IRISA

DREAM

Extension des BDI

pour la découverte de chroniques avec contraintes temporelles

Introduction

- Fouille de données temporelles
  - Complexité dûe à l'introduction du temps numérique
- Base de données inductives
  - Un cadre formel de la fouille de données

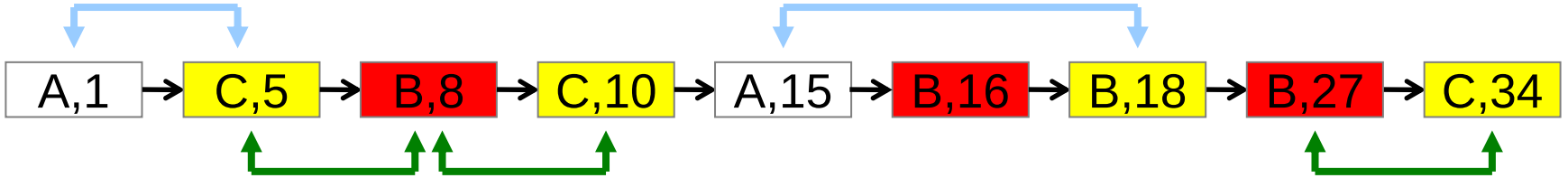
intégrer le temps dans une BDI

Plan

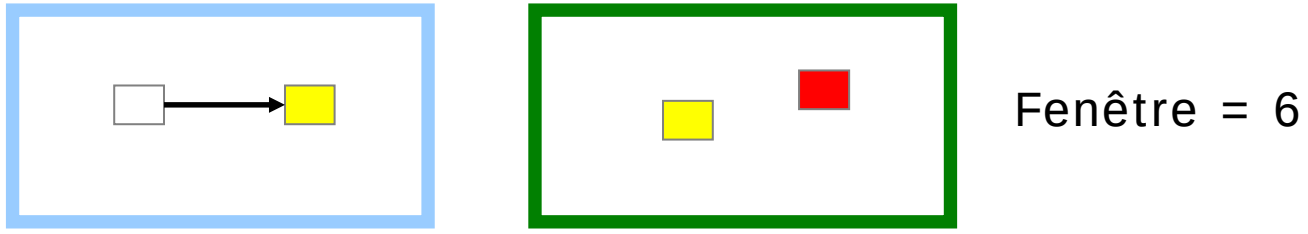
- Introduction
- Présentation générale de la fouille de données temporelles
- Base de données inductive et chroniques
  - Relation d'ordre - fréquence
- Traitement d'une requête
  - Algorithme de Mitchell
  - FACE : un outil de fouille de données temporelles

Fouille de données temporelles (1)

- A partir d'une séquence d'évènements : une liste ordonnée d'occurrences d'évènements



- On découvre des motifs



  - [*Manilla et al. 1997*]

Fouille de données temporelles (2)

- Techniques de la fouille de données « classiques »
  - Itemset ⇔ Ensemble d'évènements
  - Intérêt d'un motif essentiellement basé sur la fréquence.

Date	Enregistrement
1	A,B
2	C
5	A,C

Base de données inductive

[L. De Raedt]

- Une formalisation de la fouille de données
- Base de données « classique » : pas de motifs
- Base de données inductive : Une base de données et une base de motifs
  - La fouille de données est vue comme un processus d'extraction par **requête**
- Exemple de requêtes sur les motifs et les données :
  - Fréquence(m,D_1) > T_{min}
  - Fréquence(m,D_2) < T_{max}
  - Sous-motif(m,M_1)
  - Sous-motif(M_1,m)

