Extraction de Règles dans les Bases de Données Géographiques : Applications et Perspectives

Ansaf Salleb

Journée Fouille de Données

LIFO-MAPMO

25 mars 2004

Il était une fois...

les règles d'associations

Règles d'association

- *Item* un élément x_i de $\mathcal{I} = \{x_1, x_2, ..., x_n\}$
- *Itemset* un ensemble $X \in \mathcal{P}(\mathcal{I})$
- *Base de transactions*

$$\mathcal{D} = \{(tid, X_t) \ /tid \in T, \ X_t \in \mathcal{P}(\mathcal{I})\}$$

- *Support(X)*$=\frac{|\{(t,X_t)\in\mathcal{D}/X\subseteq X_t\}|}{|\mathcal{D}|}$
- *Itemset fréquent* $X \in \mathcal{P}(\mathcal{I})$ est fréquent ssi $support(X) \geq \gamma$, où γ seuil de support minimum

Règles d'association

Item	Film	Réalisateur
x_1	Harry Potter	C. Columbus
x_2	Star Wars II	G. Lucas
x_3	Attrape moi si tu peux	S. Spielberg
x_4	Un homme d'exception	R. Howard

$$\mathcal{I} = \{x_1, x_2, x_3, x_4\}$$
$$T = \{1,\ 2,\ 3,\ 4, \ldots,\ 15\}$$
$$\mathcal{D} = \{(1, x_1x_2), (2, x_1x_3), (3, x_3x_4), \ldots, (14, x_1x_3x_4), (15, x_3x_4)\}$$

$\mathcal{D}$	
Tid	**Transaction**
1	x_1, x_2
2	x_1, x_2, x_4
3	x_1, x_2
4	x_3, x_4
5	x_1, x_2
6	x_3, x_4
7	x_1, x_3, x_4
8	x_1, x_2, x_3
9	x_1, x_2
10	x_1, x_3, x_4
11	x_1, x_2
12	x_1, x_2, x_3
13	x_1, x_2
14	x_1, x_3, x_4
15	x_3, x_4

- **Phase I** Trouver les itemsets fréquents : $\mathcal{F}$

$$\mathcal{F} = \{ X \subseteq \mathcal{I} | \text{ support(X)} \geq \gamma \}$$

avec γ seuil de support minimum

- **Phase I** Trouver les itemsets fréquents : $\mathcal{F}$

$$\mathcal{F} = \{ X \subseteq \mathcal{I} | \text{ support(X)} \geq \gamma \}$$

avec γ seuil de support minimum

- **Phase II** Trouver les règles solides : $\mathcal{R}$

$$\mathcal{R} = \{ r : A \rightarrow C, \frac{supp(A \cup C)}{supp(A)} \geq \varphi \}$$

