PARIS DESCARTES

Visualisation

M2-INFO

Nicoleta ROGOVSCHI

nicoleta.rogovschi@parisdescartes.fr

Plan du cours

- Définition et objectifs
- Méthodes de visualisation des données
- Exemple de cas d'applications
- Le fléau de la dimension
- Classification de données à haute dimension
- Techniques des réduction des dimensions

Pourquoi Visualiser les Données ?

- Meilleur présentation des données =>

  Meilleur Compréhension / Analyse

- “Goal ....is to communicate information clearly and effectively through graphical means.”
  - Friedman(2008)

Motivation

- Augmentation de la puissance de calcul des ordinateurs
- Disponibilité des données
- Beaucoup de données à haute dimension
- Une composante complémentaire au techniques de classification

Motivation

Une bonne technique de visualisation doit être applicable même si:

- On a très peu de connaissance à priori sur les données ou pas du tout
- Les buts exploratoires sont vagues
- Les données sont non homogènes et bruitées

Motivation

- De nos jours les données de grande taille sont partout

- Associées à différentes taches de ML telles que : la classification supervisée, le «clustering» et la regression

Résultats de la visualisation

Les résultats de la visualisations peuvent être présentés sous forme de :

- Cartes
- Graphiques
- Tableaux de bord (« Dashboarding »)

Pour quoi la réduction de la dimension

- Le nombre de caractéristiques peut-être très grand
  - Les données génomiques: expressions des genes
    - Plusieurs milliers de variables
  - Données de type image : chaque pixel d'une image
    - Une image 64X64 = 4096 caractéristiques
  - Catégorisation des textes : fréquences des phrases (ou des mots) dans un document ou une page web
    - Plus de dix mille caractéristiques

Méthodes de visualisation des données

Il existe de nombreuses méthodes de visualisation des données selon le type de données qu'on traite.

Les données peuvent être :

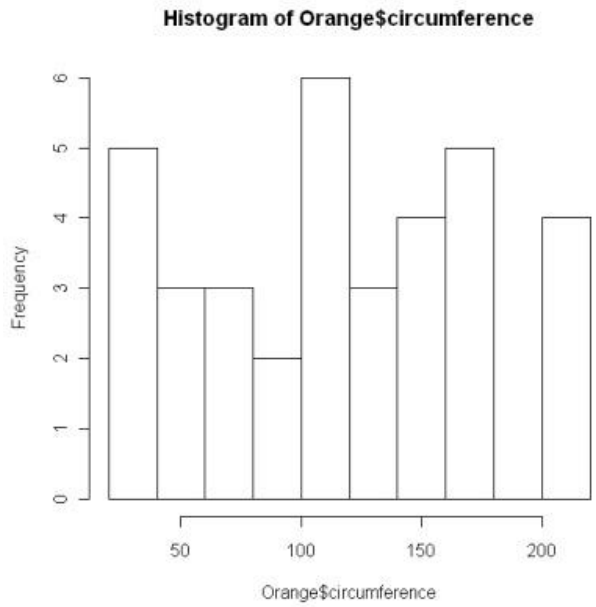
- Univariées
- Bivariées
- Multivariées

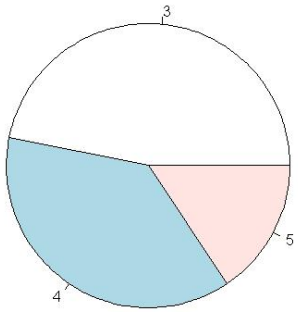
Données univariées

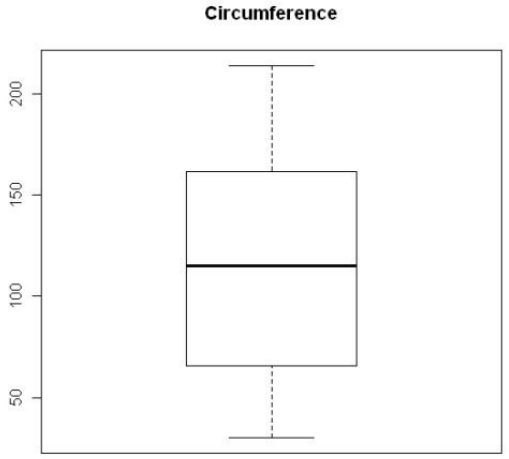
- Représente des mesure d'une variable
- D'habitude caractérisent une distribution
- Sont représentées selon les méthodes suivantes:
  - Histogramme
  - Camambert (Pie chart)
  - Boîte à moustache

Données univariées (1D)

