Méthodes statistiques et data mining

Giorgio Pauletto

pauletto[at]stanfordalumni[dot]org

Haute École de Gestion
19 février 2009

Plan

- Introduction, motivation, statistiques descriptives
  18:15–20:00
- Pause
  20:00–20:20
- Aperçu de méthodes statistiques: classification, clustering, régression
  20:20–21:30

Bibliographie

- **Hand D., Mannila H., Smyth P. (2001) *Principles of Data Mining*. MIT Press.**
- **Tan P., Steinbach M., Kumar V. (2005) *Introduction to Data Mining*. Addison-Wesley.**
- *Wikipedia*
- Tukey, J.W. (1977) *Exploratory Data Analysis*. Addison-Wesley.
- Berry M.J.A., Linoff G.S. (2004) *Data Mining Techniques*. Wiley.
- Berthold M., Hand D. (eds) (2003) *Intelligent Data Analysis*. Springer.
- Hastie T., Tibshirani R., Friedman J. (2001) *The Elements of Statistical Learning*. Springer.

Software

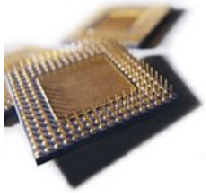
- R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
  http://www.r-project.org/
- RapidMiner (formerly YALE). Includes Weka operators.
  http://rapid-i.com/
- Weka Machine Learning Project. *Weka: Data Mining Software in Java*. The University of Waikato, New Zealand.
  http://www.cs.waikato.ac.nz/~ml/

Free/Open Source software, Multi-platform (Linux, Windows, Mac)

Partie 1: Introduction, motivation

Motivation

- Croissance de la capacité de stockage



- Croissance de la puissance de calcul

- Croissance de la bande passante

- Amélioration des algorithmes et de la recherche

- Besoins accrus d'avoir une vue globale synthétique en temps réel
- Demandes des entreprises, des gouvernements et bientôt aussi des individus

Pourquoi utiliser le data mining?

- Croissance des bases de données

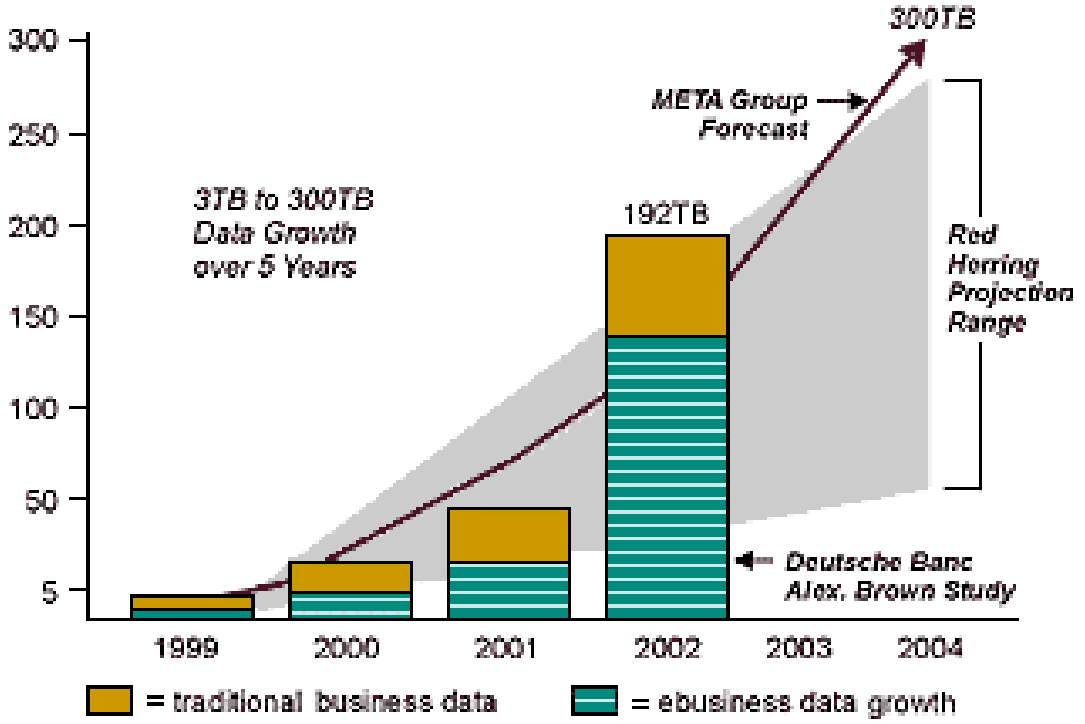


Figure 1: Data Growth Projections

Ordres de grandeur

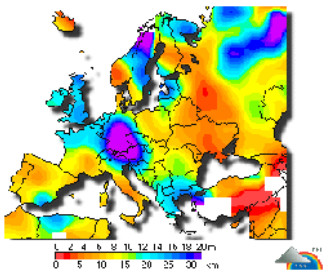
- Kilobyte = 1'000 bytes, $O(10^3)$
  - 2 KB = une page de texte dactylographié
  - 100 KB = une image basse résolution

- Megabyte = 1'000'000 bytes, $O(10^6)$
  - 1 MB = une disquette 3.5 pouces; un roman de 500 pages
  - 2 MB = une photo haute résolution 1000 × 1000 8 bits couleur
  - 5 MB = oeuvres complètes de Shakespeare; une chanson MP3 128 kb/s
  - 100 MB = 1 mètre de livres posés sur une étagère
  - 800 MB = un CD-ROM

http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/index.htm

Ordres de grandeur

- Gigabyte = 1'000'000'000 bytes, $O(10^9)$
  - 2 GB ≈ une camionnette remplie de livres
  - 5 GB ≈ un DVD simple couche
  - 20 GB ≈ une collection des oeuvres de Beethoven, DVD Blue Ray
  - 100 GB ≈ un étage de bibliothèque de journaux scientifiques
  - 200 - 500 GB ≈ un disque dur
- Terabyte = 1'000'000'000'000 bytes, $O(10^{12})$
  - 1 TB ≈ 50'000 arbres transformés en papier et imprimés
  - 2 TB ≈ une bibliothèque universitaire moyenne
  - 20 TB ≈ librairie du Congrès US
  - 400 TB ≈ ensemble données climatiques NCDC / NOAA



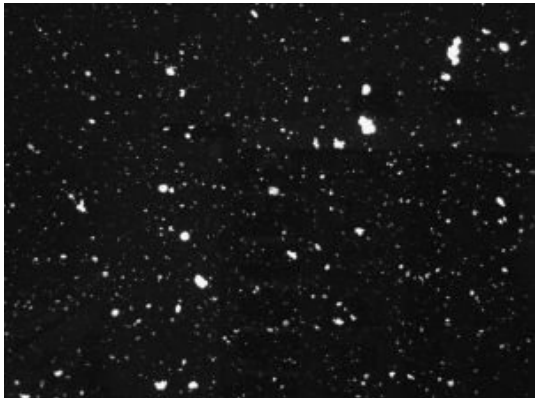
Ordres de grandeur

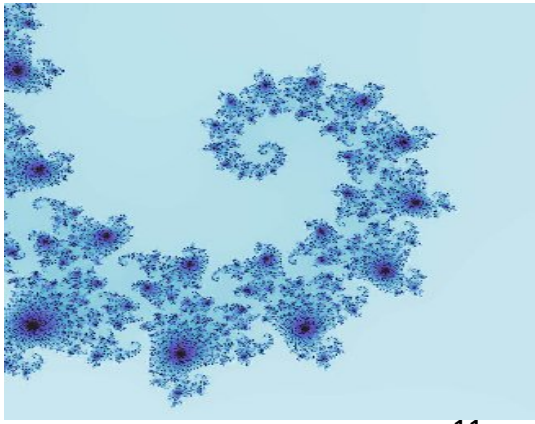
- Petabyte = 1'000'000'000'000'000 bytes, $O(10^{15})$
  - 1 PB ≈ 3 ans d'observations satellitaires de la Terre (EOS 2001)
  - 2 PB ≈ toutes le bibliothèques académiques américaines
  - 15 PB ≈ données par an issues du LHC du CERN
  - 20 PB ≈ production mondiale de disques durs 1995
  - 200 PB ≈ tous les documents imprimés dans le monde

- Exabyte = 1'000'000'000'000'000'000 b, $O(10^{18})$
  - 2 EB ≈ volume total de l'information mondialement générée en 1999
  - 5×10^{18} ≈ nombre total de mots prononcés par les êtres humains

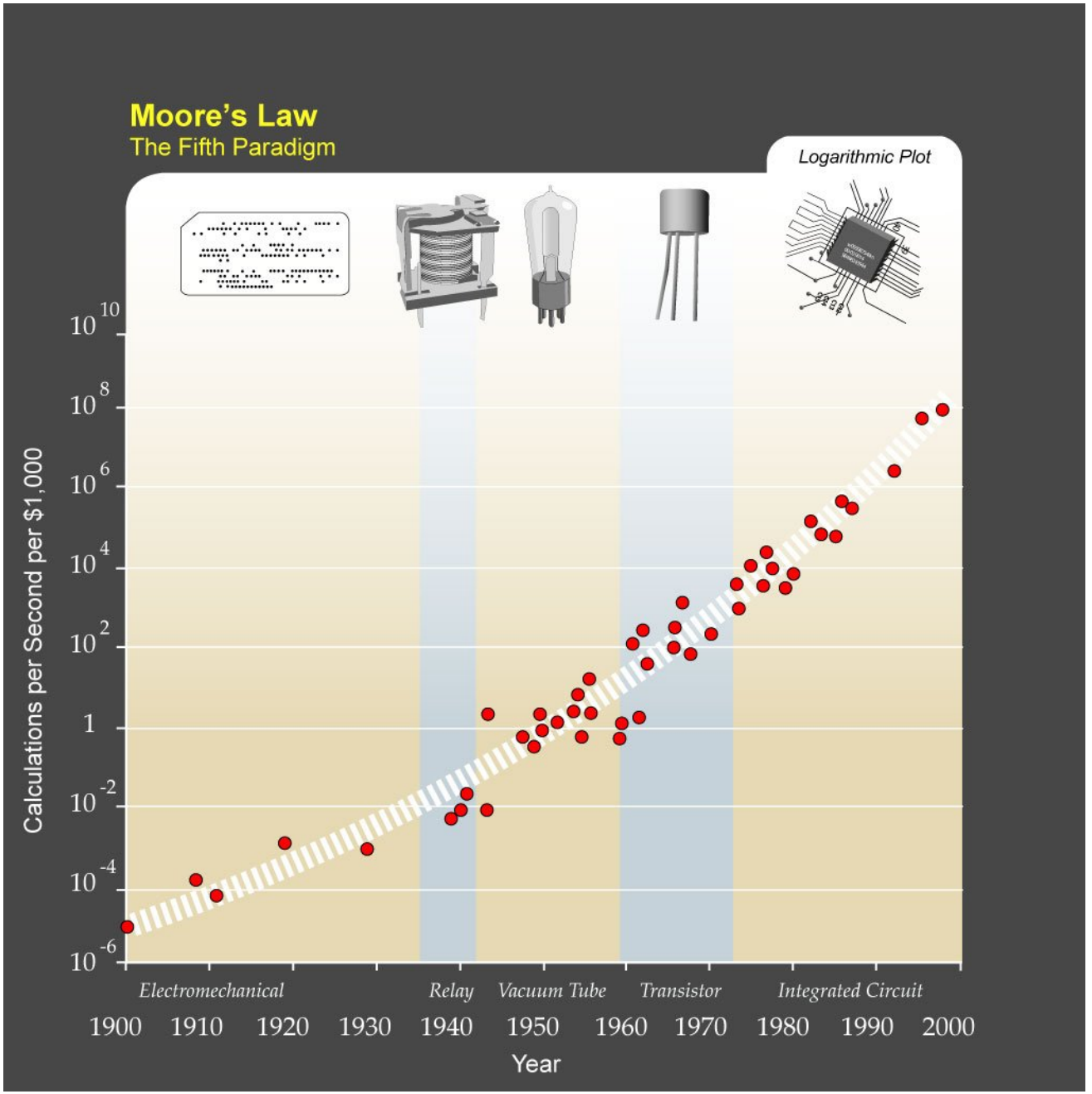
Ordres de grandeur

- Zettabyte = $O(10^{21})$
  - $70 \times 10^{21} \approx$ nombre d'étoiles de l'univers observable
  - $100 \times 10^{21} \approx$ nombre de grains de sable sur Terre
- Yottabyte $\approx O(10^{24})$





Puissance de calcul



Courtesy of Ray Kurzweil and Kurzweil Technologies, Inc.

http://en.wikipedia.org/wiki/Image:PPTMooresLawai.jpg

Une spirale de croissance

- **Loi de Moore:** le nombre de transistors sur un processeur double tous les 2 ans

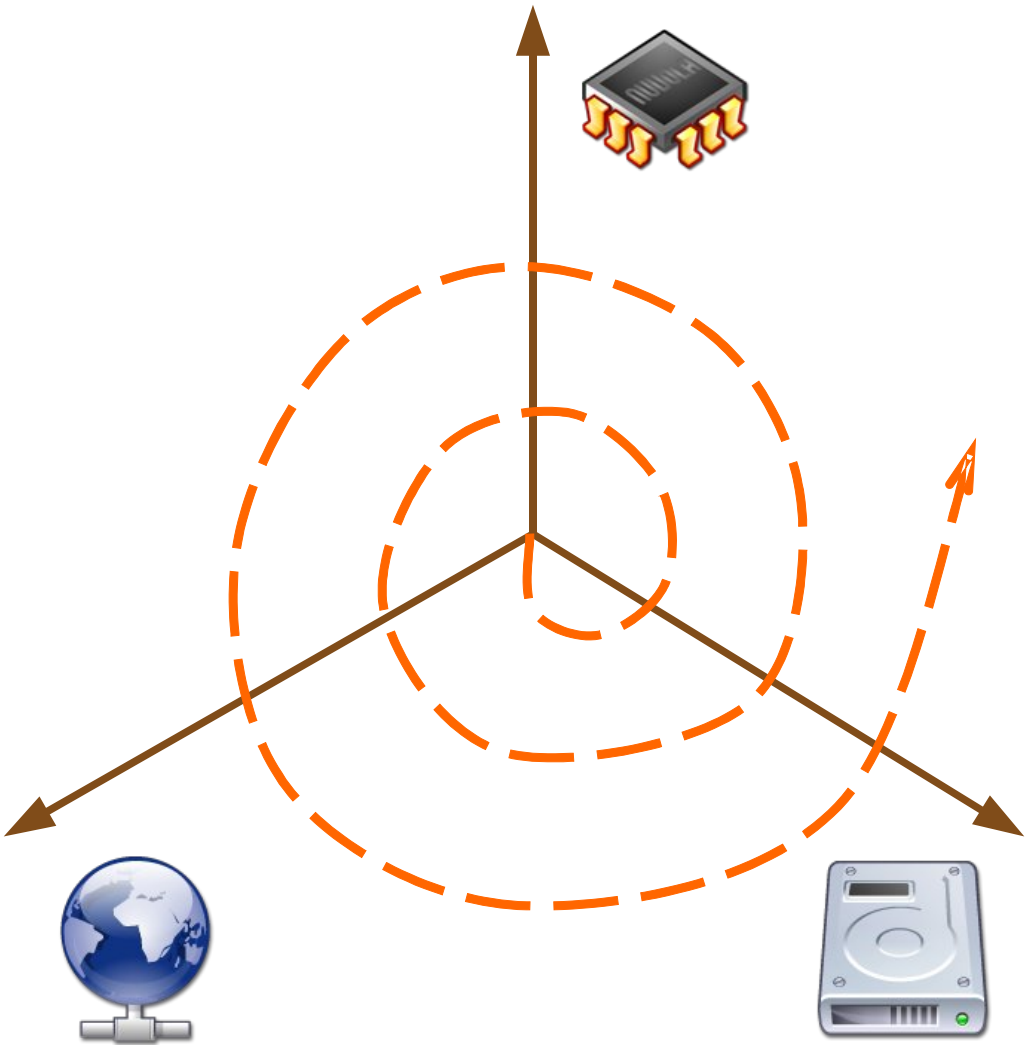
  Gordon Moore, ex CEO de Intel

- **Loi de Kryder:** la densité des supports magnétiques double chaque année

  Mark Kryder, ex CTO Seagate

- **Loi de Butter:** la bande passante des fibres optiques double tous les 9 mois

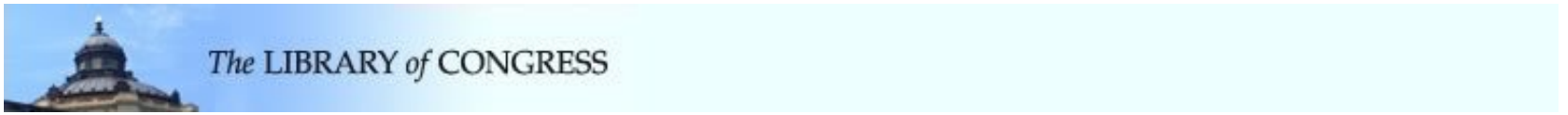
  Gerald Butter, ex dir Lab Opt Lucent



Interactif

- Parmi les bases de données des organisations suivantes, choisissez les trois les plus grandes ainsi que la moins grande
- Expliquez votre réponse
- Estimez la taille des bases de données
- Google, Librairie du congrès américain, Sprint, CIA, YouTube, ChoicePoint, AT&T, Centre mondial du climat, Internet Archive, Amazon
- *NB: Réponses estimées en 2007*

#10 Library of Congress



- 130 million items (books, photos, maps, etc)
- 29 million books
- 10,000 new items added each day
- 530 miles of shelves
- 5 million digital documents
- 20 terabytes of text data

Top 10 source
http://www.businessintelligencelowdown.com/2007/02/top_10_largest_.html

#9 Central Intelligence Agency

- 100 Freedom of Information Act items added each month
- Comprehensive statistics on more than 250 countries and entities
- Unknown number of classified information

#8 amazon.com

- 59 million active customers
- 250,000 full text books
- Users comments, forums, wishlists, etc.
- Food, clothing, electronics, home furniture, toys, tools, ...
- More than 42 terabytes of data

#7 YouTube

Broadcast Yourself™

- 100 million videos watched per day
- 65,000 videos added each day
- 60% of all videos watched online
- At least 45 terabytes of video
- Recently bought by Google

#6 ChoicePoint

- 250 terabytes of personal data
- Information on 250 million people mainly US
- Addresses, phone numbers, driving records, criminal histories, etc.
- Data sold to the highest bidders
- Able to identify 9/11 victims by matching DNA in bone fragments to information provided by victim's family members in conjunction to data found in their databases

#5 

- One of the world's largest telecommunication companies as it offers mobile services to more than 53 million subscribers
- 2.85 trillion database rows
- 365 million call detail records processed per day
- At peak, 70,000 call detail record insertions per second

#4 Google

- 91 million searches per day
- 50% of all internet searches
- 33 trillion database entries estimated
- Cache copies of pages, images, documents
- Virtual profiles of countless number of users
- Growing services: Gmail (2.5 GB/user), Blogger, Documents and spreadsheets, Ads, Calendar, Video, Maps, etc.

#3  at&t

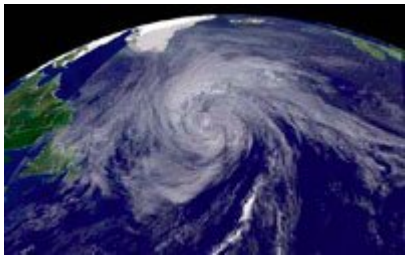
- Oldest US telecommunication company
- 323 terabytes of information
- 1.9 trillion phone call records
- Largest volume of data in one unique database and the second largest number of rows in a unique database
- If you ever made a call via AT&T less than 10 years ago, chances are they still got your information somewhere.

