De la veille à l'intelligence économique : le Data Mining et le Text Mining

Bernard DOUSSET

dousset@irit.fr http://atlas.irit.fr

Institut de Recherche en Informatique de Toulouse (IRIT)

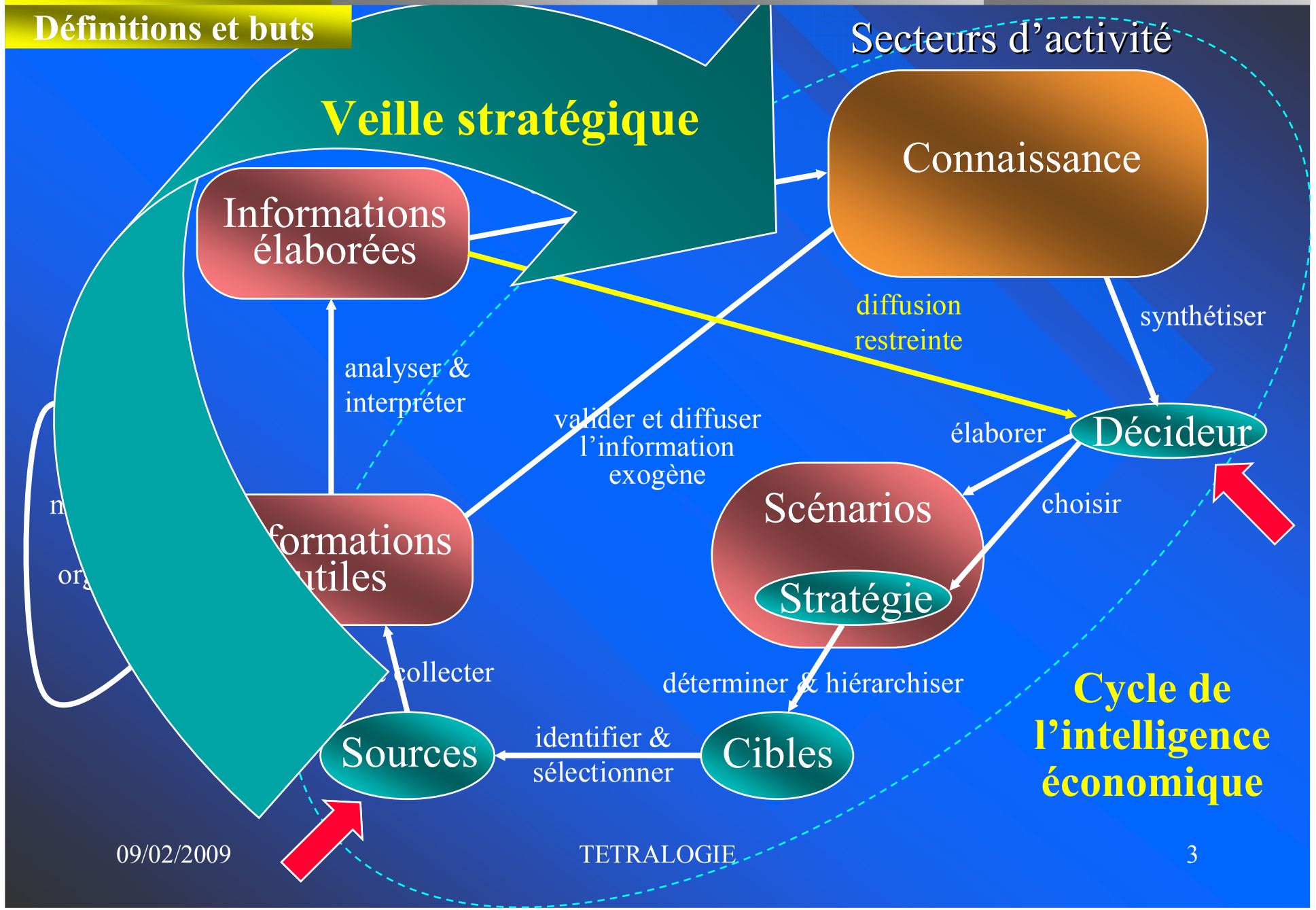
Equipe Systèmes d'Information Généralisés (SIG)

Université Paul Sabatier (Toulouse III)

Plan de la présentation

- **Historique**
  - Définitions et buts
  - La veille stratégique
  - Le multidimensionnel
  - Les classifications
- **Architecture**
  - Interactivité
  - Travail en équipe
  - Principe général
- **Contribution**
  - en analyse textuelle
  - en analyse exploratoire
  - en géostratégie
  - en dessin de graphes
- **Conclusion**
  - bilan
  - perspectives

Définitions et buts



Définitions et buts

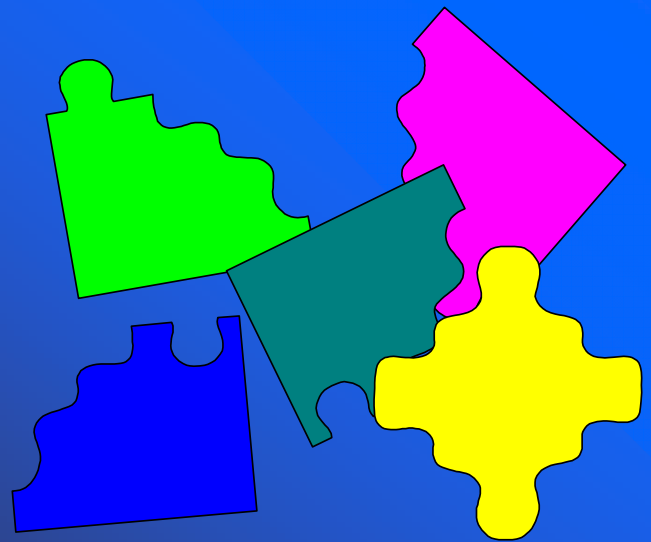
- Etudier l'environnement et l'évolution
  - Des marchés
  - Des coopérations
  - Des alliances
  - Des fusions
  - Des acquisitions
  - Des participations
  - Des implantations
  - De l'innovation
  - Des ruptures
  - Des transferts de technologie
  - Des dépôts de brevets
  - Des équipes de recherche
  - De la terminologie
  - Des sources d'information
  - Des publications
  - De la mode
  - De la publicité
  - Des appels d'offres

Définitions et buts

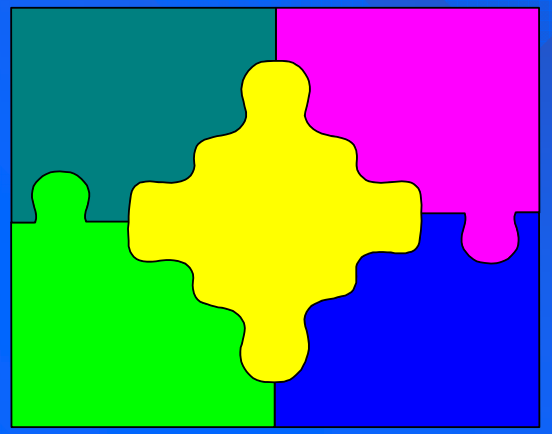
- Depuis toutes les sources d'information électronique
  - Bases bibliographiques
    - Web of science (SCI)
    - Pascal, Francis (CNRS)
    - Chemical abstract , Biosis
    - Current Contents, Inspec,
    - Cnki, Cqvip (Chine), …
  - Serveurs spécialisés
    - Dialog, Stn
    - Questel, Derwent, …
  - Presse
    - Factiva, Journaux électroniques
    - Afp, Reuters, …
  - Internet
    - Brevets : Uspto, Esp@cenet
    - Bio: Pubmed, Biospace
    - Pages web, sites web
    - Blogs, news-groups
    - Flux RSS, …
  - Intranet
    - SI propriétaire/SGBD
    - Data warehouse
    - Indexations
    - Web-logs
    - Mails, Streams, …

Définitions et buts

- Information explicite

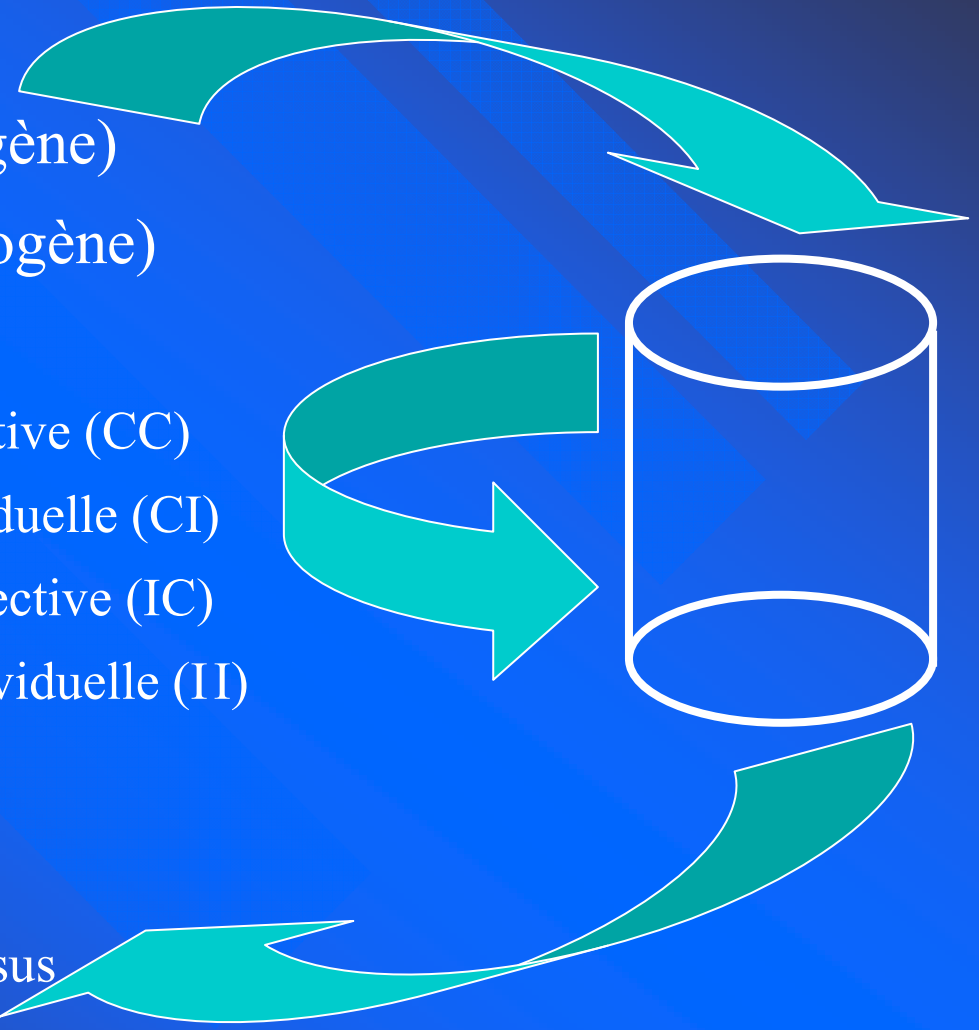


- Information implicite



Définitions et buts

- l'information explicite (exogène)
- l'information implicite (endogène)
- Maturité de l 'information
  - L'information consciente collective (CC)
  - L'information consciente individuelle (CI)
  - L'information inconsciente collective (IC)
  - L'information inconsciente individuelle (II)
- La terminologie émergente
- Les concepts émergents
  - cohérence, simultanéité, consensus



