Big Data Analytics

FOSCA GIANNOTTI AND LUCA PAPPALARDO

HTTP://DIDAWIKI.DI.UNIPI.IT/DOKU.PHP/BIGDATAANALYTICS/BDA/

DIPARTIMENTO DI INFORMATICA - Università di Pisa
anno accademico 2018/2019

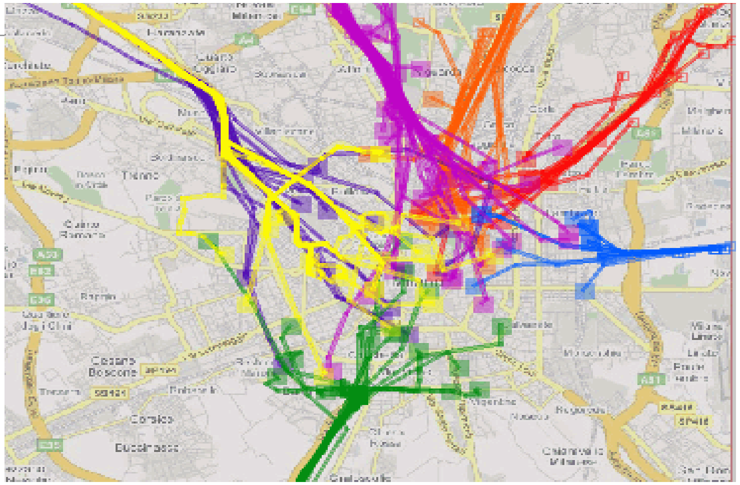
Mobility Data Mining

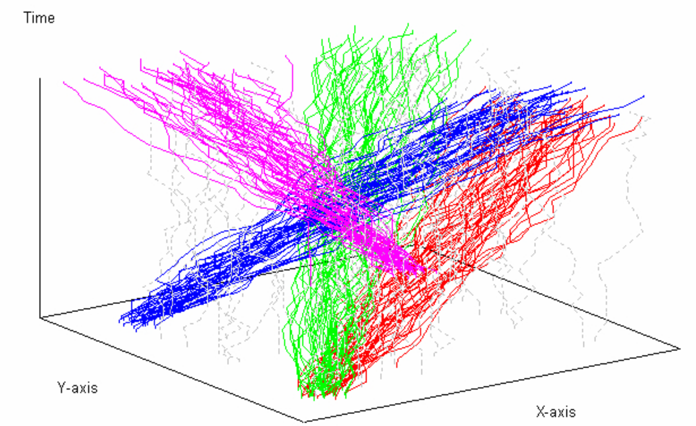
PATTERNS&MODELS

Trajectory Clustering

T-clustering

- Trajectories are grouped based on similarity
- Several possible notions of similarity
  - Start/End points
  - Shape of trajectory
  - Shape & time
  - Etc.



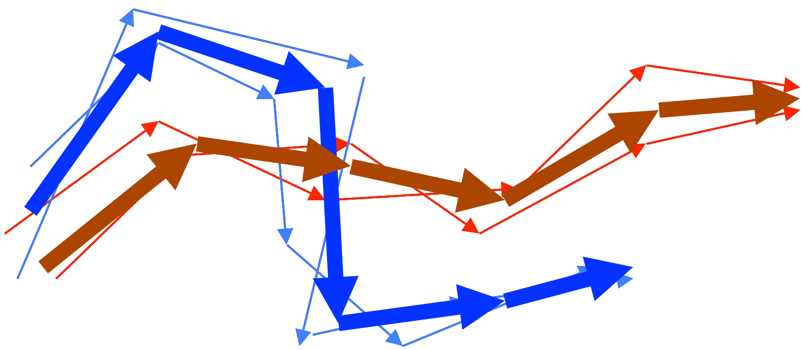


Nanni, Pedreschi. **Time-focused clustering of trajectories of moving objects**. J. of Intelligent Information Systems, 2006.

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

Trajectory Clustering

- Questions:
- Which distance between trajectories?
- Which kind of clustering?
- What is a cluster 'mean' in our case?
  - A representative trajectory?



Which distance?

- Average Euclidean distance (Spatio-temporal distance)

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

distance between moving objects τ1

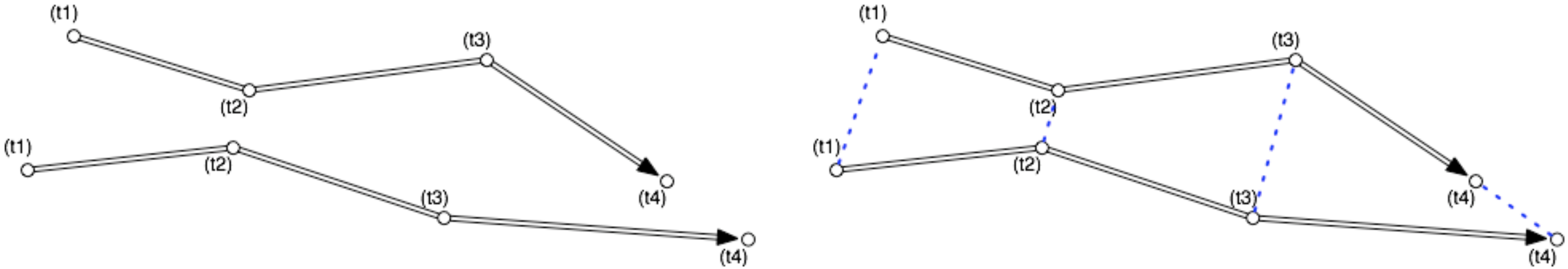
- "Synchronized" behaviour distance
  - Similar objects = almost always in the same place at the same time
- Computed on the whole trajectory

Average Euclidean Distance Sincronized

- Align point temporally

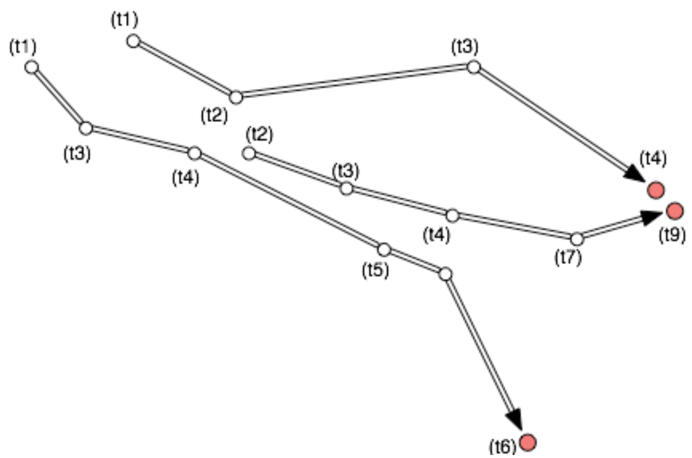
$$D(\tau_1,\tau_2)|_T = \frac{\int_T d(\tau_1(t),\tau_2(t))dt}{|T|}$$

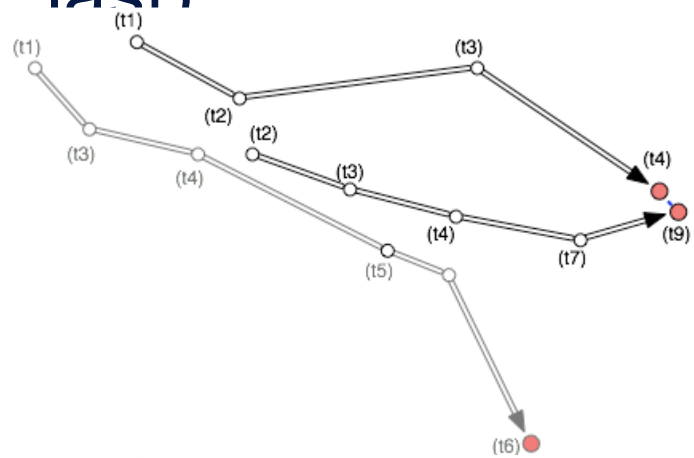
- Eventually assign penalties to non matching points

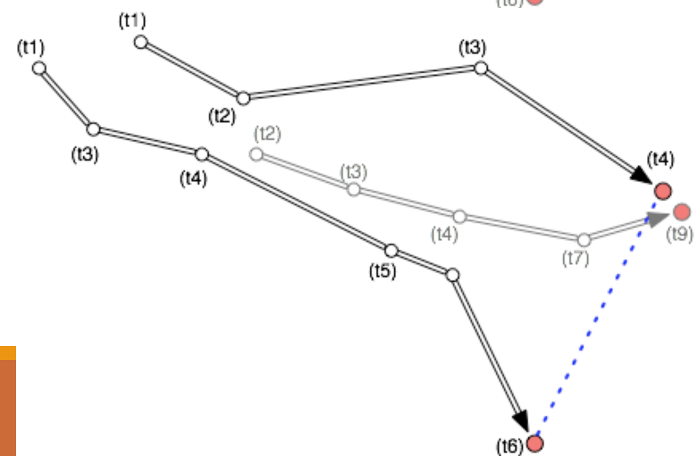


Common Destination

- Select last point *Plast* for each trajectory
- D(T,T’) = Euclidean(Plast, P’last)

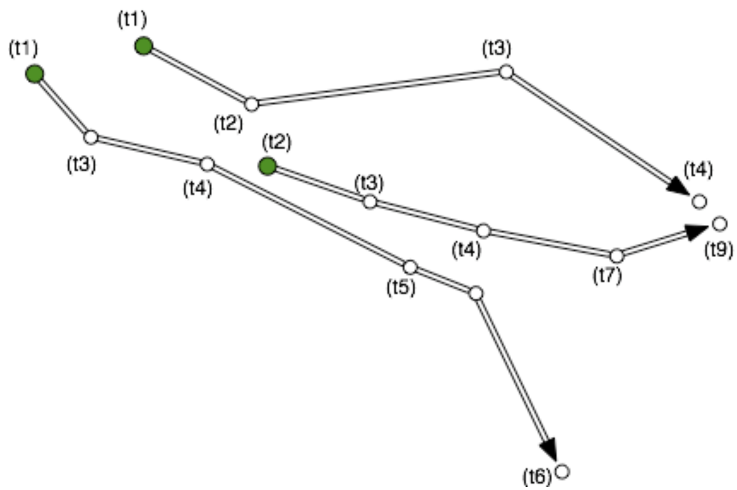


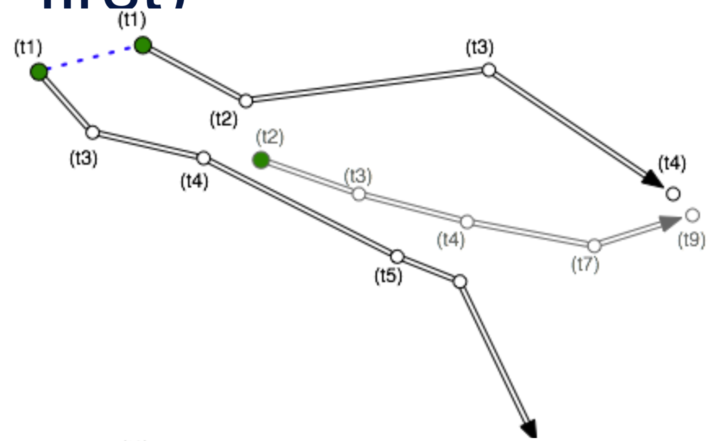


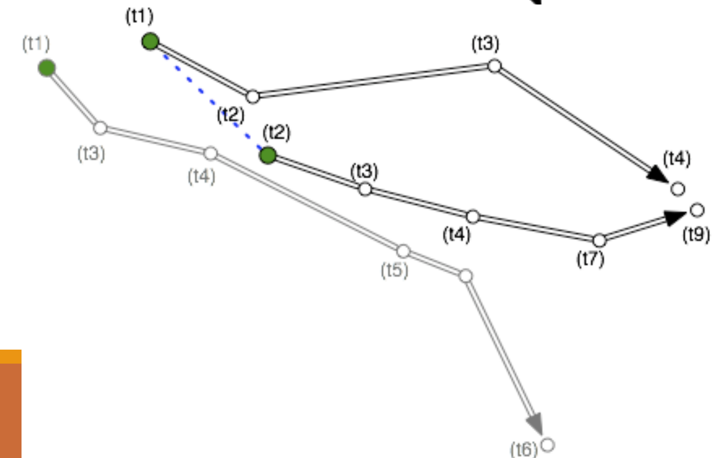


Common Origins

- Select first point *Pfirst* for each trajectory
- D(T,T') = Euclidean(Pfirst, P'first)

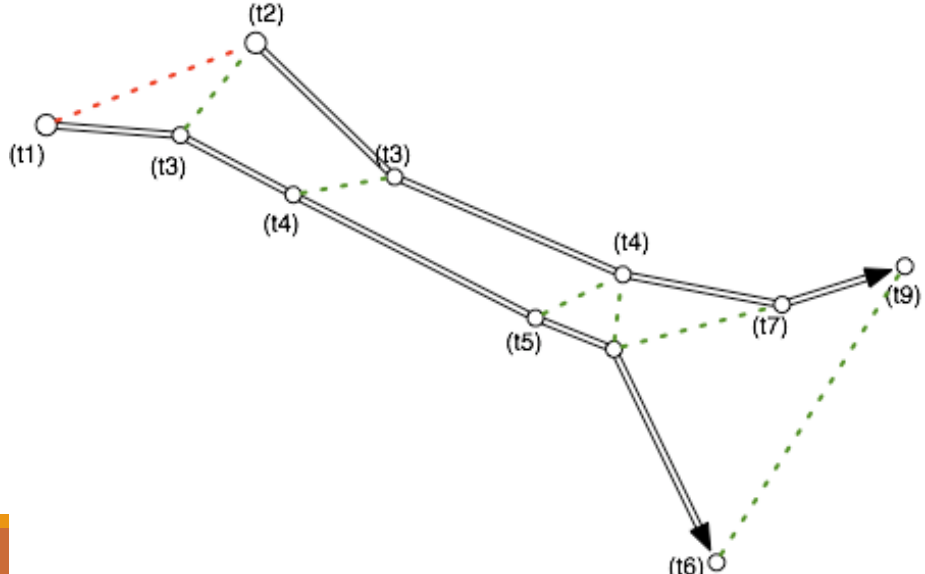






Route Similarity

- Alignment of points, multiple matches
- Average Euclidean Distance
- Penalties for non matching initial points (no penalties for destinations)

