# Classification Semi-Supervisée pour l'Identification de Cellules Phytoplanctoniques

**Guillaume Wacquet[1,2], Pierre-Alexandre Hébert [1,2], Émilie Caillault Poisson [1,2] et Denis Hamad [1,2]**

*1 Université Lille Nord de France, F-59000 Lille, France.*

*2 ULCO, LISIC, 50 rue Ferdinand Buisson, F-62228 CALAIS, France.*

*Prénom.Nom@lisic.univ-littoral.fr*



## Cadre de travail : Projet DYMAPHY

- **Projet**
  - Projet "DYMAPHY" – INTERREG IV A "2 Mers Seas Zeeën", programme de coopération transfrontalière 2007-2013, financé par le FEDER.

European Union
Co-funded by ERDF

- **Objectif**
  - Améliorer l'évaluation de la qualité des eaux marines dans la zone des 2 Mers, à travers l'étude du phytoplancton et des paramètres de l'environnement à haute résolution en utilisant une combinaison d'approches traditionnelles et nouvelles.

- **Participants**
  - France : Université du Littoral Côte d'Opale (ULCO - LISIC), Université de Lille1 (USTL - LOG), Institut français de recherche pour l'exploitation de la mer (IFREMER), Centre National de la Recherche Scientifique (CNRS),
  - Angleterre : Centre for Environment, Fisheries and Aquaculture Science (CEFAS),
  - Pays-Bas : Rijkswater Staat (RWS Zeeland).

# Cadre de travail

- **Pourquoi l'étude du Phytoplancton ?**
  - Préservation de l'environnement et de la biodiversité
    - 50% de l'activité photosynthétique totale de la planète,
    - 45% de la production primaire mondiale.
  - Tourisme, Transport, Ressources exploitables et Nourriture
    - 1er maillon de la chaîne alimentaire dans l'écosystème marin,
    - Environ 70 espèces toxiques ou nuisibles.

- **Pourquoi la classification semi-supervisée ?**
  - Partitionnement des données non étiquetées en utilisant quelques informations de groupement,
  - Données étiquetées :
    - Coût important en terme de temps et de complexité d'étiquetage (analyse microscopique),
    - Nécessité de recourir à un expert du domaine,
    - Utilisation de dispositifs spécifiques.

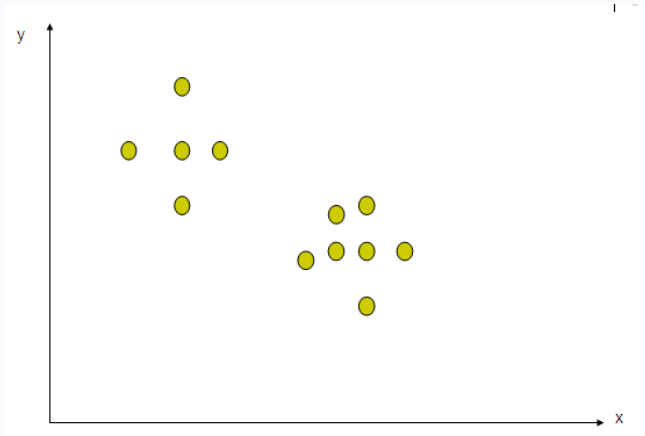


# Classification spectrale contrainte (PCSC)

- **Classification spectrale**
  - Méthode de partitionnement en *K* groupes distincts,
  - Obtention d'un espace de représentation des données,
  - Capacité de traiter des données de structure non globulaire,
  - Implémentation simple et efficace

➡ Extraction de vecteurs propres d'une matrice de similarités.

➡ *Exemple :* 12 points projetés sur un plan 2D.



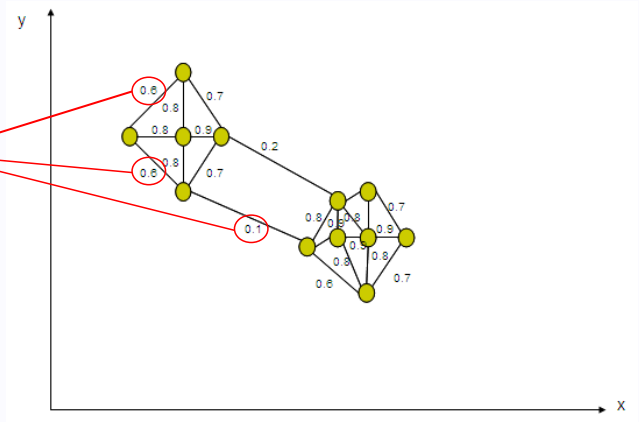
# Classification spectrale contrainte (PCSC)

- **Classification spectrale**
  - Méthode de partitionnement en *K* groupes distincts,
  - Obtention d'un espace de représentation des données,
  - Capacité de traiter des données de structure non globulaire,
  - Implémentation simple et efficace.

➡ Extraction de vecteurs propres d'une matrice de similarités.

➡ *Exemple :* Construction d'un graphe pondéré des données.

