

# Extraction de Motifs Séquentiels

Pascal Poncelet  
LIRMM

Pascal.Poncelet@lirmm.fr  
<http://www.lirmm.fr/~poncelet>



# Plan

---

- Contexte général
- Motifs séquentiels
- Extensions des motifs séquentiels
- Quelques applications des motifs
- Conclusions



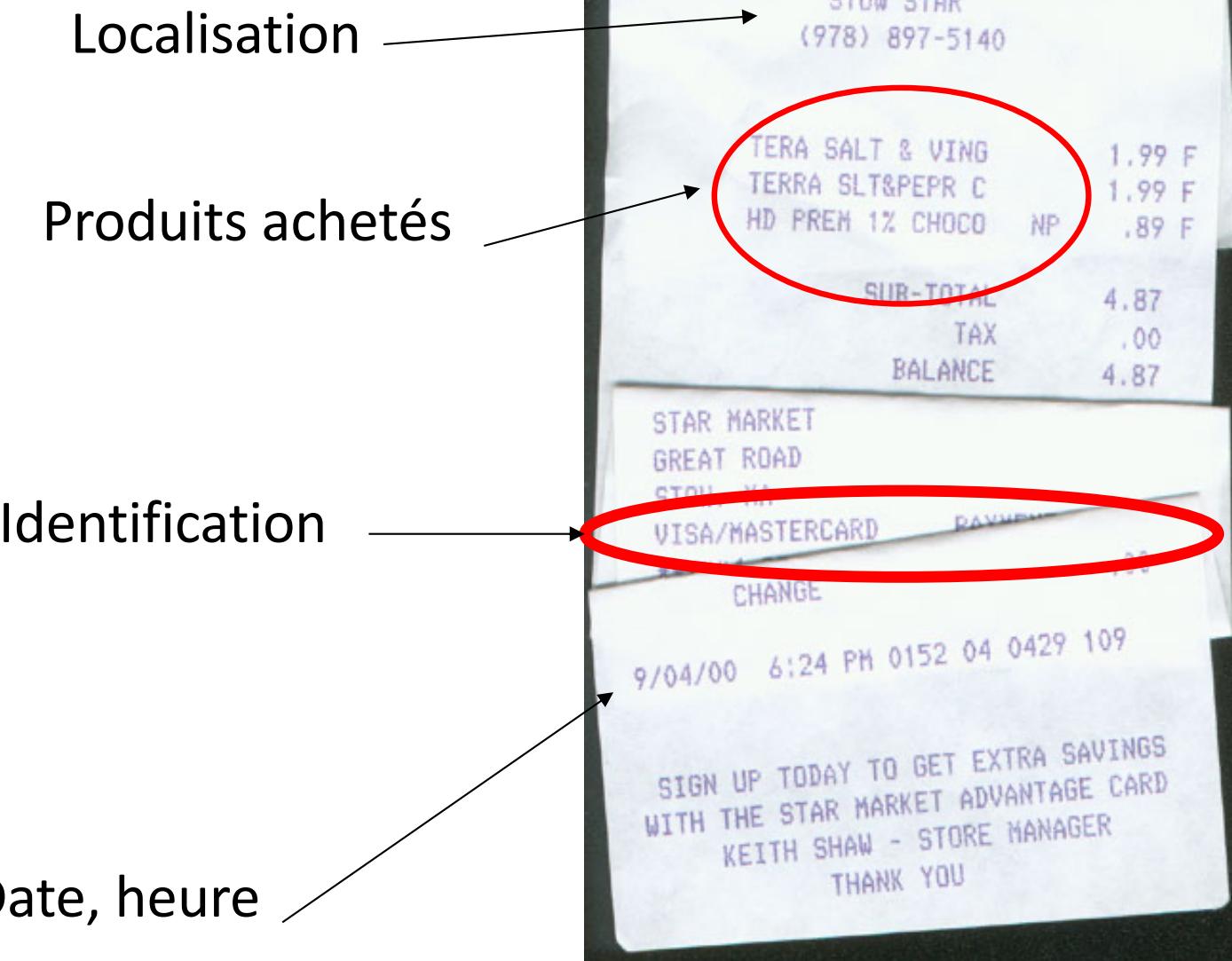
# « Panier de la ménagère »

---

- **Recherche d'associations**
  - recherche de corrélations entre attributs (items)
  - caractéristiques : « panier de la ménagère »
  - de très grandes données
  - limitations : données binaires
- **Recherche de motifs séquentiels**
  - recherche de corrélations entre attributs (items) mais en prenant en compte le temps entre items => comportement

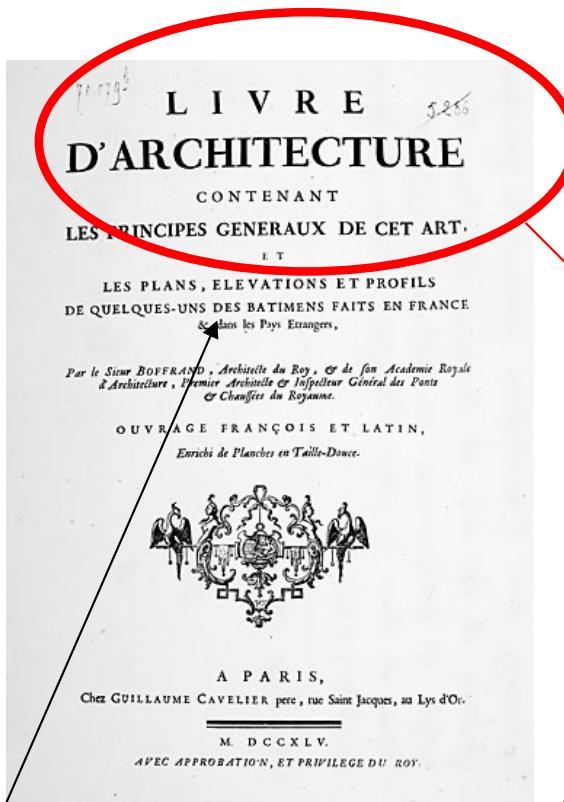


# Panier de la ménagère



# Panier de la ménagère

Localisation



Premier paragraphe

« Livre d'architecture contenant les principes généraux ... »

Position # Date

Mots # Produits



Identification

# Rappel RA

---

- $R : X \rightarrow Y (A\%, B\%)$ 
  - **Support** : portée de la règle  
*Proportion de paniers contenant tous les attributs A% des clients ont acheté les 2 articles X et Y*
  - **Confiance** :  
*Proportion de paniers contenant le conséquent parmi ceux qui contiennent l'antécédent  
B% des clients qui ont acheté X ont aussi acheté Y*
- Beurre, Pain → Lait [70%, 80%]
- Bière, Gâteaux → Couches [30%, 80%]
- Caviar → Champagne [10%, 90%]



# En résumé

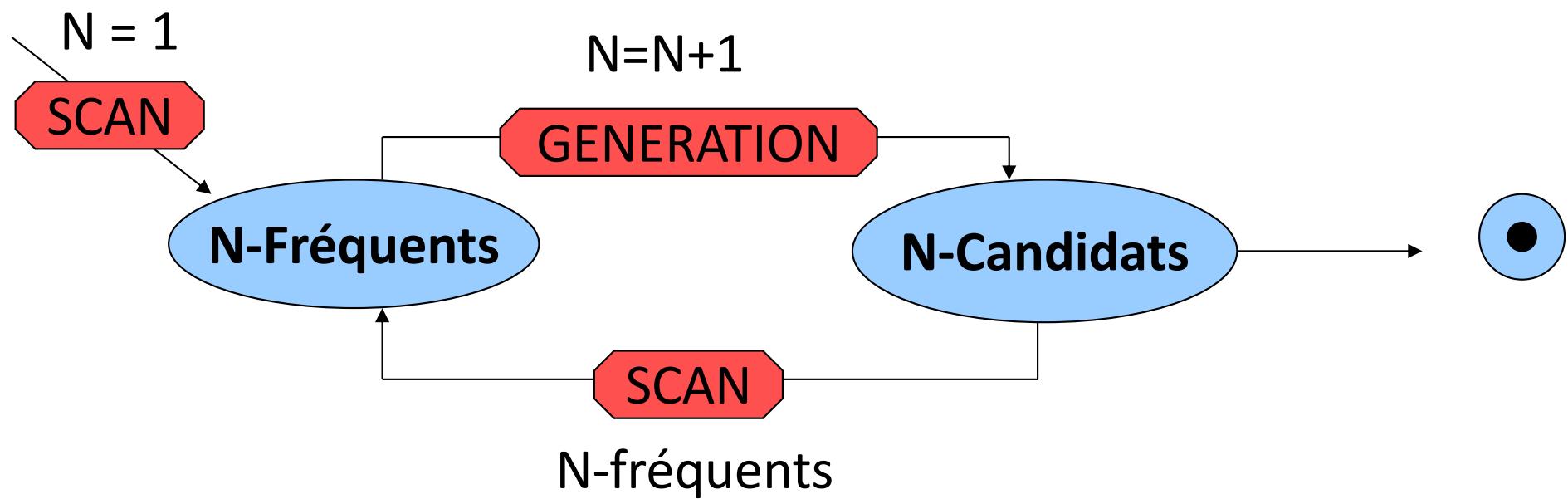
- Itemsets : A, B ou B, E, F
- Support pour un itemset
  - Supp (A,D)=1
  - Supp (A,C) = 2
- Itemsets fréquents (minSupp=50%)
  - {A,C} est un itemset fréquent
- Pour minSupp = 50% et minConf = 50%, nous avons les règles suivantes :
  - A → C [50%, 50%]
  - C → A [50%, 100%]

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

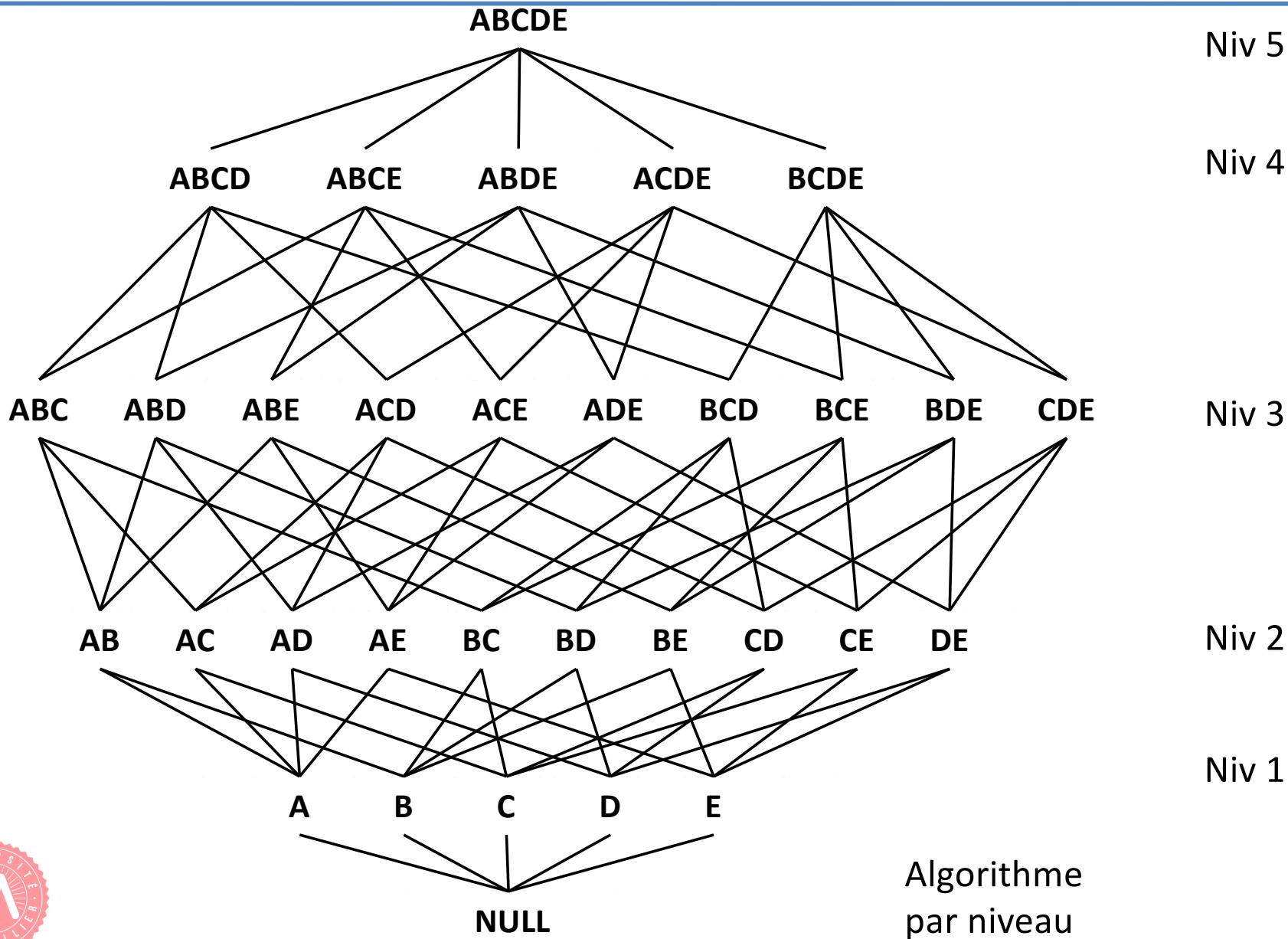


# Vers un algorithme générique

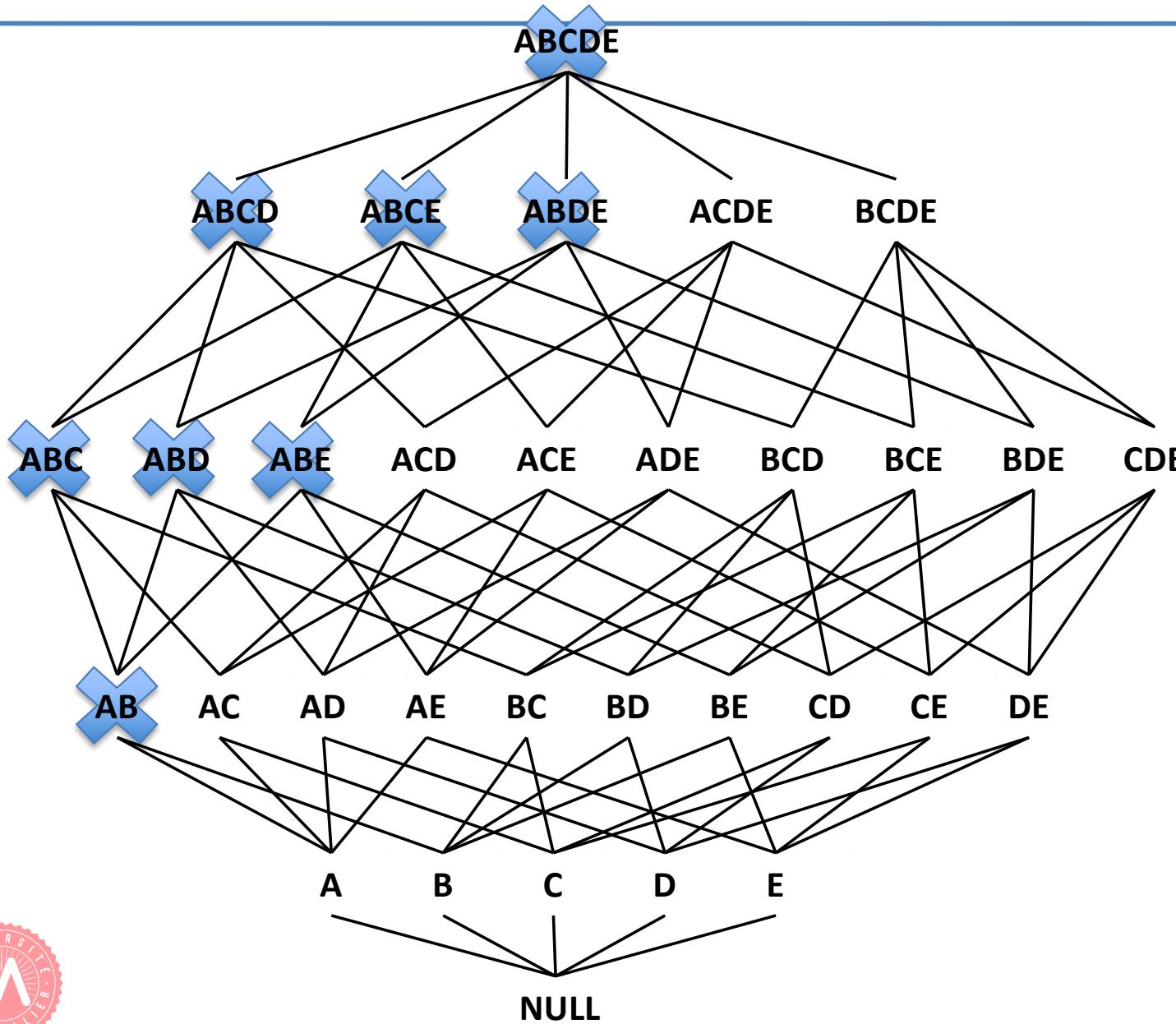
---



# Espace de recherche



# Espace de recherche



# Pourquoi la recherche de séquence ?

---

- Un important domaine de recherche pour le data mining avec de très nombreuses applications
  - Analyse des achats des clients
  - Analyse de puces ADN
  - Processus
  - Conséquences de catastrophes naturelles
  - Web mining
  - Détection de tendances dans des données textuelles



# Recherche de Motifs Séquentiels

---

- Même problématique mais avec le temps
- Item : « un article »
- Transaction : un client + un itemset + une estampille temporelle  $T = [C, (a,b,c)_5]$
- Séquence : liste ordonnée d'itemsets
- Séquence de données : « activité du client »  
Soit  $T_1, T_2, \dots, T_n$ , les transactions du client C, la séquence de données de C est :  
 $[C, <\text{itemset}(T_1) \text{ itemset}(T_2) \dots \text{ itemset}(T_n)>]$



# Recherche de Motifs Séquentiels

---

- Support minimal : nombre minimum d'occurrences d'un motif séquentiel pour être considéré comme fréquent
- Attention l'occurrence n'est prise en compte qu'une fois dans la séquence

Support (20) dans <(10) (20 30) (40) (20)>=1



# Inclusion

---

- Inclusion : Soient  $S_1 = \langle a_1 \ a_2 \ \dots \ a_n \rangle$  et  $S_2 = \langle b_1 \ b_2 \ \dots \ b_n \rangle$   $S_1 \subseteq S_2$  ssi
$$i_1 < i_2 < \dots < i_n / a_1 \subseteq b_{i1}, \dots, a_n \subseteq b_{in}$$
- $S1 = \langle (10) \ (20 \ 30) \ (40) \ (20) \rangle$

$S2 = \langle (20) \ (40) \rangle \subseteq S1$

$S3 = \langle (20) \ (30) \rangle$  n'est pas incluse dans  $S1$



# Problématique

---

- Soit D une base de données de transactions de clients. Soit  $\sigma$  une valeur de support minimal  
Rechercher toutes les séquences S telles que :  
 $\text{support}(S) \geq \sigma$  dans D
- 50% des personnes qui achètent du vin et du fromage **le lundi** achètent aussi **du pain le vendredi**

<(French wine, cheese) (bread)>



# Illustration

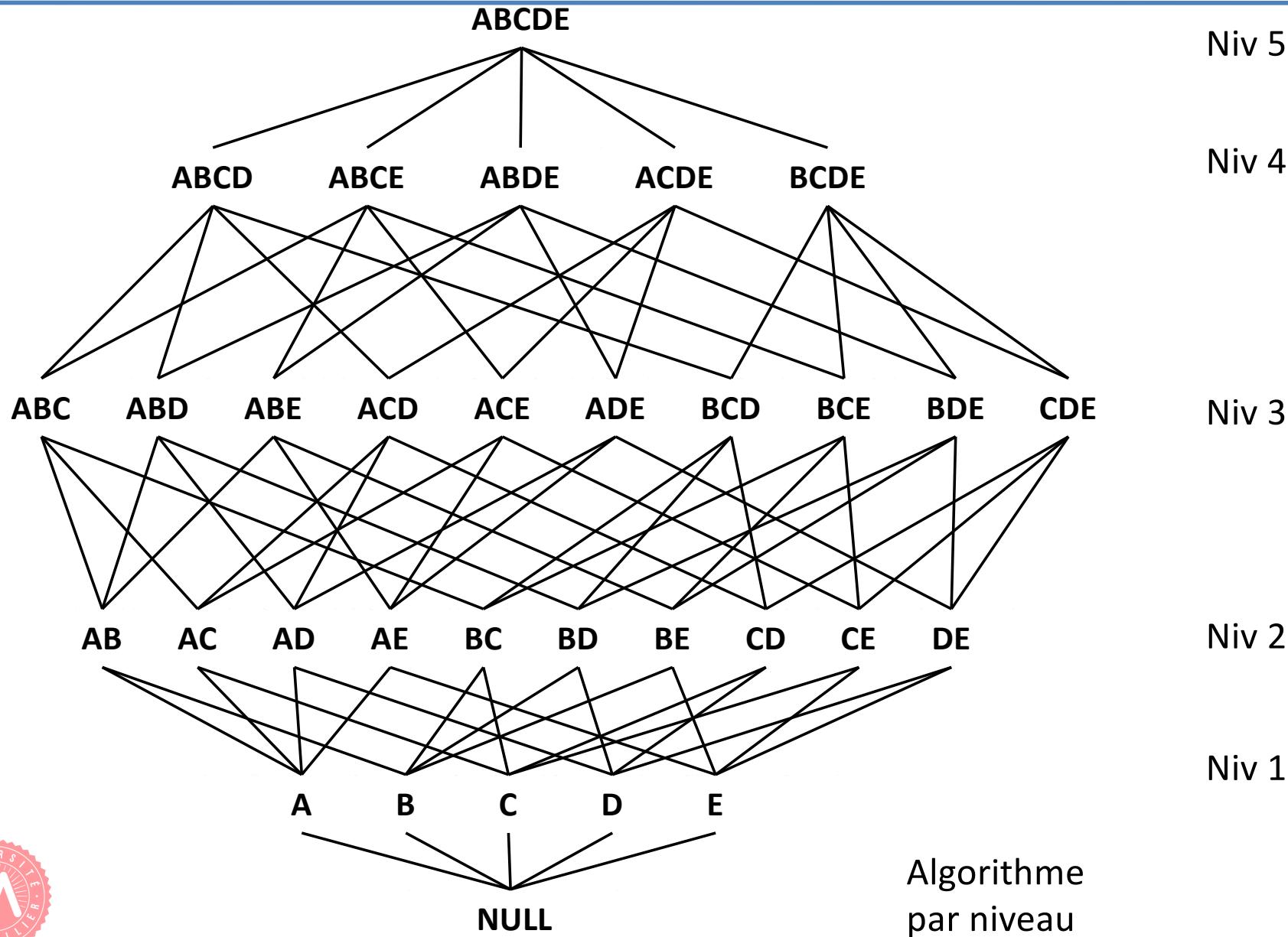
Clients	Date1	Date2	Date3	Date4
C1	10 20 30	20 40 50	10 20 60	10 40
C2	10 20 30	10 20 30		20 30 60
C3	20 30 50		10 40 60	10 20 30
C4	10 30 60	20 40	10 20 60	50

Support = 60% (3 clients) => ?

<(10 30) (20) (20 60)>

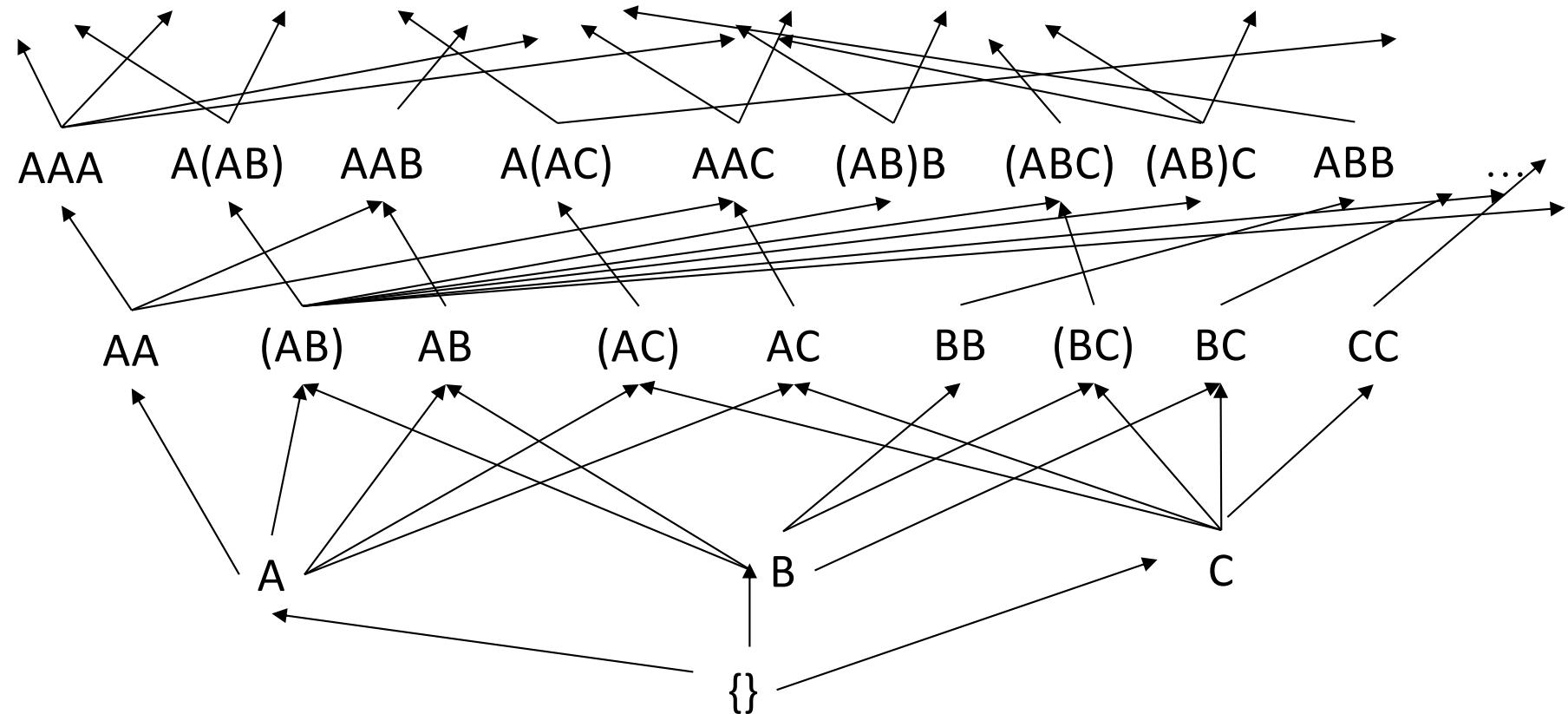


# Itemsets : Espace de recherche



# Motifs Séquentiels : l'espace de recherche

Espace de recherche théoriquement infini mais borné par la longueur de la plus grande séquence



# La propriété d'antimonotonie

---

- Une propriété essentielle (c.f. Apriori [AIS93])
  - Si une séquence n'est pas fréquente, aucune des super-séquences de S n'est fréquente!

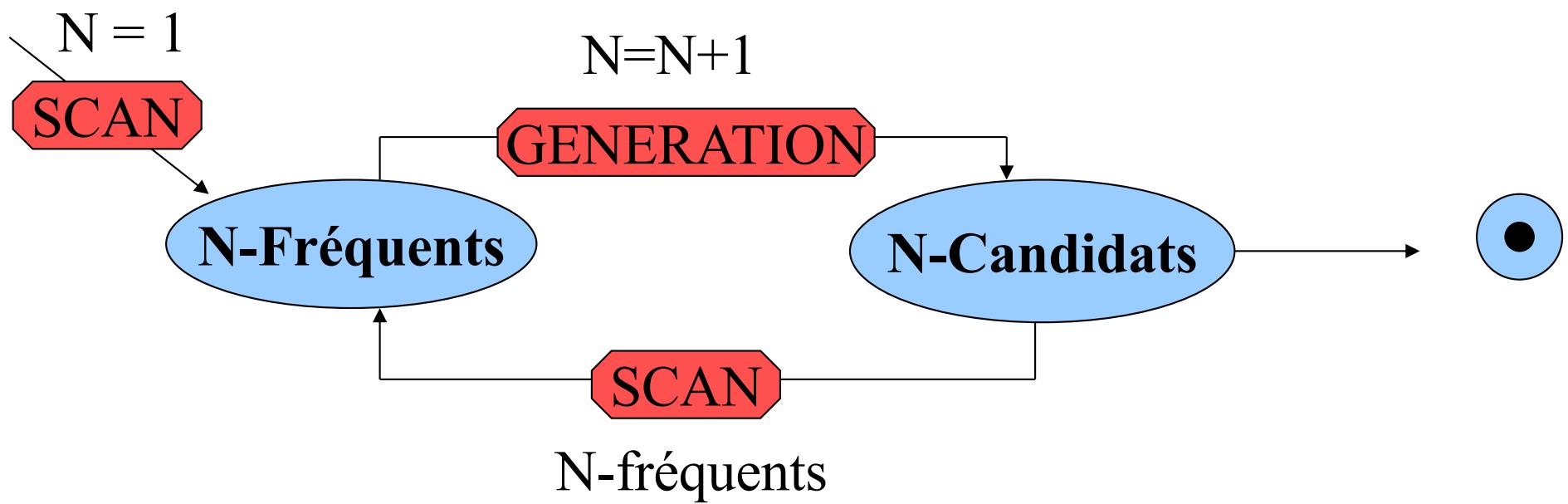
Support ( $\langle(10) (20 30)\rangle$ ) < minsupp

Support ( $\langle(10) (20 30) (40)\rangle$ ) << minsupp



# Vers un algorithme générique

---



# Quid des candidats ?

---

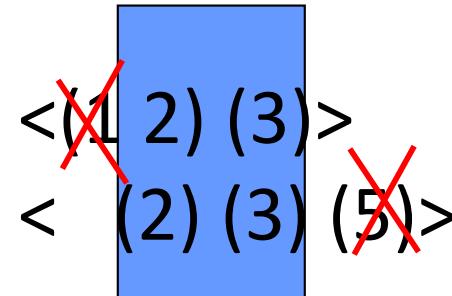
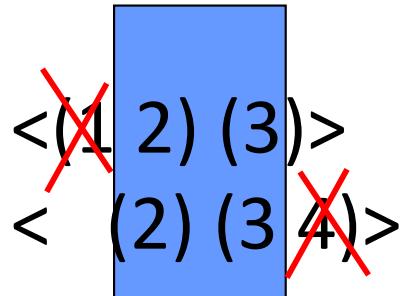
- Comment générer les candidats ?



