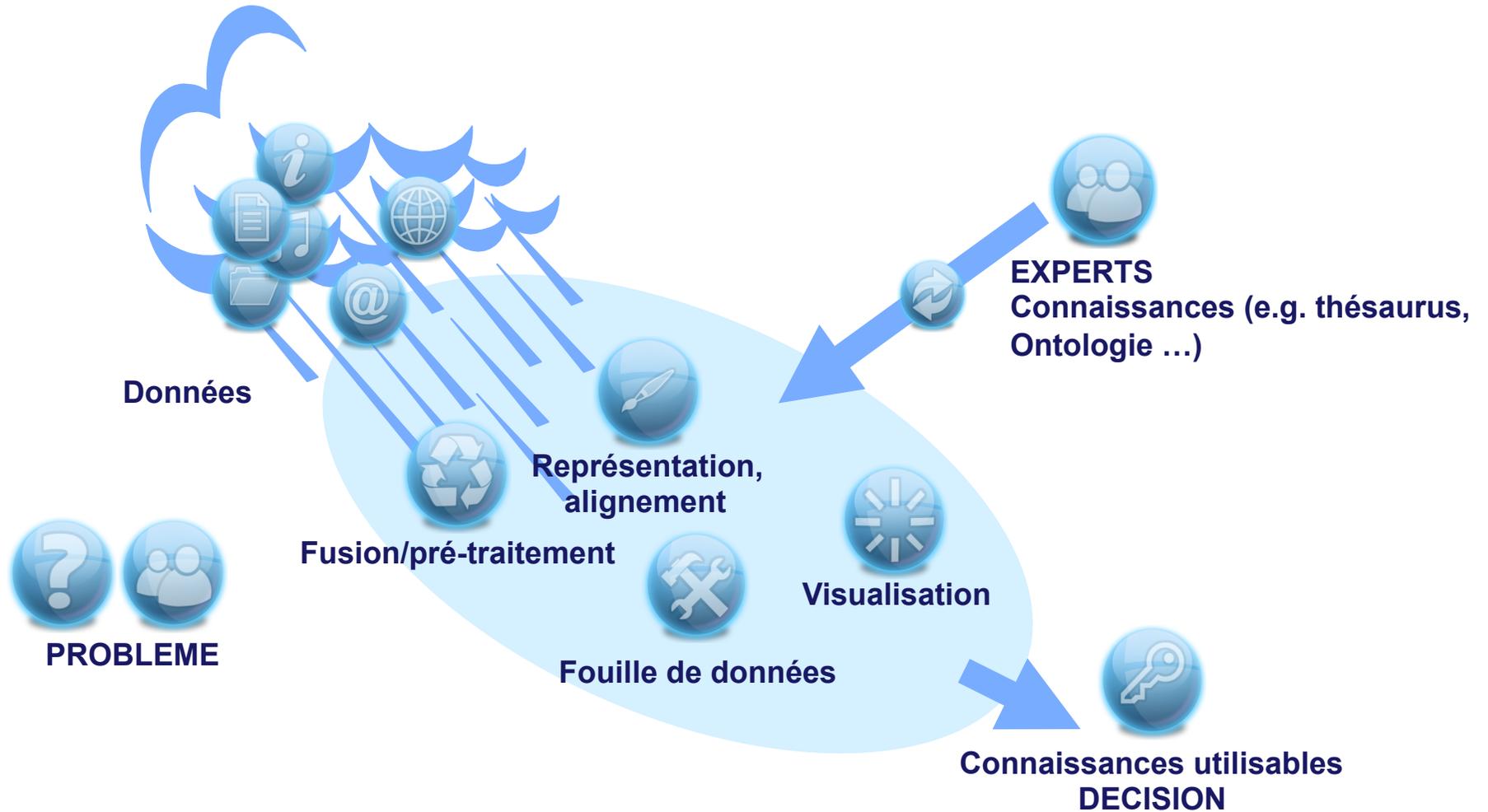


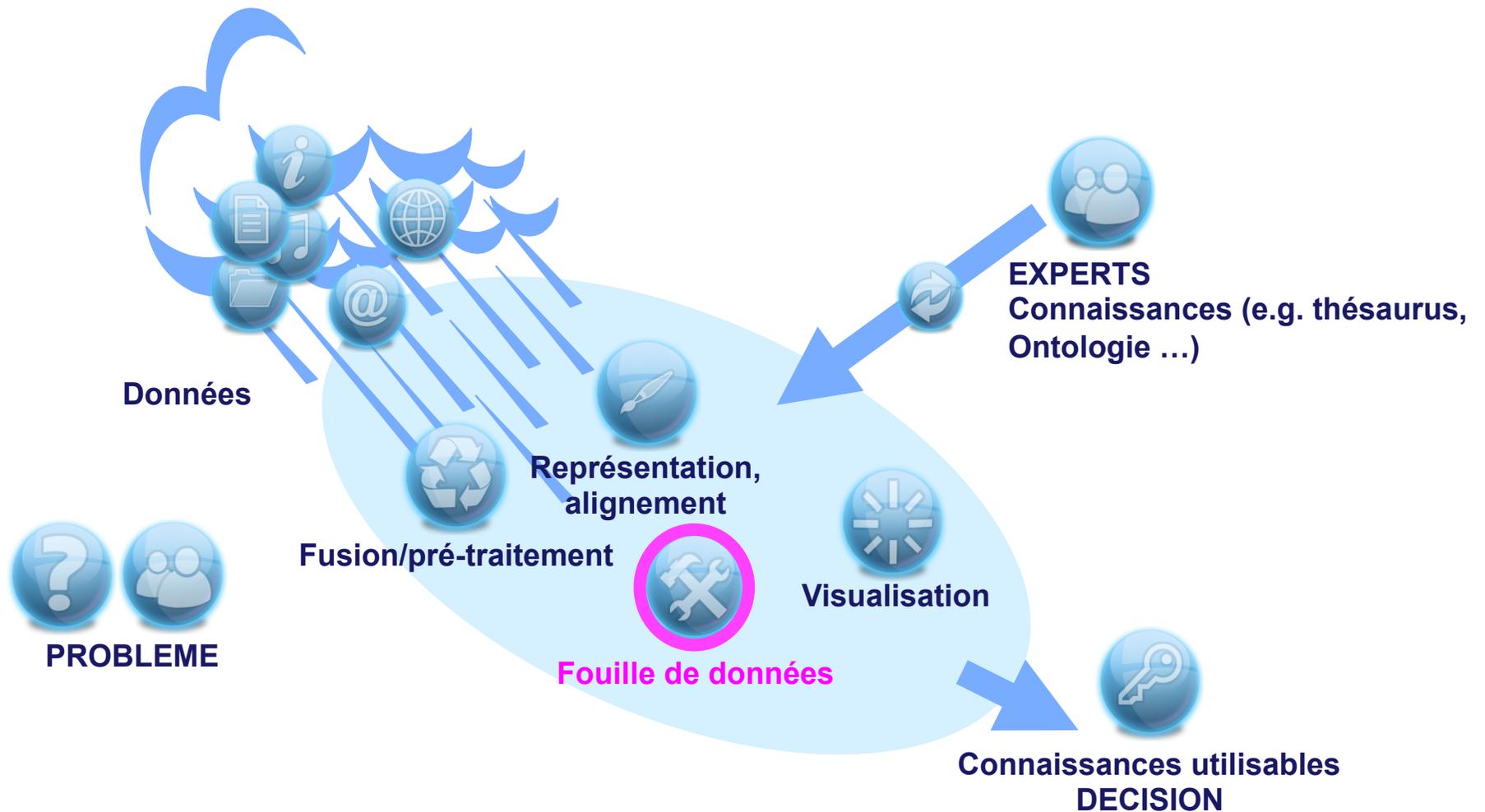
Fouille de données et Santé

sandra.bringay@univ-montp3.fr

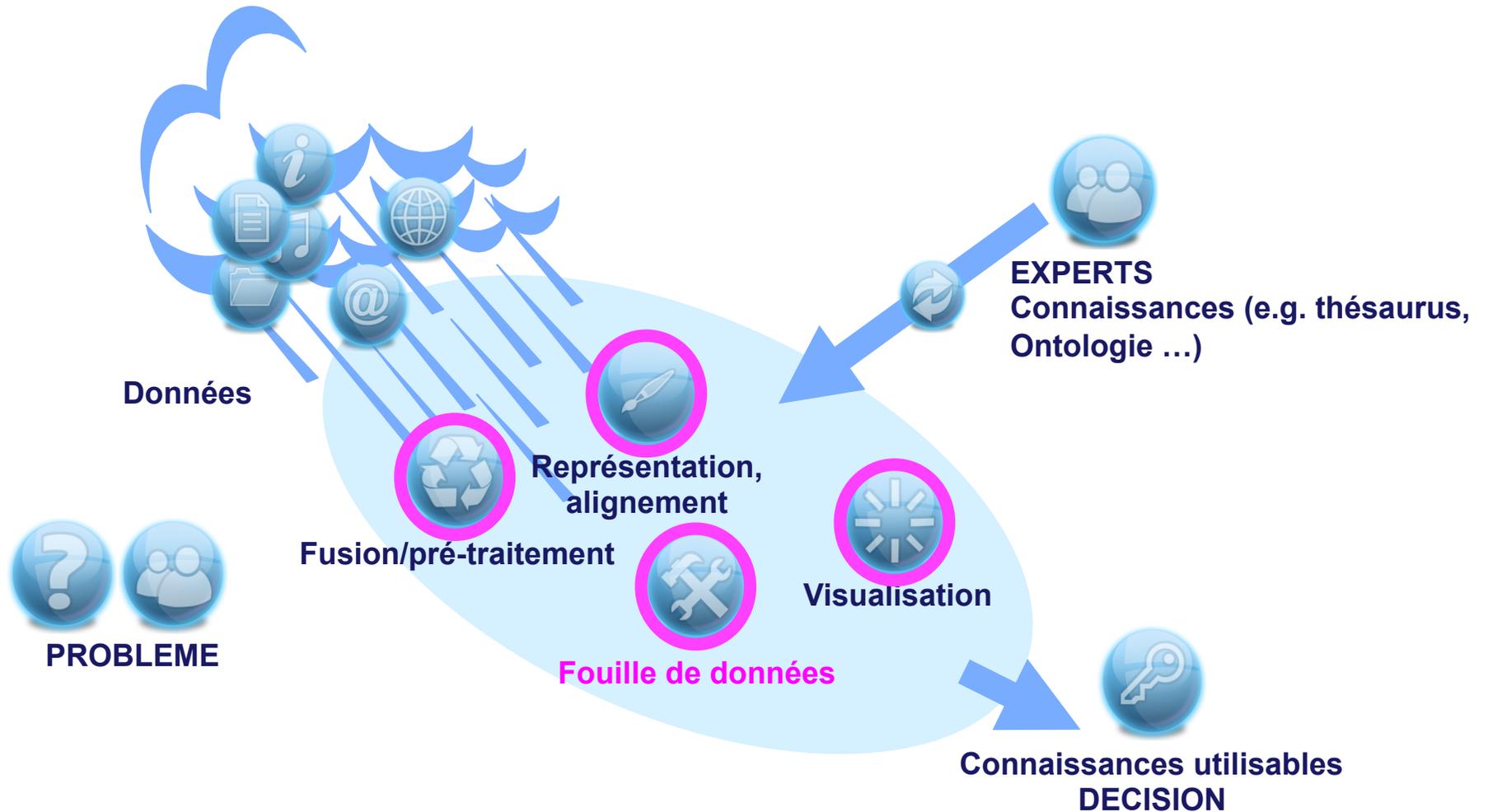
Processus d'extraction de connaissances



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



- ➔ Booléenne
- ➔ Numérique
- ➔ 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

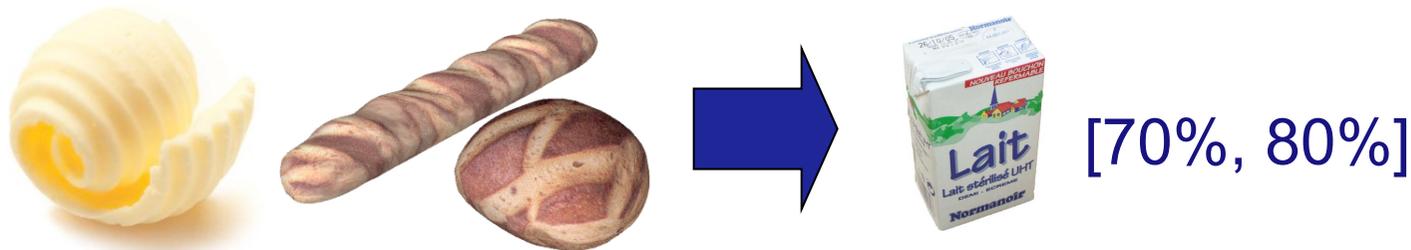


- ➔ Motifs inattendus, discriminants...
- ➔ Règles graduelles
- ➔ Trajectoires
- ➔

Règles d'association

➔ Problème du “panier de la ménagère”

ANTECEDENT → CONSEQUENT [Support, Confiance]



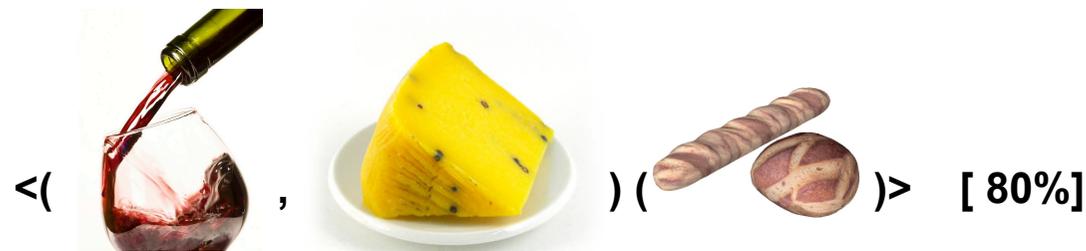
Support : 70% des clients ont acheté du beurre et du pain et du lait

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

Motifs séquentiels

⇒ Prise en compte du temps

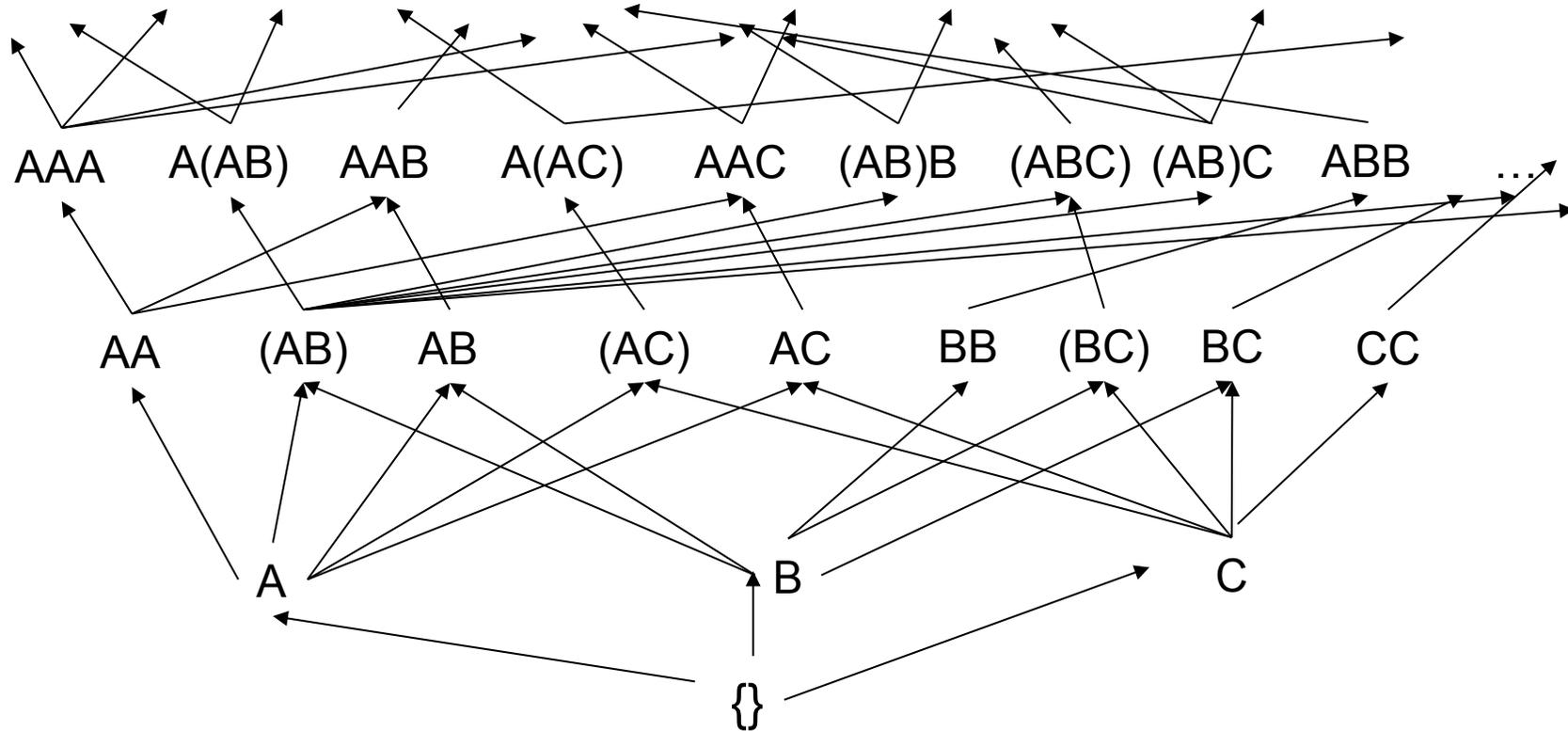
SEQUENCE OF ITEMSETS [Support]



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

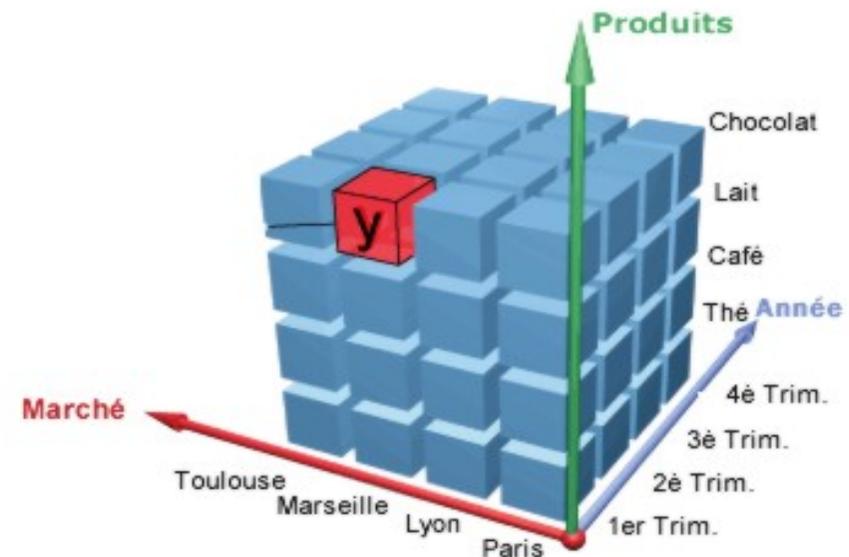


Espace de recherche

Motifs multi-dimensionnels

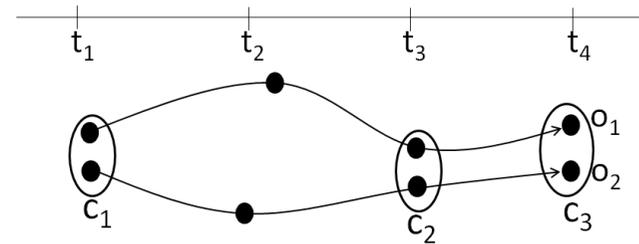
➔ Les items sont définis selon différentes dimensions

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

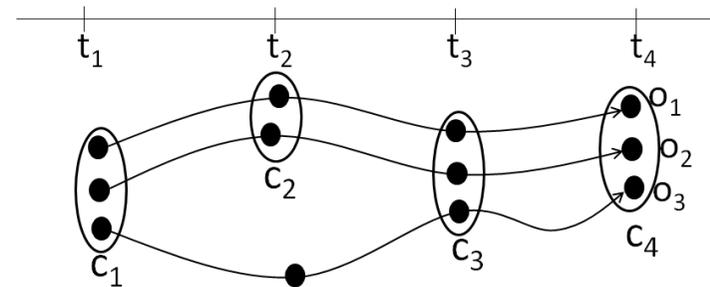


Trajectoires

➔ **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, inattendus...

