CLUSTERING CONCEPTUEL À BASE DE MOTIFS LOCALEMENT OPTIMAUX

ATELIER CLUCO, EGC 2015

frederic.pennerath@centralesupelec.fr
Equipe IMS
Centrale-Supélec, campus de Metz

Résumé

- Clustering conceptuel.

  Exemple de COBWEB (Fisher 87)

- Problèmes de sélection de motifs fréquents
- Les motifs localement optimaux (Pennerath 09)
- Limitation du modèle
- Redondance de motifs
- Les motifs δ-localement optimaux

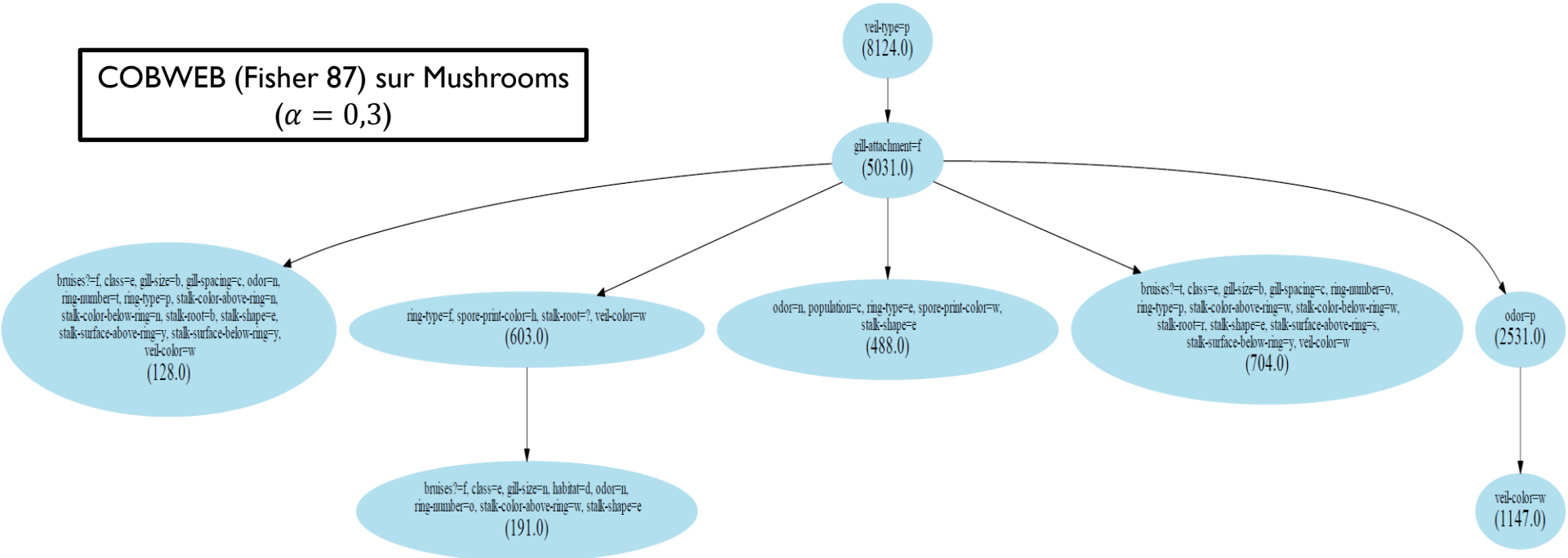
Clustering conceptuel (symbolique)

Extraire de données une hiérarchie de concepts (intension, extension) avec :