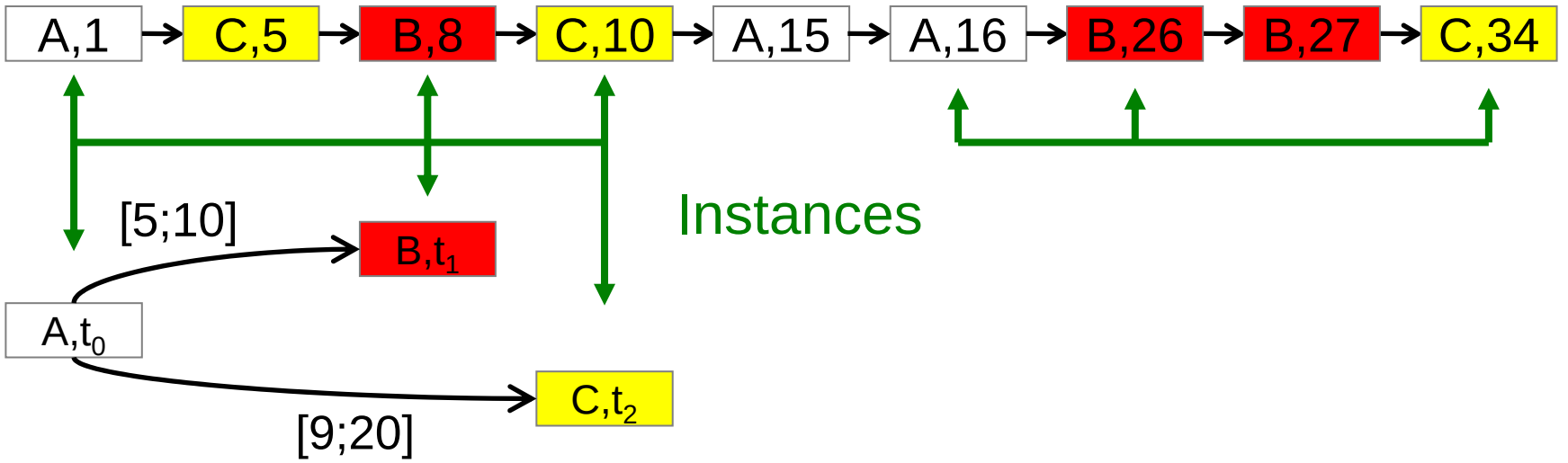
m : motifs cibles à déterminer
D_1, D_2 : données
$T_{...}$: fréquence
M_1 : motifs

Notre approche :

- Intégrer une notion numérique du temps dans les BDI
  - Les bases de données inductives actuelles n'intègrent pas de notion numérique du temps
    [Lee et De Raedt 2002, Dzeroski 2002, Imielinski et Manilla 1996]

- Les données : des séquences d'évènements
- Les motifs : des chroniques

Les motifs temporels : les chroniques



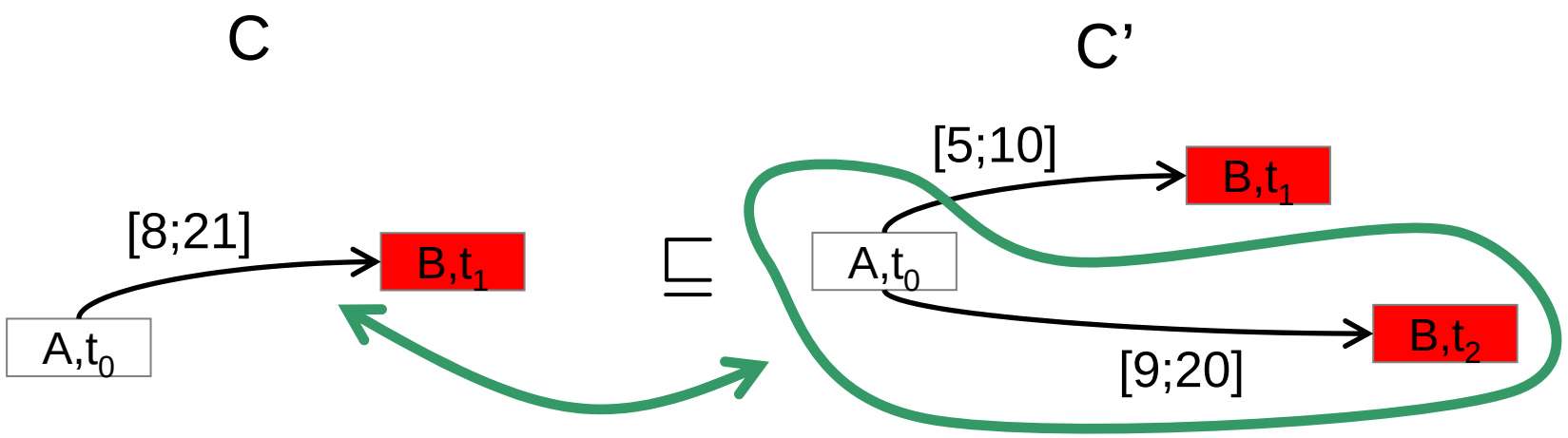
- Une chronique : Ensemble d'évènements contraints temporellement
- Une contrainte : [a,b] tel que a,b $\in \mathbb{Z}$