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 numériques)

Motifs
séquentiels

Analyse

Identification de
signatures et de
nouveau

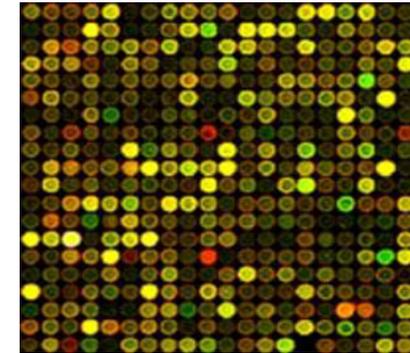


Biologistes

- [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

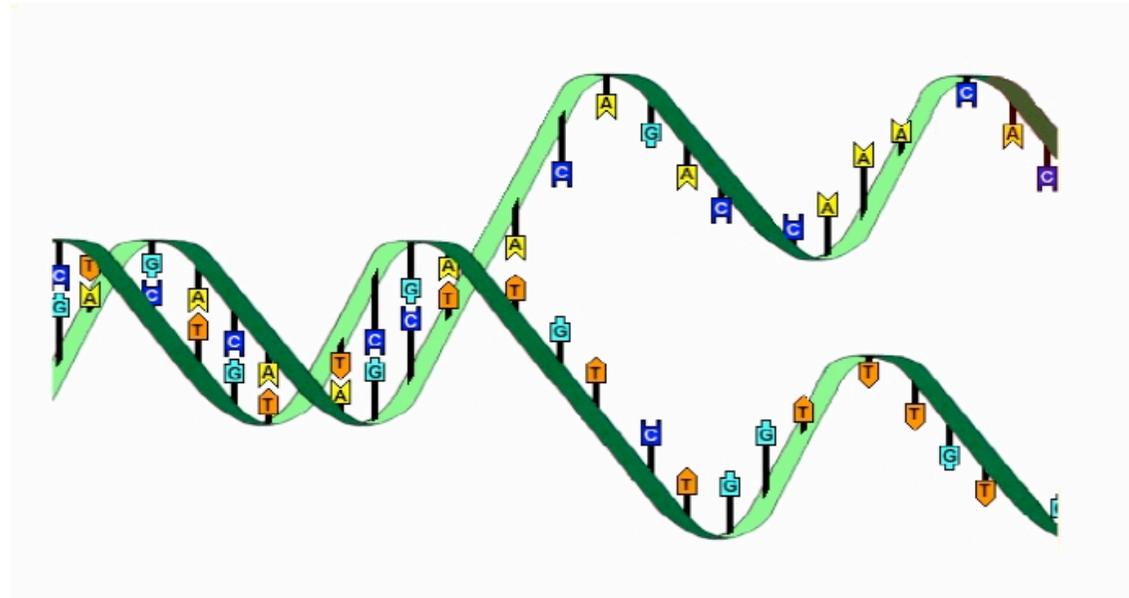
- **Mieux comprendre les maladies génétiques :**
 - perturbation des processus naturels de croissance, de division et de mort des cellules
- **Mesurer l'expression des gènes et identifier les lois suivies par ces expressions** en fonction des maladies et des traitements...
 - gènes impliqués dans la 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).

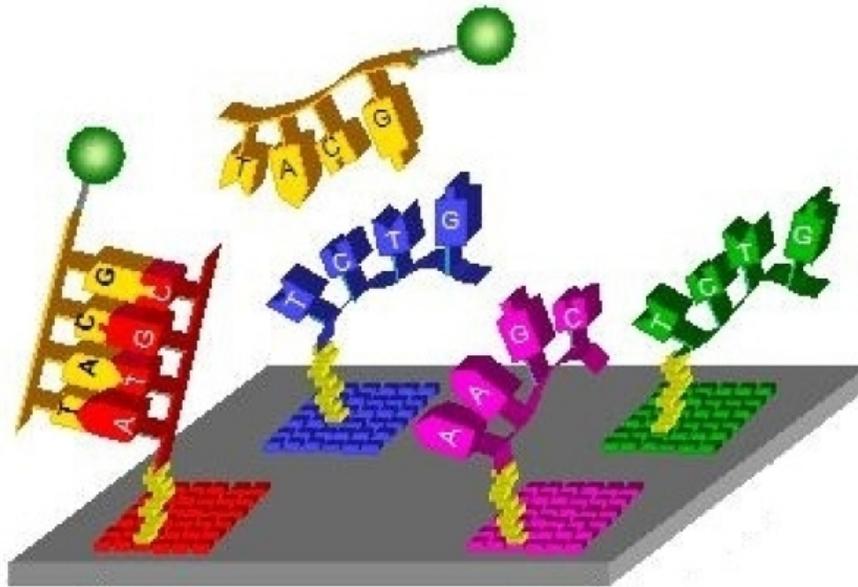
A ≡ T
T ≡ A
G ≡ C
C ≡ G



(a)

Données : puces à ADN

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



(b)

Expression (couleur) \approx
mesure de la quantité
d'ADN dénaturé qui se
reformé

Microsoft Excel - Données

File Edit View Insert Format Tools Data Help

Formulas

Home

Formulas

Formulas

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1																								
2		10150	319083	-0.21	0.90	-0.95	-0.2	1.05	-0.52	-1.25	-0.31	-0.94	0.22	-1.09	0.8	0.22	-0.95	-0.95	0.54	-0.78				
3		3078	300803	-0.67	1.1	-0.71	-1.23	1.45	-1.37	-0.88	-0.25	0.25	0.22	-0.57	0.62	0.52	0.44	-0.91	0.52	0.09				
4		628	301174	0.22	0.2	-0.27	-0.73	-0.45	-0.89	-0.83	-0.36	-0.5	-0.22	0.52	0.86	-0.41	-0.52	-1.01	-1.01	-1.47				
5		75258	319623	-0.28	0.62	-1.2	0.2	1.02	-0.2	-0.24	0.76	0.38	0.05	0.76	0.74	0.05	-0.09	0.66	0.29	-0.01				
6		34885	319628	0.12	-0.48	-0.58	0.08	-0.23	-0.37	1.01	0.85	-1.56	1.05	1.23	0.54	2.22	1.81	0.71	0.71	0.74	0.61			
7		1298	300878	0.6	0.72	-0.71	0.28	-1.73	-0.88	-0.38	0.36	1.47	-0.6	-0.71	-1.09	1.44	-0.35	-0.01	-0.16	-0.1				
8		77571	49523	-0.11	1.04	-1.92	0.48	0.74	-0.24	-0.87	-0.26	0.25	-0.34	0.45	0.77	-0.21	-0.14	-0.52	1.22	-0.17				
9		4202	305824	-0.62	0.82	-1.02	-0.22	-0.22	-0.78	-0.55	-1.08	-0.71	-2.42	1.81	0.86	0.22	-1.77	-0.15	1.01	-0.05				
10		13825	319828	-1.22	0.62	-1.22	0.58	1.22	-0.7	-0.8	0.05	0.26	-0.17	-0.82	1.06	-0.01	-1.25	-0.25	0.69	-0.59				
11		1298	300878	0.6	0.72	-0.71	0.28	-1.73	-0.88	-0.38	0.36	1.47	-0.6	-0.71	-1.09	1.44	-0.35	-0.01	-0.16	-0.1				
12		1298	300878	0.6	0.72	-0.71	0.28	-1.73	-0.88	-0.38	0.36	1.47	-0.6	-0.71	-1.09	1.44	-0.35	-0.01	-0.16	-0.1				
13		1298	300878	0.6	0.72	-0.71	0.28	-1.73	-0.88	-0.38	0.36	1.47	-0.6	-0.71	-1.09	1.44	-0.35	-0.01	-0.16	-0.1				
14		34273	319423	0.28	0.38	-0.18	-0.24	-1.32	-0.38	-1.75	-0.81	0.71	0.04	0.81	-0.41	0.81	1.01	0.31	-0.38	-1.45				
15		34273	319423	0.28	0.38	-0.18	-0.24	-1.32	-0.38	-1.75	-0.81	0.71	0.04	0.81	-0.41	0.81	1.01	0.31	-0.38	-1.45				
16		34273	319423	0.28	0.38	-0.18	-0.24	-1.32	-0.38	-1.75	-0.81	0.71	0.04	0.81	-0.41	0.81	1.01	0.31	-0.38	-1.45				
17		75653	319823	-1.54	1.7	-1.05	-0.81	1.25	-0.4	-0.53	-0.42	-0.47	1.87	-1.3	2.49	0.51	-1.61	-0.61	-2.72	0.1				
18		34388	319823	-0.28	0.77	-0.94	-1.23	0.35	-0.34	-0.28	-0.88	-0.89	0.34	0.76	1.06	-0.41	-1.96	-0.5	1.23	-0.45				
19		1315	301817	-0.28	0.14	-1.14	0.12	0.3	-0.23	-0.23	0.52	-0.83	0.48	-0.86	-0.46	-0.25	-0.04	-0.35	-0.22	0.25				
20		34285	319423	0.75	-0.48	-0.08	-0.42	-0.24	0.18	-0.84	-0.42	-0.88	0.31	0.31	-0.25	-0.22	0.22	-0.29	-0.05	-0.44				
21		30242	312842	-0.08	0.19	-1.81	0.92	1.0	0.48	-1.32	0.54	1.29	0.38	1.22	-0.82	1.75	-1.38	-0.1	-0.61	-0.57				
22		7732	309842	-0.72	0.38	-1.42	0.28	1.05	0.24	-0.49	-1.81	0.21	-0.32	0.82	0.46	0.22	-1.1	0.08	0.45	-1.11				
23		75471	319758	-0.37	1.08	-0.37	0.5	-1.12	-0.57	-1.87	0.85	-1.27	0.44	1.4	-1.72	0.82	0.81	-2.01	-0.41	0.29				
24		457	300824	-0.17	-0.47	-1.12	-0.84	0.32	-0.82	-0.88	-0.88	0.82	0.42	0.48	1.41	-1.22	-0.01	-0.21	-0.05	-0.15				
25		12785	319382	0.08	0.77	0.12	0.58	0.81	-0.38	0.81	0.51	-0.34	0.28	-0.85	-0.2	0.88	-0.77	0.52	0.75	1.13				
26		6438	307382	0.42	1.25	-2.14	-1.48	-1.85	-0.38	-1.4	0.1	1.8	1.87	1.22	-1.22	-0.22	0.29	-0.25	1.05	0.1				
27		5451	308423	1.19	-0.82	0.58	0.51	-1.12	-1.8	-0.52	-1.62	0.85	-2.87	0.6	-1.22	1.5	-2.49	0.62	2.22	1.71				
28		8228	304558	-0.94	-0.24	-0.88	0.12	1.44	-0.4	0.32	0.31	-0.75	-0.55	-0.86	1.96	1.22	-1.41	-0.98	1.49	-0.52				
29		7628	309523	-0.23	0.04	-0.58	-1.29	1.07	-0.55	-0.48	-1.22	-0.57	0.31	1.82	0.61	-0.79	0.27	-0.22	0.46	0.95				
30		1026	301588	0.14	0.58	-0.48	-0.35	0.28	-0.49	-0.3	-0.64	-0.87	0.4	0.34	1.32	0.95	0.34	-0.29	0.78	0.52				
31		7738	309848	-0.08	0.1	0.48	-0.84	-0.6	0.3	-0.17	-0.33	1.2	1.84	-1.12	-0.82	-0.88	0.89	-0.76	-0.22	0.25				
32		3428	304212	0.18	0.7	-1.2	-1.52	0.28	0.84	1.81	0.82	0.44	0.36	0.32	1.82	0.32	0.14	0.4	1.48	1.13				

Gènes

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	
1																									
2	10150	319083	-0.21	0.90	-0.95	-0.2	1.05	-0.52	0.25	0.31	-0.94	0.22	-1.09	0.8	0.22	-0.95	-0.95	0.54	-0.79						
3	3078	300803	-0.67	1.1	-0.71	-1.23	1.45	-1.37	0.86	0.25	0.25	0.22	-0.57	0.62	0.52	0.44	-0.91	0.52	0.00						
4	629	301174	0.22	0.2	-0.27	-0.73	-0.45	-0.89	0.62	0.26	-0.5	-0.22	0.52	0.99	-0.41	-0.52	-1.01	-1.01	-1.47						
5	75258	319623	-0.29	0.62	-1.2	0.2	1.02	-0.2	0.24	0.76	0.28	0.05	0.79	0.74	0.05	-0.09	0.66	0.20	-0.01						
6	34985	319628	0.12	-0.48	-0.56	0.09	-0.23	-0.37	0.21	0.02	-1.56	1.05	1.22	0.54	2.22	1.81	0.71	0.74	0.61						
7	1298	300878	0.6	0.72	-0.71	0.28	-1.73	-0.88	0.26	0.06	1.47	-0.6	-0.71	-1.09	1.44	-0.26	-0.01	-0.16	-0.1						
8	77571	49523	-0.11	1.04	-1.90	0.48	0.74	-0.24	0.87	0.26	0.25	-0.24	0.45	0.77	-0.21	-0.14	-0.52	1.22	-0.17						
9	4202	305824	-0.62	0.80	-1.02	0.22	-0.22	-0.78	0.25	0.06	-0.71	-2.42	1.81	0.89	0.22	-1.77	-0.15	1.01	-0.25						
10	18225	319828	-1.22	0.62	-1.22	0.56	1.22	-0.7	0.8	0.02	0.26	-0.17	-0.82	1.06	-0.01	-1.25	-0.26	0.69	-0.29						
11	11084	311828	-0.62	0.22	-0.02	-1.24	-2.55	0.15	0.29	0.24	-0.2	0.41	0.59	0.59	-0.11	-0.21	0.07	0.42	0.64						
12	1298	300823	0.48	0.77	-0.44	-0.77	-0.84	0.15	0.21	0.28	0.28	0.96	1.61	0.27	-0.29	0.11	-0.92	1.0	0.75						
13	34273	315428	0.28	0.28	-0.18	-0.24	-1.32	-0.28	0.75	-0.21	0.71	0.24	0.21	-0.41	0.81	1.01	0.21	-0.28	-1.45						
14	9258	311822	-1.21	1.17	-1.12	0.92	0.98	-0.74	0.47	0.28	-0.22	-0.8	0.22	1.89	1.82	-1.59	0.06	0.69	-0.29						
15	75823	319823	-1.54	1.7	-1.05	-0.81	1.25	-0.4	0.22	0.42	-0.47	1.87	-1.3	2.49	0.22	-1.61	-0.61	-2.72	0.1						
16	34288	319823	-0.25	0.77	-0.94	-1.22	0.25	-0.24	0.28	0.28	-0.89	0.24	0.79	1.06	-0.41	-1.99	-0.2	1.22	-0.45						
17	1215	301827	-0.28	0.14	-1.14	0.12	0.2	-0.22	0.22	0.02	-0.82	0.48	-0.09	-0.49	-0.22	-0.04	-0.25	-0.22	0.25						
18	34285	315421	0.75	-0.48	-0.08	-0.41	-0.24	0.15	0.24	0.42	-0.28	0.21	0.21	-0.22	-0.22	0.22	-0.29	-0.09	-0.44						
19	30242	311842	-0.09	0.79	-1.82	0.92	1.0	0.48	0.22	0.24	1.29	0.28	1.22	-0.82	1.75	-1.29	-0.1	-0.61	-0.27						
20	7122	309842	-0.72	0.28	-1.42	0.28	1.05	0.24	0.82	0.21	0.21	-0.22	0.82	0.45	0.22	-1.1	0.06	0.45	-1.11						
21	75473	319758	-0.27	1.09	-0.27	0.5	-1.12	-0.27	0.27	0.05	-1.27	0.44	1.4	-1.72	0.82	0.81	-2.01	-0.41	0.29						
22	457	300824	-0.17	-0.47	-1.12	-0.84	0.22	-0.82	0.89	0.28	0.82	0.42	0.48	1.41	-1.22	-0.01	-0.21	-0.09	-0.15						
23	12185	316028	0.06	0.77	0.12	0.58	0.82	-0.28	0.21	0.01	-0.24	0.28	-0.82	-0.2	0.98	-0.77	0.22	0.75	1.12						
24	6428	307308	0.42	1.25	-2.14	-1.48	-1.92	-0.28	0.4	0.2	1.8	1.87	1.22	-1.22	-0.22	0.29	-0.25	1.05	0.1						
25	5452	308422	1.19	-0.82	0.59	0.52	-1.12	-1.8	0.52	0.02	0.86	-2.87	0.6	-1.22	1.5	-2.49	0.62	2.22	1.71						
26	8228	304528	-0.94	-0.24	-0.88	0.12	1.44	-0.4	0.2	0.01	-0.75	-0.25	-0.26	1.96	1.22	-1.41	-0.99	1.49	-0.22						
27	7628	309522	-0.22	0.04	-0.58	-1.29	1.07	-0.55	0.89	0.22	-0.57	0.21	1.82	0.61	-0.79	0.27	-0.22	0.46	0.95						
28	1025	301588	0.14	0.59	-0.69	-0.25	0.28	-0.42	0.2	0.24	-0.87	0.4	0.24	1.22	0.25	0.24	-0.29	0.78	0.52						
29	7128	309848	-0.09	0.1	0.48	-0.84	-0.6	0.2	0.27	0.28	1.2	1.84	-1.12	-0.82	-0.28	0.89	-0.76	-0.22	0.25						
30	3428	304222	0.18	0.2	1.2	1.22	0.22	0.24	0.27	0.28	0.28	0.28	0.28	1.02	0.22	0.24	0.24	1.02	1.12						

