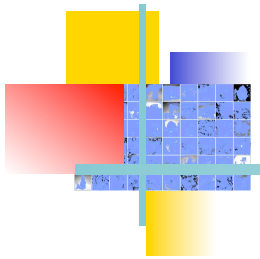
Utilisation et propriétés d'un ***précodage des données*** *par* ***motifs fréquents*** *en* ***classification supervisée***

A. Cornuéjols, S. Jouteau, M. Sebag (LRI)

Ph. Tarroux (LIMSI)

CNRS - Université de Paris-Sud, Orsay

Inra.



Motivation

Le problème étudié :

identifier des régularités dans des données en très grandes dimensions

Données en grandes dimensions

- **Définies par un très grand nombre d'attributs**

  (Note : l'un des 10 pbs soulevés lors mathématiques en 2000)

- **Exemples :**
  - Puces ADN
    - E.g. 6400 gènes,
      - ➥ organismes sains ou irradiés
  - Images
    - E.g. 256*256*(256 niveaux de gris)
      - ➥ Formes présentes dans l'image

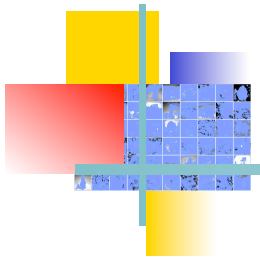
L'objectif

➥ *Identifier des régularités dans des données de très grandes dimensions*

- **Apprentissage supervisé multi-classes**

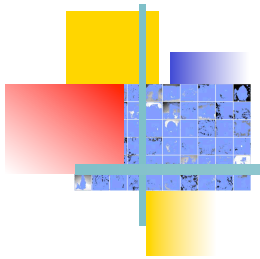
- Beaucoup de dimensions + peu d'exemples

  = Difficulté pour distinguer vraies régularités et coïncidences



Prétraitements

- **Réduction de dimension**
  - Sélection d'attributs
    - Élimination des redondances (ACP, …)
    - Recherche de corrélations
  - Modélisation : hypothèses sur la statistique du signal
    - Analyse de Fourrier
    - Analyse en ondelettes



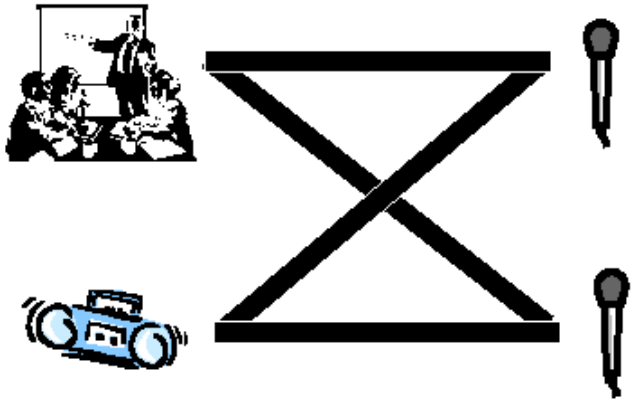
Cas de l'analyse de scènes

- Scènes naturelles ≠ scènes artificielles
- Observation neurobiologique : **codage clairsemé**
- Hypothèse : signal résultant d'une **superposition de « formes latentes »**
- ***Analyse en composantes indépendantes*** (ACI)

L'analyse en composantes indépendantes

(Introduite en 1984. Développée dans les 90s)

Hyp. de base : *les données résultent d'une combinaison linéaire de formes latentes*



➥**Recherche de ces formes latentes**

- **Mais :**
  - Inapplicable en grande dimension
  - Hypothèse de linéarité

L'ACI en analyse de scènes

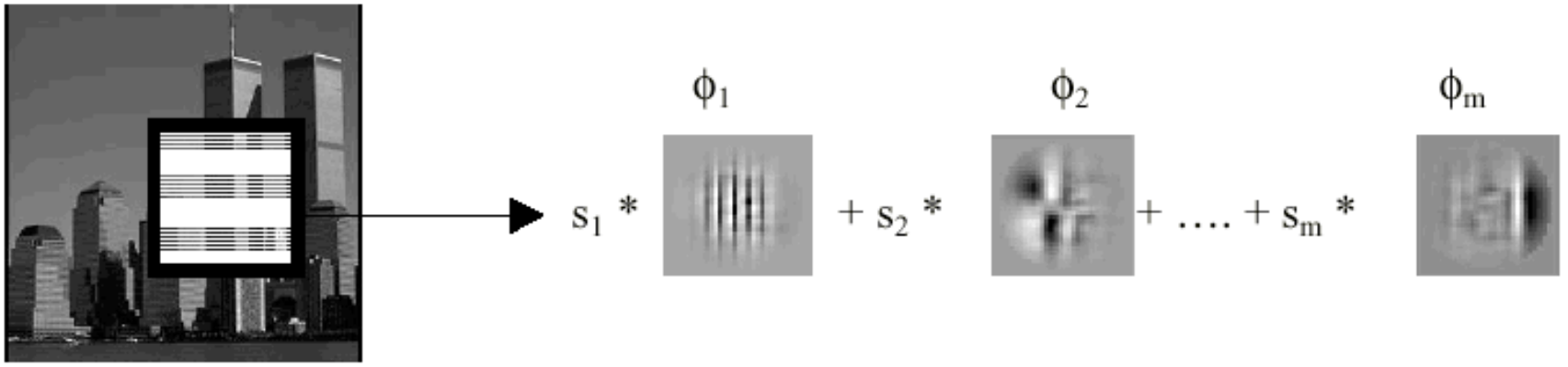
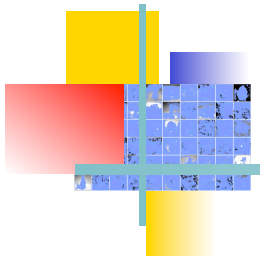


Figure 1 : Illustration de la décomposition d'une imagette dans la base Φ.

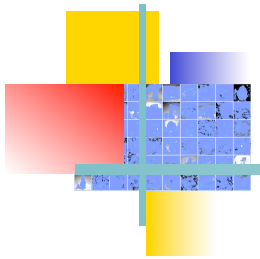
- **Les scènes sont décomposées en imagettes …**
- **… codées par des superpositions linéaires de formes latentes**



Le projet

- ***Peut-on rechercher directement un codage clairsemé ?***

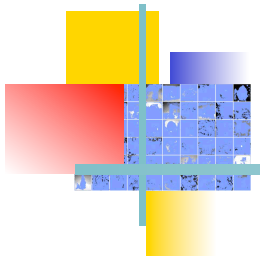
- **Idée : adapter des techniques de fouilles de données**



Les motifs fréquents

- **Le problème**
  - Étant donné une base de données consistant en tuples,
    trouver des règles d'association prédisant avec confiance quels items se trouvent souvent ensemble (Frequent ItemSets)
- **Exemple canonique (mais mythique)**
  - Les caddys dans les supermachés
  - Un tuple = ensemble d'items achetés ensemble
- **En général :**
  - Beaucoup de motifs fréquents
  - Mais peu qui soient vérifiés ensemble

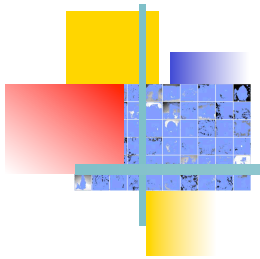
➥ **Codage clairsemé**



L'algorithme APRIORI

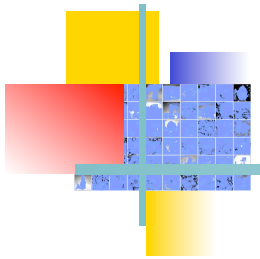
- Se donner un **taux de couverture** : ε
- Commencer avec les 1-itemsets de couverture $\geq \varepsilon$
- À chaque étape k :
  - Ajouter au k-itemsets un 1-itemset de couverture $\geq \varepsilon$
  - Éliminer les itemsets de couverture $< \varepsilon$ pour obtenir l'ensemble des (k+1)-itemsets

\+ **un usage astucieux de tables de hachage**



Adaptation de l'algorithme

- **Représentativité**
  - Chaque image correspond à un nombre suffisant de motifs
- **Codage clairsemé**
  - Chaque image correspond à un nombre limité de motifs
- **Orthogonalité des motifs**
  - Chaque couple de motifs a peu d'images en commun
- **+ Contraintes sémantiques**
  - E.g. : motifs connexes
  - E.g. : motifs en ligne
  - …



