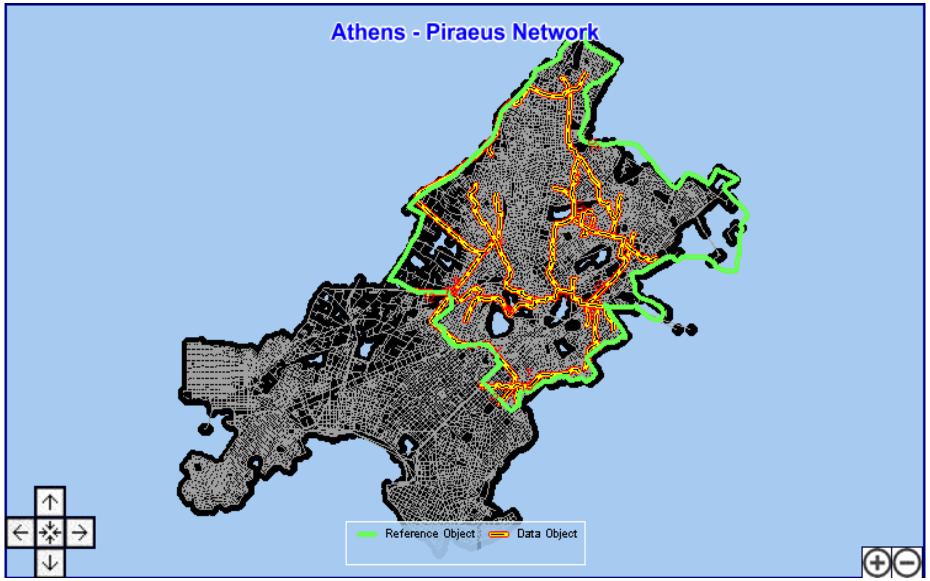
- Point query
  - Find the regions that intersect with a given trajectory
- Topological query
  - Find the regions that contain, overlap by intersect, overlap by disjoint etc with a given trajectory
- Nearest-Neighbor query
  - Find the K nearest landmarks (POIs) to a given trajectory

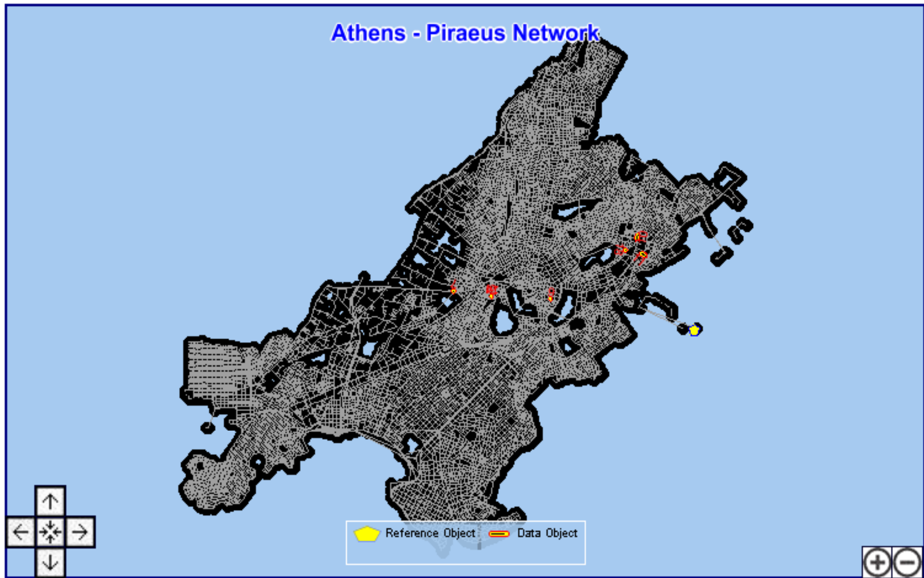




Static Spatial– Moving Point (1/2)

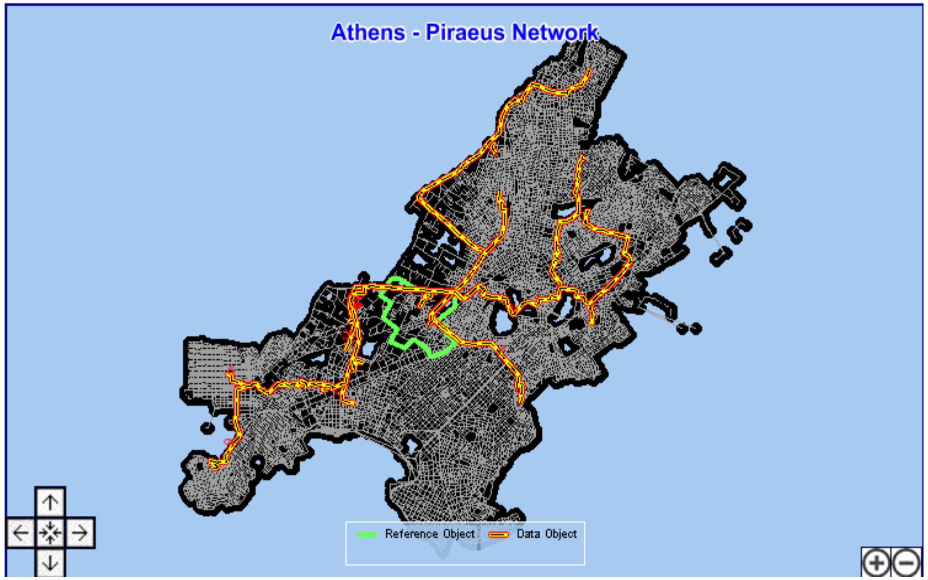
- Range query
  - Find trajectory parts fully contained in a given spatiotemporal window
- Nearest Neighbor query
  - Find the K nearest trajectory parts to a POI, within a given time period



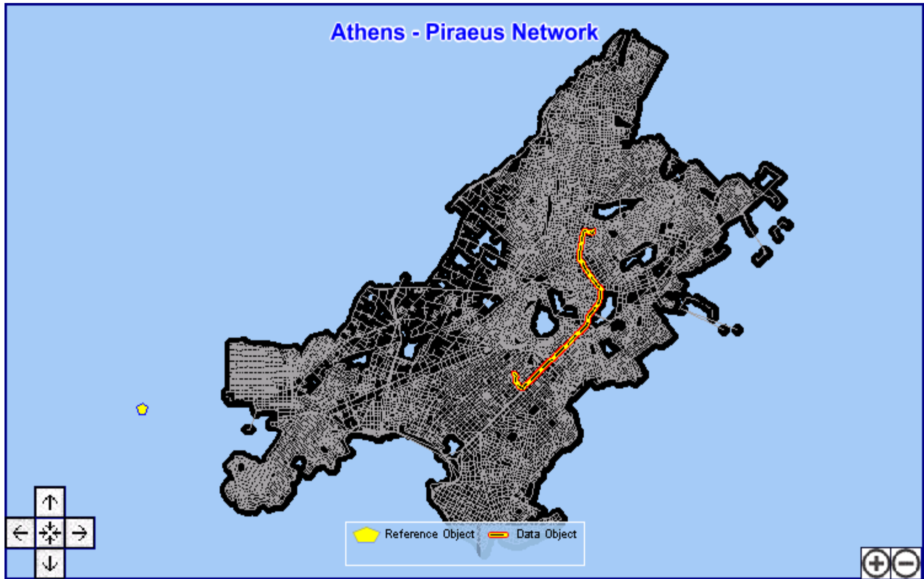


Static Spatial– Moving Point (2/2)

- Topological query
  - Find the trajectories that enter/leave an area within a given time period



- Directional query
  - Find trajectories whose location is east, west, north, south, left, right, front, behind of a POI

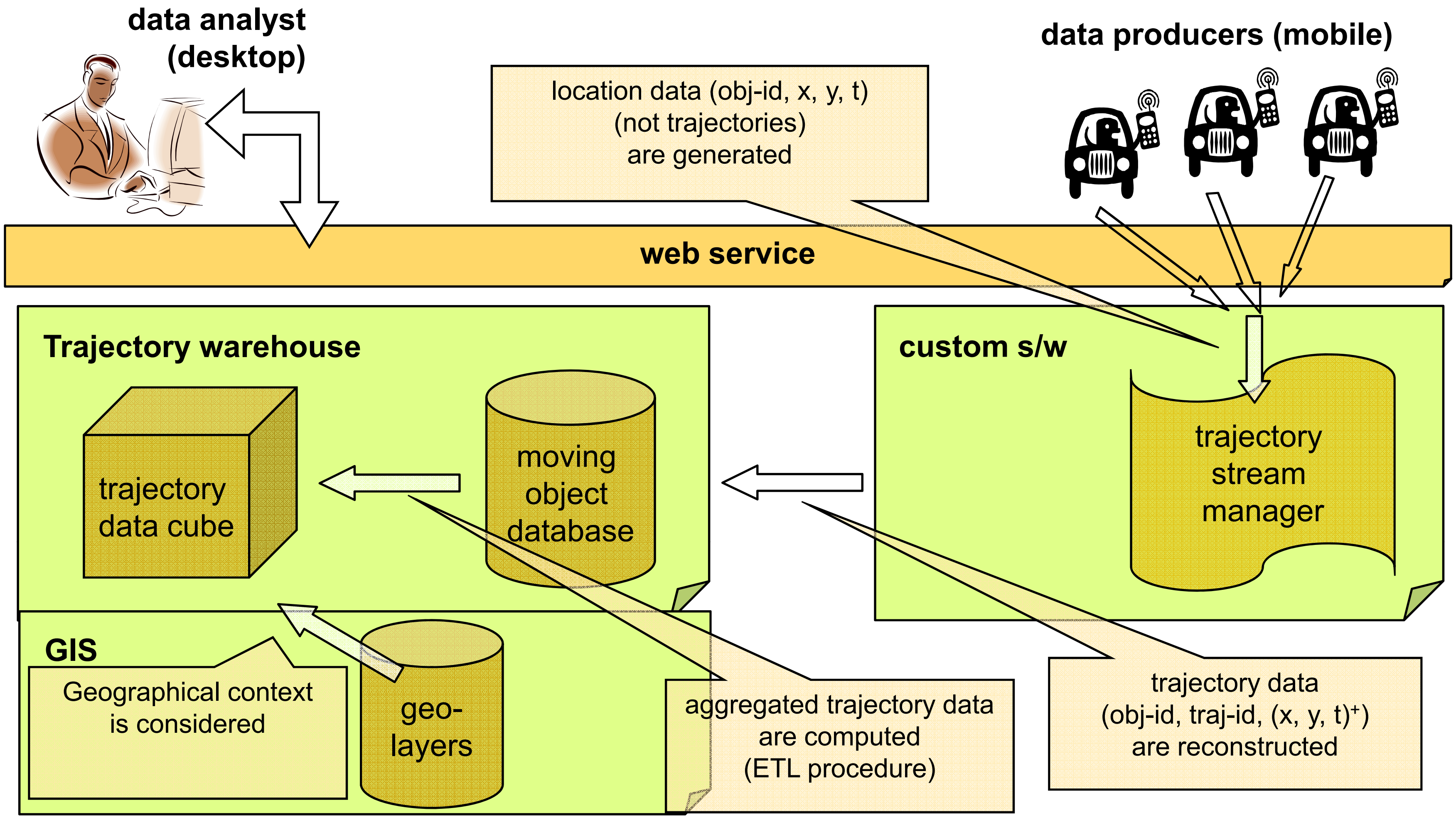


References

- Nikos Pelekis, Elias Frentzos, Nikos Giatrakos, Yannis Theodoridis: **HERMES: aggregative LBS via a trajectory DB engine**. SIGMOD Conference 2008: 1255-1258
- Elias Frentzos, Kostas Gratsias, Yannis Theodoridis: **Towards the Next Generation of Location-Based Services**. W2GIS 2007: 202-215
- Elias Frentzos, Kostas Gratsias, Nikos Pelekis, Yannis Theodoridis: **Algorithms for Nearest Neighbor Search on Moving Object Trajectories**. GeoInformatica 11(2): 159-193 (2007)
- Nikos Pelekis, Yannis Theodoridis: **Boosting location-based services with a moving object database engine**. MobiDE 2006: 3-10
- Nikos Pelekis, Yannis Theodoridis, Spyros Vosinakis, Themis Panayiotopoulos: **Hermes - A Framework for Location-Based Data Management**. EDBT 2006: 1130-1134
- Yannis Theodoridis: **Ten Benchmark Database Queries for Location-based Services**. Comput. J. 46(6): 713-725 (2003)

Trajectory Datawarehouse

A trajectory warehouse system architecture



Data warehouses (DW)

- Widely investigated for conventional, non-spatial data.
- Some research on spatial DW, pioneering work by Han et al. in 1998.
  - Spatial and non-spatial dimensions and measures.
  - OLAP operations in a spatial data cube.
- Recent research direction: developing spatio-temporal DW and supporting spatio-temporal OLAP operations in order to extract summarized spatio-temporal information.
  - Useful for: traffic supervision systems, transportation and supply chain managements, mobile e-commerce.
  - Focus on methods for an efficient implementation of spatio-temporal aggregate queries.

Trajectory data warehousing

- Trajectory data warehousing should
  - extract aggregate information from MOD
  - support a variety of dimensions (temporal, spatial, thematic, …) and measures (about space, time and their derivatives)
  - Storing measures associated with facts, concerning the set of trajs crossing the cell
    - ⇒ aggregate information in base cells
- Challenges
  - high volume and complex nature of data; special query processing requirements
- **Results so far**:
  - design of a trajectory-oriented data cube
  - extensions of traditional aggregation techniques to produce summary information for OLAP analysis

Basic definitions & schemas

- Trajectory $T_i = <(x_{i_1}, y_{i_1}, t_{i_1}), \ldots, (x_{i_{n_i}}, y_{i_{n_i}}, t_{i_{n_i}})>$

OBJECTS (<u>object-id</u>: *identifier*, description: *text*, gender: {M | F}, birth-date: *date*, profession: *text*, device-type: *text*)

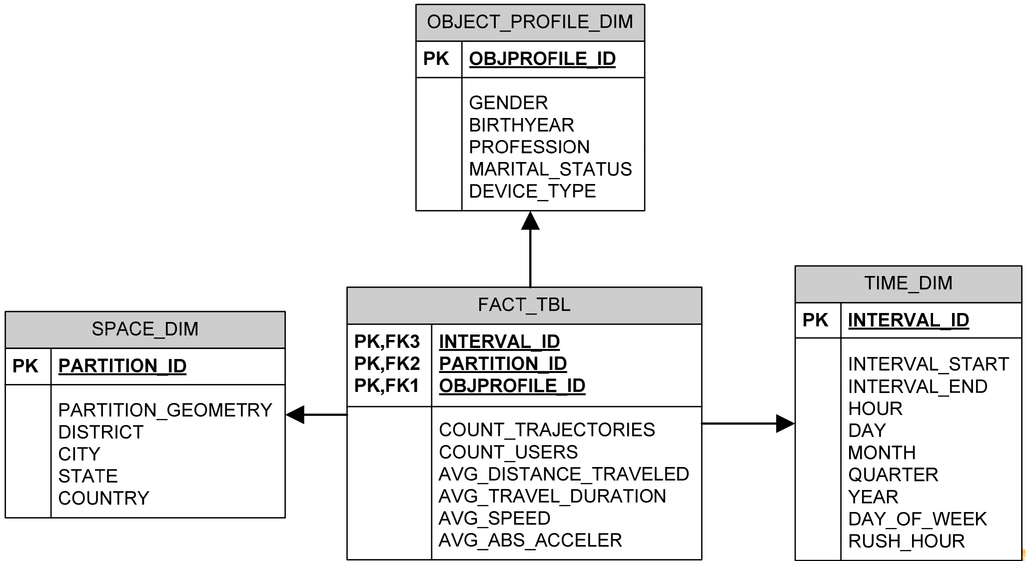
RAW_LOCATIONS (<u>object-id</u>: *identifier*, <u>timestamp</u>: *datetime*, eastings-x: *numeric*, northings-y: *numeric*, altitude-z: *numeric*)

MOD_TRAJECTORIES (<u>trajectory-id</u>: *identifier*, <u>object-id</u>: *identifier*, trajectory: *3D geometry*)

- Moving Object Database

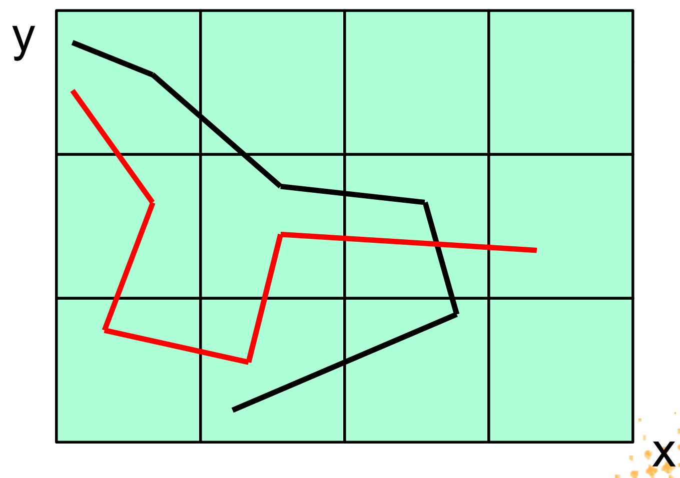
  $D = \{T_1, T_2, \ldots, T_N\}$

- Trajectory Data Warehouse
  - *Dimensions:* Spatial, Temporal, Object Profile
  - *Measures:* count (trajectories), count (users), avg (distance traveled), avg (travel duration), avg (speed), avg (abs (acceler))



ETL processing: loading

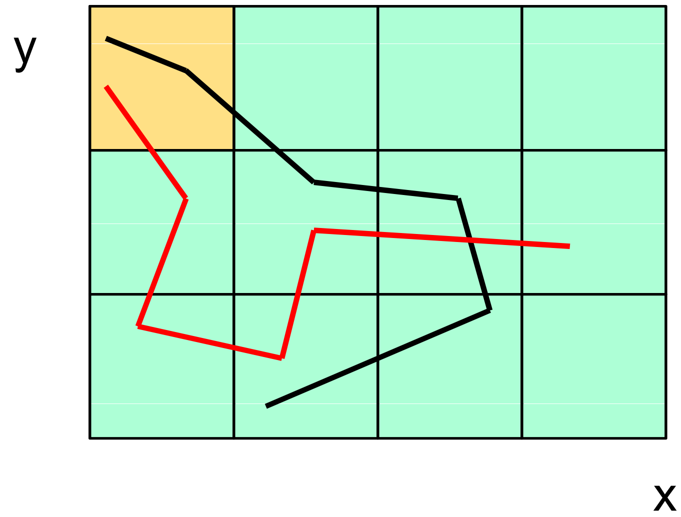
- Loading data into the dimension tables ➔ straightforward
- Loading data into the fact table ➔ complex
  - Fill in the measures with the appropriate numeric values
  - In order to calculate the measures, we have to extract the portions of the trajectories that fit into the base cells of the cube
    - We propose two alternative solutions to this problem:
      - cell-oriented
      - trajectory-oriented



ETL processing: algorithms

***Cell-oriented approach* (COA)**

- Search for the portions of trajectories that they reside inside a spatiotemporal cell
    - Perform a **spatiotemporal range query** that returns the portions of trajectories that satisfy the range constraints
    - This is efficiently supported by the **TB-tree** [VLDB'00]
- Decompose the trajectory portions with respect to the user profiles they belong to
- Compute measures for this cell
- Repeat for the next cells



COUNT_TRAJECTORIES = 2

COUNT_USERS = 2

…

ETL processing: algorithms

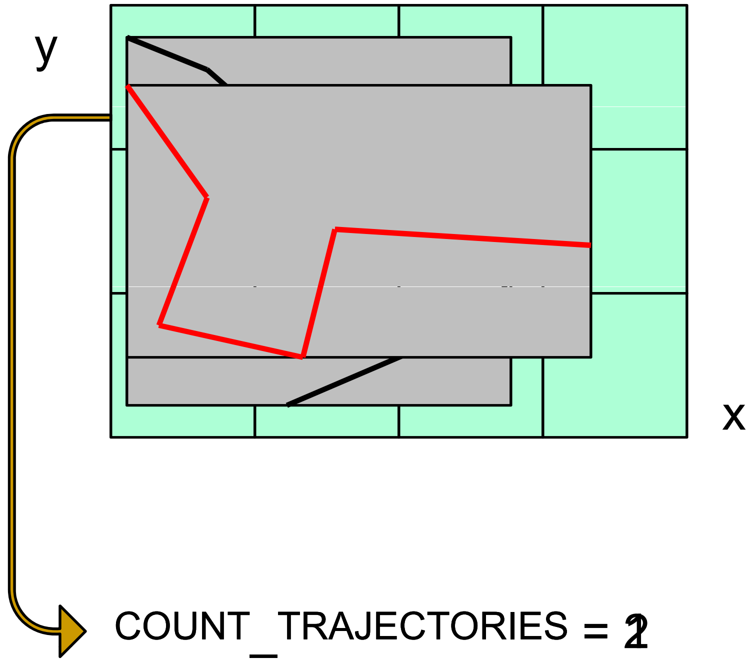
Trajectory-oriented approach **(TOA)**

- Discover the spatiotemporal cells where each trajectory resides in
  - In order to avoid checking all cells, use the trajectory MBR
- Identify the cells that overlap with the MBR and contain portions of the trajectory
- Compute measures for each cell

…

- Repeat for the next trajectories

…

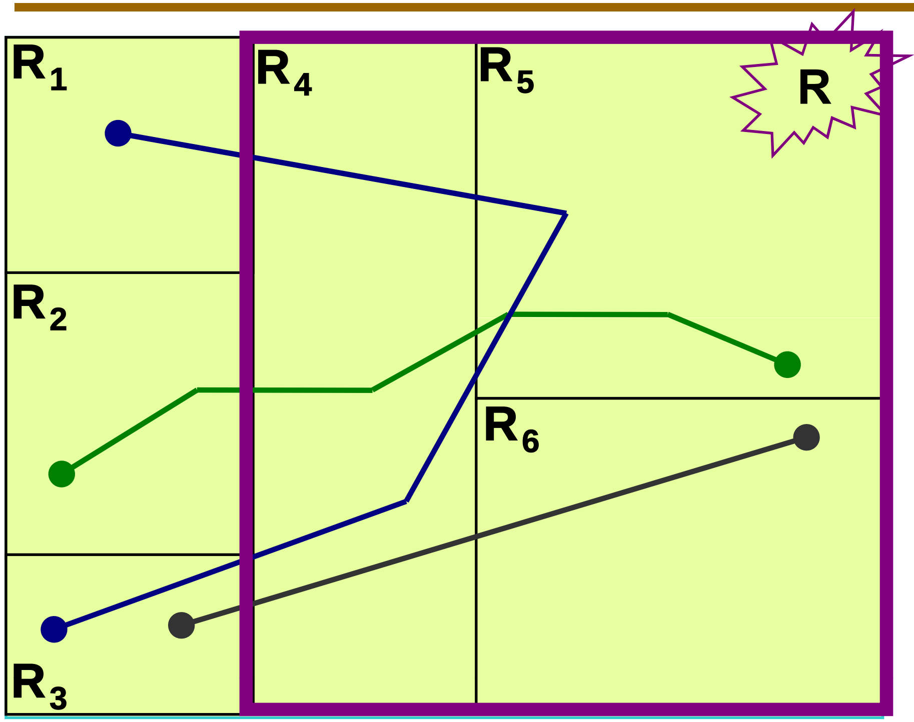


COUNT_TRAJECTORIES = 2
COUNT_USERS = 2
…

ETL processing: measures

Measure	Formula
COUNT_ TRAJECTORIES	count all distinct trajectory ids that pass through *base cell (bc)*
COUNT_USERS	count all the distinct object ids that pass through *bc*
AVG_DISTANCE_ TRAVELED	$AVG\_DISTANCE\_TRAVELED(bc) = \frac{SUM\_DISTANCE(bc)}{COUNT\_TRAJECTORIES(bc)}$ $SUM\_DISTANCE(bc) = \sum_{TP_i \in bc} len(TP_i)$
AVG_TRAVEL_ DURATION	$AVG\_TRAVEL\_DURATION(bc) = \frac{SUM\_DURATION(bc)}{COUNT\_TRAJECTORIES(bc)}$ $SUM\_DURATION(bc) = \sum_{TP_i \in bc} lifespan(TP_i)$
AVG_SPEED	$AVG\_SPEED(bc) = \frac{SUM\_SPEED(bc)}{COUNT\_TRAJECTORIES(bc)}$ $SUM\_SPEED(bc) = \sum_{TP_i \in bc} \frac{len(TP_i)}{lifespan(TP_i)}$
AVG_ABS_ ACCELER	$AVG\_ABS\_ACCELER(bc) = \frac{SUM\_ABS\_ACCELER(bc)}{COUNT\_TRAJECTORIES(bc)}$ $SUM\_ABS\_ACCELER(bc) = \sum_{TP_i \in bc} \frac{\lvert speed_{fin}(TP_i) - speed_{init}(TP_i) \rvert}{lifespan(TP_i)}$

Aggregating measures in the cube



At the lowest hierarchy level:

count of trajectories in R_4 = 3

count of trajectories in R_5 = 2

count of trajectories in R_6 = 1

Roll up in R

count of trajectories in R = 6
(according to traditional roll up)

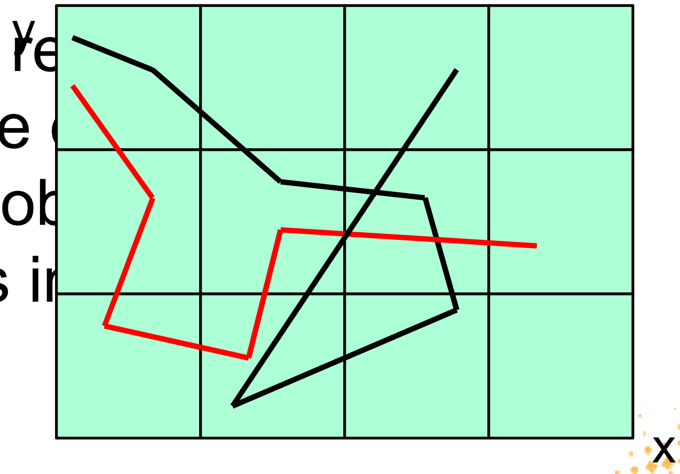
Correct answer: 3 (!!) due to the fact that the contents (trajectories) of the partitions are overlapping

How to compute the correct answer?

- A naïve solution is to query back the raw data.
- Can we do something better?

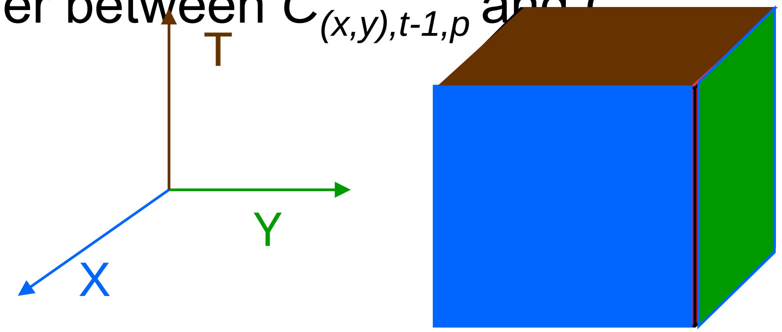
The distinct count problem: definition

- During the ETL process, measures can be computed in an accurate way by executing MOD queries
- Once the fact table has been fed, aggregate-only information is stored inside the TDW (no trajectory / user ids)
- When rolling up, COUNT_USERS, COUNT_TRAJECTORIES and, hence, all other measures defined over COUNT_TRAJECTORIES are subject to the distinct count problem [ICDE'04]:
  - if an object remains in the query re
    for several timestamps during the
    interval, instead of counting this ob
    once, it is counted multiple times i
    result



The distinct count problem: solution (1/3)

- We store in the base cells ($C_{(x,y),t,p}$) a tuple of auxiliary measures that help us correct the errors due to the duplicates when rolling-up:
  - $C_{(x,y),t,p}.Traj$: number of distinct trajectories of profile p intersecting the cell
  - $C_{(x,y),t,p}.cross\text{-}x$: number of distinct trajectories of profile p crossing the *spatial* border between $C_{(x-1,y),t,p}$ and $C_{(x,y),t,p}$
  - $C_{(x,y),t,p}.cross\text{-}y$: number of distinct trajectories of profile p crossing the *spatial* border between $C_{(x,y-1),t,p}$ and $C_{(x,y),t,p}$
  - $C_{(x,y),t,p}.cross\text{-}t$: number of distinct trajectories of profile p crossing the *temporal* border between $C_{(x,y),t-1,p}$ and C



Cell $\mathbf{C}_{(x,y),t,p}$

The distinct count problem: solution (2/3)

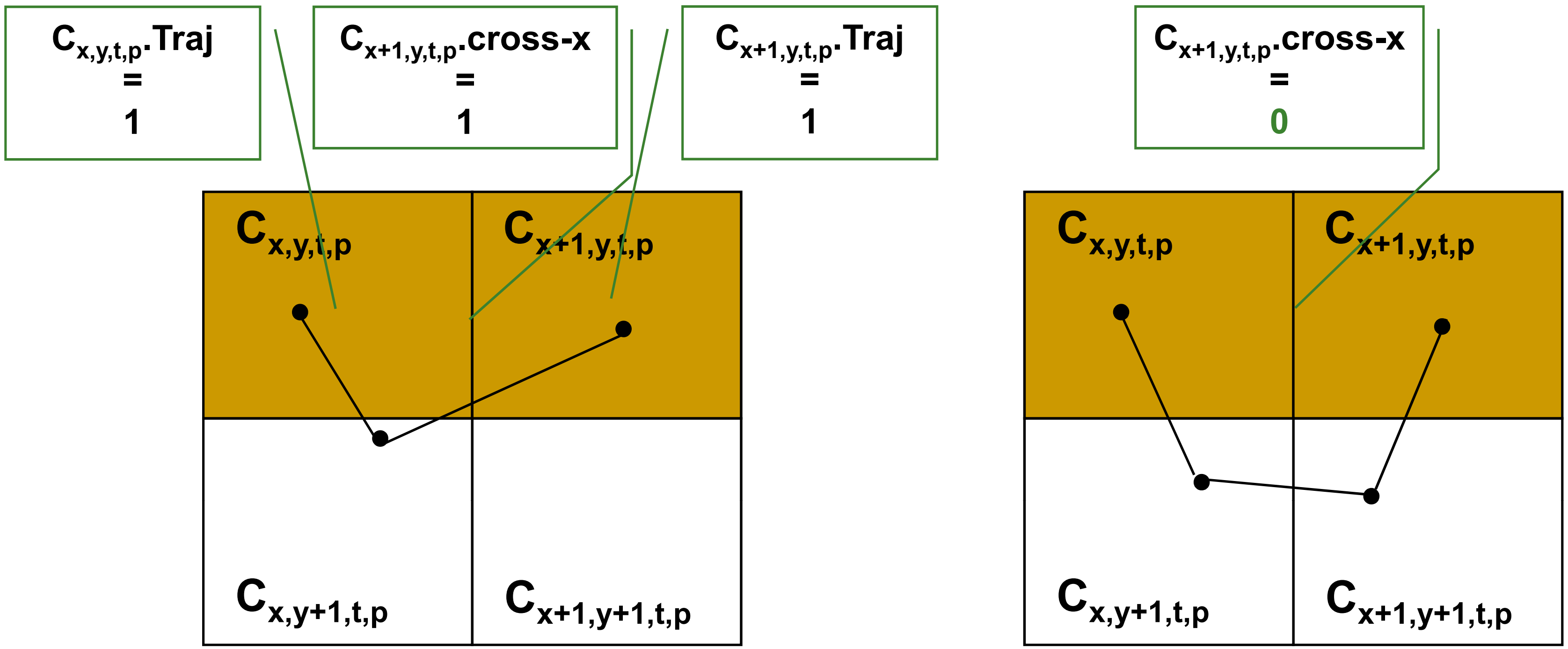
- Let $C_{(x',y'),t',p'}$ be a cell consisting of the union of two adjacent cells (i.e. $C_{(x,y),t.p} \cup C_{(x+1,y),t,p}$)
- In order to compute the **number of distinct trajectories**:

  $C_{(x',y'),t',p'}.Traj = C_{(x,y),t,p}.Traj + C_{(x+1,y),t,p}.Traj - C_{(x+1,y),t,p}.cross\text{-}x$

  - application of the well-known Inclusion/Exclusion principle for sets: $|A \cup B| = |A| + |B| - |A \cap B|$
  - **BUT** in some cases it holds that $C_{(x+1,y),t,p}.cross\text{-}x \neq |A \cap B|$ ☹
  - Example: fast and agile trajectories

The distinct count problem: solution (3/3)

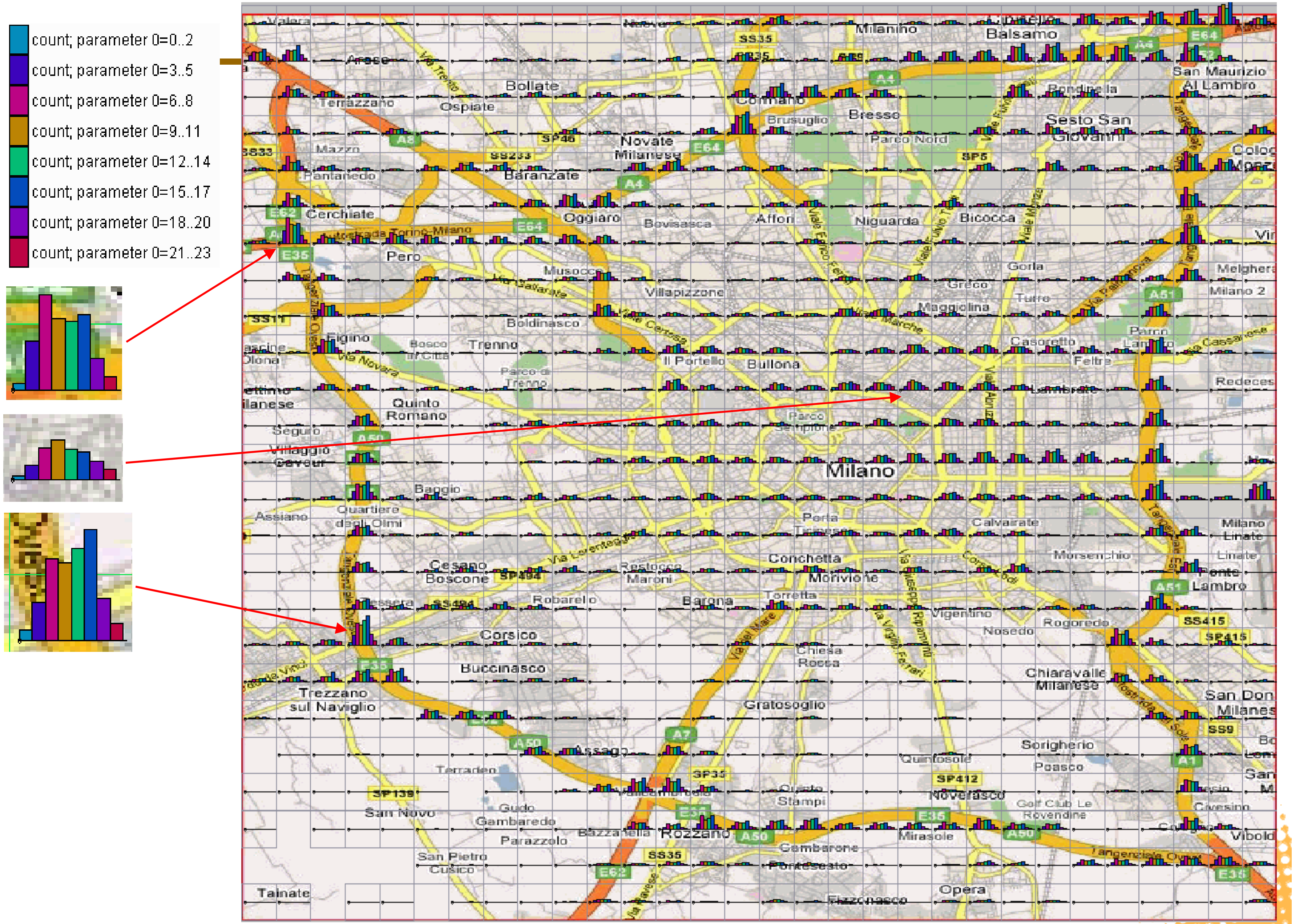
Compute the number of distinct trajectories:



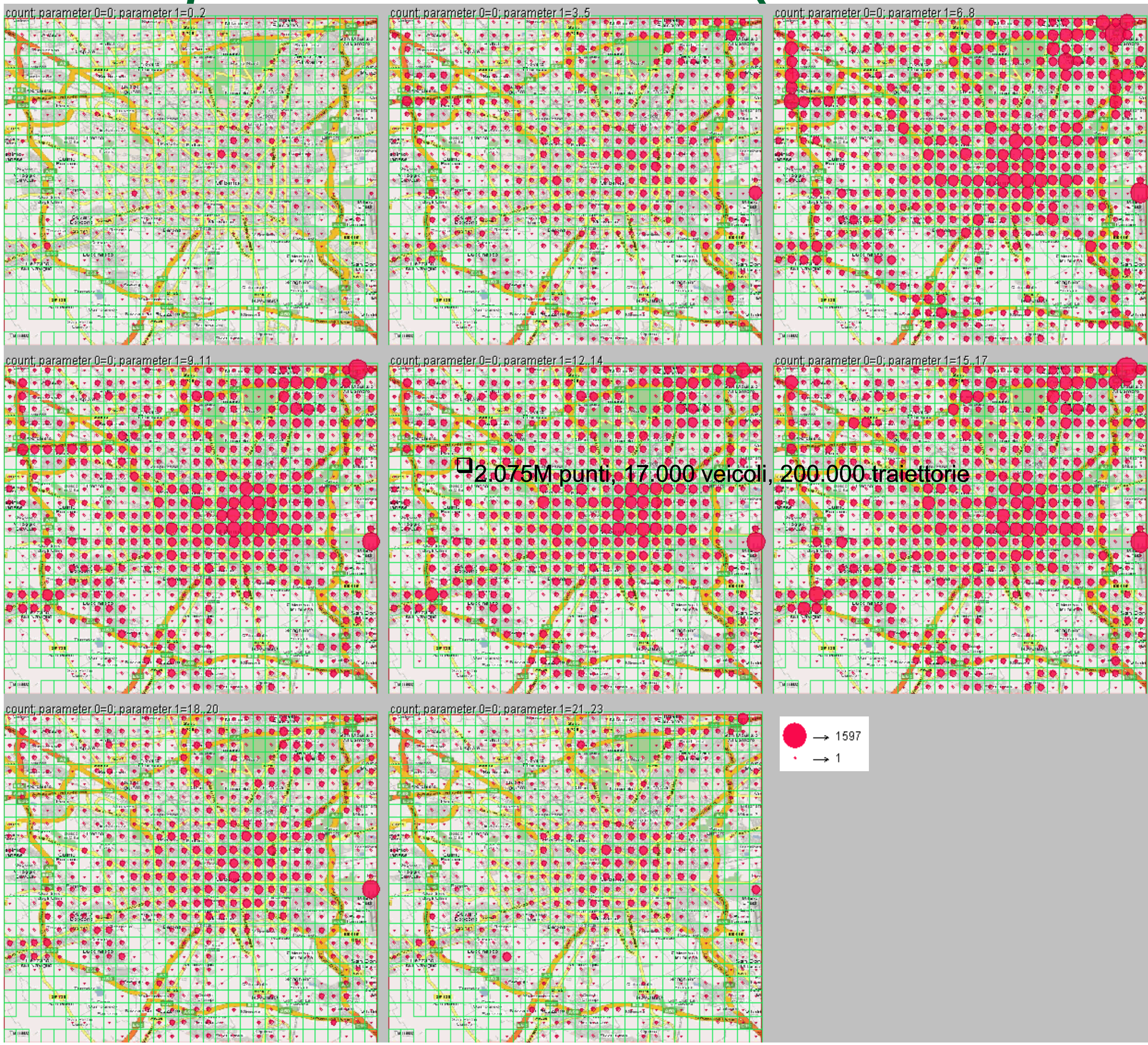
(a) Correct!

