BI = Business Intelligence
Master Data-Science
Cours 6 - Data Mining

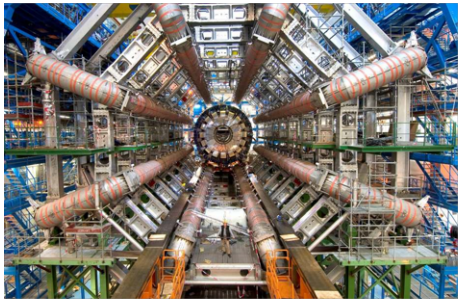
Ludovic DENOYER - ludovic.denoyer@lip6.fr
D'après Elisa Fromont

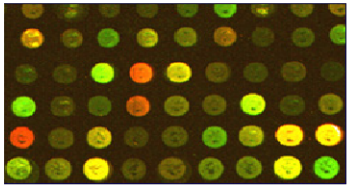
UPMC

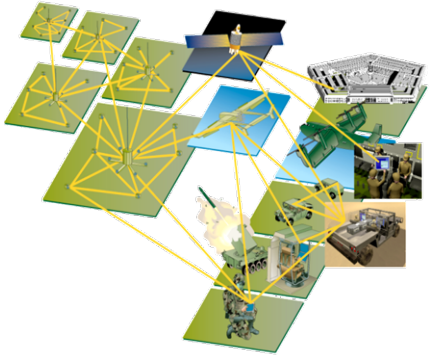
23 mars 2015

Le Data Mining

- De plus en plus de données







Le Data Mining

$\Rightarrow$ Impossibilité d'analyser les données manuellement $\Rightarrow$ Besoin de développer des outils/méthodes d'analyse de données

Data Mining

the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets
(Hand, Mannila, Smyth)

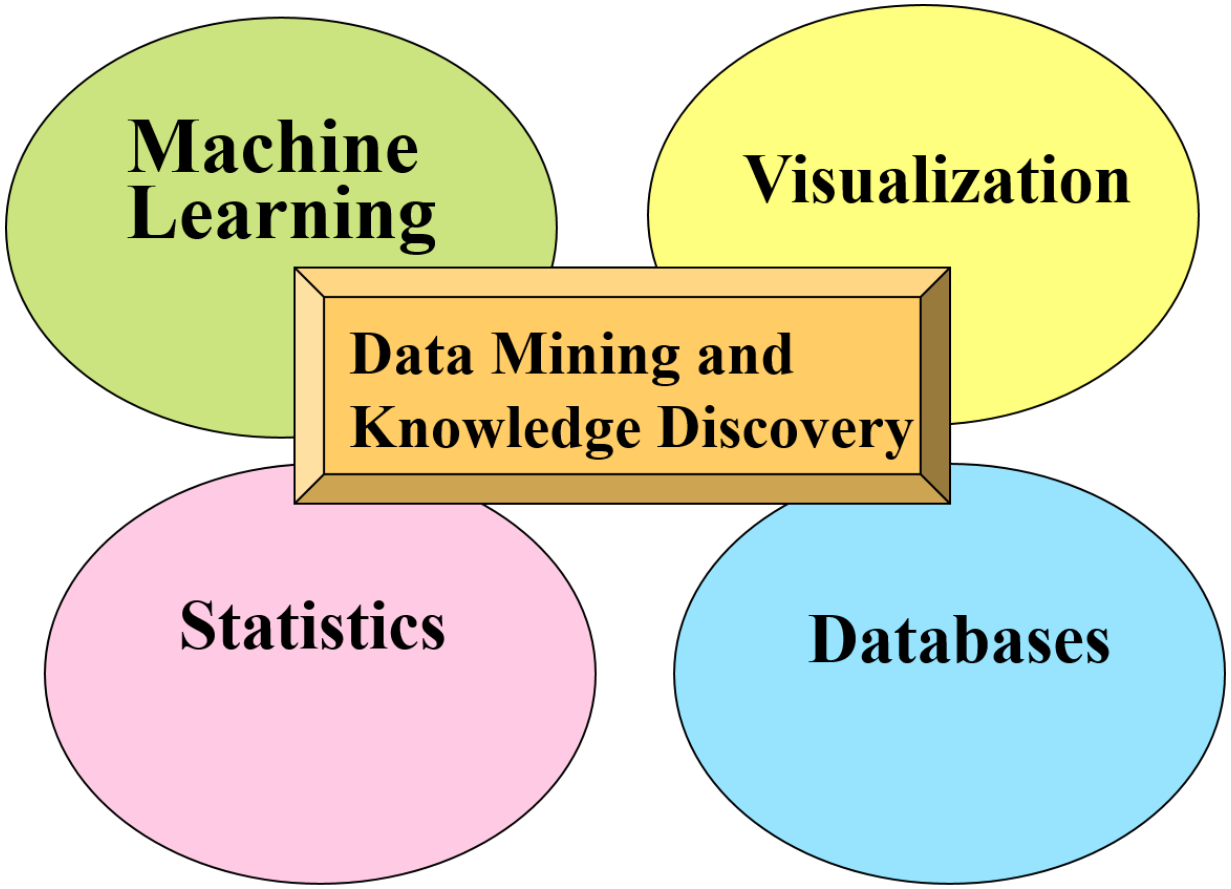
Le Data Mining

```
1110000001000000000000110000100000110000000000000010000000
0001110000011101110111000001100111000001011000001000000000
1100000001000000010000110010000000110000000000000010000100
1110000001000000010000110000000000010100000000000010000000
1110000001000000000000110000000000110000000000000010000000
0101110000001101110111001001100111000111111000001000000000
0001110000001101010111000001100111010111111000001000000000
1110000001000000000000110000100000110010000000000010100000
1100000001000000000000110000000000110000000000000010000000
1111001011000010000000111110001000110000000110001100000000
0001110000001101110111000001100111000111111000001000000000
0001110110111101111110000011111100011111110011001100000
0000001000000001100000011111000000000000001011000010000000
0000001110111100110000000000001100001100000000111001100000
0000001110111100111000000000001100000100000000111001100000
0001000010111100111000010000011000001000100001110011000000
0000000110101100101101100000001100010100000000111001100000
0000001000000010000101011110000000000000000011000110000000
0000001000000010000000111110000000000000000011000110000000
0000001000100010000000111110000000000000000010000111000000
0010001000000010000000101110000000000000000011000110000000
0001110000001101110111000001100111000111111000001000000000
0001110000001101110111000001100111000111111000001000010000
0000001000000010000000111110000000000000000011000110000000
```

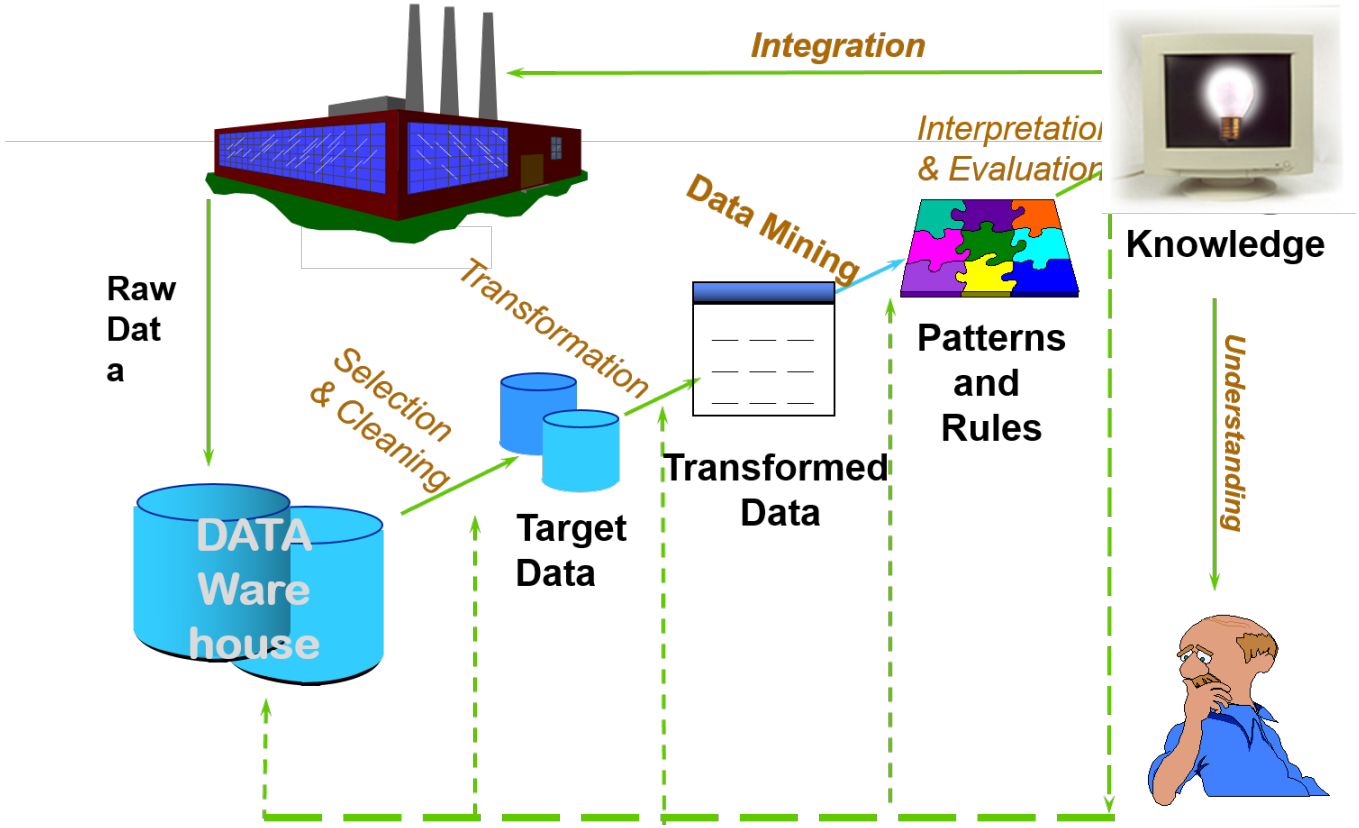
Le Data Mining

```
111101111000000000000000010000000001000000000000000000
111111111000000000000000000000000000000000000000000000
111111111000000000000010001000000000000000000000100000
111011111000000000000000000000000000000000000000000000
111111111111111111000000010000000000000000000010010000
000000111111111111000000000000000000000000000000000000
000000111110111110001000000000000010000000000000000000
000000101111111110000000000101000000000000000000000000
000000111111111110000000000000000000000000000000000000
000000111111110111000000000000000000000000000001001000
000100110111111110000000000000000000000000000000000000
000000000000000001111111111111111111111111000000000000
000000000000000001111111111111111111111111000000000010
000000000000000011011111001111111111111111000000000100
001000000100000001111111111111111111111111000000000000
000001000000000001111111111111111111101111000000000000
000000000000000001111111011111111111111111000000000000
000000000000000001111011111111111111111111000000000000
000000000000000001111111111111111011111111111111111100
000001000000100000000000000000000001111101111111111100
000000000000100000000000000000000001111111111111111100
000000001000000000010000100000000001111111110111111100
000010010000000000000000000010100000111101111111101100
000000000000010000000000000000000000111111101111111100
```