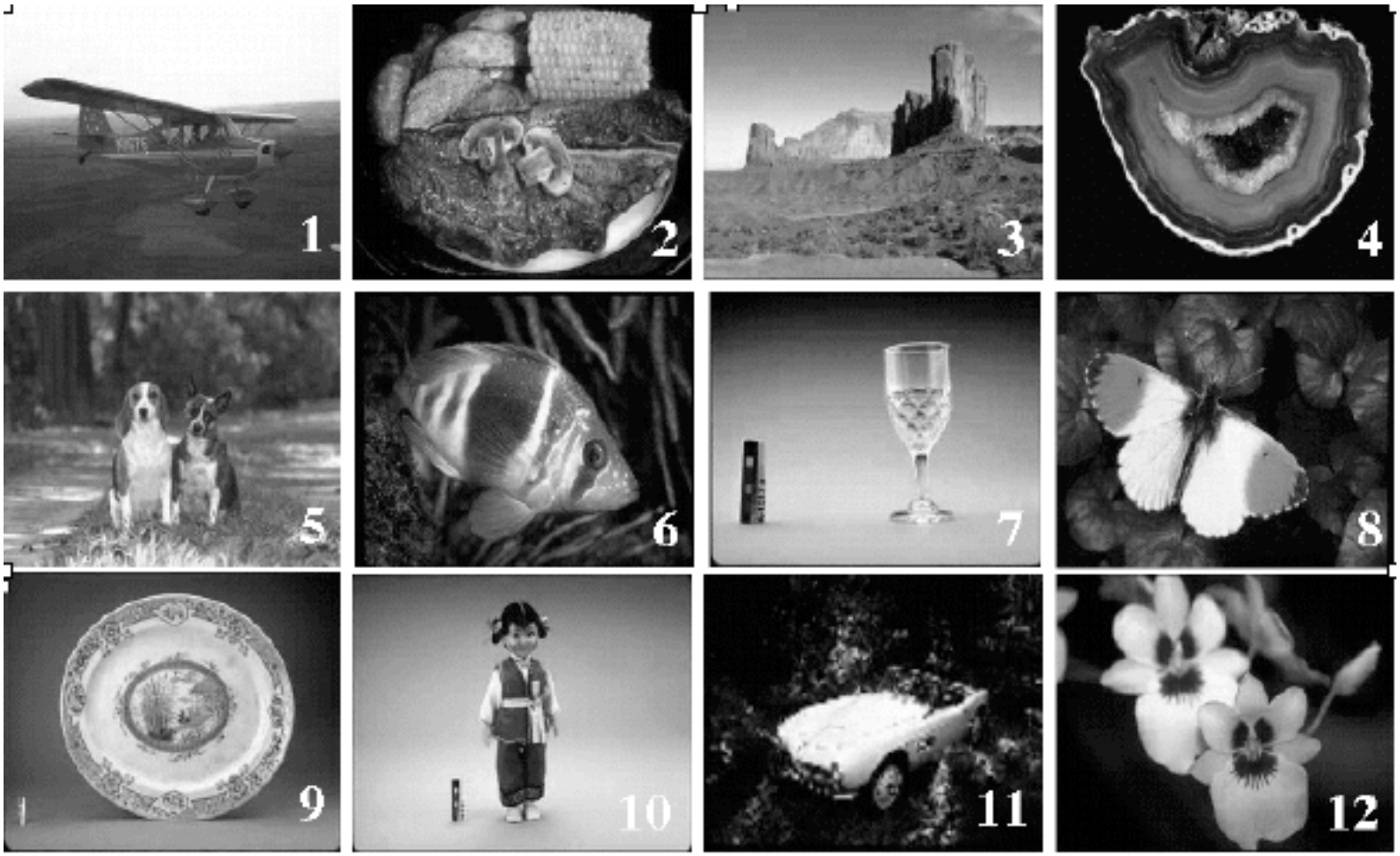
Les expériences

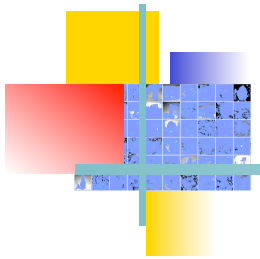
- Base d'images tirée de la base COREL
- **12 classes** différentes de scènes
- Base de **1080 images** (90 images / classe)
- 128 x 128 = 16384 en 128 niveaux de gris

  ou : 64 x 64 = 4096 en 32 ou 16 niveaux de gris

- On utilise **540 images** pour chercher **1000 motifs fréquents**

La base d'images





Constat

- L'application directe de APRIORI est impossible
- Il y a trop de motifs fréquents

Nb. élts / motif	**1**	**2**	**3**	**4**	**5**	**6**
Nb motifs	$2\ 10^3$	$110\ 10^3$	$3{,}8\ 10^6$	$80\ 10^6$	$1{,}15\ 10^9$	$12{,}5\ 10^9$

Pour images 32 x 32 en 64 niveaux de gris

➥ ***Il faut adapter l'algorithme et faire une recherche stochastique et non plus exhaustive***

Adaptation de l'algorithme

Recherche itérative et stochastique de motifs fréquents

- Paramètres : taux de couverture ε. Nombre de motifs cherchés = N
- Nombre de motifs trouvés = n

Tant que $n \leq N$ **faire**

Choix dans un exemple x_i encore peu couvert d'un premier atome a_0 présent dans au moins ε des exemples

motif <- a_0

Tant que taux de couverture de *motif* > ε **faire**

Tirer au hasard un atome a de x_i couvrant au moins ε des exemples et peu utilisé dans les motifs existants et satisfaisant les contraintes sémantiques

