

# Fouille de Données

# Data Mining

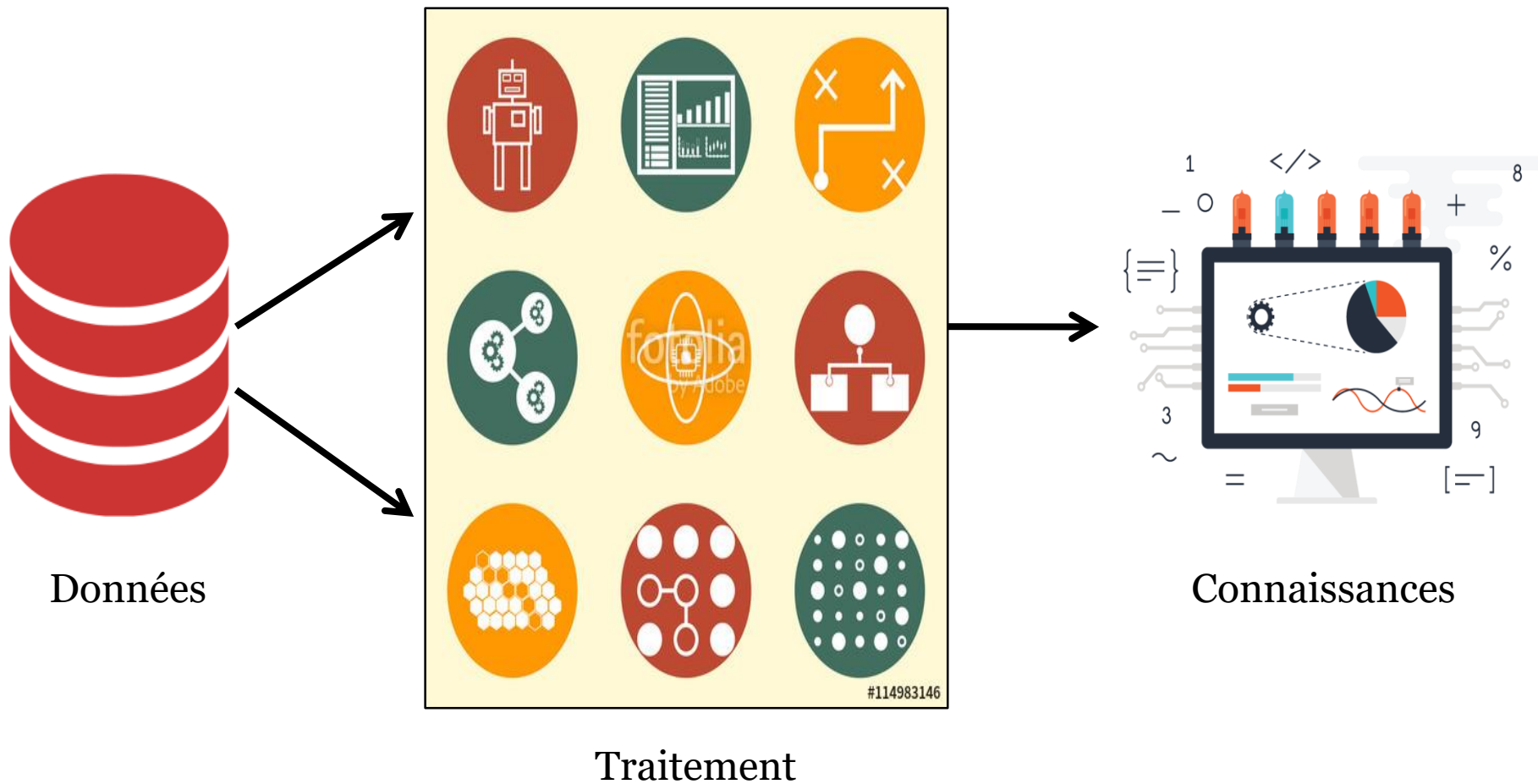
**Recherche des Motifs Fréquents**  
**et Extraction des Règles d'Association**

# Plan du cours

1. Contexte
2. Concepts de base
3. Méthodes pour la recherche des modèles fréquents
4. Types de motifs fréquents
5. Passage aux règles d'association
6. Motifs rares
7. Motifs fréquents séquentiels

# Fouille de Données - Data Mining

**SAVOIR – PREDIR/DECOUVRIR - DECIDER**

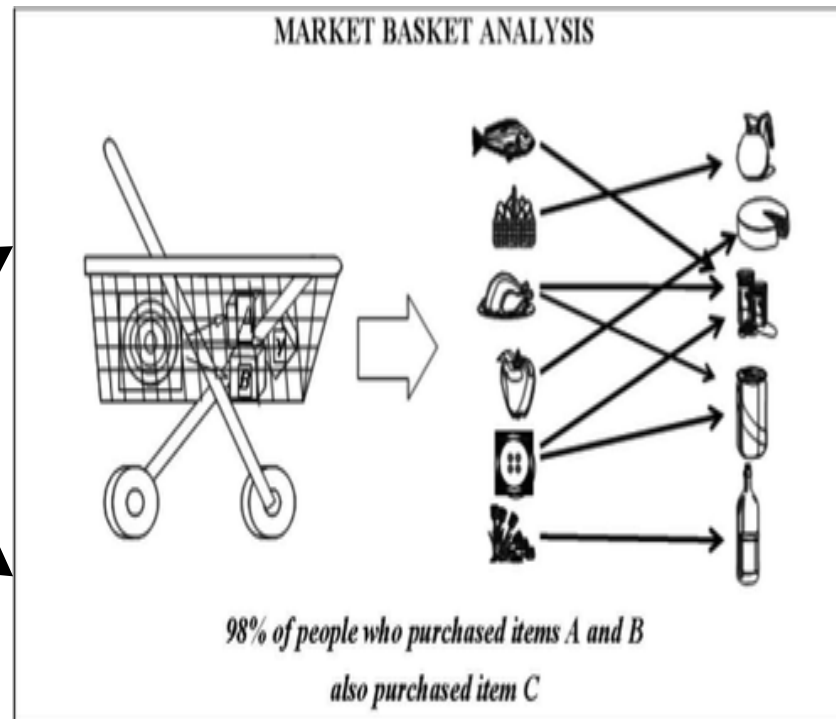
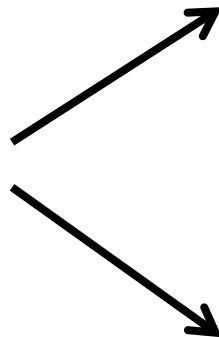