Le Data Mining



Le Data Mining

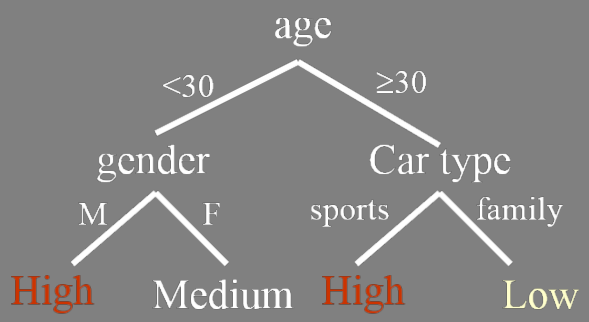


Le Data Mining

- **Trois tâches principales :**
  - Classification
  - Clustering
  - Pattern Mining
- Certaines tâches/problèmes sont **prédictifs**, d'autres sont **explicatifs**

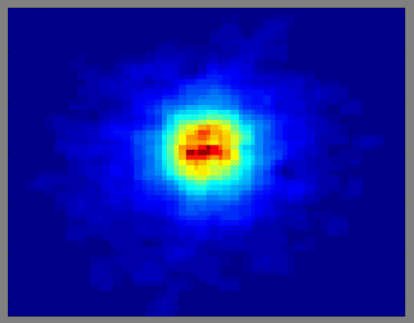
Classification

Example:



Classification

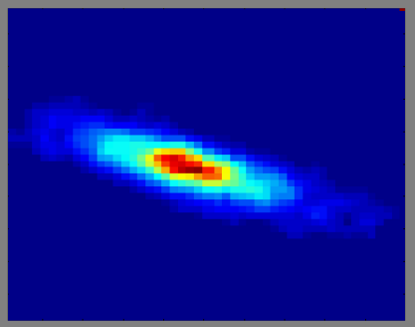
Early



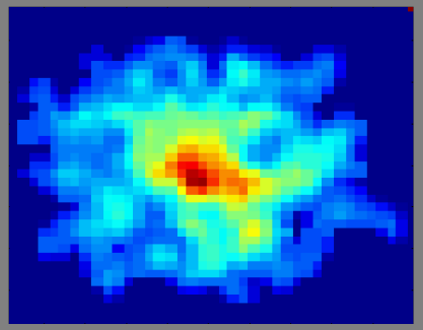
Class:
- **Phase**

Attributes:
- **image *features*,**
- **wavelengths**

Intermediate



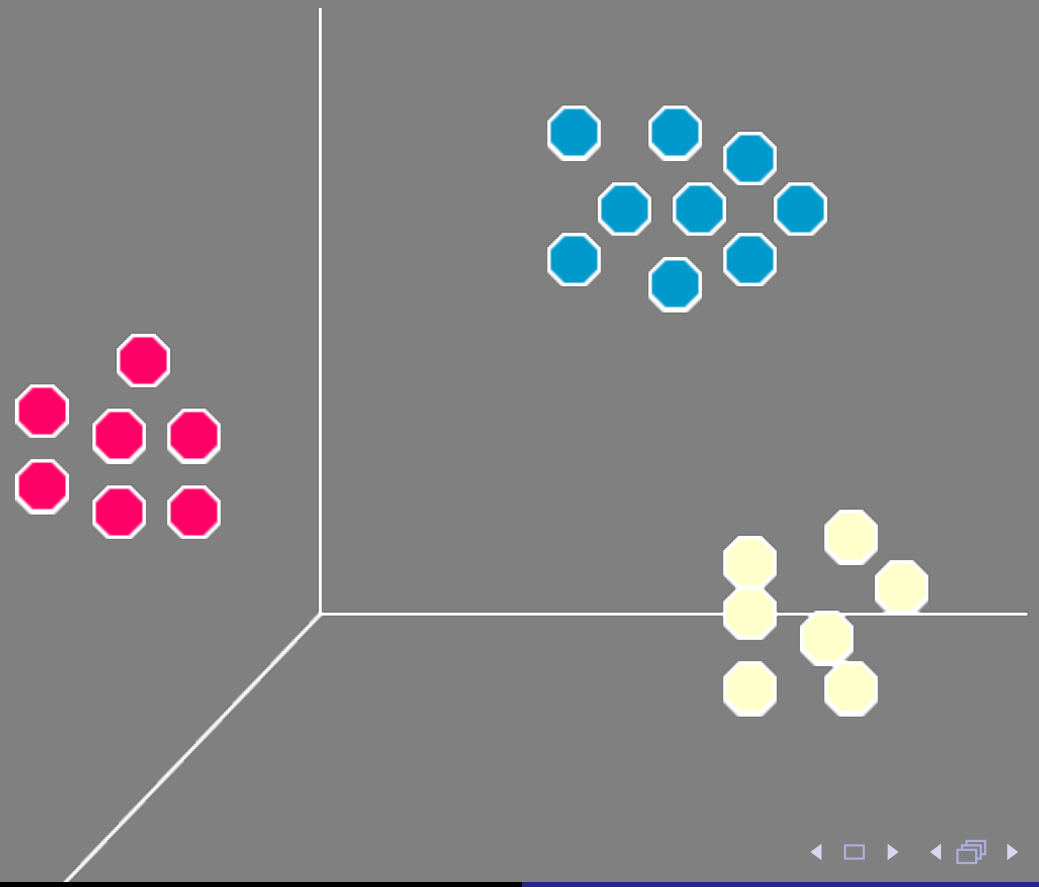
Late



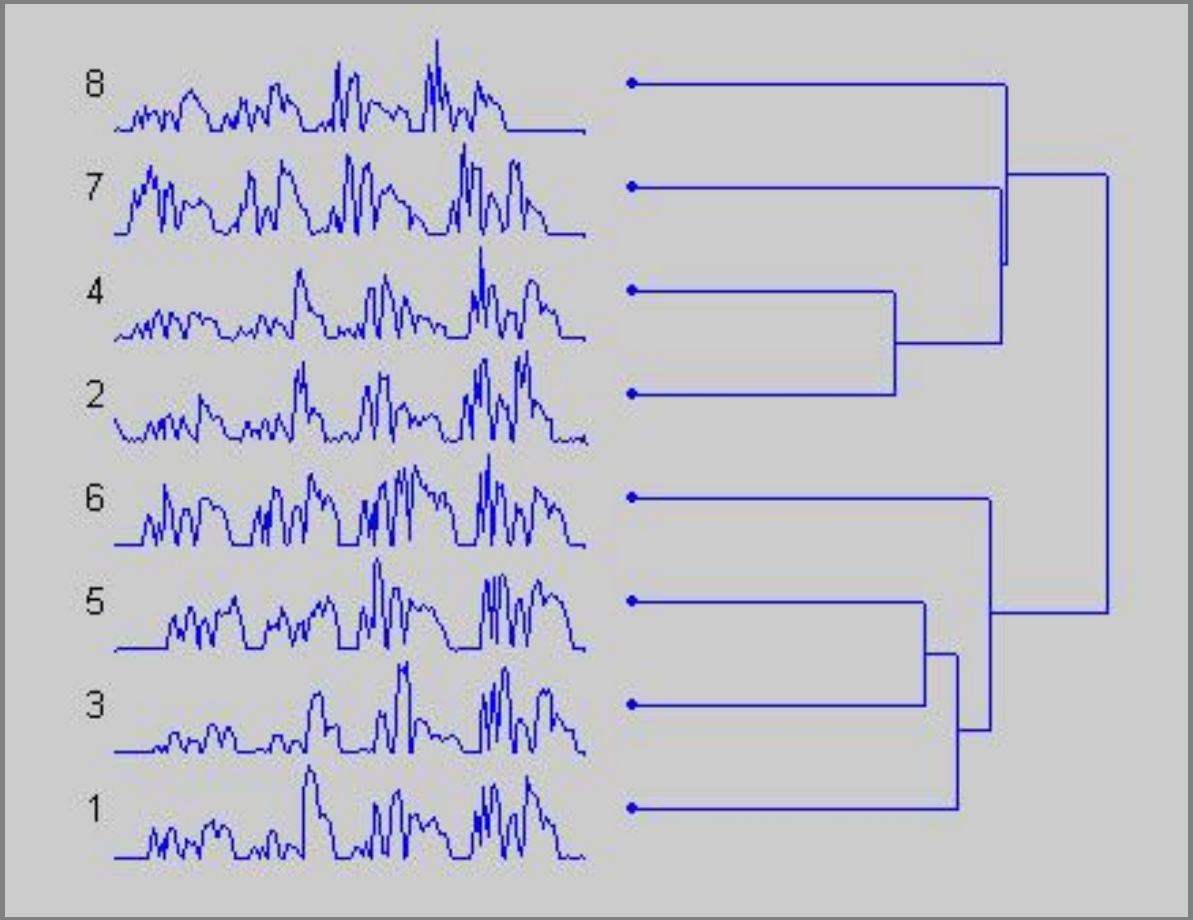
Dataset size:
- **72 million stars, 20 million galaxies**
- **Object catalog: 9 GB**
- **Image database: 150 GB**

Clustering

- Regrouper automatiquement les données qui se ressemblent
- Créer des groupes (clusters) de données, et des résumés associés à ces groupes



