Préparation des données

Camille Besse

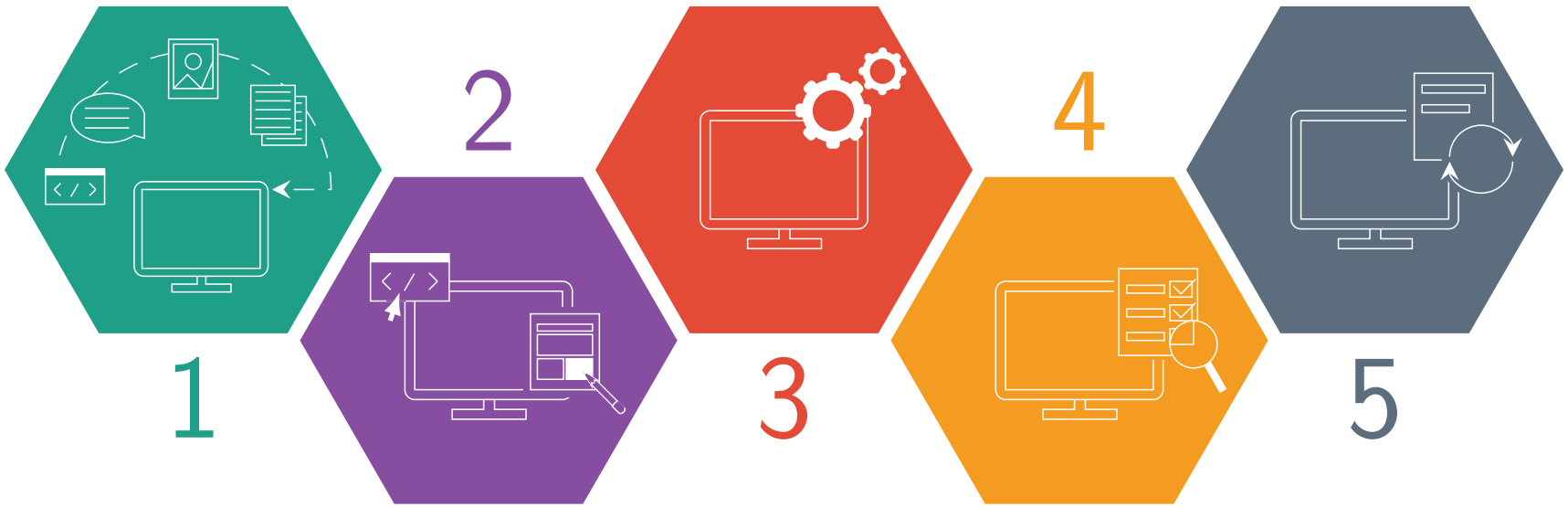
Département d'Informatique et de Génie Logiciel
Université Laval, Québec, Canada

