Ateliers R³

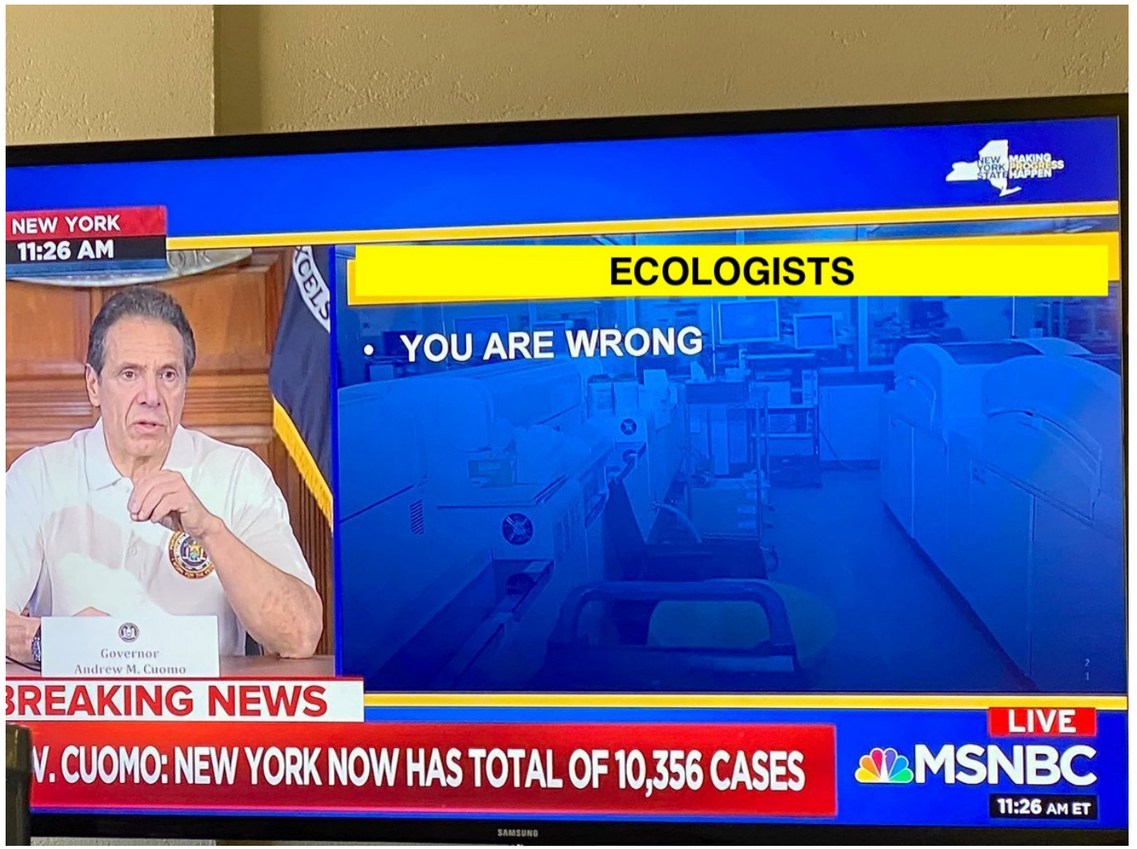
Session 5 - Visualisation de données multivariées

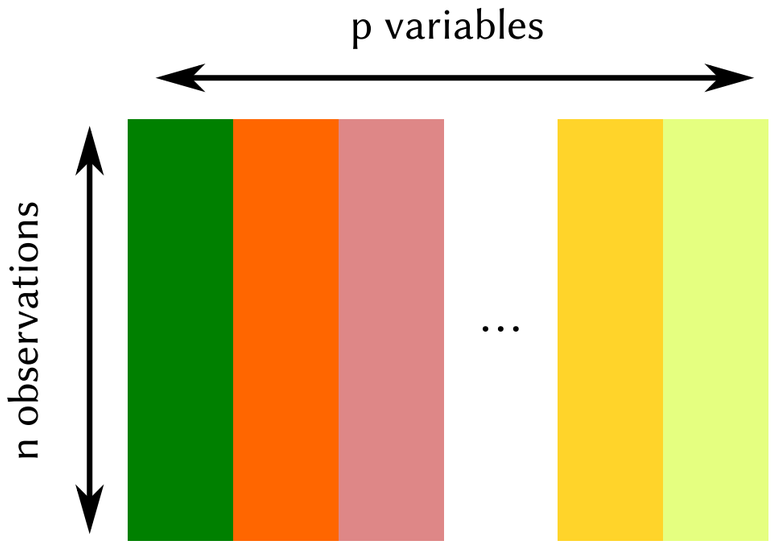
Comment visualiser les données ?

Quels grandes groupes ou variables sont importantes ?

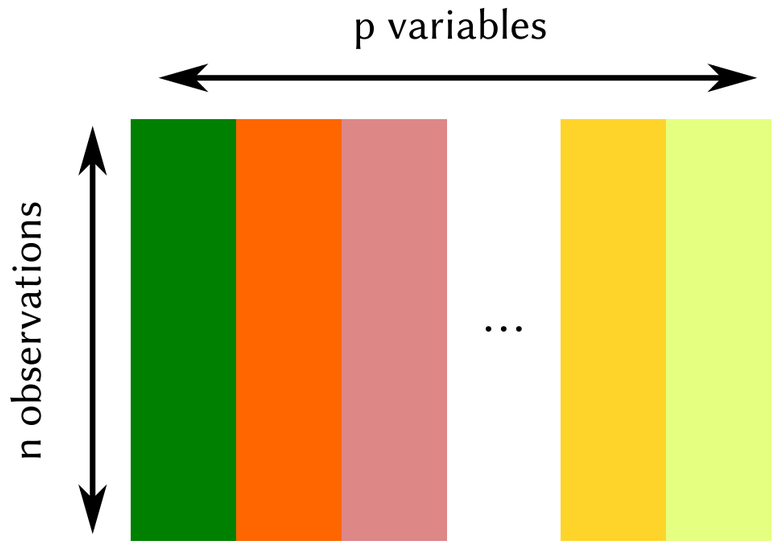
Quelles hypothèses semblent se dégager des données ?

Attention !


NEW YORK
11:26 AM
ECOLOGISTS
• YOU ARE WRONG
Governor
Andrew M. Cuomo
BREAKING NEWS
V. CUOMO: NEW YORK NOW HAS TOTAL OF 10,356 CASES
LIVE
MSNBC
11:26 AM ET



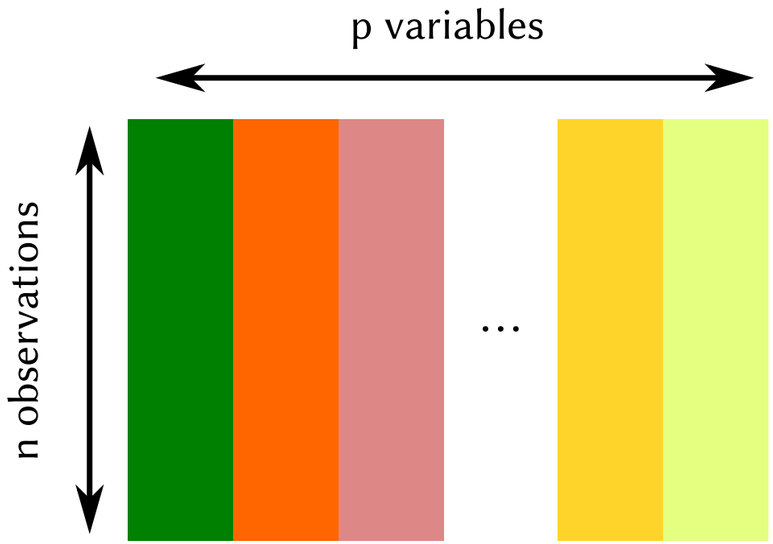
Visualisation sous forme de graphes



Visualisation sous forme de graphes

p ~= 5

ggplot2
avec des couleurs, facets, etc.



Visualisation sous forme de graphes

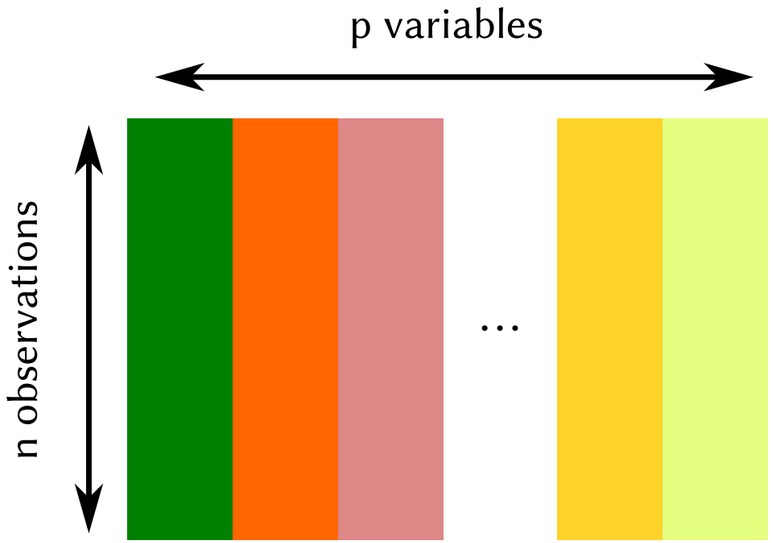
p ~= 5

ggplot2
avec des couleurs, facets, etc.

