Un algorithme de découverte de chroniques : une approche par le clustering.

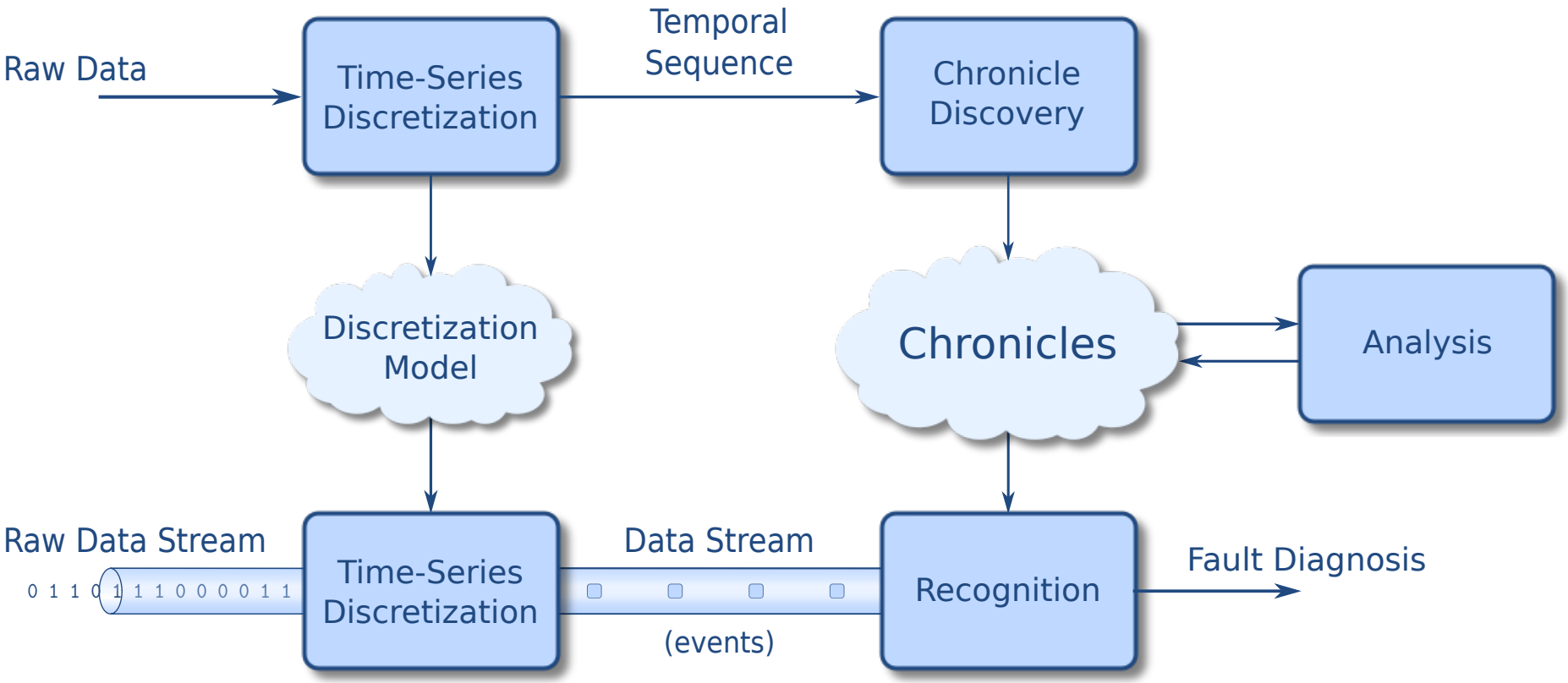
Alexandre Sahuguède

Encadré par : Euriell Le Corronc
Marie-Véronique Le Lann

LAAS-CNRS, Équipe DISCO, Université Paul Sabatier, Toulouse, FRANCE

3 Décembre 2018

Contexte


Raw Data
Time-Series Discretization
Temporal Sequence
Chronicle Discovery
Discretization Model
Chronicles
Analysis
Raw Data Stream
0 1 1 0 1 1 1 0 0 0 1 1
Time-Series Discretization
Data Stream
(events)
Recognition
Fault Diagnosis

Problématique

Problématique :

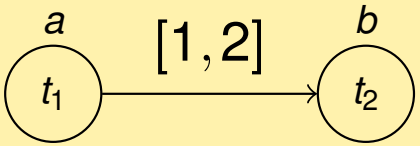
- Comment apprendre des chroniques intéressantes modélisant un ou plusieurs phénomènes, nominaux ou fautifs, à partir d'une base de données générées par le système étudié ?

Hypothèses :

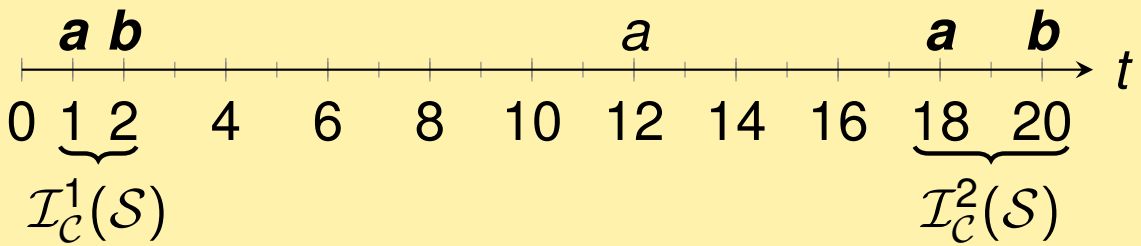
- Données d'entrées :
  - Séries temporelles multi-variées,
  - Séquence temporelle.
- Les phénomènes intéressants sont :
  - Répétés fréquemment dans les données d'entrées,
  - Sur la même granularité du temps.

