L' Apprentissage Artificiel

face aux données temporelles

Antoine Cornuéjols

AgroParisTech – INRA MIA 518

antoine.cornuejols@agroparistech.fr

Plan

Apprentissage Artificiel

et son analyse statistique

Supervisé vs. Non supervisé

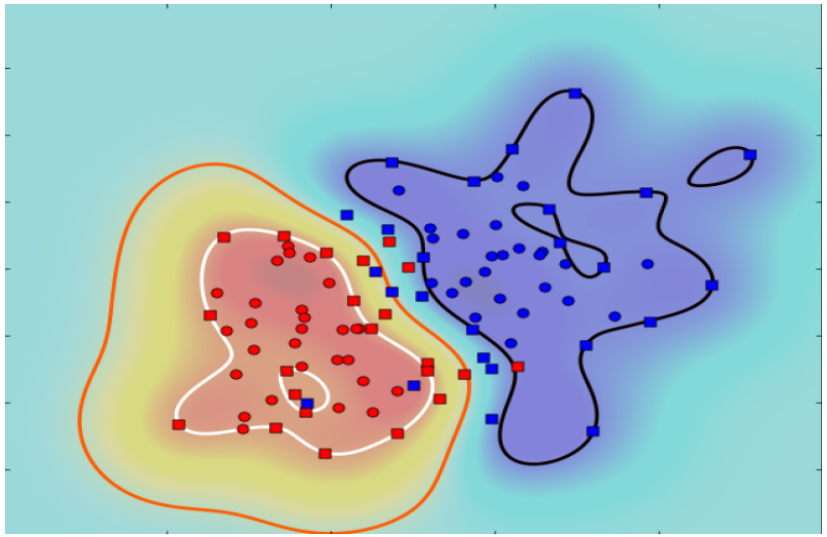
Échantillon d'apprentissage

$$S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i), \ldots, (x_m, y_m)\}$$

f : $x_i \rightarrow y_i$

h : $x_i \rightarrow y_i$

→ Fonction de décision h



Les données : organisation et types

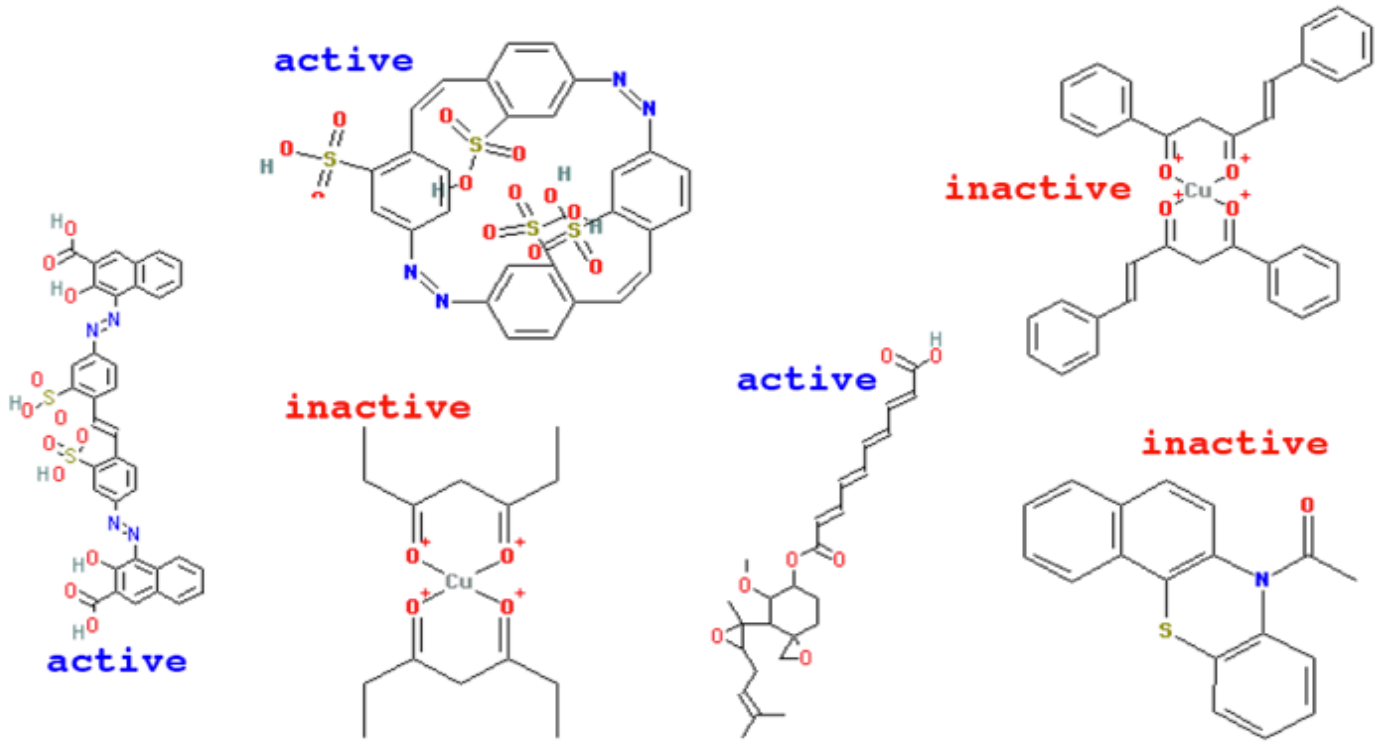
Identifieur	Genre	Age	Niveau études	Marié ?	Nb enfants	Revenu	Profession	A prospecter ?
I_21	M	43	Bac+5	Oui	3	55 000	Architecte	OUI
I_34	M	25	Bac+2	Non	0	21 000	Infirmier	NON
I_38	F	34	Bac+8	Oui	2	35 000	Chercheuse	OUI
I_39	F	67	Bac	Oui	5	20 000	Retraitée	NON
I_58	F	56	CAP	Oui	4	27 000	Ouvrière	NON
I_73	M	40	Bac+3	Non	2	31 000	Commercial	OUI
I_81	F	51	Bac+5	Oui	3	75 000	Chef d'entreprise	OUI

Exemple
(*example, instance*)

Descripteur
Attribut
(*feature*)

Étiquette
(*label*)

Apprentissage supervisé

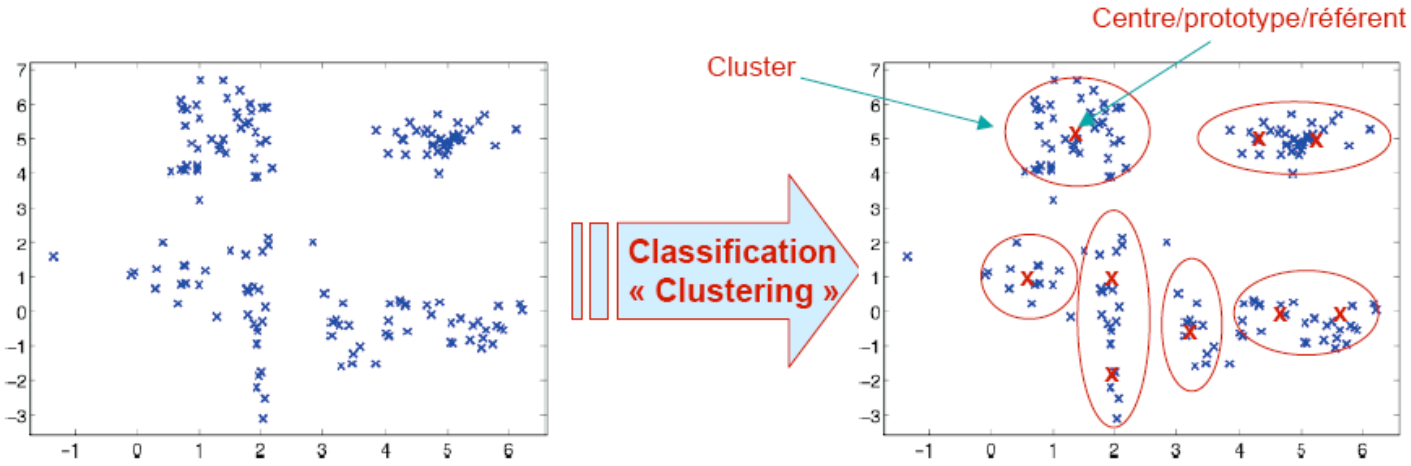


NCI AIDS screen results (from http://cactus.nci.nih.gov).

Supervisé vs. **Non** supervisé

Échantillon d'apprentissage

$$S = \{(x_1), (x_2), \ldots, (x_i), \ldots, (x_m)\}$$



Apprentissage **supervisé** : analyse statistique

$$R_{Emp}(h) = \sum_{i=1}^{m} l\left(h(\boldsymbol{x}_i), u_i\right)$$

- Principe :
  - On se donne $\mathcal{H}$
  - On cherche $h \in \mathcal{H}$ faisant peu d'erreur sur S
  - En espérant que h **fera peu d'erreur** sur des exemples à venir

$$R(h) = \int_{X \times Y} l\left(h(\boldsymbol{x}), u\right) P(\boldsymbol{x}, y)\, d\boldsymbol{x}dy$$

Fonction de perte

Étiquette prédite

Étiquette vraie (ou désirée)

Loi de probabilité jointe sur $\mathcal{X} \times \mathcal{Y}$

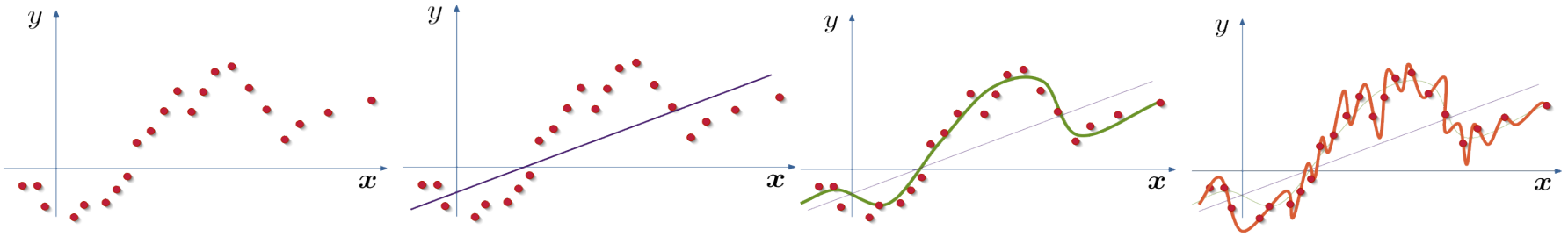
Apprentissage supervisé : analyse statistique

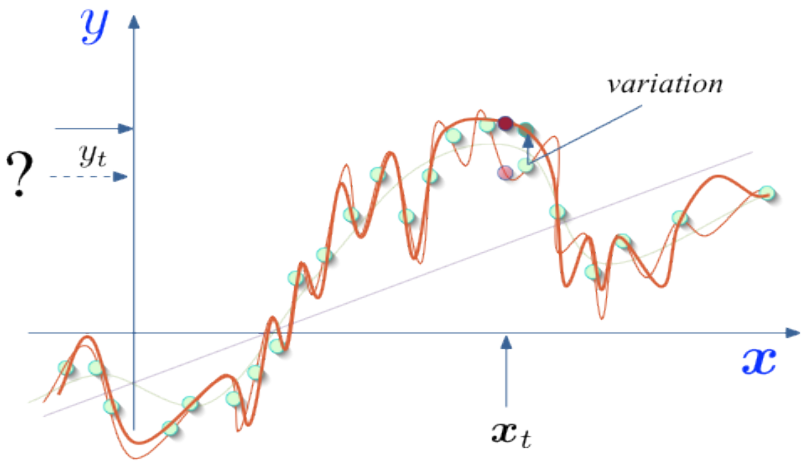
$$\forall h \in \mathcal{H} : \quad P^m \left[R(h) > R_{Emp}(h) + \mathcal{G}(\mathcal{H}, m, \delta) \right] < 1 - \delta$$

- Vrai si
  - Données i.i.d.
  - Environnement stationnaire
- Leçon essentielle :
  - Contrôler la « capacité » de $\mathcal{H}$

Apprentissage supervisé : le cas de la régression

$$h : \quad \mathbb{R}^d \rightarrow \mathbb{R}$$





Apprentissage supervisé : le cas de la régression

- Pas de vraie notion de temps
- Pas de notion de corrélation temporelle
  - Pas de notion de mémoire à contrôler
  - Tous les points sont équivalents

Tâches d'analyse de séquences (temporelles)

- **Génome**
  - Pas temporel
- **Prédiction de « clics »**
  - Temporel
- **Analyse de la parole**
  - Temporel
  - En temps réel

Apprentissage artificiel : bilan

- Prégnance de l'informatique
  - **Vocabulaire** de description **réduit**
  - Accent mis sur la **représentation** : choix de $\mathcal{H}$
- Analyse théorique
  - Surtout **apprentissage supervisé**
  - Données **i.id.**
  - Environnement **stationnaire**

Données temporelles

et tâches associées

Données temporelles : **une** séquence

- $\ldots, X_{t-n}, \ldots, X_{t-2}, X_{t-1}, X_t \ldots$
  - Séquence finie ou non
  - Les X_i peuvent être structurés
    - $X_i \in$ {symboles}
    - $X_i \in \Re$
    - X_i = vecteur
    - X_i = graphe
  - Pas d'hypothèse sur un processus sous-jacent

Données temporelles : **une** séquence

- Caractéristique
  - Les X_i ne sont pas i.i.d.

