Structuring Typical Evolutions using Temporal-Driven Constrained Clustering

Research Team Reunion

12 February 2013

Marian-Andrei Rizoiu

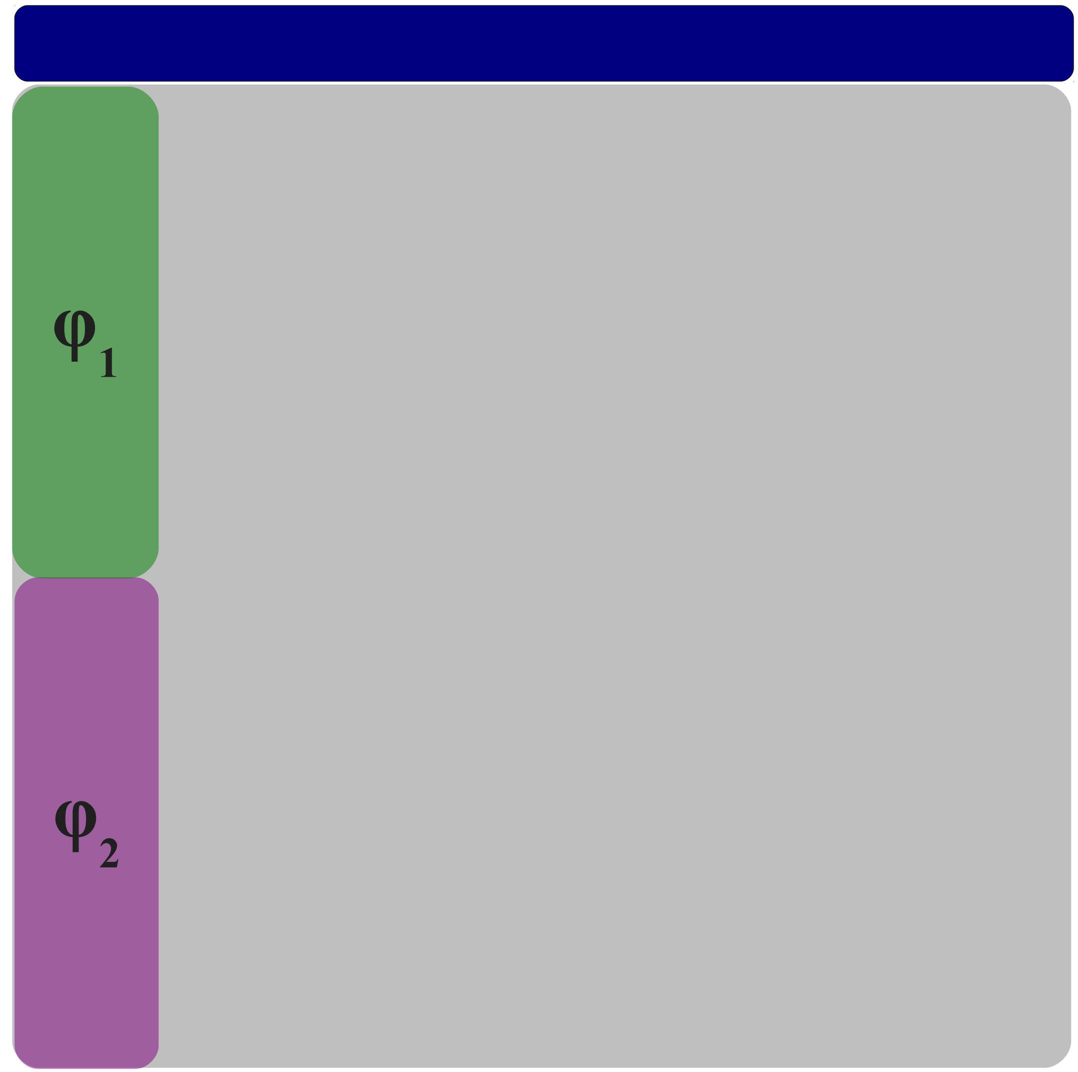
ERIC Laboratory
Université Lumière Lyon 2
France

Dataset: the values for a certain number of numerical features (x^d) for multiple entities (φ) at different moments of time (t)

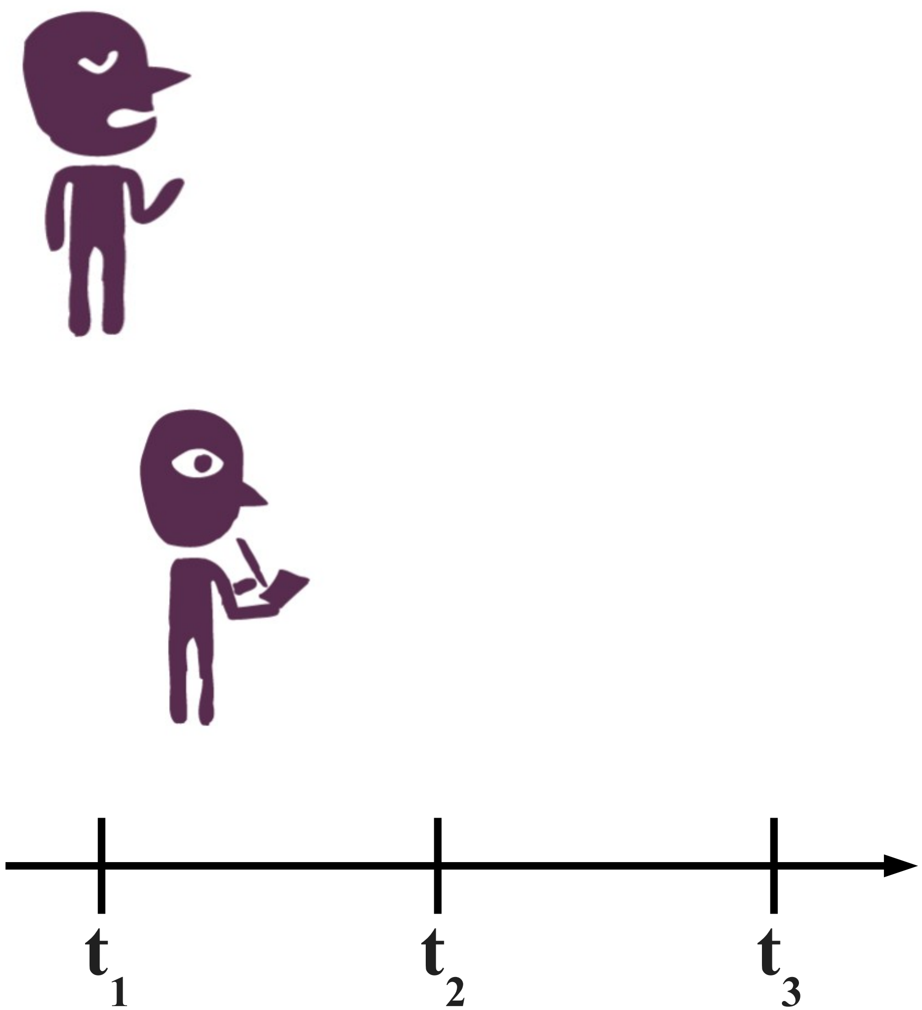
Dataset: the values for a certain number of numerical features (x^d) for multiple entities (φ) at different moments of time (t)

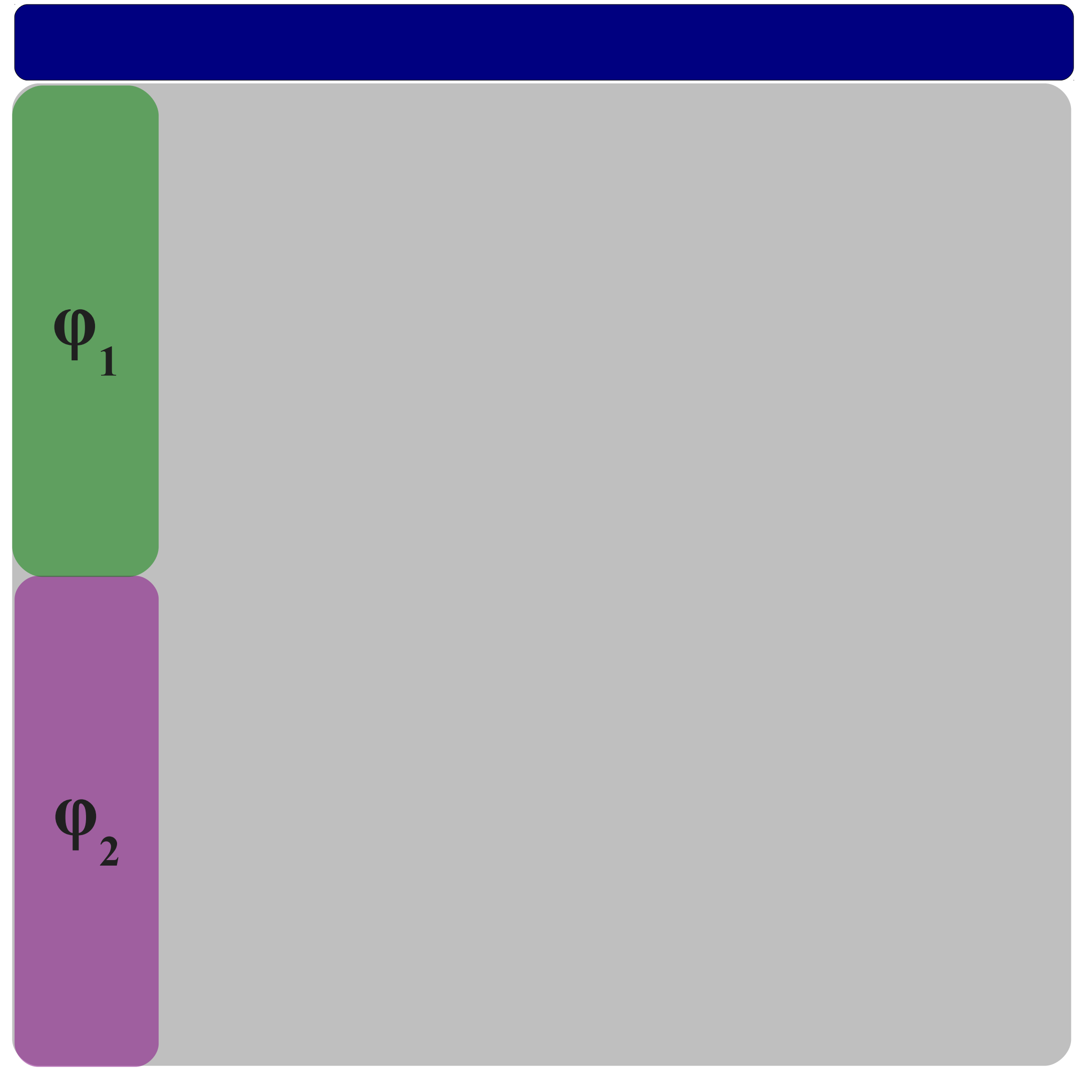
Dataset: the values for a certain number of numerical features (x^d) for multiple entities (φ) at different moments of time (t)

