Quelques Problèmes pratiques

Antoine Cornuéjols

AgroParisTech – INRAé MIA-Paris-Saclay

EKINOCS research group

antoine.cornuejols@agroparistech.fr

AgroParisTech
Talents d'une planète soutenable

INRAE

Quelques **problèmes pratiques**

1. Motivation
2. Prétraitement des données
3. Valeurs manquantes : méthodes d'imputation
4. La réduction de dimension
   a. Par sélection de variables
   b. Par projection dans un nouvel espace
5. Points aberrants
6. Classes déséquilibrées
7. Conclusion

Les « données »

- Des **formats** différents
- Des **sources** hétérogènes
- Des **descriptions** imparfaites
  - Données *manquantes*
  - Données *fausses*
  - Attributs *redondants*
  - Attributs *non pertinents*
- **Autres problèmes**
  - Classes très *déséquilibrées*
  - Points *aberrants* (*outliers*)
  - *Grandes dimensions*

Introduction : les problèmes potentiels

- Relatives aux **données**
  - Valeurs **manquantes**
  - **Bruit**
    - de description
    - de classe
  - Attributs **redondants, corrélés**
  - Attributs **non pertinents**
  - Dimensions **hétérogènes**
    - Domaines de variation très différents
    - Valeurs numériques / symboliques

Qualité des données

Qu'est-ce que c'est ?

Une **combinaison subtile de propriétés** :

- **Précision** EGC-2010 a eu lieu à Hammamet en Tunisie
- **Cohérence** Il y a une seule conférence EGC par an
- **Complétude** Chaque conférence EGC est localisée quelque part
- **Fraîcheur** Le lieu de la dernière conférence EGC était Dijon en France
- **Unicité** :
  - EGC est une conférence, pas une congrégation
  - EGC-2010 et *Extraction et Gestion des Connaissances 2010* font référence au même évènement

Intégration de données provenant de sources multiples

Opérations

Intégration et transformation. Fusionner différentes sources de données.

- Identifier les entitées, par ex. `client_id` et `cl_nr`.
- Uniformiser les données exprimées en unités diffférentes, par ex. DOLLAR et EURO
- Convertir les adresses en coordonnées
- Calculer la vente quotidienne à partir des ventes individuelles.
- Normaliser les variables entre 0 et 1.
- Remplacer toute valeur x d'un attribut A avec moyenne μ et variance σ^2 par $\frac{x-\mu}{\sigma}$ afin d'arriver à une moyenne de 0 et une variance de 1.

Qualité des données

La qualité des données est un problème fondamental

GIGO : Garbage In Garbage Out

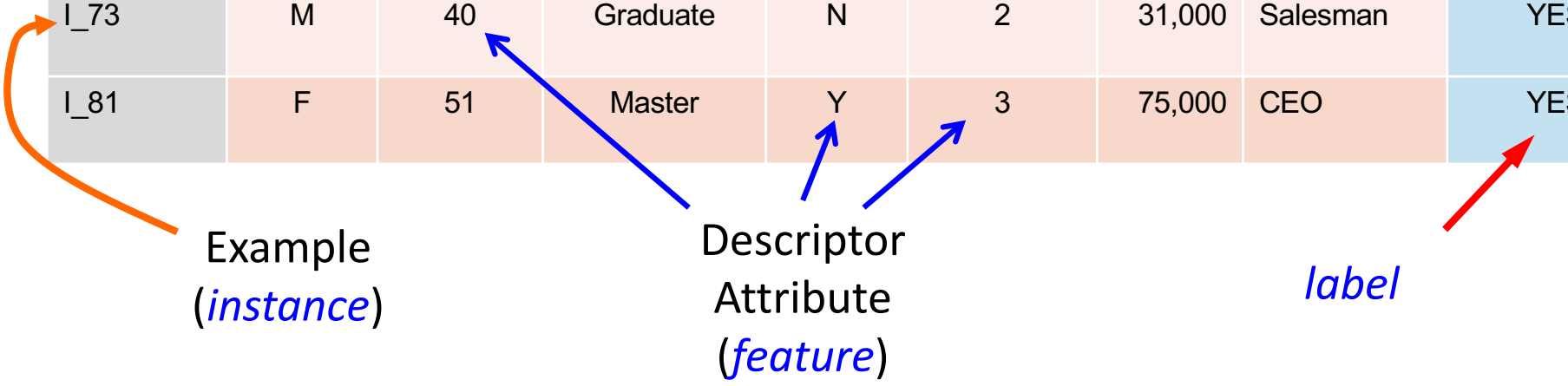
- **Omniprésent** dans toutes les applications
- **Complexe**.
  - Intrinsèque dans toute BD, entrepôt de données ou système d'information.
  - La **frontière** entre bonnes données et mauvaises données est **imprécise**.
- **Critique**. Mais coûte énormément

Quelques **problèmes pratiques**

1. Motivation
2. Prétraitement des données
3. Valeurs manquantes : méthodes d'imputation
4. La réduction de dimension
   a. Par sélection de variables
   b. Par projection dans un nouvel espace
5. Points aberrants
6. Classes déséquilibrées
7. Conclusion

Data: organization and types

Identifier	Gender	Age	Education level	Maried	Nb of children	Salary	Profession	To prospect?
I_21	M	43	Master	Y	3	55,000	Architect	YES
I_34	M	25	Sophomore	N	0	21,000	Nurse	NO
I_38	F	34	PhD	Y	2	35,000	Univ. Prof.	YES
I_39	F	67	Bachelor	Y	5	20,000	Retired	NO
I_58	F	56	Technical studies	Y	4	27,000	Employee	NO
I_73	M	40	Graduate	N	2	31,000	Salesman	YES
I_81	F	51	Master	Y	3	75,000	CEO	YES



Types de données

- **Nominales**
  - *Sexe, profession, …*
  - Nombre de valeurs **dénombrable**
  - **Aucune** relation d'ordre
  - Opérateurs arithmétiques **inapplicables**
- **Ordinales**
  - *Taille (petite, moyenne, grande)*
  - Des modalitées
  - **Relation d'ordre** entre modalités
  - **Calculs** sur des **rangs**
- **Numériques** ou **continues**
  - *Age, poids*
  - Nombre de valeurs théoriquement infini
  - **Relation d'ordre** entre les valeurs
  - Notion de **distance** et d'écart
  - **Calculs arithmétiques** possibles

Données numériques / qualitatives

- Certaines méthodes demandent des **données numériques** (fondées sur des distances)

- D'autres, des **données** symboliques ou **qualitatives** (e.g. motifs fréquents)

Les données catégorielles

ou nominales

Les **transformer** en données numériques

- Binarisation
  - Un **attribut booléen** par valeur possible
    (*N* valeurs -> *N* nouveaux attributs)

Symbolique -> numérique

- Sans proximité sémantique
  - Ex : Codage disjonctif complet (“One-hot encoding”)

Profession
Architecte
Agronome
Médecin
Pilote
Agronome

Architecte	Agronome	Médecin	Pilote
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1
0	1	0	0

Symbolique -> numérique

- Booléen -> 0, 1

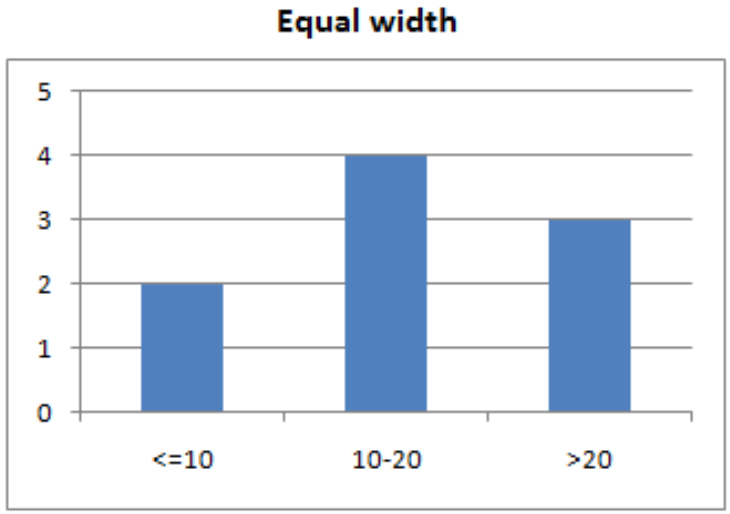
- Ensembles
  - Avec proximité sémantique -> Séquence d'entiers
    - Ex : [rouge, vert, bleu] -> [1,2,3]
  - Sans proximité sémantique
    - Ex : Code correcteur d'erreur

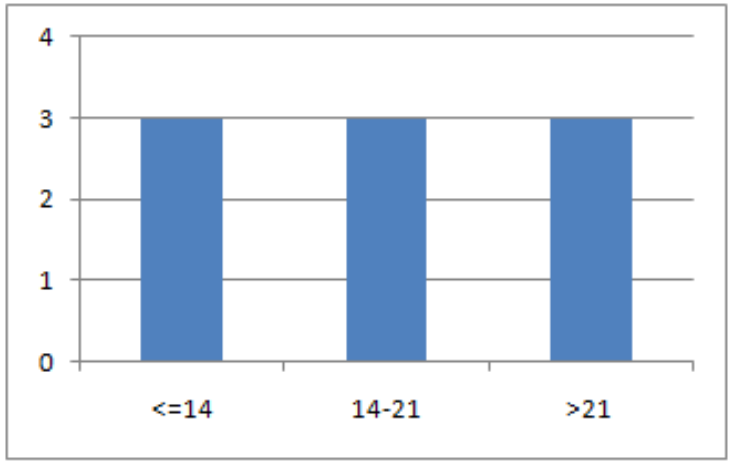
Les données numériques

Méthodes de discrétisation

- **Equal-width** binning
  - Données en *k* intervalles de **même taille**
- **Equal-frequency** binning
  - Données en *k* groupes contenant approximativement le **même nombre d'exemples**

- **Data :** 0, 4, 12, 16, 16, 18, 24, 26, 28
- **Equal width**
  - Bin 1: 0, 4 [-,10)
  - Bin 2: 12, 16, 16, 18 [10,20)
  - Bin 3: 24, 26, 28 [20,+)
- **Equal frequency**
  - Bin 1: 0, 4, 12 [-, 14)
  - Bin 2: 16, 16, 18 [14, 21)
  - Bin 3: 24, 26, 28 [21,+)





Normalisation des données

Normalisation

- Cas de **domaines de variations très différents**
  - Fausse les régularités trouvées
  - Fréquemment on utilise une ***normalisation centrée réduite***
    - **Centrée** : on ramène tous les domaines de variations autour de 0
    - **Réduite** : on divise les valeurs par l'écart-type
  - Toutes les dimensions dans [0, 1]
  - Discrétisation en un même nombre de valeurs

- Attention : ce n'est pas forcément une bonne idée

Normalisation

Pas de réponse unique

1. Clustering de **personnes**
   - Est-ce qu'un mètre est équivalent à un kilo ?
2. Clustering des **villes** au Canada
   - Les distances est-ouest > distances nord-sud
   - Sans doute une bonne idée de ne pas normaliser

Rq : l'algorithme des k-moyennes favorise les **clusters sphériques** (par rapport à la distance utilisée)

La transformation d'échelle

- Pour chaque attribut :
  1. Calculer la déviation standard
  2. Diviser sa valeur par cette déviation

```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
 Median :5.800   Median :3.000   Median :4.350   Median :1.300
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500

Created from 150 samples and 4 variables

Pre-processing:
  - ignored (0)
  - scaled (4)

  Sepal.Length    Sepal.Width      Petal.Length     Petal.Width
 Min.   :5.193   Min.   : 4.589   Min.   :0.5665   Min.   :0.1312
 1st Qu.:6.159   1st Qu.: 6.424   1st Qu.:0.9064   1st Qu.:0.3936
 Median :7.004   Median : 6.883   Median :2.4642   Median :1.7055
 Mean   :7.057   Mean   : 7.014   Mean   :2.1288   Mean   :1.5734
 3rd Qu.:7.729   3rd Qu.: 7.571   3rd Qu.:2.8890   3rd Qu.:2.3615
 Max.   :9.540   Max.   :10.095   Max.   :3.9087   Max.   :3.2798
```

La transformation centrage

- Pour chaque attribut :
  1. Calculer la moyenne
  2. La soustraire de sa valeur

```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
 Median :5.800   Median :3.000   Median :4.350   Median :1.300
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500

Created from 150 samples and 4 variables

Pre-processing:
  - centered (4)
  - ignored (0)

 Sepal.Length       Sepal.Width        Petal.Length     Petal.Width
 Min.   :-1.54333   Min.   :-1.05733   Min.   :-2.758   Min.   :-1.0993
 1st Qu.:-0.74333   1st Qu.:-0.25733   1st Qu.:-2.158   1st Qu.:-0.8993
 Median :-0.04333   Median :-0.05733   Median : 0.592   Median : 0.1007
 Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.000   Mean   : 0.0000
 3rd Qu.: 0.55667   3rd Qu.: 0.24267   3rd Qu.: 1.342   3rd Qu.: 0.6007
 Max.   : 2.05667   Max.   : 1.34267   Max.   : 3.142   Max.   : 1.3007
```

La transformation standardisation

- Pour chaque attribut :

  Combine la transformation échelle (division par la déviation standard) et le centrage

