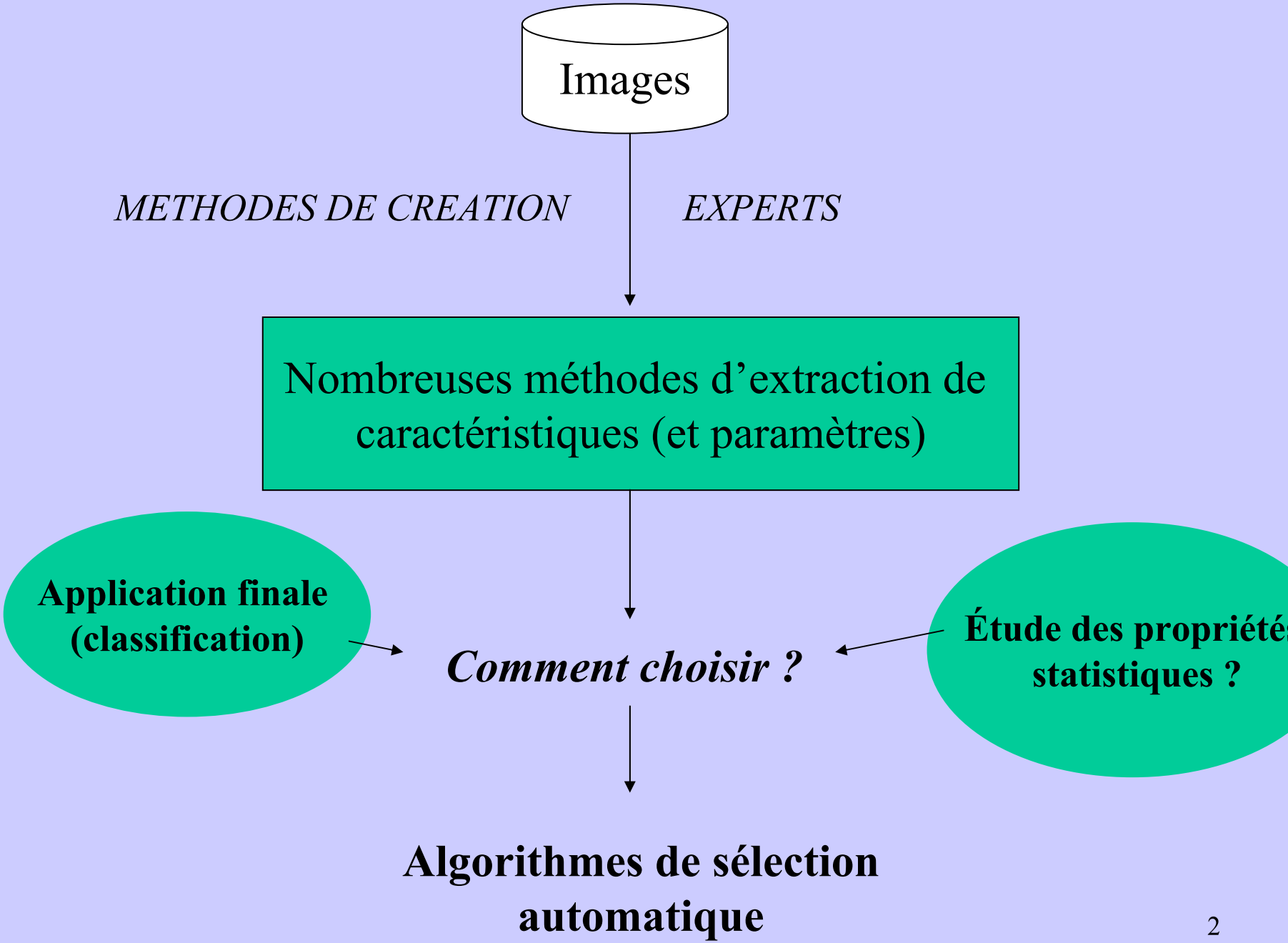
# Sélection et classification

**Marine Campedel**

29 novembre 2004

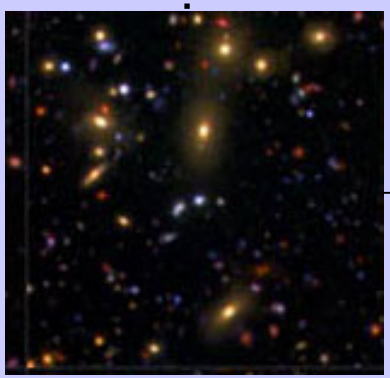
TELECOM PARIS
école nationale supérieure des télécommunications


Images
METHODES DE CREATION
EXPERTS
Nombreuses méthodes d'extraction de caractéristiques (et paramètres)
Application finale (classification)
Comment choisir ?
Étude des propriété statistiques ?
Algorithmes de sélection automatique

# Sélection de caractéristiques

- Motivation : mise au point d'une méthodologie objective pour sélectionner les caractéristiques *pertinentes* ;
- Objectif :
  - Réduction du nombre de caractéristiques stockées ;
  - Conservation/Amélioration de la performance de classification.