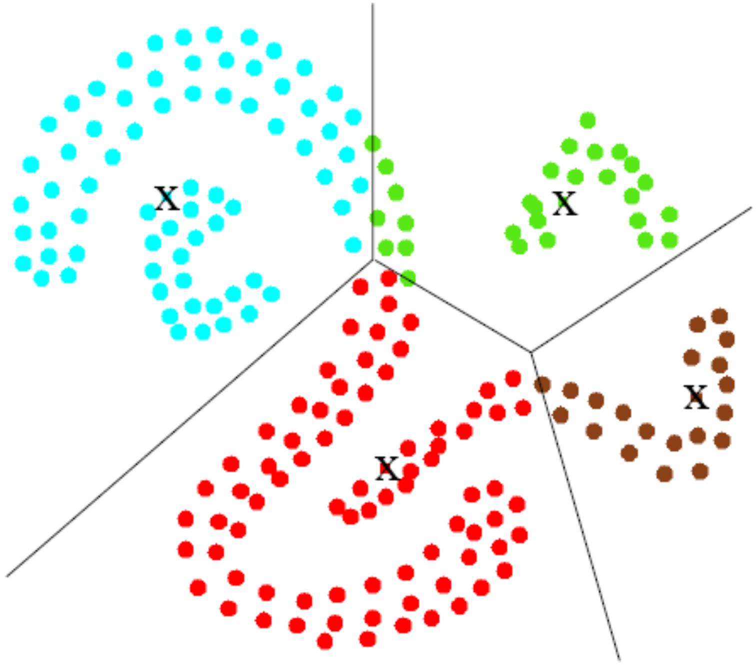


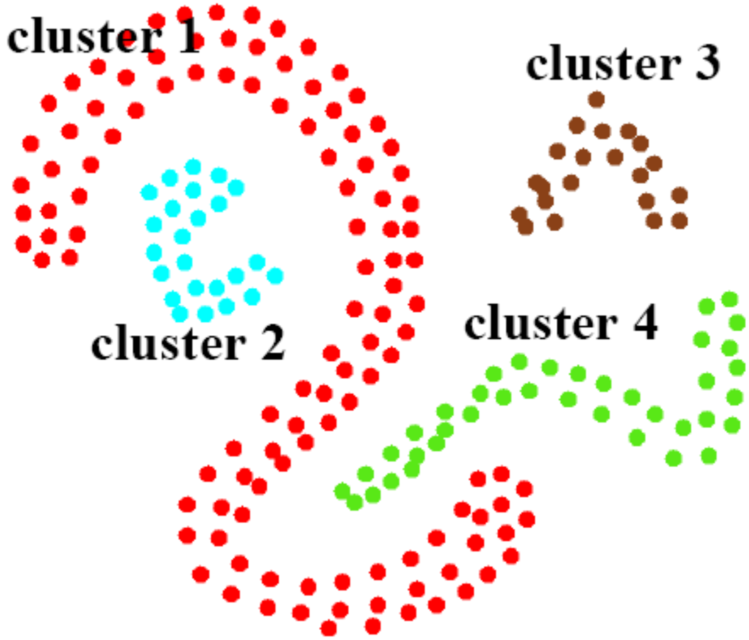
Which kind of clustering?

- 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., 1999) → **T**(rajectory)**-OPTICS**
  - Evolution of basic DBSCAN

Density Based Clustering

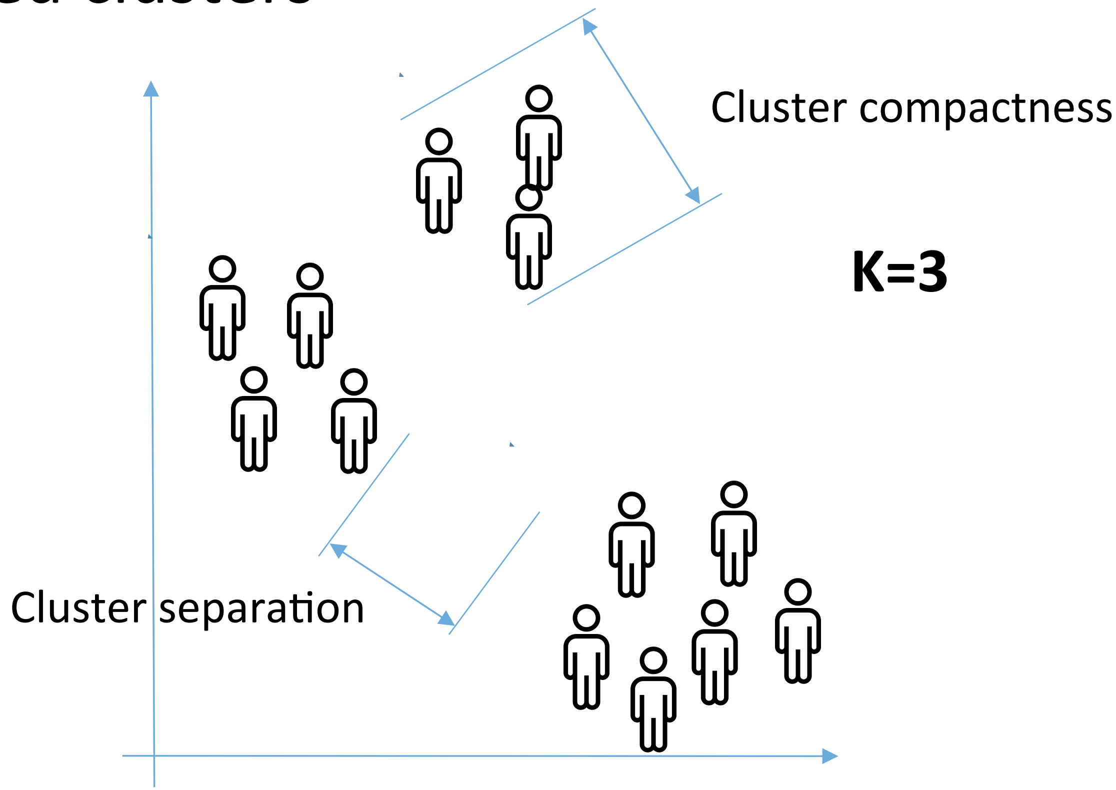
K-means	*Density-based*





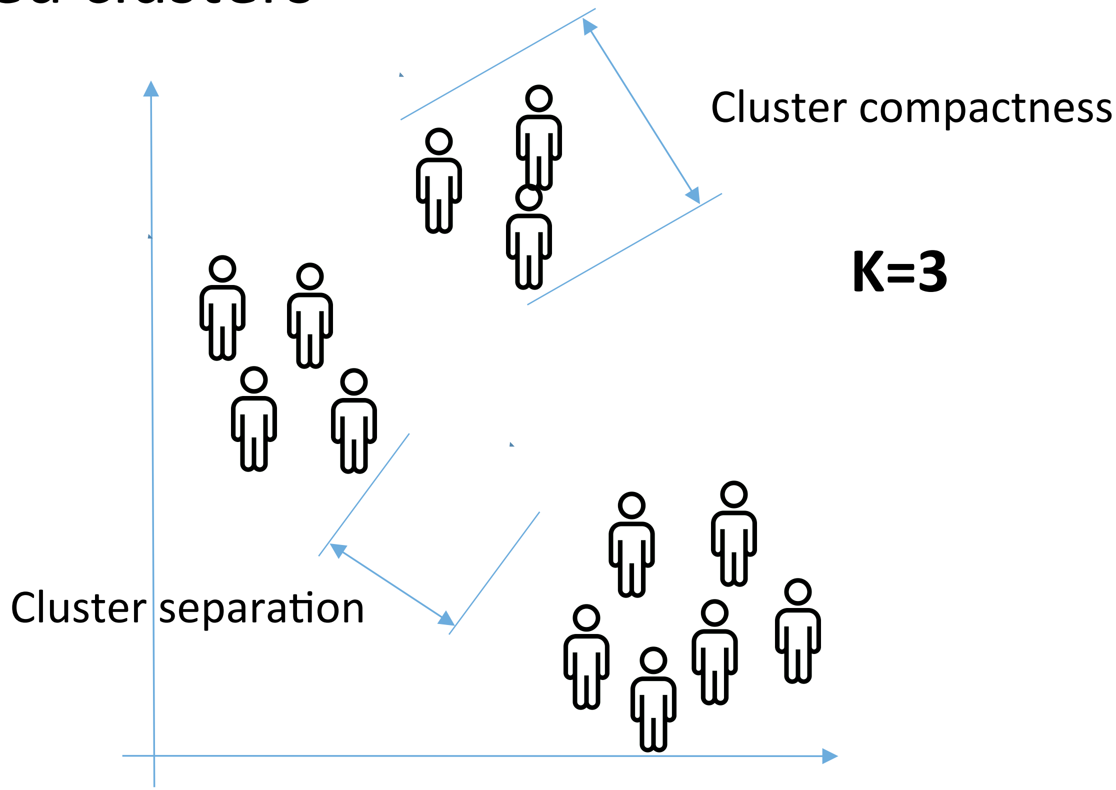
Clustering: K-means (family)

- Find k subgroups that form compact and well-separated clusters



Clustering: K-means (family)

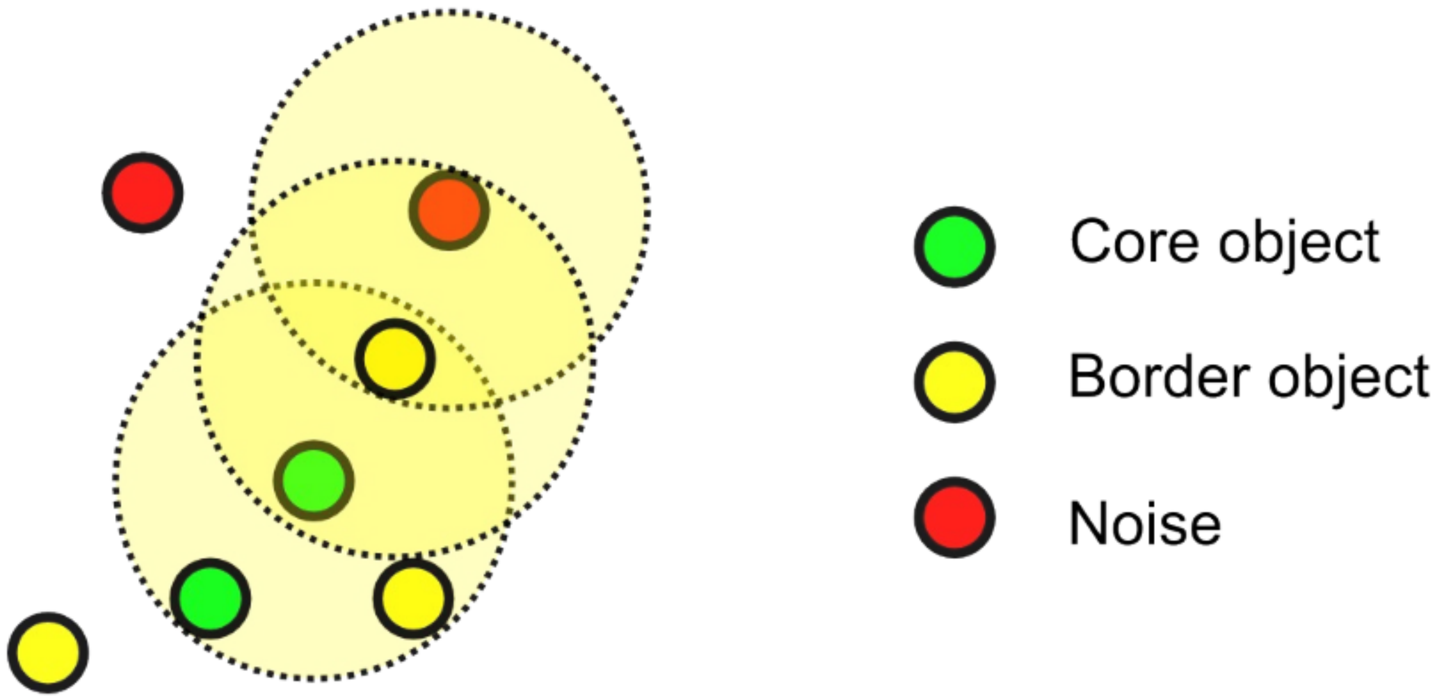
- Find k subgroups that form compact and well-separated clusters



