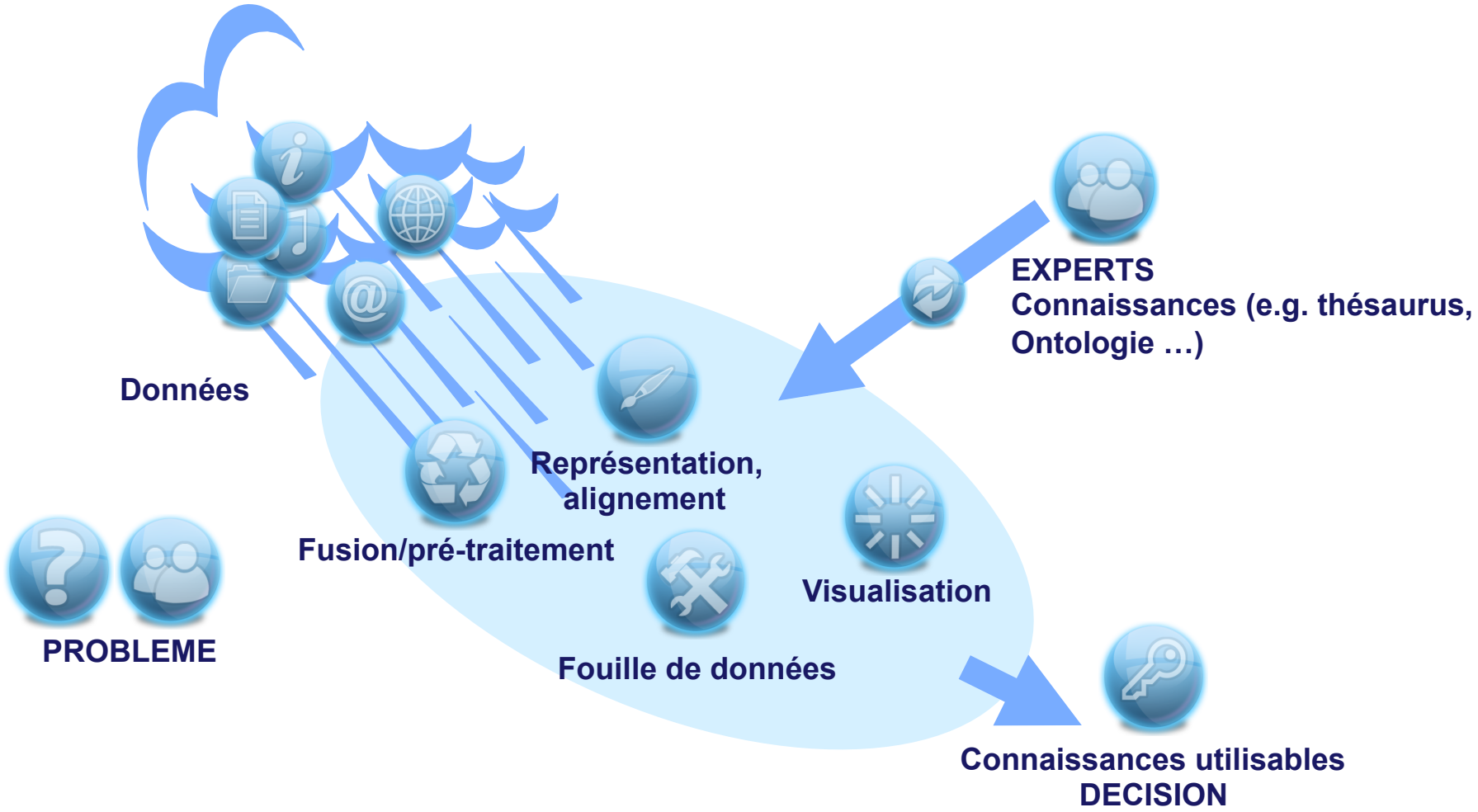
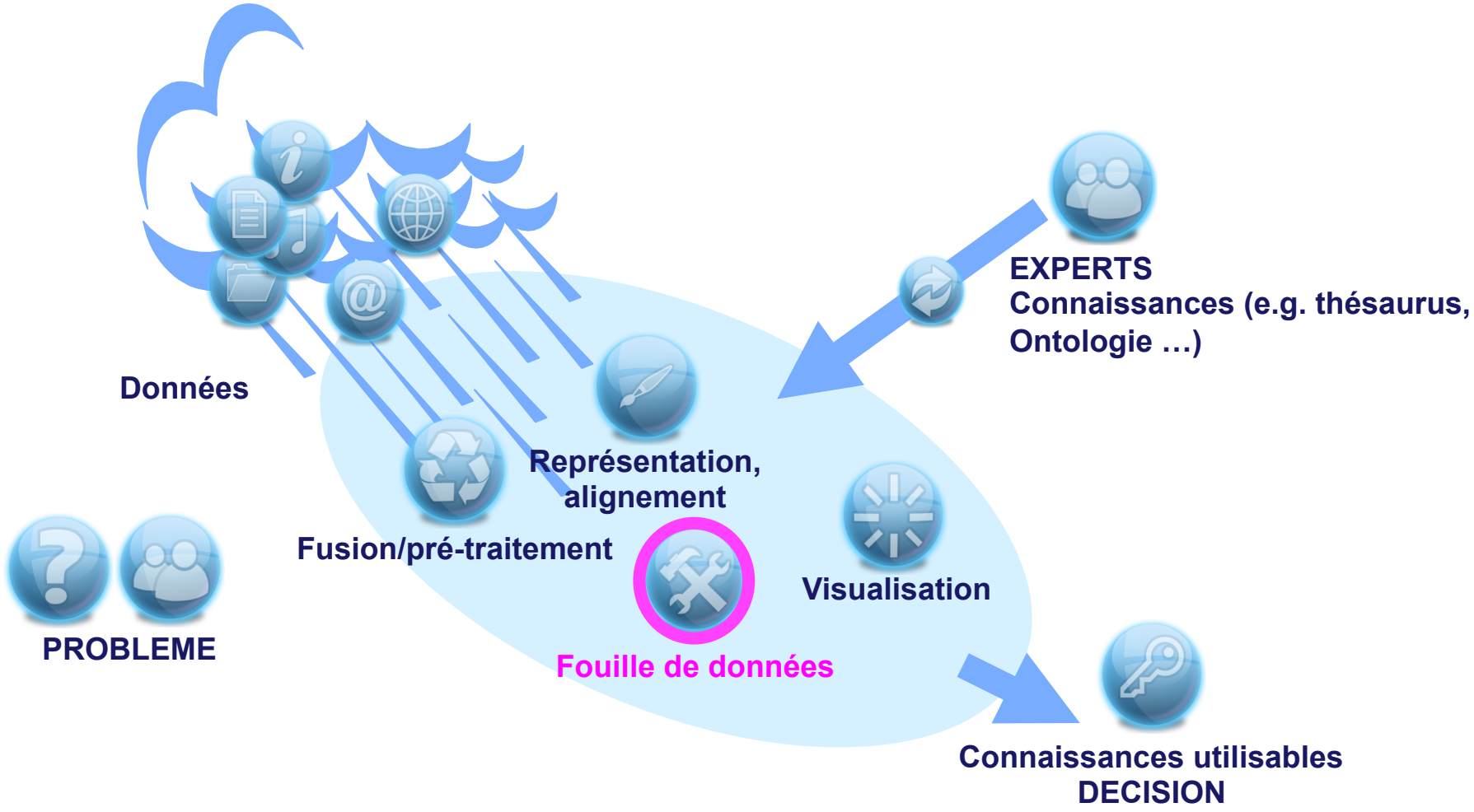
Fouille de données et Santé

sandra.bringay@univ-montp3.fr

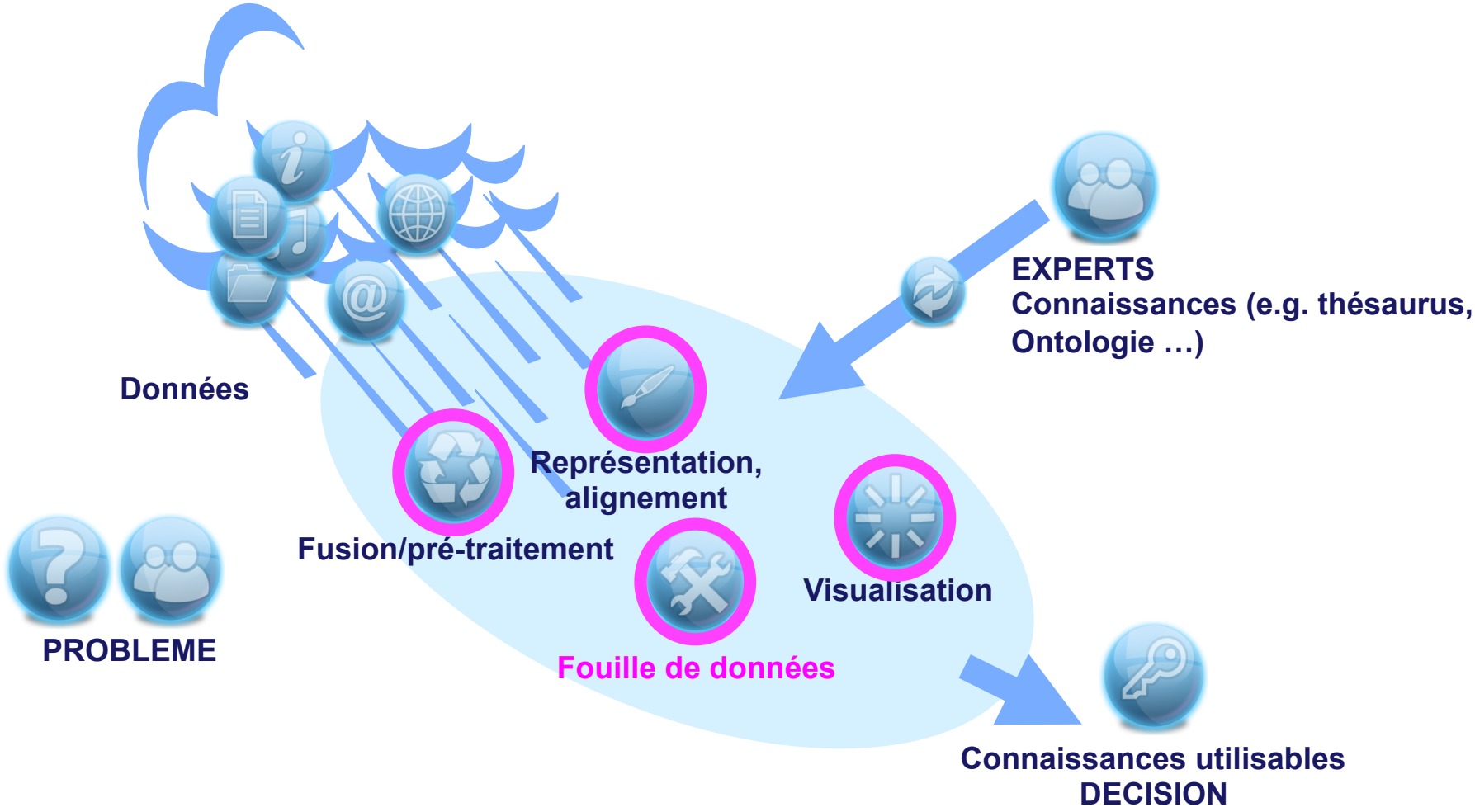
Processus d'extraction de connaissances



Processus d'extraction de connaissances



Processus d'extraction de connaissances



Un processus général

Des données complexes

Extraction de motifs

Transférer des connaissances aux experts

- Analyse
- Classification
- Prédiction

Des données complexes

Des données complexes

- Booléène
- Numerique
- Symbolique
- Multi-dimensionnelle
- Texte
- Image

Extraction de motifs

Transfert de connaissances aux experts

- Analyse
- Classification
- Prédiction

- De gros volumes
- Données bruitées
- Données manquantes
- Données dynamiques
- Flux de données

Différents types de motifs

Des données complexes

Extraction de motifs

- Corrélations
- Règles d'association
- Motifs séquentiels
- Motifs multi-dimensionnels
- Motifs contextuels
- Motifs spatio-temporels

Transfert de connaissances aux experts

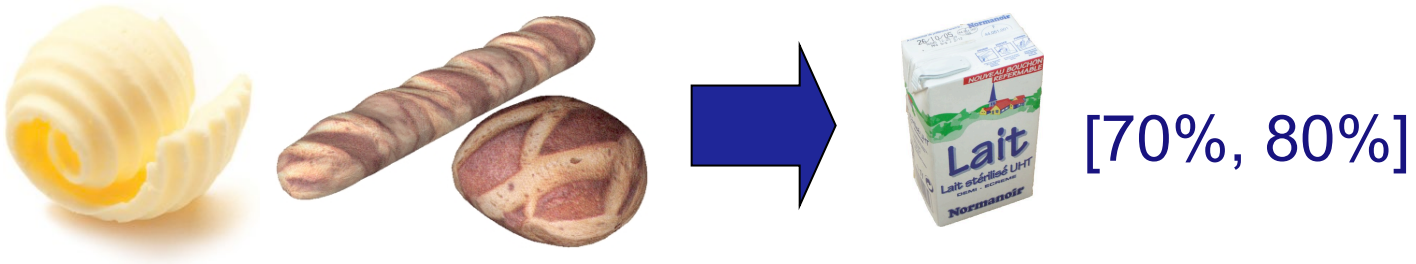
- Analyse
- Classification
- Prédiction

- Motifs inattendus, discriminants…
- Règles graduelles
- Trajectoires
- ....

Règles d'association

➲ Problème du "panier de la ménagère"

ANTECEDENT → CONSEQUENT [Support, Confiance]



[70%, 80%]

Support : 70% des clients ont acheté du beurre et du pain et du lait

Confiance : 80% des clients qui ont acheté du beurre et du pain ont acheté également du lait

Motifs séquentiels

➲ Prise en compte du temps

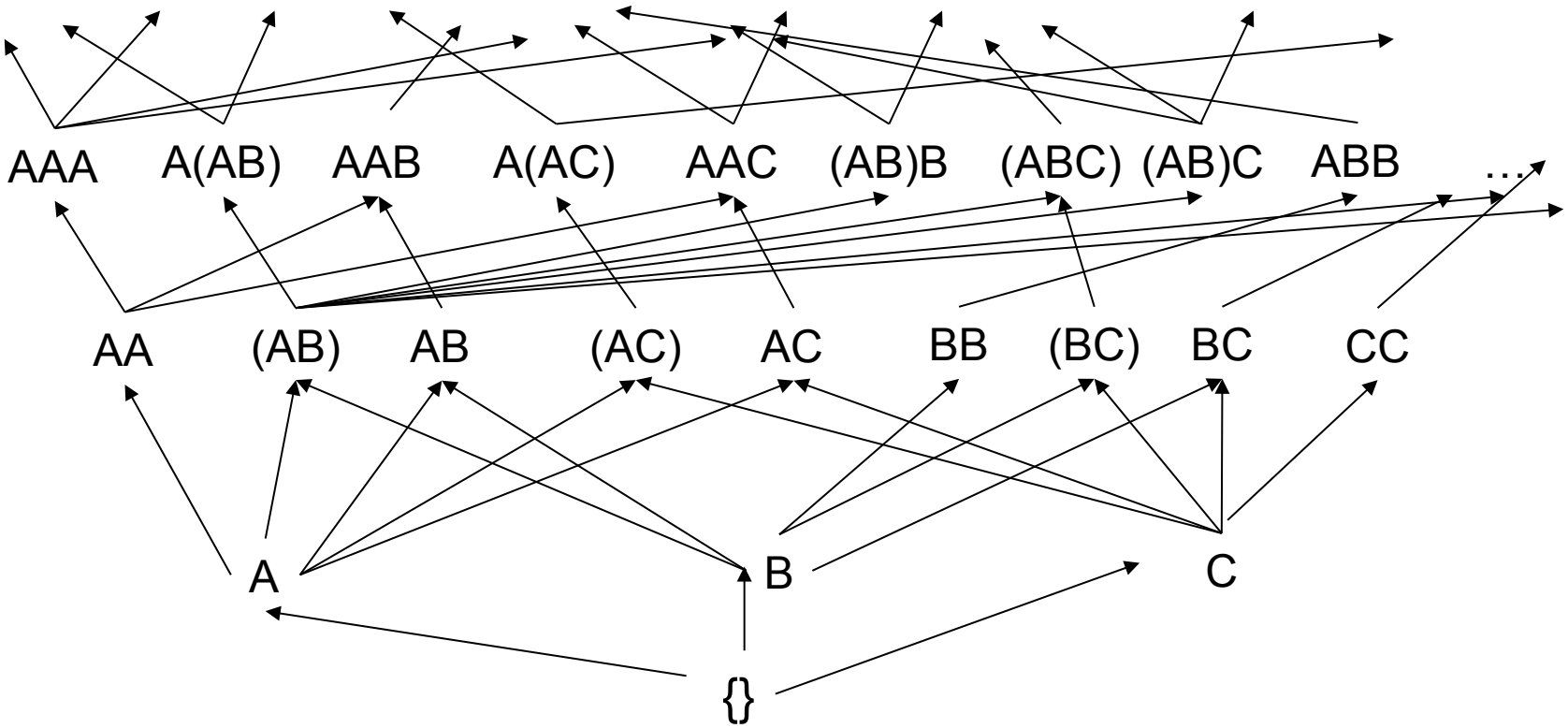
SEQUENCE OF ITEMSETS [Support]

<( **,**  **) (**  **)> [80%]**

Support : 80% des clients ont commandé du fromage et du vin ont commande plus tard du pain

Espace de recherche

Espace infini … borné par la taille de la plus longue séquence

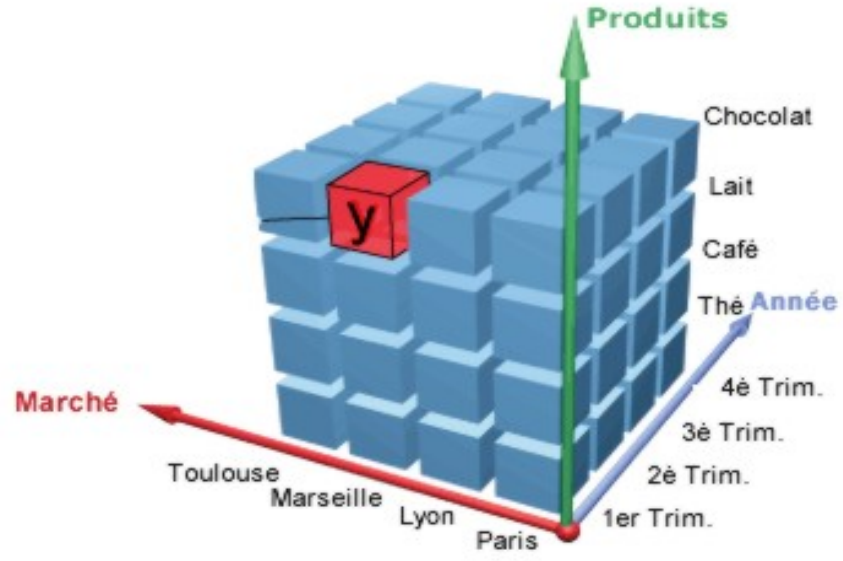


Espace de recherche

Motifs multi-dimensionnels

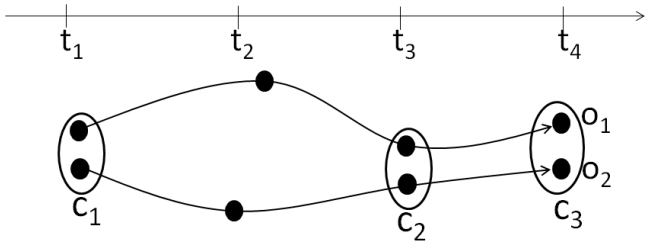
➲ **Les items sont définis selon différentes dimensions**

<([Paris, 2e Trim,  **],[*,*,**  **]) ([Lyon, *,])> [60%]**

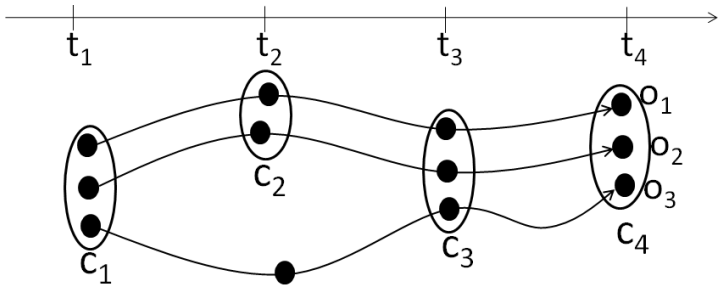


Trajectoires

➲ **Dimensions spatiale et temporelle** [Support par zone et dans le temps]



Swarm (essaim)



Convoy

Analyse des motifs

Données complexes

Extraction de motifs

Transfert de connaissances aux experts

Analyse

Classification

Prédiction

- Décrire les données
- Trouver les motifs fréquents, rares, innatendus...