- Représentation





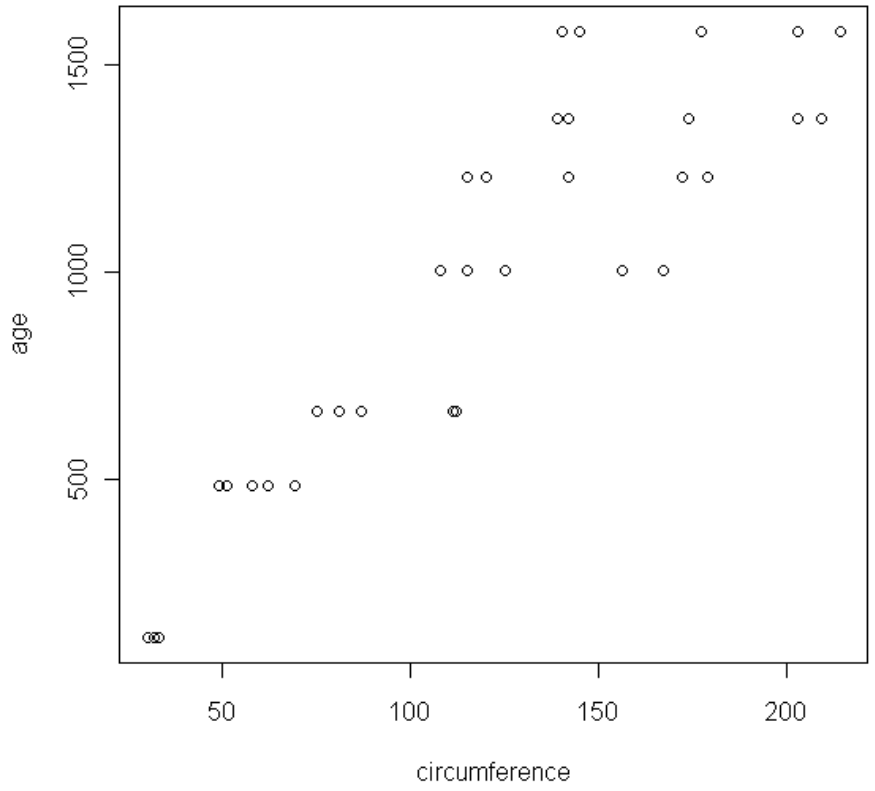


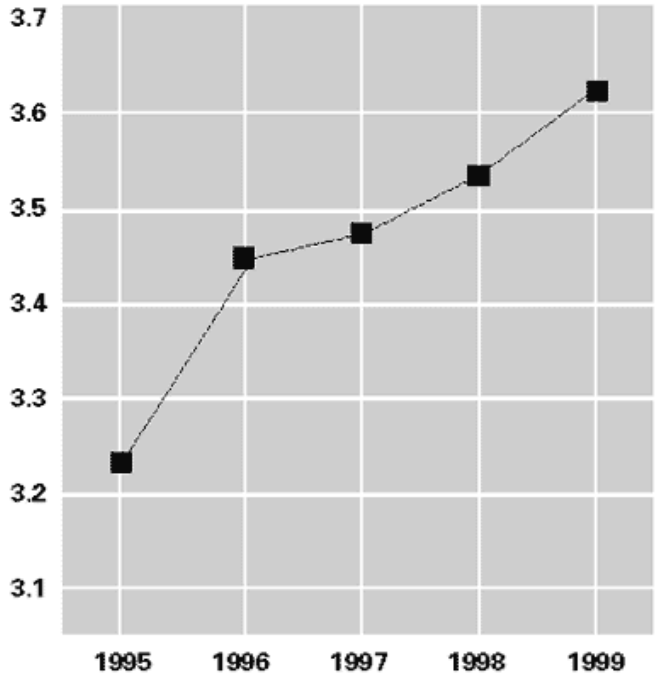
Données bivariées (2D)

- Constitue des échantillons appariés de deux variables
- Les variables sont liées
- Sont représentées selon les méthodes suivantes:
  - Scatter plot
  - Des graphiques linéaires

Données bivariées (2D)

- Représentation





Données multivariées

- Une représentation multidimensionelle de données multivariées

- Sont représentées selon les méthodes suivantes:
  - Méthodes à base d'icône
  - Méthodes à base de pixels
  - Système dynamique en coordonnées parallèles

Qu’est-ce que la visualisation ?

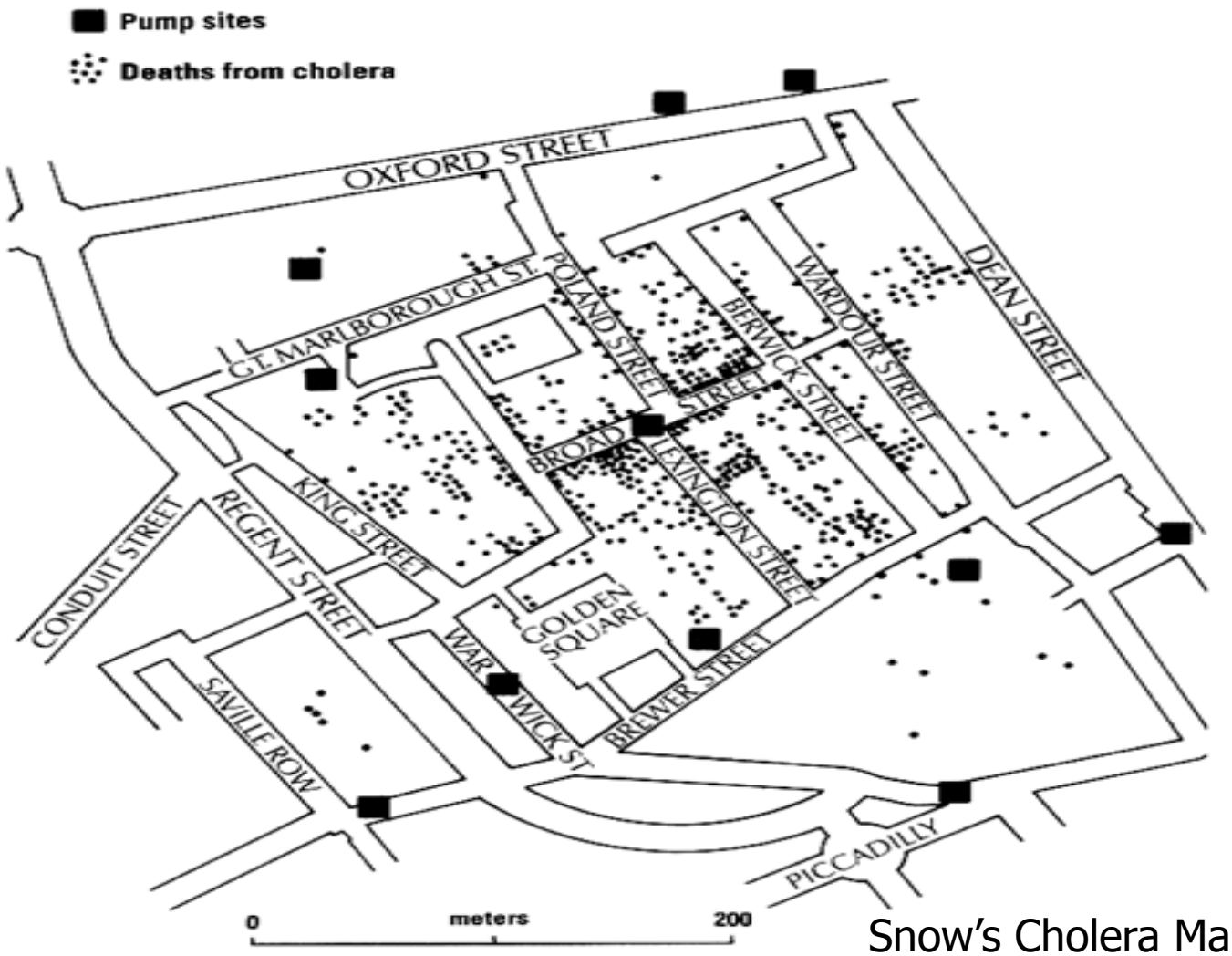
- La visualisation est le processus d’**interprétation visuelle** ou **de représentation graphique** d’un ensemble de données.

Qu’est-ce que la visualisation ?