Exemple de chroniques

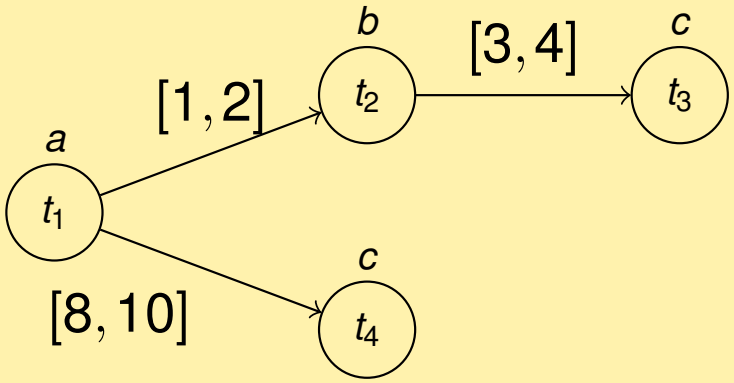
Exemple



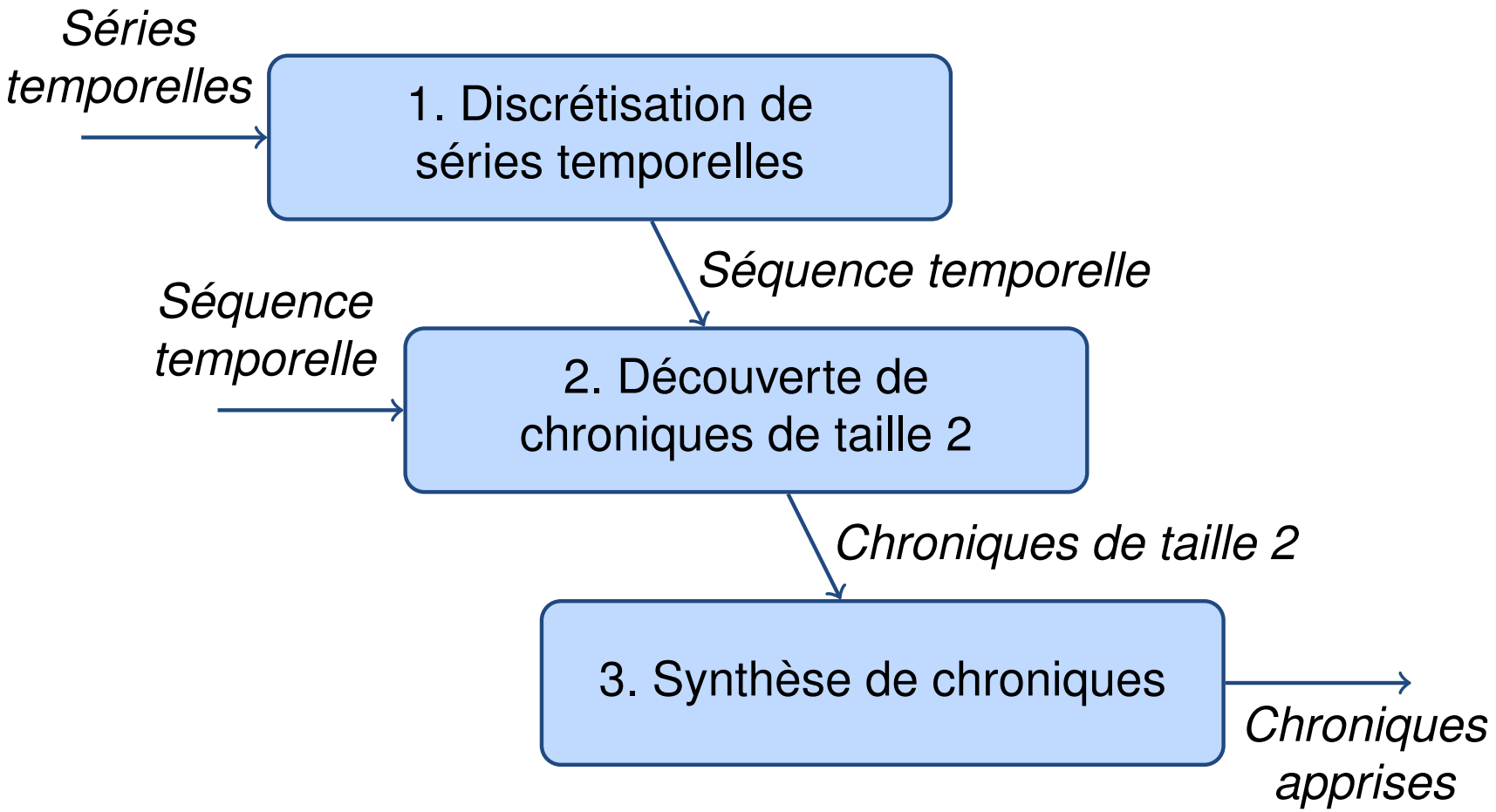
$$\mathcal{S} = \{(a,1),(b,2),(a,12),(a,18),(b,20)\}.$$



Exemple



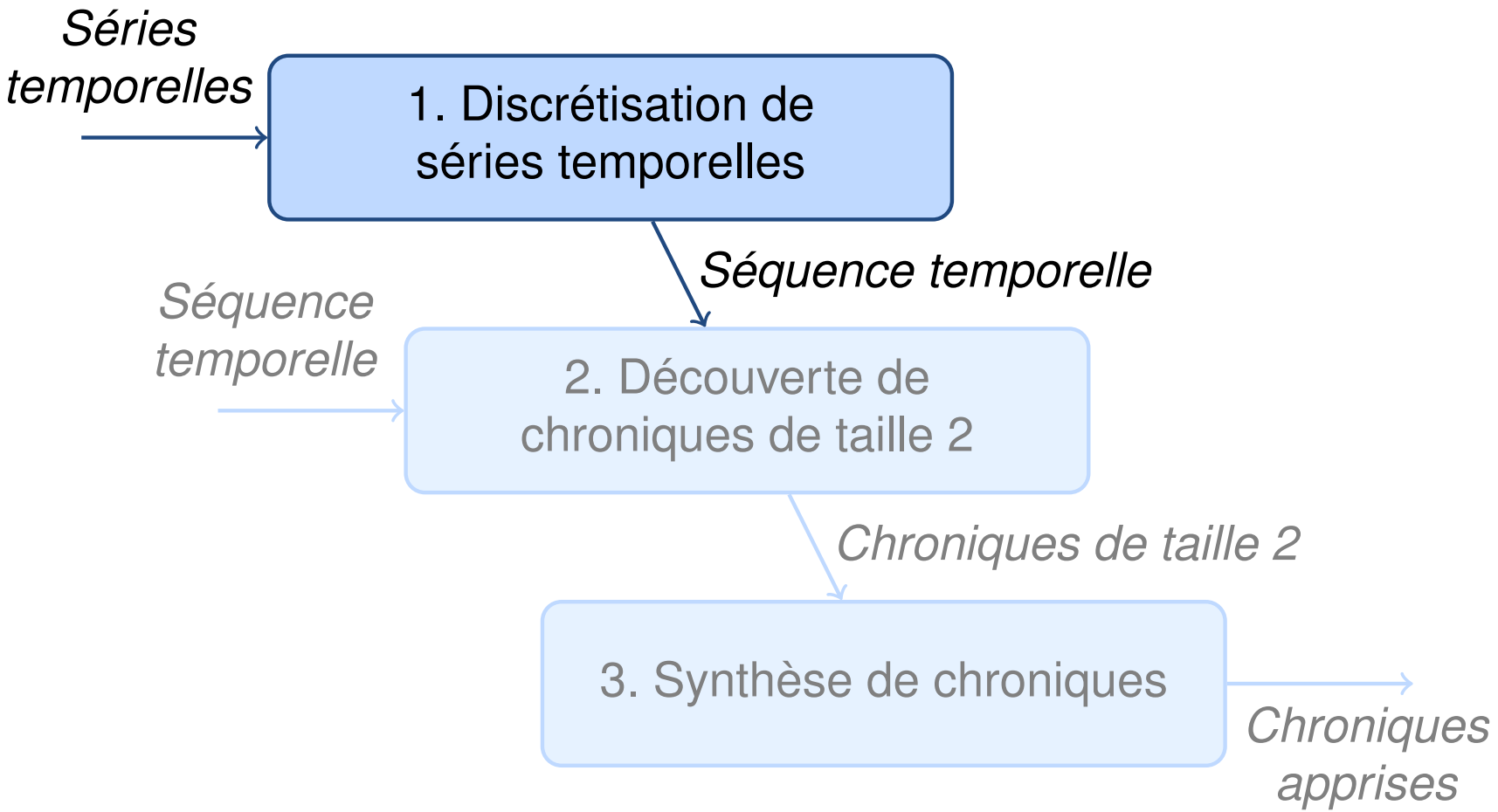
$$\mathcal{X} = \{(a,t_1),(b,t_2),(c,t_3),(c,t_4)\},$$
$$\mathcal{T} = \{\tau_{12} = x_1[1,2]x_2,$$
$$\tau_{23} = x_2[3,4]x_3,$$
$$\tau_{14} = x_1[8,10]x_4\}.$$


Séries temporelles
1. Discrétisation de séries temporelles
Séquence temporelle
Séquence temporelle
2. Découverte de chroniques de taille 2
Chroniques de taille 2
3. Synthèse de chroniques
Chroniques apprises

Introduction

Exemple du nageur

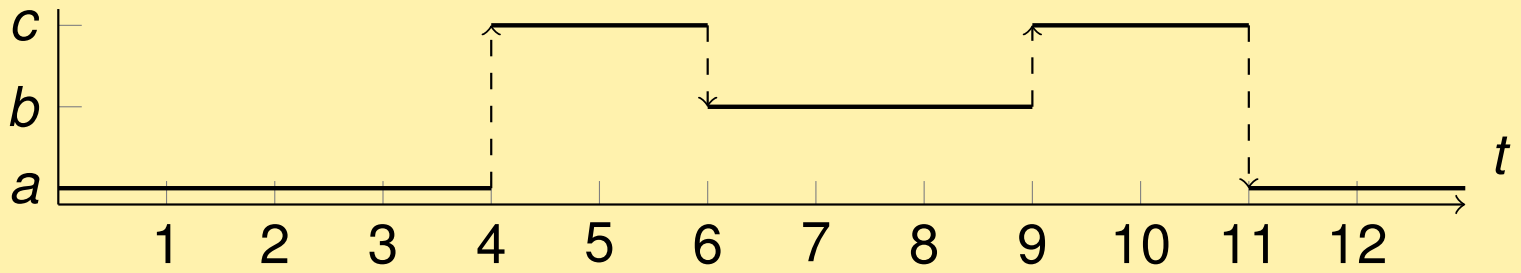
- Gant instrumenté pour la supervision de performance d'un nageur,
- Enregistrements des mouvements par 16 capteurs $\rightarrow$ 16 attributs,
- Fréquence d'échantillonnage de $50Hz$,
- Enregistrement d'un athlète nageant le crawl sur $50m$.


