Ministère de l'enseignement supérieur et de la recherche scientifique
Université de JIJEL
Faculté des Sciences Exactes et Informatique
Département d'Informatique

Analyse de Données

M1 SIAD & ILM

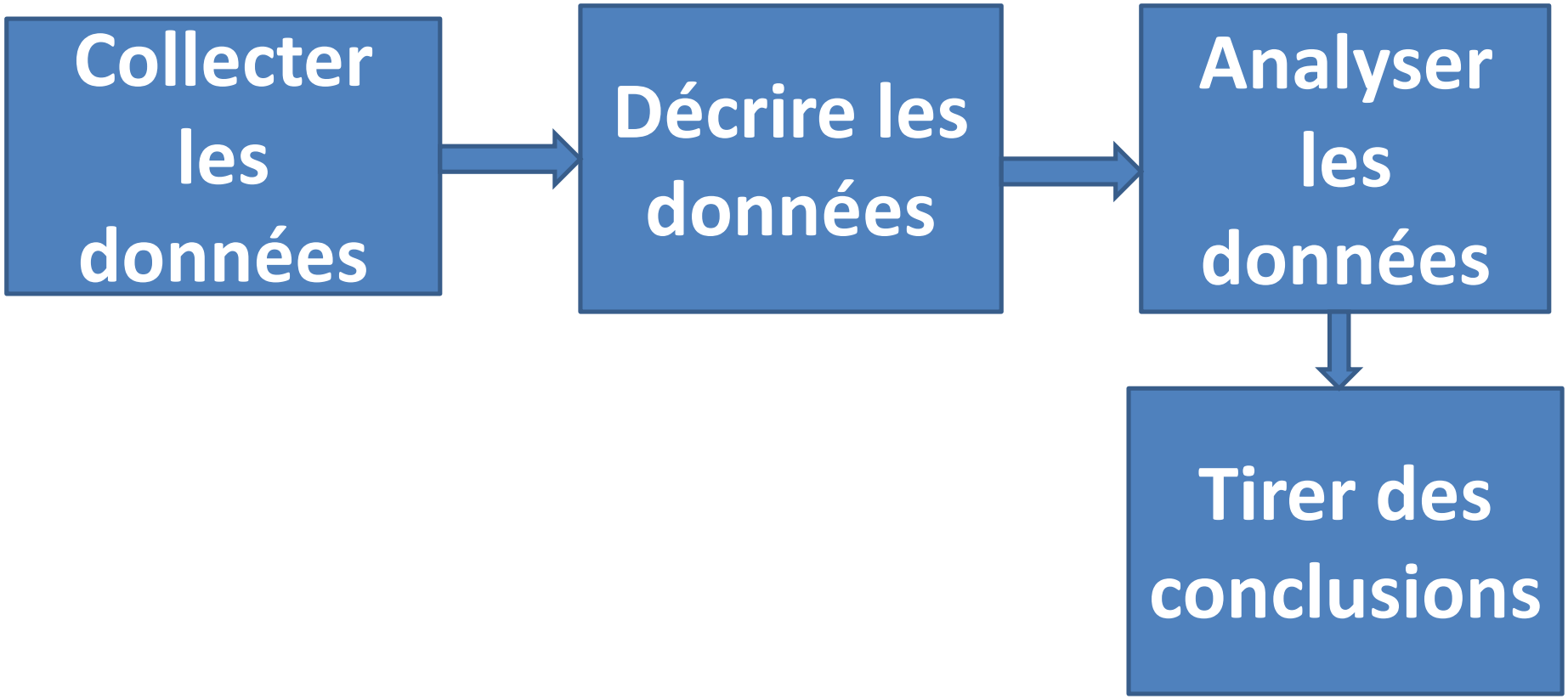
1. Statistique unidimensionnelle

C'est quoi la statistique

On appelle statistique l'ensemble de méthodes scientifiques permettant de collecter, décrire et analyser des données observées.