- Enquêtes médicales sur les patients

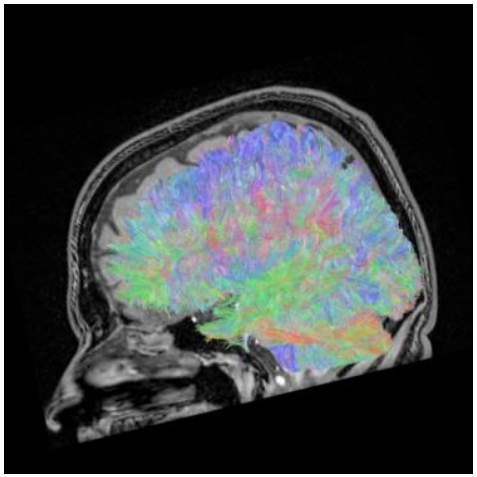
- 119 OE - pulse rate 26/07/2010 62 119 Diastolic blood pressure 26/07/2010 80 119 Systolic blood pressure 26/07/2010 120 201 Free T4 level 26/07/2010 17.8 201 Serum TSH level 26/07/2010 4.55 201 Serum calcium 26/07/2010 2.37 201 Serum inorganic phosphate 26/07/2010 1.17 201 Serum total protein 26/07/2010 64 201 Serum albumin 26/07/2010 41 201 Serum globulin 26/07/2010 23 201 Serum bilirubin level 26/07/2010 12 201 Serum alkaline phosphatase 26/07/2010 237 201 ALTSGPT serum level 26/07/2010 3 201 AST - aspartate transamSGOT 26/07/2010 24 201 International normalised ratio 26/07/2010 2 201 International normalised ratio 26/07/2010 2 580 Prostate specific antigen 26/07/2010 0.7 631 OE - pulse rate 26/07/2010 83 631 Diastolic blood pressure 26/07/2010 80 631 Systolic blood pressure 26/07/2010 133 634 Urine creatinine 26/07/2010 11.74 634 Urine microalbumin 26/07/2010 8.8 634 Urine albumin:creatinine ratio 26/07/2010 0.75 634 OE - weight 26/07/2010 84.35 634 OE - height 26/07/2010 176 634 Body Mass Index 26/07/2010 27.23 634 Diastolic blood pressure 26/07/2010 80 634 Systolic blood pressure 26/07/2010 135 786 Free T4 level 26/07/2010 12.6 786 Serum TSH level 26/07/2010 2.58 786 Plasma C reactive protein 26/07/2010 32 786 Serum calcium 26/07/2010 2.48 786 Serum inorganic phosphate 26/07/2010 0.93 786 Serum total protein 26/07/2010 71 786 Serum albumin 26/07/2010 41 786 Serum globulin 26/07/2010 30 786 Serum bilirubin level 26/07/2010 6 786 Serum alkaline phosphatase 26/07/2010 233 786 ALTSGPT serum level 26/07/2010 9 786 AST - aspartate transamSGOT 26/07/2010 18 786 Serum sodium 26/07/2010 137 786 Serum potassium 26/07/2010 4.1 786 Serum chloride 26/07/2010 103 786 Serum urea level 26/07/2010 4.7 786 Serum creatinine 26/07/2010 83 786 Total white blood count 26/07/2010 8.8 786 Red blood cell RBC count 26/07/2010 4.32 786 Haemoglobin estimation 26/07/2010 13 786 Haematocrit - PCV 26/07/2010 38.2 786 Mean corpuscular volume MCV 26/07/2010 88.4 786 Mean corpusc haemoglobinMCH 26/07/2010 30 786 Mean corpusc Hb conc MCHC 26/07/2010 34 786 Platelet count 26/07/2010 364 786 Neutrophil count 26/07/2010 6.6 786 Lymphocyte count 26/07/2010 1.5 786 Monocyte count 26/07/2010 0.5 786 Eosinophil count 26/07/2010 0.1 786 Erythrocyte sedimentation rate 26/07/2010 21 816 Serum vitamin B12 26/07/2010 348 816 Serum folate 26/07/2010 3.8 816 Total white blood count 26/07/2010 3.9 816 Red blood cell RBC count 26/07/2010 3.83 816 Haemoglobin estimation 26/07/2010 13.3 816 Haematocrit - PCV 26/07/2010 39.4 816 Mean corpuscular volume MCV 26/07/2010 103 816 Mean corpusc haemoglobinMCH 26/07/2010 34.8 816 Mean corpusc Hb conc MCHC 26/07/2010 33.8 816 Platelet count 26/07/2010 137 816 Neutrophil count 26/07/2010 2.4 816 Lymphocyte count 26/07/2010 1 816 Monocyte count 26/07/2010 0.4 816 Eosinophil count 26/07/2010 0.1 816 Body Mass Index 26/07/2010 28.99 816 OE - weight 26/07/2010 92.05 816 OE - height 26/07/2010 178.2 816 OE - pulse rate 26/07/2010 52 816 Diastolic blood pressure 26/07/2010 54 816 Systolic blood pressure 26/07/2010 107 856 Free T4 level 26/07/2010 16.4 856 Serum TSH level 26/07/2010 2.95 856 Serum calcium 26/07/2010 2.6 856 Serum inorganic phosphate 26/07/2010 0.82 856 Serum total protein 26/07/2010 72 856 Serum albumin 26/07/2010 47 856 Serum globulin 26/07/2010 25 856 Serum bilirubin level 26/07/2010 15 856 Serum alkaline phosphatase 26/07/2010 176 856 ALTSGPT serum level 26/07/2010 33 856 AST - aspartate transamSGOT 26/07/2010 23 856 Serum sodium 26/07/2010 141 856 Serum potassium 26/07/2010 4.8 856 Serum chloride 26/07/2010 102 856 Serum urea level 26/07/2010 6.3 856 Serum creatinine 26/07/2010 98 856 Total white blood count 26/07/2010 5.8 856 Red blood cell RBC count 26/07/2010 5.04 856 Haemoglobin estimation 26/07/2010 16 856 Haematocrit - PCV 26/07/2010 47.1 856 Mean corpuscular volume MCV 26/07/2010 93.4 856 Mean corpusc haemoglobinMCH 26/07/2010 31.9 856 Mean corpusc Hb conc MCHC 26/07/2010 34.1 856 Platelet count 26/07/2010 162 856 Neutrophil count 26/07/2010 3.2 856 Lymphocyte count 26/07/2010 2.1 856 Monocyte count 26/07/2010 0.3 856 Eosinophil count 26/07/2010 0.2 856 Erythrocyte sedimentation rate 26/07/2010 5 856 Diastolic blood pressure 26/07/2010 90 856 Systolic blood pressure 26/07/2010 150 1005 International normalised ratio 26/07/2010 3 1163 Serum sodium 26/07/2010 141 1163 Serum potassium 26/07/2010 5 1163 Serum chloride 26/07/2010 98 1163 Serum urea level 26/07/2010 16.8 1163 Serum creatinine 26/07/2010 159 1163 Glomerular filtration rate 26/07/2010 28 1818 Free T4 level 26/07/2010 13.6 1818 Serum TSH level 26/07/2010 1.43 1818 Total white blood count 26/07/2010 5.8 1818 Red blood cell RBC count 26/07/2010 4.65 1818 Haemoglobin estimation 26/07/2010 14.8 1818 Haematocrit - PCV 26/07/2010 43.1 1818 Mean corpuscular volume MCV 26/07/2010 92.6 1818 Mean corpusc haemoglobinMCH 26/07/2010 31.9 1818 Mean corpusc Hb conc MCHC 26/07/2010 34.5 1818 Platelet count 26/07/2010 205 1818 Neutrophil count 26/07/2010 4 1818 Lymphocyte count 26/07/2010 1.4 1818 Monocyte count 26/07/2010 0.3 1818 Eosinophil count 26/07/2010 0.1 1818 Erythrocyte sedimentation rate 26/07/2010 15 1818 Blood glucose level 26/07/2010 5.4 1818 Serum calcium 26/07/2010 2.48 1818 Serum inorganic phosphate 26/07/2010 0.89 1818 Serum total protein 26/07/2010 76 1818 Serum albumin 26/07/2010 43 1818 Serum globulin 26/07/2010 33 1818 Serum bilirubin level 26/07/2010 11 1818 Serum alkaline phosphatase 26/07/2010 287 1818 ALTSGPT serum level 26/07/2010 24 1818 AST - aspartate transamSGOT 26/07/2010 25 1818 Serum sodium 26/07/2010 138 1818 Serum potassium 26/07/2010 4.2 1818 Serum chloride 26/07/2010 105 1818 Serum urea level 26/07/2010 3.6 1818 Serum creatinine 26/07/2010 83 2714 Serum sodium 26/07/2010 134 2714 Serum potassium 26/07/2010 4.4 2714 Serum chloride 26/07/2010 104 2714 Serum urea level 26/07/2010 9.8 2714 Serum creatinine 26/07/2010 200 2714 Glomerular filtration rate 26/07/2010 29 3459 International normalised ratio 26/07/2010 2.5 3735 OE - pulse rate 26/07/2010 72 3735 Diastolic blood pressure 26/07/2010 96 3735 Systolic blood pressure 26/07/2010 169 4219 OE - pulse rate 26/07/2010 70 4219 Diastolic blood pressure 26/07/2010 71 4219 Systolic blood pressure 26/07/2010 108 4219 International normalised ratio 26/07/2010 3.3 4285 OE - pulse rate 26/07/2010 64 4285 Diastolic blood pressure 26/07/2010 90 4285 Systolic blood pressure 26/07/2010 150 4355 Diastolic blood pressure 26/07/2010 80 4355 Systolic blood pressure 26/07/2010 120 4511 International normalised ratio 26/07/2010 2.5 4763 International normalised ratio 26/07/2010 2.9 5111 International normalised ratio 26/07/2010 1.5