Si *motif*+a couvre au moins ε des exemple **alors**

motif <- *motif* +a

fin si

Fin tant que

Fin tant que

Codage clairsemé

Nb de FIS / images

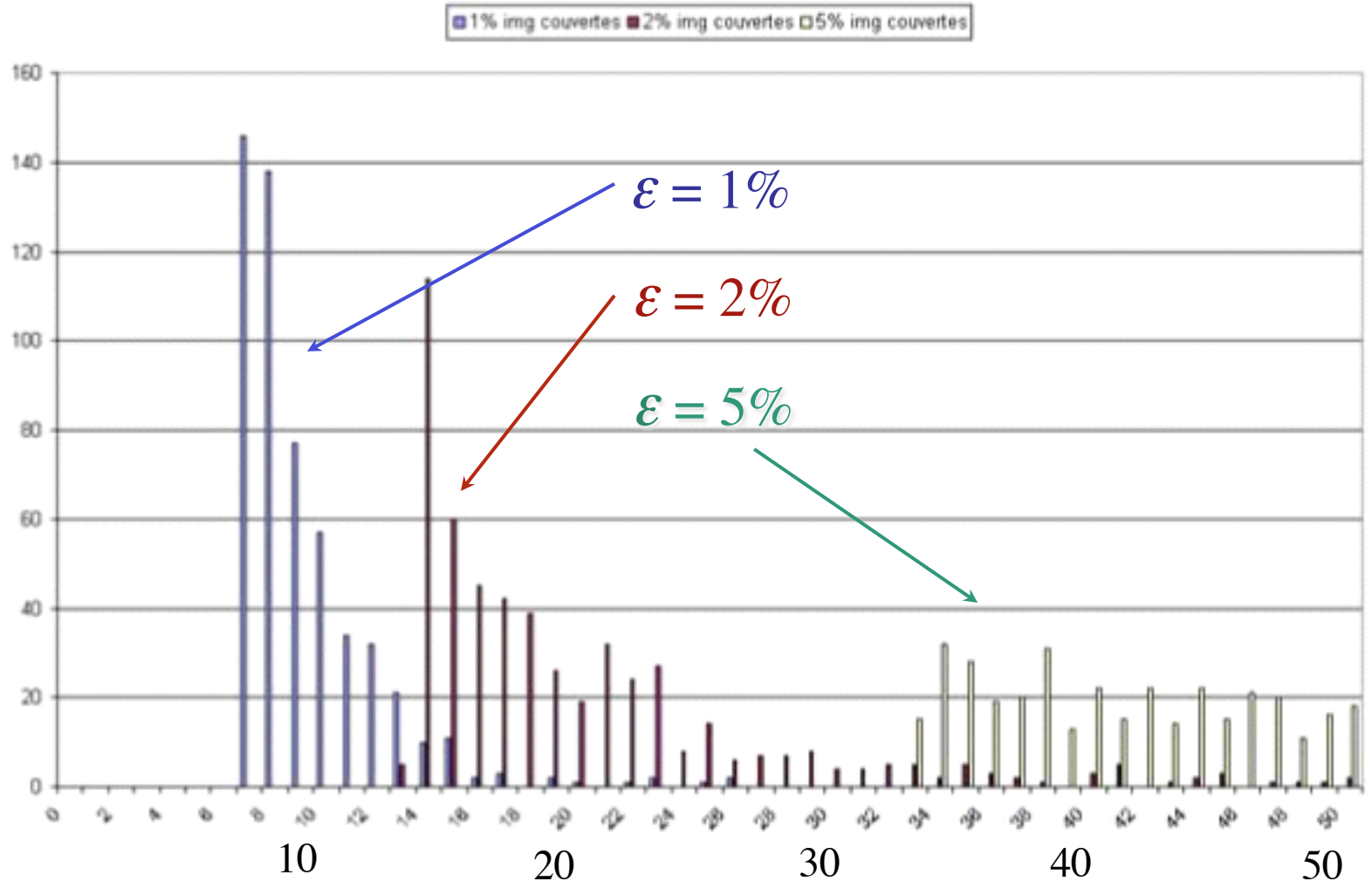


FIG. 6.22 – Histogramme représentant le nombre d'images (en ordonnée) activant N FIS (en abscisse).Ces fonctions ont été calculées à partir d'images de taille 64x64 en 16 niveaux de gris. La troisième méthode de choix des pixels (pixels se touchant) est utilisée. Un bon histogramme a le moins de valeurs possibles (très "groupé".)

Codage clairsemé

Nb de FIS / images

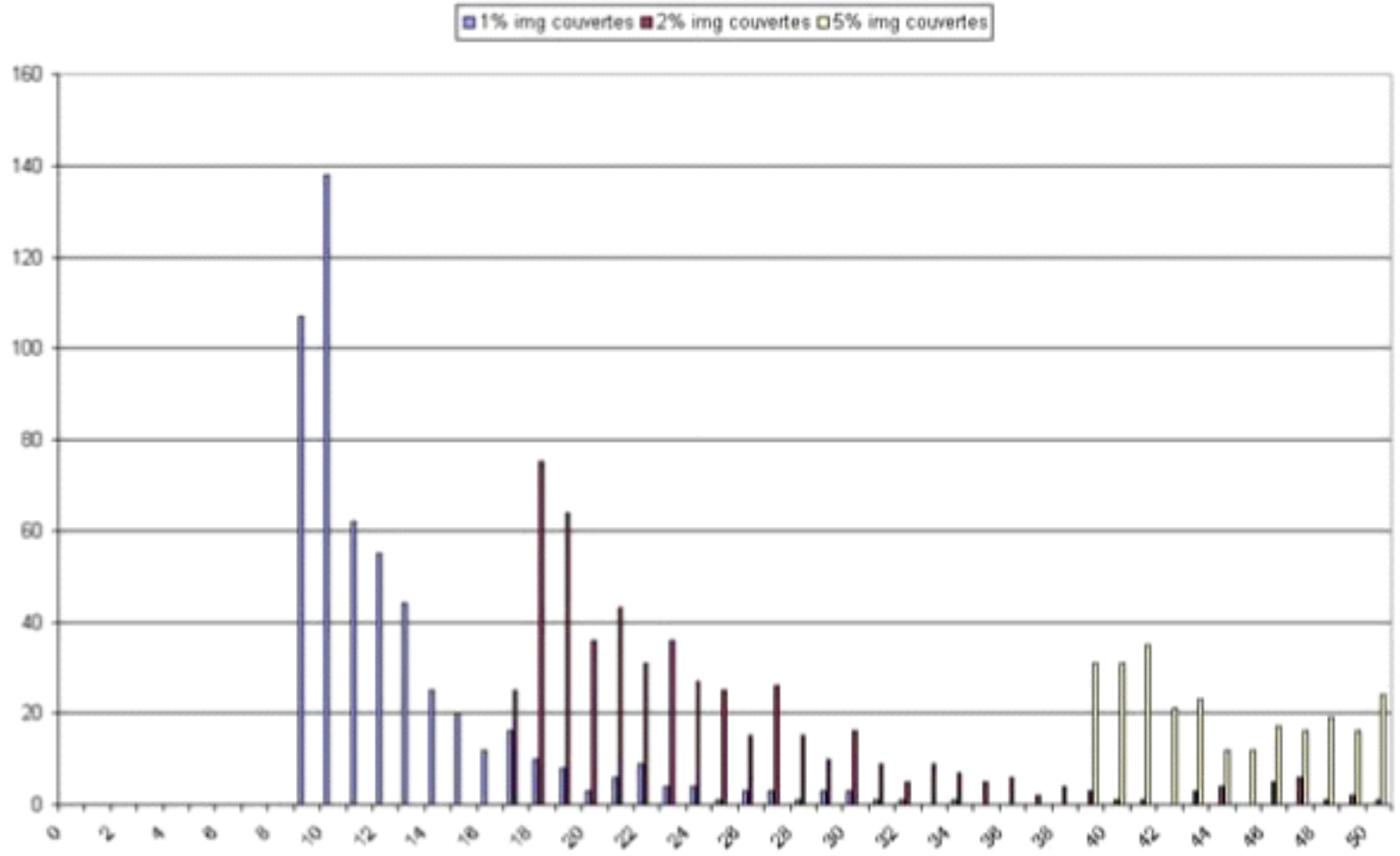


FIG. 6.23 – Histogramme représentant le nombre d'images (en ordonnée) activant N FIS (en abscisse).Ces fonctions ont été calculées à partir d'images de taille 64x64 en 16 niveaux de gris. La quatrième méthode de choix des pixels (pixels formant une ligne) est utilisée. Un bon histogramme a le moins de valeurs possibles (très "groupé".)

Orthogonalité

Nb images par couple de motifs

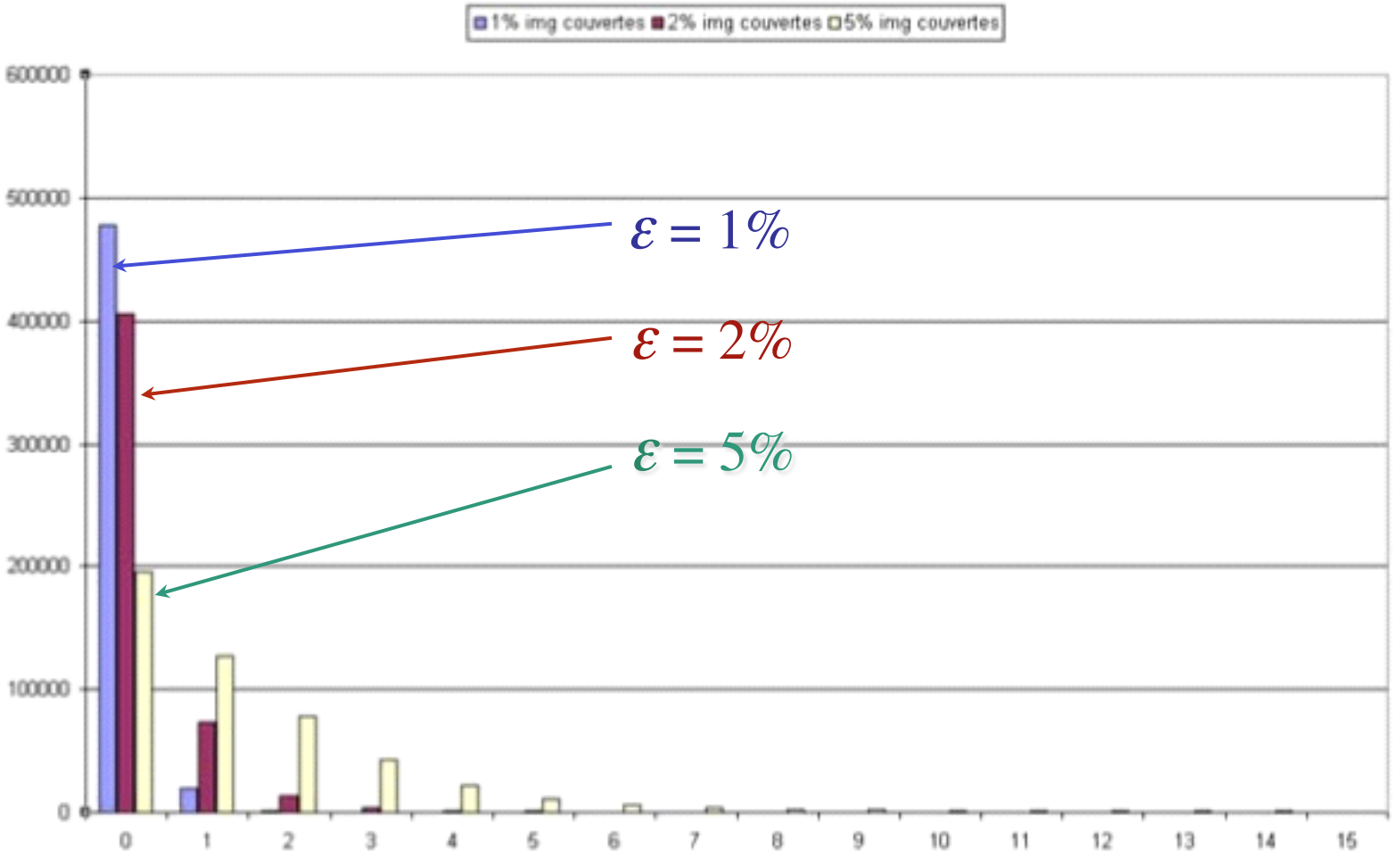
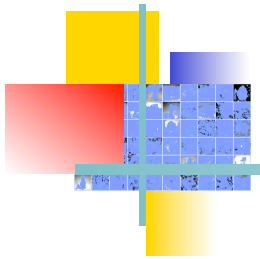