camille.besse@ift.ulaval.ca

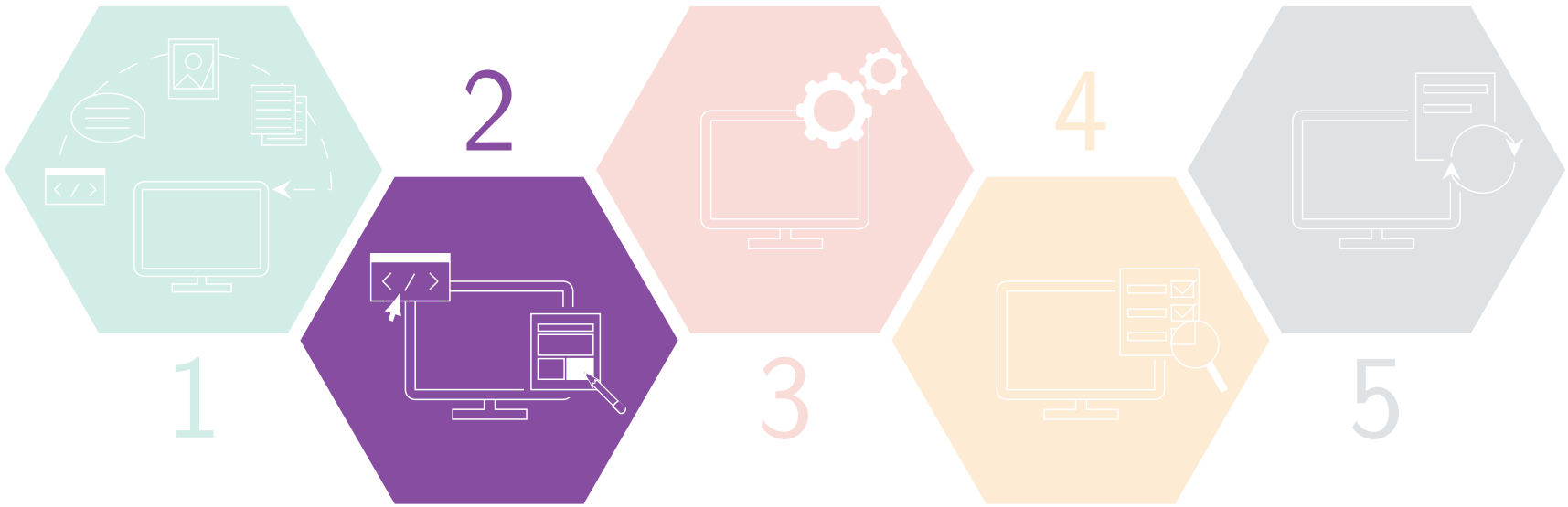
May 22, 2019

Introduction

Introduction



Introduction



Introduction

Citation

Les scientifiques de données ne travaillent que le *vendredi*
– Anonyme

Outline

Collecte & Colligation

Collecte & Colligation

1. Validité
   Cohérence et qualité de l'agrégation et de la colligation
2. Fiabilité
   Procédure d'examen périodique et maintenance
3. Temporalité
   Régularité des mesures et accessibilité en tout temps
4. Précision
   Réduire au maximum les doublons ou les absents
5. Intégrité
   Protection et/ou modification contrôlée

Nettoyage *a priori*

- Validation de la cohérence des colonnes/valeurs
- Validation de la cohérence colonnes/types
- Statistiques sur les colonnes
  Min/Max/Mean/Median cohérents
- Données manquantes en quelle proportion par ligne/colonne
- Quantité et signifiance des doublons ?

Erreurs courantes à éviter

1. Préparer sans objectif clair
2. Préparer sans idée de visualiser
3. Non contextualisation des données (éthique,culturel,. . .)
4. Saisie ou mauvaise transformation des données
5. Analyse de (trop) peu de données
6. Nommage confus des caractéristiques
7. Duplication des données
8. Altération des données
9. Agrégation des sources
10. Âgisme des données

