Réduction de dimension pour l'exploration de données de grande dimension

Sylvain Lespinats

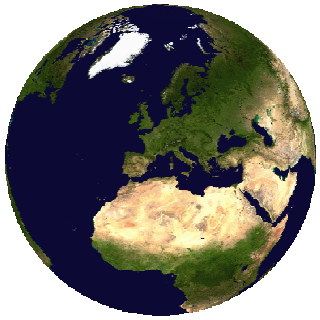
LIMA, CEA

Mail: sylvain.lespinats@cea.fr
Web: http://sy.lespi.free.fr/recherche/indexFR.html

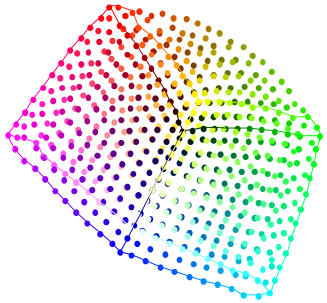
Séminaire APPRENTEO

Jeudi 29 octobre 2009

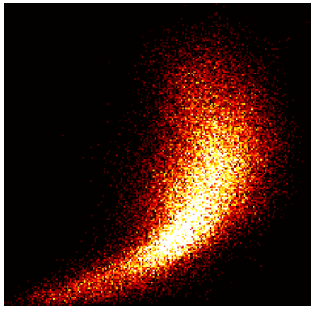
CEA, LIST, Multisensor Intelligence and Machine Learning Laboratory. F-91191 Gif-sur-Yvette, France.



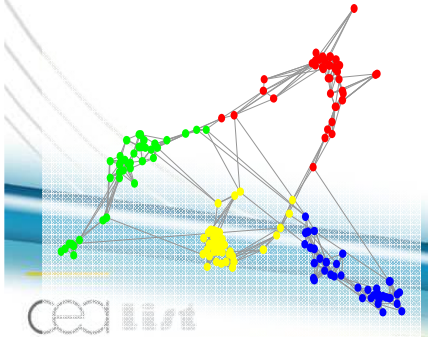
I. Géographie d'un jeu de données



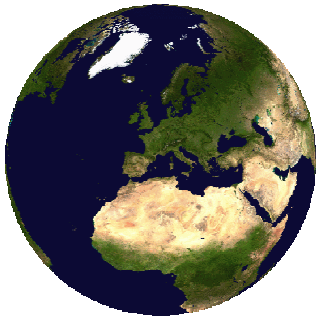
II. Réduction de dimension à partir des distances



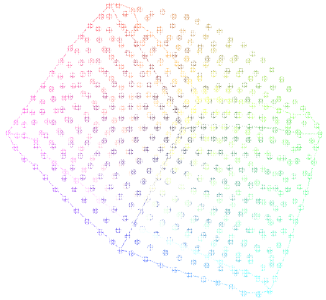
III. Evaluations des mappings



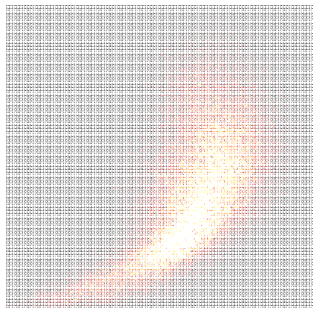
IV. Réduction de dimension à partir des rangs de voisinage



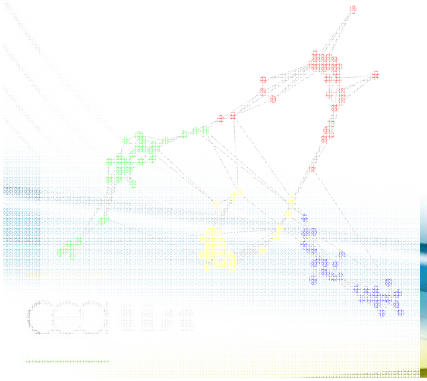
I. Géographie d'un jeu de données



II. Réduction de dimension à partir des distances

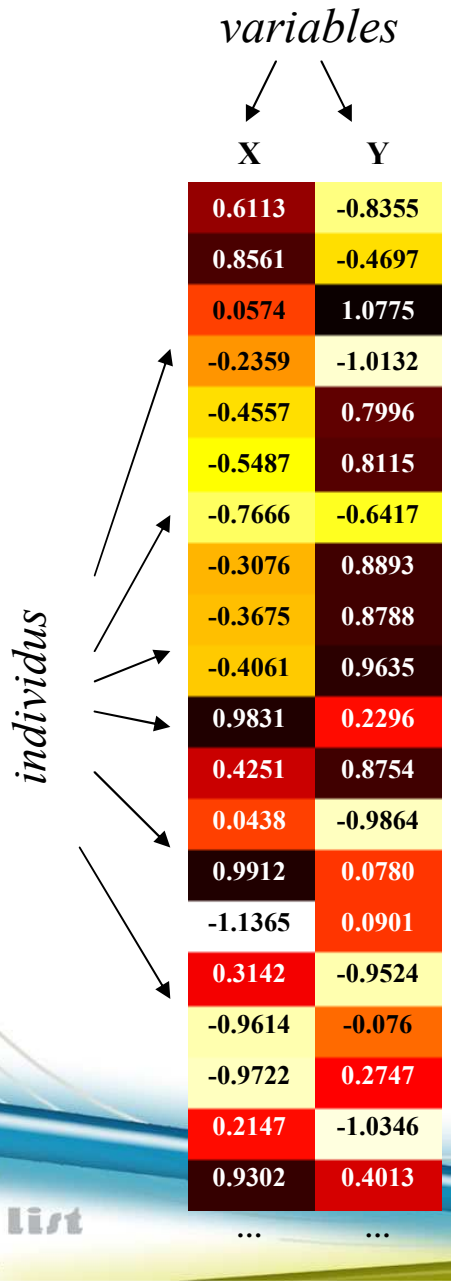


III. Evaluations des mappings



IV. Réduction de dimension à partir des rangs de voisinage

Organisation spatiale de ces données ?



2D

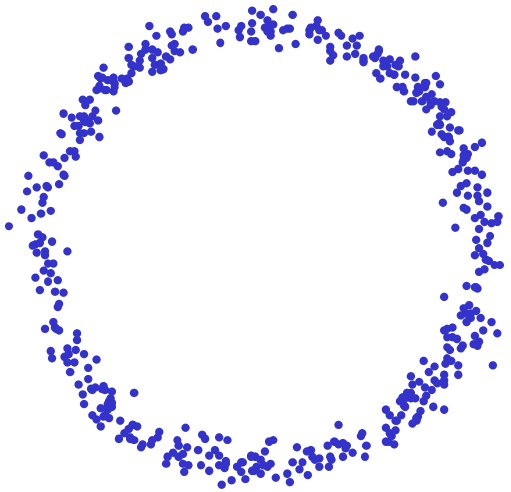
Organisation spatiale de ces données ?

variables

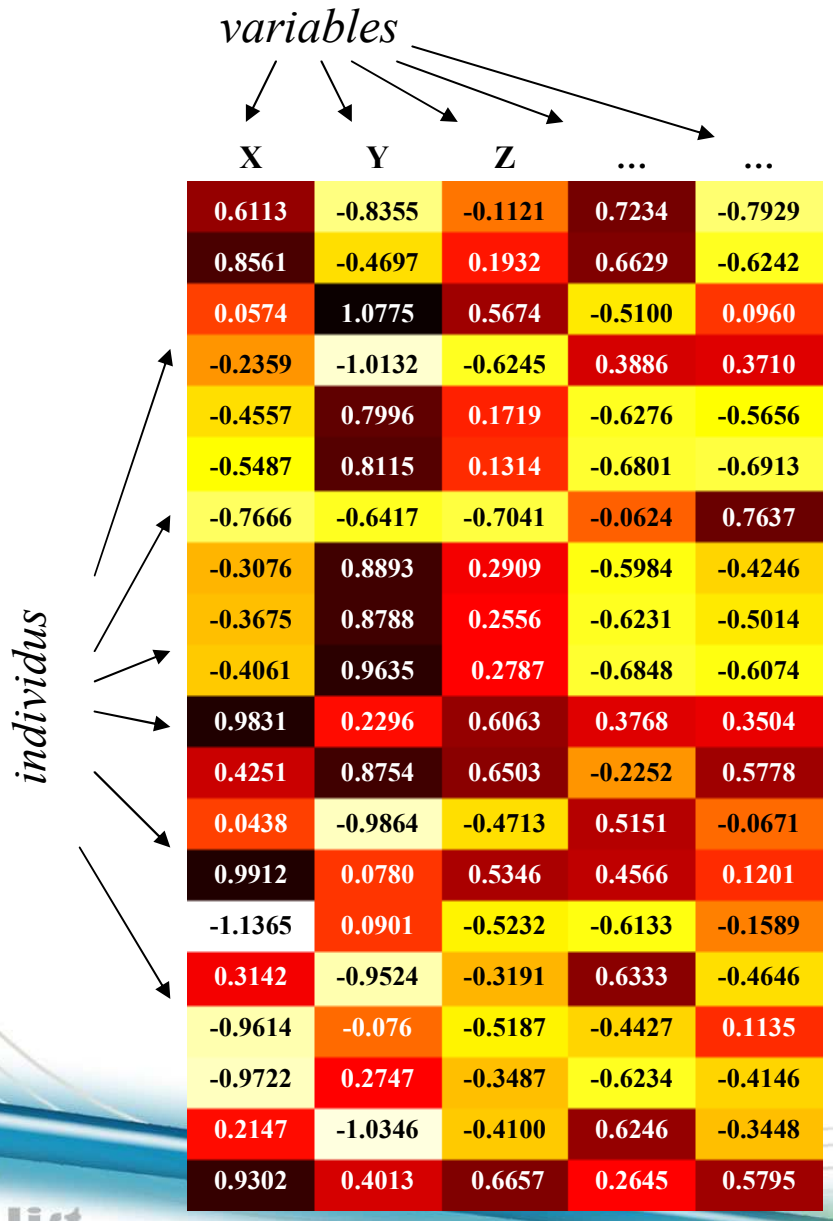
individus

X	Y
0.6113	-0.8355
0.8561	-0.4697
0.0574	1.0775
-0.2359	-1.0132
-0.4557	0.7996
-0.5487	0.8115
-0.7666	-0.6417
-0.3076	0.8893
-0.3675	0.8788
-0.4061	0.9635
0.9831	0.2296
0.4251	0.8754
0.0438	-0.9864
0.9912	0.0780
-1.1365	0.0901
0.3142	-0.9524
-0.9614	-0.076
-0.9722	0.2747
0.2147	-1.0346
0.9302	0.4013
…	…

2D



Organisation spatiale de ces données ?



5D

?

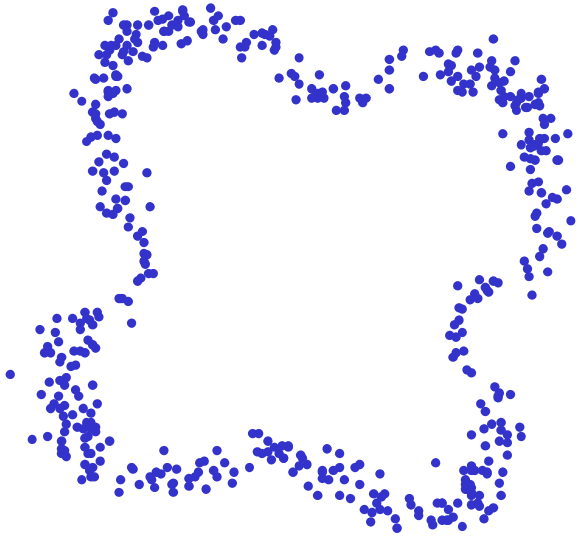
Organisation spatiale de ces données ?

variables

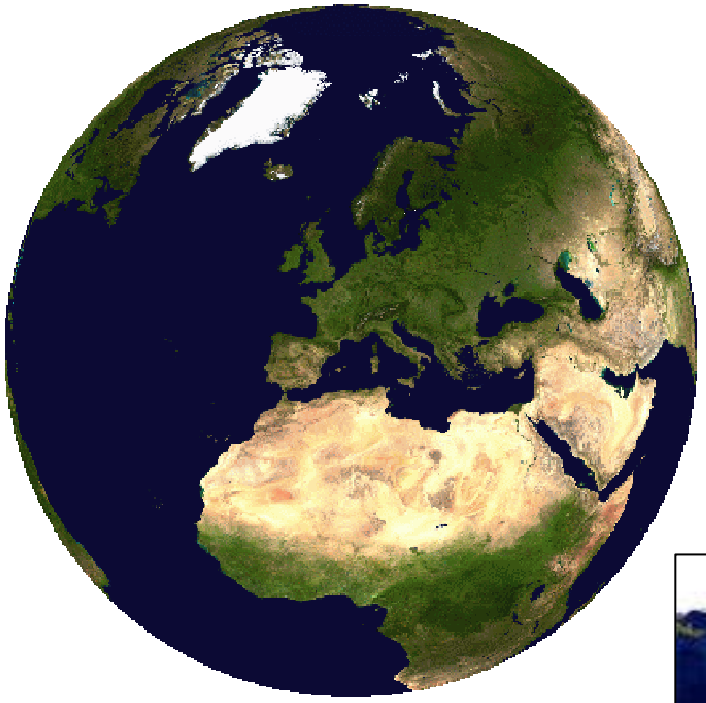
X	Y	Z	…	…
0.6113	-0.8355	-0.1121	0.7234	-0.7929
0.8561	-0.4697	0.1932	0.6629	-0.6242
0.0574	1.0775	0.5674	-0.5100	0.0960
-0.2359	-1.0132	-0.6245	0.3886	0.3710
-0.4557	0.7996	0.1719	-0.6276	-0.5656
-0.5487	0.8115	0.1314	-0.6801	-0.6913
-0.7666	-0.6417	-0.7041	-0.0624	0.7637
-0.3076	0.8893	0.2909	-0.5984	-0.4246
-0.3675	0.8788	0.2556	-0.6231	-0.5014
-0.4061	0.9635	0.2787	-0.6848	-0.6074
0.9831	0.2296	0.6063	0.3768	0.3504
0.4251	0.8754	0.6503	-0.2252	0.5778
0.0438	-0.9864	-0.4713	0.5151	-0.0671
0.9912	0.0780	0.5346	0.4566	0.1201
-1.1365	0.0901	-0.5232	-0.6133	-0.1589
0.3142	-0.9524	-0.3191	0.6333	-0.4646
-0.9614	-0.076	-0.5187	-0.4427	0.1135
-0.9722	0.2747	-0.3487	-0.6234	-0.4146
0.2147	-1.0346	-0.4100	0.6246	-0.3448
0.9302	0.4013	0.6657	0.2645	0.5795
…	…	…	…	…

