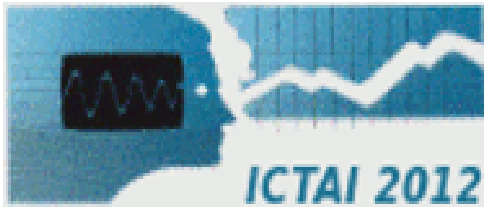


Structuring Typical Evolutions using Temporal-Driven Constrained Clustering

8 November 2012

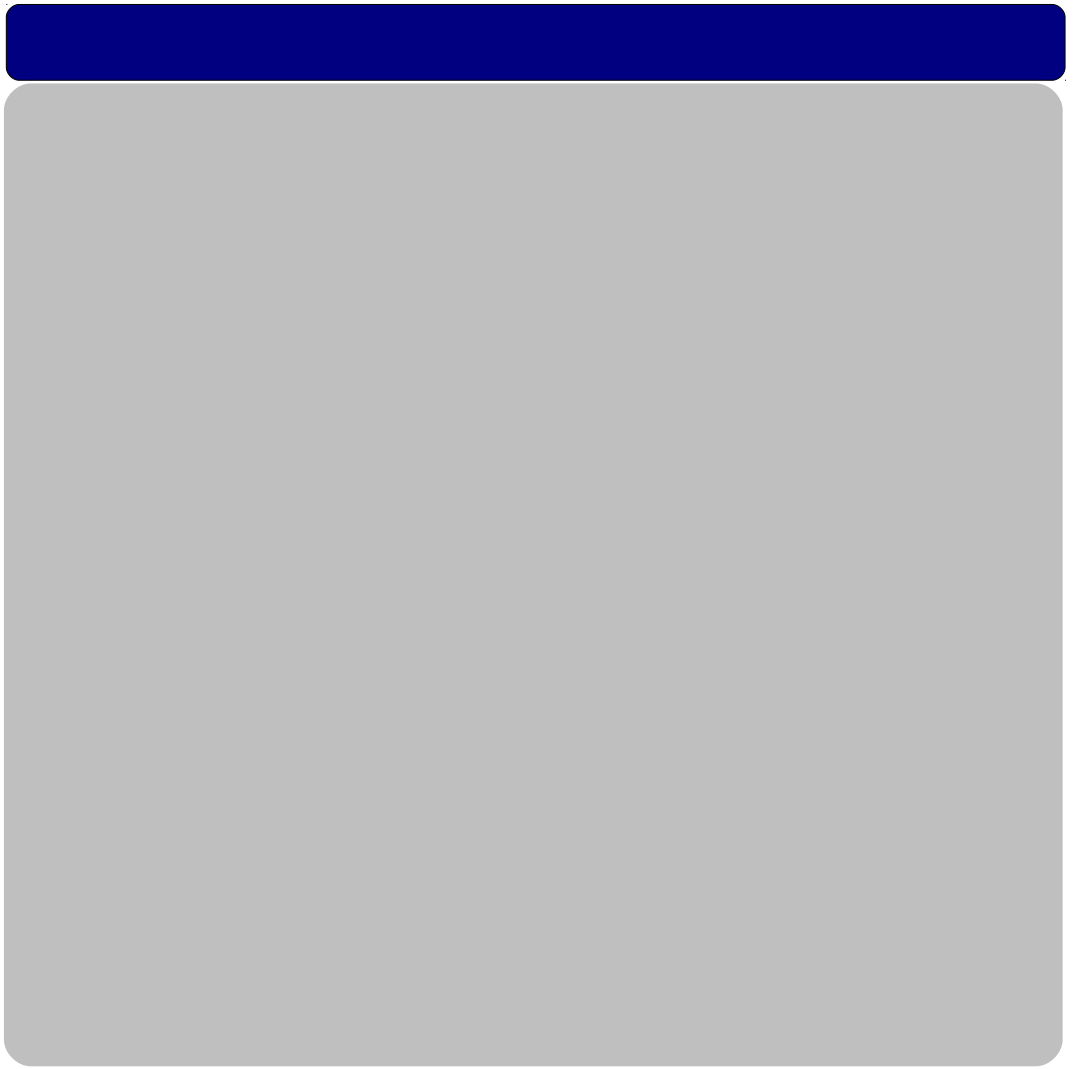
Marian-Andrei Rizoiu
Julien Velcin
Stéphane Lallich

ERIC Laboratory
Université Lumière Lyon 2
France

UNIVERSITÉ LUMIÈRE LYON 2
UNIVERSITÉ DE LYON

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)

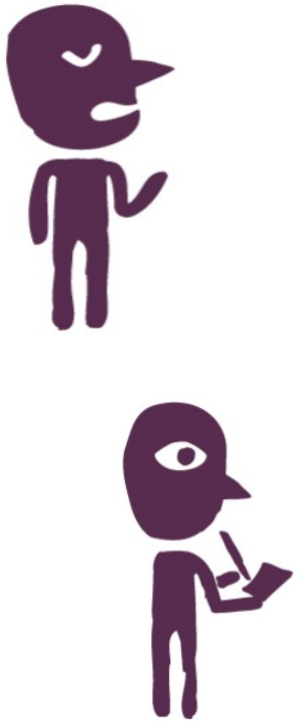


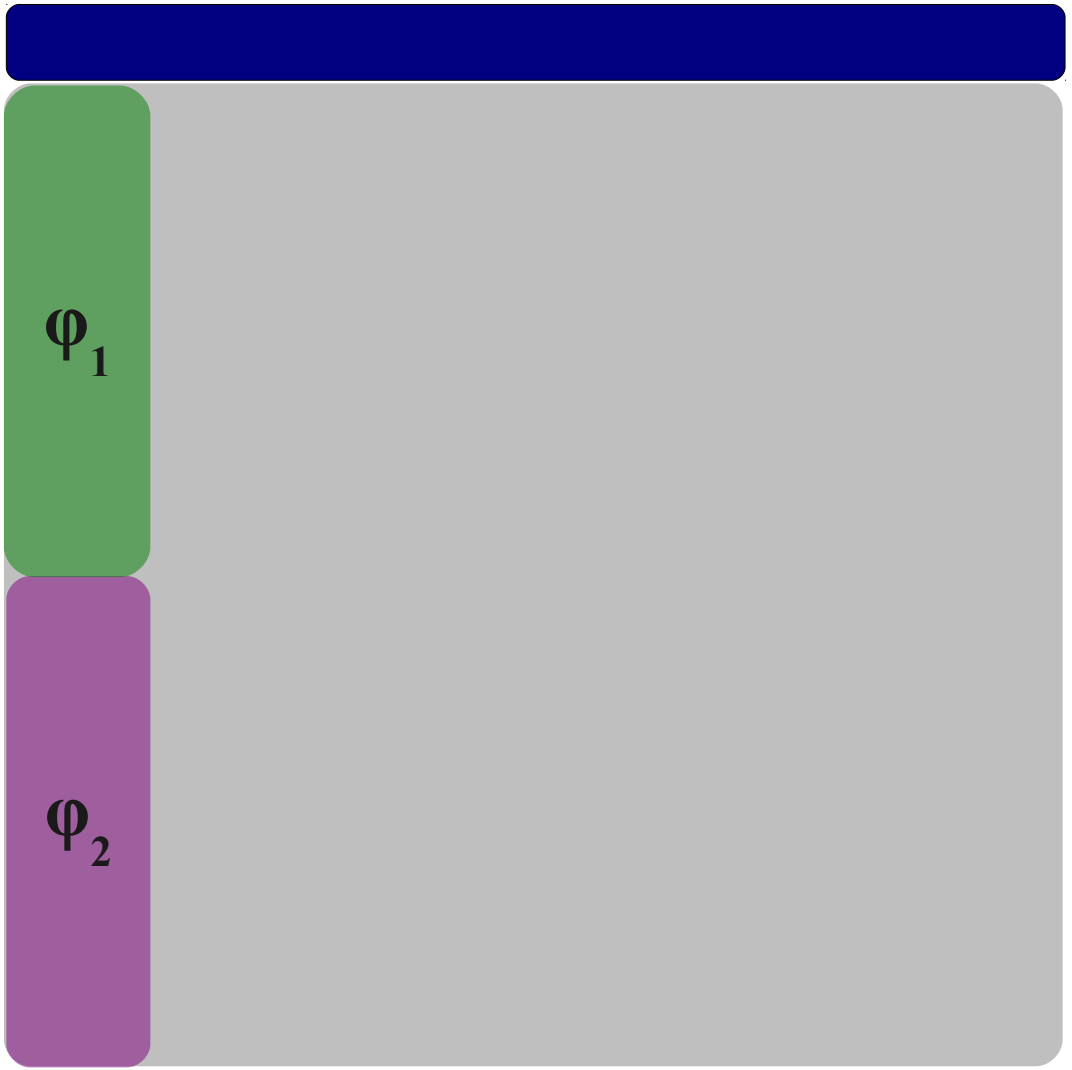
Dataset:

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

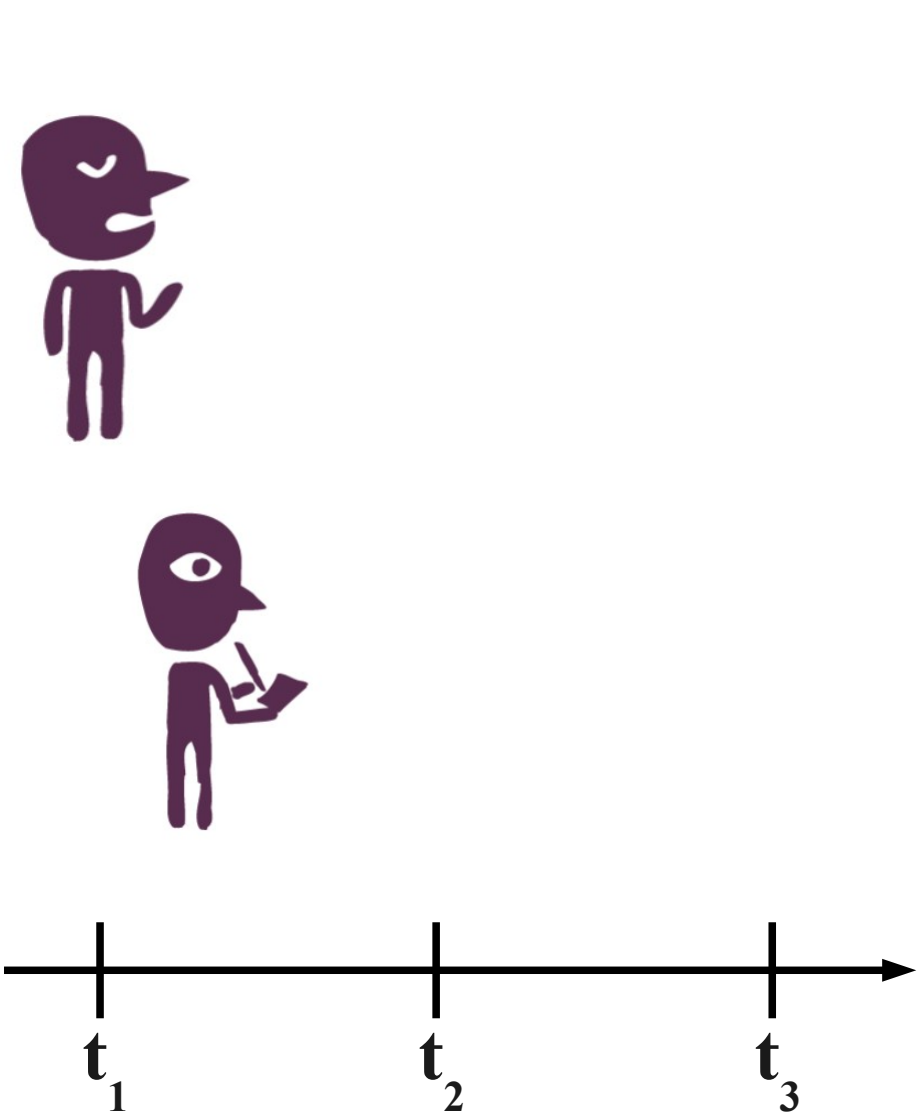
Dataset:

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

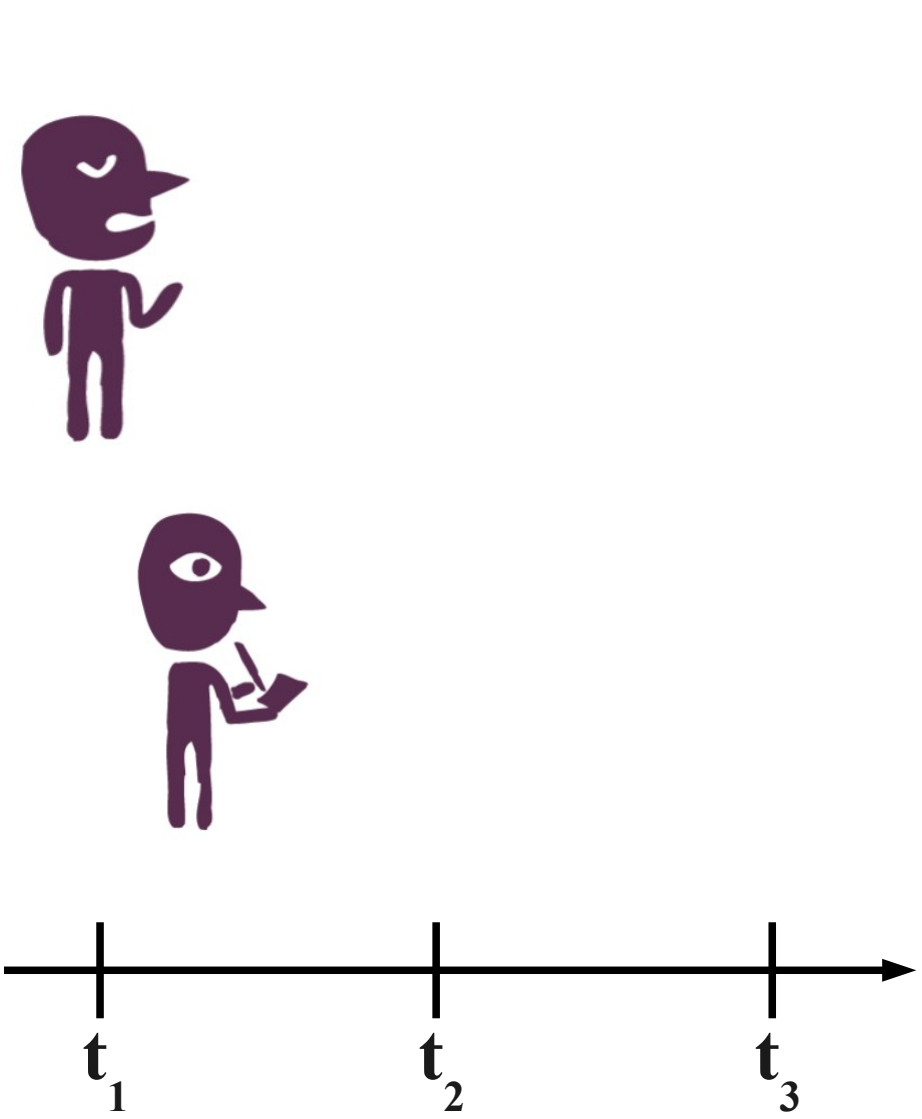


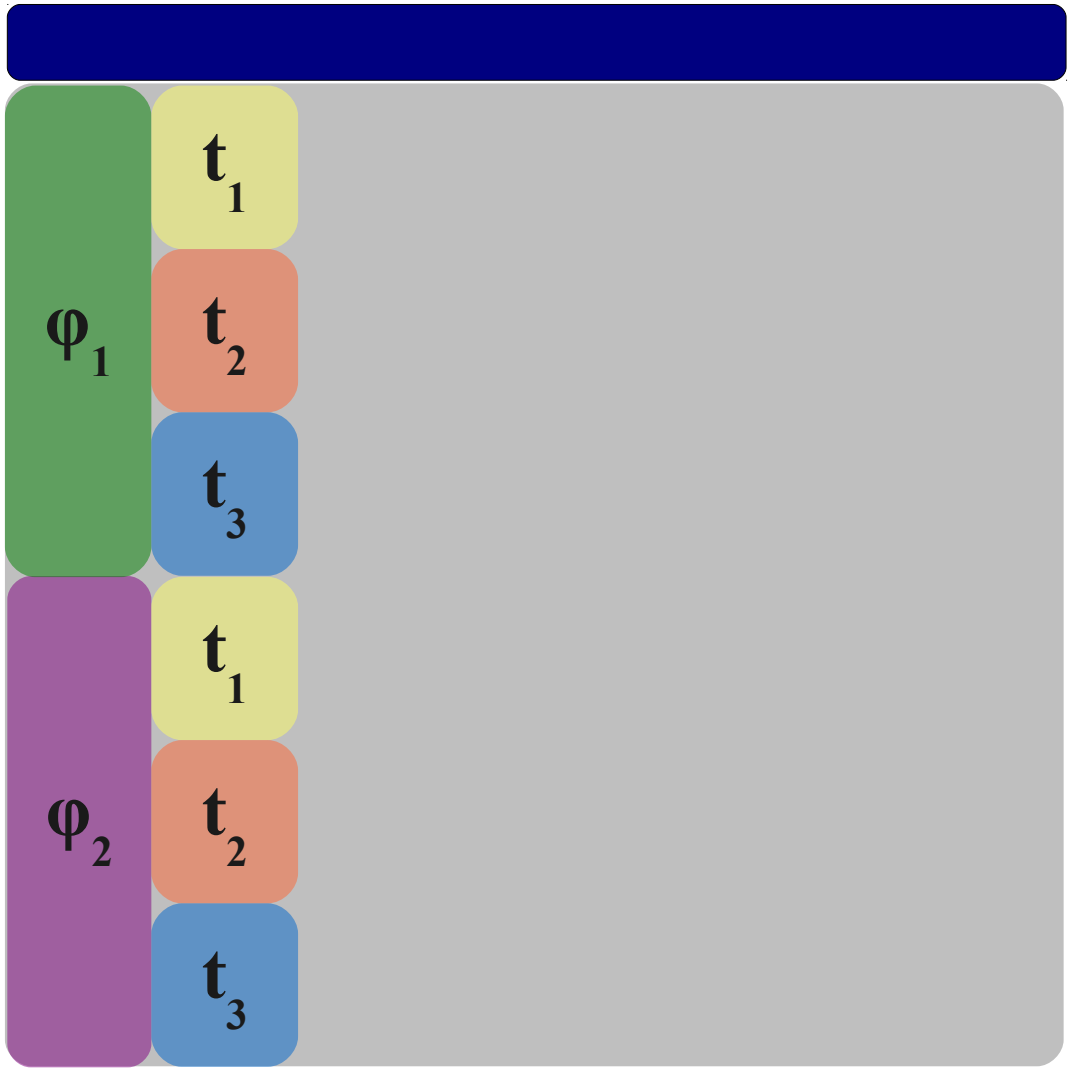


Dataset: the values for a certain number of numerical features (x^d) for multiple entities (φ) at different moments of time (t)

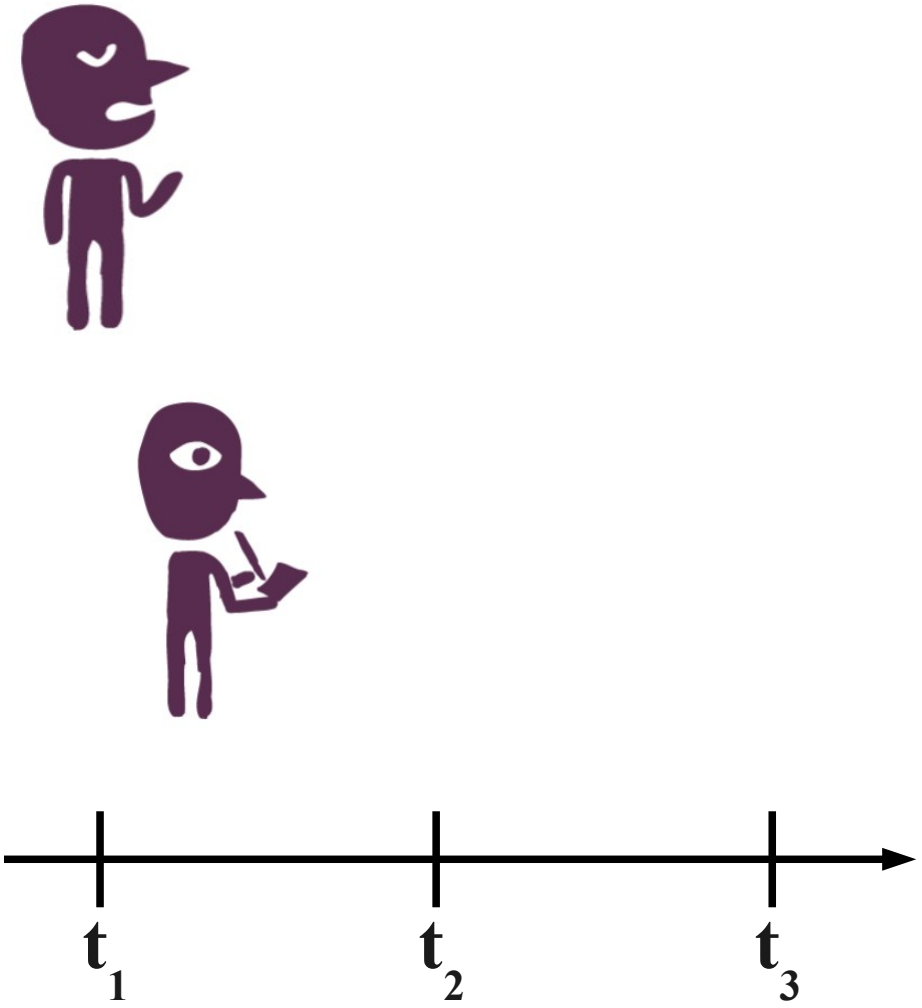


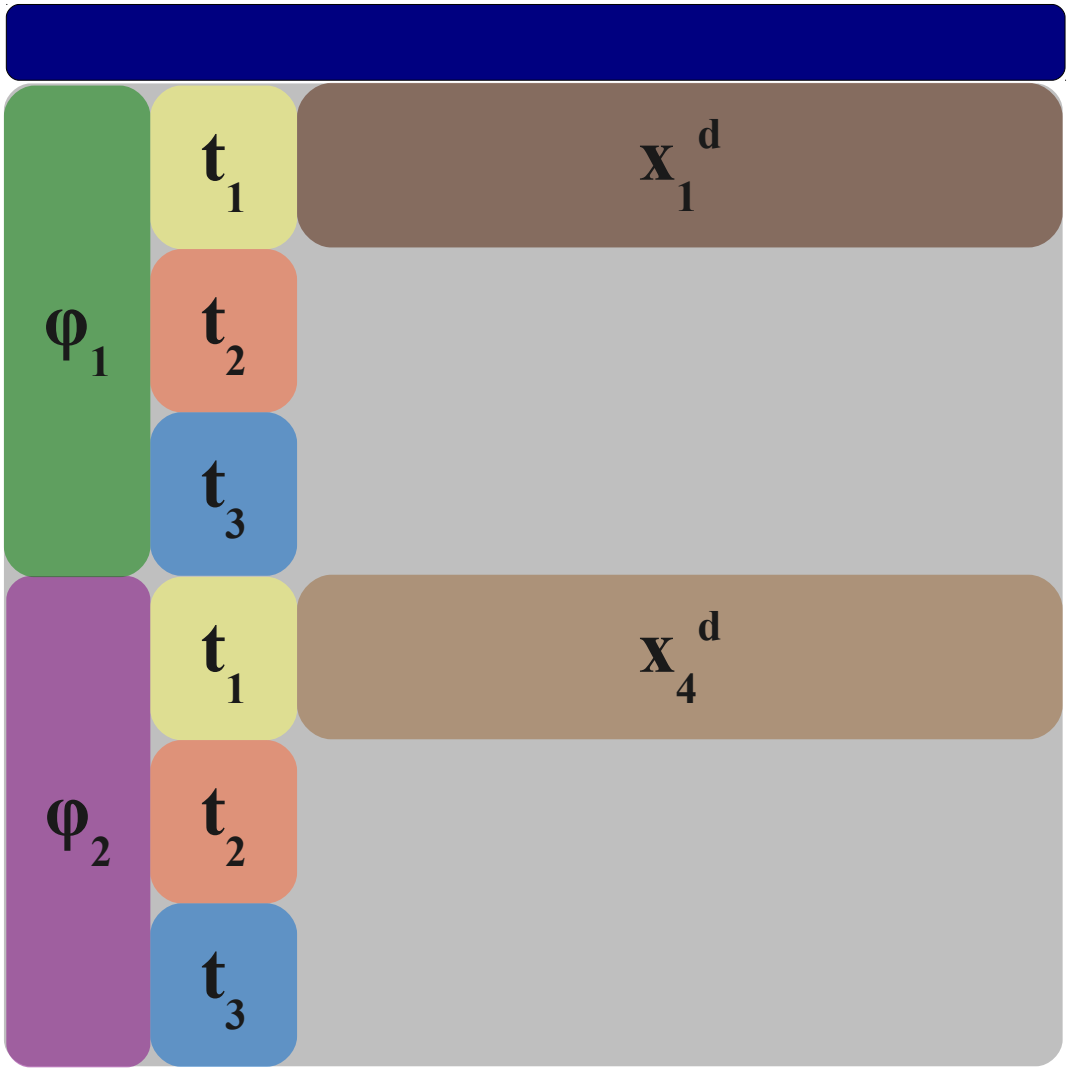
Dataset: the values for a certain number of numerical features (x^d) for multiple entities (φ) at different moments of time (t)