Classification basée sur les motifs

Données complexes

Extraction de motifs

Transfert de connaissances aux experts

- Analyse
- **Classification**
- Prédiction

➲ Associer à une donnée nouvelle une classe via les motifs utilisés comme des descripteurs

Prédiction basée sur les motifs

Données complexes

Extraction de motifs

Transfert de connaissances aux experts

- Analyse
- Classification
- **Prédiction**

- Prédire ce qui va se passer ensuite via les motifs
- Identifier des tendances

Que peut on faire des données de santé ?

Exemples de projets

Fouiller de puces à ADN

Détecter des épidémies de dengue

Identifier des cellules rares

Exploiter des données de capteurs dans un hopital

Analyser des trajectoires d'oiseaux malades

Exemples de projets

Fouiller de puces à ADN

Détecter des épidémies de dengue

Identifier des cellules rares

Exploiter des données de capteurs dans un hôpital

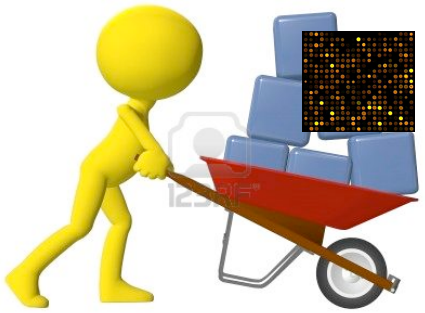
Analyser des trajectoires d'oiseaux malades

Analyse de puces à ADN

Partenaires: MMDN, IRC, IGMM, PIKKO

Pathologies ciblées : Alzheimer, Cancer, HIV

ANR Pradnet



➲Puces à ADN (données numeriques)

Motifs séquentiels

Analyse

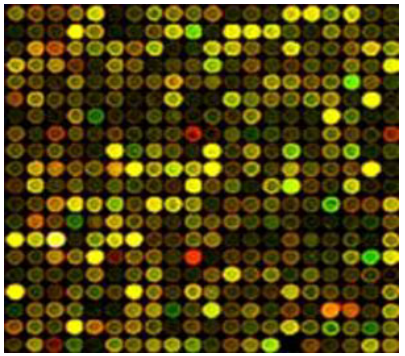
Identification de signatures et de nouveautés

Biologistes

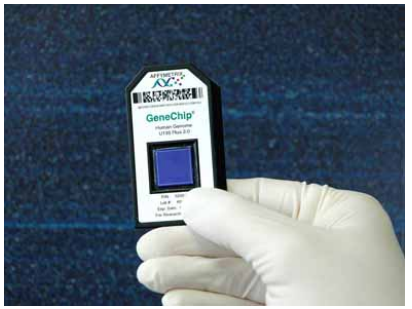
[1] S. Bringay et al. Discovering novelty in sequential patterns: application for analysis of microarray data on Alzheimer disease. MedInfo'2010, Stud Health Technol Inform. 2010;160(Pt 2):1314-8, 2010.
[2] P. Salle et al. Mining Discriminant Sequential Patterns for Aging Brain. AIME'09, 365-369.
[3] P. Salle et al. GeneMining: Identification, Visualization, and Interpretation of Brain Ageing Signatures. MIE'2009, 767-771.
[1] A. Sallaberry et al. Discovering Novelty in Gene Data: From Sequential Patterns to Visualization. ISVC10, 534-543.
[2] A. Sallaberry et al. Sequential Patterns Mining and Gene Sequence Visualization to Discover Novelty from Microarray Data. Journal of Biomedical Informatics, to appear 2011.

Données : puces à ADN

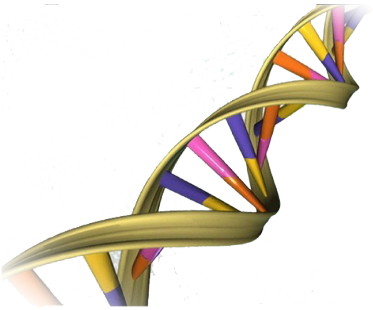
- **Mieux comprendre les maladies génétiques :**
  - perturbation des processus naturels de croissance, de division et de mort des cellules



- **Mesurer l'expression des gènes** et **identifier les lois suivies par ces expressions** en fonction des maladies et des traitements…
  - gènes impliqués dans la maladie ?
  - gènes dont les expressions sont corrélées ?
  - gènes qui inhibent ou activent une fonction ?
  - ….



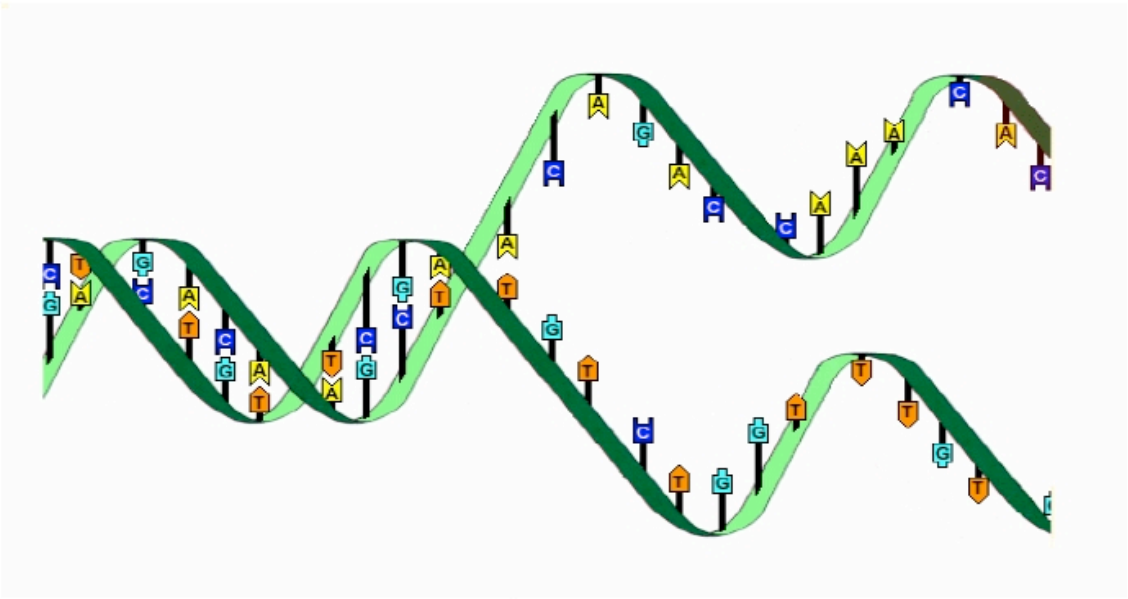
- **Difficultés** pour extraire automatiquement des connaissances liés aux **gros volumes de données**



Données : puces à ADN

Le principe : l'ADN dénaturé reforme spontanément sa double hélice lorsqu'il est porté face à un brin complémentaire (réaction d'hybridation).

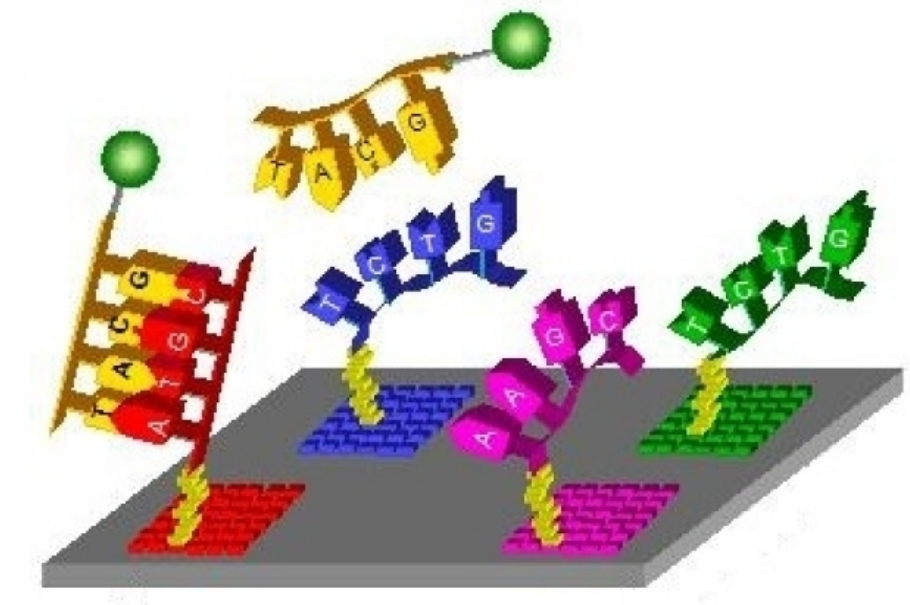
A ≡ T
T ≡ A
G ≡ C
C ≡ G



(a)

Données : puces à ADN

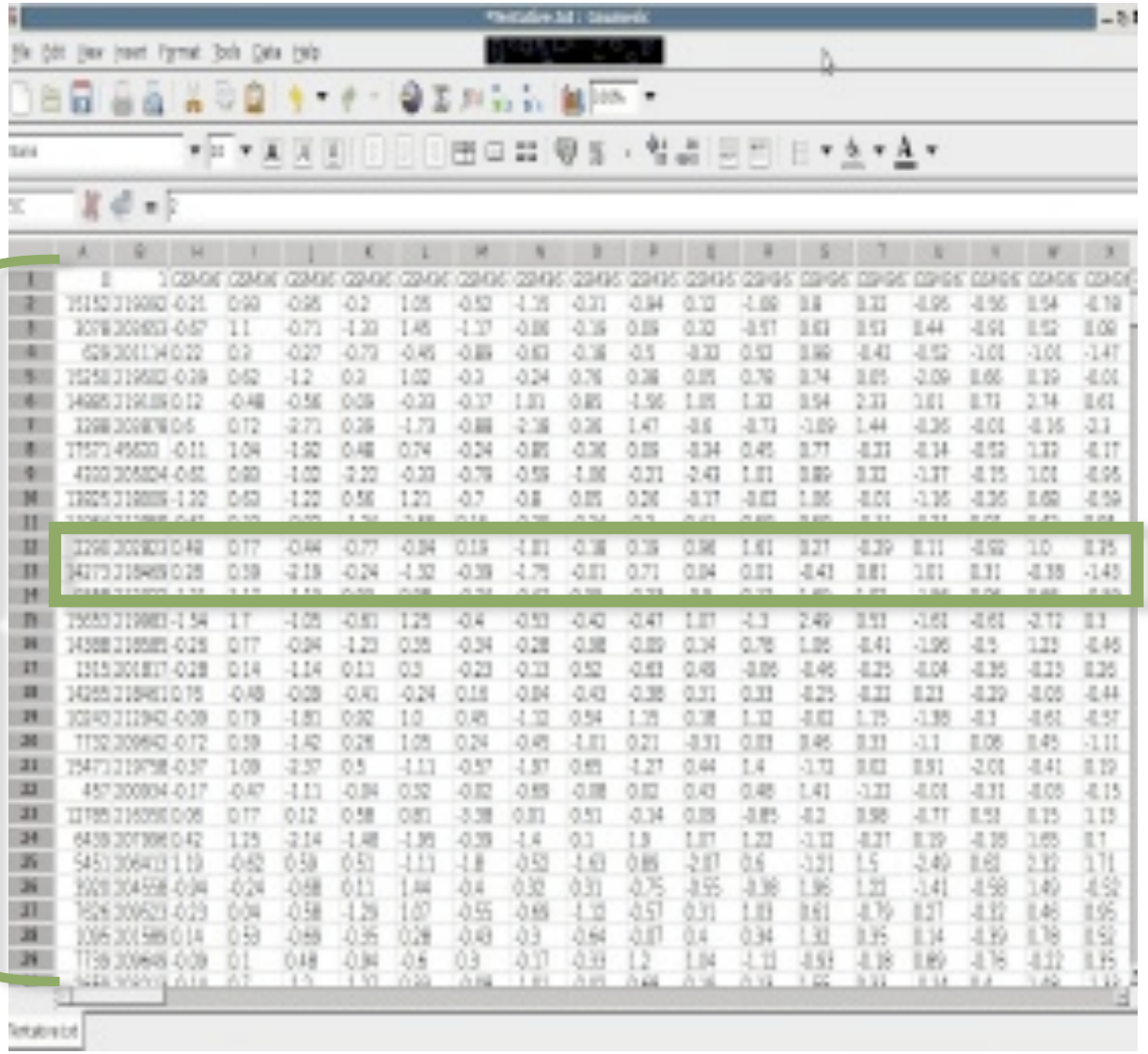
Concrètement… un ensemble de molécules d'ADN fixées en rangées ordonnées sur une petite surface

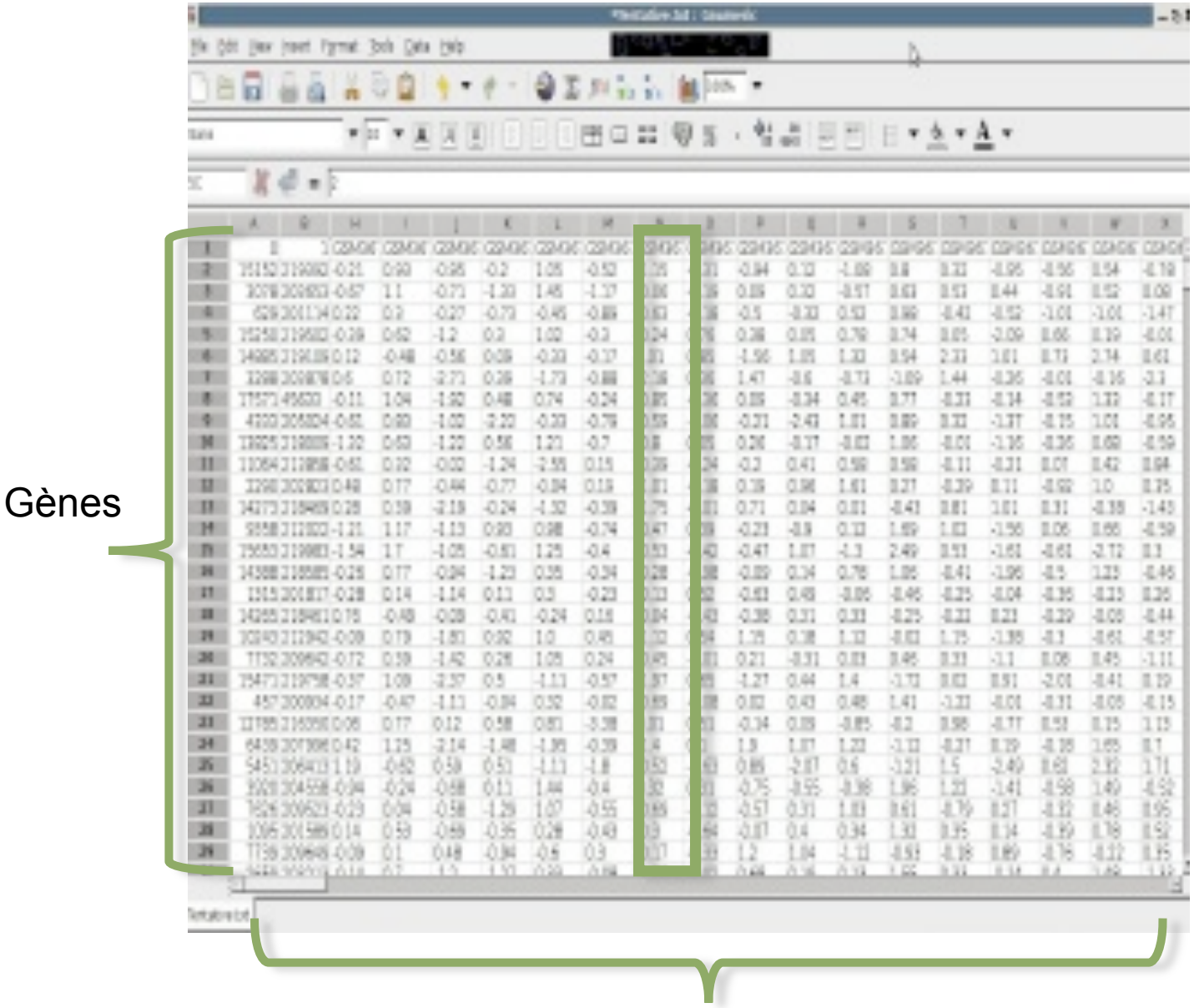


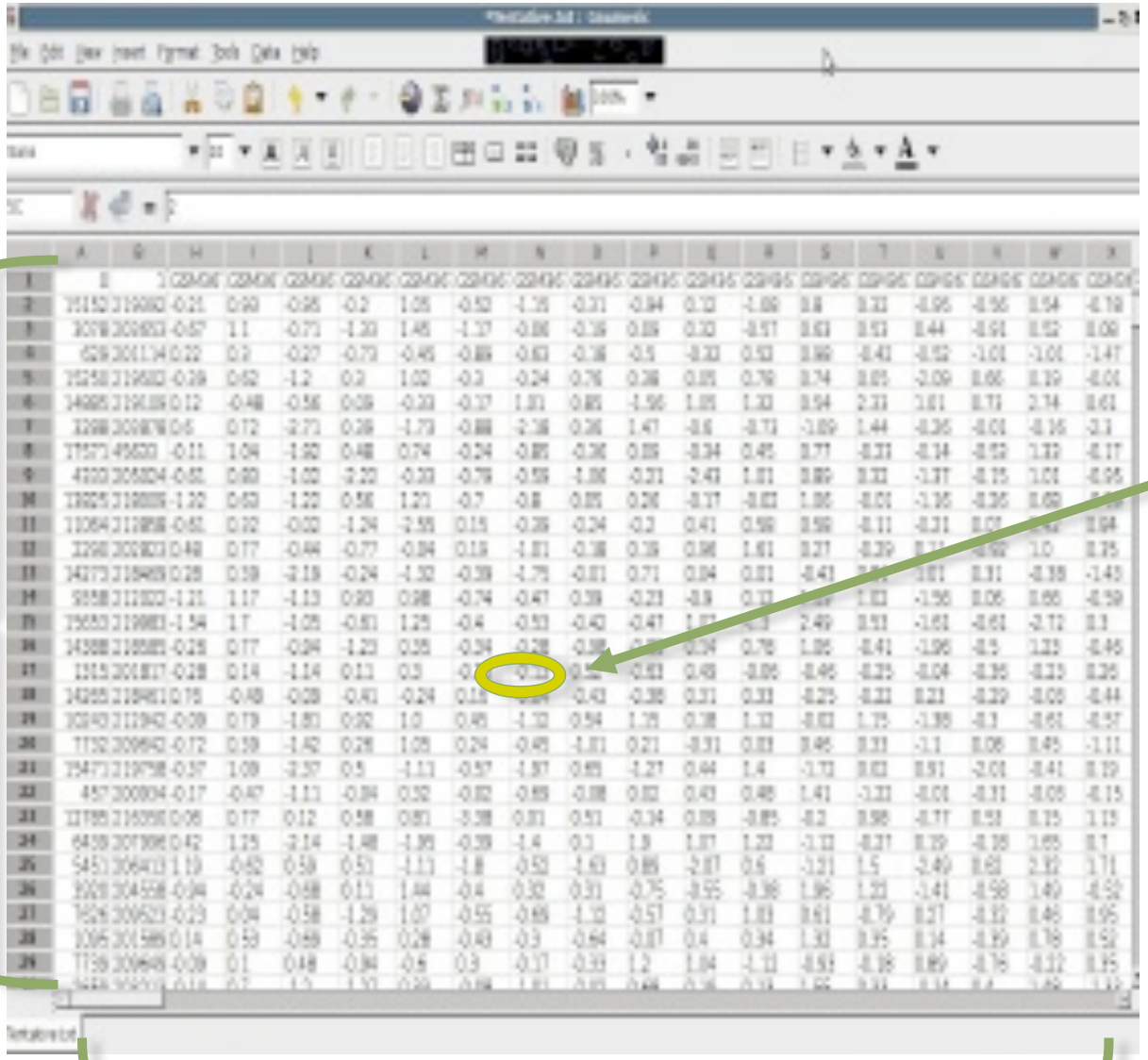
(b)