  -> Les attributs ont une moyenne de 0 et une déviation standard de 1

```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
 Median :5.800   Median :3.000   Median :4.350   Median :1.300
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500

Created from 150 samples and 4 variables

Pre-processing:
  - centered (4)
  - ignored (0)
  - scaled (4)

 Sepal.Length       Sepal.Width        Petal.Length       Petal.Width
 Min.   :-1.86378   Min.   :-2.4258   Min.   :-1.5623   Min.   :-1.4422
 1st Qu.:-0.89767   1st Qu.:-0.5904   1st Qu.:-1.2225   1st Qu.:-1.1799
 Median :-0.05233   Median :-0.1315   Median : 0.3354   Median : 0.1321
 Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
 3rd Qu.: 0.67225   3rd Qu.: 0.5567   3rd Qu.: 0.7602   3rd Qu.: 0.7880
 Max.   : 2.48370   Max.   : 3.0805   Max.   : 1.7799   Max.   : 1.7064
```

La transformation normalisation

- Pour chaque attribut :

  -> Les attributs ont une moyenne de 0 et sont dans [0, 1]

```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
 Median :5.800   Median :3.000   Median :4.350   Median :1.300
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500

Created from 150 samples and 4 variables

Pre-processing:
  - ignored (0)
  - re-scaling to [0, 1] (4)


  Sepal.Length     Sepal.Width      Petal.Length     Petal.Width
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
 1st Qu.:0.2222   1st Qu.:0.3333   1st Qu.:0.1017   1st Qu.:0.08333
 Median :0.4167   Median :0.4167   Median :0.5678   Median :0.50000
 Mean   :0.4287   Mean   :0.4406   Mean   :0.4675   Mean   :0.45806
 3rd Qu.:0.5833   3rd Qu.:0.5417   3rd Qu.:0.6949   3rd Qu.:0.70833
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000
```

Prétraitements pour les séries temporelles

Types de **prétraitements**

- Valeurs **manquantes**
  - Interpolation
- Valeurs **bruitées**
  - Binning
  - Lissage par moyenne glissante (Moving-average smoothing)
  - Lissage exponentiel (Exponential smoothing)
- **Normalisation**
- **Discrétisation**
- Changement de **représentation**
  - Transformation en ondelettes discrètes (DWT)
  - Transformation de Fourier discrète (DFT)

Valeurs **manquantes**

- Interpolation linéaire
  - Si x_i et x_j sont des valeurs pour les dates t_i et t_j, on peut estimer la valeur pour la date t avec $t_i \leq t \leq t_j$ par :

$$x = x_i + \frac{t - t_i}{t_j - t_i} \cdot (x_j - x_i)$$

- Des méthodes plus complexes peuvent aussi être utilisées
  - Interpolation **polynomiale**
  - Interpolation par **spline**

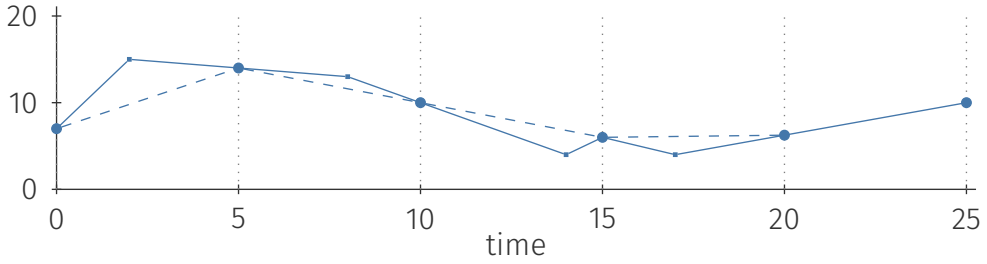
Interpolation pour synchroniser les valeurs

- Si les valeurs mesurées ne sont pas régulièrement espacées dans le temps
  - On peut utiliser une interpolation linéaire

$$x = x_i + \frac{t - t_i}{t_j - t_i} \cdot (x_j - x_i)$$

Exemple : on veut des valeurs pour les dates (0, 5, 10, 15, 20, 25)

On a les valeurs ((0, 7), (2, 15), (8, 13), (14, 4), (15, 6), (17, 4), (25, 10))



On obtient :

Valeurs **bruitées**

- Anomalies
  - Résultent de fluctuations durant le processus de **génération** des données
- Bruit
  - Causé par les artifacts de la **mesure** des données

- Techniques pour combattre le « bruit »
  - Binning
  - Lissage (*Smoothing*)

Binning

Consider a time-series $\mathcal{S}_X = \langle x_1, x_2, \ldots, x_n \rangle$, with values at each of n equally spaced timestamps $t_1, \ldots, t_n$

Binning, a.k.a. *piecewise aggregate approximation (PAA)*, divides the time-series into time intervals of size k, i.e. into intervals $[t_1, t_k], [t_{k+1}, t_{2k}], \ldots, [t_{(\lfloor n/k \rfloor - 1)k+1}, t_{\lfloor n/k \rfloor k}]$

Binned values are averages of values within each interval

$$y_i = \frac{1}{k} \sum_{r=1}^{k} x_{(i-1)k+r} \quad \text{for } i = 1, \ldots, \lfloor n/k \rfloor$$

Binning is lossy, reduces the number of points by a factor of k

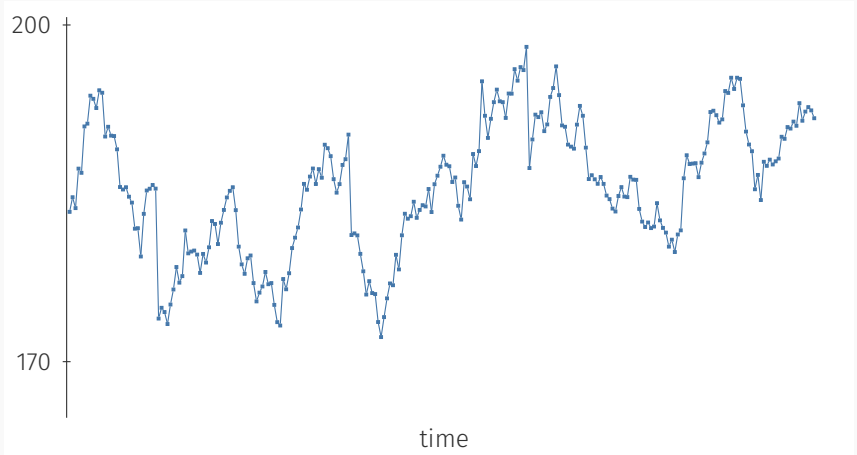
Instead of average, it is possible to take the median, which is more robust to the presence of outlier values

...

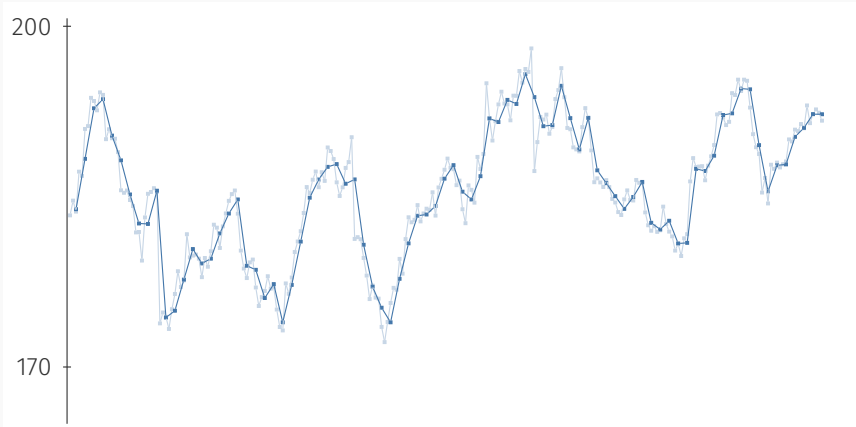
Binning : illustration

IBM stock prices from Sept. 2013 to Sept. 2014

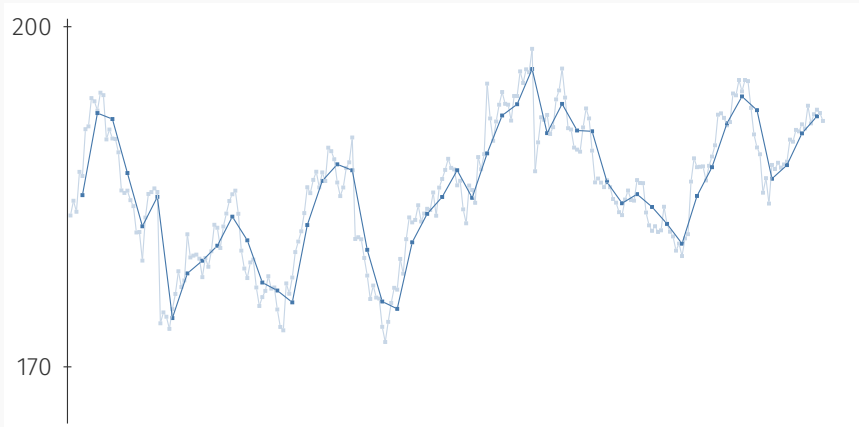
Original time-series



Binning, $k = 3$



Binning, $k = 5$



...

Lissage par moyenne mobile

Soit une série temporelle $S_X = \langle x^{(1)}, x^{(2)}, \ldots, x^{(n)} \rangle \in \mathcal{R}^n$
à valeurs régulièrement espacées aux temps $t_1, \ldots, t_n$

La **régularisation par moyenne mobile** (*Moving-average smoothing*) utilise des fenêtres de taille k qui se chevauchent $[t_1, t_k], [t_2, t_{k+1}], \ldots, [t_{n-k+1}]$

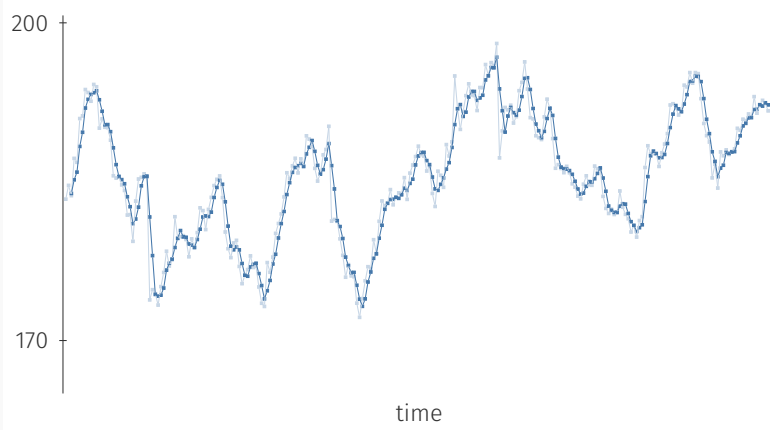
Les valeurs moyennées sont calculées comme :

$$y_i = \frac{1}{k} \sum_{r=0}^{k-1} x_{i+r} \quad \text{pour } i = 1, \ldots, n-k+1$$

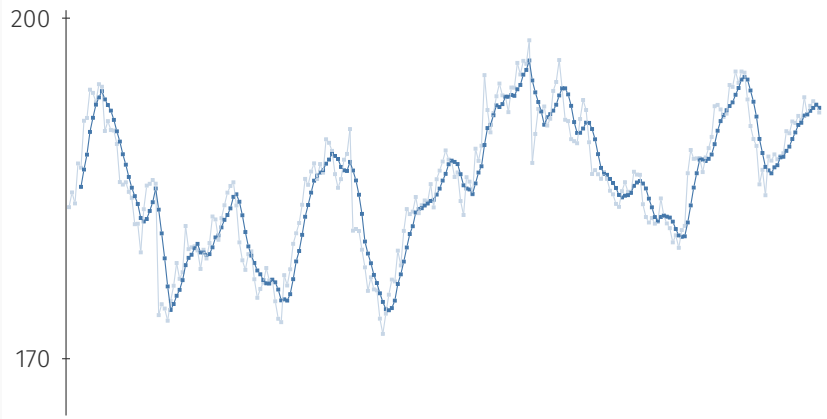
Des valeurs de k plus grandes conduisent à une régularisation plus forte.

Lissage par moyenne mobile : illustration

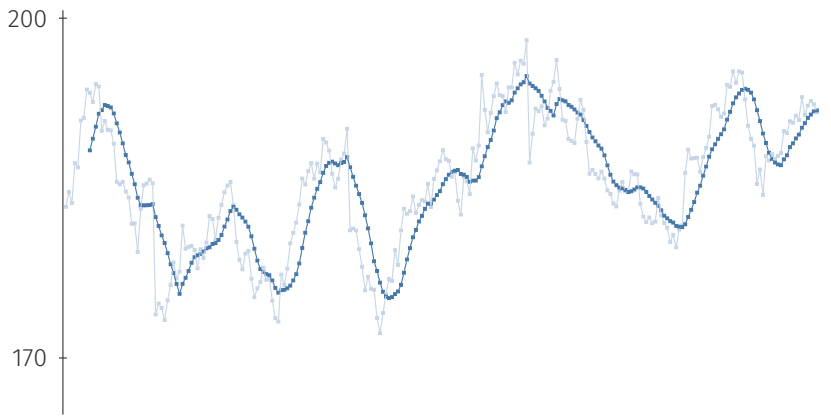
IBM stock prices from Sept. 2013 to Sept. 2014
Moving-average smoothing, $k = 3$



Moving-average smoothing, $k = 5$

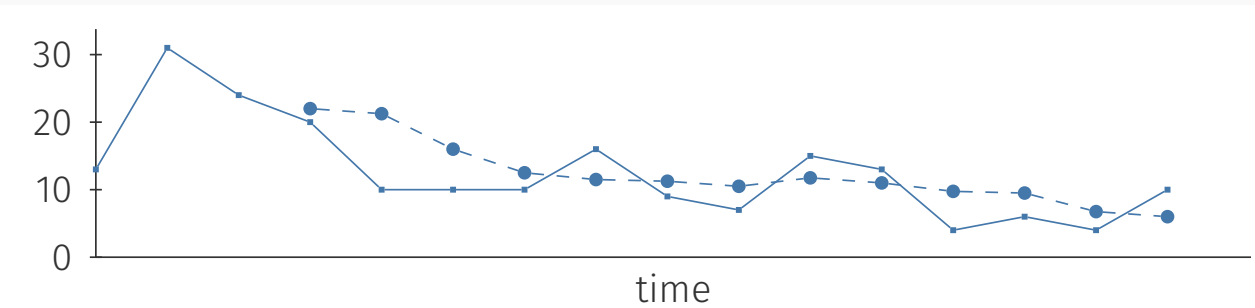


Moving-average smoothing, $k = 9$



...

