3I026 - INTRODUCTION À L'INTELLIGENCE ARTIFICIELLE ET AUX DATA SCIENCES

Vincent Guigue
Christophe Marsala

Sorbonne Université

SCIENCES SORBONNE UNIVERSITÉ

LIP6

MLIA Machine Learning & Deep Learning for Information Access

Feedback semaine précédente

UE IADS = triple difficulté

- Dimension mathématique et statistique
- Dimension algorithmique
- Dimension implémentation

Plan statistique :

- Définir une gaussienne, en 1D, en 2D
- Travailler sur une multinomiale
- Générer des points tirés selon une gaussienne en 2D

Feedback semaine précédente

UE IADS = triple difficulté

- Dimension mathématique et statistique
- Dimension algorithmique
- Dimension implémentation

Plan algorithmique :

- Gérer les itérations de l'algorithme du perceptron
- Algorithmes full gradient / batch / purement stochastique
- Notion d'*epoch*
- Détection de la convergence dans une descente de gradient

Feedback semaine précédente

UE IADS = triple difficulté

- Dimension mathématique et statistique
- Dimension algorithmique
- Dimension implémentation

Plan implémentation :

- A-t-on besoin de fournir les dimensions des entrées/sorties pour la création du classifieur ?
- Quand initialiser les paramètres du classifieur ?
- Comment gérer les epochs (solutions internes & externes)

Visualisation

- **InfoVis** = Information Visualization

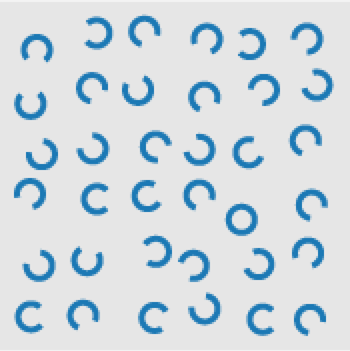
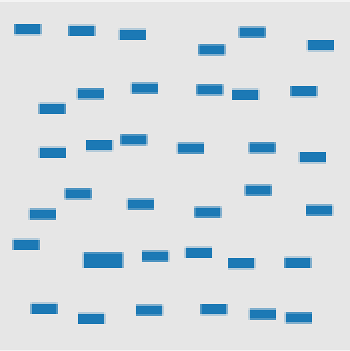
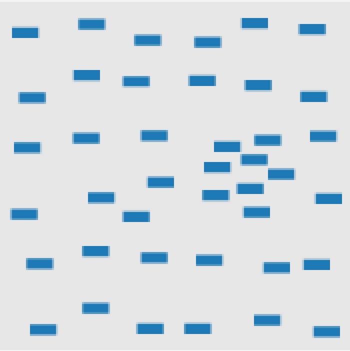
> The use of computer-supported interactive, visual representation of abstract data to amplify cognition
> Card, Mackinlay & Shneiderman

- **DataVis** = Data Visualization
- Deux problèmes extrêmement importants dans la data science
- Deux problèmes peu abordés...

Référence utile : Cours de F. Rossi
`http://apiacoa.org/teaching/visualization/index.fr.html`

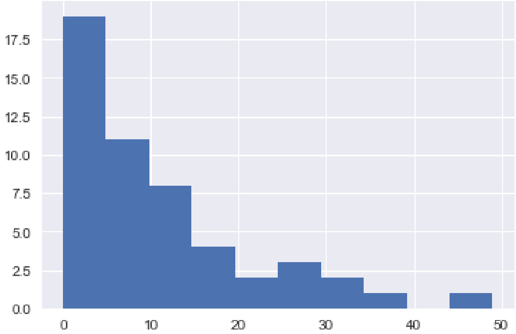
$\Rightarrow$ Lien avec l'apprentissage statistique : Quelles méthodes permettent de trouver automatiquement de bonnes visualisations des données ?

