Treillis, Classification et Images

Karell Bertet

Laboratoire L3I
Université de La Rochelle

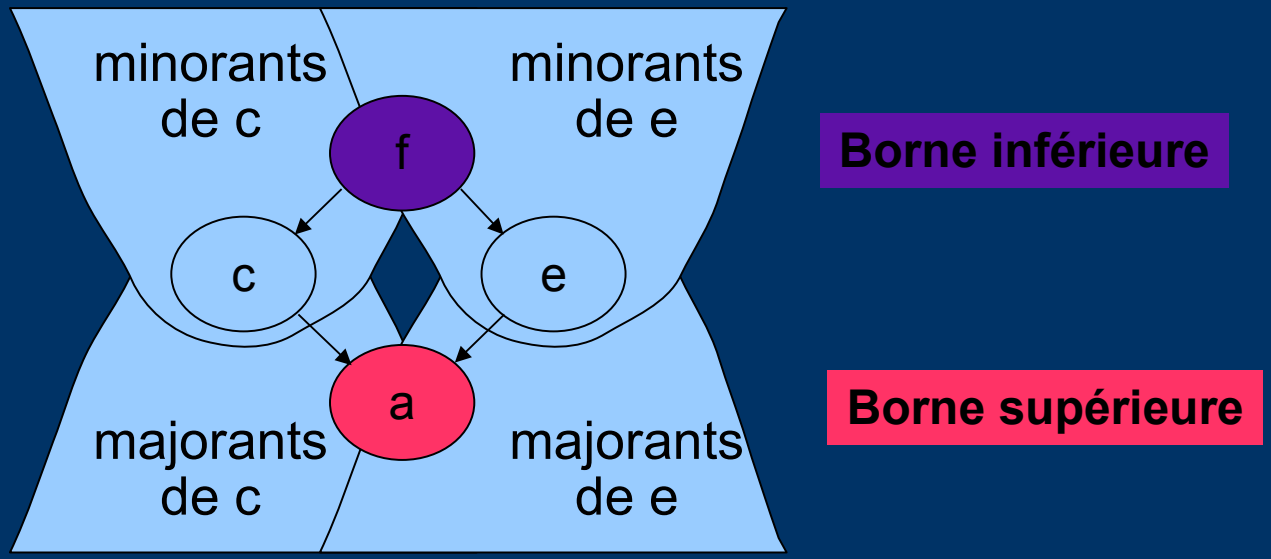
Groupe de travail Imédoc
7 Juillet 2008

Plan

- Théorie des treillis
  - Treillis de Galois
  - Règles d'association, d'implication
  - Algorithmes de génération
- Fouille de données
  - Méthodes utilisant des règles d'association
  - Méthodes utilisant un treillis
- Cas des images

Treillis: définition

- Un treillis est un ensemble muni :
  - D'une relation d'ordre : relation binaire transitive, réflexive et antisymétrique
  - D'une borne supérieure et d'une borne inférieure pour chaque paire d'éléments de l'ensemble :



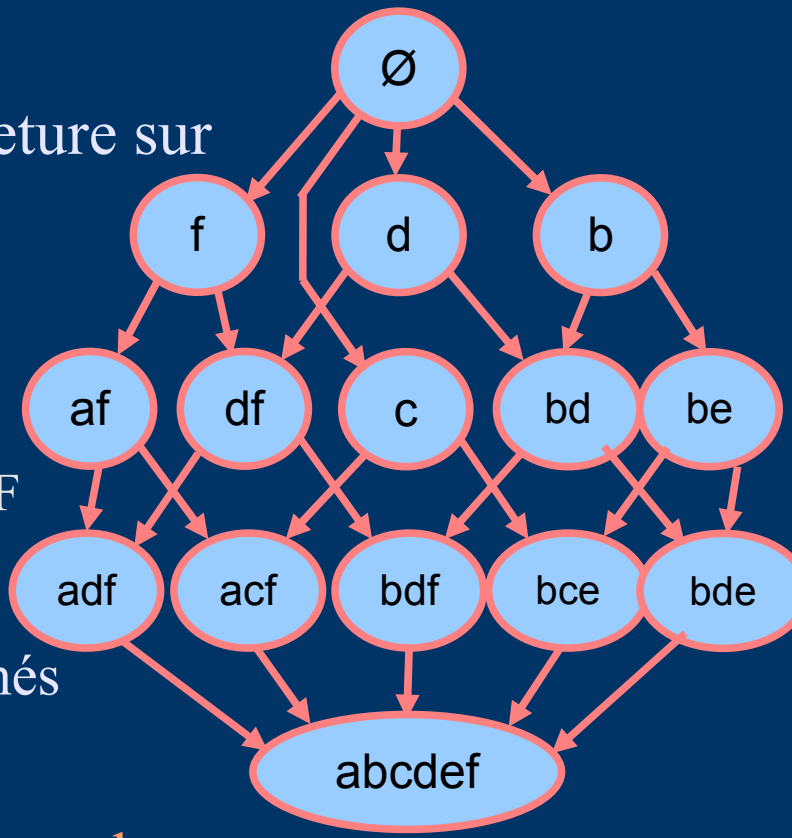
Treillis des fermés: définition [CasMon 03]

- Le treillis des fermés sur un système de fermeture sur S défini par:
  - Une famille F de fermés:

    parties de S : F={Φ,f,af, db,...} pour S={a,b,c,d,e,f}

    stables par intersection : acf ∈ F et adf ∈ F ⇒ af ∈ F

    et contenant S: S ∈ F
  - Munie de la relation d'inclusion entre les fermés

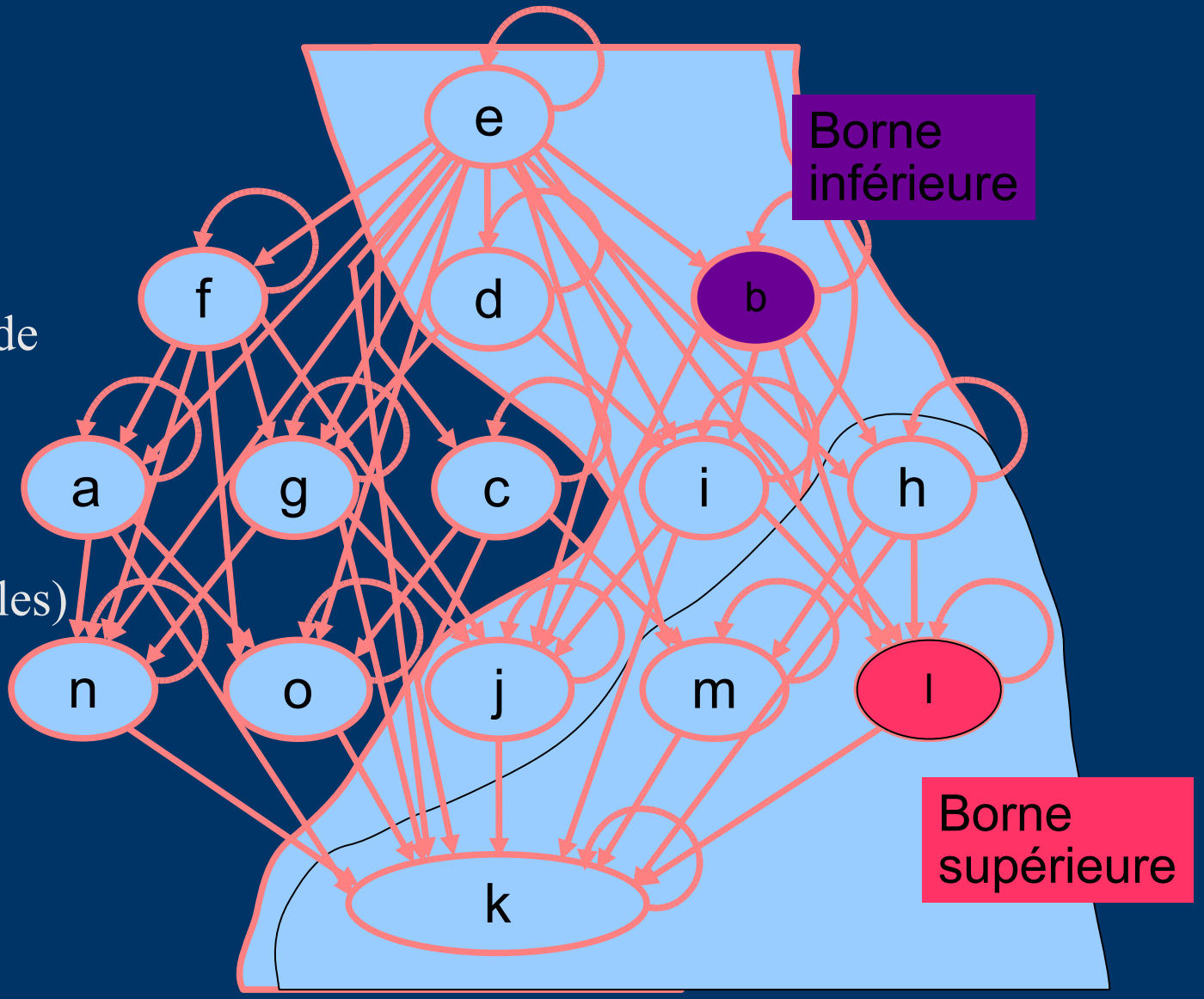


- On associe à un treillis des fermés son opérateur de fermeture φ défini sur P(S) par:
  - φ(X) est le plus petit fermé de la famille F contenant X

φ(a)=af
φ(ab)=abcde
φ(Φ)=Φ
.....