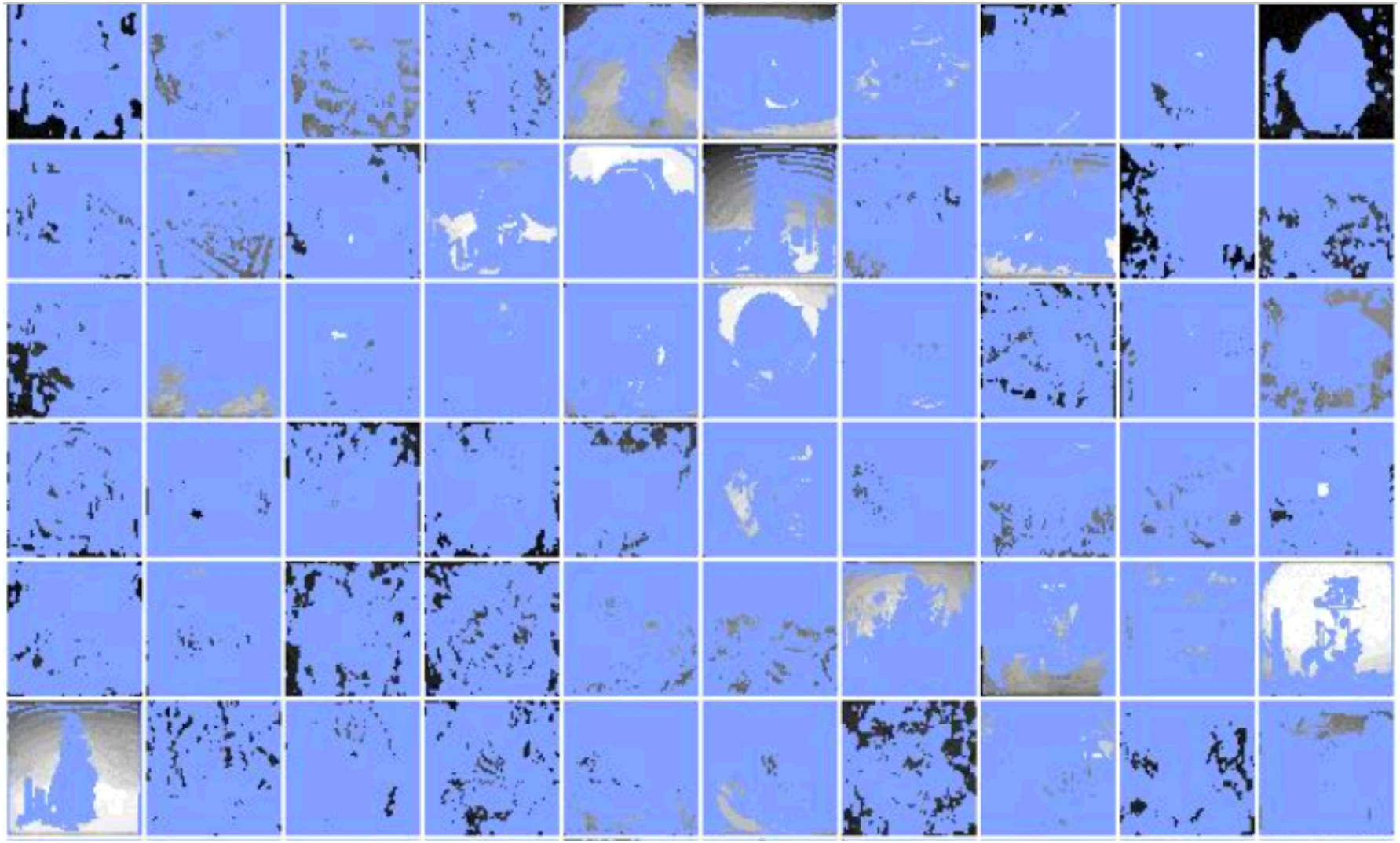
FIG. 6.18 – Histogramme représentant le nombre de couples de fonctions (en ordonnée) ayant N images en commun (en ascisse). Ces fonctions ont été calculées à partir d'images de taille 64x64 en 16 niveaux de gris. La troisième méthode de choix des pixels (pixels se touchant) est utilisée. Un bon histogramme a le moins possible de valeurs élevées.