Similarités entre points ($S_{ij}$)





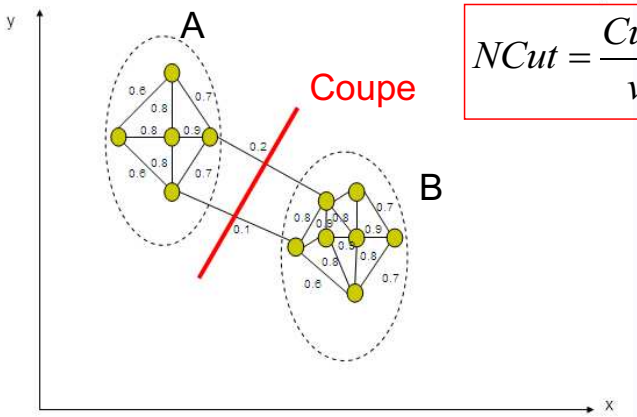


# Classification spectrale contrainte (PCSC)

- **Classification spectrale**
  - Méthode de partitionnement en *K* groupes distincts,
  - Obtention d'un espace de représentation des données,
  - Capacité de traiter des données de structure non globulaire,
  - Implémentation simple et efficace

➡ Extraction de vecteurs propres d'une matrice de similarités.


➡ *Exemple :* Recherche de la coupe normalisée optimale.



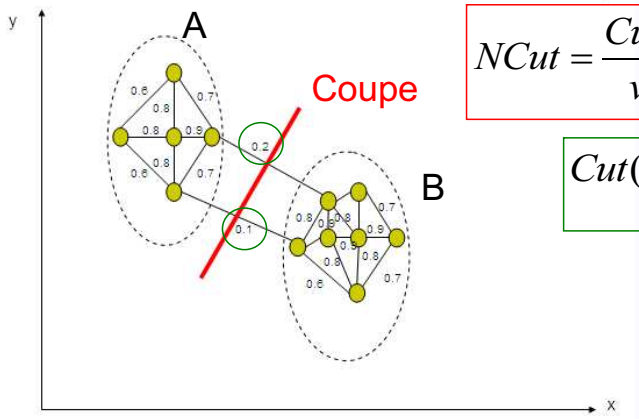

$$NCut = \frac{Cut(A,B)}{vol(A)} + \frac{Cut(A,B)}{vol(B)}$$

# Classification spectrale contrainte (PCSC)

- **Classification spectrale**
  - Méthode de partitionnement en *K* groupes distincts,
  - Obtention d'un espace de représentation des données,
  - Capacité de traiter des données de structure non globulaire,
  - Implémentation simple et efficace

➡ Extraction de vecteurs propres d'une matrice de similarités.

➡ *Exemple :* Recherche de la coupe normalisée optimale.




$$NCut = \frac{Cut(A,B)}{vol(A)} + \frac{Cut(A,B)}{vol(B)}$$

$$Cut(A,B) = \sum_{i \in A, j \in B} S_{ij}$$



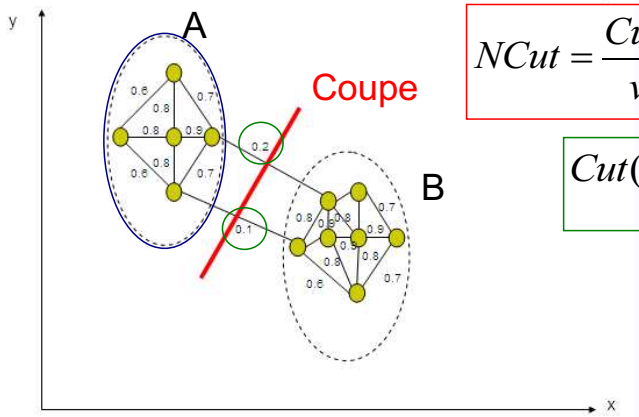

# Classification spectrale contrainte (PCSC)

- **Classification spectrale**
  - Méthode de partitionnement en *K* groupes distincts,
  - Obtention d'un espace de représentation des données,
  - Capacité de traiter des données de structure non globulaire,
  - Implémentation simple et efficace

➡ Extraction de vecteurs propres d'une matrice de similarités.

➡ *Exemple :* Recherche de la coupe normalisée optimale.


$$vol(A) = \sum_{i \in A} \sum_{j=1}^{N} S_{ij}$$





$$NCut = \frac{Cut(A,B)}{vol(A)} + \frac{Cut(A,B)}{vol(B)}$$

$$Cut(A,B) = \sum_{i \in A, j \in B} S_{ij}$$

# Classification spectrale contrainte (PCSC)

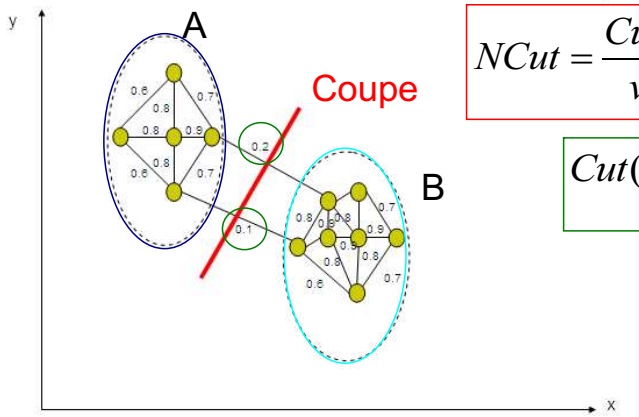
- **Classification spectrale**
  - Méthode de partitionnement en *K* groupes distincts,
  - Obtention d'un espace de représentation des données,
  - Capacité de traiter des données de structure non globulaire,
  - Implémentation simple et efficace

➡ Extraction de vecteurs propres d'une matrice de similarités.