avec φ seuil de confiance minimum

Règles d'association

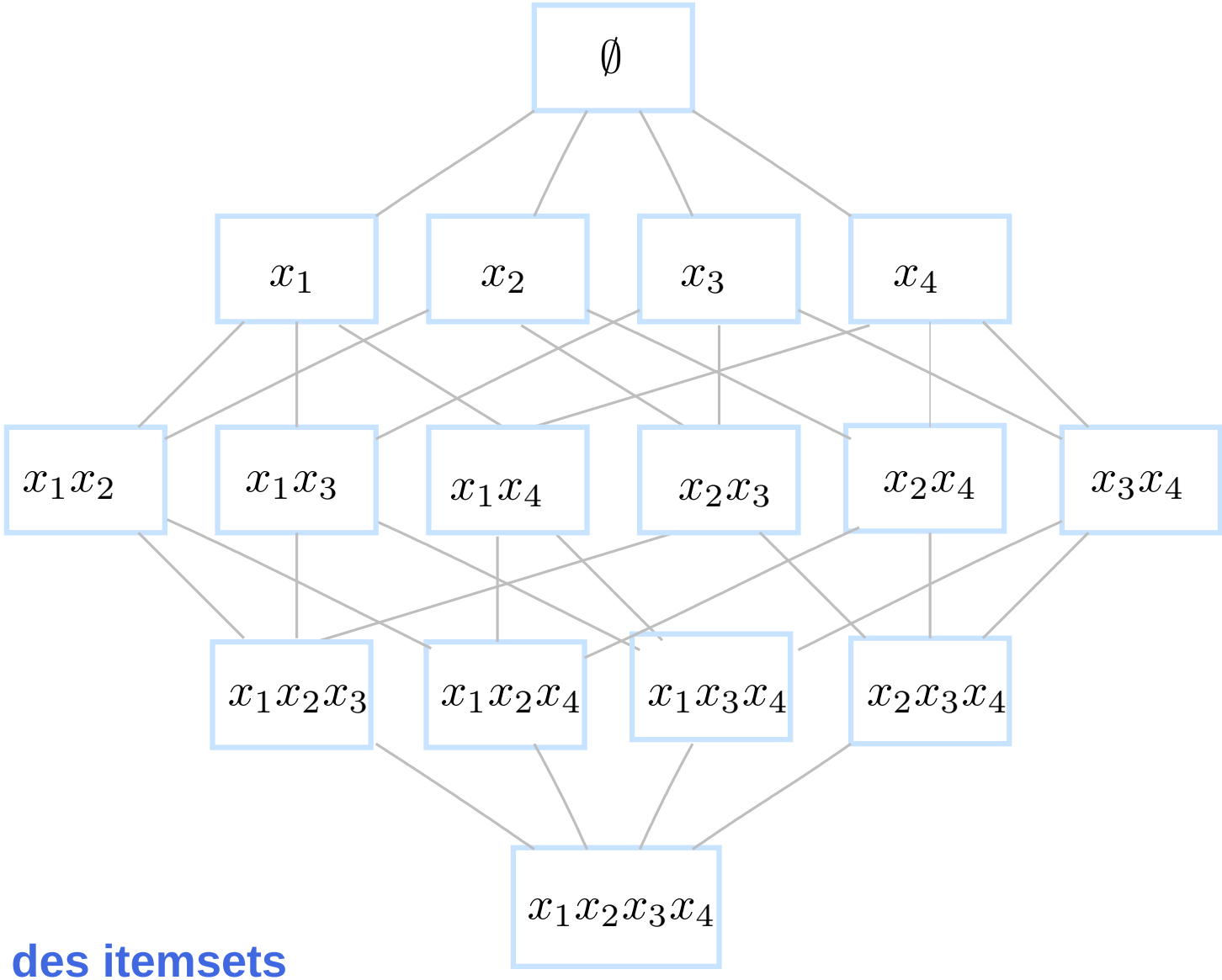
- **Phase I** $\gamma = 20\%$

$$\mathcal{F} = \{x_1(80\%),\ x_2(60\%),\ x_3(53.33\%),\ x_4(46.66\%),\ x_1x_2(60\%),\ x_1x_3(33.33\%),\ x_1x_4(26.66\%),\ x_3x_4(40\%),\ x_1x_3x_4(20\%)\}$$

