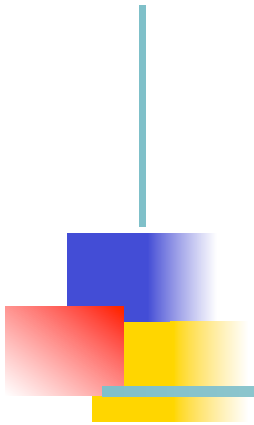


FISICA

Frequent Item Set and Independent Component Analysis

A. Cornuéjols, J. Mary & M. Sebag (LRI)

S. Jouteau, Ph. Tarroux & J-S. Liénard (LIMSI)

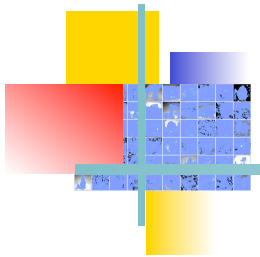
CNRS - Université de Paris-Sud, Orsay

High dimensionality data

- **Very high number of descriptors**

  (NB : One of the 10 problems for the Society, 2000))

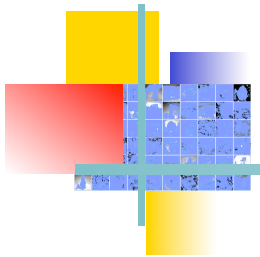
- **Examples :**
  - Microarray
    - E.g. 6400 genes,
    - ➥ tumor or sound
  - Images
    - E.g. 256*256*(256 gray levels)
    - ➥ Identification of patterns in the image



High dimensionality data

- Most data points are concentrated on a small non-linear subspace of the input space

- There must exist an efficient coding scheme
  - That is able to describe the data
  - And is economical or sparse



Sparse coding

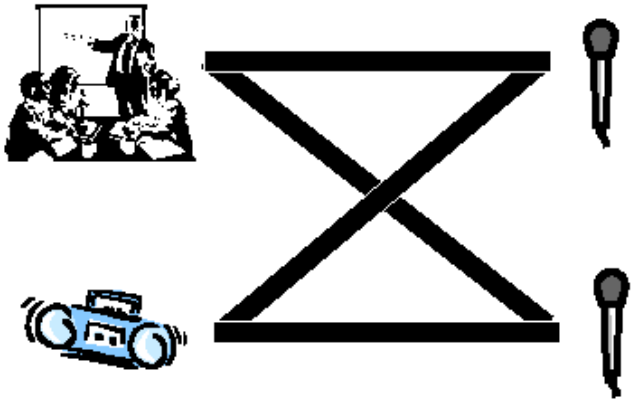
- Hypothesis :

  Each pattern is represented by a small subset of basis components in a large dictionary

Independent Component Analysis

(Introduced in 1984. Developed in the 90s)

Hypothesis: *Data can be represented by a linear combination of basis components*



- **Limits :**
  - Linearity
  - Not feasible for high number of dimensions

Exemple: ICA for scene analysis

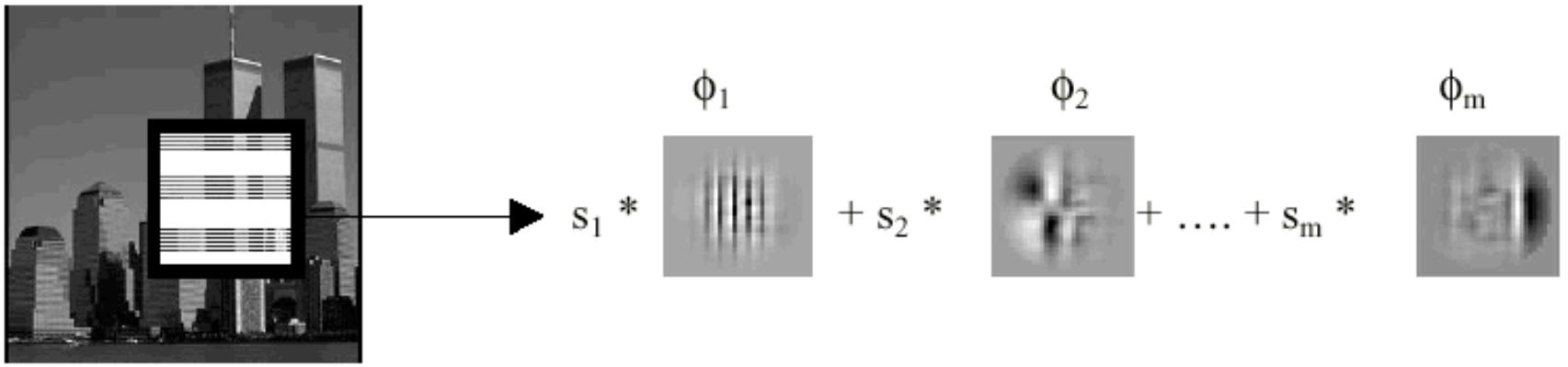


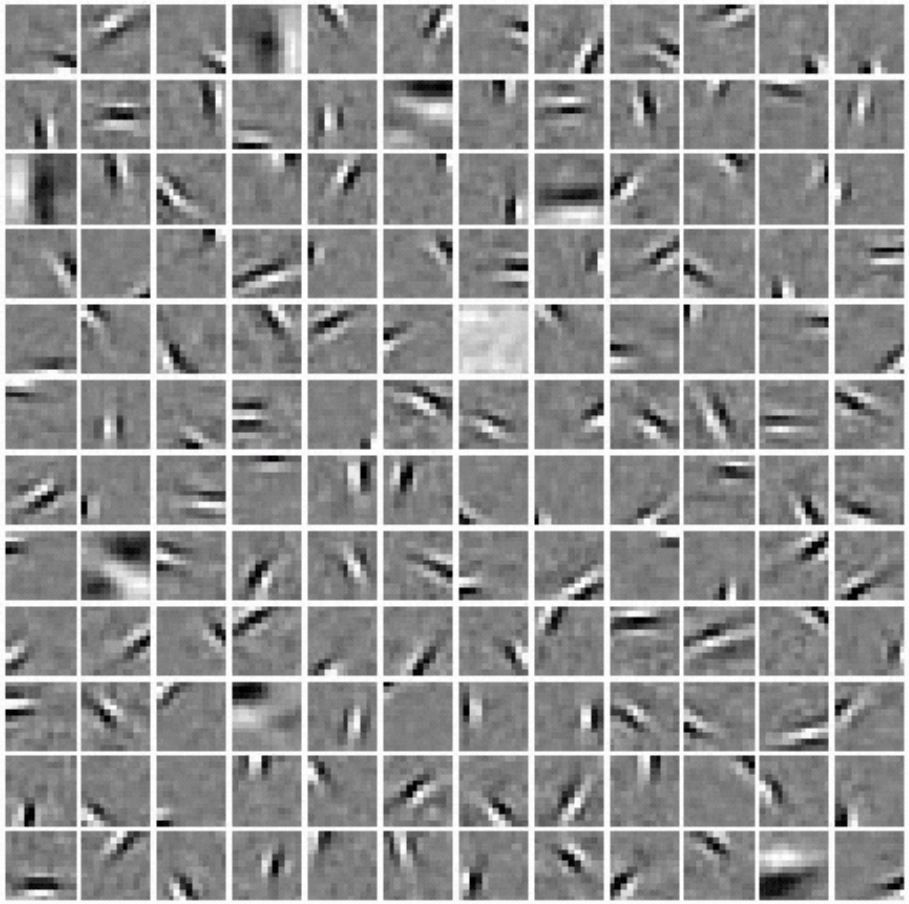
Figure 1 : Illustration de la décomposition d'une imagette dans la base Φ.

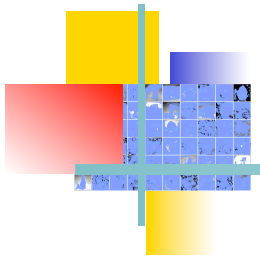
- **Scenes are decomposed in image windows …**
- **… coded by linear superposition of latent forms**

12 x 12 latent forms

[Olshausen & Field, 1996, …]

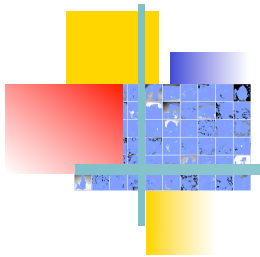
- **Most forms are**
  - Localised
  - Oriented
  - With various frequencies





FISICA

- ***Direct search of a sparse code ?***
  - Using Frequent Item sets
  - On whole patterns (e.g. whole images)

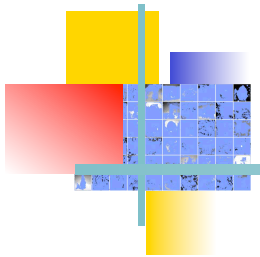


Frequent Item Sets

- **Problem**
  - Given a data base of tuples
  - Identify subsets of items that are fequently found together (Frequent ItemSets)

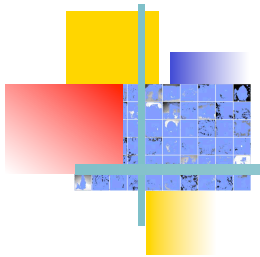
- **In general :**
  - A lot of Frequent Item Sets
  - But very few that are occur together

➥ **Sparse coding**



Constraints on the FIS

- **Representativity**
  - Each pattern corresponds to a **minimal** number of FIS
- **Sparsity**
  - Each pattern corresponds to a **limited** number of FIS
- **Orthogonality**
  - Each pair of FIS covers **few patterns in common**
- \+ **Semantical constraints**
  - E.g. : ***compact*** FIS (e.g. areas in the images)
  - E.g. : ***connected*** FIS (e.g. contours)
  - …