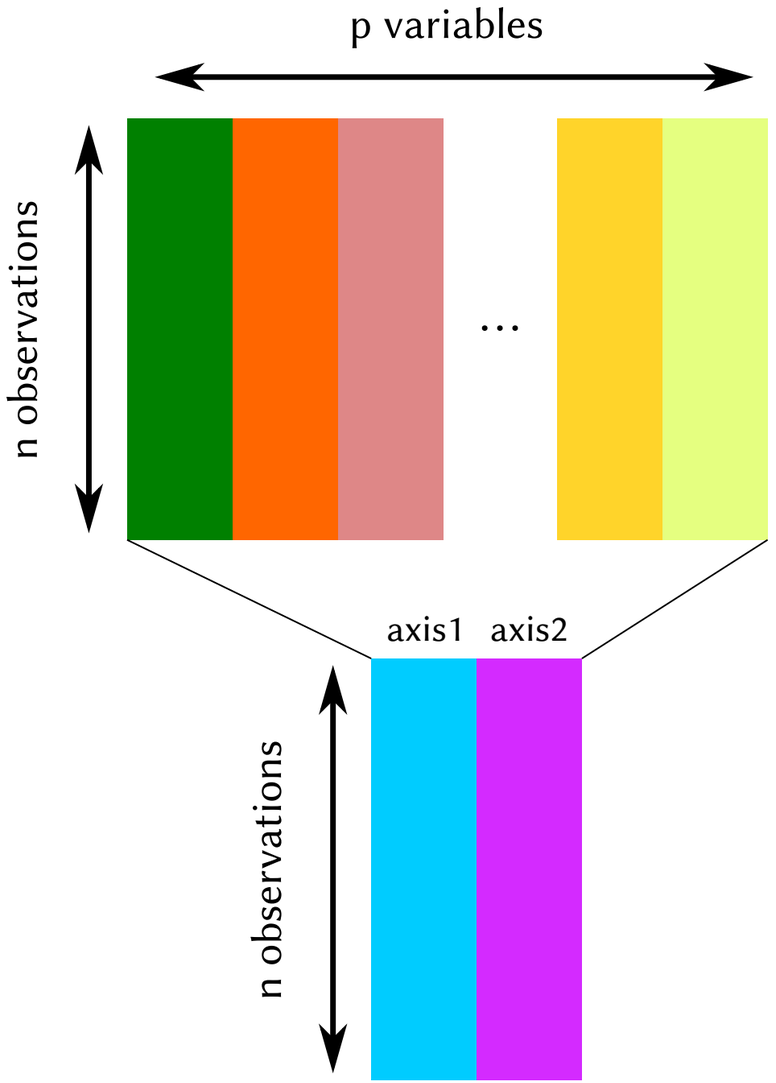
p >> 5

????

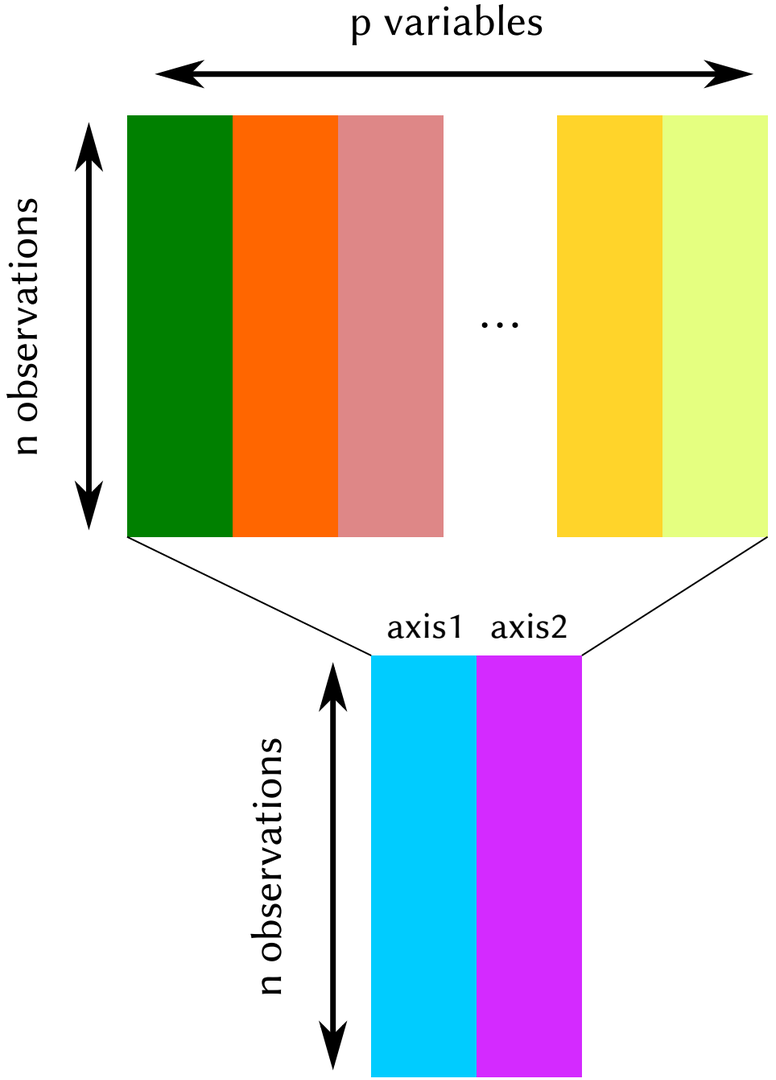
Trop de variables ? Réduire la dimensionalité



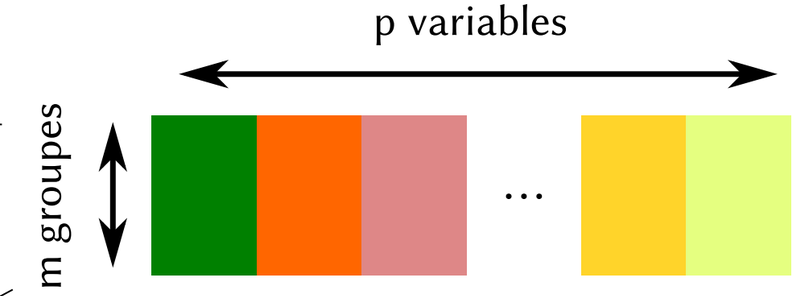
Trop de variables ? Réduire la dimensionalité