- **Phase II** $\varphi = 70\%$

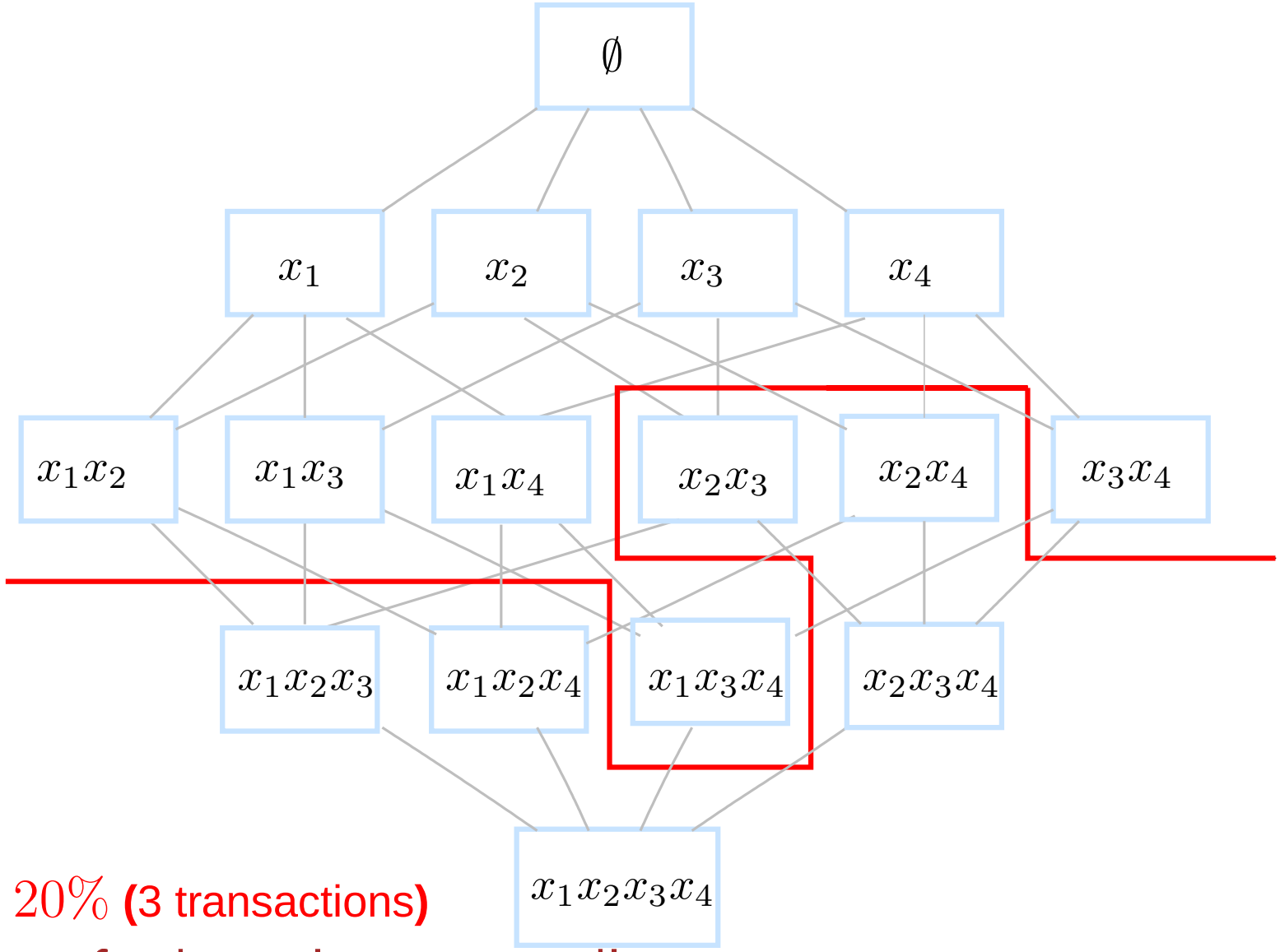
$$\begin{aligned}\mathcal{R} = \{\ & x_2 \longrightarrow x_1 && (100\%)\\ & x_4 \longrightarrow x_3 && (85,71\%)\\ & x_1 \longrightarrow x_2 && (75\%)\\ & x_3 \longrightarrow x_4 && (75\%)\\ & x_1x_4 \longrightarrow x_3 && (75\%)\\ \}\end{aligned}$$