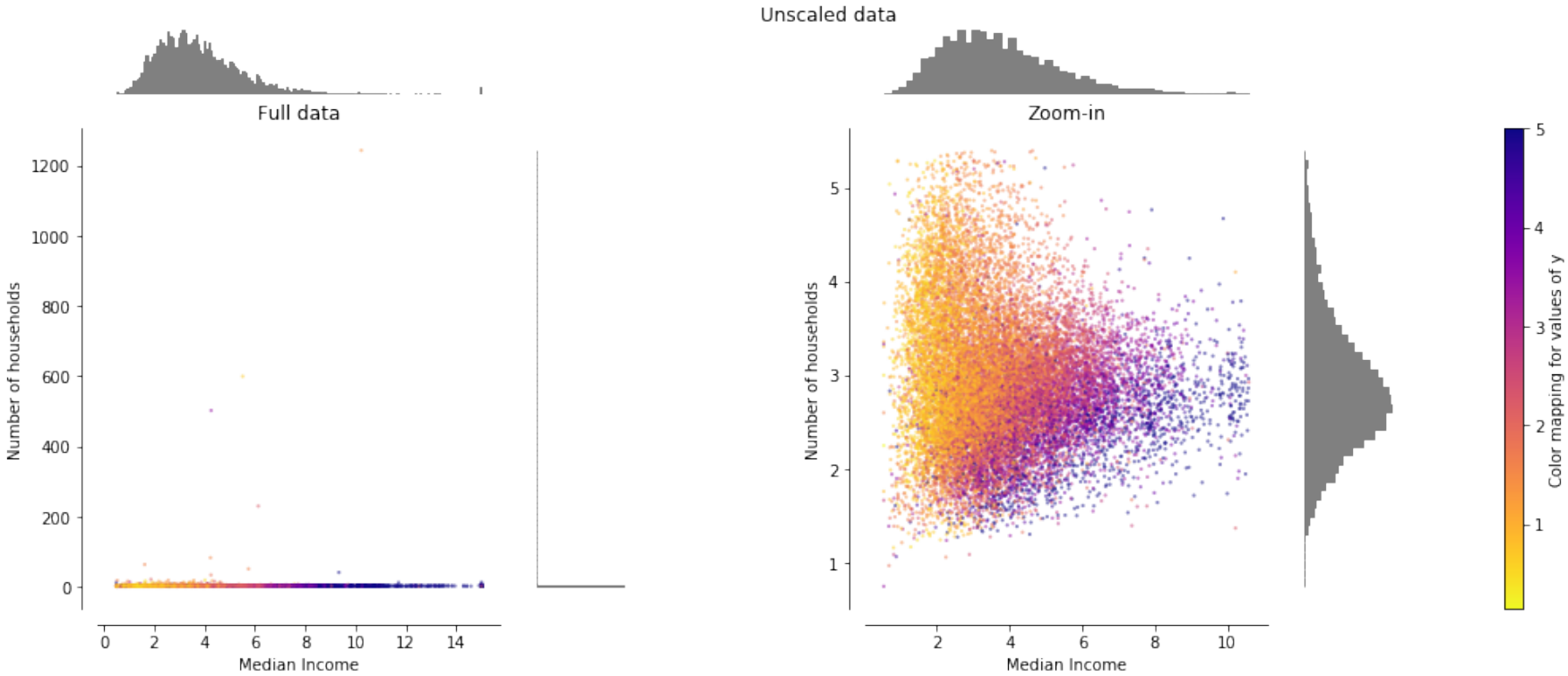
Données aberrantes

Nettoyage *a priori*

- Doit-on s'en préoccuper ?
  - Dépend de la taille du dataset
- S'il y a beaucoup de données : non.
- Si elles ont une influence sur une simple régression : peut-être.
  - Valider par visualisation.
- S'il y a peu de données:
  - Est-ce vraiment des données aberrantes ?
  - Ou juste des données débalancées ?

$\Rightarrow$ Si possible : collecter toujours plus de données !

Outliers


Full data
Number of households
Median Income
Unscaled data
Zoom-in
Number of households
Median Income
Color mapping for values of y

Identification

- Covariance Robuste : Cherche la gaussienne qui explique le mieux les données
  - Sensible aux distributions multimodales
- SVM mono-classe : séparateur a priori
  - Bon pour déterminer la nouveauté mais sensible aux données aberrantes
- Forêt d'Isolation : random forest aléatoirement appris
- Facteur local d'abérration : k-NN

Données manquantes

Imputation

Les données sont parfois incomplètes:

- Certaines colonnes sont vides
  - Imputation
  - On jette la donnée ? Combien de colonnes manquantes ? Cela change beaucoup les stats ?
- Une colonne sans aucun sens pour les données
  - Pas d'imputation (on jette la colonne ?)

Imputation

- Imputation simple : "*Filling the blanks*"
  - Statistiques : moyenne, mediane, selon la fréquence,...
  - Techniques non-paramétriques: k-nn, hot-deck (échantillonage de données similaires),...
  - Ou paramétriques: régression, ...
- Imputation multiple: "*Sampling the blanks*"
  - Faire le travail plusieurs fois et évaluer l'impact statistique

FLAG les données imputées pour éviter de les considérer comme des valeurs réelles !

Données débalancées

Données débalancées

- Lorsque c'est plus simple de juste ignorer un sous-ensemble des données
  - Qui se trouve être une catégorie complète ...
- Solution 1 : Rééchantillonnage
  - Synthetic Minority Oversampling TEchnique
    - ★ Bordeline : en privilégiant les frontières de décision
    - ★ Avec Boosting
  - Synthetic Majority Undersampling TEchnique (éventuellement informé)
    - ★ Cluster centroids : K-means
    - ★ Tomek Links

Données débalancées

- Solution 2 : Algorithmes rebalancés par le coût
  - La pénalité de mal classer un exemple de la classe minoritaire est plus grande que celui de l'autre classe
    e.g. Cost-sensitive boosting, decision trees ou neural networks
- Solution 3 : Les deux techniques précédentes dans des méthodes à base de noyau ou par apprentissage actif
    e.g. Cost sensitive undersampling SVMs
- Autres : Classification mono-classe, métriques de rang plutôt que de classe.

