le cnam

Reconnaissance des formes et méthodes neuronales

Introduction

Michel Crucianu
http://idf.pleiad.net/index.php
http://cedric.cnam.fr/~crucianm/rfmn.html

le cnam

Objectifs

- Reconnaissance des formes (*pattern recognition*) = (sens strict) identifier à quelle catégorie appartient une « forme » décrite par des données brutes
  - « Forme » = observation (ou partie d'une observation, ou ensemble d'observations)
  - Exemples : valeurs de variables décrivant un état clinique ; partie correspondant à un visage dans une image ; ensemble des valeurs prises par le cours d'une action sur une journée
- Fouille de données (*data mining*) : recherche de régularités ou de relations inconnues *a priori* dans de grands volumes de données
  - Les méthodes du cours sont également utilisées dans ce domaine

⇒ Présenter les éléments de base des méthodes d'analyse et de modélisation des données pour traiter des applications réelles

le cnam

Contenu du cours

- Introduction
- Analyse en composantes principales
- Méthodes de régression, régression linéaire
- Théorie bayesienne de la décision
- Méthodes de discrimination paramétriques et non paramétriques
- Méthodes de classification automatique
- Arbres de décision
- Perceptrons multicouches
- Cartes auto-organisatrices de Kohonen
- Programmation dynamique
- Chaînes de Markov

le cnam

Bibliographie générale

J.-P. Asselin de Beauville, F. Kettaf, *Bases théoriques pour l'apprentissage et la décision en reconnaissance des formes*. Éditions Cépaduès, 2005.

M. Crucianu, J.-P. Asselin de Beauville, R. Boné, *Méthodes factorielles pour l'analyse des données : méthodes linéaires et extensions non-linéaires*. Hermès, 2004, 288 p.

G. Dreyfus, J. Martinez, M. Samuelides, M. Gordon, F. Badran, S. Thiria, *Apprentissage statistique : Réseaux de neurones - Cartes topologiques - Machines à vecteurs supports*. Éditions Eyrolles, 2008.

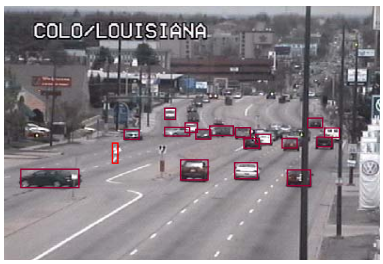
D. J. Hand, H. Mannila, P. Smyth, *Principles of Data Mining*. MIT Press, 2001.

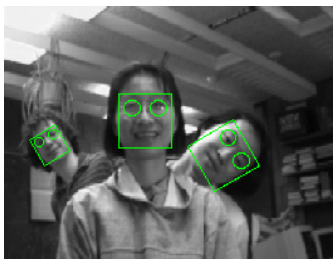
G. Saporta, *Probabilités, analyse des données et statistique*. 622 p., Éditions Technip, Paris, 2006.

le cnam

Exemple réel : détection d'objets

- Objectif de l'application : détecter des « objets » d'une certaine catégorie dans des images ou des vidéos





http://vasc.ri.cmu.edu/NNFaceDetector/

16 octobre 2014 RCP208 5

le cnam

Exemple réel : détection d'objets (2)

- Étapes de résolution (souvent plusieurs cycles) :
  - Déterminer quelles données devraient permettre de discriminer entre la classe visée et le « reste du monde »
    - Exemple : plus facile de détecter des objets en mouvement dans une vidéo (séparation facile des zones candidates du fond statique) que des objets dans une image
  - Récolter, analyser, nettoyer/corriger les données
  - Modéliser, à partir des données, la frontière de discrimination entre objet d'intérêt et autre chose
    - La construction d'un modèle n'est pas indispensable à la prise de décision, voir la méthode des *k* plus proches voisins
  - Décider, pour une région d'une nouvelle image, si elle contient ou non un objet d'intérêt

16 octobre 2014 RCP208 6

le cnam

Exemple illustratif simplifié*

- Objectif de l'application : sur des bateaux de pêche industrielle, séparer automatiquement les saumons des autres poissons pêchés
- Moyens matériels : système de vision et bras robotisé
- Hypothèses simplificatrices :
  - On considère que le système donne des valeurs à des variables décrivant l'apparence visuelle de chaque poisson et on cherche, à partir de ces valeurs, à prédire à quelle classe appartient le poisson
  - On ne s'intéresse pas ici au traitement des images,
  - Ni au contrôle du bras robotisé