➡ *Exemple :* Recherche de la coupe normalisée optimale.

$$vol(A) = \sum_{i \in A} \sum_{j=1}^{N} S_{ij}$$

$$vol(B) = \sum_{i \in B} \sum_{j=1}^{N} S_{ij}$$




$$NCut = \frac{Cut(A,B)}{vol(A)} + \frac{Cut(A,B)}{vol(B)}$$

$$Cut(A,B) = \sum_{i \in A, j \in B} S_{ij}$$




# Classification spectrale contrainte (PCSC)

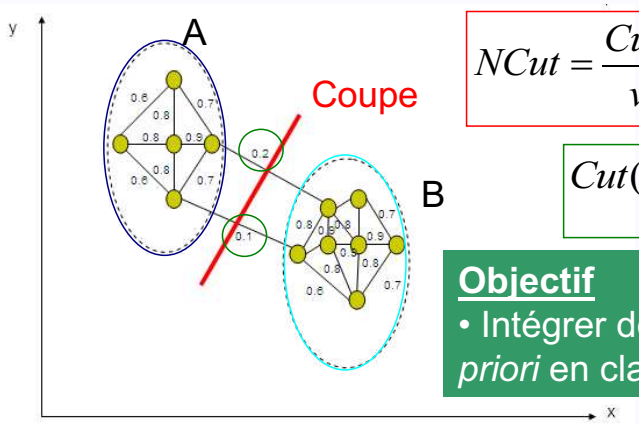
- **Classification spectrale**
  - Méthode de partitionnement en *K* groupes distincts,
  - Obtention d'un espace de représentation des données,
  - Capacité de traiter des données de structure non globulaire,
  - Implémentation simple et efficace

➡ Extraction de vecteurs propres d'une matrice de similarités.

➡ *Exemple :* Recherche de la coupe normalisée optimale.

$$vol(A) = \sum_{i \in A} \sum_{j=1}^{N} S_{ij}$$

$$vol(B) = \sum_{i \in B} \sum_{j=1}^{N} S_{ij}$$



$$NCut = \frac{Cut(A,B)}{vol(A)} + \frac{Cut(A,B)}{vol(B)}$$

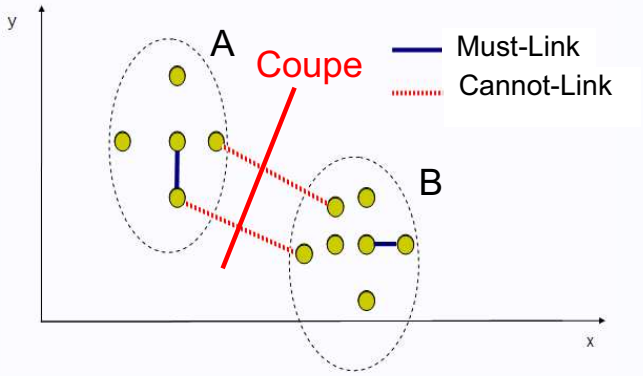
$$Cut(A,B) = \sum_{i \in A, j \in B} S_{ij}$$


**Objectif**
• Intégrer des connaissances *a priori* en classification spectrale

# Classification spectrale contrainte (PCSC)

- **Type de connaissances *a priori***
  - Contraintes de comparaison Must-Link
    → Deux objets doivent être dans le même groupe,
  - Contraintes de comparaison Cannot-Link
    → Deux objets ne doivent pas être dans le même groupe.

*Exemple :* Recherche de coupe optimale d'un graphe contraint.




**Objectifs**
- Optimiser un critère (fonction objective) intégrant les contraintes.
- Justifier l'efficacité et montrer la pertinence de notre méthode.




# Classification spectrale contrainte (PCSC)


- **Classification spectrale "classique"**

| | |
|---|---|
| $G$ | Un graphe avec N noeuds |
| $S$ | La matrice de similarités |
| $D$ | La matrice de degrés de $S$ |
| $L$ $(L_{norm})$ | La matrice Laplacienne (normalisée) |
| $y$ $(w)$ | Le vecteur indicateur de classe (normalisé) |

Avec $D = diag(D_{11},...,D_{NN})$ telle que $D_{ii} = \sum_{j=1}^{n} S_{ij}$

et $L_{norm} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$ avec $L = D - S$

et $w = D^{\frac{1}{2}} y$

# Classification spectrale contrainte (PCSC)

- **Classification spectrale "classique"**
  - Minimisation du critère de Coupe Normalisée, exprimant la cohésion interne des groupes, relativement à leur dissociation les uns des autres :

$$MNCut(G,Y) = \sum_{k=1}^{K} \frac{y_k^t (D-S) y_k}{y_k^t D y_k} = \sum_{k=1}^{K} \frac{y_k^t L y_k}{y_k^t D y_k} \quad \text{avec} \quad Y = [y_1, ..., y_K]$$

  - Changement de variable : $y = D^{-1/2} w$ pour mettre en évidence un quotient de Rayleigh :

$$MNCut(G,w) = \sum_{k=1}^{K} \frac{w_k^t D^{-\frac{1}{2}} L D^{-\frac{1}{2}} w_k}{w_k^t w_k} = \sum_{k=1}^{K} \frac{w_k^t L_{norm} w_k}{w_k^t w_k} \quad \text{avec} \quad W = [w_1, ..., w_K]$$

  - Solution à valeurs réelles (approximation de $y_k$ ) :
    - Solution donnée par les *K* vecteurs propres de $L_{norm}$ associés aux *K* plus petites valeurs propres.
  - Solution à valeurs discrètes (obtention de la partition) :
    - Application de l'algorithme des K-moyennes sur la matrice constituée des *K* vecteurs propres solutions (vecteurs colonnes).