→ 2 approches

1. **La séquence** est considérée comme **un objet**
   - On en cherche sa structure (i.e. régression)

2. **La séquence** est une **suite non i.i.d.** d'objets
   - Renouvelle l'analyse de l'apprentissage artificiel
   - Algorithmique « à la volée » (en-ligne)

Données temporelles : **une** séquence

- Tâches

1. **Prédiction** de la suite de la séquence (« *forecasting* »)

2. **Analyse** de la séquence
   - Pour interprétation
   - Indexation
   - …

3. **Apprentissage en-ligne** (« *on-line learning* »)
   - (X_{t-n}, Y_{t-n}), …, (X_{t-2}, Y_{t-2}), (X_{t-1}, Y_{t-1}), (X_t, Y_t) … (X_{t+1}, Y_{t+1}) ?
   - Changement de la règle de décision ou non
     (Environnement stationnaire ou non)

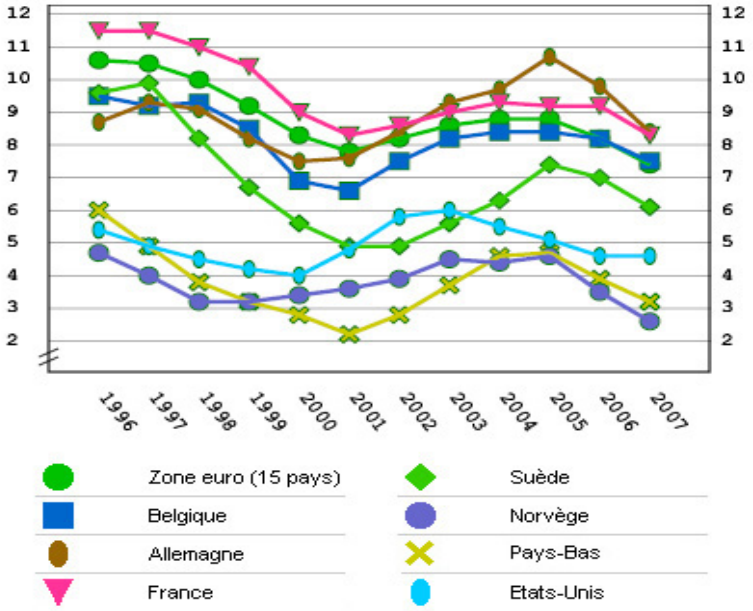
4. Recherche de **sous-séquences** fréquentes ou atypiques

Données temporelles : un **ensemble** de séquences

- Indexées ou non par les mêmes instants
- De même longueur ou non

$$\mathcal{V} = \{v^n_{1:T_n}, n = 1, \ldots, N\}$$

- E.g.
  - Cours de bourses
  - Consommations électriques
    - en différents lieux
    - en différentes saisons



Données temporelles : un **ensemble** de séquences

- Tâches

1. Classification supervisée
   - **Classer une nouvelle séquence** dans une classe connue de séquences
   - Éventuellement pour faire des **prédictions** sur l'avenir

2. Catégorisation (« *clustering* »)
   - Trouver des **catégories** de séquences

3. Requêtes
   - Chercher les séquences proches d'une séquence requête

Quels modèles

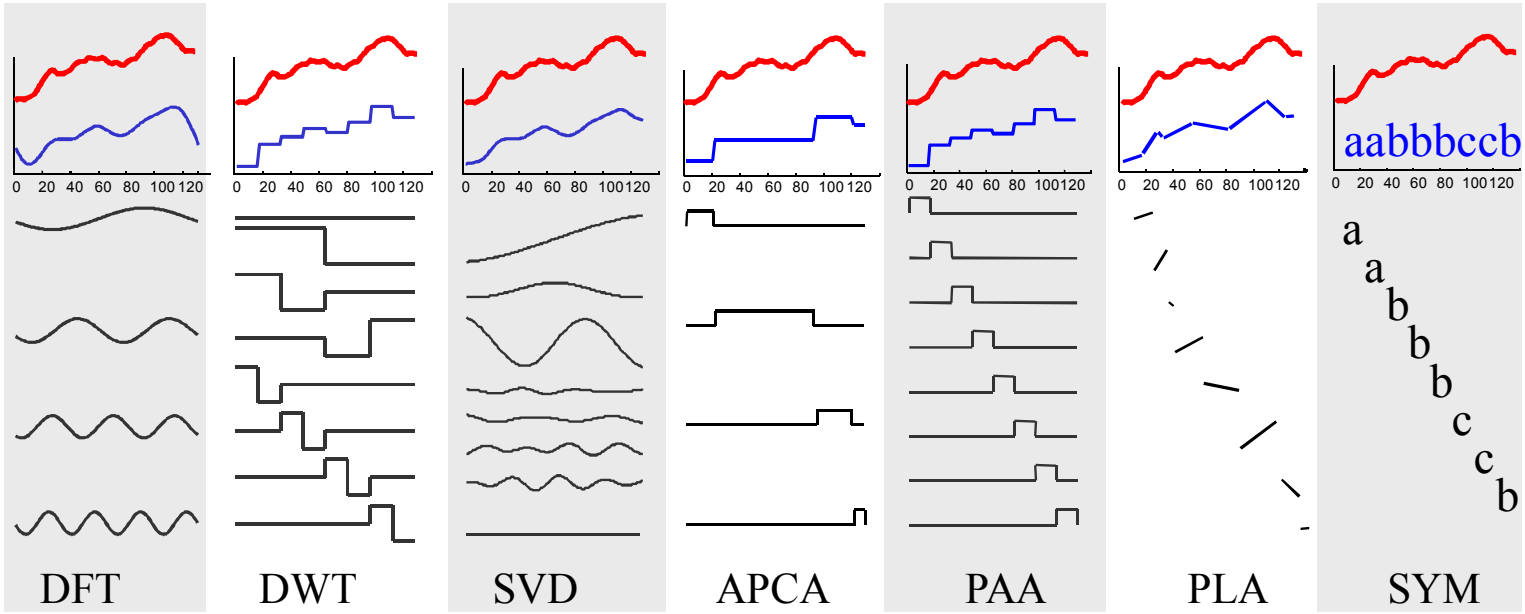
pour les séries temporelles ?

La représentation des séquences

1. **Brute**
   - Après prétraitements
2. Représentations « **analytiques** »
   - Par **combinaison de composantes** choisies dans un dictionnaire
   - Composantes orthogonales
3. Modélisation dans une **classe d'hypothèses**
   - Grammaires
   - Chaînes de Markov
   - Réseaux de neurones
   - …

Représentations analytiques

- **Transformée de Fourier** discrète
- Transformée en **ondelettes** discrète
- **Approximation agrégée** par morceaux
- …



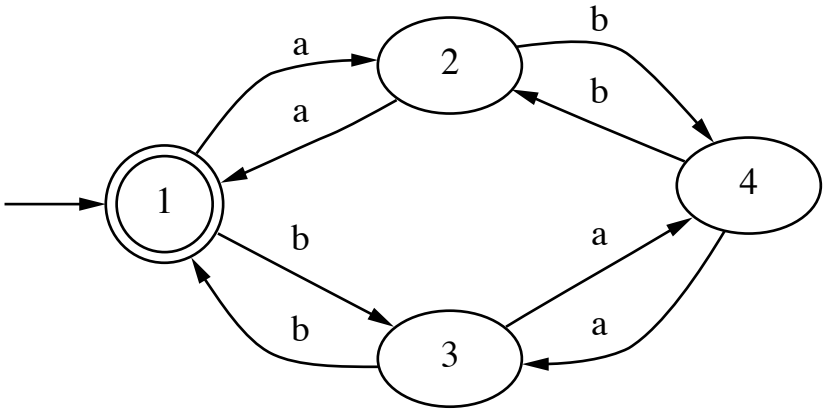
Représentation vectorielle

Modélisations des séries temporelles

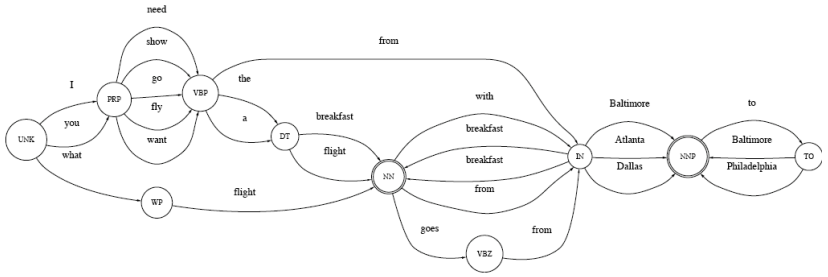
- « Classiques »
  - Grammaires
  - Chaînes de Markov
  - …
- Moins classiques
  - Réseaux de neurones récurrents
  - « Reservoir computing »
  - …

Grammaires (automates)

a b a a a b b a a b b a b b a b b a …

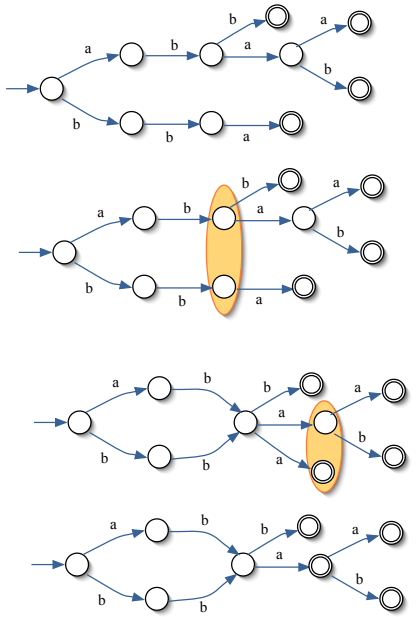


- Questions
  - Automate canonique ? (unicité ?)
  - Apprentissage ?
  - Et si bruit ?
  - Identification des états



Grammaires (automates)