Trop de variables ? Réduire la dimensionalité

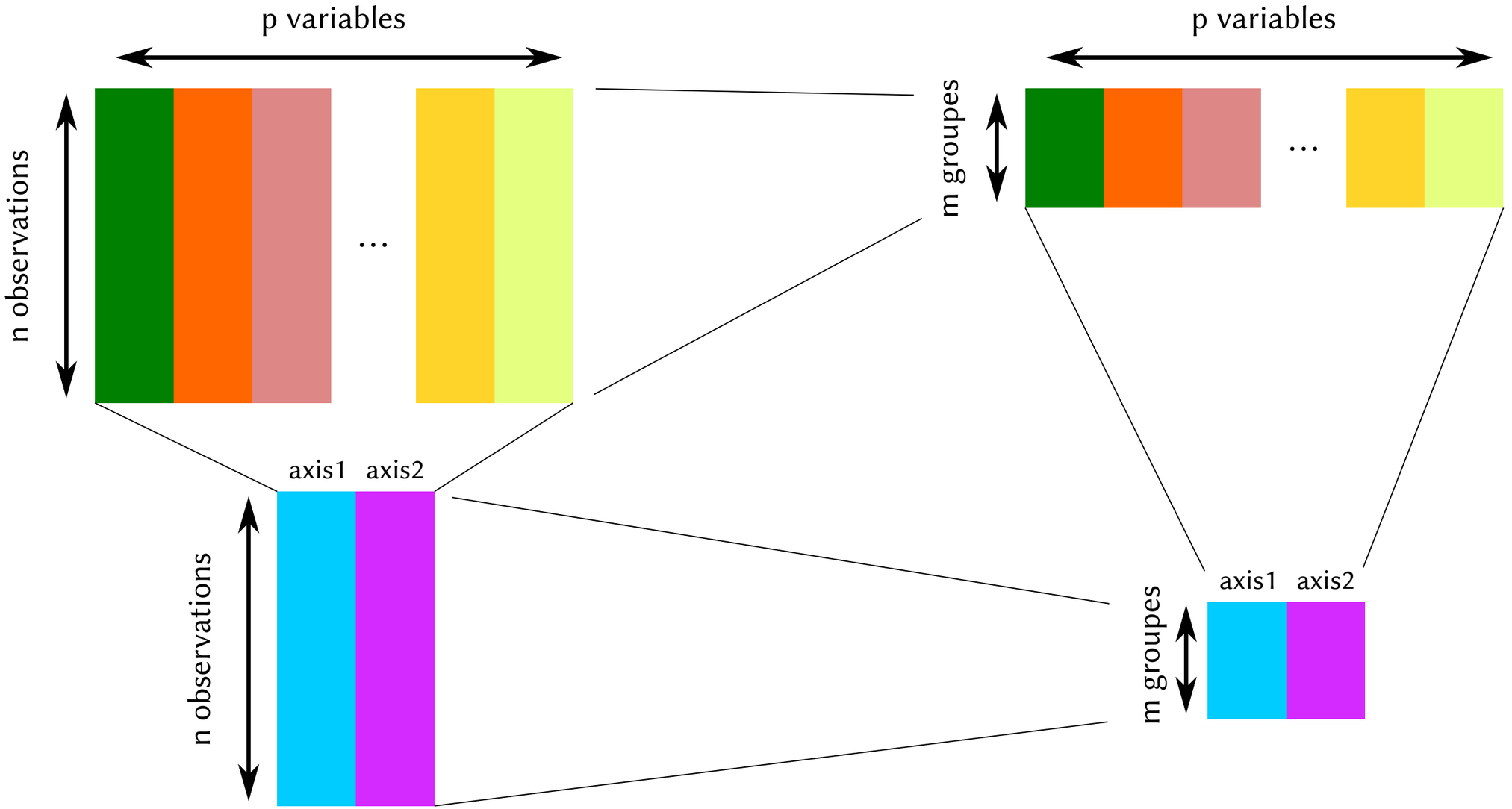


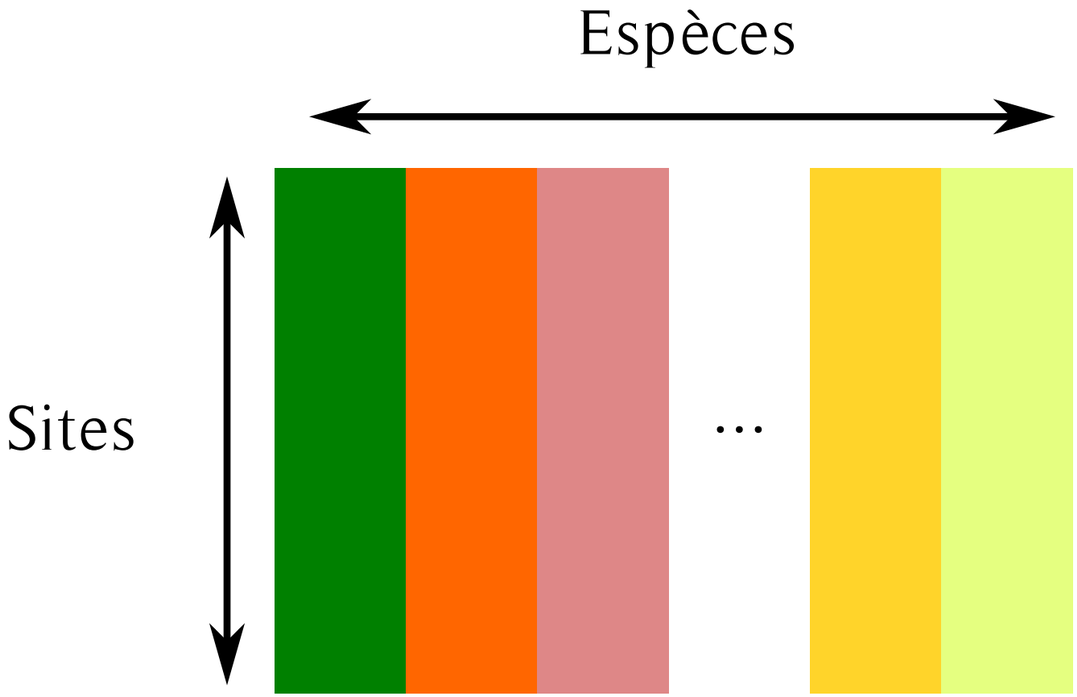
Quels grands types d'observations ? Clustering

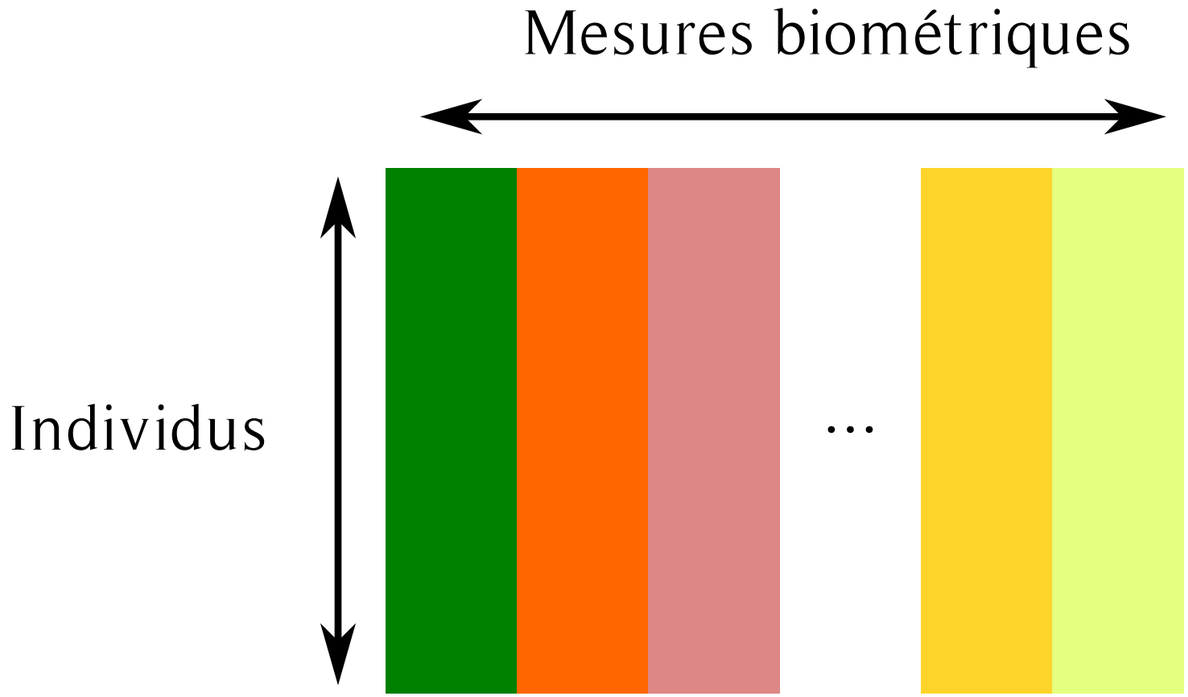


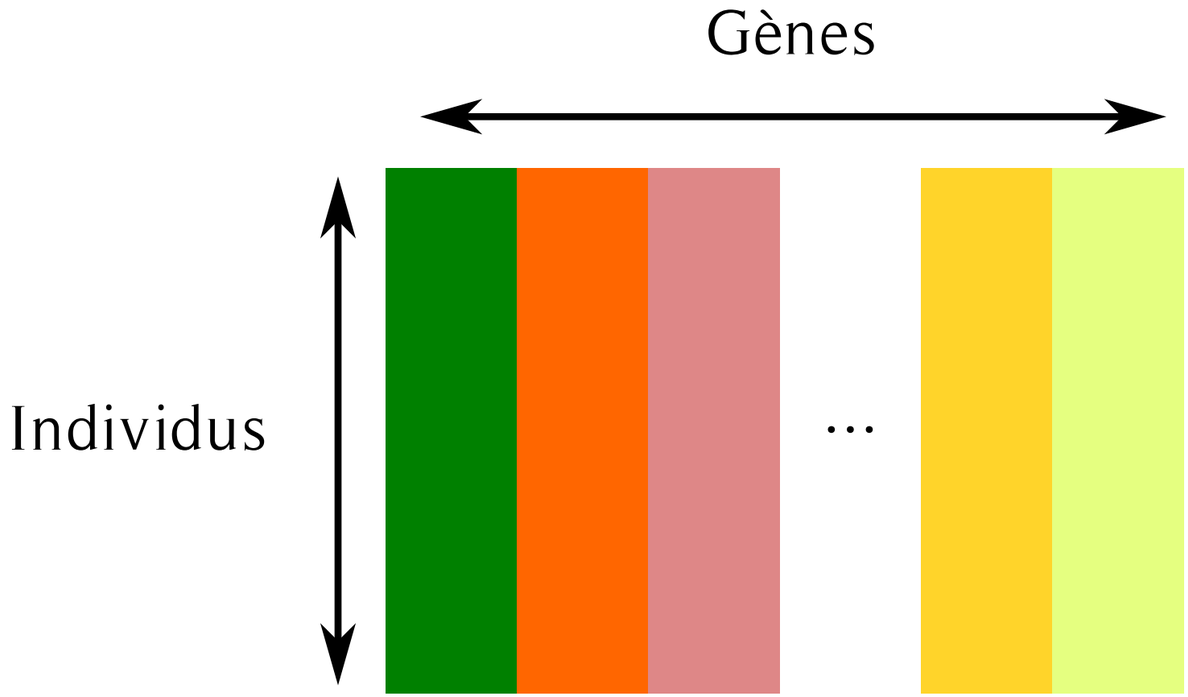
Trop de variables ? Réduire la dimensionalité