Lissage par moyenne mobile : illustration



$\langle 13, 31, 24, 20, 10, 10, 10, 16, 9, 7, 15, 13, 4, 6, 4, 10 \rangle$

Moving average smoothing, window of width 4

$$y_i = \frac{1}{k} \sum_{r=0}^{k-1} x_{i+r} \qquad \text{pour } i = 1, \ldots, n-k+1$$

...

Lissage exponentiel

Soit une série temporelle $S_X = \langle x^{(1)}, x^{(2)}, \ldots, x^{(n)} \rangle \in \mathcal{R}^n$
à valeurs régulièrement espacées aux temps $t_1, \ldots, t_n$

Dans le **lissage exponentiel**, la valeur courante lissée est définie comme une combinaison linéaire de la valeur courante originale et des valeurs lissées précédentes.

Pour un paramètre de lissage $\alpha \in [0, 1]$ et avec $y_1 = x_1$:

$$y_i = \alpha \cdot x_i + (1 - \alpha) \cdot y_{i-1} \quad \text{pour } i = 2, \ldots, n$$

...

Lissage exponentiel

Pour un paramètre de lissage $\alpha \in [0, 1]$ et avec $y_1 = x_1$:

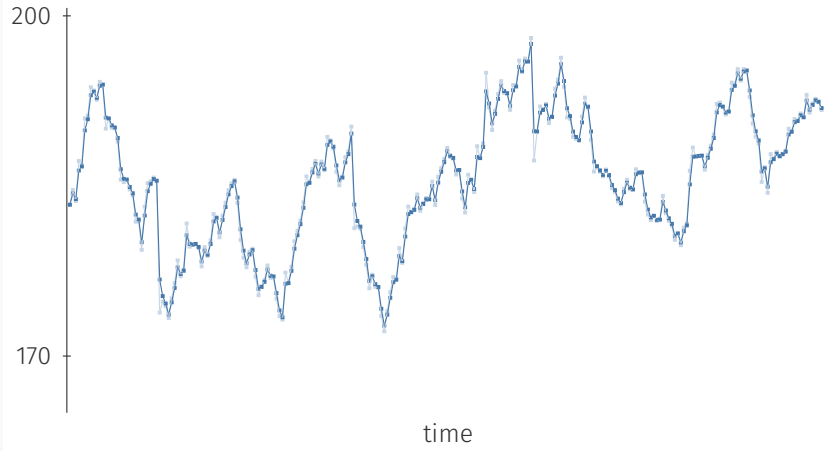
$$y_i = \alpha \cdot x_i + (1 - \alpha) \cdot y_{i-1} \quad \text{pour } i = 2, \ldots, n$$

- Les valeurs lissées peuvent être vues comme une **somme à décroissance exponentielle** des valeurs originales, donnant **plus de poids** aux valeurs **récentes**.
- Le paramètre α contrôle le facteur de décroissance
  - $\alpha = 1$: pas de décroissance
  - $\alpha = 0$: toute la série vaut la valeur de x_1

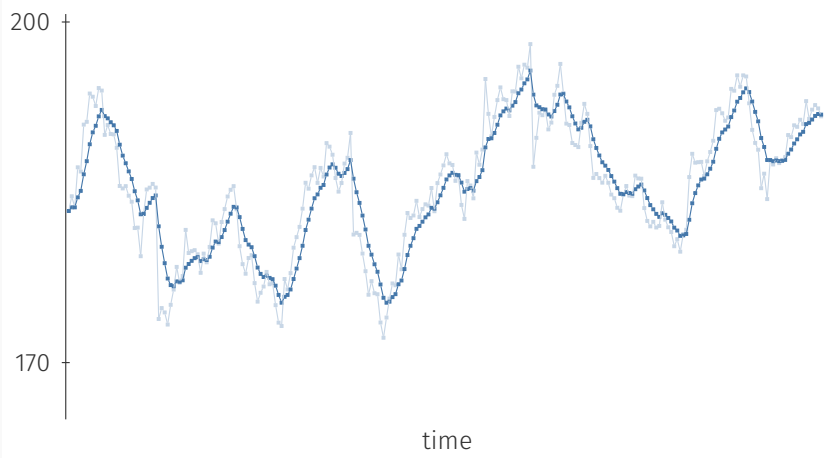
Lissage exponentiel : illustration

IBM stock prices from Sept. 2013 to Sept. 2014

Exponential smoothing, $\alpha = 0.75$

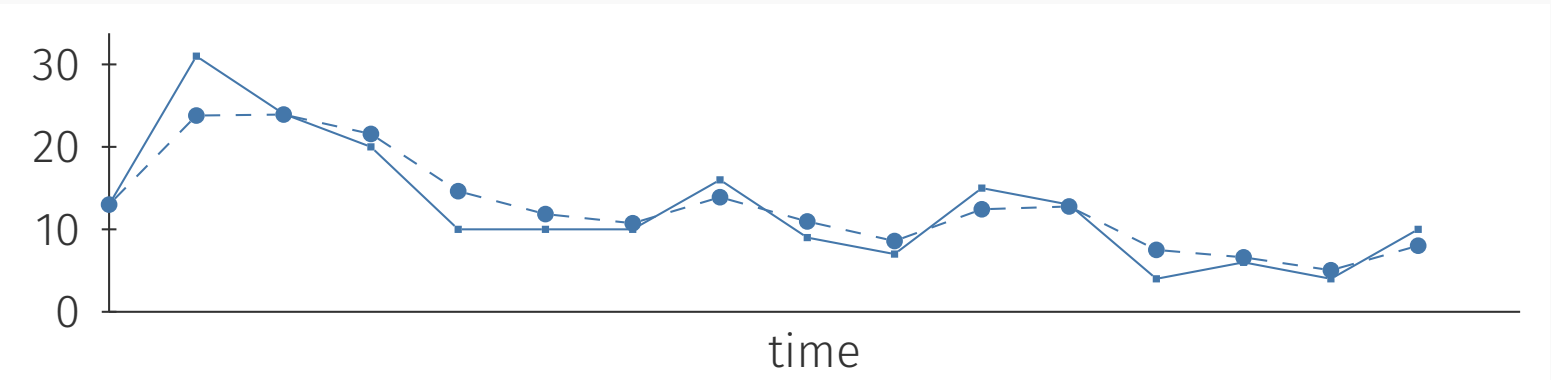


Exponential smoothing, $\alpha = 0.25$



...

Lissage exponentiel : illustration



$\langle 13, 31, 24, 20, 10, 10, 10, 16, 9, 7, 15, 13, 4, 6, 4, 10 \rangle$

Exponential smoothing, $\alpha = 0.6$

$\langle 13.00, 23.80, 23.92, 21.57, 14.63, 11.85, 10.74, 13.90, 10.96,$
$8.58, 12.43, 12.77, 7.51, 6.60, 5.04, 8.02 \rangle$

...

Normalisation

- Quand des séries temporelles sont mesurées sur des échelles différentes, il est important de les normaliser pour pouvoir les comparer

- Étant donnée une série temporelle $S_X = \langle x_1, x_2, \ldots, x_n \rangle$ prenant ses valeurs dans l'intervalle [V_{min}, V_{max}], la normalisation basée sur intervalle recode les valeurs dans l'intervalle [0, 1] selon :

$$y_i = \frac{x_i - V_{min}}{V_{max} - V_{min}}$$

Normalisation par l'écart-type

- Quand des séries temporelles sont mesurées sur des échelles différentes, il est important de les normaliser pour pouvoir en comparer les tendances

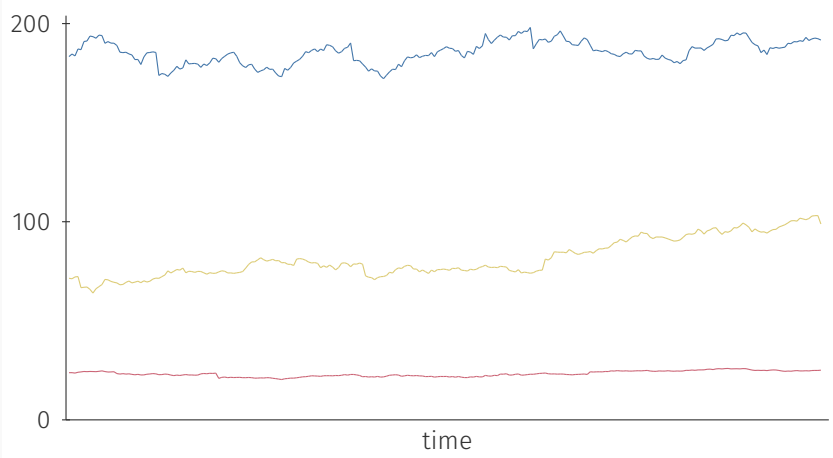
- Étant donnée une série temporelle $S_X = \langle x_1, x_2, \ldots, x_n \rangle$ de moyenne μ et d'écart-type σ, la standardisation recode les valeurs selon :

$$y_i = \frac{x_i - \mu}{\sigma}$$

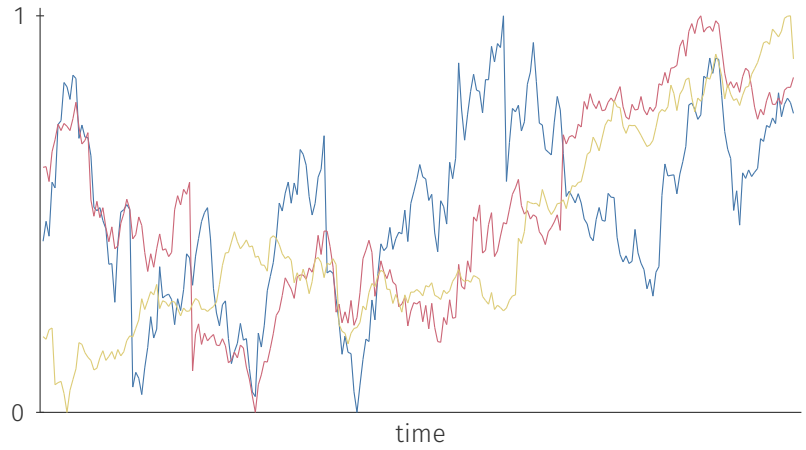
Les 2 types de normalisation : Illustration

IBM, Cisco and Apple stock prices from Sept. 2013 to Sept. 2014

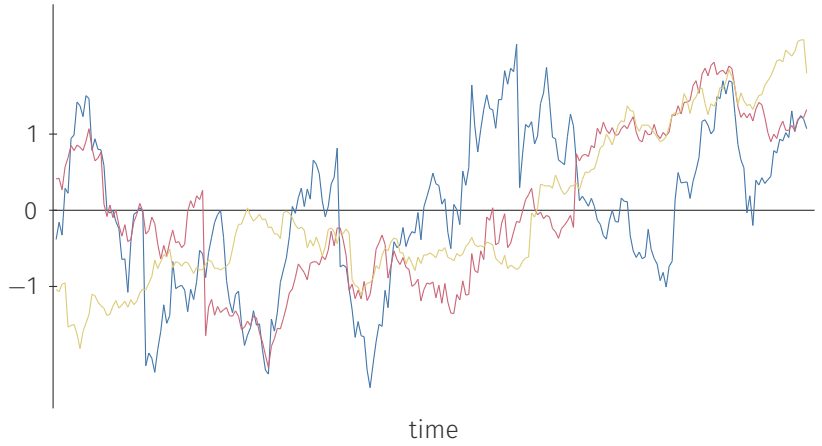
Original time-series



Range-based normalized time-series

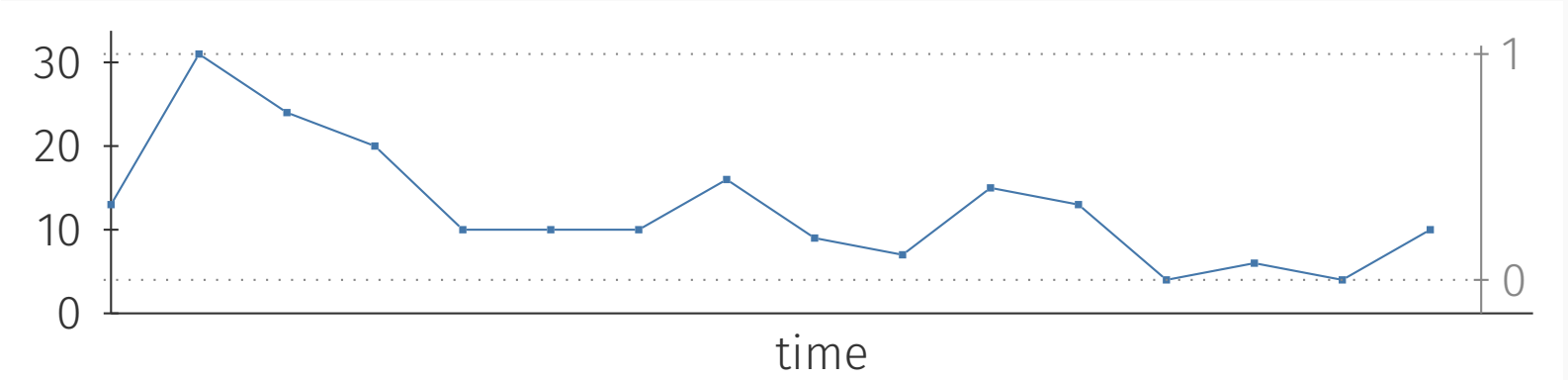


Standardized time-series



...