Transformation des données

Transformation des données

- Rééchelonnage ([-1,1] or [0,1])

$$x' = \frac{x - \min x}{\max x - \min x}$$

- Normalisation à la moyenne

$$x' = \frac{x - \bar{x}}{\max x - \min x}$$

- Standardisation (Loi de Gauss)

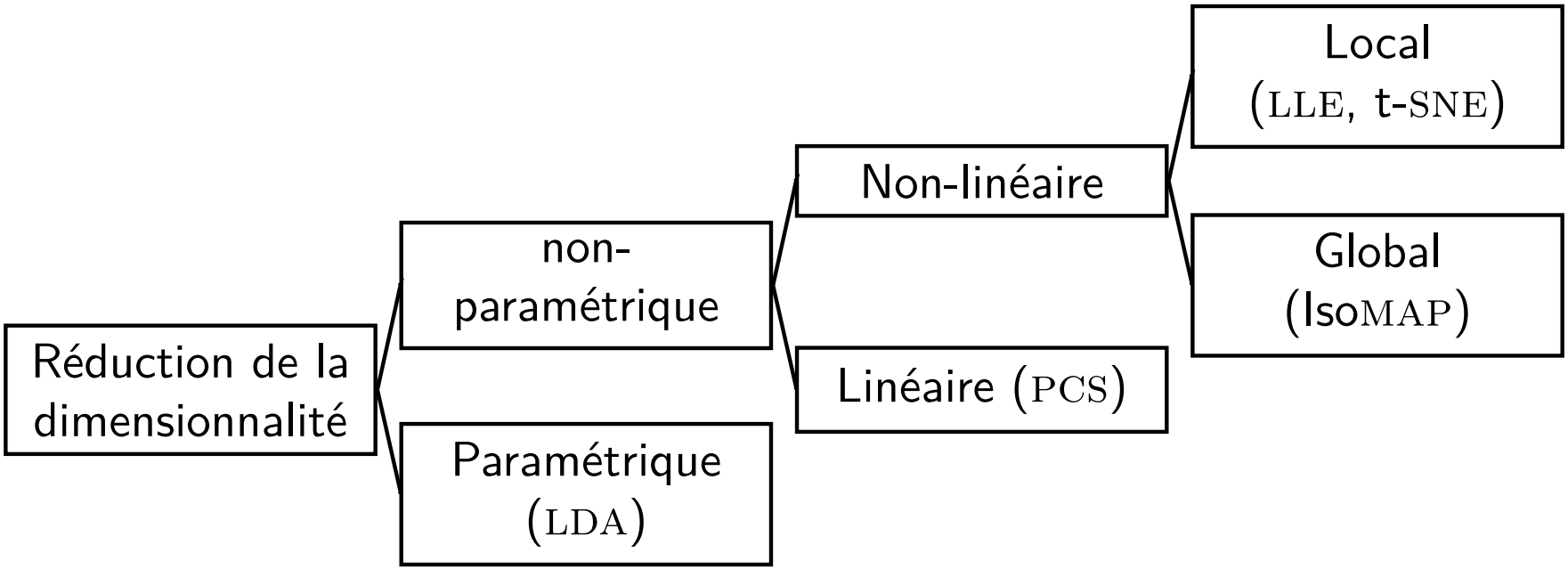
$$x' = \frac{x - \bar{x}}{\sigma_x}$$

- Normalisation à l'unité

$$x' = \frac{x}{||x||}$$

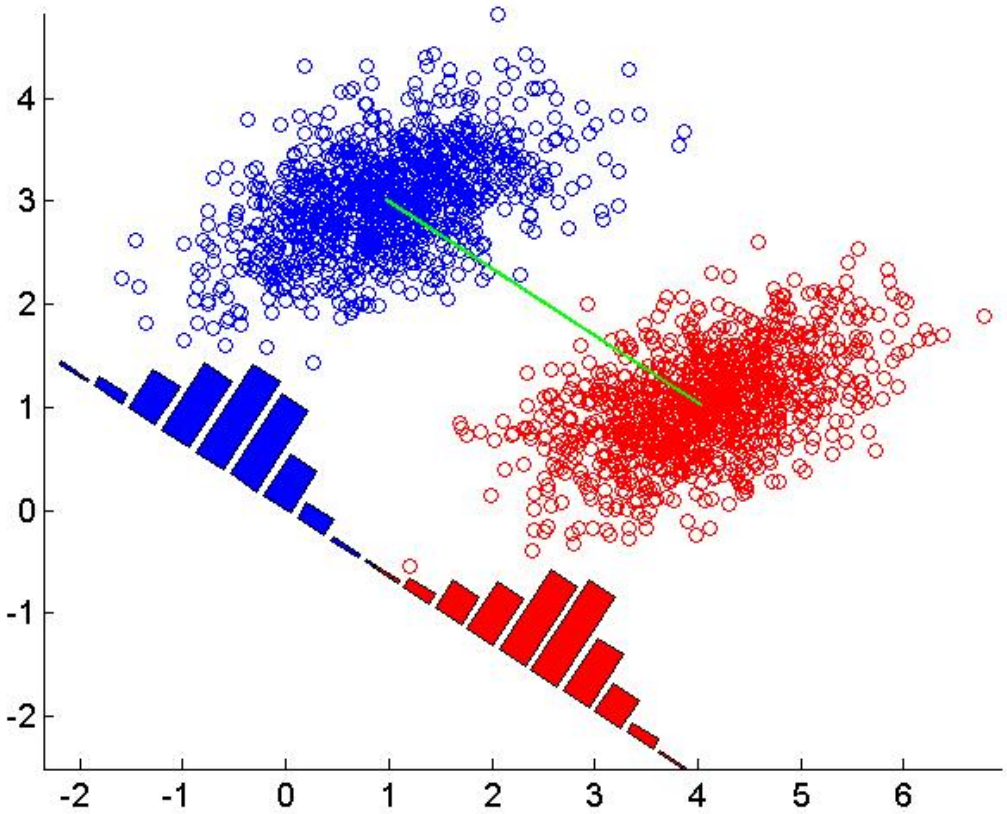
Transformation des données

- Projection d'un espace de taille C dans un sous-espace de taille $n << C$
  - Pour visualisation ($n = 2$ ou $n = 3$)
  - Pour densifier l'information
- PCA, LDA, MDS, IsoMap, LLE, LTSA, t-NE,...