- Extraction de caractéristiques de base en 200ms
- Possibilités d'analyse de densité / détection d'anomalie très rapide

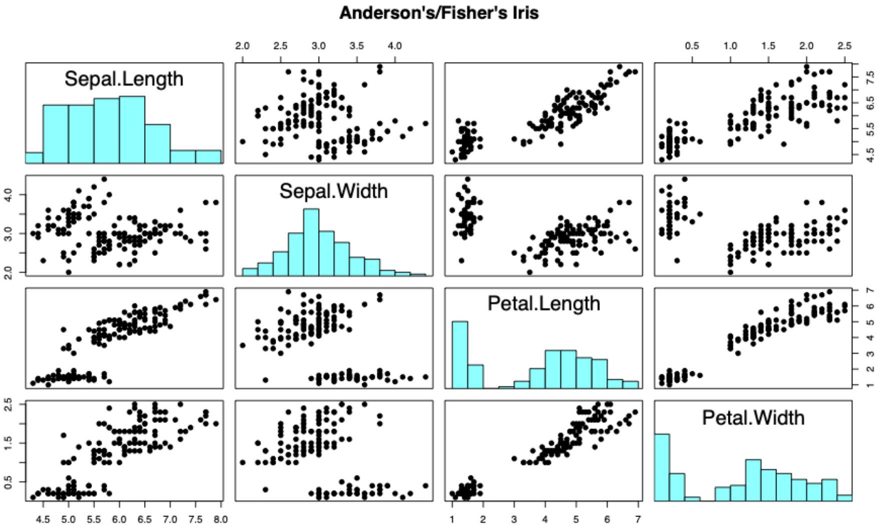
https://www.csc2.ncsu.edu/faculty/healey/PP/index.html

- Focus sur une dimension X_j
  - N Observations x_{ij}
- Solution pour la visualisation du contenu : l'histogramme

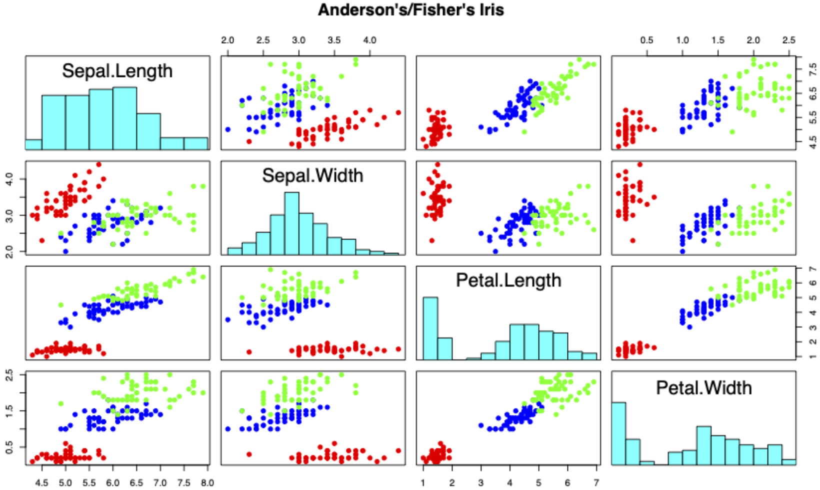


SCIENCES SORBONNE UNIVERSITÉ

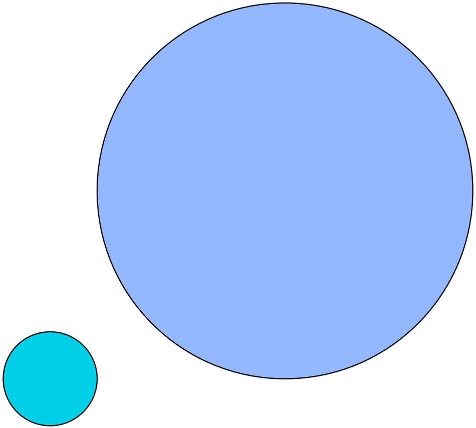
Données orgininales = Iris, 4D : comment visualiser ? ⇒ Scatter plot



Données orgininales = Iris, 4D : comment visualiser ? ⇒ Scatter plot



Avec les informations de classes



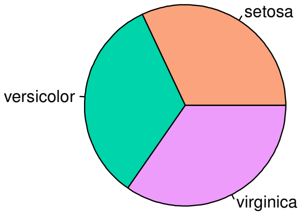
Please write down your estimation of the ratio of the areas of those disks.

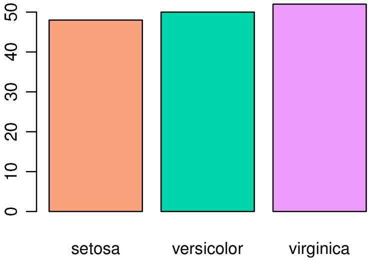
Please write down your estimation of the ratio of the lengths of those bars.