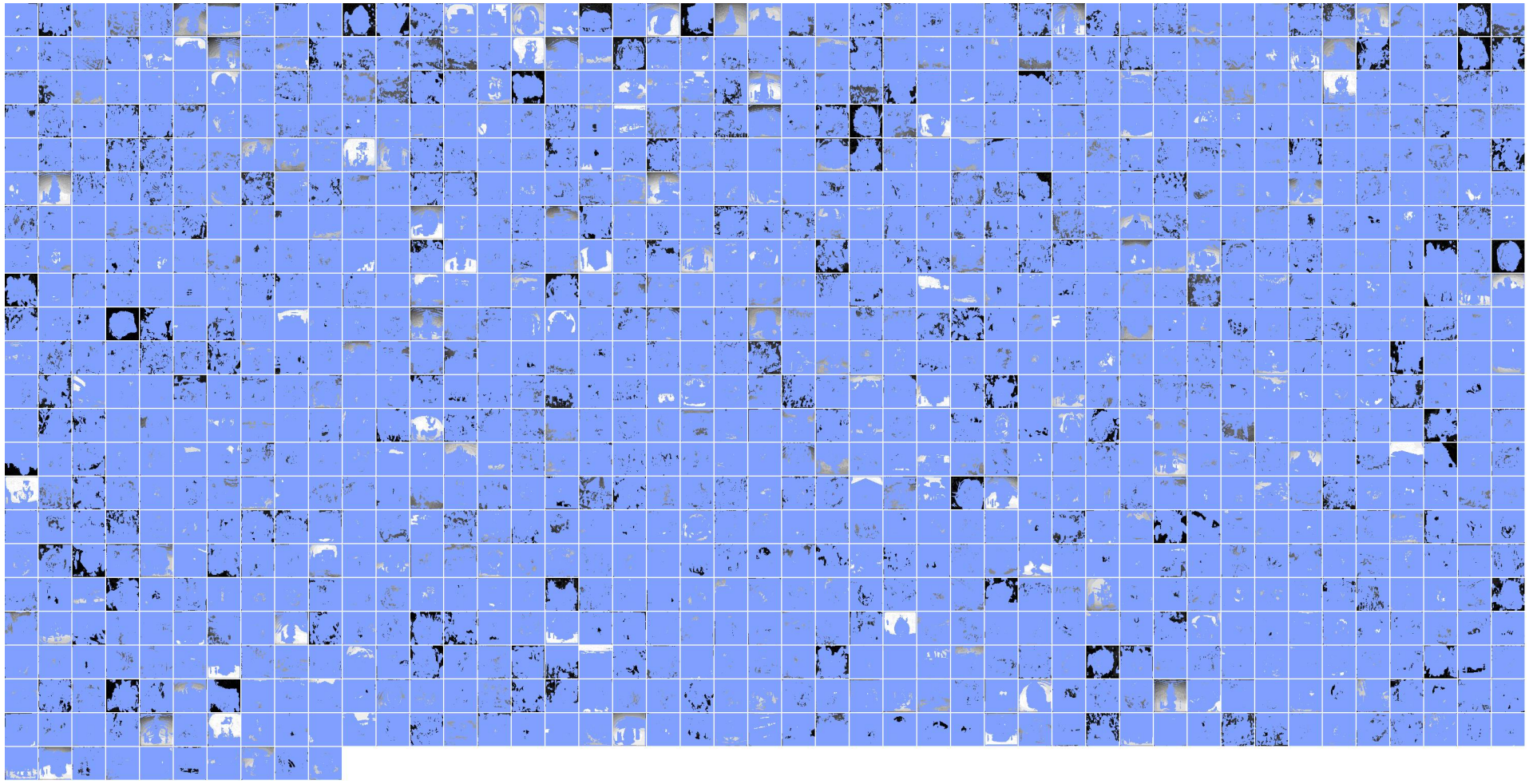
Nouvelles expériences

- **Nouvelles contraintes (choix des pixels)**
  - ***Min*** : les moins présents dans les motifs
  - ***Connexe*** : touchant les précédents
  - ***Ligne*** : formant des lignes
- **Paramètres**
  - *Taille image* : 64 x 64 x 16 (niveaux de gris)
  - *Taux de couverture* : 1, 2, 5, 10 %

