

Sciences des données

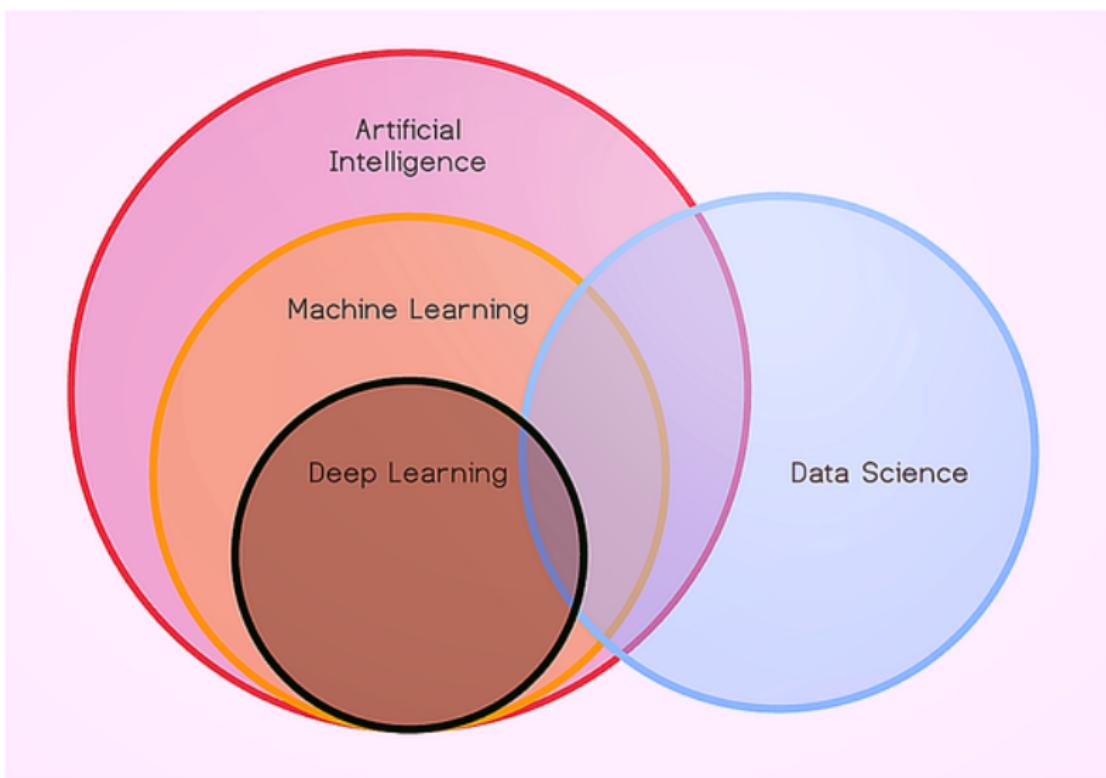
Un voyage initiatique

Cécile CAPPONI, Hachem KADRI

LIS, Aix-Marseille Université, CNRS
Equipe QARMA



M1 Informatique



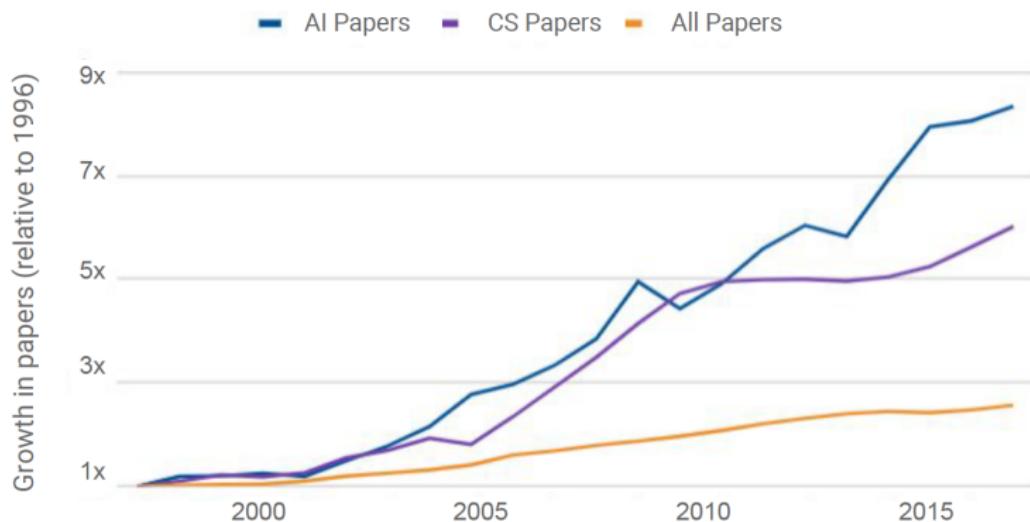
Quelques graphiques de l'« Artificial Intelligence (AI) Index 2018 Report »

<https://aiindex.org/>

Articles publiés

Growth of annually published papers by topic (1996–2017)

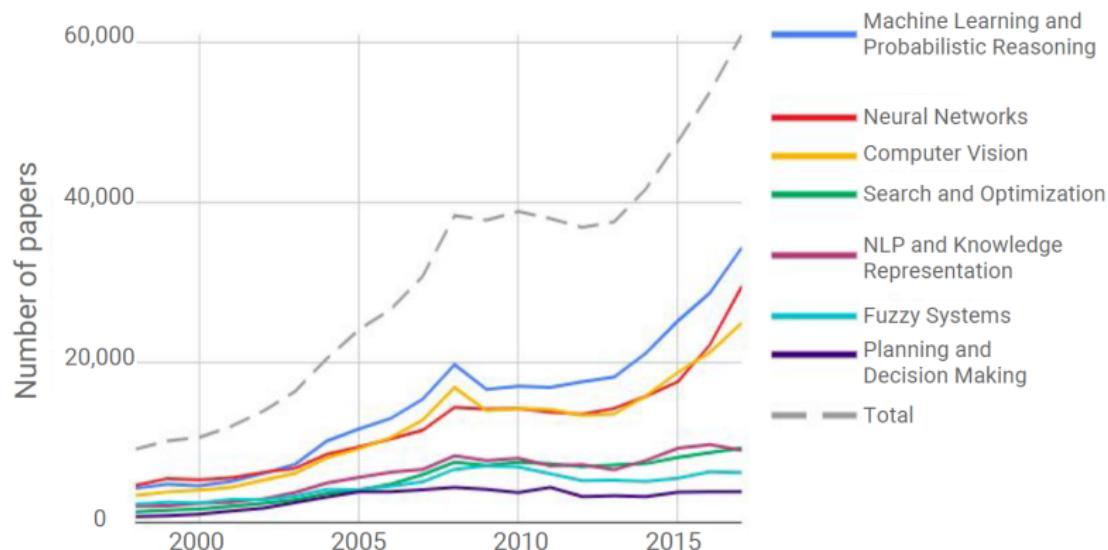
Source: Scopus



- Nombre d'articles en IA sur Scopus a augmenté de 8x depuis 1996
- Nombre d'articles en Info. a augmenté de 6x pendant la même période

Articles publiés : articles en IA par sous-catégorie

Number of AI papers on Scopus by subcategory (1998–2017)
Source: Elsevier

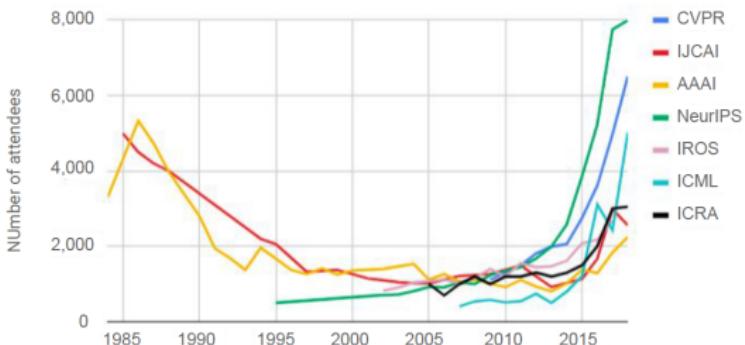


- 56% des articles sur ML, comparé à 28% en 2010.

Participation : conférences en IA

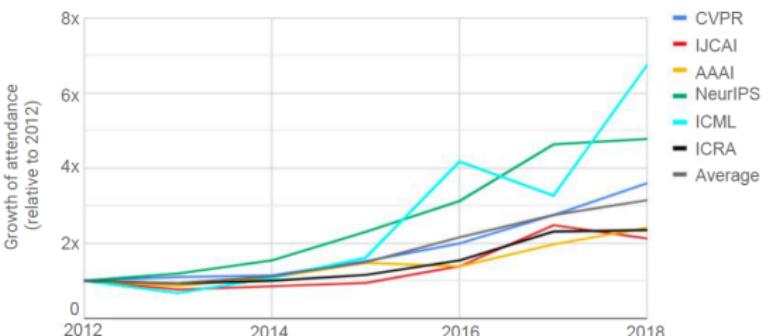
Attendance at large conferences (1984–2018)

Source: Conference provided data



Growth of large conference attendance (2012–2018)

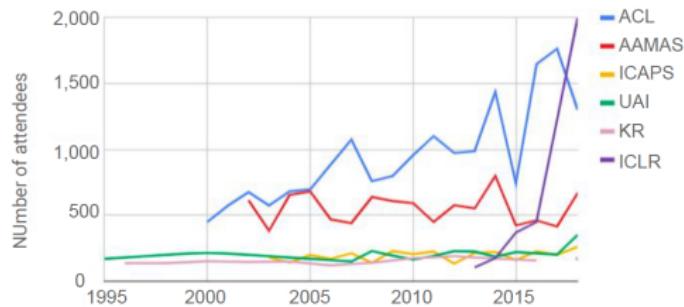
Source: Conference provided data



Participation : conférences en IA

Attendance at small conferences (1995–2018)

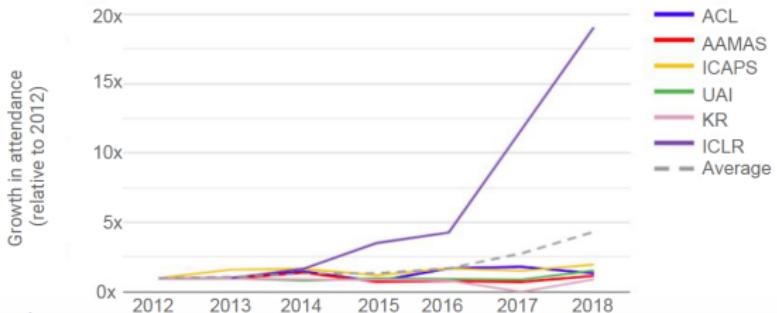
Source: Conference provided data



Note: 2018 was the first year that KR held a workshop. For consistency, workshop attendees are not included in KR's attendance count in the visual above.