Séries temporelles
1. Discrétisation de séries temporelles
Séquence temporelle
Séquence temporelle
2. Découverte de chroniques de taille 2
Chroniques de taille 2
3. Synthèse de chroniques
Chroniques apprises

Étape 1 : Discrétisation

Un événement $x = (e, t)$ est interprété comme l'instant t où les données commencent à être assignées à la classe e.

Exemple



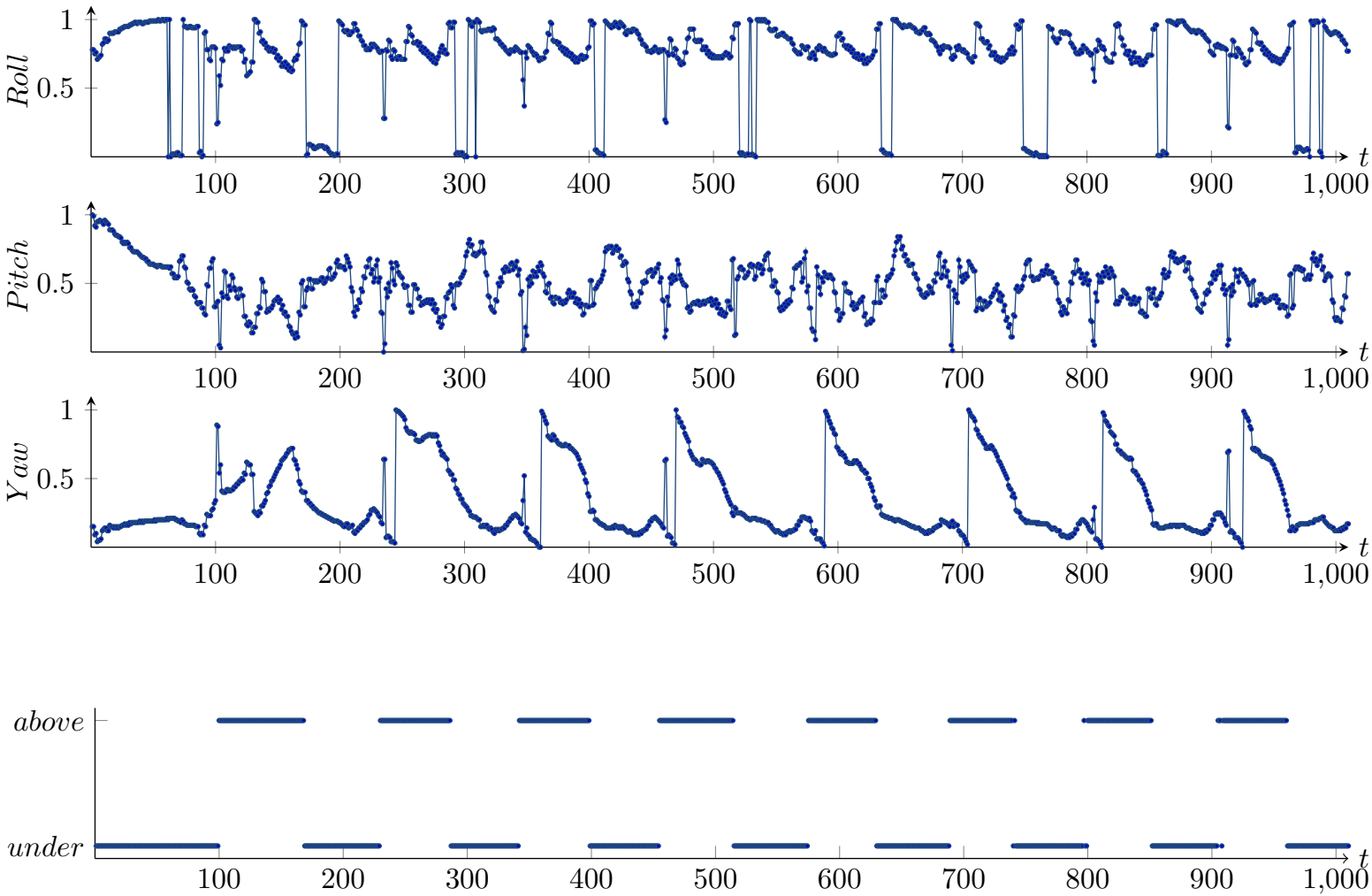
$$\Downarrow$$

$$\mathcal{S} = \{(a,0),(c,4),(b,6),(c,9),(a,11)\}$$

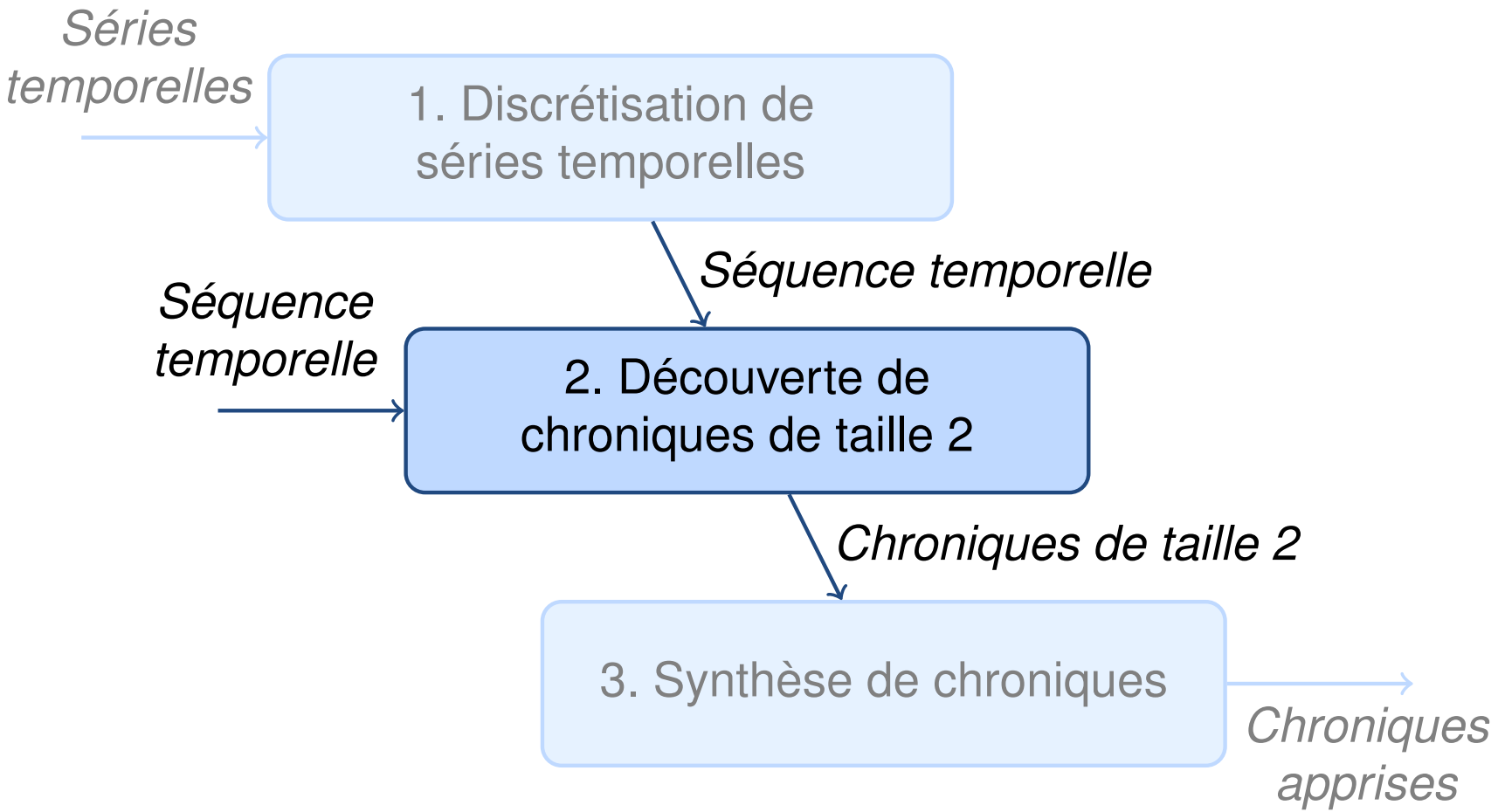
Algorithme utilisé : LAMDA

- Algorithme basé sur la logique floue,
- Peut fonctionner en auto-apprentissage (clustering),
- P3S $\rightarrow$ implémentation de LAMDA.

