Explorer le monde des données

Les prétraitements

Antoine Cornuéjols & Christine Martin

AgroParisTech – INRAé MIA-Paris-Saclay

EKINOCS research group

antoine.cornuejols,christine.martin@agroparistech.fr

AgroParisTech
Talents d'une planète soutenable

INRAE

Les prétraitements

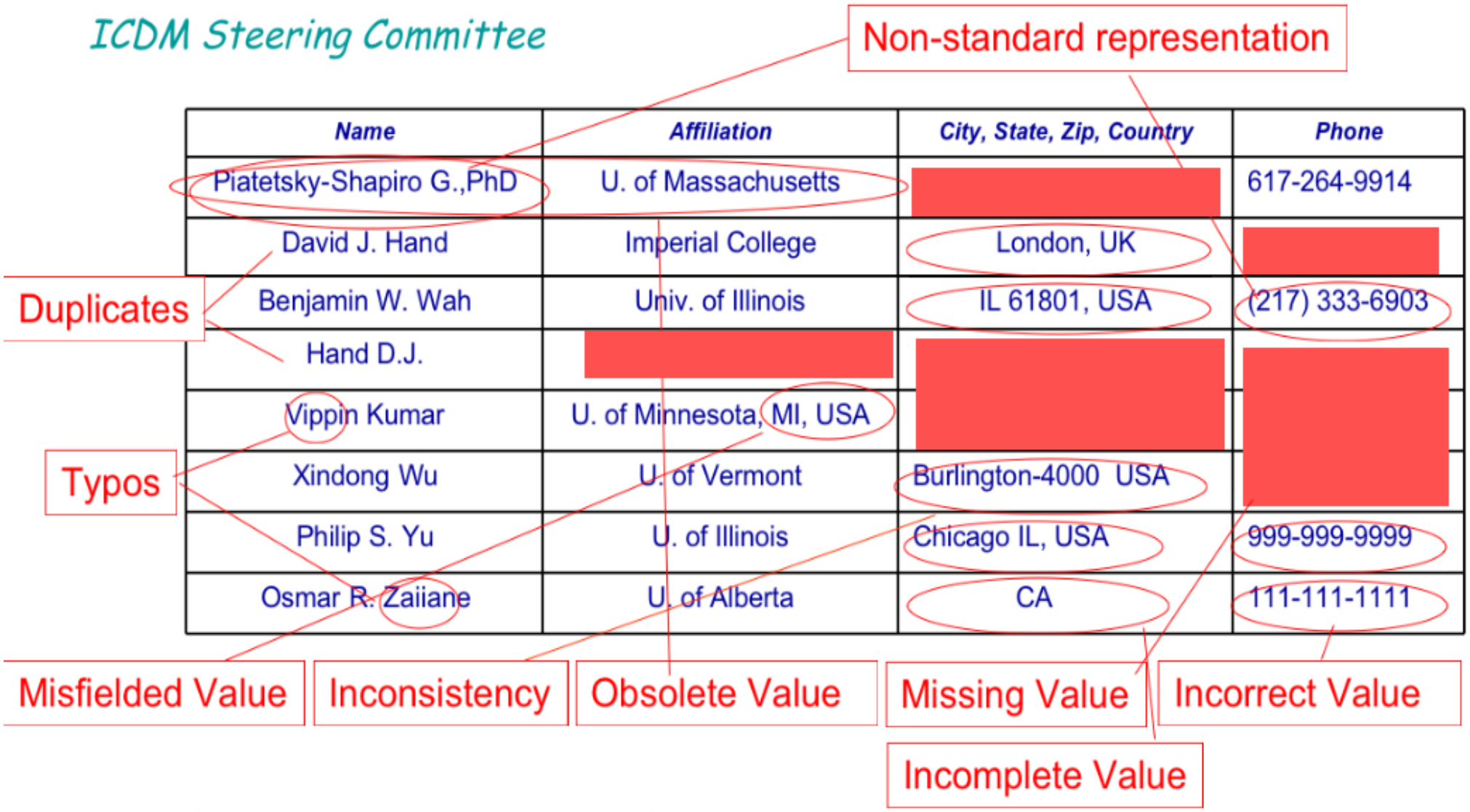
Les « données »

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

Introduction : les problèmes potentiels

- **Intrinsèques 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

La qualité des données



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
  - Détection par l'hypothèse apprise
    - Voir aussi le Boosting

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

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**. Coûte énormément

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.

Les prétraitements

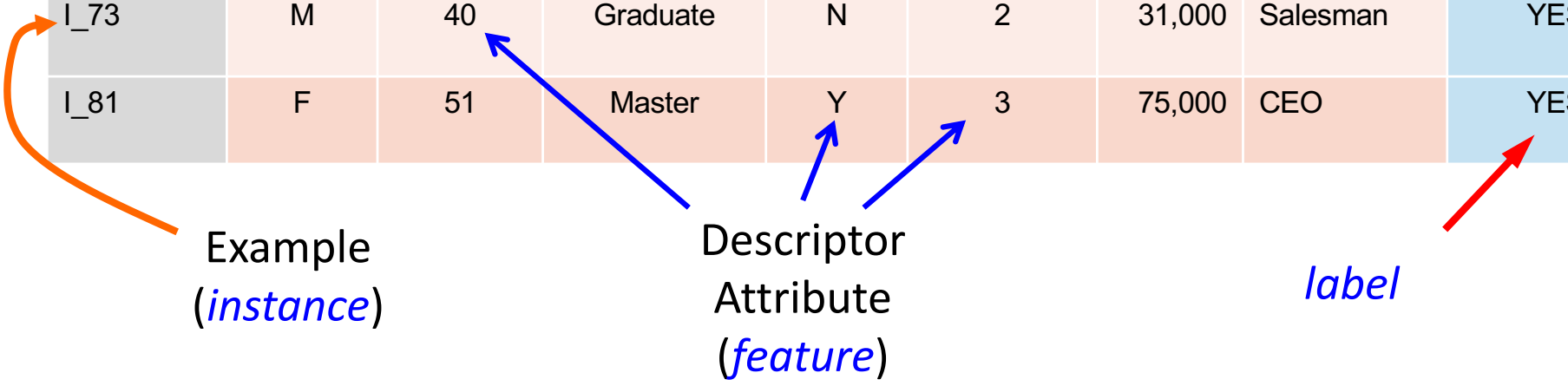
1. Motivation
2. Types de prétraitements
3. Illustration
4. La réduction de dimension
    1. Par sélection de variables
    2. Par projection dans un nouvel espace
5. Prétraitements pour les séries temporelles
6. Conclusion

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

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 of inputs

- Vectors
- Sequences
- Structured
- Temporal
- Spatial

Types of inputs

- **Vectors**
- Sequences
- Structured
- Temporal
- Spatial

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

Example (*instance*)

Descriptor Attribute (*feature*)

label

Types of inputs

- Vectors
- **Sequences**
- Structured
- Temporal
- Spatial

Protéine « sp|P00004|CYC_HORSE » is activated by …

```
 1   ttcagttgtg aatgaatgga cgtgccaaat agacgtgccg ccgccgctcg attcgcactt
61   tgctttcggt tttgccgtcg tttcacgcgt ttagttccgt tcggttcatt cccagttctt
121 aaataccgga cgtaaaaata cactctaacg gtcccgcgaa gaaaaagata aagacatctc
181 gtagaaatat taaaataaat tcctaaagtc gttggtttct cgttcacttt cgctgcctgc
 …
4021 agaacacgcc gaggctccat tcatagcacc acttcgtcgt cttaatcccc tccctcatcc
4081 gccatggcgg tgcaaaaaat aaaaagaact c
```

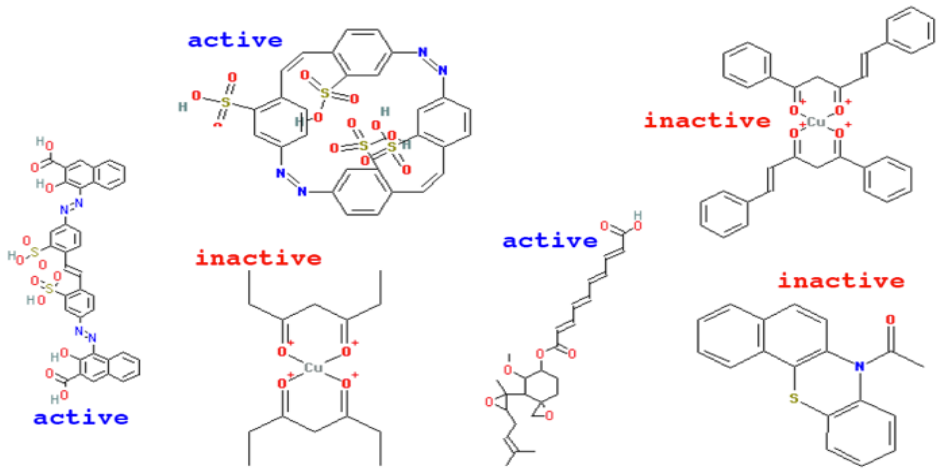
```
DEVICE=eth0
BOOTPROTO=none
ONBOOT=yes
IPADDR=192.168.0.X
NETMASK=255.255.255.O
GETEWAY=192.168.0.254
search exemple.com namserver
192.168.0.254
```

Types of inputs

- Vectors
- Sequences
- **Structured**
- Temporal
- Spatial

1st order logic

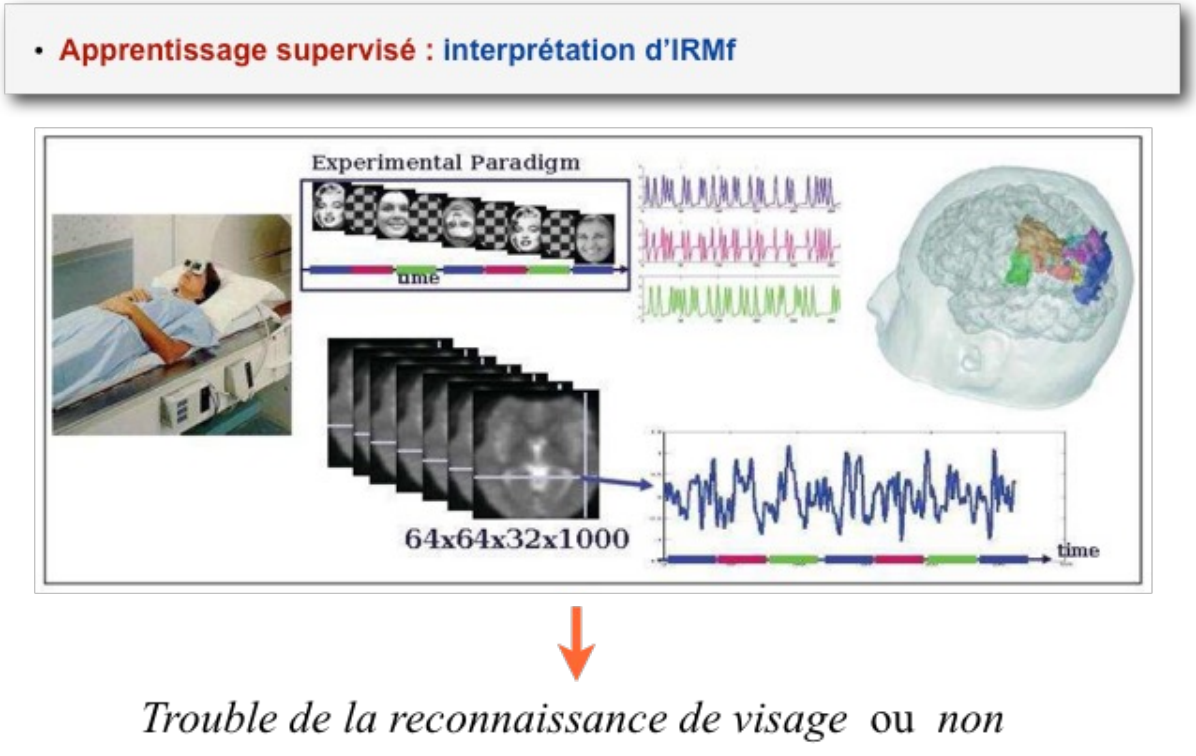
block(B1) & ontable(B2) & above(B1,B2) & …



NCI AIDS screen results (from http://cactus.nci.nih.gov).

