L'IA/AA dans tous les médias aujourd'hui

L'IA/AA est attendue **partout** demain

- Aide au **diagnostic médical**
- Aide aux **juges**
- Octroi de **prêts**
- Aide au **choix des employés**
- Évaluation du **risque de criminalité** avant l'acte criminel
- Calcul des **primes d'assurance**

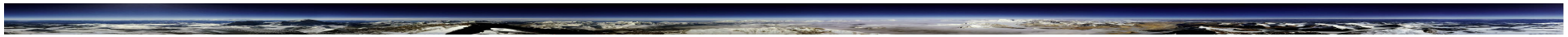
- **Assistant** personnel
- Véhicules **autonomes**
- Conduite des « **smart cities** »

De plus près

- Sommes-nous prêts ?

Qu'est-ce qu'un bon système d'apprentissage ?

La réponse a évolué avec le temps. Et demain ?



Antoine Cornuéjols

AgroParisTech – INRA MIA 518

antoine.cornuejols@agroparistech.fr

Plan

1. L'IA est **attendue partout**
2. Quelles **garanties** sur l'induction
3. Un peu d'**histoire**
4. Et **demain** ? Prémisses de changement de paradigme
5. Retour sur les **défis**
6. Conclusions … et ouverture

Un premier cas

*Prédiction du **risque de récidive** à moins de 4 ans* ('+' ou '-')

Un premier cas

*Prédiction du **risque de récidive** à moins de 4 ans* ('+' ou '-')

- > 2200 cas provenant de 26 cours juridictionnelles européennes sur 15 ans
- Un trentaine de mesures (*âge, profession, niveau de formation, lieu de résidence, style de vie, antécédents de drogues, antécédents familiaux, …*)

Un premier cas

*Prédiction du **risque de récidive** à moins de 4 ans* ('+' ou '-')

- > 2200 cas provenant de 26 cours juridictionnelles européennes sur 15 ans
- Un trentaine de mesures (*âge, profession, niveau de formation, lieu de résidence, style de vie, antécédents de drogues, antécédents familiaux, …*)
- Méthodes :
  - Arbres de décision
  - Random forests
  - SVM

Un premier cas

*Prédiction du **risque de récidive** à moins de 4 ans* ('+' ou '-')

- > 2200 cas provenant de 26 cours juridictionnelles européennes sur 15 ans
- Un trentaine de mesures (*âge, profession, niveau de formation, lieu de résidence, style de vie, antécédents de drogues, antécédents familiaux, …*)
- Méthodes :
  - Arbres de décision
  - Random forests
  - SVM

Un juge utilise l'algorithme, qui prédit pour un sujet x_i :

- Classe '+' : une récidive aura lieu d'ici 4 ans
- (Éventuellement) avec un taux de confiance de ε% (e.g. 87%)

Que doit penser le juge?

Illustration (suite)

Le juge apprend que le taux de bonne classification de la meilleure hypothèse apprise sur les données test est de 62%

- Que doit-il penser pour le **cas du sujet x_i** ?

  Classe '+' avec probabilité estimée de 0.87

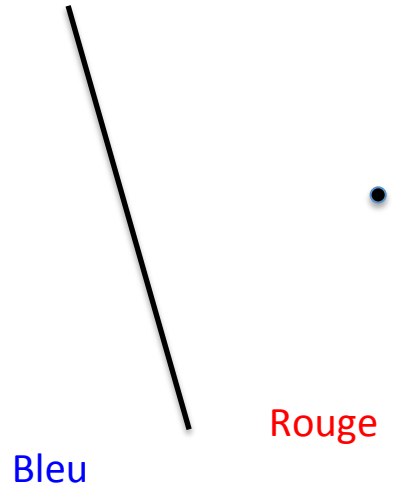
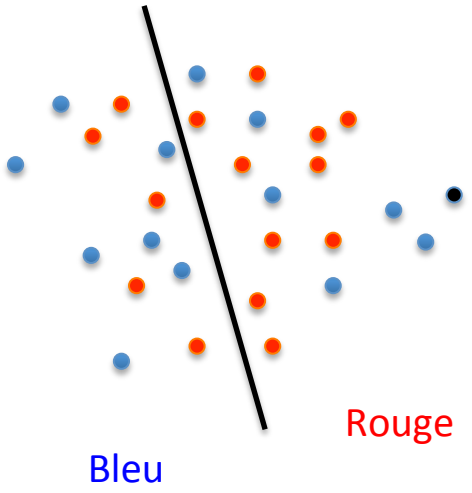


Illustration (suite)

Le juge apprend que le taux de bonne classification de la meilleure hypothèse apprise sur les données test est de 62%

- Que doit-il penser pour le **cas du sujet x_i** ?

  Classe '+' avec probabilité estimée de 0.87



Un **accident** de voiture (autonome)

Le 7 mai 2016, une voiture de la marque Tesla ***en mode « autopilot »***

percute un semi-remorque en travers de sa route

Un **accident** de voiture (autonome)

*Le 7 mai 2016, une voiture de la marque Tesla **en mode « autopilot »** percute un semi-remorque en travers de sa route*

- L'analyse révèle que :
  - Le **radar** a bien détecté le semi-remorque
    mais il y avait de nombreux panneaux routiers sur la route de « signature » radar proche
  - La **caméra** était peu sûre de ses détections en raison d'un ciel laiteux éblouissant (le semi-remorque était blanc)

Un accident de système complexe adaptatif

- L'analyse révèle que :
  - Le **radar** a bien détecté le semi-remorque
    mais …
  - La **caméra** était peu sûre de ses détections en raison de …