Density Based Clustering

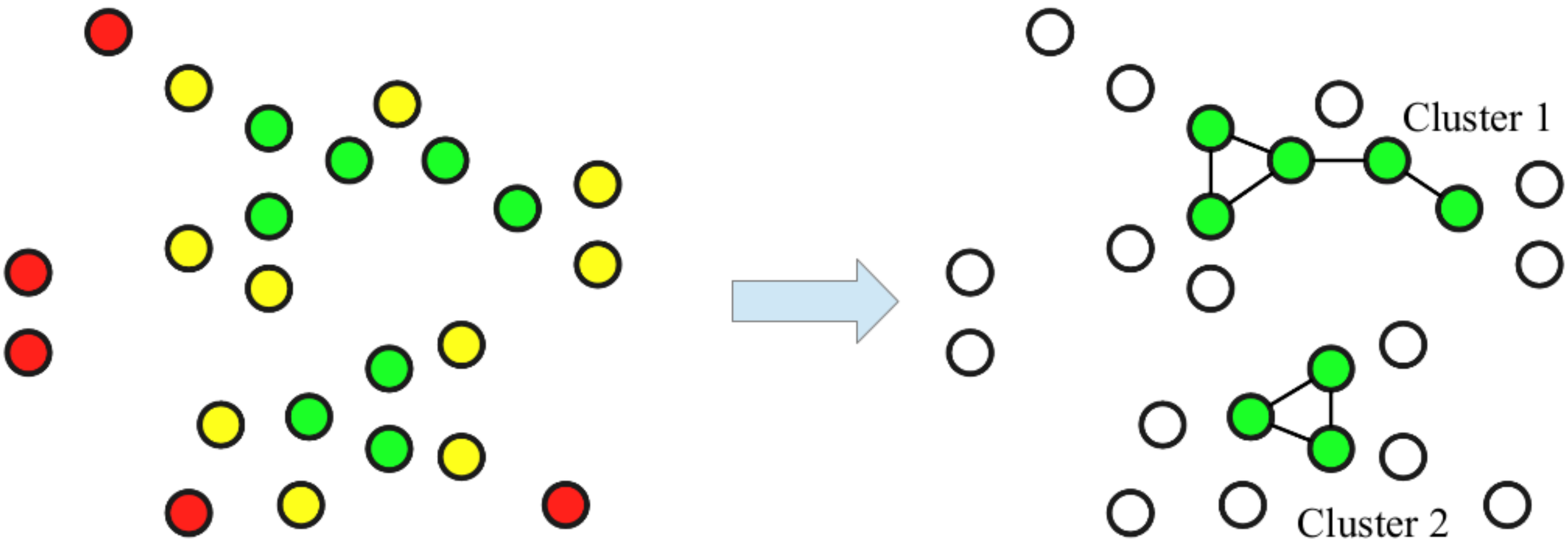
Step 1: label points as core (dense), border and noise

- Based on thresholds R (radius of neighborhood) and min_pts (min number of neighbors)



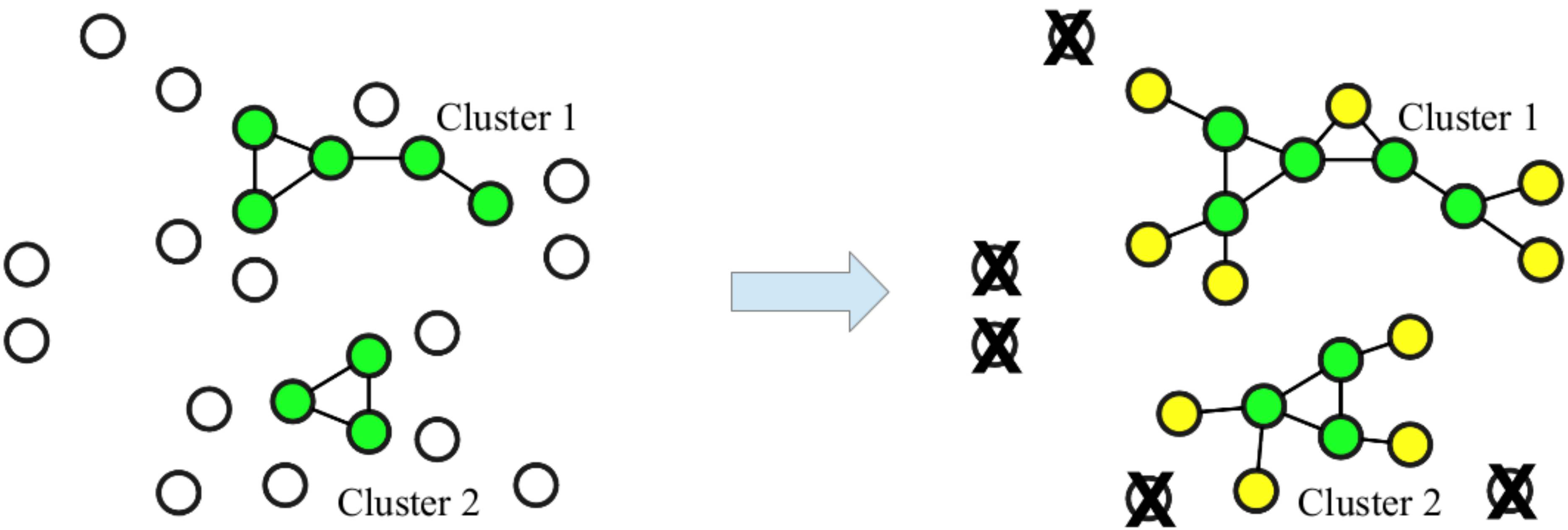
Density Based Clustering

Step 2: connect core objects that are neighbors, and put them in the same cluster



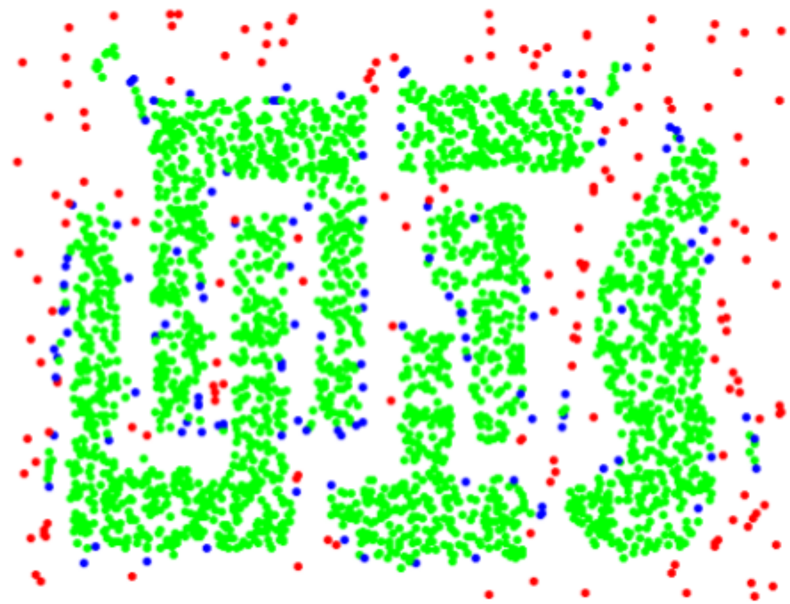
Density Based Clustering

Step 3: associate border objects to (one of) their core(s), and remove noise



Density Based Clustering

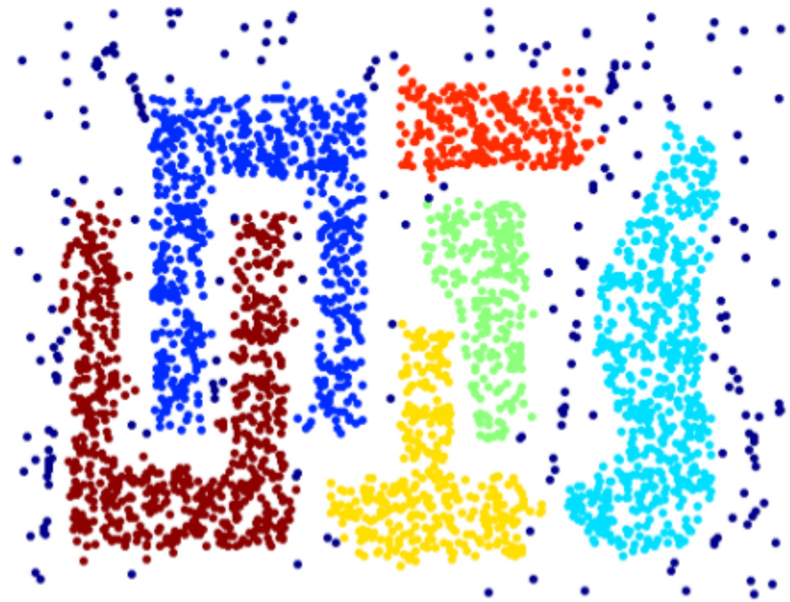
Original Points



Point types: core, border and noise

Density Based Clustering

Original Points

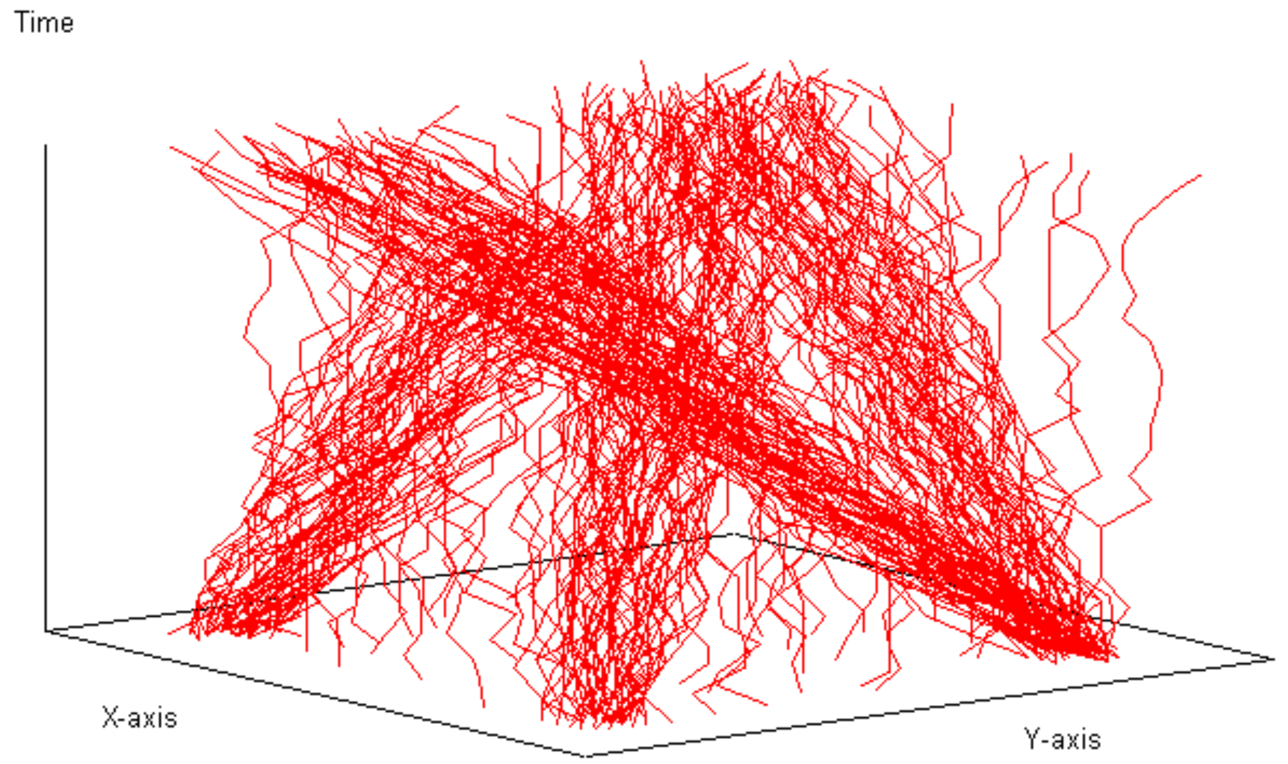


Clusters

- Resistant to Noise
- Can handle clusters of different shapes and sizes

A sample dataset

- A set of trajectories forming 4 clusters + noise (synthetic)



Ad-hoc distance functions

- Colocation –
  - Link prediction,
  - Semantic behaviors,
  - GSM data
- Spatio-temporal Colocation –
  - Link prediction,
  - Semantic behaviors,
  - GSM data

Start and End inclusion

Car Pooling Matching

Align to end –

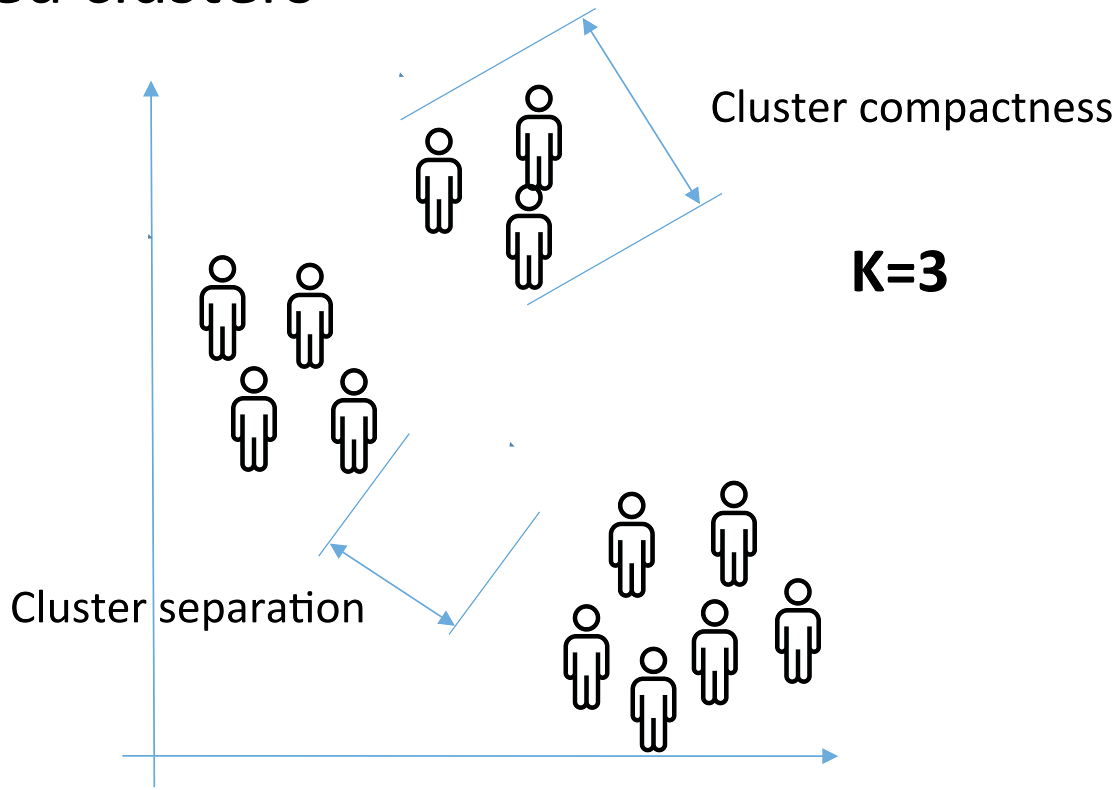
Incoming flows

Align to start –

Outcoming flows

Clustering: K-means (family)