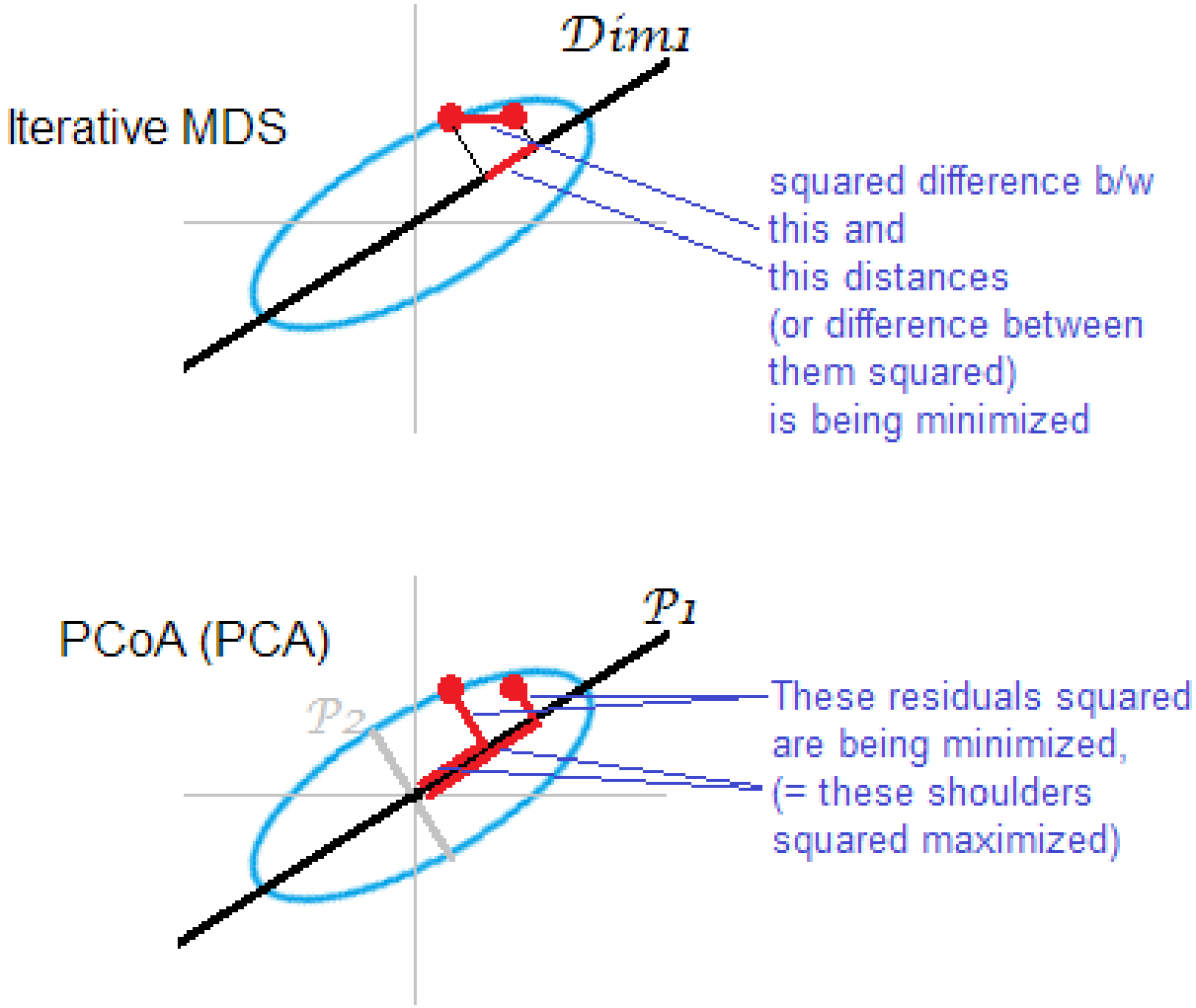
LDA : Linear Discriminant Analysis

- PCA dont l'objectif n'est pas de maximiser la variabilité, mais l'explicabilité d'une certaine variable catégorique (classe)
- Sélectionne les n variables expliquant le mieux l'étiquette



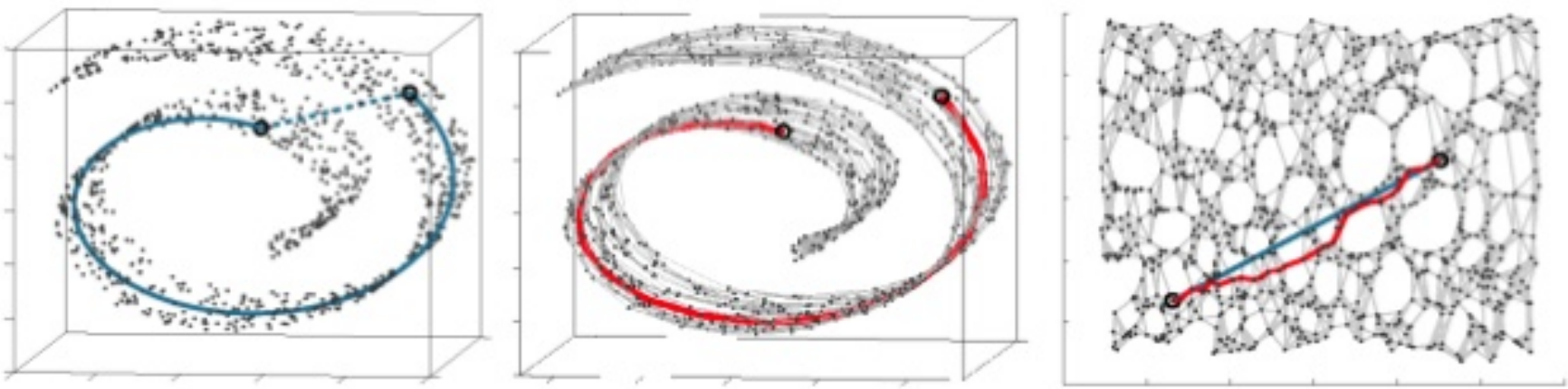
MDS : Multi-Dimensionnal Scaling

- Généralisation du PCA sur une mesure de dissimilarité entres les différents points



IsoMAP

- MDS pour lequel on utilise une autre distance que la distance Euclidienne : Distance Géodesique