- **Et si** le radar avait dit qu'il avait des doutes et élevé le seuil d'alerte de la caméra, et réciproquement ? …
- **Et si** les systèmes avaient été adaptatifs ? …

Nécessité de **garanties**

1. Pouvoir « expliquer » ses préconisations
   - Prouver que le cas entre bien dans le **domaine de validité** du système de décision

Nécessité de **garanties**

1. Pouvoir « expliquer » ses préconisations
   - Prouver que le cas entre bien dans le **domaine de validité** du système de décision

2. Comme dans le génie logiciel classique
   - Pouvoir **décomposer** en sous-modules
   - Et avoir l'équivalent de **pré-conditions** et **post-conditions**

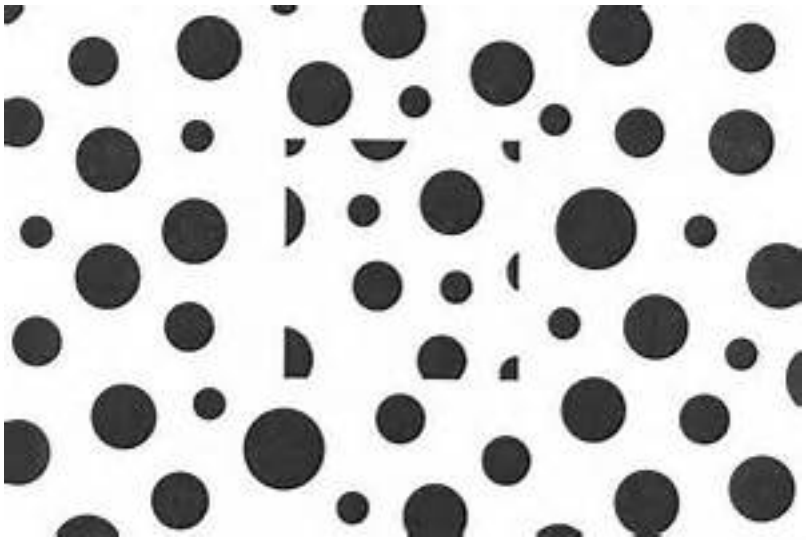
     Si je reçois cela en entrée, je garantis ceci en sortie
   - Est-ce que ce serait **suffisant** ? Faudra-t-il autre chose ?

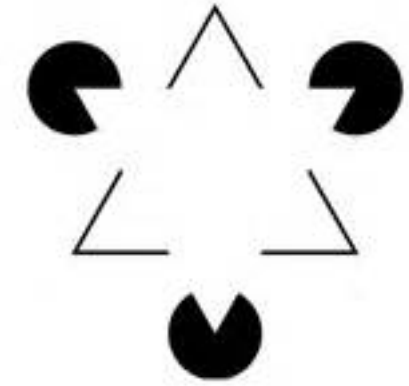
De quels outils disposons-nous pour répondre à ces questions ?

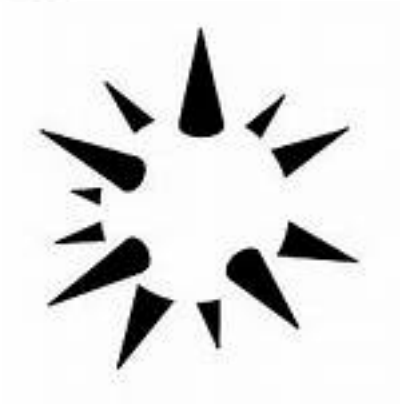
Plan

1. L'IA est **attendue partout**
2. Quelles **garanties** sur l'induction
3. Un peu d'**histoire**
4. Et **demain** ? Prémisses de changement de paradigme
5. Retour sur les **défis**
6. Conclusions … et ouverture

