Trajectory Mining

Pascal Poncelet
Univ Montpellier – LIRMM
Pascal.Poncelet@lirmm.fr
www.lirmm.fr/~poncelet

Background and Motivation

- Nowadays, many electronic devices are used for real world applications

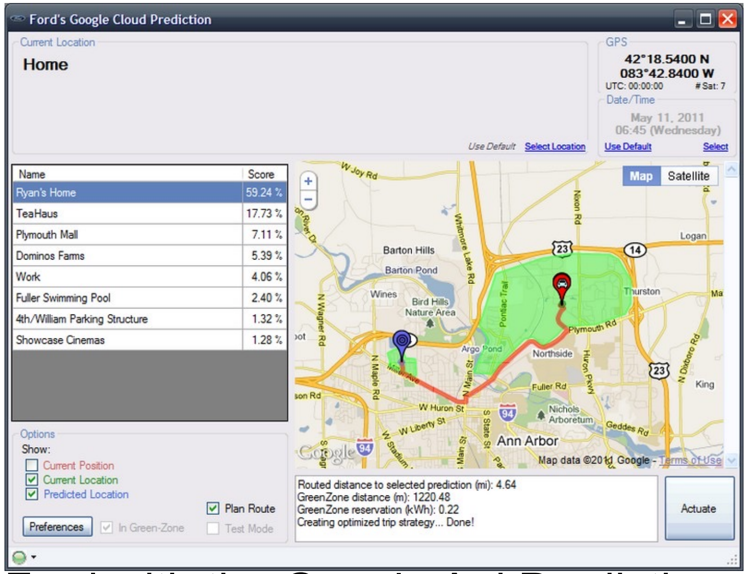
  GPS, sensor networks, mobile phone, ..

- « interesting » patterns for:

  movement pattern analysis, animal behavior, route planning and vehicle control, location prediction, …

Some Examples

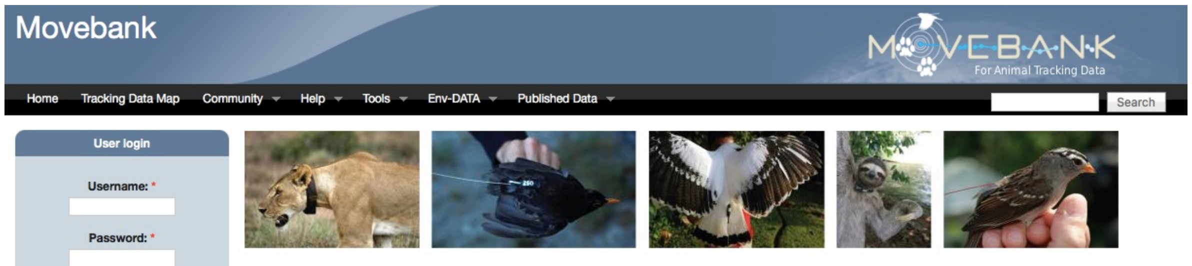
Photo: Ford and Google



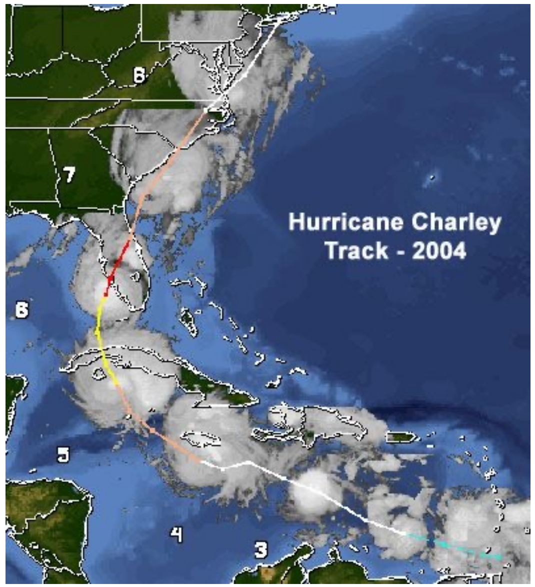
Ford with the Google Api Prediction

Photo: cdn.benzinga.com

The world's largest traffic jam in history (China)



Free, online database of animal tracking data
Animal migration analysis



Hurricane Trajectory Prediction

Some Examples

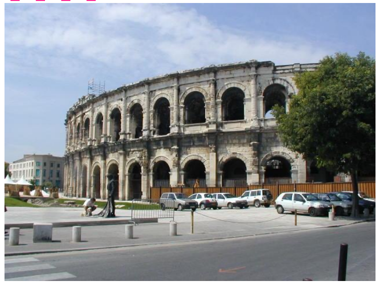
- Location Prediction

According to the frequent behavior of the Second Big Data School attendees, the next place, after São Carlos to visit is:

Montpellier and its Region !!!!

Antigone

Place de la Comédie



Nîmes

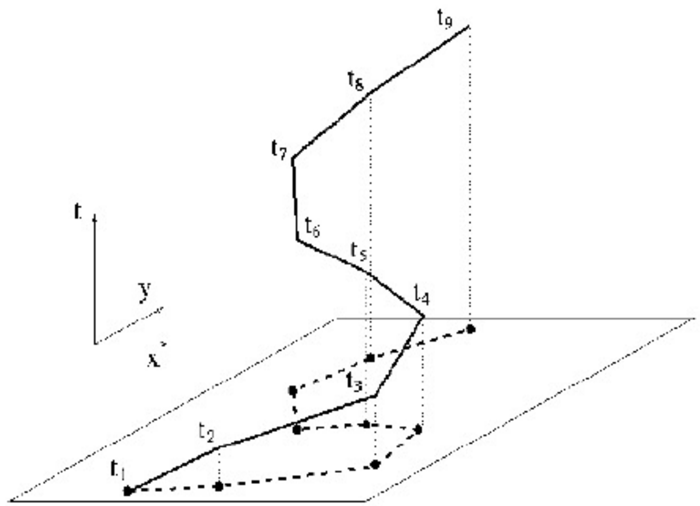
Beaches

Pont du Diable

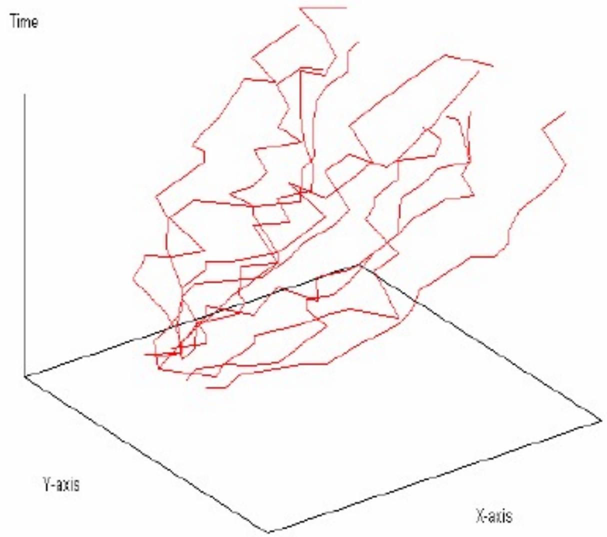
Sète

Spatio-Temporal Data

- Represented as a set of points, located in space and time.
  - $T=(x_1,y_1, t_1), \ldots, (x_n, y_n, t_n) \rightarrow$ position in space at time t_i was (x_i, y_i).



Tid	position (x,y)		time (t)
1	48.890018	2.246100	08:25
1	48.890018	2.246100	08:26
...	...		...
1	48.890020	2.246102	08:40
1	48.888880	2.248208	08:41
1	48.885732	2.255031	08:42
...	...		...
1	48.858434	2.336105	09:04
1	48.853611	2.349190	09:05
...	...		...
2	...		...



Using Ad-hoc Queries

- Spatial query extensions in GIS applications

  « find all the moving objects inside area A between 10:00 am and 2:00 pm »

  « how many cars were driven between Main Square and the Airport on Friday? »

- Find the best solution by exploring each spatial object at a specific time according to some metric distance measurement

Capture Collective Behavior

- Extracting patterns which capture ’group’ or ’common’ behaviour among moving entities:

Trajectories

- Development of approaches to identify groups of moving objects

strong relationship and interaction exist within a defined spatial region during a given time duration

Outline

- A quick reminder on pattern discovery
- The different kinds of patterns from trajectory
- Towards a unified approach for extracting trajectories
- Deal with and without a spatial component?
- A concrete illustration
- Conclusion

Outline

- **A quick reminder on pattern discovery**
  - **Association Rules**
  - **Sequential Patterns**
- The different kinds of patterns from trajectory
- Towards a unified approach for extracting trajectories
- Deal with and without a spatial component?
- A concrete illustration
- Conclusion

Association Rule Mining

- **Items :** A, B, C, D, E, F
- 4 **transactions** (subset of items)
  - T1 : {A,D}
- **Support** for an itemset
  - Supp ({A,D})=1/4
  - Supp ({A,C})=2/4
- **Frequent Itemsets** (minSupp=50%)
  - {A,C} is a frequent itemset
- **Rules :** (minSupp and minConf = 50%)
  - A $\rightarrow$ C [50%, 50%]
  - C $\rightarrow$ A [50%, 100%]

Trans ID	Items
1	A, D
2	A, C
3	A, B, C
4	A, B, E, F

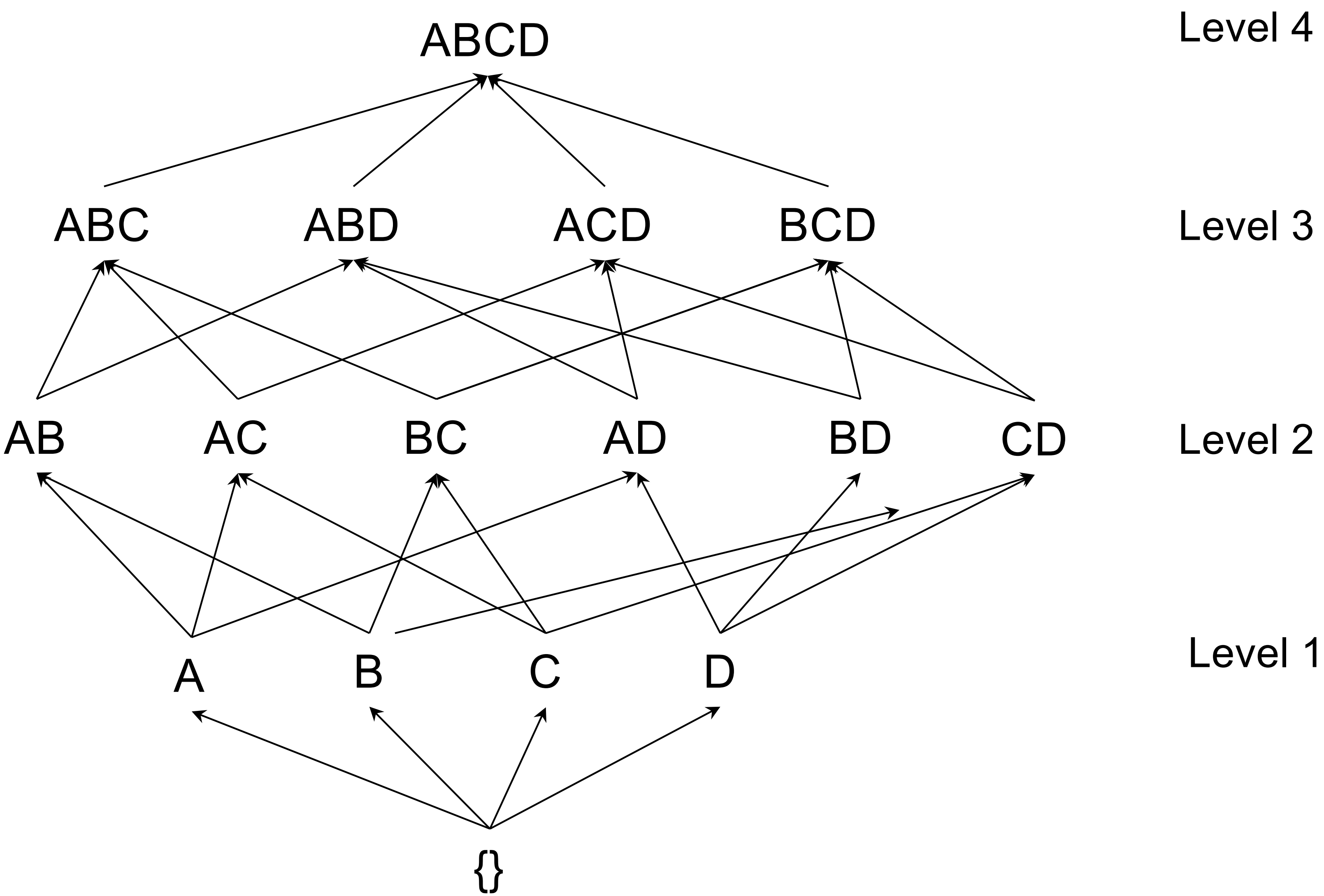
Association Rule Mining

- The corresponding matrix with the frequent itemset {A,C}

	A	B	C	D	E	F
1	1	0	0	1	0	0
2	1	0	1	0	0	0
3	1	1	1	0	0	0
4	1	1	0	0	1	1

Trans ID	Items
1	A, D
2	A, C
3	A, B, C
4	A, B, E, F

The Search Space

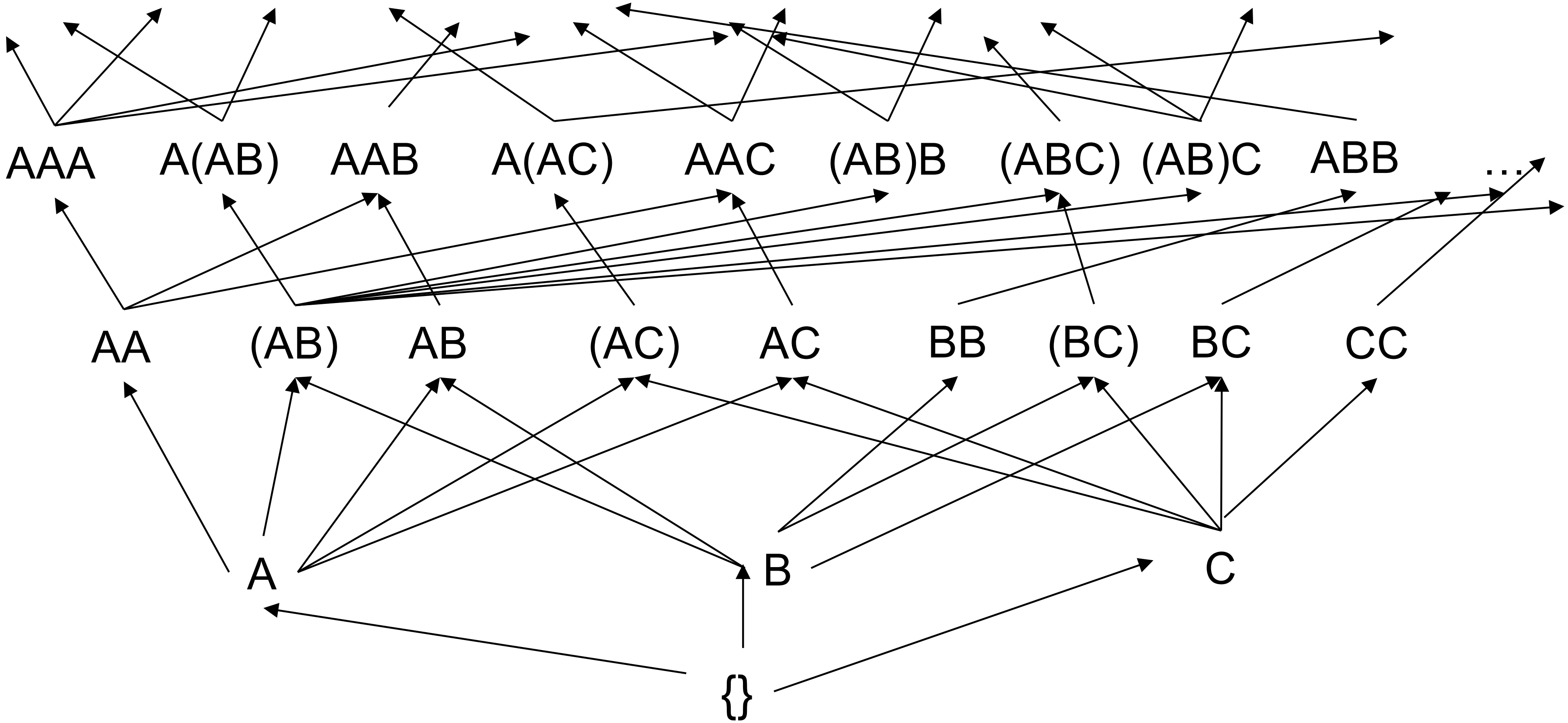


Sequential Pattern Mining

- **Items :** A, B, C, D, E, F
- **Support** of a sequence
  - Supp((A)(C))=2/4
  - Supp((A)(B)=2/4
- **Frequent sequences** (minSupp=50%) :
  - (A) (C) and (A) (B) are frequent sequences

Trans ID	Sequences
1	(A, D)
2	(A) (C) (A) (C)
3	(A) (B, C)
4	(A) (E, F) (B)

The Search Space for SP



What kinds of patterns can be mined?

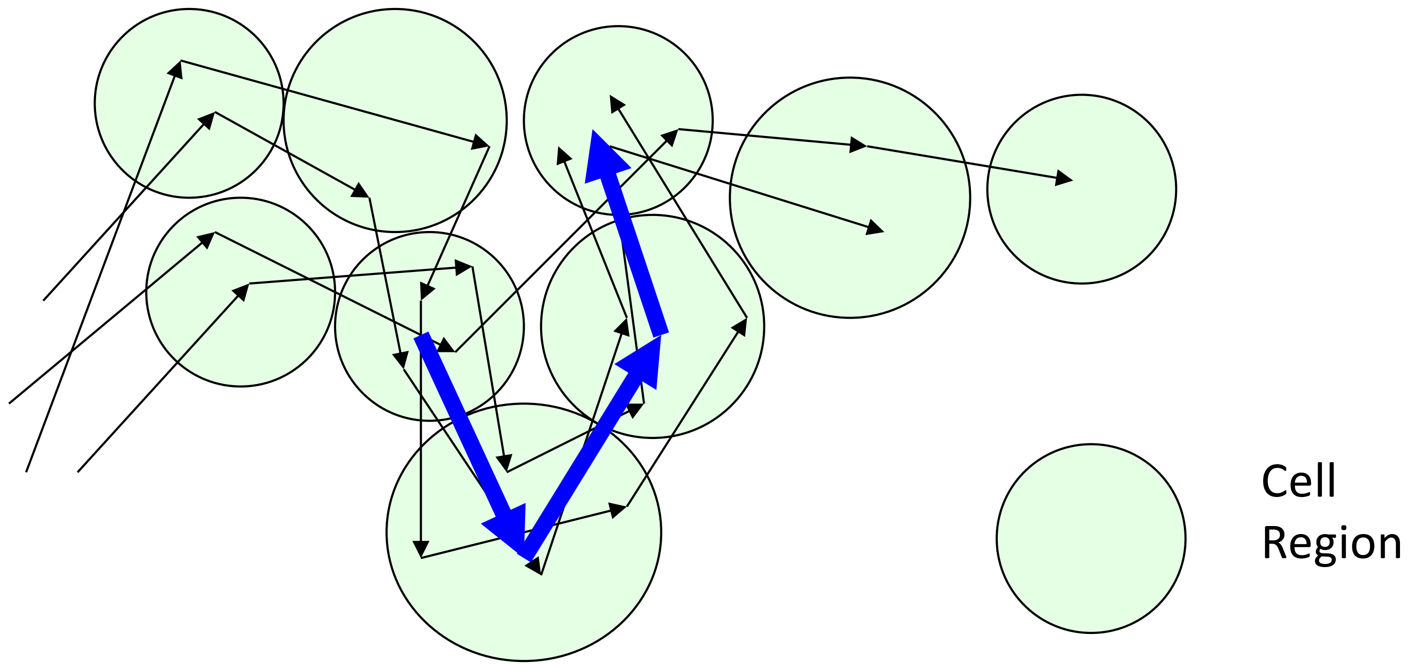
- 1, {A,B,C,D}
  - Association rules
- 1, <(A,t1) (B,C, t2) (D, t3)>
  - Sequences
- What can we do with the spatial dimension?
  - 1, <(x1,y1,t1) (x2,y2,t2) (x3,y3,t3)>

Outline

- A quick reminder on pattern discovery
- **The different kinds of patterns from trajectory**
  - **Frequent Patterns**
  - **Moving Object**
- Towards a unified approach for extracting trajectories
- Deal with and without a spatial component?
- A concrete illustration
- Conclusion

Mining Spatio-Temporal Patterns from Trajectory Data

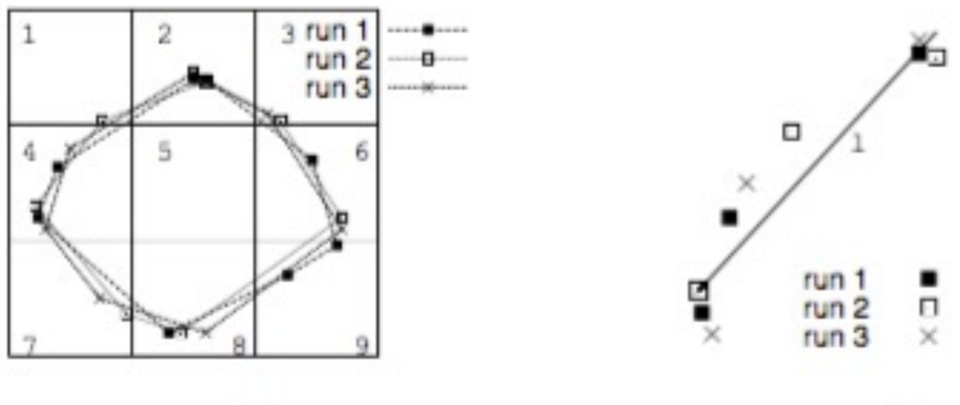
- Frequent Patterns:
  - Frequent followed paths



Frequent Spatio-Temporal Sequence

- One spatio-temporal sequence:

S = <(x1,y1,t1) (x2,y2,t2) ….. (xn,yn,tn)>



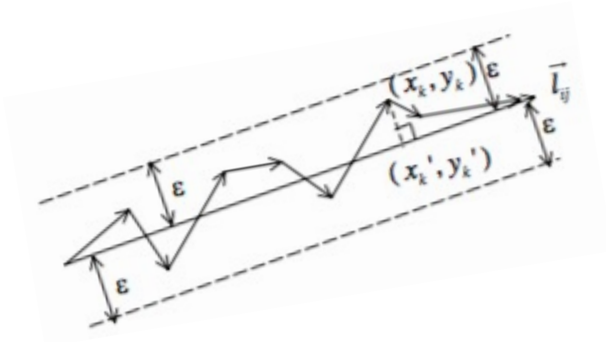
(a) (b)

- Movement patterns that occurs frequently in the sequence: Frequent spatio-temporal sequence

H. Cao et al., ICDM 05

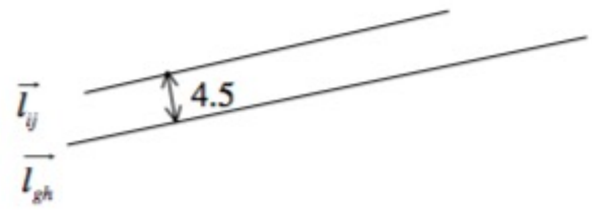
Frequent Spatio-Temporal Sequence

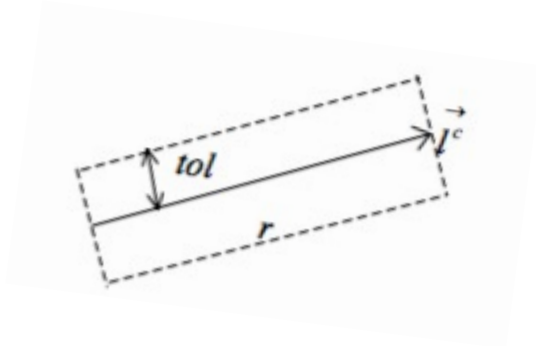
- A three step process
- Transform each trajectory in a line with several segments
  - A distance tolerance measure ε is defined
  - All trajectory points inside this distance can be projected in a representative line segment



Frequent Spatio-Temporal Sequence

- Group together similar segments and define a region
  - The similarity is based on the angle and the spatial length of the segment
  - A line segment with same angle and length is close to another one if their distance is lower than a distance threshold **d**
  - A medium segment is created, i.e. a region
- Compute frequent sequences of regions (minSupp)

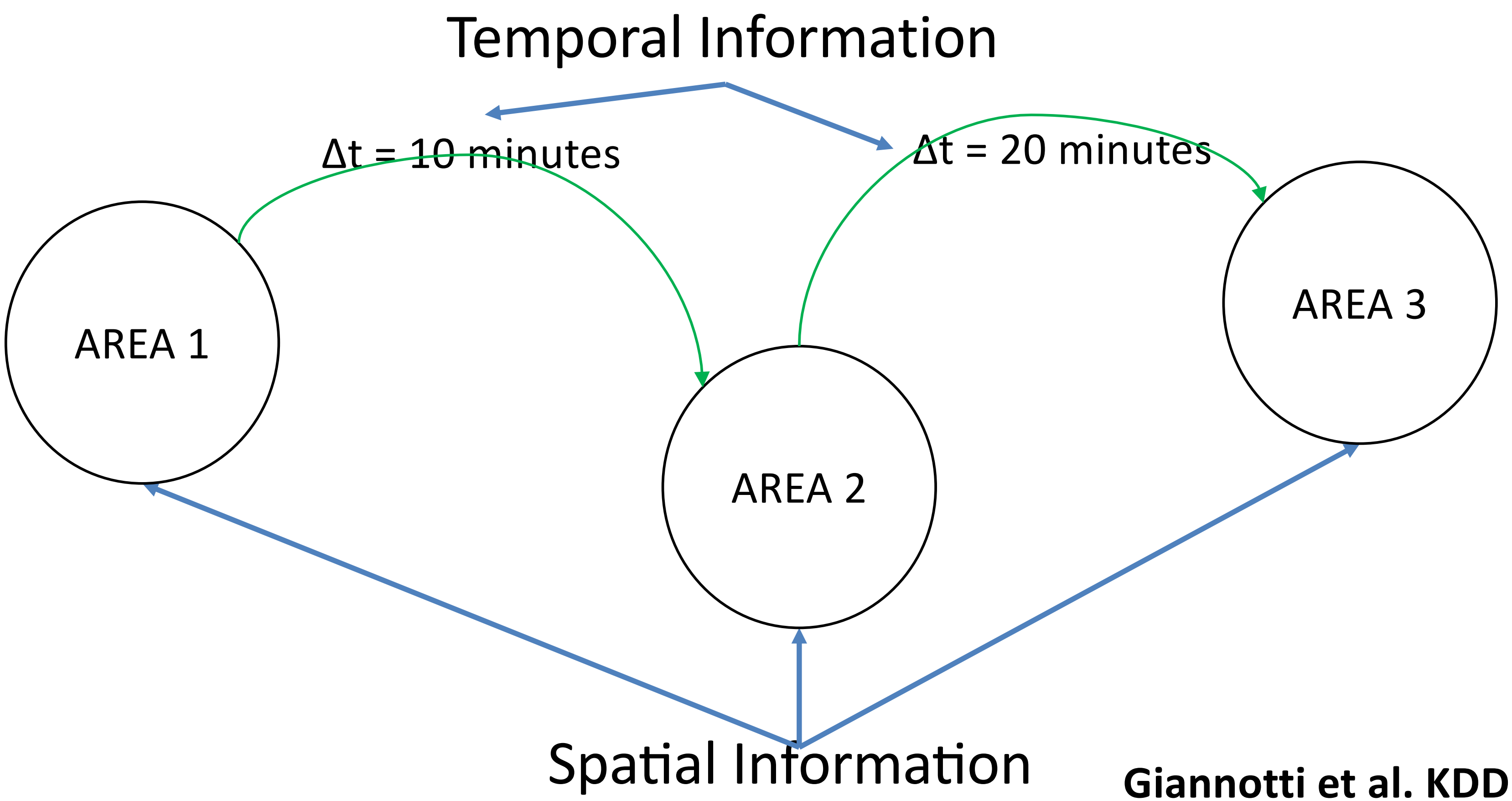




r1-r2-r3-r1-r2 => r1, r2, r1-r2
(minSupp=2)

Trajectory Pattern Mining

- A trajectory pattern describes the movements of objects both in space and in time



Giannotti et al. KDD 07

Trajectory (T-) Patterns

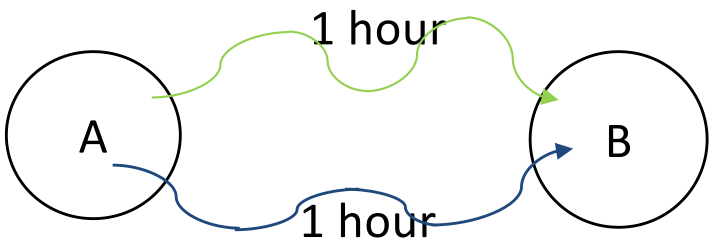
- A *Trajectory Pattern* (*T-pattern*) is a couple (s,α):
  - $s = <(x0,y0),\ldots,(xk,yk)>$ is a sequence of k+1 locations
  - $\alpha = <\alpha 1,\ldots,\alpha k>$ are the transition times

$$(x0,y0) \xrightarrow{\alpha 1} (x1,y1) \xrightarrow{\alpha 2} \ldots\ldots \xrightarrow{\alpha k} (xk,yk)$$

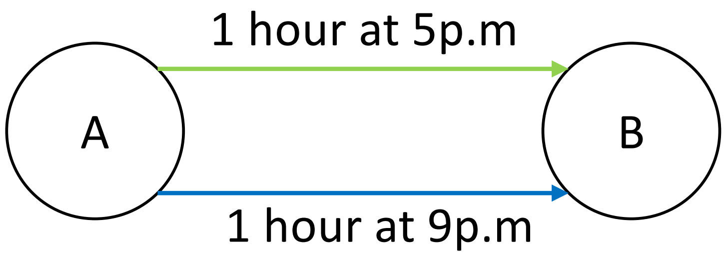
- A T-pattern *Tp occurs* in a trajectory if the trajectory contains a subsequence *S* such that:
  - Each (xi,yi) in *Tp* matches a point (xi',yi') in *S*, and the transition times in *Tp* are similar to those in *S*

Characteristics of Trajectory-Patterns

- Routes between two consecutive regions are not relevant

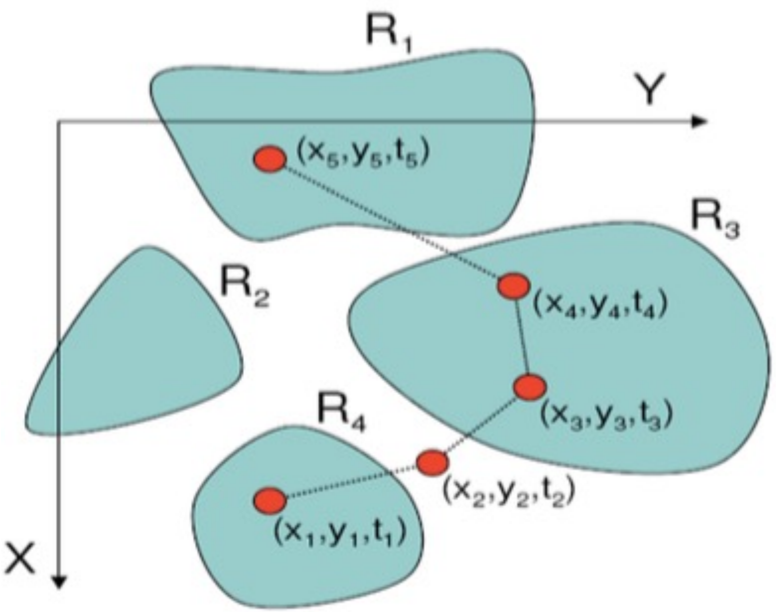


- Absolute times are not relevant



Trajectory-Pattern Mining

- A preprocessing step is performed :
  - Convert each trajectory to a sequence, i.e., by converting a location (x, y) into a region



$$S=\langle(x_1,y_1,t_1),\ \ldots,\ (x_5,y_5,t_5)\rangle$$

$$\downarrow$$

$$\langle(R_4,t_1),\ (R_3,t_3),\ (R_3,t_4),\ (R_1,t_5)\rangle$$

A Neighborhood Function

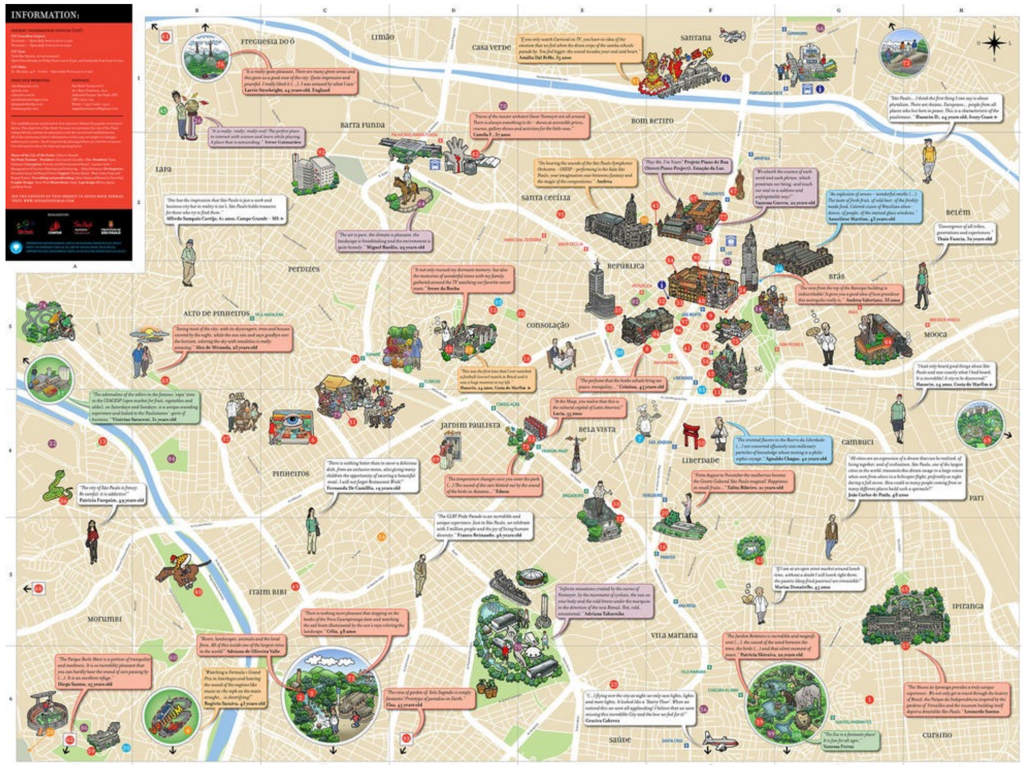
- A preprocessing step

- Neighborhood Function
  - Calculates spatial containment of regions
  - Input point to find enclosing Region of Interest
  - Defines the necessary proximity to fall into a region
  - Parameters: e – radius or necessary proximity of points

- Translate each set of points into regions
- Timestamp is selected from when the trajectory first entered the region

Using prior Knowledge

- Static Region of Interest
  - Initially receives set of **R** disjoint spatial regions
  - **R** regions are predefined based on prior knowledge
  - Each represents relevant place for processing



Popular Regions

- Data sets often do not possess predetermined regions
- Regions based on criteria of density of the trajectories
  - A grid G of n x m cells
  - A density Threshold **d**
  - Each cell with density G(I,j)

- Each region forms rectangular region
- Dense cells always contained in some region
- All regions have average density above **d**
- All regions cannot expand without their average density decreasing below **d**

Grid Density Preparation

- Split space into n x m grid with small cells
- Increment cells where trajectory passes
- Neighborhood Function NR() determines which surrounding cells
- Regression - increment continuously along trajectory

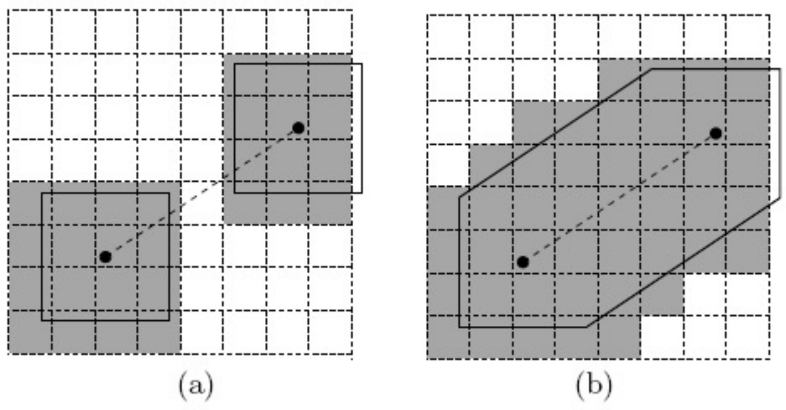


Figure 2: Density with and without regression

Popular Regions Algorithm

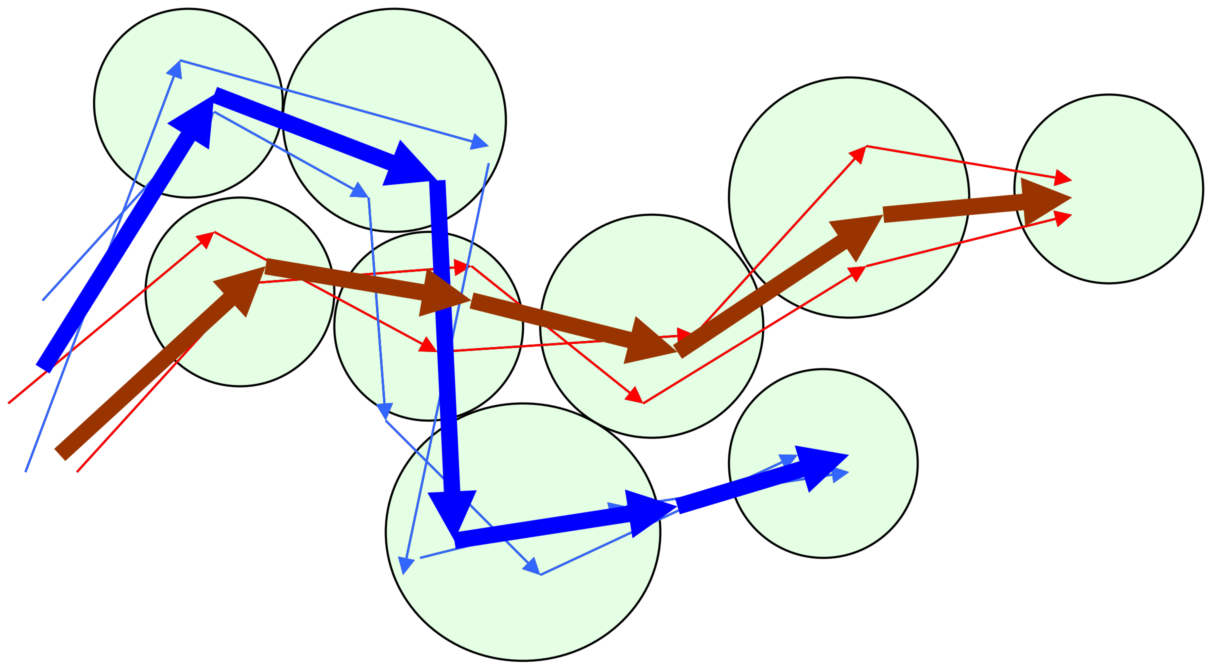
- Iteratively consider each dense cell
- For each:
  - Expands in all four directions
  - Select expansion that maximizes density
  - Repeat until expansion would decrease below density threshold

Trajectory-Pattern Mining

- After the preprocessing
- Execute a TAS (temporally annotated sequence) algorithm
  - A TAS is a sequential pattern annotated with typical transition times between its elements
  - The algorithm of TAS mining is an extension of PrefixSpan

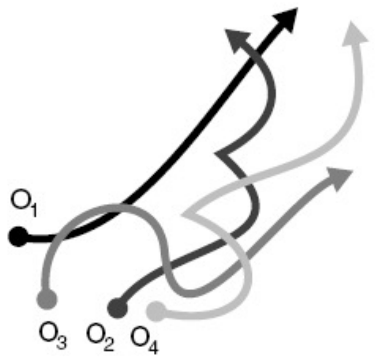
Mining Spatio-Temporal Patterns from Trajectory Data

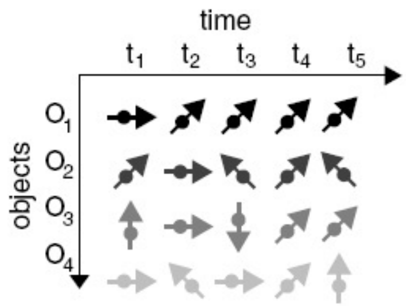
- Clustering:
  - Group together similar trajectories
  - For each group produce a summary



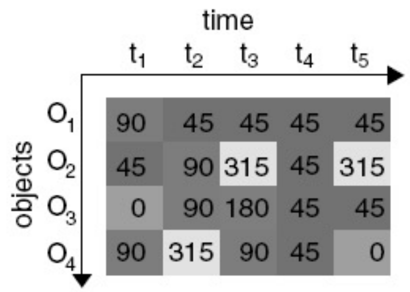
Relative Motion Patterns

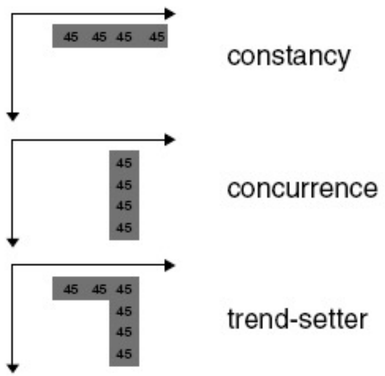
- Identify similar movements in a collection of moving-object trajectories
- Transform raw trajectories into motion attributes (speed, motion azimuth)





The REMO Matrix

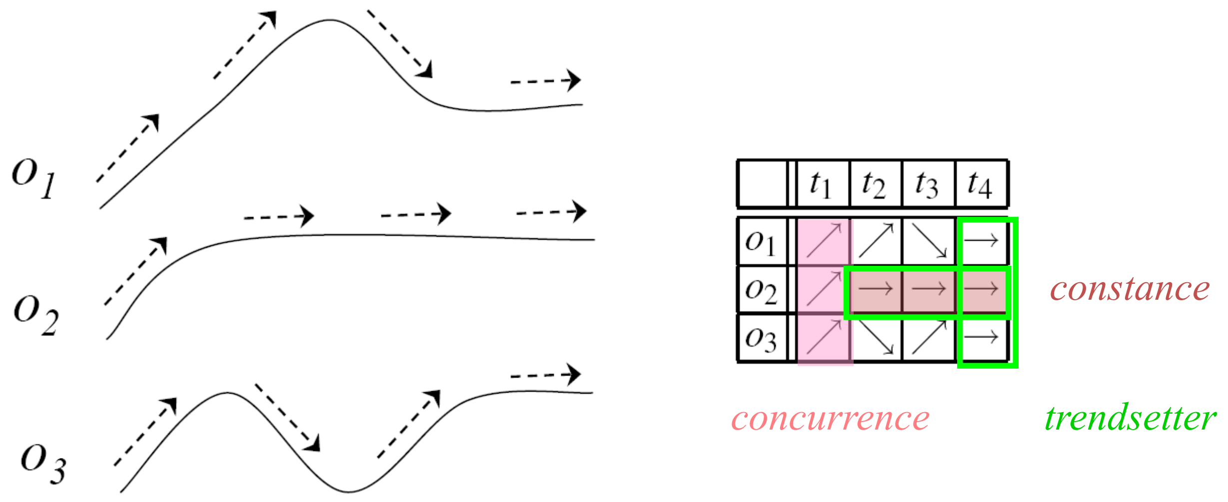




Laube, P. et al. GIScience, 2002
Laube, P. et al., IJGIS, 2005

Basic Motion Patterns

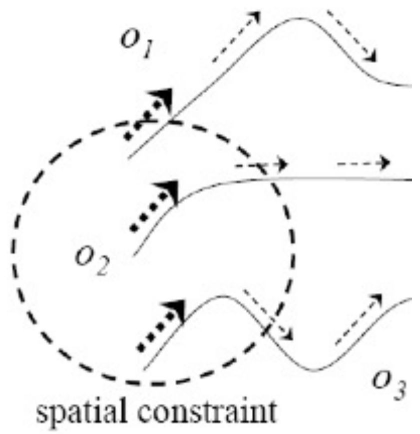
- Describing motion events, disregarding absolute positions
  - **Constance**: a sequence of equal motion attributes for consecutive times
  - **Concurrence**: the incidence of multiple objects with the same motion attributes
  - **Trendsetter**: a certain motion pattern that is shared by a set of other objects in the future. E.g., “constance” + “concurrence.”