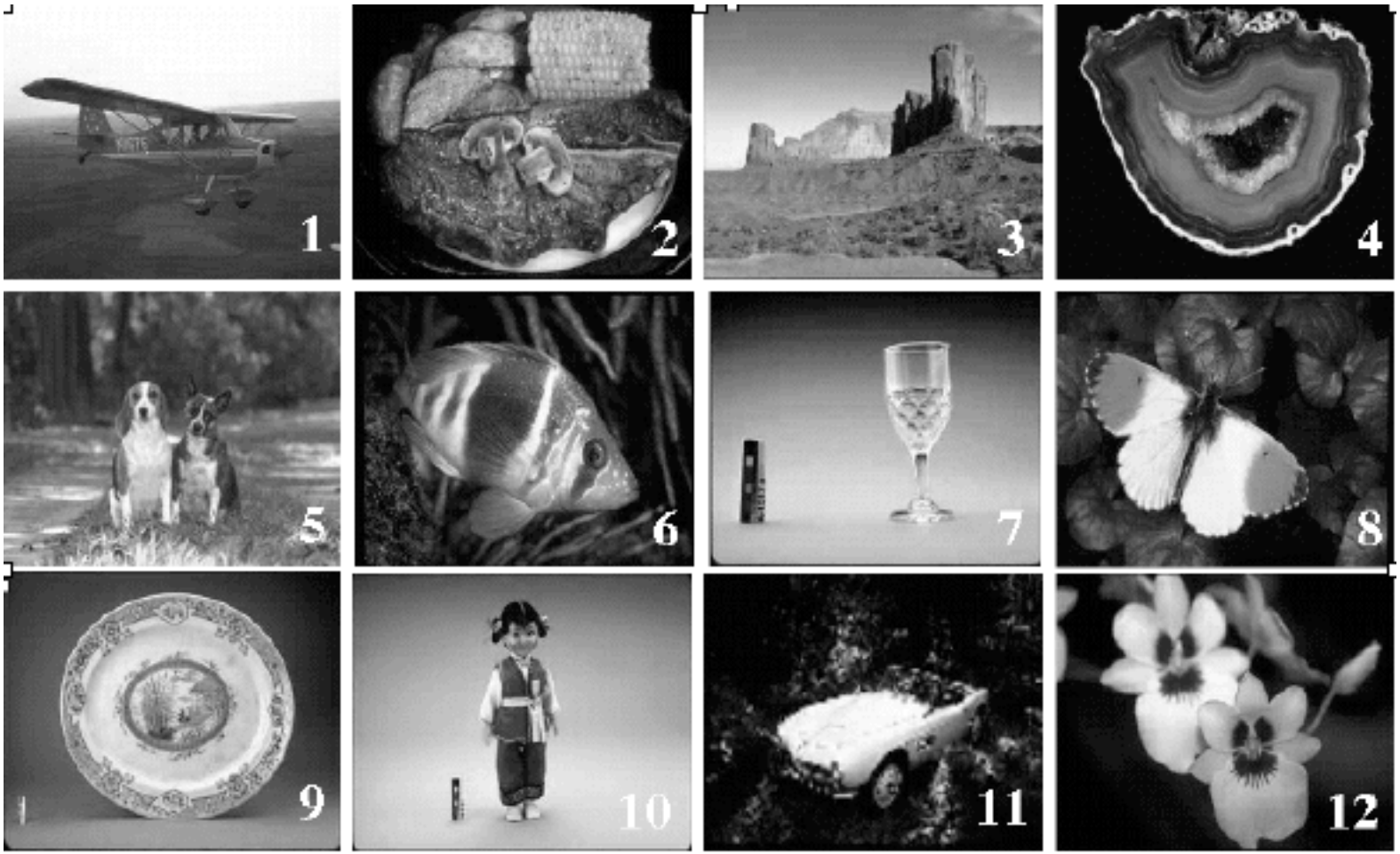
Experiments

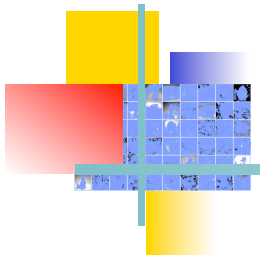
- The COREL image base
- **12 different classes**
- **1080 images** (90 images / class)
- 128 x 128 = 16384 with 128 grey levels

or : 64 x 64 = 4096 with 32 or 16 grey levels

540 images are used to identify **1000 FIS**

Image base





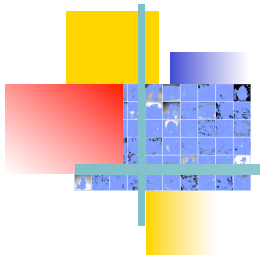
Finding

- Direct application of APRIORI is impossible
- There are too many FIS

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

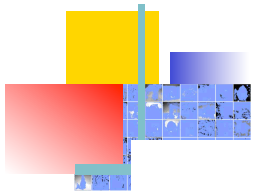
For 32 x 32 images with 64 grey levels

➥ *Adapt the algorithm with a **stochastic search strategy***



Experiments

- **Semantic constraints (choice of pixels to accrete)**
  - ***Min*** : the less present in the existing FIS
  - ***Connexe*** : Connected to the pixels of the calculated FIS
  - ***Curve*** : forming curves
- **Parameters**
  - *Image* size : 64 x 64 x 16 (grey levels)
  - *Coverage rate* : 1, 2, 5, 10 %



Sparse coding

Nb of FIS / images

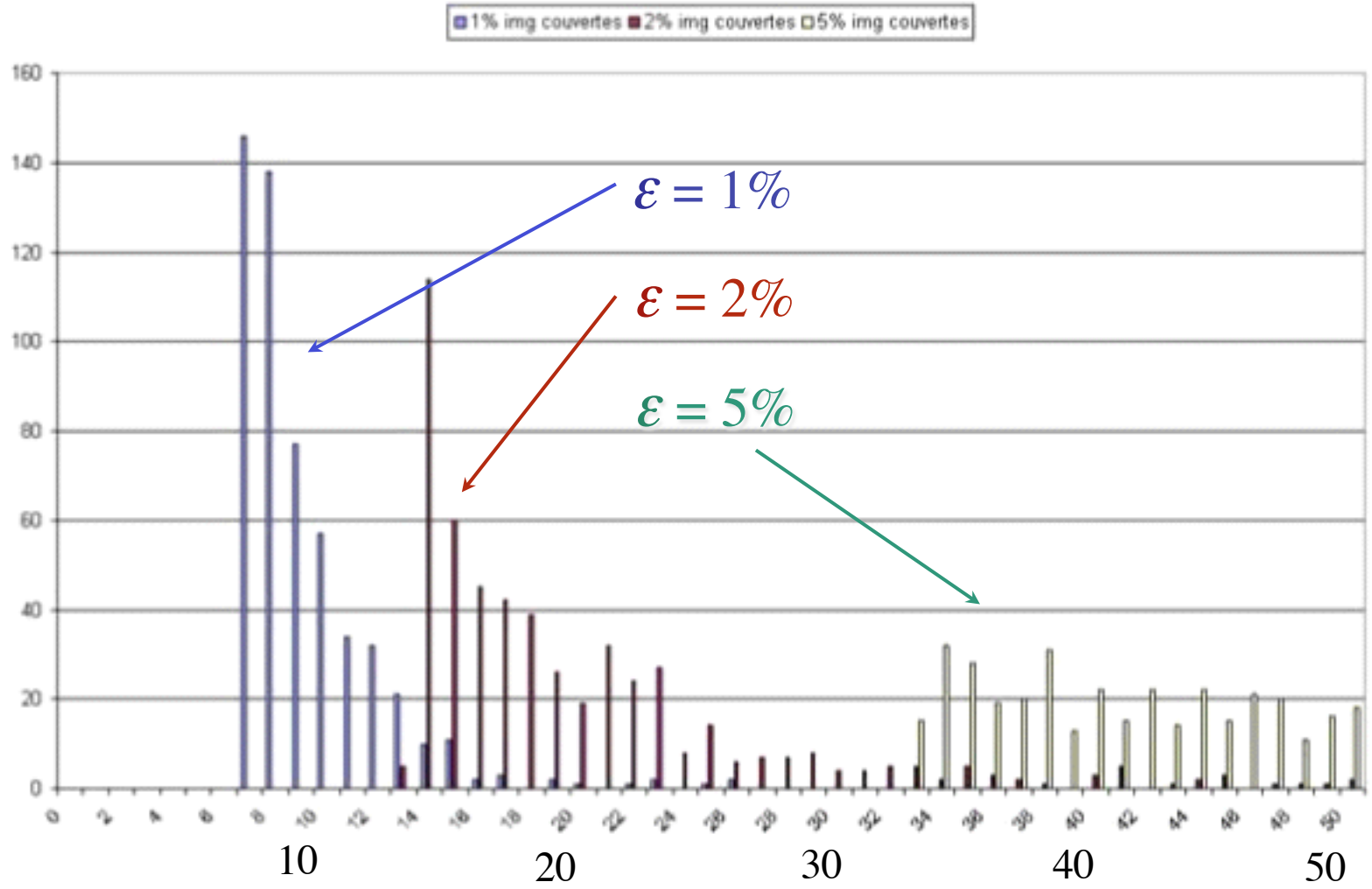


FIG. 6.22 – Histogramme représentant le nombre d'images (en ordonnée) activant N FIS (en abscisse).Ces fonctions ont été calculées à partir d'images de taille 64x64 en 16 niveaux de gris. La troisième méthode de choix des pixels (pixels se touchant) est utilisée. Un bon histogramme a le moins de valeurs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