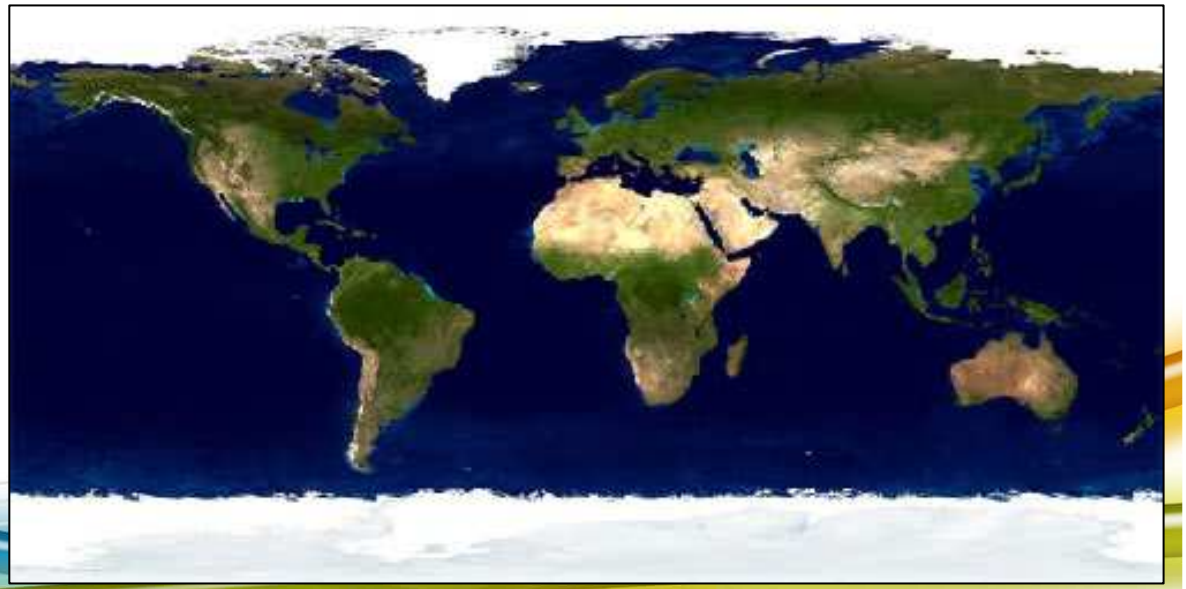
individus

5D

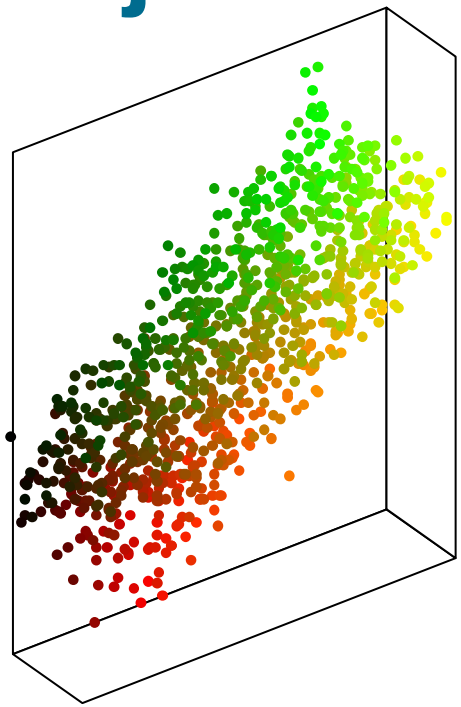


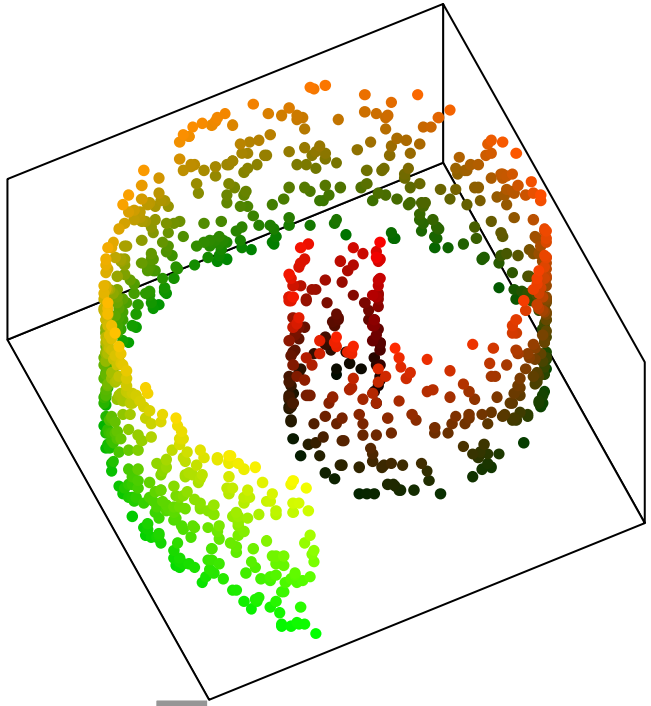
Une carte pour exprimer la "géographie" des données





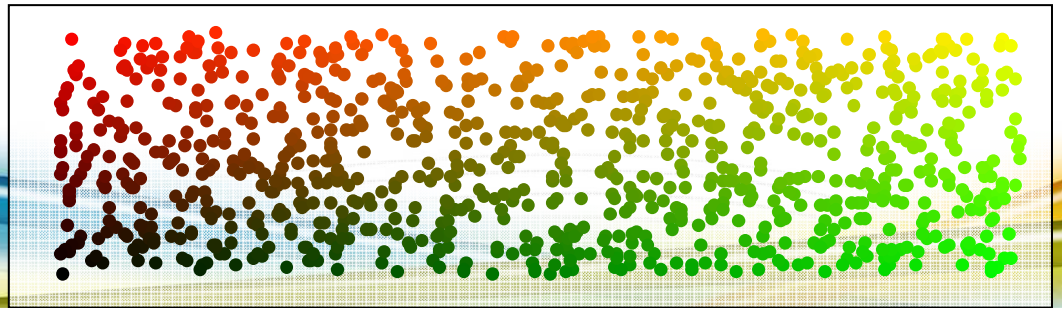
Projection linéaire ou non-linéaire





linéaire

non-linéaire



Multi Dimensional Scaling: usages

- **Visualisation des données de plus de 3 dimensions.**
- Visualisation de données issues d'espace non-euclidien.
- Plongement de données dans un espace vectoriel.
- Prétraitement (pour contourner le fléau de la dimension).

Multi Dimensional Scaling: usages

- Visualisation des données de plus de 3 dimensions.
- Plongement de données dans un espace vectoriel.

Digression: le fléau de la dimension

1) Désertification de l'espace
2) Décroissance du volume de la boule unité
3) Dépeuplement du centre des hyper-volumes
4) Concentration de la mesure

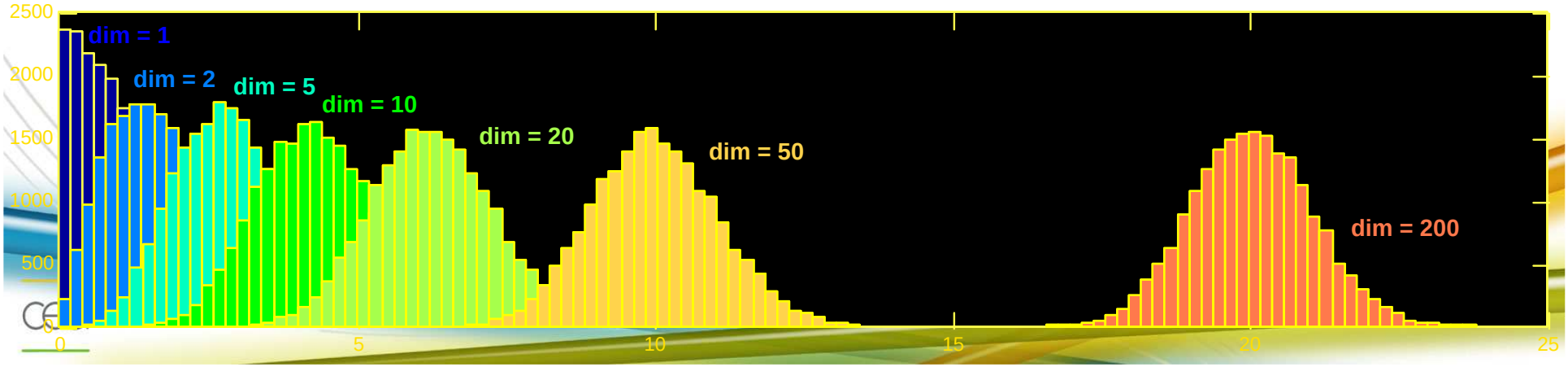
…

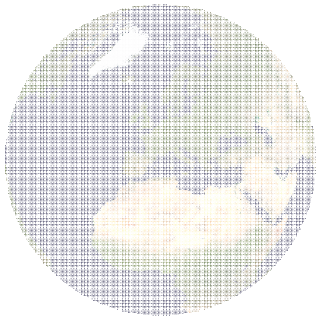
Multi Dimensional Scaling: usages

- Visualisation des données de plus de 3 dimensions.
- Visualisation de données abstraites.
- Plongement de données dans un espace vectoriel.
- Prétraitement (réduction de dimension).

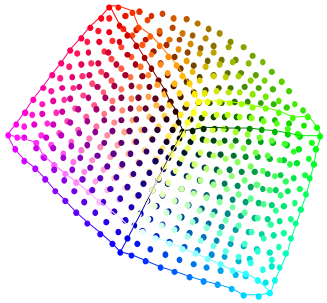
Digression: le fléau de la dimension

1) Désertification de l'espace
2) Décroissance du volume de la boule unité
3) Dépeuplement du centre des hyper-volumes
4) Concentration de la mesure

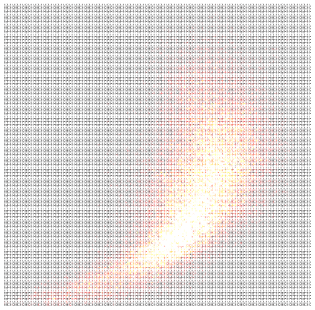




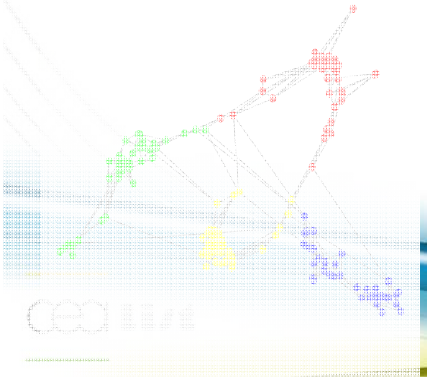
I. Géographie d'un jeu de données



II. Réduction de dimension à partir des distances



III. Evaluations des mappings

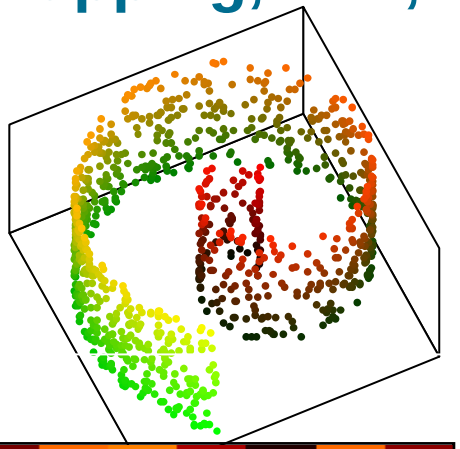


IV. Réduction de dimension à partir des rangs de voisinage

les MultiDimensional Scalings

ACP, ACC, Isomap, Sammon's mapping, LLE, …

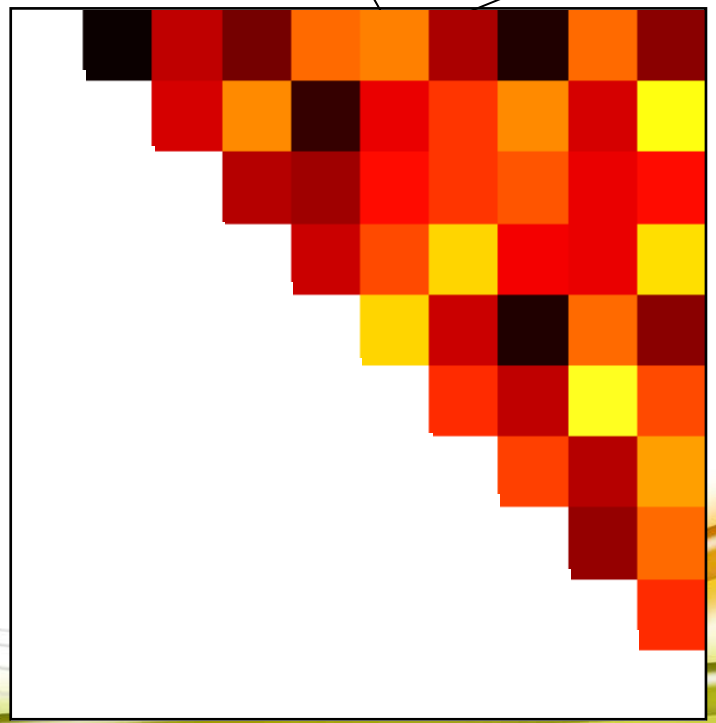
MDS : "Projection" (souvent non-linéaire) qui préserve les distances entre objets (en avantageant les distances courtes).



Une méthode possible :

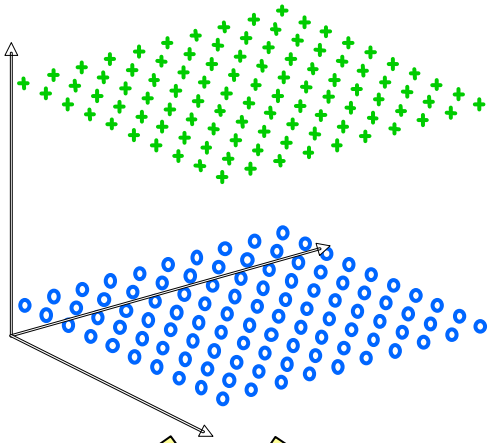
On connaît les distances entre objets.

On cherche une configuration de points dans l'espace de représentation qui préserve "au mieux" ces distances.



Défauts:
"faux-voisinages" et "déchirements"

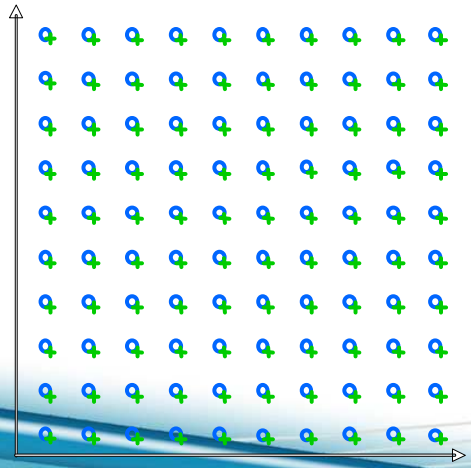
Données d'origine (3D)



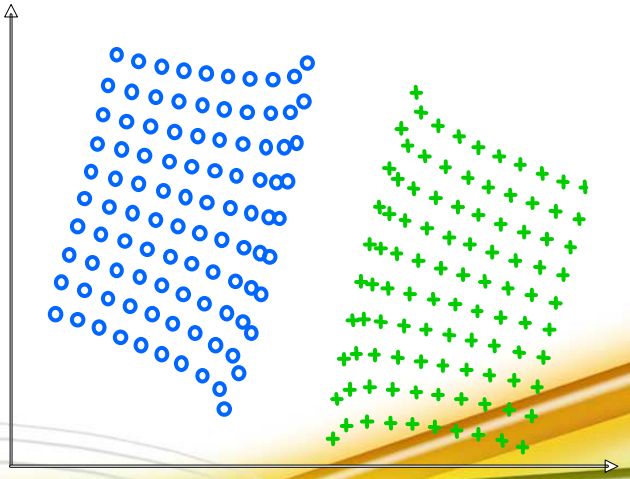
d_{ij} = *distance observée entre i et j*

δ_{ij} = *output distance dans la représentation*

mapping 2D (Analyse en composantes principales)

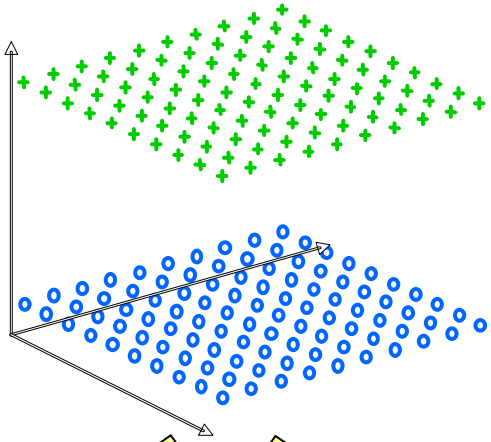


mapping 2D (Analyse en composantes curvilignes)



Défauts:
"faux-voisinages" et "déchirements"

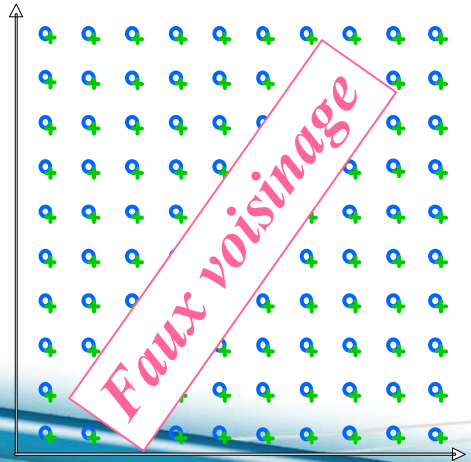
Données d'origine (3D)



d_{ij} = *distance observée entre i et j*

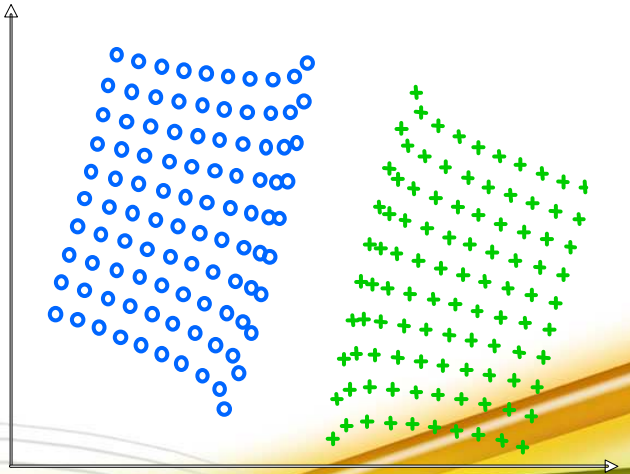
δ_{ij} = *output distance dans la représentation*

mapping 2D (Analyse en composantes principales)