L'apprentissage – une extrapolation

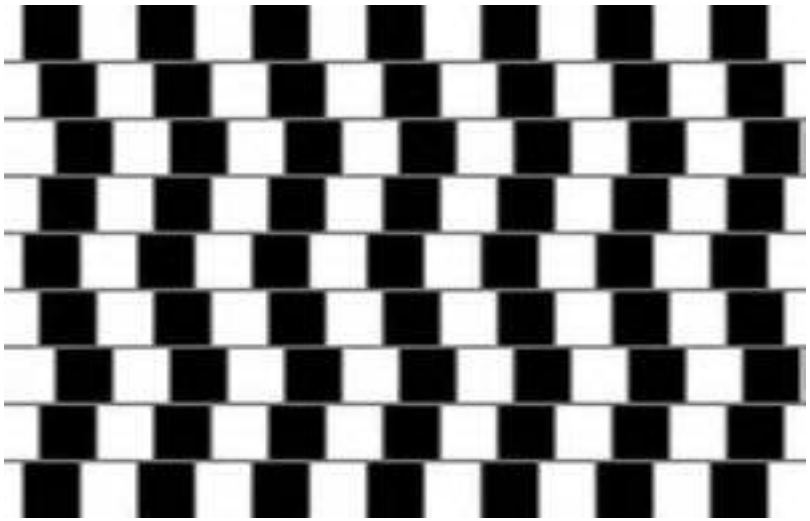




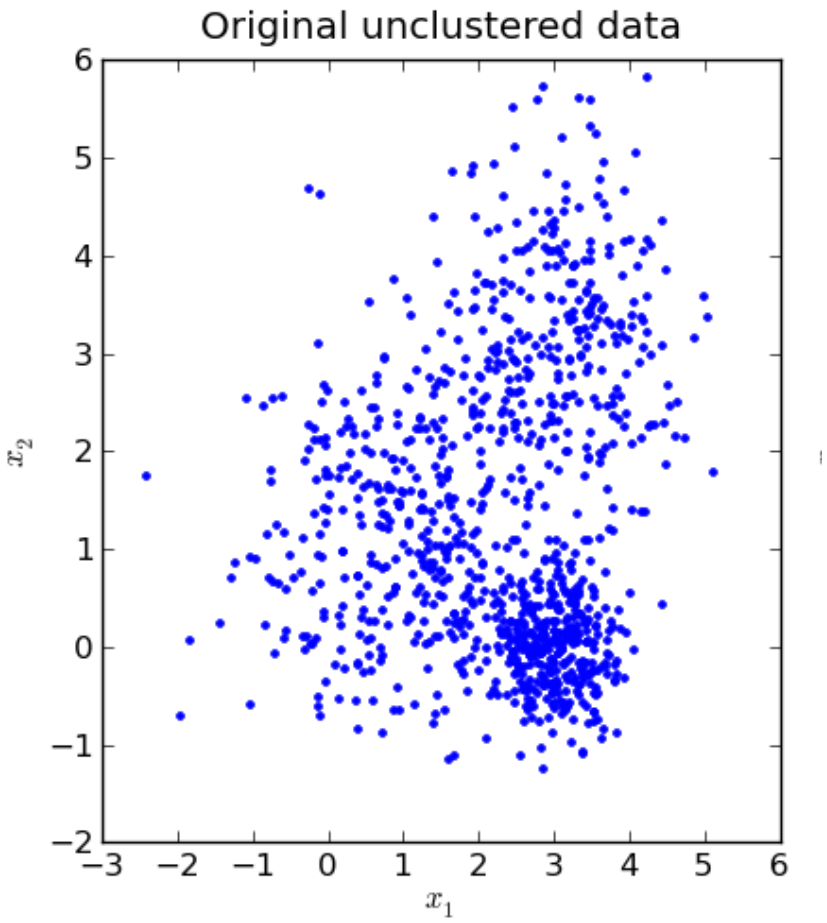


Une extrapolation – soumise à des **choix**

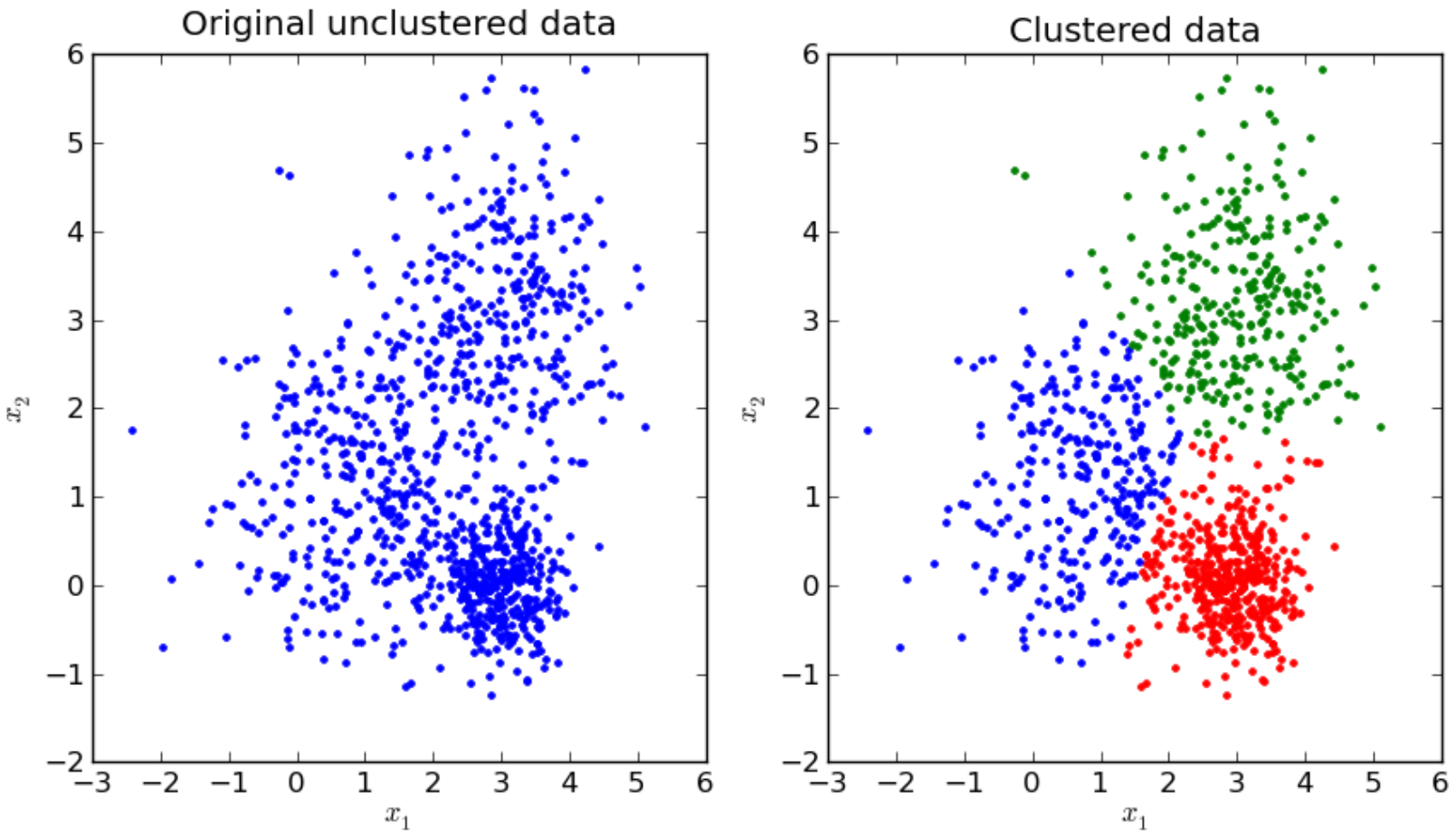
Des **biais** pouvant conduire à des illusions