Dataset: the values for a certain number of numerical features (x^d) for multiple entities (φ) at different moments of time (t)

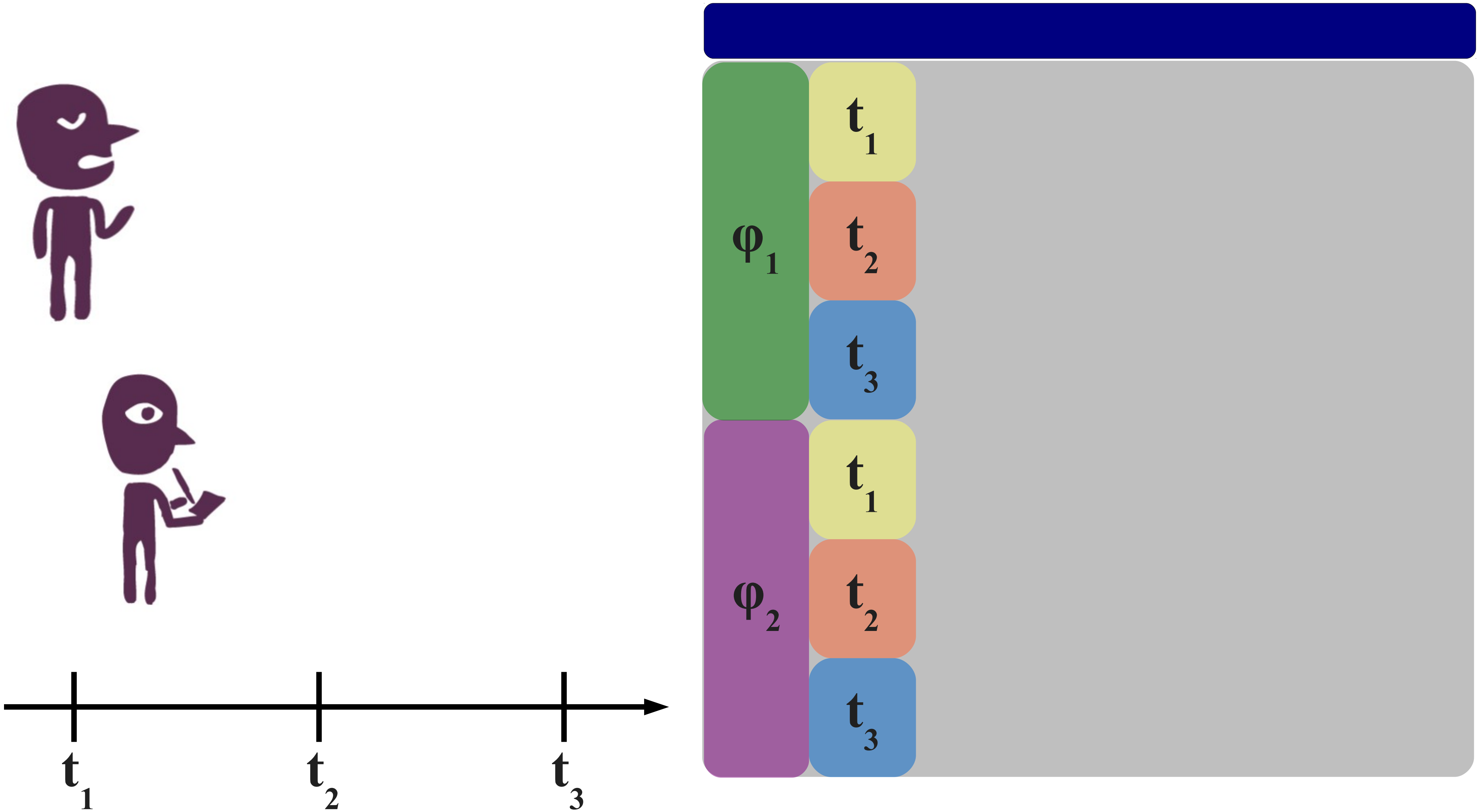


Dataset: the values for a certain number of numerical features (x^d) for multiple entities (φ) at different moments of time (t)

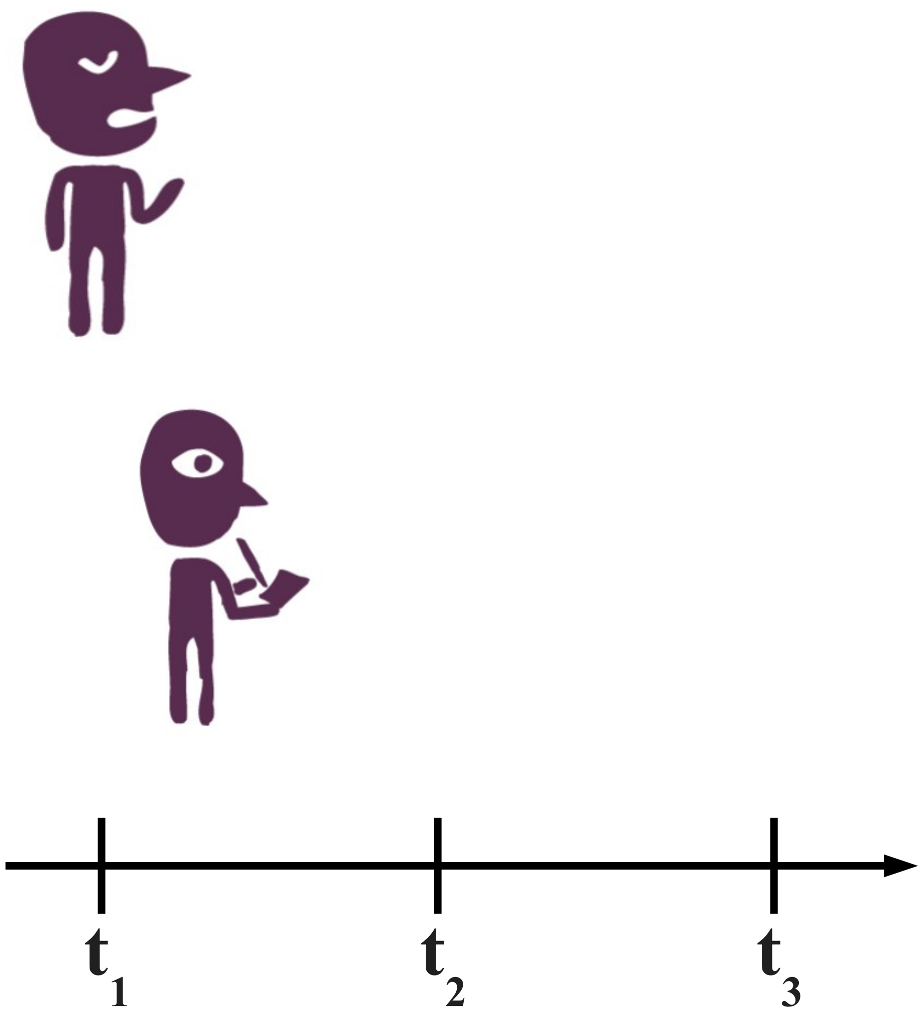


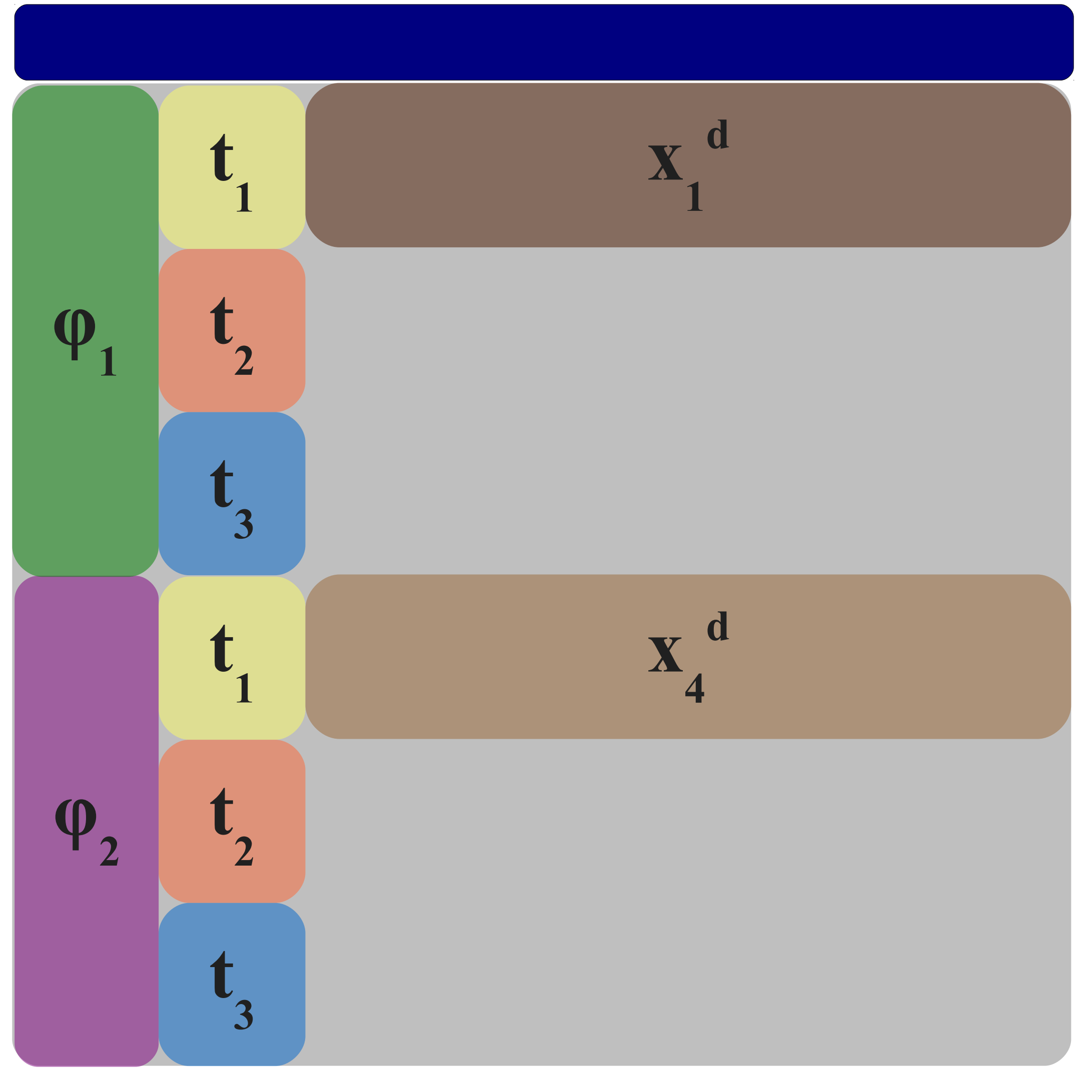


Dataset: the values for a certain number of numerical features (x^d) for multiple entities (φ) at different moments of time (t)

