R 5.12, Ressources et Culture Numériques 5

D. Fourer

BUT 3e année, IUT d'Évry-Val-d'Essonne, dept. TCJ

5 septembre 2024

Plan

Contact : **dominique.fourer@univ-evry.fr**
Suivi : `https://fourer.fr/Ens/2425/RCN5`

Objectifs du cours

- Introduction à la science des données (“big data” / “data mining”)
- Construction de tableaux de bords pertinents
- Prise en main d’un progiciel de gestion intégré (PGI ou ERP)

$\Rightarrow$ Utilisation pour une activité professionnelle.

Contenu

- 15 heures de cours/TD
- 2 devoirs surveillés

Plan

1. Introduction à la science des données
   - Théorie de l'information
   - Les données massives ou "big data"
   - Techniques de visualisation des données

2. Apprentissage automatique et fouille de données
   - Méthodes de régression
   - Apprentissage non-supervisé et *clustering*
   - Apprentissage supervisé

Plan

Données

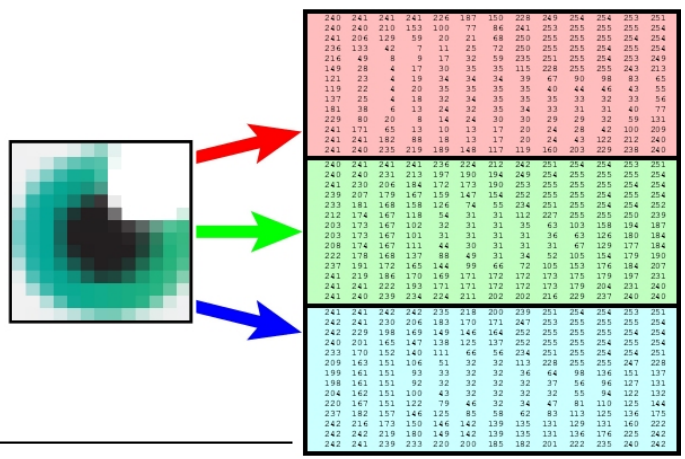
Définition

Une donnée est la représentation d'une information pouvant être traitée par un ordinateur. Cette dernière repose sur le codage binaire $b \in 0, 1$ (informatique classique) ou sur des qubits (informatique quantique) qui correspondent à une superposition des 2 états : $b_q = \alpha|0> + \beta|1>$ avec $|\alpha| + |\beta| = 1$.

Il existe différents types de données : numérique (entier, flottant, complexe), chaîne de caractères, date, audio (formats wave, mp3, ogg, etc.), image (formats jpeg, png, bmp, etc.), vidéo (formats mov,mkv, mp4 H.264, etc.), format de fichier, etc.

La quantité de données est souvent mesurée en octets.

- Exemple 1 : 01100001 correspond à la lettre "a" dans la table ASCII [1]
- Exemple 2 : Codage d'une image RGB (3 octets/pixel)



1. American Standard Code for Information Interchange

Codage de l'information

Exemple 3 : Codage des nombres entiers sur 3 bits.

Code	Nombre
000	0
001	1
010	2
011	3
100	4
101	5
110	6
111	7

Sur n bits, on peut représenter 2^n nombres différents.

Exemple 4 : Codage des jours de la semaine sur 3 bits

On peut décider d'un codage arbitraire qui associe ou pas un code pour chaque entrée du dictionnaire.

Code	Nombre
000	lundi
001	mardi
010	mercredi
011	jeudi
100	vendredi
101	samedi
110	dimanche

Codage des nombres réels

Codage en virgule flottante (norme IEEE 754)

Précision	(S)igne	(E)xposant	(M)antisse	Valeur	chiffres significatifs
Simple (32 bits)	1 bit	8 bits	23 bits	$(-1)^S \times M \times 2^{E-127}$	≈ 7
Double (64 bits)	1 bit	11 bits	52 bits	$(-1)^S \times M \times 2^{E-1023}$	≈ 16

Exemple (simple précision) : V=0011 0100 0000 0010 0110 0111 1111 0010

(S)igne	(E)xposant	(M)antisse
0	011 0100 0 (104)	000 0010 0110 0111 1111 0010 (157 682)

$$V = (-1)^0 \times 157682 \times 2^{(104-127)} = \boxed{0,0188}$$

Information et entropie

Mesure de l'information exprimée en bits Shannon.
Pour un message donné x_i, son entropie vaut :

$$H(x_i) = -\log_2(P(x_i)) \tag{1}$$

avec $P(x_i)$ la probabilité de recevoir le message x_i et $\log_2(x_i) = \frac{ln(x_i)}{ln(2)}$ le logarithme base 2.
Pour une source (discrète) X (v.a.), l'entropie correspond à son espérance :

$$H(X) = -\sum_i P(x_i)\log_2(P(x_i)) \tag{2}$$

Propriétés

- Plus un événement est fréquent, plus son entropie est faible.
- Si $P(x) = 1$, alors $H(x) = 0$ donc il est inutile de coder cette information.
- Si $P(x) \to 0^+$, alors $H(x) \to +\infty$

Codage de Huffman

Permet de produire un code de taille minimale (compression sans perte)

Algorithme

On part d'un dictionnaire X et d'un ensemble de poids p_i.

- On trie les poids $p_i \sim P(x_i)$ par ordre croissant.
- On construit un arbre binaire en partant des feuilles où chaque couple (x_i, p_i) correspond à une feuille et où chaque noeud est associé à un poids p_i. Chaque noeud de l'arbre est construit itérativement en sélectionnant systématiquement les 2 noeuds précédents ayant le poids le plus faible. Le poids associé au nouveau noeud correspond alors à la somme des poids de ses fils. Les fils du nouveau noeud pouvant être une feuille ou un noeud calculé précédemment.
- On réitère la construction des noeuds jusqu'à atteindre la racine (de poids $\sum_i p_i$).

Codage de Huffman (Exemple)

Le message **commencement** est associé au tableau de poids suivant :

x_i :	c	o	m	e	n	t
p_i :	2	1	3	3	2	1

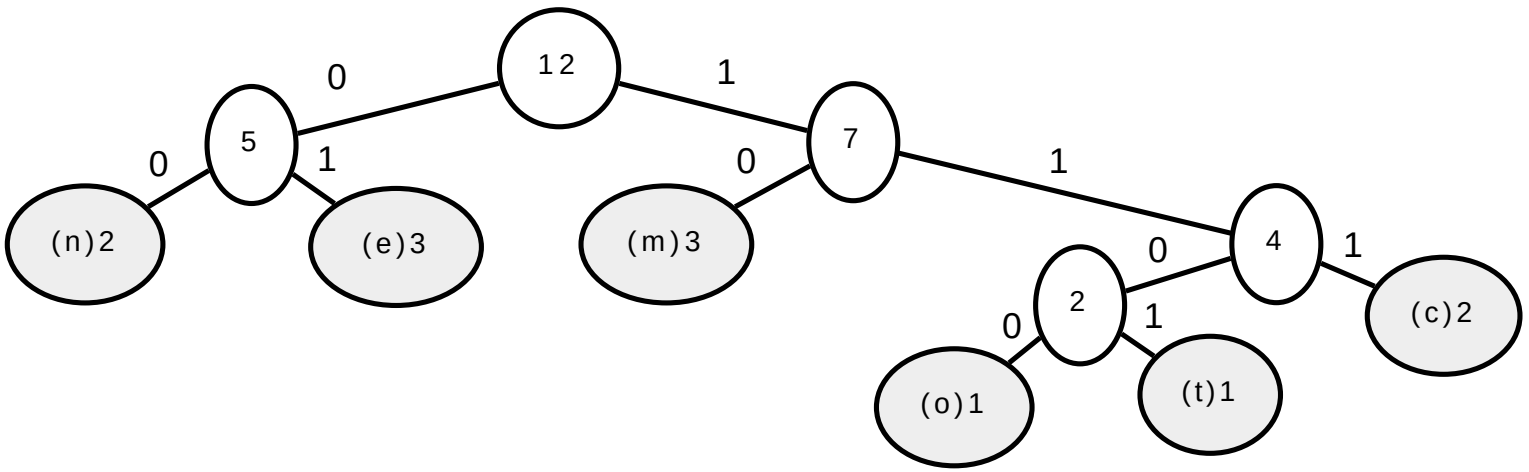


Figure : Arbre binaire du message **commencement** utilisant la convention gauche :0, droite :1 pour le codage des arêtes.

Le code binaire de chaque symbole pour le message **commencement** est donné par :

c	o	m	e	n	t
111	1100	10	01	00	1101

.

Autres techniques de compression (sans perte)

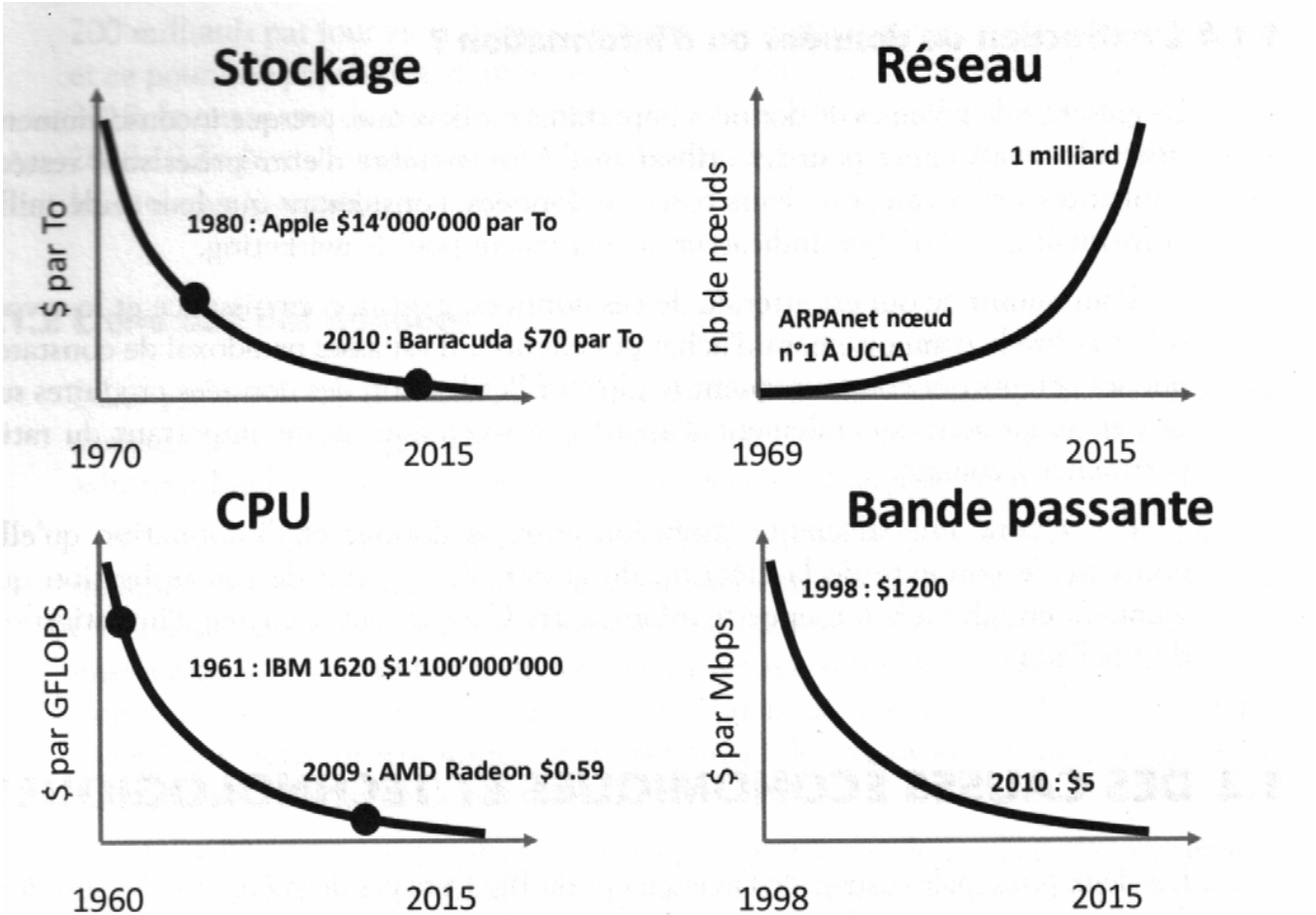
- RLE (run-length encoding) : substitue les répétitions
- CCITT : utilisée par les fax
- Baudot
- Lempel-Ziv : utilisé par le format zip
- Codage par modélisation de contexte (PPM, CM, etc.)
- etc.