Normalisation sur intervalle



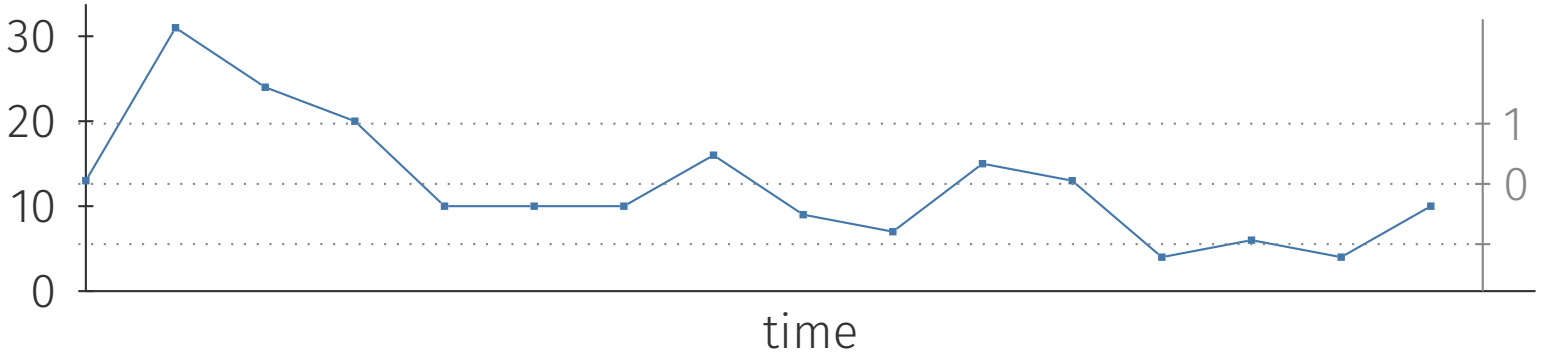
$\langle 13, 31, 24, 20, 10, 10, 10, 16, 9, 7, 15, 13, 4, 6, 4, 10 \rangle$

Range-based normalization

$$y_i = \frac{x_i - V_{min}}{V_{max} - V_{min}}$$

...

Standardisation



$$\langle 13, 31, 24, 20, 10, 10, 10, 16, 9, 7, 15, 13, 4, 6, 4, 10 \rangle$$

Standardization

$$\langle 0.05, 2.59, 1.60, 1.04, -0.37, -0.37, -0.37, 0.48, -0.51, -0.79, 0.34, 0.05, -1.22, -0.93, -1.22, -0.37 \rangle$$

...

Discrétisation

- Les séries temporelles à valeurs **numériques** peuvent être transformées en séries à valeurs **symboliques**

  - Abstraction de **valeur**
    - En divisant l'intervalle de valeurs numériques en catégories
    - E.g. {bas, moyen, élevé}

  - Abstraction de **tendance**
    - En examinant les tendances locales
    - E.g. {décroissant, stationnaire, croissant}

Discrétisation

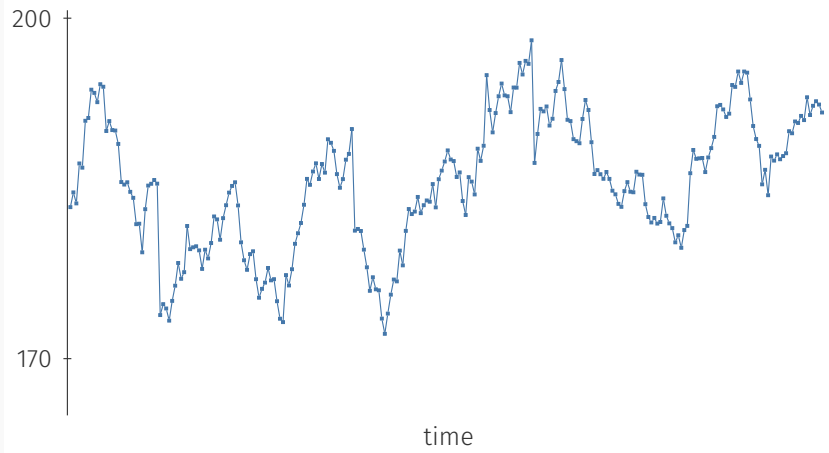
- Approximation symbolique agrégée
  (*Symbolic aggregate approximation* (SAX))
  - Prendre la **valeur moyenne** sur des intervalles successifs de même taille (calculer la « piecewise aggregate approximation » (PAA))
  - Convertir les valeurs résultantes en **valeurs discrètes** prises dans un petit ensemble de valeurs possibles
  - En prenant soin que les valeurs possibles aient une **fréquence d'apparition** à peu près égale

Remarque : SAX est une approche de « borne inférieure » : les mesures de distances calculées sur les séries recodées sont des distances inférieures à la distance dans la représentation originale

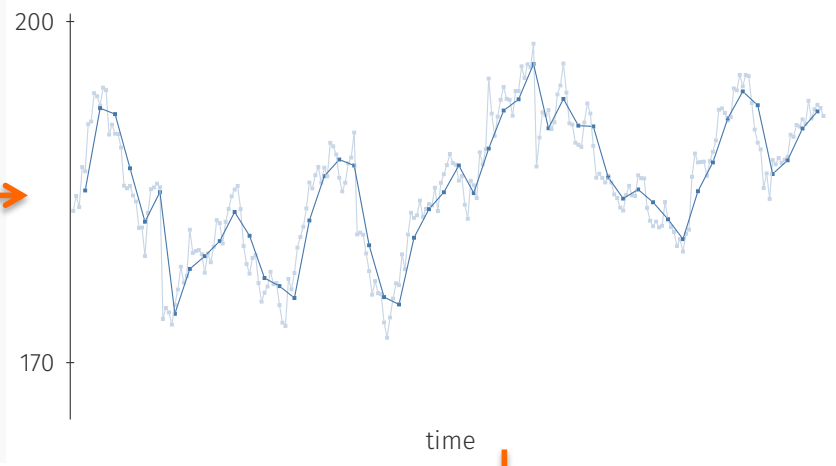
Discrétisation illustration

IBM stock prices from Sept. 2013 to Sept. 2014

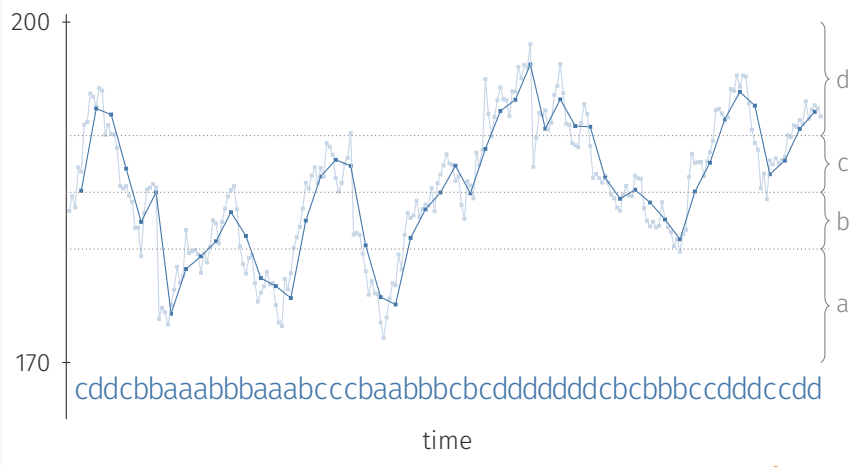
Original time-series



Binning, $k = 5$



Discretizing



...

Quelques **problèmes pratiques**

1. Motivation
2. Prétraitement des données
3. Valeurs manquantes : méthodes d'imputation
4. La réduction de dimension
   a. Par sélection de variables
   b. Par projection dans un nouvel espace
5. Points aberrants
6. Classes déséquilibrées
7. Conclusion

Les valeurs manquantes

Et leur imputation

Valeurs **manquantes**

Vache	Sexe	Age	Conso. 1er jour	Conso 2éme jour	Poids
Bérénice	F	3	3.56 kg	3.87 kg	630 kg
Marguerite	?	5	4.78 kg	?	?
Blanchette	F	?	?	5.72 kg	435 kg
Furie	M	6	6.02 kg	?	875 kg

- Valeurs souvent **indispensables**
- **Comment** faire ?

Pourquoi des valeurs manquantes

- Défaut de mesure
  - Défaut de transmission
  - Mesure en dehors des valeurs permises

  *Mesures manquantes de manière **aléatoire***

- Tous les attributs ne sont pas renseignés
  - Patients dans les hôpitaux

  *Mesures manquantes de manière **non aléatoire***

Types de sources de valeurs manquantes

1. Valeurs manquantes (non observées) mais **qui existent**
   - E.g. On n'a pas demandé son âge au patient

2. Valeurs manquantes car elles **n'auraient pas de sens**
   - E.g. Profession du conjoint quand célibataire

Ici, on se **focalise sur la classe (1)** de valeurs manquantes

Illustration : Ozone data

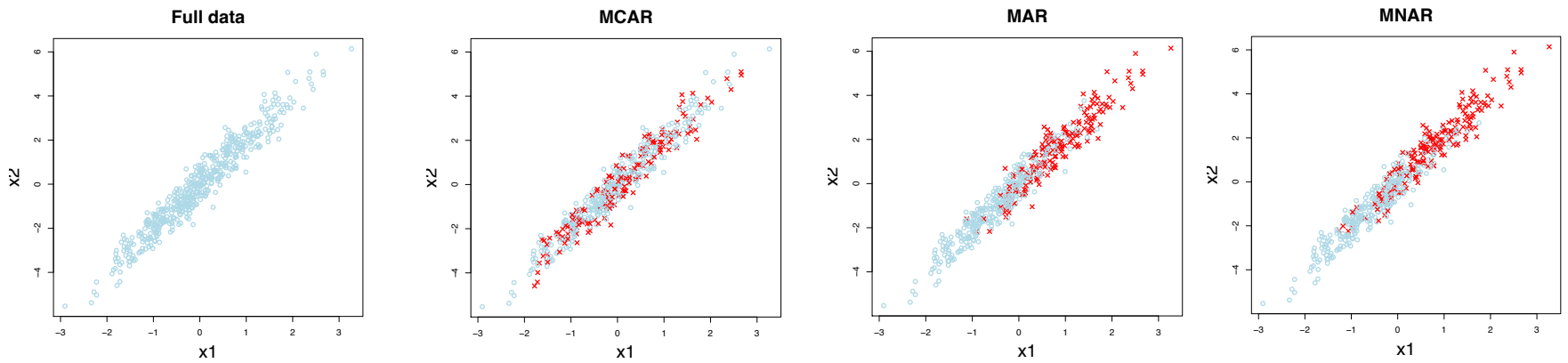
	maxO3	T9	T12	T15	Ne9	Ne12	Ne15	Vx9	Vx12	Vx15	maxO3v
0601	NA	15.6	18.5	18.4	4	4	8	NA	-1.7101	-0.6946	84
0602	82	17	18.4	17.7	5	5	7	NA	NA	NA	87
0603	92	NA	17.6	19.5	2	5	4	2.9544	1.8794	0.5209	82
0604	114	16.2	NA	NA	1	1	0	NA	NA	NA	92
0605	94	17.4	20.5	NA	8	8	7	-0.5	NA	-4.3301	114
0606	80	17.7	NA	18.3	NA	NA	NA	-5.6382	-5	-6	94
0607	NA	16.8	15.6	14.9	7	8	8	-4.3301	-1.8794	-3.7588	80
0610	79	14.9	17.5	18.9	5	5	4	0	-1.0419	-1.3892	NA
0611	101	NA	19.6	21.4	2	4	4	-0.766	NA	-2.2981	79
0612	NA	18.3	21.9	22.9	5	6	8	1.2856	-2.2981	-3.9392	101
0613	101	17.3	19.3	20.2	NA	NA	NA	-1.5	-1.5	-0.8682	NA
.	.	.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	.	.	
0919	NA	14.8	16.3	15.9	7	7	7	-4.3301	-6.0622	-5.1962	42
0920	71	15.5	18	17.4	7	7	6	-3.9392	-3.0642	0	NA
0921	96	NA	NA	NA	3	3	3	NA	NA	NA	71
0922	98	NA	NA	NA	2	2	2	4	5	4.3301	96
0923	92	14.7	17.6	18.2	1	4	6	5.1962	5.1423	3.5	98
0924	NA	13.3	17.7	17.7	NA	NA	NA	-0.9397	-0.766	-0.5	92
0925	84	13.3	17.7	17.8	3	5	6	0	-1	-1.2856	NA
0927	NA	16.2	20.8	22.1	6	5	5	-0.6946	-2	-1.3681	71
0928	99	16.9	23	22.6	NA	4	7	1.5	0.8682	0.8682	NA
0929	NA	16.9	19.8	22.1	6	5	3	-4	-3.7588	-4	99
0930	70	15.7	18.6	20.7	NA	NA	NA	0	-1.0419	-4	NA

http://www.airbreizh.asso.fr/

...