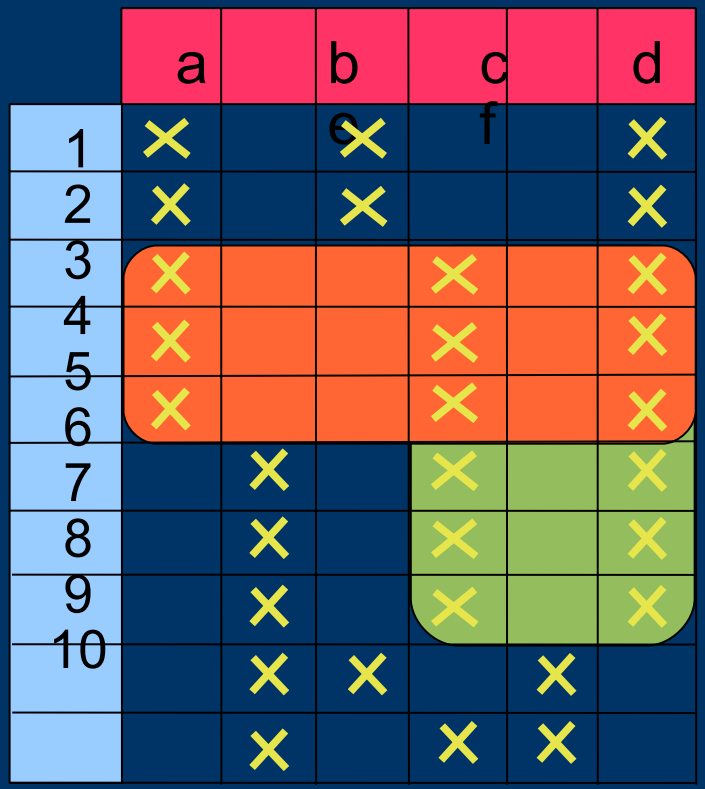
Treillis des fermés: diagramme de Hasse

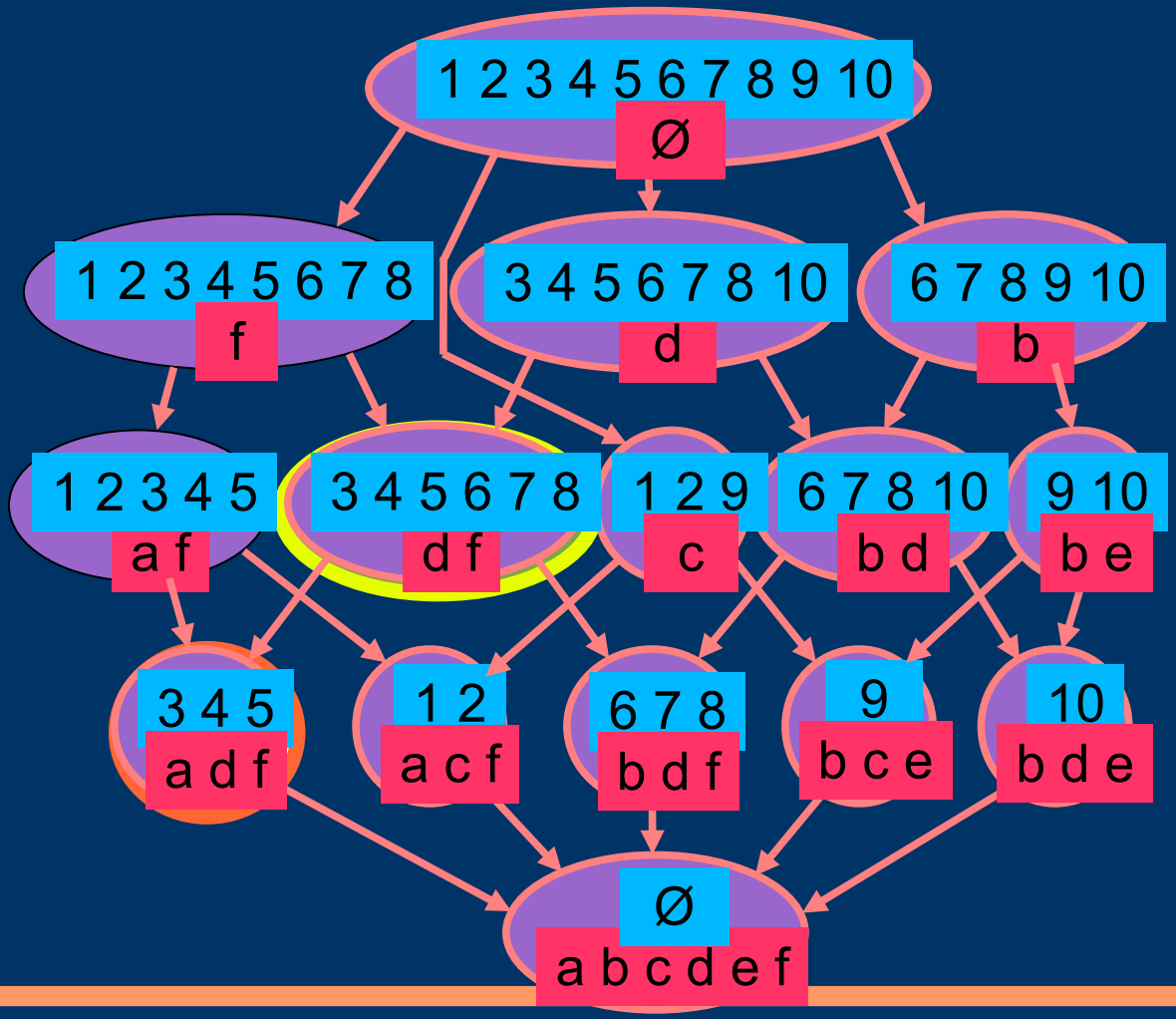
- Diagramme de Hasse:
  - Réduction transitive (suppression des arcs de transitivité)
  - Réduction réflexive (suppression des boucles)



Treillis de Galois

- Le Treillis de Galois [BarMon70] ou Treillis des concepts [Wil99] se définit à partir d'une table de données binaires





Treillis de Galois: définition

- Les données binaires sont décrites par:
  - Un ensemble O d'objets
  - Un ensemble I d'attributs
  - Une connexion de Galois (f,g) entre objets et attributs:
    - f associe aux objets leurs attributs
    - g associe aux attributs leurs objets

- Propriété:
  - f o g est un opérateur de fermeture sur les attributs,
  - g o f est un opérateur de fermeture sur les objets

Treillis de Galois: définition

- Le treillis de Galois des données (O,I,(f,g)) se définit par:
  - Un ensemble de concepts:

    un concept est une paire (A,B) avec:
    - $A \subseteq O$, $B \subseteq I$, $A = f(B)$ et $B = g(A)$

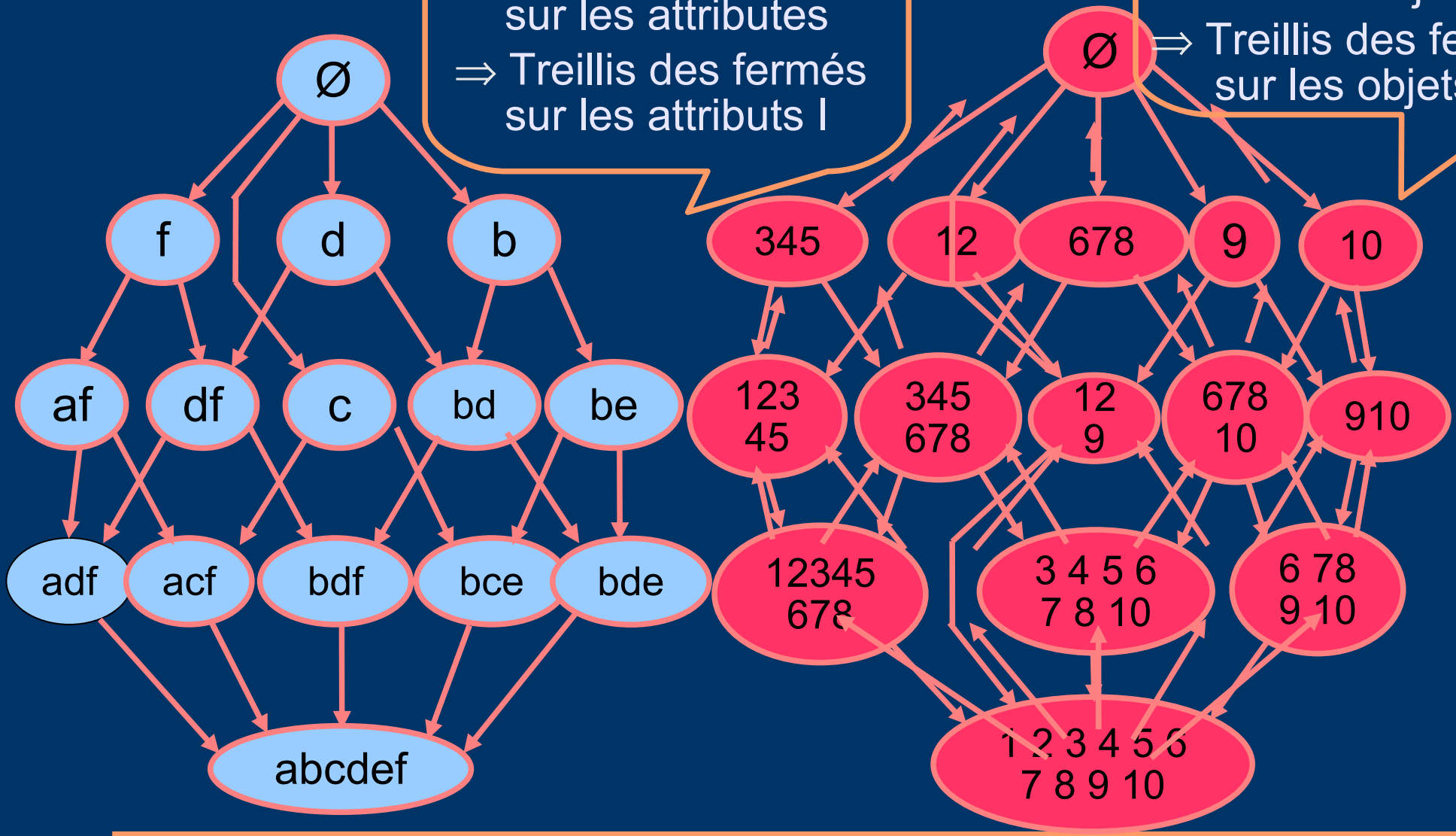
  - Muni d'une relation d'extension/subsomption $\leq$ entre les concepts:
    - $(A,B) \leqq (A',B') \Leftrightarrow A \subseteq A' \Leftrightarrow B \supseteq B'$

- Propriété: la relation $\leq$ sur l'ensemble des concepts est une relation d'ordre possédant la propriété de treillis