Quels grands types d'observations ? Clustering

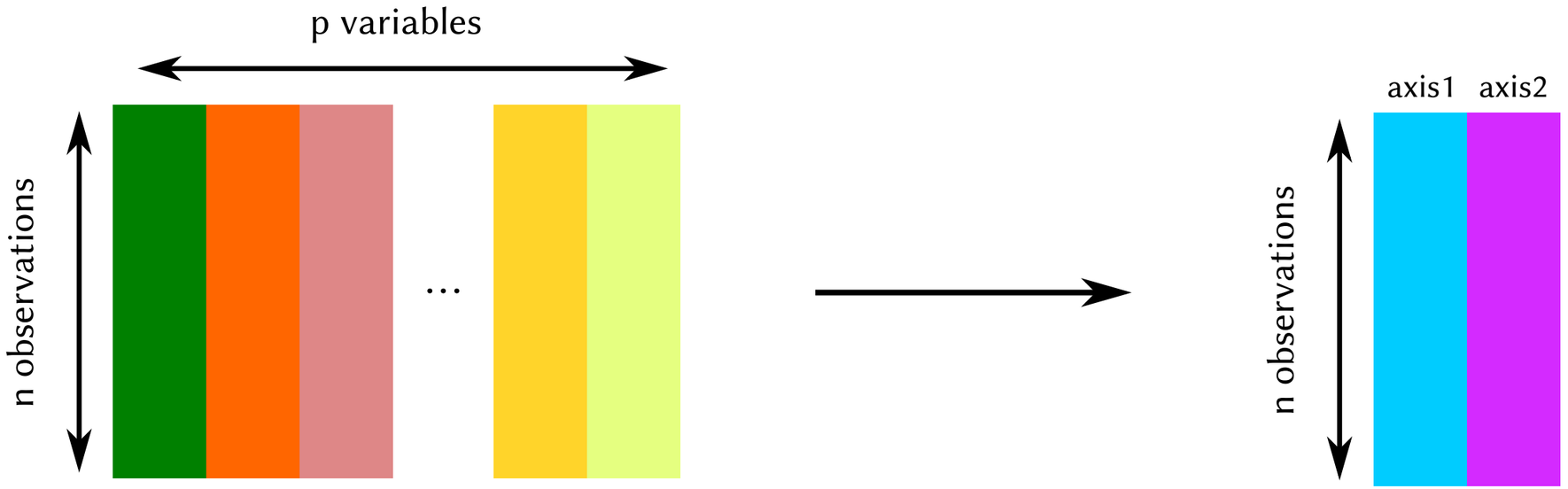


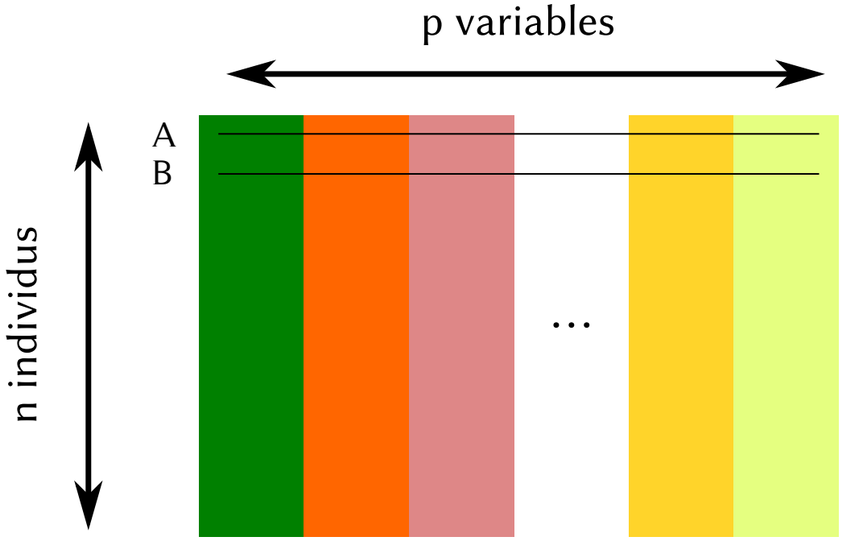

Espèces
Sites
...

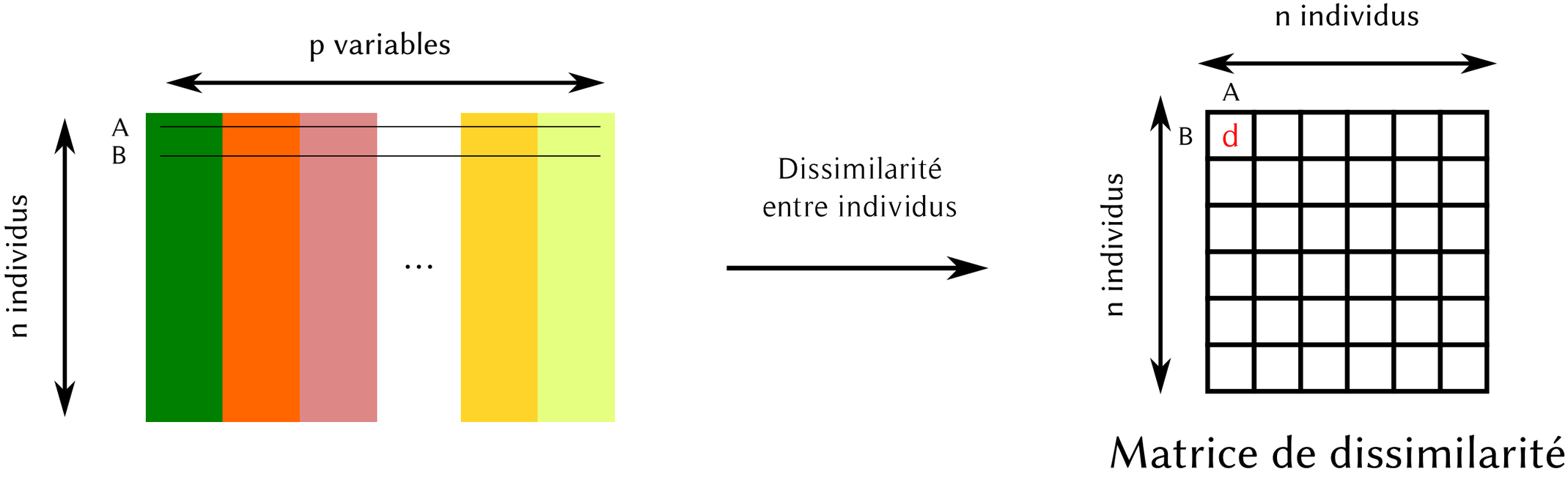

Mesures biométriques
Individus
...


Gènes
Individus
...

