

# Analyse des données nutritionnelles



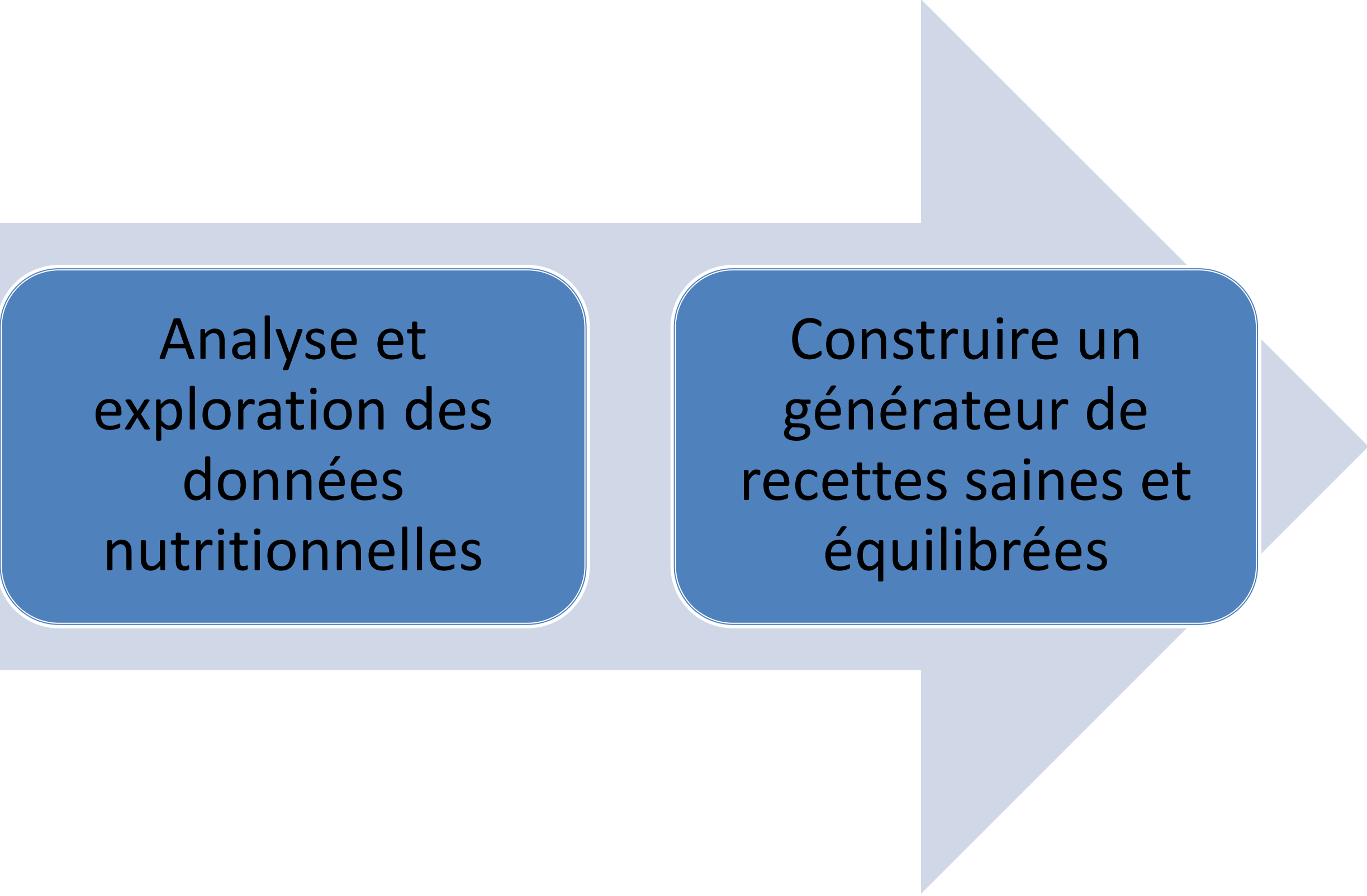
l'information alimentaire ouverte

Adil DHISSA

# Sommaire

- ① Objectif du projet
- ② Contexte
- ③ Enjeux
- ④ Méthodes de nettoyage des données
- ⑤ Exploration multivariable des données
- ⑥ Conclusions et perspectives