Exemple du nageur



$$\Rightarrow \mathcal{S} = \{(\mathit{above}, 0), (\mathit{under}, 100), (\mathit{above}, 169), (\mathit{under}, 230), (\mathit{above}, 287), \ldots\}$$


Séries temporelles
1. Discrétisation de séries temporelles
Séquence temporelle
Séquence temporelle
2. Découverte de chroniques de taille 2
Chroniques de taille 2
3. Synthèse de chroniques
Chroniques apprises

Étape 2 (1/2) : Découverte

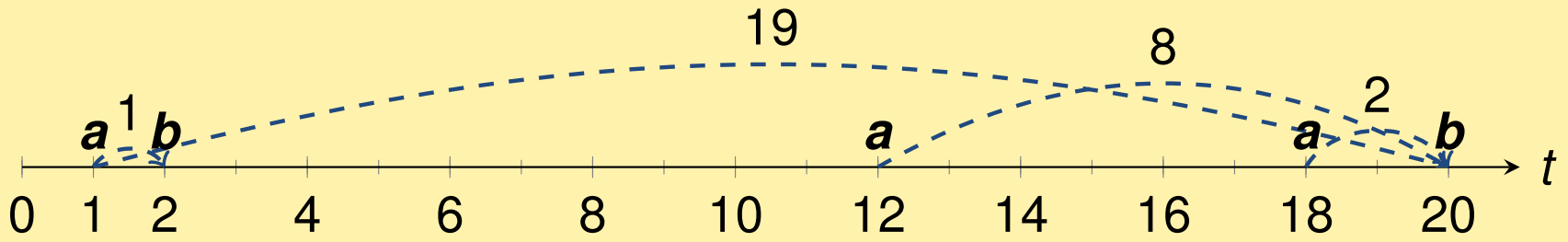
Définition : Occurrences & distances temporelles d'une paire

Soit $\mathcal{S}$ une séquence temporelle et (a, b) une paire de type d'événements tel que $a, b \in E_{\mathcal{S}}$.

$$\mathcal{O}_{ab} = \{\ \langle (a, t_i), (b, t_j) \rangle \mid \forall i, j,\ t_i < t_j, (a, t_i), (b, t_j) \in \mathcal{S}\ \}.$$

$$\mathcal{D}_{ab} = \{\ (t_j - t_i) \mid \langle (a, t_i), (b, t_j) \rangle \in \mathcal{O}_{ab}\ \}.$$

Exemple



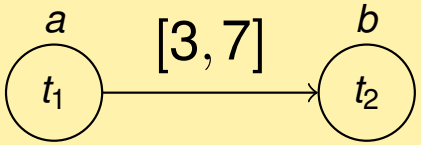
$$\mathcal{O}_{ab} = \{\langle (a,1),(b,2)\rangle, \langle (a,1),(b,20)\rangle, \langle (a,12),(b,20)\rangle, \langle (a,18),(b,20)\rangle\},$$
$$\mathcal{D}_{ab} = \{\quad 1, \quad 19, \quad 8, \quad 2 \quad\}.$$

Étape 2 (2/2) : Découverte

Propriété

À partir de l'ensemble des distances temporelles $\mathcal{D}_{ab}$ d'une paire (a, b), il est possible d'obtenir une chronique $\mathcal{C} = (\mathcal{X}, \mathcal{T})$ de taille 2 avec $\mathcal{X} = \{a, b\}$, $\mathcal{T} = \{\tau_{12} = x_1[\min\{\mathcal{D}_{ab}\}, \max\{\mathcal{D}_{ab}\}]x_2\}$ donné par la borne inférieure et supérieure de $\mathcal{D}_{ab}$.

Exemple



$\mathcal{D}_{ab} = \{4, 5, 3, 7\} \quad \Rightarrow$

$\mathcal{X} = \{x_1 = (a, t_1),\ x_2 = (b, t_2)\},$

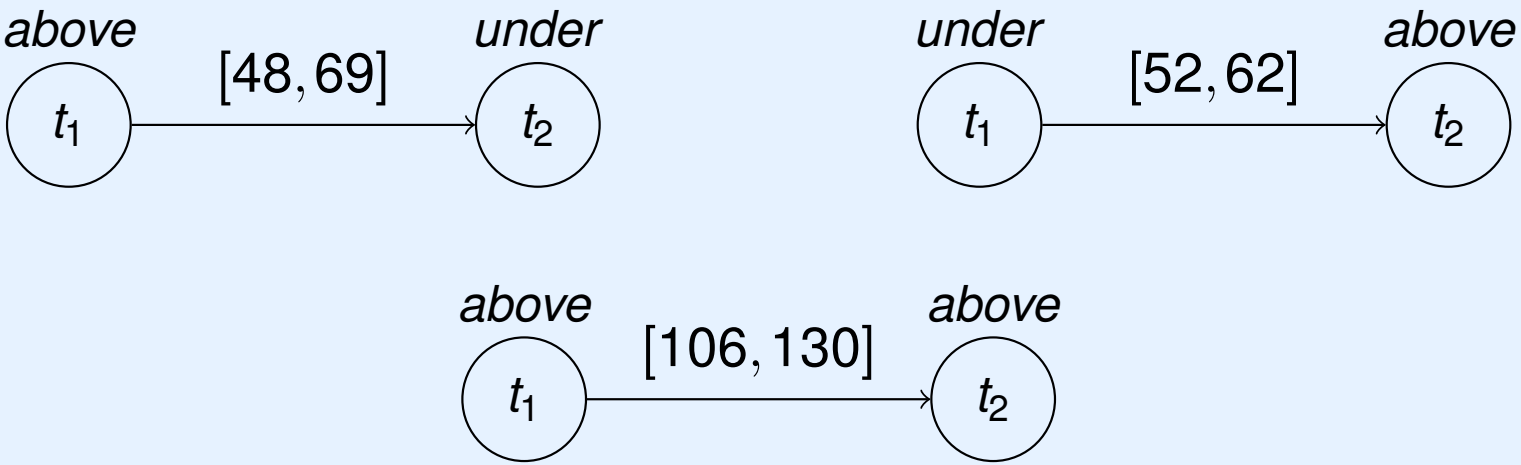
$\mathcal{T} = \{\tau_{12} = x_1[3, 7]x_2\}.$