# Méthodologie (1/2)



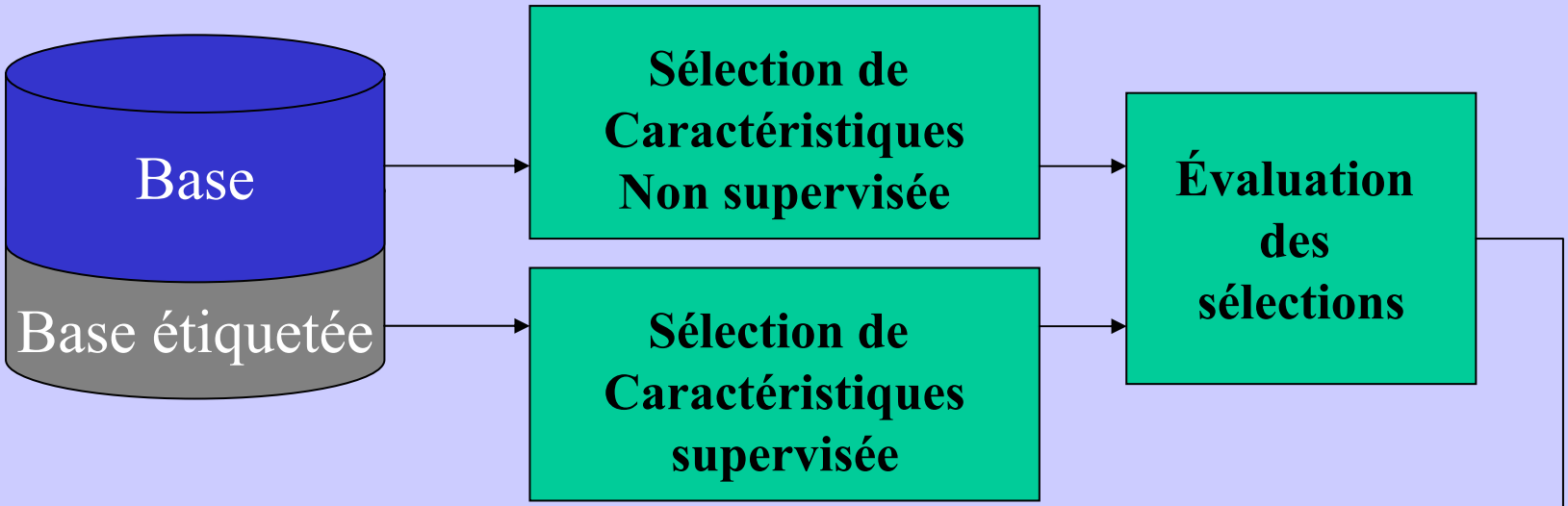
**Extraction de caractéristiques**

**Extraction de caractéristiques**

Base

Base étiquetée

# Méthodologie (2/2)




- **Ordonnancement des attributs**
- **Analyse du plus petit ensemble nécessaire**
- **Comparaison des méthodes de sélection**
- **Analyse des propriétés bruit/redondance**

# Algorithmes de sélection

- Recherche du sous-ensemble (le plus petit) de caractéristiques le plus apte à refléter la structure des données -> élimination du bruit et de la redondance ;

- Filter / Wrapper / Embedded ;
- Supervisés / non supervisés.

# Algorithmes supervisés

| | Type | Description |
|---|---|---|
| **ReliefF** | Filter | Score d'autant plus élevé que la caractéristique permet de discriminer les données de classes différentes. |
| **Fisher** | Wrapper | Analyse discriminante de Fisher. |
| **RFE** | Wrapper | Élimination récursive des caractéristiques de poids faible, à l'aide d'une SVM. |
| **AROM** | Wrapper | Approximation de la norme l0 des poids associés à chaque caractéristique, par une procédure récursive faisant intervenir une L1-SVM ou une SVM classique. |

# Algorithmes supervisés (suite)

- SVM (quadratique)

$$\min_{\mathbf{w},b} \|\mathbf{w}\|_2^2$$
avec $y_i(\mathbf{w}.\mathbf{x}_i + b) \geq 1$