# Objectif du projet

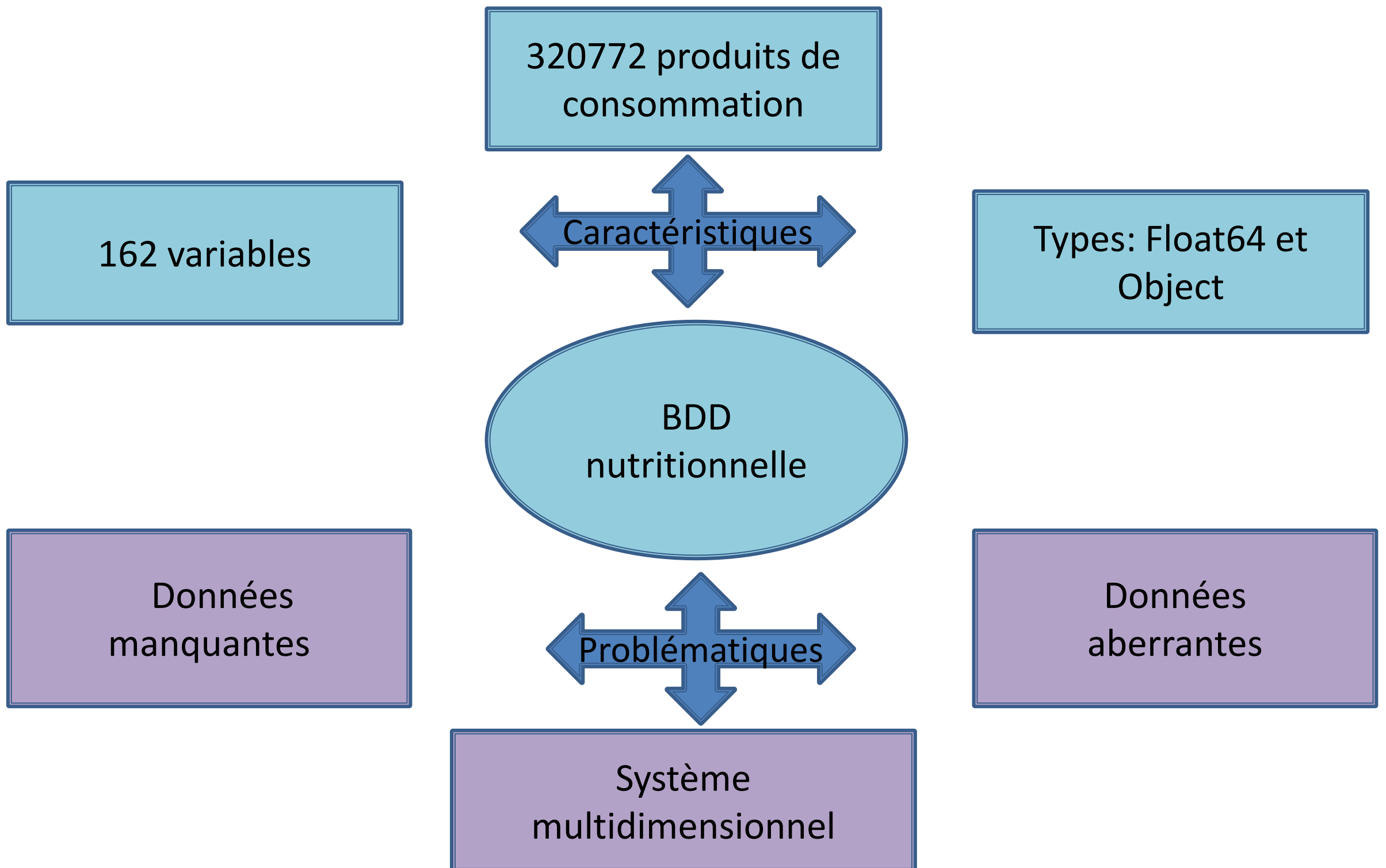


Analyse et  
exploration des  
données  
nutritionnelles

Construire un  
générateur de  
recettes saines et  
équilibrées

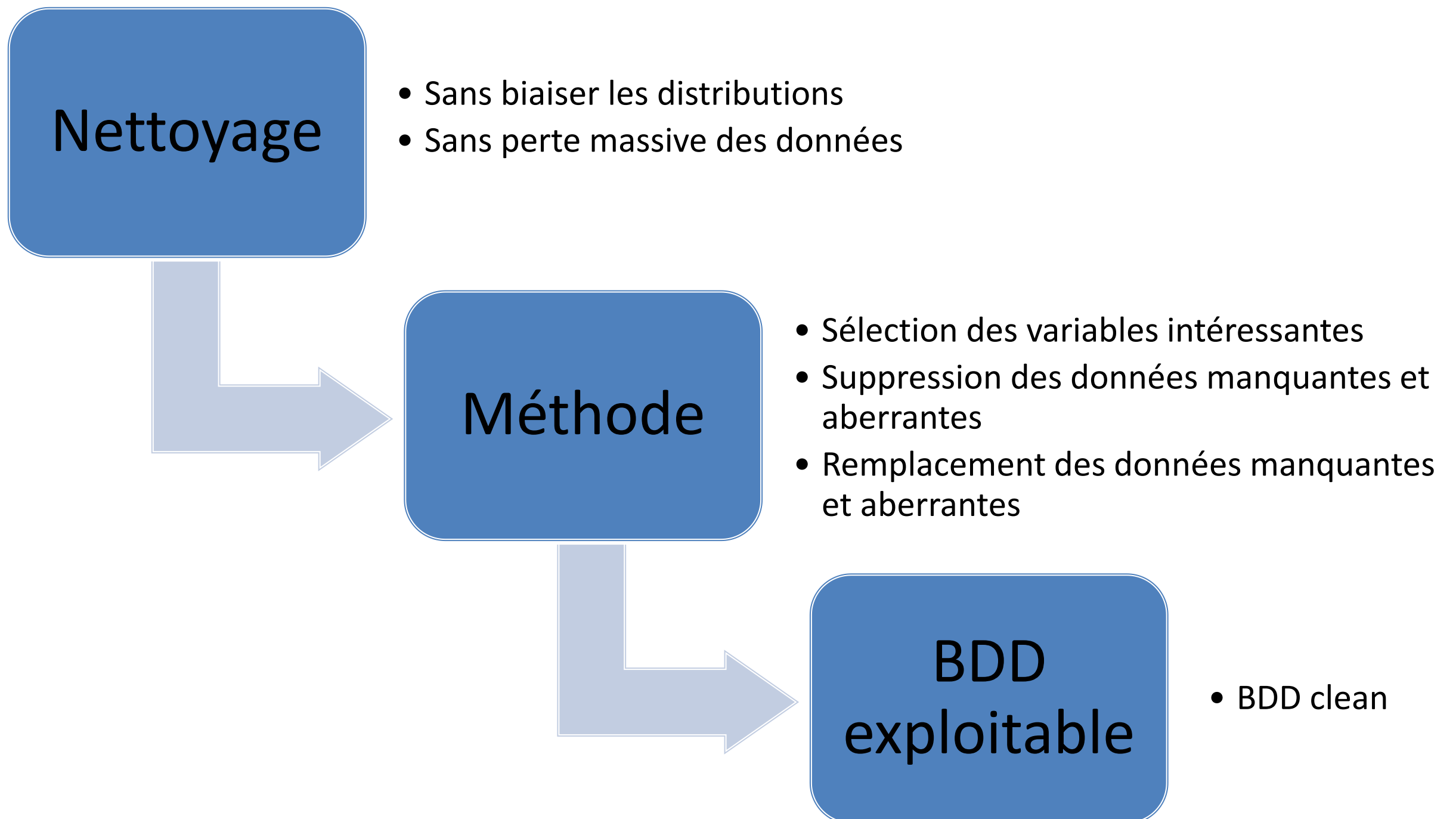
# Contexte

- ## ❑ Base de données nutritionnelles « openfoodfacts »