Ces observations consistent généralement en la mesure d'une ou plusieurs caractéristiques communes sur un ensemble de personnes ou d'objets équivalents

C'est quoi la statistique



Les deux types d'études statistiques

- La statistique **descriptive** ou statistique déductive

- La statistique **inductive** ou **inférentielle**

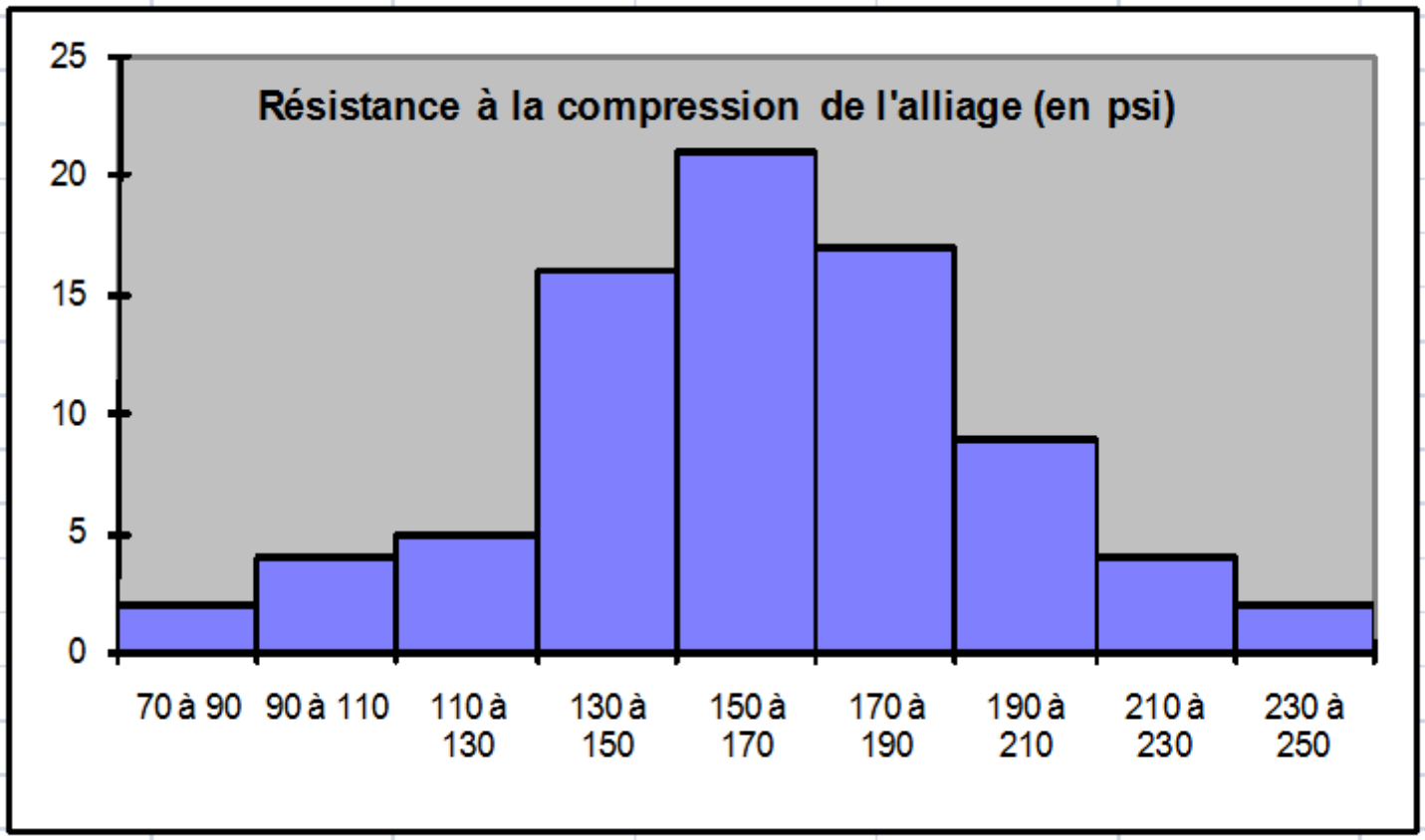
La statistique descriptive

- La **statistique descriptive** (ou statistique déductive) s'occupe de la description des données: tableau, graphique, pourcentage, ...