Sparse coding — Nb of FIS / images

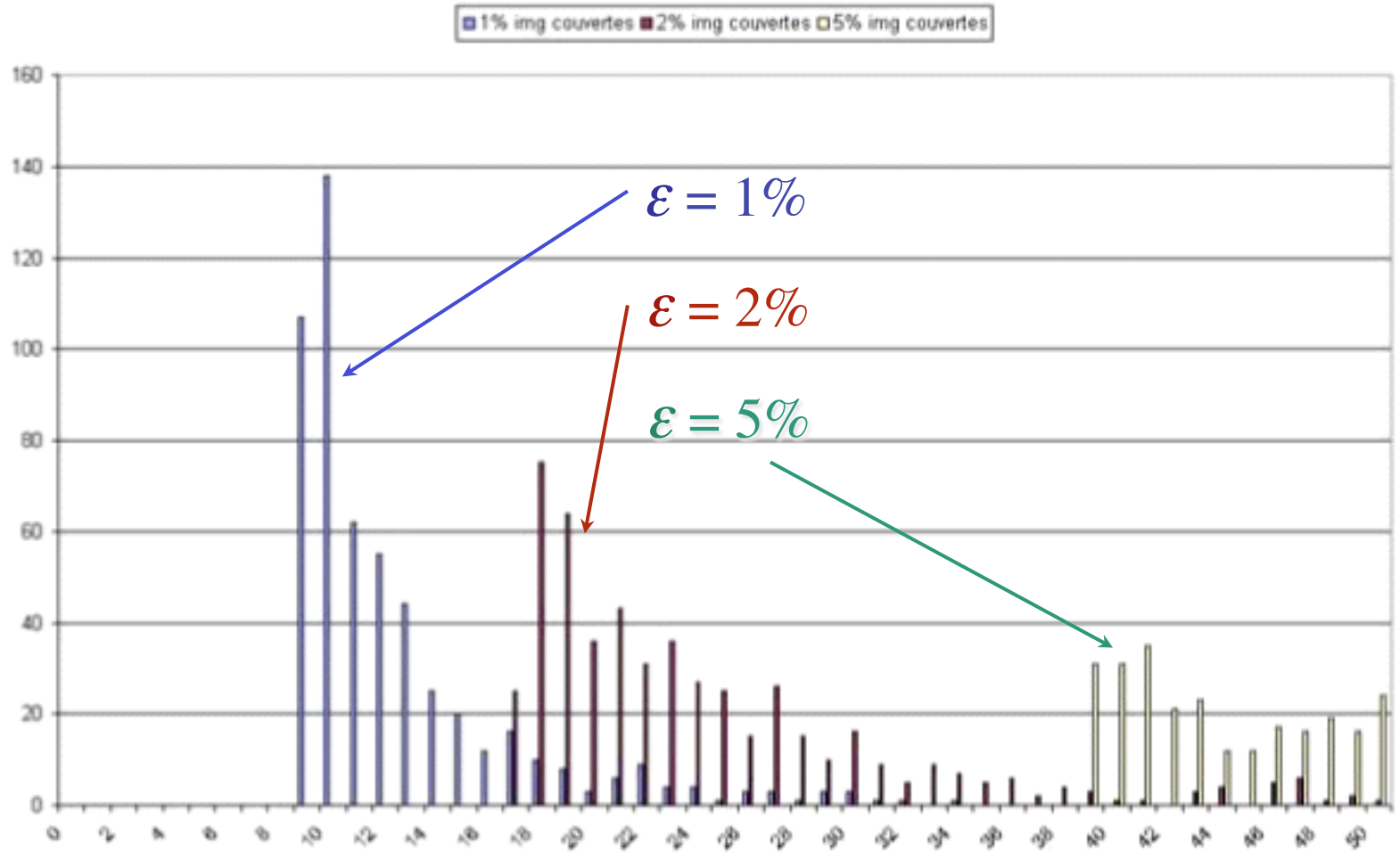


FIG. 6.23 – Histogramme représentant le nombre d'images (en ordonnée) activant N FIS (en abscisse).Ces fonctions ont été calculées à partir d'images de taille 64x64 en 16 niveaux de gris. La quatrième méthode de choix des pixels (pixels formant une ligne) est utilisée. Un bon histogramme a le moins de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

Orthogonality

Nb images / pair of FIS

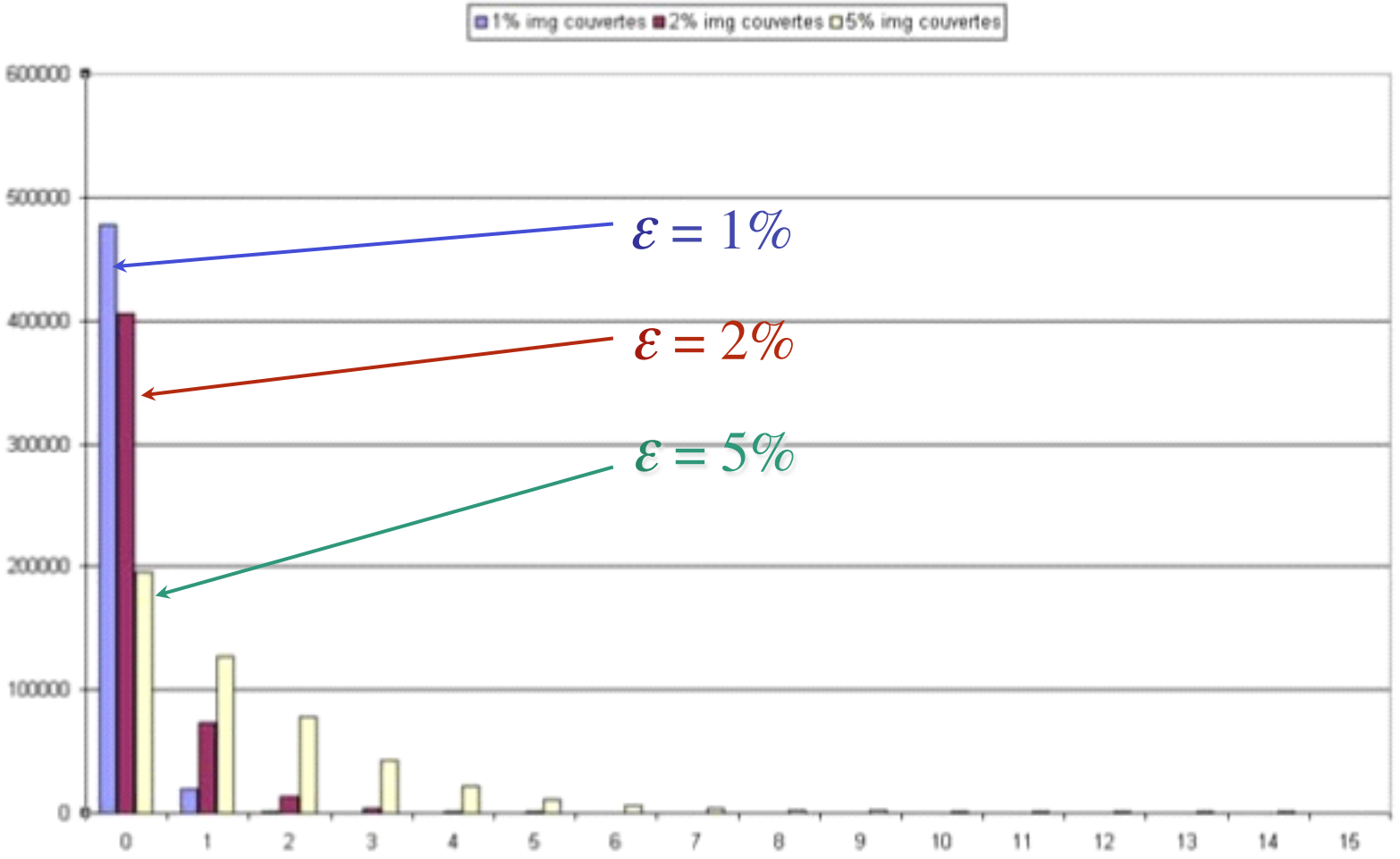
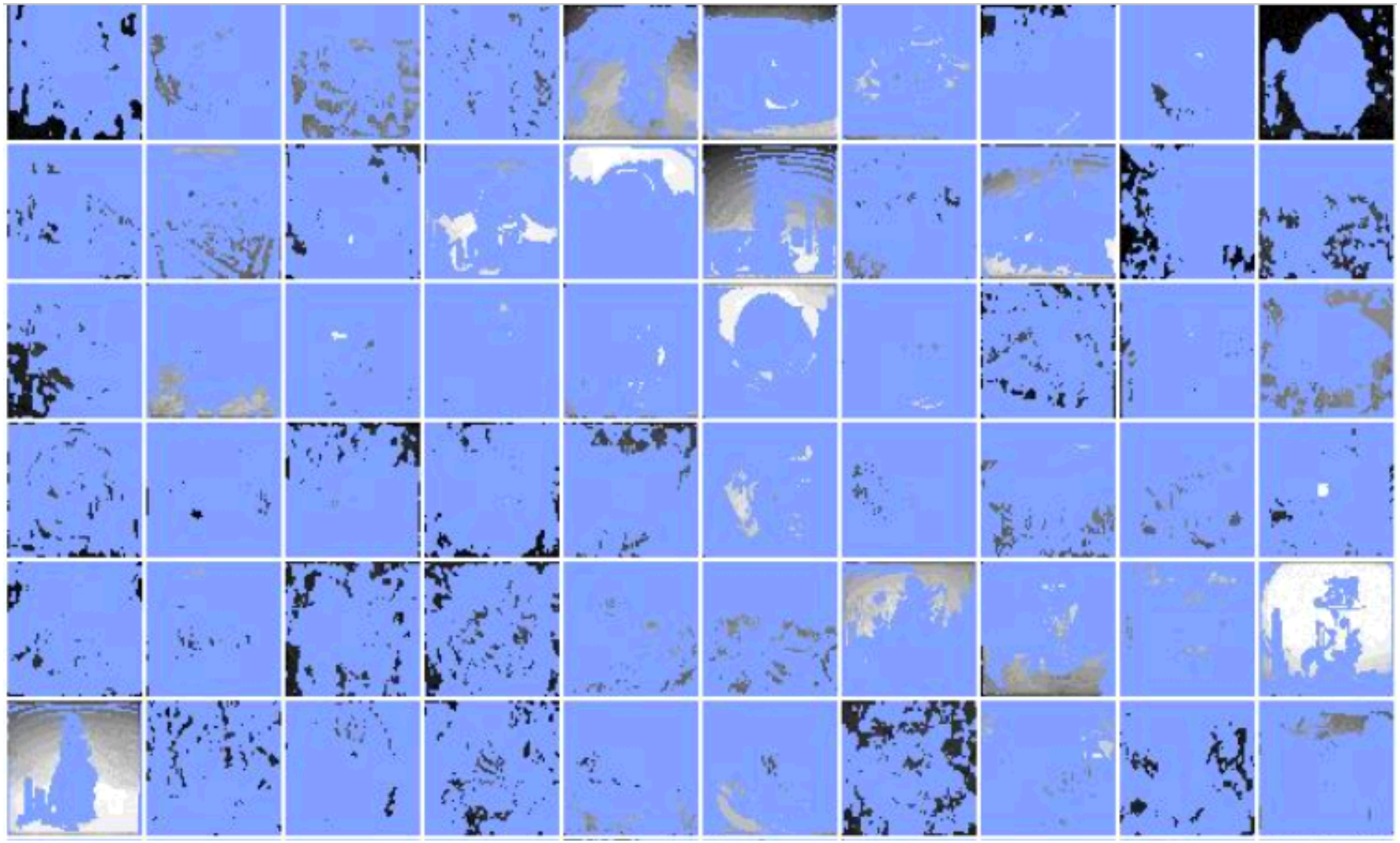
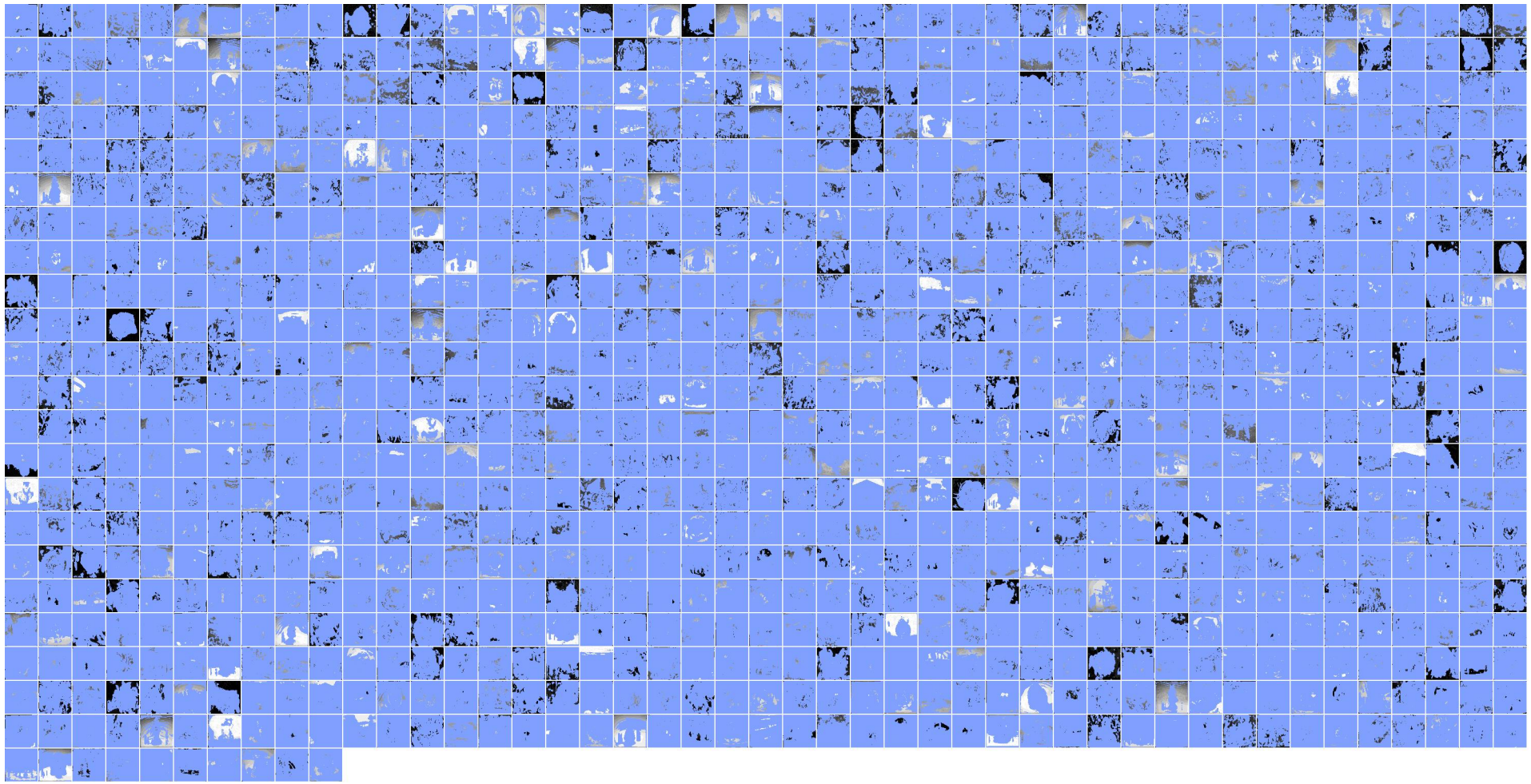