Formats de compression de fichiers usuels : zip, rar, tar, gzip, ...

Ordres de grandeur

Préfixe	Unité	Quantité	Exemple
Kilo	Ko	10^3 o	Document word (sans images) ≈ 20 *Ko*
Méga	Mo	10^6 o	Image Jpeg 4032×3024 ≈ 4 Mo
Giga	Go	10^9 o	Un film DVD ($\approx 4,7$ Go)
Téra	To	10^{12} o	Capacité d'un disque dur (entre 0,5 et 16 To)
Péta	Po	10^{15} o	Les données disponibles sur le web
Exa	Eo	10^{18} o	Les informations générées dans le monde ≈ 2*Eo*
Zetta	Zo	10^{21} o	

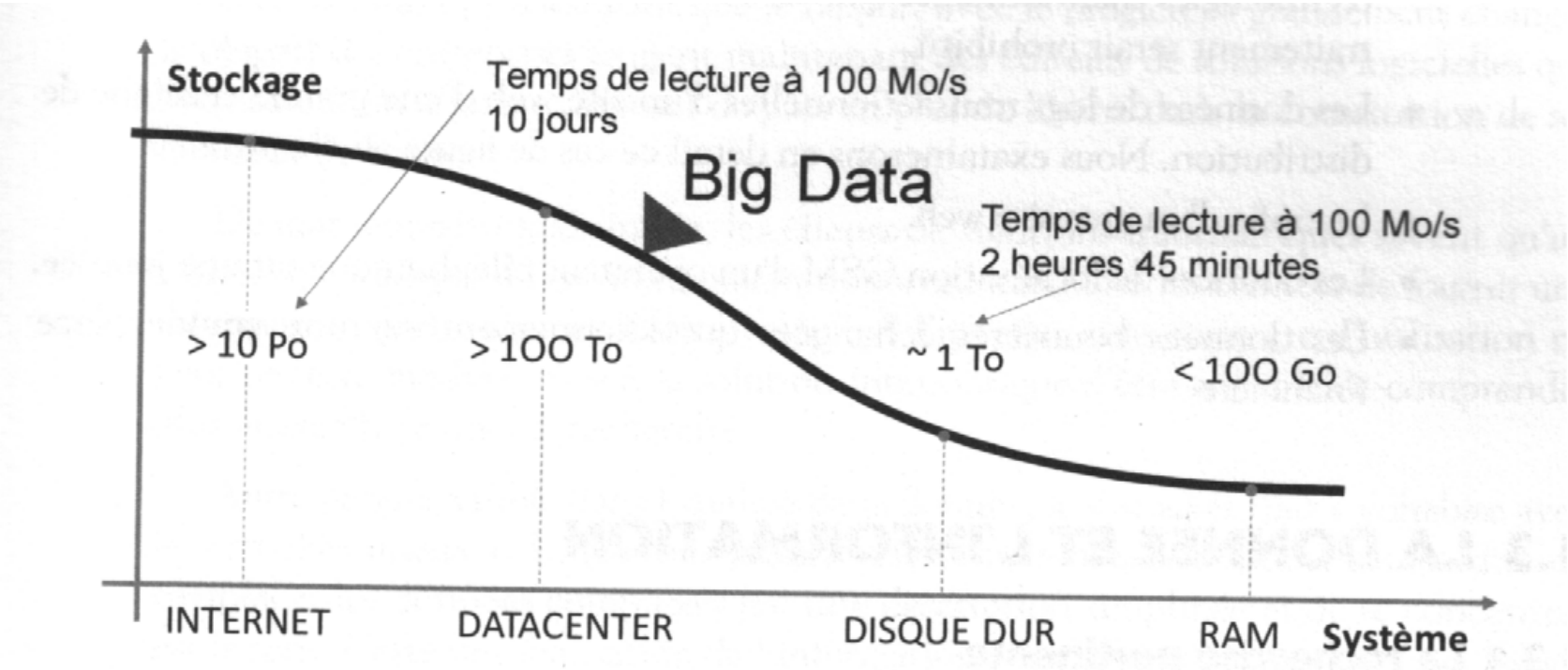
Vers le "big data"



Baisse du coût des ressources durant les dernières décénies [2].

2. source : Lemberger et al."Big Data et Machine Learning" 2nd Ed. Dunod

Données massives "Big Data"



Ordre de grandeur d'espaces de stockage et frontière approximative du "Big Data". [3].

3. source : Lemberger et al."Big Data et Machine Learning" 2nd Ed. Dunod

Le "big data" en bref

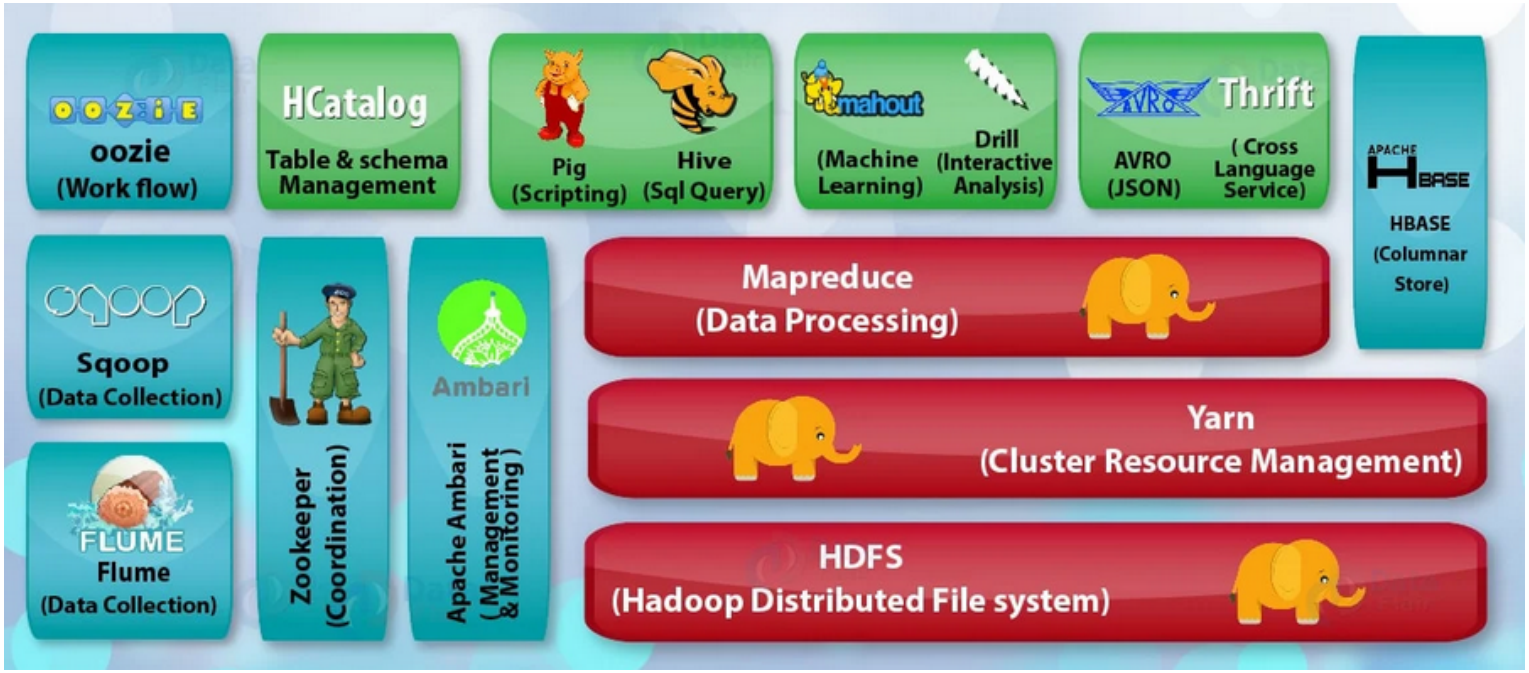
Exemples qui ne relèvent pas du "big data"

- Données pouvant être traitées sur Excel
- Données que l'on peut héberger sur un seul noeud d'une base de donnée relationnelle
- Données chères à produire (eg. recensement, sondage, INSEE, etc.)
- etc.

Exemples de données massives

- Données de traffic d'un site web important (eg. twitter, google, etc.)
- Données qu'il n'est pas possible de stocker ou de traiter avec des moyens traditionnels
- Données boursières quotidiennes (eg. CAC40, NASDAQ, etc.)
- Données de localisation GSM journalières d'un opérateur téléphonique
- etc.

Quelques outils : Hadoop



Écosystème développé par Apache en Java pour le stockage et le traitement de données massives. Hadoop est open source et repose sur des solutions standard peu coûteuses distribuées.