Types of inputs

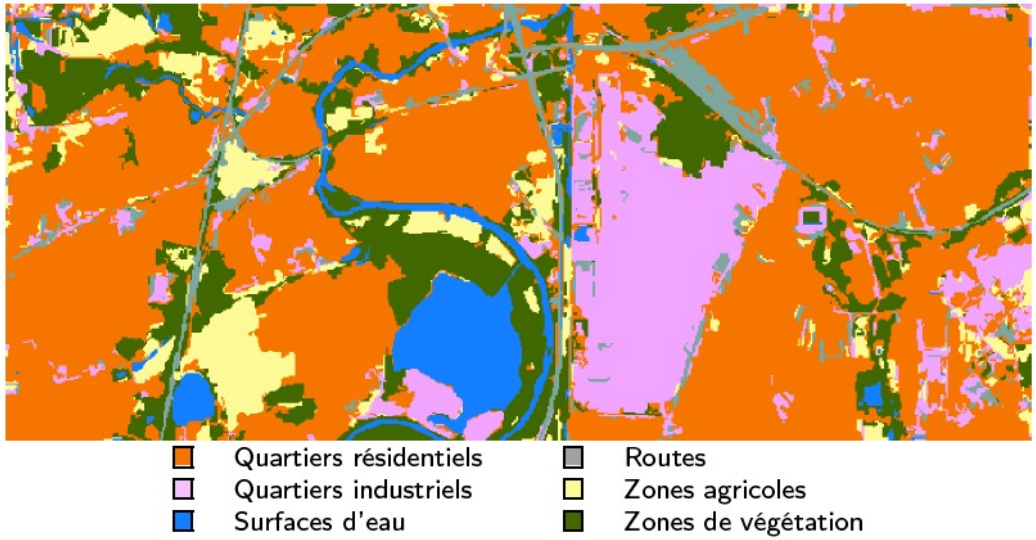
- Vectors
- Sequences
- Structured
- **Temporal**
- Spatial



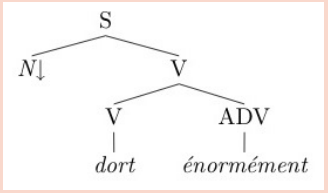
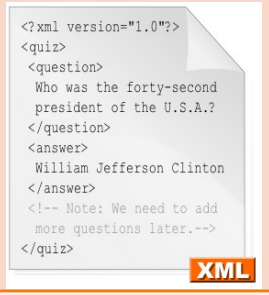
Types of inputs

- Vectors
- Sequences
- Structured
- Temporal
- **Spatial**

Image MRS



Formats

Numerical: continuous (R)	Bank account : 12,915.86 €
Numerical discreet (N ou Z)	Number of dependents : 11
Binary	Single: True
Category	Color in {rouge, vert, bleu}
Text	Protein « sp|P00004|CYC_HORSE » is activated by …
Structured data	Tree, XML expression, …  
Sequences	- Genome - Sequence of queries on a web site
Images, videos	

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

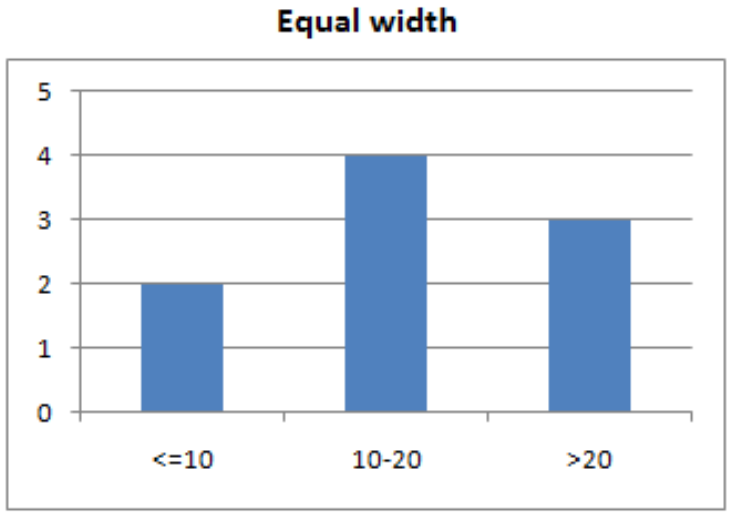
- Ensemble
  - Avec proximité sémantique -> Séquence d'entiers
    - Ex : [rouge, vert, bleu] -> [1,2,3]
  - Sans proximité sémantique
    - Ex : Codage disjonctif complet

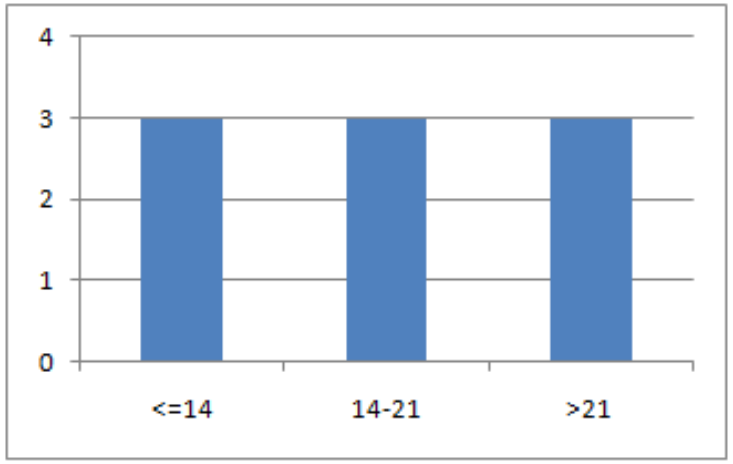
Les données numériques

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,+)





Numérique -> symbolique

- Discrétisation des domaines de variation
  - Division **uniforme** du domaine
  - Division **par sous-populations** de tailles égales
    - Ex : motifs fréquents (algorithme Apriori)
  - Division **par seuil** entre classes
    - Ex : Induction d'arbres de décision (C4.5)

