Optimisation de requêtes inductives : application à l'extraction sous contraintes de règles d'association

Baptiste Jeudy

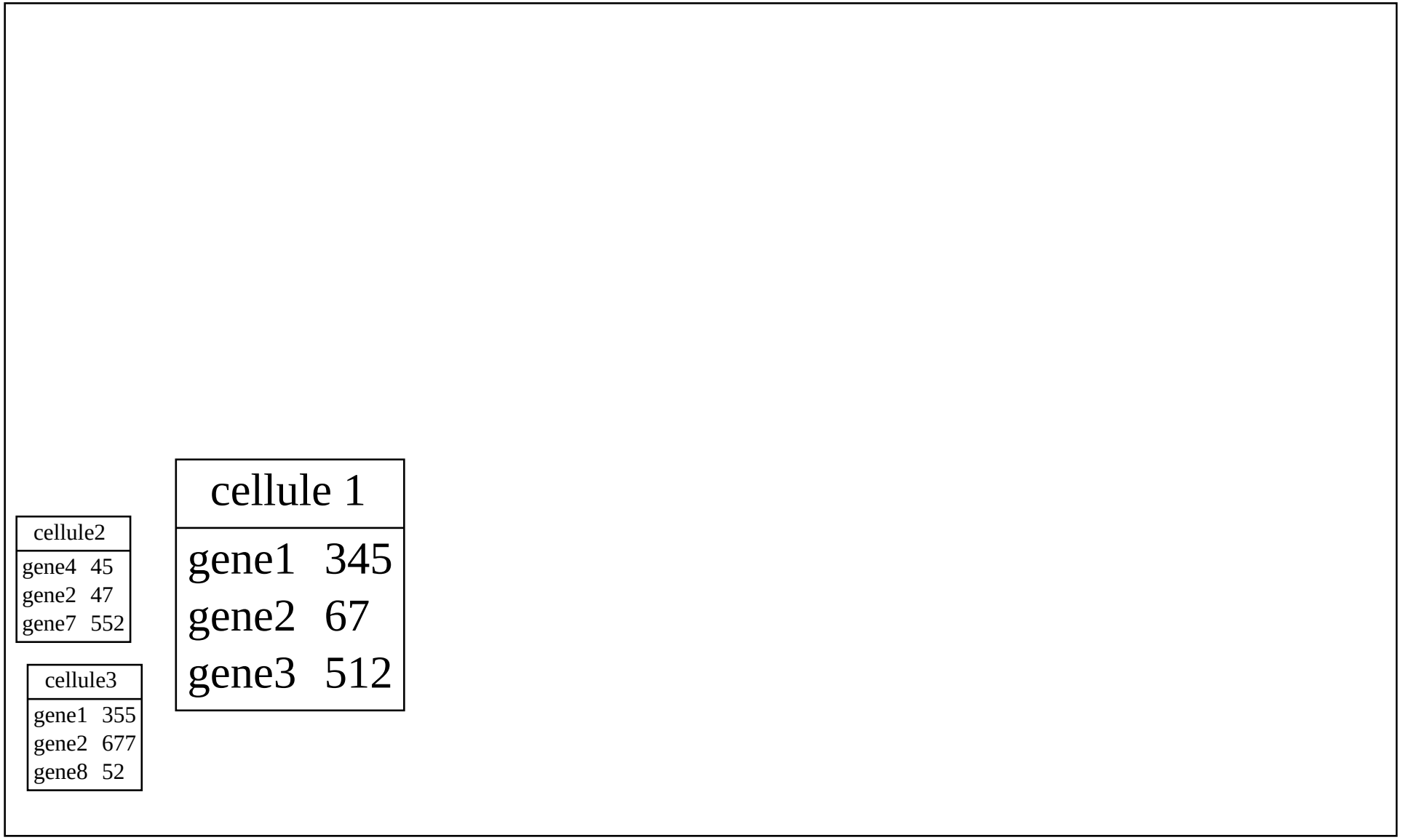
Doctorat préparé au LISI (INSA Lyon)
sous la direction de
Jean-François Boulicaut et Lionel Brunie.

Introduction

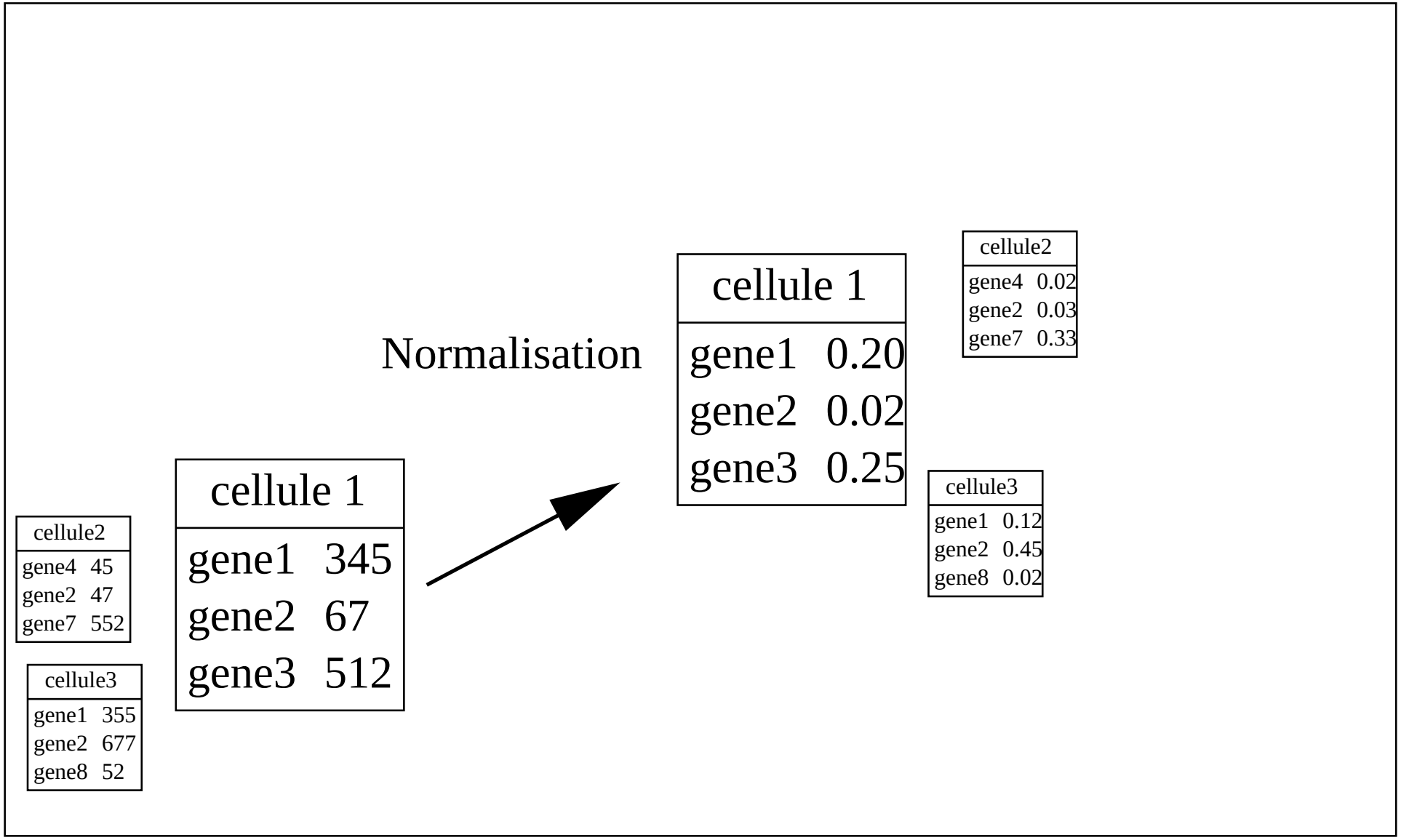
Extraction de connaissances dans les données