La statistique descriptive traite des propriétés des populations plus que des individus particuliers de ces populations.

La statistique descriptive

- Exemple:



La statistique inférentielle

- La ***statistique*** **inférentielle** (ou ***inductive***) s'occupe de tirer des conclusions générales à propos d'une population à partir d'expériences et de faire des prévisions.

Les deux types d'études statistiques

La statistique **descriptive** est utilisée pour résumer et présenter des données, tandis que la statistique **inférentielle** est utilisée pour prendre des décisions et faire des prévisions sur la base de données .

Population et individus

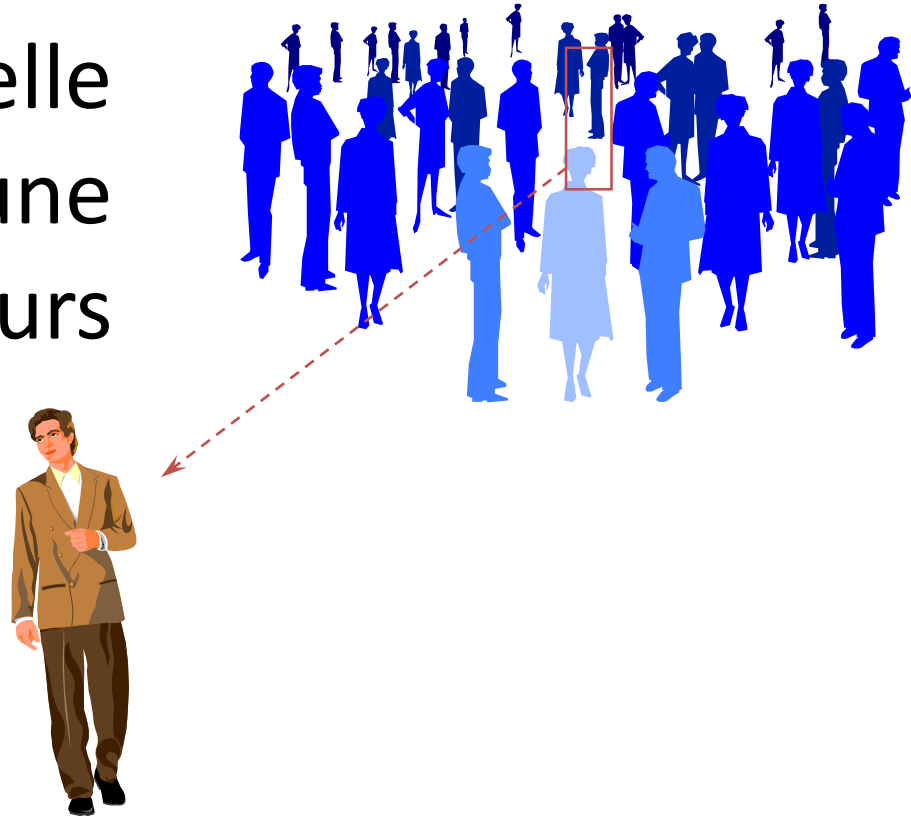
Population : ensemble des individus (ou unités statistiques) pour lequel on considère une ou plusieurs caractéristiques

- **Taille de la population:** le nombre d'individus constituant la population.

Notation : ***N***

Population et individus

Individu ou **unité statistique**: une unité distincte chez laquelle on peut observer une ou plusieurs caractéristiques données.



