

MACHINE LEARNING & VISUALISATION DE DONNÉES

Vincent Guigue
vincent.guigue@agroparistech.fr

VISUALISATION



InfoVis vs DataVis

■ 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...

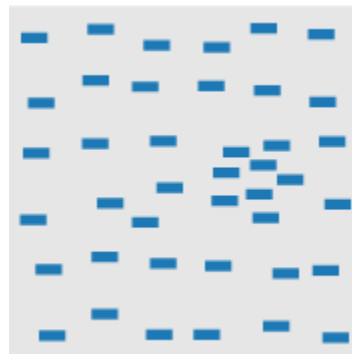
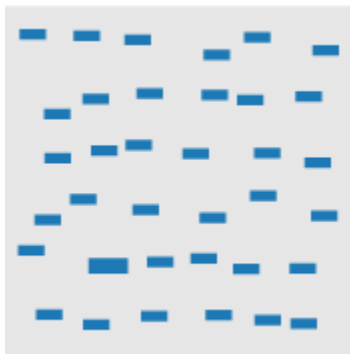
Référence utile : Cours de F. Rossi

<http://apiacoa.org/teaching/visualization/index.fr.html>

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

Humain = machine visuelle très perfectionnée

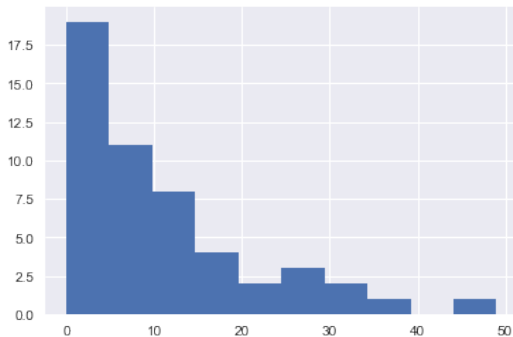
- Extraction de caractéristiques de base en 200ms
- Possibilités d'analyse de densité / détection d'anomalie très rapide



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

Visualiser une population (1D)

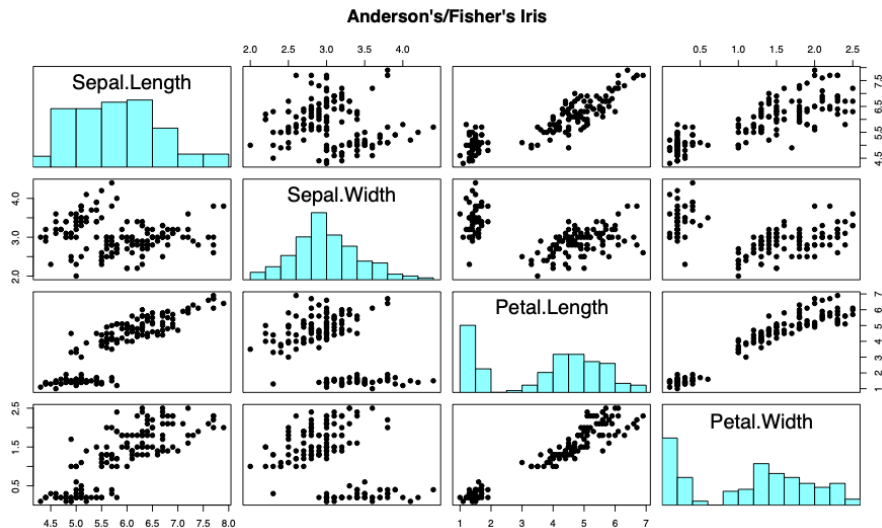
- Focus sur une dimension X_j
 - N Observations x_{ij}
- Solution pour la visualisation du contenu : l'histogramme





Visualiser une population (2D)

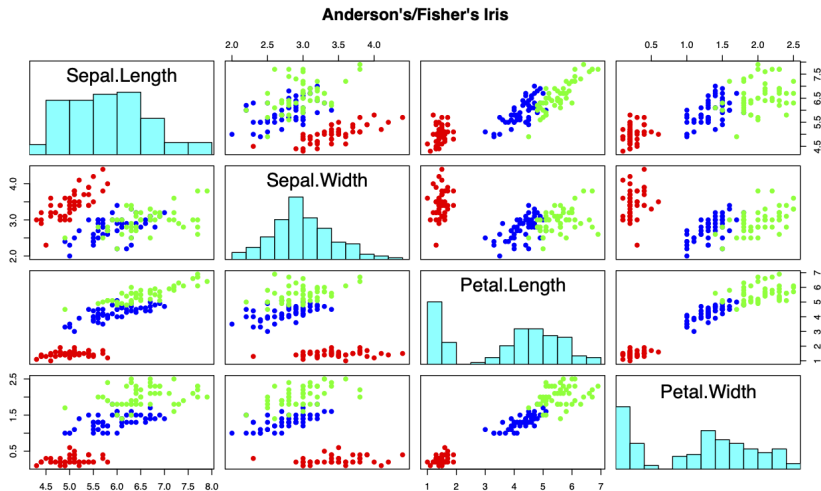
Données originales = Iris, 4D : comment visualiser ? \Rightarrow Scatter plot