(b) Not Correct!

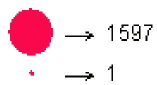
Traffic density patterns (spatio-temporal aggregation)



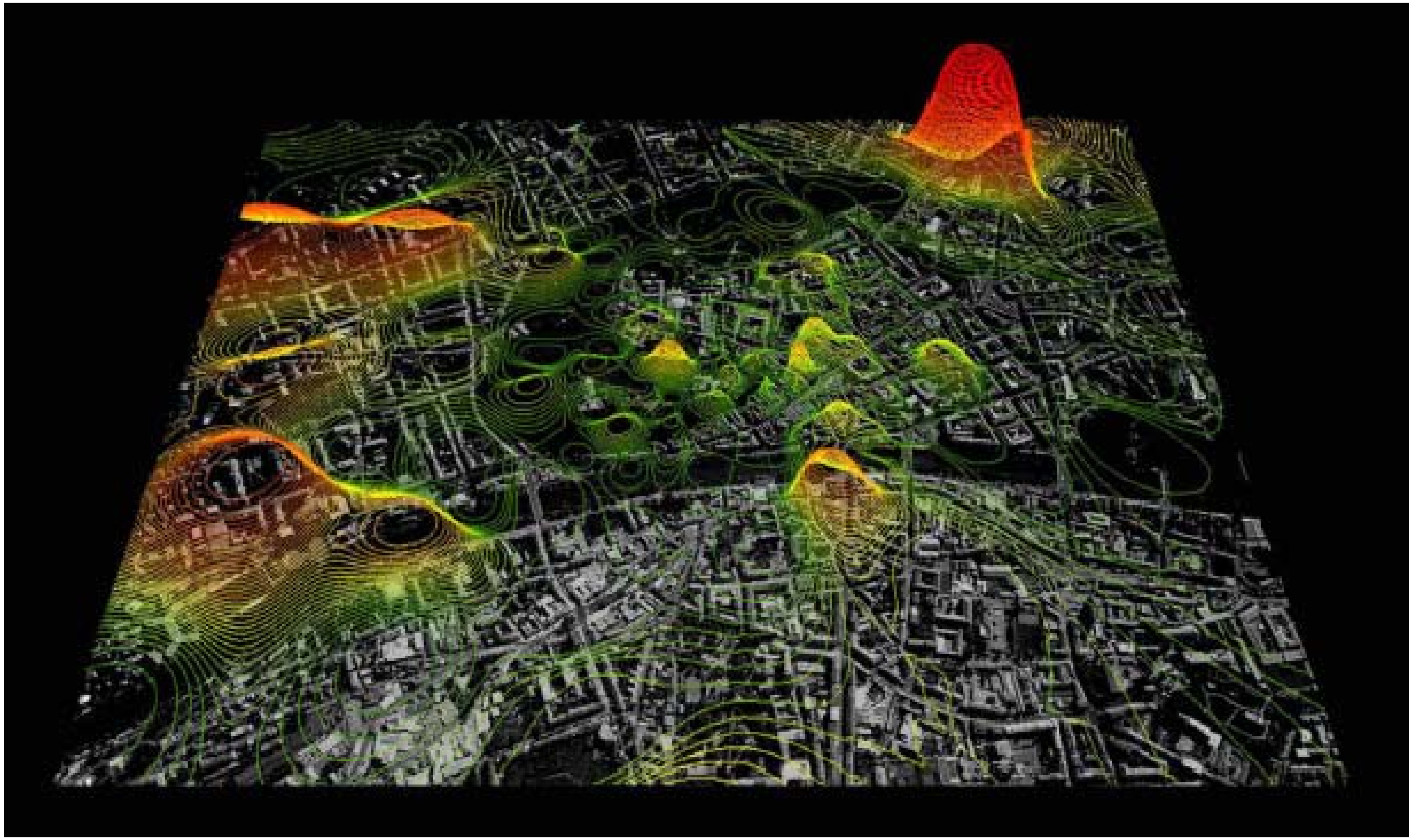
Low-speed movement (counts, 3h intervals)



- 2.075M punti, 17.000 veicoli, 200.000 traiettorie



Real-time density estimation in urban areas



The senseable project: http://senseable.mit.edu/grazrealtime/

References

- **eCourier.co.uk dataset**, http://api.ecourier.co.uk/.
- Han, J., Stefanovic, N., and Koperski, K. **Selective Materialization: An Efficient Method for Spatial Data Cube Construction**. Proc. PAKDD, 1998.
- Orlando, S., Orsini, R., Raffaetà, A., Roncato, A., and Silvestri, C. **Spatio-Temporal Aggregations in Trajectory Data Warehouses**. Proc. DaWaK, 2007.
- Gerasimos Marketos(1), Elias Frentzos(1), Irene Ntoutsi(1),Nikos Pelekis(1), Alessandra Raffaetà(2), and Yannis Theodoridis(1). **Building Real World Trajectory Warehouses.** Proc. MobiDE'08, Vancouver, Canada
- Pelekis, N., Raffaetà, A., Damiani, M.-L., Vangenot, C., Marketos, G., Frentzos, E., Ntoutsi, I., and Theodoridis, Y. **Towards Trajectory Data Warehouses**. Chapter in Mobility, Data Mining and Privacy: Geographic Knowledge Discovery. Springer-Verlag. 2008.
- Pfoser, D., Jensen, C.S., and Theodoridis, Y. **Novel Approaches to the Indexing of Moving Object Trajectories**, Proc. VLDB, 2000.
- Tao, Y., Kollios, G., Considine, J., Li, F., and Papadias, D. **Spatio-Temporal Aggregation Using Sketches**. Proc. ICDE, 2004.

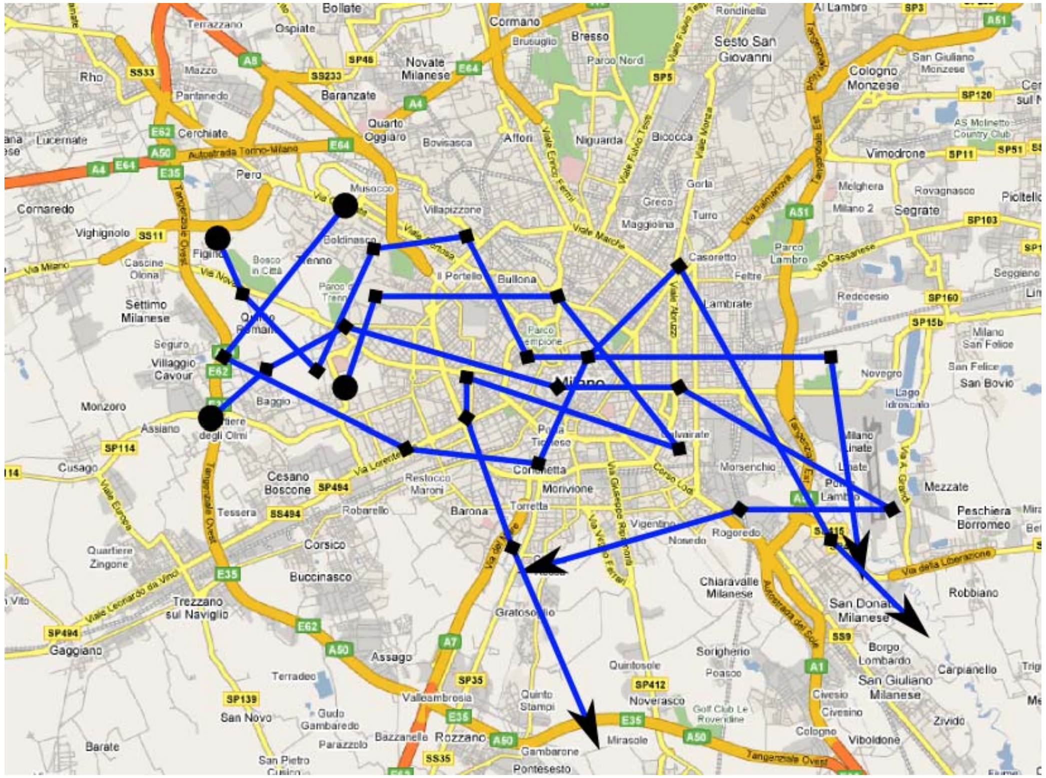
Mobility data mining

Trajectory Pattern Mining

Trajectory Classification

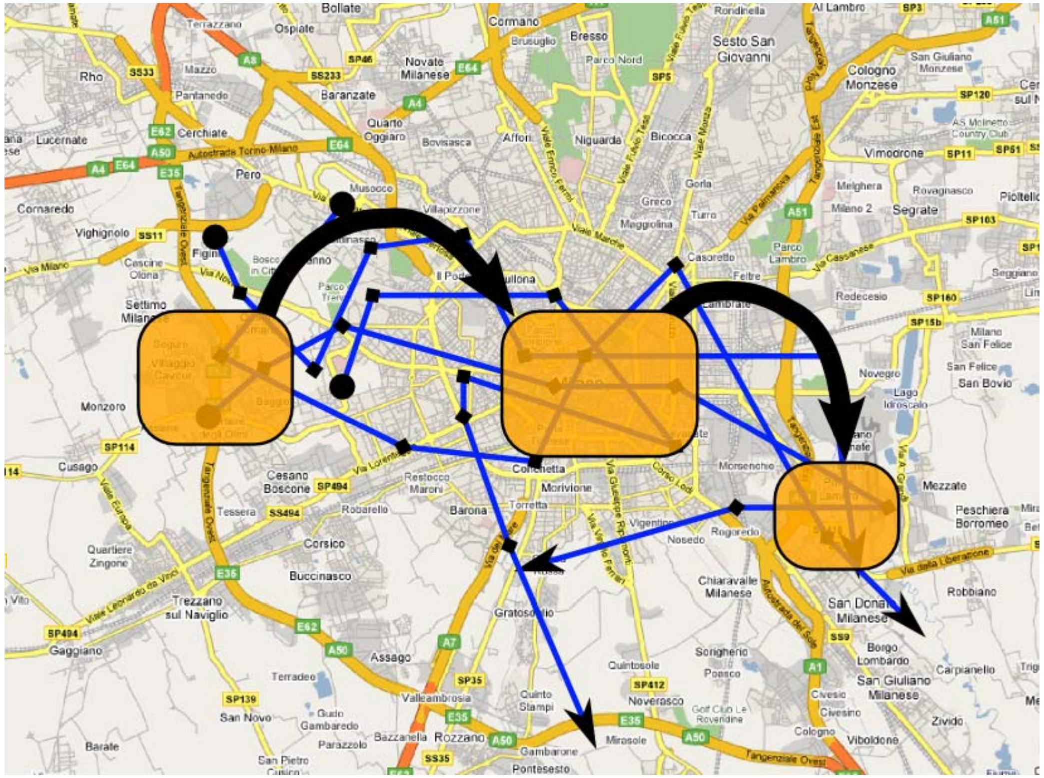
Trajectory Clustering

Q: What is a trajectory pattern?



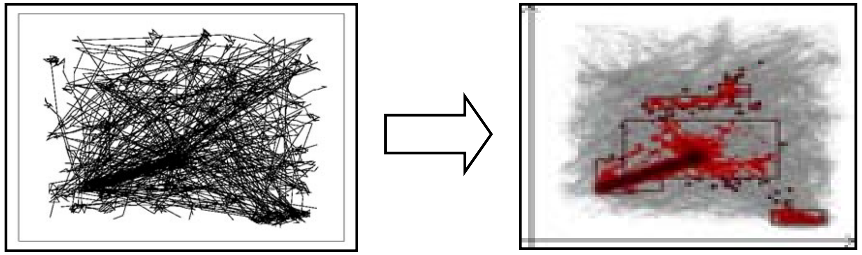
A: *A spatio-temporal sequential pattern*

- A sequence of visited regions, **frequently** visited in the **specified order** with **similar transition times**



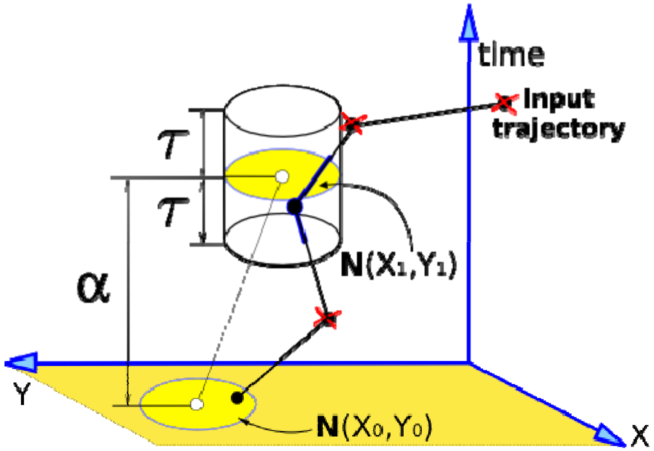
- Giannotti, Nanni, Pedreschi, Pinelli. Trajectory pattern mining. In Proc. ACM SIGKDD 2007

T-Pattern discovery

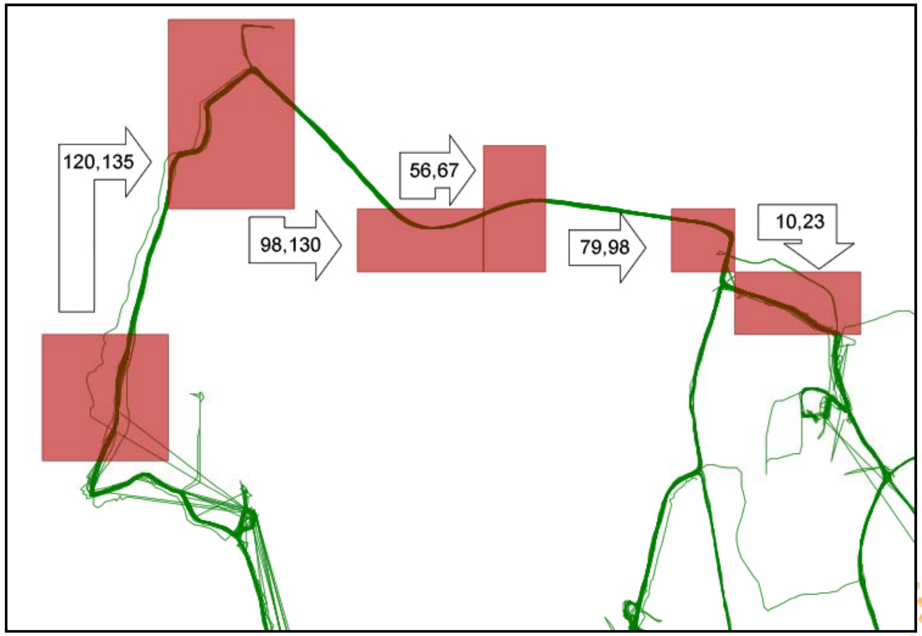


1- Find Regions of Interest

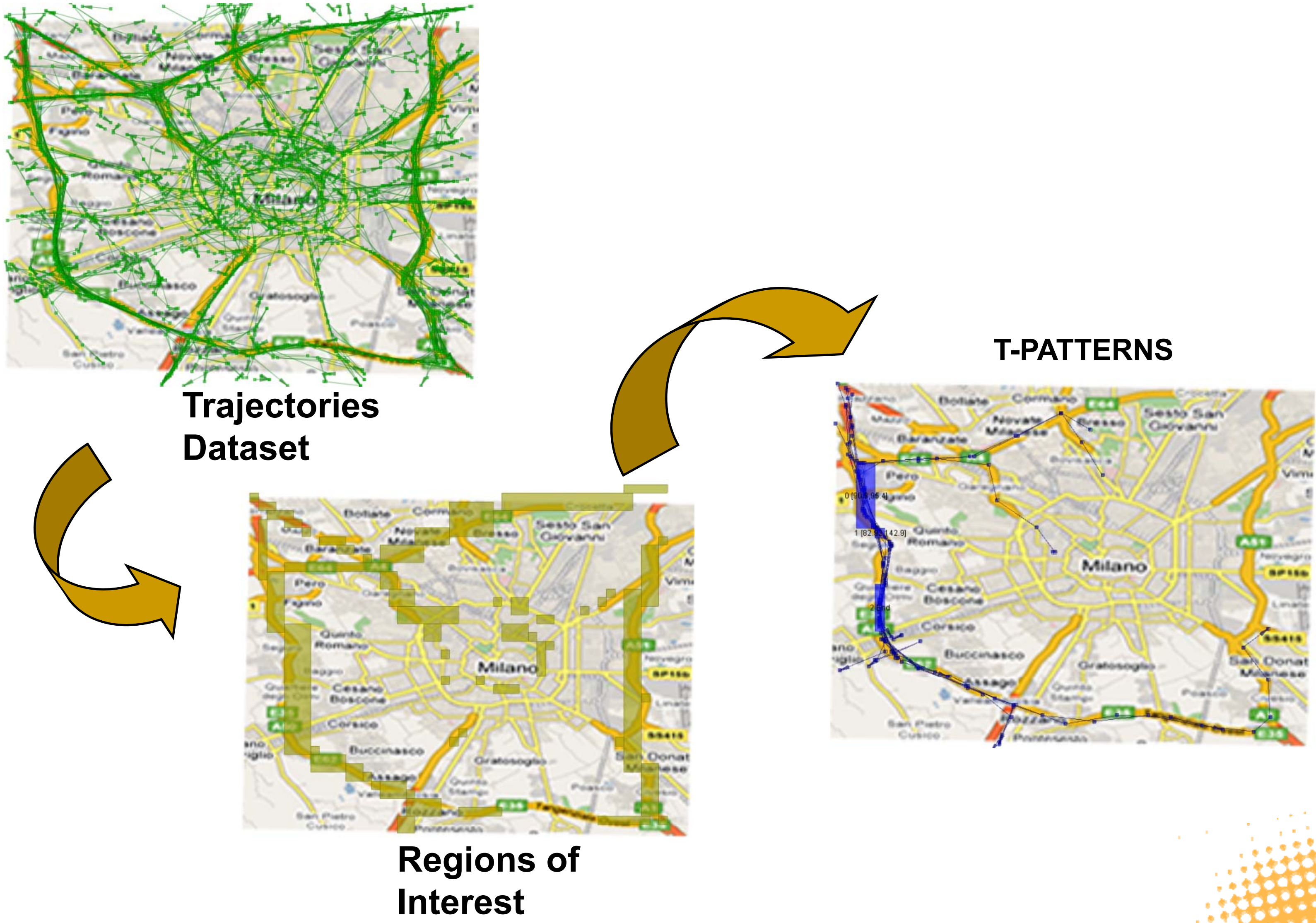
2- Find similar Trajectory in space and time



3- Extract patterns:



T-Pattern: Extraction Process



T-Patterns for trajectories

- A **Trajectory Pattern** (T-pattern) is a pair ($\mathbf{s}$, α):
  - **s** = <(x_0,y_0),..., (x_k,y_k)> is a sequence of k+1 locations
  - α = <α_1,..., α_k> are the transition times (*annotations*)

  also written as: $$(x_0, y_0) \xrightarrow{\alpha_1} (x_1, y_1) \xrightarrow{\alpha_2} \cdots \xrightarrow{\alpha_k} (x_k, y_k)$$

- A T-pattern T_p **occurs** in a trajectory if it contains a sub-sequence S such that:
  - each (x_i,y_i) in T_p matches a point (x_i',y_i') in S, and
  - the transition times in T_p are similar to those in S

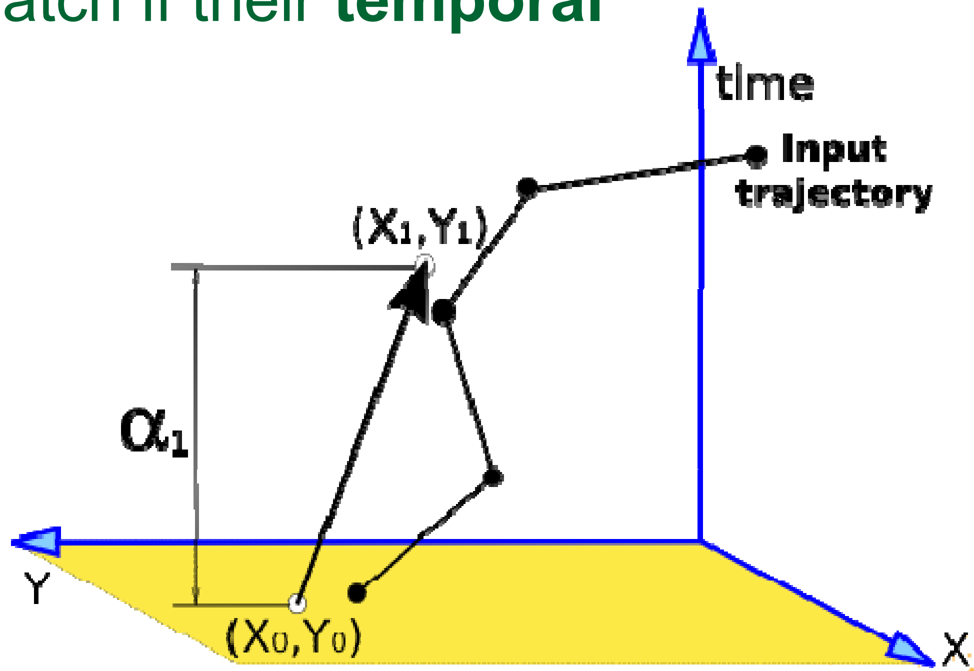
Continuity issues (space & time)

- The same exact spatial location (x,y) usually never occurs twice
- The same exact transition times usually do not occur twice
- Solution: allow approximation
  - a notion of *spatial neighborhood*
  - a notion of *temporal tolerance*

T-Pattern: approximate *occurrence*

- Two points match if one falls within a **spatial neighborhood N()** of the other
- Two transition times match if their **temporal difference is ≤ τ**
- Example:

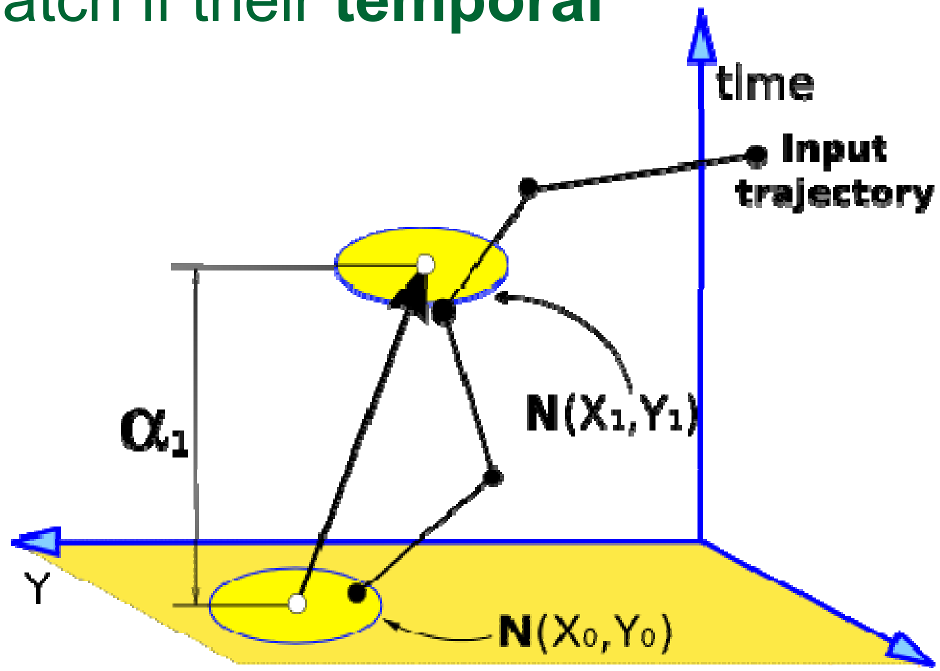
$$(x_0, y_0) \xrightarrow{\alpha_1} (x_1, y_1)$$



T-Pattern: approximate *occurrence*

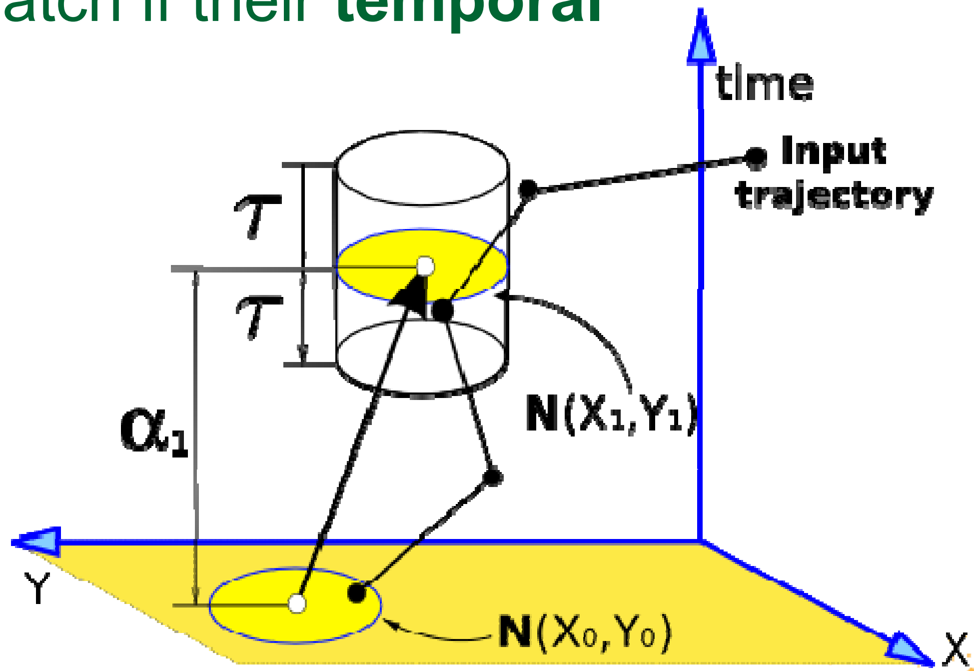
- Two points match if one falls within a **spatial neighborhood N()** of the other
- Two transition times match if their **temporal difference is ≤ τ**
- Example:

$$(x_0, y_0) \xrightarrow{\alpha_1} (x_1, y_1)$$



T-Pattern: approximate *occurrence*

- Two points match if one falls within a **spatial neighborhood N()** of the other
- Two transition times match if their **temporal difference is ≤ τ**



- Example:

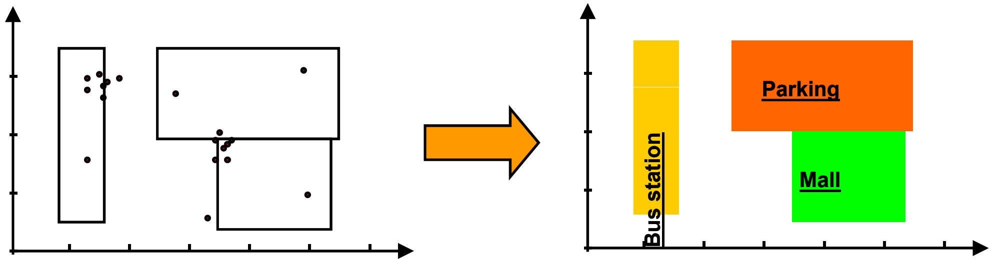
$$(x_0, y_0) \xrightarrow{\alpha_1} (x_1, y_1)$$

Computing general T-Patterns

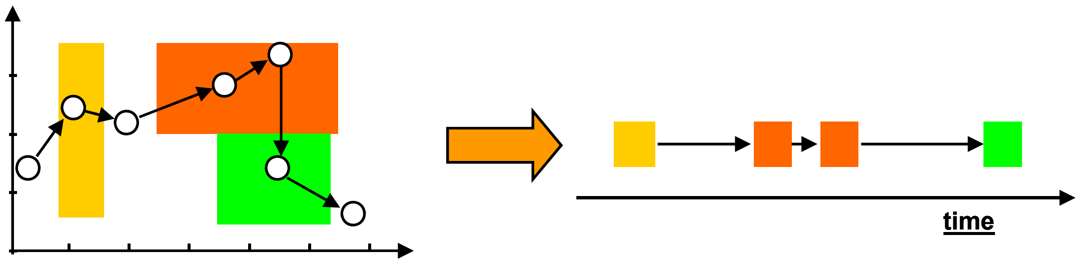
- T-pattern mining can be mapped to a density estimation problem over R^{3n-1}
  - 2 dimensions for each (x,y) in the pattern (2n)
  - 1 dimension for each transition (n-1)
- Density computed by
  - mapping each sub-sequence of n points of each input trajectory to R^{3n-1}
  - drawing an influence area for each point (composition of **N()** and **τ**)
- Too computationally expensive, heuristics needed!!!