Un exemple sur lequel nous avons travaillé

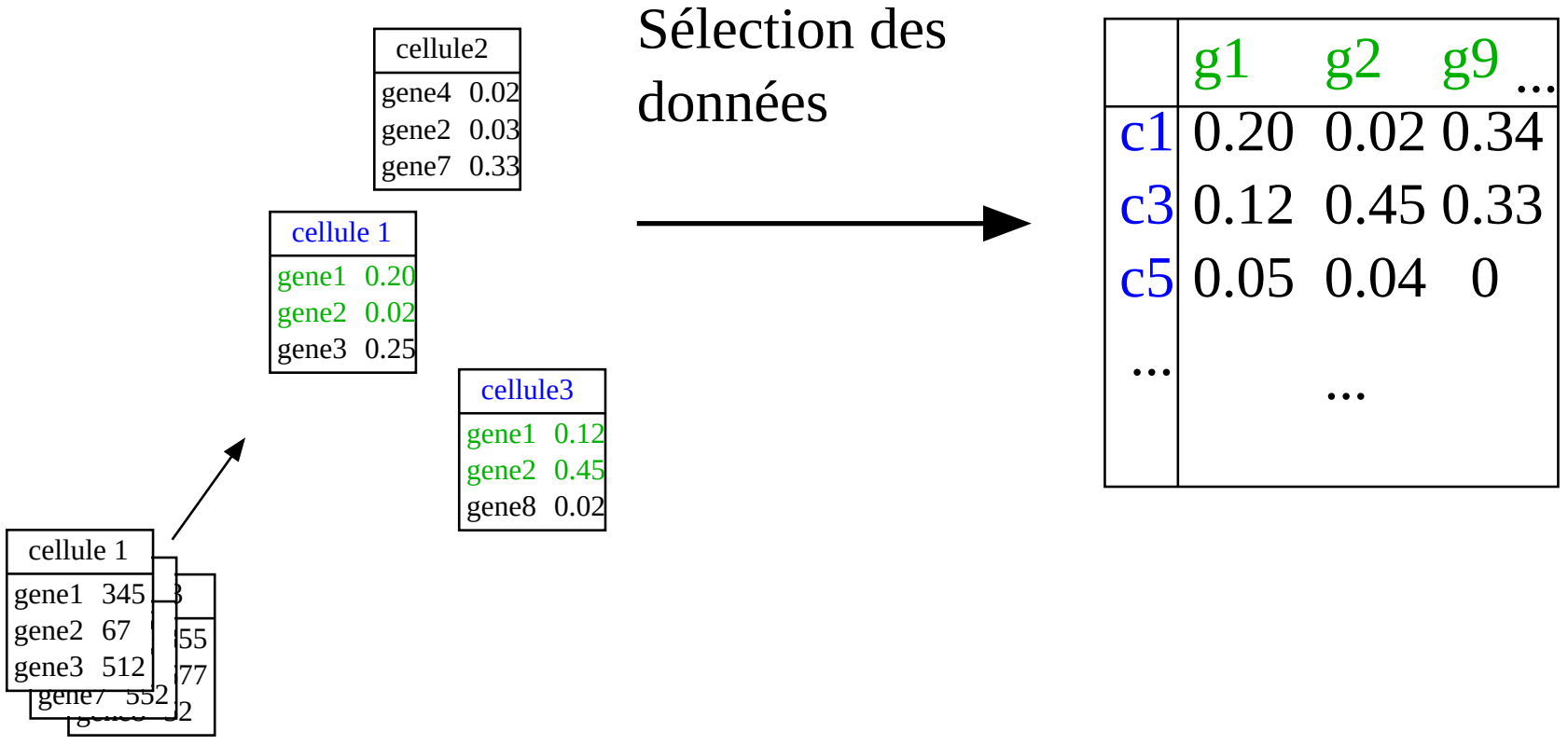
Exemple de Processus d'extraction



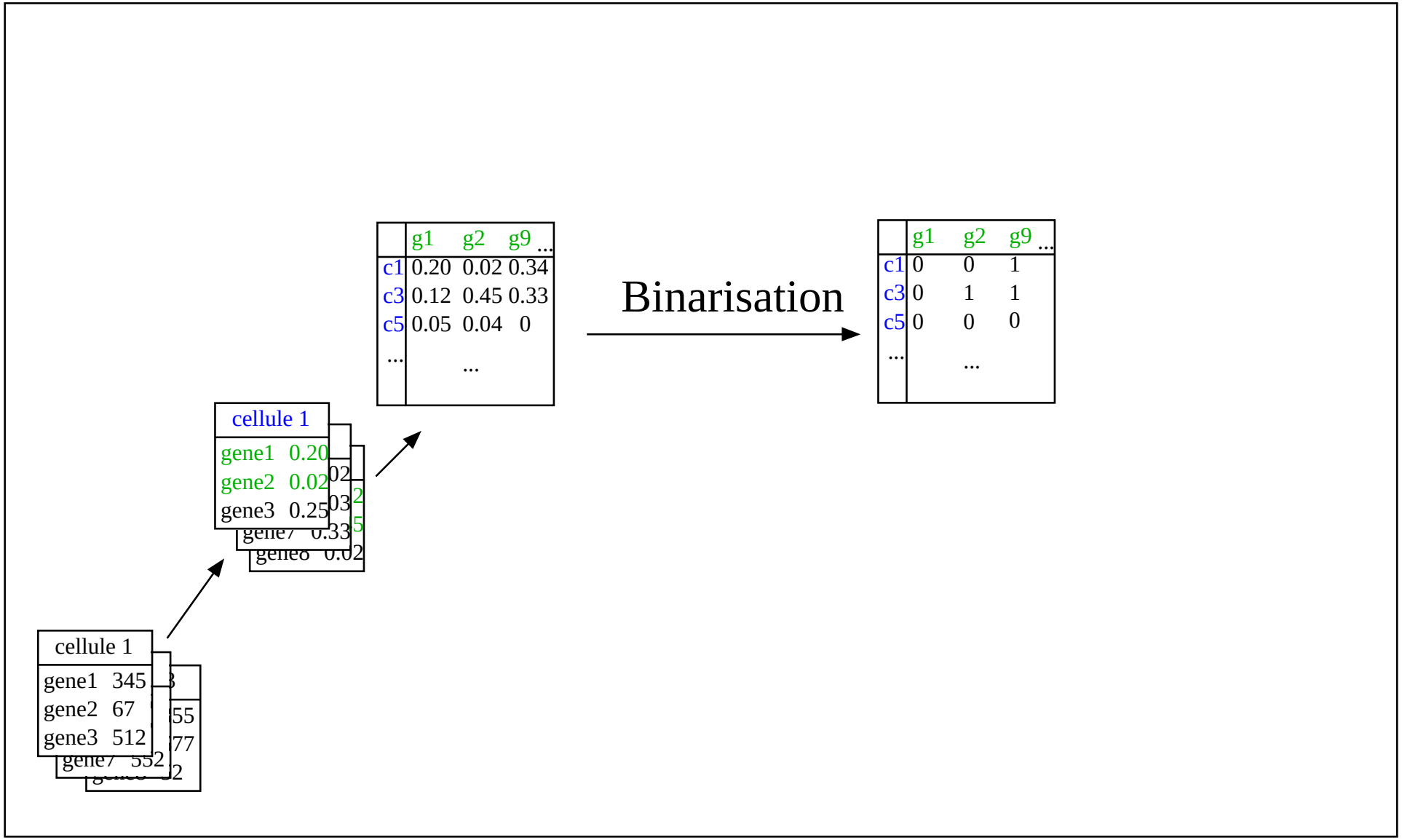
Exemple de Processus d'extraction



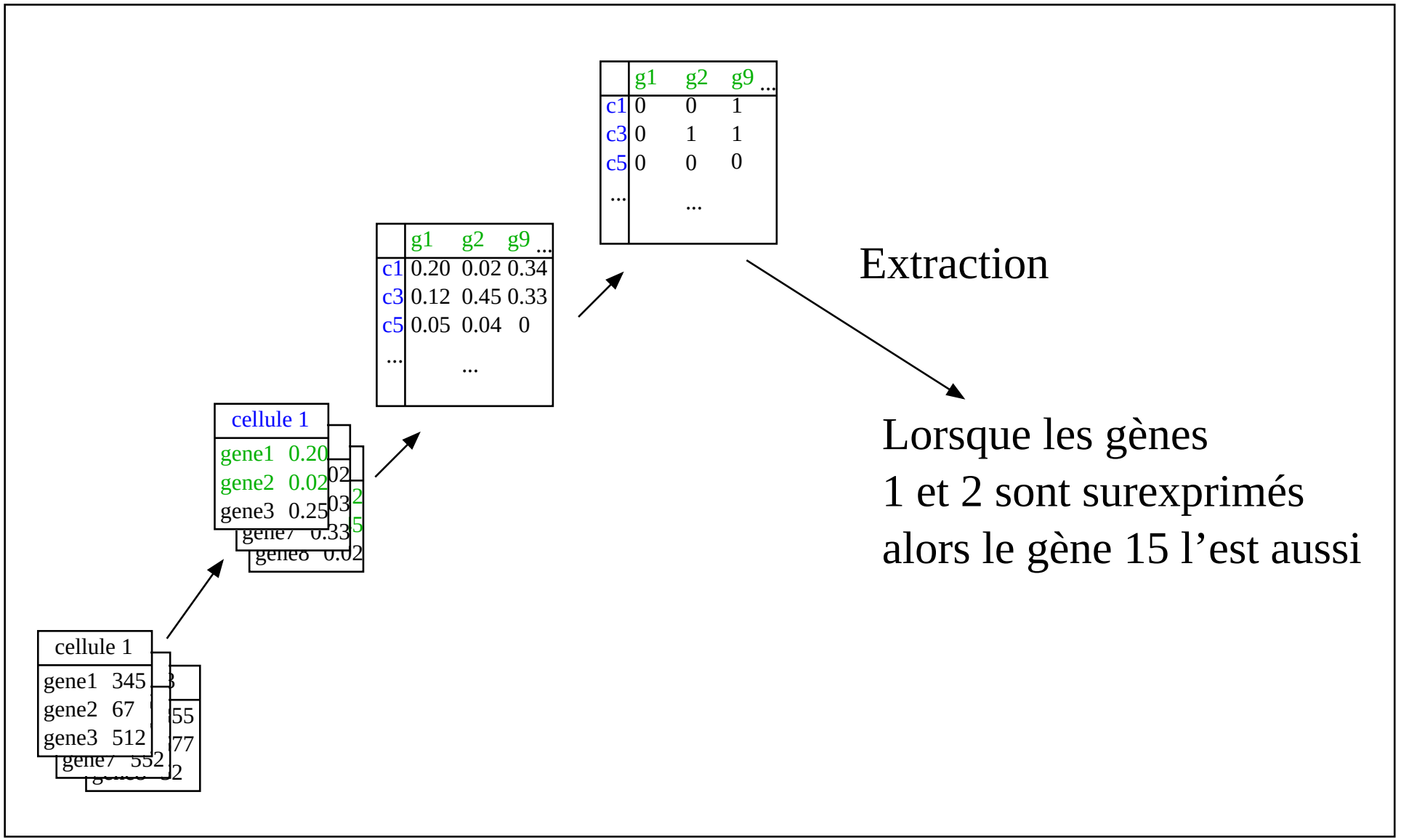
Exemple de Processus d'extraction



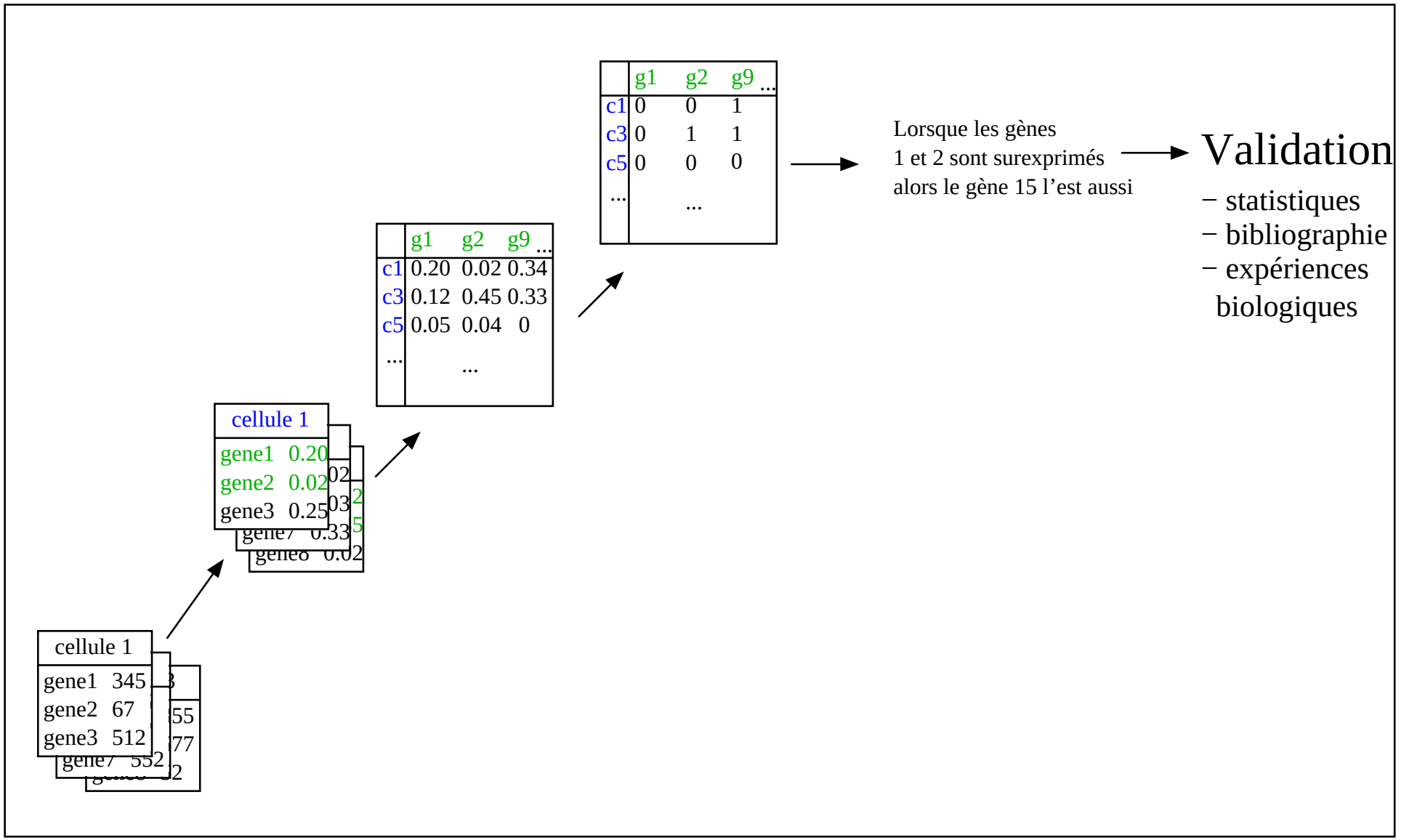
Exemple de Processus d'extraction



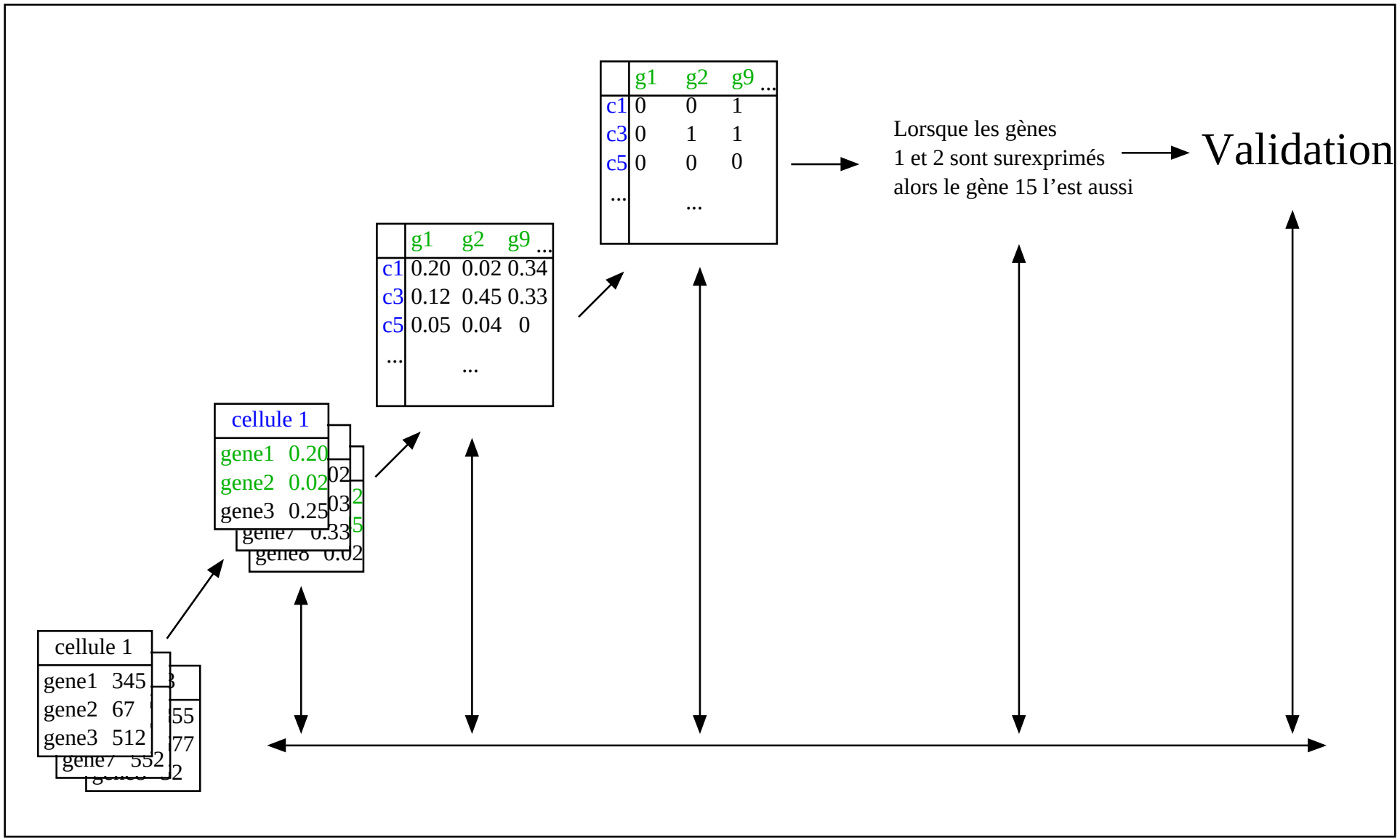
Exemple de Processus d'extraction



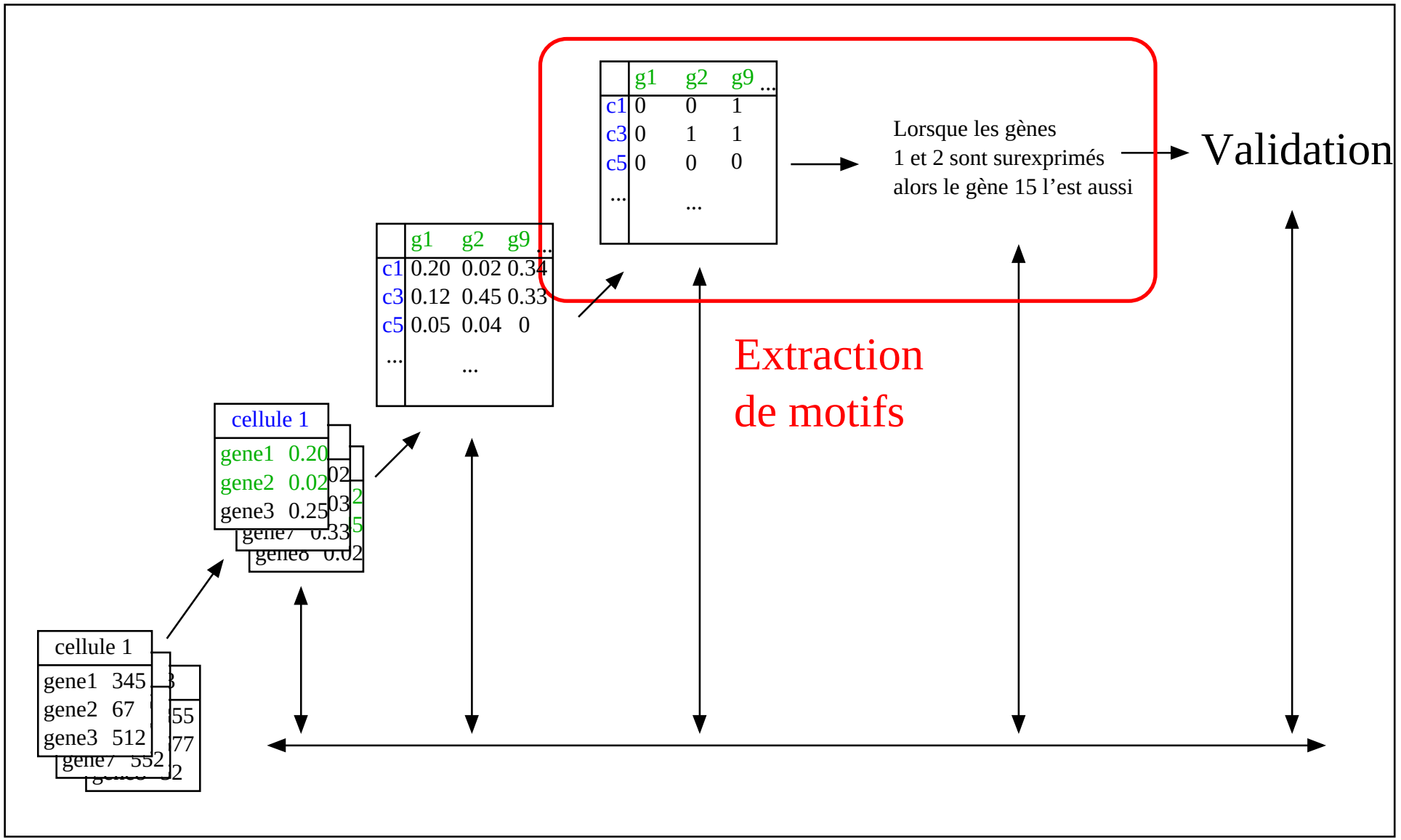
Exemple de Processus d'extraction



Exemple de Processus d'extraction



Exemple de Processus d'extraction



Bases de données booléennes

- Base de données cellules/gènes

	gène 1	gène 2	gène 3	gène 4	gène 5
cel. 1	1	1	1	1	1
cel. 2	1	1	1	1	0
cel. 3	1	0	1	1	0
cel. 4	1	1	0	0	1
cel. 5	0	0	1	1	0
cel. 6	0	0	1	0	1

- Paniers/produits
- Documents/mots-clés

Bases de données booléennes

- Hypothèses de travail
  - nb. de lignes : 10 – 10^7
  - nb. de colonnes : 10 – 10000

	A	B	C	D	E
1	1	1	1	1	1
2	1	1	1	1	0
3	1	0	1	1	0
4	1	1	0	0	1
5	0	0	1	1	0
6	0	0	1	0	1

$\Longleftrightarrow$

tid	
1	ABCDE
2	ABCD
3	ACD
4	ABE
5	CD
6	CE

Motifs utilisés

tid	
1	ABCDE
2	ABCD
3	ACD
4	ABE
5	CD
6	CE

$\Longrightarrow$

Itemset	fonctions d'évaluation freq.	clôture
A	4	A
B	3	AB
BD	2	ABCD
ACD	3	ACD
...	...	...

Règle	fonctions d'évaluation freq.	confiance
$A \Rightarrow CD$	3	75%
$BD \Rightarrow AC$	2	100%
$C \Rightarrow D$	4	80%
...	...	...

Bases de données inductives

- Nos travaux se placent dans le cadre du projet européen cInQ
- Une base de données inductive contient
  - Données
  - Motifs
- Le processus d'ECD s'articule autour de requêtes de sélection sur les motifs
  - Itemsets fréquents concernant les gènes 1 et 5
  - Règles à forte confiance dont la conclusion concerne le gène 3

Requêtes inductives simples et étendues

tid	
1	ABCDE
2	ABCD
3	ACD
4	ABE
5	CD
6	CE

$\Longrightarrow$

$\mathsf{Th}(\mathcal{C}_{2\text{-minfreq}}) =$

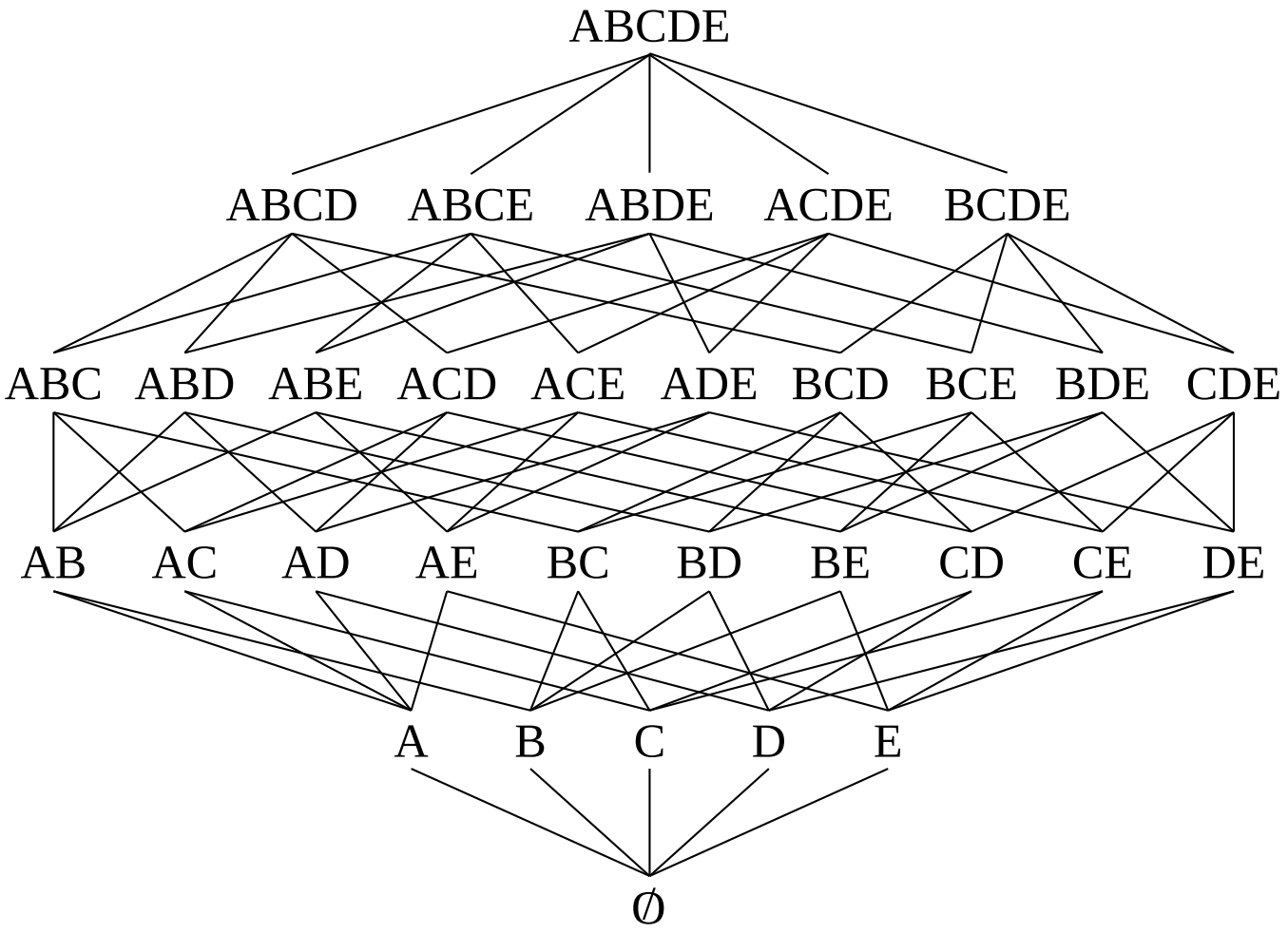
Itemset
$\emptyset$
A
B
ACD
CD
...

$\mathsf{Th}^{+}(\mathcal{C}_{2\text{-minfreq}}, \mathsf{freq}) =$

Itemset	fréq.
$\emptyset$	6
A	4
B	3
ACD	3
CD	4
...	...

Difficultés

- Le treillis des itemsets ordonné par l'inclusion



Difficultés

Prendre en compte

- Taille du treillis (espace de recherche)
- Taille de la solution

Exemples de requêtes :

- Itemsets et leurs fréquences/clôtures : infaisable
- Itemsets fréquents et fréquences : parfois faisable
- Itemsets fréquents de taille 2 contenant `A` : faisable

Utiliser les relations entre contraintes et structure du treillis

Objectifs de la thèse

- Enjeux : permettre des extractions d'itemsets dans des cas très difficiles
- Moyens :
  - Généralisations des algorithmes existants
  - Intégration des représentations condensées (Doctorat de A. Bykowski [oct. 2002])

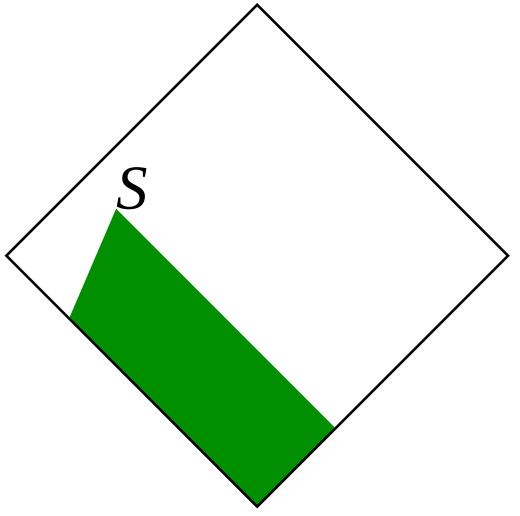
Plan

- Extraction de motifs sous contraintes [BDA 00]
- Extraction de représentations condensées sous contraintes : algorithme CoCo [IDEAS 01, journal IDA]
- Séquences de requêtes : algorithme iCoCo [PKDD 02]
- Conclusion et perspectives

Extraction de motifs sous contraintes

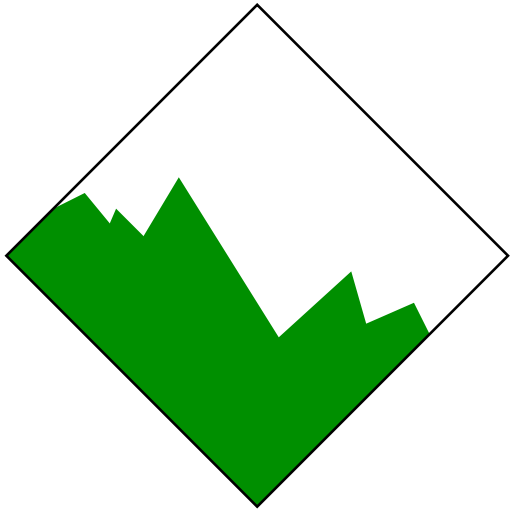
Contraintes

- $\mathcal{C}_{\mathsf{am}}$ est anti-monotone
  - $T \subseteq S$ et $\mathcal{C}_{\mathsf{am}}(S)$ alors $\mathcal{C}_{\mathsf{am}}(T)$
  - ex : $\mathcal{C}_{\gamma\text{-minfreq}}$, $S \subseteq \{\mathtt{ABC}\}$, $\mathcal{C}_{\mathsf{am}_1} \wedge \mathcal{C}_{\mathsf{am}_2}$, $\mathcal{C}_{\mathsf{am}_1} \vee \mathcal{C}_{\mathsf{am}_2}$, $\neg\mathcal{C}_{\mathsf{m}}$
- $\mathcal{C}_{\mathsf{m}}$ est monotone
  - $S \subseteq T$ et $\mathcal{C}_{\mathsf{m}}(S)$ alors $\mathcal{C}_{\mathsf{m}}(T)$
  - ex : $\mathcal{C}_{\gamma\text{-maxfreq}}$, $S \cap \{\mathtt{ABC}\} \neq \emptyset$, $\mathcal{C}_{\mathsf{m}_1} \wedge \mathcal{C}_{\mathsf{m}_2}$, $\mathcal{C}_{\mathsf{m}_1} \vee \mathcal{C}_{\mathsf{m}_2}$, $\neg\mathcal{C}_{\mathsf{am}}$



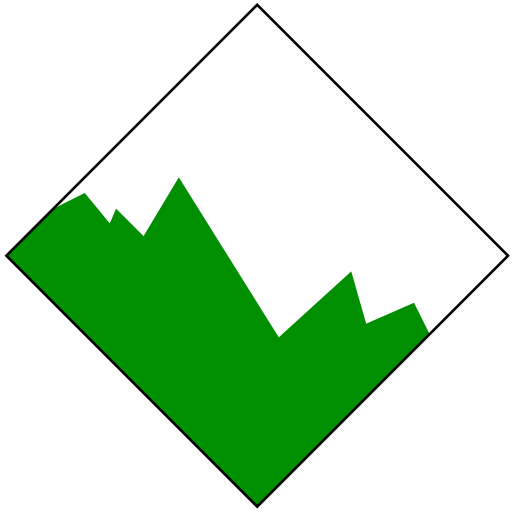
Contraintes

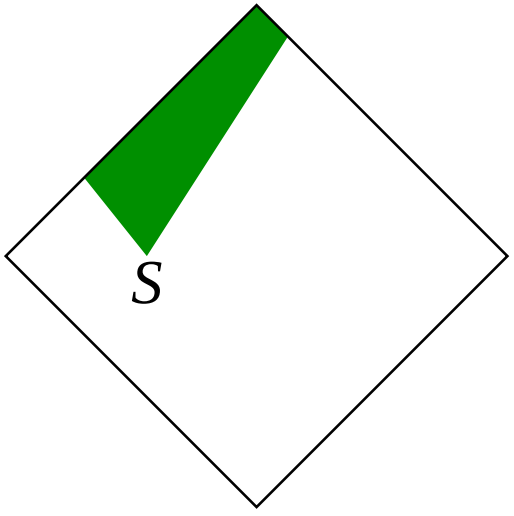
- $\mathcal{C}_{\mathsf{am}}$ est anti-monotone
  - $T \subseteq S$ et $\mathcal{C}_{\mathsf{am}}(S)$ alors $\mathcal{C}_{\mathsf{am}}(T)$
  - ex : $\mathcal{C}_{\gamma\text{-}\mathsf{minfreq}}$, $S \subseteq \{\mathtt{ABC}\}$, $\mathcal{C}_{\mathsf{am}_1} \wedge \mathcal{C}_{\mathsf{am}_2}$, $\mathcal{C}_{\mathsf{am}_1} \vee \mathcal{C}_{\mathsf{am}_2}$, $\neg\mathcal{C}_{\mathsf{m}}$
- $\mathcal{C}_{\mathsf{m}}$ est monotone
  - $S \subseteq T$ et $\mathcal{C}_{\mathsf{m}}(S)$ alors $\mathcal{C}_{\mathsf{m}}(T)$
  - ex : $\mathcal{C}_{\gamma\text{-}\mathsf{maxfreq}}$, $S \cap \{\mathtt{ABC}\} \neq \emptyset$, $\mathcal{C}_{\mathsf{m}_1} \wedge \mathcal{C}_{\mathsf{m}_2}$, $\mathcal{C}_{\mathsf{m}_1} \vee \mathcal{C}_{\mathsf{m}_2}$, $\neg\mathcal{C}_{\mathsf{am}}$