Expression (couleur) ≈ mesure de la quantité d'ADN dénaturé qui se reforme

Gènes



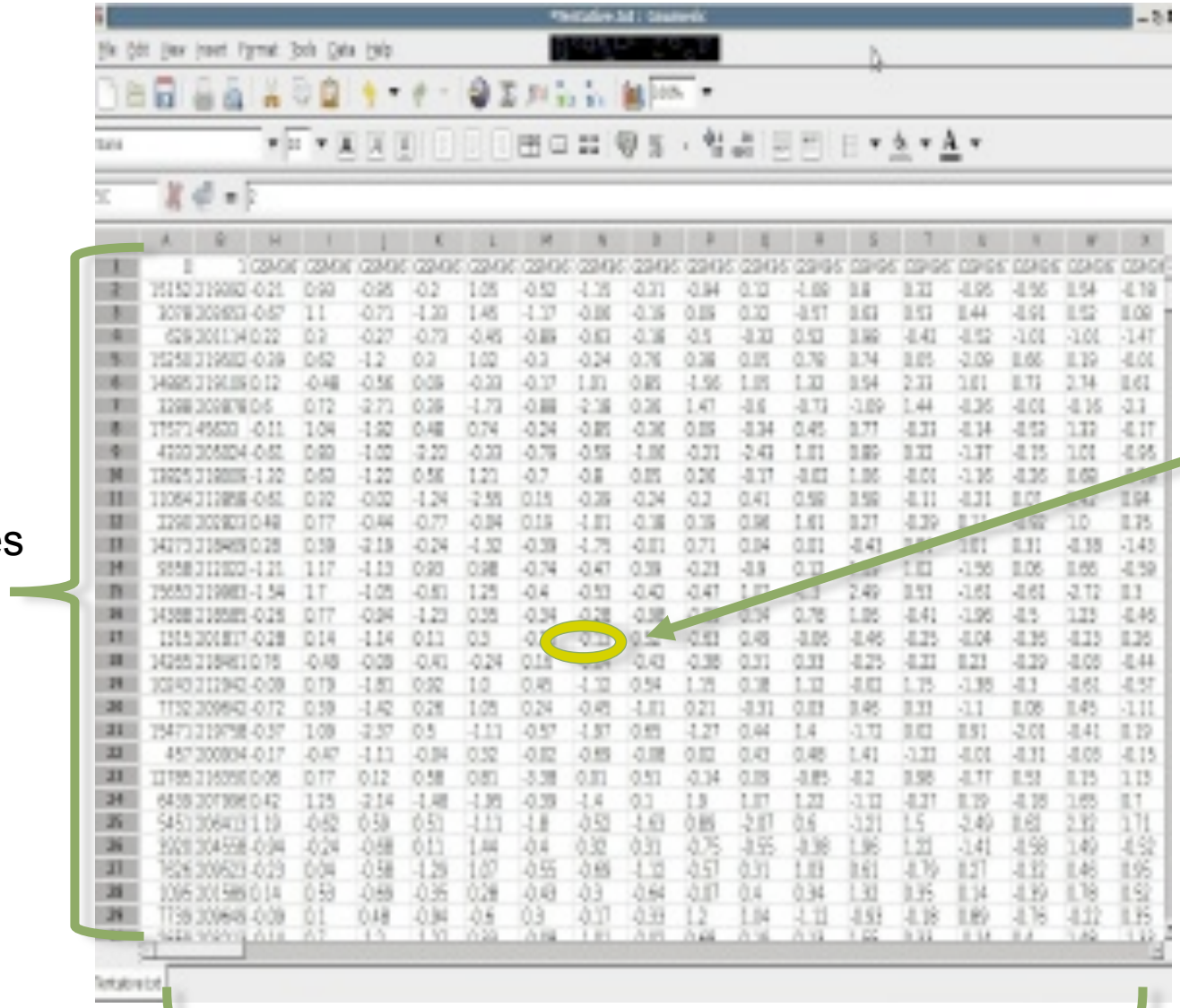

Gènes
Puces



Gènes

Intensité (expression) d'un gène mesuré par une puce

Puces



Gènes

Intensité (expression) d'un gène mesuré par une puce

Puces

Très grande densité : Affymetrix U-133 plus 2.0 Array 54,675 probesets

Alzheimer (AD)

- **Problème de société :** Forme la plus commune de démence
  - 26.6 millions de personnes atteintes (2006)
  - Augmentation du nombre de patients (*4 en 2050)

➢ Intérêt de la communauté biomédicale pour la **découverte des gènes impliqués dans le développement la maladie**

- **MMDN :** travaillent sur l'AD et le vieillissement à partir d'un modèle animal, ***Microcebus murinus***
- **Objectifs :** comparer les tissus du cortex cérébral de lémuriens jeunes (sains) avec ceux de lémuriens âgés (malades et sains)

Cancer du sein

- **Première cause de mortalité chez les 45-64 ans (2004)**
  - Perturbation de la communication cellulaire, associée à une absence de mort cellulaire, engendrant le développement d'amas de cellules cancéreuses (appelées tumeurs) qui échappent aux règles de fonctionnement du corps.

- **IRCM :** utilisent les puces ADN pour comparer les tissus issus de tumeurs du sein, répertoriées selon différents grades.
- **Objectif :** déterminer un ensemble de marqueurs pour typer ces tumeurs.
  - **Enjeu considérable :** Les thérapies sont + ou - toxiques et fonctionnent sur un patient mais pas sur un autre. Typer une tumeur s'avère crucial pour le choix d'une thérapie.

Les motifs séquentiels dans ce contexte

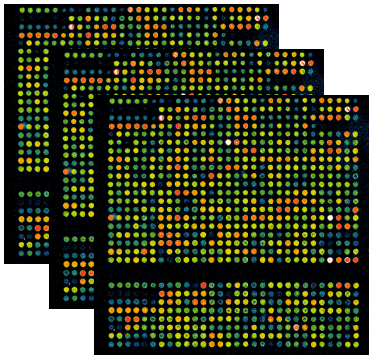
- **Motifs séquentiels :** séquences fréquentes d'itemsets ordonnés

< ( ) () >

- Dans les puces : mettre en évidence des **gènes dont les expressions sont fréquemment ordonnées de la même manière**

< (G5 G4) (G6) >

Problématique générale

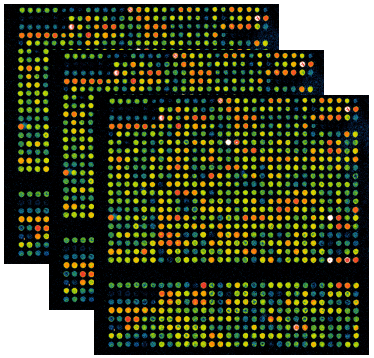


Technologies puces à ADN

Bases de connaissances et bases bibliographiques disponibles en ligne

Nouvelles connaissances

Problématique générale



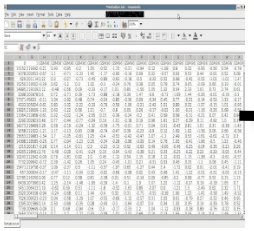
Technologies puces à ADN

Bases de connaissances et bases bibliographiques disponibles en ligne

Nouvelles connaissances

Challenge : exploiter toutes ces données en terme de signification biologique

Processus général



Fouille de données

Motifs séquentiels

Clustering et visualisation

Sélection de motifs séquentiels

Techniques d' interprétation

Nouvelles connaissances biologiques

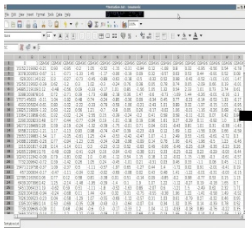
the Gene Ontology

PubMed

GEO Gene Expression Omnibus

KEGG Kyoto Encyclopedia of Genes and Genomes

Processus général



Fouille de données

Motifs séquentiels

Clustering et visualisation

Sélection de motifs séquentiels

Techniques d' interprétation

Nouvelles connaissance s biologiques

Construction de la base de séquences

Microarrays	G1	G2	G3	G4
M1	7.3	6.6	6.6	9.5
M2	5.6	7.4	5.6	5.3
M3	5.7	5.2	8.7	6.8

Construction de la base de séquences

Microarrays	G1	G2	G3	G4
M1	7.3	6.6	6.6	9.5
M2	5.6	7.4	5.6	5.3
M3	5.7	5.2	8.7	6.8

Sequences:

S1 = < (G2 G3)

Construction de la base de séquences

Microarrays	G1	G2	G3	G4
M1	7.3	~~6.6~~	~~6.6~~	9.5
M2	5.6	7.4	5.6	5.3
M3	5.7	5.2	8.7	6.8

Sequences:

S1 = < (G2 G3) (**G1**)

Construction de la base de séquences

Microarrays	G1	G2	G3	G4
M1	~~7.3~~	~~6.6~~	~~6.6~~	9.5
M2	5.6	7.4	5.6	5.3
M3	5.7	5.2	8.7	6.8

Sequences:

S*1* = < (G2 G3) (G1) (**G4**) >

Construction de la base de séquences

Microarrays	G1	G2	G3	G4
M1	7.3	6.6	6.6	9.5
M2	5.6	7.4	5.6	5.3
M3	5.7	5.2	8.7	6.8

Sequences:

S1 = < (G2 G3) (G1) (G4) >

S2 = < (G4) (G1 G3) (G2) >

S3 = < (G2) (G1) (G4) (G3)

Support d'une séquence

- **Fréquence dans la base de séquence**

Sequences:

S1 = < (**G2** G3) (**G1**) (**G4**) >
S2 = < (G4) (G1 G3) (G2) >
S3 = < (**G2**) (**G1**) (**G4**) (G3) >

Sequential pattern:

SP1 = < (**G2**) (**G1**) (**G4**) > support 2/3

Motifs séquentiels

- **Support minimum fixé par l'utilisateur**

Example:

Let be **minSupport** = 2/3

Sequential patterns:

SP1 = < (G2) (G1) (G4) > support 2/3

SP2 = < (G3 G5) > support 2/3

SP3 = < (G1) (G2) (G4) > support 1/3

Motifs séquentiels

- **Support minimum fixé par l'utilisateur**

Example:

Let be **minSupport** = 2/3

Sequential patterns:

SP1 = < (G2) (G1) (G4) > support 2/3

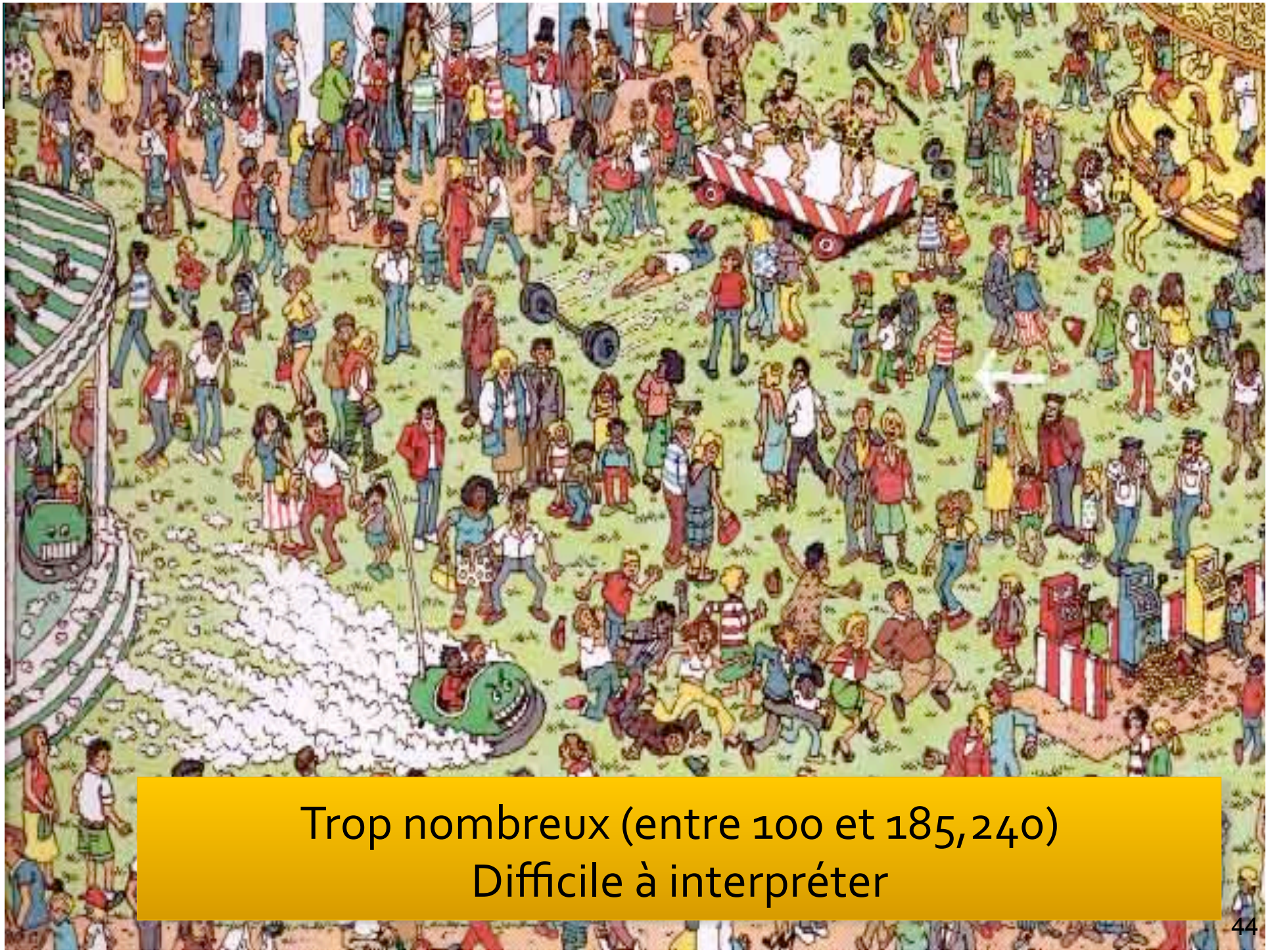
SP2 = < (G3 G5) > support 2/3

~~*SP3* = < (G1) (G2) (G4) > support 1/3~~

Motifs séquentiels +++

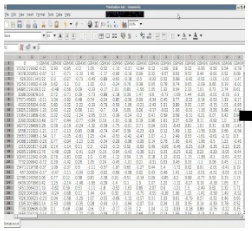
- **Les discriminants**
  - Fréquents dans une classe (malades)
  - Non fréquents dans la classe complémentaire (sains)

- **Les plus généraux, spécifiques…**



Trop nombreux (entre 100 et 185,240)
Difficile à interpréter

Processus général



Fouille de données

Motifs séquentiels

Clustering et visualisation

Sélection de motifs séquentiels

Techniques d' interprétation

Nouvelles connaissance s biologiques

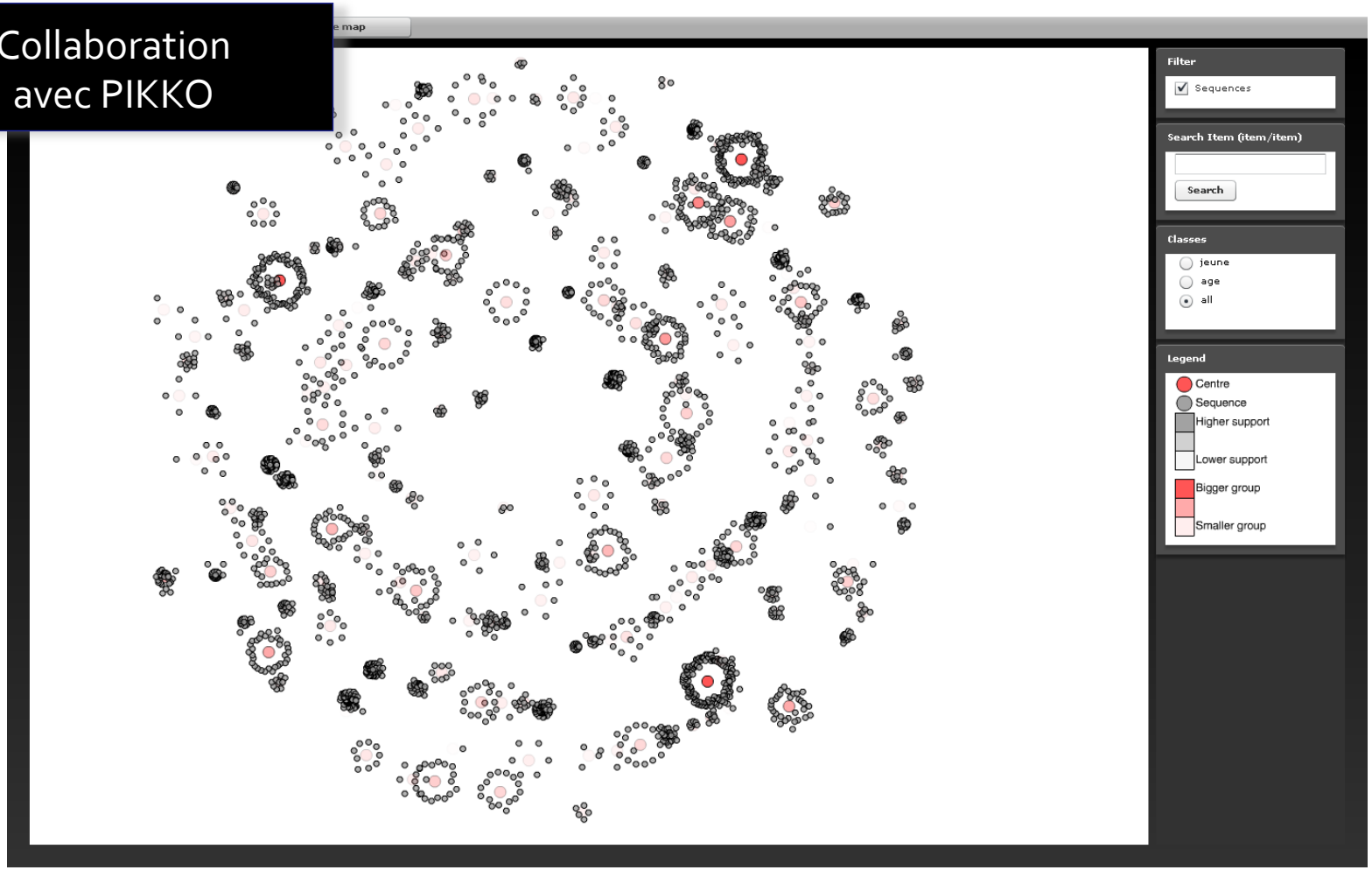
Comment comparer les motifs ? (Saneifar 2008)

S75%=<(G1)(G2 G3)>
S'75%=< (G2 G3) (G1)>

- **Mesure de similarité**
  - Gènes communs et non communs
  - Ordre des gènes
  - Support