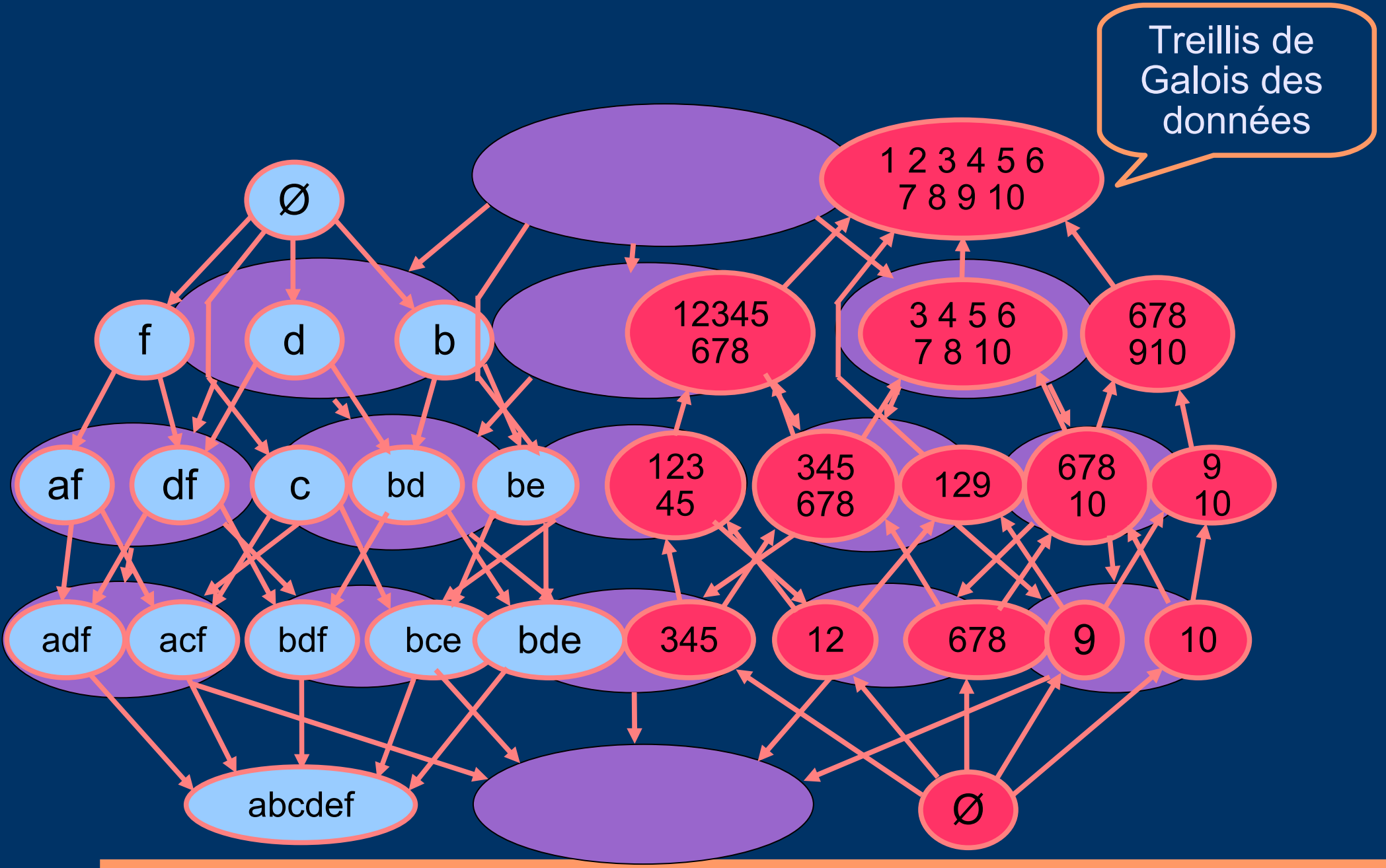
Treillis de Galois et treillis des fermés

Prop: φ=f o g est un operateur de fermeture sur les attributes
⇒ Treillis des fermés sur les attributs I

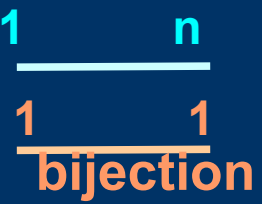
Prop: φ^{-1} est un operateur de fermeture sur les objects
⇒ Treillis des fermés sur les objets 0



Treillis de Galois et treillis des fermés



Théorie des treillis



Données binaires

Table (réduite)

Treillis de Galois
(≅ treillis des fermés)

Base canonique
(directe)

Base informative
(générique)

Règles d'implication

Règles d'association

Règles entre attributs

- Les corrélations entre les attributs peuvent s'exprimer par des règles:
  - Règle d'implication ou règle exacte:

    e $\longrightarrow$ b: « *tous les objets possédant l'attribut e possèdent également l'attribut b* »

    (les objets 9 et 10 possèdent e donc b)

  - Règle d'association ou règle approximative:

    d $\longrightarrow$ f: « une *majorité des objets possédant l'attribut d possèdent également l'attribut f* »

    (les objets 3,4,5,6,7,8,10 possèdent d, seul l'objet 10 ne possède pas f)

	a	b	c	d	e	f
1	X		X			X
2	X		X			X
3	X			X		X
4	X			X		X
5	X			X		X
6		X		X		X
7		X		X		X
8		X		X		X
9		X	X		X	
10		X		X	X	

Règles d'implication: définition [BarMon 70]

- Un système implicatif (IS) Σ est une relation binaire entre les parties d'un ensemble S:

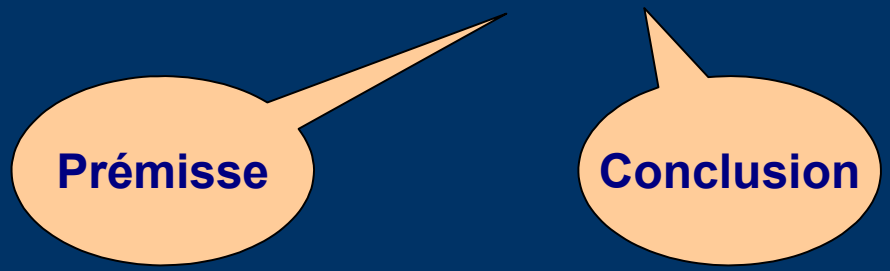
$$\Sigma \subseteq P(S) \times P(S)$$

- Un système implicatif unaire (UIS) Σ est une relation binaire entre les parties de S et S lui-même:

$$\Sigma \subseteq P(S) \times S$$

- Une règle implication est un couple d'un système implicatif (unaire)Σ:

$$(B,x) \in \Sigma \text{ noté } B \rightarrow x$$



Règles d'association: définition [Agrawal 94]

- Motif (itemset): ensemble d'attibuts
- Support du motif: proportion d'objets qui possèdent le motif par rapport à l'ensemble des objets
- Motif fréquent: son support est supérieur à un seuil de fréquence

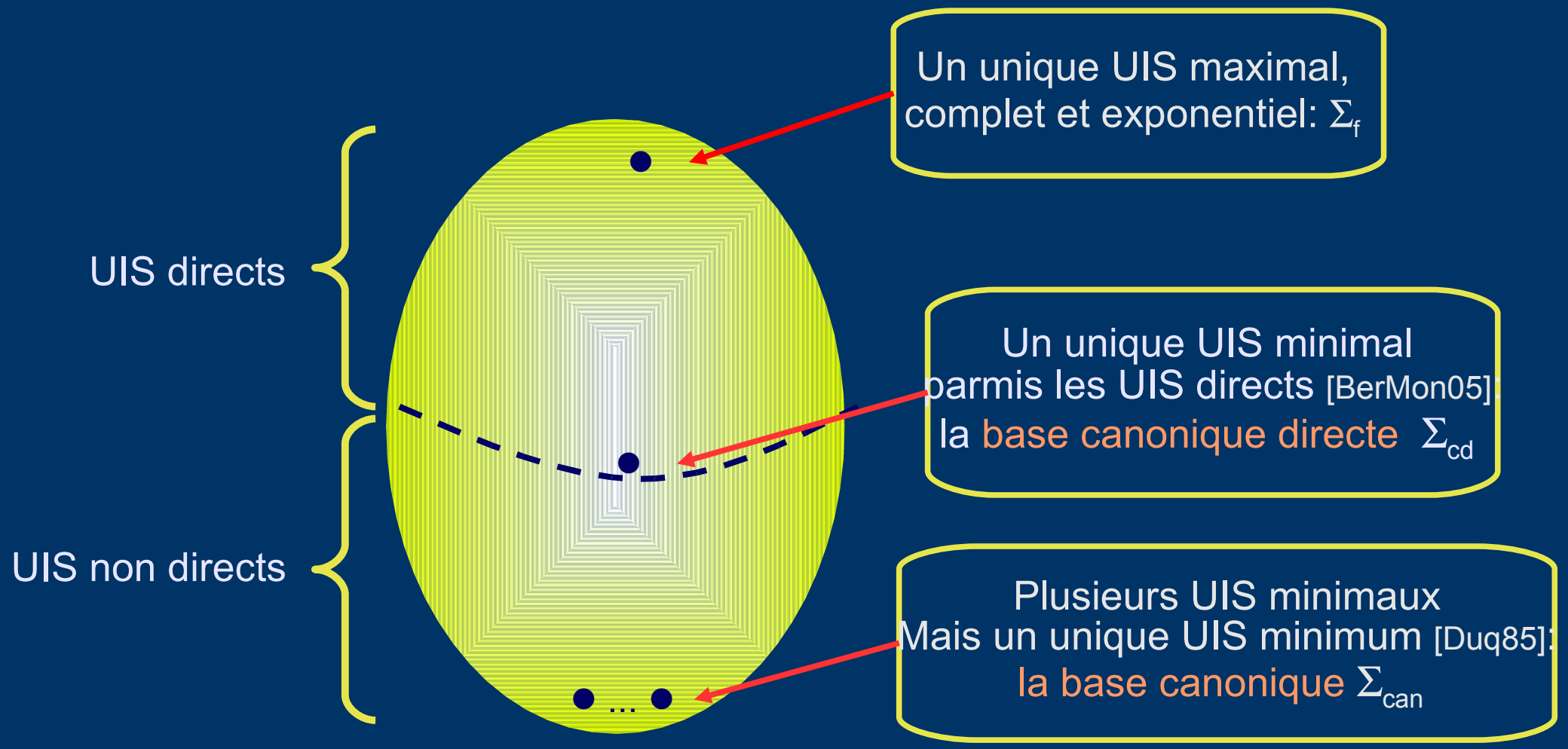
- Règle d'association: deux motifs $A \rightarrow B$
- Confiance d'une règle: support $(A \cup B)$/support (A)
- Règle valide: confiance supérieure à un seuil de confiance
- Règle exacte: confiance de 1