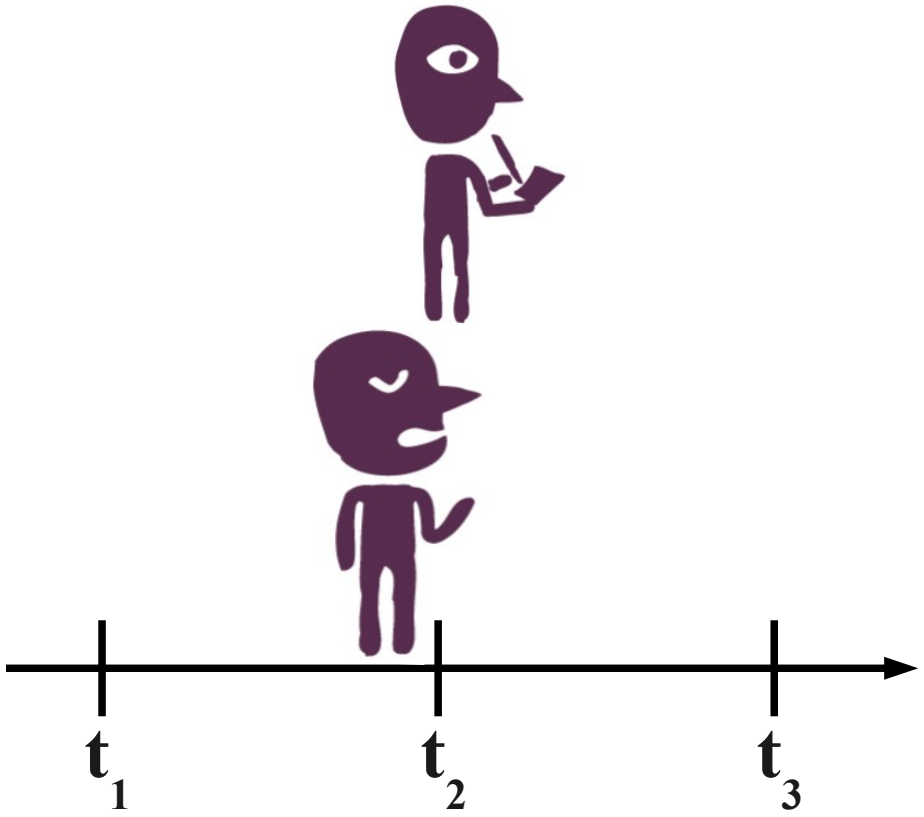


Dataset: the values for a certain number of numerical features (x^d) for multiple entities (φ) at different moments of time (t)



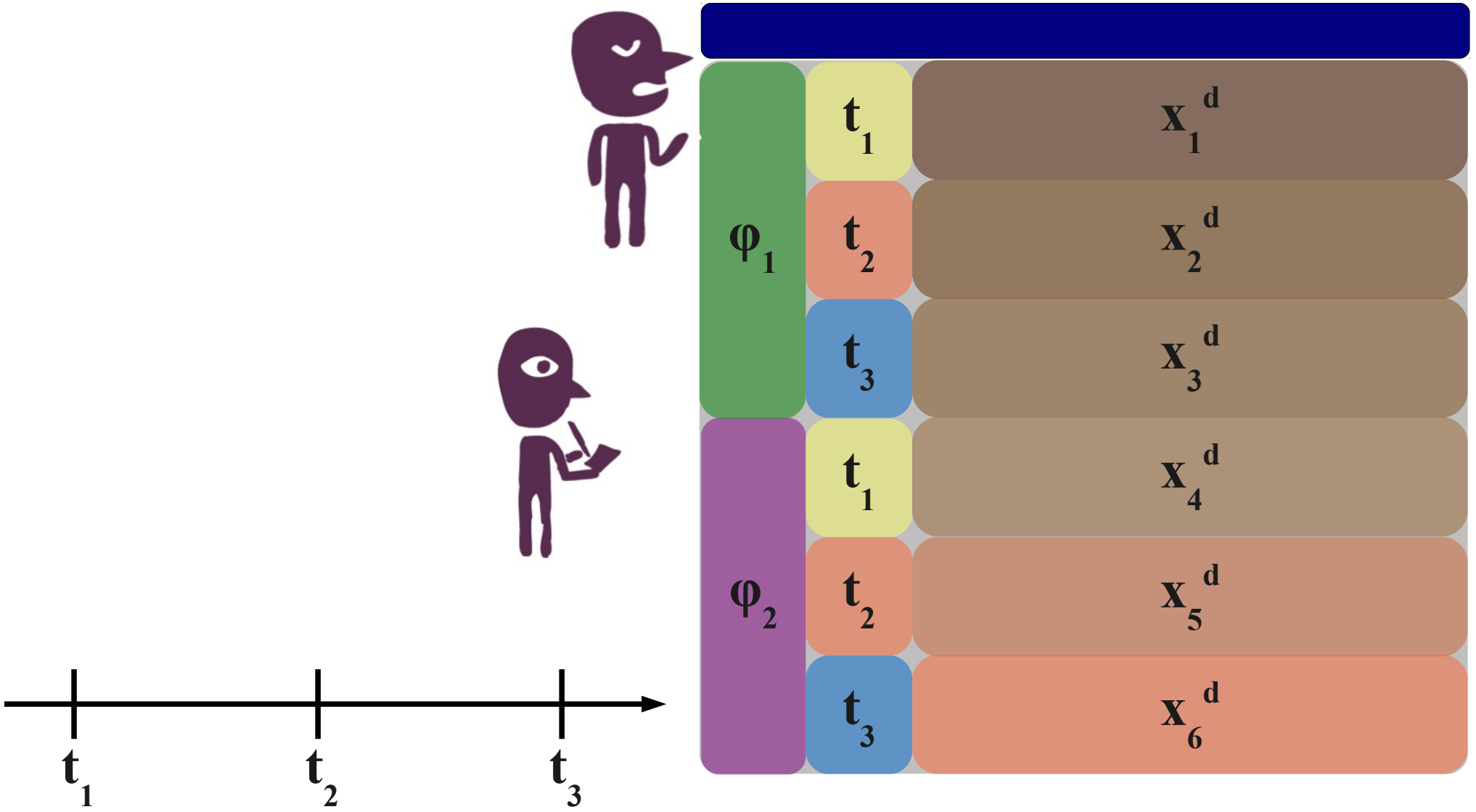


Dataset: the values for a certain number of numerical features (x^d) for multiple entities (φ) at different moments of time (t)



φ_1	t_1	x_1^d
	t_2	x_2^d
	t_3	
φ_2	t_1	x_4^d
	t_2	x_5^d
	t_3	

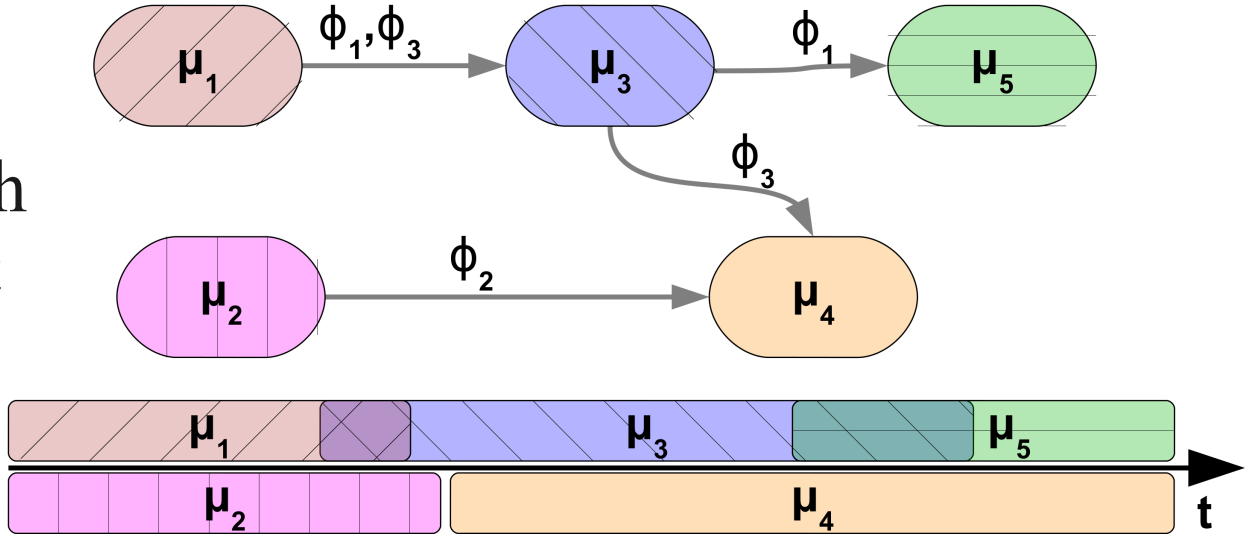
Dataset: the values for a certain number of numerical features (x^d) for multiple entities (φ) at different moments of time (t)