Growth of small conference attendance (2012–2018)

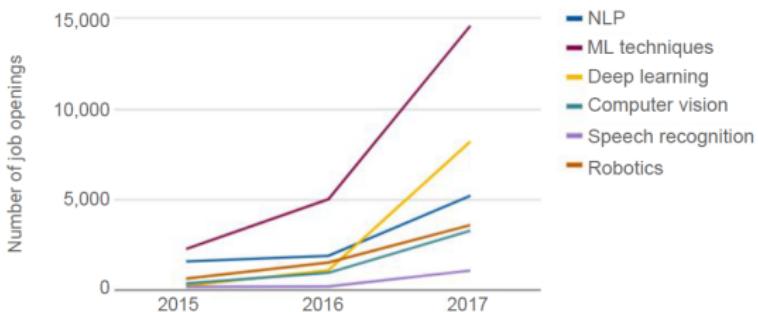
Source: Conference provided data



Marché du travail : offres d'emploi en IA

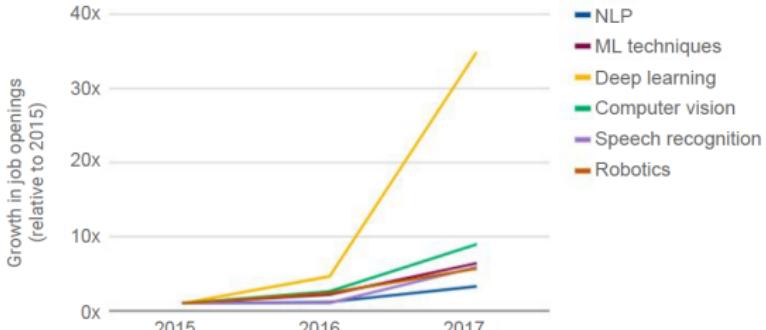
Job openings by AI skills required (2015 – 2017)

Source: Monster.com



Growth of job openings by AI skills required (2015 – 2017)

Source: Monster.com



Outline

1 | Introduction

- ## ■ Sciences des données, késako ?

2 Représentation et visualisation

- Représentations numériques des données
 - Des statistiques descriptives aux modes de visualisation
 - Quelques modes classiques de visualisation
 - Représentation et visualisation des données vectorielles

3 Analyse en composantes principales (ACP – PCA)

- Introduction
 - ACP : principes
 - Pour aller plus loin

4 Et après

Outline

1 | Introduction

- ## ■ Sciences des données, késako ?

2 Représentation et visualisation

- Représentations numériques des données
 - Des statistiques descriptives aux modes de visualisation
 - Quelques modes classiques de visualisation
 - Représentation et visualisation des données vectorielles

3 Analyse en composantes principales (ACP – PCA)

- Introduction
 - ACP : principes
 - Pour aller plus loin

4 Et après

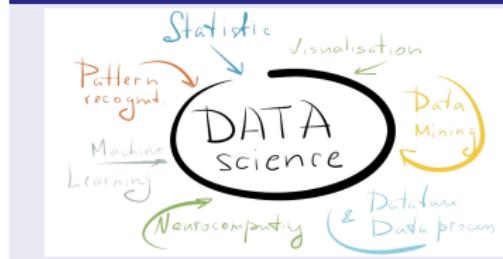
Une science récente : explosion des données numériques

De quelles données parlons-nous ?

Données d'observation, issues d'individus ou de phénomènes, anciennes ou actuelles, brutes ou travaillées, **supposées marquées par des régularités**.

- Réseaux sociaux, personnelles
 - Objets connectés (bracelet EMG, caméra surveillance, etc.)
 - Sciences (astrophysique, biologie, SHS, santé, météo, etc.)
 - Commerce (transactions, production, client, bitcoin, etc.), **Web**
 - Robots (ouvrier, drone, transports, etc.)

A l'intersection de plusieurs champs disciplinaires



Cette U.E.

- 9h cours, 18h TD/TP
 - Seulement quelques aspects
 - Cas pratiques

Qu'est-ce qu'une donnée ?



Que voyez-vous ?

- Un chat, un félin, un animal
 - Des poils noir, des yeux jaunes
 - Un malheur à venir

Qu'est-ce qu'une donnée ?



Que voit l'ordinateur ?

Qu'est-ce qu'une donnée ?



Que voit l'ordinateur ?

Qualification des données

- Qualitatives versus quantitatives
 - Catégorielles, discrètes, continues, séquentielles, vectorielles, etc.
 - Positionnées dans des taxonomies, ou pas

Qu'est-ce qu'un ensemble de données ?

Entrepôt de données : stockage



Obtention de jeu de données

- Extraction à partir d'un entrepôt (ou web)
 - Intégration des données
 - Acquisition dédiée (protocole)

Nécessité de nettoyage des données

Que faire avec des données ?

Extraction de connaissances à partir d'un jeu de données

- Les structurer, les stocker (big data, cloud...)
 - En extraire des connaissances
 - en extraire des tendances
 - reconnaître des concepts
 - les analyser, les comprendre

Aspects sociétaux : explosion des données numériques

- Droit des données, accessibilité
 - Sécurité, confidentialité
 - Aspects éthiques
 - Problématique des biais

Objectifs de ce cours (27h)

Ce que nous ne traiterons pas : big data



Extract Transform Load



Ce que nous aborderons : traitement d'un jeu de données

- Analyse préalable des données
- Visualisation des données
- Classification
- Régression
- Regroupement



Outline

1 Introduction

- Sciences des données, késako ?

2 Représentation et visualisation

- Représentaions numériques des données
- Des statistiques descriptives aux modes de visualisation
- Quelques modes classiques de visualisation
- Représentation et visualisation des données vectorielles

3 Analyse en composantes principales (ACP – PCA)

- Introduction
- ACP : principes
- Pour aller plus loin

4 Et après

Outline

1 Introduction

- Sciences des données, késako ?

2 Représentation et visualisation

- Représentaions numériques des données
- Des statistiques descriptives aux modes de visualisation
- Quelques modes classiques de visualisation
- Représentation et visualisation des données vectorielles

3 Analyse en composantes principales (ACP – PCA)

- Introduction
- ACP : principes
- Pour aller plus loin

4 Et après

Représentations presque brutes

La feuille excel pour représenter un jeu de données (e.g. open data)

- Le jeu de données Titanic
 - Colonnes typées (booléen, symbolique, réel,etc.)

Autres types de données pour les colonnes (ou groupes de colonnes)

- Texte = (longue) chaîne de caractères
 - Image = tableau de pixels à trois couleurs
 - Signal = amplitude selon le temps
 - Graphe = noeuds et arcs, matrice d'adjacence

Ensemble de données (dataset)

Un tableau de données

- Cas d'une représentation vectorielle des données : $S = \{\mathbf{x}_i\}_{i=1}^n, \mathbf{x}_i \in \mathbb{R}^d$
 - Distributions de probabilité pour chaque colonne, distributions jointes $P(A_1), P(A_2), \dots, P(A_d), P(A_1, A_2, \dots, A_d)$ où les A_j sont des variables aléatoires

Du titanic aux réseaux d'interactions biologiques

Dataset Titanic – Hétérogénéité des colonnes

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...)	female	38.0	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

- Objectif : expliquer colonne survie par les autres colonnes
 - Qualité du jeu de données
 - Colonnes Sex ou Ticket
 - Nécessité de plus d'exemples
 - Données manquantes
 - Cas d'un jeu de données étiquetées $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n, \mathbf{x}_i \in \mathcal{X}, y_i \in \mathcal{Y}$

Du titanic aux réseaux d'interactions biologiques

Texte à traduire = une donnée



Traductions

Désactiver la traduction instantanée



[Français](#) [Italien](#) [Zoulou](#) [Déterminer la langue](#)

Traduire

Uma uhielo lwethu lwezemfundu ephakeme lungakwazi ukuceqesha nokuheha abachwepheshe abanele, onjiniyela, odokotela ukuba bathuthukese ukuthuthukiswa kwama laboratories nezinkampani zaseFrance, ukugcina amatalentza ayo ekuceqesheni izizukulwane ezintsha, negeke siphumelele, hhayi ukulhanganisa lokhu ukunqotshwa kwezwelisha. Yingakha sizophindaphandla kabilu inani labafundi abaqeqeshwe eukulhakanipheni okufakelwayo, kusukela e-bachelors degree kuya ku-Ph.D, ngokusebenzisa ukuceqeshwa okutushane kokugeqeshwa, futhi kuzohlinzeka ngemali ehambisana nalesibili.

Si notre système d'enseignement supérieur est capable de former et d'attirer des techniciens qualifiés, des ingénieurs, des médecins pour améliorer le développement des laboratoires et des entreprises françaises, de garder leurs talents dans la formation des nouvelles générations, l'échouera. ne pas combiner cette conquête du nouveau monde. C'est pourquoi nous allons doubler le nombre d'étudiants formés en ingénierie artistique, du baccalauréat au doctorat, en utilisant une formation de courte durée, et vous fournira un financement de seconde main.

1

Suggérer une modification

Chaque exemple du dataset (ligne) est un document

Si notre système d'enseignement supérieur ne sait pas former et attirer assez de techniciens, d'ingénieurs, de docteurs pour alimenter le développement des laboratoires et des entreprises en France, pour garder aussi ses talents dans la formation des nouvelles générations , nous ne parviendrons pas à consolider cette conquête d'un horizon nouveau. C'est pourquoi nous doublerons le nombre d'étudiants formés à l'intelligence artificielle, depuis la licence jusqu'au doctorat en passant par les formations professionnelles courtes, et prévoirons les financements qui correspondent à ce doublement.