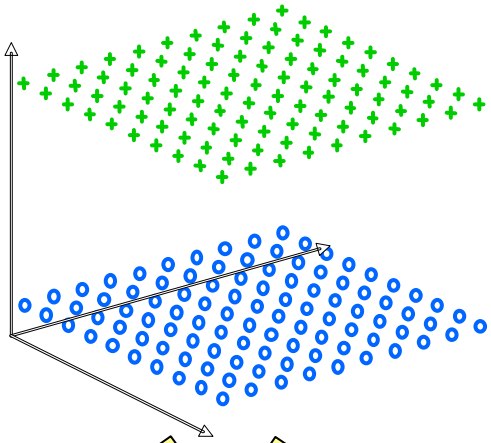
d_{ij} *grand et* δ_{ij} *petit*

mapping 2D (Analyse en composantes curvilignes)



Défauts:
"faux-voisinages" et "déchirements"

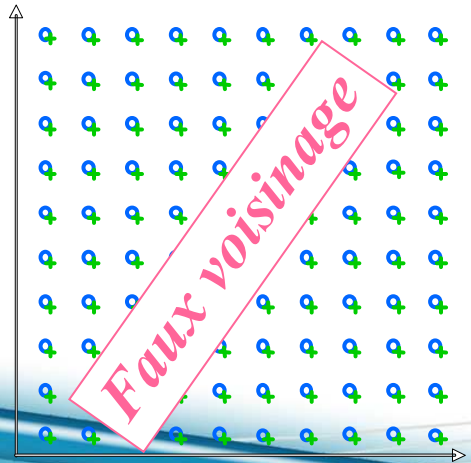
Données d'origine (3D)



d_{ij} = *distance observée entre i et j*

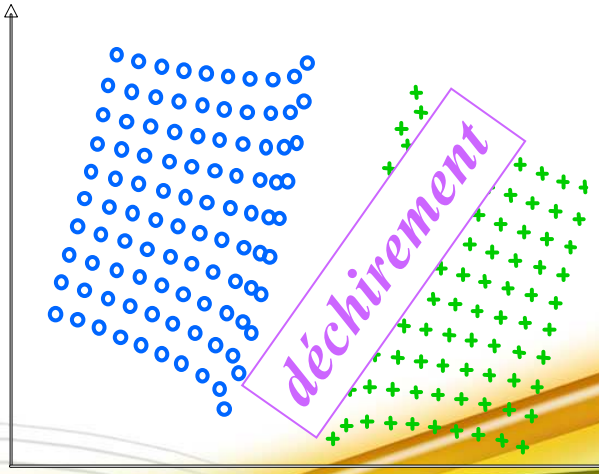
δ_{ij} = *output distance dans la représentation*

mapping 2D (Analyse en composantes principales)



d_{ij} *grand et* δ_{ij} *petit*

mapping 2D (Analyse en composantes curvilignes)



d_{ij} *petit et* δ_{ij} *grand*

MDS: Fonctions de cout

$$\zeta = \sum_{i,j} \left(d_{ij} - \delta_{ij} \right)^2 \times f\left(d_{ij} \quad ou \quad \delta_{ij} \right)$$

Pour que δ_{ij} *se rapproche de* d_{ij}.

Poids pour donner l'avantage à la préservation des voisinages

$\Rightarrow$ *f(x)* est grand si *x* est petit.

MDS: Fonctions de cout

$$\zeta = \sum_{i,j} \left(d_{ij} - \delta_{ij} \right)^2 \times f\left(d_{ij} \right)$$

(Sammon's mapping)

J.W. Sammon, 1969.

MDS: Fonctions de cout

$$\zeta = \sum_{i,j} \left(d_{ij} - \delta_{ij} \right)^2 \times f\left(d_{ij} \right)$$

(Sammon's mapping)

Hypothèse 1, déchirement : d_{ij} *petit et* δ_{ij} *grand.*

MDS: Fonctions de cout

$$\zeta = \sum_{i,j} \left(d_{ij} - \delta_{ij}\right)^2 \times f\left(d_{ij}\right)$$

(Sammon's mapping)

Hypothèse 1, déchirement : d_{ij} *petit et* δ_{ij} *grand.*

$(d_{ij} - \delta_{ij})^2$ *grand* *et* $f(d_{ij})$ *grand*

$\Rightarrow$ *Fortement pénalisé.*

MDS: Fonctions de cout

$$\zeta = \sum_{i,j} \left(d_{ij} - \delta_{ij}\right)^2 \times f\left(d_{ij}\right)$$

(Sammon's mapping)

Hypothèse 1, déchirement : d_{ij} *petit et* δ_{ij} *grand.*

$(d_{ij} - \delta_{ij})^2$ *grand et* $f(d_{ij})$ *grand*

$\Rightarrow$ *Fortement pénalisé.*

Hypothèse 2, faux voisinage : d_{ij} *grand et* δ_{ij} *petit.*

$(d_{ij} - \delta_{ij})^2$ *grand et* $f(d_{ij})$ *petit*

$\Rightarrow$ *Faiblement pénalisé.*

MDS: Fonctions de cout

$$\zeta = \sum_{i,j} \left(d_{ij} - \delta_{ij} \right)^2 \times f\left(\delta_{ij} \right)$$

(analyse en composantes curviligne)

P. Demartines P. et J. Hérault , 1997.

Faux voisinage $\Rightarrow$ *Fortement pénalisé.*

Déchirement $\Rightarrow$ *Faiblement pénalisé.*

DD-HDS: Data-Driven High Dimensional Scaling

Originalités:

1) Elle pénalise les 2 types d'erreurs ("faux voisinages" ET "déchirements").

2) Elle est adaptée aux caractéristiques des données de grande dimension (concentration de la mesure).

DD-HDS: Data-Driven High Dimensional Scaling

$$\zeta = \sum_{i,j} \left(d_{ij} - \delta_{ij} \right)^2 \times f\left(d_{ij} \quad ou \quad \delta_{ij} \right)$$

1) Pénalisation des "faux voisinages" OU des "déchirements".

DD-HDS: Data-Driven High Dimensional Scaling

$$\zeta = \sum_{i,j} \left(d_{ij} - \delta_{ij}\right)^2 \times f\left(d_{ij} \quad ou \quad \delta_{ij}\right)$$

1) Pénalisation des "faux voisinages" ET des "déchirements".

$$f\left(\min\left(d_{ij}, \delta_{ij}\right)\right)$$

$\Rightarrow$ On relâche les contraintes si les distances longues dans l'espace d'origine ET dans l'espace de représentation

DD-HDS: Data-Driven High Dimensional Scaling

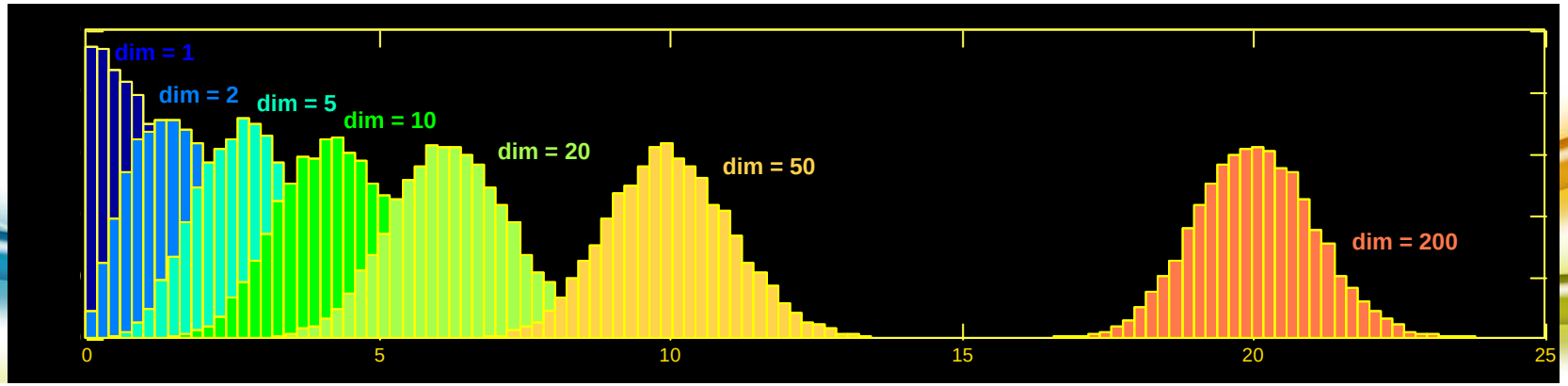
$$\zeta = \sum_{i,j} \left(d_{ij} - \delta_{ij} \right)^2 \times f\left(d_{ij} \quad ou \quad \delta_{ij} \right)$$

1) Pénalisation des "faux voisinages" ET des "déchirements".

$$f\left(\min\left(d_{ij}, \delta_{ij} \right) \right)$$

$\Rightarrow$ On relâche les contraintes si les distances longues dans l'espace d'origine ET dans l'espace de représentation

2) Risques dus au fléau de la dimension.



DD-HDS: Data-Driven High Dimensional Scaling

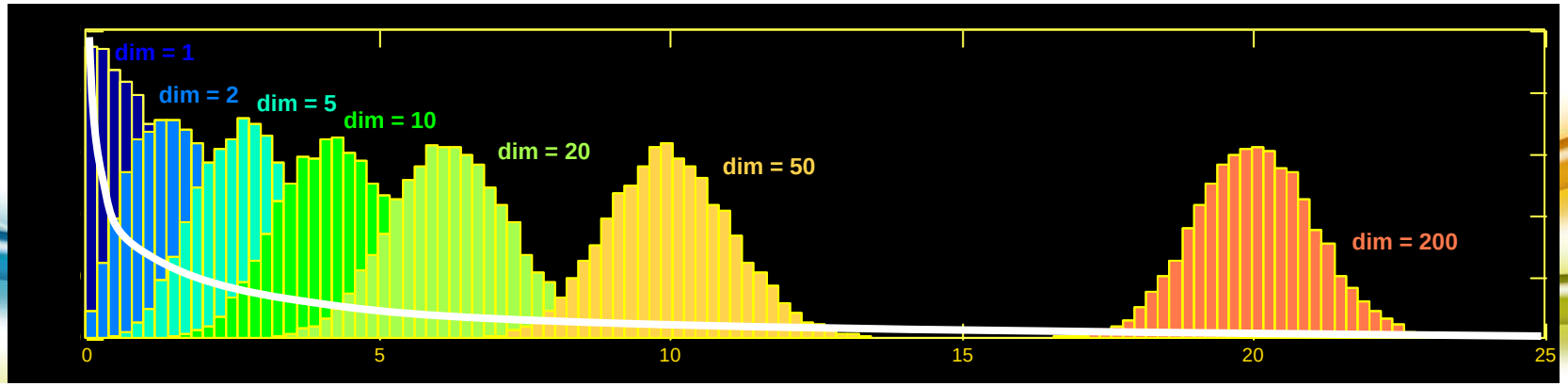
$$\zeta = \sum_{i,j} \left(d_{ij} - \delta_{ij} \right)^2 \times f\left(d_{ij} \quad ou \quad \delta_{ij} \right)$$

1) Pénalisation des "faux voisinages" ET des "déchirements".

$$f\left(\min\left(d_{ij}, \delta_{ij} \right) \right)$$

$\Rightarrow$ On relâche les contraintes si les distances longues dans l'espace d'origine ET dans l'espace de représentation

2) Risques dus au fléau de la dimension.



DD-HDS: Data-Driven High Dimensional Scaling

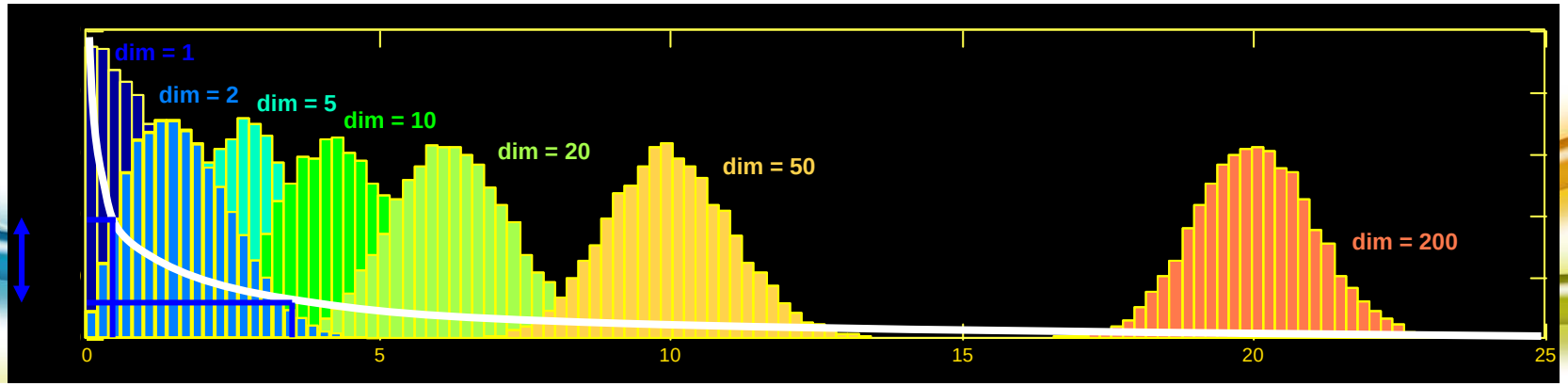
$$\zeta = \sum_{i,j} \left(d_{ij} - \delta_{ij} \right)^2 \times f\left(d_{ij} \quad ou \quad \delta_{ij} \right)$$

1) Pénalisation des "faux voisinages" ET des "déchirements".

$$f\left(\min\left(d_{ij}, \delta_{ij} \right) \right)$$

$\Rightarrow$ On relâche les contraintes si les distances longues dans l'espace d'origine ET dans l'espace de représentation

2) Risques dus au fléau de la dimension.



DD-HDS: Data-Driven High Dimensional Scaling

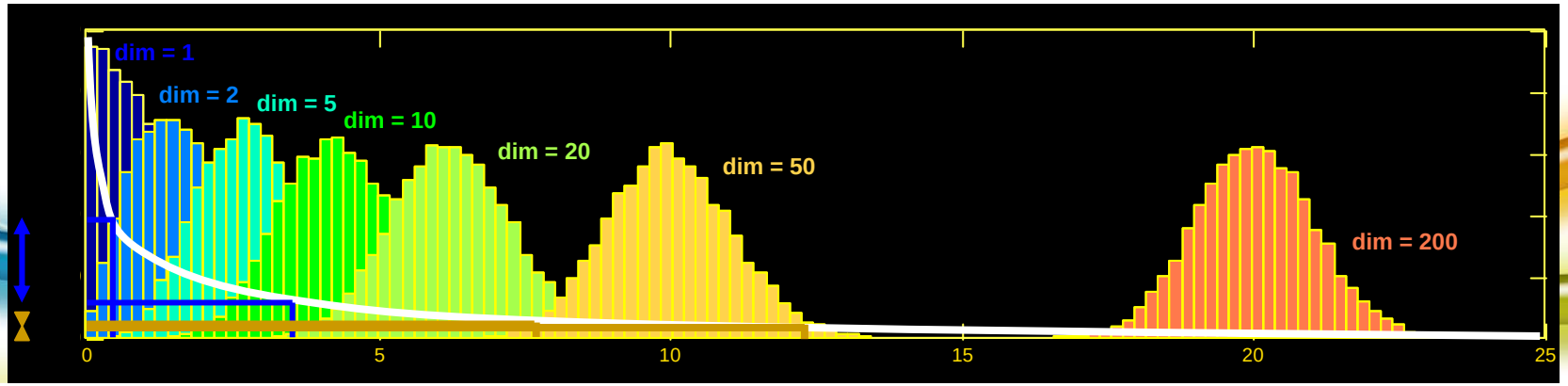
$$\zeta = \sum_{i,j} \left(d_{ij} - \delta_{ij} \right)^2 \times f\left(d_{ij} \quad ou \quad \delta_{ij} \right)$$

1) Pénalisation des "faux voisinages" ET des "déchirements".

$$f\left(\min\left(d_{ij}, \delta_{ij} \right) \right)$$

$\Rightarrow$ On relâche les contraintes si les distances longues dans l'espace d'origine ET dans l'espace de représentation

2) Risques dus au fléau de la dimension.