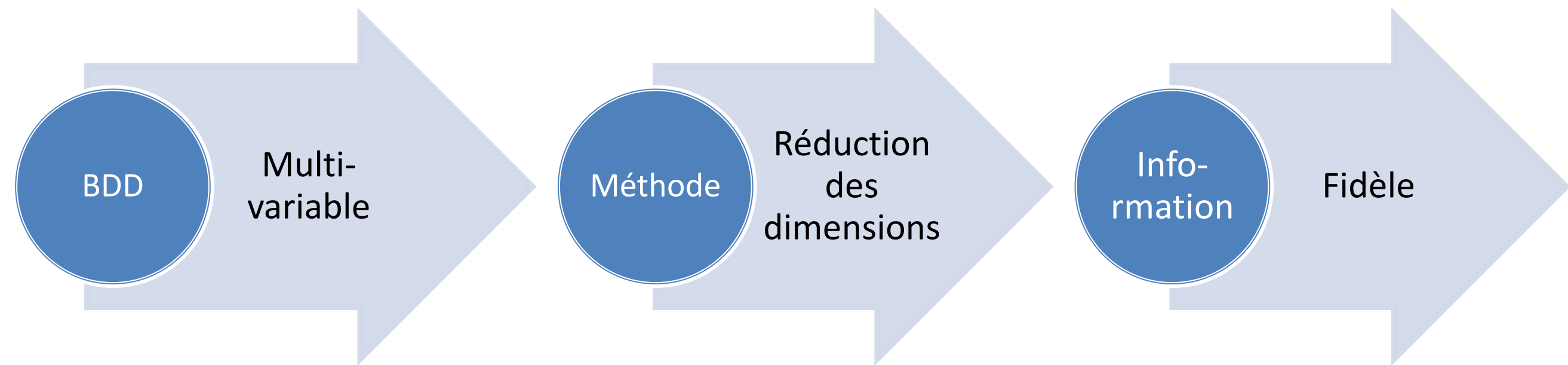
# Enjeux

## ❑ Problématique des données manquantes et aberrantes



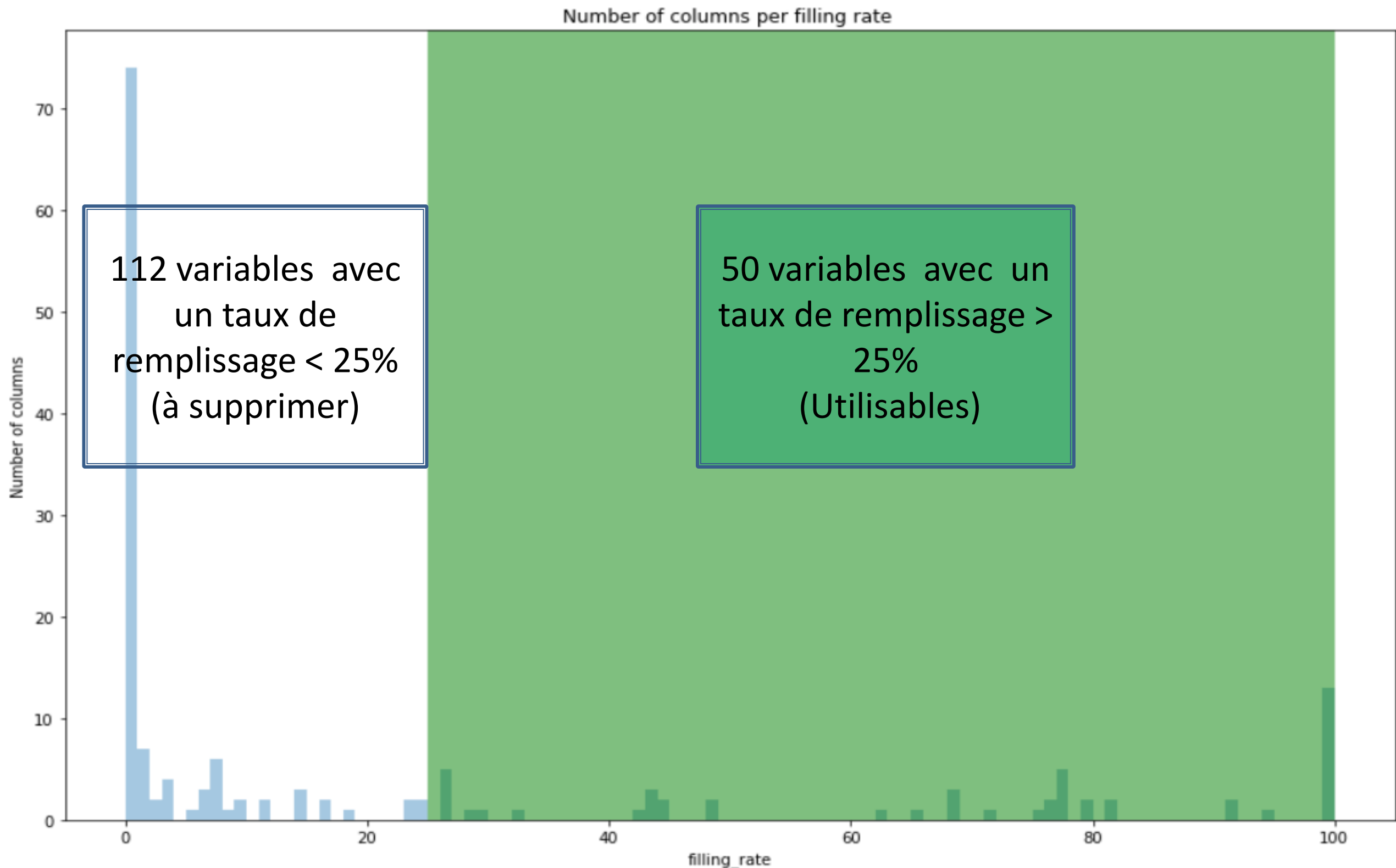
# Enjeux

## ❑ Problématique de la multi-variabilité de la BDD



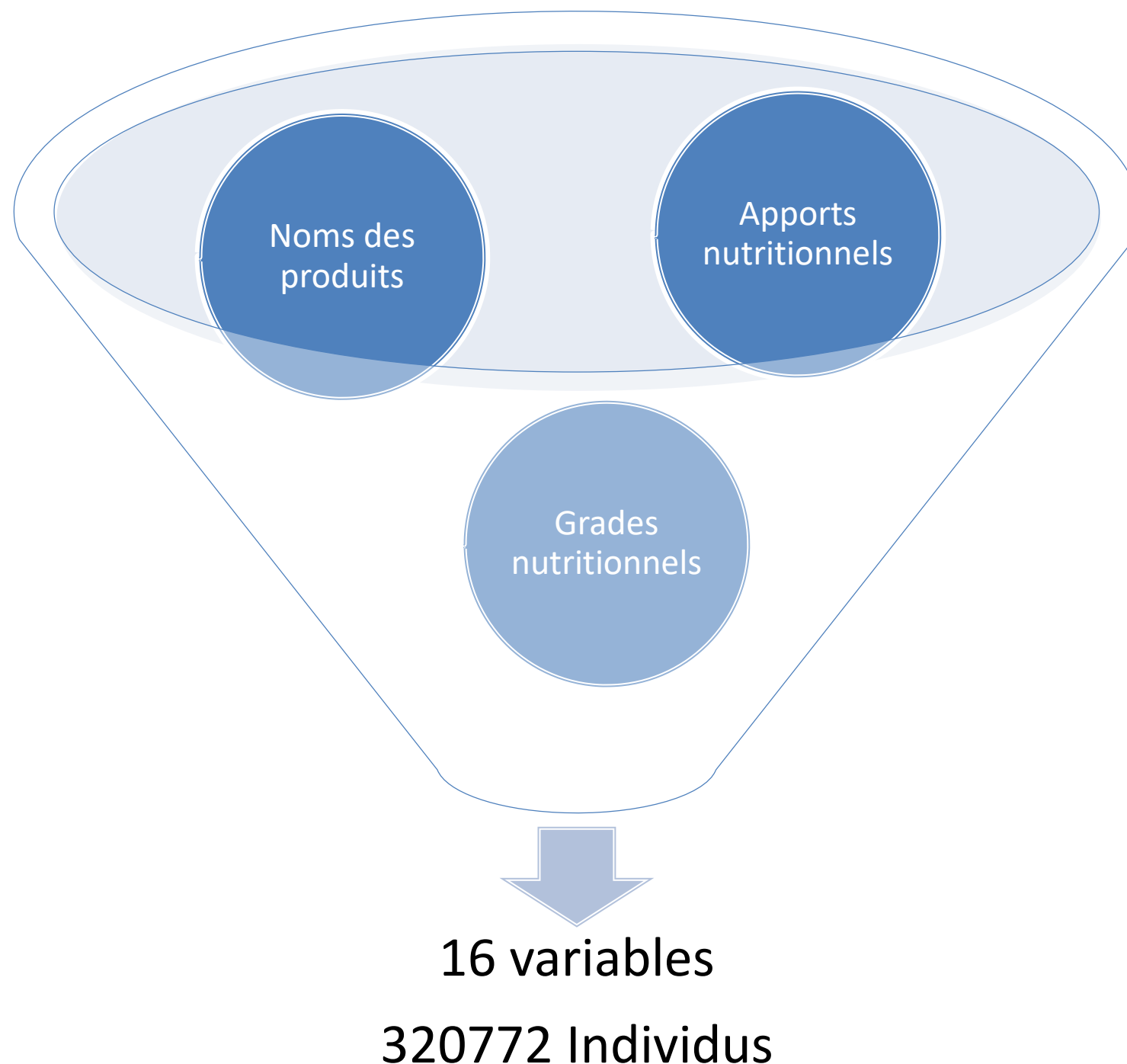
# Méthodes de nettoyage des données

## ❑ Filtrage par seuil de remplissage à 25%



# Méthodes de nettoyage des données

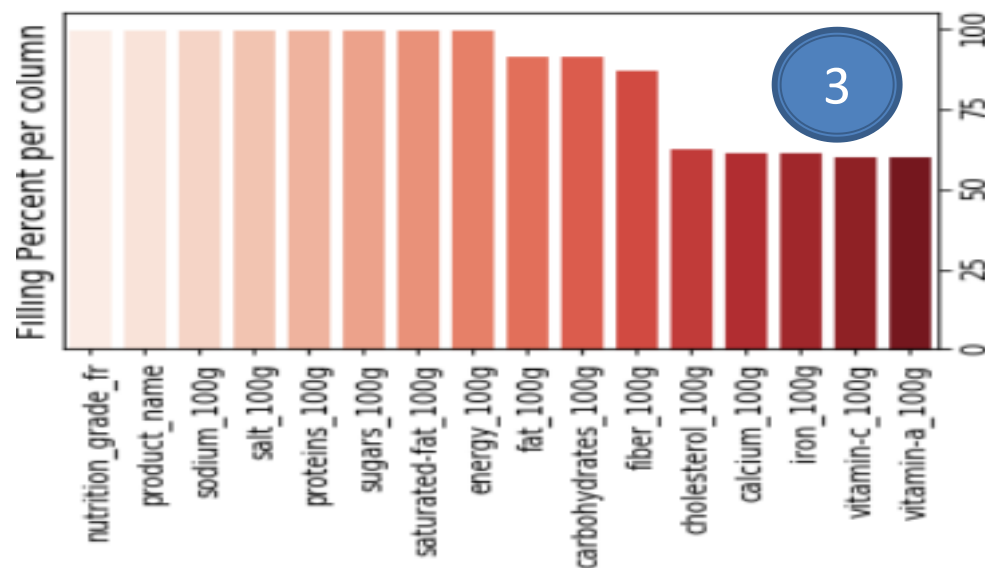
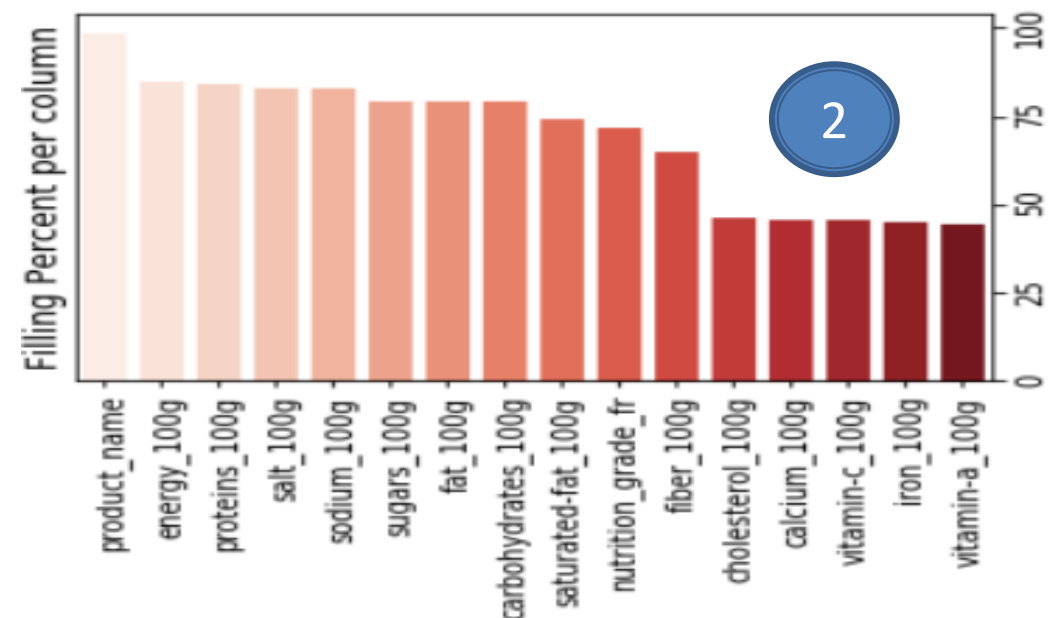
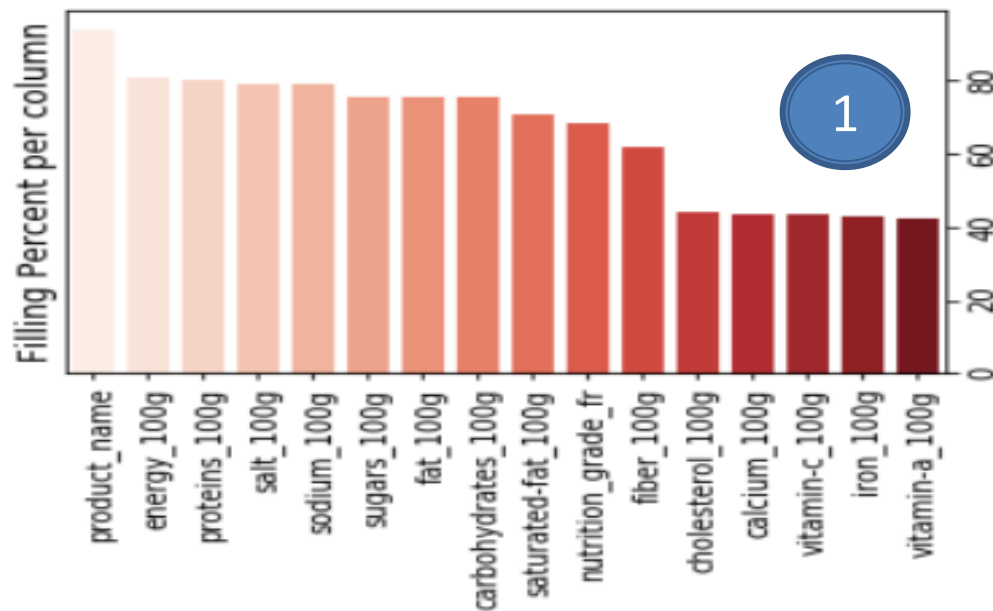
- ❑ Sélection des variables intéressantes pour l'analyse





# Méthodes de nettoyage des données

## ❑ Suppression des lignes



1. Taux de remplissage du DataFrame avec 16 variables et 320772 produits
2. Suppression des lignes contenant que des NaNs : Augmentation du taux de remplissage de toutes les variables du nouveau DataFrame avec 16 variables et 306393
3. Suppression des lignes dont le nom des produits et le grade nutritionnel sont des NaNs: Idem que 2 pour le nouveau DataFrame avec 16 variables et 218456 produits (Dimensions finales)

# Méthodes de nettoyage des données

- ❑ Remplacement des données manquantes et aberrantes

Suppression



Perte massive d'information

Remplacement par la  
moyenne-Médiane



Biaiser les distributions

# Méthodes de nettoyage des données

## ❑ Création des valeurs manquantes

	energy_100g	fat_100g	saturated-fat_100g	cholesterol_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g
count	2.182700e+05	201010.000000	218270.000000	137077.000000	200982.000000	218270.000000	191715.000000	218270.000000	218270.000000
mean	1.195373e+03	13.337201	4.954507	0.019242	33.249956	15.004324	2.829103	7.778696	1.646219
std	7.031389e+03	16.192746	7.650321	0.366696	28.341681	21.191652	13.078449	8.131484	138.277716
min	0.000000e+00	0.000000	0.000000	0.000000	0.000000	-17.860000	0.000000	-3.570000	0.000000
25%	4.520000e+02	0.833000	0.000000	0.000000	7.000000	1.280000	0.000000	1.900000	0.100000
50%	1.193000e+03	7.140000	1.790000	0.000000	24.000000	5.000000	1.500000	5.700000	0.650000
75%	1.715000e+03	21.430000	7.140000	0.020000	59.090000	23.000000	3.600000	10.710000	1.361440
max	3.251373e+06	714.290000	550.000000	95.238000	209.380000	3520.000000	5380.000000	430.000000	64312.800000