Du titanic aux réseaux d'interactions biologiques

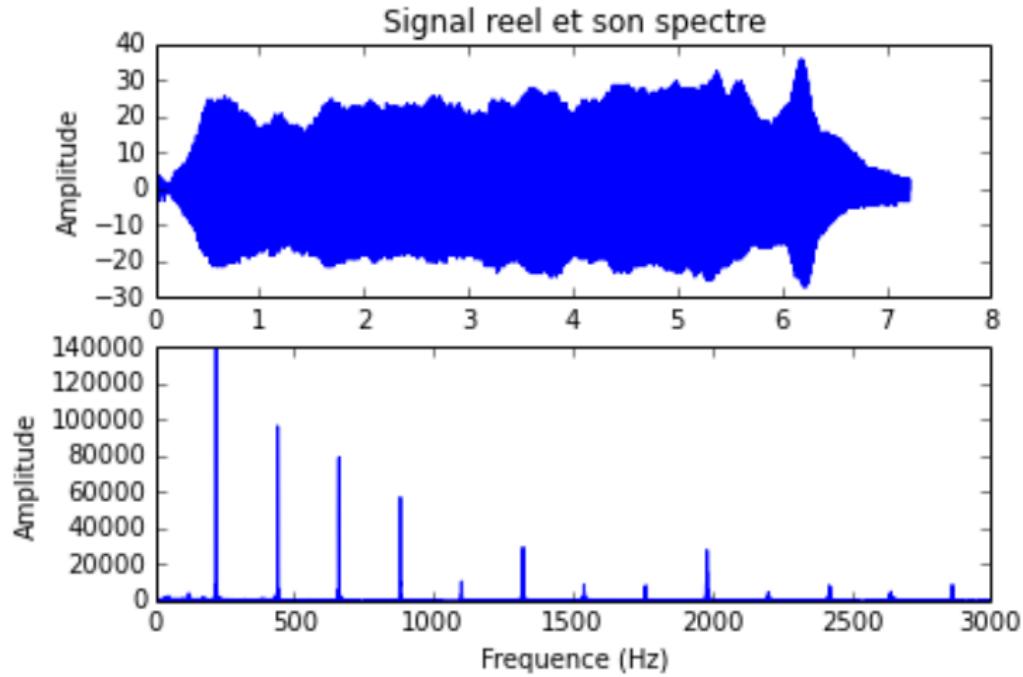
Image = une donnée

A close-up photograph of a vibrant pink peony flower with many petals, set against a dark green, slightly blurred background.

Dataset = tableau d'images (ex. scanners cérébraux)

Du titanic aux réseaux d'interactions biologiques

Signal et son spectre = une donnée

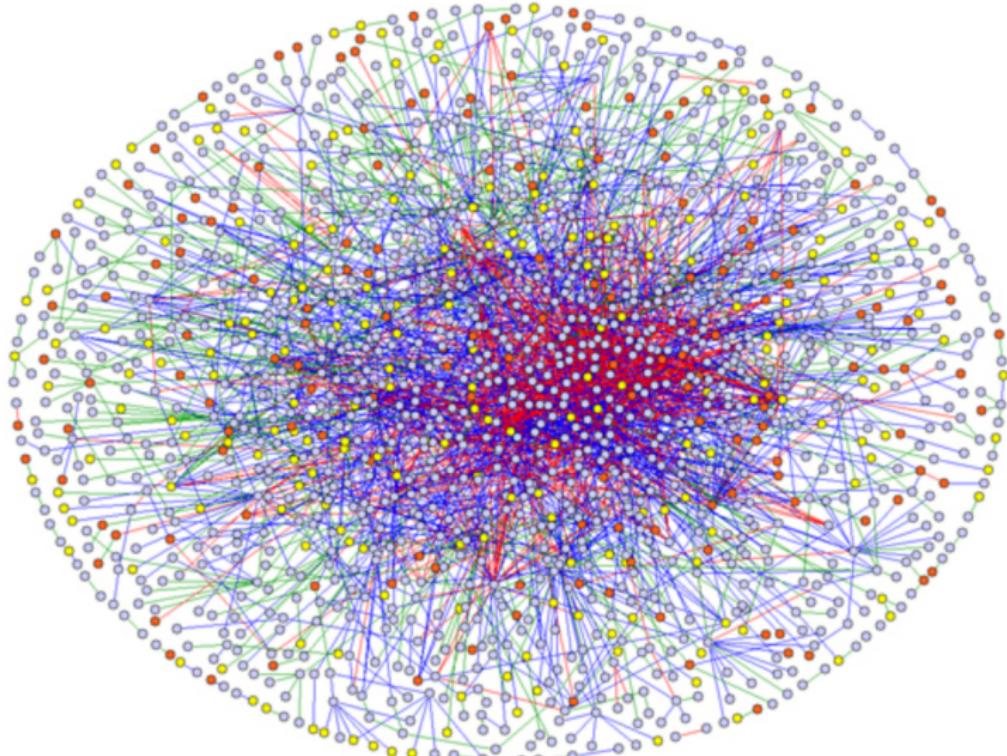


Dataset = tableau de signaux (ex. playlist musicale)
(source : tangenteX.com)

Du titanic aux réseaux d'interactions biologiques

une donnée = une interaction entre deux protéines

Dataset = Un graphe d'interactions entre protéines



Un Graal pour traiter les données : la représentation vectorielle

Fondements mathématiques

Bases éprouvées, riches et solides

- Distances et similarités entre deux vecteurs
- Transformations linéaires ou non, dérivées

$$\begin{matrix} x_{absolu(hom)} \\ y_{absolu(hom)} \\ z_{absolu(hom)} \\ w_{absolu(hom)} \end{matrix} = \begin{matrix} \text{rotation} & \text{scale} & \text{translation} \end{matrix} \times \begin{matrix} M_{1,1} & M_{1,2} & M_{1,3} & M_{1,4} \\ M_{2,1} & M_{2,2} & M_{2,3} & M_{2,4} \\ M_{3,1} & M_{3,2} & M_{3,3} & M_{3,4} \\ M_{4,1} & M_{4,2} & M_{4,3} & M_{4,4} \end{matrix} \begin{matrix} x_{local} \\ y_{local} \\ z_{local} \\ 1 \end{matrix}$$

- Algèbre linéaire, statistiques, topologie
- Propriétés algorithmiques (parcimonie, arithmétique, etc.)

Alternatives algorithmiques

Séquences, sacs, arbres, automates, graphes

Outline

1 Introduction

- Sciences des données, késako ?

2 Représentation et visualisation

- Représentaions numériques des données
- Des statistiques descriptives aux modes de visualisation
- Quelques modes classiques de visualisation
- Représentation et visualisation des données vectorielles

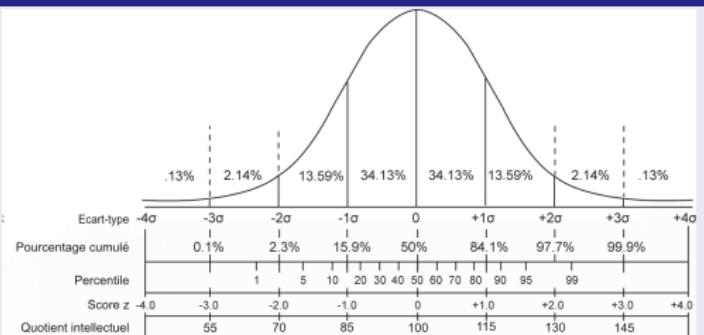
3 Analyse en composantes principales (ACP – PCA)

- Introduction
- ACP : principes
- Pour aller plus loin

4 Et après

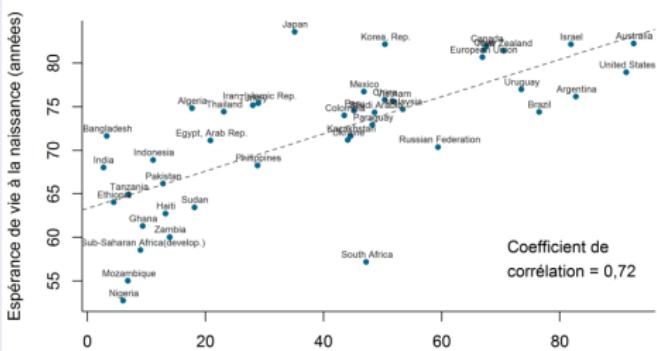
Exemples d'échantillons statistiques (jeu de données)

Echantillon à une seule variable

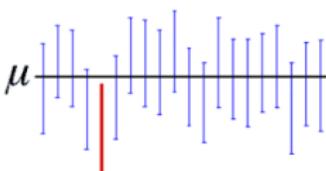


Echantillon à deux variables : variable facteur et variable à expliquer

Espérance de vie et consommation de viande en 2014 par pays



Statistiques descriptives simples (1)



Estimation de propriétés statistiques simples

- Propriétés fondées sur des distributions de probabilités
- Nous ne disposons que d'un échantillon : distribution inconnue
- **Estimateurs de ces propriétés**, notion de biais
- Utilité pour avoir un aperçu statistique de l'échantillon et des variables

Sur échantillon avec une seule variable x

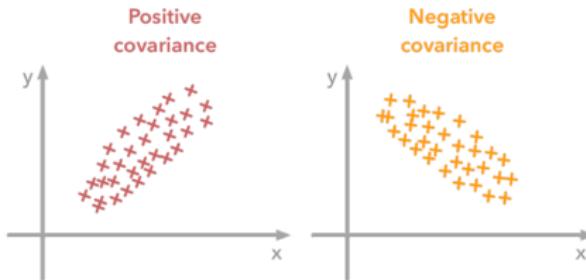
Echantillon de variables numériques $S = \{x_i\}_{i=1}^n, x_i \in \mathbb{R}$ (série statistique)

- Moyenne de x sur S : $\mu_S(x) = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
- Variance de x sur S : $V_S(x) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ (indicateur de dispersion).
- Ecart-type de x sur S : $\sigma_S(x) = \sqrt{V_S(x)}$

Statistiques descriptives simples (2)

Sur échantillon avec deux variables $x, y : S = \{(x_i, y_i)\}_{i=1}^n, (x_i, y_i) \in \mathbb{R}^2$

- Covariance des variables dans S pour quantifier les écarts conjoints de x et y par rapport à leurs moyennes respectives :
 $\text{cov}_S(x, y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$
- Coefficient de corrélation de S pour mesurer l'intensité d'une relation entre x et y , estimateur de Bravais-Pearson : $r_p(x, y) = \frac{\text{cov}_S(x, y)}{\sigma_x \sigma_y}$
- Fonction expliquant y par x (facteur) dans S : $y = f(x)$ (en régression linéaire : $y = ax + b$)
- Coefficient de détermination de S par f :
 $R^2 = 1 - \frac{\sum_{i=1}^n (y_i - f(x_i))^2}{n V_S(y)} = r_p^2(f(x), y)$



Statistiques descriptives simples (3)

Avec $m > 2$ variables : Statistiques multi-variées

$$\mathcal{S} = \{(x_{i,1}, x_{i,2}, \dots, x_{i,d})\}_{i=1}^n, \text{ avec } x_{i,j} \in \mathbb{R}$$