Clustering

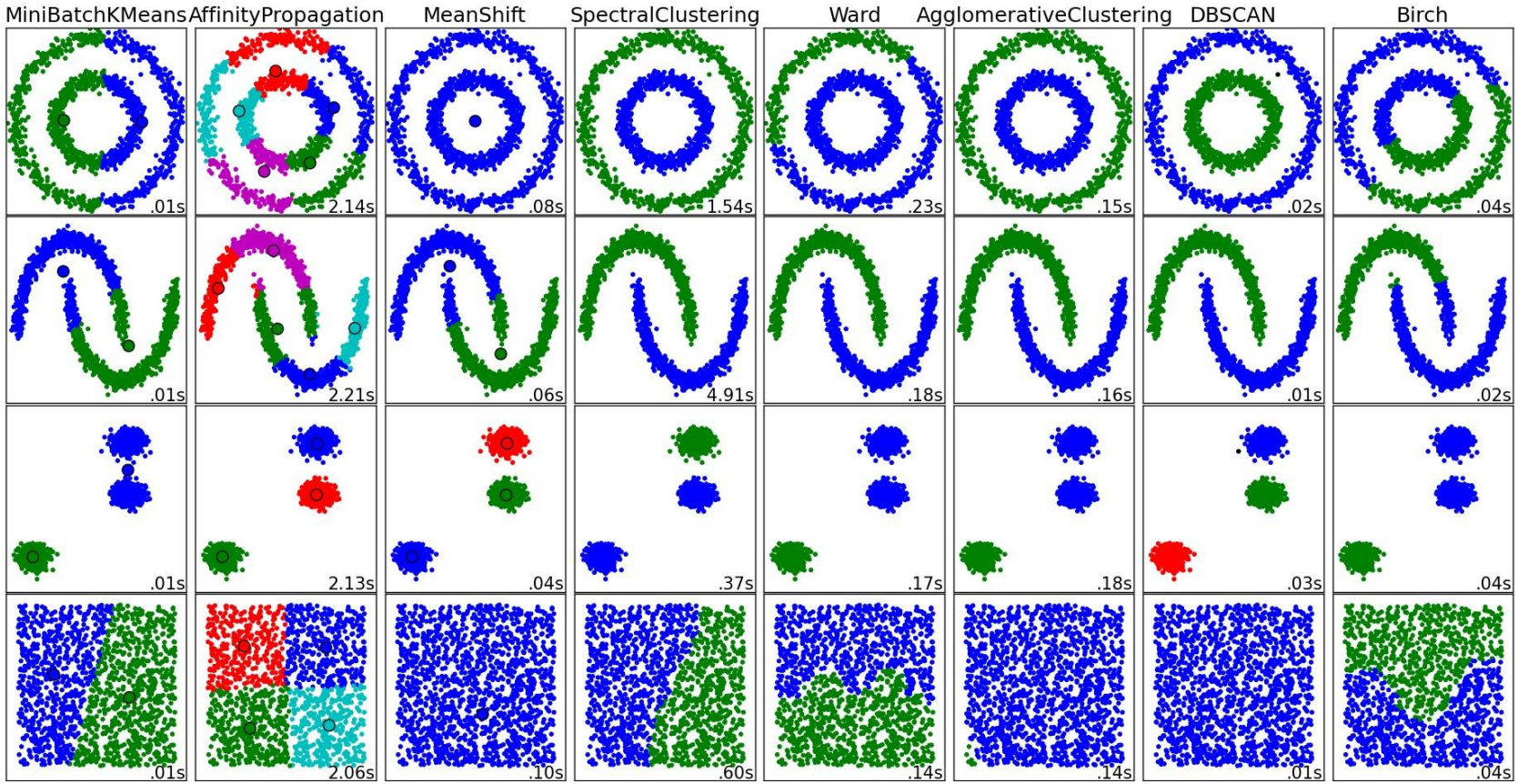


Clustering



Clustering

Dépend beaucoup des biais a priori



Comment fonder l'induction ?

Illustration : apprendre à **classer des exemples**

- Comment faire ?