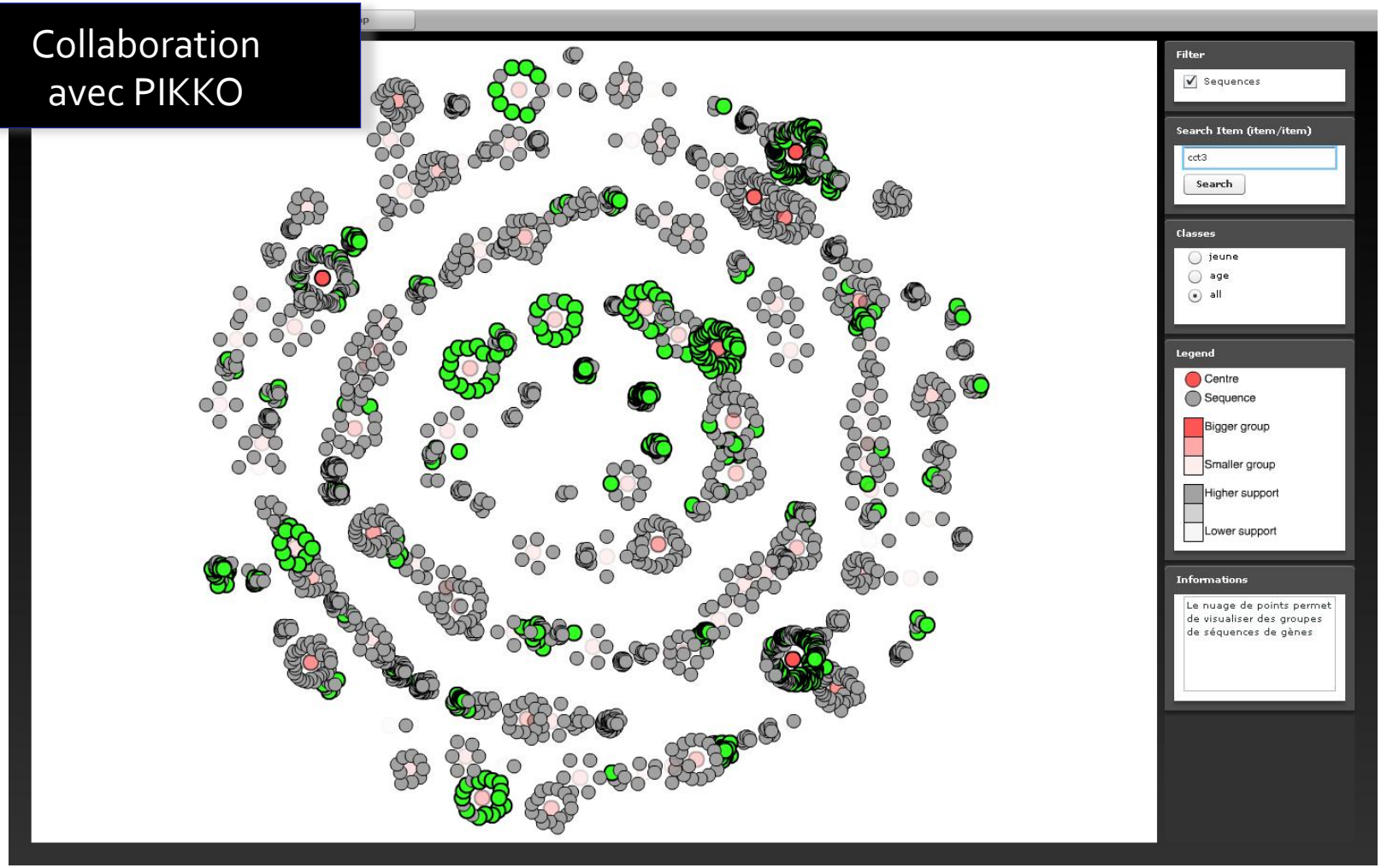
Clustering k-means

Collaboration avec PIKKO



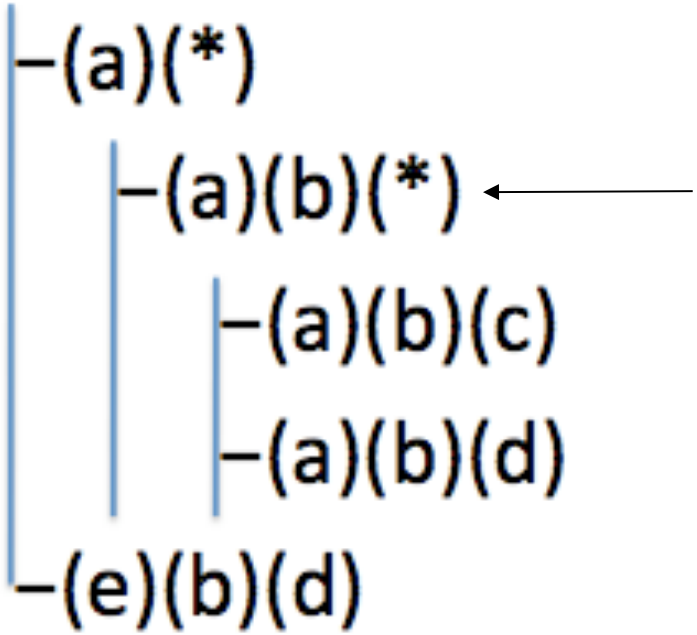
Clustering k-means

Collaboration avec PIKKO



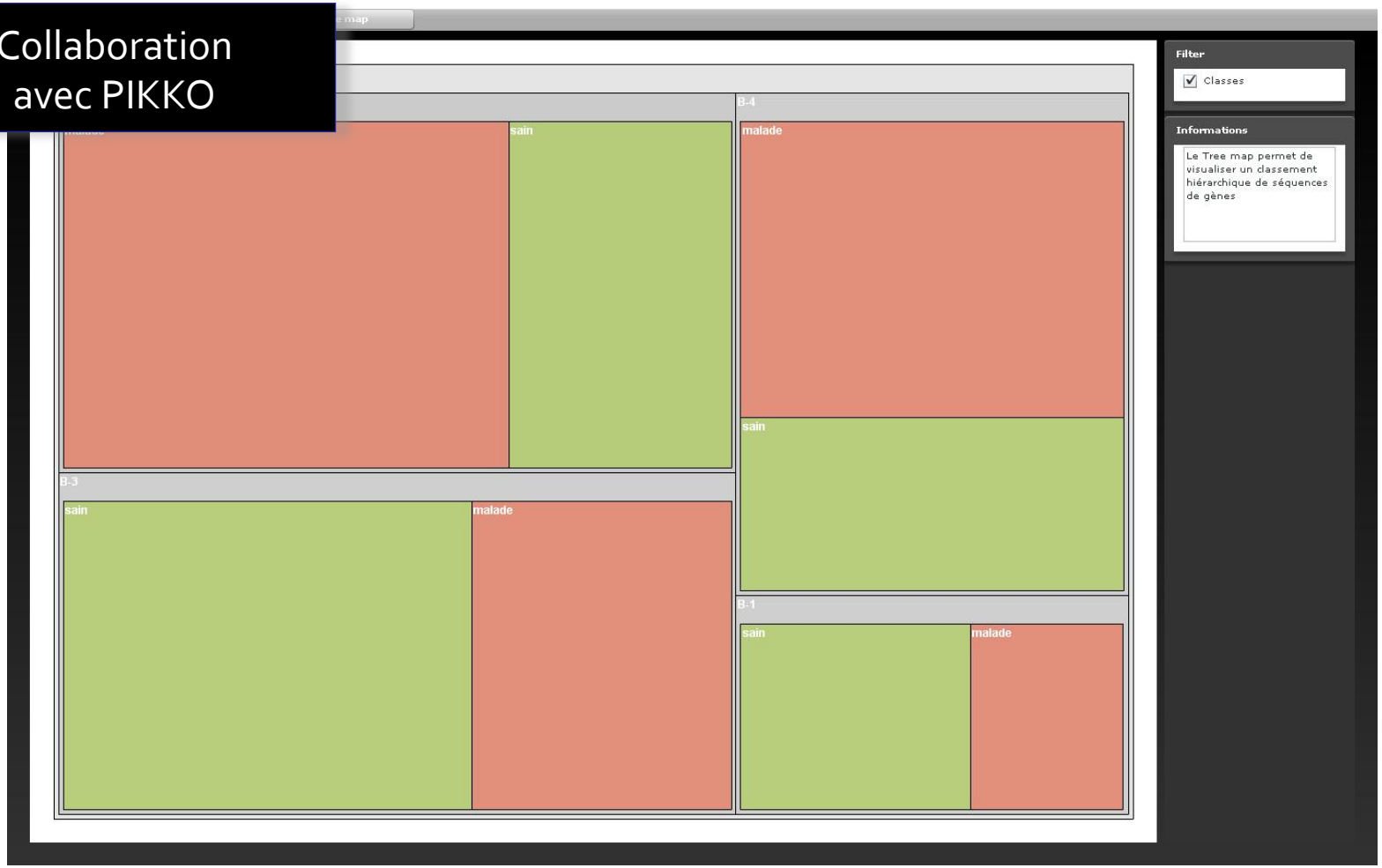
Clustering hiérarchique (Nin Guerero 2009)

- **Exemple: (a)(b)(c), (a)(b)(d), (e)(b)(d)**



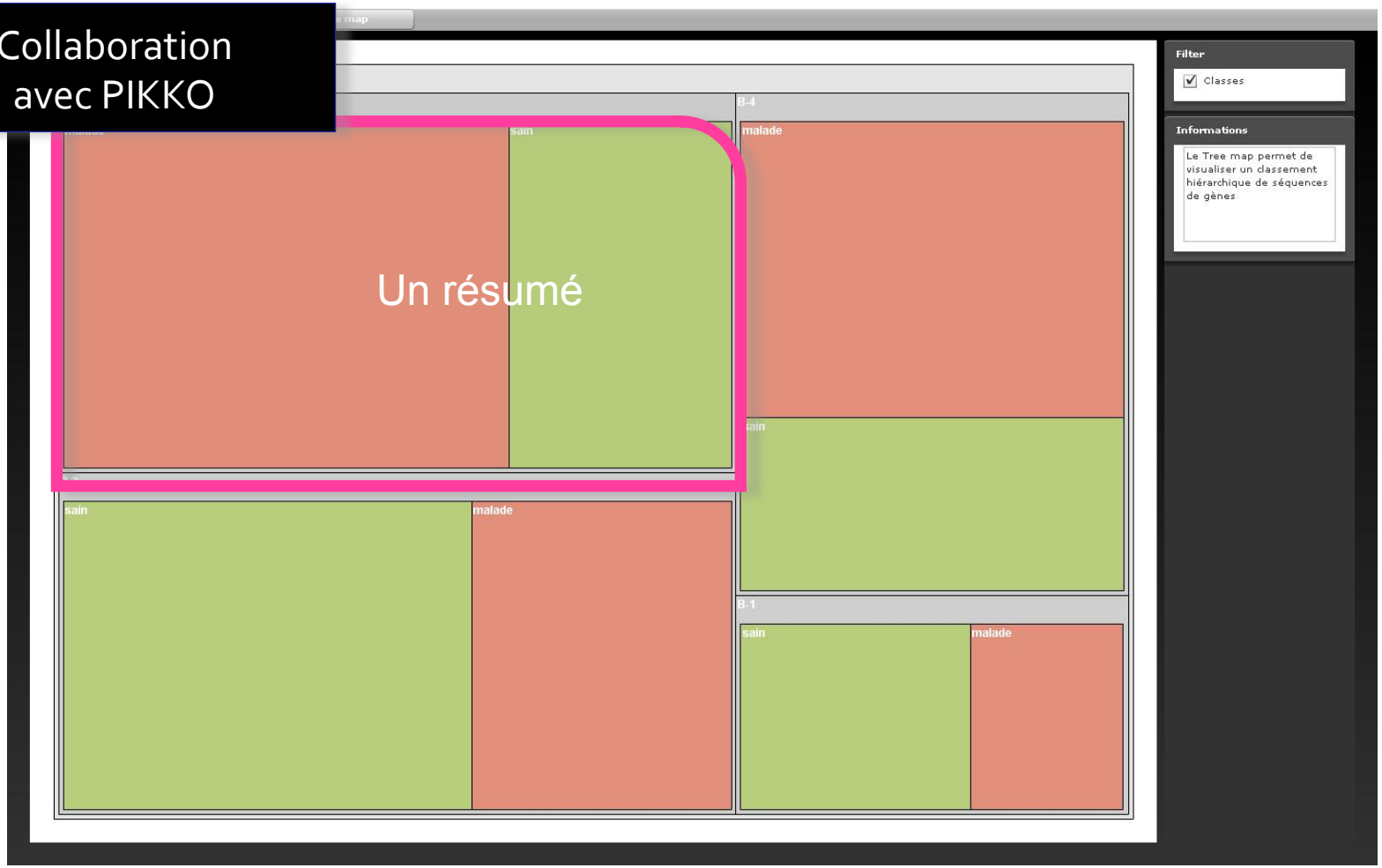
Clustering hiérarchique (Nin Guerero 2009)

Collaboration
avec PIKKO



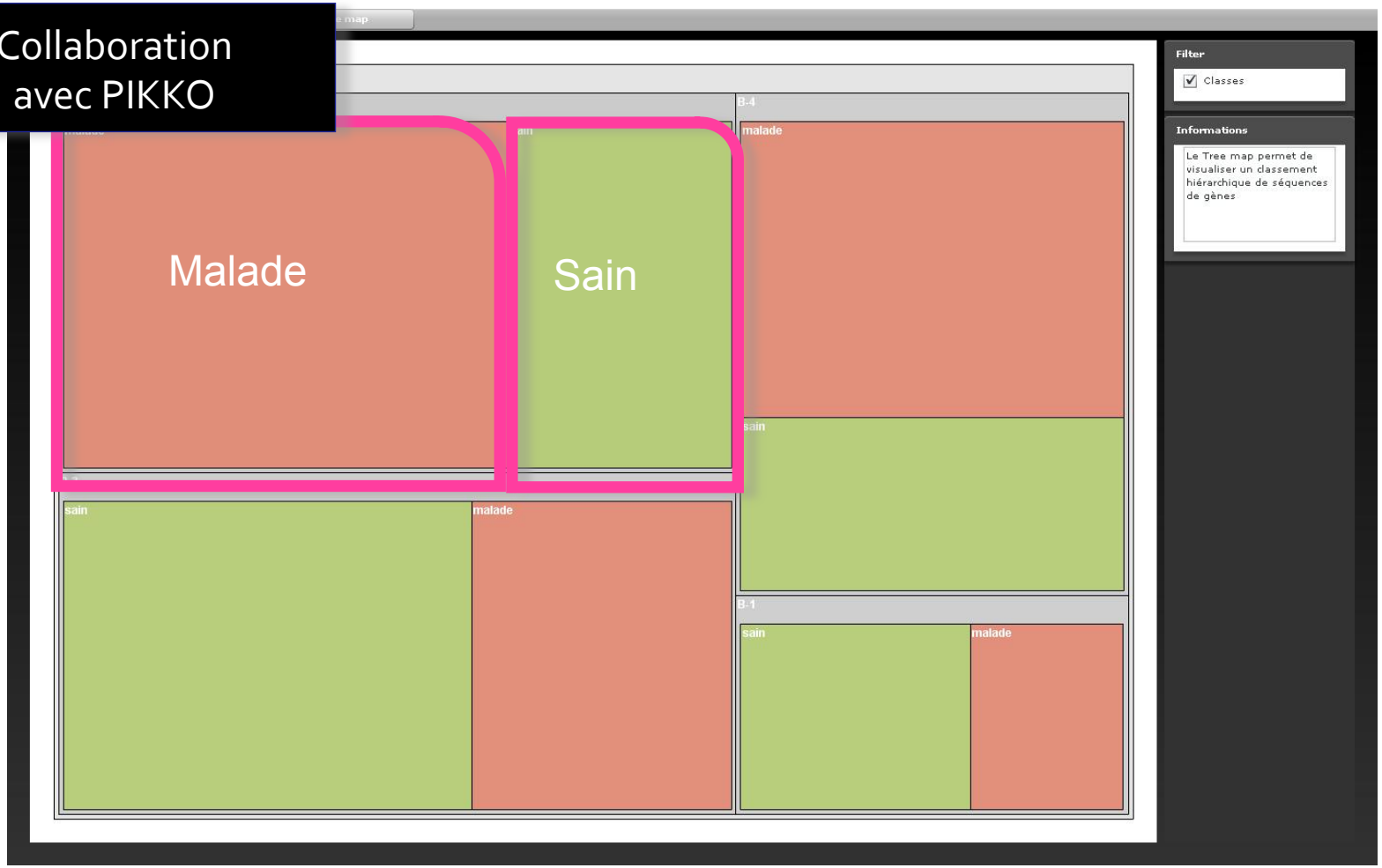
Clustering hiérarchique (Nin Guerero 2009)

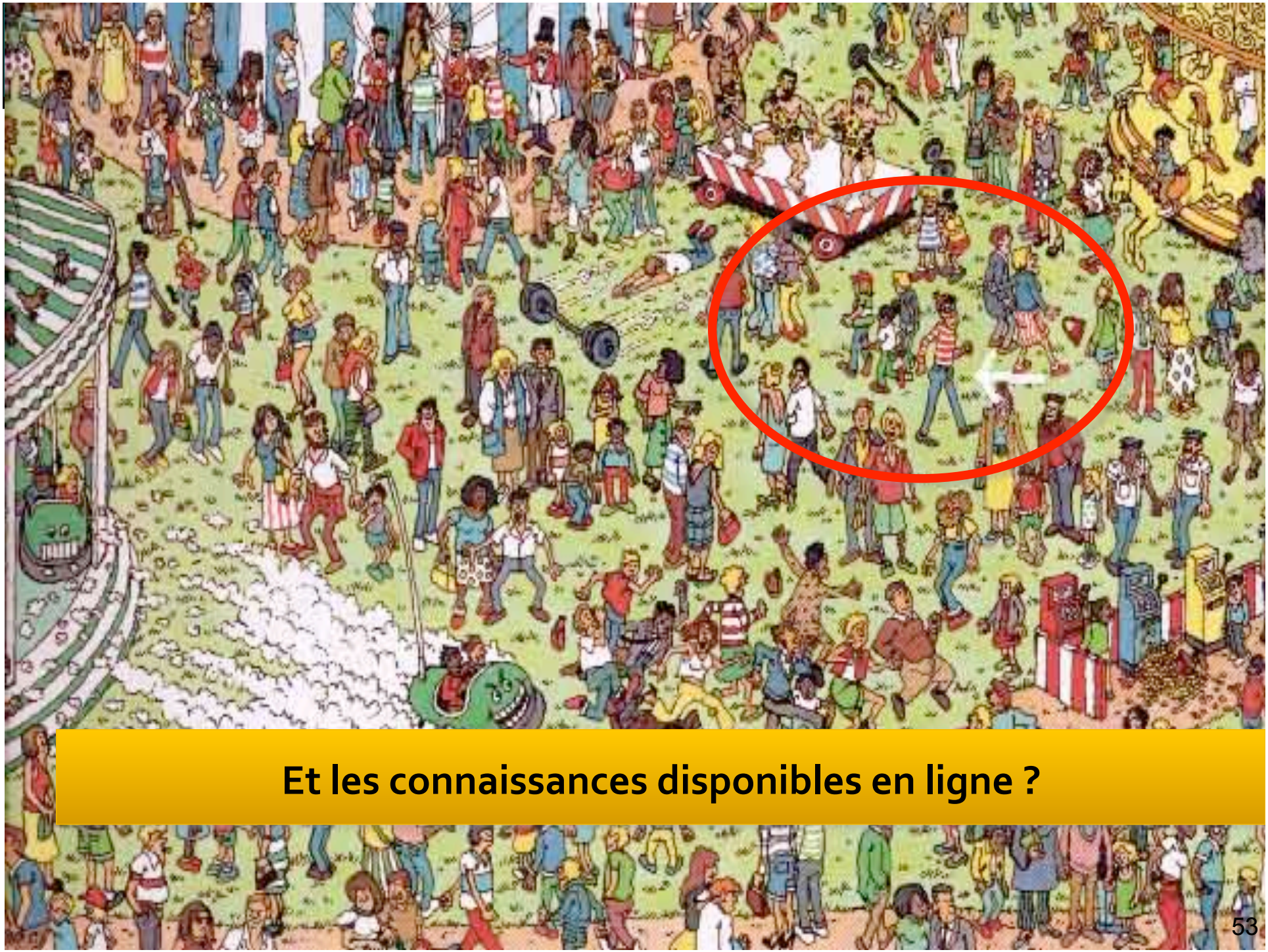
Collaboration
avec PIKKO



Clustering hiérarchique (Nin Guerero 2009)

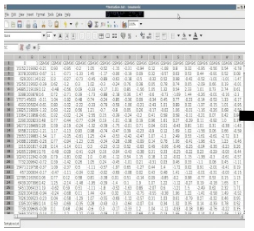
Collaboration
avec PIKKO





Et les connaissances disponibles en ligne ?