Clustering



Clustering

bush | Search | Advanced Search | Preferences

George W. **Bush** - Wikipedia, the free encyclopedia
Open-source encyclopedia article provides personal, business and political information about the President, his policies, and public perceptions and ...
en.wikipedia.org/wiki/George_W._**Bush** - 459k - Cached - Similar pages - Note this

Bush (band) - Wikipedia, the free encyclopedia
Bush was a post-grunge band from the UK, formed in 1992. Their debut album was the self-released Sixteen Stone in 1994. They have sold well over 10 million ...
en.wikipedia.org/wiki/**Bush**_(band) - 60k - Cached - Similar pages - Note this
More results from en.wikipedia.org »

President of the United States - George W. **Bush**
The Oval Office contains speeches and statements of President **Bush**, a description of policy priorities, biographies, and photo essays.
www.whitehouse.gov/president/ - 21k - Cached - Similar pages - Note this
More results from www.whitehouse.gov »

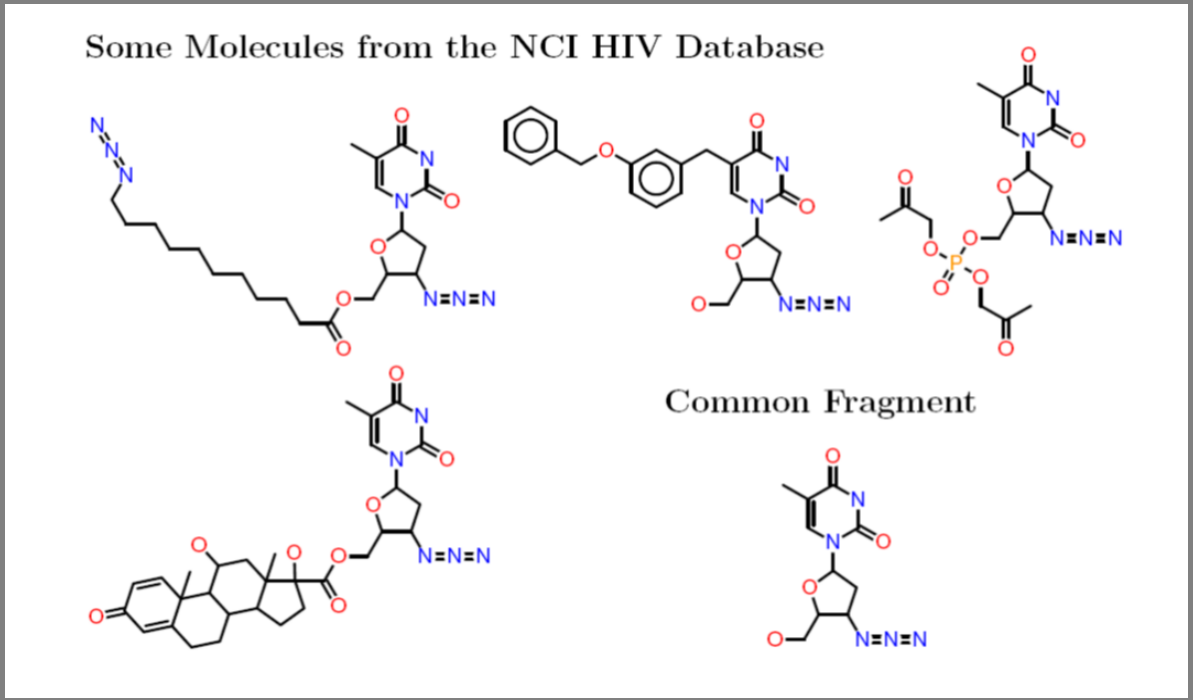
Gavin Rossdale: gavinrossdalefans.com
The former lead singer of **BUSH**, the platinum selling alt rock juggernaut, Gavin can now be seen UP CLOSE at this intimate Past Show. ...
gavinrossdalefans.com/ - 38k - Cached - Similar pages - Note this

Bush Furniture, Inc
Bush designs and manufactures quality, ready to assemble, entertainment centers, TV stands, home office and business furniture.
www.**bush**furniture.com/ - 26k - Cached - Similar pages - Note this

Pattern Mining

- Trouver des motifs fréquents/récurrents dans un ensemble de données
- Panier de la ménagère

Pattern Mining



Dangers du Data Mining

- Implication et Causalité
- Paradoxe de Simpson
- Data dredging (nettoyage)
- Redondance
- Pas d'Informations Nouvelles
- Sur-apprentissage (modèles prédictifs)

Implication et Causalité

- Coca-Cola Light $\rightarrow$ Obésité
- Soins intensifs $\rightarrow$ Mort
- A la plage :
  - Ventes de glaces en hausse $\Rightarrow$ Nombre de noyés en hausse
  - Nombre de noyés en hausse $\Rightarrow$ Ventes de glaces en hausse

http://www.tylervigen.com/

Paradoxe de Simpson

Exemple

On considère deux contributeurs de Wikipédia : Lisa et Bart. La première semaine, Lisa améliore 60% des articles qu'elle édite alors que Bart améliore 90% des articles qu'il édite. La deuxième semaine, Lisa n'améliore que 10% des articles et Bart s'en tient à un score de 30%. Les deux fois, Bart obtient un meilleur score que Lisa. Mais lorsque les deux actions sont combinées, Lisa a amélioré un plus grand pourcentage que Bart. Comment est-ce possible ?

Paradoxe de Simpson

Exemple

On considère deux contributeurs de Wikipédia : Lisa et Bart. La première semaine, Lisa améliore 60% des articles qu'elle édite alors que Bart améliore 90% des articles qu'il édite. La deuxième semaine, Lisa n'améliore que 10% des articles et Bart s'en tient à un score de 30%. Les deux fois, Bart obtient un meilleur score que Lisa. Mais lorsque les deux actions sont combinées, Lisa a amélioré un plus grand pourcentage que Bart. Comment est-ce possible ?

	Semaine 1	**Semaine 2**	**Total**
Lisa	60/100 = 60 %	1/10 = 10 %	61/110 = 55,45 %
Bart	9/10 = 90 %	30/100 = 30 %	39/110 = 35,45 %

Data Dredging

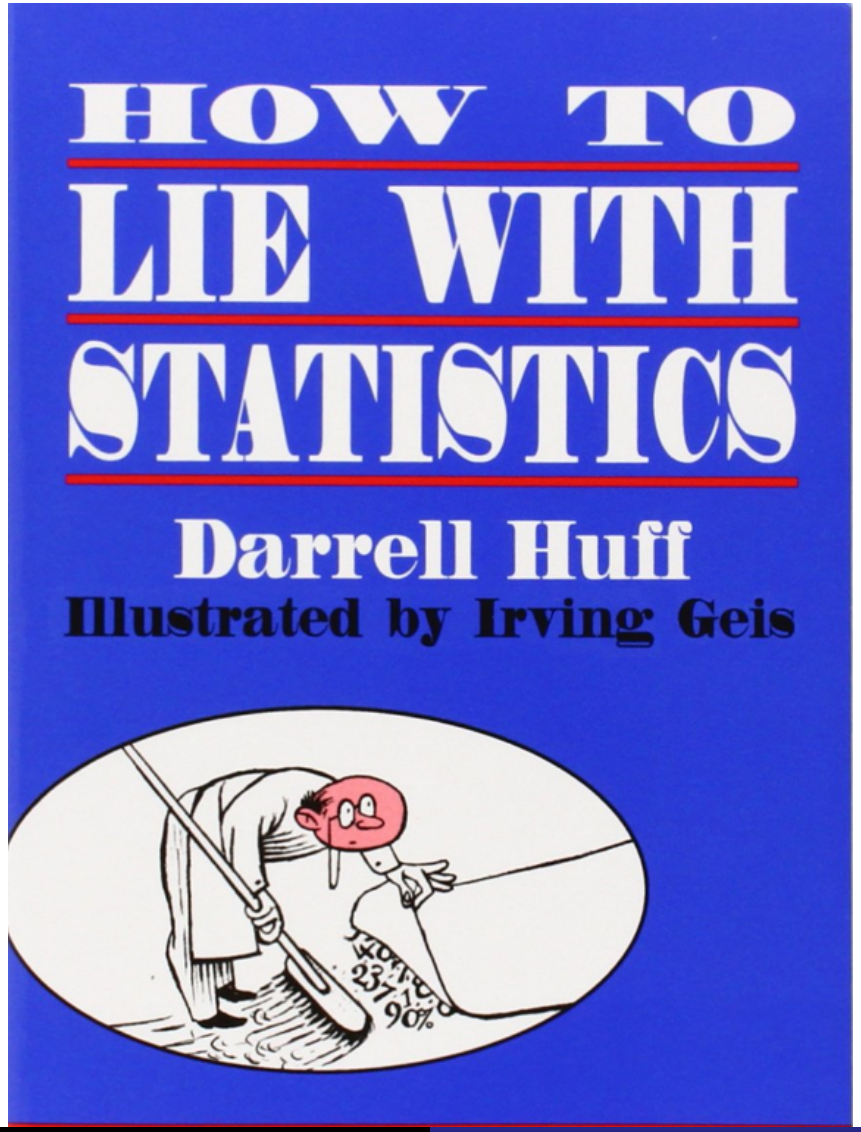
- Torturer les donées jusqu'à confession