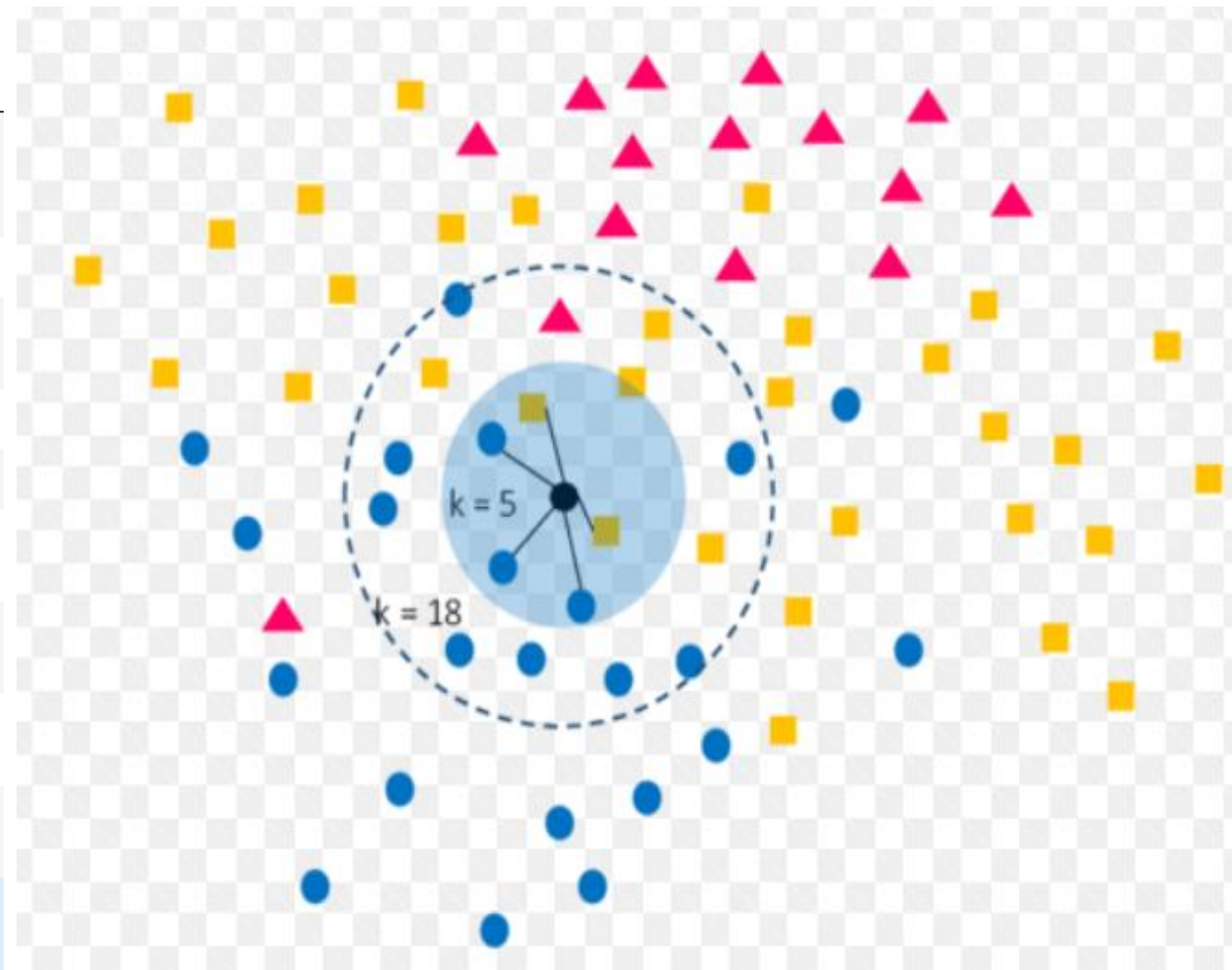
- Remplacement des valeurs négatives par NaNs
- L'écart inter-quartile est la différence entre le 3e quartile et le 1e quartile :  $IQ = Q3 - Q1$
- Outliers: Valeurs  $> Q3 + 1,5IQ \Rightarrow$  Remplacement par des NaNs

# Méthodes de nettoyage des données

## ❑ Méthode KNN

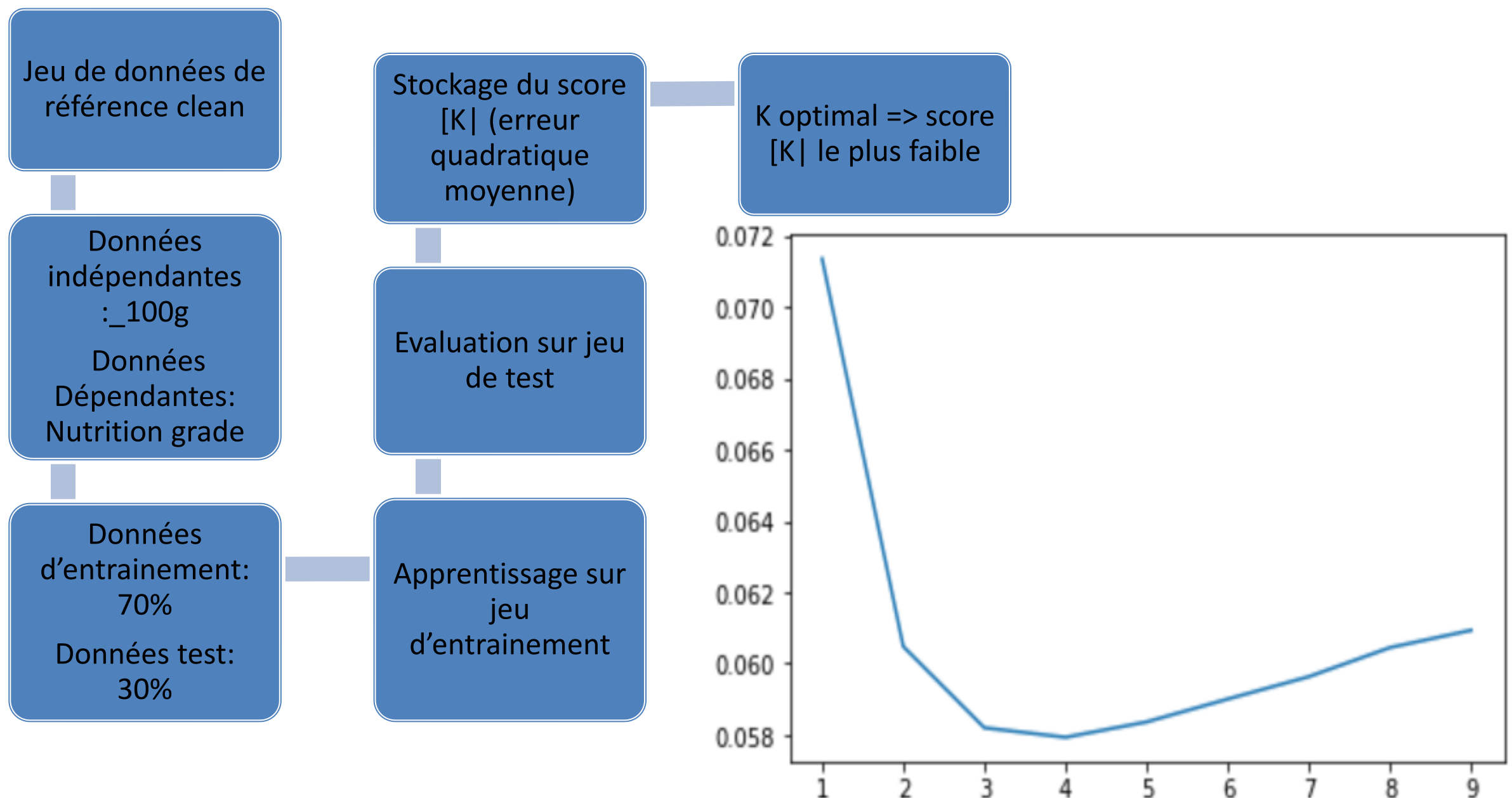
- Méthode d'apprentissage supervisé
- Calcul de la distance euclidienne
- Si 5-NN, la valeur imputée est la moyenne des 5 valeurs les plus proches

proteins_100g	salt_100g	sodium_100g	vitamin-a_100g	vitamin-c_100g	calcium_100g	iron_100g
3.570	0.00000	0.000000	0.000000	0.0214	0.000	0.00129
17.860	0.63500	0.250000	0.000000	0.0000	0.071	0.00129
17.860	1.22428	0.482000	NaN	NaN	0.143	0.00514
14.060	0.13970	0.055000	NaN	NaN	0.062	0.00422
16.670	1.60782	0.633000	NaN	NaN	0.133	0.00360
14.550	0.02286	0.009000	0.000273	NaN	NaN	0.00131
14.290	0.01016	0.004000	NaN	0.0064	0.107	0.00450
16.670	0.02540	0.010000	NaN	NaN	0.048	0.00429
7.500	0.28448	0.112000	0.000075	0.0225	0.100	0.00180
13.330	0.46482	0.183000	0.000100	NaN	0.067	0.00240