Processus général



Fouille de données

Motifs séquentiels

Clustering et visualisation

Sélection de motifs séquentiels

Techniques d' interprétation

Nouvelles connaissances biologiques

the Gene Ontology

PubMed

GEO Gene Expression Omnibus

KEGG Kyoto Encyclopedia of Genes and Genomes

Interprétation des motifs via les documents (Bringay 2010)

S75%,25%=<(G1)(G2 G3)>

Textes

PubMed

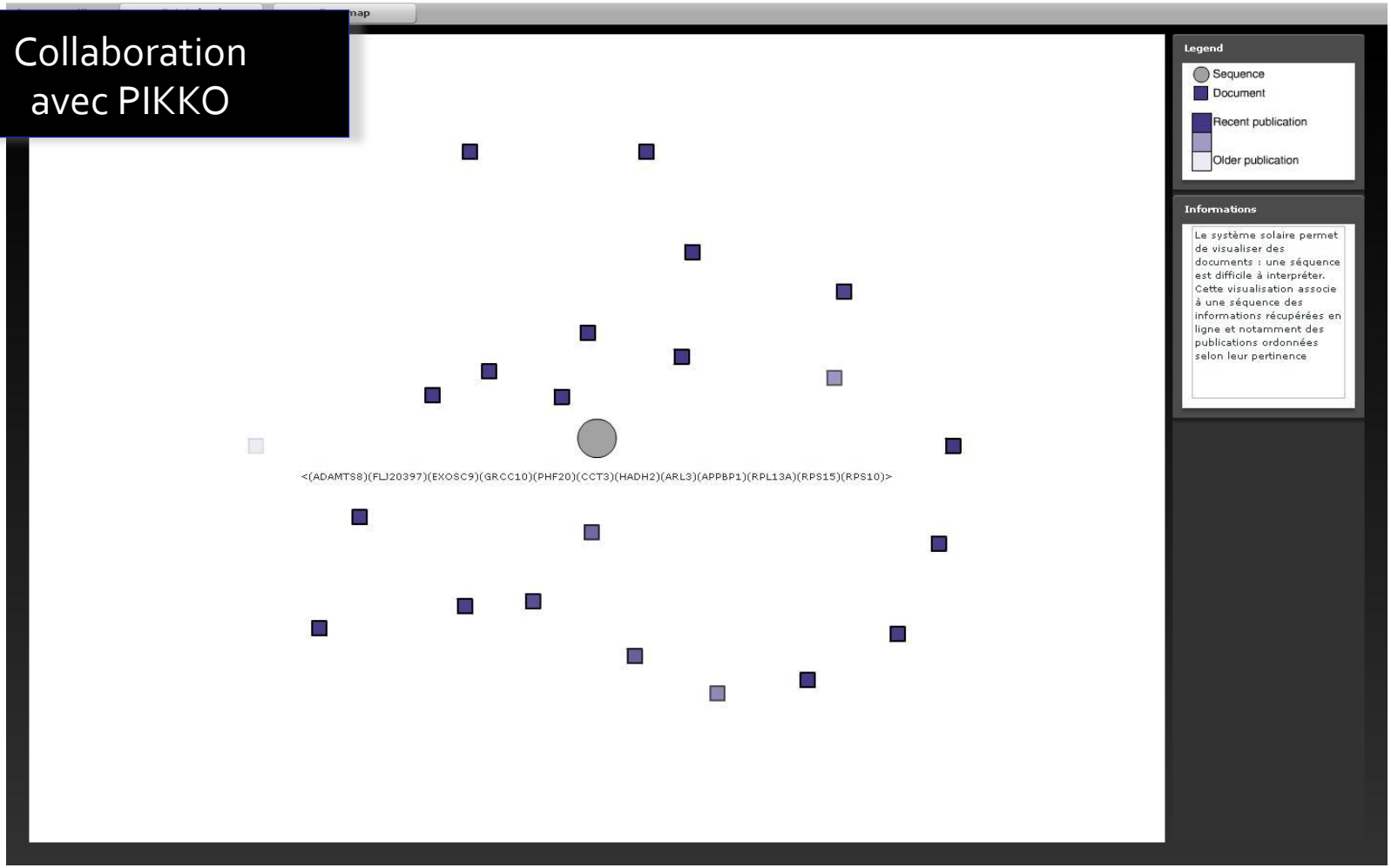
- **Recherche de documents associés avec les gènes des motifs**

- **Objectifs: validation + recherche de nouveautés**

  - **Séquences populaires et innovantes**

Visualisation de documents (Sallaberry 2010)

Collaboration avec PIKKO



Séquences innovantes associées avec des documents = nouvelle connaissance ayant un signification biologique

Exemple de motif pertinent

S_{75}=<(MRVI1)(PGAP1)(PLA2R1)(A2M)(GSK3B)>

- Protéines impliquées dans les mécanismes de **signalisation** et du **métabolisme**
- Certaines interfèrent avec les événements cellulaires de la maladie d'**Alzheimer**

Conclusions préliminaires

- De **nouvelles connaissances pour les biologistes** qui leur permettent d'étudier l'impact de l'expression des gènes sur les maladies
- Un **outil** pour rendre ces données manipulables

  - **What else?**

… motifs difficiles à interpréter…

Et les experts …

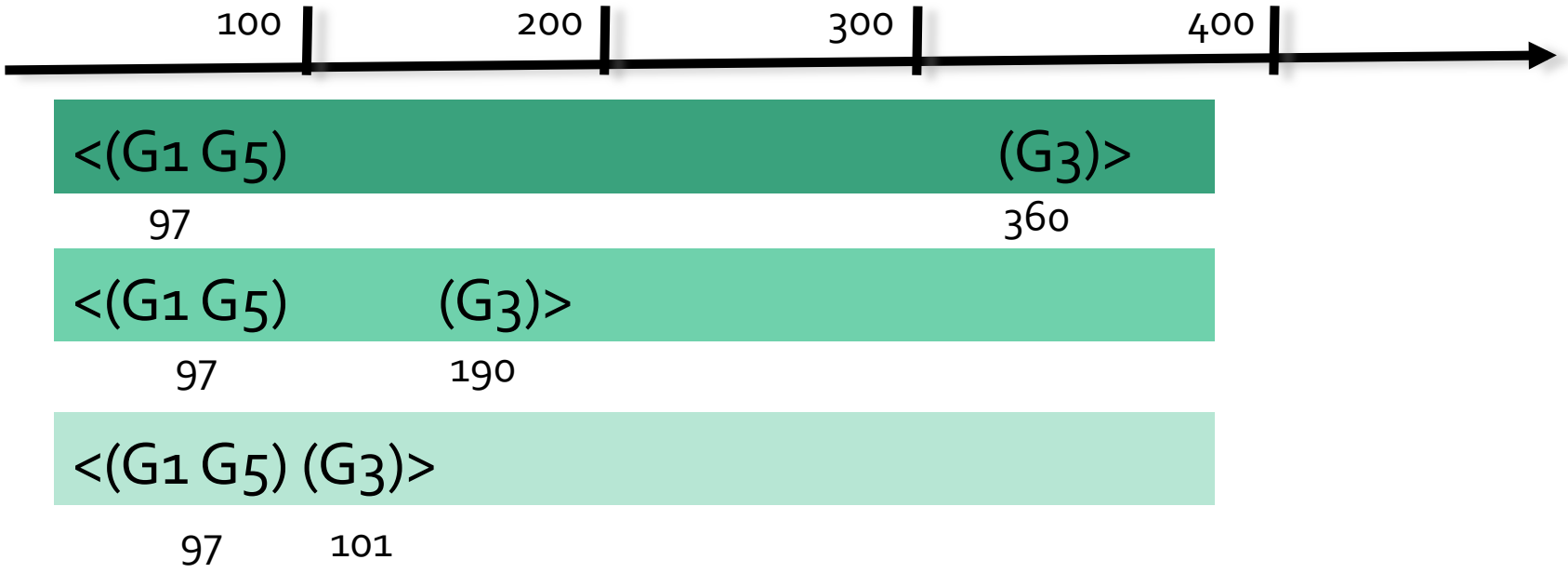
- **Motifs séquentiels :** ne sont pas facilement
  - **Compréhensibles** et **manipulables** par les experts
  - **Interprétables** d'une manière linguistique sans la définition d'une partition stricte des valeurs d'expression des gènes

<(G1 G5)(G3)>

Et les experts …

- **Motifs séquentiels :** ne sont pas facilement
  - **Compréhensible** et **manipulables** par les experts
  - **Interprétables** d'une manière linguistique sans la définition d'une partition stricte des valeurs d'expression des gènes

<(G1 G5)(G3)>



Motifs à écarts flous (Bringay 2009)

- **Motifs séquentiels :** ne sont pas facilement
  - **Compréhensible** et **manipulables** par les experts
  - **Interprétables** d'une manière linguistique sans la définition d'une partition stricte des valeurs d'expression des gènes

<(G1 G5)(G3)>

<(G1 G5)(very over expressed 0,8) (G3)>

G3 is far much expressed compared to G1 and G5, which are expressed in a similar way

Recherche des motifs à écarts flous

Puce	Séquence de gènes
M1	<(G2)(G1 G5)(G3)(G4)> 3.7 4 4.3 5 7
M2	<(G2)(G1 G5)(G4)(G3)> 3.2 4.2 4.7 10 12

< (G1 G5) (G3) >

Recherche des motifs à écarts flous

Puces	Séquences de gènes
M1	<(G2)(G1 G5)(G3)(G4)> 3.7 4 4.3 5 7
M2	<(G2)(G1 G5)(G4)(G3)> 3.2 4.2 4.7 10 12

- **Différence d'Itemset δ(it2; it1):** la valeur absolue de la différence entre l'intensité du gène it2 et it1

- **Exemple:**

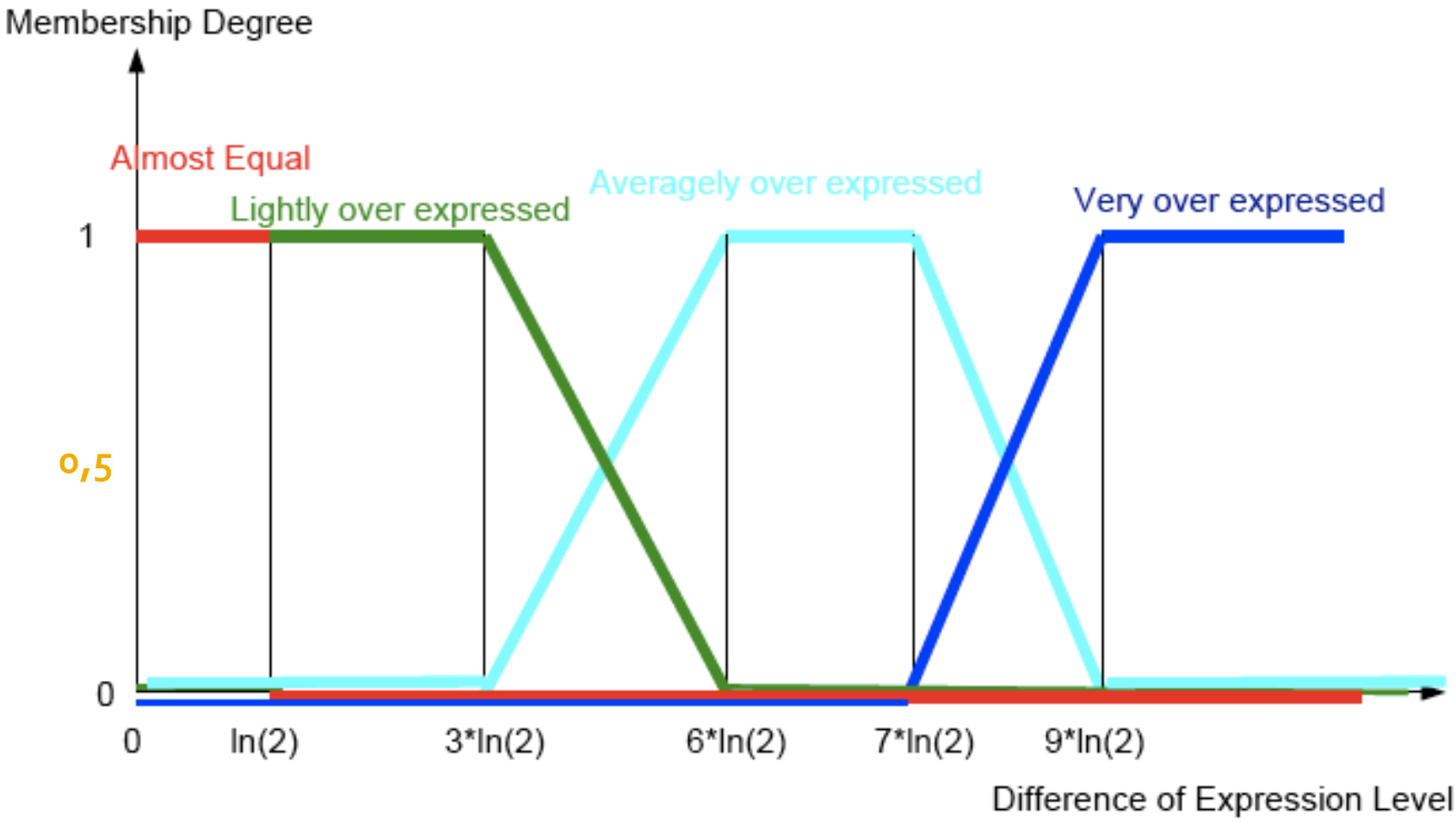
M1: δ((G3); (G1 G5))

= I5-4I=1

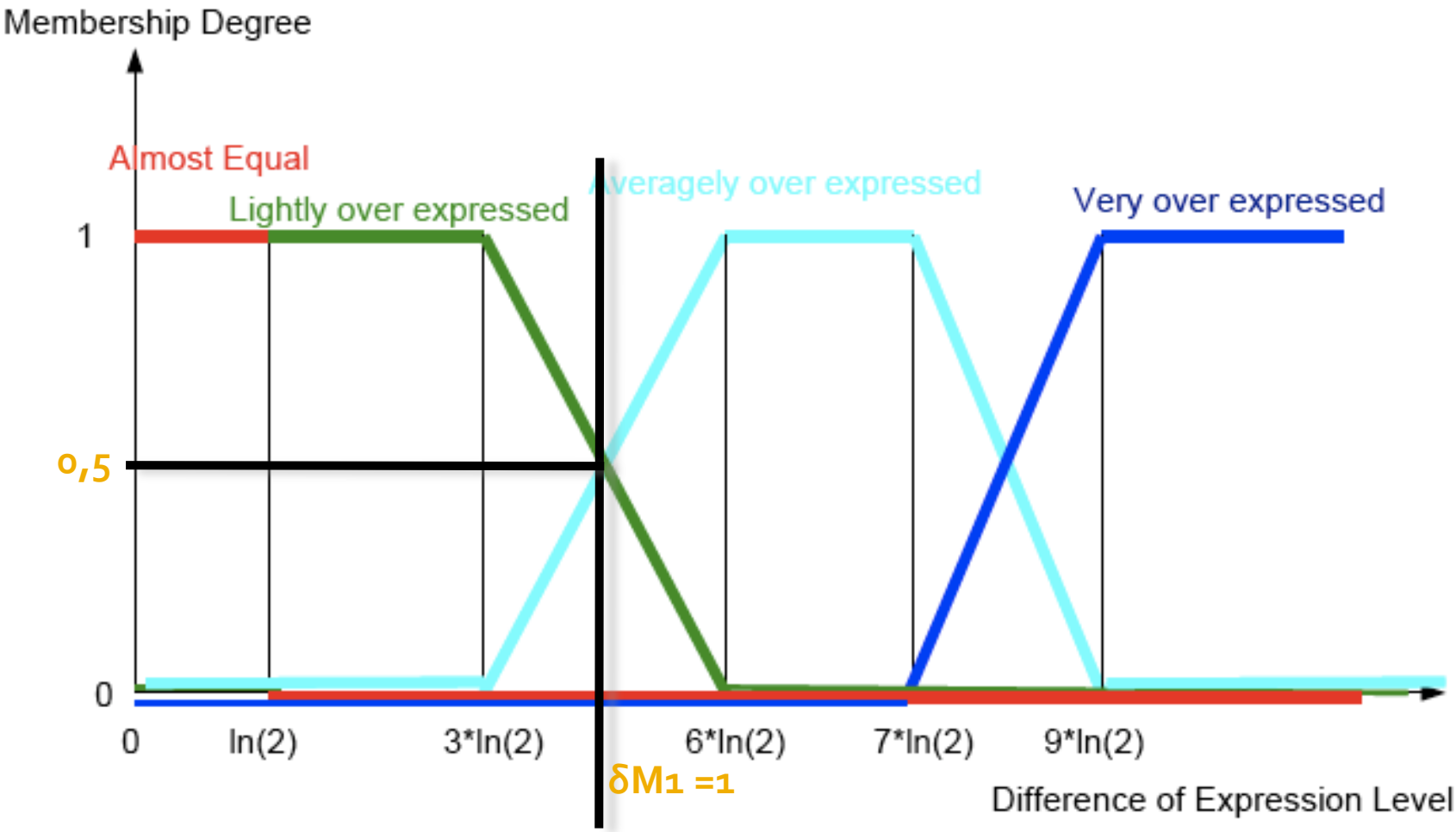
M2: δ((G3); (G1 G5))

= I12-4.2I=7,8

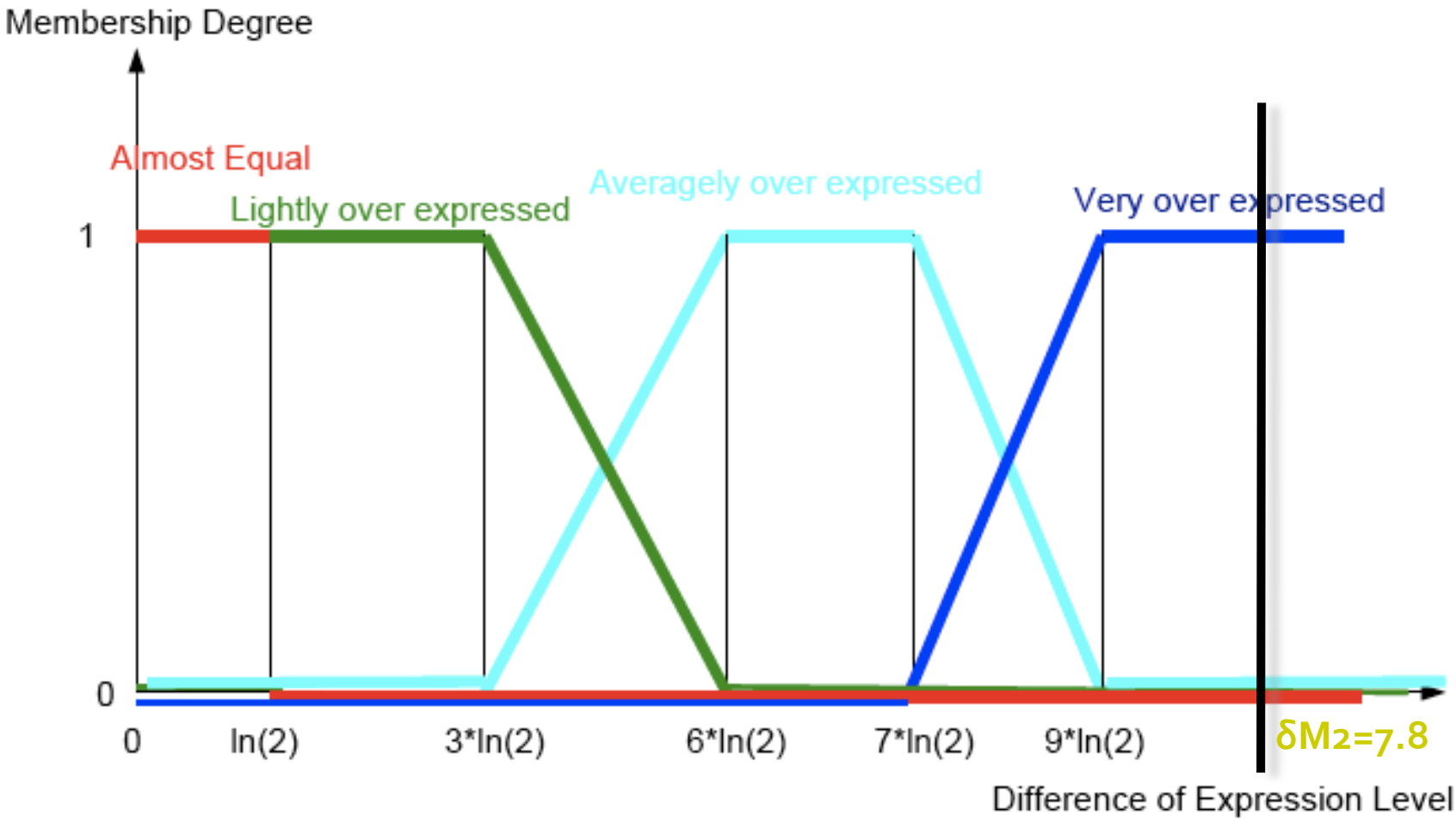
Recherche des motifs à écarts flous



Recherche des motifs à écarts flous



Recherche des motifs à écarts flous



Recherche des motifs à écarts flous

Puces	Séquences de gènes
M1	<(G2)(G1 G5)(G3)(G4)> 3.7 4 4.3 5 7 (moyennement sur-exprimé; 0,5)
M2	<(G2)(G1 G5)(G4)(G3)> 3.2 4.2 4.7 10 12 (très sur exprimé ;1)

Degré d'une séquence à écart flou :

$$F_{SFG}(M) = \top(d1, \ldots, dn\text{-}1)$$

T-norm appliquée à tous les degrés des écarts de la séquence

$$F_{(G3)(G1\ G5)}(M1) = 0{,}5$$
$$F_{(G3)(G1\ G5)}(M2) = 1$$

Recherche des motifs à écarts flous

Puces	Séquences de gènes
M1	<(G2)(G1 G5)(G3)(G4)> 3.7 4 4.3 5 7 (moyennement sur-exprimé; 0,5)
M2	<(G2)(G1 G5)(G4)(G3)> 3.2 4.2 4.7 10 12 (très sur exprimé ;1)