Gènes

Puces

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	
1																										
2	10150	319083	-0.21	0.90	-0.95	-0.2	1.05	-0.52	-1.25	-0.31	-0.94	0.12	-1.09	0.8	0.23	-0.95	-0.95	0.54	-0.19							
3	3078	300803	-0.67	1.1	-0.71	-1.23	1.45	-1.37	-0.88	-0.25	0.25	0.22	-0.57	0.63	0.53	0.44	-0.91	0.52	0.09							
4	629	301174	0.22	0.2	-0.27	-0.73	-0.45	-0.89	-0.83	-0.38	-0.5	-0.22	0.53	0.99	-0.41	-0.52	-1.01	-1.01	-1.47							
5	35258	319623	-0.29	0.62	-1.2	0.2	1.02	-0.2	-0.24	0.76	0.38	0.05	0.79	0.74	0.05	-0.09	0.66	0.29	-0.01							
6	34985	319635	0.12	-0.48	-0.56	0.09	-0.23	-0.37	1.01	0.85	-1.56	1.05	1.23	0.54	2.23	1.81	0.71	2.74	0.61							
7	3298	300878	0.6	0.72	-0.73	0.28	-1.73	-0.88	-0.38	0.26	1.47	-0.6	-0.73	-1.09	1.44	-0.35	-0.01	-0.16	-0.1							
8	17573	49523	-0.11	1.04	-1.92	0.48	0.74	-0.24	-0.87	-0.26	0.25	-0.34	0.45	0.77	-0.21	-0.14	-0.52	1.22	-0.17							
9	4323	305824	-0.62	0.82	-1.02	-0.22	-0.23	-0.78	-0.55	-1.06	-0.71	-2.43	1.81	0.89	0.22	-1.77	-0.15	1.01	-0.05							
10	13925	319825	-1.22	0.62	-1.22	0.56	1.22	-0.7	-0.8	0.05	0.26	-0.17	-0.82	1.06	-0.01	-1.35	-0.35	0.69	0.09							
11	11084	311828	-0.62	0.22	-0.02	-1.24	-2.55	0.15	-0.28	-0.24	-0.2	0.41	0.59	0.99	-0.11	-0.21	0.01	0.01	1.04							
12	3298	300823	0.48	0.77	-0.44	-0.77	-0.84	0.15	-1.81	-0.38	0.28	0.96	1.61	0.27	-0.29	0.11	-0.99	1.0	0.75							
13	34273	319429	0.25	0.39	-0.19	-0.24	-1.32	-0.39	-1.75	-0.81	0.71	0.04	0.81	-0.41	0.01	0.01	0.01	-0.38	-1.45							
14	5958	311822	-1.21	1.17	-1.13	0.92	0.98	-0.74	-0.47	0.39	-0.23	-0.9	0.27	0.01	1.02	-1.59	0.06	0.69	-0.59							
15	7583	319823	-1.54	1.7	-1.05	-0.81	1.25	-0.4	-0.53	-0.42	-0.47	1.07	0.1	2.49	0.51	-1.61	-0.61	-2.72	0.1							
16	34388	319825	-0.25	0.77	-0.94	-1.23	0.35	-0.34	-0.28	-0.38	-0.21	-0.24	0.76	1.06	-0.41	-1.99	-0.5	1.23	-0.49							
17	1315	301817	-0.28	0.14	-1.14	0.13	0.3	-0.22	0.22	-0.21	-0.81	0.48	-0.09	-0.46	-0.25	-0.04	-0.35	-0.22	0.29							
18	34285	319421	0.75	-0.48	-0.08	-0.41	-0.24	0.15	0.01	-0.43	-0.38	0.31	0.31	-0.25	-0.21	0.21	-0.29	-0.09	-0.44							
19	30242	311842	-0.09	0.19	-1.81	0.92	1.0	0.49	-1.32	0.54	1.29	0.28	1.32	-0.82	1.75	-1.39	-0.1	-0.61	-0.57							
20	1132	309842	-0.72	0.39	-1.42	0.28	1.05	0.24	-0.49	-1.81	0.21	-0.21	0.82	0.46	0.22	-1.1	0.08	0.45	-1.11							
21	35473	319758	-0.37	1.09	-0.37	0.5	-1.13	-0.57	-1.87	0.85	-1.27	0.44	1.4	-1.73	0.82	0.81	-2.01	-0.41	0.29							
22	457	300824	-0.17	-0.47	-1.13	-0.84	0.32	-0.82	-0.89	-0.88	0.82	0.42	0.48	1.41	-1.22	-0.01	-0.21	-0.09	-0.15							
23	33185	318328	0.06	0.77	0.12	0.58	0.81	-0.38	0.81	0.51	-0.34	0.29	-0.85	-0.2	0.98	-0.77	0.52	0.75	1.15							
24	6438	307308	0.42	1.25	-2.14	-1.48	-1.95	-0.39	-1.4	0.1	1.9	1.87	1.22	-1.22	-0.21	0.29	-0.29	1.65	0.1							
25	5453	306423	1.19	-0.82	0.59	0.51	-1.13	-1.8	-0.52	-1.63	0.85	-2.87	0.6	-1.21	1.5	-2.49	0.62	2.22	1.71							
26	3228	304558	-0.94	-0.24	-0.88	0.13	1.44	-0.4	0.32	0.31	-0.75	-0.55	-0.36	1.96	1.22	-1.41	-0.99	1.49	-0.52							
27	1628	309523	-0.23	0.04	-0.58	-1.29	1.07	-0.55	-0.49	-1.22	-0.57	0.31	1.83	0.61	-0.79	0.27	-0.22	0.46	0.95							
28	3095	301585	0.14	0.59	-0.69	-0.35	0.28	-0.49	-0.3	-0.64	-0.87	0.4	0.34	1.31	0.95	0.34	-0.29	0.78	0.52							
29	1138	309648	-0.09	0.1	0.48	-0.84	-0.6	0.3	-0.17	-0.33	1.2	1.84	-1.11	-0.93	-0.28	0.89	-0.76	-0.22	0.25							
30	3428	304213	0.18	0.7	1.2	1.52	0.28	0.84	1.81	0.82	0.44	0.28	0.21	1.02	0.22	0.14	0.2	1.69	1.15							

Gènes

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

Puces

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1																								
2		10150	319083	-0.21	0.90	-0.95	-0.2	1.05	-0.52	-1.25	-0.31	-0.94	0.13	-1.09	0.8	0.23	-0.95	-0.95	0.54	-0.70				
3		3078	300803	-0.67	1.1	-0.71	-1.20	1.45	-1.37	-0.88	-0.35	0.35	0.20	-0.57	0.63	0.53	0.44	-0.91	0.52	0.00				
4		629	301174	0.22	0.2	-0.27	-0.73	-0.45	-0.89	-0.83	-0.36	-0.5	-0.23	0.53	0.80	-0.41	-0.52	-1.01	-1.01	-1.47				
5		75258	319603	-0.29	0.62	-1.2	0.2	1.02	-0.2	-0.24	0.76	0.38	0.05	0.70	0.74	0.05	-0.09	0.66	0.30	-0.01				
6		34085	319618	0.12	-0.48	-0.58	0.08	-0.23	-0.37	1.01	0.85	-1.50	1.05	1.23	0.54	2.23	1.81	0.71	2.74	0.61				
7		1298	300878	0.6	0.72	-0.73	0.28	-1.73	-0.88	-0.38	0.36	1.47	-0.6	-0.73	-1.09	1.44	-0.35	-0.01	-0.35	-0.3				
8		77573	49503	-0.11	1.04	-1.90	0.48	0.74	-0.24	-0.87	-0.30	0.05	-0.34	0.45	0.77	-0.31	-0.34	-0.52	1.22	-0.17				
9		4303	305804	-0.62	0.80	-1.02	0.20	-0.23	-0.78	-0.55	-1.00	-0.71	-2.43	1.81	0.80	0.23	-1.77	-0.35	1.01	-0.05				
10		13025	319005	-1.20	0.62	-1.22	0.58	1.23	-0.7	-0.8	0.05	0.26	-0.77	-0.83	1.00	-0.01	-1.35	-0.35	0.60					
11		11084	311808	-0.62	0.20	-0.02	-1.24	-2.55	0.15	-0.28	-0.24	-0.2	0.41	0.90	0.90	-0.11	-0.31	0.03	-0.1	1.04				
12		1298	300803	0.48	0.77	-0.44	-0.77	-0.84	0.15	-1.81	-0.38	0.38	0.98	1.61	0.27	-0.29	0.11	-0.90	1.0	0.75				
13		34173	319403	0.28	0.38	-0.18	-0.24	-1.32	-0.38	-1.75	-0.81	0.71	0.04	0.81	-0.41	0.01	1.81	0.31	-0.38	-1.45				
14		9058	311803	-1.21	1.17	-1.13	0.93	0.98	-0.74	-0.47	0.38	-0.23	-0.3	0.37	0.01	1.81	-1.50	0.06	0.60	-0.59				
15		7503	319083	-1.54	1.7	-1.05	-0.81	1.25	-0.4	-0.53	-0.42	-0.47	1.07	0.3	2.49	0.51	-1.61	-0.61	-2.72	0.1				
16		34308	319503	-0.25	0.77	-0.94	-1.23	0.35	-0.34	-0.28	-0.38	0.01	-0.24	0.70	1.00	-0.41	-1.90	-0.5	1.23	-0.40				
17		1315	301817	-0.28	0.14	-1.14	0.13	0.3	-0.1	0.1	0.1	-0.81	0.48	-0.00	-0.40	-0.25	-0.04	-0.30	-0.23	0.20				
18		34285	319403	0.75	-0.48	-0.08	-0.41	-0.24	0.15	0.01	-0.43	-0.30	0.31	0.31	-0.25	-0.21	0.21	-0.29	-0.00	-0.44				
19		30340	311842	-0.09	0.79	-1.81	0.92	1.0	0.40	-1.32	0.54	1.29	0.38	1.32	-0.83	1.75	-1.30	-0.1	-0.61	-0.57				
20		7132	309842	-0.72	0.38	-1.62	0.28	1.05	0.24	-0.49	-1.81	0.21	-0.31	0.83	0.40	0.33	-1.1	0.08	0.45	-1.11				
21		74473	319798	-0.37	1.08	-0.37	0.5	-1.13	-0.57	-1.87	0.85	-1.27	0.44	1.4	-1.73	0.83	0.81	-2.01	-0.41	0.29				
22		457	300804	-0.17	-0.47	-1.13	-0.84	0.30	-0.82	-0.88	-0.88	0.82	0.43	0.48	1.41	-1.23	-0.01	-0.31	-0.00	-0.15				
23		11785	318008	0.06	0.77	0.12	0.58	0.81	-0.38	0.81	0.51	-0.34	0.05	-0.85	-0.2	0.90	-0.77	0.53	0.75	1.15				
24		6438	307308	0.42	1.25	-2.14	-1.48	-1.95	-0.39	-1.4	0.3	1.8	1.87	1.23	-1.23	-0.21	0.29	-0.20	1.05	0.1				
25		5453	308413	1.19	-0.82	0.59	0.51	-1.13	-1.8	-0.52	-1.63	0.85	-2.87	0.6	-1.21	1.5	-2.40	0.60	2.22	1.71				
26		8028	304558	-0.94	-0.24	-0.88	0.13	1.44	-0.4	0.32	0.31	-0.75	-0.55	-0.30	1.96	1.23	-1.41	-0.90	1.40	-0.52				
27		7628	309523	-0.23	0.04	-0.58	-1.29	1.07	-0.55	-0.68	-1.32	-0.57	0.31	1.83	0.61	-0.70	0.27	-0.22	0.46	0.95				
28		1005	301585	0.14	0.59	-0.68	-0.35	0.28	-0.49	-0.3	-0.64	-0.87	0.4	0.34	1.31	0.95	0.34	-0.30	0.78	0.52				
29		7138	309648	-0.08	0.1	0.48	-0.84	-0.6	0.3	-0.37	-0.33	1.2	1.84	-1.11	-0.83	-0.38	0.89	-0.76	-0.22	0.95				
30		3448	304513	0.18	0.7	1.5	1.57	0.28	0.84	1.81	0.83	0.44	0.36	0.33	1.00	0.33	0.14	0.7	1.60	1.10				

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

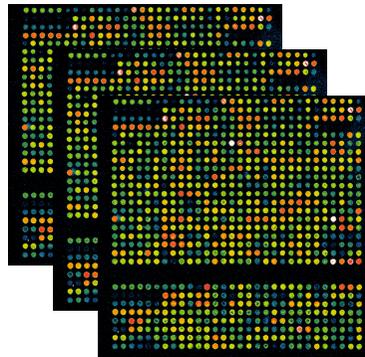
- **Motifs séquentiels** : séquences fréquentes d'itemsets ordonnés



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

< (G5 G4) (G6) >

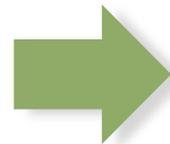
Problématique générale



Technologies puces à ADN



Bases de connaissances et bases
bibliographiques disponibles en ligne

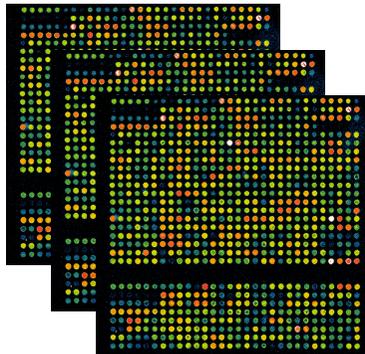


Nouvelles
connaissances





Problématique générale



Technologies puces à ADN



Bases de connaissances et bases
bibliographiques disponibles en ligne

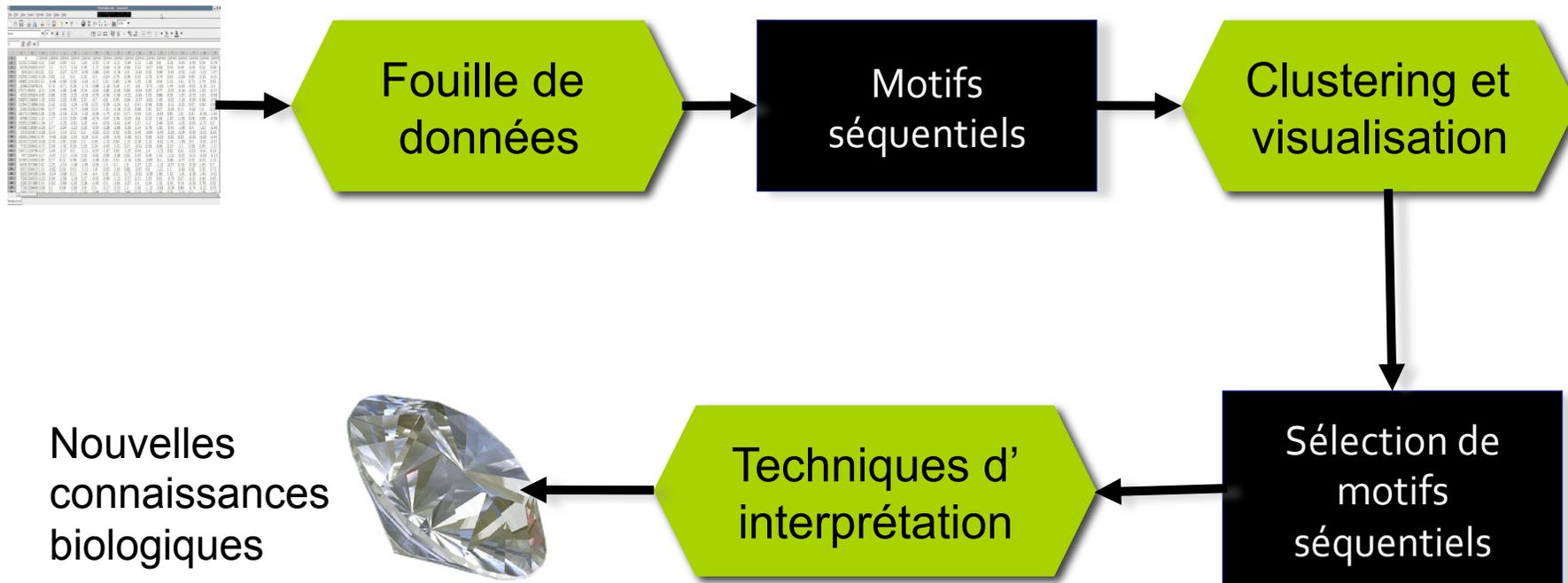


Nouvelles
connaissances

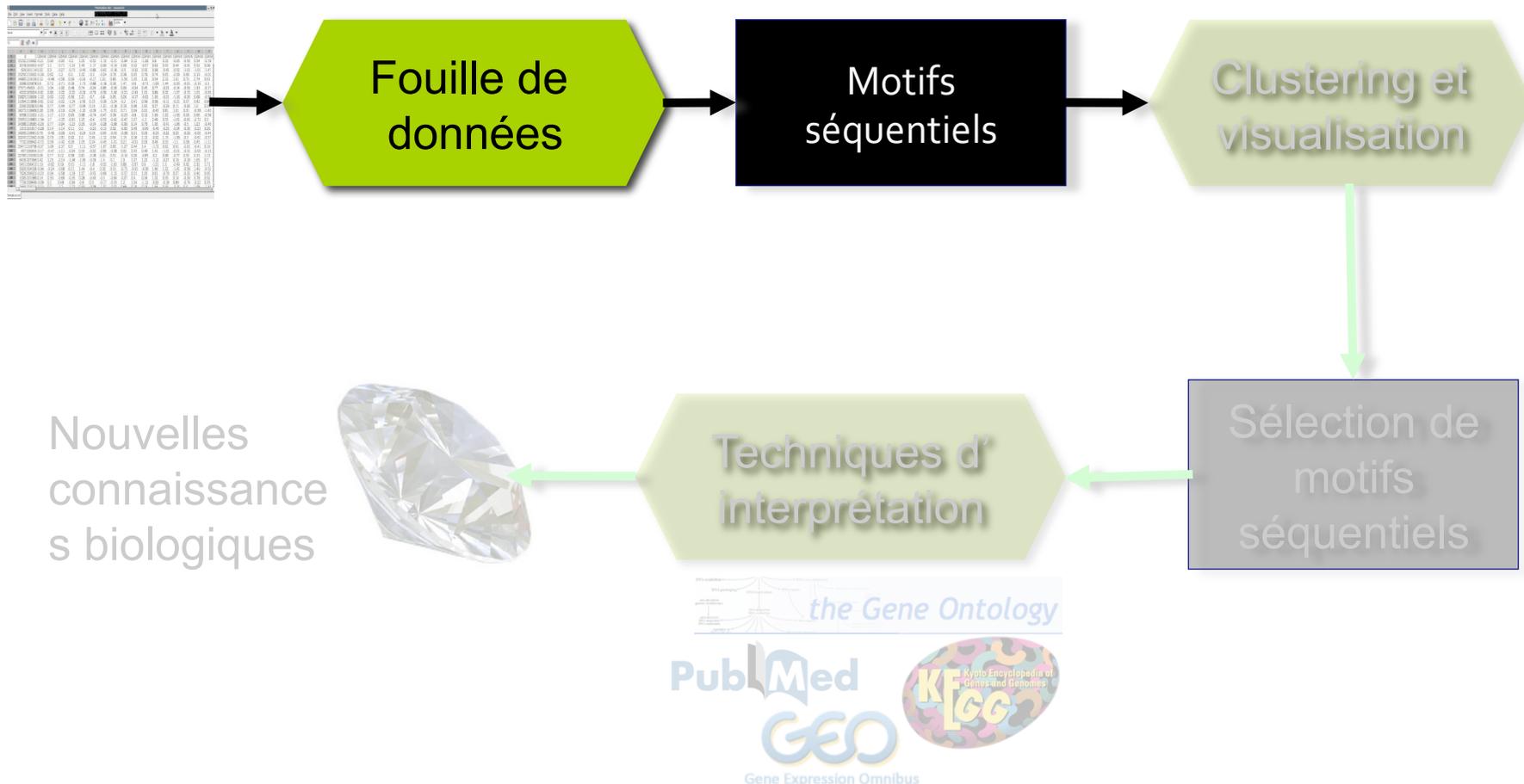


Challenge : exploiter toutes ces données en terme
de signification biologique

Processus général



Processus général



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:

$$S1 = \langle (G2 \ G3) \rangle$$

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:

$$S1 = \langle (G2 \ G3) \ (\mathbf{G1}) \rangle$$

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:

$$S1 = \langle (G2 \ G3) (G1) (\mathbf{G4}) \rangle$$

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:

$$S1 = \langle (G2 \ G3) (G1) (G4) \rangle$$

$$S2 = \langle (G4) (G1 \ G3) (G2) \rangle$$

$$S3 = \langle (G2) (G1) (G4) (G3) \rangle$$

Support d'une séquence

- **Fréquence dans la base de séquence**

Sequences:

$$S1 = < (\mathbf{G2} \ G3) (\mathbf{G1}) (\mathbf{G4}) >$$

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

$$S3 = < (\mathbf{G2}) (\mathbf{G1}) (\mathbf{G4}) (G3) >$$

Sequential pattern:

$$SP1 = < (\mathbf{G2}) (\mathbf{G1}) (\mathbf{G4}) > \text{ support } 2/3$$

Motifs séquentiels

- Support minimum fixé par l'utilisateur

Example:

Let be **minSupport** = 2/3

Sequential patterns:

$SP1 = \langle (G2) (G1) (G4) \rangle$ support 2/3

$SP2 = \langle (G3 G5) \rangle$ support 2/3

$SP3 = \langle (G1) (G2) (G4) \rangle$ support 1/3

Motifs séquentiels

- Support minimum fixé par l'utilisateur

Example:

Let be **minSupport** = 2/3

Sequential patterns:

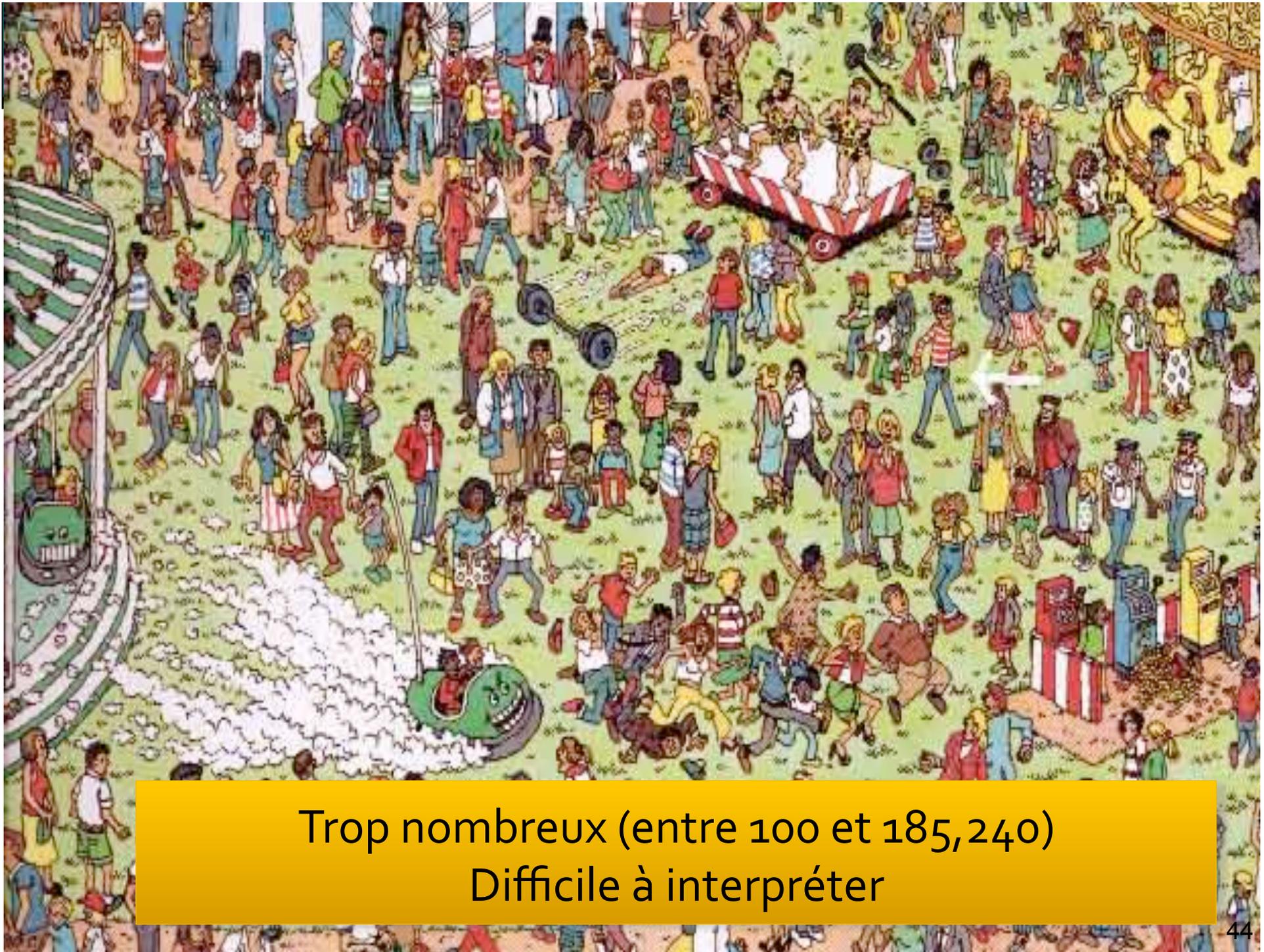
$SP1 = \langle (G2) (G1) (G4) \rangle$ support 2/3

$SP2 = \langle (G3) (G5) \rangle$ support 2/3

~~$SP3 = \langle (G1) (G2) (G4) \rangle$ support 1/3~~

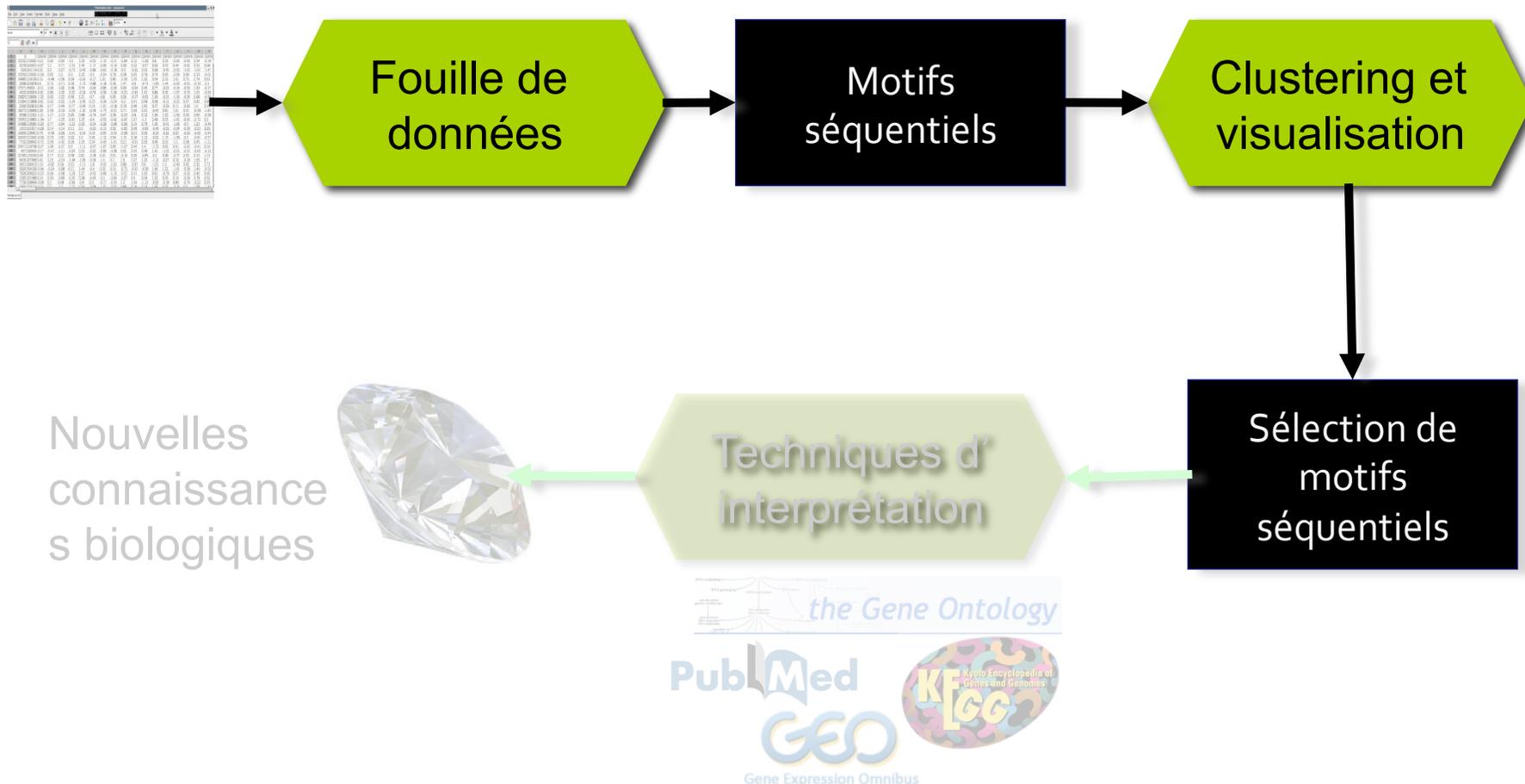
Motifs séquentiels +++

- **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



Comment comparer les motifs ? (Saneifar 2008)

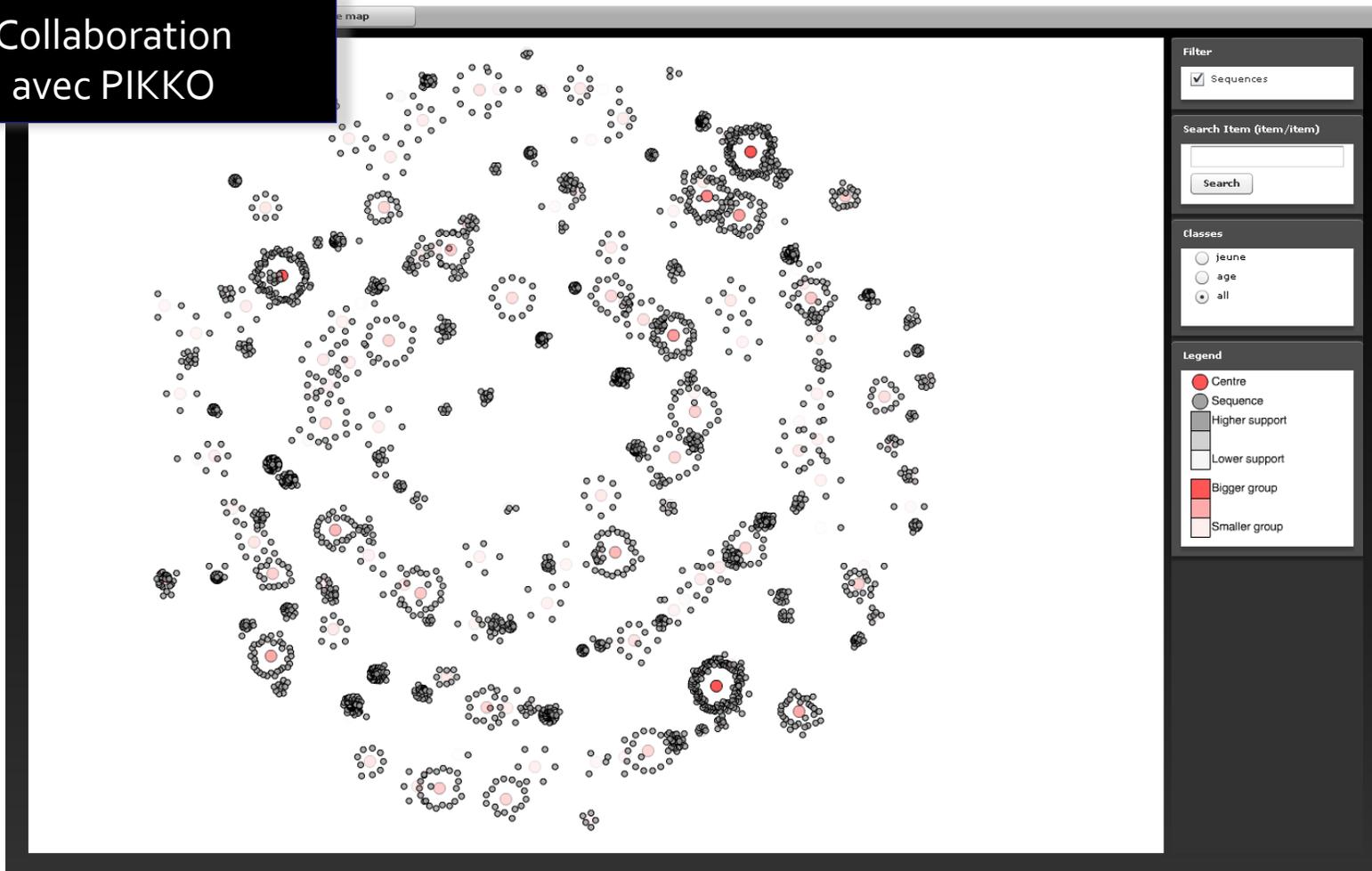
$$S_{75\%} = \langle (G_1)(G_2 G_3) \rangle$$
$$S'_{75\%} = \langle (G_2 G_3) (G_1) \rangle$$

■ **Mesure de similarité**

- Gènes communs et non communs
- Ordre des gènes
- Support

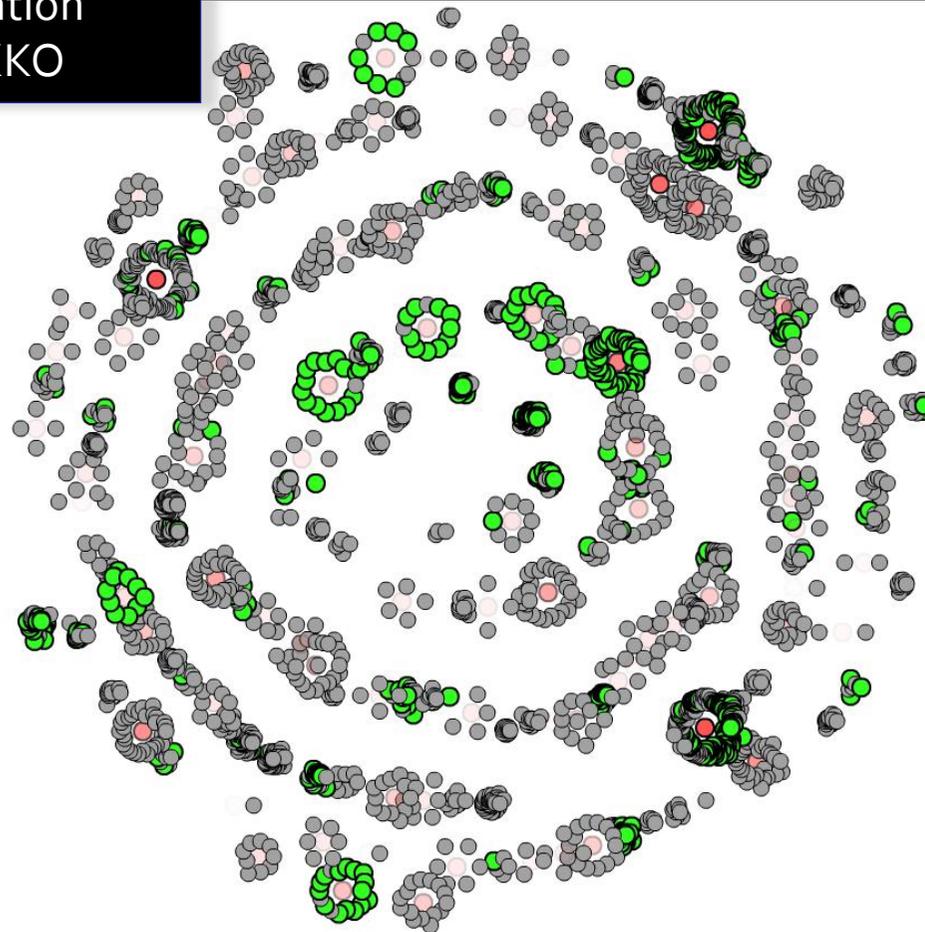
Clustering k-means

Collaboration
avec PIKKO



Clustering k-means

Collaboration
avec PIKKO



Filter

Sequences

Search Item (item/item)

Classes

jeune
 age
 all

Legend

Centre
 Sequence

Bigger group
 Smaller group

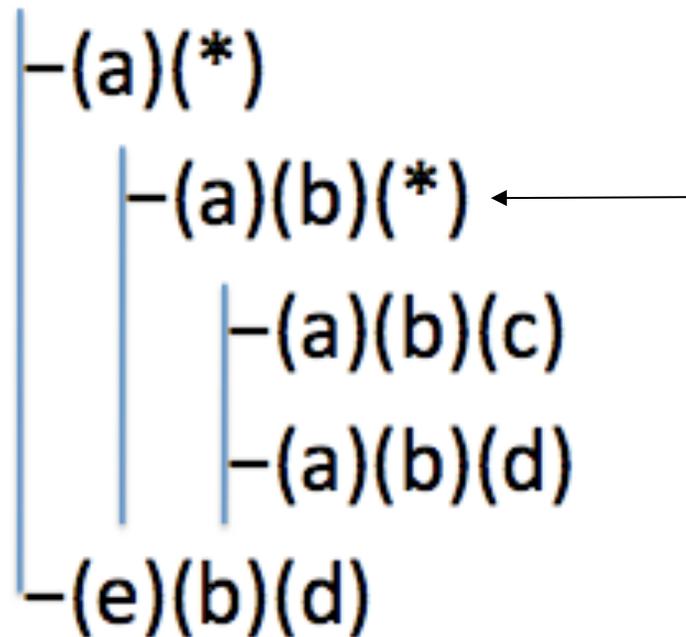
Higher support
 Lower support

Informations

Le nuage de points permet de visualiser des groupes de séquences de gènes

Clustering hiérarchique (Nin Guerero 2009)

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



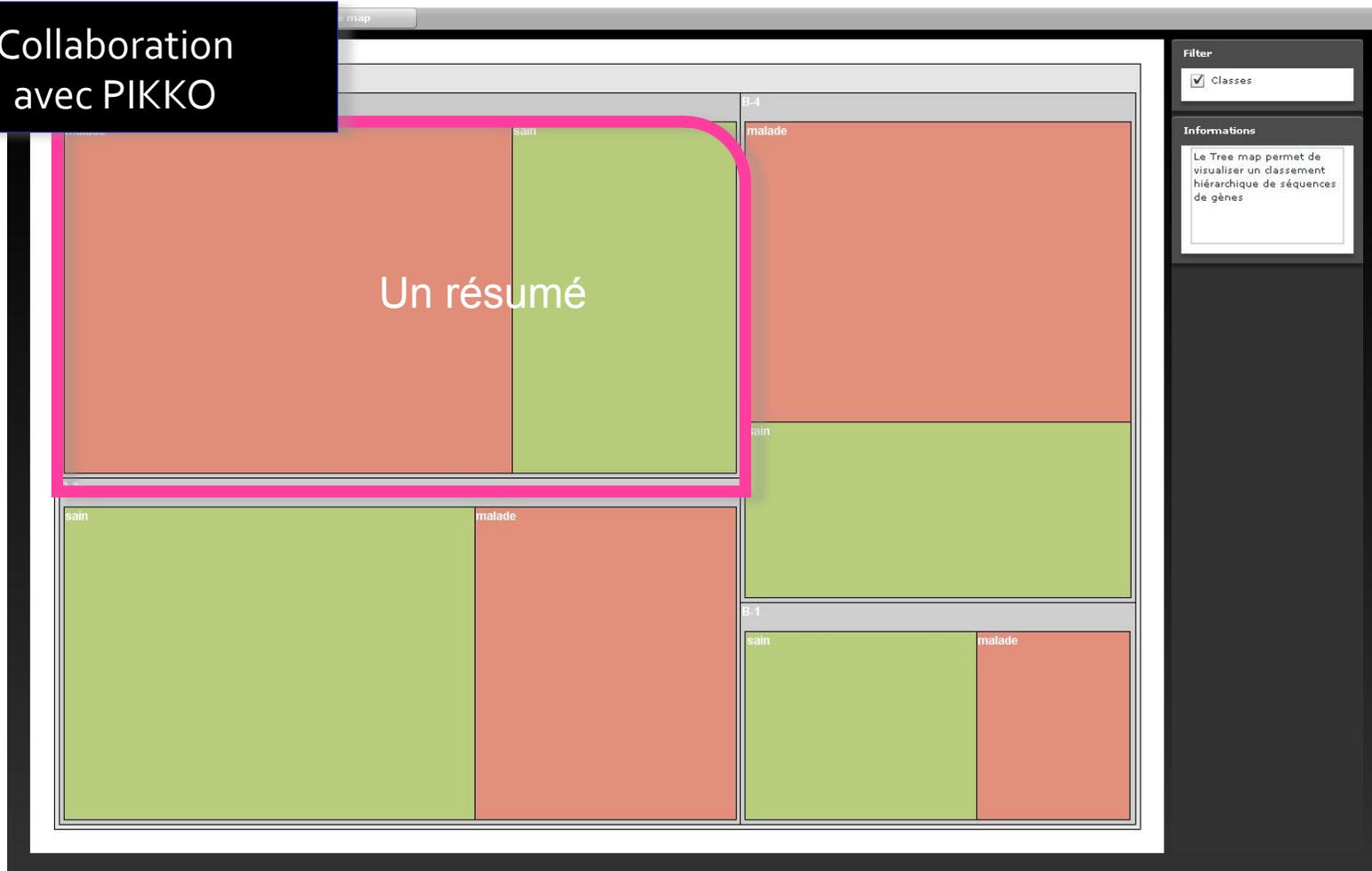
Clustering hiérarchique (Nin Guerero 2009)