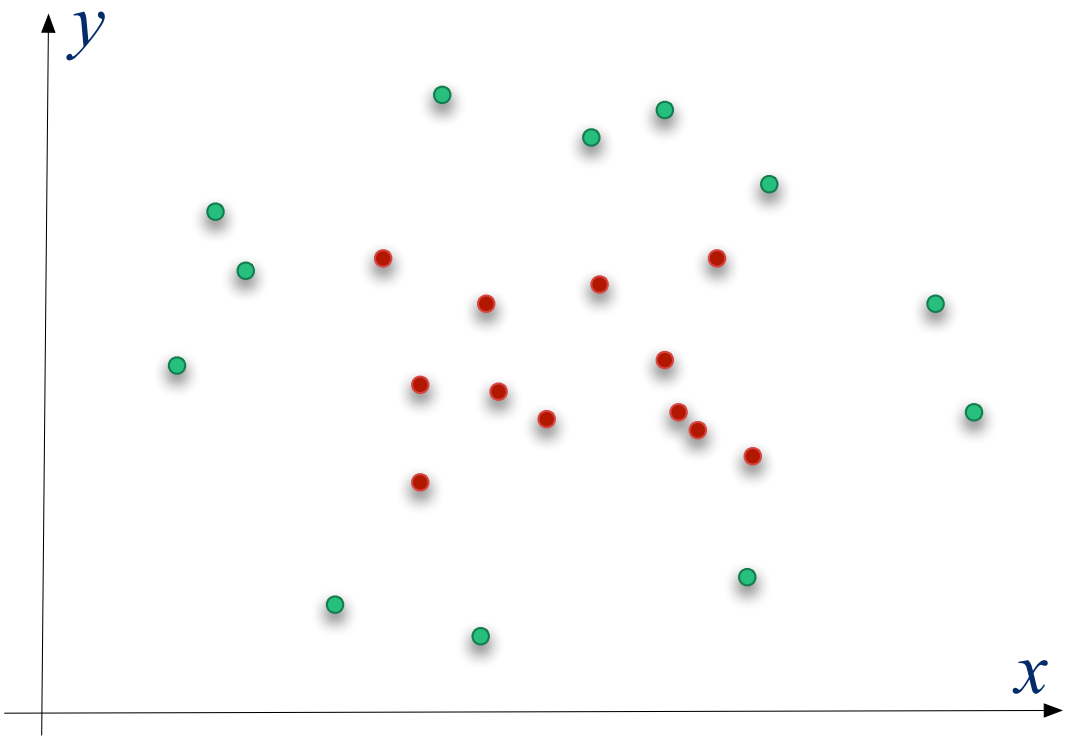


Illustration : apprendre à **classer des exemples**

- Comment faire ?

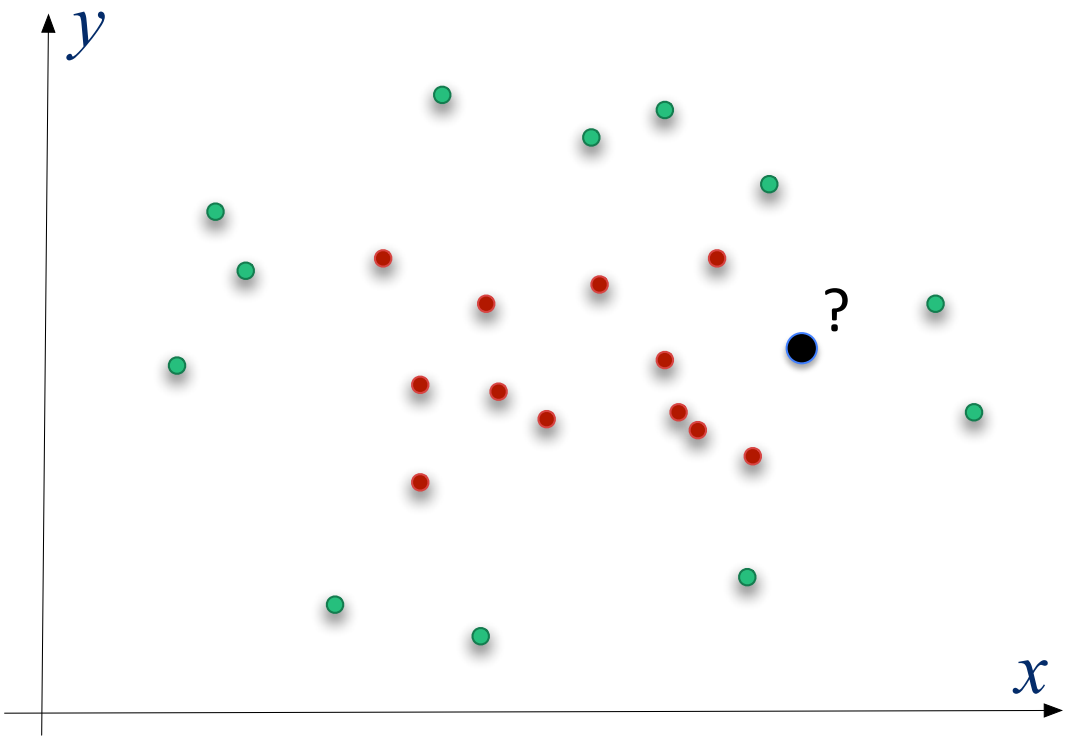
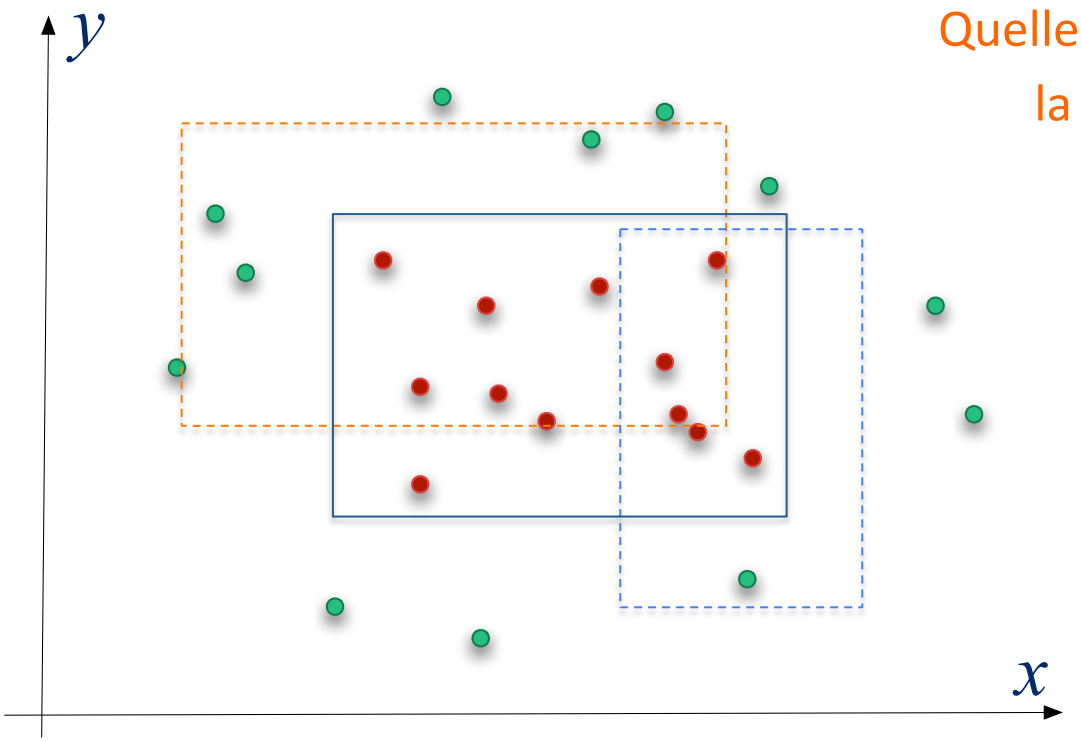