Relation de généralité

- C plus générale que C’ $(C \sqsubseteq C')$

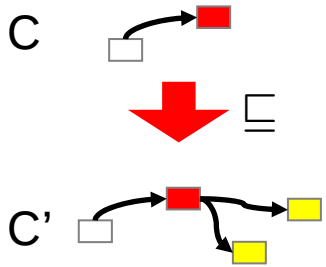
$\Leftrightarrow$

Il existe un sous-graphe de C’ tel que les contraintes de C sont égales ou plus larges que celles de ce sous-graphe.

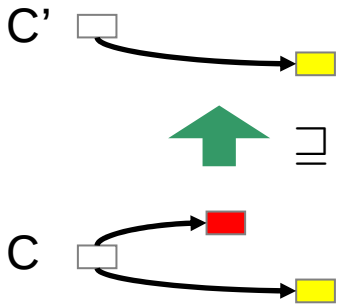


Fréquence et relation d'ordre

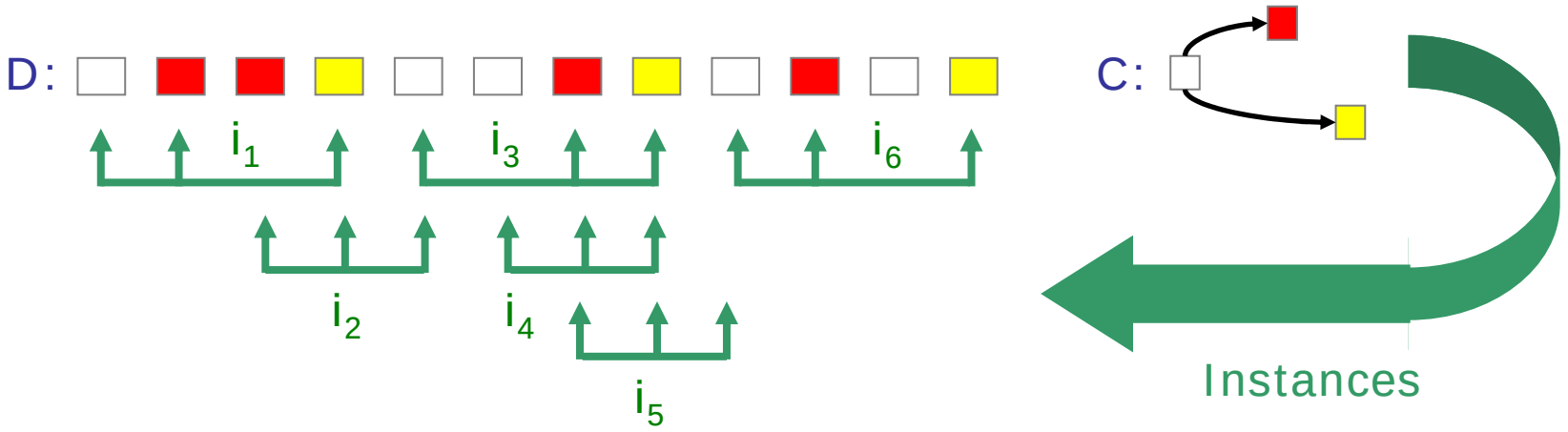
- Contrainte monotone
  - Ex : freq(m,D) $\leq$ T
    - $C \in m \wedge C' \sqsubseteq C \Rightarrow C' \in m$
    - $Freq(C') \leq Freq(C)$



- Contrainte anti-monotone
  - Ex : freq(m,D) $\geq$ T
    - $C \in m \wedge C' \sqsubseteq C \Rightarrow C' \in m$
    - $Freq(C') \geq Freq(C)$



Fréquence d'une chronique



- $I_C(D) = \{i_1, i_2, i_3, i_4, i_5, i_6\}$
- Critère de reconnaissance Q
  - $Freq(C,D) = |E|, E \subseteq I_C(D), Q(E)$