# Génération des candidats

---

- S-Extension : ajout d'une séquence
- I-Extension : ajout d'un itemset



---

<(1 2) (3 4)>  
I-Extension

<(1 2) (3) (5)>  
S-Extension

# GSP

---

- A la APRIORI [Srikant, Agrawal, EDBT' 96]

$L=1$

While ( $\text{Result}_L \neq \text{NULL}$ )

Candidate Generate

Prune

Test

$L=L+1$



# Recherche des séquences de taille 1

- Candidats initiaux : toutes les séquences réduites à un item
  - $\langle a \rangle, \langle b \rangle, \langle c \rangle, \langle d \rangle, \langle e \rangle, \langle f \rangle, \langle g \rangle, \langle h \rangle$
- Un passage sur la base pour compter le support des candidats

Seq. ID	Séquence
10	$\langle (bd)cb(ac) \rangle$
20	$\langle (bf)(ce)b(fg) \rangle$
30	$\langle (ah)(bf)abf \rangle$
40	$\langle (be)(ce)d \rangle$
50	$\langle a(bd)bcb(ade) \rangle$

$$\minSupp = 2$$

Cand	Sup
$\langle a \rangle$	3
$\langle b \rangle$	5
$\langle c \rangle$	4
$\langle d \rangle$	3
$\langle e \rangle$	3
$\langle f \rangle$	2
$\langle g \rangle$	1
$\langle h \rangle$	1

# Le Processus

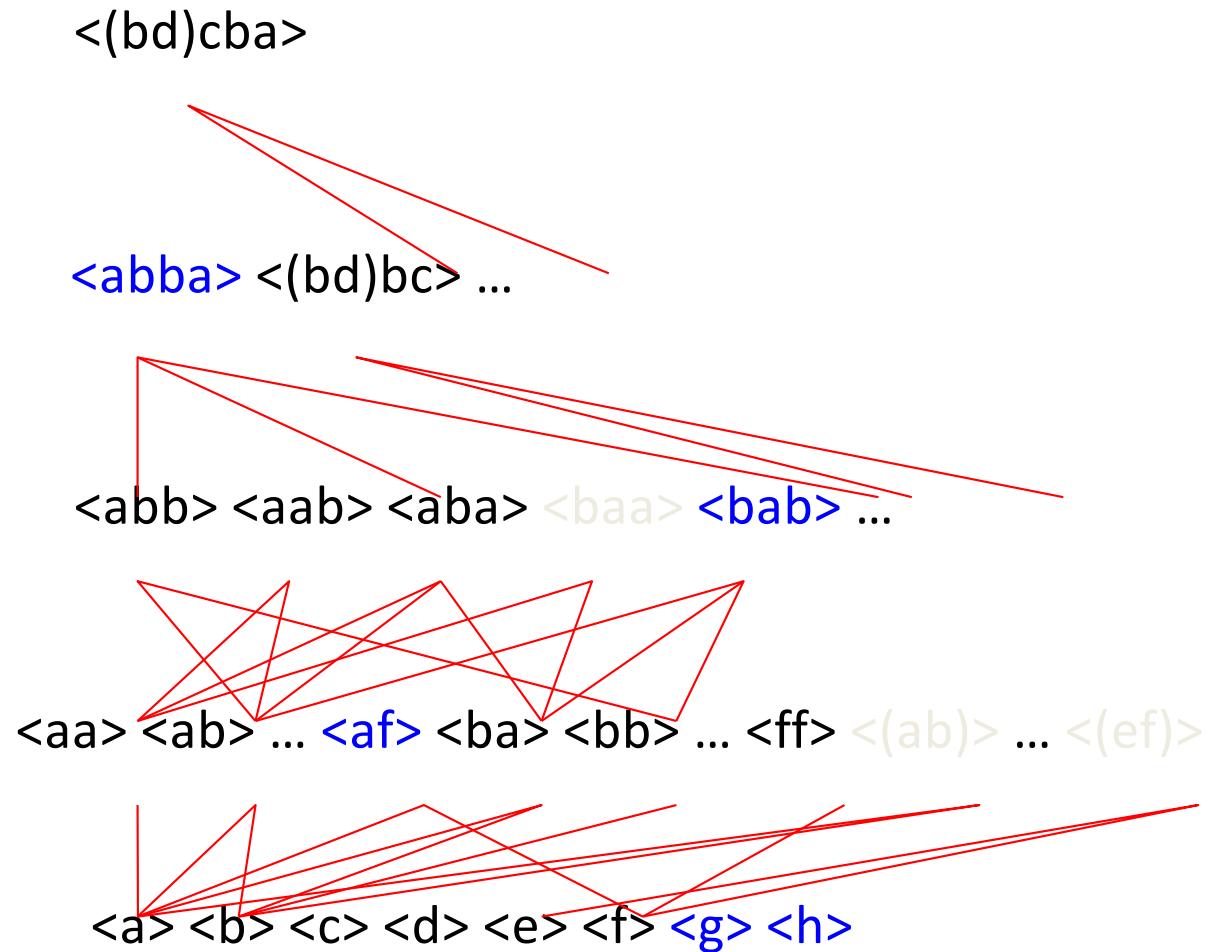
**5th scan :** 1 candidate  
1 length-5 seq pattern

**4th scan :** 8 candidates  
6 length-4 seq pat

**3rd scan :** 46 candidates  
19 length-3 seq pat.

**2<sup>nd</sup> scan :** 51 candidates  
19 length-2 seq pat.

**1st scan :** 8 candidates  
6 length-1 seq pattern



# Génération des candidats de taille 2

51 2-Candidats

I-Extension

S-Extension

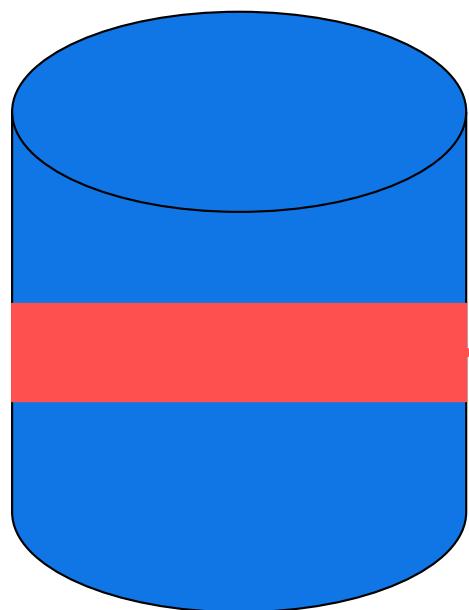
	<a href="#">&lt;a&gt;</a>	<a href="#">&lt;b&gt;</a>	<a href="#">&lt;c&gt;</a>	<a href="#">&lt;d&gt;</a>	<a href="#">&lt;e&gt;</a>	<a href="#">&lt;f&gt;</a>
<a href="#">&lt;a&gt;</a>	<a href="#">&lt;aa&gt;</a>	<a href="#">&lt;ab&gt;</a>	<a href="#">&lt;ac&gt;</a>	<a href="#">&lt;ad&gt;</a>	<a href="#">&lt;ae&gt;</a>	<a href="#">&lt;af&gt;</a>
<a href="#">&lt;b&gt;</a>	<a href="#">&lt;ba&gt;</a>	<a href="#">&lt;bb&gt;</a>	<a href="#">&lt;bc&gt;</a>	<a href="#">&lt;bd&gt;</a>	<a href="#">&lt;be&gt;</a>	<a href="#">&lt;bf&gt;</a>
<a href="#">&lt;c&gt;</a>	<a href="#">&lt;ca&gt;</a>	<a href="#">&lt;cb&gt;</a>	<a href="#">&lt;cc&gt;</a>	<a href="#">&lt;cd&gt;</a>	<a href="#">&lt;ce&gt;</a>	<a href="#">&lt;cf&gt;</a>
<a href="#">&lt;d&gt;</a>	<a href="#">&lt;da&gt;</a>	<a href="#">&lt;db&gt;</a>	<a href="#">&lt;dc&gt;</a>	<a href="#">&lt;dd&gt;</a>	<a href="#">&lt;de&gt;</a>	<a href="#">&lt;df&gt;</a>
<a href="#">&lt;e&gt;</a>	<a href="#">&lt;ea&gt;</a>	<a href="#">&lt;eb&gt;</a>	<a href="#">&lt;ec&gt;</a>	<a href="#">&lt;ed&gt;</a>	<a href="#">&lt;ee&gt;</a>	<a href="#">&lt;ef&gt;</a>
<a href="#">&lt;f&gt;</a>	<a href="#">&lt;fa&gt;</a>	<a href="#">&lt;fb&gt;</a>	<a href="#">&lt;fc&gt;</a>	<a href="#">&lt;fd&gt;</a>	<a href="#">&lt;fe&gt;</a>	<a href="#">&lt;ff&gt;</a>

	<a href="#">&lt;a&gt;</a>	<a href="#">&lt;b&gt;</a>	<a href="#">&lt;c&gt;</a>	<a href="#">&lt;d&gt;</a>	<a href="#">&lt;e&gt;</a>	<a href="#">&lt;f&gt;</a>
<a href="#">&lt;a&gt;</a>		<a href="#">&lt;(ab)&gt;</a>	<a href="#">&lt;(ac)&gt;</a>	<a href="#">&lt;(ad)&gt;</a>	<a href="#">&lt;(ae)&gt;</a>	<a href="#">&lt;(af)&gt;</a>
<a href="#">&lt;b&gt;</a>			<a href="#">&lt;(bc)&gt;</a>	<a href="#">&lt;(bd)&gt;</a>	<a href="#">&lt;(be)&gt;</a>	<a href="#">&lt;(bf)&gt;</a>
<a href="#">&lt;c&gt;</a>				<a href="#">&lt;(cd)&gt;</a>	<a href="#">&lt;(ce)&gt;</a>	<a href="#">&lt;(cf)&gt;</a>
<a href="#">&lt;d&gt;</a>					<a href="#">&lt;(de)&gt;</a>	<a href="#">&lt;(df)&gt;</a>
<a href="#">&lt;e&gt;</a>						<a href="#">&lt;(ef)&gt;</a>
<a href="#">&lt;f&gt;</a>						

Sans la propriété d'anti-monotonie  
 $8*8+8*7/2=92$  candidats

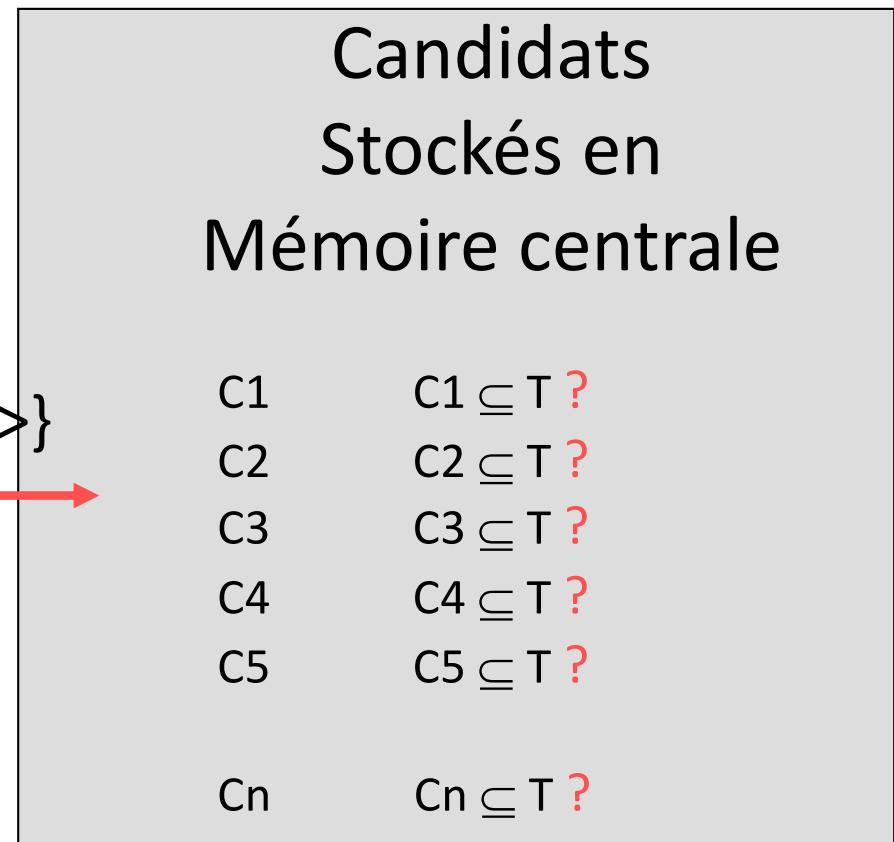


# Comptage des supports des candidats



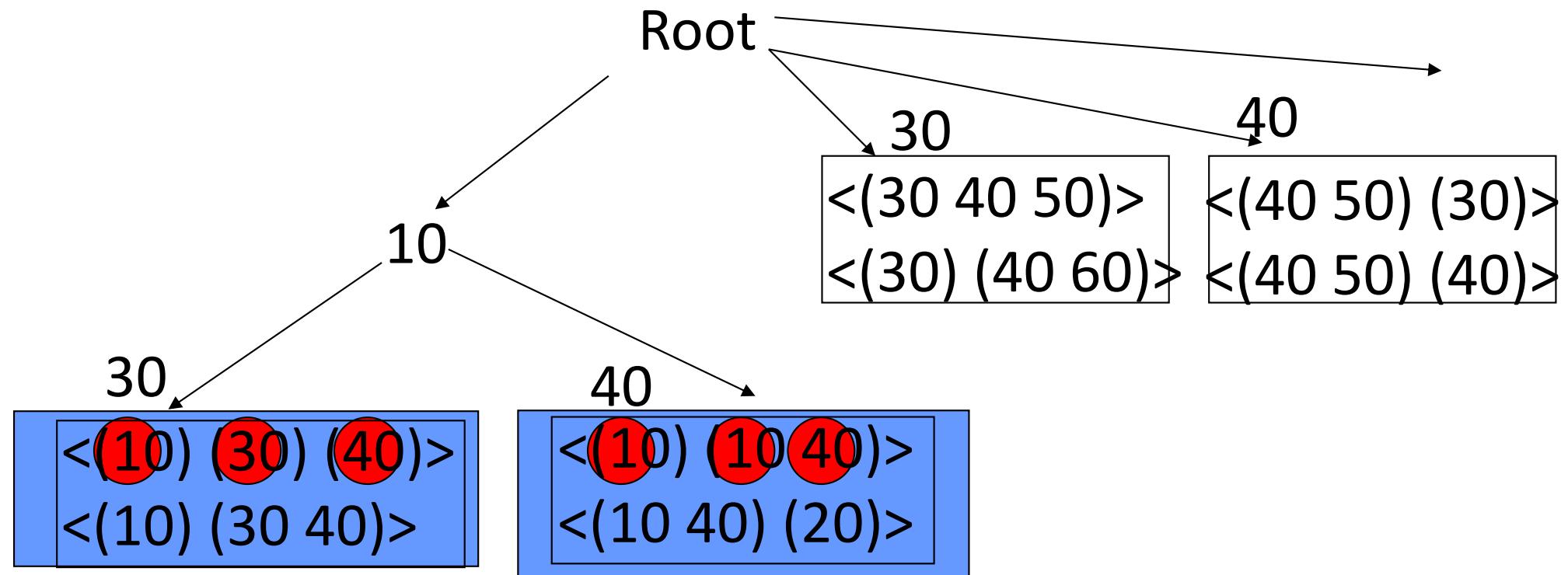
Un tuple  
 $T=\{cid, \langle(a) (bc) (d)\rangle\}$

BASE DE DONNEES



MEMOIRE CENTRALE

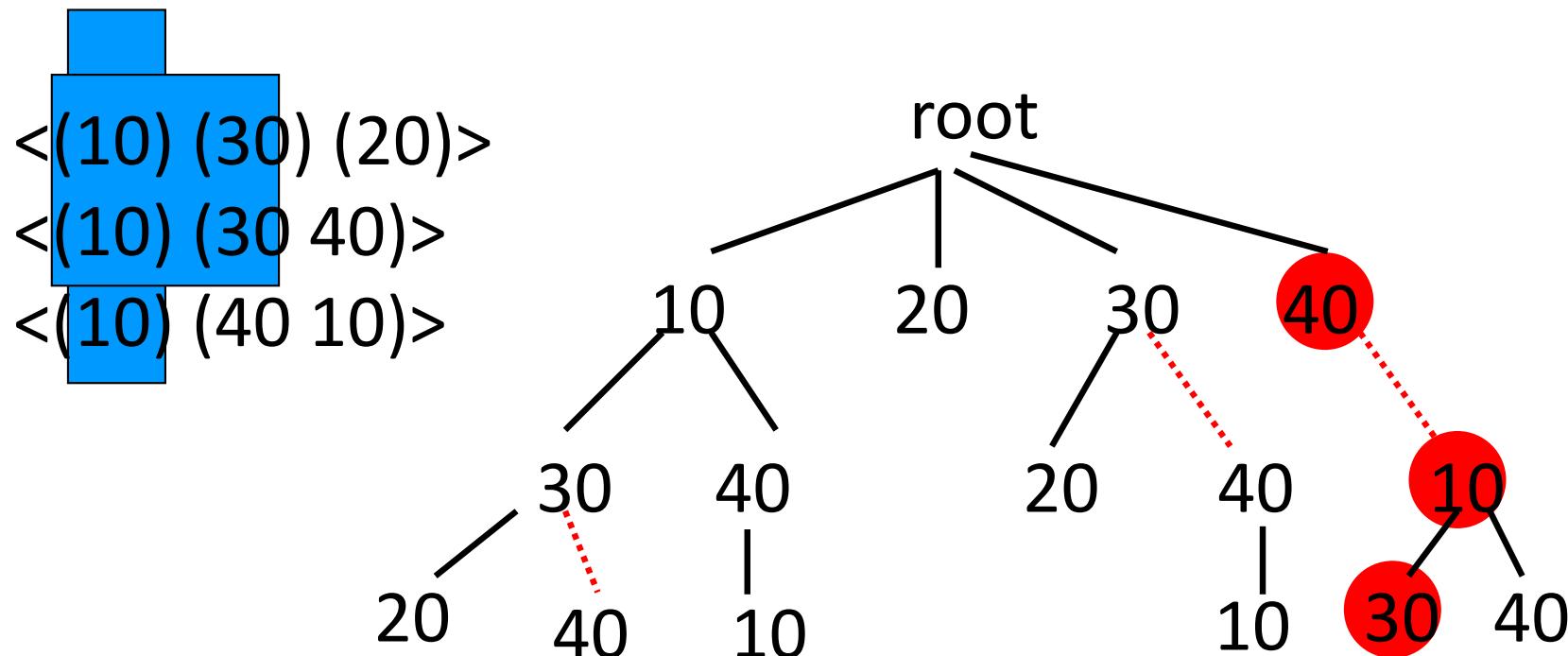
# Stockage des candidats



$$S = \langle(10) (30) (10 40)\rangle$$

# PSP (Prefix Tree for Sequential Patterns)

- Vers une structure plus efficace : prefix tree

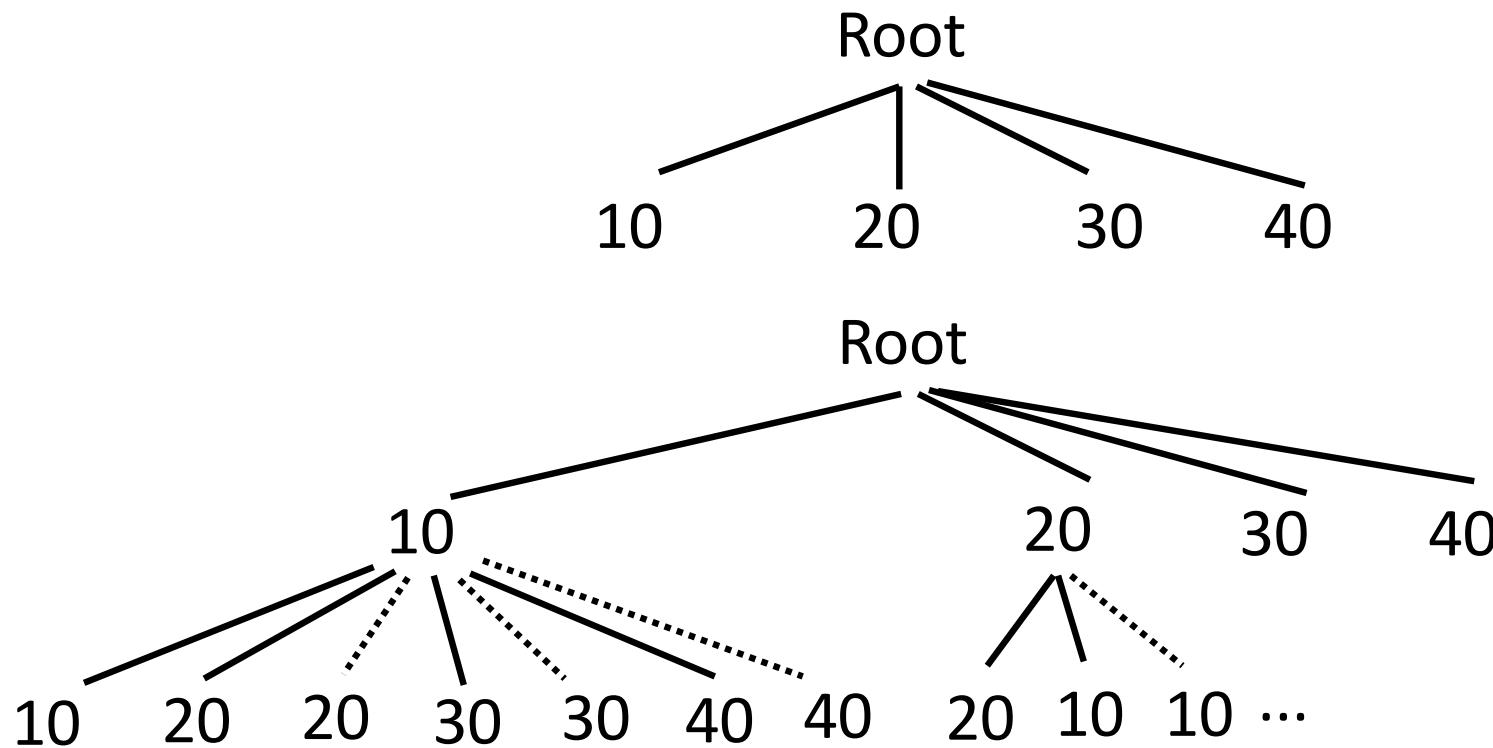


<(30 50 40 10) (30) (20 10)>

# PSP (cont.)

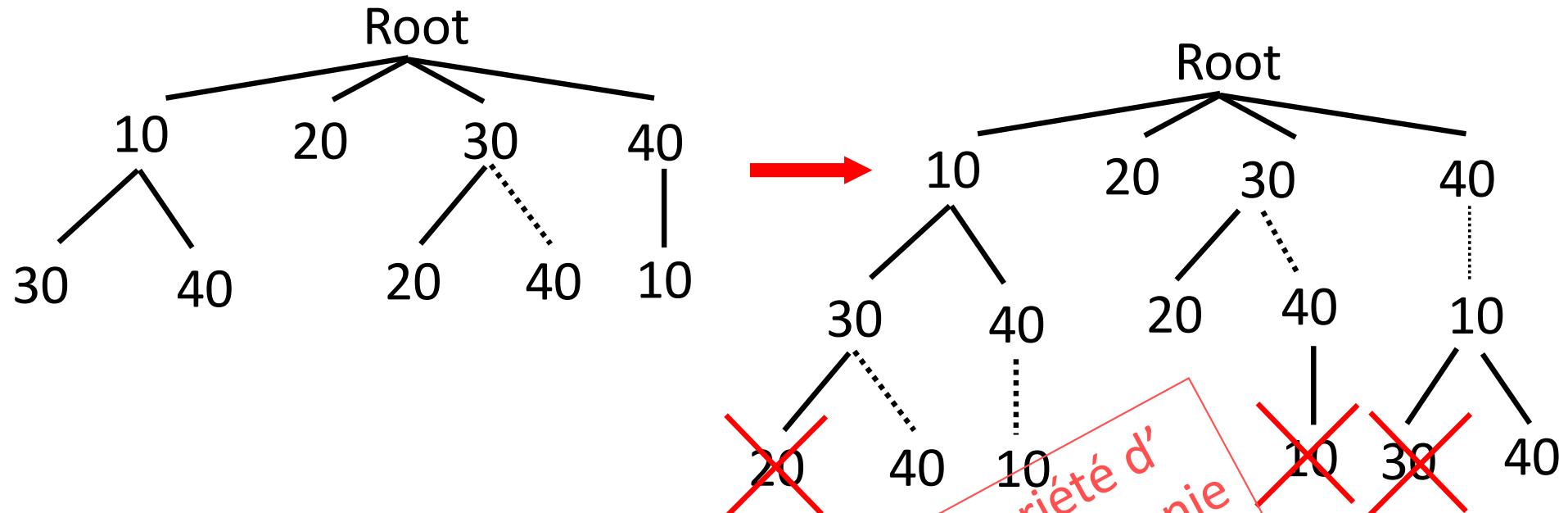
---

## □ Génération des candidats de taille 2



# PSP (cont.)

## □ Génération des candidats de taille $> 2$



Candidats et fréquents  
dans le même arbre



# SPAM

---

- Utilisation de bitmaps pour rechercher les motifs fréquents
- Hypothèse : la base tient toujours en mémoire
- On construit d'un arbre lexicographique contenant toutes les branches possibles – élimination des branches en fonction du support
- Nouvelle représentation des données



# SPAM (cont.)

---

- Représentation verticale des données

$$C1 = \langle (1)_3 \ (1)_5 \rangle$$

<b>C1</b>	(1)	
	T1	0
	T2	0
	T3	1
	T4	0
	T5	1



# SPAM (cont.)

---

- Toutes les données

$$C_2 = \langle (1)_2 (2)_4 \rangle$$

- S-Extension
- I-Extension

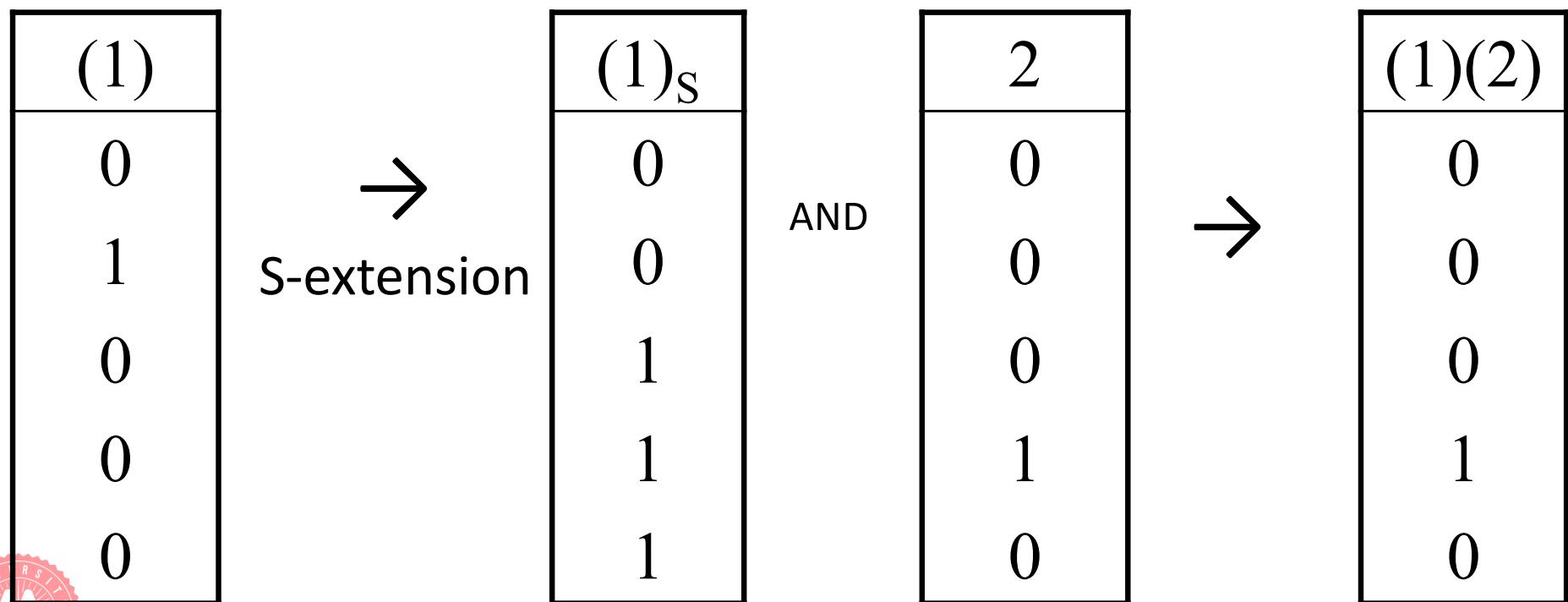
		(1)	(2)
	T1	0	0
C2	T2	1	0
	T3	0	0
	T4	0	1
	T5	0	0



# SPAM (cont.)

- S-Extension : un bitmap transformé + AND
- I-Extension : AND
- Exemple : recherche du candidat (1) (2)

Pour le client C1 précédent



# L'algorithme SPADE

---

- SPADE (*Sequential PAttern Discovery using Equivalent Class*) - M. Zaki (Machine Learning 01)
- Représentation verticale des données
- Une base de séquence est transformée en :
  - Item : <SID, EID>
- La recherche des motifs est réalisée en étendant les sous séquences un item à la fois via la génération des candidats d'Apriori



# L'algorithme SPADE

Sid	Sequence
1	(A) (A B C) (A C) (D) (C F)
2	(A D) (C) (B C) ( A E)
3	(E F) (A B) (D F) (C) (B)
4	(E) (G) (A F) (C) (B) (C)



SID	Eid	Items
1	1	A
1	2	A B C
1	3	A C
1	4	D
1	5	C F
2	1	A D
2	2	C
2	3	B C
2	4	A E
3	1	E F
3	2	A B
3	3	D F
3	4	C
3	5	B
....	...	37



# L'algorithme SPADE

<b>SID</b>	<b>Eid</b>	<b>Items</b>
1	1	A
1	2	A B C
1	3	A C
1	4	D
1	5	C F
2	1	A D
2	2	C
2	3	B C
2	4	A E
3	1	E F
3	2	A B
3	3	D F
3	4	C
3	5	B
		...

<b>A</b>	
<b>Sid</b>	<b>eid</b>
1	1
1	2
1	3
2	1
2	4
3	2
4	3

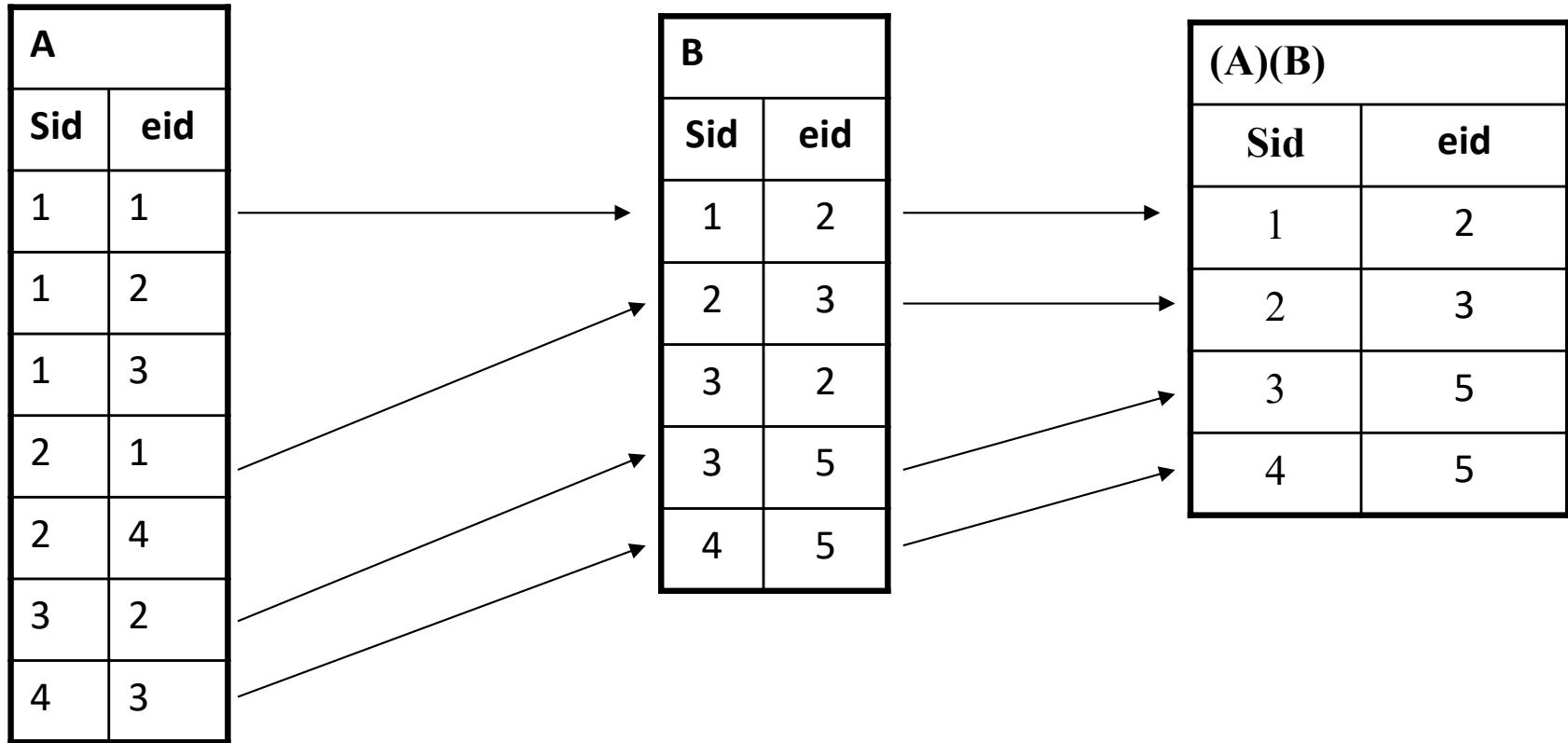
<b>B</b>	
<b>Sid</b>	<b>eid</b>
1	2
2	3
3	2
3	5
4	5

<b>(A)(B)</b>	
<b>Sid</b>	<b>eid</b>
1	2
2	3
3	5
4	5

Listes d'occurrences



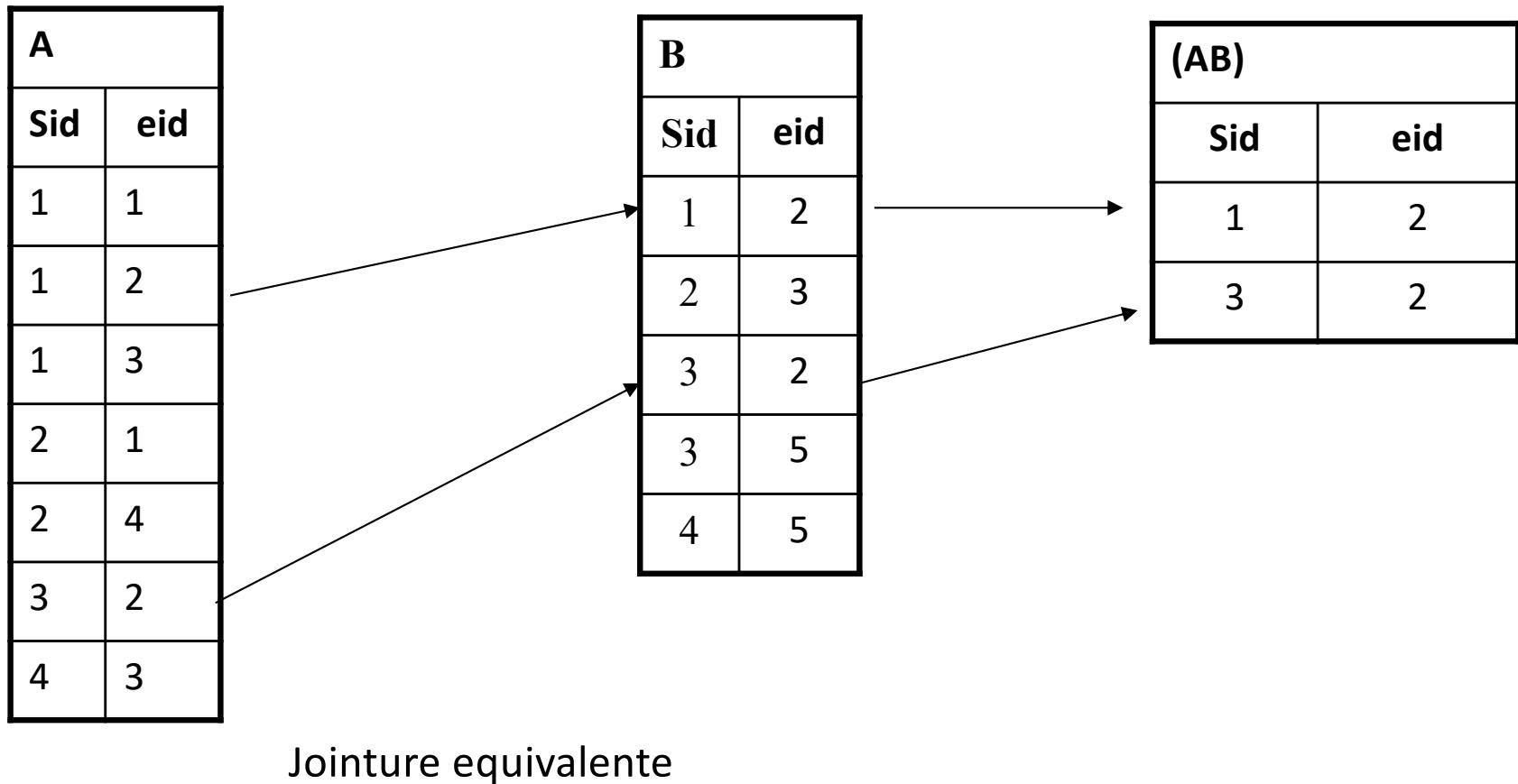
# L'algorithme SPADE



Jointure temporelle

Seuls les couples  $\langle(1,1), (1,2)\rangle$ ,  $\langle(2,1) (2,3)\rangle$ ,  $\langle(3,2), (3,5)\rangle$   
et  $\langle(4,3), (4,5)\rangle$  donnent lieu à de nouvelles occurrences du motifs  
(A) (B)