Collaboration
avec PIKKO



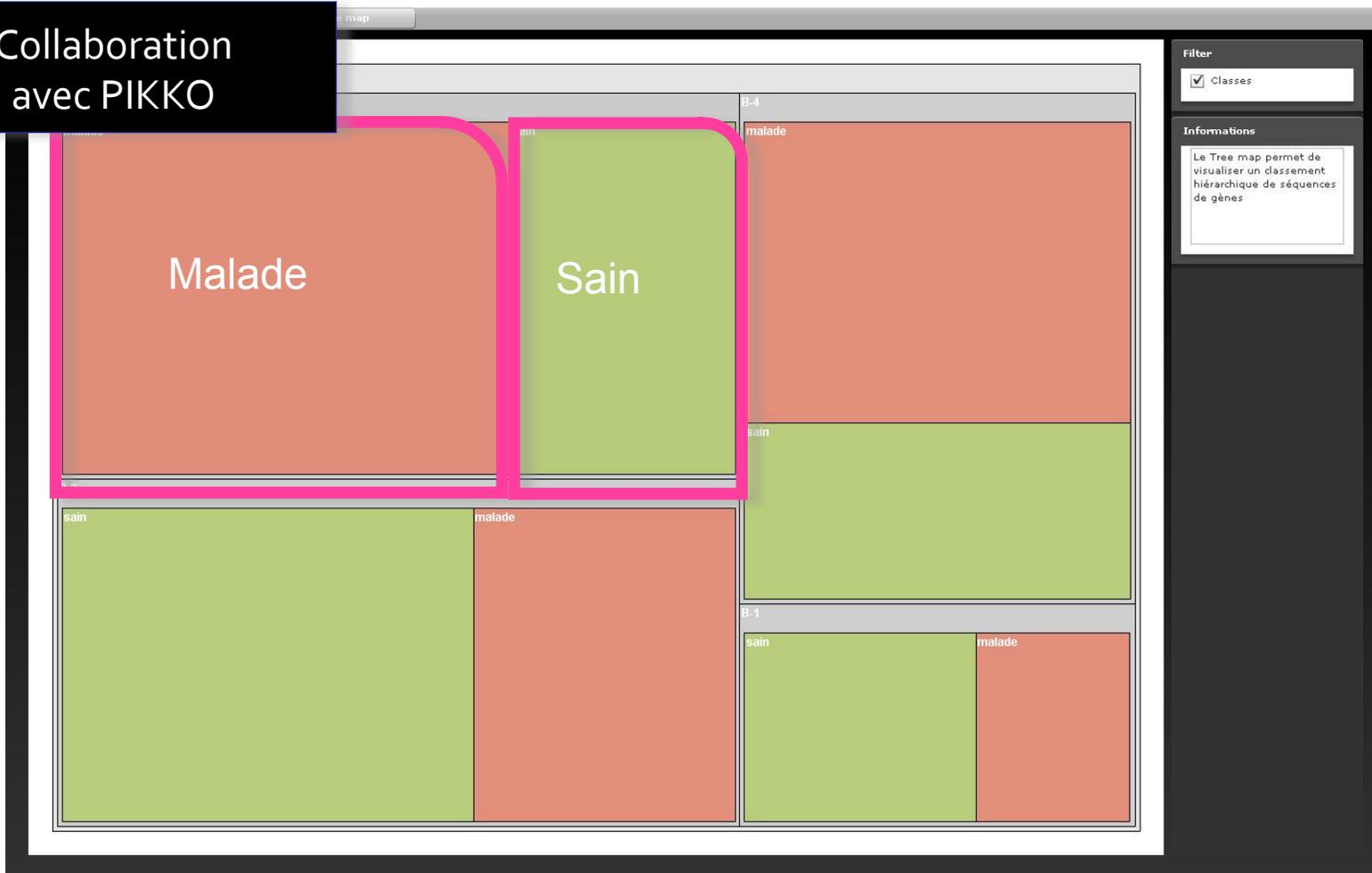
Clustering hiérarchique (Nin Guerero 2009)

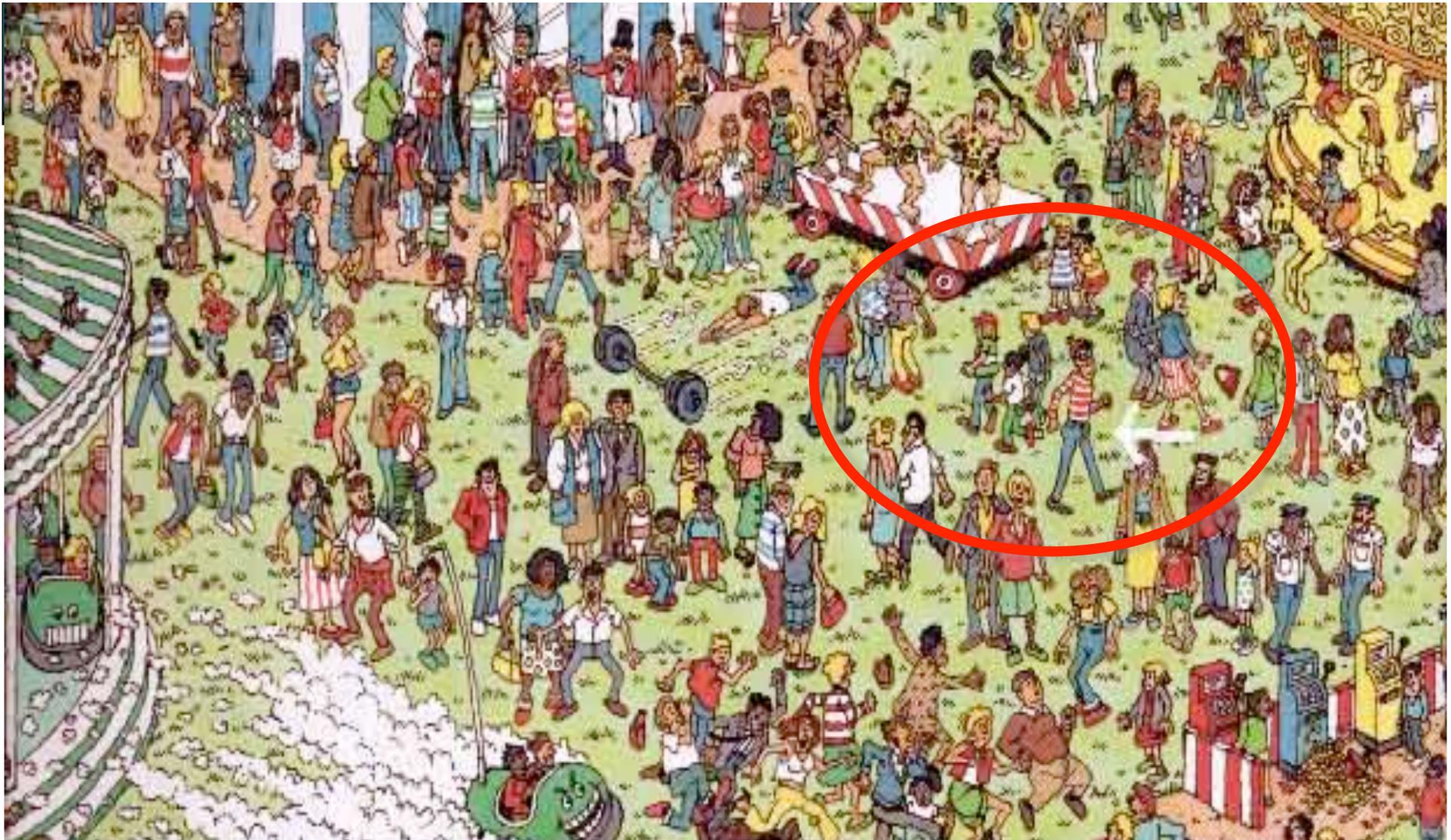
Collaboration
avec PIKKO



Clustering hiérarchique (Nin Guerero 2009)

Collaboration
avec PIKKO

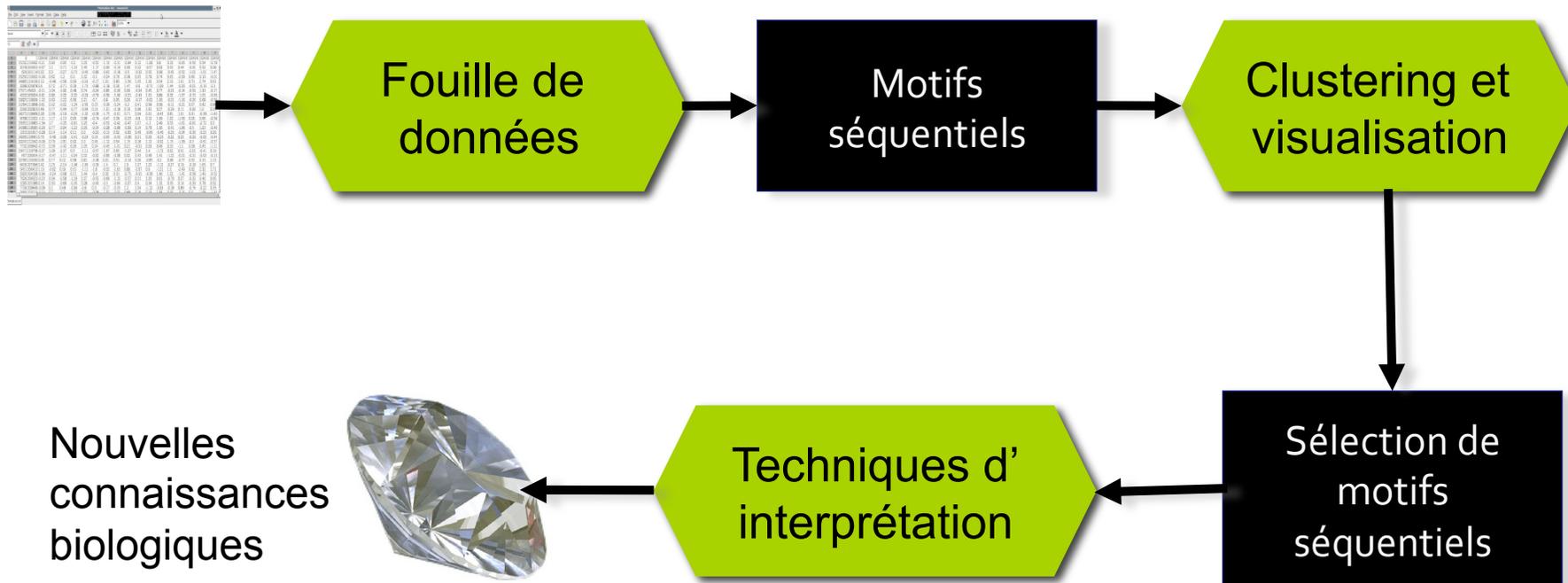




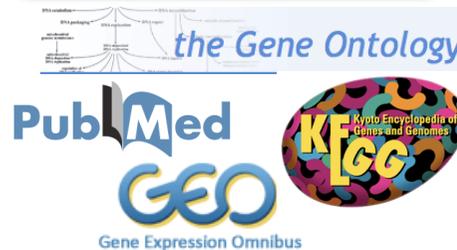
Et les connaissances disponibles en ligne ?



Processus général



Nouvelles
connaissances
biologiques



Interprétation des motifs via les documents (Bringay 2010)

$S75\%, 25\% = \langle (G1)(G2 G3) \rangle$



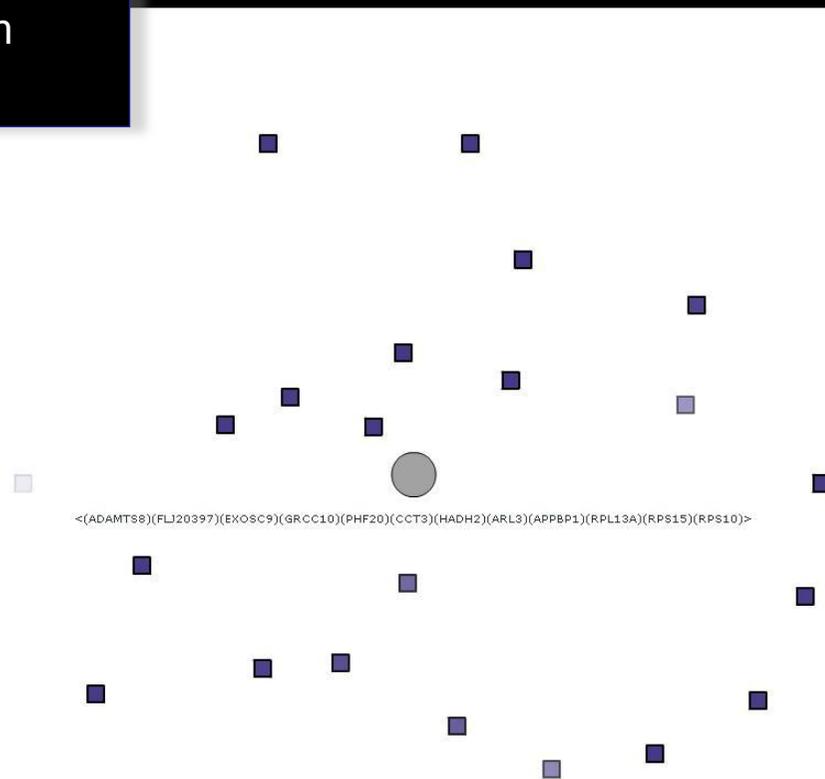
Textes



- 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



Legend

- Sequence
- Document
- Recent publication
- Older publication

Informations

Le système solaire permet de visualiser des documents : une séquence est difficile à interpréter. Cette visualisation associe à une séquence des informations récupérées en ligne et notamment des publications ordonnées selon leur pertinence.



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

Exemple de motif pertinent

$$S_{75} = \langle (MRVI_1)(PGAP_1)(PLA_2R_1)(A_2M)(GSK_3B) \rangle$$

- 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 ...

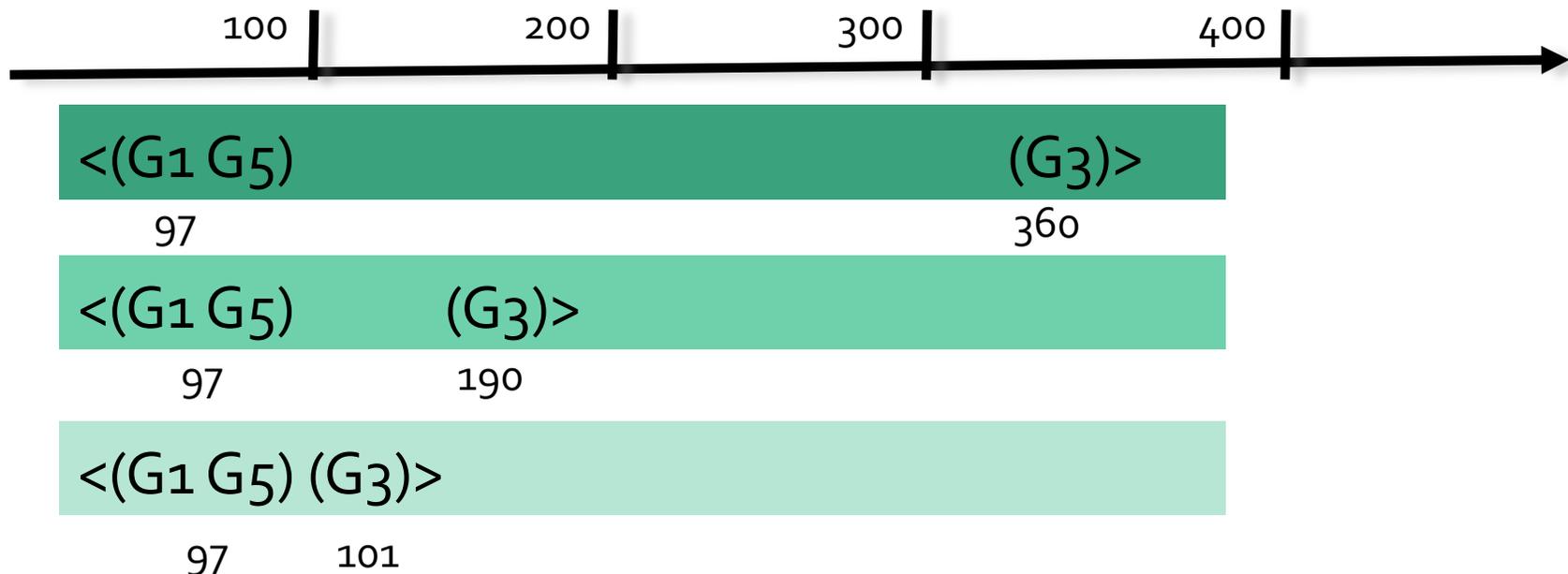
- **Motifs séquentiels** : ne sont pas facilement
 - **Compréhensibles** et **manipulables** par les experts
 - **Interprétables** d'une manière linguistique sans la définition d'une partition stricte des valeurs d'expression des gènes

$\langle (G_1 G_5)(G_3) \rangle$

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

$\langle (G_1 G_5)(G_3) \rangle$



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

$\langle (G_1 G_5)(G_3) \rangle$



$\langle (G_1 G_5)(\text{very over expressed } 0,8) (G_3) \rangle$

*G_3 is far much expressed compared to G_1 and G_5 ,
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	$\langle (G2)(G1 \ G5)(G3)(G4) \rangle$ 3.7 4 4.3 5 7 
M2	$\langle (G2)(G1 \ G5)(G4)(G3) \rangle$ 3.2 4.2 4.7 10 12 

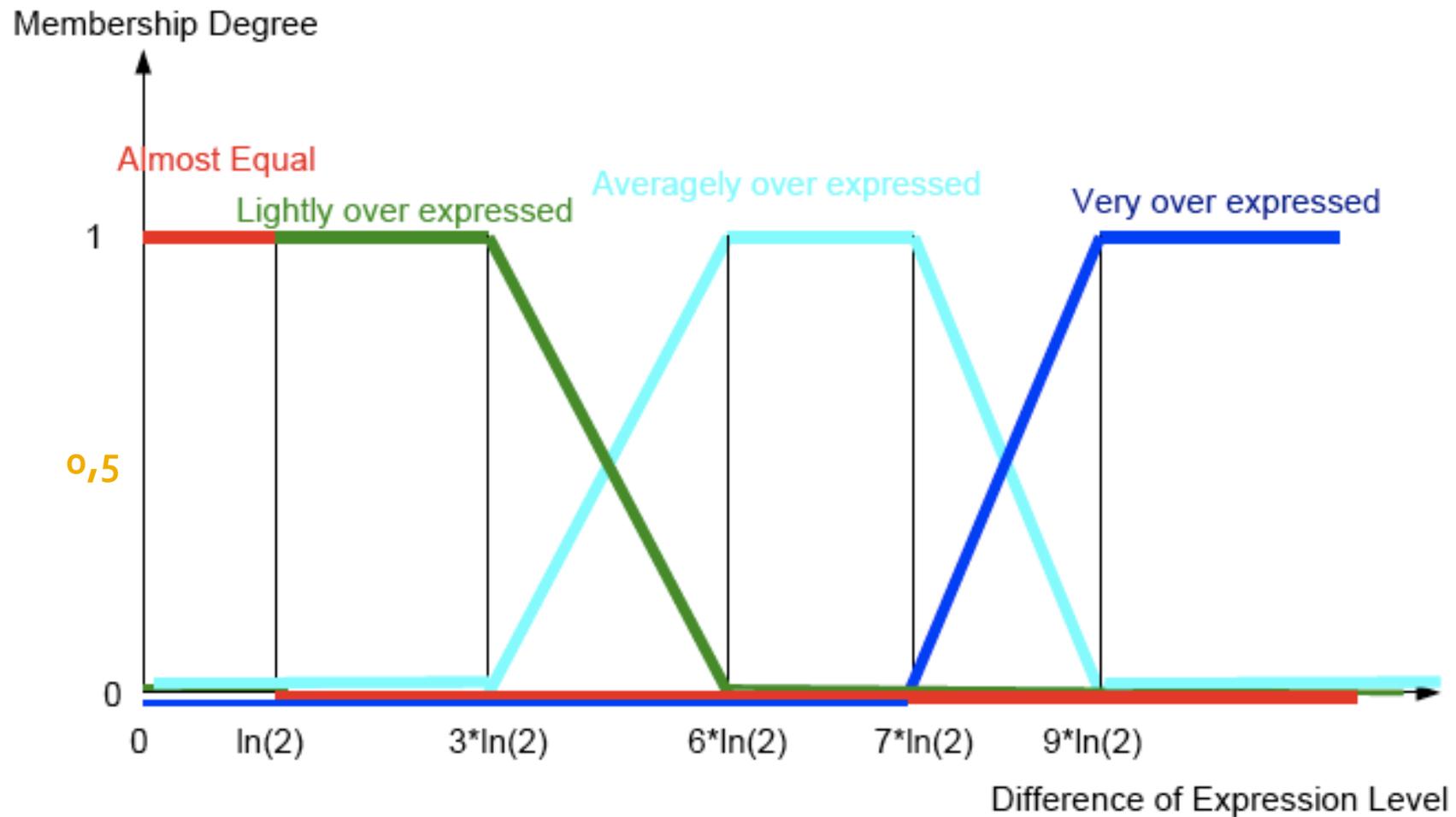
- **Différence d'Itemset $\delta(it2; it1)$** : la valeur absolue de la différence entre l'intensité du gène $it2$ et $it1$