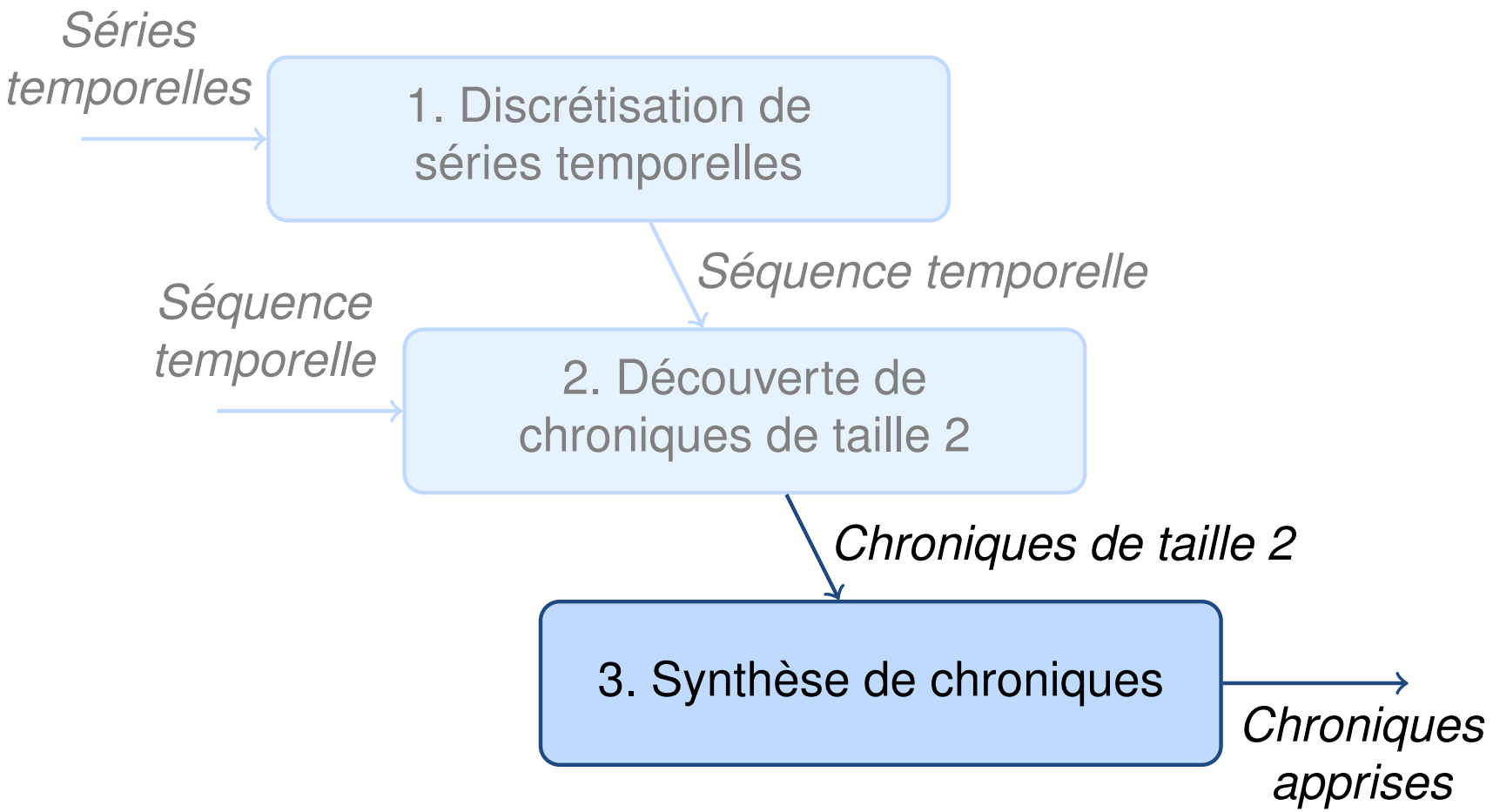
Exemple du nageur

Résultats

- 62 chroniques de taille 2 ont été découvertes,
- Fréquences variant de 2 à 17.

Avec $f = 17$:



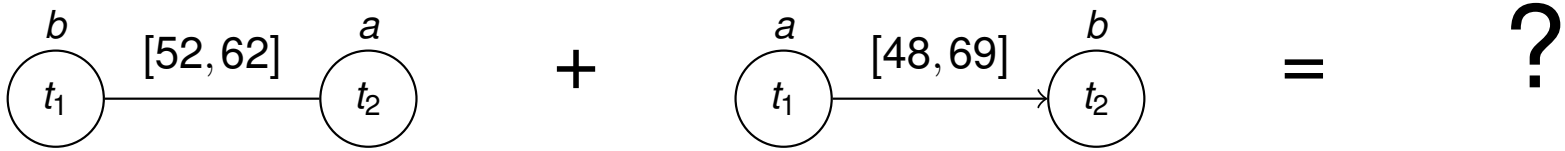

Séries temporelles
1. Discrétisation de séries temporelles
Séquence temporelle
Séquence temporelle
2. Découverte de chroniques de taille 2
Chroniques de taille 2
3. Synthèse de chroniques
Chroniques apprises

Étape 3 (1/3) : Synthèse

De la découverte de chroniques de taille 2 à la synthèse de chroniques

- Découverte de chroniques de taille 2 $\rightarrow$ nombreux petit schémas temporel,
- Sur quel critère peut-on dire que deux chroniques représentent deux sous-ensembles du même phénomène ?

$\rightarrow$ Indice de Jaccard.



Étape 3 (2/3) : Synthèse

Définition : Occurrences d'un événement

Soit $\mathcal{C} = (\mathcal{X}, \mathcal{T})$ une chronique et $\mathcal{S}$ une séquence temporelle, $\mathcal{O}_i$ est l'ensemble des temps d'occurrences de l'événement x_i dans toutes les instances de chroniques $\mathcal{I}_{\mathcal{C}}(\mathcal{S})$.

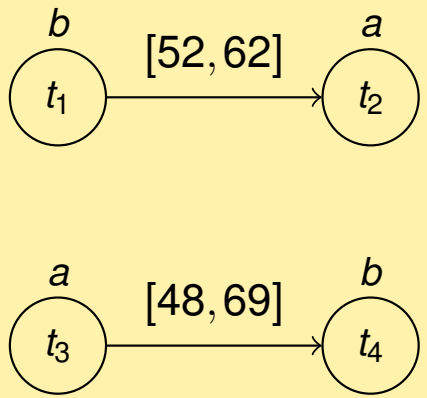
$$\mathcal{O}_i = \{\ t_i \mid \forall \mathcal{I}_{\mathcal{C}}(\mathcal{S}), x_i = (e, t_i) \in \mathcal{X}\ \}.$$

Définition : Indice de Jaccard

L'indice de Jaccard entre deux événements x_i et x_j est donné par :

$$J(x_i, x_j) = \frac{|\mathcal{O}_i \cap \mathcal{O}_j|}{|\mathcal{O}_i \cup \mathcal{O}_j|}.$$

Exemple



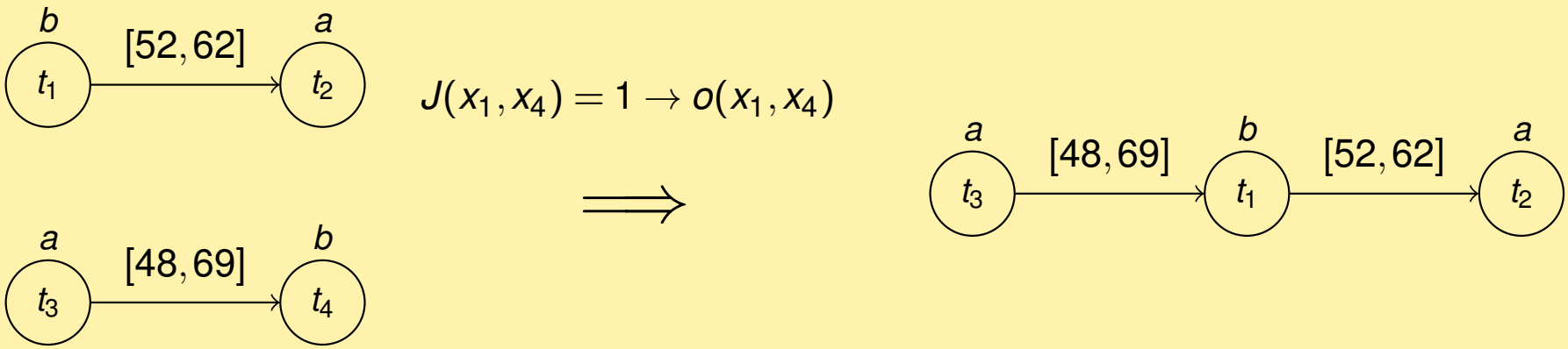
$$\mathcal{O}_1 = \{169, 287, \ldots, 1845, 1962\},$$
$$\mathcal{O}_4 = \{169, 287, \ldots, 1845, 1962\},$$
$$J(x_1, x_4) = \frac{|\mathcal{O}_1 \cap \mathcal{O}_4|}{|\mathcal{O}_1 \cup \mathcal{O}_4|} = \frac{17}{17} = 1,$$
$$\mathcal{O}_2 = \{230, 342, \ldots, 1903, 2017\},$$
$$\mathcal{O}_3 = \{100, 230, \ldots, 1797, 1903\},$$
$$J(x_2, x_3) = \frac{|\mathcal{O}_2 \cap \mathcal{O}_3|}{|\mathcal{O}_2 \cup \mathcal{O}_3|} = \frac{16}{18} = 0.88.$$

Étape 3 (3/3) : Synthèse

Définition : Opération

Une opération est la fusion de 2 nœuds d'une chronique. Une opération est appliquée lorsque l'indice de Jaccard est au dessus d'un seuil défini *minJac*.