Approach 1: predefined regions

- Fix a set of pre-defined regions of interest



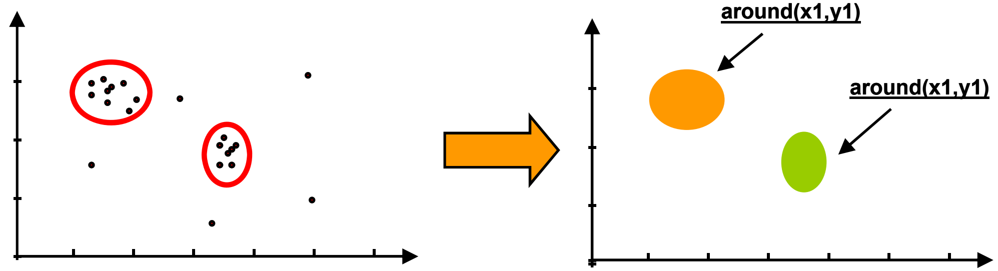
- Map each (x,y) of the trajectory to its region



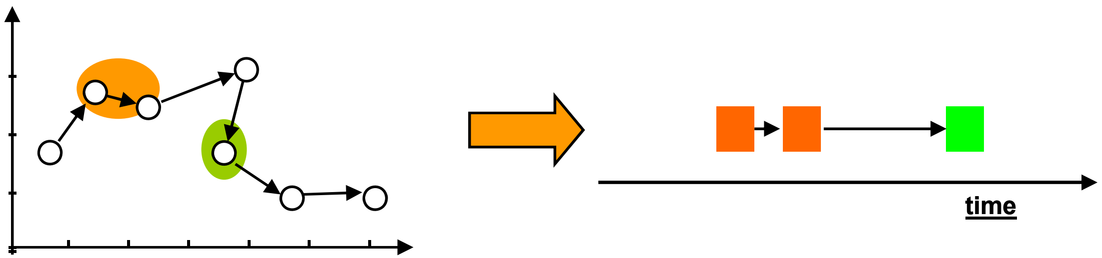
- **Sample pattern:** $$\textit{Bus station} \xrightarrow{20\,\text{min.}} \textit{Mall}$$

Approach 2: static discovered regions

- Detect significant regions thru spatial clustering



- **Map each (x,y) of the trajectory to its region**

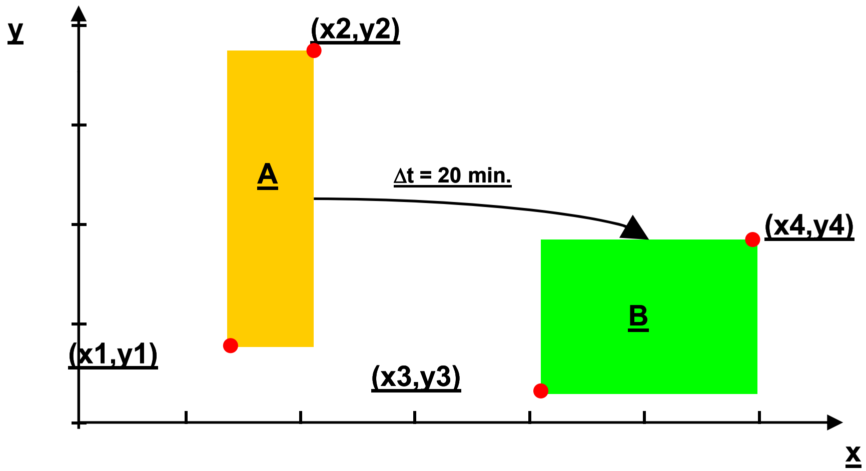


- **Sample pattern:** $around(x_1, y_1) \xrightarrow{20\,\mathrm{min.}} around(x_2, y_2)$

Approach 3: dynamic discovered regions

- Dynamic discovering of dense regions
  - Regions are located at each step of the pattern generation

- Sample pattern: $$(x, y) \in A \xrightarrow{20\,\text{min.}} (x, y) \in B$$



≡

1.Considering all trajectories, A is a cluster/dense region
2.Considering only trajectories that visit A, B is a cluster
3.”20 mins” is a typical time for pattern A→B

Static Neighborhoods

Regions-of-Interest (RoI)

- Given a set of *Regions of Interest R,* define the neighborhood of (x,y) as:

$$N_R(x,y) = \begin{cases} A & \text{if } A \in R \;\&\; (x,y) \in A \\ \emptyset & \text{otherwise} \end{cases}$$

  - Neighbors ⇔ belong to the same region
  - Points in no region have no neighbors

From ST-sequences to sequences

- With static neighborhoods $N_R()$ ST-sequences replaced by corresponding seqs of regions:

  A T-pattern ($\mathbf{s},\alpha$) is contained in a ST-sequence $S=<(x_1,y_1,t_1), ..., (x_n,y_n,t_n)>$ ⇔ the TAS ($\mathbf{s}',\alpha$) is contained in sequence S'

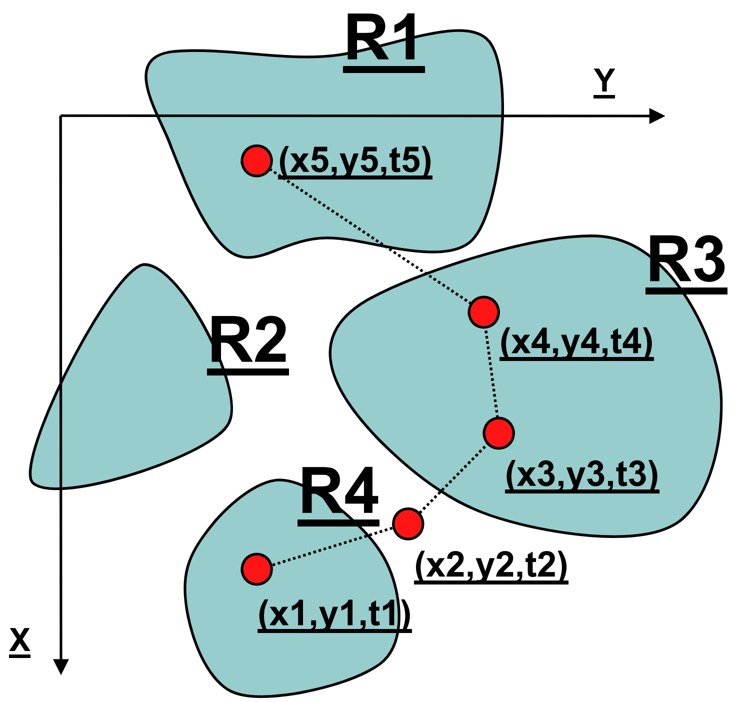
  - ***s'*** (resp. *S'*) is obtained by mapping each element (x,y) of ***s*** (resp. *S*) to $N_R(x,y)$
  - TAS = Temporally annotated seq. of labels

  - E.g.: $$s_0 \xrightarrow{\alpha_1} s_1 \xrightarrow{\alpha_2} \cdots \xrightarrow{\alpha_n} s_n$$

  - *Fosca Giannotti, Mirco Nanni, Dino Pedreschi.* Efficient Mining of Temporally Annotated Sequences. *SIAM-DM 2006.*

Translating ST-sequences

Example



S=<(x1,y1,t1), ..., (x5,y5,t5)>

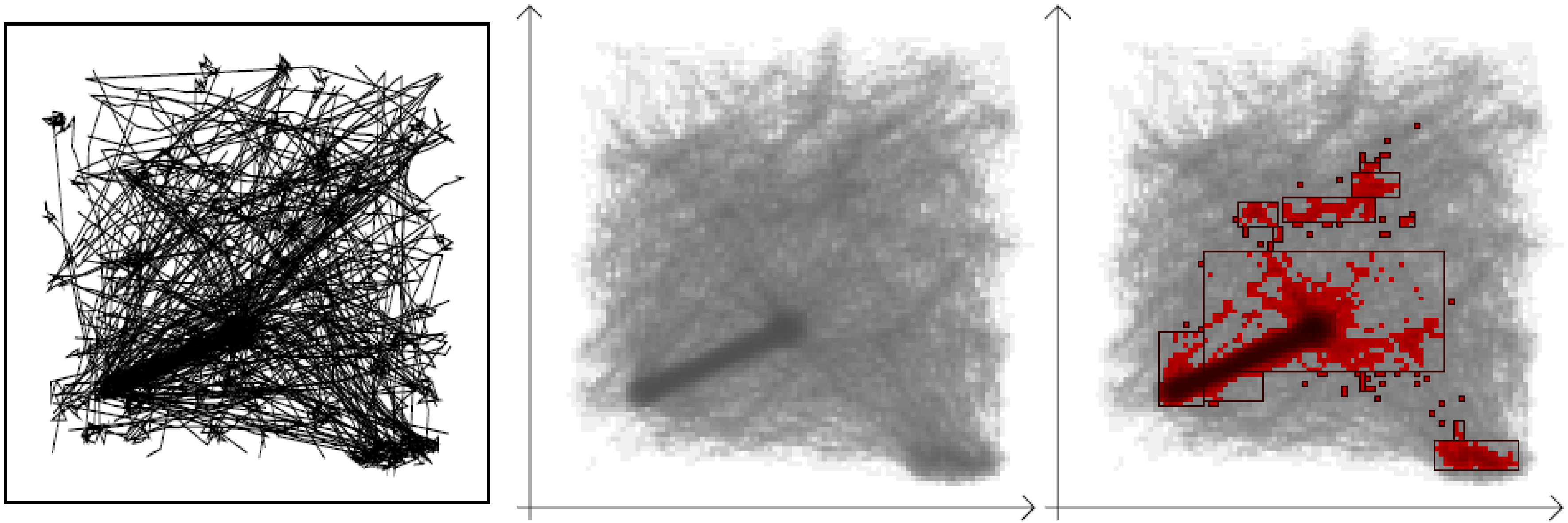
<(R4,t1), (R3,t3), (R3,t4), (R1,t5)>

Static Neighborhoods: issue

- What if RoI are not known a priori?
- Solution: define heuristics for automatic RoI extraction from data
- Wide range of heuristics:
  - Geography-based (e.g., crossroads)
  - Usage-based (e.g., popular places)
  - Mixed (e.g., popular squares)

Static Neighborhoods

A usage-based heuristic



(a) input trajectories (b) density distribution (c) dense cells and extracted RoI

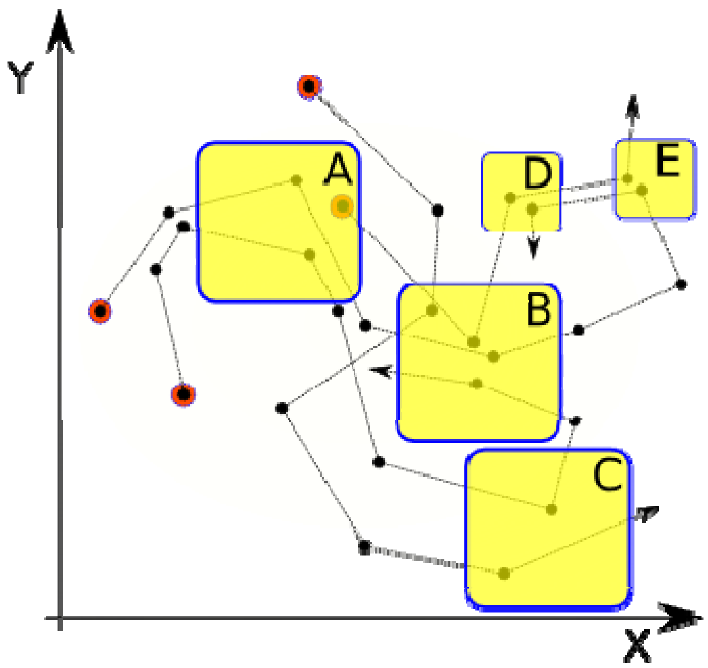
1. Impose a regular grid over space
2. Find dense cells (i.e., touched by many trajs.)
3. Coalesce cells into rectangles of bounded size

Multi-step refinement RoI

- Static RoI
  - Cells approximate single points, regions group points that are likely to form similar patterns
  - Yet, they should regard only trajectories that support the discovered pattern, not all database
- Towards general T-patterns
  - Check & update dense cells and regions of each pattern against the trajectories that support it
  - Approximation: Perform the update as step-wise refinement as patterns grow

Step-wise dynamic RoI

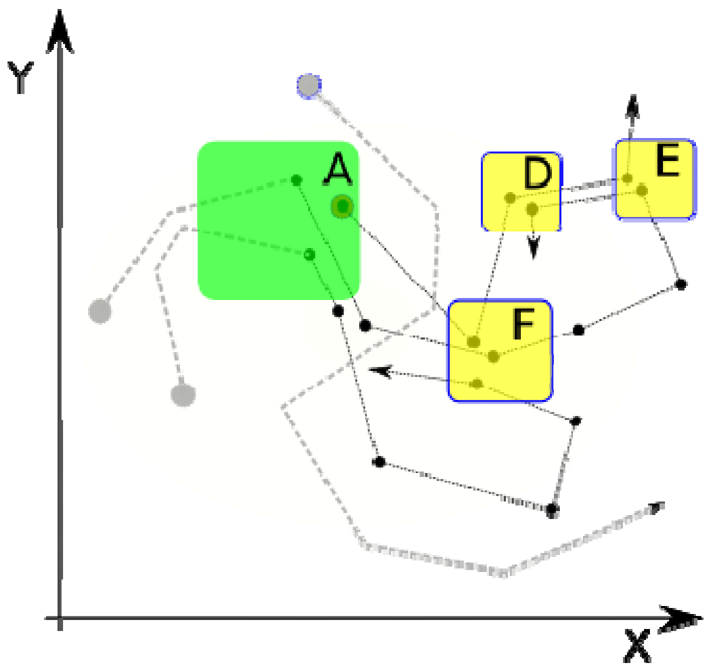
Example



- Start computing regions as basic RoI approach
- Regions describe interesting places of *everybody*

Step-wise dynamic RoI

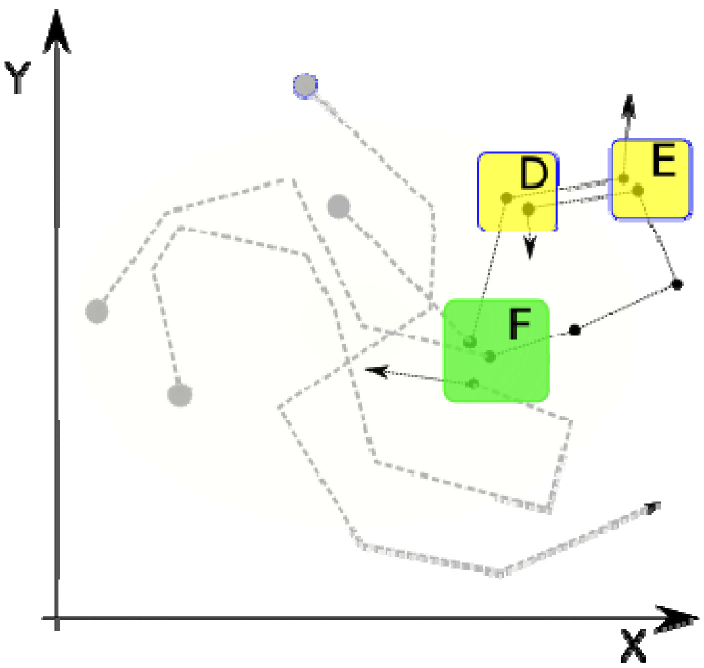
Example



- Focusing on A, we consider only the subset of relevant trajectories
- Regions can change (usually shrink/split)
- They are interesting only for who passes thru A

Step-wise dynamic RoI

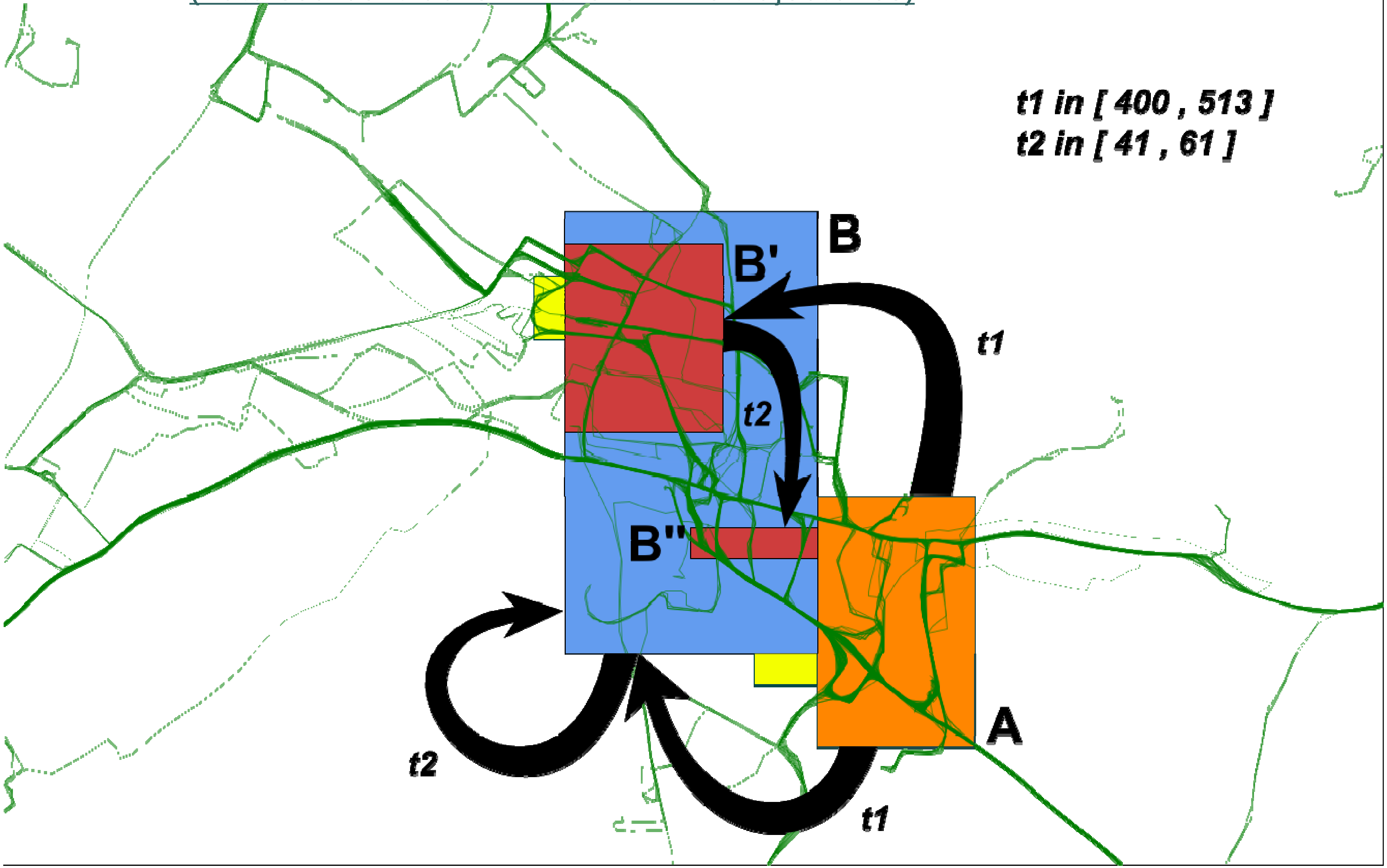
Example



- Focusing on A->F (with some transition time), we further restrict the set of trajectories involved
- The process is repeated as far as possible

Sample T-patterns

(Data source: trucks in Athens – 273 trajectories)



Related works on T-patterns

- H. Cao, N. Mamoulis, and D. W. Cheung. Mining frequent spatio-temporal sequential patterns. ICDM'05.
  - *patterns are in the form of sequences of trajectory segments, and their approximate instances are searched in the data*
- P. Kalnis, N. Mamoulis, and S. Bakiras. On discovering moving clusters in spatio-temporal data. SSTD'05.
  - *patterns are in the form of moving regions within time intervals, such as spatio-temporal cylinders or tubes. Instances are trajectory segments fully contained in the moving regions*
- N. Mamoulis, H. Cao, G. Kollios, M. Hadjieleftheriou, Y. Tao, and D. Cheung. Mining, indexing, and querying historical spatiotemporal data. KDD'04.
  - *maximal periodic patterns, treating discrete time and continuous spatial locations that are discretized dynamically through density-based clustering*

Related works on T-patterns

- J. Yang and M. Hu. TrajPattern: Mining sequential patterns from imprecise trajectories of mobile objects. EDBT'06.
  - patterns in the form of sequences of locations are mined, and also the uncertainty of object locations is considered from a probabilistic viewpoint
- H. Cao, N.Mamoulis, and D.W. Cheung. Discovery of collocation episodes in spatiotemporal data.ICDM'06.
  - input objects are associated to an object type (e.g., deers, pumas, etc.), and then patterns describing the proximity (i.e., collocation) between object types are mined

Ongoing work

- Application-oriented assessments on large, real datasets show that T-patterns are many and difficult to evaluate
  - A starting point for further model construction, rather than a final product
- Simplification of output transition times
  - The most complex info for end users
- Study relations with
  - Geographic background knowledge, such as points of interests and road network
  - Privacy issues – are T-patterns safe? Can we use T-patterns to protect (anonymize) original data?
  - Reasoning on trajectories and patterns

Mobility data mining

Trajectory Pattern Mining

Trajectory Classification

Trajectory Clustering

Location prediction based on T-patterns

F. Pinelli, A. Monreale, R. Trasarti, F. Giannotti

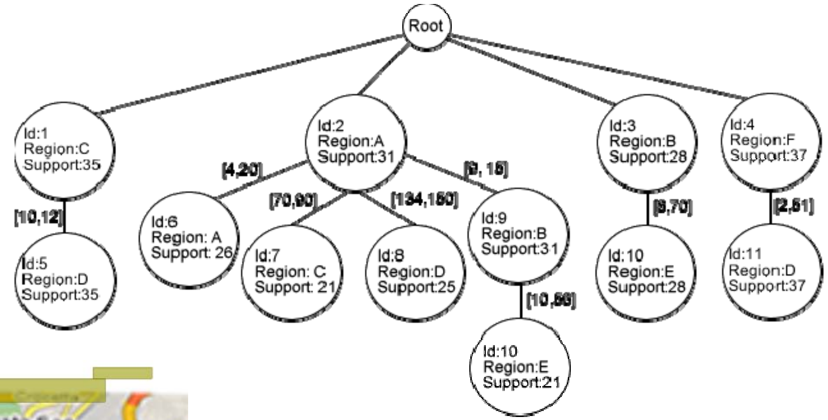
Location prediction within the mobility data analysis environment Daedalus

Workshop on Intelligent Transportation Systems @MDM 2008

Location Prediction: Idea

T-Pattern extracts a set of local patterns from a global set of data.

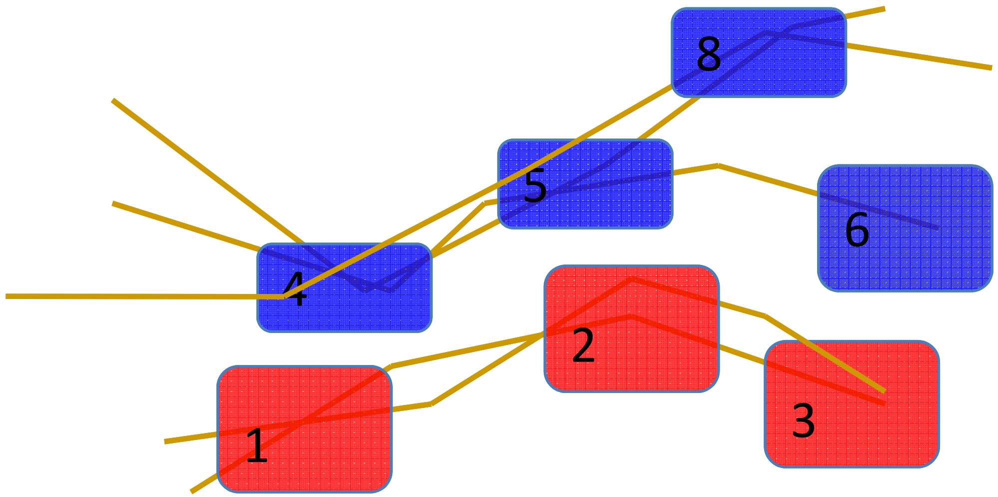
Can we use these patterns to build a global model to predict the next location?



Local patterns (T-pattern)

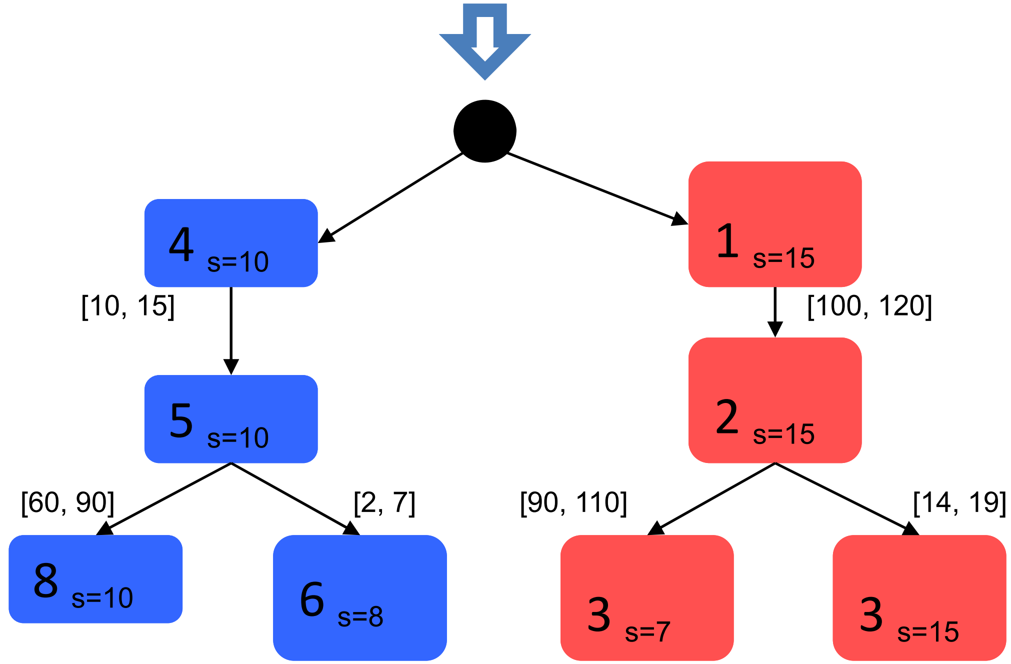
Global model (Ptree)

Location Prediction: Building Ptree



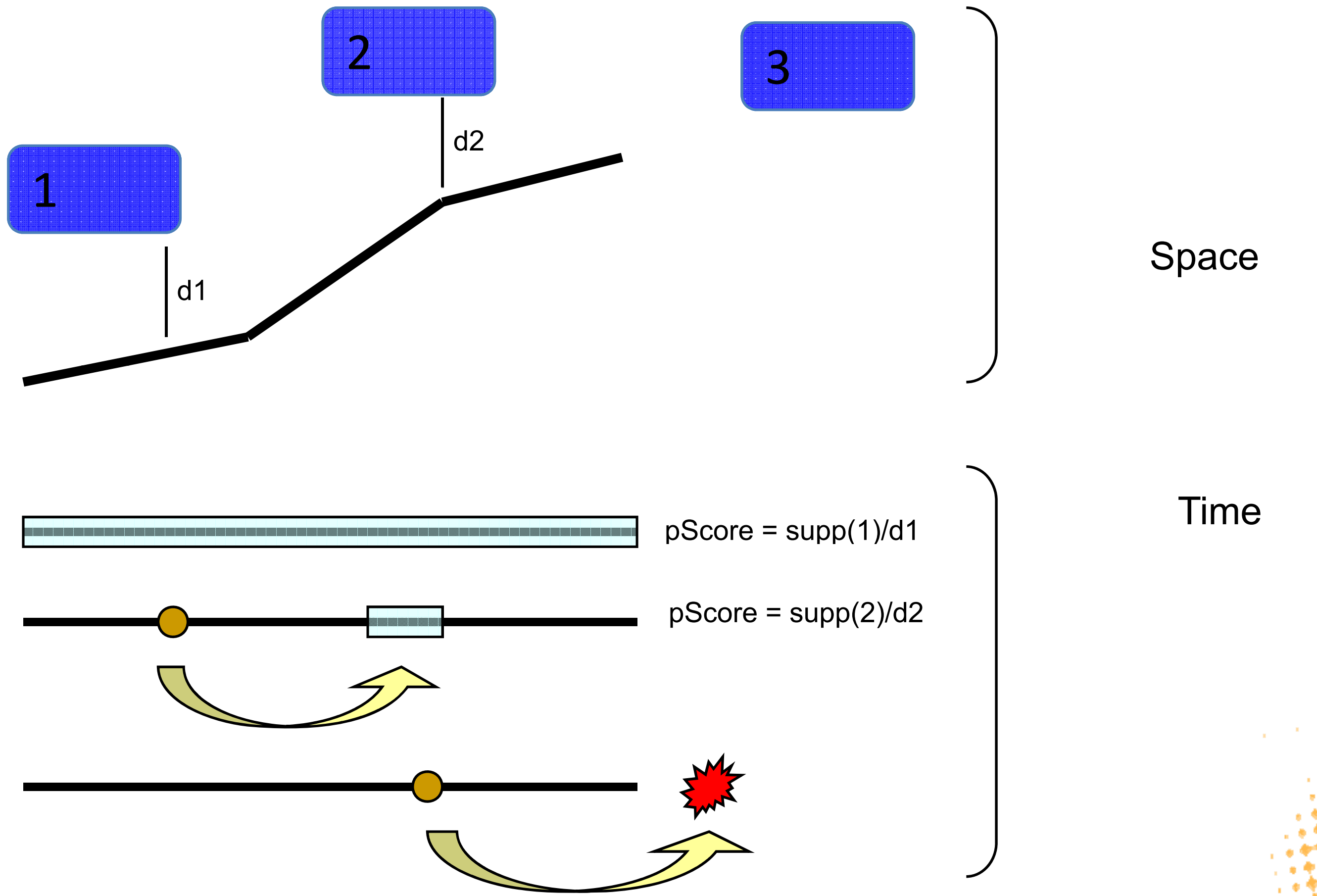
T-Pattern results:

4 [10, 15] → 5 [60, 90] → 8 s.10
4 [10, 15] → 5 [2, 7] → 6 s. 8
1 [100, 120] → 2 [90, 110] → 3 s. 7
1 [100, 120] → 2 [14, 19] → 3 s.15

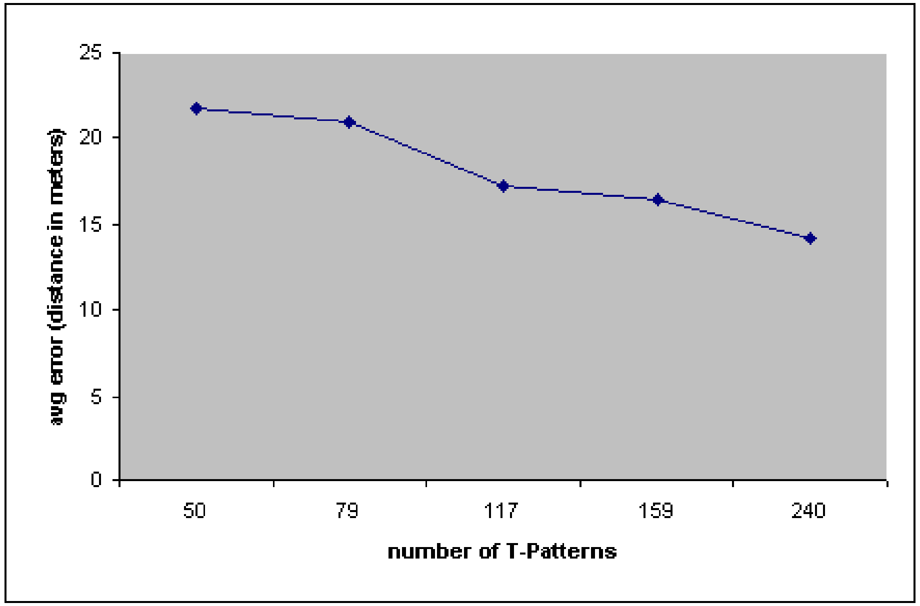


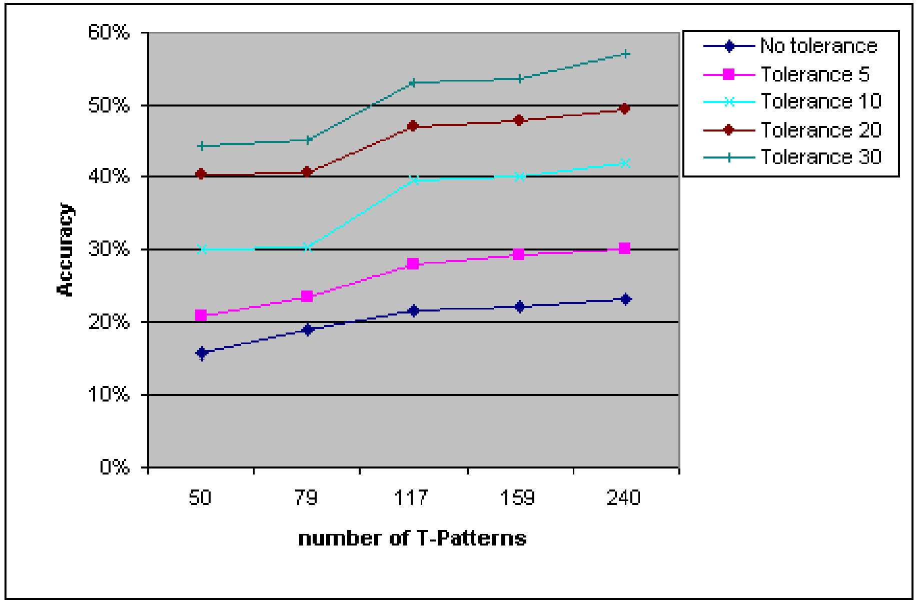
Location Prediction

The idea is to find the pattern that best matches a given trajectory computing the ***puntual score*** for each admissible node in the Ptree and then the ***score*** of a path on it.



Experiments





Works on location prediction

- B. Xu and O. Wolfson. Time-series prediction with applications to traffic and moving objects databases. MobiDE, 2003.
- G. Yavas, D. Katsaros, O. Ulusoy, Y. Manolopoulos. A data mining approach for location prediction in mobile environments. Data Knowl. Eng., 54(2):121–146, 2005.
- M. Morzy. Prediction of moving object location based on frequent trajectories. ISCIS 2006, LNCS 4263 Springer.
- M. Morzy. Mining frequent trajectories of moving objects for location prediction. MLDM 2007, LNCS 4571 Springer.
- H. Jeung, Q. Liu, H. T. Shen, and X. Zhou. A hybrid prediction model for moving objects. ICDE, 2008.