#2 NERSC
#2 Internet Archive

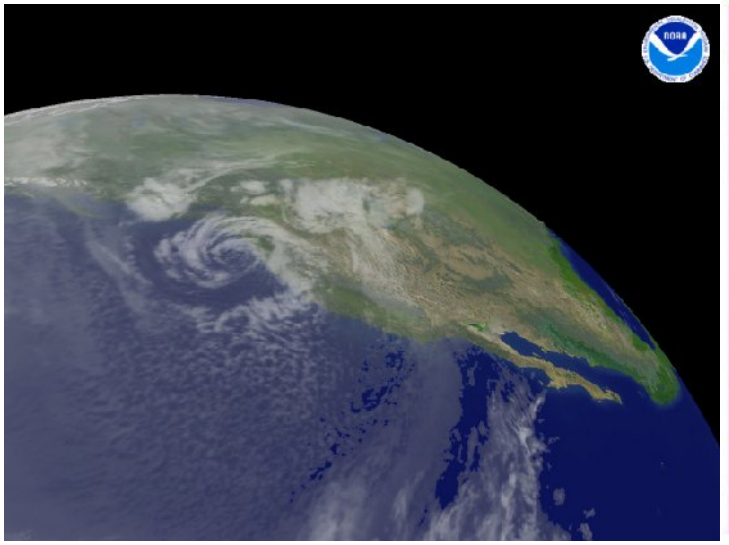
- National Energy Research Scientific Computing Center
- 2.8 petabytes of data
- Operated by 2,000 computational scientists

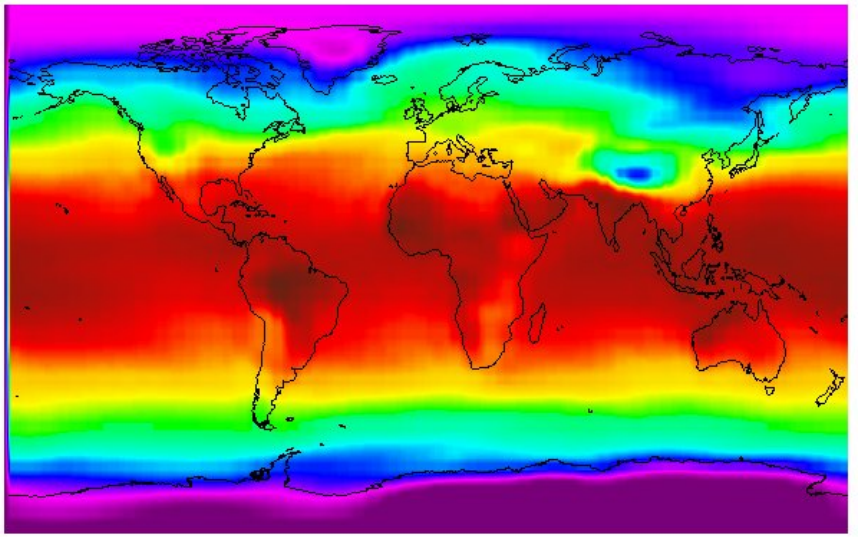
- Wayback machine
- 2 petabytes of data
- Growing at 20 terabytes per month

#1 World Data Centre for Climate



- 220 terabytes of web data
- 6 petabytes of additional data





Qu'est-ce que le data mining ?

- C'est un processus automatisé d'analyse exploratoire et de modélisation prédictive sur des grands ensembles de données.
- « Le data mining est l'analyse de grands ensembles de données observées afin de découvrir et de résumer des relations de façon nouvelle qui soient compréhensibles et utiles à leur propriétaire. »

Hand, Mannila, Smyth (2001)

Qu'est-ce que le data mining ?

- Intersection de plusieurs disciplines

Calcul scientifique

Reconnaissance de formes

Visualisation de données

Intelligence artificielle

Statistiques

Data mining

Apprentissage automatique

Analyse données

Gestion bases données

Business intelligence

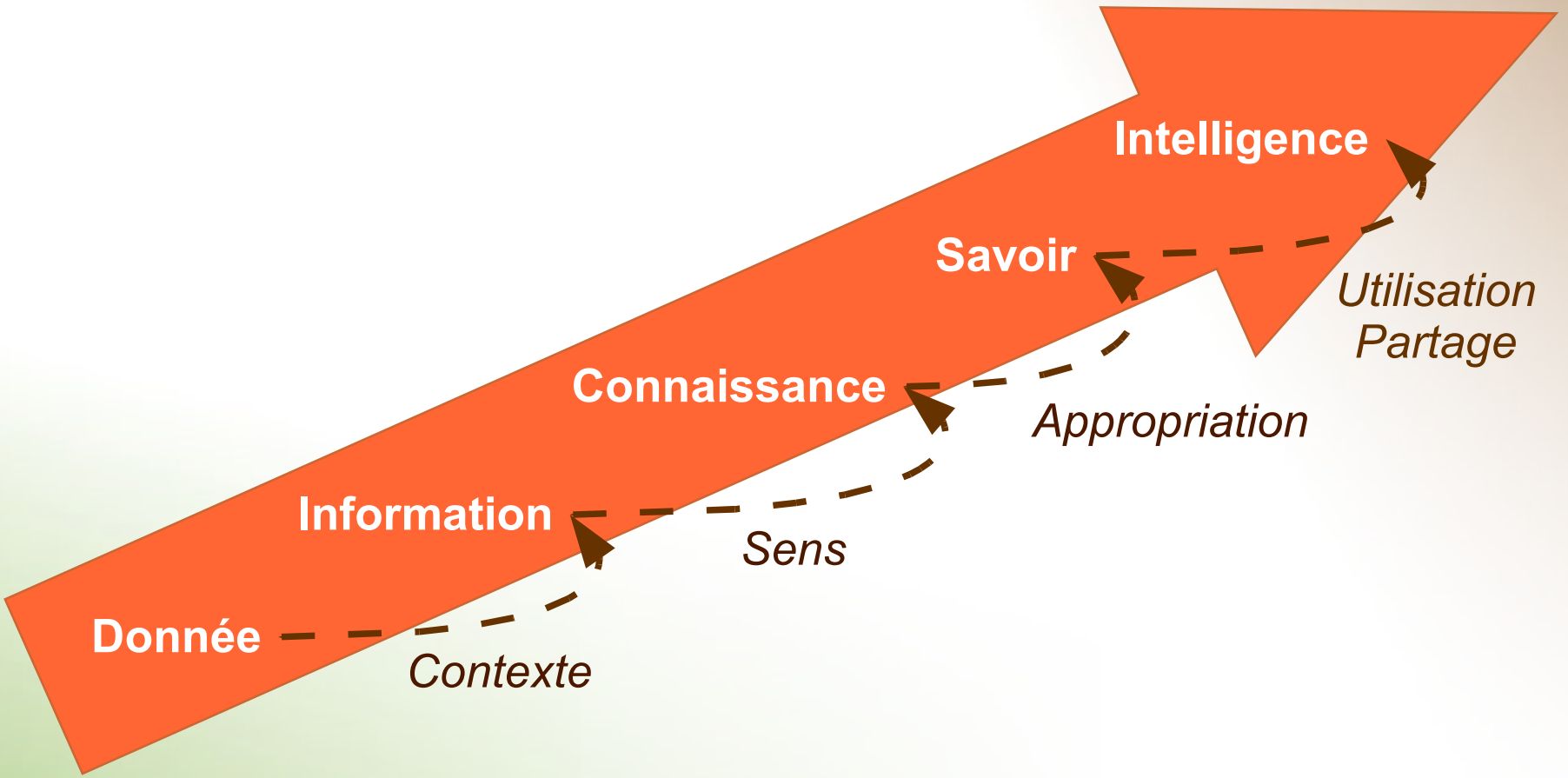
Gestion des systèmes d'information

Gestion des connaissances

Terminologie

- Data Mining
- Knowledge Discovery in Databases (KDD)
- Business Intelligence (BI)
- Fouille de données
- Extraction de connaissances
- Exploration de données
- ...

Humain



Machine

A quoi ça sert ?

- Gestion de la relation client (GRC)

  Customer Relationship Management (CRM)

  - Vision unifiée du client de l'entreprise pour mieux le servir
- Marketing direct, Vente croisée
  - A quels clients dois-je offrir un nouveau service ?
  - Quels autres produits puis-je offrir à mon client avec le plus de succès ?
- Rétention de clientèle (churn, attrition)
  - Quels clients sont prêts à fuir chez un concurrent ?

A quoi ça sert ?

- Détection de fraudes
  - Quels comportements d'achat par carte de crédit sont suspects ?
  - Peut-on repérer les cas de fraude fiscale ?
  - Quels messages emails sont des spams ?
  - Détection d'intrusions (ou de tentatives) sur un site Web ou un système informatique
- Scoring
  - Quelle est la tarification adaptée pour un certain type d'acheteur ?

A quoi ça sert ?

- Aide à la décision
  - Quels diagnostics médicaux sont les plus probables avec tels symptômes ?
- Fourniture de meilleurs services
  - Demandes de subsides pour les chômeurs qui sont les plus susceptibles d'avoir des difficultés à retrouver un emploi

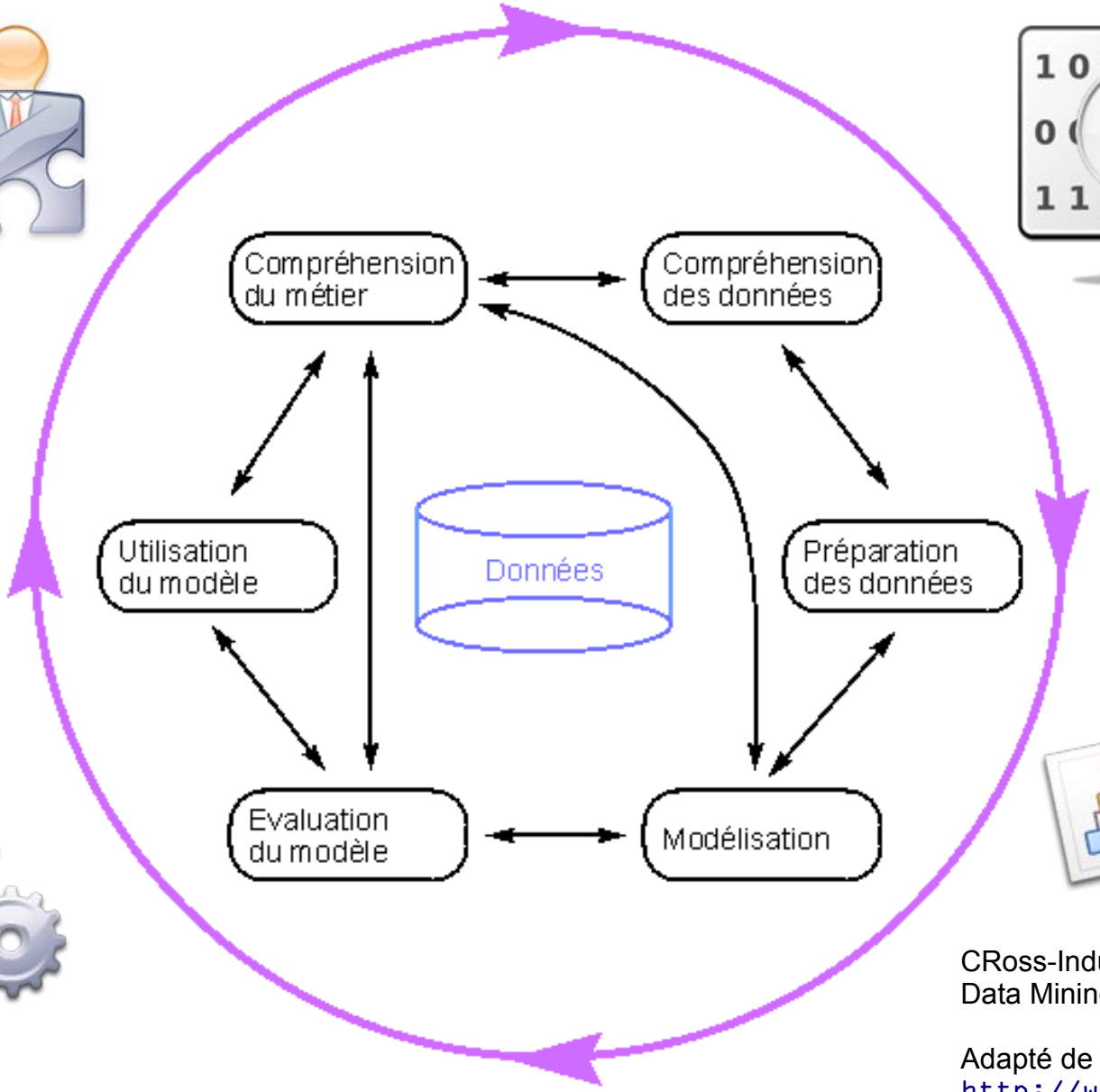
Aspects légaux

- La sphère privée et la personnalité sont protégées en Suisse
- Préposé fédéral à la protection des données et à la transparence
- Constitution Art. 13 et Code Civil Art. 28 : *toute personne a le droit d'être protégée contre l'emploi abusif des données qui la concernent*

Aspects légaux

- Loi sur la protection des données (LPD) et ordonnance relative
- Art. 3 a : *données personnelles (données), toutes les informations qui se rapportent à une personne identifiée ou identifiable*
- Pratiquement toutes les données personnelles peuvent être considérées comme sensibles
- Droit d'être renseigné sur ses propres données
- Aspects éthiques !

Processus de data mining



CRoss-Industry Standard Process for Data Mining

Adapté de CRISP-DM 1.0
http://www.crisp-dm.org/

Processus de data mining

Compréhension du métier

Cette première phase est essentielle et doit permettre de **comprendre les objectifs et les besoins métiers** afin de les intégrer dans la définition du projet DM et de décliner un plan permettant de les atteindre et les satisfaire.

Compréhension des données

Il s'agit de collecter et de **se familiariser avec les données à disposition**. Il faut également identifier le plus tôt possible les **problèmes de qualité** des données, développer les premières intuitions, détecter les premiers ensembles et hypothèses à analyser.

Préparation des données

Cette phase comprend toutes les étapes permettant de construire le jeu de données qui sera utilisé par le(s) modèle(s). Ces étapes sont souvent exécutées plusieurs fois, en fonction du modèle proposé et du retour des analyses déjà effectuées. Il s'agit entre autres d'**extraire, transformer, mettre en forme, nettoyer et de stocker de façon pertinente les données**. La préparation des données peut constituer environ **60 à 70% du travail total**.

Modélisation

C'est ici qu'entrent en jeu les méthodologies de modélisation issues notamment de la statistique. Les modèles sont souvent validés et construits avec l'**aide d'analystes du côté métier et d'experts en méthodes quantitatives**. Il y a dans la plupart des cas plusieurs façons de modéliser le même problème de DM et plusieurs techniques pour arriver à ajuster au mieux un modèle aux données. La boucle de feedback vers les points précédents est fréquemment utilisée pour améliorer le modèle.

Évaluation du modèle

Une fois arrivés à cette phase, un ou plusieurs modèles sont construits. Il faut s'assurer que les **résultats** sont jugés **satisfaisants** et sont **cohérents** notamment vis-à-vis des objectifs métier.

Utilisation du modèle

La mise au point du modèle n'est pas la fin du processus de DM. Une fois les **connaissances extraites des données**, elles doivent encore être organisées et présentées de façon à les rendre **utilisables par les destinataires**. Cela peut être aussi simple que de fournir une synthèse descriptive des données ou aussi complexe que de mettre en oeuvre un processus complet de fouille de données pour l'utilisateur métier final. Il est néanmoins toujours **important que l'utilisateur comprenne les limites des données et de l'analyse pour que ses interprétations et ses décisions soient judicieuses**.

Questions ?

Statistiques descriptives

- D'après John W. Tukey on passe constamment entre deux approches en statistiques:
  les statistiques *exploratoires* et
  les statistiques *confirmatoires*

« *Far better an approximate answer to the right question, which is often vague, than the exact answer to the wrong question, which can always be made precise.* »

Données

- Les données sont un ensemble de mesures fournies par un processus qui les génère
- En général, on observe *n* individus sur un certain nombre de variables *p*
- On range souvent ces données sous forme d'un tableau avec *n* lignes et *p* colonnes

Variable, attribut

Donnée suspecte

ID	Age	Sexe	Etat civil	Education	Revenu
248	54	M	Marié	Post oblig	100000
249	??	F	Marié	Post oblig	12000
250	29	M	Marié	Université	23000
251	9	M	Célibataire	Scolarisé	0
252	85	F	Célibataire	Post oblig	19798
253	40	M	Marié	Post oblig	40100
254	38	F	Célibataire	Obligatoire	2691
255	7	M	??	Scolarisé	0
256	49	M	Marié	Post oblig	30000
257	76	M	Marié	Université	30686

Individu, observation, objet

Valeur manquante

Nature des données

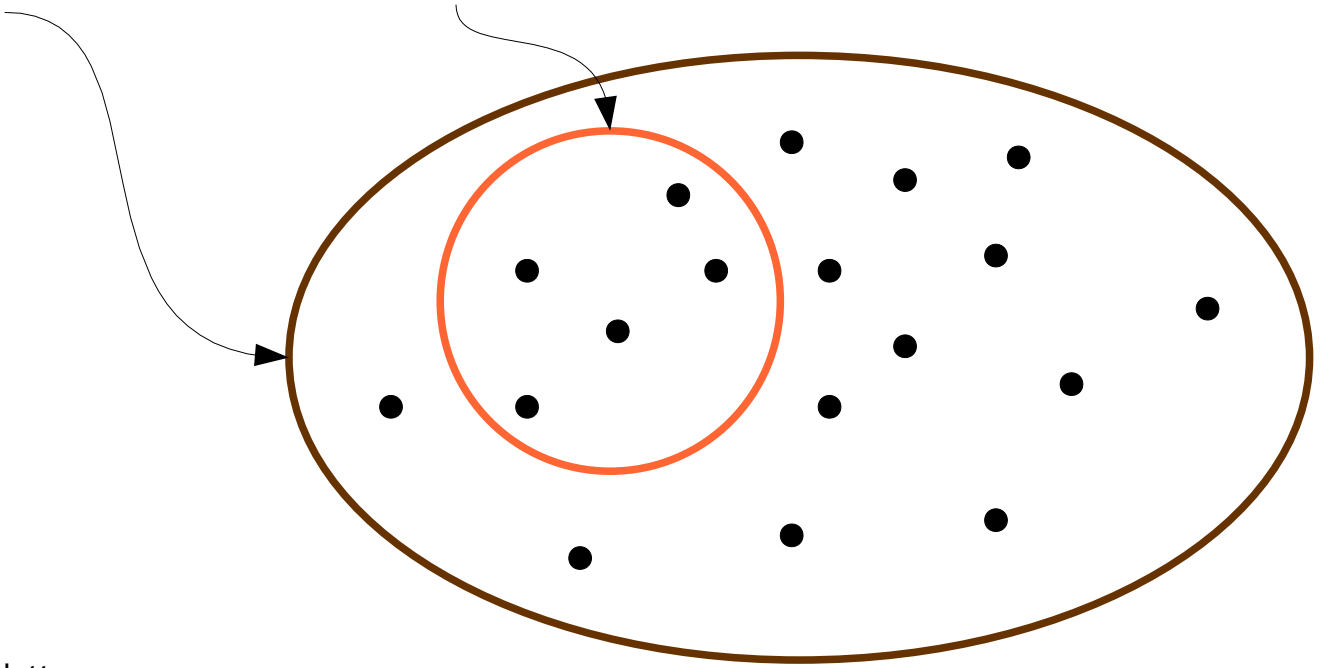
- Variables quantitatives
  - Valeurs numériques et sommables, discrètes ou continues
    Ex: 18, 7654.43, -0.762, 0, 9999
- Variables qualitatives
  - Ordinales
    Ex: petit, moyen, grand, très grand
  - Nominales (catégories ou modalités)
    Ex: féminin, masculin
    célibataire, marié, divorcé, veuf