Contexte

**SAVOIR – **DECOUVRIR** - DECIDER**



Données



Algorithmes

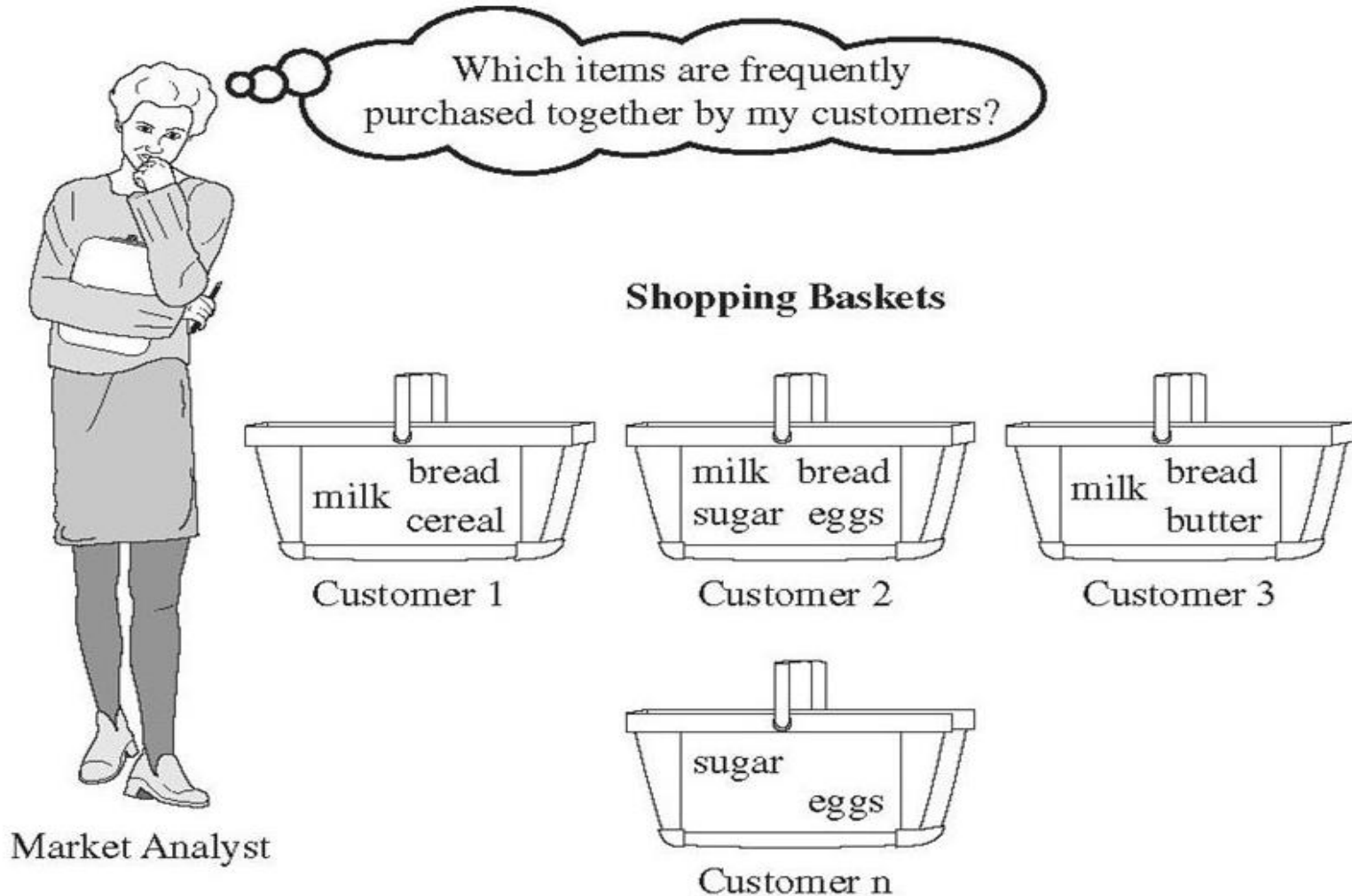


Motifs  
Fréquents et  
Règles  
d'association

## Contexte

### Analyse du panier de marché

Quels sont les produits qui apparaissent fréquemment (ensemble) dans un ensemble de données ?



## Contexte

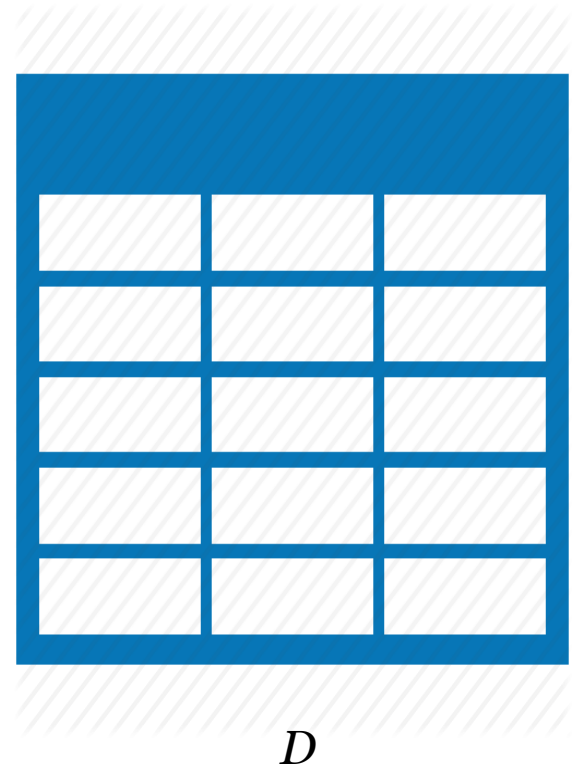
- Analyse des données médicales : Quelles sont les maladies qui apparaissent fréquemment (ensemble) ?
- Analyse d'ADN en biologie afin de comprendre les propriétés génétiques des espèces.
- L'analyse du climat en météorologie afin de mieux orienter l'agriculture ?

# Concepts de base

## 1 - Base de données formelle



Données

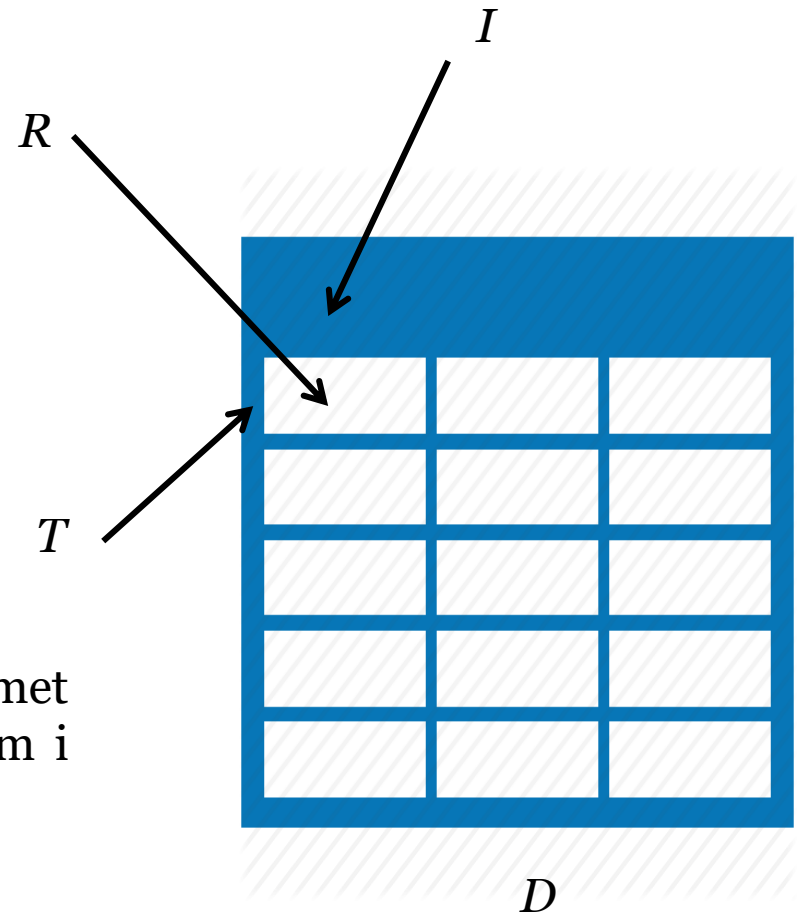


# Concepts de base

## 1-Base de données formelle

définie par le triplet :  $(T, I, R)$

- ✓  $T$ : ensemble fini d'instances
- ✓  $I$ : ensemble fini d'items
- ✓  $R$ : relation sur  $T * I$ , qui permet d'indiquer si une instance  $x$  a un item  $i$  (noté  $xRi$ ) ou non (1 ou 0).



Ne tenir compte que de la présence des items pas de leur quantité.



## Concepts de base

Exemple :

Base de transactions



TID	Items
1	Pain, Cacahuètes, Lait, Fruits, Confiture
2	Pain, Confiture, Soda, Chips, Lait, Fruit
3	Biscuits, Confiture, Soda, Chips, Pain
4	Confiture, Soda, Cacahuètes, Lait, Fruits
5	Confiture, Soda, Chips, Lait, Pain
6	Fruits, Soda, Chips, Lait
7	Fruits, Soda, Cacahuètes, Lait
8	Fruits, Cacahuètes, Fromage, Yaourt

## Concepts de base

Exemple :

Base formelle :



- **T** : toutes les transactions d'achat
- **I** : tous les produits/articles
- **R** : produit acheté ou non dans la transaction

## Concepts de base

Exemple :

Base formelle



<b>R</b>	<b>Pain</b>	<b>Lait</b>	<b>Fruits</b>	<b>Chips</b>	<b>Biscuits</b>	<b>...</b>
<b>T1</b>	1	1	1	0	0	
<b>T2</b>	1	1	1	1	0	
<b>T3</b>	1	0	0	1	1	
<b>T4</b>	0	1	1	0	0	
<b>T5</b>	1	1	0	1	0	
...						

# Concepts de base

## 2 - Motif

- ✓ = Itemset
- ✓ Un sous ensemble de  $I$ .
- ✓ Une collection d'un ou de plusieurs items.
- ✓ Ex : {Pain}, {Pain, Lait}, {Confiture, Soda, Chips}
- ✓ **k**-motif / **k**-itemset : un motif qui contient k items.
- ✓ Ex : k=3, Motifs de taille **3** : {Pain, Lait, Confiture}, {Lait, Fruits, Soda}.

## Concepts de base

### 3 - Support d'un motif

- Mesure la fréquence d'un motif dans une base.
- Support Count (  $\sigma$  )
  - Fréquence d'apparition d'un motif.
  - Ex :  $\sigma(\{\text{Lait, Pain}\}) = 3$   
 $\sigma(\{\text{Soda, Chips}\}) = 4$
- Support
  - Fraction des transactions contenant un motif.
  - Ex :  $s(\{\text{Lait, Pain}\}) = 3/8$   
 $s(\{\text{Soda, Chips}\}) = 4/8$

TID	Items
1	<b>Pain</b> , Cacahuètes, <b>Lait</b> , Fruits, Confiture
2	<b>Pain</b> , Confiture, Soda, Chips, <b>Lait</b> , Fruit
3	Biscuits, Confiture, Soda, Chips, Pain
4	Confiture, Soda, Cacahuètes, Lait, Fruits
5	Confiture, Soda, Chips, <b>Lait</b> , <b>Pain</b>
6	Fruits, Soda, Chips, Lait
7	Fruits, Soda, Cacahuètes, Lait
<b>8</b>	Fruits, Cacahuètes, Fromage, Yaourt

## Concepts de base

### 4 – Motif fréquent

- Un motif dont le support est  $\geq$  à un seuil **minsup**.
- Sinon, il est dit non fréquent.
- **Propriété** : Si  $m$  est un motif fréquent, alors tout sous-ensemble de  $m$  est également un motif fréquent.
- Ex : {Pain, Lait } est un motif fréquent  $\Rightarrow$  {Pain} et {Lait} sont fréquents.