# L'algorithme SPADE



Seuls les couples  $\langle(1,2), (1,2)\rangle$  et  $\langle(3,2), (3,2)\rangle$  donnent lieu à de nouvelles occurrences du motifs (A B)

# FreeSpan

---

- Han et al. (KDD 00)
- Stratégie : « Divide and Conquer »
  - Projeter la base de données de séquences récursivement en un ensemble de bases plus petites
  - Fouiller chaque base projetée pour rechercher les sous ensembles d'un motif



# FreeSpan

---

- Préfixe et Suffixe (Projection)
- $\langle a \rangle$ ,  $\langle aa \rangle$ ,  $\langle a(ab) \rangle$  et  $\langle a(abc) \rangle$  sont des préfixes de la séquence  $\langle a(abc)(ac)d(cf) \rangle$
- Soit la séquence  $\langle a(abc)(ac)d(cf) \rangle$

Préfixe	<u>Suffixe</u> (Prefix-Based <u>Projection</u> )
$\langle a \rangle$	$\langle (abc)(ac)d(cf) \rangle$
$\langle aa \rangle$	$\langle (_bc)(ac)d(cf) \rangle$
$\langle a(ab) \rangle$	$\langle (_c)(ac)d(cf) \rangle$



# FreeSpan

- **Etape 1 :** rechercher toutes les séquences de taille 1 et les lister en les triant en fonction de la valeur du support

$f\_list = A:4, B:4, C:4, D:3, E:3, F:3; G:1$

- **Etape 2:** Diviser l'espace de recherche. L'ensemble de motifs peut être partitionné en 6 ensembles disjoints (parcourir la  $f\_list$ ) :
  - Ceux qui contiennent uniquement le préfixe A
  - Ceux qui contiennent uniquement le préfixe B
  - Ceux qui contiennent uniquement le préfixe C
  - Ceux qui contiennent uniquement le préfixe D
  - Ceux qui contiennent uniquement le préfixe E
  - Ceux qui contiennent le préfixe F

Sid	Séquence
10	(A) (A B C) (A C) (D) (C F)
20	(A D) (C) (B C) ( A E)
30	(E F) (A B) (D F) (C) (B)
40	(E) (G) (A F) (C) (B) (C)



# FreeSpan

- Considérer uniquement les projections pour  $\langle A \rangle$
- $\langle A \rangle$ -projected database:
  - $\langle (ABC)(AC)D(CF) \rangle, \langle (\_D)C(BC)(AE) \rangle, \langle (\_B)(DF)CB \rangle, \langle (\_F)CBC \rangle$
- Rechercher tous les 2-seq. pat. ayant comme préfixe  $\langle A \rangle$  :  $\langle AA \rangle:2, \langle AB \rangle:4, \langle (AB) \rangle:2, \langle AC \rangle:4, \langle AD \rangle:2, \langle AF \rangle:2$ 
  - Partitionner en 6 sous-ensembles
    - Ayant le préfixe  $\langle AA \rangle$ ;
    - ...
    - Ayant le préfixe  $\langle AF \rangle$

Sid	Sequence
10	(A) (A B C) (A C) (D) (C F)
20	(A D) (C) (B C) ( A E)
30	(E F) (A B) (D F) (C) (B)
40	(E) (G) (A F) (C) (B) (C)

F\_list = A:4, B:4, C:4, D:3, E: 3, F: 3



# FreeSpan

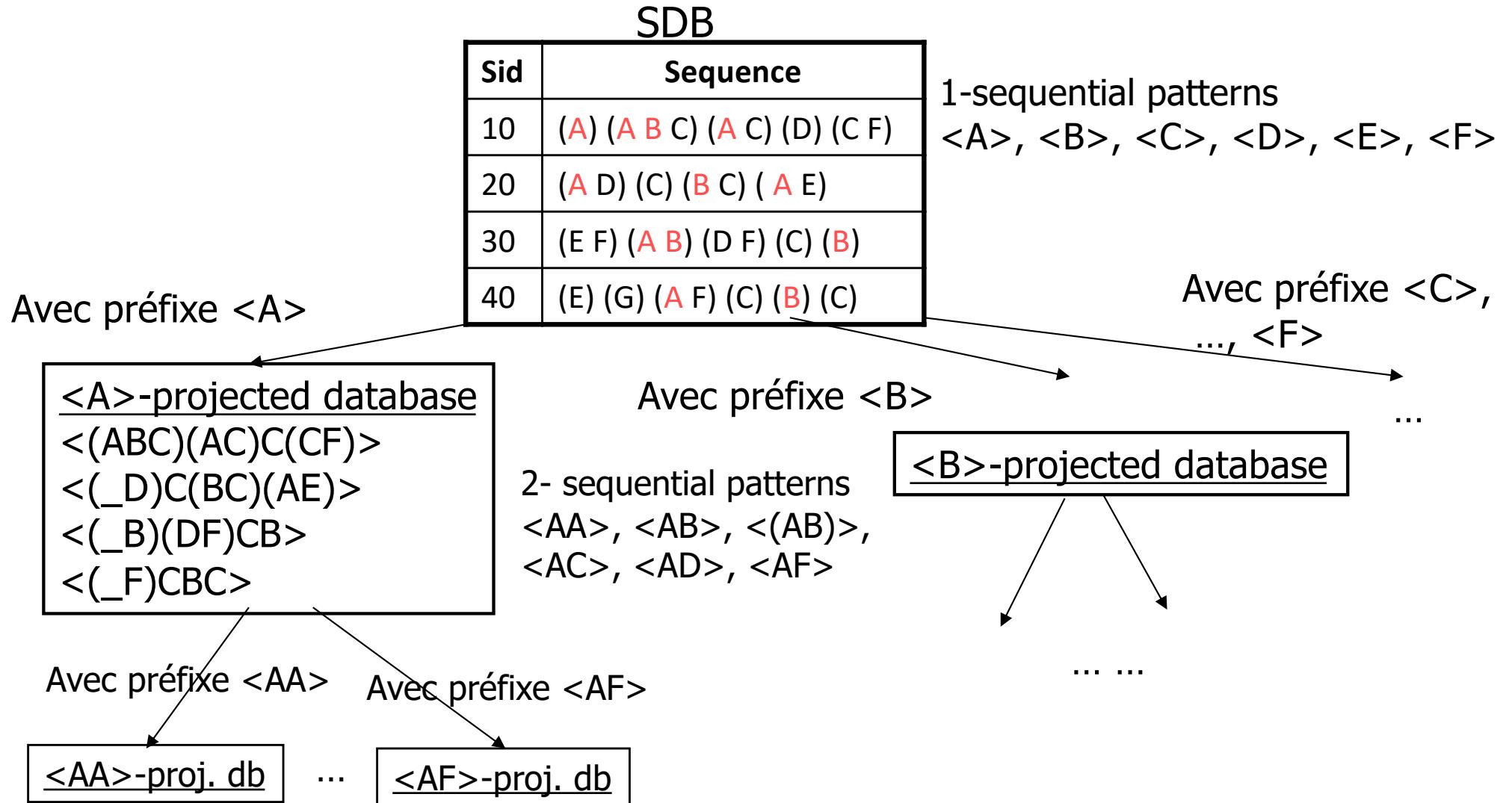
Sid	Sequence
10	(A) (A B C) (A C) (D) (C F)
20	(A D) (C) (B C) ( A E)
30	(E F) (A B) (D F) (C) (B)
40	(E) (G) (A F) (C) (B) (C)

F\_list = A:4, B:4, C:4, D:3, E: 3, F: 3

- Considérer uniquement les projections pour <AA>
- <AA>-projected database:
  - (\_BC)(AC)D(CF) et (\_E)
- ...



# Complétude de FreeSpan



# Plan

---

- Contexte général
- Motifs séquentiels
- Extensions des motifs séquentiels
- Quelques applications des motifs
- Conclusions



# Motifs généralisés

---

- Pour certains domaines d'applications il est nécessaire de limiter les résultats  
*corrélations entre achat du caviar le 1er janvier et de champagne le 31 décembre ?*
- Contraintes de temps
  - windowSize* : regrouper des événements
  - minGap* : considérer des événements comme trop proches
  - maxGap* : considérer des événements comme trop éloignés



# Illustration

Client	Date	Items
C1	1	<i>Ringworld</i>
C1	2	<i>Foundation</i>
C1	15	<i>Ringworld Engineers, Second Foundation</i>
C2	1	<i>Foundation, Ringworld</i>
C2	20	<i>Foundation and Empire</i>
C2	50	<i>Ringworld Engineers</i>

Support = 50% : <(Ringworld) (Ringworld Engineers)> et  
 <(Foundation) (Ringworld Engineers)>

windowSize=7 jours : <(Foundation, Ringworld) (Ringworld  
 Engineers)>



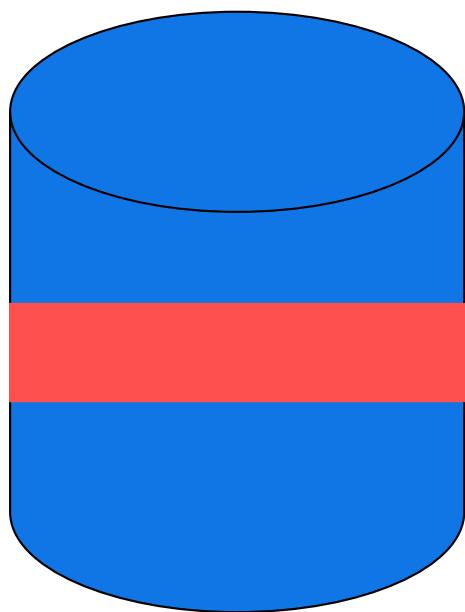
# Contraintes temporelles

---

- $d = \langle^1(1) ^2(2\ 3) ^3(4) ^4(5\ 6) ^5(7) \rangle$
- Candidat :  $C = \langle(1\ 2\ 3\ 4)\ (5\ 6\ 7) \rangle$ 
  - $\text{windowSize} = 3$ ,  $\text{minGap}=0$ ,  $\text{maxGap} = 5$ ,
  - $d = \langle(1\ 2\ 3\ 4)\ (5\ 6\ 7) \rangle$  - Donc C est inclus dans d
- Candidat :  $C = \langle(1\ 2\ 3)\ (6\ 7) \rangle$ 
  - $\text{windowSize} = 1$ ,  $\text{minGap}=3$ ,  $\text{maxGap} = 4$ ,
  - $d = \langle(1\ 2\ 3)\ (4)\ (5\ 6\ 7) \rangle$
  - $\text{minGap}$  pas respecté entre 3 et 5 ! C pas inclus dans d



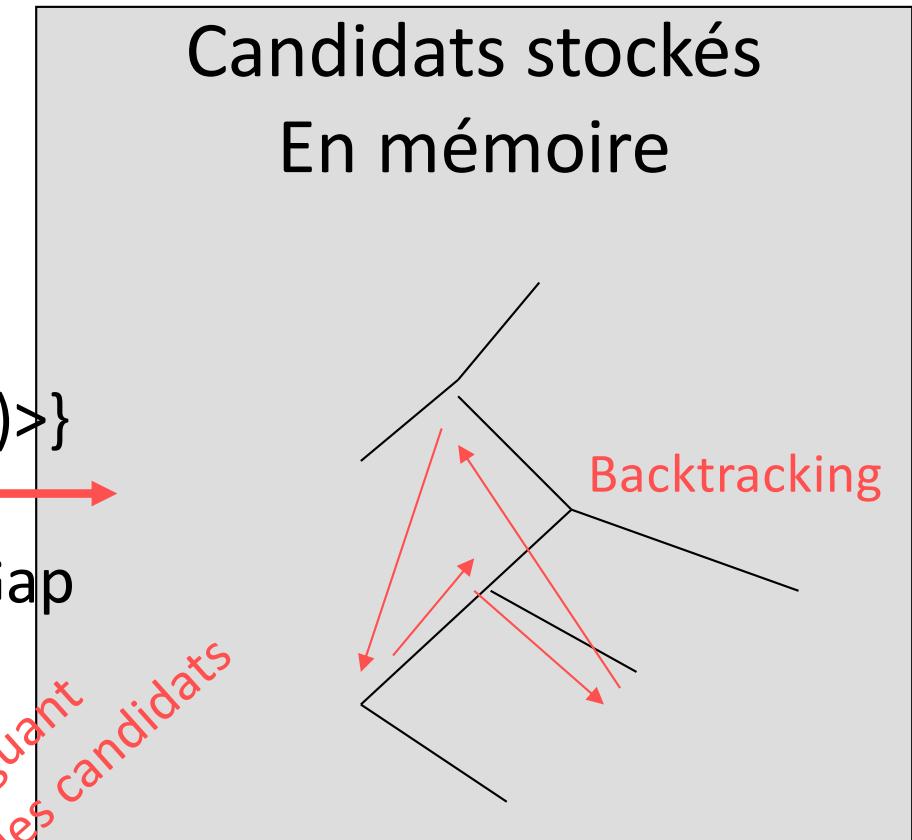
# Comment gérer les contraintes ?



BASE DE DONNEES

Un tuple  
 $T=\{cid, \langle(a) (bc) (d)\rangle\}$   
+  
wS, minGap, maxGap

En naviguant  
Entre les candidats



MEMOIRE CENTRALE

# Inclusion des contraintes

---

Client	Date	Items
C1	1	10
C1	7	20
C1	13	30
C1	17	40
C1	18	50
C1	24	60

minGap=1  
windowSize=5



<(10) (20) (30) (50) (60)>  
<(10) (20) (30) (40) (60)>  
**<(10) (20) (30) (40 50) (60)>**  
<(10) (20) (30 40) (60)>  
**<(10) (20) (30 40 50) (60)>**

# Recherche des inclusions

Date	1	7	13	17	18	24
C	1	2	3	4	5	6

windowSize = 5, minGap = 1

Via minGap

$\langle(1) (2) (3) (4) (6)\rangle$

$\langle(1) (2) (3) (5) (6)\rangle$

Puis avec windowSize

$\langle(1) (2) (3) (4\ 5) (6)\rangle$

$\langle(1) (2) (3\ 4) (6)\rangle$

$\langle(1) (2) (3\ 4\ 5) (6)\rangle$

En fait:

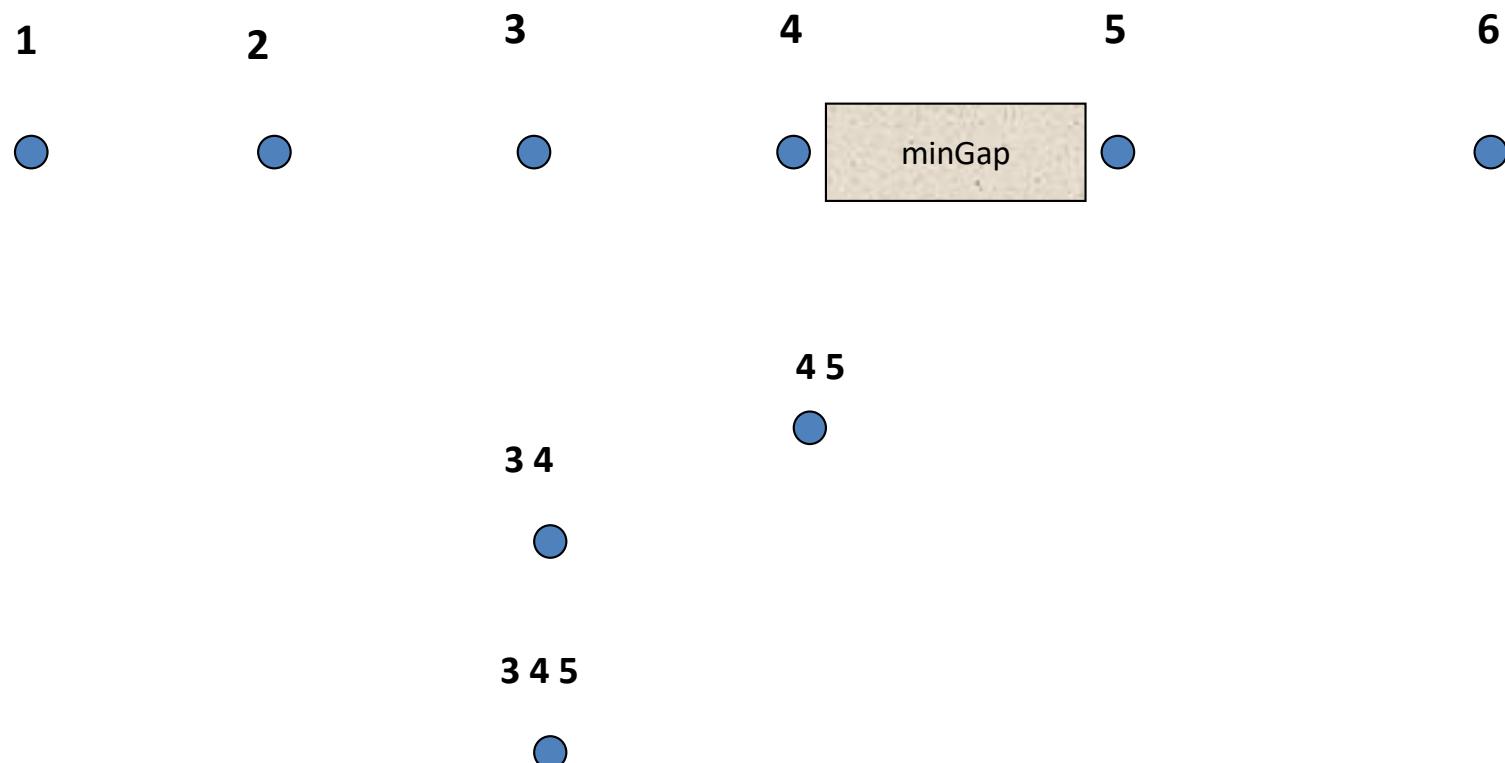
$\langle(1) (2) (3) (4\ 5) (6)\rangle$

et  $\langle(1) (2) (3\ 4\ 5) (6)\rangle$

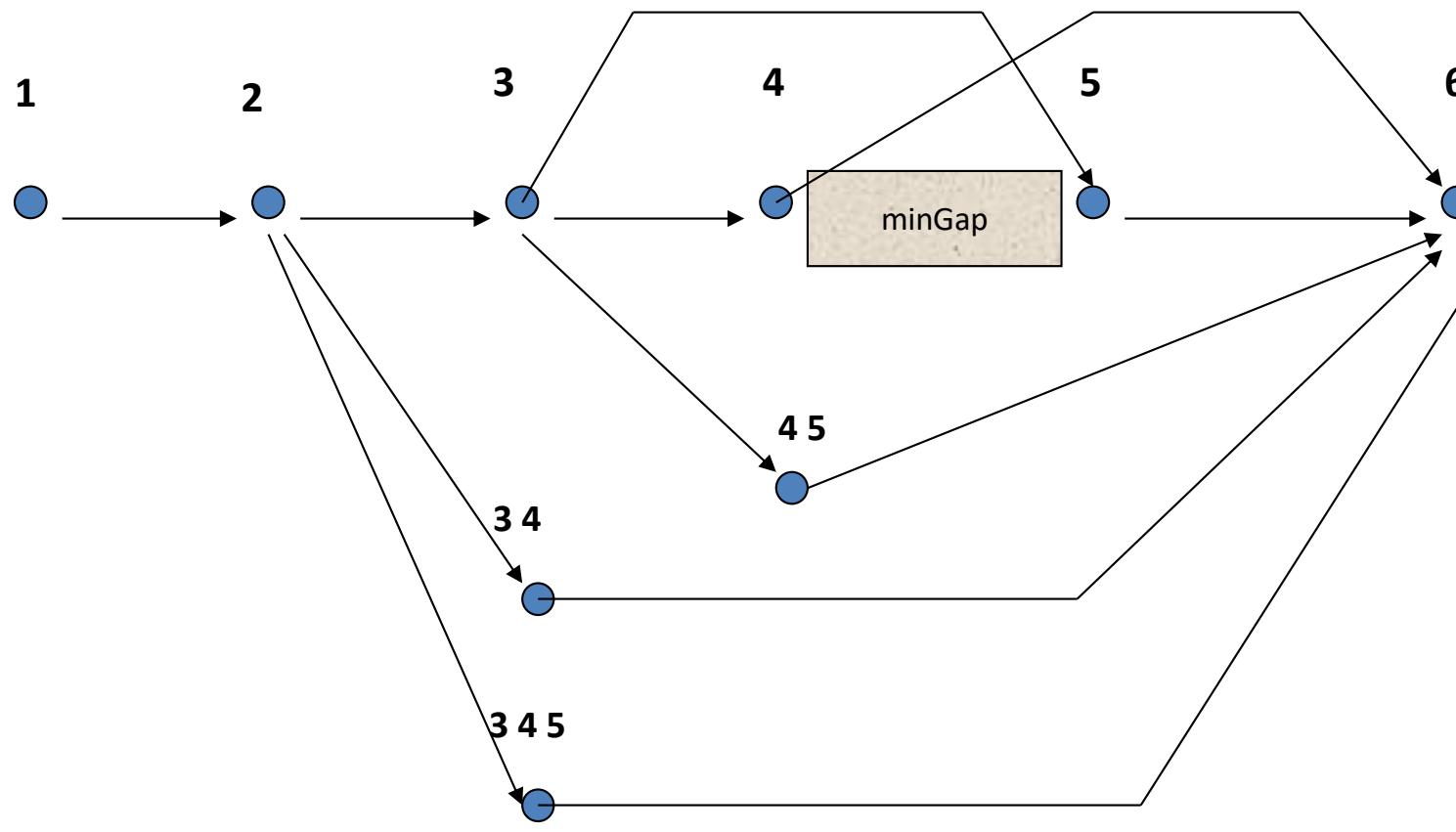


# Recherche des inclusions (cont.)

Date	1	7	13	17	18	24
C	1	2	3	4	5	6



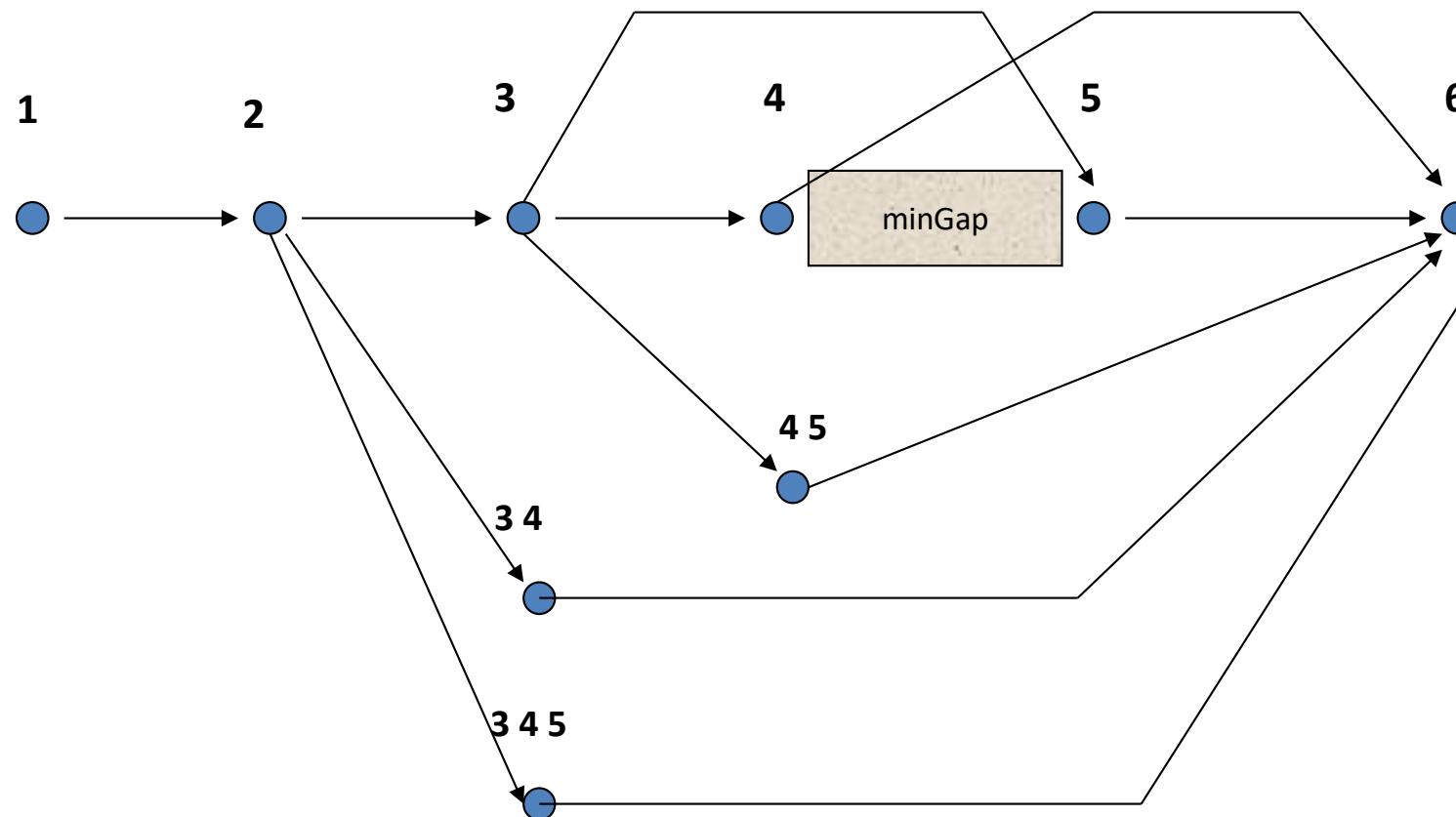
# Recherche des inclusions (cont.)



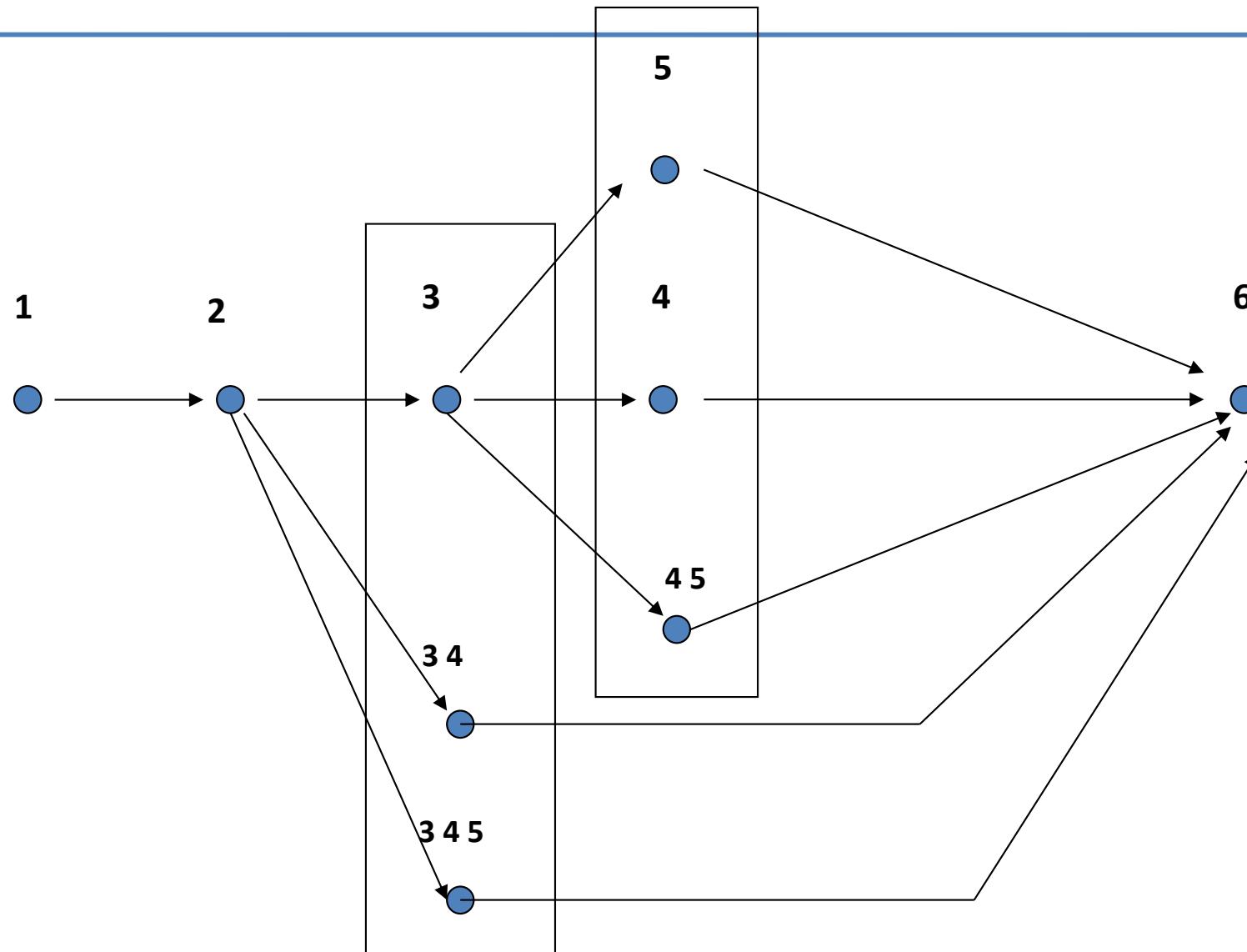
Tous les chemins mais quid des inclusions

# Recherche des inclusions (cont.)

---

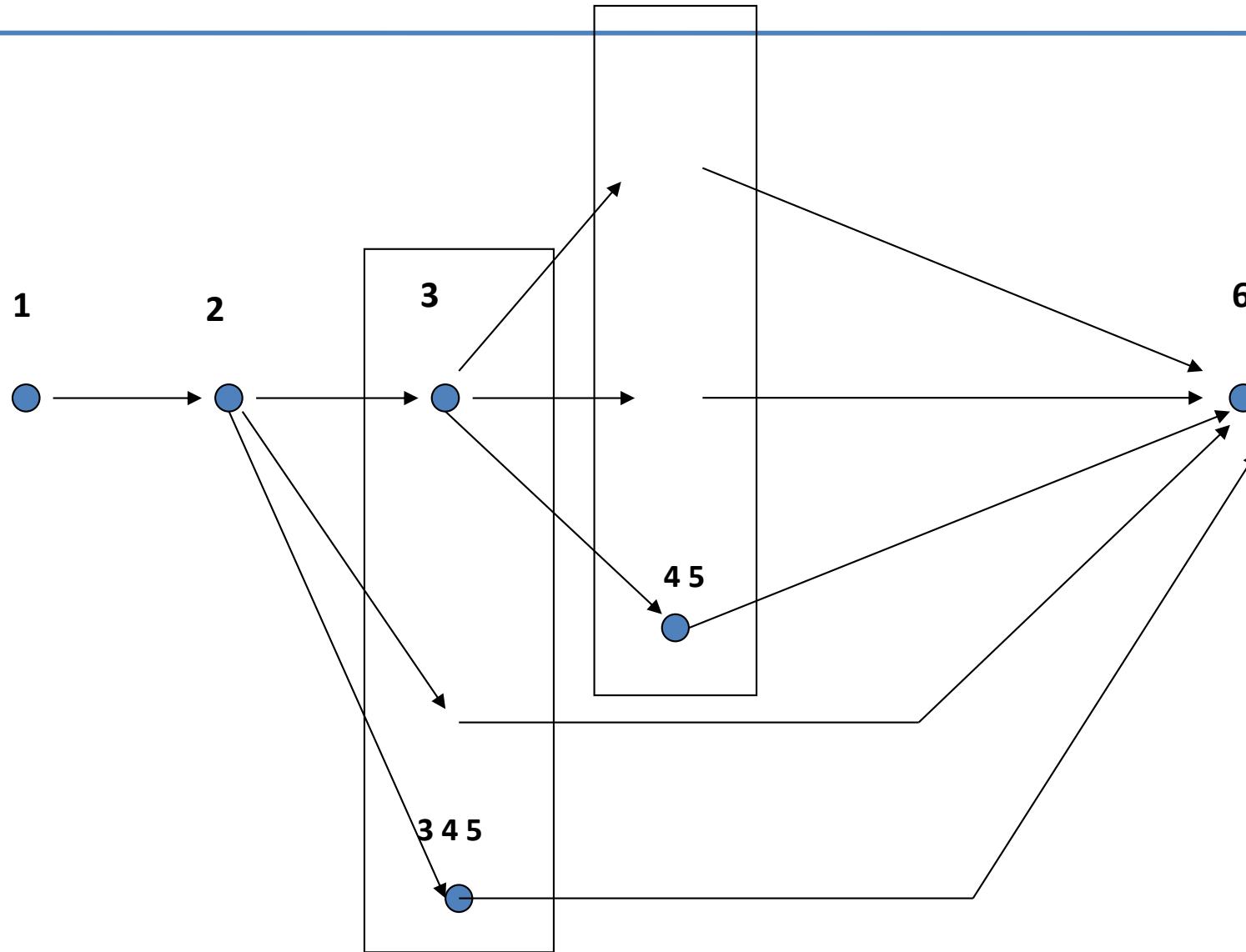


# Recherche des inclusions (cont.)



Regrouper ensemble les nœuds de même origine

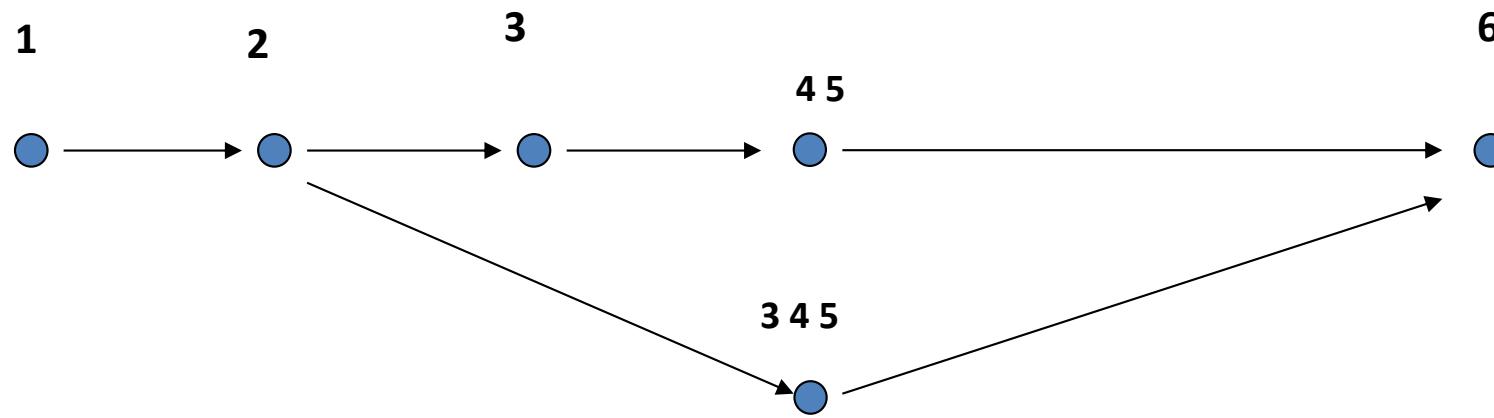
# Recherche des inclusions (cont.)