Goal: Detect typical evolution patterns of individuals in the dataset

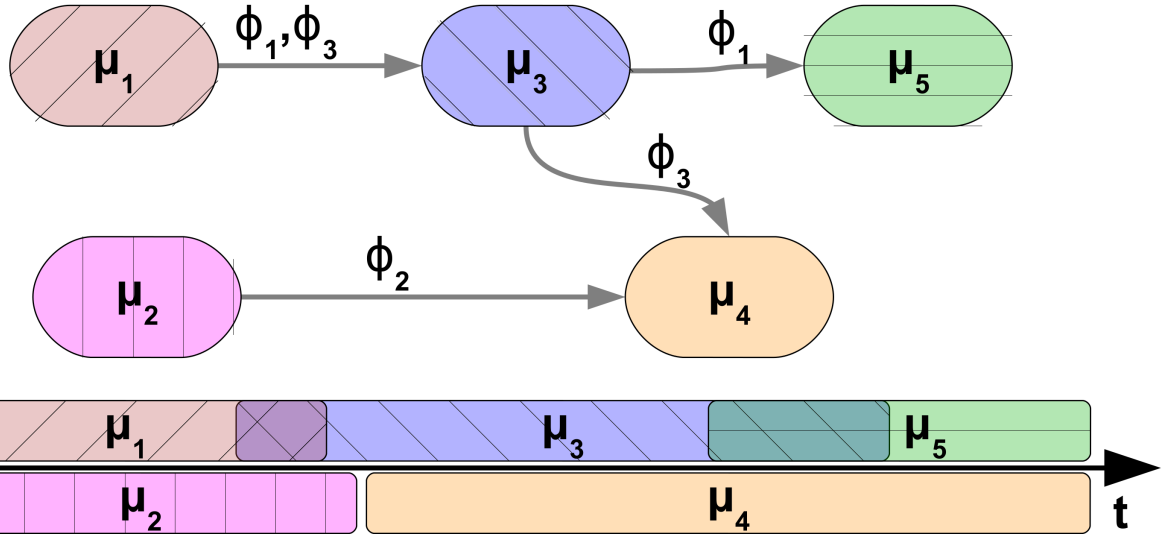
Goal: Detect typical evolution patterns of individuals in the dataset

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

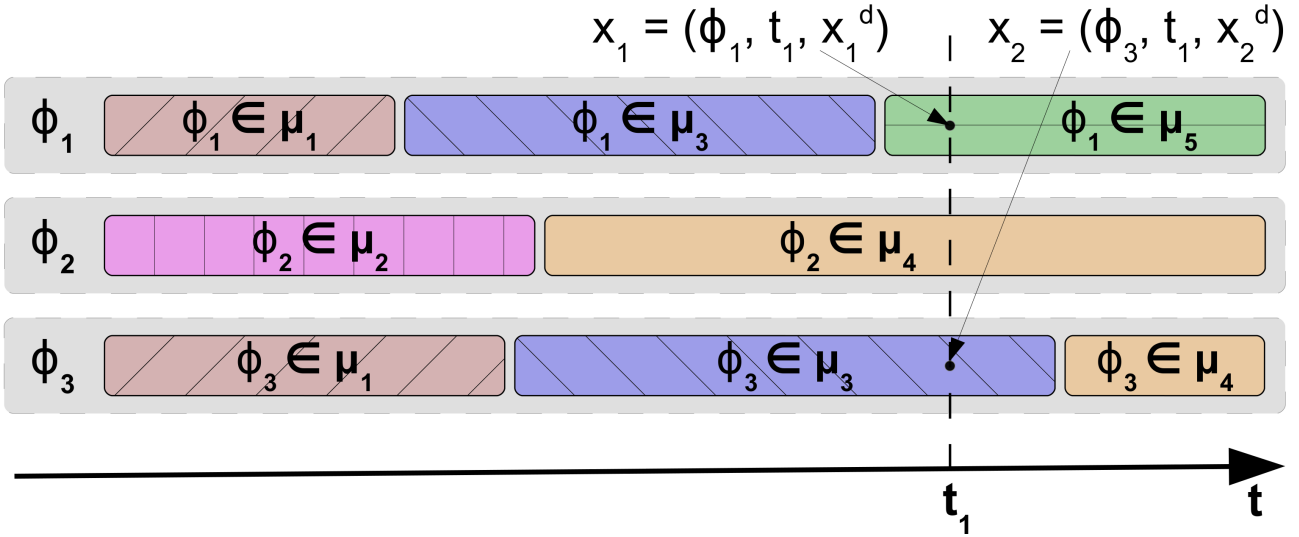


Goal: Detect typical evolution patterns of individuals in the dataset

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



b) the trajectory of entities through the different phases



Summary:

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 ⟶ 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 → pair-wise constraints → apply penalty when constraints are broken

[Wagstaff & Cardie '00]

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**

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:

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

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

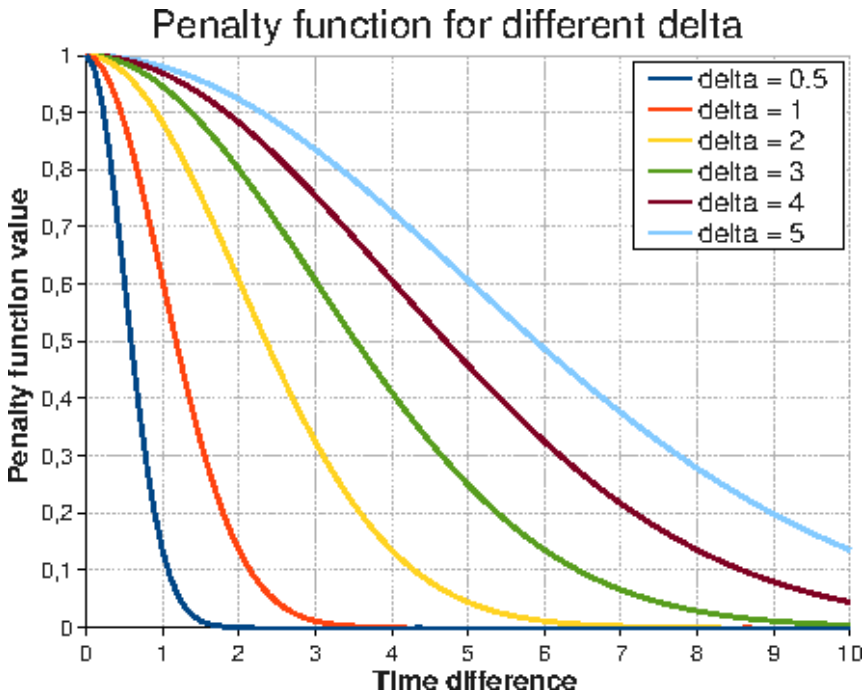
Semi-Supervised clustering *[Wagstaff & Cardie '00]* → pair-wise constraints → apply penalty when constraints are broken

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:

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

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;

} variance → { MDvar, Tvar

- continuous segmentation of observations belonging to an entity.

Shannon Entropy A → B → A → B ???

Partition evaluation measures

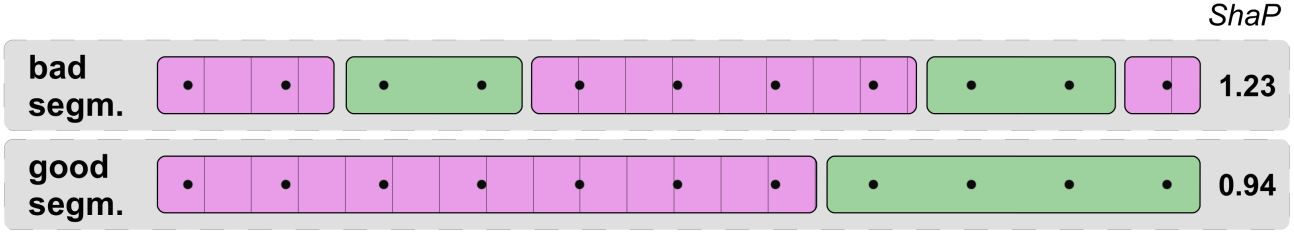
→ descriptive coherence of clusters;
→ temporal coherence of clusters;
} variance → { MDvar, Tvar

→ continuous segmentation of observations belonging to an entity.

Shannon Entropy A → B → A → B ???