DD-HDS: Data-Driven High Dimensional Scaling

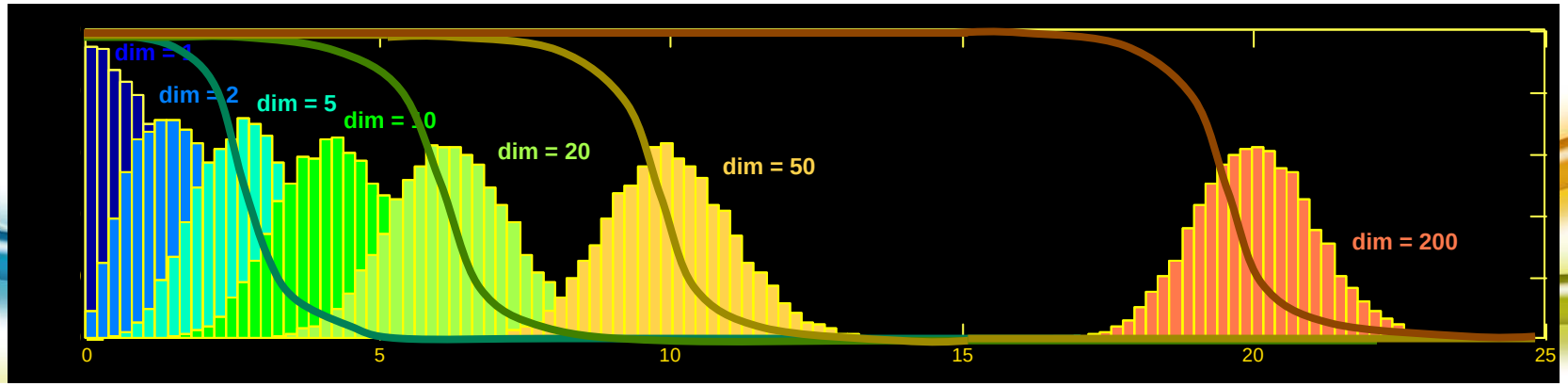
$$\zeta = \sum_{i,j} \left(d_{ij} - \delta_{ij} \right)^2 \times f\left(d_{ij} \quad ou \quad \delta_{ij} \right)$$

1) Pénalisation des "faux voisinages" ET des "déchirements".

$$f\left(\min\left(d_{ij}, \delta_{ij} \right) \right)$$

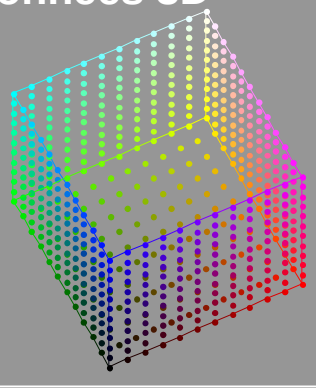
$\Rightarrow$ On relâche les contraintes si les distances longues dans l'espace d'origine ET dans l'espace de représentation

2) Risques dus au fléau de la dimension.



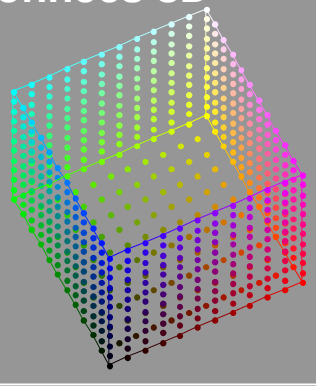
Exemple de projection : le cube ouvert

Données 3D

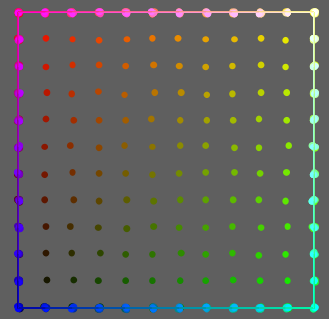


Exemple de projection : le cube ouvert

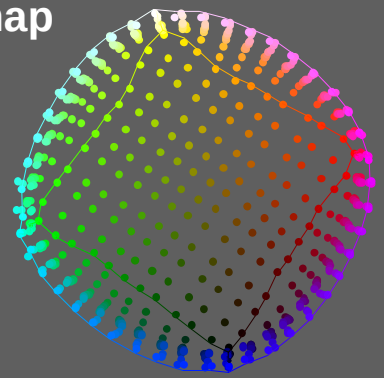
Données 3D



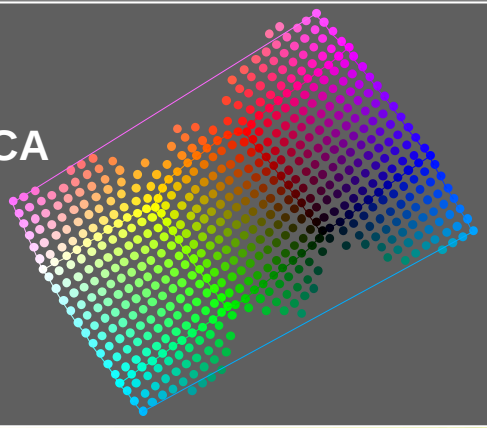
ACP



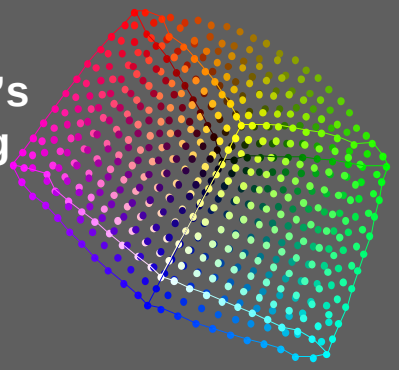
Isomap



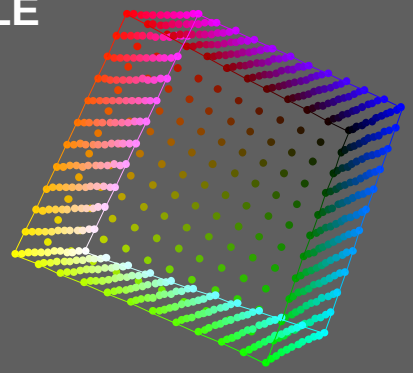
CCA



Sammon's mapping

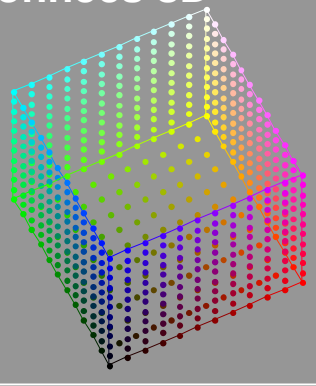


LLE

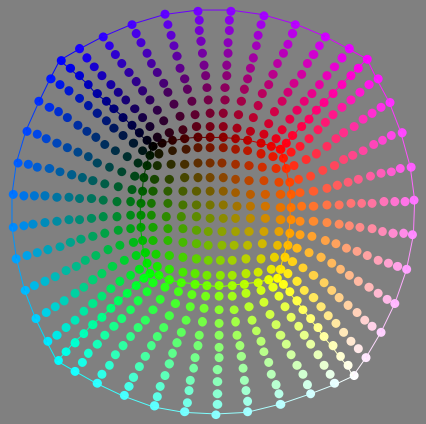


Exemple de projection : le cube ouvert

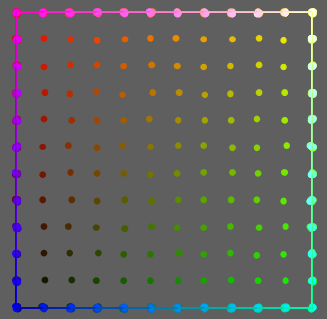
Données 3D



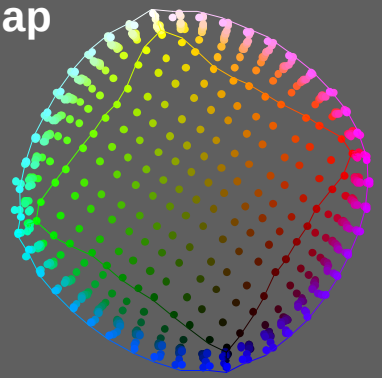
DD-HDS



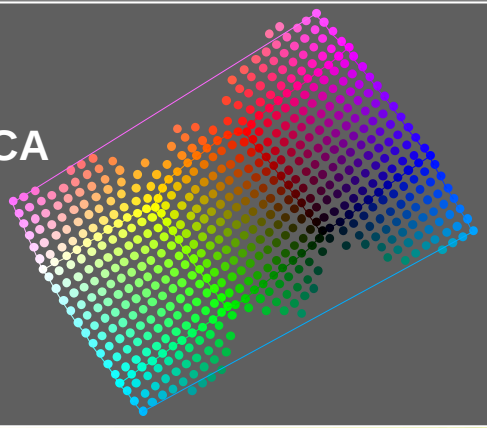
ACP



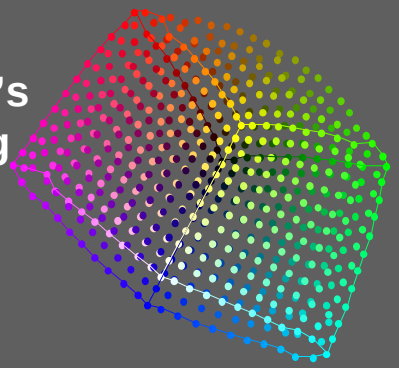
Isomap



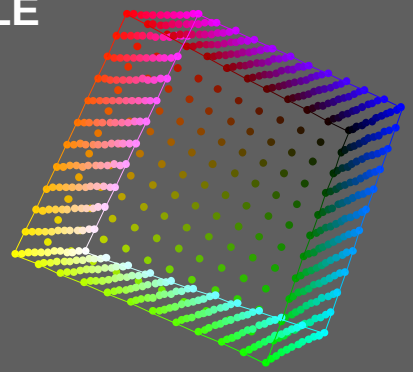
CCA

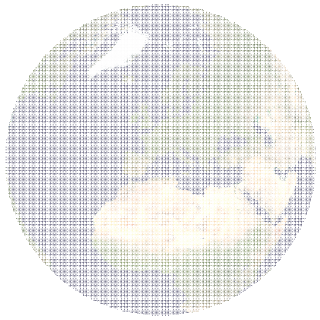


Sammon's mapping

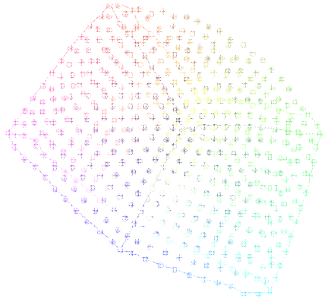


LLE

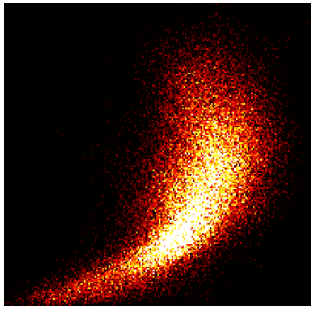




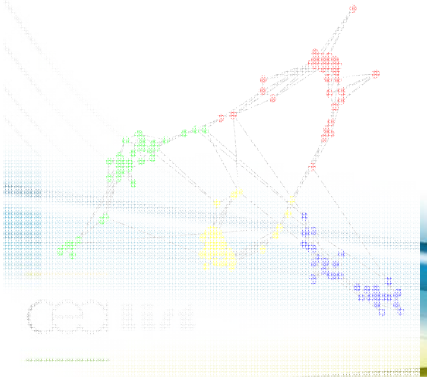
I. Géographie d'un jeu de données



II. Réduction de dimension à partir des distances

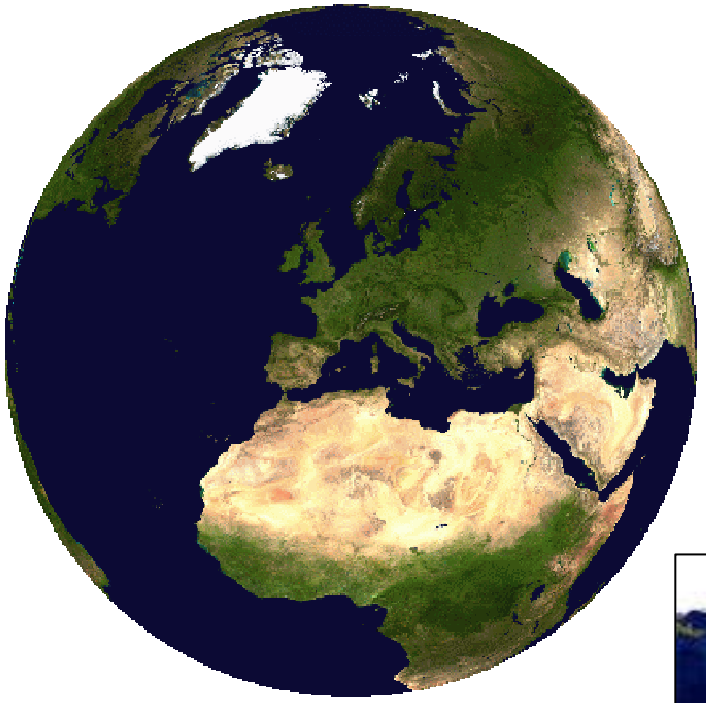


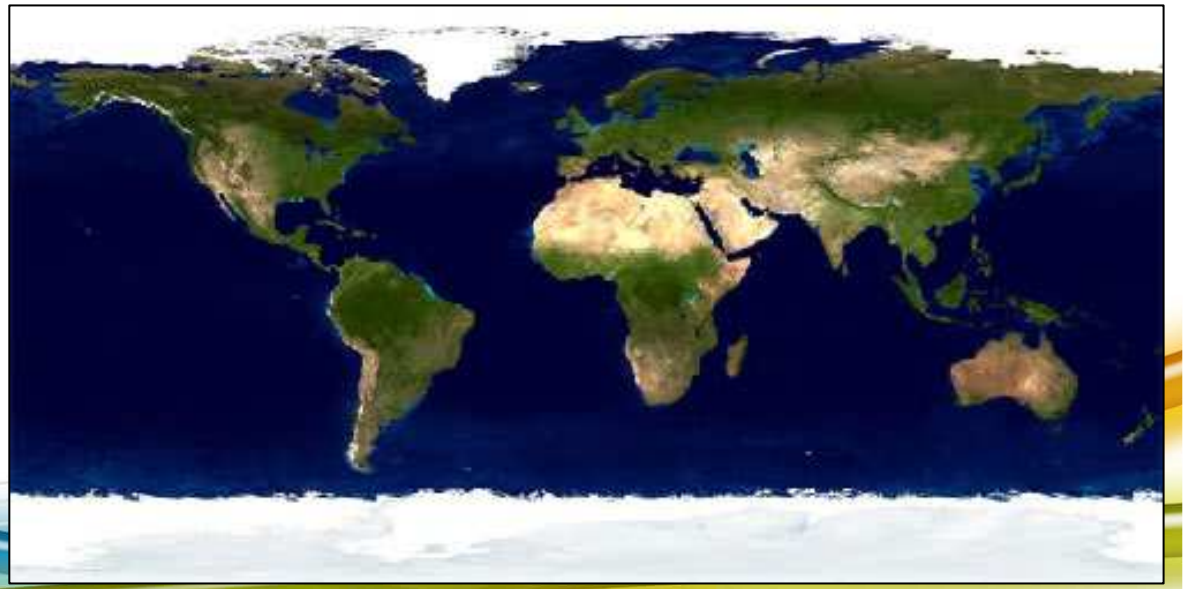
III. Evaluations des mappings



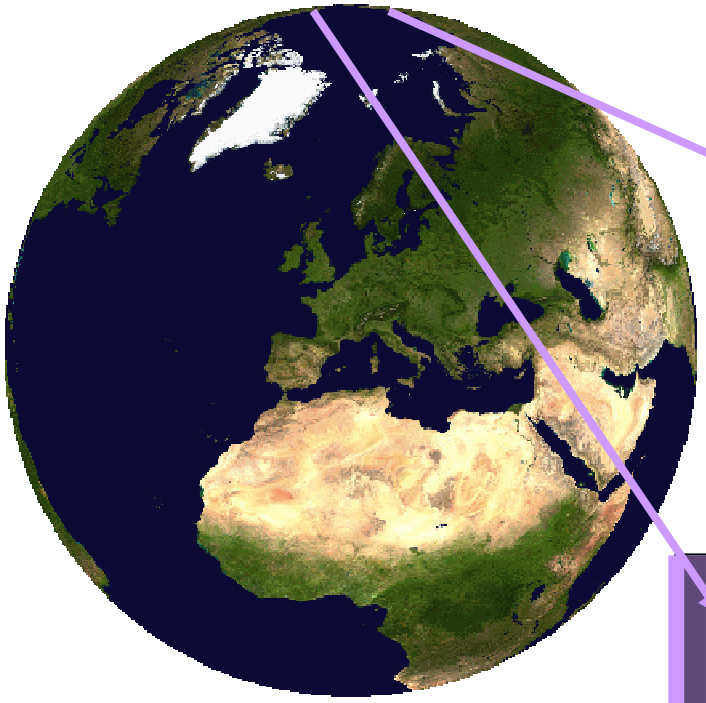
IV. Réduction de dimension à partir des rangs de voisinage