- **Exemple:**

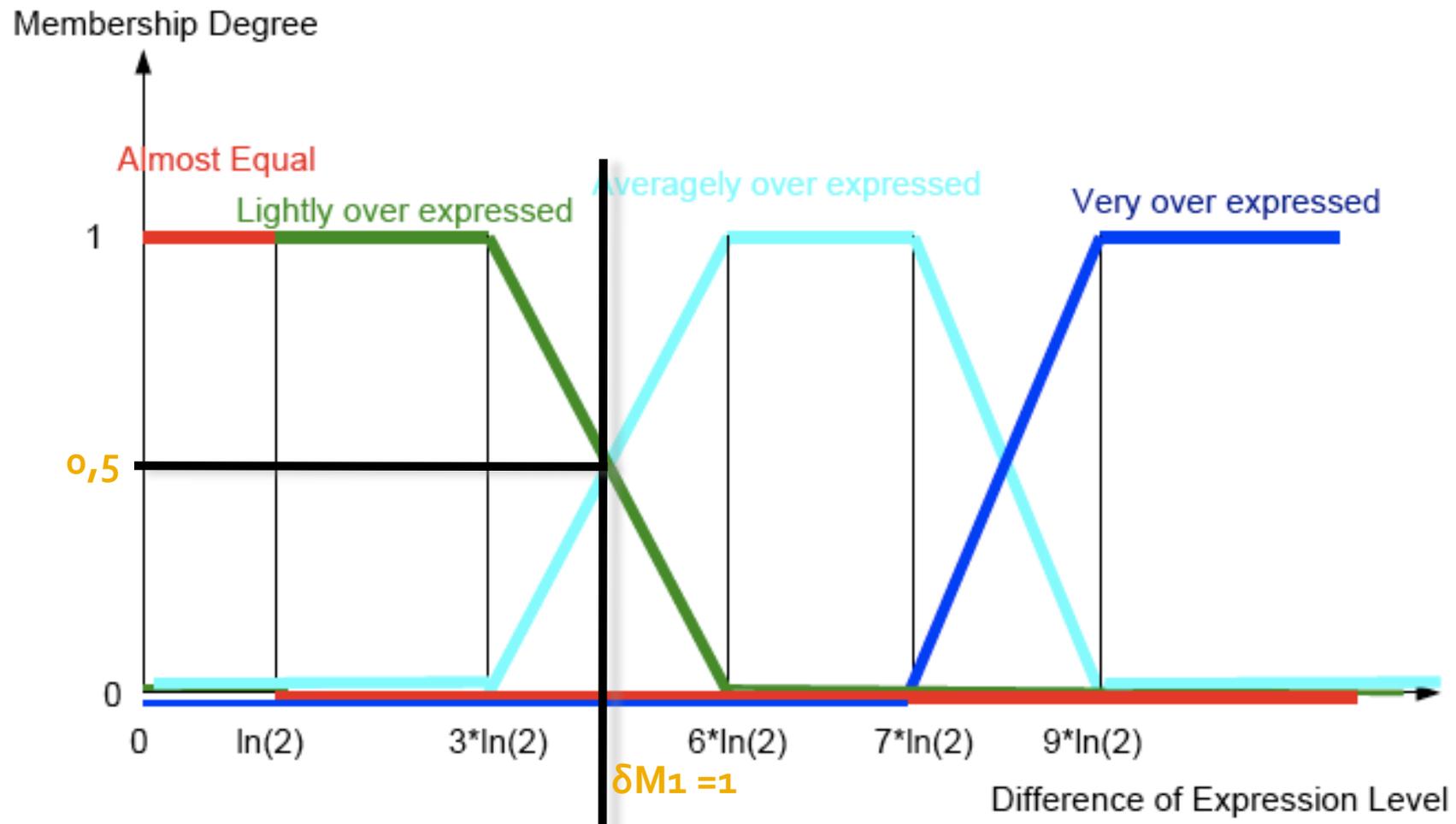
$$M1: \delta((G3); (G1 \ G5)) \\ = |5-4|=1$$

$$M2: \delta((G3); (G1 \ G5)) \\ = |12-4.2|=7,8$$

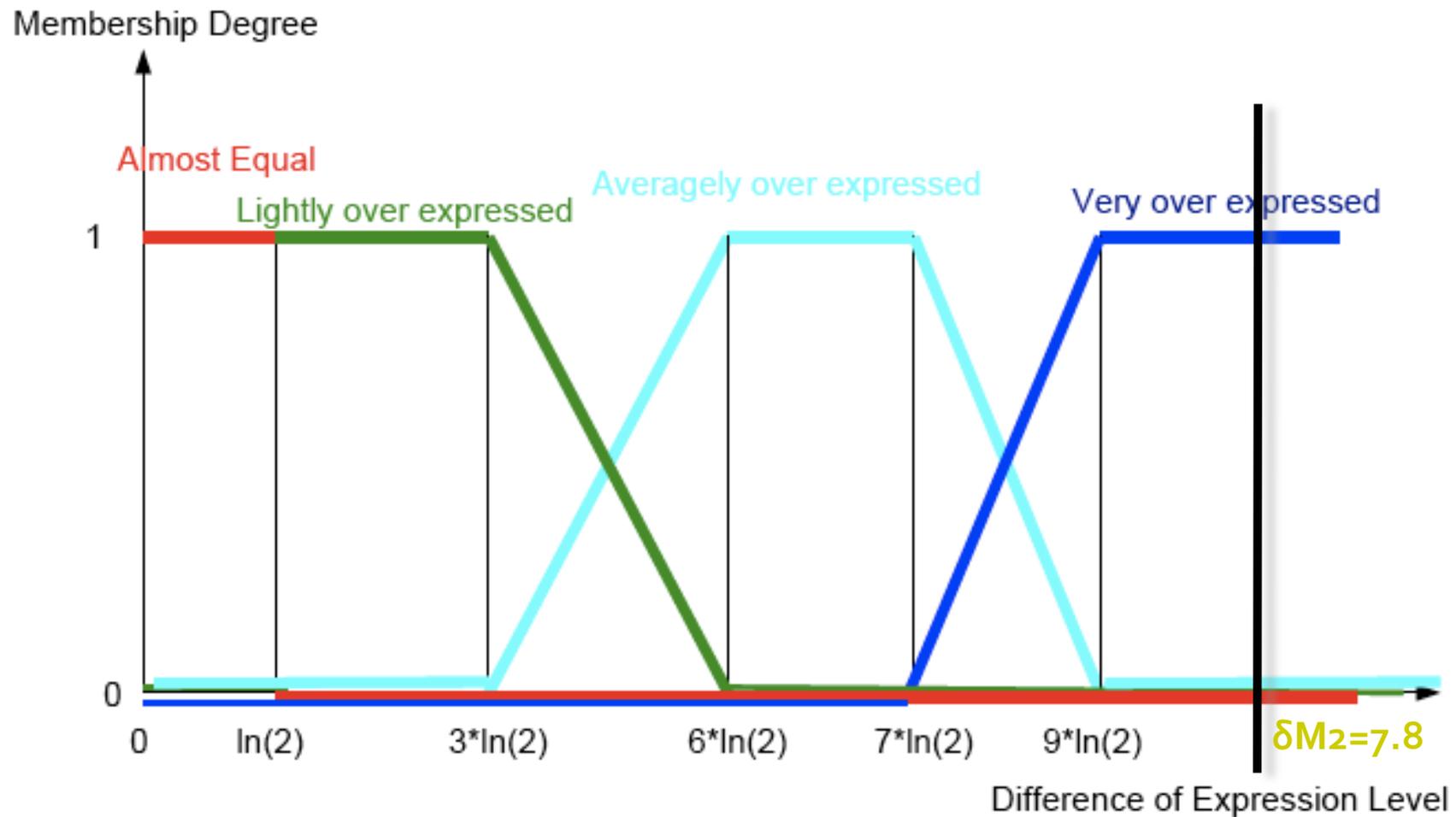
Recherche des motifs à écarts flous



Recherche des motifs à écarts flous



Recherche des motifs à écarts flous



Recherche des motifs à écarts flous

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

Degré d'une séquence à écart flou :

$$F_{SFG}(M) = \overline{\overline{\top}} (d_1, \dots, d_{n-1})$$

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

Support d'une séquence à écart flou

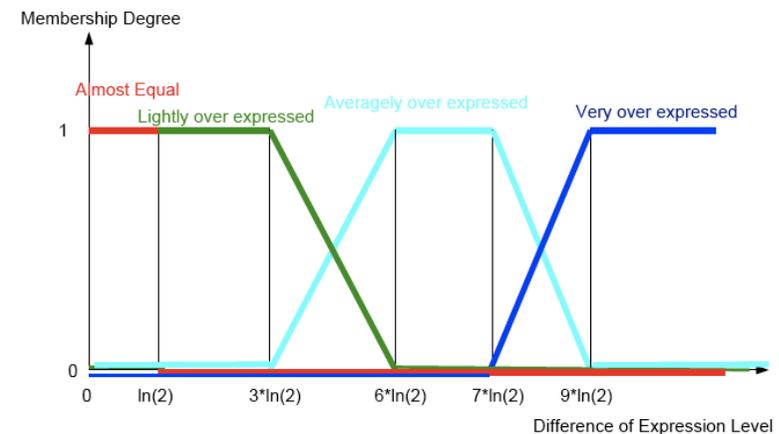
$$Freq(s_{FG}) = \frac{\sum_{o \in \mathcal{O}} [F_{s_{FG}}(s_o)]}{|\mathcal{O}|}$$

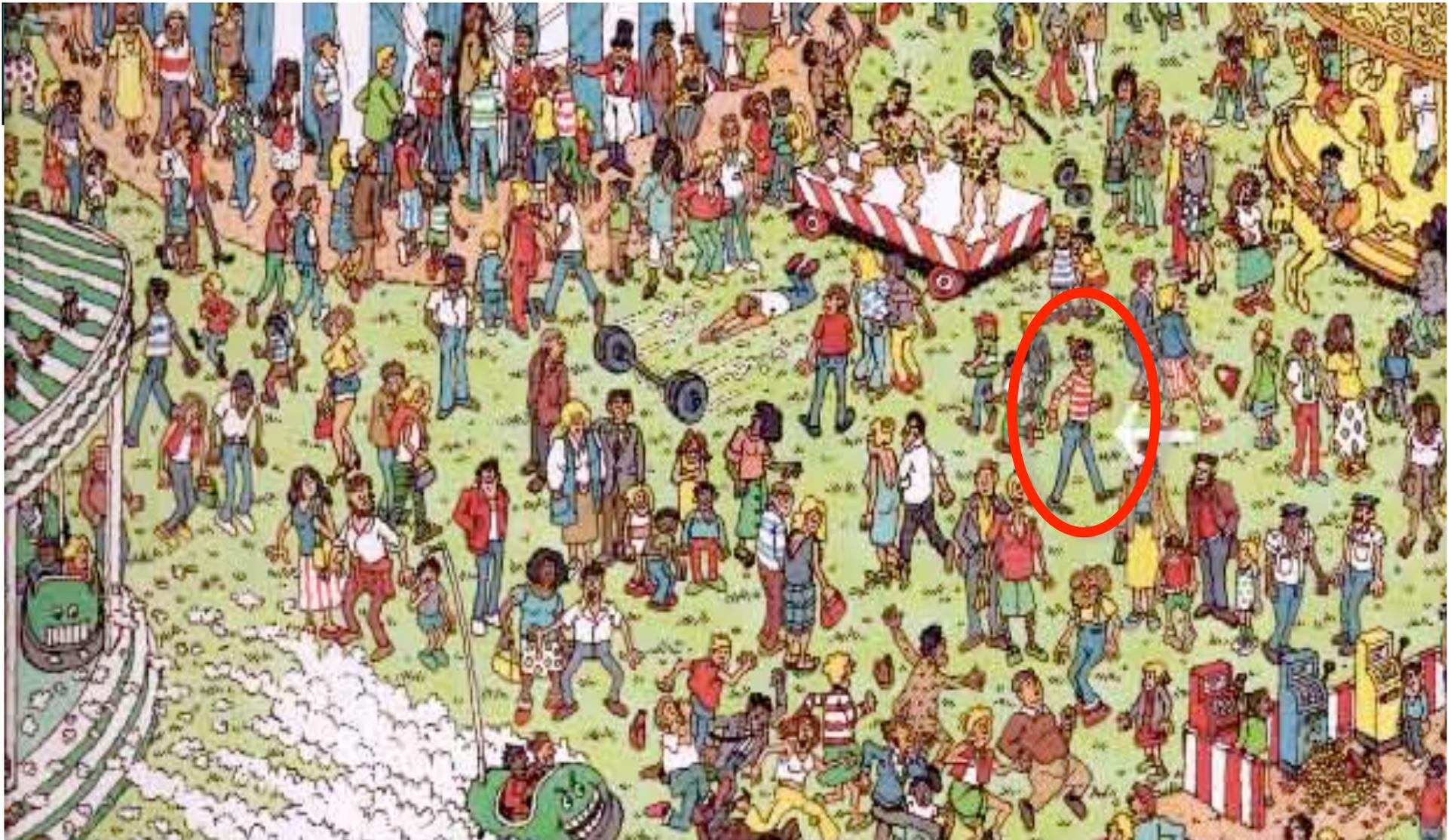
Pourcentage de puces vérifiant le motif à écarts flous

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

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





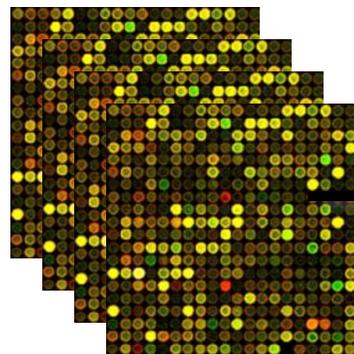
Séquences plus facilement interprétables



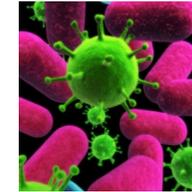
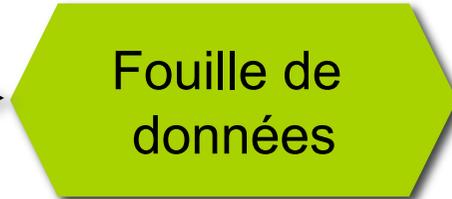
... utiliser les motifs pour prédire le
grade d'un cancer

Définition du problème

Phase 1 : apprentissage du modèle



Données



Grade 1



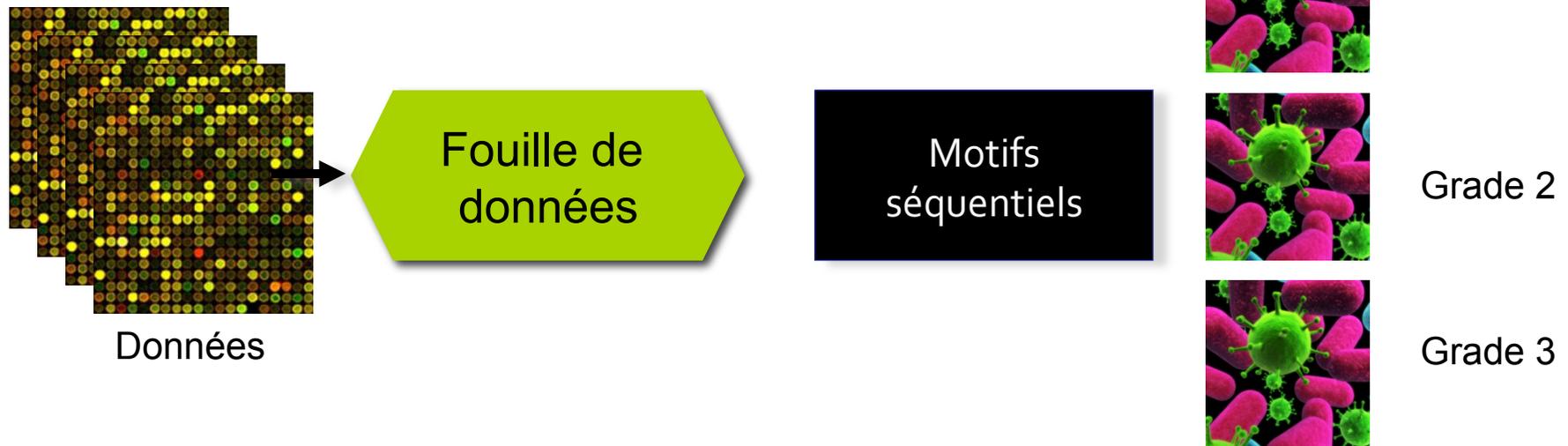
Grade 2



Grade 3

Définition du problème

Phase 1 : apprentissage du modèle



Phase 2 : prédiction du grade



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**

$S1 = \langle (G4) (G5 G6) (G3) (G1) (G2) \rangle$

- **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

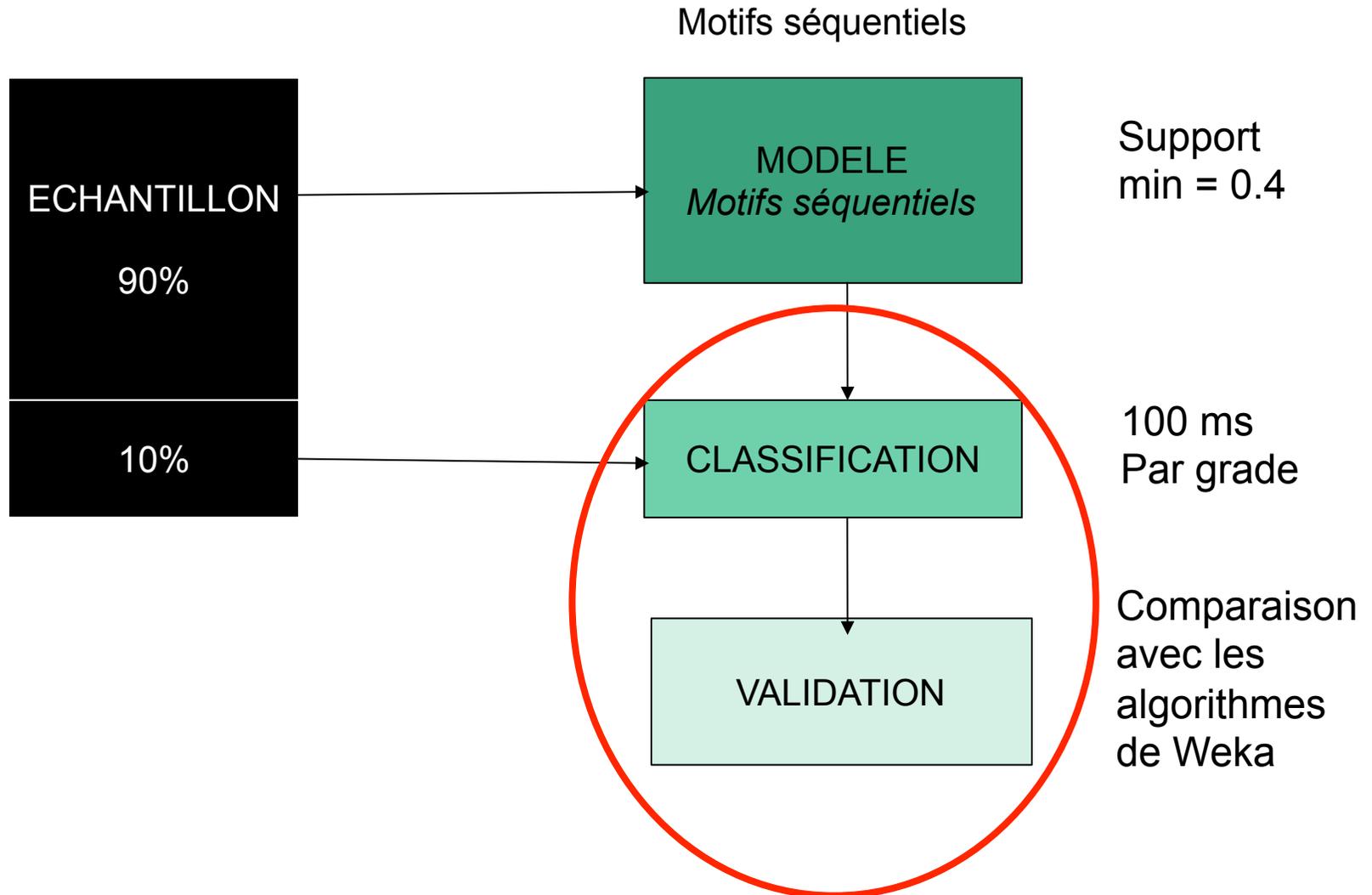
Données

- **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



Evaluation

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

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

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 épidémiologique

➔ Données environnementale

(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
épidémies



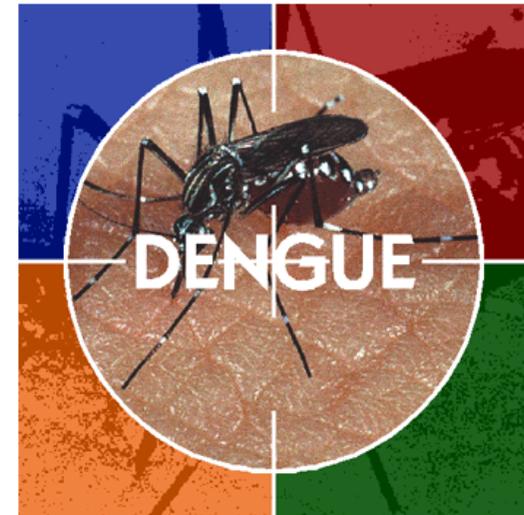
Health
Professionals

[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. Alatrasta-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