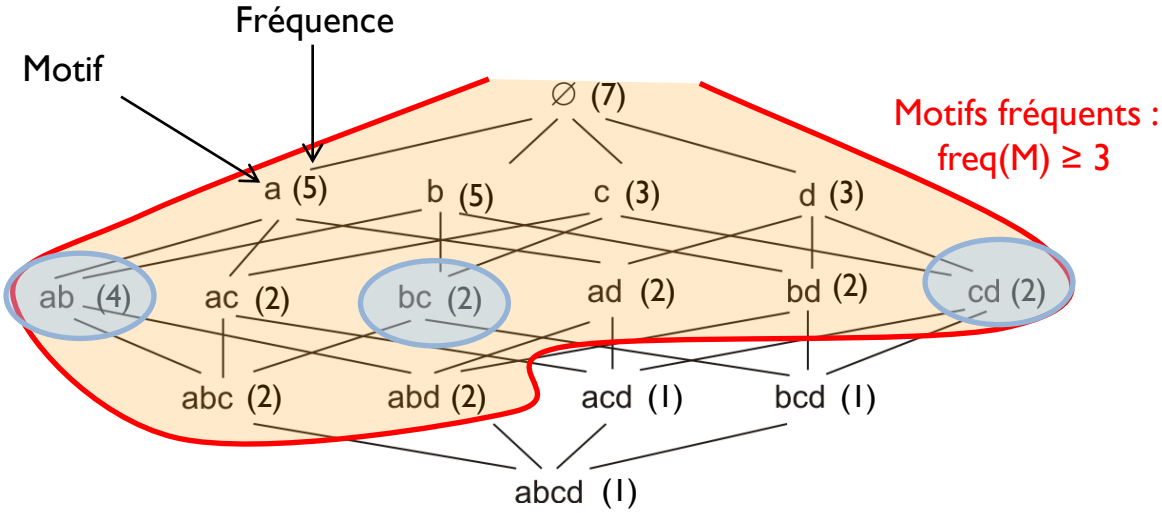
- Des concepts représentatifs des données → de grandes extensions
- Des concepts descriptifs des données → de grandes intensions

→ Problème de compromis

COBWEB (Fisher 87) sur Mushrooms
($\alpha = 0{,}3$)



Problèmes de sélection de motifs fréquents



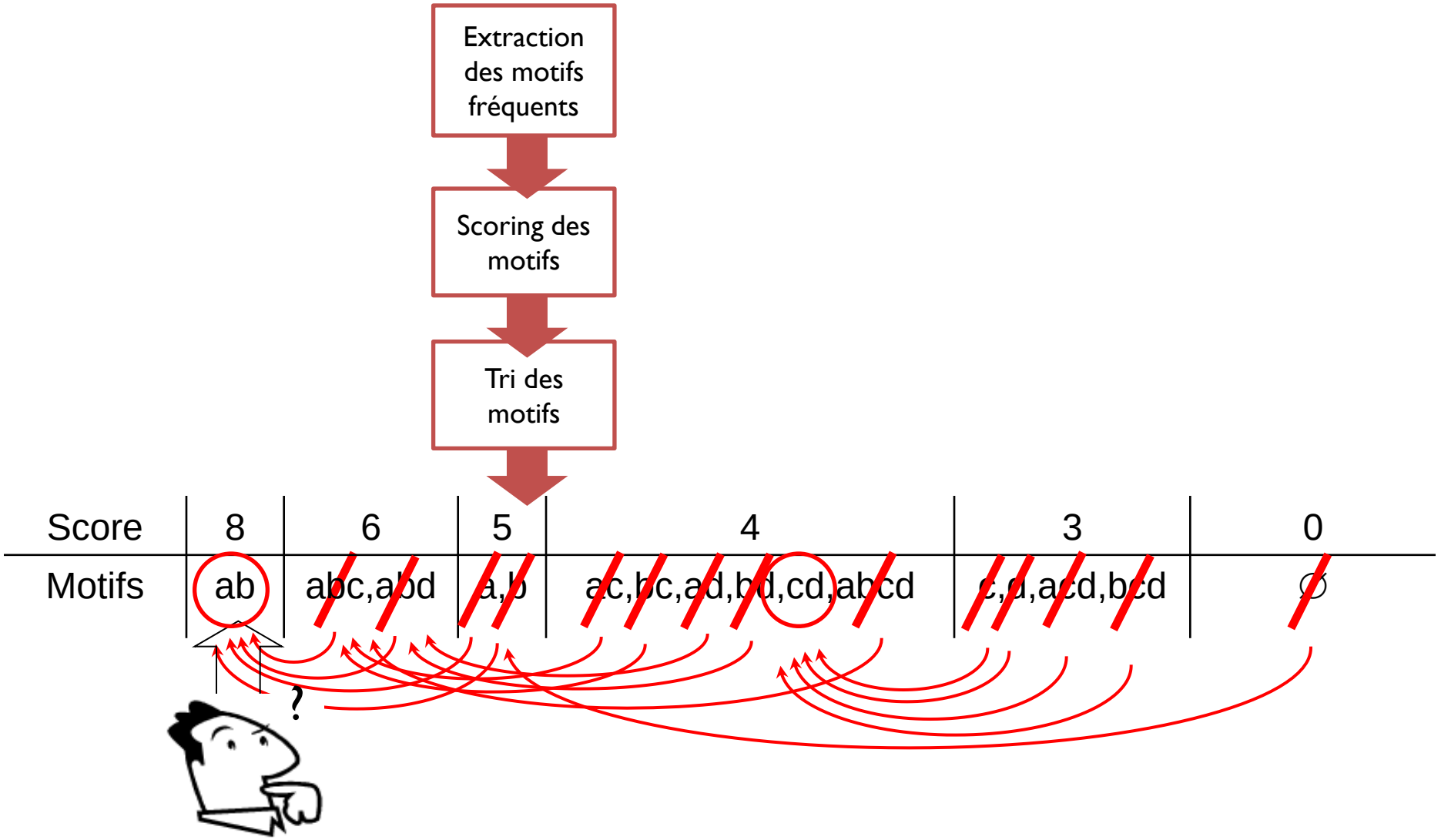
Extraire un ensemble de motifs $\{M_i\}$ parmi les fréquents