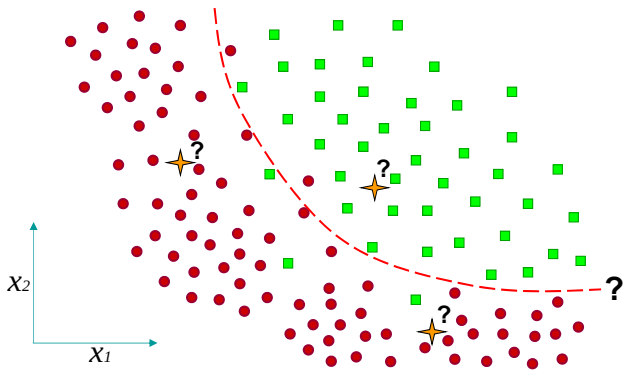
* Inspiré de : R.O. Duda, P.E. Hart, D.G. Stork, *Pattern Classification*, Wiley-Interscience, 2001.

16 octobre 2014 RCP208 7

le cnam

Exemple illustratif simplifié (2)

- **Pour faciliter l'illustration, on considère que 2 variables (x_1, x_2) suffisent à décrire un poisson**



16 octobre 2014 RCP208 8

le cnam

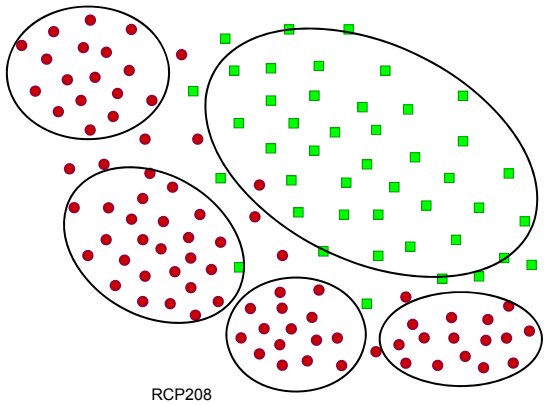
Types d'analyse/modélisation

- Exploratoire ↔ confirmatoire
  - Exploratoire : rechercher des régularités dans les données
  - Confirmatoire : répondre à des questions précises
- Descriptive ↔ décisionnelle
  - Descriptive : caractériser les données observées
  - Décisionnelle (prédictive) : aller au-delà des données observées

le cnam

Analyse/modélisation exploratoire

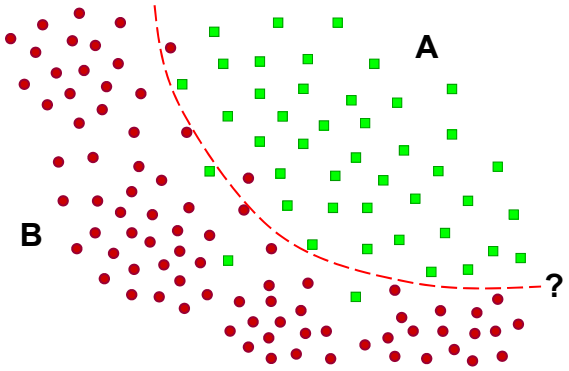
- Rechercher des régularités dans les données
  - Exemple : regroupements



le cnam

Analyse/modélisation confirmatoire

- Répondre à des questions précises
  - Exemple : peut-on séparer le groupe A du groupe B ?

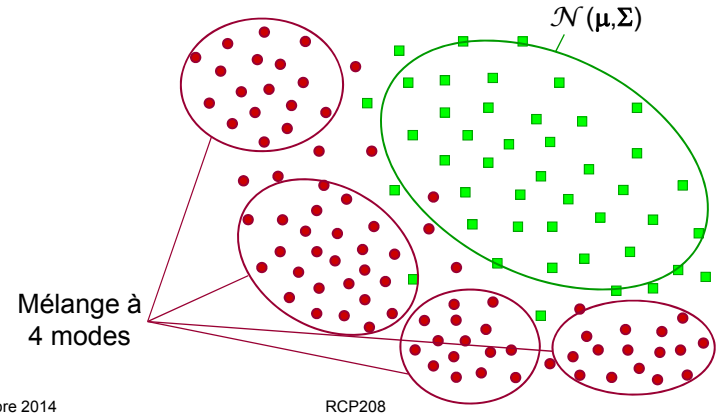


16 octobre 2014 RCP208 11

le cnam

Analyse/modélisation descriptive

- Caractériser les données observées
  - Exemple : que caractérise chaque classe de données ?