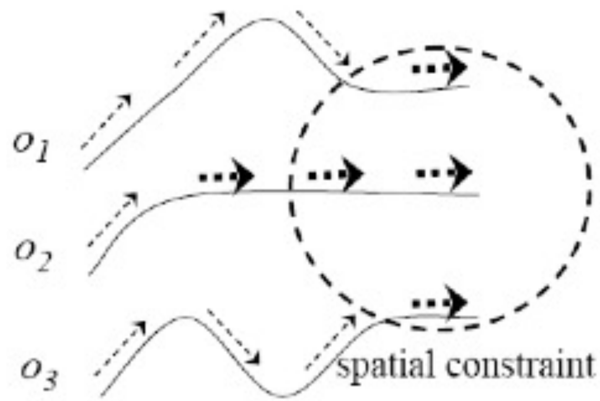
(a) raw trajectories (b) motion-azimuth trajectories

Relative Motion Patterns

- *Flock* ($m > 1$, $r > 0$): At least *m* entities are within a circular region of radius ***r*** and they move in the same direction ("constance" + a spatial constraint)

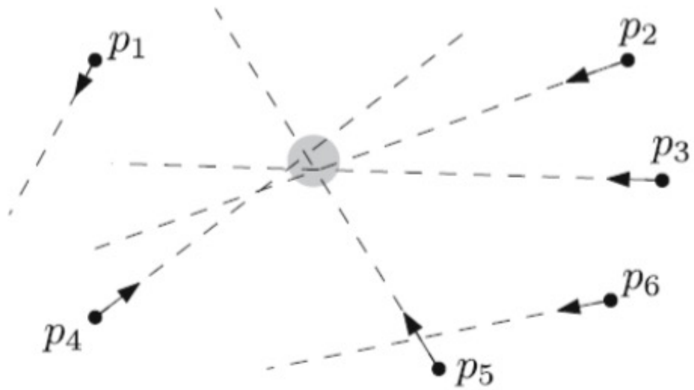


- *Leadership* ($m > 1$, $r > 0$, $s > 0$) At least *m* entities are within a circular region of radius ***r***, they move in the same direction, and at least one of the entities was already heading in this direction for at least ***s*** time steps

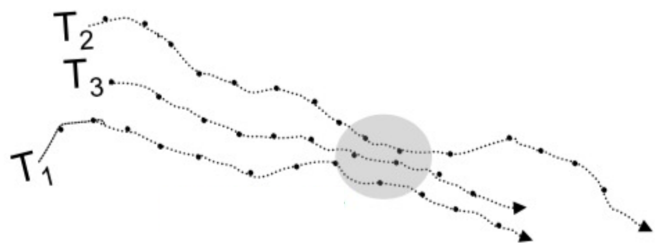


Relative Motion Patterns

- *Convergence* ($m > 1$, $r > 0$) At least ***m*** entities will pass through the same circular region of radius ***r*** (assuming they keep their direction)



- *Encounter* ($m > 1$, $r > 0$) At least ***m*** entities will be simultaneously inside the same circular region of radius ***r*** (assuming they keep their speed and direction)



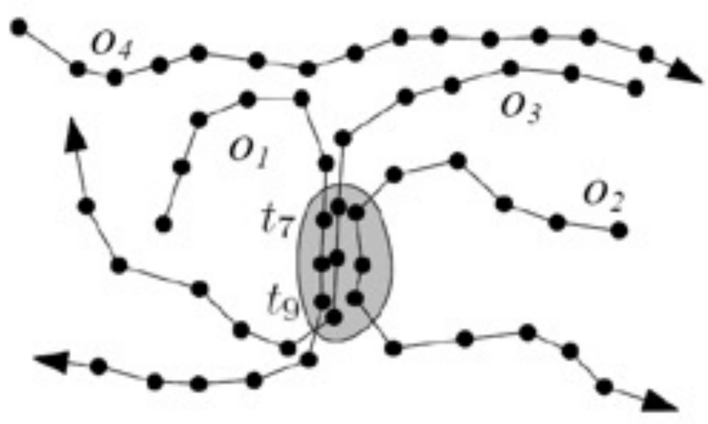
Disc-Based Trajectory Patterns

- Circular spatial constraint is considered (Disc-Based)
- Basic relative motion patterns are no longer considered
- Integration of time constraints in pattern

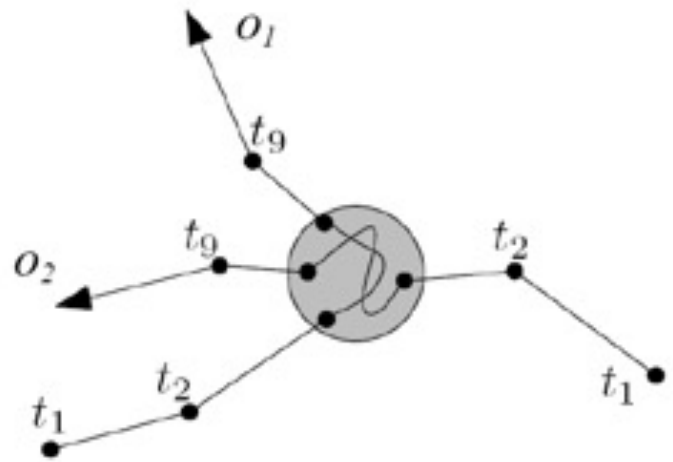
Gudmundsson et al. GIS'06,
Benkert et al. SAC'07
Vieira, M. Et al. SIGSPATIAL'09

Disc-Based Trajectory Patterns

- *Flock* (m,k,r): a group of at least m objects that move together for at least k consecutive time points, while staying within a disc with radius r

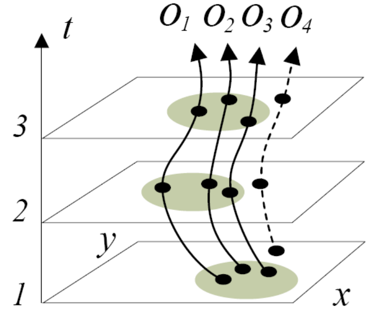


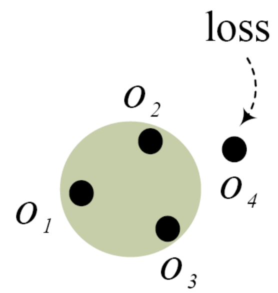
- Meet (m,k,r): a group of at least m objects that stay together in a stationary disc with radius r for at least k consecutive time points



Density-Based Trajectory Patterns

- The selection of a proper disc size *r* is difficult
  - A large *r* may capture objects that are intuitively not in the same group
  - A small *r* may miss some objects that are intuitively in the same group
- *lossy-flock* problem

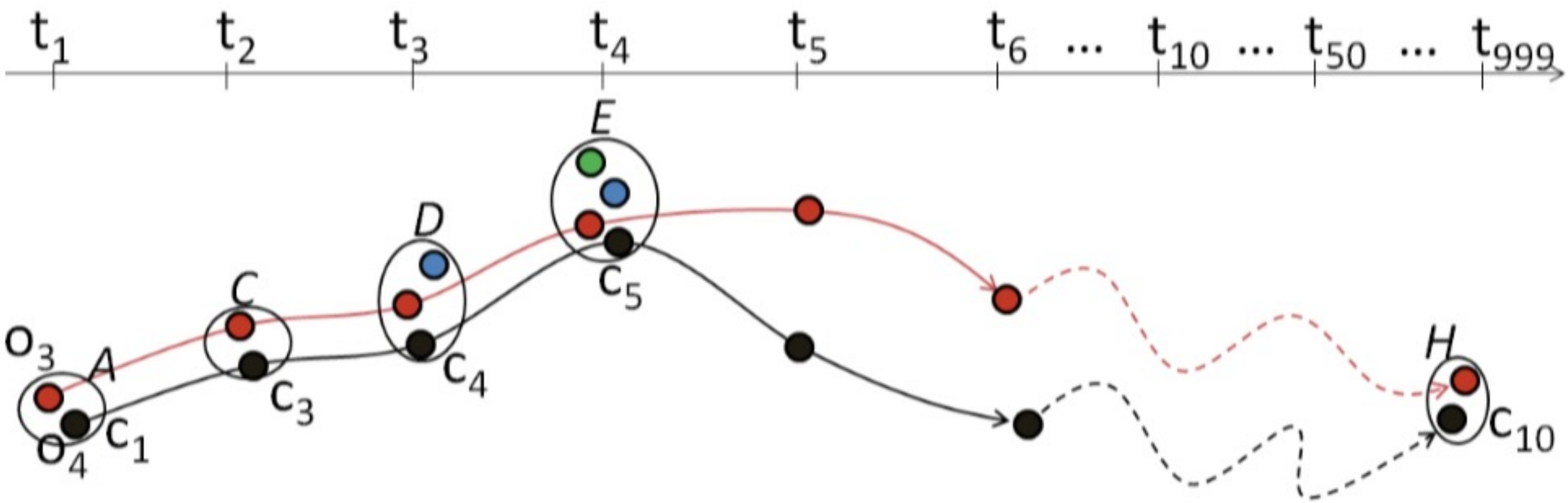




- Employ density concepts

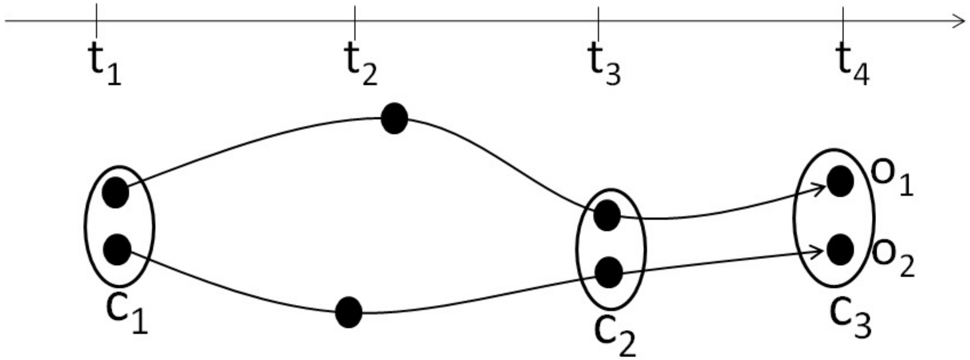
Swarm

- Swarm : At least Ɛ objects move together (density-connected objects) during min_t timestamps



Li et al. VLDB'10

Swarm – Closed Swarm



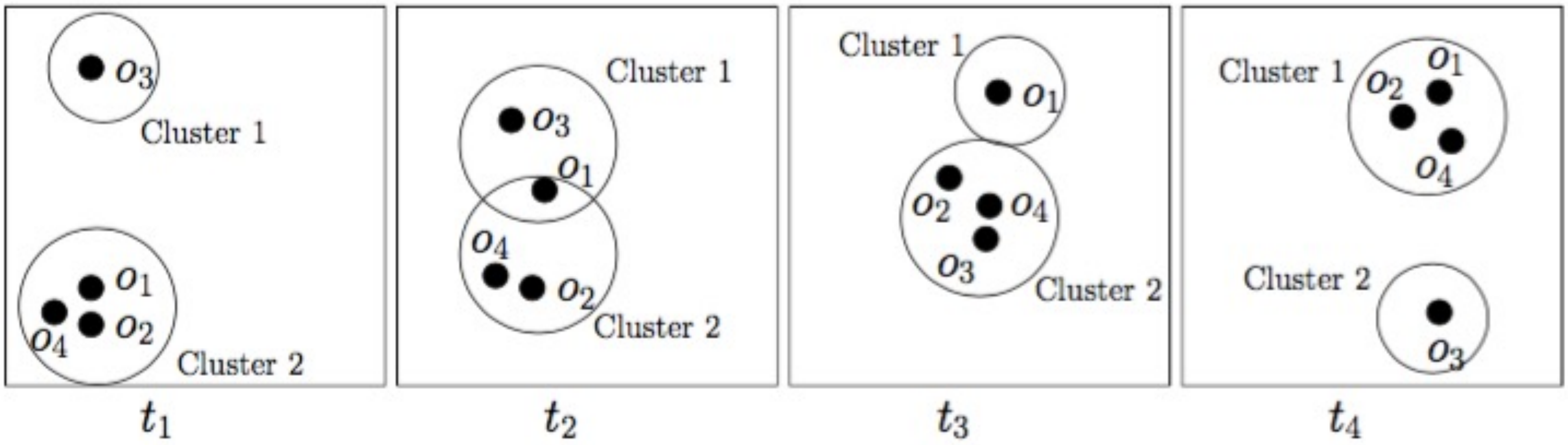
$$(\{o_1, o_2\}, \{t_1, t_3, t_4\})$$

DEFINITION 1. *Swarm [6]. A pair (O, T) is a swarm if:*

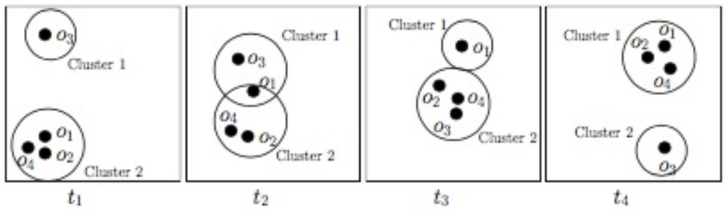
$$\begin{cases} (1) : \forall t_{a_i} \in T, \exists c \ s.t. \ O \subseteq c, \ c \text{ is a cluster.} \\ \textit{There is at least one cluster containing} \\ \textit{all the objects in } O \textit{ at each timestamp in } T. \\ (2) : |O| \geq \varepsilon. \\ \textit{There must be at least } \varepsilon \textit{ objects.} \\ (3) : |T| \geq min_t. \\ \textit{There must be at least } min_t \textit{ timestamps.} \end{cases}$$

An illustrative Example

- How many Swarms?



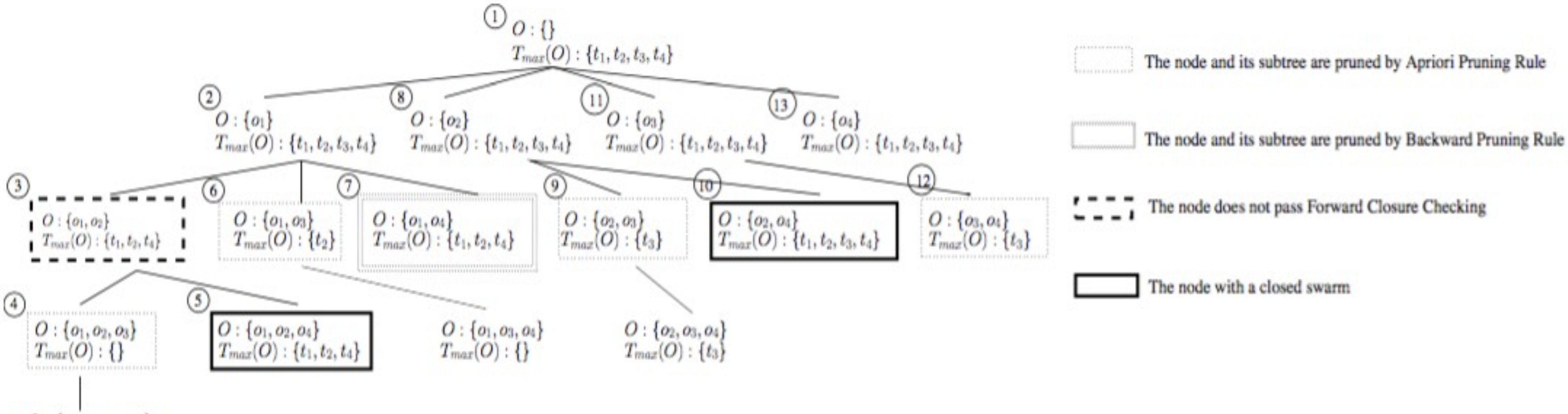
The Algorithm



2 closed swarms: *({o_2,o_4},{t_1,t_2,t_3,t_4})*
({o_1,o_2,o_4},{t_1,t_2,t_4})

A Depth First Search:
Get the maximal Timeset

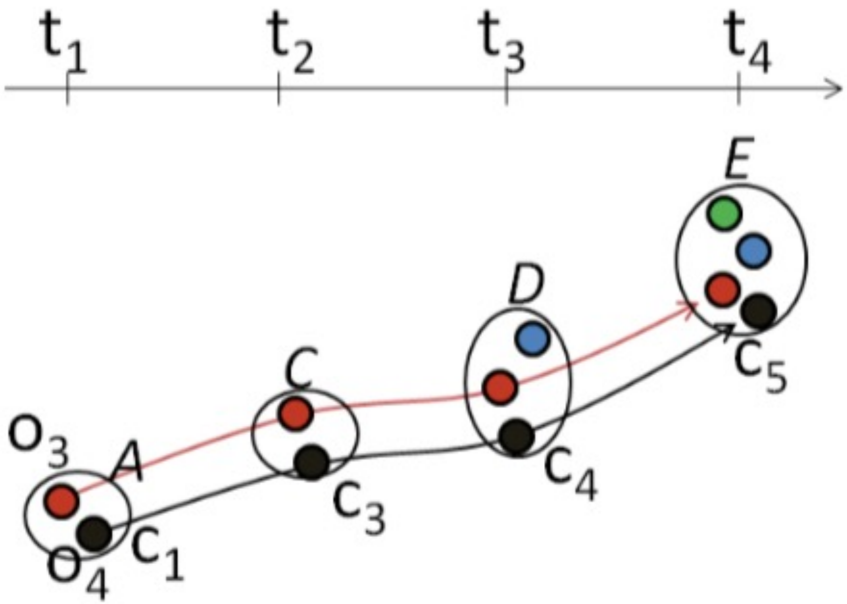
- Pruning Rules:
  - A Priori Pruning
  - Backward Pruning



ObjectGrowth Search Space $(min_o = 2, min_t = 2)$

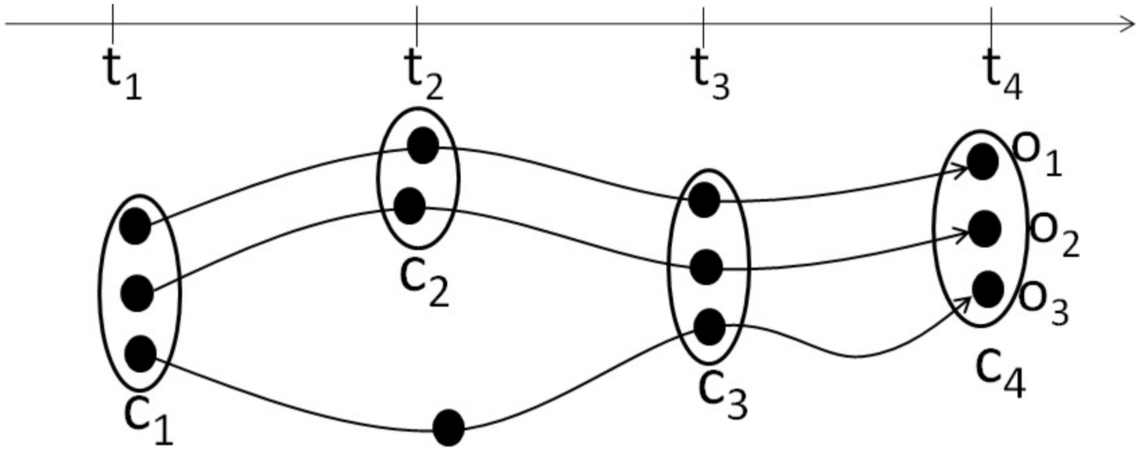
Convoy

- Convoy: At least Ɛ objects move together (density-connected objects) during min_t consecutive timestamps



Jeung et al. ICDE’08 & VLDB’08

Convoy



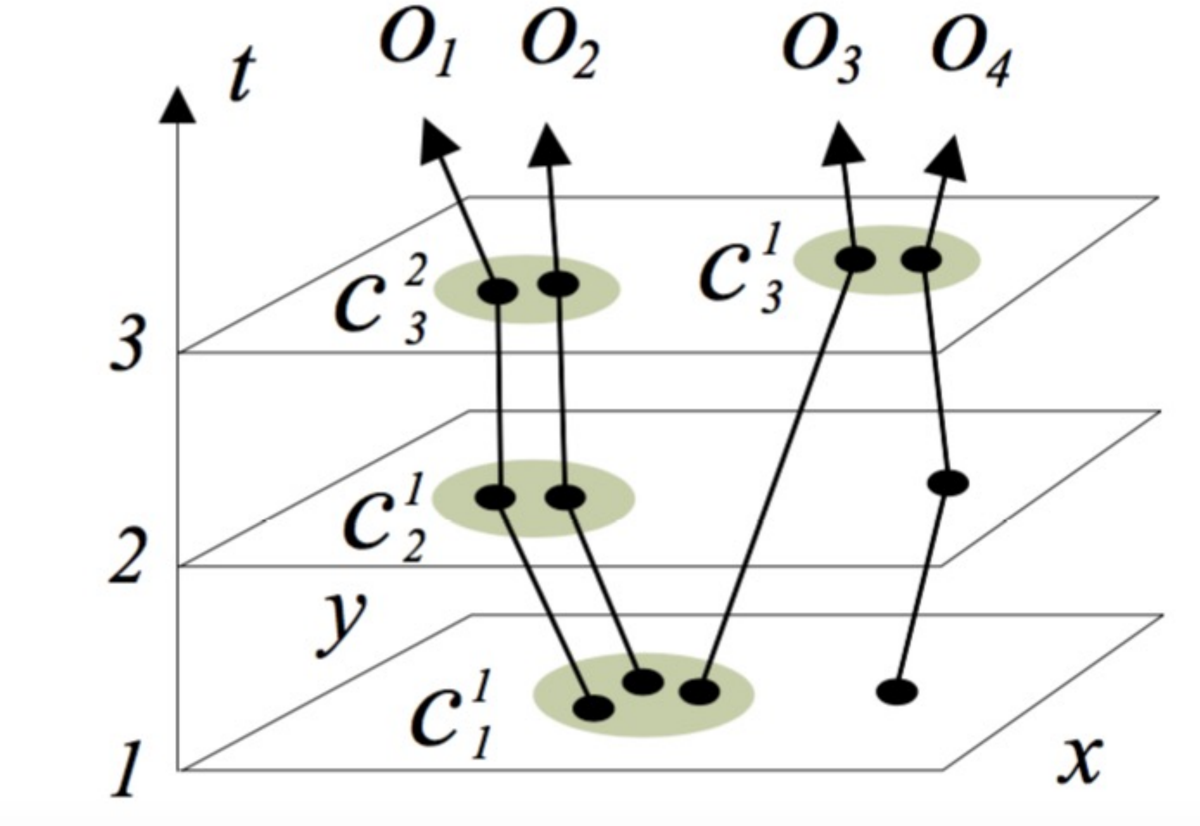
$(\{o_1, o_2\}, \{t_1, t_2, t_3, t_4\})$

Definition 3. *Convoy [3]. A pair* (O, T)*, is a convoy if:*

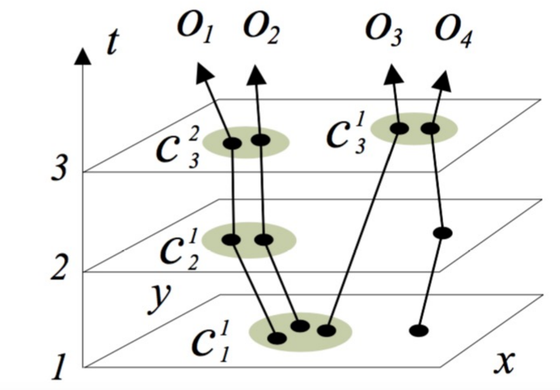
$$\begin{cases} (1) : (O,T) \ \textit{is a swarm.} \\ (2) : \forall i, 1 \leq i < |T|, \ t_{a_i}, t_{a_{i+1}} \ \textit{are consecutive.} \end{cases}$$

An illustrative Example

- How many Convoys?



The Algorithm



Applying DBSCAN at each timestamp

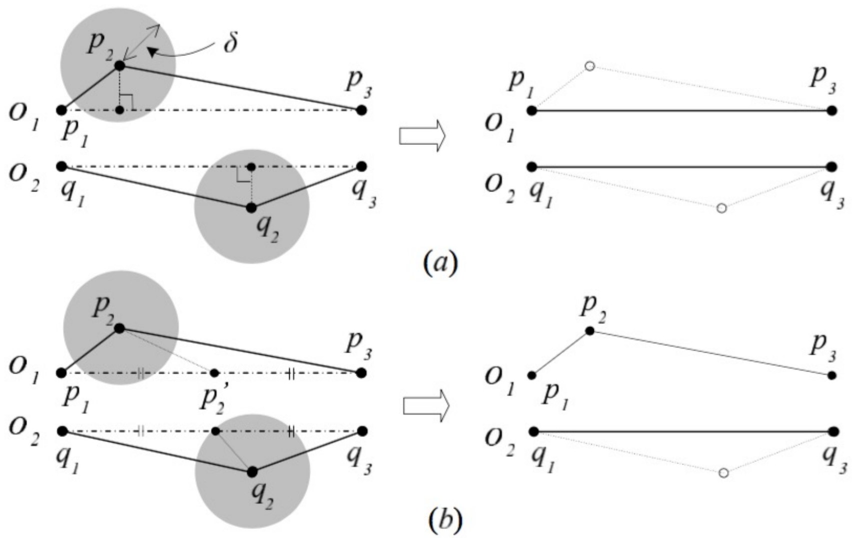
1 convoy: *$(\{o_2,o_3\},\{t_1,t_3\})$*

Timestamp	Clusters	Candidate set V
t_1	c_1^1	$v_1 = c_1^1$
t_2	c_2^1	$v_1 = c_1^1 \cap c_2^1$
t_3	c_3^1, c_3^2	$v_1 = c_1^1 \cap c_2^1 \cap c_3^2, v_2 = c_3^1$

$v_1=\{o_1,o_2,o_3\}$
$v_1=\{o_1,o_2\}$
$v_1=\{o_1,o_2\}$

Algorithm Improvements

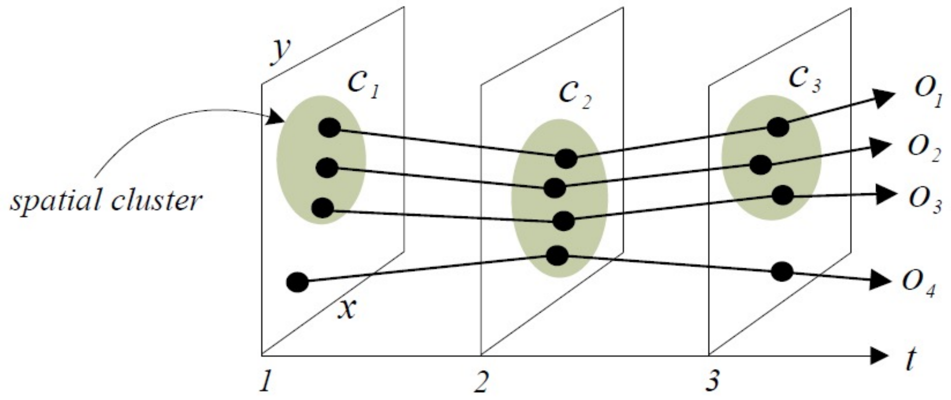
- Reducing the search space:
  - Simplifying the original trajectories



- Apply DBSCAN at each timestamp: CUTS, CUTS+

Moving Clusters

- Moving Cluster : At least Ɛ common objects move together (density-connected objects) during min_t consecutive timestamps



$$\frac{|c_t \cap c_{t+1}|}{|c_t \cup c_{t+1}|} \geq \theta$$

- Convoys + additional condition that they need to share some objects between two consecutive timestamps

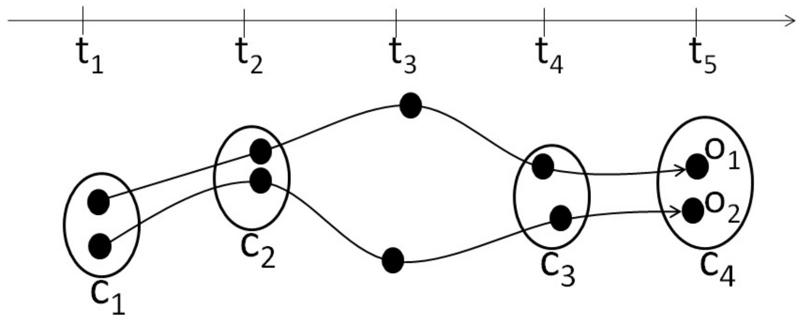
Kalnis, SSTD 2005

Group Pattern

- Combination of both convoy and closed swarm
  - A closed swarm of disjointed convoys

$T_S = \{(\{o_1, o_2\}, \{t_1, t_2\}), (\{o_1, o_2\}, \{t_4, t_5\})\}$

$(\{o_1, o_2\}, T_S)$ is a closed swarm of convoys



Definition 4. *Group Pattern*[21]. Given a set of objects O, a minimum weight threshold min_{wei}, a set of disjointed convoys $T_S = \{s_1, s_2, \ldots, s_n\}$, a minimum number of convoys min_c. (O, T_S) is a group pattern if:

$$\begin{cases} (1) : (O, T_S) \; \textit{is a closed swarm w.r.t } min_c. \\ (2) : \frac{\sum_{i=1}^{|T_S|} |s_i|}{|T_{DB}|} \geq min_{wei}. \end{cases}$$

Note that min_c is only applied for T_S (i.e. $|T_S| \geq min_c$).

Wang, DKE 2006

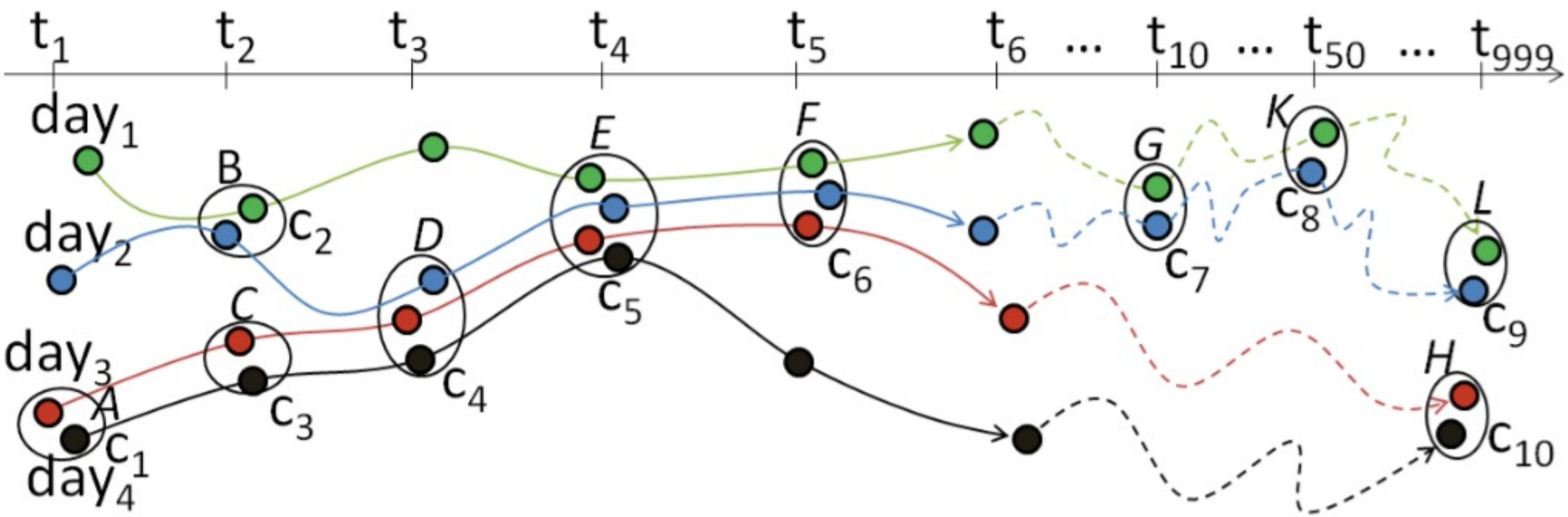
Periodic Patterns

- An object follows the same routes (approximately) over regular time intervals
  - People wake up at the same time and generally follow the same route to their work everyday
- Given an object's trajectory including N time-points, T_P which is the number of timestamps that a pattern may re-appear.
- The object's trajectory is decomposed into N/ T_P sub-trajectories
- T_P is data dependent
  - T_P = 'a day' in traffic control applications since many vehicles have daily patterns
  - T_P = 'a year' for annual animal migration patterns

N. Mamoulis, KDD04

Periodic Patterns

- A periodic pattern is a closed swarm discovery from N/ TP sub-trajectories.



4 daily sub-trajectories

We extract potential periodic patterns such as {c1,c3,c4,c5}, {c2,c5,c6}

Outline

- A quick reminder on pattern discovery
- The different kinds of patterns from trajectory
- **Towards a unified approach for extracting trajectories**
  - **Cluster Matrix**
  - **Properties**
  - **GetMove**
  - **How to Manage Blocks**
  - **Lord of the Rings**
  - **Gradual Spatio-Temporal Rules**
  - **The top-k informative Patterns**
- Deal with and without a spatial component?
- An illustration
- Conclusion

Motivations

- Clustering:
  - Flock, convoy, moving cluster, swarm, closed swarm, k-Stars, ...
- Different approaches to mine them:
  - CuTS*, ObjectGrowth, CMC, Vg-Growth, ...
  - Complexity, computation cost, time consuming!!!
- *How about unifying algorithm?*
- What happen when new data arriving?
  - *How about incremental algorithm?*

The main intuition

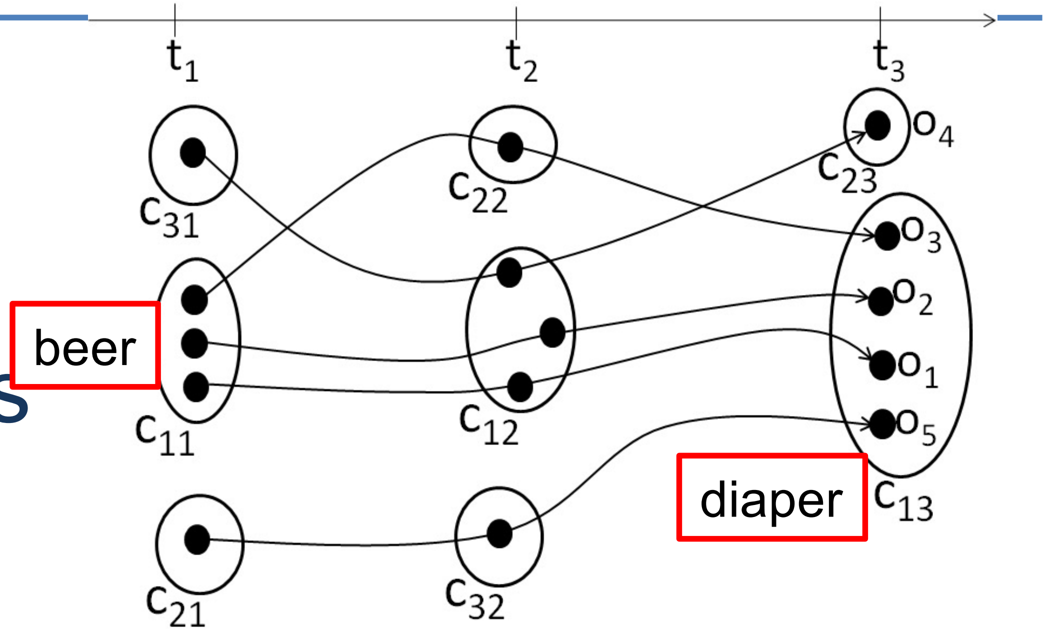
- Very efficient approaches have been defined for extracting a set of association rules
  – Apriori,Closed, Derivable, …

	A	B	C	D	E	F
1	1	0	0	1	0	0
2	1	0	1	0	0	0
3	1	1	1	0	0	0
4	1	1	0	0	1	1

- *Is it possible to adapt a same kind of approach for extracting trajectories?*

Cluster Matrix

- Patterns:
  - Evolution of clusters
- Objects: transactions
- Clusters: items



Cluster Matrix

T_{DB}		t_1			t_2			t_3	
Clusters C_{DB}		beer	c_{21}	c_{31}	c_{12}	c_{22}	c_{32}	diaper	c_{23}
O_{DB}	o_1	1			1			1	
	o_2	1			1			1	
	o_3	1				1		1	
	o_4			1	1				1
	o_5		1				1	1	

Frequent Closed Itemset from Cluster Matrix

Frequent Closed Itemset $\Upsilon = \{c_{t_{a_1}}, c_{t_{a_2}}, \ldots, c_{t_{a_p}}\}$

Life Time $T_\Upsilon = \{t_{a_1}, t_{a_2}, \ldots, t_{a_p}\}$

Set of Objects $O(\Upsilon) = \bigcap_{i=1}^{p} c_{t_{a_i}}$ Support $\sigma(\Upsilon)$

$$\Upsilon = \{c_{11}, c_{13}\} \; (O(\Upsilon) = \{o_1, o_2, o_3\}, T_\Upsilon = \{t_1, t_3\})$$

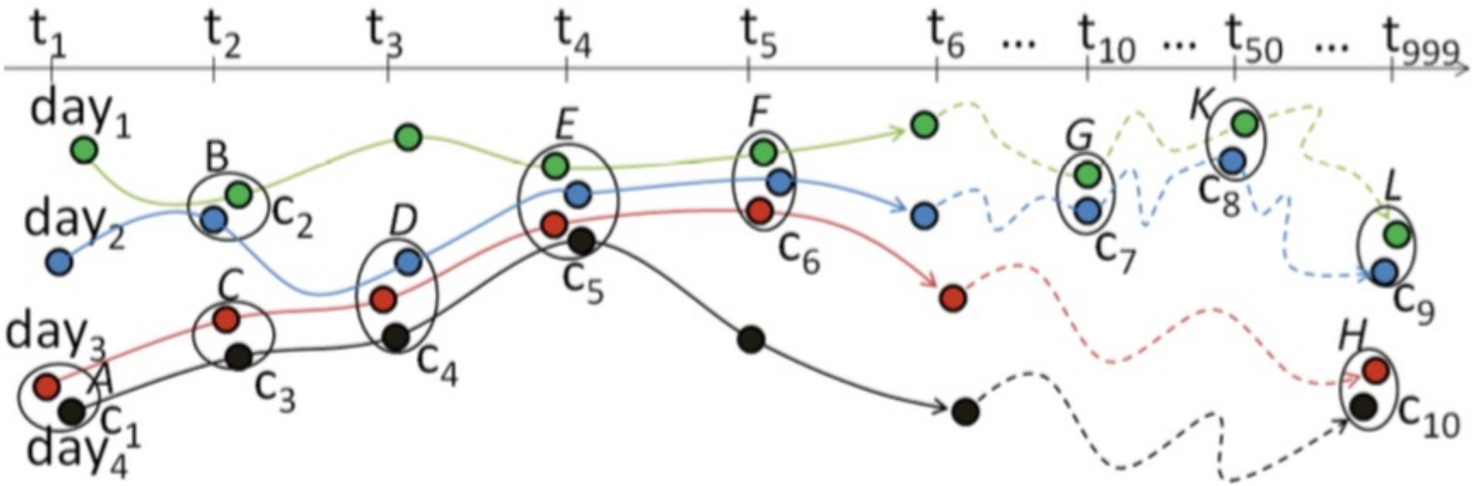
$$\sigma(\Upsilon) = 3$$

$$|\Upsilon| = 2$$

Cluster Matrix

T_{DB}		t_1			t_2			t_3	
Clusters C_{DB}		c_{11}	c_{21}	c_{31}	c_{12}	c_{22}	c_{32}	c_{13}	c_{23}
O_{DB}	o_1	1			1			1	
	o_2	1			1			1	
	o_3	1				1		1	
	o_4			1	1				1
	o_5		1				1	1	

Periodic Cluster Matrix



T_{db}		t_1	t_2		t_3	t_4	t_5	t_{10}	t_{50}	t_{999}	
Clusters C_{db}		c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}
ST_{db}	st_1		1			1	1	1	1	1	
	st_2		1		1	1	1	1	1	1	
	st_3	1		1	1	1	1				1
	st_4	1		1	1	1					1

The main intuition (following…)

- We are now able to extract itemsets corresponding to a set of clusters occurring over time
- Not trajectories yet!
- *What about properties on Itemsets?*

Property for Swarm

DEFINITION 1. *Swarm [6]. A pair (O, T) is a swarm if:*

$$\begin{cases} (1): \forall t_{a_i} \in T, \exists c \ s.t. \ O \subseteq c, \ \ c \ is \ a \ cluster. \\ There \ is \ at \ least \ one \ cluster \ containing \\ all \ the \ objects \ in \ O \ at \ each \ timestamp \ in \ T. \\ (2): |O| \geq \varepsilon. \\ There \ must \ be \ at \ least \ \varepsilon \ objects. \\ (3): |T| \geq min_t. \\ There \ must \ be \ at \ least \ min_t \ timestamps. \end{cases}$$

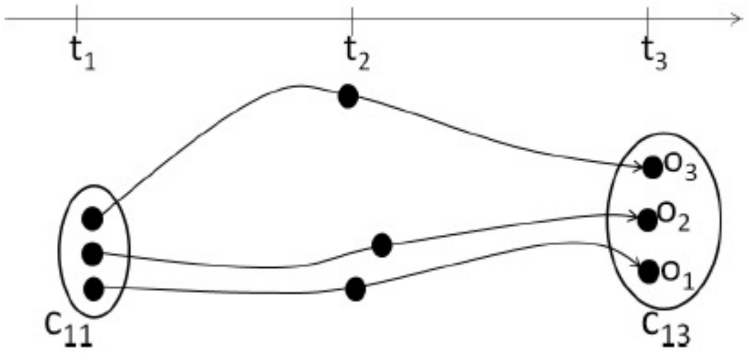
PROPERTY 1. *Swarm. Given a frequent itemset $\Upsilon = \{c_{t_{a_1}}, c_{t_{a_2}}, \ldots, c_{t_{a_p}}\}$. $(O(\Upsilon), T_\Upsilon)$ is a swarm if, and only if:*

$$\begin{cases} (1): \sigma(\Upsilon) \geq \varepsilon \\ (2): |\Upsilon| \geq min_t \end{cases}$$

Swarm

Property 1. *Swarm.* *Given a frequent itemset* $\Upsilon = \{c_{t_{a_1}}, c_{t_{a_2}}, \ldots, c_{t_{a_p}}\}$. $(O(\Upsilon), T_\Upsilon)$ *is a swarm if, and only if:*

$$\begin{cases} (1) : \sigma(\Upsilon) \geq \varepsilon \\ (2) : |\Upsilon| \geq min_t \end{cases}$$



Cluster Matrix

T_{DB}		t_1			t_2			t_3	
Clusters C_{DB}		c_{11}	c_{21}	c_{31}	c_{12}	c_{22}	c_{32}	c_{13}	c_{23}
O_{DB}	o_1	1			1			1	
	o_2	1			1			1	
	o_3	1				1		1	
	o_4			1	1				1
	o_5		1				1	1	

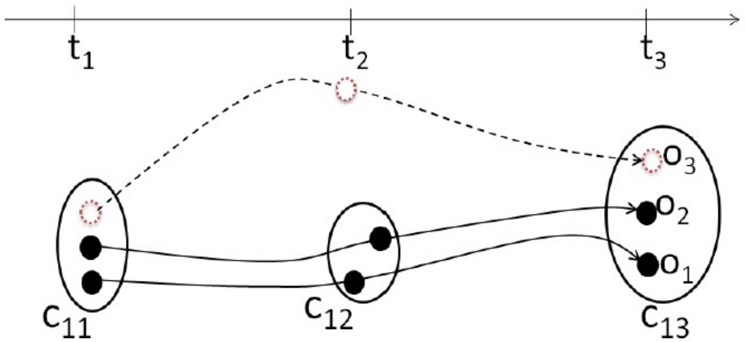
$$\Upsilon = \{c_{11}, c_{13}\} \, (O(\Upsilon) = \{o_1, o_2, o_3\}, T_\Upsilon = \{t_1, t_3\})$$

$$\sigma(\Upsilon) = 3 > \varepsilon \text{ and } |\Upsilon| = 2 \geq min_t$$

Convoy

PROPERTY 3. *Convoy. Given a frequent itemset* $\Upsilon = \{c_{t_{a_1}}, c_{t_{a_2}}, \ldots, c_{t_{a_p}}\}$. $(O(\Upsilon), T_\Upsilon)$ *is a convoy if and only if:*

$$\begin{cases} (1) : (O(\Upsilon), T_\Upsilon) \text{ is a swarm.} \\ (2) : \forall j, 1 \leq j < p : t_{a_j}, t_{a_{j+1}} \text{ are consecutive.} \end{cases}$$



Cluster Matrix

T_{DB}		t_1			t_2			t_3	
Clusters C_{DB}		c_{11}	c_{21}	c_{31}	c_{12}	c_{22}	c_{32}	c_{13}	c_{23}
O_{DB}	o_1	1			1			1	
	o_2	1			1			1	
	o_3	1				1		1	
	o_4			1	1				1
	o_5		1				1	1	

$$\Upsilon = \{c_{11}, c_{12}, c_{13}\}$$
$$(O(\Upsilon) = \{o_1, o_2\}, T_\Upsilon = \{t_1, t_2, t_3\})$$

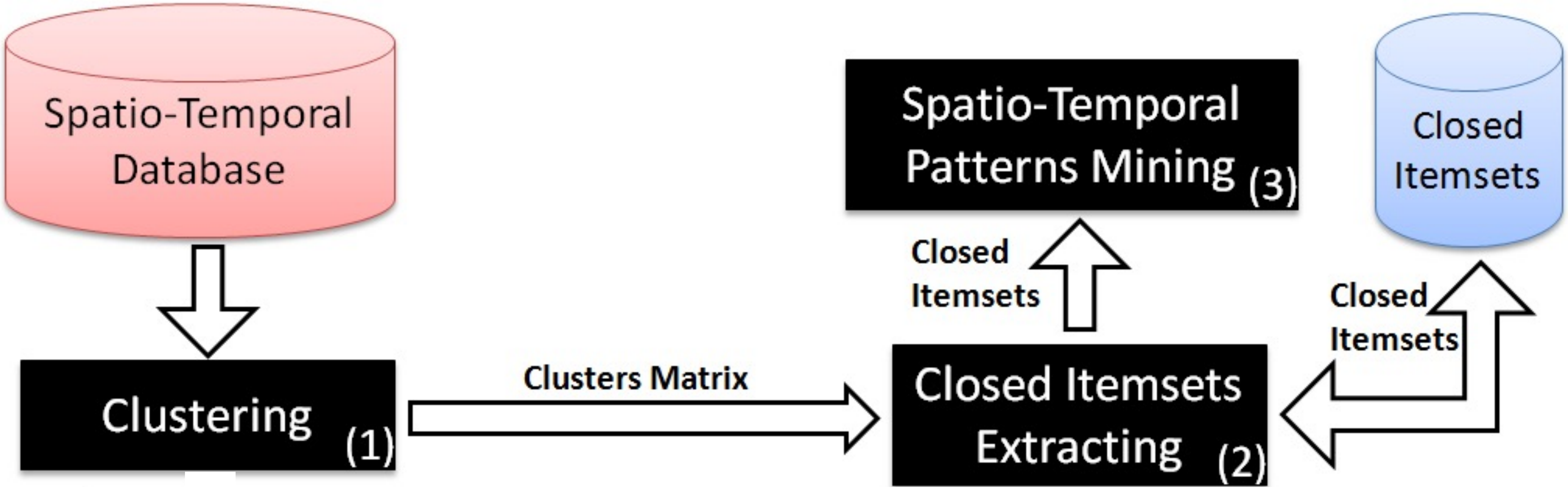
Properties

- In the same way it is possible to define properties for:

  Swarm, Closed Swarm, Convoy, Moving Cluster, Periodic Pattern, …

- We are now able to extract trajectories!