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. Calcule la **déviation standard**
  2. Le **divise** 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. Calcule la **moyenne**
  2. La **soustrait**

```
  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

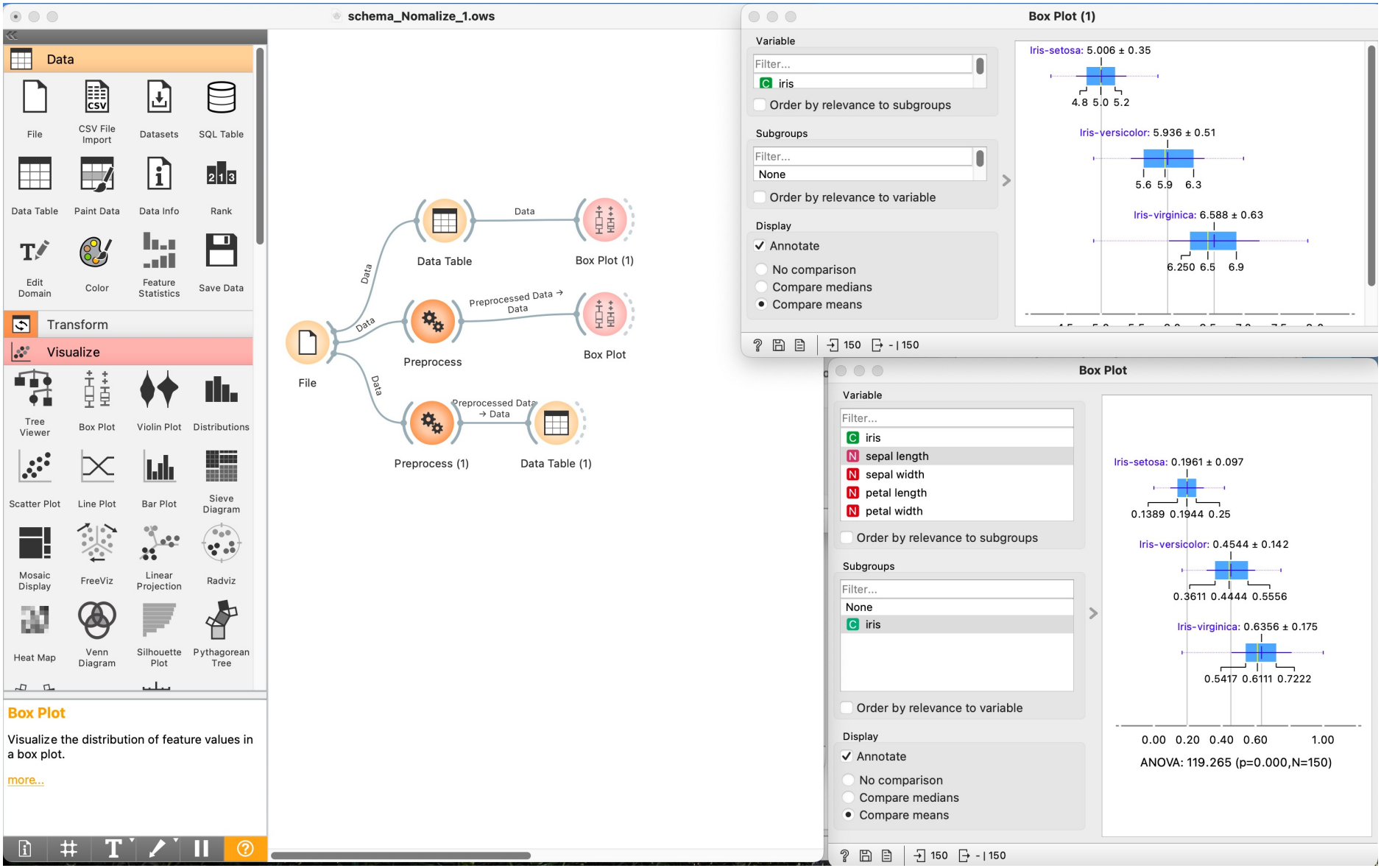
```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
 Median :5.800   Median :3.000   Median :4.350   Median :1.300
 Mean   :5.843   Mean   :3.057   Mean   :3.758   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
```


schema_Nomalize_1.ows
Data
File
CSV File Import
Datasets
SQL Table
Data Table
Paint Data
Data Info
Rank
Edit Domain
Color
Feature Statistics
Save Data
Transform
Visualize
Tree Viewer
Box Plot
Violin Plot
Distributions
Scatter Plot
Line Plot
Bar Plot
Sieve Diagram
Mosaic Display
FreeViz
Linear Projection
Radviz
Heat Map
Venn Diagram
Silhouette Plot
Pythagorean Tree
Box Plot
Visualize the distribution of feature values in a box plot.
more...
Data
Data Table
Box Plot (1)
Preprocessed Data → Data
Preprocess
Box Plot
File
Preprocess (1)
Data Table (1)
Box Plot (1)
Variable
Filter...
iris
Order by relevance to subgroups
Subgroups
None
Order by relevance to variable
Display
Annotate
No comparison
Compare medians
Compare means
Iris-setosa: 5.006 ± 0.35
4.8 5.0 5.2
Iris-versicolor: 5.936 ± 0.51
5.6 5.9 6.3
Iris-virginica: 6.588 ± 0.63
6.250 6.5 6.9
150
- | 150
Box Plot
Variable
Filter...
iris
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
- | 150

...

- Mise à l'échelle min-max :
  - Convient aux algorithmes qui exigent que les caractéristiques d'entrée se situent dans une plage spécifique (e.g. SVM, Réseaux de neurones)
  - Il faut **traiter** les **valeurs aberrantes**, car elles peuvent affecter la mise à l'échelle.
- Normalisation par score Z :
  - Quand les données suivent une **distribution normale**
  - Ou pour utiliser des algorithmes tels que le **clustering k-means**, la **régression linéaire** et la **régression logistique**.
  - Elle aboutit à une distribution centrée autour de 0 avec un écart-type de 1,
  - Idéale pour les algorithmes qui supposent que les données sont normalement distribuées.

Les valeurs manquantes

Et leur imputation

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***

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 ?

Valeurs **manquantes**

1. **Élimination** des données touchées

2. **Remplacement** par

   - valeur **la plus fréquente** (en particulier si données catégorielles)
   - valeur calculée par **interpolation**
     - **Moyenne** des valeurs de la colonne
     - Interpolation **linéaire** (si cela a un sens)

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 »

Les points aberrants

« *outliers* »

Traitement des valeurs aberrantes

- Utilisez des **diagrammes en boîte** (box-plots), des **histogrammes** ou des **nuages de points** (scatter plots) pour *visualiser la distribution* des caractéristiques numériques et **repérer visuellement** les valeurs aberrantes potentielles.