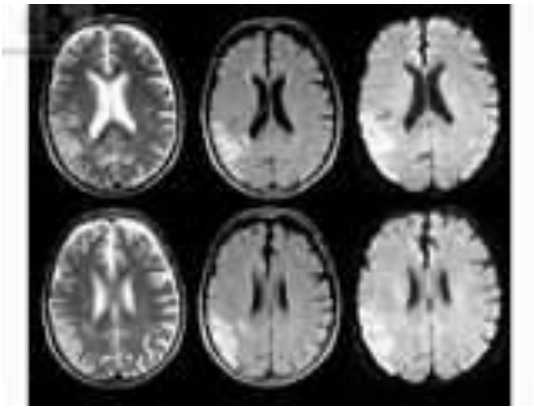
Qu’est-ce que la visualisation ?

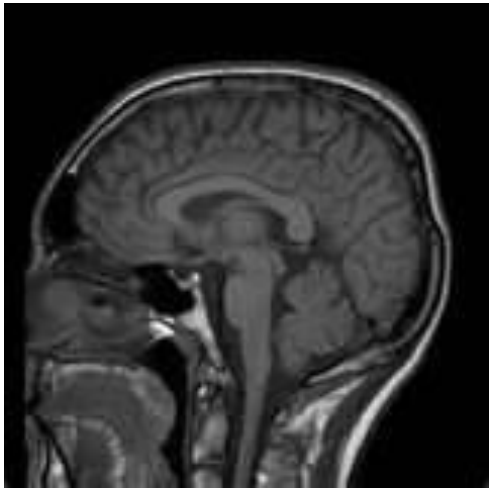


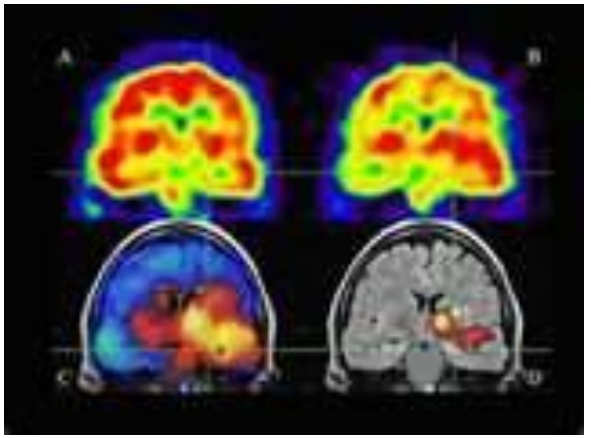
Snow’s Cholera Map, 1855

Qu’est-ce que la visualisation ?

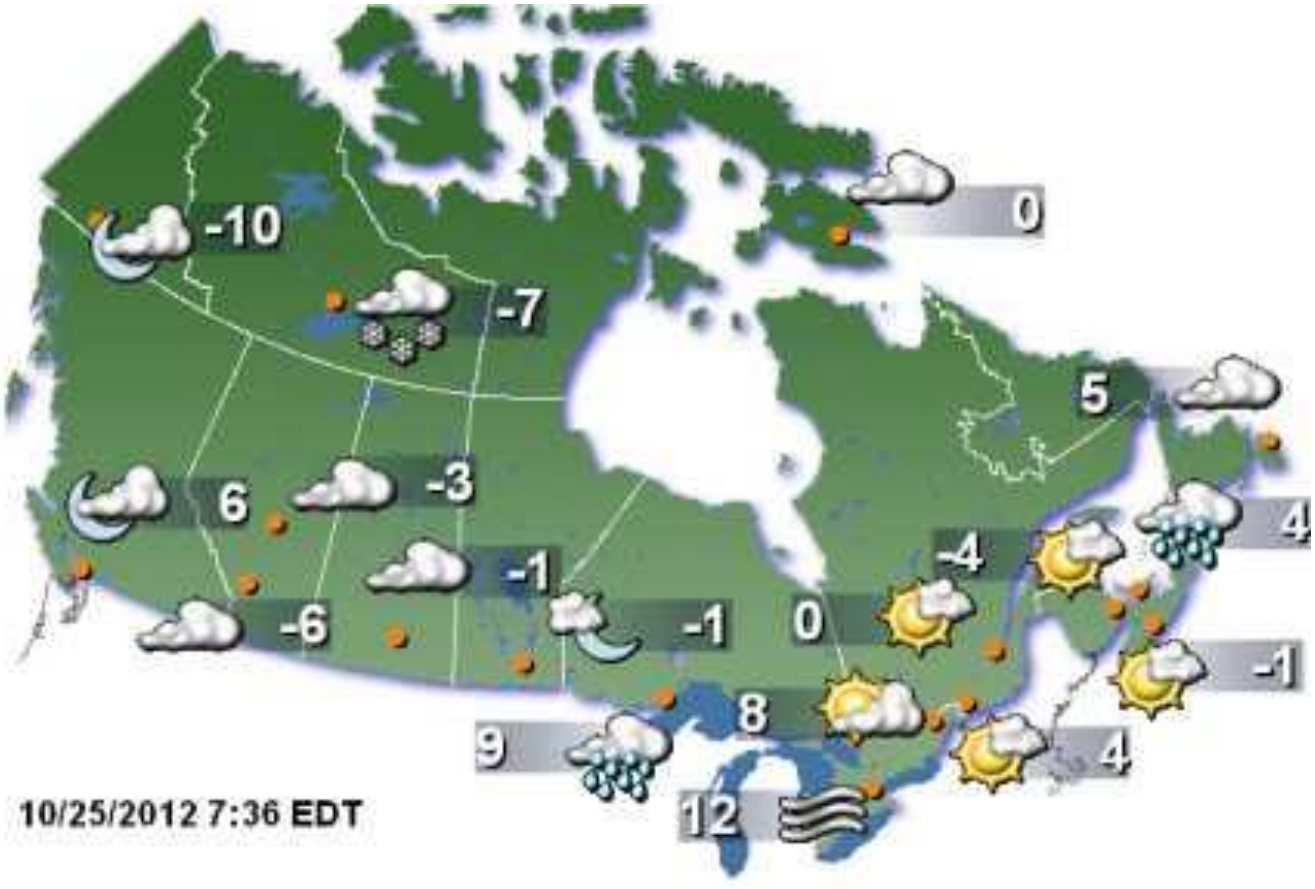








Qu’est-ce que la visualisation ?