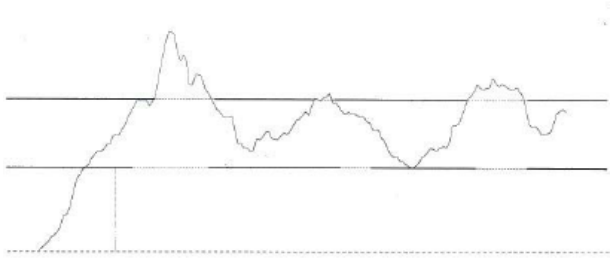
- **Apprentissage** de grammaires
  - À partir d'un échantillon de séquences positives (et négatives)
  - Algorithmes d'inférence grammaticale



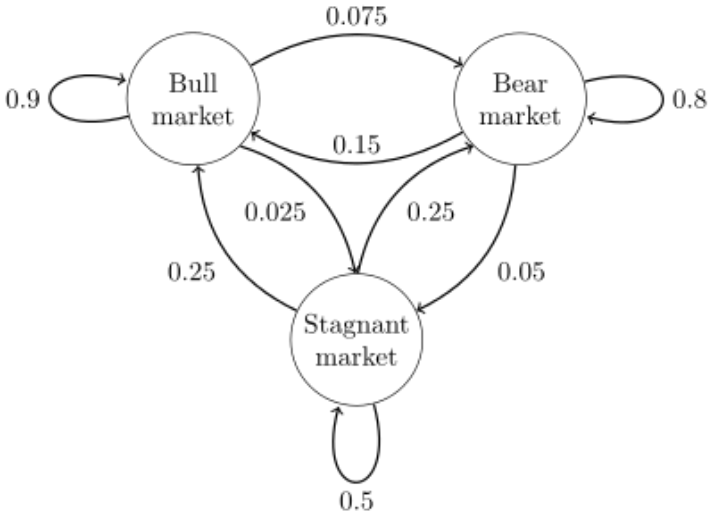
Peut servir à la prédiction

Chaînes de Markov

- Introduction de probabilités sur les transitions
  - Meilleure tolérance au bruit
  - Apprentissage de matrice de transition



- **Suppose**
  - Processus stationnaire
  - Mémoire à un état

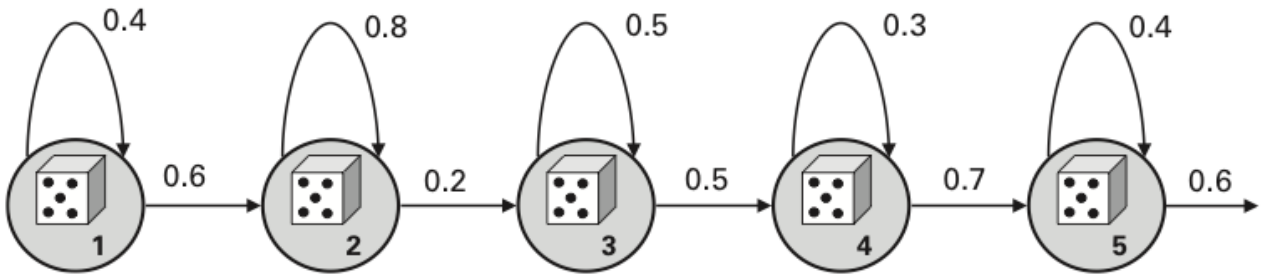


HMM (Hidden Markov Chains)

1 3 4 5 3 4 6 2 4 5 2 4 5 1 1 1 2 4 5 3 2 5 2 2 5 4 3

Calculer la séquence la plus probable d'états cachés (dés pipés) ayant engendré cette séquence de tirages

E.g. On sait que la séquence est issue du tirage par **5 dés** successivement, chacun étant biaisé d'une manière spécifique



Outcomes: 1 3 4 5 3 4 6 2 4 5 2 4 5 1 1 1 2 4 5 3 2 5 2 2 5 4 3

States: ← 1 →← 2 →← 3 →← 4 →← 5 →

Probability: 0.0001245

Modèles de Markov

- **Apprentissage**
  - À partir d'un ensemble de séquences
  - Par algorithme EM (Expectation-Maximization)

- **Peut servir à** :
  - Engendrer une séquence
  - Prédire la suite
  - Inférer la séquence d'états cachées la plus probable étant donnée une séquence d'observations

Analyse

de séquences individuelles

Plan

1. L'**apprentissage artificiel** et son analyse statistique
2. **Données temporelles** et tâches associées
    1. Tâches
    2. Modélisations des séquences
3. Analyses de **séquences individuelles**
    1. Compréhension
    2. Prédiction
    3. Apprentissage en-ligne
    4. Identification de sous-séquences
4. Analyses d'**ensemble de séquences**
    1. Clustering
    2. Classification supervisée
5. **Conclusions**

Analyser pour comprendre

Quelles données temporelles ?

- Analyse de séquences génomiques
  - Analyse de la structure
  - Classification introns vs. Exons

- Textes / documents
  - Abstract ; introduction ; …
  - Types d'arguments

- Séquences vidéo
  - Journal télévisé ; pub ; …

Analyse / compréhension d'une séquence

1. Méthodes analytiques
2. Analyse en Comp. Indep. (ICA)

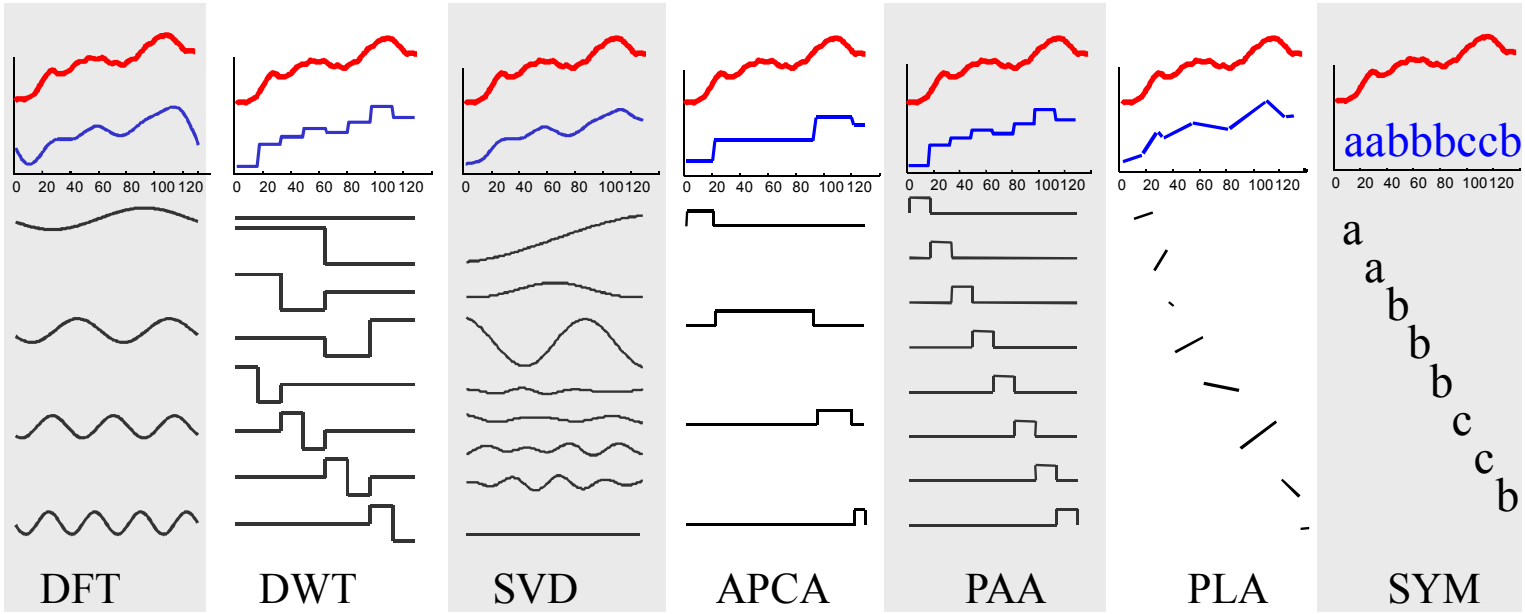
Données continues

3. Recherche de structures
   - Dans classe de modèles
     - Grammaires
     - Modèles de Markov

Données discrètes

Représentations analytiques

- **Transformée de Fourier** discrète
- Transformée en **ondelettes** discrète
- **Approximation agrégée** par morceaux
- …

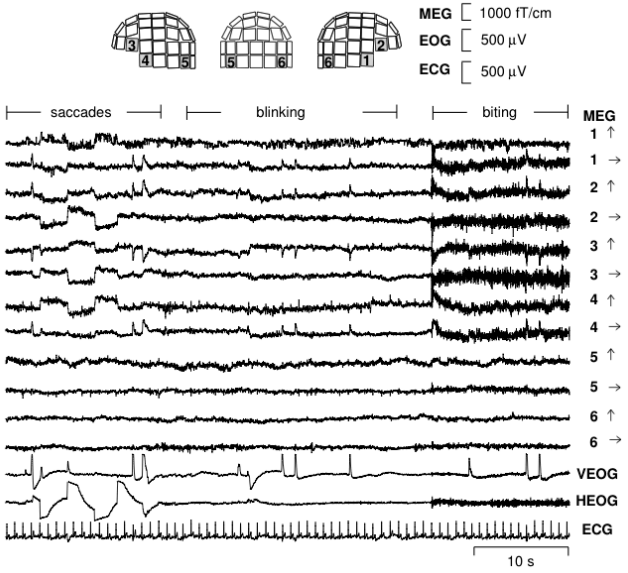


Représentation vectorielle

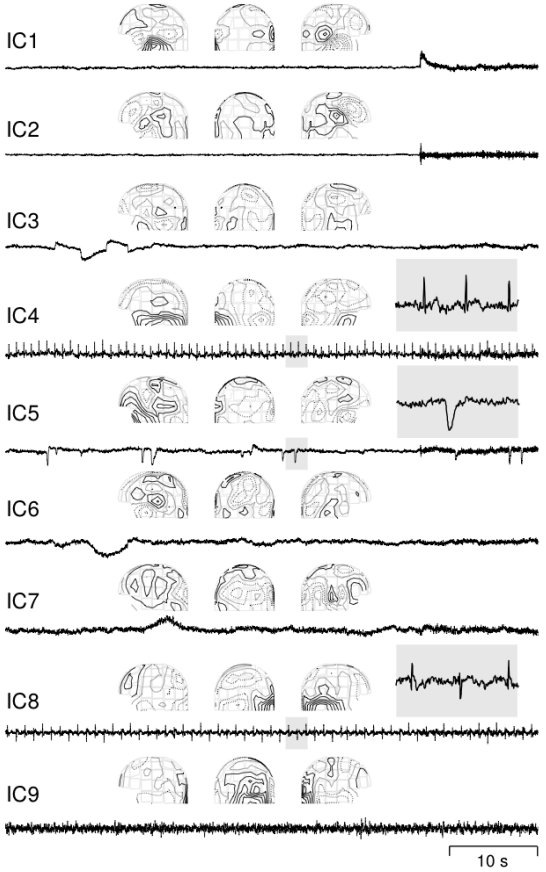
Analyser la séquence

- ICA

  (Independent Component Analysis)



(From Vigário et al, 1998). *Samples of MEG signals, showing artifacts produced by blinking, saccades, biting and cardiac cycle. For each of the 6 positions shown, the two orthogonal directions of the sensors are plotted.*



(From Vigário et al, 1998). *Nine independent components found from the MEG data. For each component the left, back and right views of the field patterns generated by these components are shown — full line stands for magnetic flux coming out from the head, and dotted line the flux inwards.*

Prédiction de la suite de la séquence

Quelles données temporelles ?