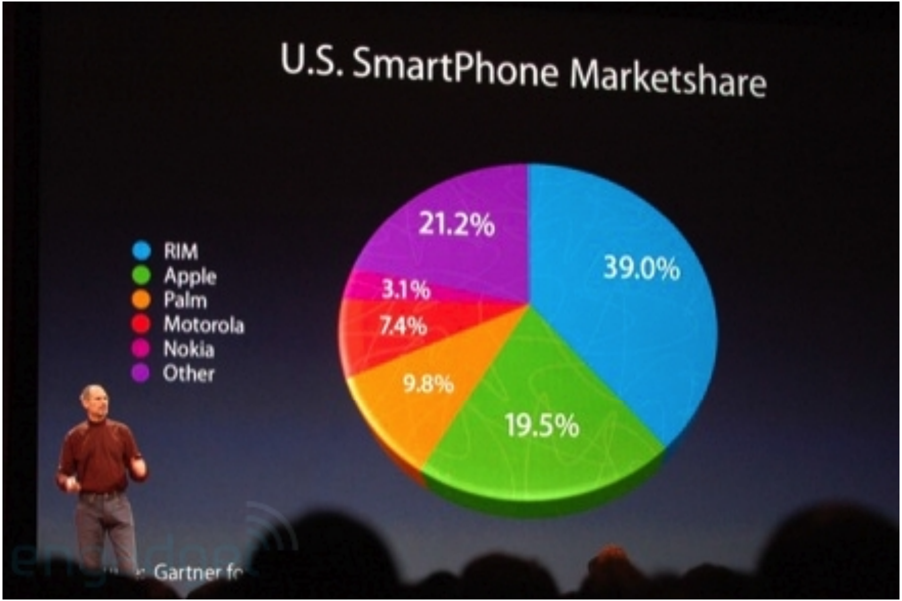
Another visual abstraction

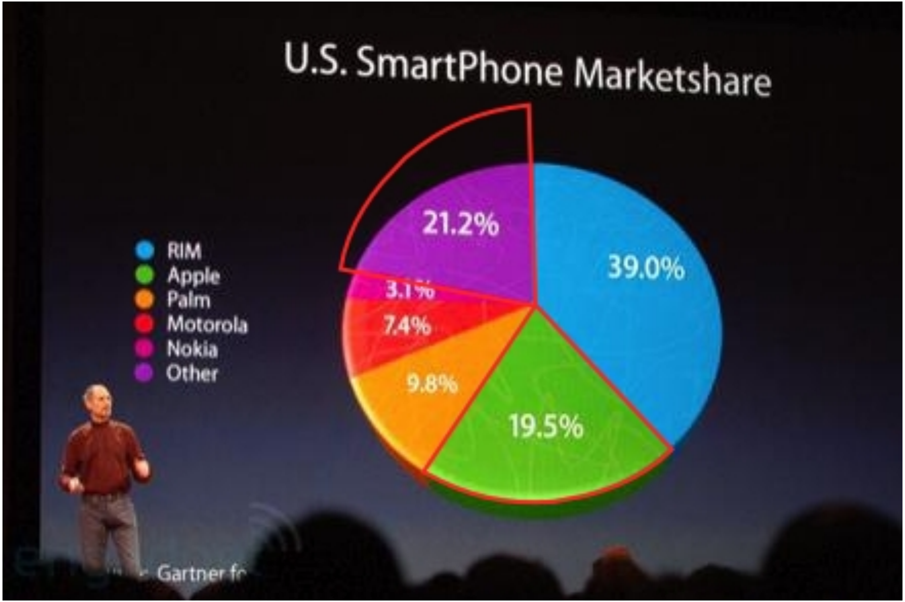
Using the same counting data, replace the Q pie slices by Q bars with length/height proportional to N_q

And the views are






U.S. SmartPhone Marketshare
RIM
Apple
Palm
Motorola
Nokia
Other
21.2%
39.0%
3.1%
7.4%
9.8%
19.5%
Gartner f


U.S. SmartPhone Marketshare
RIM
Apple
Palm
Motorola
Nokia
Other
21.2%
39.0%
7.4%
9.8%
19.5%
Gartner