Idées de base

- Traiter les données par lots séparés avant de fusionner les informations
- Synthétiser les informations pertinentes, travailler sur des représentations synthétiques
- Distribuer le stockage et le traitement sur plusieurs serveurs

Objectifs

- Représenter les informations de façon synthétique
- Extraire les informations pertinentes à partir de données brutes
- Inférence statistique et prise de décision

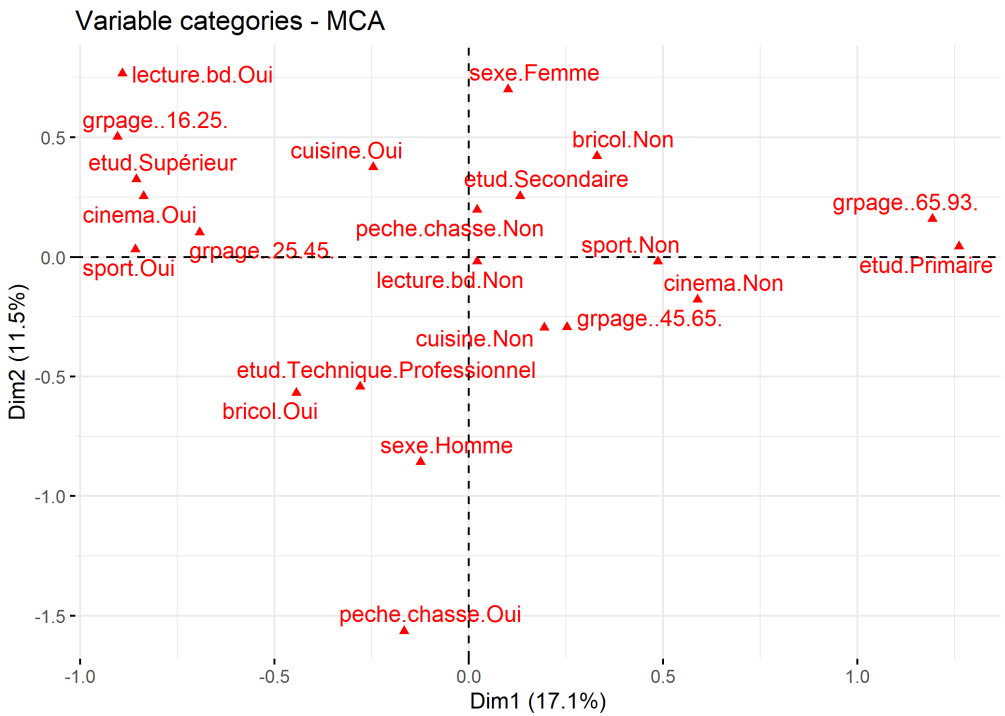


Figure : Exemple d'Analyse des correspondances multiples [4].

4. **source** :https://larmarange.github.io/analyse-R/analyse-des-correspondances-multiples.html

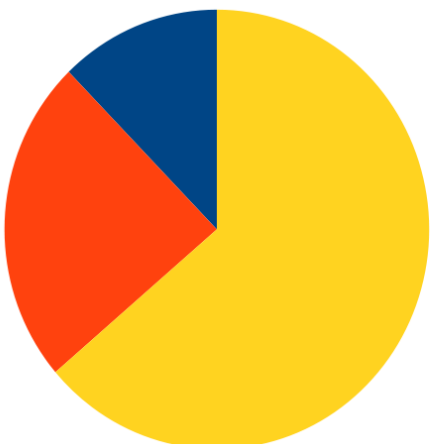
Table de contingence

Répartition d'une population sur 2 variables statistiques.

	Salaire en euros		
X \ Y	<1000	[1000;2500]	>2500
Ville 1	2100	4100	10900
Ville 2	1300	8300	2100
Ville 3	3200	5400	8300
Ville 4	5600	7100	2900

Ville 1

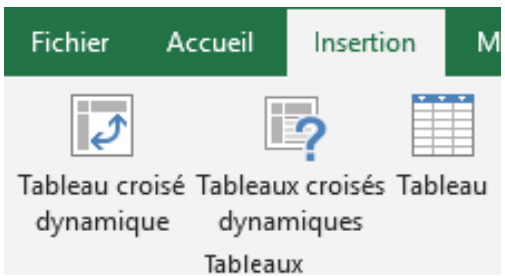
Répartition de la population en fonction des revenus



Ville 4

Répartition de la population en fonction des revenus

Tableau croisé dynamique



Création

1. Sélection de la source de données.
2. Choix de l'emplacement du tableau croisé dynamique
3. Choix des critères de regroupement

- **lignes** : nom des champs
- **colonnes** : hiérarchie entre les champs si plusieurs critères
- $\sum$: valeurs numériques (nombre, somme, moyenne, min, max, etc.)

Histogramme et loi empirique 1/2

L'histogramme est la représentation graphique des effectifs n_i d'une population de taille N relative à une ou plusieurs variables. Il nécessite le choix d'un pas de quantification (nombre de classes $I \ll N$, tel que $i \in [1, I]$) pour représenter chaque variable quantitative.
On peut déduire la loi empirique des données en divisant les effectifs par la taille de la population étudiée.

$$\hat{P}(x_i) = f_i = \frac{n_i}{N} \tag{3}$$

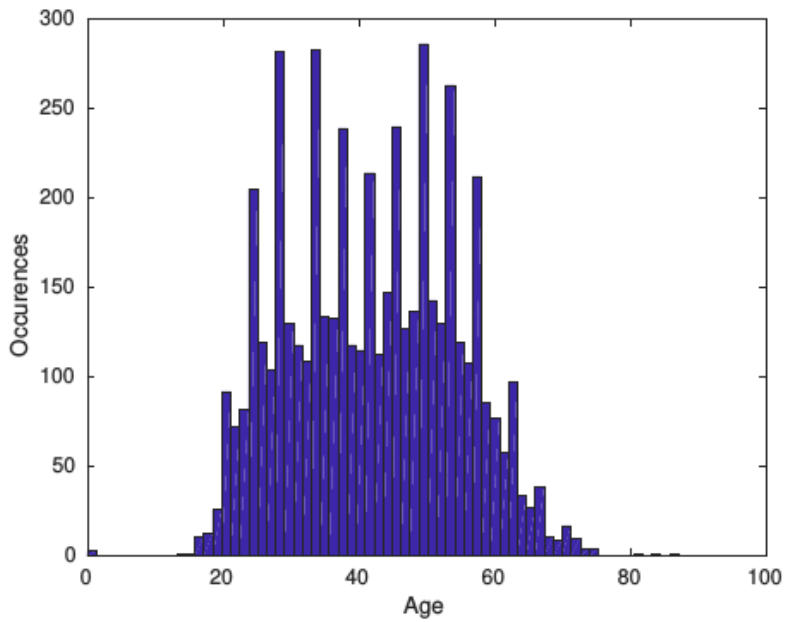
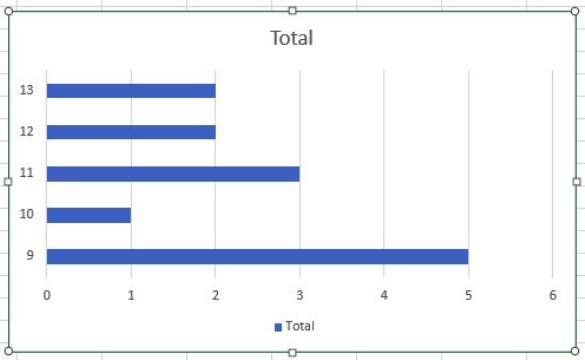


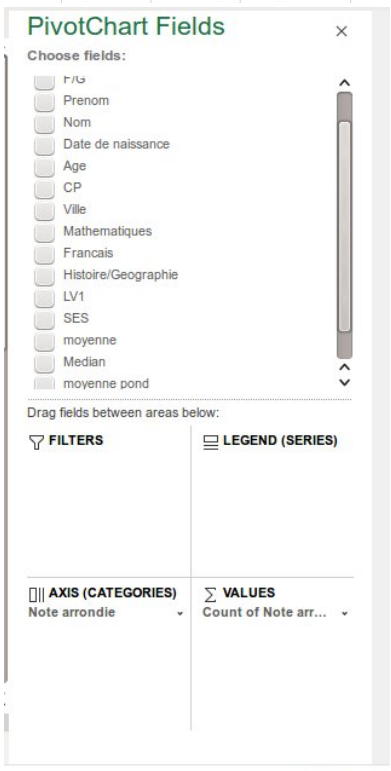
Figure : Exemple d'histogramme ($N = 5105$).

Exemple : Histogramme des notes dans une classe

Prenom	Nom	Date de naissance	Age	CP	Ville	Mathematiques	Francais	Histoire/Geo	LV1	SES	moyenne	Median	moyenne pond	Note arrondie
Patrick	Adler	10/28/95	23	27000	ÉVREUX	13,9	9,5	5,5	8,1	9,3	9,26	9,3	9,97	10
Matthieu	Blondlot	09/09/96	23	59190	HAZEBROUCK	9,25	17	8	15,3	5,75	11,06	9,25	10,82	11
Amber	Brousseau	04/24/95	24	64000	PAU	10,5	14,3	13,4	15,17	11,95	13,064	13,4	12,6855	13
Dominique	Duranseau	10/10/95	23	91000	EVRY	16,95	10	5,5	11,21	9,3	10,592	10	11,4515	11
Belda	Grimard	03/04/95	24	13100	AIX-EN-PROVENCE	5,8	14	8,6	8,9	7,03	8,866	8,6	8,571	9
Josette	Guemon	03/08/97	22	77200	TORCY	14,5	15,8	14,5	9,3	10,34	12,888	14,5	13,148	13
Orane	Lamy	03/16/94	25	93100	MONTREUIL	12,4	11,7	10,3	5,9	19,4	11,94	11,7	12,37	12
Alain	Langelier	04/11/99	20	68200	MULHOUSE	5,6	14,2	10,25	10,5	7,5	9,61	10,25	9,1325	9
Marc	Marois	07/11/94	25	75003	PARIS	6,2	9,75	9	13,41	8	9,272	9	8,7715	9
Philippe	Noël	10/20/95	23	55100	VERDUN	10,18	8,72	11,5	10	4,5	8,98	10	8,923	9
Norbert	Orville	04/08/95	24	59400	CAMBRAI	12,3	19,5	2,3	4,5	4,03	8,526	4,5	9,416	9
Anouk	Poissonnier	05/25/95	24	94270	LE KREMLIN-BICÊTRE	13,85	18,75	8,34	5	7	10,588	8,34	11,306	11
Karlotta	Turgeon	07/11/98	21	91000	EVRY	10,75	8,4	13,5	13,5	13,5	11,93	13,5	11,655	12
									TOTAL:		10,5058462	10,18	10,63230769	

Row Labels	Count of Note arrondie
9	5
10	1
11	3
12	2
13	2
Grand Total	13





Statistiques descriptives

Objectif : synthétiser l'information contenue dans un ensemble de données.