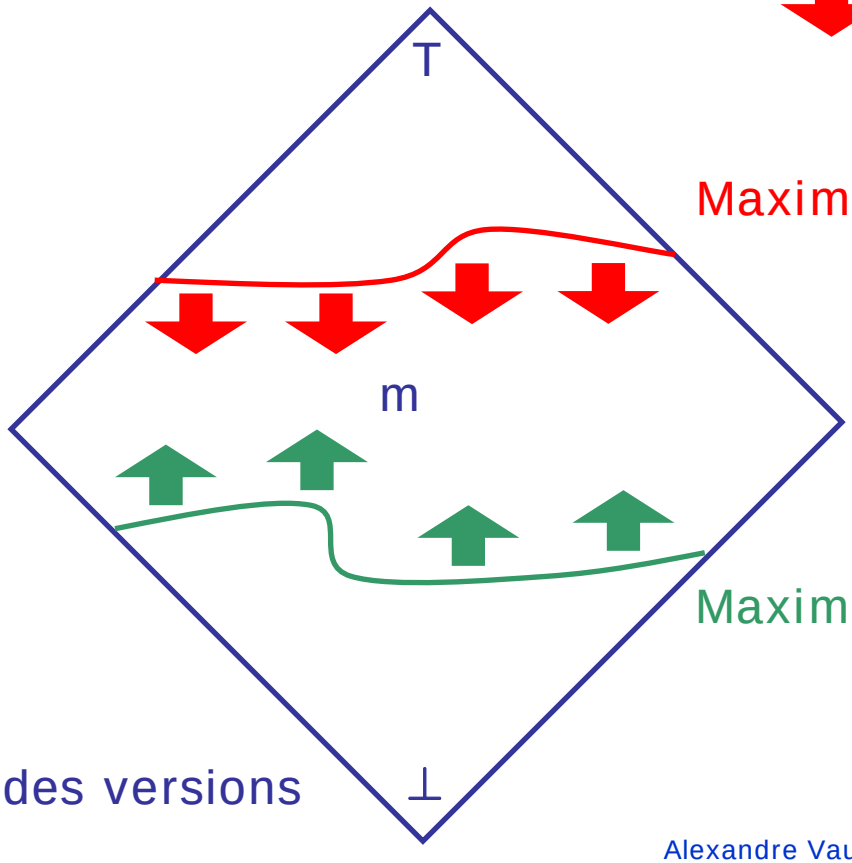
E est unique

(Anti)monotonie des contraintes sur la fréquence

- Exemple, $Q_{d\&t}$: critère d'instances disjointes au plus tôt
  - $E = \{i_1, i_3, i_6\}$ et $F(C,D) = 3$