Pas d'informations nouvelles

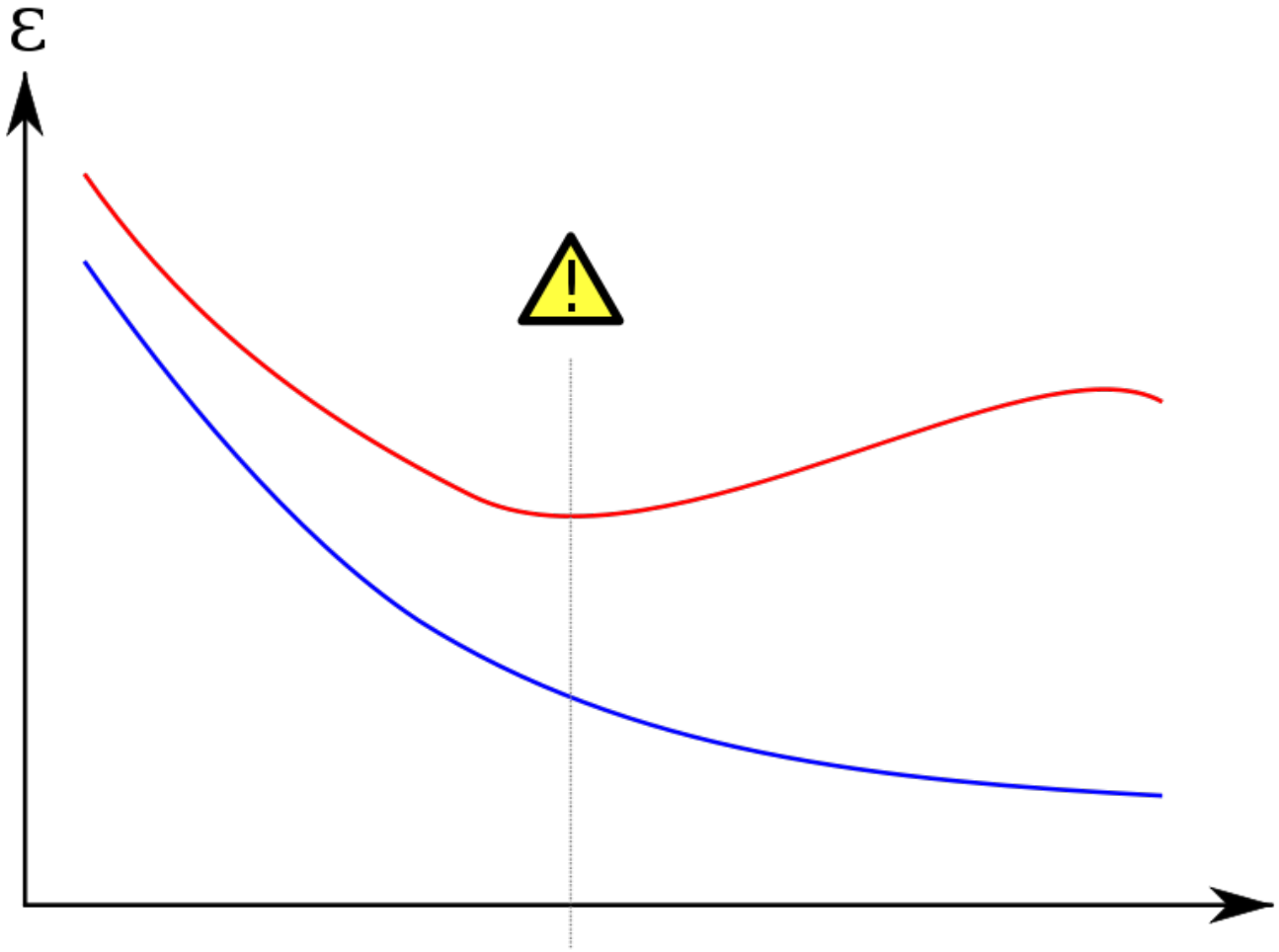
- Motifs les plus fréquents = motifs les plus connus
- Beaucoup de motifs intéressants sont peu fréquents, sinon, on les connaitrait déjà

Rule $X \rightarrow Y$	Supp(XY)	Co
solar_alt→temp	33%	66
precip→cloud	21%	64
cloud→precip	21%	48
w_speed→precip	19%	44
cloud, w_speed→precip	13%	57

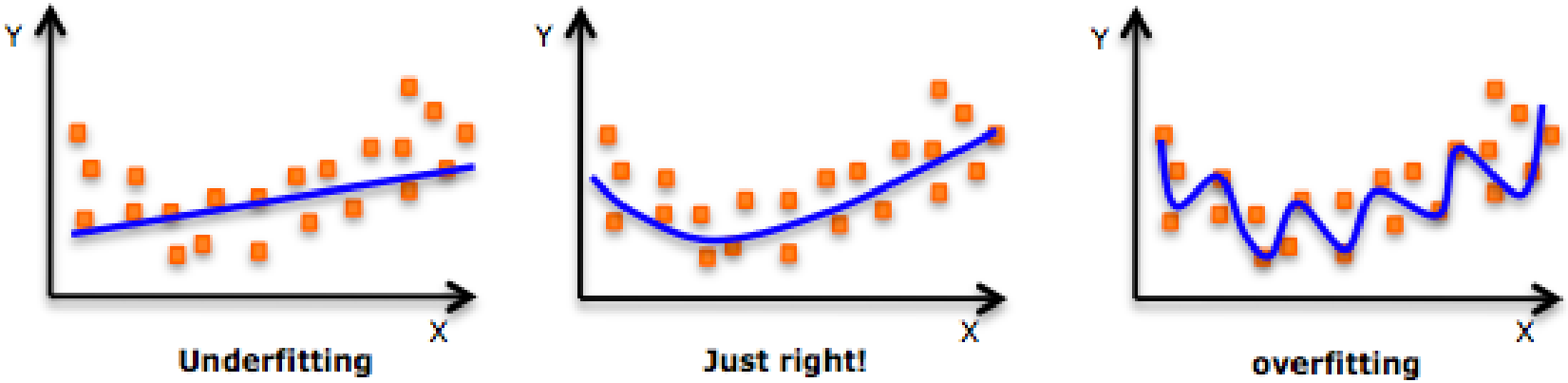
Et pleins d'autres



Sur-apprentissage



Sur-apprentissage



Détection de règles d'association

Histoire

Le concept de règle d'association a été popularisé, en particulier, par un article de Rakesh Agrawal de 1993. Mais il est possible que cette notion ait été découverte sous le nom de GUHA en 1966 par Petr Hájek et ses collègues. Rakesh Agrawal et son équipe présentent des règles d'association dont le but est de découvrir des similitudes entre des produits dans des données saisies sur une grande échelle dans les systèmes informatiques des points de ventes des chaînes de supermarchés - panier de la ménagère

Détection de règles d'association

- Données transactionnelles

- Recherche d'associations inconnues/intéressantes dans ces bases transactionnelles
- Très (très) grand volumne de données

Détection de règles d'association

- Applications à la recomendation

- En classification : découverte de motifs fréquents dans une classe, mais pas dans les autres

Frequent Item Sets

TID	Produce
1	MILK, BREAD, EGGS
2	BREAD, SUGAR
3	BREAD, CEREAL
4	MILK, BREAD, SUGAR
5	MILK, CEREAL
6	BREAD, CEREAL
7	MILK, CEREAL
8	MILK, BREAD, CEREAL, EGGS
9	MILK, BREAD, CEREAL

Frequent Item Sets

TID	Products
1	A, B, E
2	B, D
3	B, C
4	A, B, D
5	A, C
6	B, C
7	A, C
8	A, B, C, E
9	A, B, C

ITEMS:

A = milk
B= bread
C= cereal
D= sugar
E= eggs

Frequent Item Sets

TID	Products
1	A, B, E
2	B, D
3	B, C
4	A, B, D
5	A, C
6	B, C
7	A, C
8	A, B, C, E
9	A, B, C

TID	A	B	C	D	E
1	1	1	0	0	1
2	0	1	0	1	0
3	0	1	1	0	0
4	1	1	0	1	0
5	1	0	1	0	0
6	0	1	1	0	0
7	1	0	1	0	0
8	1	1	1	0	1
9	1	1	1	0	0

Frequent Itemsets

- Soit $I = \{i_1, ...i_m\}$ un ensemble d'items
- Soit $T = \{t_1, ..., t_n\}$ un ensemble de transactions où t_i est un sous-ensemble de I

Support

Le support $Supp(x)$ avec $x \subseteq I$ est le nombre de transactions qui contiennet x.

$$Supp(x) = \frac{x.count}{Card(T)}$$

Un *Frequent itemset* est un sous-ensemble de I dont le support est supérieur à une certaine valeur

$$Supp(x) \geq minsup$$

Frequent Itemsets

TID	Products
1	A, B, E
2	B, D
3	B, C
4	A, B, D
5	A, C
6	B, C
7	A, C
8	A, B, C, E
9	A, B, C

- Quel est le support de $\{A\}, \{B, D\}, \{A, B, E\}$?
- Quels sont les itemsets fréquents avec $minsup = 4$

Frequent Itemsets

TID	Products
1	A, B, E
2	B, D
3	B, C
4	A, B, D
5	A, C
6	B, C
7	A, C
8	A, B, C, E
9	A, B, C

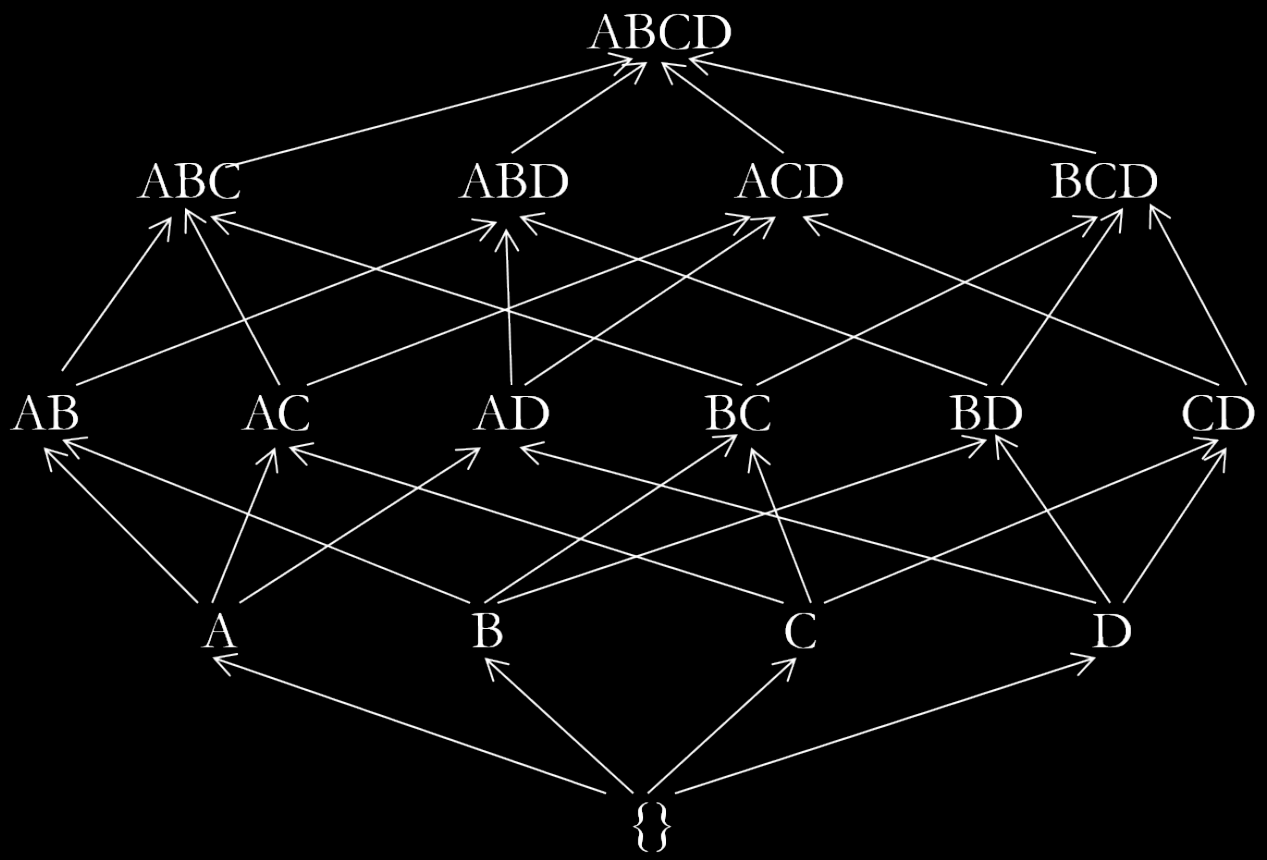
- Quel est le support de $\{A\}, \{B, D\}, \{A, B, E\}$?
- Quels sont les itemsets fréquents avec $minsup = 4$