Espérance (moyenne pondérée)

$$E[X] = \sum_i P(x_i)x_i = \bar{x} \tag{4}$$

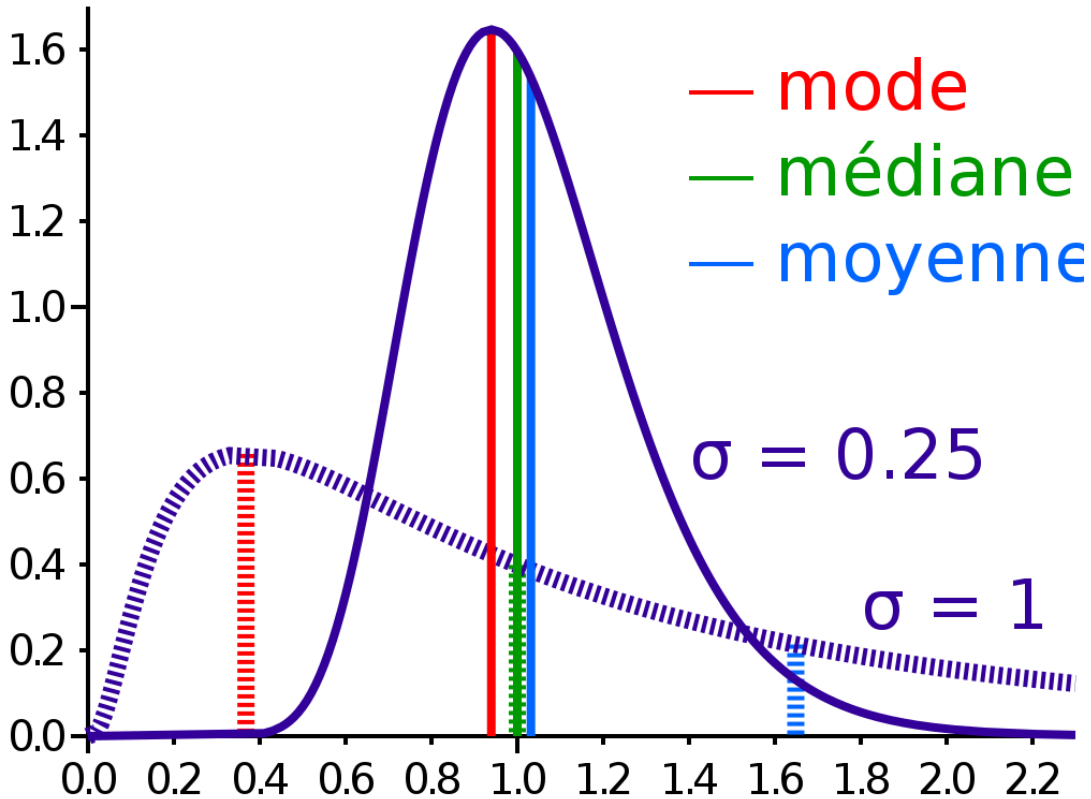
Variance

$$\mathrm{Var}(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2 = \sum_i P(x_i)(x_i - \bar{x})^2 \tag{5}$$

Écart-type

$$\sigma_X = \sqrt{\mathrm{Var}(X)} \tag{6}$$

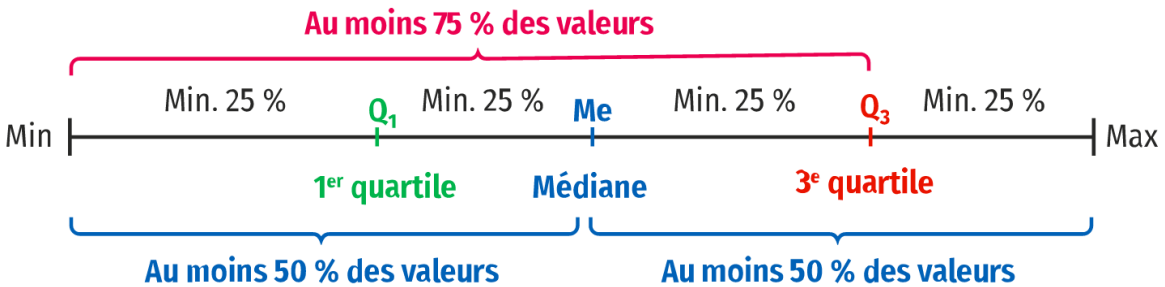
Loi de probabilité, Mode, Médiane



Définition

- Mode : valeur **m** qui maximise une loi de probabilité : $\mathbf{m} = \arg\max_x P(X = x)$
- Médiane : valeur m qui sépare la population en deux sous-effectifs de même taille (i.e $P(X \geq m) = P(X \leq m) \geq \frac{1}{2}$)

Quartiles, quantiles

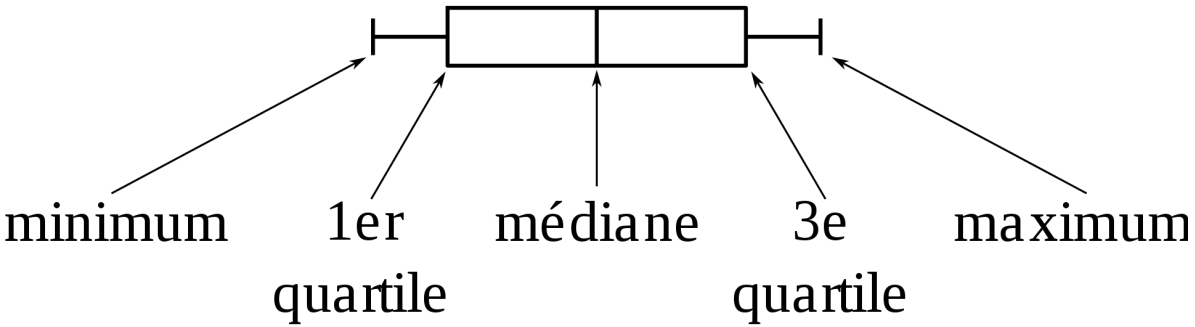


Définition

- Quartile : Divise une distribution en 4 parties égales ;
  $P(X \leq q_1) = P(q_1 \leq X \leq q_2) = P(q_2 \leq X \leq q_3) = P(X \geq q_4)$
- Quantile : Généralise la notion de quartile pour un nombre arbitraire de parties égales dans une distribution.

Remarque : la médiane correspond au second quartile q_2

Boîte à moustache



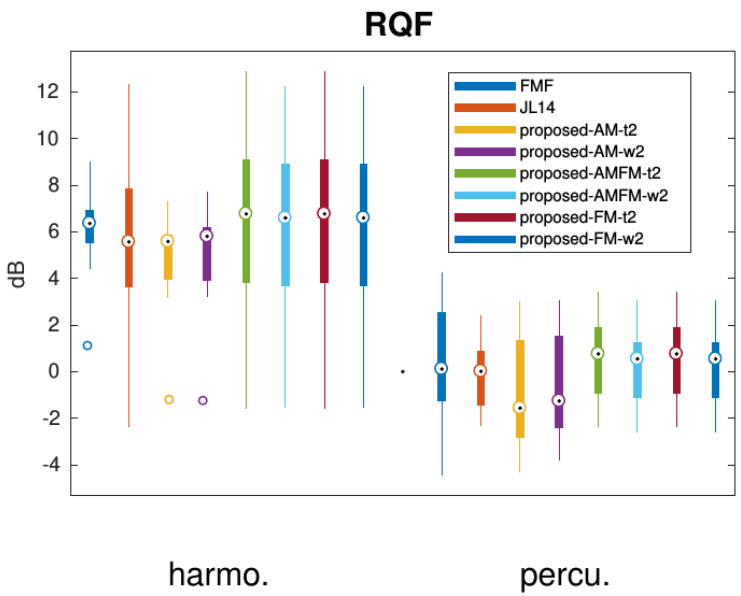


Figure : Résultats comparatifs de plusieurs techniques exprimées en dB. [5]

5. D. Fourer, Separation de Sources Harmoniques / Percussives utilisant des estimateurs locaux de modulation lineaire AM-FM. Proc. GRETSI 2022. Nancy, France.

Analyse en composantes principales

- Réduire la dimension des données
- Décorréler les individus selon des axes principaux
- Obtenir k nouvelles variables combinaison linéaires des p variables initiales

$\Rightarrow$ Les k nouvelles variables sont les composantes principales. Elles définissent les axes principaux.

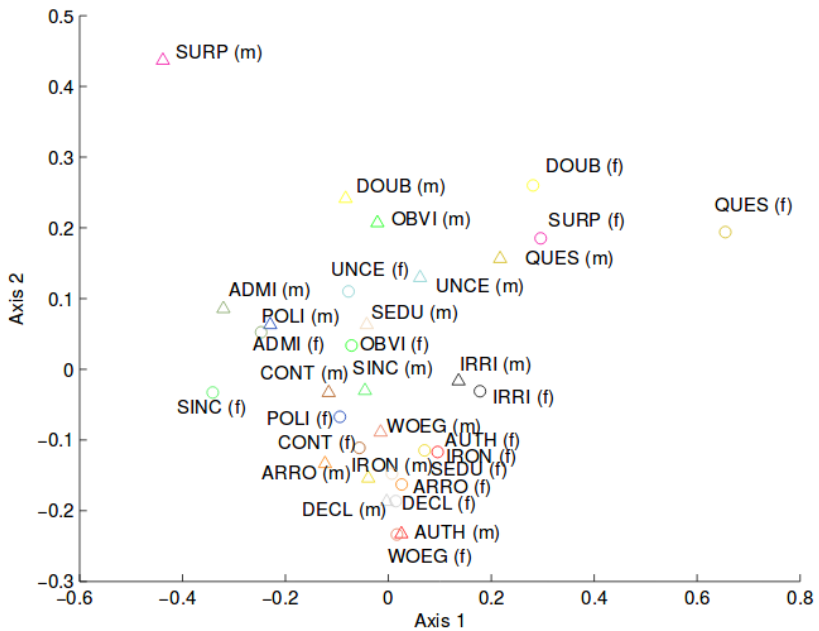


Figure : ACP des attitudes socio-culturelles en japonais. [6]

6. D. Fourer, T. Shochi, JL Rouas, JJ Aucouturier and M Guerry. Prosodic analysis of spoken Japanese attitudes. Proc. Speech Prosody 7. May 20-23 2014. Dublin, Ireland.