- Peu nombreux
- Descriptifs ≈ faible distance intra-classe
- Non redondants ≈ forte distance inter-classe

Nombreuses propositions :

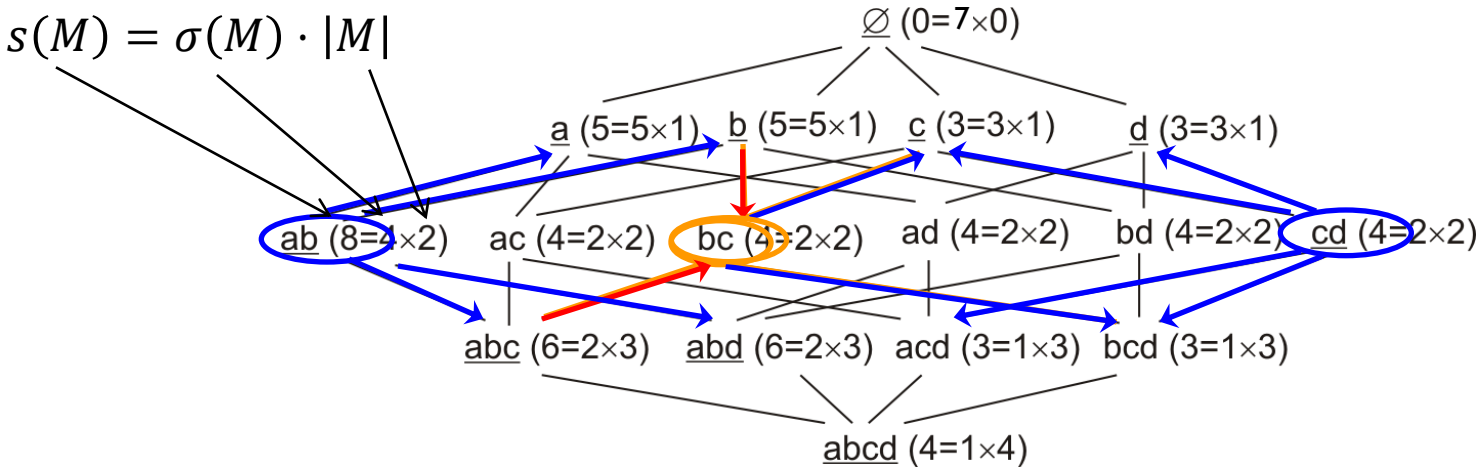
Mampaey11, Van Leeuwen11, Knobbe10, Tatti10, **Pennerath09**, Knobbe06, etc

La sélection (naïve) de motifs par un expert



Le modèle des motifs localement optimaux (MLOs)