Malédiction de la dimensionnalité

Objectifs de la visualisation de données

1. Eliminer du bruit
   - Labélisation
   - captation

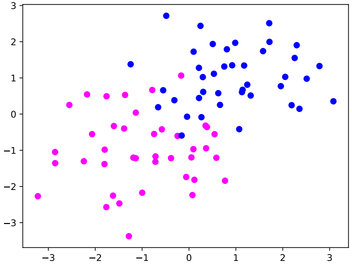
   $\Rightarrow$ Transformation légère
2. Comprendre des modes de fonctionnement des données
   - Zones de densités
   - Feature engineering

   $\Rightarrow$ Transformation avancée

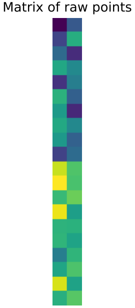
SCIENCES SORBONNE UNIVERSITÉ

A classical toy example to illustrate the curse of dimensionality :

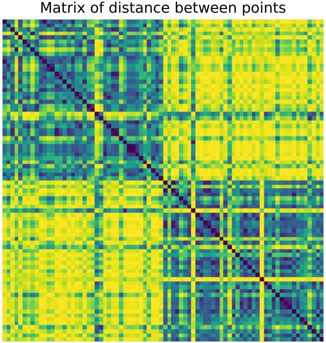
Original dataset :



Matrix view



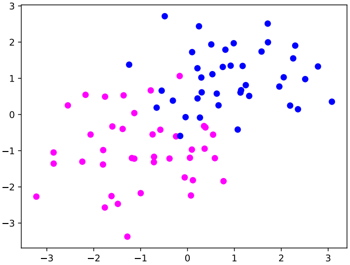
Distance matrix



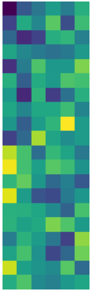
Easy problem / classes are clearly separated

A classical toy example to illustrate the curse of dimensionality :

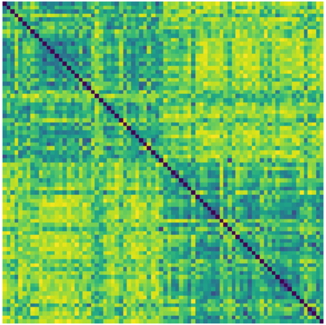
Original dataset :



Matrix view



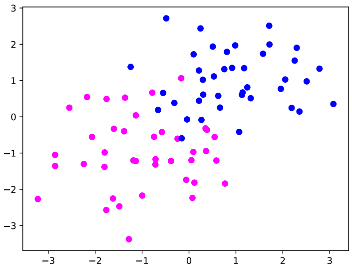
Distance matrix



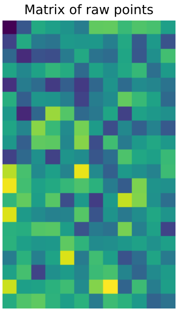
Adding some noisy dimensions in the dataset

A classical toy example to illustrate the curse of dimensionality :

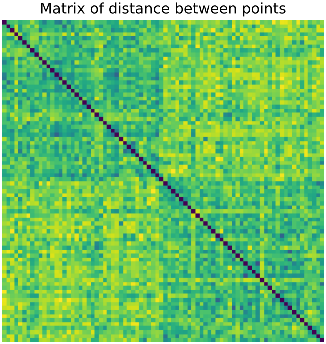
Original dataset :



Matrix view



Distance matrix

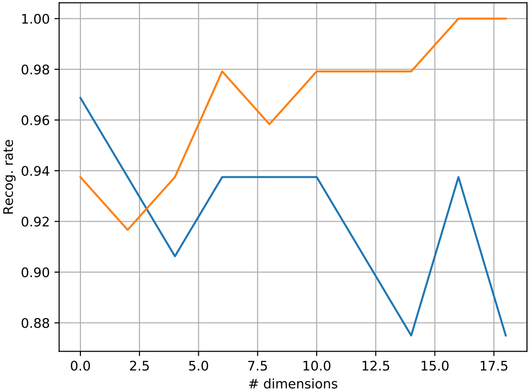


Adding **more** noisy dimensions in the dataset

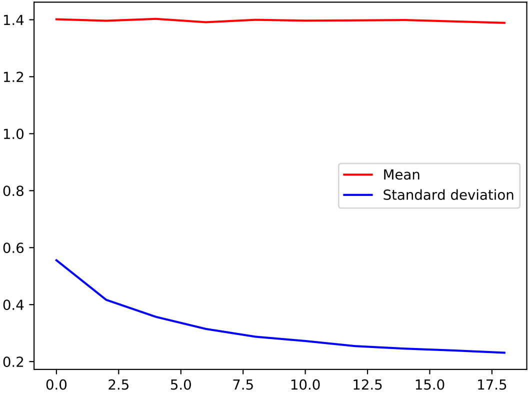
$\Rightarrow$ Euclidian distance is very sensitive to the dimensionality issue