Population et individus

Exemple 1 : paramètre étudié : note d'un étudiant dans un groupe de TD.
un individu = un étudiant

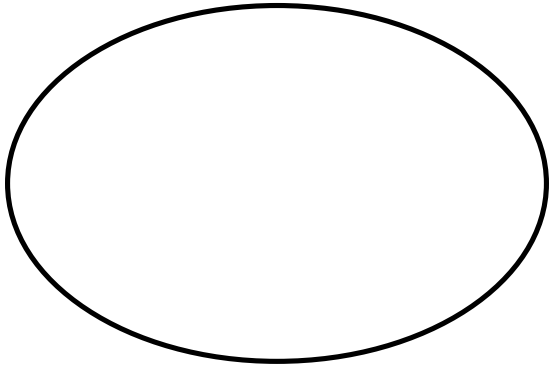
Exemple 2 : paramètre étudié : note moyenne de chaque groupe de TD d'étudiant inscrit dans une licence.
un individu = un groupe de TD

Échantillon

Dans la plupart des cas, il est difficile d'obtenir l'information à partir de la population dans son ensemble (ce serait alors un recensement). On se restreint à un sous-ensemble, l'échantillon pour tirer des conclusions sur la population.

Échantillon

Sous-groupe d'une population donnée.
(Une observation par individu)



Taille de l'échantillon : le nombre d'observations dans l'échantillon. Notation : ***n***

Caractères statistiques

Caractère ou **Variable** = propriété observable des individus qui prend différents états appelés **modalités**.

Plusieurs catégories de caractères => **méthodes statistiques différentes**.

- Les modalités d'une variable qualitative sont les différentes valeurs que peut prendre celle-ci. Par exemple. Les modalités de la variable "sexe" sont : féminin, masculin

Caractères statistiques

Caractères qualitatifs

Ne résultent ni d'une mesure par un instrument ni d'un comptage.

✓ **nominale** : modalités exprimables par des noms et non hiérarchisées. (dichotomique = 2 modalités).

Ex: Couleur des yeux, nationalité, présence/absence d 'une maladie

✓ **ordinale** : traduit le degré d'un état sans que ce degré ne puisse être défini par un nombre. Modalités hiérarchisées.

Ex : qualification professionnelle (travail d'un potier) 'non qualifié', 'semi - qualifié', 'qualifié '

Caractères statistiques

Caractères quantitatifs

= mesurables ; résultent d'une mesure ou d'un comptage.

✓ **discret** : il peut prendre seulement certaines valeurs = résulte d'un comptage.

Ex : Nombre d'objets dans un dépôt,

✓ **continu** : peut prendre n'importe quelle valeur dans un intervalle donné.

Ex: Le poids, l'âge,

Indicateurs statistiques

1. Indicateurs de position

- **La moyenne arithmétique**

Soit un échantillon de *n* valeurs observées *x*1, *x*2, ….,*xi*,….,*xn*

Données non groupées $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Données groupées pour caractère discret $\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i$

Indicateurs statistiques

1. Indicateurs de position

- **La moyenne arithmétique**

- Facile à calculer
- La somme des écarts à la moyenne est nulle:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

- Fortement influencée par les valeurs extrêmes
- Représente mal une population hétérogène (polymodale)

Indicateurs statistiques

2. Indicateurs de dispersion

- **La variance**

Soit un échantillon de *n* valeurs observées *x*1, *x*2, ….,*xi*,….,*xn*

Données non groupées $$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2$$

Données groupées discrètes $$\sigma^2 = \frac{1}{n}\sum_{i=1}^{k}n_i(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{k}n_i x_i^2 - \bar{x}^2$$

Indicateurs statistiques

2. Indicateurs de dispersion

- **L'écart-type**

C'est la racine carrée de la variance: $\sigma = \sqrt{\sigma^2}$

(en anglais « standard deviation », « s.d. »)

Indicateurs statistiques

Exemple

Variance d'une série statistique

X : Notes à un devoir dans une classe de 10 élèves

07, 08, 10, 11, 11, 13, 13, 14, 15, 18

$\overline{x} = 12$

03, 04, 04, 07, 07, 17, 19, 19, 20, 20

$\overline{x} = 12$

11, 11, 12, 12, 12, 12, 12, 12, 13, 13

$\overline{x} = 12$

Indicateurs statistiques

Exemple

Variance d'une série statistique

$$\left\{ \begin{array}{c} X : \text{Notes à un devoir dans une classe de 10 élèves} \\ 07, 08, 10, 11, 11, 13, 13, 14, 15, 18 \\ \overline{x} = 12 \qquad \overline{X^2} = 153{,}8 \end{array} \right\}$$

$$\text{Var}(X) = \frac{1}{n}\sum_{i=1}^{i=n}(X_i - \overline{x})^2 = \overline{X^2} - \overline{x}^2$$

$$\text{Var}(X) = 9{,}8$$

Limite de la statistique descriptive

En statistique descriptive, on s'intéresse aux caractéristiques de tendance centrale, de dispersion, de forme, les liaisons entre deux variables.