# Classification spectrale contrainte (PCSC)

- **Intégration des contraintes par paires**
  - Connaissances additionnelles sous la forme de deux ensembles de paires d'objets :
    - les paires d'objets appartenant à des groupes différents : $\{x_i, x_j\} \in CL$ avec $\{x_i, x_j\} \subseteq X$, *CL* pour Cannot-Link,
    - les paires d'objets appartenant au même groupe : $\{x_i, x_j\} \in ML$ avec $\{x_i, x_j\} \subseteq X$, *ML* pour Must-Link.
  - Coût de pénalisation $PC$, fonction des distances dans l'espace spectral des $y_k$, dans le but de rapprocher/éloigner les projections spectrales des points appariés en *ML* ou en *CL* :

$$PC(CL, ML, Y, \alpha, \beta) = \sum_{k=1}^{K} [-\frac{\alpha}{|CL|} \sum_{\{x_i, x_j\} \in CL} (y_{ik} - y_{jk})^2 + \frac{\beta}{|ML|} \sum_{\{x_i, x_j\} \in ML} (y_{ik} - y_{jk})^2]$$

    - Contraintes sur les projections spectrales.

# Classification spectrale contrainte (PCSC)

- **Intégration des contraintes par paires**

$$PC(CL,ML,Y,\alpha,\beta)=\sum_{k=1}^{K}[-\frac{\alpha}{|CL|}\sum_{\{x_i,x_j\}\in CL}(y_{ik}-y_{jk})^2+\frac{\beta}{|ML|}\sum_{\{x_i,x_j\}\in ML}(y_{ik}-y_{jk})^2]$$

$$=\sum_{k=1}^{K}[\frac{1}{2}\sum_{i,j}(y_{ik}-y_{jk})^2 Q_{ij}] \text{ avec } Q_{ij}=\begin{cases}-\alpha/|CL| \text{ si } \{x_i,x_j\}\in CL\\ +\beta/|ML| \text{ si } \{x_i,x_j\}\in ML\\ 0 \text{ sinon}\end{cases}$$

Forme matricielle de *PC* :

$$PC(CL,ML,W)=\sum_{k=1}^{K}[w_k^t D^{-\frac{1}{2}}(R-Q)D^{-\frac{1}{2}}w_k] \text{ avec } w_k=D^{\frac{1}{2}}y_k$$

$$\text{et } R_{ii}=\sum_{j=1}^{n}Q_{ij}$$

Quotient de Rayleigh :

$$PC(CL,ML,W)=\sum_{k=1}^{K}\frac{w_k^t D^{-\frac{1}{2}}(R-Q)D^{-\frac{1}{2}}w_k}{w_k^t w_k}$$



# Classification spectrale contrainte (PCSC)

- **Intégration des contraintes par paires**
  - Fonction objective (critère à minimiser) :

$$J(G,CL,ML,W)=MNCut(G,W)+PC(CL,ML,W)$$

$$=\sum_{k=1}^{K}\frac{w_k^t(D^{-\frac{1}{2}}LD^{-\frac{1}{2}})w_k+w_k^t(D^{-\frac{1}{2}}(R-Q)D^{-\frac{1}{2}})w_k}{w_k^t w_k} \text{ avec } L=D-S$$

$$J(G,CL,ML,W)=\sum_{k=1}^{K}\frac{w_k^t(I-D^{-\frac{1}{2}}(S+Q-R)D^{-\frac{1}{2}})w_k}{w_k^t w_k}$$

Optimisation de la fonction *J* :

Processus de minimisation du critère identique à celui de la classification spectrale « classique » mais avec une matrice de similarités modifiée.

$$S^*=S+Q-R$$

# Applications : Protocole expérimental

- **Comparaison de PCSC avec 2 algorithmes récents utilisant les contraintes par paires**
  - "Spectral Learning" [Kamvar et al., 2003] (**SL**)

    Modification des valeurs de similarités :
    - Si $\{x_i, x_j\} \in CL$ alors $S_{ij} = 0$,
    - Si $\{x_i, x_j\} \in ML$ alors $S_{ij} = 1$.
  - "Flexible Constrained Spectral Clustering" [Wang et al, 2010] (**FCSC**)
    - Formalisation Lagrangienne : $\min MNCut$ sc. $ML$ et $CL$
    - Résolution d'un système de valeurs propres généralisées :

$$L_{norm} v = \lambda (D^{-\frac{1}{2}} Q D^{-\frac{1}{2}} - \beta I) v$$

Avec $Q_{ij} = \begin{cases} -1 \text{ si } \{x_i, x_j\} \in CL \\ +1 \text{ si } \{x_i, x_j\} \in ML \\ 0 \text{ sinon} \end{cases}$, $L_{norm} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$ : matrice Laplacienne normalisée,

$\beta$ : paramètre de pondération fixé *a priori.*



# Applications : Protocole expérimental

- **Variantes (pour expérimentations)**
  - Projection de l'espace spectral sur la sphère-unité, juste avant l'étape de discrétisation (**SL**, **FCSC**, et **PCSC**) [Ng et al., 2002],
  - Utilisation de la matrice de contraintes de [Wang et al., 2010] (**FCSC**) :

    Valeurs des contributions ML et CL identiques : $\frac{\alpha}{|CL|} = \frac{\beta}{|ML|} = 1$
  - Spécification de l'importance relative des deux objectifs (**PCSC**) :