Semantic annotation of mobility raw data

- many applications in the mobility domain require a semantic interpretation of movement information
  - traffic management, site evaluation, LBS, advertisement

- physical trajectories can be retrieved by GPS loggers

- **obtaining semantic trajectories is a challenge**

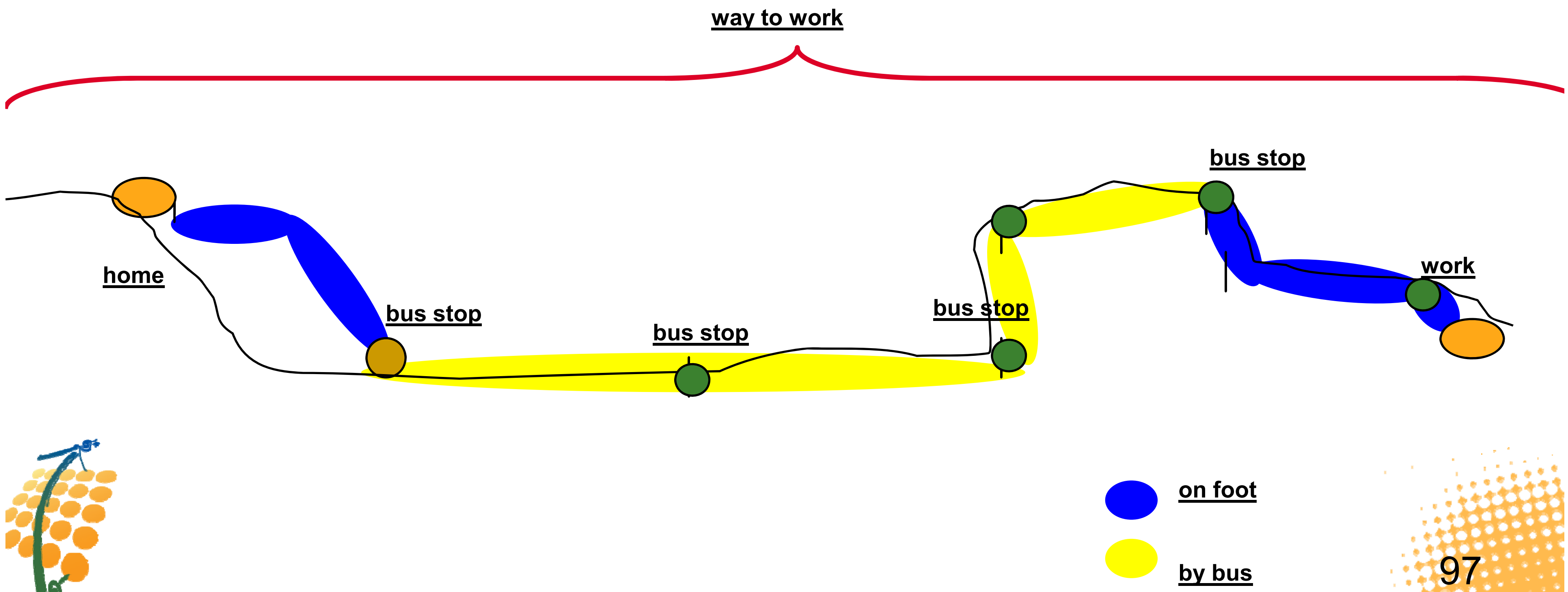
Semantic Annotation of GPS Trajectories

Physical Trajectory:

- e.g. GPS recording over some period of time

Semantic Trajectory:

- places where a person stayed
- means of transportation
- combination of above elements for higher-level description



Semantic Annotation of GPS Trajectories

Barış Güç, Michael May, Yücel Saygın, Christine Körner

AGILE Conference, 2008

Related Work

- many studies show inconsistencies between GPS trajectories and travel diaries (Stopher 2007, Zmund 2003)
- automatic annotation of trajectories using background information and land uses (Axhausen 2003, Wolf et al. 2001, Wolf 2000) is limited in several aspects
  - focus on vehicular movement
  - distinguish only few trip purposes
  - ambiguous results possible due to land use data
  - the purpose of a trip can be irrelevant to its destination

- Axhausen, K.W., S. Schönfelder, J. Wolf, M. Oliveira and U. Samaga: 80 weeks of GPS-traces: Approaches to enriching the trip information, Arbeitsbericht Verkehrs- und Raumplanung, 2003.
- Wolf, J., Guensler R. and Bachman, W.: Elimination of the Travel Diary: An Experiment to Derive Trip Purpose from GPS Travel Data, Transportation Research Record, 1768, 125-134, 2001
- Wolf, J.: Using GPS data loggers to replace travel diaries in the collection of travel data, Dissertation, 2000
- Zmud, J. and Wolf, J.: Identifying the Correlations of Trip Misreporting – Results fro the California Statewide Household Travel Survey GPS Study. In: Proc. of the 10th International Conference on Travel Behaviour Research, 2003.

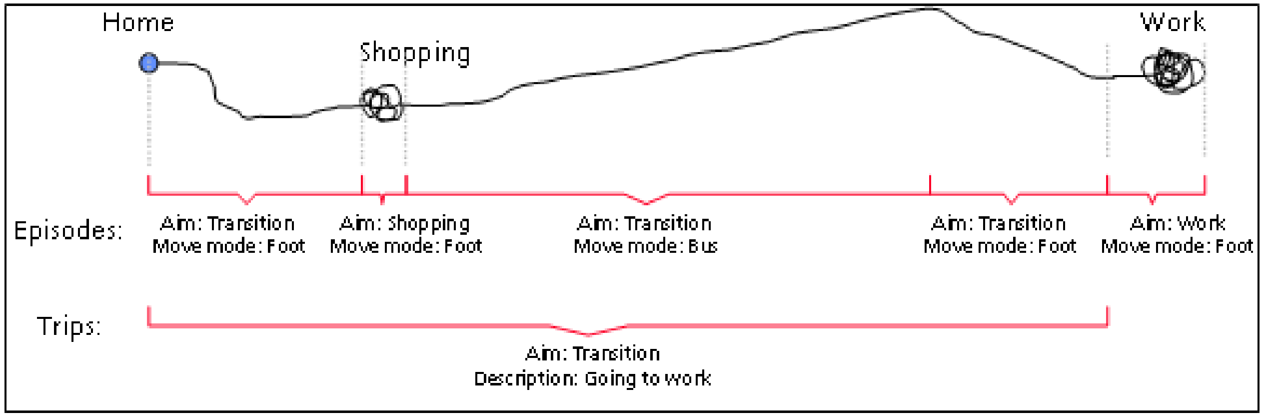
Aim

- ensure the accurate annotation of a trajectory by the user
  - present the physical trajectory in geographic and temporal context
  - assist the user during the annotation process
  - ensure consistency among users

- a tool to visualize, annotate and store GPS trajectory data

Annotation Model

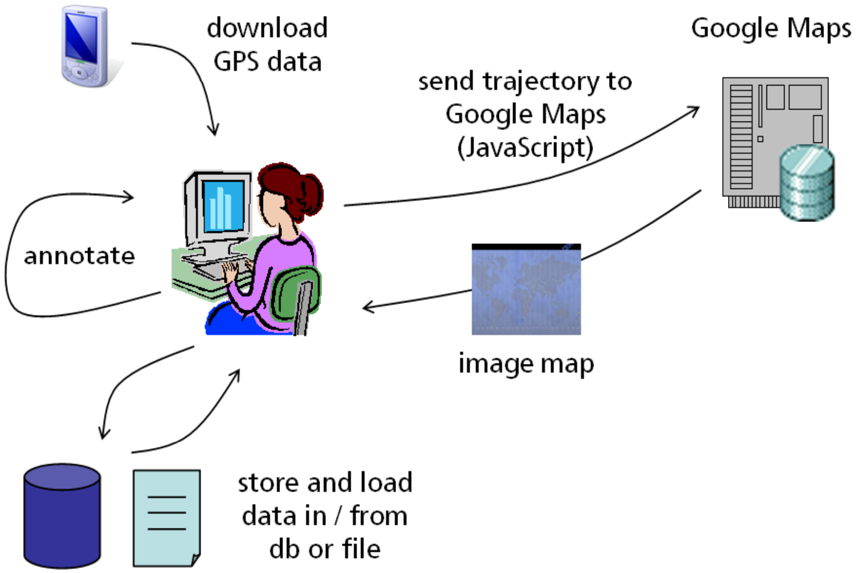
- annotation model follows the concept of episodes (Mountain 2001)
- semantic episodes are homogeneous sections of a trajectory with respect to
  - purpose of the movement (e.g. working, shopping, transition)
  - mode of transportation (e.g. by car, bus, foot)
- "Trips" for aggregating episodes on a higher semantic level
  - e.g.: all episodes on the way to work can be grouped into a common trip



- Mountain, D. M. and Raper J. F.: Modelling Human Spatio-Temporal Behaviour: A Challenge for Location-based Services. In: Proc. of the 6th International Conference on GeoComputation, 2001.

Annotation Workflow

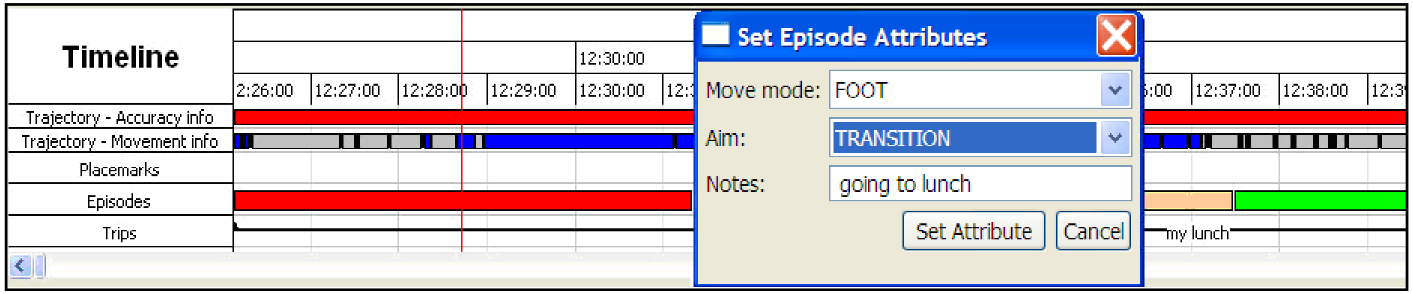
- Download data from GPS device
- Visualize trajectories using Google Maps
- Annotate on a “timeline”
- Store annotation and GPS raw data on central database

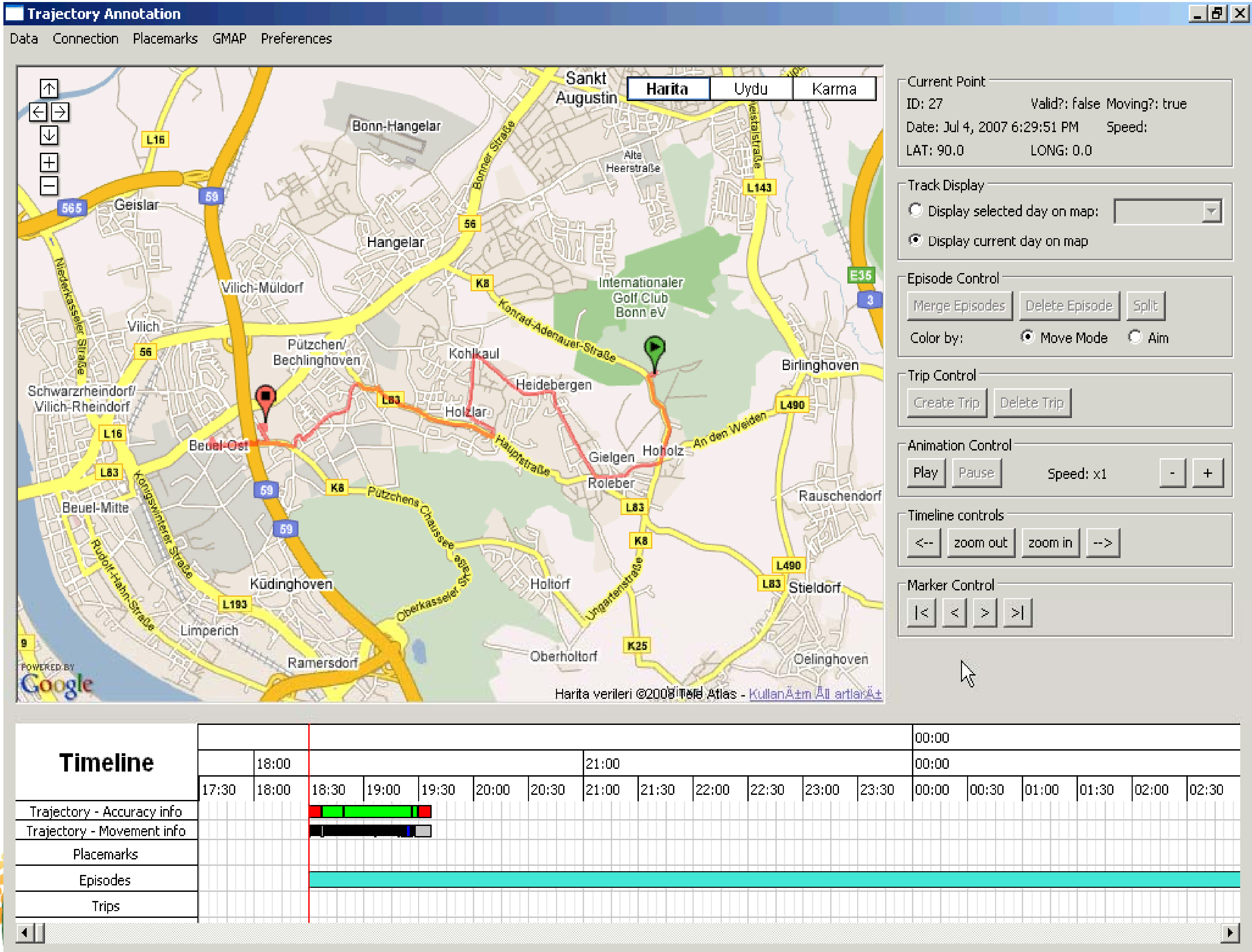


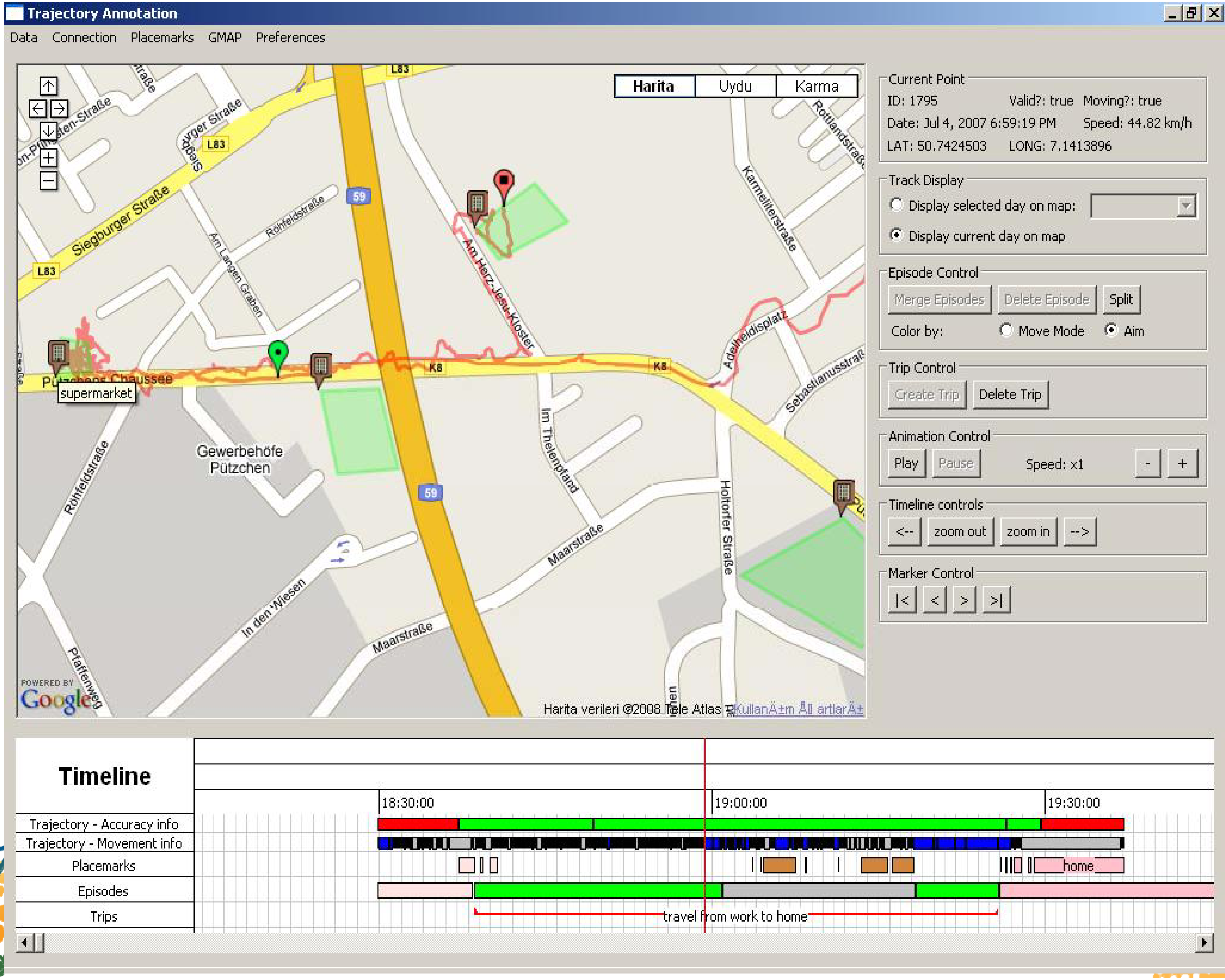
Interface Functionality

- Annotation
  - Annotate on the timeline by partitioning trajectory into episodes
  - Interface ensures consistency between users
  - flexible

- “Placemarks”
  - Users mark favorite places on the map
  - Display visited placemarks on the timeline




Trajectory Annotation
Data Connection Placemarks GMAP Preferences
Harita Uydu Karma
Current Point
ID: 27 Valid?: false Moving?: true
Date: Jul 4, 2007 6:29:51 PM Speed:
LAT: 90.0 LONG: 0.0
Track Display
Display selected day on map:
Display current day on map
Episode Control
Merge Episodes Delete Episode Split
Color by: Move Mode Aim
Trip Control
Create Trip Delete Trip
Animation Control
Play Pause Speed: x1 - +
Timeline controls
<-- zoom out zoom in -->
Marker Control
|< < > >|
Timeline
Trajectory - Accuracy info
Trajectory - Movement info
Placemarks
Episodes
Trips
17:30 18:00 18:30 19:00 19:30 20:00 20:30 21:00 21:30 22:00 22:30 23:00 23:30 00:00 00:30 01:00 01:30 02:00 02:30


Trajectory Annotation
Data
Connection
Placemarks
GMAP
Preferences
Harita
Uydu
Karma
Siegburger Straße
Pützchens Chaussee
supermarket
Gewerbehöfe Pützchen
Am Herz-Jesu-Kloster
Im Thelenpfad
Holtorfer Straße
Maarstraße
In den Wiesen
Röhfeldstraße
Am Langen Graben
Karmeliterstraße
Adelheidisplatz
Sebastianusstraße
Pfaffenweg
POWERED BY Google
Harita verileri ©2008 Tele Atlas
Current Point
ID: 1795
Valid?: true
Moving?: true
Date: Jul 4, 2007 6:59:19 PM
Speed: 44.82 km/h
LAT: 50.7424503
LONG: 7.1413896
Track Display
Display selected day on map:
Display current day on map
Episode Control
Merge Episodes
Delete Episode
Split
Color by:
Move Mode
Aim
Trip Control
Create Trip
Delete Trip
Animation Control
Play
Pause
Speed: x1
Timeline controls
zoom out
zoom in
Marker Control
Timeline
18:30:00
19:00:00
19:30:00
Trajectory - Accuracy info
Trajectory - Movement info
Placemarks
Episodes
Trips
home
travel from work to home

Challenges

- Extend approach to
  - automatic extraction of frequently visited places
  - automatically derive the means of transportation
  - provide the user with a possible annotation

- Use data with data mining and machine learning techniques for automatic annotation/classification

The Challenge of Trajectory Classification

- Build a predictive model that associates a trajectory with a class from a given set
  - E.g.: { car, motorbike, truck }

    { dangerous, non-dangerous }

- The model relies only on the movement described by the trajectory
  - Possibly with background knowledge about context

Features for trajectory classification

- Key phase in classification: represent trajectories through an alphabet of *behaviours*
  1. extract significant (frequent, discriminative, etc.) patterns emerging from data
  2. describe each trajectory in terms of which patterns it follows
  3. extract rules correlating descriptive patterns and target label

- From local patterns to global (predictive) models

Works on trajectory classification

- Scarce results so far, e.g.
- Fraile, R. and Maybank, S. J., “Vehicle Trajectory Approximation and Classification,” In *Proc. 9th British Machine Vision Conf.,* Southampton, UK, pp. 832–840, Sept. 1998.

Mobility data mining

Trajectory Pattern Mining

Trajectory Classification

Trajectory Clustering

Works on Trajectory Clustering

- Gaffney, S. and Smyth, P., Trajectory Clustering with Mixtures of Regression Models, ACM SIGKDD 1999.
- Gaffney, S., Robertson, A., Smyth, P., Camargo, S., and Ghil, M., Probabilistic Clustering of Extratropical Cyclones Using Regression Mixture Models, Tech. Rep. UCI-ICS 06-02, 2006.
- Nanni, M., Pedreschi, D. Time-focused clustering of trajectories of moving objects. J. of Intelligent Information Systems, 2006.
- Lee, J.-G., Han, J., and Whang, K.-Y., Trajectory Clustering: A Partition-and-Group Framework, SIGMOD 2007.
- Rinzivillo, Pedreschi, Nanni, Giannotti, Andrienko, Andrienko. Visually-driven analysis of movement data by progressive clustering. J. of Information Visualization, 2008

Which distance between trajectories?

- **Average Euclidean distance**

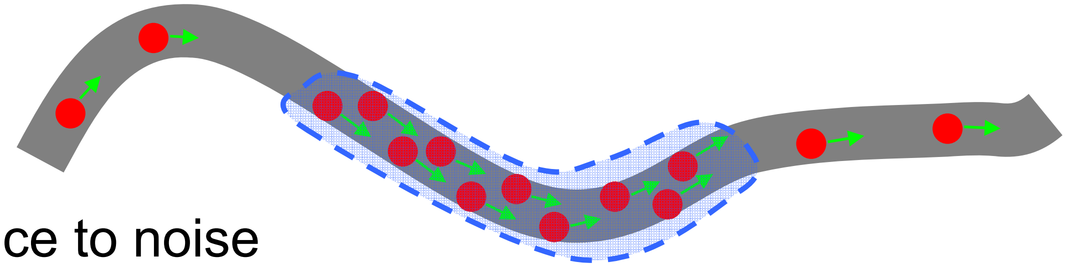
$$D(\tau_1, \tau_2)\,|_T = \frac{\int_T \boxed{d(\tau_1(t), \tau_2(t))}\, dt}{|T|}$$

distance between moving objects τ1 and τ2 at time *t*

- "Synchronized" behaviour distance
  - Similar objects = almost always in the same place at the same time
- Computed on the whole trajectory
- Computational aspects:
  - Cost = O(|τ1| + |τ2|)   (|τ| = number of points in τ)
  - It is a metric => efficient indexing methos allowed

Which kind of clustering?

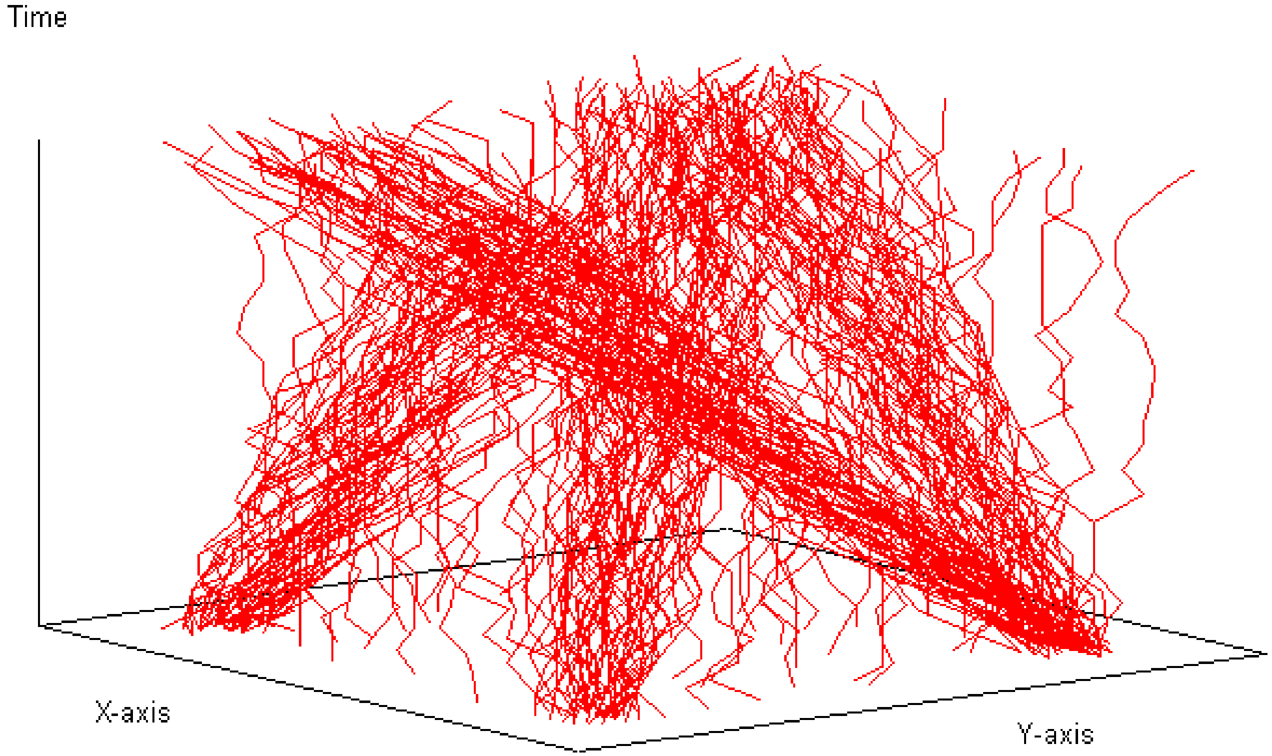
- General requirements:
  - Non-spherical clusters should be allowed
    - E.g.: A traffic jam along a road = “snake-shaped” cluster



  - Tolerance to noise
  - Low computational cost
  - Applicability to complex, possibly non-vectorial data
- A suitable candidate: Density-based clustering
  - OPTICS (Ankerst et al., SIGMOD 99
    - ➔ **T**(rajectory)-**OPTICS**

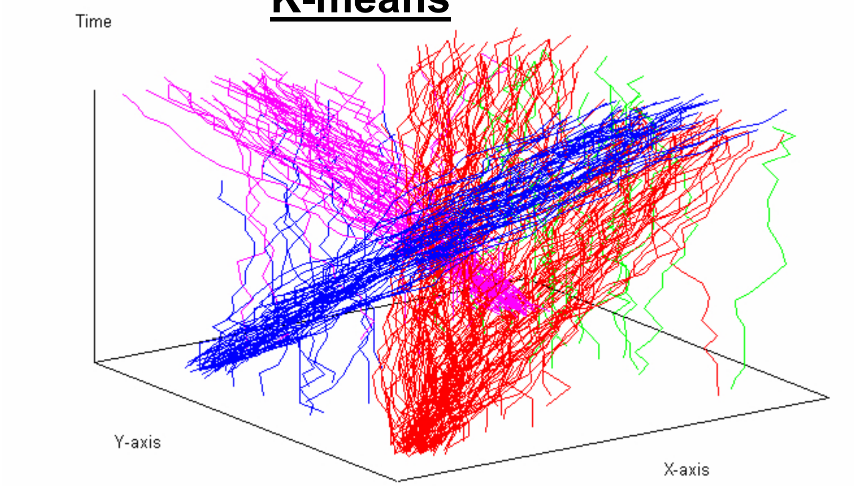
A sample dataset

- Set of trajectories forming 4 clusters + noise (synthetic)

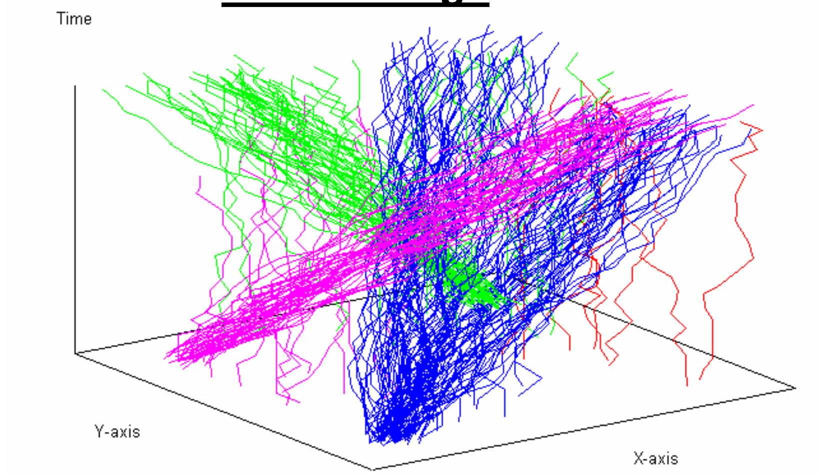


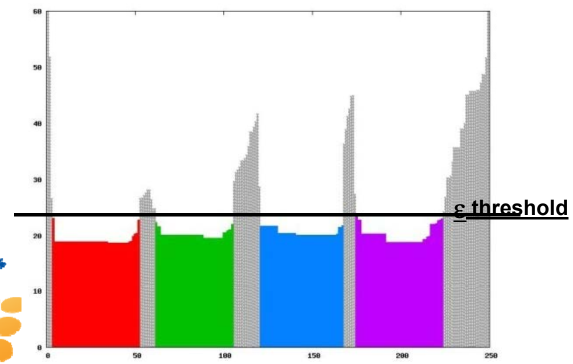
T-OPTICS vs. HAC & K-means

K-means



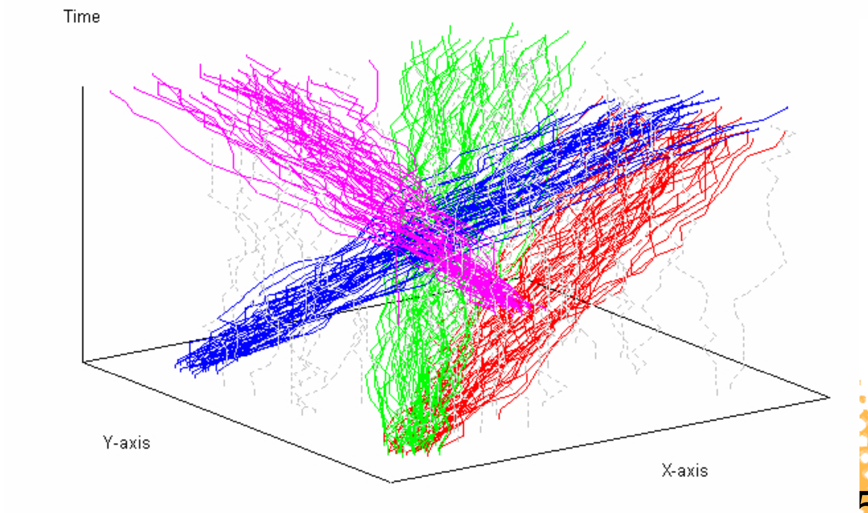
HAC-average





T-OPTICS

Reachability plot
(= objects reordering for distance distribution)



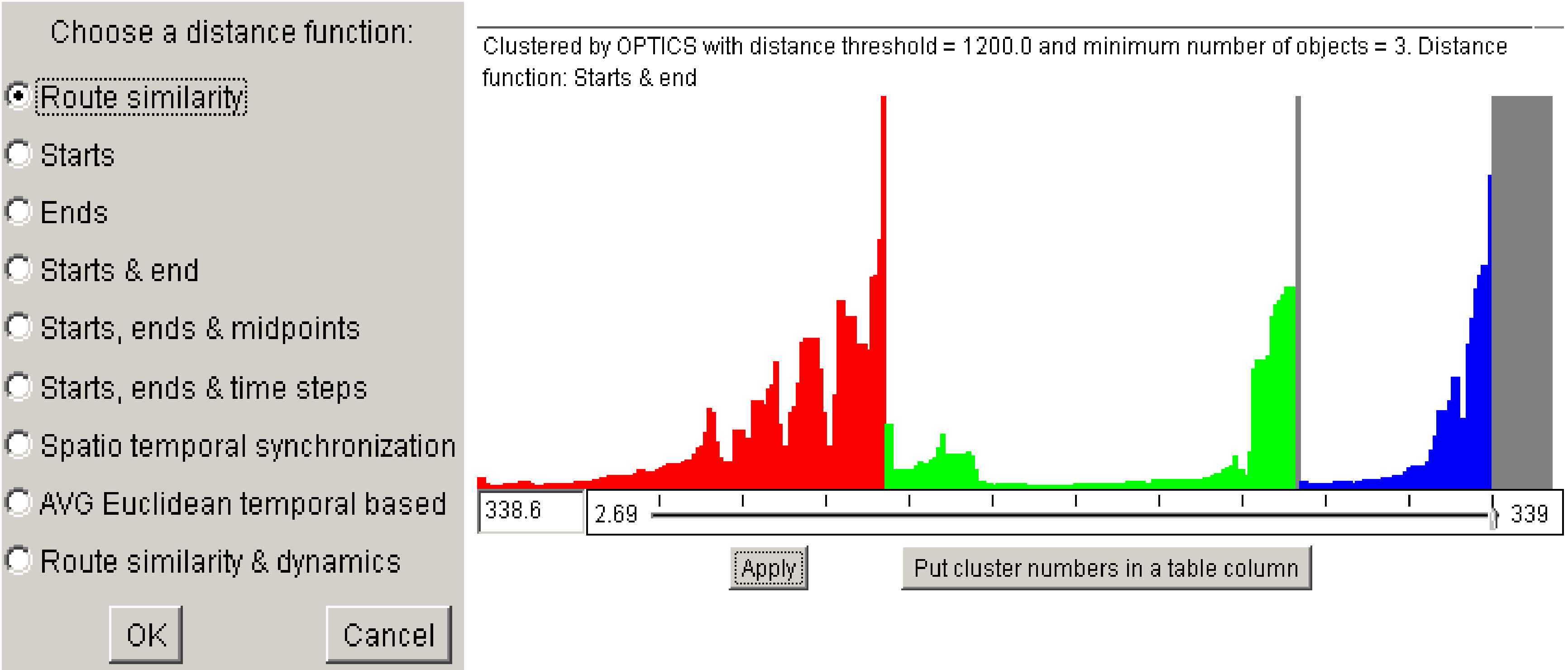
Temporal focusing

- Different time intervals can show different behaviours
  - E.g.: objects that are close to each other within a time interval can be much distant in other periods of time
- The time interval becomes a parameter
  - E.g.: rush hours vs. low traffic times
- Already supported by the distance measure
  - Just compute $D(\tau_1, \tau_2)\,|_{T}$ on a time interval $T' \subseteq T$
- Problem: significant T' are not always known *a priori*
  - An automated mechanism is needed to find them