Traitement d'une requête

- $freq(m,D_1) \geq \mathcal{T}_1 \wedge freq(m,D_2) \leq \mathcal{T}_2$



- $freq(m,D_1) \leq \mathcal{T}_1 \wedge freq(m,D_2) \geq \mathcal{T}_2$

Maximalement général

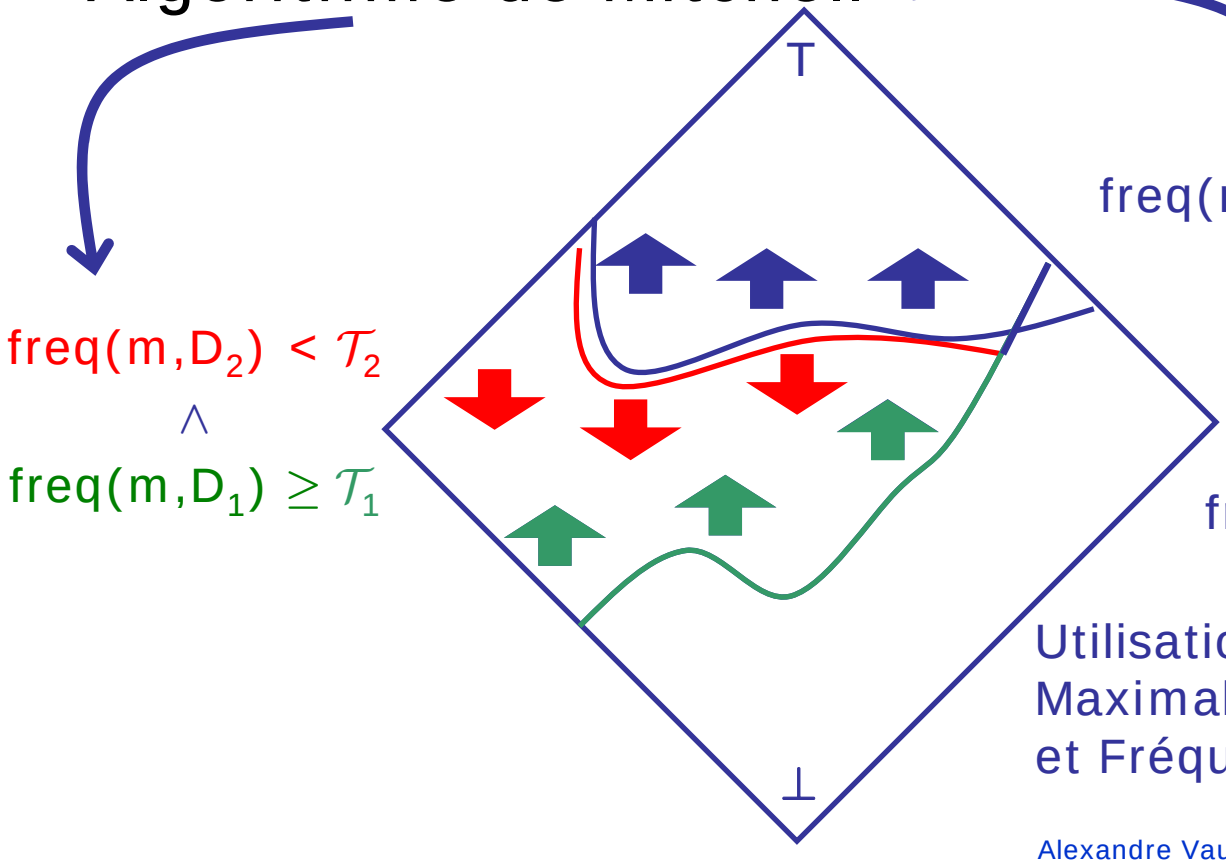
Le calcul des bords suffit

Maximalement spécifique

Espace des versions

Calcul des bords de l'espace des versions

- $freq(m,D_1) \geq \mathcal{T}_1 \wedge freq(m,D_2) < \mathcal{T}_2$
  - Algorithme de Mitchell



$freq(m_2,D_2) \geq \mathcal{T}_2$

+

$freq(m_1,D_1) \geq \mathcal{T}_1$

$freq(m,D_2) < \mathcal{T}_2$

$\wedge$

$freq(m,D_1) \geq \mathcal{T}_1$

Utilisation des Chroniques Maximalement spécifiques et Fréquentes (CMFs)

FACE

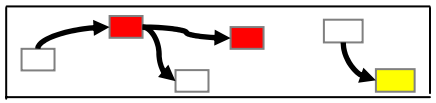
Frequency Analyser for Chronicle Extraction

[Dousson & Thang Vu Duong]