Illustration : apprendre à classer des exemples en 2D

- Comment faire ?



Quelle est l'hypothèse la meilleure ?

Quelle hypothèse choisir ?

Quelle qualité pour chaque hypothèse candidate ?

Quelle hypothèse choisir ?

Quelle performance ?

- Coût d'**une erreur** de prédiction
  - La ***fonction de perte***

$$\ell\big(h(\mathbf{x}), y\big)$$

Quelle hypothèse choisir ?

Quelle performance ?

- Coût d'**une erreur** de prédiction
  - La ***fonction de perte***

$$\ell\big(h(\mathbf{x}), y\big)$$

- Quel **coût à venir** (espérance) si je choisis *h* ?
  - Espérance de coût : le « ***risque réel*** »

$$R(h) = \int_{\mathcal{X}\times\mathcal{Y}} \ell\big(h(\mathbf{x}), y\big)\ \mathbf{p}_{\mathcal{X}\mathcal{Y}}(\mathbf{x}, y)\ d\mathbf{x}\, dy$$

Quelle hypothèse choisir ?

Comment trouver *h** (ou une bonne hypothèse) alors que l'on n'a accès qu'à un **échantillon d'apprentissage limité** ?

Quelle hypothèse choisir ?

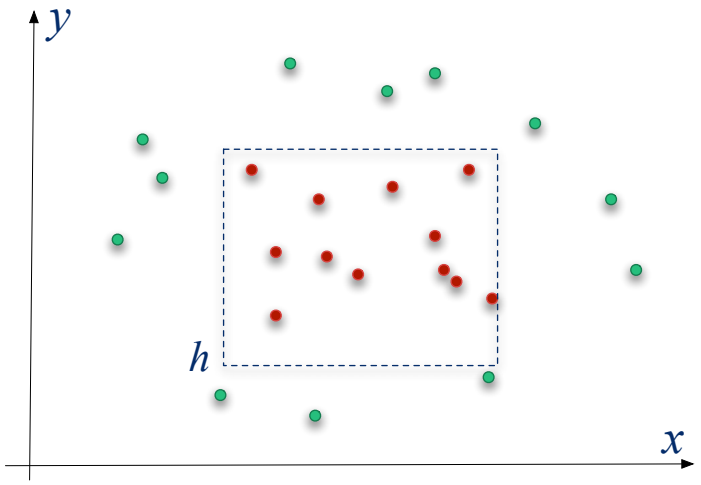
Comment trouver *h** (ou une bonne hypothèse) alors que l'on n'a accès qu'à un **échantillon d'apprentissage limité** ?

- Critère inductif : $\mathcal{H}$ x S -> valeur(h)

- Le plus naturel : ERM
  - La Minimisation du Risque Empirique

Quelle hypothèse choisir ?

- Quelle performance attendue pour *h* ?
  - Erreur moyenne sur l'échantillon d'apprentissage *S*



Le « risque empirique »

$$\hat{R}(h) = \frac{1}{m} \sum_{i=1}^{m} \ell\big(h(\mathbf{x}_i), y_i\big)$$

- A-t-on raison d'utiliser l'ERM ?

Un exemple qui dit beaucoup …

- Exemples décrits par :

  Number (1 or 2); ***size*** (small or large); ***shape*** (circle or square); ***color*** (red or green)

Description	**Your** prediction	**True** class
1 large red square		-
1 large green square		+
2 small red squares		+
2 large red circles		-
1 large green circle		+
1 small red circle		+

How many possible functions altogether from *X* to *Y* ? $2^{2^4} = 2^{16} = 65{,}536$

How many functions do remain after 6 training examples? $2^{10} = 1024$

Induction: impossible de gagner ?

- **Un biais est nécessaire**

- **Types** de biais
  - **De représentation** (déclaratif)
  - **De recherche** (procédural)

One example that tells a lot …

- Examples described using:

  ~~***Number***~~ (1 or 2); ***size*** (small or large); ~~***shape***~~ (circle or square); ***color*** (red or green)

Description	**Your** prediction	**True** class
~~1~~ large red ~~square~~		-
~~1~~ large green ~~square~~		+
~~2~~ small red ~~squares~~		+
~~2~~ large red ~~circles~~		-
~~1~~ large green ~~circle~~		+
~~1~~ small red ~~circle~~		+

How many possible functions with 2 descriptors from *X* to *Y* ? $2^{2^2} = 2^4 = 16$

How many functions do remain after 3 ≠ training examples? $2^1 = 2$

L'analyse « PAC learning »

- On arrive à :

$$\forall h \in \mathcal{H}, \forall \delta \leq 1 : \quad \mathbf{P}^m \left[R_{\text{Réel}}(h) \leq R_{\text{Emp}}(h) + \overbrace{\frac{\log |\mathcal{H}| + \log \frac{1}{\delta}}{m}}^{\varepsilon} \right] > 1 - \delta$$