Contraintes

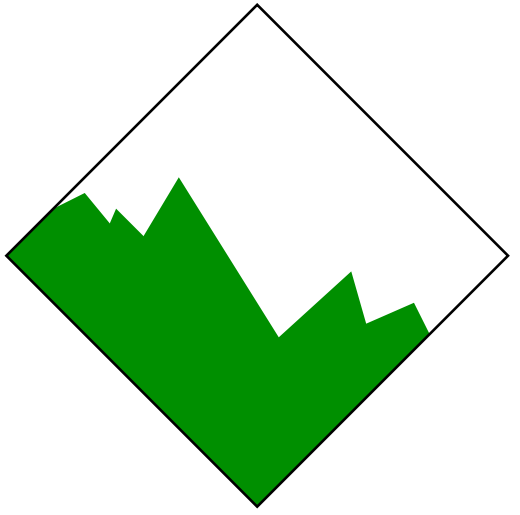
- $\mathcal{C}_{\mathsf{am}}$ est anti-monotone
  - $T \subseteq S$ et $\mathcal{C}_{\mathsf{am}}(S)$ alors $\mathcal{C}_{\mathsf{am}}(T)$
  - ex : $\mathcal{C}_{\gamma\text{-}\mathsf{minfreq}}$, $S \subseteq \{\mathtt{ABC}\}$, $\mathcal{C}_{\mathsf{am}_1} \wedge \mathcal{C}_{\mathsf{am}_2}$, $\mathcal{C}_{\mathsf{am}_1} \vee \mathcal{C}_{\mathsf{am}_2}$, $\neg\mathcal{C}_{\mathsf{m}}$
- $\mathcal{C}_{\mathsf{m}}$ est monotone
  - $S \subseteq T$ et $\mathcal{C}_{\mathsf{m}}(S)$ alors $\mathcal{C}_{\mathsf{m}}(T)$
  - ex : $\mathcal{C}_{\gamma\text{-}\mathsf{maxfreq}}$, $S \cap \{\mathtt{ABC}\} \neq \emptyset$, $\mathcal{C}_{\mathsf{m}_1} \wedge \mathcal{C}_{\mathsf{m}_2}$, $\mathcal{C}_{\mathsf{m}_1} \vee \mathcal{C}_{\mathsf{m}_2}$, $\neg\mathcal{C}_{\mathsf{am}}$

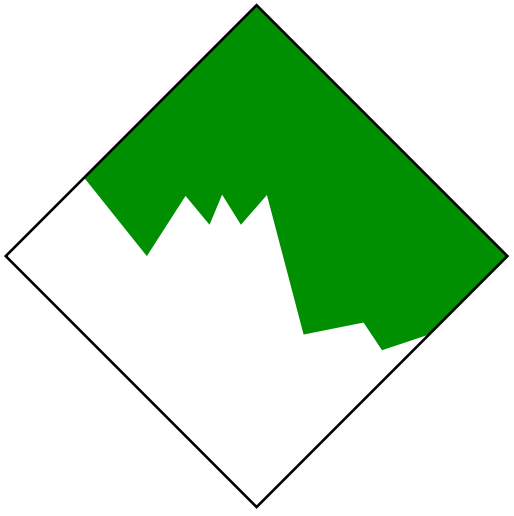




Contraintes

- $\mathcal{C}_{\mathsf{am}}$ est anti-monotone
  - $T \subseteq S$ et $\mathcal{C}_{\mathsf{am}}(S)$ alors $\mathcal{C}_{\mathsf{am}}(T)$
  - ex : $\mathcal{C}_{\gamma\text{-minfreq}}$, $S \subseteq \{\mathtt{ABC}\}$, $\mathcal{C}_{\mathsf{am}_1} \wedge \mathcal{C}_{\mathsf{am}_2}$, $\mathcal{C}_{\mathsf{am}_1} \vee \mathcal{C}_{\mathsf{am}_2}$, $\neg\mathcal{C}_{\mathsf{m}}$
- $\mathcal{C}_{\mathsf{m}}$ est monotone
  - $S \subseteq T$ et $\mathcal{C}_{\mathsf{m}}(S)$ alors $\mathcal{C}_{\mathsf{m}}(T)$
  - ex : $\mathcal{C}_{\gamma\text{-maxfreq}}$, $S \cap \{\mathtt{ABC}\} \neq \emptyset$, $\mathcal{C}_{\mathsf{m}_1} \wedge \mathcal{C}_{\mathsf{m}_2}$, $\mathcal{C}_{\mathsf{m}_1} \vee \mathcal{C}_{\mathsf{m}_2}$, $\neg\mathcal{C}_{\mathsf{am}}$



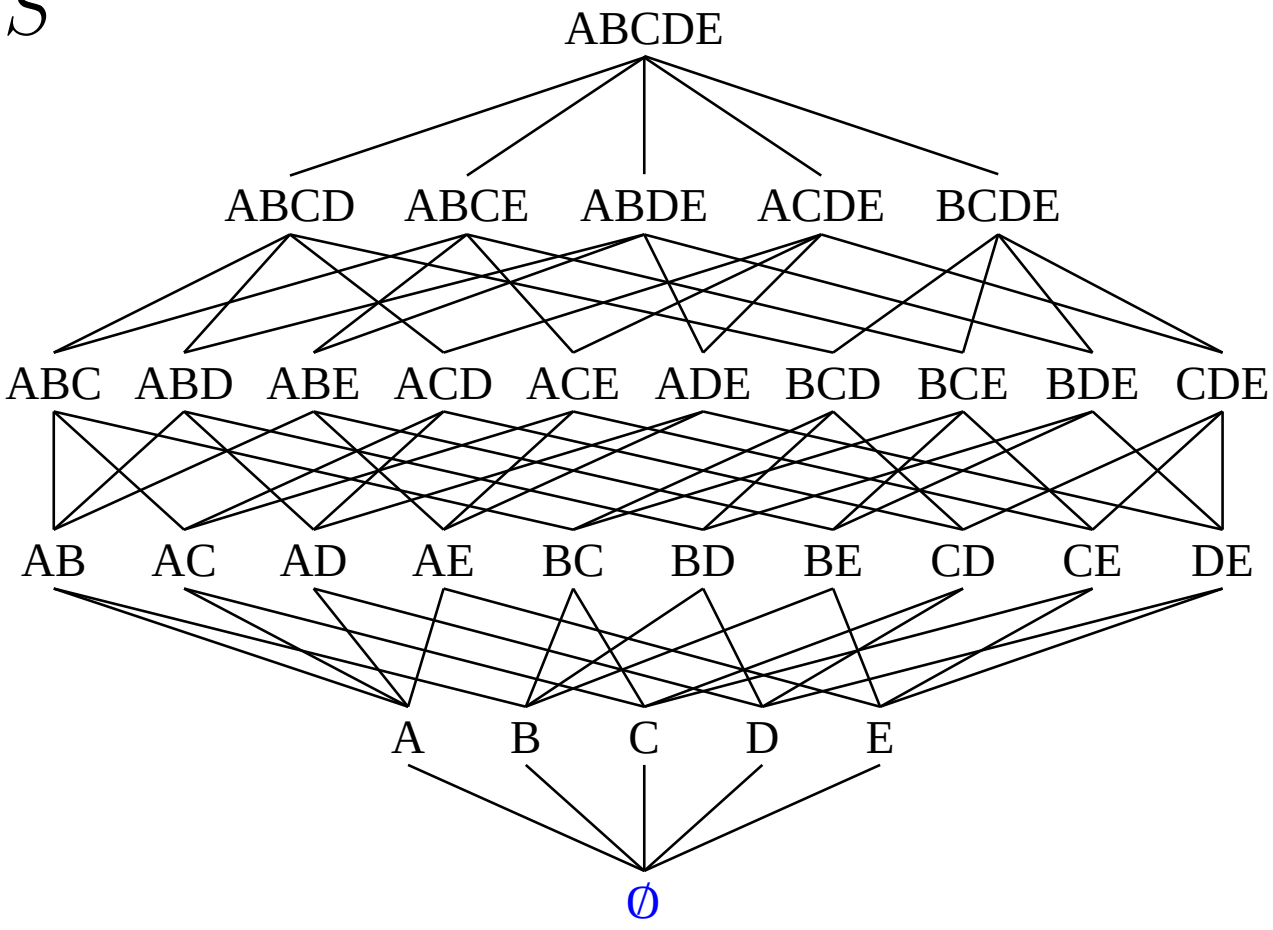


Exploitation des contraintes anti-monotones

Algorithme niveau par niveau dans le cas de contraintes anti-monotones.

Ex: $\mathcal{C}_{2\text{-minfreq}} \wedge \mathtt{E} \notin S$

tid	
1	ABCDE
2	ABC
3	ACD
4	ABE
5	CD
6	CE

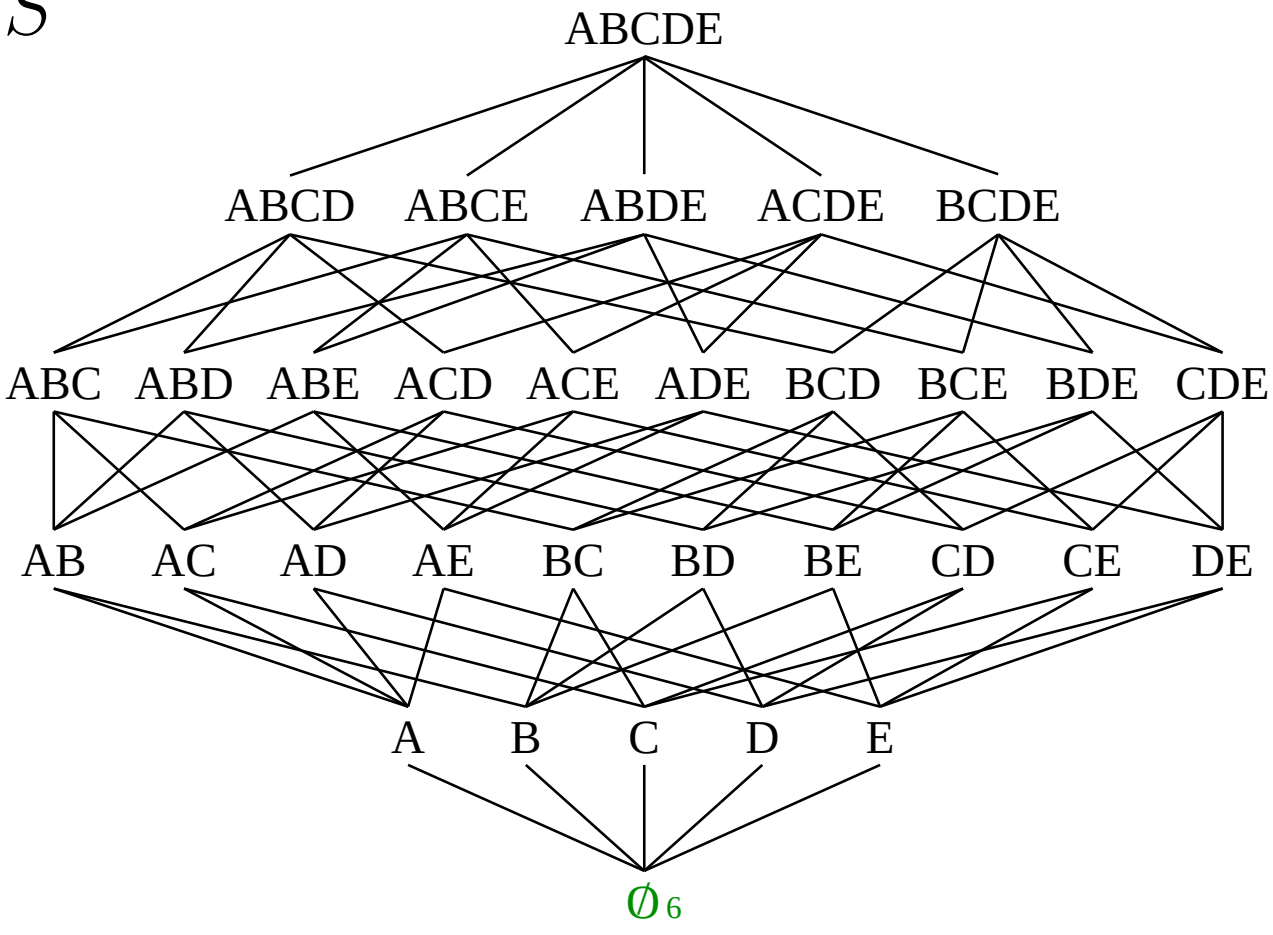


Exploitation des contraintes anti-monotones

Algorithme niveau par niveau dans le cas de contraintes anti-monotones.

Ex: $\mathcal{C}_{2\text{-minfreq}} \wedge \mathtt{E} \notin S$

tid	
1	ABCDE
2	ABC
3	ACD
4	ABE
5	CD
6	CE

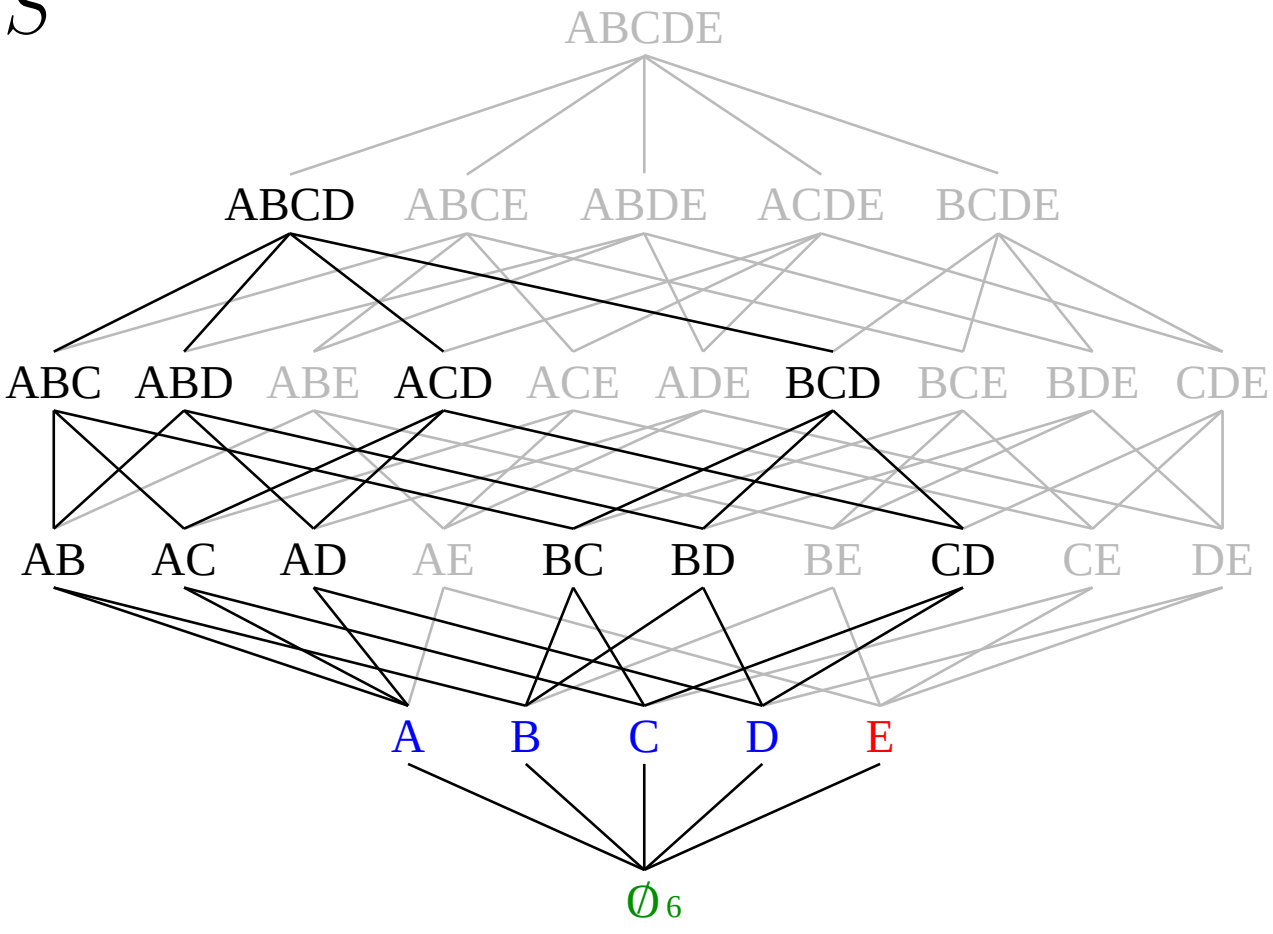


Exploitation des contraintes anti-monotones

Algorithme niveau par niveau dans le cas de contraintes anti-monotones.

Ex: $\mathcal{C}_{2\text{-minfreq}} \wedge \mathtt{E} \notin S$

tid	
1	ABCDE
2	ABC
3	ACD
4	ABE
5	CD
6	CE

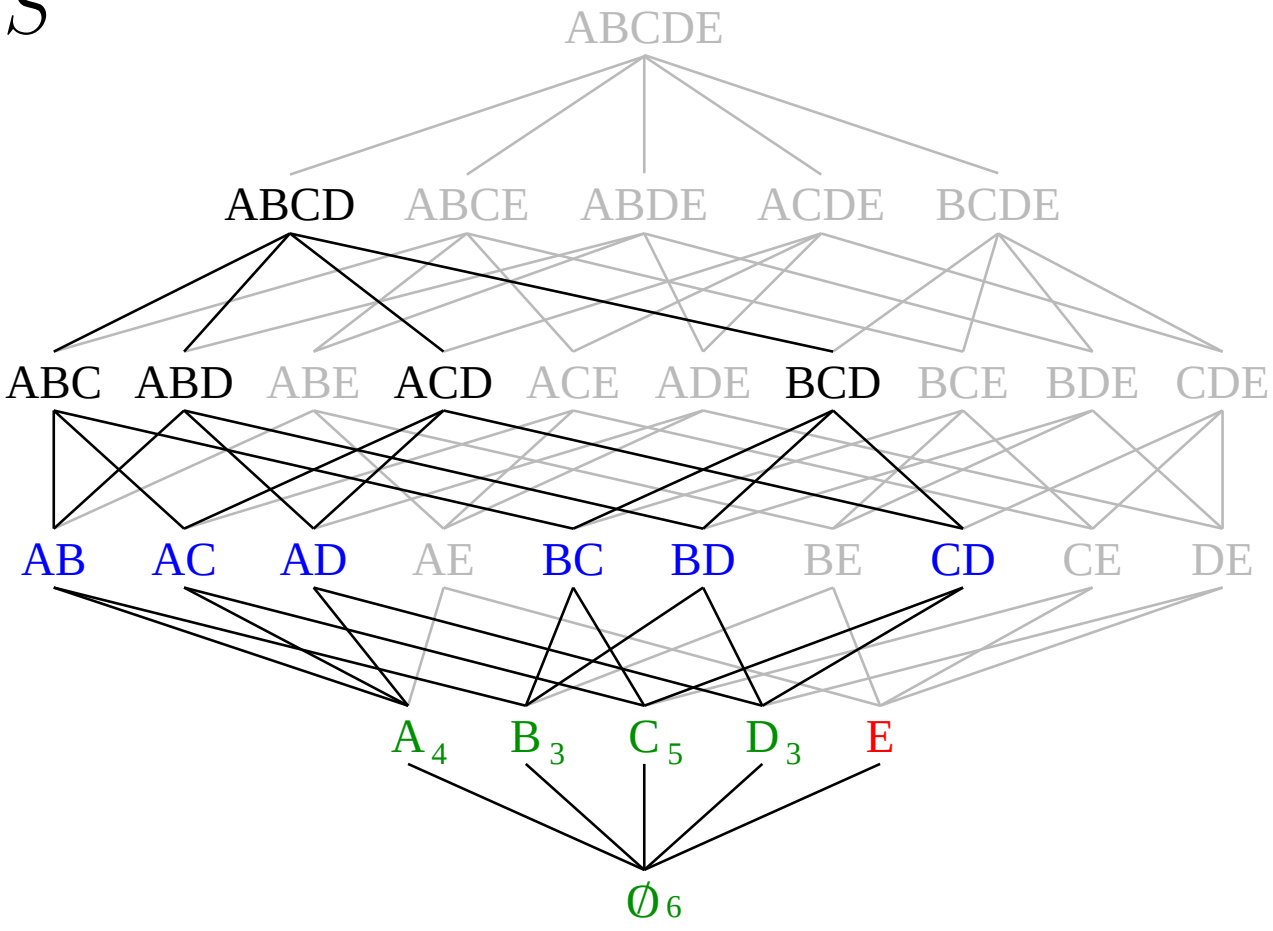


Exploitation des contraintes anti-monotones

Algorithme niveau par niveau dans le cas de contraintes anti-monotones.