    $0 \le MNCut \le 1$ et $\lambda_{Q\min} \le PC \le \lambda_{Q\max}$

    → Normalisation de Q pour avoir $0 \le PC_{norm} \le 1$ : $Q_{norm} = \frac{Q - \lambda_{Q\min}}{\lambda_{Q\max} - \lambda_{Q\min}}$

    Paramètre de pondération $\gamma$ :

$$J(G, CL, ML, \gamma, W) = (1-\gamma).MNCut(G, W) + \gamma.PC_{norm}(CL, ML, W)$$

# Application : Bases de données UCI

- **Quatre bases de données tirées des archives UCI**
  - "Glass" : 214 objets, 9 attributs (2 classes),
  - "Hepatitis" : 80 objets, 19 attributs (2 classes),
  - "Dermatology" : 366 objets, 34 attributs (6 classes),
  - "Wine" : 178 objets, 13 attributs (3 classes).

- **Matrice de similarités**
  - Construite à l'aide d'un noyau Gaussien : $S_{ij} = \exp(\frac{-\|x_i - x_j\|^2}{2\sigma^2})$

    avec σ : paramètre d'échelle (ici, moyenne des variances des attributs).

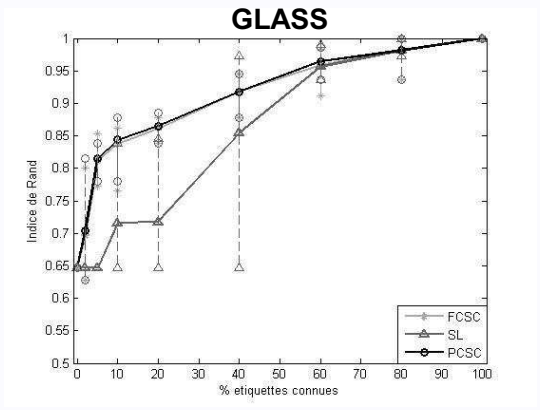
- **Optimisation des paramètres de pondération $\gamma$ et $\beta$ *a posteriori***
  - Discrétisation de l'intervalle de définition en 100 valeurs équidistantes,
  - Sélection de la valeur de $\gamma$ (ou $\beta$) maximisant le critère :

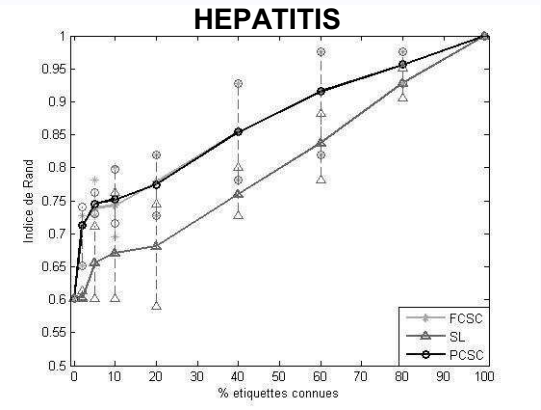
$$Crit = (1 - MNCut) + ML_{resp} + CL_{resp}$$

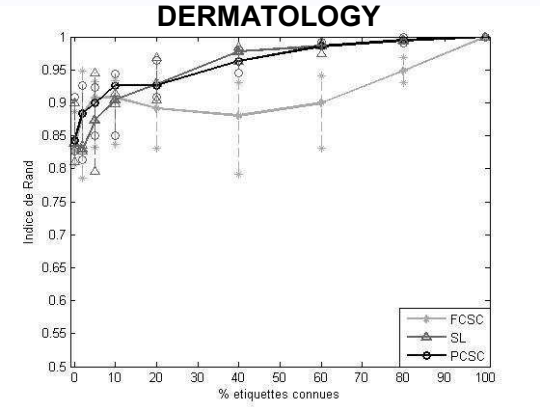
avec $ML_{resp}$ et $CL_{resp}$ : taux de respect des contraintes *ML* et *CL*.