Cependant, le statisticien peut se trouver devant un tableau contenant plusieurs variables et individus.

Dans ce tableau, il cherche à dégager par exemple la tendance globale des données.

Limite de la statistique descriptive

Par exemple:

- Les variables qui sont liées, les individus qui se ressemblent.
- Regrouper les individus suivant leur proximité au vue des variables.

Limite de la statistique descriptive

Dans ces situations, la statistique descriptive reste limitée.

On passe donc aux méthodes *d'analyse des données multidimensionnelles*:

c'est la « grande statistique descriptive ».

2. Analyse de données multidimensionnelles

Analyses de données multidimensionnelles

Définition : statistiques descriptives multidimensionnelles (beaucoup de dimensions)

Objectif : extraire l'information principale d'un tableau à double entrée, y compris quand il est très grand

Méthode : consentir une perte d'information pour gagner en efficacité

Domaines d’application

- ✓Démographie , Économie, Études de marché
- ✓Assurances, Agriculture, Finance,
- ✓ Transport , Communications – etc.

Principe général d'ADD

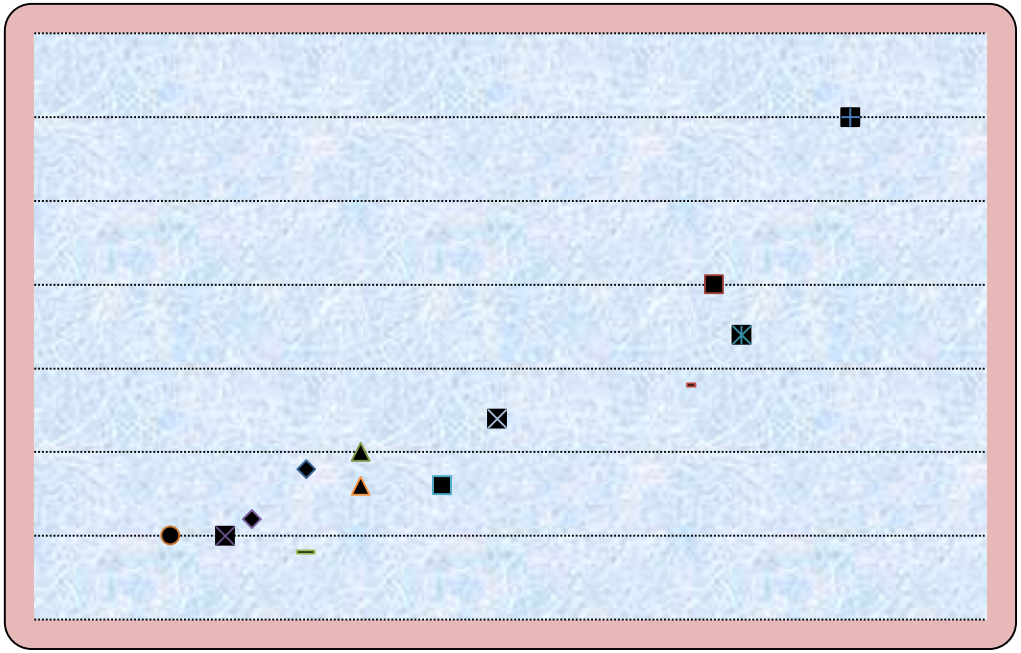
- Si n individus et seulement 2 variables X et Y, il est facile de représenter l'ensemble des données sur un graphique plan : chaque individu i est un point de coordonnées X_i et Y_j → **nuage**
- L'allure du nuage renseigne sur l'intensité et la nature de la relation entre X et Y.
- **Si plus de 3 variables**, il faut **trouver de « bonnes » approximations du nuage** pour l'appréhender dans sa globalité.