Ex: $\mathcal{C}_{2\text{-minfreq}} \wedge \mathtt{E} \notin S$

tid	
1	ABCDE
2	ABC
3	ACD
4	ABE
5	CD
6	CE

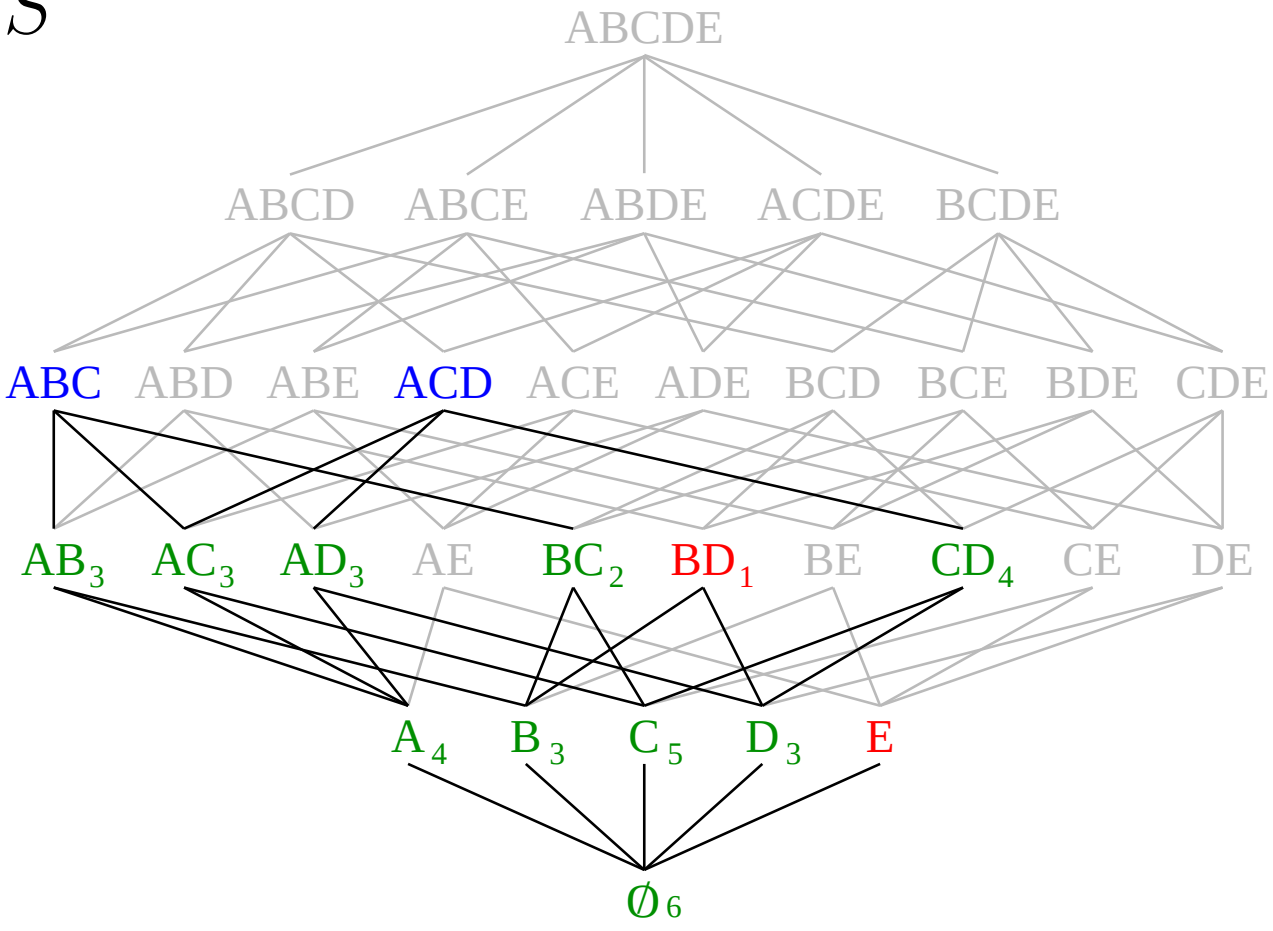


Exploitation des contraintes anti-monotones

Algorithme niveau par niveau dans le cas de contraintes anti-monotones.

Ex: $\mathcal{C}_{2\text{-minfreq}} \wedge \mathtt{E} \notin S$

tid	
1	ABCDE
2	ABC
3	ACD
4	ABE
5	CD
6	CE

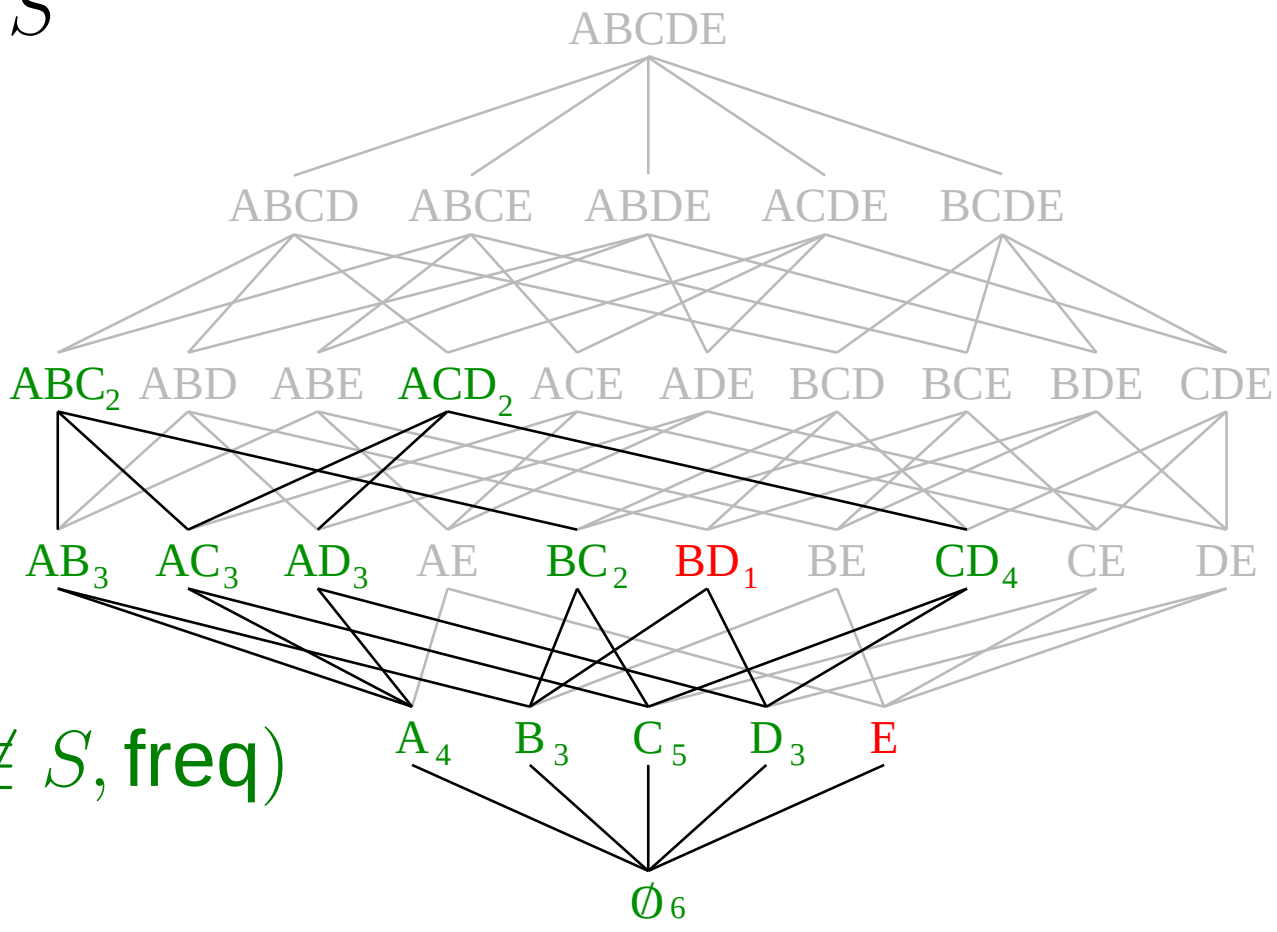


Exploitation des contraintes anti-monotones

Algorithme niveau par niveau dans le cas de contraintes anti-monotones.

Ex: $\mathcal{C}_{2\text{-minfreq}} \wedge \mathtt{E} \notin S$

tid	
1	ABCDE
2	ABC
3	ACD
4	ABE
5	CD
6	CE



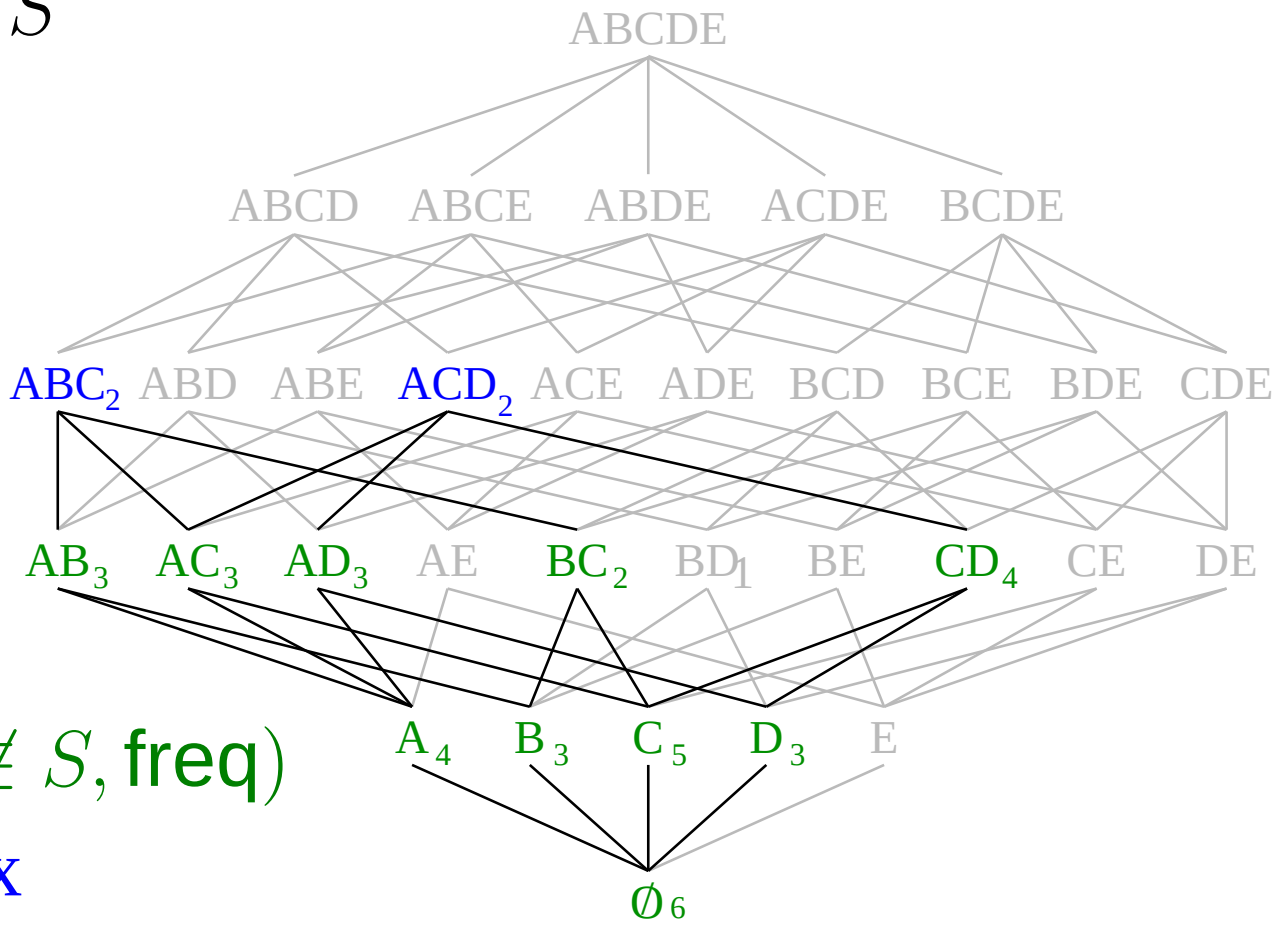
$\mathrm{Th}^{+}(\mathcal{C}_{2\text{-minfreq}} \wedge \mathtt{E} \notin S, \mathrm{freq})$

Exploitation des contraintes anti-monotones

Algorithme niveau par niveau dans le cas de contraintes anti-monotones.

Ex: $\mathcal{C}_{2\text{-minfreq}} \wedge \mathtt{E} \notin S$

tid	
1	ABCDE
2	ABC
3	ACD
4	ABE
5	CD
6	CE

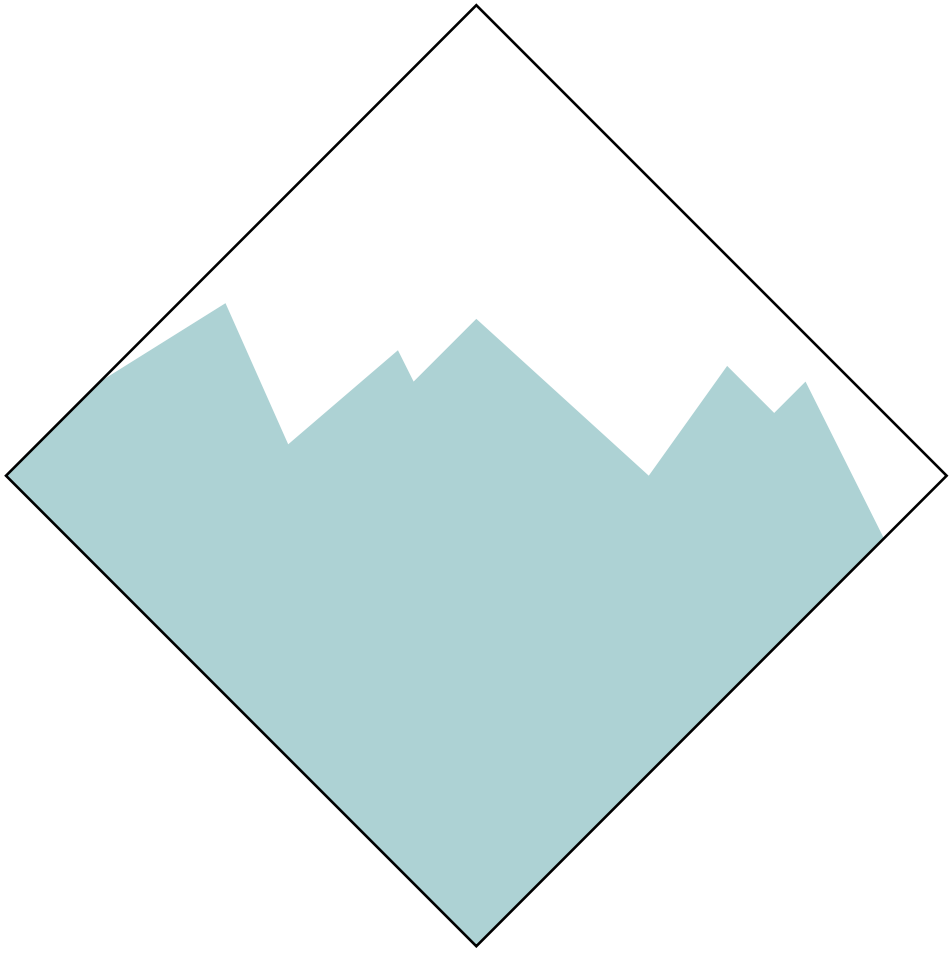


$\mathsf{Th}^{+}(\mathcal{C}_{2\text{-minfreq}} \wedge \mathtt{E} \notin S, \mathsf{freq})$

Éléments maximaux

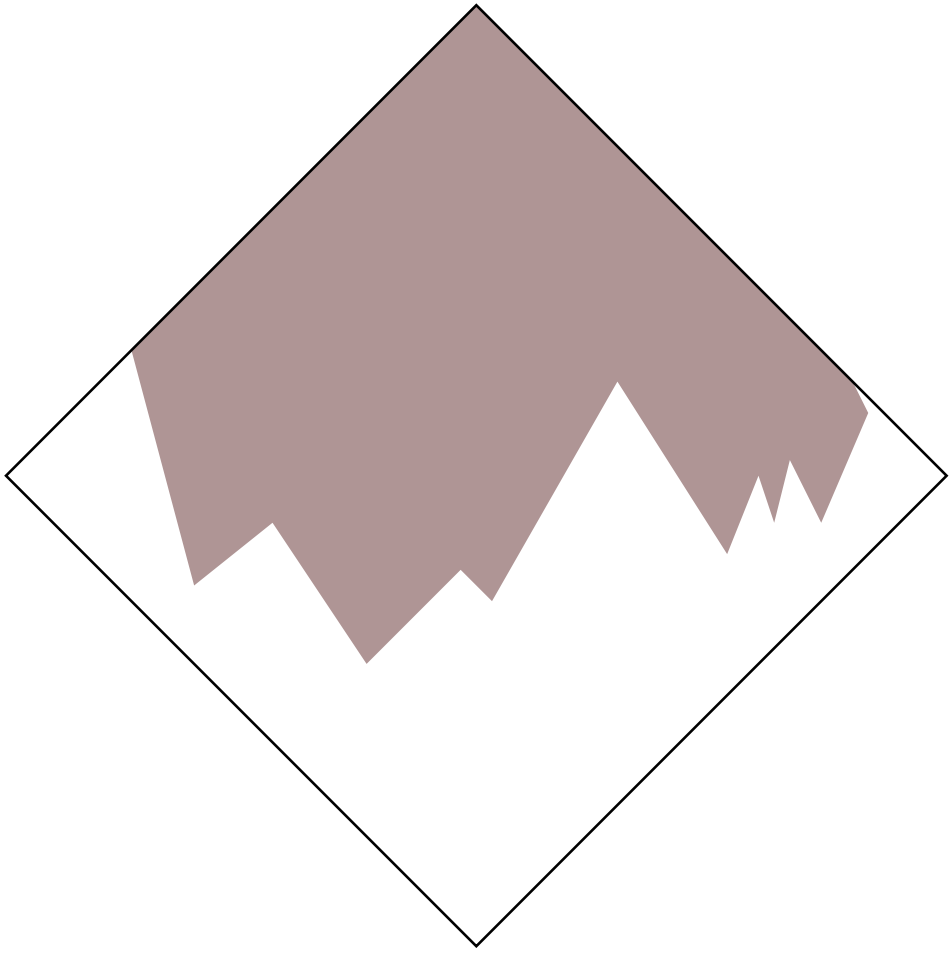
Structure de $\mathsf{Th}(\mathcal{C}_{\mathbf{am}} \wedge \mathcal{C}_{\mathbf{m}})$

$\mathsf{Th}(\mathcal{C}_{\mathsf{am}})$



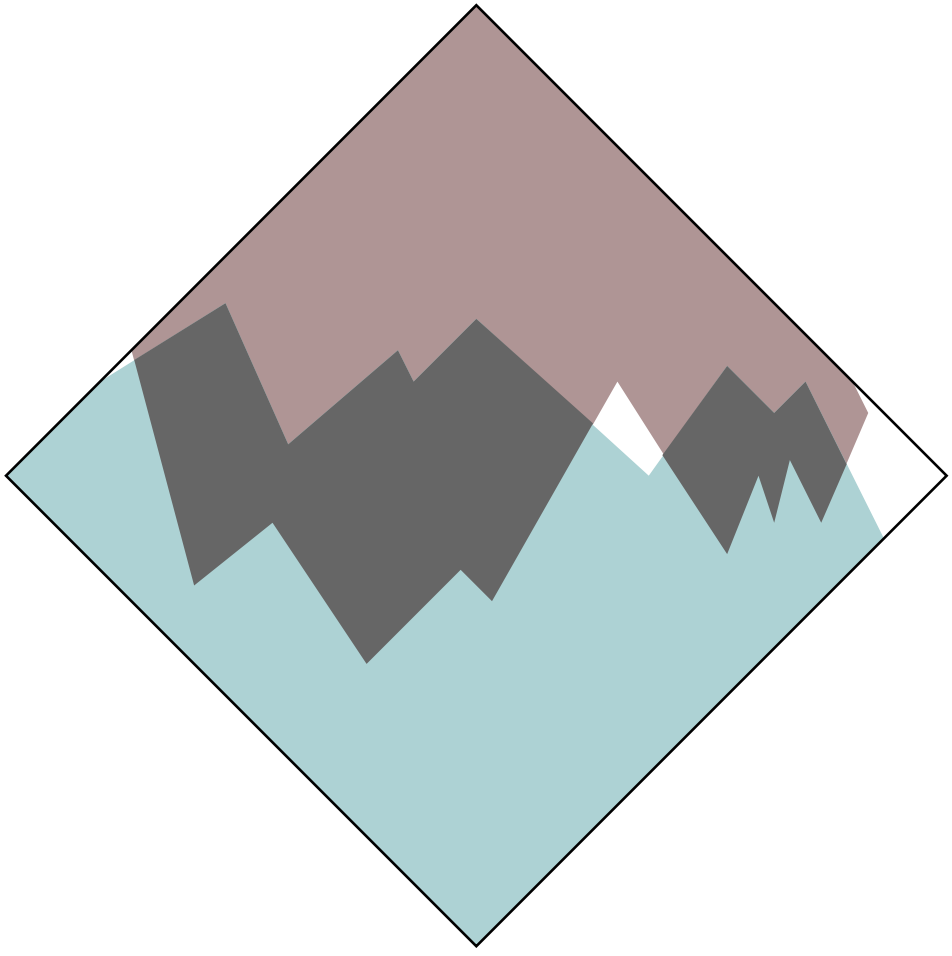
Structure de $\mathsf{Th}(\mathcal{C}_{\mathbf{am}} \wedge \mathcal{C}_{\mathbf{m}})$

$\mathsf{Th}(\mathcal{C}_{\mathsf{m}})$



Structure de $\mathsf{Th}(\mathcal{C}_{\mathsf{am}} \wedge \mathcal{C}_{\mathsf{m}})$

$\mathsf{Th}(\mathcal{C}_{\mathsf{am}} \wedge \mathcal{C}_{\mathsf{m}})$

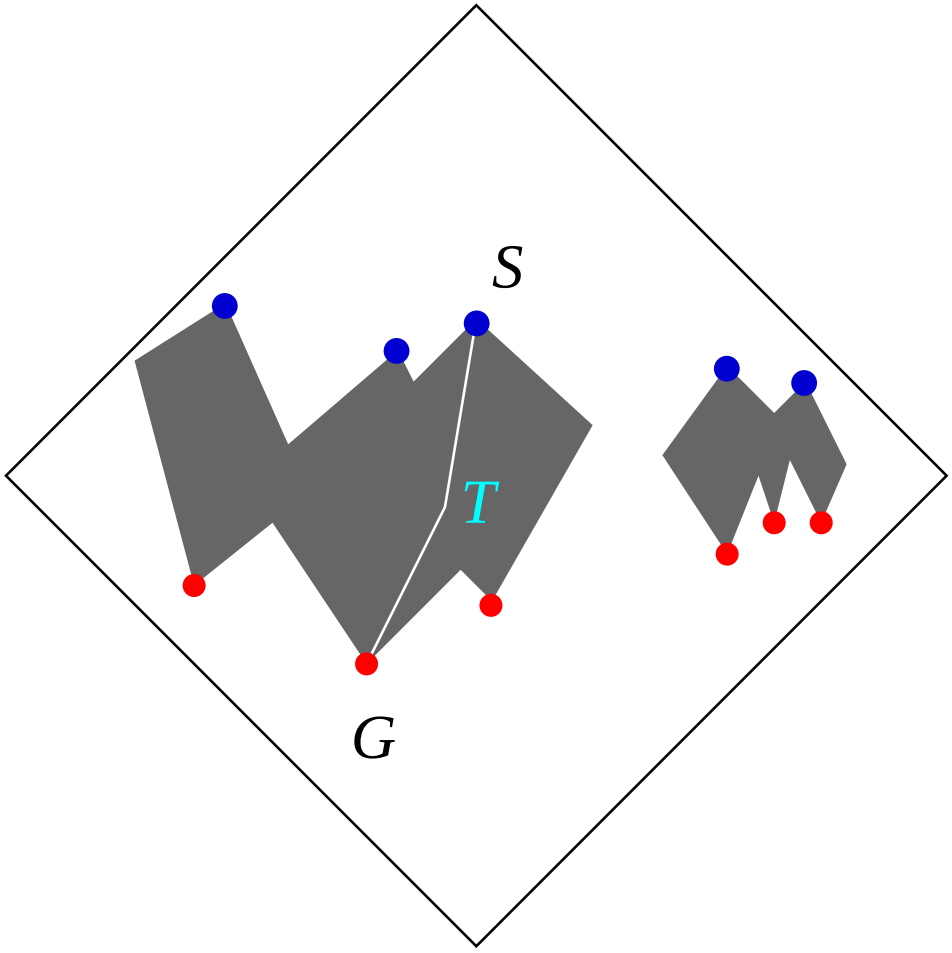


Structure de $\mathsf{Th}(\mathcal{C}_{\mathsf{am}} \wedge \mathcal{C}_{\mathsf{m}})$