Règles d'association



Treillis des itemsets

Règles d'association

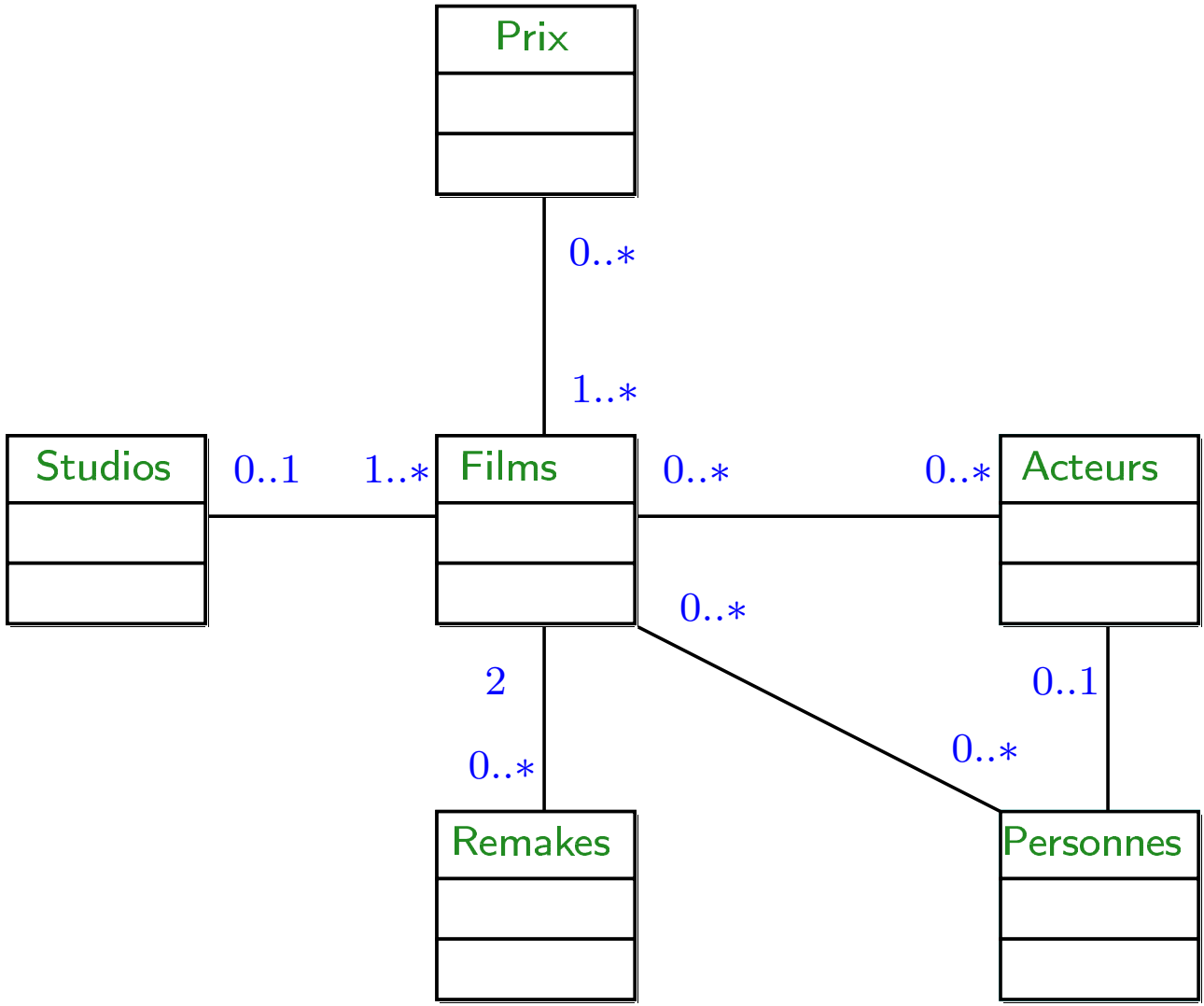


$\gamma = 20\%$ **(3 transactions)**

Élagage fondé sur la propriété d'antimonotonie

- **Problèmes**
  - Le nombre de règles peut être très très grand !
  - Tâche coûteuse en temps et espace mémoire
  - Extension au relationnel difficile...
- **Solutions possibles**
  - Cibler la recherche, augmenter les seuils, utiliser d'autres mesures,...
  - Algorithmes de recherche intelligents
  - Aplatir les relations mais perte d'information :(

Règles de caractérisation : exemple



Puis naquit...

CaractériX

Règles de caractérisation

Turmeaux, Salleb, Vrain, Cassard (PKDD'03, EGC'03)

- *Soit une BDR décrivant des objets $\mathcal{E}$ et leurs relations $\mathcal{R}$*
- *$\mathcal{E}$ un ensemble d'objets*

  $\mathcal{E} = \mathcal{E}_1 \cup \mathcal{E}_2 \cdots \cup \mathcal{E}_n$ *où chaque* $\mathcal{E}_i$ *de même type* T_i
- *$\mathcal{R}$ ensemble de relations binaires*
- *Caractériser* $\mathcal{E}_{cible}$
  - *propriétés des objets cibles*
  - *propriétés des objets liés.*

Règles de caractérisation

- Cible $\mathcal{E}_{cible}$
- Chemin quantifié δ

$$\delta = Q_1\ \mathcal{E}_1 \ldots Q_n\ \mathcal{E}_n$$

$F : \forall R$, $F : \exists R$, $P_{nom=Hitchcock} : \forall F \forall R$

- Propriété p

$R.type \in \{oscar, palme\ d'or\}$

- Règle caractéristique

$$\boxed{\mathcal{E}_{cible} : \delta :: p}$$

$P_{nom=Hitchcock} : \forall F :: F.genre = Suspense$

Règles de caractérisation

- Couverture

$$couverture(r, \mathcal{E}_{cible}) = \frac{|\{o | o \in \mathcal{E}_{cible} \ et \ \mathcal{V}_r(o) = vrai\}|}{|\mathcal{E}_{cible}|}$$

- Exemples

$$P_{nom=Hitchcock} : \forall F :: F.genre = Suspense$$

$$P_{nom=Hitchcock} : \exists F :: F.genre = Suspense$$

$$P_{réalisateurs} : \exists F \exists R :: R.type \in \{oscar, palme \ d'or\}$$

Applications aux données géographiques

Le SIG Andes

SIG

Un Système d'Information Géographique (SIG) désigne un système permettant de gérer des données géographiques spatialement référencées, organisées en couches, ainsi que les descriptions des objets.