Nature des données

- Textes
  - Corpus documentaires, bases de connaissances, sites web (blogs, forums), etc.
  - Ex: PageRank de Google
- Transactions
  - Liste d'achats, visites de sites web, mouvements de fonds, etc.
  - Ex: Amazon, cartes de fidélité
- Multimédia: images, sons, vidéos

Nature des données

- Données expérimentales vs données observées
  - Contrôle, plans d'expériences
  - Données non reproductibles
- Population vs échantillon



Statistiques descriptives

- Caractéristiques de tendance centrale
  - Moyenne: somme des valeurs divisée par leur nombre $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - Médiane: valeur qui partage l'effectif en deux
  - (Mode: valeur la plus fréquente, utile pour les données nominales)
- Caractéristiques de dispersion
  - Variance et écart-type:

$$s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad s = \sqrt{s^2}$$

Exemple

- Données: 2, 3, 4, 5, 2
- Moyenne: $(2 + 3 + 4 + 5 + 2) / 5 = 16/5 = 3.2$
- Médiane: ordre croissant 2, 2, 3, 4, 5
- Mode: valeur la plus fréquente 2
- Variance:
  $((2 - 3.2)^2 + (3 - 3.2)^2 + (4 - 3.2)^2 + (5 - 3.2)^2 + (2 - 3.2)^2)/(5-1) =$
  $((-1.2)^2 + (-0.2)^2 + (0.8)^2 + (1.8)^2 + (-1.2)^2) / 4 =$
  $(1.44 + 0.04 + 0.64 + 3.24 + 1.44) / 4 = 6.8 / 4 = 1.7$
- Écart-type: $\sqrt{1.7} = 1.3038$

Exemple

- Pour pouvoir comparer des variables qui sont dans des échelles différentes et qui ont des moyennes différentes on standardise les variables.
  (2 3 4 5 2)
- Centrer: Enlever la moyenne
  (-1.2 -0.2 0.8 1.8 -1.2)

$$x_c = x - \bar{x}$$

- Réduire: Diviser par l'écart-type
  (-0.920358 -0.153393 0.613572 1.380537 -0.920358)

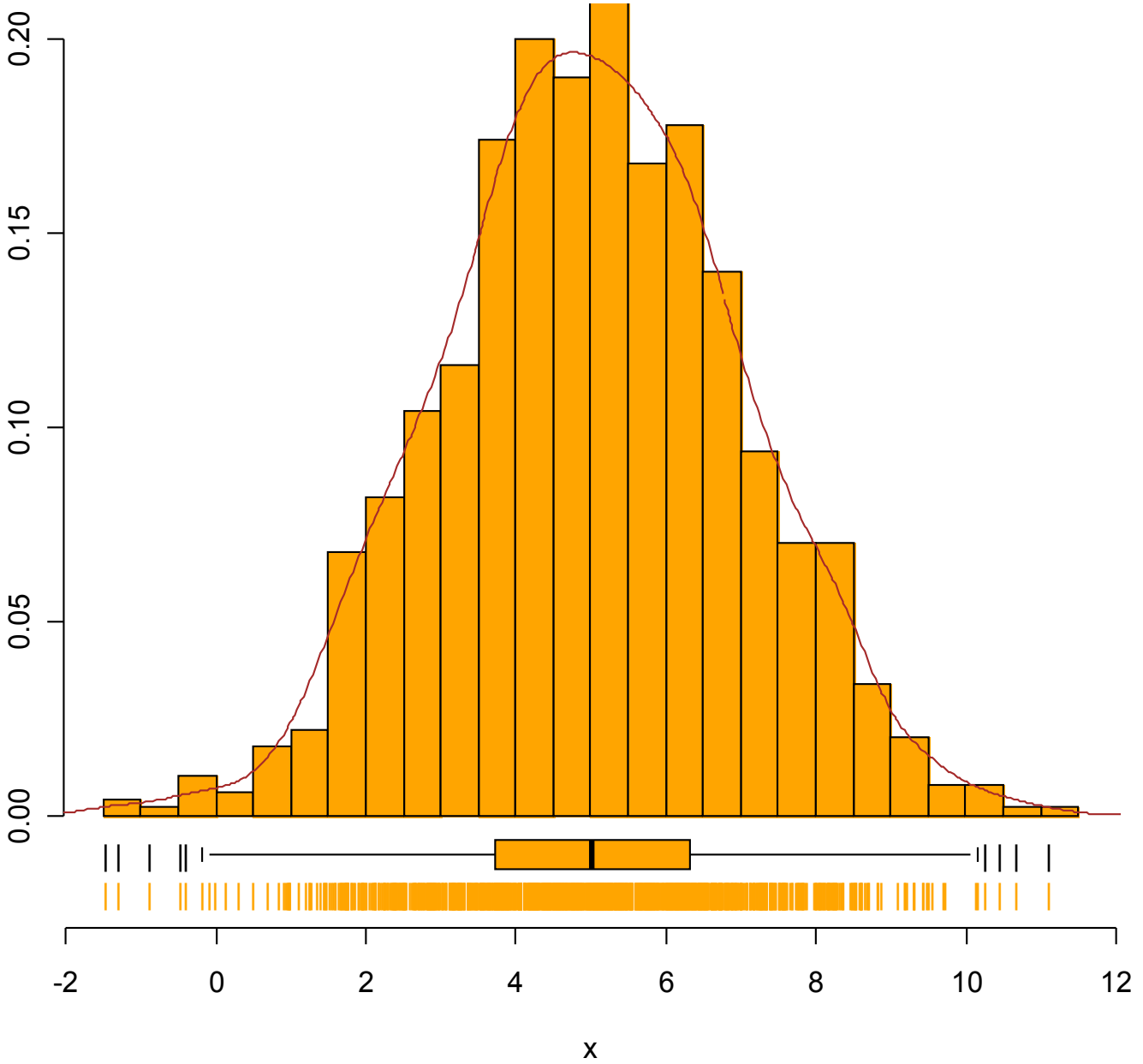
$$x_r = \frac{x_c}{s}$$

Autre exemple

```
 [1]  4.74279234  8.90074709  7.06175647  2.80194483  4.19497661  4.30344155
 [7]  5.58483474  2.30323484  5.66284107  4.40083221  3.36392437  3.99950448
[13]  3.84358469  4.08730476  8.99779318  4.72991012  4.25692068  4.57993247
[19]  6.50889816  2.92288236  9.20517739  4.29231163  3.39349137  8.87978777
[25]  5.21750942  1.23355214  3.56117390  3.06277902  5.16581773  6.51331144
[31]  7.65417446  3.06673214  6.85696290  6.23328467  4.65693313  3.89305573
[37]  6.25453751  5.52968492  5.93383069  1.20633128  8.16230204  3.58200150
[43]  4.32520576  3.71162361  6.43827240  5.93904064  3.17376193  3.91862340
[49]  6.38859032  7.02370090  7.64838419  6.65913238  2.23316516  4.86409606
 ...
[991]  3.26833660  3.10757379  3.68926328  6.29135011  2.90976308  3.93447886
[997]  2.36438509  7.11194721  8.43741944  2.98630325
```

Comment décrire ces données ?

- Moyenne: 4.9859
- Médiane: 4.9915
- Écart-type: 2.01114
- ?



Min	q1	Médiane	Moyenne	q3	Max
-1.477	3.724	5.014	5.006	6.323	11.110

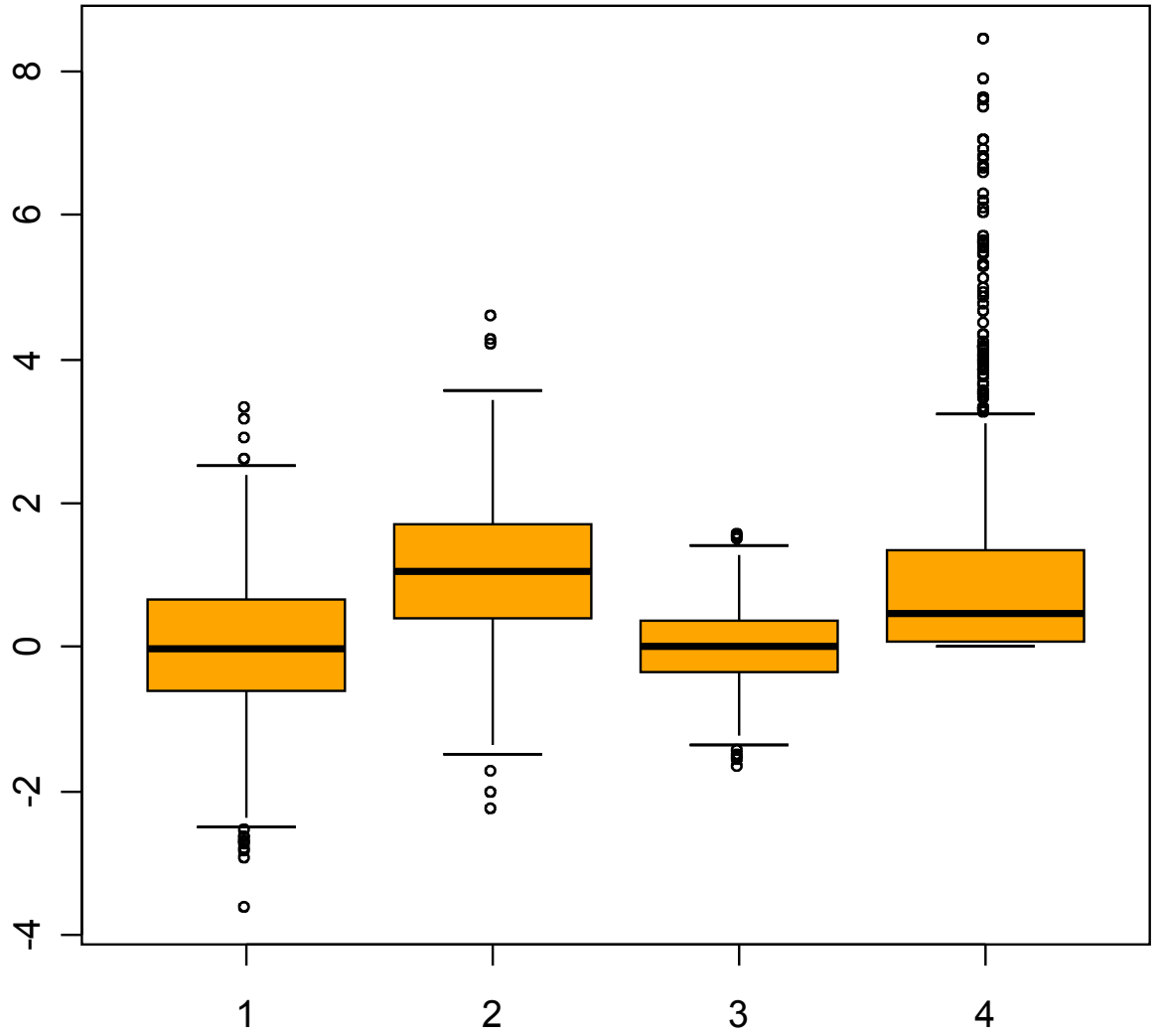
Représentations graphiques

- Box plot
- Histogramme
- Diagramme en bâtons
- Diagramme de dispersion

Box plot

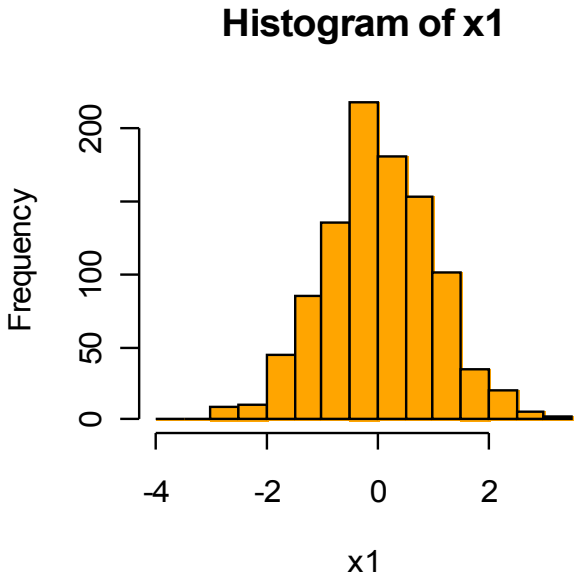
- Résumé à 5 valeurs
- Médiane
- 1er et 3e quartiles q1, q3
- Min et Max
- 1.5 fois intervalle interquartile (q3 – q1)

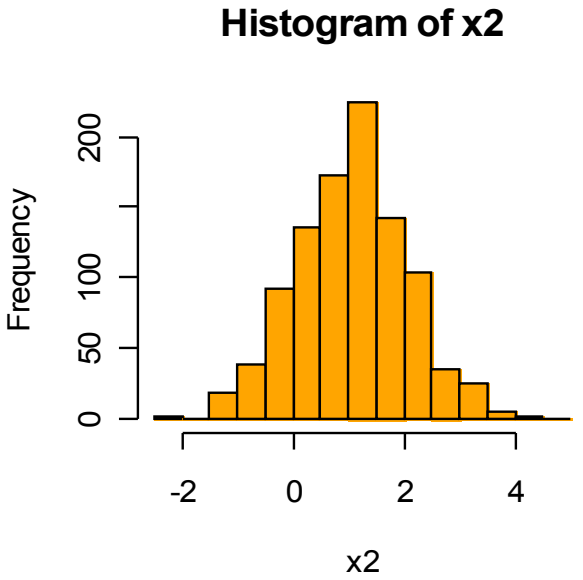
X1 ~ N(0,1)
X2 ~ N(1,1)
X3 ~ N(0,0.5)
X4 ~ Chi2(1)

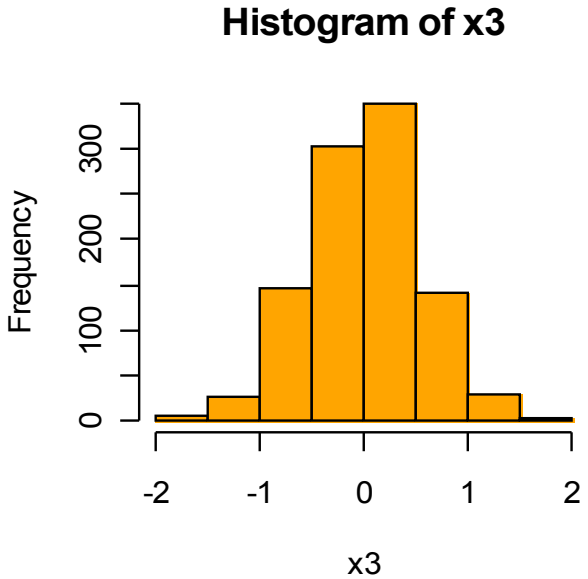


Histogramme

- Surface proportionnelle aux fréquences
- Pas d'espace entre les rectangles
- Nombre de classes à choisir







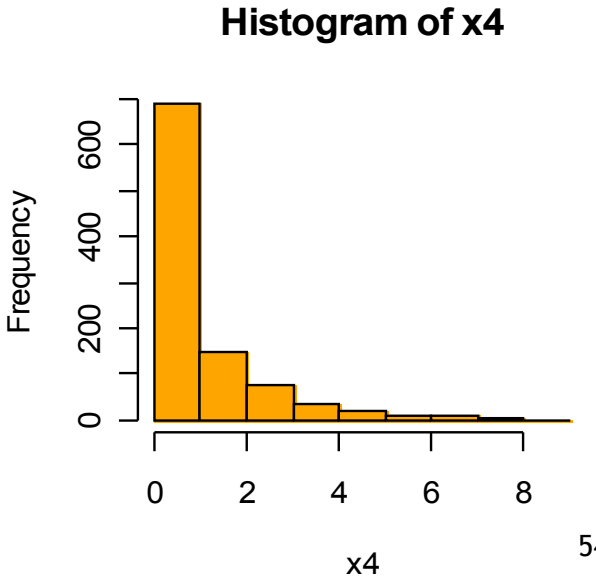
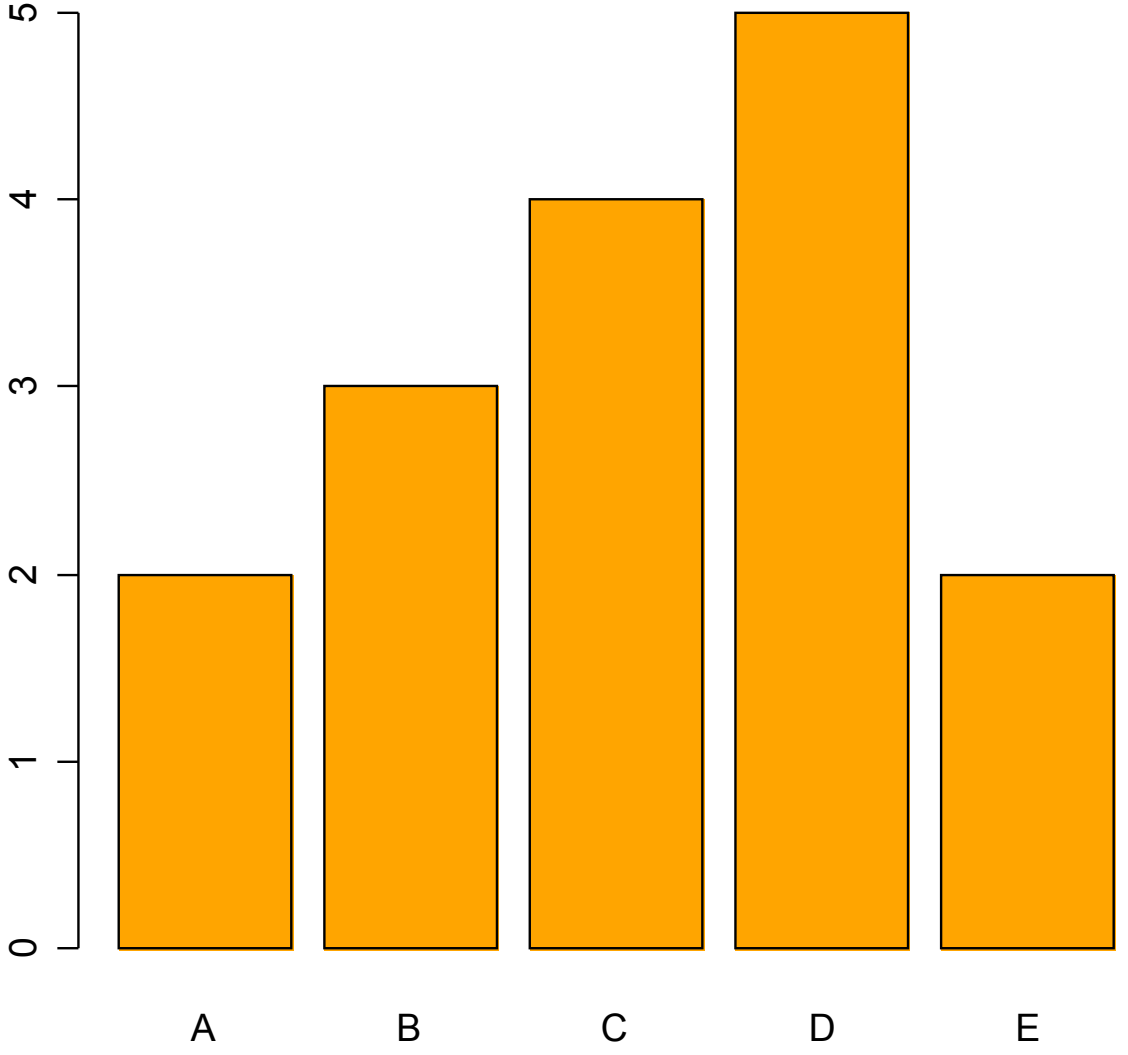