Calcul d'une ACP

Données réparties dans une matrice de dimension $N \times p$. N lignes (individus) et p variables quantitatives.

$$X = \begin{pmatrix} X_{1,1} & X_{1,2} & \dots & X_{1,p} \\ X_{2,1} & \vdots & \dots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ X_{N,1} & X_{N,2} & \dots & X_{N,p} \end{pmatrix} \tag{7}$$

Étapes

- Calcul de la matrice de covariance $M = \frac{1}{N} X^T X$
- Décomposition en valeurs propres λ_i et vecteurs propres V tels que $M = V \Lambda V^{-1}$
- On projette chaque individu sur la nouvelle base formée par les vecteurs propres
- On réduit le nombre d'axes principaux à $q < k$ dont l'inertie cumulée vaut $C_q = \frac{\sum_{i=1}^{q} \lambda_i}{\sum_{i=1}^{k} \lambda_i}$

Première partie

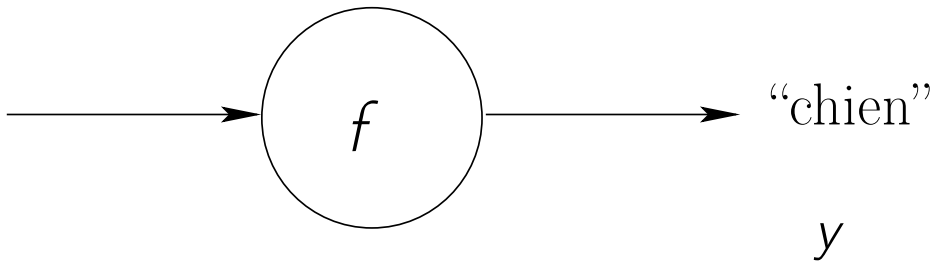
Fin de la première partie du cours

Plan

Objectifs

Comment construire f qui fonctionne pour n'importe quel x ?

x



- Interprétation des données : régression
- Détection d'agrégats : *clustering*, segmentation
- Prédiction : classification automatique

Plusieurs configurations possibles

- Supervisé : on dispose d'exemples annotés permettant d'entraîner un modèle : (x^{train}, y^{train}).
- Non-supervisé : on exploite la structure des données sans annotation (apprentissage d'une représentation et d'un espace métrique).

Régression linéaire (rappels)

Objectif

Construire un modèle permettant d'approximer les données. Modèle linéaire de la forme $y = \tilde{a}x + \tilde{b}$.

- x, y : variables observées
- moyennes $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$, $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$
- écart-type : $\sigma_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x - \bar{x})^2}$
- Covariance : $\mathrm{Cov}(x, y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$

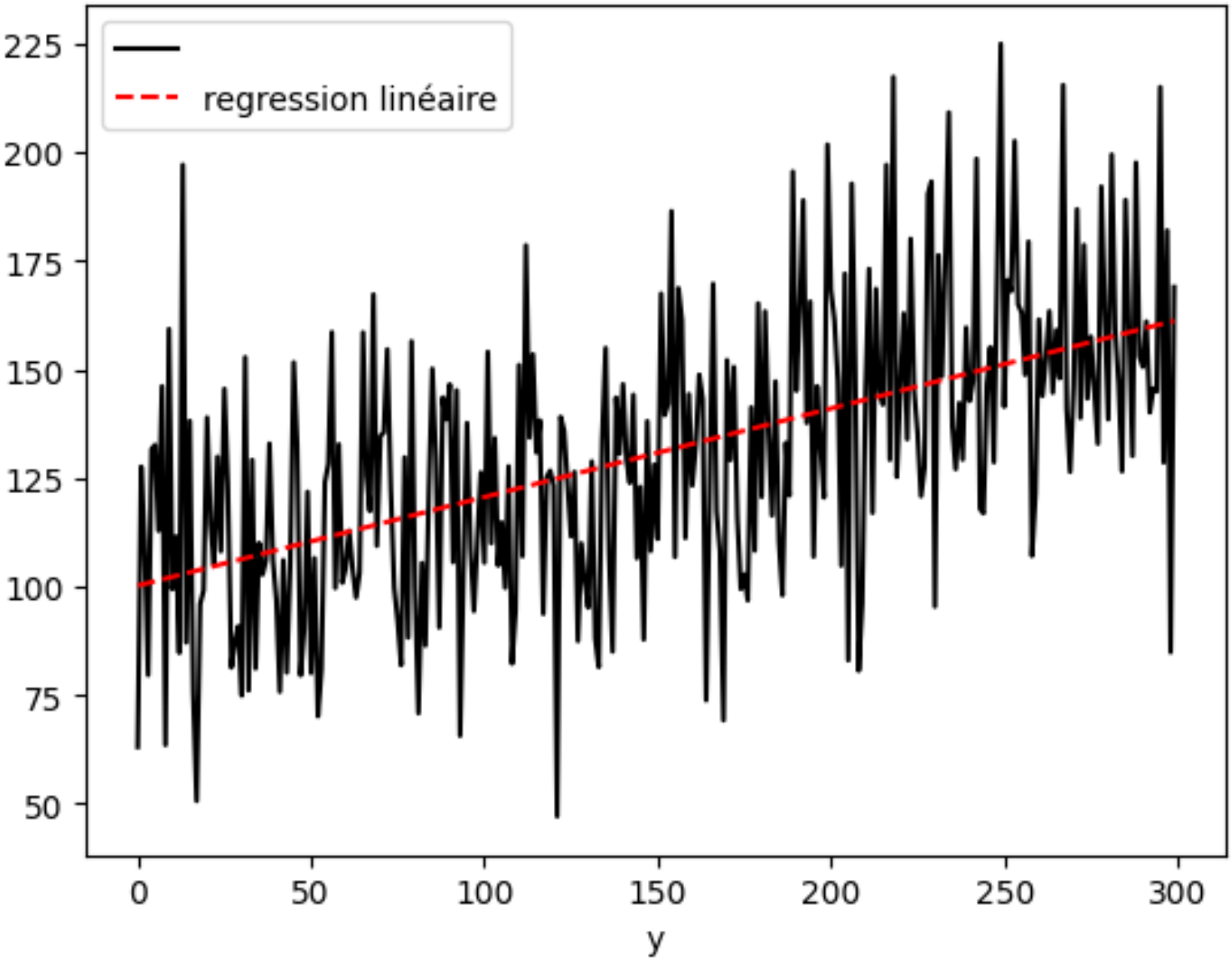
Solution des moindres carrés

$$\tilde{a} = \frac{\mathrm{Cov}(x, y)}{\sigma_x^2} \tag{8}$$

$$\tilde{b} = \bar{y} - \tilde{a}\bar{x} \tag{9}$$

Minimise $E[(y - \hat{y})^2]$ avec $\hat{y} = \tilde{a}x + \tilde{b}$

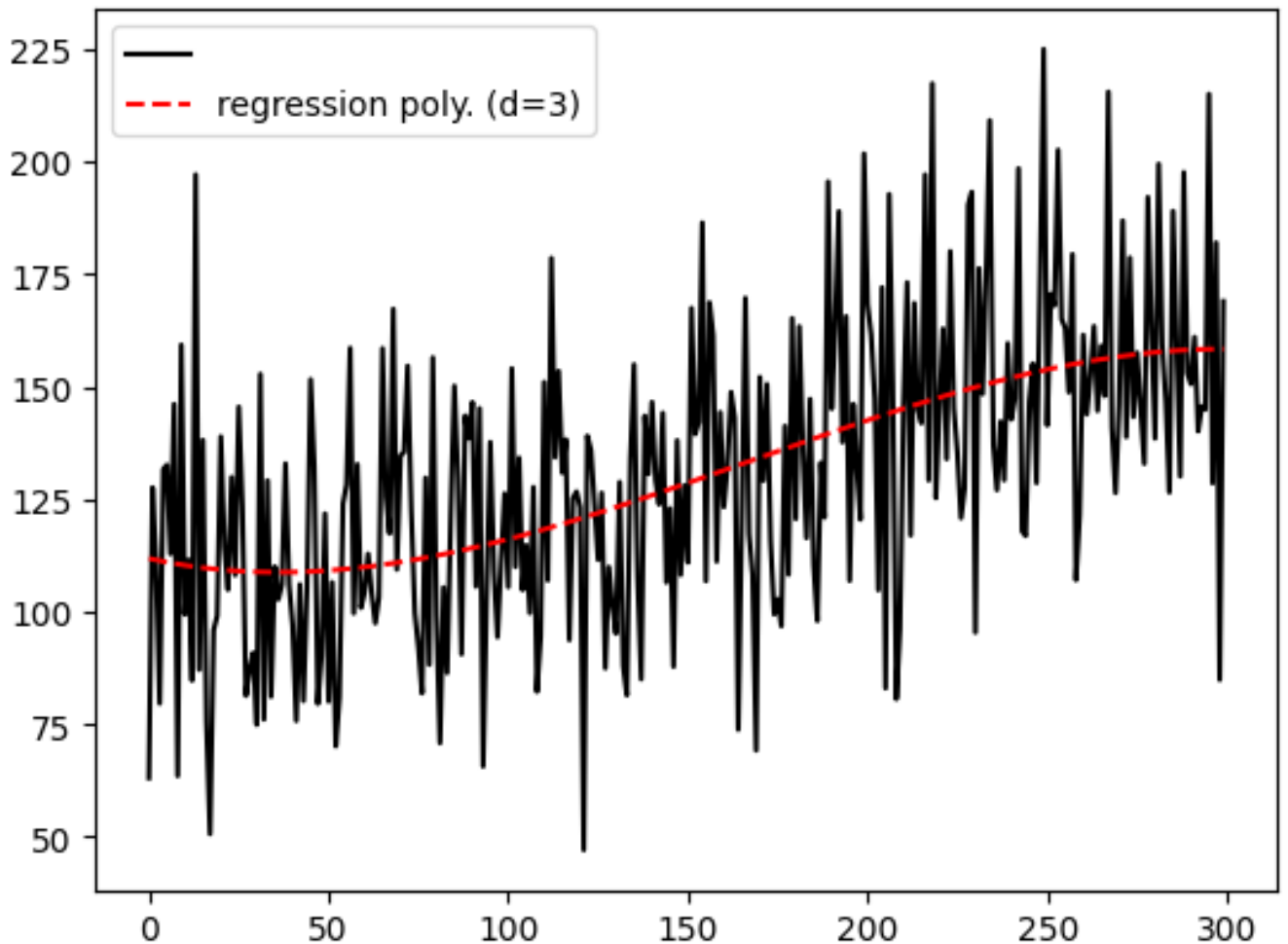
Exemple de droite de régression linéaire



Régression polynomiale (généralisation)

Modèle polynomial de degré d de la forme :

$$\hat{y} = \sum_{i=0}^{d} a_i x^i = a_d x^d + a_{d-1} x^{d-1} + \ldots + a_1 x + a_0 \tag{10}$$



Classification Ascendante Hiérarchique

Objectif

Pour une métrique donnée, regrouper itérativement deux à deux les individus les plus proches.

Algorithme

On choisit une fonction de distance entre 2 individus $d(x_i, x_j)$:

1. Calculer la matrice de distance $D \in \mathbb{R}+^{N \times N}$ entre tous les individus.
2. Regrouper les deux individus les plus proches dans le même agrégat (choisir un nouveau représentant pour chaque agrégat).
3. Itérer [2] jusqu'à obtenir le nombre d'agrégats souhaité.