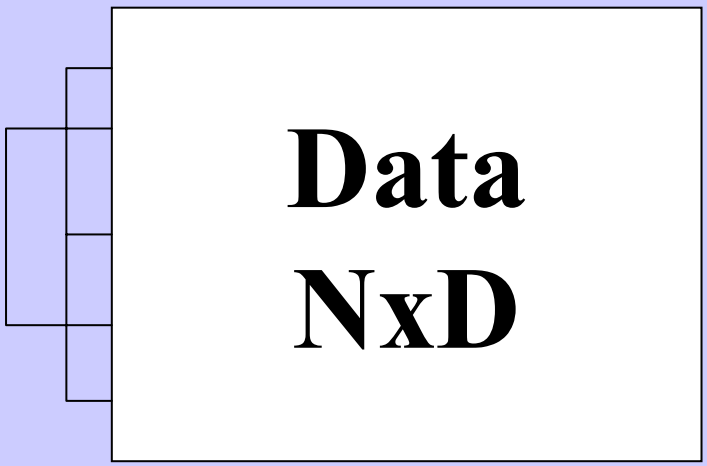
- L1-SVM (linéaire)

$$\min_{\mathbf{w},b} |\mathbf{w}|$$
avec $y_i(\mathbf{w}.\mathbf{x}_i + b) \geq 1$

- AROM -> L1-AROM et L2-AROM
- Sélection L1 : une boucle de sélection L1-AROM

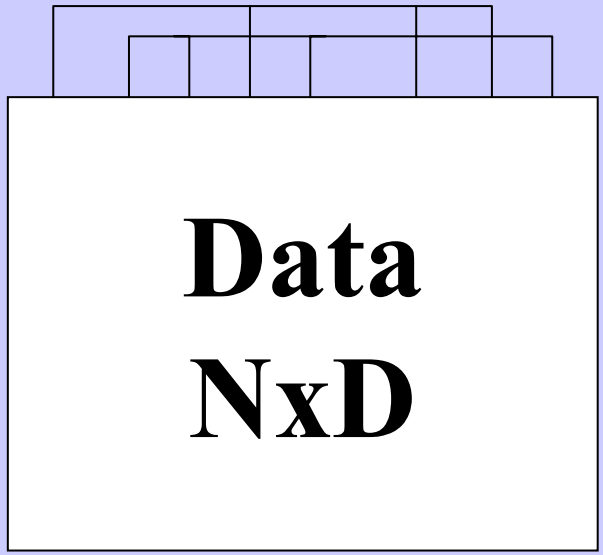
# Principe de la sélection non supervisée