Trois types de mécanismes

1. MCAR (Missing Completely At Random)
   - Données perdues
2. MAR (Missing At Random)
   - E.g. une machine tombe en panne au-delà d'une certaine température
3. MNAR (Missing Not At Random)
   - E.g. un thermomètre ne fonctionne plus au-delà d'une certaine température



X_1 always observed
X_2 incomplete

1. MCAR
   - Les exemples sont représentatifs des données. **Pas de biais**.
2. MAR
   - Les exemples ne sont **pas représentatifs** des données. Des inférences valides peuvent être obtenues en **modélisant la matrice des données**.
3. MNAR
   - Les exemples ne sont **pas représentatifs**. Des inférences valides demandent de **modéliser le mécanisme de perte** de données.

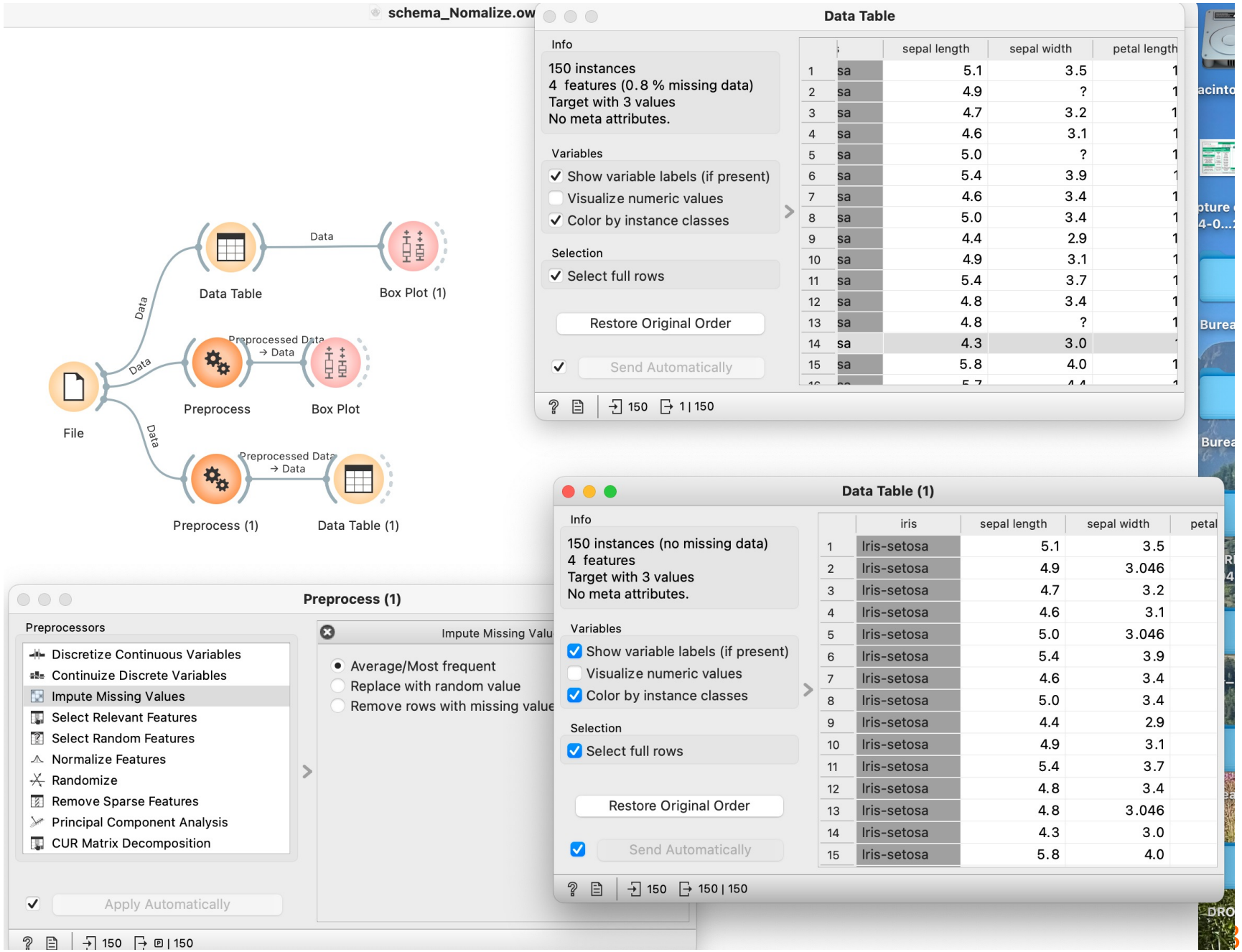
L'imputation des valeurs manquantes

- **Retrait** de l'attribut
  - **Retire de l'information**. Catastrophique si attribut important ou si beaucoup d'attributs sont touchés
- Par la **valeur moyenne** pour cet attribut
  - Peut modifier considérablement la distribution des valeurs, en particulier en **sous-estimant la variance**
- Par la **valeur moyenne** pour cet attribut dans **cette classe**
  - Peut propager des erreurs
- Par **interpolation**
  - Essayer de prédire la valeur manquante **à partir des autres attributs**
  - Ou à partir de **règles** : « si étudiant => âge = moins de 25 ans »

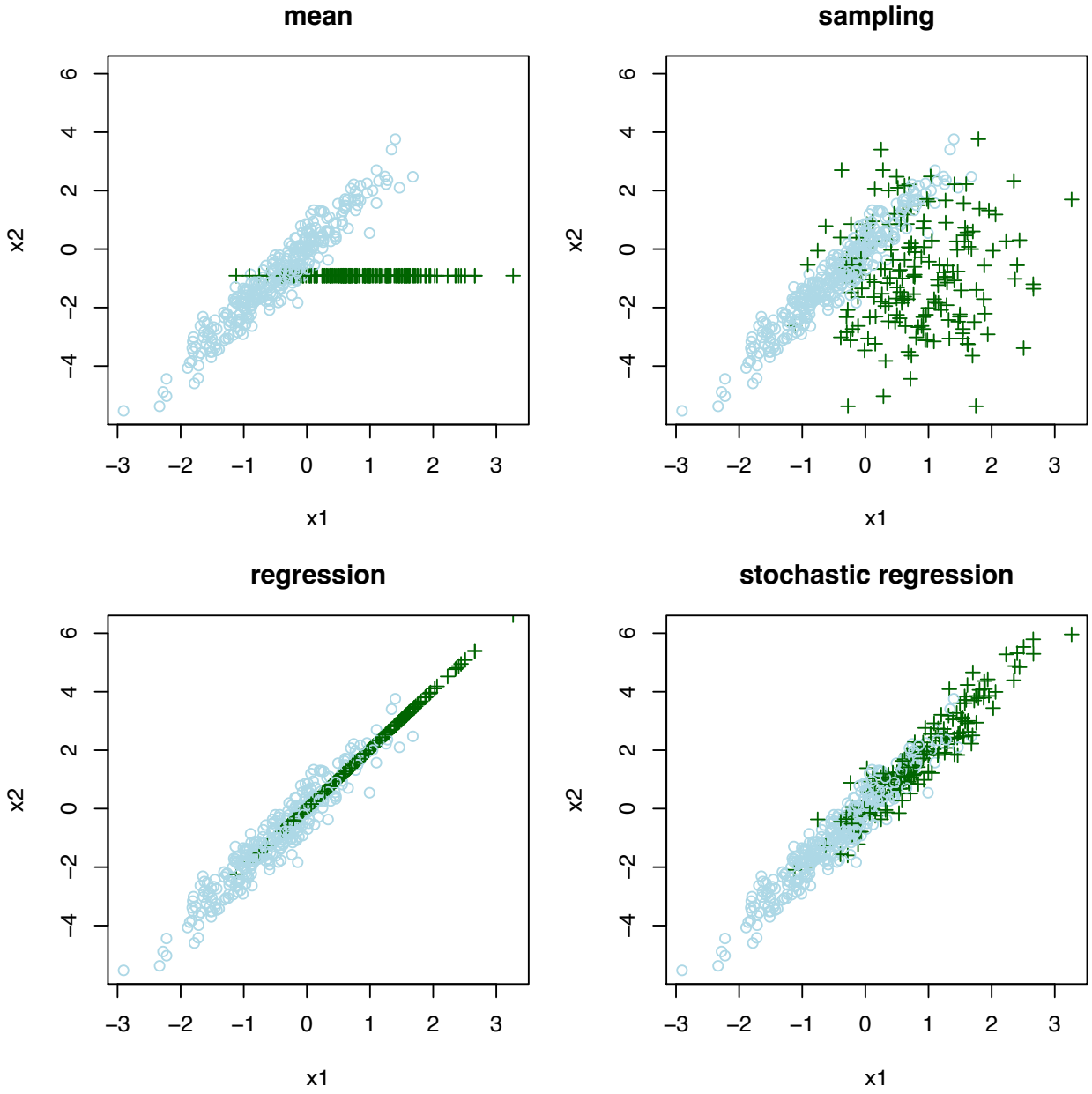
Valeurs **manquantes**

1. **Élimination** des données touchées

2. **Remplacement** par
   - valeur **la plus fréquente**
   - valeur calculée par **interpolation**
     - **Moyenne** des valeurs de la colonne
     - Interpolation **linéaire** (si cela a un sens)


schema_Nomalize.ow
Data Table
Info
150 instances
4 features (0.8 % missing data)
Target with 3 values
No meta attributes.
Variables
Show variable labels (if present)
Visualize numeric values
Color by instance classes
Selection
Select full rows
Restore Original Order
Send Automatically
sepal length
sepal width
petal length
1 sa 5.1 3.5 1
2 sa 4.9 ? 1
3 sa 4.7 3.2 1
4 sa 4.6 3.1 1
5 sa 5.0 ? 1
6 sa 5.4 3.9 1
7 sa 4.6 3.4 1
8 sa 5.0 3.4 1
9 sa 4.4 2.9 1
10 sa 4.9 3.1 1
11 sa 5.4 3.7 1
12 sa 4.8 3.4 1
13 sa 4.8 ? 1
14 sa 4.3 3.0
15 sa 5.8 4.0 1
150
1 | 150
Data
Data Table
Box Plot (1)
Preprocessed Data
→ Data
Preprocess
Box Plot
File
Preprocessed Data
→ Data
Preprocess (1)
Data Table (1)
Data Table (1)
Info
150 instances (no missing data)
4 features
Target with 3 values
No meta attributes.
Variables
Show variable labels (if present)
Visualize numeric values
Color by instance classes
Selection
Select full rows
Restore Original Order
Send Automatically
iris
sepal length
sepal width
petal
1 Iris-setosa 5.1 3.5
2 Iris-setosa 4.9 3.046
3 Iris-setosa 4.7 3.2
4 Iris-setosa 4.6 3.1
5 Iris-setosa 5.0 3.046
6 Iris-setosa 5.4 3.9
7 Iris-setosa 4.6 3.4
8 Iris-setosa 5.0 3.4
9 Iris-setosa 4.4 2.9
10 Iris-setosa 4.9 3.1
11 Iris-setosa 5.4 3.7
12 Iris-setosa 4.8 3.4
13 Iris-setosa 4.8 3.046
14 Iris-setosa 4.3 3.0
15 Iris-setosa 5.8 4.0
150
150 | 150
Preprocess (1)
Preprocessors
Discretize Continuous Variables
Continuize Discrete Variables
Impute Missing Values
Select Relevant Features
Select Random Features
Normalize Features
Randomize
Remove Sparse Features
Principal Component Analysis
CUR Matrix Decomposition
Impute Missing Valu
Average/Most frequent
Replace with random value
Remove rows with missing value
Apply Automatically
150
| 150

Exemples : valeur de x2 manquante (pour des valeurs de x1 > -1)



...

Méthodes d'imputation

- Paramétriques
  - E.g. régression stochastique
    - Attention au choix du modèle
- Non-paramétriques
  - E.g. kNN, random forest, …
    - Requiert un nombre important d'observations complètes
- Semi-paramétriques
  - E.g. prédictive mean matching

Un paradoxe

- **Si** l'on est capable d'inférer la valeur manquante d'une variable,

  c'est que les autres variables contiennent déjà cette information !!!