- Prédiction de cours boursiers
  - Par extrapolation
  - Par comparaison de courbes et classification supervisée ou non
- Reconnaissance de la parole
  - Non i.i.d.
  - Non stationnaire
  - Traitement « à la volée »
  - Recherche et identification de sous-séquences
  - Grande variabilité intra et inter-locuteur
  - Données supervisées disponibles

Prédiction de la suite de la séquence

- Méthodes classiques linéaires et non linéaires
  - Autoregressive : ARMA ; ARIMA ; …
  - ARCH, GARCH, …

- Autres méthodes : linéaires et non linéaires
  - SVM
  - TDNN
  - Reservoir computing
  - …

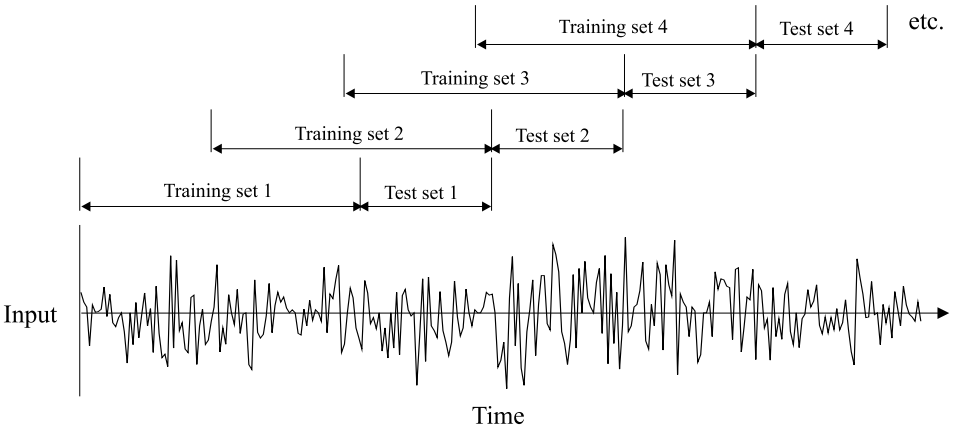
SVM et **prédiction** de séquence

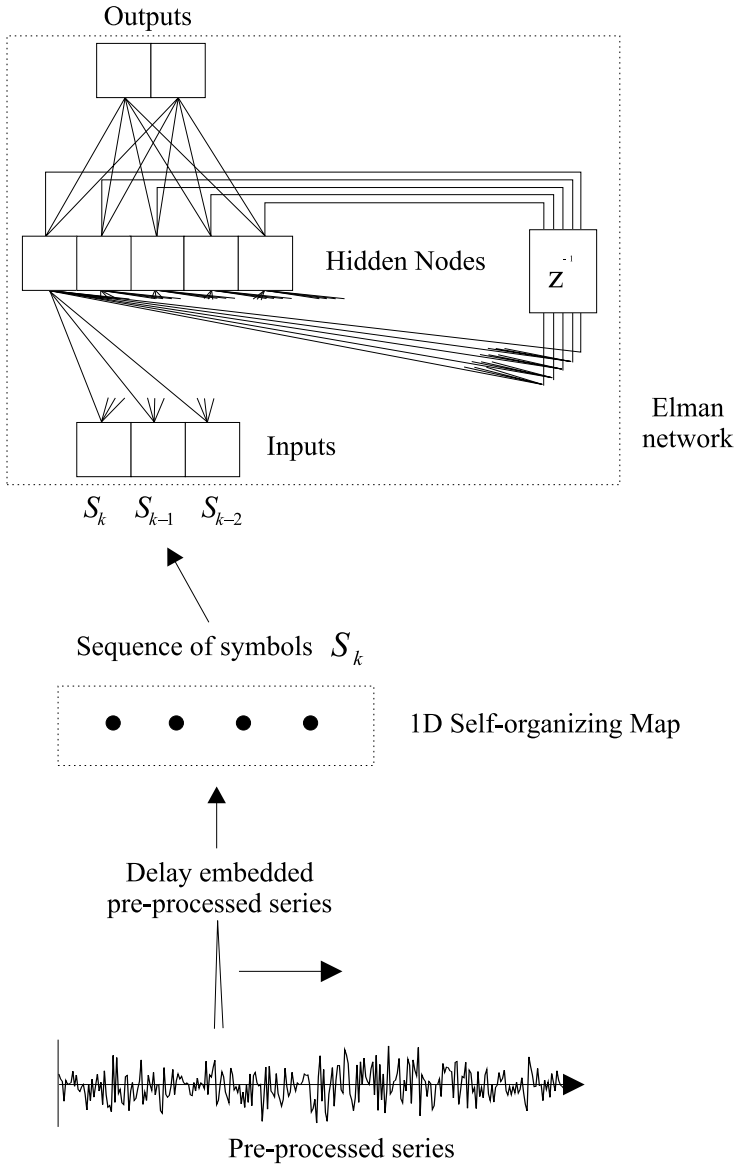
- **Motivation**
  - Problème : représentation en **grande dimension** : « curse of dimensionality »
  - Mais **les SVM ne sont** (presque) **pas sensibles à la dimension** de l'espace

- **Idée**
  - Prendre les entrées sur une fenêtre glissante $\{X_{t-n}, X_{t-n+1}, \ldots, X_{t-1}\}$
  - Pour prédire X_t ou $X_{t+\tau}$
  - SVM linéaire ou non
  - Problèmes :
    - Les $X_{t-n}, X_{t-n+1}, \ldots, X_{t-}$ ne sont pas i.i.d. (pas conforme à la théorie)
    - Réglage de n (prise en compte du passé)
    - Choix du noyau

- Performances obtenues
  - Bonnes
  - Mais pas bien meilleures que modèles AR

TDNN (« Time Delay Neural Networks »)





Étapes

1. Prétraitements
2. Discrétisation (SOM)
3. TDNN (Elman)
4. Extraction d'automates
5. Traduction en règles

Prédiction : « reservoir computing »

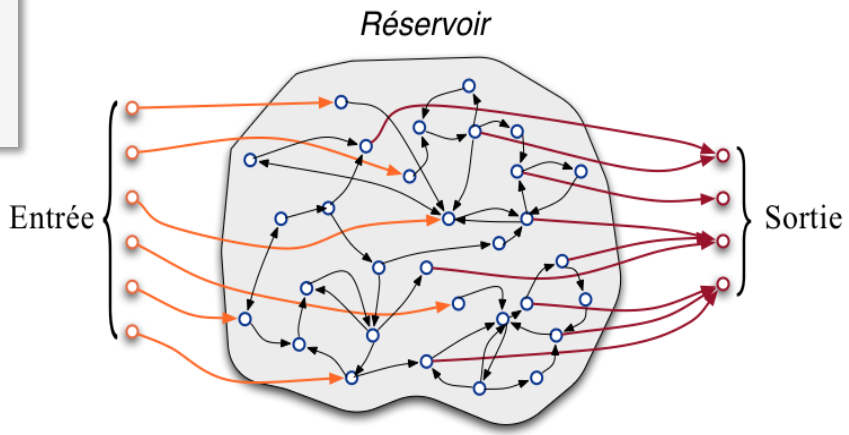
Les RN récurrents sont (très) difficiles à entraîner (par gradient)

Principe

- Utilisation d'un réseau récurrent **non adaptatif !!**
- **Apprendre** une **couche de sortie** linéaire

Performances

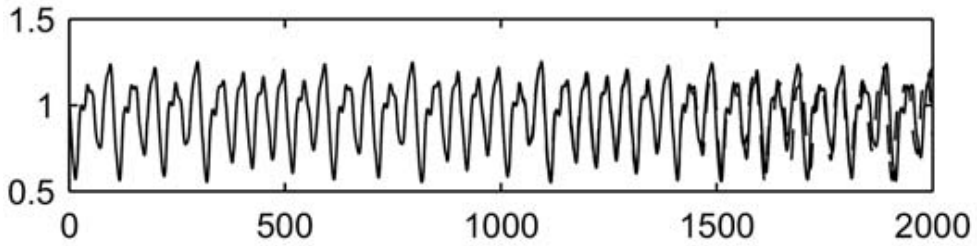
- 1000 fois plus précis sur prédiction de certains systèmes dynamiques
- 1er sur tâches diverses
  - Reconnaissance de voyelles en japonais
  - Reconnaissance de chiffres parlés
  - Prédiction de cours boursiers
  - …

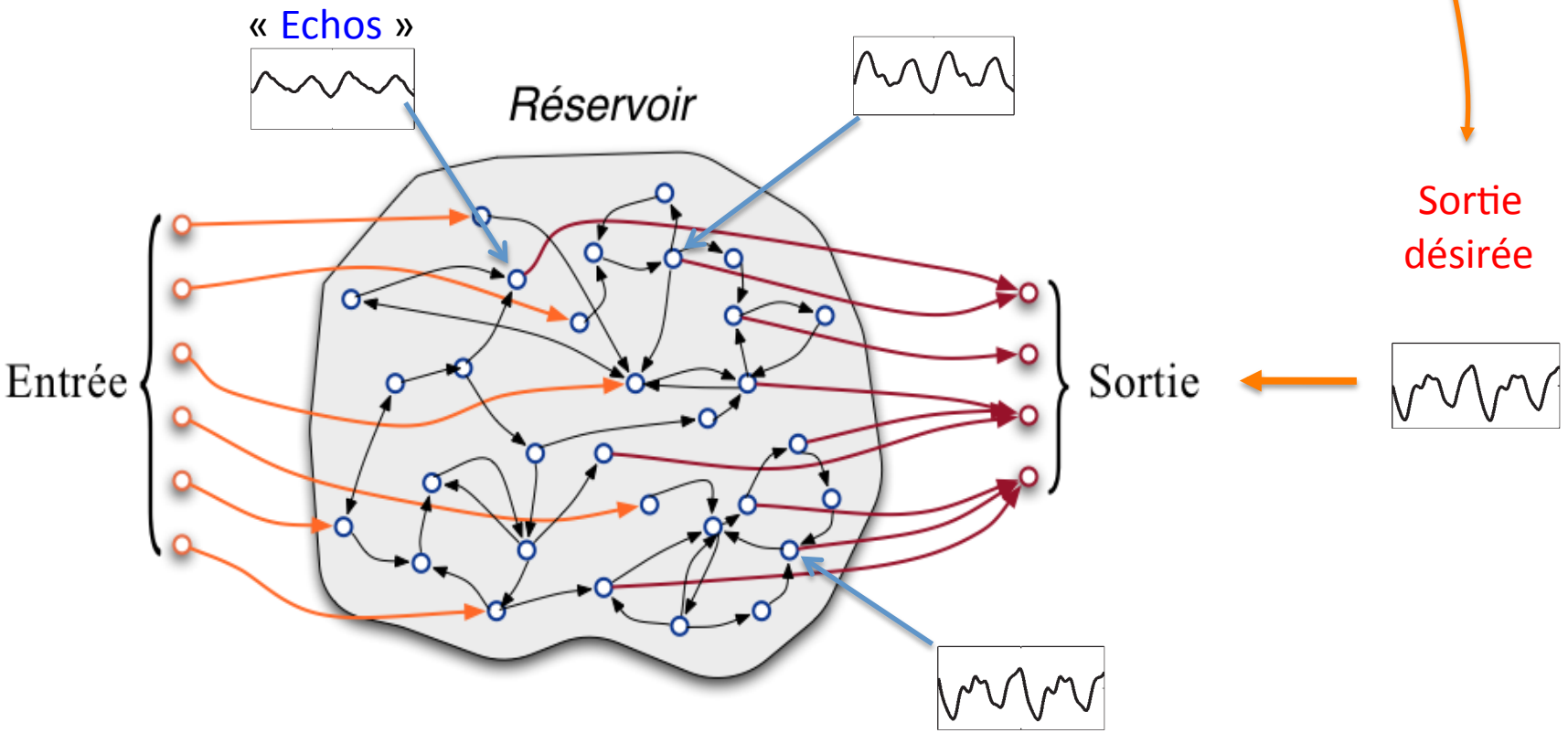


[Lukosevicius & Jaeger « Reservoir computing approaches to recurrent neural networks training », Computer Science Review, 2009]