Exemples de distance :

- Distance Manhattan : $d(x, y) = \sum_{i=1}^{N} |x_i - y_i|$
- Distance euclidienne : $d(x, y) = \sqrt{\sum_{i=1}^{N} |x_i - y_i|^2}$
- Similarité cosinus : $d_{cos}(x, y) = \frac{x \cdot y^T}{||x||||y||}$ avec $||x|| = \sqrt{\sum_i x_i^2}$

Dendrogramme (exemple)

Données :

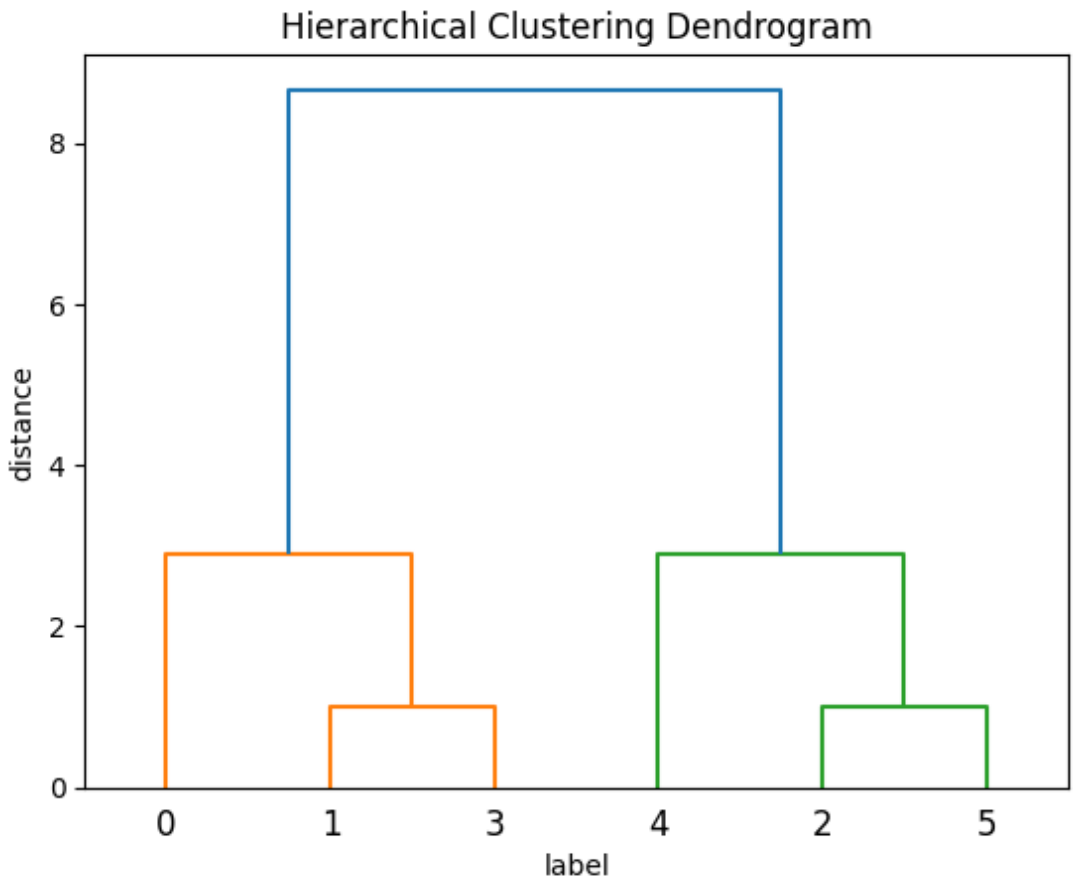
label	X
0	2
1	4
2	10
3	5
4	7
5	9

Distance euclidienne :

$$d(x,y) = \sqrt{(x-y)^2}$$

Matrice de distances :

$$D = \begin{pmatrix} 0 & 2 & 8 & 3 & 5 & 7 \\ 2 & 0 & 6 & 1 & 3 & 5 \\ 8 & 6 & 0 & 5 & 3 & 1 \\ 3 & 1 & 5 & 0 & 2 & 4 \\ 5 & 3 & 3 & 2 & 0 & 2 \\ 7 & 5 & 1 & 4 & 2 & 0 \end{pmatrix}$$



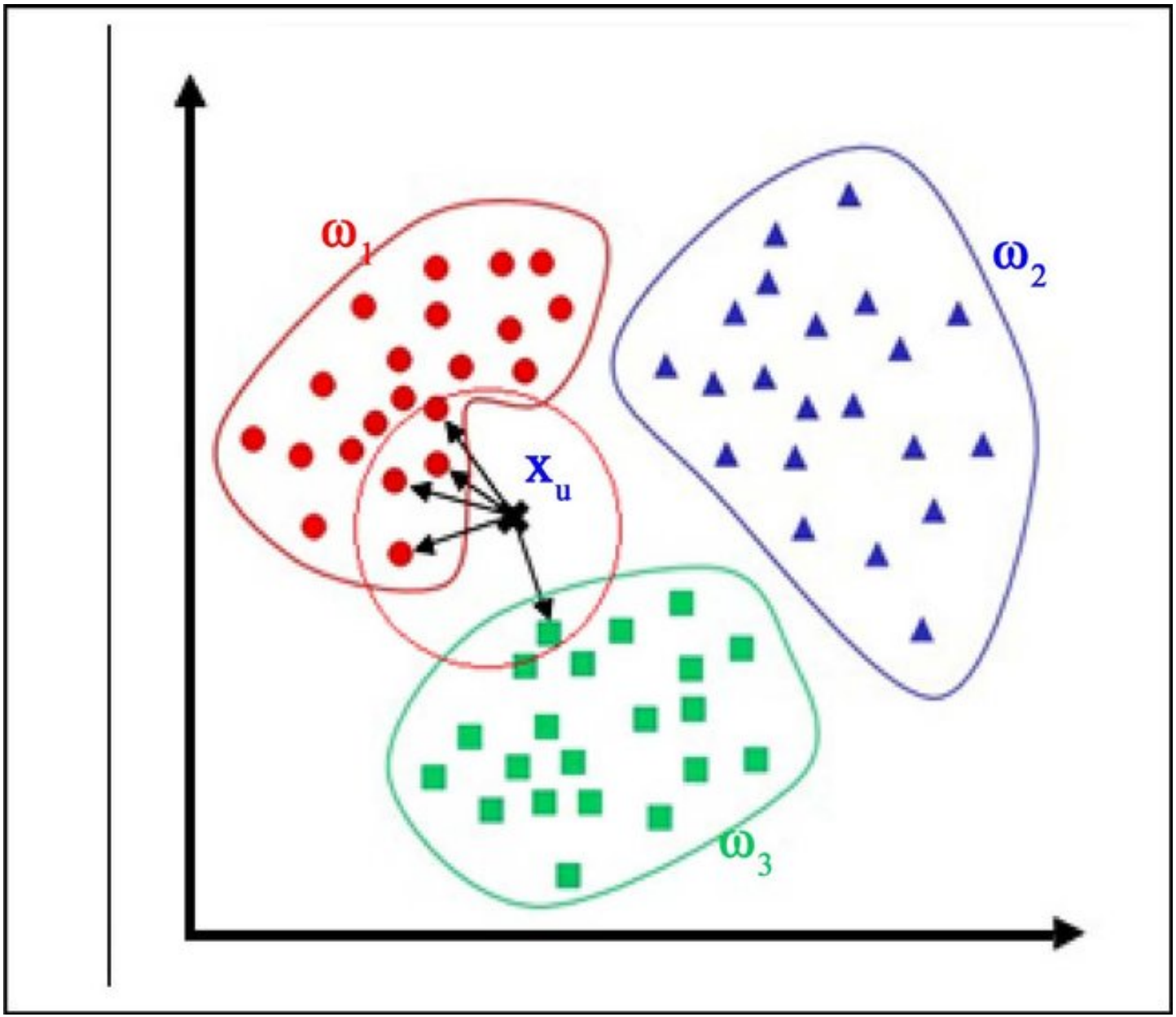
K plus proches voisins (KNN)

On dispose de paires annotées (x^{train}, y^{train}).
On choisit un nombre entier $K \geq 1$ (impair de préférence) et une fonction de distance d

Méthode

1. Pour un individu x_i quelconque, choisir les K individus x_j^{train} $(j \neq i)$ les plus proches.
2. Prédiction par vote majoritaire (ie. le y le plus représenté dans le K-voisinage de x) :

Illustration de l'algorithme KNN



Classification naïve bayésienne

Théorème de Bayes

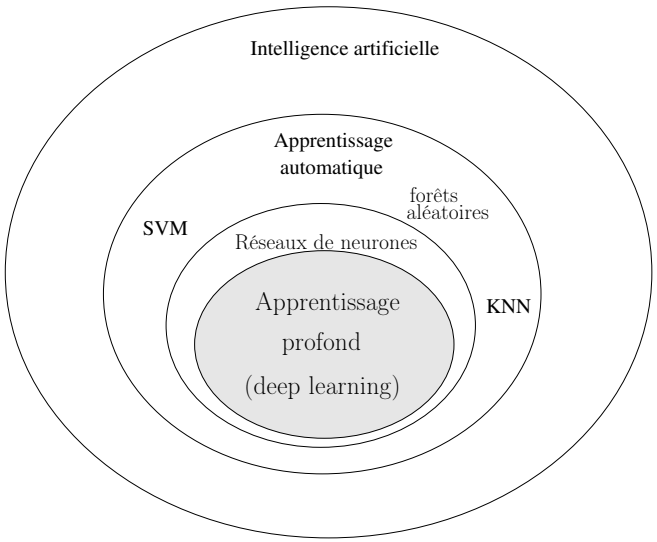
$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} \tag{11}$$

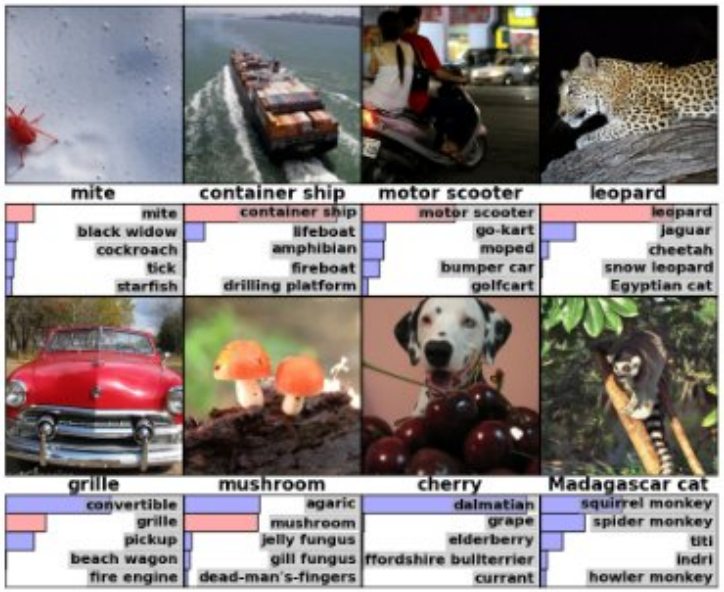
- $P(x)$: a priori sur les données (loi empirique)
- $P(y)$: a priori sur les labels (loi uniforme)
- $P(x|y)$: loi a priori sur les données d'une classe (loi empirique des x_i de la classe concernée)

Inférence bayésienne :

- Maximum a posteriori (MAP) : $\hat{y} = \arg\max_y P(y|x)$
- Espérance : $\hat{y} = E_{y|x}(y) = \sum_i y_i P(y_i|x))$