LLE : Locally Linear Embedding

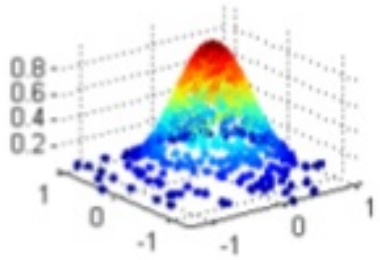
- Exprime un point selon une combinaison linéaire de ses voisins dans l'espace original

$$E(W) = \sum_i |\mathbf{X}_i - \sum_j \mathbf{W}_{ij}\mathbf{X}_j|^2$$

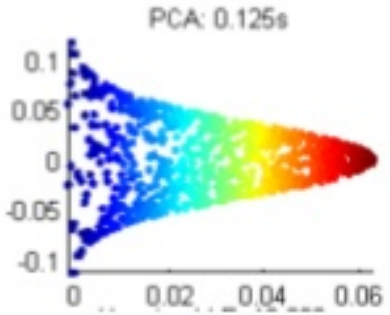
- Recalcul la position du point dans l'espace d'arrivée selon cette même combinaison

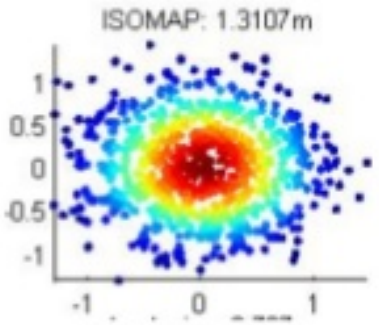
$$C(Y) = \sum_i |\mathbf{Y}_i - \sum_j \mathbf{W}_{ij}\mathbf{Y}_j|^2$$