Temporal focusing

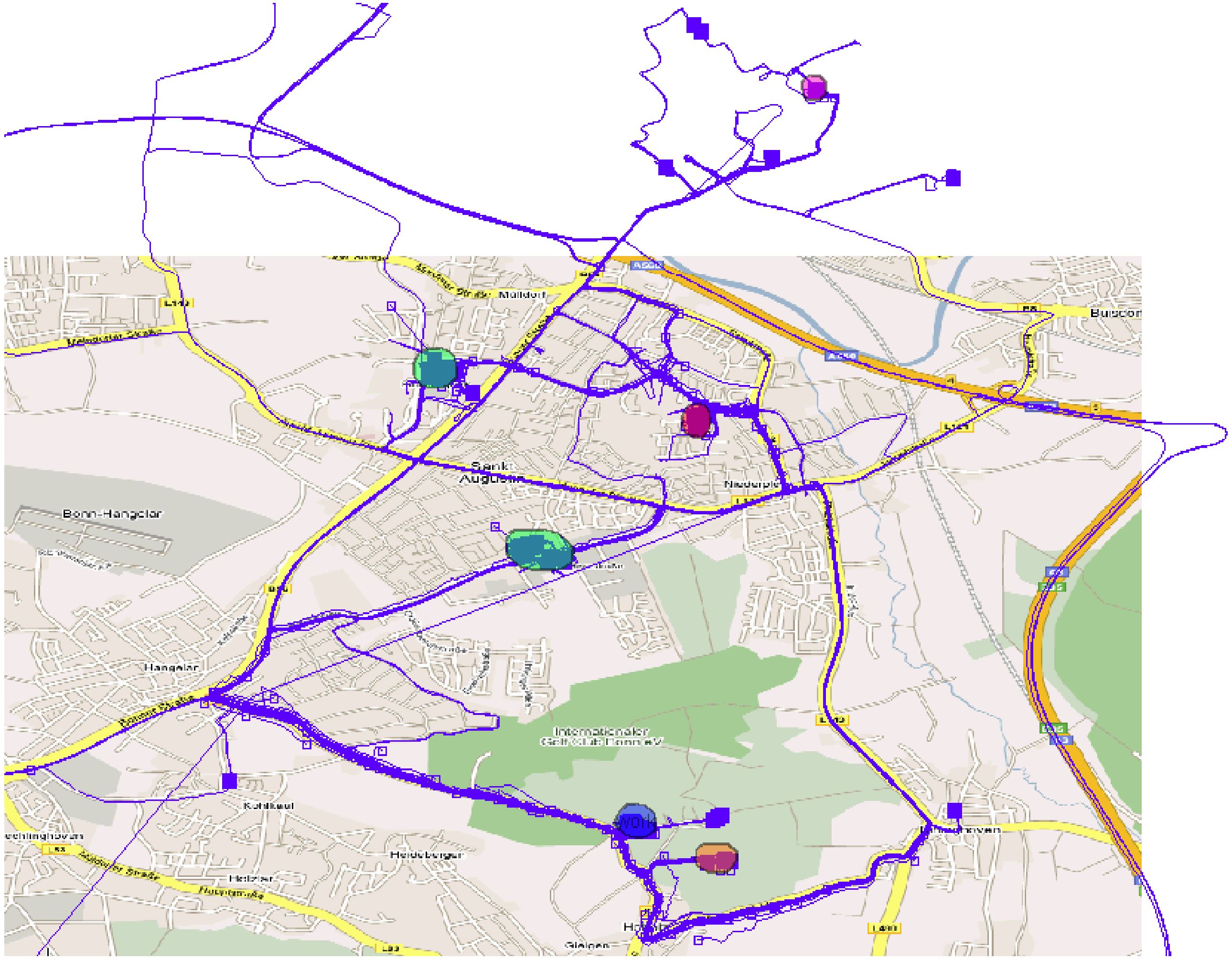
1. Provide a notion of **interestingness** to be associated with time intervals
   - Defined in terms of **estimated quality** of the clustering extracted on the given time interval
2. Formalize the **Temporal focusing** task as an optimization problem
   - Discover the time interval that maximizes the interestingness measure

Interactive density-based trajectory clustering

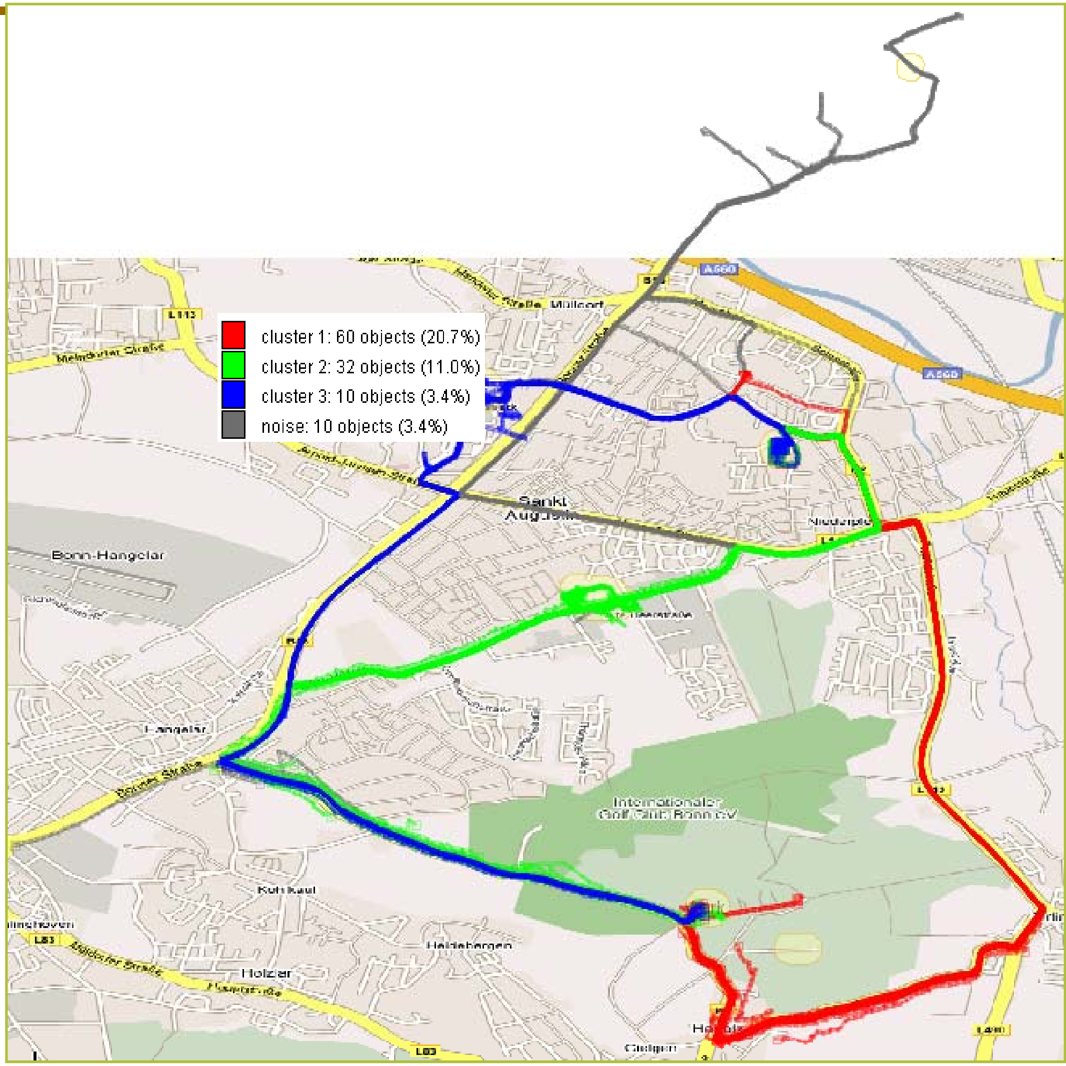


- More trajectory distance functions
- Rinzivillo, Pedreschi, Nanni, Giannotti, Andrienko, Andrienko. **Visually-driven analysis of movement data by progressive clustering**. J. of Information Visualization, 2008

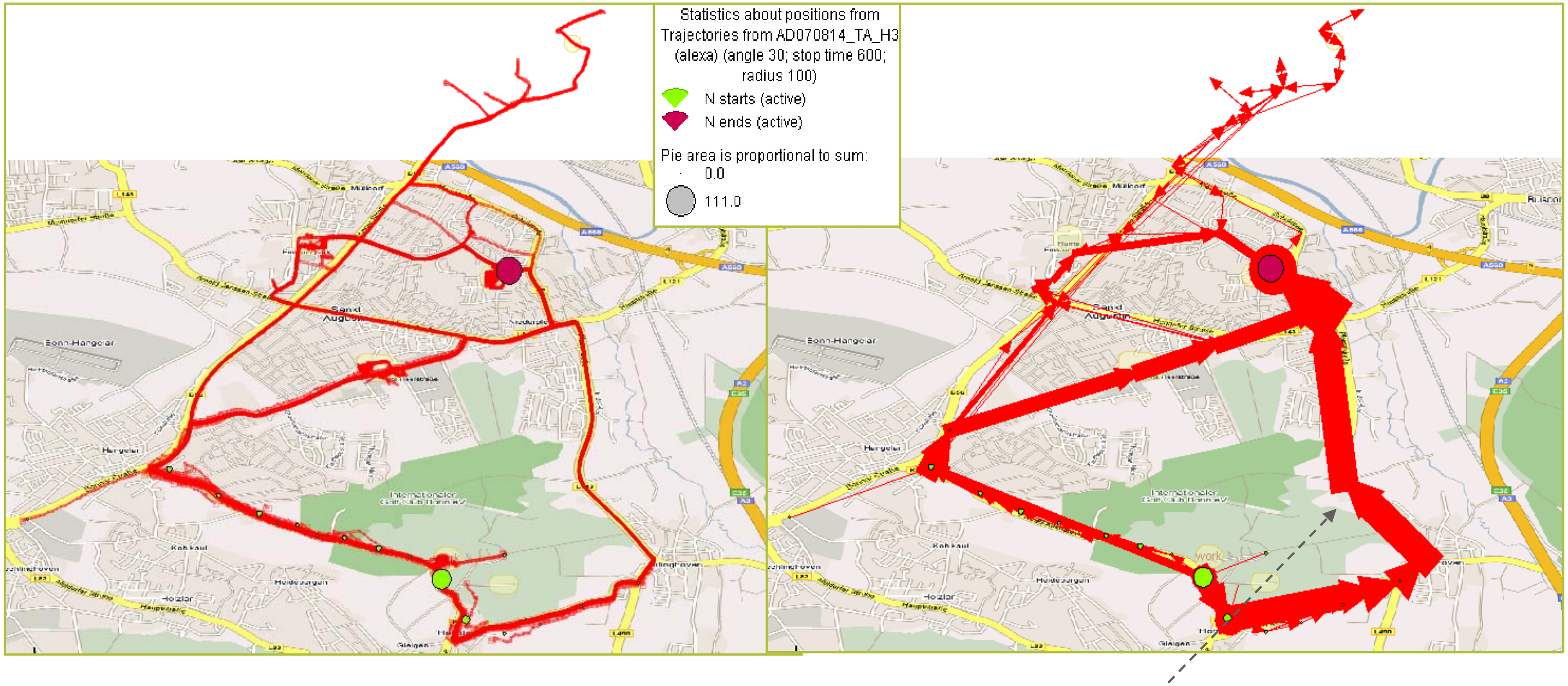
Looking for frequent stops & moves



Clusters of typical trips

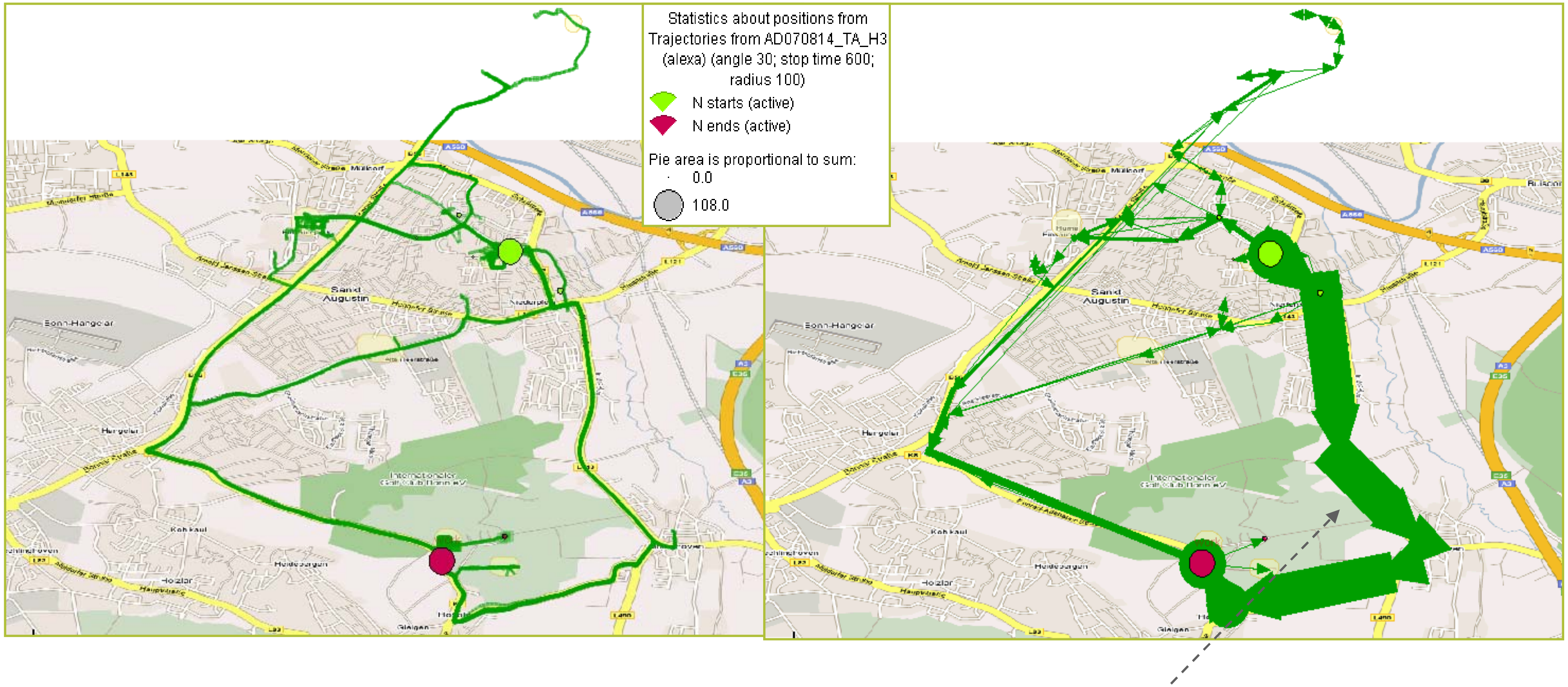


Cluster 1: from work to home



Observation: the eastern route is chosen more often

Cluster 2: from home to work



Observation: the eastern route is chosen much more often

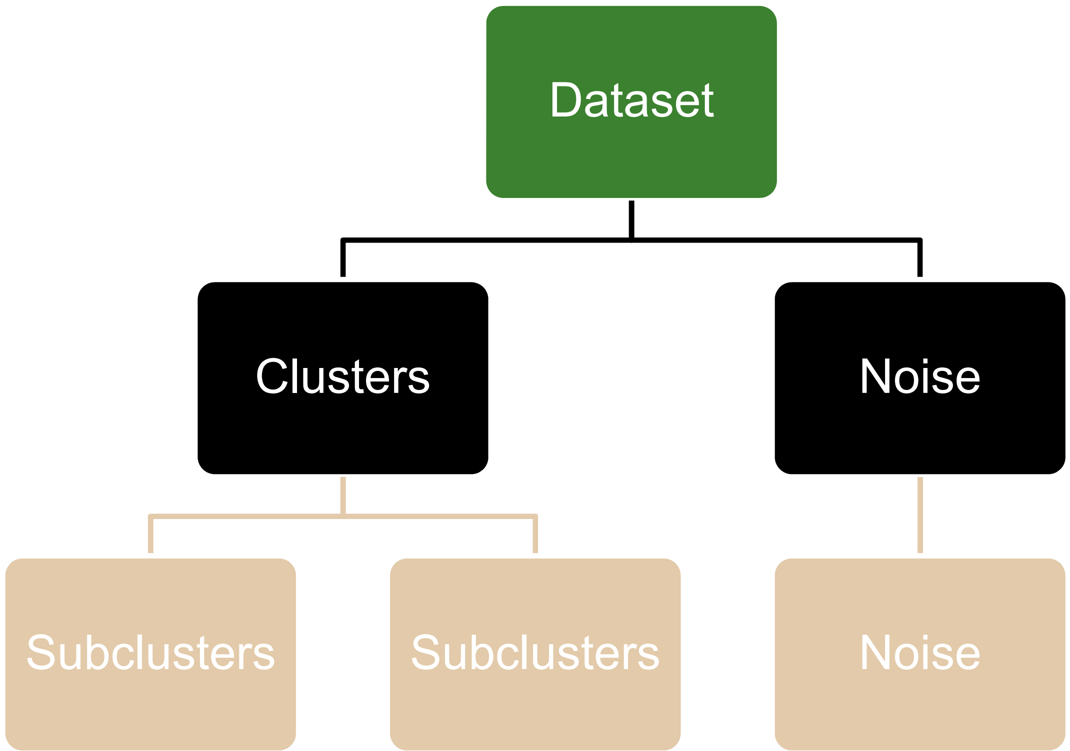
Progressive clustering

- Provide the analyst with a library of distance functions, each with a clear meaning
- Step refined analysis through the successive application of several distance measures
  - Start with simple and efficient measures (common ends)
  - Refine the obtained clusters with more sophisticated functions

Process Overview

Simple and very efficient distance measure

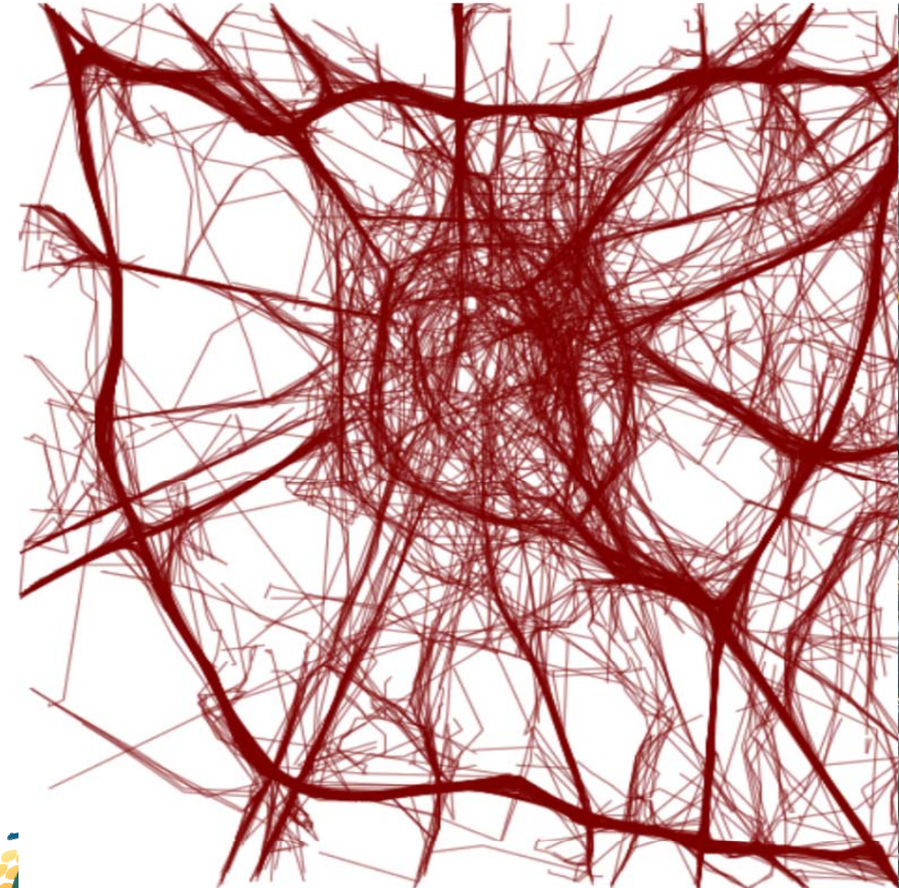
More selective and particular distance functions (or more restrictive parameters)

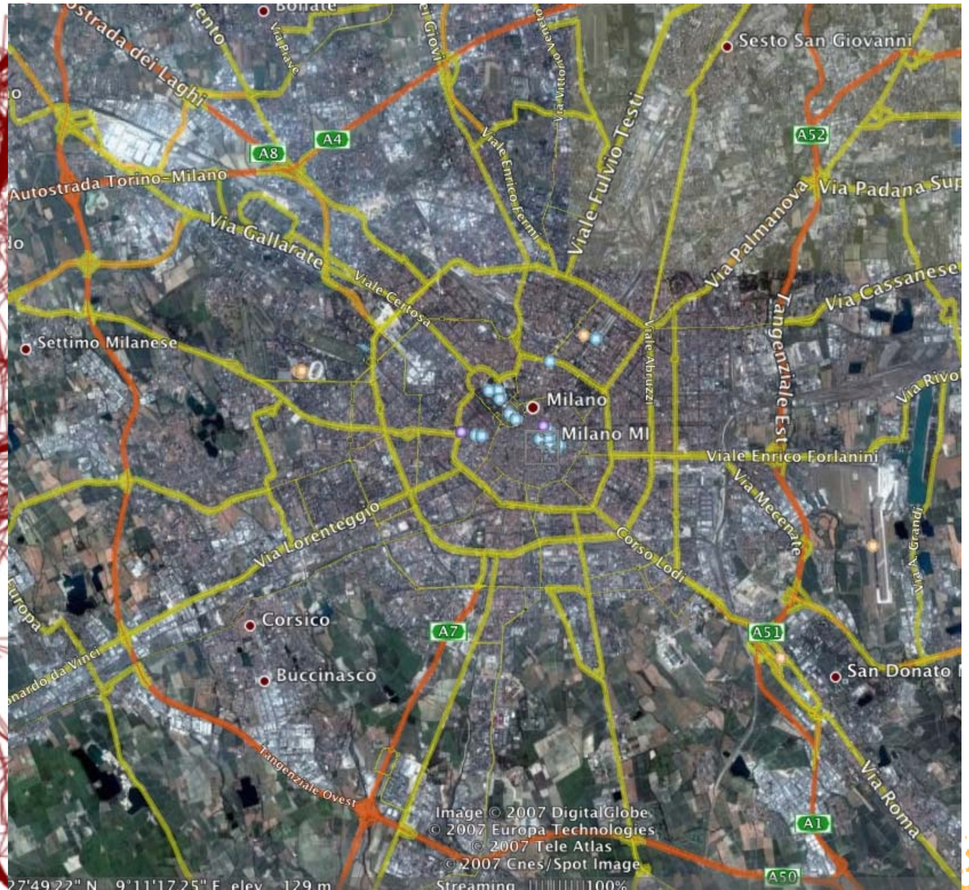


Mobility data analysis on a realistic GPS dataset

- WIND Telecomunicazioni spa (major telecom provider, GeoPKDD partner)
  - GSM data (Handover data: aggregated flows between adjacent cells)
- Other collaborations:
  - Comune di Milano, Mobility Agency
  - Infoblu and OctoTelematics (GPS receivers on board of cars with special insurance contract)
- Experience on a a dataset of
  - 2 M positions,
  - 17 K vehicles,
  - 200 K trajectories

MILANO: data on the map



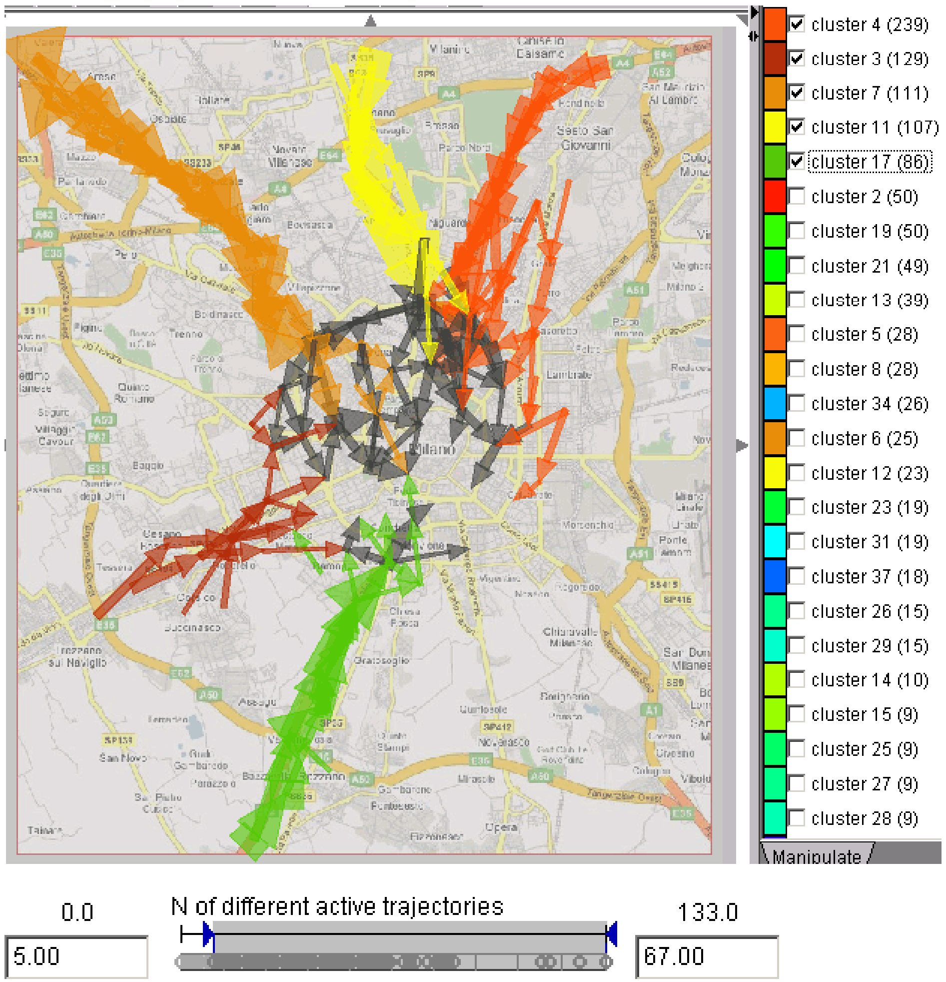


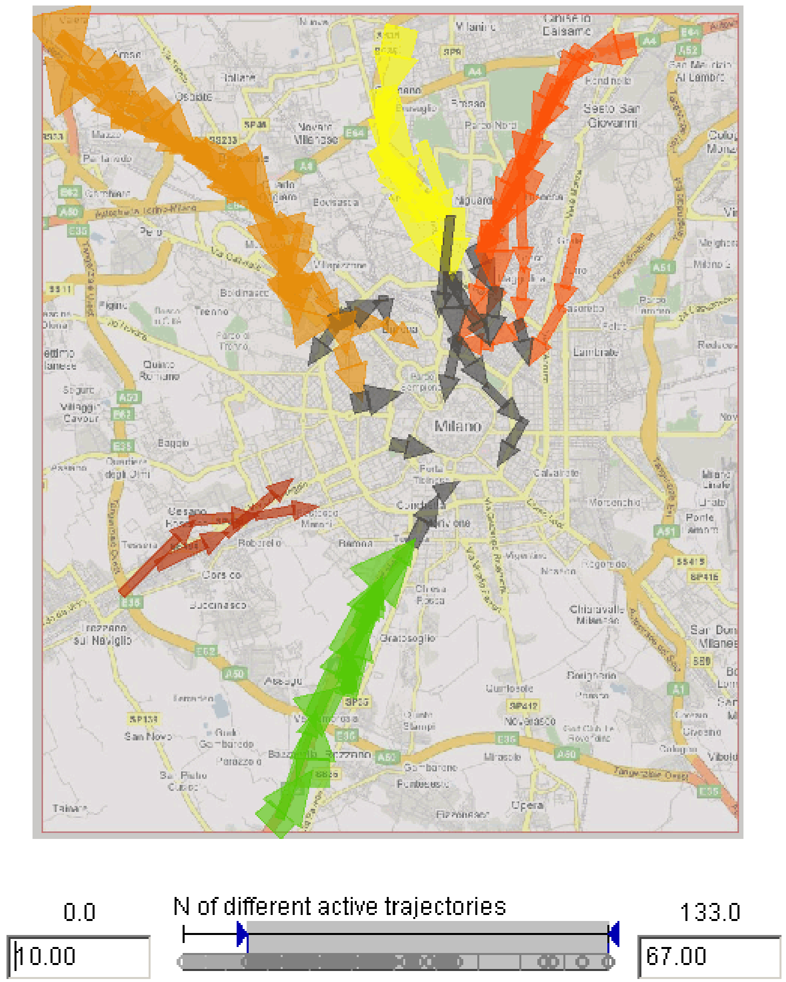
Progressive clustering

- First, create a large clusters of trajectories using the “common ends” distance function,
- Concentrate on the (big) cluster of inward trajectories (routes towards the city center)
- Refine by creating subclusters using a more sophisticated distance function (route similarity)

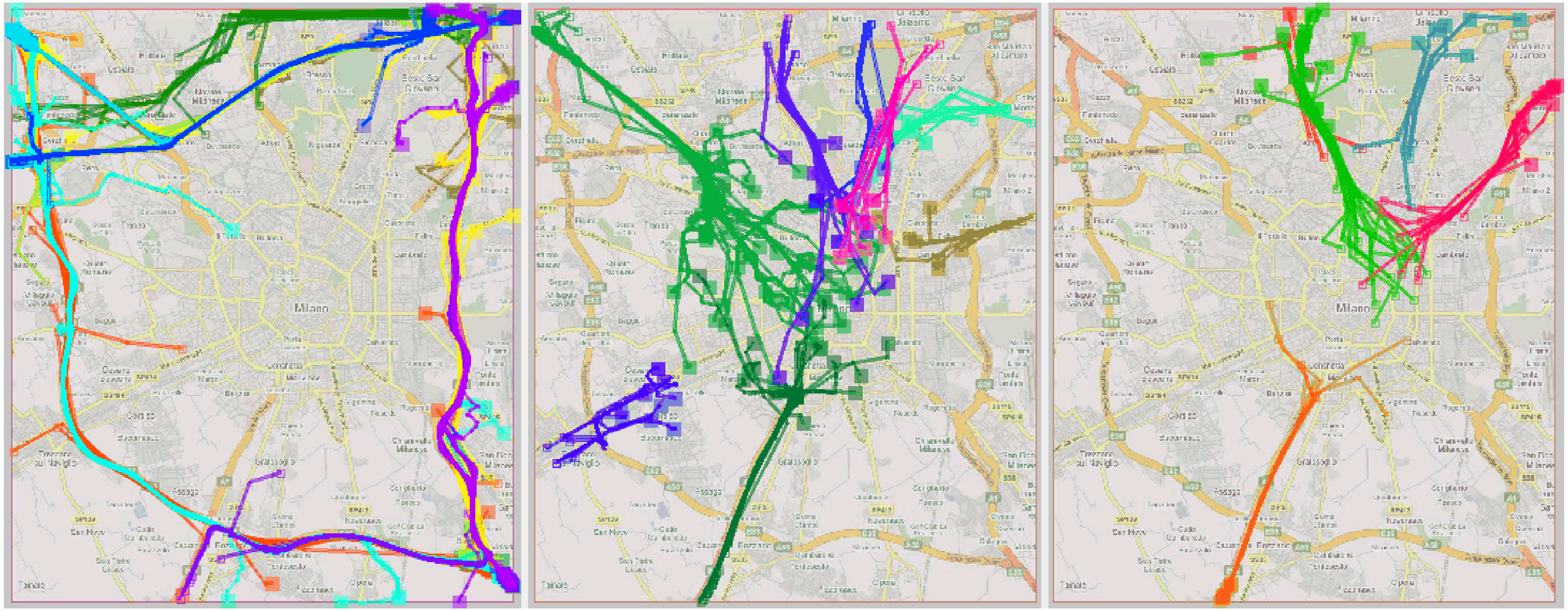
5 biggest (sub-)clusters of trajectories towards the city centre

Dark grey: moves occurring in trajectories from several clusters





Clustering trajectories on "route similarity"



Left: peripheral routes; middle: inward routes; right: outward routes.