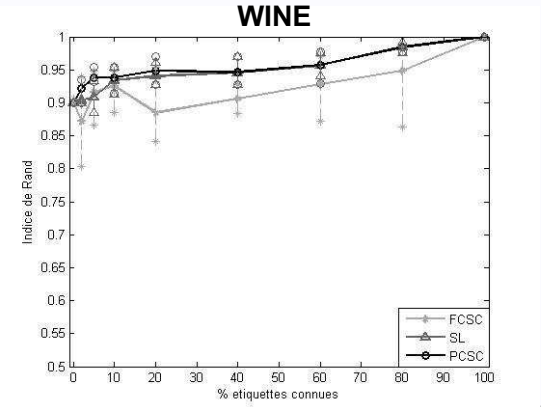


# Application : Bases de données UCI



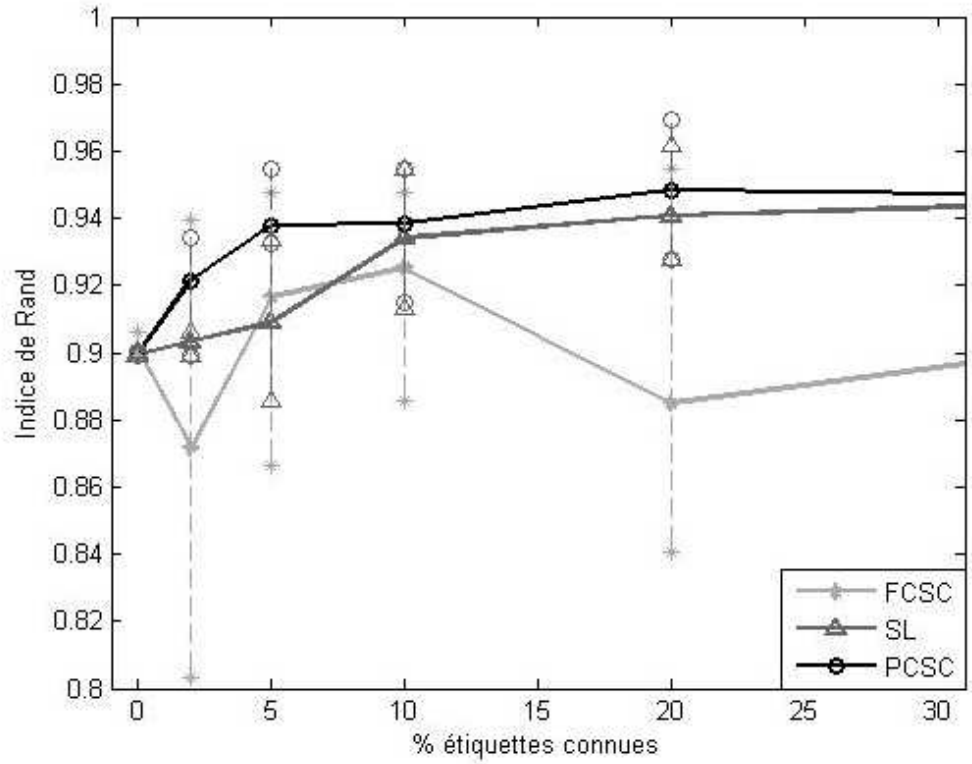



# Application : Bases de données UCI

**ZOOM SUR WINE**

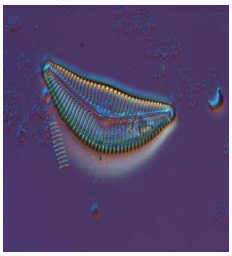




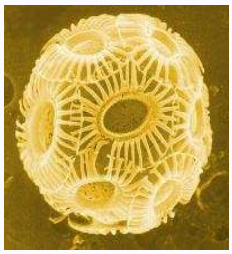

# Application : Identification des cellules phytoplanctoniques

- **Phytoplancton**
  - Plancton végétal microscopique qui erre au gré des courants,
  - Longueur : entre 1 µm et plusieurs mm,
  - Environ 6000 espèces au niveau mondial dont 70 toxiques ou nuisibles,
  - Responsable de 45% de la production primaire mondiale.

**DIATOMEE**

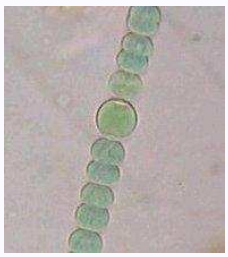


**COCCOLITHOPHORIDE**

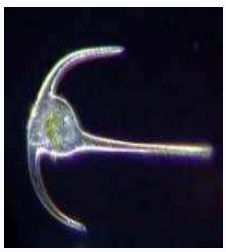


- **Enjeux**
  - Plan écologique et climatique :
    Préservation de l'environnement,
    Préservation de la biodiversité.
  - Plan économique :
    Tourisme, transport,
    Ressources exploitables, nourriture.

**CYANOBACTERIE**

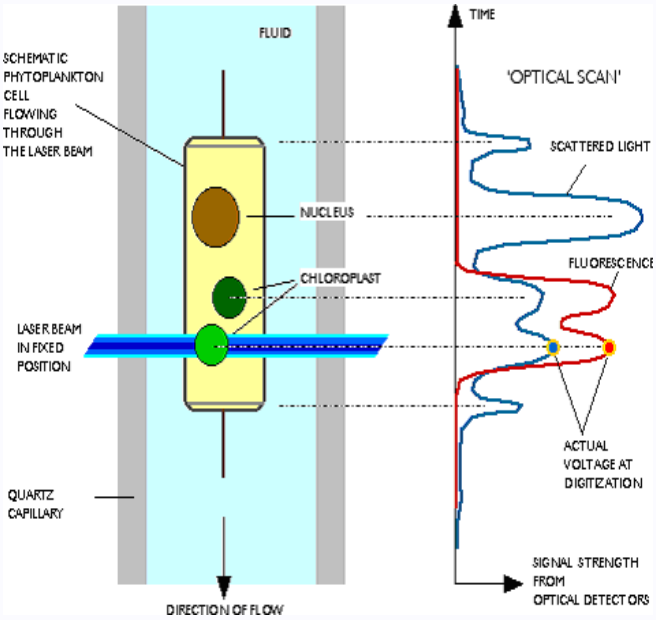


**DINOFLAGELLE**



# Application : Identification des cellules phytoplanctoniques

- **Cytomètrie en flux**




- Obtention de mesures de caractéristiques de certaines propriétés physiques et optiques des particules par un écoulement à grande vitesse devant un faisceau laser.

→ Signaux temporels

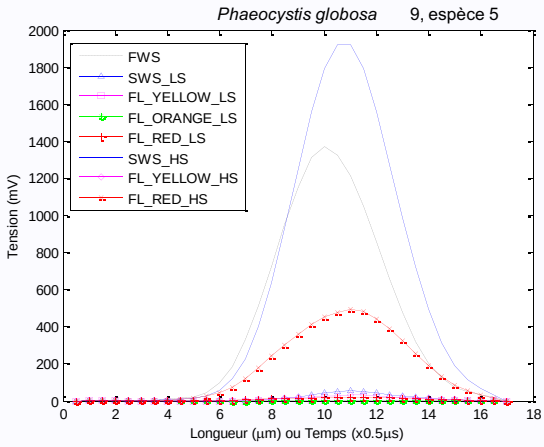
- Analyse de la diffusion et de la fluorescence des particules.