$$\{\}, \{A\}, \{B\}, \{C\}, \{A, B\}, \{A, C\}, \{B, C\}$$

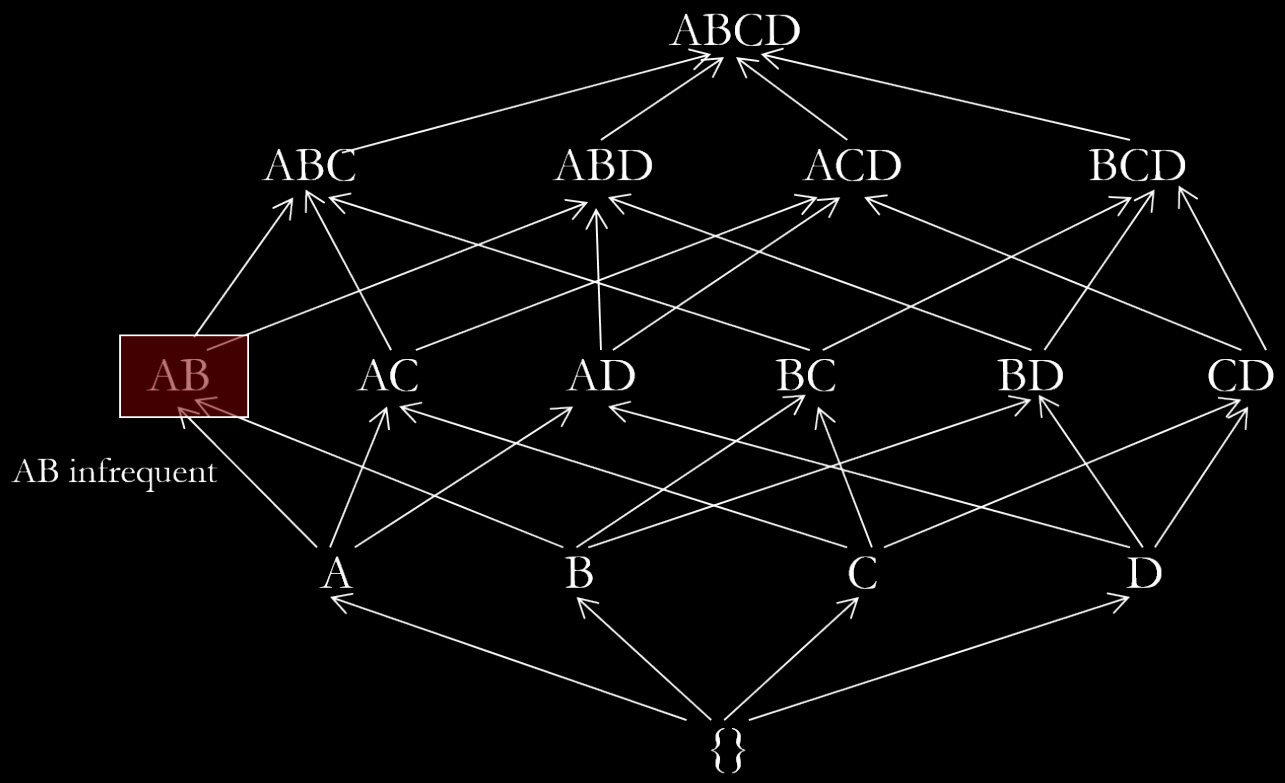
Monotonie

- **Tout sous-ensemble d'un itemset fréquent est aussi un itemset fréquent**
- Imaginez que $\{A, B\}$ apparaissent 10 fois, alors $\{A\}$ et $\{B\}$ apparaissent au moins 10 fois !
- Le principe de monotonie va permettre le développement d'algorithmes performants capables d'analyser de très grandes masses de données

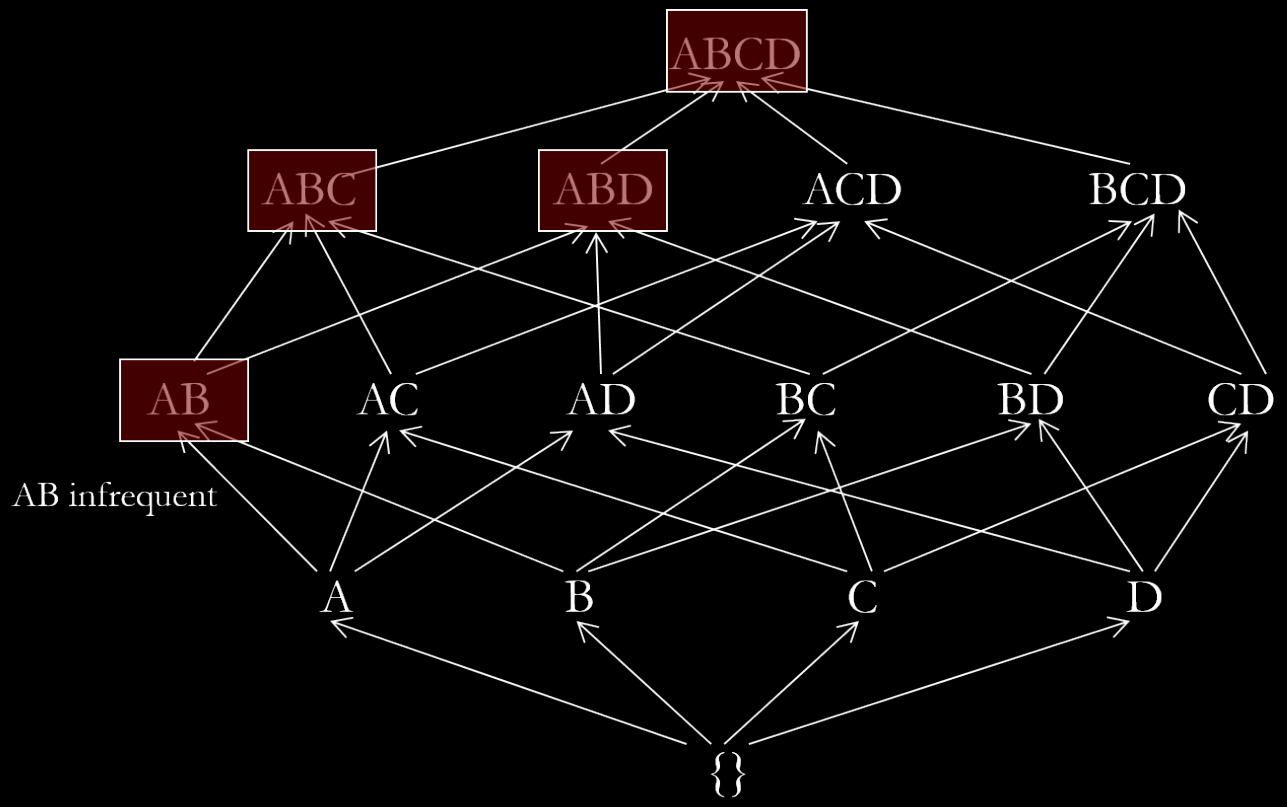
Treillis



Treillis



Treillis



Règles d'association

- **Définition :** $x \Rightarrow y$
- x et y sont des itemset $\subseteq I$
- x et y sont distinct : $x \cap y = 0$
- y est non vide : $y \neq 0$
- **Signification :** Si une transaction contient x, alors elle contient y aussi
- Exemple : $\{A, C\} \Rightarrow \{D, E, F\}$

Règles d'association

- Si $R : x \Rightarrow y$ alors $Supp(R) = Supp(x) \cup Supp(y)$
- Confiance : $Conf(R) = Supp(R)/Supp(x)$
  - La confiance mesure la fraction de transaction $x \cup y$ par rapport à celles qui ont x
- Les règles avec un haut support, et une confiance élevée sont appelées "règles fortes" (strong rules)

TID	List of items
1	**A, B, E**
2	B, D
3	B, C
4	A, B, D
5	A, C
6	B, C
7	A, C
8	**A, B, C, E**
9	A, B, C

- Considérons l'ensemble $\{A, B, E\}$, quelles sont les règles de $minsup = 2$ et de $minconf = 50\%$?

Découverte de règles d'association

- Etant donné :
  - *minsup*, *minconf*, et un ensemble de transactions T
- Problème :
  - Trouver les règles d'association R de T telles que $Supp(R) \geq minsup$ et $Conf(R) \geq minconf$
- Problème combinatoire complexe

Découverte de règles d'association

- Approche typique :
  - Une règle $R : X \Rightarrow Y$ satisfie *minsup* et *minconf* ssi :
    - $Supp(X \cup Y) \geq minsup$
    - $Supp(X \cup Y)/Supp(X) \geq minconf$
  - On va chercher tous les Z tels que $Supp(Z) \geq minsup$
  - Ensuite, pour chaque Z :
    - On découpe Z en X et Y tel que $Z = X \cap Y$
    - On teste pour savoir si $Supp(X \cup Y)/Supp(X) \geq minconf$

Monotonie des règles d'association

Exercice

Tout comme pour les itemsets, une propriété de monotonie peut être trouvée dans les règles d'association. Cherchez là...