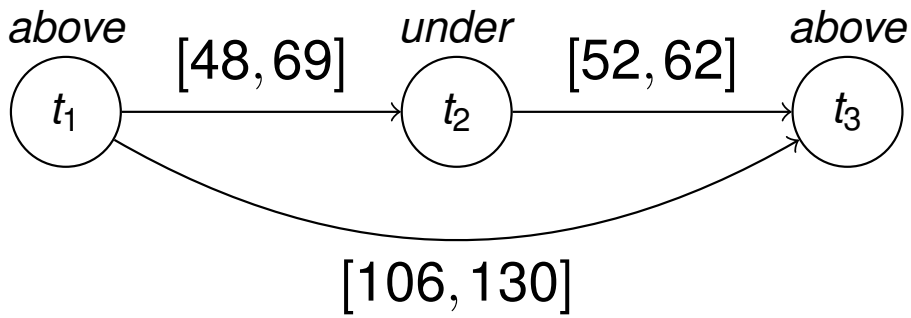
Exemple



Exemple du nageur

Résultats

- Avec $minJac = 1$,
- 16 chroniques avec leurs tailles variant entre 3 et 6.



Physiquement :

$$\begin{aligned} \text{Phase au dessus de l'eau} &\rightarrow [0.96, 1.38]\ \textit{secondes}, \\ \text{Phase sous l'eau} &\rightarrow [1.04, 1.24]\ \textit{secondes}, \\ \text{Séquence complète} &\rightarrow [2.12, 2.6]\ \textit{secondes}. \end{aligned}$$

Pourquoi un ordre sur les opérations ?

Avec le seuil $minJac = 1$

- Résultat unique,
- Hypothèse forte,
- Problème de robustesse.

Avec le seuil $minJac < 1$

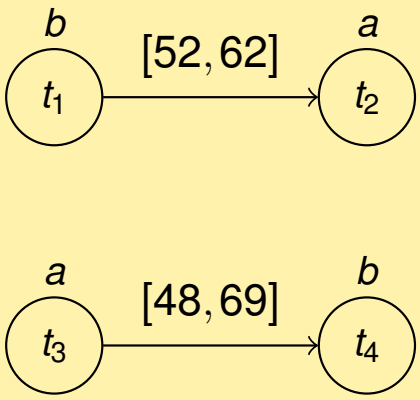
- Certaine opérations ne sont pas applicables (ne respecte pas le formalisme des chroniques),
- Le résultat n'est plus unique,
- L'ordre des opérations est importante,
- Amélioration de la robustesse.

Types d'opérations

Indices de Jaccard entre 2 chroniques

- 4 indices à calculer,
- Divisées en 3 types :
  - Type 1 : entre 2 nœuds sources,
  - Type 2 : entre 2 nœuds destinations,
  - Type 3 : entre 1 nœud source et 1 nœud destination.

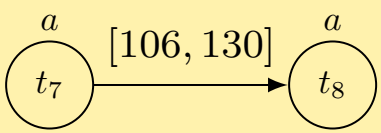
Exemple

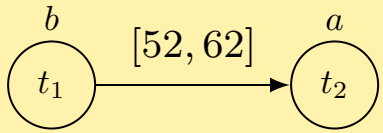


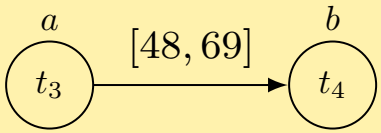
- Type 1 (2 nœuds sources) :
  - $J(x_1, x_3) = 0$.
- Type 2 (2 nœuds destinations) :
  - $J(x_2, x_4) = 0$.
- Type 3 (1 nœud source et 1 nœud destination) :
  - $J(x_2, x_3) = 0.88$,
  - $J(x_4, x_1) = 1$.

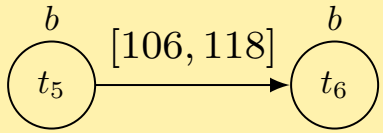
Exemple d'ordre des opérations (1/7)

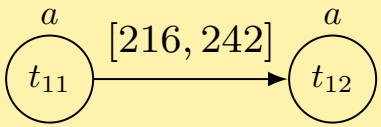
Exemple : Un ordre possible

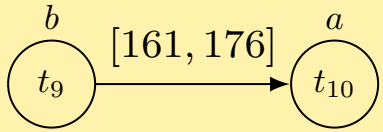












1 - Type 1 (2 nœuds sources),

↓

2 - Type 3 (1 nœud source et 1 nœud destination),

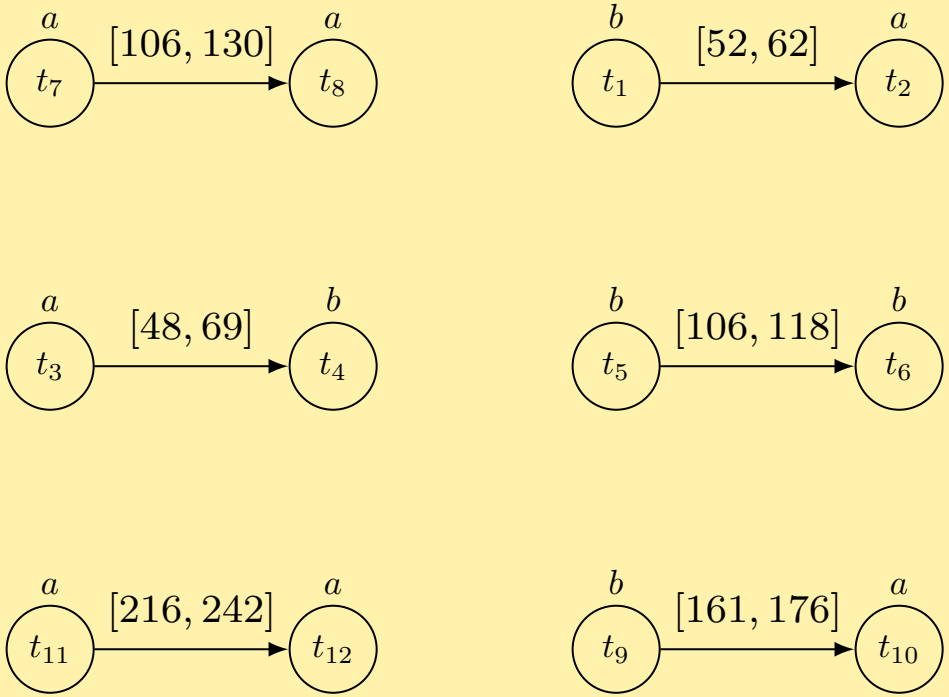
↓

3 - Type 2 (2 nœuds destinations).

- 60 indices de Jaccard calculés,
- 17 au dessus du seuil $minJac = 0.9$.

Exemple d'ordre des opérations (2/7)

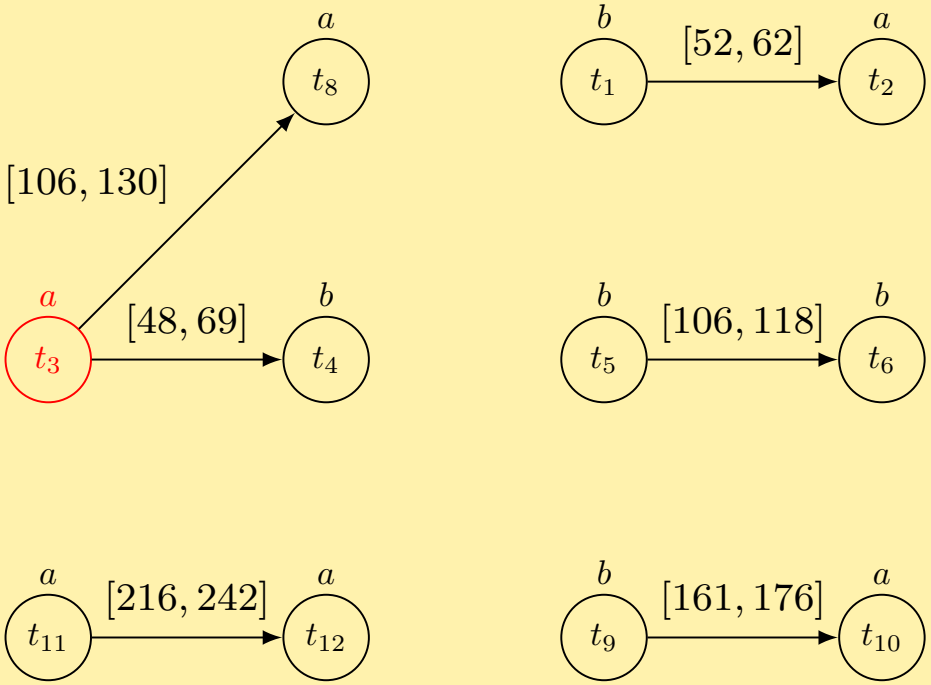
Exemple : 1 - Type 1 (2 nœuds sources)



$$J(x_3, x_7) = 1,$$
$$J(x_5, x_9) = 1,$$
$$J(x_1, x_5) = 0.94,$$
$$J(x_1, x_9) = 0.94,$$
$$J(x_3, x_{11}) = 0.94,$$
$$J(x_7, x_{11}) = 0.94.$$