→ Classer la population suivant plusieurs critères.




# Application : Identification des cellules phytoplanctoniques


- **Courbes cytomètriques**
  - 8 signaux par cellule,
  - Conditions expérimentales identiques (fréquence d'échantillonnage, seuils de détection, etc.).




- 1 signal pour la diffusion à petits angles (FWS),
  → structure externe de la cellule;
- 2 signaux pour la diffusion à 90° (SWS),
  → structure interne de la cellule;
- 2 signaux pour la fluorescence rouge (FLR)
  → présence de pigments de chlorophylle;
- 1 signal pour la fluorescence orange (FLO),
  → âge, présence de pigments spécifiques;
- 2 signaux pour la fluorescence jaune (FLY),
  → présence de pigments spécifiques.

**PHOTO MICROSCOPIQUE**

# Application : Identification des cellules phytoplanctoniques

- **Base de données phytoplanctoniques**
  - 700 cellules étiquetées provenant d'échantillons de culture
  - 7 espèces représentées par 100 cellules chacune

    *Chaetoceros socialis*,
    *Emiliania huxleyi*,
    *Lauderia annulata*,
    *Leptocylindrus minimus*,
    *Phaeocystis globosa*,
    *Skeletonema costatum*
    *Thalassiosira rotula*.

**Objectifs**
- Rendre le processus de reconnaissance des espèces automatique.
- Retrouver la composition du prélèvement d'eau à partir des signaux bruts issus d'un Cytomètre en flux et en introduisant la connaissance de quelques paires de contraintes.

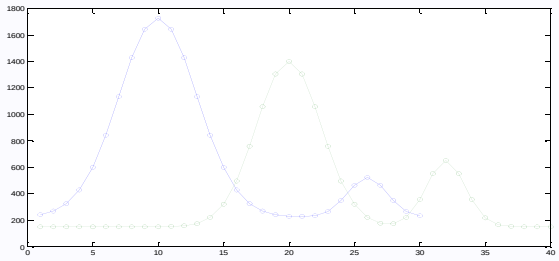


# Application : Identification des cellules phytoplanctoniques

- **Matrice de similarités**
  - Provient de [Caillault et al, 2009] (papier présenté à STIC et Environnement 2009),
  - Utilisation d'une mesure de type Dynamic Time Warping (DTW).

**Objectif**
- Rendre compte de la similarité de forme de deux ensembles de profils, en tolérant des déformations temporelles contrôlées.

**COURBES INITIALES**

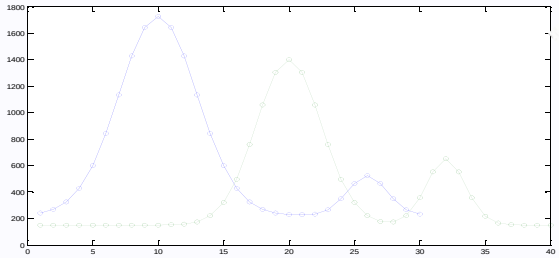


# Application : Identification des cellules phytoplanctoniques

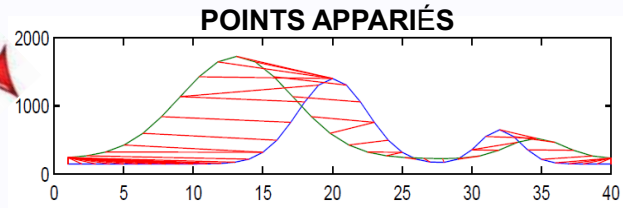
- **Matrice de similarités**
  - Provient de [Caillault et al, 2009] (papier présenté à STIC et Environnement 2009),
  - Utilisation d'une mesure de type Dynamic Time Warping (DTW).

**Objectif**
• Rendre compte de la similarité de forme de deux ensembles de profils, en tolérant des déformations temporelles contrôlées.

**COURBES INITIALES**



**POINTS APPARIÉS**



**COURBES OBTENUES**

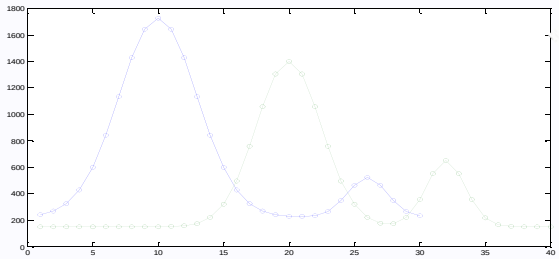


# Application : Identification des cellules phytoplanctoniques

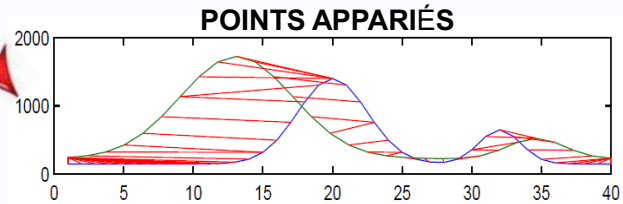
- **Matrice de similarités**
  - Provient de [Caillault et al, 2009] (papier présenté à STIC et Environnement 2009),
  - Utilisation d'une mesure de type Dynamic Time Warping (DTW).