Qu’est-ce que la visualisation ?

Qu'est-ce que la visualisation ?

- La visualisation est une branche de l'informatique regroupant le **traitement, l'analyse et la représentation graphique de données** provenant de divers domaines : les sciences sociales, les finances, la médecine, le divertissement, etc.
- Il y deux champs de compétence qui sont particulièrement sollicités en visualisation : l'**infographie et les statistiques.**

Qu’est-ce que la visualisation ?

- Il est important de savoir distinguer les domaines du traitement d’image, de l’infographie et de la visualisation.
- Le **traitement d’image est l’étude des images 2D pour en extraire de** l’information ou pour en modifier les caractéristiques.

Qu'est-ce que la visualisation ?

- L'**infographie permet de créer des images de toute pièce à l'aide d'un** ordinateur, qu'il s'agisse d'images 2D dessinées par un artiste ou de complexes scènes 3D.

- La **visualisation quant à elle permet l'exploration** de données représentées sous une forme visuelle afin d'aider notre compréhension du phénomène illustré.

Qu’est-ce que la visualisation ?

- Un des objectif de la visualisation est de représenter visuellement des données qui ne possèdent pas nécessairement d’**interprétation géométrique naturelle.**

Acquisition des données

L'**acquisition des données brutes peut être faite de différentes** façons :

- par des simulations (calculs informatiques)
- des enquêtes statistiques
- des bases de données historiques
- des capteurs de mesures réelles, etc.

Acquisition des données

Sources d'erreurs

- L'échantillonnage est-il assez précis pour qu'on soit en mesure d'obtenir l'information souhaitée ? Inversement, est-il trop fin ? Il ne faut pas considérer des données inutiles qui ne feraient qu'alourdir les calculs.
- Est-ce que la quantification se fait avec assez de précision pour être en mesure de faire ressortir les caractéristiques souhaitées ?

Filtrage des données

Sources d'erreurs

- Conserve-t-on les données importantes et significatives ? Au contraire, élimine-t-on les données non pertinentes à l'extraction des caractéristiques souhaitées ?
- Si on rajoute des données, les données ajoutées sont elles représentatives du reste ?

Données à haute dimension

- L'augmentation de la puissance de calcul et de l'espace de stockage des ordinateurs a mené à une augmentation de la taille des ensembles de données.
- Ainsi, plusieurs champs des sciences reposent maintenant sur notre capacité à analyser et visualiser des données à haute dimension.

Le fléau de la dimension

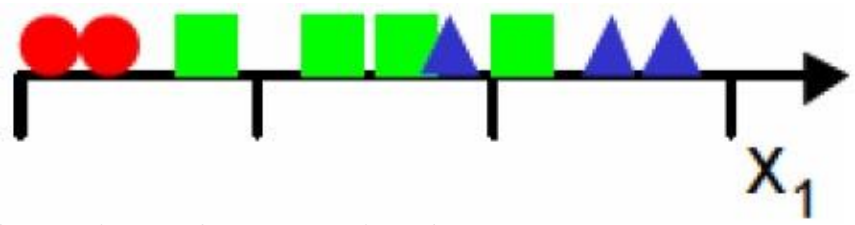
Le fléau de la dimension

Le fléau de la dimension (« Curse of dimensionality »)

- Un terme inventé en 1961 par Bellman
- Se réfère au problème de l'augmentation explosive du volume de données associée à l'ajout de dimensions supplémentaires dans un espace mathématique
- On va illustrer ce problème avec un simple exemple

Exemple jouet

- Problème de reconnaissance des formes à 3 classes
- On dispose de 9 observations1D (le long d'un axe)



Une approche simple serait de

- Diviser l'espace des variables dans des cases uniformes
- Calculer le taux des exemples pour chaque classe de chaque case et
- Pour un nouvel exemple, trouver sa case et choisir la classe prédominante de la case

Exemple jouet

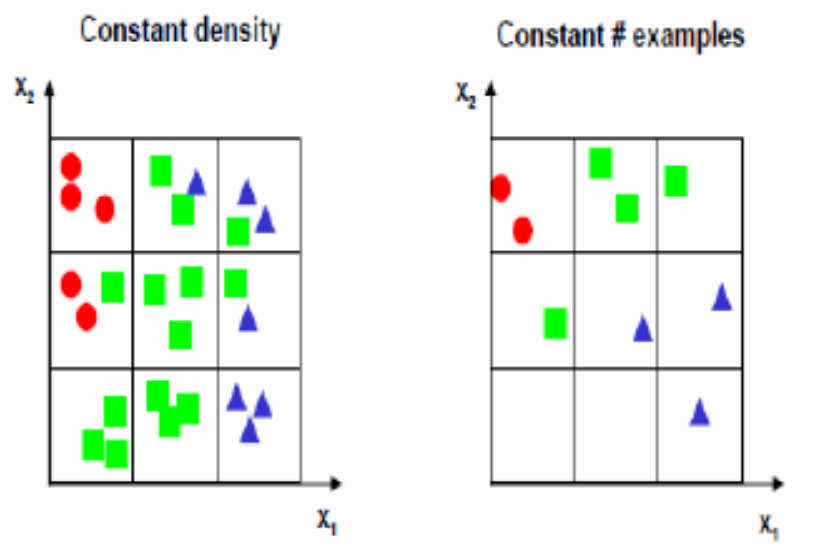
- Pour notre exemple jouet on commence avec une seule dimension et on divise l'axe en 3 segments. On constate qu'on a une moyenne de 3 exemples par région.
- Par la suite, on observe qu' il existe trop d'intersections entre les classes, on a décidé donc de rajouter une deuxième dimension pour essayer d'améliorer la séparabilité des classes.

Exemple jouet (2D)

- Si on rajoute une $2^{ième}$ dimension on passe de 3 cases (en 1D) à $3^2=9$ (en 2D).
- Il y a un autre problème qui se pose : On maintient la densité des exemples par case ou on garde le même nombre d'exemple qu'on a utilisé en 1D?