Vin	Bel.	N.L.	RFA	Ita.	UK	Sui.	USA	Can.
CHMP	7069	3786	12578	8037	13556	9664	10386	206
MOS1	2436	586	2006	30	1217	471	997	51
MOS2	3066	290	10439	1413	7214	112	3788	330
ALSA	2422	1999	17183	57	1127	600	408	241
GIRO	22986	22183	21023	56	30025	6544	13114	3447
BOJO	17465	19840	72977	2364	39919	17327	17487	2346
BORG	3784	2339	4828	98	7885	3191	11791	1188

$$n = 7, d = 8, x_{2,5} = 1217$$

La matrice de **covariance** mesure, pour chaque couple de variables différentes, leur propension à varier ensemble dans le jeu de données.
(<http://www.info.univ-angers.fr/~gh/Datasets/vins.htm>)

La matrice de covariance C de $S = \{\mathbf{x}_i\}$

Définition

- Matrice $X = (x_{i,j})$ la description du jeu de données S , de taille $n \times d$
- $C = X^T X$ est sa matrice de covariance, de taille $d \times d$: variance de chaque variable sur la diagonale, covariances des variables 2 à 2 ailleurs
- covariance entre variables normalisées : a et $b = 0$ si a et b varient indépendamment, 1 (ou -1) si variables proportionnelles (colinéaires)
- $X_{j,i}^T = X_{i,j}$, C est symétrique, donc diagonalisable
- C inversible sauf si deux colonnes sont colinéaires

$$X = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,d} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,d} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,d} \end{pmatrix} \quad X^T = \begin{pmatrix} x_{1,1} & x_{2,1} & \cdots & x_{n,1} \\ x_{1,2} & x_{2,2} & \cdots & x_{n,2} \\ \cdots & \cdots & \cdots & \cdots \\ x_{1,d} & x_{2,d} & \cdots & x_{n,d} \end{pmatrix}$$

$$C = X^T X = \begin{pmatrix} \sum_{i=1}^n (x_{i,1})^2 & \sum_{i=1}^n x_{i,1} x_{i,2} & \cdots & \sum_{i=1}^n x_{i,1} x_{i,d} \\ \sum_{i=1}^n x_{i,2} x_{i,1} & \sum_{i=1}^n (x_{i,2})^2 & \cdots & \sum_{i=1}^n x_{i,2} x_{i,d} \\ \cdots & \cdots & \cdots & \cdots \\ \sum_{i=1}^n x_{i,d} x_{i,1} & \sum_{i=1}^n x_{i,d} x_{i,2} & \cdots & \sum_{i=1}^n (x_{i,d})^2 \end{pmatrix}$$

Disgression : ne pas confondre corrélation et causalité

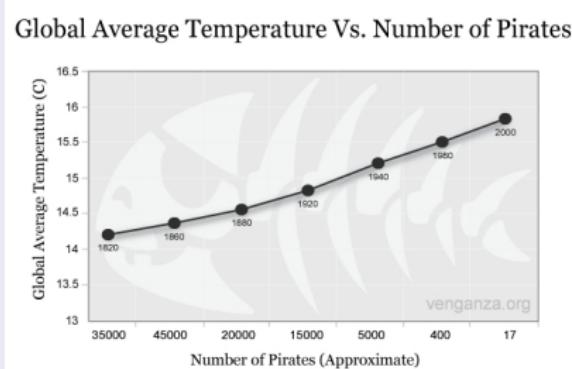
Cum hoc ergo propter hoc

■ Si A et B sont corrélés

- 1 A est la cause de B ?
- 2 B est la cause de A ?
- 3 Autorenforcement du système (1 ET 2)
- 4 Il existe un troisième facteur (inconnu) étant la cause commune de A et B
- 5 Une coïncidence

Du pastafarisme aux corrélations fallacieuses

tylervigen.com/spurious-correlations



Exercice : calcul de ces stats descriptives élémentaires sur un exemple simple

<http://www.info.univ-angers.fr/~gh/Datasets/bumpus.htm>, caractéristiques physiologiques d'oiseaux échoués. Extrait :

LOT	AIL	TET	HUM	BRE
156	245	31.6	18.5	20.5
154	240	30.4	17.9	19.6
153	240	31.0	18.4	20.6
153	236	30.9	17.7	20.2
155	243	31.5	18.6	20.3
163	247	32.0	19.0	20.9
157	238	30.9	18.4	20.2

- 1 Estimer moyenne, variance et écart-type de chaque variable
- 2 Calculer la matrice de covariance : quels sont les couples de variables les plus covariantes ?
- 3 Régression $LOT = f(HUM)$ (intuitivement, graphiquement) : existe-t-il une corrélation linéaire, et si oui quelle est approximativement son équation ?

Correction, avec du python !

```
oiseaux = np.array([[156,245,31.6,18.5,20.5],  
                   [154 , 240 , 30.4 , 17.9 , 19.6],  
                   [153 , 240 , 31.0 , 18.4 , 20.6],  
                   [153 , 236 , 30.9 , 17.7 , 20.2],  
                   [155 , 243 , 31.5 , 18.6 , 20.3],  
                   [163 , 247 , 32.0 , 19.0 , 20.9],  
                   [157 , 238 , 30.9 , 18.4 , 20.2]])
```

```
namevar = ['LOT','AIL','TET','HUM','BRE']  
for ivariable in np.arange(5):  
    x = oiseaux[:,ivariable]  
    print(namevar[ivariable]+': moyenne=%1.2f, variance=%1.2f, ecarttype=%1.2f'  
          % (np.mean(x), np.var(x), np.std(x)))
```

```
LOT: moyenne=155.86, variance=10.41, ecarttype=3.23  
AIL: moyenne=241.29, variance=13.06, ecarttype=3.61  
TET: moyenne=31.19, variance=0.25, ecarttype=0.50  
HUM: moyenne=18.36, variance=0.16, ecarttype=0.40  
BRE: moyenne=20.33, variance=0.14, ecarttype=0.38
```

```
cov_oiseaux = np.dot(np.transpose(oiseaux), oiseaux)  
print(cov_oiseaux)
```

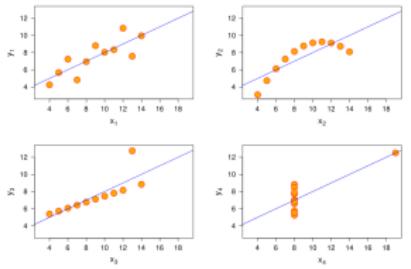
```
[[170113. 263300. 34031.7 20034.7 22183.4 ]  
 [263300. 407623. 52683.1 31013.7 34340.5 ]  
 [ 34031.7 52683.1 6809.59 4008.55 4438.85]  
 [ 20034.7 31013.7 4008.55 2360.03 2613.03]  
 [ 22183.4 34340.5 4438.85 2613.03 2893.75]]
```

Limites des analyses statistiques de base

Le quartet d'Anscombe [E. Tufte]

Fonction 1	Fonction 2	Fonction 3	Fonction 4
(x, y)	(x, y)	(x, y)	(x, y)
(10.0, 8.04)	(10.0, 9.14)	(10.0, 7.46)	(8.0, 6.58)
(8.0, 6.95)	(8.0, 8.14)	(8.0, 6.77)	(8.0, 5.76)
...
(7.0, 4.82)	(7.0, 7.26)	(7.0, 6.42)	(8.0, 7.91)
(5.0, 5.68)	(5.0, 4.74)	(5.0, 5.73)	(8.0, 6.89)

- 4 jeux de données aux mêmes propriétés statistiques simples
- données très différentes



Moyenne x et y	9.0 et 7.5
Variance x et y	10 et 3.75
Corrélation x et y	0.816
Eq. droite régression	$y = \frac{1}{2}x + 3$
Coeff de détermination	0.67

Un peu de python

```
import numpy as np

def predict(x):
    return 3 + 0.5 * x

def coeffdet(x,y):
    cfdet = 0.
    nbe = len(x)
    tsq = np.var(y)*nbe

    err = np.ndarray(nbe)
    for i in range(nbe):
        err[i] = (y[i]-predict(x[i]))**2
    cfdet = np.sum(err)
    R2 = 1-cfdet/tsq
    return R2

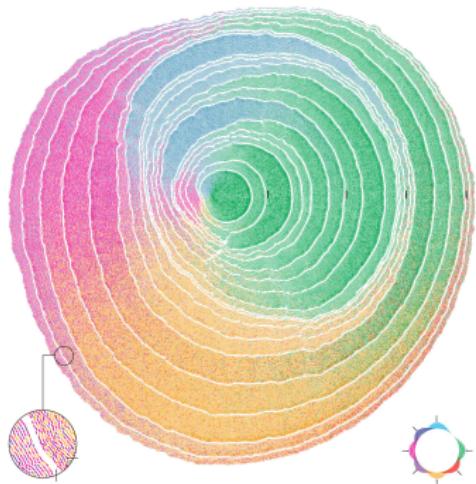
x1 = np.array([10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5])
x2 = np.array([10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5])
x3 = np.array([10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5])
x4 = np.array([8, 8, 8, 8, 8, 8, 19, 8, 8, 8])
y1 = np.array([8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68])
y2 = np.array([9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74])
y3 = np.array([7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73])
y4 = np.array([6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89])
pairs = (x1, y1), (x2, y2), (x3, y3), (x4, y4)

for x, y in pairs:
    print('meanx=%1.2f, stdx=%1.2f, meany=%1.2f, stdy=%1.2f, pearson=%1.2f, coeffdet=%1.2f '
          '(np.mean(x), np.std(x), np.mean(y), np.std(y), np.corrcoef(x, y)[0][1], coeffdet(x,y)))
```

```
meanx=9.00, stdx=3.16, meany=7.50, stdy=1.94, pearson=0.82, coeffdet=0.67
```

Les différents diagrammes de visualisation

Le SWD Challenge une visualisation raconte une histoire



Nombre et provenance des immigrés aux USA, depuis 1800 (un cercle concentrique par décennie)

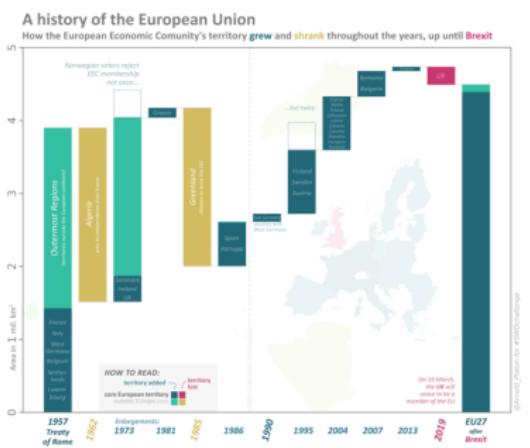
source : <http://www.storytellingwithdata.com>)

Bases = Couleurs, barres, camembert, intervalles de confiance, etc.