# Méthodes de nettoyage des données

## ❑ K Optimal



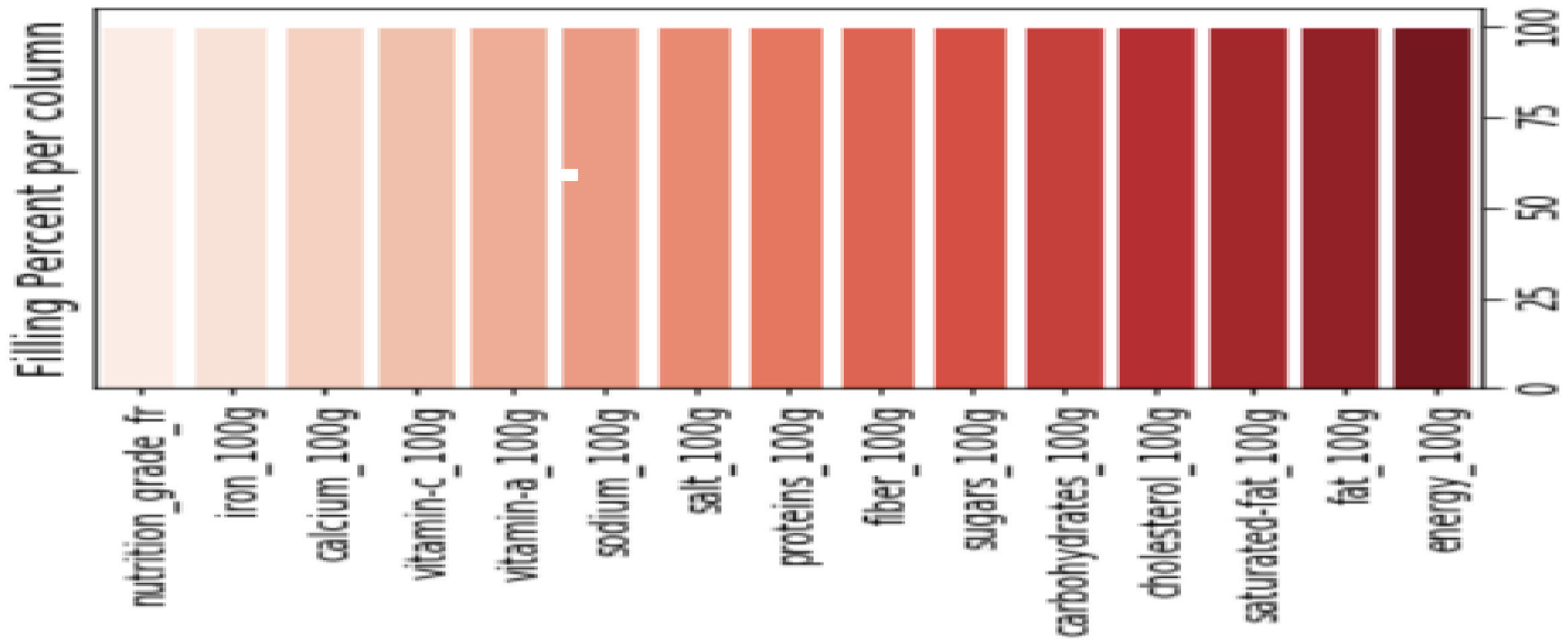


# Méthodes de nettoyage des données

## ❑ Résultats

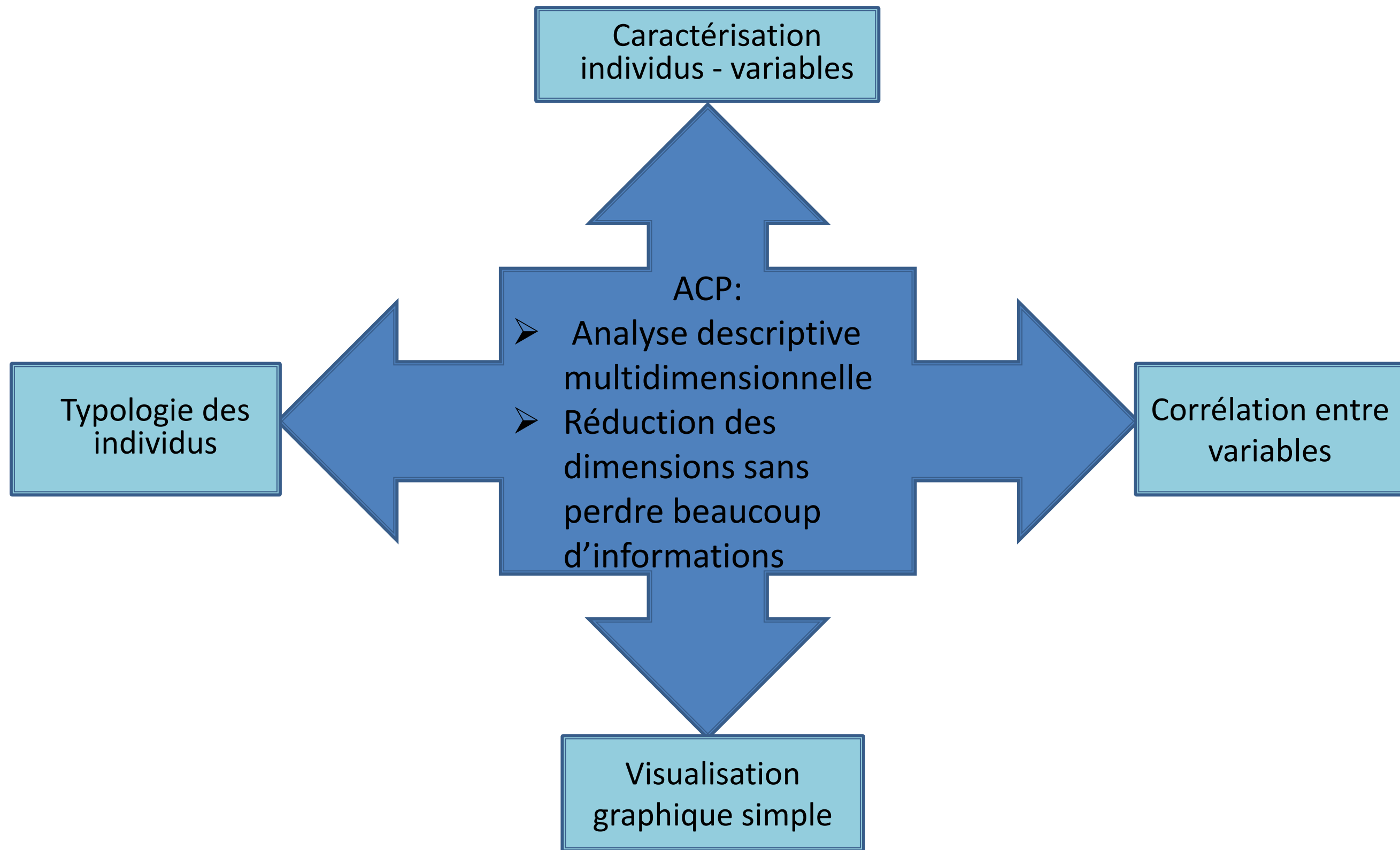
vitamin-a_100g	vitamin-c_100g	vitamin-a_100g	vitamin-c_100g
0.000000	0.0214	0.000000	0.02140
0.000000	0.0000	0.000000	0.00000
NaN	NaN	0.000000	0.00020
NaN	NaN	0.000171	0.00342
NaN	NaN	0.000643	0.02142

Dimensions finales:  
218456 individus  
16 variables



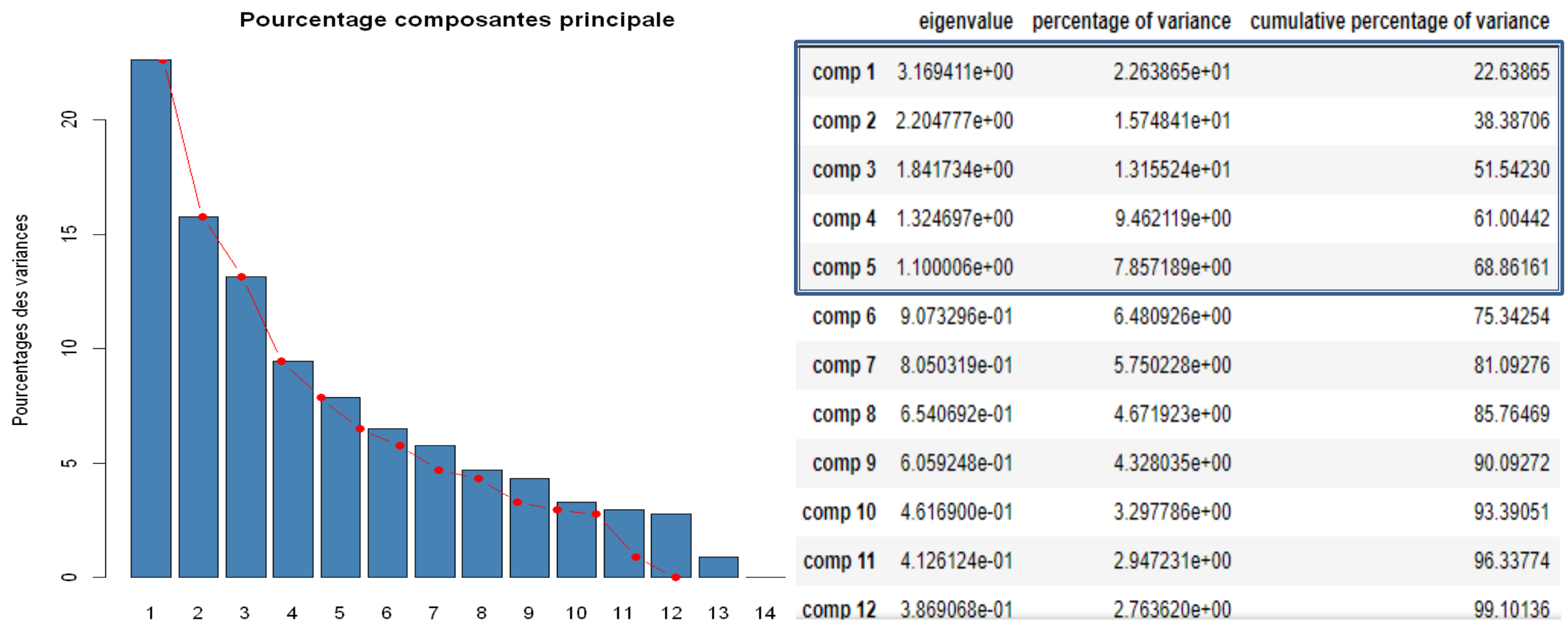
# Exploration multivariables des données