- Calculez des **statistiques** telles que la **moyenne**, l'**écart-type**, les **quartiles** et l'**écart interquartile** (EI). Les valeurs aberrantes sont souvent définies comme les points de données situés en dessous de Q1 - 1,5 * EI ou au-dessus de Q3 + 1,5 * EI.

- Appliquez des transformations telles que la **transformation logarithmique** ou la transformation de Box-Cox afin de rendre la distribution des données **plus normale** et de réduire l'impact des valeurs aberrantes.

Pour toute valeur de x positive, on définit la transformée de Box-Cox de la manière suivante

$$B(x,\lambda) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{si } \lambda \neq 0 \\ \log(x) & \text{si } \lambda = 0 \end{cases}$$

si $\lambda > 1$ cela amplifie les grandes valeurs de x

si $\lambda < 1$ cela réduit les grandes valeurs de x

Les doublons

« *duplicates* »

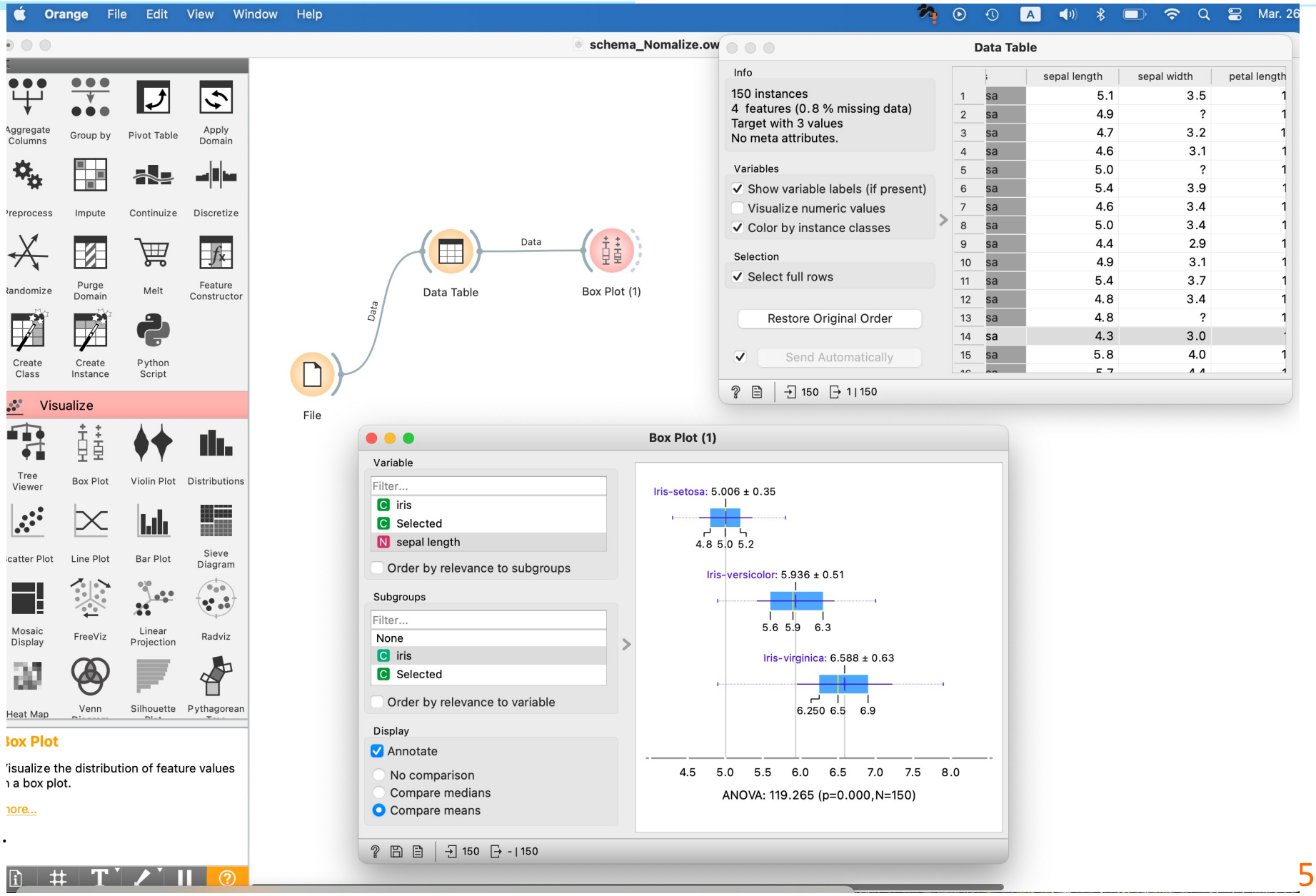
- Ils peuvent **fausser** les **statistiques**
- Ils ne sont **pas** toujours **évidents** à **identifier**
- Il existe des **fonctions spéciales** dans des logiciels tels que *pandas*
  - Des fonctions comme duplicated() en pandas peuvent identifier des doublons sur des lignes ou des colonnes spécifiées.
  - Ces doublons peuvent être éliminés en utilisant la fonction drop_duplicates()

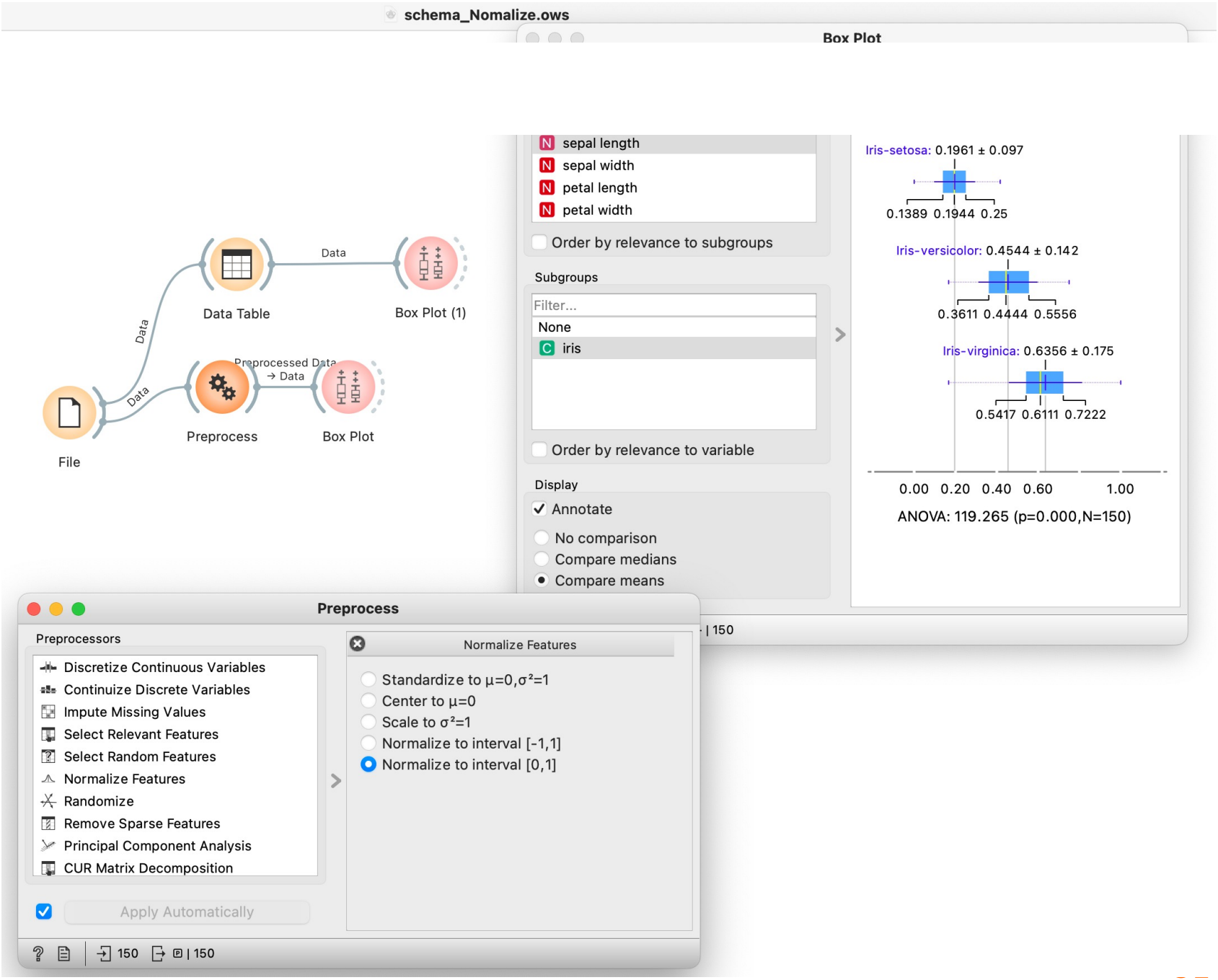
Des méthodes similaires existent dans d'autres logiciels

Les prétraitements

1. Motivation
2. Types de prétraitements
3. Illustration
4. La réduction de dimension
    1. Par sélection de variables
    2. Par projection dans un nouvel espace
5. Prétraitements pour les séries temporelles
6. Conclusion

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

Les prétraitements

1. Motivation
2. Types de prétraitements
3. Illustration
4. La **réduction** de dimension
    1. Par sélection de variables
    2. Par projection dans un nouvel espace
5. Prétraitements pour les séries temporelles
6. 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 ?

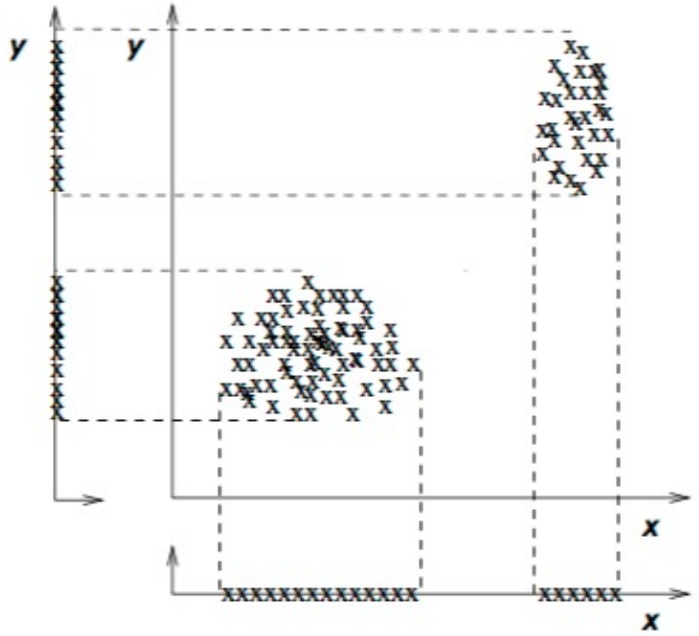
- Par projection dans un nouvel **espace de plus petite dimension**
  - Analyse en **composantes**
    - **Principales** (ACP)
    - Pour des valeurs catégorielles : ACM
    - Indépendantes (ACI)
    - ...
  - Projections **non linéaires**
    - **t-SNE**, IsoMap, Locally Linear Embedding, ...

Les prétraitements

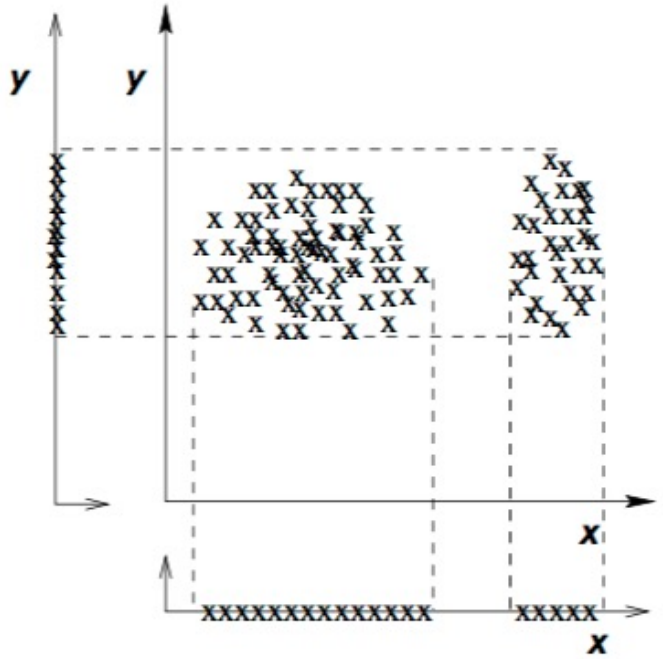
1. Motivation
2. Types de prétraitements
3. Illustration
4. La réduction de dimension
    1. Par **sélection** de variables
    2. Par projection dans un nouvel espace
5. Prétraitements pour les séries temporelles
6. Conclusion

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

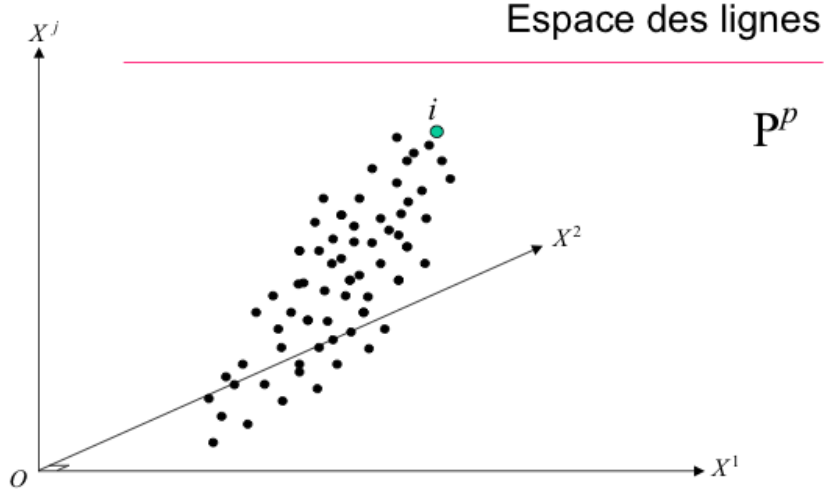
  - **Classification** : isoler l'**information** permettant de **classer les exemples et ceux à venir**
    - Peut résulter d'un apprentissage

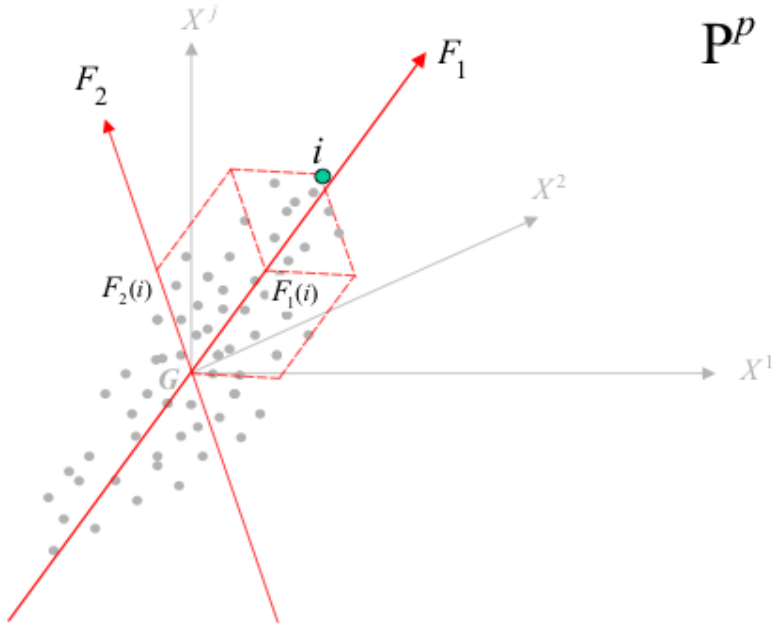
Les prétraitements

1. Motivation