Difficultés : rendu correct d'une analyse, importance de la perception humaine, difficultés d'appréhension, précision, etc.

Les différents diagrammes de visualisation

Le SWD Challenge une visualisation raconte une histoire



L'Europe vue par ses territoires géographiques perdus et gagnés : évolution depuis 1957 jusqu'au Brexit

Bases = Couleurs, barres, camembert, intervalles de confiance, etc.

Difficultés : rendu correct d'une analyse, importance de la perception humaine, difficultés d'apprehension, précision, etc.

Outline

1 Introduction

- Sciences des données, késako ?

2 Représentation et visualisation

- Représentaions numériques des données
- Des statistiques descriptives aux modes de visualisation
- **Quelques modes classiques de visualisation**
- Représentation et visualisation des données vectorielles

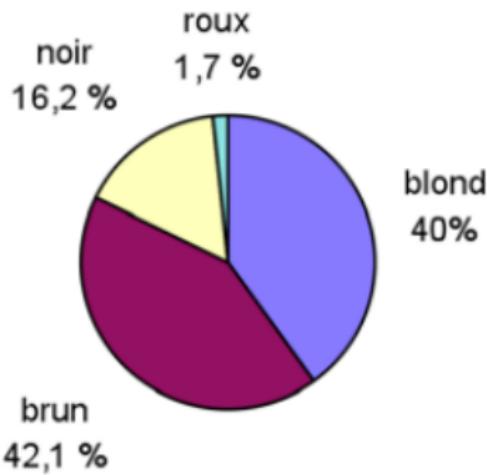
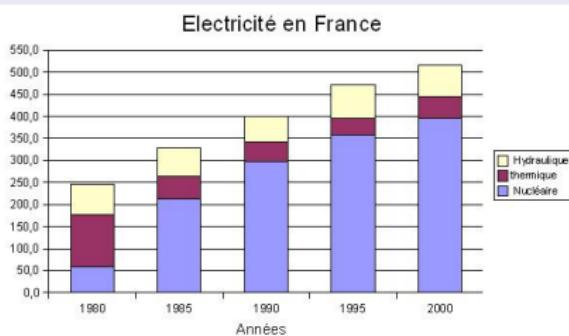
3 Analyse en composantes principales (ACP – PCA)

- Introduction
- ACP : principes
- Pour aller plus loin

4 Et après

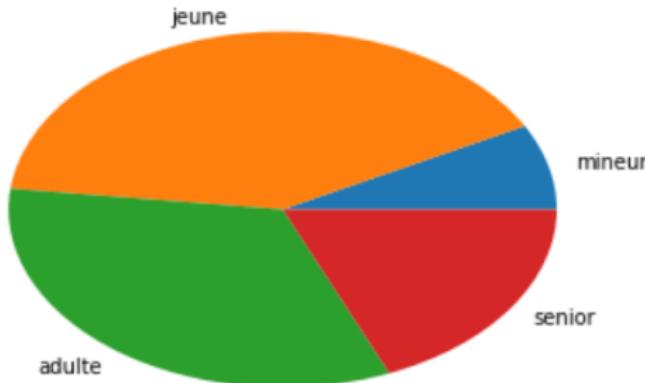
Visualisation effectifs/fréquences (1)

Données qualitatives : barres et camembert



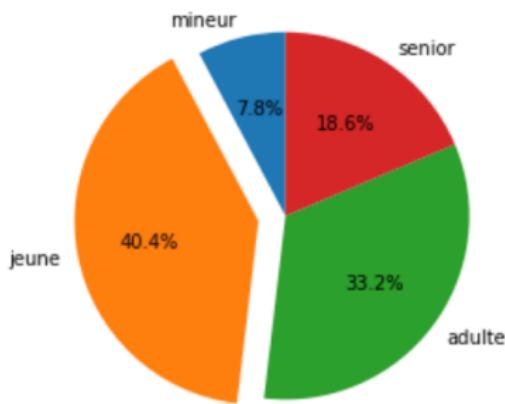
Outils Python (bruts et affinés)

```
import matplotlib.pyplot as plt
import random
name = ['mineur', 'jeune', 'adulte', 'senior']
data = [5000, 26000, 21400, 12000]
plt.pie(data, labels=name)
plt.show()
```



Outils Python (bruts et affinés)

```
import matplotlib.pyplot as plt
import random
name = ['mineur', 'jeune', 'adulte', 'senior']
data = [5000, 26000, 21400, 12000]
explode=(0, 0.15, 0, 0)
plt.pie(data, explode=explode, labels=name, autopct='%.1f%%',
         startangle=90)
plt.axis('equal')
plt.show()
```

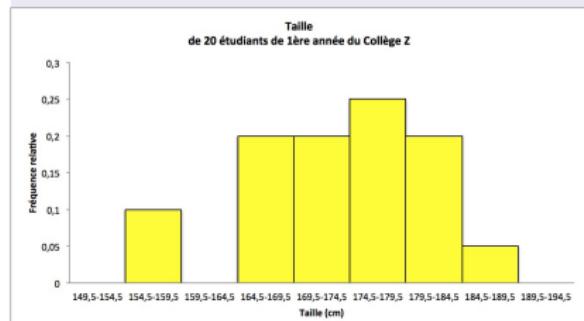
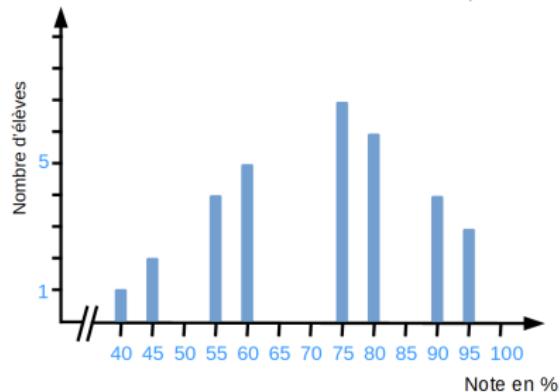


Visualisation effectifs/fréquences (2)

Données quantitatives

Diagrammes en bâtons (un bâton par valeur discrète), ou histogramme lorsque les données sont classées (ou avec intervalle de valeurs)

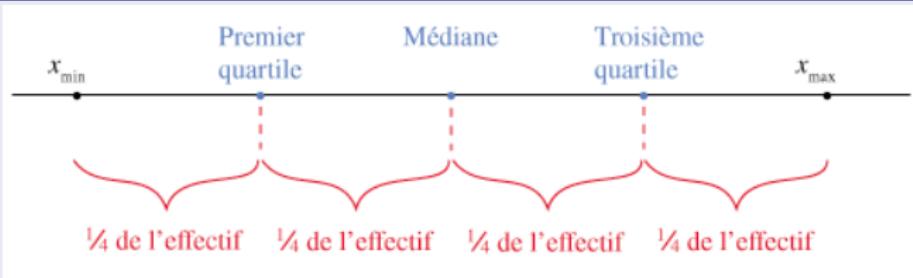
Résultats au dernier test de mathématiques



Médiiane et quartiles

Pour un échantillon S , une seule variable x , estimateurs de distributions de probabilités

La médiane de S et les 3 quartiles



- Médiiane = valeur m_S de x telle qu'il y a autant d'individus dans S pour lesquels $x < m_S$ que d'individus avec $x > m_S$
- Quartile : même principe, mais division en 4 des valeurs prises par S : même quantité d'individus dans chaque partie définie par les quartiles. Une quartile est une valeur de x . Il existe donc 3 quartiles : Q1, Q2 (médiane), Q3

Les percentiles

Les percentiles

Un percentile est un pourcentage d'individus dans S en dessous d'une certaine valeur de x

Poids-pour-l'âge FILLES

De la naissance à 2 ans (percentiles)

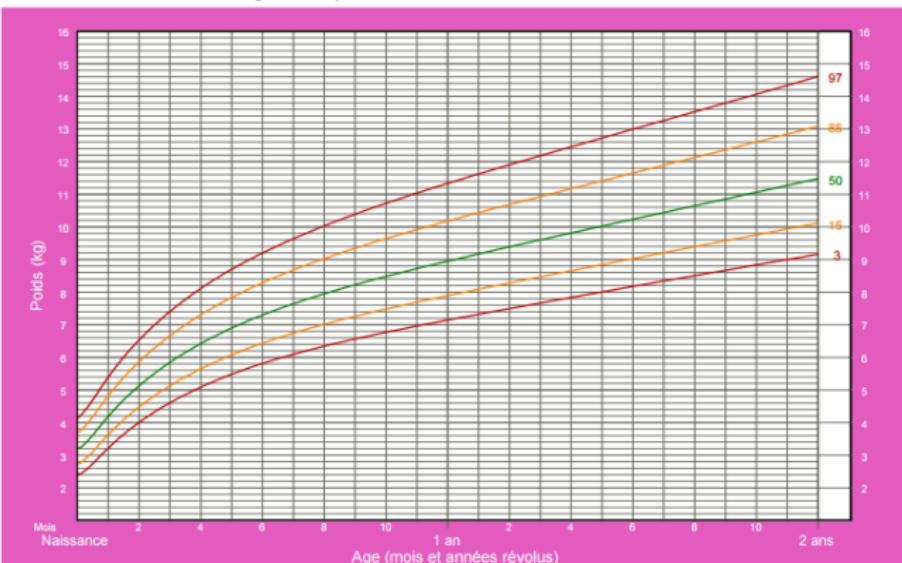


Diagramme de Tukey : visualisation de ces distributions

Pour un échantillon S , une seule variable x

Une boîte à moustache !

- Indication de la médiane (et parfois la moyenne), des deux autres quartiles, valeurs maximum et minimum
- Représentation graphique respectant les écarts entre ces valeurs (et non pas la proportion d'individus)
- Dérives : on y note parfois certains percentiles

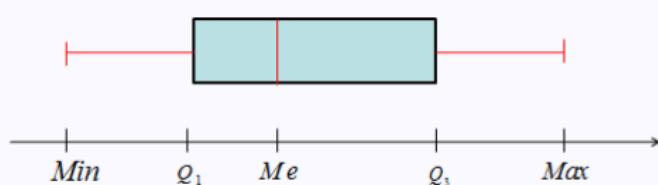
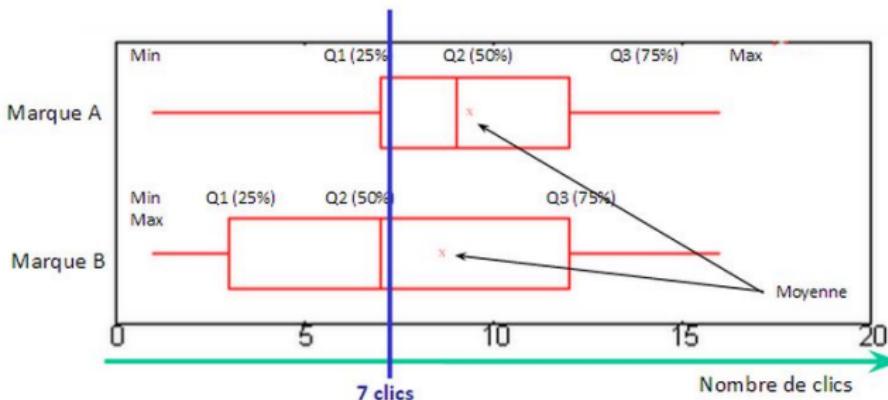


Diagramme de Tukey : visualisation de ces distributions

Pour un échantillon S , une seule variable x

Une boîte à moustache !