Supprimer les nœuds inclus ayant même destination

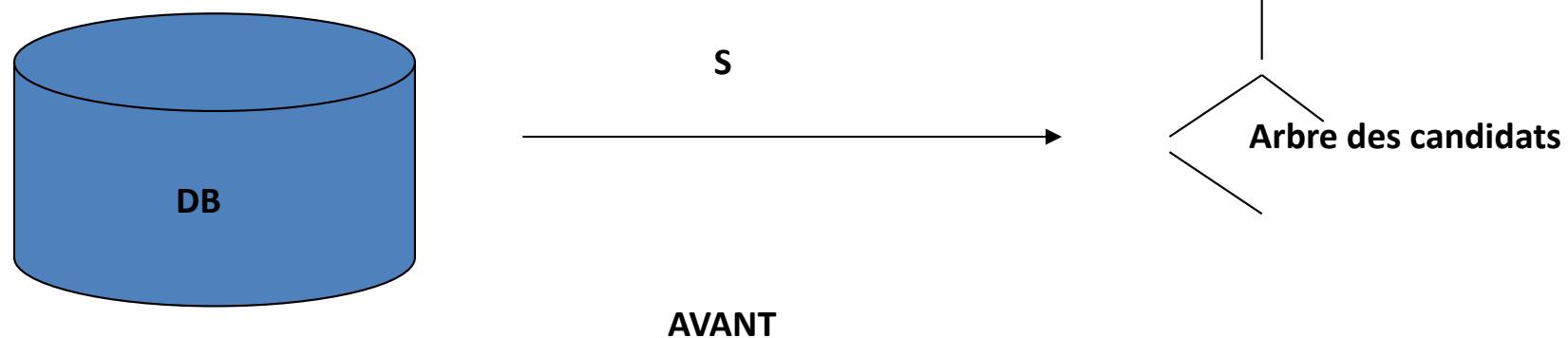
# Recherche des inclusions (cont.)

---



$\langle(1)(2)(3)(4\ 5)(6)\rangle$  et  $\langle(1)(2)(3\ 4\ 5)(6)\rangle$

# Recherche des inclusions (cont.)



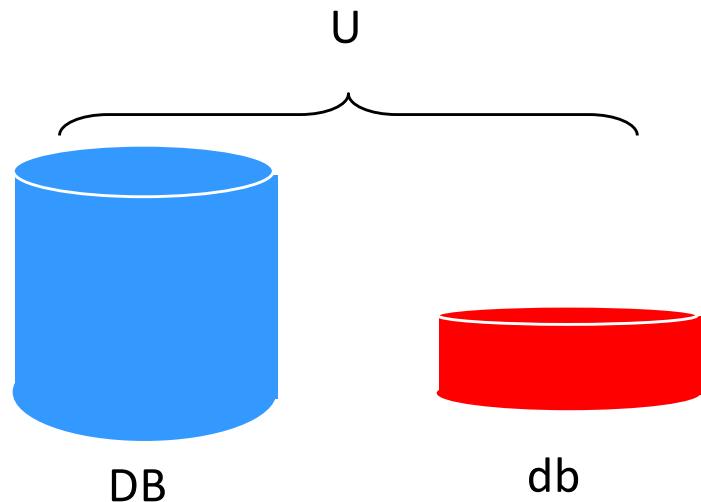
# Recherche Incrémentale

---

- Quid des connaissances extraites au préalable ?
- Est-il possible d'éviter de recommencer une extraction lors de l'ajout de nouvelles données ?
- L'approche Ise (*Incremental Sequential Extraction*)
- Point de départ :
  - des séquences fréquentes avec leur support
  - une base d'origine
  - une base incrément



# ISE



$L^{DB}$ =fréquents de DB

$k$  = taille du plus grand fréquent de  $L^{DB}$

- 1) rechercher toutes les nouvelles séquences de taille  $j \leq k+1$
- 2) rechercher toutes les séquences de taille  $j > k+1$



- 1<sup>ère</sup> itération
  - rechercher les 1-séquences dans db
  - combiner avec  $L_{DB}^1$ 
    - 1-séquences fréquentes sur U
- {
  - générer des 2-candidats
  - rechercher les sous-séquences fréquentes de  $L_{DB}$  qui précédent les items fréquents de db
    - séquences fréquentes de taille  $j \leq k+1$  (*freqSeed*)
    - 2-séquences fréquentes dans db (2-*freqExt*)



# ISE

---

- $j^{\text{ième}}$  itération ( $j < k+1$ )
  - générer des candidats à partir de freqSeed et de j-freqExt ( $candInc$ ) –
  - générer des  $(j+1)$ -candExt
    - => Extensions possibles :  $(j+1)$ -freqExt
    - => freqInc
- $L^U : L^{DB}$  et les séquences maximales des ensembles précédents
- Extraire les  $(k+1)$ -séquences



# Conclusion

---

- $t(DB) + t(db) < t(DB+db)$
- Indépendant de l'algorithme d'extraction
- Sans doute trop de passage sur la base
- Quand exécuter ISE ?



# Recherche Incrémentale

---

- IncSpan
- Idée : Maintenir un ensemble de séquences « presque fréquentes »
  - Séquence **fréquente** :  $\text{support}(s) \geq \sigma$
  - Séquence **semi-fréquente** :  $\text{support}(s) < \sigma$  et  $\text{support}(s) \geq \mu$  ( $0 \leq \mu \leq 1$ )
  - Séquence **non fréquente** :  $\text{support}(s) < \mu * \sigma$



# IncSpan

---

- Principe : Appliquer PrefixSpan en gérant les différents ensembles fréquents et semi fréquents
- Avantages : pas de parcours sur la base car pas de candidats à gérer



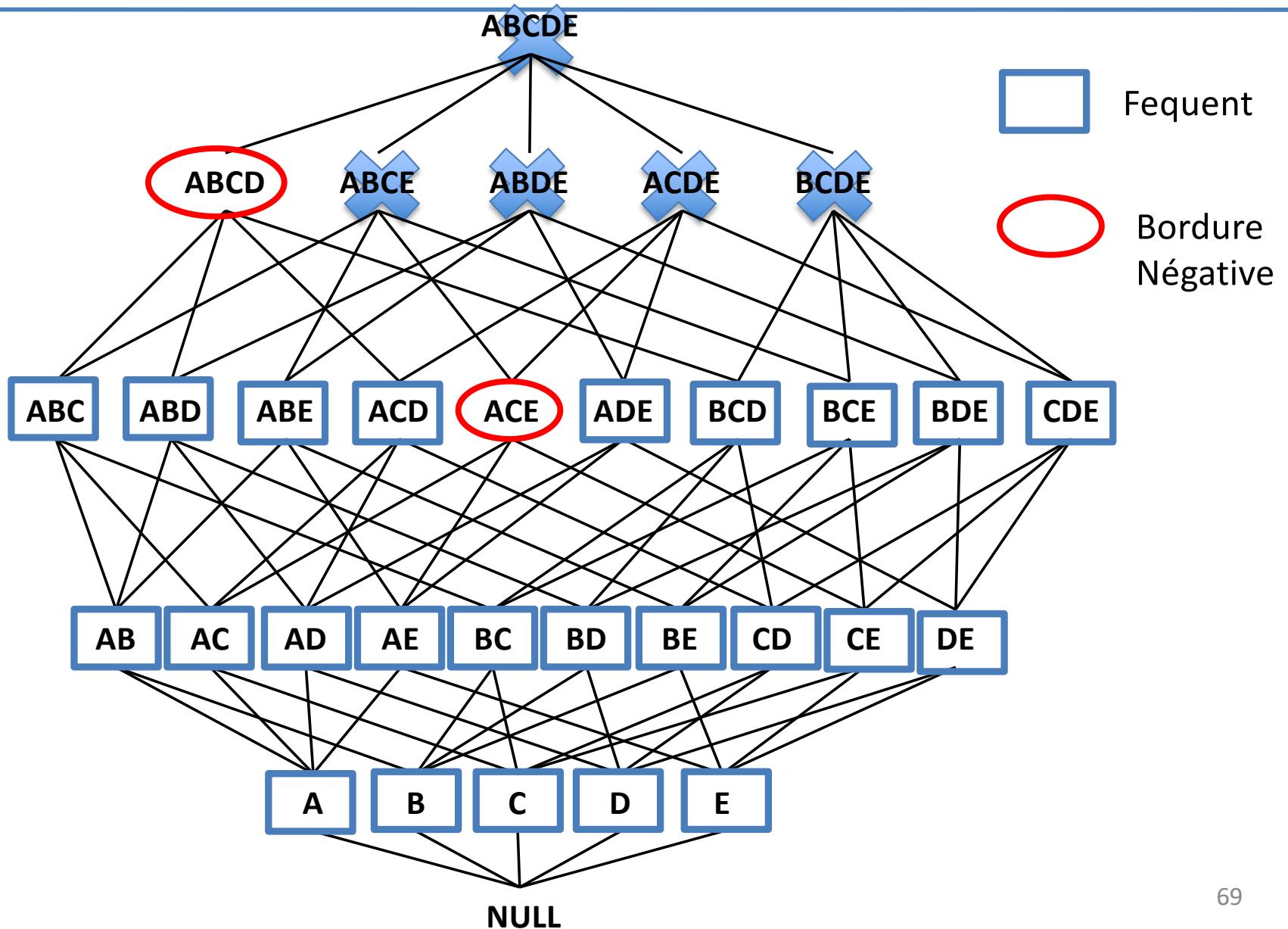
# Recherche Incrémentale

---

- ISM [S. Parhasarathy et al, 99]
- Idée : tenir compte de la bordure négative
- Avantages : éviter de repartir de l'origine !



# Bordure Négative



# ISM

---

- Incremental sequence lattice( ISL)
- Toutes les séquences fréquentes et toutes les séquences dans la bordure négatives
- 2 phases
  - Phase 1 : Mettre à jour les supports des éléments dans *NB* et *FS*.
  - Phase 2 : Prendre en compte les éléments qui ne l'étaient pas précédemment



# ISM

---

- Avantage : efficace
- Inconvénients : Trop d'éléments à stocker
- La bordure négative peut être très importante



# Approches incrémentales

---

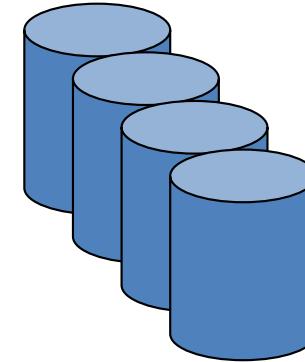
- Quelques remarques
- Peu d'approches pourtant le problème est important
- Quand exécuter l'algorithme
  - Trop tôt : inutile
  - Trop tard : inutile



# Motifs séquentiels & BD distribuées

- Un ensemble de bases de données

$$\text{DB} = \text{DB}_1 \cup \text{DB}_2 \dots \text{DB}_n$$



- Extraire les séquences fréquentes

Pour c1 : Alice (1)<sub>1</sub>, Bob (2)<sub>2</sub>, Carol (7)<sub>4</sub>, Alice (3)<sub>5</sub>

CID	Alice	Bob	Carol
1	(1) <sub>1</sub> (3) <sub>5</sub>	(2) <sub>2</sub>	(7) <sub>4</sub>
2	(2) <sub>4</sub>	(1) <sub>3</sub>	(3) <sub>6</sub>
3	(2) <sub>6</sub> (3) <sub>7</sub>		(1) <sub>2</sub> (7) <sub>3</sub>

# Motifs séquentiels collaboratifs

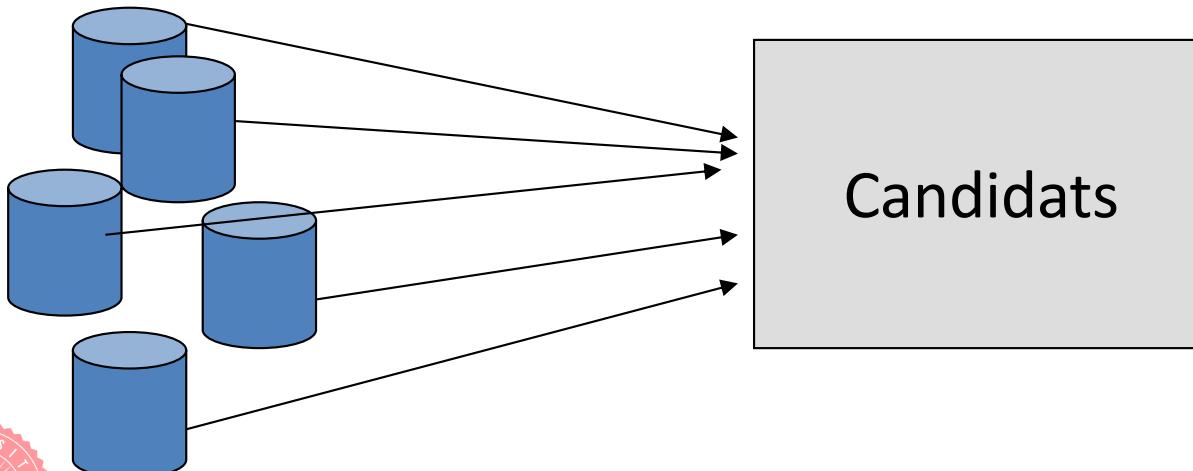
- Une représentation verticale des items en fonction des bases

		$V_1^1$	$V_1^D$
C1	T1	0	1
	T2	0	0
	T3	1	0
	T4	1	1
			...
C2	T1	0	1
	T2	1	0
	T3	0	0
	T4	0	1

# Motifs séquentiels collaboratifs (cont.)

---

- Hypothèse : un algorithme de génération-vérification est disponible
- Génération : combiner les  $k-1$  séquences fréquentes pour générer des candidats de taille  $k$
- Améliorer la phase de vérification



# Motifs séquentiels collaboratifs (cont.)

---

- Exemple : Recherche de (1) (2) dans  $DB_1, DB_2$ 
  1. Demander à  $DB_1$  et  $DB_2$  leur vecteur correspondant à l'item spécifique (1) ( $V_1^1$  et  $V_1^2$ )
  2. Faire un *OR* logique entre  $V_1^1$  et  $V_1^2$
  3. Utiliser la S-extension de SPAM (fonction F)
  4. Demander à  $DB_1$  et  $DB_2$  le vecteur pour (2) ( $V_2^1$  et  $V_2^2$ )
  5. Faire un *OR* logique entre  $V_2^1$  et  $V_2^2$
  6. Appliquer un *AND* entre les deux (fonction G)
  7. Convertir le bitmap en entier et compter le nombre de 1 (fonction  $\Sigma$ )



# Motifs séquentiels collaboratifs (cont.)

		$V_1^1$
	T1	0
C1	T2	0
	T3	1
	T4	0
	T1	0
C2	T2	1
	T3	0
	T4	0

OR

		$V_1^D$
C1	T1	0
	T2	1
	T3	1
	T4	0
C2	T1	1
	T2	0
	T3	0
	T4	0

$$Z1 = f(V_1^1 \\ OR V_1^2)$$

S-Extension

		$Z1$
C1	T1	0
	T2	0
	T3	1
	T4	1
C2	T1	0
	T2	1
	T3	1
	T4	1

		$Z1$
C1	T1	0
	T2	0
	T3	1
	T4	1
C2	T1	0
	T2	1
	T3	1
	T4	1

AND

		$Z2 = V21 OR V22$
C1	T1	0
	T2	0
	T3	0
	T4	1
C2	T1	1
	T2	1
	T3	1
	T4	1

$$\rightarrow G = (Z1 \\ AND Z2)$$

		$Z3$
C1		1
C2		1

$$\rightarrow \sum 2$$



# Préservation de la vie privée

---

- Une contrainte forte :

Alice, Bob et Carol ne peuvent pas donner d'information sur le contenu de leur base

- Problème : Evaluation d'une séquence candidate
  - Est-ce que l'item 1 pour le client 1 appartient à la base de Carol ?



# Préservation de la vie privée (cont.)

---

- Collaboratif ?
- Considérer de nouvelles fonctions sûres ( $\text{AND}^S$ ,  $\text{OR}^S$ ,  $\text{G}^S$ ,  $\text{F}^S$ ,  $\sum^S$ )
- Réaliser la vérification sans fournir d'information des bases de données sources (Alice, Bob, Carol)
- Une nouvelle architecture
  - Trois sites ne collaborant pas (semi-honnêtes)



# Préservation de la vie privée (cont.)

Site « Data Miner »

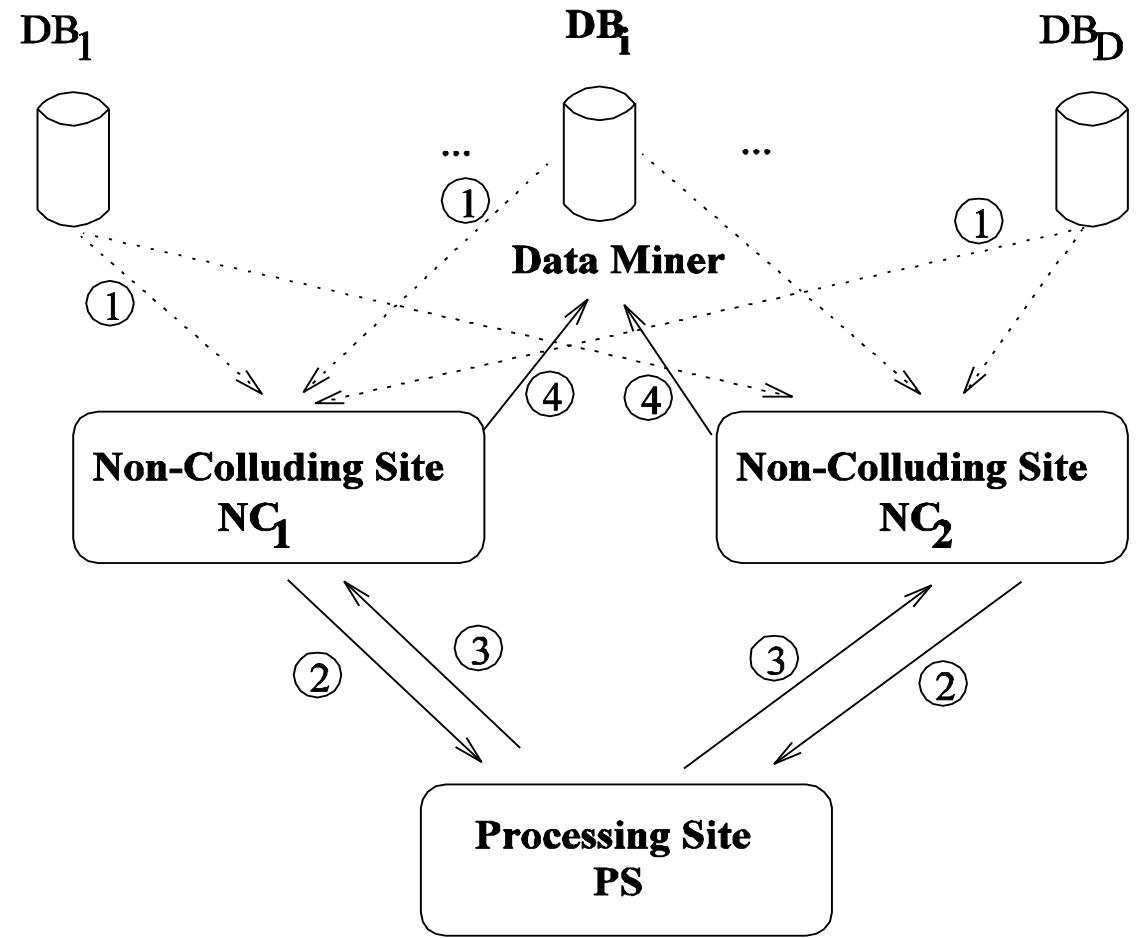
Réalise les opérations de fouille

Sites « Non Colluding »

Réalisent les opérations sûres

Site « Processing »

Calcul des fonctions



# Préservation de la vie privée (cont.)

---

- Etape de prétraitements
- Ajout de faux clients dans les bases de données sources
  - Idée : on augmente le nombre de clients pour éviter qu'une partie puisse obtenir le bon résultat par rapport au support
- Permuter la liste de clients
  - Idée : éviter pour un site de savoir de quel client il traite
- Comptage du support des clients ajoutés
  - Idée : supprimer le bruit au final



# Préservation de la vie privée (cont.)

---

- Rappel :

<b>XOR</b>	1	0
1	0	1
0	1	0



# Préservation de la vie privée (cont.)

---

- Envoi des données des BD sources aux sites « Non Colluding »
- $NC_1$  et  $NC_2$  doivent avoir le minimum d'information
- Pour chaque base  $DB_i$ , pour chaque item  $it$ , générer un vecteur de bits aléatoirement de la même taille que le vecteur  $it$  ( $R_{DBi}$ )
- $Z_{DBi} = V_{it} \text{ XOR } R_{DBi}$
- Envoyer  $Z_{DBi}$  à  $NC_1$  et  $R_{DBi}$  à  $NC_2$  (et vice versa)
- $NC_1$  et  $NC_2$  :  $R_{DBi}$  ou vecteurs XOR-isés



# Préservation de la vie privée (cont.)

---

- Un exemple : le protocole AND<sup>S</sup>
- Entrée :  $(X^+, Y^+ \mid X^-, Y^-)$  sont des bits tels que  $X^+$  et  $Y^+$  appartiennent à  $NC_1$  et  $X^-$  et  $Y^-$  appartiennent à  $NC_2$
- Sortie :  $(A^R \mid B^R)$  sont tels que :  
$$A^R \text{ } XOR \text{ } B^R = (X^+ \text{ } XOR \text{ } X^-) \text{ } AND \text{ } (Y^+ \text{ } XOR \text{ } Y^-)$$



# Préservation de la vie privée (cont.)

---

- NC<sub>1</sub> et NC<sub>2</sub> génèrent mutuellement et s'échangent des nombres aléatoires R<sub>A</sub>, R'<sub>A</sub>, R<sub>B</sub> et R'<sub>B</sub> tels que : X<sup>+</sup>' = X<sup>+</sup> XOR R<sub>A</sub>, Y<sup>+</sup>' = Y<sup>+</sup> XOR R'<sub>A</sub>, X<sup>-</sup>' = X<sup>-</sup> XOR R<sub>B</sub> et Y<sup>-</sup>' = Y<sup>-</sup> XOR R'<sub>B</sub>
- NC1 envoie X<sup>+</sup>' et Y<sup>+</sup>' à PS
- NC2 envoie X<sup>-</sup>' et Y<sup>-</sup>' à PS
- PS calcule : C<sup>+</sup> = X<sup>+</sup>' AND Y<sup>+</sup>' et C<sup>-</sup> = X<sup>-</sup>' AND Y<sup>-</sup>' ainsi qu'un nombre aléatoire R<sub>PS</sub>



# Préservation de la vie privée (cont.)

---

- PS envoie  $A'_{PS} = C^+ XOR R_{PS}$  à NC<sub>1</sub> et  $B'_{PS} = C^- XOR R_{PS}$  à NC<sub>2</sub>
- NC<sub>1</sub> calcule  $A^R = A'_{PS} XOR (X^+ AND R'_B) XOR (Y^+ AND R_B) XOR (X^+ AND Y^+) XOR (R_B AND R'_A)$
- NC<sub>2</sub> calcule  $B^R = B'_{PS} XOR (X^- AND R'_A) XOR (Y^- AND R_A) XOR (X^- AND Y^-) XOR (R_A AND R'_B)$
- Résultat final (AR | BR) tels que  $A^R XOR B^R = (X^+ XOR X^-) AND (Y^+ XOR Y^-)$



# Préservation de la vie privée (cont.)

---

- Résultat final
- $A^R \text{ XOR } B^R = (X^+ \text{ AND } R'_B) \text{ XOR } (Y^+ \text{ AND } R_B) \text{ XOR } (X^+ \text{ AND } Y^+) \text{ XOR } (R_B \text{ AND } R'_A) \text{ XOR } (X^- \text{ AND } R'_A) \text{ XOR } (Y^- \text{ AND } R_A) \text{ XOR } (X^- \text{ AND } Y^-) \text{ XOR } (R_A \text{ AND } R'_B) \text{ XOR } (X^- \text{ AND } R'_B) \text{ XOR } (Y^+ \text{ AND } R_B) \text{ XOR } (X^- \text{ AND } R'_A) \text{ XOR } (Y^- \text{ AND } R_A) \text{ XOR } (X^- \text{ AND } Y^-) \text{ XOR } (X^- \text{ AND } Y^-) \text{ XOR } (R_A \text{ AND } R'_B) \text{ XOR } (R_B \text{ AND } R'_A) \text{ XOR } R_{PS} \text{ XOR } R_{PS}$



# Préservation de la vie privée (cont.)

---

- Propriété du XOR :
- $R \text{ } XOR \text{ } R = 0$



# Préservation de la vie privée (cont.)

---

- $A^R \text{ XOR } B^R = (X^+ \text{ AND } R'_B) \cancel{\text{XOR}} (Y^+ \text{ AND } R_B) \cancel{\text{XOR}} (X^+ \text{ AND } Y^+) \text{ XOR } (R_B \text{ AND } R'_A) \text{ XOR } (X^- \text{ AND } R'_A) \text{ XOR } (Y^- \text{ AND } R_A) \text{ XOR } (X^- \text{ AND } Y^-) \cancel{\text{XOR}} (R_A \text{ AND } R'_B) \text{ XOR } (X^- \text{ AND } R'_B) \cancel{\text{XOR}} (Y^+ \text{ AND } R_B) \text{ XOR } (X^- \text{ AND } R'_A) \text{ XOR } (Y^- \text{ AND } R_A) \text{ XOR } (X^- \text{ AND } Y^-) \cancel{\text{XOR}} (R_A \text{ AND } R'_B) \text{ XOR } (R_B \text{ AND } R'_A) \cancel{\text{XOR}} R_{PS} \text{ XOR } \cancel{R_{PS}}$



# Préservation de la vie privée (cont.)

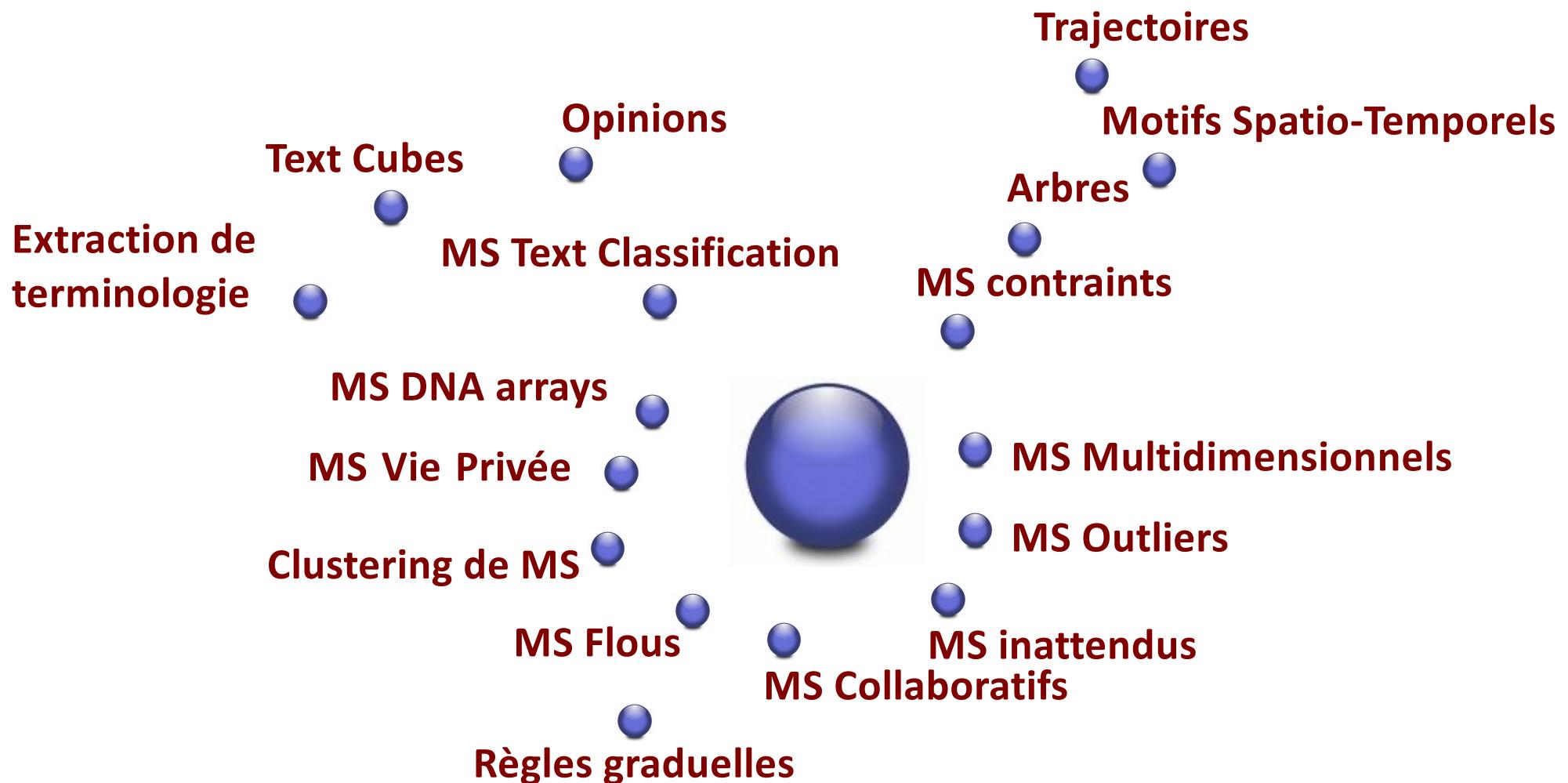
---

- Le résultat final :

$$A^R \text{ } XOR \text{ } B^R = (X^+ \text{ } XOR \text{ } X^-) \text{ AND } (Y^+ \text{ } XOR \text{ } Y^-)$$



# Autour des Motifs Séquentiels



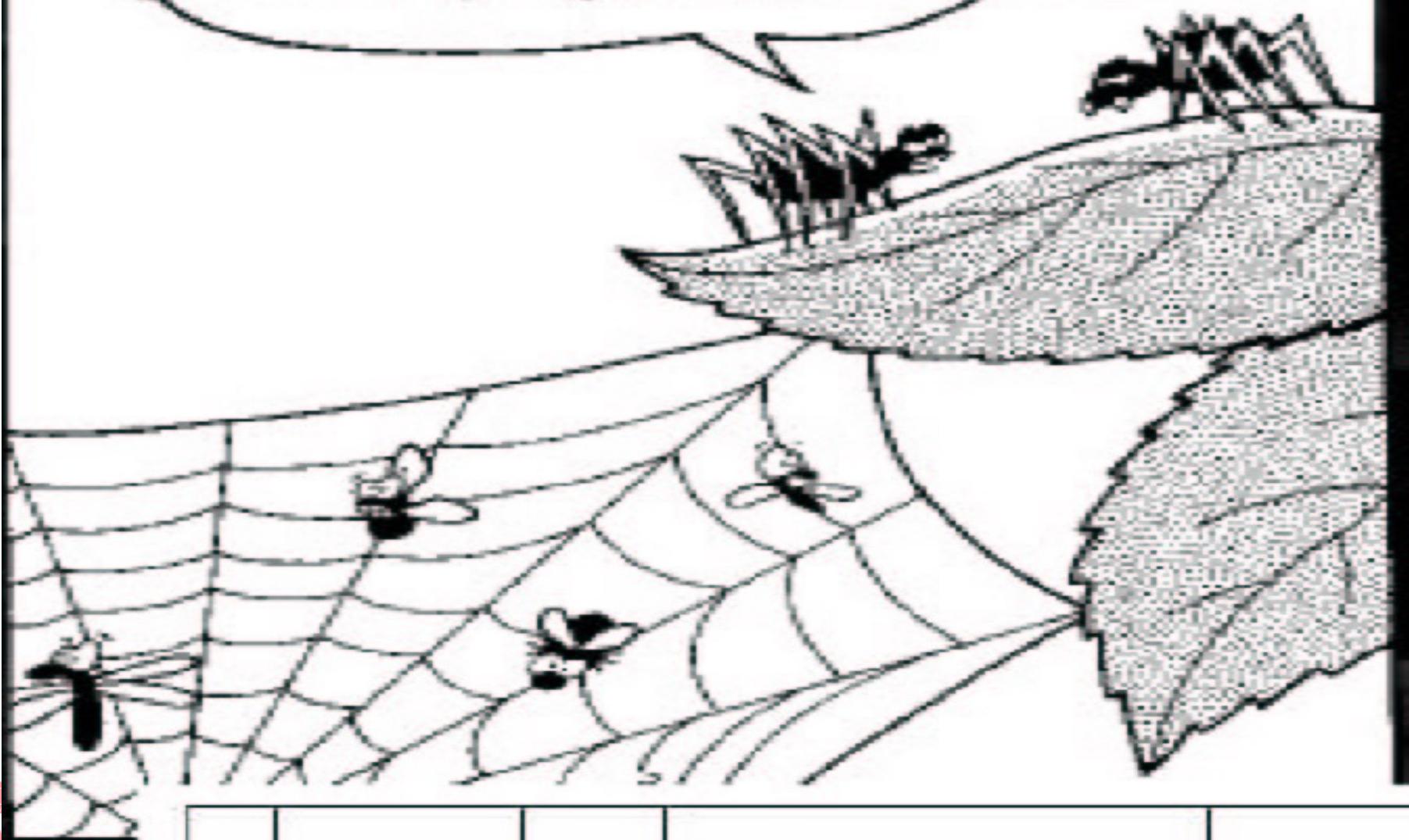
# Plan

---

- Contexte général
- Motifs séquentiels
- Extensions des motifs séquentiels
- Quelques applications des motifs
- Conclusions



EXCUSE ME A SEC... I WANT  
TO CHECK HOW MANY HITS I GOT  
ON MY WEBSITE...