Diagramme en bâtons

- Utile pour les variables discrètes ou catégorielles



Cleveland Dot Plots

- Aussi appelé Dot Chart

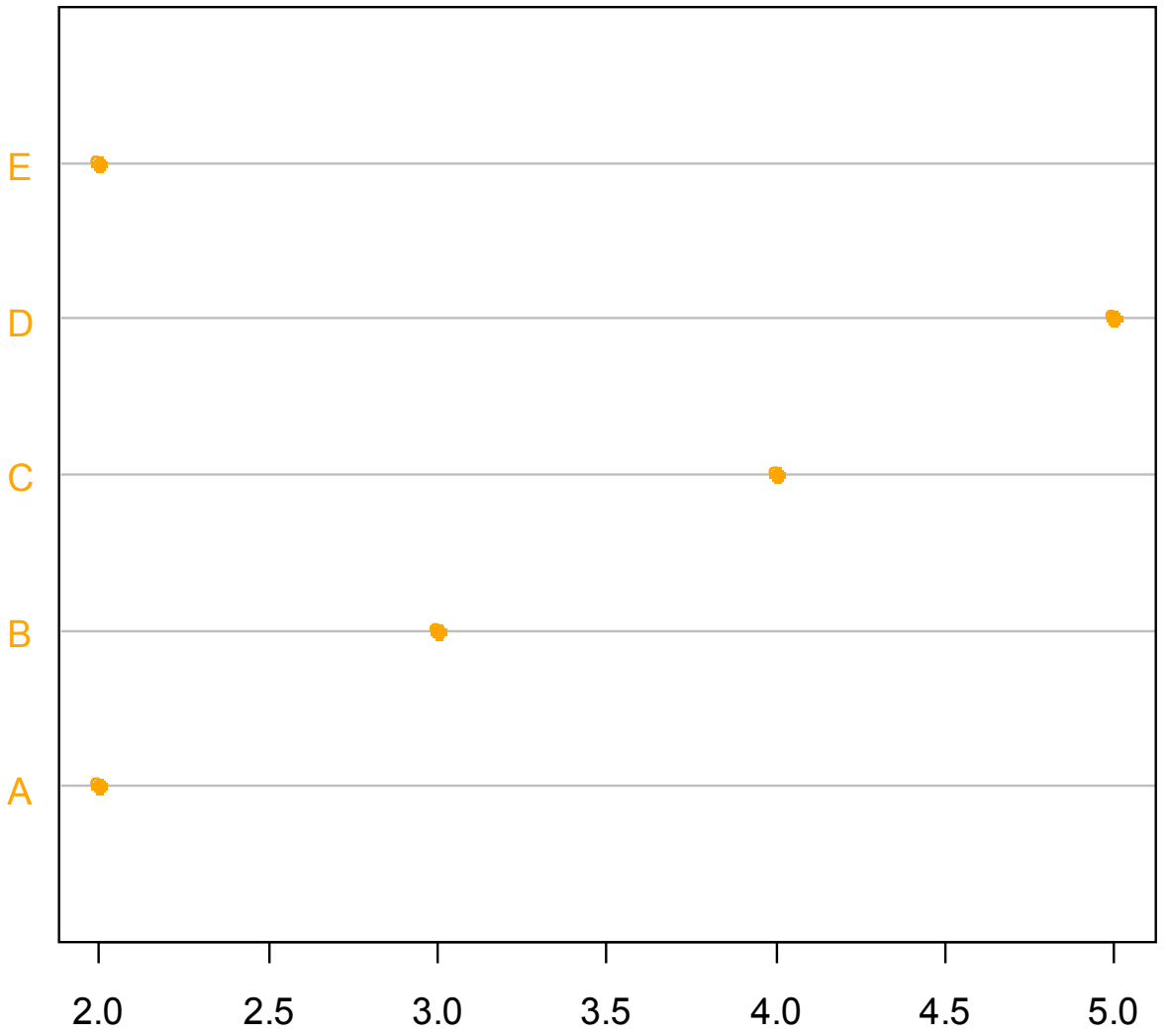
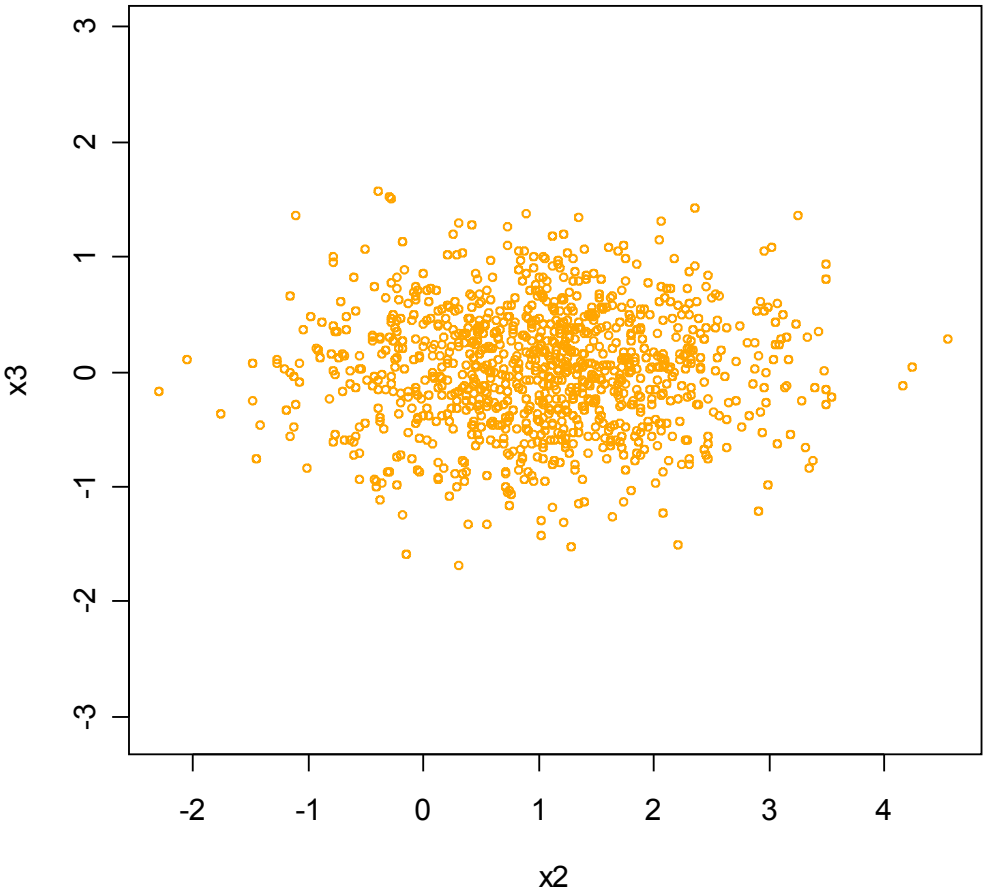
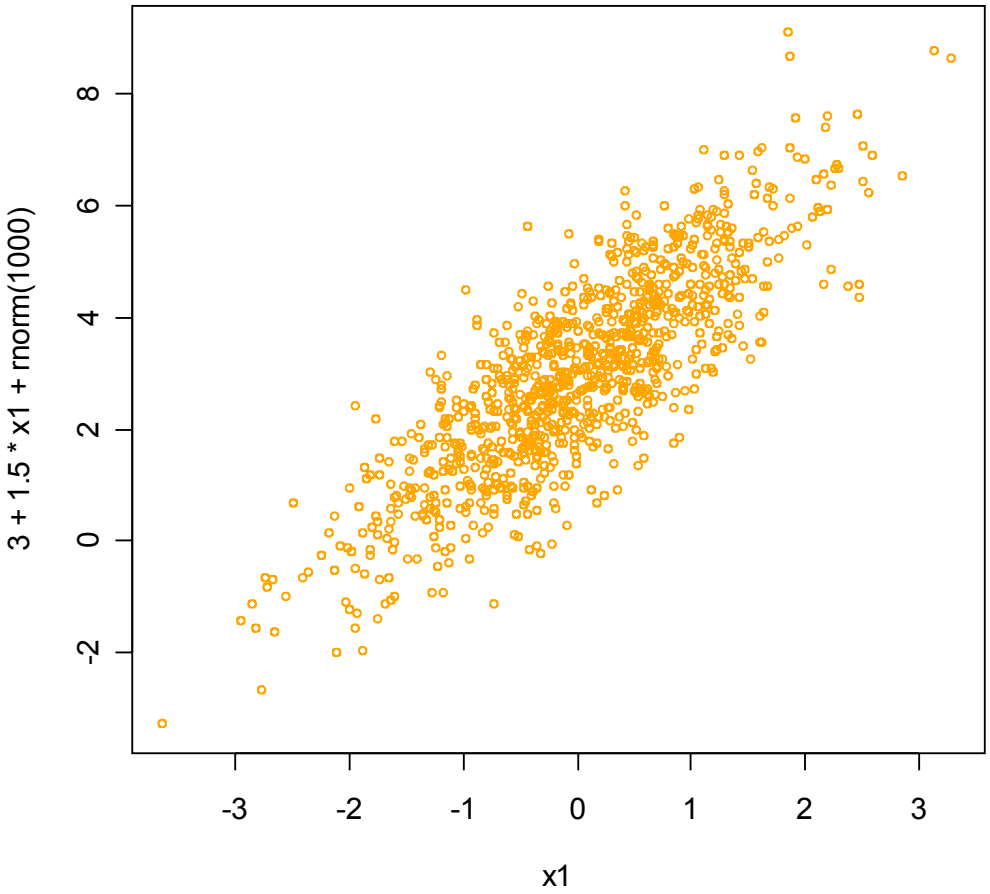
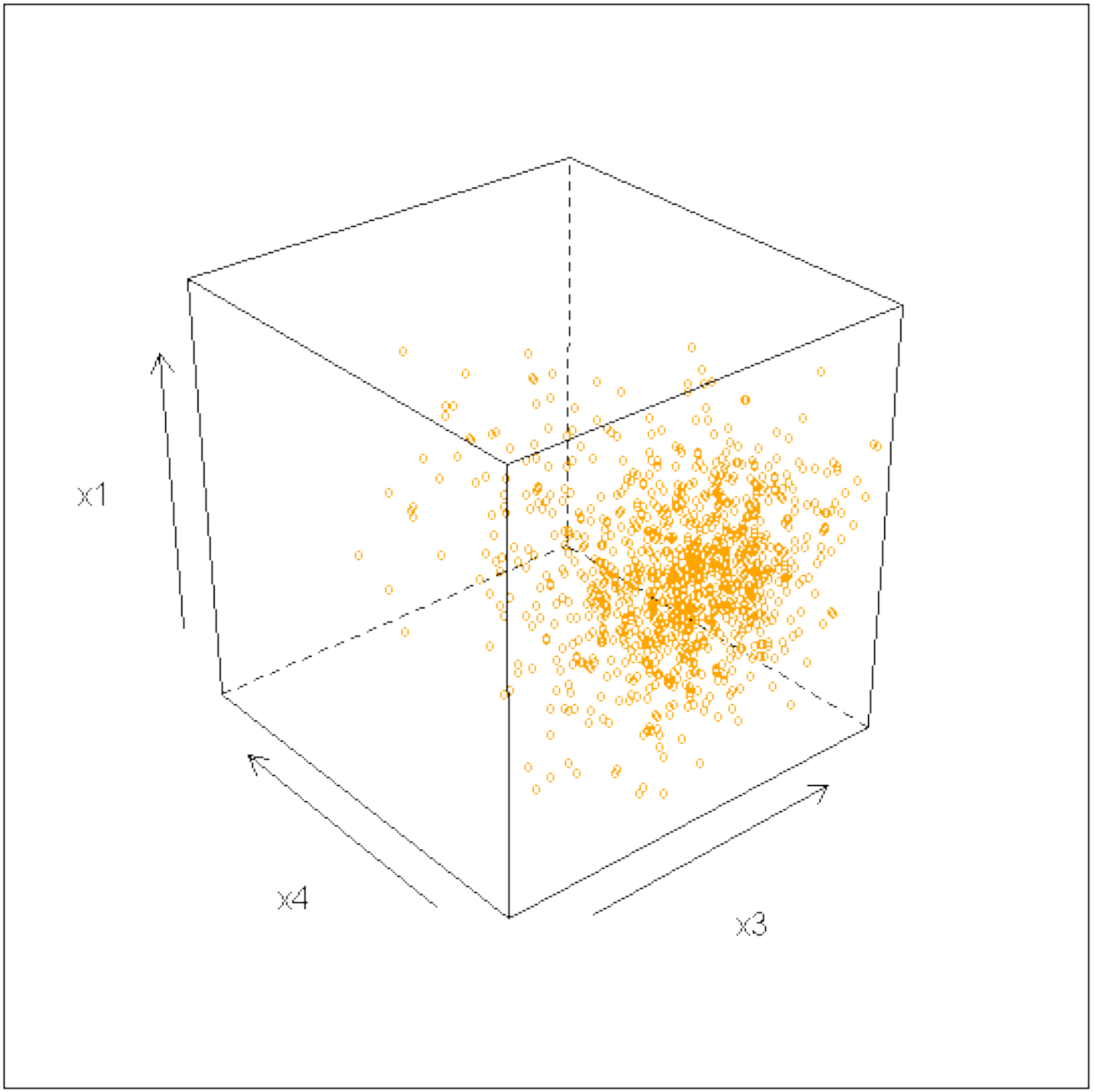