Proposal: Correct the Shannon entropy to penalize changes

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



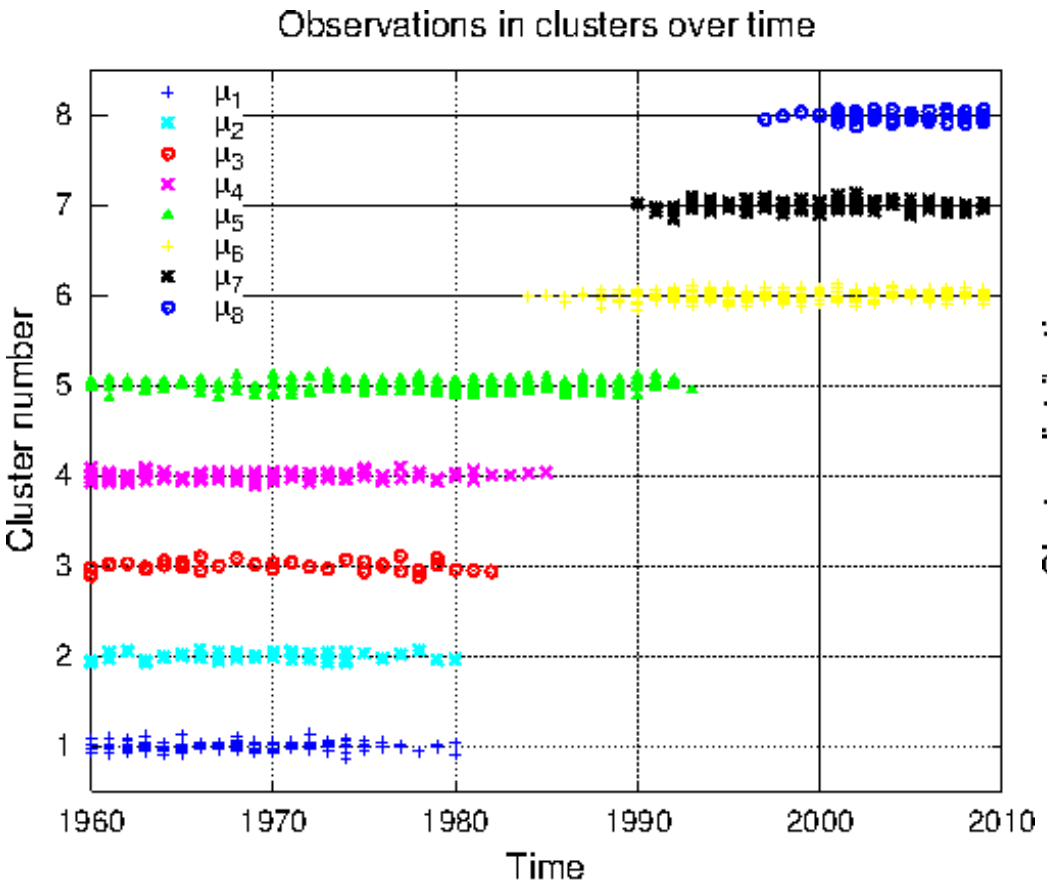
Summary:

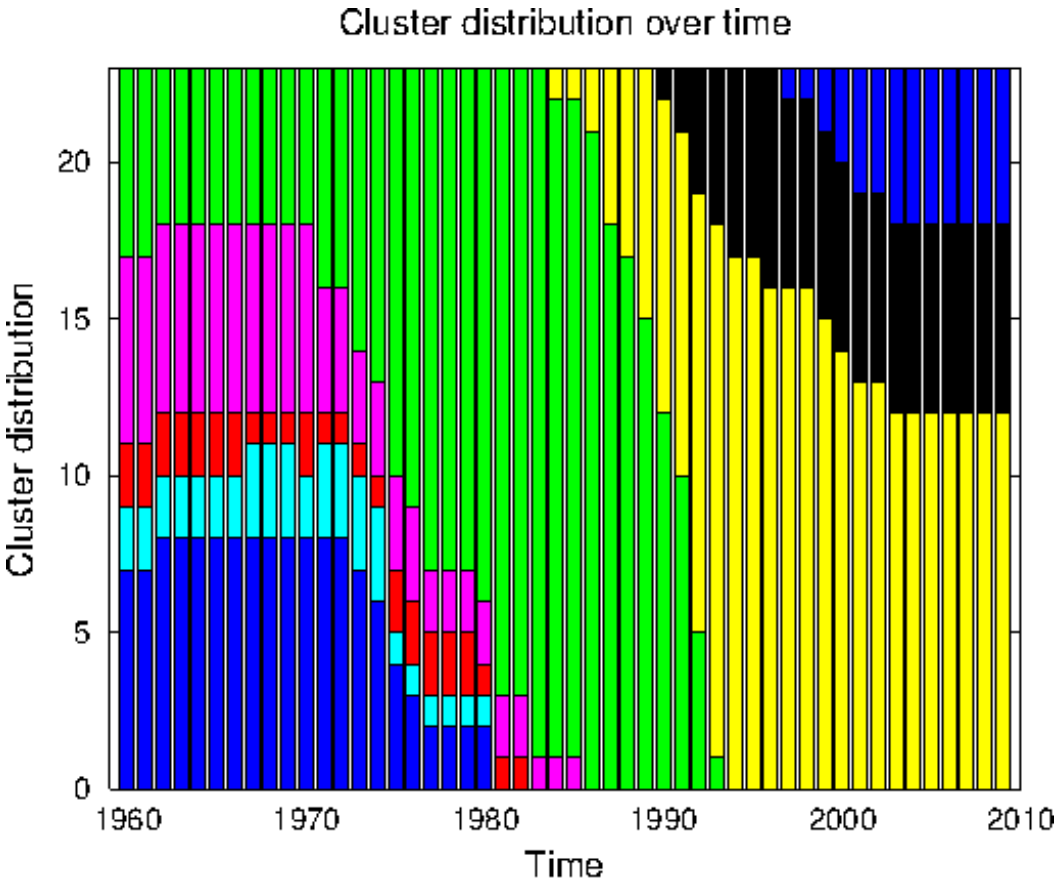
Compared Political Dataset I

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 $\delta = 3$)