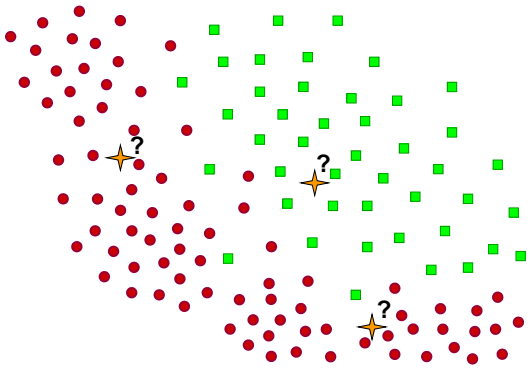


16 octobre 2014 RCP208 12

le cnam

Analyse/modélisation décisionnelle

- Prédire les valeurs de certaines variables à partir des autres variables
  - Exemple : à quelle classe appartient une nouvelle donnée ?



le cnam

La nature des données

- Les données : valeurs que prennent un ensemble de variables (attributs, traits...) pour un ensemble d'observations (objets, entités, individus, enregistrements…)
- Typologie
  - Quantitatives (numériques) ↔ Qualitatives (catégorielles)
    - Continues ↔ Discrètes
    - Ordinales ↔ Nominales

Ex. :	Ex. :	Ex. :	Ex. :
Longueur	Age	Classement,	Nom de marque
Durée	Population	Échelle de Lickert*	Catégorie socio-prof.
Température			
Concentration			

* Exemple d'échelle de Likert : Pas du tout d'accord / Pas d'accord / Ni en désaccord ni d'accord / D'accord / Tout à fait d'accord

le cnam

La nature des données

- Typologie (suite)
  - Séquentielles (ex. taux de change, débit rivière) ↔ non séquentielles
    - Séquentielles : en général, valeurs successives non indépendantes
  - Spatiales (ex. fertilité du sol) ↔ non spatiales
    - Spatiales : en général, valeurs spatialement proches non indépendantes
  - Structurées (ex. phrases) ↔ non structurées
    - Structurées : la structure est importante pour l'utilisation des données
- La nature des variables doit être en accord avec la nature des propriétés des objets que les variables représentent !
  - Question : quelles conséquences si on représente par des valeurs numériques
    - des modalités (par ex. catégorie socio-professionnelle) ?
    - des catégories ordonnées (par ex. échelle de Lickert) ?

16 octobre 2014 RCP208 15

le cnam

La réalité des données

- Données inadaptées
- Données non représentatives
- Données affectées de bruit
- Données mal enregistrées
- Données aberrantes (*outliers*)
- Données manquantes
- Malédiction de la dimension

⇒ Indispensable de « se familiariser avec les données » par des études exploratoires préalables, visualisation...

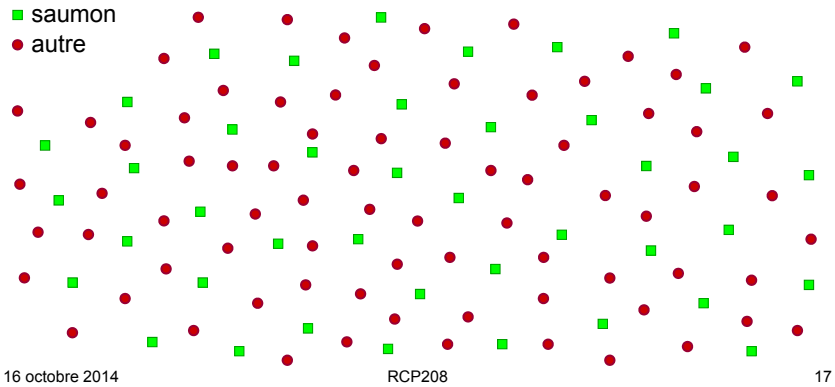
16 octobre 2014 RCP208 16

le cnam

Données inadaptées

- Variables importantes absentes ⇒ on ne peut pas répondre à la question posée (l'information utile est absente des données)