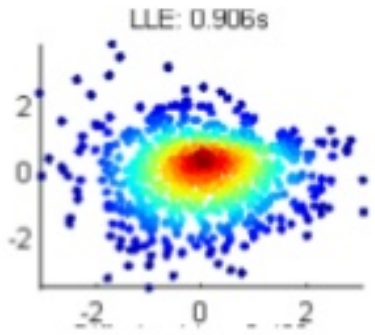
LLE : Locally Linear Embedding

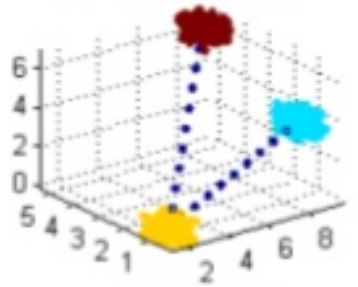


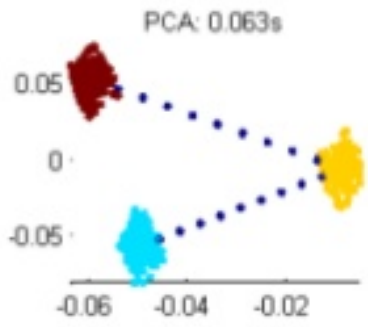
PCA vs. ISOMAP vs. LLE

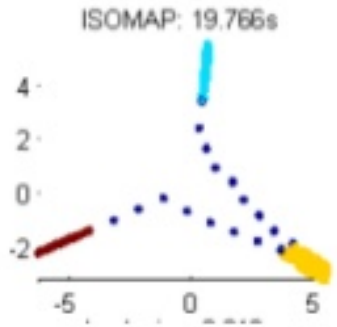


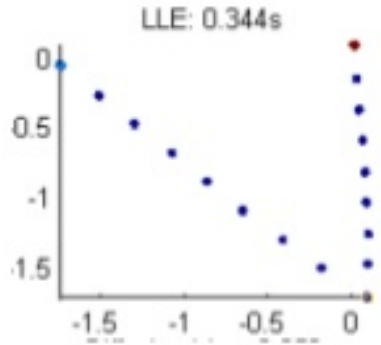






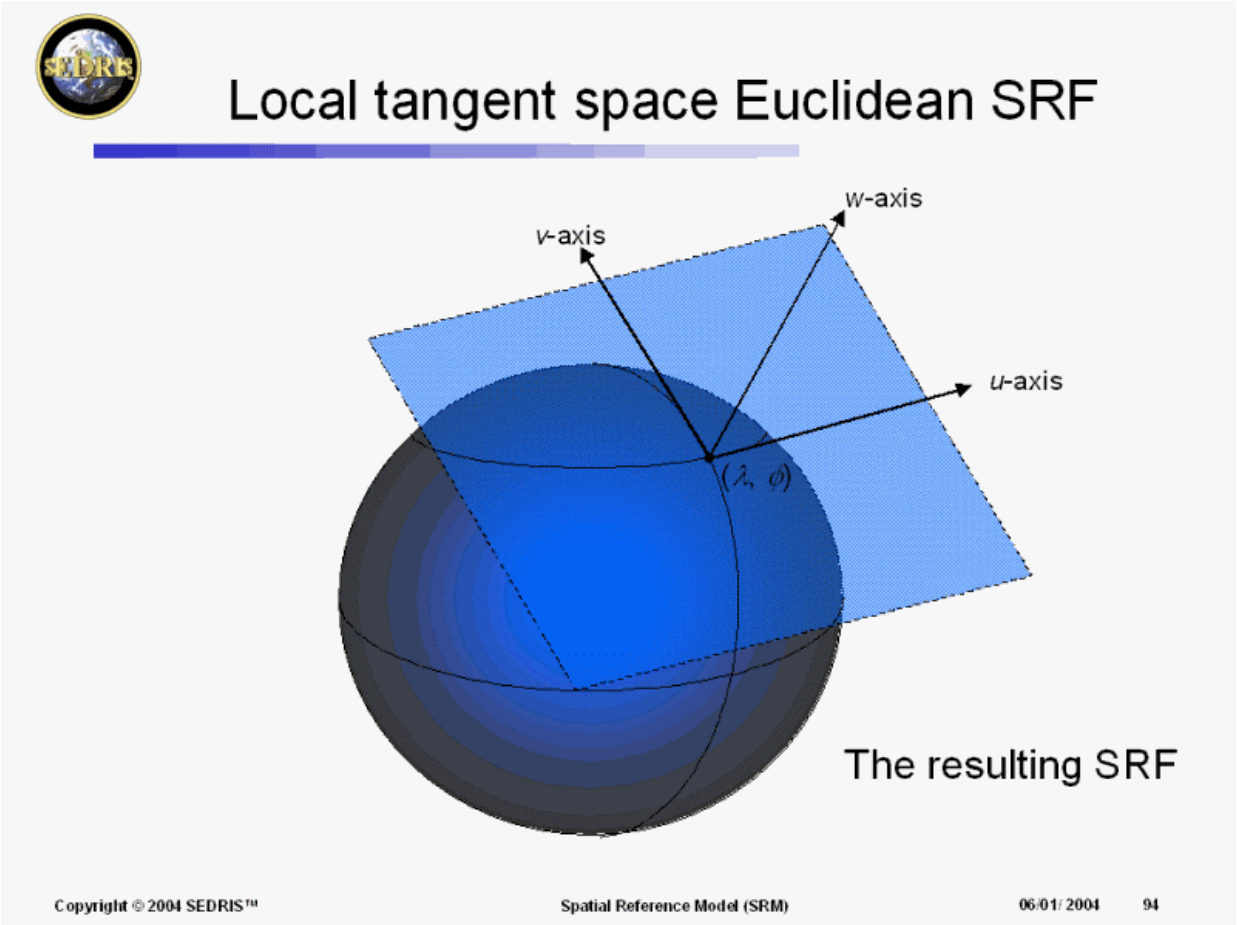






LTSA : Local Tangent Space Alignment

- Basée sur l'idée qu'un espace correctement "déplié" a tous ses hyperplans tangents alignés
- Calcule donc un hyperplan tangent pour chaque point calcule une représentation qui les aligne.



t-SNE : t-distributed Stochastic Neighbour Embeddin

- Transformation non linéaire des données dans un sous-espace de ou 3 dimensions
  - Très souvent pour de la visualisation
- Produit deux distributions de probabilité sur les paires de points proportionnelle à leur similarité:
  - Dans l'espace à grand dimension
  - Dans l'espace transformé
- Minimise la KL-divergence entre les deux distributions selon leur position dans l'espace transformé

t-SNE : t-distributed Stochastic Neighbour Embeddin

https://distill.pub/2016/misread-tsne/

Step
1,750

Points arranged in 3D, on two linked circles. Different runs may give different results.

Share this view

Number Of Points 50

Perplexity 10

Epsilon 5

MARTIN WATTENBERG
Google Brain

FERNANDA VIÉGAS
Google Brain

IAN JOHNSON
Google Cloud

Oct. 13
2016

Citation:
Wattenberg, et al., 2016

That's all folks !

Questions ?