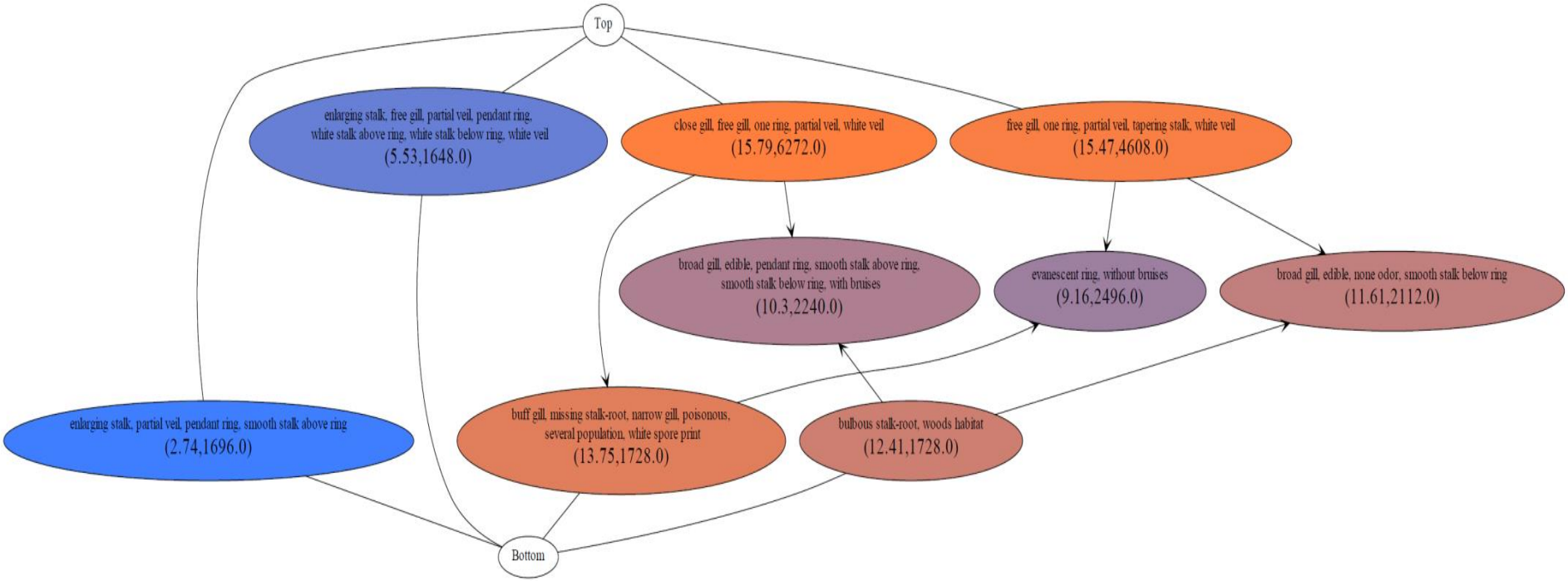
- **Ordre des motifs** $(\mathcal{M}, \subseteq_{\mathcal{M}})$
- **Fonction de score** $s : (\mathcal{M}, \subseteq_{\mathcal{M}}) \rightarrow (\mathcal{S}, \leq_{\mathcal{S}})$
- **Voisinage** d'un motif M : prédécesseurs et successeurs immédiats de M
- M' **domine** M si M' et M sont voisins et si $s(M) <_{\mathcal{S}} s(M')$
- Un motif est **localement optimal** s'il n'est dominé par aucun motif.