- Approche classique




=> Recherche du sous-ensemble donnant la meilleure clusterisation des exemples

- Notre approche

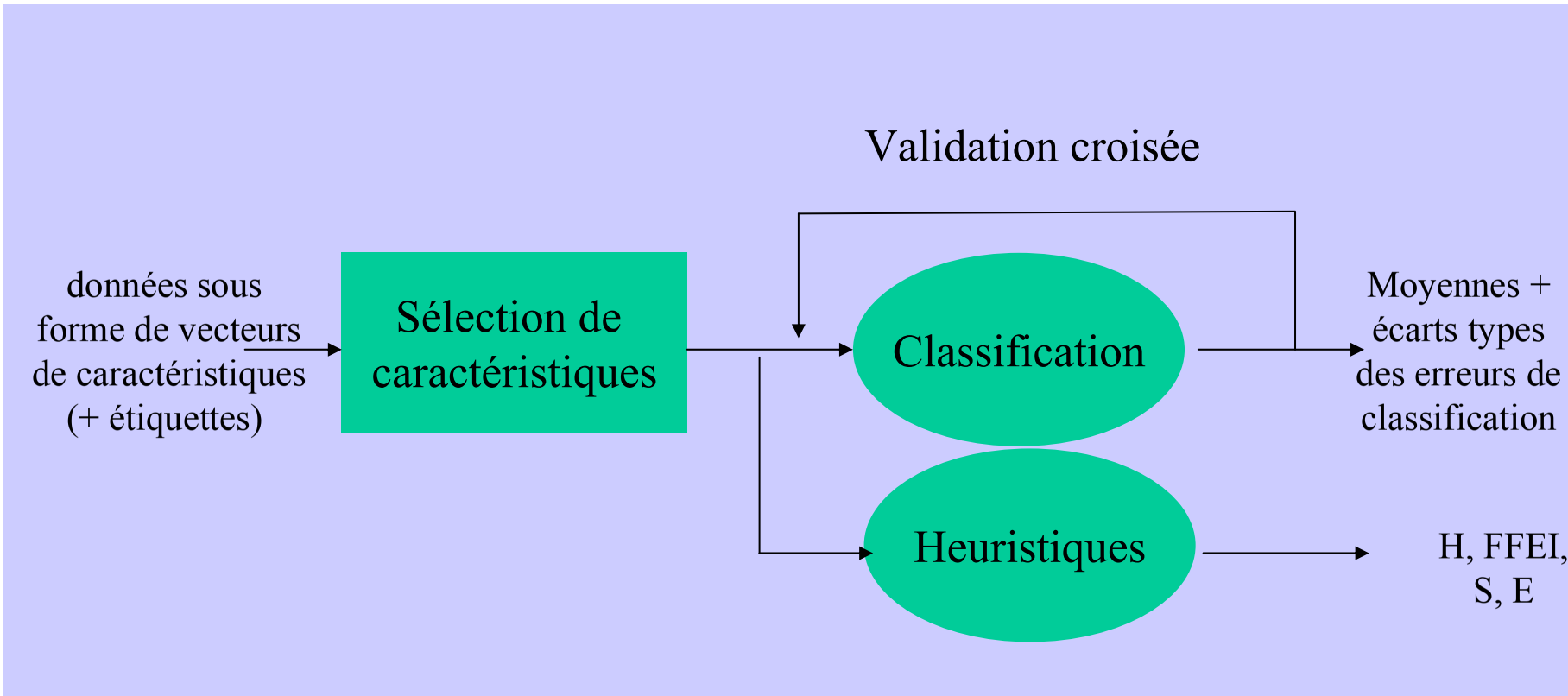



=> Choix des représentants des clusters de caractéristiques

# Algorithmes non supervisés

| | **Type** | **Description** |
|---|---|---|
| **Schéma général (lourd)** | Wrapper ou embedded (non implanté) | Recherche du sous-ensemble guidée par l'évaluation de la qualité d'une clusterisation des données. Utilisation d'algorithmes exploratoires (Greedy ou GA). |
| **MIC** | Filter | Utilisation d'une clusterisation KPPV des caractéristiques + choix d'un représentant par cluster. |
| **kMeans-FS** | Filter | Utilisation d'une clusterisation KMoyennes des caractéristiques +sélection de la caractéristique la plus proche du centroïde. |
| **SVC-FS** | Filter | Utilisation des vecteurs de support issus d'une classification 1-classe (clusterisation SVC) |