Bases de règles

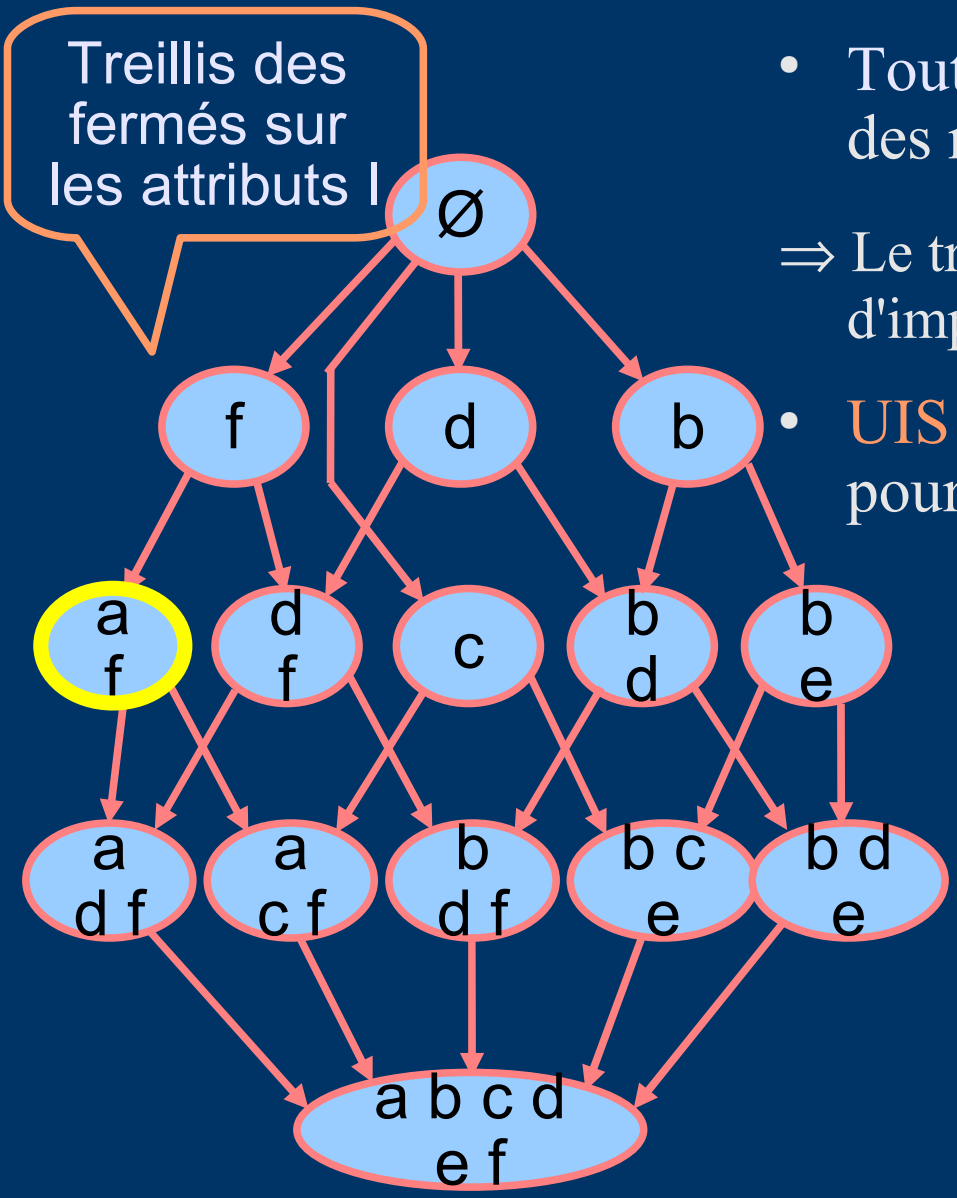
- Base: ensemble minimal de règles à partir desquelles on peut retrouver toutes les règles possibles par un mécanisme d'inférence (axiomes d'Armstrong, ...)
- Bases de règles d'implication:
  - La base canonique [Duq85] (ou Stem base, ou base de Duquenne-Guigues)
  - La base canonique directe que l'on retrouve [BerMon05]:
    - sous différentes terminologies: dépendances fonctionnelles, base minimale à gauche, base faible d'implications, ...
    - sous différentes formes: clauses de Horn, générateurs minimaux ...
- Bases de règles d'association: plusieurs, parmis lesquelles:
  - La base générique informative [Gas06:IGB] définie à partir des générateurs minimaux (prémisses de la base canonique directe) [BerHam08]

Bases de règles d'implication

Ensemble des UIS équivalents (i.e. représentant les mêmes données et le même treillis) ordonnés par inclusion:



Règles d'implication et treillis des fermés



- Tout fermé φ(X) s'obtient par « application » des règles d'implication

⇒ Le treillis des fermés est associé au système d'implication

- UIS direct: une seule passe sur les règles suffit pour calculer un fermé

Base canonique directe:

Σ_{cd} = {a→f, e→b, ef→a, cf→a,
cd→b, af→c, ab→c, bf→d,
ef→d, cd→e, bc→e, ab→e, cd→f }

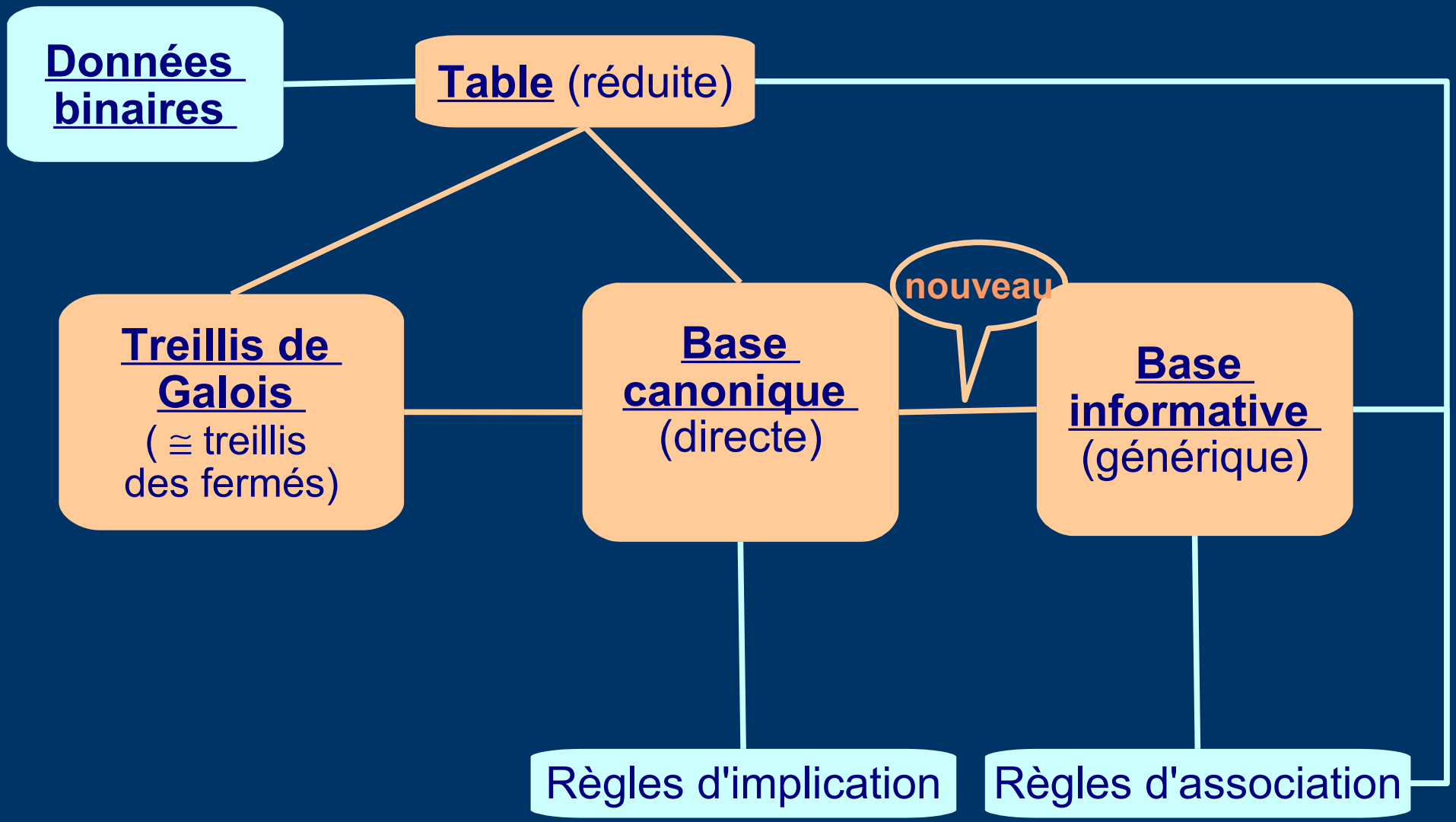
Fermé φ(a) = plus petit fermé contenant a
= application de la règle a→f

Théorie des treillis

1 n

1 1

bijection



Algorithmes de génération

- Risque algorithmique d'explosion combinatoire:
  - Nombre exponentiel de règles/concepts dans le pire des cas,

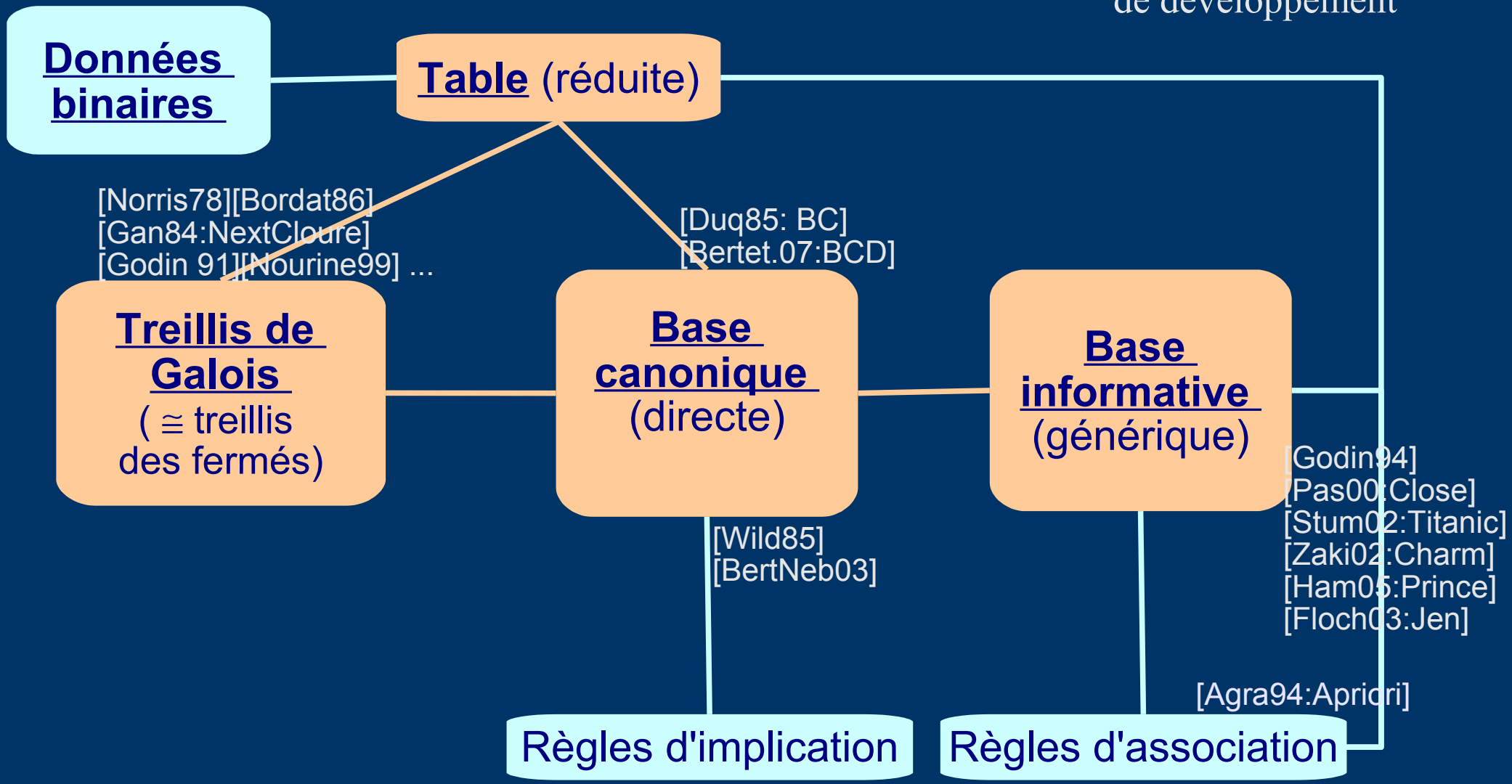
    .... mais polynimial en pratique
  - Génération polynomiale d'un concept.
  - Génération exponentielle d'une règle d'une base (problème ouvert)