- Indication de la médiane (et parfois la moyenne), des deux autres quartiles, valeurs maximum et minimum
- Représentation graphique respectant les écarts entre ces valeurs (et non pas la proportion d'individus)
- Dérives : on y note parfois certains percentiles



Outline

1 Introduction

- Sciences des données, késako ?

2 Représentation et visualisation

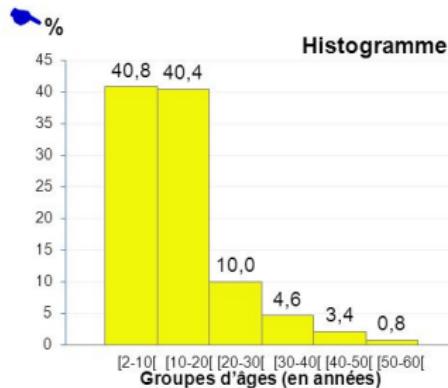
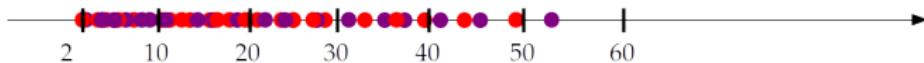
- Représentaions numériques des données
- Des statistiques descriptives aux modes de visualisation
- Quelques modes classiques de visualisation
- Représentation et visualisation des données vectorielles

3 Analyse en composantes principales (ACP – PCA)

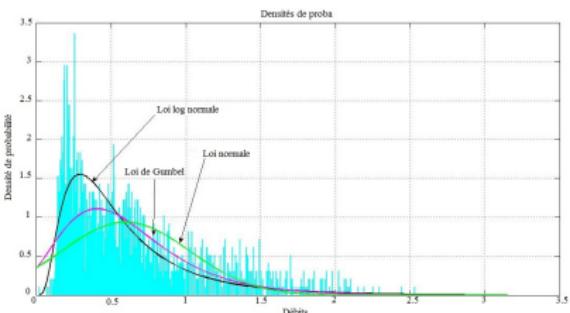
- Introduction
- ACP : principes
- Pour aller plus loin

4 Et après

Visualisation brute d'un ensemble de données 1D : exemples

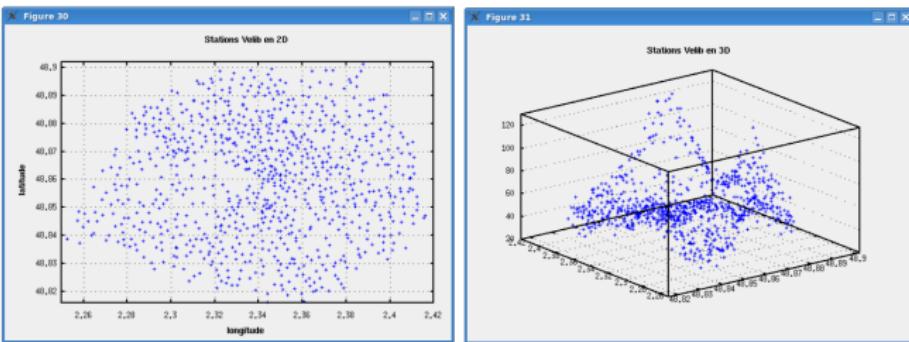


Répartition des patients en fonction de l'âge



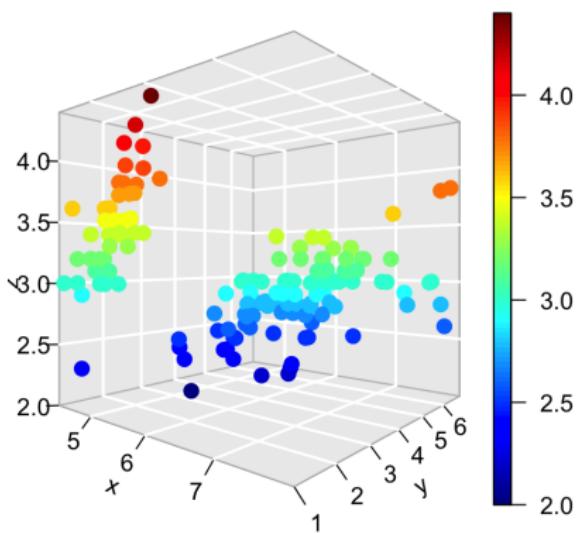
(source : enseeiht)

Visualisation brute d'un ensemble de données 2D, 3D : exemples



(source : N. Cheifetz, 2009)

Visualisation d'un ensemble de données 4D : exemples

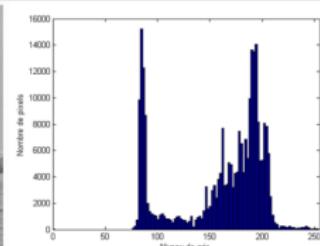
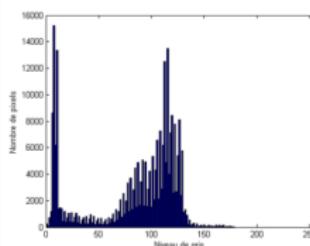


(source : STHDA)

Au delà de 4D : difficile !

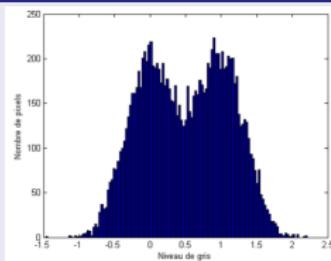
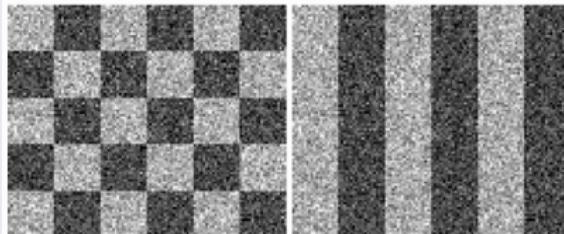
Représentation vectorielle d'une donnée image : exemple

Du tableau de pixels à des représentations spécifiques



Histogramme : observation statistique d'un seul critère (ici, le niveau de gris)
(source : B. Perret)

Limites de l'histogramme

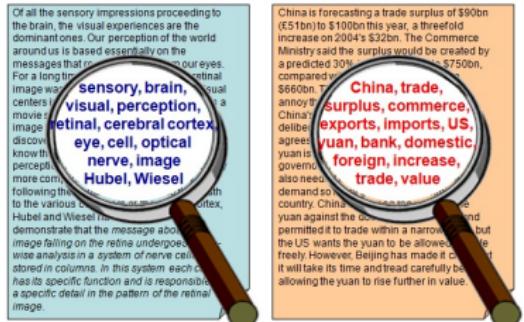


(source : B. Perret)

Représentation vectorielle d'une donnée texte : exemple

Notion de dictionnaire : espace vectoriel

- Sac de mots = représentation d'un texte par les mots qui le composent, sans ordre
- Vecteur : chaque mot du dictionnaire est une composante de l'espace
- Valeur d'une composant : présence/absence, nombre d'occurrences, fréquences, etc.
- Alternatives nombreuses : *n*-grams, word embeddings, etc.



Document 1

The quick brown
fox jumped over
the lazy dog's
back.

Document 2

Now is the time
for all good men
to come to the
aid of their party.

Term	Document 1	Document 2
aid	0	1
all	0	1
back	1	0
brown	1	0
come	0	1
dog	1	0
fox	1	0
good	0	1
jump	1	0
lazy	1	0
men	0	1
now	0	1
over	1	0
party	0	1
quick	1	0
their	0	1
time	0	1

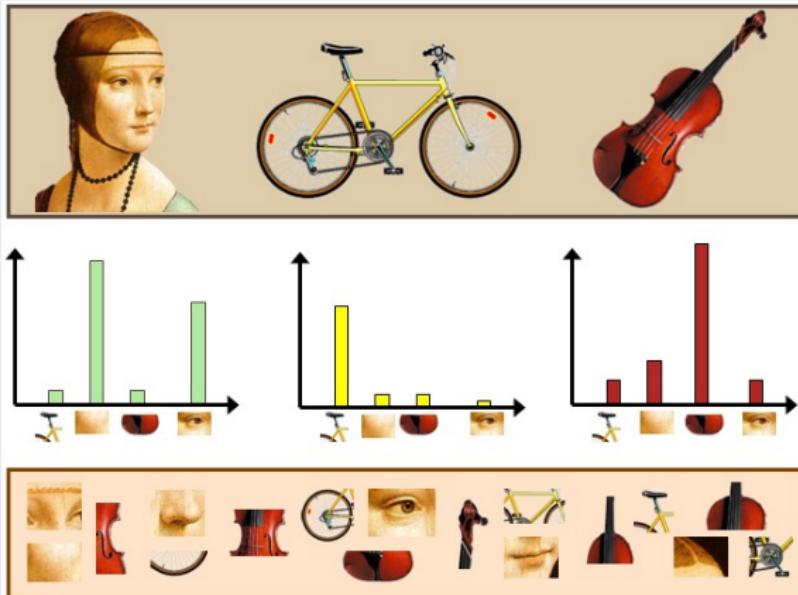
Stopword List

for
is
of
the
to

Extension des sacs de mots aux images

Dictionnaire de mots visuels = espace vectoriel

Toujours une histoire d'histogrammes



(source : Gil's CV blog)

Visualisation d'un jeu de données au delà de 4D

Une réalité

$$S = \{\mathbf{x}_i\}_{i=1}^n, \text{ avec } \mathbf{x}_i \in \mathbb{R}^d$$

- Iris dataset, $n = 150$, $d = 4$
 - Animal with Attributes, $n = 30K$, d de $3 \times 256 =$ de 768 (HOC) à 4000 (BOW), selon espace vectoriel de description
 - Titanic, $n = 500$, $d = 15$

Réduction de dimensions pour un aperçu plus synthétique

- Projection sur deux ou trois variables d'intérêt, lesquelles ?
 - Analyse en composantes principales, pour dégager des combinaisons informatives de composantes
 - Analyse discriminante en cas de supervision : données appartenant à des groupes identifiés : $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, avec $y_i \in \{1, 2, \dots, k\}$

Outline

1 Introduction

- Sciences des données, késako ?