# Évaluation des sélections

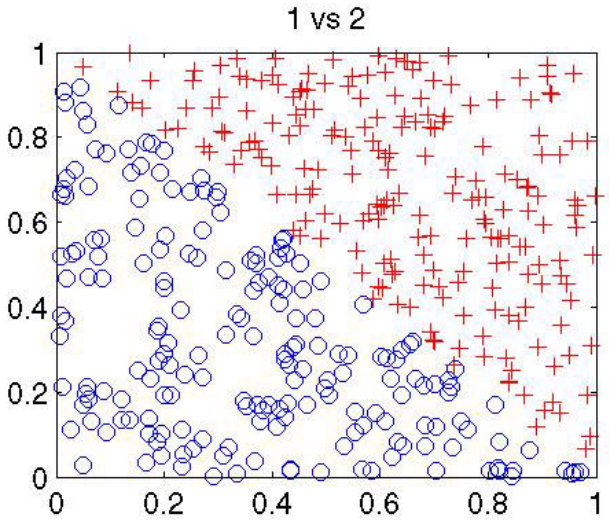


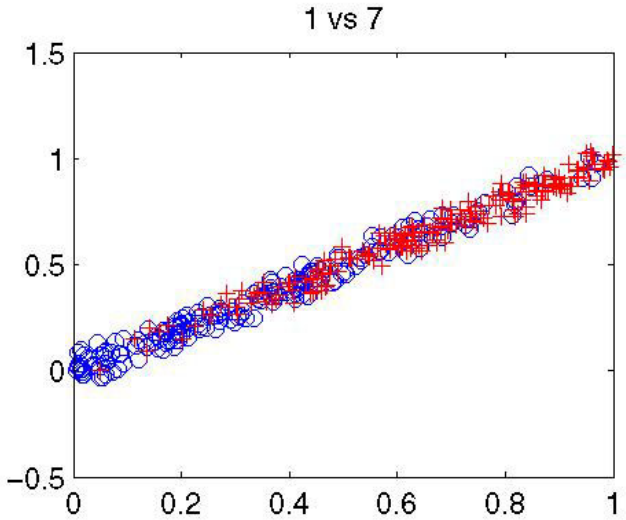
# Évaluation des sélections

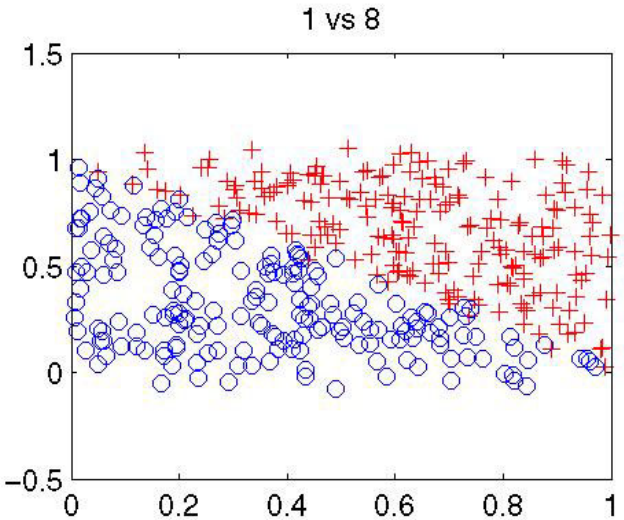
| | | |
|---|---|---|
| **Classificateurs** | **KNN** | K = sqrt( Ntrain ) |
| | **Fisher** | |
| | **SVM** | Noyau linéaire (svmlib2.5, C=1000) |
| | **L1-SVM** | |
| **Heuristiques** | **H** | Entropie de représentation |
| | **S** | Séparabilité des classes |
| | **FFEI** | Indice de flou |
| | **E** | Entropie |

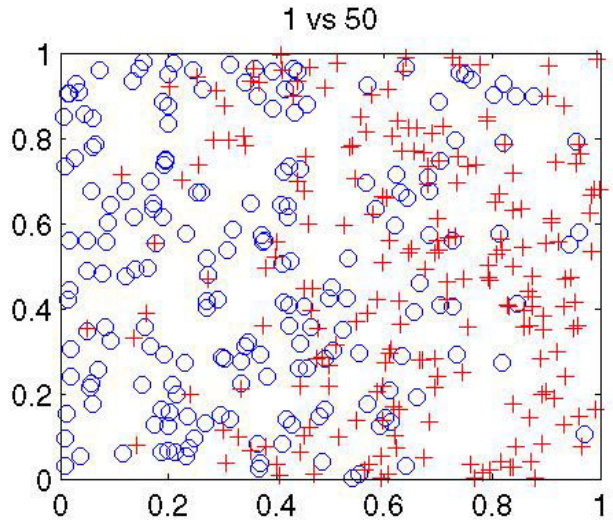
# Cas synthétique

- 50 attributs (distribution uniforme sur [0,1])
- 2 classes, 2 attributs pertinents ;
- 6 attributs corrélés à l'un ou l'autre des attributs pertinents ;
- Les méthodes non supervisées non adaptées car ne peuvent que réduire la redondance ;
- Fiabilité des sélections supervisées étudiées sur 50 tirages de données synthétiques