Les défauts d’un planisphère

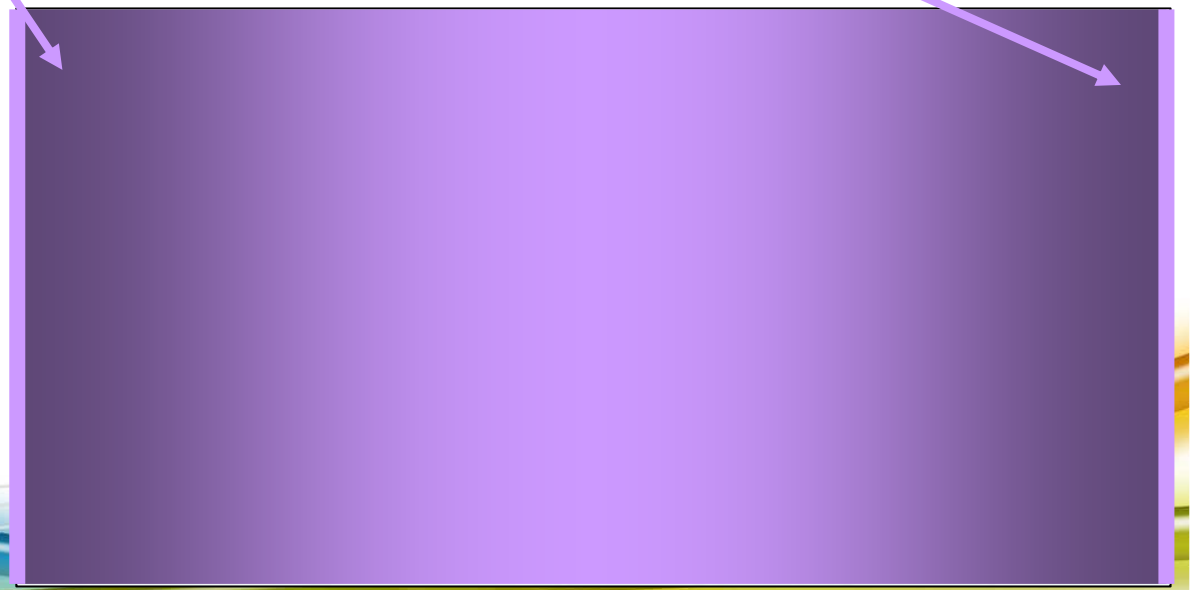




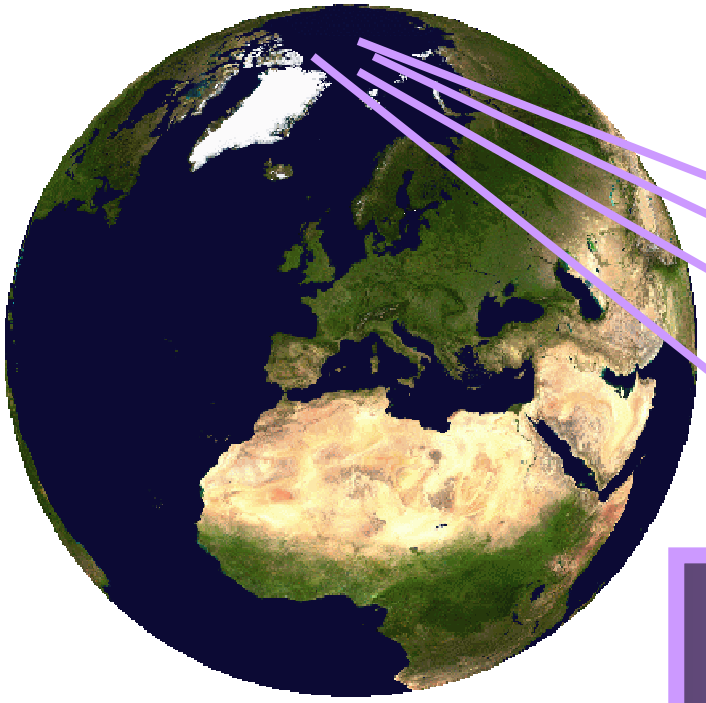
Les défauts d'un planisphère



déchirements



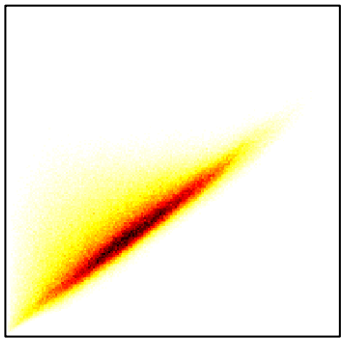
Les défauts d’un planisphère



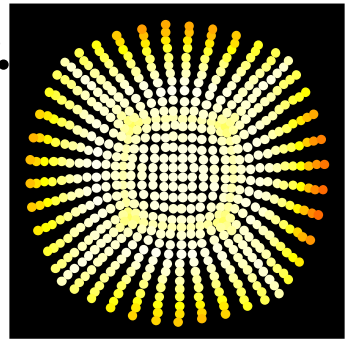
déchirements

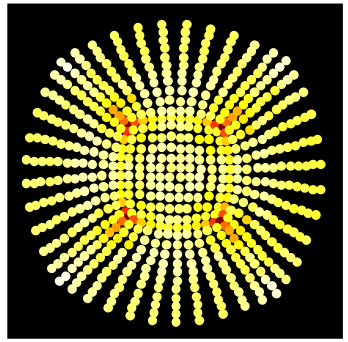
Analyse des défauts: une stratégie en 3 étapes

1) *Analyse globale.*

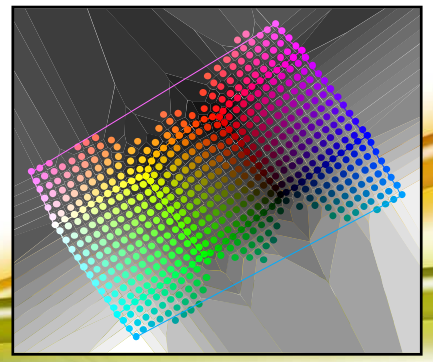


2) *Evaluation locale du niveau d'erreur.*

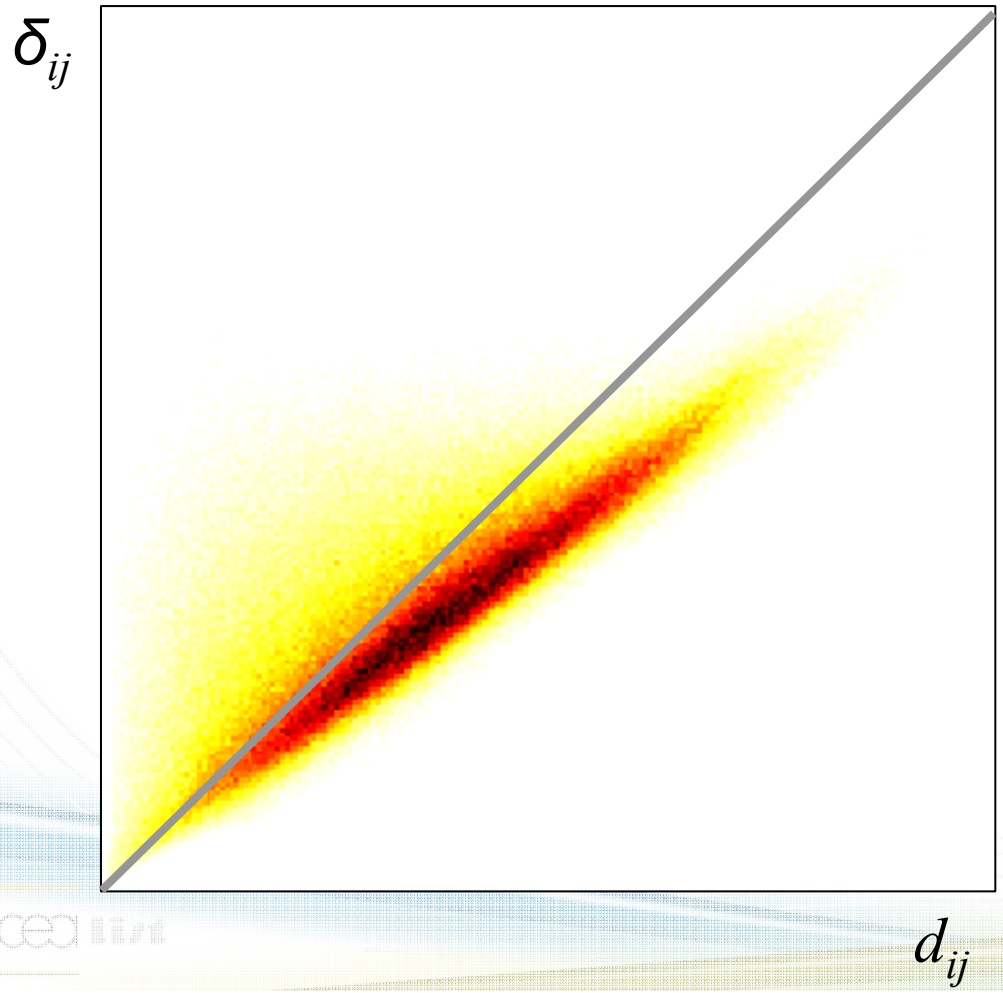




3) *Etudes des proximités selon un point de vue.*

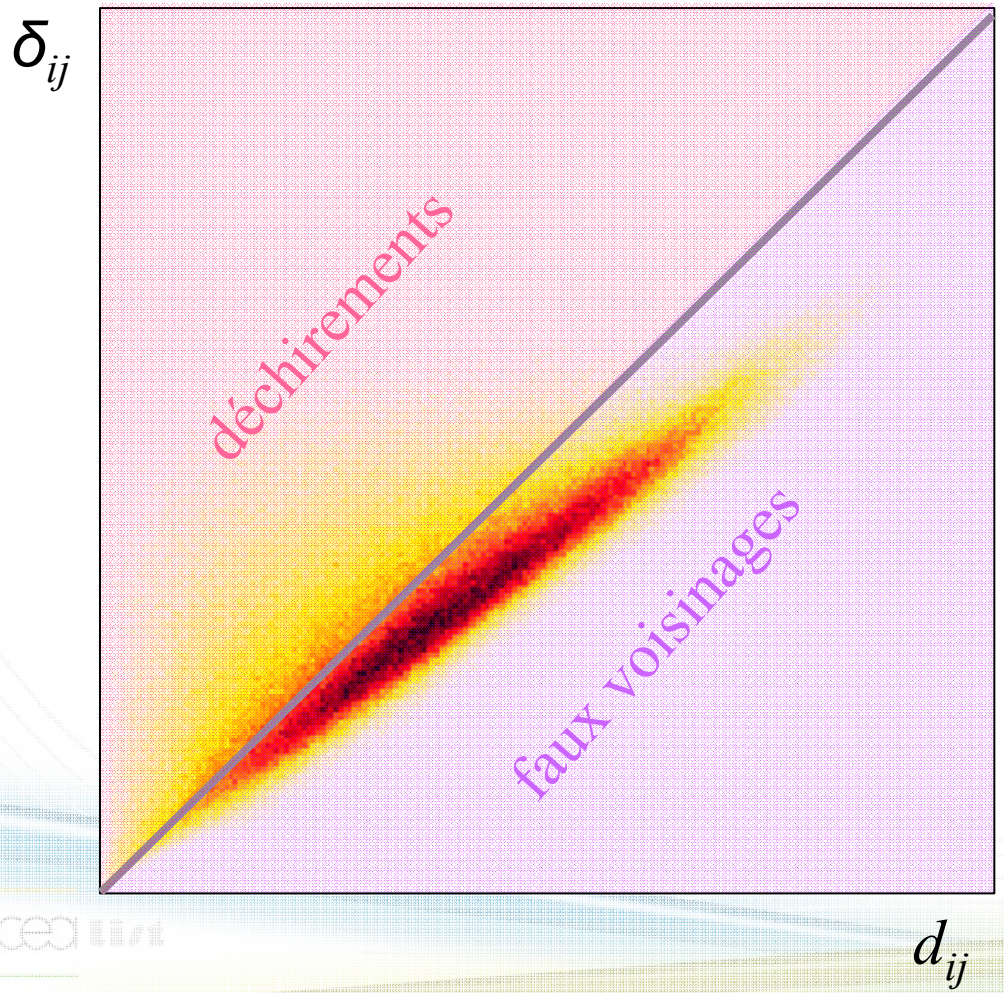


Y a-t-il des défaut dans une carte: le diagramme de Shepard



Permet de visualiser la qualité d'une représentation

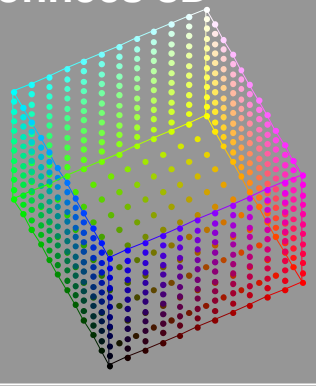
Y a-t-il des défaut dans une carte: le diagramme de Shepard



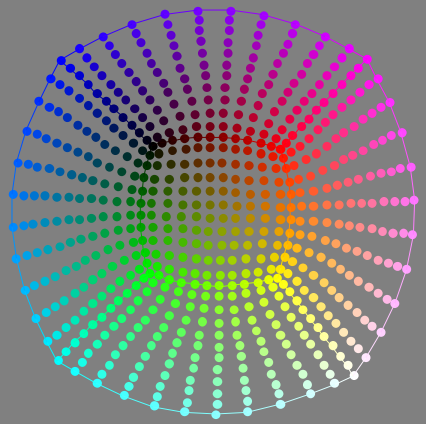
Permet de visualiser la qualité d'une représentation