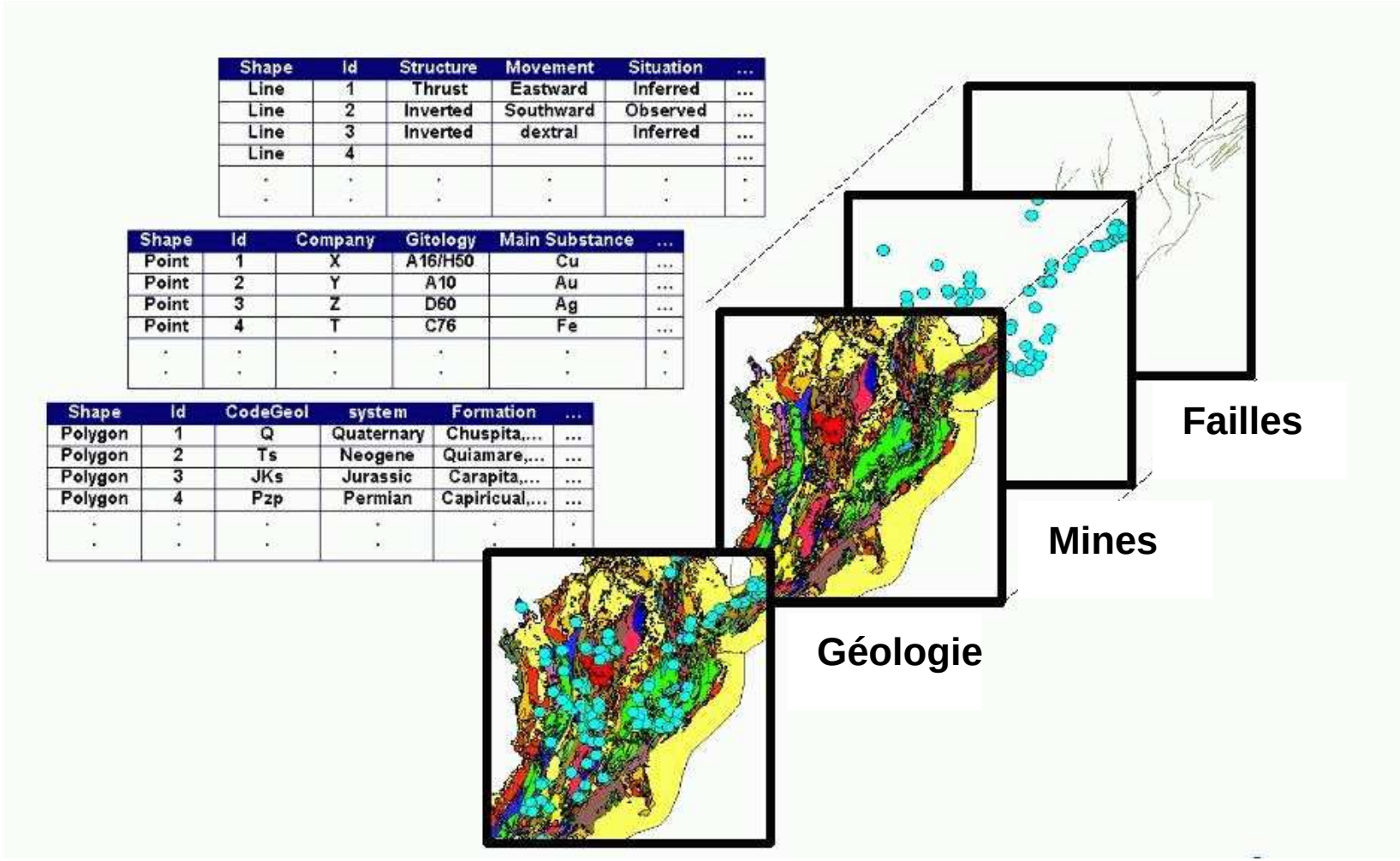
Le SIG Andes

SIG Un Système d'Information Géographique (SIG) désigne un système permettant de gérer des données géographiques spatialement référencées, organisées en couches, ainsi que les descriptions des objets.

SIG Andes Le SIG Andes est un Système d'Information homogène à vocation géologique et métallogénique, portant sur la totalité de la Cordillère des Andes (Cassard 1999). Parmi les 14 couches :

- Couche géographique
- Couche MNT (modèle numérique de terrain)
- Couche de synthèse géologique des Andes à l'échelle de 1/2 000 000
- Couche sismique
- Couche volcanique
- Couche gravimétrique
- Couche géochimique
- Couche des gisements
- etc.

Applications au SIG Andes (BRGM REM/VADO)



Shape	Id	Structure	Movement	Situation	...
Line	1	Thrust	Eastward	Inferred	...
Line	2	Inverted	Southward	Observed	...
Line	3	Inverted	dextral	Inferred	...
Line	4				...
.	.	.	.	.	.
.	.	.	.	.	.