Prédiction : « reservoir computing »

Signal à « apprendre »



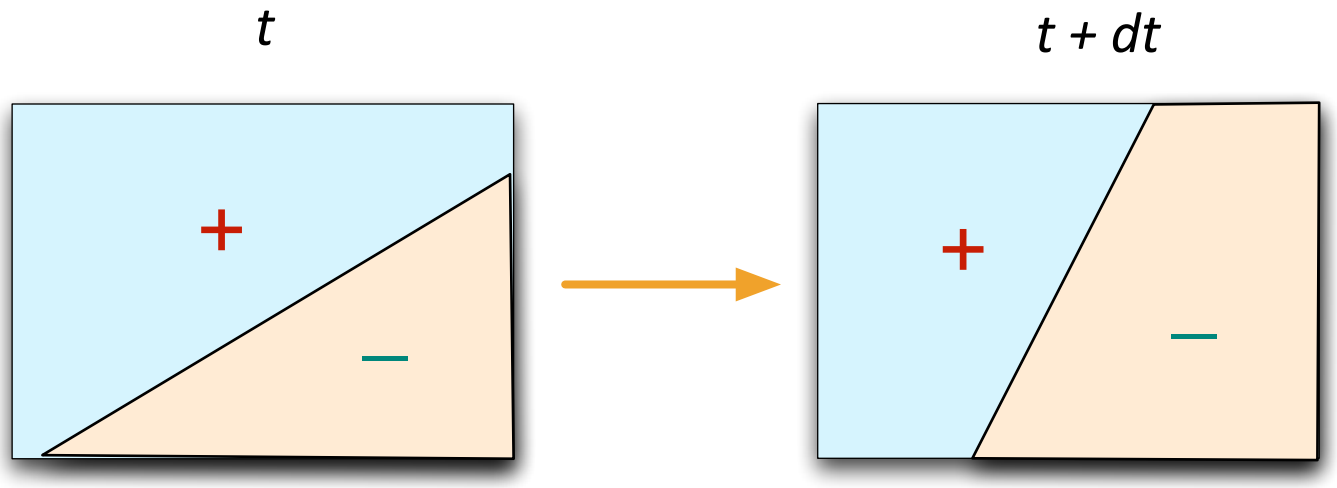


Apprentissage **en-ligne**

(« *on-line learning* »)

L'apprentissage « en-ligne »

- Le scénario
  - $(X_{t-n}, Y_{t-n}), \ldots, (X_{t-2}, Y_{t-2}), (X_{t-1}, Y_{t-1}), (X_t, Y_t) \ldots (X_{t+1}, Y_{t+1})$?



- Mesure de performance
  - Σ erreurs ; Moyenne erreurs

L'apprentissage « en-ligne »

- Apprentissage contre toute séquence
  - Plus d'hypothèse de stationnarité
  - Ni d'aucune régularité temporelle

- Comment savoir si l'algorithme est bon ?

- Idée de comité d'« experts »

Utilisation d'un **comité d'experts**

	Expert_1	Expert_2	Expert_3	Expert_4	Expert_5	Expert_6
J_1	1	0	0	1	0	1
J_2	1	1	1	0	0	0
J_3	1	0	0	0	1	1
J_4	1	0	0	1	1	1
J_5	1	1	0	1	1	1
J_6	1	0	0	0	1	0

Quel algorithme de choix à chaque J ?

Algorithme de sélection d'expert

- Choix d'**un expert a priori sans changement**
  - Propriétés ?
    - Possibilité de perte ∞

Peut-on faire mieux ?

Utilisation d'un **comité d'experts**

	Expert_1	Expert_2	Expert_3	Expert_4	Expert_5	Expert_6
J_1	1	0	0	1	0	1
J_2						
J_3						
J_4						
J_5						
J_6						

Utilisation d'un **comité d'experts**

	Expert_1	Expert_2	Expert_3	Expert_4	Expert_5	Expert_6
J_1	1	0	0	1	0	1
J_2	1	1	1	0	0	0
J_3						
J_4						
J_5						
J_6						

Utilisation d'un **comité d'experts**

	Expert_1	Expert_2	Expert_3	Expert_4	Expert_5	Expert_6
J_1	1	0	0	1	0	1
J_2	1	1	1	0	0	0
J_3	1	0	0	0	1	1
J_4						
J_5						
J_6						

Algorithme glouton déterministe

Algorithme de sélection d'expert

- Algorithme **glouton déterministe**
  - ***Propriétés*** ?
    - Peut être très bon
    - **Pire cas ?**

Algo glouton déterministe : **pire cas**

	Expert_1	Expert_2	Expert_3	Expert_4	Expert_5	Expert_6
J_1	1	0	0	0	0	0
J_2	0	1	0	0	0	0
J_3	0	0	1	0	0	0
J_4	0	0	0	1	0	0
J_5	0	0	0	0	1	0
J_6	0	0	0	0	0	1

$$L \leq N(L^*) + N - 1$$

Perte algo

Perte meilleur expert

Apprentissage **en-ligne**

- Pourquoi comparer avec le meilleur expert ?
  - Notion de « regret »

- Pourquoi pas avec le meilleur algorithme possible ?

Algorithme de sélection d'expert

- Algorithme **glouton aléatoire**
  - ***Propriétés*** ?
    - Peut être très bon
    - **Pire cas ?**

$$L_{RG} \leq (\ln N + 1)(L^*) + \ln N$$

Le « cas **réalisable** »

- Classification à 2 classes
- ∃ un expert *i* inconnu ne faisant jamais d'erreur : $h_{i,t}(x_t) = y_t \quad \forall t$
- Quelle stratégie ?
  - On assigne **un poids** $w_t = 1$ à tous les experts
  - À chaque *t*
    - Prédire la classe majoritaire dans le vote : $H(x_t)$
    - Comparer la prédiction $h_{i,t}(x_t)$ avec y_t
    - Assigner $w_t = 0$ à tous les experts ayant fait une erreur

$$L_{CR} \leq \lfloor \log_2 N \rfloor$$

Cas **réalisable** : preuve

- Initialement : $W_0 = N$
- À chaque étape : $W_t \leq W_{t-1}/2$

$$L_{CR} \leq \lfloor \log_2 N \rfloor$$

Le cas **non** réalisable

- À t=0, $W_0 = N$
- À chaque *t* :

$$w_i(t) = \begin{cases} w_i(t) & \text{si } y(t) = h_{i,t}(\mathbf{x}_t) \\ \beta\, w_i(t) & \text{si } y(t) \neq h_{i,t}(\mathbf{x}_t) \end{cases}$$

$$W(t) \leq W(t-1)/2 + \beta\, W(t-1)/2$$

$$W(t) \leq W_0 \frac{(1+\beta)^t}{2^t} \qquad \text{et} \qquad W(t) \geq \beta^{L^*(t)}$$

$$\beta^{L^*(t)} \leq W_0\, (1+\beta)^t / 2^m$$

$$L_{CR} \leq \left\lfloor \frac{\log_2 N + L^* \log_2(1/\beta)}{\log_2 \frac{2}{1+\beta}} \right\rfloor$$

Bilan sur ce type d'analyse

- Permet d'obtenir des **théorèmes** !!
- Mais **trop exigeant** et peu réaliste
- Idée intéressante : **comité d'experts**

Environnement **non stationnaire**

- **Co-variate shift** $\mathbf{p}_{\mathcal{X}}$
  - Dérive virtuelle
  - Non i.i.d.

- **Changement de concept** $\mathbf{p}_{\mathcal{Y}|\mathcal{X}}$
  - *Concept drift*
  - Non i.i.d. + non stationnaire

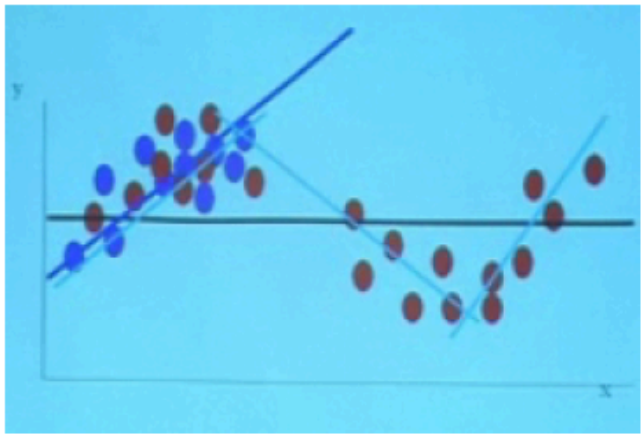
Co-variate shift

No longer a “direct” link between empirical risk and real risk

Modify the inductive criterion

The performance for the target distribution $\mathbf{P}'_{\mathcal{X}}$ (*generalization*) depends on :

- The performance for $\mathbf{P}_{\mathcal{X}}$ (*learning*)
- The similarity between $\mathbf{P}_{\mathcal{X}}$ and $\mathbf{P}'_{\mathcal{X}}$