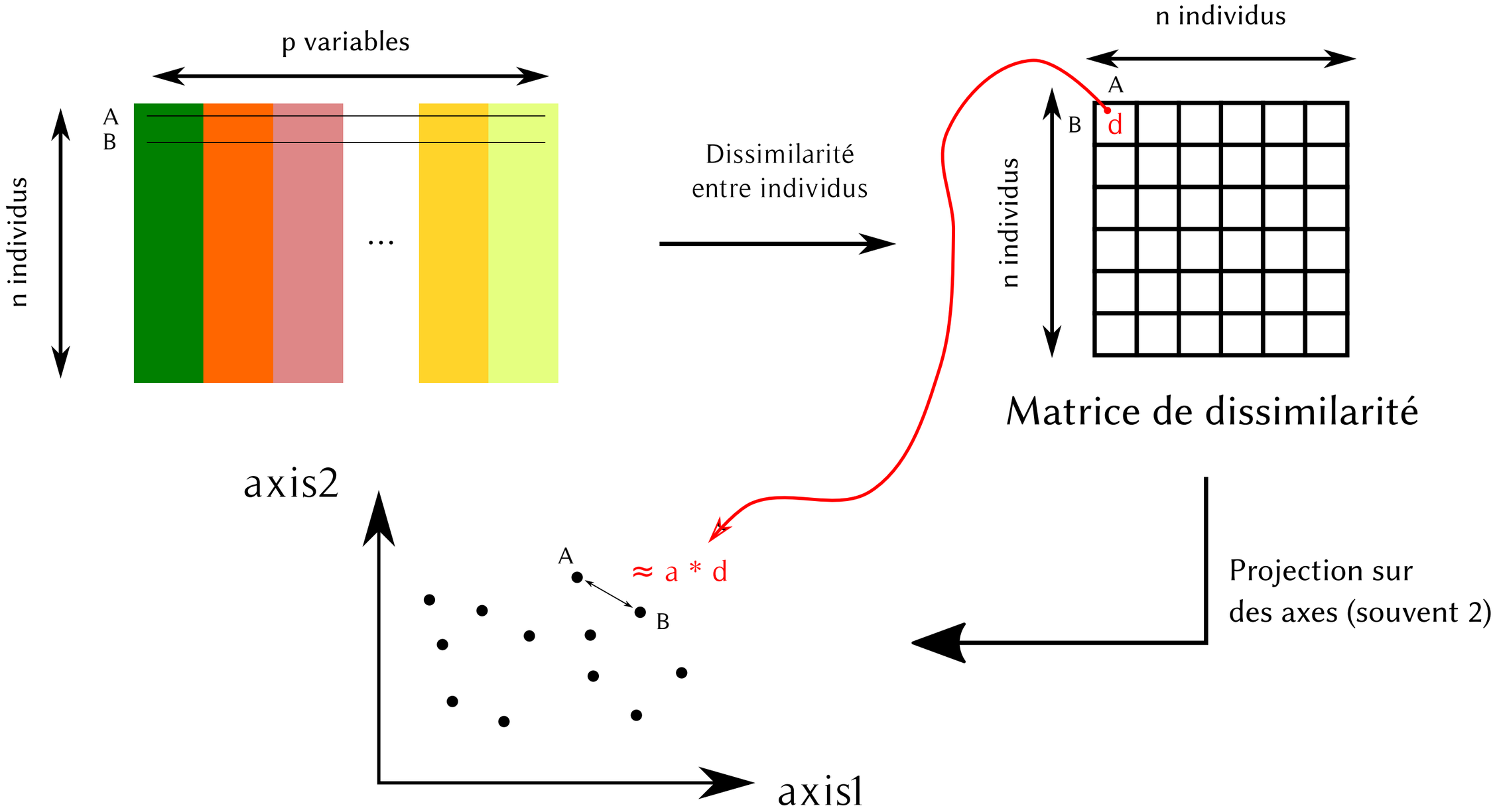
Réduction de dimensionalité

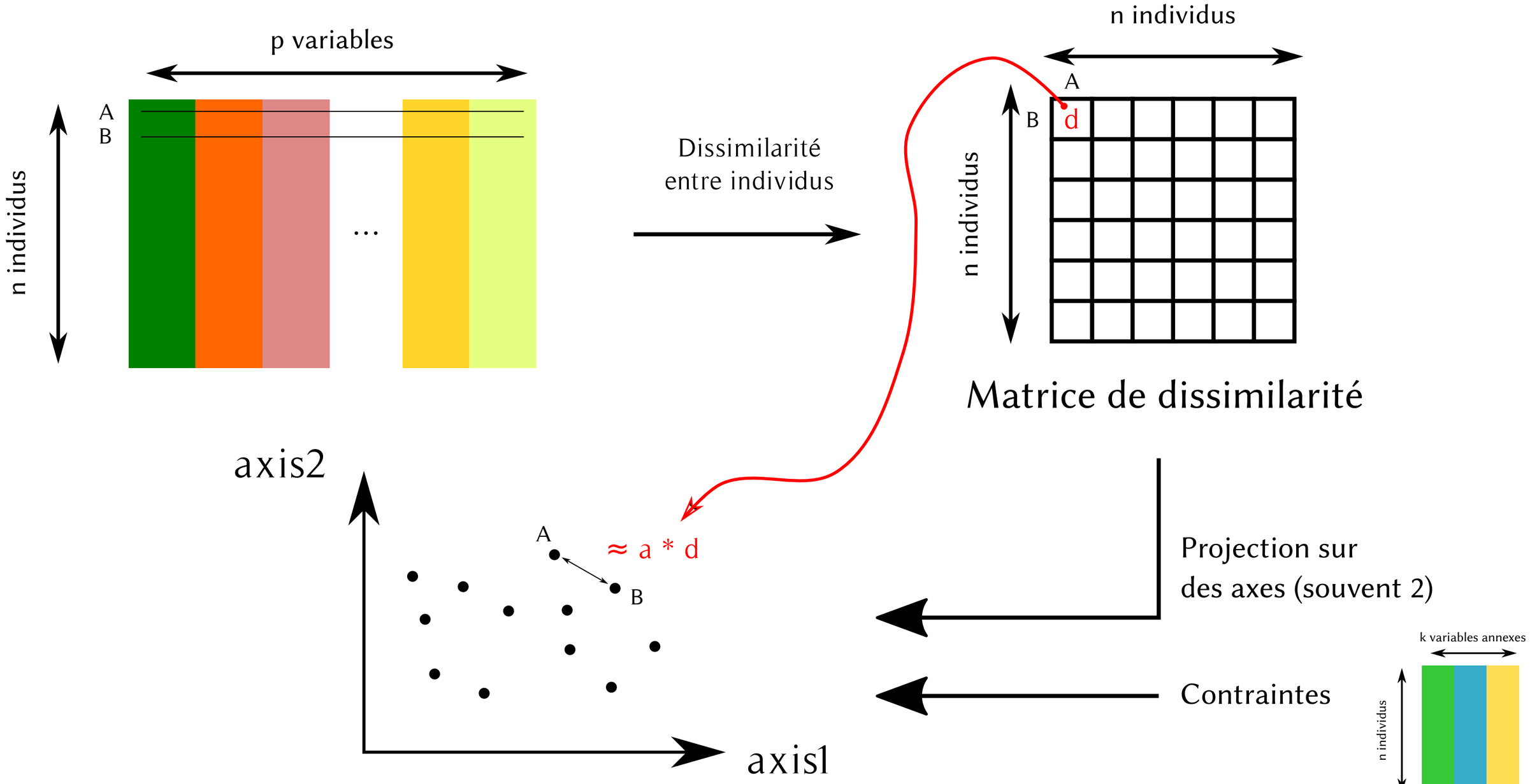


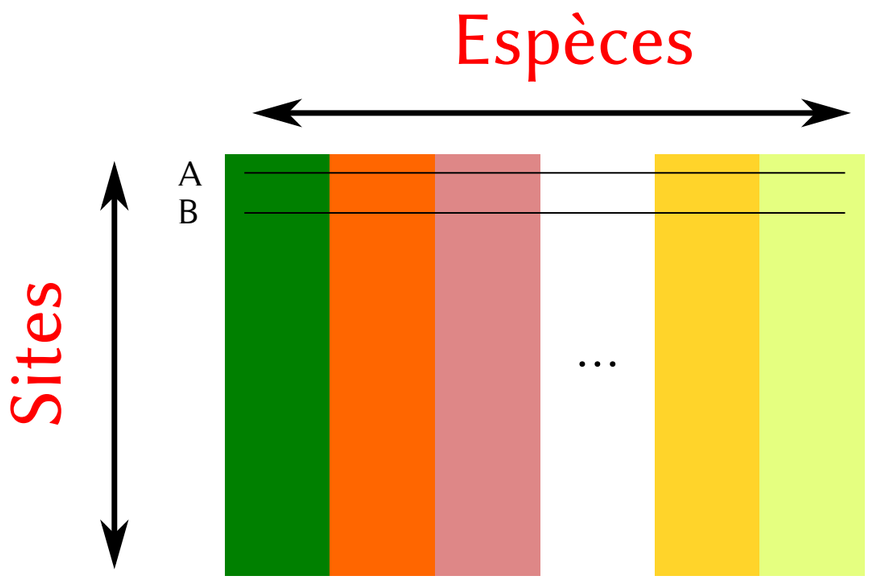

p variables
n individus
A
B
...