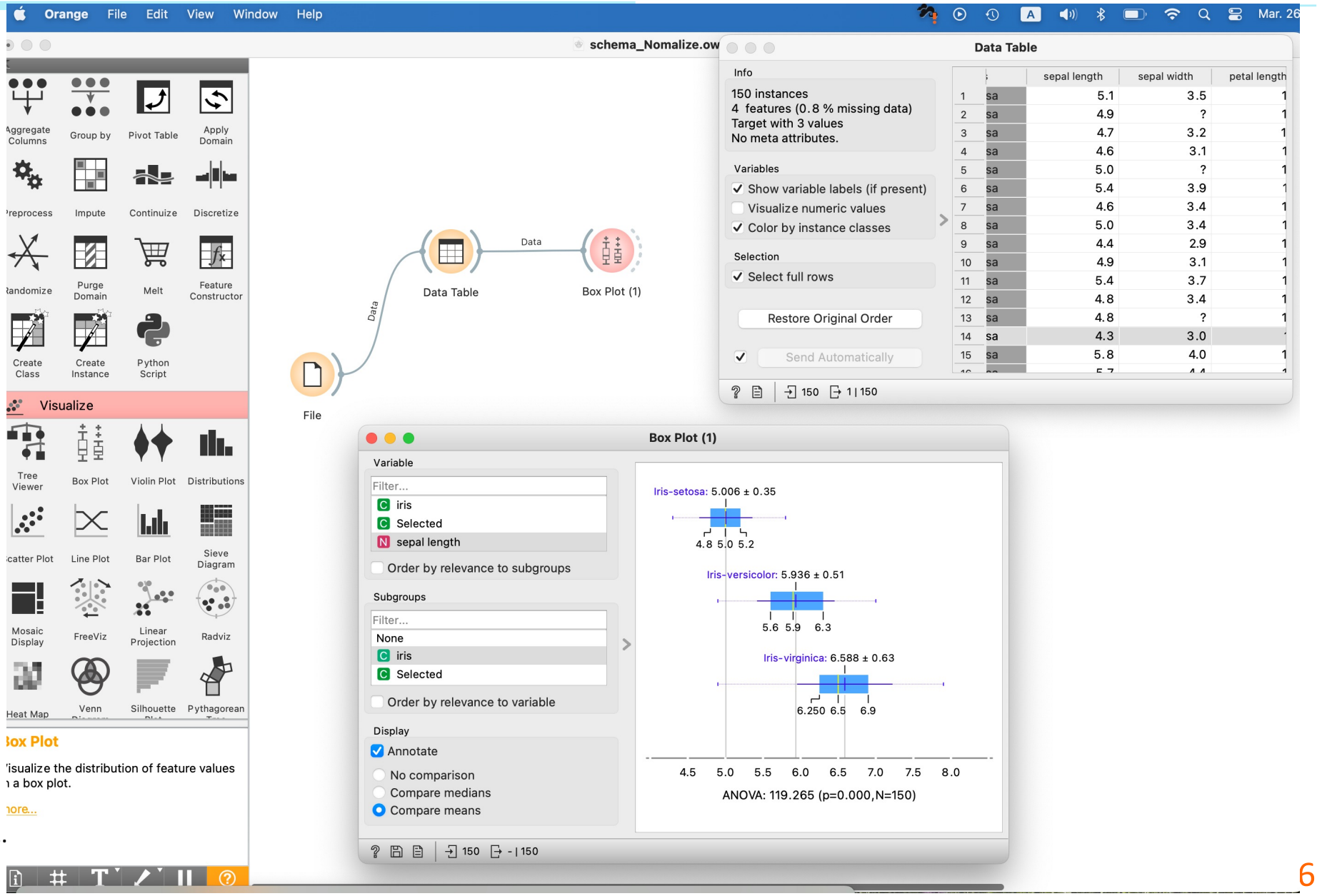
Imputation **unique** vs. Imputation **multiple**

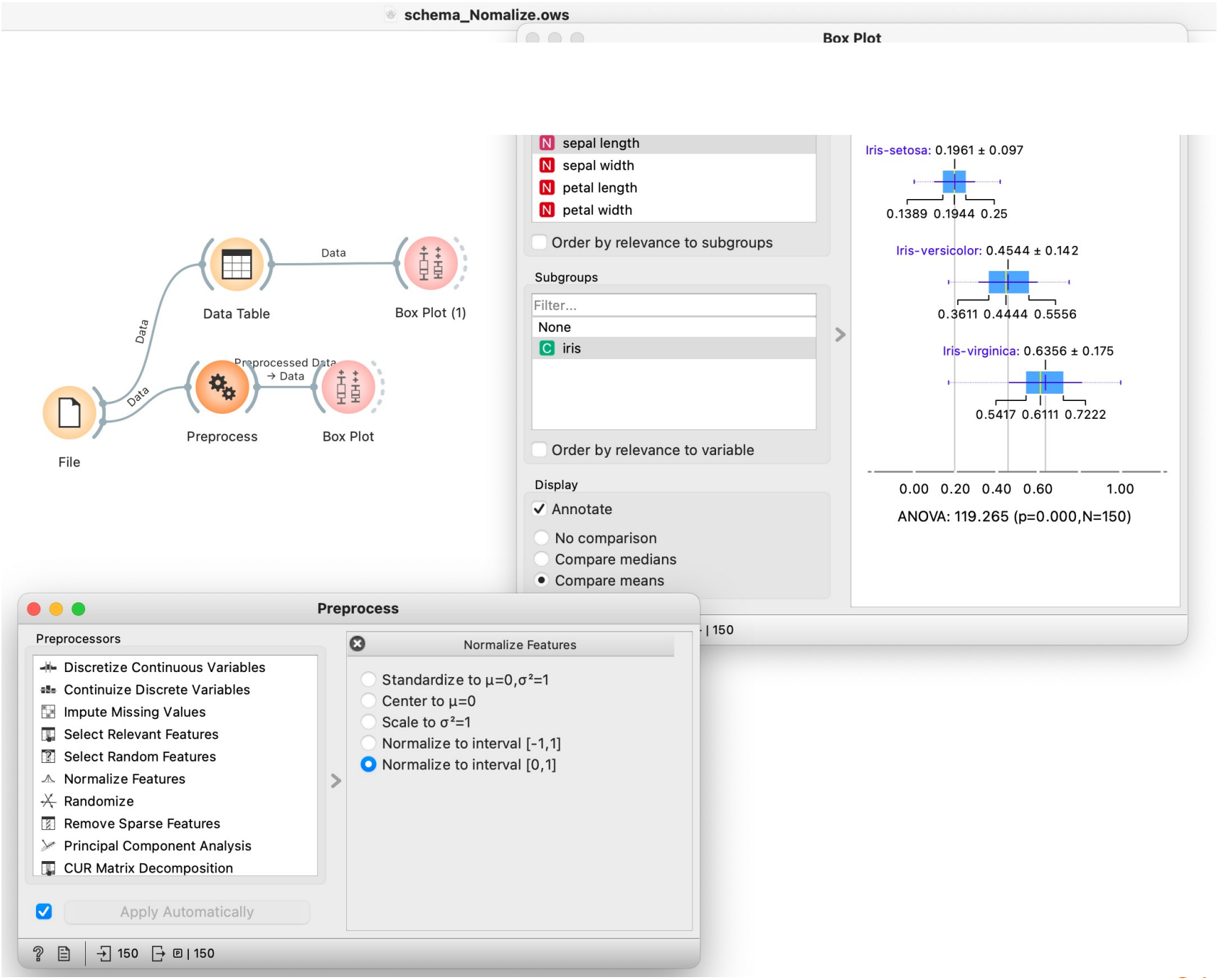
- Les **imputations uniques** n'estiment pas correctement la variance des résultats, même si l'espérance des valeurs imputées est correcte

- Des méthodes d'**imputation multiple** permettent de rendre compte de cette variance
  - Utiliser **plusieurs modèles** (paramétriques)
    ou **méthodes d'imputation** (en variant leurs hyper-paramètres)

    - Voir le tutoriel de Vincent Audigier à SFC-2024
    - https://vincentaudigier.weebly.com/uploads/1/7/3/1/17317324/public_sfc_2024.zip

Utilisation du logiciel Orange




schema_Nomalize.ows
Box Plot
Data
Data Table
Box Plot (1)
Data
Preprocessed Data
→ Data
Data
Preprocess
Box Plot
File
sepal length
sepal width
petal length
petal width
Order by relevance to subgroups
Subgroups
Filter...
None
iris
Order by relevance to variable
Display
Annotate
No comparison
Compare medians
Compare means
Iris-setosa: 0.1961 ± 0.097
0.1389 0.1944 0.25
Iris-versicolor: 0.4544 ± 0.142
0.3611 0.4444 0.5556
Iris-virginica: 0.6356 ± 0.175
0.5417 0.6111 0.7222
0.00 0.20 0.40 0.60 1.00
ANOVA: 119.265 (p=0.000,N=150)
| 150
Preprocess
Preprocessors
Discretize Continuous Variables
Continuize Discrete Variables
Impute Missing Values
Select Relevant Features
Select Random Features
Normalize Features
Randomize
Remove Sparse Features
Principal Component Analysis
CUR Matrix Decomposition
Normalize Features
Standardize to μ=0,σ²=1
Center to μ=0
Scale to σ²=1
Normalize to interval [-1,1]
Normalize to interval [0,1]
Apply Automatically
150
| 150

- Notebook
  - pandas_test_2.ipynb

Quelques **problèmes pratiques**

1. Motivation
2. Prétraitement des données
3. Valeurs manquantes : méthodes d'imputation
4. La réduction de dimension
   a. Par **sélection** de variables
   b. Par **projection** dans un nouvel espace
5. Points aberrants
6. Classes déséquilibrées
7. Conclusion

Les données en **grandes dimensions**

- Des **phénomènes surprenants**
  - Sur les **distances** entre points

- Obstacle à l'**interprétation**
  - **Visualisation**
  - Recherche de **régularités**
    - Risque accru de trouver des **corrélations « accidentelles »**
      - Propres à ce jeu de données

Approches

- Par sélection d'attributs
  - Comment faire ?

- Par projection dans un nouvel **espace de plus petite dimension**
  - Analyse en **composantes**
    - **Principales** (ACP)
    - Pour des valeurs catégorielles : ACM
    - Indépendantes (ACI)
    - ...
  - Projections **non linéaires**
    - **t-SNE**, IsoMap, Locally Linear Embedding, ...

Quelques **problèmes pratiques**

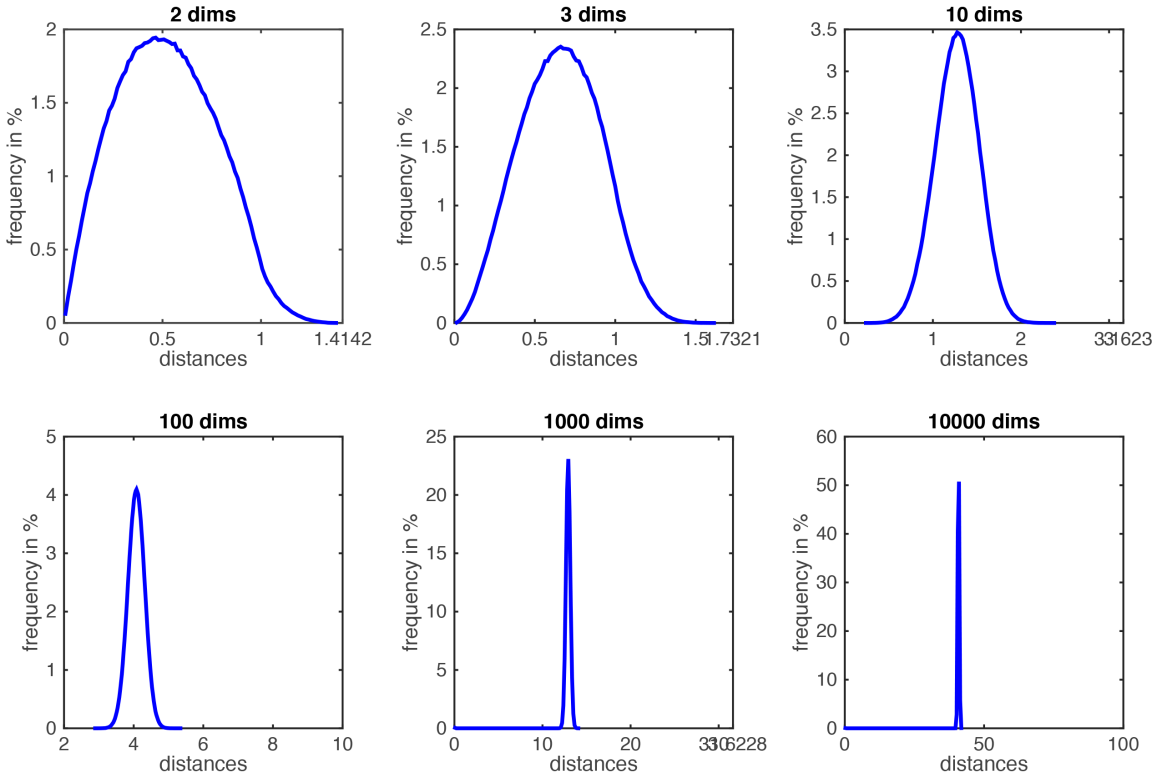
1. Motivation
2. Prétraitement des données
3. Valeurs manquantes : méthodes d'imputation
4. La réduction de dimension
   a. Par **sélection** de variables
   b. Par **projection** dans un nouvel espace
5. Points aberrants
6. Classes déséquilibrées
7. Conclusion

Réduction de l'espace d'entrée

- Attributs **redondants**
- Attributs **non pertinents**
- Attributs **peu informatifs**

Réduction de l'espace de description

- La malédiction de la dimensionnalité
  - En stat : il faut un **nombre exponentiel de données** (en la dimension de l'espace) pour approcher une distribution
  - En apprentissage fondé sur les distances : un **phénomène de concentration**



Distributions des distances de paires de points tirées aléatoirement dans le cube unité en fonction de d

Réduction de l'espace de description

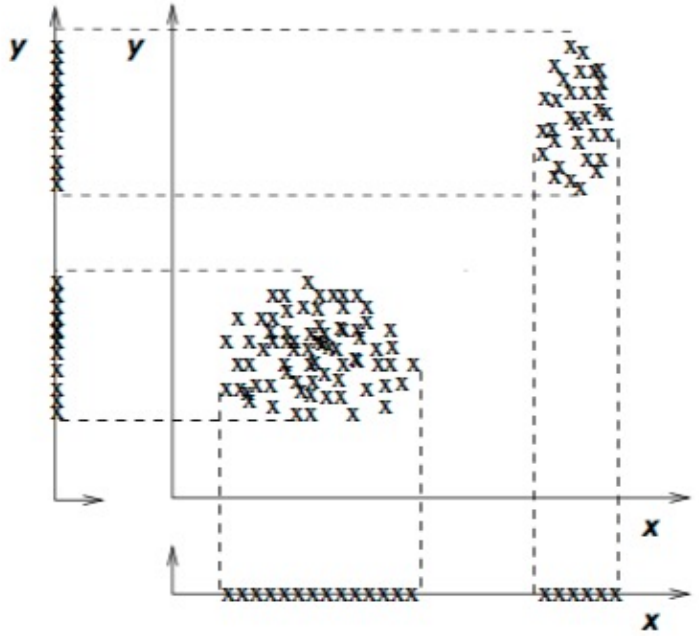
Deux grandes classes de méthodes

1. **Sélection** de variables
   - Le plus souvent en contexte d'apprentissage supervisé
   - Exemples
     - ANOVA
     - RELIEF
     - Méthodes de régularisation : Ridge, LASSO, Elastic Net, ...