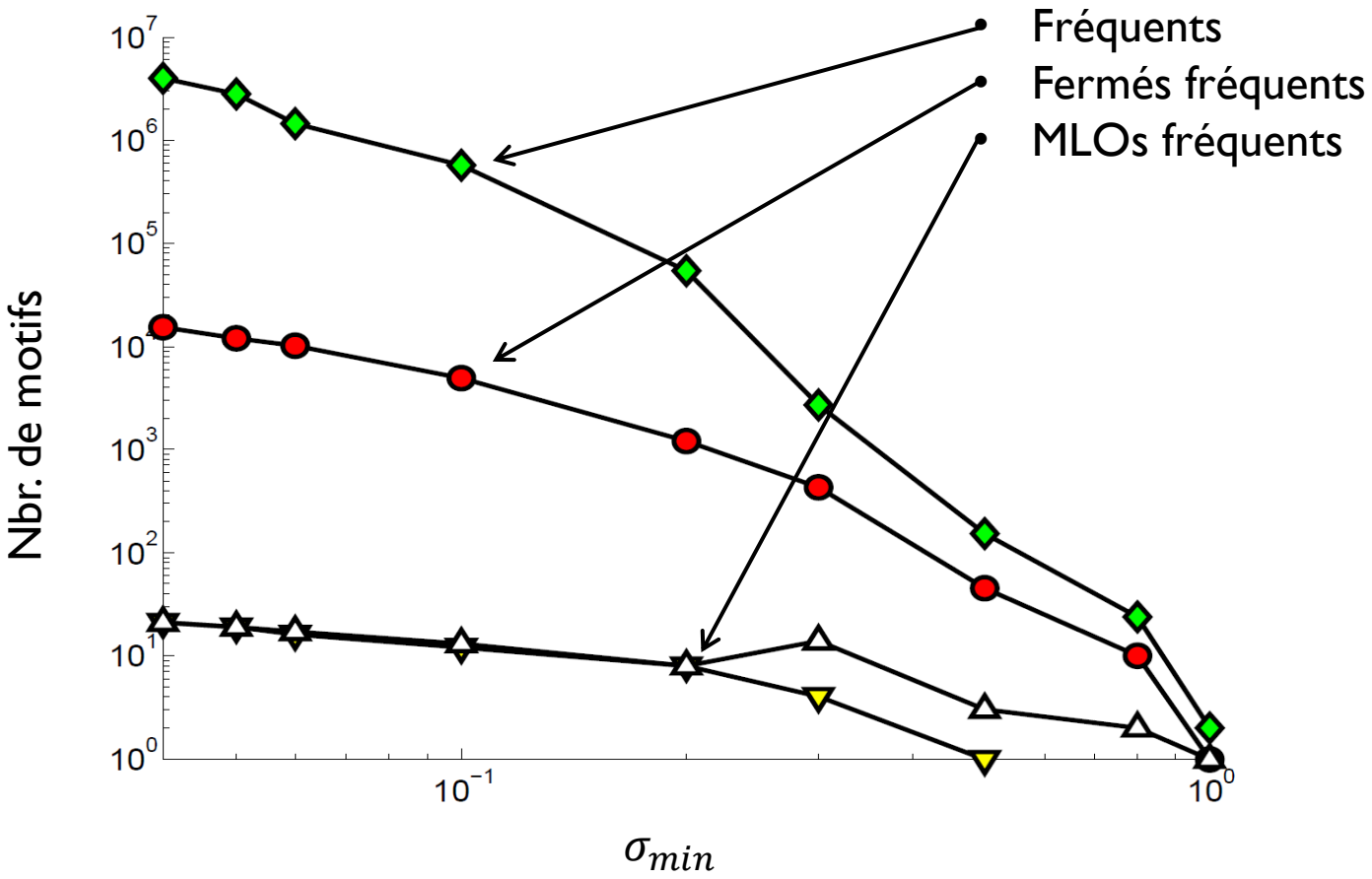
Exemple de Mushrooms

Pour $\sigma_{min} = 0{,}2$:

- >53000 motifs fréquents
- >1200 motifs fermés fréquents
- 9 motifs localement optimaux fréquents pour $s(M) = I(M) \cdot \sigma(M)$



Exemple de Mushrooms

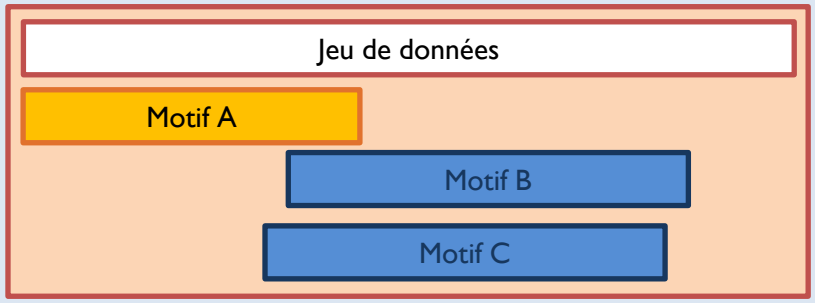


Comment est définie la redondance entre motifs ?