Diagramme de dispersion





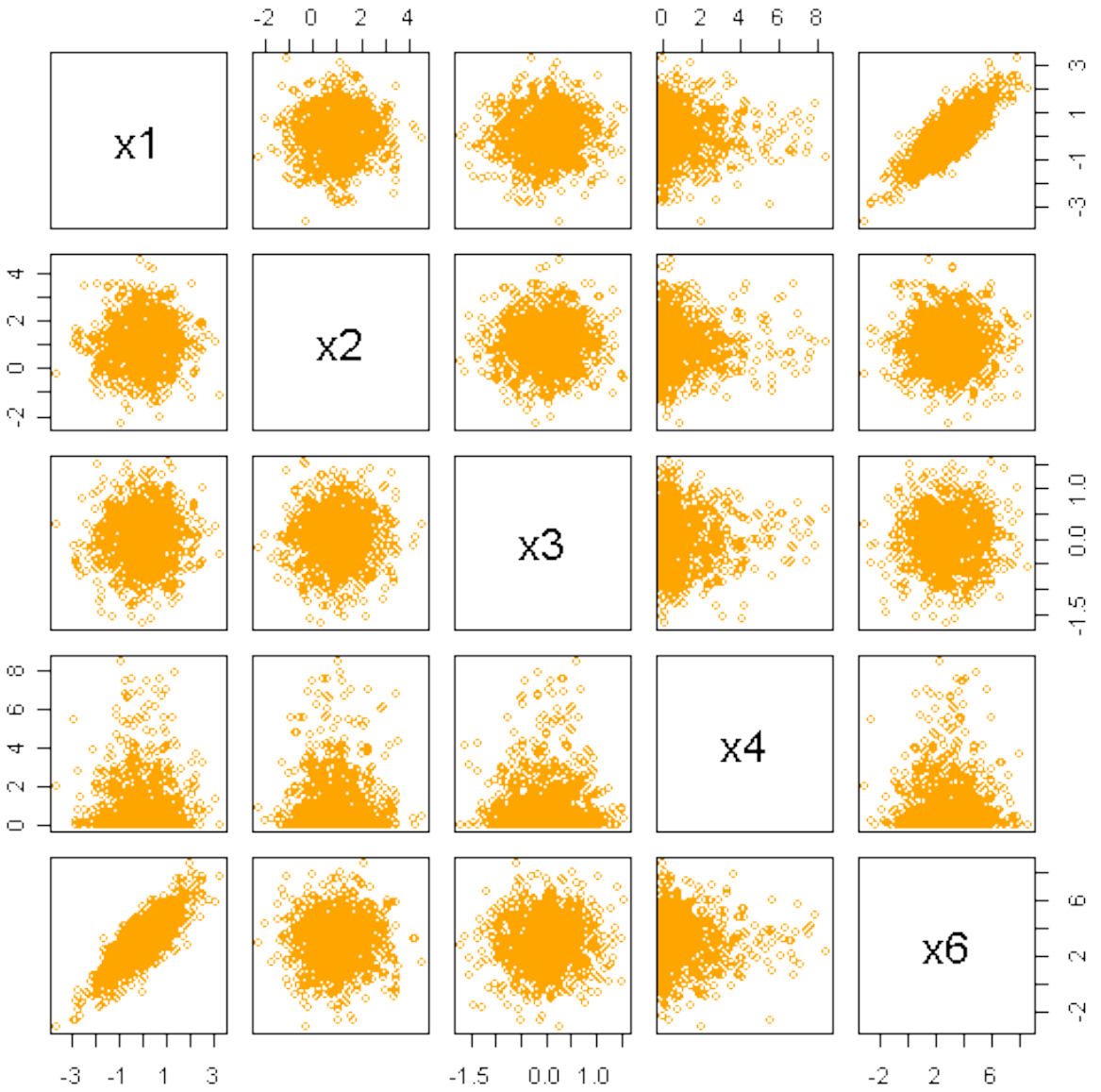
Problème de surimpression

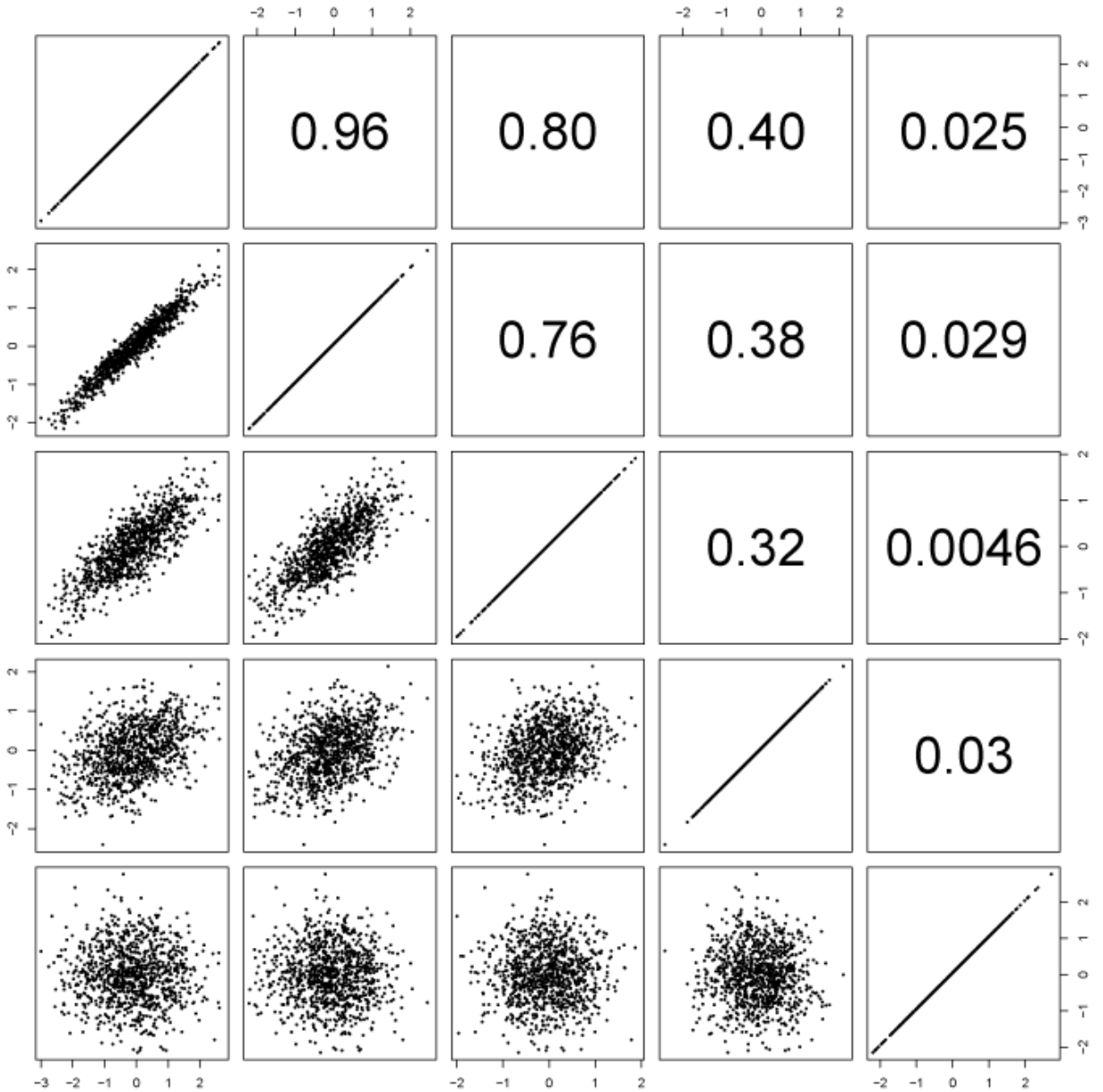


Corrélation

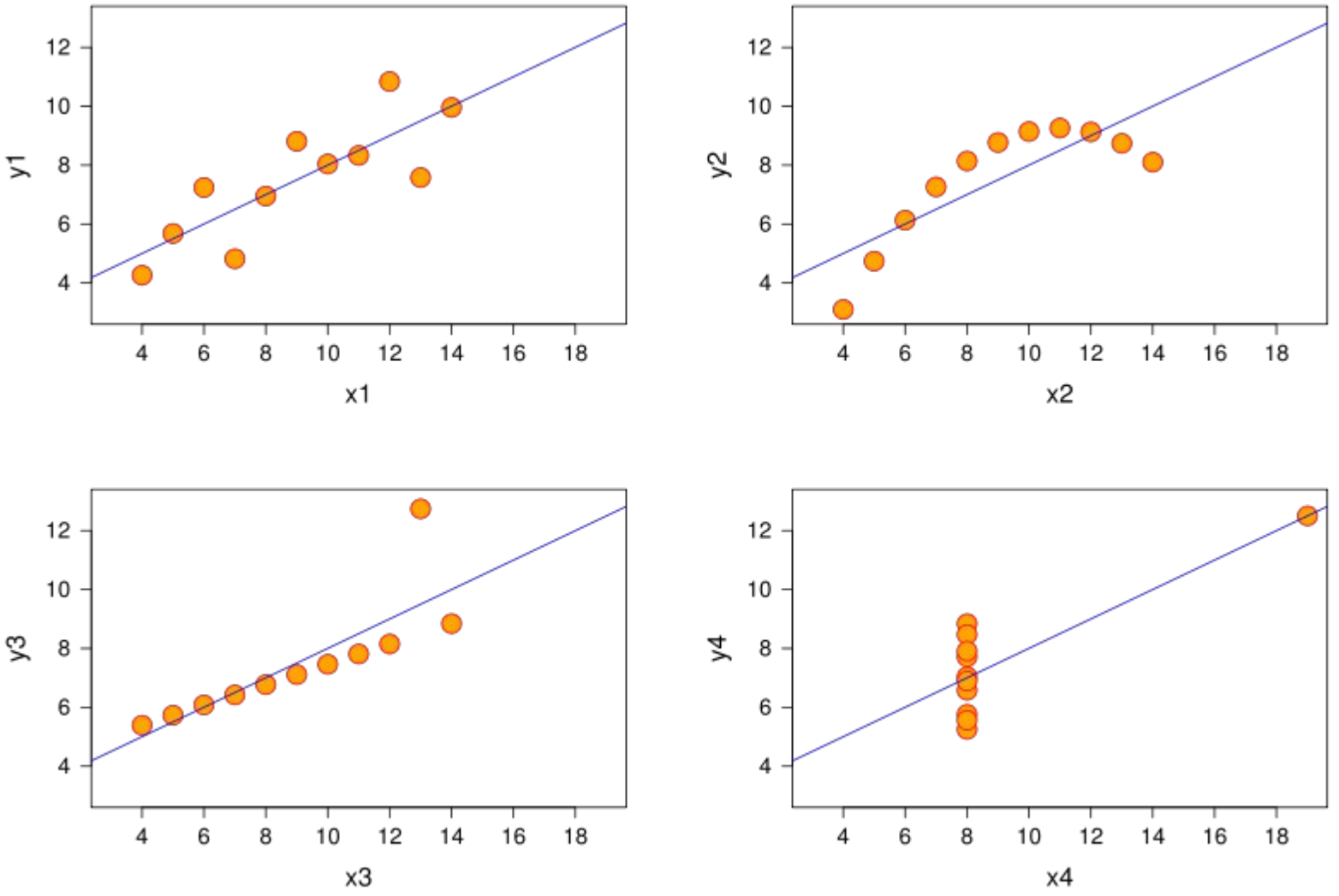
- Mesure de l'association linéaire entre deux variables
- Covariance: $$cov(x, y) = \frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$
- Corrélation: $$\rho(x, y) = \frac{cov(x, y)}{s_x s_y}$$

Matrice de diagrammes





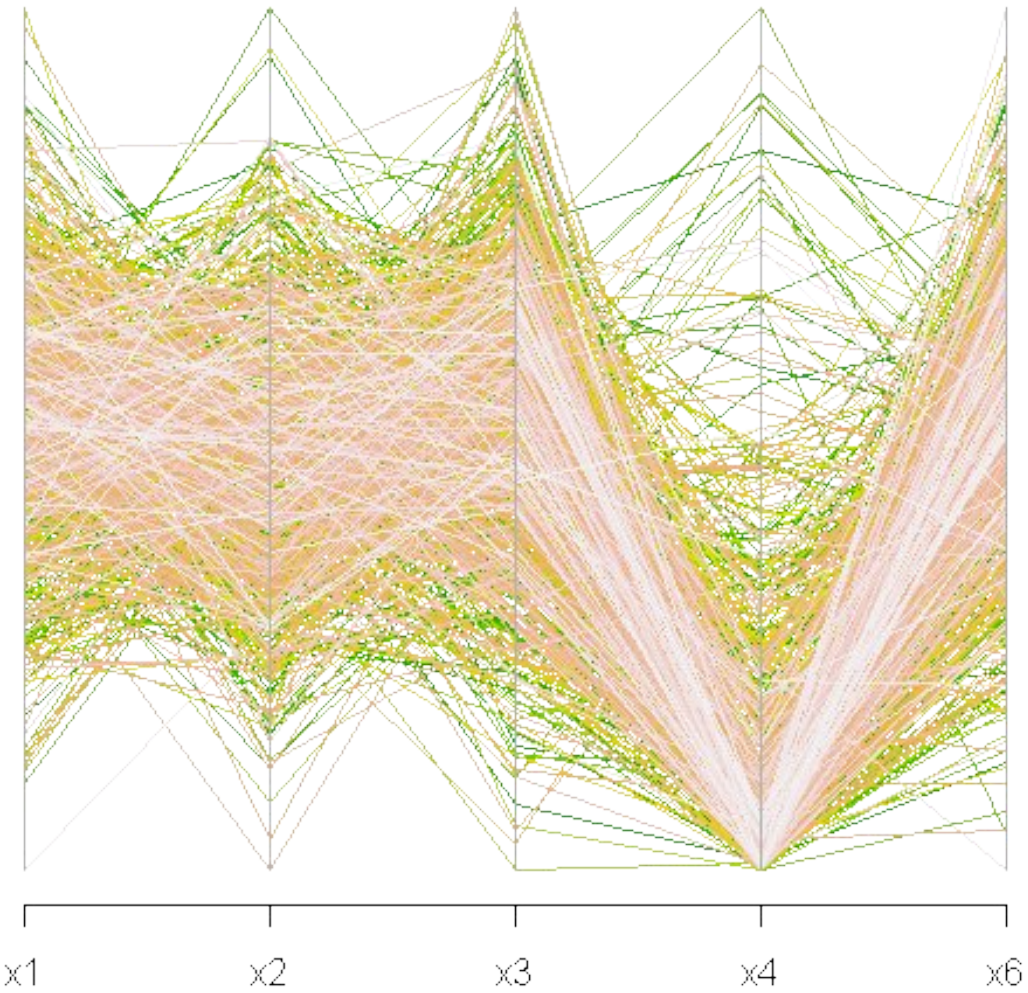
http://en.wikipedia.org/wiki/Image:Corr-example.png



Exemple de Anscombe, corrélation = 0.81

Coordonnées parallèles

- Utile pour représenter plus de 3 variables
- Influence de l'ordre
- Influence de l'échelle



Incertitude

- Variable aléatoire

  X comportement inconnu, x sa réalisation

- Densité

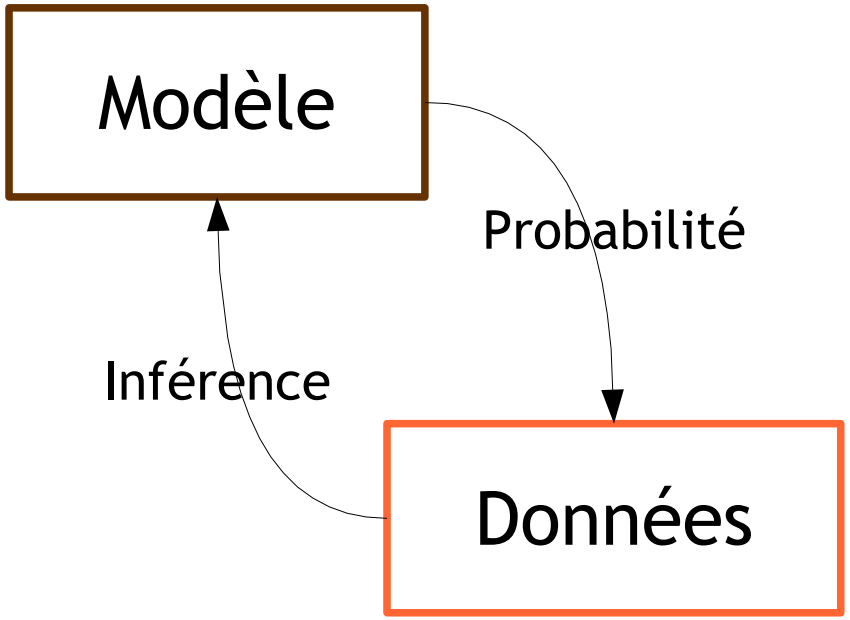
  $f(x)$ décrit la densité de probabilité de X

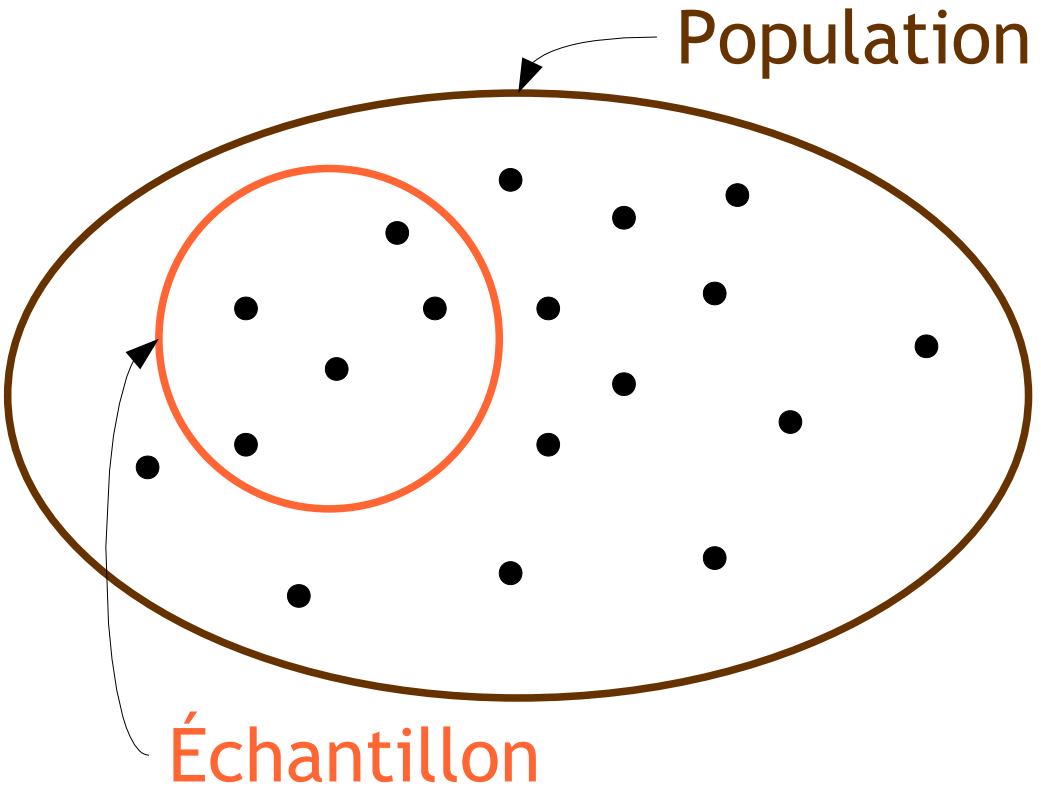
  $F(x)$ est appelée distribution de probabilité de X

$$P(X<a) = \int_{-\infty}^{a} f(x)dx = F(a)$$

$$f(x)\geq 0, \quad P(a<X<b)=\int_{a}^{b} f(x)dx, \quad \int_{-\infty}^{+\infty} f(x)dx=1$$

Inférence





Modèle

- Un modèle est une représentation simplifiée de la réalité
- Données : matrice individus variables
- Processus inconnu ayant généré ces données
- Modèle M : probabilité P d'observer les données D en ayant le modèle M $\quad P(D \mid M)$
- On explicite le modèle M en lui donnant une forme paramétrique $\quad M = M(\theta) \quad P(D \mid M, \theta)$

Estimation

- Maximum de vraisemblance
- Processus de génération des données

$$P(D \mid M, \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

- Valeurs des paramètres du modèle qui donnent la plus grande probabilité d'observer les données existantes

$$L(\theta \mid M, D) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

- Estimation:

$$\hat{\theta} \rightarrow \theta$$

Espérance et variance

- Espérance

$$E(X) = \int_{-\infty}^{+\infty} x\, f(x)\, dx$$

- Variance

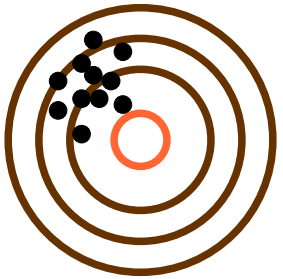
$$V(X) = \int_{-\infty}^{+\infty} (x - E(X))^2\, f(x)\, dx$$

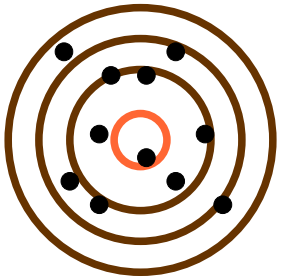
- Biais

$$Biais(\hat{\theta}) = E(\hat{\theta} - \theta)$$

- Erreur quadratique moyenne

$$E\left[(\hat{\theta} - \theta)^2\right] = \left(Biais(\hat{\theta})\right)^2 + V(\hat{\theta})$$





Questions ?

Aperçu de méthodes statistiques

Techniques statistiques

- Classification
  - Arbres de décision / rule-based, nearest neighbors, bayesian, neural nets, support vector machines, etc.
- Clustering
  - K-moyennes / DBScan
- Régression
  - linéaire, non linéaire, logistique, simple, multiple

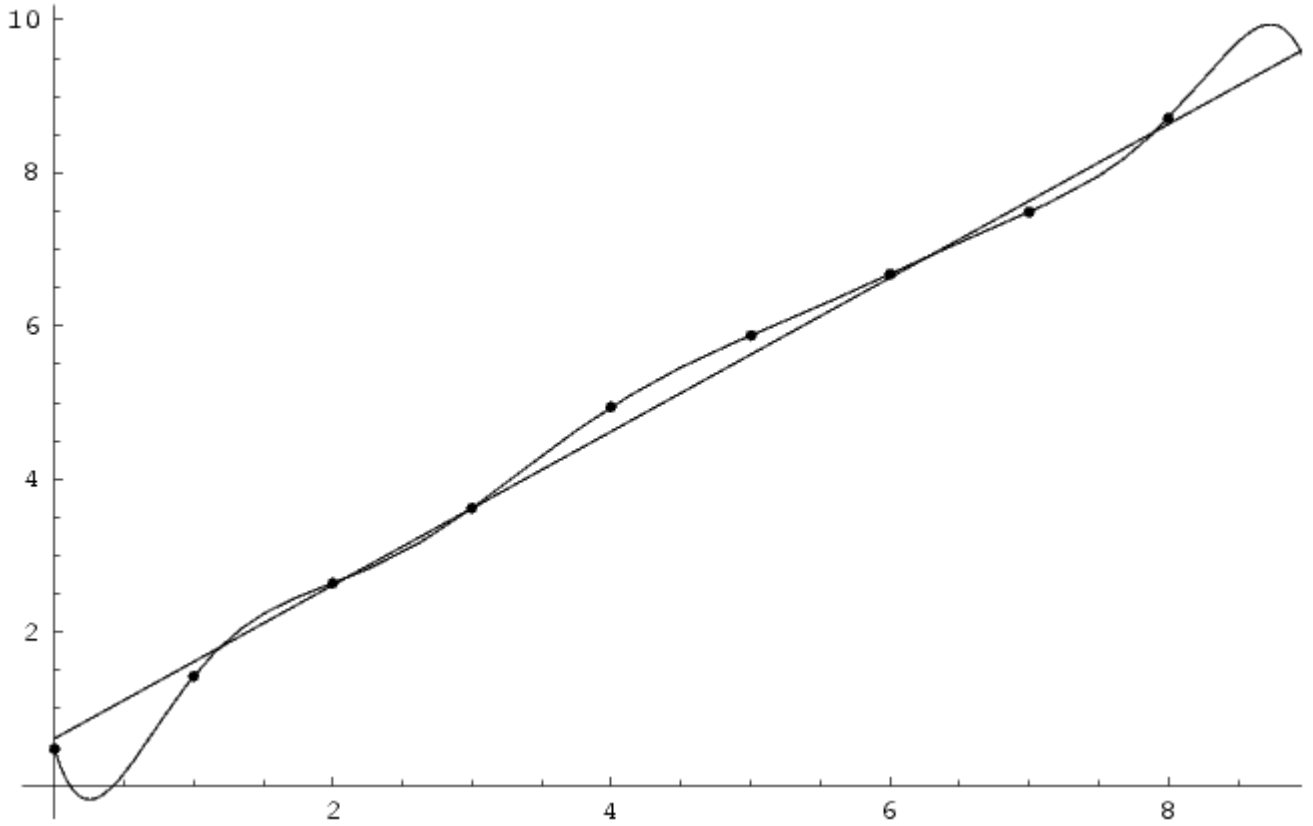
Type d'apprentissage

- Supervisé : finalité explicative, modélisation, une variable supposée influencée par les autres.
  Ex : classification, régression
- Non supervisé : finalité prédictive, interprétation moins importante, recherche d'une typologie entre les variables.
  Ex : clustering

Pièges

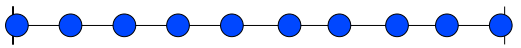
- Acharnement : data snooping, data dredging, overfitting, surapprentissage
  - On favorise trop la diminution de biais par rapport à la variance. Ex : introduire trop de variables
- Malédiction de la dimensionnalité : curse of dimensionality
  - La complexité du calcul croît de façon combinatoire avec le nombre de données
  - Plus un espace est de grande dimension, plus il est creux