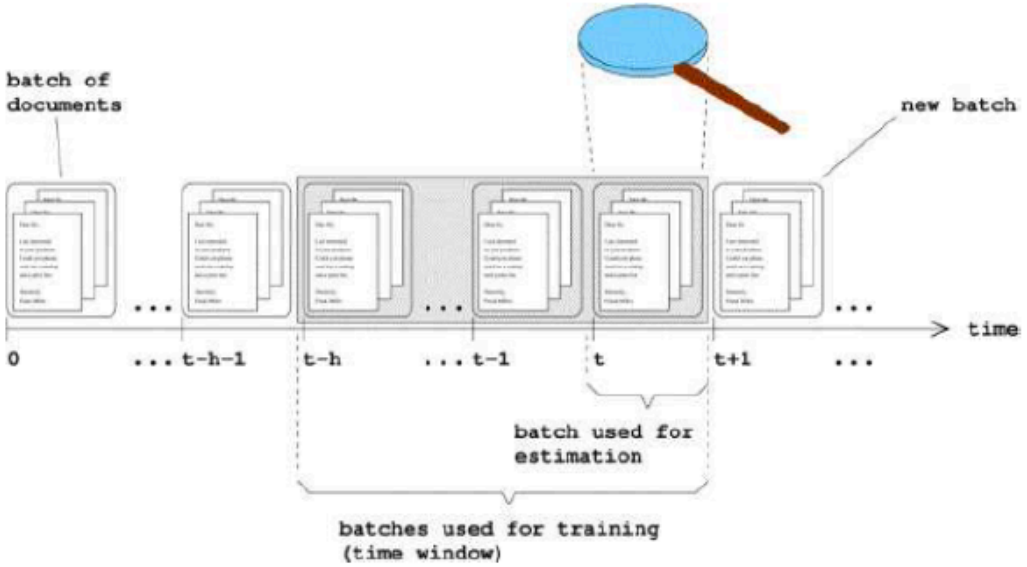


Dérive de concept

Drift of $\mathbf{P}_{\mathcal{X}|\mathcal{Y}}$

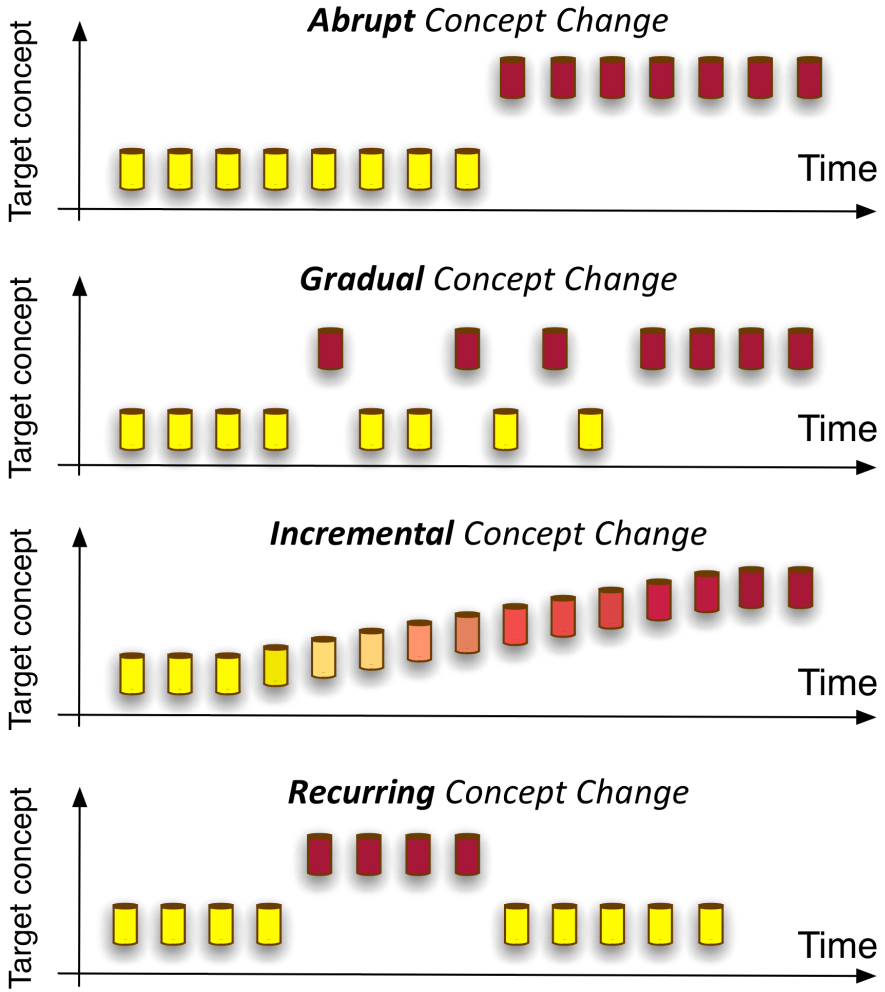
Exemples :

- Profiles of customers (purchases function of *income, age, ...*)
- Document filtering function of the interests of the user



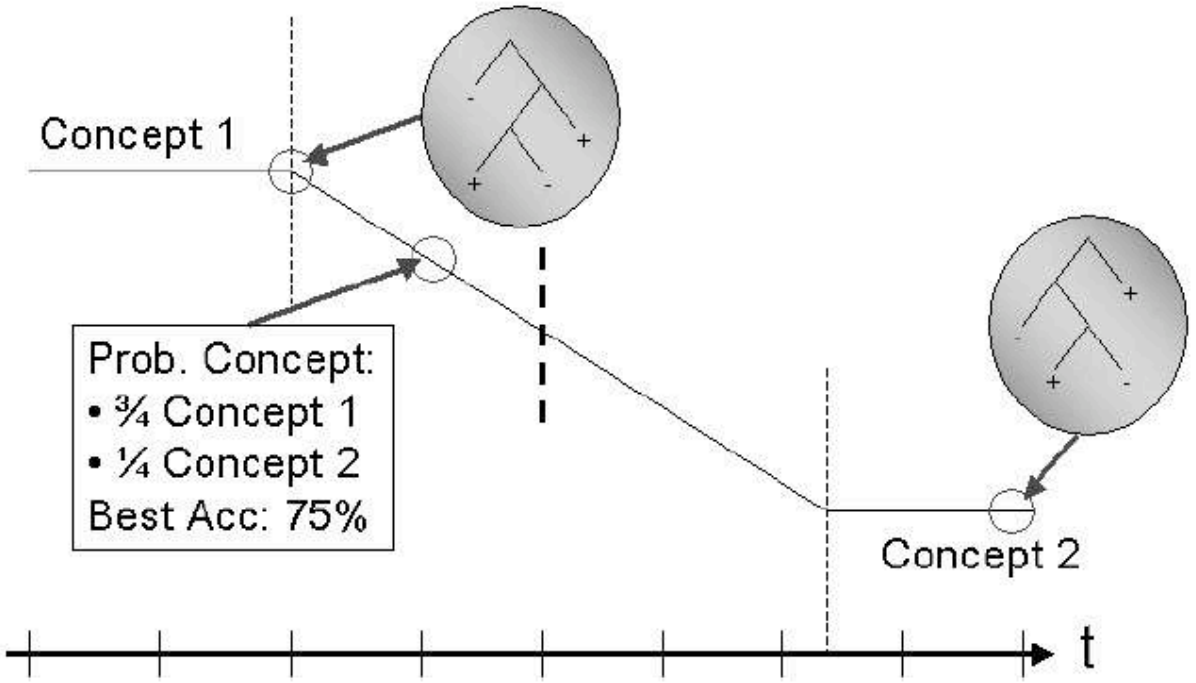
Changement de concept

Types de changements de concepts



Changement de concept

- Exemple



Changement de concept

- … à nouveau le problème du contrôle de la mémoire

Le dilemme ***plasticité-stabilité***

Dérive de concept

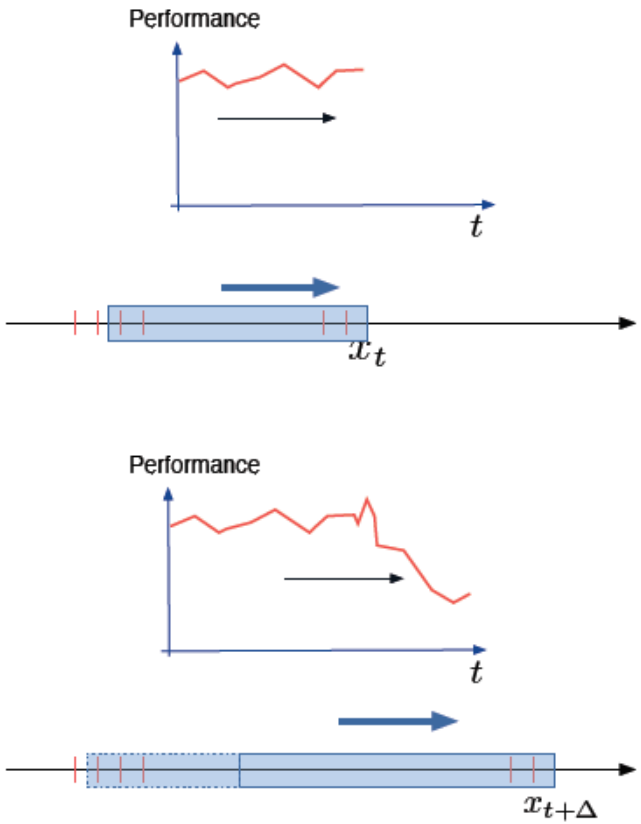
- Problèmes
  - **Apprendre le modèle le plus précis** possible
    - *Longue mémoire*
  - **Être réactif** mais en résistant au bruit
    - *Oublier rapidement*

Dilemme stabilité-plasticité

- **Approches**
  1. ***Fenêtre** glissante*
  2. ***Poids** sur les données*
  3. *Méthodes **d'ensemble***

Dérive de concept : **fenêtres glissantes**

Principle:



WK96 G. Widmer and M. Kubat (1996) *"Learning in the presence of concept drift ans hidden contexts"* Machine Learning 23: 69–101, 1996.

Dérive de concept : **méthodes d'ensemble**

- Apprendre des **experts** sur des **fenêtres différentes**
- **Pondérer** les experts en fonction de leur performance (récente)
- **Remplacer** les plus mauvais experts

Dérive de concept : **méthodes d'ensemble**

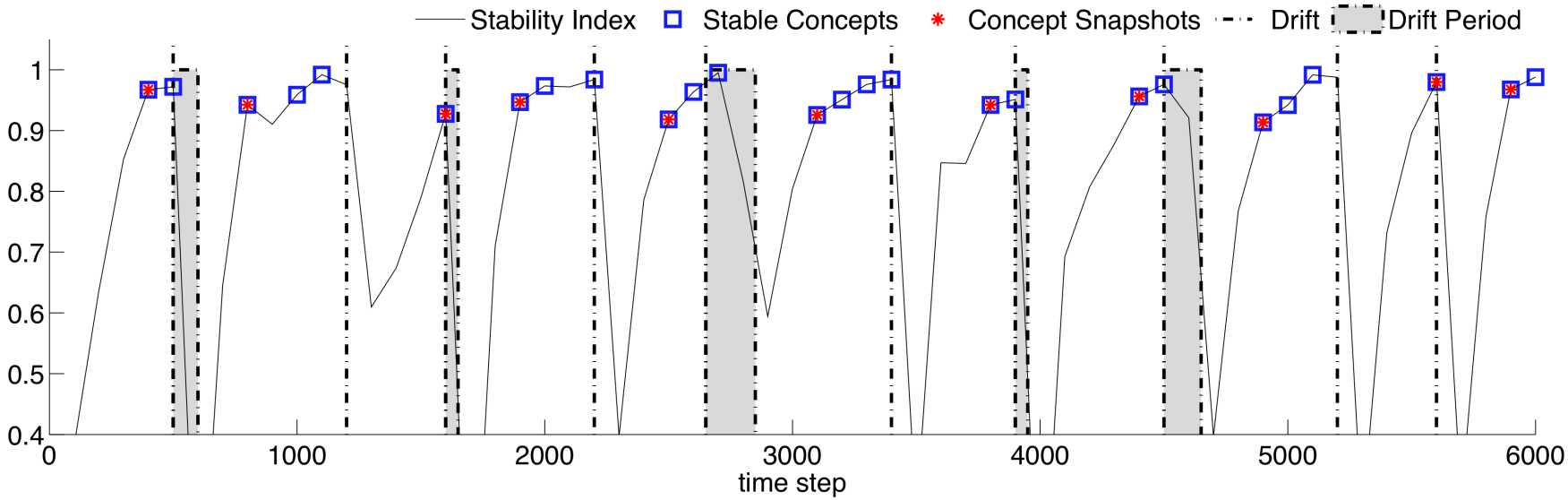
Dynamic weighted majority

- Classifiers in ensemble have initially a weight of 1
- For each new instance:
  - If a ***classifier predicts incorrectly***, reduce its weight
  - If ***weight drops below threshold***, remove classifier
  - If ***ensemble*** then ***predicts incorrectly***, install new classifier
  - Finally, all classifiers are (incrementally) updated by considering new instance

---

KM03 Kolter, Maloof (2003) *"Dynamic weighted majority: a new ensemble method for tracking concept drift"* ICDM 2003, 123-130.

De l'adaptation à l'anticipation



Stream		Adaptation			Anticipation	Total gain		Dues to anticipation		Due to recurring	
name	size	base learner	mean	std-dev	predictor	mean	std-dev	mean	std-dev	mean	std-dev
10-D Low	7,150	perceptron	**112.7**	4.6	Elman net	**1.0**	0.0	0.0	0.0	1.0	0.0
10-D Med.	7,150	perceptron	**826.7**	36.1	Elman net	**311.8**	33.9	60.6	23.2	251.2	20.6
10-D High	7,150	perceptron	**904.8**	33.4	Elman net	**345.4**	25.2	87.8	26.4	257.6	13.3
Robot	753	decision tree	**43.0**	2.6	-	**9.0**	1.9	-	-	9.0	1.9

Table 1: Summary of the experiments and the measured gains in prediction errors wrt. an adaptive only strategy.

G. Jaber, A. Cornuéjols, and P. Tarroux, "**Anticipative and Dynamic Adaptation to Concept Changes," in** *Proc. ECML-PKDD-2013* (Workshop ``Real-World Challenges for Data Stream Mining''), Prague, Czech Republic, 2013.