Redondance dans l'espace des données :

Deux motifs sont redondants s'ils ont approximativement le même support

(i.e. ont un indice de Jaccard proche de 1).



Avantages :

- Indicateurs calculables facilement et rapidement
- Facile à justifier du point de vue des statistiques

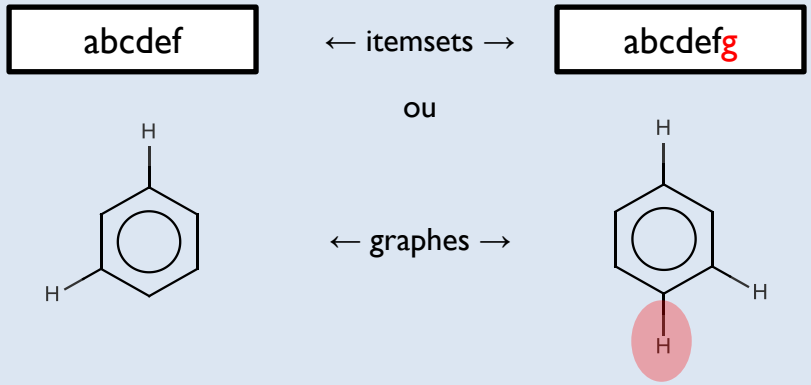
Inconvénients :

- Aveugle aux connaissances du domaine
- Souvent hiérarchie de motifs horizontale (partition des données)

Redondance dans l'espace des motifs :

Deux motifs sont redondants s'ils sont structurellement semblables

(i.e. ont une faible distance d'édition)



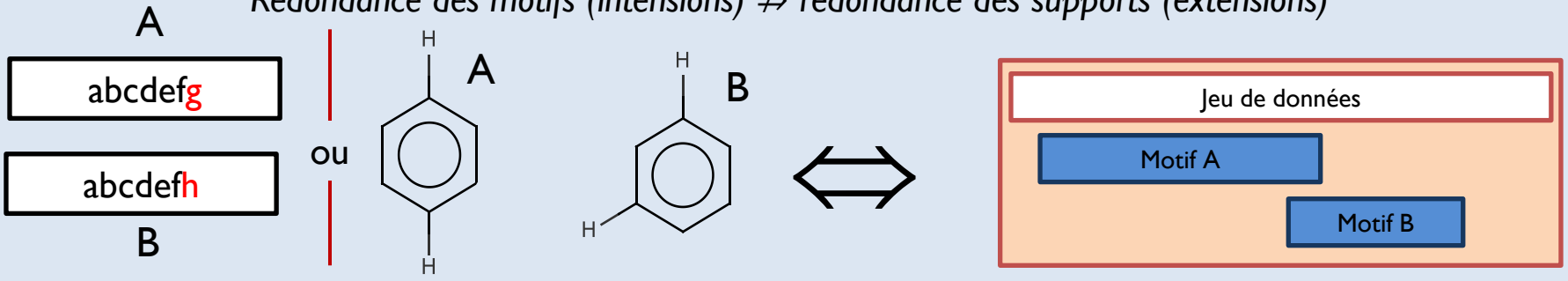
Avantages :

- Agit dans le domaine d'interprétation de l'expert.
- La distance d'édition peut intégrer des connaissances du domaine.
- Compatible avec une hiérarchie verticale