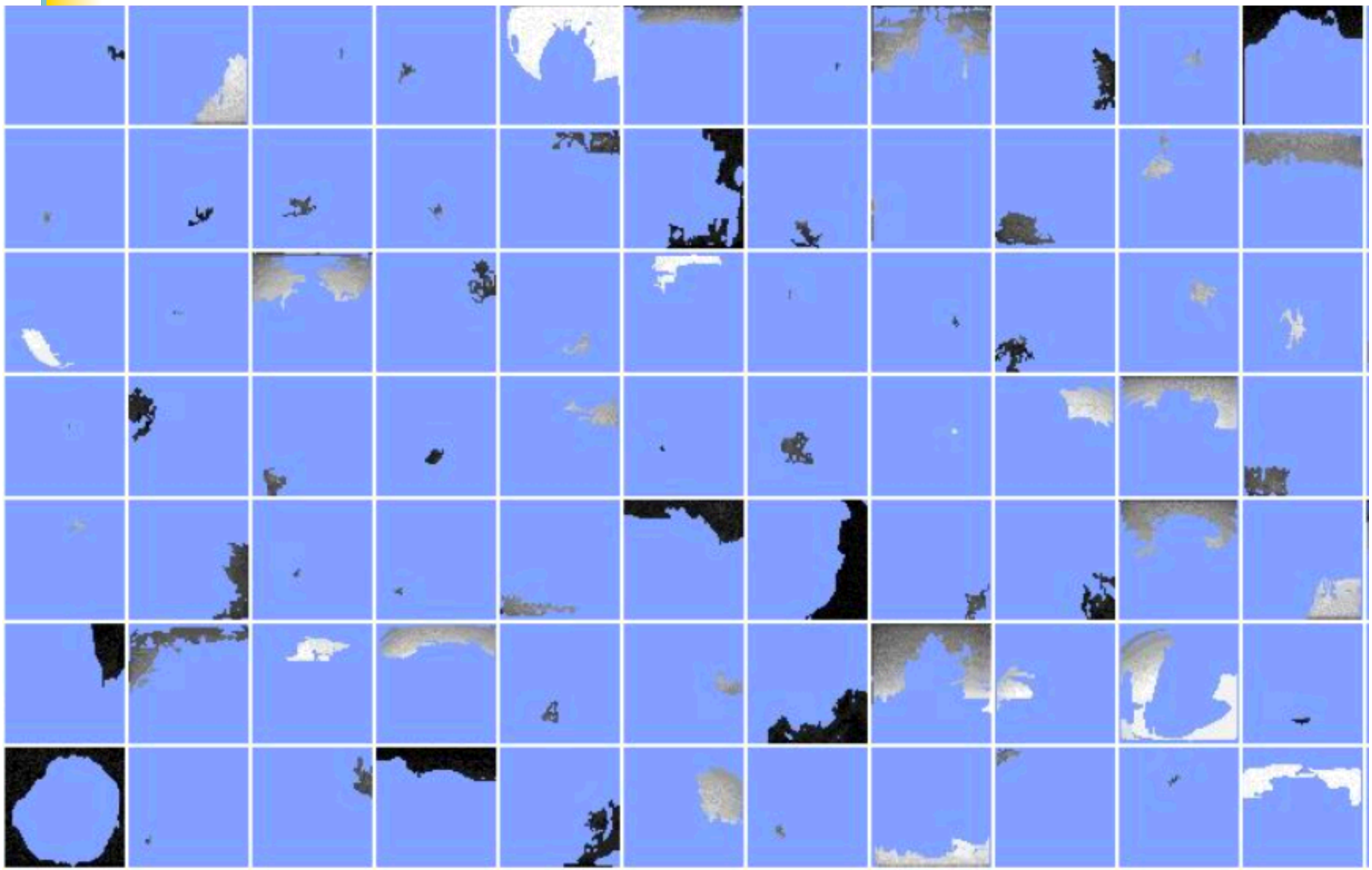
FIS : min_1%



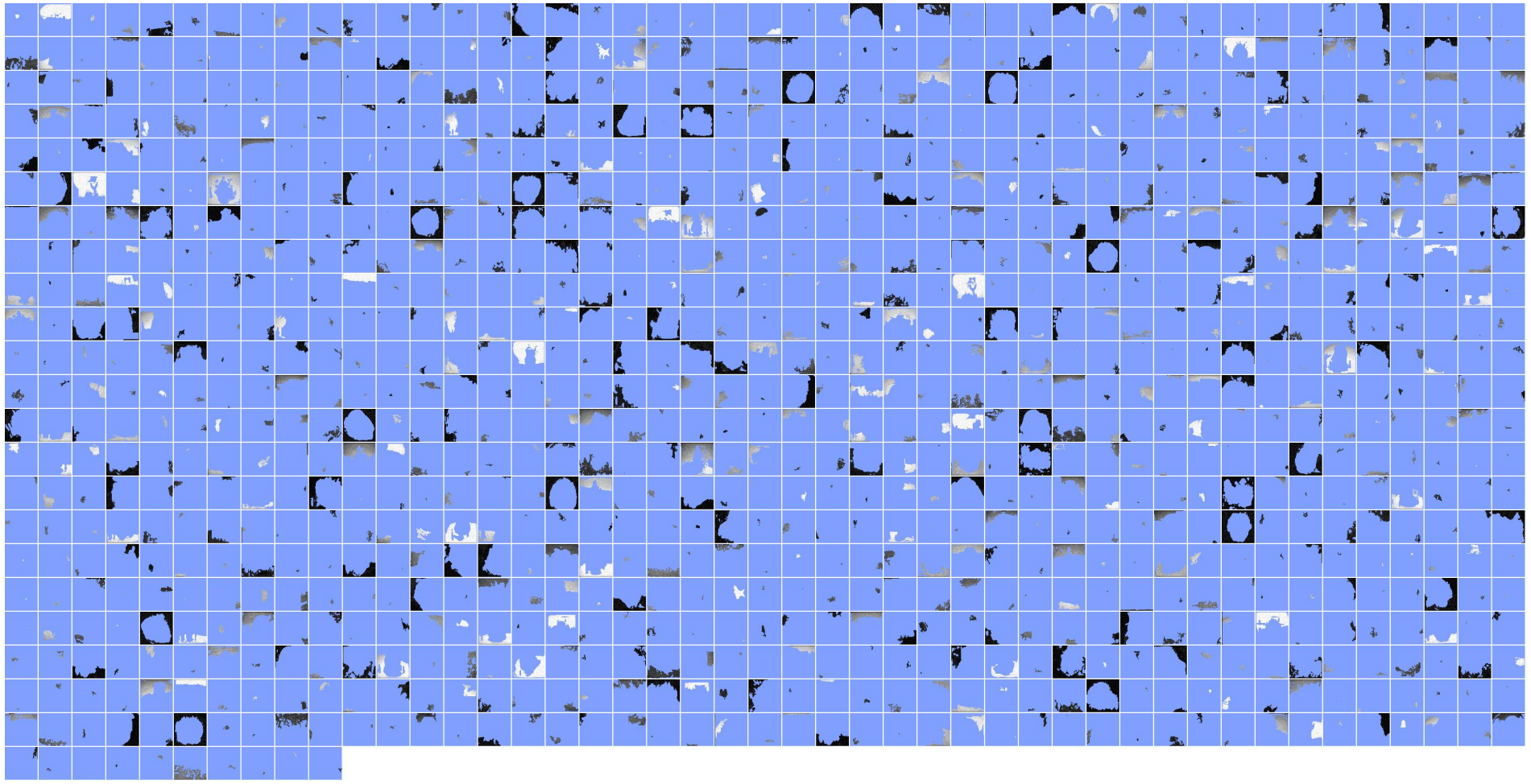
FIS : min_1%



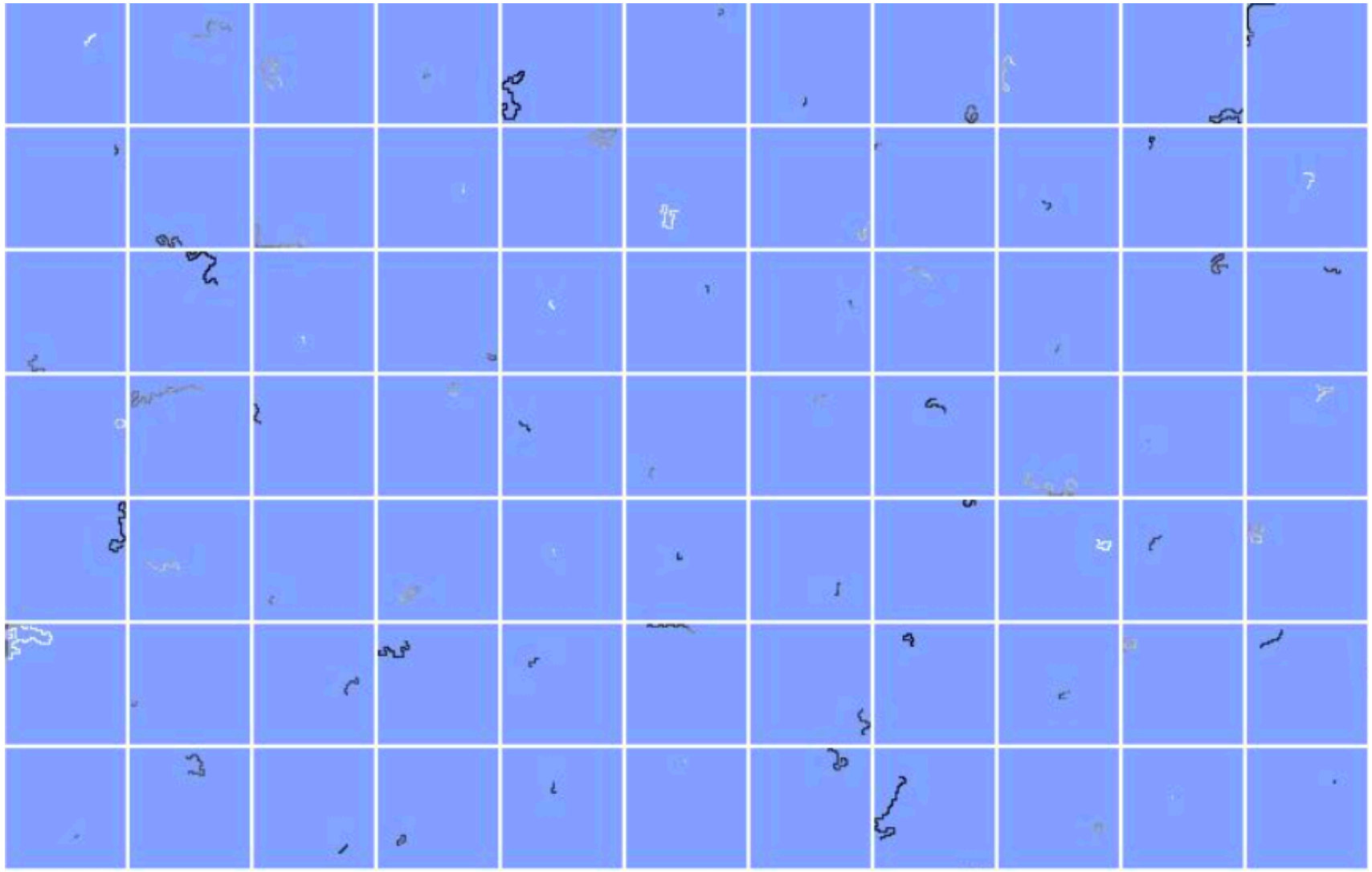
FIS : connexe_1%



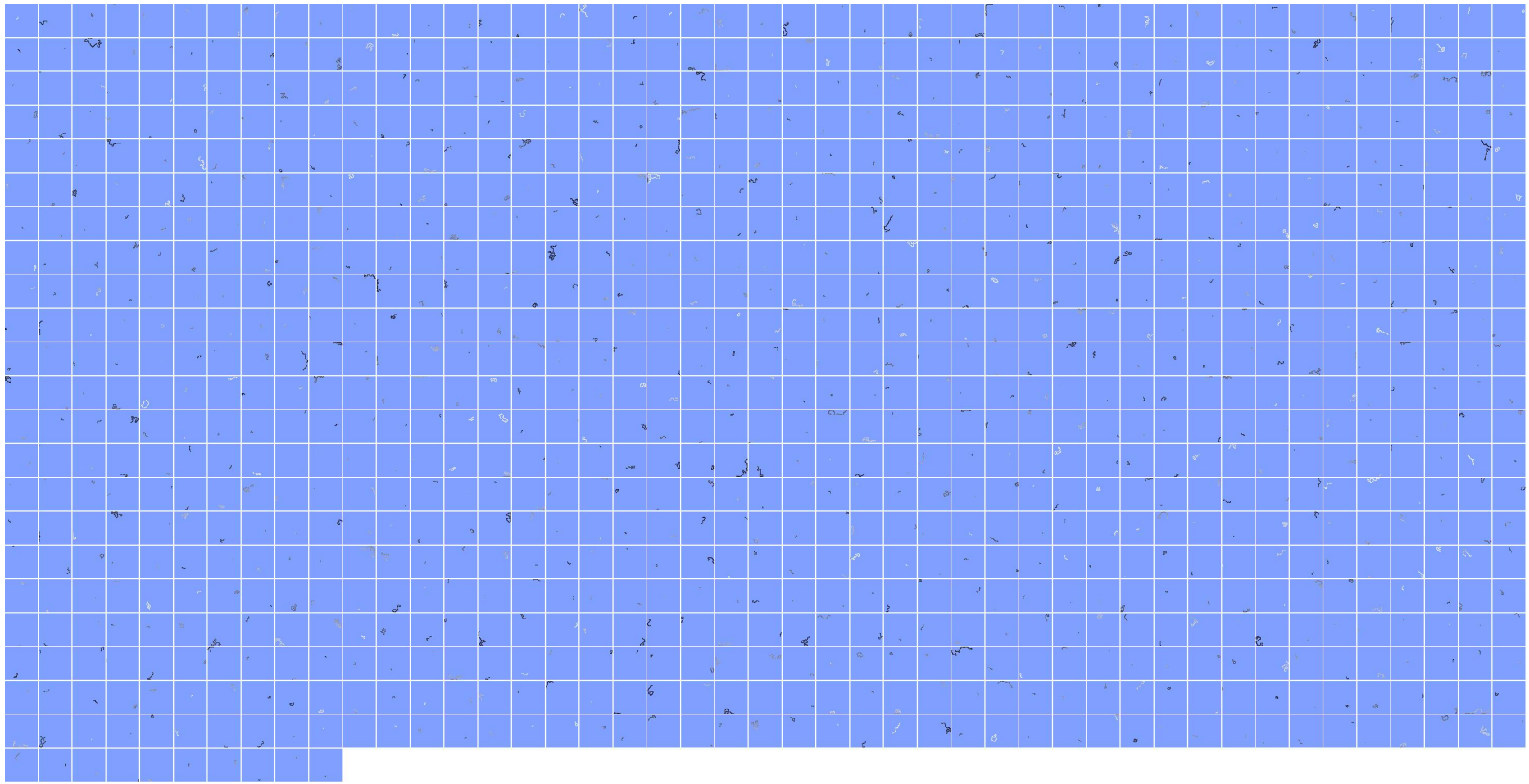
FIS : connexe_1%

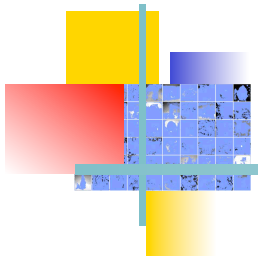


FIS : ligne_1%



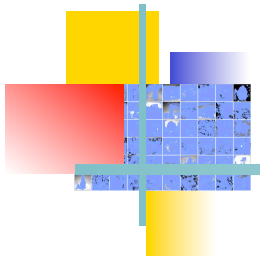
FIS : ligne_1%





Analyse

- **Difficilement interprétables !!**
- **Pas de contours**, même quand contraintes dans ce sens
- Malheureusement **pas de comparaison possible avec ACI** puisque ACI non praticable