⇒ Refaire une collecte de données après une analyse du problème



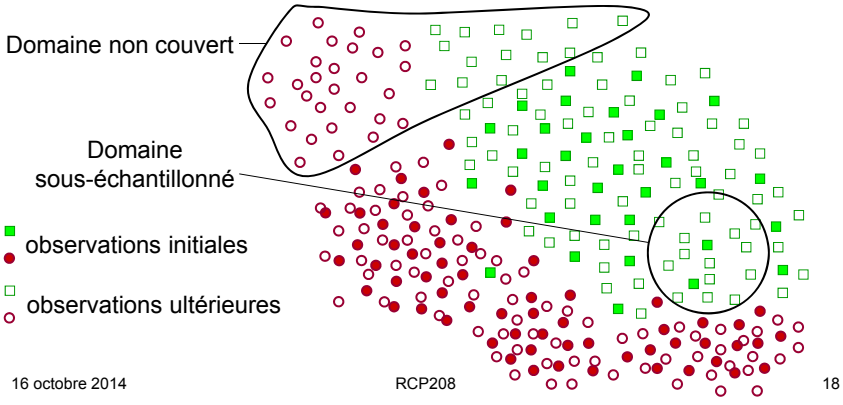
16 octobre 2014 RCP208 17

le cnam

Données non représentatives

- Couverture partielle du domaine ou problème d'échantillonnage ?
- Échantillonnage : souvent de convenance ou opportuniste

⇒ Comprendre comment les données ont été collectées



16 octobre 2014 RCP208 18

le cnam

Données non représentatives (2)

- Approches :
  - Restreindre le modèle aux régions avec un bon échantillonnage
  - Employer des méthodes de modélisation permettant d'obtenir des intervalles de confiance
  - Compléter la collecte de données pour mieux couvrir le domaine visé

le cnam

Données manquantes

- Pour certaines observations, les valeurs de certaines variables peuvent manquer
  - Exemple (pêche) : certains capteurs ne répondent pas
  - Exemple (sondages) : pas de réponse à certaines questions
- Les observations à valeurs manquantes peuvent être éliminées, mais dans certains cas il peut être pertinent de remplacer une valeur manquante par une estimation
  - Remplacement par la moyenne de la variable
  - Remplacement par un prototype de *cluster* (résultant d'une classification automatique tenant compte des valeurs de toutes les variables)
  - Utilisation d'algorithmes de type EM estimant en parallèle, par itérations successives, le modèle et les données manquantes (voir RCP209)

le cnam

Données manquantes (2)

- Le fait qu'une donnée soit manquante n'est pas toujours indépendant de la valeur qu'aurait prise cette donnée
  - Exemple (pêche) : réponse des mesures optiques liée à la présence de brouillard et/ou réponse en température des capteurs, à leur tour liées à la région géographique et donc aux spécificités des données
  - Exemple (sondages) : données manquantes expliquées par des réticences à donner certaines réponses à certaines questions

→ les méthodes simples de remplacement biaisent les résultats

⇒ Pour employer une méthode plus élaborée il est important de comprendre pourquoi des valeurs manquent

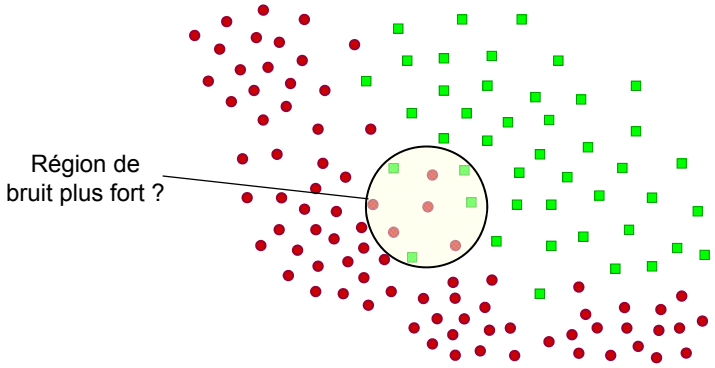
16 octobre 2014 RCP208 21

le cnam

Données affectées de bruit

- Le niveau du bruit qui affecte des données dépend parfois de leurs valeurs

⇒ Identifier les sources de bruit et leurs spécificités

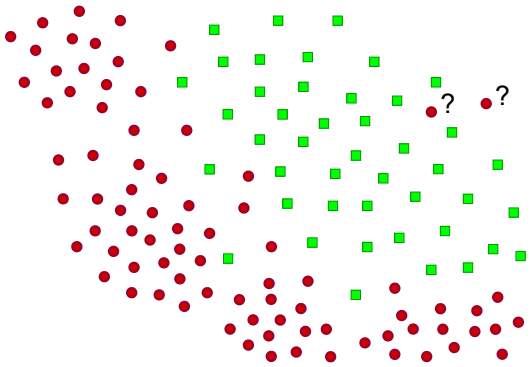


16 octobre 2014 RCP208 22

le cnam

Données aberrantes (*outliers*)