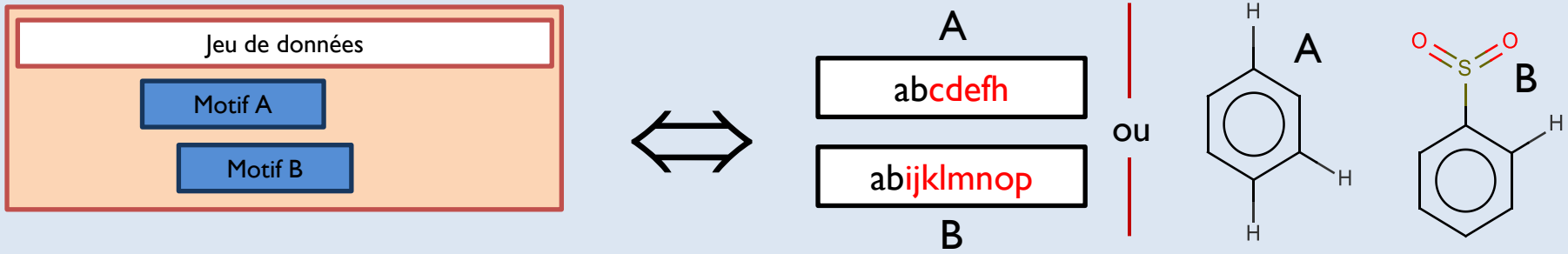
Les deux formes de redondance sont-elles redondantes ?

Souvent mais pas toujours…

Redondance des motifs (intensions) ⇏ redondance des supports (extensions)



Redondance des supports ⇏ redondance des motifs



Relation de δ-dominance

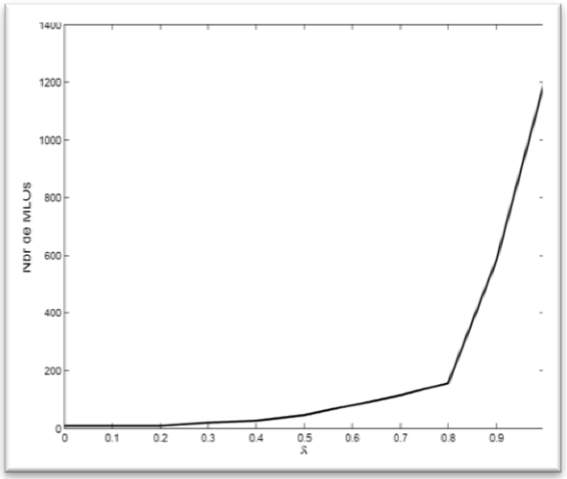
Définition : pour $\delta \in [0,1]$

M' **δ-domine** M si $s(M') >_S s(M)$ et si M et M' sont redondants

- dans l'espace des motifs, i.e. M et M' sont voisins
- dans l'espace des données, i.e. $J(M_1, M_2) \stackrel{\text{def}}{=} J\big(ext(M_1), ext(M_2)\big) \geq \delta$

Propriétés :

- Si $\delta = 0$, les δ-MLOs sont les MLOs
- Si $\delta = 1$, les δ-MLOs sont les fermés
- Si $\delta_1 \leq \delta_2$, { δ_1-MLOs } $\subseteq$ { δ_2-MLOs }

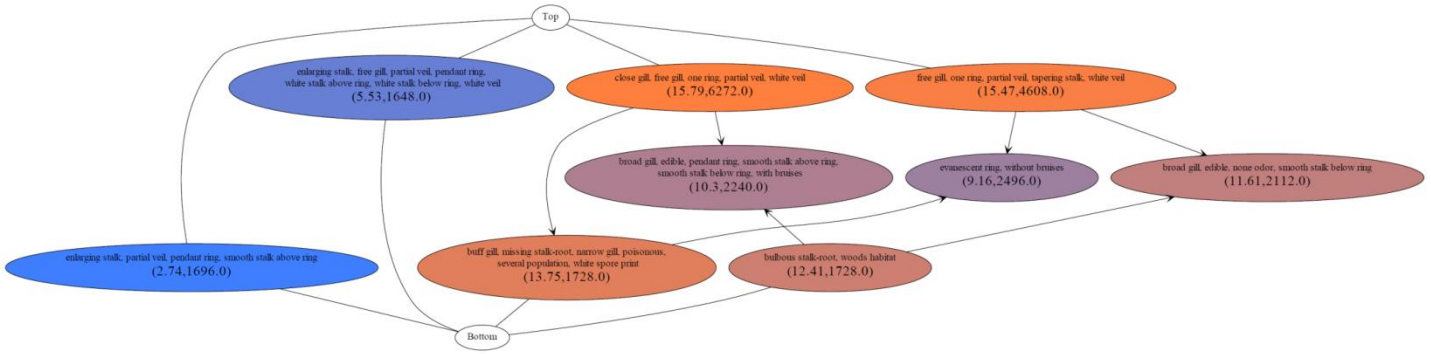


- Immédiat à prendre en compte :