$\mathsf{Th}(\mathcal{C}_{\mathsf{am}} \wedge \mathcal{C}_{\mathsf{m}})$

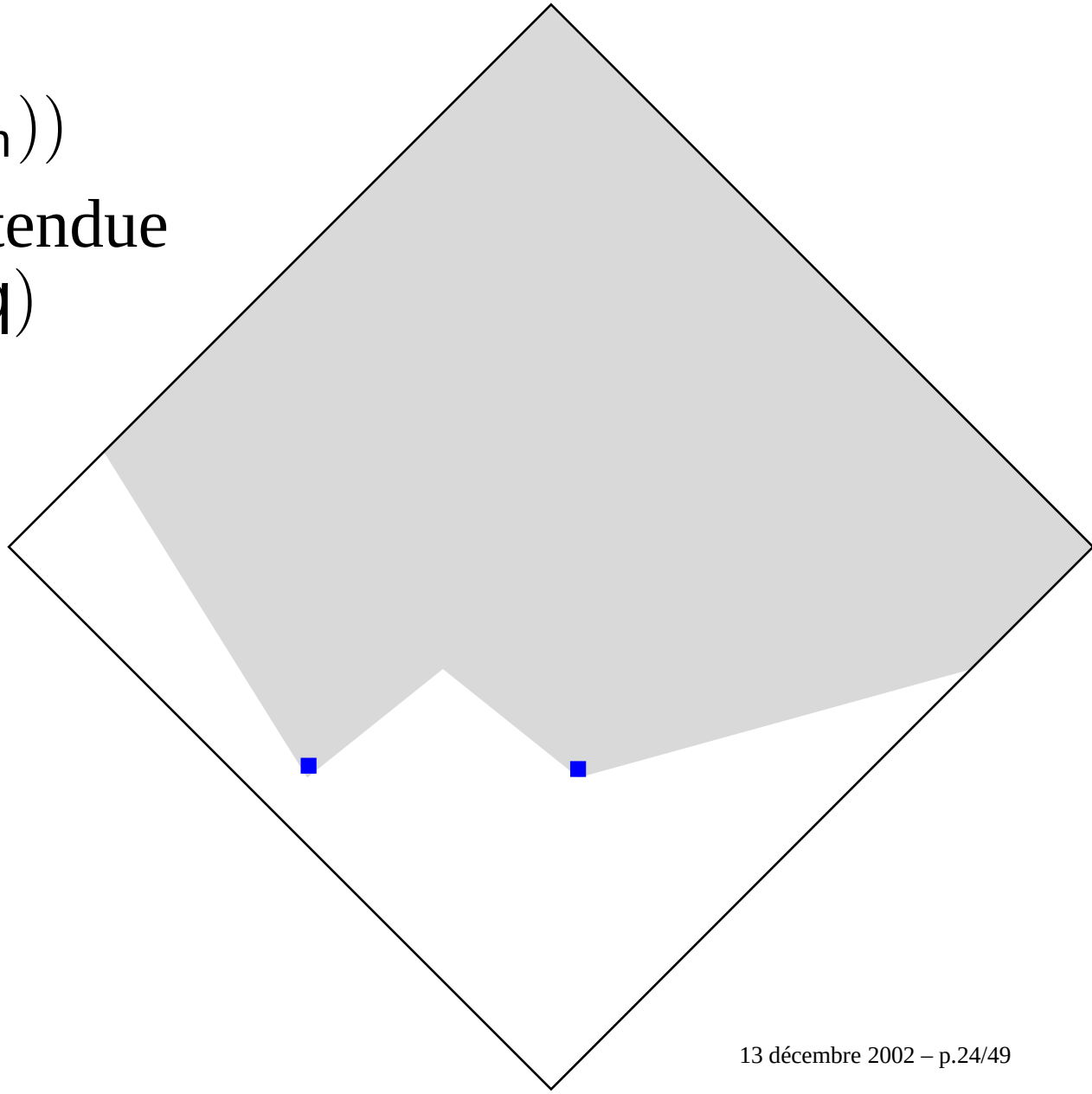
- $\mathsf{S} = \max_{\subseteq} \mathsf{Th}(\mathcal{C}_{\mathsf{am}} \wedge \mathcal{C}_{\mathsf{m}})$
- $\mathsf{G} = \min_{\subseteq} \mathsf{Th}(\mathcal{C}_{\mathsf{am}} \wedge \mathcal{C}_{\mathsf{m}})$
- Caractérisation :
  $S \in \mathsf{S}$
  $G \in \mathsf{G}$
  $G \subseteq T \subseteq S$

Espace convexe ou
Espace des versions



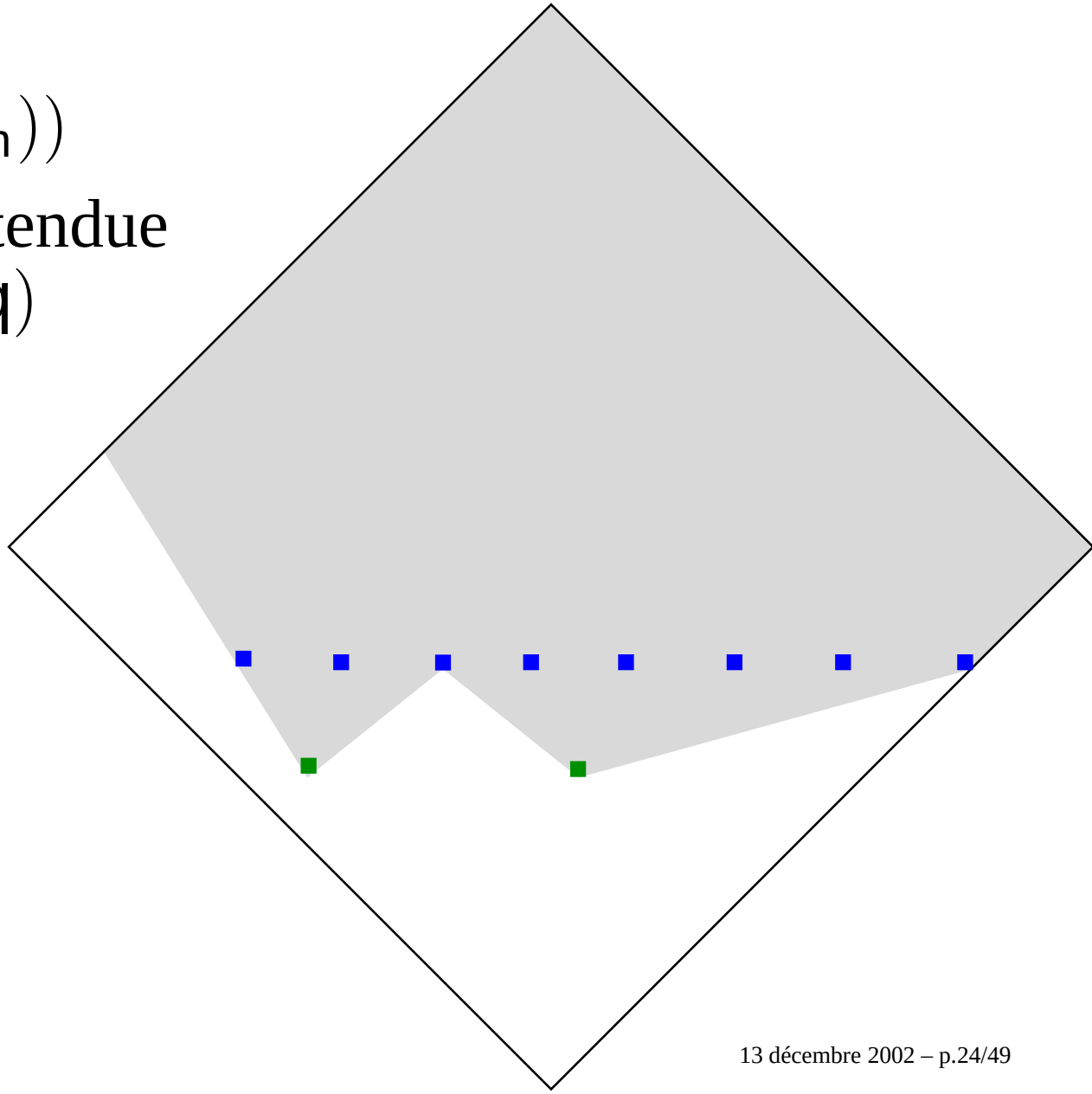
Calcul de $\mathsf{Th}^+(\mathcal{C}_{\mathsf{am}} \wedge \mathcal{C}_{\mathsf{m}}, \mathsf{freq})$ [BDA 00]

- Niveau par niveau
  à partir de $\mathsf{G}(\mathsf{Th}(\mathcal{C}_{\mathsf{m}}))$
- Calcule la théorie étendue
  $\mathsf{Th}^+(\mathcal{C}_{\mathsf{am}} \wedge \mathcal{C}_{\mathsf{m}}, \mathsf{freq})$



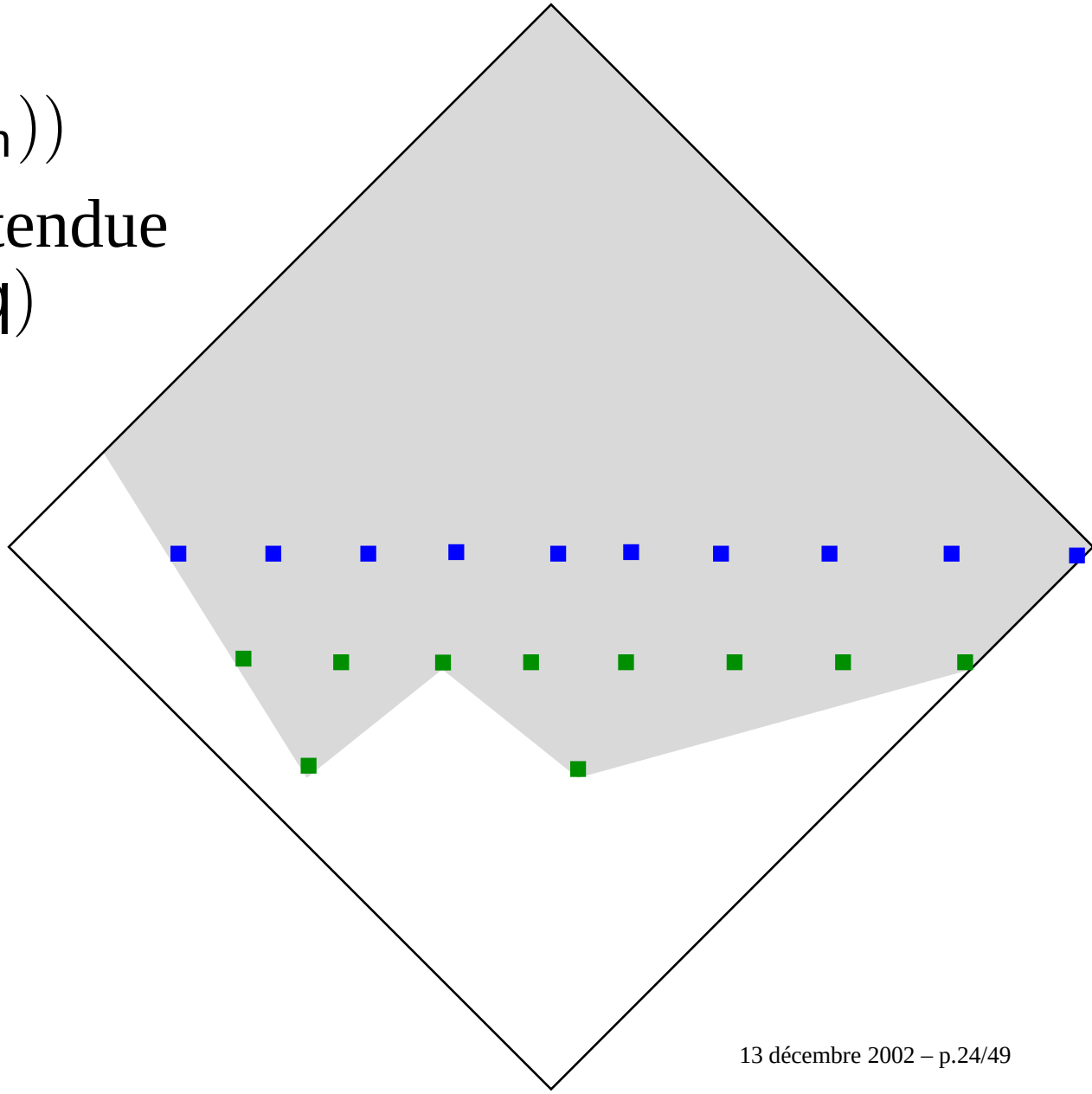
Calcul de $\mathsf{Th}^+(\mathcal{C}_{\mathsf{am}} \wedge \mathcal{C}_{\mathsf{m}}, \mathsf{freq})$ [BDA 00]

- Niveau par niveau
  à partir de $\mathsf{G}(\mathsf{Th}(\mathcal{C}_{\mathsf{m}}))$
- Calcule la théorie étendue
  $\mathsf{Th}^+(\mathcal{C}_{\mathsf{am}} \wedge \mathcal{C}_{\mathsf{m}}, \mathsf{freq})$



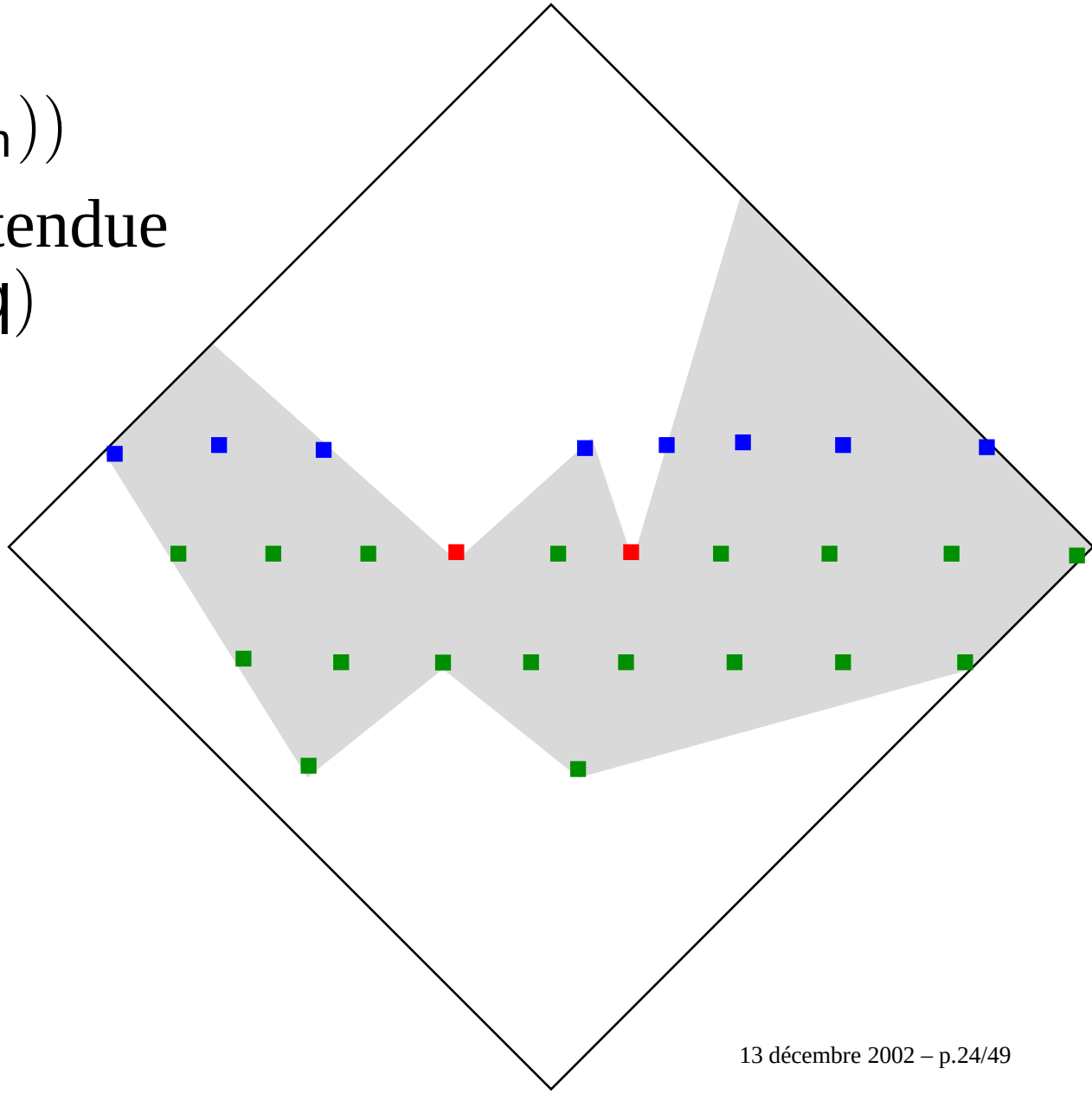
Calcul de $\mathsf{Th}^+(\mathcal{C}_{\mathsf{am}} \wedge \mathcal{C}_{\mathsf{m}}, \mathsf{freq})$ [BDA 00]

- Niveau par niveau
  à partir de $\mathsf{G}(\mathsf{Th}(\mathcal{C}_{\mathsf{m}}))$
- Calcule la théorie étendue
  $\mathsf{Th}^+(\mathcal{C}_{\mathsf{am}} \wedge \mathcal{C}_{\mathsf{m}}, \mathsf{freq})$



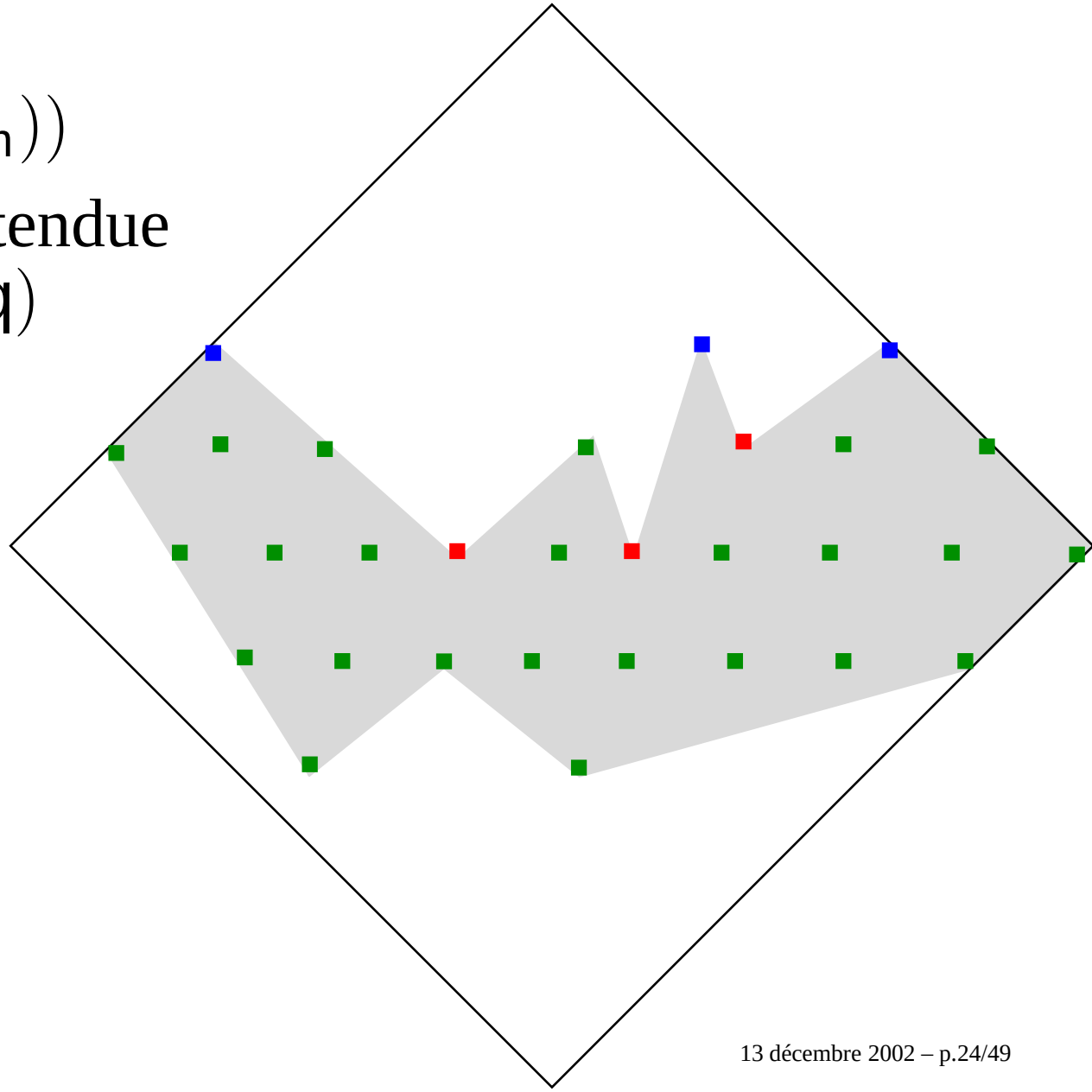
Calcul de $\mathsf{Th}^+(\mathcal{C}_{\mathsf{am}} \wedge \mathcal{C}_{\mathsf{m}}, \mathsf{freq})$ [BDA 00]