Monotonie des règles d'association

Exercice

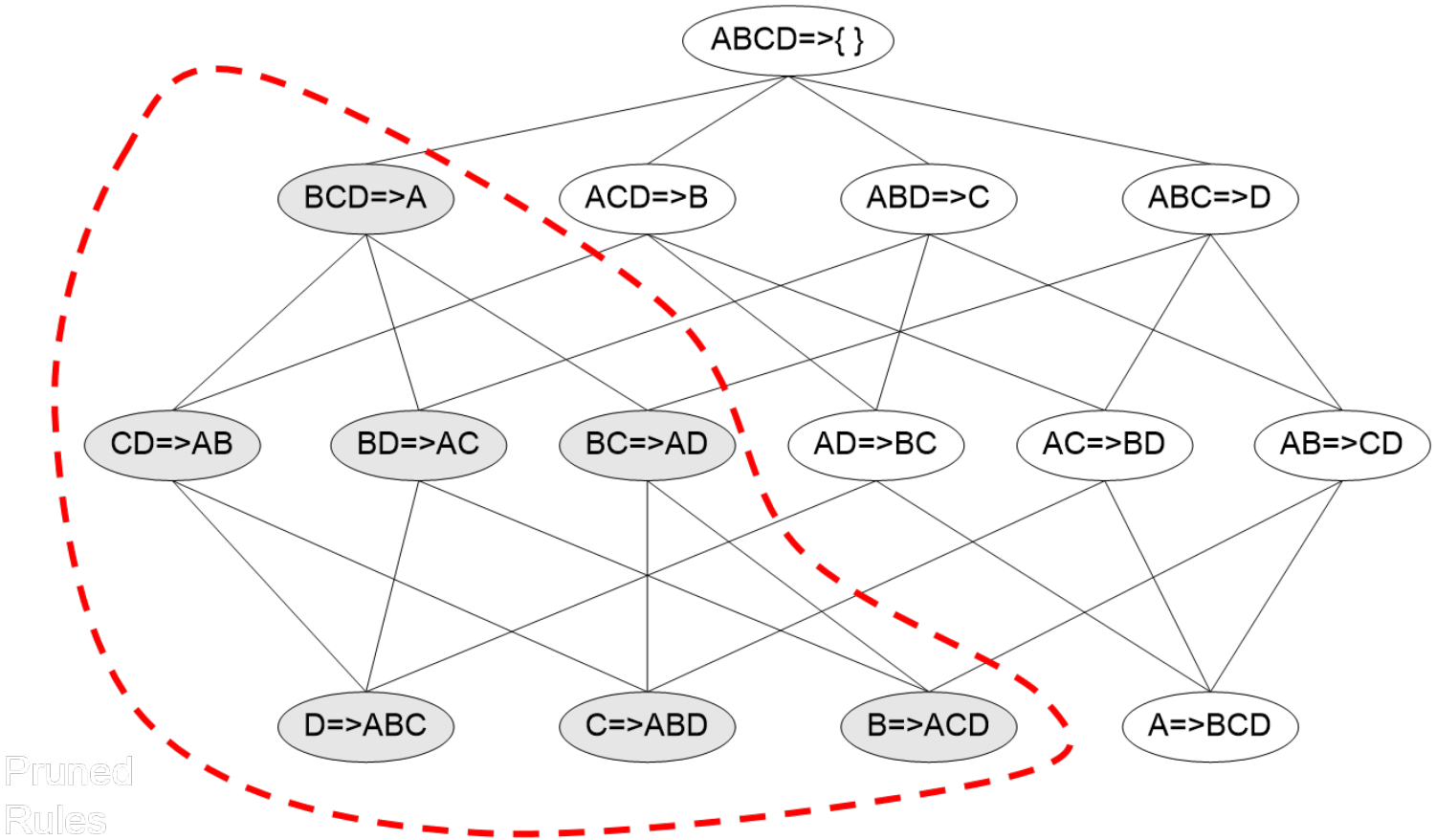
Tout comme pour les itemsets, une propriété de monotonie peut être trouvée dans les règles d'association. Cherchez là...

Solution

Soit $Z = X \cap Y = X' \cap Y'$ avec $X' \subseteq X$ alors :

- $Supp(X \Rightarrow Y) = Supp(X' \Rightarrow Y')$
- $Conf(X \Rightarrow Y) \geq Conf(X' \Rightarrow Y')$

Monotonie des règles d'association



Algorithme APriori

APriori

L'algorithme APriori est un algorithme d'exploration de données conçu en 1994, par Rakesh Agrawal et Ramakrishnan Sikrant, dans le domaine de l'apprentissage des règles d'association. Il sert à reconnaitre des propriétés qui reviennent fréquemment dans un ensemble de données et d'en déduire une catégorisation.

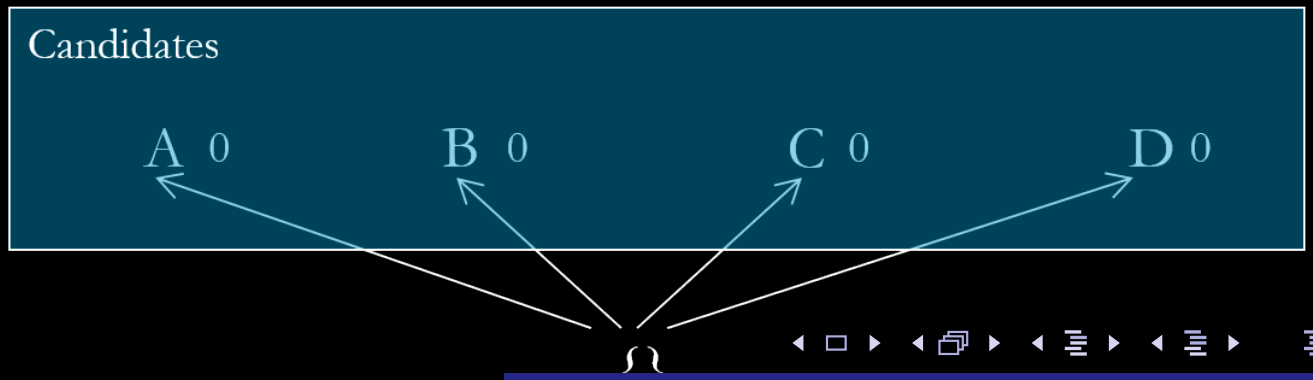
L'algorithme Apriori s'execute en deux étapes :

- Soient minsupp l'indice de support minimum donné, et minconf l'indice de confiance donné.
- Génération de tous les itemsets fréquents c'est-à-dire
- Génération de toutes les règles d'associations de confiance à partir des itemsets fréquents, c'est-à-dire