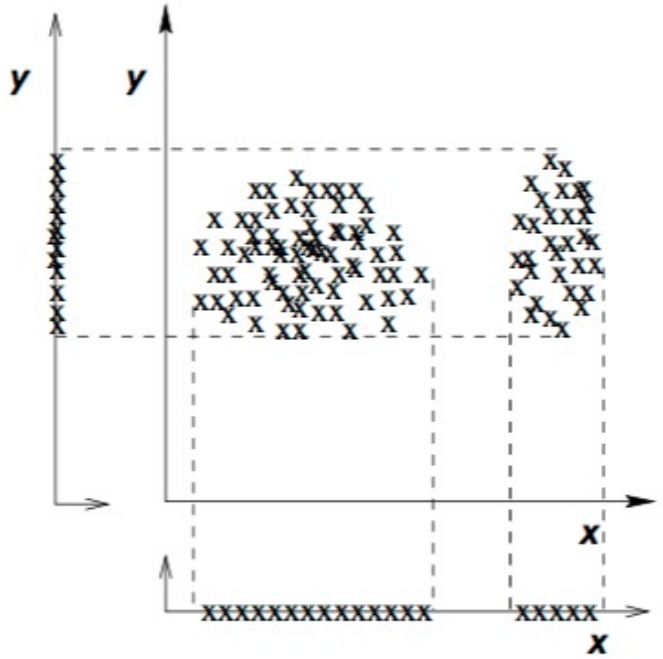
2. **Changement** d'espace de représentation
   - **Analyses en composantes** : principales (ACP), indépendantes (ACI), composantes non négatives (NMF), Analyse sémantique latente (LSA), ...
   - **Manifold learning** : ISOMAP, tSNE

Sélection d'attributs

Attributs **redondants** ou **non informatifs**



Attributs **redondants** : Les deux attributs apportent la même information



y **non informatif** : il ne permet pas de distinguer les deux clusters

From [J. DY & C. Brodley (2004) « Feature selection for unsupervised learning », JMLR 5 (2004) 845-889]

Mesure de pertinence des attributs

- **Pertinence** ⟷ apport d'**information sur l'objectif**

  - **Description** : conserver l'**information sur la forme de la distribution** des exemples
    - Dépend d'hypothèses faites *a priori* sur la distribution

  - **Classification** : isoler l'**information** permettant de **classer les exemples et ceux à venir**
    - Peut résulter d'un apprentissage

Sélection de variables

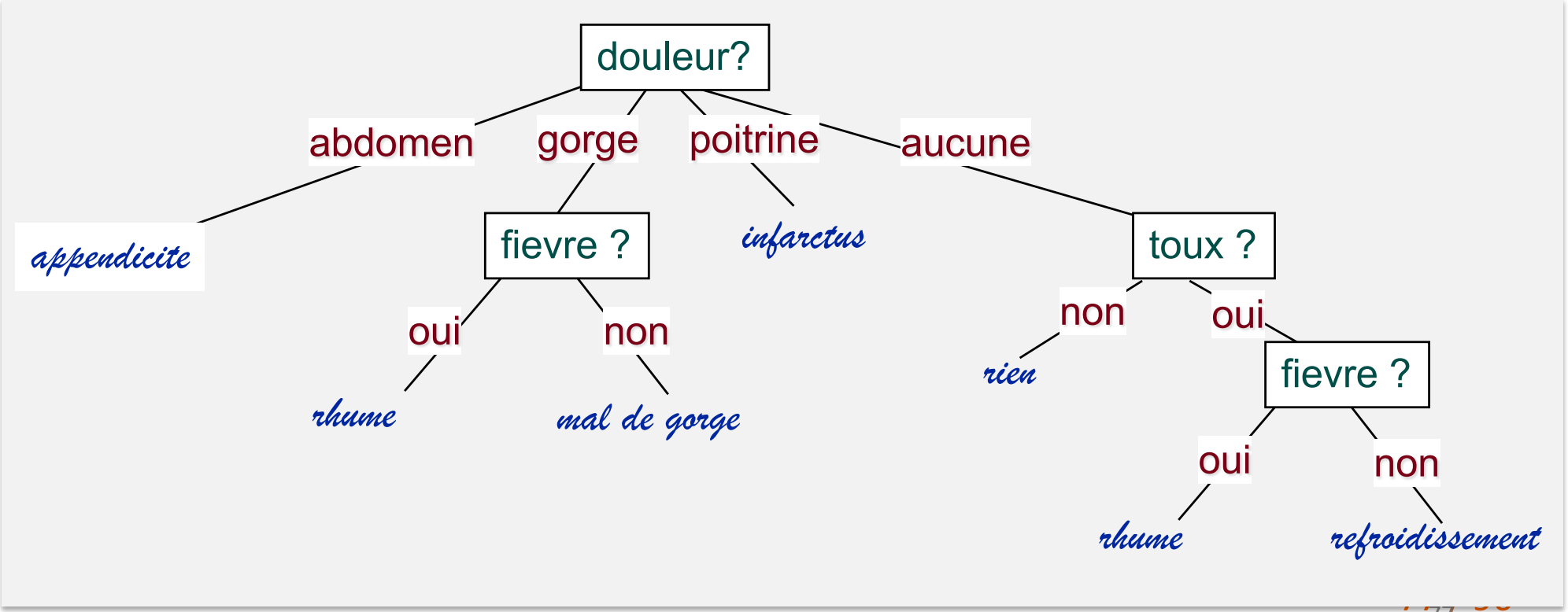
1. Méthodes **intégrées** (*embedded*)
   - La méthode d'apprentissage **sélectionne** les descripteurs **par elle même**
   - E.g. Arbres de décision

2. Méthodes « **symbiose** » (*wrapper*)
   - **Exploration des sous-ensembles de descripteurs** en observant la performance d'une méthode de classification

3. Méthodes de **filtre** (*filter*)
   - Sélection des variables **indépendamment** les unes des autres
   - ANOVA, RELIEF, ...

Méthodes « intégrées »

- L'apprentissage conduit à distinguer les attributs pertinents
  - Ex : arbres de décision



Méthode « symbiose » (*wrapper*)

- On utilise un algorithme d'apprentissage pour **évaluer la pertinence de l'ensemble d'attributs** sélectionné

  - Avantage : on optimise bien ce que l'on veut optimiser
  - Inconvénient : espace de recherche gigantesque
    - Il faut avoir recours à une exploration heuristique
      - Ascendante (« *forward selection* »)
      - Descendante (« *backward selection* »)
    - Coûteux

Classification : Méthodes de filtres

- Beaucoup de techniques
  - **ANOVA** (ANalysis Of VAriance)
    - Comparer la variabilité intra-classe à la variabilité inter-classe
    - Méthode paramétrique : suppose des lois normales

$$S^2_{int} = \frac{SC_{int}}{\sharp \text{ degrés de liberté}} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i^2)}{n-k}$$

$$S^2_{ent} = \frac{\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2}{k-1}$$

Plus S_{ent} grand par rapport à S_{int}, moins la variance peut-être expliquée par la variabilité de l'échantillonnage et plus c'est indicatif d'un effet classe.

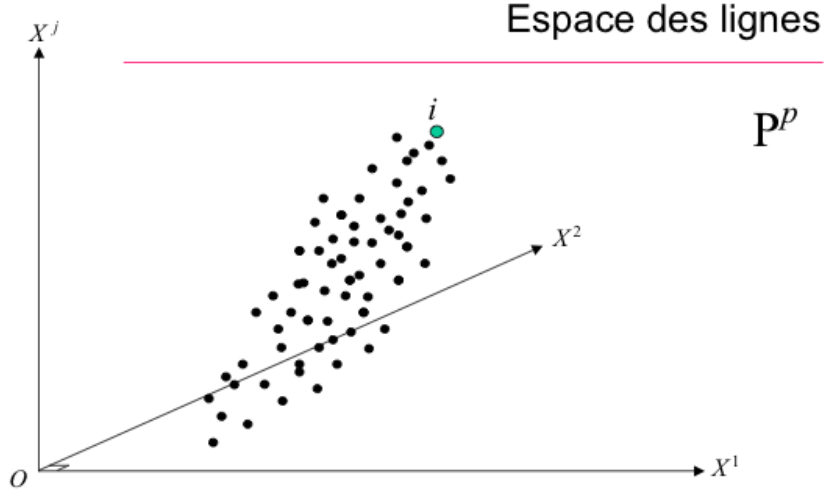
  - **RELIEF**
    - Principe sous-jacent similaire : mesurer en quoi la connaissance de l'attribut apporte une information sur la classe des exemples
    - Méthode non paramétrique

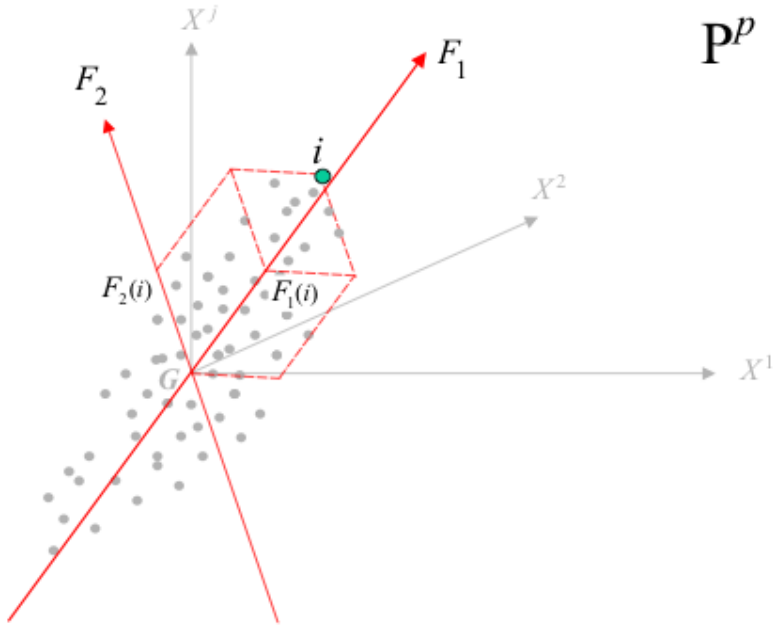
Quelques **problèmes pratiques**

1. Motivation
2. Prétraitement des données
3. Valeurs manquantes : méthodes d'imputation
4. La réduction de dimension
   a. Par **sélection** de variables
   b. Par **projection** dans un nouvel espace
5. Points aberrants
6. Classes déséquilibrées
7. Conclusion

Projection dans nouvel espace

L'Analyse en Composantes Principales





- Méthode linéaire
- Optimisant un critère quadratique
  - Très sensible aux points aberrants

L'Analyse en Composantes Principales

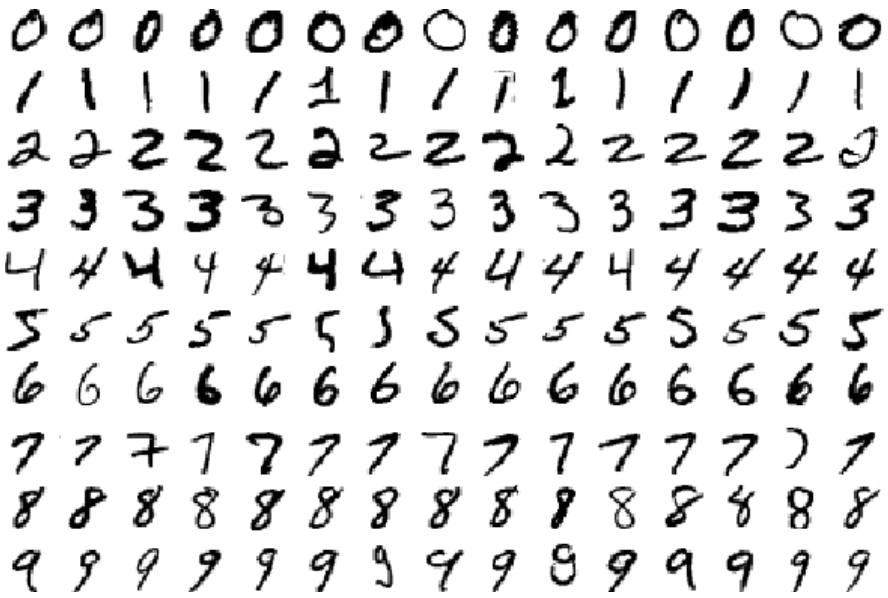
Calcule les **composantes principales** (axes orthogonaux)

Préservant l'essentiel de l'inertie (la variance) du jeu de données

```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
Created from 150 samples and 5 variables

Pre-processing:
  - centered (4)
  - ignored (1)
  - principal component signal extraction (4)
  - scaled (4)

PCA needed 2 components to capture 95 percent of the variance

       Species         PC1                PC2
 setosa    :50   Min.   :-2.7651   Min.   :-2.67732
 versicolor:50   1st Qu.:-2.0957   1st Qu.:-0.59205
 virginica :50   Median : 0.4169   Median :-0.01744
                 Mean   : 0.0000   Mean   : 0.00000
                 3rd Qu.: 1.3385   3rd Qu.: 0.59649
                 Max.   : 3.2996   Max.   : 2.64521
```