Shape	Id	Company	Gitology	Main Substance	...
Point	1	X	A16/H50	Cu	...
Point	2	Y	A10	Au	...
Point	3	Z	D60	Ag	...
Point	4	T	C76	Fe	...
.	.	.	.	.	.
.	.	.	.	.	.

Shape	Id	CodeGeol	system	Formation	...
Polygon	1	Q	Quaternary	Chuspita,...	...
Polygon	2	Ts	Neogene	Quiamare,...	...
Polygon	3	JKs	Jurassic	Carapita,...	...
Polygon	4	Pzp	Permian	Capiricual,...	...
.	.	.	.	.	.
.	.	.	.	.	.

Organisation en couches du SIG Andes

SIG Andes : extraction de règles d'association

Application I ARGIS et ARGIS_PPV Salleb et Vrain (PKDD'00)

Forme	id_{Mine}	Gitologie	Substance
point	1	A16	Cu
point	2	A10	Au
point	3	D60	Ag
point	4	C76	Fe

Forme	$id_{Géol}$	Code	Système
polygone	1	Q	Quaternaire
polygone	2	Ts	Néogène
polygone	3	Qv	Quaternaire
polygone	4	Pzp	Permien

id_{Mine}	relation spatiale	$id_{Géol}$
1	inclus_dans	350
2	inclus_dans	102
2	proche_de	3
3	inclus_dans	1890

SIG Andes : extraction de règles d'association

Cassard et al. (EUG XI), Lips et al. (Gis in Geology 02)

- **Règles statistiques** :

$$Mine(x) \wedge Gitologie(x, A) \rightarrow Gitologie(x, A1) \quad (92.12\%)$$

- **Règles de contrôle** :

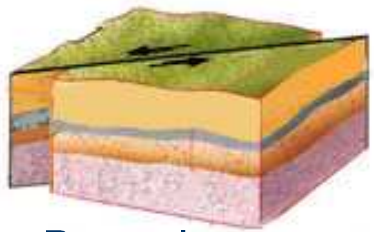
$$Mine(x) \wedge Gitologie(x, H12) \rightarrow \ Substance\_principale(x, AU) \quad (89.32\%)$$

- **Nouvelles règles** :

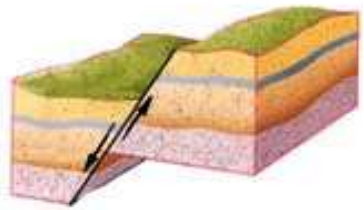
$$Mine(x) \wedge Faille(z) \wedge Gitologie(x, porphyrique) \wedge Proche\_de(x, z)$$

$$\rightarrow Structure(z, décrochement) \quad (43.75\%)$$

Interprétation



Décrochement



Faille normale

Faille inverse

SIG Andes : extraction de règles de caractérisation

Application II CaracteriX Turmeaux, Salleb, Vrain et Cassard (PKDD'03, EGC'03)