FIG. 6.18 – Histogramme représentant le nombre de couples de fonctions (en ordonnée) ayant N images en commun (en ascisse). Ces fonctions ont été calculées à partir d'images de taille 64x64 en 16 niveaux de gris. La troisième méthode de choix des pixels (pixels se touchant) est utilisée. Un bon histogramme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

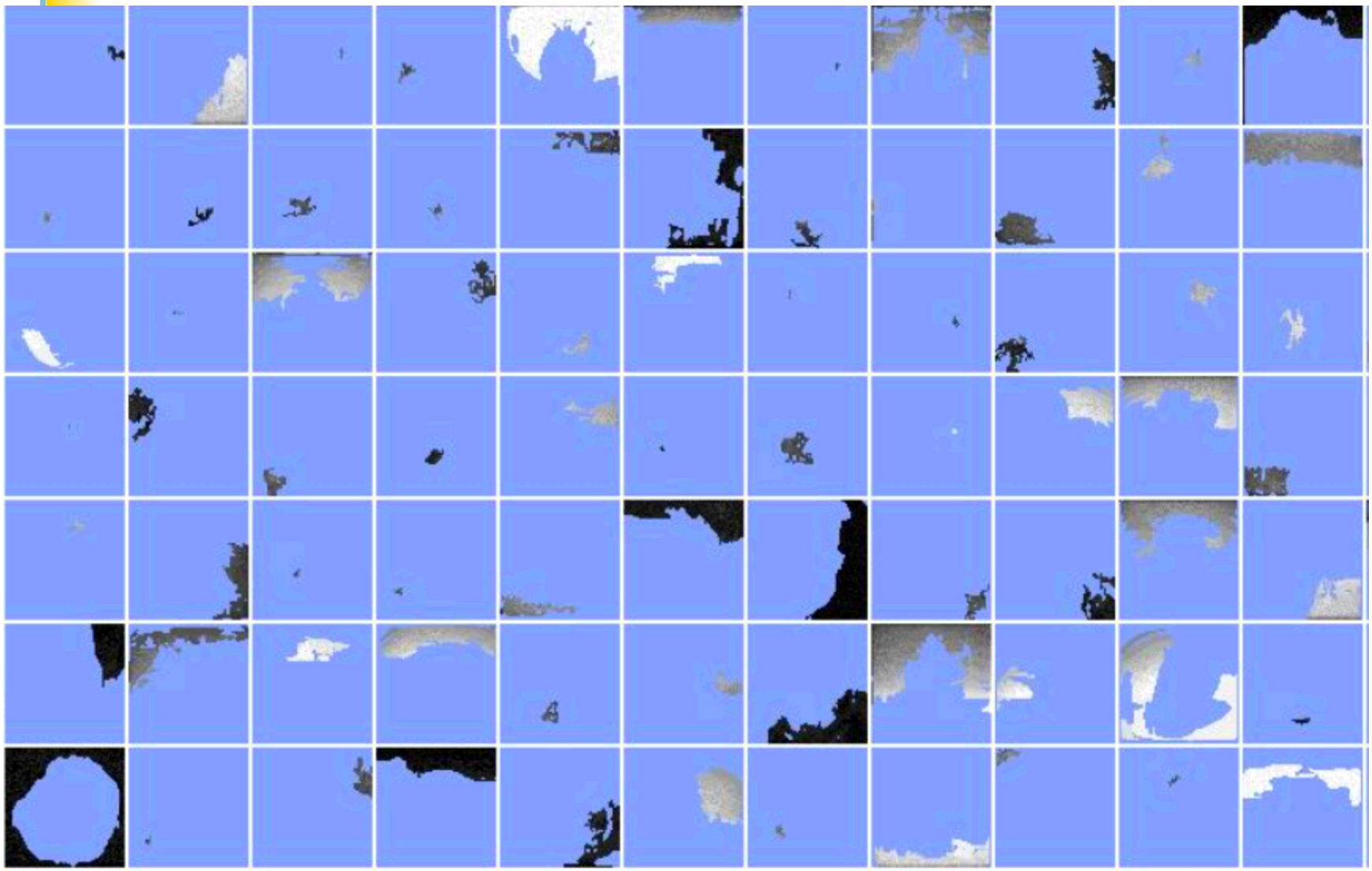
FIS : min_1%



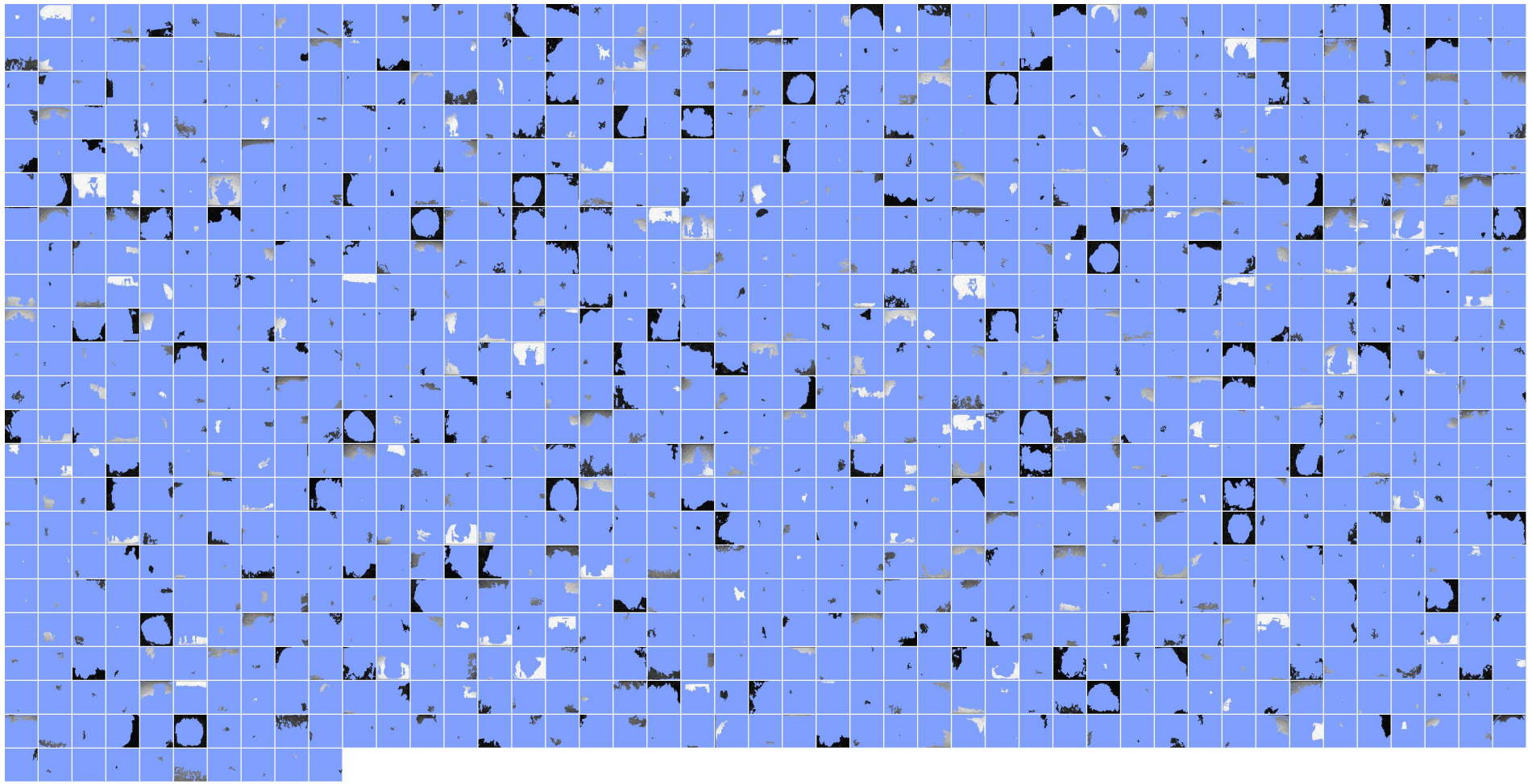
FIS : min_1%