Exemple de projection : le cube ouvert

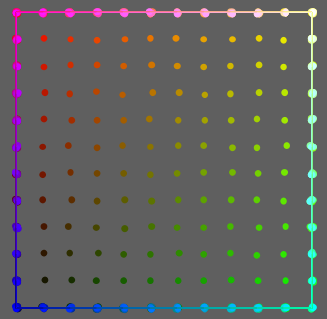
Données 3D



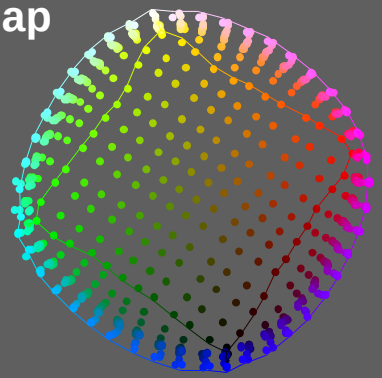
DD-HDS



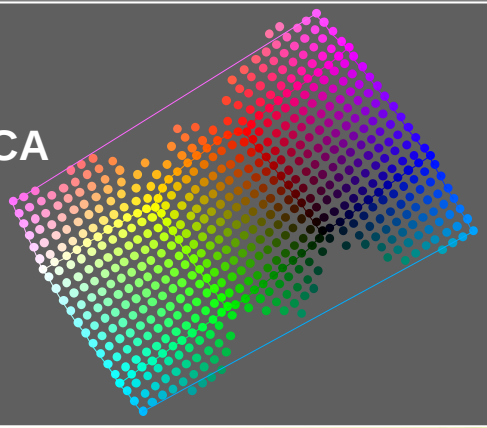
ACP



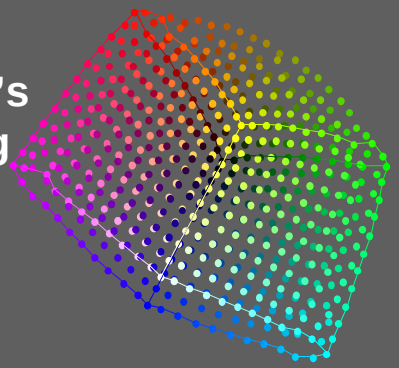
Isomap



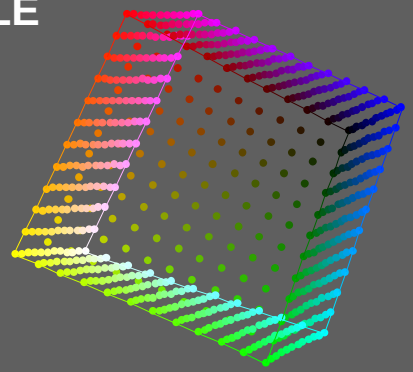
CCA



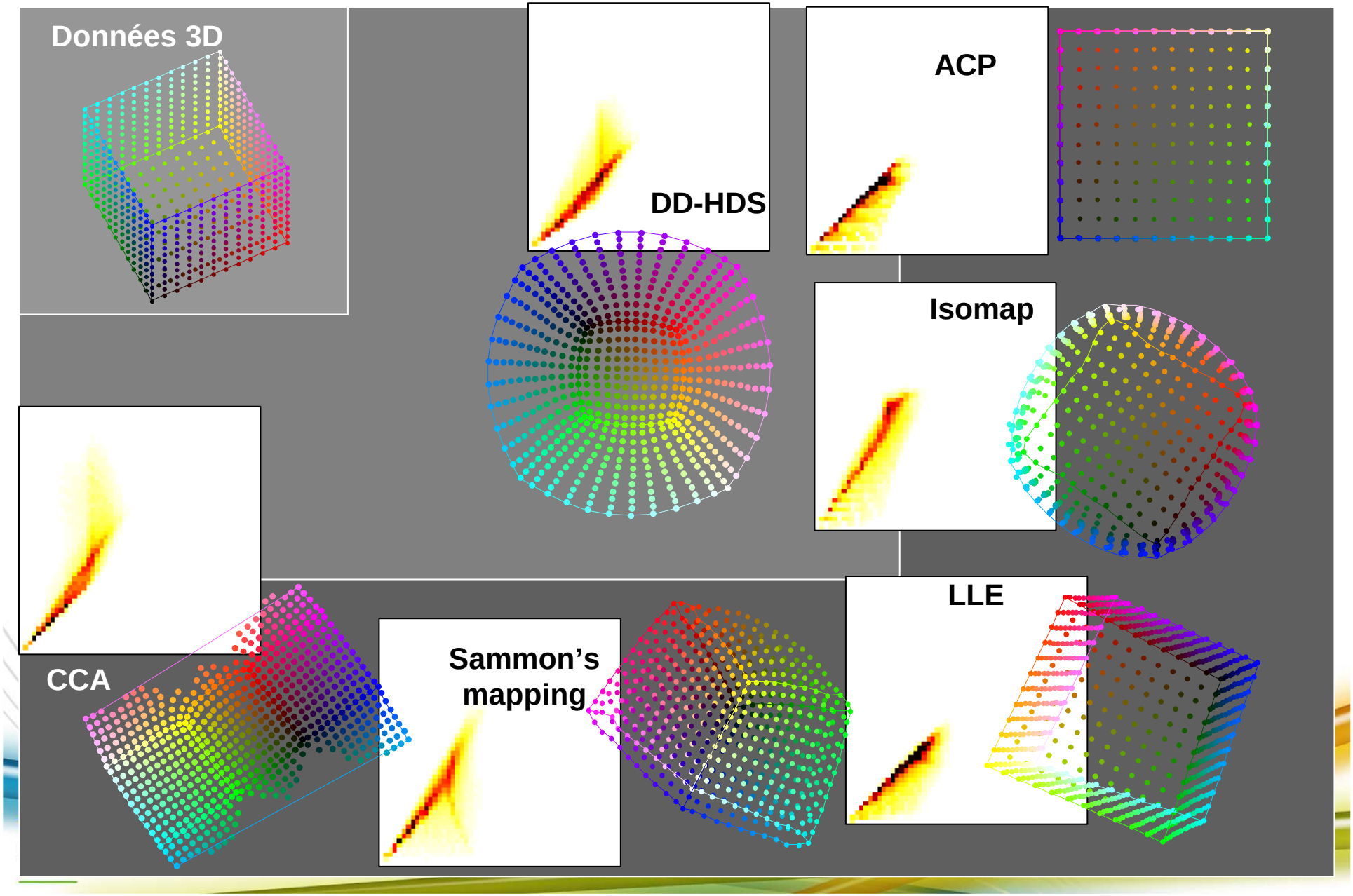
Sammon's mapping



LLE

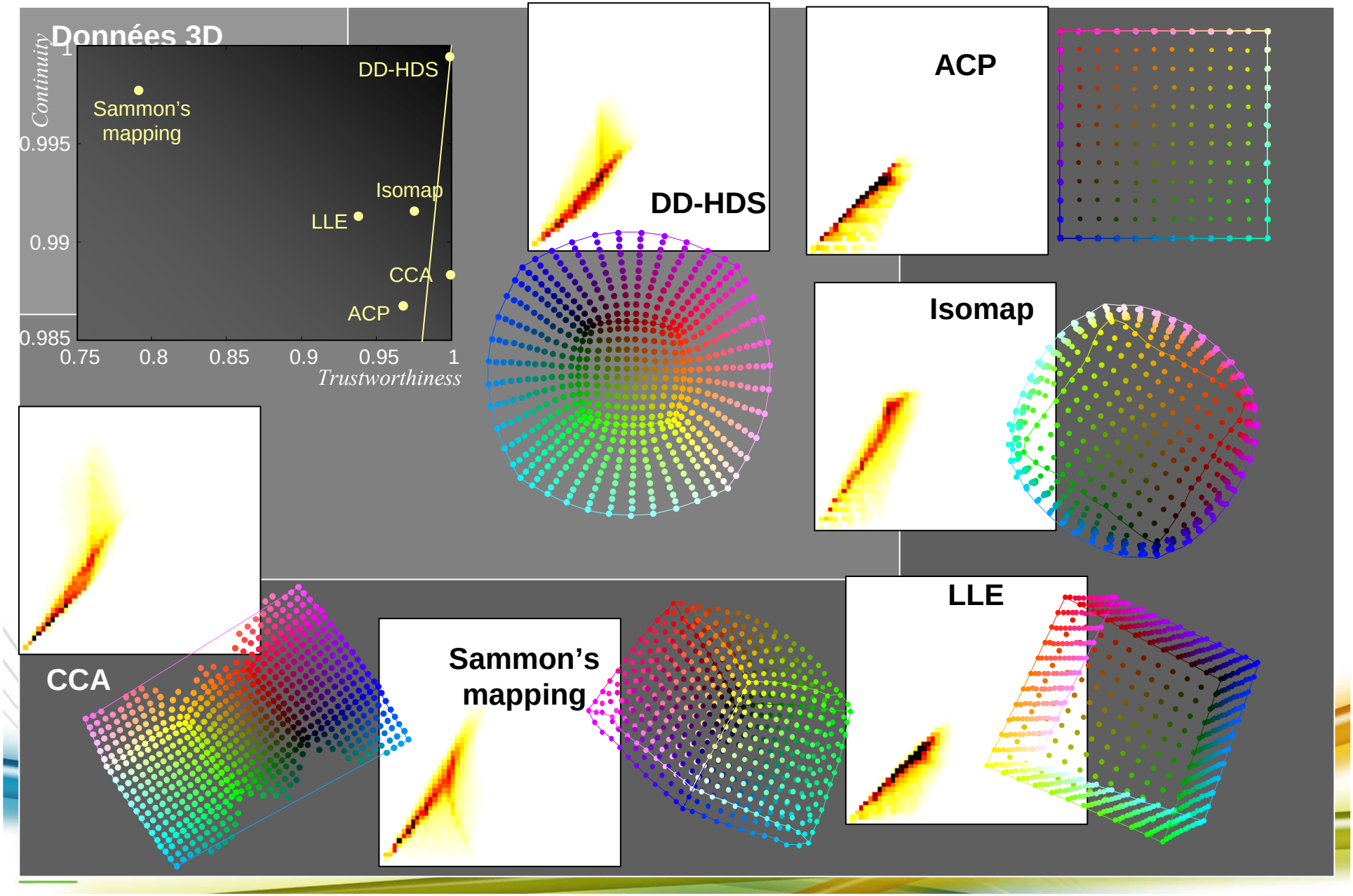


Exemple de projection : le cube ouvert



Exemple de projection : le cube ouvert

J. Venna J. et S. Kaski, 2001.



Indices quantifiant les faux voisinages et déchirements

$$\zeta = \sum_{i,j} \left(d_{ij} - \delta_{ij}\right)^2 \times f\left(d_{ij}\right)$$

Sensible aux déchirements

$$\zeta = \sum_{i,j} \left(d_{ij} - \delta_{ij}\right)^2 \times f\left(\delta_{ij}\right)$$

Sensible aux faux-voisinages

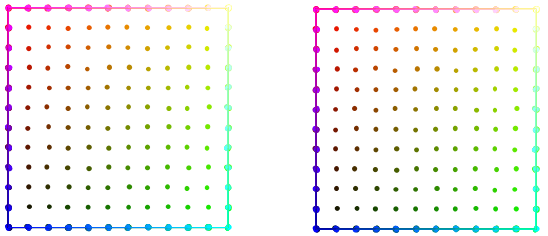
Ils peuvent être mesurés localement

$$\zeta_i = \sum_{j} \left(d_{ij} - \delta_{ij}\right)^2 \times f\left(d_{ij}\right)$$

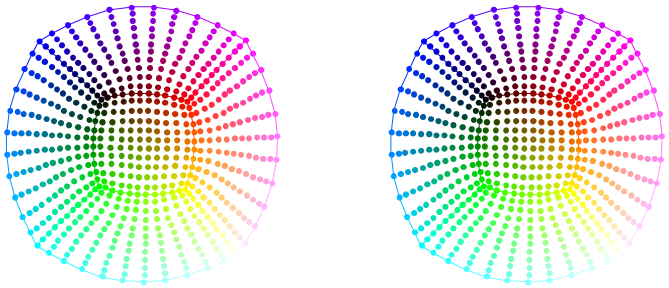
$$\zeta_i = \sum_{j} \left(d_{ij} - \delta_{ij}\right)^2 \times f\left(\delta_{ij}\right)$$