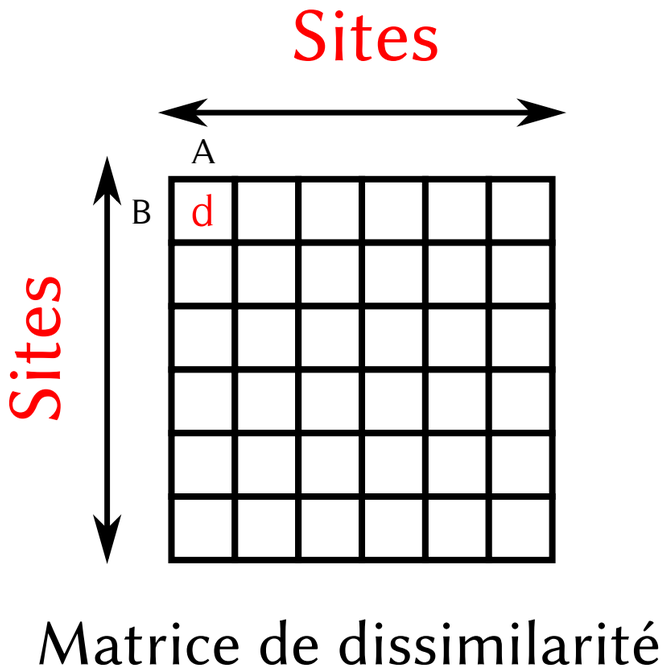
Matrice de dissimilarité


p variables
n individus
A
B
...
Dissimilarité
entre individus
n individus
n individus
A
B
d
Matrice de dissimilarité
axis2
A
≈ a * d
B
axis1
Projection sur
des axes (souvent 2)

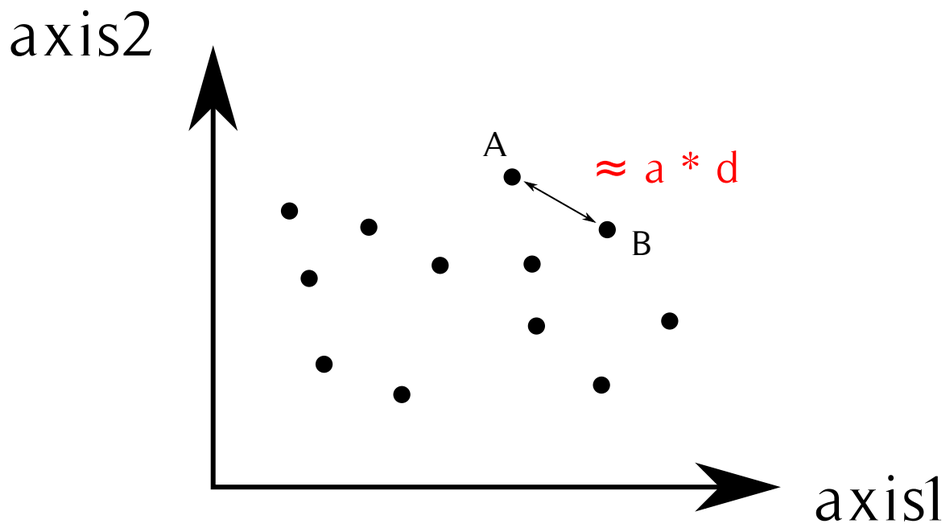

p variables
n individus
A
B
...
Dissimilarité
entre individus
n individus
n individus
A
B
d
Matrice de dissimilarité
axis2
A
≈ a * d
B
axis1
Projection sur
des axes (souvent 2)
Contraintes
k variables annexes
n individus



Dissimilarité
entre **sites**



Matrice de dissimilarité



Projection sur
des axes (souvent 2)

Contraintes

Mesures environementales,
Modalités expérimentales,

(pour des données quantitatives principalement)

Dissimilarité \ Contraintes	Non	Oui
Distance euclidienne données continues	Analyse en Composantes Principales **PCA**	Analyse de redondance Redundancy analysis **RDA**
Distance du χ^2 données >0	Analyse factorielle des correspondances **CA**	Analyse canonique des correspondance **CCA**
Autre dissimilarité	Analyse en coordonnées principales Metric dimensional scaling Principal Coordinates Analysis **PCoA/MDS**	Analyse canonique des coordonnées principales ? Distance-based redundancy analysis Constrained Analysis of Principal Coordinates **dbRDA / Capscale**

Distance euclidienne

site 1

A	B	C	D
80%	10%	0%	0%

20

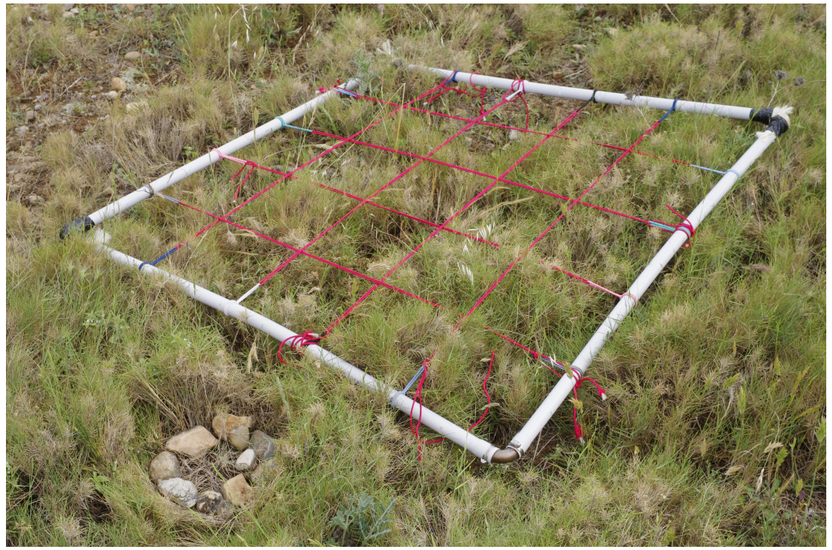
site 2

A	B	C	D
70%	20%	0%	0%

10

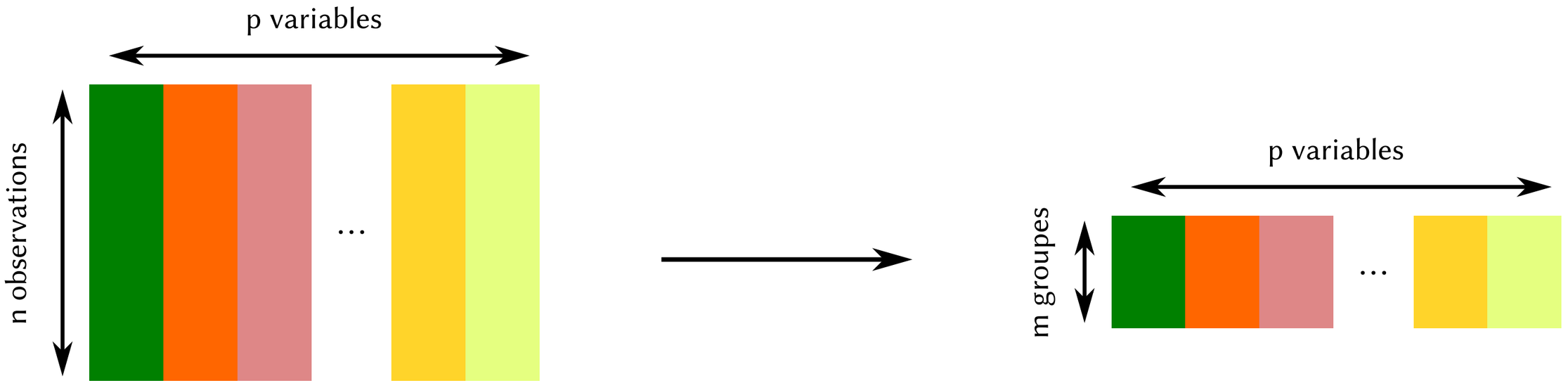
site 3

A	B	C	D
80%	10%	5%	5%

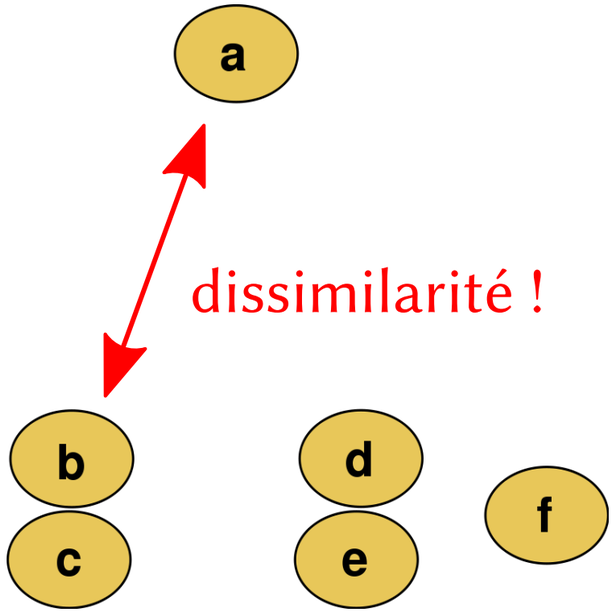


En pratique...