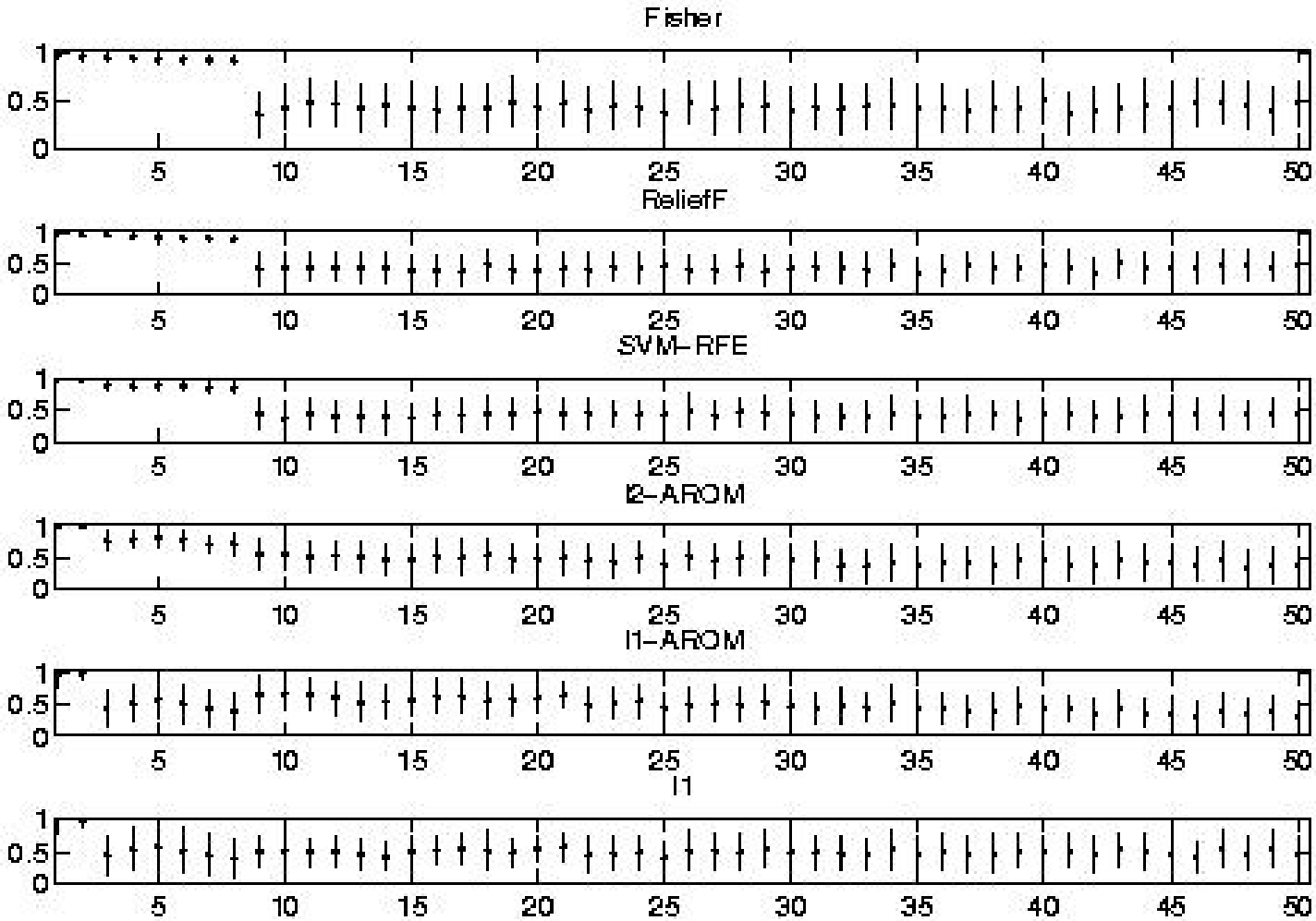

1 vs 2
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1


1 vs 7
1.5
1
0.5
0
−0.5
0
0.2
0.4
0.6
0.8
1


1 vs 8
1.5
1
0.5
0
−0.5
0
0.2
0.4
0.6
0.8
1


1 vs 50
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1

# Cas synthétique - Stabilité



# Cas synthétique - Évaluation

| Pour un tirage synthétique | | **Composantes sélectionnées** |
|---|---|---|
| **Sélection de 2** | **Fisher** | 4,6 |
| | **ReliefF** | 2,4 |
| | **SVM-RFE** | **1,2** |
| | **L2-AROM** | **1,2** |
| | **L1-AROM** | **1,2** |
| | **L1** | **1,2** |

# Cas synthétique – Heuristiques

| Synthétique | | E | H | S | FFEI |
|---|---|---|---|---|---|
| **Sélection de 2** | **Fisher** | **0.566** | 0.030 | 0.498 | **0.457** |
| | **ReliefF** | **0.563** | 0.017 | 0.403 | **0.460** |
| | **SVM-RFE** | 0.629 | **0.691** | **1.932** | 0.467 |
| | **L2-AROM** | 0.629 | **0.691** | **1.932** | 0.467 |
| | **L1-AROM** | 0.629 | **0.691** | **1.932** | 0.467 |
| | **L1** | 0.629 | **0.691** | **1.932** | 0.467 |