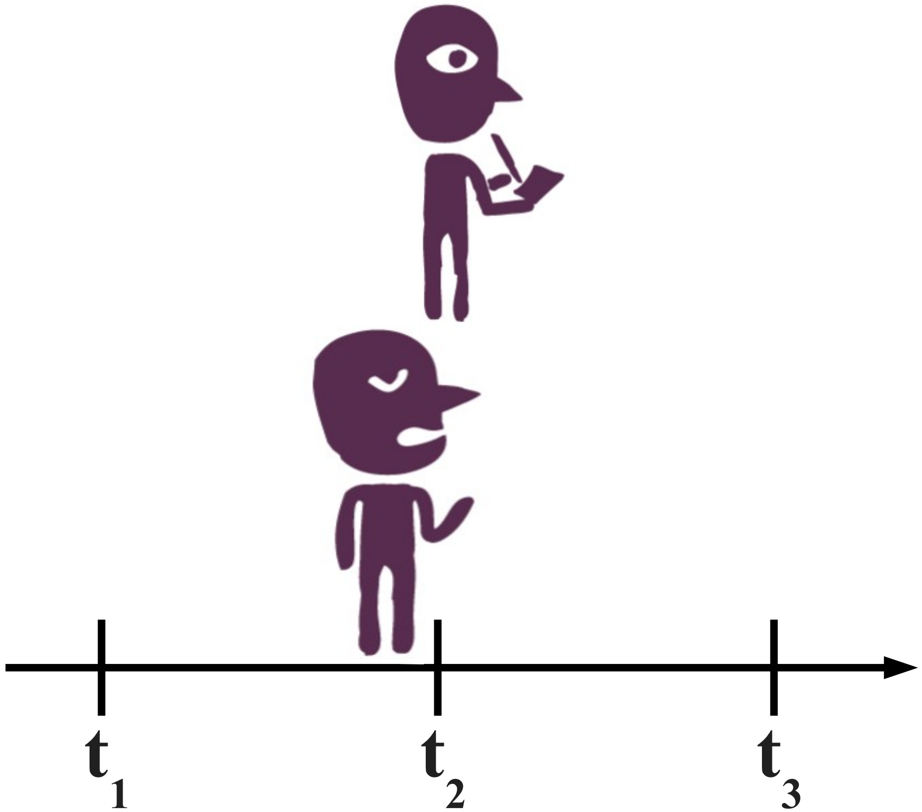


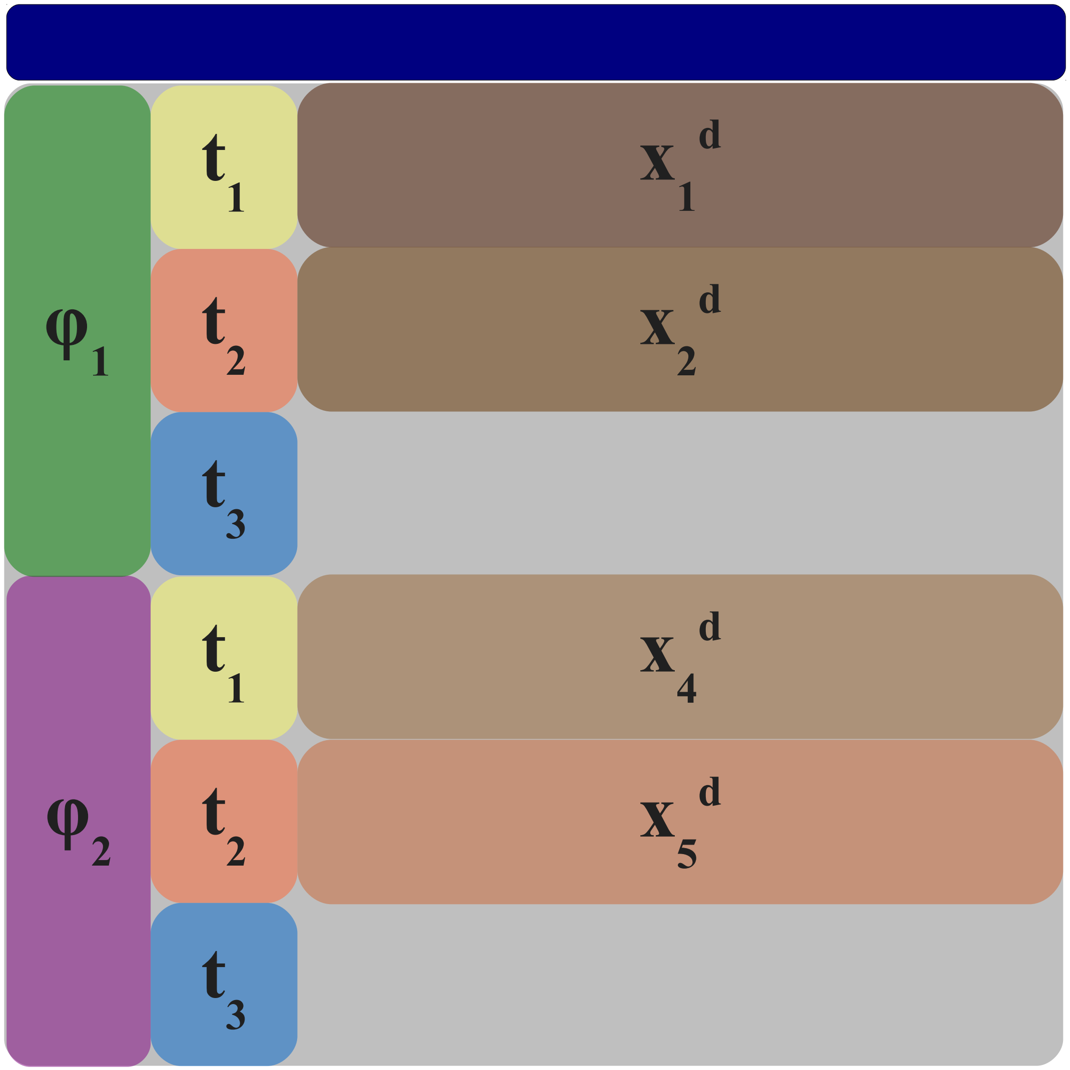
Dataset: the values for a certain number of numerical features (x^d) for multiple entities (φ) at different moments of time (t)



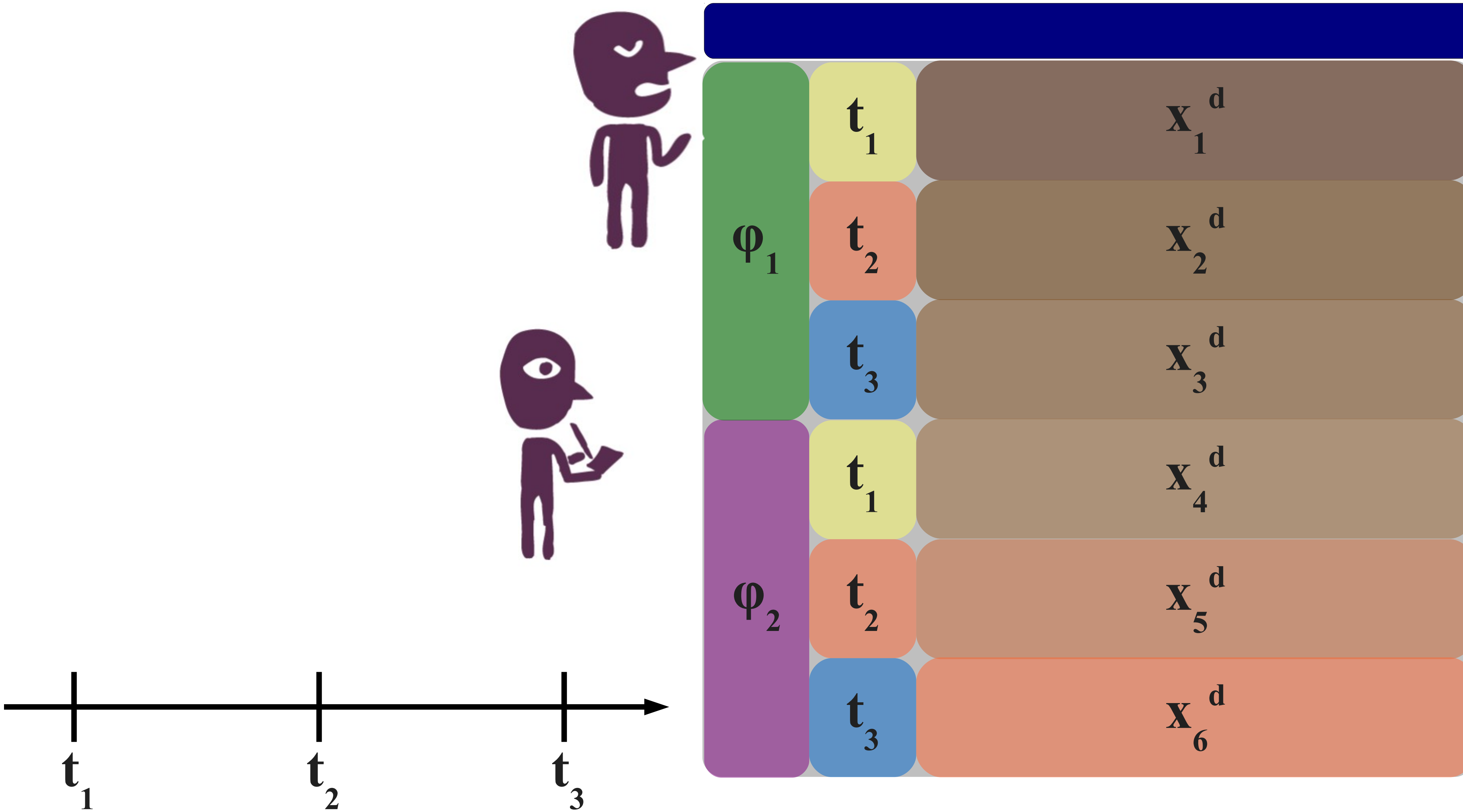


Dataset: the values for a certain number of numerical features (x^d) for multiple entities (φ) at different moments of time (t)