Approches heuristiques de l'apprentissage en-ligne : **bilan**

- Efficaces dans certaines situations
- Demandent le réglage de paramètres
- En plein essor

- Manque des fondations théoriques solides

Sous-séquences

fréquentes ou atypiques

Recherche de sous-structures fréquentes

- Soit une base de données $\mathcal{D}$
- *m* « transactions » *s* constituées d'items *b*
- La recherche de sous-structures fréquentes (e.g. sous-séquences) est de complexité calculatoire exponentielle

Recherche de sous-séquences fréquentes

- **Ensemble d'items** $\mathcal{I} = \{b_i\}$.
- $\mathcal{E} = \mathcal{P}(\mathcal{I})$ est l'ensemble de tous les **évènements** possibles
- Soit α_i un événement
- Une **séquence** est une liste ordonnée $\alpha_1 \rightarrow \alpha_2 \rightarrow \ldots \rightarrow \alpha_n$
  - $\mathcal{I}$ = {A, B, C, D, E, F}
  - A -> AB -> BCD -> AE
- **Sous-séquences**
  - A -> A
  - A -> E
  - AB -> B -> E
  - ~~AE~~

Recherche de sous-séquences fréquentes

- Recherche des sous-séquences de support > min_support

TID	Transaction
1	$A \rightarrow AB \rightarrow BCD \rightarrow E$
2	$CE \rightarrow AB \rightarrow F \rightarrow CDE$
3	$BE \rightarrow B \rightarrow AF \rightarrow ACE$
4	$A \rightarrow E \rightarrow BF$
5	$BCD \rightarrow AF \rightarrow ABF$

TID	Transaction
1	$A \rightarrow AB \rightarrow BCD \rightarrow E$
2	$CE \rightarrow AB \rightarrow F \rightarrow CDE$
3	$BE \rightarrow B \rightarrow AF \rightarrow ACE$
4	$A \rightarrow E \rightarrow BF$
5	$BCD \rightarrow AF \rightarrow ABF$

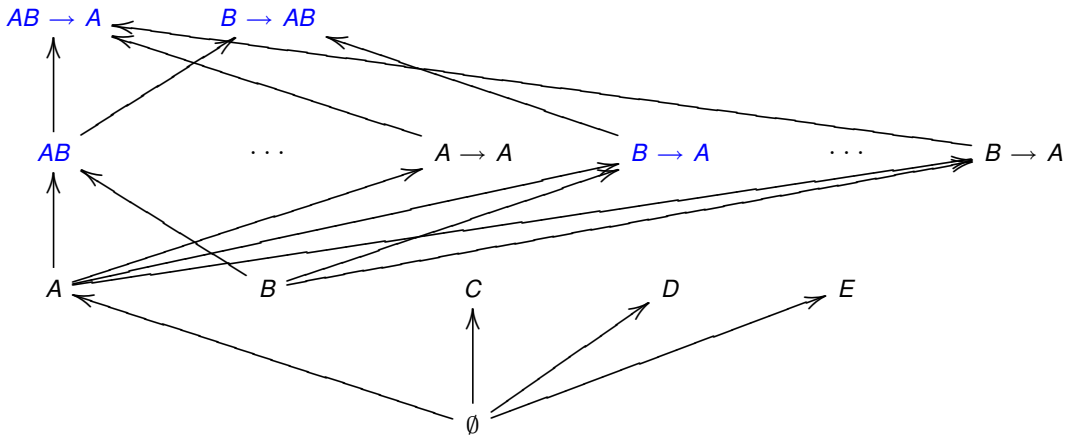
La sous-séquence *A ->A*

a un Min_support = 3

Recherche de sous-séquences fréquentes

- Exploitation
  - D'un treillis sur les séquences
  - D'une propriété d'anti-monotonicité

- **Algorithmes**
  - Généralisations de l'algorithme Apriori
  - GSP
  - Spade
  - PrefixSpan
  - …



Analyse

d'**ensembles** de séquences

Types de tâches

1. Apprentissage non supervisé (clustering)
   - E.g. recherche de profils types de consommateurs

2. Apprentissage supervisé
   - Classement de nouvelles séquences
     - E.g. consommation électrique élevée / non élevée entre 18h et 20h

Questions

- Trouver les *k* séquences les plus proches d'une séquence donnée
- Trouver les séquences à moins de ε d'une séquence donnée

- Distances classiques
  - Euclidienne
  - Normes diverses
  - Présupposent un espace vectoriel

→ Il faut **transformer** la représentation des séquences

- Distances pour les séquences
  - *Dynamic Time Warping* (DTW)
  - *Distances d'édition*
  - *Sous-séquence commune la plus longue*
  - …

Questions

1. **Représentation** des séquences

2. **Alignement** entre séquences

3. Mesure de similarité ou de **distance**

La représentation des séquences

1. Brute

2. Représentations « analytiques »
   - Par combinaison de composantes choisies dans un dictionnaire
   - Composantes orthogonales

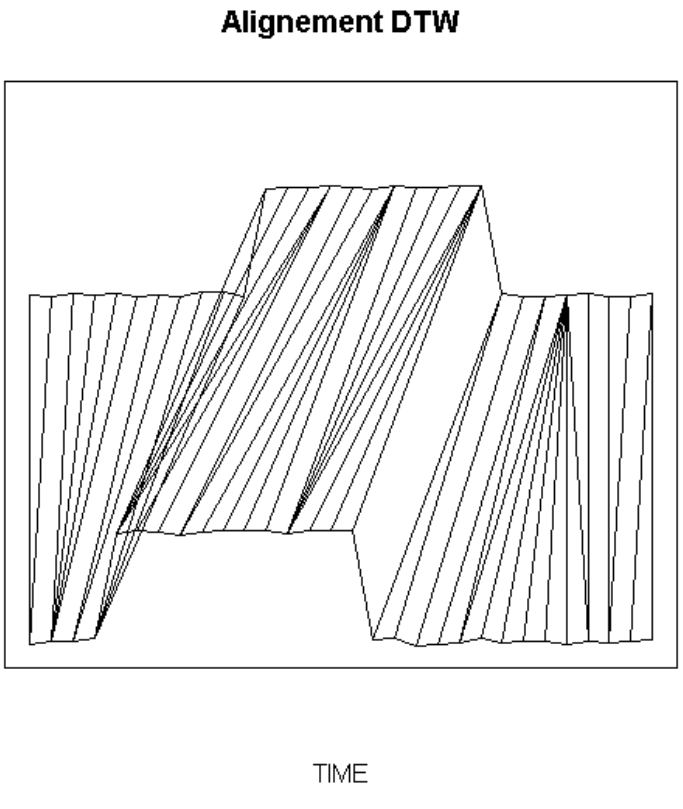
3. Régularité supposée
   - Grammaires
   - Chaînes de Markov
   - Réseaux de neurones
   - …
   - Problème de la comparaison

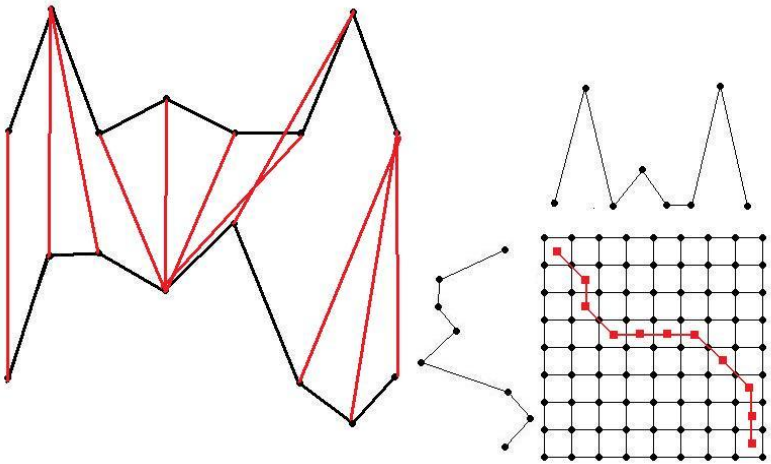
L'alignement

- **Représentation brute**
  - Problèmes
    - Très grande dimension
    - Pas nécessairement même dimension
  - DTW (Dynamic Time Warping)
    - Principe
      - Coût associé à appariement local (e.g. même profil local)
      - Coût associé à déplacement non simultané dans le temps
      - Minimisation de la somme des coûts
    - Calcul
      - Programmation dynamique
      - Réalisable en-ligne

Alignement par DTW

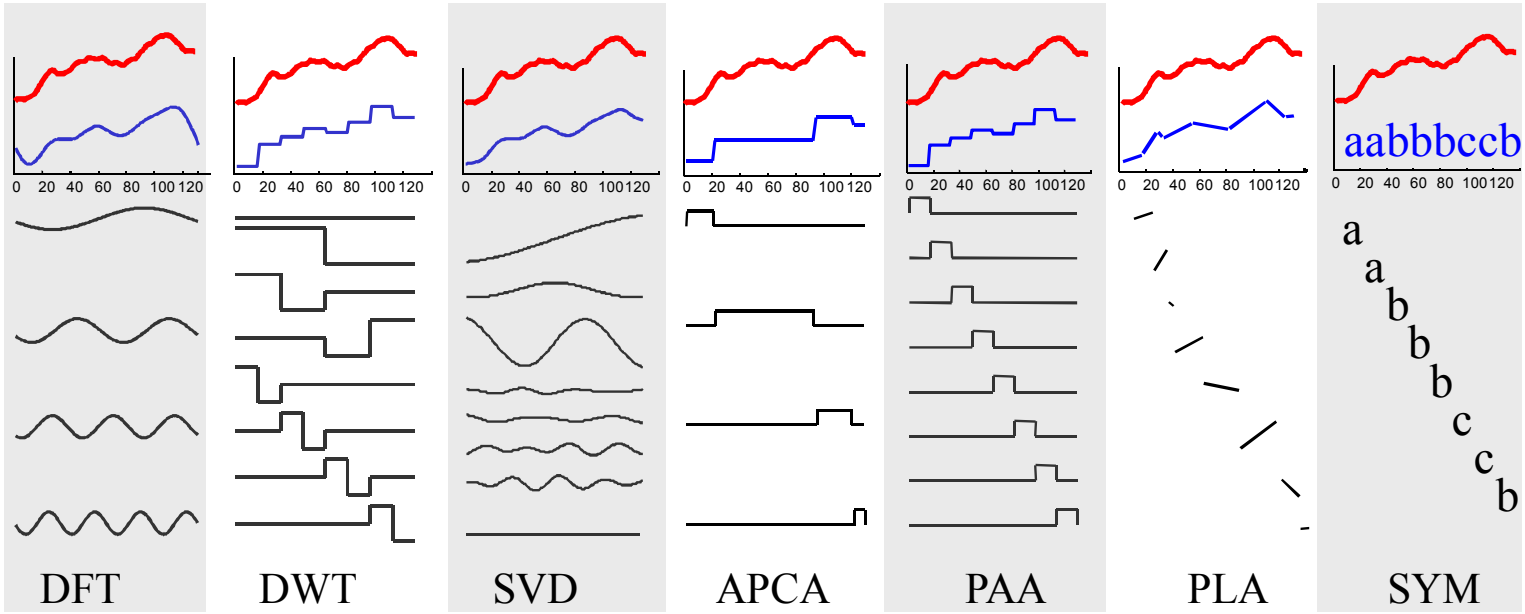
Sankoff, D. et Kruskal, J. (1983). Time warps, string edits, and macromolecules : the theory and practice of sequence comparison. *Reading : Addison-Wesley Publication, 1983, edited by Sankoff, David ; Kruskal, Joseph B., 1*





Représentations analytiques

- Transformée de Fourier discrète
- Transformée en ondelettes discrète
- Approximation agrégée par morceaux
- …