Visualiser une population (2D)

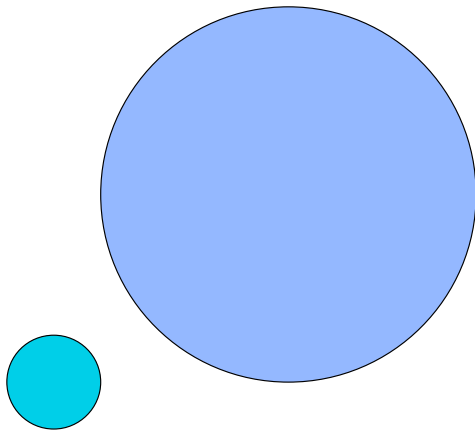
Données originales = Iris, 4D : comment visualiser ? \Rightarrow Scatter plot



Avec les informations de classes



Limites humaines



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



Limites humaines



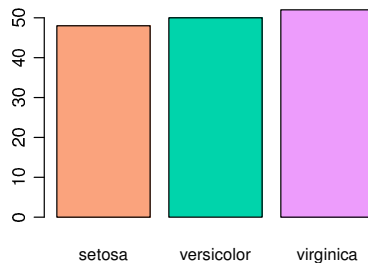
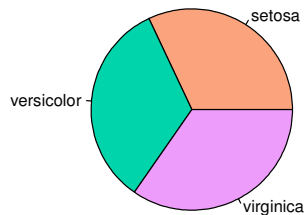
Please write down your estimation of the ratio of the lengths of those bars.