Surapprentissage

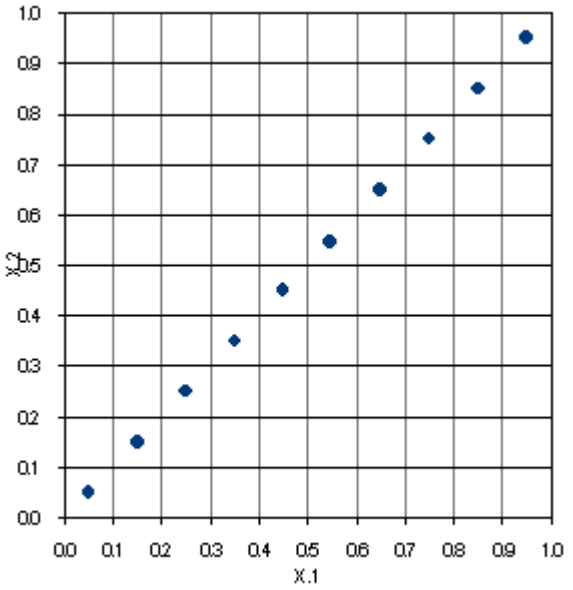


Théorème de Stone-Weierstrass :
Toute fonction continue définie sur un compact peut être approchée aussi près que l'on veut par une fonction polynomiale

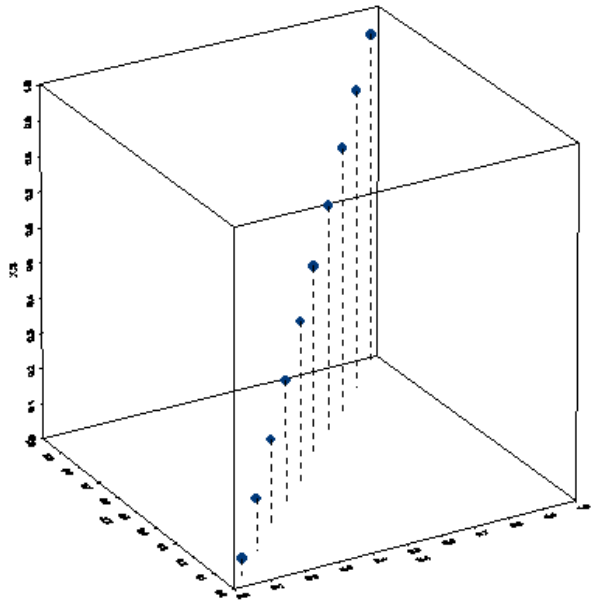
Malédiction de la dimensionnalité



d = 1
10 échantillons
10%



d = 2
10 échantillons
1%



d = 3
10 échantillons
0.1%

Pour couvrir 10% d'un espace de dimension *d*, il faut 10^d échantillons

Classification

- Ranger les individus dans des classes prédéfinies en fonction de leurs variables
- Apprentissage supervisé
- Ex : Modéliser une variable catégorielle en fonction de variables quantitatives

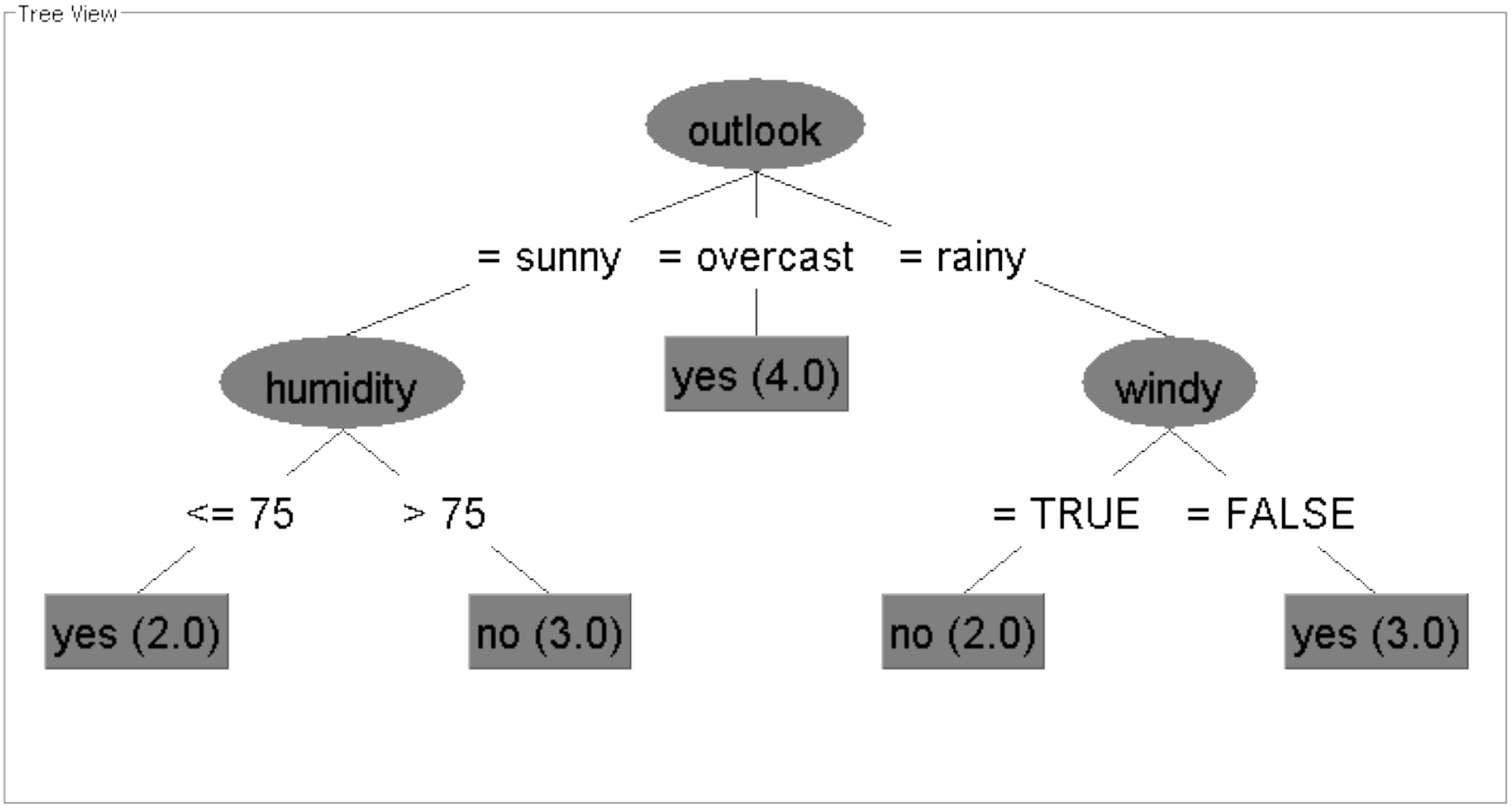
$$P(classe \mid x) = f(x, \theta)$$

  - Analyse discriminante
  - Régression logistique
  - Arbres
  - Classification de Bayes
  - ...

Arbres

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Arbres



Clustering

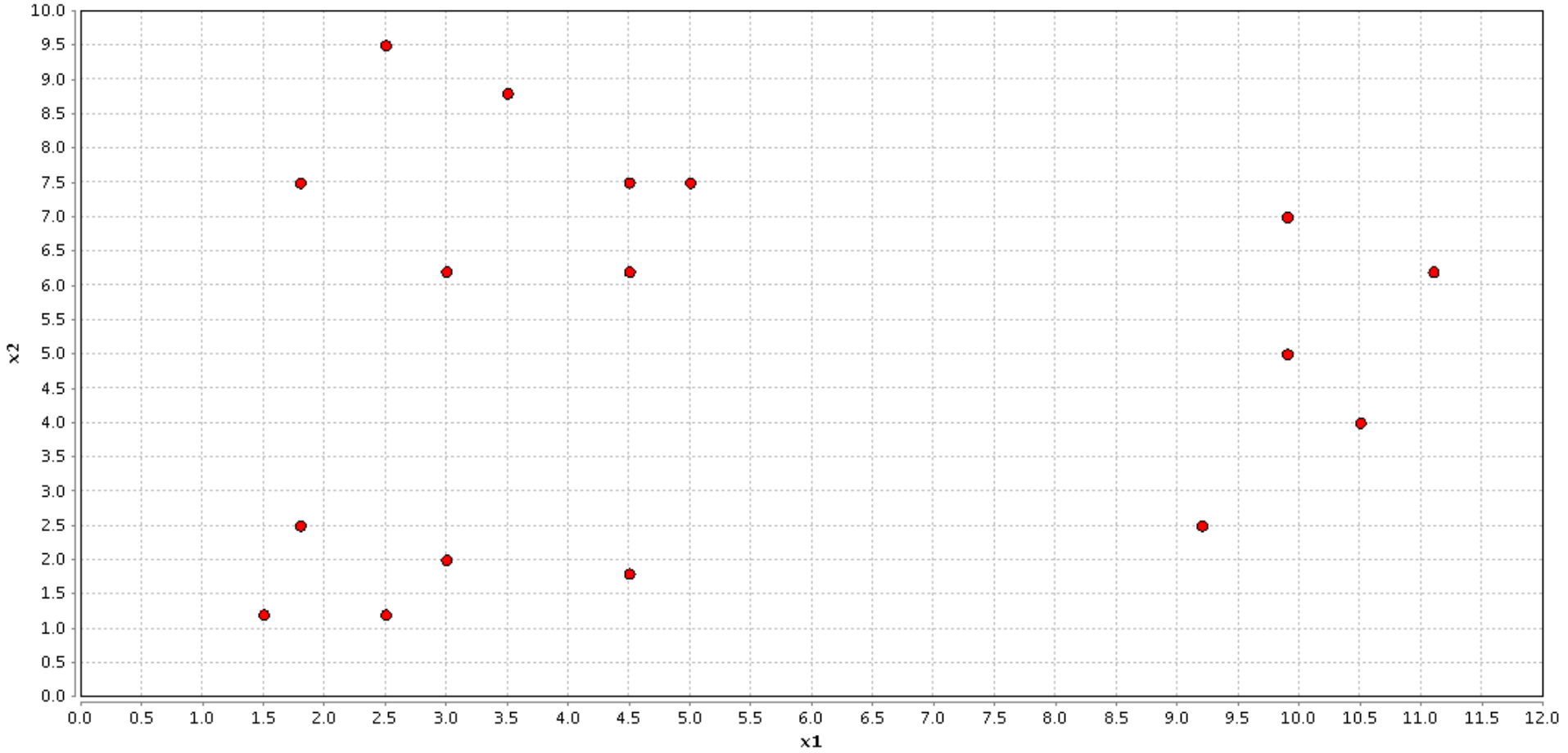
- Trouver quels groupes ressortent des données sans les connaître *a priori*
- Apprentissage non supervisé
- Méthodes dépendent du type de données
- Exemples :
  - k-means (k-moyennes)
  - DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

k-means

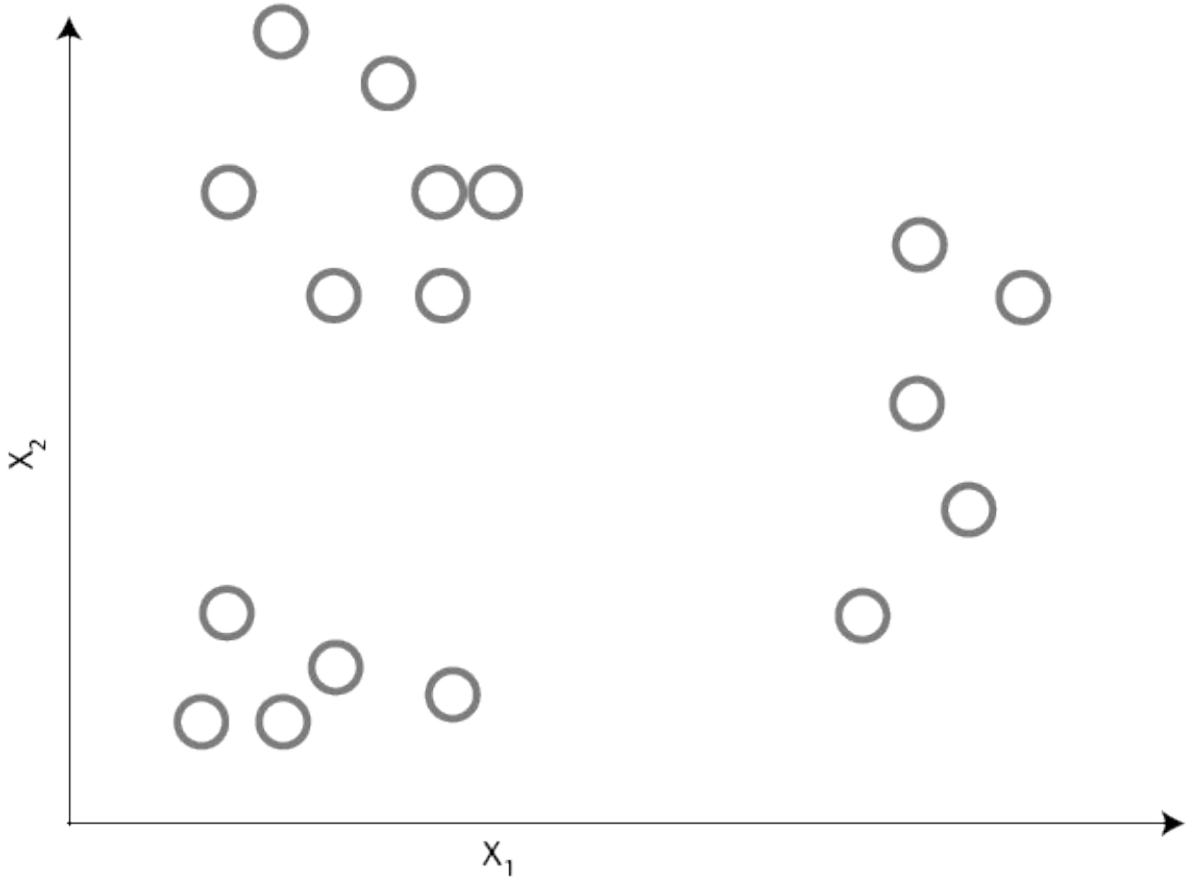
1. Placer les *k* centres initiaux
2. Assigner chaque individu au groupe qui est le plus proche du centre
3. Lorsque tous les individus sont assignés, recalculer les position des *k* centres
4. Répéter les étapes 2 et 3 jusqu'à convergence des centres