Support d'une séquence à écart flou

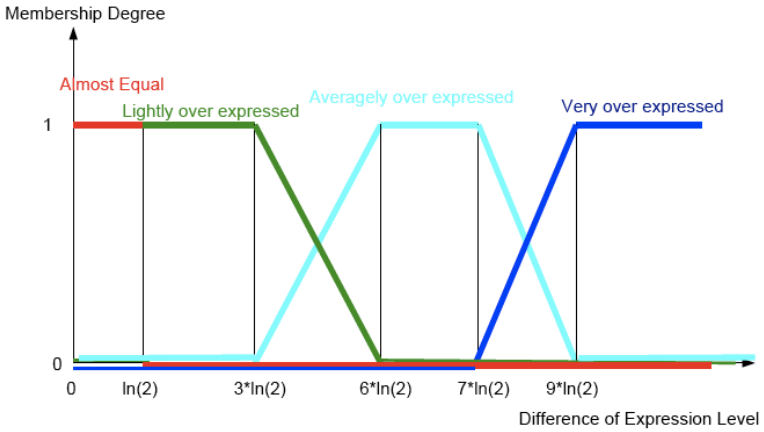
$$Freq(s_{FG}) = \frac{\sum_{o \in \mathcal{O}} \left[F_{s_{FG}}(s_o) \right]}{|\mathcal{O}|}$$

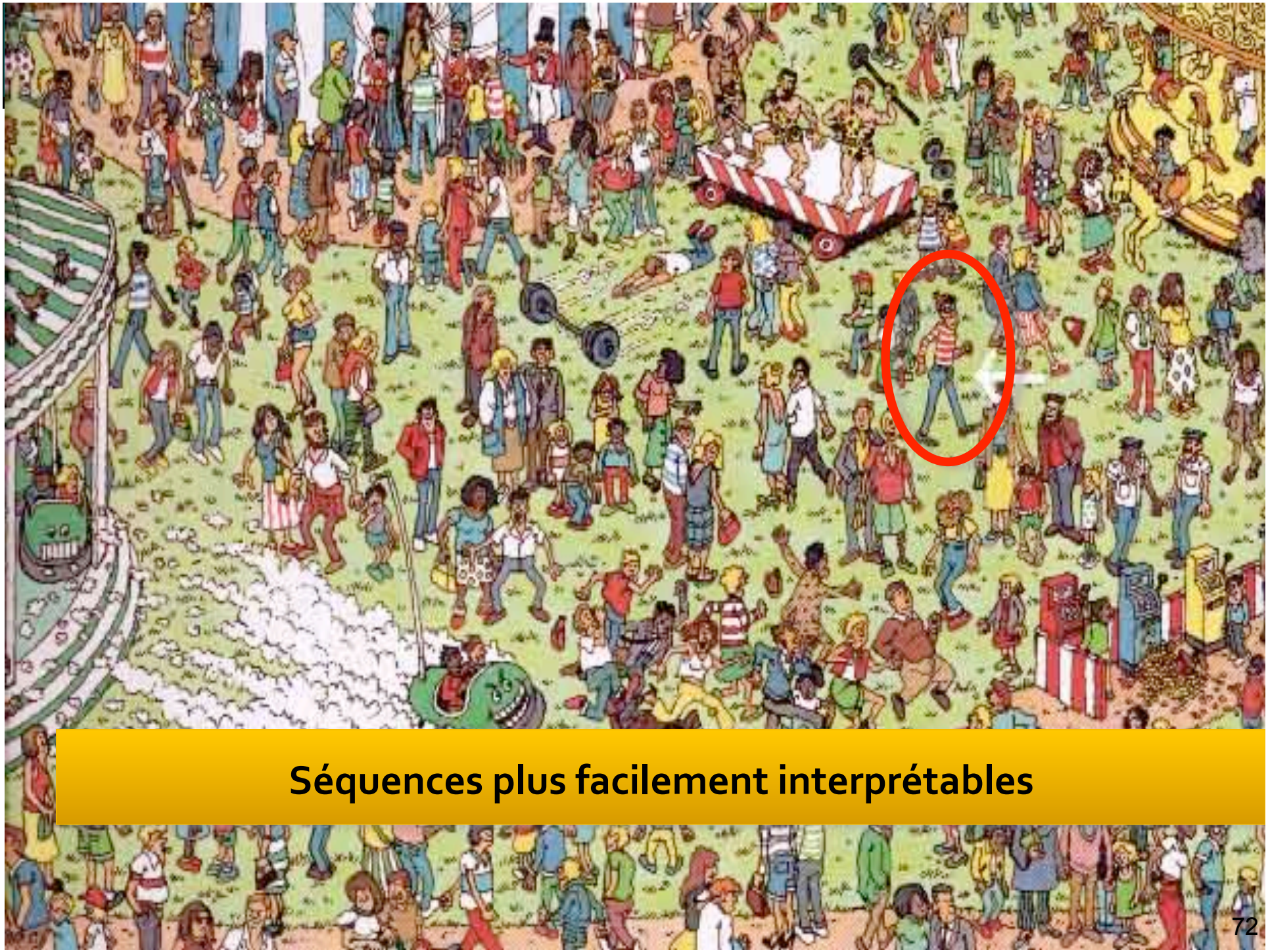
Pourcentage de puces vérifiant le motif à écarts flous

$Freq_{(G3)very(G1\ G5)} = 1/2 = 0,5$

Conclusions et perspectives

- Motifs à écarts flous **plus compréhensibles** et **manipulables** par les experts
- **Simple à calculer** (post-traitement)
- Utiliser les propriétés des contraintes liées au flou pour améliorer les performances de notre algorithme
- **Nouvelle information :** motifs à écart flous discriminant
- prédiction des types de cancer



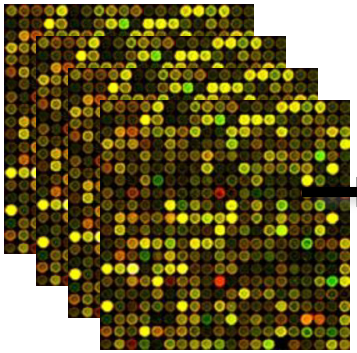


Séquences plus facilement interprétables

… utiliser les motifs pour prédire le grade d’un cancer

Définition du problème

Phase 1 : apprentissage du modèle



Données

Fouille de données

Motifs séquentiels

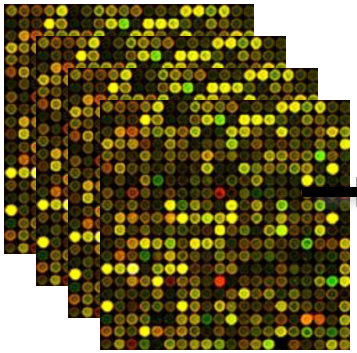
Grade 1

Grade 2

Grade 3

Définition du problème

Phase 1 : apprentissage du modèle



Données

Fouille de données

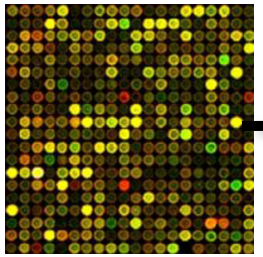
Motifs séquentiels

Grade 1

Grade 2

Grade 3

Phase 2 : prédiction du grade



Classification

Grade 3

Apprentissage du modèle

- **Etape 1 : Extraction des motifs**
- Pour chaque motif, calcul du **support pour chaque type de tumeurs**

	Support Grade 1	Support Grade 2	Support Grade 3
SP1	90%	75%	78%
SP2	84%	60%	38%
SP3	80%	25%	33%

Apprentissage du modèle

- **Etape 2 : Calcul de la « qualité » d'un motif**
- Pour chaque motif, calcul de l'écart entre les 2 plus haut supports

  - **SP3 est le plus discriminant**

	Support Grade 1	Support Grade 2	Support Grade 3	Discriminance
SP1	**90%**	75%	78%	90-78 = **12**
SP2	**84%**	60%	38%	84-79 = **24**
SP3	**80%**	25%	33%	80-33 = **47**

Apprentissage du modèle

- **Résultat**
- Pour chaque grade, une liste des k meilleurs motifs

Grade 1	Grade 2	Grade 3
(G1 G3)	(G4)(G3)	(G5 G6)(G1)
(G5)(G2)	(G5 G6)(G2)	(G2)(G1)
(G2)(G4)(G3)	(G3)(G2)	(G2)(G4)

Classification

- **Entrée : séquence à classer**

S1 = < (G4) (G5 G6) (G3) (G1) (G2) >

- **Etape 1 :** Pour chaque séquence, on teste l'inclusion des motifs

Grade 1	Grade 2	Grade 3
(G1 G3)	(G4)(G3)	(G5 G6)(G1)
(G5)(G2)	(G5 G6)(G2)	(G2)(G1)
(G2)(G4)(G3)	(G3)(G2)	(G2)(G4)
1	3	0

Classification

- **Attribution d'un groupe à une séquence**
  - Au plus fort score

Grade 1	Grade 2	Grade 3
(G1 G3)	(G4)(G3)	(G5 G6)(G1)
(G5)(G2)	(G5 G6)(G2)	(G2)(G1)
(G2)(G4)(G3)	(G3)(G2)	(G2)(G4)
1	3	0

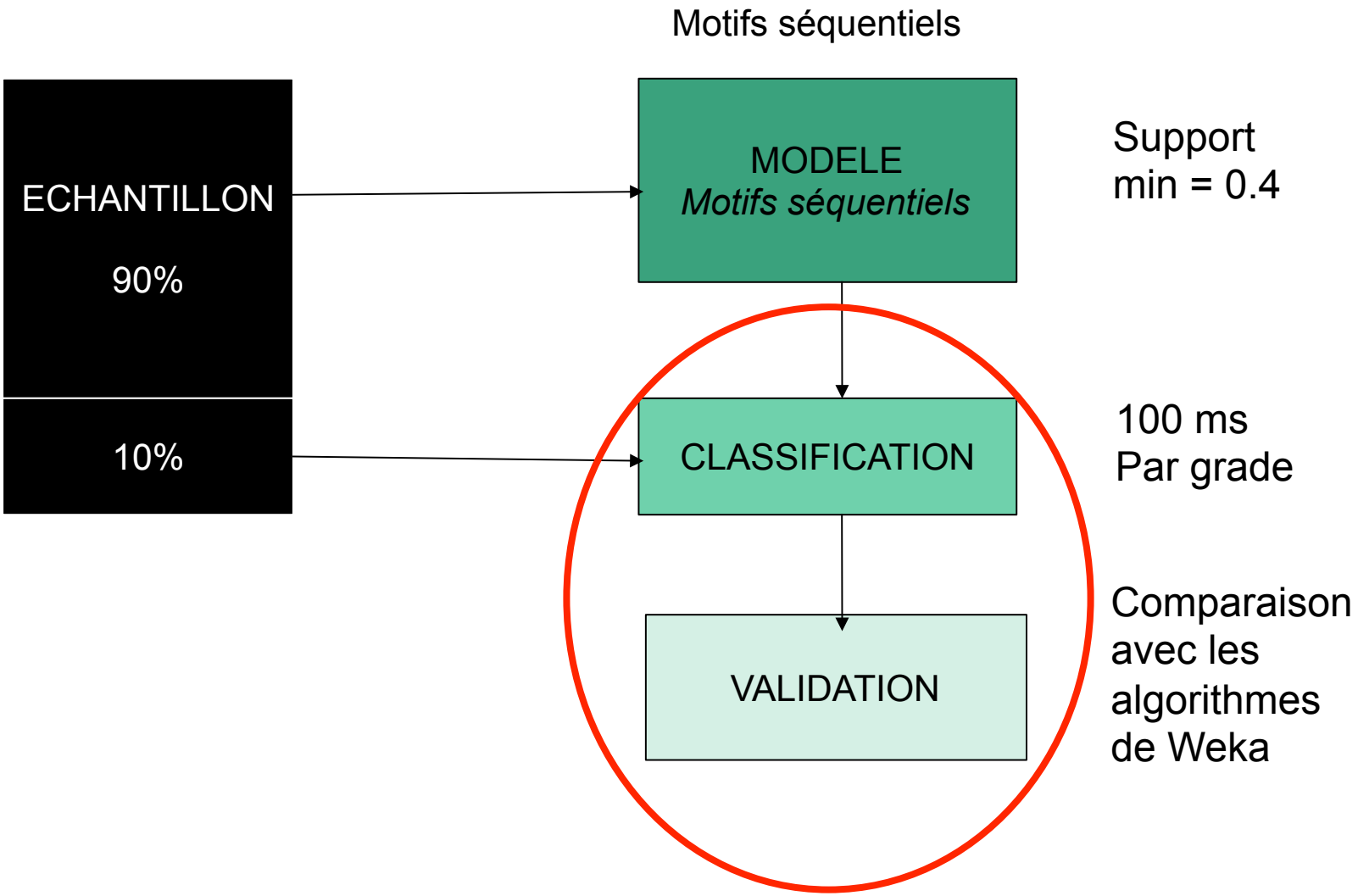
Données

- **Données disponibles en ligne**
  - NCBI http://www.ncbi.nlm.nih.gov
    - KJX64-KJ125
    - TAM
    - TBG
- **Données de l'IRCM**

	Grade 1	Grade 2	Grade 3
Microarrays	162	274	185

- **Sélection 28 genes** Sotiriou and al. (2006)
  - Connus pour leurs implications dans le cancer du sein

Protocole


Motifs séquentiels
ECHANTILLON
90%
10%
MODELE
Motifs séquentiels
CLASSIFICATION
VALIDATION
Support
min = 0.4
100 ms
Par grade
Comparaison
avec les
algorithmes
de Weka

Evaluation

$$Rappel = \frac{Séquences\ correctement\ attribuées}{Séquences\ appartenant\ à\ la\ tumeur}$$

$$Précision = \frac{Séquences\ correctement\ attribuées}{Séquences\ attribuées}$$

Résultats

- **Tous les jeux de données**

	Grade 1	Grade 2	Grade 3
Recall	.81	.15	.86
Precision	.39	.76	.54
F-Measure	.53	.25	.66

- Tumeurs de grade 2 : profils hétérogènes

Résultats

	online datasets			IRCM and online datasets			IRCM		
	R	P	F	R	P	F	R	P	F
Rules (9)	0.924	0.906	0.915	0.922	0.922	0.922	0.954	0.967	0.959
Bayes (8)	0.920	0.918	0.819	0.819	0.826	0.822	0.921	0.932	0.926
Functions (5)	0.917	0.918	0.917	0.902	0.901	0.902	0.955	0.970	0.962
Lazy (4)	0.937	**0.943**	**0.940**	0.922	0.931	0.926	0.909	0.939	0.923
Misc (3)	0.932	0.936	0.934	0.937	0.934	0.936	0.934	0.942	0.937
Tree (12)	**0.938**	**0.943**	0.940	**0.943**	**0.944**	**0.944**	0.928	0.944	0.936
SP	0.891	0.880	0.885	0.896	0.910	0.903	**0.962**	**0.974**	**0.968**

- **IRCM : jeu de données bien réparti et homogène**

Suivre, détecter et prédire l'évolution de la Dengue

Partenaires : INVS Guyane, UNC
Institut Pasteur
Pathologie visée : Dengue

Motifs spatio-temporels

Données dans le temps et l'espace

- Données épidemiologique
- Données environmentale (météo, entomologique...)

Analysis

Evolution de la dengue dans l'espace et le temps

Impact environnemental

Classification

Détection des pics épidémiques

Prediction

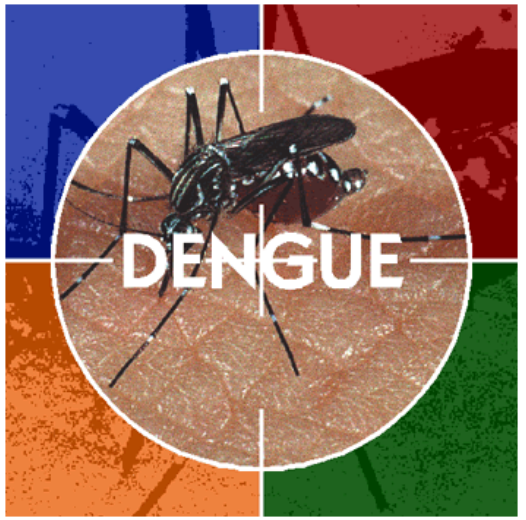
Prédiction de l'évolution des épidémie

Health Professionals

[1] C. Flamand et al. The Epidemiologic Surveillance of Dengue-Fever in French Guyana : When achievements trigger higher goals. 23rd European Medical Informatics conference, MIE2011, Oslo, Norway,, 2011.
[2] H. Alatrista-Salas, S. Bringay, F. Flouvat, N. Selmaoui-Folcher, M. Teisseire: «The Pattern Next Door: Towards Spatio-Sequential Pattern Discovery». In PAKDD'12. Kuala Lumpur, Malaysia, Jun. 2012

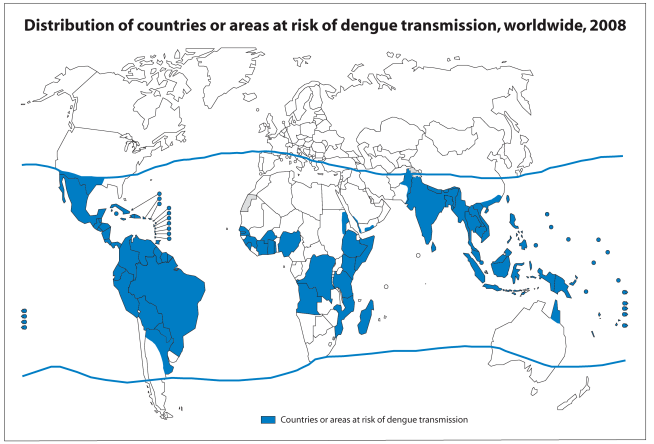
Quelques définitions - épidémiologie

- **Epidémie :** Fréquence plus élevée que celle attendue d'une maladie (TLP)
- **Incidence :** Nombre de nouveaux cas de maladies par unité de temps
- **Seuil épidémique :** Valeur quantitative au-delà de laquelle on considère qu'un nombre de cas est anormalement élevé (situation potentiellement épidémique)
- **Risque :** Probabilité de développer une maladie pendant une période donnée.
- **Alerte :** Evènement sanitaire anormal représentant un risque potentiel pour la santé publique
- **Surveillance épidémiologique :** Observation vigilante et continue de la distribution et des tendances de l'incidence d'une pathologie par la collecte systématique, consolidation et évaluation de données pertinentes



La dengue

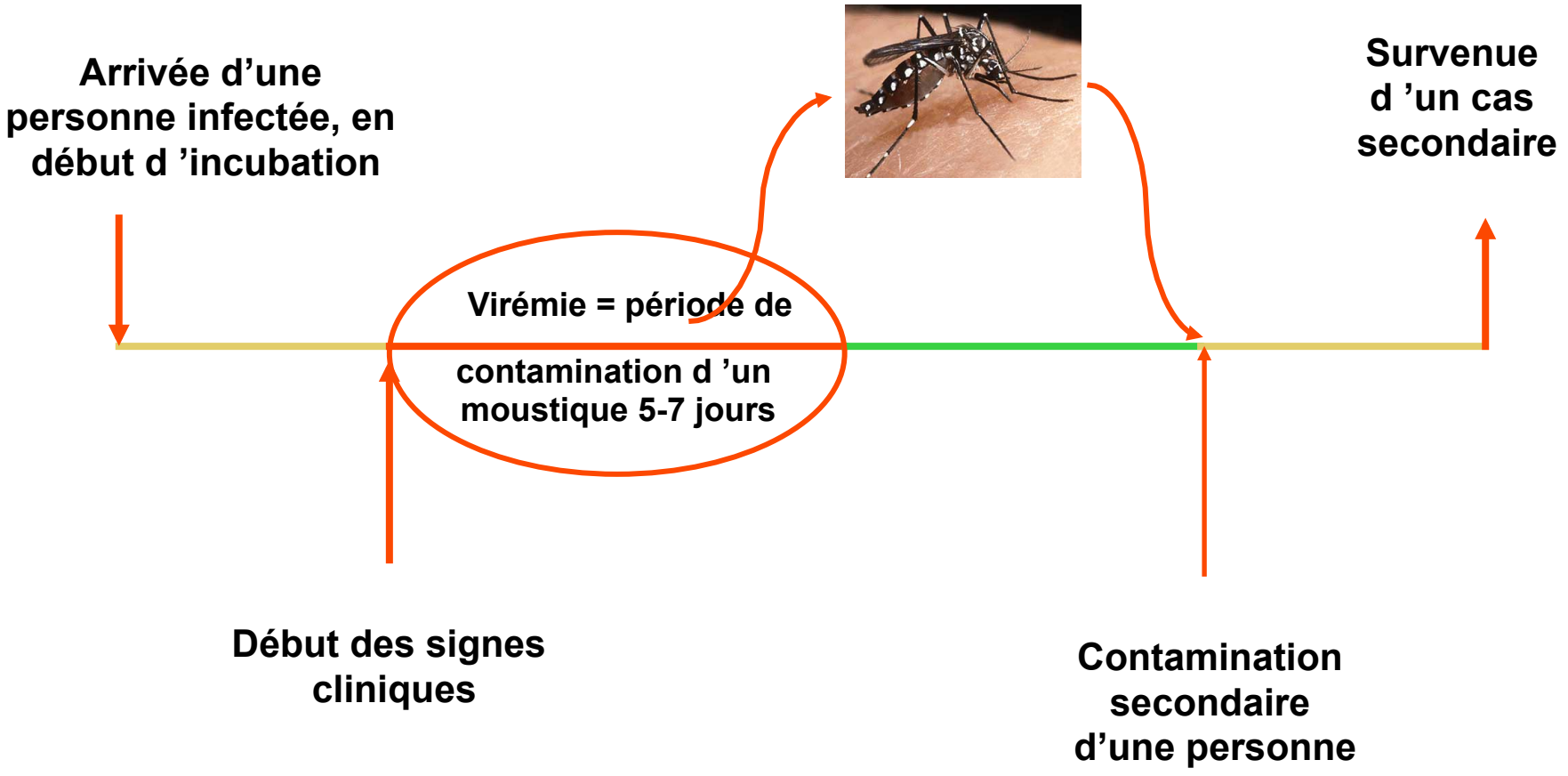
- **Grippe tropicale** transmise à l'homme par le **moustique Aedes aegypti** ou **Aedes albopictus**
- **Arbovirose (**transmise par les suceurs de sang) **n°1 chez l'Homme**
  - 50 à 100 millions de personnes affectées/an, dans le monde
  - 200 000 à 500 000 hospitalisations DFH (Dengue à Fièvre Hémorragique) /an, dans le monde
  - 20 000 décès/an, dans le monde (principalement en Asie du Sud Est)
- **4 sérotypes** (DENV-1, DENV-2, DENV-3,DENV-4)
- **Endémique** (persistente) **dans les pays tropicaux.**



Data Source: World Health Organization
Map Production: Control of Neglected Tropical Diseases (NTD)
World Health Organization

World Health Organization

Cycle de transmission

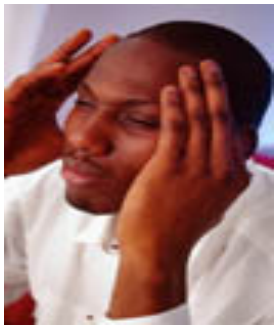


- Pas de transmission interhumaine, ni par l'air ou par l'eau.