Exemple d'ordre des opérations (3/7)

Exemple : 1 - Type 1 (2 nœuds sources)



$J(x_3, x_7) = 1,$

$J(x_5, x_9) = 1,$

$J(x_1, x_5) = 0.94,$

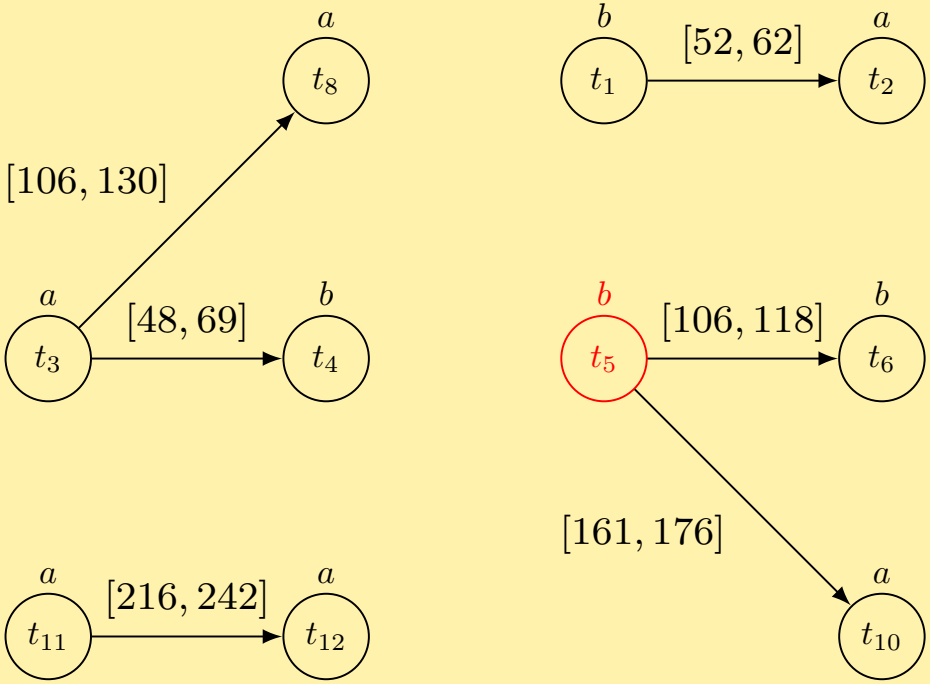
$J(x_1, x_9) = 0.94,$

$J(x_3, x_{11}) = 0.94,$

~~$J(x_3, x_{11}) = 0.94.$~~

Exemple d'ordre des opérations (4/7)

Exemple : 1 - Type 1 (2 nœuds sources)



$J(x_3, x_7) = 1,$

$J(x_5, x_9) = 1,$

$J(x_1, x_5) = 0.94,$

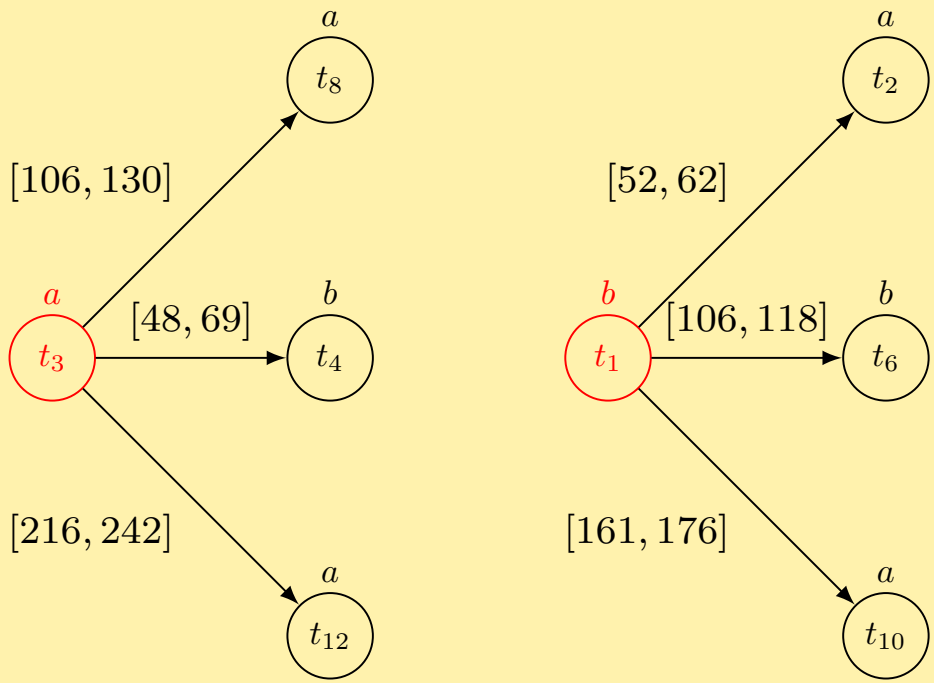
~~$J(x_1, x_5) = 0.94,$~~

$J(x_3, x_{11}) = 0.94,$

~~$J(x_3, x_{11}) = 0.94.$~~

Exemple d'ordre des opérations (5/7)

Exemple : 1 - Type 1 (2 nœuds sources)

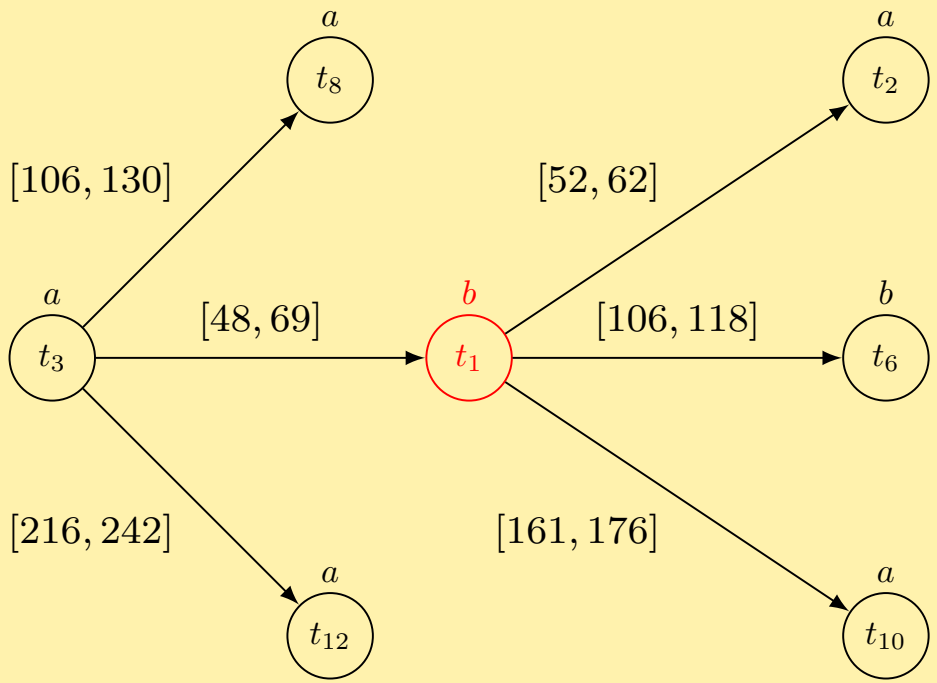


$$J(x_3, x_7) = 1,$$
$$J(x_5, x_9) = 1,$$
$$J(x_1, x_5) = 0.94,$$
$$\cancel{J(x_1, x_5) = 0.94,}$$
$$J(x_3, x_{11}) = 0.94,$$
$$\cancel{J(x_3, x_{11}) = 0.94.}$$

Étape suivante : 2 - Type 3 (1 nœud source et 1 nœud destination)...