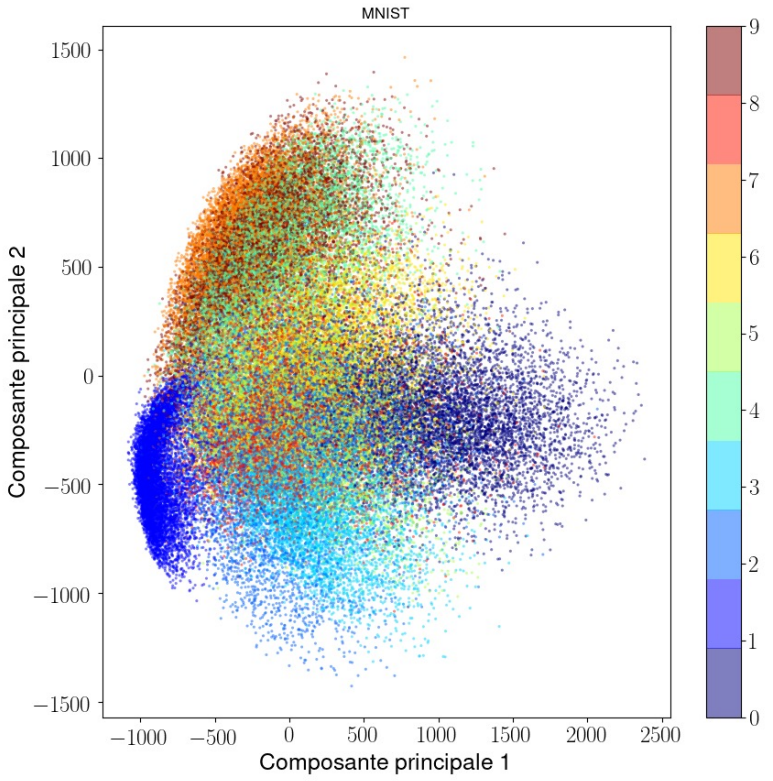
Analyse en **composantes principales**

- Identifier les axes principaux d'inertie
- Limite : méthode linéaire



d = 584

60 000 images



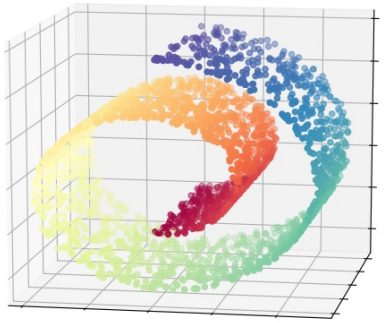
d = 2

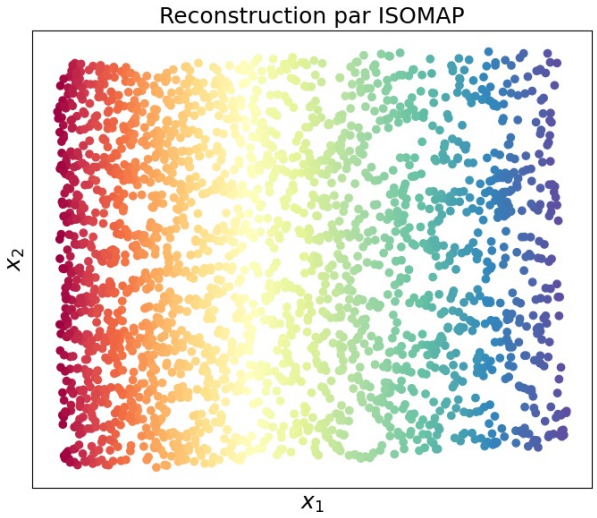
Analyse en Composantes Principales (ACP)

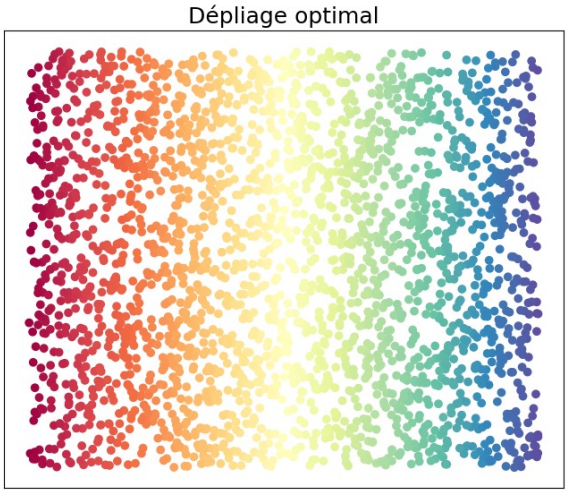
...

Cas **non linéaire** : ISOMAP

- Tente de préserver les propriétés de la variété sous-jacente en identifiant les paires de points proches et en reportant leur distance euclidienne

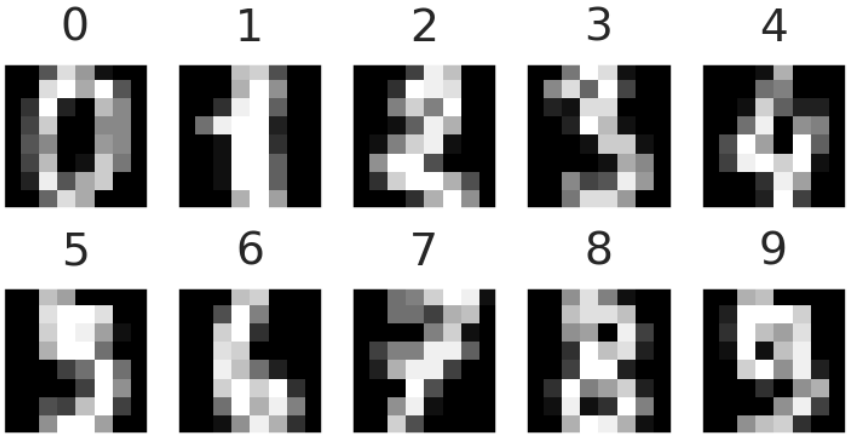




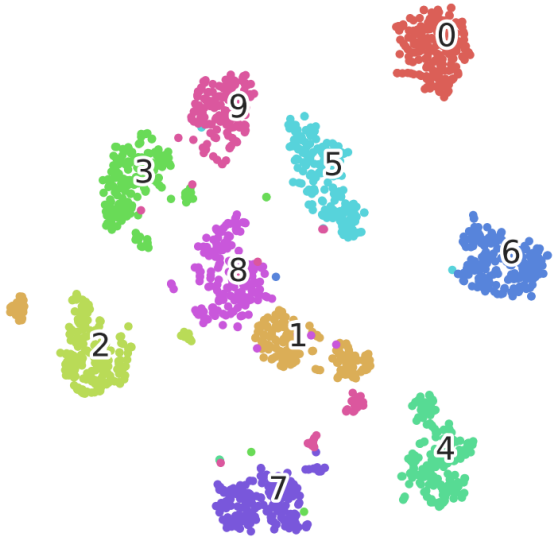


Cas **non linéaire** : tSNE

- À partir de calcul de probabilités conditionnelles
- Projette en **dimension 2** ou **3**
- Complexité en m^2 : donc recourir à des optimisations (m : nb d'exemples)

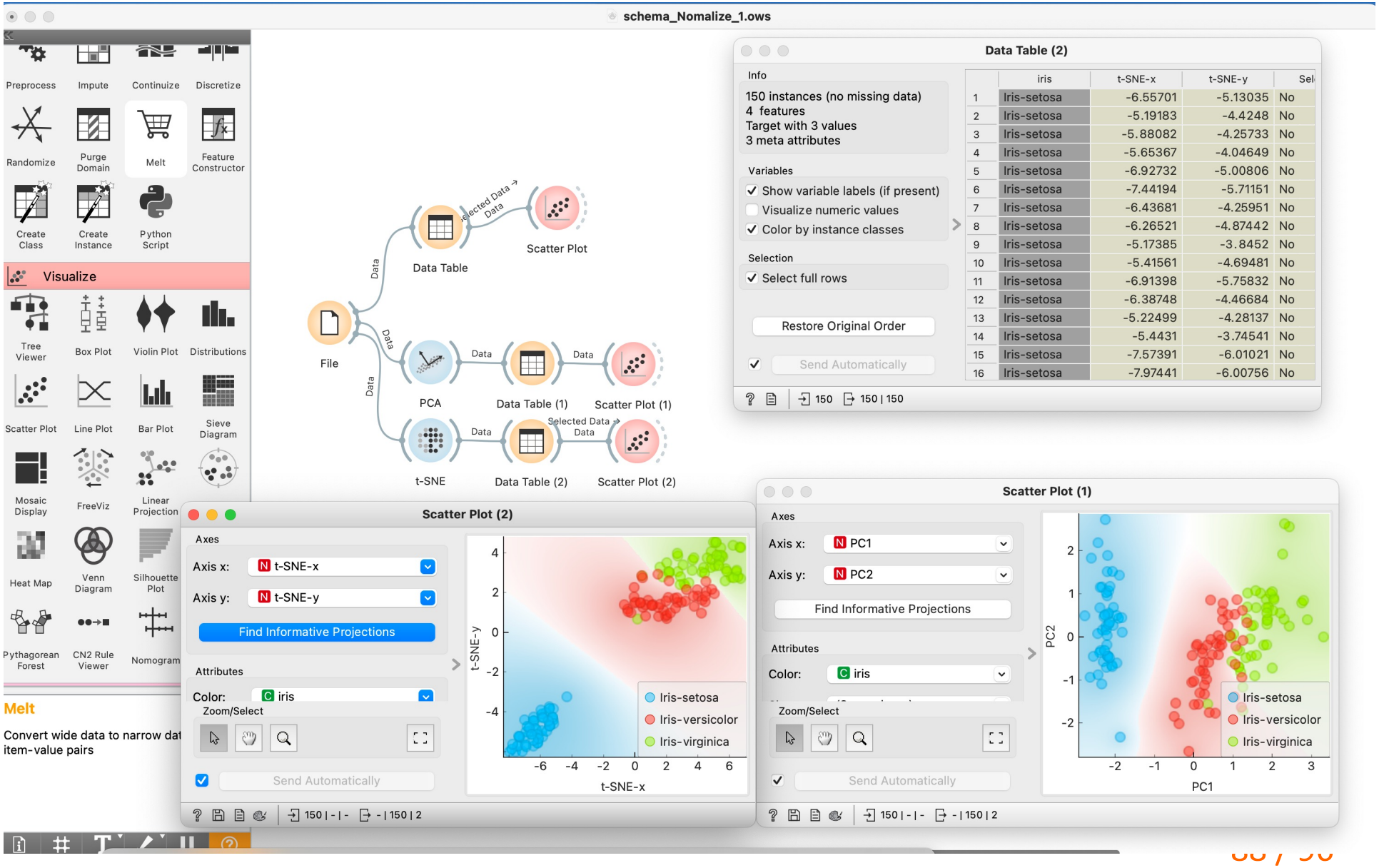


d = 8 x 8 = 64



d = 2

Utilisation de t-SNE



- Notebook
  - demo_pca_tsne_umap.ipynb

Quelques **problèmes pratiques**

1. Motivation
2. Prétraitement des données
3. Valeurs manquantes : méthodes d'imputation
4. La réduction de dimension
   a. Par sélection de variables
   b. Par projection dans un nouvel espace
5. Points aberrants
6. Classes déséquilibrées
7. Conclusion

Données **aberrantes**

- **Valeurs** aberrantes
  - Ex : âge = 123 ans ; tél : 999-999-999
  - Détection par comparaison à contraintes d'intégrité

- **Points** aberrants
  - Sur plusieurs dimensions
  - Peuvent être de vrais points mais faussant les statistiques

Les **points aberrants**

- Des **erreurs** ou des **exceptions**
  - Exemples :
    - salaires des diplômés Agro à la sortie
    - Fraudes
    - Attaques sur réseau

- Méthodes
  - Expertise
  - Contraintes d'intégrité
  - Détection par méthode d'apprentissage
    - One-class SVM
    - Boosting
    - ...

Quelques **problèmes pratiques**

1. Motivation
2. Prétraitement des données
3. Valeurs manquantes : méthodes d'imputation
4. La réduction de dimension
   a. Par sélection de variables
   b. Par projection dans un nouvel espace
5. Points aberrants
6. Classes déséquilibrées
7. Conclusion

Données **déséquilibrées** selon les classes

1. **Sous-échantillonner** la classe majoritaire
   - Pas adapté si peu de données

2. Créer des **exemples fictifs** de la classe minoritaire
   - E.g. par bruitage des exemples existants

3. **Modifier les coûts** de mauvaise classification
   - E.g. coût(faux positif) >> coût(faux négatif)

Quelques **problèmes pratiques**

1. Motivation
2. Prétraitement des données
3. Valeurs manquantes : méthodes d'imputation
4. La réduction de dimension
   a. Par sélection de variables
   b. Par projection dans un nouvel espace
5. Points aberrants
6. Classes déséquilibrées
7. Conclusion

À retenir

1. Les prétraitements sont **essentiels**
   - **Nettoyage** des données
     et **changements** de représentation
     sont **essentiels** pour les traitements ultérieurs

2. **Nombreux** problèmes et nombreuses techniques

3. Demande **réflexion** et **soin**