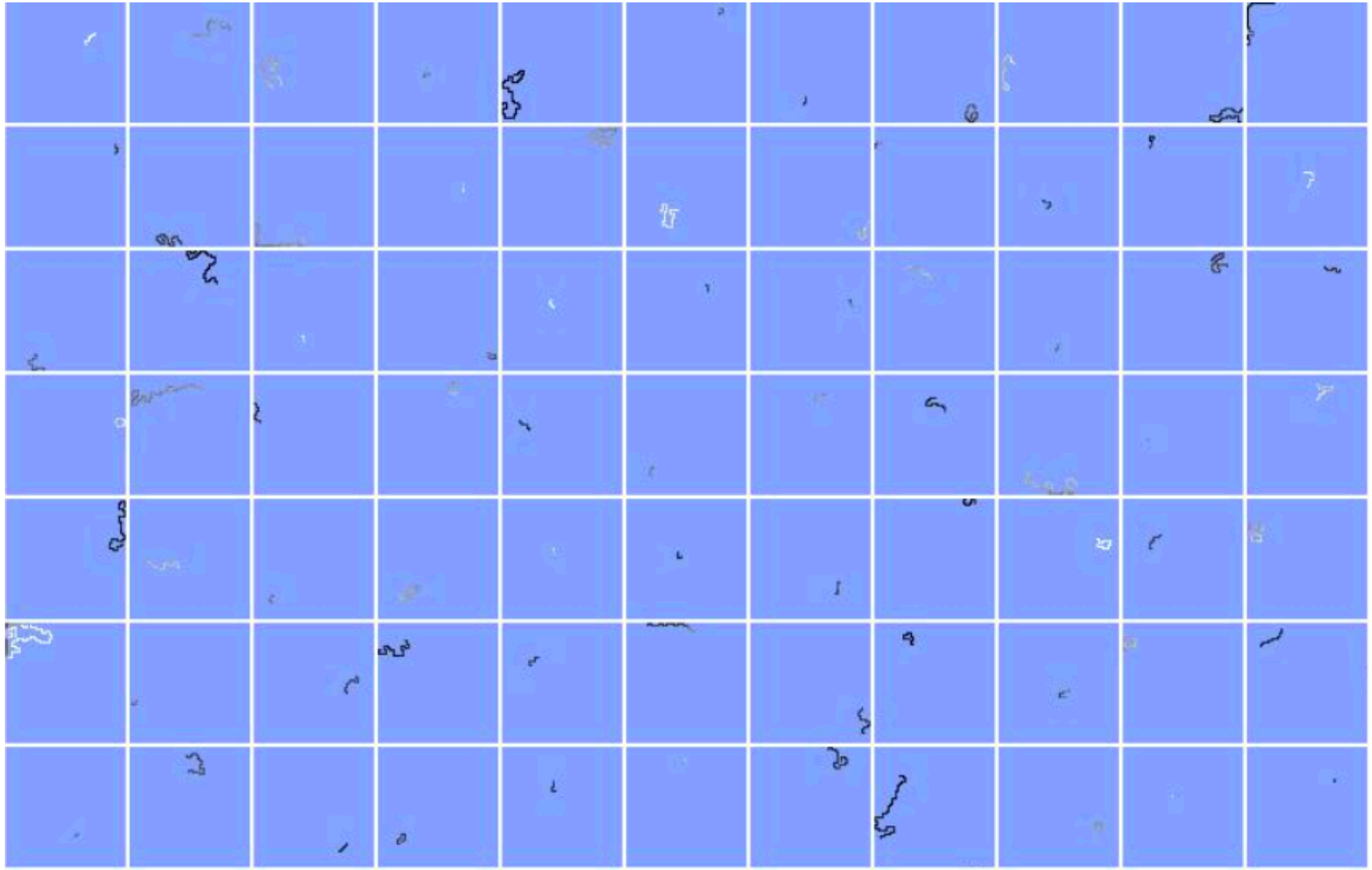
FIS : connected_1%



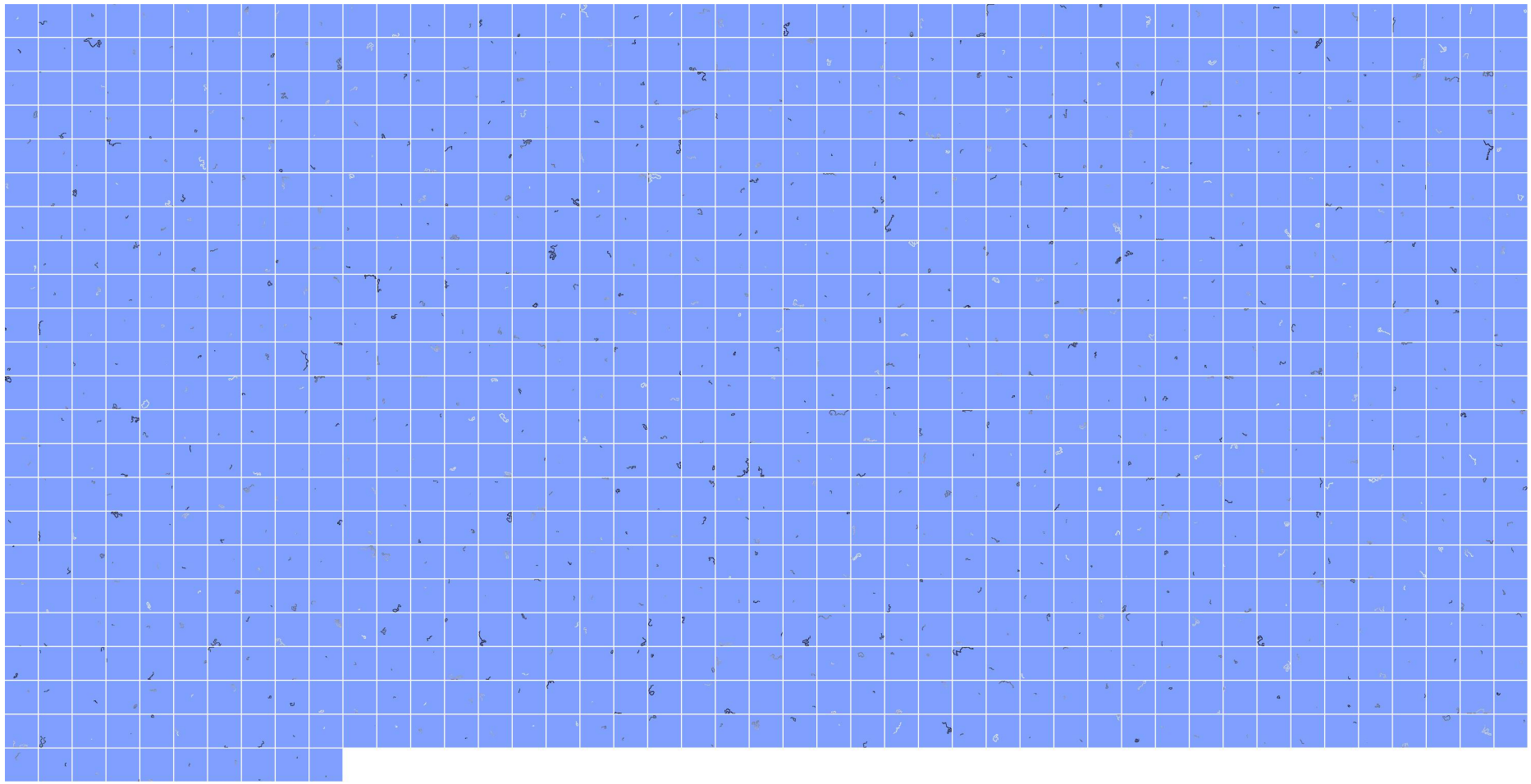
FIS : connected_1%



FIS : curve_1%



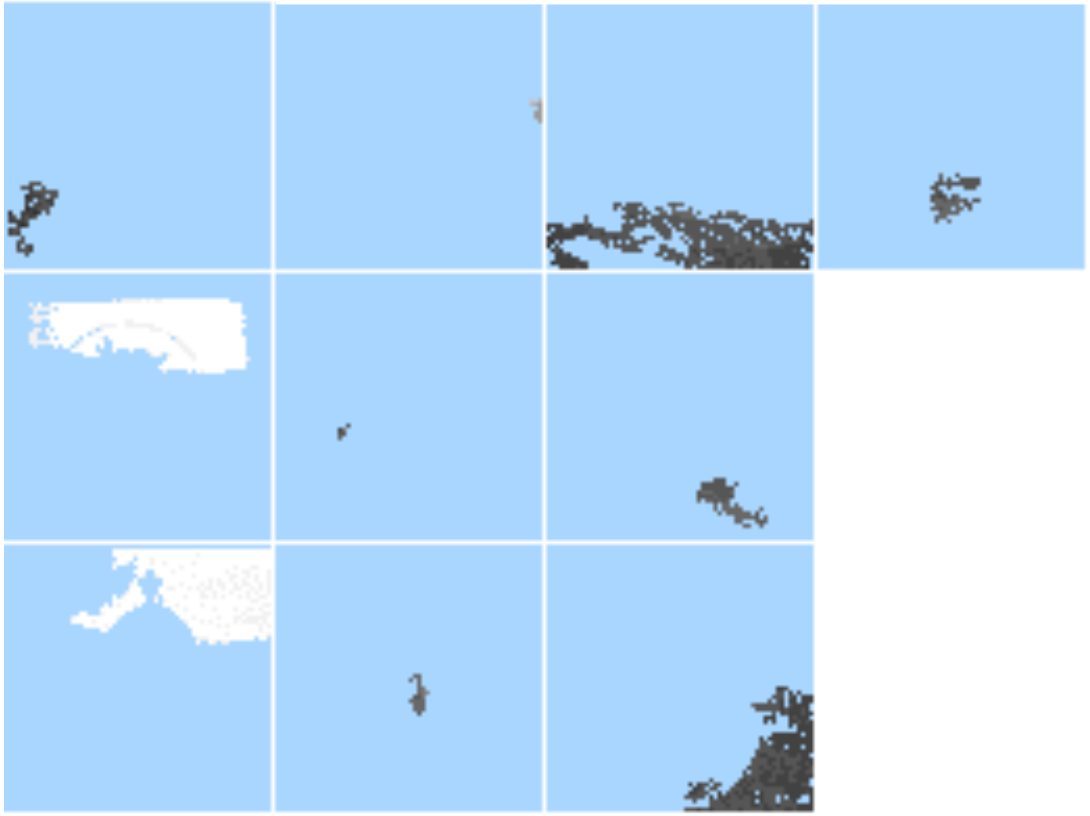
FIS : curve_1%

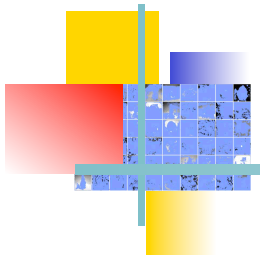


Code for an image

Image

The FIS taken in the dictionary





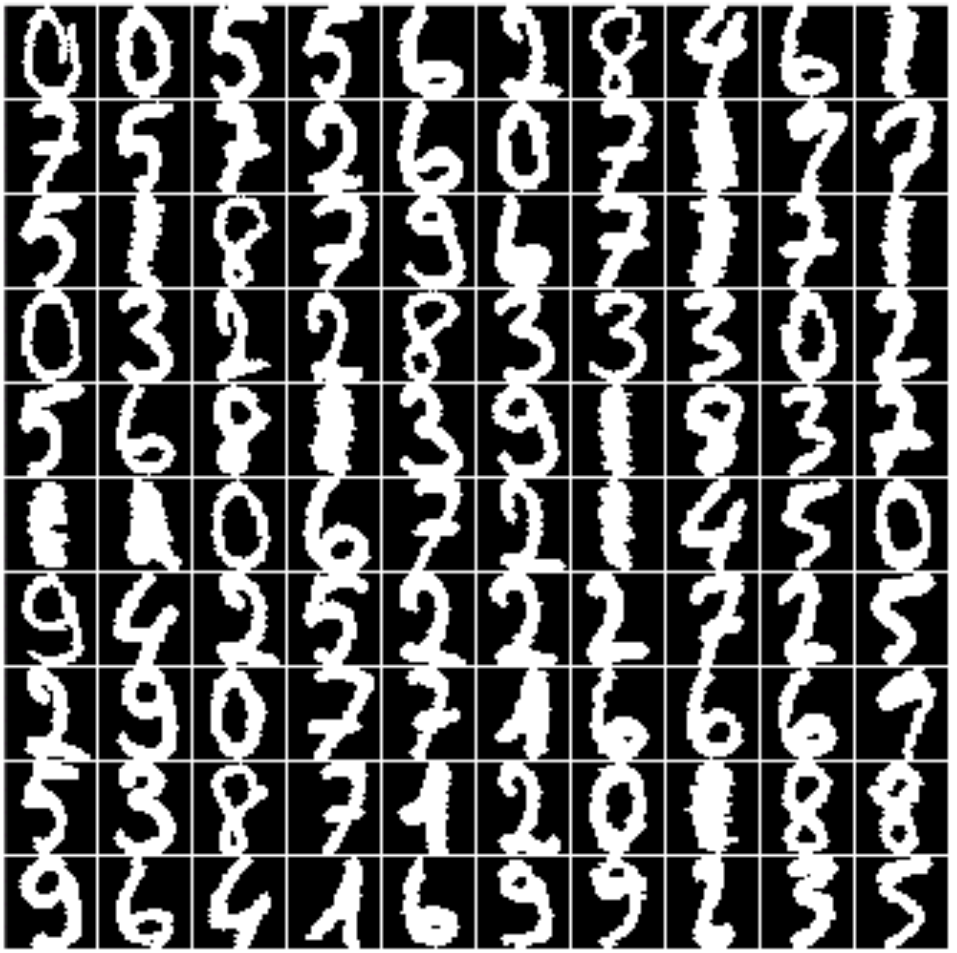
Analysis

- **Quite difficult to interpret !!**
- **No contours**, even with constraints to this effect