Exemple jouet (2D)

- Si on choisit de maintenir la densité, on augmente le nombre d'exemple de 9 (en 1D) à 27 (en 2D)
- Si on choisit de maintenir le nombre d'exemple, la projection des données en 2D est très « sparse »



Exemple jouet (3D)

Si on rajoute une 3^{ième} dimension les choses s'empirent

- Le nombre des cases augmente

jusqu'à $3^3=27$

- Pour garder la même densité

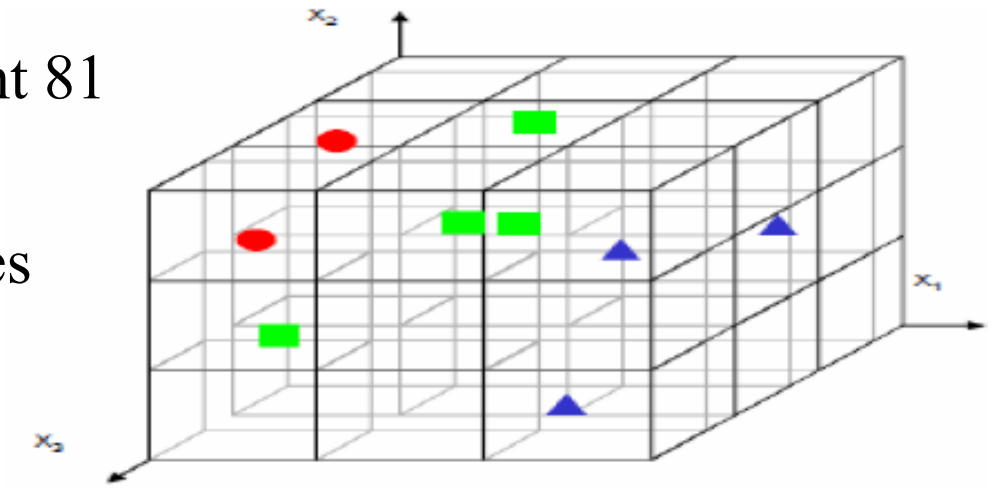
d'exemples le nombre

d'observations exigées devient 81

- Pour le même nombre

d'exemples, les cases obtenues

sont presque vides



Le fléau de la dimension

L'approche réalisé sur l'exemple jouet est inefficace

- Il y a d'autres approches moins influencées par le fléau de la dimension, mais le problème existe toujours

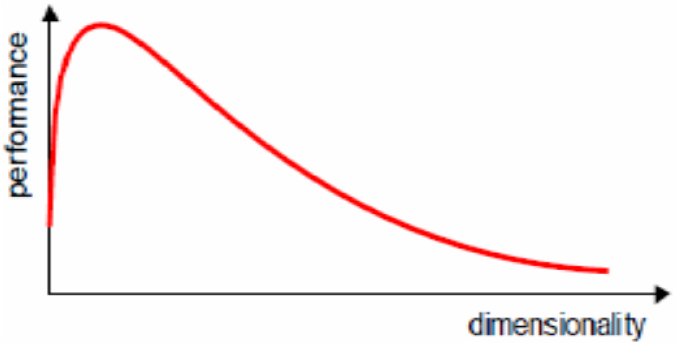
Le fléau de la dimension

Comment peut-on traiter le «fléau de la dimension»?

- En rajoutant à priori de l'information
- En réalisant un lissage plus ample de la fonction cible
- En réduisant la dimension des données

Le fléau de la dimension

- En pratique le fléau de la dimension signifie que pour une taille donnée de l'échantillon, il existe un nombre maximal de variables au delà duquel la performance du notre classifieur va dégrader plutôt qu'augmenter.



Le fléau de la dimension

Le fléau de la dimension génère plusieurs phénomènes dont :

- La concentration de la mesure
- La désertification de l'espace
- La décroissance du volume de la boule unité
- Dépeuplement du centre des hyper-volumes

Conséquences

Le fléau de la dimension a plusieurs conséquences:

1. Une augmentation exponentielle du nombre d'exemples est nécessaire afin de maintenir une certaine densité de l'échantillon (Pour une densité de N observations/case et D dimensions, le nombre total d'exemples est N^D)

Conséquences

2. Une augmentation exponentielle de la complexité de la fonction cible (celle qui estime la densité). Pour faire un bon apprentissage, une fonction cible plus complexe nécessite un échantillon de points beaucoup plus dense

Conséquences

3. Pour une dimension on peut trouver dans la littérature une multitude de fonctions de densité, mais pour les grandes dimensions on a seulement la densité Gaussienne multivariée.

   De plus, pour des grandes valeurs de D, on peut traiter la densité Gaussienne seulement dans une forme simplifié.

Le fléau de la dimension

- Ces constats suggèrent qu'on a besoin d'un traitement spécial pour manipuler les données volumineuses, qui diffère de celui qu' on a pour les données de faible dimension
- On rencontre les mêmes problèmes dans d'autres types de distribution de données

Classification des données à haute dimension

Classification des données à haute dimension

Méthodes :

- **Subspace-clustering**: on cherche des clusteurs qui existent dans des sous-espaces de l'espace des données de départ
  - CLIQUE, ProClus et des approches de co-clustering

- **Techniques de réduction de la dimension** : Construire un espace de dimension beaucoup plus faible et chercher les clusters dans cette espace (on peut construire de nouvelles dimensions en combinant certaines dimensions dans les données d'origine)

Subspace-clustering

Méthodes de Subspace Clustering

- Méthodes de recherche des sous-espaces : On cherche différents sous-espaces pour trouver des clusters
  - Approches “Bottom-up”
  - Approches “Top-down”
- Méthodes de clustering basées sur la corrélation
  - Par exemple : les approches à base de l’ACP
- Méthodes de Co-clustering
  - Méthodes basées sur l’optimisation (Cheng and Church, ISMB’2000)
  - Méthodes d’énumération (Pei et al., ICDM’2003)

Méthodes de recherche des sous-espaces

- Méthodes de recherche des sous-espaces
- *Approches "Bottom-up"*
  - On commence par des petits sous-espaces et on cherche des sous-espaces plus grands seulement quand il peut y avoir des clusters dans des tels sous-espaces
  - Diverses techniques d'elagage pour reduire le nombre des sous-espaces plus grands qu'on cherche
  - Ex. CLIQUE (Agrawal et al. 1998)
- *Approches "Top-down"*
  - On commence par l'espace de départ et on cherche des sous espaces plus petits d'une manière recursive
  - Le sous-espace d'un cluster peut être déterminée par le voisinage local
  - Ex. PROCLUS (Aggarwal et al. 1999): une approche semblable à *k*-medoid

Méthodes basées sur la corrélation

- Méthodes de clustering basées sur la corrélation : Des approches à base de modèles de corrélation avancés
- Ex : approches à base d'ACP :
  - On applique l'ACP pour génerer un ensemble de nouvelles dimensions non-correlées
  - Ensuite on gère les clusters dans le nouvel espace ou ses sous-espaces