2 Représentation et visualisation

- Représentaions numériques des données
- Des statistiques descriptives aux modes de visualisation
- Quelques modes classiques de visualisation
- Représentation et visualisation des données vectorielles

3 Analyse en composantes principales (ACP – PCA)

- Introduction
- ACP : principes
- Pour aller plus loin

4 Et après

Outline

1 Introduction

- Sciences des données, késako ?

2 Représentation et visualisation

- Représentaions numériques des données
- Des statistiques descriptives aux modes de visualisation
- Quelques modes classiques de visualisation
- Représentation et visualisation des données vectorielles

3 Analyse en composantes principales (ACP – PCA)

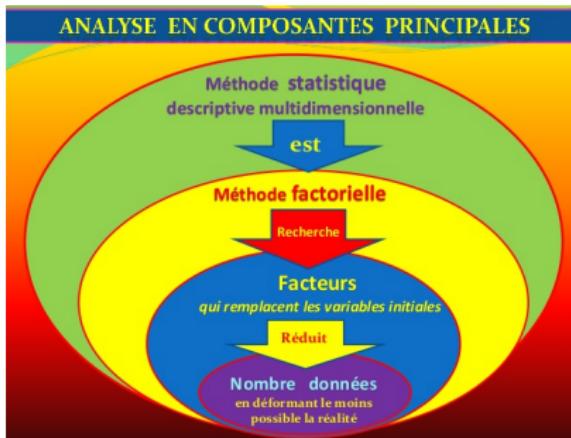
- Introduction
- ACP : principes
- Pour aller plus loin

4 Et après

Analyse en composantes principales : introduction

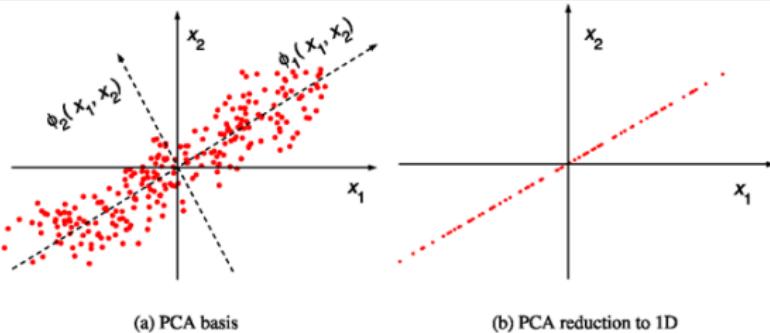
Plus d'un siècle d'existence (Pearson, 1901)

- Statistique multivariée, analyse factorielle
 - Transformation de composantes (axes, variables) corrélées entre elles (ex. $d_3 = ad_1 + bd_2 + c$) en nouvelles composantes décorrélées (=composantes principales)
 - Réduction de dimensions, élimination de redondances, débruitage, donc visualisation et pré-traitement
 - Compression des données



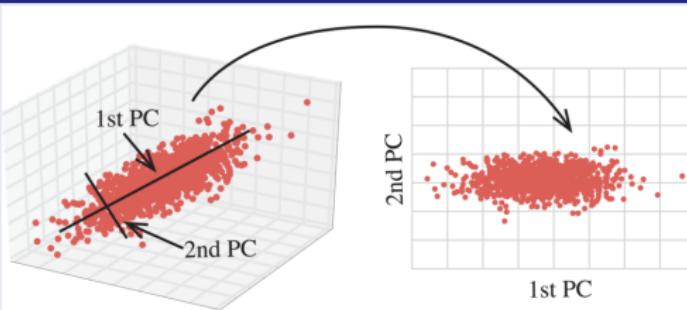
Travail sur un nuage de points à d dimensions

Cas de $d = 2$, visualisation des axes principaux, et projection



(source : In Depth Tutorial)

Cas de $d = 3$, réduction vers $d = 2$



L'Analyse en Composantes Principales : matrice de données

Matrice d'entrées

- X matrice de taille $n \times d$ (32×12), à valeurs réelles (pour l'instant). Un individu (=donnée) par ligne, une variable par colonne
- $x_{i,j}$ est la valeur de la j ème variable pour le i ème individu
- Comparaison de deux lignes = comparaison de deux individus dans l'espace des variables \mathbb{R}^d
- Comparaison de deux colonnes = comparaison de deux variables dans l'espace des individus \mathbb{R}^n
- **Comparaisons** = distances (ressemblances), dépendances (relations) sauce covariance

Exemple

x	sport	sommeil	lecture	internet	repas	...	ménage
x_1	0.04	0.27	0.09	0.11	0.03	...	0.08
x_2	0.11	0.21	0.01	0.08	0.09	...	0.11
x_3	0.03	0.26	0.08	0.12	0.02	...	0.07
...
x_n	0.01	0.31	0.13	0.13	0.08	...	0.02

Outline

1 Introduction

- Sciences des données, késako ?

2 Représentation et visualisation

- Représentaions numériques des données
- Des statistiques descriptives aux modes de visualisation
- Quelques modes classiques de visualisation
- Représentation et visualisation des données vectorielles

3 Analyse en composantes principales (ACP – PCA)

- Introduction
- ACP : principes
- Pour aller plus loin

4 Et après

Comparaisons entre individus

Distance entre deux individus

Ici, distance euclidienne : deux points sont d'autant plus voisins que leurs coordonnées (*activités quotidiennes*) sont proches.

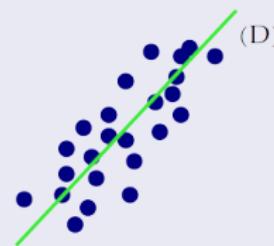
$$d^2(x_1, x_2) = \sum_{j=1}^d (x_{1,j} - x_{2,j})^2$$

Projection des points sur une droite (D)

- Obtenir une meilleure image approchée du nuage de points
- Réfléter la dispersion des points sur cette droite (inertie)
- Minimiser la distance entre chaque point et son projeté

axe principal = D telle que

$$\operatorname{argmax}_D \left\{ \sum_{i=1}^n \sum_{i'=1}^n d_D^2(x_i, x_{i'}) \right\}$$



A la recherche des axes principaux

Principe de l'ACP

- Chercher une représentation alternative des n individus dans un sous-espace vectoriel (F_k) de dimension k , avec k petit (2 ou 3 pour la visualisation)
- = définition de k **nouvelles variables** qui sont des **combinaisons linéaires** des d variables initiales, en perdant le moins d'information possible

Définitions

- *composantes principales* : les nouvelles variables
- *axes principaux* : les axes que les composantes déterminent (dans F_k)
- *facteurs principaux* : les formes linéaires associées

Perdre le moins d'informations possibles

- F_k s'ajuste au nuage des individus
- le nuage (=individus) projeté sur F_k a une grande dispersion

La dispersion mesurée par l'inertie

Inertie d'un nuage de points

$$I_g = \frac{1}{n} \sum_{i=1}^n d^2(x_i, g) \text{ où } g \text{ est le centre de gravité}$$

Soit p_i le projeté orthogonal de la variable x_i sur le sous-espace F

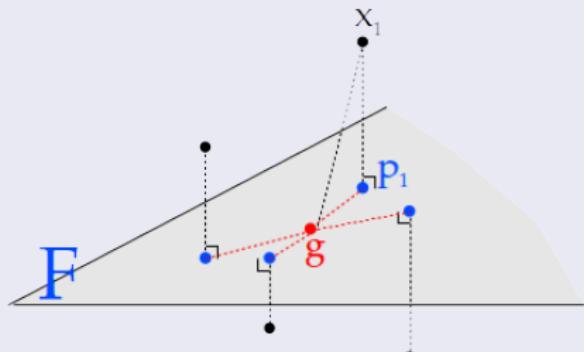
$$d^2(x_i, g) = d^2(x_i, p_i) + d^2(p_i, g)$$

On cherche F tel que

$$\sum_{i=1}^n d^2(x_i, p_i) \text{ soit minimale}$$

donc par Pythagore

$$\text{variance } \sum_{i=1}^n d^2(p_i, g) \text{ maximale}$$



Axes principaux, vecteurs et valeurs propres

Les d axes principaux d'inertie

Axes de direction des **vecteurs propres** de la matrice de covariance, normés à 1

- 1 Premier axe $u_1 = (u_{1,1}, u_{1,2} \dots u_{1,d})$: vecteur associé à la plus grande **valeur propre** λ_1 (sa variance)
- 2 axe u_2 : celui associé à la deuxième plus grande **valeur propre** λ_2
- 3 etc.

A chaque axe principal : une **composante** principale

Une variable obtenue par combinaison linéaire des variables initiales

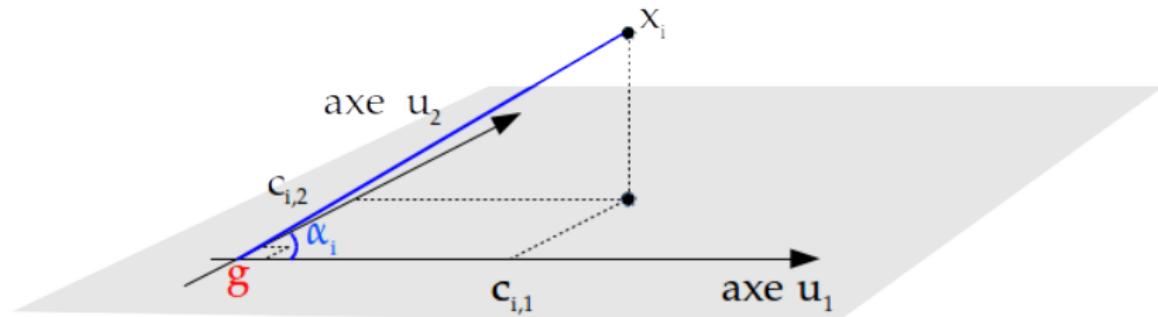
$$c_1 = u_{1,1}d_1 + u_{1,2}d_2 + \dots + u_{1,d}d_d$$

- 1 Composante c_1 : vecteur portant les coordonnées des projections des x_i sur l'axe u_1
- 2 c_2 : vecteur portant les coordonnées des projections des x_i sur l'axe u_2
- 3 etc.