# Web Usage Mining

---

- Analyse de l'usage des visiteurs sur un site Web
- Les pages contiennent l'information
- Les liens sont des « routes » (hyperliens)
- Comment les personnes naviguent-elles sur Internet ?
  - Web Usage Mining (Clickstream Analysis)
  - Information sur les chemins de navigation disponibles dans des fichiers logs.
- Principe :  
intégrer et « fouiller » ces données pour en produire de l'information et de la connaissance



# Web Usage Mining

---

- Pourquoi analyse l'usage des sites Web ?
- La connaissance sur la manière dont les visiteurs utilisent un site Web permet de :
  - Fournir une aide pour réorganiser site
  - Aider le concepteur à positionner l'information importante que les visiteurs recherchent.
  - Précharger et cacher les pages
  - Fournir des sites adaptatifs (personnalisation)
  - Eviter le « zapping »
- Utile dans le cas du e-commerce



# Exemple d'utilisation

---

Statistiques générales	Performance du site	Retenir les clients
Analyse du contenu	Groupement des clients	Campagne adaptée
Point d'entrée	Ciblages des clients	Campagne ciblée
Parcours	Comportement des clients	Modification dynamique



# Web Usage Mining

---

- De nombreux outils disponibles
- Statistiques générales :
  - Nombre de hits
  - Quelle est la page la plus populaire du site ?
  - Qui a visité le site ?
  - Qu'est ce qui a été téléchargé ?
  - Quels sont les mots clés utilisés pour venir sur le site ?



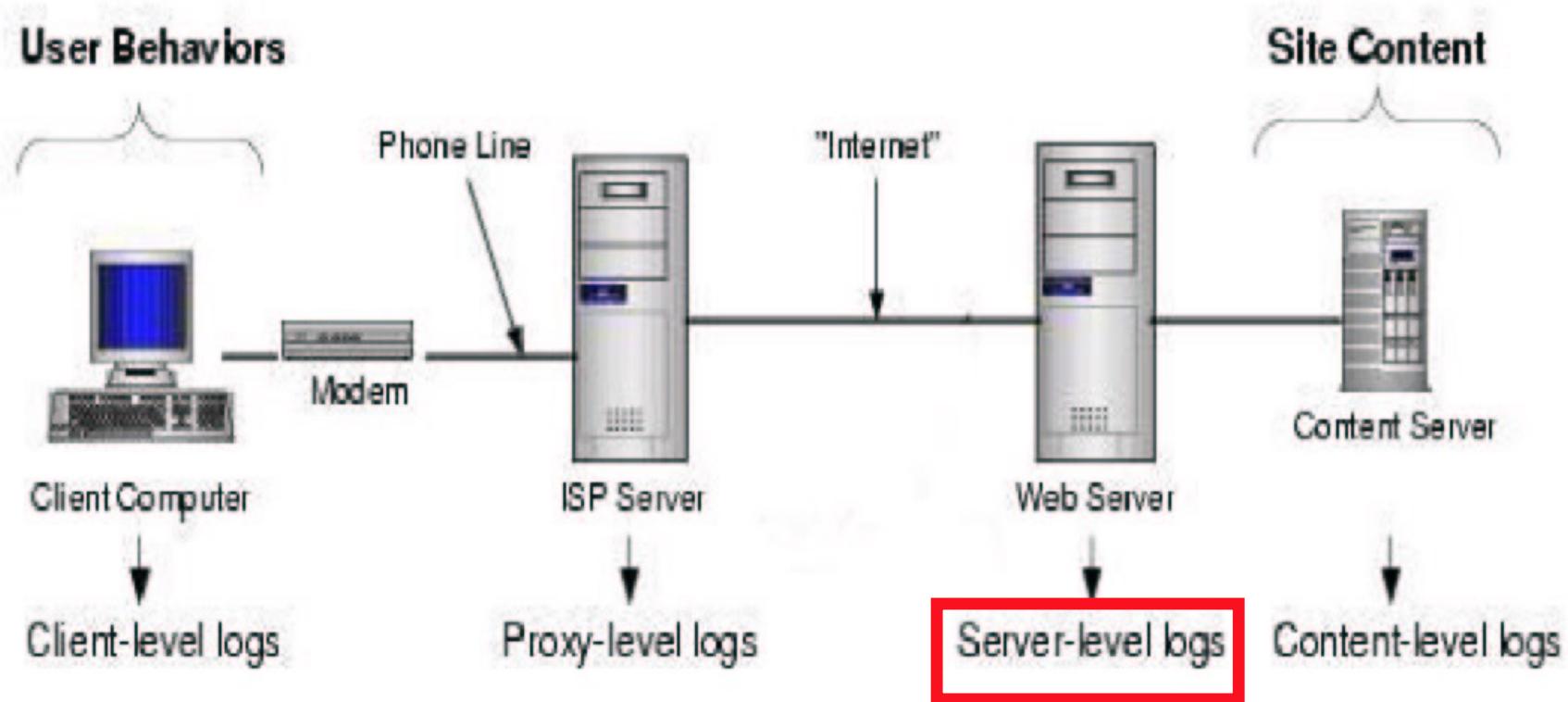
# Web Usage Mining



« 75% des **parisiens** qui achètent une **raquette de tennis** achètent **trois mois** après des **chaussures** »  
Modification dynamique

# Log or Logs?

Information sur les chemins de navigation dans les fichiers logs



# Web logs

---

IP or domain name      User Id      Date and Time      Request  
123.456.78.9 -- [24/Oct/1999:19:13:44 -0400] "GET /Images/tagline.gif HTTP/1.0"

Status      File Size      Referrer URL      Browser      Cookies  
200 1449 <http://www.teced.com/> "Mozilla/4.51 [en] (Win98;I)"



# Web logs

IP Address	Time	Method/URL/Protocol	Sta	Size	Referre d	Agent
123.456.78.9	[25/Apr/1998:03:04:41 –0500	GET A.html HTTP/1.0	200	3290	-	Mozilla/3.01 (Win95, I)
123.456.78.9	[25/Apr/1998:03:05:34 –0500	GET B.html HTTP/1.0	200	2050	A.html	Mozilla/3.01 (Win95, I)
123.456.78.9	[25/Apr/1998:03:05:39 –0500	GET L.html HTTP/1.0	200	4130	-	Mozilla/3.01 (Win95, I)
123.456.78.9	[25/Apr/1998:03:06:02 –0500	GET F.html HTTP/1.0	200	5096	B.html	Mozilla/3.01 (Win95, I)
123.456.78.9	[25/Apr/1998:03:06:58 –0500	GET A.html HTTP/1.0	200	3290	-	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
123.456.78.9	[25/Apr/1998:03:07:42 –0500	GET B.html HTTP/1.0	200	2050	A.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
123.456.78.9	[25/Apr/1998:03:07:55 –0500	GET R.html HTTP/1.0	200	8140	L.html	Mozilla/3.01 (Win95, I)
123.456.78.9	[25/Apr/1998:03:09:50 –0500	GET C.html HTTP/1.0	200	1820	A.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
123.456.78.9	[25/Apr/1998:03:10:02 –0500	GET O.html HTTP/1.0	200	2270	F.html	Mozilla/3.01 (Win95, I)
123.456.78.9	[25/Apr/1998:03:10:45 –0500	GET J.html HTTP/1.0	200	9430	C.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
123.456.78.9	[25/Apr/1998:03:12:23 –0500	GET G.html HTTP/1.0	200	7220	B.html	Mozilla/3.01 (Win95, I)
123.456.78.9	[25/Apr/1998:05:05:22 –0500	GET A.html HTTP/1.0	200	3290	-	Mozilla/3.01 (Win95, I)



Clients

Dates

Items

# Web logs

IP Address	Time	Method/URL/Protocol	Sta	Size	Referer	Agent
123.456.78.9	[25/Apr/1998:03:04:41 –0500]	GET A.html HTTP/1.0	200	3290	-	Mozilla/3.01 (Win95, I)
123.456.78.9	[25/Apr/1998:03:05:34 –0500]	GET B.html HTTP/1.0	200	2050	A.html	Mozilla/3.01 (Win95, I)
123.456.78.9	[25/Apr/1998:03:05:39 –0500]	GET L.html HTTP/1.0	200	4130	-	Mozilla/3.01 (Win95, I)
123.456.78.9	[25/Apr/1998:03:06:02 –0500]	GET F.html HTTP/1.0	200	5096	B.html	Mozilla/3.01 (Win95, I)
123.456.78.9	[25/Apr/1998:03:06:58 –0500]	GET A.html HTTP/1.0	200	3290	-	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
123.456.78.9	[25/Apr/1998:03:07:42 –0500]	GET B.html HTTP/1.0	200	2050	A.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
123.456.78.9	[25/Apr/1998:03:07:55 –0500]	GET R.html HTTP/1.0	200	8140	L.html	Mozilla/3.01 (Win95, I)
123.456.78.9	[25/Apr/1998:03:09:50 –0500]	GET C.html HTTP/1.0	200	1820	A.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
123.456.78.9	[25/Apr/1998:03:10:02 –0500]	GET O.html HTTP/1.0	200	2270	F.html	Mozilla/3.01 (Win95, I)
123.456.78.9	[25/Apr/1998:03:10:45 –0500]	GET J.html HTTP/1.0	200	9430	C.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
123.456.78.9	[25/Apr/1998:03:12:23 –0500]	GET G.html HTTP/1.0	200	7220	B.html	Mozilla/3.01 (Win95, I)
123.456.78.9	[25/Apr/1998:05:05:22 –0500]	GET A.html HTTP/1.0	200	3290	-	Mozilla/3.01 (Win95, I)

# Web logs

IP Address	Time	Method/URL/Protocol	Sta	Size	Referre d	Agent
123.456.78.9	[25/Apr/1998:03:04:41 –0500]	GET A.html HTTP/1.0	200	3290	-	Mozilla/3.01 (Win95, I)
123.456.78.9	[25/Apr/1998:03:05:34 –0500]	GET B.html HTTP/1.0	200	2050	A.html	Mozilla/3.01 (Win95, I)
123.456.78.9	[25/Apr/1998:03:05:39 –0500]	GET L.html HTTP/1.0	200	4130	-	Mozilla/3.01 (Win95, I)
123.456.78.9	[25/Apr/1998:03:06:02 –0500]	GET F.html HTTP/1.0	200	5096	B.html	Mozilla/3.01 (Win95, I)
123.456.78.9	[25/Apr/1998:03:06:58 –0500]	GET A.html HTTP/1.0	200	3290	-	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
123.456.78.9	[25/Apr/1998:03:07:42 –0500]	GET B.html HTTP/1.0	200	2050	A.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
123.456.78.9	[25/Apr/1998:03:07:55 –0500]	GET R.html HTTP/1.0	200	8140	L.html	Mozilla/3.01 (Win95, I)
123.456.78.9	[25/Apr/1998:03:09:50 –0500]	GET C.html HTTP/1.0	200	1820	A.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
123.456.78.9	[25/Apr/1998:03:10:02 –0500]	GET O.html HTTP/1.0	200	2270	F.html	Mozilla/3.01 (Win95, I)
123.456.78.9	[25/Apr/1998:03:10:45 –0500]	GET J.html HTTP/1.0	200	9430	C.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
123.456.78.9	[25/Apr/1998:03:12:23 –0500]	GET G.html HTTP/1.0	200	7220	B.html	Mozilla/3.01 (Win95, I)
123.456.78.9	[25/Apr/1998:05:05:22 –0500]	GET A.html HTTP/1.0	200	3290	-	Mozilla/3.01 (Win95, I)

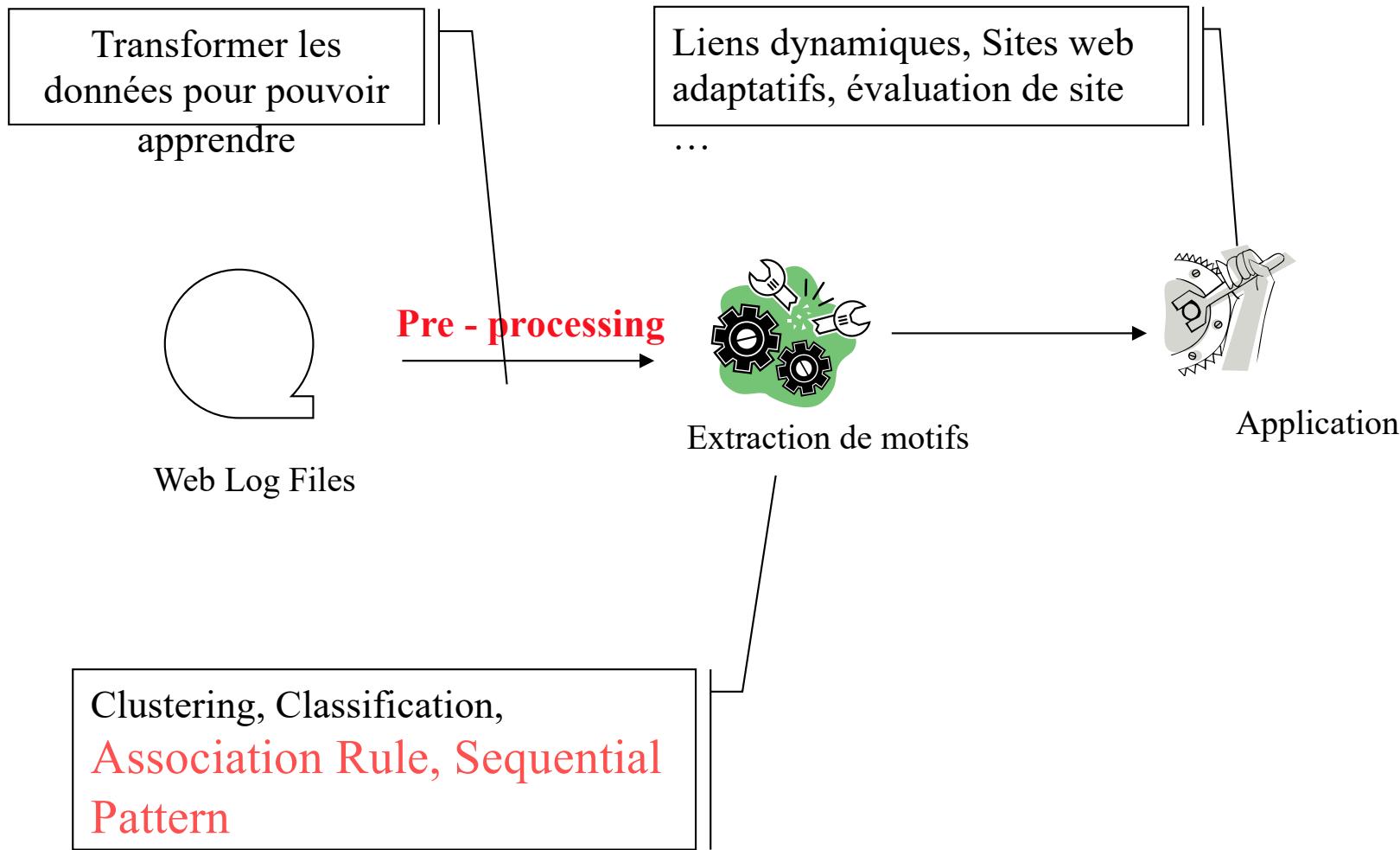


Clients

Dates

Items

# KDD pour WUM ?



# Pre-traitements

---

- **Data Filtering - Data Cleaning**
  - Status Code (1xx: Informational, 2xx: Success, 3xx: Redirection, 4xx: Client Error, 5xx: Server Error)
  - Requêtes automatiques (bots, performance monitoring systems)
  - Suppression des entrées concernant des requêtes pour des fichiers graphiques, des frames ...
  - Suppression des entrées générées par des spiders/crawlers (utilisés par les moteurs de recherche)



# Web Usage Mining

---

- Préparation des données (suffixe, éliminations des robots – agents de moteurs)
- Identification de l'utilisateur

Tout n'est pas dans le fichier Access Log

Utilisation d'heuristiques :

*Si une page est demandée et qu'elle n'est pas directement liée aux autres pages, il est probable qu'il existe différents utilisateurs sur la même machine*

*Utilisation des informations sur l'IP, le nom de la machine, le navigateur, des informations temporelles ...*



# Web Usage Mining

---

- Problèmes :
  - ID utilisateurs supprimées pour des raisons de sécurité
  - IP individuelles cachées par les proxys
  - Les caches des proxy et du côté clients
- Solutions actuelles :
  - Enregistrement de l'utilisateur – pratique ??
  - Cookies – difficile ??
  - « Cache busting » - augmente le trafic sur le réseau (inutile avec certains proxy)



# Web Usage Mining

---

- Sessions : Comment identifier/définir une transaction d'un visiteur ?
- « Time Oriented »
  - Durée totale d'une session :  $\leq 30$  minutes
  - Par temps passé sur une page :  $\leq 10$  minutes/page
- « Navigation Oriented »
  - Le « referrer » est la page précédente, ou le « referrer » n'est pas défini mais demandé dans les 10 secondes, ou le lien de la page précédente à la page courante dans le site web



# Web Usage Mining

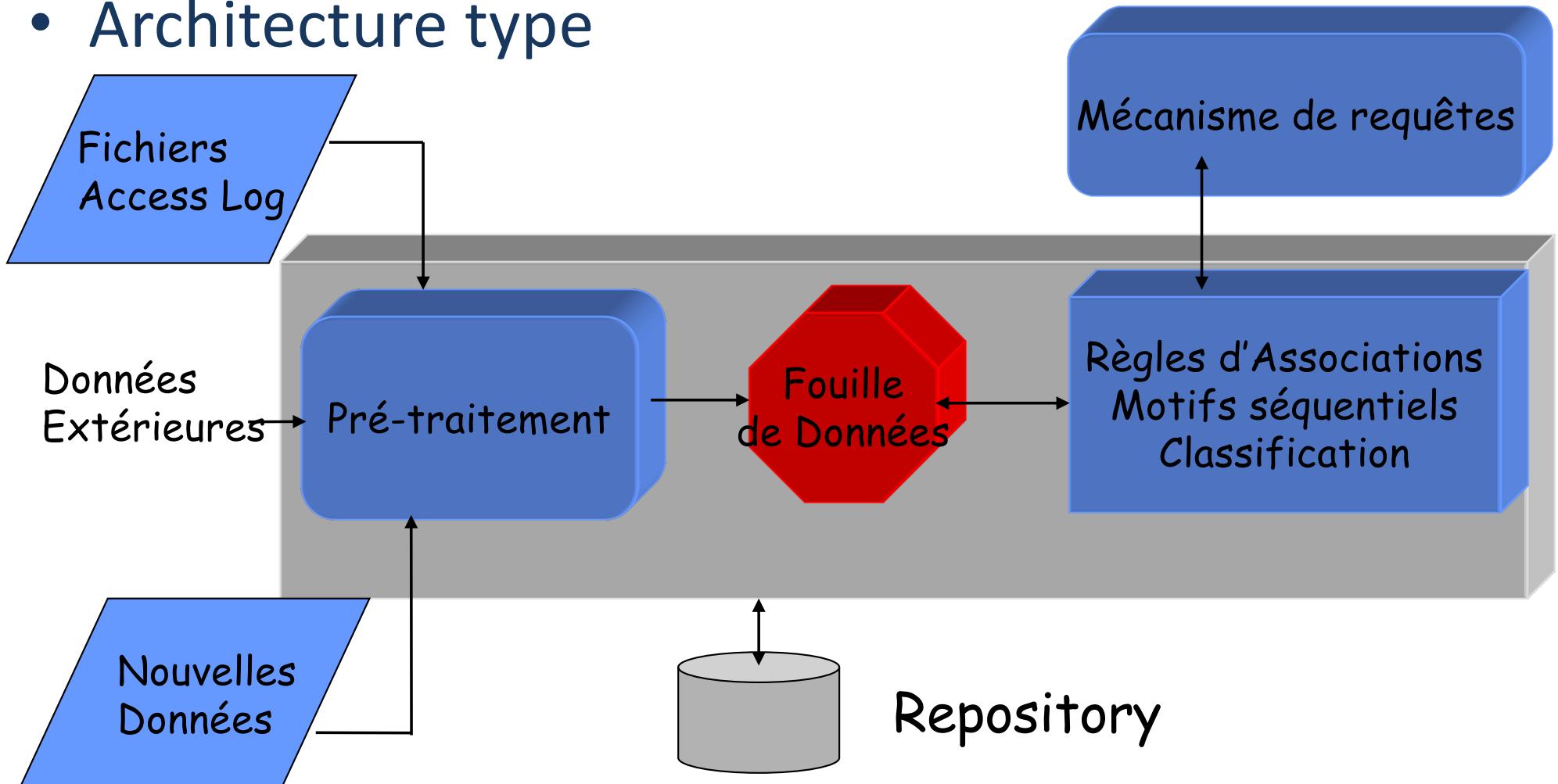
---

- Sources de données
  - Utilisation de fichiers logs
  - Mais aussi cookies, bases de données des clients,
  - ....



# Web Usage Mining

- Architecture type



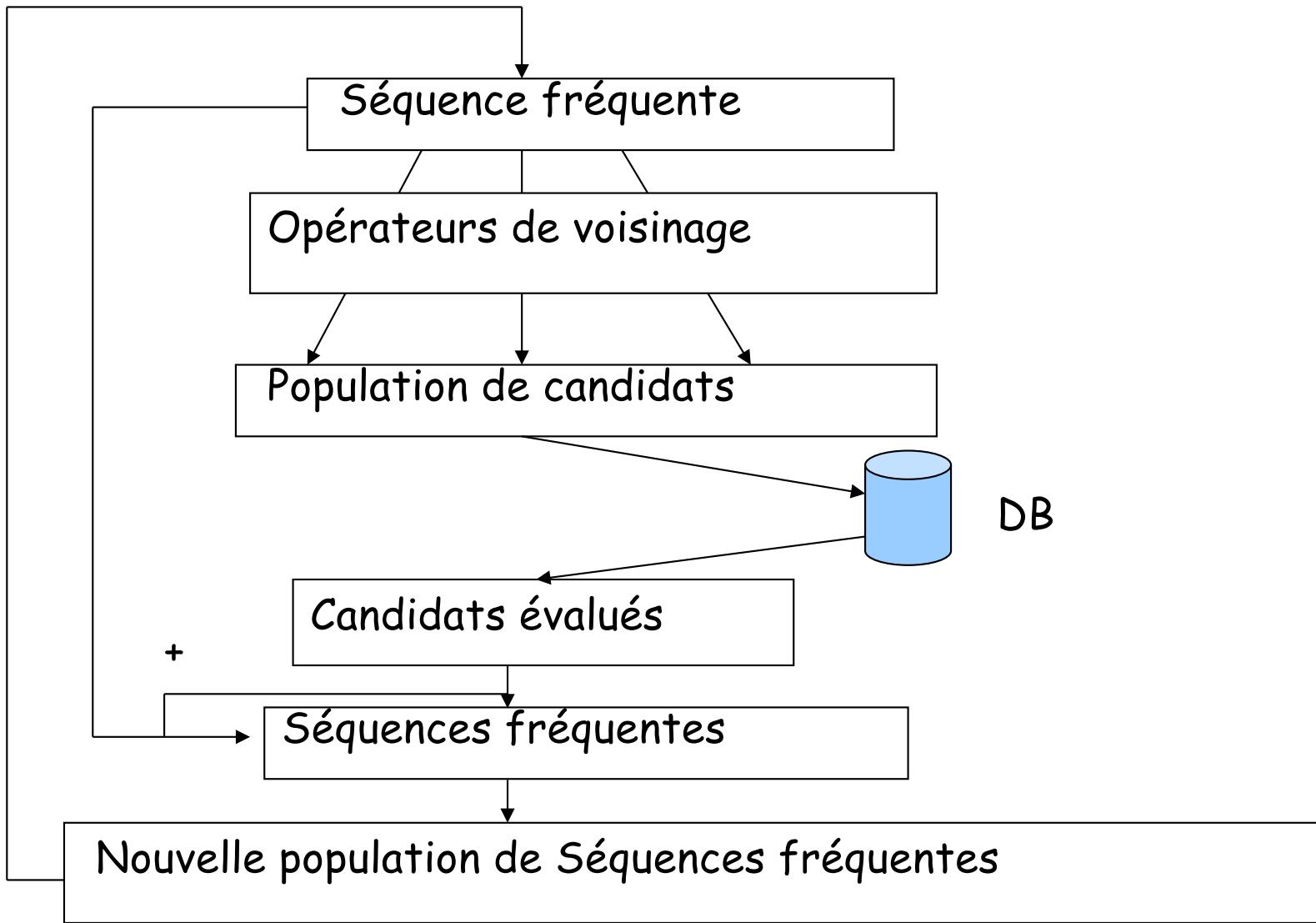
# Web Usage Mining

---

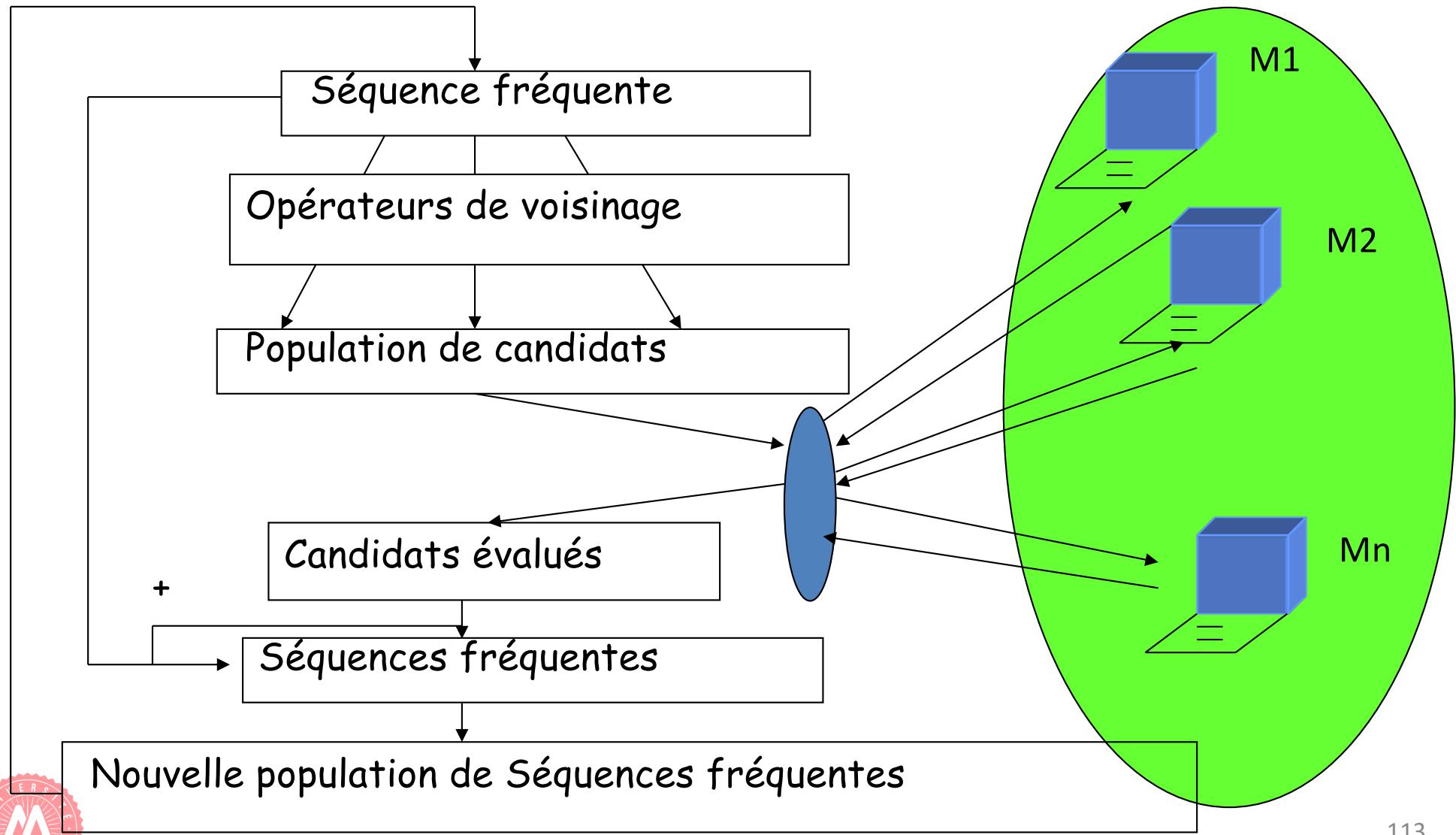
- Limites de l'approche : analyse en différée
  - Vers une approche « temps réel »
  - Pourquoi ne pas extraire les motifs séquentiels avec une méta-heuristique ?
- 
- Inspiration des algorithmes génétiques
  - Principe assez similaire



# Web Usage Mining temps réel



# Web Usage Mining temps réel



# Web Usage Mining temps réel

---

- Opérateurs de voisinage
- Ajout d 'item
  - <(a) (b) (d)> avec l 'item c
    - < (c) (a) (b) (d)> <(a) (c) (b) (d)>
    - <(a) (b) (c) (d)> <(a) (b) (d) (c)>
- Nouvel item fréquent : candidats de taille 2



# Web Usage Mining temps réel

---

- Croisement
  - <(a) (b) (g) (h)> et <(e) (f) (c) (d)>
  - <(a) (b) (c) (d)> <(e) (f) (g) (h)>
- Croisement amélioré
  - la coupure se fait après le plus long préfixe commun
- Dernier croisement
  - la seconde séquence est choisie pour son préfixe
- Extension de séquence
  - mélange entre ajout d ’item et nouvel item fréquent



# Avantages du WUM temps réel

---

- Disponibilité immédiate
  - en tant que pré-calcul ou comme résultat direct
- Un nouveau type de séquence
  - « la période du 02 au 07 janvier abrite le fréquent  $\langle(x)(y)(z)$ , avec un support de 70% »
- Résultats dédiés aux utilisateurs connectés
- Puissance de calcul inépuisable
- Data Mining interactif
- Extraction de fréquents de grande taille



# Du WUM au P2P Mining

---

- Les systèmes d'échanges pair-à-pair (P2P) :
  - Offrir à une large communauté des applications pour partager des fichiers, partager des calculs, dialoguer ou communiquer en temps réel, ...
  - **bonne infrastructure** pour les opérations sur de grandes masses de données ou avec de **très nombreux calculs**,... la fouille de données
- Un constat : la distribution “*Mandriva Linux 2005*” est souvent téléchargée avec *CD1.iso*, puis *CD2.iso* et finalement *CD3.iso*.



# P2P Mining

---

- Connaissance pour
  - Aider à rechercher des documents liés
  - Éviter des opérations de broadcast
  - Web Usage Mining vs. P2P Usage Mining
- Des motifs ... oui mais .... attention
  - Systèmes très dynamiques
    - Les noeuds agissent indépendamment les uns des autres
  - Comportement très dynamique car cible ouverte
  - Quand un noeud disparaît, les séquences de ce noeud disparaissent également de la base distribuée
    - Quid de la connaissance extraite ?



# P2P Mining : problématique

---

- Un item  $i \Rightarrow [op; i]$  ou  $op=\{d \text{ (download)}, r \text{ (request)}\}$
- Une séquence maintenant  $S=< ([d; 3]) ([d; 4] [r; 5]) ([d; 8]) >$

*Un utilisateur a téléchargé 3, puis il a téléchargé 4 et interrogé sur 5 en même temps et enfin il a téléchargé 8*



# P2P Mining : problématique (cont.)

---

- Soit  $D_t$  la base au temps  $t$ , nous avons :  
Pour un nœud  $u$ , nous notons sa partition (ses données)  $D_t^u$ .  
Nous avons donc, pour tous les nœuds connectés à un instant  $t$

$$D_t = \cup D_t^u$$

- Problème : trouver les séquences dont le nombre d'occurrences dans  $D_t$  est supérieur ou égal à  $minSupp$ .