## Concepts de base

Exemple :

On pose **minsup** = 3

{Cacahuètes, Lait, Confiture}	2	Non Fréquent
{Soda, Chips}	4	Fréquent
{Pain}	4	Fréquent
{Fromage, Yaourt, Fruits}	1	Non Fréquent

TID	Items
1	Pain, Cacahuètes, Lait, Fruits, Confiture
2	Pain, Confiture, Soda, Chips, Lait, Fruit
3	Biscuits, Confiture, Soda, Chips, Pain
4	Confiture, Soda, Cacahuètes, Lait, Fruits
5	Confiture, Soda, Chips, Lait, Pain
6	Fruits, Soda, Chips, Lait
7	Fruits, Soda, Cacahuètes, Lait
8	Fruits, Cacahuètes, Fromage, Yaourt

# Méthodes pour la recherche des modèles fréquents

## **Approche naïve**

- Parcourir l'ensemble de tous les motifs possibles ;
- Calculer le support de chaque motif ;
- Comparer le support au minsup ;
- Ne garder que les motifs fréquents parmi cet ensemble.



# Méthodes pour la recherche des modèles fréquents

## Approche naïve

- Pour  $n$  items dans une base formelle  $\Rightarrow 2^n$  motifs candidats possibles.
- Cas exemple : 10 items  $\Rightarrow 2^{10}$
- En pratique : Base peut avoir plusieurs milliers d'items et plusieurs millions d'instances  $\Rightarrow$  Nombre de motifs trop grand.
- Consommatrice en temps et en ressource. Complexité temporelle.

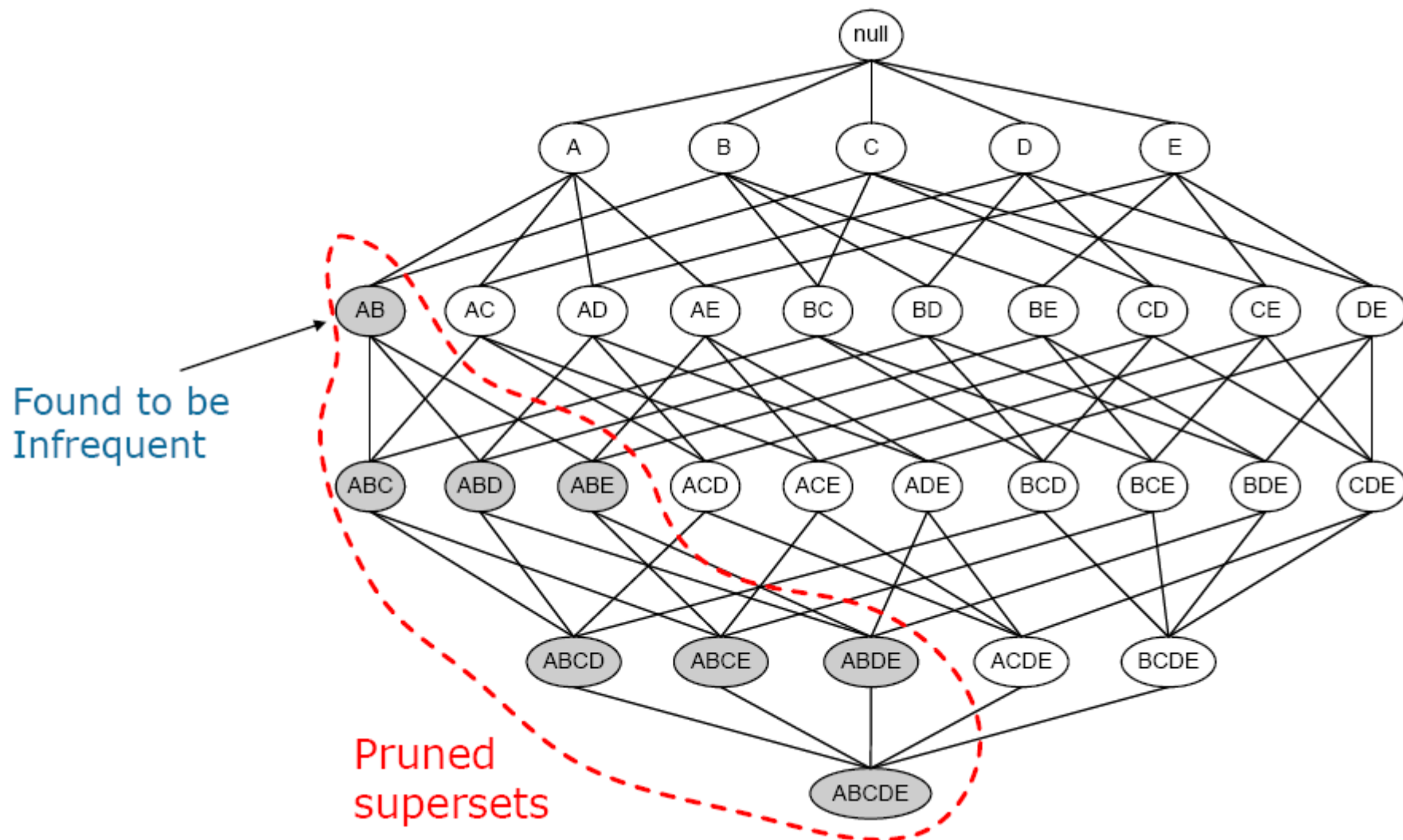
# Méthodes pour la recherche des modèles fréquents

## **Algorithme Apriori**

- Proposé par Agrawal et ses co-auteurs, 1994.
- S'appuie sur les deux principes suivants :
  1. Tout sous-motif d'un motif fréquent est fréquent.
  2. Tout sur-motif d'un motif non fréquent est non fréquent.

# Méthodes pour la recherche des modèles fréquents

## Algorithme Apriori



# Méthodes pour la recherche des modèles fréquents

## Algorithme Apriori

- Effectue l'extraction par niveaux :
  - ✓ Chercher les motifs fréquents de longueur 1;
  - ✓ Combiner ces motifs pour obtenir des motifs de longueur 2 et ne garder que les fréquents parmi eux;
  - ✓ Combiner ces motifs pour obtenir des motifs de longueur 3 et ne garder que les fréquents parmi eux;
  - ✓ Continuer jusqu'à la longueur maximale...

## Méthodes pour la recherche des modèles fréquents

### Algorithme Apriori – Exemple

TID	Items
T1	a, c, d
T2	b, c, e
T3	a, b, c, e
T4	b, e
T5	a, b, c, e
T6	b, c, e

# Méthodes pour la recherche des modèles fréquents

## Algorithme Apriori – Exemple

1- Base Formelle :

TID	Items
T1	a, c, d
T2	b, c, e
T3	a, b, c, e
T4	b, e
T5	a, b, c, e
T6	b, c, e

R	a	b	c	d	e
T1	1	0	1	1	0
T2	0	1	1	0	1
T3	1	1	1	0	1
T4	0	1	0	0	1
T5	1	1	1	0	1
T6	0	1	1	0	1

# Méthodes pour la recherche des modèles fréquents

## Algorithme Apriori – Exemple

2- L'ensemble des motifs fréquents :