The Main Process



GeT_Move – Pruning Rule

- Spatio-temporal patterns:
  - Clusters (items) must belong to different timestamps
  - Items which form FCIs must be in different timestamps
- FCIs include more than 1 item in the same timestamp will be discarded

Cluster Matrix

T_{DB}		t_1			t_2			t_3	
Clusters C_{DB}		c_{11}	c_{21}	c_{31}	c_{12}	c_{22}	c_{32}	c_{13}	c_{23}
O_{DB}	o_1	1			1			1	
	o_2	1			1			1	
	o_3	1				1		1	
	o_4			1	1				1
	o_5		1				1	1	

GeT_Move

- Frequent closed itemset mining algorithm to extract FCIs:

  LCM Algorithm (Linear time Closed itemset Miner)

- Extract patterns from FCIs

$$\xrightarrow[|X| \geq min_t \text{ : Closed Swarm}]{X}$$

GeT_Move

- Frequent closed itemset mining algorithm to extract FCIs:

  LCM Algorithm (Linear time Closed itemset Miner)

- Extract patterns from FCIs

$$X$$

$$Consecutiveness,\ object\ correctness,\ min_t$$

Convoy

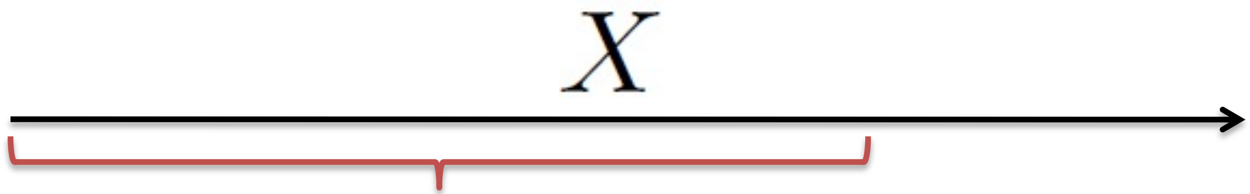
Set of disjointed convoys is a Group Pattern

GeT_Move

- Frequent closed itemset mining algorithm to extract FCIs:

  LCM Algorithm (Linear time Closed itemset Miner)

- Extract patterns from FCIs



Consecutiveness, integrity proportion

Moving Cluster

Two remaining problems

- Association rule mining algorithms have some problems when dealing with a huge number of items, i.e. long transactions!

- *What happens with new data?*

Towards an Incremental Approach

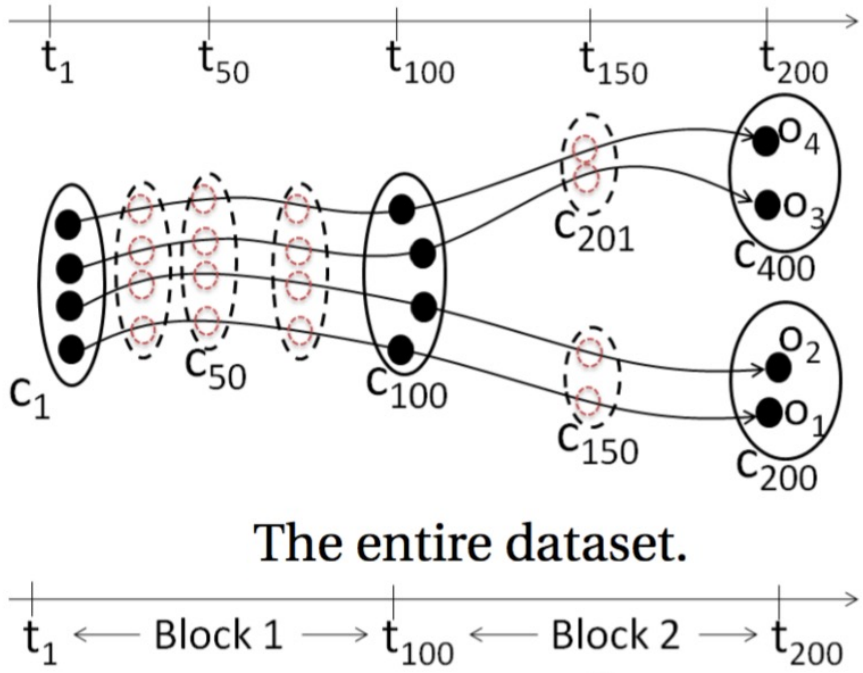
- Main idea: “Compress the dataset”
  - Shorten the transactions by splitting the trajectory matrix into short intervals called Blocks
  - Apply FCI mining on blocks (local FCI)

Maximal size of

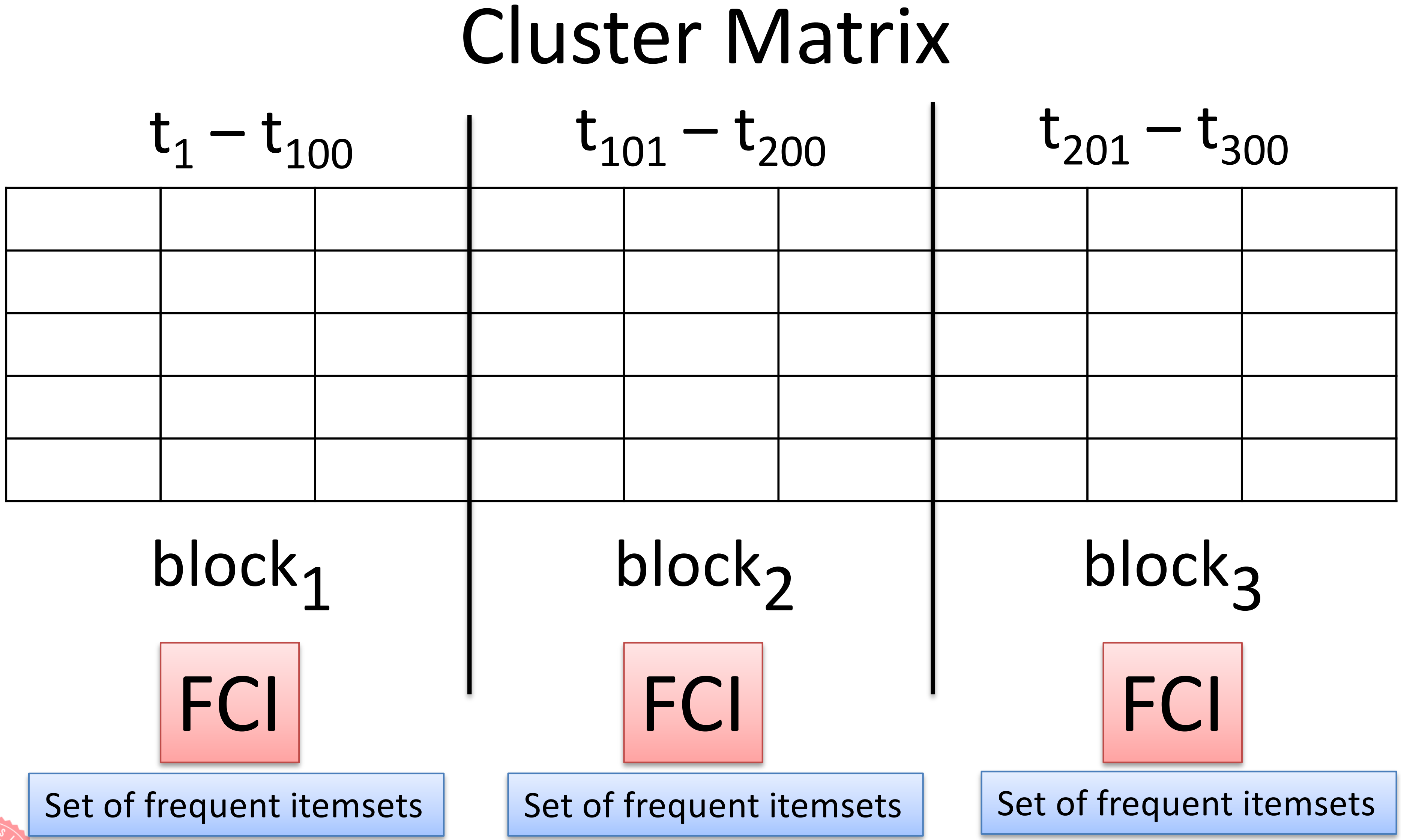
Itemset on the dataset=200

Maximal size of itemset

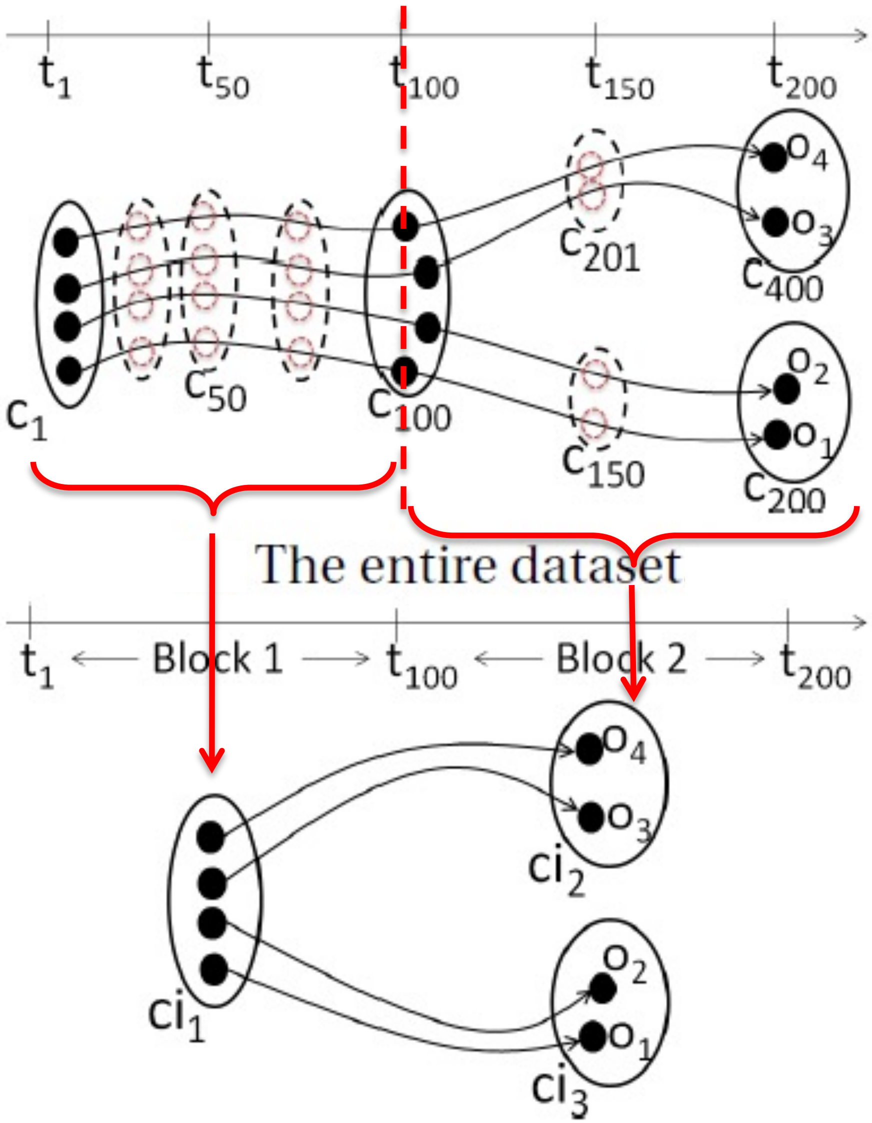
in the dataset=100



Block

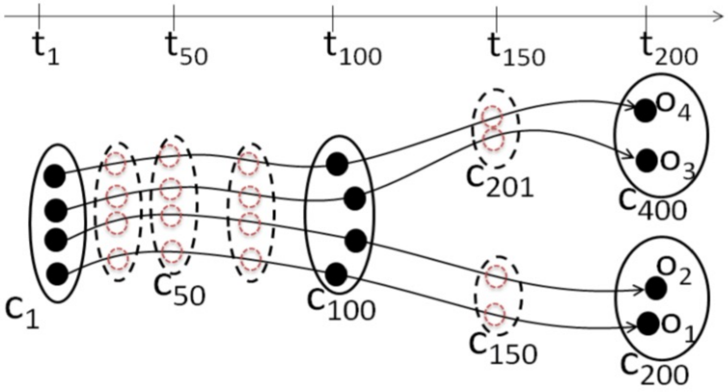

Cluster Matrix
$t_1 - t_{100}$
$t_{101} - t_{200}$
$t_{201} - t_{300}$
$block_1$
$block_2$
$block_3$
FCI
FCI
FCI
Set of frequent itemsets
Set of frequent itemsets
Set of frequent itemsets

Incremental GeT_Move

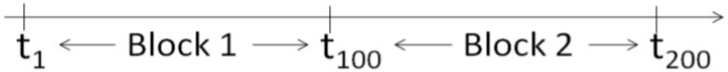


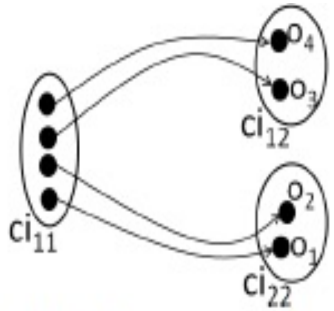
Data after applying frequent closed itemsets mining on Blocks

Closed Itemset Matrix



The entire dataset.





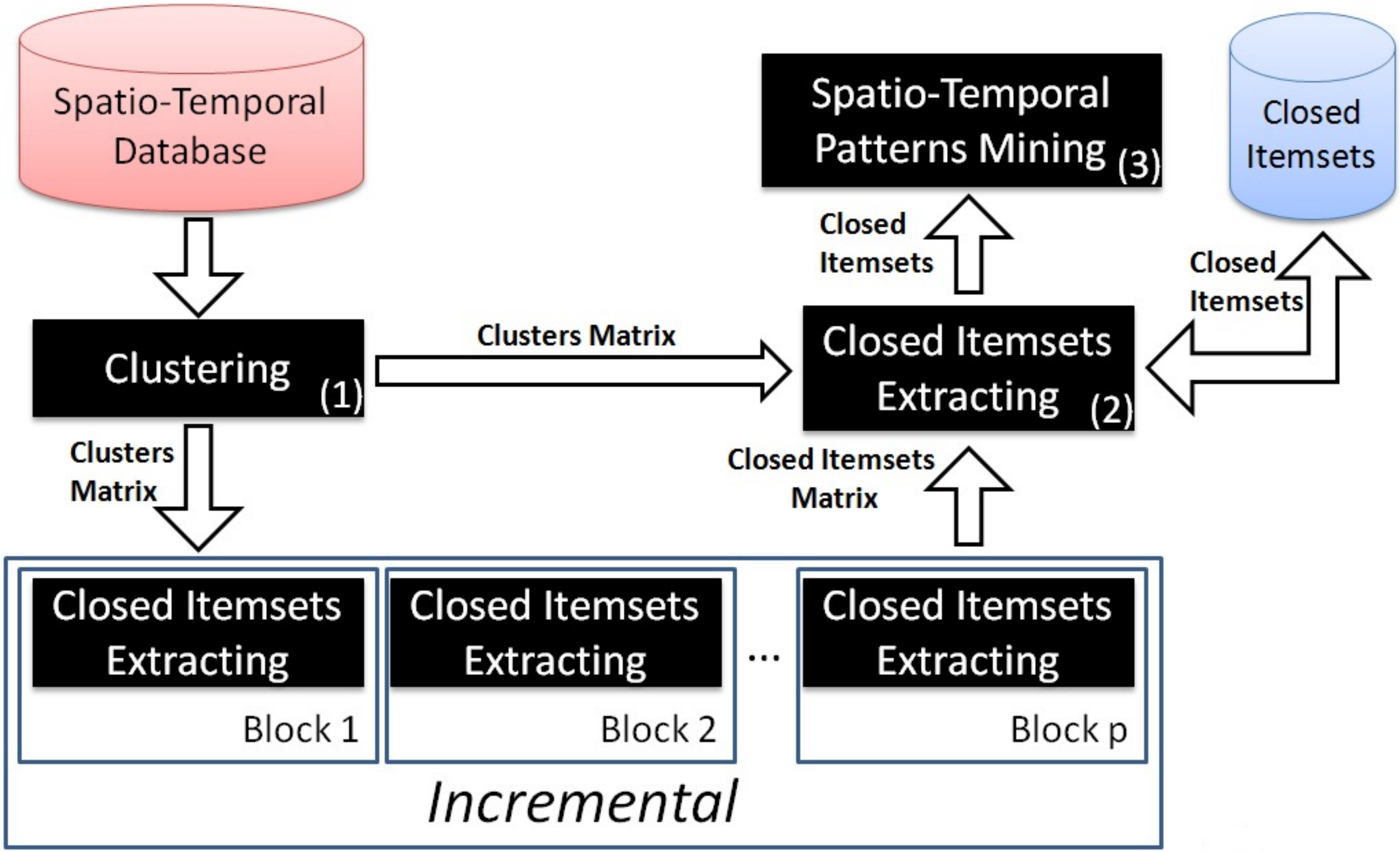
Data after applying frequent closed itemsets mining on Blocks.

Closed Itemset Matrix.

Block B		b_1	b_2	
Frequent Closed Itemsets CI		ci_1	ci_2	ci_3
O_{db}	o_1	1		1
	o_2	1		1
	o_3	1	1	
	o_4	1	1	

Definition 8. *Closed Itemset Matrix (CIM). Closed itemset matrix is a cluster matrix with some differences as follows: 1) Timestamp* t *now becomes a block* b*, 2) Item* c *is a frequent closed itemset* ci.

The Main Process



Experimental Results

- Datasets:

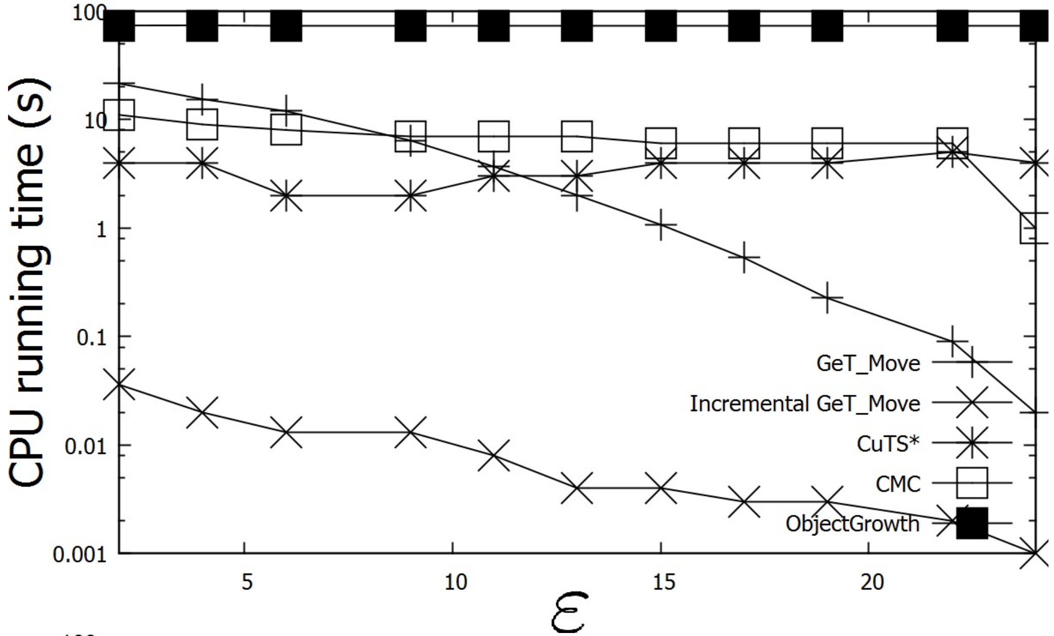
	#objects	#timestamps
Swainsoni	43	4,425
Buffalo	165	3,000
Synthetic*	500	10,000
Synthetic 2	50,000	10,000

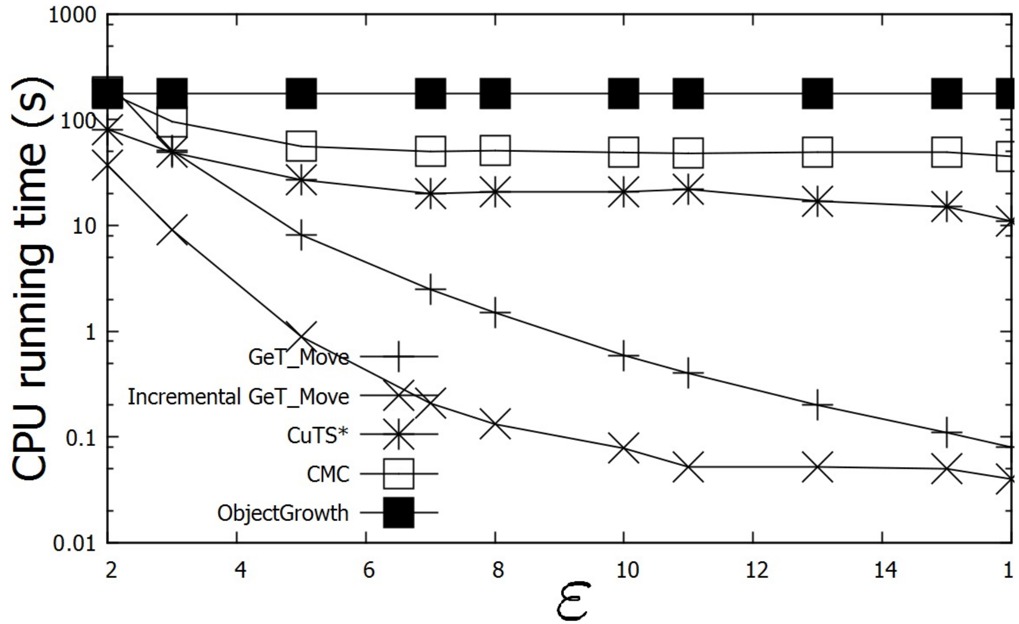
- Competitive algorithms:

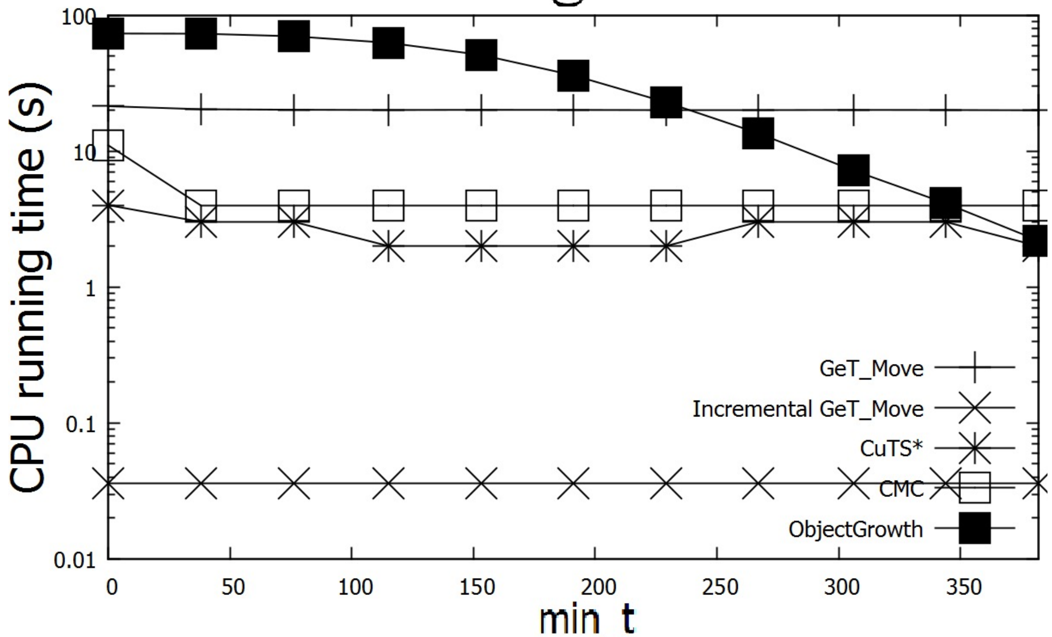
	CMC	**CuTS***	**ObjectGrowth**	**Vg-Growth**	***Incremental GeT_Move***
Convoys	X	X			X
Closed Swarms			X		X
Group Patterns				X	X

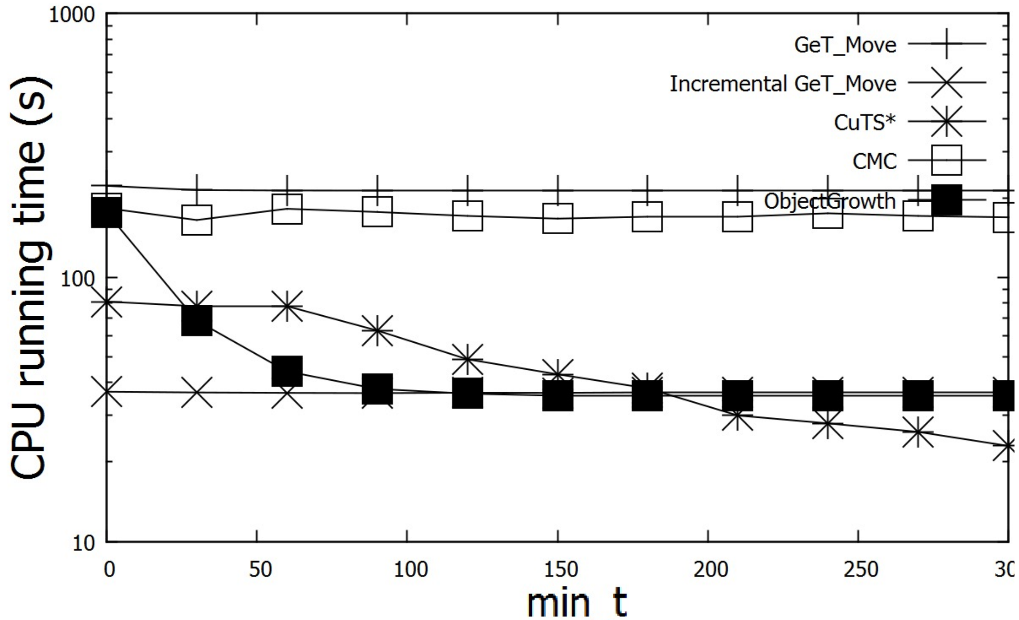
* http://iapg.jade-hs.de/personen/brinkhoff/generator/

Swainsoni - Buffalo





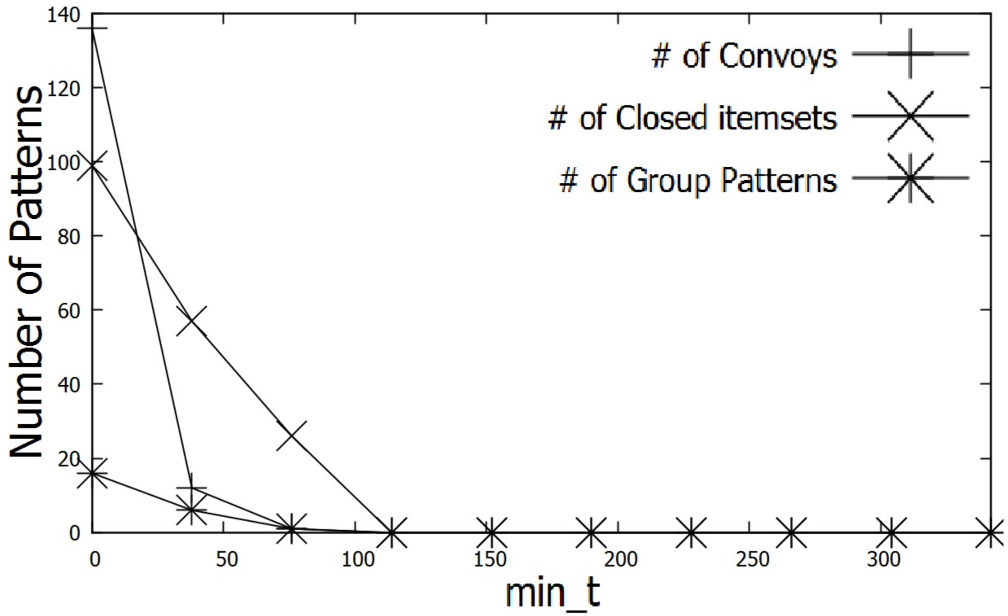




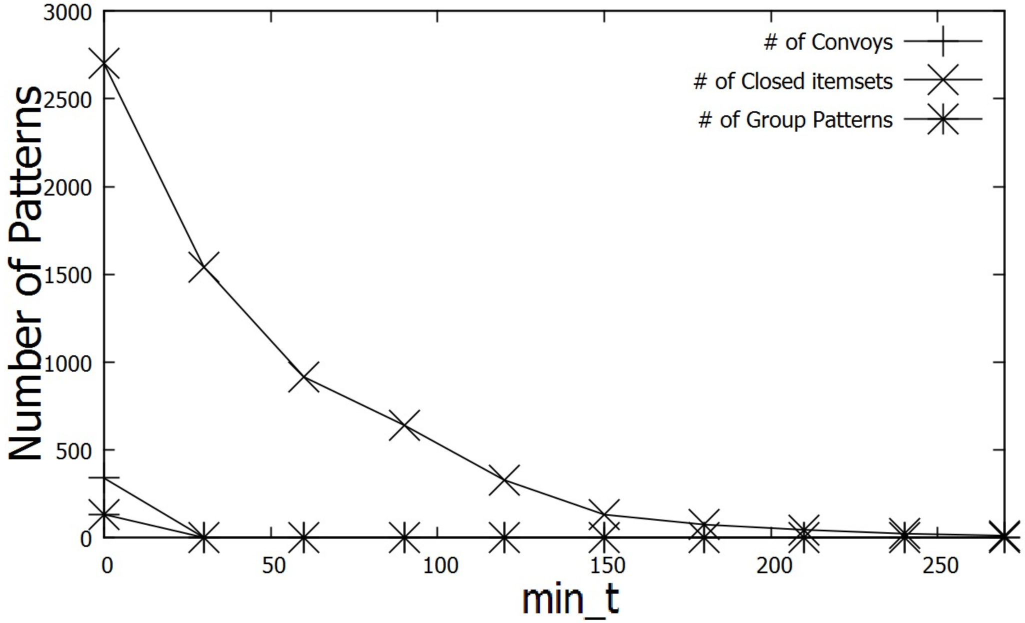
Swainsoni

Buffalo

Swainsoni - Buffalo

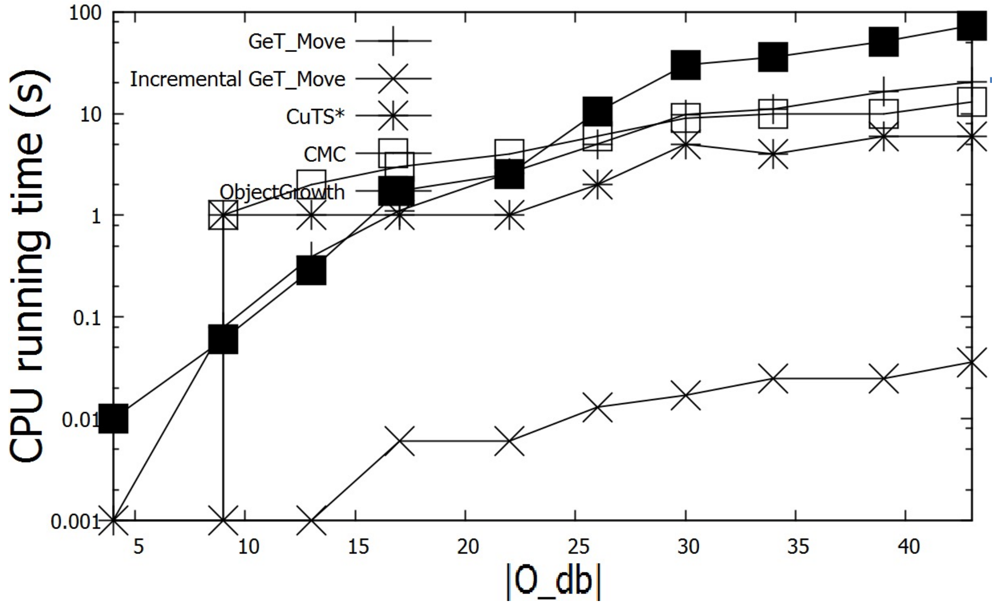


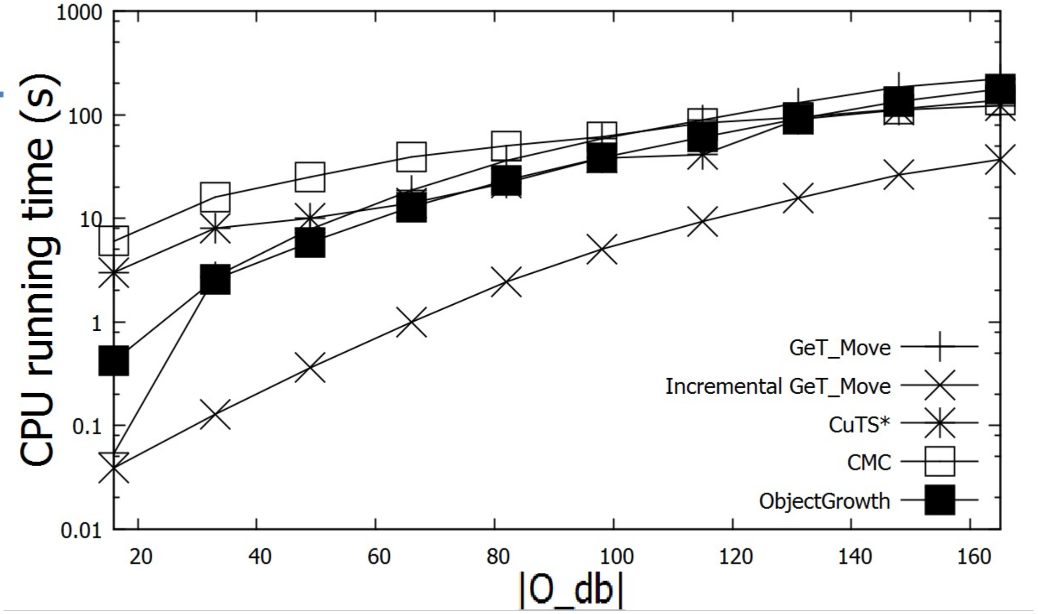
Swainsoni

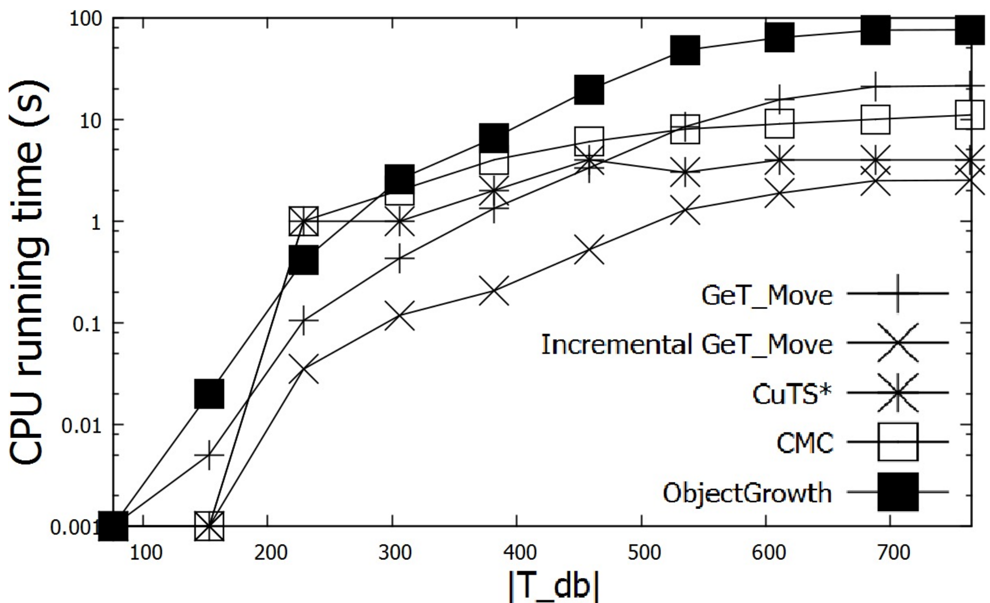


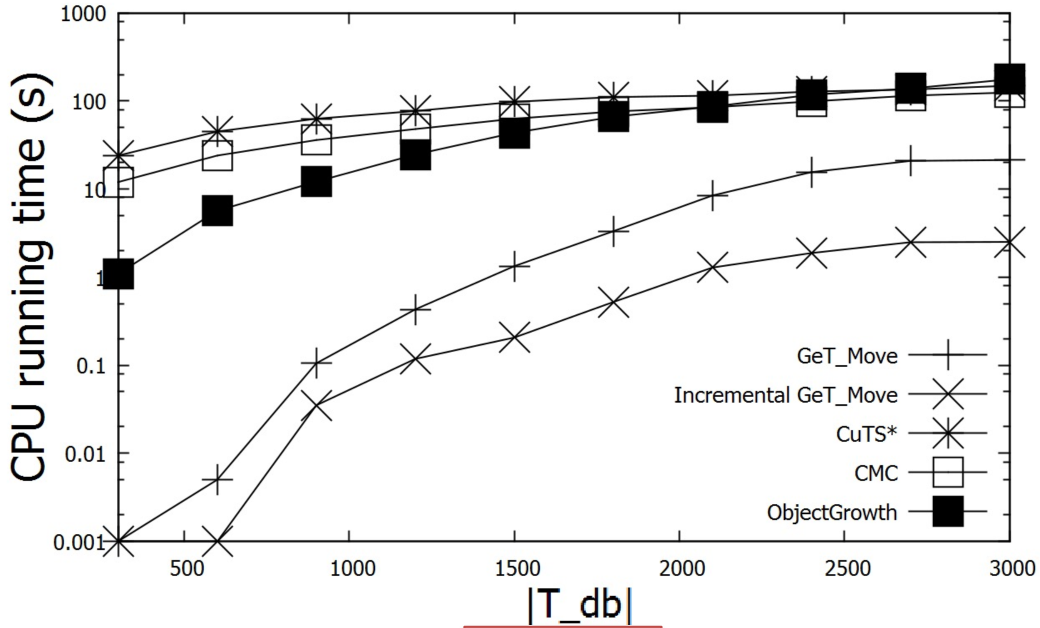
Buffalo

Swainsoni - Buffalo



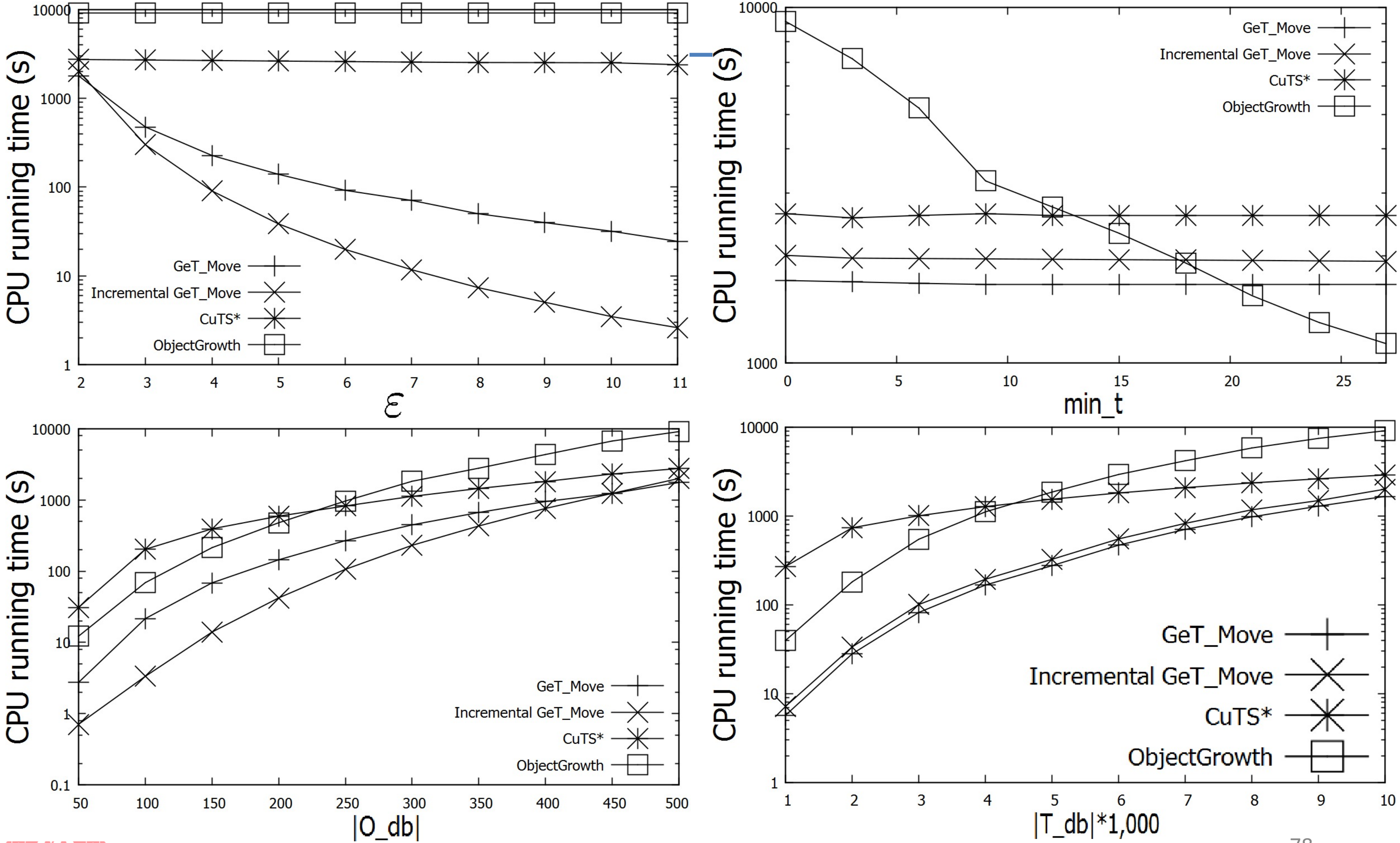






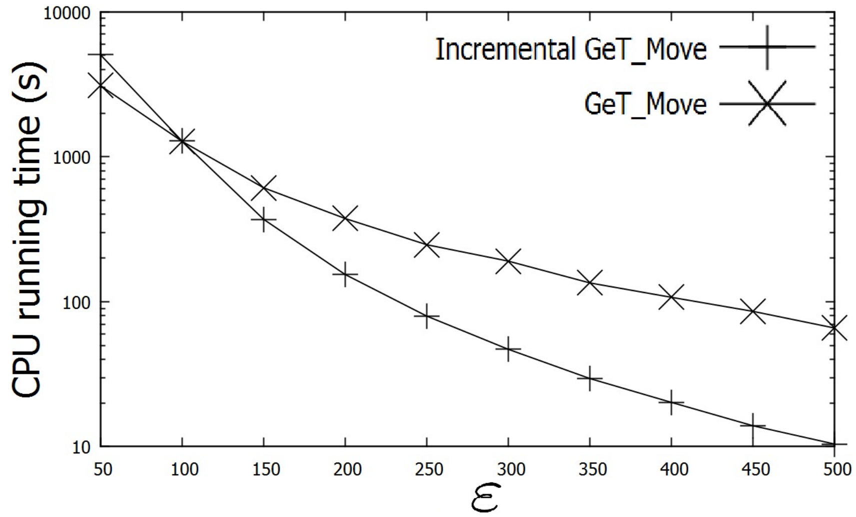
Swainsoni **Buffalo**

Synthetic Data



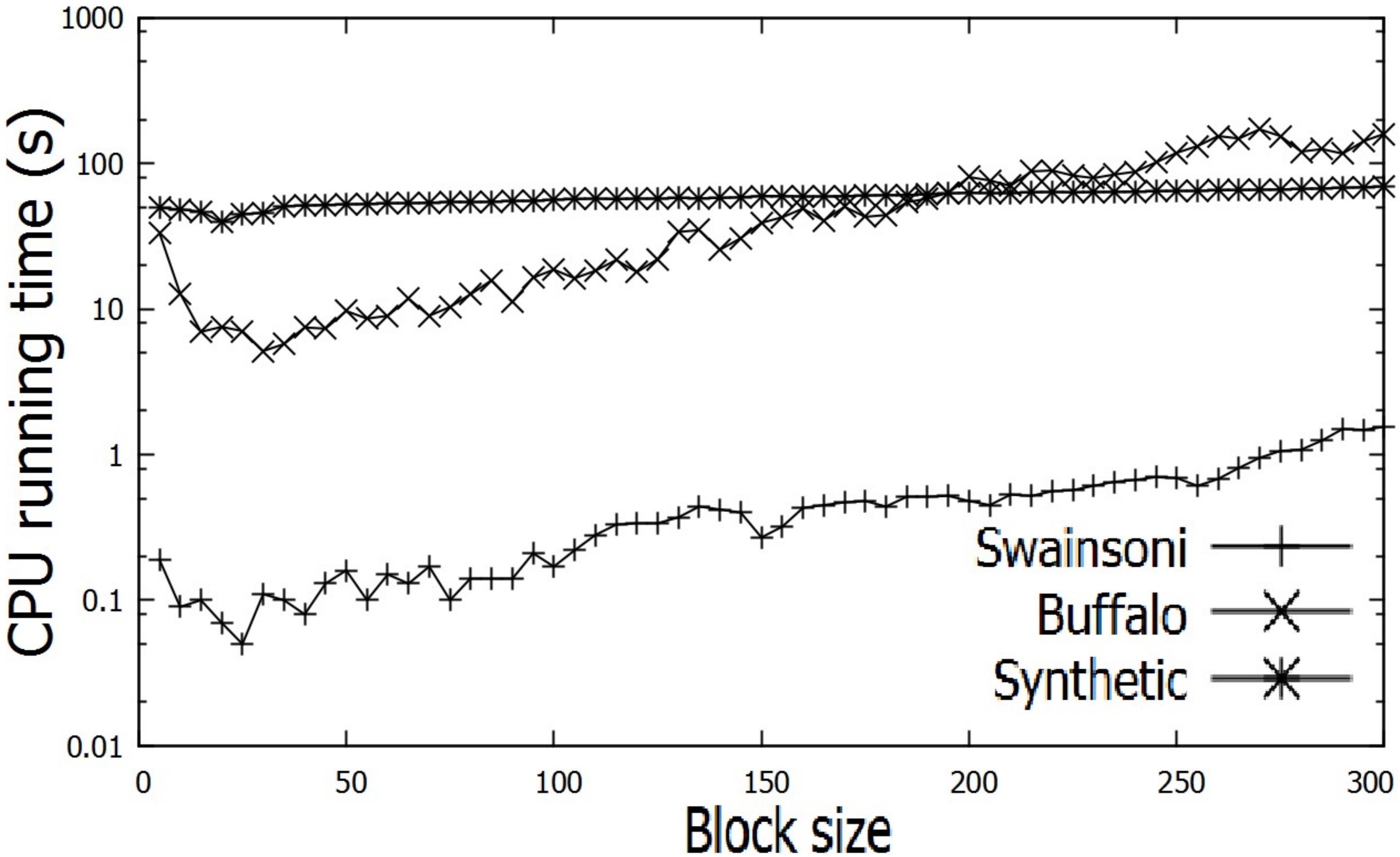
Scalability

- Synthetic data:
  - 50,000 objects in 10,000 timestamps
  - 500 million locations in total
  - CMC and CuTS* stop due to a lack of memory capacity after processing 300 million locations
  - ObjectGrowth cannot provide the results after 1day running



Block Size

- Different block sizes:
  - Range 20-30.