# P2P Mining - Hypothèse

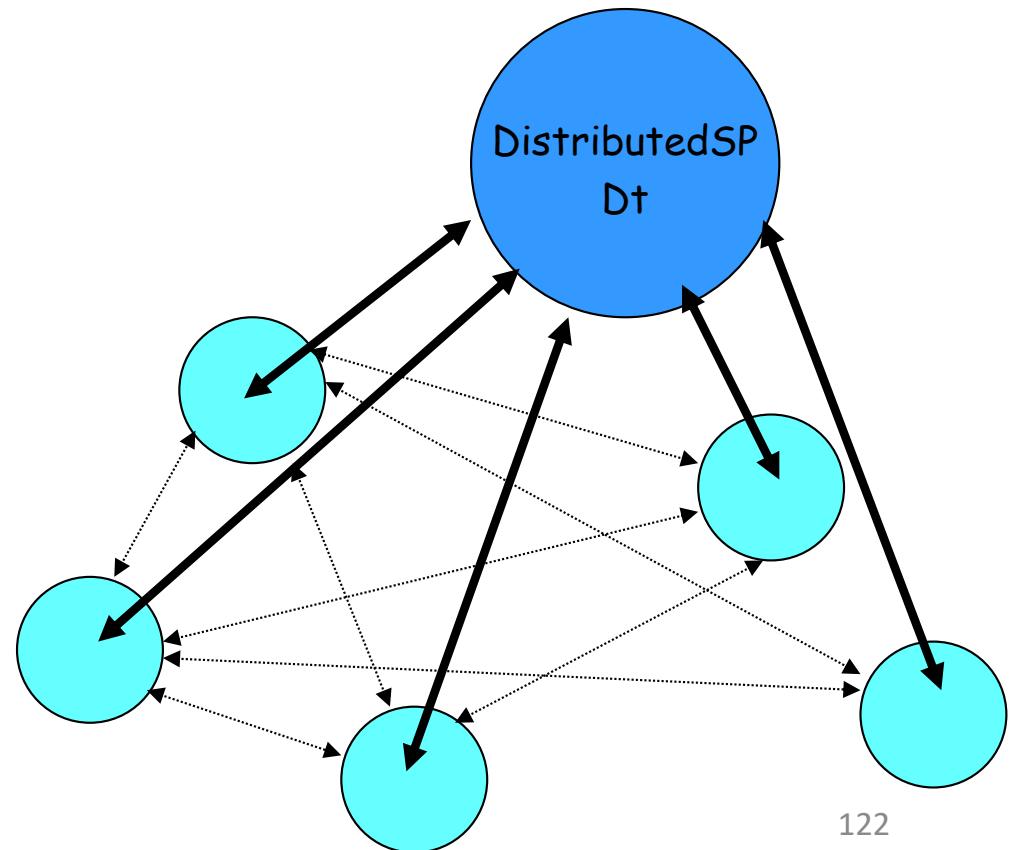
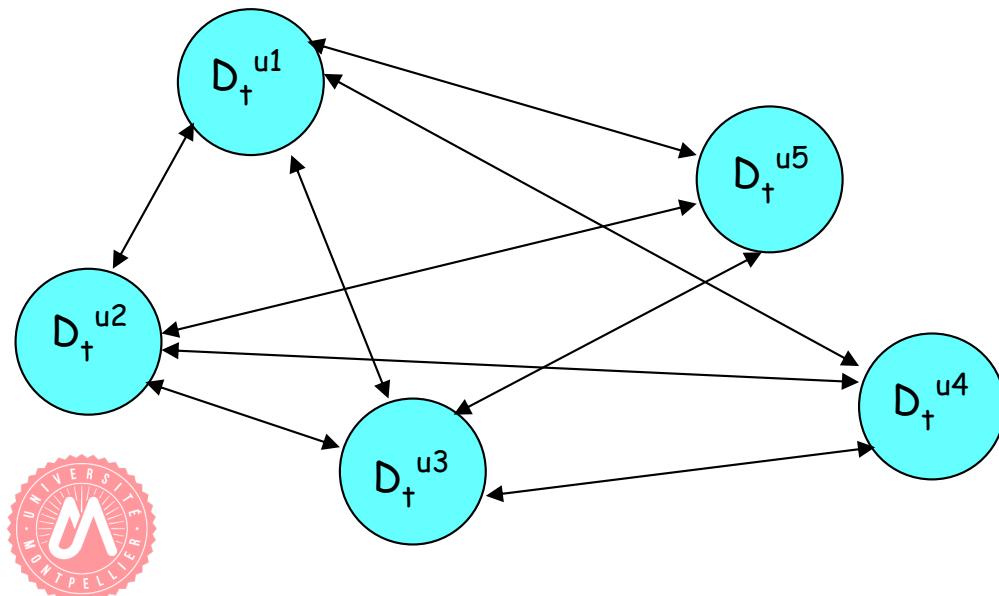
---

- Hypothèses : Réseau de pairs non structuré
- Pair  $D_t^u$  est capable de recevoir des séquences candidates, d'évaluer leur support dans  $D_t^u$  et de retourner le résultat



# P2P Mining – Hypothèse (cont.)

- Un pair spécial ( “*DistributedSP* ”) qui est connecté à tous les nouveaux pairs qui arrivent sur le réseau



# Une nouvelle approche

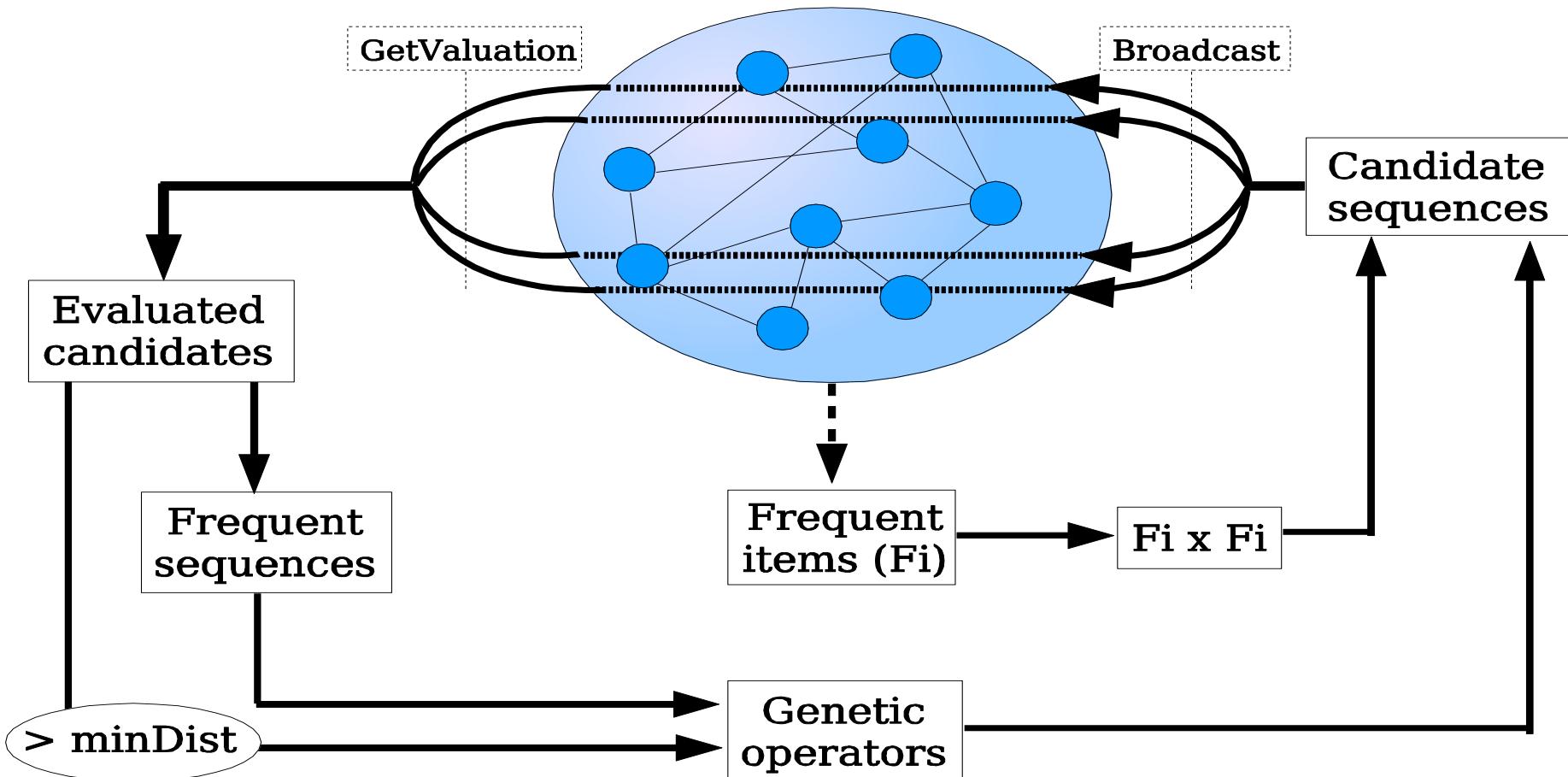
---

- Principe général : Générer (distribution de séquences candidates) Elaguer
  - 1) L'ensemble des items fréquents est extrait des pairs connectés.
  - 2) L'ensemble de tous les candidats de taille 2 est généré. Ces candidats sont évalués par les pairs connectés pour connaître ceux qui ont un nombre d'occurrences suffisant sur toute la base
  - 3) Les résultats sont récupérés par le pair *DistributedSP*
  - 4) L'heuristique, basée sur des opérateurs génétiques est alors appliquée et le nouvel ensemble de candidats est envoyé aux pairs connectés pour évaluation.

Processus répété tant qu'il existe des nœuds connectés.



# Une nouvelle approche



# Text Mining

---

- Qu'est ce que le Text Mining
  - L'extraction de connaissance à partir de données textuelles (découvertes de tendances, classification/organisation, ....)
- Les BD textuelles sont omniprésentes
  - Bases de données de bibliothèques, bases de données de documents, mails, WWW, ...
  - Les données textuelles sont structurées ou semi-structurées
- La quantité de données textuelles augmente très rapidement « Le texte est facile à produire.



# Text Mining vs. Recherche d'information

---

- Recherche d'Information (*Information Retrieval*)  
Domaine développé en parallèle des bases de données  
L'information est organisée dans (un grand nombre de) documents  
Pb : localiser les documents pertinents en se basant sur l'entrée de l'utilisateur (mots clés ou documents exemples)
- Quid de la connaissance supplémentaire ?
- Requête vs Data Mining



# Text Mining - Classification

---

- Classification automatique
  - Classification automatique d'un grand nombre de documents (pages Web, mails, fichiers textuels) basée sur un échantillon de documents pré-classifié
- Mise en oeuvre
  - *Echantillon* : des experts génèrent l'échantillon
  - *Classification* : l'ordinateur découvre les règles de classification
  - *Application* : les règles découvertes peuvent être utilisées pour classer des nouveaux documents et les affecter à la bonne classe



# Text Mining - Classification

---

- Quelques problèmes
  - Synonymie : un mot T peut ne pas apparaître dans un document mais si le document est très lié à T (data mining / software product)
  - Polysémie : le même mot peut avoir plusieurs sens (mining)
  - Représentation des documents (vecteurs de termes, choix des termes représentatifs, calcul de la distance entre un vecteur représentant le groupe de documents et celui du nouveau document, ...)
  - Evolution des classes dans le temps



# Text Mining - Corrélations

---

- Analyse d'associations basée sur des mots clés
  - Rechercher des associations/correlations parmi des mots clés ou des phrases
- Mise en œuvre
  - *Pré-traitement des données* : parser, supprimer les mots inutiles (le, la, ...) => prise en compte d'une analyse morpho-syntaxique (e.g. lemmatiseur)
  - Un document est représenté par : (document\_id, {ensemble de mots clés})
  - Appliquer des algorithmes de recherche de règles d'association



# Text Mining - Corrélations

---

- Quelques problèmes
  - Ceux du traitement de la langue naturelle
  - Les mots inutiles (ordinateur ? Utile ?) – Réduction de l'espace de recherche
  - Les associations de mots, phrase, paragraphe, ...



# Text Mining – Analyse de tendances

---

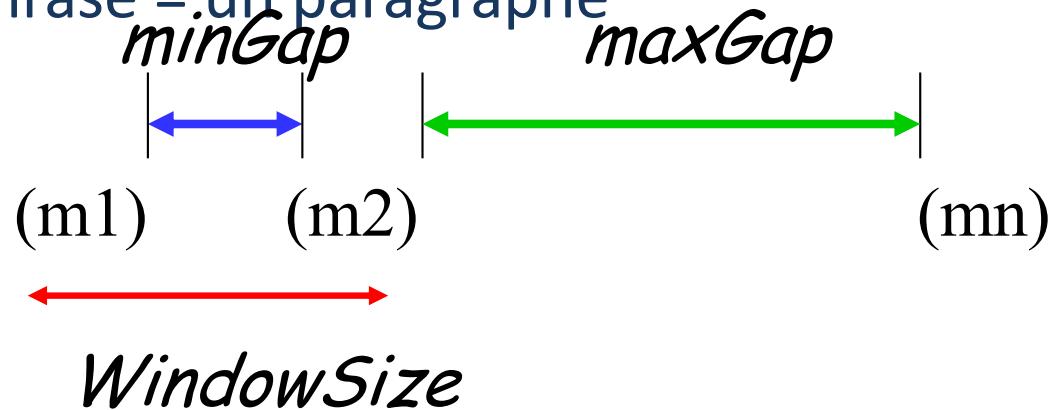
- Rechercher des tendances dans les documents
- Mise en œuvre
  - Pré-traitement : attention l'ordre est important
  - Document représenté par : (document\_id, <phrases simplifiées : ensemble de mots ordonnés>)
  - Appliquer des algorithmes de motifs séquentiels
  - Générer l'historique des phrases
  - Recherche les phrases qui correspondent à des tendances



# Text Mining – Analyse de tendances

---

- Principes
  - Un mot : (m)
  - Une phrase : <(m1) (m2) (m3) ... (mn)>
  - Paramètres : WindowSize, MaxGap, MinGap) Une phrase = une phrase
    - Une phrase = un ensemble de mots proches
    - Une phrase = un paragraphe



# Text Mining – Analyse de tendances

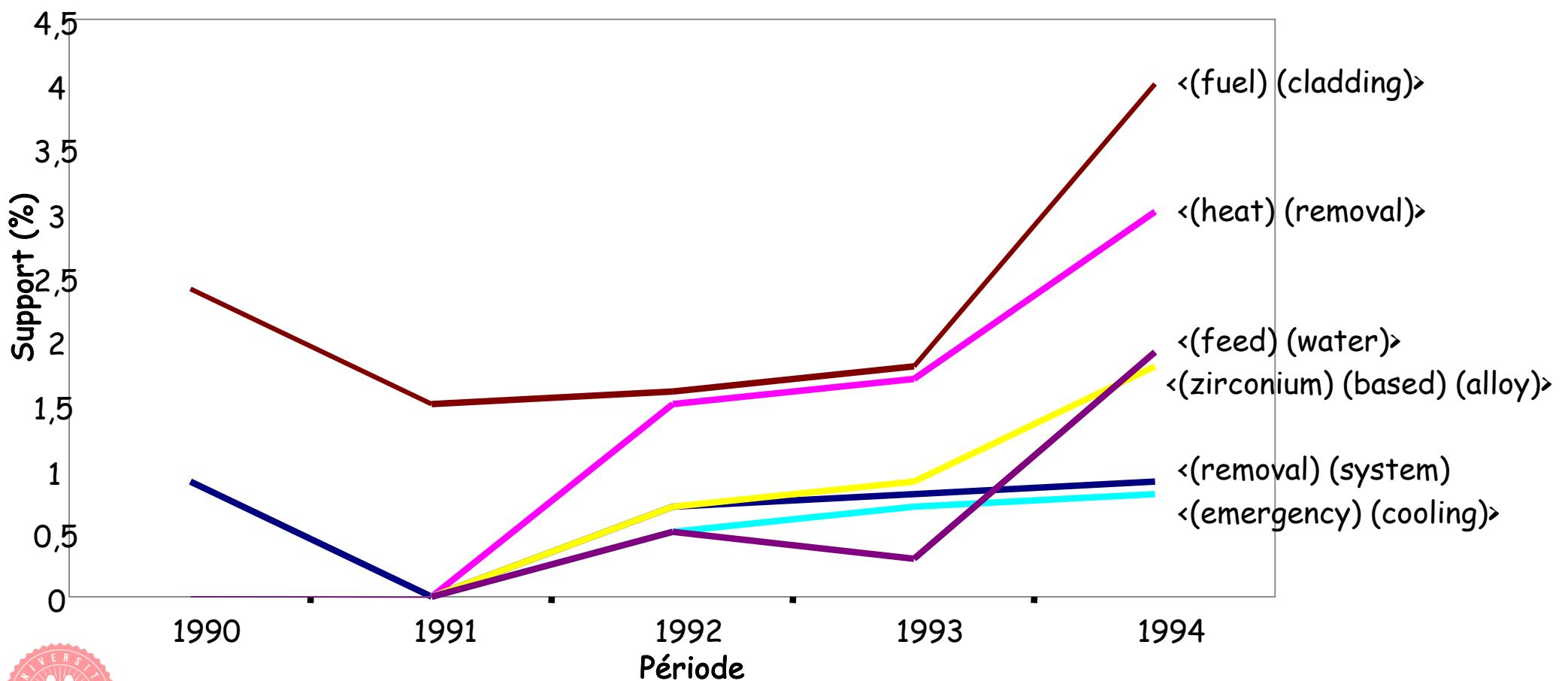
---

- Gérer l'historique des phrases
  - Partitionner les documents en fonction de leur estampille (ex : année pour les brevets, mois pour des documents sur le Web)
  - Pour chaque partition, génération des ensembles fréquents de phrases
  - Maintenir l'historique des supports pour chaque phrase
  - Interroger l'historique des phrases pour connaître les tendances (tendance récente à monter, transition récente, résurgence d'usage, ....)



# Text Mining – Analyse de tendances

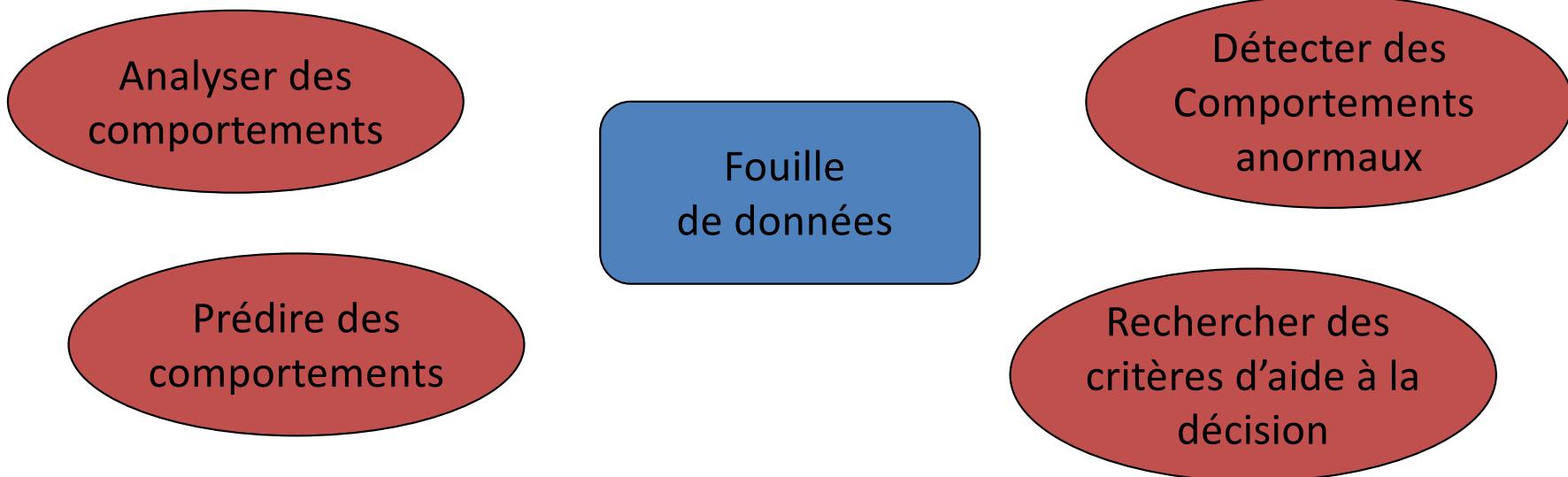
Brevets classés dans la catégorie « Induced Nuclear Reactions: Processes, Systems and Elements »



# Fouille de données de santé

---

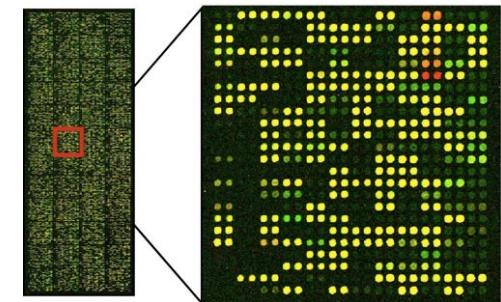
- **Données particulières:** hétérogènes, souvent imprécises, subjectives, non déterministes, bruitées, avec des valeurs manquantes et des erreurs



# Puces à ADN

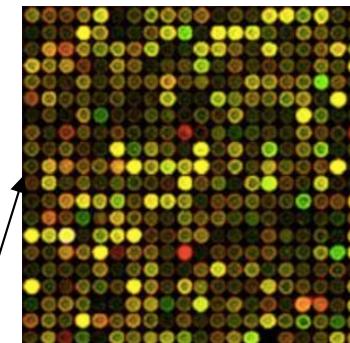
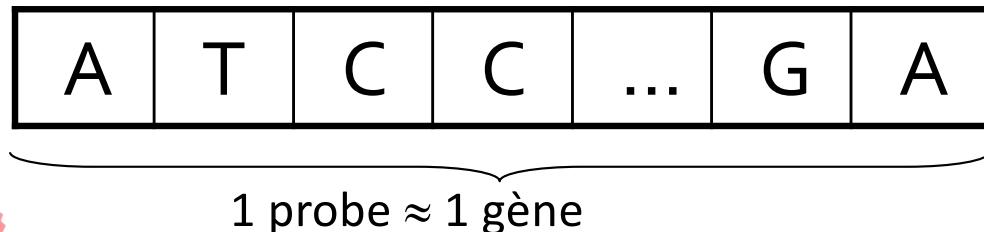
---

- **Incontournables pour comprendre les maladies génétiques complexes** : perturbation des processus naturels de croissance, de division et de mort des cellules
- **Utilisées par les biologistes** pour acquérir de grandes quantités de données sur l'expression des gènes et identifier les lois suivies par ces expressions en fonction des maladies et des traitements :
  - gènes impliqués dans la maladie ?
  - gènes dont les expressions sont corrélées ?
  - gènes qui inhibent ou activent une fonction ?
  - ....
- **Difficultés pour extraire automatiquement** des connaissances liés aux gros volumes de données



# Puces à ADN

- **Le principe :** propriété de l'ADN dénaturé de reformer spontanément sa double hélice lorsqu'il est porté face à un brin complémentaire (réaction d'hybridation).
- $A \equiv T$
- $T \equiv A$
- $G \equiv C$
- $C \equiv G$
- **Concrètement...** un ensemble de molécules d'ADN fixées en rangées ordonnées sur une petite surface



**Expression (couleur) ≈**  
mesure de la quantité  
d'ADN dénaturé qui se  
reforme

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1		B	3	03406	03406	03406	03406	03406	03406	03406	03406	03406	03406	03406	03406	03406	03406	03406	03406	03406	03406	03406	03406	03406
2	21152	319680	-0.21	0.93	-0.38	-0.2	1.05	-0.32	-1.35	-0.31	-0.34	0.03	-1.06	0.11	1.33	-0.35	-0.36	0.54	-1.19					
3	307830483	-0.57	11	-0.71	-1.20	1.45	-1.37	-0.88	-0.38	-0.89	0.03	-0.51	0.03	1.31	0.44	-0.91	0.52	0.09						
4	429304114	0.22	0.2	-0.27	-0.73	-0.45	-0.88	-0.83	-0.38	-0.5	-0.30	0.53	1.06	-1.41	-0.52	-1.01	-1.01	-1.47						
5	752583119602	-0.39	0.62	-1.2	0.2	1.02	-0.3	-0.24	0.76	0.38	0.03	0.76	1.24	1.05	-0.09	0.66	0.39	-0.01						
6	149883119608	0.12	-0.48	-0.58	0.68	-0.23	-0.37	1.01	0.85	-1.96	1.05	1.30	1.54	2.33	1.81	0.71	2.34	0.61						
7	3298309876	0.05	0.72	-2.71	0.28	-1.73	-0.88	-2.38	0.38	1.47	-0.8	-0.71	-1.89	1.44	-0.35	-0.01	-0.16	-2.1						
8	7757149630	-0.11	1.04	-1.02	0.48	0.74	-0.24	-0.85	-0.36	0.85	-0.34	0.45	0.71	-0.31	-0.14	-0.52	1.33	-0.17						
9	4200305804	-0.52	0.93	-1.02	-0.22	-0.23	-0.78	-0.55	-1.06	-0.21	-0.43	1.01	0.89	0.33	-0.37	-0.15	1.01	-0.95						
10	132053119888	-1.22	0.63	-1.22	0.56	1.23	-0.7	-0.8	0.85	0.26	-0.37	-0.03	1.06	-0.01	-0.16	-0.35	1.69	-0.59						
11	212643119886	-0.47	0.19	0.25	1.76	0.16	-0.28	0.24	0.21	0.03	0.65	1.11	0.71	0.01	0.01	0.47	1.11	0.54						
12	3298309803	0.48	0.17	-0.94	-0.77	-0.84	0.13	1.81	-0.38	0.38	0.38	1.81	0.21	-0.39	0.11	-0.52	1.10	0.35						
13	74773119868	0.28	0.38	-2.18	-0.24	-1.32	-0.38	-4.75	-0.81	0.71	0.04	0.31	-1.41	1.81	1.81	0.31	-0.38	-1.45						
14	34883119883	-0.17	0.77	-0.94	-1.23	0.38	-0.34	-0.28	-0.38	-0.29	0.34	0.76	1.06	-1.41	-1.96	-0.5	1.25	-0.45						
15	1215304817	-0.28	0.14	-1.14	0.13	0.3	-0.23	0.32	0.52	-0.83	0.48	-0.06	-0.46	-0.25	-0.04	-0.35	-0.23	0.26						
16	142853119861	0.75	-0.48	-0.08	-0.43	-0.24	0.18	-0.84	-0.43	-0.38	0.31	0.31	-0.25	-0.21	0.21	-0.29	-0.08	-0.44						
17	3040311984	-0.09	0.73	-1.81	0.02	1.0	0.48	-1.32	0.54	1.35	0.38	1.12	-0.03	1.75	-1.35	-0.1	-0.61	-0.57						
18	1132309862	-0.72	0.38	-1.42	0.28	1.05	0.24	-0.45	-1.81	0.21	-0.31	0.03	0.46	0.33	-1.1	0.06	0.45	-1.11						
19	75471119798	-0.57	1.08	-2.37	0.5	-1.13	-0.57	-1.87	0.85	-1.21	0.44	1.4	-1.71	0.03	0.91	-2.01	-0.41	0.19						
20	457300804	-0.17	-0.47	-1.13	-0.34	0.32	-0.82	-0.88	0.08	0.03	0.43	0.48	1.41	-0.31	-0.01	-0.31	-0.05	-0.15						
21	317853119880	0.08	0.77	0.12	0.58	0.81	-0.38	0.31	0.51	-0.34	0.03	-0.85	-0.2	0.98	-0.77	0.53	0.35	1.13						
22	6458307388	0.42	1.25	-2.14	-1.48	-0.38	-1.4	0.3	1.8	1.01	1.23	-1.12	-0.21	-0.21	0.39	-0.28	1.85	0.1						
23	5451306413	1.19	-0.52	0.59	0.51	-1.13	-1.8	-0.52	1.6	0.86	-2.07	0.6	-0.21	1.5	-0.49	0.61	0.32	1.71						
24	3209304558	-0.04	-0.58	0.15	1.48	-0.4	0.03	0.31	0.75	-0.95	-0.36	1.96	1.21	-1.41	-0.58	1.49	-0.52							
25	1406300623	0.04	-0.58	-1.29	1.07	-0.55	-0.88	1.0	-0.51	0.31	1.01	0.61	-0.39	0.21	-0.15	0.46	0.95							
26	2005301588	0.14	0.59	-0.58	0.35	0.28	-0.49	0.03	-0.64	0.03	0.4	0.34	1.31	0.35	0.14	-0.19	0.38	0.51						
27	7159309868	0.09	0.01	0.48	0.34	-0.6	0.03	0.31	0.39	1.2	1.04	-1.11	-0.93	-0.18	0.89	0.76	-0.31	0.35						
28	3444304101	0.07	1.5	-1.10	0.01	0.04	1.81	0.15	0.64	0.5	0.14	1.62	1.04	1.14	1.14	1.49	1.49	1.49						

Gènes

Gènes

Puces

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1		1	0.098	0.098	0.098	0.098	0.098	0.098	0.098	0.098	0.098	0.098	0.098	0.098	0.098	0.098	0.098	0.098	0.098	0.098	0.098	0.098	0.098	
2	1152319803-021	0.99	-0.98	-0.2	1.05	-0.92	-0.25	0.71	-0.94	0.12	-0.99	0.18	0.32	-0.95	-0.95	0.54	-0.19							
3	307830483-057	1.1	-0.73	-1.23	1.48	-1.37	0.98	0.38	0.98	0.32	-0.57	0.03	0.51	0.44	-0.91	0.52	0.09							
4	628301134022	0.2	-0.27	-0.73	-0.46	-0.89	0.03	0.38	-0.5	-0.33	0.53	0.86	-0.41	-0.42	-1.01	-1.01	-1.47							
5	25258319803-029	0.92	-1.2	0.2	1.02	-0.2	0.24	0.76	0.38	0.35	0.78	0.74	0.05	-0.09	0.66	0.19	-0.01							
6	14985319408012	-0.48	-0.58	0.68	-0.33	-0.27	0.71	0.65	-1.96	1.05	1.33	0.54	2.33	1.61	0.71	0.74	0.61							
7	329830987005	0.72	-0.73	0.38	-1.73	-0.88	0.08	0.76	1.47	-0.5	-0.73	-1.09	1.44	-0.35	-0.01	-0.16	-0.3							
8	17571456303-011	1.04	-1.92	0.48	0.74	-0.24	0.85	0.38	0.38	-0.34	0.45	0.77	-0.31	-0.14	-0.53	1.37	-0.17							
9	4720305804-055	0.93	-1.02	-0.23	-0.23	-0.79	0.58	0.08	-0.21	-0.43	1.01	0.89	0.33	-1.37	-0.15	1.01	-0.95							
10	19225319808-120	0.93	-1.22	0.58	1.21	-0.7	0.18	0.75	0.26	-0.37	-0.83	1.06	-0.01	-1.16	-0.26	0.69	-0.59							
11	110643112888-045	0.72	-0.02	-1.24	-2.55	0.15	0.28	0.24	-0.2	0.41	0.58	0.58	-0.11	-0.31	0.01	0.42	0.94							
12	3298309830448	0.77	-0.46	-0.77	-0.34	0.18	0.71	0.38	0.38	0.38	1.81	0.71	-0.39	0.11	-0.69	1.0	0.25							
13	14373318489328	0.39	-0.18	-0.24	-1.32	-0.39	0.73	0.71	0.71	0.34	0.81	-0.41	0.81	0.81	0.81	0.81	-0.38	-1.45						
14	59583112002-121	1.17	-1.13	0.33	0.98	-0.74	0.47	0.78	-0.21	-0.8	0.33	1.89	1.03	-1.36	0.06	0.86	-0.59							
15	25653319803-154	1.7	-1.05	-0.81	1.25	-0.4	0.73	0.70	-0.41	1.07	-0.3	2.49	0.71	-1.61	-0.61	-2.72	0.3							
16	14388318883-028	0.77	-0.26	-1.23	0.35	-0.38	0.28	0.38	-0.29	0.34	0.78	1.05	-0.41	-1.96	-0.5	1.23	-0.46							
17	1315301817-028	0.14	-1.14	0.13	0.3	-0.23	0.73	0.72	-0.83	0.48	-0.86	-0.46	-0.25	-0.04	-0.35	-0.23	0.26							
18	14265318481075	-0.48	-0.28	-0.61	-0.24	0.18	0.74	0.70	-0.38	0.31	0.33	-0.25	-0.22	0.21	-0.29	-0.05	-0.44							
19	102403112840-009	0.79	-1.80	0.32	1.0	0.48	0.72	0.74	1.25	0.38	1.32	-0.03	1.75	-1.38	-0.1	-0.61	-0.57							
20	1132309840-072	0.39	-1.42	0.28	1.05	0.24	0.75	0.71	0.21	-0.31	0.03	0.46	0.33	-1.1	0.06	0.45	-1.11							
21	154733129798-037	1.08	-0.37	0.5	-1.11	-0.57	0.77	0.73	-1.21	0.44	1.4	-0.73	0.03	0.81	-2.01	-0.41	0.19							
22	457300804-017	-0.47	-1.13	-0.34	0.32	-0.82	0.65	0.08	0.02	0.43	0.48	1.41	-1.31	-0.01	-0.31	-0.05	-0.15							
23	11785316288008	0.77	0.12	0.58	0.81	-0.38	0.71	0.71	-0.34	0.38	-0.85	-0.2	0.98	-0.77	0.58	0.75	0.15	0.15						
24	6438307388042	1.25	-0.14	-1.48	-1.38	-0.38	0.7	0.73	1.3	1.07	1.23	-0.13	-0.21	0.39	-0.28	0.85	0.1							
25	5451306403119	-0.42	0.58	0.51	-1.11	-1.8	0.75	0.70	0.86	-2.0	0.6	-0.21	1.5	-0.49	0.61	0.39	0.71							
26	3208304598-034	-0.24	-0.68	0.11	1.04	-0.4	0.70	0.71	-0.75	-0.55	-0.36	1.96	1.21	-1.41	-0.98	1.49	-0.58							
27	1626309820-029	0.04	0.58	-1.26	1.07	-0.55	0.76	0.70	-0.51	0.31	1.0	-0.61	-0.79	0.21	-0.32	0.46	0.95							
28	1095301588014	0.59	-0.68	-0.35	0.28	-0.40	0.73	0.74	-0.07	0.4	0.34	1.31	0.95	0.14	-0.39	0.78	0.52							
29	113930988-009	0.1	0.48	-0.84	0.5	0.73	0.70	0.73	1.2	1.04	-1.11	-0.89	-0.18	0.89	-0.78	-0.12	0.35							
30	144430400-014	0.7	1.3	1.0	0.29	0.94	0.70	0.74	0.76	0.76	0.73	0.73	0.73	0.73	0.73	0.73	0.73	0.73						