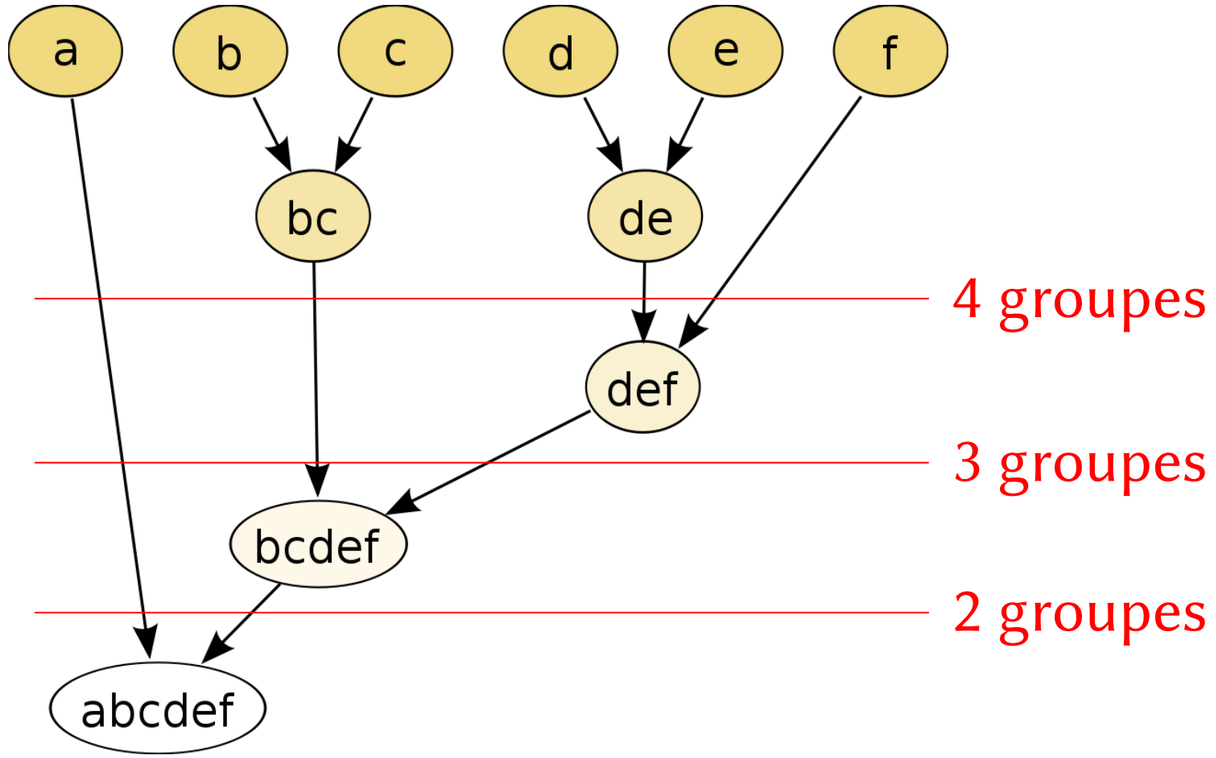
Clustering (partitionnement de données)

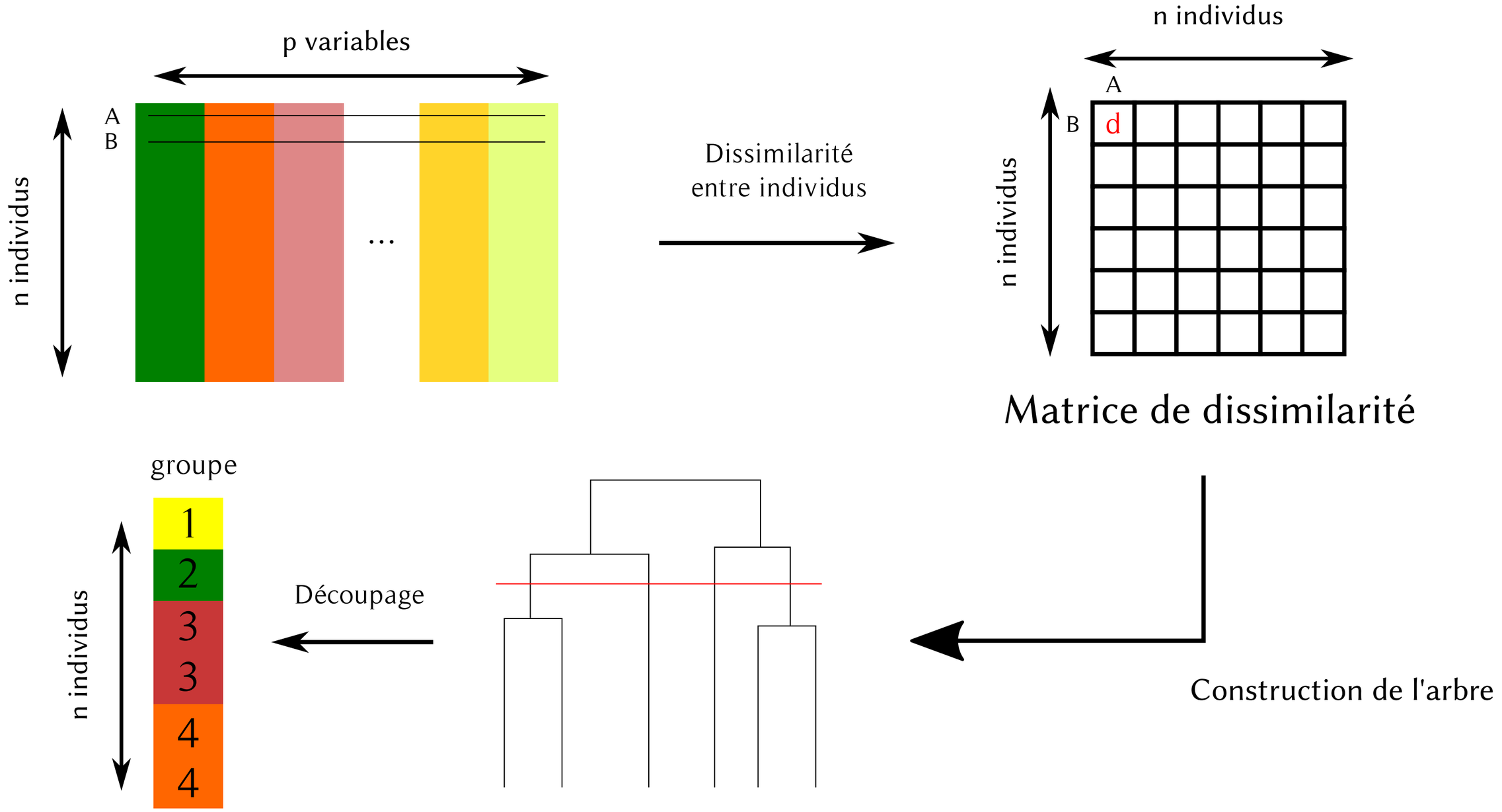


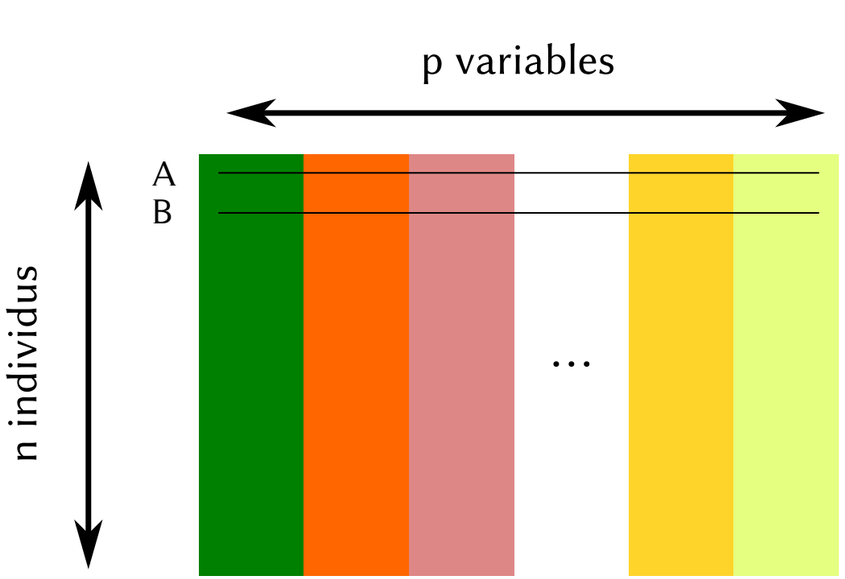
Clustering



(wikipedia)