Effectiveness

One of discovered closed swarm

One of discovered convoy

Effectiveness

One of discovered group pattern

Towards a Parameter Free

- How to specify the optimal size of blocks?

Reordering the dataset

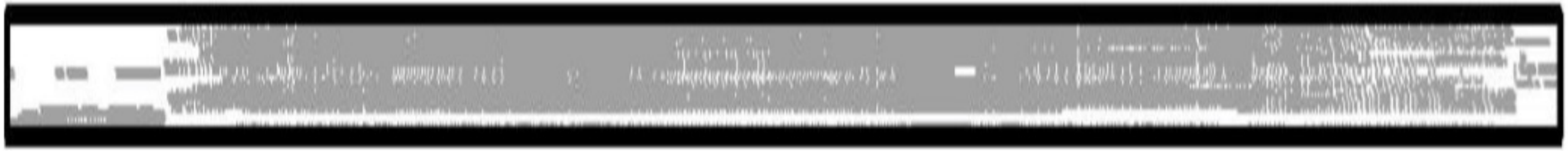
- LCM is very efficient for dense data
- Reorganizing the data

(a) (b) (c)

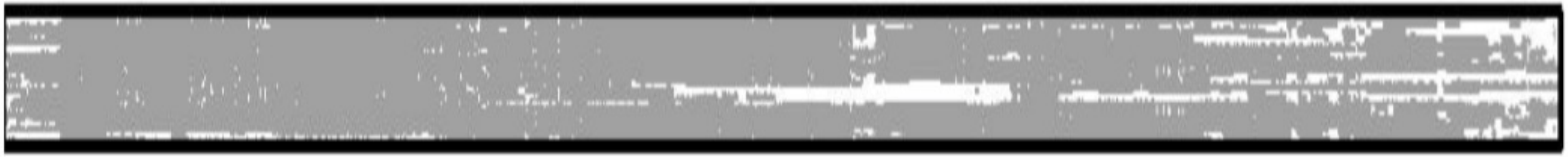
Examples of non-nested , almost nested, fully nested datasets [37]. Black = 1, white = 0. (a) Original, (b) Almost nested, (c) Fully nested.

H. Mannila, E. Terzi. *Nestedness and Segmented Nestedness*. In KDD'07

The Original Distributions



Original Swainsoni cluster matrix.



Original Buffalo cluster matrix.

Original Synthetic cluster matrix.

An Example of Nested

(a) Original Swainsoni cluster matrix.

(b) Nested Swainsoni cluster matrix.

(c) Original Buffalo cluster matrix.

(d) Nested Buffalo cluster matrix.

(e) Original Synthetic cluster matrix.

(f) Nested Synthetic cluster matrix.

Dataset	Matrix fill	#Nested blocks	avg.length
Swainsoni	17.8%	102	4.52
Buffalo	7.2%	602	2.894
Synthetic	0.1%	8	2.00

Reordering the dataset

- New definition of cluster matrix

Definition 9. Fully nested cluster matrix (resp. block). An $n \times m$ 0-1 block b is fully nested if for any two column r_i and r_{i+1} ($r_i, r_{i+1} \in b$) we have $r_i \cap r_{i+1} = r_{i+1}$.

The Algorithm

They share the same objects. It is a block.

They do not share the same objects. They are stored in a separate place that will be considered at the end

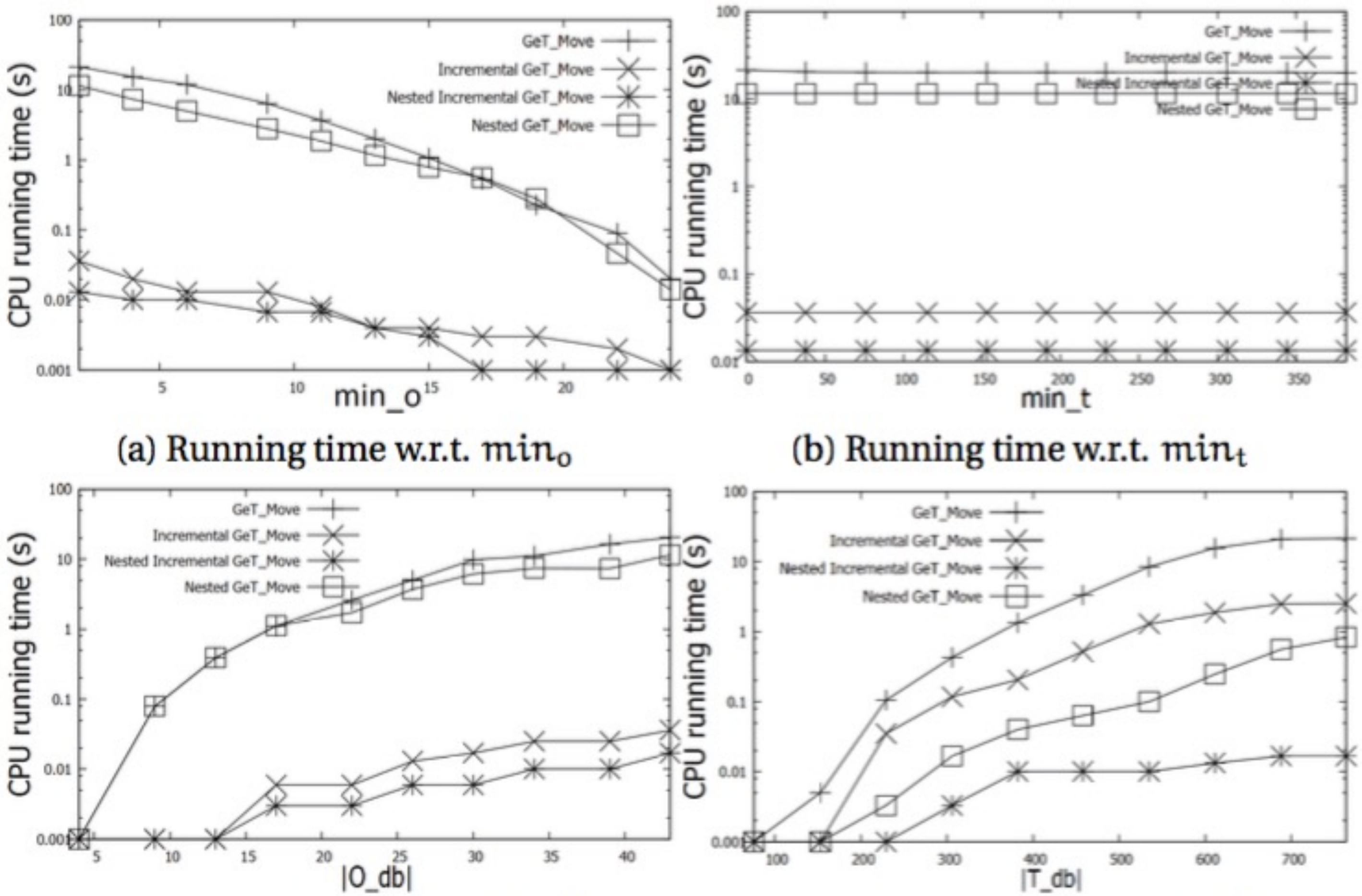
Algorithm 3: Fully Nested Block Partition

```
Input   : a nested cluster matrix CM_N
Output: a set of blocks B
begin
    B := ∅; NestedB := ∅; SpareB := ∅;
    foreach item i ∈ CM_N do
        if i ∩ (i+1) = (i+1) then
            NestedB := NestedB ∪ i;
        else
            NestedB := NestedB ∪ i;
            if |NestedB| ≤ 1 then
                SpareB.push_all(NestedB);
                NestedB := ∅
            else
                B := B ∪ NestedB;
                NestedB := ∅
    return B := B ∪ SpareB;
where the purpose SpareB.push_all(NestedB) function is to put
NestedB to SpareB.
```

The Global Approach

- *Apply the nested and segment nested Greedy* algorithm
- Apply the partition algorithm to get blocks
- Apply incremental mining with FCI

Experimental Results



(a) Running time w.r.t. min_o

(b) Running time w.r.t. min_t

(c) Running time w.r.t. $|O_{db}|$

(d) Running time w.r.t. $|T_{db}|$

Figure 3.17: Running time on Swainsoni dataset.

Experimental Results

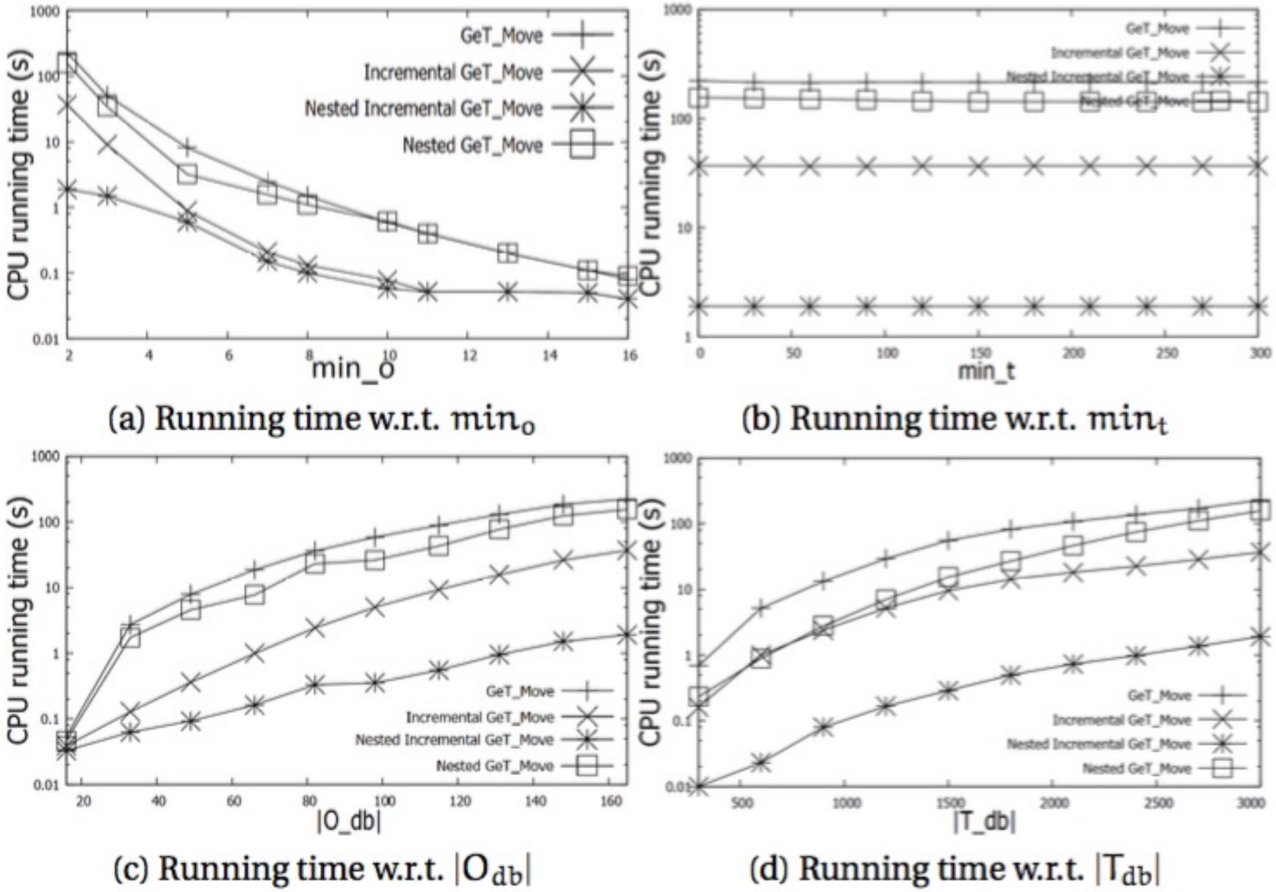
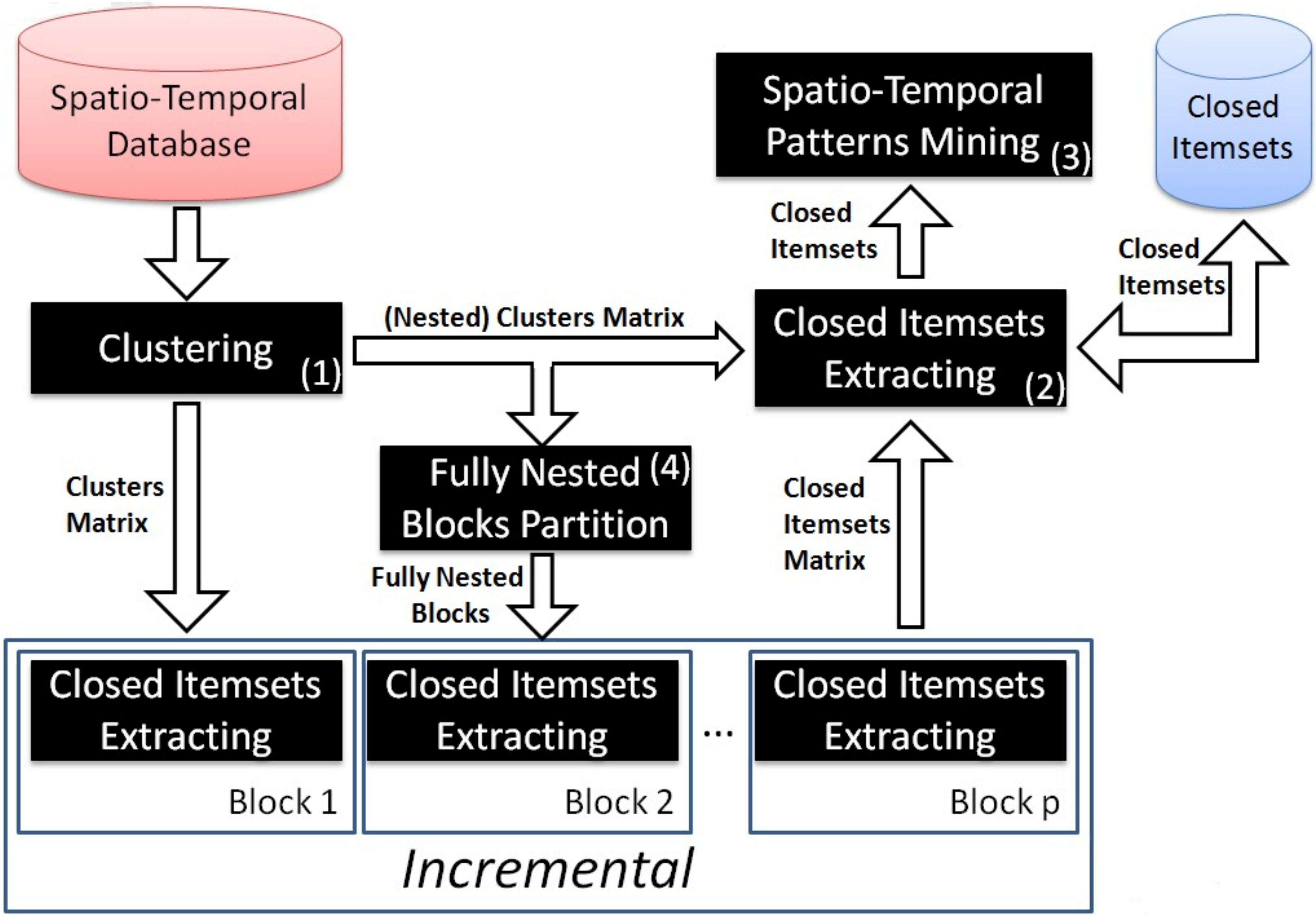
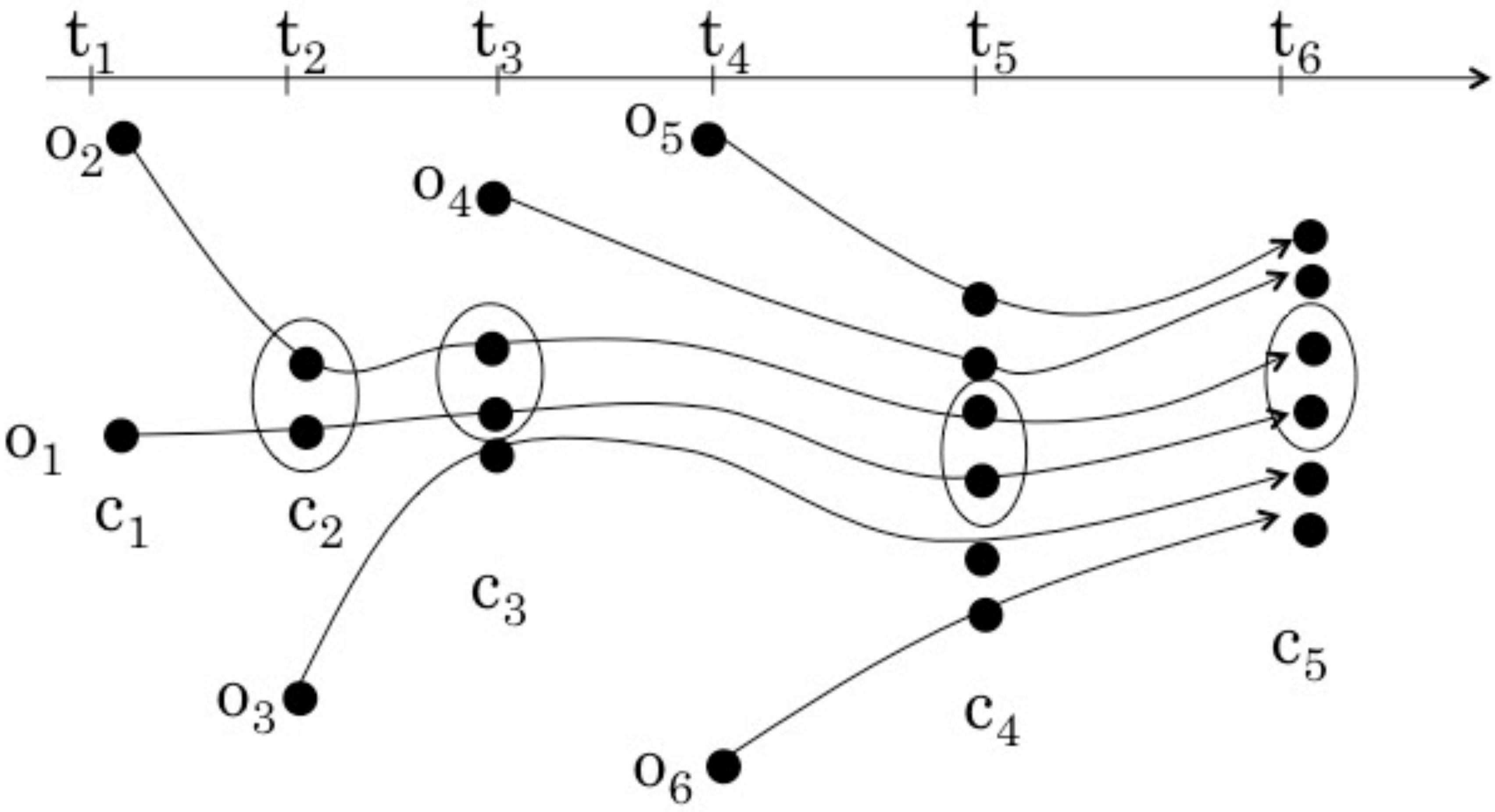


Figure 3.18: Running time on Buffalo dataset.

The Full Process

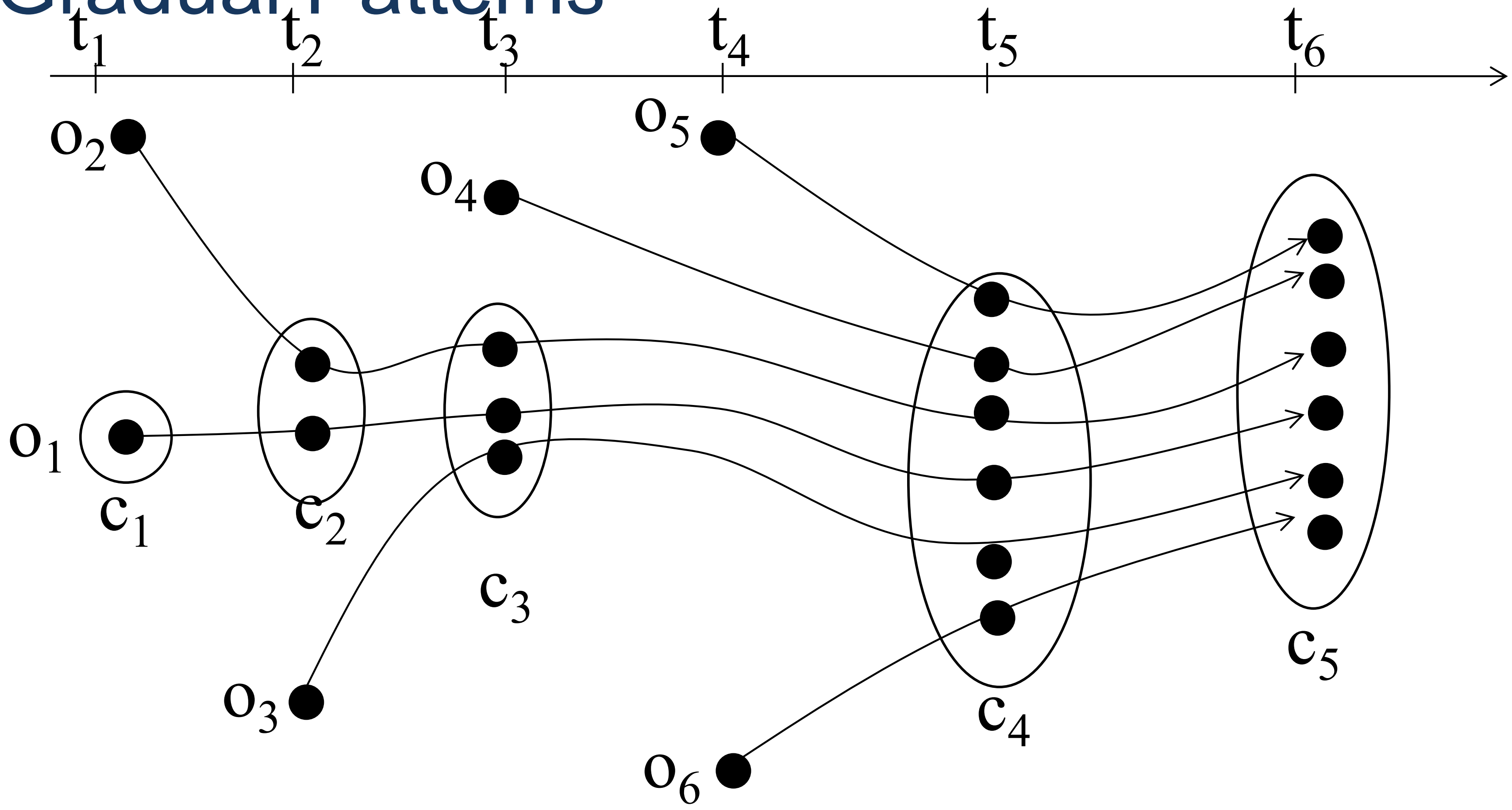


What’s about lord of the Rings?

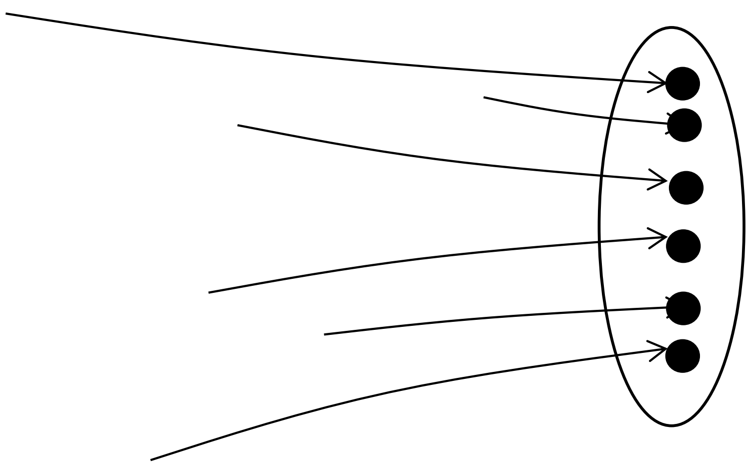


What's about lord of the Rings?

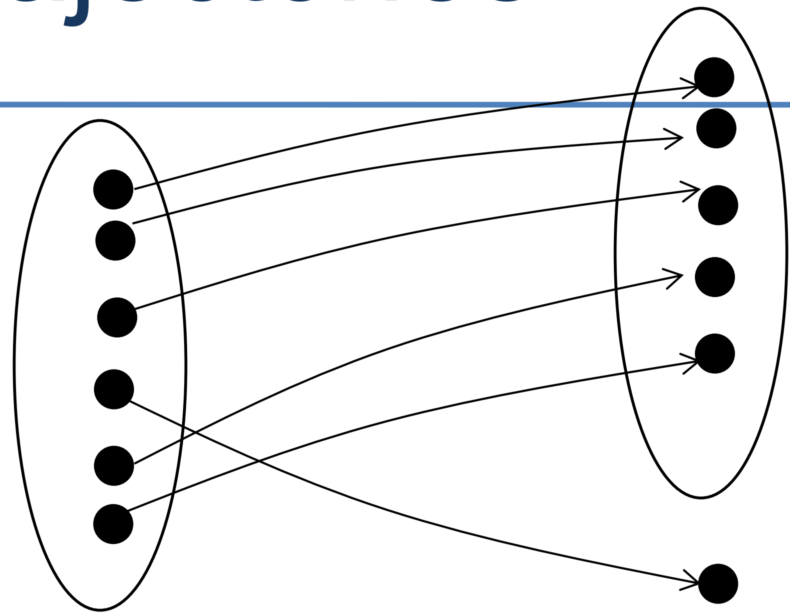
- Gradual Patterns



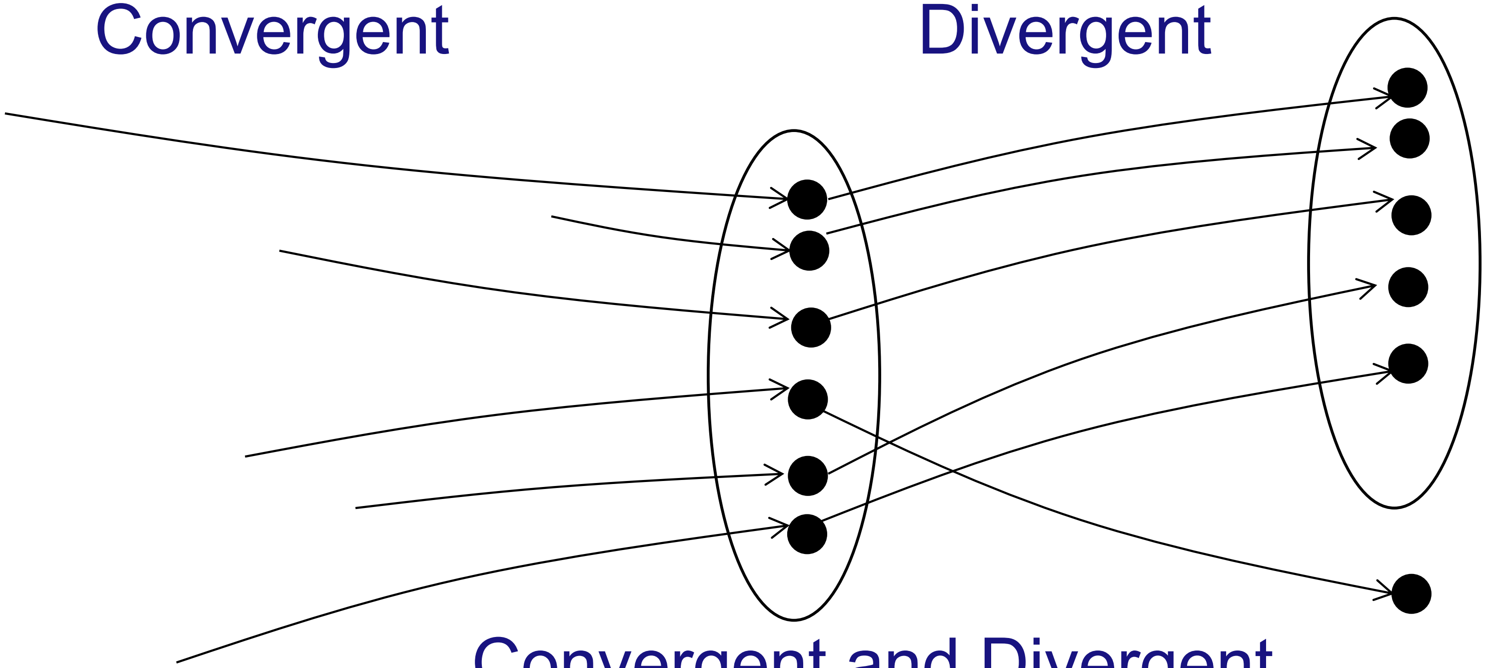
…Gradual Trajectories



Convergent

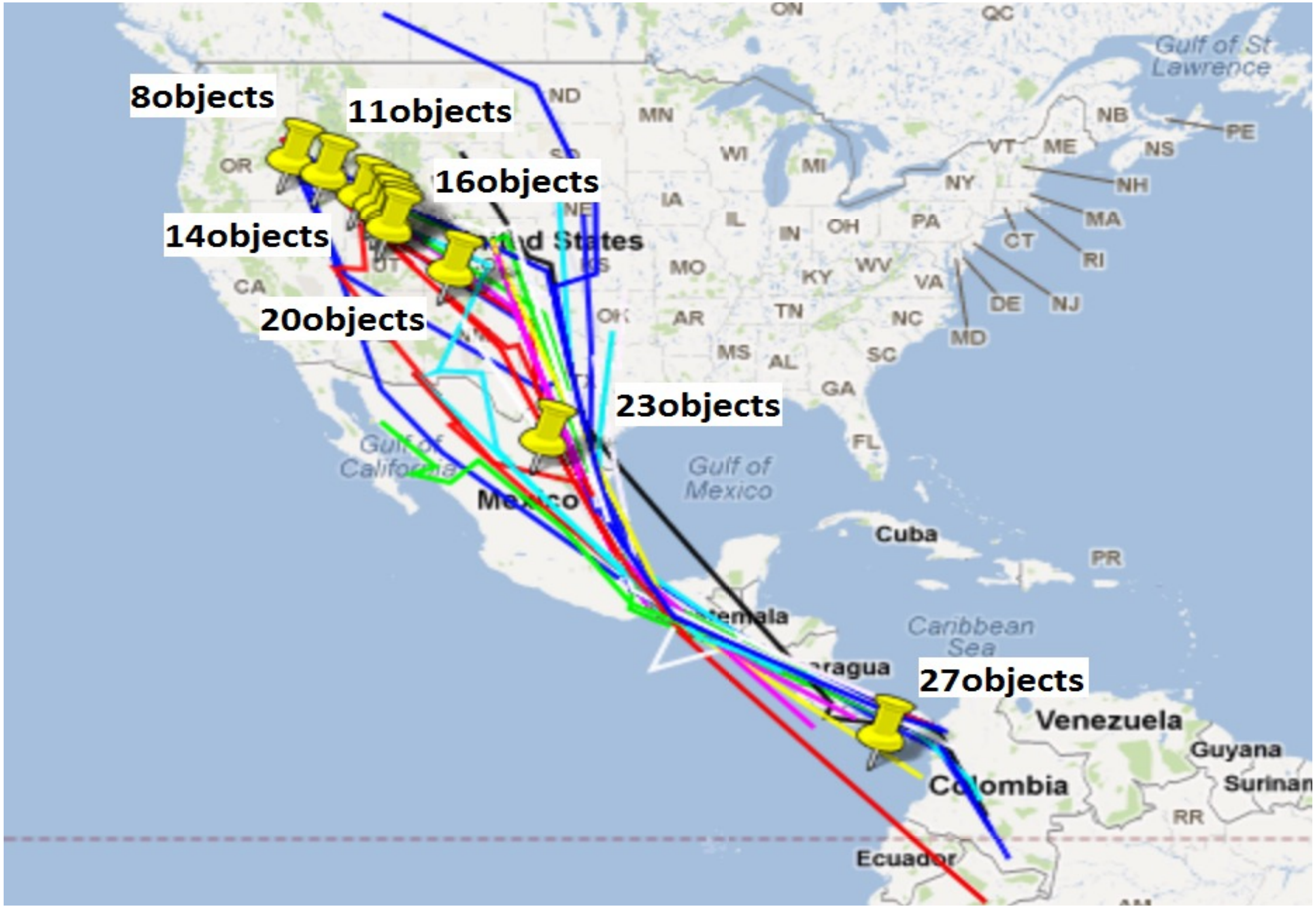


Divergent



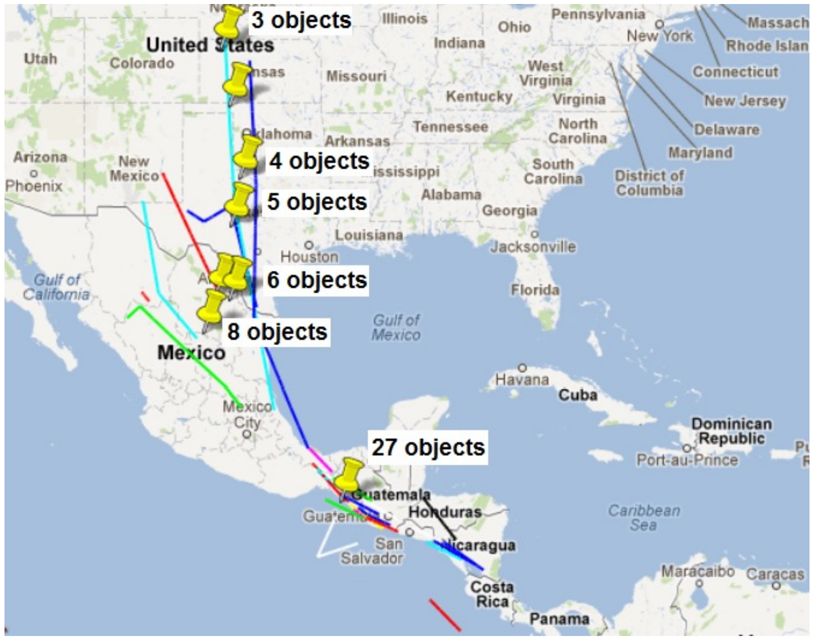
Convergent and Divergent

More seriously



Pattern Definition

- The objects still remain in the next cluster
- The number of objects is increasing (resp. decreasing)
- At least a number of certain timestamps



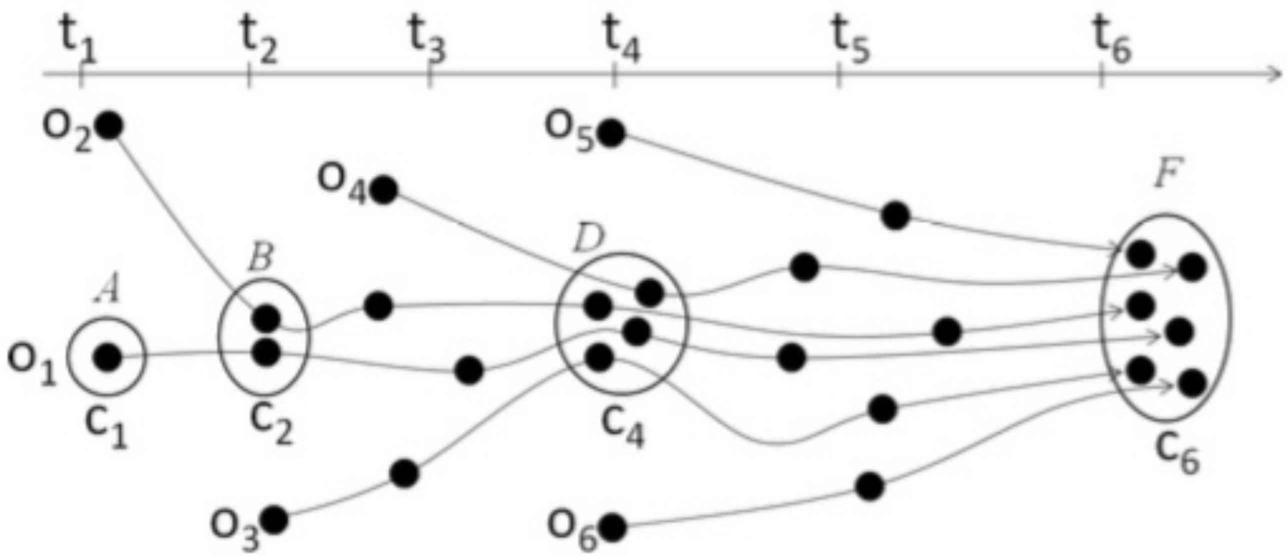
rGpattern

Definition 17. *rGpattern. Given a list of clusters* $C^* = \{c_1, \ldots, c_n\}$ *and a minimum threshold* min_t. C^* *is a rGpattern if:*

$$C^* = C^{\geq} : \begin{cases} (1) : |C^*| \geq min_t. \\ \forall i \in \{1, \ldots, n-1\}, \\ (2) : c_i \subseteq c_{i+1}. \\ (3) : |c_n| > |c_1|. \end{cases}$$

$$C^* = C^{\leq} : \begin{cases} (1) : |C^*| \geq min_t. \\ \forall i \in \{1, \ldots, n-1\}, \\ (2) : c_i \supseteq c_{i+1}. \\ (3) : |c_n| < |c_1|. \end{cases}$$

Example



$C_1^{\geq} = \{c_1, c_2, c_4\}$ is a rGpattern

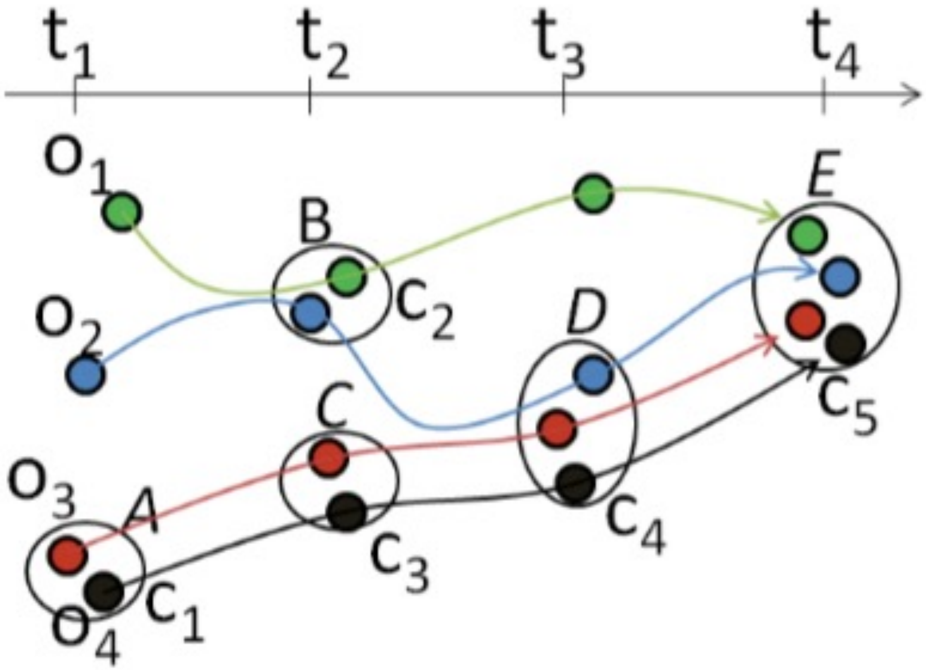
$|C_1^{\geq}| \geq min_t$, $c_1 \subset c_2 \subset c_4$ and $|c_4| = 4 > |c_1| = 1$

$C_1^{\geq} = \{c_1, c_2, c_4\}$, $C_2^{\geq} = \{c_1, c_2, c_6\}$, $C_3^{\geq} = \{c_2, c_4, c_6\}$ and $\boxed{C_4^{\geq} = \{c_1, c_2, c_4, c_6\}}$

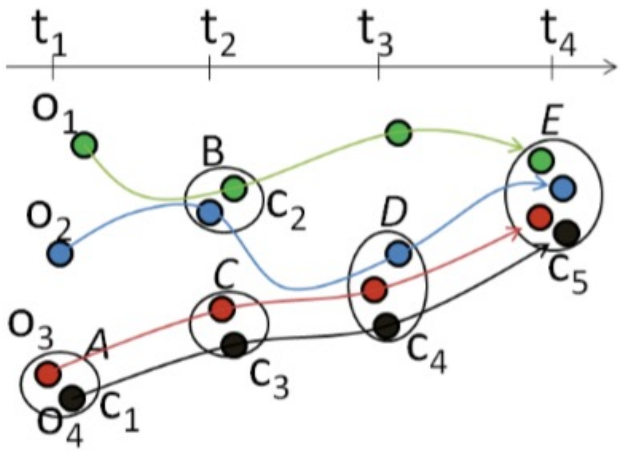
Definition 18. *Maximal rGpattern. Given a rGpattern* $C^* = \{c_1, \ldots, c_n\}$. C^* *is maximal if* $\nexists C'^*, C^* \subset C'^*$ *and* C'^* *is a rGpattern.*

An Illustrative Example

- How many maximal rGpatterns?