Limites humaines

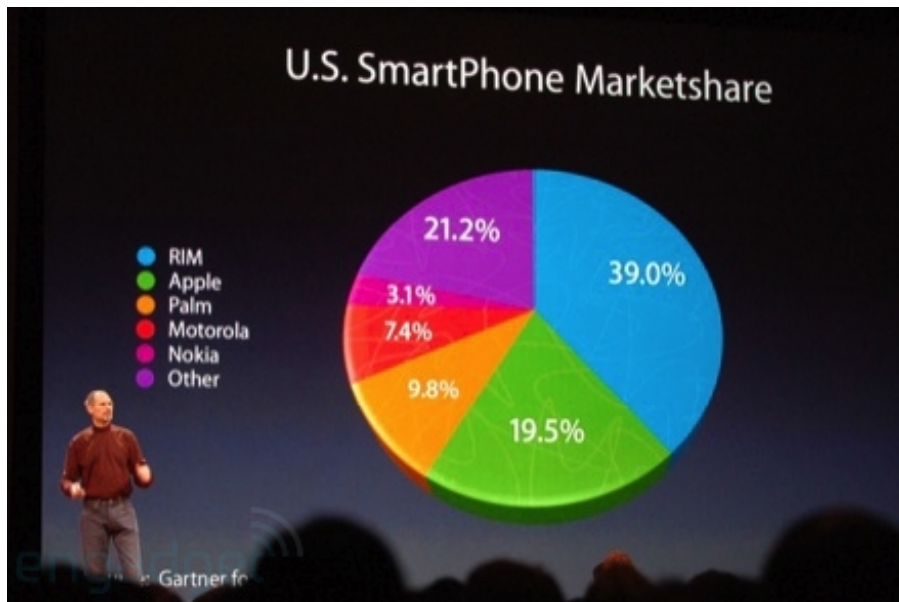
Another visual abstraction

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

And the views are

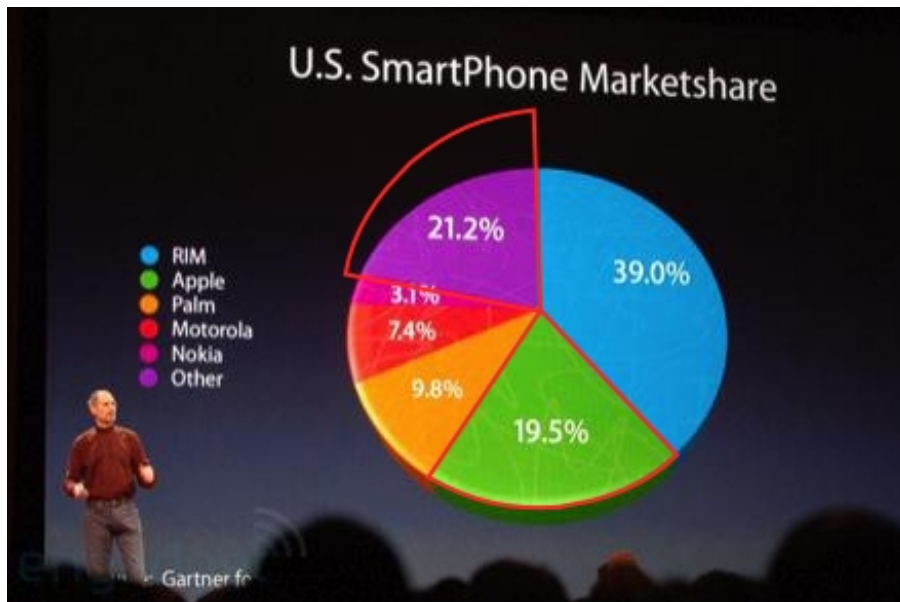


Limites humaines



Steve Jobs' keynote at Macworld 2008, source:

Limites humaines



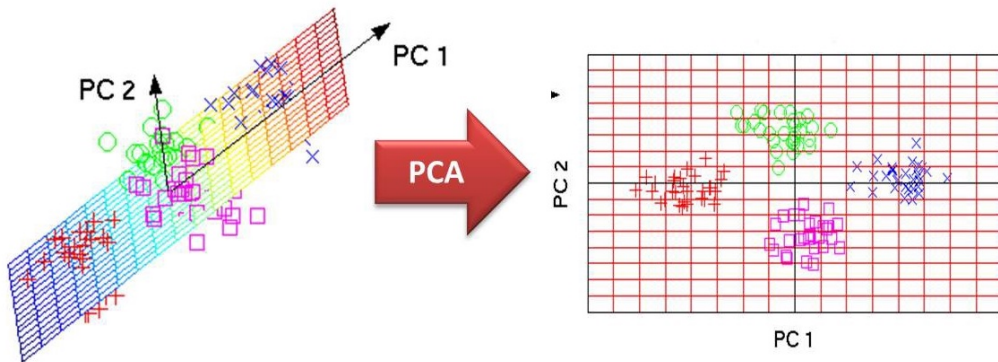
Steve Jobs' keynote at Macworld 2008, source:

TRANSFORMATIONS AVANCÉES

ACP : analyse en composantes principales

ACP (PCA) = outil de base pour

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



ACP : analyse en composantes principales

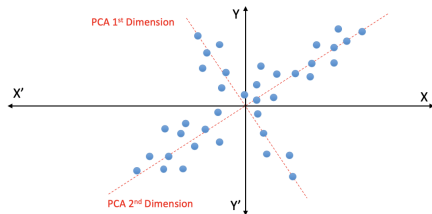
ACP (PCA) = outil de base pour

Idée : trouver des axes qui maximise la variance
⇒ projeter sur ces axes

- Transformation non supervisée
- Transformation applicable sur de nouveaux points

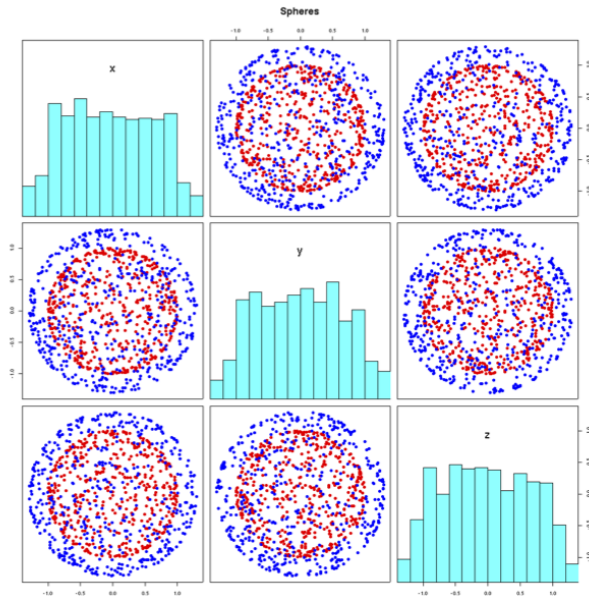
- 1 $X \in \mathbb{R}^{N \times d}$
- 2 ACP sur $X^T X \in \mathbb{R}^{d \times d}$
- 3 Récupération de $\{V_i \in \mathbb{R}^d, \lambda_i \in \mathbb{R}_+\}_{i=1,\dots,d}$
- 4 d Axes de projection $V_i \dots$ associés à leur force d'explication λ_i
- 5 Utilisation des V_i sur les données de test