A classical toy example to illustrate the curse of dimensionality :

Basic classifier on those datasets



Distances between points in the dataset



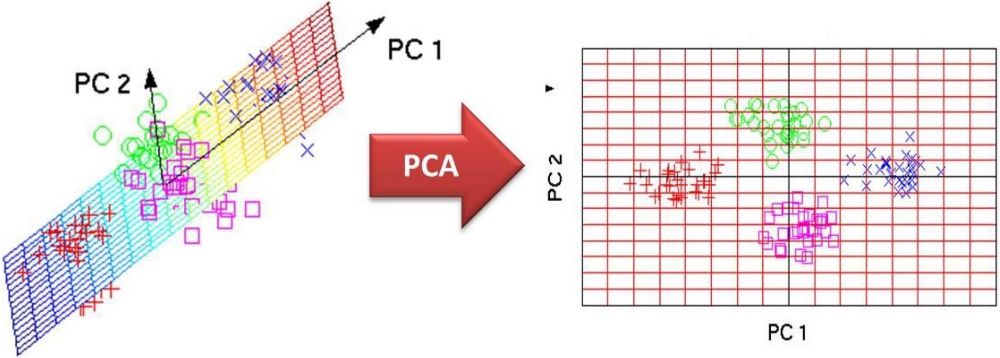
$\Rightarrow$ Learn accuracy $\nearrow$, test accuracy $\searrow$ = **overfitting**

$\Rightarrow$ All points tend to lay on an hypersphere (they become equidistant)

Transformations avancées

ACP (PCA) = outil de base pour

1. La visualisation de données en grande dimension
2. La réduction de la dimension et du bruit

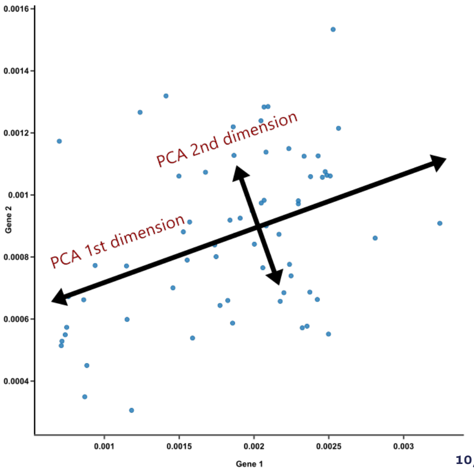


ACP (PCA) = outil de base pour

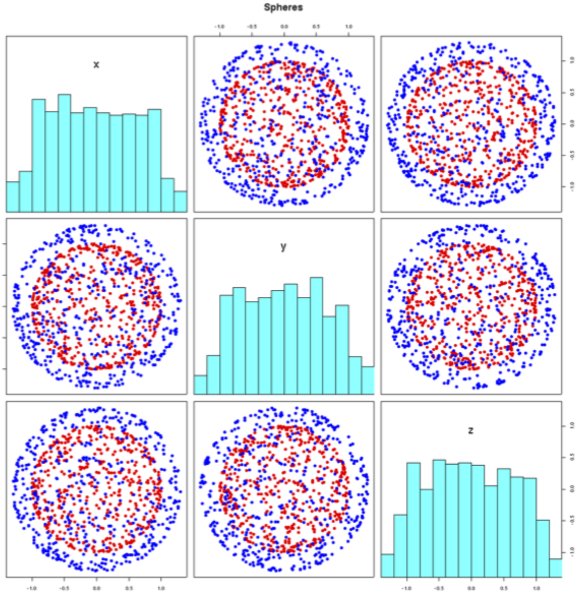
1. La visualisation de données en grande dimension
2. La réduction de la dimension et du bruit