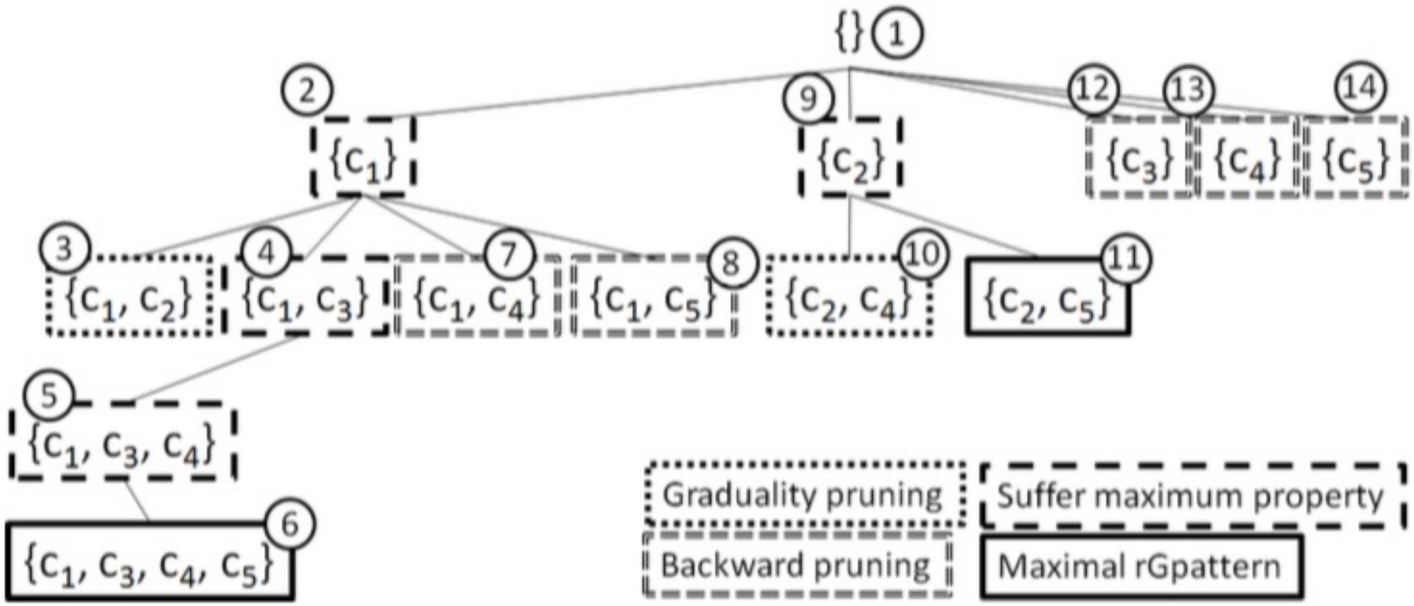


The Algorithm



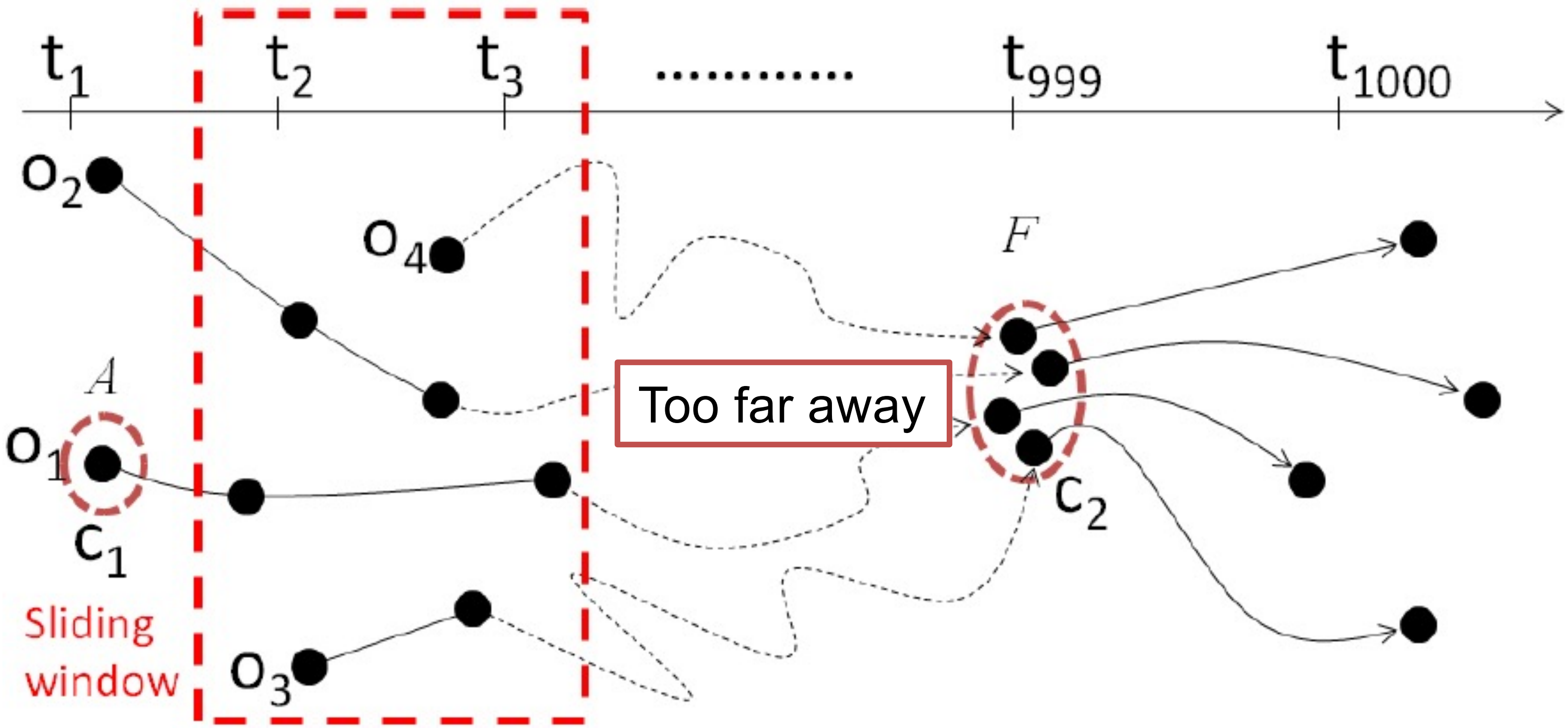
A Depth First Search:

- 2 Pruning Rules (Graduality, Backward)



Time Relaxed Gradual Trajectories

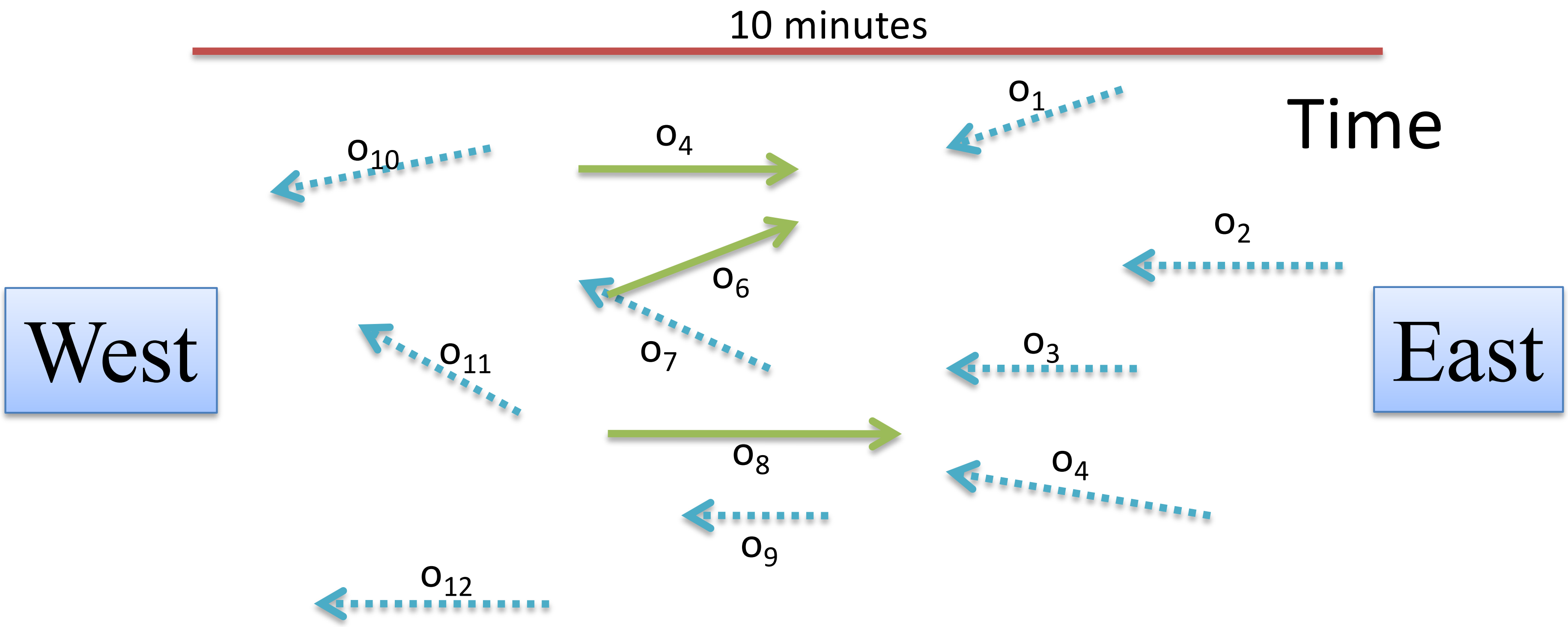
- Timestamps can be:
  - non-consecutive
  - within a sliding time window



Gradual Rules

- Gradual rules highlight complex order correlations of the form *"The more/less X, then the more/less Y"*
  – The nearer to city center, the higher the rent
- Mining gradual patterns plays a crucial role in many applications (numerical data): biological databases, survey databases, data streams…

Issues and Motivations

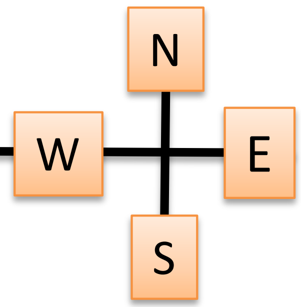


The more time passes the greater the number of objects that move from east to west	Combination of gradual rule and spatio-temporal pattern "*gradual-spatio-temporal*" pattern

Rule Definition

- Set of timestamps: $T = \{t_1, t_2, \dots, t_n\}$
- Set of objects: $O = \{o_1, o_2, \dots, o_z\}$
- A variation: $* \in \{+, -\}$
- Set of directions: $\Delta = \{\Delta_{EW}, \Delta_{WE}, \Delta_{NS}, \Delta_{SN}\}$

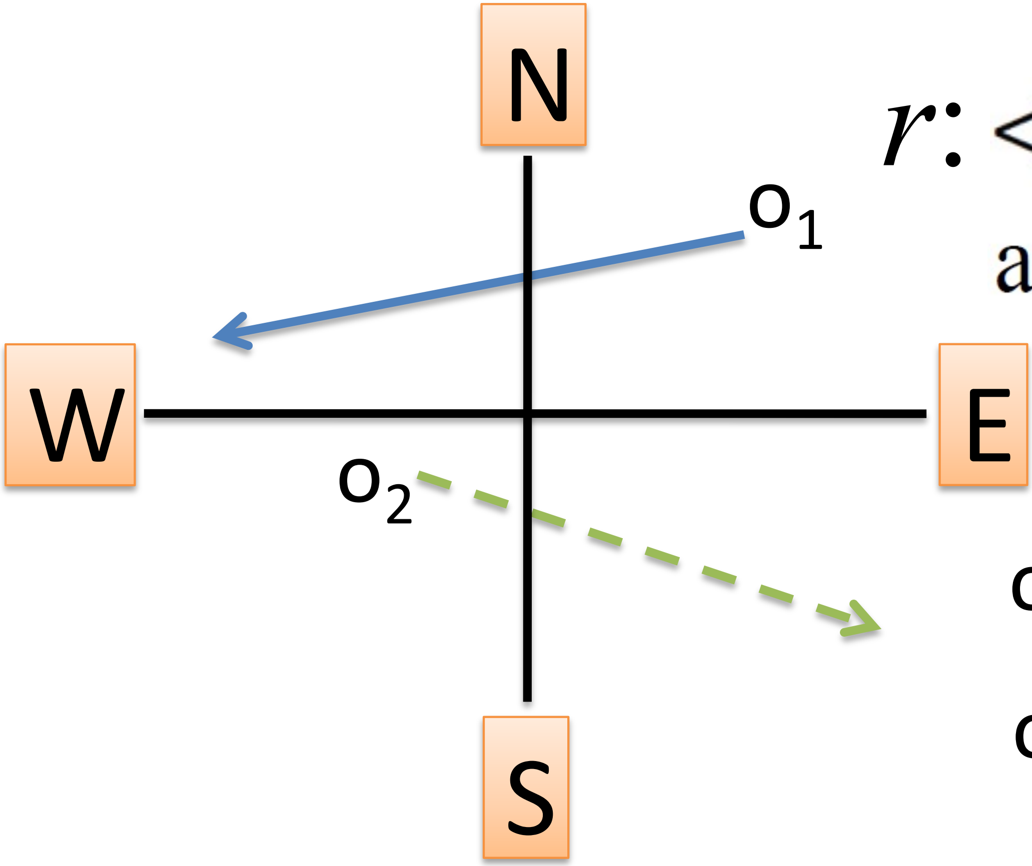
$$< T^+, O^* \rightarrow \Delta_q >$$



"the more time passes
the greater the number of objects
that move from east to west"

$$\boxed{< T^+, O^+ \rightarrow \Delta_{EW} >}$$

Positive Moving Objects



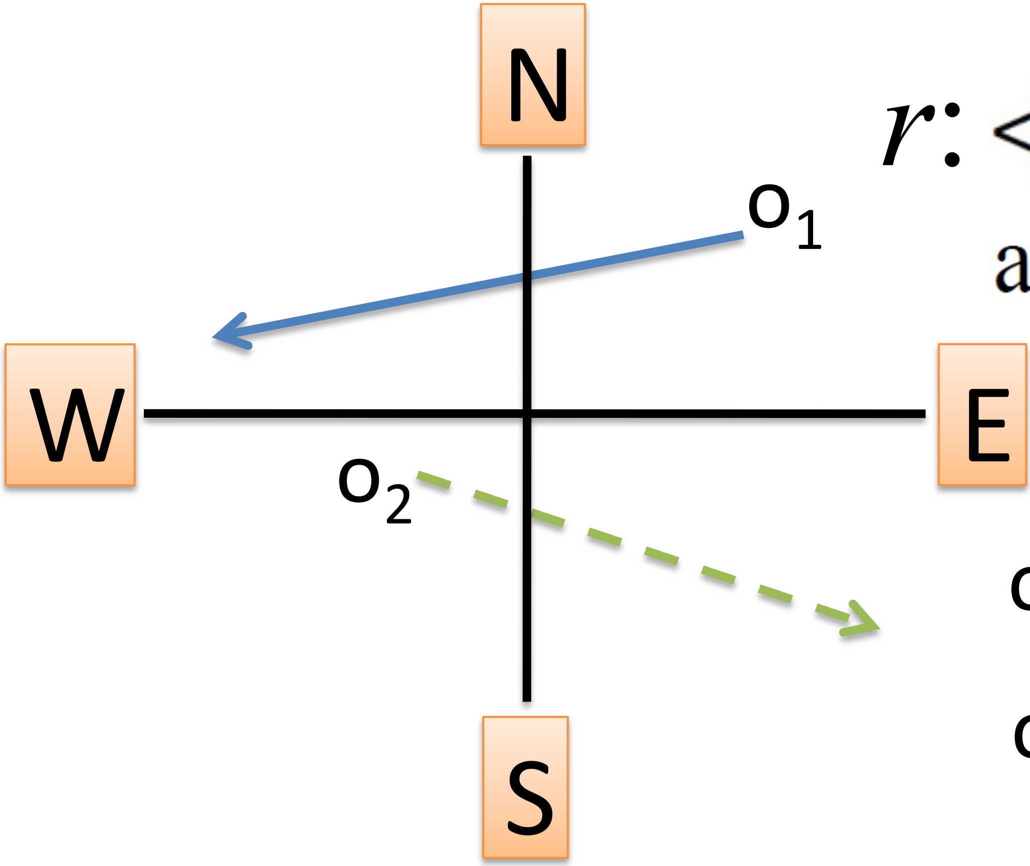
$r: < T^+, O^+ \rightarrow \Delta_{EW} >$

a time segment $[t_i, t_{i+1}]$

o_1: $p_{moving,r_{[t_i,t_{i+1}]}}$

o_2: $n_{moving,r_{[t_i,t_{i+1}]}}$

Support of a Rule on a Timestamp



$r: < T^{+}, O^{+} \rightarrow \Delta_{EW} >$

a time segment $[t_i, t_{i+1}]$

o1: $p_{moving, r_{[t_i, t_{i+1}]}}$

o2: $n_{moving, r_{[t_i, t_{i+1}]}}$

$$\sigma(r)_{[t_i, t_{i+1}]} = \frac{p_{moving, r_{[t_i, t_{i+1}]}}}{p_{moving, r_{[t_i, t_{i+1}]}} + n_{moving, r_{[t_i, t_{i+1}]}}}$$

Supporting Time Segment

- Supporting time segment, non supporting time segment

$$r: < T^+, O^* \rightarrow \Delta_q >$$

$[t_i, t_{i+1}]$ is a *supporting time segment*

$$if\ O^* = O^+ then \begin{cases} p_{moving, r_{[t_i, t_{i+1}]}} \geq p_{moving, r_{[t_{i-1}, t_i]}} & : condition(1) \\ \sigma(r)_{[t_i, t_{i+1}]} \geq \sigma(r)_{[t_{i-1}, t_i]} & : condition(2) \\ \sigma(r)_{[t_i, t_{i+1}]} \geq \sigma_0 & : condition(3) \end{cases}$$

$$if\ O^* = O^- then \begin{cases} p_{moving, r_{[t_i, t_{i+1}]}} \leq p_{moving, r_{[t_{i-1}, t_i]}} & : condition\ (1) \\ \sigma(r)_{[t_i, t_{i+1}]} \leq \sigma(r)_{[t_{i-1}, t_i]} & : condition(2) \\ \sigma(r)_{[t_i, t_{i+1}]} \geq \sigma_0 & : condition(3) \end{cases}$$

Supporting Time Pattern

$$p_s = (t_i, t_{i+k}), |p_s| = k \ (k \geq 1)$$

p_s is a *k-supporting time pattern* if and only if

$\forall j \ (0 \leq j < k), \ [t_{i+j}, t_{i+j+1}]$ is a *supporting time segment*.

$p_{ns} = (t_i, t_{i+k})$ is a *k-non-supporting time pattern* if and only if

$\forall j \ (0 \leq j < k), [t_{i+j}, t_{i+j+1}]$ is a *non-supporting time segment*.

$p_{nt} = (t_i, t_{i+k})$ is a *k-neutral time pattern* if and only if

$\forall j \ (0 \leq j < k), [t_{i+j}, t_{i+j+1}]$ is a *neutral time segment*.

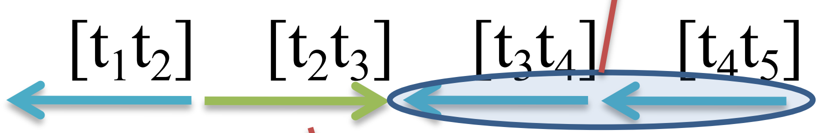
A Running Example

	R1	R2	R3	R4	R5	R6	R7	R8	R9
t_1								o_3	o_1, o_2
t_2	o_9	o_4, o_5		o_7		o_6	o_3	o_1, o_2	
t_3		o_9	o_4, o_5		o_6		o_1, o_2	o_3	o_8
t_4		o_6	o_7	o_4, o_5		o_1, o_2	o_3	o_8	
t_5	o_6	o_7			o_1, o_2, o_4, o_5	o_3	o_8		

$$< T^+, O^+ \rightarrow \Delta_{EW} >$$

$$p_{moving,r_{[t_3,t_4]}} = 5 \text{ (i.e. } o_7, o_1, o_2, o_3, o_8)$$
$$n_{moving,r_{[t_3,t_4]}} = 2 \text{ (i.e. } o_4, o_5)$$
$$\sigma(r)_{[t_3,t_4]} = {}^5/_7 > \sigma_0 = 0.3$$

$p_{s,1} = (t_1, t_2)$ $p_{s,2} = (t_3, t_5)$



$p_{ns,1} = (t_2, t_3)$

Support and Confidence of Rules

a set of timestamps $T = \{t_1, t_2, \dots, t_n\}$

$r{:}\ < T^+, O^* \rightarrow \Delta_q >$

- *We have already see the support of a Rule on one Timestamp:*

$$\sigma(r)_{[t_i,t_{i+1}]} = \frac{p_{moving,r_{[t_i,t_{i+1}]}}}{p_{moving,r_{[t_i,t_{i+1}]}} + n_{moving,r_{[t_i,t_{i+1}]}}}$$

Support and Confidence of Rules

a set of timestamps $T = \{t_1, t_2, \dots, t_n\}$

$r: < T^+, O^* \rightarrow \Delta_q >$

- p_s : is k-supporting time pattern
- p_{ns} : is k-non supporting time pattern
- p_{nt} : is k-neutral time pattern

$$\sigma(r)_T = \sum_{i=1}^{n} |p_{s,i}|$$

$$c(r)_T = \frac{\sum_{i=1}^{n} |p_{s,i}|}{\sum_{i=1}^{n} |p_{s,i}| + \sum_{j=1}^{m} |p_{ns,j}| + \sum_{k=1}^{l} |p_{nt,k}|}$$

A Running Example

	R1	R2	R3	R4	R5	R6	R7	R8	R9
t_1								o_3	o_1, o_2
t_2		o_4, o_5		o_7		o_6	o_3	o_1, o_2	
t_3			o_4, o_5		o_6		o_1, o_2	o_3	o_8
t_4		o_6	o_7	o_4, o_5		o_1, o_2	o_3	o_8	
t_5	o_6	o_7			o_1, o_2, o_4, o_5	o_3	o_8		

$$< T^+, O^+ \rightarrow \Delta_{EW} >$$

$$p_{s,1} = (t_1, t_2) \quad p_{s,2} = (t_3, t_5)$$

$[t_1t_2]$ $[t_2t_3]$ $[t_3t_4]$ $[t_4t_5]$

$$p_{ns,1} = (t_2, t_3)$$

$$\sigma(r)_T = \sum_{i=1}^{2} |p_{s,i}| = 3$$

$$c(r)_T = \frac{\sum_{i=1}^{2} |p_{s,i}|}{\sum_{i=1}^{2} |p_{s,i}| + \sum_{j=1}^{1} |p_{ns,j}|} = \frac{3}{4} = 0.75$$

The Algorithm

Compute support value of each direction at each timestamp.

If the support is larger than predefined threshold then it's considered as a rule candidate.

Compute the supporting time patterns, non-supporting time patterns, neutral time patterns.

Compute support and confidence value for each rule.

Algorithm GSTD: *Gradual-Spatio-Temporal rule Discovering*
Input: a set of objects O, a set of timestamps $T = \{t_1, t_2, \dots, t_n\}$, a set of directions $\Delta = \{\Delta_{EW}, \Delta_{WE}, \Delta_{NS}, \Delta_{SN}\}$, a minimum support threshold σ_0, a minimum moving objects threshold θ.
Output: a set of rules r.

$candidateSet = \emptyset;$
for each timestamp $t_i \in T$
  for each direction $\Psi \in \Delta$
    $r < T^+, O^+ \rightarrow \Psi >.p_{[t_i,t_{i+1}]} \leftarrow ComputePositiveMovingObject(O);$
    $r < T^+, O^+ \rightarrow \Psi >.n_{[t_i,t_{i+1}]} \leftarrow ComputeNegativeMovingObject(O);$
    $r < T^+, O^+ \rightarrow \Psi >.\sigma_{[t_i,t_{i+1}]} \leftarrow \frac{p}{p+n};$
    if $r \notin candidateSet$ **then**
      if $r < T^+, O^+ \rightarrow \Psi >.\sigma_{[t_i,t_{i+1}]} \geq \sigma_0$ **then**
        $r < T^+, O^+ \rightarrow \Psi >.t_{st} \leftarrow t_i;$ // t_{st} is the starting point
        $candidateSet \leftarrow candidateSet \cup r < T^+, O^+ \rightarrow \Psi >;$
      end if
    end if
    for each rule $r \in candidateSet$ // $r < T^+, O^+ \rightarrow \Psi >$
      if $p_{[t_i,t_{i+1}]} \geq p_{[t_{i-1},t_i]}$ *and* $\sigma(r)_{[t_i,t_{i+1}]} \geq \sigma(r)_{[t_{i-1},t_i]}$
        and $\sigma(r)_{[t_i,t_{i+1}]} \geq \sigma_0$ *and* $(p_{[t_i,t_{i+1}]} + n_{[t_i,t_{i+1}]}) \geq \theta$ **then**
        $r.sp \leftarrow r.sp + 1;$ //$r.sp$ is the supporting time
      end if
    end for
  end for
end for
for each rule $r \in candidateSet$ // $r < T^+, O^+ \rightarrow \Psi >$
  $\sigma(r)_T \leftarrow r.sp;$
  $c(r)_T \leftarrow \frac{r.sp}{|T|};$
  return r;
end for

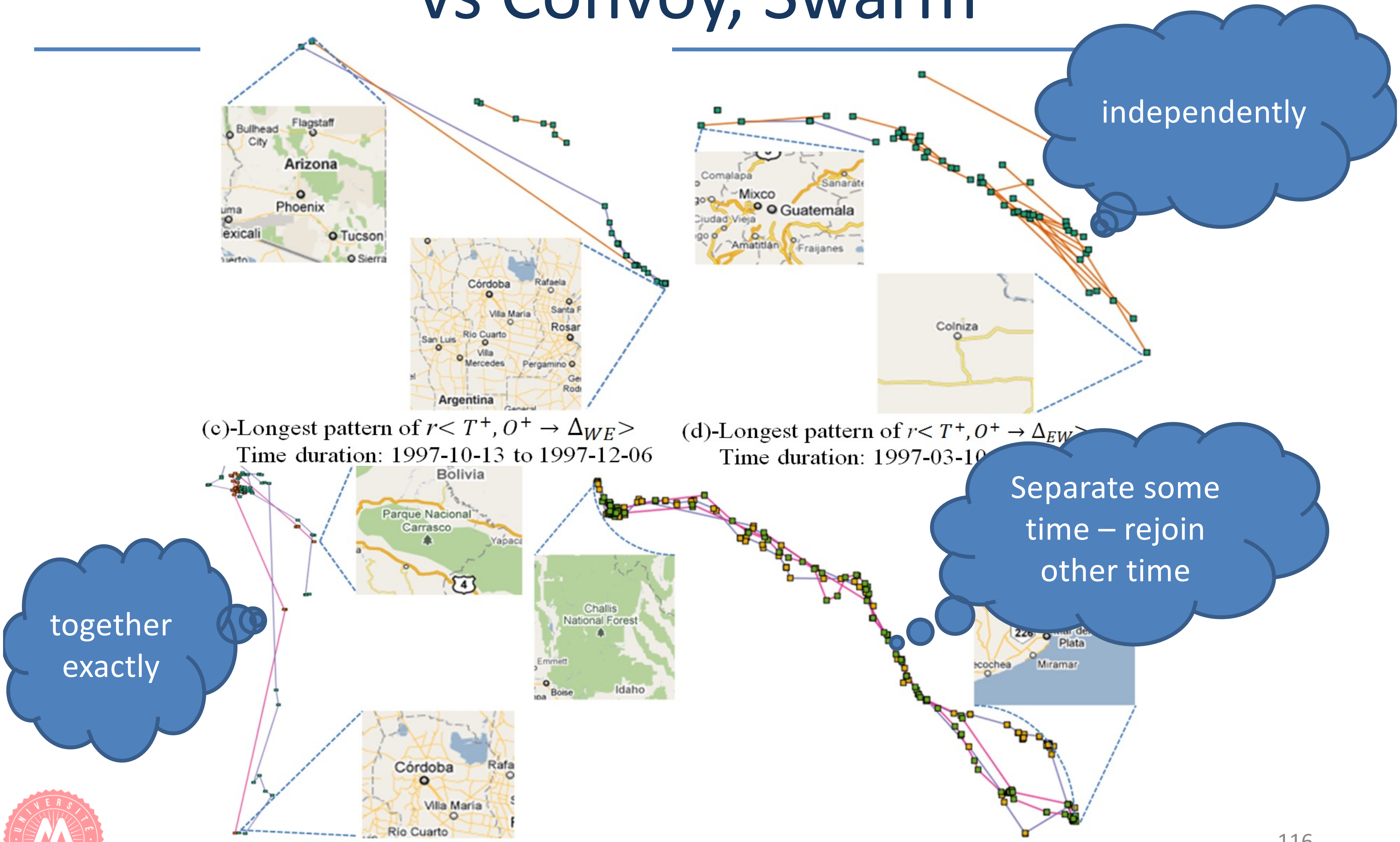
Experimental Results

- Swainsoni dataset (43 objects evolving over time and 4225 different timestamps). July 1995 to June 1998.

SUPPORT AND CONFIDENCE FOR EACH RULE.

Rule	$\sigma(r)_T$	$c(r)_T$
$< T^+, O^+ \rightarrow \Delta_{NS} >$	2653	0.6416
$< T^+, O^+ \rightarrow \Delta_{SN} >$	1944	0.4617
$< T^+, O^+ \rightarrow \Delta_{WE} >$	2725	0.6472
$< T^+, O^+ \rightarrow \Delta_{EW} >$	2060	0.4903

Gradual-Spatio-Temporal Rules vs Convoy, Swarm



(c)-Longest pattern of $r < T^+, O^+ \rightarrow \Delta_{WE} >$
Time duration: 1997-10-13 to 1997-12-06

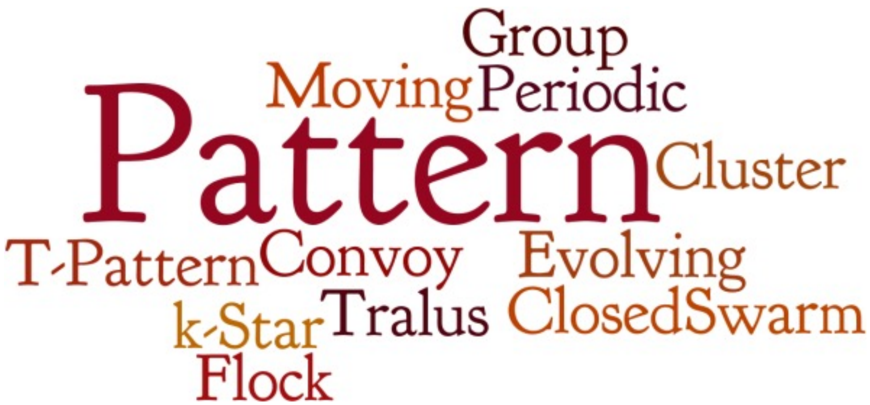
(d)-Longest pattern of $r < T^+, O^+ \rightarrow \Delta_{EW} >$
Time duration: 1997-03-10

(e)-Convoy
Time duration: 1996-11-17 to 1997-02-28

(f)-Closed swarm
Time duration: 1996-08-29 to 1997-06-17

The Most Informative Patterns

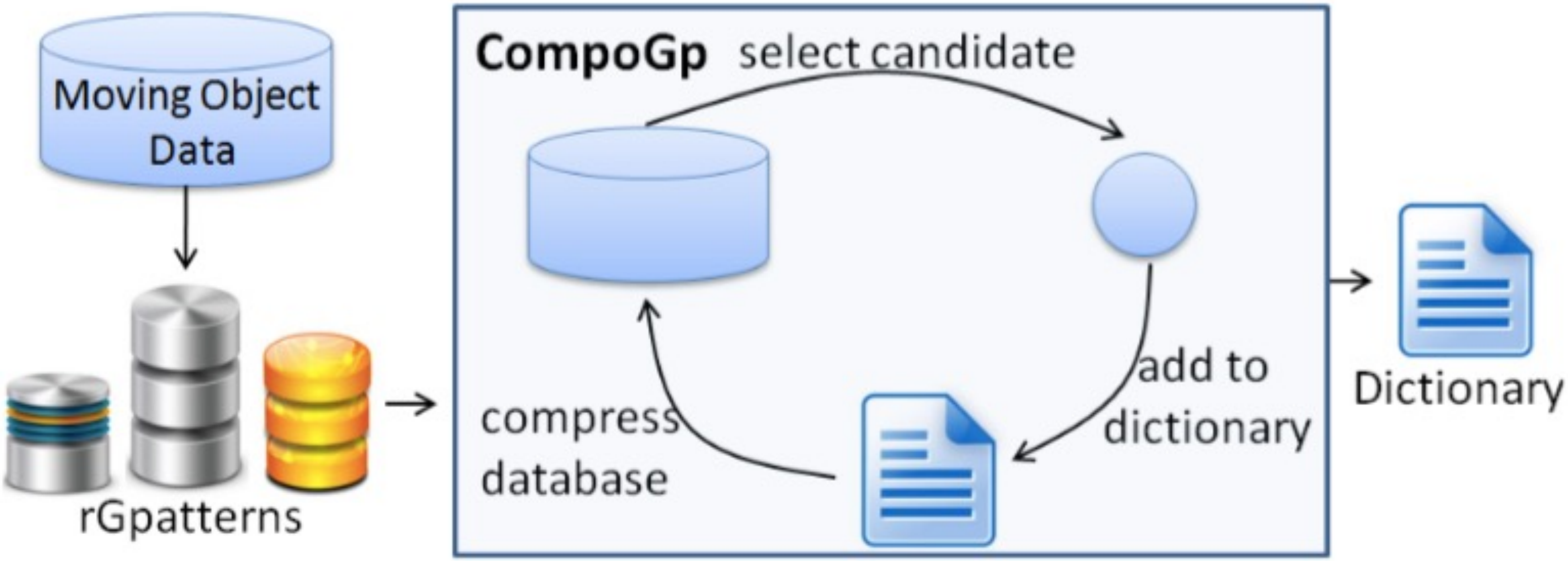
Patterns



data

T_{DB}		t_1			t_2		t_3	t_4	t_5		
Clusters C_{DB}		c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}
O_{DB}	o_1	1			1		1	1	1		
	o_2			1	1		1	1			1
	o_3						1				
	o_4		1			1		1		1	
	o_5		1			1		1		1	

The Main Idea



Problem Statement

- Given a spatio-temporal DB O_{db} and a set of patterns F (extracted from O_{db})
- Discover the optimal dictionary P (subset of F)
  - compresses the data best w.r.t. the given encoding schema

MDL approach: $L_P(O_{db}) = L(P) + L(O_{db}|P)$

$$\arg\min_{P \subset F} \Big(\sum_{p \in P} L(p) + L(O_{DB}|P) \Big)$$

- $L(p)$: number of bits to encode the pattern p + extra bit to encode the type of pattern
- $L(O_{db}|P)$: number of bits to encode the dataset O_{db} given P

Encoding Example

T_{DB}		t_1			t_2		t_3	t_4	t_5		
Clusters C_{DB}		c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}
O_{DB}	o_1	1			1		1	1	1		
	o_2			1	1		1	1			1
	o_3						1				
	o_4		1			1		1		1	
	o_5		1			1		1		1	

O_{DB}	Patterns $\mathcal{F}$
$o_1 = c_1c_4c_6c_7c_8$	
$o_2 = c_3c_4c_6c_7c_{10}$	$p_1 = c_1c_4c_6, \bar{1}$
$o_3 = c_6$	$p_2 = c_2c_5c_7c_9, \bar{0}$
$o_4 = c_2c_5c_7c_9$	$p_3 = c_7c_8, \bar{2}$
$o_5 = c_2c_5c_7c_9$	$P_4 = c_4c6c_7\bar{0}$

Encoding Example

T_{DB}		t_1			t_2		t_3	t_4	t_5		
Clusters C_{DB}		c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}
O_{DB}	o_1	1			1		1	1	1		
	o_2			1	1		1	1			1
	o_3						1				
	o_4		1			1		1		1	
	o_5		1			1		1		1	

Encoded O_{DB}	Dictionary $\mathcal{P}$
$o_1 = [p_1, 0][p_3, 0]$	
$o_2 = c_3[p_1, 1][p_3, 0]c_{10}$	$p_1 = c_1c_4c_6, \bar{1}$
$o_3 = [p_1, 2]$	$p_2 = c_2c_5c_7c_9, \bar{0}$
$o_4 = p_2$	$p_3 = c_7c_8, \bar{2}$
$o_5 = p_2$	$p_4 = c_4c_6c_7, \bar{0}$

L(O$_{DB}$/P) = 4 + 6 + 2 + 1 + 1 = 14
L(P) = 4 + 5 + 3 + 4 = 16
L$_P$(O$_{DB}$) = 30

Encoded O_{DB}	Dictionary $\mathcal{P}$
$o_1 = [p_1, 0]c_7c_8$	
$o_2 = c_3[p_1, 1]c_7c_{10}$	$p_1 = c_1c_4c_6, \bar{1}$
$o_3 = [p_1, 2]$	$p_2 = c_2c_5c_7c_9, \bar{0}$
$o_4 = p_2$	
$o_5 = p_2$	

L(O$_{DB}$/P) = 4 + 5 + 2 + 1 + 1 = 13
L(P) = 4 + 5 = 9
L$_P$(O$_{DB}$) = 22

The Algorithm

Compute for each pattern its compression size and select the most efficient

Compute the compression size for each object of the DB. This can be done efficiently by using pointers rather than true value

Algorithm 9: NaiveCompo

Input : Database O_{db}, set of patterns F, int K
Output: Compressing patterns $\mathcal{P}$

begin
  $\mathcal{P} \longleftarrow \emptyset$;
  while $|\mathcal{P}| < K$ **do**
    foreach $p \in F$ **do**
      $O^{d}_{db} \longleftarrow O_{db}$;
      $L^{*}(O^{d}_{db}|p) \longleftarrow \mathtt{CompressionSize}(O^{d}_{db}, p)$;
    $p^{*} \longleftarrow \operatorname{argmin}_{p} L^{*}(O^{d}_{db}|p)$;
    $\mathcal{P} \longleftarrow p^{*}$; $F \longleftarrow F \setminus \{p^{*}\}$;
    Replace all instances of p^{*} in O_{db} by its pointers;
    Replace all instances of p^{*} in F by its pointers;
  output $\mathcal{P}$;

CompressionSize(O^{d}_{db}, p)
begin
  $size \longleftarrow 0$;
  foreach $o \in O_{db}$ **do**
    if $p.involved(o) = true$ **then**
      Replace instance of p in o by its pointers;
  foreach $o \in O_{db}$ **do**
    $size \longleftarrow size + |o|$;
  $size \longleftarrow size + |p| + 1$;
  output $size$;

The 3-patterns

(a) rGpattern ≥ (b) Closed swarm (c) rGpattern≤

From the running Swansoni example

Outline

- A quick reminder on pattern discovery
- The different kinds of patterns from trajectory
- Towards a unified approach for extracting trajectories
- **Deal with and without a spatial component?**
- A concrete illustration
- Conclusion

Spatial or not spatial?

- Trajectories have been defined for dealing with spatio-temporal data

- What are exactly trajectories?