## ❑ Analyse en composantes principales



# Exploration multivariable des données

## ❑ Choix du nombre de composantes Principales



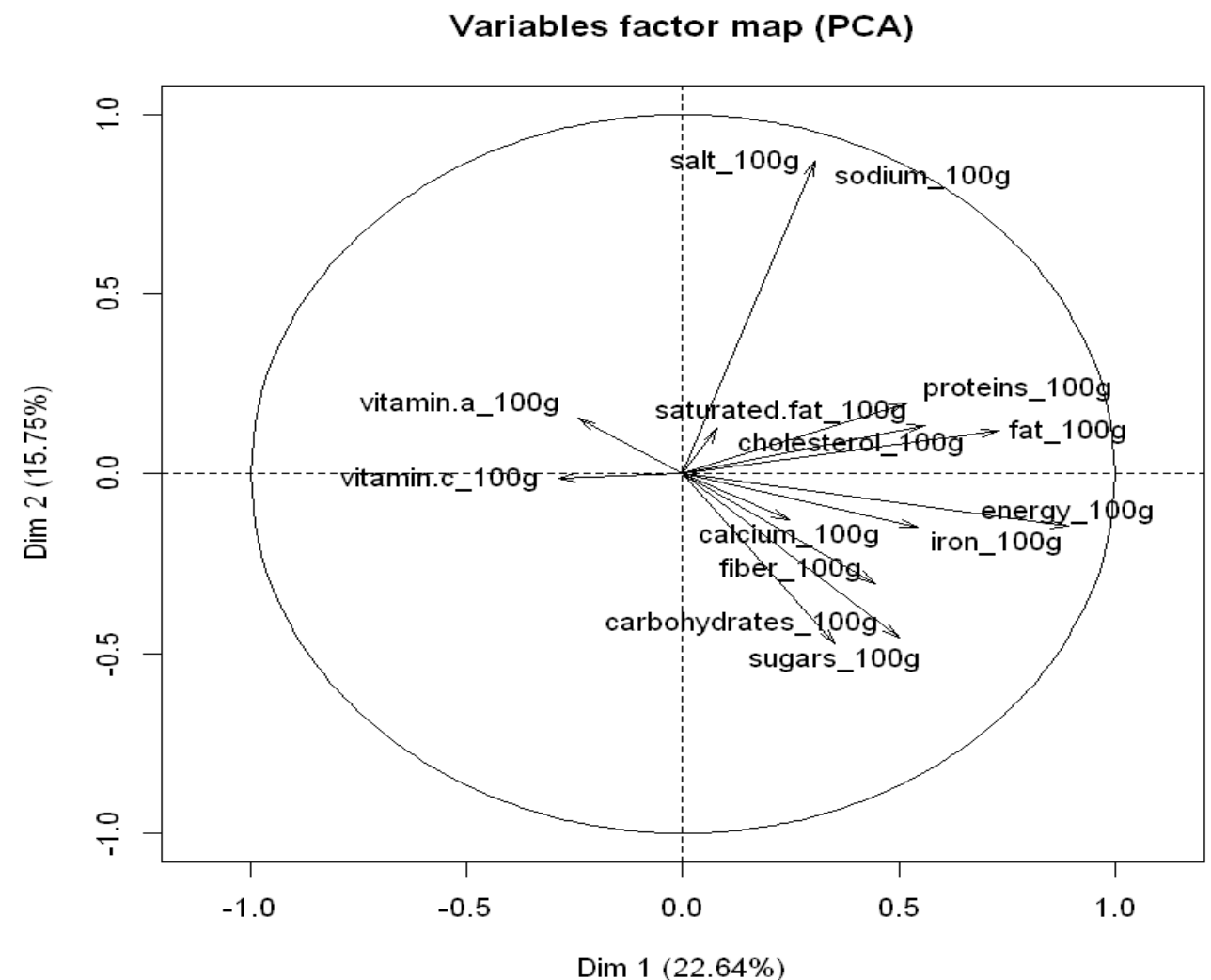
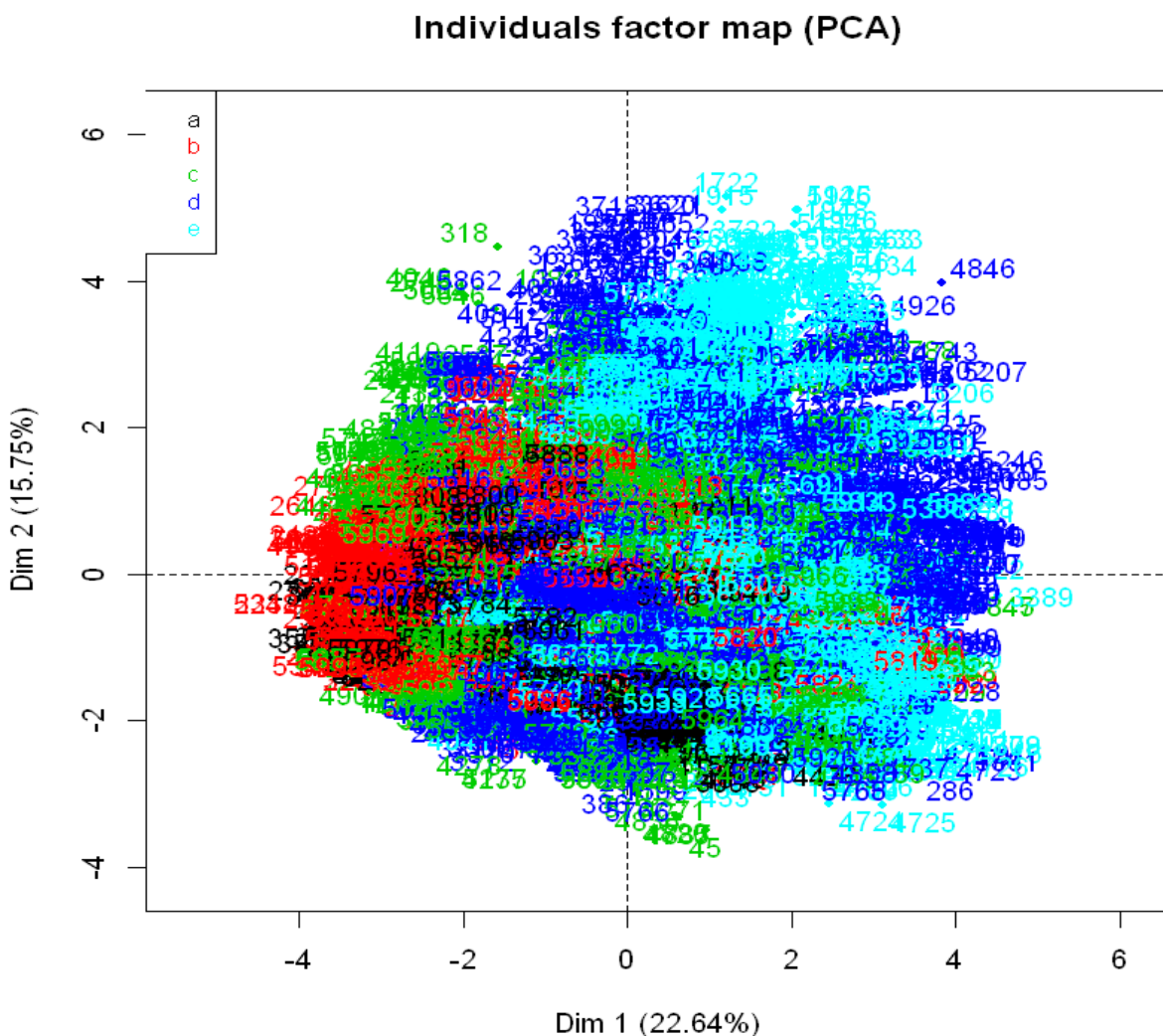
Critère de Kaiser:

- 5 composantes principales
- Explication de 68,86 % de l'information de la BDD



# Exploration multivariable des données

## □ Représentation graphique des composantes 1 et 2



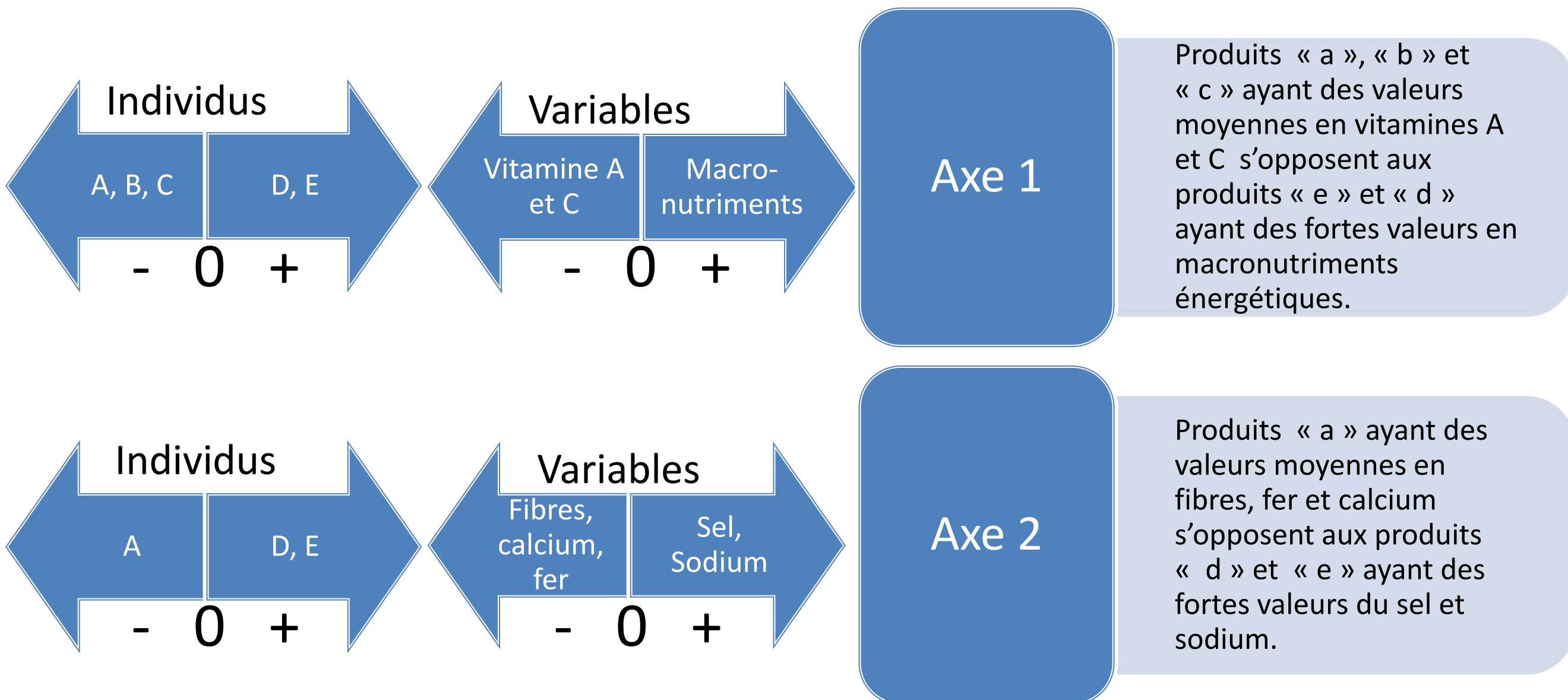
- 2 variables proches
- 2 variables opposées
- 2 variables orthogonales
- Produits-variable occupent même espace



- Corrélées positivement
- Corrélées négativement
- Non corrélées
- Produits caractérisés par des fortes valeurs de la variable

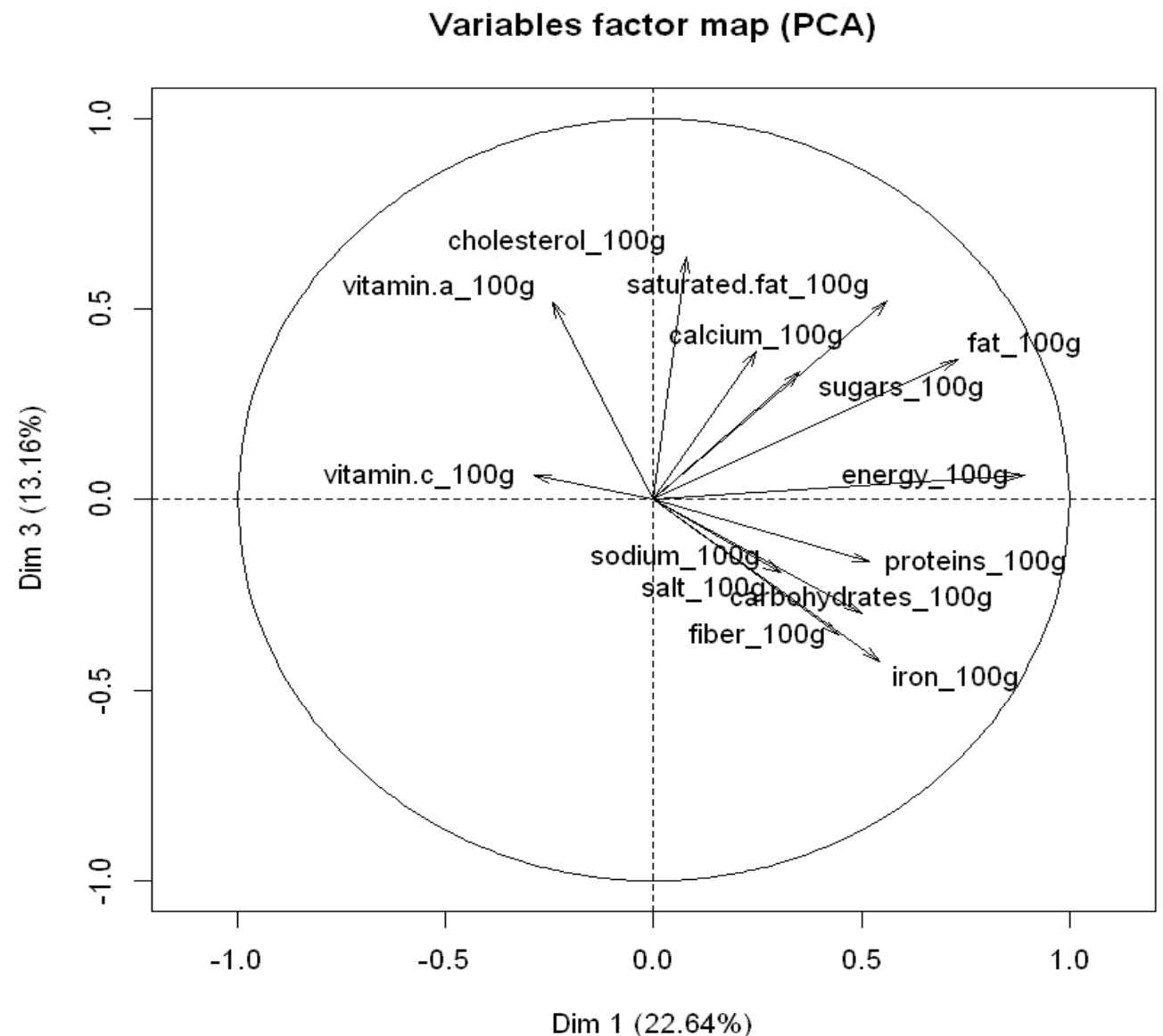
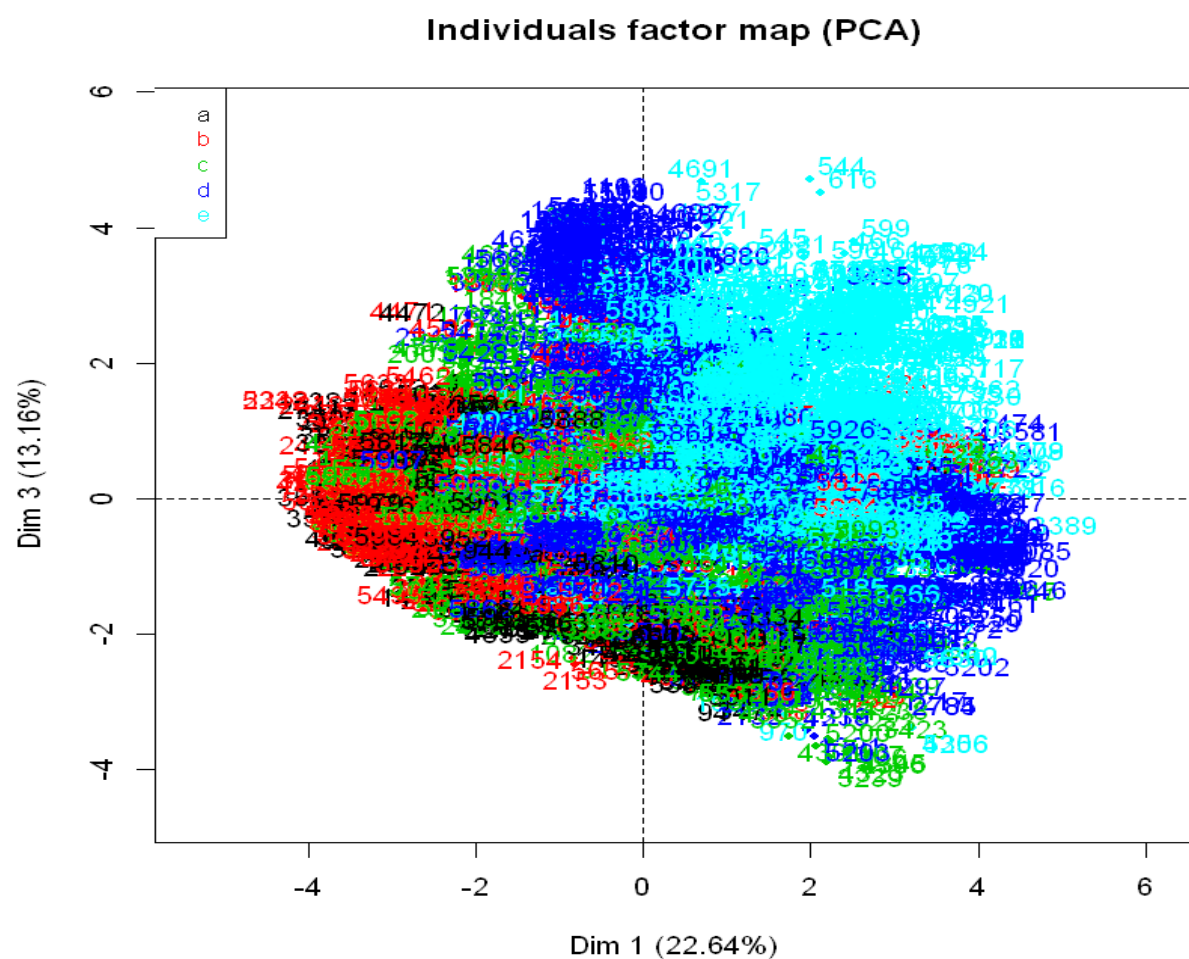
# Exploration multivariable des données