- Niveau par niveau
  à partir de $\mathsf{G}(\mathsf{Th}(\mathcal{C}_{\mathsf{m}}))$
- Calcule la théorie étendue
  $\mathsf{Th}^+(\mathcal{C}_{\mathsf{am}} \wedge \mathcal{C}_{\mathsf{m}}, \mathsf{freq})$



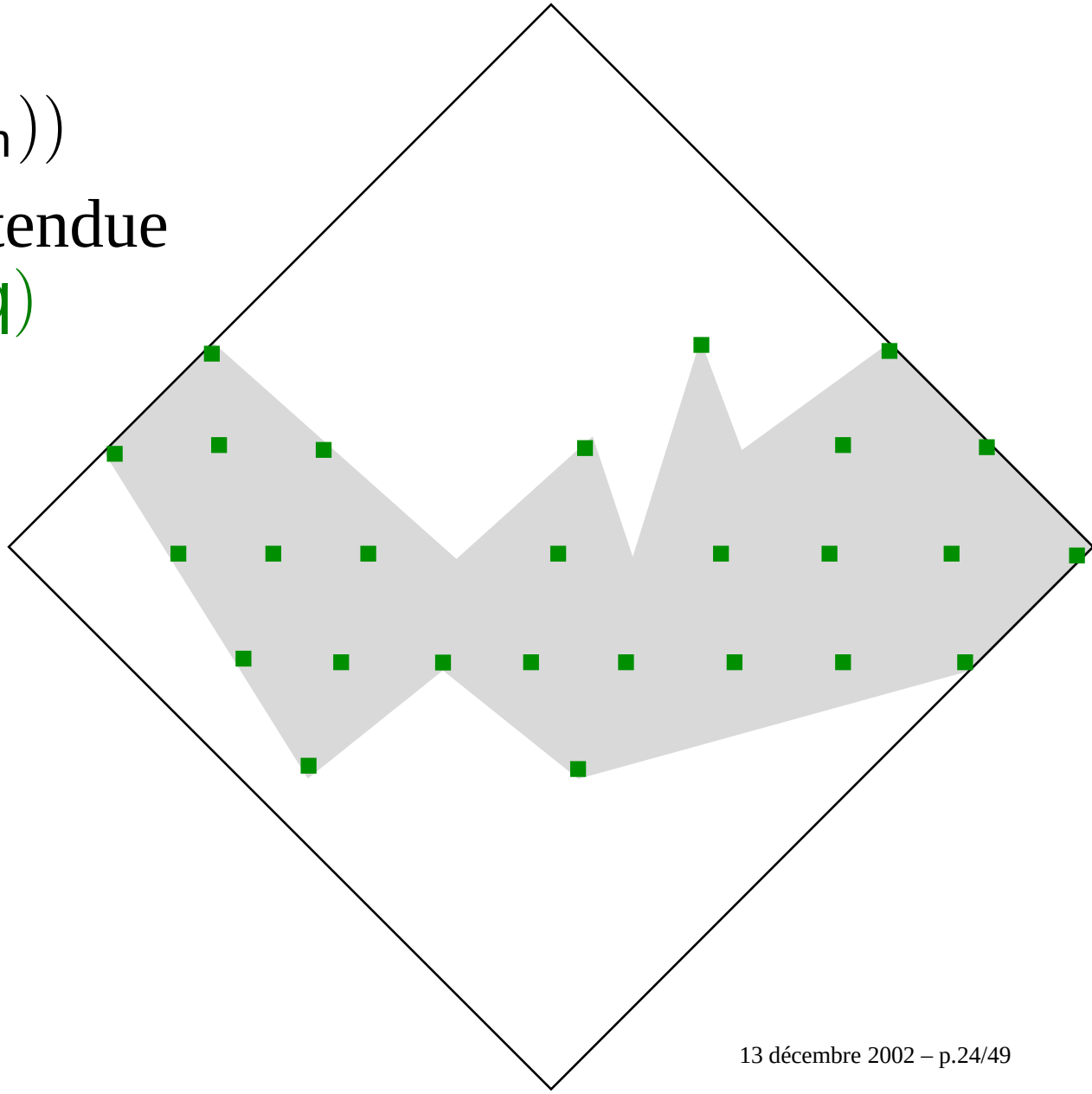
Calcul de $\mathsf{Th}^+(\mathcal{C}_{\mathsf{am}} \wedge \mathcal{C}_{\mathsf{m}}, \mathsf{freq})$ [BDA 00]

- Niveau par niveau
  à partir de $\mathsf{G}(\mathsf{Th}(\mathcal{C}_{\mathsf{m}}))$
- Calcule la théorie étendue
  $\mathsf{Th}^+(\mathcal{C}_{\mathsf{am}} \wedge \mathcal{C}_{\mathsf{m}}, \mathsf{freq})$



Calcul de $\mathsf{Th}^+(\mathcal{C}_{\mathsf{am}} \wedge \mathcal{C}_{\mathsf{m}}, \mathsf{freq})$ [BDA 00]

- Niveau par niveau
  à partir de $\mathsf{G}(\mathsf{Th}(\mathcal{C}_{\mathsf{m}}))$
- Calcule la théorie étendue
  $\mathsf{Th}^+(\mathcal{C}_{\mathsf{am}} \wedge \mathcal{C}_{\mathsf{m}}, \mathsf{freq})$

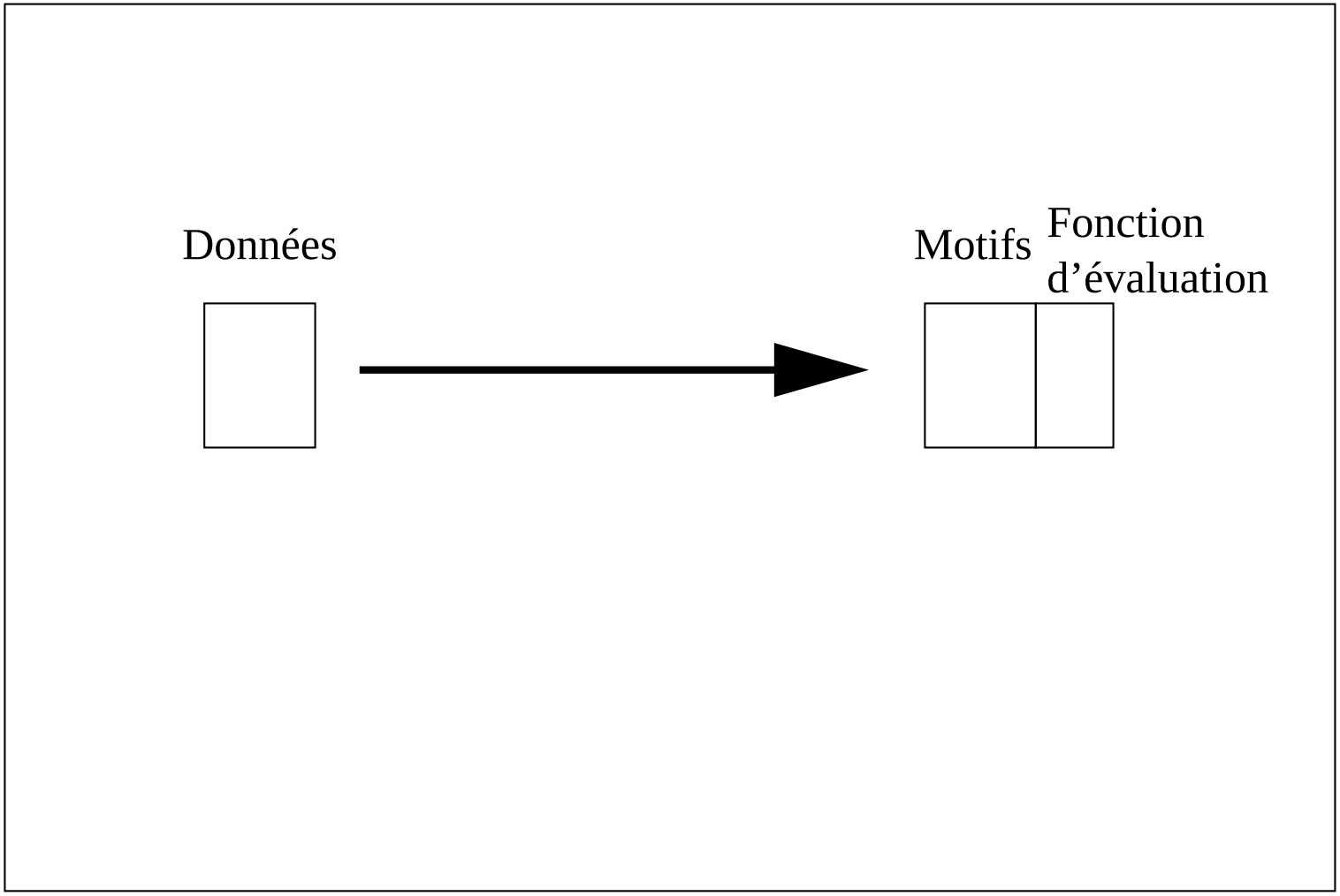


Limitations

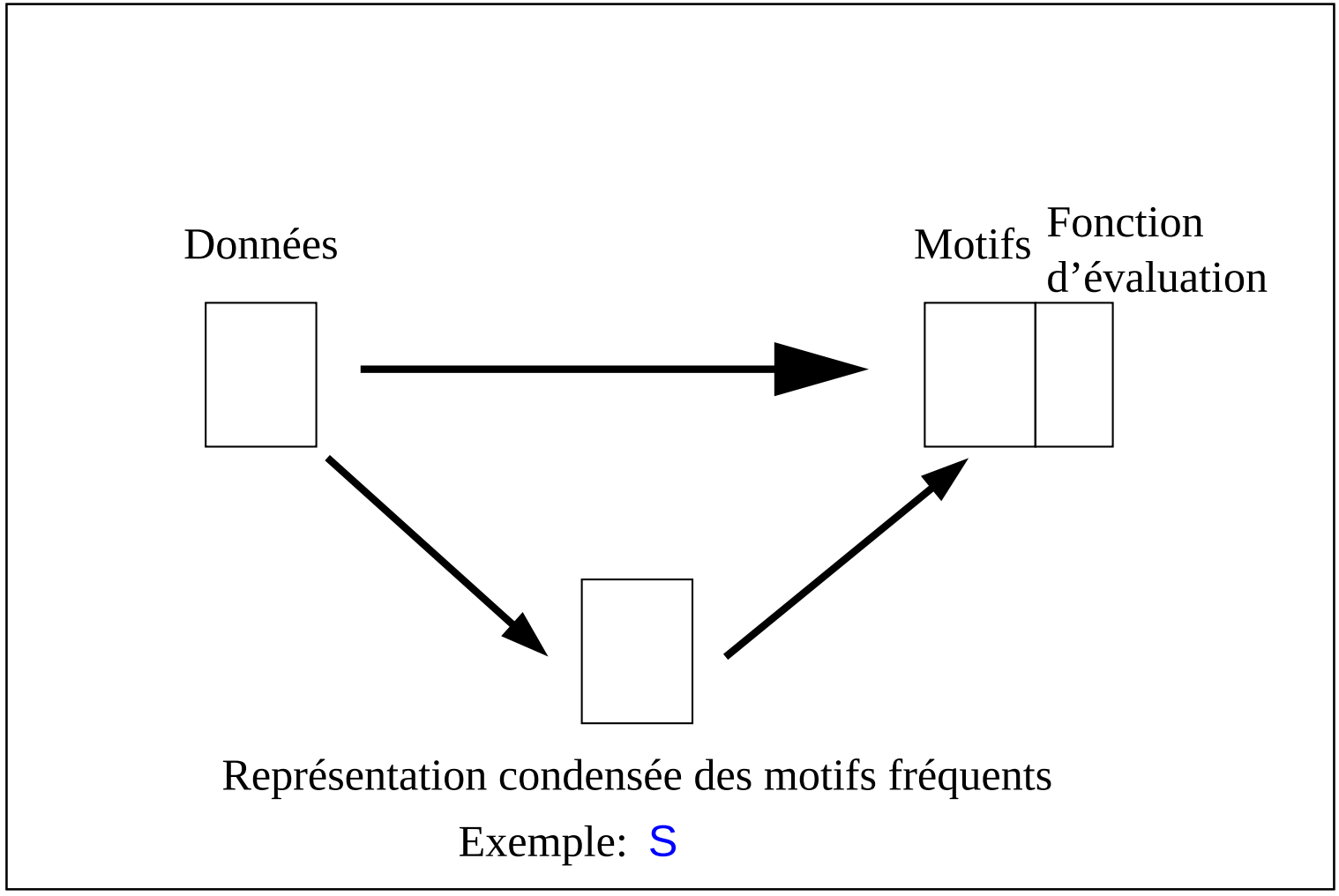
- Décomposition $\mathcal{C} = \mathcal{C}_{\mathsf{am}} \wedge \mathcal{C}_{\mathsf{m}} \wedge \mathcal{C}_{autre}$
- Lorsque $\mathsf{Th}(\mathcal{C}_{\mathsf{am}} \wedge \mathcal{C}_{\mathsf{m}})$ est trop grand
  - si $\mathbf{r}$ est dense/corrélée
- Faut-il “pousser” $\mathcal{C}_{\mathsf{m}}$?
  Ex : $\mathcal{C}(S) = \mathcal{C}_{\gamma\text{-}\mathsf{minfreq}}(S) \wedge (|S| > 2)$

Extraction de représentations condensées sous contraintes

Représentations condensées des motifs fréquents

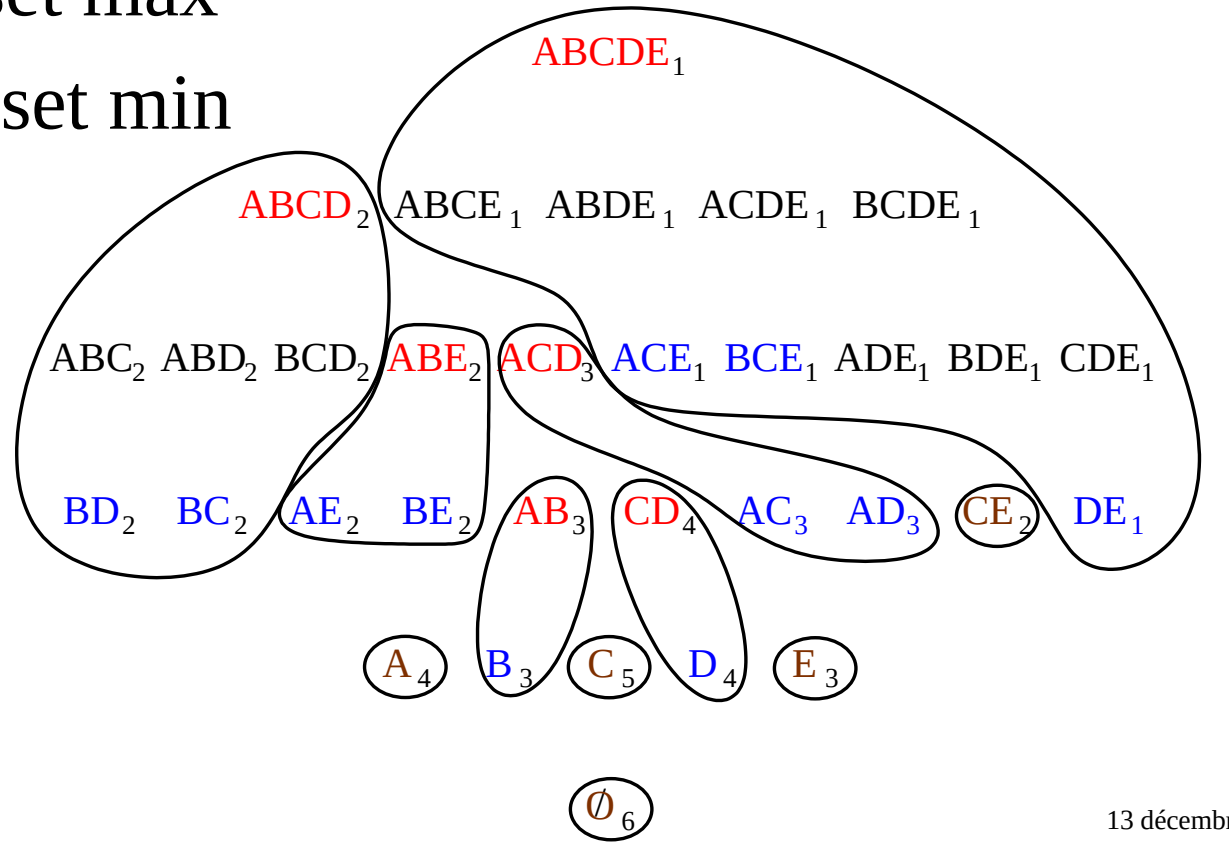


Représentations condensées des motifs fréquents



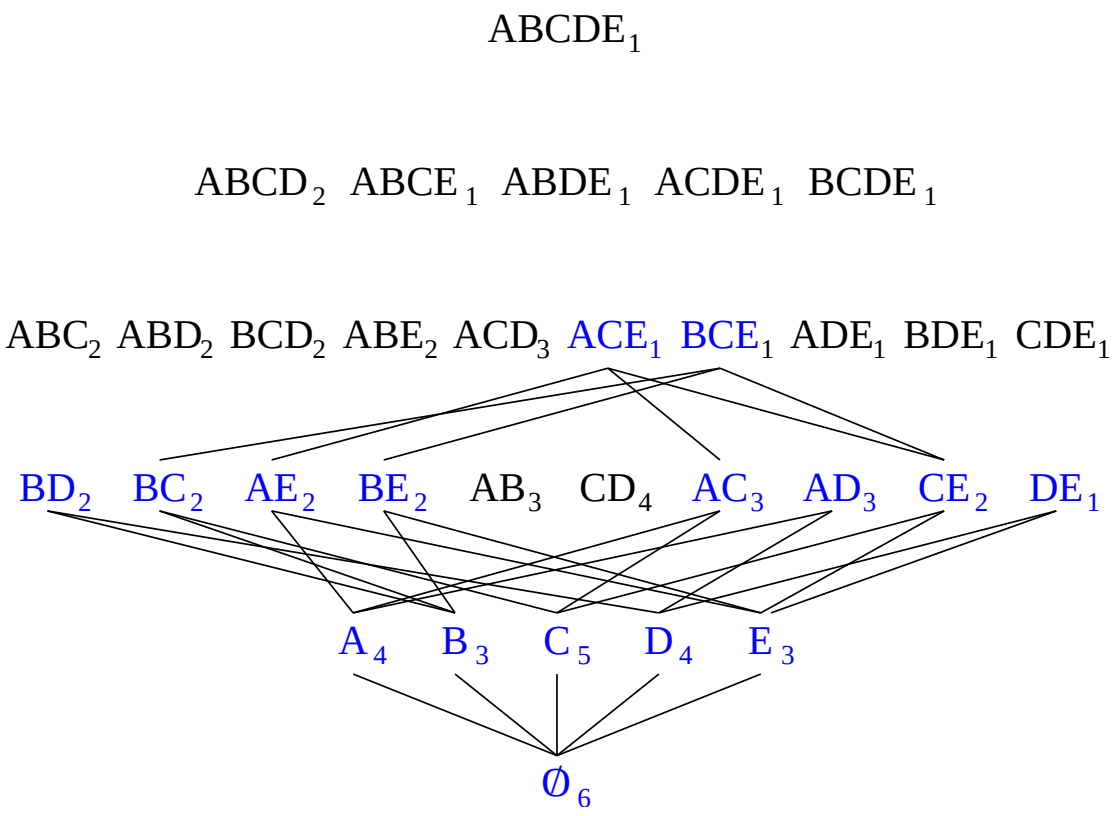
Itemsets libres et clos

- $\mathsf{cl}(S)$: plus grand sur ensemble F de S tel que $\mathsf{freq}(F) = \mathsf{freq}(S)$
- relation $V \sim W \Leftrightarrow \mathsf{cl}(V) = \mathsf{cl}(W)$
- Clos : itemset max
- Libre : itemset min