La classification : le protocole

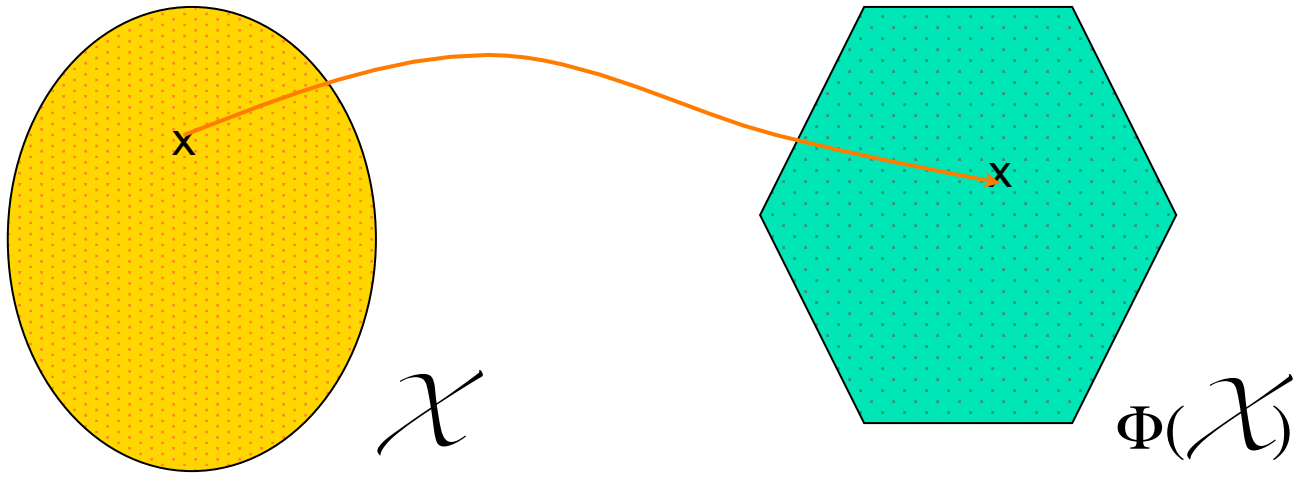
- **Apprentissage d'une base de 1000 motifs sur 540 images**
- **Les paramètres :**
  - Taille image (32 x 32, 64 x 64 ou 128 x 128)
  - Niveaux de gris (16, 32 ou 64)
  - Taux de couverture (1%, 2%, 5% ou 10%)
- **Test sur les 540 images restantes** (répété 10 fois)

  - Note : Tous les résultats sont disponibles sur :

    http://www.eleves.iie.cnam.fr/jouteau

La classification : la méthode

- **Chaque exemple (dans $\mathcal{X}$) est décrit par ses motifs (dans $\Phi(\mathcal{X})$)**
- **Un nouvel exemple est classé par une méthode de plus proches voisins (dans l'espace de redescription $\Phi(\mathcal{X})$)**
  - 1-ppv
  - ou *k*-ppv avec pondération en fonction de la distance



Performances (ε = 5%)

	Av	Pl	Ut	Mi	Ch	Po	Ve	Pa	Por	Fi	Vo	Fl
Avi	**67%**	2%	-	-	2%	2%	10%	10%	4%	2%	-	-
Pla	-	**21%**	-	2%	7%	19%	10%	12%	5%	-	19%	5%
Uta	17%	-	**33%**	-	7%	-	-	3%	10%	10%	13%	7%
Min	-	-	-	**100%**	-	-	-	-	-	-	-	-
Chi	26%	5%	7%	-	**14%**	9%	12%	9%	5%	2%	12%	-
Poi	5%	13%	3%	8%	-	**13%**	18%	21%	-	3%	10%	8%
Ver	2%	2%	-	-	10%	7%	**43%**	-	21%	5%	7%	2%
Pap	6%	6%	-	-	2%	14%	14%	**35%**	6%	-	12%	4%
Por	2%	2%	-	-	-	2%	-	12%	**70%**	10%	-	2%
Fig	-	-	-	-	-	-	6%	-	24%	**70%**	-	-
Voi	21%	6%	-	-	4%	4%	8%	4%	4%	29%	**19%**	-
Fle	2%	9%	-	-	-	9%	21%	14%	-	-	16%	**28%**

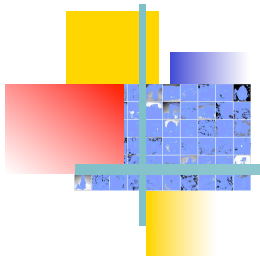
Avec un réseau de neurones RBF

	Av	Pl	Ut	Mi	Ch	Po	Ve	Pa	Por	Fi	Vo	Fl	Rj
Avi	**50.7**	-	3.3	-	-	-	-	-	-	-	1.3	-	44.7
Pla	-	-	-	6.3	0.8	-	-	-	-	-	-	1.6	91.3
Uta	1.1	-	**23.3**	-	3.3	1.1	-	-	2.2	-	-	-	68.9
Min	0.8	-	0.8	**28.8**	-	2.3	-	0.8	-	-	-	3.0	63.6
Chi	-	-	4.0	0.8	**11.1**	2.4	-	0.8	-	-	3.2	0.8	77
Poi	-	-	3.7	2.2	-	**0.7**	-	-	-	-	0.7	0.7	91.9
Ver	-	-	2.9	-	0.7	-	**9.4**	-	20.3	15.9	-	-	50.7
Pap	-	-	-	7.3	-	1.3	-	**13.3**	1.3	-	-	-	76.7
Por	2.0	-	0.7	-	-	-	0.7	-	**45.3**	4.7	-	-	46.7
Fig	-	-	-	-	-	-	18.7	-	6.7	**42.0**	0.7	-	32
Voi	4.1	-	-	-	2.0	-	-	-	-	2.0	**34.0**	-	57.8
Fle	-	-	-	2.1	-	0.7	-	1.4	-	-	0.7	**24.8**	70.2

Comparaison

	Av	Pl	Ut	Mi	Ch	Po	Ve	Pa	Por	Fi	Vo	Fl
Avi	**67%**	2%	-	-	2%	2%	10%	10%	4%	2%	-	-
Pla	-	**21%**	-	2%	7%	19%	10%	12%	5%	-	19%	5%
Uta	17%	-	**33%**	-	7%	-	-	3%	10%	10%	13%	7%
Min	-	-	-	**100%**	-	-	-	-	-	-	-	-
Chi	26%	5%	7%	-	**14%**	9%	12%	9%	5%	2%	12%	-
Poi	5%	13%	3%	8%	-	**13%**	18%	21%	-	3%	10%	8%
Ver	2%	2%	-	-	10%	7%	**43%**	-	21%	5%	7%	2%
Pap	6%	6%	-	-	2%	14%	14%	**35%**	6%	-	12%	4%
Por	2%	2%	-	-	-	2%	-	12%	**70%**	10%	-	2%
Fig	-	-	-	-	-	-	6%	-	24%	**70%**	-	-
Voi	21%	6%	-	-	4%	4%	8%	4%	4%	29%	**19%**	-
Fle	2%	9%	-	-	-	9%	21%	14%	-	-	16%	**28%**