**minsup = 2**

TID	Items
T1	a, c, d
T2	b, c, e
T3	a, b, c, e
T4	b, e
T5	a, b, c, e
T6	b, c, e

## Méthodes pour la recherche des modèles fréquents

### Algorithme Apriori – Exemple

2- L'ensemble des motifs fréquents : **minsup = 2**

TID	Items
T1	a, c, d
T2	b, c, e
T3	a, b, c, e
T4	b, e
T5	a, b, c, e
T6	b, c, e

**C<sub>1</sub>**

1-Itemset	Sup Count
{a}	3
{b}	5
{c}	5
{d}	1
{e}	5



## Méthodes pour la recherche des modèles fréquents

### Algorithme Apriori – Exemple

2- L'ensemble des motifs fréquents : **minsup = 2**

TID	Items
T1	a, c, d
T2	b, c, e
T3	a, b, c, e
T4	b, e
T5	a, b, c, e
T6	b, c, e

**C<sub>1</sub>**

1-Itemset	Sup Count
{a}	3
{b}	5
{c}	5
{d}	1
{e}	5

**L<sub>1</sub>** = {{a},{b},{c},{e}}

## Méthodes pour la recherche des modèles fréquents

### Algorithme Apriori – Exemple

2- L'ensemble des motifs fréquents : **minsup = 2**

$$\mathbf{L_1} = \{\{a\}, \{b\}, \{c\}, \{e\}\}$$

**C<sub>2</sub>**

2-Itemset	Sup Count
{a, b}	2
{a, c}	3
{a, e}	2
{b, c}	4
{b, e}	5
{c, e}	4

TID	Items
T1	a, c, d
T2	b, c, e
T3	a, b, c, e
T4	b, e
T5	a, b, c, e
T6	b, c, e

## Méthodes pour la recherche des modèles fréquents

### Algorithme Apriori – Exemple

2- L'ensemble des motifs fréquents : **minsup = 2**

TID	Items
T1	a, c, d
T2	b, c, e
T3	a, b, c, e
T4	b, e
T5	a, b, c, e
T6	b, c, e

**C<sub>2</sub>**

2-Itemset	Sup Count
{a, b}	2
{a, c}	3
{a, e}	2
{b, c}	4
{b, e}	5
{c, e}	4

**L<sub>2</sub>** = { {a, b}, {a, c}, {a, e},  
          {b, c}, {b, e}, {c, e} }

## Méthodes pour la recherche des modèles fréquents

### Algorithme Apriori – Exemple

2- L'ensemble des motifs fréquents : **minsup = 2**

**L<sub>2</sub>** = { {a, b}, {a, c}, {a, e}, {b, c}, {b, e}, {c, e} }

TID	Items
T1	a, c, d
T2	b, c, e
T3	a, b, c, e
T4	b, e
T5	a, b, c, e
T6	b, c, e

**C<sub>3</sub>**

3-Itemset	Sup Count
{a, b, c}	2
{a, b, e}	2
{a, c, e}	2
{b, c, e}	4

## Méthodes pour la recherche des modèles fréquents

### Algorithme Apriori – Exemple

2- L'ensemble des motifs fréquents : **minsup = 2**

TID	Items
T1	a, c, d
T2	b, c, e
T3	a, b, c, e
T4	b, e
T5	a, b, c, e
T6	b, c, e

**C<sub>3</sub>**

3-Itemset	Sup Count
{a, b, c}	2
{a, b, e}	2
{a, c, e}	2
{b, c, e}	4

**L<sub>3</sub>** = {{a, b, c}, {a, b, e},  
          {a, c, e}{b, c, e}}

## Méthodes pour la recherche des modèles fréquents

### Algorithme Apriori – Exemple

2- L'ensemble des motifs fréquents : **minsup = 2**

$$\mathbf{L_3} = \{\{a, b, c\}, \{a, b, e\}, \{a, c, e\}, \{b, c, e\}\}$$

TID	Items
T1	a, c, d
T2	b, c, e
T3	a, b, c, e
T4	b, e
T5	a, b, c, e
T6	b, c, e

**C4**

4-Itemset	Sup Count
{a, b, c, e}	2

$$\mathbf{L_4} = \{\{a, b, c, e\}\}$$

## Méthodes pour la recherche des modèles fréquents

### Algorithme Apriori – Exemple

2- L'ensemble des motifs fréquents : **minsup = 2**

$$\mathbf{L_4} = \{\{a, b, c, e\}\}$$

**C<sub>5</sub>**

5-Itemset	Sup Count
$\phi$	

TID	Items
T1	a, c, d
T2	b, c, e
T3	a, b, c, e
T4	b, e
T5	a, b, c, e
T6	b, c, e

## Méthodes pour la recherche des modèles fréquents

### Algorithme Apriori – Exemple

2- L'ensemble des motifs fréquents : **minsup = 2**

**$L_1 \cup L_2 \cup L_3 \cup L_4$**

TID	Items
T1	a, c, d
T2	b, c, e
T3	a, b, c, e
T4	b, e
T5	a, b, c, e
T6	b, c, e

$$L_1 = \{\{a\}, \{b\}, \{c\}, \{e\}\}$$

$$L_2 = \{\{a, b\}, \{a, c\}, \{a, e\}, \{b, c\}, \{b, e\}, \{c, e\}\}$$

$$L_3 = \{\{a, b, c\}, \{a, b, e\}, \{a, c, e\}, \{b, c, e\}\}$$

$$L_4 = \{\{a, b, c, e\}\}$$



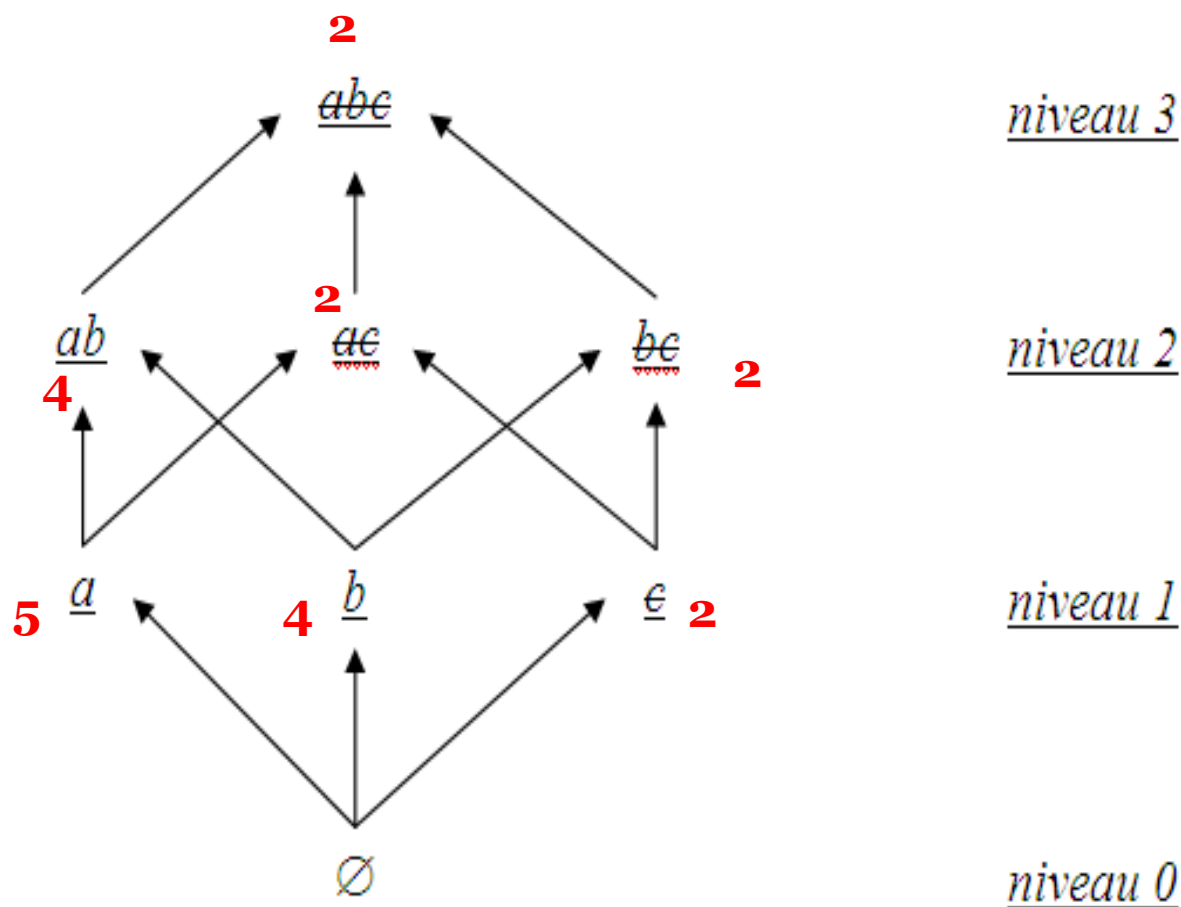
# Méthodes pour la recherche des modèles fréquents

## Algorithme Apriori

Remarque 1 - Treillis

Exemple :  $I = \{a, b, c\}$

minsup = 3



# Méthodes pour la recherche des modèles fréquents

## Algorithme Apriori

### Remarque 2

- ✓ **Elagage** du parcours du treillis.
- ✓ Un des objectifs des optimisations de cet algorithme est de diminuer le nombre d'accès à la base de données.
- ✓ Le seuil minsup est fixé par l'analyste.
- ✓ Celui-ci peut suivre une approche itérative en fixant un seuil au départ et, en fonction du résultat, changera la valeur du seuil

# Méthodes pour la recherche des modèles fréquents

## **Algorithme Apriori**

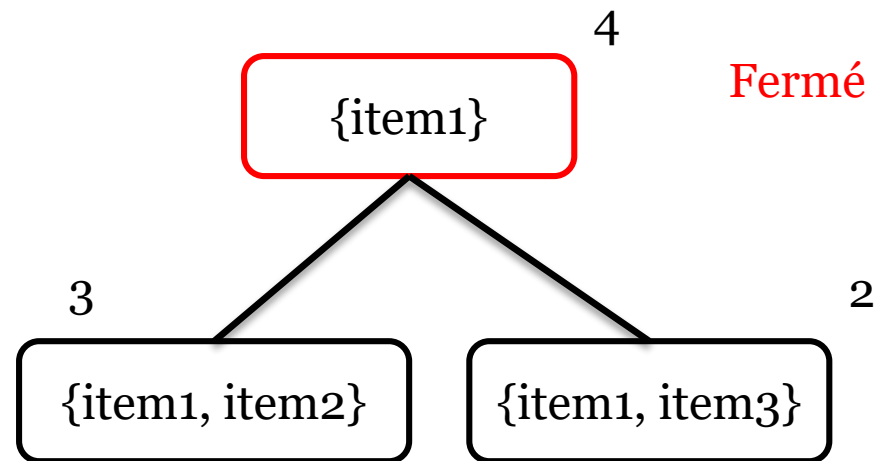
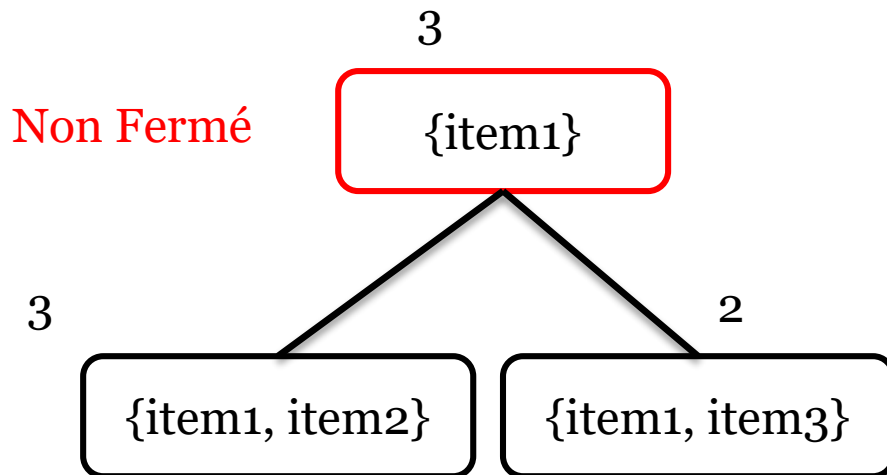
### Remarque 2