- Utilisation en fouille de données:
  - du treillis de Galois (ou treillis des fermés)
  - des règles d'association (ou bases génériques)

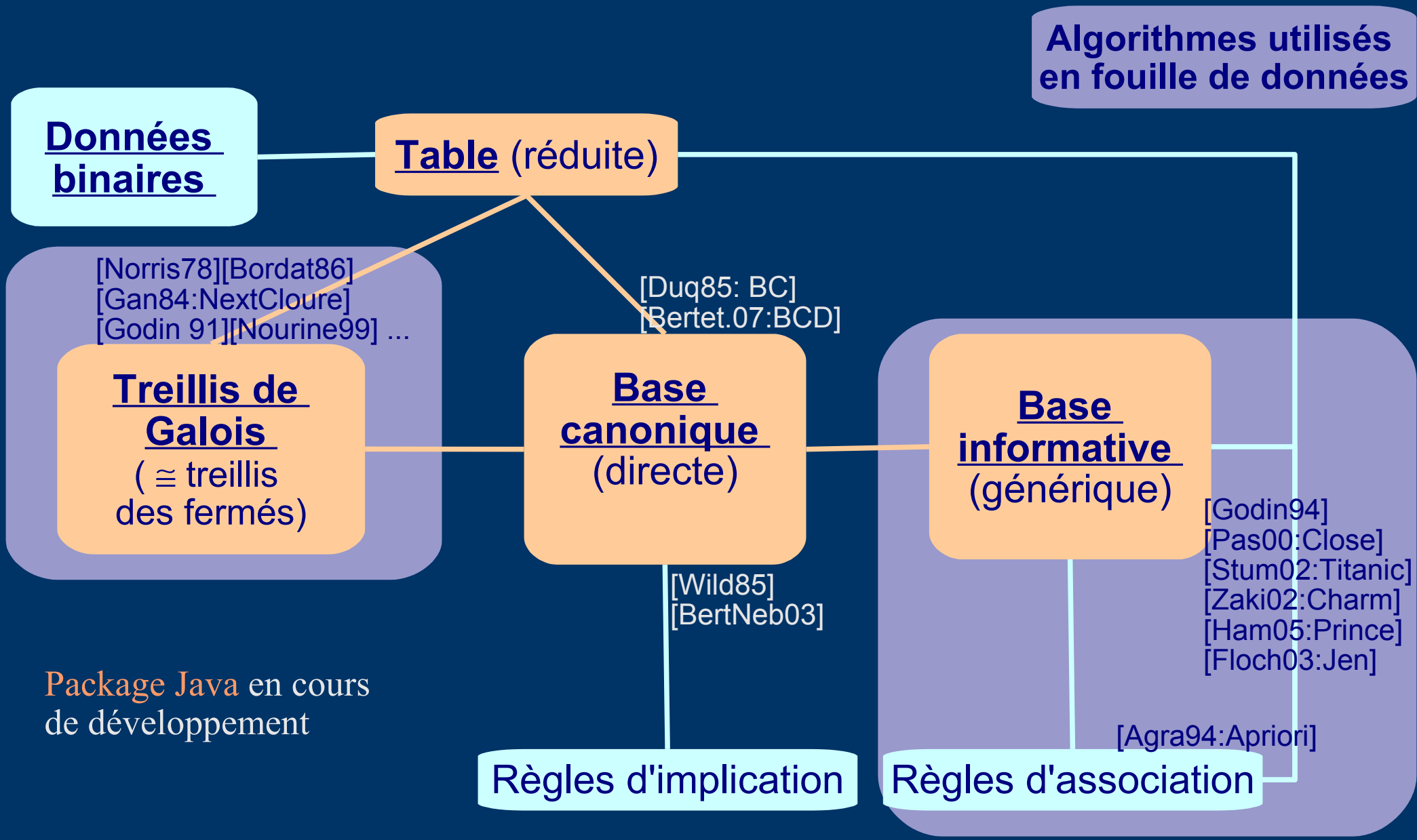
$\Rightarrow$ Intérêt algorithmiques croissant

Algorithmes de génération

Package Java en cours de développement



Algorithmes de génération



Package Java en cours de développement

Plan

- Théorie des treillis
  - Treillis de Galois
  - Règles d'association, d'implication
  - Algorithmes de génération
- Fouille de données
  - Méthodes utilisant des règles d'association
  - Méthodes utilisant un treillis
- Cas des images

Objectifs de la fouille de données

- Données divisées entre données d'apprentissage, de validation et de test
- Clasification supervisée:
  - Apprentissage: Construire une description de chaque classe à partir de données d'apprentissage
  - Classification: Associer une classe à des données de test
- Classification non supervisée ou segmentation:
  - Apprentissage: Regrouper les données d'apprentissage en clusters homogènes
  - Classification: Associer un cluster à des données de test
- Indexation:
  - Apprentissage: Identifier des clés (éléments significatifs) pour chaque classe (supervisé) ou cluster (non supervisé)
  - Indexation: Utiliser la clé pour retrouver des données de test

Méthodes de fouille de données

- Méthodes numériques:
  - Données: Pour des données numériques seulement
  - Méthodes: Réseaux bayésien, réseaux de neurones, k-ppv, algorithmes génétiques .....
- Méthodes symboliques:
  - Données: Intégration de données symboliques et numériques ... après discrétisation pour obtenir une table binaire
  - Méthodes: Arbre de décision, règles ou bases d'association, treillis de Galois, ...
- Comparaison:
  - Méthodes symboliques: plus lisible, permet le sélection d'attributs
  - Méthodes numériques: pas de perte d'information liée à la discrétisation

Fouille et règles d'association

- Apprentissage:
  1. Génération:
     - de l'ensemble des règles d'association (Apriori)
     - ou d'une base générique pour en limiter le nombre
  2. Sélection / élagage des règles par apprentissage / confiance / ...
  3. Possibilité de tri des règles par ordre de confiance / classe / cluster
- Classification:
  - Selon la première règle vérifiée / la règle majoritaire / la classe majoritaire / le cluster maoritaire

Fouille et règles d'association

- Méthodes:
  - Premières méthodes utilisent des règles d'association:

    CBA[Liu 98], CMAR[Li 01], ARC[AZ 02], ...
  - Plus récemment, utilisation d'une base générique:

    CPAT[XY 03], GARCm[BEBY 06] , ...
- Résultats:

  Taux comparables, voire supérieurs à la méthode numérique Bayésien et à la méthode symbolique C4.5 (arbre de décision).

Fouille et treillis de Galois

- Apprentissage:
  1. Calcul du treillis de Galois:
     - plusieurs concepts pour une même classe / cluster
  2. Possibilités de sélection de concepts:
     - élagage du treillis (Iceberg) / sélection de concepts / extraction de règles à partir des concepts
- Classification:
  - Par sélection: vote majoritaire à partir des concepts / règles sélectionnés
  - Par navigation: navigation type « arbre de décision » par validation d'attributs jusqu'à atteindre un concept associé à une classe / un cluster
    - Bouclage de pertinence (avec un utilisateur)
    - Génération à la demande des concepts (gain de temps et de place)

Fouille et treillis de Galois

- Méthodes:
  - Classification par sélection:

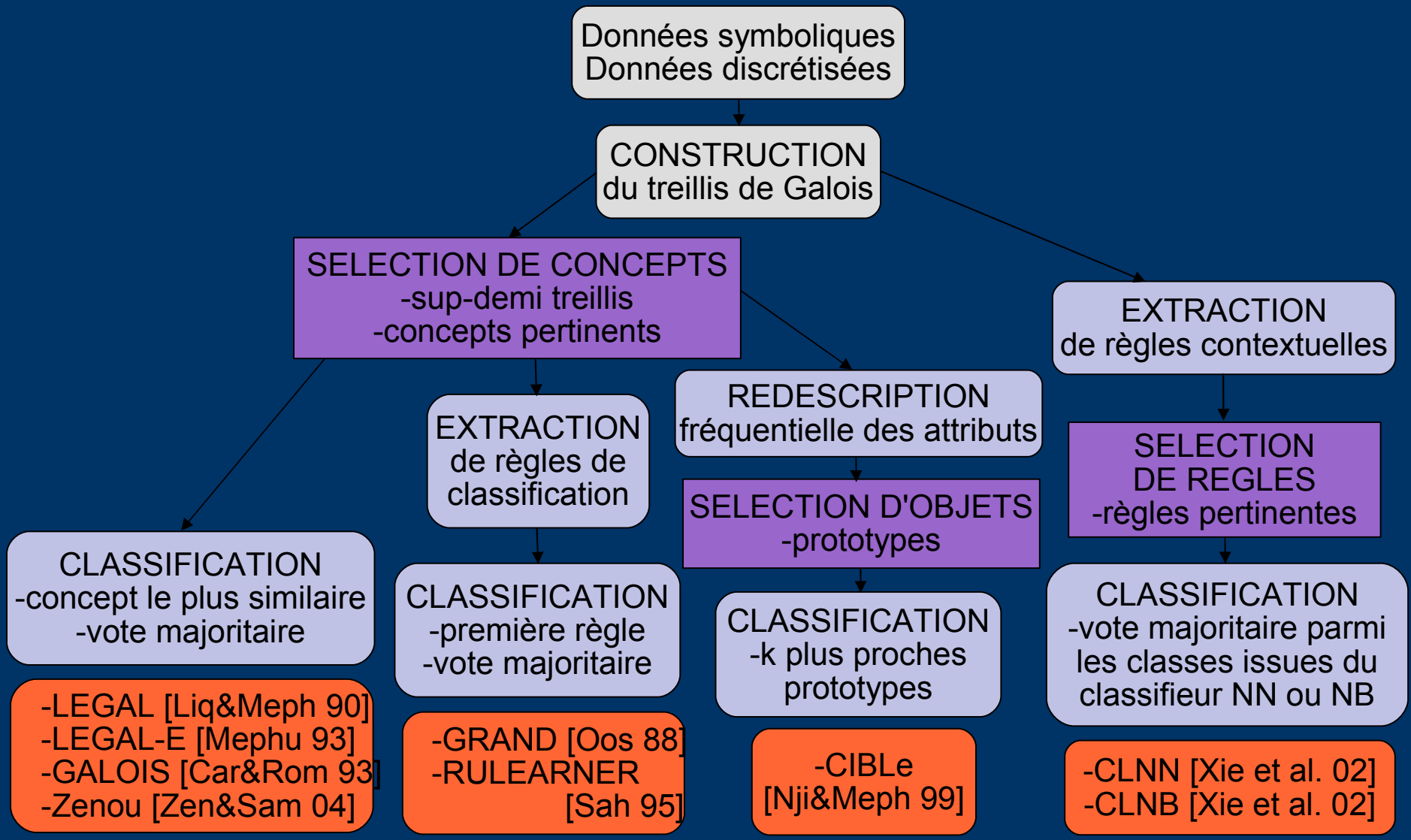
    Legal [Liq&Meph 90], Galois [Car&Rom 93], Zenou [Zen&Sam 04], Grand [Oos 88], Rulearner [Sah 95], Cible [Nji&Meph 99], CNN [Xie et al. 02]
  - Classification par navigation:

    Navigala [Guillas 07], Exploration [Eklund 06]

- Résultats:

  Taux comparables, voire supérieurs aux méthodes numériques: Bayésien, Perceptron et aux méthodes symboliques: C4.5 (arbre de décision), C4.5 Rules, AQ15, AQR (règles), Pebls (k-PPV)

Fouille et treillis de Galois



Plan

- Théorie des treillis
  - Treillis de Galois
  - Règles d'association, d'implication
  - Algorithmes de génération
- Fouille de données
  - Méthodes utilisant des règles d'association
  - Méthodes utilisant un treillis
- Cas des images

Cas des images: signatures

1. Extraction de signature:
   - Grande variété de signatures possibles:
     - globale / locale,
     - décrivant la couleur, la forme, la texture
     - dépendantes des images
     - parfois difficiles à paramétrer
     - ....

   $\Rightarrow$ Images décrites par un vecteur numérique de caractéristques

Cas des images: discrétisation

1. Discrétisation: Regroupement des caractéristiques en intervalles (flous)
   - Deux critères de discrétisation:
     - Critère de segmentation des données

     intègre ou non l'information de classe (cas supervisé ou non)
     - Critère d'arrêt

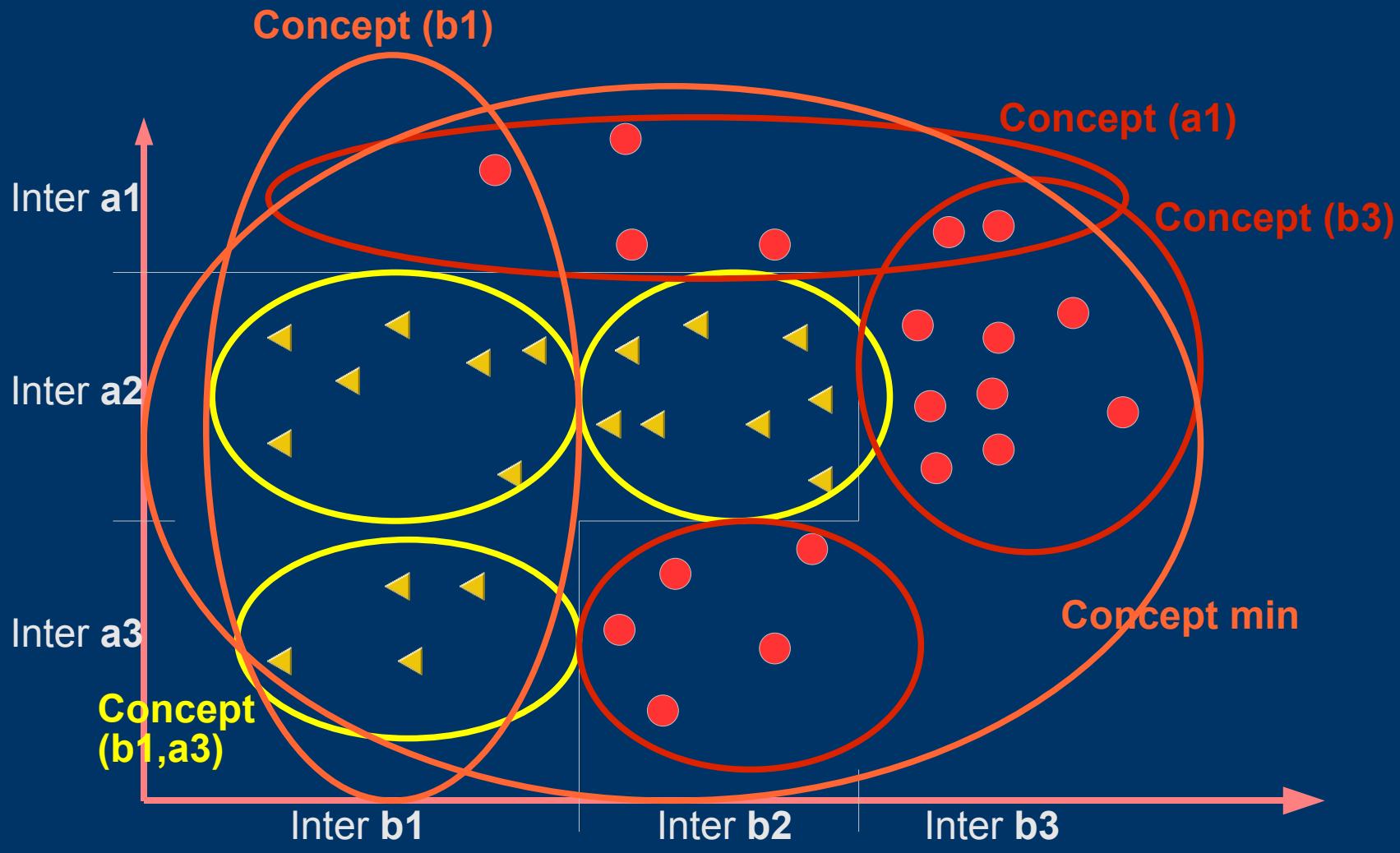
     nb d'étapes, séparation de classes, ...

   - Induit une sélection de caractéristiques:

   sélection des caractéristiques séparées en plusieurs intervalles
   - Pour des données linéairement séparables par morceaux

⇒ Image décrite par un ensemble d'intervalles

Données séparables par morceaux



Cas des images: treillis ou règles

1. Treillis ou règles d'assocation:
   - Les ensembles d'intervalles décrivant les images peuvent s'organiser sous forme d'une table:
     - Objets O : images ;
     - Attributs I : intervalles (flous) ;
     - Connexion de Galois (f,g) : association entre les images et les intervalles

⇒ Génération du treillis de Galois ou des règles d'association

Cas des images: utilisation

- Classement des images (supervisé ou non):
  - avec un treillis (sélection ou navigation)

    Reconnaissance de symboles détériorés: Navigala [Guillas 07]
  - avec des règles d'association

    Categorisation of documents: ARC [AZ 02]

- Indexation / exploration:
  - Avec un treillis:

    Exploration d'une base d'images [Eklund 06]

Cas des images: utilisation

- Sélection d'attributs:
  - Par discrétisation (caractéristiques séparées en plusieurs intervalles)
  - Par sélection de concepts ou de règles (intervalles associés)

    Amers visuels décrivants des pièces [Zen&Sam 04]

  Possibilité d'utiliser le treillis seulement pour sélectionner des attributs

  La sélection intègre les corrélation entre attributs

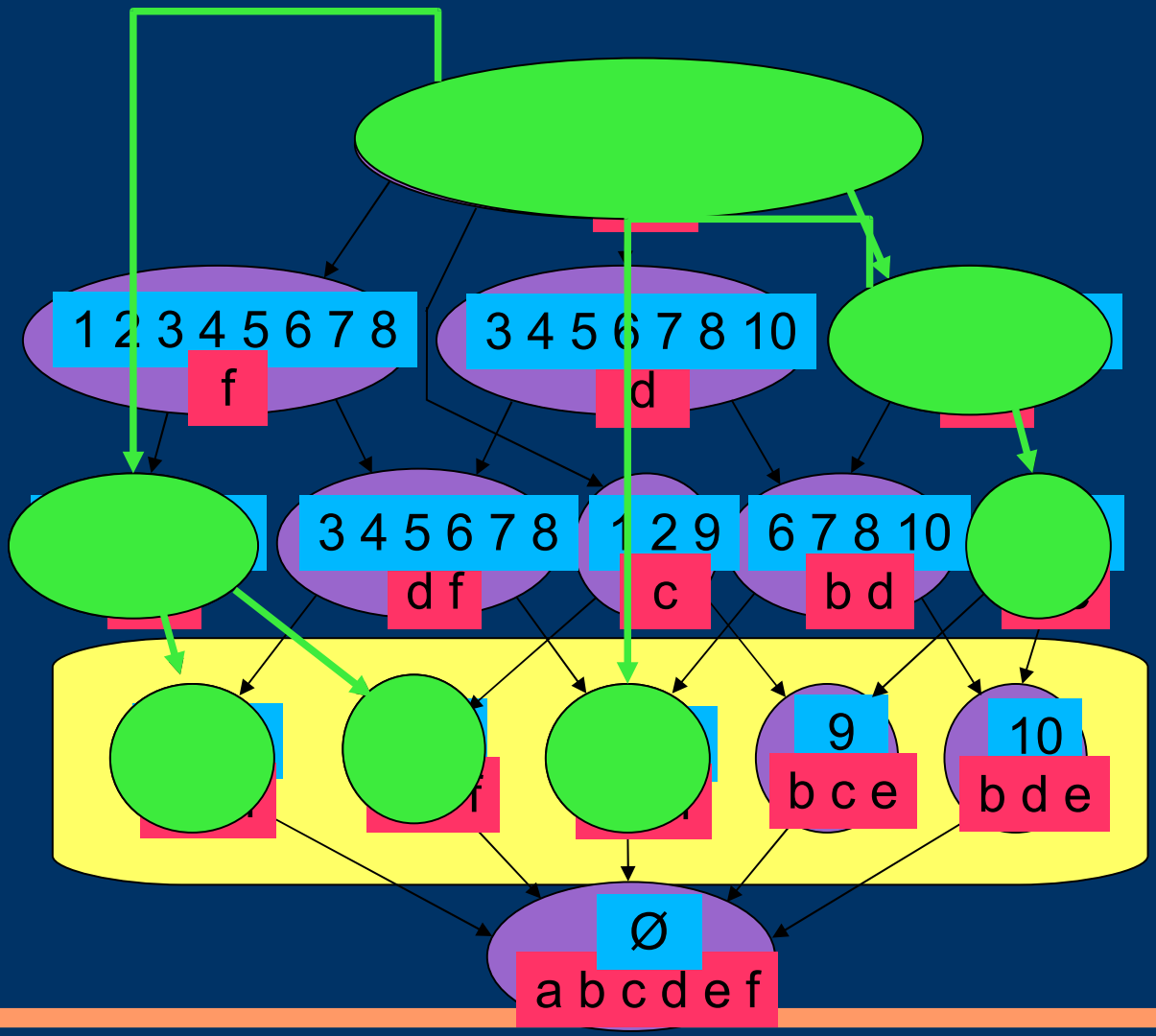
Treillis de Galois

Le treillis de Galois ainsi associé à des images possède deux propriétés:
co-atomisticité et inf-complémentarité