Gènes

Puces

Intensité (expression) d'un gène mesuré par une puce

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X		
1		G3M06																								
2	71152310082	-0.21	0.99	-0.38	-0.2	1.05	-0.52	-1.25	-0.31	-0.84	0.12	-1.09	0.8	0.33	-0.95	-0.96	0.54	-0.19								
3	3078300633	-0.57	11	-0.71	-1.20	1.48	-1.37	-0.36	-0.38	0.05	0.10	-0.51	0.01	0.51	1.44	-0.91	0.52	0.09								
4	028301134022	0.2	-0.27	-0.73	-0.46	-0.89	-0.03	-0.38	-0.5	-0.33	0.52	0.86	-0.41	-0.52	-1.01	-1.01	-1.47									
5	72458319602	-0.28	0.62	-1.2	0.2	1.02	-0.2	-0.24	0.76	0.36	0.01	0.79	0.24	0.05	-2.09	0.66	0.39	-0.01								
6	14088319608012	-0.48	-0.58	0.08	-0.33	-0.37	1.01	0.85	-1.96	1.05	1.30	0.54	2.31	0.81	0.71	2.74	0.61									
7	329830087805	0.72	-0.73	0.38	-1.73	-0.88	-2.38	0.38	1.47	-0.6	-0.73	-0.89	1.44	-0.35	-0.01	-0.16	-0.3									
8	7757149530	-0.11	1.04	-1.32	0.48	0.74	-0.24	-0.85	-0.36	0.05	-0.34	0.45	0.77	-0.31	-0.14	-0.52	0.32	-0.17								
9	4220305024-0.62	0.89	-1.02	-0.23	-0.33	-0.79	-0.55	-1.06	-0.21	-0.21	1.01	0.89	0.33	-0.37	-0.15	1.01	-0.05									
10	138253100000	-1.22	0.63	-1.22	0.58	1.21	-0.7	-0.8	0.05	0.26	-0.27	-0.03	1.06	-0.01	-1.16	-0.26	0.69									
11	110843121988-0.62	0.72	-0.02	-1.24	-2.58	0.15	-0.38	-0.24	-0.2	0.41	0.58	0.58	-0.11	-0.21	0.21	0.09	1.94									
12	3298300803048	0.77	-0.44	-0.77	-0.34	0.18	-1.81	-0.38	0.38	0.36	1.61	0.21	-0.39	0.21	-0.66	1.0	0.25									
13	14273319889028	0.59	-0.19	-0.24	-1.32	-0.39	-1.75	-0.81	0.71	0.04	0.81	-0.41	0.84	0.81	0.31	-0.38	-1.43									
14	59583121021-1.21	1.17	-1.13	0.83	0.98	-0.74	-0.47	0.38	-0.23	-0.8	0.72	-0.89	1.03	-1.56	0.06	0.68	-0.59									
15	75653319883-1.54	17	-1.05	-0.81	1.25	-0.4	-0.53	-0.43	-0.41	1	-0.3	2.49	0.31	-1.61	-0.61	-2.72	0.3									
16	14388318888-0.28	0.77	-0.24	-1.23	0.35	-0.34	-0.28	-0.28	-0.09	0.34	0.78	1.06	-0.41	-1.96	-0.5	1.23	-0.45									
17	1315301817-0.28	0.14	-1.14	0.13	0.3	-0.21	-0.21	-0.21	-0.83	0.48	-0.86	-0.46	-0.25	-0.04	-0.35	-0.23	0.26									
18	14285318881075	-0.48	-0.28	-0.43	-0.24	0.19	-0.51	-0.43	-0.38	0.31	0.31	-0.23	-0.25	-0.22	0.21	-0.29	-0.05	-0.44								
19	302403121942-0.09	0.79	-1.83	0.02	1.0	0.48	-1.53	0.54	1.25	0.38	1.32	-0.03	1.75	-1.38	-0.3	-0.61	-0.57									
20	11132309863-0.72	0.59	-1.42	0.28	1.05	0.24	-0.45	-1.81	0.21	-0.31	0.83	0.46	0.33	-1.1	0.06	0.45	-1.11									
21	75473319798-0.57	1.09	-0.37	0.5	-1.11	-0.57	-1.87	0.65	-1.27	0.44	1.4	-1.73	0.03	0.91	-2.01	-0.41	0.19									
22	457300804-0.17	-0.47	-1.13	-0.34	0.32	-0.82	-0.85	-0.38	0.03	0.43	0.48	1.41	-1.31	-0.01	-0.31	-0.05	-0.15									
23	13785316288008	0.77	0.12	0.58	0.83	-0.38	0.01	0.51	-0.34	0.08	-0.85	-0.2	0.98	-0.77	0.93	0.75	1.15									
24	6438307988042	1.25	-0.24	-1.48	-1.38	-0.38	-1.4	0.3	1.8	1.81	1.23	-1.13	-0.31	0.39	-0.38	0.85	0.1									
25	5451306403119	-0.62	0.59	0.51	-1.11	-1.8	-0.55	-1.6	0.86	-2.01	0.6	-1.21	1.5	-2.49	0.61	0.32	1.71									
26	3029304588-0.04	-0.24	-0.68	0.13	1.44	-0.4	0.32	0.31	-0.75	-0.95	-0.36	1.96	1.23	-1.41	-0.98	1.49	-0.52									
27	7526309529-0.29	0.04	0.58	-1.28	1.07	-0.55	-0.98	-1.3	-0.51	0.31	1.03	-0.61	-0.79	0.21	-0.32	0.46	0.95									
28	1095301588014	0.59	0.69	-0.35	0.28	0.48	0.3	-0.64	0.07	0.4	0.34	1.31	0.35	0.34	-0.39	0.78	0.95									
29	TT3930968-0.09	0.01	0.48	-0.36	-0.5	0.3	-0.37	0.39	1.2	1.04	-1.11	-0.87	-0.18	0.89	-0.76	-0.12	0.85									
30	5446309701014	0.7	1.3	1.31	0.29	0.94	1.81	0.16	0.44	0.16	0.34	1.06	0.34	0.34	0.34	0.34	1.49	1.4								

Gènes

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X		
1		123456	123456	123456	123456	123456	123456	123456	123456	123456	123456	123456	123456	123456	123456	123456	123456	123456	123456	123456	123456	123456	123456			
2	731523119802	-0.21	0.99	0.98	-0.2	1.05	-0.52	-1.25	0.31	-0.34	0.12	-1.09	0.8	1.22	-0.95	-0.95	1.54	-1.19								
3	3078309853	-0.57	1.11	-0.71	-1.20	1.48	-1.37	-0.88	-0.38	0.39	0.30	-0.51	0.63	0.51	0.44	-0.91	0.52	0.09								
4	129301114022	0.2	-0.27	-0.79	-0.45	-0.89	-0.82	-0.38	-0.5	-0.33	0.53	0.86	-0.41	-0.52	-1.01	-1.01	-1.47									
5	752583119802	-0.29	0.62	-1.2	0.2	1.02	-0.3	-0.24	0.76	0.38	0.35	0.79	0.24	0.05	-2.09	0.66	0.19	-0.01								
6	340853119808	0.12	-0.48	-0.56	0.08	-0.33	-0.37	1.01	0.85	-1.96	1.30	0.54	2.31	1.81	0.71	2.34	0.61									
7	3298309878046	0.72	-2.71	0.26	-1.73	-0.88	-2.38	0.26	1.47	-0.8	-0.73	-1.89	1.44	-0.26	-0.01	-0.16	-0.3									
8	1757149630	-0.11	1.04	-1.52	0.48	0.74	-0.24	-0.85	0.36	0.35	-0.34	0.45	0.71	-0.33	-0.34	-0.57	1.32	-0.17								
9	4720308824	-0.62	0.99	-1.02	-0.20	-0.39	-0.79	-0.55	-1.06	-0.21	-0.43	1.01	0.86	0.33	-0.37	-0.15	1.01	-0.56								
10	18253119888	-1.20	0.63	-1.22	0.56	1.23	-0.7	-0.8	0.35	0.26	-0.37	-0.83	1.86	-0.01	-1.35	-0.26	0.67	-0.39								
11	110843119888	-0.62	0.32	-0.32	-1.24	-2.55	0.15	-0.38	-0.24	-0.2	0.41	0.58	0.56	-0.11	-0.21	0.52	0.42	1.94								
12	3298309803	0.48	0.77	-0.44	-0.77	-0.34	0.18	-1.81	0.38	0.38	0.36	1.61	0.21	-0.29	-0.19	-0.92	1.0	0.75								
13	141733119889	0.28	0.39	-2.18	-0.24	-1.32	-0.38	-4.75	0.31	0.71	0.04	0.81	0.67	0.81	1.81	0.31	-0.38	-1.45								
14	91583119802-1	21	1.17	-1.13	0.93	0.98	-0.74	-0.47	0.38	-0.21	-0.1	0.85	1.03	-1.56	0.05	1.85	-0.59									
15	750583119880-1	54	1.17	-1.05	-0.81	1.25	-0.4	-0.53	-0.42	-0.47	-0.87	-1.1	2.49	1.51	-1.51	-0.61	-2.72	1.1								
16	145883119880	-0.28	0.77	-0.34	-1.23	0.35	-0.34	-0.28	-0.39	0.34	0.78	1.05	-0.41	-1.95	-0.45	1.25	-0.45									
17	1315301810-0.28	0.14	-1.14	0.11	0.3	-0.01	-0.01	-0.01	-0.81	0.48	-0.06	-0.46	-0.25	-0.04	-0.35	-0.13	0.25									
18	1425531198810	78	-0.48	-0.38	-0.41	-0.24	0.39	-0.87	-0.41	-0.38	0.31	0.31	-0.25	-0.22	0.21	-0.29	-0.05	-0.44								
19	301903119882	-0.09	0.79	-1.83	0.92	1.0	0.46	-1.10	0.54	1.25	0.38	1.32	-0.82	1.75	-1.35	-0.1	-0.61	-0.57								
20	11132309882	-0.12	0.39	-1.42	0.28	1.05	0.24	-0.48	-1.81	0.21	-0.31	0.83	0.46	0.33	-1.1	0.06	0.45	-1.11								
21	154713119798	-0.57	1.09	-2.37	0.5	-1.11	-0.57	-1.87	0.65	-1.21	0.44	1.4	-0.72	0.83	0.81	-2.01	-0.41	0.19								
22	457300884	-0.17	-0.47	-1.13	-0.84	0.32	-0.82	-0.85	0.08	0.02	0.43	0.48	1.41	-0.21	-0.01	-0.31	-0.05	-0.19								
23	1378531198880	0.06	0.77	0.12	0.58	0.81	-0.38	0.31	0.51	-0.34	0.38	-0.85	-0.2	0.98	-0.77	0.53	0.25	1.13								
24	6438307988042	125	-2.14	-1.48	-1.38	-0.39	-1.4	0.1	1.8	1.87	1.22	-0.13	-0.21	0.29	-0.18	0.85	0.1									
25	545130640119	-0.62	0.59	0.51	-1.11	-1.8	-0.52	-1.03	0.85	-2.07	0.6	-0.21	1.5	-0.49	0.81	2.32	1.71									
26	30283045880	-0.24	-0.58	0.15	1.46	-0.4	0.38	0.31	-0.75	-0.95	-0.36	1.96	1.21	-1.41	-0.98	1.49	0.52									
27	1926309829	-0.19	0.04	-0.58	-1.29	1.07	-0.55	-0.88	-1.2	-0.51	0.31	1.83	-0.61	-0.79	0.27	-0.12	0.46	0.95								
28	3005301588014	0.59	-0.68	-0.35	0.28	-0.49	-0.03	-0.64	-0.07	0.4	0.34	1.31	0.35	0.14	-0.39	0.78	0.92									
29	1138309848	-0.09	0.01	0.48	-0.34	-0.05	0.03	-0.17	0.38	1.2	1.94	-1.11	-0.91	-0.18	0.89	-0.16	-0.12	0.85								
30	34443041014	0.7	1.5	-1.31	0.26	0.94	1.81	0.15	0.48	0.26	0.14	1.95	0.14	0.14	0.14	1.49	1.0									

Intensité  
(expression)  
d'un gène  
mesuré par  
une puce

Puces

Très grande densité : Affymetrix U-133 plus 2.0 Array 54,675 probesets

# Les motifs séquentiels dans ce contexte...

---

- **Motifs séquentiels** : séquences fréquentes d'itemsets ordonnés

< (   ) (  ) >

- Rechercher des motifs séquentiels pour **mettre en évidence des gènes dont les expressions sont fréquemment ordonnées de la même manière**

< (G5 G4) (G6) >

- **2 exemples avec**
  - MMDN sur la maladie d'Alzheimer
  - IRCM sur le cancer du sein.

# Maladie d'Alzheimer : problème majeur de la société moderne

---

- **Maladie d'Alzheimer (AD)** : la forme la plus commune de démence
  - 26.6 millions de personnes atteintes (2006)
  - Augmentation du nombre de patients (\*4 en 2050)
- Intérêt de la communauté biomédicale pour la découverte de gènes impliqués dans le développement la maladie
- **MMDN** : travaillent sur l'AD et sur le vieillissement à partir d'un modèle animal, *Microcebus murinus*
- Objectifs : **comparer les tissus du cortex cérébral** de lémuriens jeunes (sains) avec ceux de lémuriens âgés (malades) pour étudier le vieillissement (la maladie d'Alzheimer)



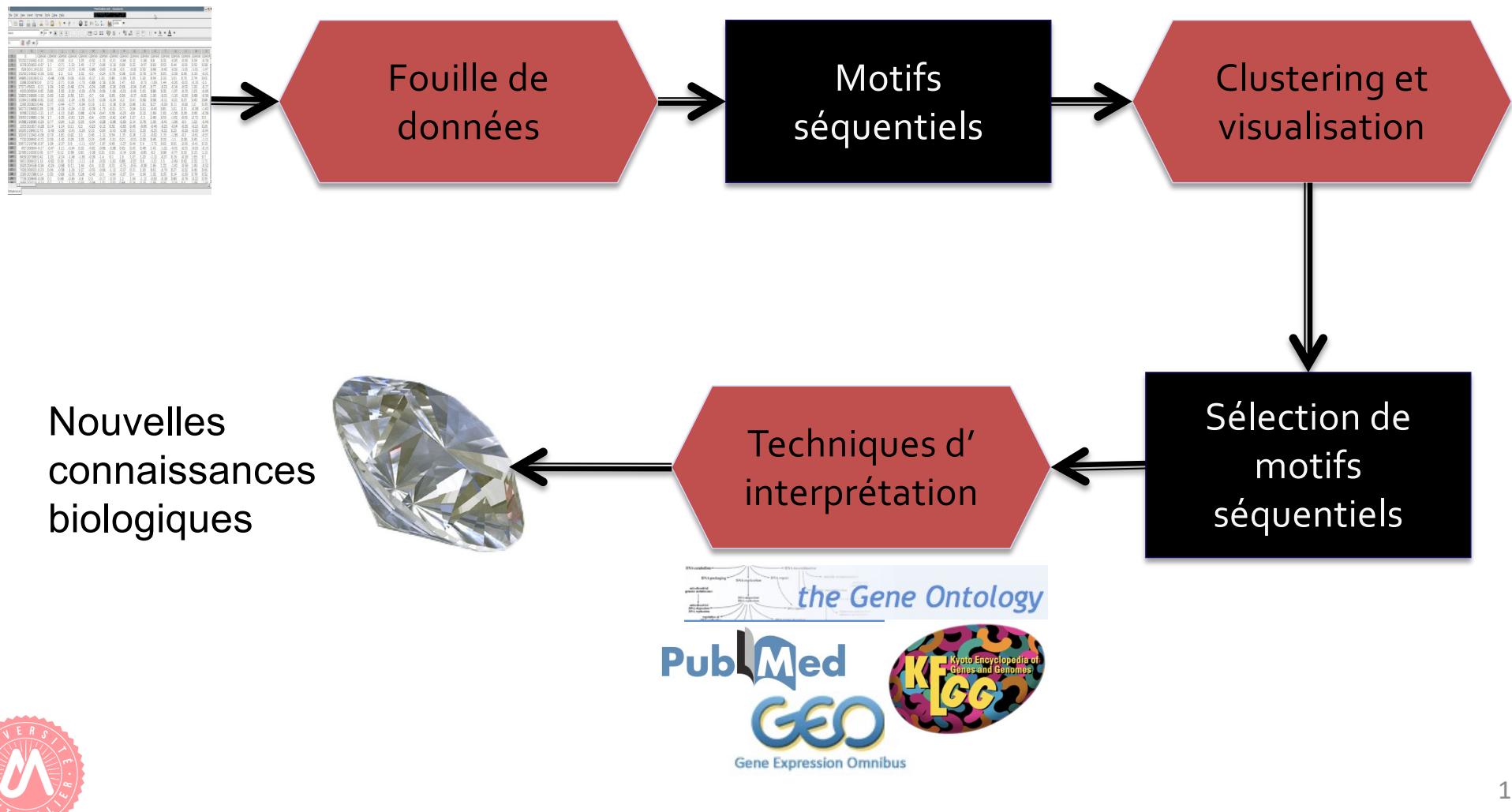
# Cancer du sein : première cause de mortalité entre 45 et 64 ans (2004)

---

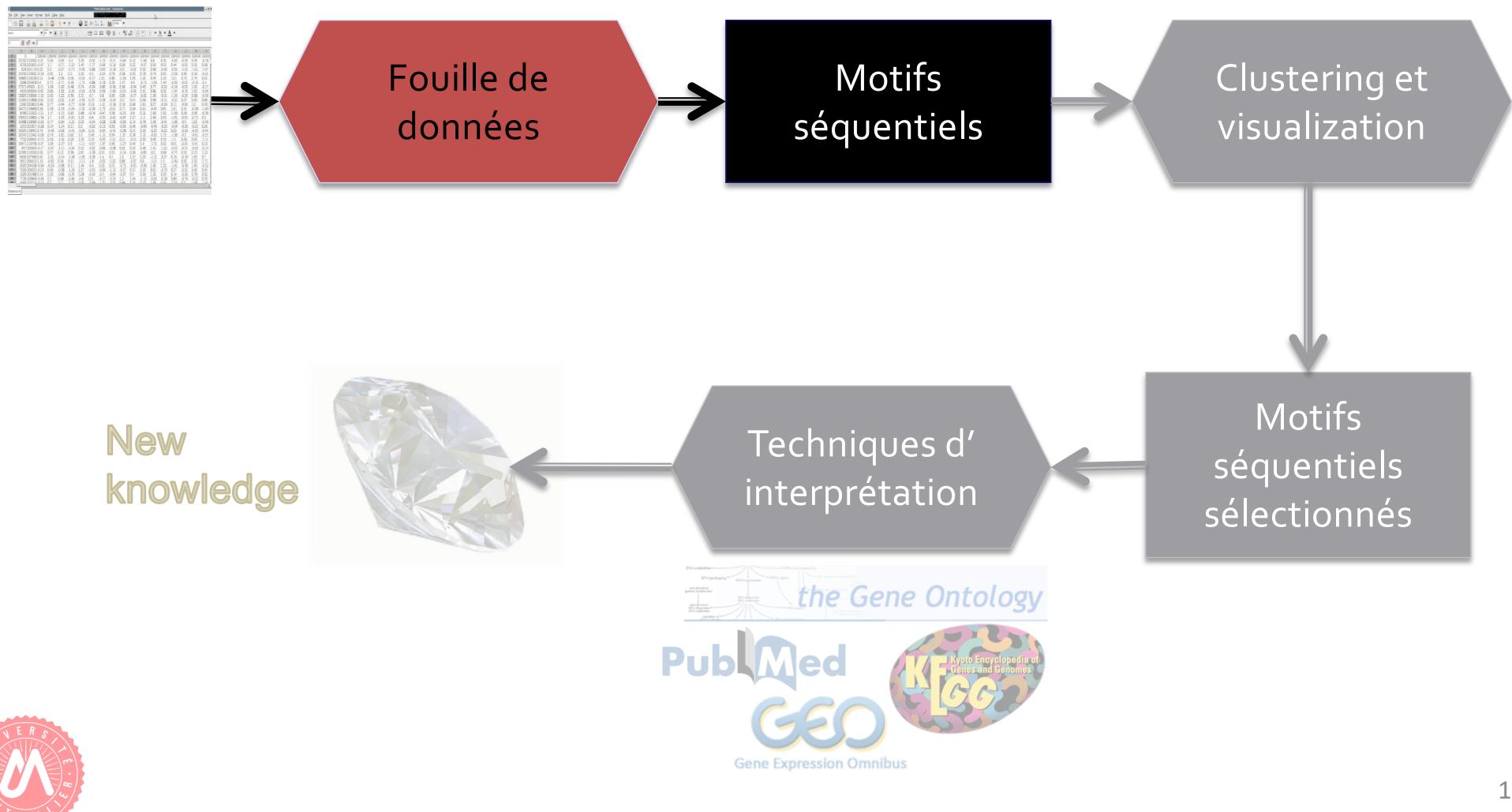
- **Perturbation de la communication cellulaire**, associée à une **absence de mort cellulaire**, engendrant le **développement d'amas de cellules cancéreuses** (appelées tumeurs) qui échappent aux règles de fonctionnement du corps.
- **IRCM** : utilisent les puces ADN pour comparer les tissus issus de tumeurs du sein, répertoriés selon différents grades.
- **Objectif** : déterminer un ensemble de biomarqueurs suffisants pour **typer ces tumeurs**.
- **Enjeu considérable** : Les thérapies sont + ou - toxiques et fonctionnent sur un patient mais pas sur un autre. Typer une tumeur s'avère crucial pour le choix d'une thérapie.



# Processus général



# Processus général



# Recherche de motifs séquentiels

Puces	Séquences de gènes
M <sub>1</sub>	<(G <sub>2</sub> )(G <sub>1</sub> G <sub>5</sub> )(G <sub>3</sub> )(G <sub>4</sub> )>
M <sub>2</sub>	<(G <sub>2</sub> )(G <sub>1</sub> G <sub>5</sub> )(G <sub>4</sub> )(G <sub>3</sub> )>
M <sub>3</sub>	<(G <sub>2</sub> )(G <sub>4</sub> )(G <sub>1</sub> G <sub>5</sub> )(G <sub>3</sub> )>
M <sub>4</sub>	<(G <sub>2</sub> )(G <sub>3</sub> )(G <sub>1</sub> G <sub>5</sub> )(G <sub>4</sub> )>

<(G<sub>2</sub>)(G<sub>1</sub> G<sub>5</sub>)(G<sub>3</sub>)>

- Le gène G<sub>2</sub> a une expression plus petite que les gènes G<sub>1</sub> et G<sub>5</sub> qui ont une expression similaire et plus petite que le gène G<sub>3</sub>



# Recherche de motifs séquentiels

Puces	Séquences de gènes
M <sub>1</sub>	<(G <sub>2</sub> )(G <sub>1</sub> G <sub>5</sub> )(G <sub>3</sub> )(G <sub>4</sub> )>
M <sub>2</sub>	<(G <sub>2</sub> )(G <sub>1</sub> G <sub>5</sub> )(G <sub>4</sub> )(G <sub>3</sub> )>
M <sub>3</sub>	<(G <sub>2</sub> )(G <sub>4</sub> )(G <sub>1</sub> G <sub>5</sub> )(G <sub>3</sub> )>
M <sub>4</sub>	<(G <sub>2</sub> )(G <sub>3</sub> )(G <sub>1</sub> G <sub>5</sub> )(G <sub>4</sub> )>

<(G<sub>2</sub>)(G<sub>1</sub> G<sub>5</sub>)(G<sub>3</sub>)>

Support = 3/4



# Recherche de motifs séquentiels

Puces	Séquences de gènes
M <sub>1</sub>	<(G <sub>2</sub> )(G <sub>1</sub> G <sub>5</sub> )(G <sub>3</sub> )(G <sub>4</sub> )>
M <sub>2</sub>	<(G <sub>2</sub> )(G <sub>1</sub> G <sub>5</sub> )(G <sub>4</sub> )(G <sub>3</sub> )>
M <sub>3</sub>	<(G <sub>2</sub> )(G <sub>4</sub> )(G <sub>1</sub> G <sub>5</sub> )(G <sub>3</sub> )>
M <sub>4</sub>	<(G <sub>2</sub> )(G <sub>3</sub> )(G <sub>1</sub> G <sub>5</sub> )(G <sub>4</sub> )>

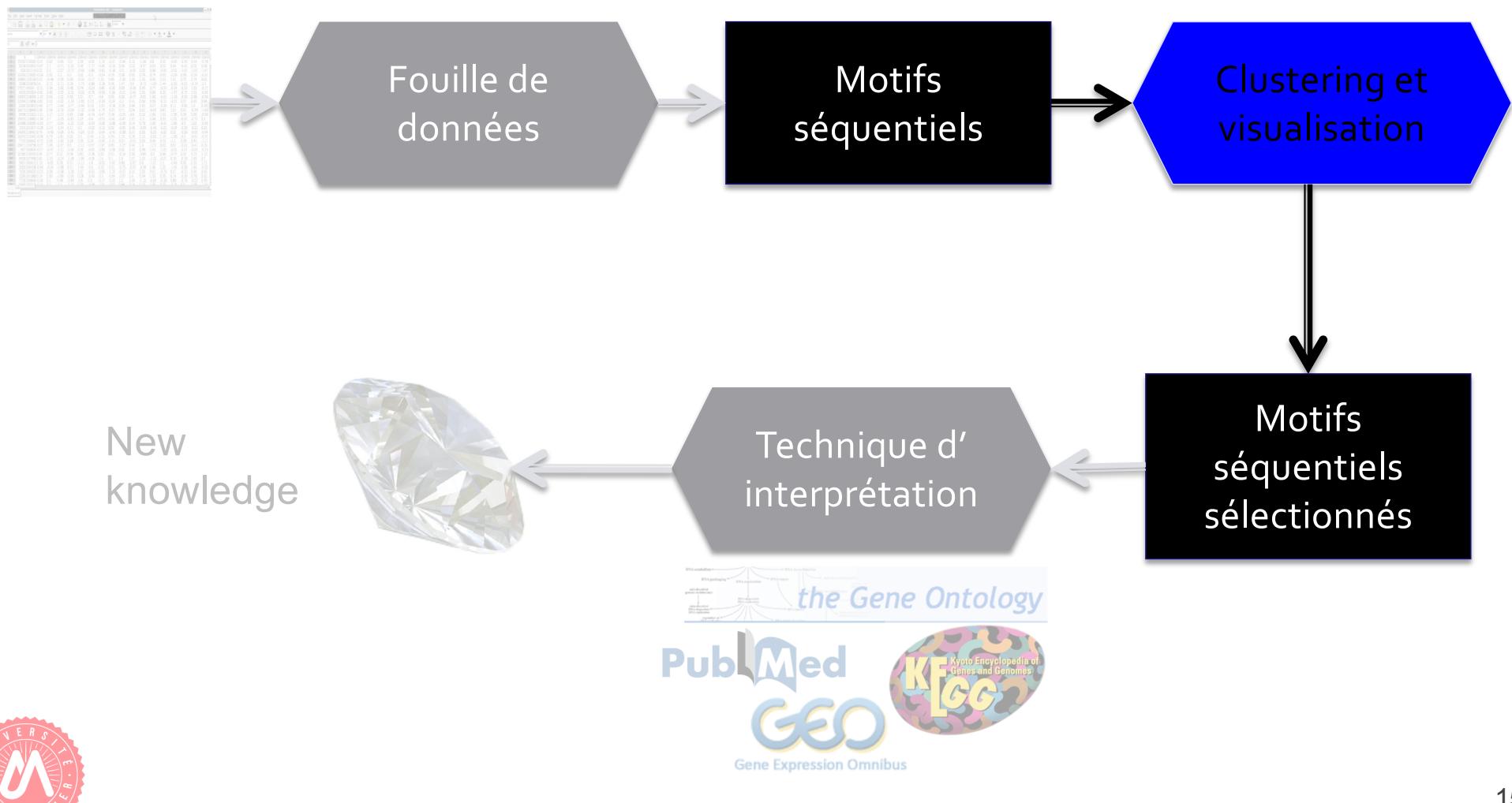
<(G<sub>2</sub>)(G<sub>1</sub> G<sub>5</sub>)(G<sub>3</sub>)>

Support = 3/4

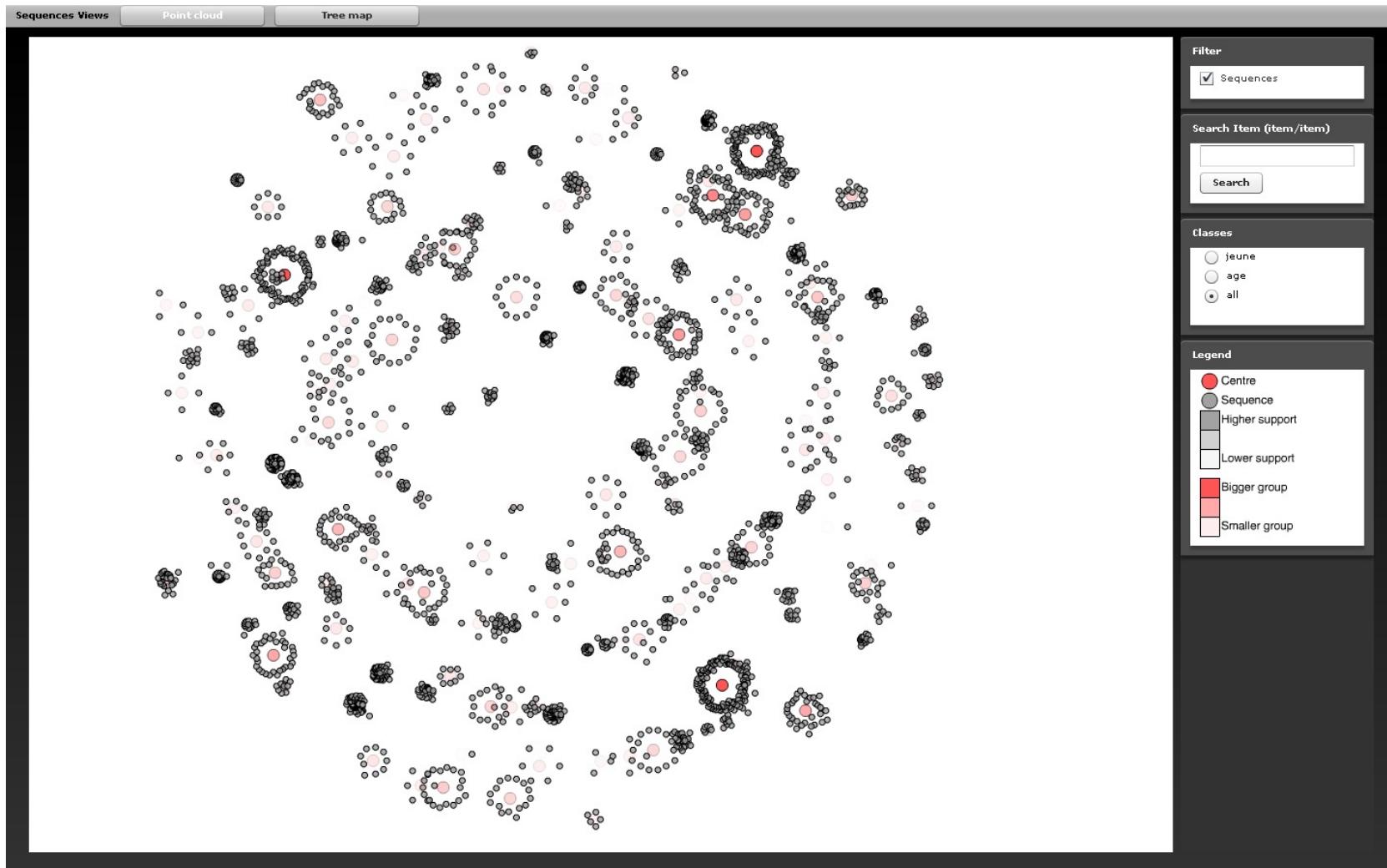
- Motifs séquentiels discriminants
  - Fréquents dans une classe (malades)
  - Non fréquents dans la classe complémentaire (sains)



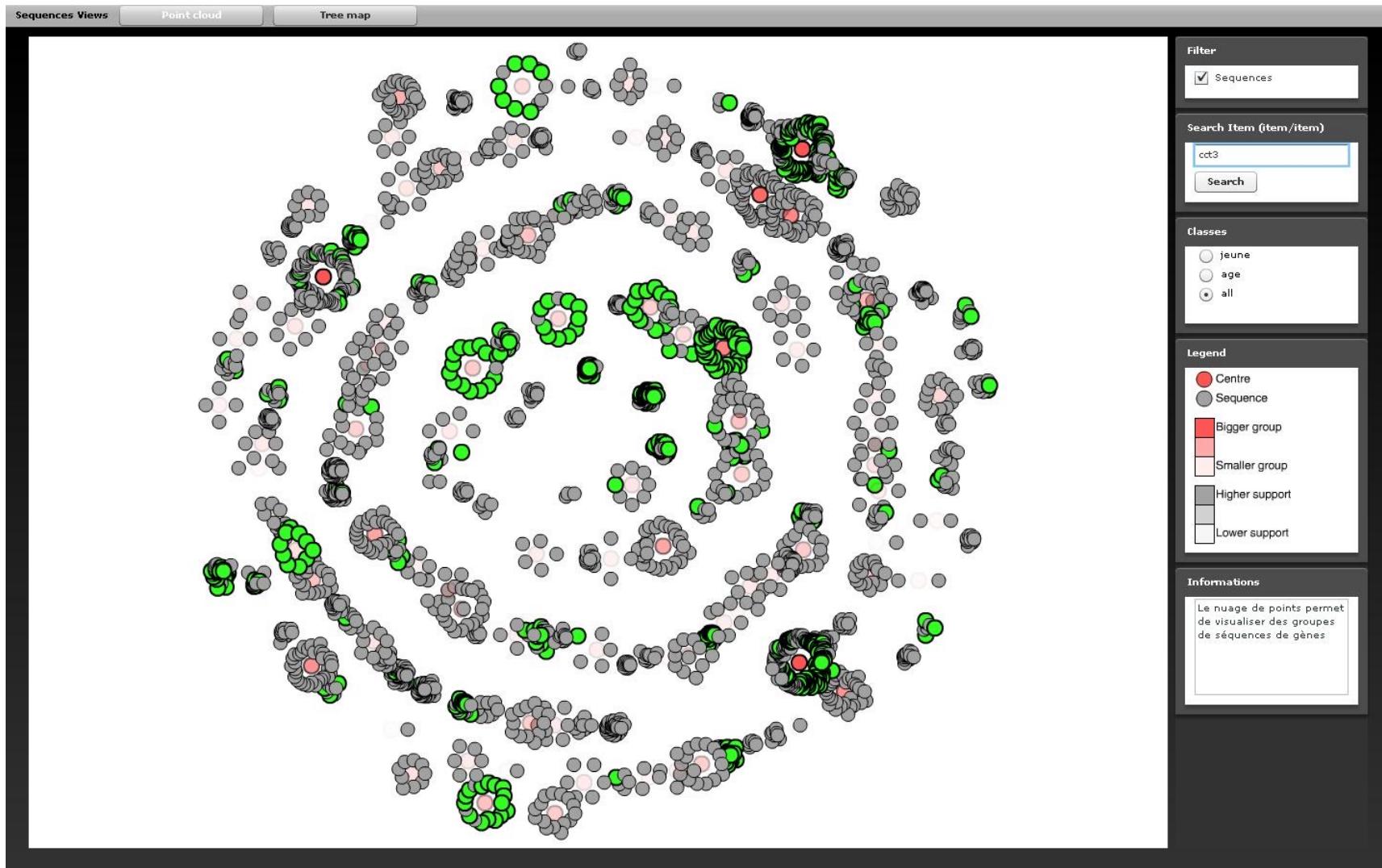
# Processus général



# Clustering simple (k-means)



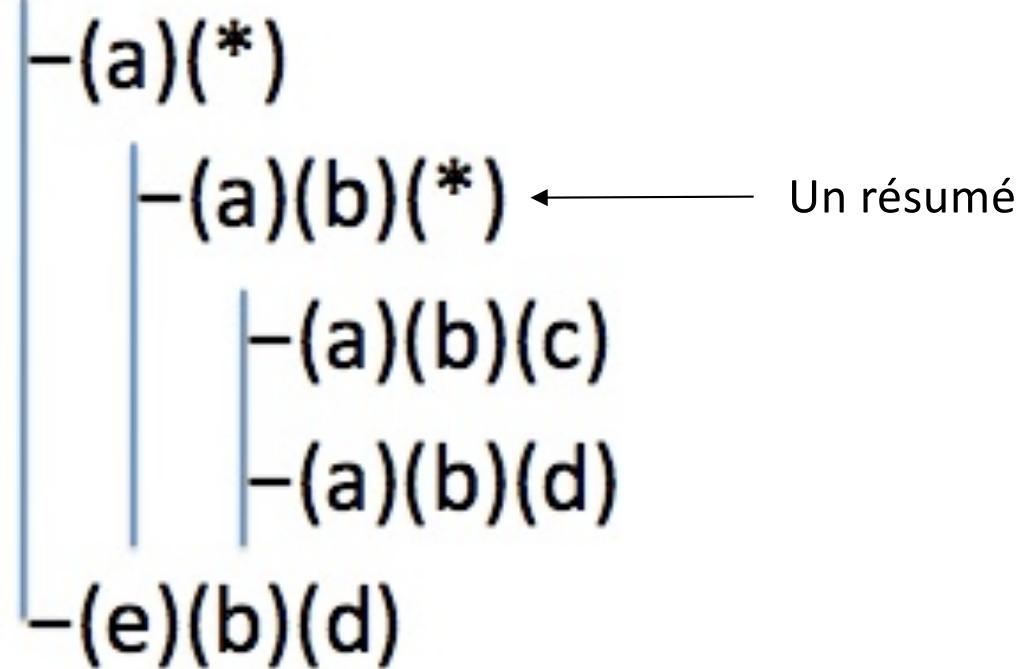
# Clustering simple (k-means)