La veille stratégique

- Le processus de veille stratégique comporte 5 phases :
  - la sélection des données utiles :
    - terminologie, dates, acteurs (auteurs, organismes, pays…)
  - la préparation des données :
    - pré-traitements : nettoyages [Zipf49], synonymies [Porter80]
    - transformations : modèle de représentation des données [Salton89]
  - l'analyse des données:
    - classifications, règles d'associations, séquences, ...
  - l'interprétation et l'évaluation :
    - basées en grande partie sur les visualisations
  - l'exploitation et donc la prise de décision

- Les analyses multidimensionnelles
  - Fondements : Spearman & Pearson dès 1904
  - Analyse canonique et ACP : [Hotteling35]
  - Analyse des correspondances : [Hirschfeld35] et [Guttman41]
  - Analyse procustéennes : [Schonemann65]
  - En France
    - Analyse Factorielle des Correspondances : [Benzécri73]
    - Analyse de deux tableaux : [Lafosse85]
    - Compléments en analyse procustéenne : [Fichet87]

Les classifications

- Classifications
  - Taxonomie par CAH [Sokal63]
  - Centres mobiles [Forgy65]
  - Nuées dynamiques [Diday71]
  - Partitionnements
    - Itératifs : Minimisation de la coupure [KL70], [FM82]
    - Spectraux : VP de la matrice de Laplace [Hall70], [Hagen91]
    - Multi-niveaux : regroupement + itératif KMETIS [Karypis98]
    - Stochastiques : Markov CLustering [Van Dongen00]

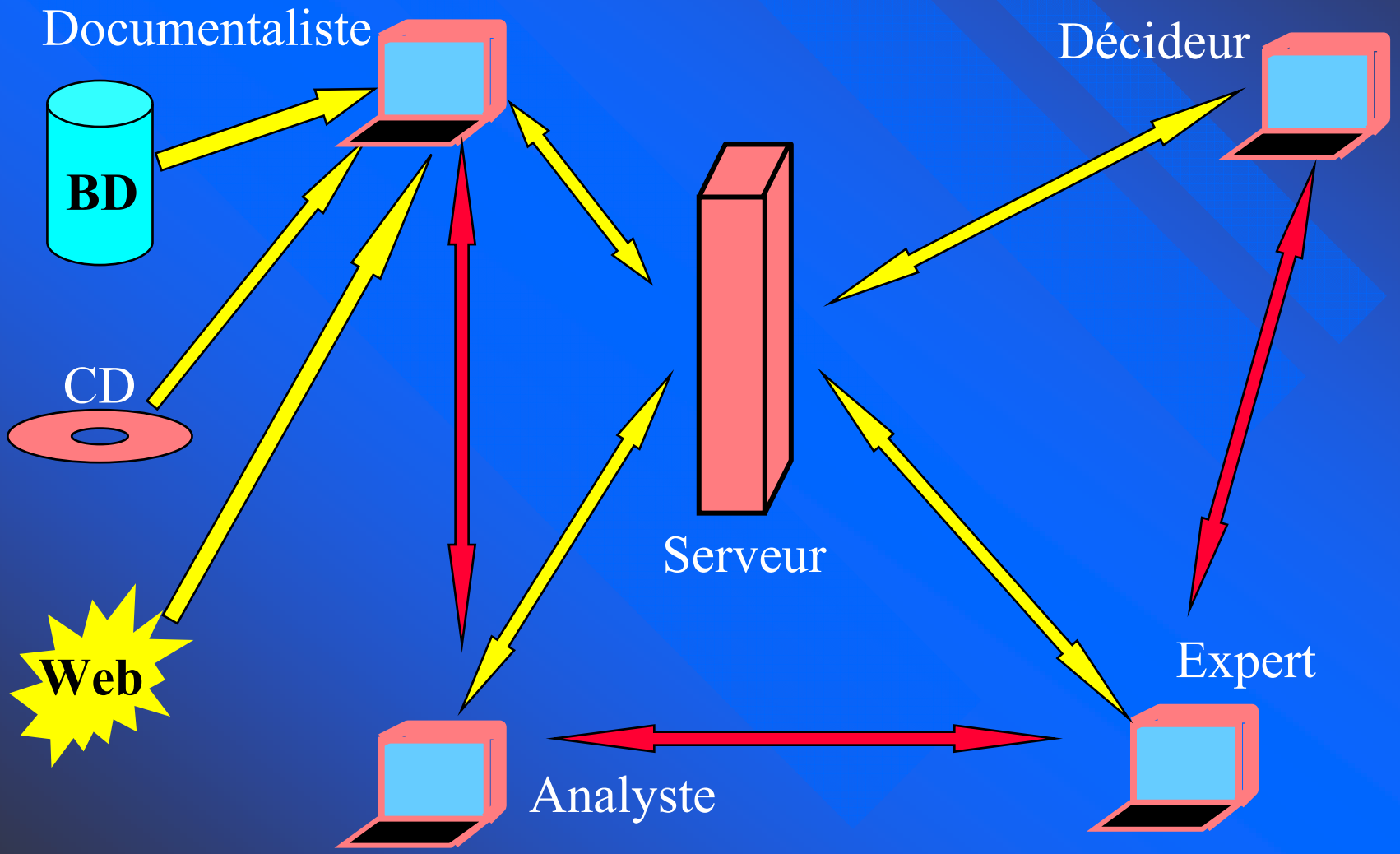
Pourquoi un système interactif?

- En Amont de l'analyse l'interactivité sert à :
  - choisir les sources d'information
  - mettre au point les équations de recherche
  - évaluer les corpus (volume, pertinence, bruit)
  - choisir et valider le format optimal de sortie
  - superviser la collecte du corpus
  - contrôler les dictionnaires (forme et volume)
  - choisir les filtres (+ et -, thématiques, granularité, ...)
  - valider les synonymies et l'indexation (Multi-termes)

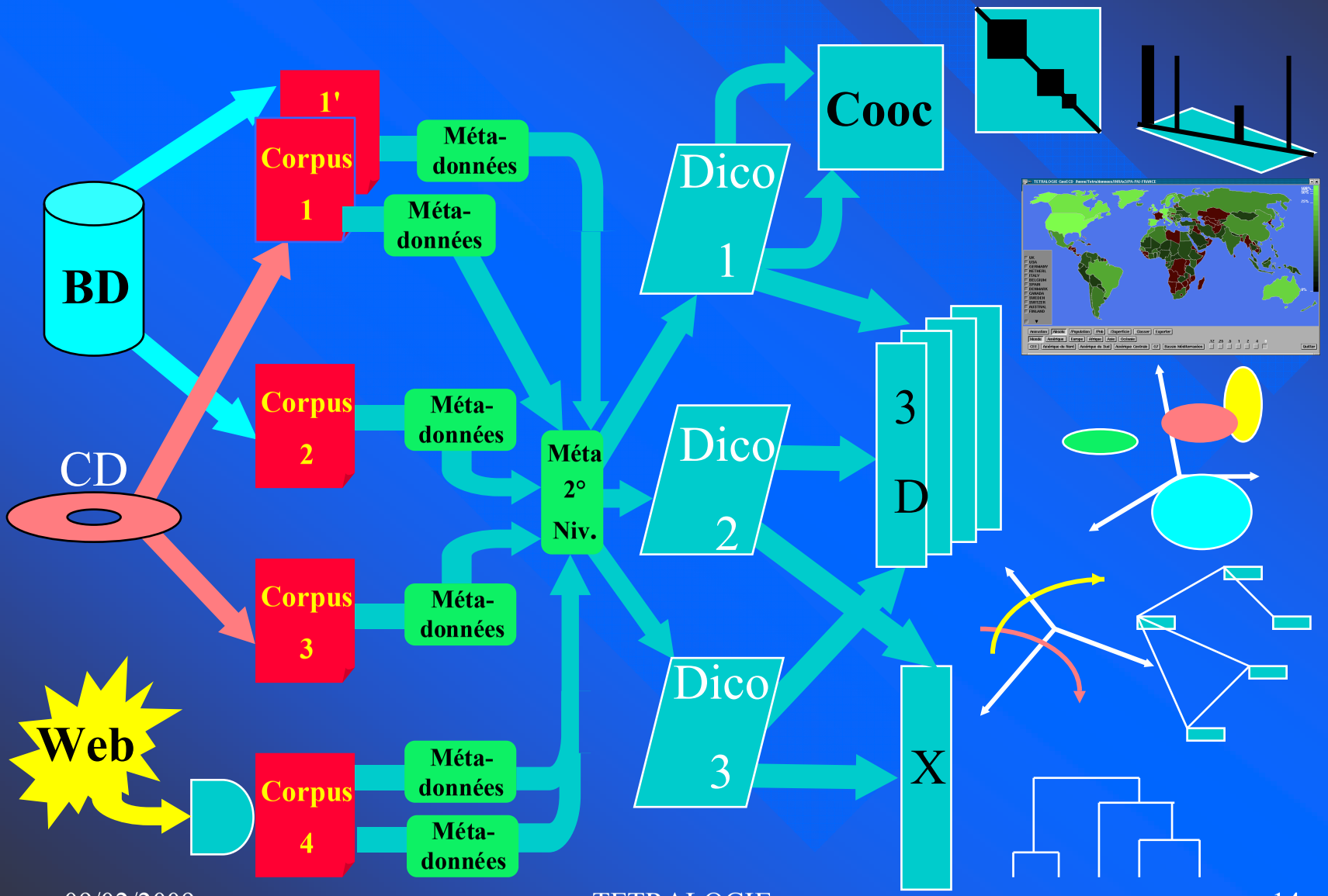
Pourquoi un système interactif?