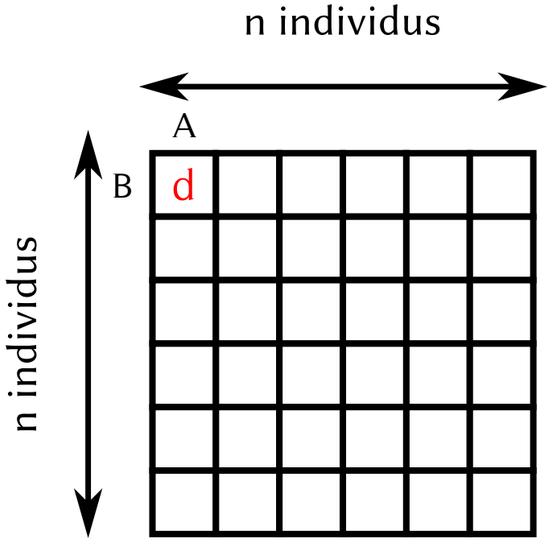



p variables
n individus
A
B
...
Dissimilarité
entre individus
n individus
n individus
A
B
d
Matrice de dissimilarité
Construction de l'arbre
Découpage
groupe
n individus
1
2
3
3
4
4



Dissimilarité
entre individus

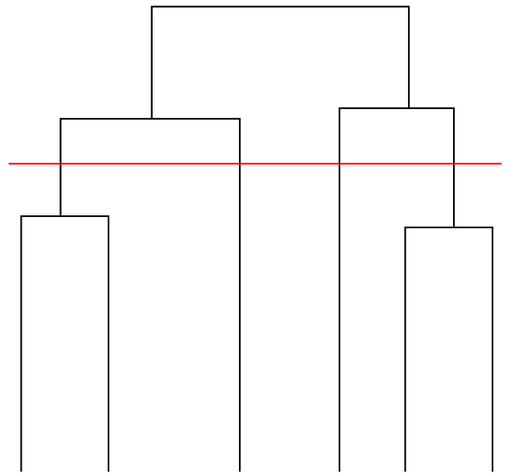
Dissimilarité ?



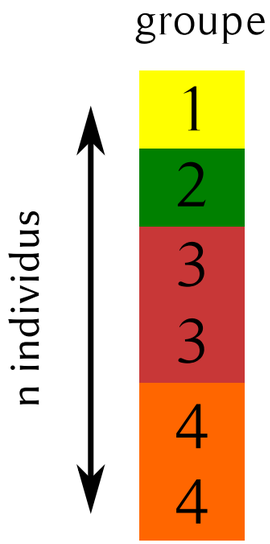
Matrice de dissimilarité

Méthode
d'agglomération ?

Construction de l'arbre

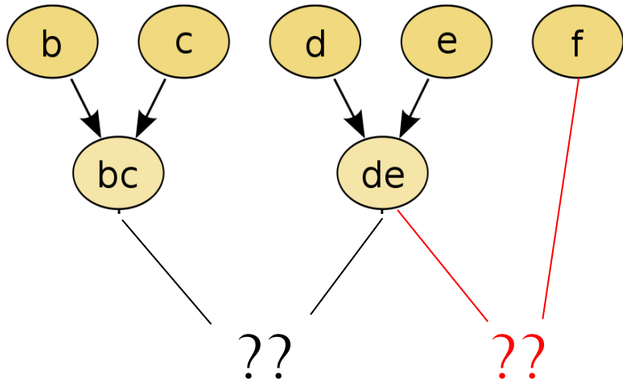


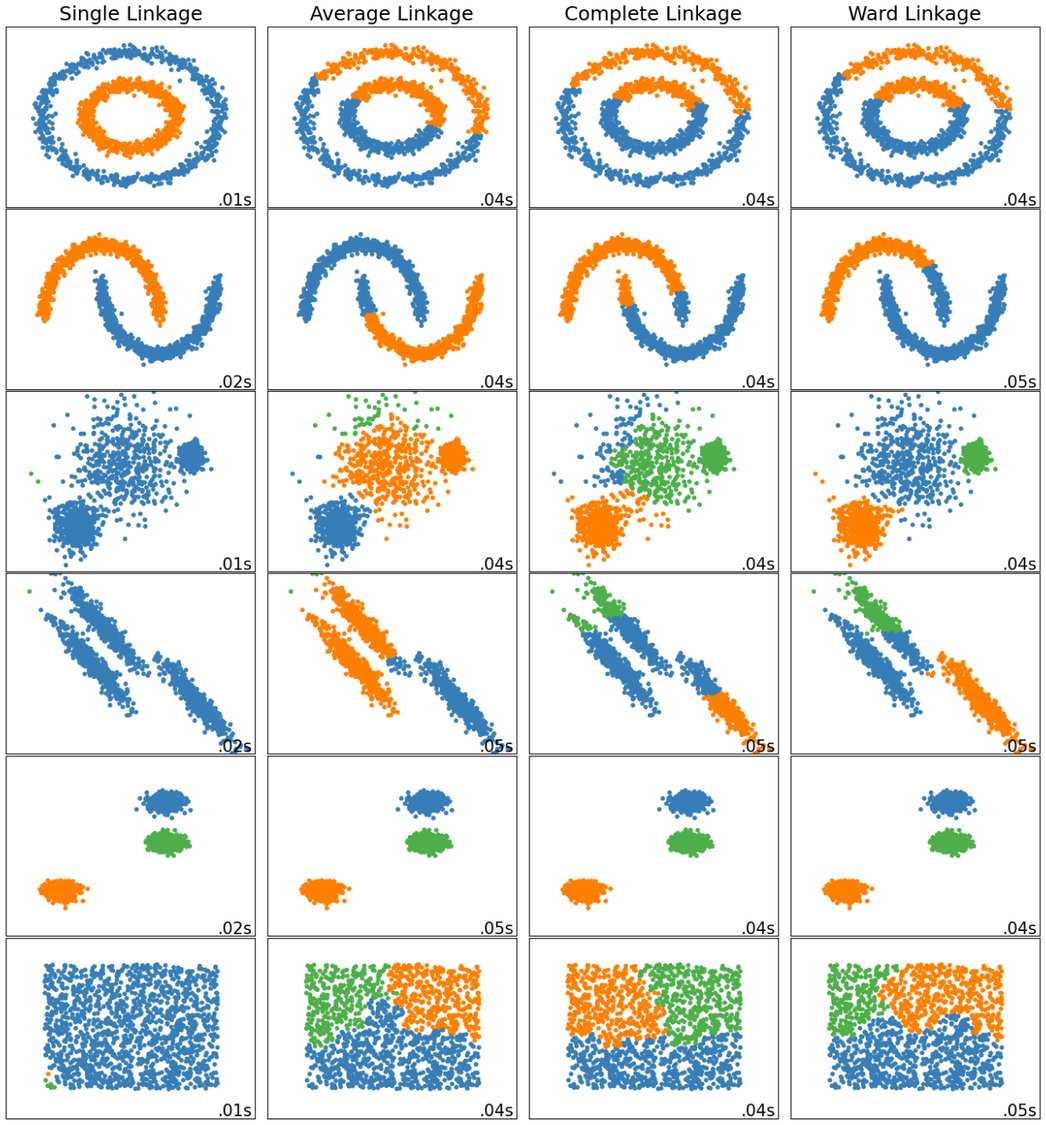
Découpage



Dissimilarité ?

Méthode d'agglomération ?





https://scikit-learn.org/stable/auto_examples/cluster/plot_linkage_comparison.html#sphx-glr-auto-examples-cluster-plot-linkage-comparison-py

En pratique...

Dissimilarité ?

Méthode d'agglomération ?

Combien de cluster dans mes données ?

Quels grands groupes étant donné un certain nombre de clusters ?

Le monde des analyses multivariées est vaste

On cherche à comprendre comment est structuré un jeu de données, pas à faire des tests statistiques

Est-ce que je sais quelque chose de mon objet d'étude qui me permet de privilégier un choix ou un autre ?

Que font les autres dans mon domaine ?

Est-ce que les techniques utilisées font des hypothèses intéressantes pour moi ?

Recap:

- **Visualiser un jeu de données de faible dimensionalité: ggpairs**
- **Un usage rapide de la rda, cca**
- **Un usage rapide du clustering hierarchique**

Atelier: jeudi 12 novembre, 14h !

Tous les exercices et infos sur https://rrr.mbb.cnrs.fr