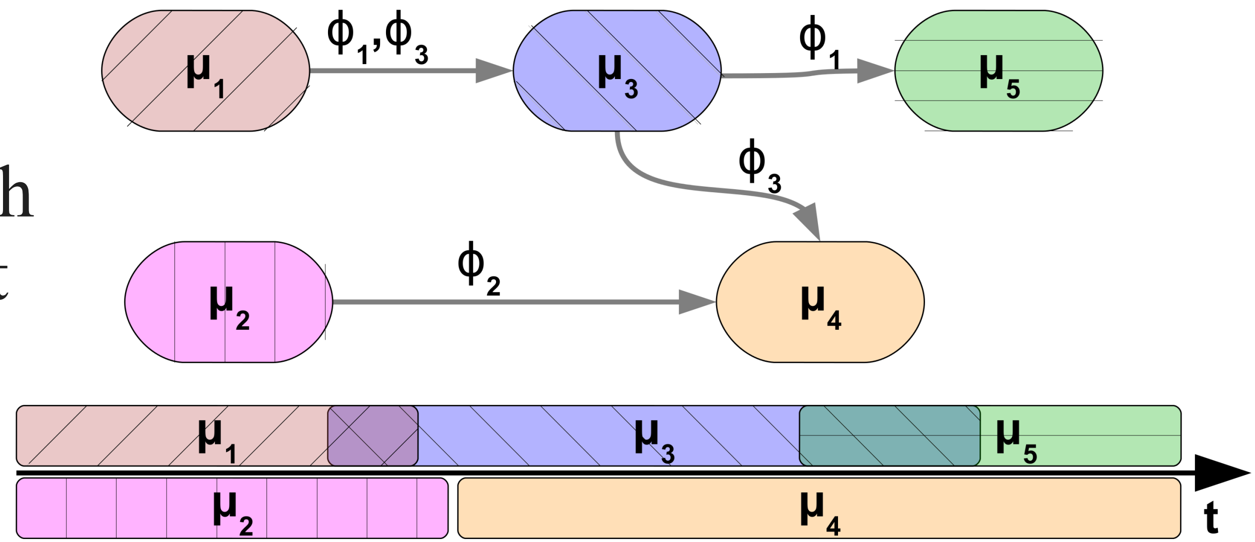
Dataset: the values for a certain number of numerical features (x^d) for multiple entities (φ) at different moments of time (t)



Goal: Detect typical evolution patterns of individuals in the dataset

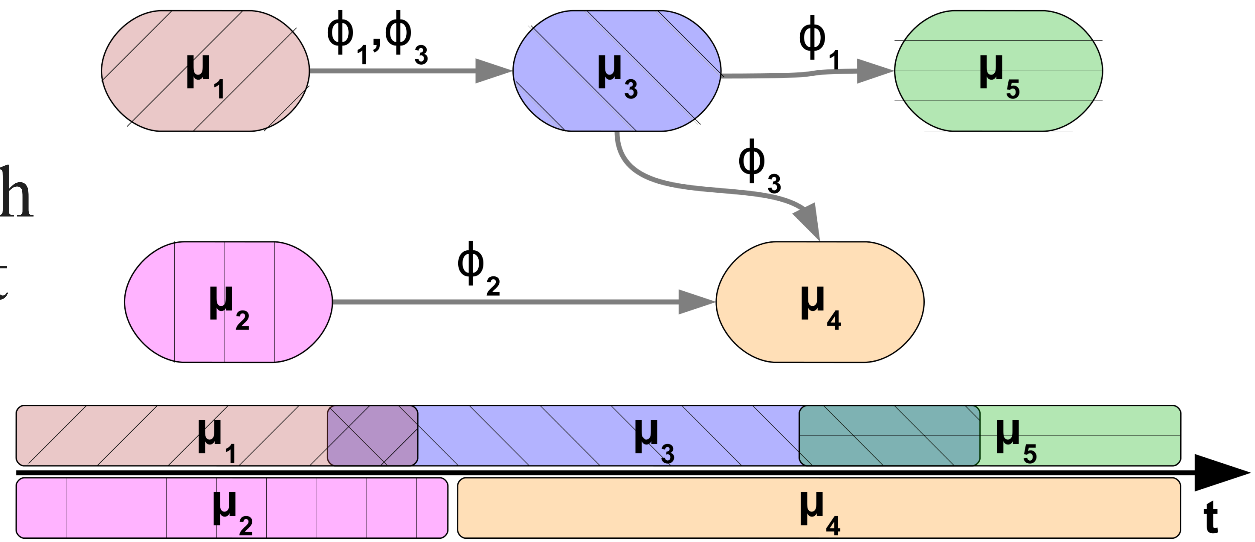
Goal: Detect typical evolution patterns of individuals in the dataset

a) the phases through which the entity collection went over time

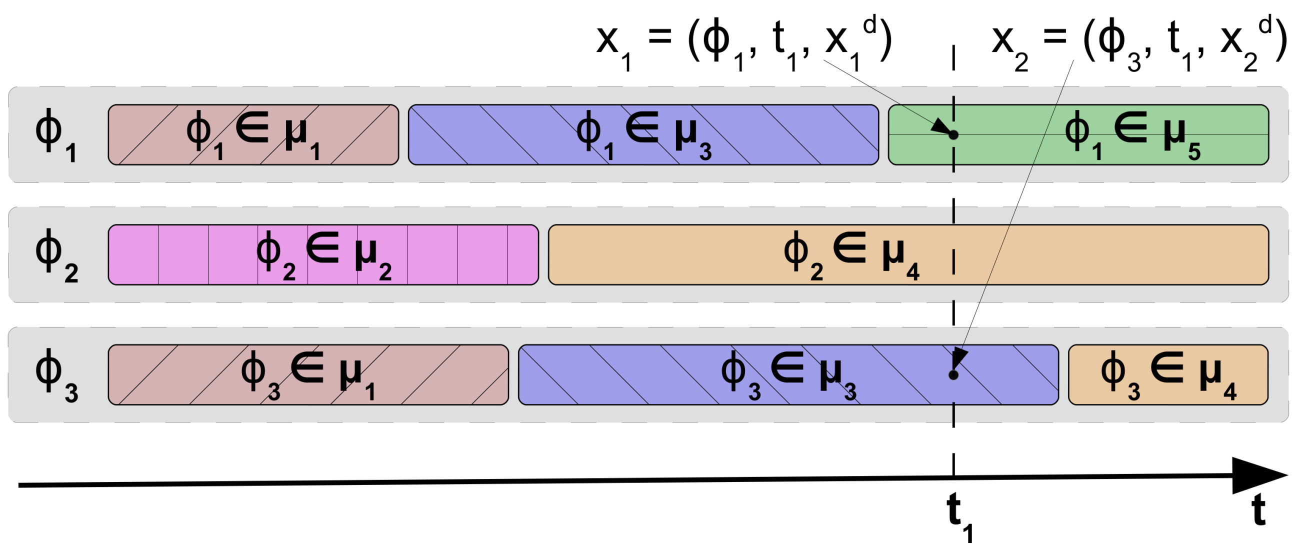


Goal: Detect typical evolution patterns of individuals in the dataset

a) the phases through which the entity collection went over time



b) the trajectory of entities through the different phases



Summary:

1. Problem
   1.1 Data
   1.2 Goal

2. Proposed solutions:
   2.1 A clustering solution
   2.2 Temporal-Aware Dissimilarity Measure
   2.3 Contiguity Penalty Measure
   2.4 TDCK-Means algorithm
   2.5 Evaluation measures

3. Experiments
   3.1 Qualitative evaluation
   3.2 Quantitative evaluation

4. Conclusion and perspectives

Proposed solution: A temporal-aware constrained clustering algorithm, resulted clusters serve as phases.

Proposed solution: A temporal-aware constrained clustering algorithm, resulted clusters serve as phases.

The resulted partition must ensure:

- the descriptive coherence of clusters;
- the temporal coherence of clusters;

- continuous segmentation of observations belonging to an entity.

Proposed solution: A temporal-aware constrained clustering algorithm, resulted clusters serve as phases.

The resulted partition must ensure:

- the descriptive coherence of clusters;
- the temporal coherence of clusters;

→ Temporal-aware dissimilarity measure

- continuous segmentation of observations belonging to an entity.

→ Contiguity penalty measure

Proposed solution: A temporal-aware constrained clustering algorithm, resulted clusters serve as phases.

The resulted partition must ensure:

- the descriptive coherence of clusters;
- the temporal coherence of clusters;

→ Temporal-aware dissimilarity measure

- continuous segmentation of observations belonging to an entity.

→ Contiguity penalty measure

K-Means like algorithm. Objective function to minimize:

$$J = \sum_{\mu_j \in M} \sum_{x_i \in C_j} \left(\| x_i - \mu_j \|_{TE} + \sum_{(x_k \notin C_j) \wedge (x_k^{\varphi} = x_i^{\varphi})} w(x_i, x_k) \right)$$

Proposed solution: A temporal-aware constrained clustering algorithm, resulted clusters serve as phases.

The resulted partition must ensure:

- the descriptive coherence of clusters;
- the temporal coherence of clusters;

→ Temporal-aware dissimilarity measure 1

- continuous segmentation of observations belonging to an entity.

→ Contiguity penalty measure 2

K-Means like algorithm. Objective function to minimize:

$$J = \sum_{\mu_j \in M} \sum_{x_i \in C_j} \left(\underbrace{\|x_i - \mu_j\|_{TE}}_{1} + \underbrace{\sum_{(x_k \notin C_j) \wedge (x_k^{\varphi} = x_i^{\varphi})} w(x_i, x_k)}_{2} \right)$$

Euclidean distance $\longrightarrow$ distance in the description space

Euclidean distance → distance in the description space

Temporal-aware dissimilarity measure → distance in both description space and temporal space

Euclidean distance → distance in the description space

Temporal-aware dissimilarity measure → distance in both description space and temporal space

$$\|x_i - x_j\|_{TE} = 1 - \left(1 - \frac{\|x_i^d - x_j^d\|^2}{\Delta x_{max}}\right)\left(1 - \frac{|x_i^t - x_j^t|^2}{\Delta t_{max}}\right)$$

Euclidean distance → distance in the description space

Temporal-aware dissimilarity measure → distance in both description space and temporal space

$$\|x_i - x_j\|_{TE} = 1 - \left(1 - \frac{\|x_i^d - x_j^d\|^2}{\Delta x_{max}}\right)\left(1 - \frac{|x_i^t - x_j^t|^2}{\Delta t_{max}}\right)$$

Properties:

- $\|x_i - x_j\|_{TE} \in [0,1], \forall x_i, x_j \in X$
- $\|x_i - x_j\|_{TE} = 0 \Leftrightarrow x_i^d = x_j^d \wedge x_i^t = x_j^t$
- $\|x_i - x_j\|_{TE} = 1 \Leftrightarrow \|x_i^d - x_j^d\| = \Delta x_{max} \vee |x_i^t - x_j^t| = \Delta t_{max}$