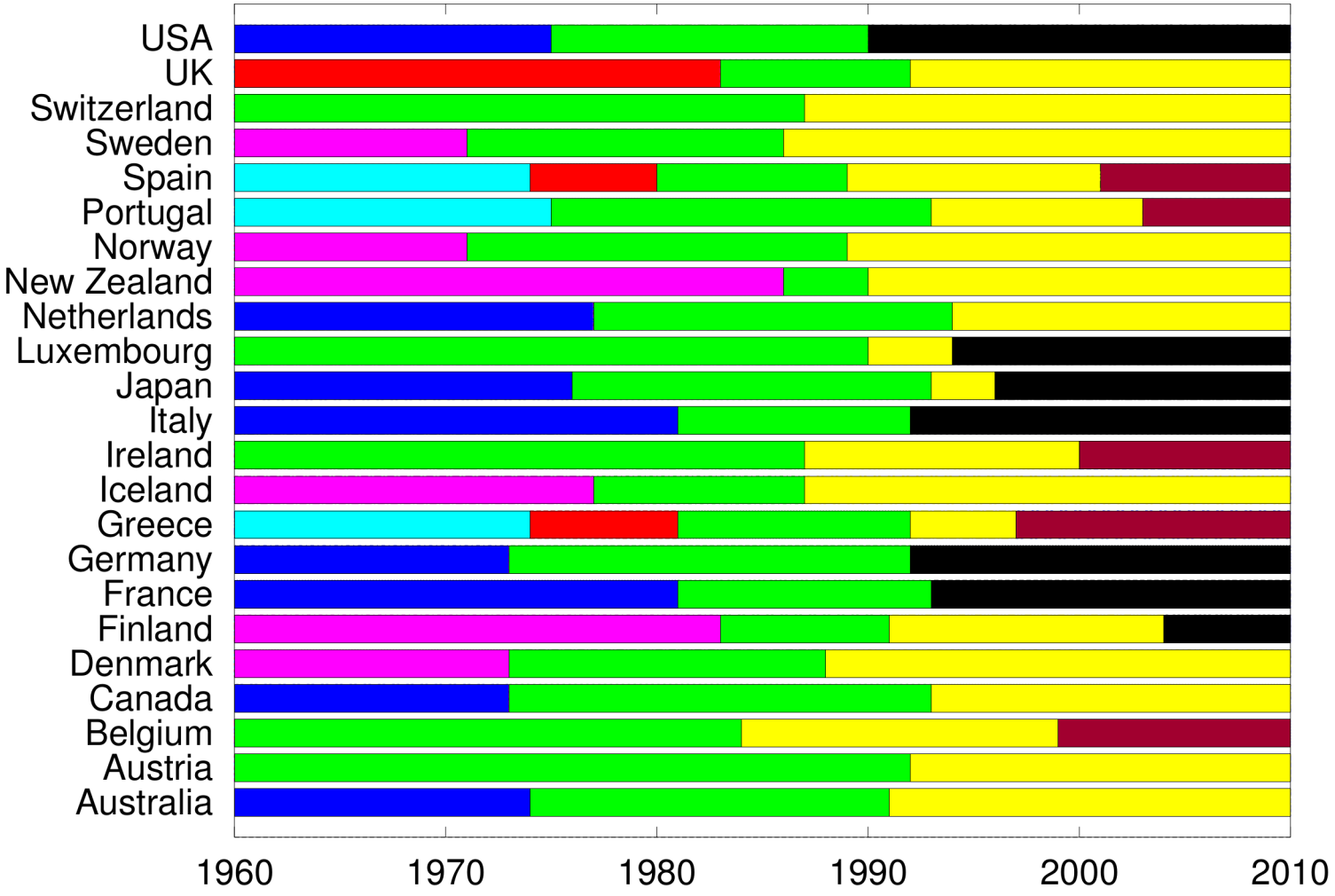




Compared Political Dataset I

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

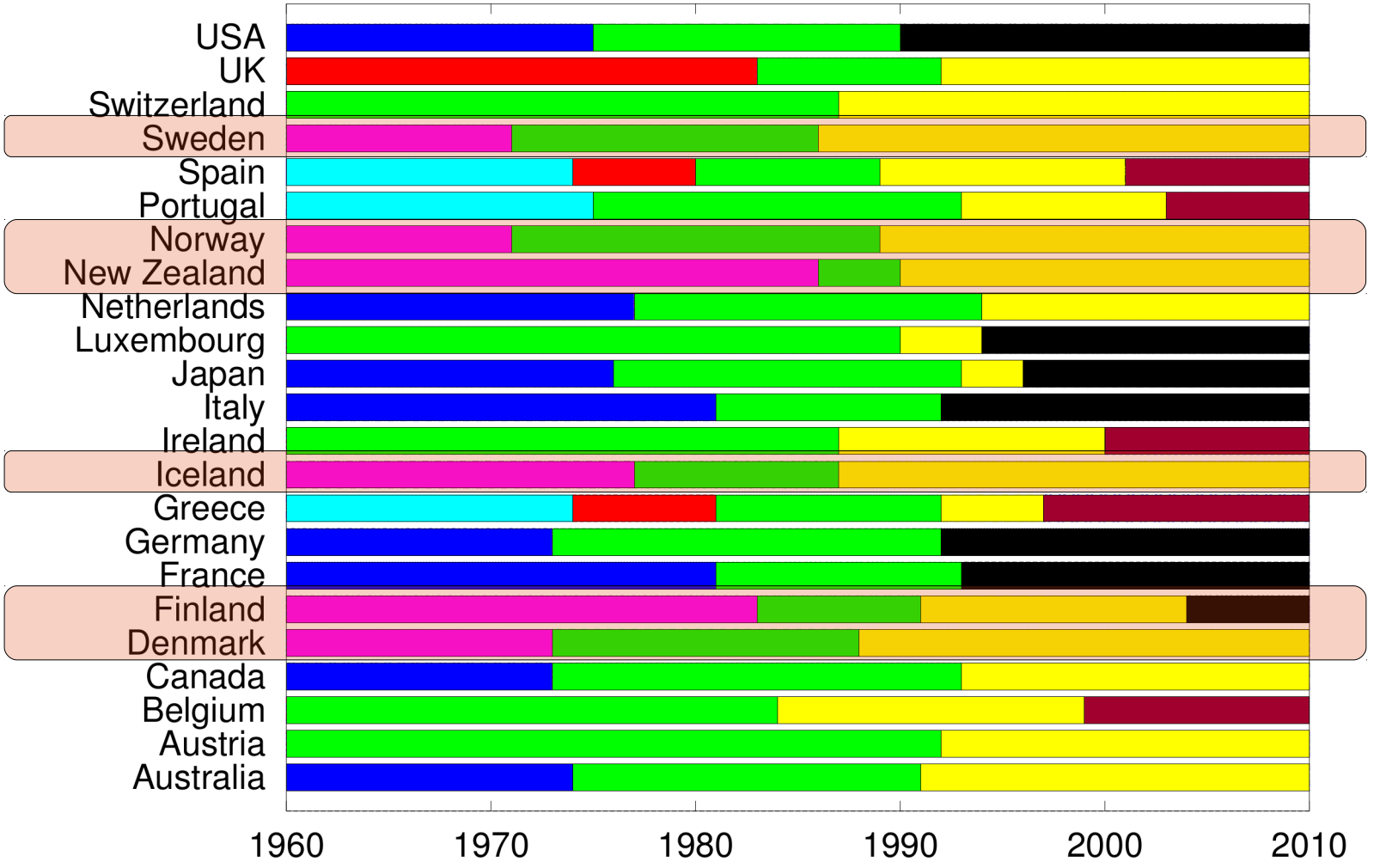
Execution TDCK-Means (8 clusters, $\beta = 0.003$ and $\delta = 3$)



Compared Political Dataset I

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

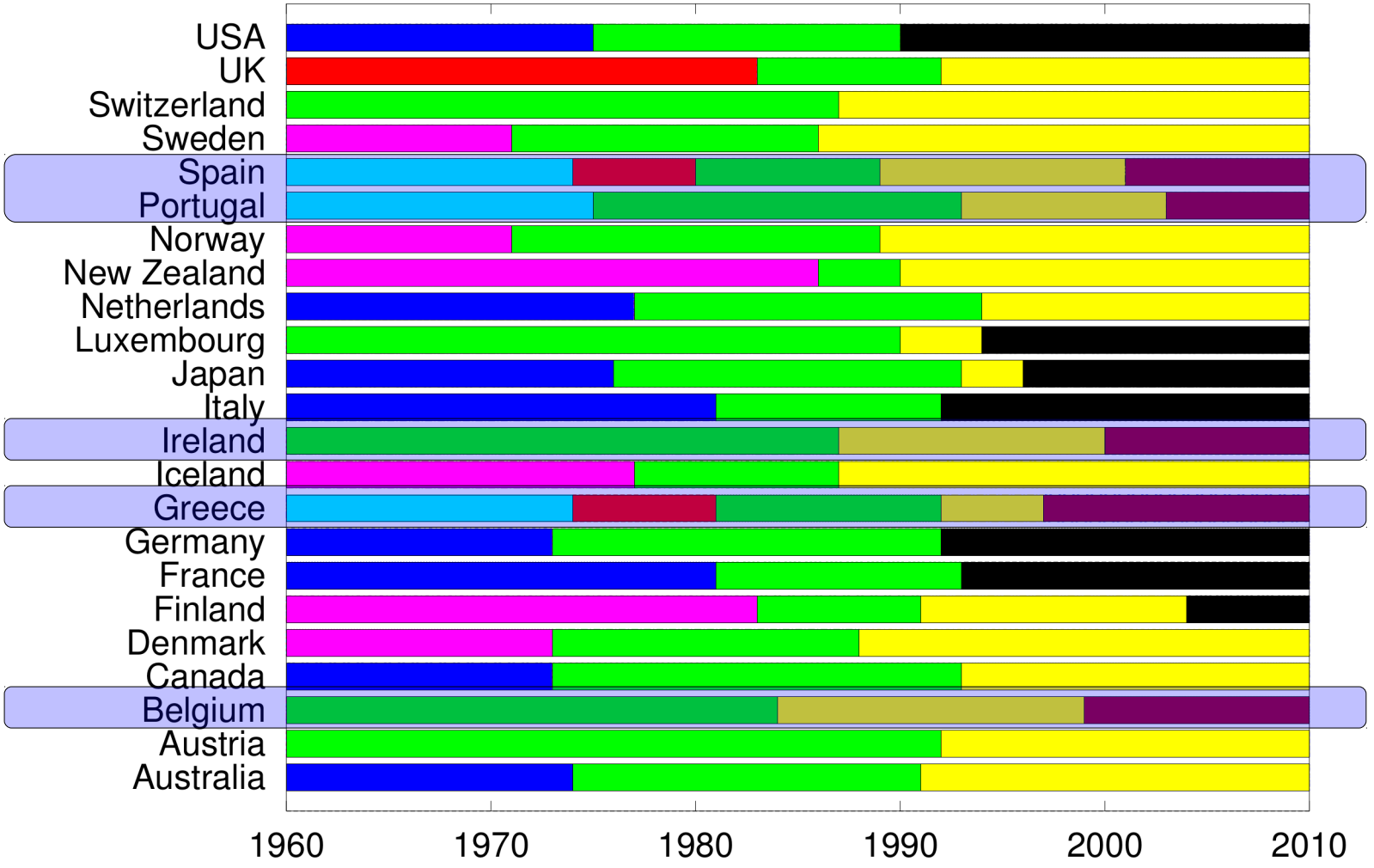
Execution TDCK-Means (8 clusters, $\beta = 0.003$ and $\delta = 3$)