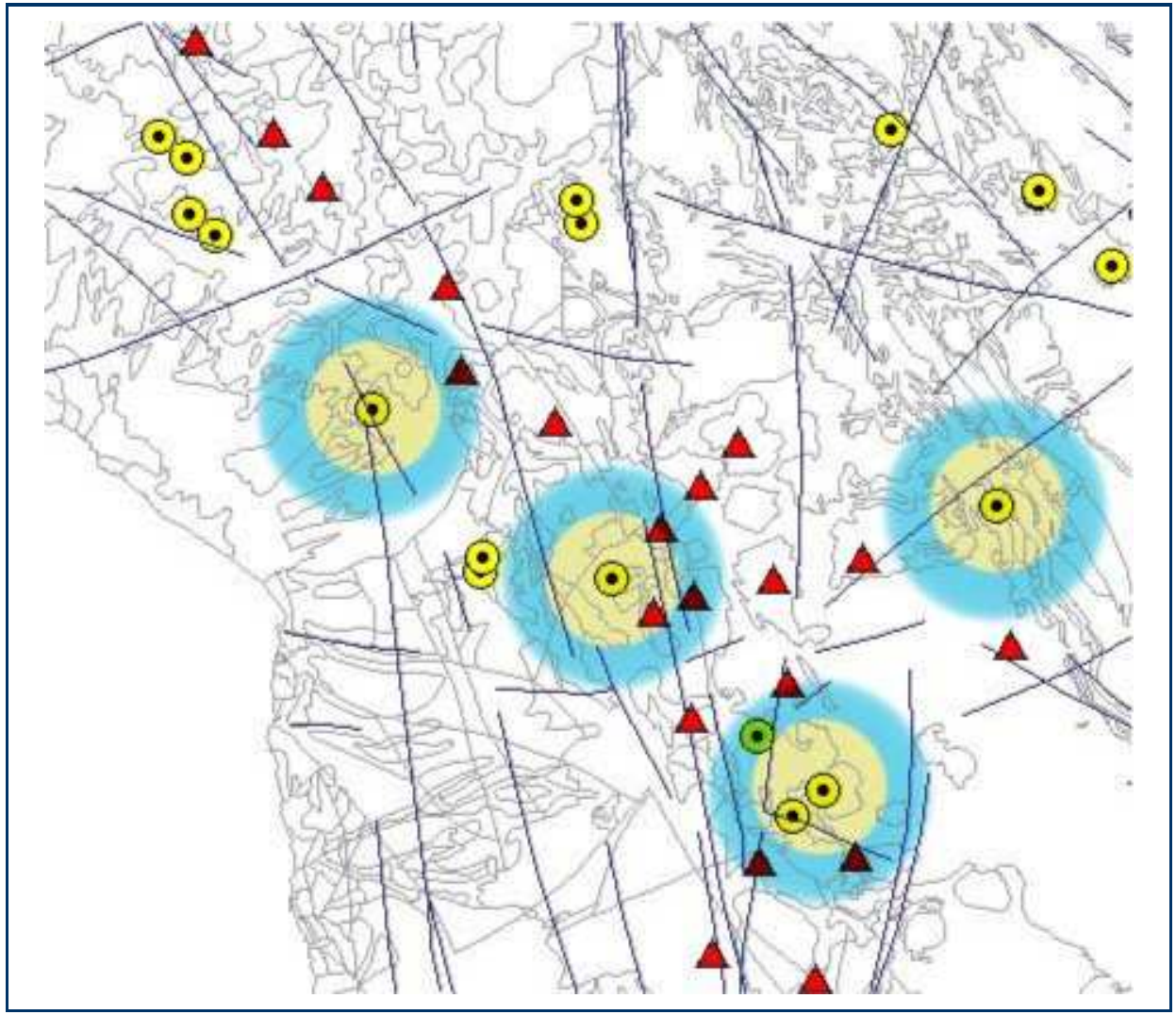
- $\mathcal{E} = \mathcal{E}_1 \cup \mathcal{E}_2 \cup \mathcal{E}_3 \cup \mathcal{E}_4$:
  - $\mathcal{E}_1$: *mines*
  - $\mathcal{E}_2$: *géologie*
  - $\mathcal{E}_3$: *volcans*
  - $\mathcal{E}_4$: *failles*
- $\mathcal{R}$: relations de distance entre objets

Caractériser $\mathcal{E}_{cible} = \{$ *mines d'or* $\} \subseteq \quad \mathcal{E}_1$

- Comment ? construction progressive de buffers croissants autour des objets cibles

$$M : \forall_{3Km} V \succeq M : \forall_{5Km} V \succeq M : \forall_{10Km} V$$
$$M : \exists_{10Km} V \succeq M : \exists_{5Km} V \succeq M : \exists_{3Km} V$$