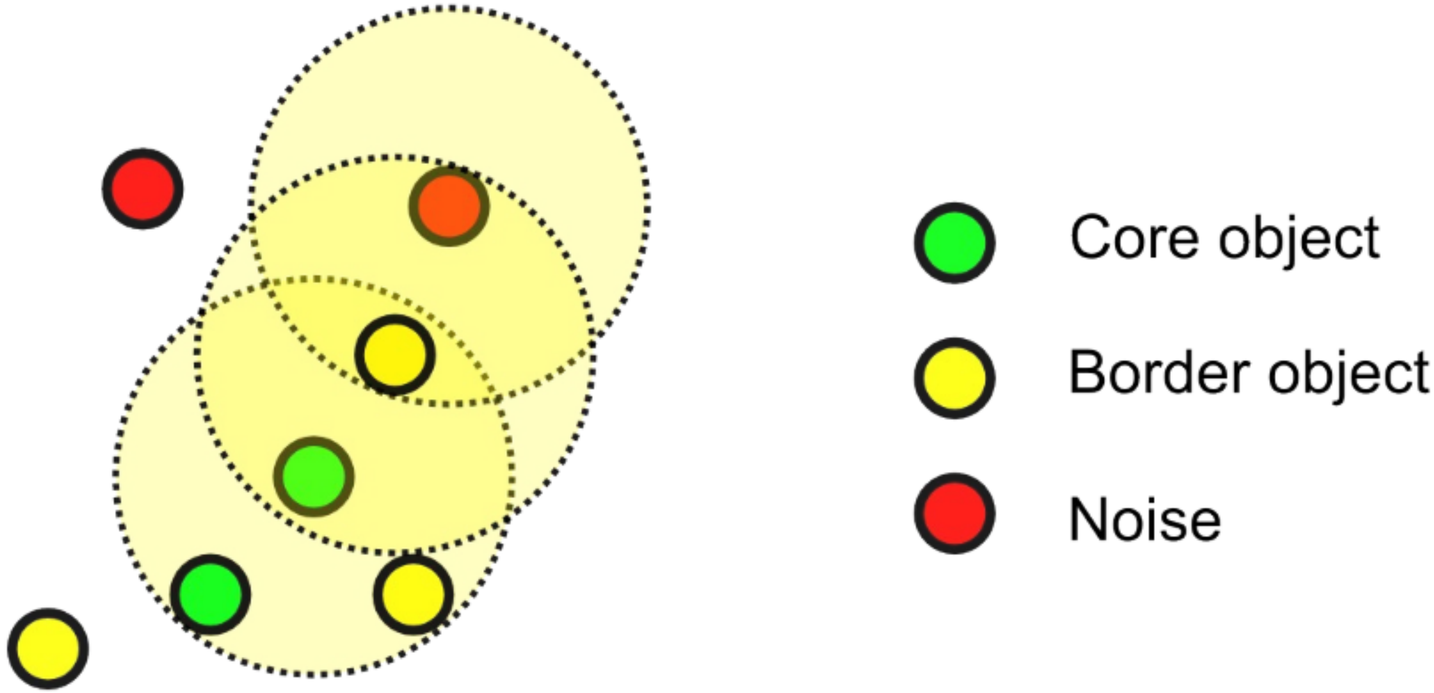
- Find k subgroups that form compact and well-separated clusters



Density Based Clustering

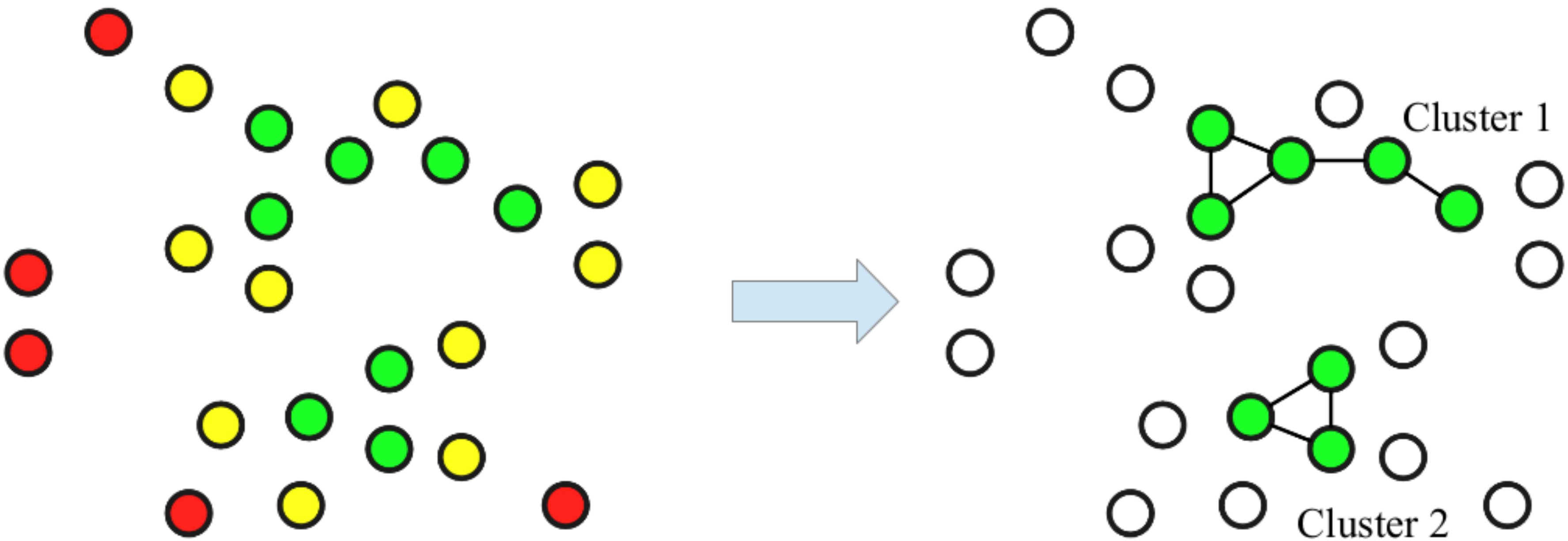
Step 1: label points as core (dense), border and noise

- Based on thresholds R (radius of neighborhood) and min_pts (min number of neighbors)



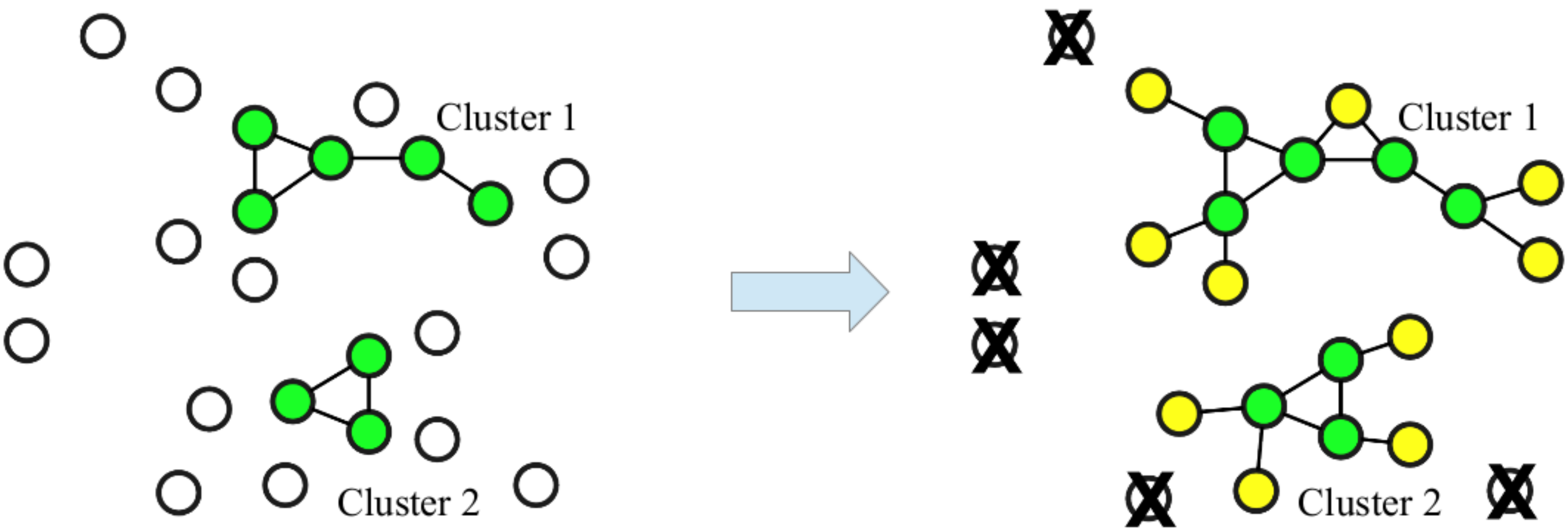
Density Based Clustering

Step 2: connect core objects that are neighbors, and put them in the same cluster