- Pendant l'analyse elle permet de
  - sélectionner et contrôler les matrices de croisement
  - choisir les méthodes à déployer
  - piloter ces méthodes (localement ou à distance)
  - extraire les connaissances via les visualisations
  - choisir les vues à conserver
  - consulter les experts via le réseau
  - restituer l'analyse au décideur

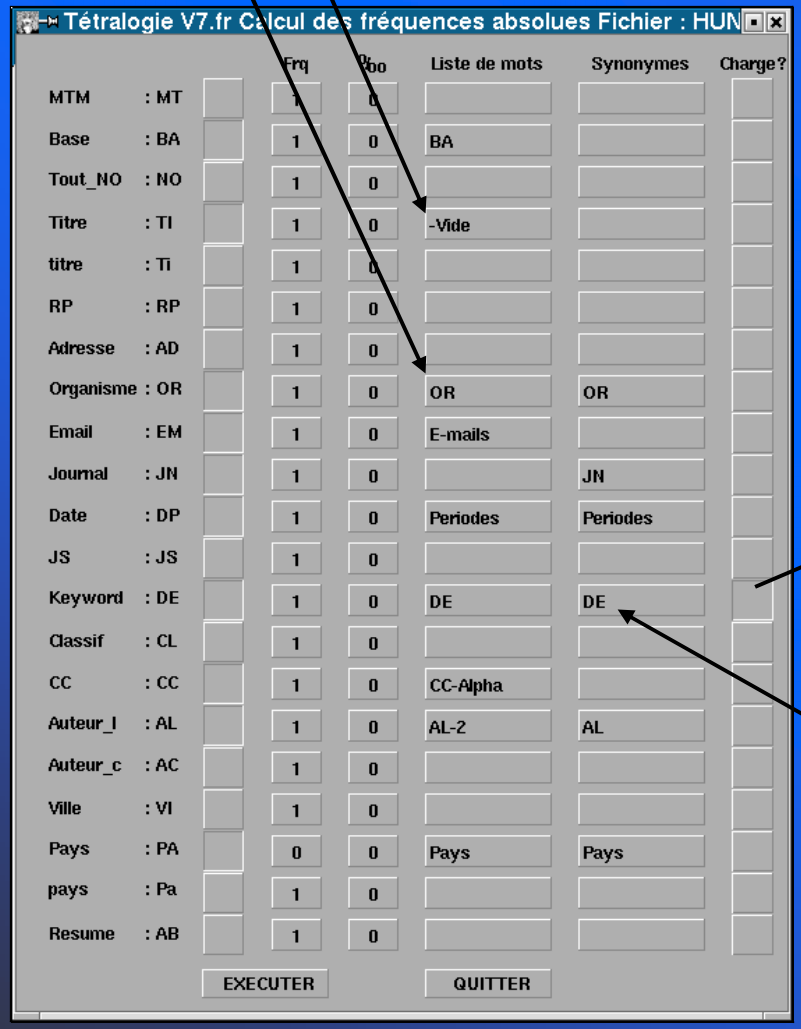
- Connexions entre acteurs de la veille



Principe général



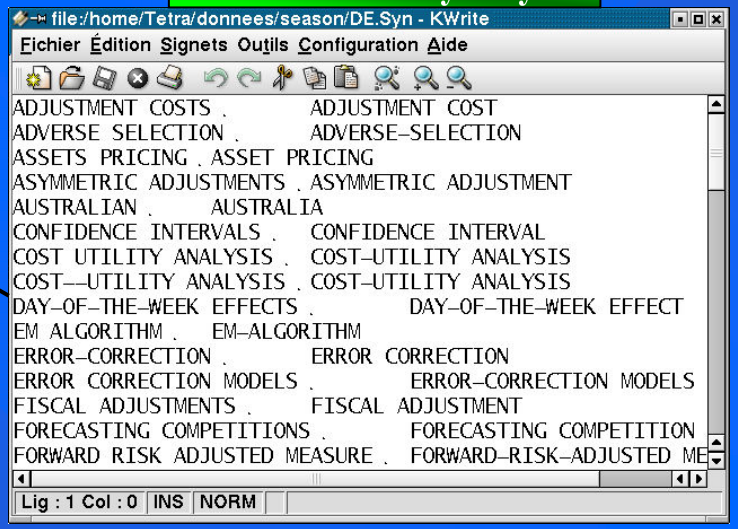
Dictionnaire + ou -



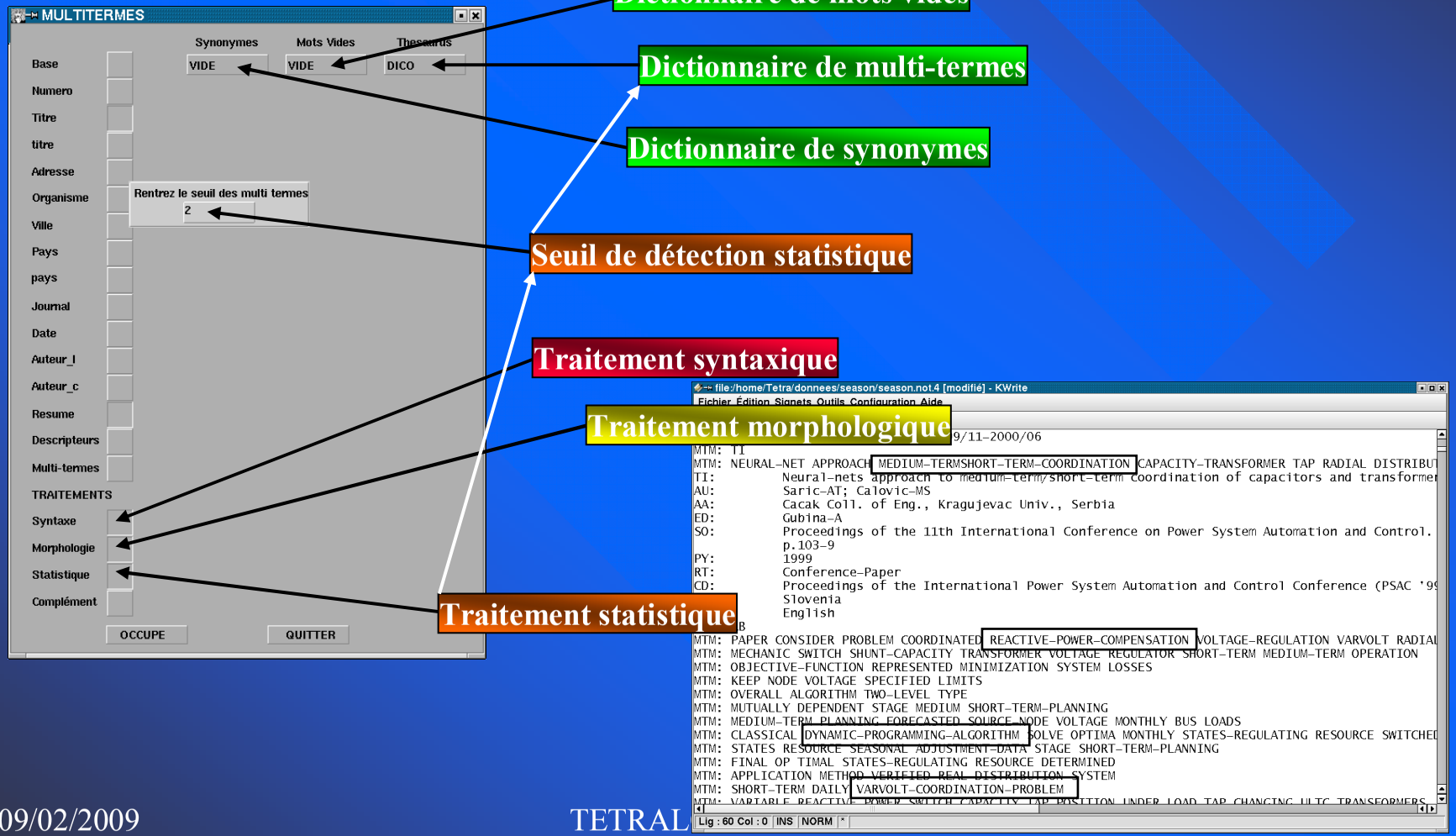
- Extraction des dictionnaires

Courbe de charge

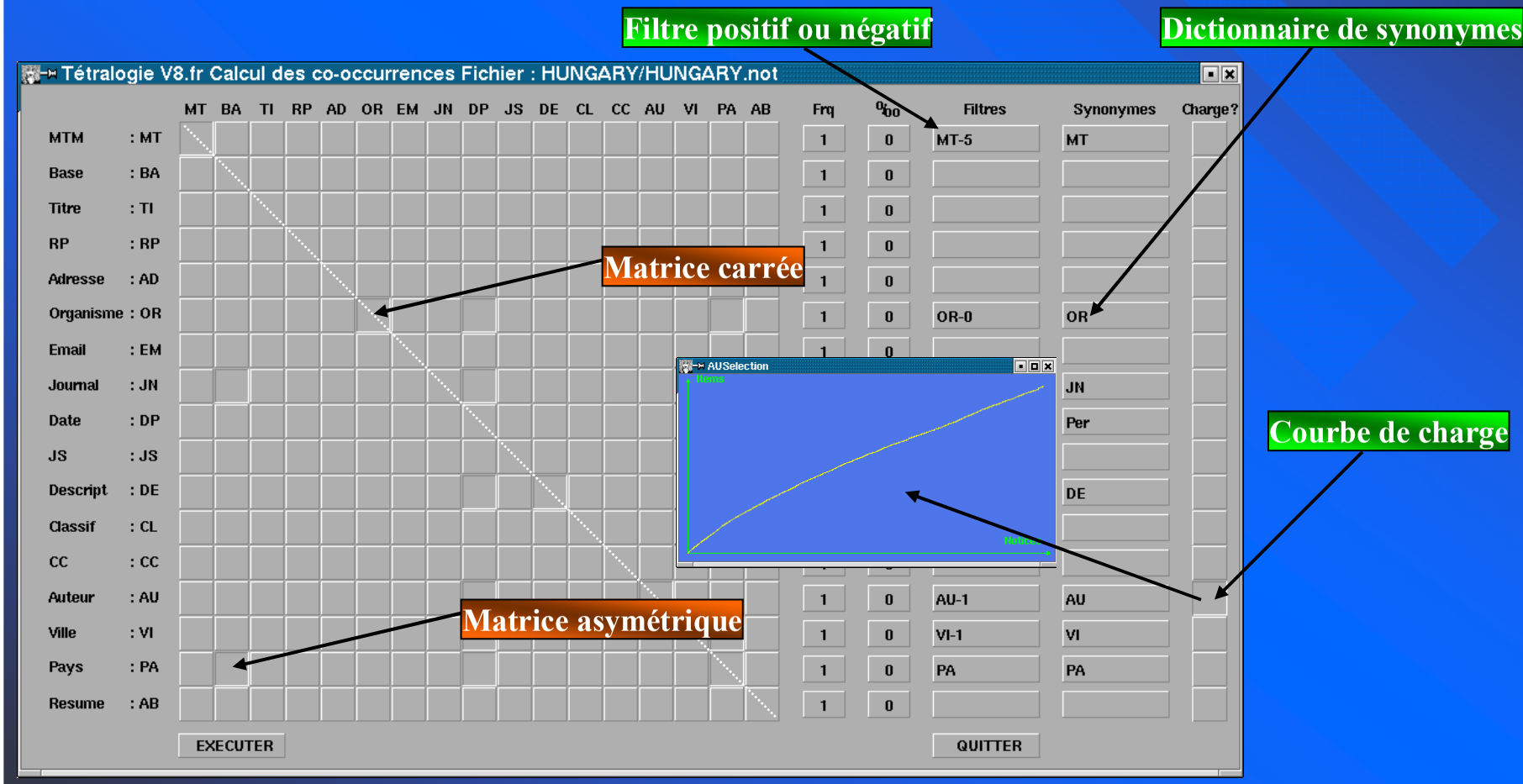
Dictionnaire de synonymes



Détection des multi-termes



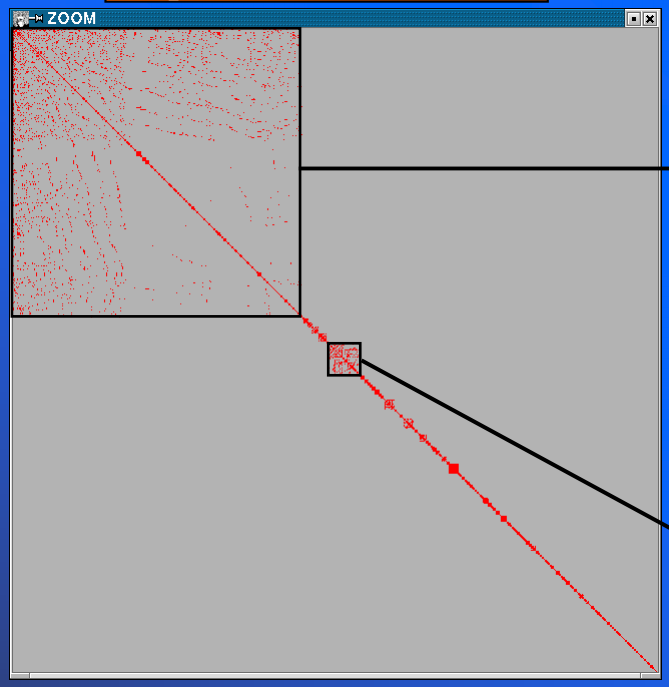
- Croisements 2D



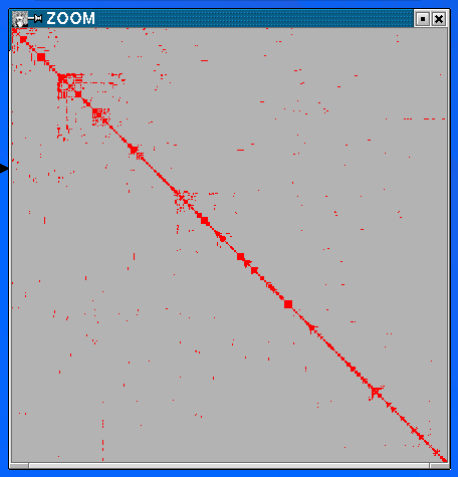
- pour le traitement des matrices
  - proposition de plusieurs algorithmes de tris
  - génération supervisée de matrices de croisement