TID	A	B	C	D
1	0	1	1	0
2	0	1	1	0
3	1	0	1	1
4	1	1	1	1
5	0	1	0	1

minsup=2

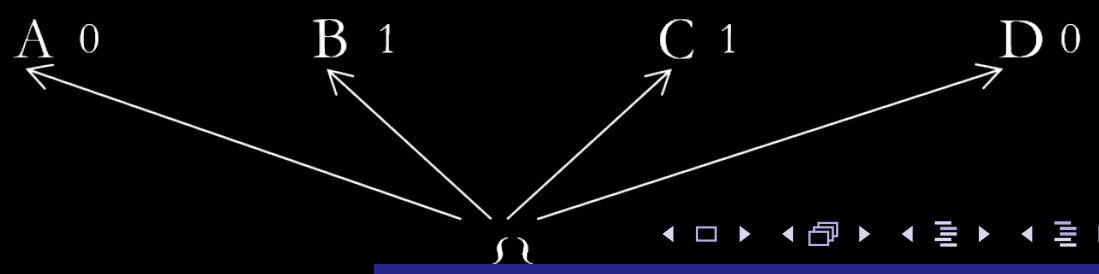
Apriori



TID	A	B	C	D
1	0	1	1	0
2	0	1	1	0
3	1	0	1	1
4	1	1	1	1
5	0	1	0	1

minsup=2

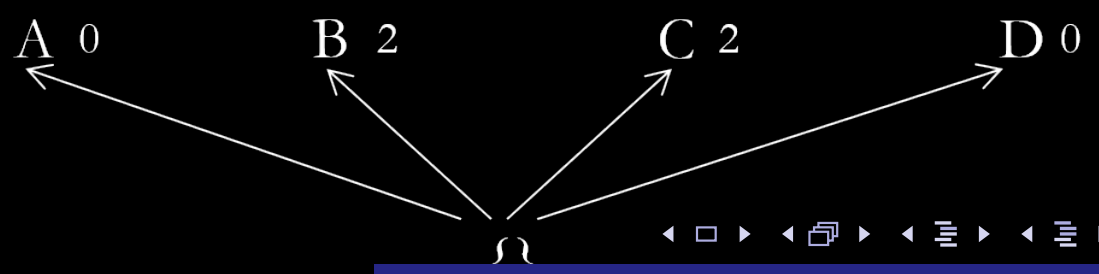
Apriori



TID	A	B	C	D
1	0	1	1	0
2	0	1	1	0
3	1	0	1	1
4	1	1	1	1
5	0	1	0	1

Apriori

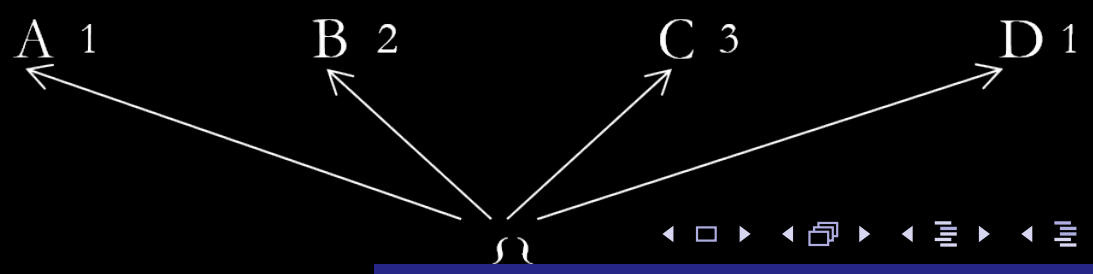
minsup=2



TID	A	B	C	D
1	0	1	1	0
2	0	1	1	0
3	1	0	1	1
4	1	1	1	1
5	0	1	0	1

minsup=2

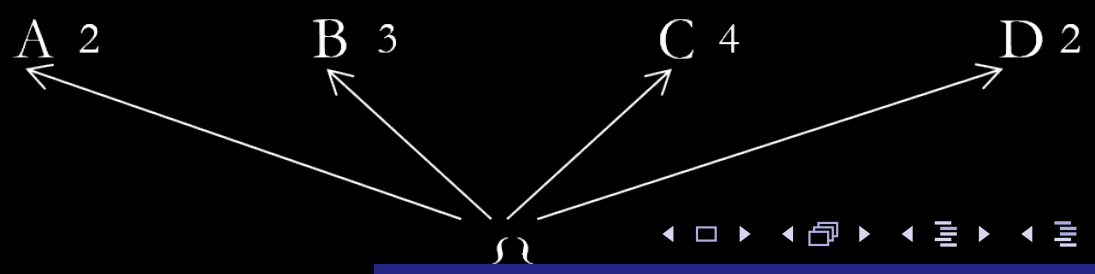
Apriori



TID	A	B	C	D
1	0	1	1	0
2	0	1	1	0
3	1	0	1	1
4	1	1	1	1
5	0	1	0	1

minsup=2

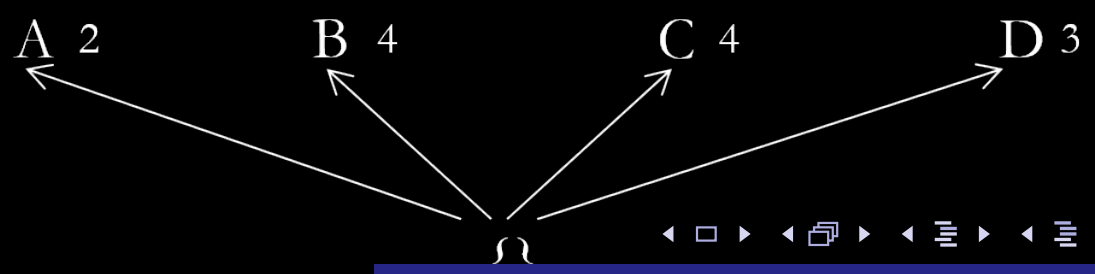
Apriori



TID	A	B	C	D
1	0	1	1	0
2	0	1	1	0
3	1	0	1	1
4	1	1	1	1
5	0	1	0	1

Apriori

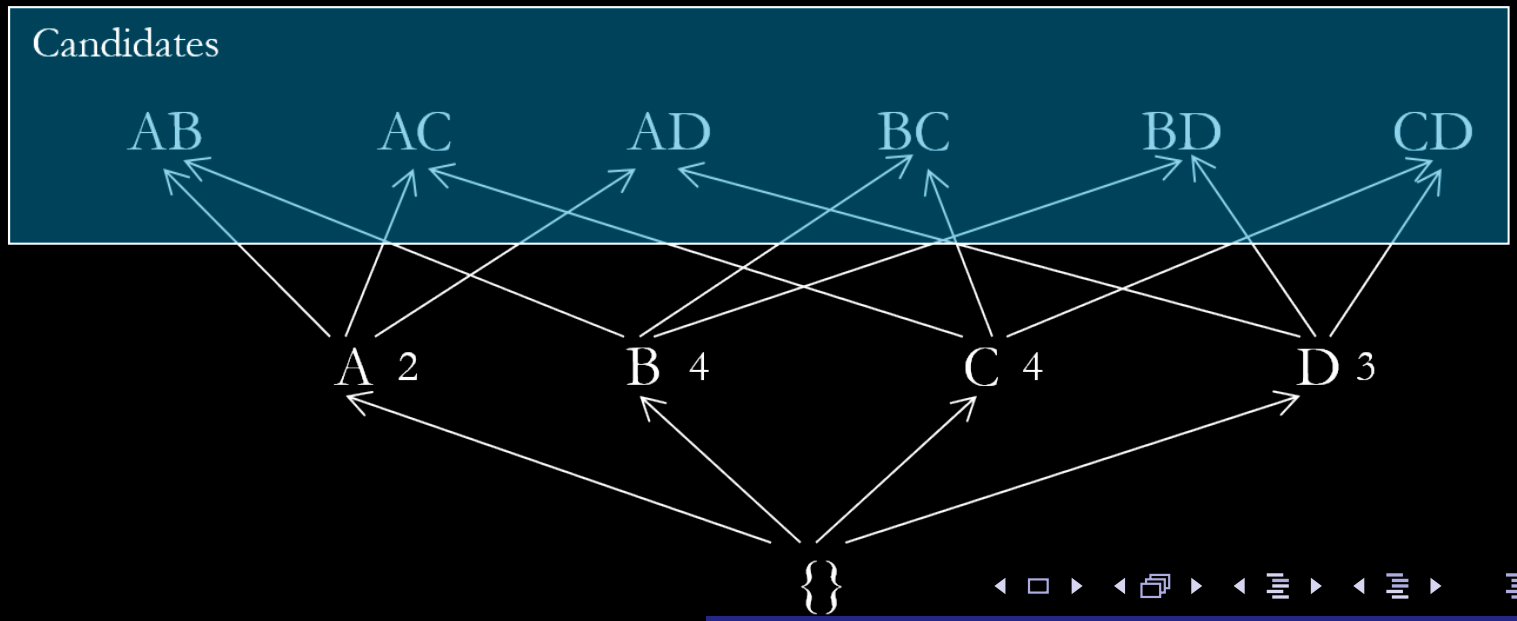