2. Types de prétraitements
3. Illustration
4. La réduction de dimension
    1. Par **sélection** de variables
    2. Par **projection** dans un nouvel espace
5. Prétraitements pour les séries temporelles
6. Conclusion

Projection dans nouvel espace

L'Analyse en Composantes Principales





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

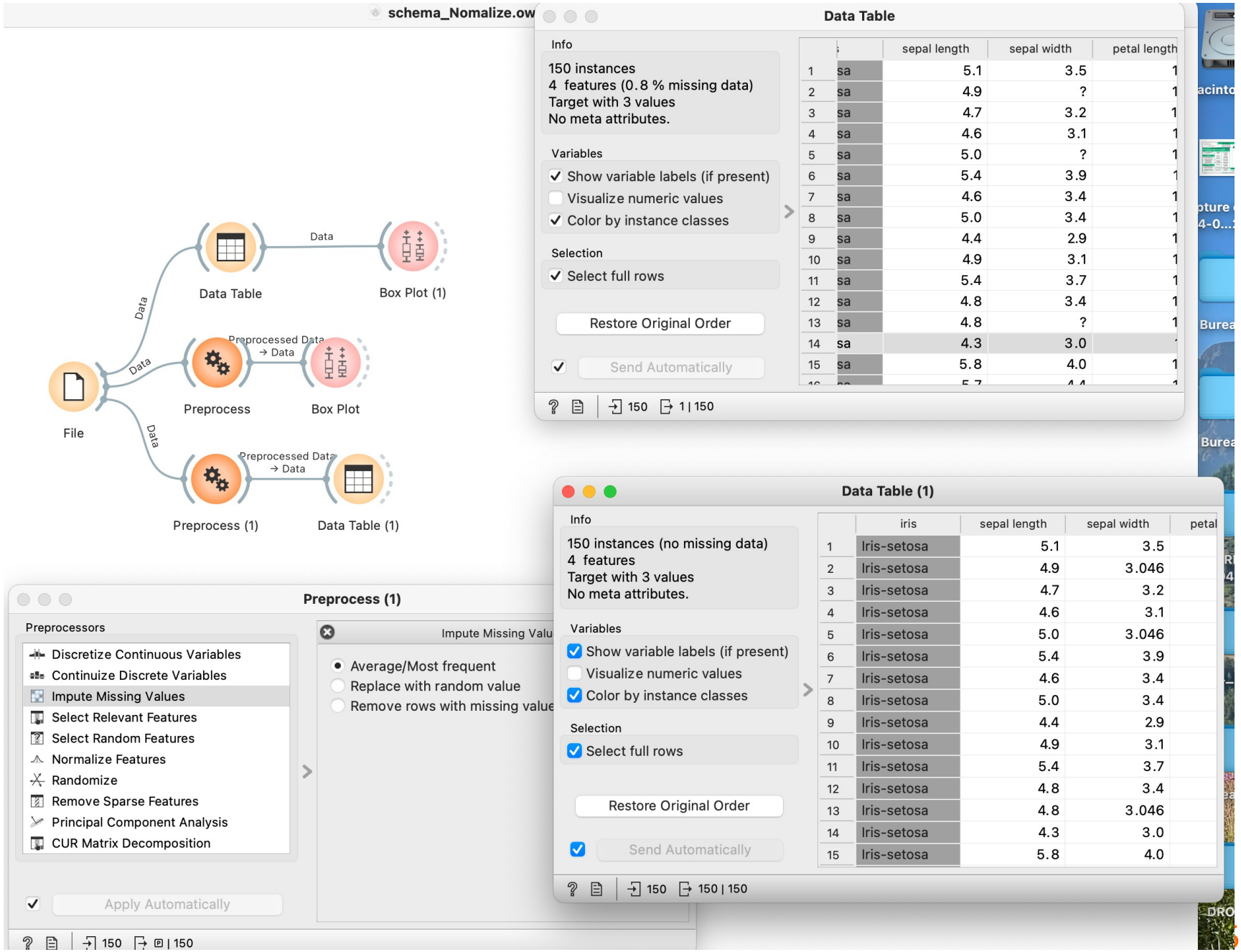
L'Analyse en Composantes Principales

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

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

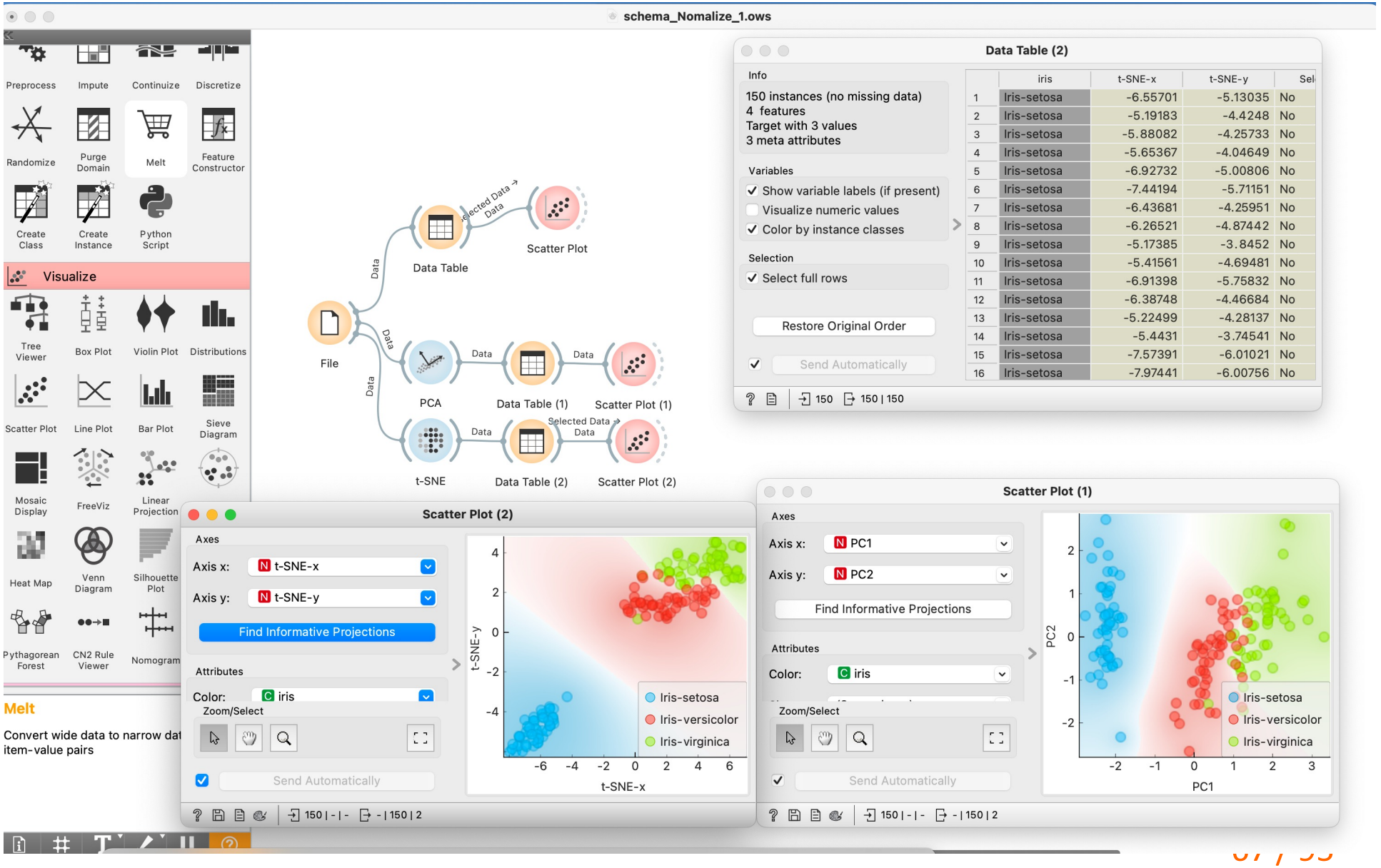
```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   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
```

- Notebook
  - demo_pca_tsne_umap.ipynb


schema_Nomalize.ow
Data Table
Box Plot (1)
Preprocess
Box Plot
File
Preprocess (1)
Data Table (1)
Data
Preprocessed Data
→ Data
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
150
1 | 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
Data Table (1)
150 instances (no missing data)
4 features
Target with 3 values
No meta attributes.
iris
sepal length
sepal width
petal
Iris-setosa
150 | 150


Data Table
Info
150 instances (no missing data)
4 features
Target with 3 values
No meta attributes.
Restore Original Order
Send Automatically
iris
sepal length
sepal width
petal length
petal width
Iris-setosa
Data Table (1)
Info
150 instances (no missing data)
4 features
Target with 3 values
2 meta attributes
Variables
Restore Original Order
Send Automatically
variance
PC1 0.727705
PC2 0.230305
Scatter Plot (1)
Axes
Axis x: PC1
Axis y: PC2
Find Informative Projections
Attributes
Color: iris
Zoom/Select
Send Automatically
Iris-setosa
Iris-versicolor
Iris-virginica
Scatter Plot
Axis x: sepal length
Axis y: sepal width
Shape: (Same shape)
Preprocess
Impute
Continuize
Discretize
Randomize
Purge Domain
Melt
Feature Constructor
Create Class
Create Instance
Python Script
Visualize
Tree Viewer
Box Plot
Violin Plot
Distributions
Scatter Plot
Line Plot
Bar Plot
Sieve Diagram
Mosaic Display
FreeViz
Heat Map
Venn Diagram
Pythagorean
CN2 Rule
Scatter Plot
Interactive scatter plot
intelligent data visualiz
more...
File
Data Table
Scatter Plot
PCA
Data Table (1)
Scatter Plot (1)
Data

...


schema_Nomalize_1.ows
Preprocess
Impute
Continuize
Discretize
Randomize
Purge Domain
Melt
Feature Constructor
Create Class
Create Instance
Python Script
Visualize
Tree Viewer
Box Plot
Violin Plot
Distributions
Scatter Plot
Line Plot
Bar Plot
Sieve Diagram
Mosaic Display
FreeViz
Linear Projection
Heat Map
Venn Diagram
Silhouette Plot
Pythagorean Forest
CN2 Rule Viewer
Nomogram
Melt
Convert wide data to narrow dat
item-value pairs
File
Data Table
Scatter Plot
PCA
Data Table (1)
Scatter Plot (1)
t-SNE
Data Table (2)
Scatter Plot (2)
Data
Selected Data
Data Table (2)
Info
150 instances (no missing data)
4 features
Target with 3 values
3 meta attributes
Variables
Show variable labels (if present)
Visualize numeric values
Color by instance classes