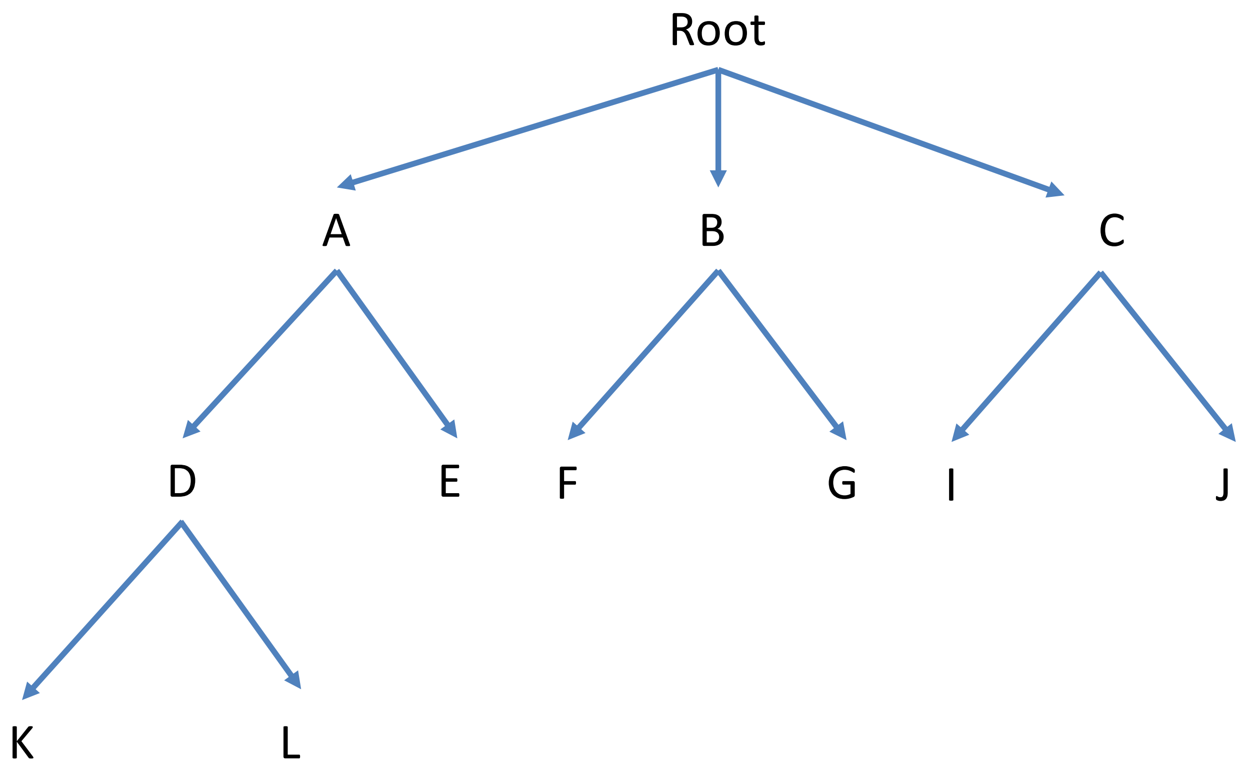
Web Usage Mining

- Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications. (wikipedia)
- Use log entries

123.456.78.9 - - [24/Apr/2018:19:13:44 –0400] “GET /Images/tagline.gif HTTP/1.0”

200 1449 http://www.teced.com/ “Mozilla/4.51 [en] (Win98;I)”

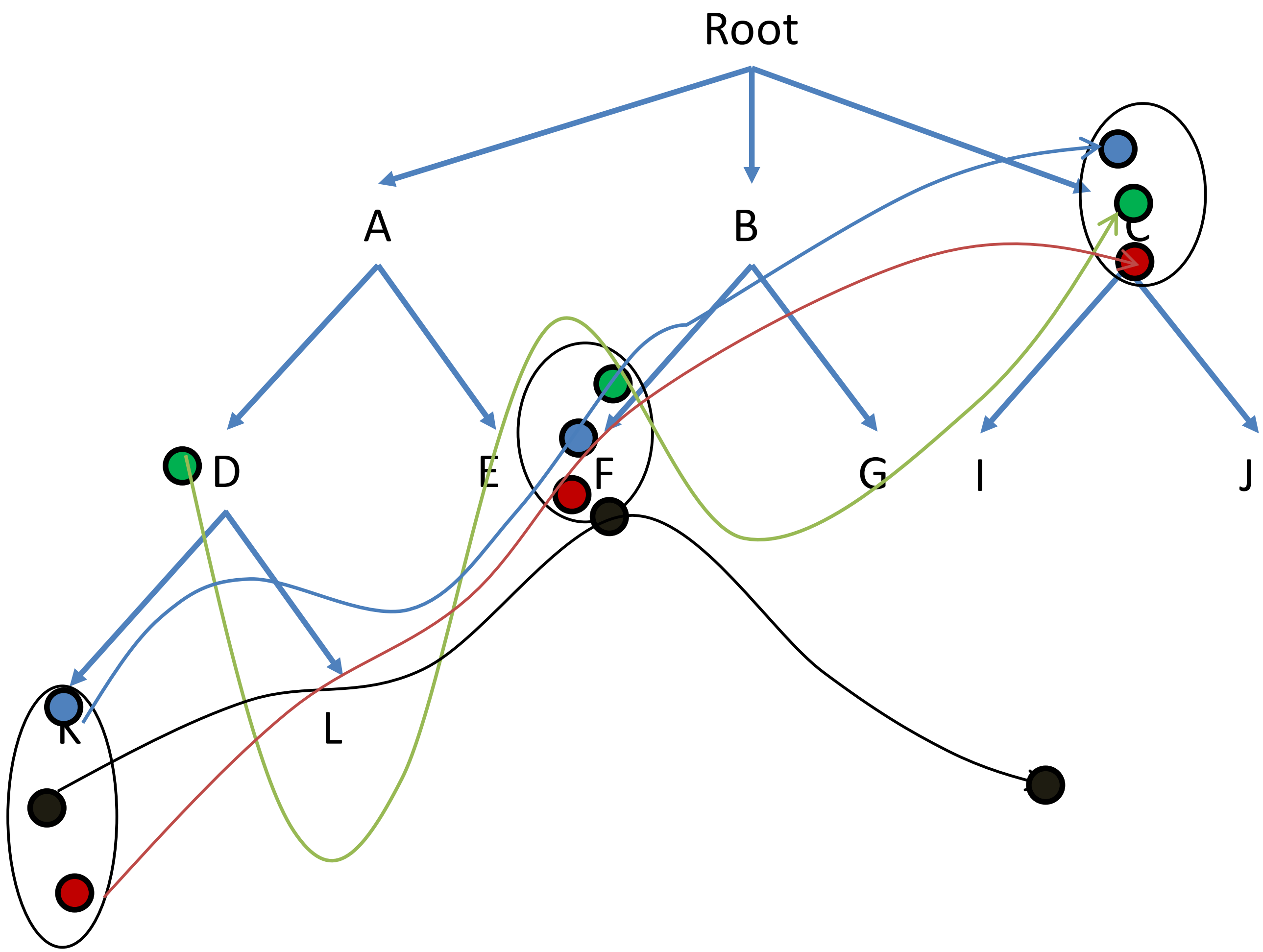
An Illustrative Example



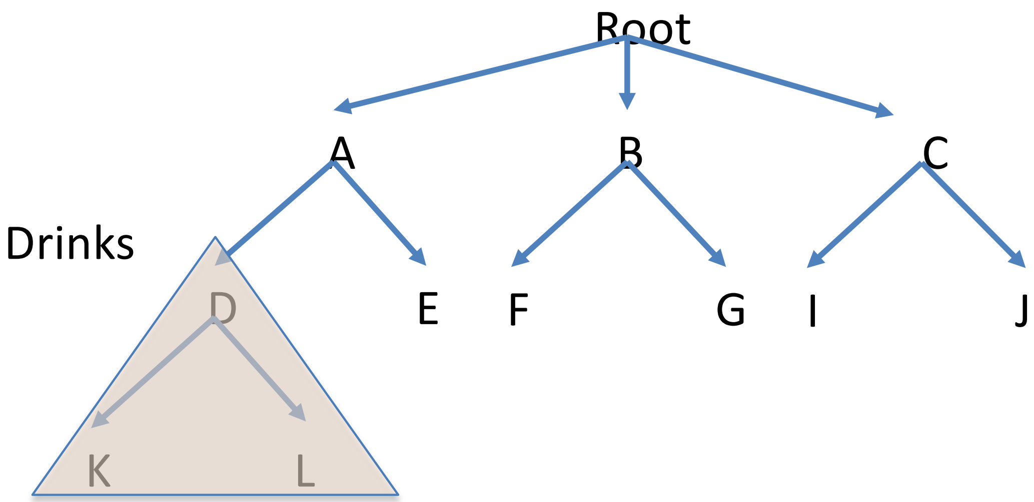
The Cluster Matrix

	T1												T2												T3													
	A	B	C	D	E	F	G	H	I	J	K	L	A	B	C	D	E	F	G	H	I	J	K	L	A	B	C	D	E	F	G	H	I	J	K	L	A	..
C1	1														1																	1						
C2	1													1																	1							
C3		1												1																		1						
C4		1													1														1									
C5				1											1																1							
C6				1												1															1							
C7				1												1															1							
C8								1													1													1				
…																																						

An Illustrative Example

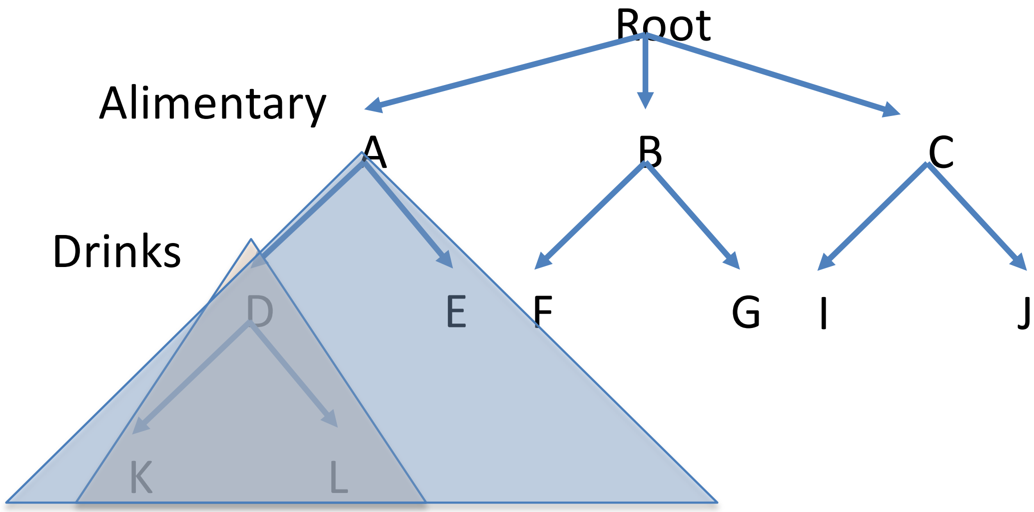


At Different Levels of Granularities



	T1										T2										T3											
	A	B	C	D	E	F	G	H	I	J	A	B	C	D	E	F	G	H	I	J	A	B	C	D	E	F	G	H	I	J	A	..
C1	1												1															1				
C2	1											1															1					
C3		1										1																1				
C4		1											1												1							
C5				1									1														1					
C6				1										1													1					
C7				1										1													1					
C8								1											1											1		
…																																

At Different Levels of Granularities



Group at the upper level

	T1			T2			T3				
	A	B	C	A	B	C	A	B	C	A	..
C1	1					1			1		
C2	1				1				1		
C3		1			1			1			
C4		1				1			1		
C5	1					1		1			
C6	1				1				1		
C7	1			1				1			
C8	1			1					1		
…											

Usefulness

- Personalization: Probabilistic Latent Semantic Analysis

- Efficiency: preload pages, site organization

- E-commerce: Ad placements

PMSI Data

- Using data from the French hospital database (PMSI) to estimate **coronary thrombosis**
- PMSI is a coding system mandatory for hospital public service and private care facilities

Patient general status

D 001	A S A 1	001 BE
D 002	A S A 2	012 BE
D 003	A S A 3	014 BE
D 004	A S A 4	020 BE

Anaesthesia type and anaesthesia recovery

1) General anaesthesia

No addition between all these items

D 010	General anaesthesia	000 BE
D 011	General anaesthesia with intubation	030 BE
D 018	General anaesthesia with blood transfusion	075 BE
D 012	General anaesthesia in Trendelenbourg position	030 BE

D 040	30 min	025 BE	D 049	5 h	250 BE
D 041	1 h	050 BE	D 050	5 h 30 min	275 BE
D 042	1 h 30 min	075 BE	D 051	6 h	300 BE
D 043	2 h	100 BE	D 052	6 h 30 min	325 BE
D 044	2 h 30 min	125 BE	D 053	7 h	350 BE

Clusters

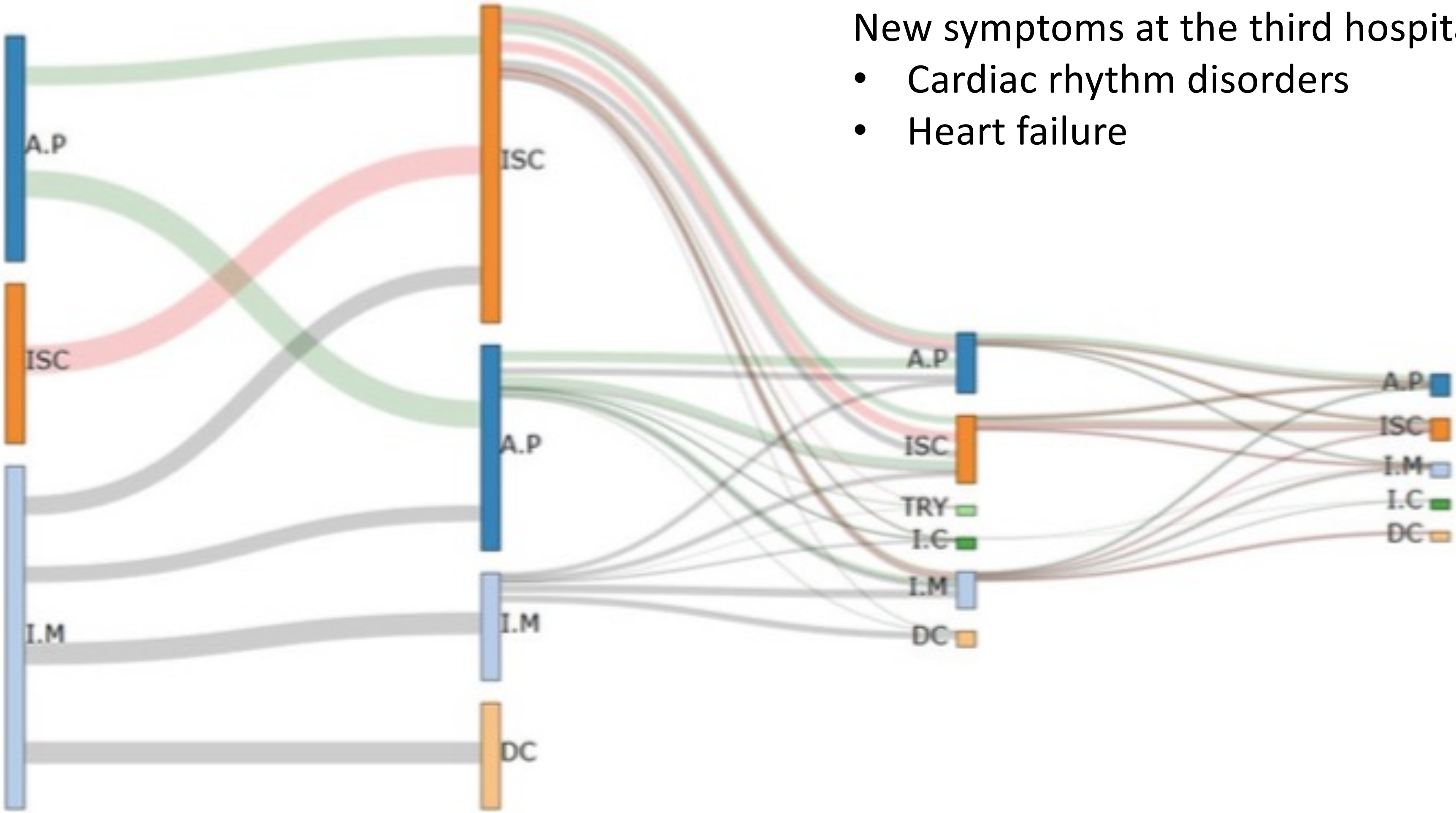
- Clusters are the different types of cares (main diagnosis)

I21: coronary thrombosis
I20: Angina pectoris
R07: Chest pain
E14: Diabetes
I70: Atherosclerosis

Patient	t_0	t_1	t_2	t_3
P_1	I21	E14	I20	I70
P_2	I21	R07	I20	
P_3	I21	E14	I20	I70
P_4	I21	R07	I20	

- From 2009 to 2014, all patients having coronary thrombosis as main diagnosis

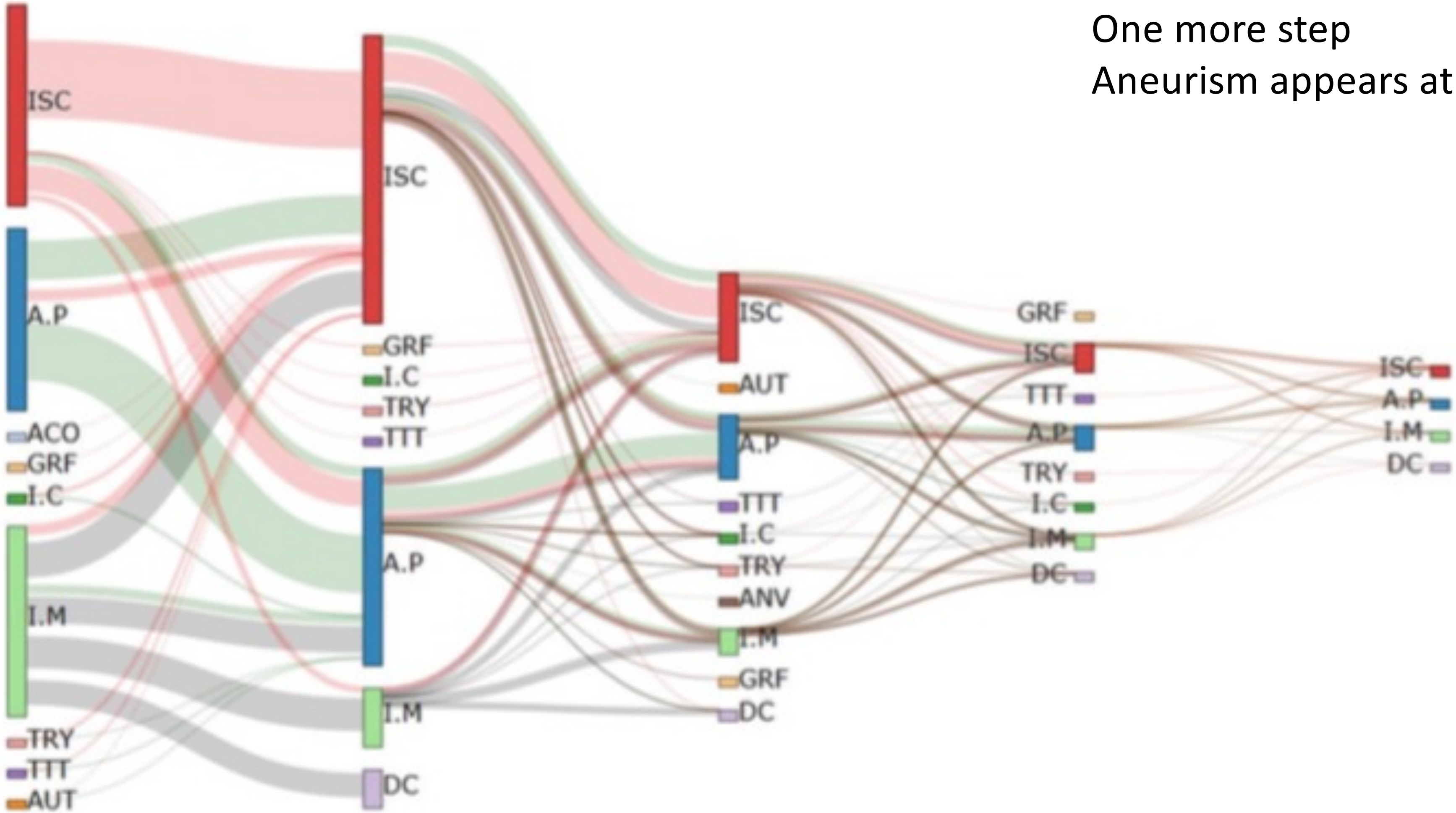
Women > +65



New symptoms at the third hospitalization:

- Cardiac rhythm disorders
- Heart failure

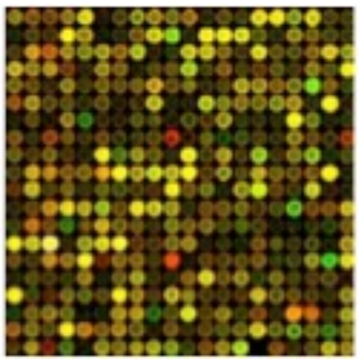
Men >+65



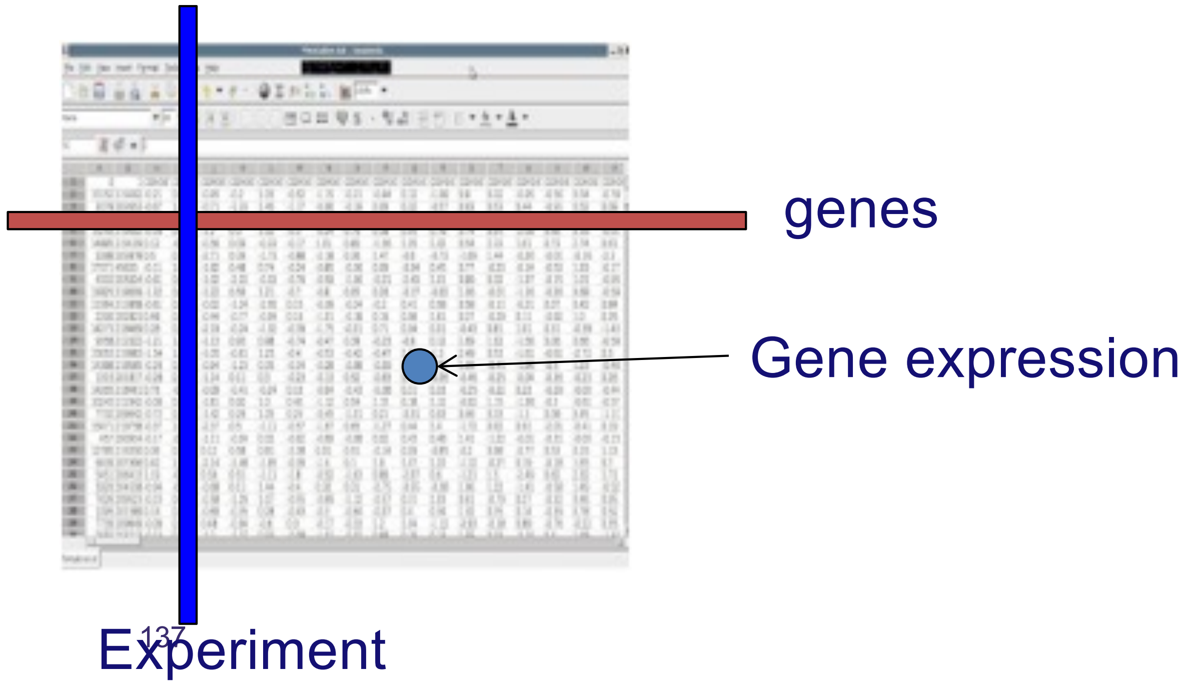
One more step
Aneurism appears at time 3

DNA Arrays Analysis

- DNA Arrays analysis: understanding the reaction of genes on different conditions
- Experiments conducted on Alzheimer disease, Cancer and HIV



Affymetrix U-133 plus 2.0 Array
54,675 probesets



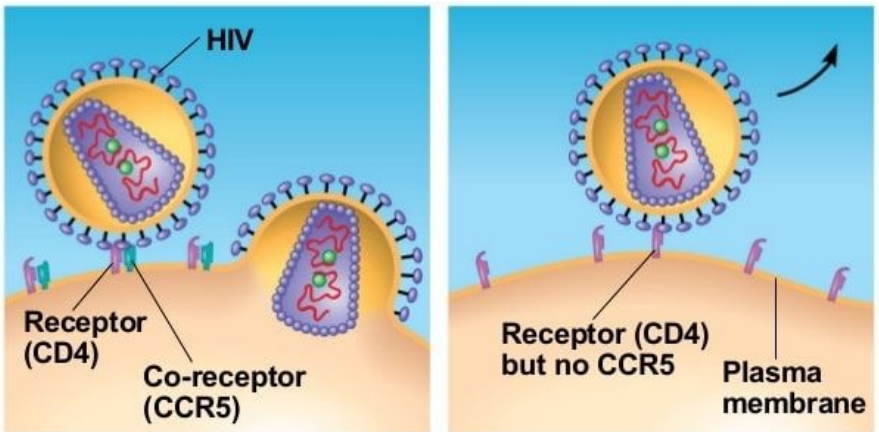
Dealing with HIV Data

- Two majors type of HIV viruses exist: type-1 (HIV-1) and type-2 (HIV- 2)
- HIV-2 is a close relative of the prototype AIDS virus HIV-1
- HIV-2 is biologically similar to HIV-1, but precision concerning clinical outcomes of HIV-2 infected individuals are lacking

Co-receptor CXCR4 – CCR5

- HIV-1 cells invasion is enabled by the binding of envelope glycoproteins to the receptor CD4 and a co-receptor, principally CXCR4 or CCR5, according to the viral strain (X4 or R5, respectively).

Figure 7.11



HIV can infect a cell that has CCR5 on its surface, as in most people.

HIV cannot infect a cell lacking CCR5 on its surface, as in resistant individuals.

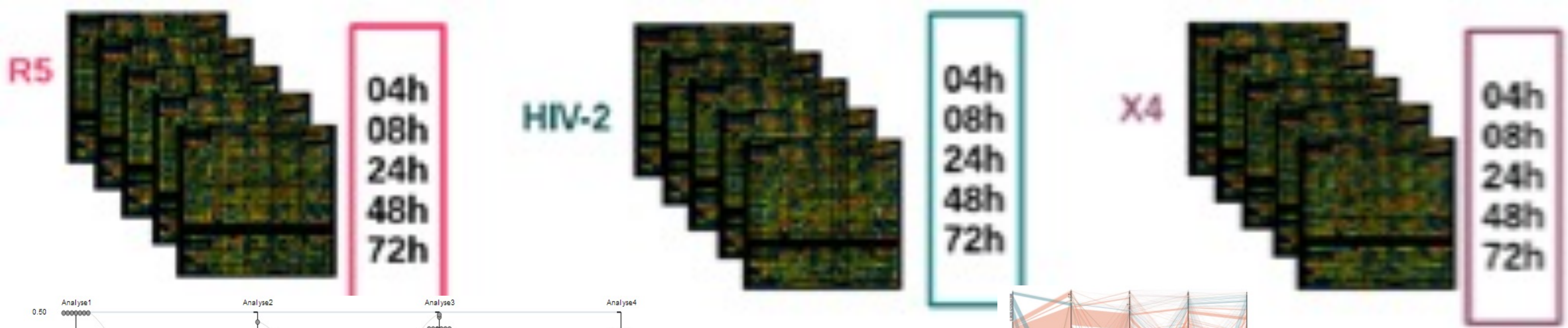
Experiments

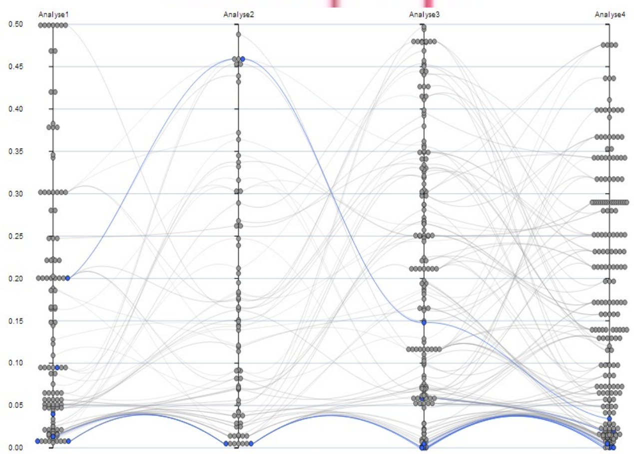
- Expression profile of about 19.000 human genes at 5 points timestamps after infection by one of 3 strains of HIV (HIV-2, X4, R5)

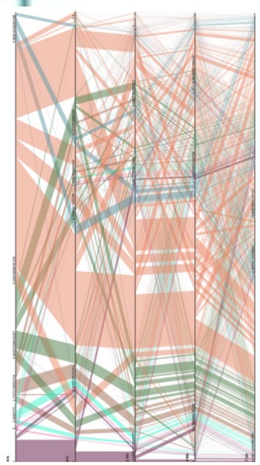
T_{db}		04h	08h		24h	48h	72h
Clusters C_{db}		c_1	c_2	c_3	c_4	c_5	c_6
Genes	$gene_1$		1			1	1
	$gene_2$		1		1	1	1
	$gene_3$	1		1	1	1	1
	$gene_4$	1		1	1	1	

Extracting patterns

- « Real unknown and interesting phenomena. Correlations between biological functions » dixit Biologist. Under investigation







Trajectories for Tweets

- A tweet:
  - 140 characters
  - Tags (e.g. #saopaulo18, @user), RT (retweet)
  - Metatags (location, date, user-id, …)

- Cluster of users
- Cluster of words

Using Tweets

- Identification of patterns of retweets and to understand how information spreads over time in Twitter
- Tweetprofile a research platform for extracting, storing and analyzing the Portuguese Twittosphere for research and journalistic purposes

RetweetPatterns: detection of spatio-temporal patterns of retweets

Tomy Rodrigues[1], Tiago Cunha[1], Dino Ienco[2], Pascal Poncelet[3], and Carlos Soares[1]

Experiments

- A set of retweets extracted at the time of the protests in Brazil
- From June 2013 and July 2013 - 17083 tweets extracted from Twitter during a protests period in Brazil
- After a preprocessing : 260 retweets

Results

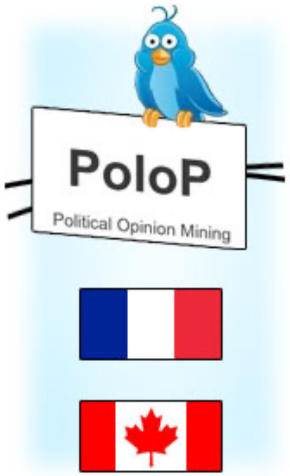
- 18 closed swarms (84 retweets), 5 convoys, 5 moving clusters
- the majority of paths found are located in Rio de Janeiro and São Paulo

Closed Swarm patterns found

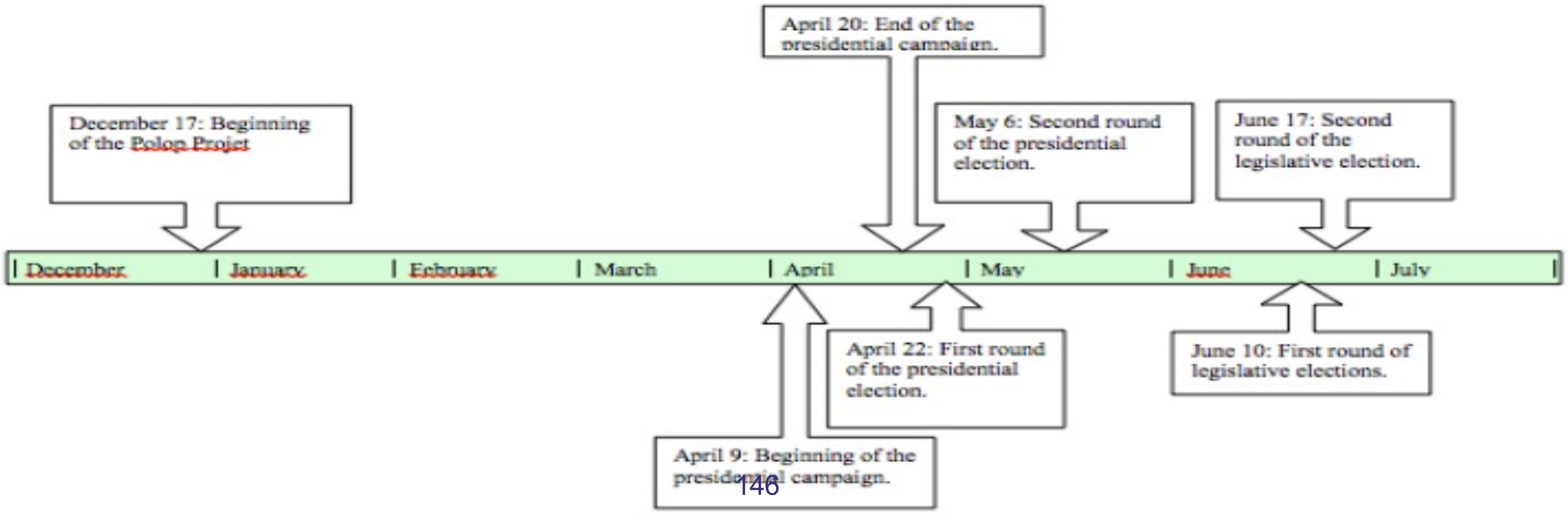
Timeline and wordcloud for the Closed Swarm patterns found

Tweets for Communities

- A set of users sharing the same interest

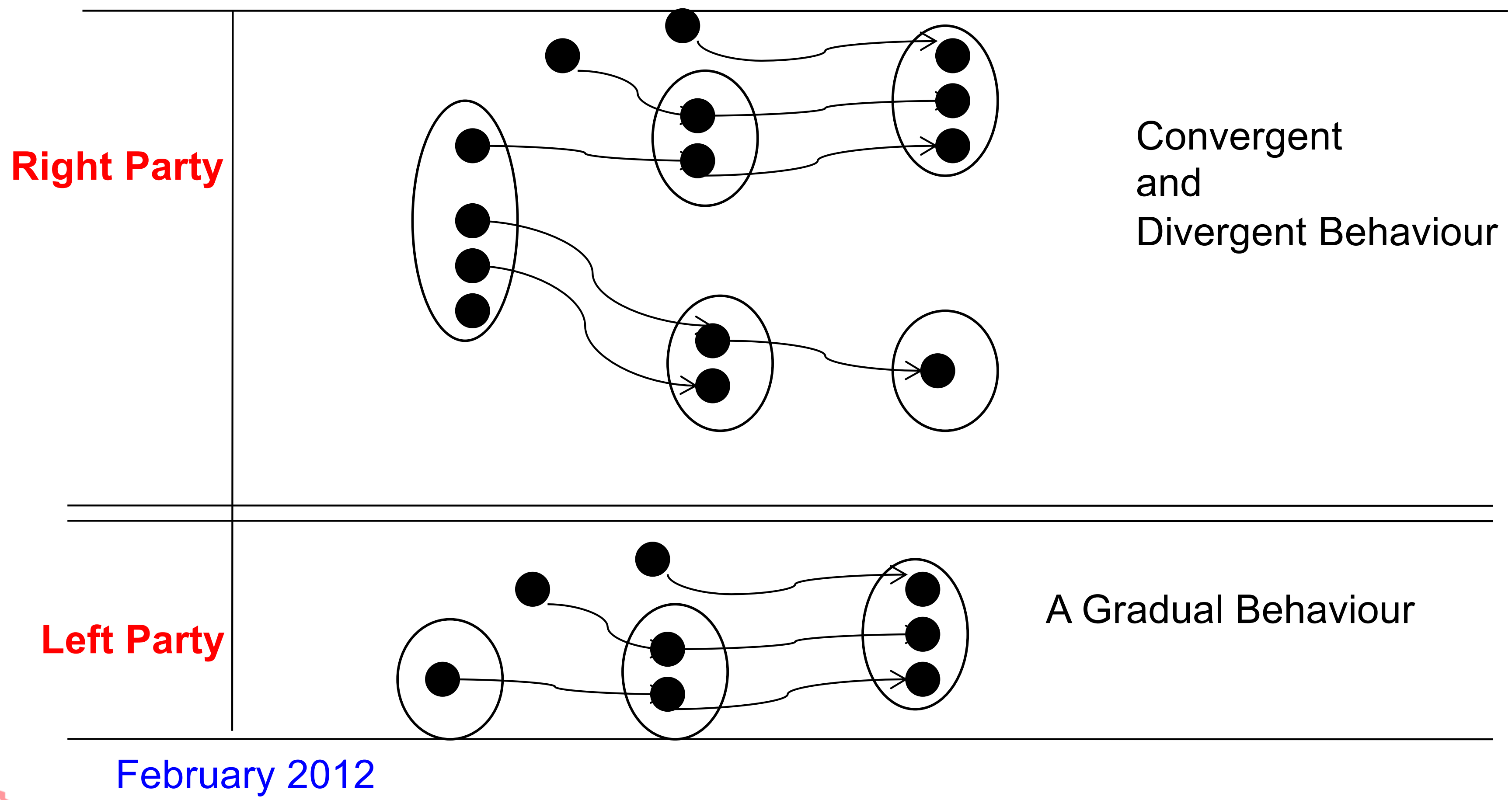


200 Political Leaders
April 2012: 1440471 Tweets



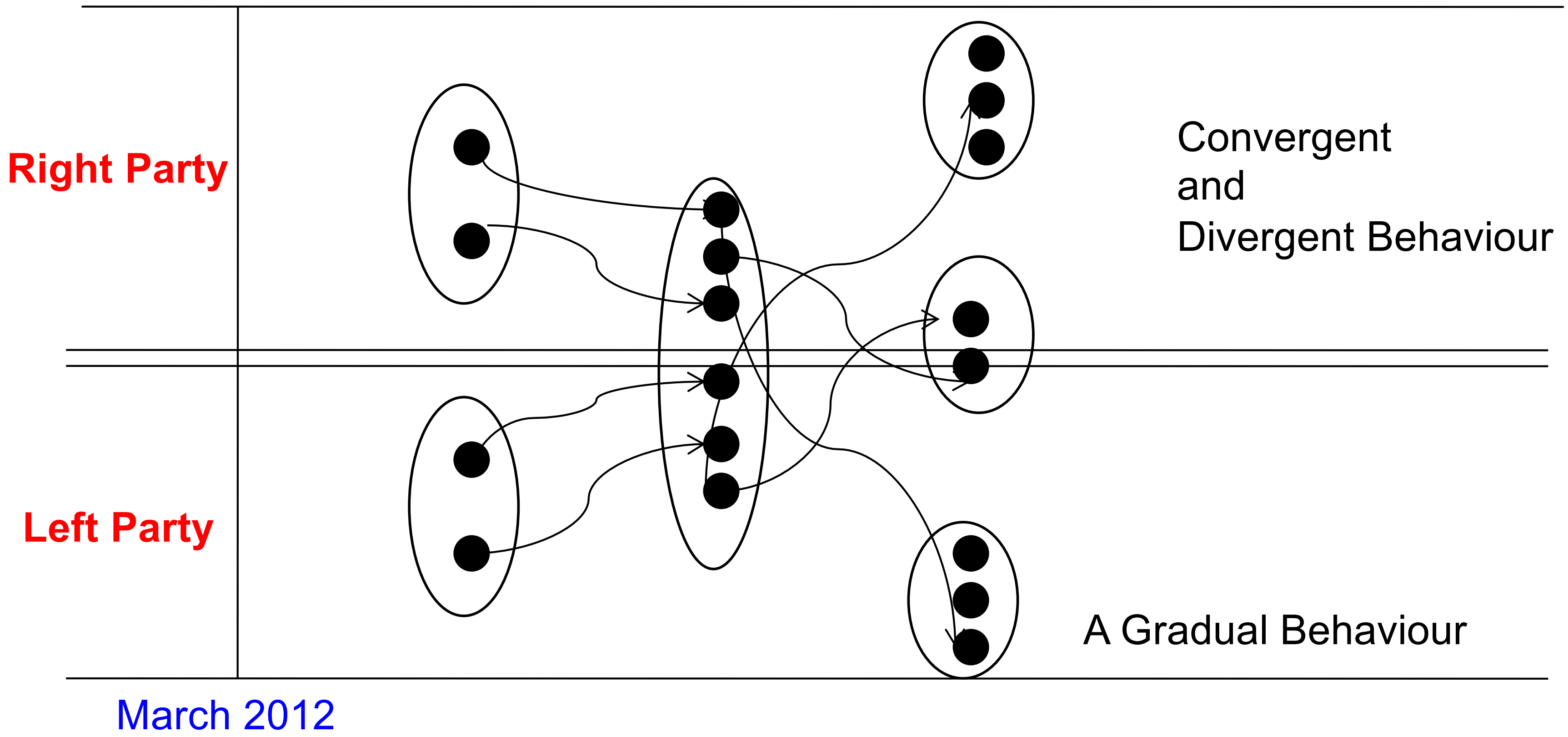
One trajectory

- “*The Debate - The focus of the French election campaign suddenly shifts from the economy to racism and national identity*” (07/02/2012)



One trajectory

- *"Shootings in Toulouse and Montauban: The victims - Seven people have been killed and two wounded in serial gun attacks in south-western France"* (3/2012)



A typical French trajectory

President
May 2012 – May 2017
.

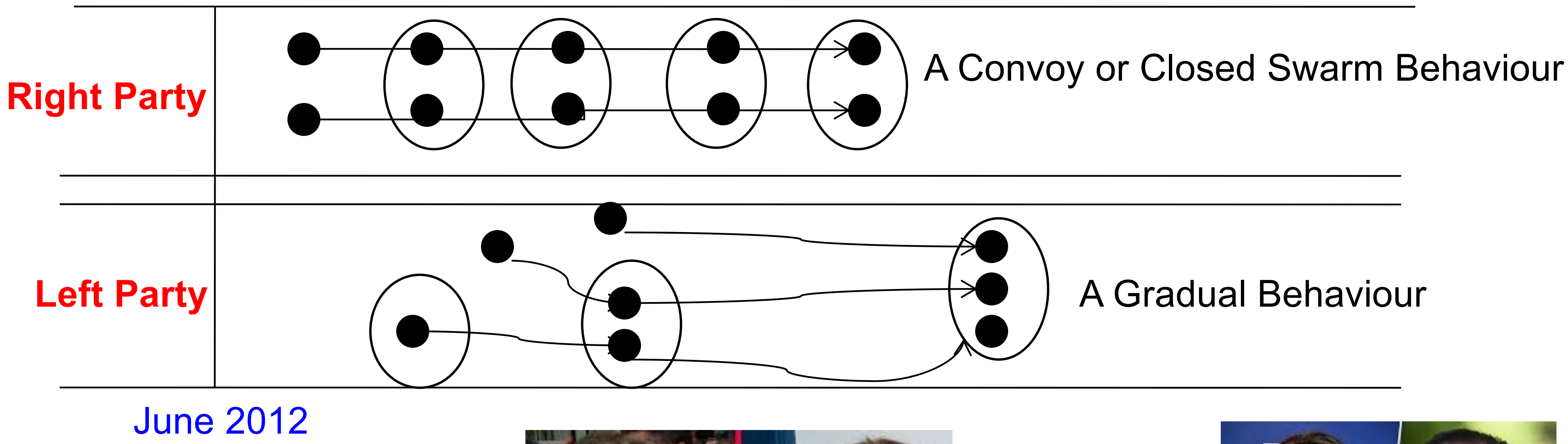
1980 - 2007
Ségolène Royal

2005 - 2014
Valerie Trieweiler

2013 -
Julie Gayet

One trajectory

- *"Courage to Olivier Falorni who has not been unworthy, who has battled alongside La Rochelle residents for so many years with unselfish commitment"* Valerie Trierweiler, First Lady/Girlfriend of the President François Holllande



Left Party

Left Party

Valerie Trieweiler

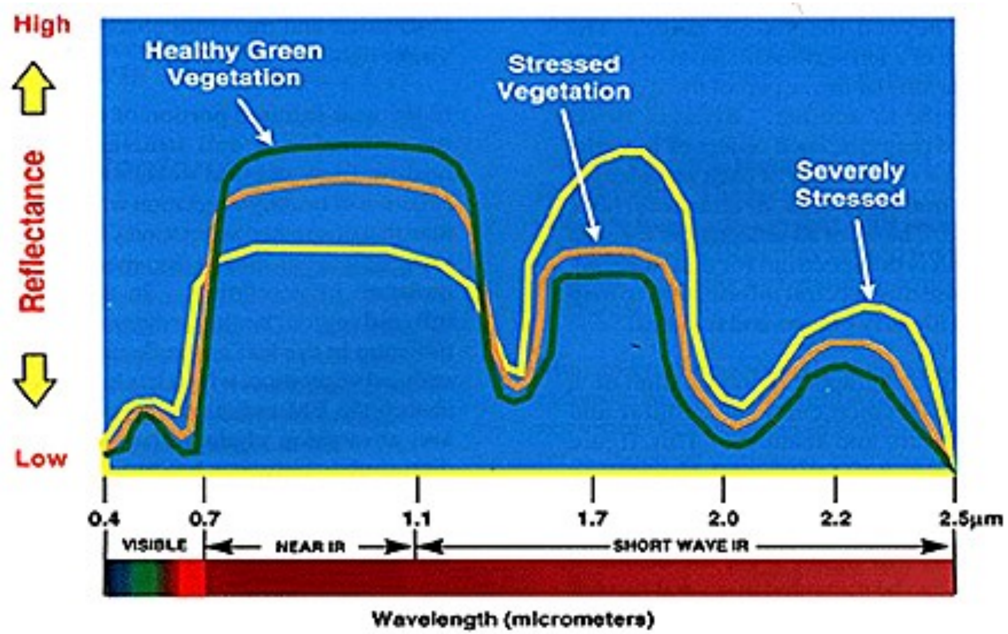
Ségolène Royal

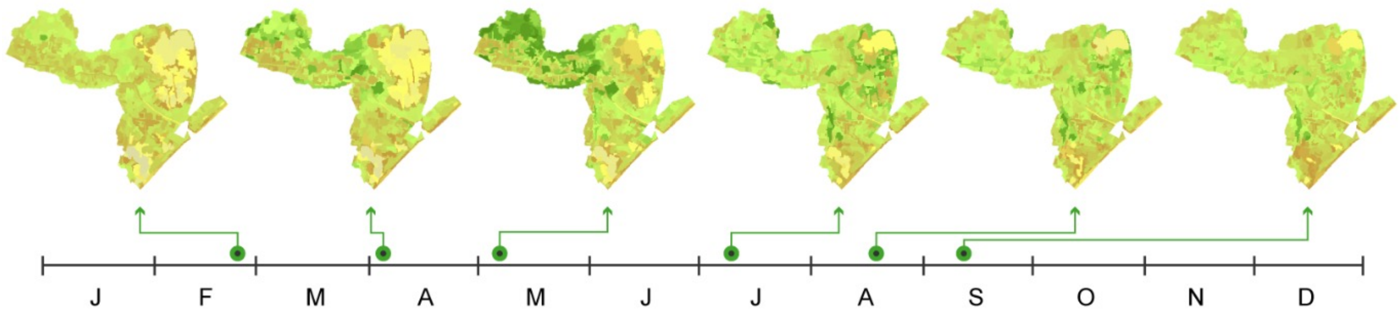
Julie Gayet

Satellite Remote Sensing Time Series

Remote sensing possibilities

- Spectral response of natural vegetation
  - Vegetation types
- Textural response
  - Vegetation structure (mapping physiognomic classes, e.g. shrubland and grassland)
- Temporal response
  - Vegetation structure dynamics