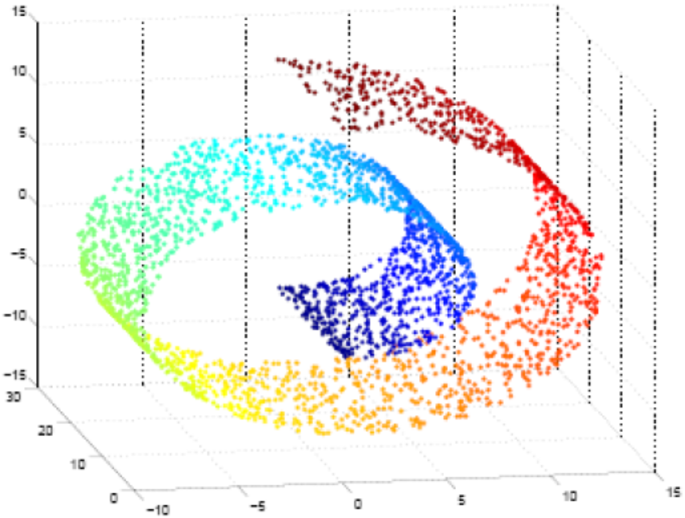
Idée : trouver des axes qui maximise la variance $\Rightarrow$ projeter sur ces axes

- Transformation non supervisée
- Transformation applicable sur de nouveaux points
  1. $X \in \mathbb{R}^{N \times d}$
  2. ACP sur $X^T X \in \mathbb{R}^{d \times d}$
  3. Récupétation de $\{V_i \in \mathbb{R}^d, \lambda_i \in \mathbb{R}_+\}_{i=1,\dots,d}$
  4. d Axes de projection V_i... associés à leur *force d'explication* λ_i
  5. Utilisation des V_i sur les données de test