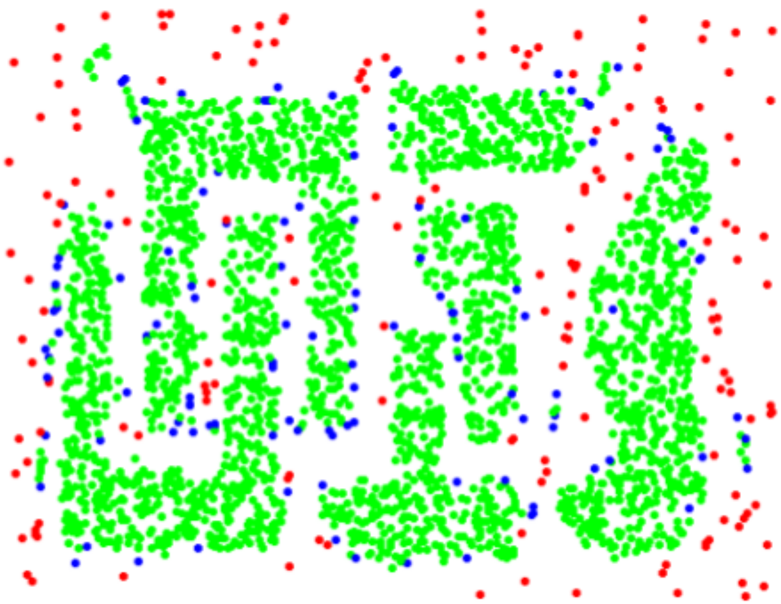
Density Based Clustering

Step 3: associate border objects to (one of) their core(s), and remove noise



Density Based Clustering

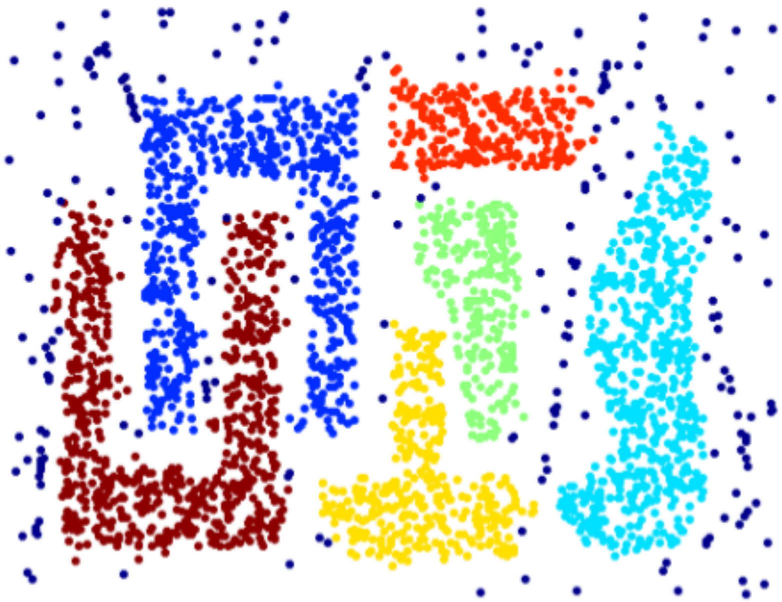
Original Points



Point types: core, border and noise

Density Based Clustering

Original Points

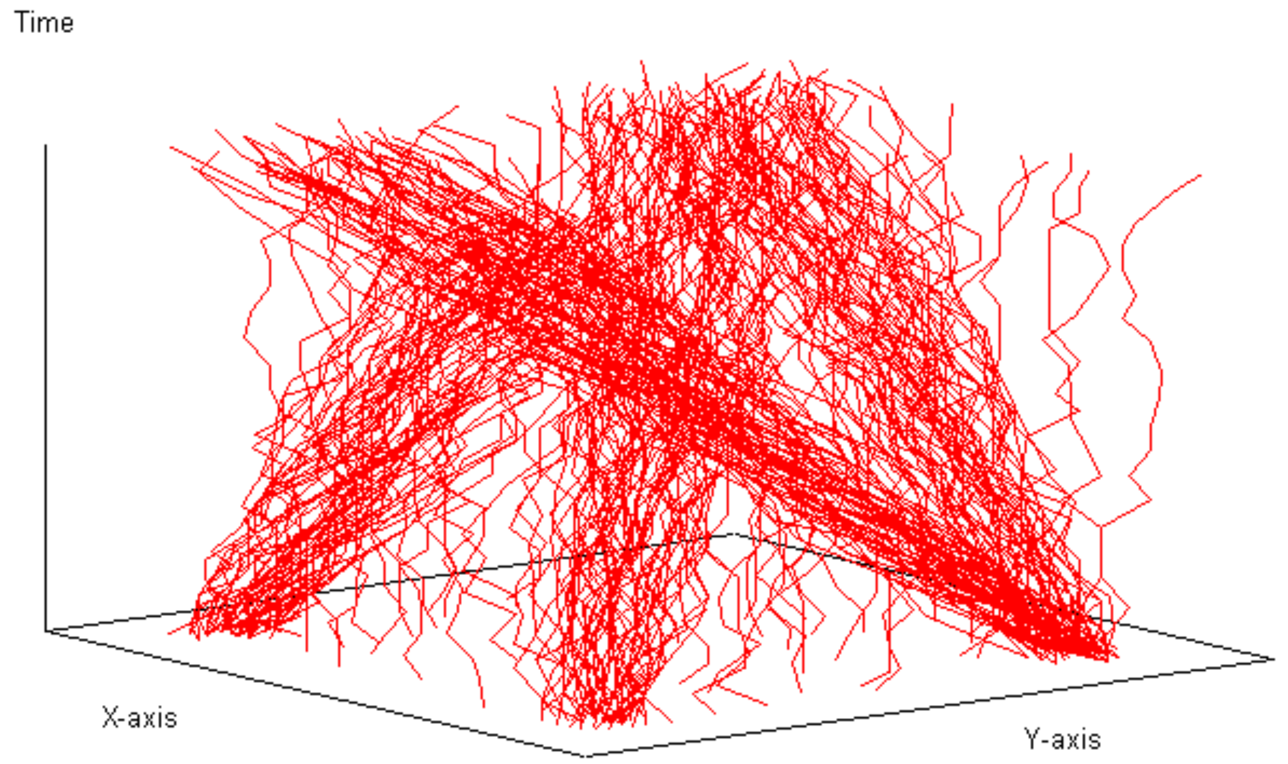


Clusters

- Resistant to Noise
- Can handle clusters of different shapes and sizes

A sample dataset

- A set of trajectories forming 4 clusters + noise (synthetic)



Ad-hoc distance functions

- Colocation –
  - Link prediction,
  - Semantic behaviors,
  - GSM data
- Spatio-temporal Colocation –
  - Link prediction,
  - Semantic behaviors,
  - GSM data

Start and End inclusion

Car Pooling Matching

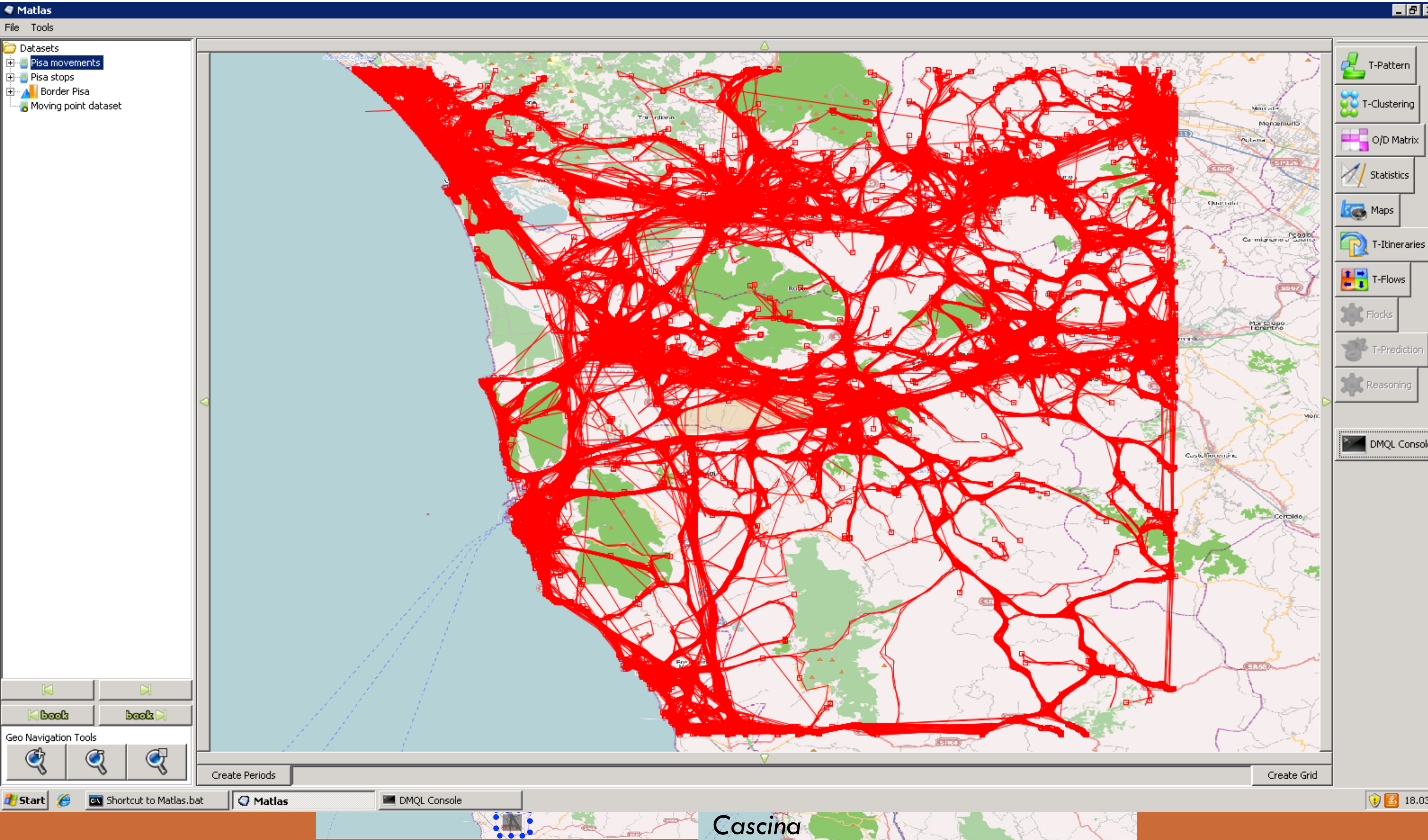
Align to end –

Incoming flows

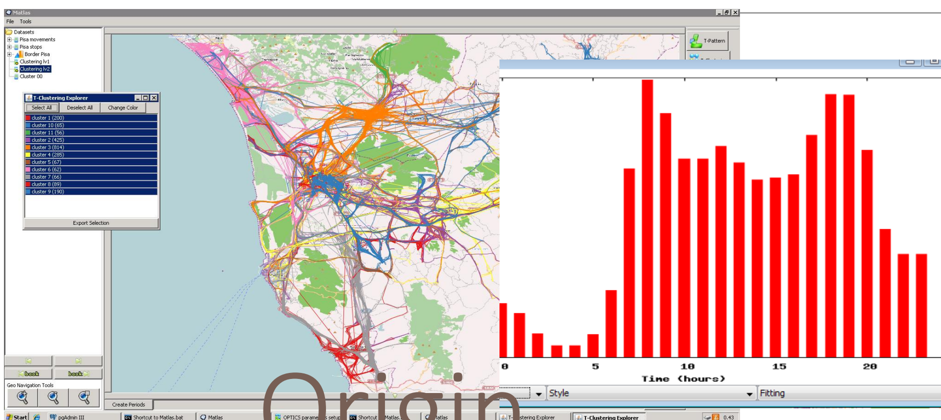
Align to start –

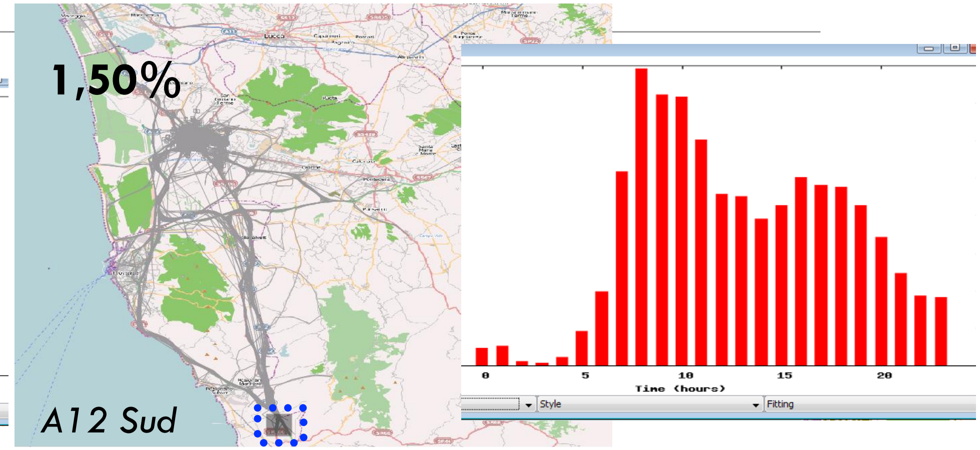
Outcoming flows

Access patterns using T-clustering

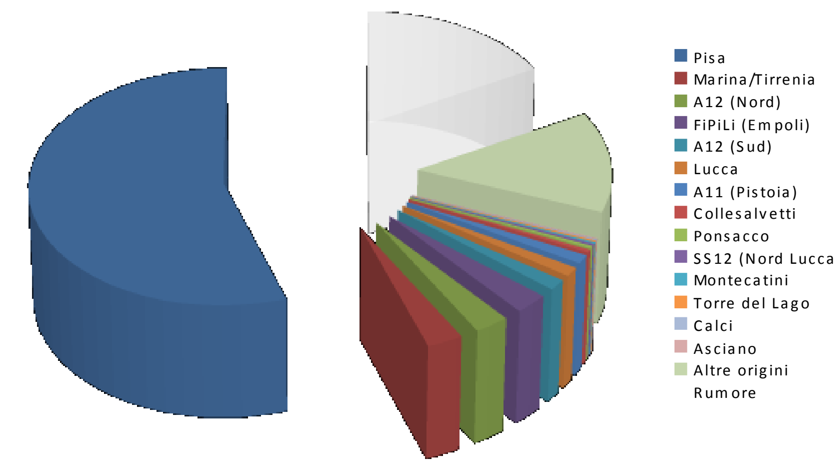


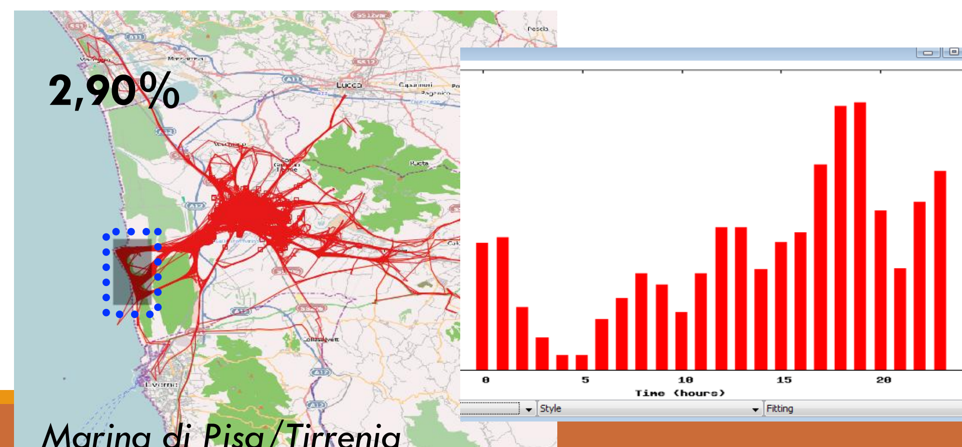
Characterizing the access patterns: origin & time