Principe général d'ADD

Exemple:

On dispose de deux variables:

revenu et **consommation** sur 13 ménages.



MENAGE	REVENU	CONSOMMATION
1	10	9
2	25	20
3	12	10
4	7	5
5	26	17
6	5	5
7	30	30
8	24	14
9	10	4
10	8	6
11	15	8
12	12	8
13	17	12

Principe général d'ADD

- Si nous avons trois variables : Revenu, consommation et nombre personnes dans le ménage.

  On peut faire un graphique à **trois dimensions**.

- Si nous avons plusieurs variables (par exemple plus de 15) sur plusieurs individus alors on ne peut plus faire des graphique à 15 dimensions!

D'où l'utilisation des **méthodes de projection**.

Principe général de l'ADD

L'Analyse des données (ADD): l'ensemble de méthodes **descriptives ayant pour objectif de résumer et visualiser l'information contenue dans un grand tableau de données**

Principe général de l'ADD

«L'analyse des données est un ensemble de techniques pour découvrir la structure, éventuellement compliquée, d'un tableau de nombres à plusieurs dimensions et de traduire par une structure plus simple et qui la résume au mieux. Cette structure peut le plus souvent, être représentée graphiquement'» (J-P. Fénelon)

Les principaux objectifs de l'ADD

-Répondre aux problèmes posés par des tableaux de grandes dimensions

-Résumer les informations contenues dans un grand tableau sous forme d'une matrice

-Organiser et visualiser les informations

Méthodes de l'ADD

Les principales méthodes de l'ADD se séparent en deux groupes:

- Les méthodes de classification,
- Les méthodes factorielles

Les méthodes de classification

- Les méthodes de classification visant à réduire la taille de l'ensemble des individus en formant des groupes homogènes d'individus ou de variables.
- Ces groupes on les appelle aussi des classes, ou familles, ou segments, ou clusters.
- La classification est appelée aussi Segmentation ou Clustering.

Les méthodes factorielles

Les méthodes factorielles cherchent à réduire le nombre de variables en les résumant par un petit nombre de composantes synthétiques en utilisant essentiellement des outils de l'algèbre linéaire et donnant lieu à des représentations graphiques dans lesquelles les objets à décrire se transforment en des points sur des axes et des plans.

Les principales techniques factorielles

- L'analyse en composantes principales (hotelling, 1933) qui analyse un ensemble de données (observations) faites sur un ensemble de variables **quantitatives** (numériques)

Les principales techniques factorielles

- L'analyse des correspondances (Benzekri, 1964) qui est une technique de base pour analyser des tables de contingence qui peut être utilisé pour des variables qualitatives ou quantitatives positives de nature très divers.

Les principales techniques factorielles

- L’analyse canonique (Hotelling) qui contient à la régression multiple et l’analyse discriminante comme des cas particulier.

Les méthodes factorielles

- Si on travaille avec un tableau de variables **numériques,** on utilisera **l'analyse en composantes principales,**
- Si on travaille avec des variables **qualitatives**, on utilisera **l'analyse des correspondances.**
- Les **liens** entre **deux groupes de variables** peuvent être traites par **l'analyse canonique**.

Analyse factorielle

- Etude de la position d'un **nuage de points** dans l'espace et **description de sa forme**
- Pour mieux voir :
  - **se placer au milieu du nuage**, c'est-à-dire déplacer l'origine au centre de gravité (= individu fictif « moyen »)
  - **regarder dans les directions d'allongement principal**, c'est-à-dire changer d'axes
- Techniquement, **changer de repère** (→ diagonaliser une matrice)

De l'image à la réalité: les outils d'interprétation.

Ce que nous observons sur les photos peuvent être trompeuse.

⇒ Il nous faut des outils d'aide à interprétation.

Les outils:

- **Les Cosinus carré**: (CO^2), qualité de la représentation.
- Le contribution **(CTR)**: permet de mesurer la part des variables ou individus dans la formation des axes.
- **Disto**: distance d'un individu à l'individu moyen.