  - tableur 3D adapté aux matrices de grande taille (zooms 2 et 3D)
- pour les analyses multidimensionnelles
  - visualisations interactives en 3D et 4 D
  - synchronisation de cartes locales ou distantes
  - visualisation de trajectoires et rotations procustéennes (AFCM)
- pour les classifications
  - arbres hiérarchiques interactifs avec exportation des classes
  - partitionnement de graphes, graphes de classes
  - segmentation pour les cartes géographiques

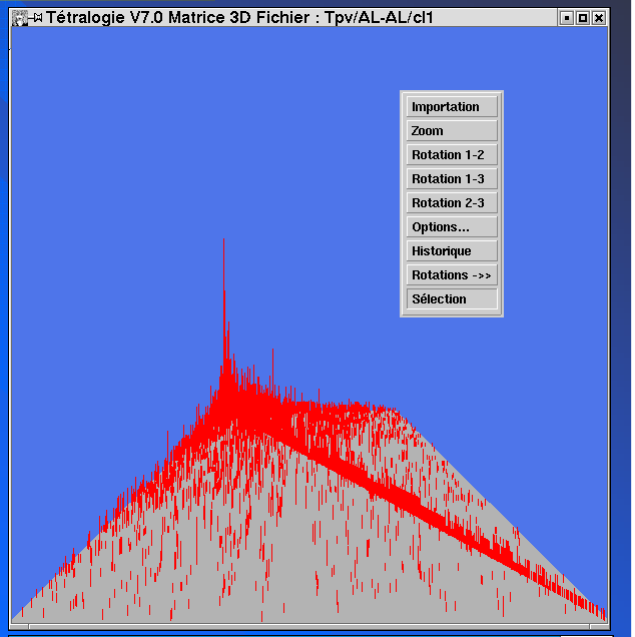
Algorithmes de tris de matrices

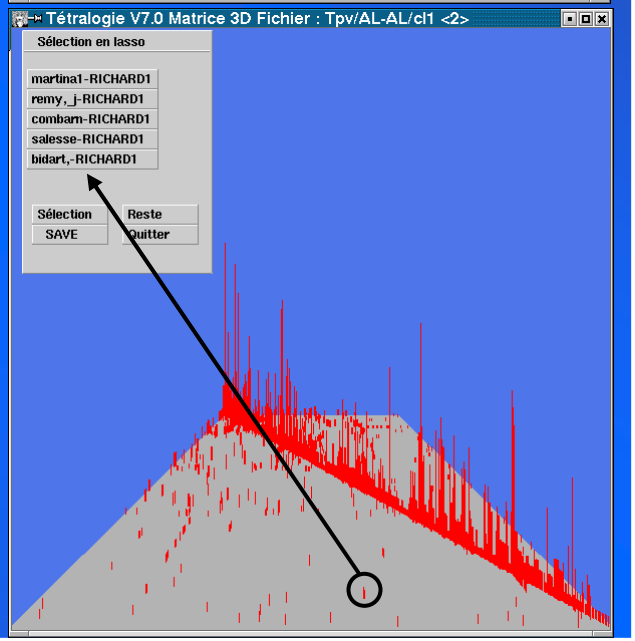
Tri par classes de connexité



Tri par blocs diagonaux







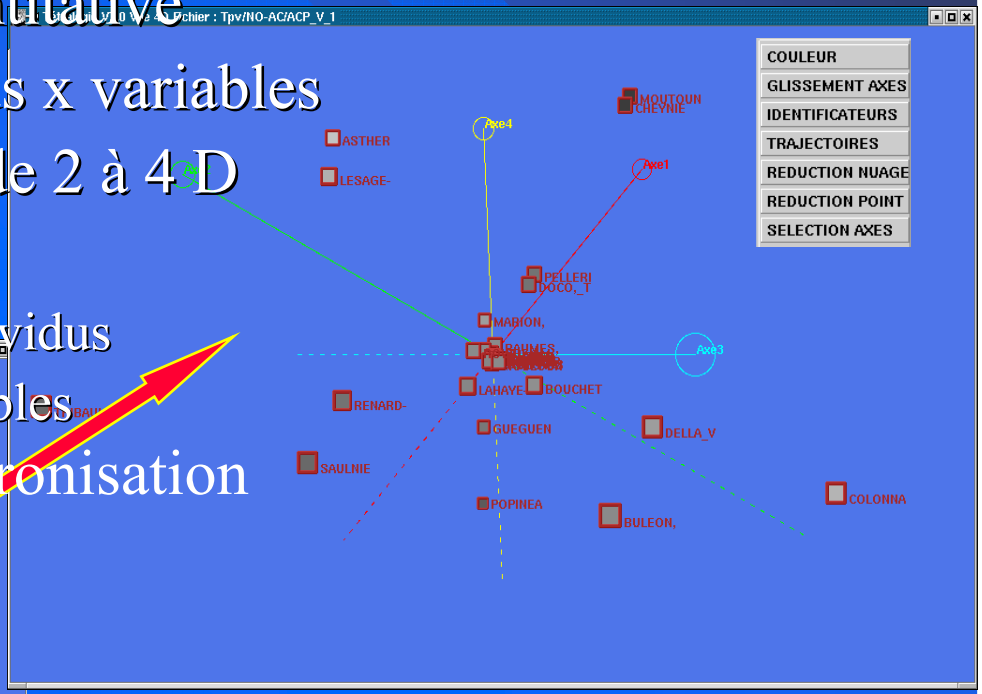
en analyse exploratoire

- Analyse en composantes principales (ACP)
  - s'applique aux données quantitative comme les tableaux individus x variables
  - elle utilise un espace réduit de 2 à 4 D
  - deux cartes en sont déduites