Produit une partition des individus dépendant de la distance utilisée

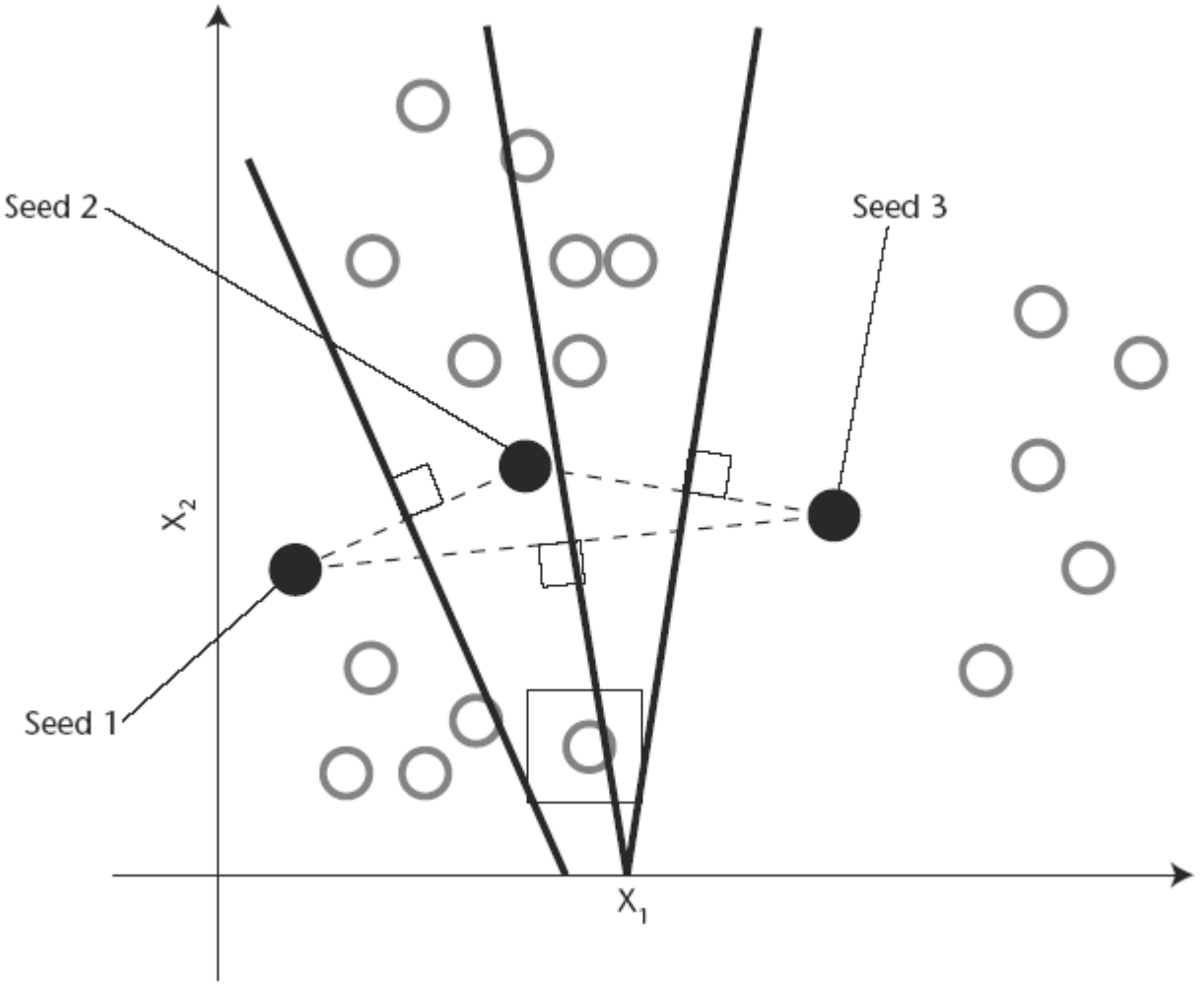
Exemple

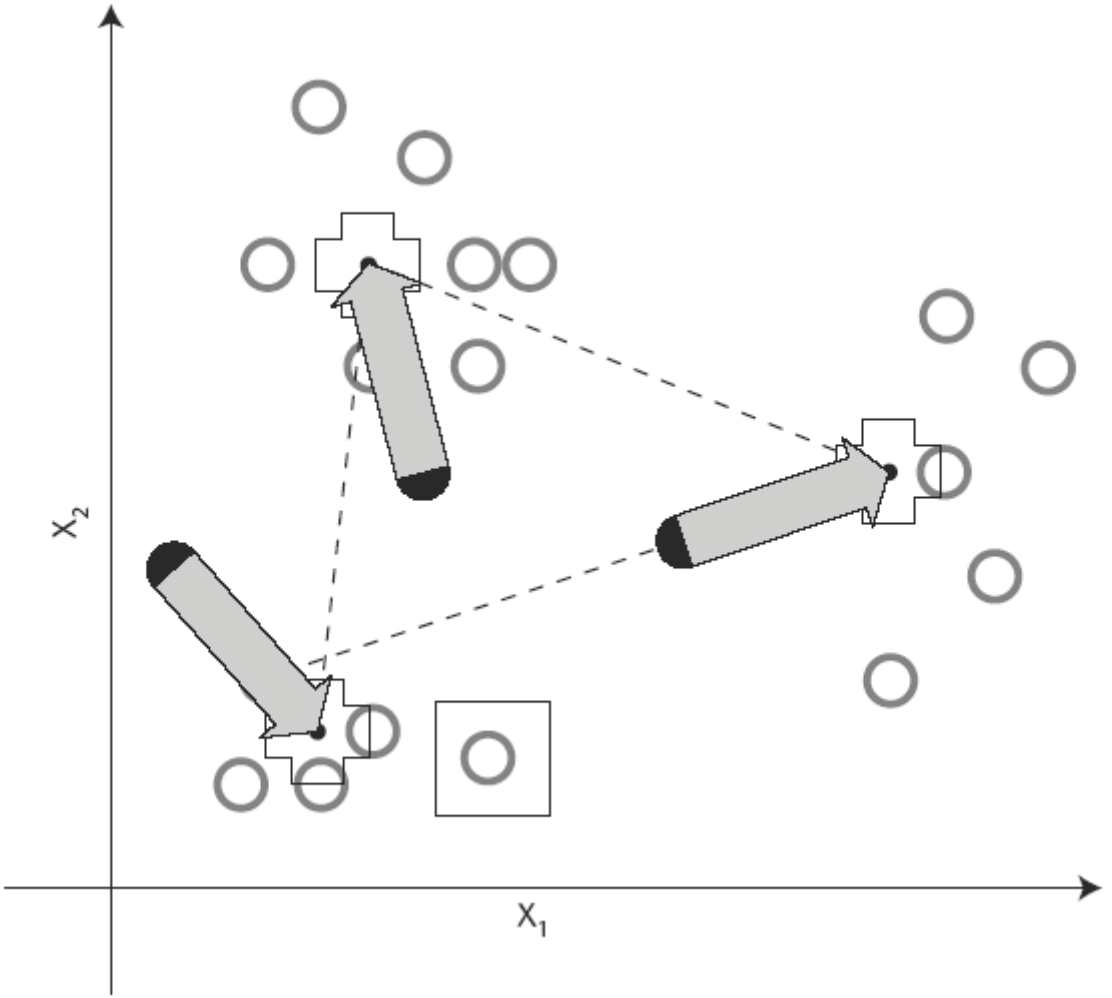


k-means

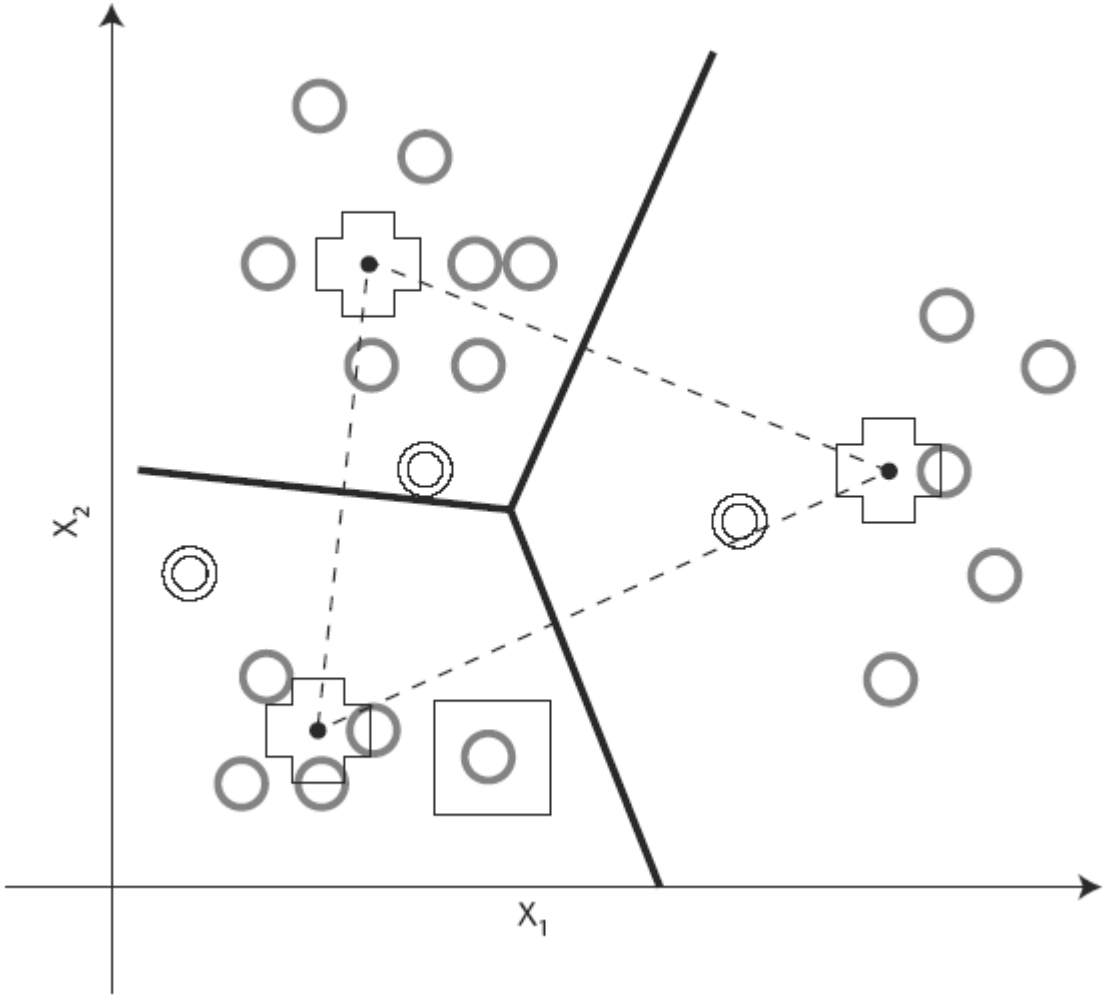


k-means

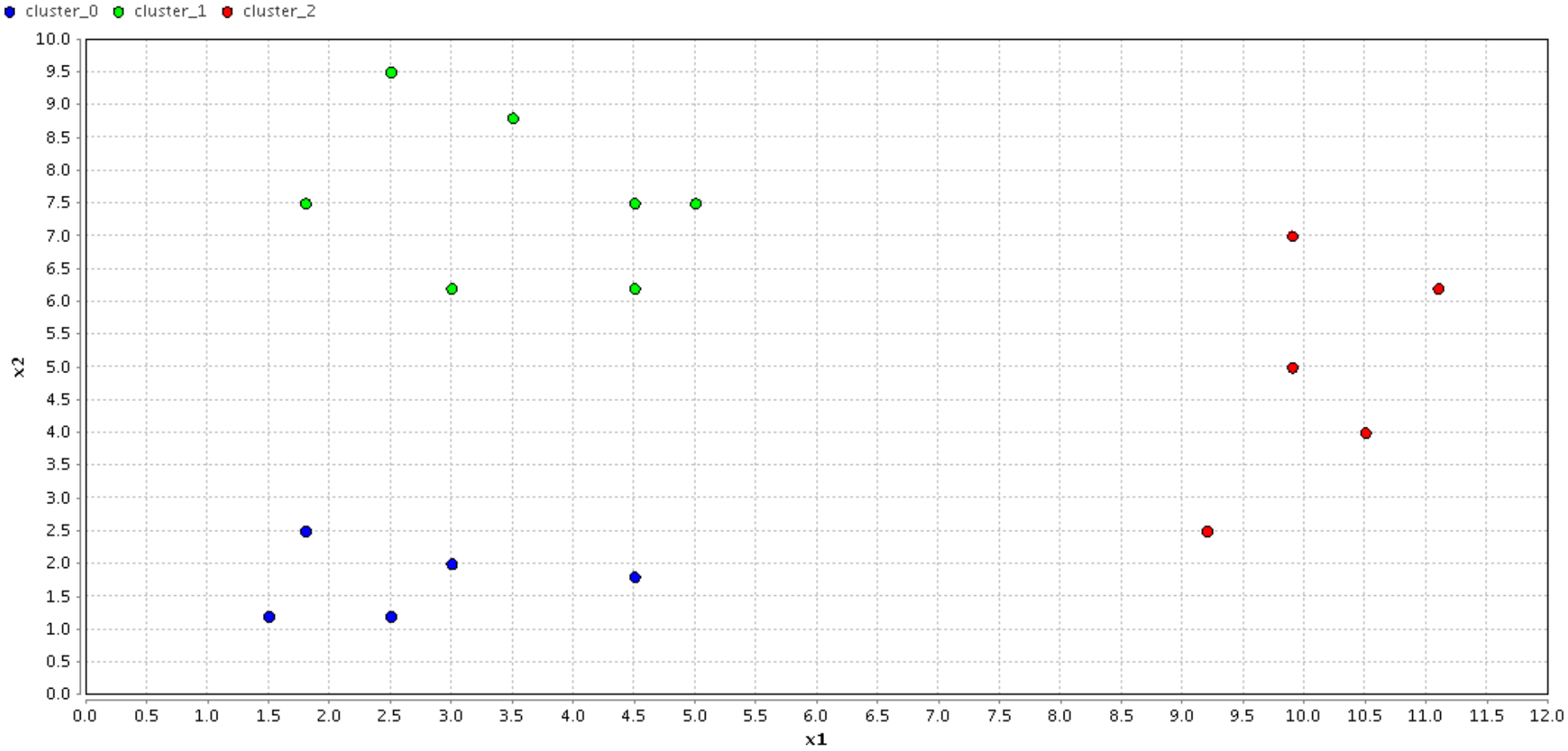



X2
X1

k-means



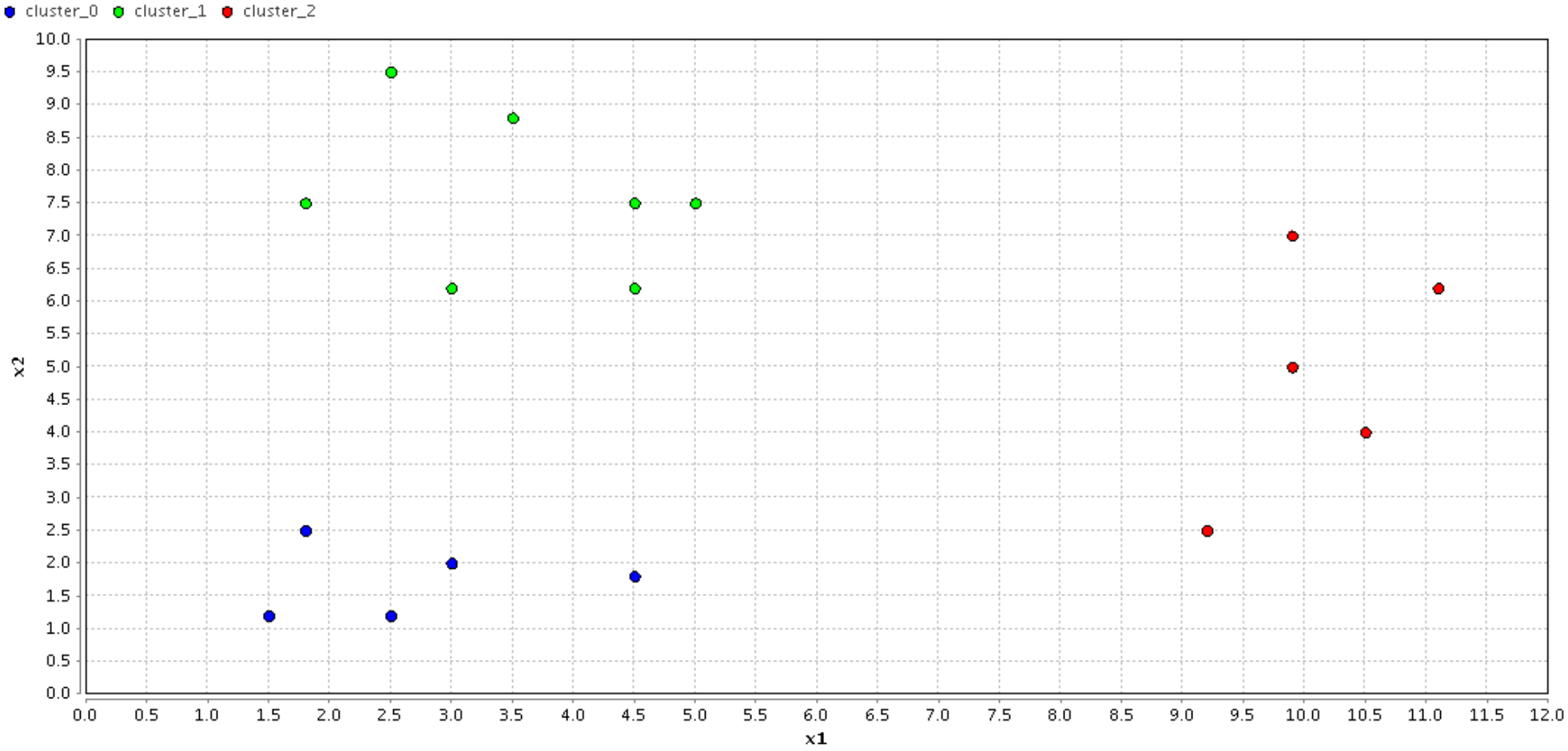
k-means



DBSCAN

1. Initialiser les paramètres ε et *minP*
2. Pour chaque point non visité *P*
    1. Compter le nombre de points *N* dans le voisinage de *P* défini par ε
    2. Si $N < minP$, alors le *P* est marqué comme bruit
    3. Sinon *P* est ajouté au cluster *C*
    4. Continuer la visite des points du voisinage
3. Fin

DBSCAN



Avantages de DBSCAN

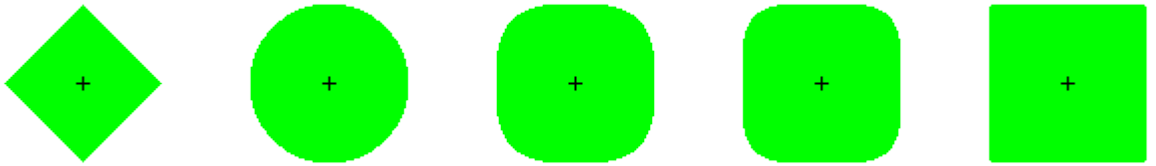
- Pas d'initialisation du nombre de clusters a priori
- Pas de biais quant à la forme ou la taille des clusters
- Robuste au bruit, perturbations

Inconvénients de DBSCAN

- Nécessite une bonne définition de la distance utilisée pour le voisinage
- Complexité de temps quadratique $O(N^2)$, N étant le nombre de points du cluster
- Pas adapté aux ensembles de points hiérarchiques, avec densité variable

Extensions

- Distances $d_p(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}$



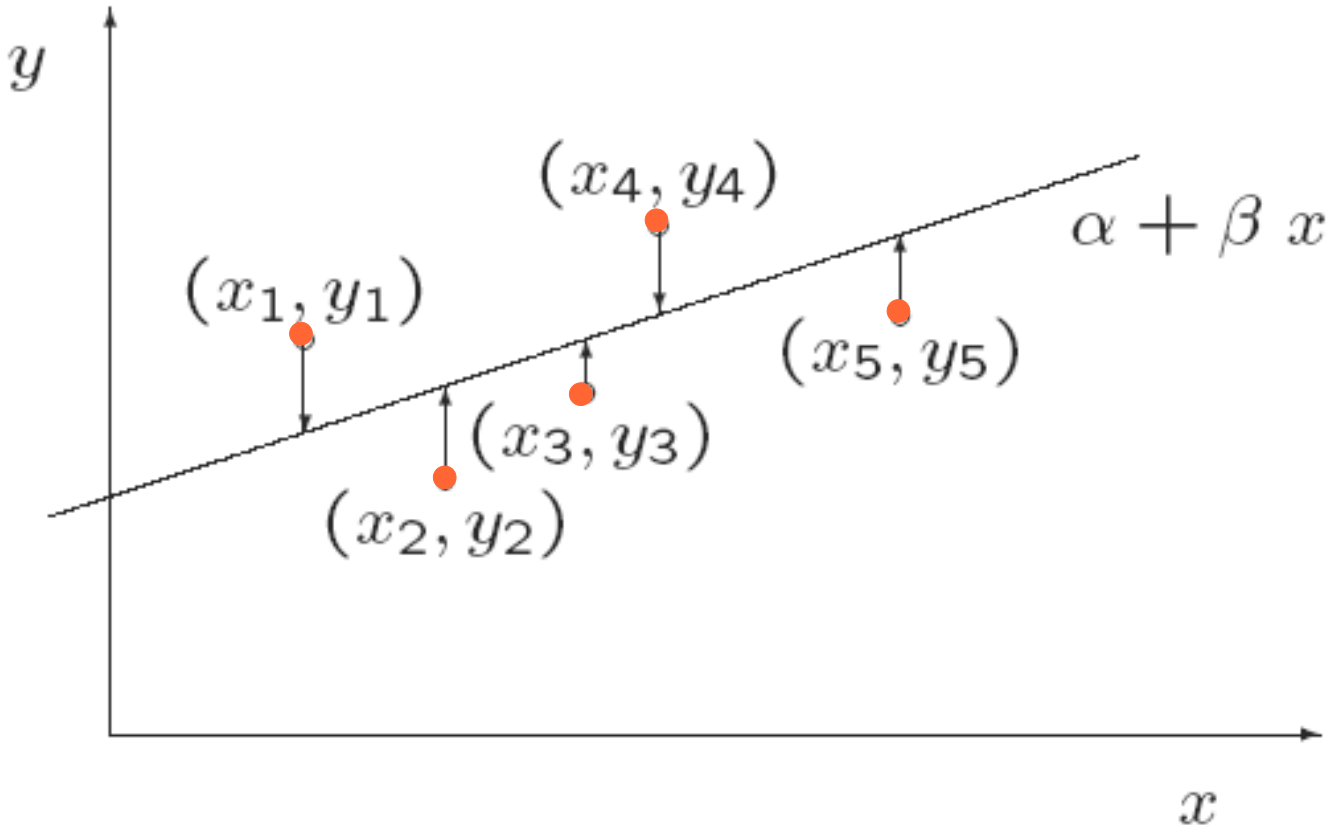
$p = 1$ $p = 2$ $p = 3$ $p = 4$ $p = \infty$

- Autres distances: Mahalanobis, Manhattan, ...
- Score à minimiser:
  - Distance maximum d'un individu au centre
  - Somme des moyennes des distances aux centres
  - Somme des variances des distances aux centres
- Frontières non linéaires

Régression

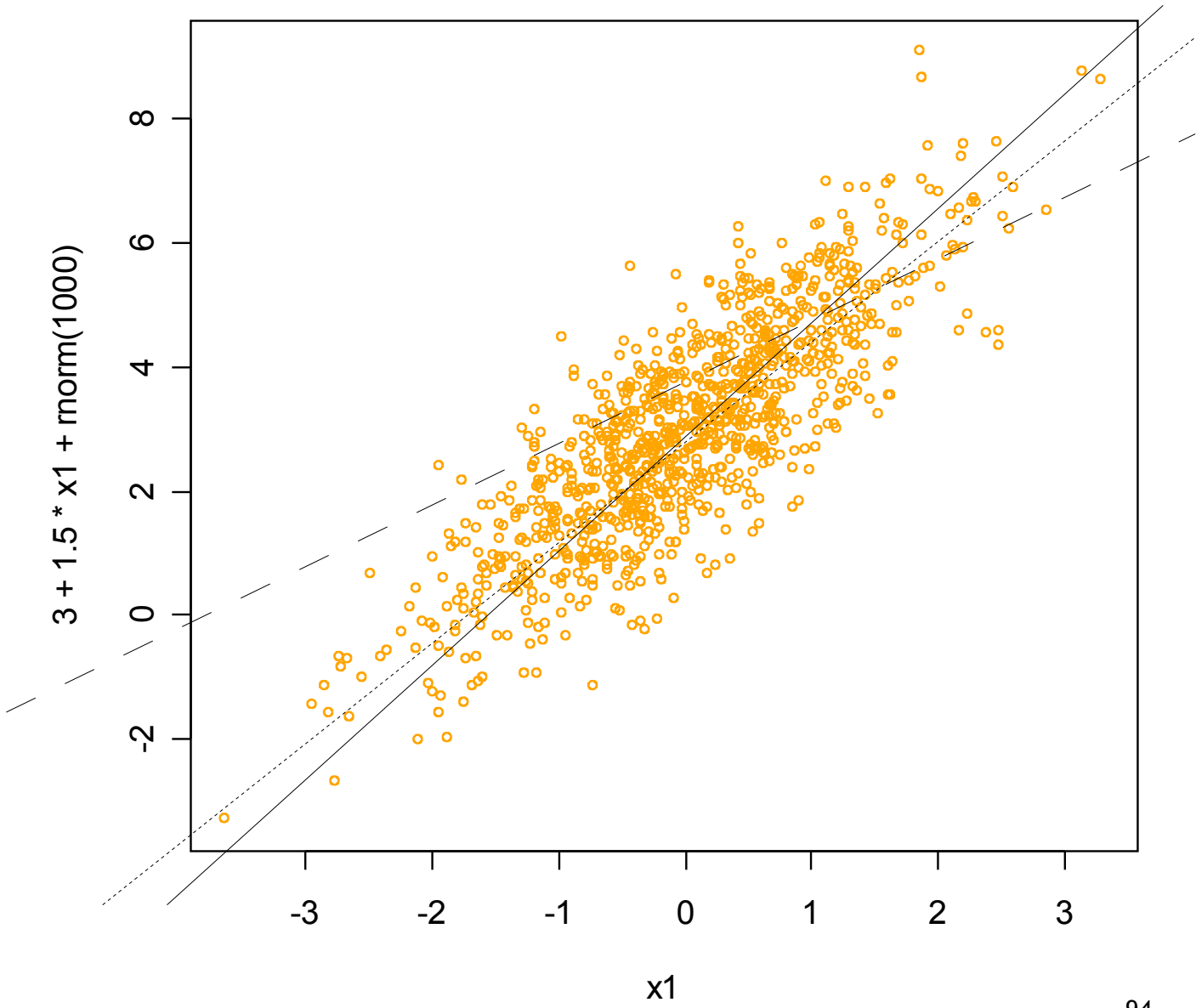
- Apprentissage supervisé, modèle descriptif, explicatif et prédictif
- Très largement utilisée pour les données quantitatives
- Une variable est expliquée par les autres
- Exemples :
  - Quel est la dépense moyenne attendue en fonction du revenu ?
  - Nombre de personnes cliquant sur une publicité sur le Web en fonction du placement ?