Non corrélation des composantes principales

Représentation des individus

Soit c_j la j ème composante principale : $c_j = \begin{pmatrix} c_{1,j} \\ c_{2,j} \\ \vdots \\ c_{n,j} \end{pmatrix}$ = coordonnées des m individus projetés sur l'axe principal j : $p_i = \langle c_{i,1}, c_{i,2}, \dots, c_{i,d} \rangle$



Pour obtenir une représentation humainement visible, plane, on ne garde que les deux premières composantes

Un algorithme pour calculer les premières composantes principales ?

Evidemment ! Les q premières CP...

Soit S l'échantillon de données (matrice X), n individus, d variables

- 1 Centrez et réduisez les données : pour chaque variable k de chaque individu i dans S , on recalcule X

$$x_i^k \leftarrow \frac{x_i^k - \bar{x}^k}{\sigma_k}$$

- 2 Calculer C la matrice de covariance de X centrée-réduite
- 3 Calculer les valeurs propres de X et leurs vecteurs associés
- 4 Prenez les q plus grandes valeurs propres λ , et les q plus grands axes principaux
- 5 Calculer M la nouvelle représentation matricielle de S dans cette nouvelle représentation

Oui, mais, comment obtient-on les valeurs propres ?

- Inversion de la matrice pour calculer son déterminant
- Tirer partie des propriétés de la matrice de covariance (diago ?)

Outline

1 Introduction

- Sciences des données, késako ?

2 Représentation et visualisation

- Représentaions numériques des données
- Des statistiques descriptives aux modes de visualisation
- Quelques modes classiques de visualisation
- Représentation et visualisation des données vectorielles

3 Analyse en composantes principales (ACP – PCA)

- Introduction
- ACP : principes
- Pour aller plus loin

4 Et après

A la recherche des valeurs propres de la matrice de covariance C

Rappels : définition simplifiée des valeurs et vecteurs propres

Soit une transformation linéaire $f : \mathbb{R}^d \mapsto \mathbb{R}^d$ de matrice carrée A

- Transformée du vecteur \vec{a} vers le vecteur $\vec{b} : \vec{b} = A\vec{a}$
- Lorsqu'il existe $\vec{a}, \lambda, \vec{b} = A\vec{a}$ tels que $\vec{b} = \lambda\vec{a}$ (a et son transformé b colinéaires : même direction), alors λ est une valeur propre, et \vec{a} est un vecteur propre de $A : \lambda\vec{a} = A\vec{a}$
- Caractériser les (λ, \vec{a}) pour lesquels A est une simple homothétie (étirement sans rotation)

$$\det(A - \lambda I) = 0$$

(équation polynomiale de degré d)

Rappels : propriété dans le cas des matrices symétriques

Une matrice carrée M est symétriquessi $M = M^T$

- ses valeurs propres λ sont toutes réelles
- ses vecteurs propres issus des différentes λ sont orthogonaux, et forment une base orthonormée dans laquelle l'application f représentée par M admet une matrice diagonale (théorème spectral)

Exercice de (re-)découverte

Soit l'application linéaire $f : \mathbb{R}^2 \mapsto \mathbb{R}^2$ représentée par la matrice

$$A = \begin{pmatrix} 0 & -\sqrt{3} \\ -\sqrt{3} & -2 \end{pmatrix}$$

En partant du vecteur $\vec{v} = (0, 1)$, et de l'ensemble vide Λ

- 1 calculer $\vec{v}_t = A\vec{v}$
 - 2 est-ce que \vec{v} et \vec{v}_t sont colinéaires, et si oui, rajouter λ dans Λ tel que $\vec{v}_t = \lambda\vec{v}$
 - 3 $\vec{v} \leftarrow \text{rot}(\vec{v}, 30)$ (rotation dans le sens trigonométrique)
 - 4 recommencer en (1) une quinzaine de fois

Quels sont les valeurs propres et les vecteurs propres de A ? Quelle est la matrice diagonale de f dans la nouvelle base orthonormée?

Et en python ?

Learn english

Visualisation

<https://python-graph-gallery.com/> et **matplotlib**

Statistiques descriptives élémentaires

<https://docs.scipy.org/doc/scipy/reference/stats.html>

ACP

[http://scikit-learn.org/stable/modules/generated/
sklearn.decomposition.PCA.html](http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html)

Outline

1 Introduction

- Sciences des données, késako ?

2 Représentation et visualisation

- Représentaions numériques des données
- Des statistiques descriptives aux modes de visualisation
- Quelques modes classiques de visualisation
- Représentation et visualisation des données vectorielles

3 Analyse en composantes principales (ACP – PCA)

- Introduction
- ACP : principes
- Pour aller plus loin

4 Et après

En continuant dans cette science

Dans les chapitres suivants

Dans les prochains chapitres :

- Algorithmes simples d'apprentissage pour
 - la classification supervisée
 - la régression
 - le regroupement (clustering)
- Protocoles généraux d'expérimentation
- Mesures de performances

Et en TD/TP

- Python par la pratique (alternative demandée par employeurs = R, parfois Java)
- Librairies utiles
- Participation à un challenge par équipes de 2 à 4

Au delà ce cours (pour aller plus loin)

- Introduction à l'apprentissage automatique (M1 – S2)
- Master IAAA (M2) à Marseille !
- Stages de pratique recommandés (chez Qarma ou ailleurs)

Ecosystème Python pour la Data Science

Tout au long de ce cours, nous utiliserons principalement

- NumPy : multidimensional array package
- SciPy : scientific computing package
- Matplotlib : plotting library for visualization
- pandas : data analysis library
- scikit-learn : machine learning library

Installer Python et les packages Data Science

Anaconda Python distribution

Anaconda est une distribution libre et open source du langage de programmation Python appliquée au développement d'applications dédiées à la science des données et à l'apprentissage automatique (traitement de données à grande échelle, analyse prédictive, calcul scientifique), qui vise à simplifier la gestion des paquets et de déploiement.

Anaconda installer

<https://www.anaconda.com/download/>

Anaconda quick-start guide

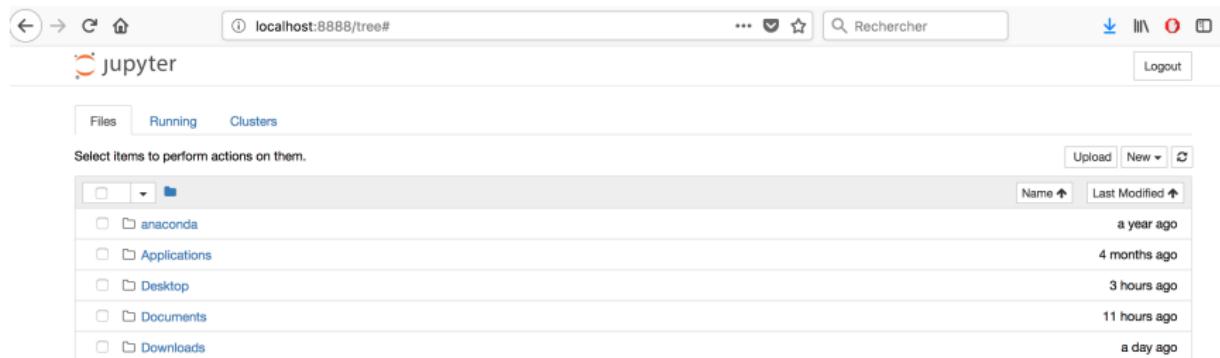
<https://conda.io/docs/user-guide/getting-started.html>

Jupyter Notebooks

- Environment interactif de calcul
- Peut rassembler, dans le même document, du texte, des images, des formules mathématiques et du code informatique exécutable.
- Installé par défaut avec la distribution Anaconda

Pour lancer Jupyter notebook, exécutez la commande suivante sur le terminal :

```
$ jupyter notebook
```



Colaboratory

Un outil google offrant un environnement Jupyter Notebook qui s'exécute dans le cloud et stocke les Notebooks sur Google Drive.

<https://colab.research.google.com/>

The screenshot shows the Google Colaboratory interface. At the top, there's a toolbar with File, Edit, View, Insert, Runtime, Tools, Help, and a status bar showing 'CONNECT' and 'EDITING'. Below the toolbar, there are tabs for CODE, TEXT, CELL, and DISCARD CHANGES. A sidebar on the left displays a 'Welcome to Colaboratory!' message and a snippet of TensorFlow code demonstrating matrix addition. The main workspace contains several code cells. One cell shows the addition of two 3x3 matrices:

$$\begin{bmatrix} 1 & 1 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 2 & 3 \end{bmatrix} = \begin{bmatrix} 2 & 3 & 4 \end{bmatrix}$$
$$\begin{bmatrix} 1 & 1 & 1 \end{bmatrix} + \begin{bmatrix} 4 & 5 & 6 \end{bmatrix} = \begin{bmatrix} 5 & 6 & 7 \end{bmatrix}$$

Another cell imports TensorFlow and NumPy, and runs a session to add two tensors. A third cell imports matplotlib and plots a scatter plot with a linear regression line. The bottom of the screen shows a progress bar at 28% completion.

```
[ ] print('Hello, Colaboratory!')  
Hello, Colaboratory!  
  
Colaboratory allows you to execute TensorFlow code in your browser with a single click. The example below adds two matrices.  
  
[ ] 
$$\begin{bmatrix} 1 & 1 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 2 & 3 \end{bmatrix} = \begin{bmatrix} 2 & 3 & 4 \end{bmatrix}$$
  

$$\begin{bmatrix} 1 & 1 & 1 \end{bmatrix} + \begin{bmatrix} 4 & 5 & 6 \end{bmatrix} = \begin{bmatrix} 5 & 6 & 7 \end{bmatrix}$$
  
  
[ ] import tensorflow as tf  
import numpy as np  
  
with tf.Session():  
    input1 = tf.placeholder(tf.float32, [2, 3])  
    input2 = tf.placeholder(tf.float32, [2, 3])  
    output = tf.add(input1, input2)  
    result = output.eval()  
print(result)  
  
[[ 2,  3,  4],  
 [ 5,  6,  7]]  
  
Colaboratory includes widely used libraries like matplotlib, simplifying visualization.  
  
[ ] import matplotlib.pyplot as plt  
import numpy as np  
  
x = np.arange(20)  
y = np.random.rand(20) + np.random.randn(1).ravel()  
a, b = np.polyfit(x, y, 1)  
plt.scatter(x, y, 'o', np.arange(20)+b, a*np.arange(20)+b, '^')
```