2 variantes

- **Forme classique**

- Manifestation brutale après 5 à 7 jours d'incubation;
- Symptômes : Forte fièvre, maux de tête, Frissons, Nausées, Vomissements, Douleurs articulaires et musculaires, Rash cutané (apparition de tâches rouges)
- Évolution : Amélioration au bout d'une semaine et Rémission en quelques semaines.



- **Forme hémorragique**

- Débute sous la forme classique, Rémission passagère, Aggravation brusque vers le 3ème ou 4ème jour
- Symptômes : Hémorragies digestives, nasales….
- Évolution : Guérison rapide avec asthénie (affaiblissement de l'organisme) pendant 3 à 6 mois, aggravation, état de choc et parfois décès.

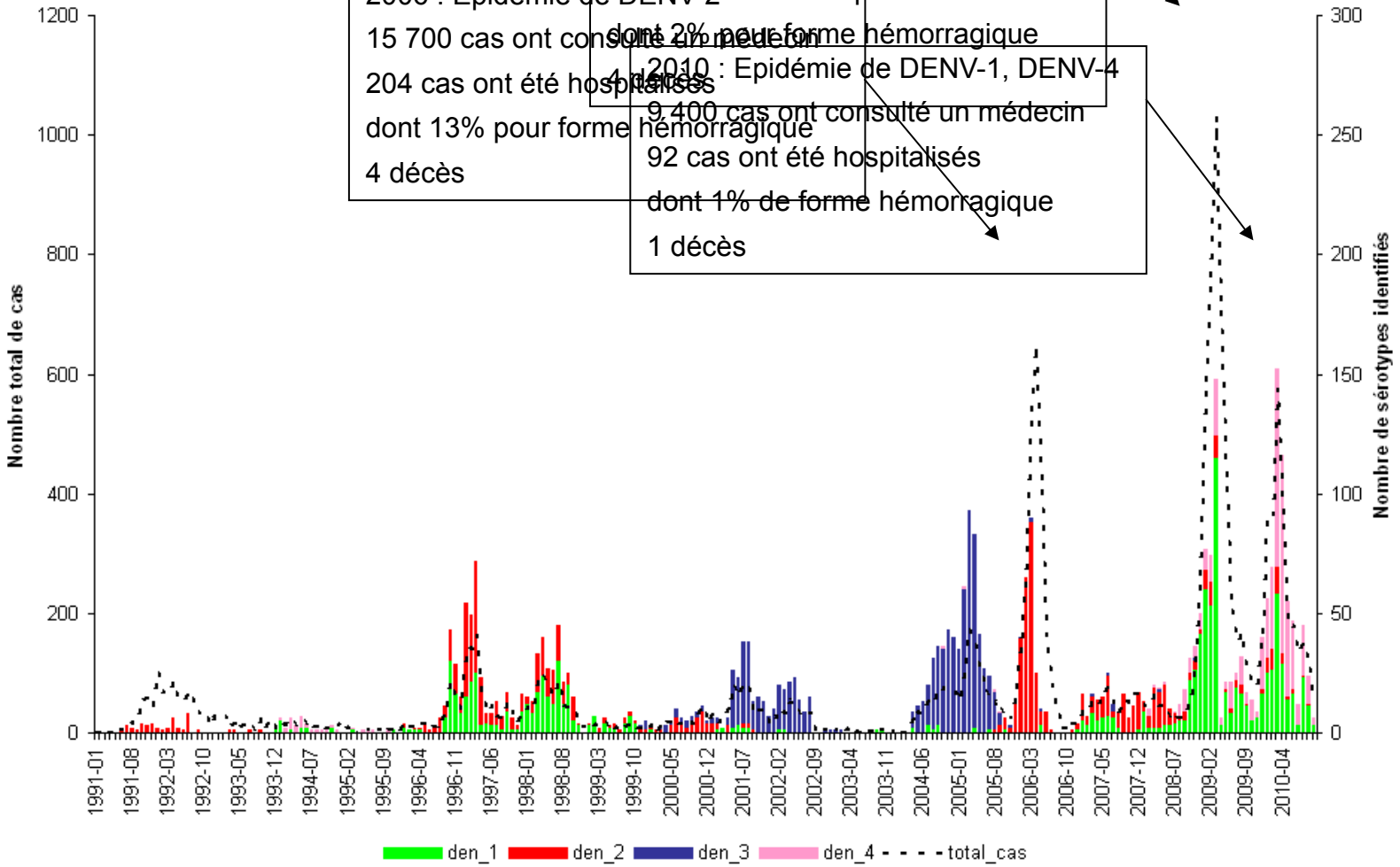
La dengue en Guyane

- **Circulation des 4 sérotypes**
- **Mode endémo-épidémique en Guyane**

2006 : Epidémie de DENV-2
15 700 cas ont consulté un médecin
204 cas ont été hospitalisés
dont 13% pour forme hémorragique
4 décès

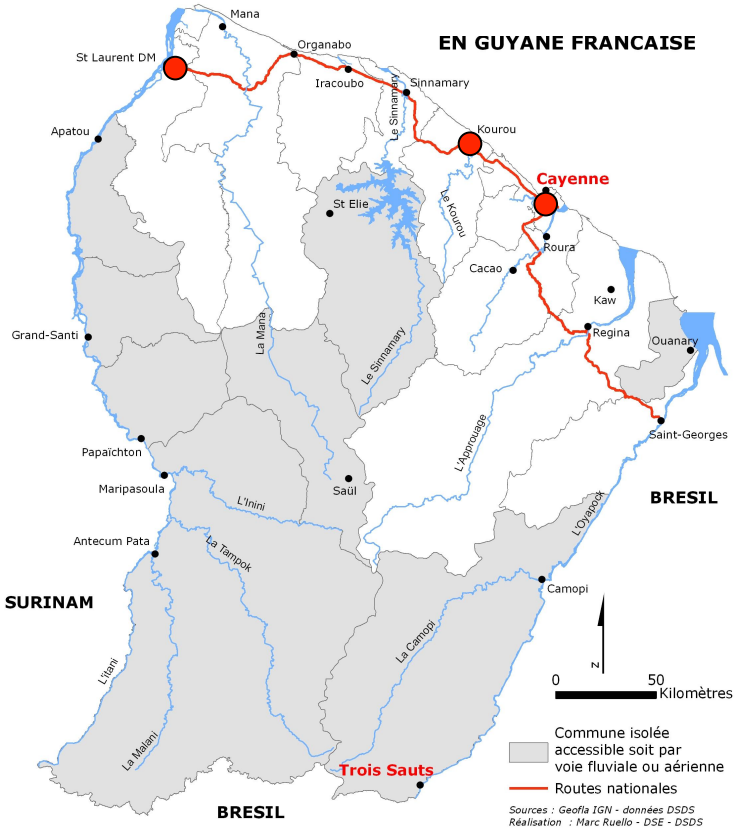
2009 : Epidémie de DENV-1
13 900 cas ont consulté un médecin
241 cas ont été hospitalisés
dont 2% pour forme hémorragique
4 décès

2010 : Epidémie de DENV-1, DENV-4
9 400 cas ont consulté un médecin
92 cas ont été hospitalisés
dont 1% de forme hémorragique
1 décès



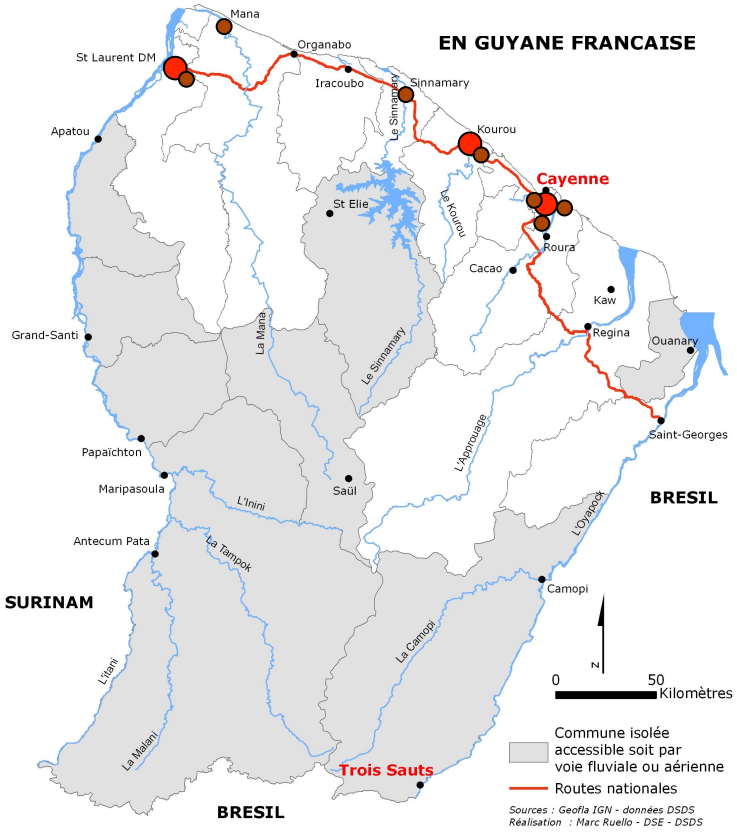
Sources des données

- Laboratoires : 7 laboratoires répartis sur 3 communes du Littoral (Cas biologiquement confirmés)



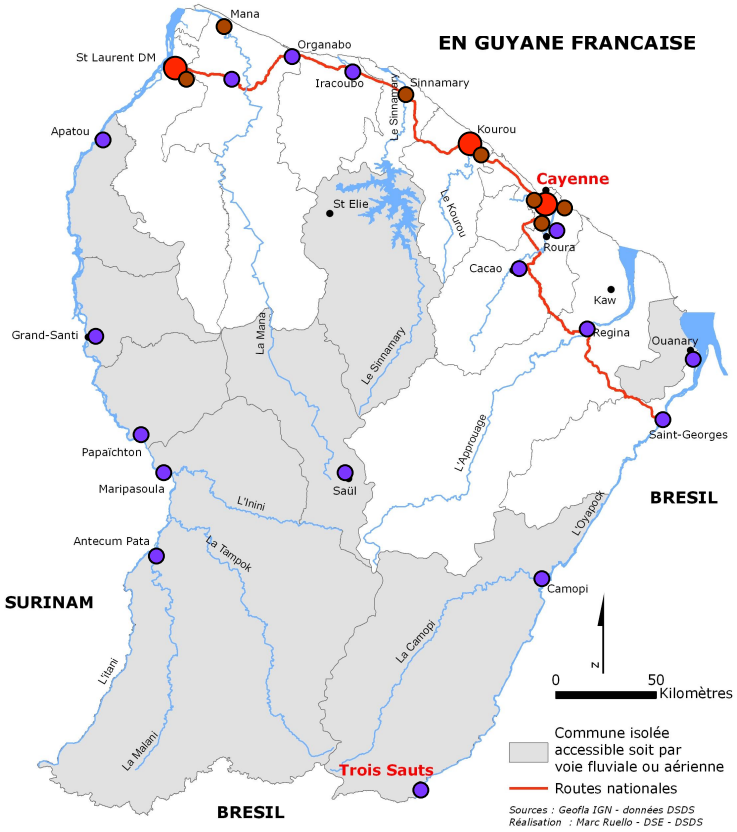
Sources des données

- Médecins sentinelles : 25 médecins généralistes (33%) sur 7 com. du littoral (cas cliniquement évocateurs)

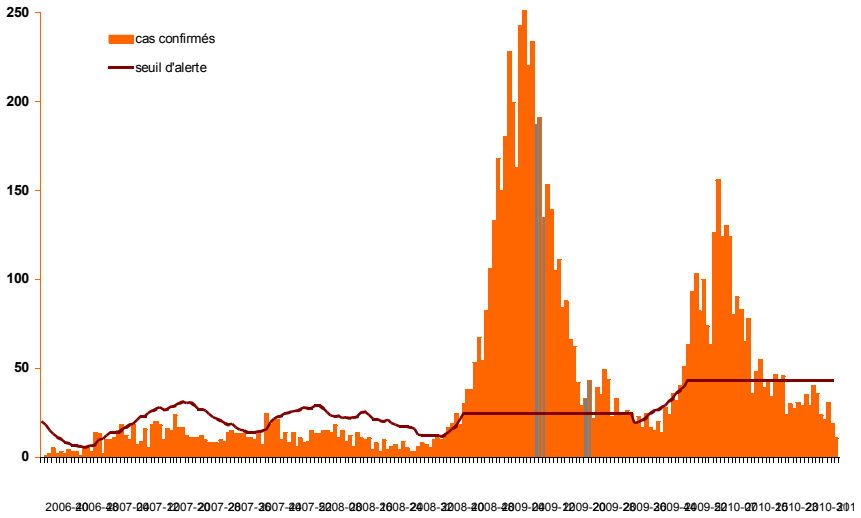


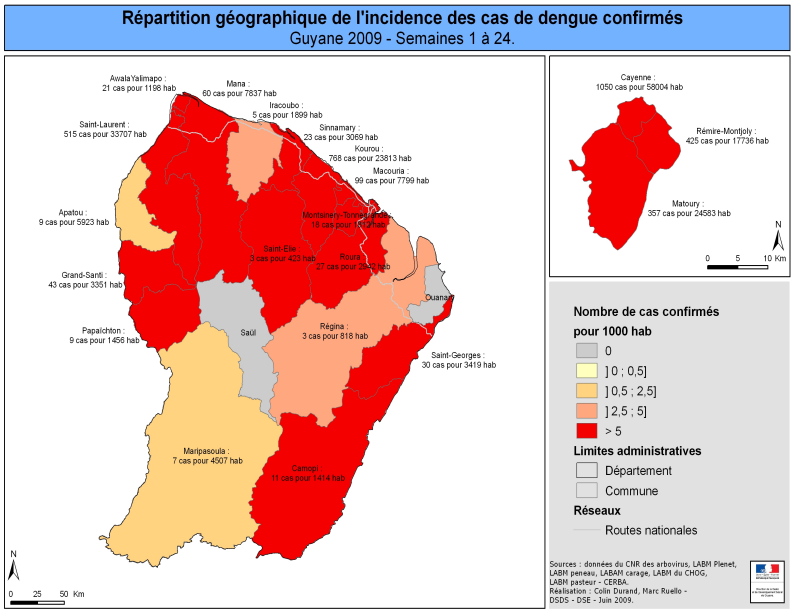
Sources des données

- 17 Centres et Postes de santé (cas cliniquement évocateurs)

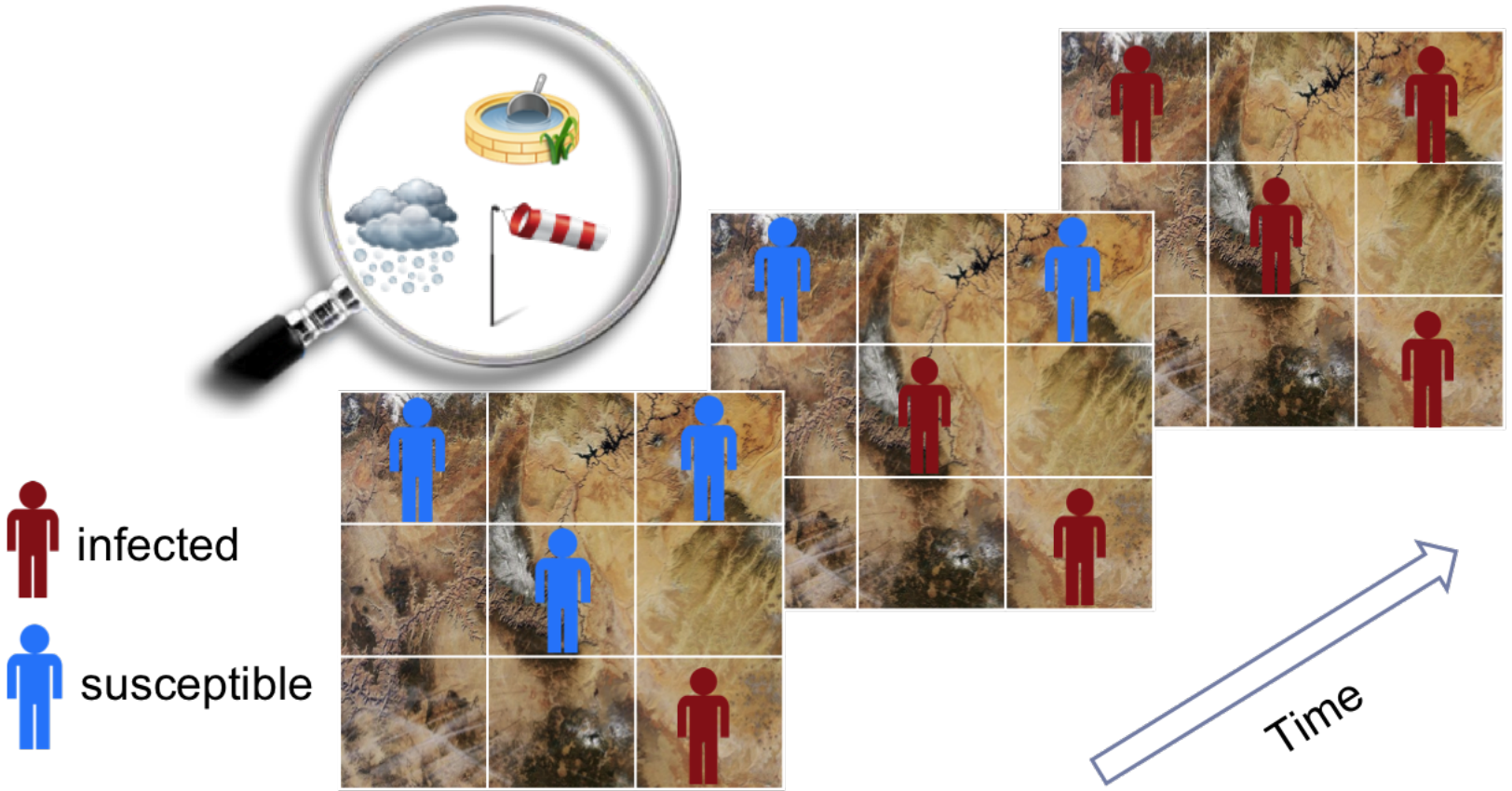


Surveillance à l'échelle de la commune





Notre problématique



Base de données

area	date	temperature	rain	wind (km/h)
Port Klang	1	low	low	35
Port Klang	2	low	high	-
Kelang	1	low	mean	12
Kelang	2	high	mean	-
Kapar	1	low	-	-
Kapar	2	low	low	23