Cartes des coordonnées en 4D

- celle des coordonnées des individus
- cercle de corrélation des variables

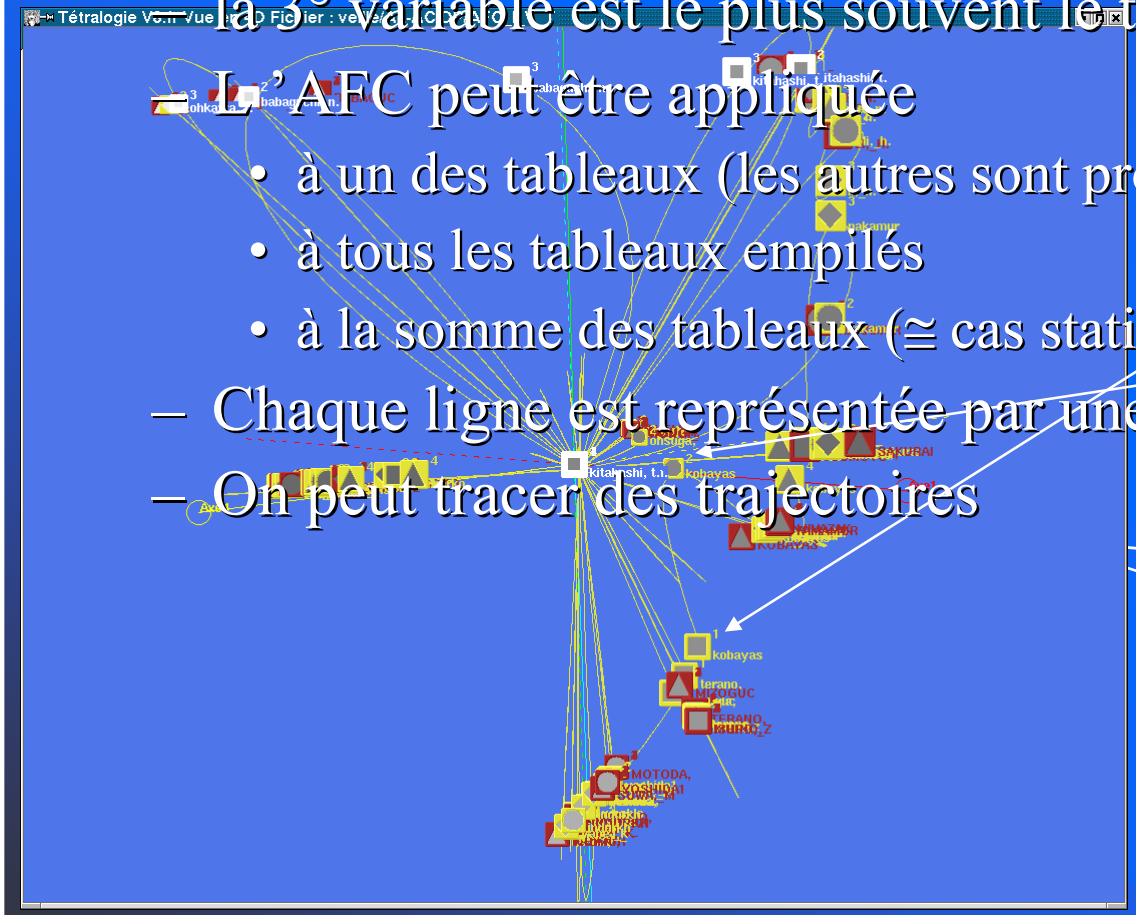
Synchronisation

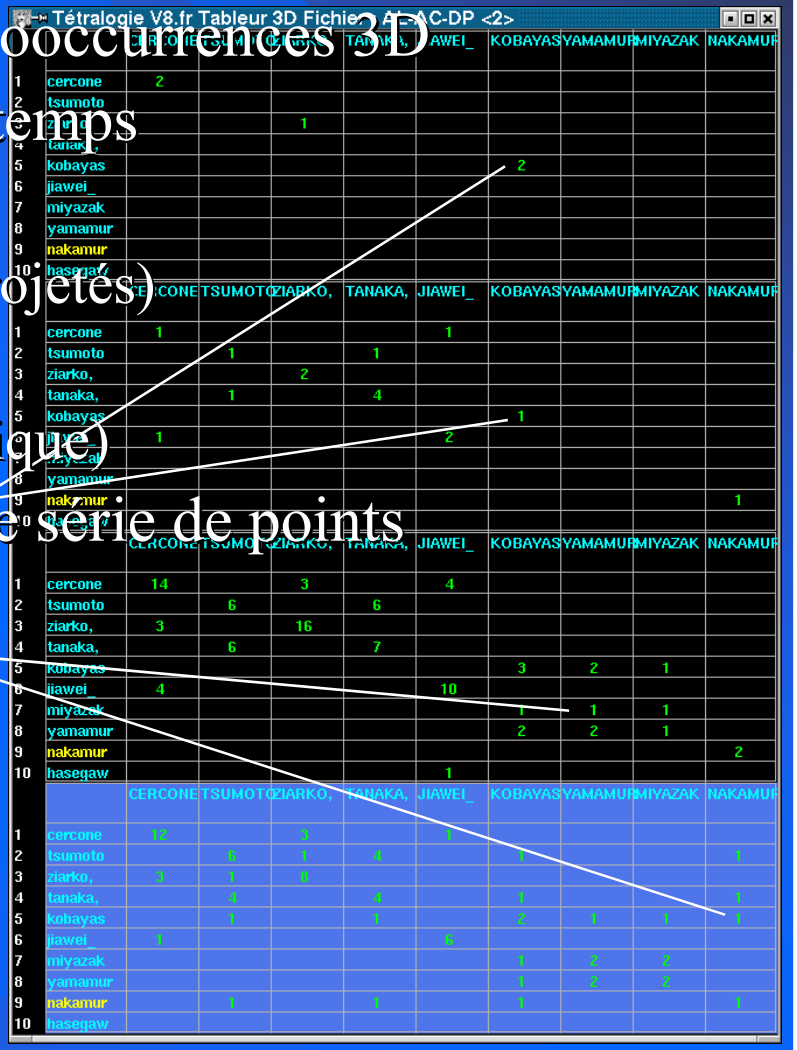


Cercle des corrélations

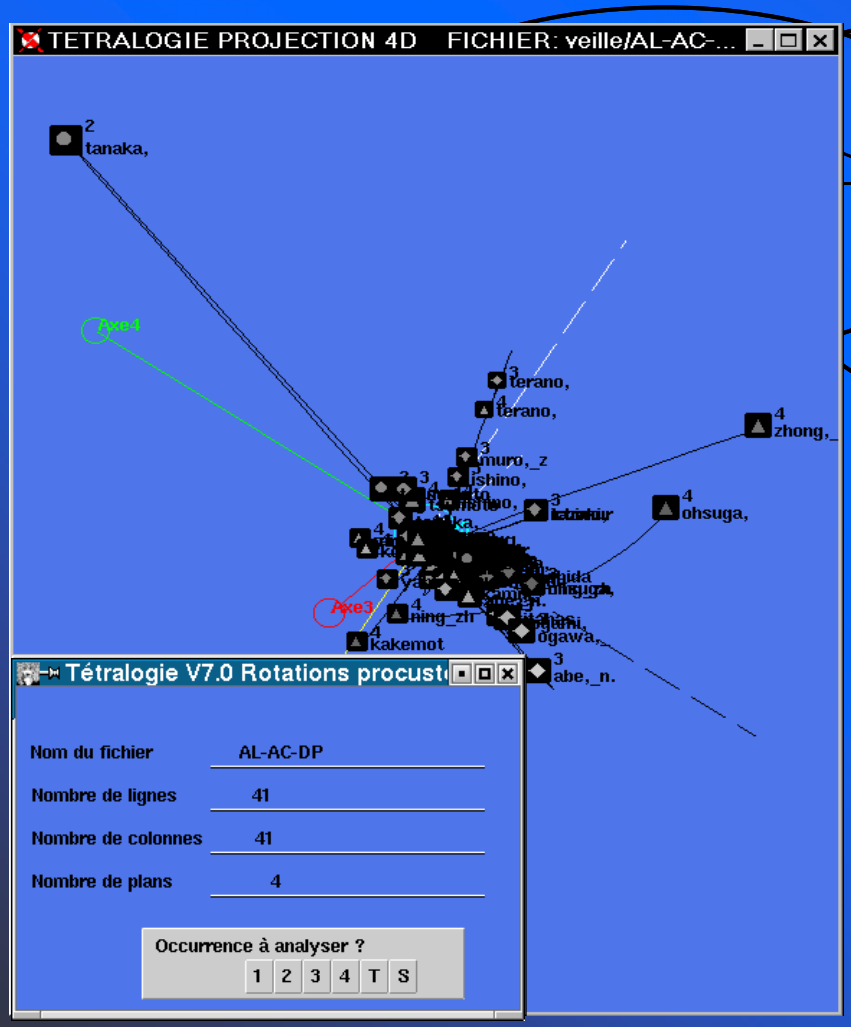
en analyse exploratoire

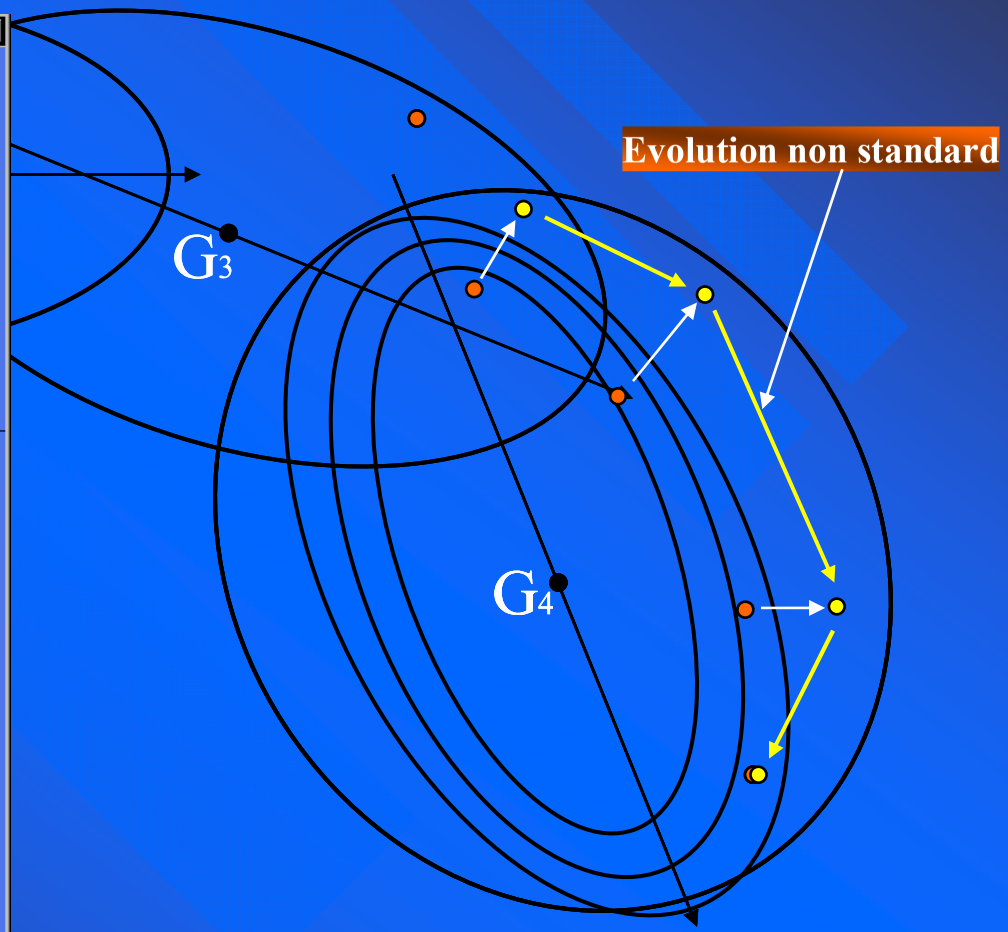
- Analyse factorielle des correspondances multiple (AFCM)
  - nous l'appliquons aux matrices de cooccurrences 3D
  - la 3° variable est le plus souvent le temps
  - L'AFC peut être appliquée
    - à un des tableaux (les autres sont projetés)
    - à tous les tableaux empilés
    - à la somme des tableaux (≅ cas statique)
  - Chaque ligne est représentée par une série de points
  - On peut tracer des trajectoires