Séquence d'évènements

FACE

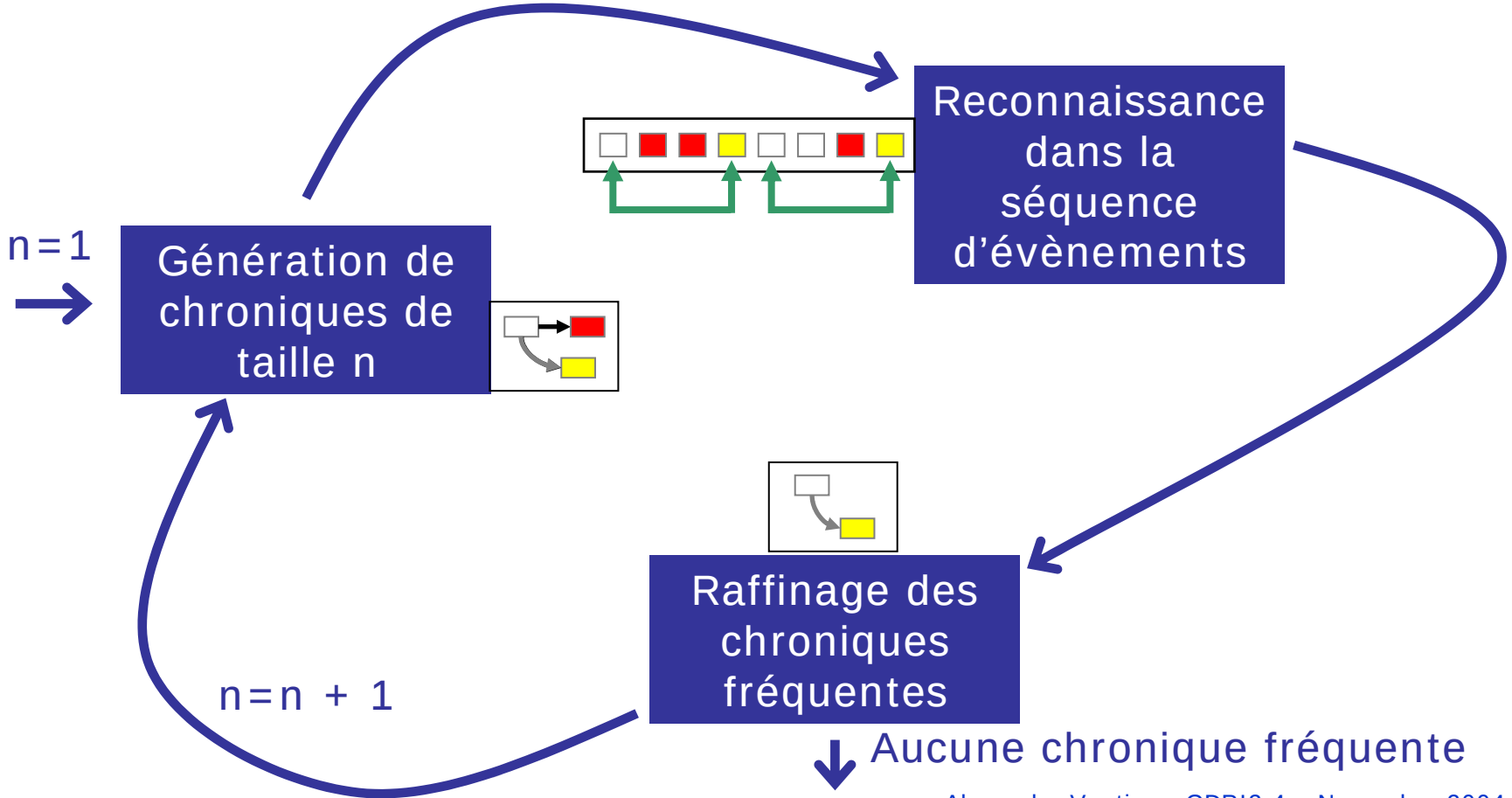
Chroniques représentatives



- Principe :
  - Freq(m,D) $\geq \mathcal{T}$
    - Contrainte Anti-monotone
    - Une chronique peut être fréquente si toutes ses sous-chroniques sont fréquentes.