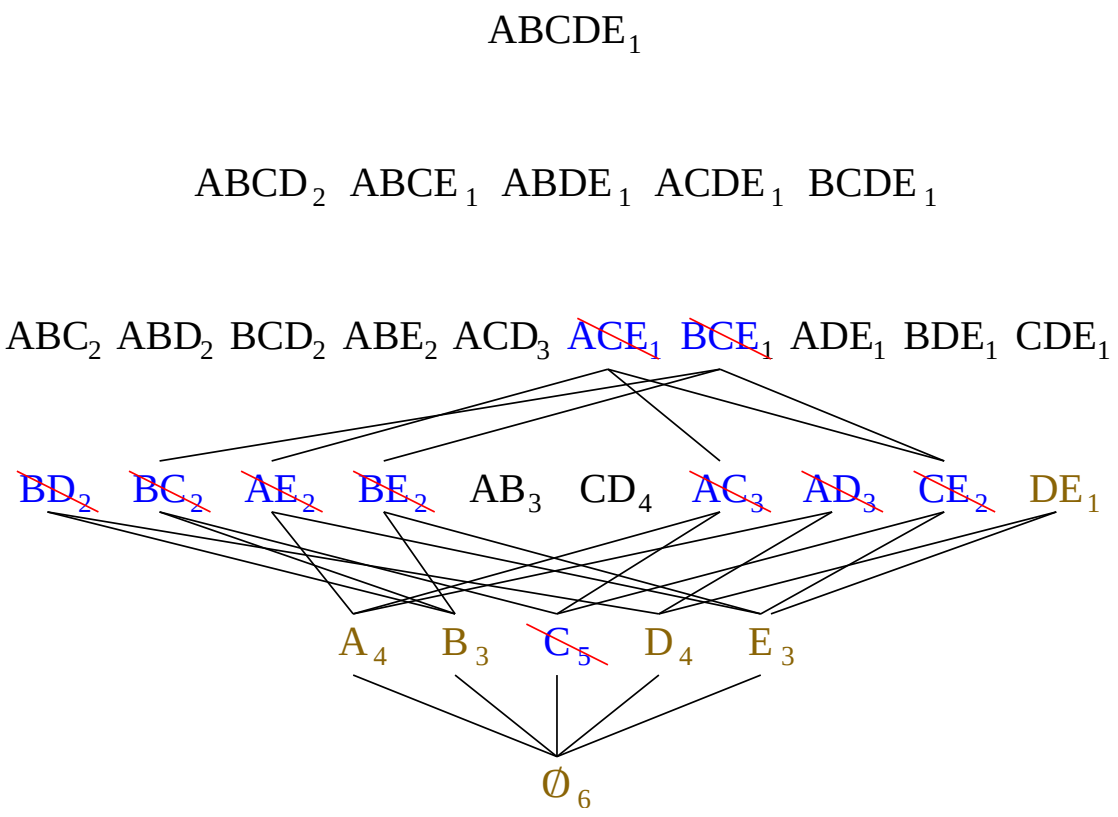
Itemsets δ-libres

- S est libre : $\forall T \subset S,\ \mathsf{freq}(T) \neq \mathsf{freq}(S)$
- Généralisation avec δ entier positif
  S est δ-libre : $\forall T \subset S,\ \mathsf{freq}(T) - \mathsf{freq}(S) > \delta$
  ex : 1-libres



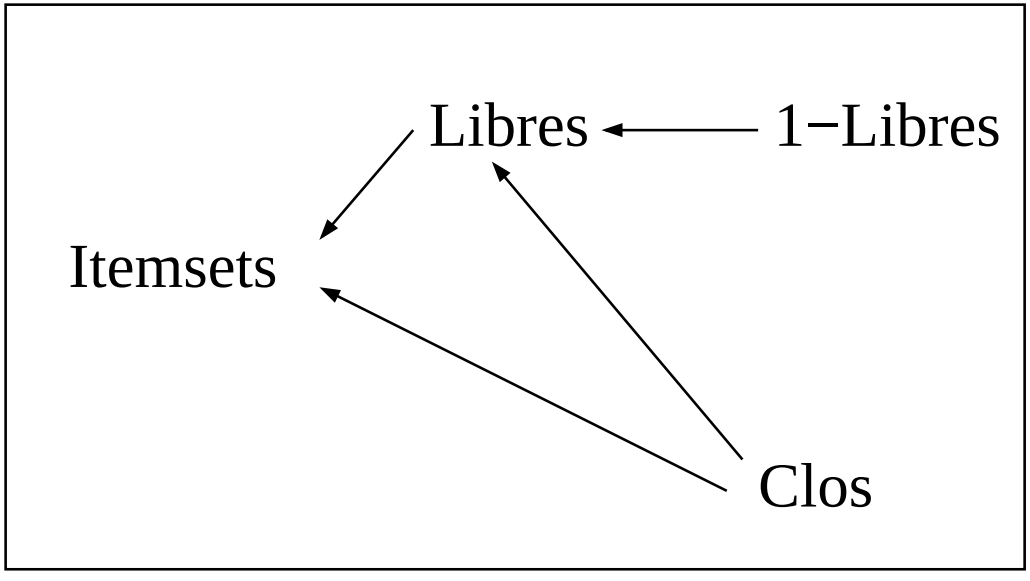
Itemsets δ-libres

- S est libre : $\forall T \subset S,\ \mathsf{freq}(T) \neq \mathsf{freq}(S)$
- Généralisation avec δ entier positif
  S est δ-libre : $\forall T \subset S,\ \mathsf{freq}(T) - \mathsf{freq}(S) > \delta$
  ex : 1-libres



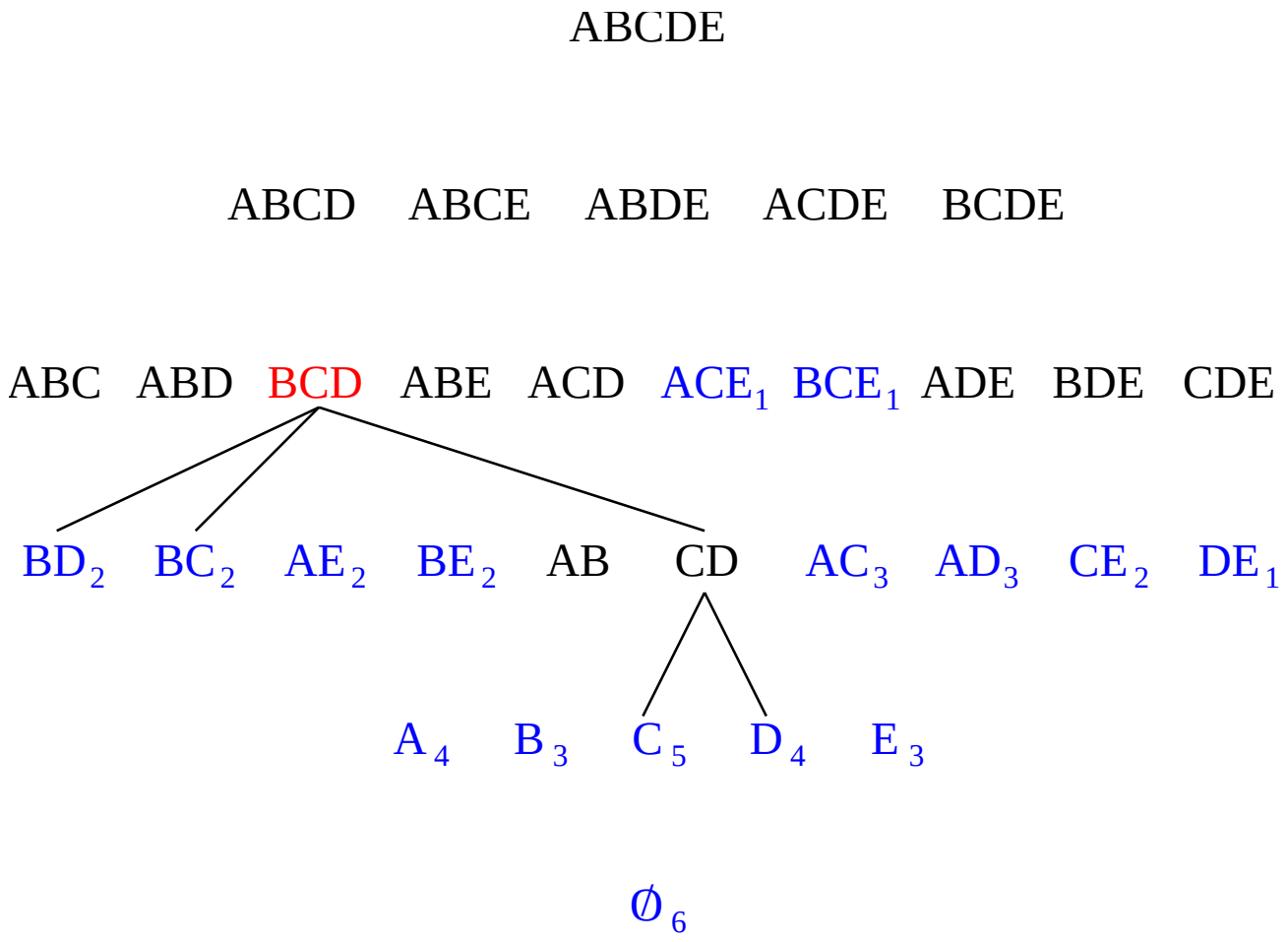
Différentes représentations condensées

- itemsets : 32
- clos : 11
- libres : 16
- 1-libres : 6



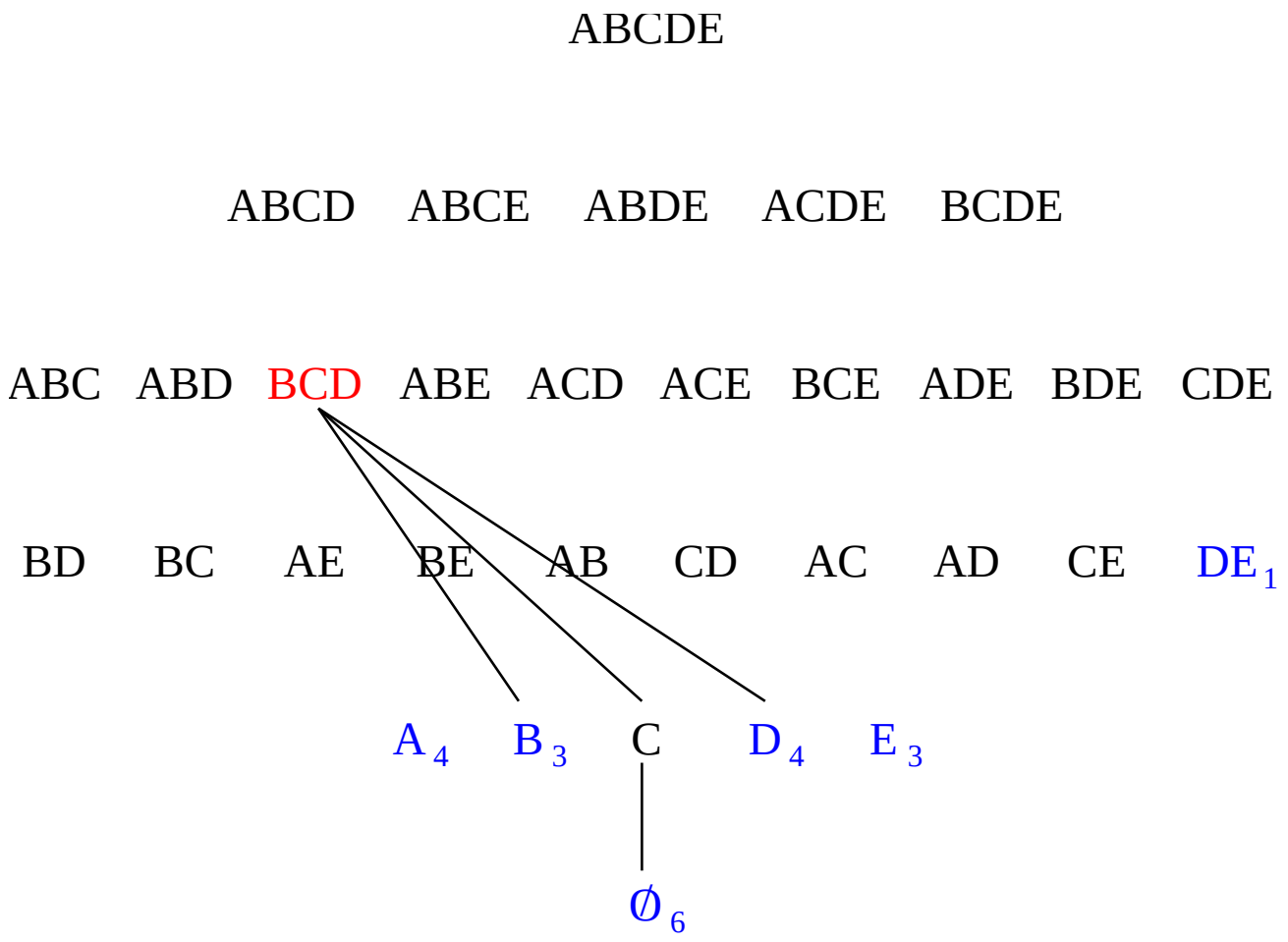
Régénération

- Régénération des fréquences des itemsets à partir des libres et des δ-libres



Régénération

- Régénération des fréquences des itemsets à partir des libres et des δ-libres



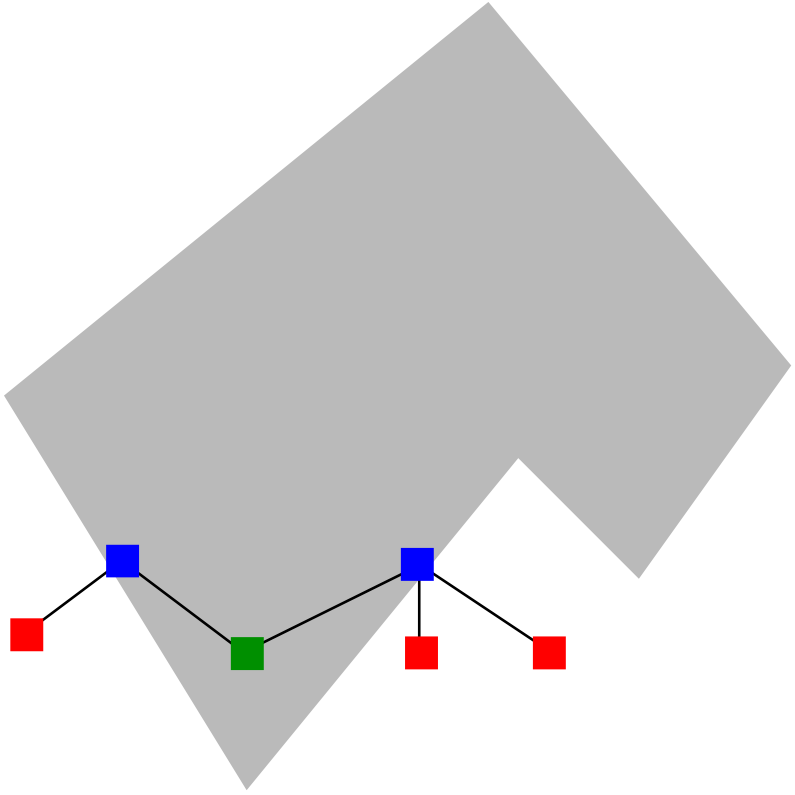
Contraintes et algorithmes

- Définition des contraintes
  - $\mathcal{C}_{\text{clos}}$
  - $\mathcal{C}_{\text{libre}}$ et $\mathcal{C}_{\delta\text{-libre}}$: anti-monotones
- Algorithmes d'extraction de motifs fréquents
  - clos fréquents : $\mathcal{C}_{\gamma\text{-minfreq}} \wedge \mathcal{C}_{\text{clos}}$
  - libres fréquents : $\mathcal{C}_{\gamma\text{-minfreq}} \wedge \mathcal{C}_{\text{libre}}$
  - δ-libres fréquents : $\mathcal{C}_{\gamma\text{-minfreq}} \wedge \mathcal{C}_{\delta\text{-libre}}$
- Notre algorithme (CoCo) [IDEAS 01, Journal IDA]
  - Combiner représentations condensées et $\mathcal{C}_{\text{am}} \wedge \mathcal{C}_{\text{m}}$

Difficultés

Dues aux contraintes monotones

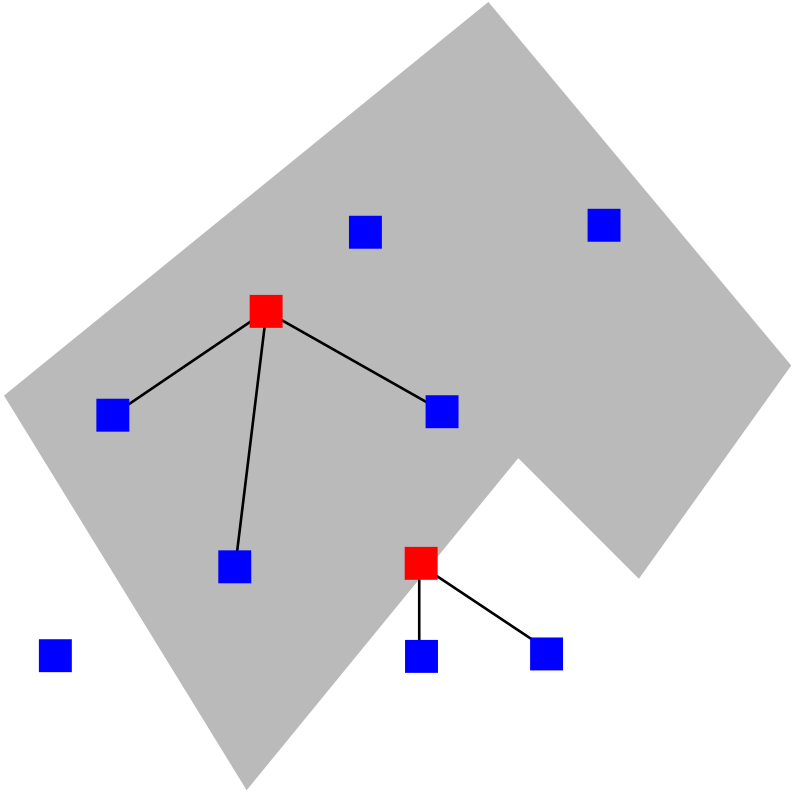
- Test de la contrainte $\mathcal{C}_{\delta\text{-libre}}$



Difficultés

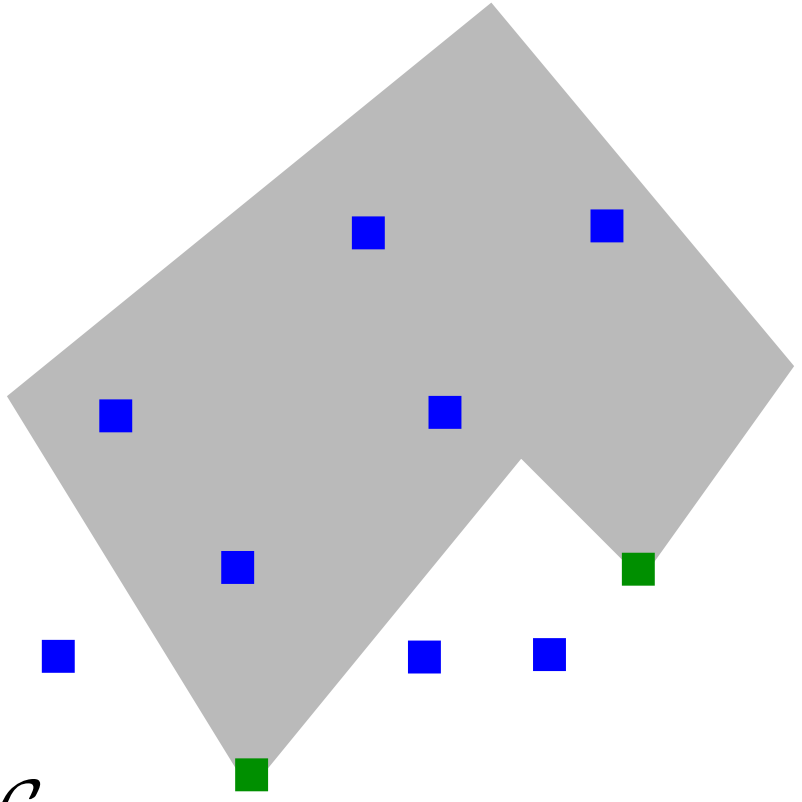
Dues aux contraintes monotones

- Test de la contrainte $\mathcal{C}_{\delta\text{-libre}}$
- Régénération des itemsets à partir des δ-libres



δ-libres contextuels [IDEAS 01]

- Définition: S est δ-libre contextuel relativement à $\mathcal{C}_{\mathsf{m}}$:
  $\forall T \subset S$ tel que $\mathcal{C}_{\mathsf{m}}(T)$, $\mathsf{freq}(T) - \mathsf{freq}(S) > \delta$
- $\mathcal{C}_{\delta-\text{libre-c}}$ est anti-monotone



- Notre algorithme CoCo:
  Extraction de $\mathcal{C}_{\delta-\text{libre-c}} \wedge \mathcal{C}_{\mathsf{m}} \wedge \mathcal{C}_{\mathsf{am}}$

Application/Évaluation [FQAS 00]

- Complétion de la base de données

tid	
1	AB
2	ADE
3	BCD

$\Longrightarrow$

tid	
1	AB$\overline{\text{CDE}}$
2	ADE$\overline{\text{BC}}$
3	BCD$\overline{\text{AE}}$

  Extraction très difficile

- Quels sont les itemsets fréquents contenant au moins 3 items positifs ?
  Contrainte $\mathcal{C}_{\text{am3p}}$ (monotone)

$$\mathcal{C}_{\text{am3p}}(\text{ABD}\overline{\text{E}}) = \text{vrai}, \qquad \mathcal{C}_{\text{am3p}}(\overline{\text{A}}\text{BC}) = \text{faux}$$

Comparaison de différentes stratégies

Stratégies utilisant des techniques existantes :