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



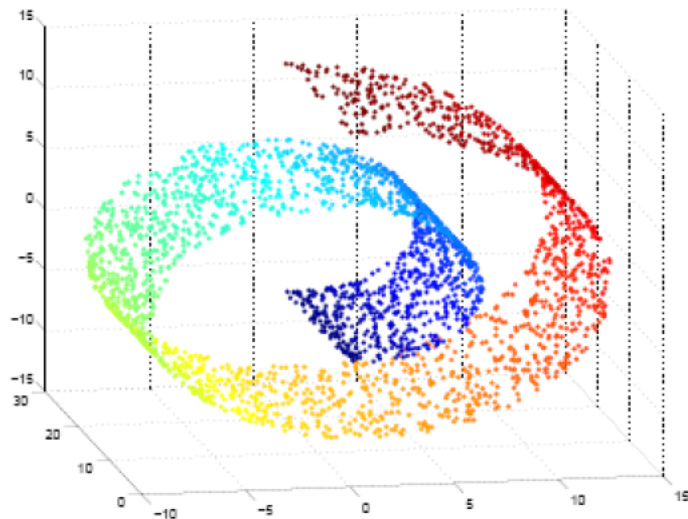


Limite de la sélection de variables



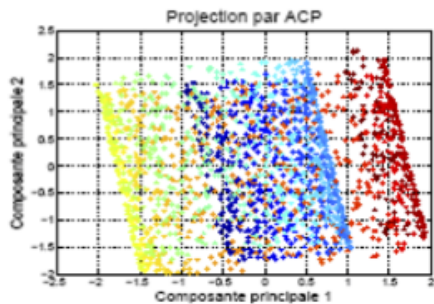
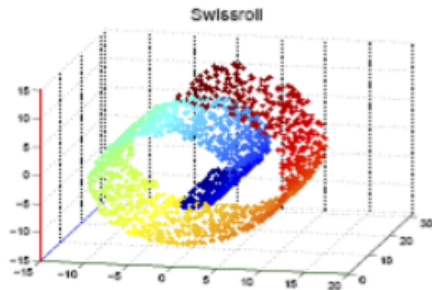