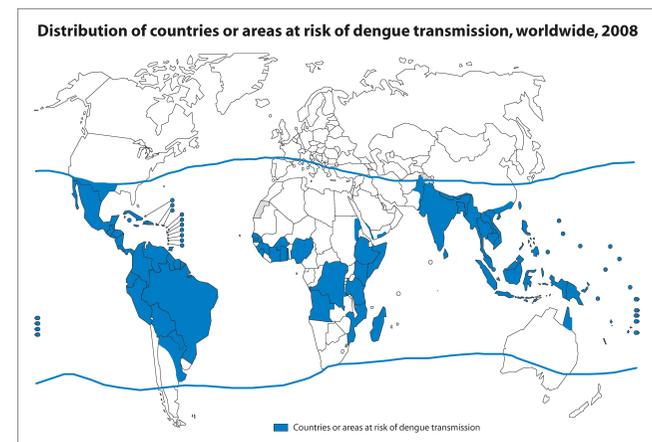
Quelques définitions - épidémiologie

- **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

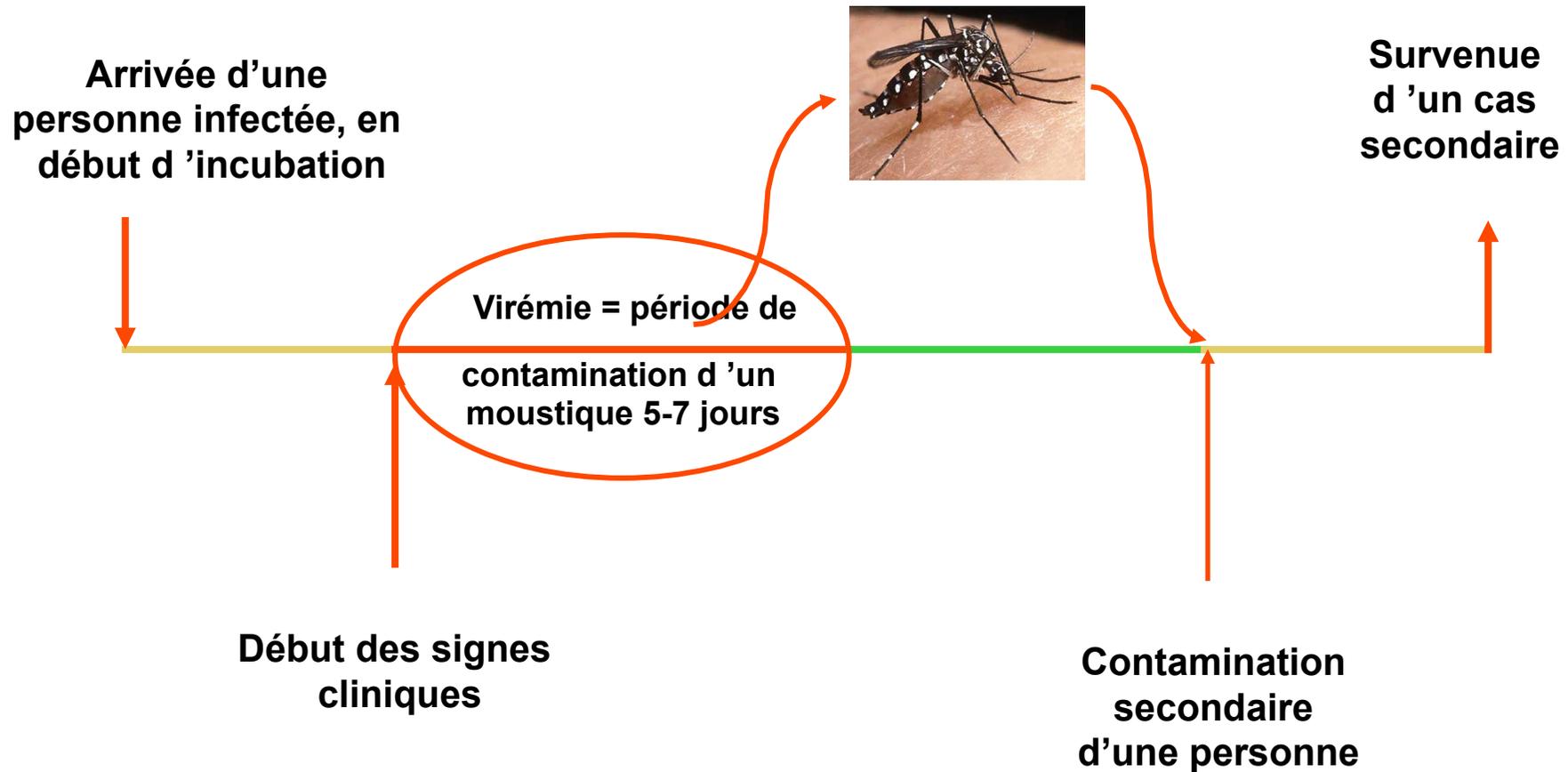
- **Grippe tropicale** transmise à l'homme par le **moustique Aedes aegypti** ou **Aedes albopictus**
- **Arbovirose** (transmise par les suceurs de sang) **n°1 chez l'Homme**
 - 50 à 100 millions de personnes affectées/an, dans le monde
 - 200 000 à 500 000 hospitalisations DFH (Dengue à Fièvre Hémorragique) /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.**



The boundaries and names shown on this map do not imply the expression of any opinion whatsoever on the part of the World Health Organization concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted lines on maps represent approximate border lines for which there may not yet be full agreement. © WHO 2010. All rights reserved.

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



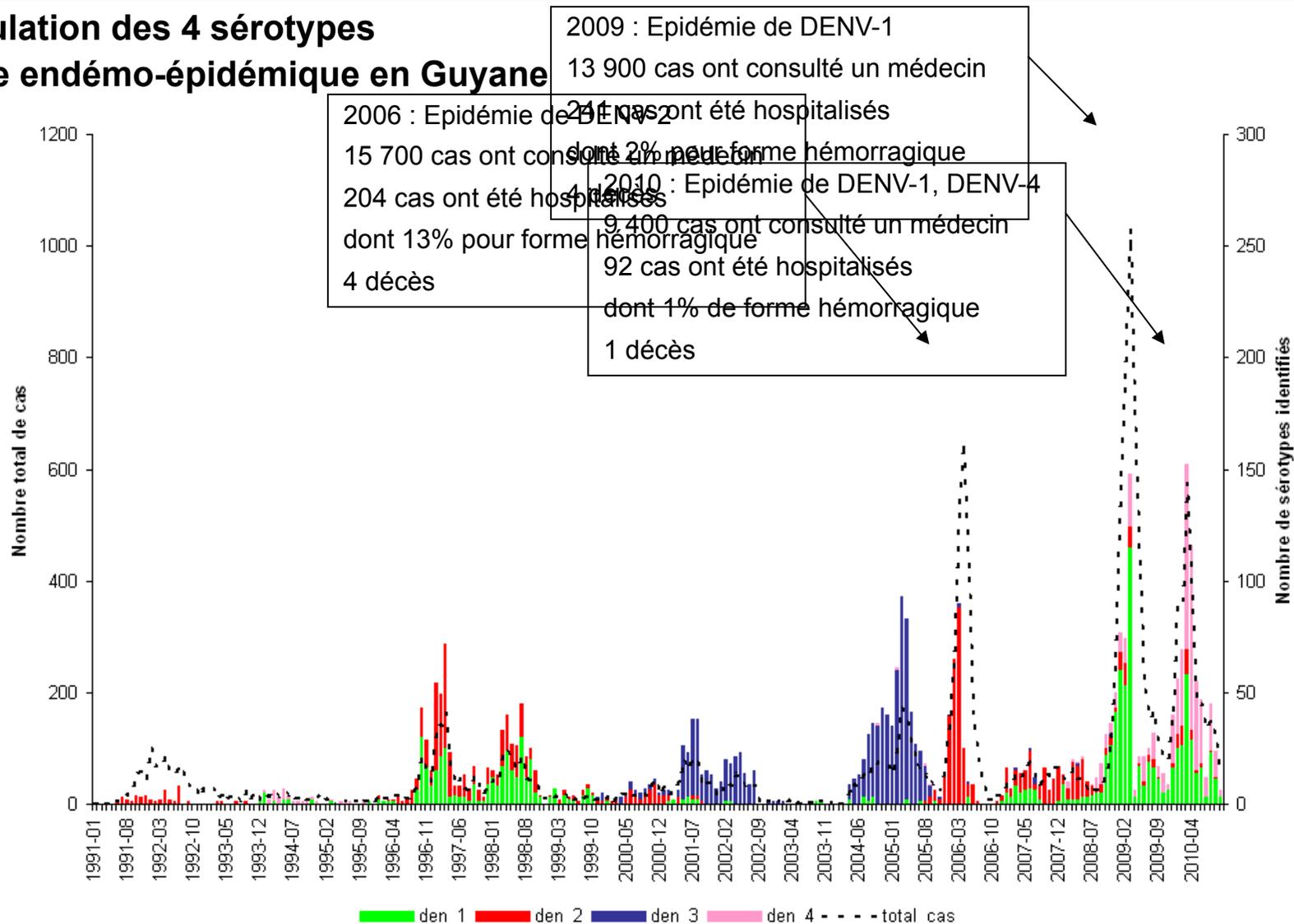
▪ **Forme hémorragique**

- 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.



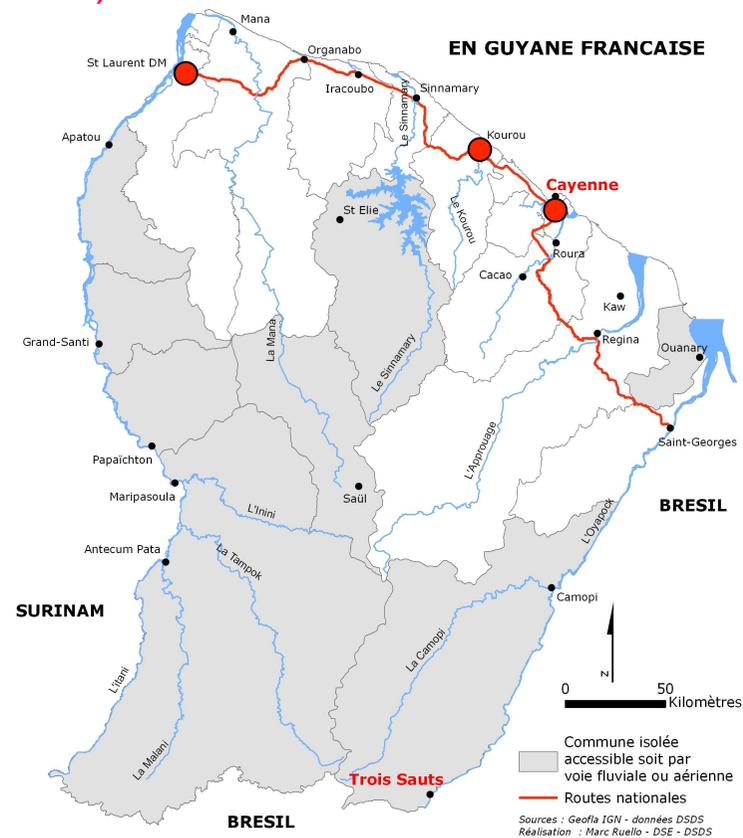
La dengue en Guyane

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



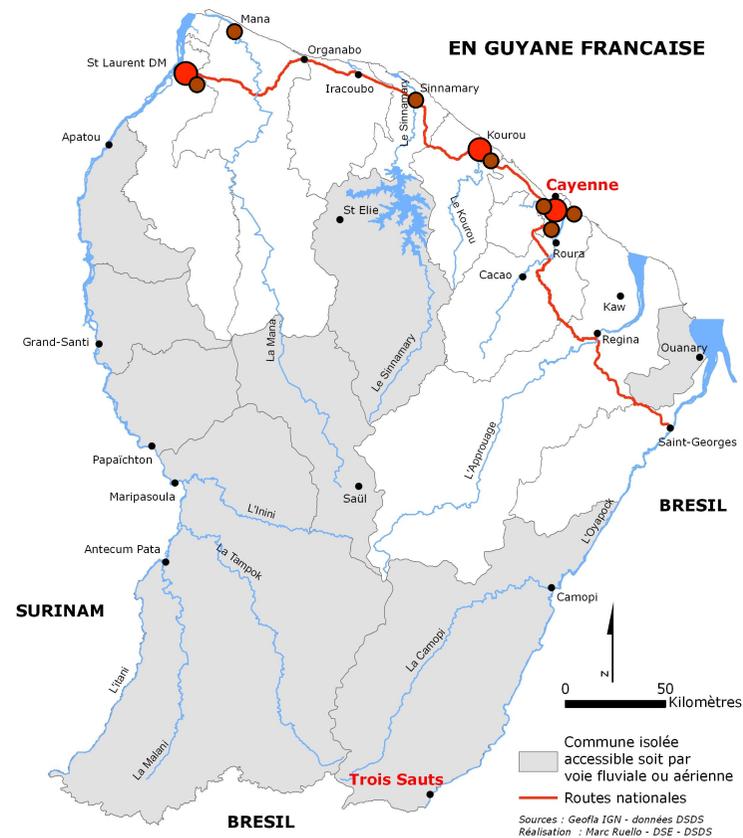
Sources des données

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



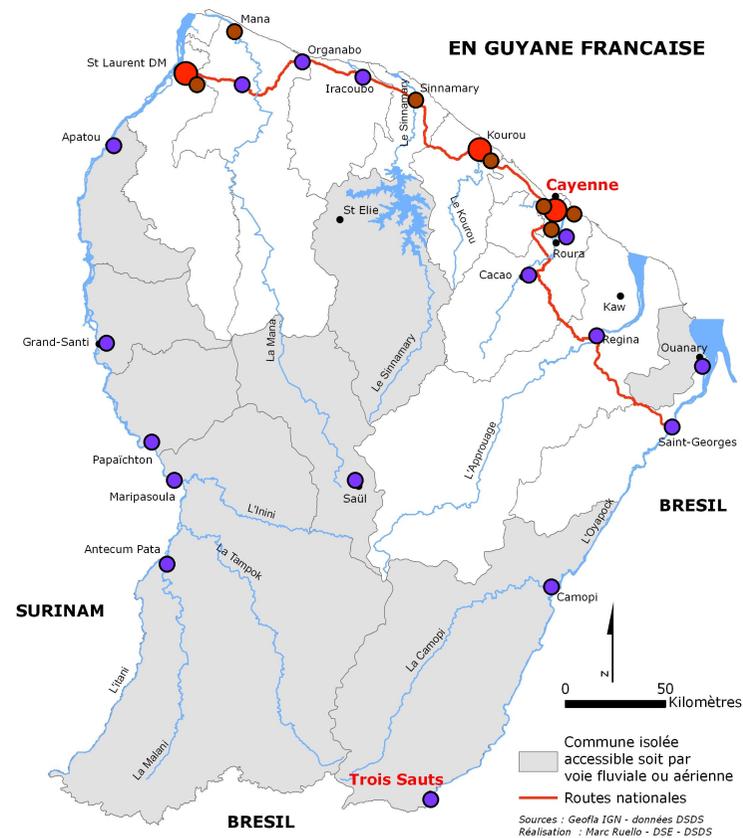
Sources des données

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

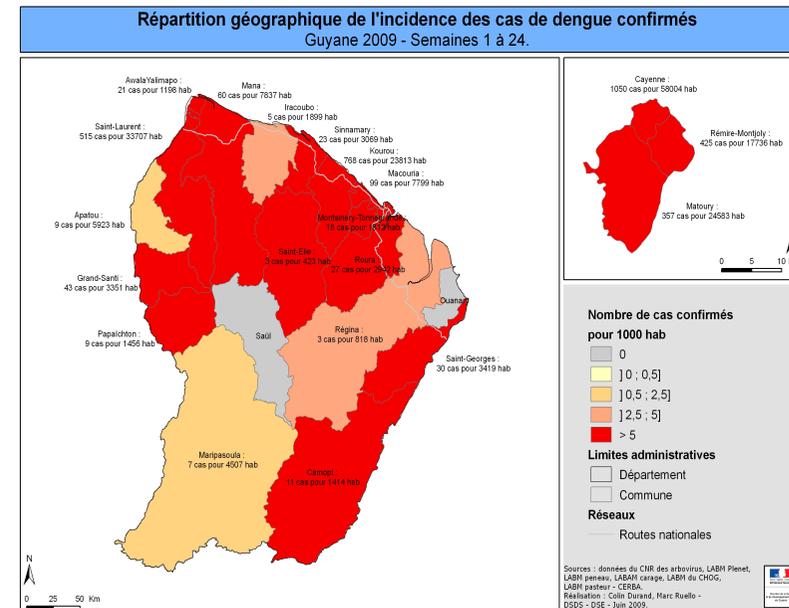
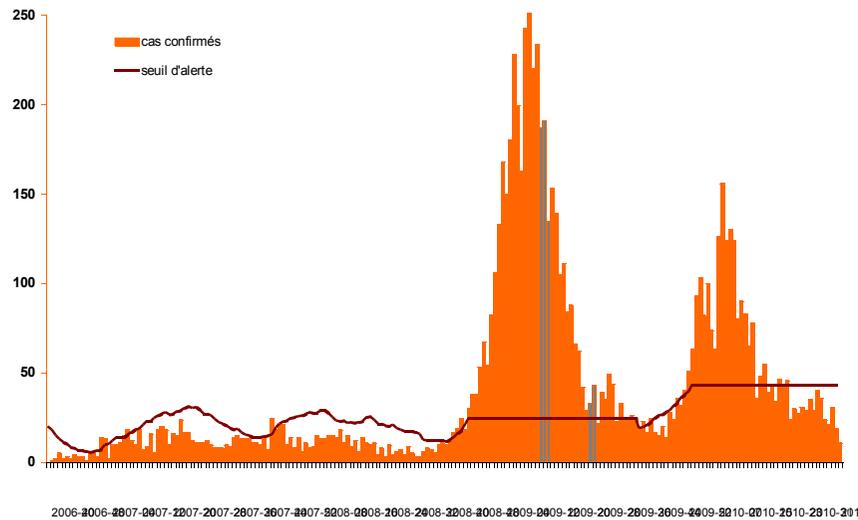