- Analyse procustéenne



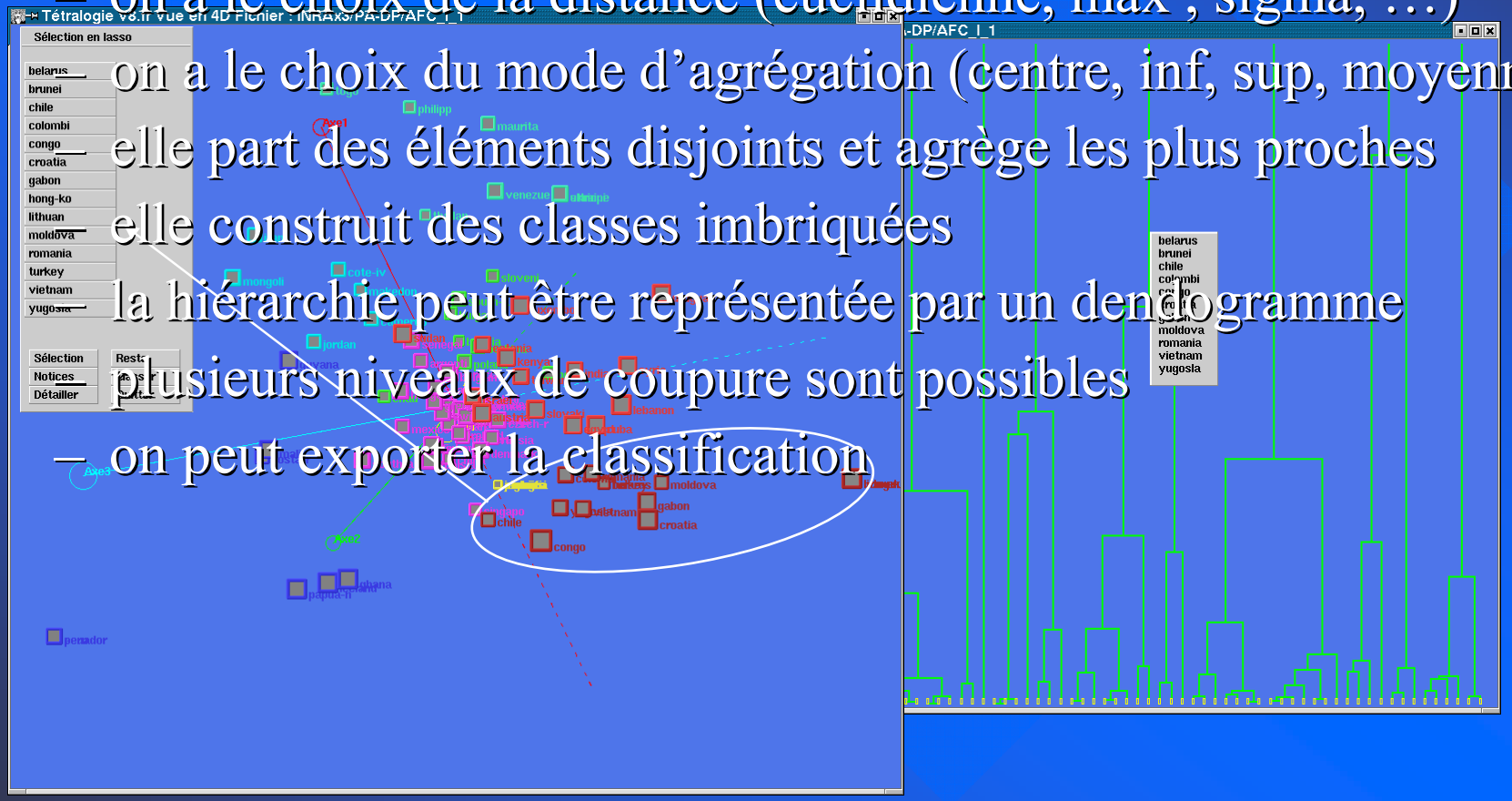


en analyse exploratoire

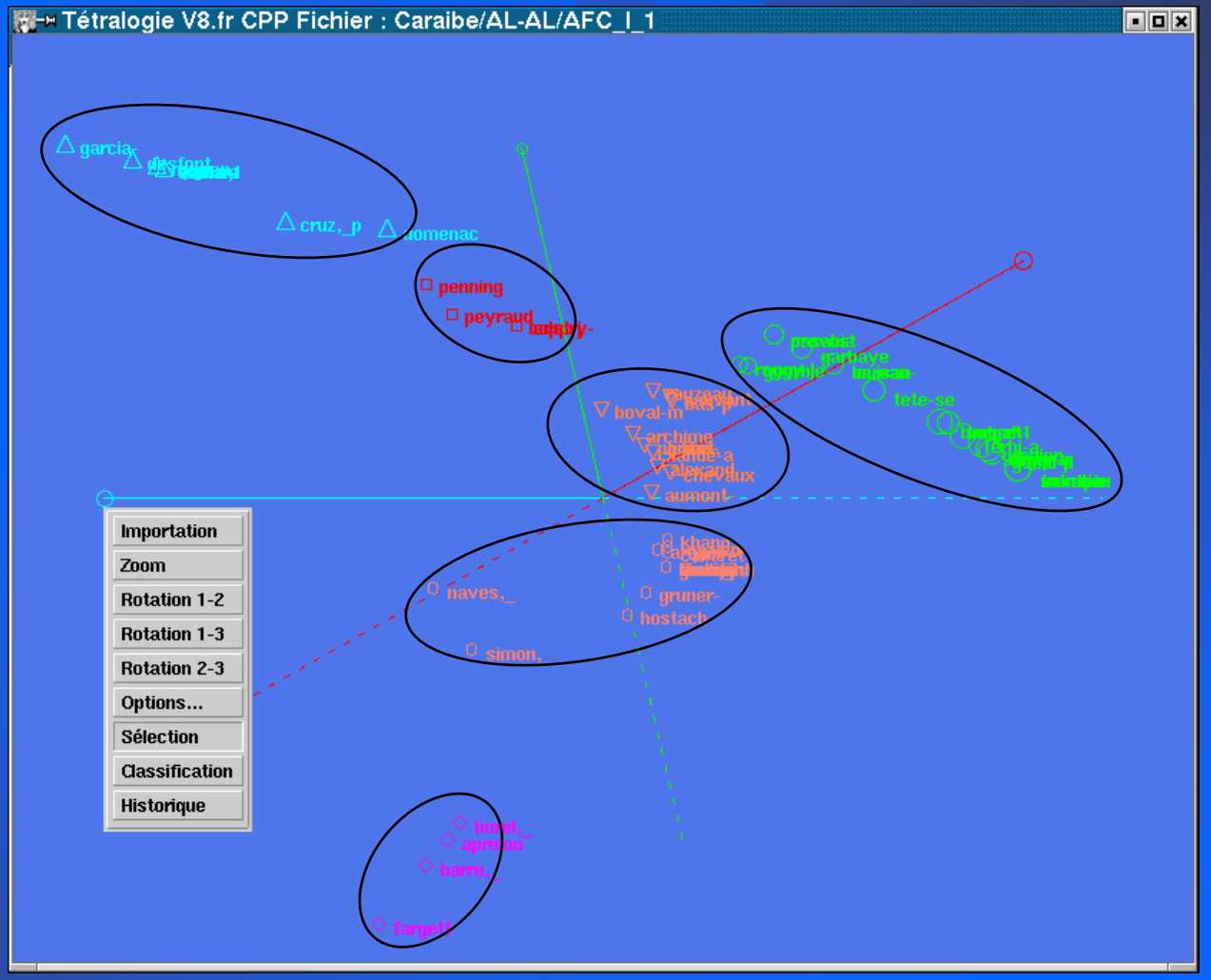
- Classification ascendante hiérarchique (CAH)
  - méthode non supervisée
  - on a le choix de la distance (euclidienne, max , sigma, …)

$$d_e(A,B)=\sqrt{\sum_{i=1,n}(x_i-y_i)^2} \qquad d_m(A,B)=\min_{k,l}\{d_e(x_k,y_l)\} \qquad d_M(A,B)=\max_{k,l}\{d_e(x_k,y_l)\} \qquad d_\mu(A,B)=\frac{\sum_k\sum_l d_e(x_k,y_l)}{|A|\,|B|}$$

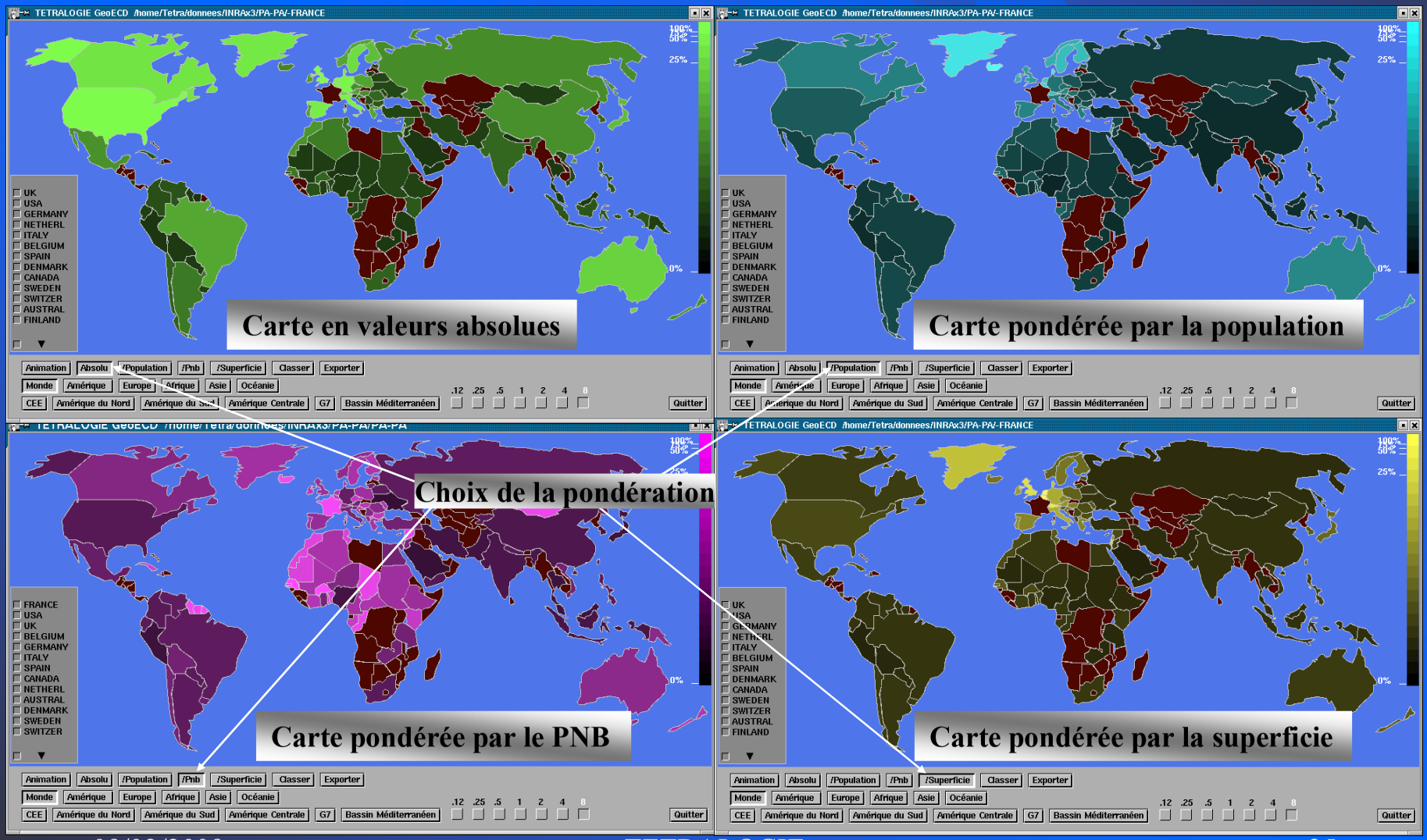
  - on a le choix du mode d'agrégation (centre, inf, sup, moyenne)
  - elle part des éléments disjoints et agrège les plus proches
  - elle construit des classes imbriquées
  - la hiérarchie peut être représentée par un dendrogramme
  - plusieurs niveaux de coupure sont possibles
  - on peut exporter la classification