Le succès du *deep learning* depuis 2012





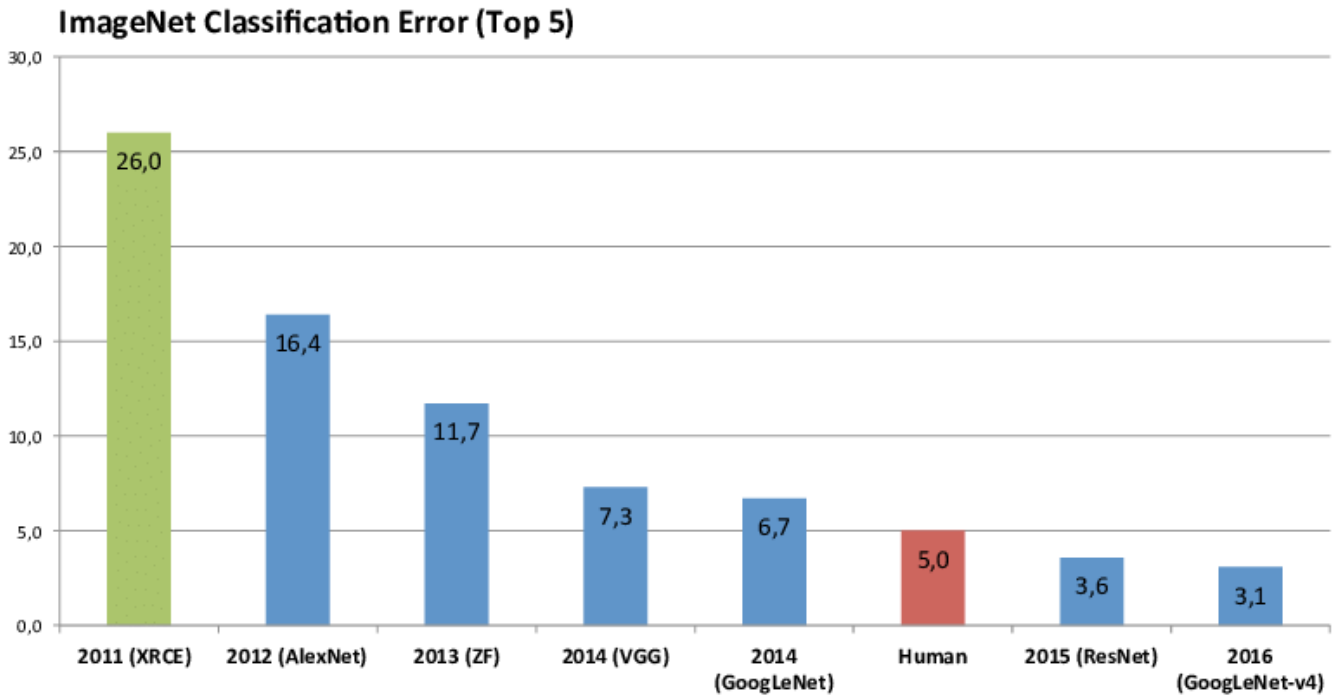
`http://image-net.org/challenges/LSVRC/`

$\approx$ 1,2 millions d'images, 1000 classes [ILSVRC2012] [Krizhevsky et al. NIPS'12]

- Excellentes performances
- Grande masse de données annotées
- Parallélisation des algorithmes
- Moyens de calculs accessibles (e.g. GPU, clusters, etc.)

	Affiliation	Taux d'erreur	Description
1	U. Toronto	15,31 % (-10,8%)	**Deep CNN**
2	U. Tokyo	26,172 %	**desc. + classif.**
3	U. Oxford	26,979 %	**desc. + classif.**
4	Xerox/INRIA	27,058 %	**desc. + classif.**

Avantages du *deep learning*



- État de l'art en vision par ordinateur [Azizpour et al., 2016]
- Forte capacité de généralisation (sans sur-apprentissage)
- Apprend automatiquement une représentation discriminante (*deep features*)
- Surpasse l'humain dans de plus en plus de domaines (e.g. reconnaissance d'images, jeu d'échec, Alphago, Alphastar, etc.)
- Une infinité de domaines d'application (e.g. biomédical, finance, astronomie, reconnaissance vocale, etc.)

Exemple 1 : Reconnaissance d'objets

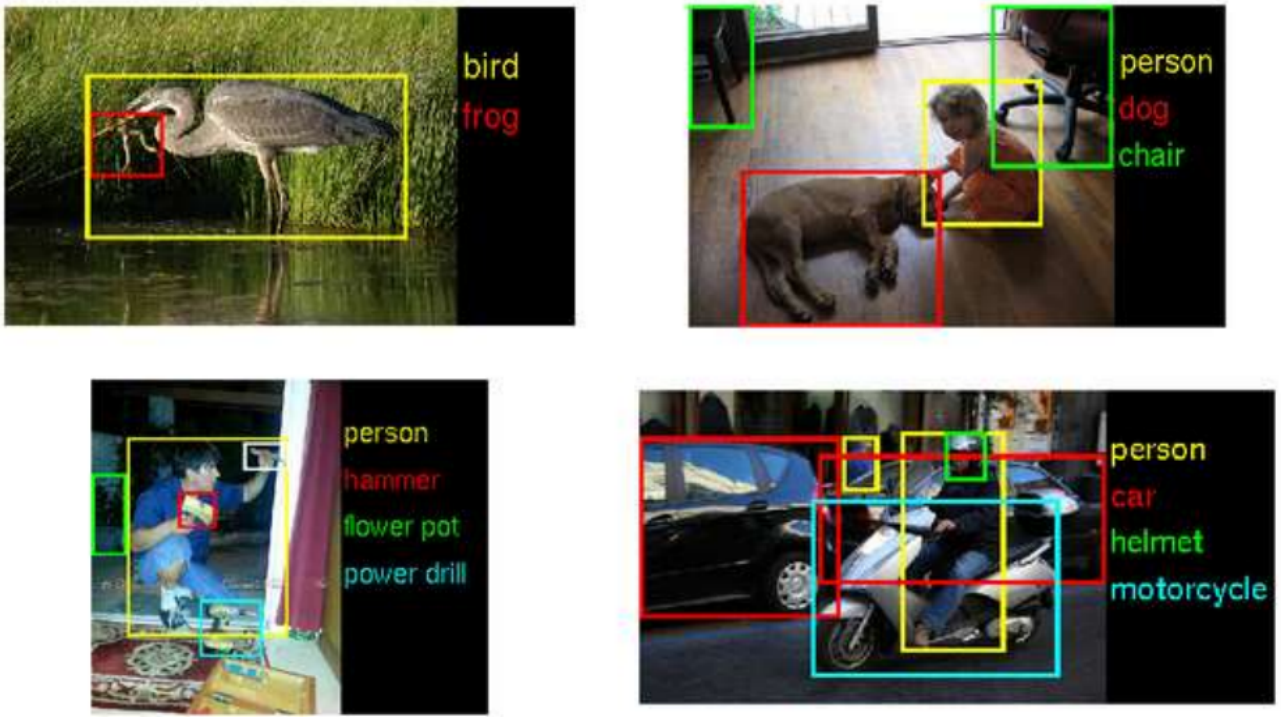
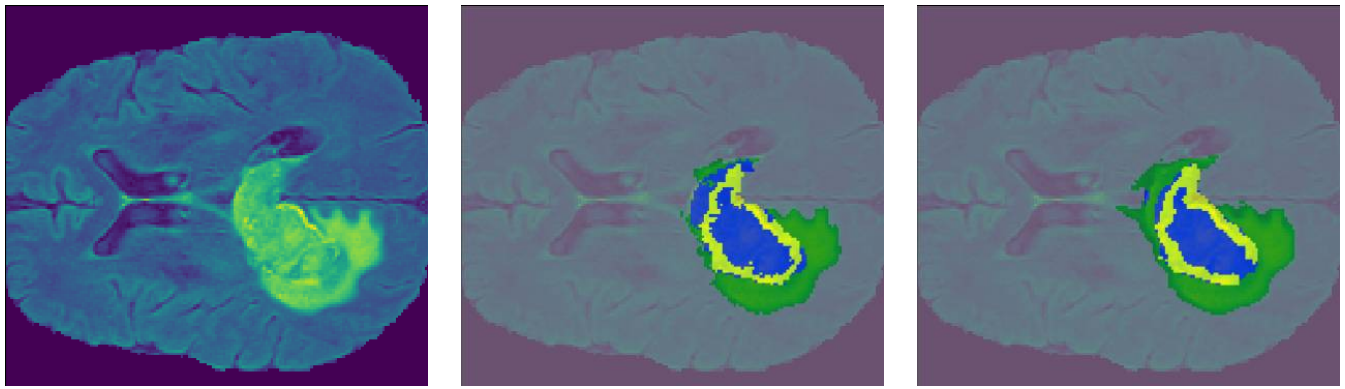


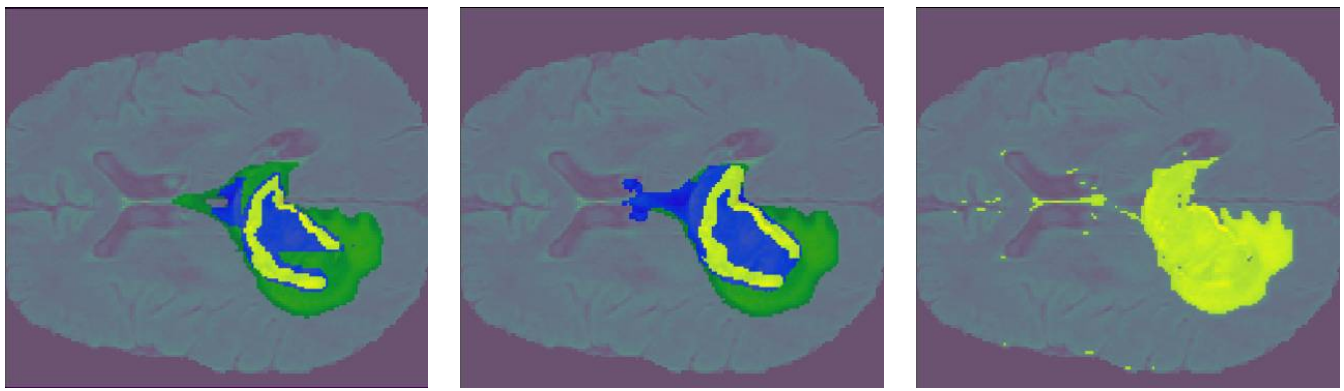
Figure : Taux d'erreur : Humain : $\approx 5\%$, RNP [7] : $\leq 4\%$