Sources des données

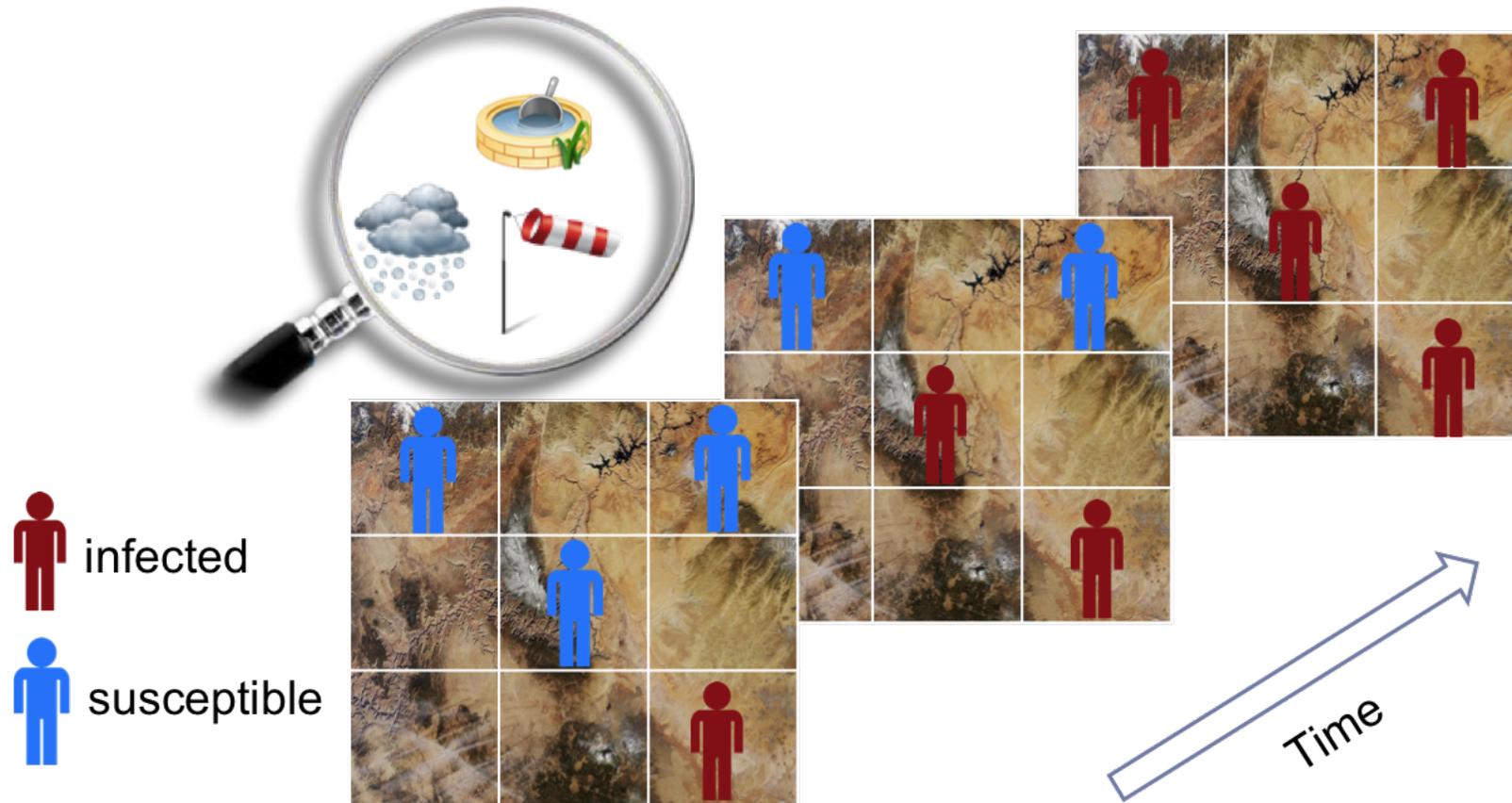
- 17 Centres et Postes de santé (cas cliniquement évocateurs)



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

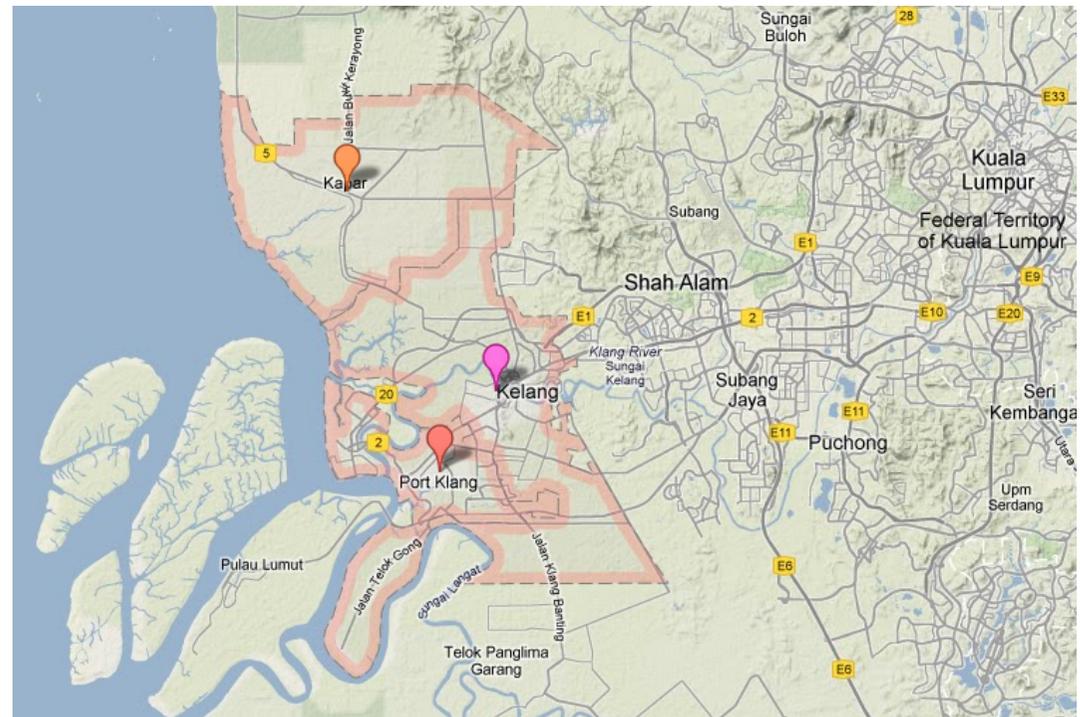
On groupe les dimensions d'analyse par date et zone pour obtenir une **base de séquences sDB**

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_-) <u>(temp_low rain_low wind_23)</u>

itemset

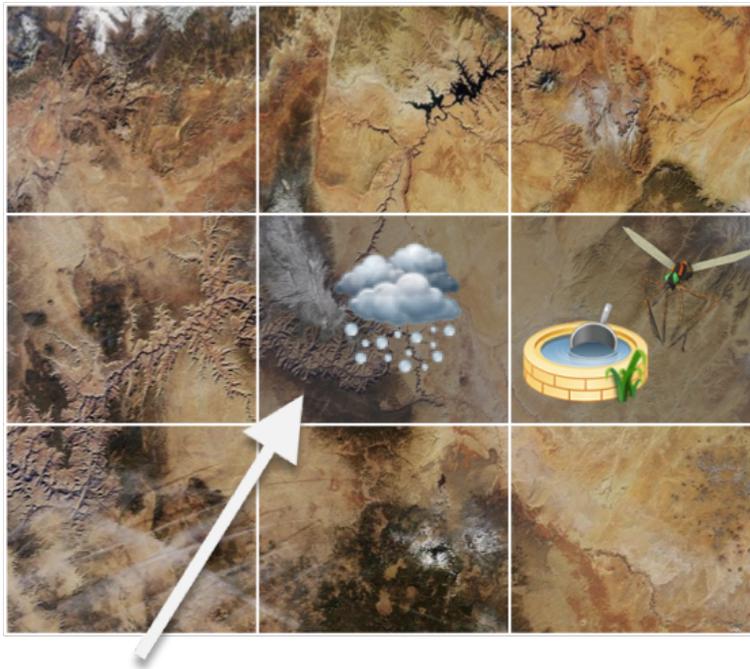
Zonage

area 1	area 2
Port Klang Kelang	Kelang Kapar



Itemset spatial

2 items spatialement proches



Dynamique spatiale

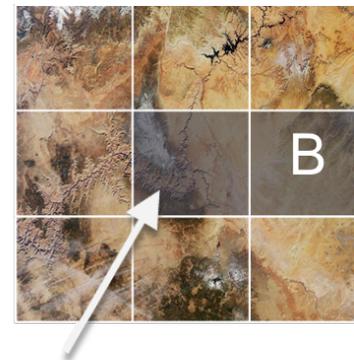
Opérateur spatial • (près de)

Opérateur de groupage []

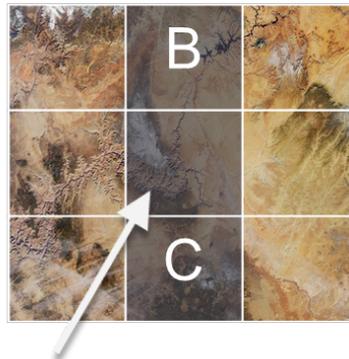
Symbole représentant l'absence d'itemsets θ



$(A \bullet B)$



$(\theta \bullet B)$

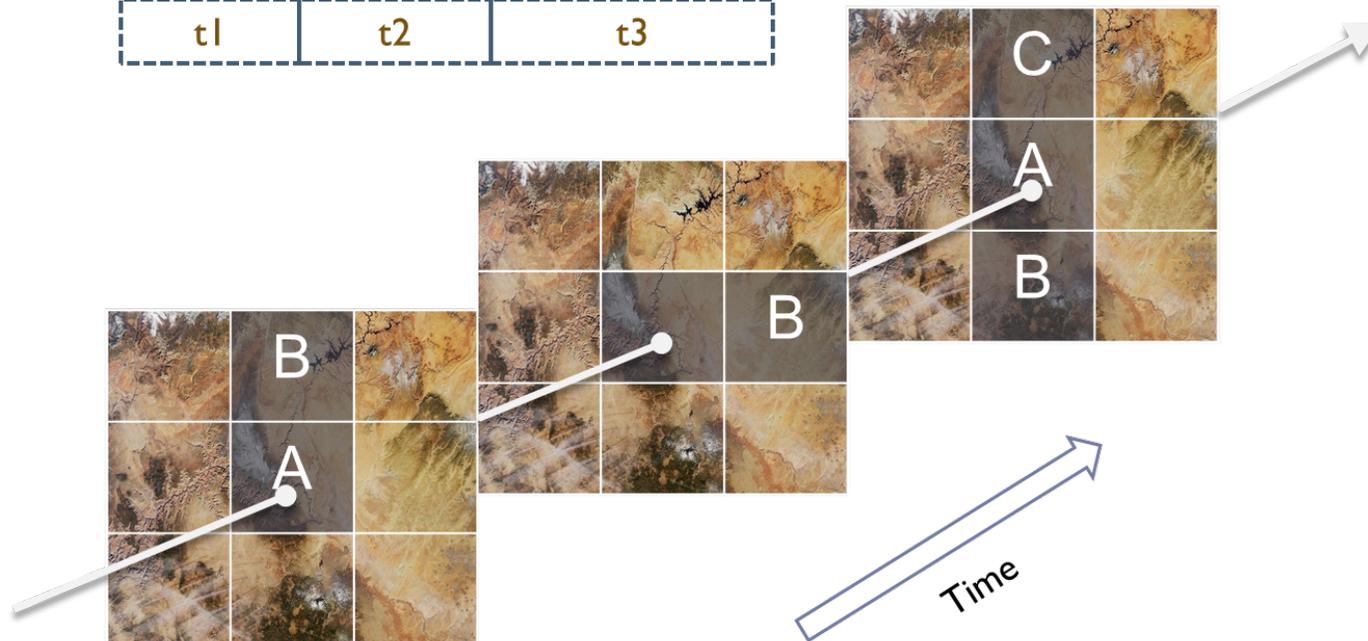


$(\theta \bullet [B ; C])$

Spatial Sequences 2S

Séquence d'itemsets spatiaux

$\langle (A \bullet B) (\theta \bullet B) (A \bullet [B ; C]) \rangle$



Fouille de séquence 2S

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

$$\text{support}(2S) \geq \alpha$$

Avec α le support minimum

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

Support

area	sequence of itemsets
Port Klang	$\langle (\text{temp_low } \text{rain_low } \text{wind_35}) (\text{temp_low } \text{rain_high } \text{wind_}) \rangle$
Kelang	$\langle (\text{temp_low } \text{rain_mean } \text{wind_12}) (\text{temp_high } \text{rain_mean } \text{wind_}) \rangle$
Kapar	$\langle (\text{temp_low } \text{rain_} \text{wind_}) (\text{temp_low } \text{rain_low } \text{wind_23}) \rangle$

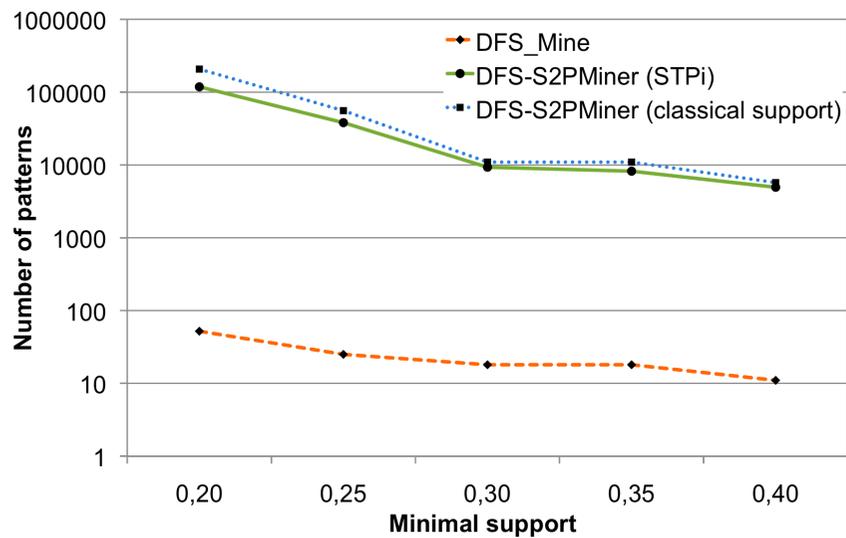
support ($\langle (\text{rain_low } \text{rain_high}) \rangle$) = $\frac{2}{3}$, the sequence appears for 66% of areas

support ($\langle (\text{temp_low}) \rangle$) = 1, the sequence appears in all areas

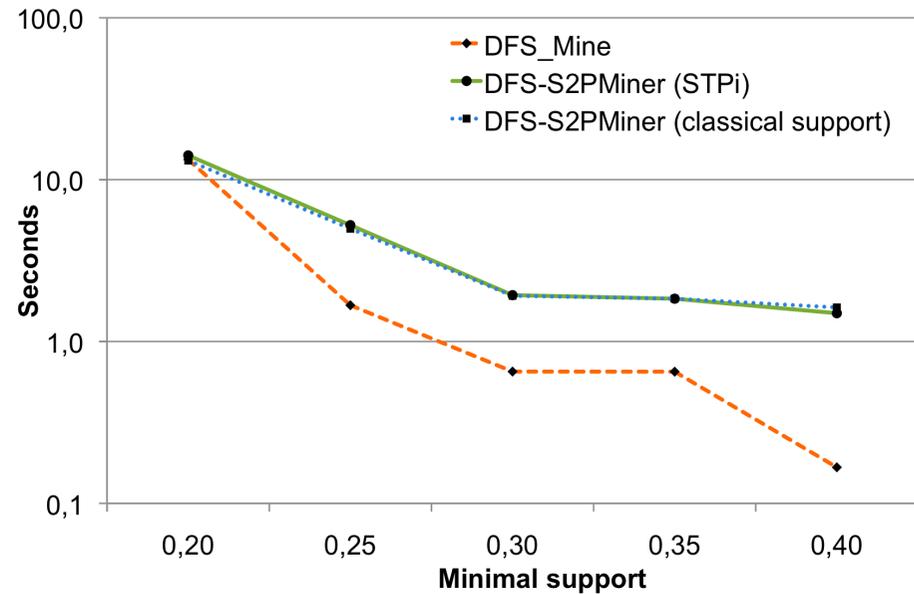
! *The support uses only the spatial component of dataset*

Evaluation

Number of patterns extracted by DFS_Mine and DFS-S2PMiner on Dengue dataset



Execution runtime of DFS_Mine and DFS-S2PMiner on Dengue dataset



... 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 Cytometry

Clustering of
rare cells

Analysis

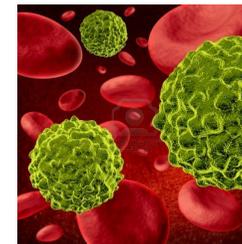
→ Cluster of different kinds
of cells

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

STREAM



[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.

Trends of Health News on Tweets

Objectives : Trend Detections on Social Networks and Tweets



Text Cubes

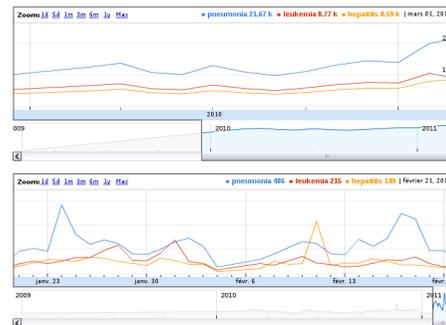
- Stream of Tweets
- MeSH Hierarchy

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.

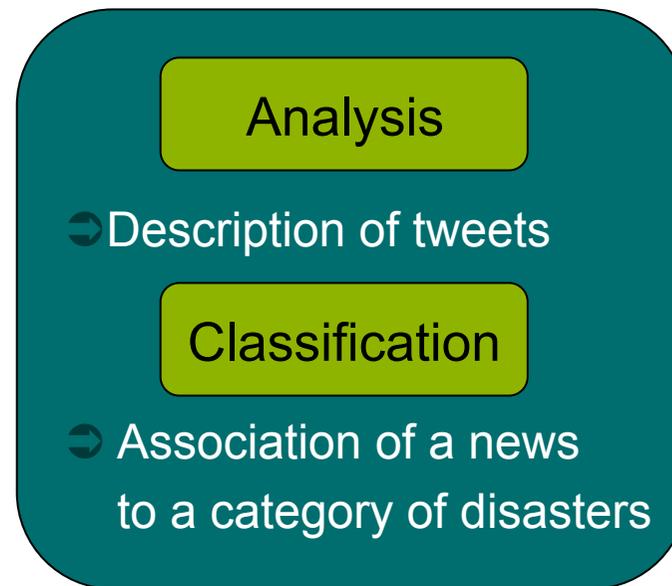
Detection of Natural Disasters

Partners: WebReport



Textual
Patterns

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



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), 2010.112