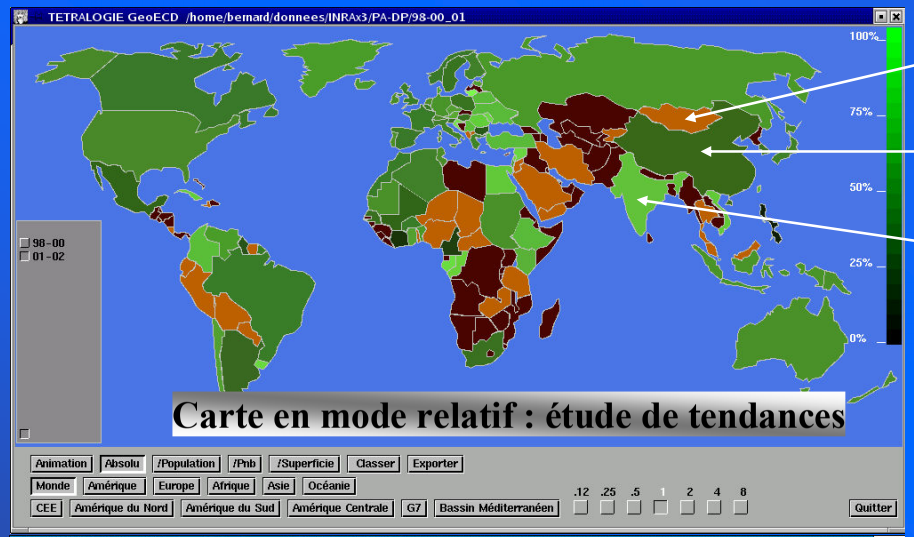
- Classification par partition CPP (Centres mobiles)
  - supervisée
  - k classes
  - 1 représentant
  - choix en 3D
  - coloration
  - icônes
  - 4 classes
  - 6 classes



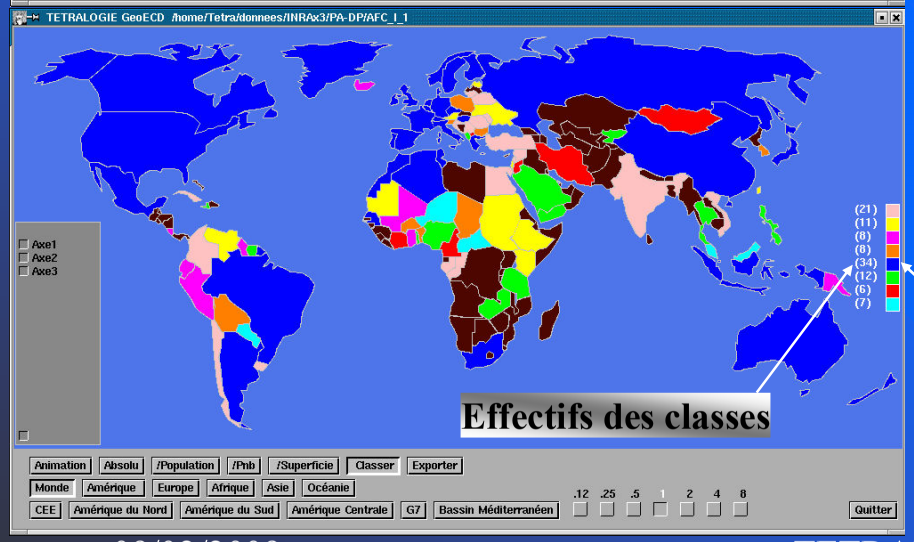
Pondération par des données externes



Segmentation et importation de classes



Carte en mode relatif : étude de tendances

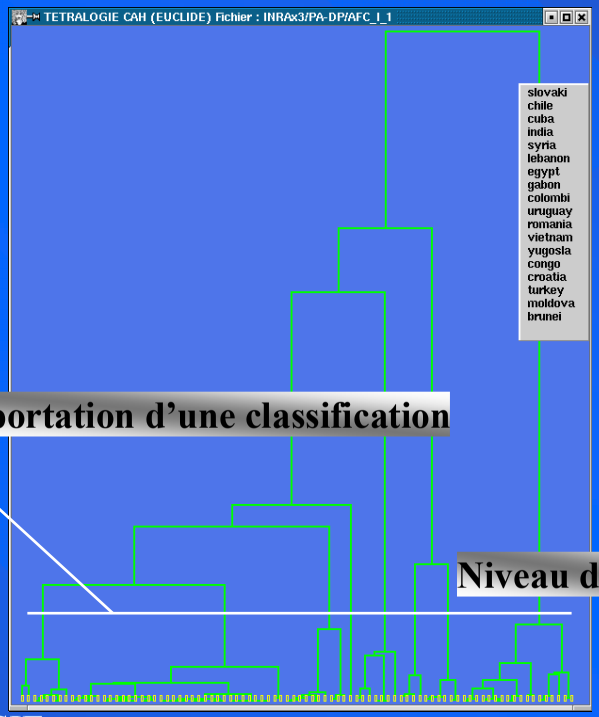


Effectifs des classes

Pays non présents dans la dernière période

Pays en récession

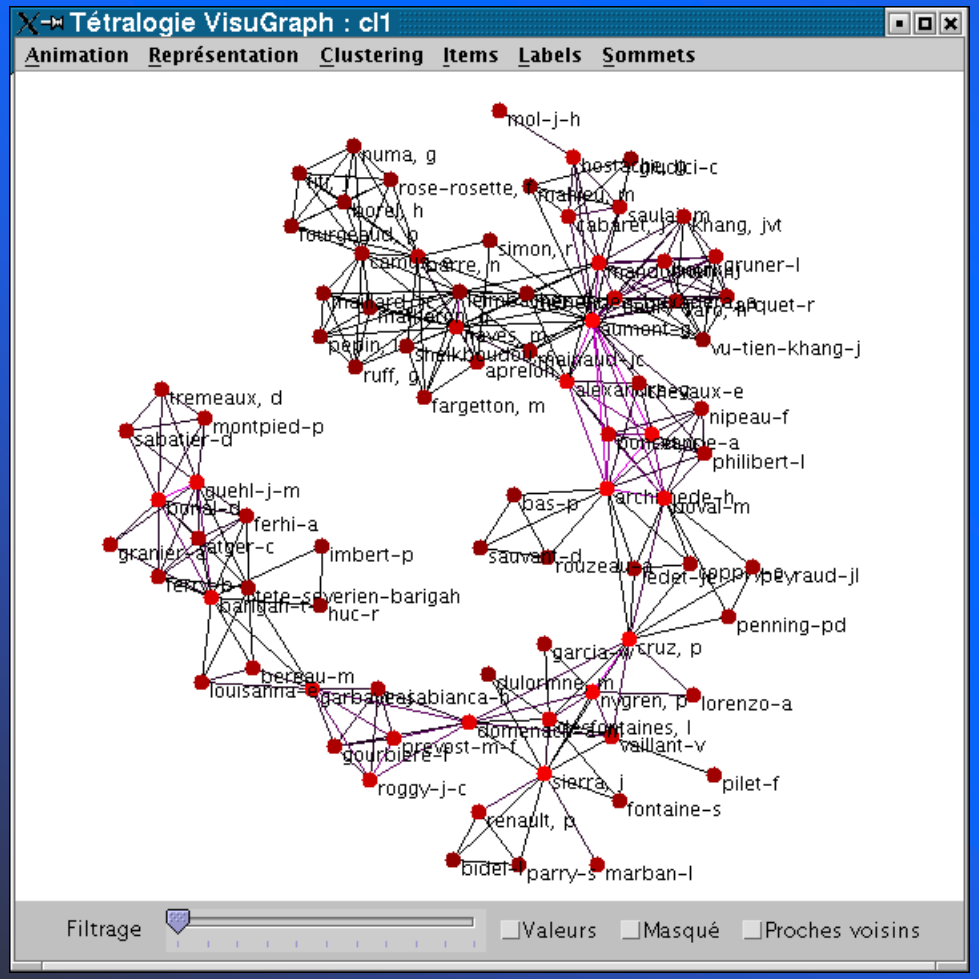
Pays en forte croissance



Exportation d'une classification

Niveau de coupure

Convergence rapide du placement des sommets



≈Algorithme de [Fruchterman91]

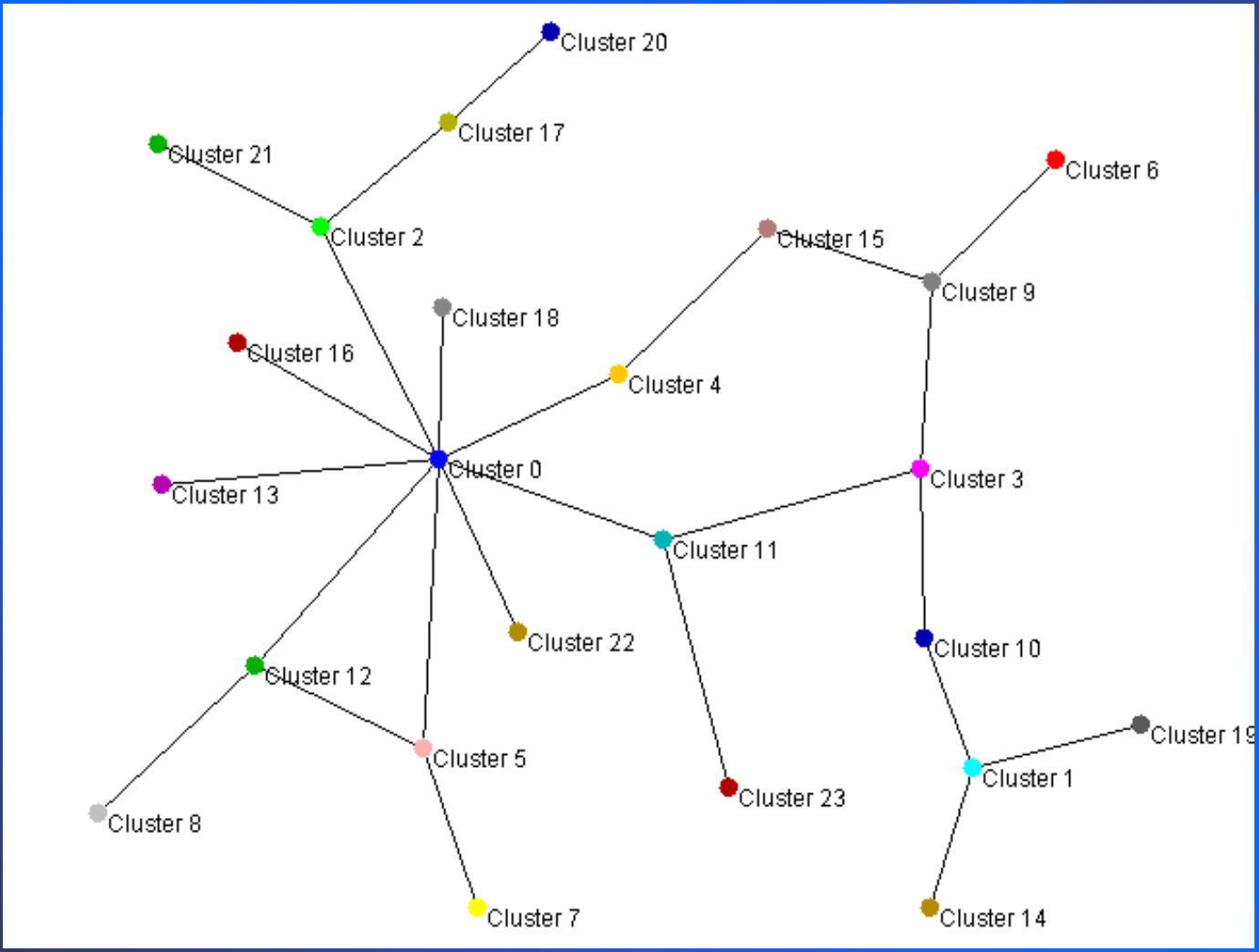
Attraction :