- Rinzivillo, Pedreschi, Nanni, Giannotti, Andrienko, Andrienko **Visually-driven analysis of movement data by progressive clustering**. J. of Information Visualization, 2008

Challenges of visually-driven clustering

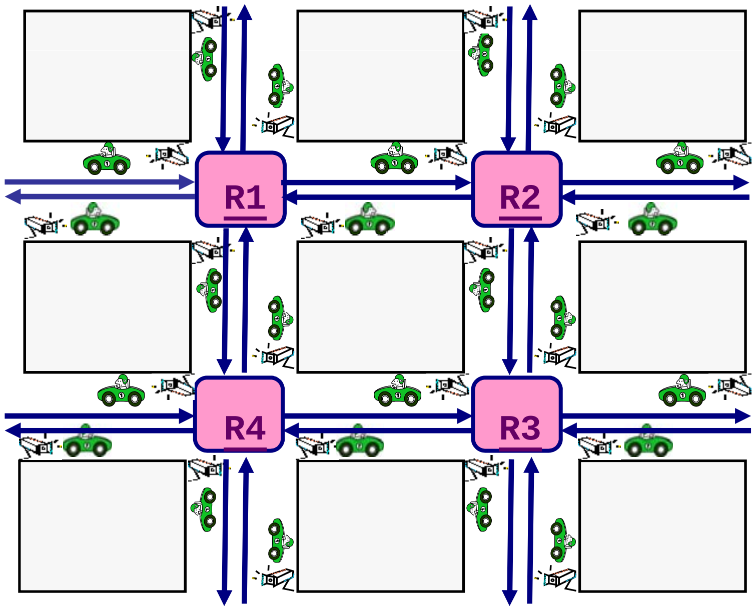
- Progressive refinement through visually-driven exploration
  - Progressively complex similarity functions
- Scalability
  - Index structures to support efficient neighborhood queries for trajectory clustering (Nanni, Pedreschi, Pelekis, Theodoridis, 2008)
  - Progressive clustering by sampling
- Incremental clustering and concept drift

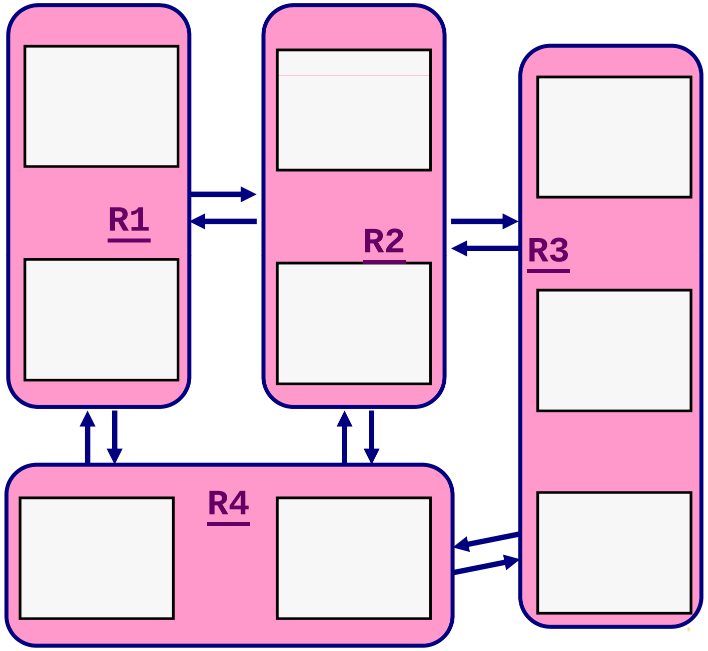
Traffic mining on road network

Mining (typically clustering) of aggregate traffic data over road networks

Network Traffic

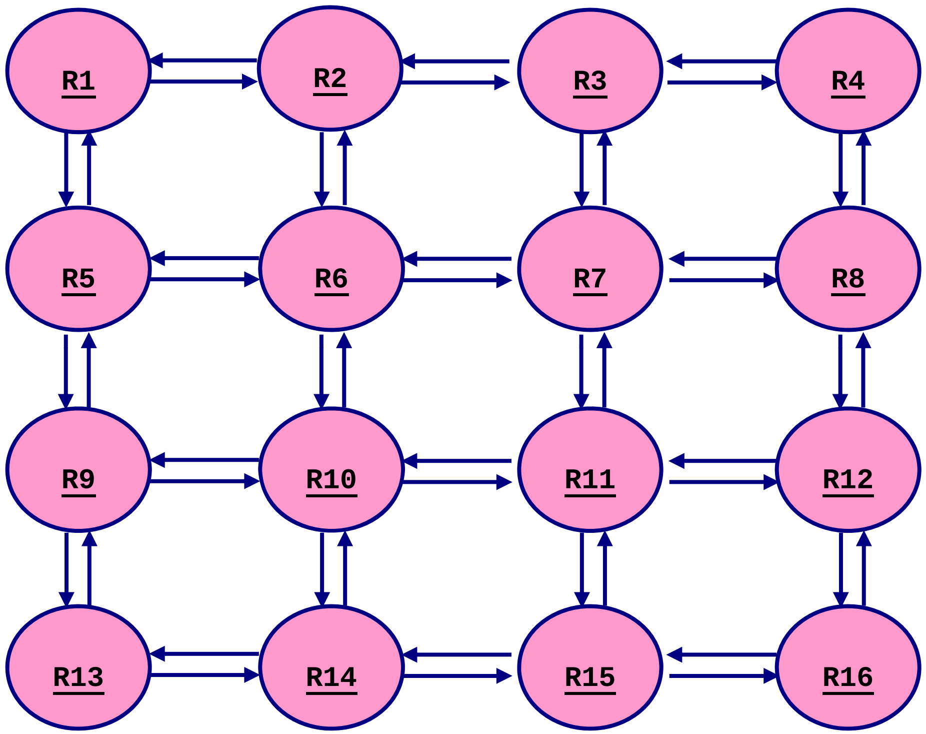
- Consider a fixed network consisting of a set of non-overlapping regions. Regions could be
  - road intersections (e.g. Via del Corso – Via del Tritone)
  - landmarks of interest (e.g. Colosseo, Parlamento)
  - or even greater areas (e.g. Centro Storico Roma)





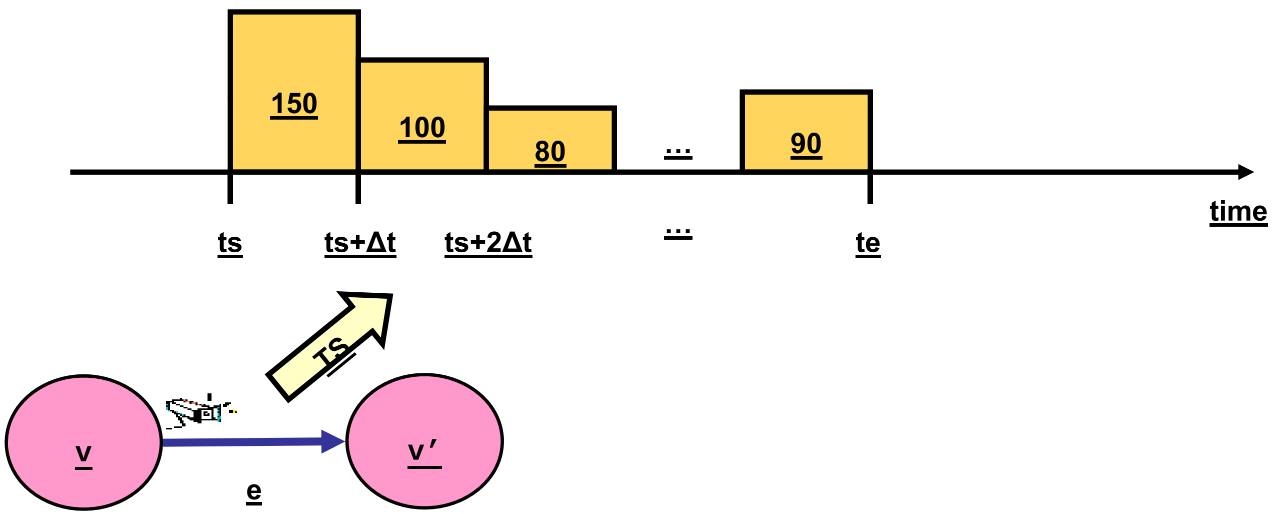
Network Graph

- The network is modeled as a directed graph G=(V,E)
  - nodes V → regions
  - edges E → direct connections between regions



Capturing traffic through sensors

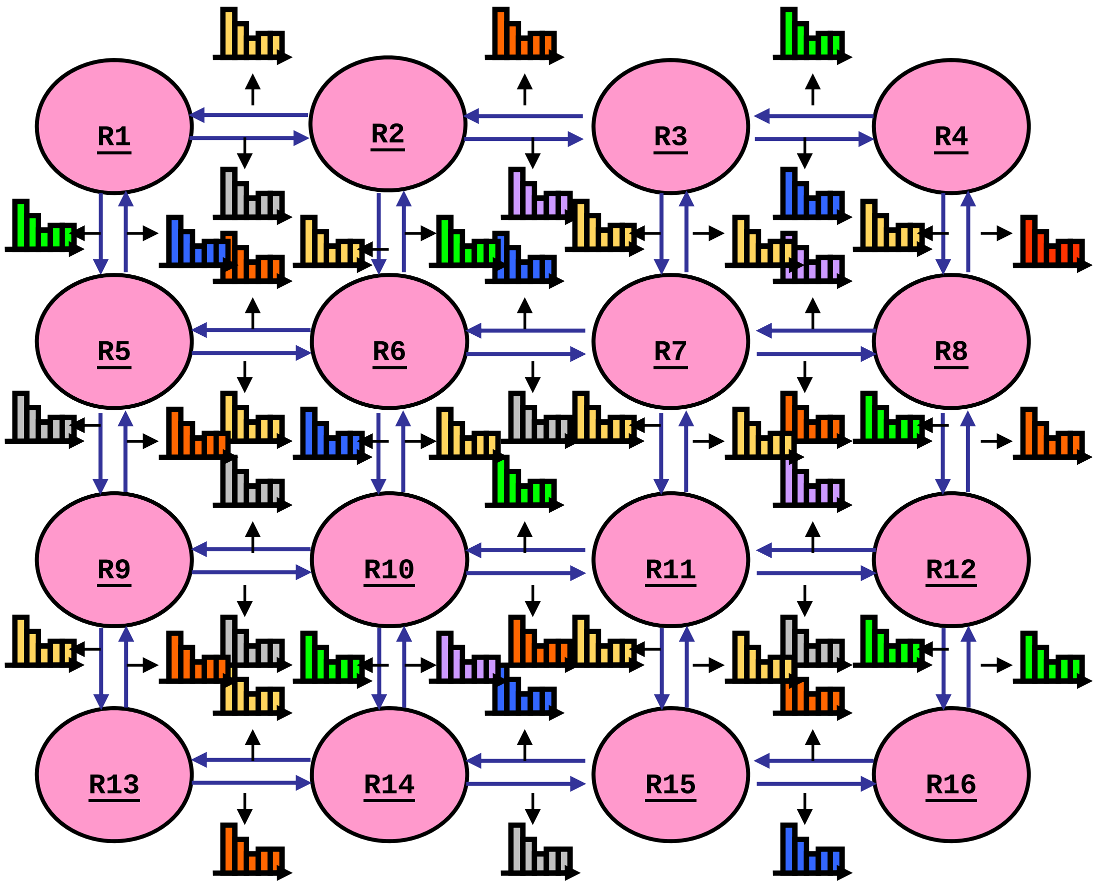
- Each edge e=(v, v') is equipped with sensor technology that captures the movement from region v to region v'.
- Definition: The **traffic series of a sensor** $s \in S$ during a time period $[t_s, t_e]$ consists of the number of cars passed through this sensor during this period, recorded at Δt intervals and ordered in time:
  - $TS_s = \{v_i, t_i\}$, $t_s \le t_i \le t_e$, $\Delta t = t_i - t_{i-1}$ the *transmission rate* of the sensor



134

Network Traffic

- **Traffic series of the network:** $TS = \{TS_s, s \in S\}$

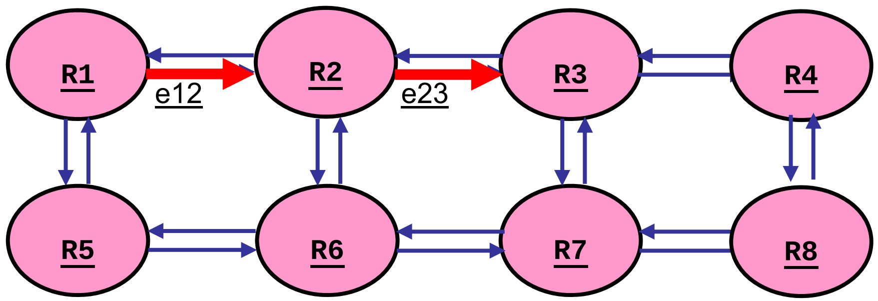


Works on Traffic Mining over Road Netwoks

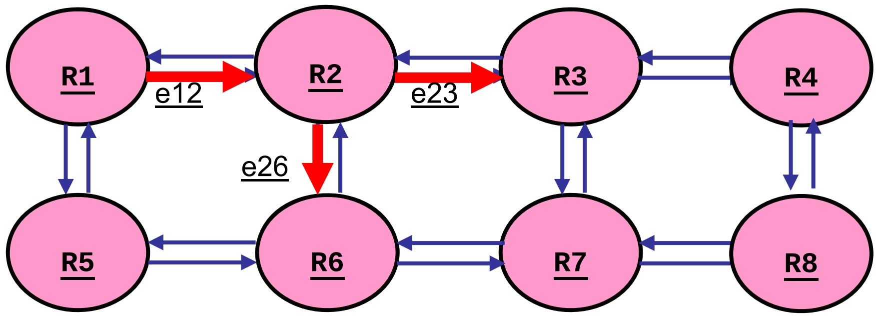
- Xiaolei Li, Jiawei Han, Jae-Gil Lee and Hector Gonzalez. Traffic Density-Based Discovery of Hot Routes in Road Networks. STD 2007 (Advances in Spatio-Temporal Databases).
- Hector Gonzalez, Jiawei Han, Xiaolei Li, Margaret Myslinska, John Paul Sondag. Adaptive Fastest Path Computation on a Road Network: A Traffic Mining Approach. VLDB 2007
- Irene Ntoutsi, Nikos Mitsou, Gerasimos Marketos,Yannis Theodoridis. Mining Traffic Flow in a Road Network: How does the traffic flow? Int. Journal of Business Intelligence and Data Mining, 2008

Traffic relationships

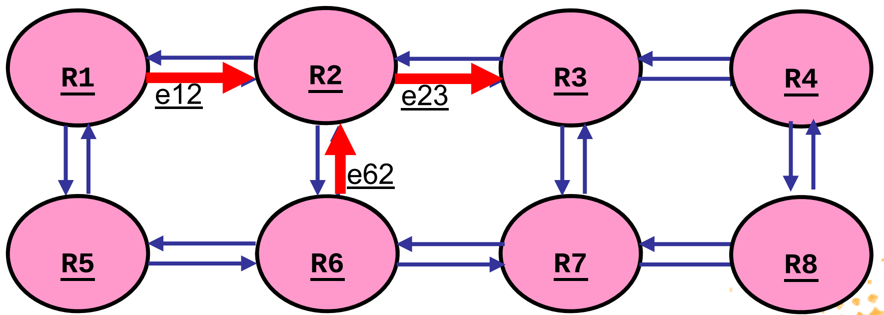
- **Traffic propagation**
  - traffic from e_{12} is propagated to e_{23}
  - This might indicate objects that continue moving in a highway



- **Traffic split/ spread**
  - traffic from e_{12} is split into e_{23} and e_{26}
  - This might indicate objects that leave a highway and follow different directions to their destination



- **Traffic merge**
  - traffic to e_{23} merges traffic from e_{12} and e_{62}
  - This might indicate objects that enter a highway from different directions



A three-level clustering algorithm

- A *divisive hierarchical clustering* algorithm to detect different behaviors of traffic flow

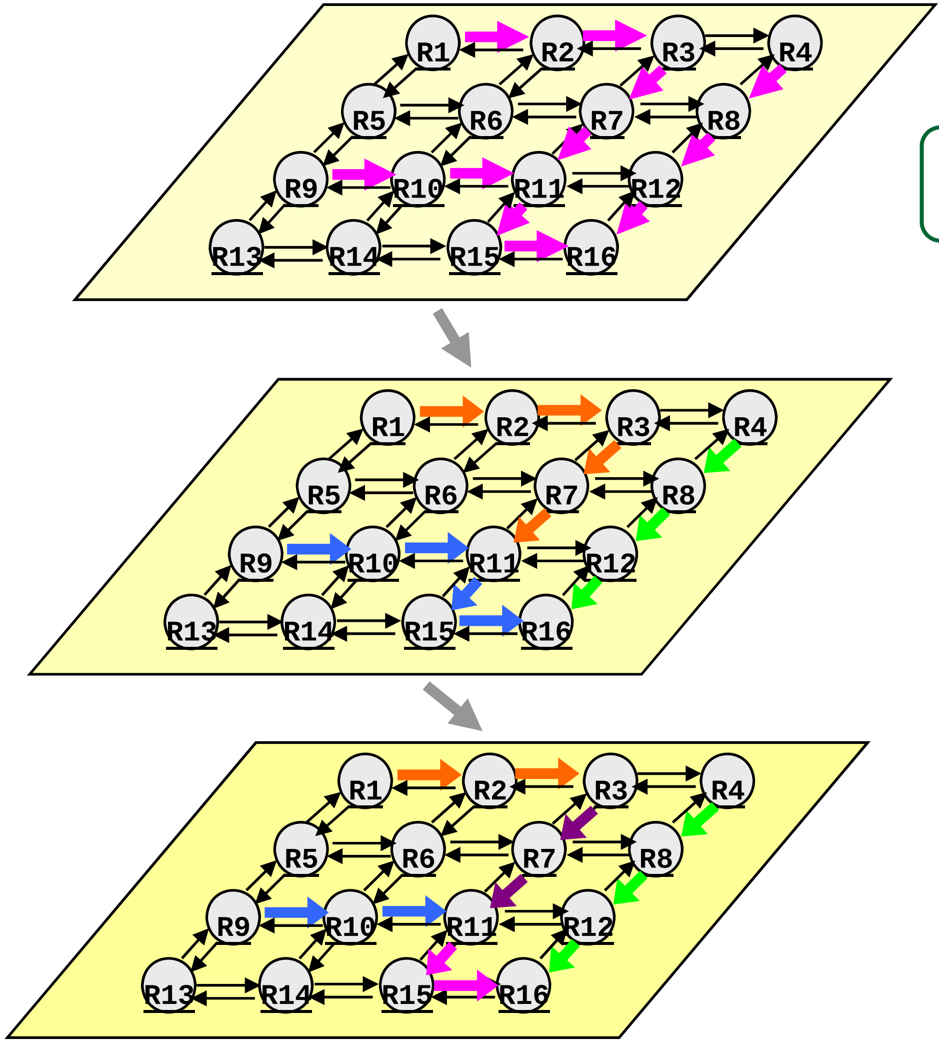
- Three different distance measures: $dis_{value}(e_1, e_2)$, $dis_{shape}(e_1, e_2)$, $dis_{struct}(e_1, e_2)$ capture different aspects of (dis-)similarity of traffic flow between two edges/ road networks:

  - edges with *similar traffic shape* // dis_{shape}

  - edges located *nearby* // dis_{struct}

  - edges with *similar traffic values* // dis_{value}

A hierarchical view of the traffic edges

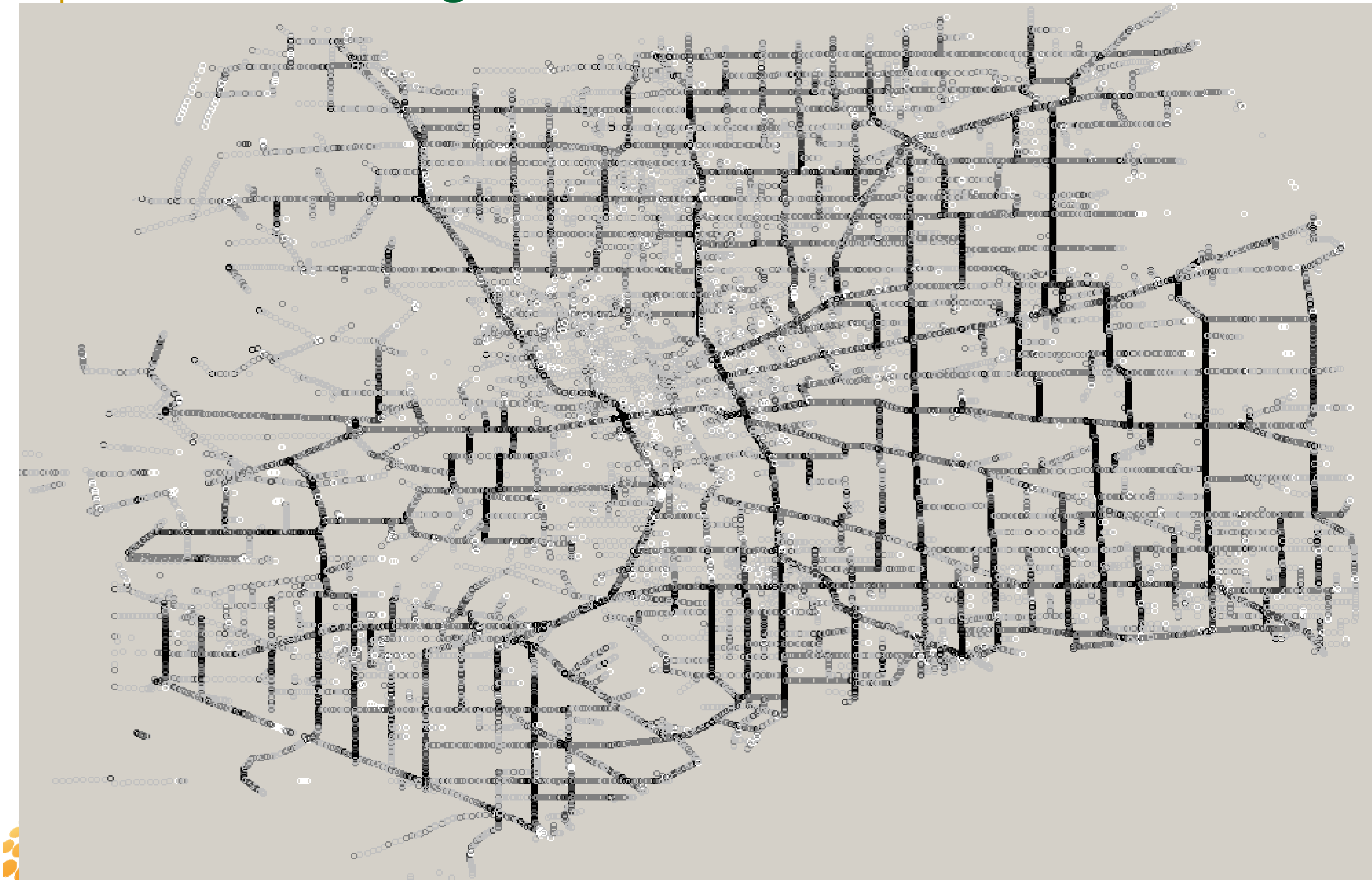


L1: edges with similar traffic shape

L2: edges with similar traffic shape that are also nearby in the network

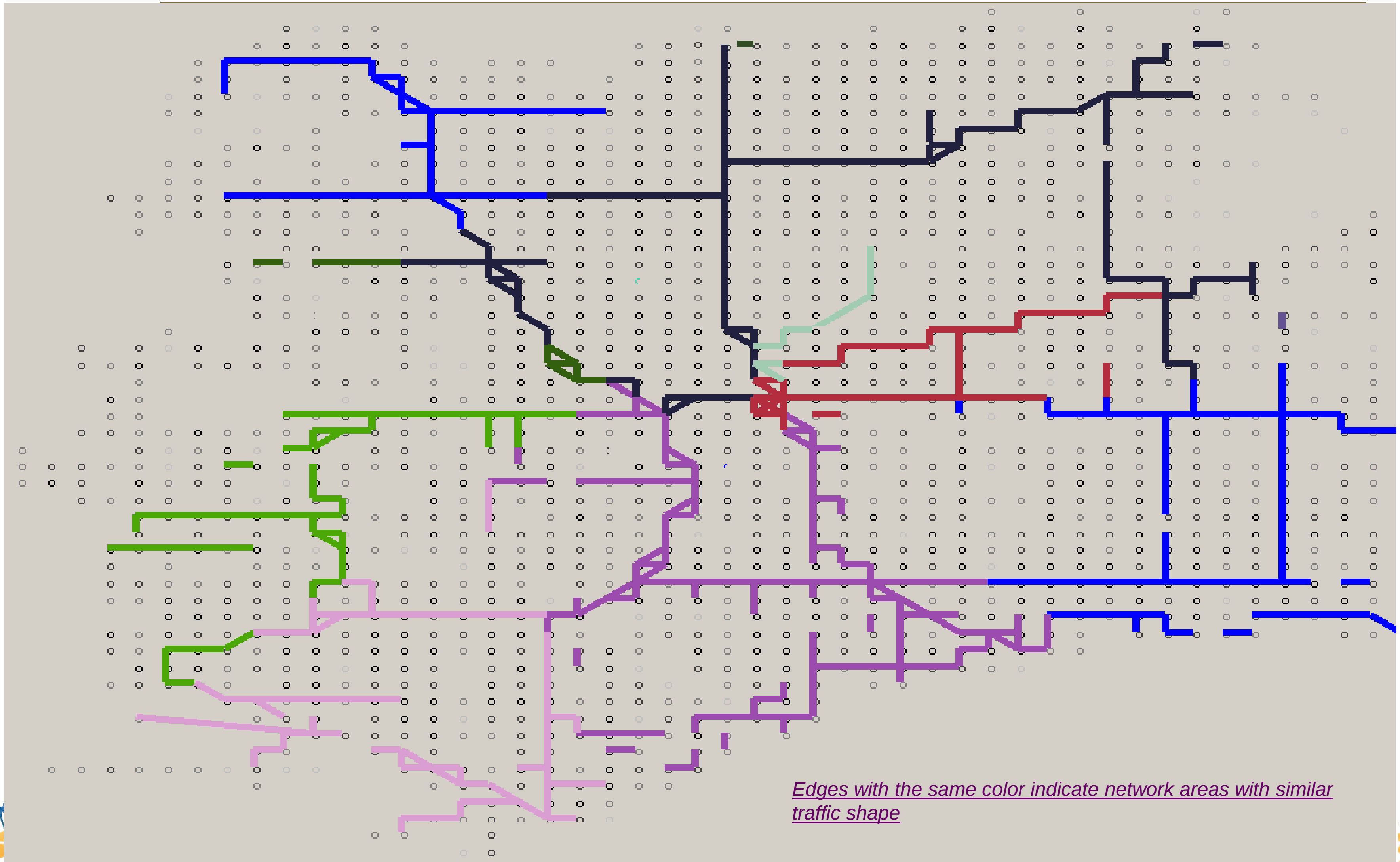
L3: edges with similar traffic values

The original traffic network

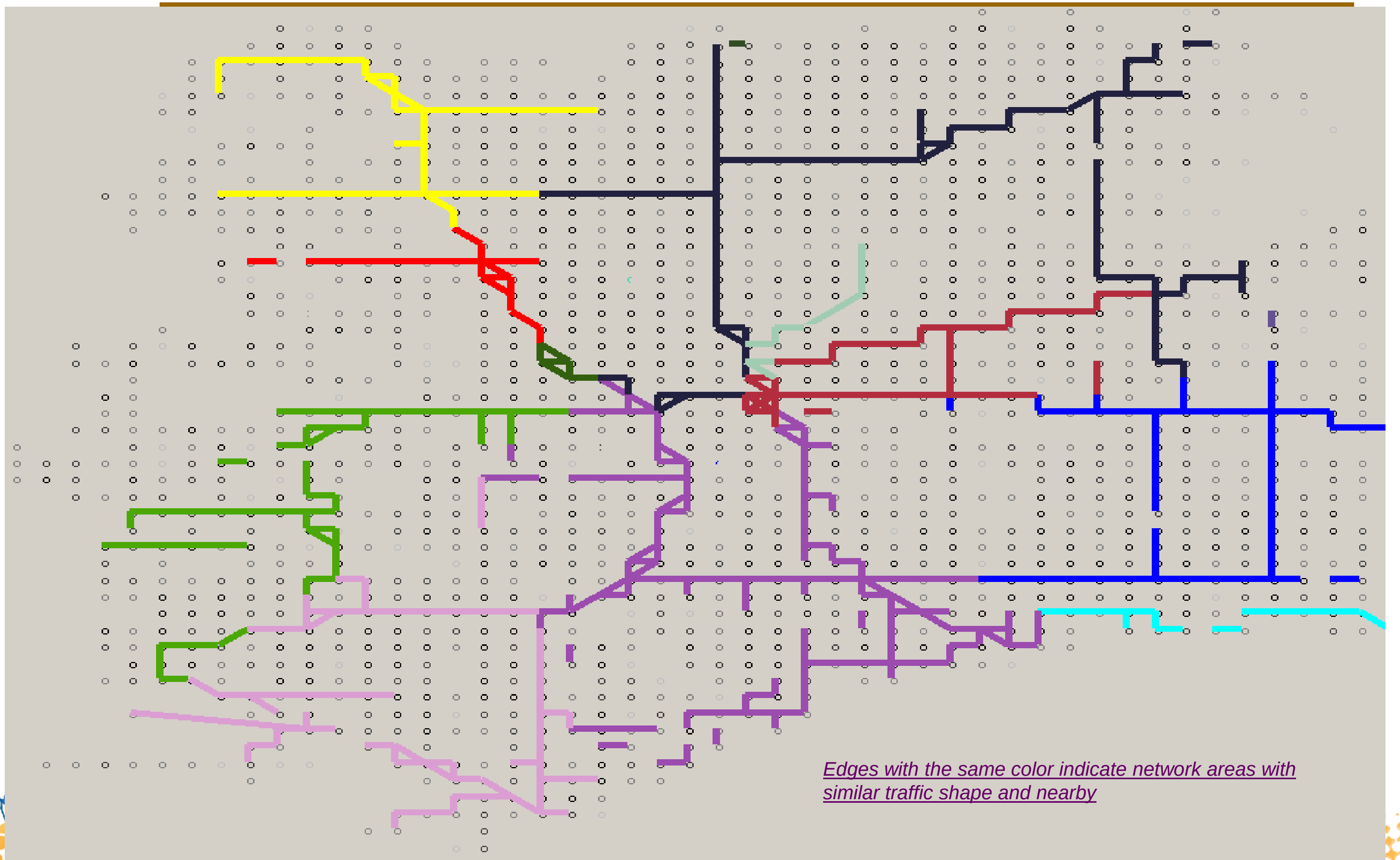


140

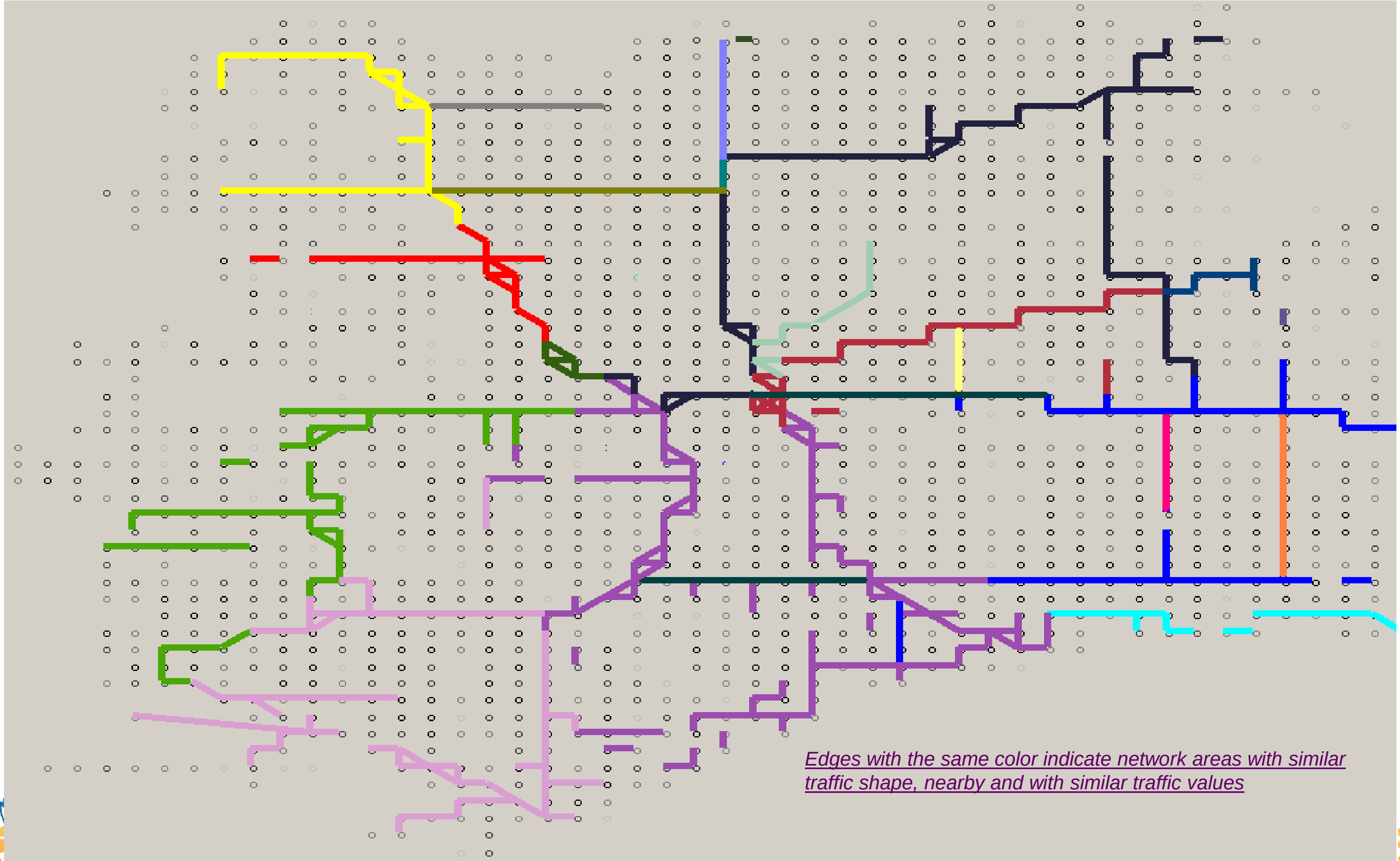
Clustering results – L_1



Clustering results – L_2



Clustering results – L_3



February 8, 2008 5:56 PM PST

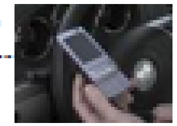
Nokia turns people into traffic sensors

Posted by Erica Ogg | 8 comments

UNION CITY, Calif.--On a cool, overcast morning in the parking lot of a Lowe's hardware store, 100 UC Berkeley students lined up in rows ready to jump into a bevy of idling vehicles.

With media and VIPs from companies like Nokia, Navteq, General Motors, BMW, and CalTrans looking on, wave after wave of students left the parking lot to drive a 10-mile stretch of the nearby 880 freeway as part of a large-scale experiment to test how cell phones can monitor and predict traffic.

The test, conducted all day Friday, was put on by the California Center for Innovative Transportation (CCIT) as a joint project between Nokia, CalTrans, and Berkeley's Department of Civil and Environmental Engineering.



Video: Using cell phones to track traffic

Each student car was issued a Nokia N95 phone with GPS and special traffic-monitoring software developed by Nokia's Palo Alto, Calif.-based research lab--plus a Bluetooth headset. As the students drove the freeway, the phone sent data about each car's speed and position back to the company's research facility. The data is compiled and used to predict traffic patterns and help drivers get where they need to be quickly. Nokia hopes that one day the system could be a significantly cheaper way to track traffic than the permanent sensors installed in roadways or next to them because it uses equipment most people already own: cell phones.

Alex Bayen, a professor of civil and environmental engineering and lead researcher on the project for Berkeley, called the experiment "a glimpse into the future of traffic information

About News Blog

Recent posts on technology, trends, and more.

Subscribe to this blog
Click this link to view this blog as XML.

Add this feed to your online news reader

- Add to Google
- Add to my Yahoo
- Add to MSN
- Add to Bloglines
- Add to Newsgator

An archaeology of the present

- 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.

- **Mobility data mining**

From opportunities to threats

- Personal mobility data, as gathered by the wireless networks, are extremely sensitive
- Their disclosure may represent a brutal violation of the privacy protection rights, i.e., to keep confidential
  - the places we visit
  - the places we live or work at
  - the people we meet
  - …

Privacy-preserving mobility data mining

The naive scientist's view

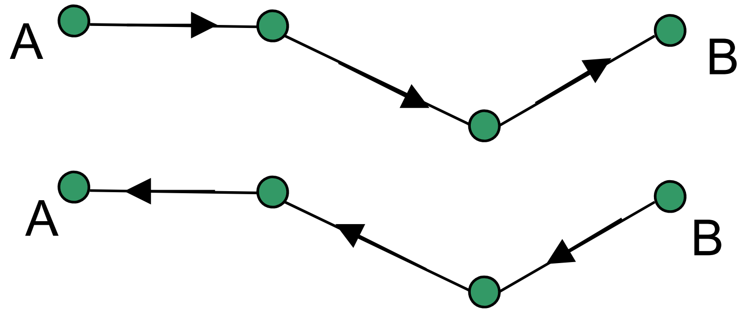
- Knowing the exact identity of individuals is not needed for **analytical** purposes
  - De-identified mobility data are enough to reconstruct aggregate movement behaviour, pertaining to groups of people.
- Reasoning coherent with European data protection laws: personal data, **once made anonymous**, are not subject to privacy law restrictions
- Is this reasoning correct?

Unfortunately not!

- Making data (reasonably)a nonymous is not easy.
- Sometimes, it is possible to reconstruct the exact identities from the de-identified data.
- Many famous example of re-identification
  - Dalenius …
  - Governor of Massachusetts' clinical records (Sweeney's experiment, 2001)
  - America On Line August 2006 crisis: user re-identified from search logs
- Two main sources of danger:
  - **Many observations** on the same "anonymous" subject
  - **Linking data**, after joining separate datasets

Spatio-temporal linkage in Mobility Data

Id:
34567



[almost every day mon-fri
between 7:45 – 8:15]

[almost every day mon-fri
between 17:45 – 18:15]

- By intersecting the phone directories of locations A and B we find that only one individual lives in A and works in B.
- Id:34567 = Prof. Smith
- Then you discover that on Saturday night Id:34567 usually drives to the city red lights district…

Basic ideas for anonymity preserving data analysis

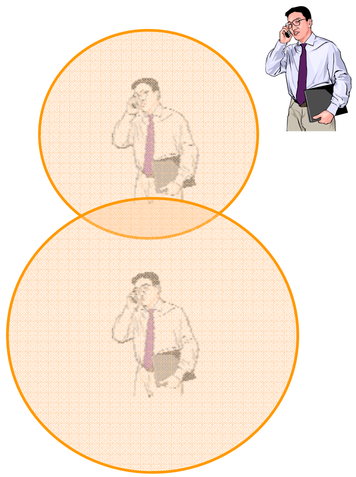
How do people (try to) stay anonymous?