Semi-Supervised clustering *[Wagstaff & Cardie '00]* → pair-wise constraints → apply penalty when constraints are broken

Semi-Supervised clustering → pair-wise constraints → apply penalty when constraints are broken

[Wagstaff & Cardie '00]

Segmentation contiguity → soft MUST-LINK pair-wise constraints → time-dependent **Contiguity Penalty Function**

Semi-Supervised clustering *[Wagstaff & Cardie '00]* → pair-wise constraints → apply penalty when constraints are broken

Segmentation contiguity → soft MUST-LINK pair-wise constraints → time-dependent **Contiguity Penalty Function**

Contiguity Penalty Function:

$$w(x_i, x_j) = \beta * e^{\frac{-1}{2}\left(\frac{|x_i^t - x_j^t|}{\delta}\right)^2}$$

$$\text{for } x_i^{\varphi} = x_j^{\varphi}$$

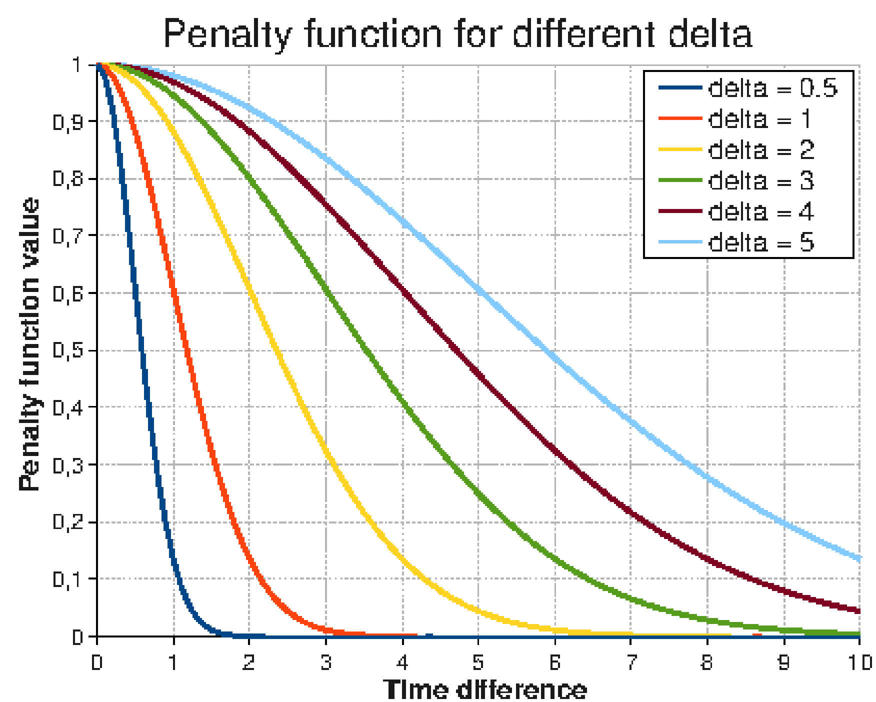
Semi-Supervised clustering *[Wagstaff & Cardie '00]* → pair-wise constraints → apply penalty when constraints are broken

Segmentation contiguity → soft MUST-LINK pair-wise constraints → time-dependent **Contiguity Penalty Function**

Contiguity Penalty Function:

$$w(x_i, x_j) = \beta * e^{\frac{-1}{2}\left(\frac{|x_i^t - x_j^t|}{\delta}\right)^2}$$

$$for\ x_i^{\varphi} = x_j^{\varphi}$$



The TDCK-Means algorithm: Inspired from K-Means. Iteratively recomputes centroids and assignments of observations to clusters.

Uses the Temporal-Aware Dissimilarity Function and the Contiguity Penalty Function.

Centroids: (μ_j^t, μ_j^d)

The TDCK-Means algorithm: Inspired from K-Means. Iteratively recomputes centroids and assignments of observations to clusters.

Uses the Temporal-Aware Dissimilarity Function and the Contiguity Penalty Function.

Centroids: (μ_j^t, μ_j^d)

Centroids update:

$$\underbrace{\mu_j^d = \frac{\sum_{x_i \in C_j} x_i^d * \left(1 - \frac{|x_i^t - \mu_j^t|^2}{\Delta t_{max}^2}\right)}{\sum_{x_i \in C_j} \left(1 - \frac{|x_i^t - \mu_j^t|^2}{\Delta t_{max}^2}\right)} \qquad \mu_j^t = \frac{\sum_{x_i \in C_j} x_i^t * \left(1 - \frac{\|x_i^d - \mu_j^d\|^2}{\Delta x_{max}^2}\right)}{\sum_{x_i \in C_j} \left(1 - \frac{\|x_i^d - \mu_j^d\|^2}{\Delta x_{max}^2}\right)}}_{\text{Weighted averages}}$$

Partition evaluation measures

→ descriptive coherence of clusters;
→ temporal coherence of clusters;

→ continuous segmentation of observations belonging to an entity.

Partition evaluation measures

- descriptive coherence of clusters;
- temporal coherence of clusters;

$\xrightarrow{\text{variance}}$ MDvar, Tvar

- continuous segmentation of observations belonging to an entity.

Shannon Entropy $\quad A \rightarrow B \rightarrow A \rightarrow B$???

Partition evaluation measures

- descriptive coherence of clusters;
- temporal coherence of clusters;

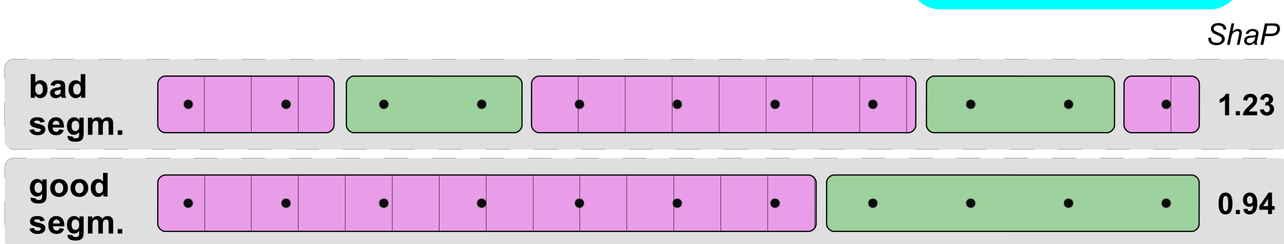
$\xrightarrow{\text{variance}}$ MDvar, Tvar

- continuous segmentation of observations belonging to an entity.

Shannon Entropy A → B → A → B ???

Proposal: Correct the Shannon entropy to penalize changes

$$ShaP = \sum_{x_i \in X} \sum_{j=1}^{k} \left(-p(\mu_j) * \log_2(p(\mu_j)) * \boxed{\left(1 + \frac{n_{ch} - n_{min}}{n_{obs} - 1}\right)} \right)$$



Summary:

1. Problem
  1.1 Data
  1.2 Goal

2. Proposed solutions:
  2.1 A clustering solution
  2.2 Temporal-Aware Dissimilarity Measure
  2.3 Contiguity Penalty Measure
  2.4 TDCK-Means algorithm
  2.5 Evaluation measures

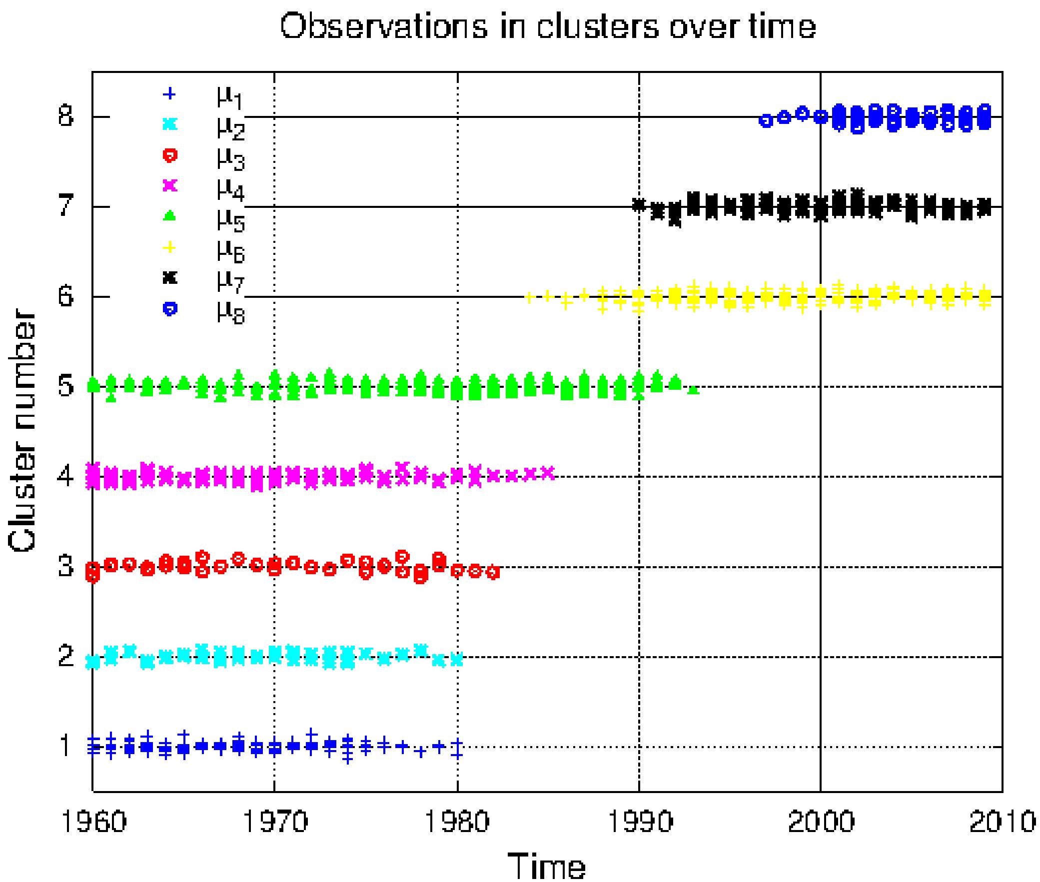
3. Experiments
  3.1 Qualitative evaluation
  3.2 Quantitative evaluation