Algorithme général de FACE

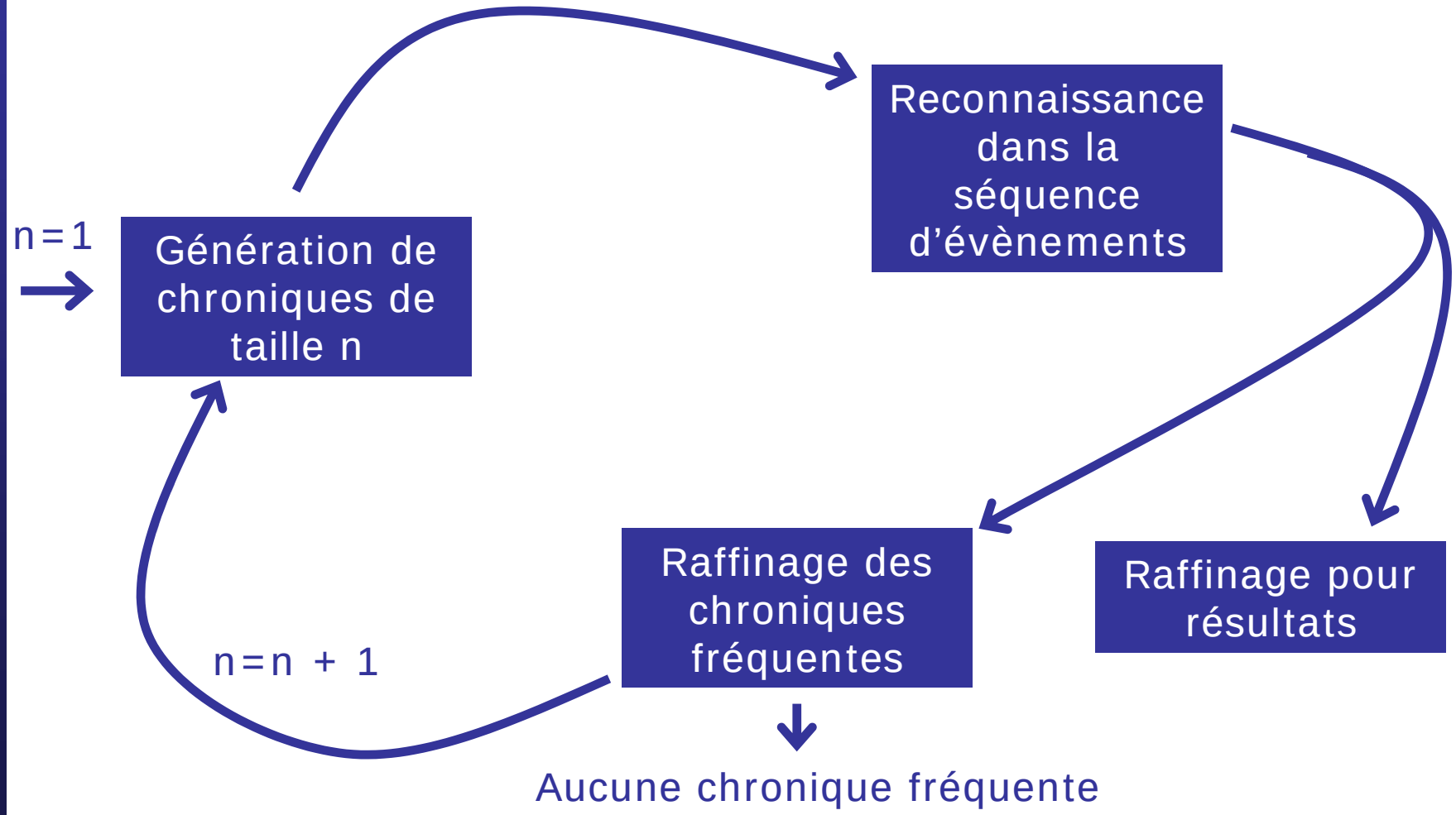
- Adaptation de l'algorithme *A PRIORI*



FACE…
…Un outil de fouille de données

- Optimisé pour la génération d'un nombre minimum de chroniques
  - Synthèse efficace ≠ recherche des CMFs
- Un extracteur complet et correct d'instances !
  - À partir des instances reconnues on peut retrouver les chroniques maximalement spécifiques et fréquentes (CMFs)

Algorithme modifié de FACE



Raffinage pour résultats

- ⇔ Recherche des motifs fréquents sur des données numériques
  - Très coûteux : en temps, en espace
- Introduction d'un nouveau critère d'intérêt : la densité
  - Meilleure caractérisation des chroniques intéressantes
  - Réduction du nombre de CMFs
- Utilisation et adaptation d'algorithmes de clustering
  - Basé sur la densité, algorithmes hiérarchiques...

Conclusion

- Extension des BDIs à la recherche de motifs intégrant une notion temporelle
  - Nécessité de calculer seulement les CMFs de chaque séquence d'évènements
- Formalisation de la notion de chronique
  - Relation d'ordre
  - Fréquence, critère de reconnaissance
- Utilisation d'un outil de fouille de données existant : FACE

Perspectives

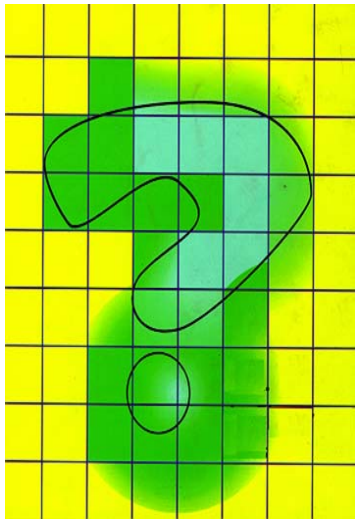
- Poursuivre la réalisation des bases de données inductives étendues au temps
- Utilisation d'autres mesures d'intérêt
  - Fréquence sans critère de reconnaissance
  - Autre que la fréquence
- Application dans le domaine de la détection d'intrusions dans les réseaux de télécommunications