Représentation vectorielle

Clustering de séquences

La comparaison : mesures de distance

- Représentations fondées sur les **valeurs**
  - Représentations « brutes » ou « analytiques »
  - Sans *time warping*
    - toutes les normes Lp
  - Avec *time warping*
    - Dynamic time warping
    - Distance d'édition
    - Distance de Fréchet
- Représentations fondées sur les **comportements**
  - Signe de la pente ; dérivées
  - Coefficients de Pearson

Clustering en représentation vectorielle

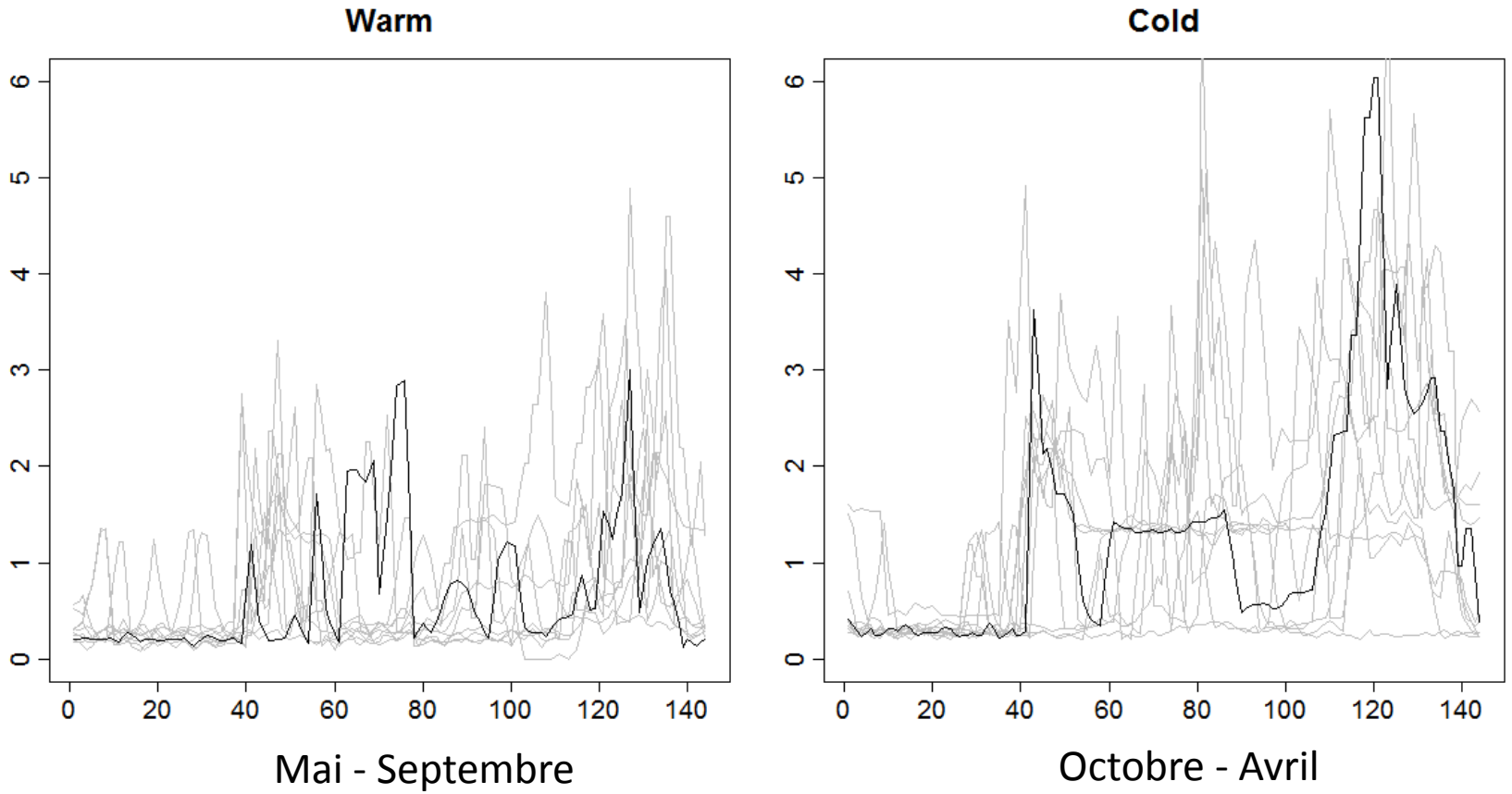
- *K*-moyennes
- Classification hiérarchique ascendante
- Clustering spectral
- …

Classification supervisée de séquences

Classification supervisée de séquence

- Le problème
  - E.g. Consommation électrique pour des séries des classes Warm et Cold du jeu de données consseason.

    [Hebrail, G. et al., Exploratory analysis of functional data via clustering and optimal segmentation. 2010]



Classification supervisée de séquence

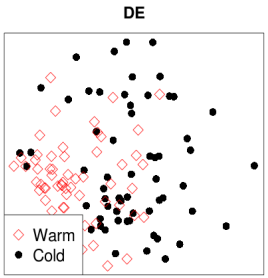
- Approches
  - k-NN
  - SVM
  - …
- Questions :
  - Quelle représentation ?
    - Des séquences
    - Des classes
  - Quelle distance ?

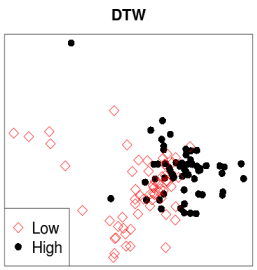
Classification supervisée : exemple

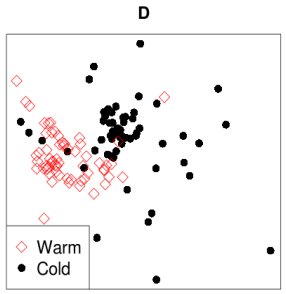
- *Thèse de Cédric Frambourg (LIG, 2013)*
  - Apprentissage
    - de **l'appariement DTW** pour augmenter le contraste entre classes
      - minimisant la variance intra-classes et maximisant la variance inter-classes
    - de **poids associés aux instants** dans une métrique induite (-> instants discriminants)
    - Calcul d'un **profil moyen / classe**
  - Utilisation de *k*-ppv
    - Avec une distance induite par la phase d'apprentissage
  - Tests
    - Sur courbes artificielles
    - Sur courbes de consommation électrique

Classification supervisée : exemple

- Apprentissage des « saisons »
  - **Données**
    - 349 jours
    - Relevé toutes les 10mn : 144 mesures / jour
    - Échantillons d'apprentissage = 60 séquences
    - Échantillons de test = 30 séquences

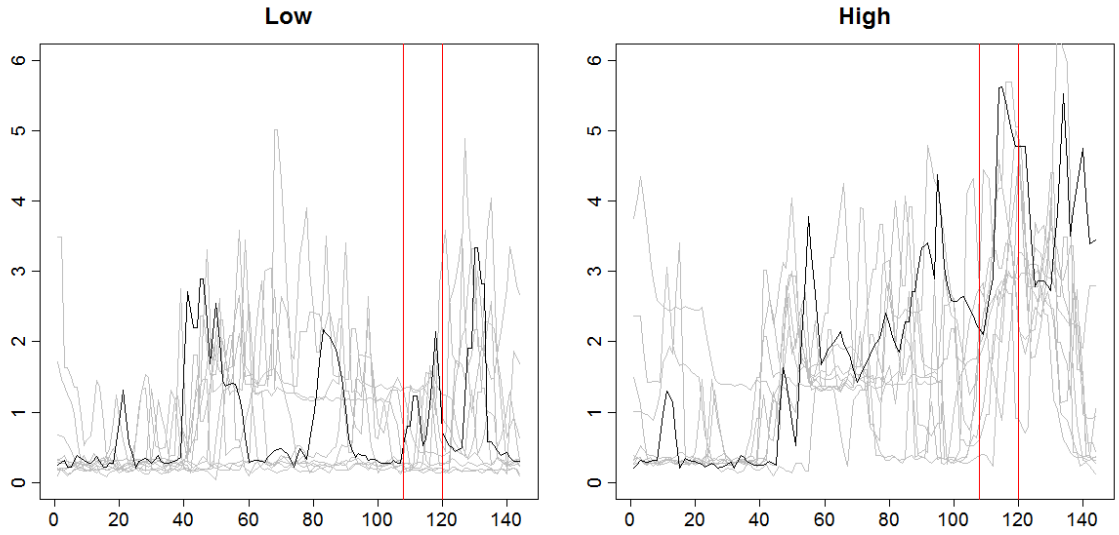






k	DE	DTW	D
1NN	23.9	28.3	**9.4**
3NN	22.8	31.1	12.8
5NN	20.0	30.0	20.5
7NN	22.2	30.6	11.1

Le cas de la « classification précoce »

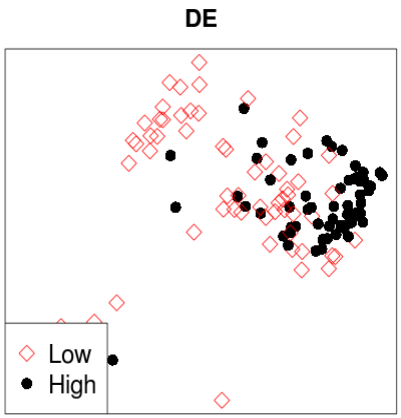


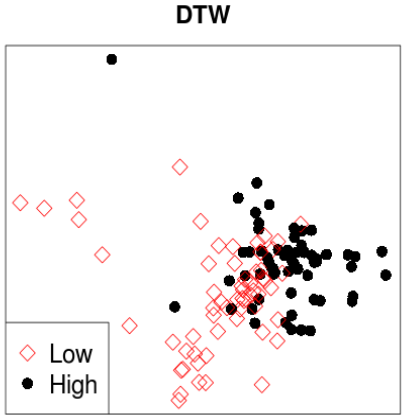
- Le **dilemme**
  - Plus on attend
    - Meilleure est la prédiction
    - Moins intéressante elle est

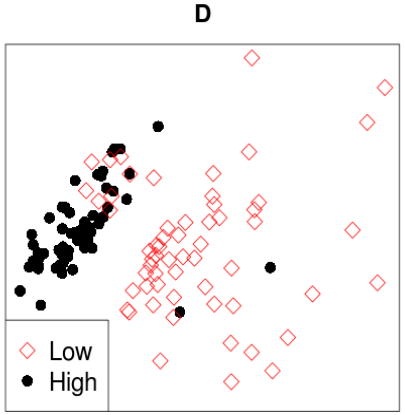
→ Maximiser une sorte d'**espérance de gain**

Le cas de la « classification précoce »

- Apprentissage sur les instants avant 16h







k	DE	DTW	D
1NN	30.6	28.9	5.6
3NN	26.7	26.1	4.4
5NN	23.3	23.9	2.8
7NN	23.3	23.3	**1.7**

Approche « Frambourg » : bilan

- **Performant**
- **Intéressant**
  - Apprentissage de profil de classe
  - Apprentissage de fonction d'appariement
  - Apprentissage d'une distance induite
- À **explorer davantage**
  - Complexité computationnelle
  - Processus itératif à optima multiples
    - Dépendant de l'initialisation
  - Pas mal de paramètres
  - Risque de sur-apprentissage

Conclusions

Bilan

- Domaine encore assez neuf pour l'AA
  - Peu focalisé sur la prédiction de la suite d'une séquence

- Source de questions intéressantes
  - Données **non vectorielles**
  - Données **non i.i.d.**
  - Apprentissage **en-ligne**
    - Algorithmes efficaces
    - Contrôle de la mémoire
    - Adaptation de l'hypothèse et de $\mathcal{H}$ en-ligne
  - Lié à problèmes généraux
    - **Transfert** entre tâches