Selection
Select full rows
Restore Original Order
Send Automatically
150
150 | 150
Scatter Plot (2)
Axes
Axis x: t-SNE-x
Axis y: t-SNE-y
Find Informative Projections
Attributes
Color: iris
Zoom/Select
Send Automatically
Iris-setosa
Iris-versicolor
Iris-virginica
150 | - | -
- | 150 | 2
Scatter Plot (1)
Axis x: PC1
Axis y: PC2
Find Informative Projections
Attributes
Color: iris
Zoom/Select
Send Automatically
PC1
PC2
150 | - | -
- | 150 | 2

Les prétraitements

1. Motivation
2. Types de prétraitements
3. Illustration
4. La réduction de dimension
    1. Par sélection de variables
    2. Par projection dans un nouvel espace
5. Prétraitements pour les **séries temporelles**
6. Conclusion

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**

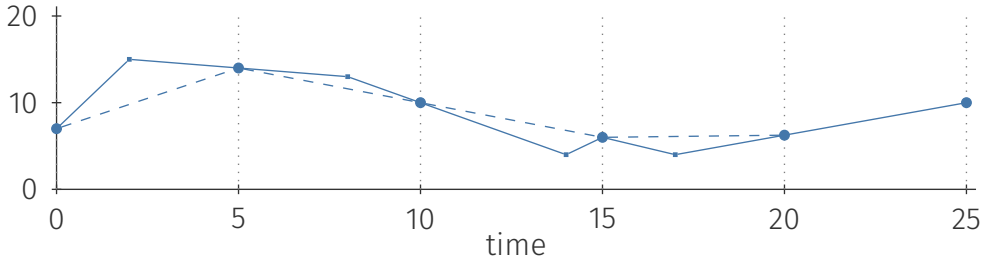
Interpolation pour synchroniser les valeurs

- 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

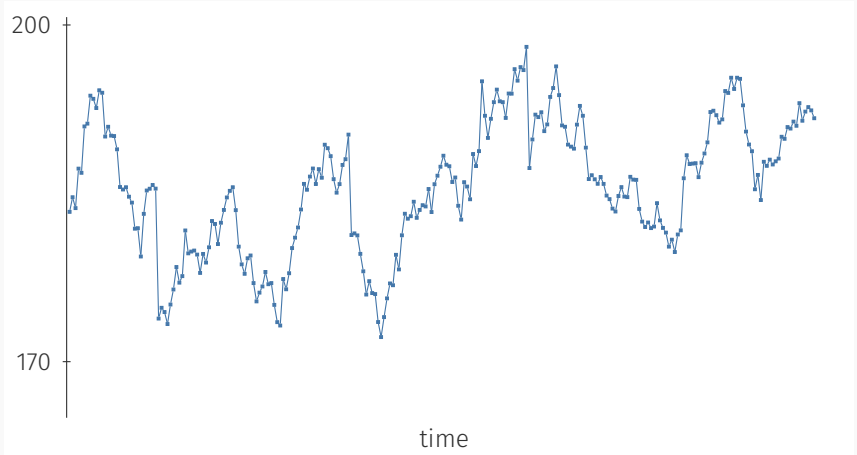
Instead of average, it is possible to take the median, which is more robust to the presence of outlier values

...

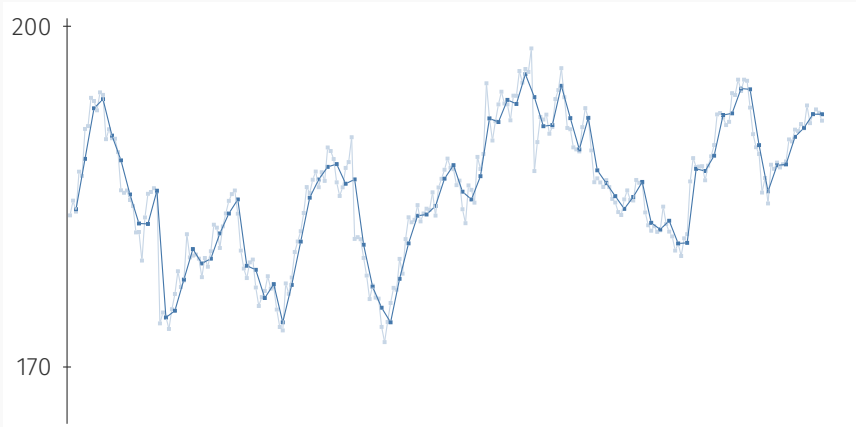
Binning : illustration

IBM stock prices from Sept. 2013 to Sept. 2014

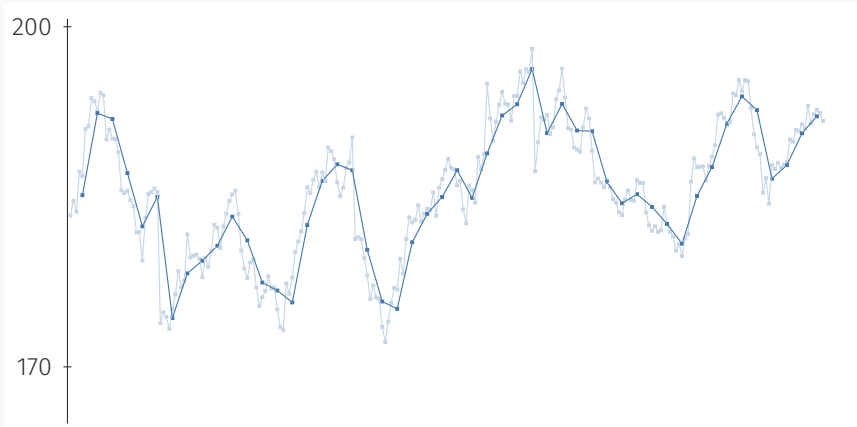
Original time-series



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 :

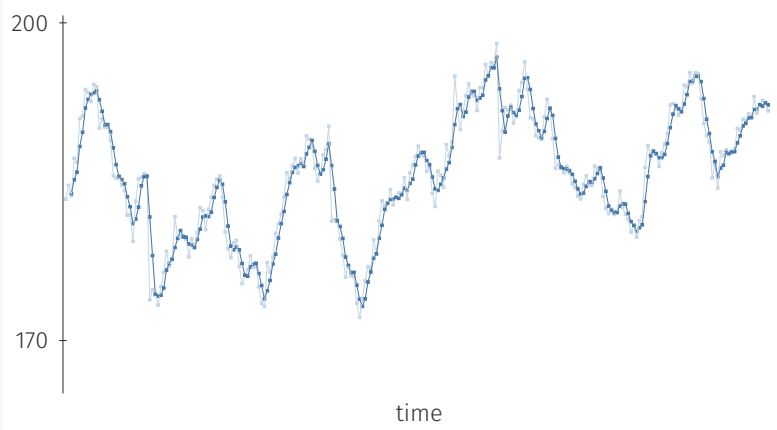
$$y_i = \frac{1}{k} \sum_{r=0}^{k-1} x_{i+r} \quad \text{pour } i = 1, \ldots, n-k+1$$

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

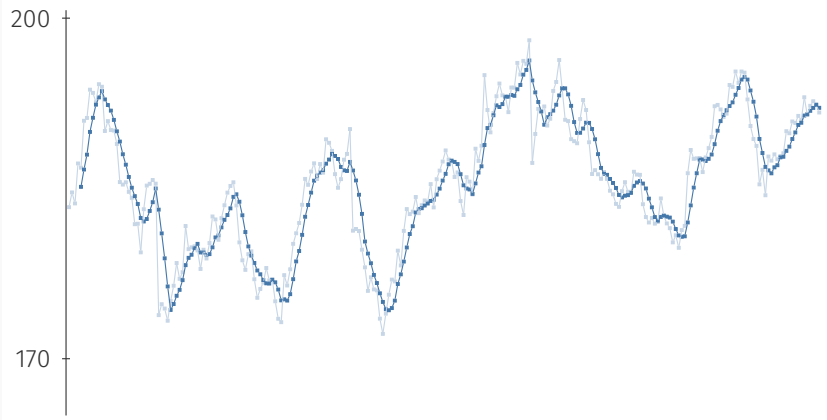
Lissage par moyenne mobile : illustration

IBM stock prices from Sept. 2013 to Sept. 2014

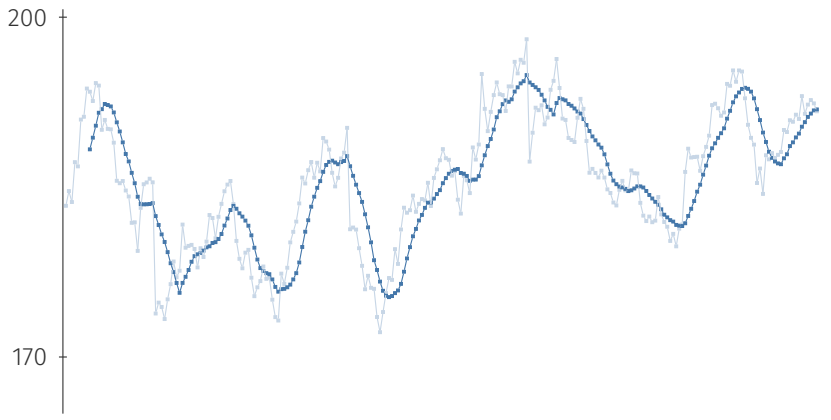
Moving-average smoothing, $k = 3$



Moving-average smoothing, $k = 5$

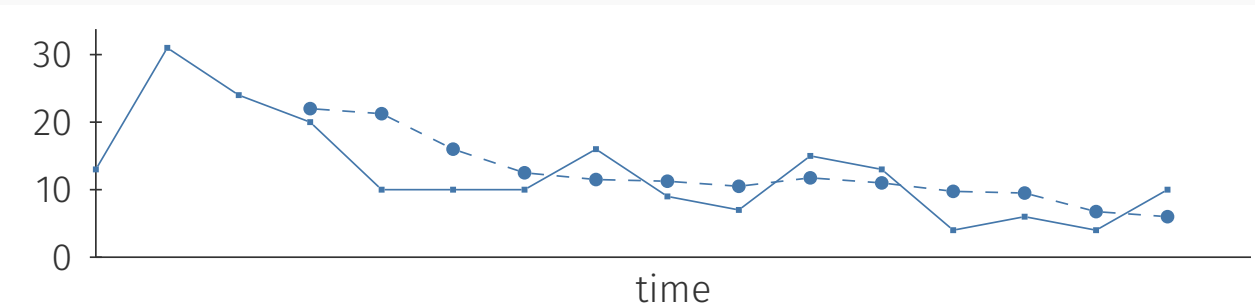


Moving-average smoothing, $k = 9$



...

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 $y_1 = x_1$:

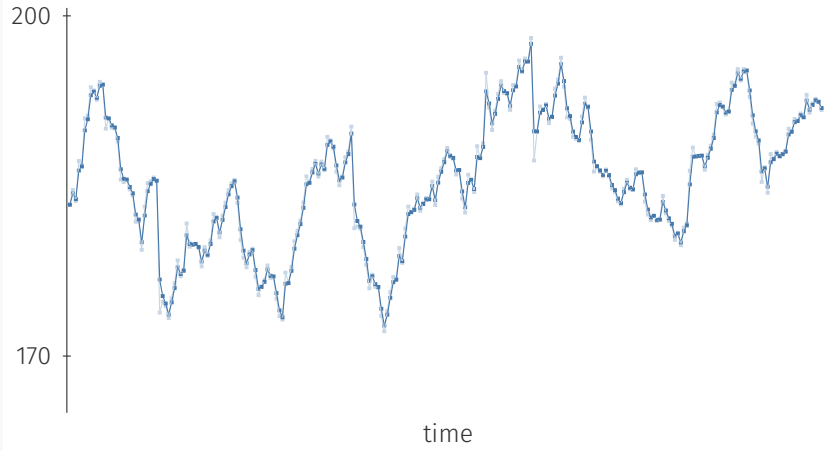
$$y_i = \alpha \cdot x_i + (1 - \alpha) \cdot y_{i-1} \quad \text{pour } i = 2, \ldots, n$$

- 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

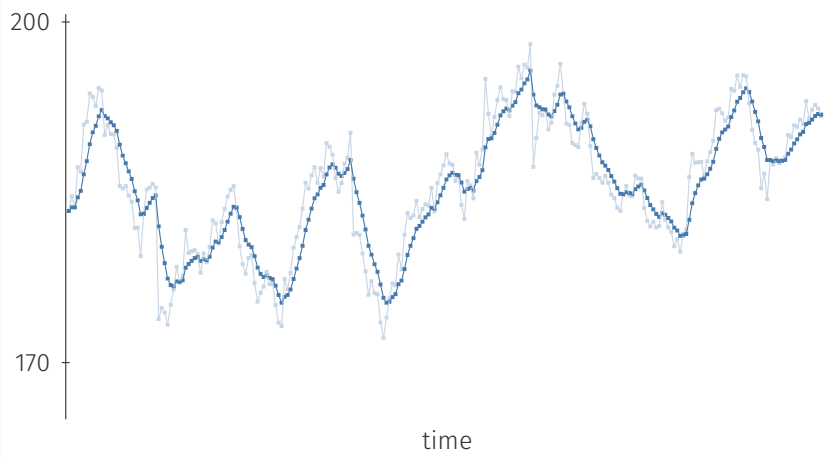
Lissage exponentiel : illustration

IBM stock prices from Sept. 2013 to Sept. 2014

Exponential smoothing, $\alpha = 0.75$

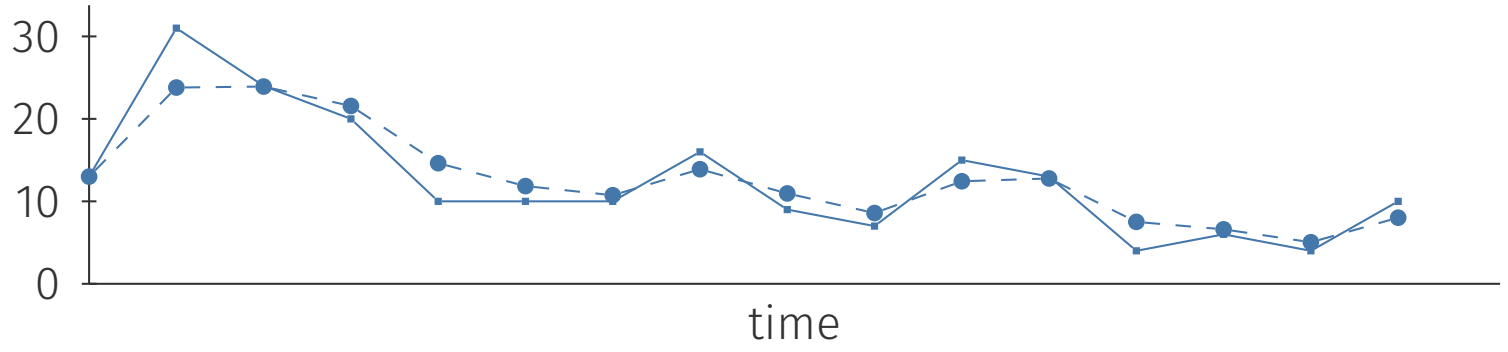


Exponential smoothing, $\alpha = 0.25$



...

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

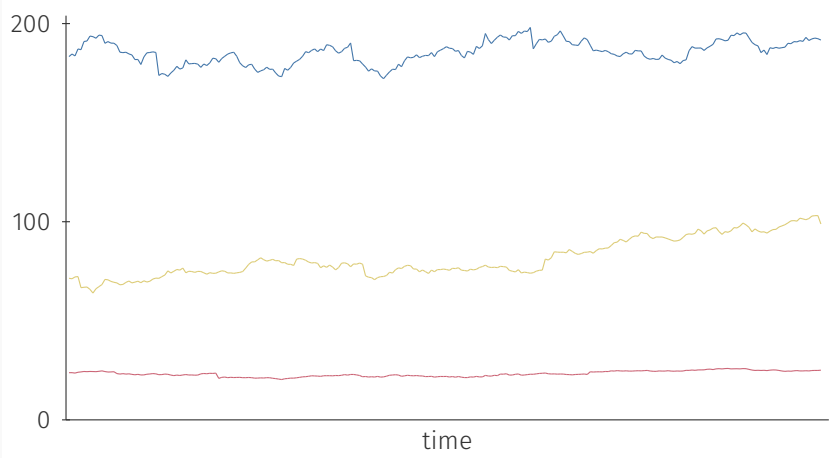
- Étant donnée une série temporelle $S_X = \langle x_1, x_2, \ldots, x_n \rangle$ de moyenne μ et d'écart-type σ, la standardisation recode les valeurs selon :

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

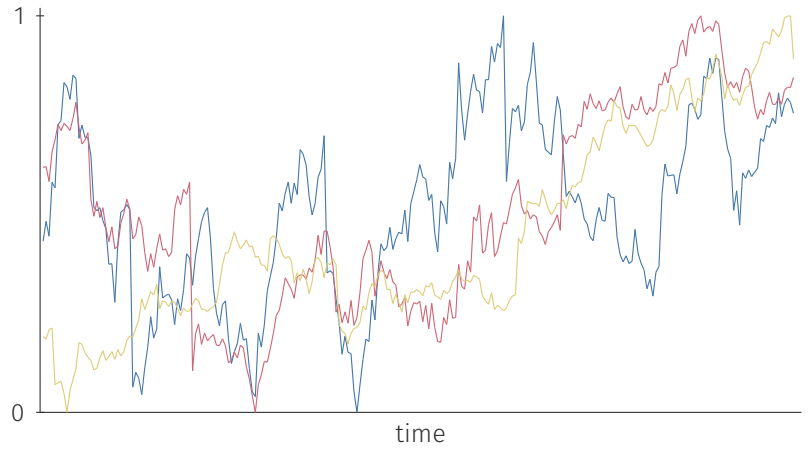
Les 2 types de normalisation : Illustration

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

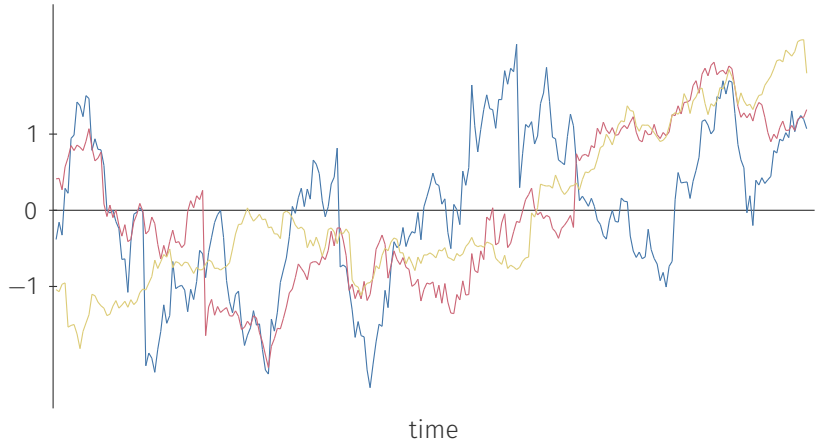
Original time-series



Range-based normalized time-series

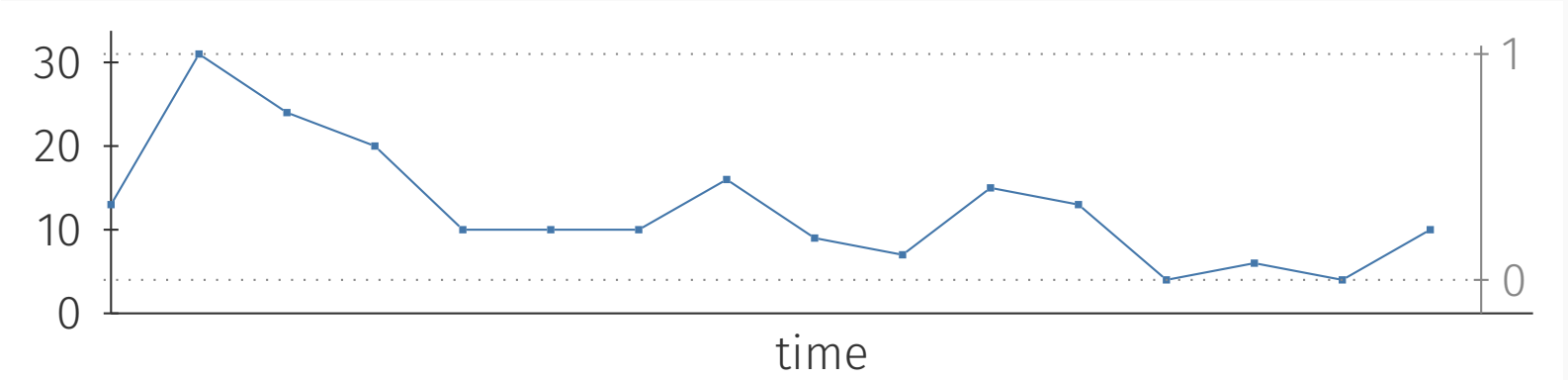


Standardized time-series



...

Normalisation sur intervalle