- Unfortunately **no possible comparison with ICA** since ICA is impracticable

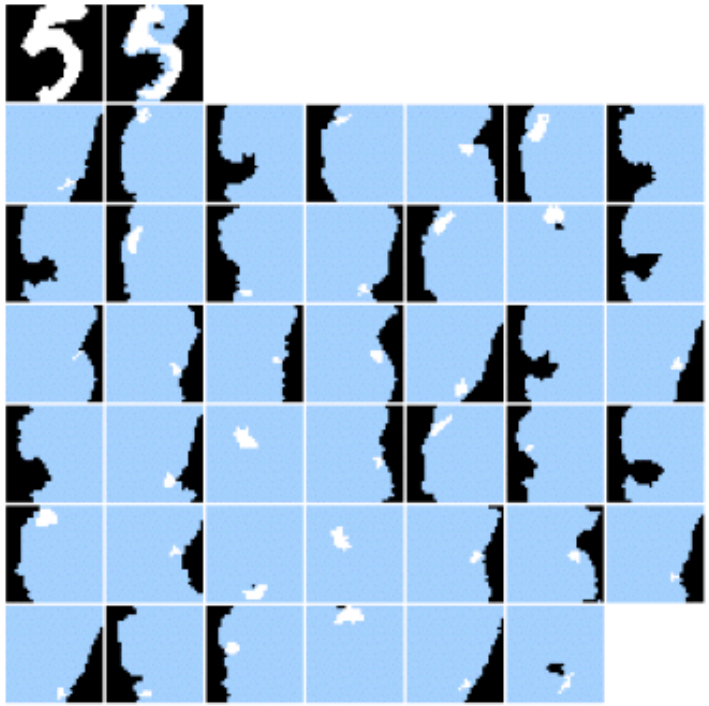
Hand-written digits

- **3000 examples**
- **B/W**

url : http://
ftp.ics.uci.edu/pub/
machine-learning-
databases/optdigits/



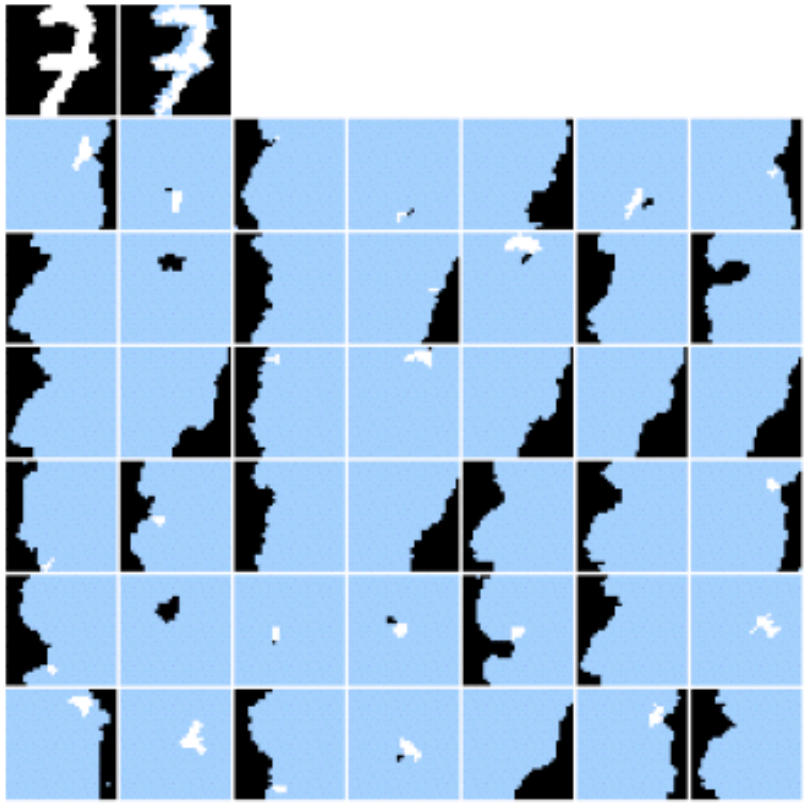
Example codes for the 5



FIS of coverage 5%
(41 / 1000)

FIS of coverage 1%
(10 / 1000)

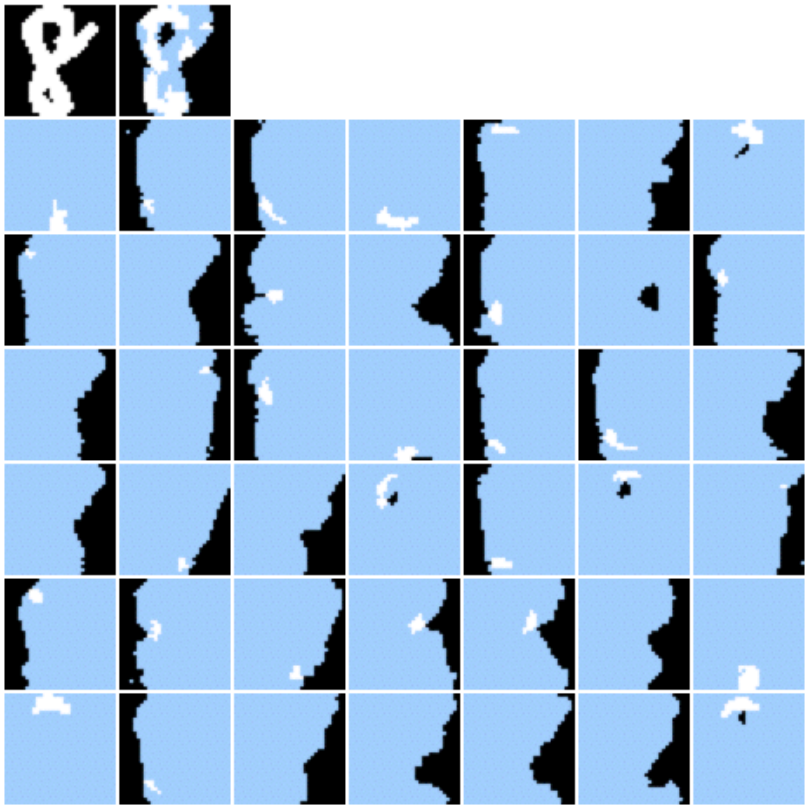
Example codes for the 7



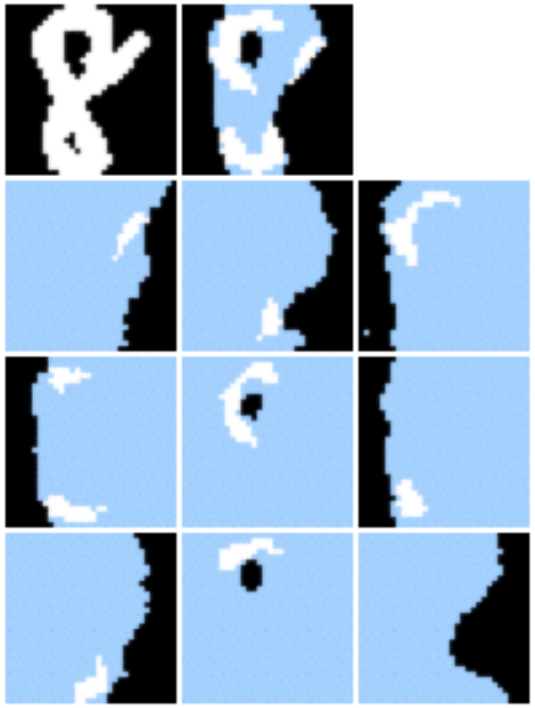
FIS of coverage 5%
(42 / 1000)