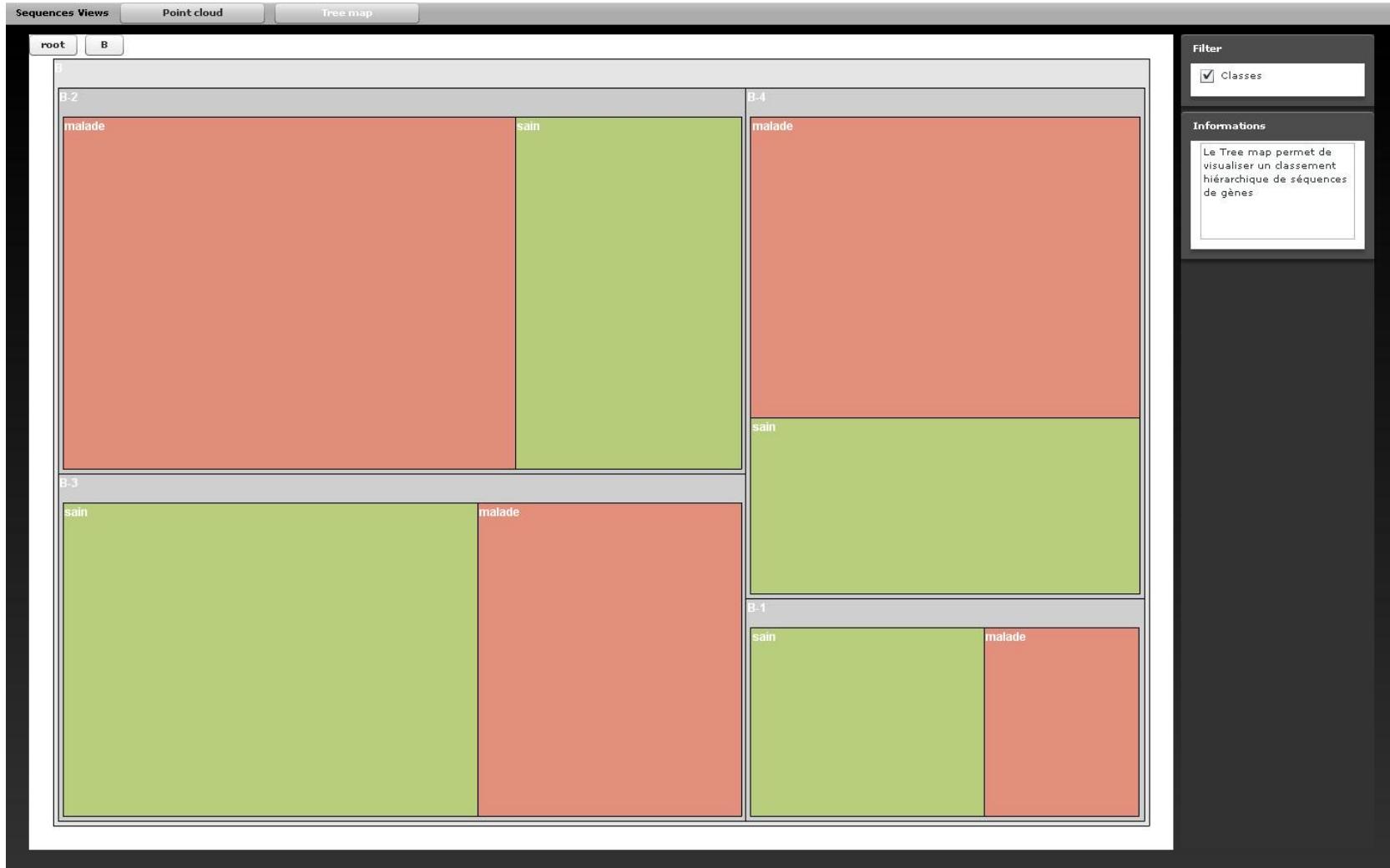
# Clustering hiérarchique

---

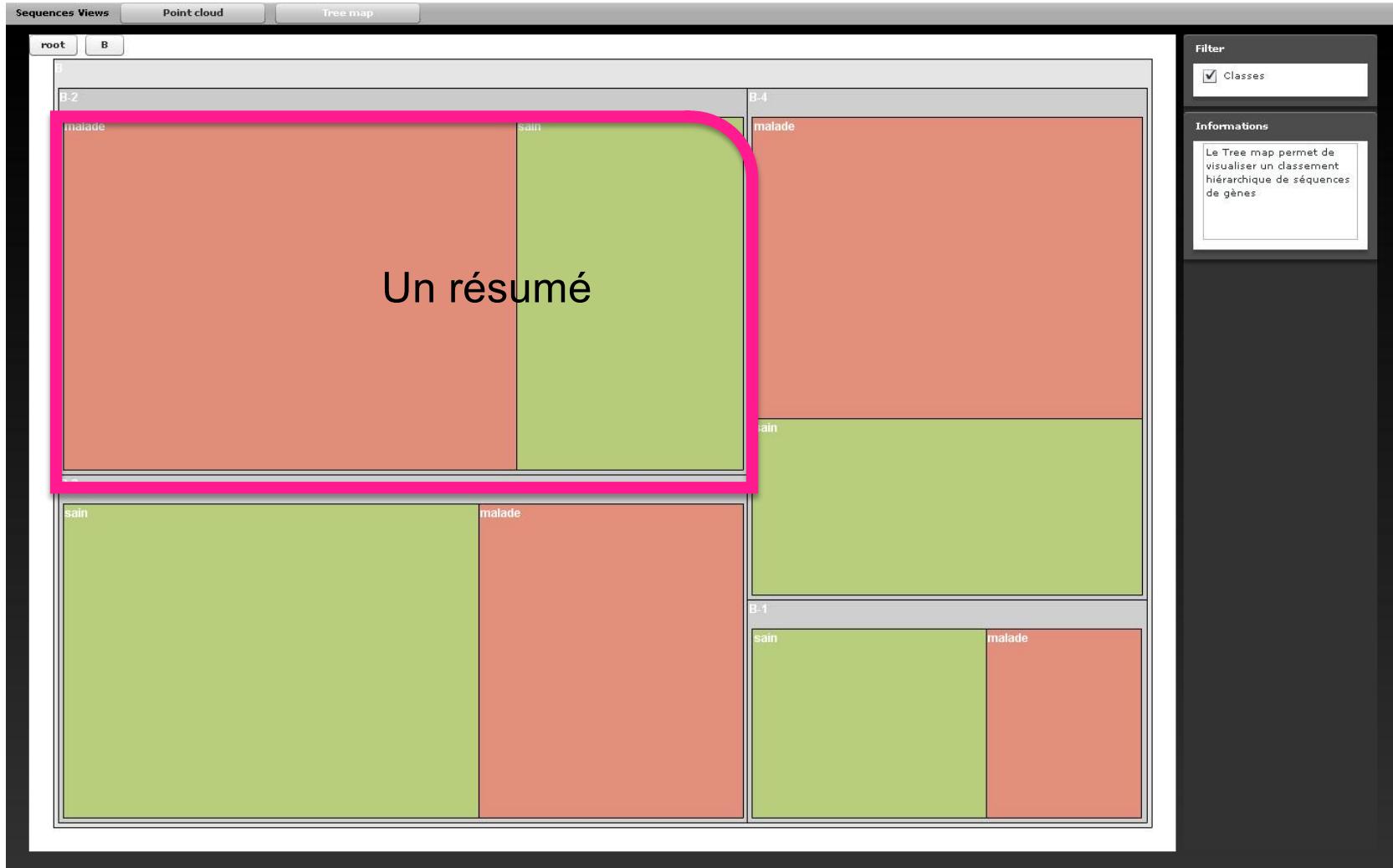
- Exemple:  $(a)(b)(c)$ ,  $(a)(b)(d)$ ,  $(e)(b)(d)$



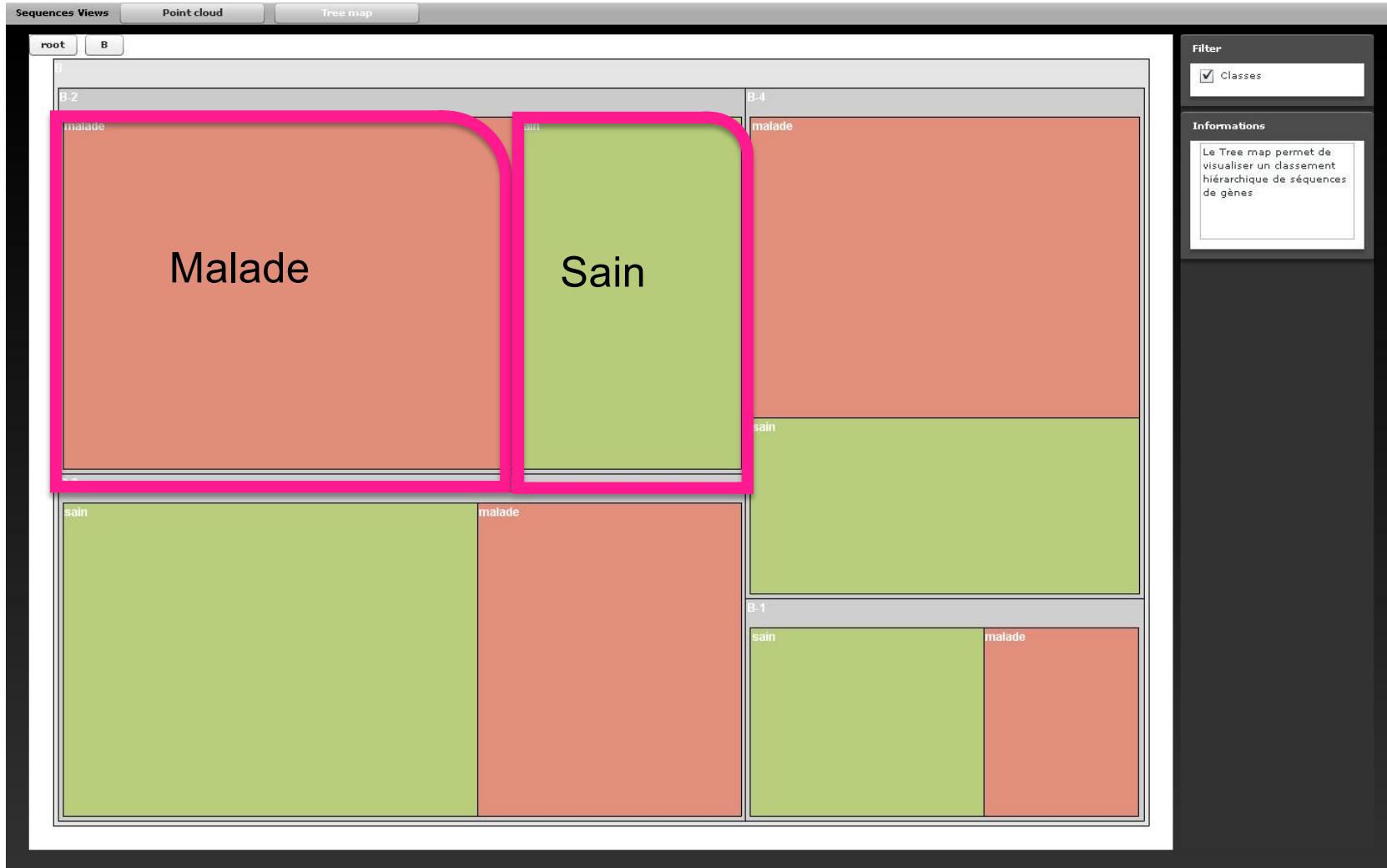
# Clustering hiérarchique



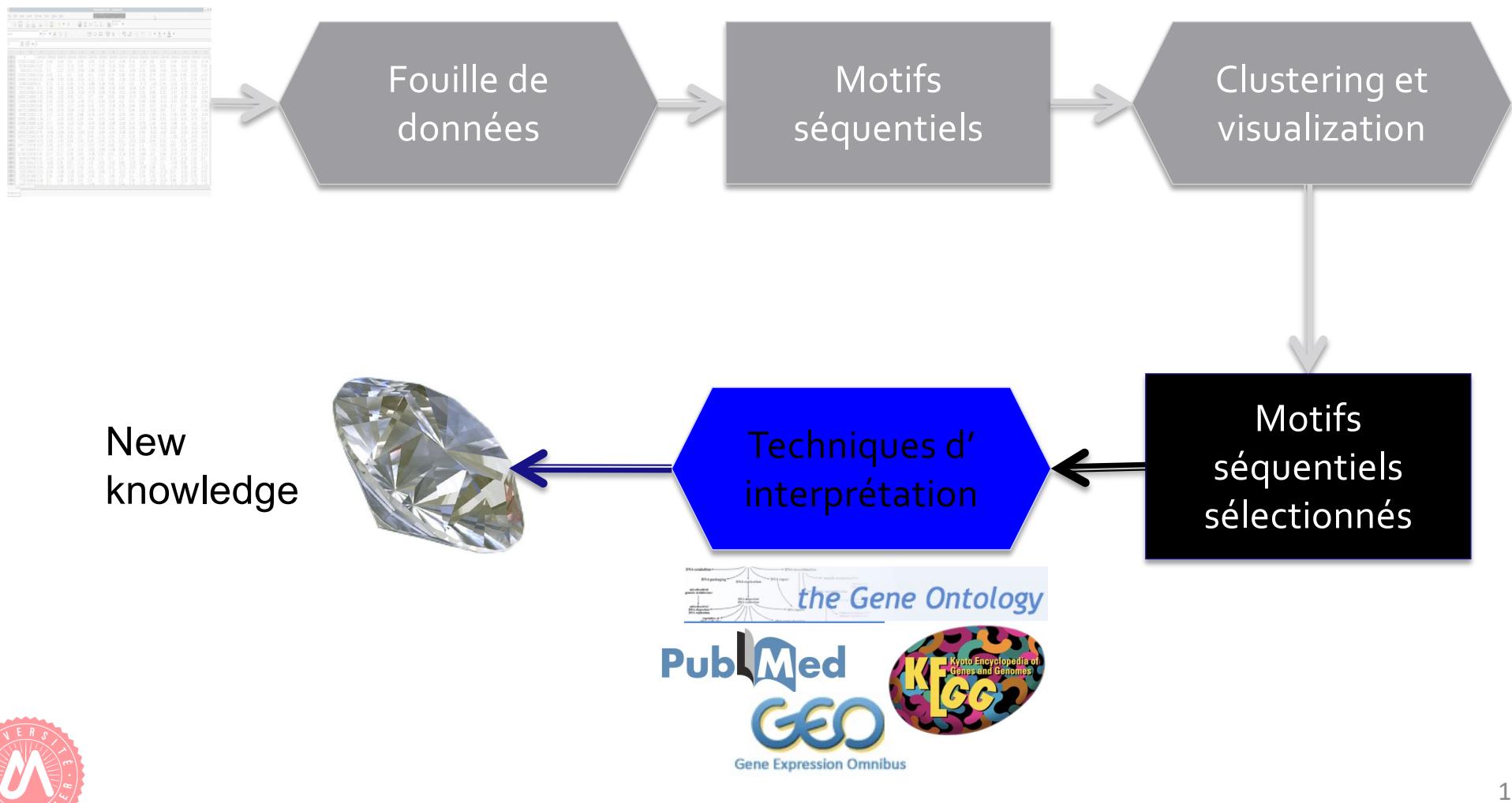
# Clustering hiérarchique



# Clustering hiérarchique



# Processus général



# Interprétation des motifs via les documents

---

$S_{75\%,25\%} = \langle (G_1)(G_2 G_3) \rangle$



Textes

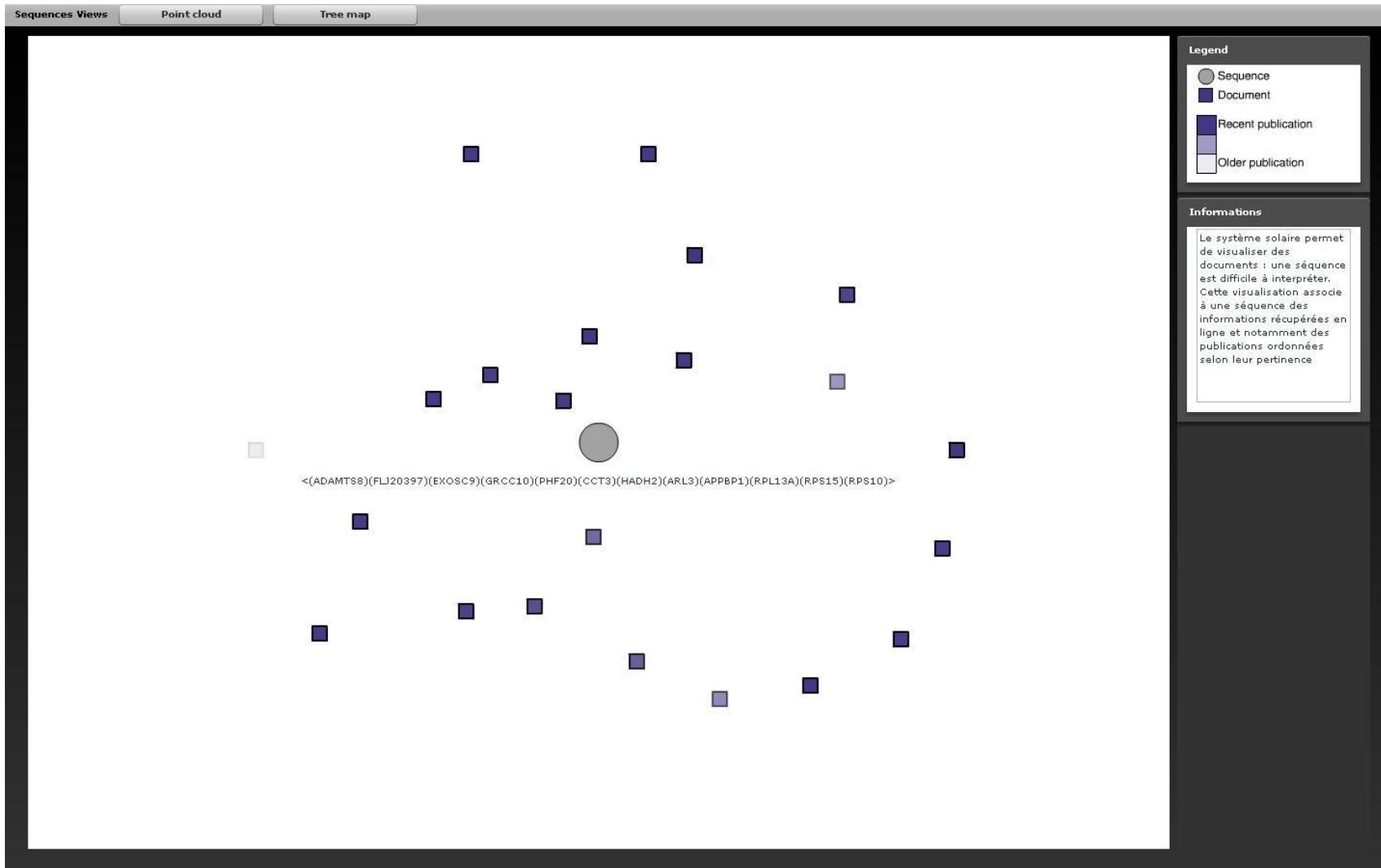


- Recherche de documents associés avec les gènes des motifs
- **Objectifs:** validation + recherche de nouveautés

## Séquences populaires et innovantes

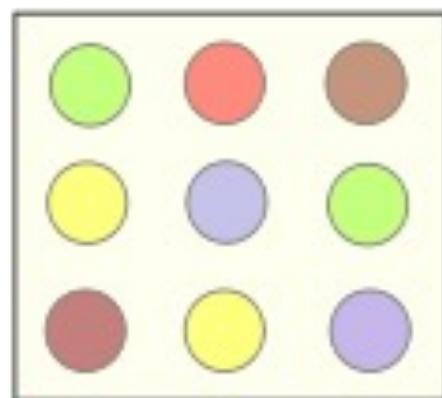


# Visualisation de documents



# Typer les tumeurs à partir des motifs

---



Tumeur bénigne

Tumeur maligne

Classement grâce aux motifs séquentiels

# Classement selon 3 classes

---

- Classe 1 : faible
- Classe 2 : moyen
- Classe 3 : forte

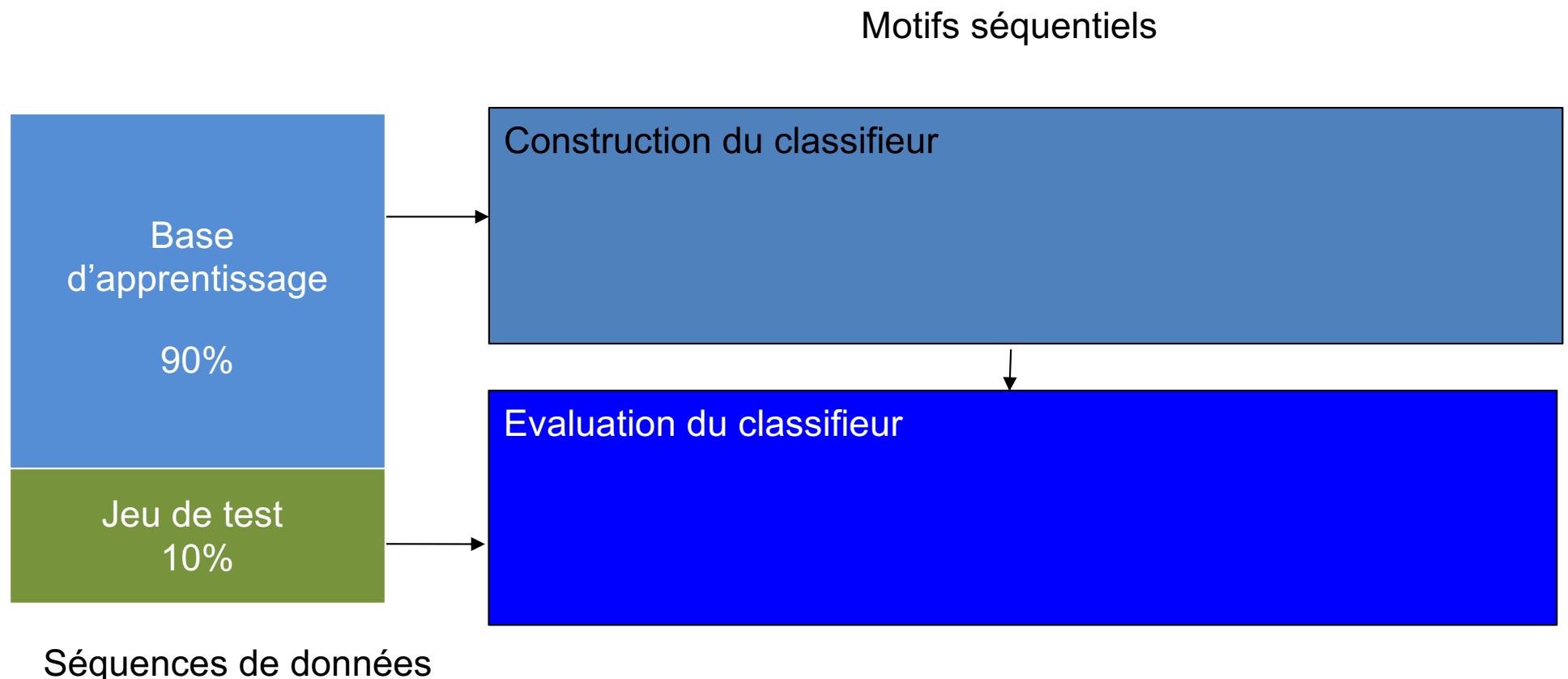
**0.96% de rappel et 0.97% de précision  
selon le jeu de données**

The screenshot shows a software interface for sequence analysis. At the top, there are tabs for "Séquences", "Motifs", "Classification", and "Résultats". The "Classification" tab is active. Below it, there is a large text area containing XML code representing a new pattern. At the bottom, there is a table titled "Classification" with columns "Classe" and "Support". The data in the table is as follows:

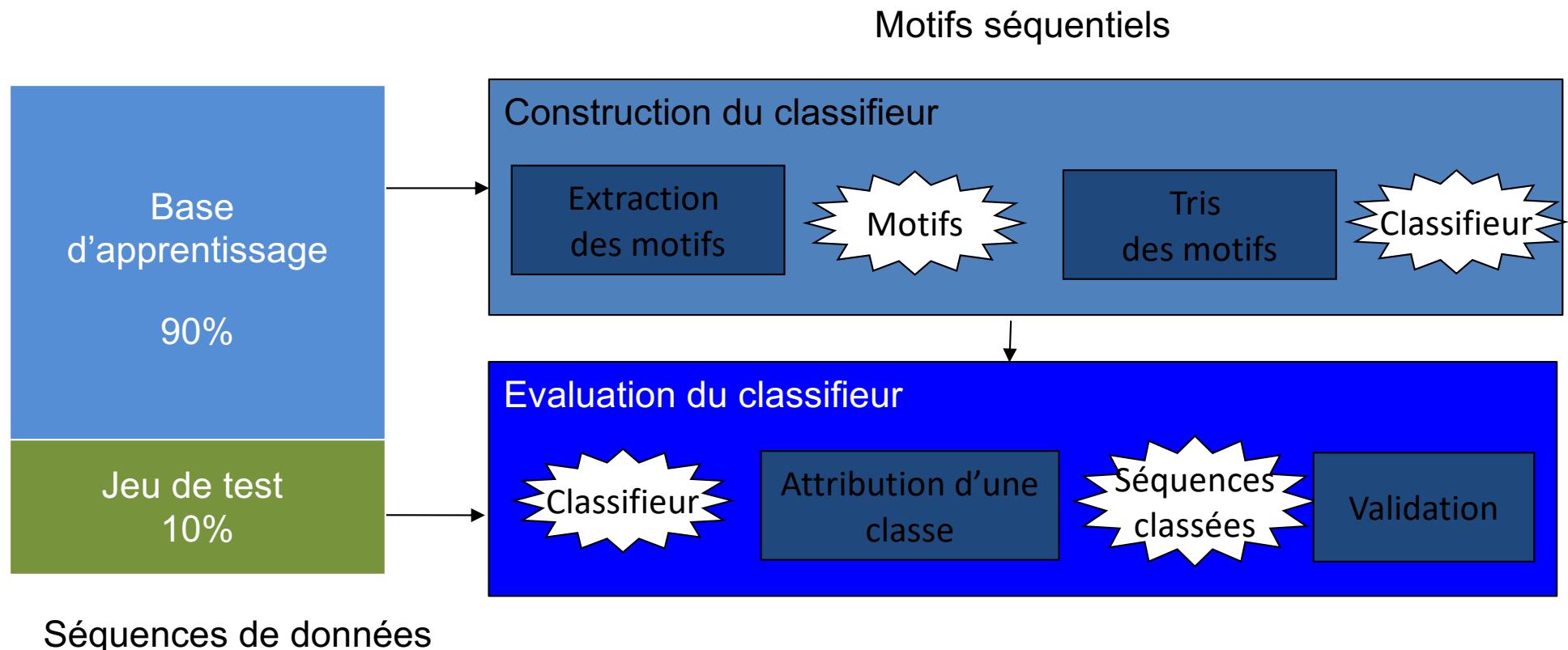
Classe	Support
1 1	0.727273
2 2	0.145833
3 3	0.0



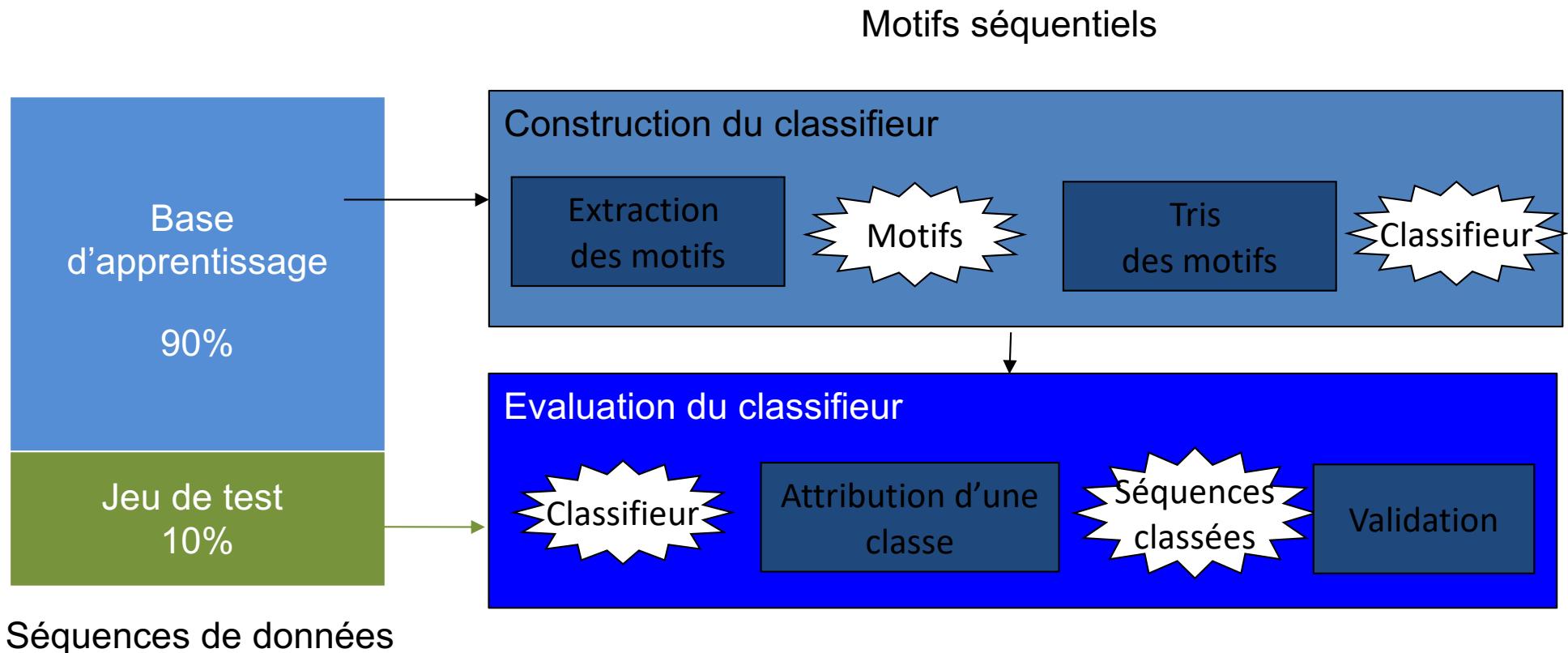
# Schéma général



# Schéma général



# Schéma général



**Répétition pour validation croisée**

# Logiciels disponibles

---

- **Weka (un peu)**
- **SPMF: An Open-Source Data Mining Library**

[http://www.philippe-fournier-viger.com/  
spmf/index.php?link=algorithms.php](http://www.philippe-fournier-viger.com/spmf/index.php?link=algorithms.php)

(le plus complet)

- Pas d'implémentation directe dans Scikit-Learn
- Solutions ad-hoc
- Intégré dans des solutions industrielles



# Voir notebook

---

- ExtractSequentialPattern.ipynb
- Utilisation de AprioriAll, PrefixSpan (Python-Java)



# Conclusions

---

- Les motifs séquentiels sont une petite partie des patterns à extraire ...
  - Arbres, graphes, multigraphes ...
  - De nombreuses approches existent
- Ce qu'il faut retenir : les patterns sont différents, les usages sont différents mais les contraintes existent aussi quelques soient les types de patterns