## ❑ Interprétation Axe 1 et 2



# Exploration multivariable des données

## □ Interprétation Axe 1 et 3



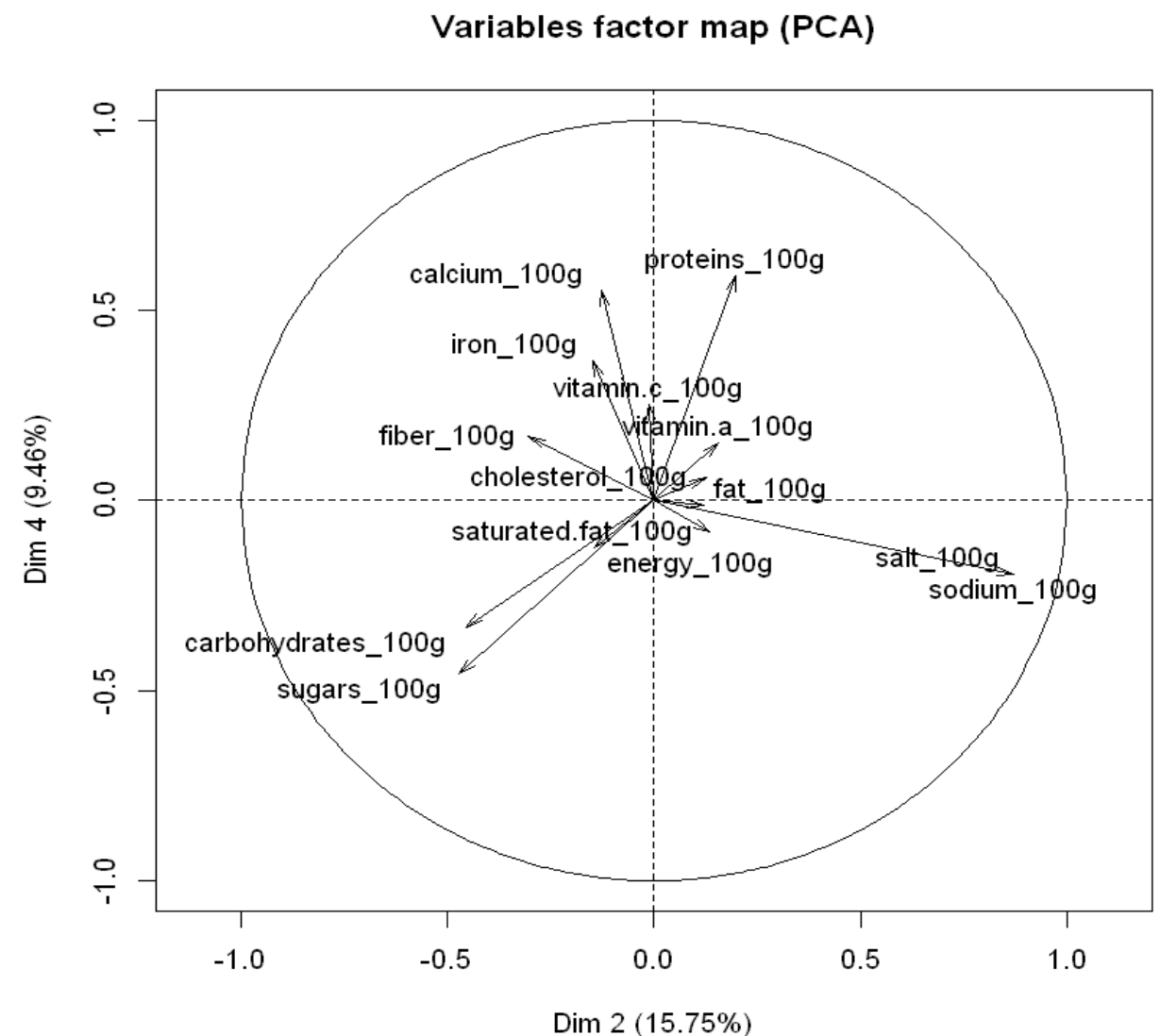
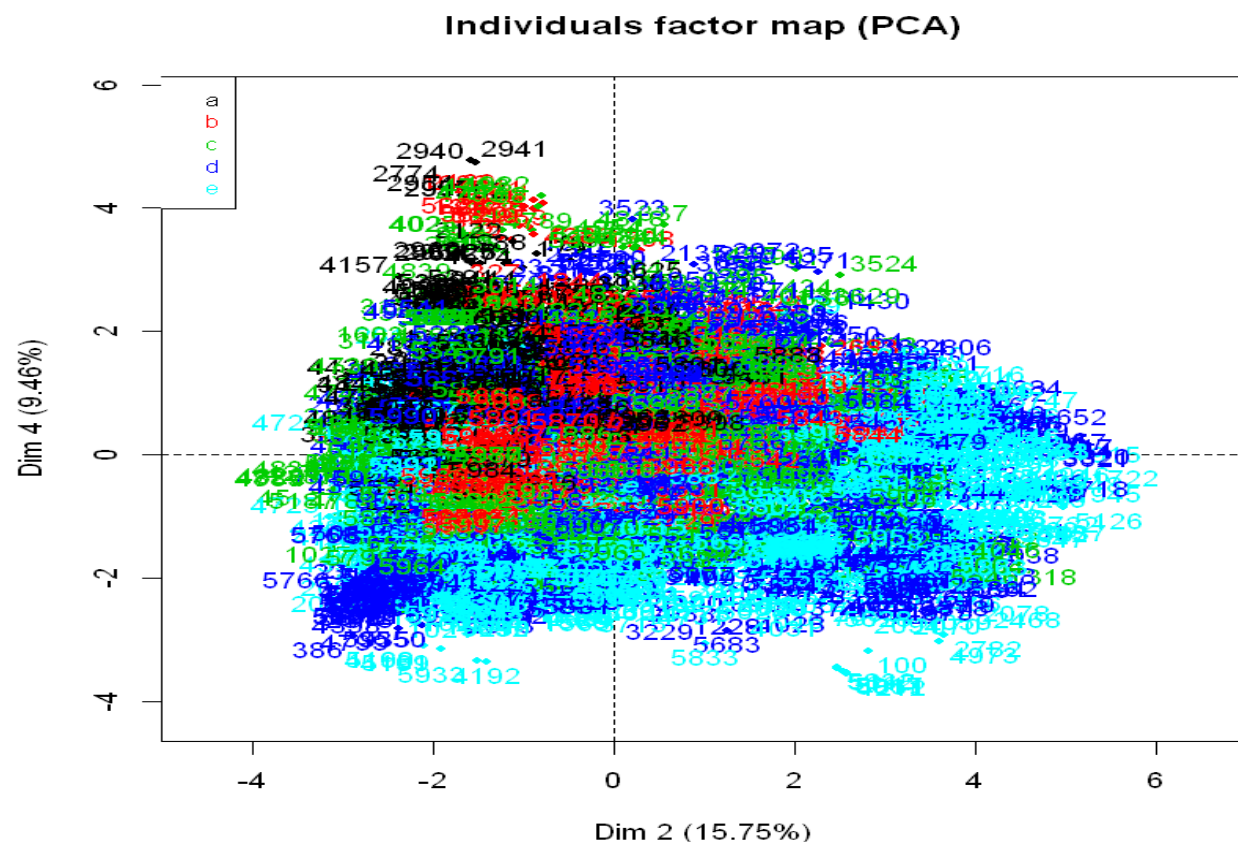
## □ Conclusions Axe 1 et 3

- L'axe 1: Produits «a», «b» et «c» ayant des valeurs moyennes de vitamines A et C s'opposent aux produits «e» et «d» ayant des fortes valeurs en macronutriments énergétiques.
- L'axe 3: Produits «a» et «c» ayant des fortes valeurs en fibres et fer s'opposent aux produits «d» et «e» ayant une forte présence des macronutriments énergétiques et le calcium.



# Exploration multivariable des données

## ❑ Interprétation Axe 2 et 4



## ❑ Conclusions Axe 2 et 4

- L'axe 4: Produits « a », « b » et « c » ayant une présence moyenne en calcium, fer et fibres et une faible présence en vitamines A et C s'opposent aux produits « e » et « d » ayant une forte présence en protéines, sel, sodium et peu de matières grasses.
- L'axe 2: Produits « a », « b » et « c » ayant une présence moyenne en fibres, fer et calcium s'opposent aux produits « d » et « e » ayant une forte présence en sucre, glucide, sel, et sodium.

# Exploration multivariable des données

## ❑ Recettes saines et équilibrées

Valeurs moyennes:  
Protéines, Matières  
grasses, Glucides

Valeurs moyennes:  
Vitamines A, C  
Fer, Calcium, Fibres

Faibles valeurs:  
Sucre, Glucides  
Ne pas cumuler les 2

Faibles valeurs:  
Sel, Sodium  
Ne pas cumuler

Faibles valeurs:  
Matières grasses  
Cholestérol  
Ne pas cumuler les 2

Ne pas mélanger les  
Macronutriments et le  
sel

# Conclusions et perspectives

KNN et ACP: Méthodes efficaces pour l'analyse et le traitement des données

Enrichir la BDD:  
Vitamines D, E, B...  
Minéraux: Zn, Mg, Ph...

Améliorer la qualité des données:

Bloquer les valeurs (-)

Fixer les limites sup

Garder les variables utiles pour les études nutritionnelles