- Données éloignées de toutes les autres ou seulement de celles de leur classe



16 octobre 2014 RCP208 23

le cnam

Données aberrantes (2)

- Approches :
  - Ignorer les données aberrantes :
    - De façon implicite : méthodes robustes (par ex. modèles basés sur des distributions à queues longues)
    - De façon explicite : détection suivie de suppression
  - Comprendre les données aberrantes :
    - Comment les valeurs aberrantes ont été obtenues → possibilité de correction des valeurs aberrantes ?
    - Classes ayant plusieurs modes éloignés ?
    - Explications liées à la compréhension du problème ?

⇒ Erreurs d'enregistrement, bruit ou phénomène significatif ?

16 octobre 2014 RCP208 24

le cnam

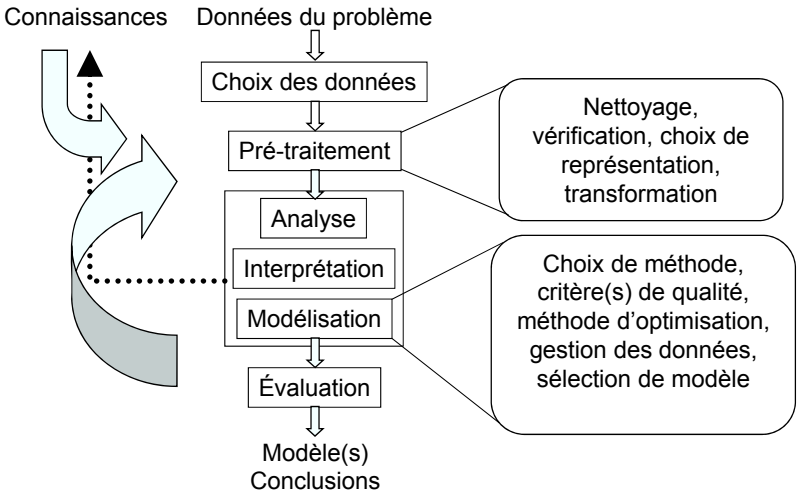
Malédiction de la dimension

(*curse of dimensionality*)

- Dimension *d* des données : nombre de variables unidimensionnelles qui caractérisent les données
- Quelques difficultés engendrées par *d* élevé
  - Complexité algorithmique proportionnelle au carré (et parfois à la puissance 3) de la dimension *d*
    → coût excessif
  - *n* points (*n* fixé) dans un hypercube de côté *L* en dimension *d* : la densité des points diminue exponentiellement avec *d*
    → estimation de densité impossible au-delà d'une certaine valeur pour *d*
  - Distribution uniforme de *n* points dans un hypercube de côté *L* en dimension *d* : la variance des distances entre points diminue avec *d*
    → décision peu fiable sur la base des *k* plus proches voisins

le cnam

Quelle approche générale ?



le cnam

Choix de la méthode : qualité

- Précision : erreur minimale (en test), taux minimal de faux positifs…
  - Cette performance ne doit pas être l'unique critère de choix !
- Lisibilité (~*explicability*) : résultats/décisions interprétables
  - Pour des applications critiques (par ex. contrôle réaction chimique à risque) on ne peut se contenter d'une solution « boîte noire »
  - La lisibilité rend possible la vérification/validation *a priori*
  - Solutions pour rendre lisibles des modèles qui ne le sont pas *a priori* (par ex. extraction de règles d'un réseau de neurones)
- Rapidité de la construction du modèle, de la prise de décision
  - Contraintes de temps sur la (re)construction du modèle ou sur la prise de décision ?
- Autres critères spécifiques à l'application