Exemple d'ordre des opérations (6/7)

Exemple : 2 - Type 3 (1 nœud source et 1 nœud destination)



$J(x_1, x_4) = 1,$

~~$J(x_1, x_6) = 0.94,$~~

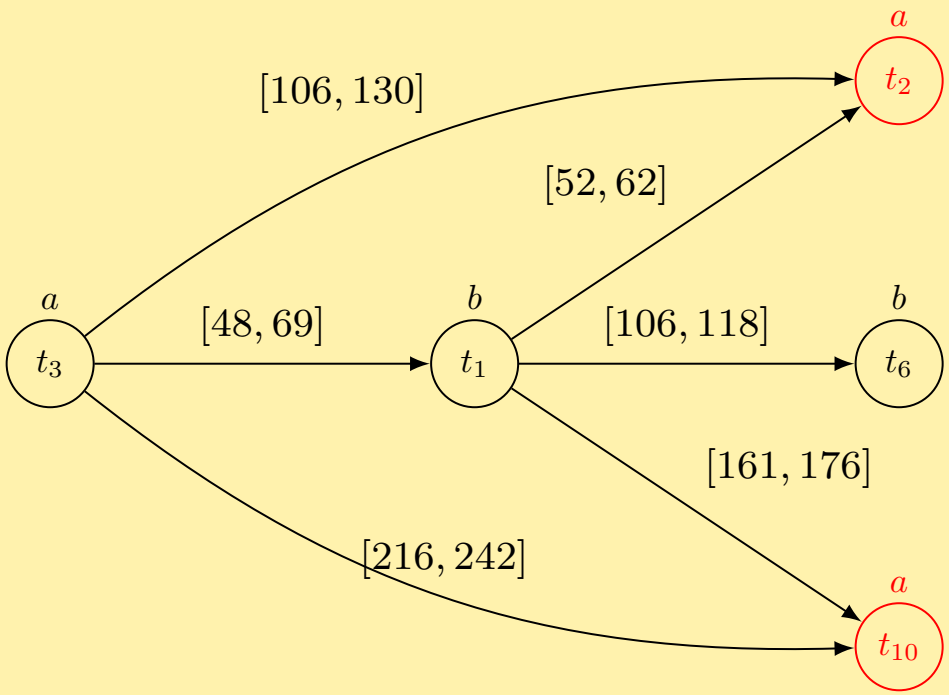
~~$J(x_1, x_4) = 0.94,$~~

~~$J(x_1, x_4) = 0.94.$~~

Étape suivante : 3 - Type 2 (2 nœuds destinations)...

Exemple d'ordre des opérations (7/7)

Exemple : 3 - Type 2 (2 nœuds destinations)



~~$J(x_2, x_1) = 1,$~~
$J(x_{10}, x_{12}) = 1,$
$J(x_2, x_8) = 0.94,$
~~$J(x_2, x_{10}) = 0.94,$~~
~~$J(x_1, x_1) = 0.94,$~~
~~$J(x_1, x_{10}) = 0.94,$~~
~~$J(x_1, x_{12}) = 0.94.$~~

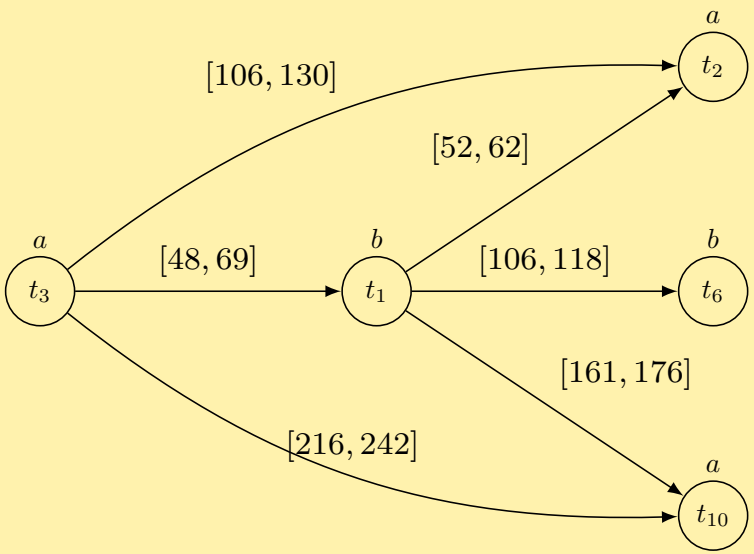
Expériences (1/2)

Définition : Complexité de chronique

Cette métrique de performance est définie par la taille d'un chronique n et par le nombre de ses contraintes temporelles m. Elle est calculée par la formule suivante :

$$cc = \frac{2m}{n-1}.$$

Exemple



$$m = 6,\ n = 5,$$

$$cc = \frac{2m}{n-1} = 3.$$

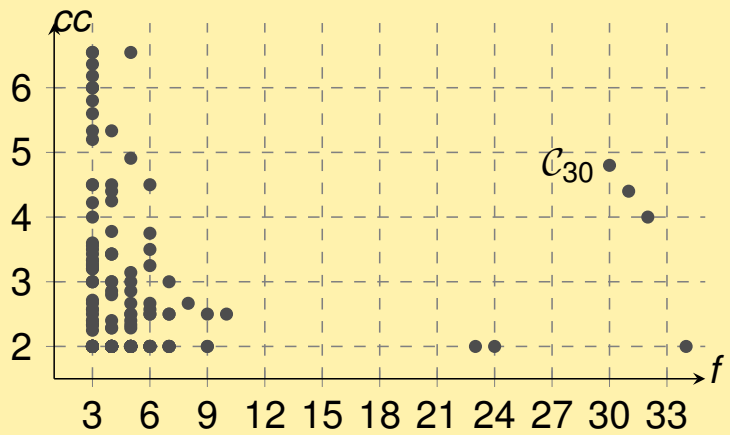
Expériences (2/2)

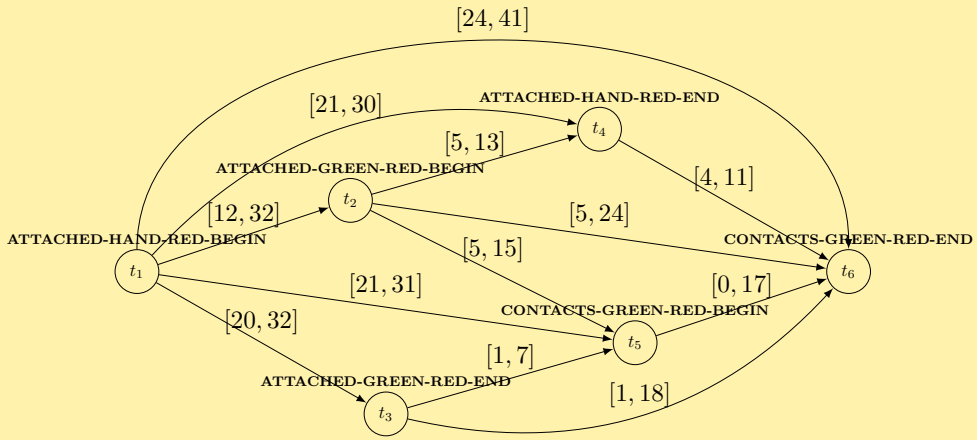
Blocks dataset [1]

$minJac = 0.9$, DBSCAN : $\varepsilon = 10$, $minPts = 3$

Temporal Sequence	Events	Event types	# chronicles learned
assemble	528	12	5995
disassemble	486	12	3147
pick-up	184	6	127
put-down	***186***	***6***	***180***
stack	362	10	1349
unstack	364	10	1119

Exemple : *put-down*





1. F. Mörchen and D. Fradkin : Robust mining of time intervals with semi-interval partial order patterns, SIAM, 2010

Conclusions et perspectives

Ce qui as été fait :

- Une nouvelle approche au problème de découverte de chroniques,
- Une métrique de performance pour aider l'expert à évaluer les chroniques apprises.

Avantages & Inconvénients

- Ne nécessite pas de fréquence objective pour les chroniques apprises,
- Les paramètres de l'algorithme de clustering choisi influence beaucoup le temps d'exécution et la qualité des résultats.

Ce qui reste à faire :

- Comparaison avec d'autres algorithmes de découverte de chroniques,
- Discrétisation de séries temporelles $\rightarrow$ Amélioration possible des résultats ?

Merci pour votre attention !

Un algorithme de découverte de chroniques : une approche par le clustering.

Alexandre Sahuguède

Encadré par : Euriell Le Corronc
Marie-Véronique Le Lann

LAAS-CNRS, Équipe DISCO, Université Paul Sabatier, Toulouse, FRANCE

3 Décembre 2018