- ✓ Si trop de motifs fréquents ont été trouvés, il augmentera le seuil ; dans le cas inverse, il le diminuera.
- ✓ Le temps de calcul de l'algorithme Apriori décroît avec le seuil.
- ✓ Si l'analyste fixe une valeur de seuil trop grande, cela gaspillera moins de temps que s'il en fixe un trop petit.
- ✓ Optimisation Apriori : AprioriTID, apriori partitionné, comptage dynamique, etc.

## Types de motifs fréquents

### Motif Fréquent **Fermé** :

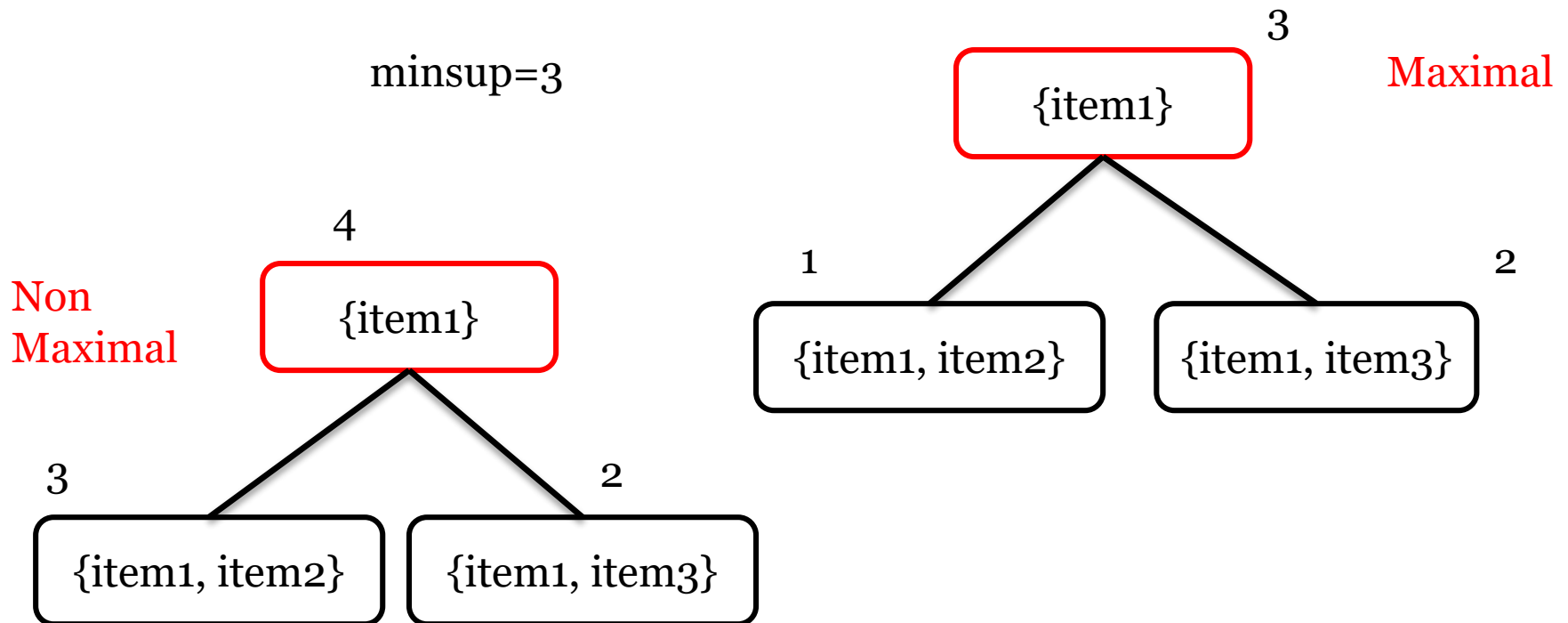
- Motif fréquent dont aucun de ses sur-motifs immédiats n'a un support identique.



## Types de motifs fréquents

### Motif Fréquent **Maximal** :

- Motif fréquent dont aucun de ses sur-motifs immédiats n'est fréquent.



## Types de motifs fréquents

### Motif Fréquent **Fermé** :

- Motif fréquent dont aucun de ses sur-motifs immédiats n'a un support identique.

### Motif Fréquent **Maximal** :

- Motif fréquent dont aucun de ses sur-motifs immédiats n'est fréquent.

Les motifs F. maximaux  $\subset$  Les motifs F. fermés  $\subset$  Les motifs fréquents

## Types de motifs fréquents

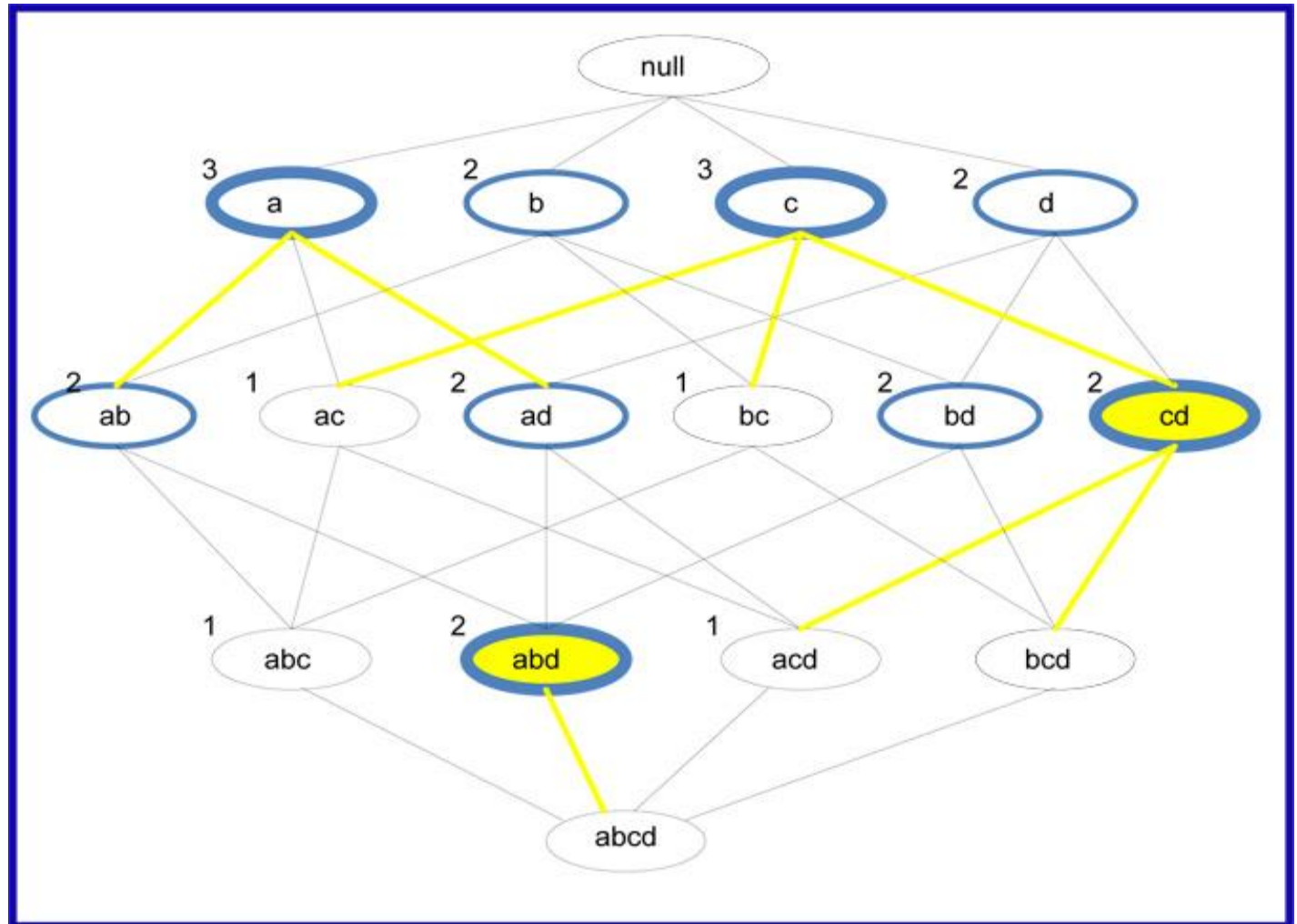
Exemple :