FIS of coverage 1%
(11 / 1000)

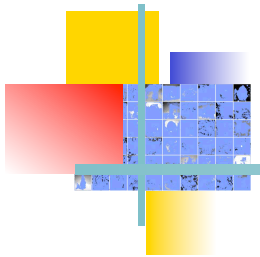
Example codes for the 8



FIS of coverage 5%
(42 / 1000)



FIS of coverage 1%
(9 / 1000)



Classification performance: protocole

- **Learning 1000 FIS on 540 training images**
- **Parameters :**
  - Image size (32 x 32, 64 x 64 ou 128 x 128)
  - Grey levels (16, 32 ou 64)
  - Coverage rates (1%, 2%, 5% ou 10%)

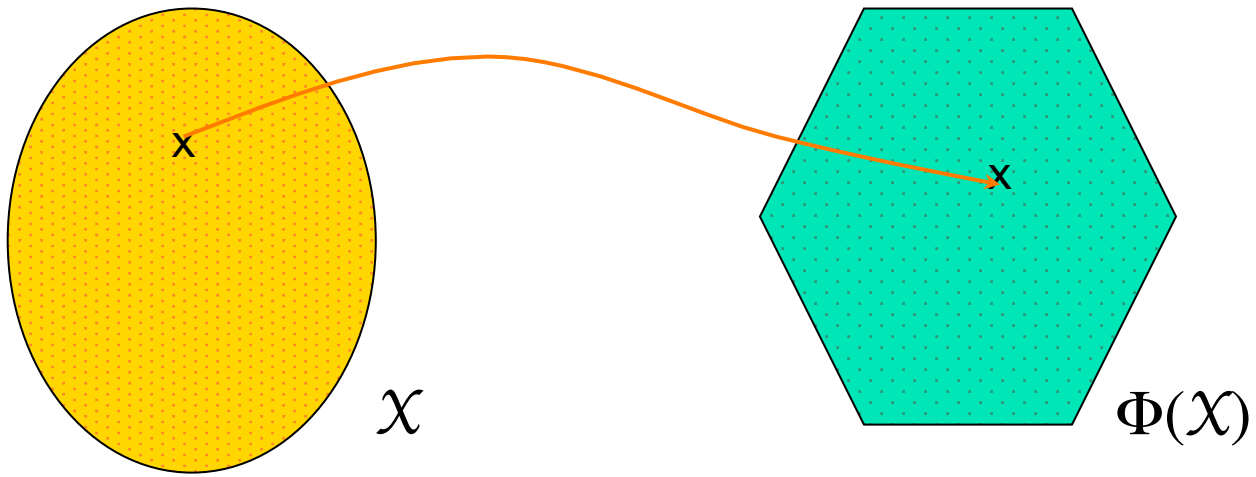
➥ **Test sur les 540 images restantes** (répété 10 fois)

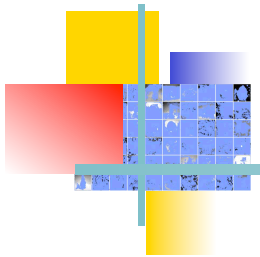
- NB : all results available on:

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

Classification : method

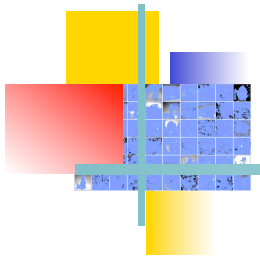
- **Each *example* (in $\mathcal{X}$) is *described by its FIS* (in $\Phi(\mathcal{X})$)**
- **A new example is labeled by a *k*-NN method (in the FIS space $\Phi(\mathcal{X})$)**
  - 1-ppv
  - or *k*-ppv weighted according to the distance





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%**



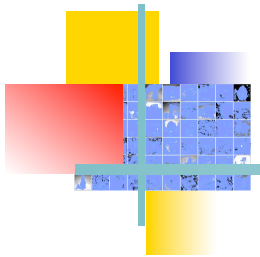
With a RBF neural network

	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

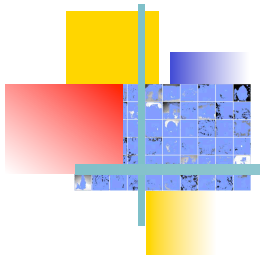
- Results
  - Best results for ε = 2 or 5 %
  - Similar for : *min, connexe, ligne*
  - Quite superior to RN or SVM (+ Gabor)
- Can be improved upon …
  - With a adaptive matching

Results for the digits

- **1000 FIS ; 10 ppv**

	0	1	2	3	4	5	6	7	8	9
0	**90**	1	2	.	.	6	1	.	.	.
1	.	**93**	6	.	.	1	.	.	.	.
2	.	2	**88**	.	.	4	.	6	.	.
3	.	3	11	**72**	.	2	1	12	.	.
4	1	8	.	.	**88**	2	.	.	1	.
5	.	1	2	2	.	**96**	.	.	.	.
6	4	4	5	.	.	1	**86**	.	.	.
7	.	6	.	.	.	5	1	**88**	.	.
8	.	21	13	9	.	21	1	13	**20**	1
9	2	3	4	21	4	10	1	11	.	**44**

trace 7.6548



Results for the digits

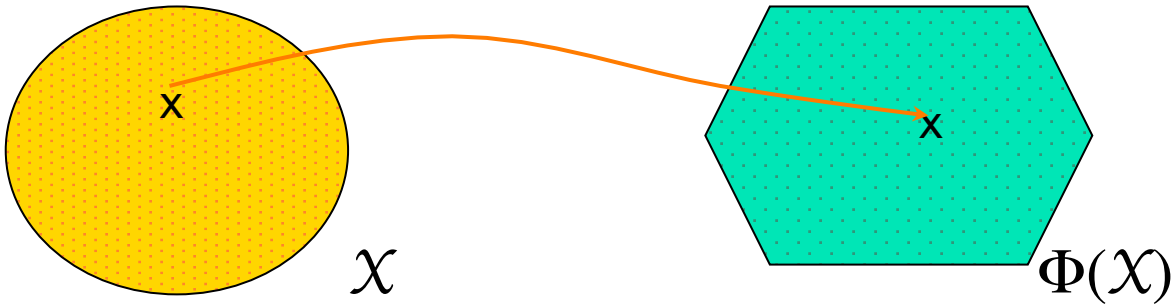
- **10 ppv in the input space**

	0	1	2	3	4	5	6	7	8	9
0	**100**	.	.	.	.	.	.	.	.	.
1	.	**100**	.	.	.	.	.	.	.	.
2	.	2	**97**	.	.	.	.	1	.	.
3	.	1	.	**97**	.	.	.	2	1	.
4	.	3	.	.	**97**	.	.	.	.	.
5	.	.	.	.	1	**98**	1	.	.	1
6	1	.	.	.	.	.	**99**	.	.	.
7	.	.	.	.	.	.	.	**98**	1	2
8	.	14	1	2	.	.	.	1	**80**	2
9	.	.	.	3	.	2	.	.	2	**94**

= 9.5993

Analysis

- **Properties of this coding scheme ?**
  - Unsupervised coding !!
  - And a *k*-NN classification method

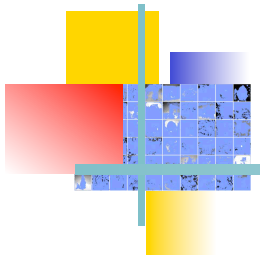


Properties ?

Approches classiques

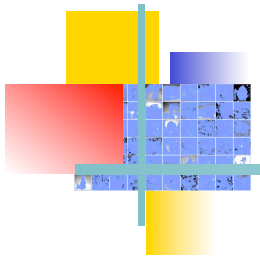
… et moins classiques

	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**



Conclusion

- **Analyse théorique en cours**
  - Ne rentre pas dans le cadre des théories de l'approximation

- **Expérimentations**
  - sur les ***scènes naturelles*** (poursuite du travail)
    - Contraindre la formation des FIS suivant les **règles de la Gestalt Theory**