Méthodes de Co-clustering

- **Co-clustering** : On classifie les objets et les attributs d'une manière simultanée (on traite les objets et les attributs symmetriquement)
- Méthodes basées sur l'optimisation
  - On essaie de trouver une sous-matrice quand elle aquiert la meilleure signification en tant que bi-cluster
  - En raison du coût de calcul, une recherche de type « glouton » est utilisée pour trouver des bi-clusters locaux optimaux
- Méthodes d'énumération
  - On utilise un seuil de tolerance pour specifier le degrée de bruit permis dans les co-clusters qu'on va traiter
  - Ensuite on essaye d'énumerer toutes les sous-matrices commes des bi-clusters qui satisfons les contraintes posées

Techniques de réduction de la dimension

Réduction des dimensions

- Les données dans un espace de grande dimension ne sont pas distribuées uniformément
- La réduction des dimension est une technique assez utilisée pour traiter « le fléau de la dimension»

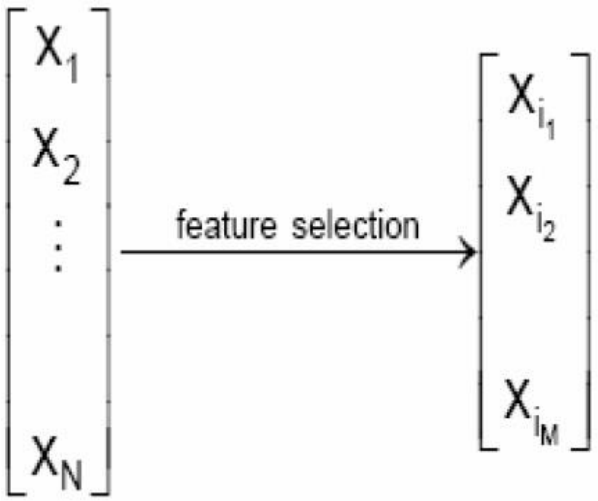
Réduction des dimensions

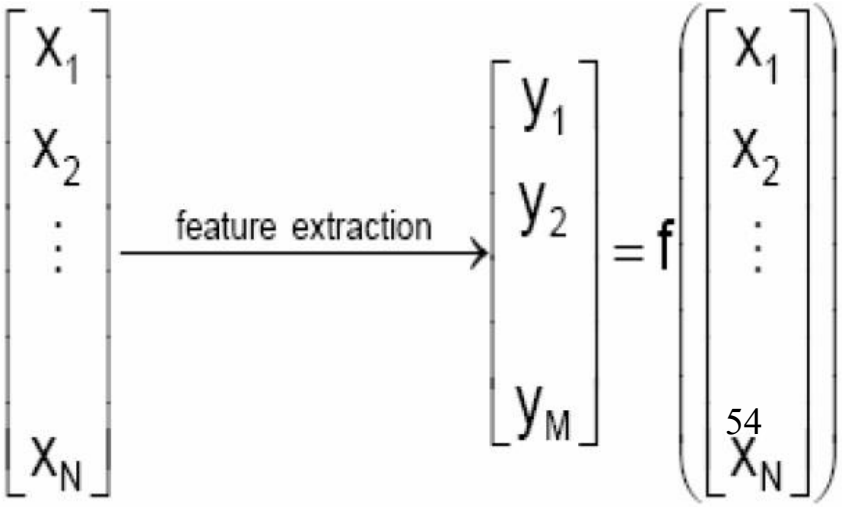
Il existe une multitude de techniques de réduction des dimensions :

- Linéaires vs non-linéaires
- Déterministes vs probabilistes
- Supervisées vs non supervisées

Réduction des dimensions

- Réduction des dimensions : Méthodologies
  - Sélection des variables (« feature selection ») : choisir un sous-ensemble de toutes les variables
  - Extraction des variables (« feature extraction ») : créer un sous-ensemble de nouvelles variables par combinaison





Réduction des dimensions par Sélection de variables

Sélection de variables

- Définition :

La sélection de variables est un procédé permettant de choisir un sous-ensemble optimal de variables pertinentes, à partir d'un ensemble de variables, selon un certain critère de performance.

On peut se poser trois questions essentielles :

Q1 : Comment mesurer la pertinence des variables ?

Q2 : Comment former le sous-ensemble optimal ?

Q3 : Quel critère d'optimalité utiliser ?

Sélection de variables

La réponse à Q1 consiste à trouver une mesure de pertinence ou un *critère d'évaluation* *J(X)* permettant de quantifier l'importance d'une variable ou d'un ensemble de variables .

Q2 évoque le problème du choix de la *procédure de recherche ou de constitution du sous-ensemble optimal* des variables pertinentes.

Q3 demande la définition d'un *critère d'arrêt* de la recherche

Sélection de variables

Critère d'évaluation

➢ Pour un problème de classement, on teste, par exemple, la qualité de discrimination du système en présence ou en absence d'une variable.

➢ Pour un problème de régression, on teste plutôt la qualité de prédiction par rapport aux autres variables.

Sélection de variables

Une alternative consiste à utiliser une méthode de recherche de type *Branch & Bound.*
Cette méthode de recherche permet de restreindre la recherche et donne le sous-ensemble optimal de variables, sous l'hypothèse de monotocité du critère de sélection *J(X)*.

Le critère est dit monotone si :

$$X_1 \subset X_2 \subset K \subset X_m \Rightarrow J(X_1) \leq J(X_2) \leq K \leq J(X_m)$$

où X_k est l'ensemble contenant k variables sélectionnées.

Sélection de variables

Problème : la plupart des critères d'évaluation ne sont pas monotones

Recours à des méthodes sous-optimales :

- *Sequential Forward Selection (SFS)*
- *Sequential Backward Selection (SBS)*
- *Bidirectional Selection (BS)*

Sélection de variables

- ***Sequential Forward Selection (SFS)***

Soit X l'ensemble des variables,

Au départ l'ensemble des variables sélectionnées est vide.

A chaque étape k :

- on sélectionne la variable X_i qui optimise le critère d'évaluation $J(X_k)$

$$J(X_k) = \max_{x_i \in (X - X_{k-1})} J\left(X_{k-1} \cup \{x_i\}\right)$$

✓ liste ordonnée des variables selon leur importance

Sélection de variables

- *Sequential Backward Selection (SBS)*

On part de l'ensemble complet des variables X et on procède par élimination :

à chaque étape :

- la variable X_i la moins importante selon le critère d'évaluation $J(X_k)$ est éliminée

$$J(X_k) = \max_{x_i \in X_{k+1}} J\left(X_{k+1} - \{x_i\}\right)$$

✓ liste ordonnée des variables selon leur importance : Les variables les plus pertinentes sont alors les variables qui se trouvent dans les dernières positions de la liste.

Sélection de variables

La procédure *BS* effectue sa recherche dans les deux directions (Forward et Backward) d'une manière concurrentielle.

La procédure s'arrête dans deux cas :

(1) quand une des deux directions a trouvé le meilleur sous ensemble de variables avant d'atteindre le milieu de l'espace de recherche

(2) quand les deux directions arrivent au milieu.

Les ensembles de variables sélectionnées trouvés respectivement par *SFS* et par *SBS* ne sont pas égaux à cause de leurs différents principes de sélection.

Cette méthode réduit le temps de recherche puisque la recherche s'effectue dans les deux directions et s'arrête dès qu'il y a une solution quelle que soit la direction.