Exemple de projection : le cube ouvert

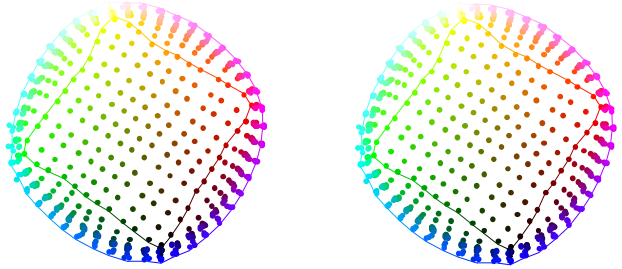
ACP

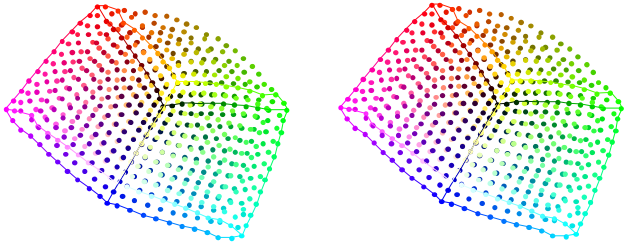


DD-HDS

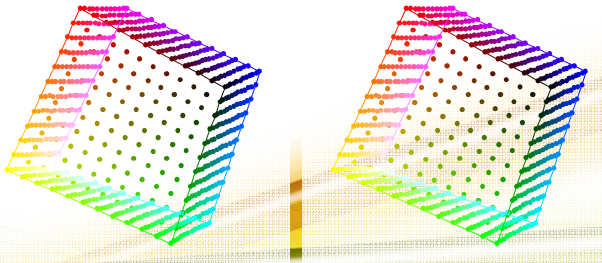


Isomap



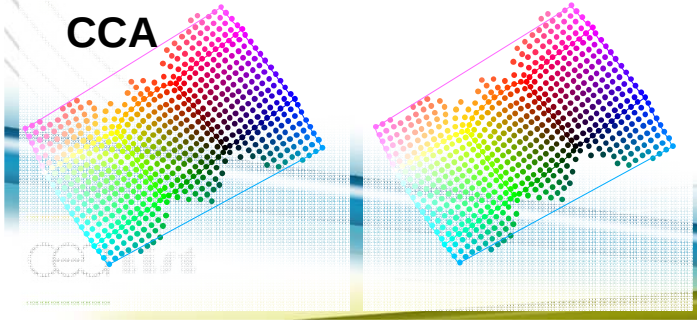


LLE



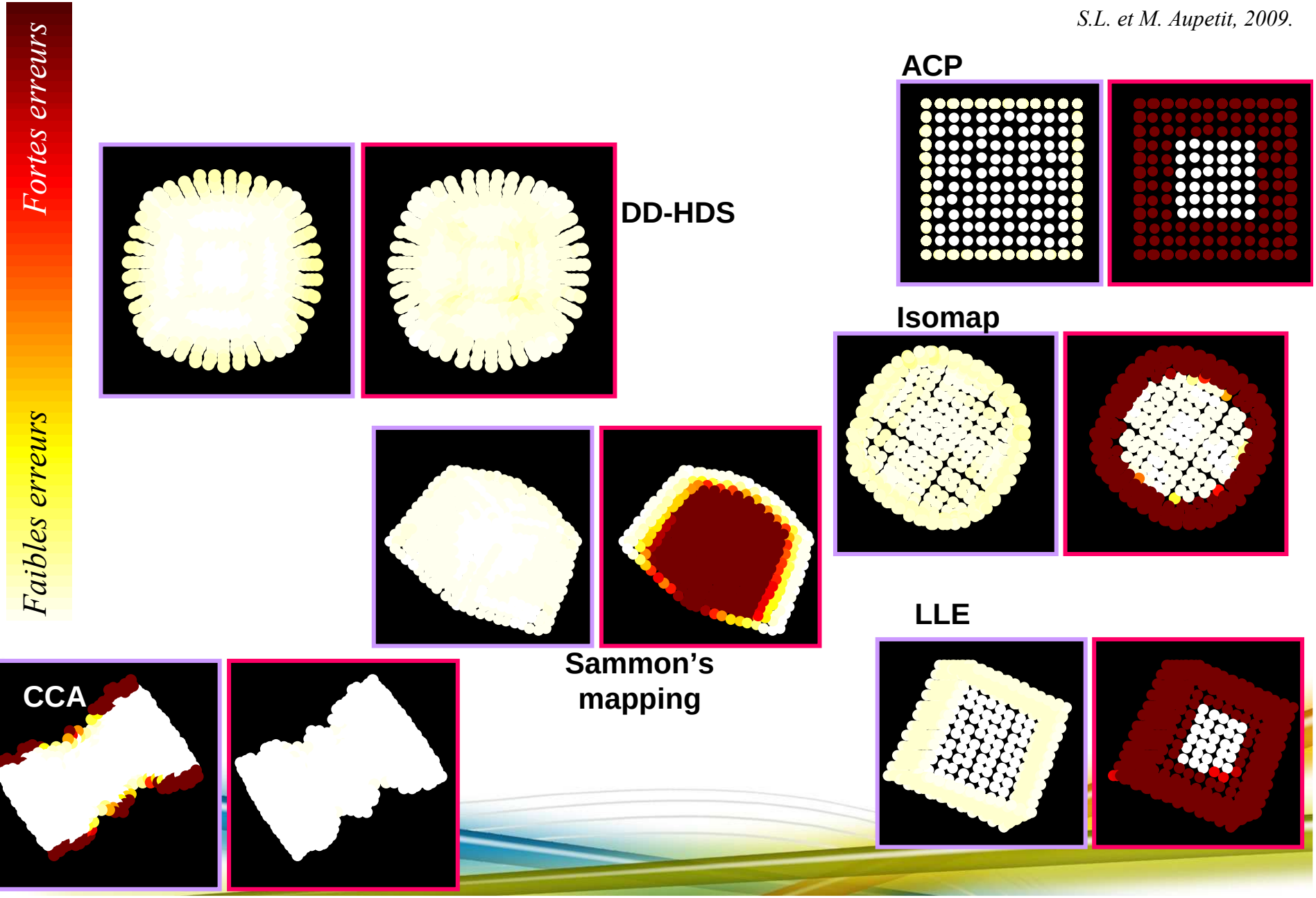
Sammon's mapping

CCA



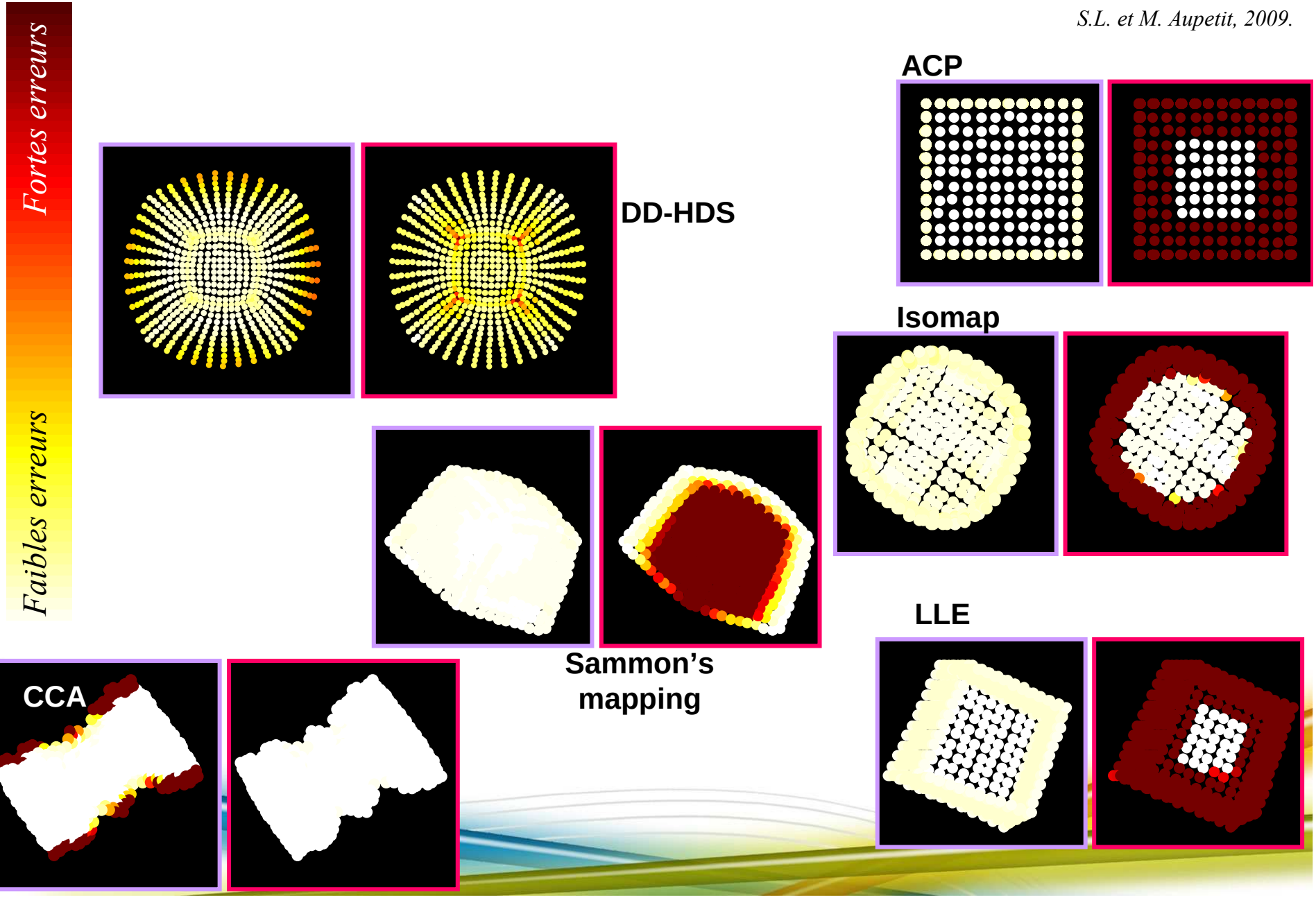
Exemple de projection : le cube ouvert

S.L. et M. Aupetit, 2009.

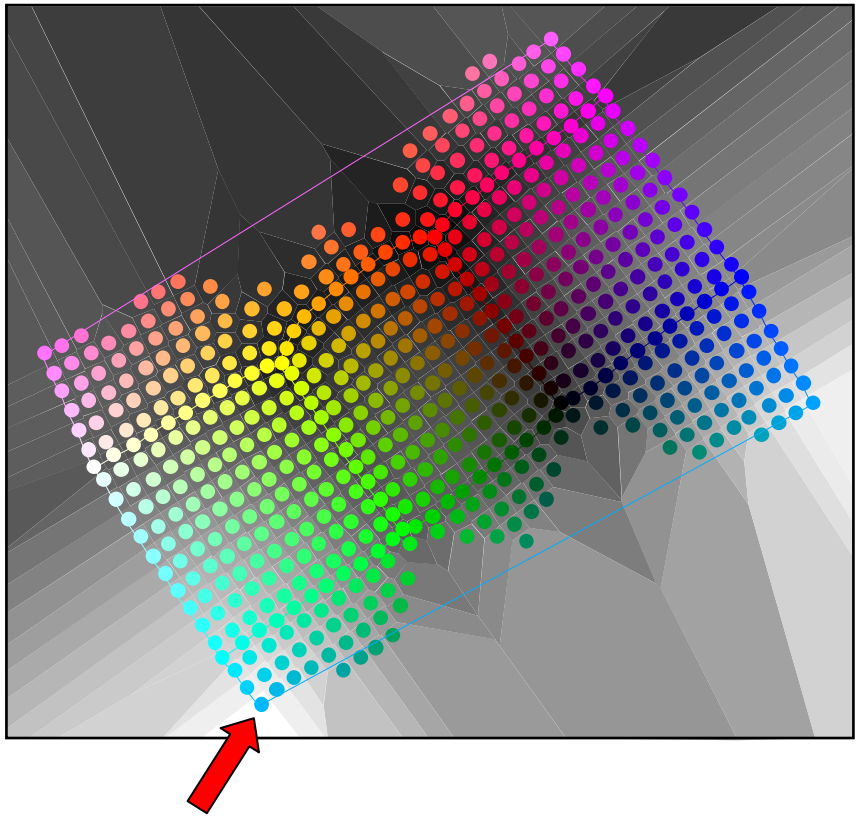


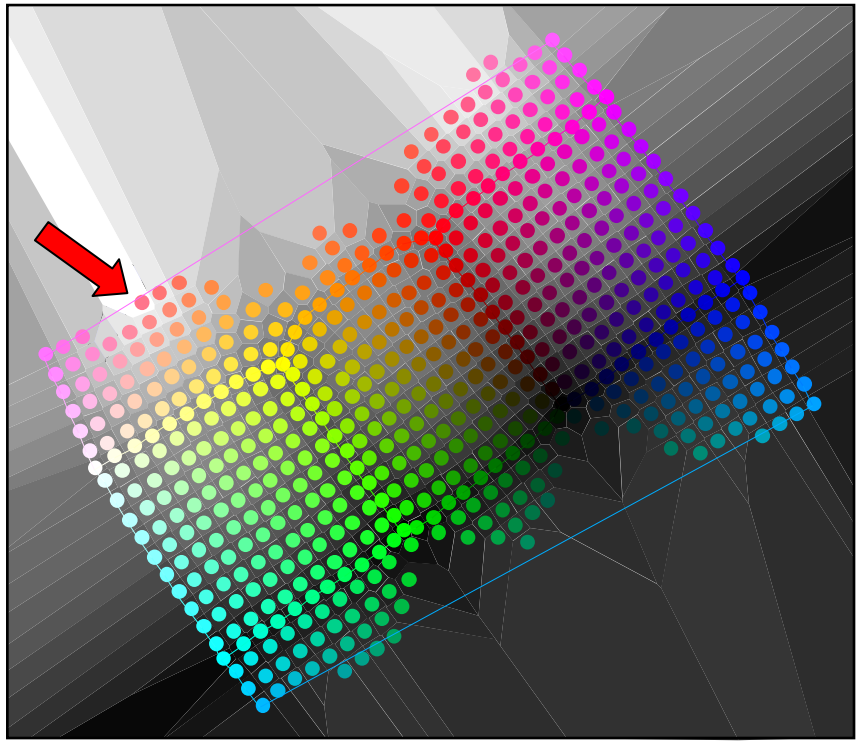
Exemple de projection : le cube ouvert

S.L. et M. Aupetit, 2009.



Proximités selon un point de vue



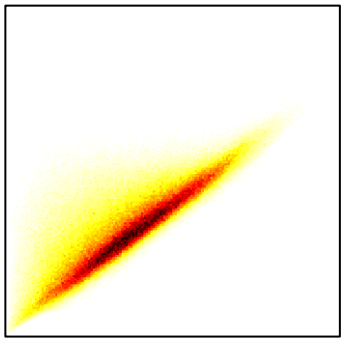


Niveau de gris : distance de la référence aux autres données

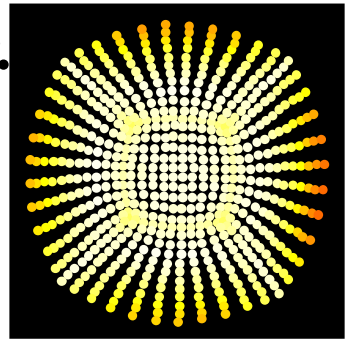
M. Aupetit, 2007.

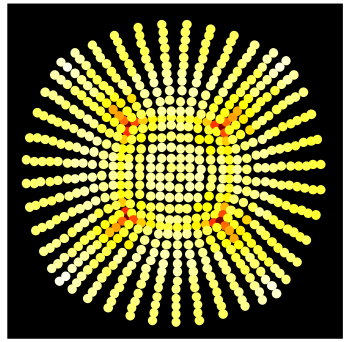
Analyse des défauts: une stratégie en 3 étapes

1) *Analyse globale.*

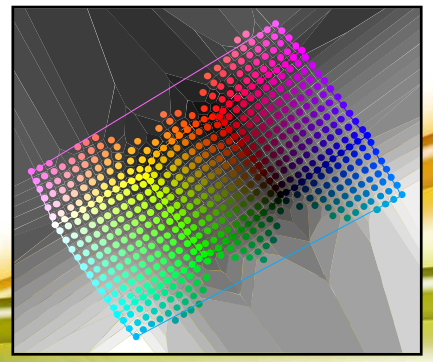


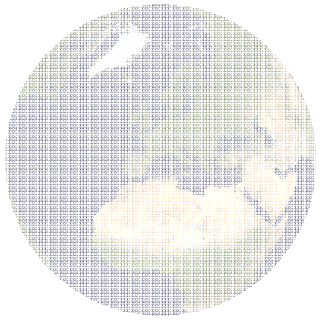
2) *Evaluation locale du niveau d'erreur.*



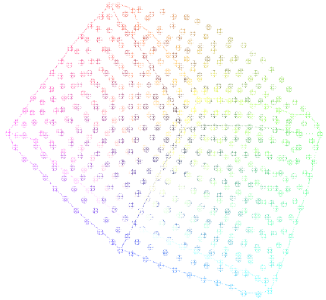


3) *Etudes des proximités selon un point de vue.*

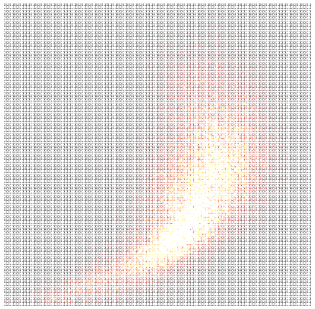




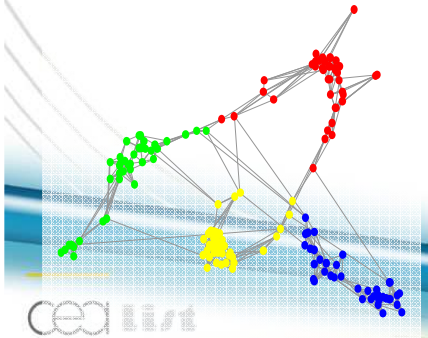
I. Géographie d'un jeu de données



II. Réduction de dimension à partir des distances



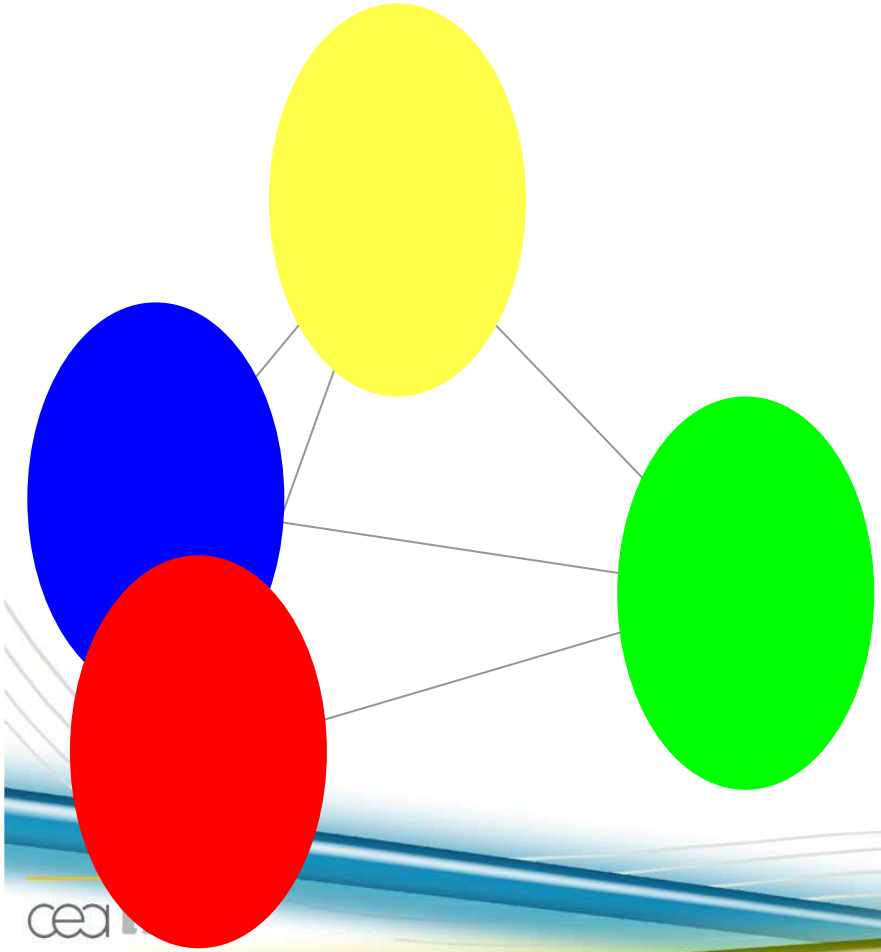
III. Evaluations des mappings



IV. Réduction de dimension à partir des rangs de voisinage

Limites des MDS

Nous constatons que sur certains jeux de données complexes, les MDS expriment mal certaines relations entre données.



Exemple de données: 10D, 4 classes générées par des tirages gaussiens centrés sur les sommets d'une pyramide régulière (variance plus élevée sur une dimension).

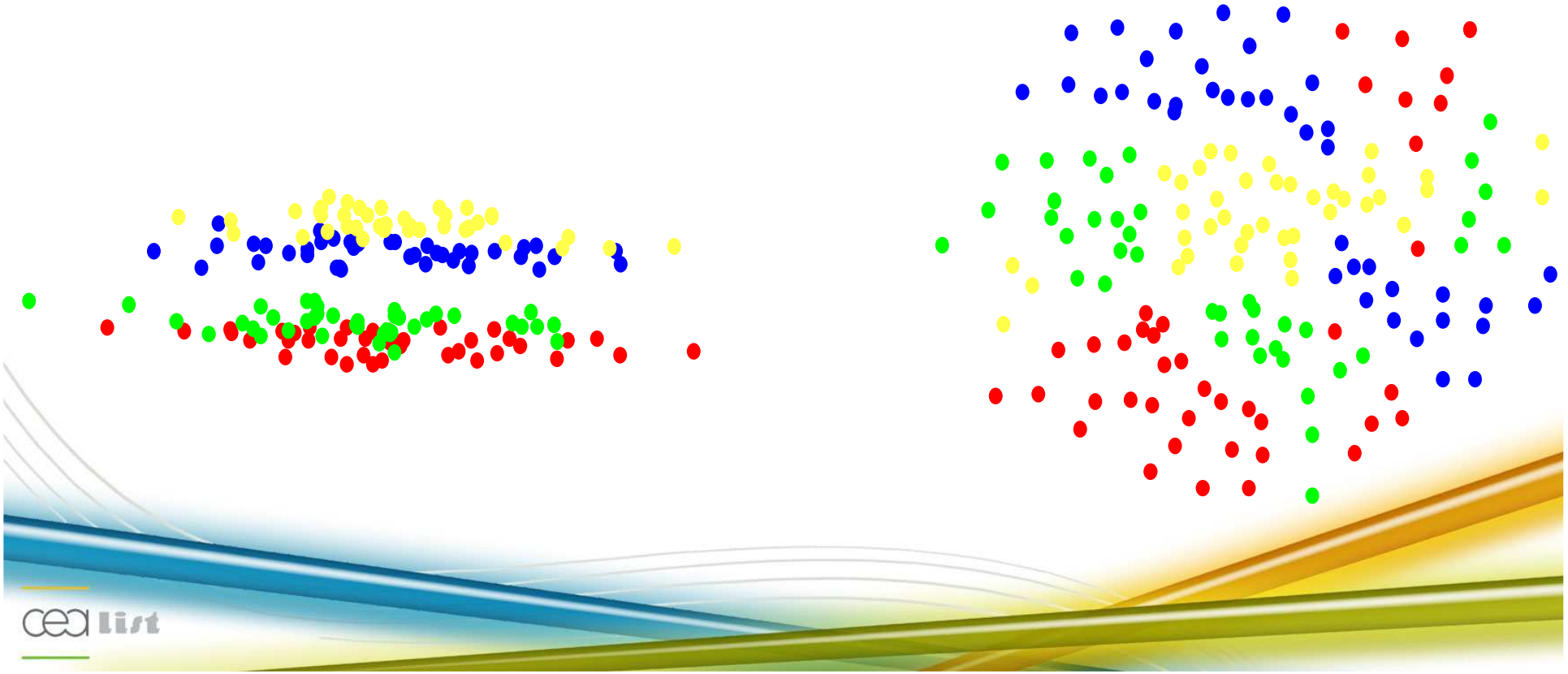
Grande dimension – Anisotropie - Classes à égale distance les unes des autres - Distances intra-classes et inter-classes du même ordre de grandeur.