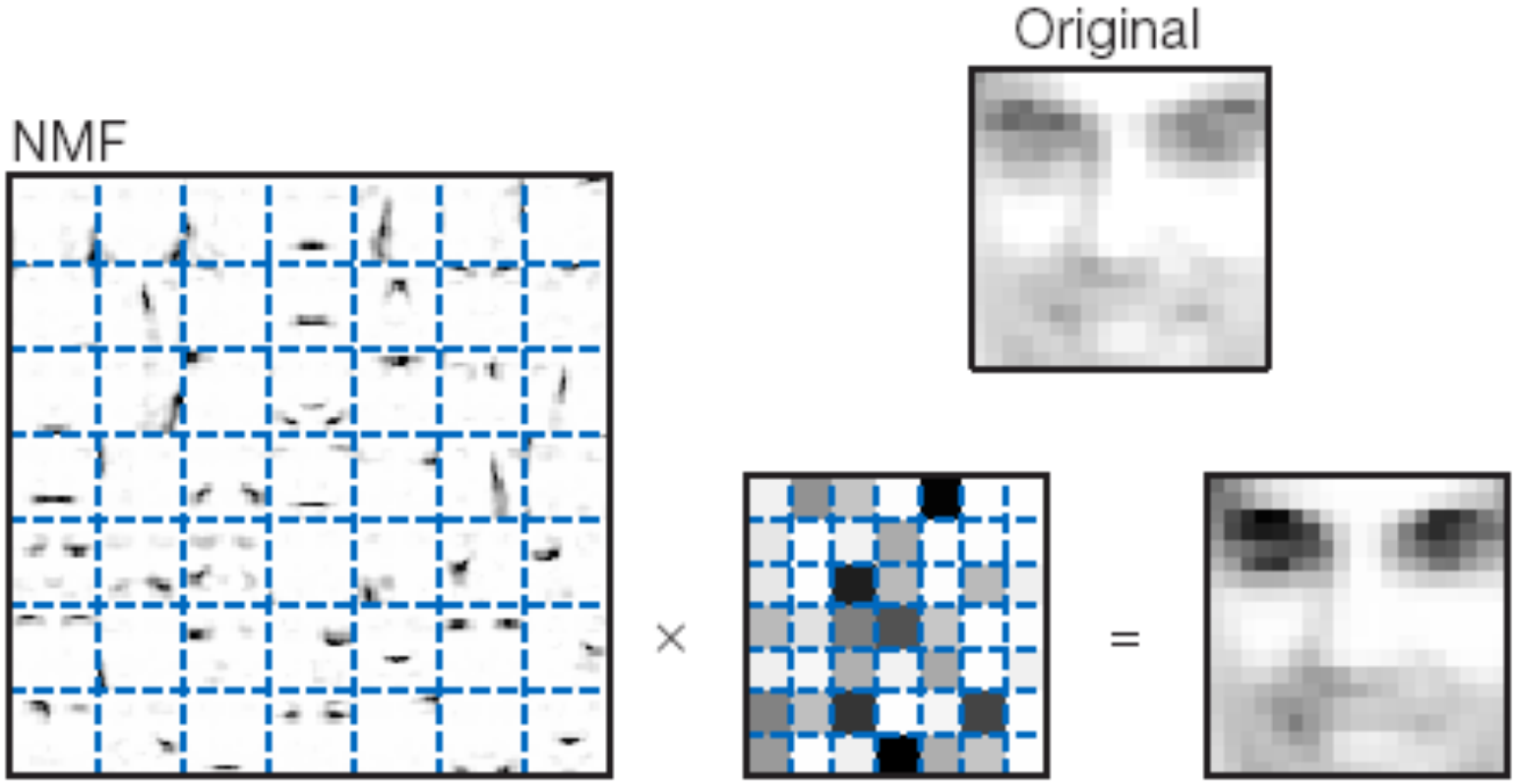
      Proximité / continuité / co-linéarité / co-circularité / régularité / clôture / …

    - Utiliser une **étape de prétraitement**

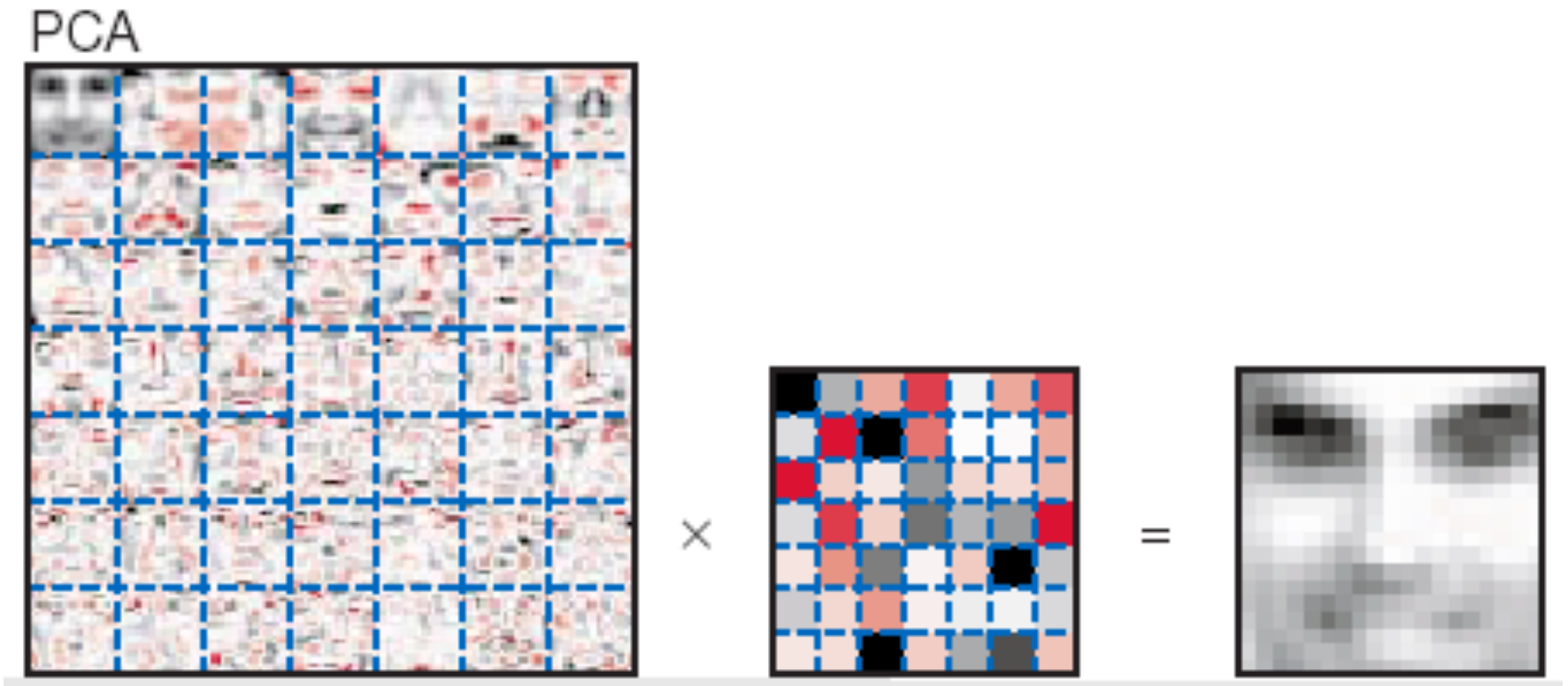
    - Comparaison avec la **NMF** (Non Negative Matrix Factorization)

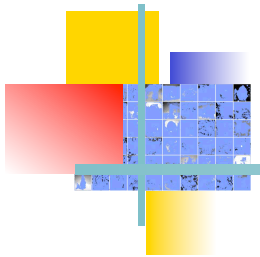
Non Negative Matrix Factorization

[Lee & Seung, Nature, 1999]



ACP





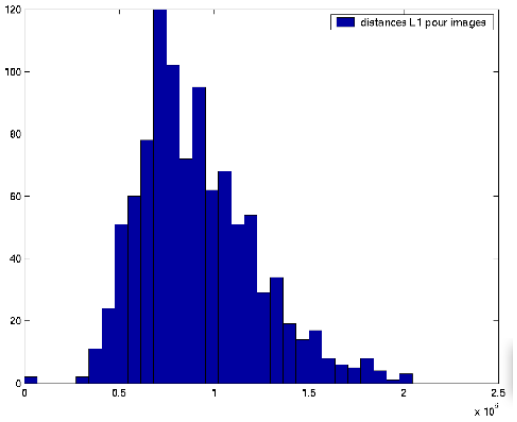
Conclusion

- sur les ***puces ADN***
  - Problème : très peu d'exemples
- sur la ***classification de textes*** de NewsGroups
  - Donne de bons résultats

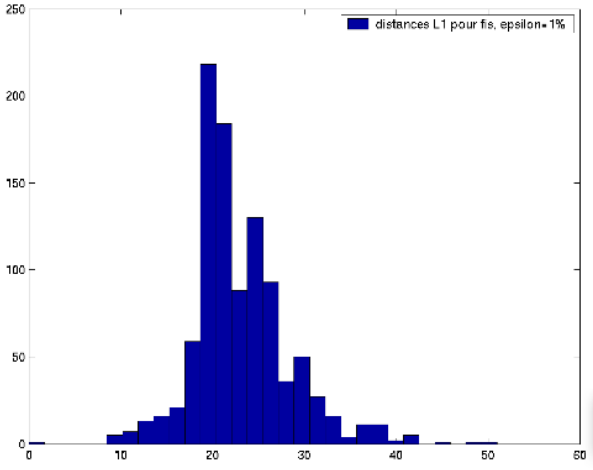
➠ ***Peut-être un nouveau type de traitement du signal***

Analyse théorique

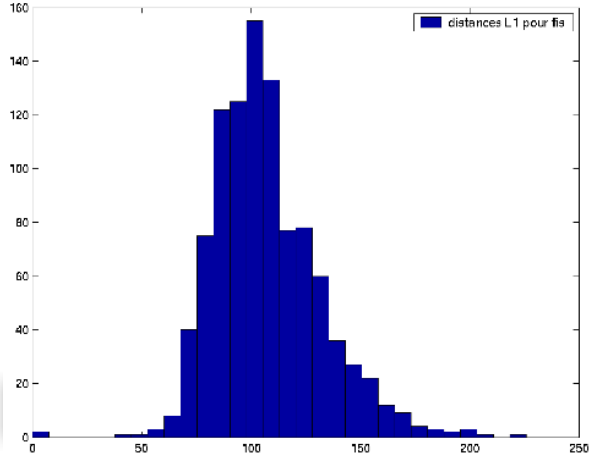
Histogrammes des distances entre images



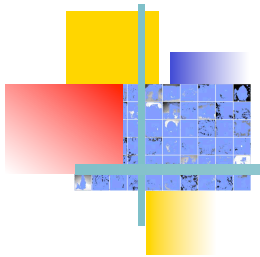
Espace des pixels



Espace des FIS_1%



Espace des FIS_5%



Bibliographie

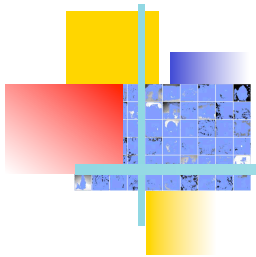
- Huang J. & Mumford D. **:** ***The statistics of natural images*****.**
- Olshausen B. & Field D. (1996) : ***Natural image statistics and efficient coding*****. Network.**

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



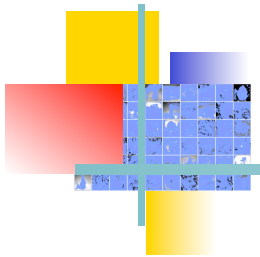
L'objectif

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

- **Apprentissage supervisé multi-classes**

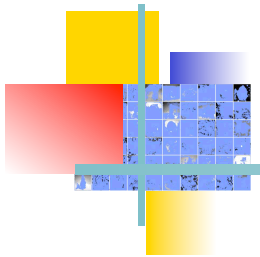
- Beaucoup de dimensions + peu d'exemples

  = Difficulté pour distinguer vraies régularités et coïncidences



Prétraitements

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



Cas de l'analyse de scènes

- Scènes naturelles ≠ scènes artificielles
- Observations neurobiologiques : **codage clairsemé**
- Hypothèse : signal résultant d'une **superposition de « formes latentes »**

  ➥ ***Analyse en composantes indépendantes*** (ACI)