- Spatial dimension
- Temporal dimension (timestamps)
- Analysis dimension

Base de séquences spatio-temporelles

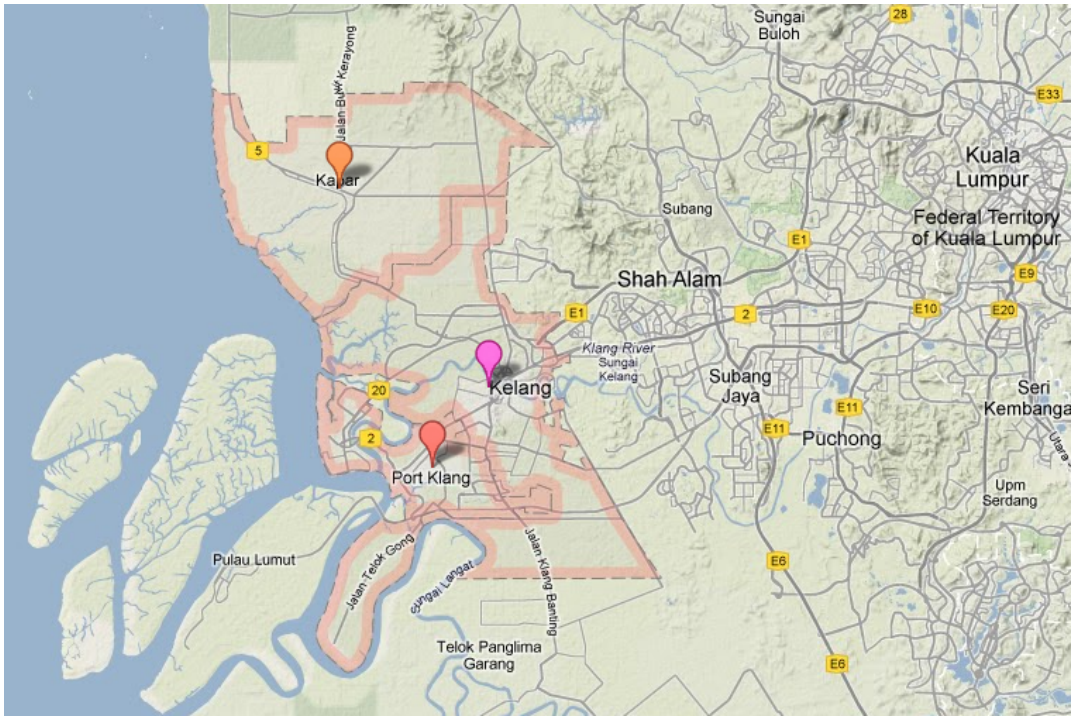
On groupe les dimensions d'analyse par date et zone pour obtenir une **base de séquences sDB**

area	sequence of itemsets
Port Klang	(temp_low rain_low wind_35) (temp_low rain_high wind_-)
Kelang	(temp_low rain_mean wind_12) (temp_high rain_mean wind_-)
Kapar	(temp_low rain_- wind_-) (temp_low rain_low wind_23)

itemset

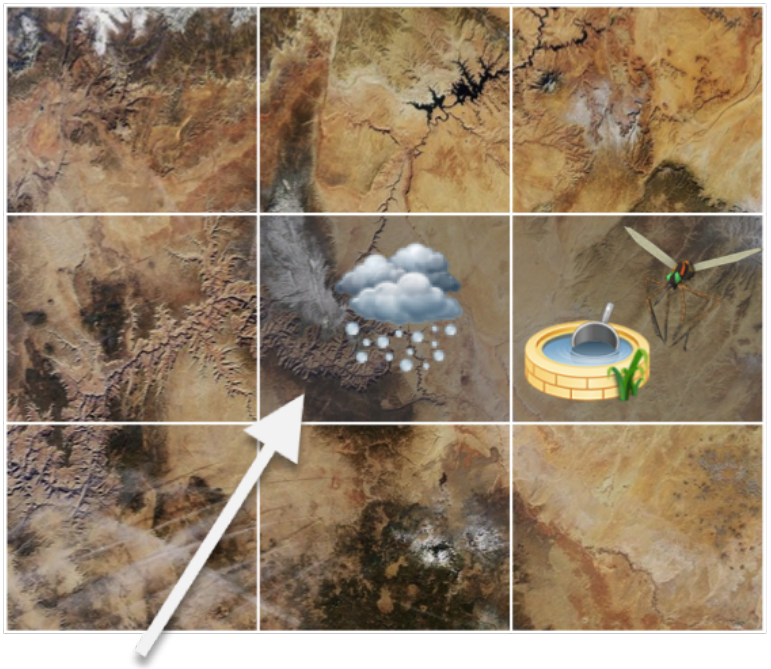
Zonage

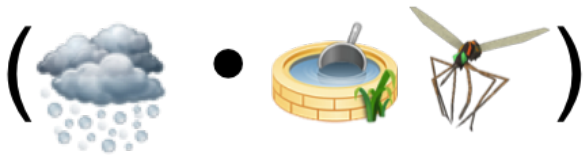
area 1	area 2
Port Klang Kelang	Kelang Kapar



Itemset spatial

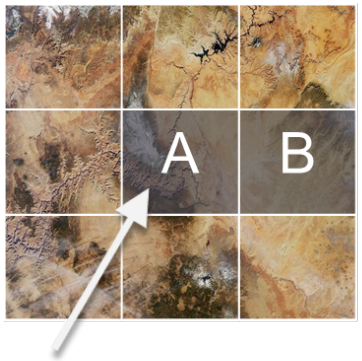
2 items spatialement proches



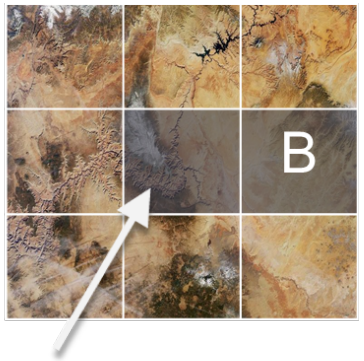


Dynamique spatiale

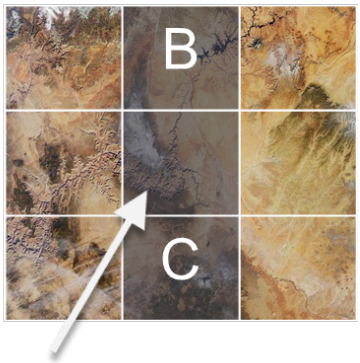
Opérateur spatial • (près de)
Opérateur de groupage [ ]
Symbole representant l'absence d'itemsets θ



(A • B)



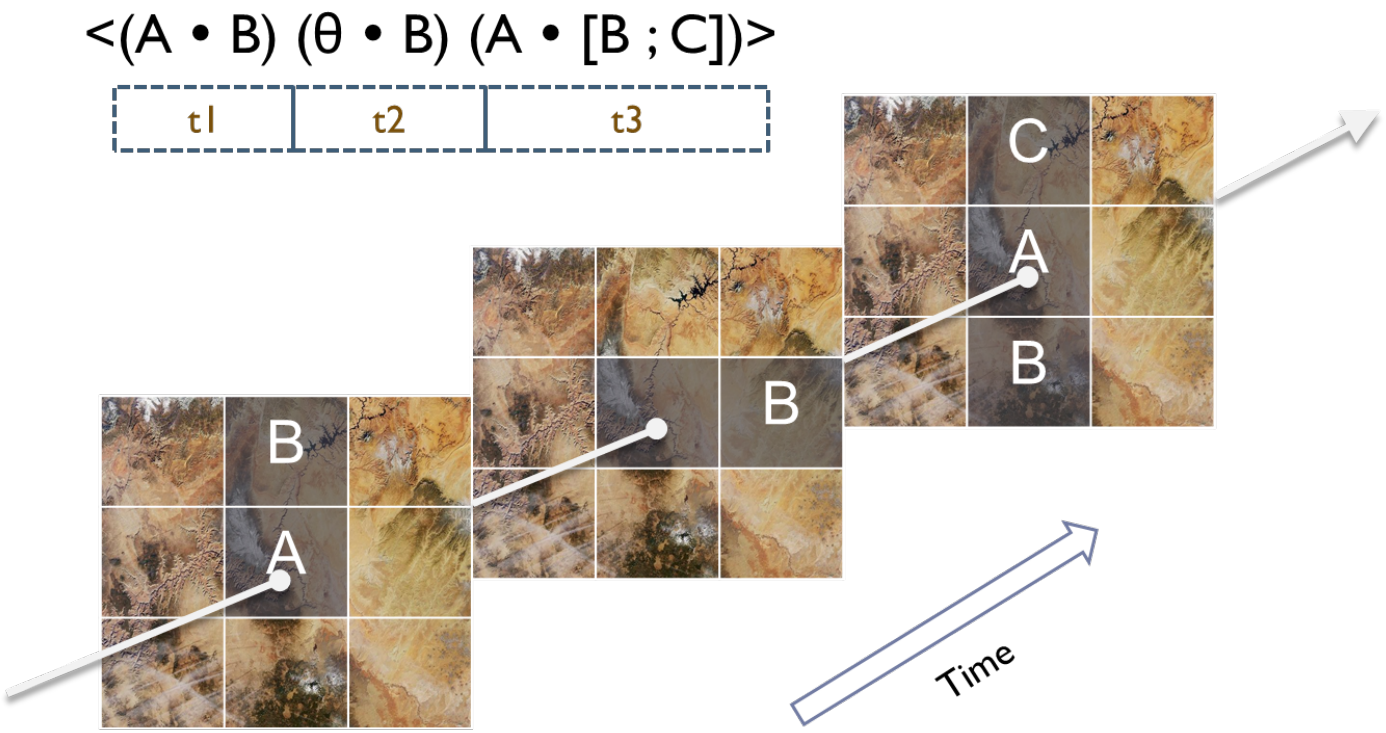
(θ• B)



(θ • [B ; C])

Spatial Sequences 2S

Séquence d'itemsets spacials



Fouille de séquence 2S

Extraire un ensemble de 2S fréquentes dans la base de séquences sDB :

$$support(2S) >= \alpha$$

Avec α le support minimum

Une 2S fréquente est également appelée motif spatio-temporel ou S2P

Support

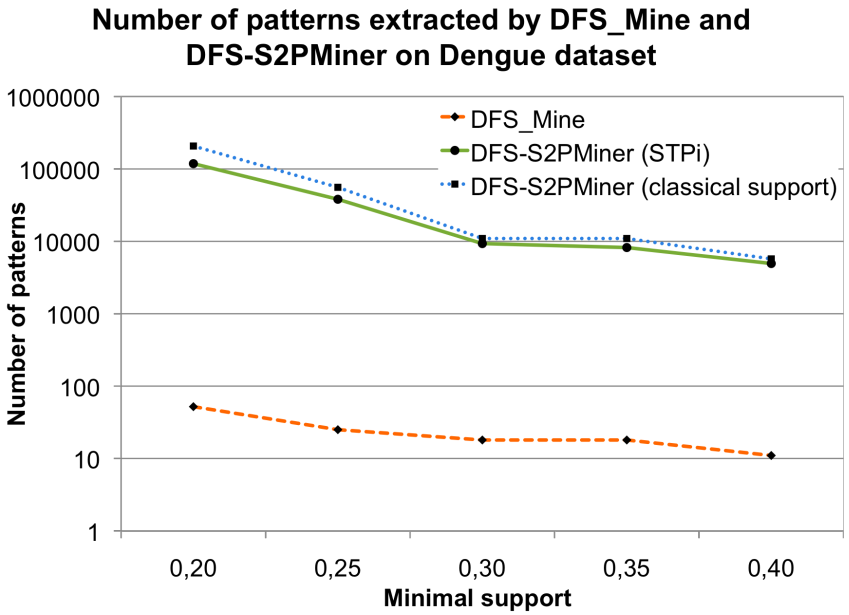
area	sequence of itemsets
Port Klang	<(temp_low rain_low wind_35) (temp_low rain_high wind_-)>
Kelang	<(temp_low rain_mean wind_12) (temp_high rain_mean wind_-)>
Kapar	<(temp_low rain_- wind_-) (temp_low rain_low wind_23)>

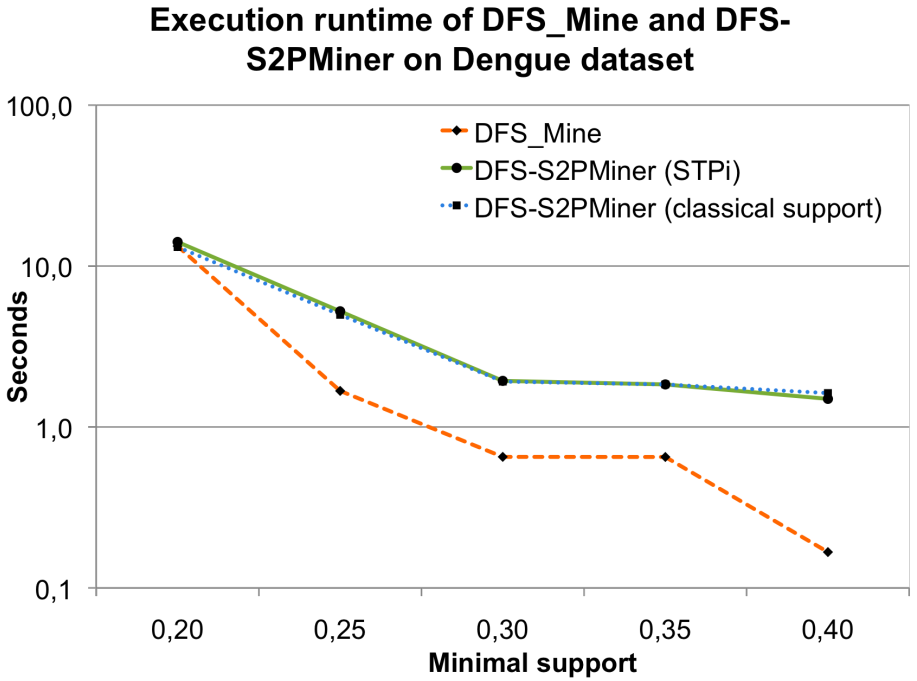
support (<(rain_low rain_high)>) $= \frac{2}{3}$, the sequence appears for 66% of areas

support (<(temp_low)>) $= 1$, the sequence appears in all areas

! *The support uses only the spatial component of dataset*

Evaluation


Number of patterns extracted by DFS_Mine and DFS-S2PMiner on Dengue dataset
1000000
100000
10000
1000
100
10
1
Number of patterns
DFS_Mine
DFS-S2PMiner (STPi)
DFS-S2PMiner (classical support)
0,20
0,25
0,30
0,35
0,40
Minimal support


Execution runtime of DFS_Mine and DFS-S2PMiner on Dengue dataset
100,0
10,0
1,0
0,1
Seconds
DFS_Mine
DFS-S2PMiner (STPi)
DFS-S2PMiner (classical support)
0,20
0,25
0,30
0,35
0,40
Minimal support

… mais encore…

HIV1 vs. HIV2

Partners : Institut de Génétique Moléculaire de Montpellier

Targeted pathologies: HIV

Discriminant Sequential Patterns

- DNA microarrays
- Experiments **over time**

Analysis

- HIV1 and HIV2 characteristics

Classification

- Identification of targeted genes

Biologists

Rare Cells

Partners : Zenith, I3M

Objectives : extract 10 cells over billions of cells

➲Flow Cytrometry

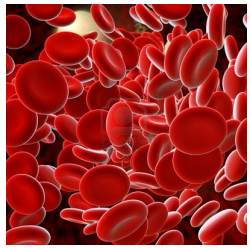
Clustering of rare cells

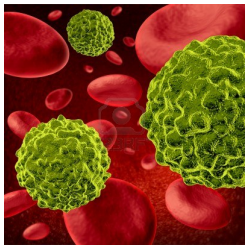
Analysis

➲Cluster of different kinds of cells

Classification

➲ Signature of rare cells (cancer, CVA, …)





Health professionals

Projet VIPP : Vers une identification précoce des pathologies : découverte de critères pour la caractérisation de cellules rares dans le sang
Financement via le Labex NUMEV

Sensors Data Analysis

Partners : CHU Martinique, ANR MIDAS

Objectives : Stream Summarization

➲Streams of Sensors Data

Multi-dimensional Patterns

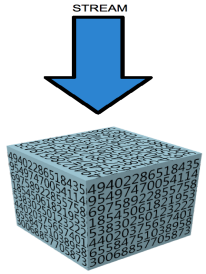
Analysis

- ➲Alarm Triggering
- ➲Past Data Requests
- ➲Contextual Hierarchies

Classification

- ➲ Class attribution according to a context

Health professionals



[1] Y. Pitarch et al. Context-Aware Generalization for Cube Measures. 13th ACM International DOLAP 2010, Toronto, Canada, October 2010, pp. 99-104.
[2] Y. Pitarch, et al. Summarizing Multidimensional Data Streams: A Hierarhy-Graph-Based Approach. 14th Pacific-Asia Conference on Knowledge Discovery and Mining (PAKDD 2010), Hyderabad, India, June 2010, pp. 335-342.

Follow a H1N1 Epidemic in News

Partners : Nevantropic
Objectives : H1N1 propagation

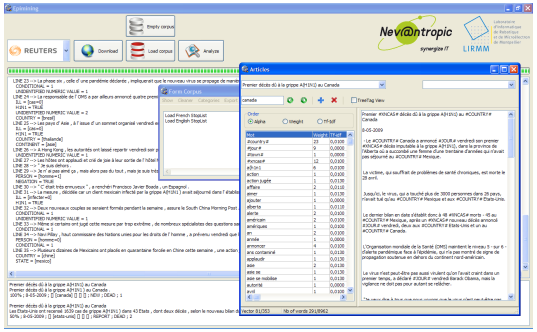
Textual Patterns

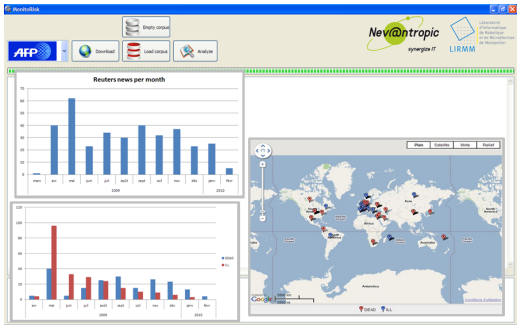
➲News (Reuters, AFP,…)

Classification

➲New cases, bilans

Health professionals





[1] D. Breton et al. Analyse de dépêches pour l'épidémiologie. 21ièmes Journées Francophones d'Ingénierie des Connaissances (IC 2010), Nîmes, France, Juin 2010.

Trends of Health News on Tweets

Objectives : Trend Detections on Social Networks and Tweets

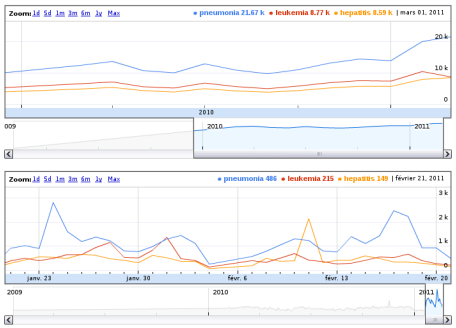
- Stream of Tweets
- MeSH Hierarchy

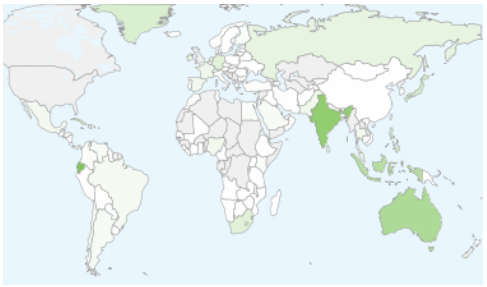
Text Cubes

Analysis

- Symptoms detection
- Trends Analysis
- Sentiments Analysis

Health professionals





[1] S. Bringay et al.. Towards an On-Line Analysis of Tweets Processing. 22nd International Conference DEXA 2011, Toulouse, France, August 2011.

Detection of Natural Disasters

Partners: WebReport

- Tweets
- Class of disasters (tornado, flood, earthquake...)

Textual Patterns

Analysis

- Description of tweets

Classification

- Association of a news to a category of disasters

Journalists

[1] B. Rosoor et al. A la recherche des tweets porteurs d'informations journalistiques. Démonstration présentée à la conférence EGC 2011, Brest (France), 25 au 28 janvier 2011, 283-286, 2011.
[2] B. Rosoor et al. Quand un tweet détecte une catastrophe naturelle... VSST'2010, Toulouse (France), 2010.