7. RNP : Réseaux de neurones profonds

Exemple 2 : Segmentation d'IRM



(a) Entrée (FLAIR) (b) Vérité terrain (c) 2D U-Net+CL

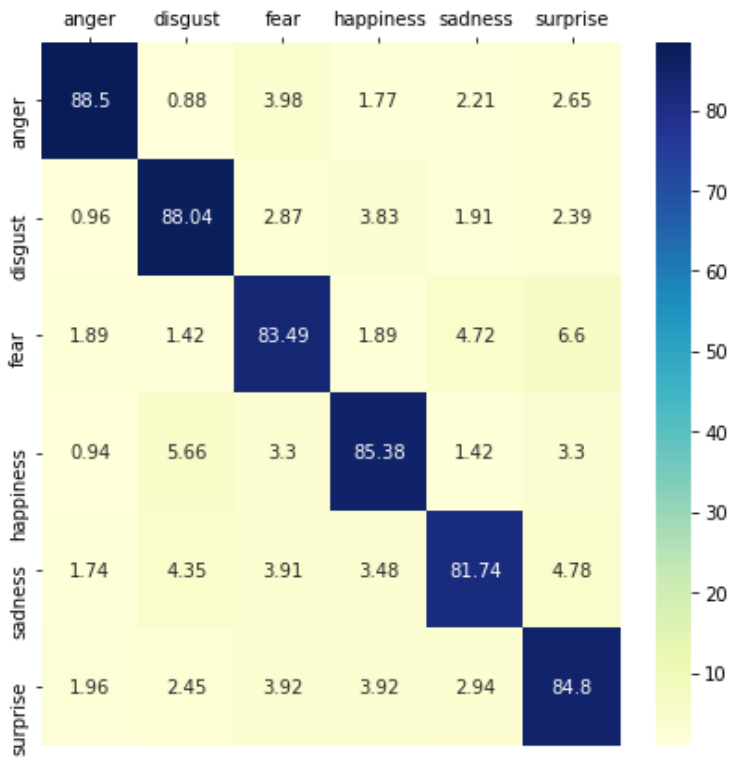


(d) 3D U-Net (e) Cascaded Network (f) K-Means

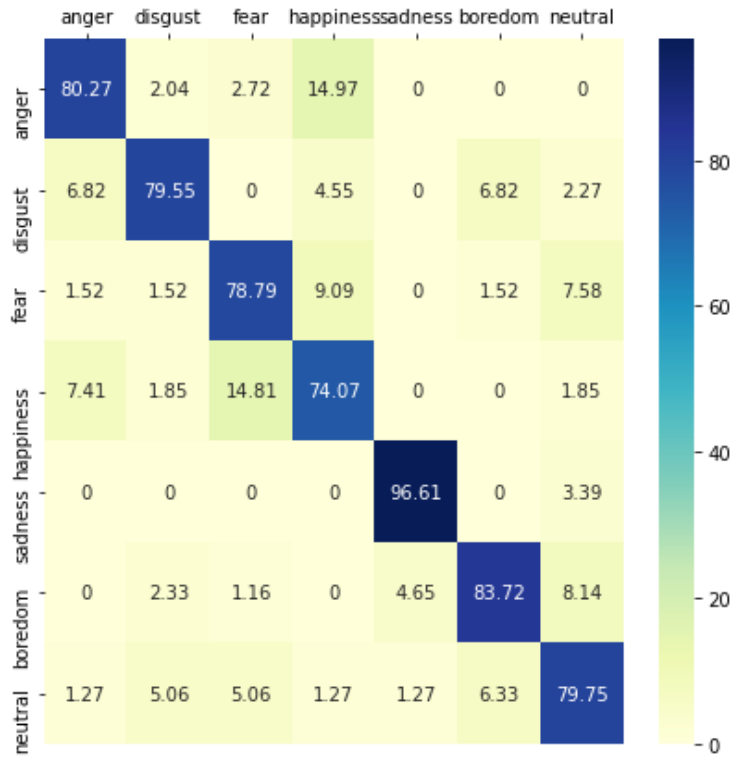
8

8. Ikram Brahim, Dominique Fourer, Vincent Vigneron and Hichem Maaref. Deep Learning Methods for MRI Brain Tumor Segmentation : a comparative study. Proc. IEEE IPTA 2019. Istanbul, Turkey

Exemple 3 : Reconnaissance des émotions dans la voix



(g) eNTERFACE05, STFT-Alex+RCS41 (Acc. 85.33%)



(h) EMO-DB, STFT-Alex+RCS19 (Acc. 81.82%)

9

9. Xia, S., Fourer, et al. (2022, May). Speech Emotion Recognition using Time-frequency Random Circular Shift and Deep Neural Networks. In Speech Prosody 2022.

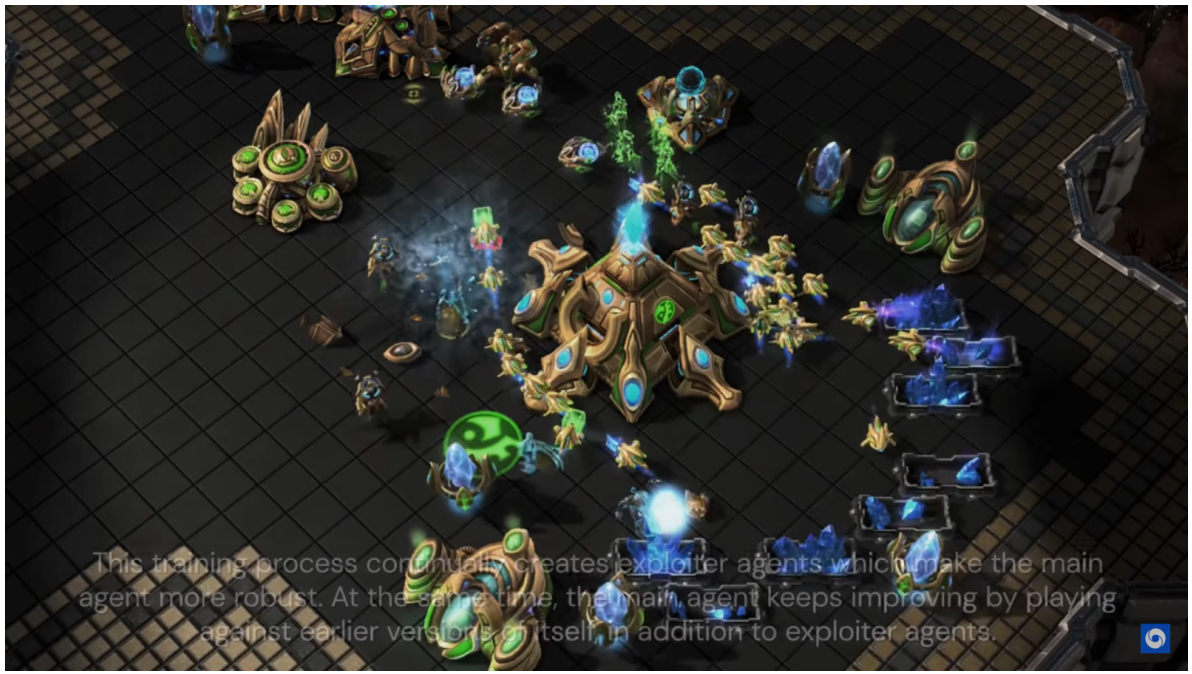
Exemple 4 : Transfert de style

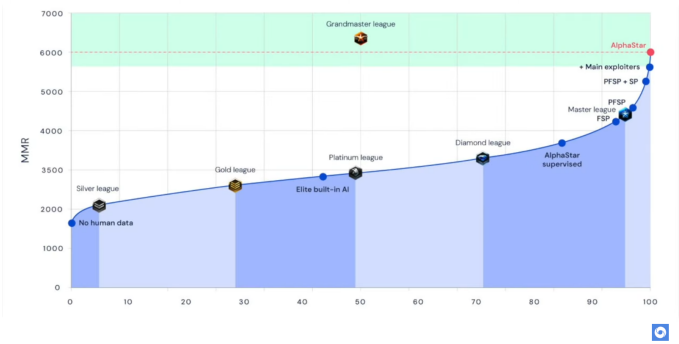
https://colab.research.google.com/github/tensorflow/hub/blob/master/examples/colab/tf2_arbitrary_image_stylization.ipynb

Exemple 5 : Apprentissage par renforcement (alphago)

https://www.deepmind.com

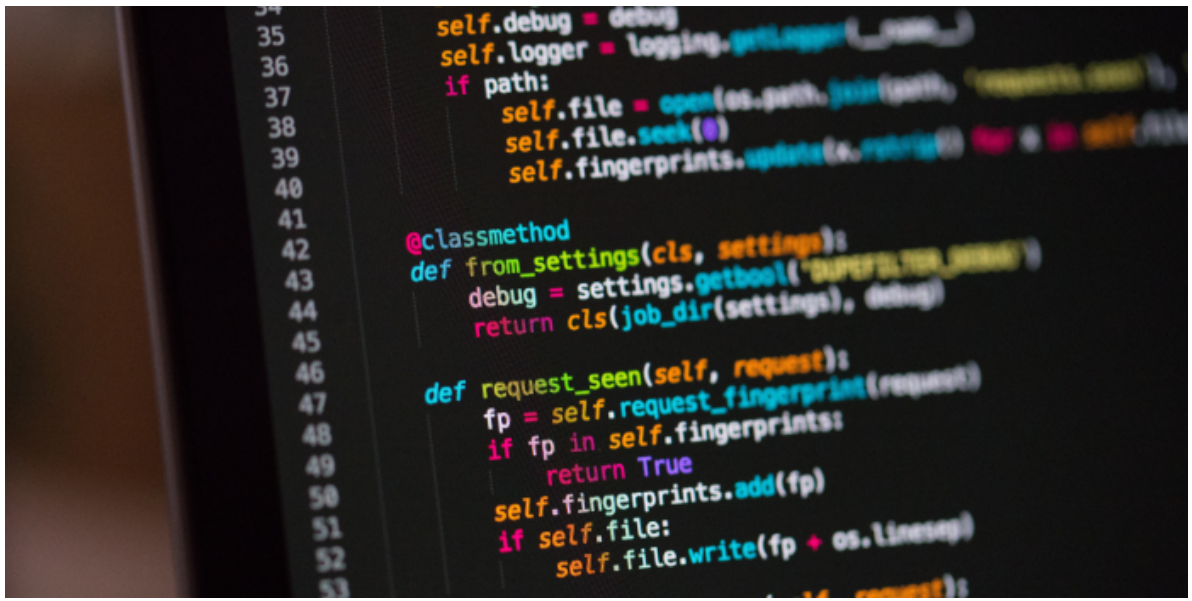
Exemple 6 : Apprentissage par renforcement (alphastar)





https://www.deepmind.com/open-source/alphastar-resources

Exemple 7 : Génération automatique de code



https://alphacode.deepmind.com/

etc.

De nombreuses autres applications...
Actu IA : `https://www.actuia.com/`
Data IA : `https://www.dataia.eu/`
Paper with code : `https://paperswithcode.com/`
axe IA (IBISC) : `https://ia.fourer.fr`