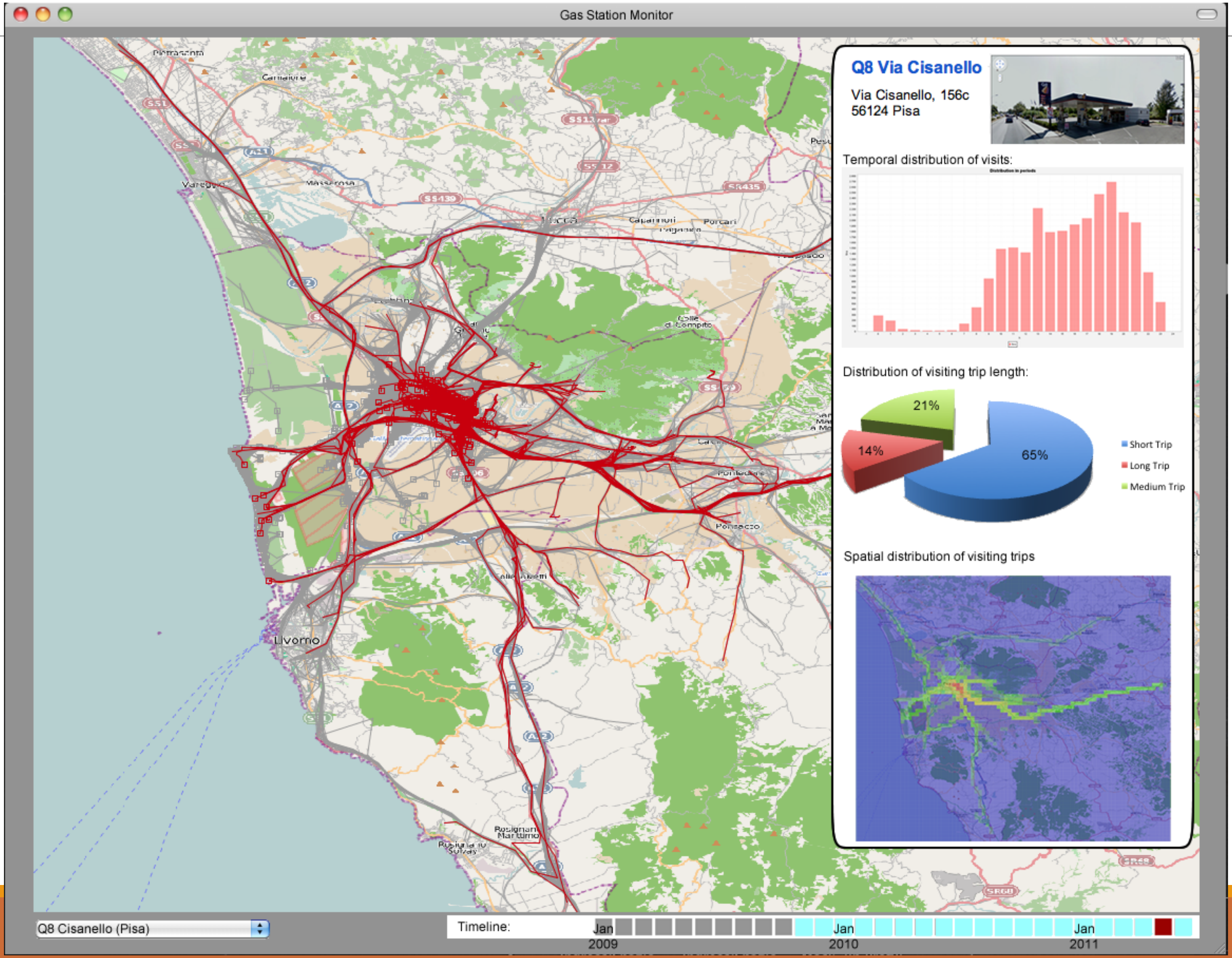


Origin distribution



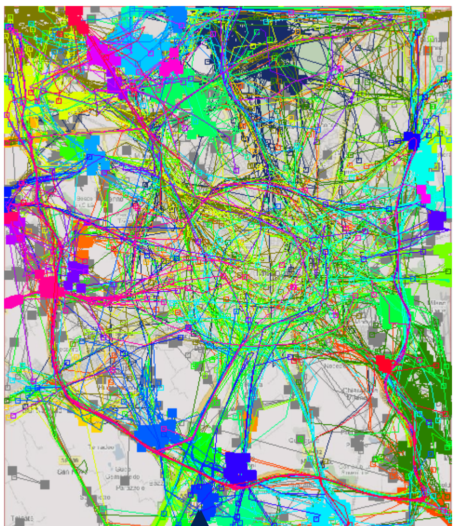


Studying the attractiveness/efficiency of a service with GPS tracks



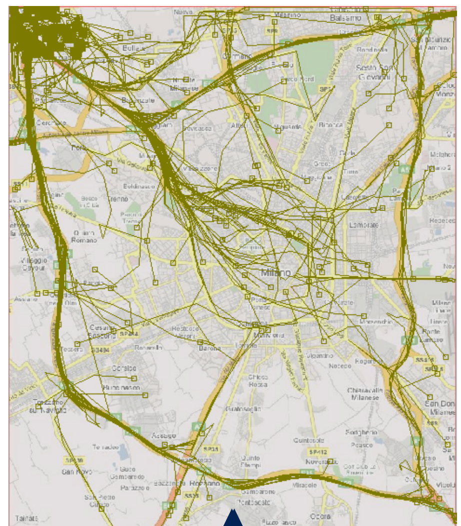
Progressive clustering

- *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)*

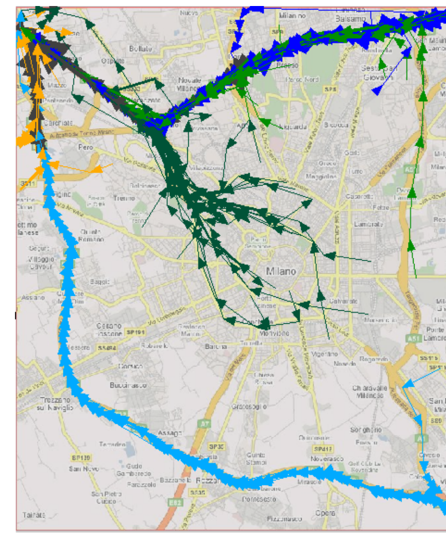


ring Data

n Destination)

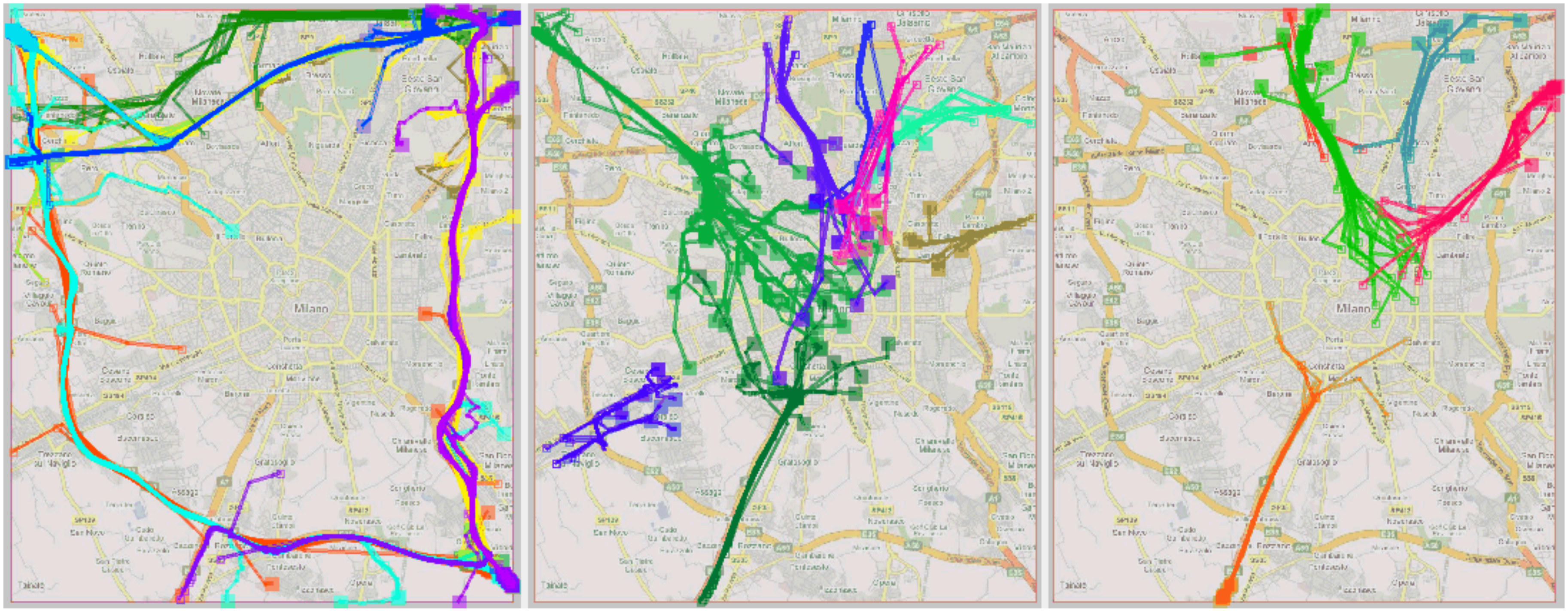


Select a Cluster



Clustering Data

(route similarity)



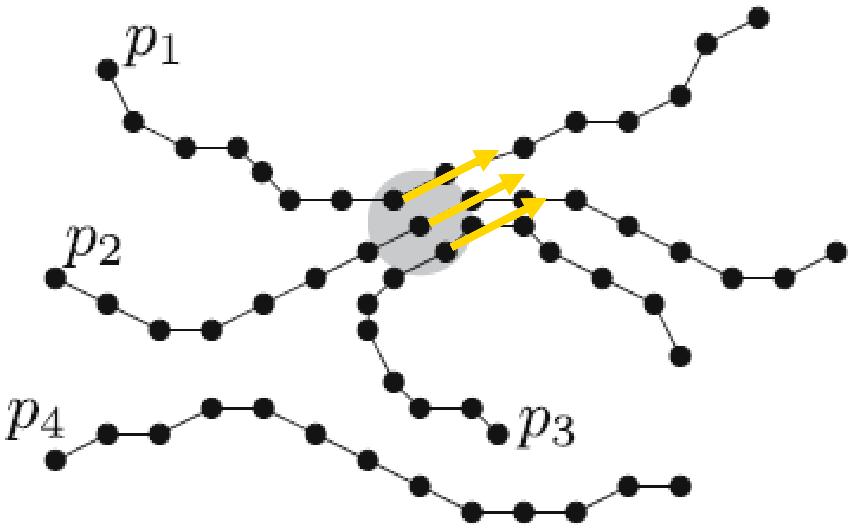
Left: peripheral routes; middle: inward routes; right: outward routes.

Trajectory patterns

Are there groups of objects that move together for some time or in a similar way?

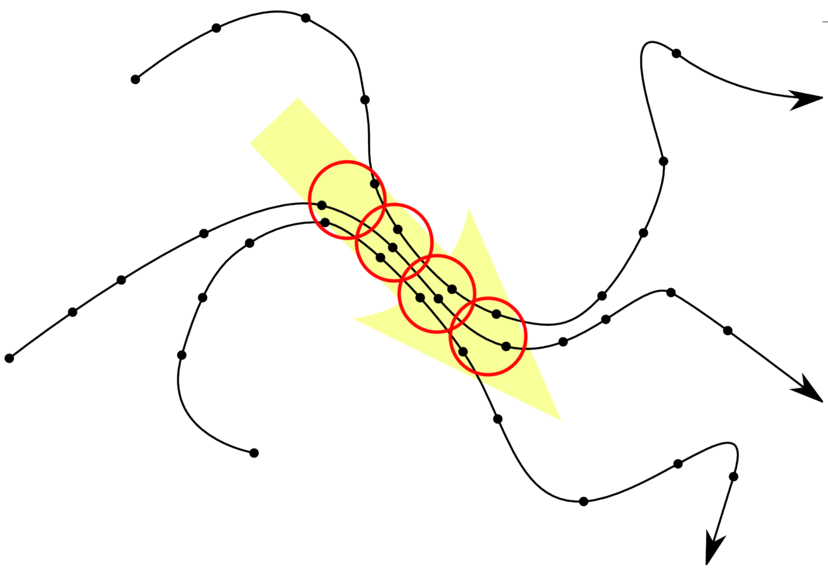
umundsson et al. 07)

- ***Flock*** (*m* > 1, *r* > 0): At least *m* entities are within a circular region of **radius *r*** an



An example of a ***flock*** pattern for *p1*, *p2*, and *p3* at 8th time step; also a ***leadership*** p

Moving Trajectory Flocks

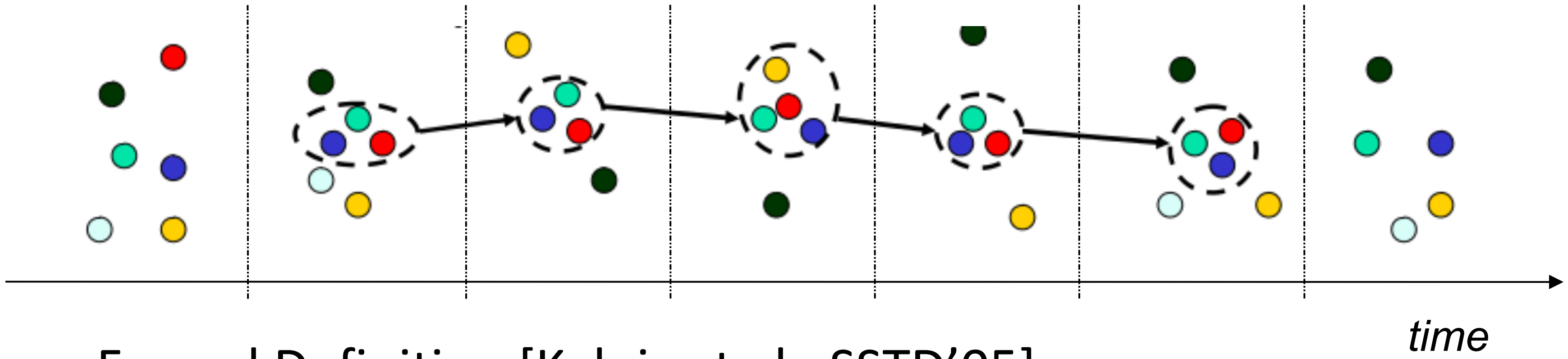


- Group of objects that move together (close to each other) for a time interval

- Discover all possible:
  - sets of objects O, with |O| > min_size and
  - time intervals T, with |T| > min_duration
- such that for all timestamps t ∈T the points in O|t are contained in a circle of radius *r*