Limites des MDS

4 classes très différenciées que l'on retrouve mal à travers les MDS

ACP

DD-HDS



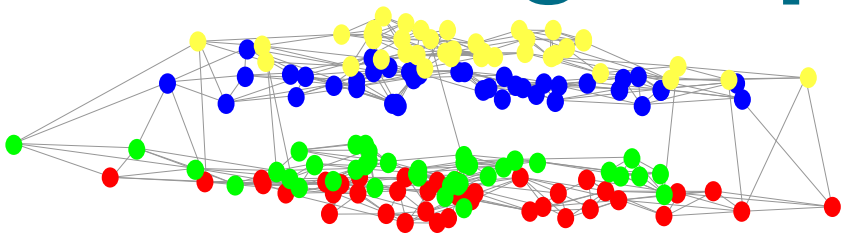
RankVisu : préservation des rangs de voisinage

Rang 1 -> le plus proche voisin
Rang 2 -> le 2eme plus proche voisin
…

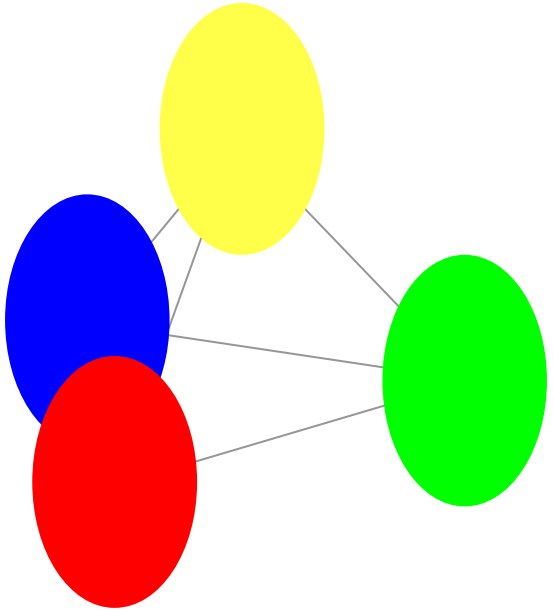
Objectif: préservation des rangs de voisinage en avantageant les petits rangs.

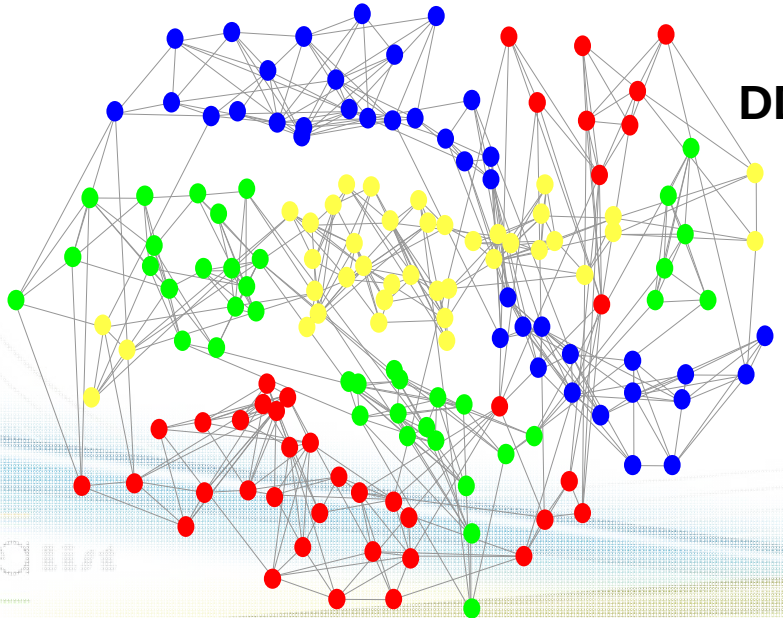
S.L., B. Fertil, P. Villemain et J. Hérault, 2009.

Les 4 groupes anisotropes



ACP

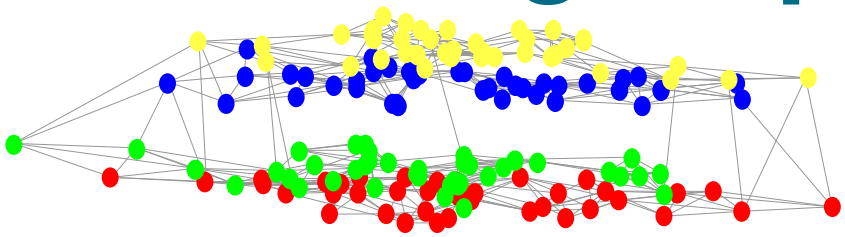




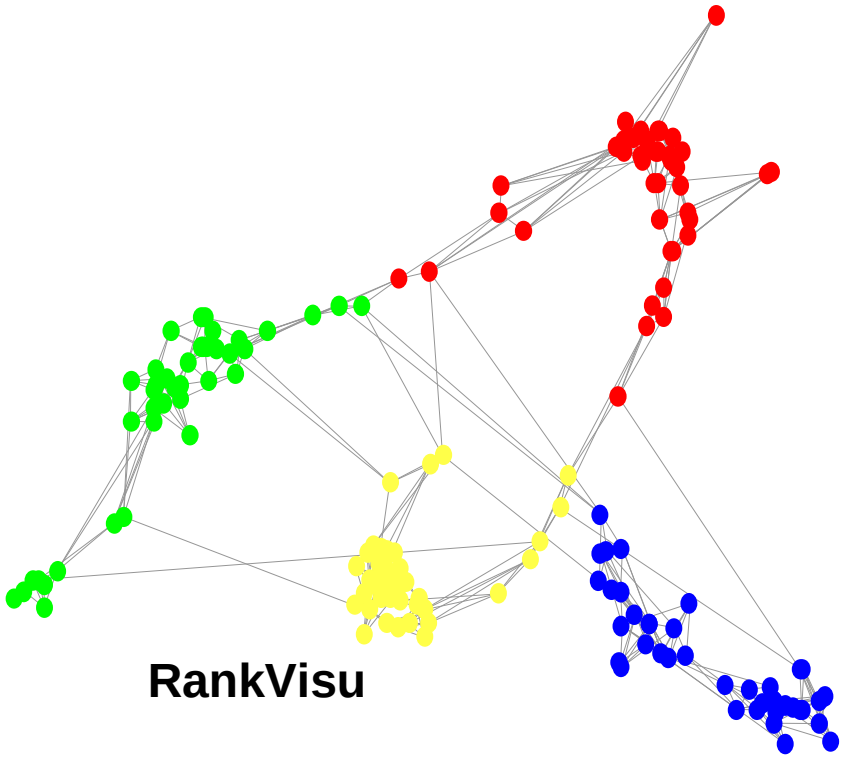
DD-HDS

Chaque point est connecté à ses 5 plus proches voisins

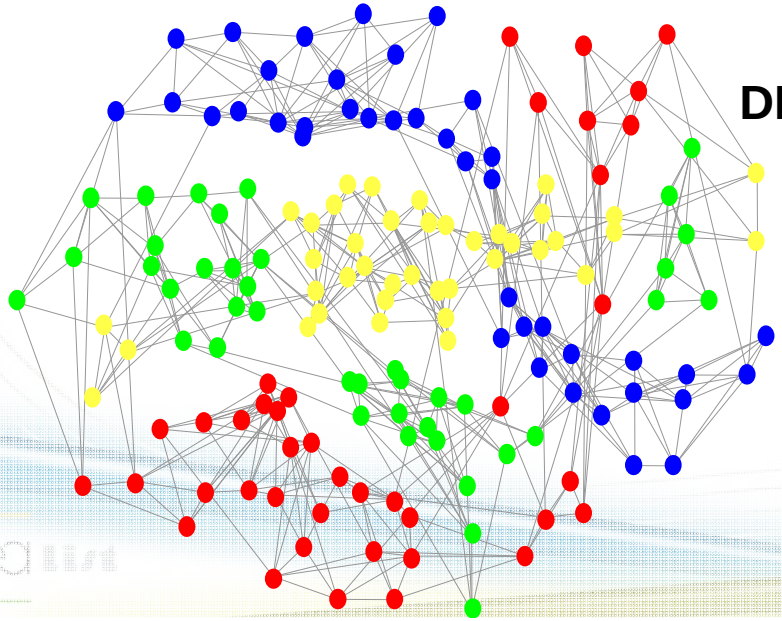
Les 4 groupes anisotropes



ACP



RankVisu

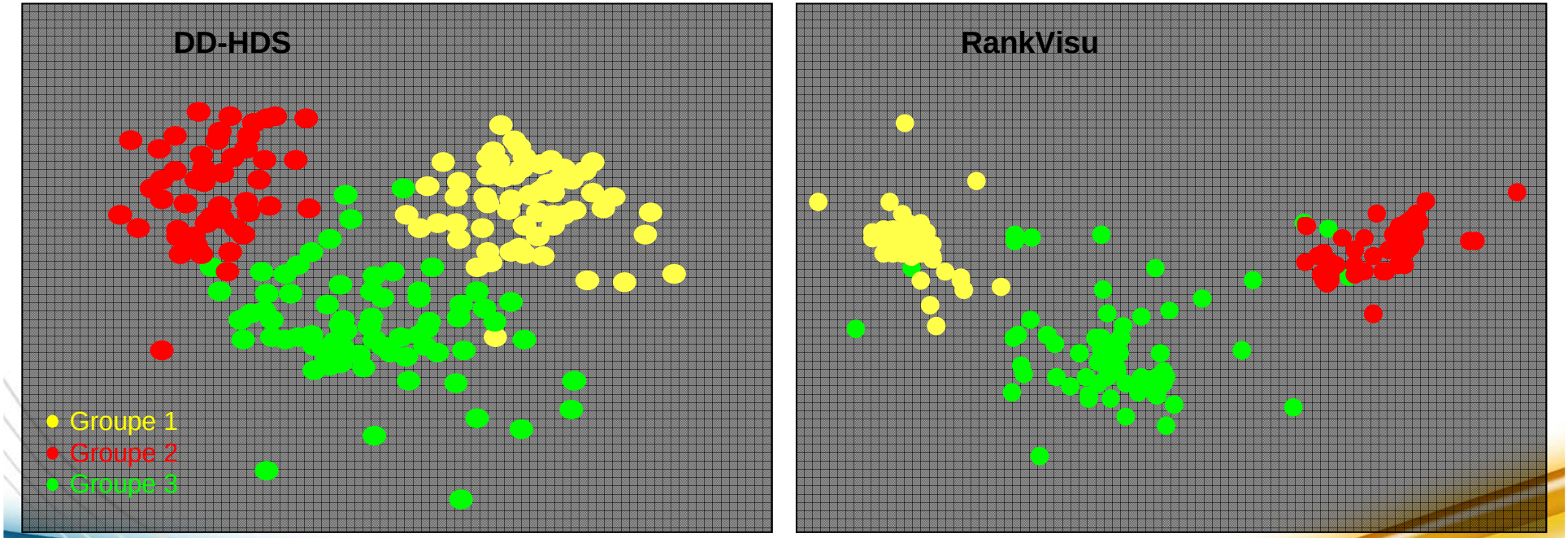


DD-HDS

Chaque point est connecté à ses 5 plus proches voisins

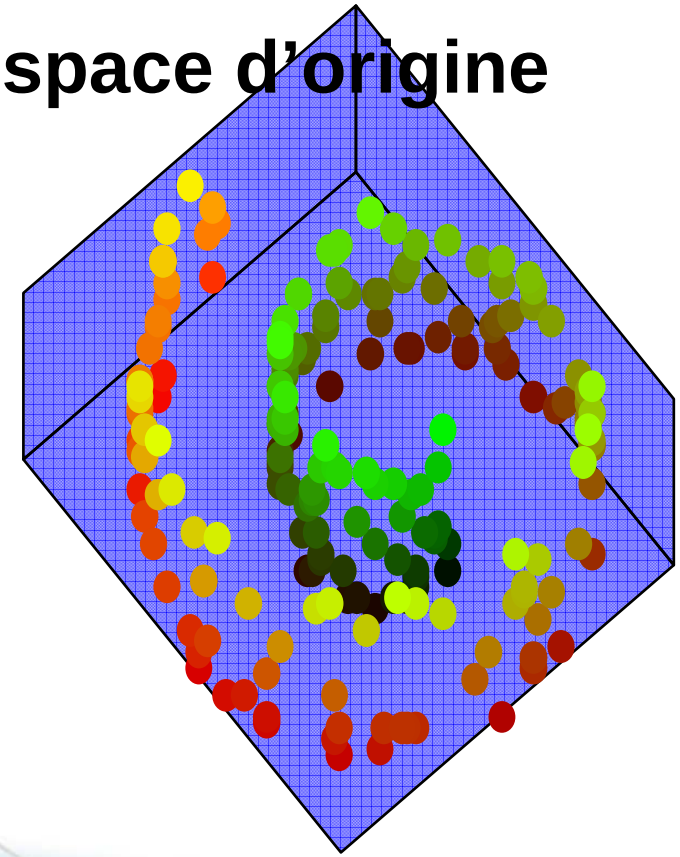
"Wine data"

mesures chimiques sur les vins de 3 producteurs

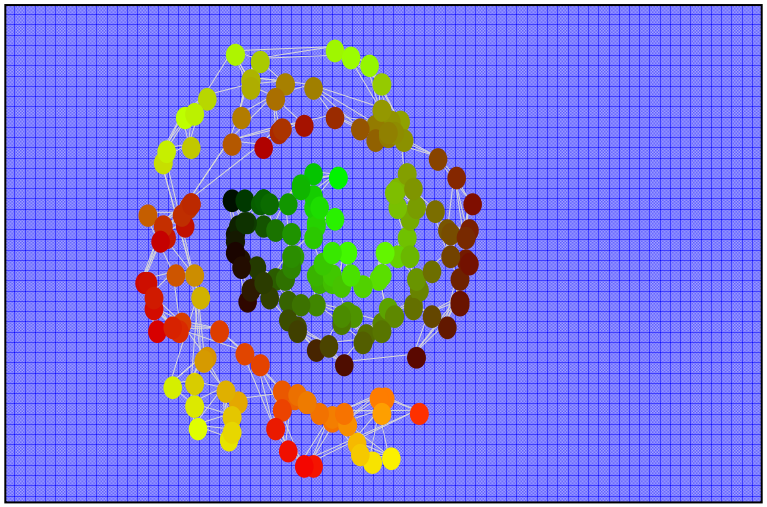


Exemple sur le "swiss roll"

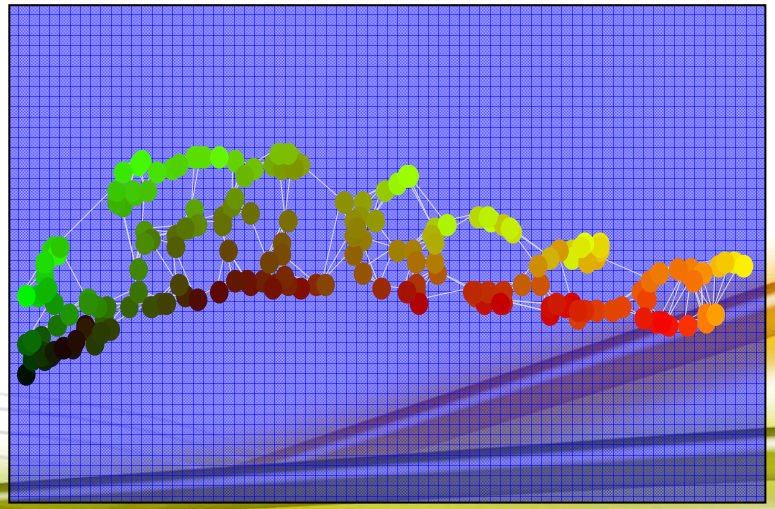
Espace d'origine



DD-HDS



RankVisu



Conclusion

Les réductions de dimension permettent de découvrir l'organisation spatiale des données.

J'ai détaillé ici deux méthodes particulièrement efficaces quand on est face à des données de grande dimension.

L'évaluation des résultats est essentielle. Nous préconisons une procédure en 3 temps: évaluation globale → locale → selon quelques points de vue.