$$f_a(v_i, v_j) = \frac{a_{ij} \times d_{ij}^{\alpha_a}}{k}$$

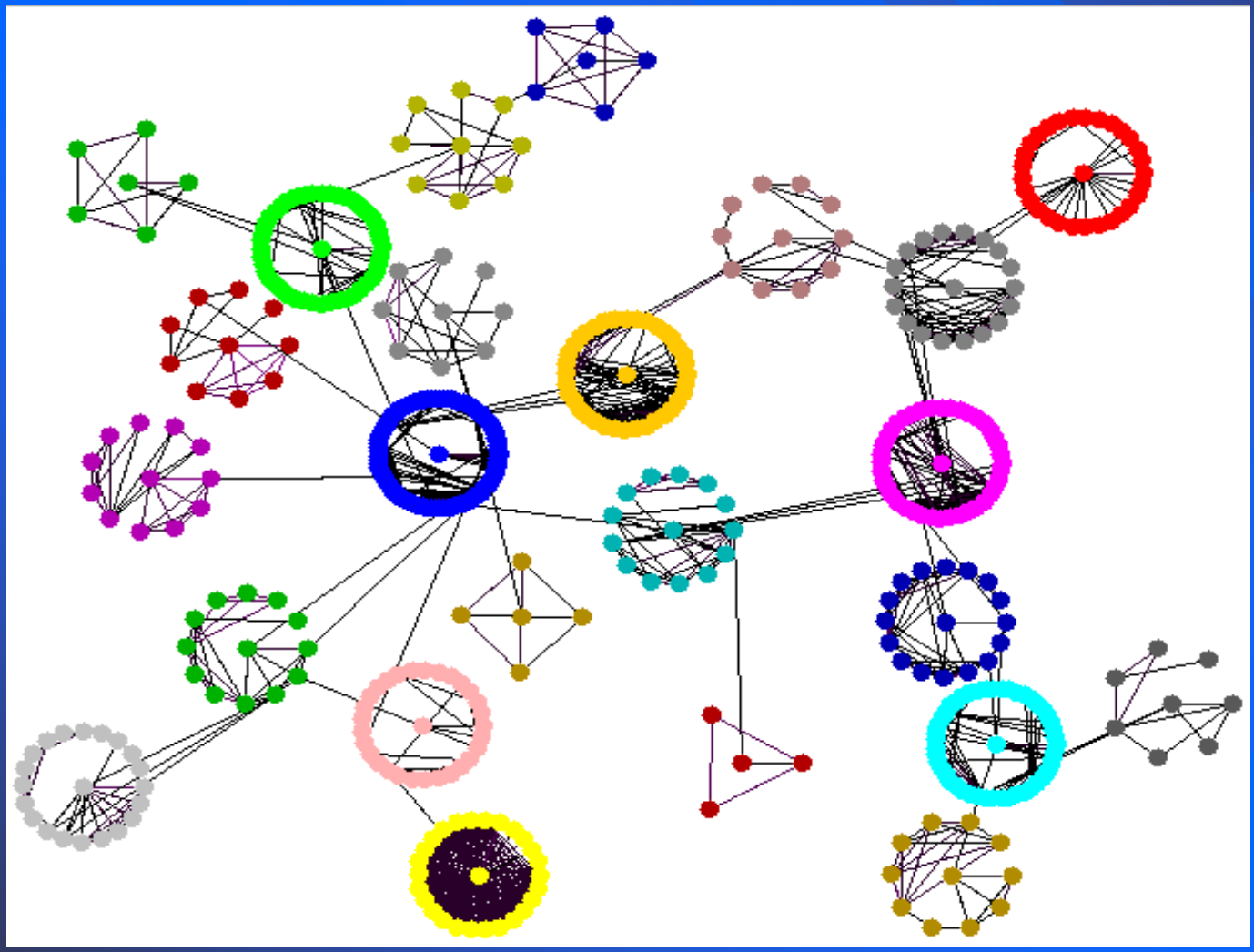
Répulsion :

$$f_r(v_i, v_j) = -\frac{k^2}{d_{ij}^{\alpha_r}}$$

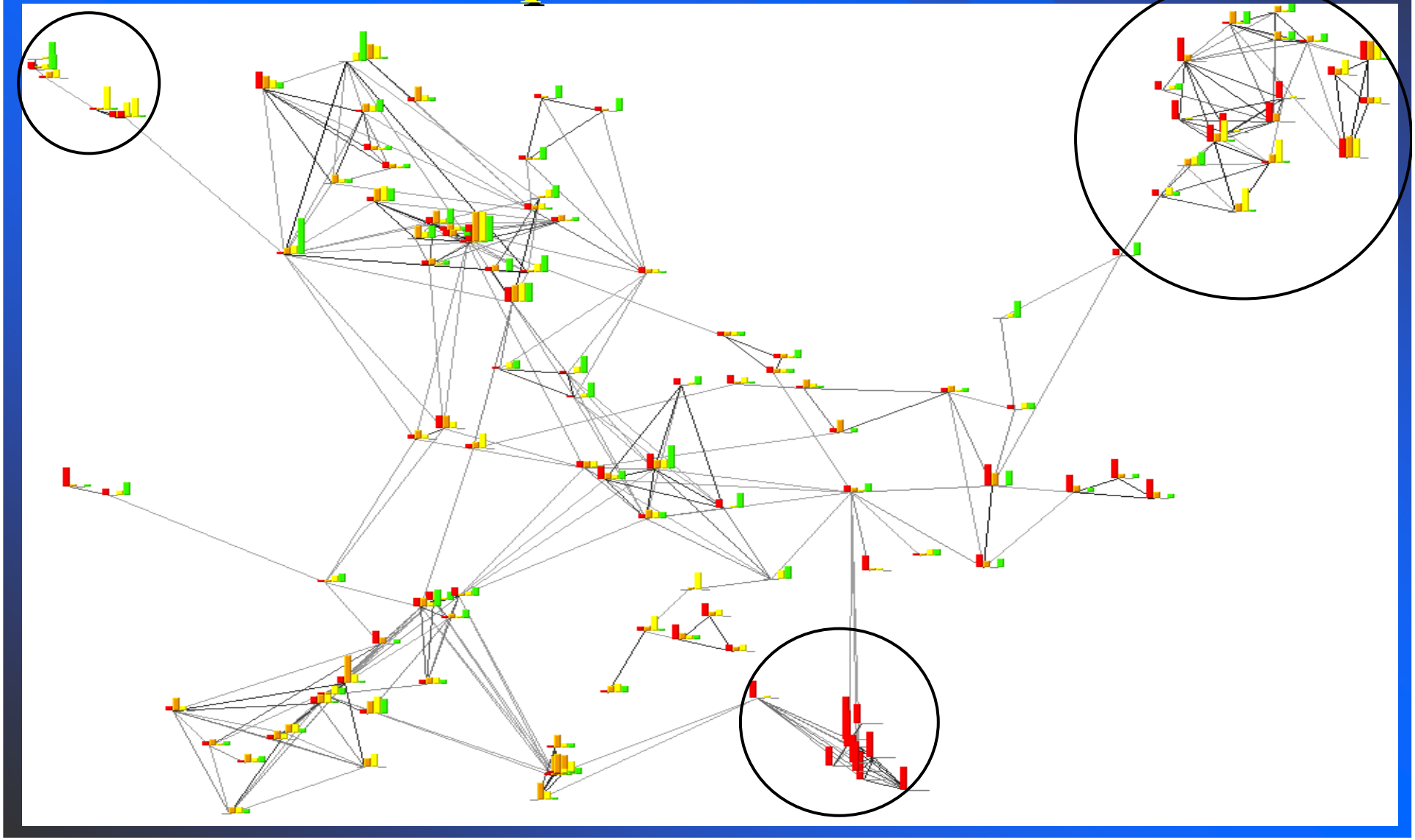
Graphe de clusters



Graphe partitionné



Graphe évolutif



- Philosophie de notre plate-forme
  - c'est un ensemble cohérent de prototypes inter-opérants,
  - qui utilise un standard unique pour le format des données,
  - son interface graphique est homogène,
  - elle permet de partager ressources et méthodes via le réseau.
- Son utilisation en recherche
  - support d'évaluation de méthodes, d'outils et de produits d'IE,
  - vaste base d'exemples, à l'échelle, déjà analysés.
- Domaine d'application : la veille stratégique
  - veille scientifique (scientométrie, indicateurs, évaluation)
  - veille technologique (propriété industrielle, produits, procédés)
  - veille économique (marchés, concurrents, substituts, …)

Dans le cadre de l'Intelligence économique,

cette approche doit permettre

- le recueil, le suivi, le traitement, la diffusion ciblée et la protection de l'information stratégique
- l'accès systématique à l'information implicite
- l'émulation des compétences via les réseaux
- une culture collective interne (partage et émulation)
- une culture collective externe (logique de réseau d'acteurs)
- éventuellement , une culture proactive (influence, lobbying, contre-information/rumeurs, extraction des signaux faibles, décryptage des réseaux,…)