IRISA

DREAM

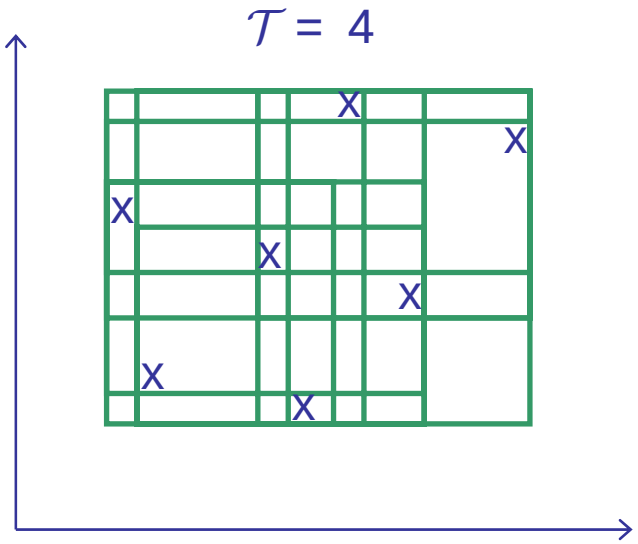
Extension des BDI

pour la découverte de chroniques avec contraintes temporelles



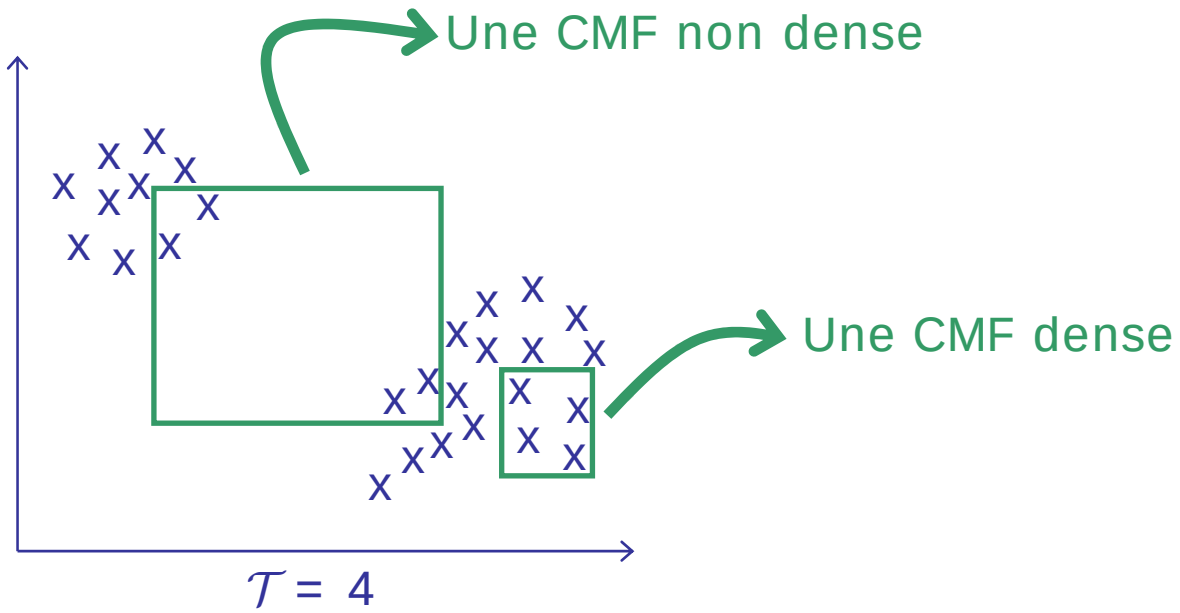
Raffinage pour résultats

- Une instance $\Leftrightarrow$ un point
- Une chronique $\Leftrightarrow$ un hypercube
- Rechercher les hypercubes minimaux englobant au moins $\mathcal{T}$ points



- Très coûteux
  - En temps
  - En espace
- Recherche des motifs fréquents sur des données numériques

Densité des CMFs (1)



- Chronique intéressante :
  - Fréquente
  - Maximalement spécifique
  - **dense**

Densité des CMFs (2)

- Utilisation et adaptation d'algorithmes de clustering
  - Basé sur la densité, algorithmes hiérarchiques...
- Réduction du nombre de CMFs
- Meilleure caractérisation des chroniques intéressantes