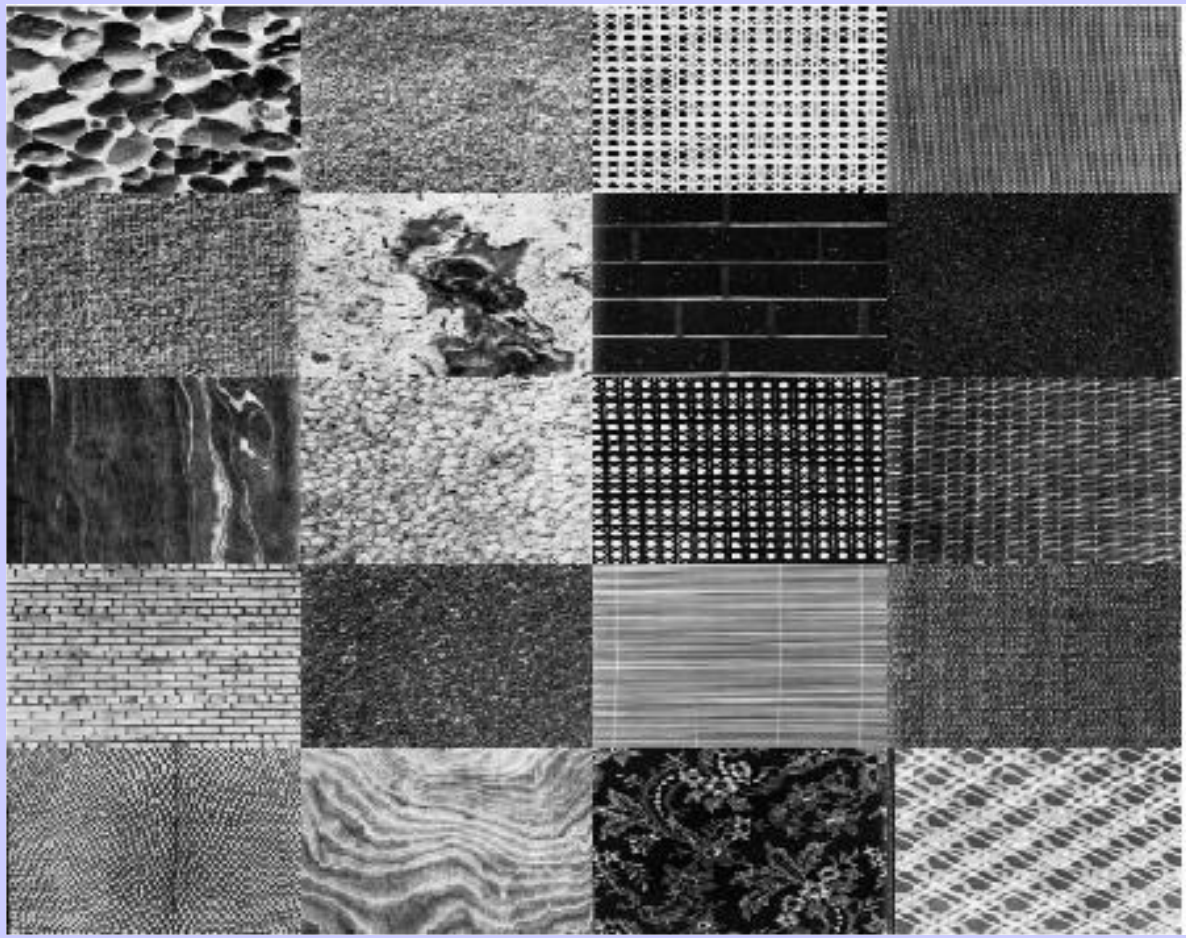
# Cas synthétique – Classificateurs

| Synthétique | | kppv | fisher | svm | L1-svm |
|---|---|---|---|---|---|
| **Tous les attributs** | | 12.3±3.8 | 5.3±1.0 | 7.0±2.2 | 3.8±2.2 |
| **Sélection de 2attributs** | **Fisher** | 30.0±3.9 | 26.0±3.9 | 25.8±4.8 | 26.0±3.6 |
| | **ReliefF** | 29.8±4.0 | 26.8±4.0 | 27.8±4.2 | 27.8±4.2 |
| | **SVM-RFE** | **3.8±3.5** | **2.8±2.9** | **0.5±0.6** | **0.5±0.6** |
| | **L2-AROM** | **3.8±3.5** | **2.8±2.9** | **0.5±0.6** | **0.5±0.6** |
| | **L1-AROM** | **3.8±3.5** | **2.8±2.9** | **0.5±0.6** | **0.5±0.6** |
| | **L1** | **3.8±3.5** | **2.8±2.9** | **0.5±0.6** | **0.5±0.6** |
| **Sélection de 8 attributs** | **Fisher** | **4.8±2.9** | **3.3±2.9** | 0.8±0.5 | 0.5±0.6 |
| | **ReliefF** | **4.8±2.9** | **3.3±2.9** | 0.8±0.5 | 0.5±0.6 |
| | **SVM-RFE** | 6.3±0.5 | 13.5±3.3 | **0.0±0.0** | 0.5±0.6 |
| | **L2-AROM** | 10.3±2.2 | **3.3±2.1** | 0.3±0.5 | **0.0±0.0** |
| | **L1-AROM** | 10.5±1.7 | 14.0±2.9 | **0.0±0.0** | **0.0±0.0** |
| | **L1** | 7.5±1.7 | 7.8±2.4 | 1.0±0.8 | 1.3±1.3 |