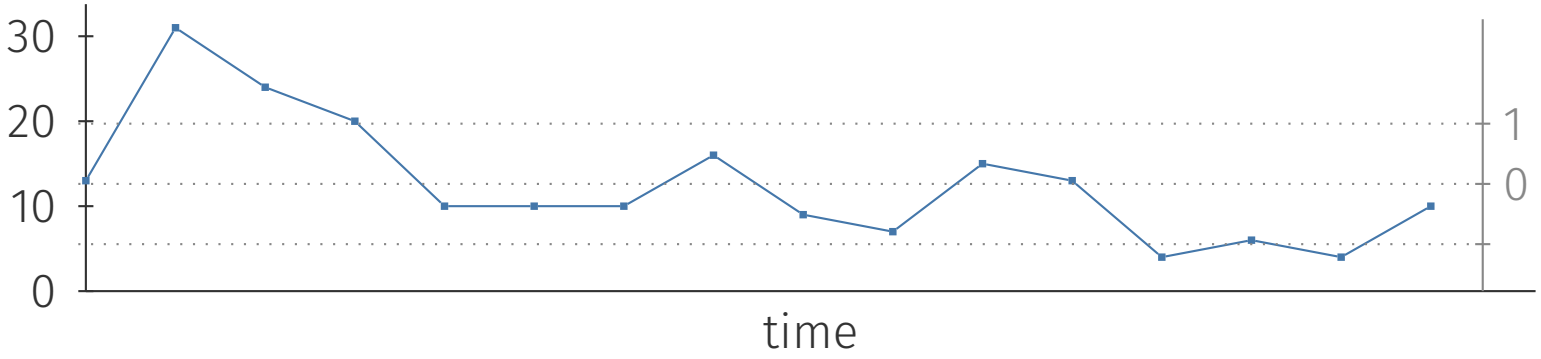
$\langle 13, 31, 24, 20, 10, 10, 10, 16, 9, 7, 15, 13, 4, 6, 4, 10 \rangle$

Range-based normalization

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

...

Standardisation

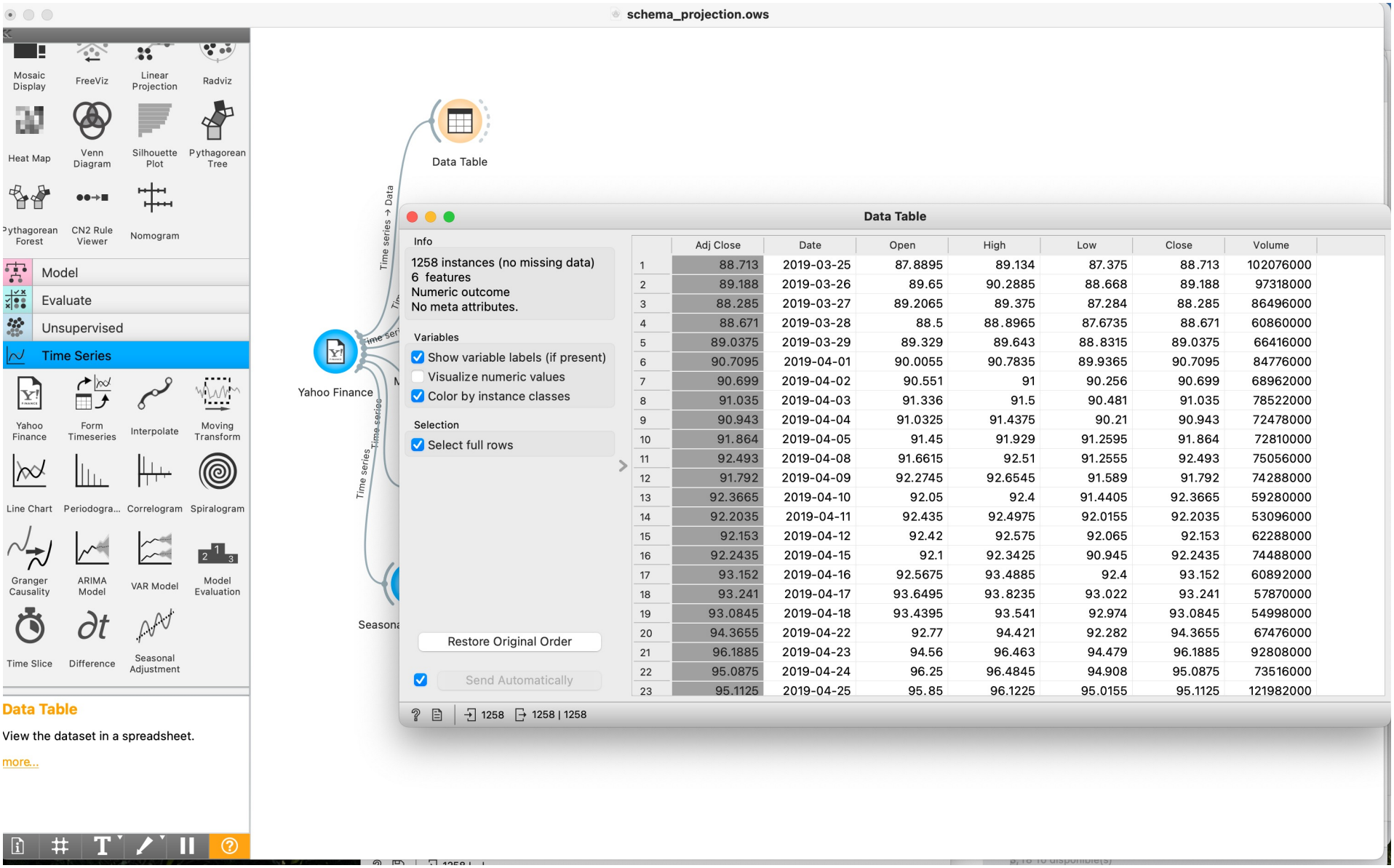


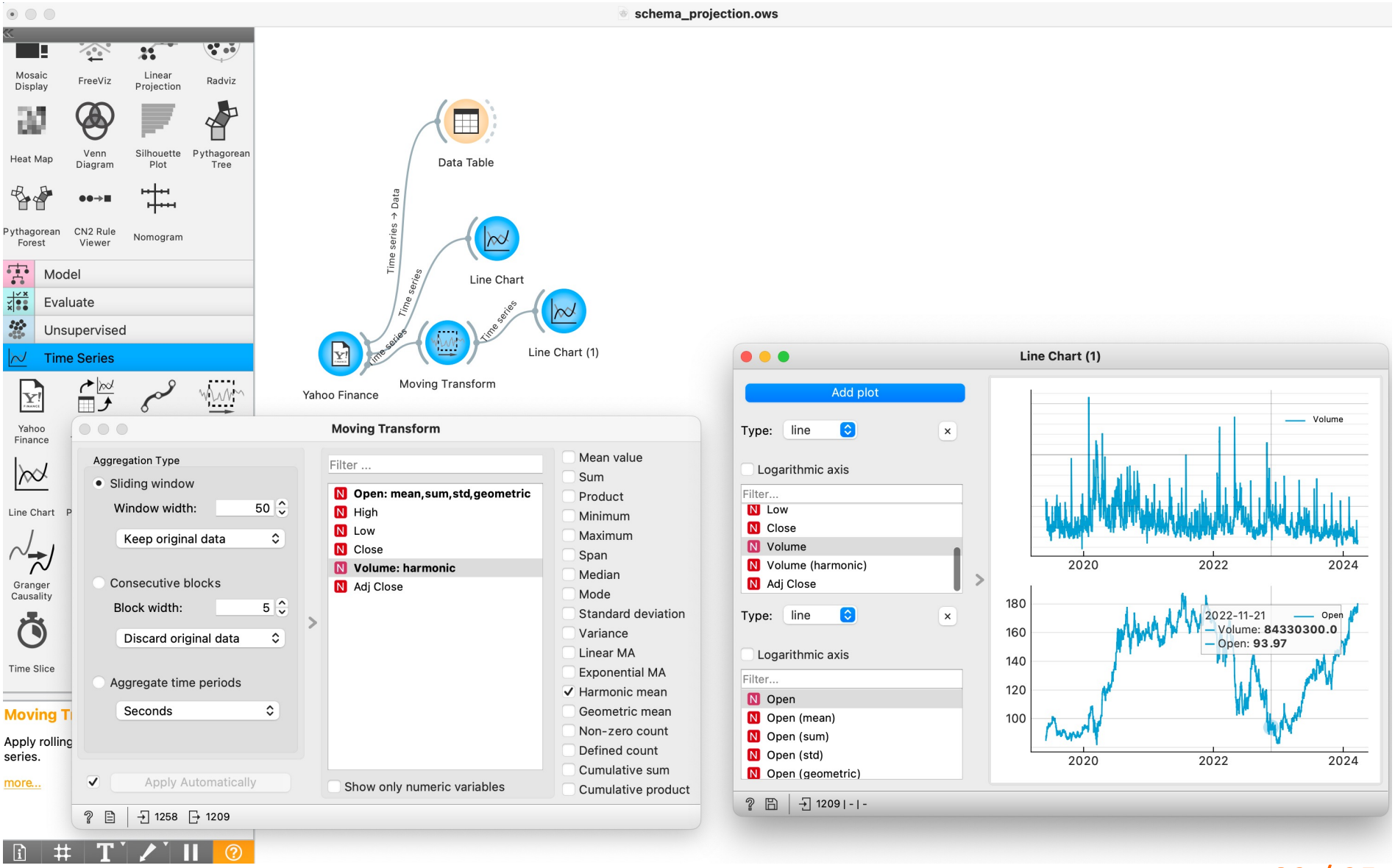
$$\langle 13, 31, 24, 20, 10, 10, 10, 16, 9, 7, 15, 13, 4, 6, 4, 10 \rangle$$

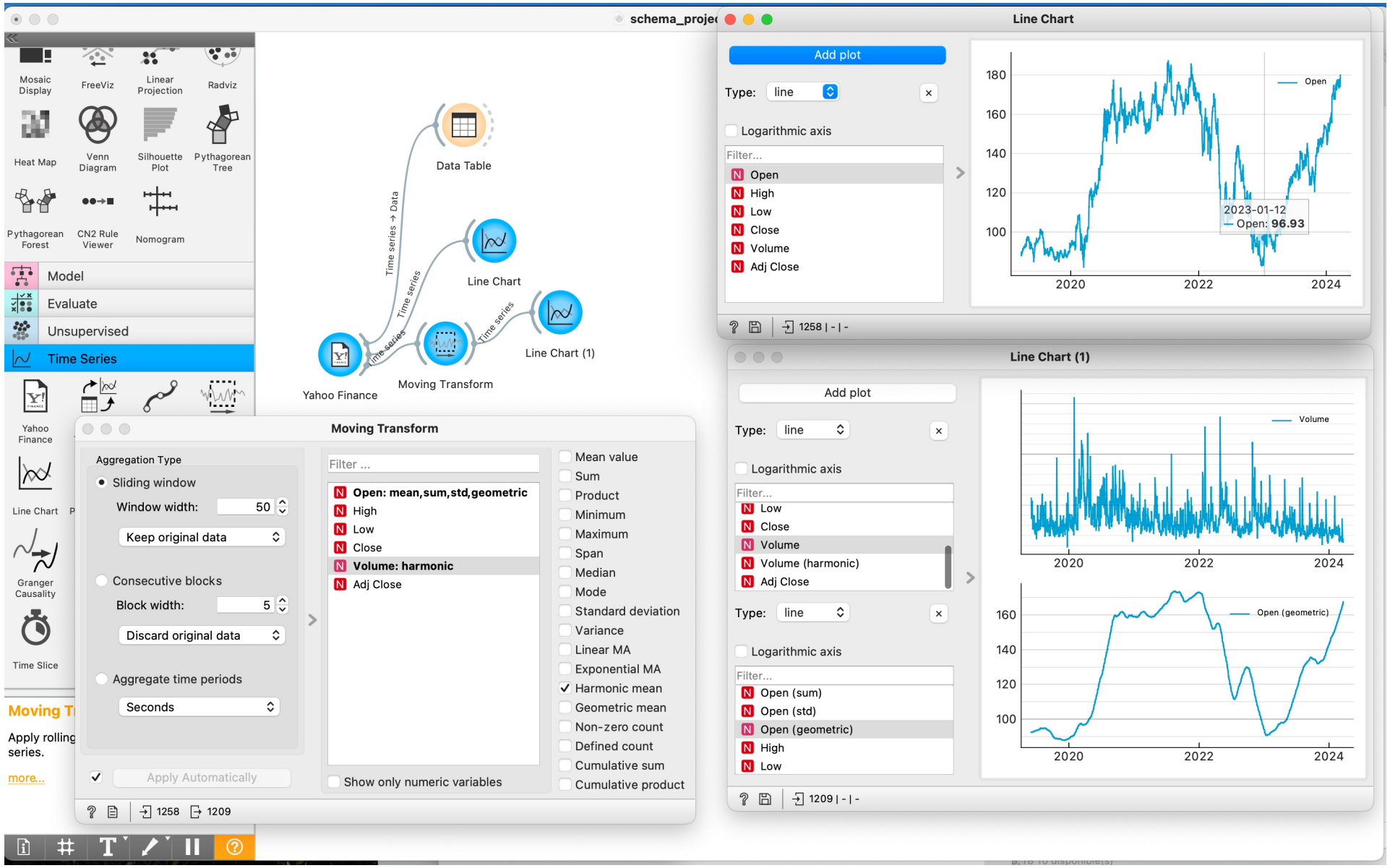
Standardization

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

...


schema_projection.ows
Mosaic Display
FreeViz
Linear Projection
Radviz
Heat Map
Venn Diagram
Silhouette Plot
Pythagorean Tree
Pythagorean Forest
CN2 Rule Viewer
Nomogram
Model
Evaluate
Unsupervised
Time Series
Yahoo Finance
Form Timeseries
Interpolate
Moving Transform
Line Chart
Periodogra...
Correlogram
Spiralogram
Granger Causality
ARIMA Model
VAR Model
Model Evaluation
Time Slice
Difference
Seasonal Adjustment
Data Table
View the dataset in a spreadsheet.
more...
Data Table
Yahoo Finance
Time series → Data
Info
1258 instances (no missing data)
6 features
Numeric outcome
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
1258
1258 | 1258
Adj Close Date Open High Low Close Volume
1 88.713 2019-03-25 87.8895 89.134 87.375 88.713 102076000
2 89.188 2019-03-26 89.65 90.2885 88.668 89.188 97318000
3 88.285 2019-03-27 89.2065 89.375 87.284 88.285 86496000
4 88.671 2019-03-28 88.5 88.8965 87.6735 88.671 60860000
5 89.0375 2019-03-29 89.329 89.643 88.8315 89.0375 66416000
6 90.7095 2019-04-01 90.0055 90.7835 89.9365 90.7095 84776000
7 90.699 2019-04-02 90.551 91 90.256 90.699 68962000
8 91.035 2019-04-03 91.336 91.5 90.481 91.035 78522000
9 90.943 2019-04-04 91.0325 91.4375 90.21 90.943 72478000
10 91.864 2019-04-05 91.45 91.929 91.2595 91.864 72810000
11 92.493 2019-04-08 91.6615 92.51 91.2555 92.493 75056000
12 91.792 2019-04-09 92.2745 92.6545 91.589 91.792 74288000
13 92.3665 2019-04-10 92.05 92.4 91.4405 92.3665 59280000
14 92.2035 2019-04-11 92.435 92.4975 92.0155 92.2035 53096000
15 92.153 2019-04-12 92.42 92.575 92.065 92.153 62288000
16 92.2435 2019-04-15 92.1 92.3425 90.945 92.2435 74488000
17 93.152 2019-04-16 92.5675 93.4885 92.4 93.152 60892000
18 93.241 2019-04-17 93.6495 93.8235 93.022 93.241 57870000
19 93.0845 2019-04-18 93.4395 93.541 92.974 93.0845 54998000
20 94.3655 2019-04-22 92.77 94.421 92.282 94.3655 67476000
21 96.1885 2019-04-23 94.56 96.463 94.479 96.1885 92808000
22 95.0875 2019-04-24 96.25 96.4845 94.908 95.0875 73516000
23 95.1125 2019-04-25 95.85 96.1225 95.0155 95.1125 121982000


schema_projection.ows
Mosaic Display
FreeViz
Linear Projection
Radviz
Heat Map
Venn Diagram
Silhouette Plot
Pythagorean Tree
Pythagorean Forest
CN2 Rule Viewer
Nomogram
Model
Evaluate
Unsupervised
Time Series
Yahoo Finance
Line Chart
Granger Causality
Time Slice
Data Table
Line Chart
Line Chart (1)
Moving Transform
Yahoo Finance
Time series → Data
Time series
Time series
Time series
Moving Transform
Aggregation Type
Sliding window