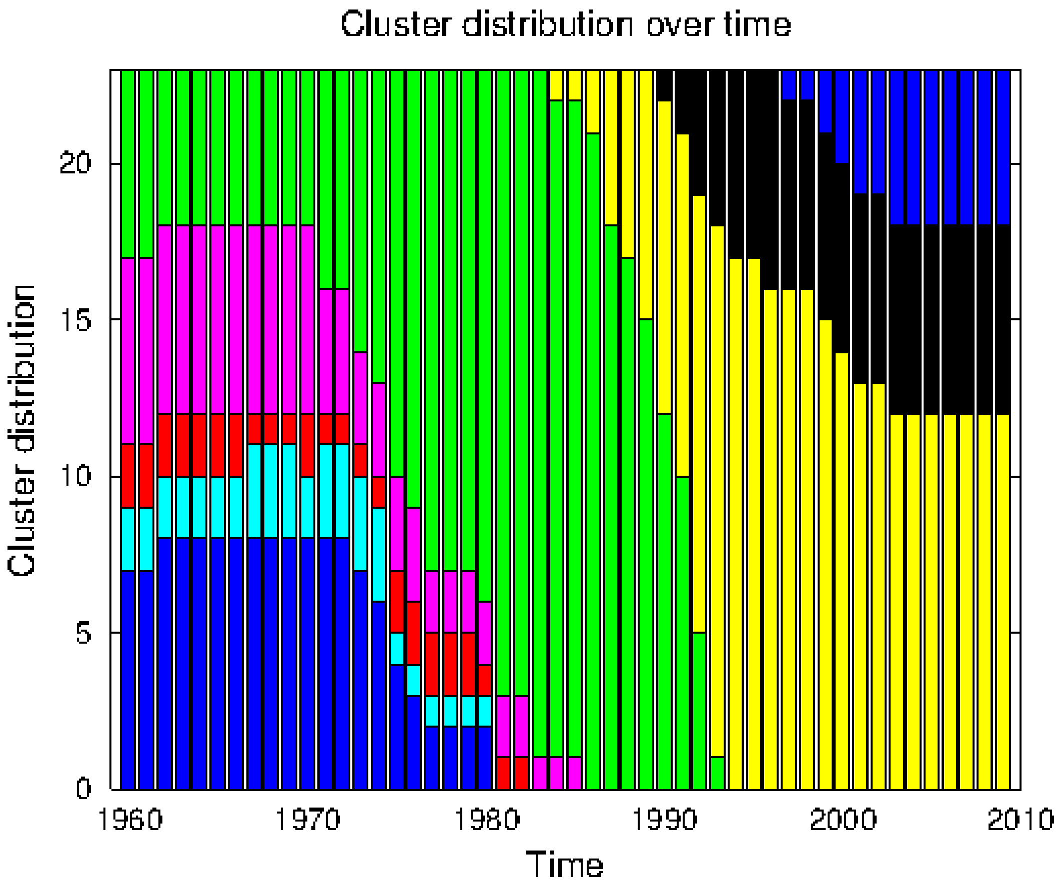
4. Conclusion and perspectives

Compared Political Dataset I 23 countries, 60 years, 207 political, demographic, social and economic variables.

Compared Political Dataset I 23 countries, 60 years, 207 political, demographic, social and economic variables.

Execution TDCK-Means (8 clusters, $\beta = 0.003$ and $\delta = 3$)

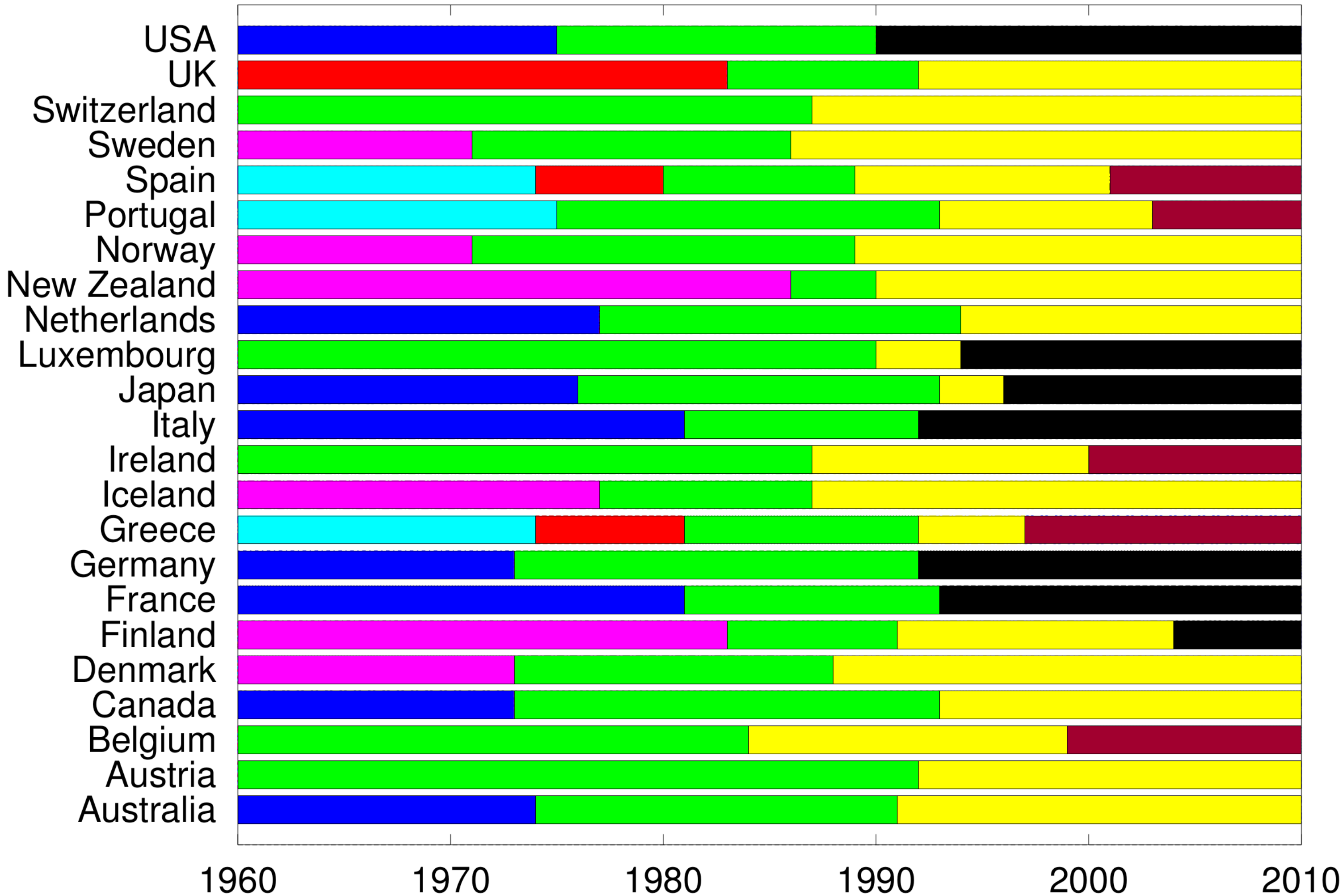




Compared Political Dataset I

23 countries, 60 years, 207 political, demographic, social and economic variables.

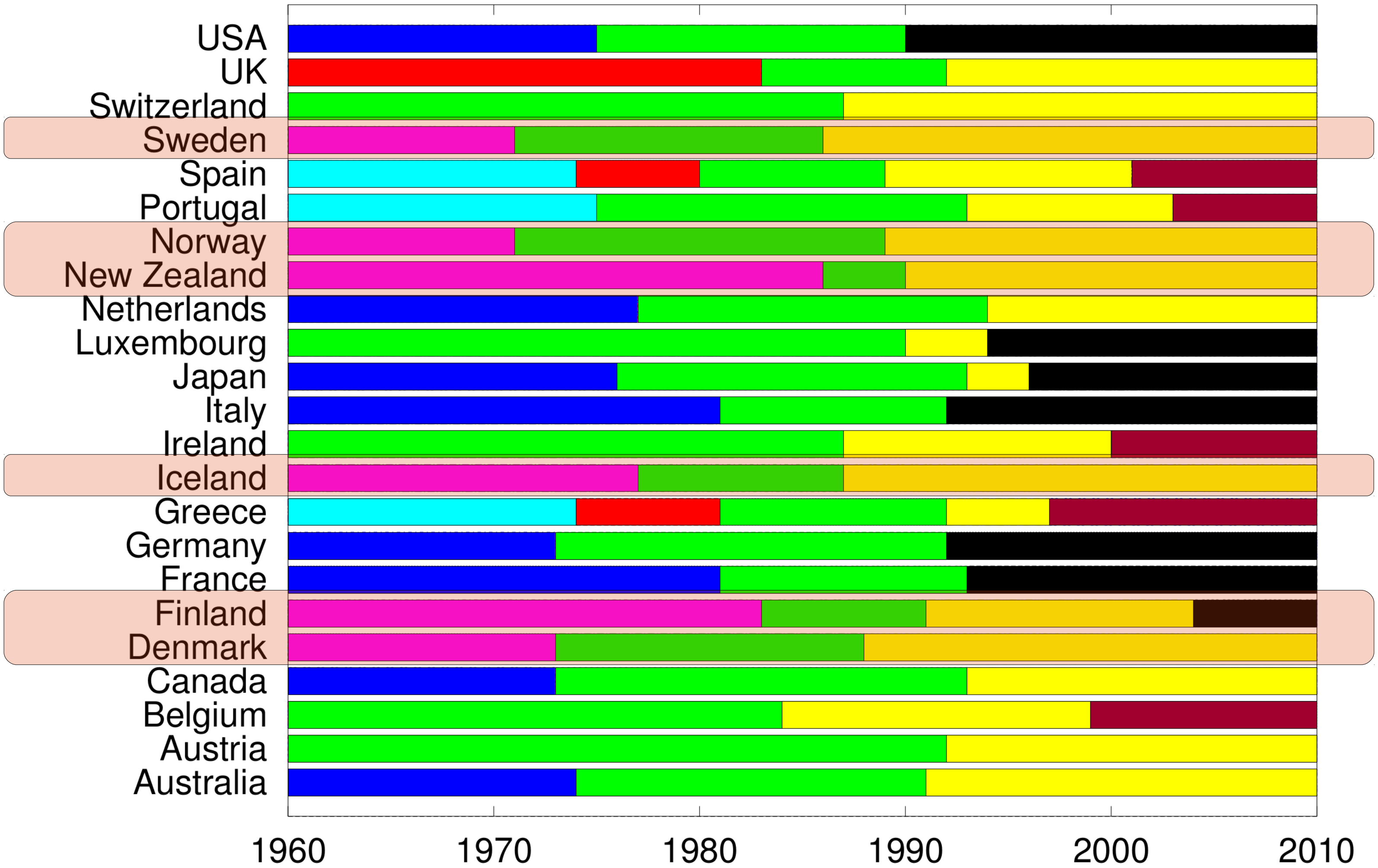
Execution TDCK-Means (8 clusters, $\beta = 0.003$ and $\delta = 3$)



Compared Political Dataset I

23 countries, 60 years, 207 political, demographic, social and economic variables.