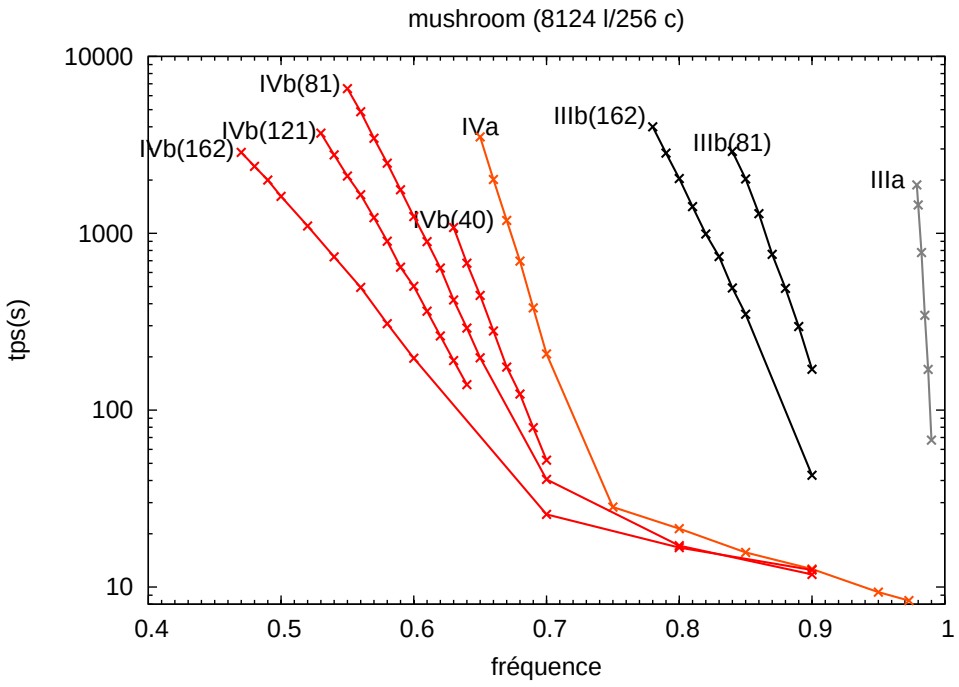
- Strat. I : $\mathcal{C}_{\gamma\text{-freq}}$ puis test $\mathcal{C}_{\text{am3p}}$
- Strat. II : $\mathcal{C}_{\gamma\text{-freq}} \wedge \mathcal{C}_{\text{am3p}}$
- Strat. IIIa : $\mathcal{C}_{\gamma\text{-freq}} \wedge \mathcal{C}_{\text{libre}}$ puis test $\mathcal{C}_{\text{am3p}}$
- Strat. IIIb : $\mathcal{C}_{\gamma\text{-freq}} \wedge \mathcal{C}_{\delta\text{-libre}}$ puis test $\mathcal{C}_{\text{am3p}}$

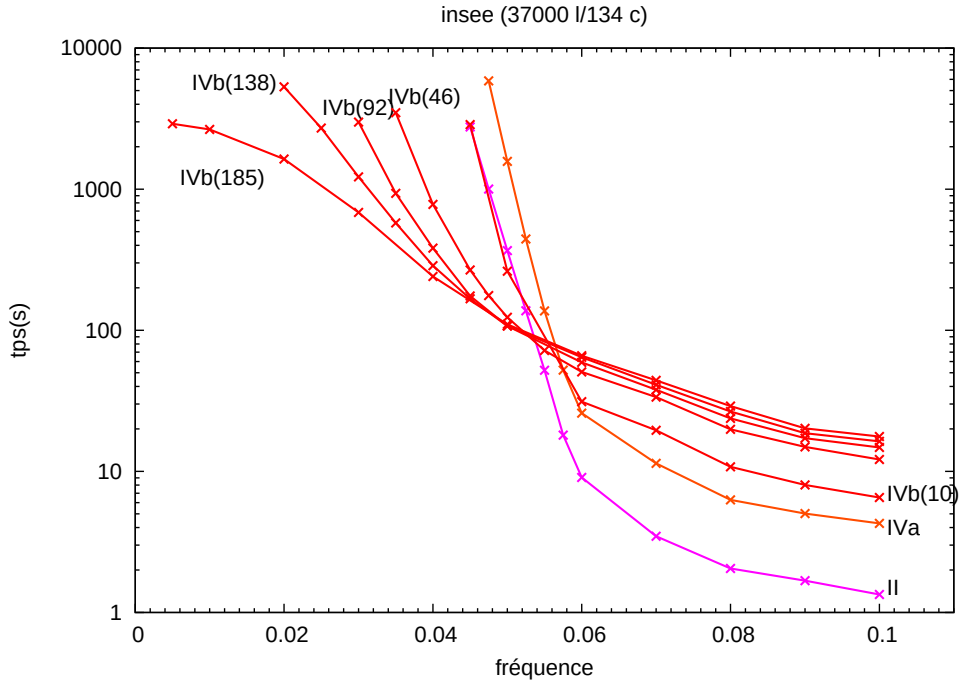
Nouvelles stratégies utilisant notre algorithme CoCo :

- Strat. IVa : $\mathcal{C}_{\gamma\text{-freq}} \wedge \mathcal{C}_{\text{libre-c}} \wedge \mathcal{C}_{\text{am3p}}$
- Strat. IVb : $\mathcal{C}_{\gamma\text{-freq}} \wedge \mathcal{C}_{\delta-\text{libre-c}} \wedge \mathcal{C}_{\text{am3p}}$

Performances

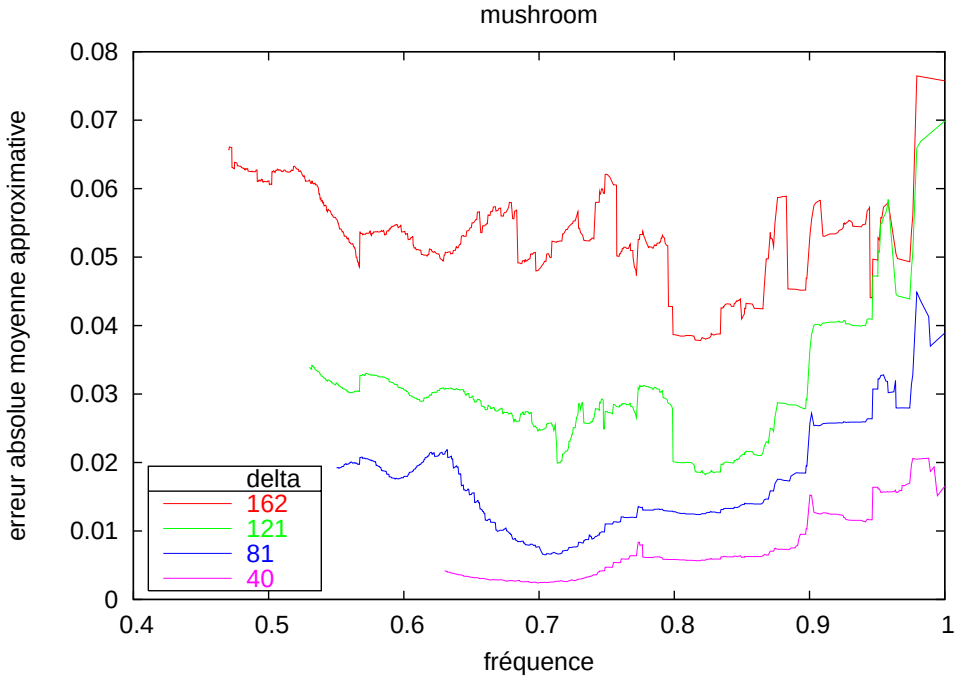
- Temps d'extraction en fonction du seuil de fréquence

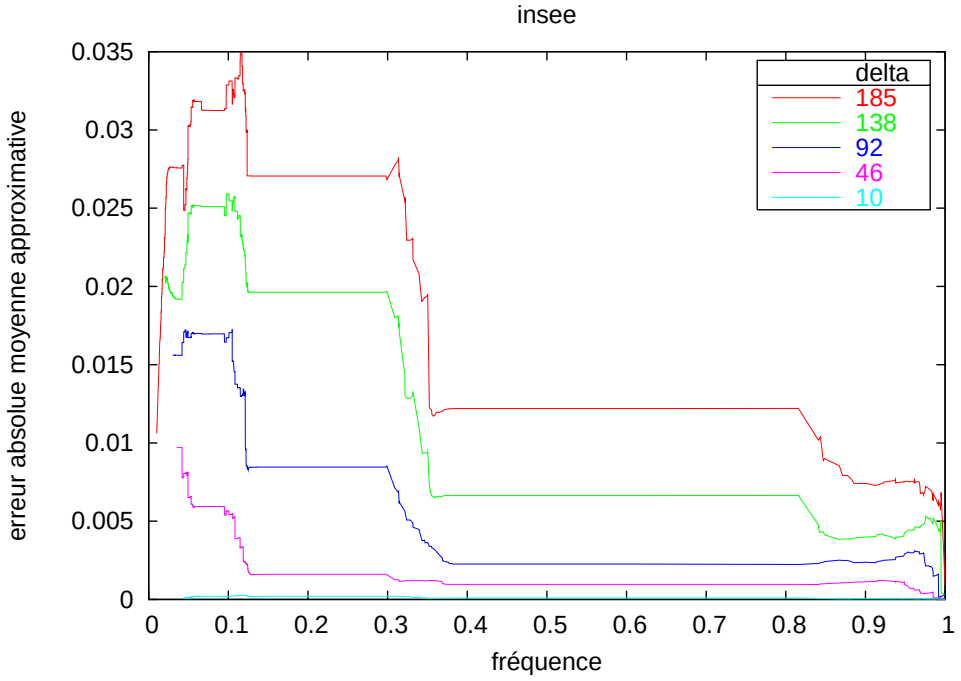




Erreurs d'approximation

- Erreur moyenne en fonction de δ et de la fréquence





Application biologique

Collaboration avec le CGMC (Lyon I, O. Gandrillon)
Données SAGE : cellules/gènes chez l'homme

- Extraction des libres fréquents : infaisable
- Dans un contexte restreint ($\sim$1000 gènes) : faisable
  Grande homogénéité des motifs (itemsets et clôtures) permettant [Genome Biology] :
  - découverte d'erreur d'attribution d'une fonction biologique
  - hypothèse sur la fonction biologique d'un gène
- Extraction dans le contexte complet : extraction sous contraintes indispensable

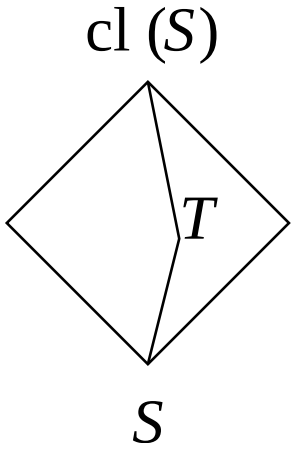
Séquences de requêtes

Motivation

- Processus d'extraction itératifs
  $\Longrightarrow$ réutilisation de résultats de requêtes précédentes
- Approches existantes : stockage des résultats de requêtes, combinaisons de requêtes
  - Volumineux
  - Infaisable dans des données denses/corrélées
- Notre approche : utilisation d'une représentation condensée comme cache [PKDD 02]
  - Moins volumineux
  - Profite de notre algorithme CoCo

Description de notre cache

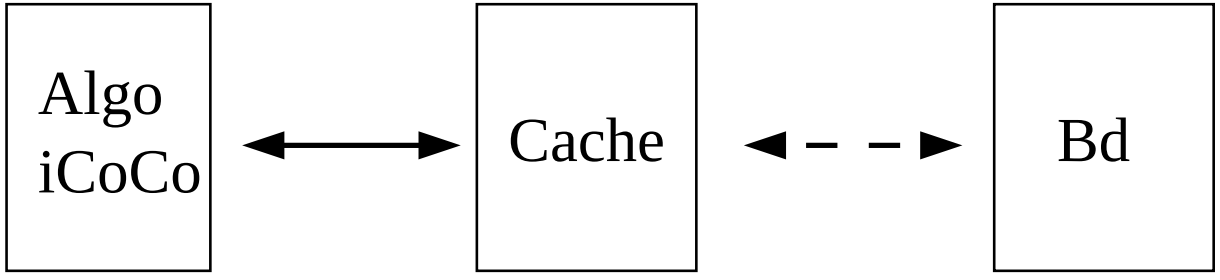
- Contient des triplets $(S, \mathsf{freq}(S), \mathsf{cl}(S) \setminus S)$
- Permet de répondre à la question “quelles sont la fréquence et la clôture de T ?” pour $S \subseteq T \subseteq \mathsf{cl}(S)$



- Pour ces itemsets T, le cache remplace la base de données

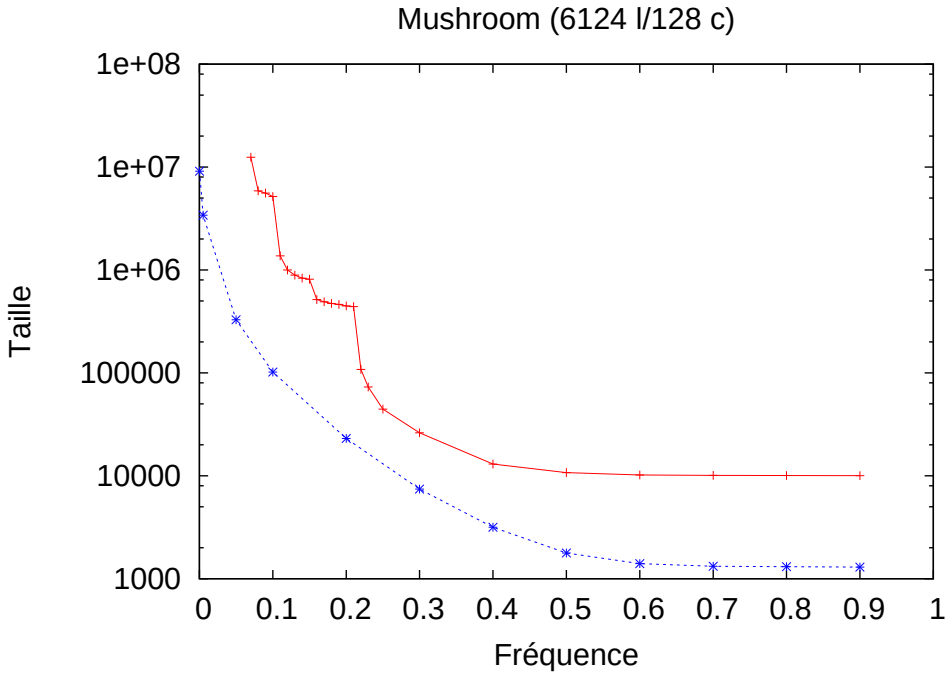
Algorithme iCoCo [PKDD 02]

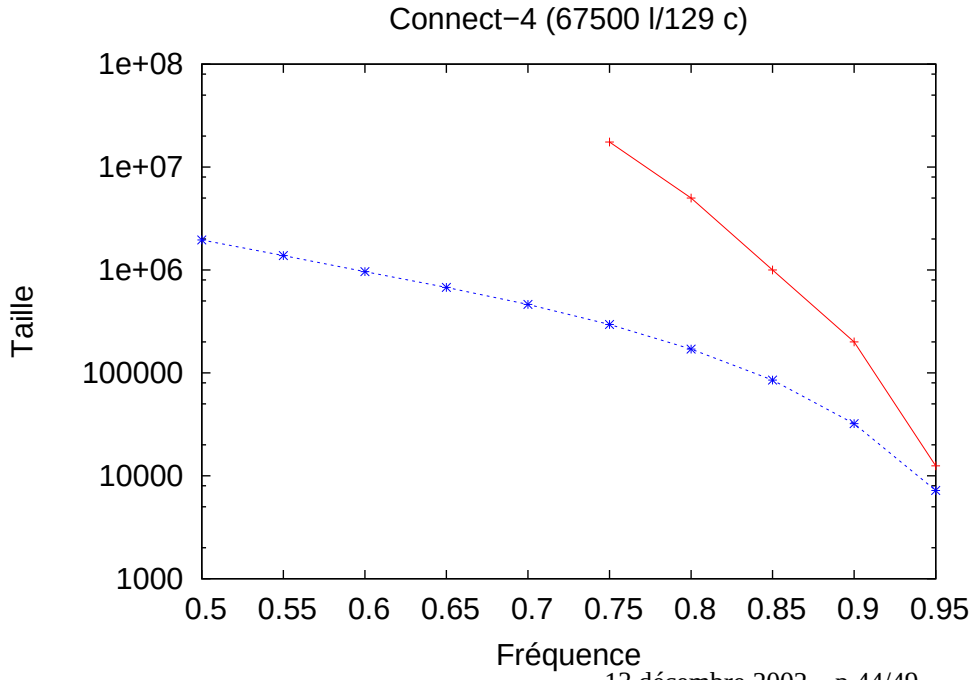
- Fondé sur notre algorithme CoCo
  $\mathcal{C}_{\text{am}} \wedge \mathcal{C}_{\text{m}} \wedge \mathcal{C}_{\text{libre-c}}$
- Utilise le cache à la place de la base de données lorsque cela est possible
- Notre implémentation actuelle n'utilise que les contraintes anti-mononotones



Évaluation de la taille du cache

- Taille de la collection $\mathcal{F}_\gamma$ des itemsets γ-fréquents
  $taille(\mathcal{F}_\gamma) = \sum_{S \in \mathcal{F}_\gamma} (|S| + 1)$
- Taille de notre cache
  $taille(cache) = \sum_{S \in FreqLibre_\gamma} (|S| + |\mathbf{cl}(S) \setminus S| + 1)$



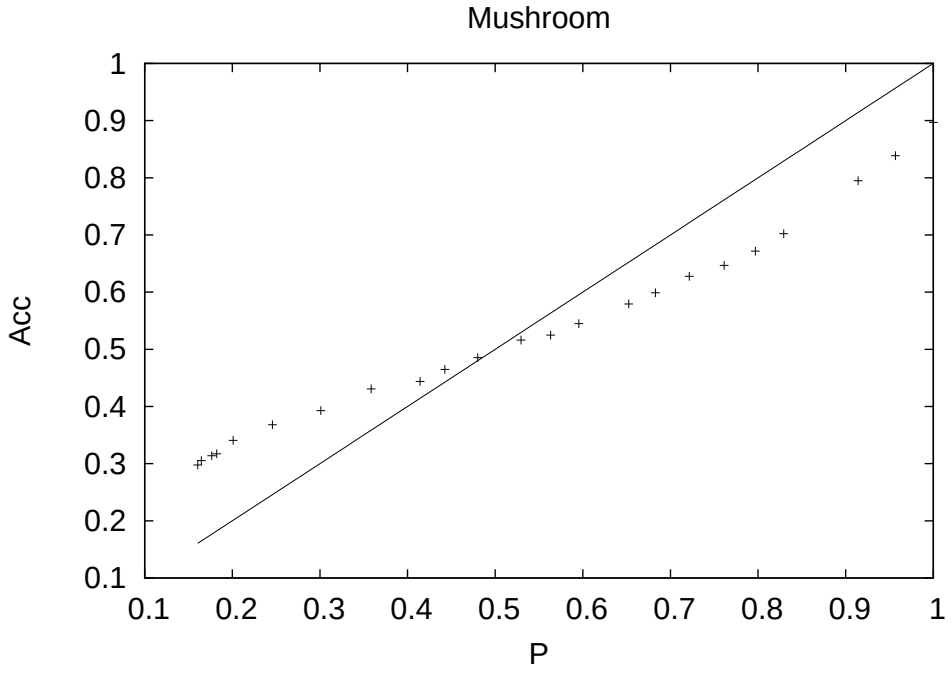


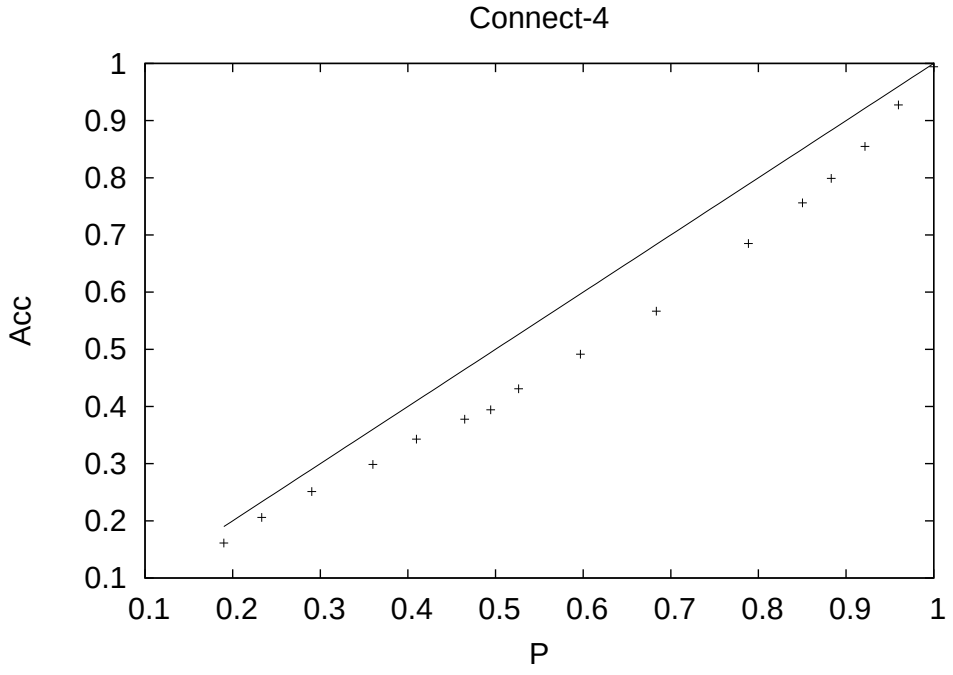
Performances

Deux extractions avec chaque contrainte $\mathcal{C}_{\gamma\text{-minfreq}}$: (1) sans cache, (2) avec un cache C

- $\mathrm{P} = \frac{\left|\mathsf{Th}(\mathcal{C}_{\gamma\text{-minfreq}}) \cap C\right|}{\left|\mathsf{Th}(\mathcal{C}_{\gamma\text{-minfreq}})\right|}$

$$\mathrm{Acc} = \frac{t_{sans} - t_{avec}}{t_{sans}}$$





Conclusion et perspectives

Contributions principales

Bases de données inductives

- Théorie
  - Algorithme générique pour les requêtes étendues : itemsets satisfaisant $\mathcal{C}_{\mathsf{am}} \wedge \mathcal{C}_{\mathsf{m}}$
  - Algorithme CoCo : représentations condensées sous contraintes
  - Algorithme iCoCo : utilisation d'un cache
- Application
  - Données d'expression de gènes

Perspectives

- Usages multiples des motifs fréquents
- cInQ : définition de stratégies d'évaluation
  - Quand faut-il pousser les contraintes ?
  - Contraintes plus générales
  - Liens avec l'apprentissage et la PLI
- Intégration dans le langage d'extraction de règles d'association MINE-RULE (Unito)

Vers des bases de données inductives dédiées ?

- Molfea [ALU-FR] : extraction de fragments moléculaires
- Analyse du transcriptome :
  - Contexte Spécifique
  - Nécessité d'utiliser les contraintes
  - Définition des contraintes