Propriété:
tout arbre de décision
est **inclus** dans un
treillis **V-complémentaire**

Théorème:
un treillis **V-complémentaire**
est la **fusion** de tous les
arbres de décision

Propriété:
on retrouve toutes les
classes dans les concepts
couvrant le concept max
d'un treillis **coatomistique**



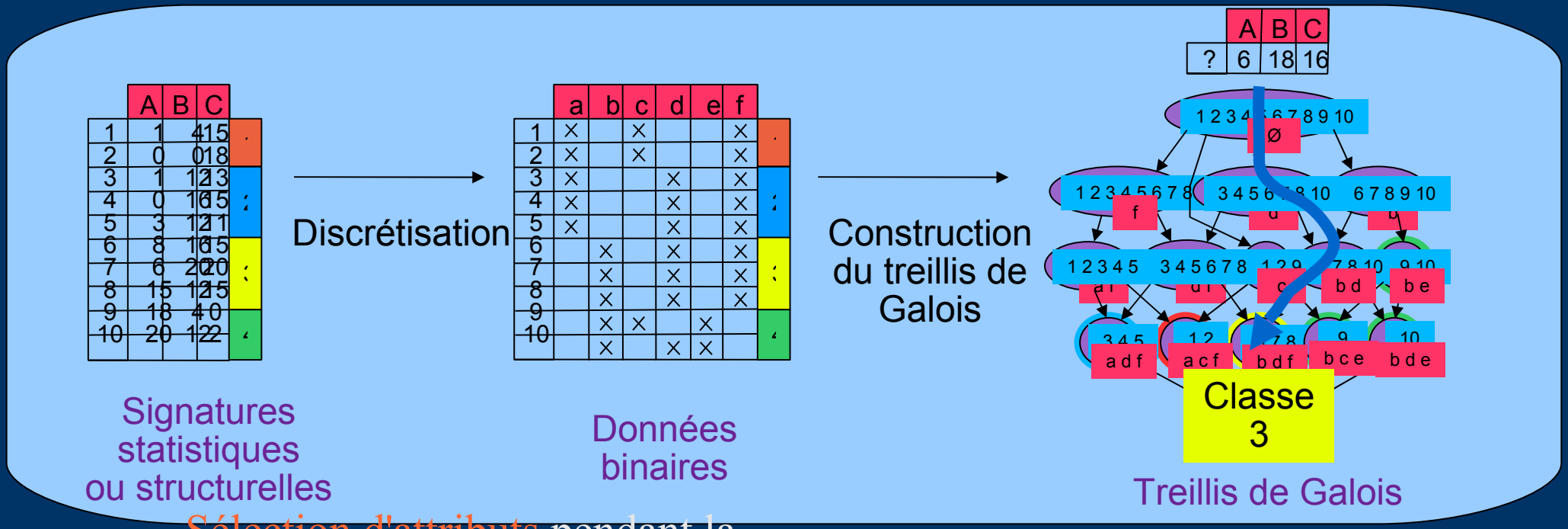
Méthode de classification Navigala [Guillas 07]

- Apprentissage:

  à partir de signatures de symboles

- Reconnaissance par navigation

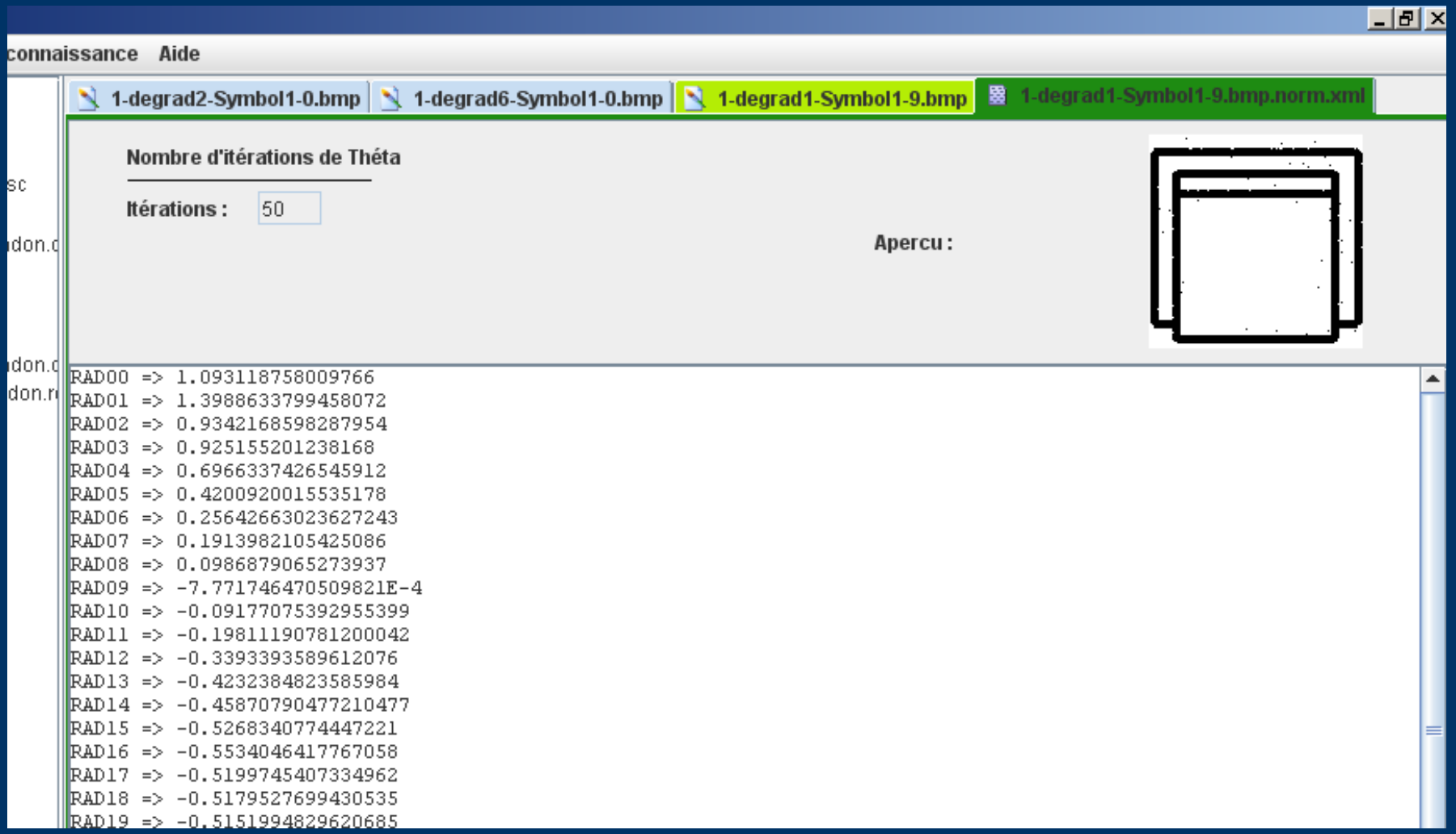
  d'un symbole détérioré



Sélection d'attributs pendant la discrétisation seulement

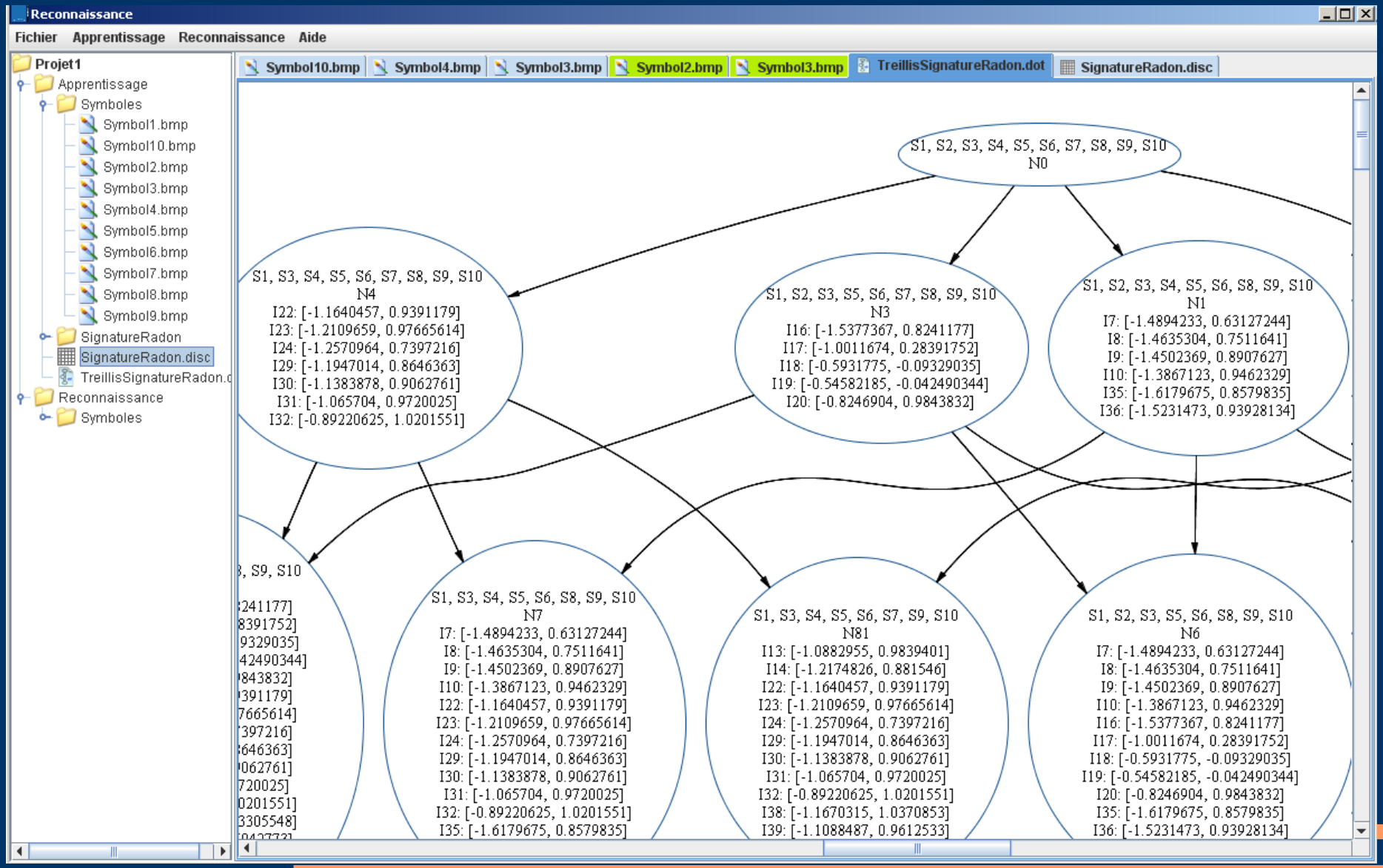
Méthode de classification Navigala

- Logiciel



Méthode de classification Navigala

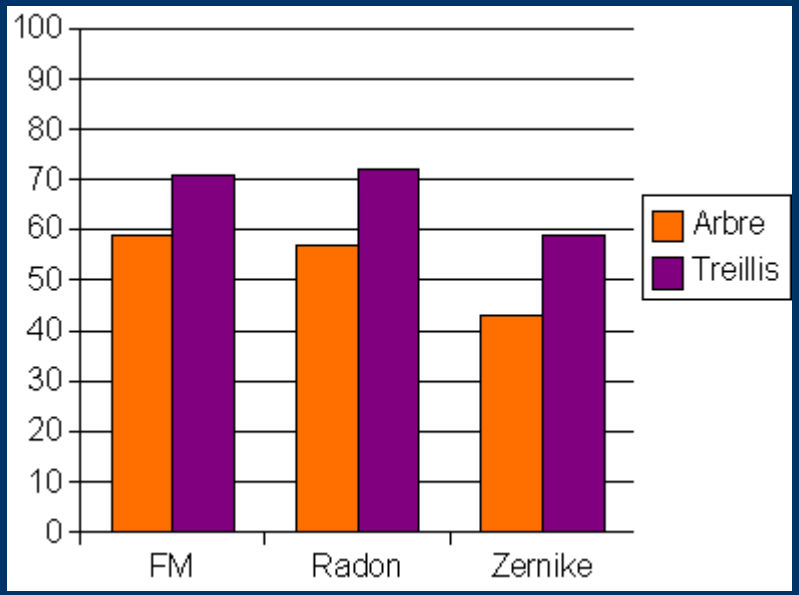
- Logiciel



Méthode de classification Navigala

- Comparaison expérimentale avec un arbre de décision

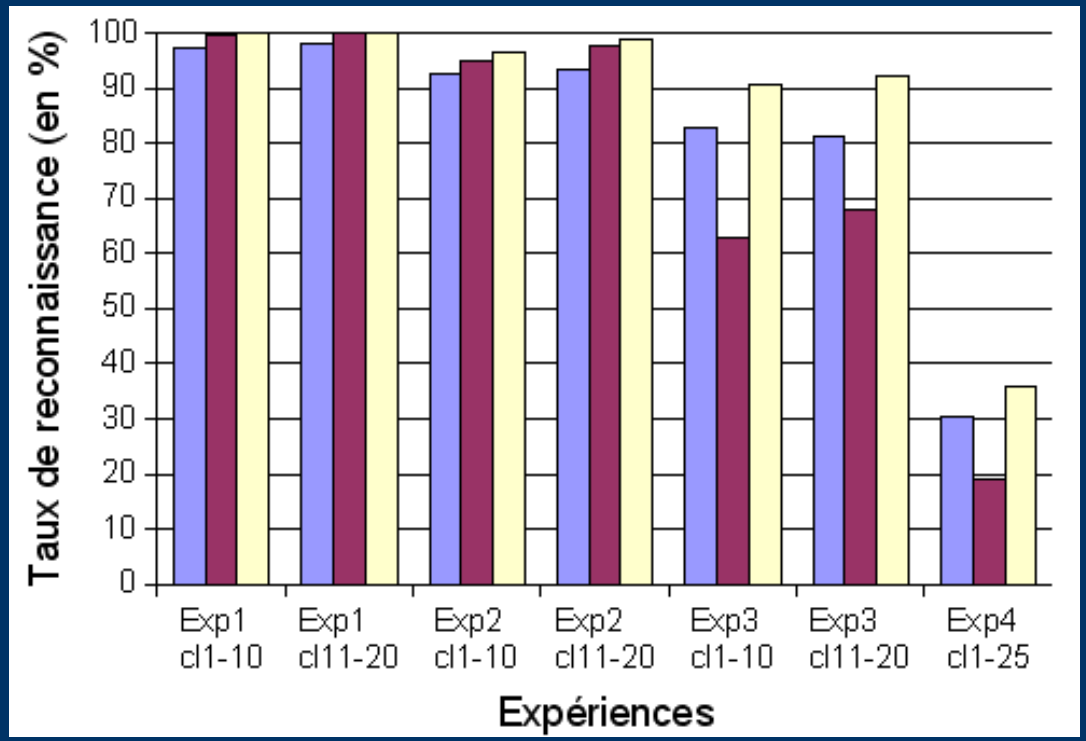
  Pourcentage de reconnaissance obtenus par un arbre de décision (CART) et Navigala à partir d'une même table discrétisée (symboles GREC 2003).



10 symboles modèles (10 classes)
900 symboles bruités

Méthode de classification Navigala

- Comparaison avec des classifieurs usuels en reconnaissance de formes



Navigala	7-8 attributs
Bayésien	50 attributs
K-PPV (k=1)	50 attributs

GREC 2003 : 2 x 10 classes
Exp1 : 5 blocs de 182 symboles
Exp2 : 10 blocs de 91 symboles
Exp3 : 26 blocs de 35 symboles

GREC 2005 : 25 classes
Exp4 : 5 blocs de 35 symboles

Conclusion

- Comment bien utiliser treillis et règles d'association ?
  - Dépend du problème:

    sélection ? classification ? indexation ?
  - Depend des données:

    sont-elles linéairement séparables ?

    une phase de sélection est-elle nécessaire ?
- Comment utiliser le bon algorithme ?
  - Quelques notions de théorie des treillis
  - Package Java en cours de développement

Perspectives

- Deux packages Java en cours de développement:
  1. Génération de quelques signatures à partir d'images (fini)
  2. Génération du treillis ou de la base de règles d'association à partir d'une table binaire (fichier texte)
- Sélection de caractéristiques:
  - étude expérimentale à envisager (SFS, SBS, SFFS, SFBE,...)
  - combinaison de caractéristques
- Algorithme incrémentaux
- Importance de la discrétisation en lien avec les travaux sur l'arbre de décision (thèse de Nathalie)
- Treillis associé à des données ordinales (nouveau)

Données ordinales [Nourine à paraitre]

	c1	c2
a	2	1
b	3	0
c	1	2
d	0	2

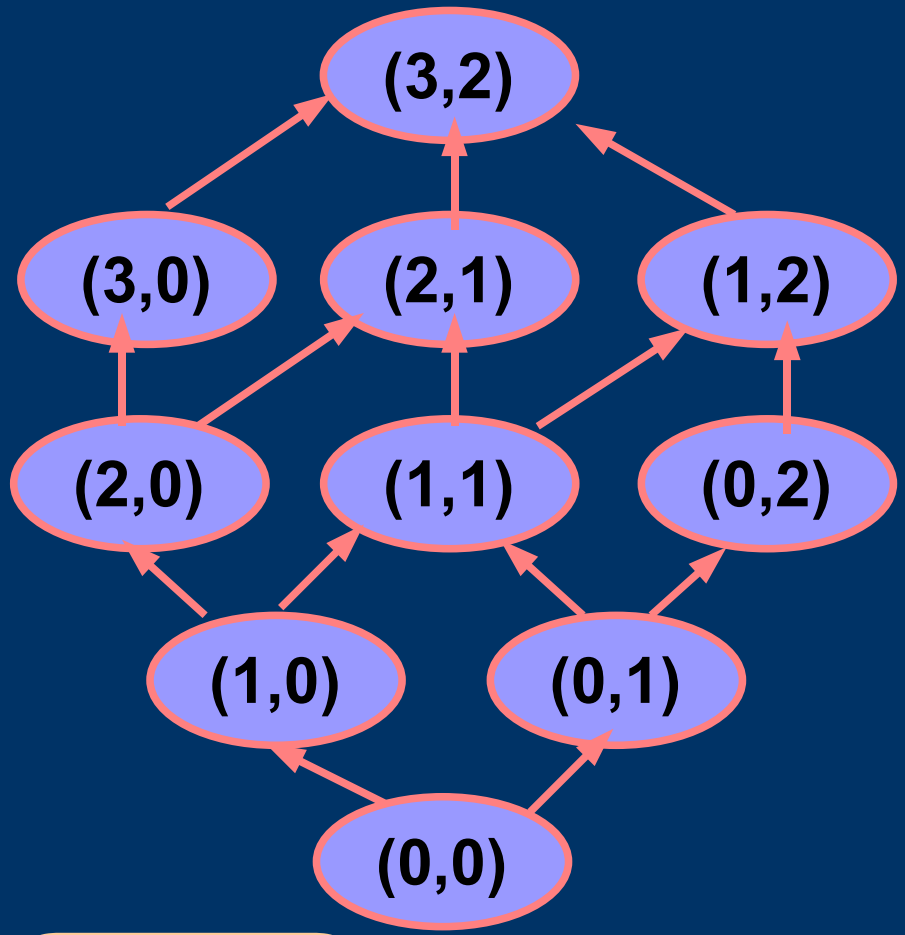


Table ordinale réduite
(≅ relation multi-valuée)

1 — **bijection** — 1

Treillis

Bibliographie

- Théorie des treillis: définitions
  - [BerMon 05] [CasMon 03] [Wille 99] [BarMon 70] [Duq 85] [Gas06:IGB] [Agra94] [Gas06:IGB]
- Théorie des treillis: algorithmes
  - [Bordat 86] [Ganter 84:NextClosure] [Godin 91] [Norris 78] [Duq 85] [Nourine99] [Wild 95] [TaouilBas 02] [BertNeb 03] [Floch:Jen 03]
- Classifiaction et data-mining:
  - [Car&Rom 93] [Ducrou 06] [Duquenne 07] [Zen&Sam 04] [Eklund 06] [Liq&Meph 90] [Rak 97] [Mephu 93] [Nji&Meph 99] [Oos 88][Sah 95] [Xie et al. 02] [Nourine à paraitre]
- Images:
  - [Guillas 05] [Guillas 07:Navigala] [Adam 01] [Llados 07] [Collin 93] [Fons 05] [Tabbone 06]