Le principe de minimisation du risque empirique

n'est sain **que si** il y a des contraintes sur l'espace des hypothèses

Recette pour … **concevoir des algorithmes** d'apprentissage

1. Définir un **critère inductif régularisé**
   a. Exprimer le coût d'erreur de prédiction en une fonction de perte
   b. Définir un **terme de régularisation** qui exprime
      les attendus sur les régularités du monde
   c. Si possible, rendre convexe le problème d'**optimisation** résultant

$$h_{opt} = \underset{h \in \mathcal{H}}{\text{ArgMin}} \left[\underbrace{\frac{1}{m} \sum_{i=1}^{m} l(h(\mathbf{x}_i), y_i)}_{\text{empirical risk}} + \lambda \underbrace{reg(\mathcal{H})}_{\text{bias on the world}} \right]$$

2. Utiliser ou développer un **algorithme d'optimisation efficace**

Learning **sparse linear** approximator

- The **hypothesis** is of the form $h(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x}$
- A priori assumption: few non zero coefficients

Ridge regression

$$\mathbf{w}^*_{\text{ridge}} = \underset{\mathbf{w}}{\text{Argmin}} \left\{ \sum_{i=1}^{m} (y_i - \mathbf{w}\,\mathbf{x}_i)^2 + \lambda ||\mathbf{w}||_2^2 \right\}$$

Lasso regression

$$\mathbf{w}^*_{\text{lasso}} = \underset{\mathbf{w}}{\text{Argmin}} \left\{ \sum_{i=1}^{m} (y_i - \mathbf{w}\,\mathbf{x}_i)^2 + \lambda ||\mathbf{w}||_1 \right\}$$

Risque empirique régularisé

Expression de **substitution** du risque régularisé

Optimisation

3.3 du chapitre 3. Ainsi, étant donnés un échantillon source étiqueté $S = \{(x_i^s, y_i^s)\}_{i=1}^m$ constitué de m exemples *i.i.d.* selon P_S et un échantillon cible non étiqueté $T = \{(x_i^t)\}_{i=1}^m$ composé de m exemples *i.i.d.* selon D_T, en posant $S_u = \{x_i^s\}_{i=1}^m$ l'échantillon S privé de ses étiquettes, on veut minimiser :

$$\min_{\mathbf{w}} \; c\,m\,\mathrm{R}_S(G_{\rho_\mathbf{w}}) + a\,m\,\mathrm{dis}_{\rho_{\mathbf{w}^2}}(S_u, T_u) + \mathrm{KL}(\rho_\mathbf{w} \| \pi_0), \tag{7.5}$$

où $\mathrm{dis}_{\rho_{\mathbf{w}^2}}(S_u, T_u) = \left| \mathop{\mathrm{E}}_{(h,h')\sim\rho_{\mathbf{w}^2}} \mathrm{R}_{S_u}(h,h') - \mathop{\mathrm{E}}_{(h,h')\sim\rho_{\mathbf{w}^2}} \mathrm{R}_{T_u}(h,h') \right|$ est le désaccord empirique entre S_u et T_u spécialisé à une distribution $\rho_\mathbf{w}$ sur l'espace $\mathcal{H}$ des classifieurs linéaires considéré. Les réels $a > 0$ et $c > 0$ sont des hyperparamètres de l'algorithme. Notons que les constantes A et C du théorème 7.7 peuvent être retrouvées à partir de n'importe quelle valeur de a et c. Étant donnée la fonction $\ell_{\mathrm{dis}}(x) = 2\,\ell_{\mathbf{Erf}}(x)\,\ell_{\mathbf{Erf}}(-x)$ (illustrée sur la figure 7.1) , pour toute distribution D sur X, on a :

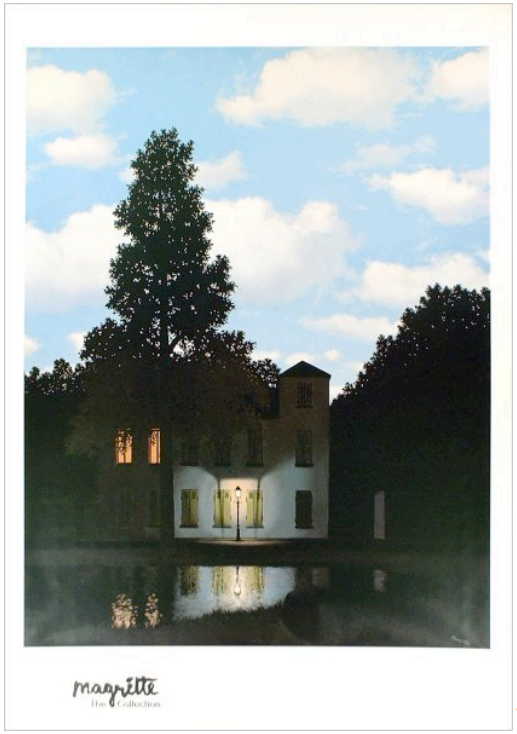
$$\begin{aligned}
\mathop{\mathrm{E}}_{(h,h')\sim\rho_{\mathbf{w}^2}} \mathrm{R}_D(h,h') &= \mathop{\mathrm{E}}_{\mathbf{x}\sim D} \mathop{\mathrm{E}}_{(h,h')\sim\rho_{\mathbf{w}^2}} \mathrm{I}\left[h(\mathbf{x}) \neq h'(\mathbf{x})\right] \\
&= 2 \mathop{\mathrm{E}}_{\mathbf{x}\sim D} \mathop{\mathrm{E}}_{(h,h')\sim\rho_{\mathbf{w}^2}} \mathrm{I}\left[h(\mathbf{x}) = 1\right] \mathrm{I}\left[h'(\mathbf{x}) = -1\right] \\
&= 2 \mathop{\mathrm{E}}_{\mathbf{x}\sim D} \mathop{\mathrm{E}}_{h\sim\rho_\mathbf{w}} \mathrm{I}\left[h(\mathbf{x}) = 1\right] \mathop{\mathrm{E}}_{h'\sim\rho_\mathbf{w}} \mathrm{I}\left[h'(\mathbf{x}) = -1\right] \\
&= 2 \mathop{\mathrm{E}}_{\mathbf{x}\sim D} \ell_{\mathbf{Erf}}\left(\frac{\langle \mathbf{w}, \mathbf{x}\rangle}{\|\mathbf{x}\|}\right) \ell_{\mathbf{Erf}}\left(-\frac{\langle \mathbf{w}, \mathbf{x}\rangle}{\|\mathbf{x}\|}\right) \\
&= \mathop{\mathrm{E}}_{\mathbf{x}\sim D} \ell_{\mathrm{dis}}\left(\frac{\langle \mathbf{w}, \mathbf{x}\rangle}{\|\mathbf{x}\|}\right).
\end{aligned}$$