	Av	Pl	Ut	Mi	Ch	Po	Ve	Pa	Por	Fi	Vo	Fl	Rj
Avi	**50.7**	-	3.3	-	-	-	-	-	-	-	1.3	-	44.7
Pla	-	-	-	6.3	0.8	-	-	-	-	-	-	1.6	91.3
Uta	1.1	-	**23.3**	-	3.3	1.1	-	-	2.2	-	-	-	68.9
Min	0.8	-	0.8	**28.8**	-	2.3	-	0.8	-	-	-	3.0	63.6
Chi	-	-	4.0	0.8	**11.1**	2.4	-	0.8	-	-	3.2	0.8	77
Poi	-	-	3.7	2.2	-	**0.7**	-	-	-	-	0.7	0.7	91.9
Ver	-	-	2.9	-	0.7	-	**9.4**	-	20.3	15.9	-	-	50.7
Pap	-	-	-	7.3	-	1.3	-	**13.3**	1.3	-	-	-	76.7
Por	2.0	-	0.7	-	-	-	0.7	-	**45.3**	4.7	-	-	46.7
Fig	-	-	-	-	-	-	18.7	-	6.7	**42.0**	0.7	-	32
Voi	4.1	-	-	-	2.0	-	-	-	-	2.0	**34.0**	-	57.8
Fle	-	-	-	2.1	-	0.7	-	1.4	-	-	0.7	**24.8**	70.2

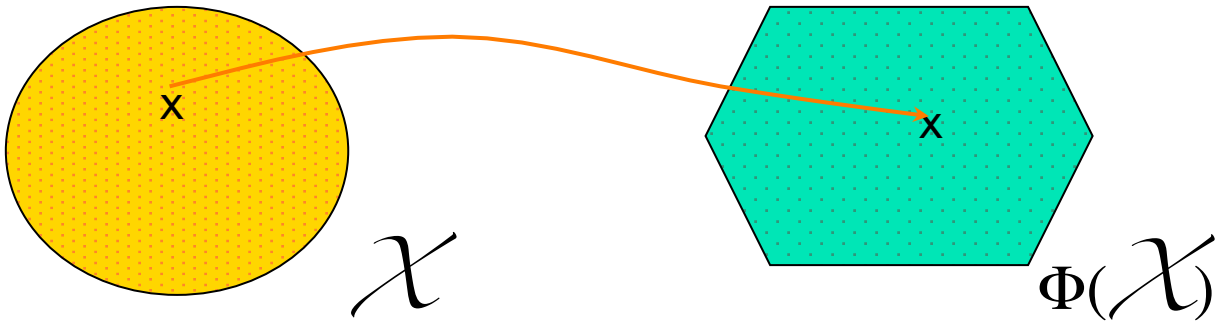


Performances en classification

- Résultats
  - Meilleurs résultats pour ε = 2 ou 5 %
  - Assez comparable : *min, connexe, ligne*
  - Bien meilleurs que méthode RN
- Peut mieux faire …
  - Avec un appariement plus souple

Analyse

- **Pourquoi ça marche (si bien) ?**
  - Recodage non supervisé !!
  - Puis une méthode de plus proche(s) voisin(s)



Quelles sont les propriétés de ce recodage ?

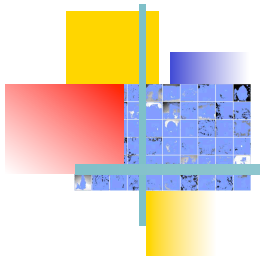
Approches classiques

… et moins classiques

	Réduction	Orthogonalité	Indép. des données	Approximation
Analyse fonctionnelle	+/-	✔	✔	✔
PCA	✔	✔		✔
Apprent. artificiel	✔			✔
ICA				✔

Le codage par motifs fréquents

- **Ne permet pas la reconstruction des entrées**
- **Les motifs sont orthogonaux : mais par rapport aux exemples d'apprentissage !!**
- **Espace** C_{100}^{10}
- **Tous les points d'apprentissage sont orthogonaux dans cet espace**



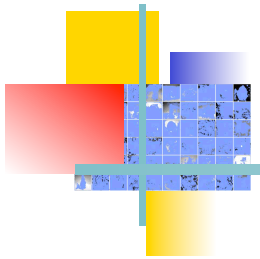
Analyse (suite)

- ***Pourquoi aurait-on de bonnes garanties en généralisation ?***

➥ **Capacité limitée ?** (dimension de Vapnik-Chervonenkis)

- **Conjecture initiale**
  - Peu de cellules de Voronoi
  - « Parois » définies par peu de paramètres (≈ 20)
  - donc d_{VC} très limitée
  - ➢ **… Mais faux**

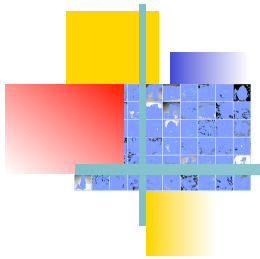
Autre analyse en cours



Conclusion

- **Analyse théorique en cours**
- **Expérimentations**
  - sur les puces ADN
  - sur la classification de textes de NewsGroups

➠ ***Peut-être un nouveau type de traitement du signal***



ICA : Analyse en Composantes. Indép.

(Introduite en 1984. Développée dans les 90s)

Hyp. de base : *les données résultent d'une combinaison linéaire de formes latentes*

- **Recherche de ces formes latentes**
  - Exemples
    - Images
    - Puces ADN

      ➔ ***Codage clairsemé des données***

- MAIS :
  - Les algorithmes existants **ne fonctionnent pas** en grande dimension
  - Hypothèse de linéarité

ε = 5% ; min *(voisinage discret)*

Matrice booléenne

	Avions	Plats	Utah	Minéraux	Chiens	Poissons	Verres	Papillons	Porcelaines	Figures	Voitures	Fleurs	Rejet
Avions	**31%**	10%	3%	2%	4%	4%	6%	5%	10%	-	22%	3%	-
Plats	-	**21%**	1%	7%	3%	23%	2%	18%	5%	-	12%	9%	-
Utah	12%	7%	**40%**	-	3%	17%	-	7%	-	-	13%	2%	-
Minéraux	-	2%	-	**68%**	-	16%	3%	8%	-	-	-	3%	-
Chiens	5%	11%	5%	2%	**26%**	14%	5%	15%	9%	-	8%	1%	-
Poissons	5%	13%	3%	3%	6%	**12%**	11%	16%	-	-	11%	20%	-
Verres	10%	-	-	-	2%	7%	**33%**	4%	17%	2%	21%	4%	-
Papillons	4%	10%	-	1%	5%	14%	9%	**28%**	3%	-	17%	9%	-
Porcelaines	2%	2%	-	-	5%	5%	4%	7%	**68%**	5%	2%	1%	-
Figures	4%	-	-	-	-	-	9%	-	23%	**64%**	-	-	-
Voitures	10%	8%	2%	-	8%	6%	5%	3%	9%	17%	**26%**	7%	-
Fleurs	-	6%	-	5%	3%	9%	9%	29%	-	-	18%	**20%**	-

ε = 5% ; min *(voisinage continu)*

Matrice continue

	Avions	Plats	Utah	Minéraux	Chiens	Poissons	Verres	Papillons	Porcelaines	Figures	Voitures	Fleurs	Rejet
Avions	**67%**	4%	6%	-	4%	-	8%	4%	-	-	6%	-	-
Plats	5%	**24%**	2%	2%	7%	21%	2%	10%	5%	-	14%	7%	-
Utah	17%	-	**50%**	-	-	3%	-	3%	-	3%	17%	7%	-
Minéraux	-	-	-	**100%**	-	-	-	-	-	-	-	-	-
Chiens	14%	7%	9%	-	**23%**	5%	12%	14%	2%	-	12%	2%	-
Poissons	3%	18%	-	13%	10%	**23%**	8%	10%	-	-	10%	5%	-
Verres	7%	5%	-	5%	2%	7%	**33%**	-	26%	7%	7%	-	-
Papillons	8%	4%	-	-	6%	2%	14%	**53%**	6%	-	4%	2%	-
Porcelaines	-	-	-	4%	-	-	-	14%	**72%**	8%	-	2%	-
Figures	-	-	-	-	-	-	2%	-	22%	**76%**	-	-	-
Voitures	31%	4%	2%	-	15%	-	2%	17%	2%	17%	**10%**	-	-
Fleurs	5%	2%	-	-	5%	2%	23%	33%	-	-	12%	**19%**	-