le cnam

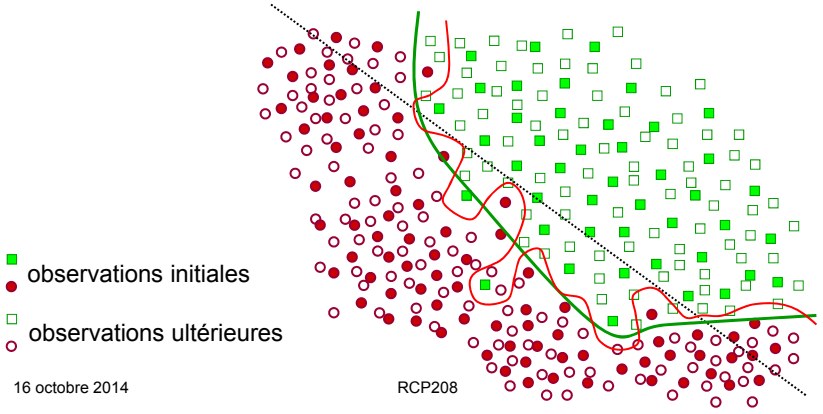
Choix de la méthode : ingénierie

- *Usability* : facilité d'emploi
  - Exemple : un expert est indispensable pour mettre au point le modèle et pour toute évolution ultérieure ?
- *Embedability* : facilité d'introduction dans un système global
  - Exemple : la méthode impose des contraintes fortes sur l'échange de données (par ex. codage des données d'entrée, de sortie) ?
- Flexibilité : adaptation facile au changement de spécifications
  - Exemples : faut-il repartir de zéro si les conditions de mesure changent ou si un capteur est remplacé par un autre de courbe de réponse différente ?
- Passage à l'échelle (*scalability*)
  - Exemple : gros volumes/débits de données

le cnam

Décision et généralisation

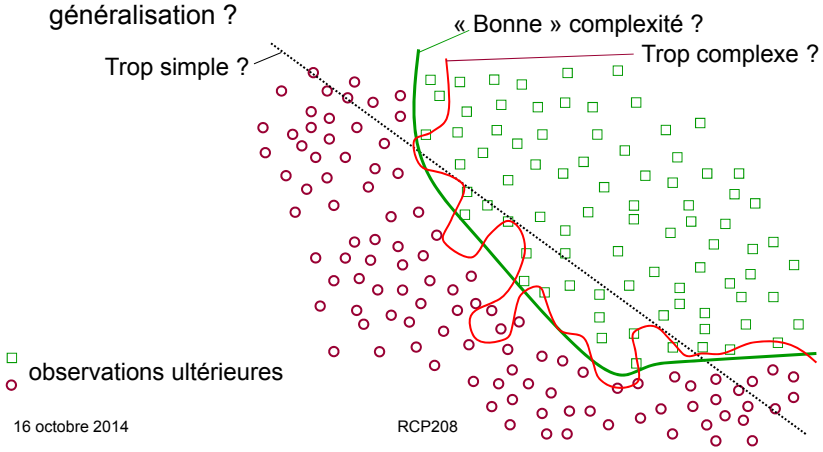
- Modèle mis au point sur les observations initiales (d'apprentissage)
- Utilisation pour décision si bonne généralisation (modèle « valable » au-delà des observations initiales)



le cnam

Décision et généralisation (2)

- Hypothèse : même distribution pour apprentissage et test
- Peut-on borner l'écart entre l'erreur d'apprentissage et l'erreur de généralisation ?



le cnam

Généralisation et choix de modèle

- Complexité des modèles
  - Faible : les modèles peuvent s'avérer insuffisants pour modéliser de façon satisfaisante les observations initiales (ou « données d'apprentissage ») ⇒ on peut les rejeter sur cette base
  - Élevée : le problème de détermination des paramètres du modèle est mal posé, il y a de nombreuses solutions équivalentes sur les observations initiales ⇒ des critères externes sont nécessaires (connaissances *a priori* du problème, rasoir d'Occam…)
- Un cas extrême : famille très riche de modèles très complexes, observations initiales très peu nombreuses ⇒ on trouve un « bon » modèle (sur les observations initiales…) pour tout ! (*data dredging*)
  - Exemple : prévision quasi-parfaite des moyennes annuelles du S&P500 à partir de la production de beurre, de fromage et de viande de mouton du Bangladesh et des États-Unis dans les années précédentes… mais sur les données d'apprentissage seulement !