$\text{minsup} = 2$

Fréquent

Fermé

Maximal

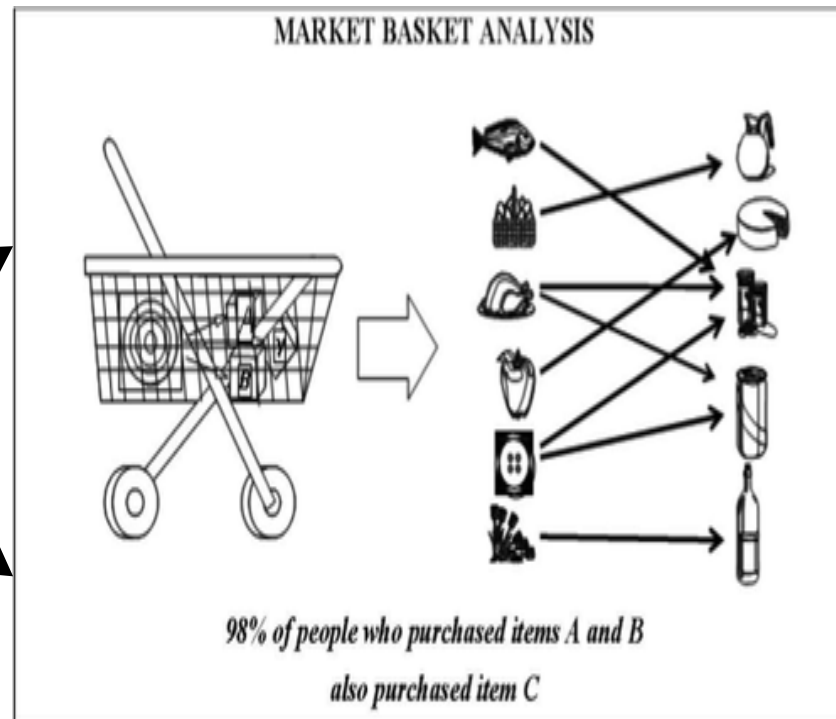


Contexte

## SAVOIR – DECOUVRIR - DECIDER



Données



Algorithmes



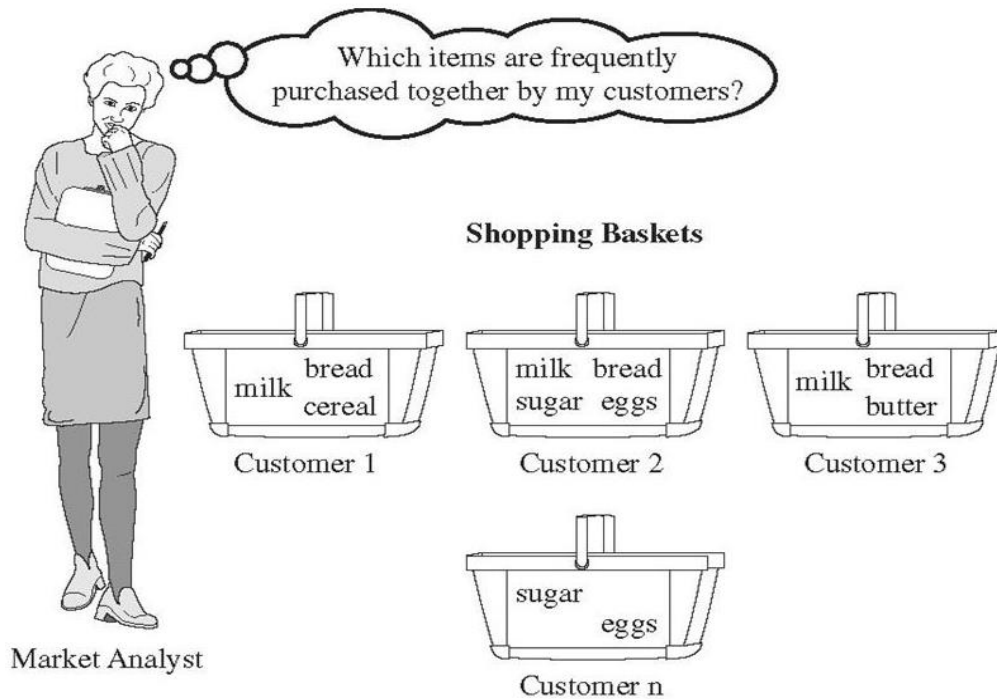
**Motifs  
Fréquents**



# Passage aux règles d'association

## Analyse du panier de marché

Quels sont les produits qui sont achetés fréquemment et simultanément ?



Transcrire la  
connaissance  
sous forme de  
**règle  
d'association.**

## Passage aux règles d'association

- La découverte des règles d'association : Phase qui suit la phase de recherche des motifs fréquents.
- Trouver toutes les règles qui existent entre les motifs fréquents.
- Les règles ont la forme suivante :

Si    antécédent    alors    conséquent

Ex :      Si    Lait et Beurre    alors    Pain    /

{Lait, Beurre} => {Pain}

## Passage aux règles d'association

- Soit la règle d'association suivante :

Si X alors Y /  $\mathbf{X} \Rightarrow \mathbf{Y}$  (ou X et Y des motifs fréquents)

- Mesures d'évaluation d'une règle :

1. Support

- 2. Confiance**

- Support : un indicateur de fiabilité de la règle.
- Confiance : un indicateur de précision de la règle.

## Passage aux règles d'association

- Soit la règle d'association suivante :

Si  $X$  alors  $Y$  /  $X \Rightarrow Y$  (ou  $X$  et  $Y$  des motifs fréquents)

- Confidence: Mesure à quelle fréquence les items de  $Y$  apparaissent dans les transactions qui contiennent  $X$ .

- **Confiance ( $X \Rightarrow Y$ )** = 
$$\frac{\text{Nombre de transactions contenant } (X \cup Y)}{\text{Nombre de transactions contenant } X}$$

$$\text{Confiance}(\{Pain\} \Rightarrow \{Lait\}) = \frac{\# \text{ de transactions contenant } \{Pain, Lait\}}{\# \text{ de transactions contenant } \{Pain\}}$$

## Passage aux règles d'association

### Apriori

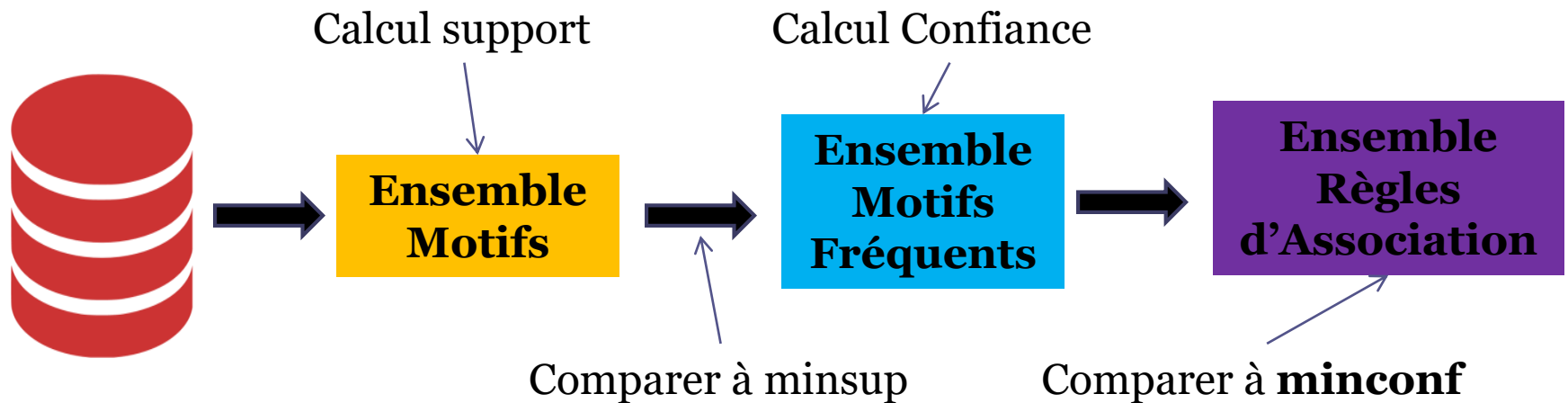
- L'ensemble des règles d'association peut être trouvé en calculant le support et la confiance de toutes les combinaisons possibles des motifs fréquents;
- Puis, prendre celle dont la confiance est importante.
- Des seuils peuvent être fixés : ***minsup*** et ***minconf***.
- Les règles d'association qui dépassent un minimum de support **et** minimum de confiance sont appelées règles **solides/fortes**.