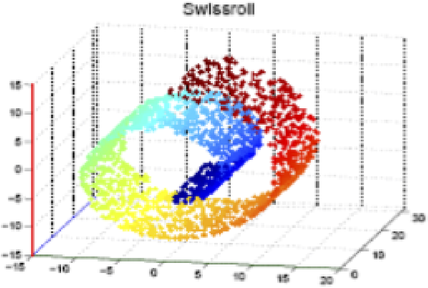


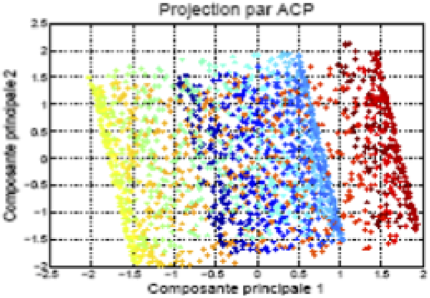
Limite de la sélection de variables



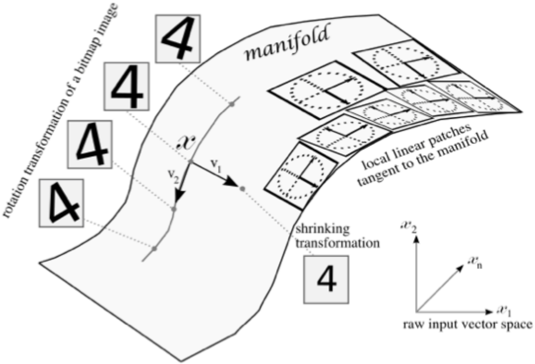

15
10
5
0
-5
-10
-15
30
20
10
0
-10
-5
0
5
10
15

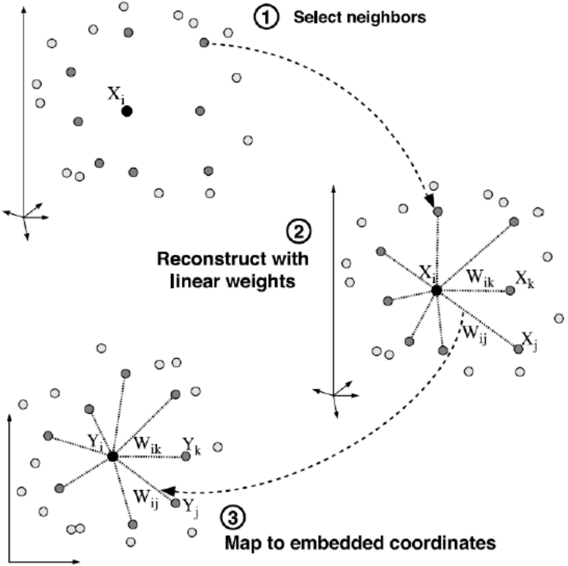
LLE : local linear embedding





Idée : Les données sont organisées selon une variété

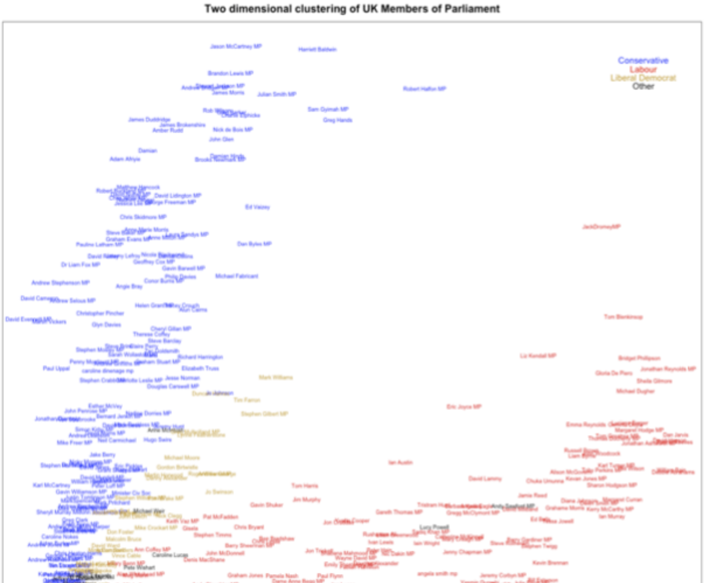




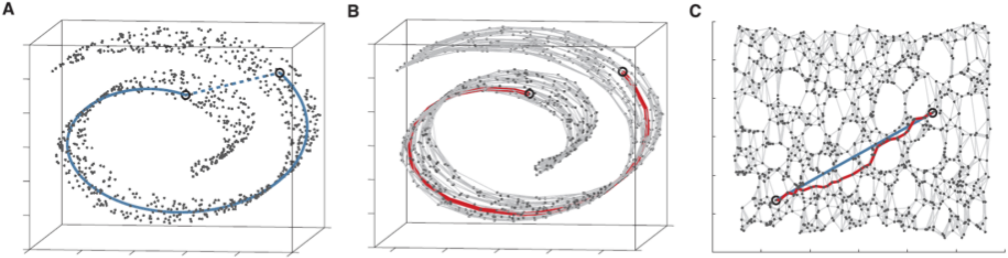
1 Select neighbors
X_i
2 Reconstruct with linear weights
X_i
W_ik
X_k
W_ij
X_j
Y_i
W_ik
Y_k
W_ij
Y_j
3 Map to embedded coordinates

MDS : multi-dimensional scaling

	Atlanta (1)	Boston (2)	Cincinnati (3)	Columbus (4)	Dallas (5)	Indianapolis (6)	Little Rock (7)	Los Angeles (8)	Memphis (9)	St. Louis (10)	Spokane (11)	Tampa (12)
(1)	0											
(2)	1068	0										
(3)	461	867	0									
(4)	549	769	107	0								
(5)	805	1819	943	1050	0							
(6)	508	941	108	172	882	0						
(7)	505	1494	618	725	325	562	0					
(8)	2197	3052	2186	2245	1403	2080	1701	0				
(9)	366	1355	502	586	464	436	137	1831	0			
(10)	558	1178	338	409	645	234	353	1848	294	0		
(11)	2467	2747	2067	2131	1891	1959	1988	1227	2042	1820	0	
(12)	467	1379	928	985	1077	975	912	2480	779	1016	2821	0

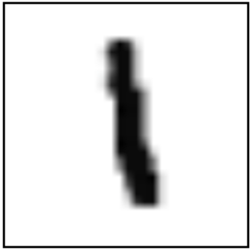

Two dimensional clustering of UK Members of Parliament
Conservative
Labour
Liberal Democrat
Other

ISOMAP



Step		
1	Construct neighborhood graph	Define the graph G over all data points by connecting points i and j if [as measured by $d_X(i,j)$] they are closer than ϵ (ϵ-Isomap), or if i is one of the K nearest neighbors of j (K-Isomap). Set edge lengths equal to $d_X(i,j)$.
2	Compute shortest paths	Initialize $d_G(i,j) = d_X(i,j)$ if i,j are linked by an edge; $d_G(i,j) = \infty$ otherwise. Then for each value of $k = 1, 2, \ldots, N$ in turn, replace all entries $d_G(i,j)$ by $\min\{d_G(i,j), d_G(i,k) + d_G(k,j)\}$. The matrix of final values $D_G = \{d_G(i,j)\}$ will contain the shortest path distances between all pairs of points in G (*16*, *19*).
3	Construct d-dimensional embedding	Let λ_p be the p-th eigenvalue (in decreasing order) of the matrix $\tau(D_G)$ (*17*), and v_p^i be the i-th component of the p-th eigenvector. Then set the p-th component of the d-dimensional coordinate vector $\mathbf{y}_i$ equal to $\sqrt{\lambda_p} v_p^i$.