Limite de la sélection de variables



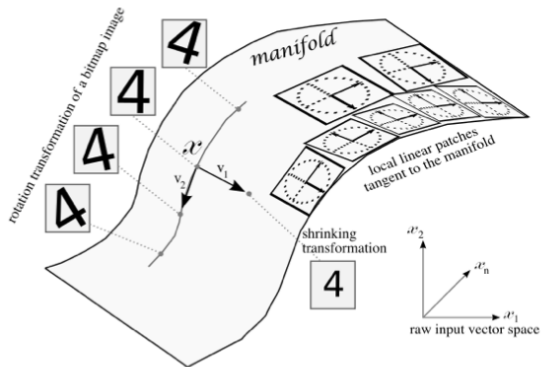


LLE : local linear embedding



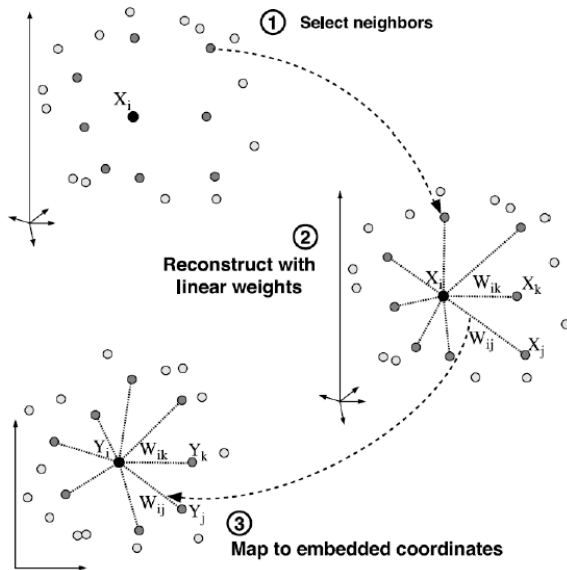
LLE : local linear embedding

Idée : Les données sont organisées selon une variété





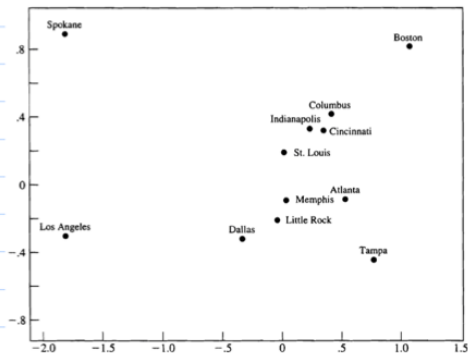
LLE : local linear embedding





MDS : multi-dimensional scaling

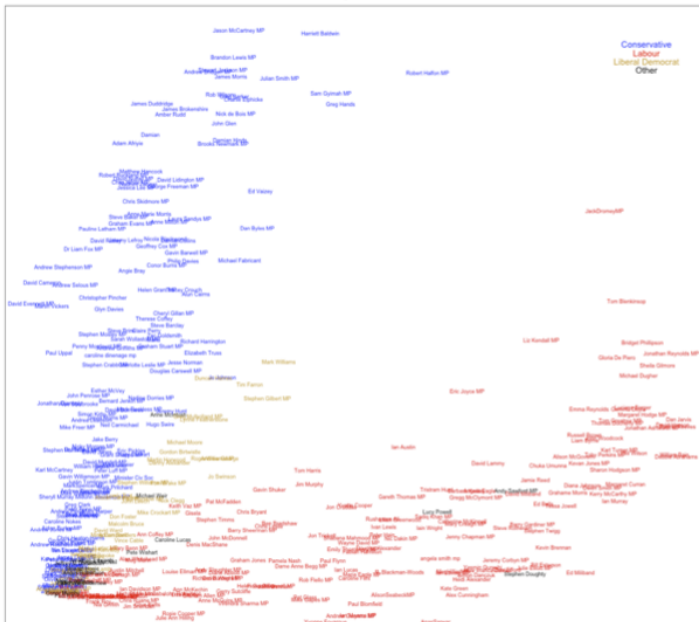
	Atlanta (1)	Boston (2)	Cincinnati (3)	Columbus (4)	Dallas (5)	Indianapolis (6)	Little Rock (7)	Los Angeles (8)	Memphis (9)	St. Louis (10)	Spokane (11)	Tampa (12)
(1)	0											
(2)	1068	0										
(3)	461	867	0									
(4)	549	769	107	0								
(5)	805	1819	943	1050	0							
(6)	508	941	108	172	882	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

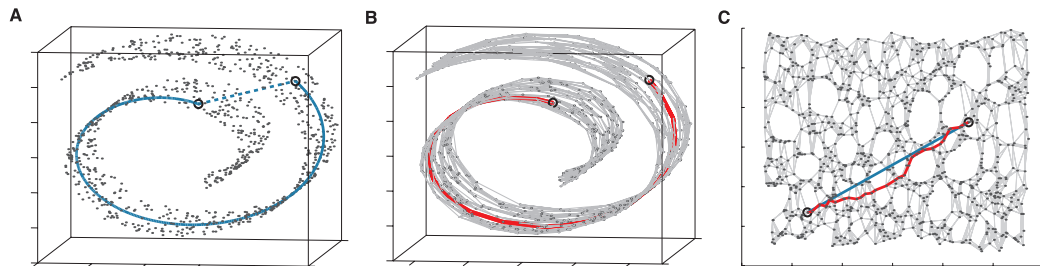




MDS : multi-dimensional scaling

Two dimensional clustering of UK Members of Parliament





- Se déplacer dans le graphe des plus proches voisins
- Reconstruire les distances dans le graphe (plutôt que dans l'espace d'origine)



ISOMAP

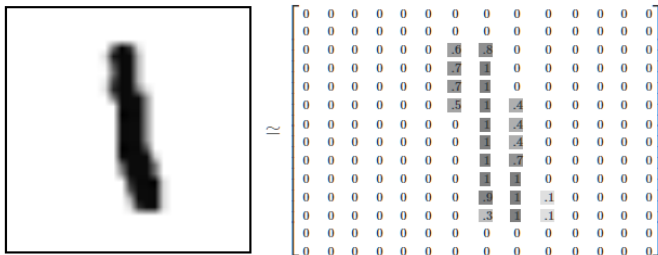
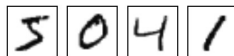
Step

- | | | |
|---|--------------------------------------|--|
| 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 $k = 1, 2, \dots, N$ 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 $\sqrt{\lambda_p} v_p^i$. |

Sur USPS

Que se passe-t-il sur des données USPS ou MNIST ?