## Passage aux règles d'association

### Génération des règles d'association :

Démarche en deux étapes

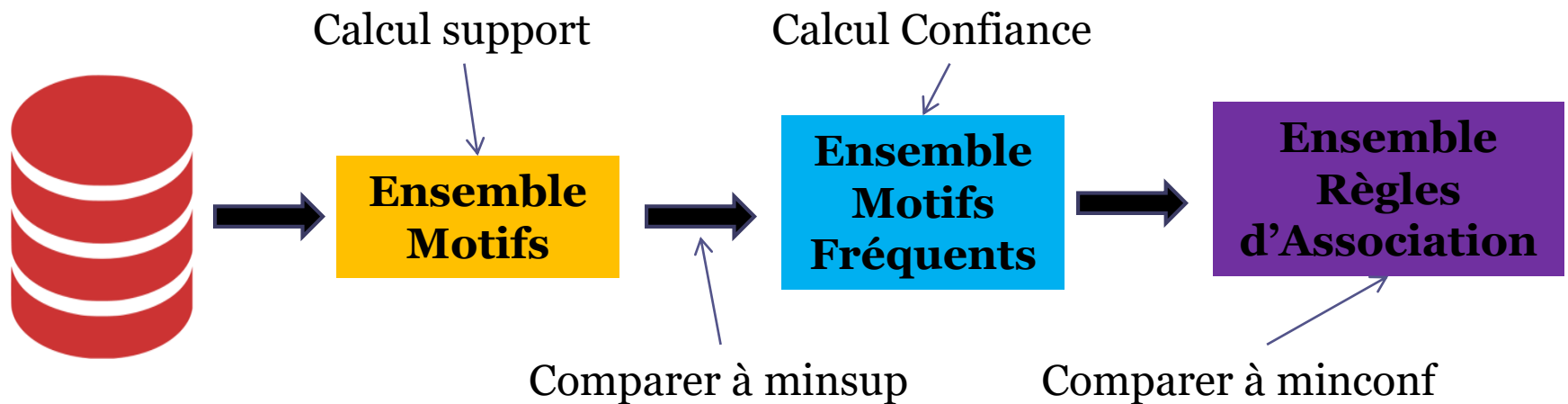
1. recherche des itemsets fréquents ( $\text{support} \geq \text{minsup}$ )
2. à partir des itemsets fréquents, extraire les règles ( $\text{sup} \geq \text{minsup}$  et  $\text{conf} \geq \text{minconf}$ )



## Passage aux règles d'association

### Génération des règles d'association :

- Pour chaque motif fréquent ***m***, générer tous les sous ensembles non vide de ***m*** ;
- Pour chaque sous ensemble ***s*** non vide, enregistrer la règle (***s*** => ***m-s***) si :  
***Support(s => m-s) >= minsup***  
et  
***Confiance (s => m-s) >= minconf***



## Exemple - Démo

Base de transactions

On pose **minsup** = 33,34%, **minconf** = 60%



TID	Items
T1	Xbox, Casque, Smartwatch
T2	Xbox, Casque
T3	Xbox, Tablette, SDCard
T4	SDCard, Tablette
T5	SDCard, Smartwatch
T6	Xbox, Tablette, SDCard



## Exemple - Démo

On pose **minsup** = 2, **minconf** = 60%

1 – Extraire les motifs fréquents

**L1 U L2 U L3**

**L1** = {{Xbox},{Casque},{Smartwatch},{Tablette}}

**L2** = {{Xbox, Casque},  
          {Xbox, Tablette},  
          {Xbox, SDCard},  
          {Tablette, SDCard} }

**L3** = {{Xbox, Tablette, SDCard}}

## Exemple - Démo

On pose **minsup** = 33.34%, **minconf** = 60%

### 2 – Générer les règles d'association

{Xbox, Casque}

$$\left\{ \begin{array}{l} \{Xbox\} \Rightarrow \{Casque\} - \text{sup} = 2/6=0.33 \quad \text{et} \quad \text{conf} = 2/4 = 0.5 \\ \{Casque\} \Rightarrow \{Xbox\} - \text{sup} = 2/6=0.33 \quad \text{et} \quad \text{conf} = 2/2 = 1 \end{array} \right.$$

{Xbox, Tablette}

$$\left\{ \begin{array}{l} \{Xbox\} \Rightarrow \{Tablette\} - \text{sup} = 0.33 \quad \text{et} \quad \text{conf} = 0.5 \\ \{Tablette\} \Rightarrow \{Xbox\} - \text{sup} = 2/6=0.33 \quad \text{et} \quad \text{conf} = 2/3=0.66 \end{array} \right.$$

## Exemple - Démo

On pose **minsup** = 33.34%, **minconf** = 60%

### 2 – Générer les règles d'association

{Xbox, SDCard}

$$\left\{ \begin{array}{ll} \{Xbox\} \Rightarrow \{SDCard\} & - \text{ sup} = 0.33 \quad \text{et} \quad \text{conf} = 0.5 \\ \{SDCard\} \Rightarrow \{Xbox\} & - \text{ sup} = 2/6=0.33 \quad \text{et} \quad \text{conf} = 2/4=0.5 \end{array} \right.$$

{Tablette, SDCard}

$$\left\{ \begin{array}{ll} \{Tablette\} \Rightarrow \{SDCard\} & - \text{ sup} = 3/6=0.5 \quad \text{et} \quad \text{conf} = 3/3=1 \\ \{SDCard\} \Rightarrow \{Tablette\} & - \text{ sup} = 3/6=0.5 \quad \text{et} \quad \text{conf} = 3/4=0.75 \end{array} \right.$$

## Exemple - Démo

On pose **minsup** = 33.34%, **minconf** = 60%

### 2 – Générer les règles d'association

{Xbox, Tablette, SDCard}

{Xbox} => {Tablette, SDCard} - 2/6=0.33	conf = 2/4=0.5
<b>{Tablette} =&gt; {SDCard, Xbox} - 2/6=0.33</b>	<b>conf = 2/3=0.66</b>
{SDCard} => {Tablette, Xbox} - 2/6=0.33	conf = 2/4=0.5
<b>{Xbox, Tablette} =&gt; {SDCard} - 2/6=0.33</b>	<b>conf = 2/2=1</b>
<b>{Xbox, SDCard} =&gt; {Tablette} - 2/6=0.33</b>	<b>conf = 2/2=1</b>
<b>{Tablette, SDCard} =&gt; {Xbox} - 2/6=0.33</b>	<b>conf = 2/3=0.66</b>

## Exemple - Démo

On pose **minsup** = 33.34%, **minconf** = 60%

### 2 – Générer les règles d'association

Règles solides

[ {Casque} => {Xbox} - sup = 2/6=0.33 et **conf = 2/2 = 1**  
{Tablette} => {Xbox} - sup = 2/6=0.33 et **conf = 2/3=0.66**  
{Tablette} => {SDCard}  
{SDCard} => {Tablette}  
{Tablette} => {SDCard, Xbox} - 2/6=0.33 **conf = 2/3=0.66**  
{Xbox, Tablette} => {SDCard} - 2/6=0.33 **conf = 2/2=1**  
{Xbox , SDCard} => {Tablette} - 2/6=0.33 **conf = 2/2=1**  
{Tablette, SDCard} => {Xbox} - 2/6=0.33 **conf = 2/3=0.66**

## Motifs Rares

- Les motifs rares représentent les motifs qui apparaissent rarement dans un ensemble de données.
- Rareté ~ exceptions
- Découverte intéressante pour certains domaines : médecine, biologie, etc.
- Découverte de symptômes non usuels ou effets indésirables exceptionnels.
- Ex : En pharmacovigilance (i.e. Partie de la pharmacologie dédiée à la détection et l'étude des effets indésirables des médicaments.)
- L'extraction des motifs rare ~ Associer des médicaments avec des effets indésirables ~ trouver des cas où un médicament avait des effets mortels ou indésirables sur les patients.