Sélection de variables

- *Critères d'arrêt*

✓ Le nombre optimal de variables n'est pas connu a priori, l'utilisation d'une règle pour contrôler la sélection-élimination de variables permet d'arrêter la recherche lorsque aucune variable n'est plus suffisamment informative.

✓ Le critère d'arrêt est souvent défini comme une combinaison de la procédure de recherche et du critère d'évaluation.

✓ Une heuristique, souvent utilisée, consiste à calculer pour les différents sous-ensembles de variables sélectionnées une estimation de l'erreur de généralisation par validation croisée.

✓ Le sous-ensemble de variables sélectionnées est celui qui minimise cette erreur de généralisation.

Réduction des dimensions par extraction de caractéristiques

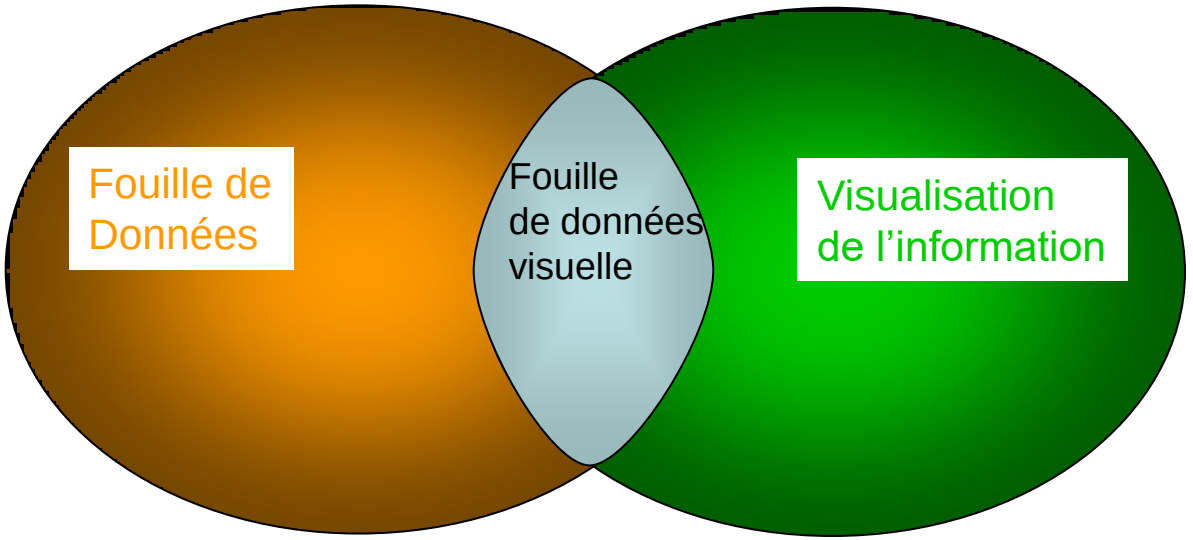
Réduction des dimensions par extraction de caractéristiques

Deux grandes familles de méthodes :

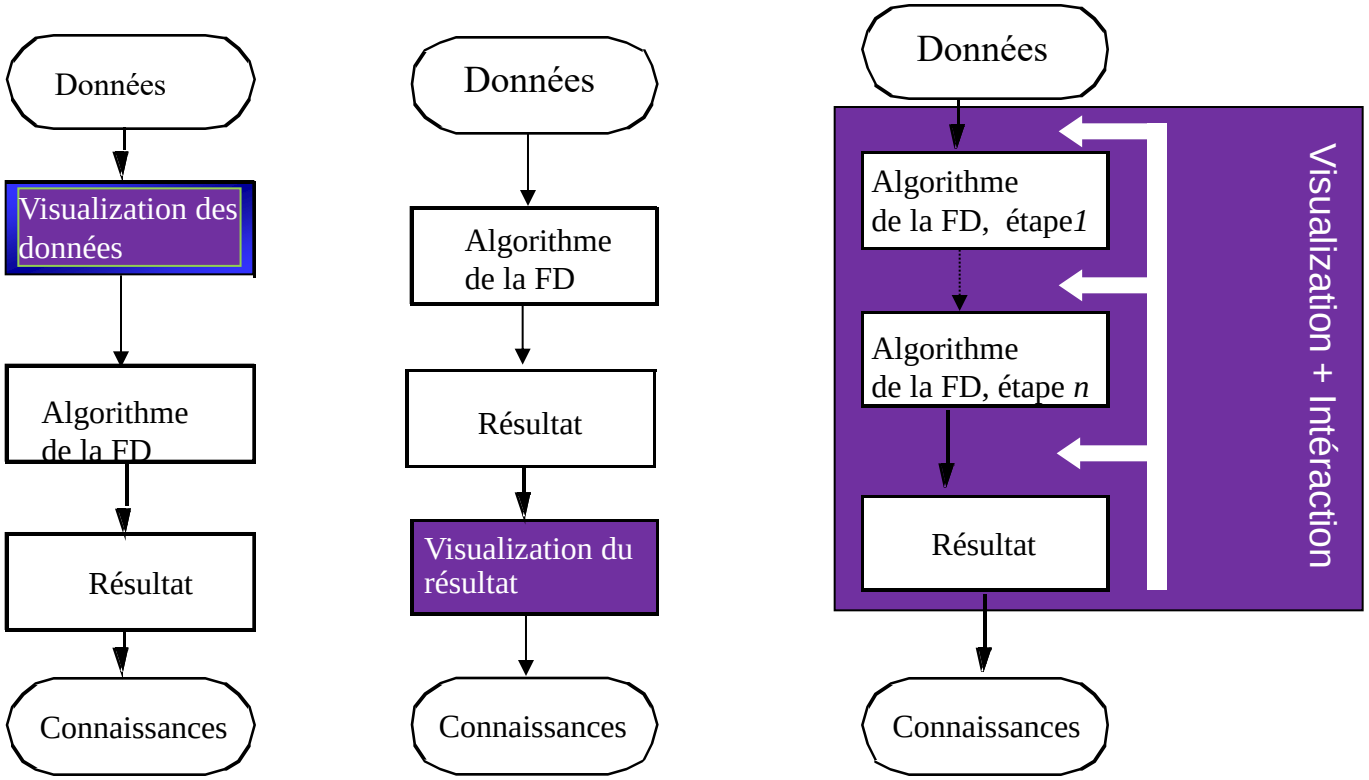
- **Méthodes linéaires**
  - Analyse en Composantes Principales (ACP)
  - Analyse Discriminante Linéaire (ADL)
  - Multi-Dimensional Scaling (MDS)
  - …
- **Méthodes non-linéaires**
  - Isometric feature mapping (Isomap)
  - Locally Linear Embedding (LLE)
  - Kernel PCA
  - Segmentation spectrale (spectral clustering)
  - Methodes supervisées (S-Isomap)
  - …

Fouille de données visuelle

Fouille de données visuelle



Fouille de données visuelle: Schéma



Logiciels de visualisation

- Graphviz
- Tulip
- Knime
- R
- ...

www.KDnuggets.com/software/visualization.html