- either by **camouflage**
  - pretending to be someone else or somewhere else
- or by **hiding in the crowd**
  - becoming indistinguishable among many others

Concepts for Location Privacy

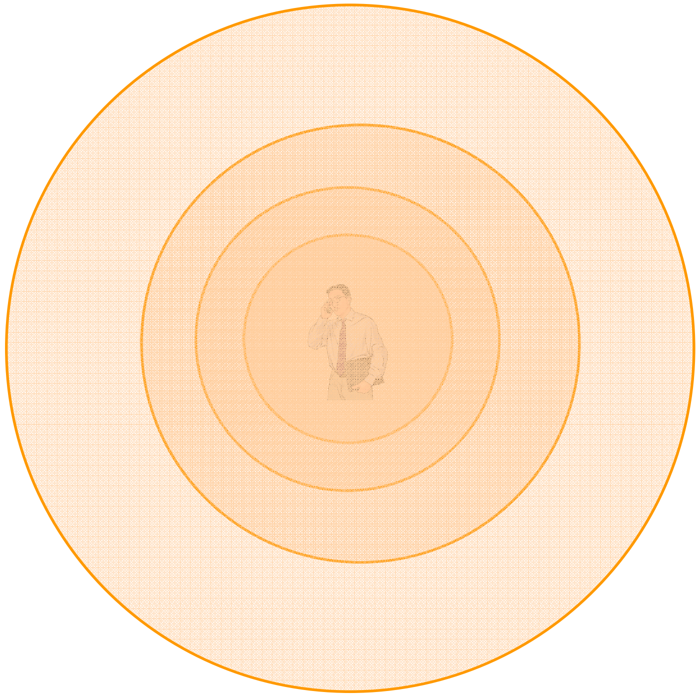
Location Perturbation – Randomization

- The user location is represented with a **fake** value
- Privacy protection is achieved from the fact that the reported location is false
- The accuracy and the amount of privacy mainly depends on how far is the reported location from the exact location



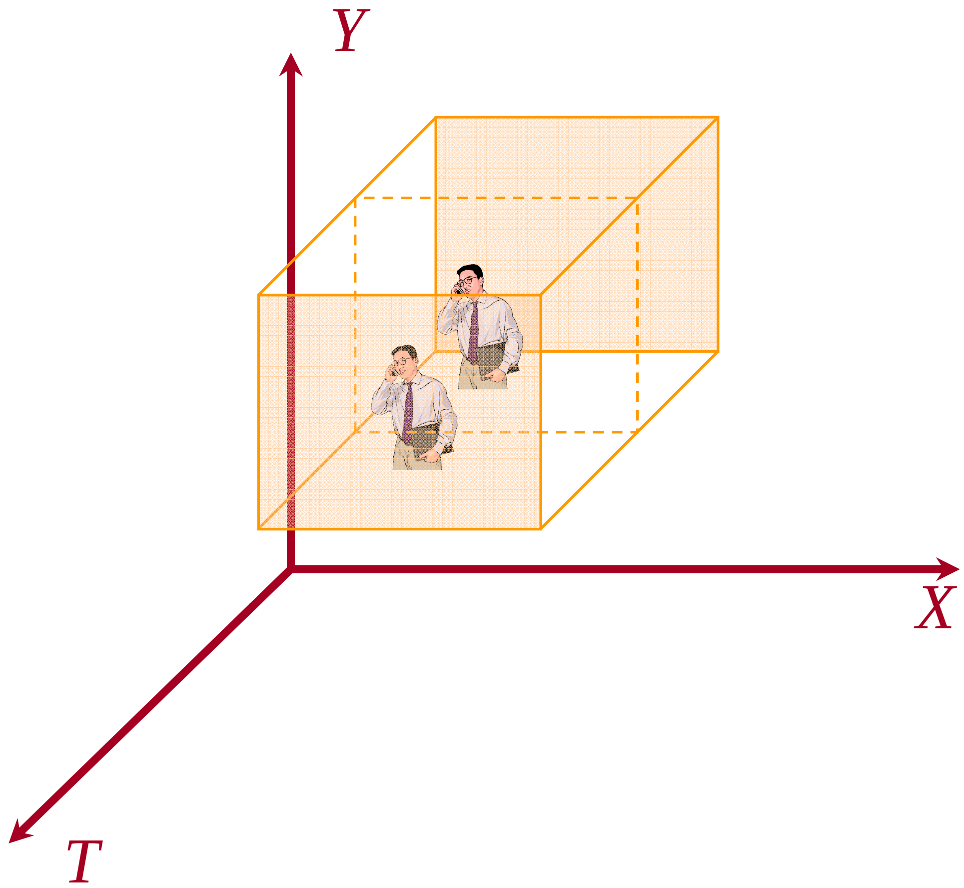
Concepts for Location Privacy
Spatial Cloaking – Generalization

- The user exact location is represented as a region that includes the exact user location
- An adversary does know that the user is located in the region, but has no clue where the user is exactly located
- The area of the region achieves a trade-off between user privacy and accuracy



Concepts for Location Privacy
Spatio-temporal generalization

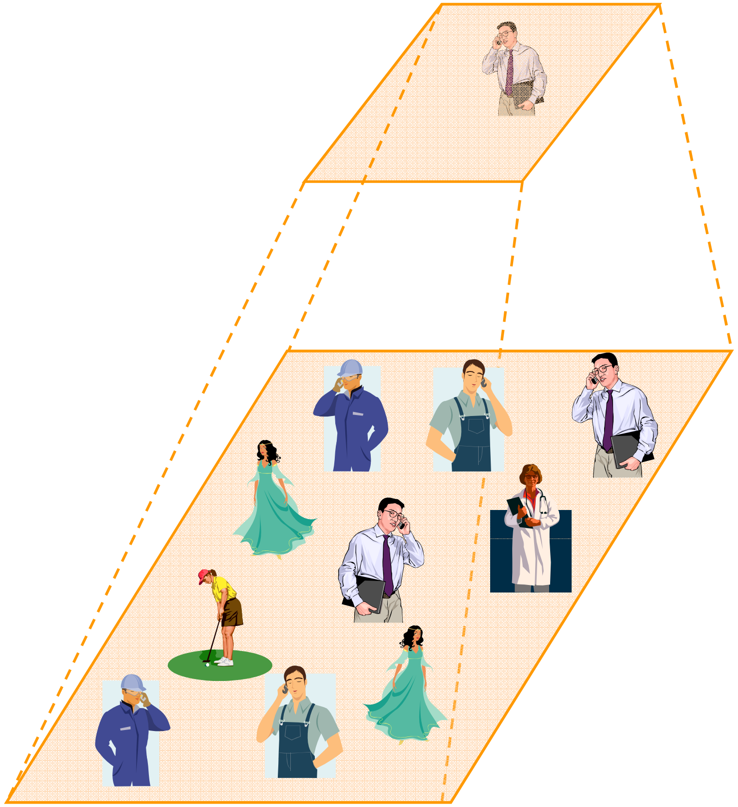
- In addition to the spatial dimension, generalize also the temporal dimension



Concepts for Location Privacy

***k*-anonymity**

- User's position is generalized to a region containing **at least *k* users**
- The user is indistinguishable among other *k* users
- The area largely depends on the surrounding environment.
- A value of *k* =100 may result in a very small area downtown Hong Kong, or a very large area in the desert.



10-anonymity

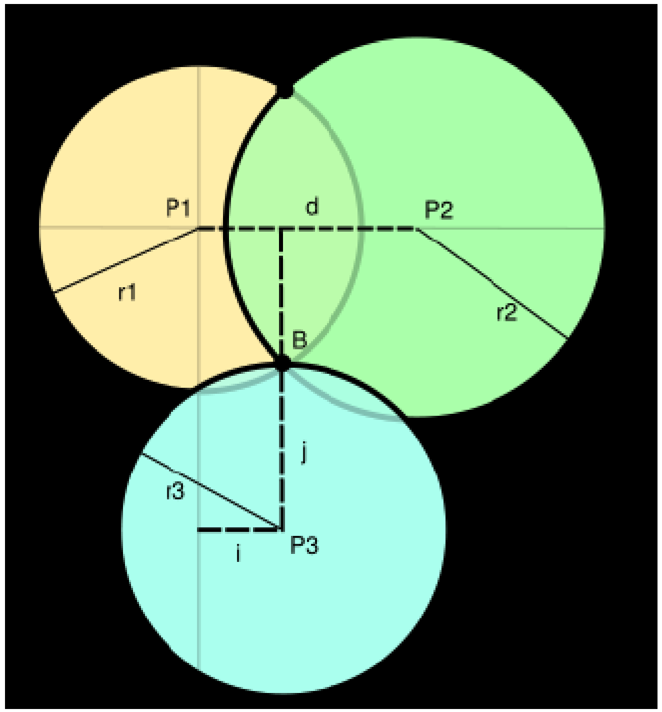
Privacy- preserving spatio-temporal data mining

Trajectory randomization is risky!

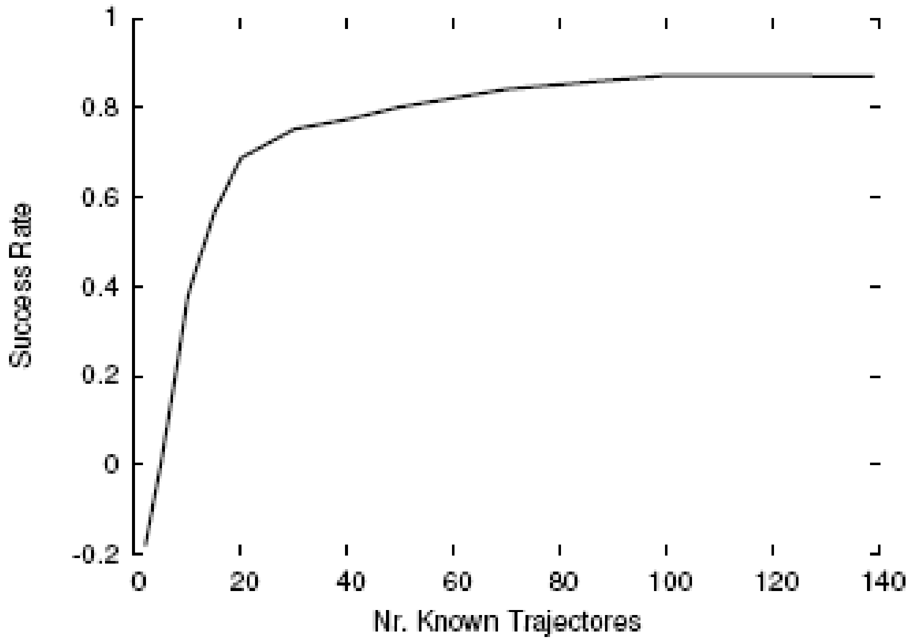
Trajectory anonymization

A subtle re-identification attack

- Disclosure Risks of Distance Preserving Data Transformations
  - Erkay Savas, Yucel Saygin, Emre Kaplan, and Thomas B. Pedersen (Sabanci Univ., Istanbul)
- What if the attacker knows:
  - Some trajectories
  - All mutual distances
- Hyper-lateration
  - Works in d dimensions given d + 1 points
  - If known trajectories are few, then approximate!



Red: true traj Blue: approx traj



(b) Success-rate vs. number of known trajectories (Each sample is the average of 60 experiments run for 50.000 iterations).

Privacy- preserving spatio-temporal data mining

Trajectory randomization is risky!

Trajectory anonymization

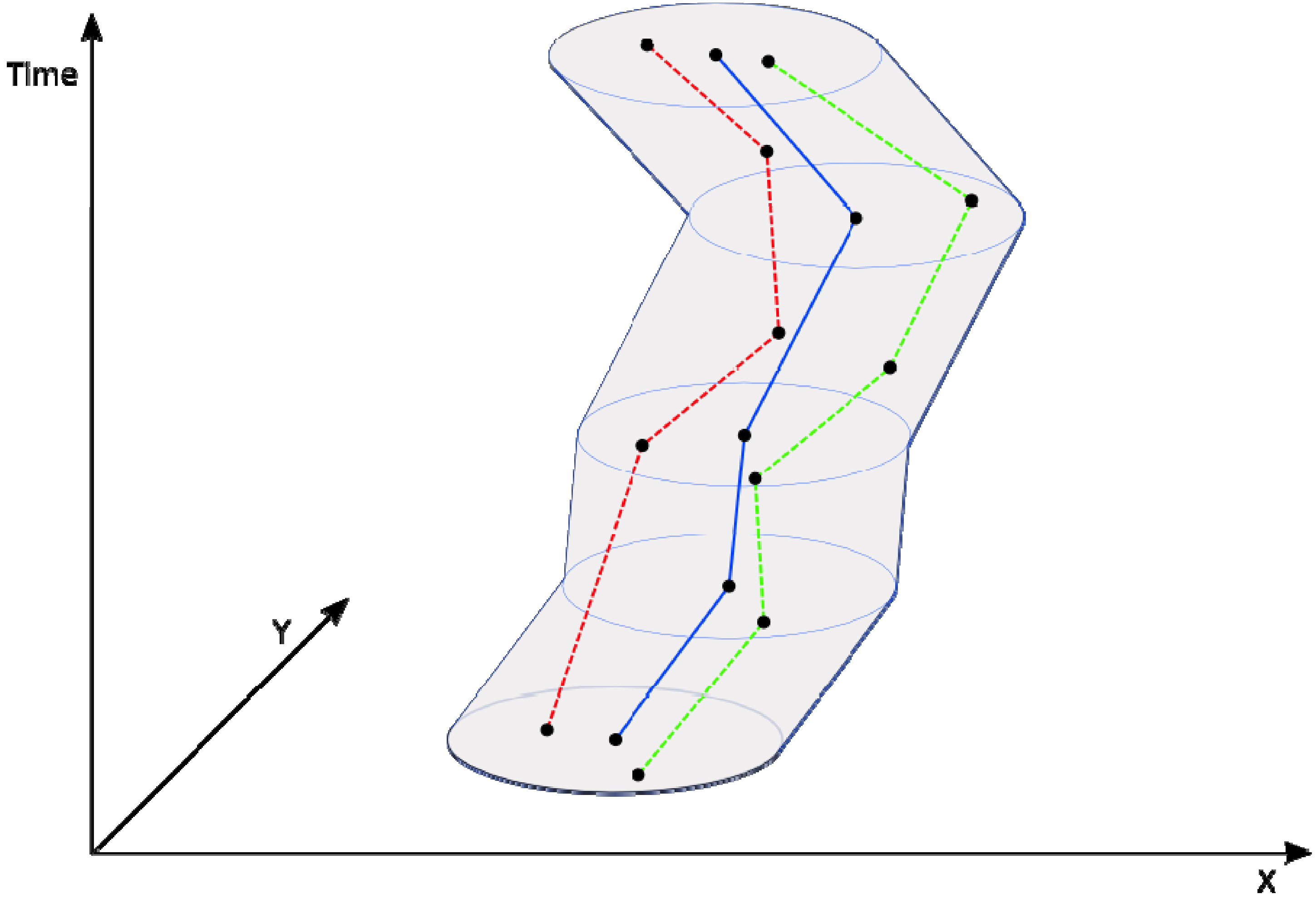
Trajectory anonymization

- Several variants developed in GeoPKDD:
  - Abul, Bonchi, Nanni (Pisa KDD LAB). Int. Conf. Data Engineering ICDE 2008
  - Nergiz, Atzori, Saygin (Sabanci Univ. + Pisa KDD LAB). 2007 (submitted)
  - Gkoulalas-Divanis, Verykios (Univ. Thessaly). 2007 (submitted)
  - Pensa , Monreale, Pinelli, Pedreschi (Pisa KDD LAB) PiLBA Int. Workshop on Privacy in Location-Based Applications @ ESORICS 2008
- Common goal: construct an anonymized version of a trajectory dataset, preserving some target analytical properties
- Different techniques adopted

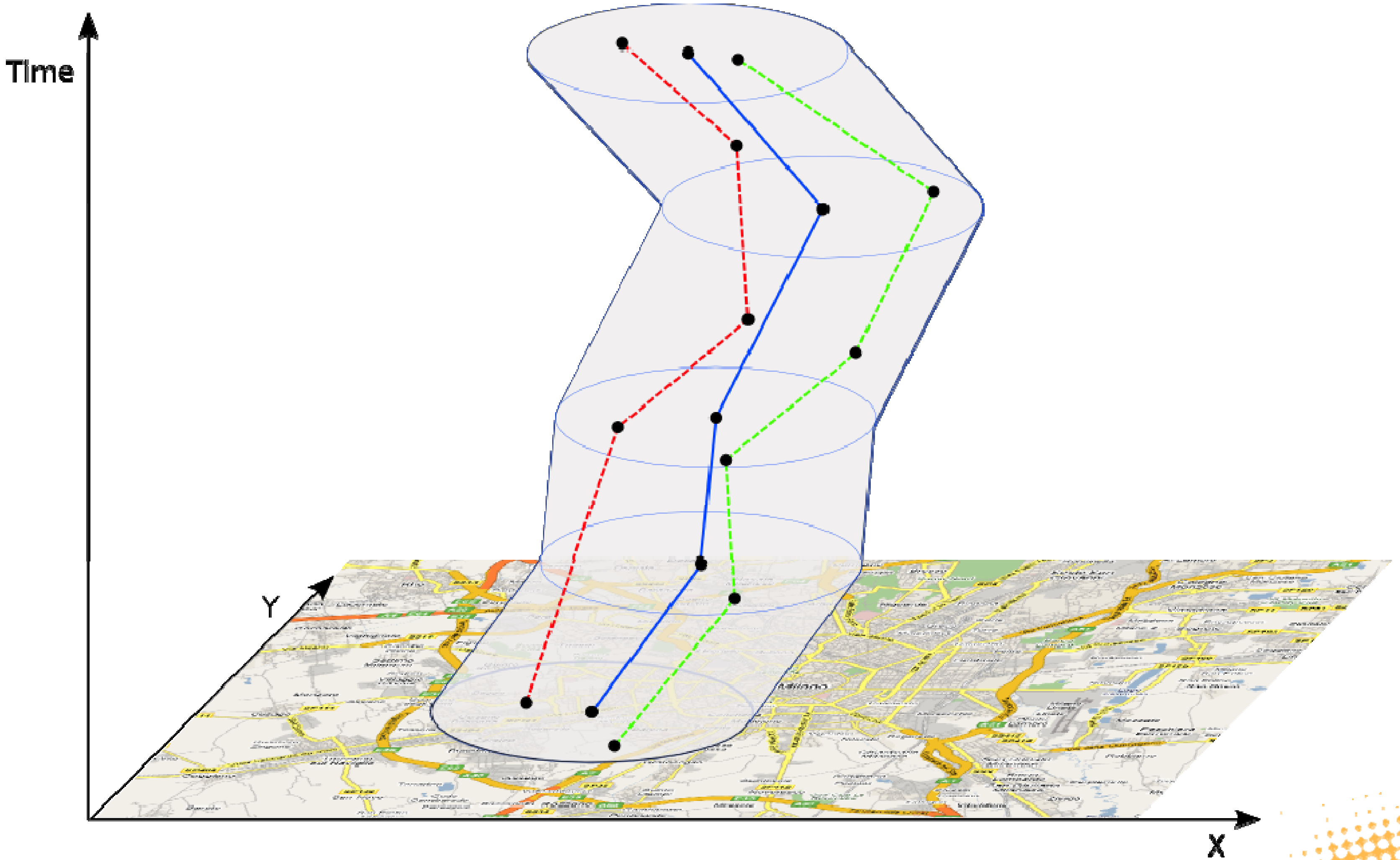
Example result: Never Walk Alone

- Bonchi, Abul, Nanni. *Never Walk Alone: Uncertainty for Anonymity in Moving Objects Databases*. ICDE 2008
- Basic ideas:
  - Trade uncertainty for anonymity: trajectories that are close up the uncertainty threshold are indistinguishable
  - Combine k-anonymity and perturbation
- Two steps:
  - Cluster trajectories into groups of k similar ones (removing outliers)
  - Perturb trajectories in a cluster so that each one is close to each other up to the uncertainty threshold

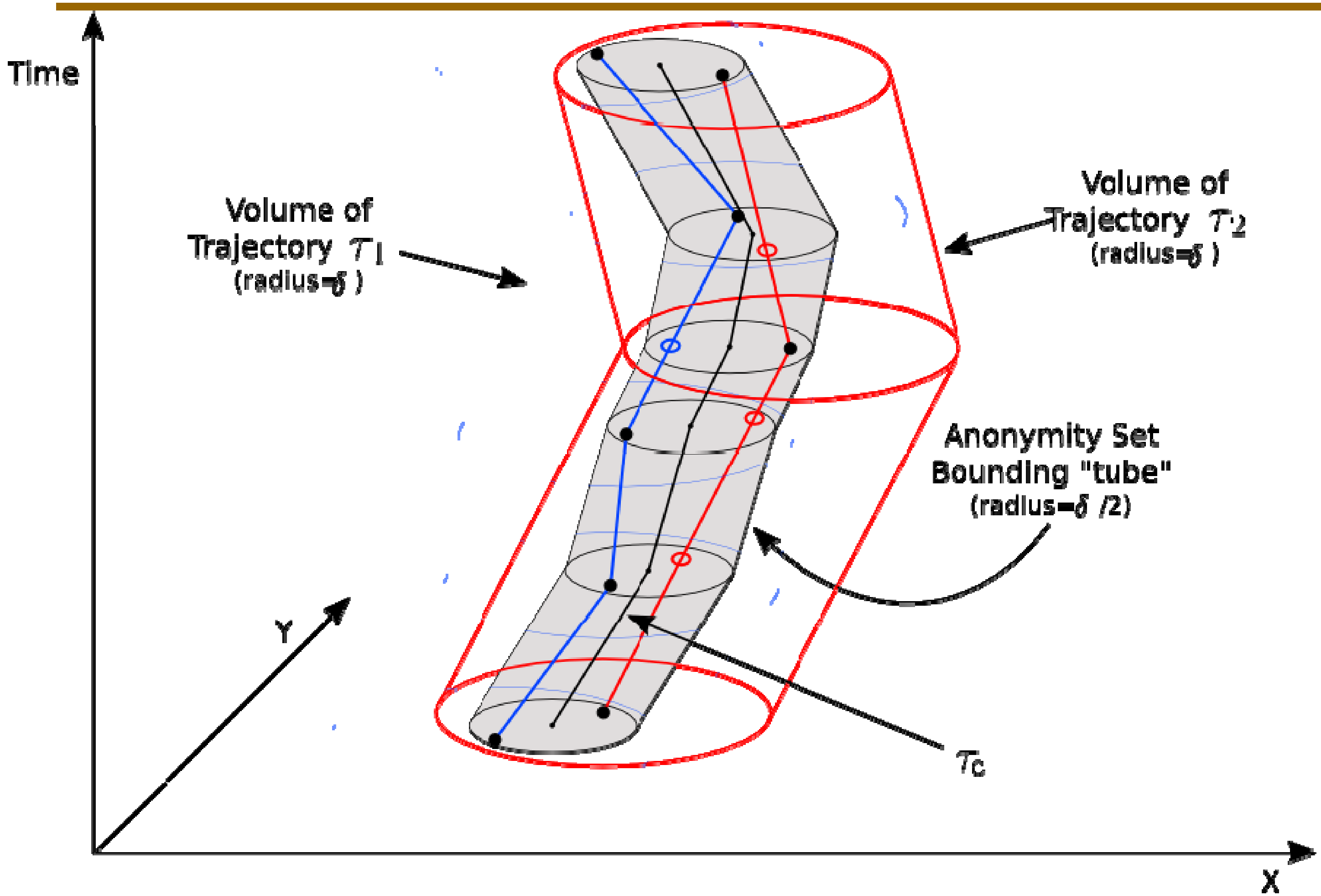
Trajectory cluster



Trajectory cluster



(K,δ) –anonymity set



- K = minimum number of trajectories in the set
- δ = uncertainty threshold (e.g., measurement error of GPS device)

Quality of anonymized datasets

- For reasonable values of K and δ, some interesting analytical properties of the original dataset are preserved by the anonymized trajectories :
  - **density** (aggregate count of mobile users in the spatio-temporal dimension)
  - **Clustering** (to some extent …)
  - **T-patterns: NOT!**
- Prototype **trajectory anonymity toolkit** available

Pattern-Preserving k-Anonymization of Sequences and its Application to Mobility Data Mining

Ruggero G. Pensa, Anna Monreale, Fabio Pinelli, Dino Pedreschi

PiLBA 08 – Int. Workshop on Privacy in Location-Based Applications @ ESORICS 2008

k-Anonymization of sequences

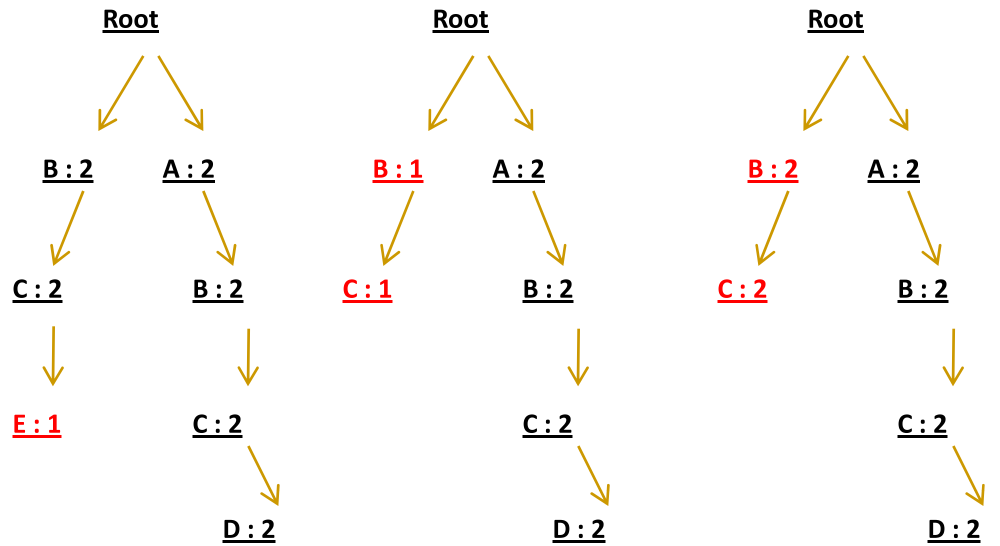
- Idea : each infrequent subsequence is potentially dangerous
- Goal: providing an anonymized dataset of sequences, while preserving frequent sequential pattern results
- Given a dataset of sequences *D*
- Provide a dataset of sequences *D'* s.t.
  1. D' does not contain any *k*-infrequent subsequence
  2. The collection of *k*-frequent pattern in *D'* is « similar » to the collection of *k*-frequent pattern in *D*

k-Anonymization of sequences /2

- Prefix-tree based anonymization algorithm
  1. Build the prefix-tree from *D*
  2. Prune-away all *k*-infrequent subtrees
  3. Re-build the tree by updating the support of existing nodes belonging to pruned subsequences
  4. Generate the anonymized dataset *D’*

Example (k=2)

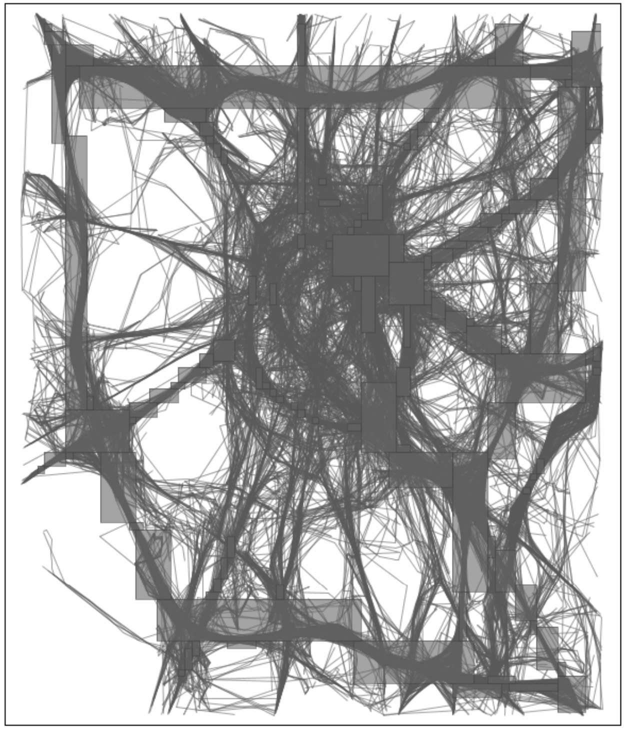
Dataset D
B C
A B C D
A B C D
B C E



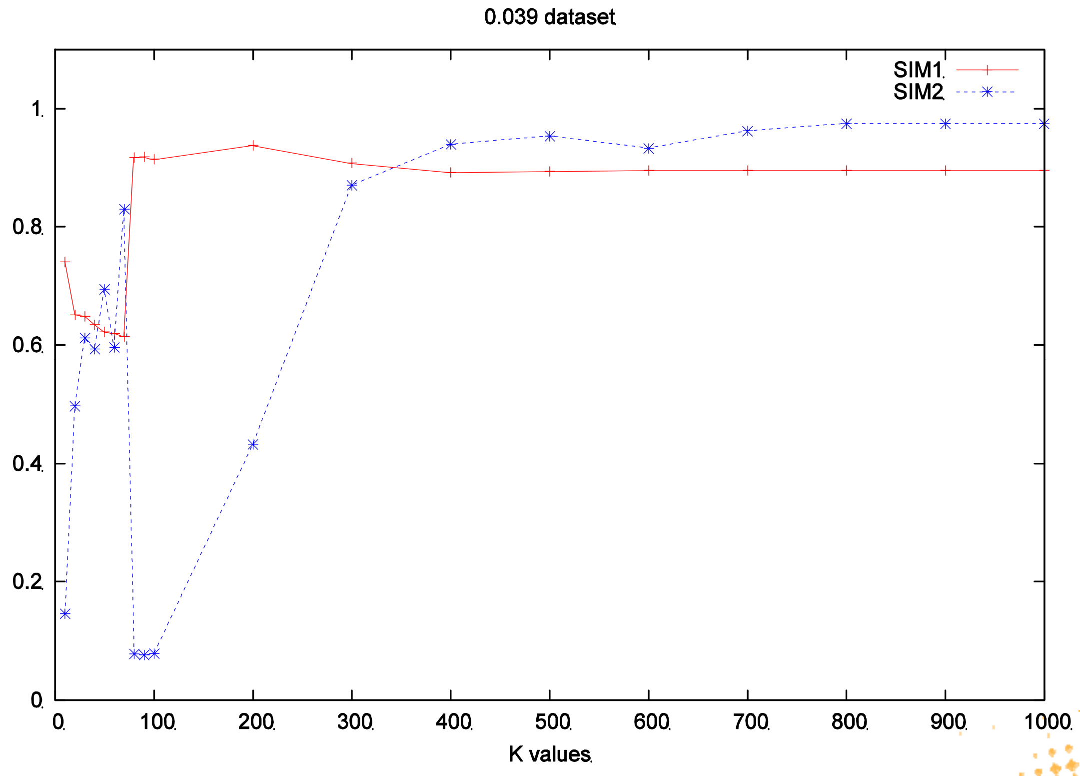
Dataset D'
B C
A B C D
A B C D
B C

Infrequent sequences:
B C E : 1

Experimental results (Milan traffic data)



Pattern support
Pattern collection size



Key open challenges

- Define an acceptable formal measure of anonymity protection:
  - Probability of re-identification (in a given context)
  - A (technically supported) juridical issue!
- Sampling: a necessity **and** an opportunity!
  - Necessary for performance/feasibiliy of data mining from massive mobility datasets
  - Good for anonymity (re-identification probability decreases)

Visual analytics for mobility data

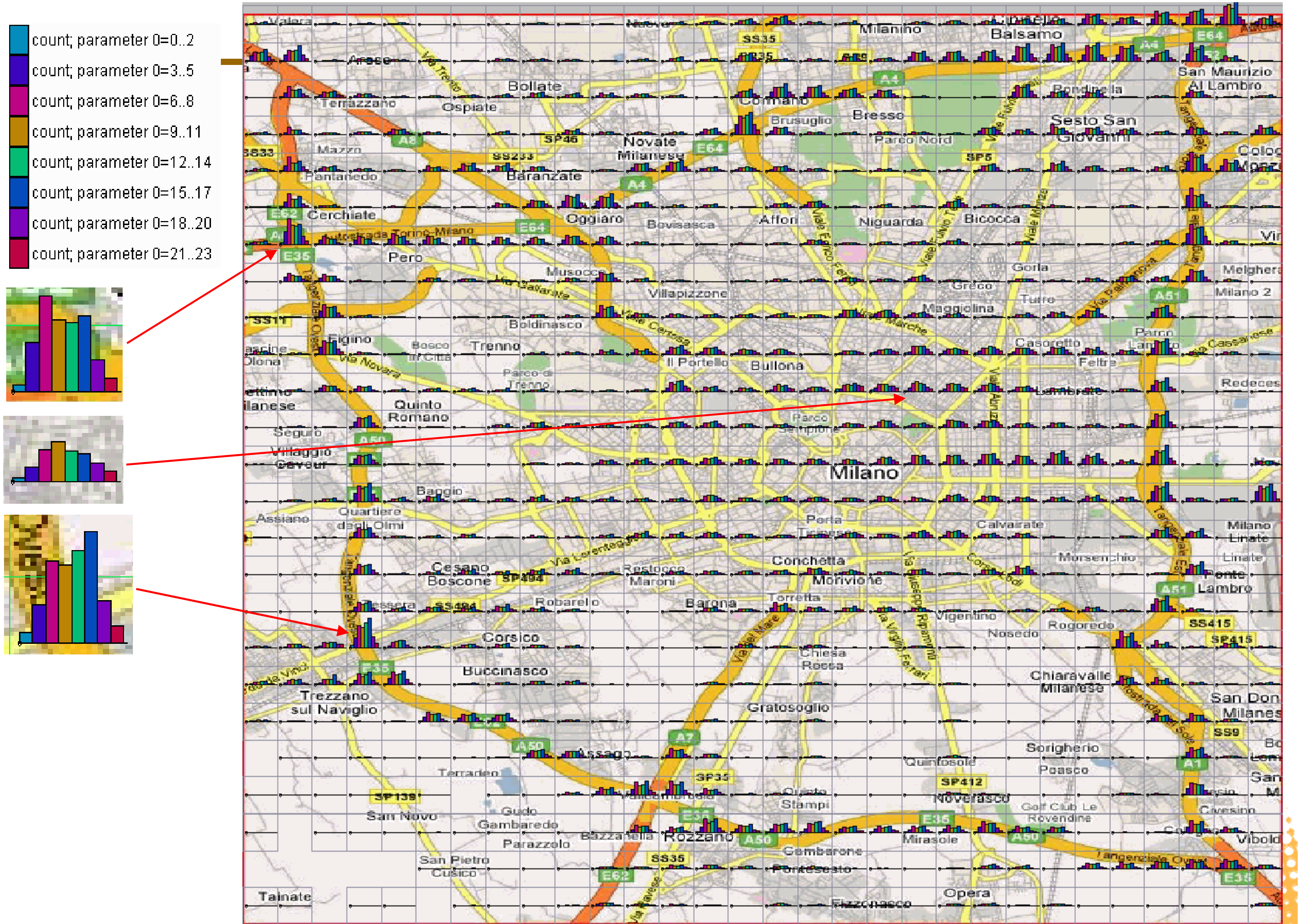
Visual analytics for mobility data

- A synergy of
  - interactive visualization,
  - database processing and
  - data mining
- helps to make sense from large amounts of movement data by interactive, visually-driven exploratory data analysis
- Prototype created in GeoPKDD.eu, based on the Common-GIS system developed at Fraunhofer (Gennady and Natalia Andrienko)