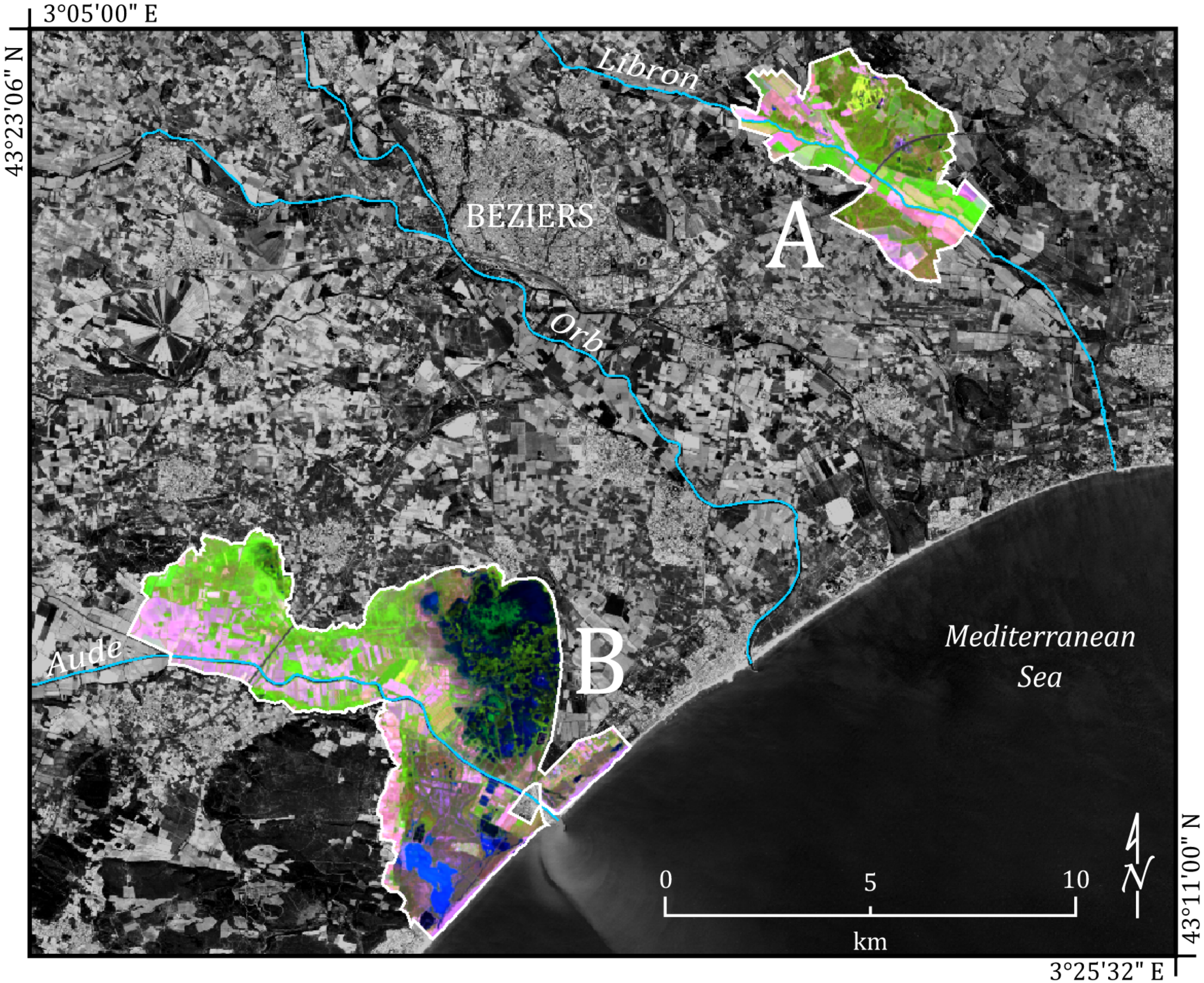


Motivations

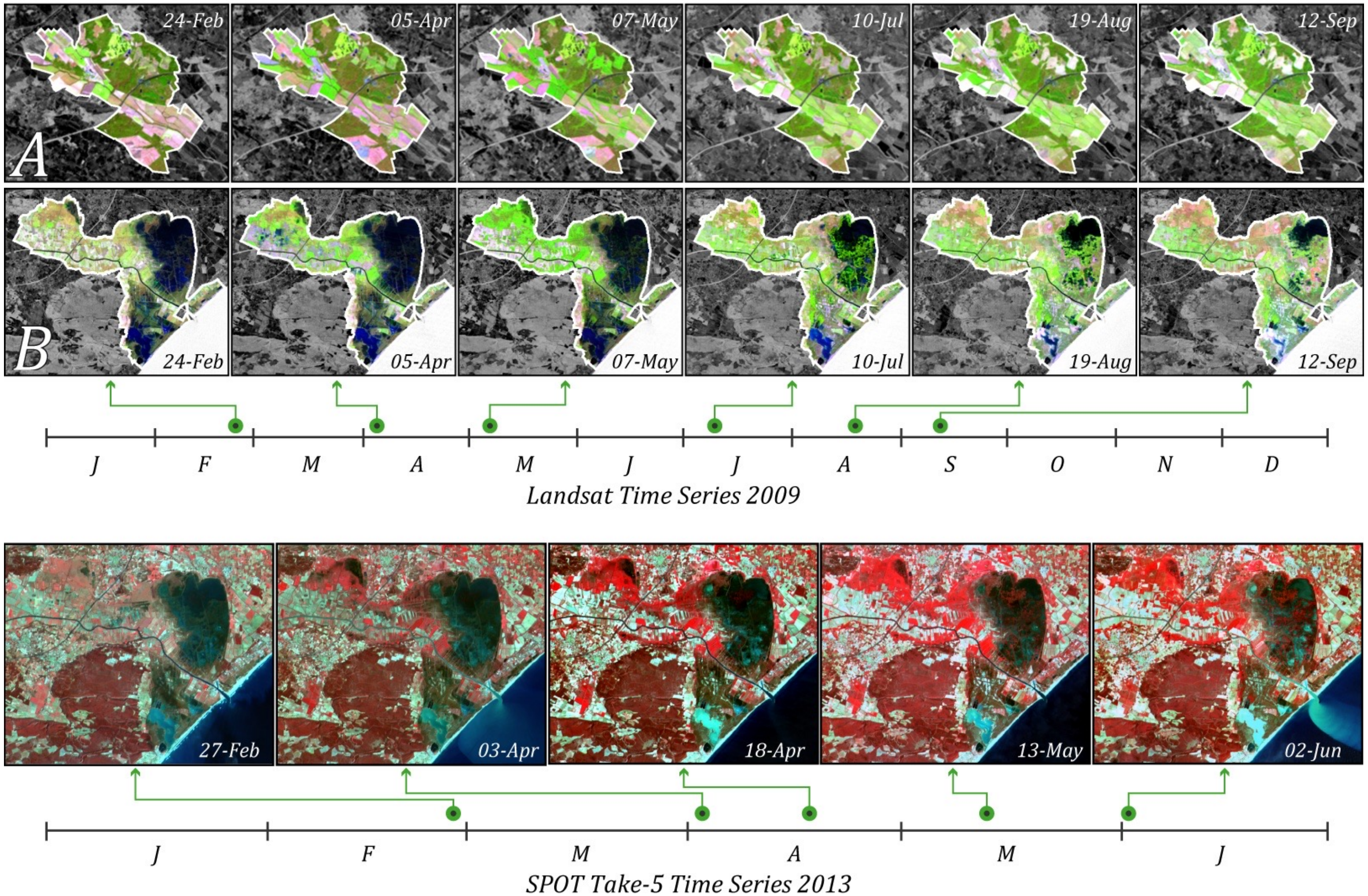
- Automatically detect spatiotemporal evolutions (and their related dynamics)
  - how an entity (i.e. a lake, a saltmarsh area, a crop field) evolves along the time

- Provided useful information for natural habitats monitoring and mapping

A – Libron Valley | B – Low Aude Plane



A – Libron Valley | B – Low Aude Plane

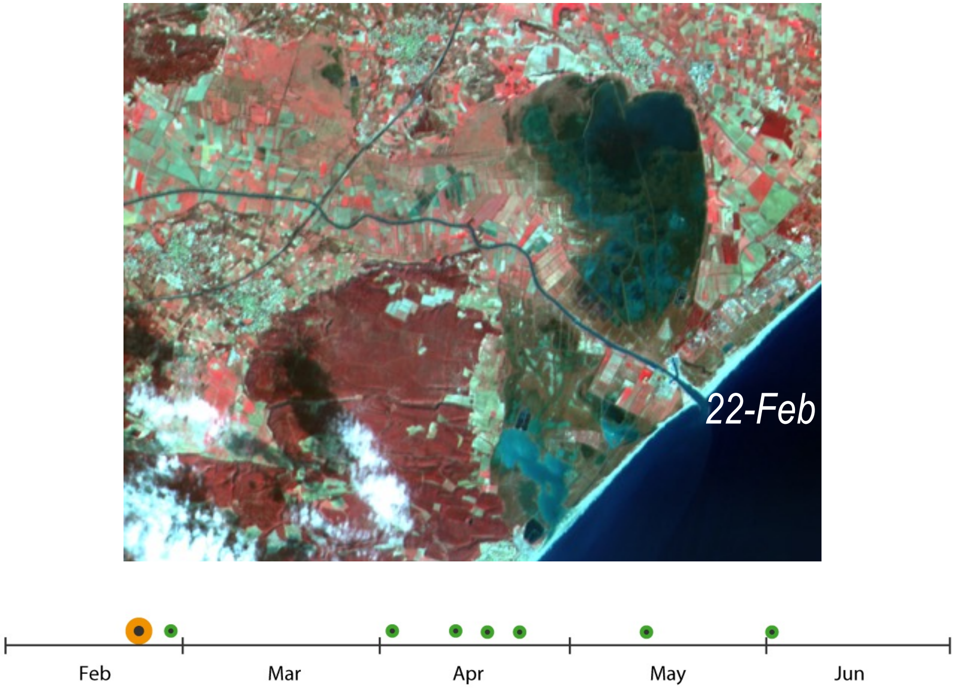


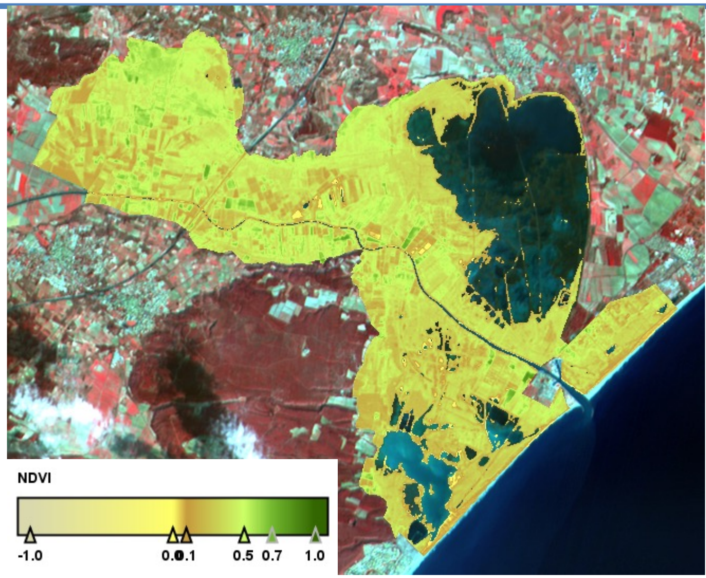
30m

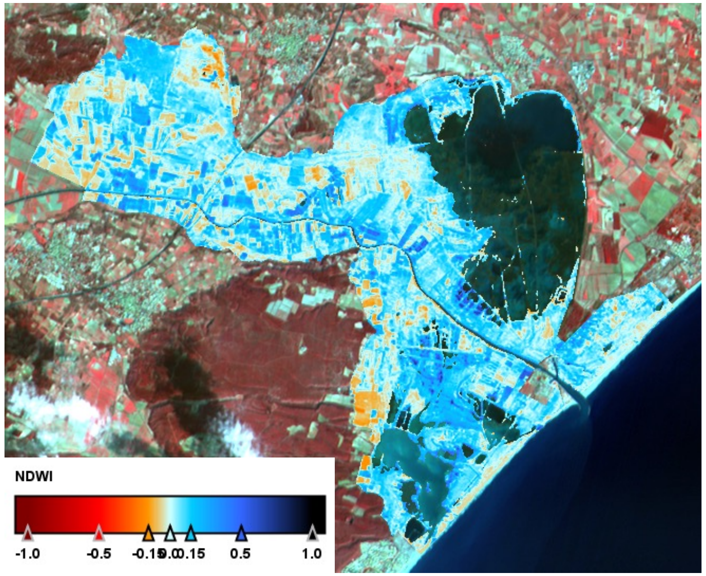
5m

General Evolution of the Vegetation

NDVI: Normalized Difference Vegetation Index
Sensitive to the amount of photosynthetically active vegetation present in the plant canopy
NDWI: Normalized Difference Water Index
sensitive to changes in liquid water content of vegetation canopies

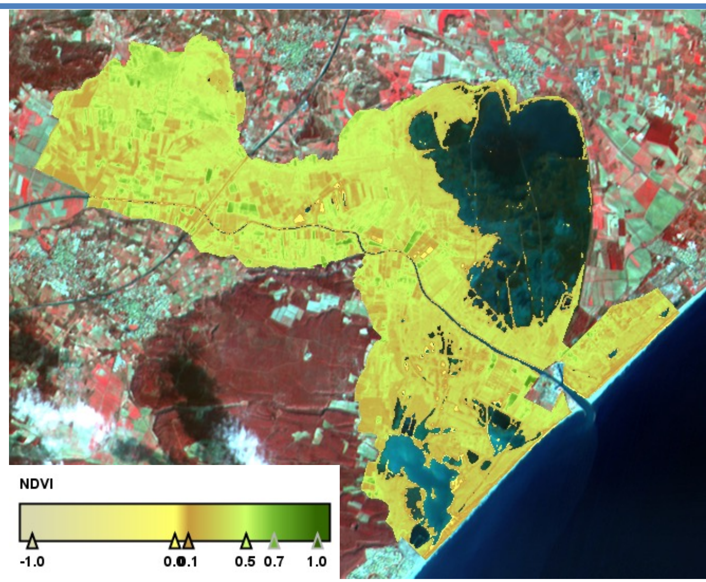


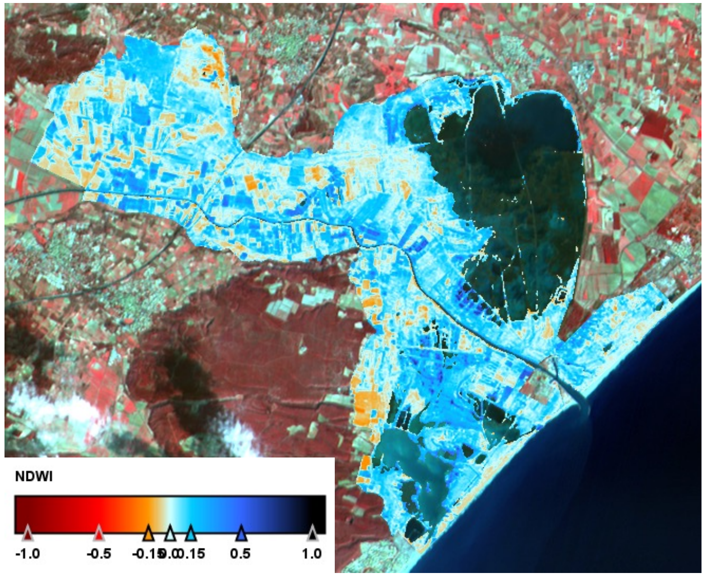




General Evolution of the Vegetation

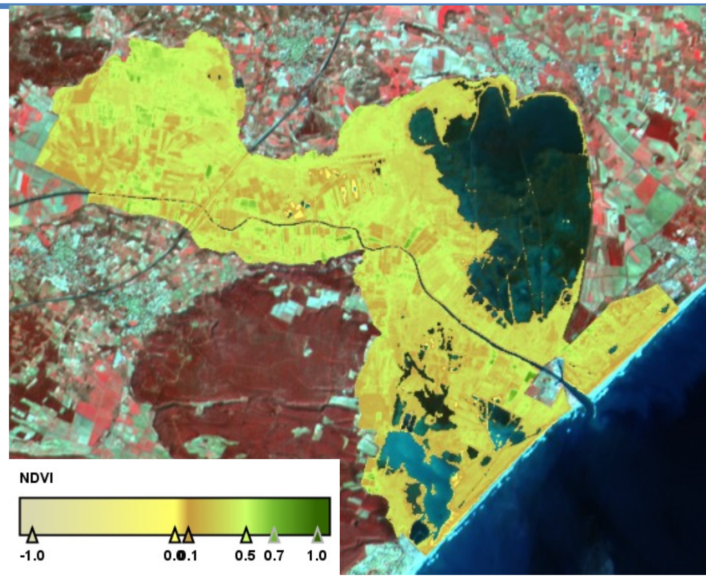
NDVI and NDWI Evolutions

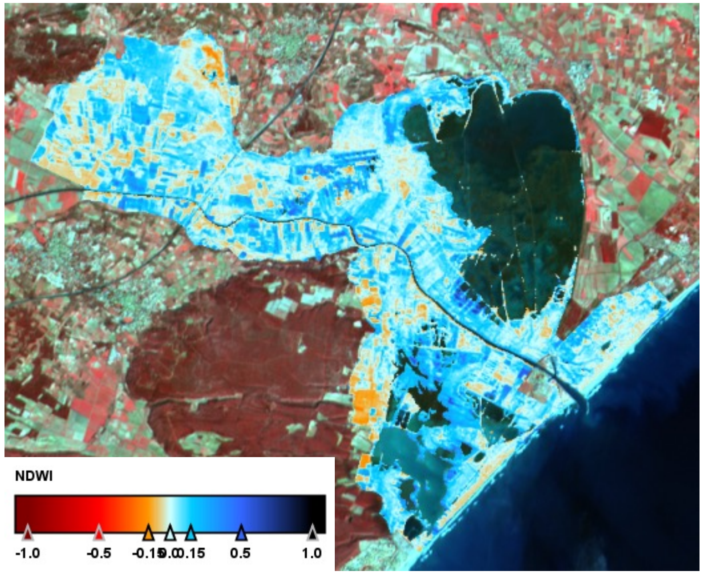




General Evolution of the Vegetation

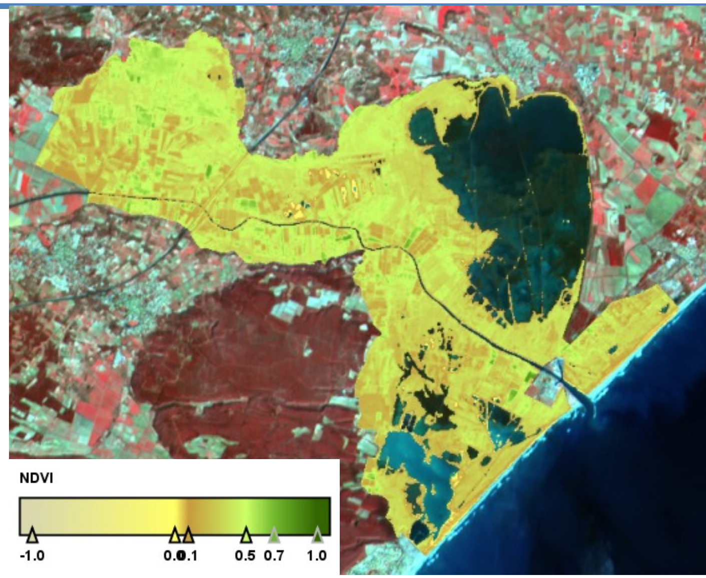
NDVI and NDWI Evolutions

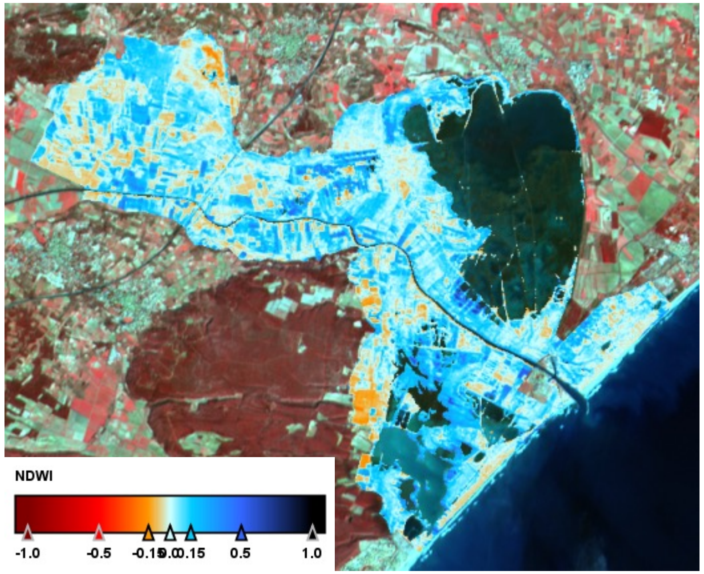




General Evolution of the Vegetation

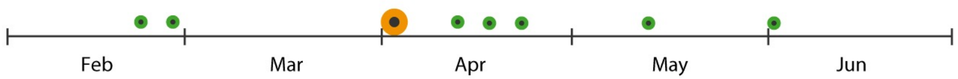
NDVI and NDWI Evolutions

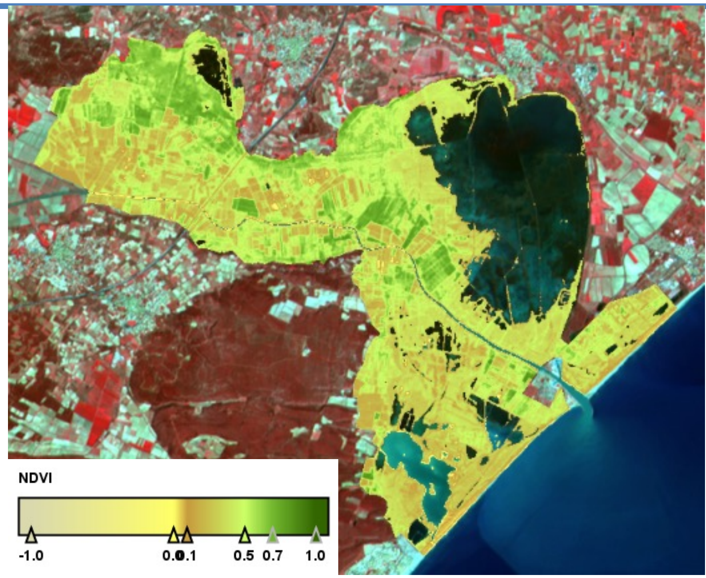


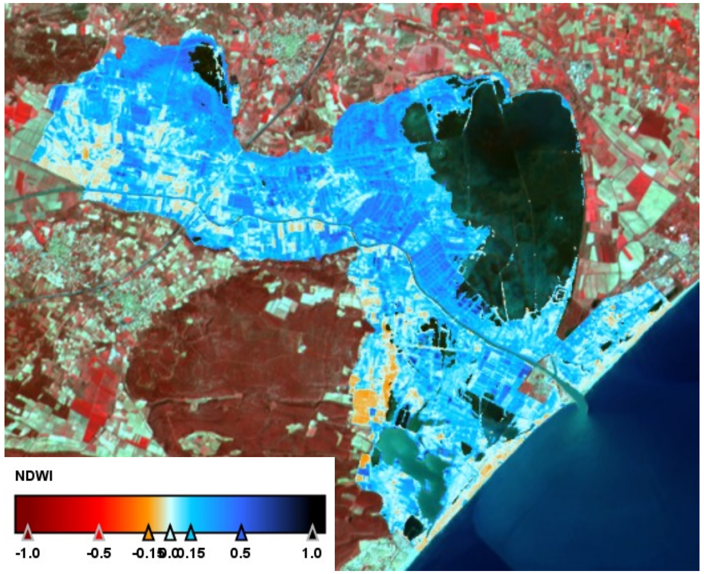


General Evolution of the Vegetation

NDVI and NDWI Evolutions

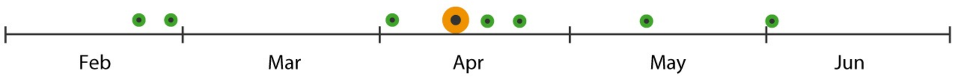


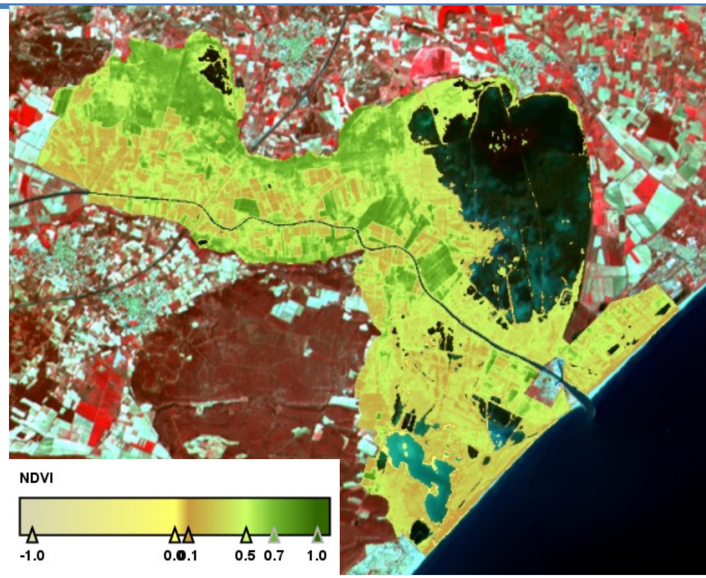


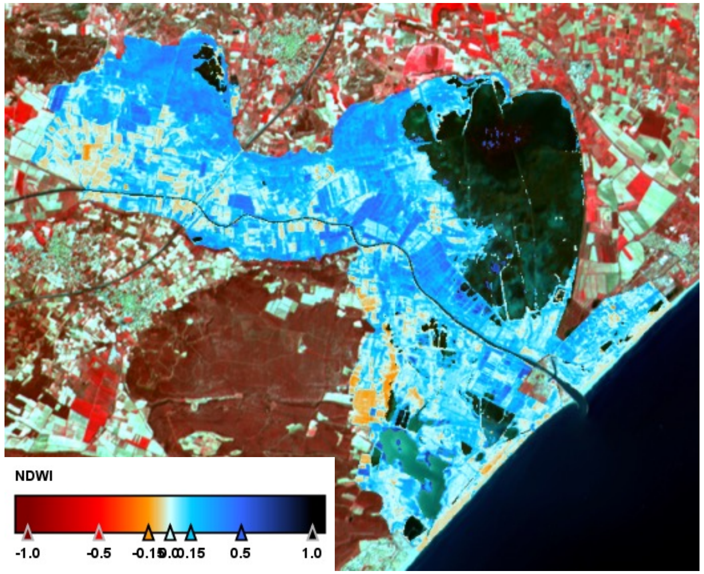


General Evolution of the Vegetation

NDVI and NDWI Evolutions

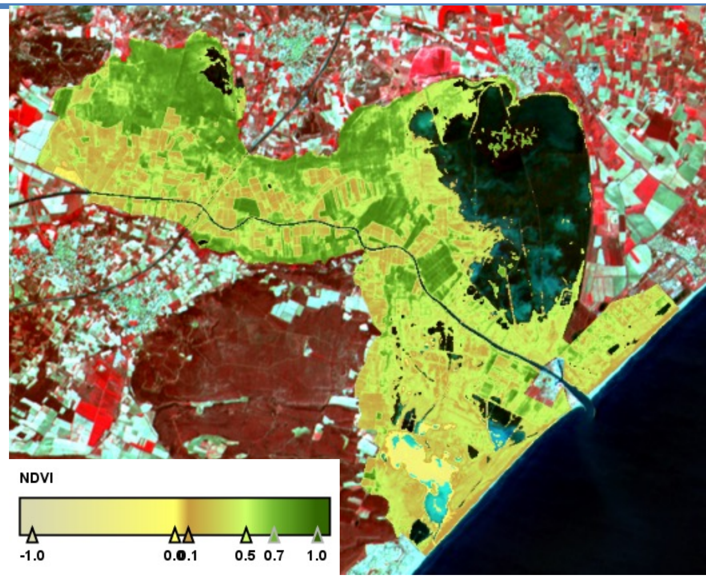


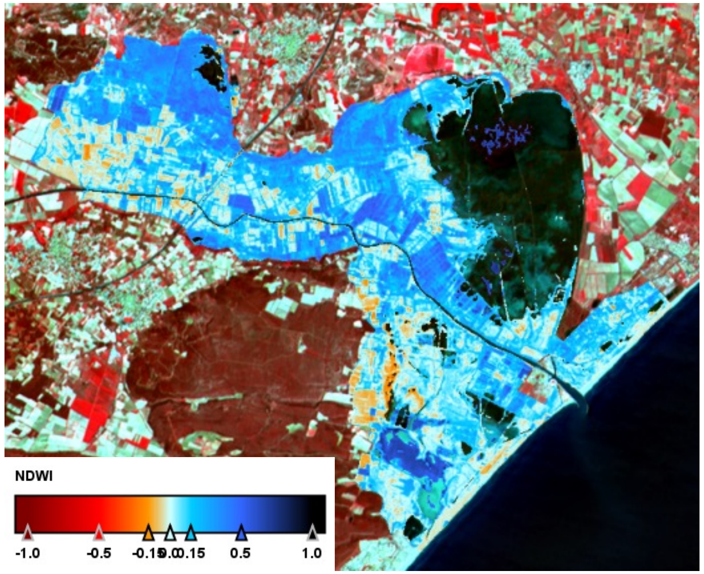




General Evolution of the Vegetation

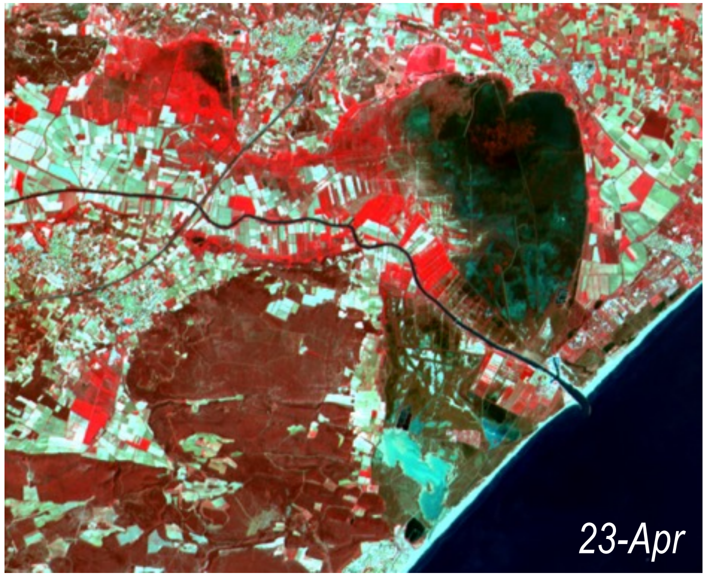
NDVI and NDWI Evolutions

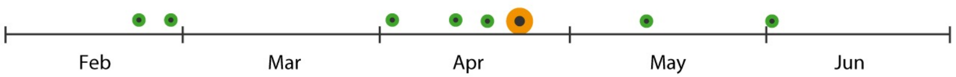


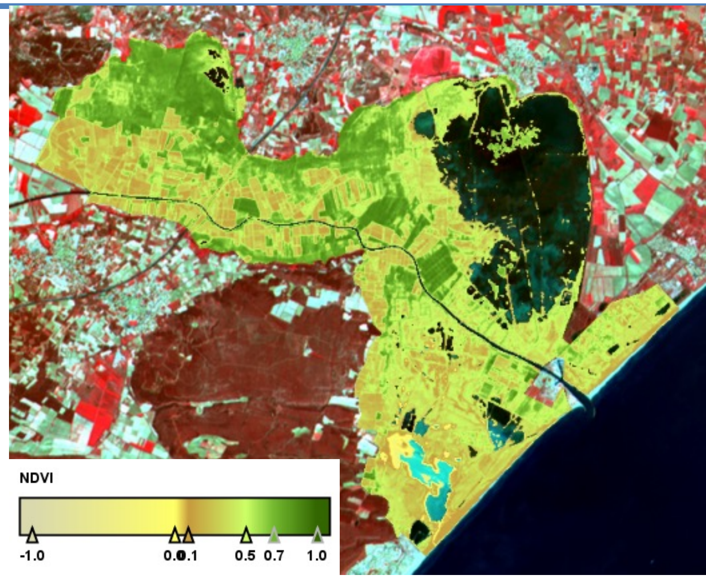


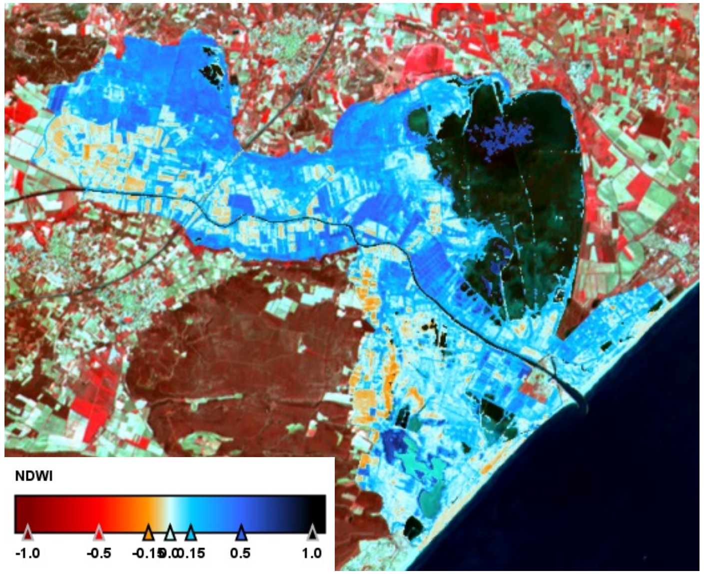
General Evolution of the Vegetation

NDVI and NDWI Evolutions



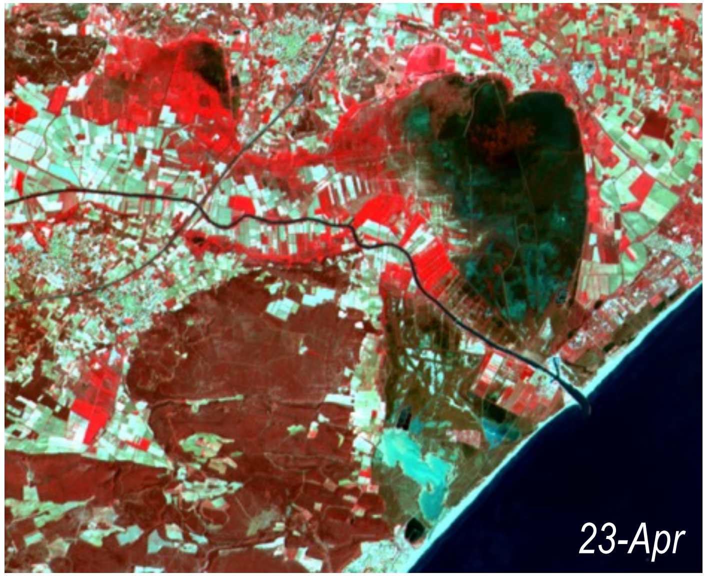


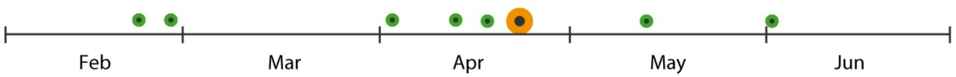


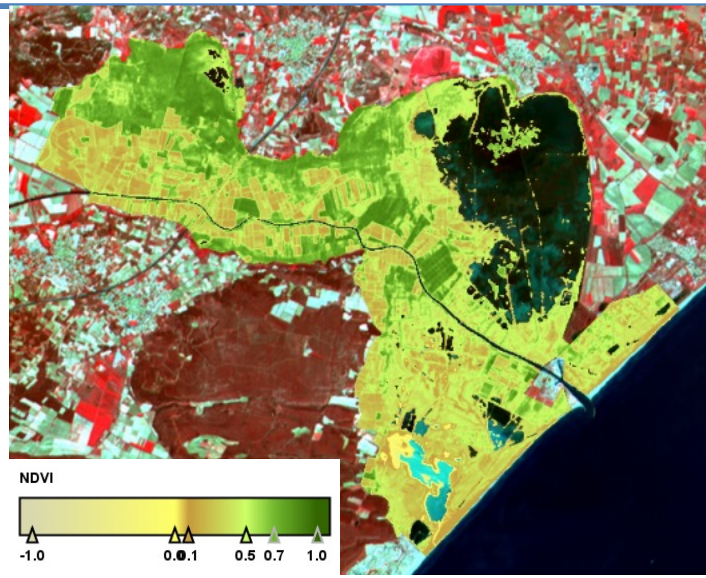


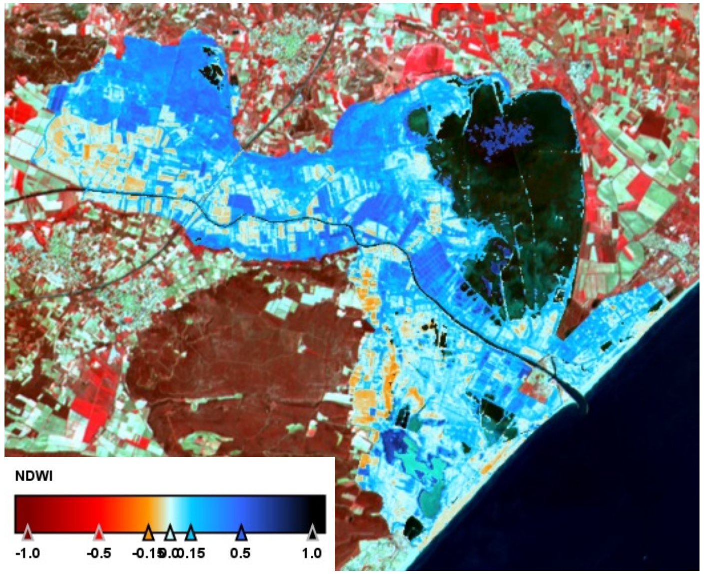
General Evolution of the Vegetation

NDVI and NDWI Evolutions



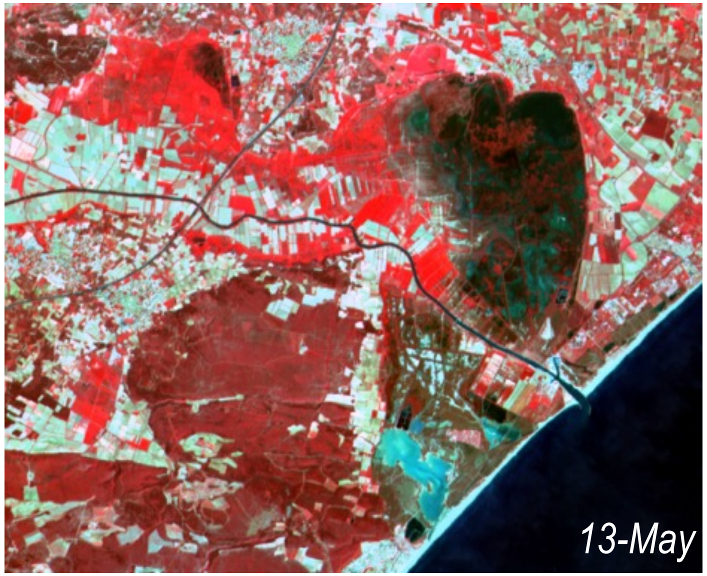


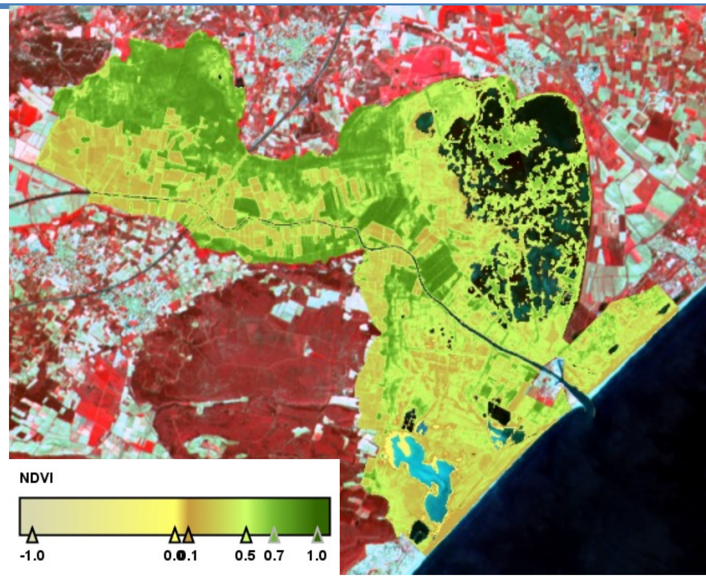


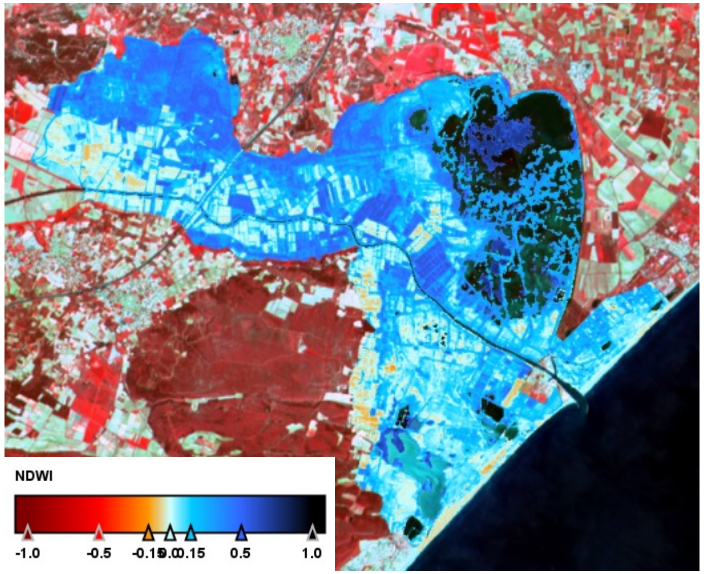


General Evolution of the Vegetation

NDVI and NDWI Evolutions

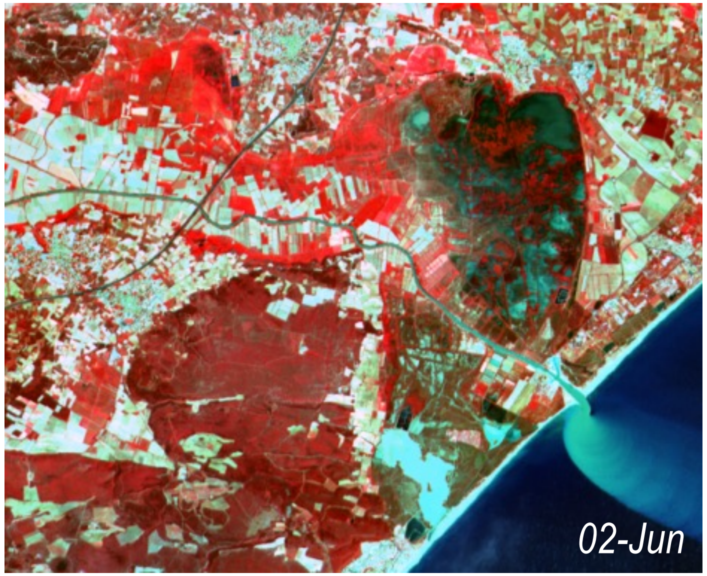


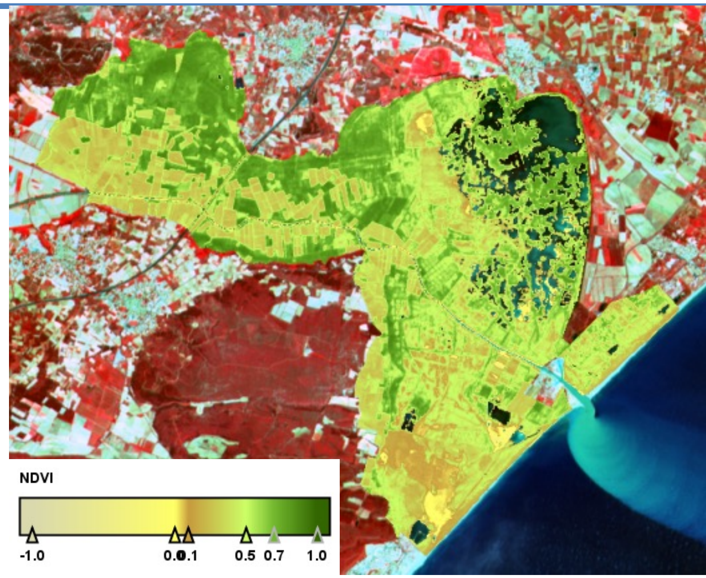


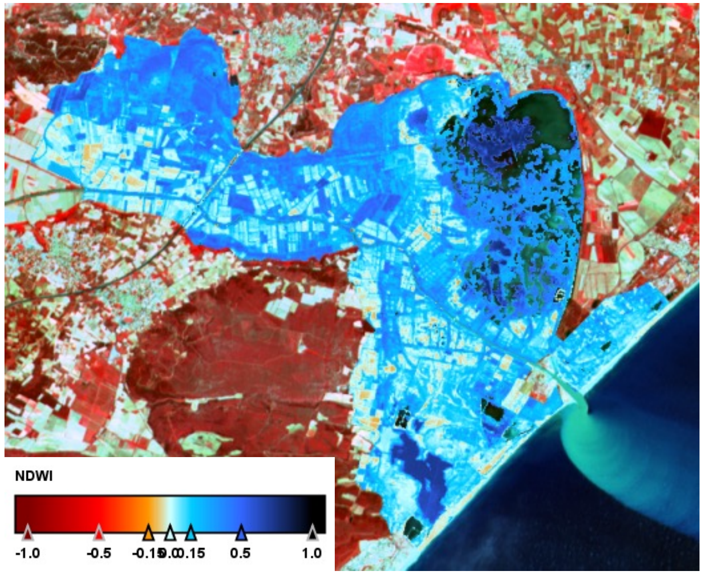


General Evolution of the Vegetation

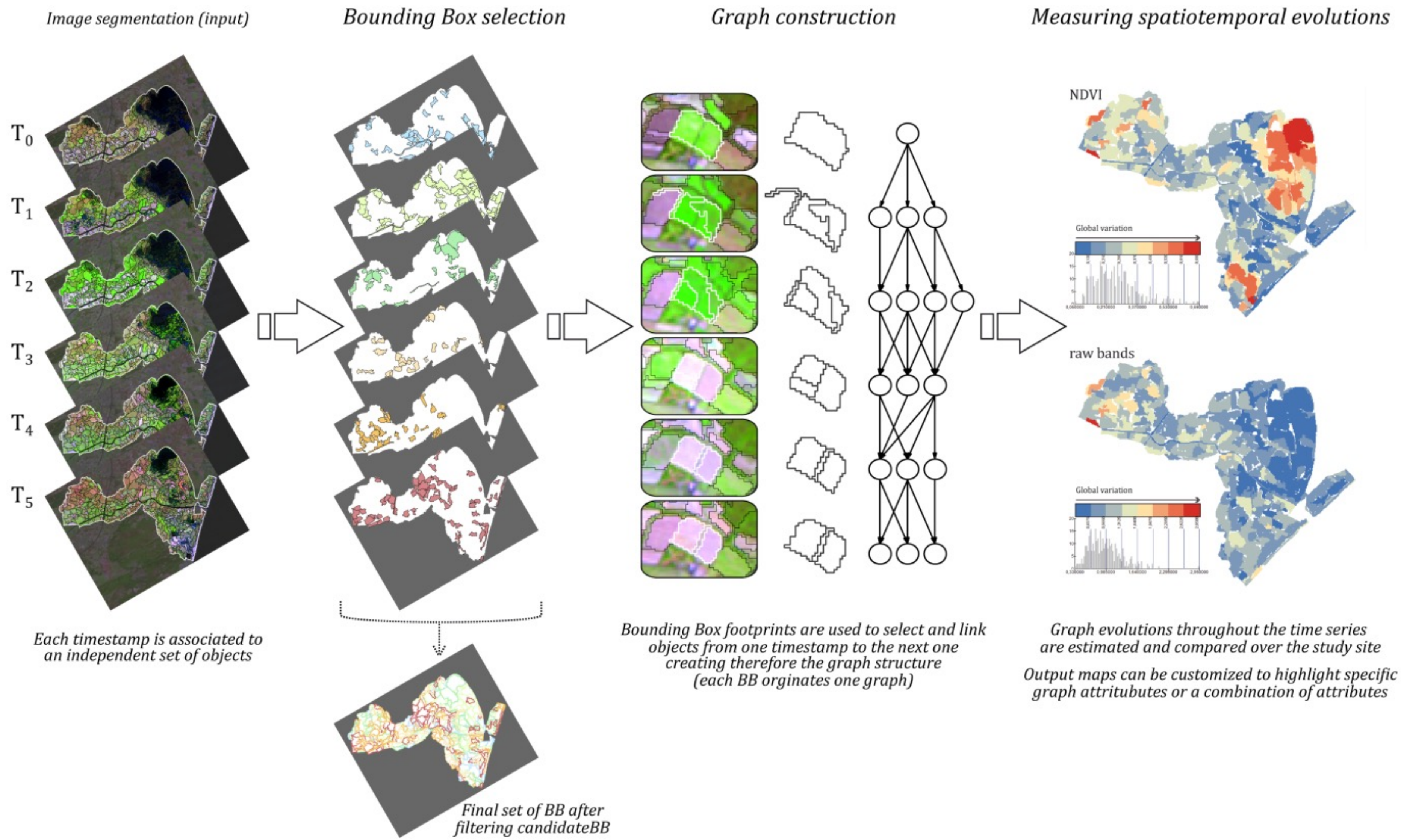
NDVI and NDWI Evolutions







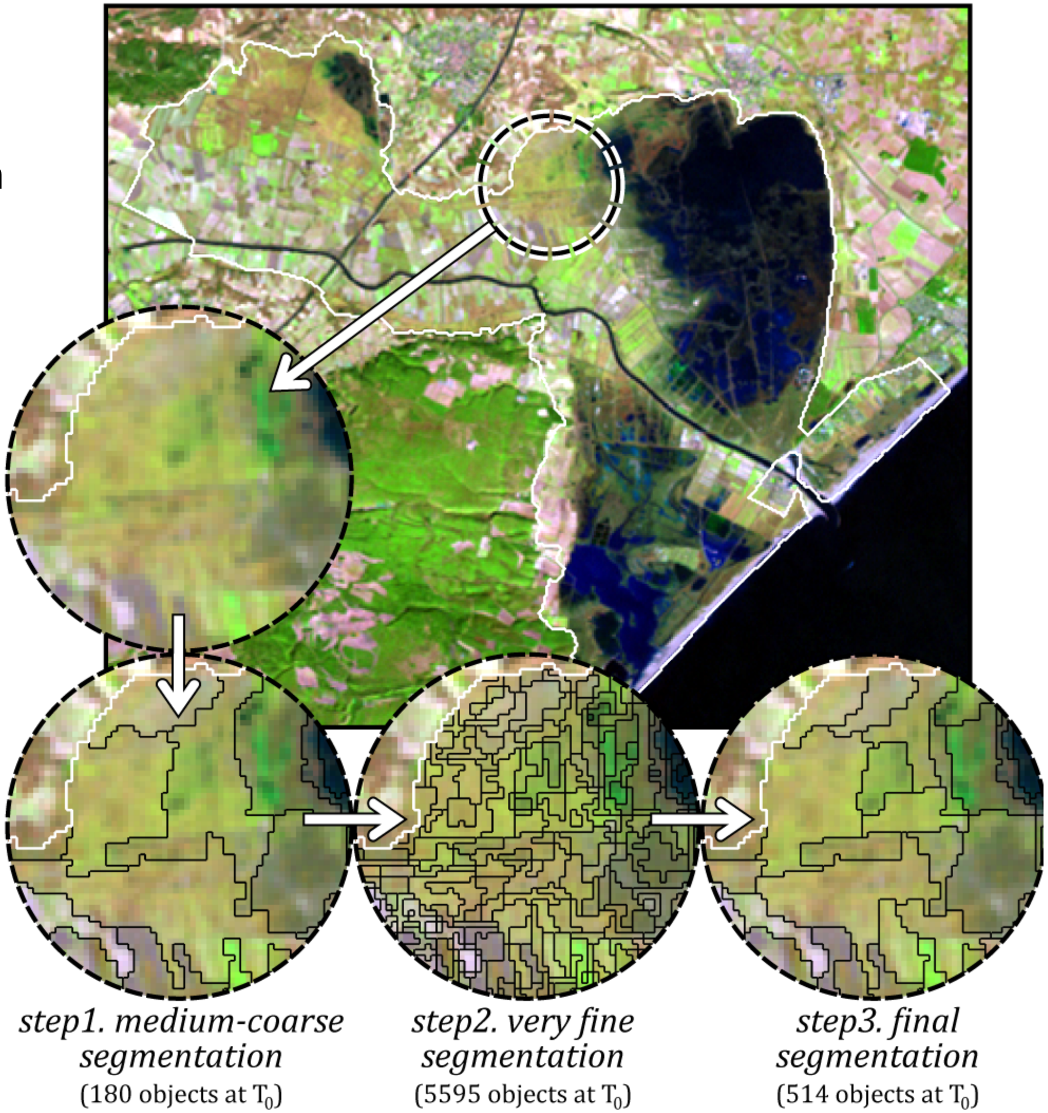
The Main Approach



The Main Steps

Step 1

- *Preprocessing & Segmentation*
  - data standardization
  - fine geometrical registration
  - spatial subset
  - spectral indices
  - …