Window width: 50
Keep original data
Consecutive blocks
Block width: 5
Discard original data
Aggregate time periods
Seconds
Apply Automatically
Filter ...
Open: mean,sum,std,geometric
High
Low
Close
Volume: harmonic
Adj Close
Mean value
Sum
Product
Minimum
Maximum
Span
Median
Mode
Standard deviation
Variance
Linear MA
Exponential MA
Harmonic mean
Geometric mean
Non-zero count
Defined count
Cumulative sum
Cumulative product
Show only numeric variables
1258
1209
Line Chart (1)
Add plot
Type: line
Logarithmic axis
Filter...
Low
Close
Volume
Volume (harmonic)
Adj Close
Open
Open (mean)
Open (sum)
Open (std)
Open (geometric)
1209 | - | -
Volume
2020
2022
2024
180
160
140
120
100
2022-11-21
Open
Volume: 84330300.0
Open: 93.97


schema_proje
Mosaic Display
FreeViz
Linear Projection
Radviz
Heat Map
Venn Diagram
Silhouette Plot
Pythagorean Tree
Pythagorean Forest
CN2 Rule Viewer
Nomogram
Model
Evaluate
Unsupervised
Time Series
Yahoo Finance
Line Chart
Granger Causality
Time Slice
Moving T
Apply rolling series.
more...
Data Table
Line Chart
Line Chart (1)
Yahoo Finance
Moving Transform
Time series -> Data
Time series
Time series
Time series
Moving Transform
Aggregation Type
Sliding window
Window width: 50
Keep original data
Consecutive blocks
Block width: 5
Discard original data
Aggregate time periods
Seconds
Apply Automatically
Filter ...
Open: mean,sum,std,geometric
High
Low
Close
Volume: harmonic
Adj Close
Show only numeric variables
Mean value
Sum
Product
Minimum
Maximum
Span
Median
Mode
Standard deviation
Variance
Linear MA
Exponential MA
Harmonic mean
Geometric mean
Non-zero count
Defined count
Cumulative sum
Cumulative product
1258
1209
Line Chart
Add plot
Type: line
Logarithmic axis
Filter...
Open
High
Low
Close
Volume
Adj Close
1258 | - | -
2023-01-12
Open: 96.93
Line Chart (1)
Add plot
Type: line
Logarithmic axis
Filter...
Low
Close
Volume
Volume (harmonic)
Adj Close
Type: line
Logarithmic axis
Filter...
Open (sum)
Open (std)
Open (geometric)
High
Low
1209 | - | -
Volume
Open (geometric)

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

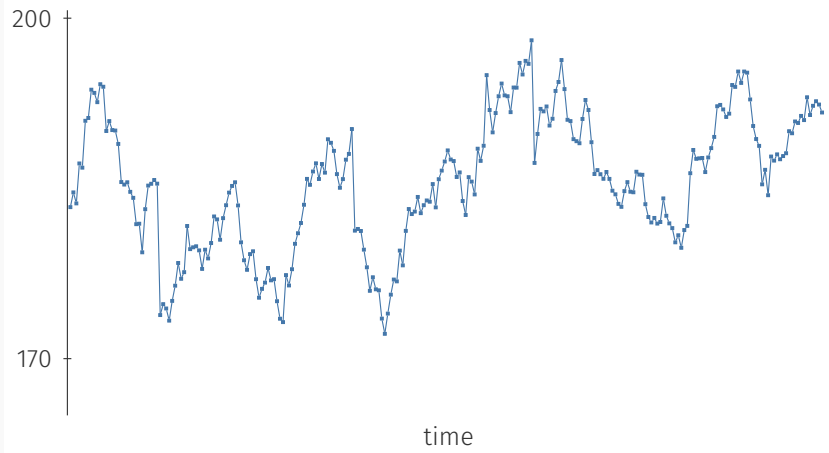
- Approximation symbolique agrégée
  (*Symbolic aggregate approximation* (SAX))
  - Prendre la **valeur moyenne** sur des intervalles successifs de même taille (calculer la « piecewise aggregate approximation » (PAA))
  - Convertir les valeurs résultantes en **valeurs discrètes** prises dans un petit 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

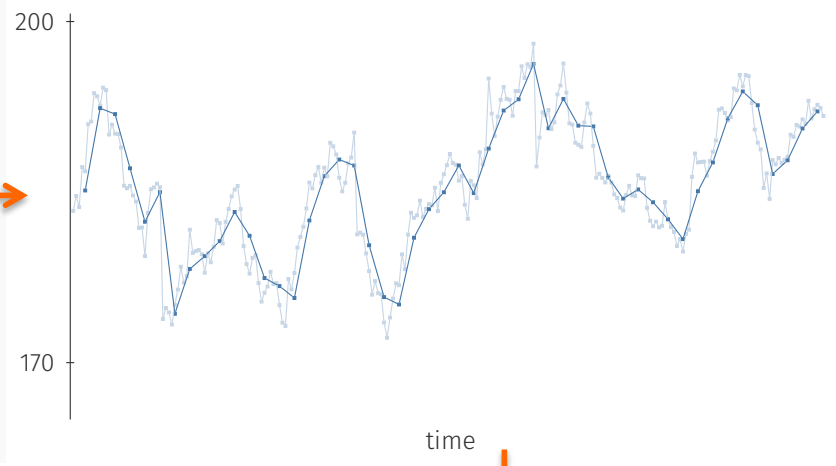
Discrétisation illustration

IBM stock prices from Sept. 2013 to Sept. 2014

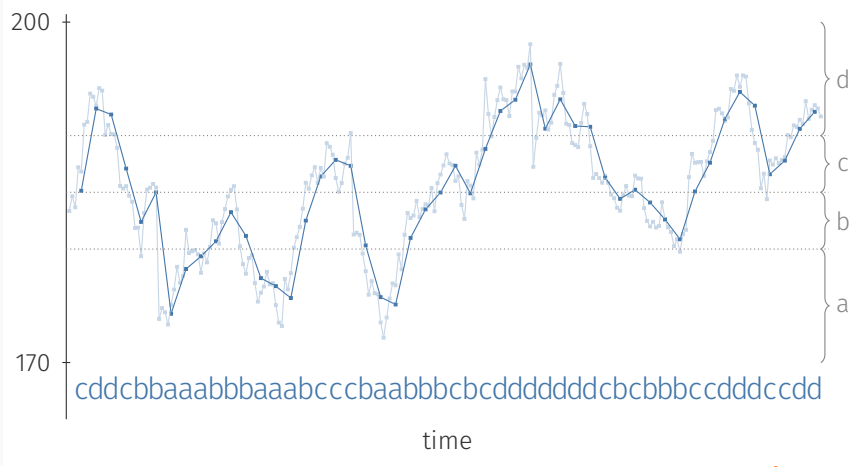
Original time-series



Binning, $k = 5$



Discretizing



...

Les prétraitements

1. Motivation
2. Types de prétraitements
3. Illustration
4. La réduction de dimension
    1. Par sélection de variables
    2. Par projection dans un nouvel espace
5. Prétraitements pour les séries temporelles
6. 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**