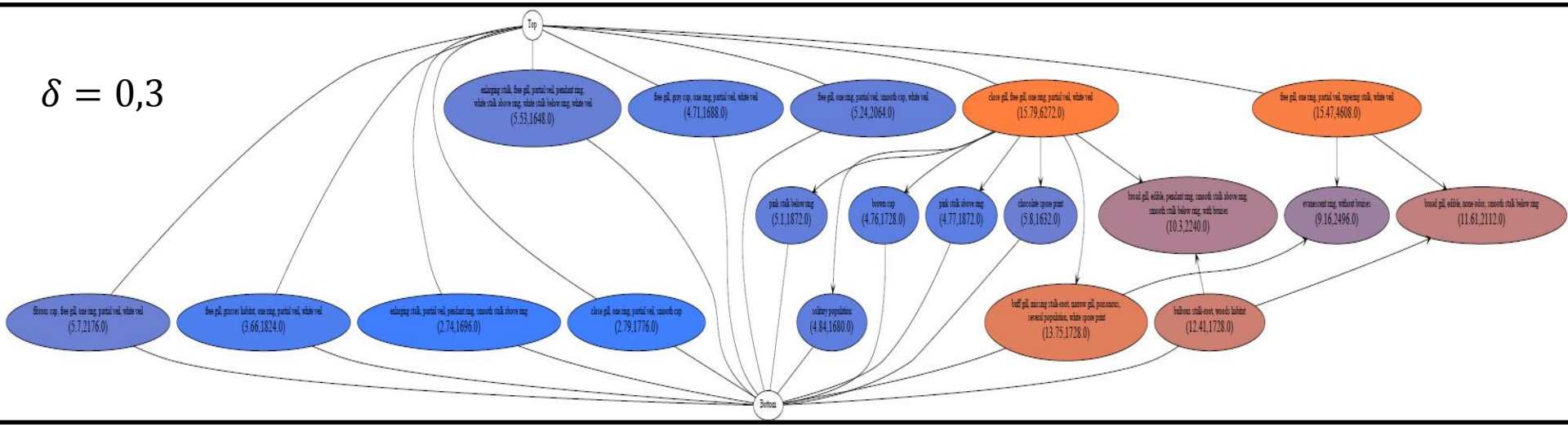
$$J(M_1, M_2) = J(M_1, M_1 \cup \{i\}) = \frac{|ext(M_1) \cap ext(M_1 \cup \{i\})|}{|ext(M_1) \cup ext(M_1 \cup \{i\})|} = \frac{|ext(M_1 \cup \{i\})|}{|ext(M_1)|} = \frac{\sigma(M_2)}{\sigma(M_1)}$$

δ-MLOs de Mushrooms ($\sigma_{min} = 0{,}2$)

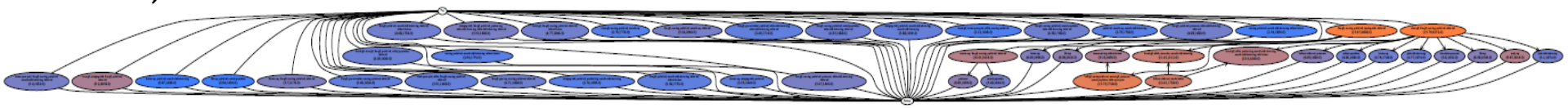
$\delta = 0{,}2$



$\delta = 0{,}3$



$\delta = 0{,}5$



Conclusions

- δ-MLOs définissent un clustering conceptuel satisfaisant
  - Peu de motifs non redondants
  - Niveau de granularité modulable
  - Développement horizontal et vertical équilibré
- Avantage des δ-MLOs par rapport à COBWEB et aux autres méthodes de sélection de motifs :
  - Prise en compte de la redondance dans l'espace des motifs
  - Prise en compte des connaissances métier (fonction de score, voisinage)
  - Algorithme exact sans heuristique
- Perspective : comparaison plus systématique avec COBWEB