Compared Political Dataset I

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

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



Quantitative evaluation

5 algorithms:

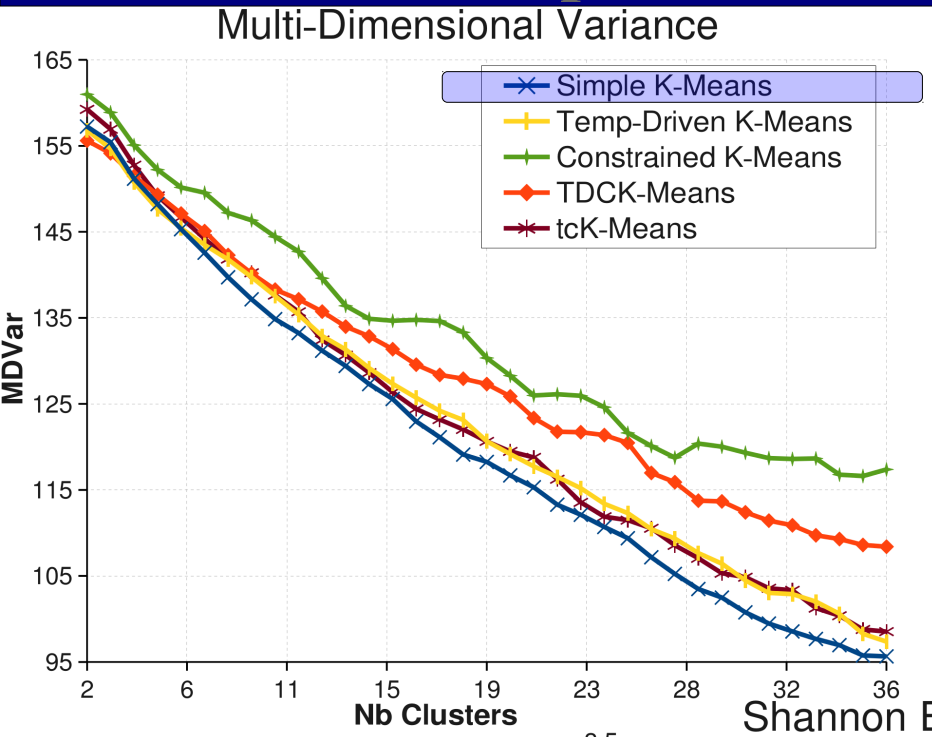
- K-Means *[MacQueen '67]*;
- tcK-Means *[Lin and Hauptmann '10]*

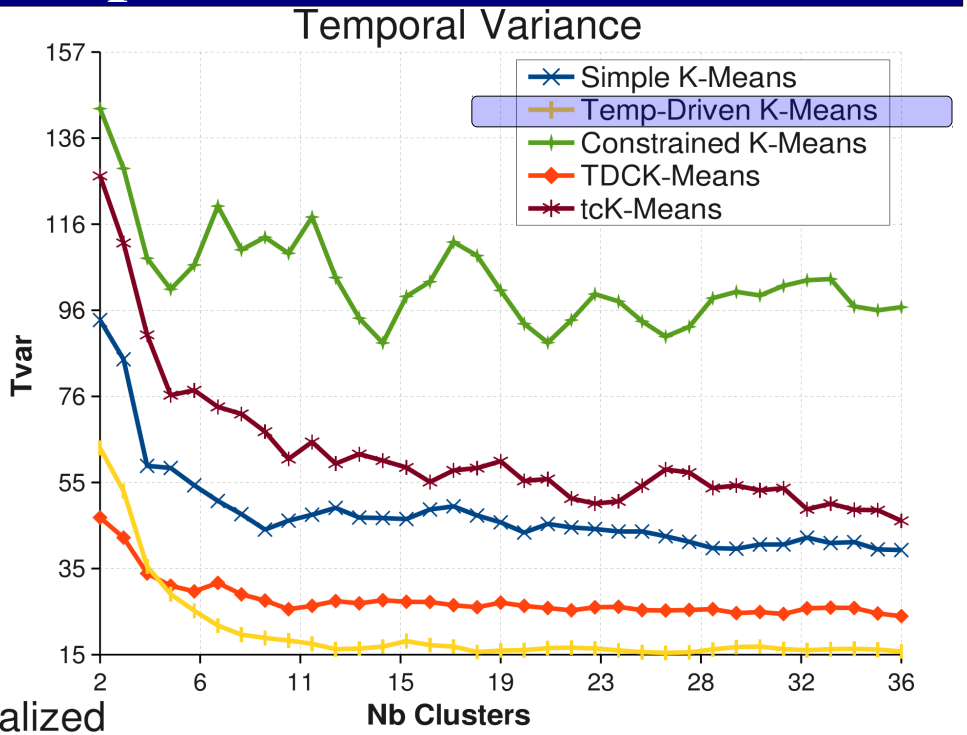
- Temporal-Driven K-Means;
  (uses Temporal-Aware Measure)
- Constrained K-Means;
  (uses Contiguity Penalty Function)

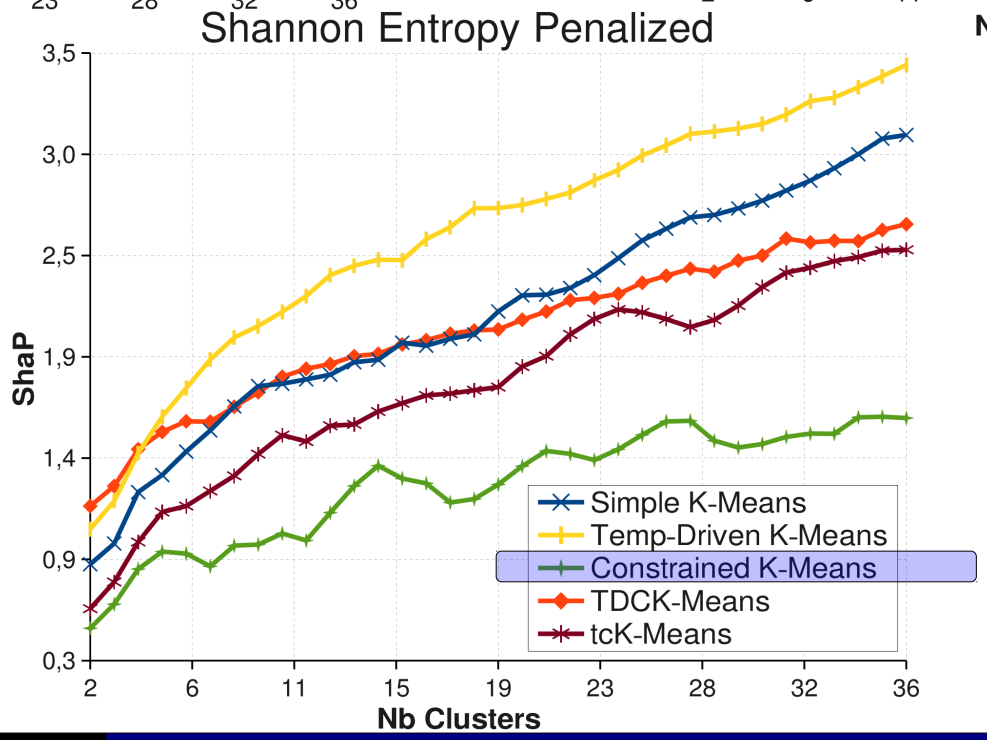
- **TDCK-Means;**
  (combines the two above)

3 measures:

- MDvar
- Tvar
- ShaP


Multi-Dimensional Variance
Simple K-Means
Temp-Driven K-Means
Constrained K-Means
TDCK-Means
tcK-Means
MDVar
Nb Clusters
165
155
145
135
125
115
105
95
2
6
11
15
19
23
28
32
36


Temporal Variance
Simple K-Means
Temp-Driven K-Means
Constrained K-Means
TDCK-Means
tcK-Means
Tvar
Nb Clusters
157
136
116
96
76
55
35
15
2
6
11
15
19
23
28
32
36


Shannon Entropy Penalized
Simple K-Means
Temp-Driven K-Means
Constrained K-Means
TDCK-Means
tcK-Means
ShaP
Nb Clusters
3,5
3,0
2,5
1,9
1,4
0,9
0,3
2
6
11
15
19
23
28
32
36

Summary:

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.*

Perspectives:

- Generating the evolution graph;
- Automatic description of generated evolution phases (clusters);
- Flexible configuring the ration between the descriptive component and the temporal component in the dissimilarity measure.

Thank you!

Questions?

Impact of parameters β and δ