## Motifs Rares

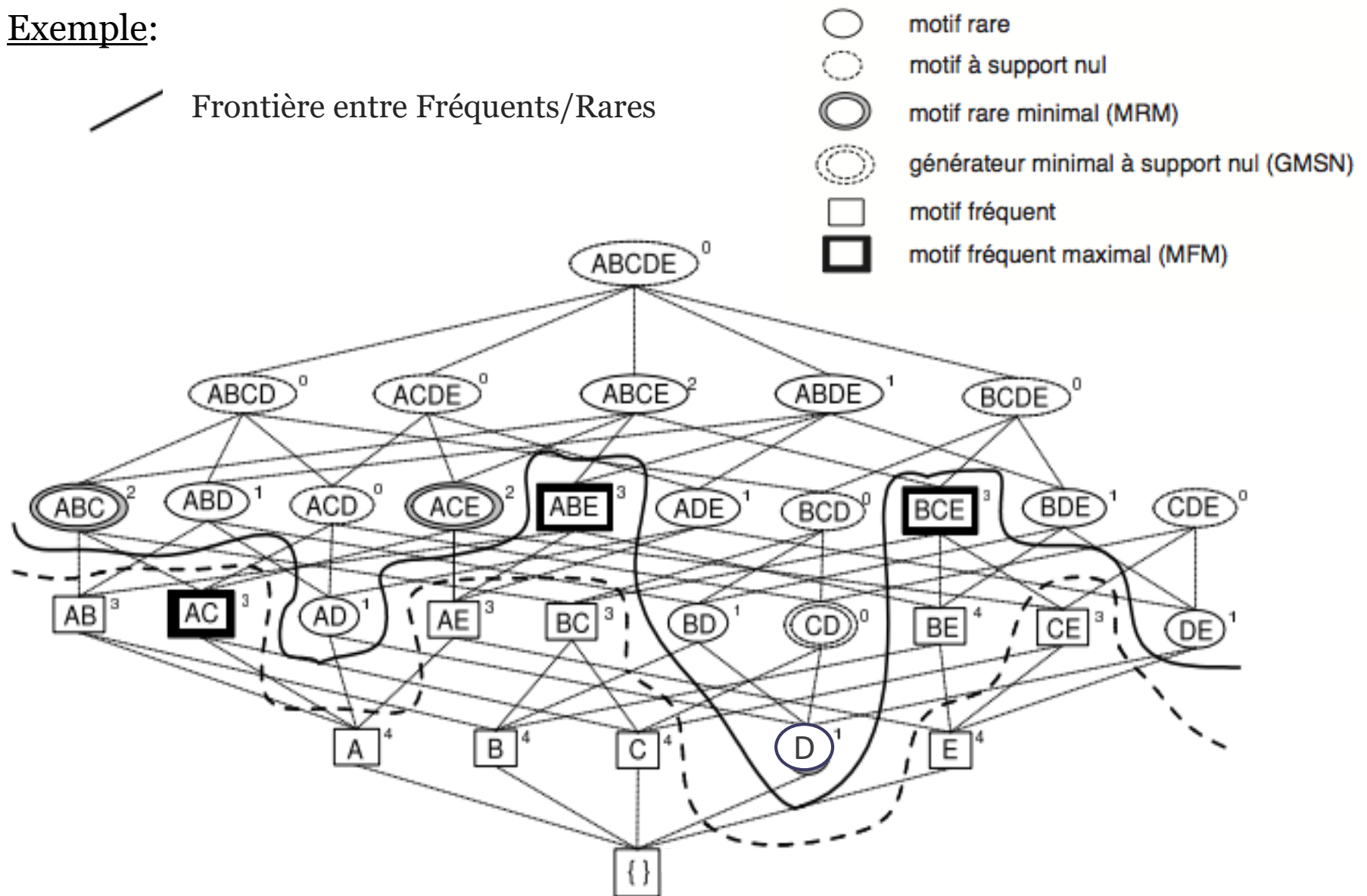
### Concepts :

- Un motif est dit rare s'il n'est pas fréquent.
- Son support est inférieur ou égal à un support maximum  $\text{maxsup}$ , avec  $\text{maxsup} = \text{minsup} - 1$ .
- Motif rare minimal (**MRM**): un motif rare dont tous ses sous-motifs sont fréquents.
- Motif fréquent maximal (MFM): un motif fréquent dont tous ses sur-motifs sont rares.
- Générateur minimal à support nul (**GMSN**) : un motif à support nul, ainsi tous ses sur-motifs sont des motifs à support nul, et si tous ses sous-motifs sont des motifs à support non nul.

# Motifs Rares

## Exemple:

Frontière entre Fréquents/Rares





## Motifs fréquents séquentiels

- Prise en compte de la dimension temporelle.
- Intégrer les contraintes temporelles (succession) dans la recherche des motifs.
- Extraire des enchainements d'ensembles d'items, couramment associés sur une période de temps bien spécifiée.
- Exemple : Un client achète un laptop, puis une souris, puis casque audio dans les jours qui suivent.
- Marketing, finance, détection des symptômes précédant une maladie, etc.

## Motifs fréquents séquentiels

- **Séquence** : une liste ordonnée par dates croissantes, non vide, d'items, notée :

$$s = \langle m_1 \ m_2 \ \dots \ m_n \rangle, \text{ où } m \text{ est un motif.}$$

- Ex :  $\langle a \ (ab) \ (ac) \ d \ (cf) \rangle$
- Une séquence : une suite de transactions qui apporte une relation d'ordre entre les achats d'un client.
- Un item peut apparaître au plus une fois dans un motif, mais plusieurs fois dans une séquence.
- Une séquence de taille  $l$  :  **$l$ -séquence**.
- **Base de données séquentielle** est composée d'éléments ordonnés, des tuples  $\langle \text{SID}, s \rangle$ , où SID : ID séquence, et  $s$  : séquence.

## Motifs fréquents séquentiels

### Sous-séquence :

Soit  $s_1 = \langle a_1 a_2 \dots a_n \rangle$  et  $s_2 = \langle b_1 b_2 \dots b_m \rangle$  deux séquences de données.  $s_1$  est une sous séquence de  $s_2$  si et seulement si il existe  $i_1 < i_2 < \dots < i_n$  des entiers tels que  $a_1 \subset b_{i_1}, a_2 \subset b_{i_2}, \dots, a_n \subset b_{i_n}$ .

### Exemples :

-  $s_1 = \langle (C)(DE)(H) \rangle$  sous-séquence de  $s_2 = \langle (G)(CH)(I)(DEF)(H) \rangle$   
car :

- $(C) \subset (CH)$ ;
- $(DE) \subset (DEF)$  ;
- $(H) \subset (H)$ .

-  $s_1 = \langle (C)(E) \rangle$  n'est pas une sous-séquence de  $s_2 = \langle (CE) \rangle$

## Motifs fréquents séquentiels

- Un client supporte une séquence  $\mathbf{s}$  (fait partie du support pour  $s$ ) si  $\mathbf{s}$  est une sous séquence de la séquence de données de ce client.
- Le support d'une séquence  $\mathbf{s}$  est calculé comme étant le pourcentage des clients qui supportent  $\mathbf{s}$ .
- Une séquence dont le support est supérieur ou = à  $\text{minsup}$  est une **séquence fréquente**, appelée *Sequential Pattern*.
- Si une séquence  $\mathbf{s}$  n'est pas fréquente, aucune de ses sur-séquences n'est fréquente.
- Si une séquence  $\mathbf{s}$  est fréquente, alors toutes ses sous-séquences le sont aussi.
- Une séquence est maximale si elle n'est une sous-séquence d'aucune autre séquence.

## Motifs fréquents séquentiels

### Exemple:

SID	Sequence
1	<a( <u>ab</u> c)(a <u>c</u> )d(cf)>
2	<(ad)c(bc)(ae)>
3	<(ef)( <u>ab</u> )(df) <u>c</u> b>
4	<eg(af)c <b>b</b> c>

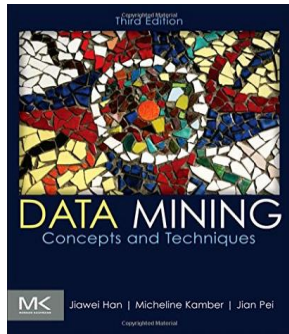
- Longueur de la séquence 1 : 9
- La séquence 1 contient plusieurs fois 'a', elle ne contribuera qu'une seule fois au support de <a>.
- La séquence <a(bc)df> est une sous-séquence de la séquence 1.
- Support (<(ab)c>) est égal à 2 (Présent dans 1 et 3).

# Motifs fréquents séquentiels

## Algorithme GSP :

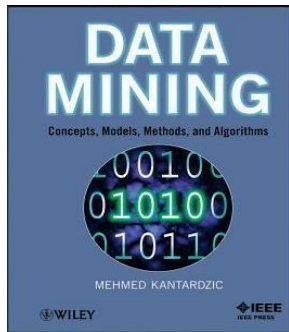
- Generalized Sequential Patterns.
- Proposition pour la recherche des motifs séquentiels fréquents.
- Basé sur l'algorithme Apriori.
- Suit les étapes suivantes :

# Ressources



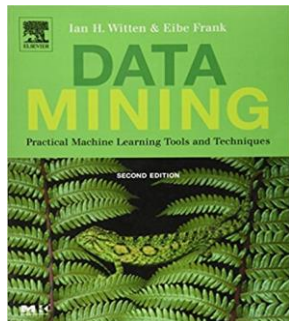
## **Data Mining : concepts and techniques, 3rd Edition**

- ✓ Auteur : Jiawei Han, Micheline Kamber, Jian Pei
- ✓ Éditeur : Morgan Kaufmann Publishers
- ✓ Edition : Juin 2011 - 744 pages - ISBN 9780123814807



## **Data Mining : concepts, models, methods, and algorithms**

- ✓ Auteur : Mehmed Kantardzi
- ✓ Éditeur : John Wiley & Sons
- ✓ Edition : Aout 2011 – 552 pages - ISBN : 9781118029121



## **Data Mining: Practical Machine Learning Tools and Techniques**

- ✓ Auteur : Ian H. Witten & Eibe Frank
- ✓ Éditeur : Morgan Kaufmann Publishers
- ✓ Edition : Juin 2005 - 664 pages - ISBN : 0-12-088407-0

# Ressources

Cours – Abdelhamid DJEFFAL – Fouille de données avancée

✓ [www.abdelhamid-djeffal.net](http://www.abdelhamid-djeffal.net)

WekaMOOC – Ian Witten – Data Mining with Weka

✓ <https://www.youtube.com/user/WekaMOOC/featured>

Cours - Laboratoire ERIC Lyon - DATA MINING et DATA SCIENCE

✓ [https://eric.univ-lyon2.fr/~ricco/cours/supports\\_data\\_mining.html](https://eric.univ-lyon2.fr/~ricco/cours/supports_data_mining.html)

Gregory Piatetsky-Shapiro - KDNuggets

✓ <http://www.kdnuggets.com/>