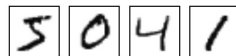
256/384 dimensions \Rightarrow 2D !



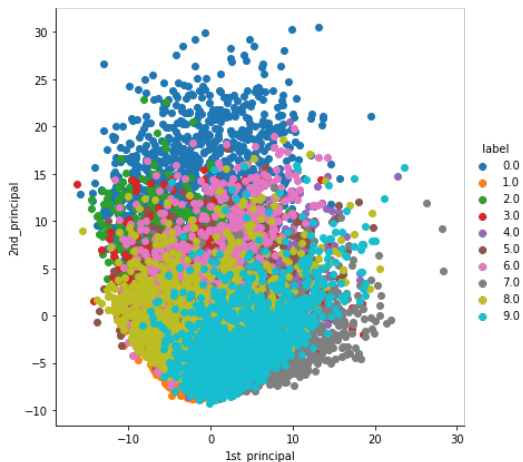
Sur USPS

Que se passe-t-il sur des données USPS ou MNIST ?

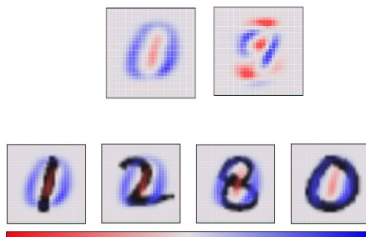
256/384 dimensions \Rightarrow 2D !



ACP/PCA



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



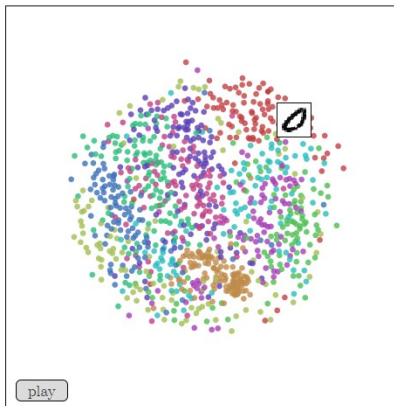
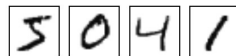


Sur USPS

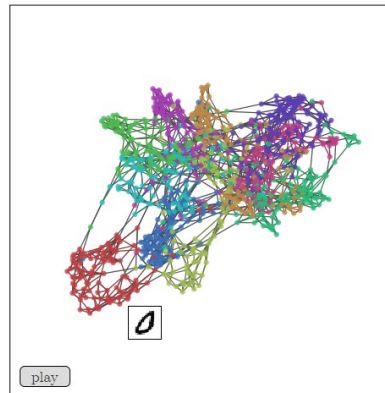
Que se passe-t-il sur des données USPS ou MNIST ?

256/384 dimensions \Rightarrow 2D !

Projection non linéaire



Visualizing MNIST with MDS



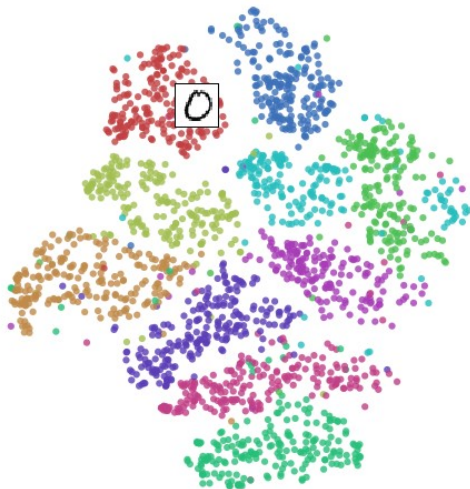
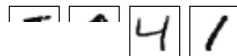
Visualizing MNIST as a Graph

Sur USPS

Que se passe-t-il sur des données USPS ou MNIST ?

256/384 dim

T-SNE



A t-SNE plot of MNIST



Conclusion

- Un outil pour comprendre les données
 - Identifier les classes avec plusieurs modes
 - Les points aberrants
 - Anticiper les difficultés (ou les facilités)
- Un outil pour analyser les erreurs des modèles... Puis améliorer les modèles
 - Présenter les résultats
 - Comprendre les erreurs