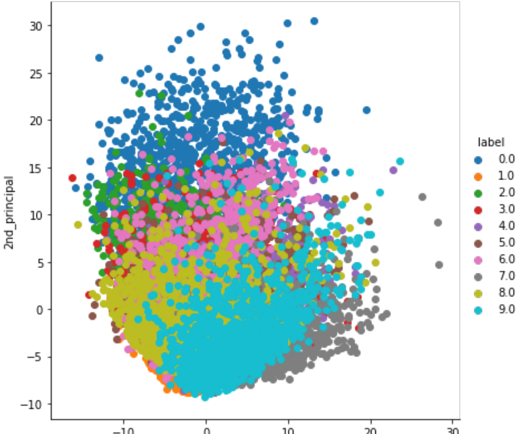
Que se passe-t-il sur des données USPS ou MNIST ? 256/384 dimensions $\Rightarrow$ 2D !



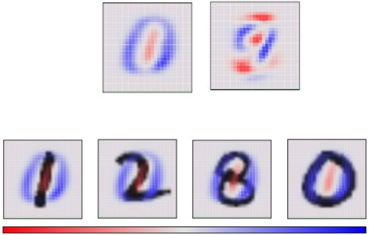
$$\simeq \begin{bmatrix}
0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&.6&.8&0&0&0&0&0&0\\
0&0&0&0&0&0&.7&1&0&0&0&0&0&0\\
0&0&0&0&0&0&.7&1&0&0&0&0&0&0\\
0&0&0&0&0&0&.5&1&.4&0&0&0&0&0\\
0&0&0&0&0&0&0&1&.4&0&0&0&0&0\\
0&0&0&0&0&0&0&1&.4&0&0&0&0&0\\
0&0&0&0&0&0&0&1&.7&0&0&0&0&0\\
0&0&0&0&0&0&0&1&1&0&0&0&0&0\\
0&0&0&0&0&0&0&.9&1&.1&0&0&0&0\\
0&0&0&0&0&0&0&.3&1&.1&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0
\end{bmatrix}$$

Que se passe-t-il sur des données USPS ou MNIST ? 256/384 dimensions $\Rightarrow$ 2D !

ACP/PCA



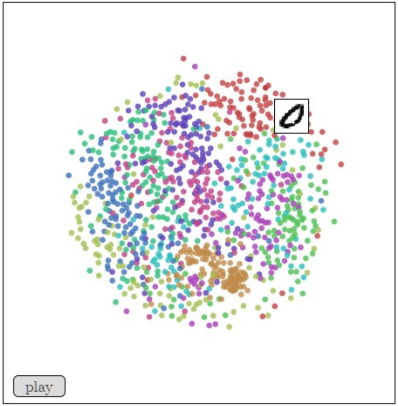
$\Rightarrow$ Pas de miracle... Mais pas si mal !



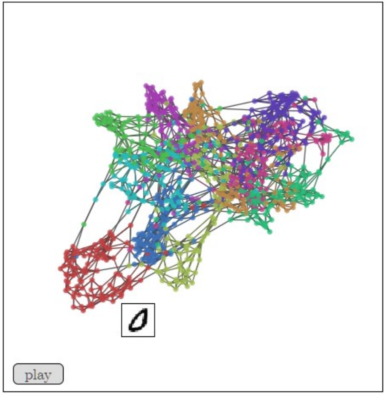
Sur USPS

Que se passe-t-il sur des données USPS ou MNIST ? 256/384 dimensions $\Rightarrow$ 2D !

Projection non linéaire



Visualizing MNIST with MDS



Visualizing MNIST as a Graph

Que se passe-t-il sur des données USPS ou MNIST ? 256/384 dimensions ⇒

T-SNE

A t-SNE plot of MNIST