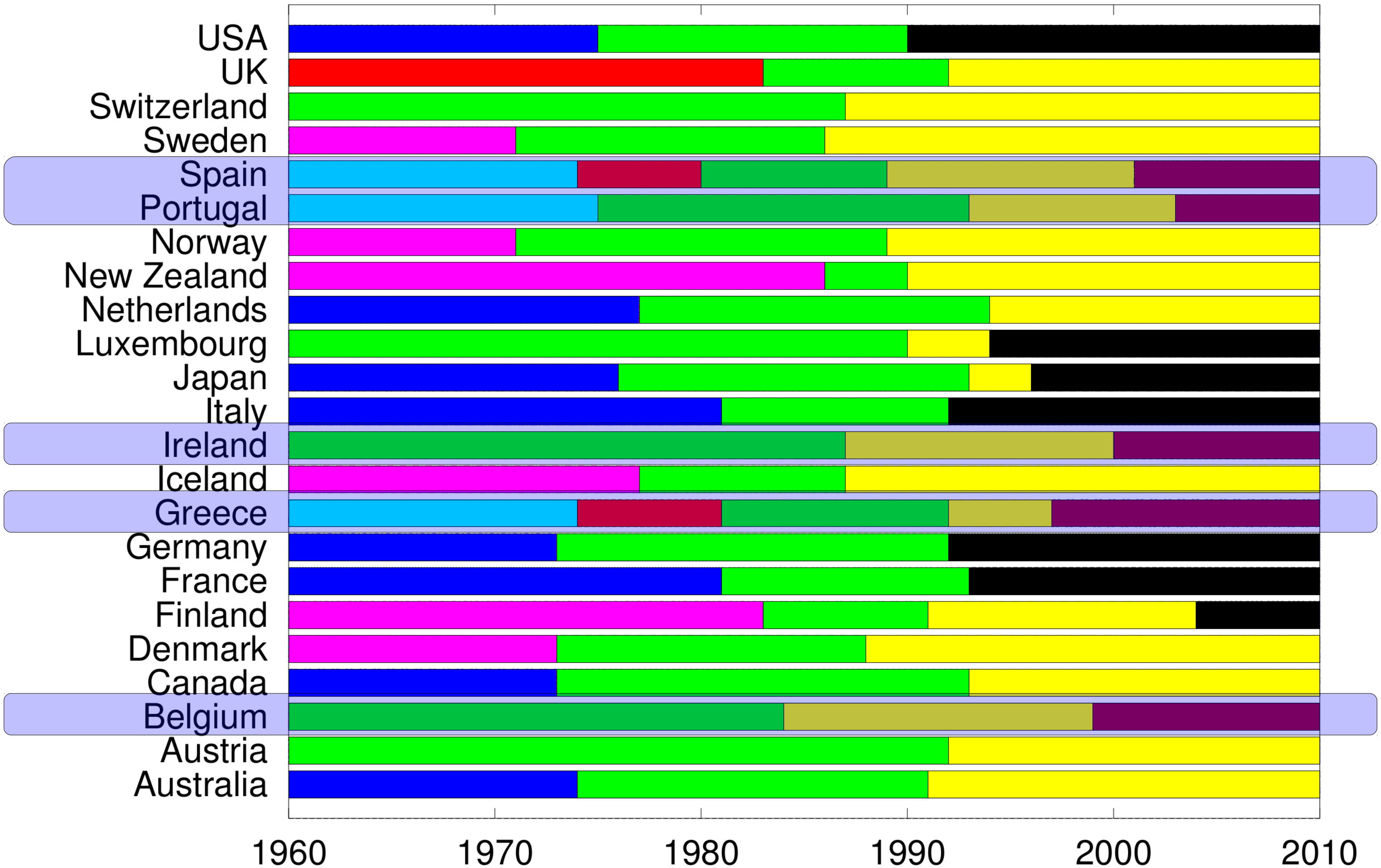
Execution TDCK-Means (8 clusters, $\beta = 0.003$ and $\delta = 3$)



Compared Political Dataset I

23 countries, 60 years, 207 political, demographic, social and economic variables.

Execution TDCK-Means (8 clusters, β = 0.003 and δ = 3)



Summary:

1. Problem
   - 1.1 Data
   - 1.2 Goal
2. Proposed solutions:
   - 2.1 A clustering solution
   - 2.2 Temporal-Aware Dissimilarity Measure
   - 2.3 Contiguity Penalty Measure
   - 2.4 TDCK-Means algorithm
   - 2.5 Evaluation measures
3. Experiments
   - 3.1 Qualitative evaluation
   - 3.2 Quantitative evaluation
4. Conclusion and perspectives

Conclusion:

- Studied the detection of typical evolutions starting from a collection of observations corresponding to entities;
- Proposed a new **Temporal-Aware Measure**;
- Proposed a new **Contiguity Penalty Function**;
- Proposed a new algorithm for detecting evolutions: **TDCK-Means;**
- Other applications: political careers, life trajectories *etc.*

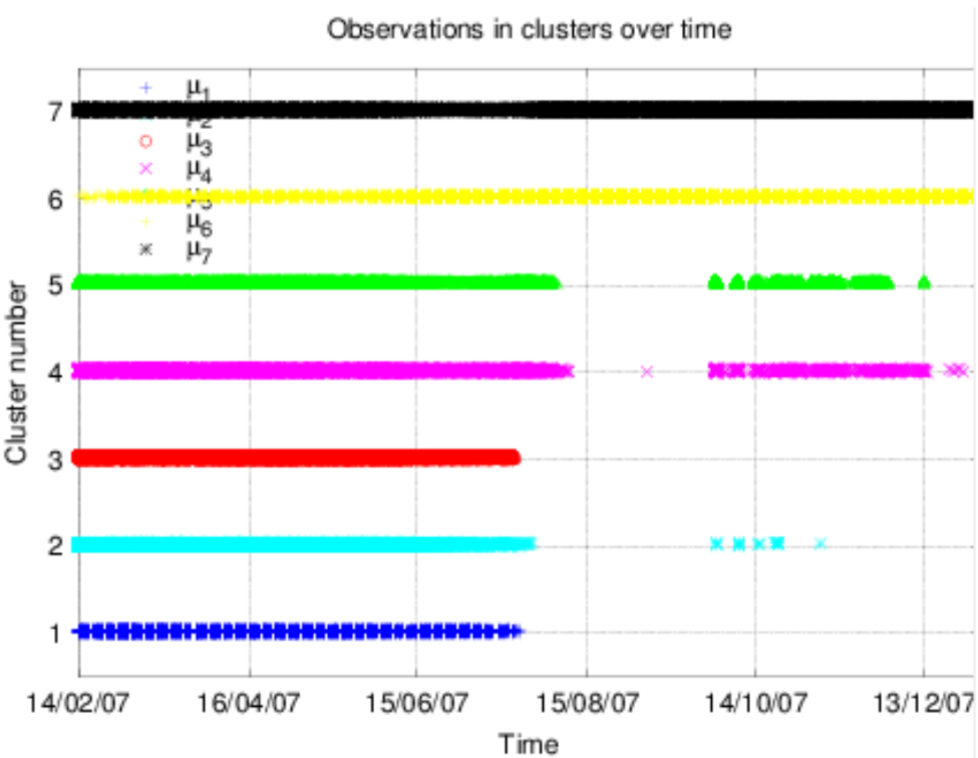
Current work: Apply the TDCK-Means to detect user social roles

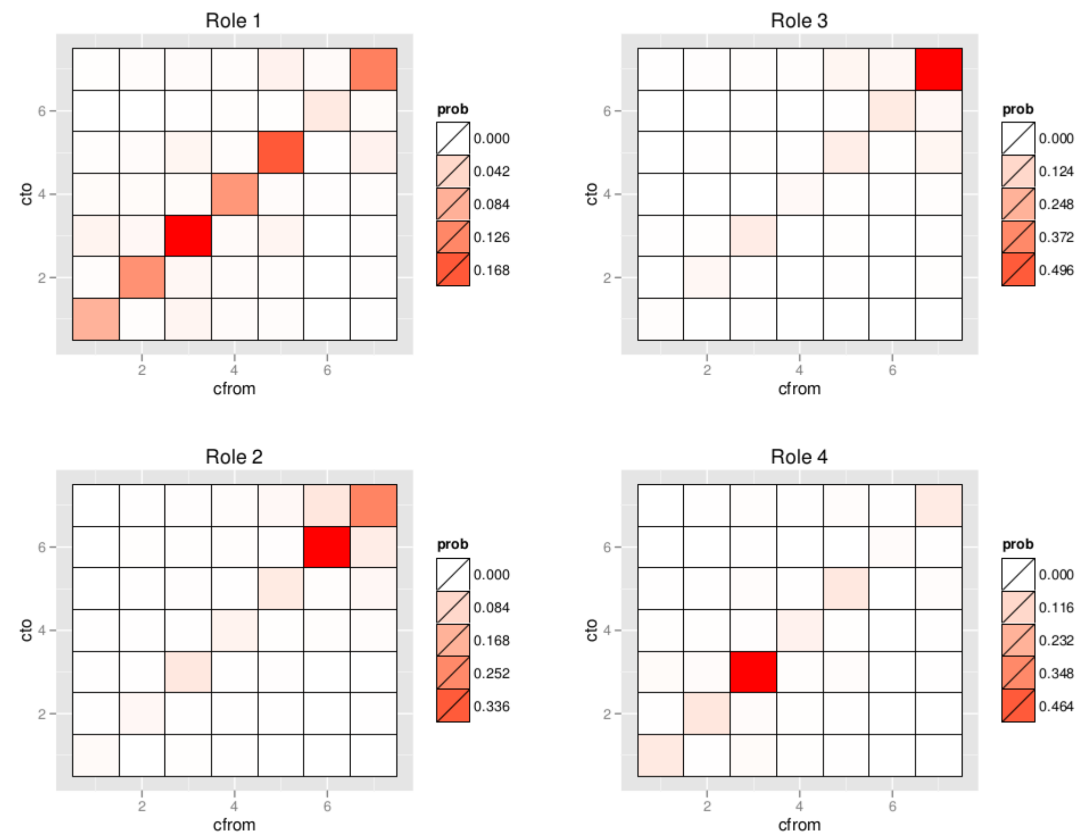
(in collaboration with Technicolor laboratories, Rennes)

Current work: Apply the TDCK-Means to detect user social roles

(in collaboration with Technicolor laboratories, Rennes)

Regroup user activity into temporal contiguous clusters
Interpret transitions between clusters as user roles.





Current work: Infer graph structure for clusters during the clustering

(Research group Julien V., Stéphane B., Stéphane L. and Rizoiu M-A.)

Current work: Infer graph structure for clusters during the clustering

(Research group Julien V., Stéphane B., Stéphane L. and Rizoiu M-A.)