Régression linéaire simple



Régression linéaire

- $y = f(x)$
- $y = \alpha + \beta\, x$
- Estimer α et β



Régression linéaire multiple

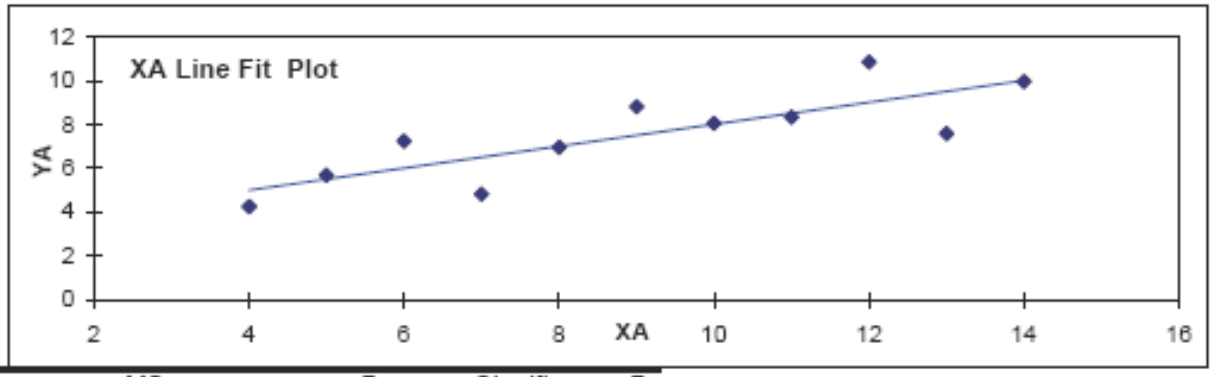
- Variable expliquée y et variables explicatives x
- Modèle choisi $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$
- Minimiser la somme des écarts verticaux au carré entre les données et le modèle (principe des moindres carrés)
- Estimation des paramètres $\hat{\beta} \rightarrow \beta$
- Tester la qualité des résultats, la sensibilité

Régression linéaire

- Diagnostics:
  - Forme du graphique des résidus
  - Indicateur R^2
  - Tests de significativité statistique des paramètres
  - etc.

SUMMARY OUTPUT **A**

Regression Statistics	
Multiple R	0.816420516
R Square	0.66654246
Adjusted R Square	0.629491622
Standard Error	1.236603323
Observations	11



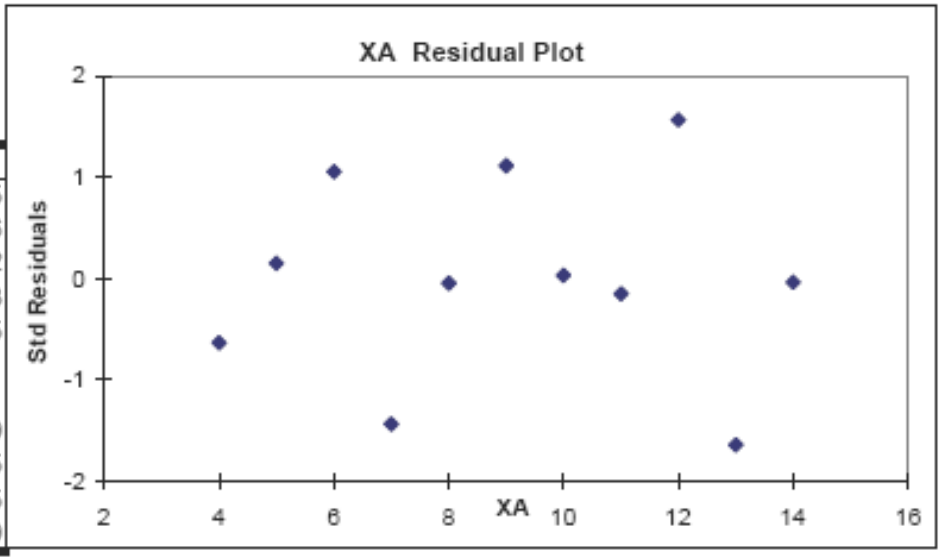
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	27.51000091	27.51000091	17.98994297	0.002169629
Residual	9	13.76269	1.529187778		
Total	10	41.27269091			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	3.000090909	1.124746791	2.667347828	0.025734051	0.455734961	5.544446857	0.455734961	5.544446857
XA	0.500090909	0.117905501	4.241455289	0.002169629	0.233369933	0.766811885	0.233369933	0.766811885

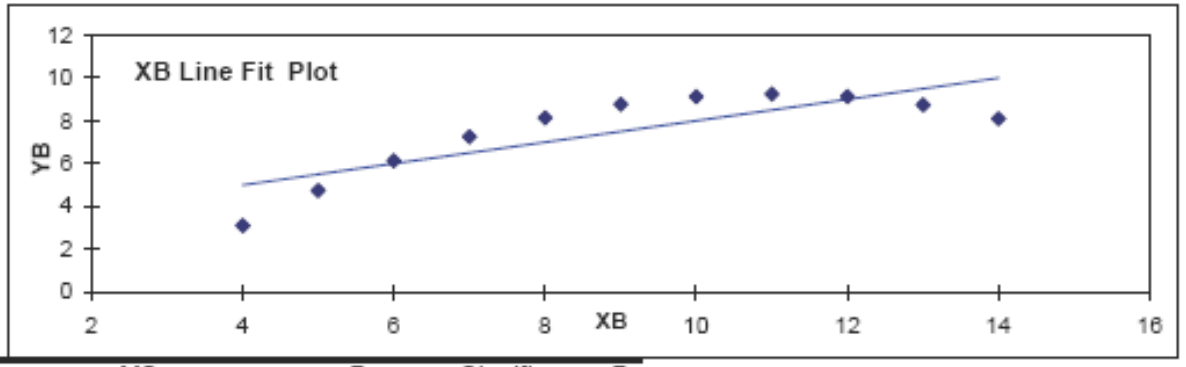
RESIDUAL OUTPUT

Observation	*Predicted YA*	*Residuals*	*Standard Residuals*
1	8.001	0.039	0.033243975
2	7.000818182	-0.050818182	-0.043317906
3	9.501272727	-1.921272727	-1.637711332
4	7.500909091	1.309090909	1.115881668
5	8.501090909	-0.171090909	-0.145839535
6	10.00136364	-0.041363636	-0.035258761
7	6.000636364	1.239363636	1.056445471
8	5.000454545	-0.740454545	-0.631170569
9	9.001181818	1.838818182	1.567426285
10	6.500727273	-1.680727273	-1.432668075
11	5.500545455	0.179454545	0.152968779



SUMMARY OUTPUT **B**

Regression Statistics	
Multiple R	0.816236506
R Square	0.666242034
Adjusted R Square	0.629157815
Standard Error	1.237214205
Observations	11



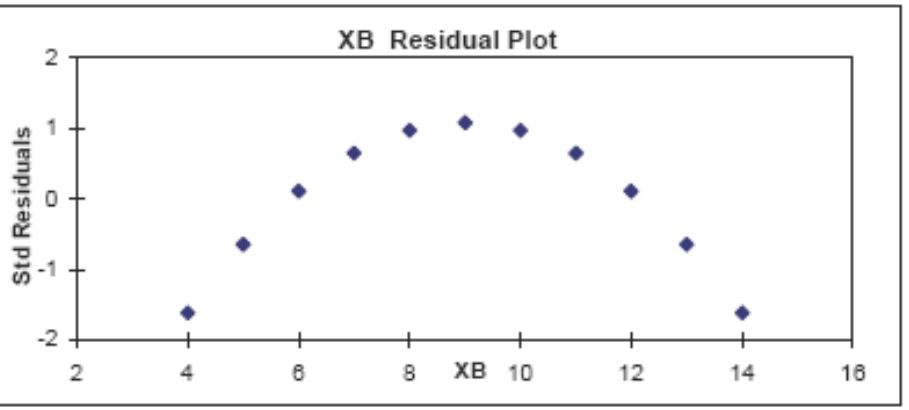
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	27.5	27.5	17.96564849	0.002178816
Residual	9	13.77629091	1.53069899		
Total	10	41.27629091			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	3.000909091	1.125302416	2.666757884	0.025758941	0.45529623	5.546521952	0.45529623	5.546521952
XB	0.5	0.117963746	4.23859039	0.002178816	0.233147264	0.766852736	0.233147264	0.766852736

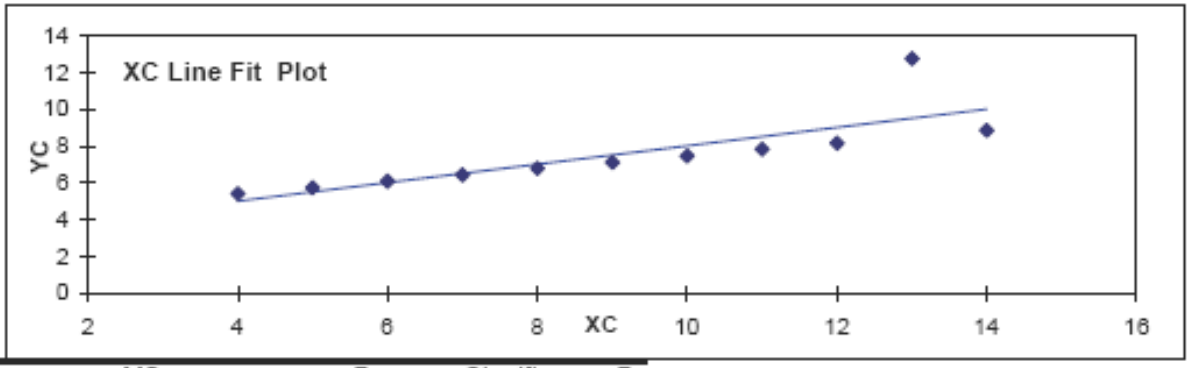
RESIDUAL OUTPUT

Observation	*Predicted YB*	*Residuals*	*Standard Residuals*
1	8.000909091	1.139090909	0.970492611
2	7.000909091	1.139090909	0.970492611
3	9.500909091	-0.760909091	-0.648285966
4	7.500909091	1.269090909	1.081251146
5	8.500909091	0.759090909	0.646736896
6	10.00090909	-1.900909091	-1.619553113
7	6.000909091	0.129090909	0.109983999
8	5.000909091	-1.900909091	-1.619553113
9	9.000909091	0.129090909	0.109983999
10	6.500909091	0.759090909	0.646736896
11	5.500909091	-0.760909091	-0.648285966



SUMMARY OUTPUT **C**

Regression Statistics	
Multiple R	0.816286739
R Square	0.666324041
Adjusted R Square	0.629248935
Standard Error	1.236311351
Observations	11



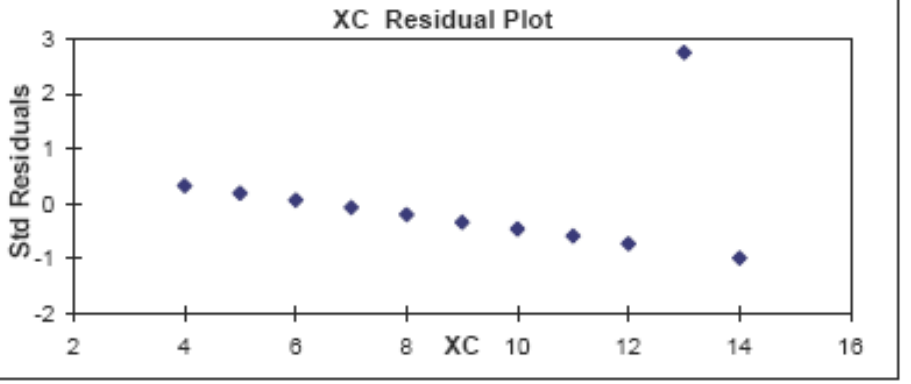
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	27.47000818	27.47000818	17.97227582	0.002176305
Residual	9	13.75619182	1.528465758		
Total	10	41.2262			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	3.002454545	1.12448123	2.670079737	0.025619109	0.458699339	5.546209752	0.458699339	5.546209752
XC	0.499727273	0.117877662	4.239372102	0.002176305	0.233069272	0.766385274	0.233069272	0.766385274

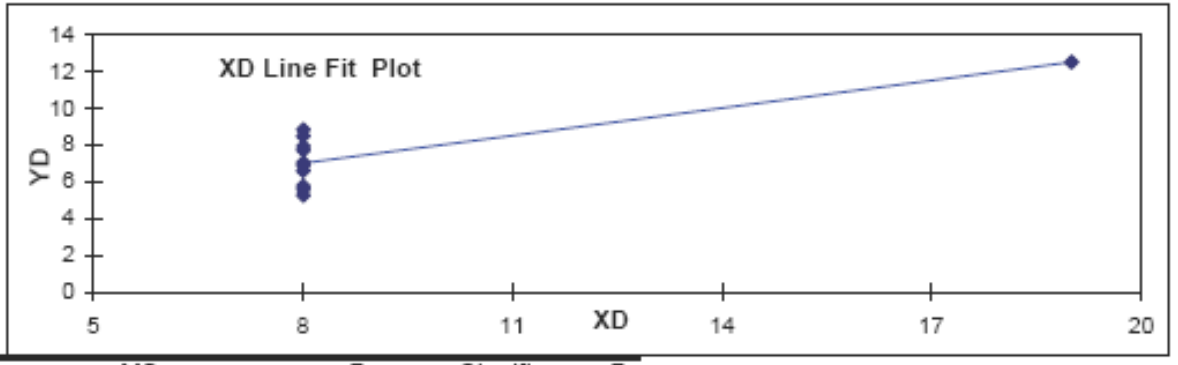
RESIDUAL OUTPUT

Observation	*Predicted YC*	*Residuals*	*Standard Residuals*
1	7.999727273	-0.539727273	-0.460177364
2	7.000272727	-0.230272727	-0.196333041
3	9.498909091	3.241090909	2.763389488
4	7.5	-0.39	-0.332518257
5	8.499454545	-0.689454545	-0.587836471
6	9.998636364	-1.158636364	-0.987866011
7	6.000818182	0.079181818	0.067511283
8	5.001363636	0.388636364	0.331355606
9	8.999181818	-0.849181818	-0.724021688
10	6.500545455	-0.080545455	-0.068673934
11	5.501090909	0.228909091	0.19517039



SUMMARY OUTPUT **D**

Regression Statistics	
Multiple R	0.816521437
R Square	0.666707257
Adjusted R Square	0.62967473
Standard Error	1.235695486
Observations	11



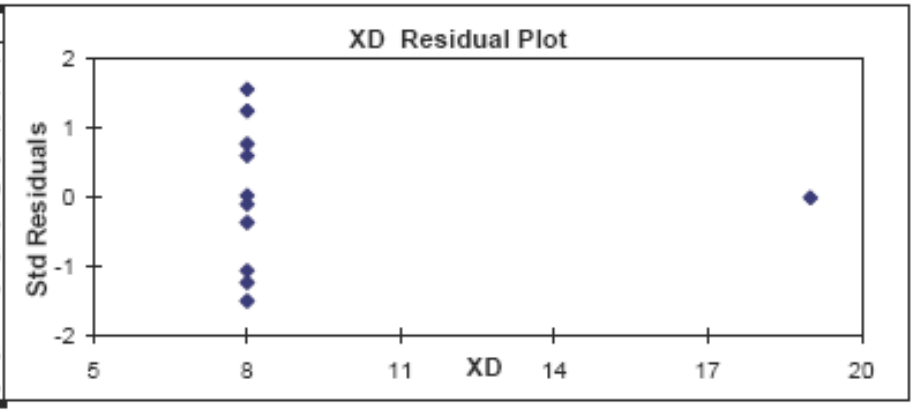
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	27.49000091	27.49000091	18.00328821	0.002164602
Residual	9	13.74249	1.526943333		
Total	10	41.23249091			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	3.001727273	1.123921072	2.670763408	0.025590425	0.459239232	5.544215314	0.459239232	5.544215314
XD	0.499909091	0.117818942	4.243028189	0.002164602	0.233383925	0.766434257	0.233383925	0.766434257

RESIDUAL OUTPUT

Observation	*Predicted YD*	*Residuals*	*Standard Residuals*
1	7.001	-0.421	-0.359128094
2	7.001	-1.241	-1.058617494
3	7.001	0.709	0.60480242
4	7.001	1.839	1.568732935
5	7.001	1.469	1.253109669
6	7.001	0.039	0.033268398
7	7.001	-1.751	-1.49366578
8	7.001	-1.441	-1.229224665
9	7.001	0.909	0.775409591
10	7.001	-0.111	-0.09468698
11	12.5	-3.55271E-15	-3.03059E-15



Algorithmes de data mining

1. But : visualisation, classification, description, prédiction, etc.
2. Forme du modèle : linéaire, non linéaire, hiérarchique, arbre, etc.
3. Fonction de score : moindres carrés, fonction de perte robuste, etc.
4. Méthode d'optimisation : locale, globale, combinatoire, aléatoire, etc.
5. Gestion des données

Directions et défis futurs

- Analyse d'images, sons, vidéos, etc.
- Analyse de réseaux sociaux
- Analyse bio-informatique, génétique, pharma
- Traitement de flux plutôt que de bases de données stockées
- Prouver un retour sur investissement
- Protection de la sphère privée

Questions ?

Exemple

- Choisir un secteur d'activité
  - Santé, bien-être
  - Alimentation, grande distribution
  - Médias, communication
  - Éducation
  - Infrastructures, construction
  - ...
- Décrire un scénario d'analyse de data mining