minsup=2



TID	A	B	C	D
1	0	1	1	0
2	0	1	1	0
3	1	0	1	1
4	1	1	1	1
5	0	1	0	1

Apriori

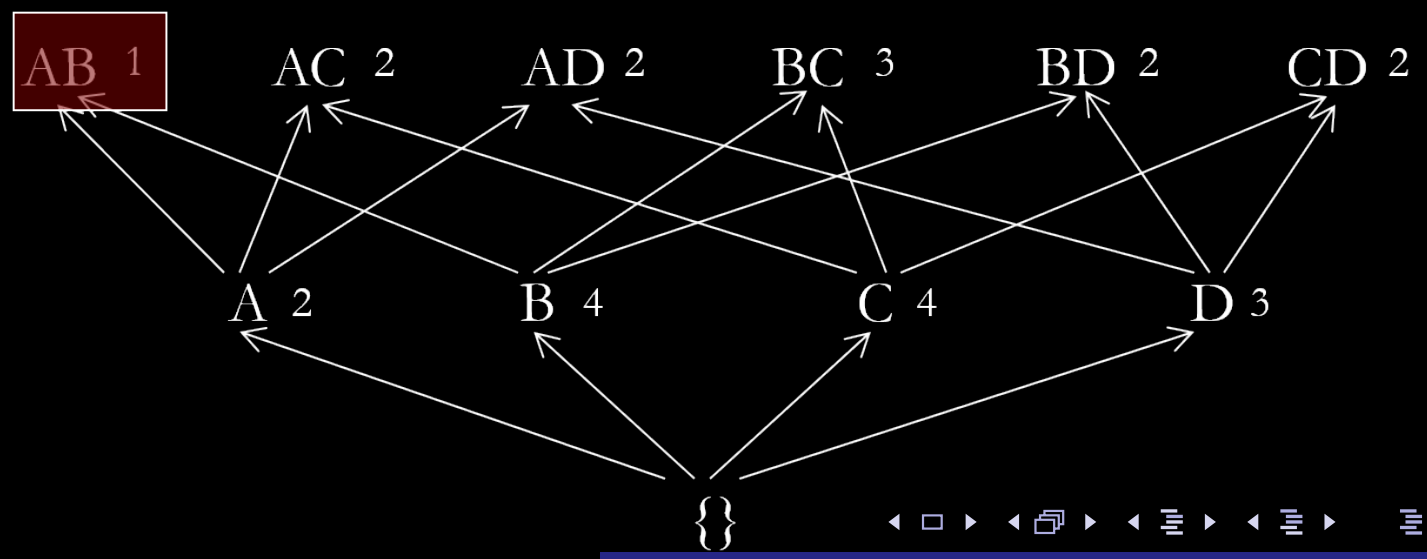
minsup=2



TID	A	B	C	D
1	0	1	1	0
2	0	1	1	0
3	1	0	1	1
4	1	1	1	1
5	0	1	0	1

Apriori

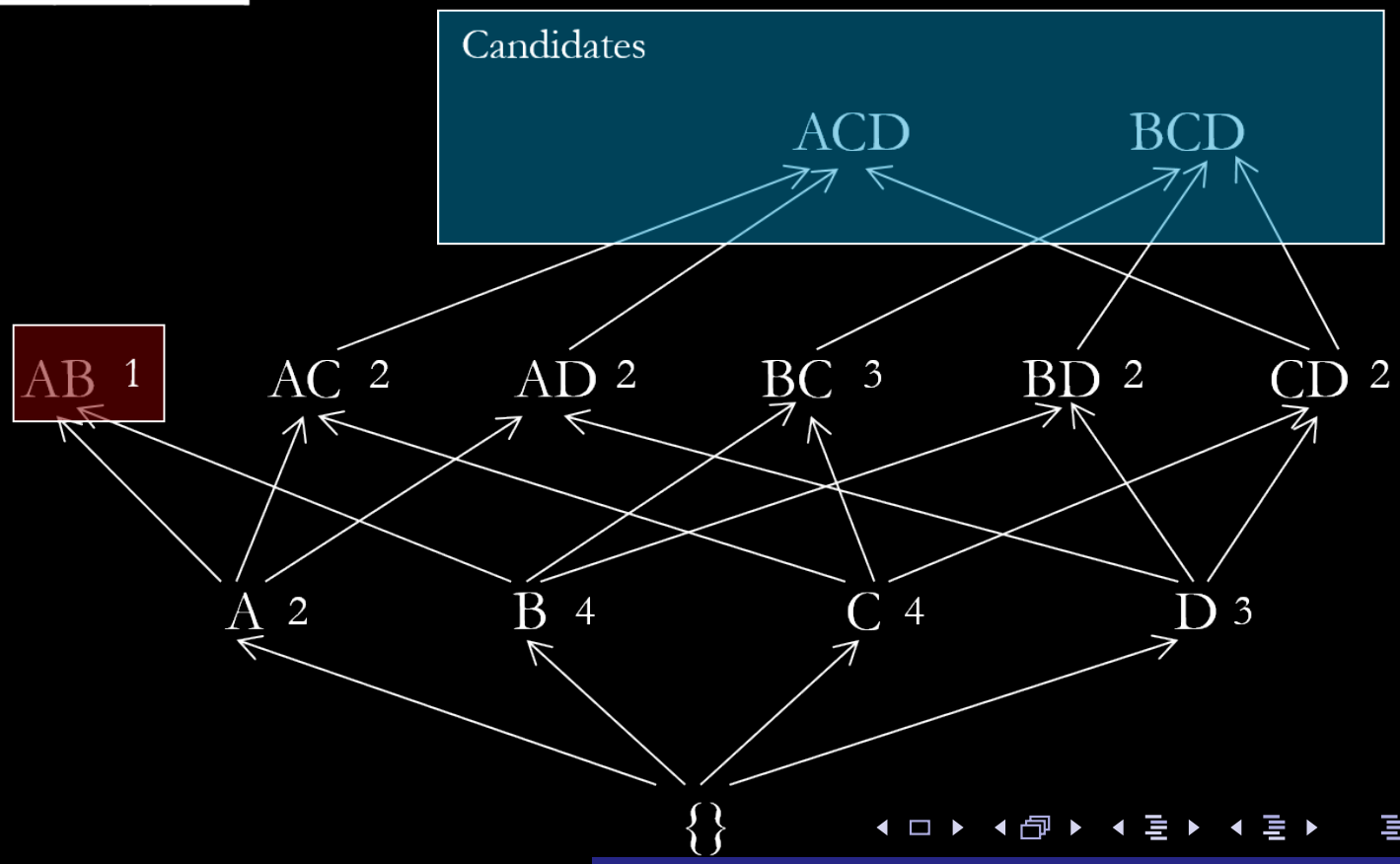
minsup=2



TID	A	B	C	D
1	0	1	1	0
2	0	1	1	0
3	1	0	1	1
4	1	1	1	1
5	0	1	0	1

Apriori

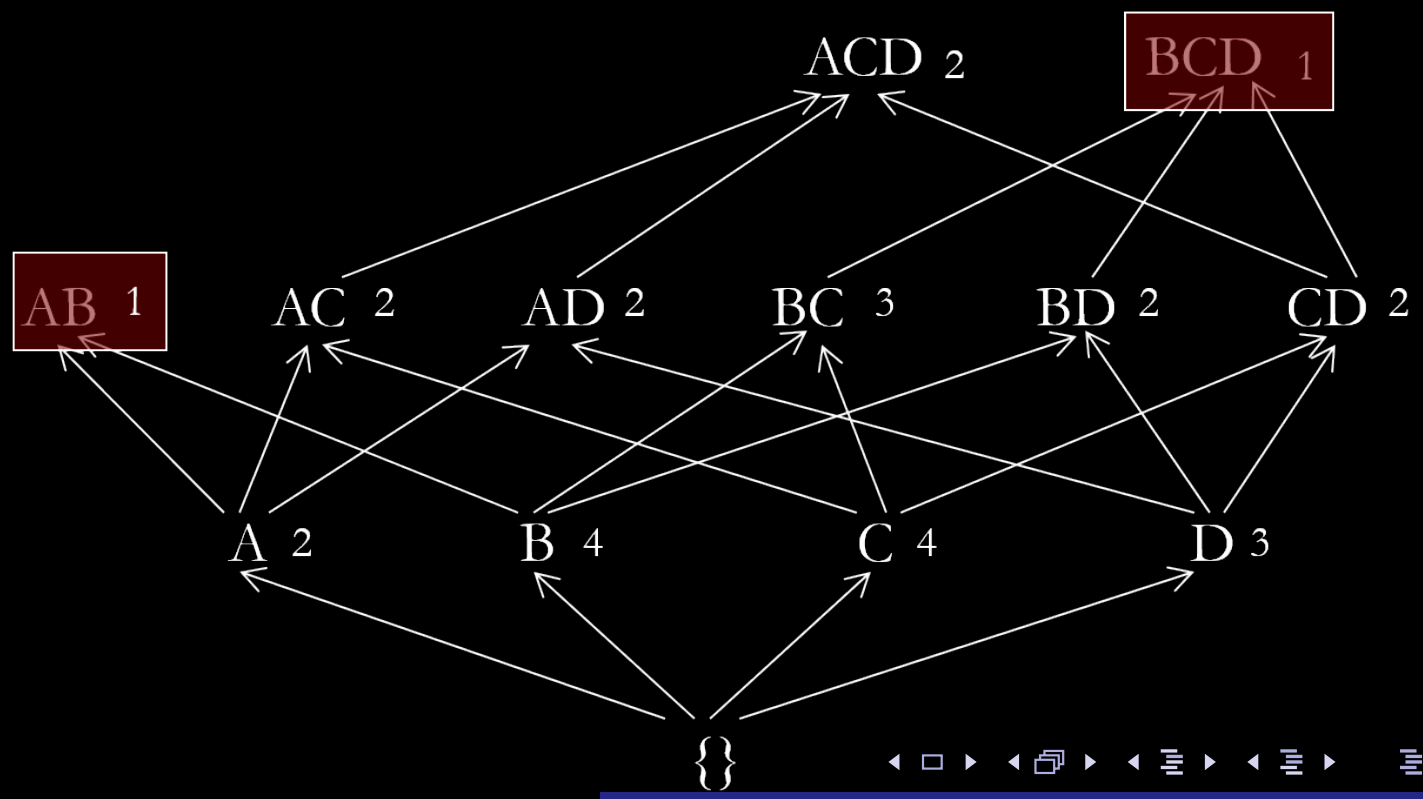
minsup=2



TID	A	B	C	D
1	0	1	1	0
2	0	1	1	0
3	1	0	1	1
4	1	1	1	1
5	0	1	0	1

Apriori

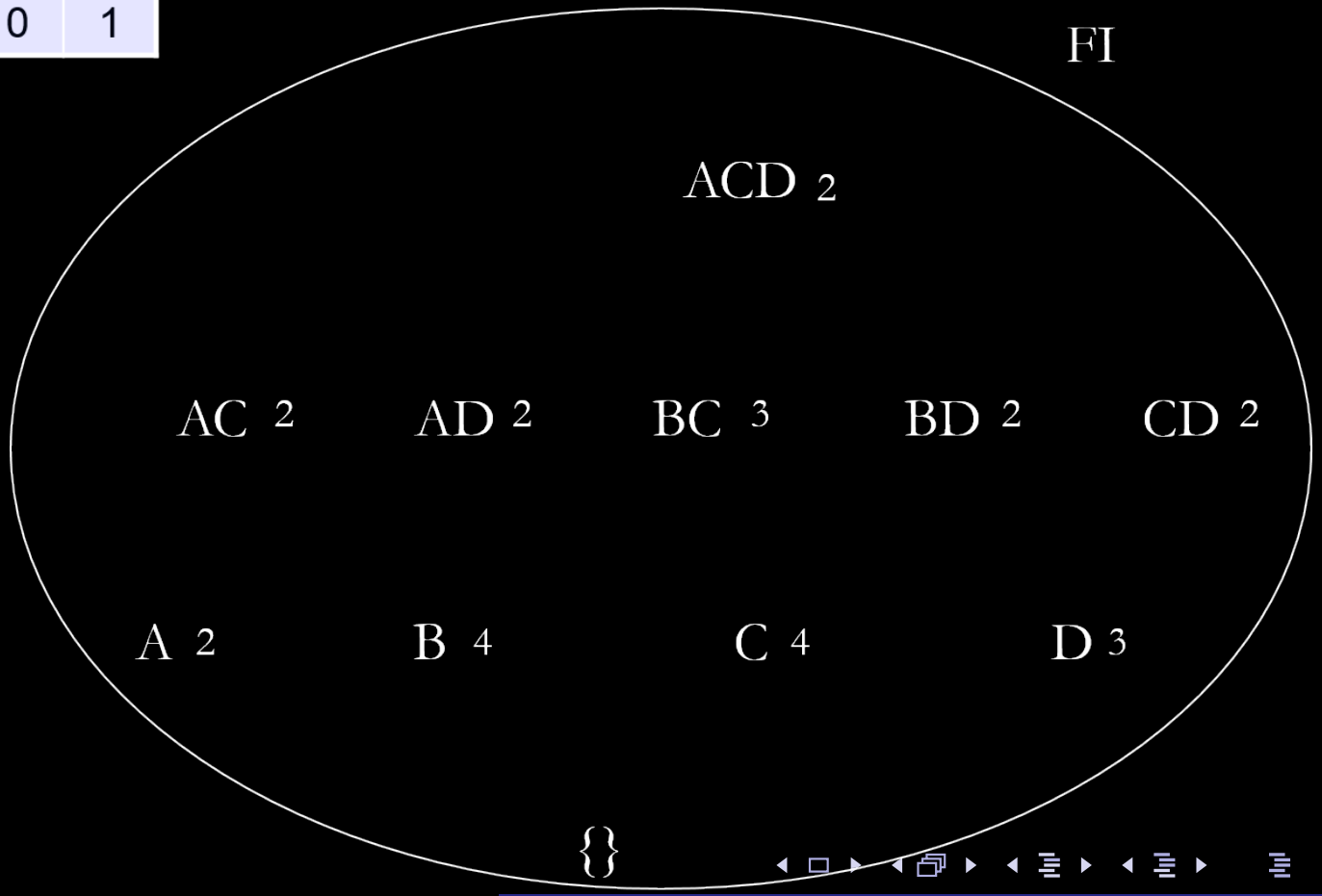
minsup=2



TID	A	B	C	D
1	0	1	1	0
2	0	1	1	0
3	1	0	1	1
4	1	1	1	1
5	0	1	0	1

minsup=2

Apriori



Exercice

TID	Items
1	a, b, c
2	b, c, d, e
3	c, d
4	a, b, d
5	a, b, c

Trouvez toutes les règles d'association pour $minsup = 2$ et $minconf = 0.7$