Moving Clusters

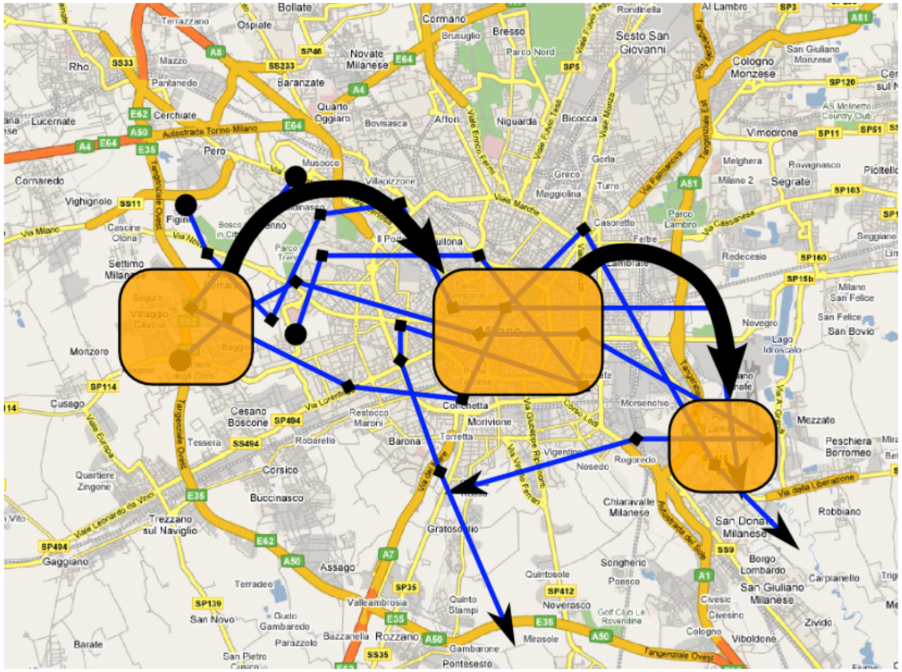
- A ***moving cluster*** is a set of objects that move close to each other f



- Formal Definition [Kalnis et al., SSTD'05]:
  - A ***moving cluster*** is a sequence of (snapshot) clusters ***c1, c2, ..***

T-Patterns

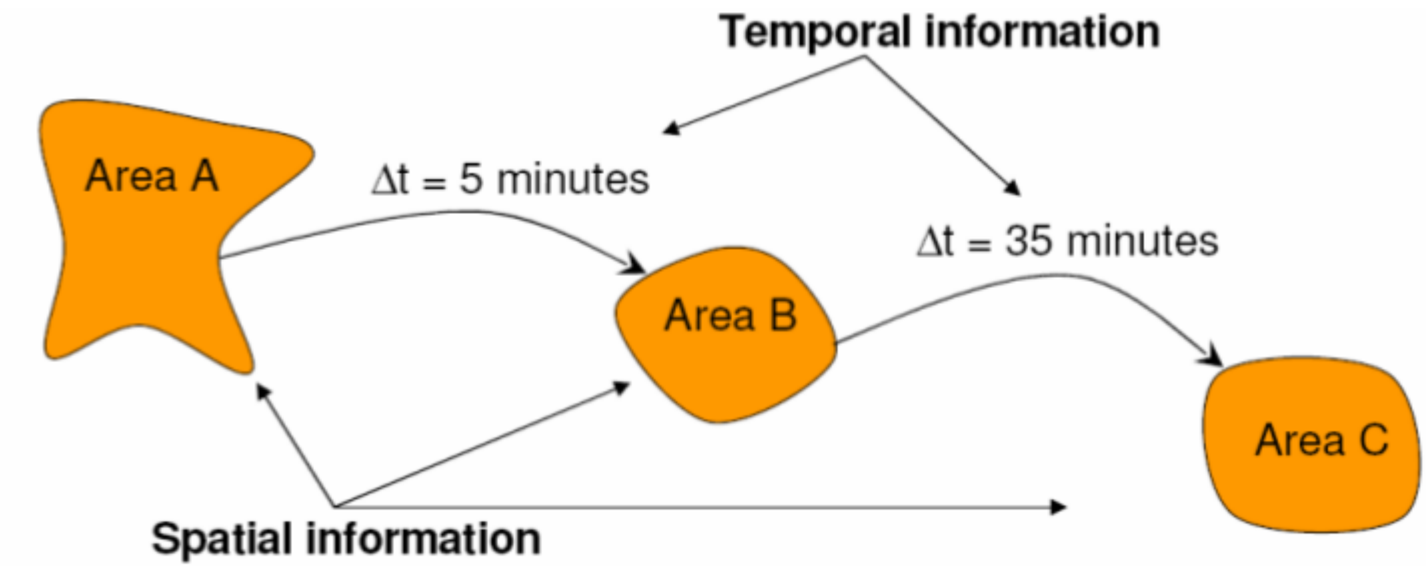
- A sequence of visited regions, **frequently** visited in the **specified order** with **similar transition times**



T-Patterns

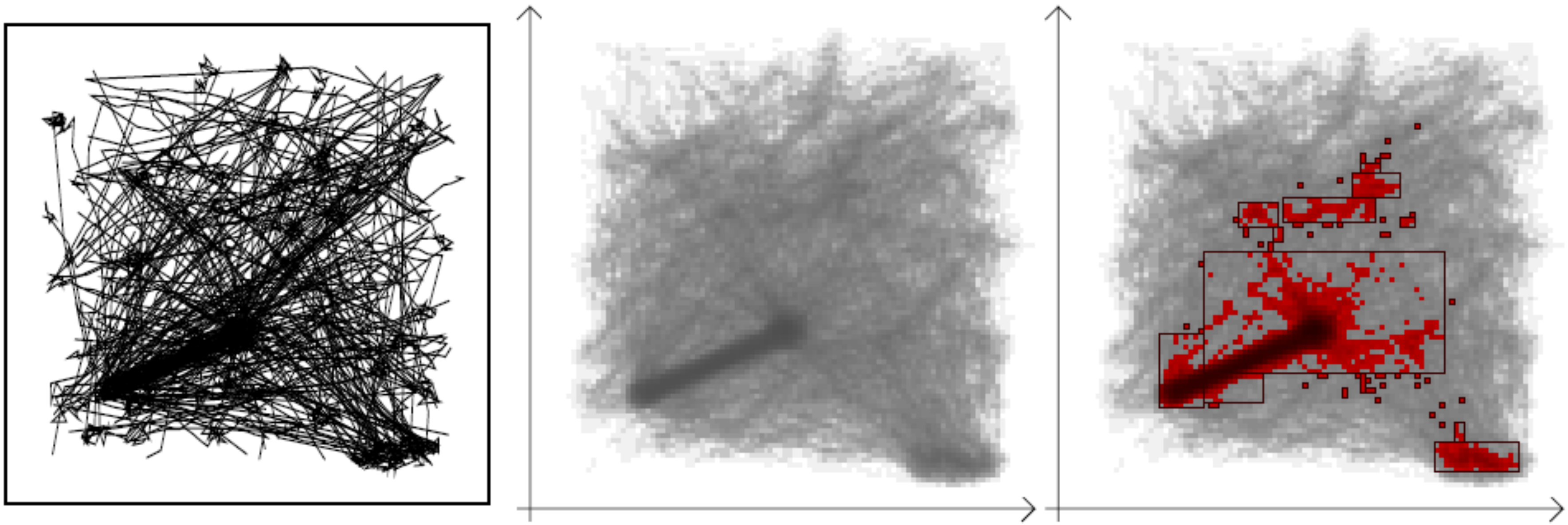
$$A_0 \xrightarrow{t_1} A_1 \xrightarrow{t_2} \dots \quad A_{n-1} \xrightarrow{t_n} A_n$$

- $\boldsymbol{t_i}$ = transition time, $\boldsymbol{A_i}$ = spatial region



$$\mathit{Station} \xrightarrow{20\text{min.}} \mathit{Castle} \xrightarrow{65\text{min.}} \mathit{Museum}$$

Finding regions
A usage-based heuristic



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

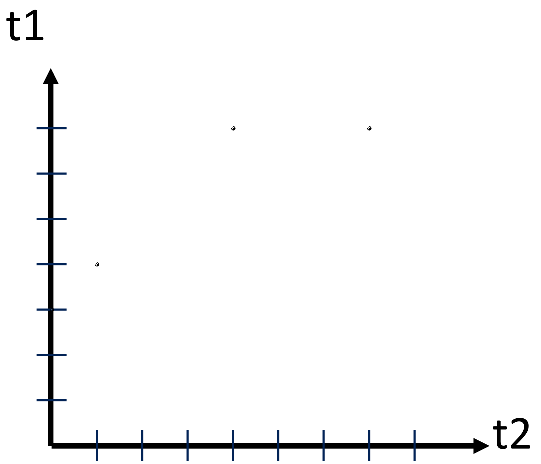
1. Impose a regular grid over space
2. Find dense cells (i.e., touched by many trajs.)
3. Coalesce cells into rectangles of bounded size

Temporal component

Discover frequent time interval using a tolerance (5 min):

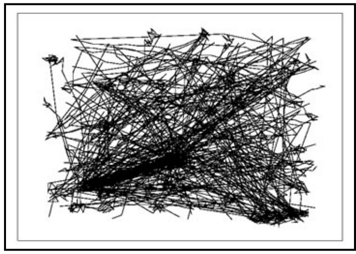
A t1 B t2 C

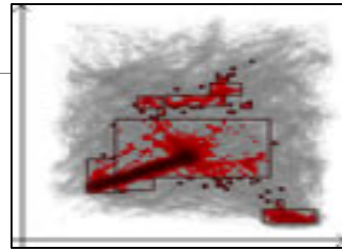
t1	t2
1. 10 min	1. 5 min
2. 15 min	2. 10 min
3. 30 min	3. 25 min
4. 30 min	4. 40 min



A 10-15 min B 5-10 min C

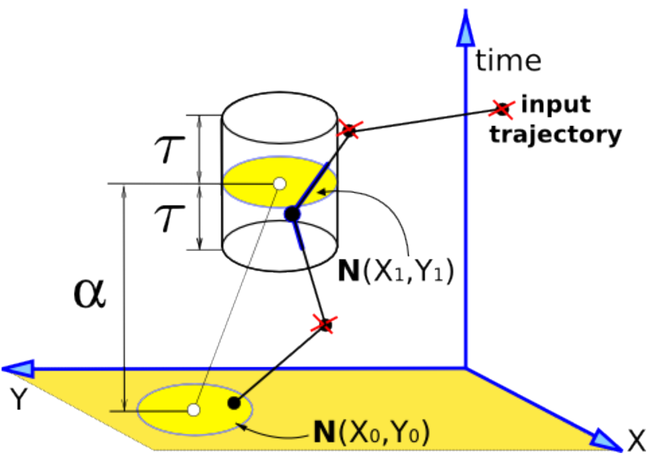
T-Pattern discovery





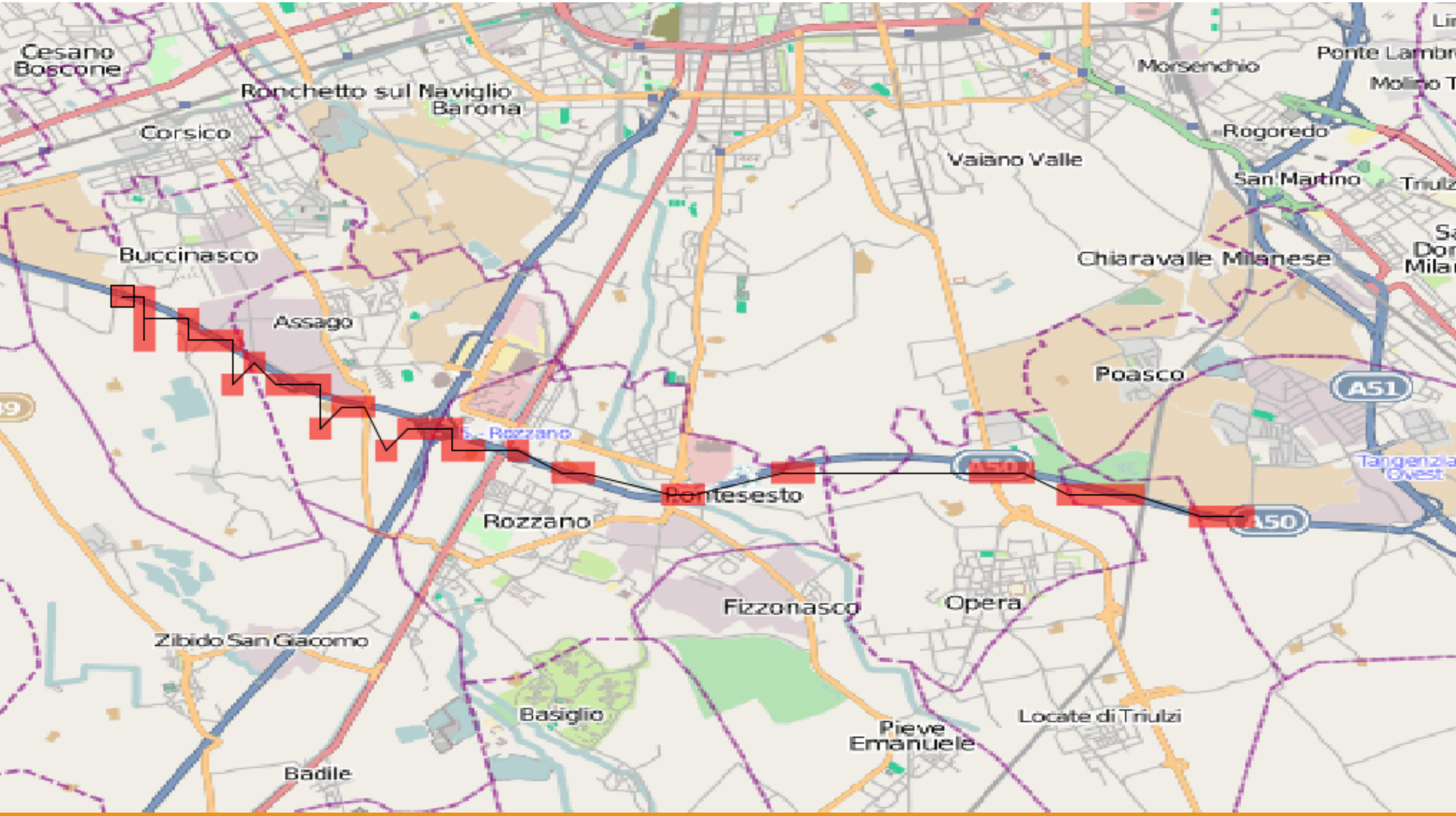
1- Find Regions of Interest

2- Find similar Trajectory in space and time



3- Extract patterns:

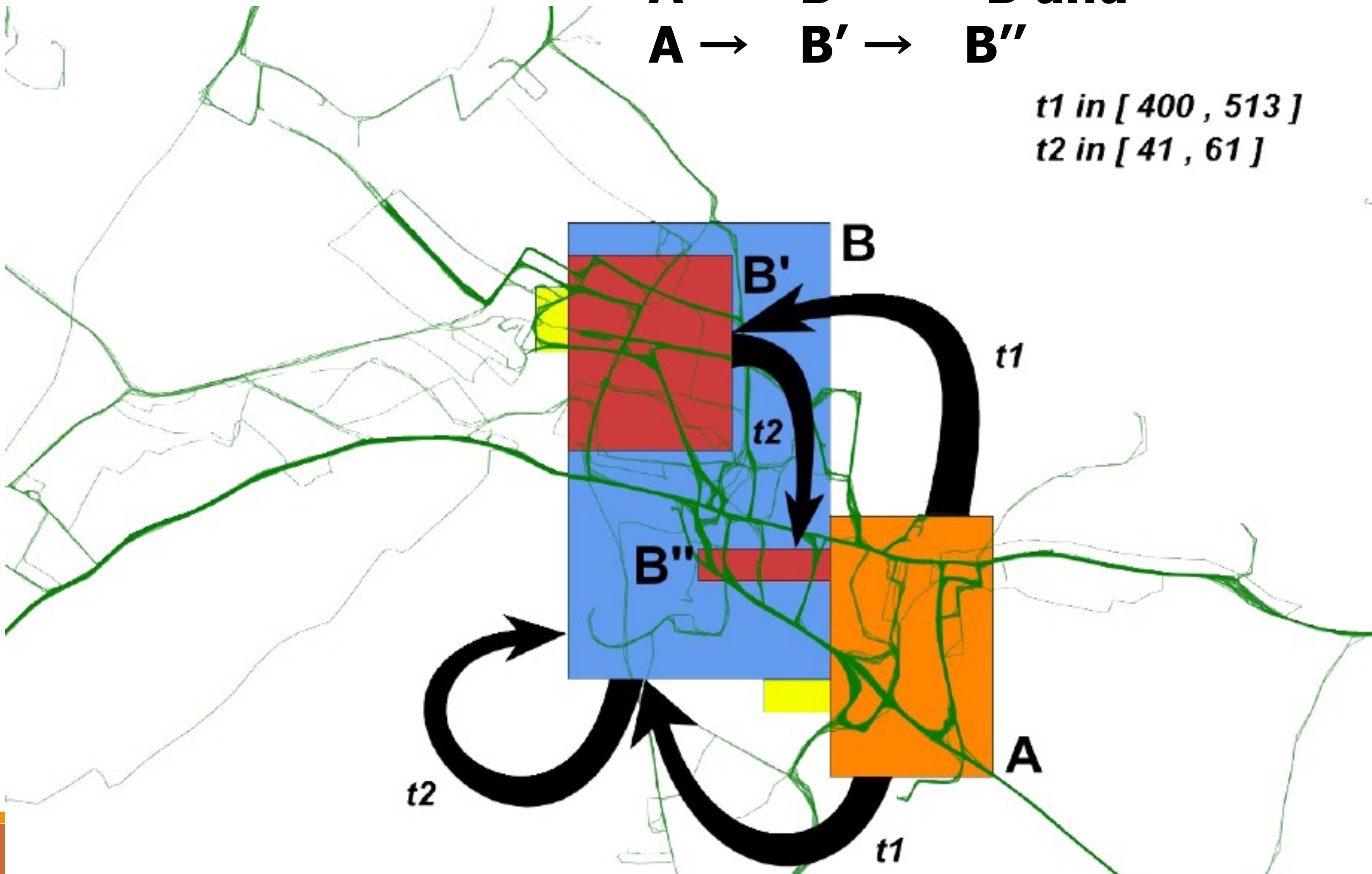
A T-pattern


Cesano Boscone
Ronchetto sul Naviglio
Barona
Corsico
Morsenchio
Rogoredo
Vaiano Valle
San Martino
Buccinasco
Chiaravalle Milanese
Assago
Poasco
A51
Rozzano
A50
Pontesesto
Tangenziale Ovest
Fizzonasco
Opera
Zibido San Giacomo
Basiglio
Pieve Emanuele
Locate di Triulzi
Badile

Sample Trajectory Pattern

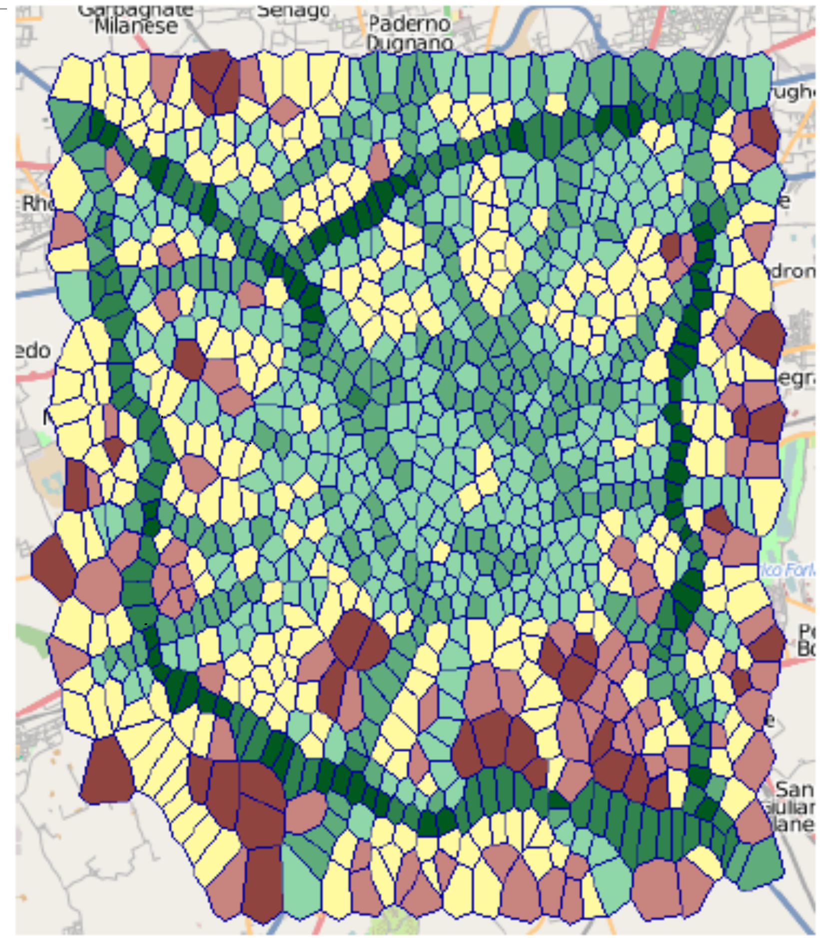
Data Source: Trucks in Athens – 273 trajectories)

A → B → B and
A → B′ → B″



Simpler case: GSM trajectories

- Each trajectory in the database is a time-stamped sequence of **predefined areas** (antennas)



- A T-pattern is a sequence of such areas that occurs

Short bibliography

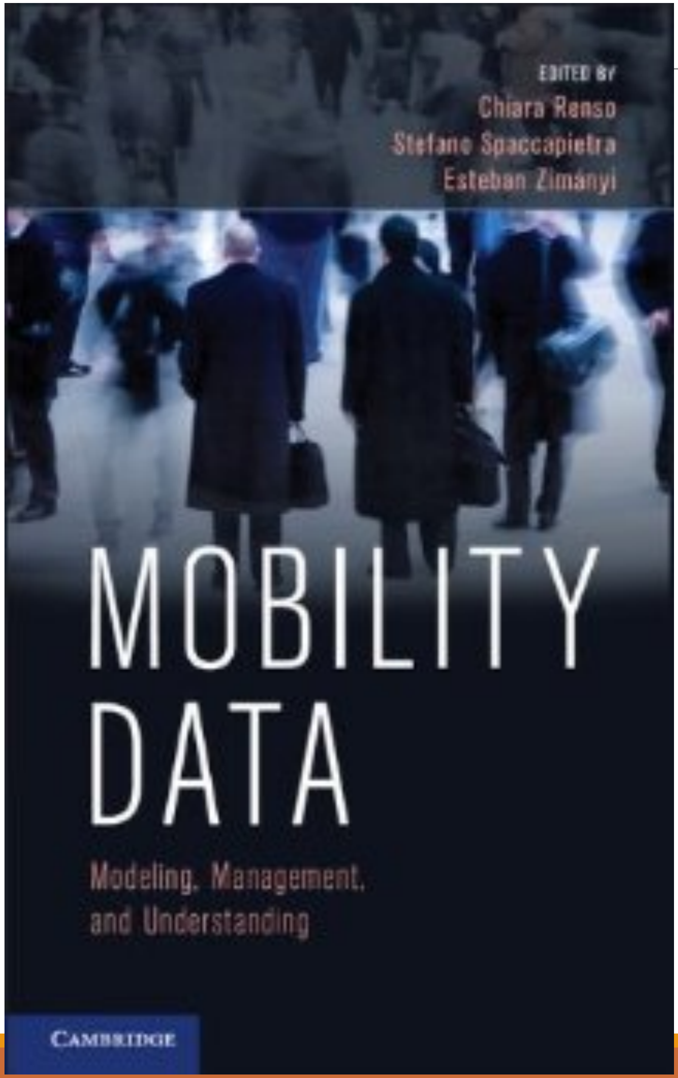
- G. Andrienko, N. Andrienko, S. Rinzivillo, M. Nanni, D. Pedreschi, and F. Giannotti. Interactive visual clustering of large collections of trajectories. In Visual Analytics Science and Technology, 2009. VAST2009. IEEE Symposium on , pages 3{10, 2009.

- F. Giannotti, M. Nanni, D. Pedreschi, F. Pinelli, C. Renso, S. Rinzivillo, and R. Trasarti. Unveiling the complexity of human mobility by querying and mining massive trajectory data. VLDB J. , 20(5):695{719, 2011.

- Fosca Giannotti, Mirco Nanni, Fabio Pinelli, and Dino Pedreschi. Trajectory pattern mining. In KDD , 2007.

- M.Nanni, R.Trasarti, G.Rossetti, and D.Pedreschi. Ecient distributed computation of human mobility aggregates through user mobility proles. In UrbComp12 , 2012.

Mobility data: Modeling, Managing and understanding, Cambridge press.



I. **Mobility Data Modeling and Representation**
Trajectories and their Representations, S. Spaccapietra, C. Parent, L. Spinsanti
Trajectory Collection and Reconstruction, G. Marketos, M.L Damiani, N. Pelekis, Y. Theodoridis, Z
Trajectory Databases, R.H. Guting, T. Behr, C. Duntgen
Trajectory Data Warehouses, A.A. Vaisman, E. Zimányi
Mobility and Uncertainty, C. Silvestri, A.A. Vaisman

II. **Mobility Data Understanding**
Mobility Data Mining, M. Nanni
Understanding Human Mobility using Mobility Data Mining, C. Renso, R. Trasarti
Visual Analytics of Movement: A Rich Palette of Techniques to Enable Understanding, N. Andrien
Mobility Data and Privacy, F. Giannotti, A. Monreale, D. Pedreschi

III. Mobility Applications
Car Traffic Monitoring, D. Janssens, M. Nanni, S. Rinzivillo
Maritime Monitoring, T. Devogele, L. Etienne, C. Ray
Air Traffic Analysis, C. Hurter, G. Andrienko, N. Andrienko, R.H. Guting, M. Sakr
Animal Movement, S. Focardi, F. Cagnacci
Person Monitoring with Bluetooth Tracking, M. Versichele, T. Neutens, N. Van de Weghe

IV. Future Challenges and Conclusions
A Complexity Science Perspective on Human Mobility, F. Giannotti, L. Pappalardo, D. Pedreschi, I
Mobility and Geo-Social Networks, L. Spinsanti, M. Berlingerio, L. Pappalardo
Conclusions, C. Renso, S. Spaccapietra, E. Zimànyi

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