Modify the objective function to take into account a graph structure

$$f_{opt} = \lambda_1 \sum_{p=1}^{k} \sum_{x_i \in X_p} ||x_i - \mu_p||_{TE} + \lambda_2 \sum_{p=1}^{k} \sum_{q=1}^{k} a_{pq}^2 d\_T(c_p, c_q)^2 + \lambda_3 \sum_{p=1}^{k} \sum_{q=1}^{k} a_{pq}^2 inter_{\phi}(c_p, c_q)^2$$

Current work: Infer graph structure for clusters during the clustering

(Research group Julien V., Stéphane B., Stéphane L. and Rizoiu M-A.)

Modify the objective function to take into account a graph structure

$$f_{opt} = \lambda_1 \sum_{p=1}^{k} \sum_{x_i \in X_p} ||x_i - \mu_p||_{TE} + \lambda_2 \sum_{p=1}^{k} \sum_{q=1}^{k} \boxed{a_{pq}^2 d\_T(c_p, c_q)^2} + \lambda_3 \sum_{p=1}^{k} \sum_{q=1}^{k} a_{pq}^2 inter_{\phi}(c_p, c_q)^2$$

Temporal distance between clusters

Current work: Infer graph structure for clusters during the clustering

(Research group Julien V., Stéphane B., Stéphane L. and Rizoiu M-A.)

Modify the objective function to take into account a graph structure

$$f_{opt} = \lambda_1 \sum_{p=1}^{k} \sum_{x_i \in X_p} ||x_i - \mu_p||_{TE} + \lambda_2 \sum_{p=1}^{k} \sum_{q=1}^{k} \underbrace{a_{pq}^2 d\_T(c_p, c_q)^2}_{\text{Temporal distance between clusters}} + \lambda_3 \sum_{p=1}^{k} \sum_{q=1}^{k} \underbrace{a_{pq}^2 inter_{\phi}(c_p, c_q)^2}_{\text{Intersection of clusters}}$$

Current work: Infer graph structure for clusters during the clustering

(Research group Julien V., Stéphane B., Stéphane L. and Rizoiu M-A.)

Modify the objective function to take into account a graph structure

$$f_{opt} = \lambda_1 \sum_{p=1}^{k} \sum_{x_i \in X_p} ||x_i - \mu_p||_{TE} + \lambda_2 \sum_{p=1}^{k} \sum_{q=1}^{k} \underbrace{a_{pq}^2 d\_T(c_p, c_q)^2}_{\text{Temporal distance between clusters}} + \lambda_3 \sum_{p=1}^{k} \sum_{q=1}^{k} \underbrace{a_{pq}^2 inter_{\phi}(c_p, c_q)^2}_{\text{Intersection of clusters}}$$

Estimate the adjacency matrix during the Objective Function optimization

a_{ij} – link between clusters *i* and *j*

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix}$$

Enforce the link between two clusters when:

- the two clusters are close in time;
- the two clusters share multiple entities.

Enforce the link between two clusters when:

- the two clusters are close in time;
- the two clusters share multiple entities.

Re-calculate centroids:

- gradient descent;
- Lagrange multiplicators.

Thank you!

Questions?

Quantitative evaluation

5 algorithms:

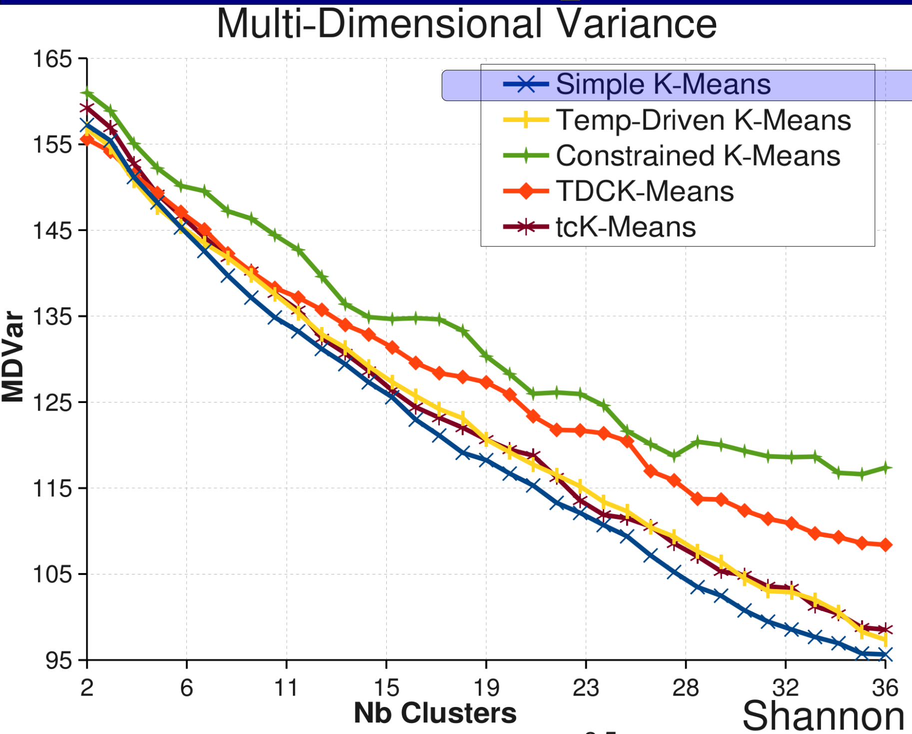
- K-Means *[MacQueen '67]*;
- tcK-Means *[Lin and Hauptmann '10]*

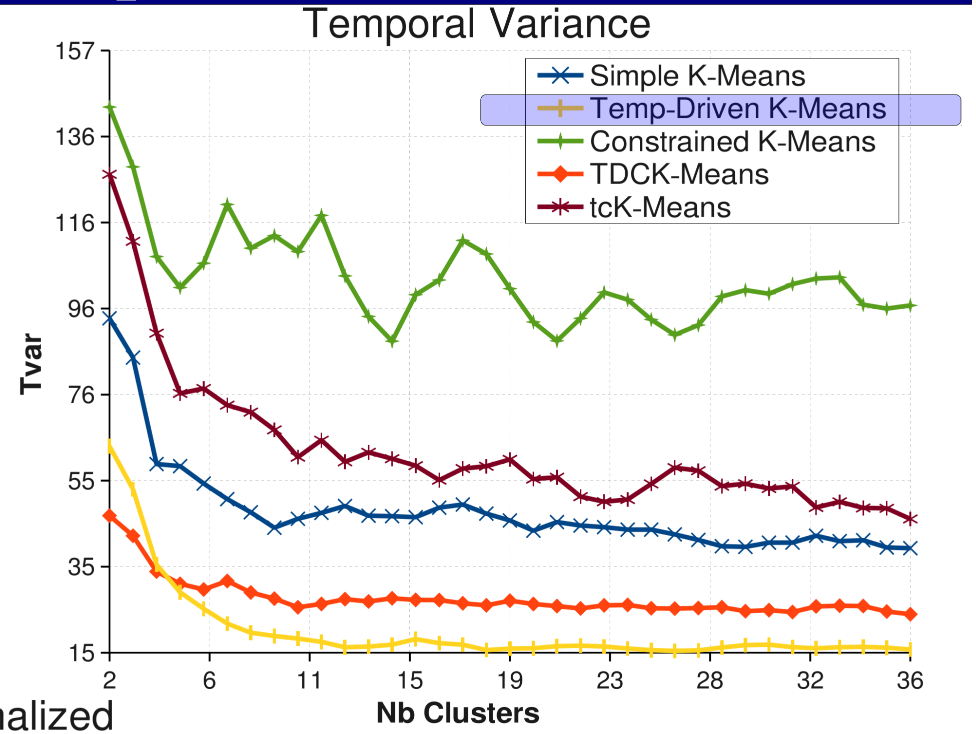
- Temporal-Driven K-Means;
  (uses Temporal-Aware Measure)
- Constrained K-Means;
  (uses Contiguity Penalty Function)

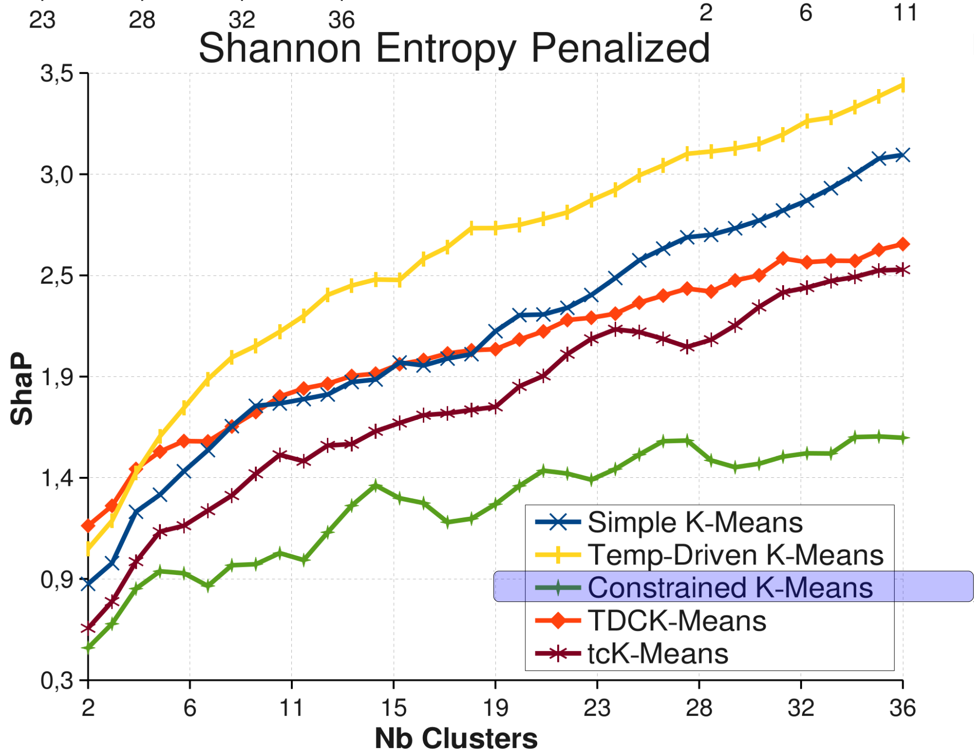
- **TDCK-Means;**
  (combines the two above)

3 measures:

- MDvar
- Tvar
- ShaP


Multi-Dimensional Variance
MDVar
Nb Clusters
Simple K-Means
Temp-Driven K-Means
Constrained K-Means
TDCK-Means
tcK-Means


Temporal Variance
Tvar
Nb Clusters
Simple K-Means
Temp-Driven K-Means
Constrained K-Means
TDCK-Means
tcK-Means


Shannon Entropy Penalized
ShaP
Nb Clusters
Simple K-Means
Temp-Driven K-Means
Constrained K-Means
TDCK-Means
tcK-Means