The Main Steps

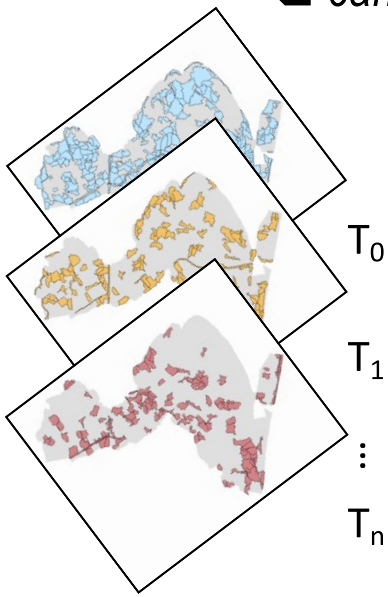
Step 1

- *Preprocessing & Segmentation*

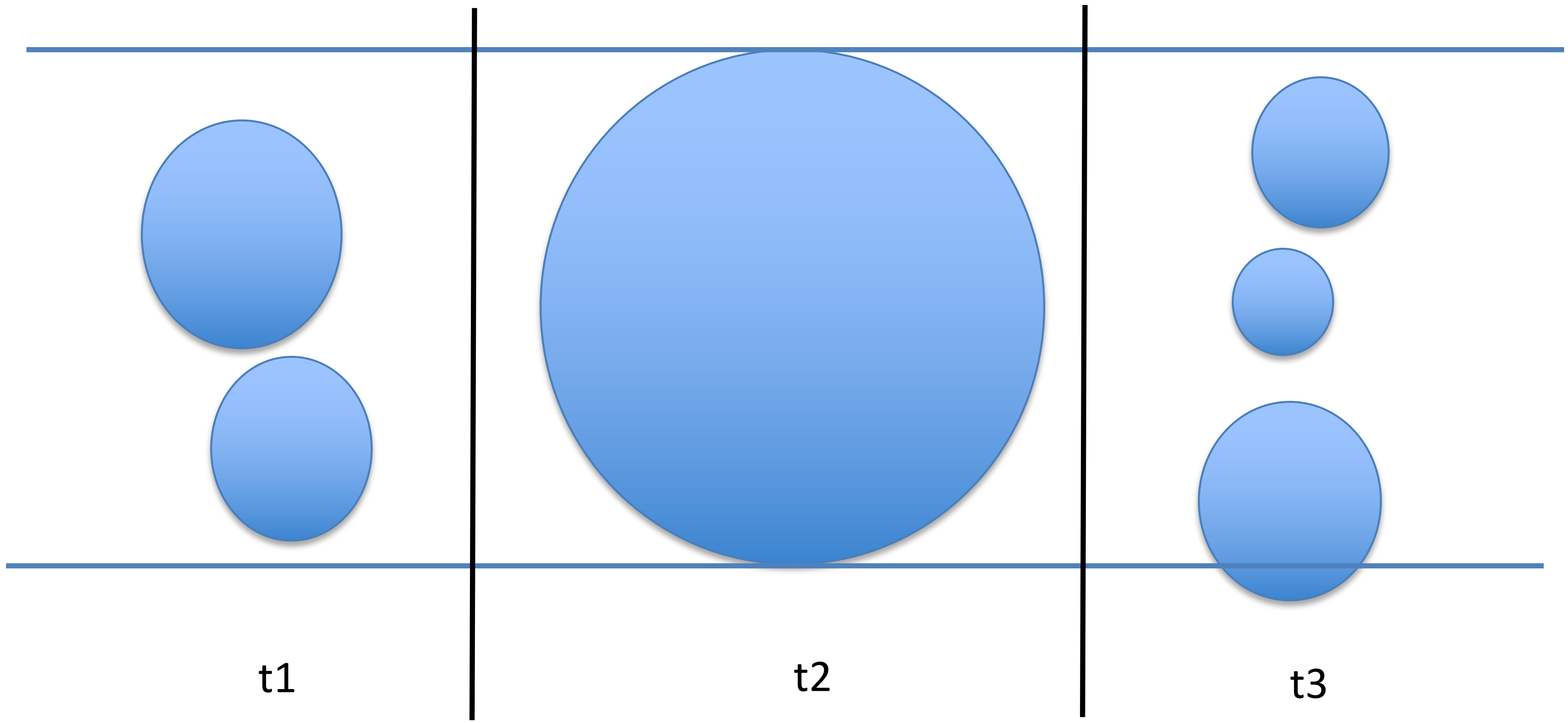
Step 2

- ***BB selection***

- Retrieval of the Maximal Spatial Extents
  - On the whole study area
  - Along the 6 timestamps
  - The result is a list of objects
    ➘ *candidateBB*



Selection of a Bounding Box



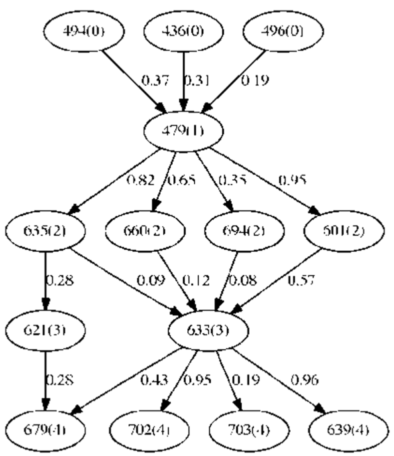
A temporary lake: At t2 it reaches its maximal spatial extent

The Main Steps

Step 1
- *Preprocessing & Segmentation*

Step 2
- *BB selection*

Step 3
- ***Graph construction***



- Each *BB is used to build an evolution graph*
- The objects of each graph are chosen following 2 conditions:
  - objet ∩ *BB* ≥ 25% of the object surface
  - objet ∩ *BB* ≥ 20% of the BB surface

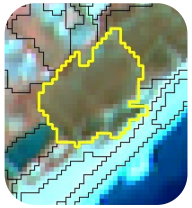
The Main Steps

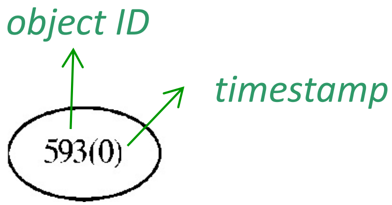
Step 1 • *Preprocessing & Segmentation*

Step 2 • *BB selection*

Step 3 • ***Graph construction***

T_0





T_1

T_2

T_3

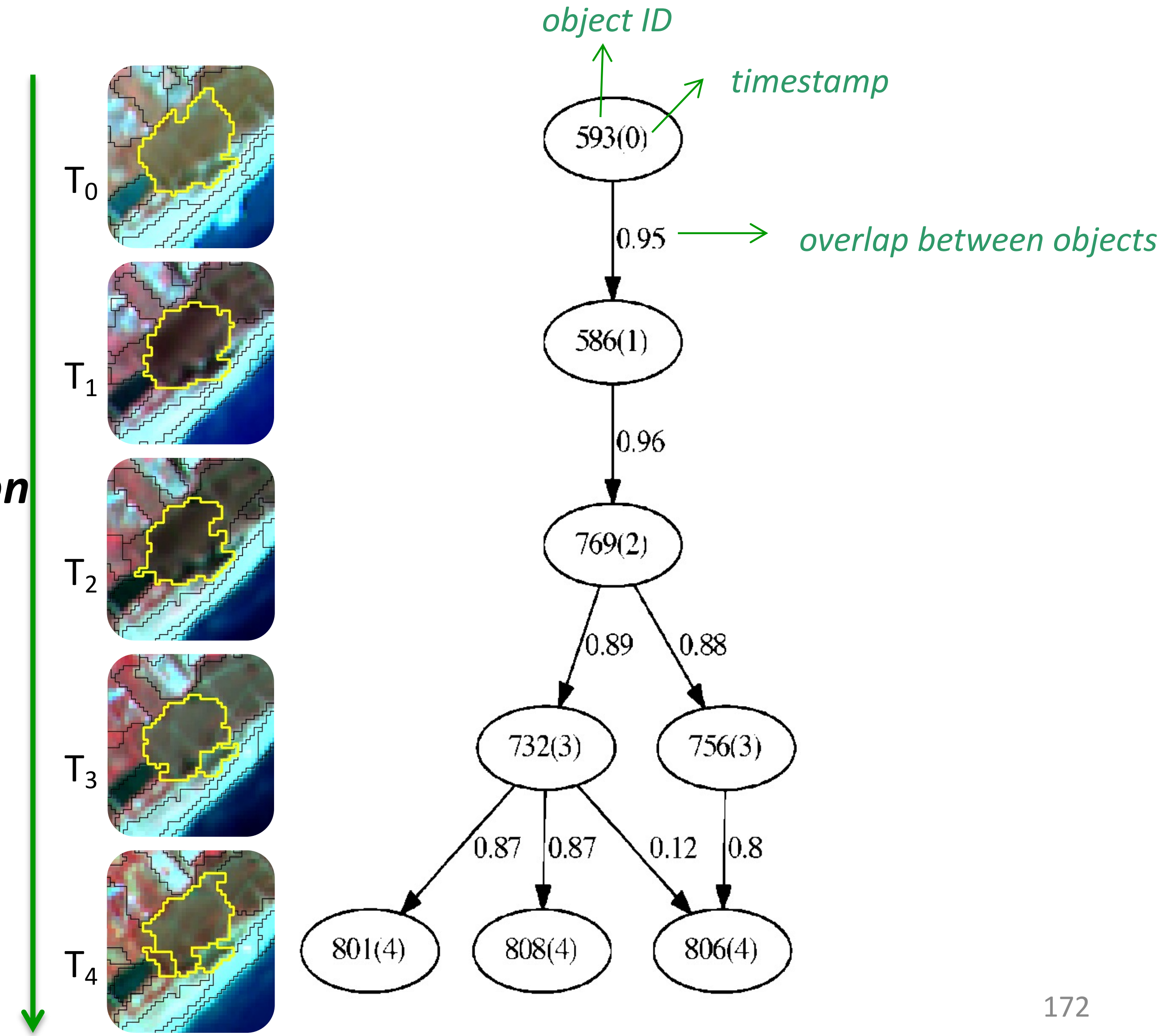
T_4

The Main Steps

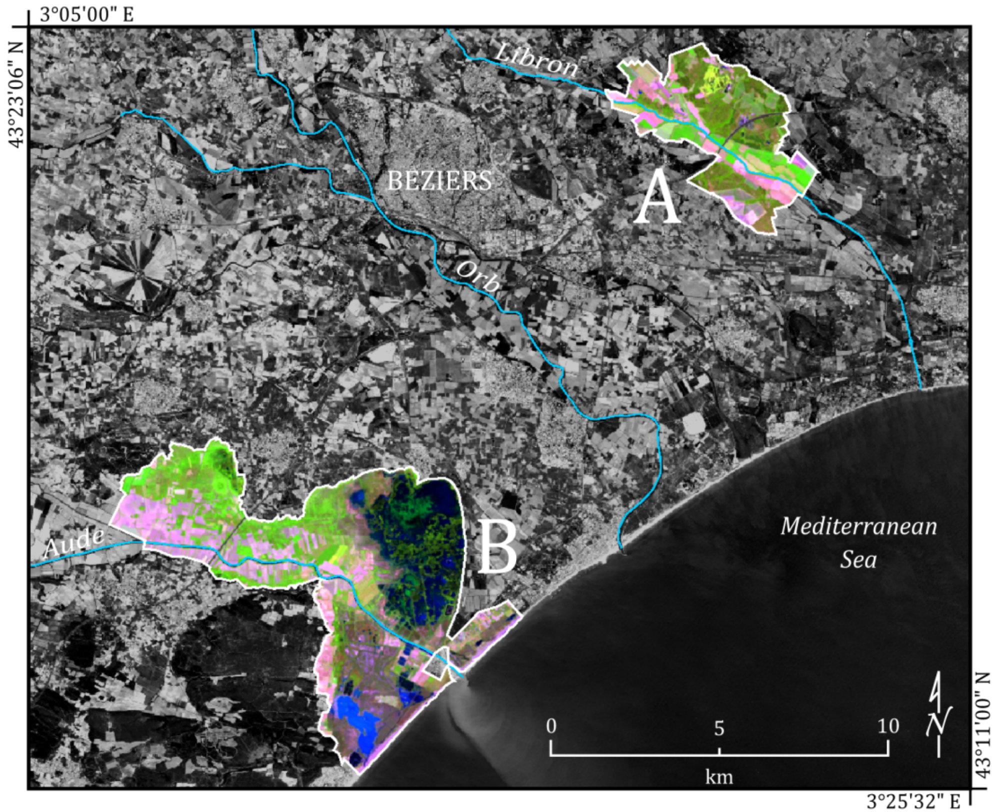
Step 1 — •*Preprocessing & Segmentation*

Step 2 — •*BB selection*

Step 3 — •***Graph construction***



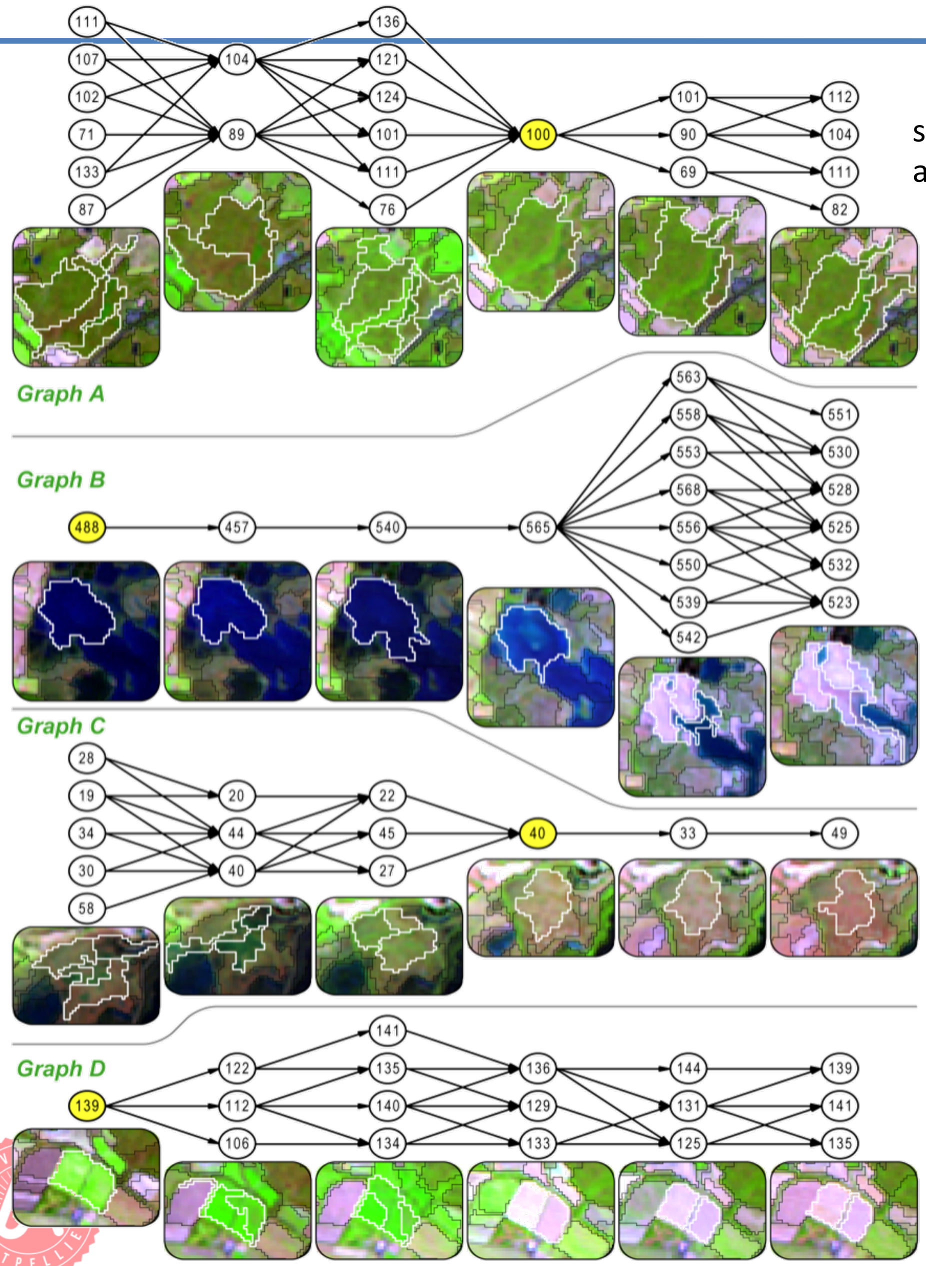
Results



Landsat 2009

- Number of objets
  - Libron = 1218 (~ 203 per timestamps)
  - BPA = 3 373 (~ 562 per timestamps)
- Number of extracted graphs (evolutions)
  - Libron = 142
  - BPA = 340

Evolution Graphs



scrubland and forest

coastal lagoon

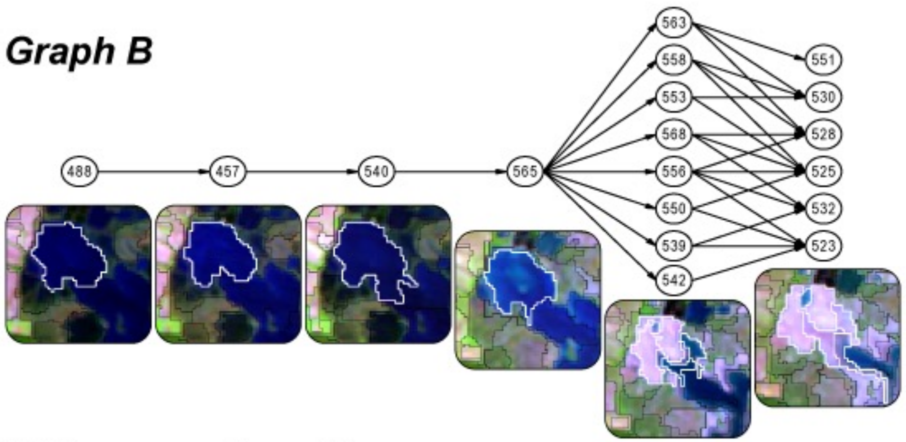
saltmarsh and salt meadow area

Cereal crop fiels

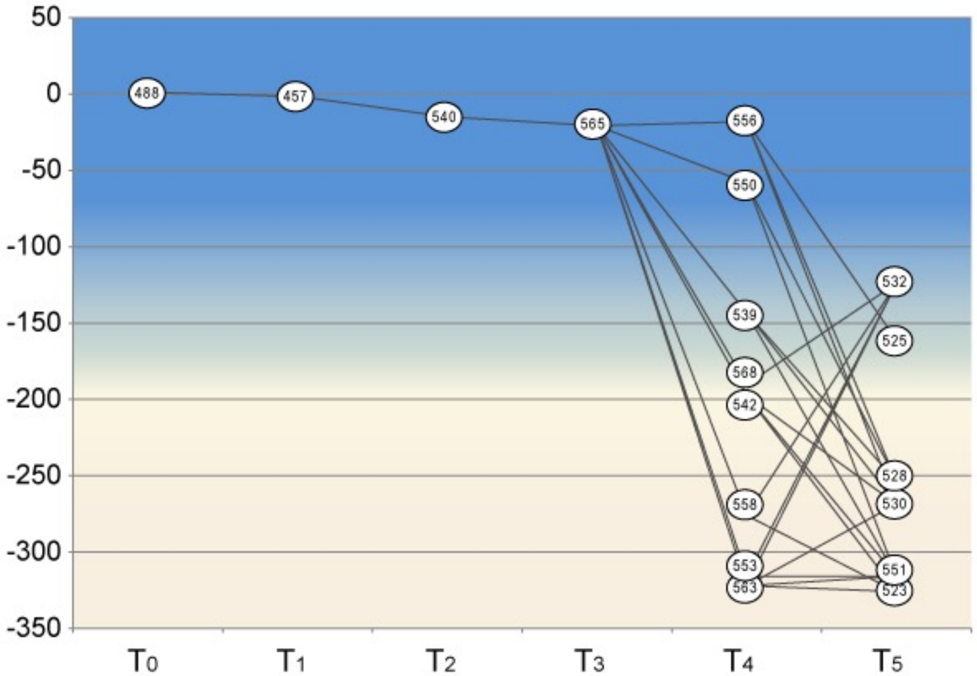
- Each node contains a set of attributes
  - Red
  - Green
  - …
  - NDVI
  - NDWI
  - VSDI (index conceived to monitor drought)
  - …
- The graphs are also described by statistics
  - # of nodes
  - # of edges
  - # of sequences
  - …

Evolution Graphs

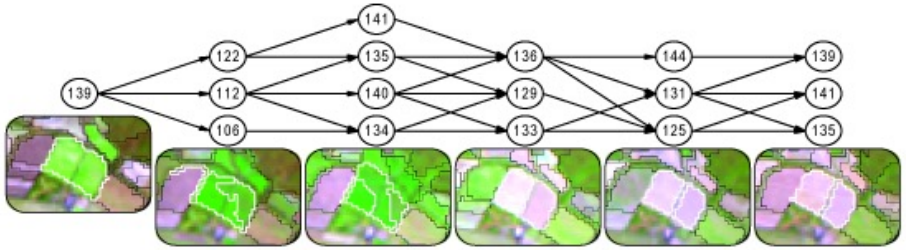
Graph B



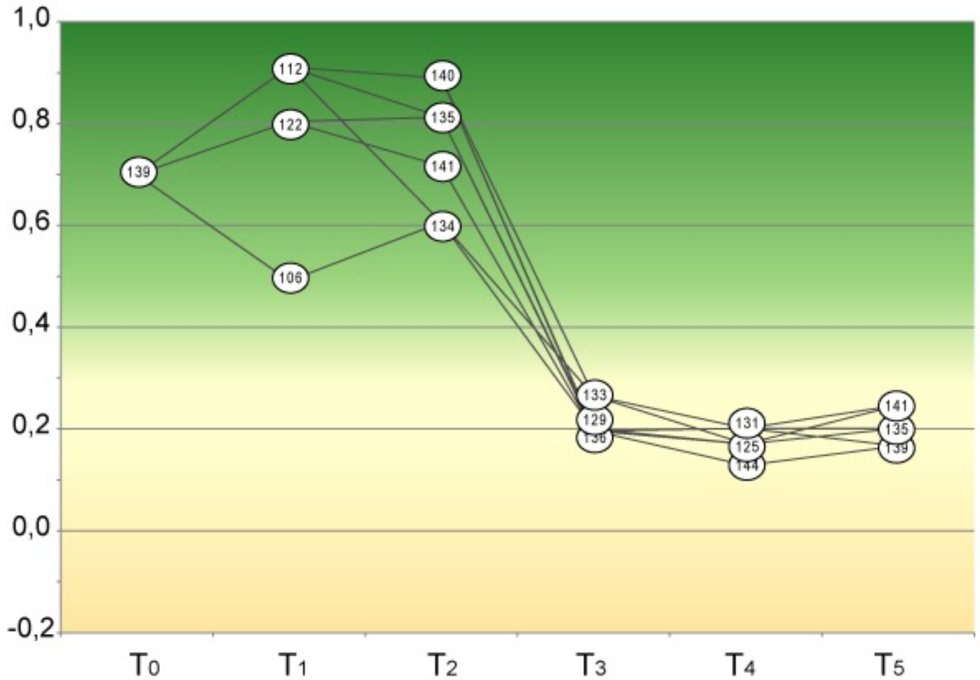
VSDI temporal profile



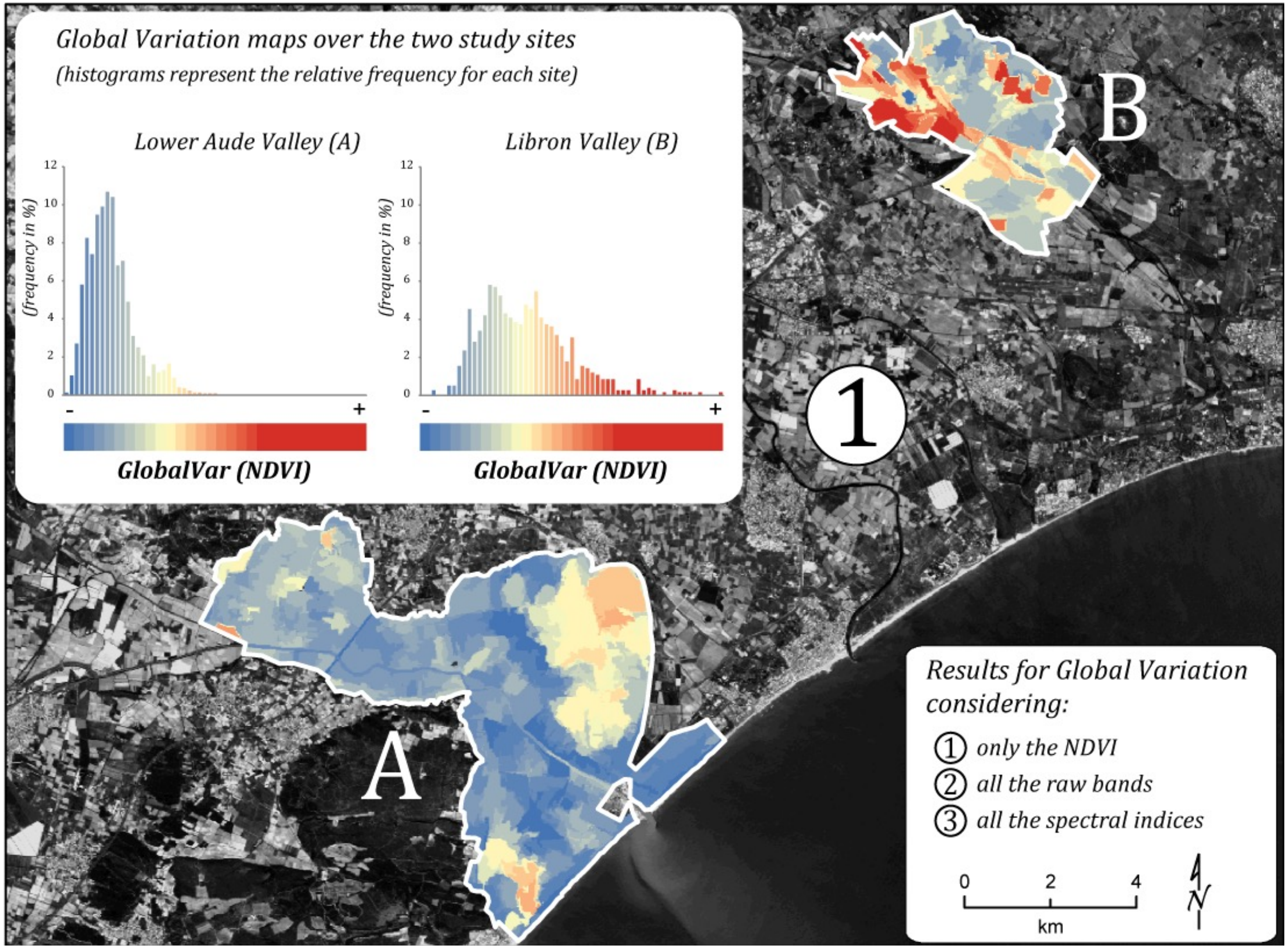
Graph D



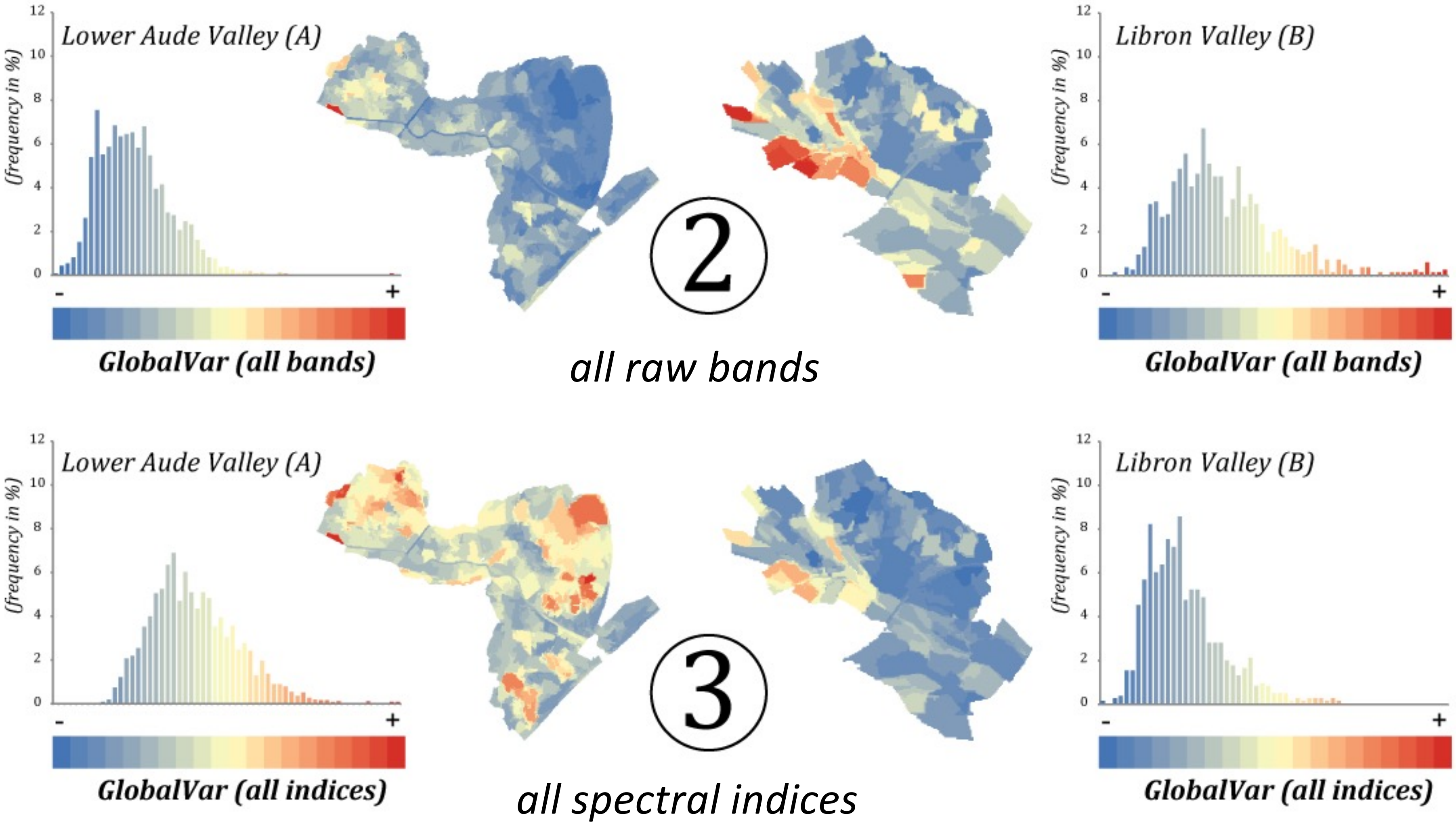
NDVI temporal profile



Evolution Intensity



Evolution Intensity



Similar Behaviors

- Clustering approaches to find the region having the same behaviors
- At every timestamp in the graph, the set of attributes is available
- Considering the number of objects that may appear in a graph

Outline

- A quick reminder on pattern discovery
- The different kinds of patterns from trajectory
- Towards a unified approach for extracting trajectories
- Deal with and without a spatial component?
- **A concrete illustration**
- Conclusion

Now let's have a concrete example

jupyter TrajectoriesNoteBook Last Checkpoint: il y a 4 heures (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run Markdown

This notebook is an example of applying getmove on real trajectories.

Installation [...]

Running trajectories

In [ ]:
```
import matplotlib.pyplot as plt
import matplotlib.cm as cm
from mpl_toolkits.basemap import Basemap
import numpy as np
import datetime
from sklearn.cluster import DBSCAN
import pandas as pd
from matplotlib import interactive
import json
import os
from pprint import pprint
```

Outline

- A quick reminder on pattern discovery
- The different kinds of patterns from trajectory
- Towards a unified approach for extracting trajectories
- Deal with and without a spatial component?
- A concrete illustration
- **Conclusion**

Conclusions

- Trajectories are useful patterns for …lots of applications
- They propose new kinds of patterns even when the spatial component is not present
- They have been really well studied and many approaches exist
- There are more and more applications

Conclusions

- Considering the Pattern Mining issue:
  - Incremental (Objects)
  - Stream Mining
  - New measures for selecting the most interesting patterns
  - Outlier detection
  - Privacy
  - Classification/Clustering of trajectories
  - Interactive Mining
  - Considering the clustering during the process rather than before
  - …

Conclusions

- Big Data Issues:
  - Volume
  - Variety
  - Velocity
  - Veracity

Special Thanks …

- Jérôme Azé, Nicolas Bechet, Elnaz Bigdeli, Flavien Bouillot, Sandra Bringay, Fati Chen, Tiago Cunha, Jacques Fize, Fabio Gunter, Dino Ienco, Diana Inkpen, Stan Matwin, Jordi Nin, Phan Nhat-Hai, Jessica Pinaire, Yoann Pitarch, Tomy Rodrigues, Mathieu Roche, Arnaud Sallabery, Carlos Soares, Maguelonne Teisseire, …
- And many others ….

Some Related Publications

- N. Phan, P. Poncelet and M. Teisseire. "*All in One: Mining Multiple Movement Patterns*". International Journal of Information Technology and Decision Making (IJITDM), June 2016, pp. 1-42
- P. N. Hai, Pascal Poncelet, Maguelonne Teisseire. *"An Efficient Spatio-Temporal Mining Approach to Really Know Who Travels with Whom!".* ISI special issue, selected, 2013
- P. N. Hai, Dino Ienco, Pascal Poncelet, Maguelonne Teisseire. *"Mining Representative Movement Patterns through Compression".* PAKDD 2013.
- A.Z.E. Aabidine, A. Sallaberry, S. Bringay, M. Fabregue, C. Lecellier, P. N. Hai, P. Poncelet. *"Co^{2}Vis: A Visual Analytics Tool for Mining Co-expressed And Co-regulated Genes Implied in HIV Infections"*. IEEE BioVis 2013.
- P. N. Hai, Dino Ienco, Pascal Poncelet, Maguelonne Teisseire. *"Mining Fuzzy Moving Object Clusters".* ADMA 2012.
- P. N. Hai, Dino Ienco, Pascal Poncelet, Maguelonne Teisseire. *"Mining Time Relaxed Gradual Moving Object Clusters".* ACM GIS 2012.
- F. Bouillot, P. N. Hai, N. Béchet, S. Bringay, D. Ienco, S. Matwin, P. Poncelet, M. Roche, and M. Teisseire. *"How to Extract Relevant Knowledge from Tweets?".* ISIP 2012.
- P. N. Hai, Pascal Poncelet, Maguelonne Teisseire. *"GET_MOVE: An Efficient and Unifying Spatio-Temporal Pattern Mining Algorithm for Moving Objects".* IDA 2012.
- P. N. Hai, Pascal Poncelet, Maguelonne Teisseire. *"An Efficient Spatio-Temporal Mining Approach to Really Know Who Travels with Whom!".* BDA 2012.
- P. N. Hai, Dino Ienco, Pascal Poncelet,Maguelonne Teisseire. *"Extracting Trajectories through an Efficient and Unifying Spatio-Temporal Patten Mining System".* ECML-PKDD 2012.
- P. N. Hai, Pascal Poncelet, Maguelonne Teisseire. *"MovingObjects: Combining Gradual Rules and Spatio-Temporal Patterns".* IEEE ICSDM 2011.

All the papers are available on my web page

Bibliography

- M. Andersson, J. Gudmundsson, P., Laube, T. Wolle. "Reporting leadership patterns among trajectories". In: SAC, 2007.
- M. Andersson, J. Gudmundsson, P., Laube, T. Wolle. "Reporting leaders and followers among trajectories of moving point objects". GeoInformatica 12(4), 2008.
- M. Benkert, J. Gudmundsson, F. Hbner, T. Wolle. "Reporting flock patterns". Computational Geometry 41(1), 2008.
- J. Damascène Mazimpaka and S. Timpf. "Trajectory data mining: A review of methods and applications". In Journal of Spatial Information Science, N° 13, pp. 61–99, 2016.
- Z. Feng and Y. Zhu. "A Survey on Trajectory Data Mining: Techniques and Applications.", IEEE Access 2016.
- F. Giannotti, M. Nanni, F. Pinelli, D. Pedreschi. "Trajectory pattern mining" Proceedings of SIGKDD 2007.
- F. Giannotti, M. Nanni, and D. Pedreschi. "Efficient Mining of Sequences with Temporal Annotations". In Proc. SIAM Conference on Data Mining, 2006.
- J. Gudmundsson, M. van Kreveld, B. Speckmann. "Efficient detection of motion patterns in spatio-temporal data sets" In: Proceedings of the ACM international symposium on Advances in geographic information systems, 2004.
- J. Gudmundsson and M. V. Kreveld. "Computing longest duration flocks in trajectory data". In ACM GIS, pages 35–42, 2006.
- J. Han, Z. Li, and L. A. Tang. "Mining moving object, trajectory and traffic data". In Database Systems for Advanced Applications, volume 5982 of Lecture Notes in Computer Science, pages 485–486. 2010.
- H. Jeung, M. L. Yiu, X. Zhou, C. S. Jensen, and H. T. Shen. "Discovery of convoys in trajectory databases". *Proc. VLDB Endow.*, 1(1):1068–1080, August 2008.

Bibliography

- P. Kalnis, N. Mamoulis, and S. Bakiras. "On discovering moving clusters in spatio-temporal data". In *SSTD ’05*, pages 364–381, 2005
- P. Laube, S. Imfeld. “Analyzing relative motion within groups of trackable moving point objects". In GIScience, 2002.
- P. Laube, S. Imfeld, R. Weibel. "Discovering relative motion patterns in groups of moving point objects". International Journal of Geographical Information Science 19(6), 2005.
- P. Laube, M. van Kreveld, S., Imfeld. “Finding remo - detecting relative motion patterns in geospatial lifelines. In: Proceedings of the International Symposium on Spatial Data Handling, 2004.
- P. Laube, R.S. Purves. "An approach to evaluating motion pattern detection techniques in spatio-temporal data". Computers, Environment and Urban Systems 30(3), 2006.
- Z. Li, B. Ding, J. Han, and R. Kays. Swarm: mining relaxed temporal moving object clusters. *Proc. VLDB Endow.*, 3(1-2):723–734, 2010.
- L. A. Tang, Y. Zheng, J. Yuan, J. Han, A. Leung, C.-C. Hung, and W.-C. Peng. On discovery of traveling companions from streaming trajectories. In *ICDE*, pages 186–197, 2012.
- M.R. Vieira, P. Bakalov, V.J. Tsotras. "On-line discovery of flock patterns in spatio-temporal data". In Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2009.
- Y. Wang, E. P. Lim, and S. Y. Hwang. "Efficient mining of group patterns from user movement data". *Data Knowl. Eng.*, 57(3):240–282, June 2006.
- K. Zheng, Y. Zheng, N.J.Yuan, S. Shang, X. Zhou "Online discovery of gathering patterns over trajectories". IEEE Trans Knowl Discov Data Eng 26:1974–1988, 2014
- Y. Zheng. "Trajectory Data Mining: An Overview". ACM Transactions on Intelligent Systems and Technology (TIST), 6(3), 2015.

- Questions ?