# Cas synthétique – 8 attributs

| Synthétique | | kppv | fisher | svm | L1-svm | H | S |
|---|---|---|---|---|---|---|---|
| **8 premiers attributs** | | 6.0±2.6 | 4.5±2.4 | 1.5±2.4 | 0.3±0.5 | | |
| **Sélection de 2** | **Fisher** | 30.5±4.0 | 25.3±3.6 | 26.5±4.4 | 26.5±4.4 | 0.017 | 0.49 |
| | **ReliefF** | **3.8±3.5** | **2.8±2.9** | **0.5±0.6** | **0.5±0.6** | **0.691** | **1.93** |
| | **SVM-RFE** | **3.8±3.5** | **2.8±2.9** | **0.5±0.6** | **0.5±0.6** | **0.691** | **1.93** |
| | **L2-AROM** | **3.8±3.5** | **2.8±2.9** | **0.5±0.6** | **0.5±0.6** | **0.691** | **1.93** |
| | **L1-AROM** | **3.8±3.5** | **2.8±2.9** | **0.5±0.6** | **0.5±0.6** | **0.691** | **1.93** |
| | **L1** | **3.8±3.5** | **2.8±2.9** | **0.5±0.6** | **0.5±0.6** | **0.691** | **1.93** |
| | **kMeans-FS** | 6.5±3.1 | 6.8±2.9 | 6.0±3.4 | 6.0±3.4 | **0.691** | 1.84 |
| | **MIC** | 31.0±4.6 | 26.5±4.4 | 27.5±4.5 | 27.5±4.5 | 0.016 | 0.40 |
| | **SVC-FS** | 5.0±3.4 | 3.8±2.1 | 2.8±1.7 | 2.8±1.7 | **0.691** | **1.92** |

# 20 images de Brodatz



# Résultats de classification

| Brodatz | Nb | kppv | fisher | svm | L1-svm |
|---|---|---|---|---|---|
| **Tout** | **180** | **4.2±1.3** | **4.4±2.3** | **2.6±2.1** | **1.2±1.3** |
| **Haralick** | **78** | **3.4±2.1** | **4.6±2.3** | **1.6±1.7** | **1.4±0.5** |
| **Gabor** | **24** | **4.8±1.6** | 6.4±2.2 | 6.4±4.0 | 12.4±1.7 |
| **Steer** | **26** | 13.8±3.6 | 8.2±3.6 | 7.6±1.8 | 7.8±2.3 |
| **Contourlet** | **26** | 14.4±1.8 | **4.4±2.6** | 8.4±3.8 | 8.6±3.8 |
| **Qmf** | **26** | 14.0±5.6 | **5.0±2.2** | 13.8±3.5 | 14.8±1.8 |

# Les coefficients d'Haralick

| Brodatz | | kppv | fisher | svm | L1-svm | H | S |
|---|---|---|---|---|---|---|---|
| **78 attributs** | | 3.4±2.1 | 4.6±2.3 | 1.6±1.7 | 1.4±0.5 | | |
| **Sélection de 20** | **Fisher** | **2.6±2.1** | **3.4±2.1** | **1.4 ±1.6** | **1.8±2.1** | 1.54 | 26.69 |
| | **ReliefF** | 6.6±3.4 | 12.8±3.6 | 3.0±1.2 | 6.4±3.3 | 1.17 | 25.66 |
| | **SVM-RFE** | 3.2±1.5 | 8.8±3.3 | 2.2±1.5 | 2.8±0.9 | 1.28 | 22.53 |
| | **L2-AROM** | 3.6±1.1 | 8.4±3.5 | 1.8±1.5 | 2.8±1.3 | 1.50 | 26.39 |
| | **L1-AROM** | 3.8±0.9 | 4.6±2.4 | **1.4±1.5** | 2.6±1.3 | 1.69 | **28.03** |
| | **L1** | 3.4±1.8 | 5.2±1.6 | **1.0±1.2** | 3.0±1.4 | 1.63 | 27.97 |
| | **kMeans-FS** | 3.2±1.5 | **4.6±2.3** | 2.4±2.9 | **1.8±2.2** | 1.64 | 27.10 |
| | **MIC** | **2.4±1.3** | 6.8±2.8 | 2.6±1.8 | 2.8±1.8 | 1.69 | 25.57 |
| | **SVC-FS** | **2.6±1.7** | **4.4±2.4** | 2.0±1.4 | **2.0±1.4** | **1.82** | 24.51 |

# Conclusion

- Les procédures de sélection automatiques permettent une comparaison objective des caractéristiques ;
- Les procédures non supervisées sont rapides et aussi efficaces que les procédures supervisées sur les bases (réelles) testées ;
- Nécessité de coupler les choix de l'algorithme de sélection et du classificateur dans l'application finale.
- S'applique à n'importe quel type de données.

# Perspectives à court terme

- Évaluation du nombre minimal de caractéristiques à sélectionner ;
- Extension de SVC-FS et KMeans-FS à la suppression des caractéristiques non pertinentes ;
- Mise au point d'une procédure d'évaluation complètement non supervisée (utilisation de clusterisation des données + heuristiques d'évaluation de la qualité des clusters).