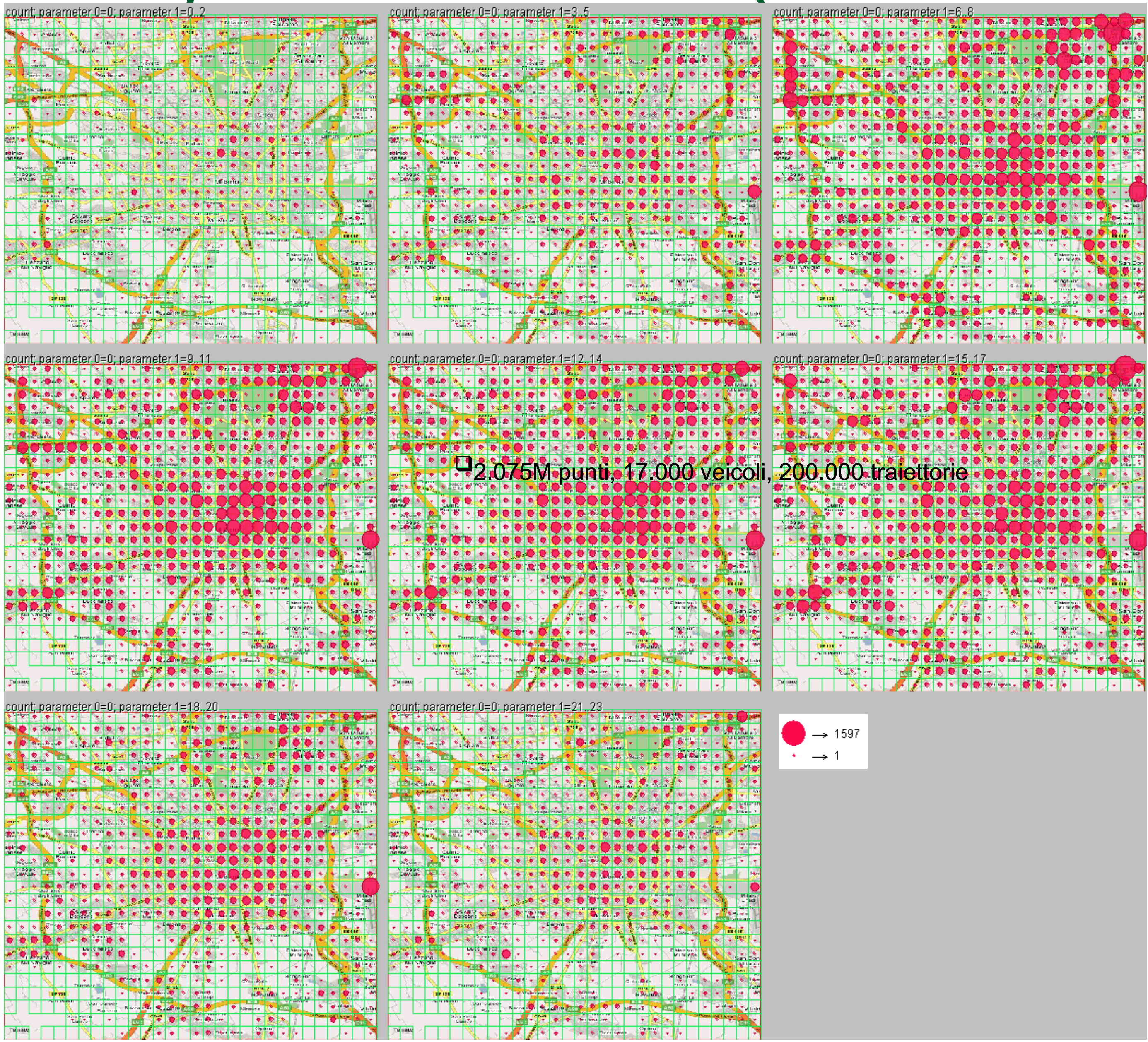
Major techniques

- Aggregation:
  - Traffic-oriented view: by time intervals; by space compartments; by movement direction; by other point-related movement attributes
  - Trajectory-oriented view: by time intervals; by general (trajectory-related) attributes; by starts and ends; by route similarity (through clustering)
- Summarization:
  - Numeric: count, mean, median, …
  - Spatial: aggregated moves
- Visualization and interaction:
  - Multiple coordinated views: animated and static maps, non-cartographic displays
  - Interactive filtering: by time, space, cluster membership, attribute values
  - Dynamic aggregates reacting to the filtering

Traffic density patterns (spatio-temporal aggregation)



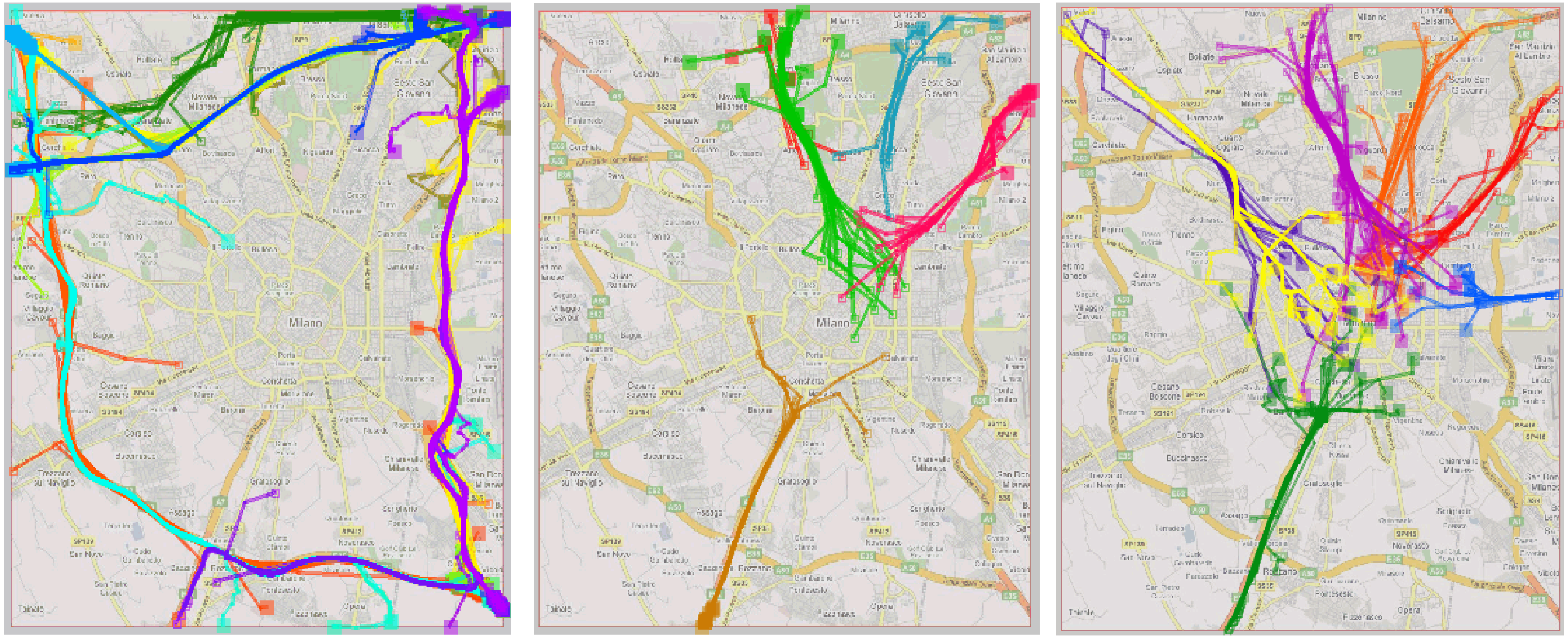
Low-speed movement (counts, 3h intervals)



- 2.075M punti, 17.000 veicoli, 200.000 traiettorie

→ 1597
→ 1

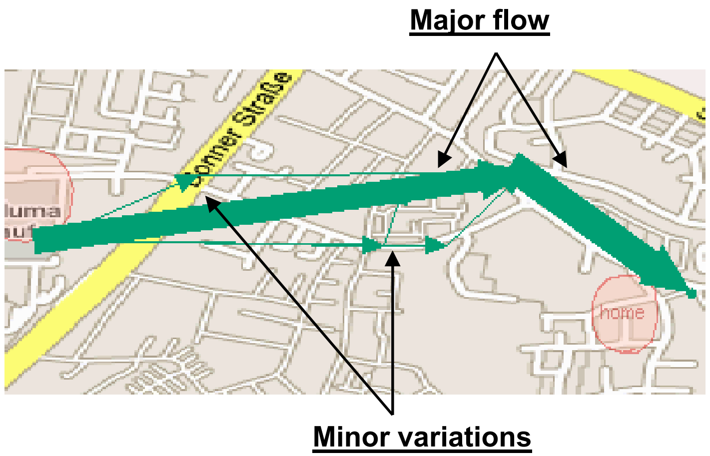
Examples of clusters of trajectories

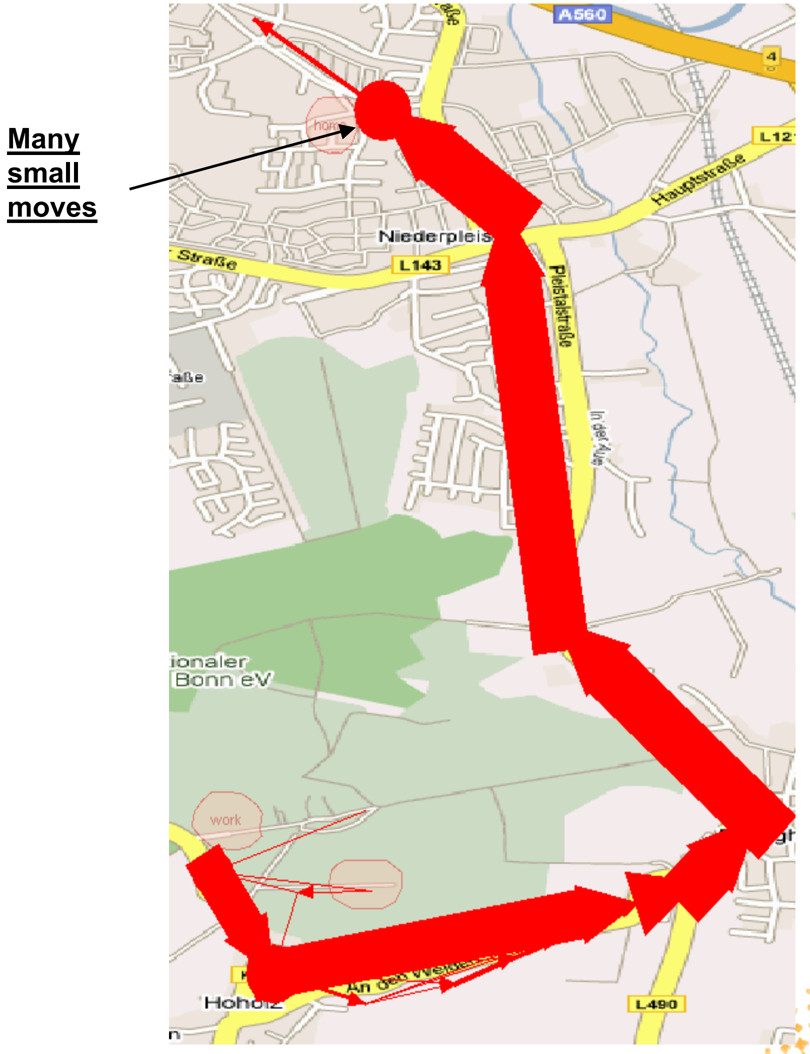


What is an appropriate way to visualize groups of trajectories?

Summarized representation of a bunch of trajectories

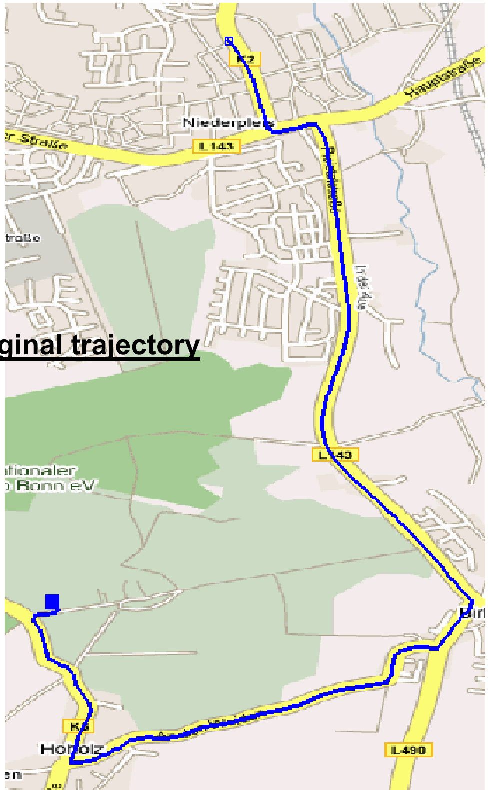
1) Trajectories → sequences of "moves" between "places"
2) For each pair of "places" compute the number of "moves"
3) Represent by vectors (arrows) with proportional widths





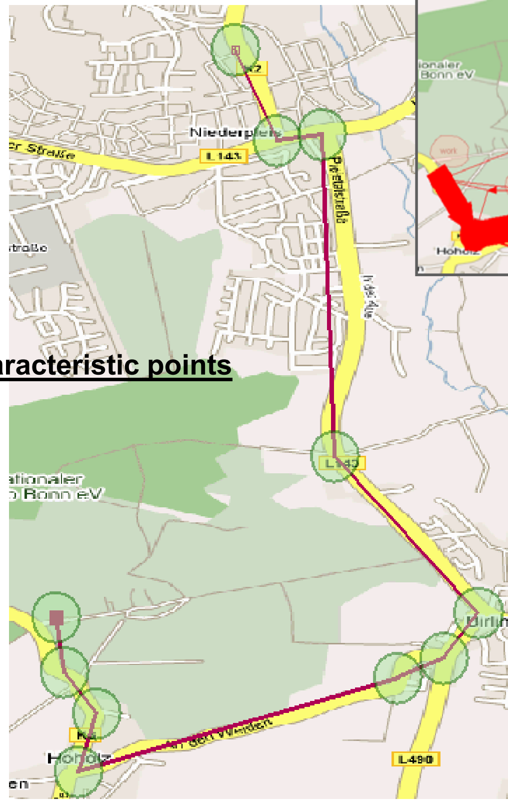
Defining "places"

1) Extract characteristic points from all trajectories
2) Build areas (e.g. circles) around groups of points and isolated points



Original trajectory

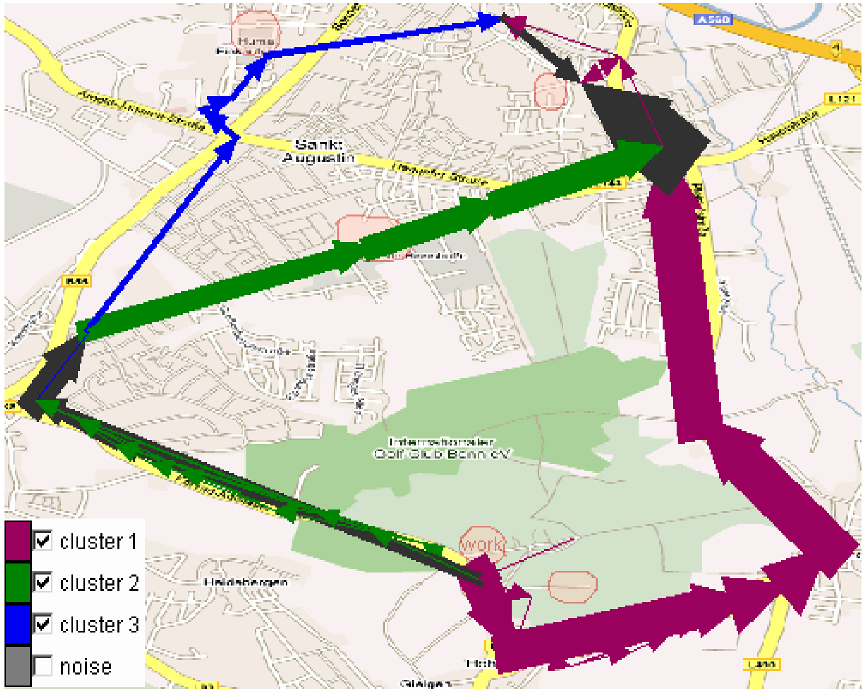
Simplified trajectory

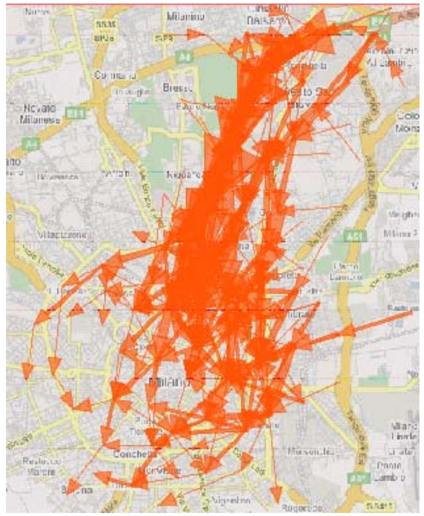


Characteristic points

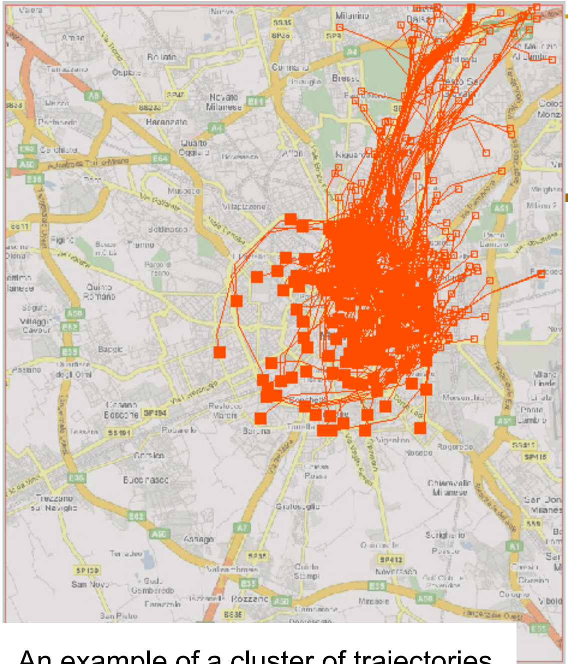
Dynamic aggregation of moves

Each aggregated move is an active object reacting to selection (filtering) of the source data by changing the thickness, color, or visibility of the respective vector.
In particular, aggregated moves react to selection of clusters.

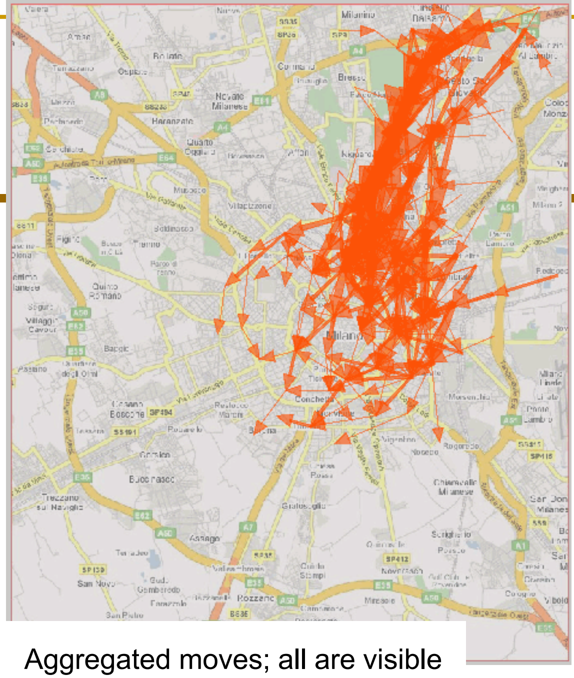




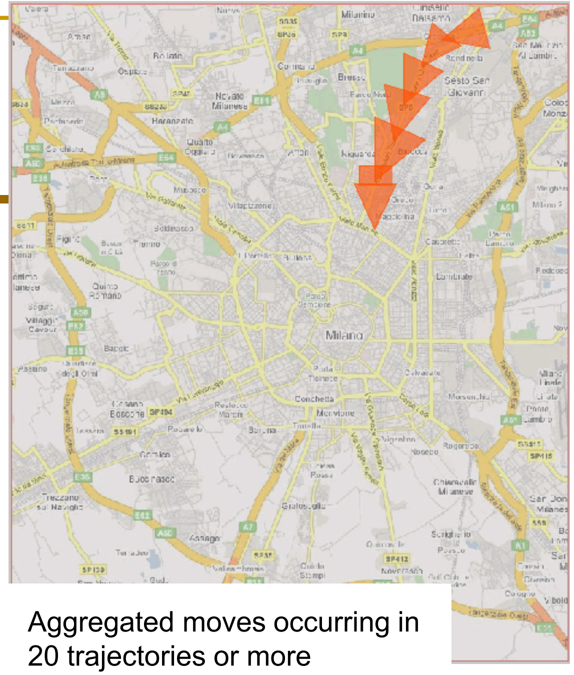
But: not always is a cluster clearly seen…
Possible solution: filter aggregated moves by the number of elementary moves (i.e. trajectory fragments) they include



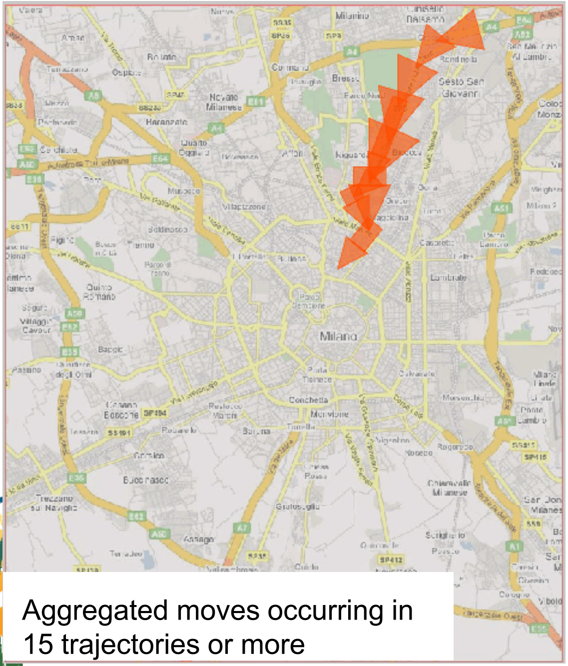
An example of a cluster of trajectories



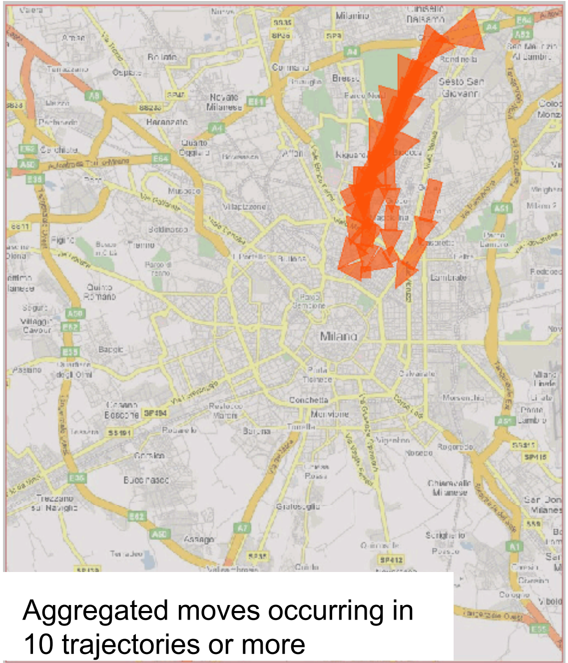
Aggregated moves; all are visible



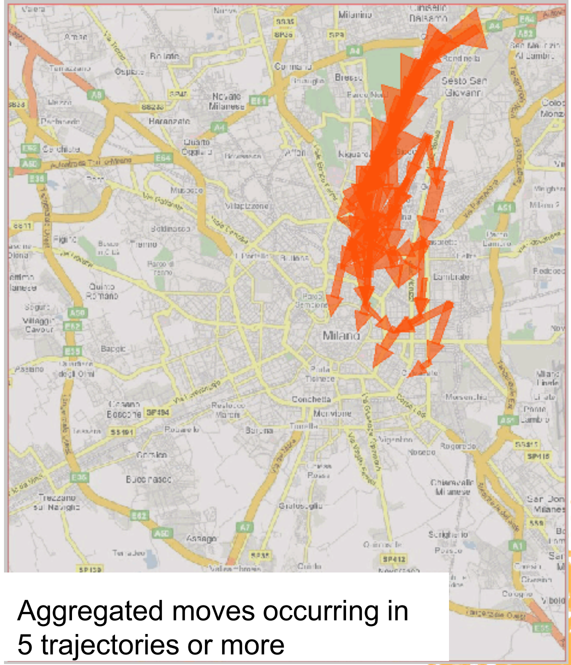
Aggregated moves occurring in 20 trajectories or more



Aggregated moves occurring in 15 trajectories or more

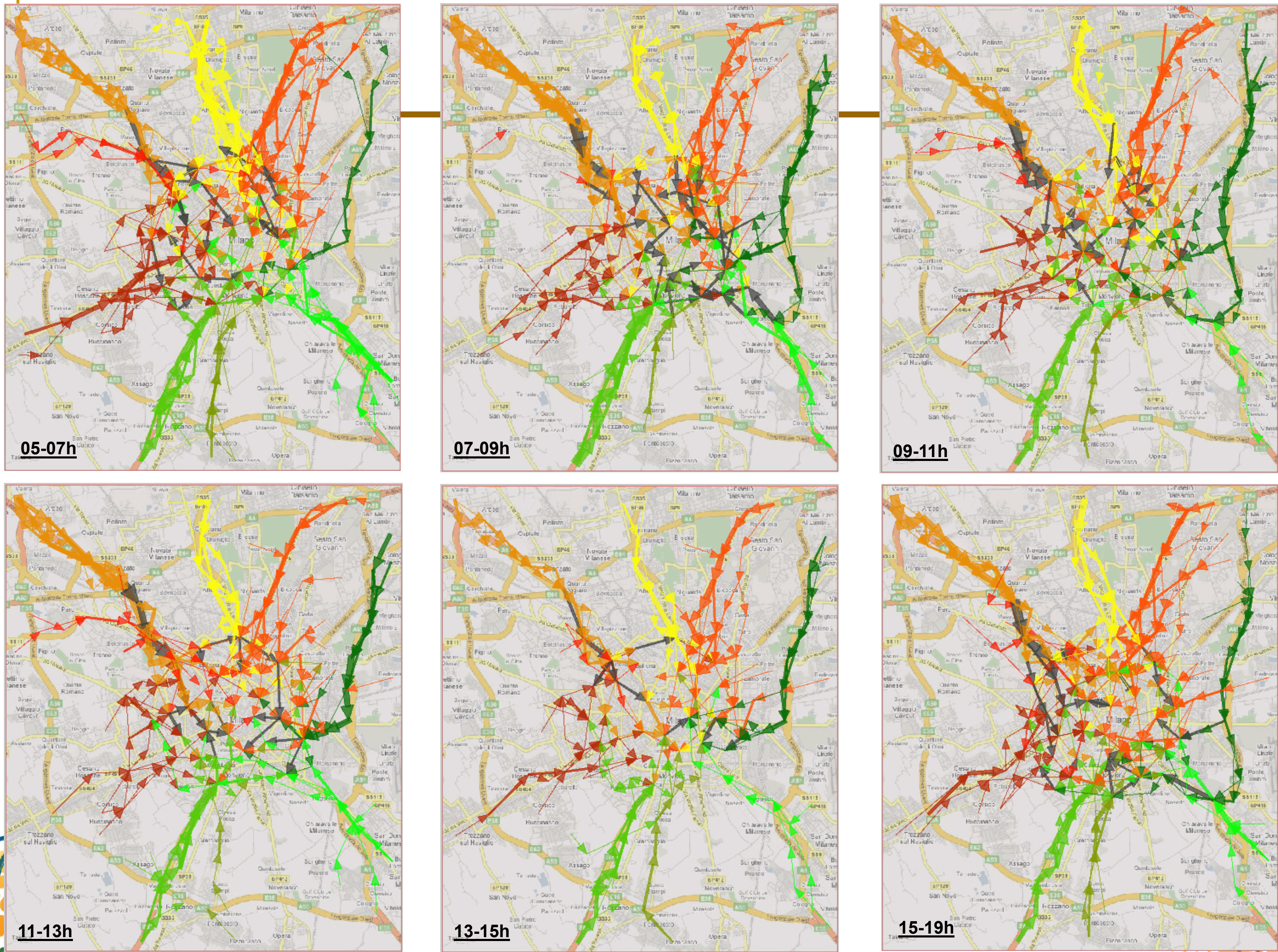


Aggregated moves occurring in 10 trajectories or more



Aggregated moves occurring in 5 trajectories or more

Exploration of the use of the most popular routes towards the centre by times of the day

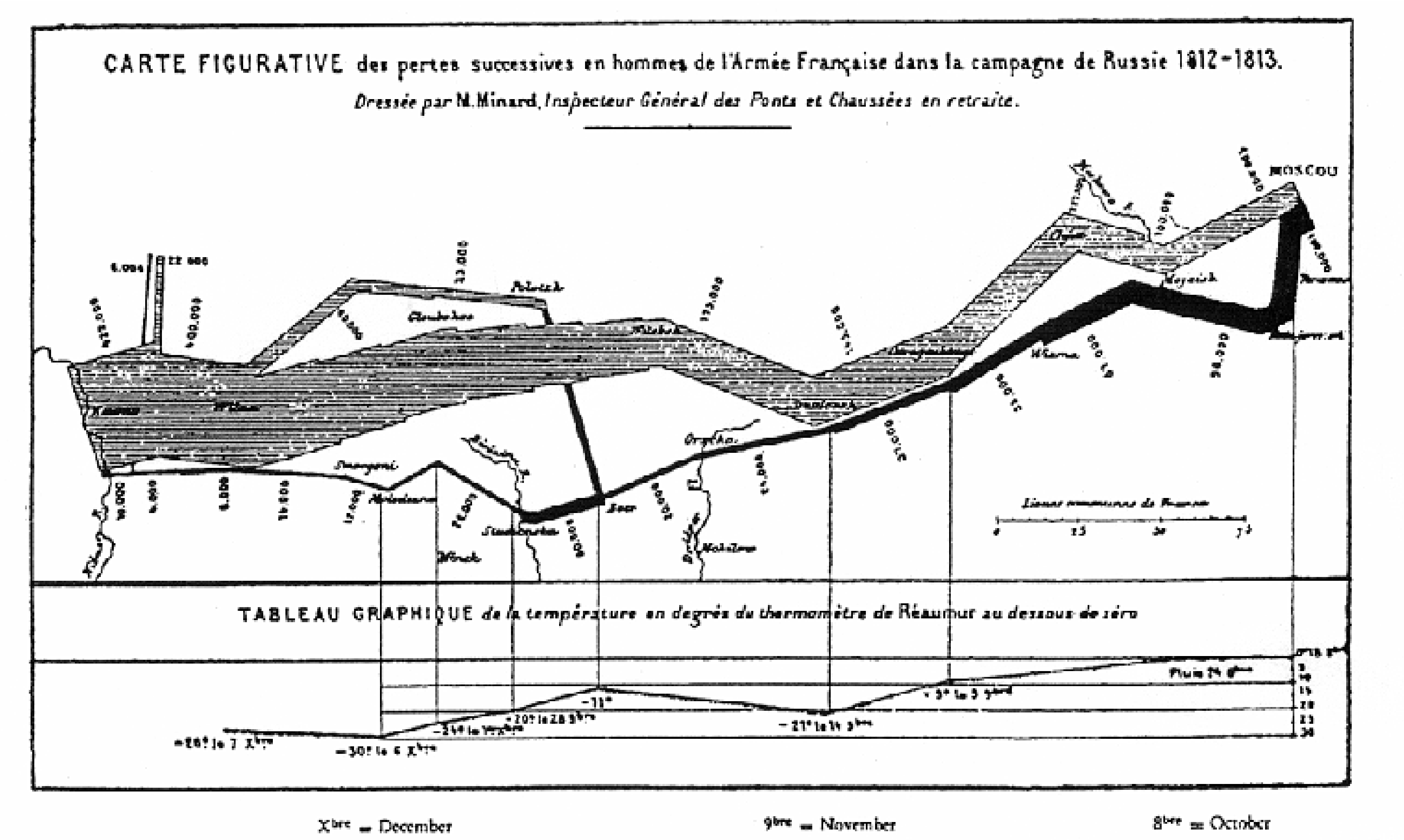


Conclusions

Privacy-preserving Mobility Data Mining strives for a win-win situation

- Obtaining the advantages of collective mobility knowledge without disclosing inadvertently any individual mobility knowledge.
- A word of wisdom: solutions can only be obtained via an alliance of technology, legal regulations, and social norms (Rakesh Agrawal)
- **GeoPKDD.eu** is in the mix, shaping up the area of PP mobility data mining
- Challenge: UbiComp will flood us with new complex data (in a decentralized setting)
  - data miners have only begun to scratch the surface of this problem

… trying to accomplish a long-time dream



The representation of Napoleon's Russian campaign of 1812 produced by Charles Joseph Minard in 1861

Acknowledgements

- We are grateful to all the GeoPKDD researchers, who made the project successful by their results and contributed actively to this tutorial
  - They're too many to be listed here, their work has been cited along these notes
  - Thanks folks!
- GeoPKDD is a project in the Future and Emerging Technologies programme of the Sixth Framework Programme for Research of the European Commission, FET-Open contract n: 014915

Giannotti · Pedreschi (Eds.)

Mobility, Data Mining and Privacy

The technologies of mobile communications and ubiquitous computing pervade our society, and wireless networks sense the movement of people and vehicles, generating large volumes of mobility data. This is a scenario of great opportunities and risks: on one side, mining this data can produce useful knowledge, supporting sustainable mobility and intelligent transportation systems; on the other side, individual privacy is at risk, as the mobility data contain sensitive personal information. A new multidisciplinary research area is emerging at this crossroads of mobility, data mining, and privacy.

This book assesses this research frontier from a computer science perspective, investigating the various scientific and technological issues, open problems, and roadmap. The editors manage a research project called GeoPKDD, Geographic Privacy-Aware Knowledge Discovery and Delivery, funded by the EU Commission and involving 40 researchers from 7 countries, and this book tightly integrates and relates their findings in 13 chapters covering all related subjects, including the concepts of movement data and knowledge discovery from movement data; privacy-aware geographic knowledge discovery; wireless network and next-generation mobile technologies; trajectory data models, systems and warehouses; privacy and security aspects of technologies and related regulations; querying, mining and reasoning on spatiotemporal data; and visual analytics methods for movement data.

This book will benefit researchers and practitioners in the related areas of computer science, geography, social science, statistics, law, telecommunications and transportation engineering.

ISBN 978-3-540-75176-2

9 783540 751762

springer.com

Giannotti
Pedreschi (Eds.)

Mobility, Data Mining
and Privacy

Fosca Giannotti
Dino Pedreschi (Eds.)

Mobility, Data Mining and Privacy

Geographic Knowledge Discovery

Springer