SIG Andes : extraction de règles de caractérisation



SIG Andes : extraction des règles de caractérisation

$M : \exists_{10km}\ G ::$

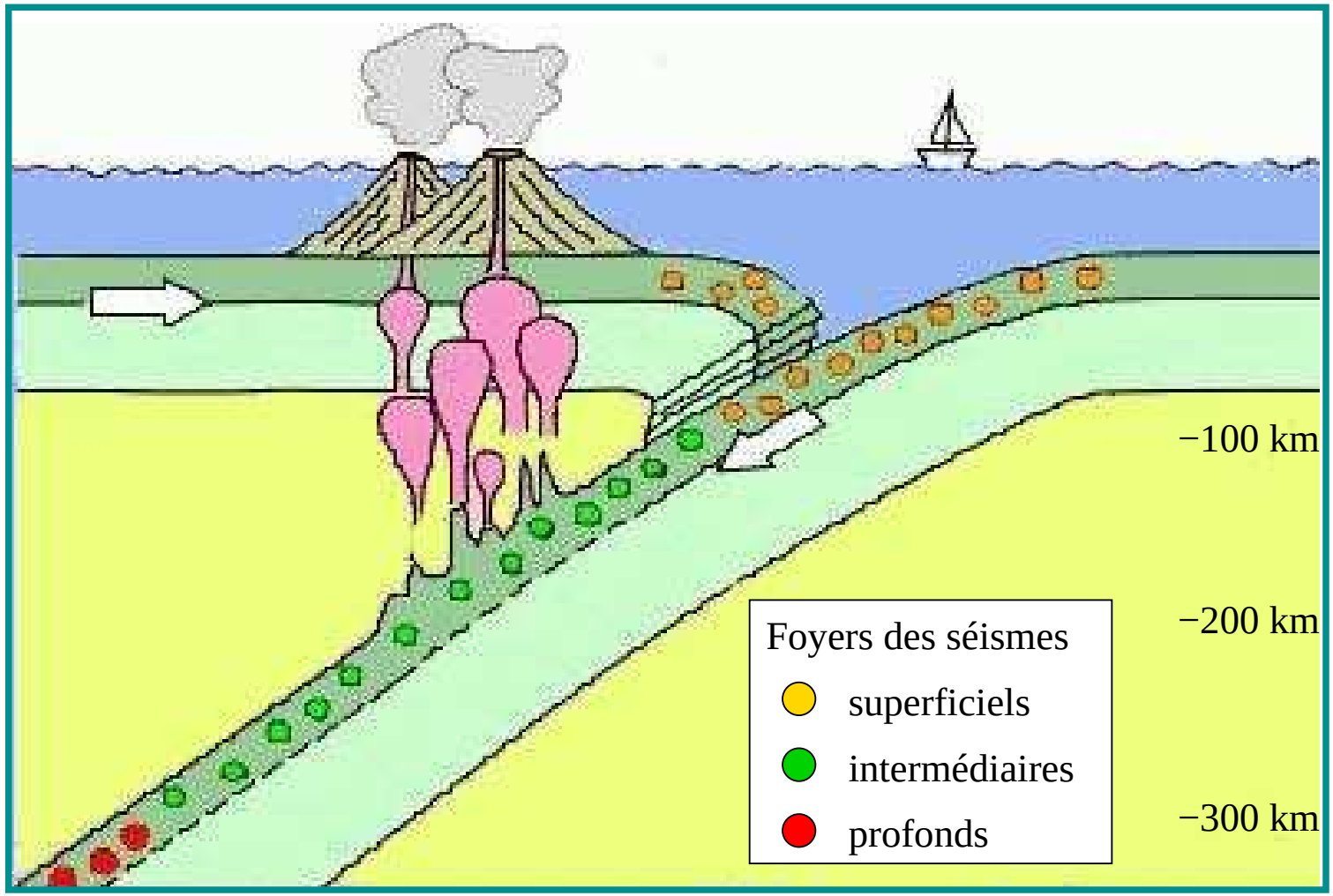
$M.\text{Substance principale} = \text{Au} \wedge$
$M.\text{Profondeur\_Benioff} \in [75\ ..\ 150] \wedge$
$M.\text{Distance\_Benioff} \in [170\ ..\ 275] \wedge$
$M.\text{Pente\_Benioff} \in [8°\ ..\ 16°] \wedge$
$M.\text{Lithologie} = \text{volcanique} \wedge$
$M.\text{Gitologie} = \text{épithermale} \wedge$
$M.\text{Morphologie} = \text{veines} \wedge$
$G.\text{Système} = \text{Volcanique} \wedge$
$G.\text{Age} = \text{tertiaire}$

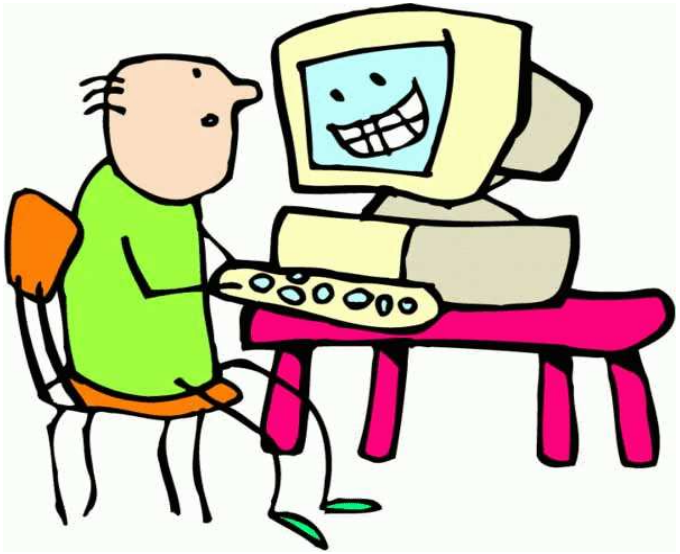
SIG Andes : extraction des règles de caractérisation

Interprétation

SIG Andes : extraction des règles de caractérisation

Interprétation





© prescolaire.grandmonde.com

Data Miner

Real Miner

à suivre...