**Objectif**
• Rendre compte de la similarité de forme de deux ensembles de profils, en tolérant des déformations temporelles contrôlées.

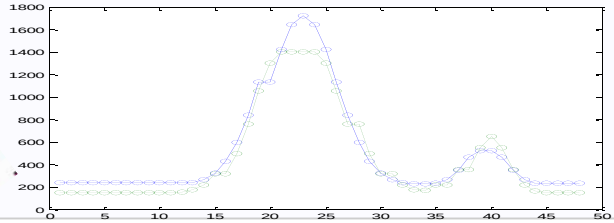
**COURBES INITIALES**



**POINTS APPARIÉS**



**COURBES OBTENUES**

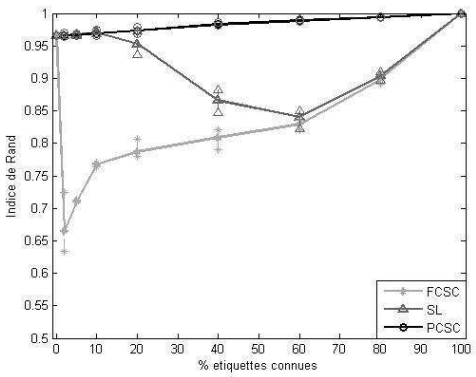


$$S(x_{ih}, y_{jh}) = \frac{d(x_{ih}, y_{jh})}{\max\{d(x_{ih}, 0), (y_{jh}, 0)\}}$$

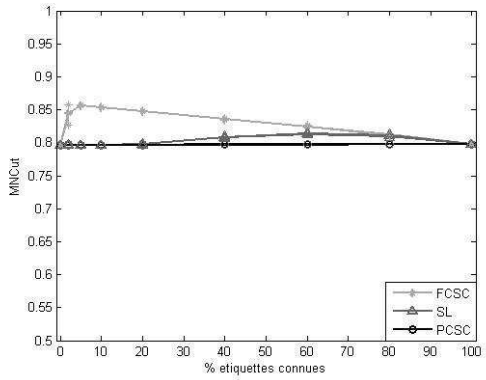
Sciences et Techniques de l'Information et de la Communication pour l'Environnement
École Nationale Supérieure des Mines de Saint-Etienne – 11-13 mai 2011

# Application : Identification des cellules phytoplanctoniques

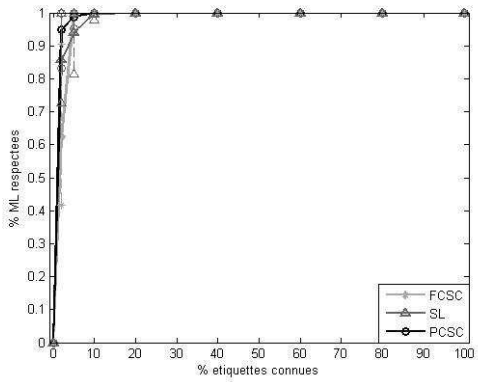
**INDICE DE RAND**



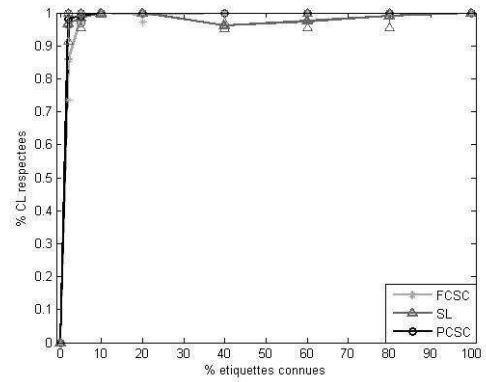

**MNCUT**




**% ML RESPECTEES**




**% CL RESPECTEES**






# Conclusions

- **Apports du papier**
  - Algorithme de classification spectrale intégrant des connaissances *a priori* sous forme de contraintes de comparaison entre paires d'objets (Must-Link et Cannot-Link).
  - Utilisation d'un paramètre jouant le rôle de balance entre la structure originale des données et l'impact des contraintes par paires définies.
  - Définition d'un critère d'optimisation explicite et simple de résolution (fournissant toujours une solution).

# Conclusions

- **Évaluation de la méthode proposée**

  – Bases de données UCI :
    - Performances supérieures aux autres méthodes (en terme d'indice de Rand et de taux de respect des contraintes)
    - Ajout d'une faible quantité de contraintes

      → Gain de performance considérable.

  – Identification de cellules phytoplanctoniques :
    - Performances supérieures ou identiques aux autres méthodes.
    - Gain moins important que pour les bases UCI.

      → Score déjà très élevé de la classification spectrale non contrainte.

      → Forte variabilité inter-espèces des profils de cellules phytoplanctoniques.

      → Paires sélectionnées aléatoirement => besoin d'un très grand nombre pour constituer une information utile.



# Perspectives

- **Apprentissage actif**
  – Soumission à l'expert, d'un ensemble de paires de profils judicieusement choisies.

  Objectif : amélioration considérable de la tâche d'étiquetage semi-automatique des cellules phytoplanctoniques.

- **Application à des données naturelles**
  – Application de l'algorithme de classification spectrale semi-supervisée proposée sur des données phytoplanctoniques provenant du milieu naturel.

  Objectif : montrer l'impact de la méthode proposée sur un ensemble d'échantillons marins naturels plus varié en nombre d'espèces et plus riche en information.

**Merci de votre attention.**

# Questions ?