Ainsi, trouver la solution optimale de l'équation (7.5) revient à chercher le vecteur $\mathbf{w}$ qui minimise :

$$c \sum_{i=1}^m \ell_{\mathbf{Erf}}\left(y_i^s \frac{\langle \mathbf{w}, \mathbf{x}_i^s\rangle}{\|\mathbf{x}_i^s\|}\right) + a \left| \sum_{i=1}^m \left[\ell_{\mathrm{dis}}\left(\frac{\langle \mathbf{w}, \mathbf{x}_i^s\rangle}{\|\mathbf{x}_i^s\|}\right) - \ell_{\mathrm{dis}}\left(\frac{\langle \mathbf{w}, \mathbf{x}_i^t\rangle}{\|\mathbf{x}_i^t\|}\right) \right] \right| + \frac{\|\mathbf{w}\|^2}{2}. \tag{7.6}$$

L'équation précédente est fortement non convexe. Afin de rendre sa résolution plus facilement contrôlable, nous remplaçons la fonction $\ell_{\mathbf{Erf}}(\cdot)$ par sa relaxation convexe $\ell_{\mathbf{Erf}_{\mathrm{cvx}}}(\cdot)$ (comme pour PBGD3 et illustrée sur la figure 7.1). L'optimisation se réalise ensuite par une descente de gradient. Le gradient de l'équation 7.6 étant :

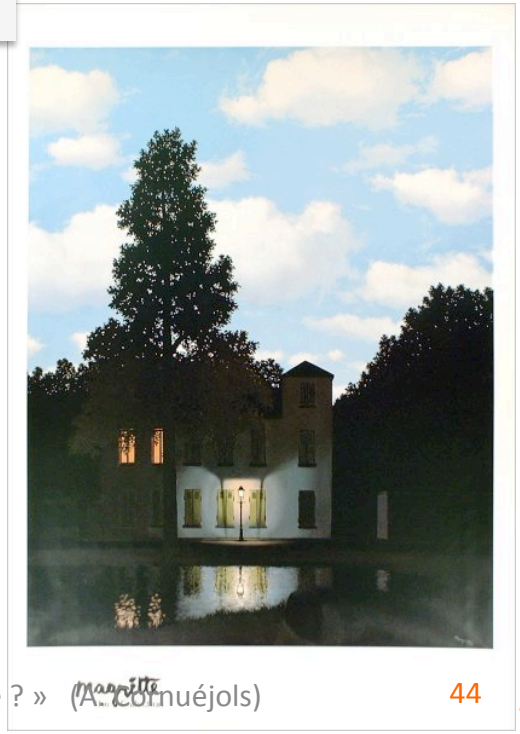
Des garanties « de lampadaire »



Des garanties « de lampadaire »

(*Quasi*) garantie que :

- **Si** le monde satisfait **mes attentes** sur lui
- **Alors** l'algorithme d'apprentissage produira une bonne hypothèse (proche de la vraie)



Des garanties « de lampadaire »

(*Quasi*) garantie que :

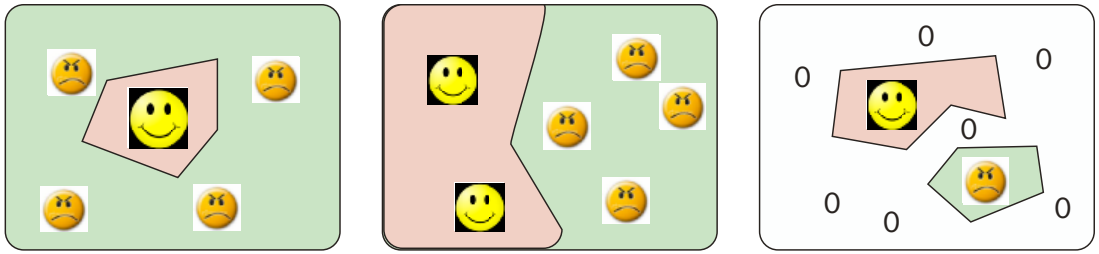
- **Si** le monde satisfait **mes attentes** sur lui
- **Alors** l'algorithme d'apprentissage produira une bonne hypothèse (proche de la vraie)

- **Autrement** l'apprentissage peut conduire à de très mauvaises hypothèses

  (ex. *Si le monde **n'est pas** parcimonieux*)

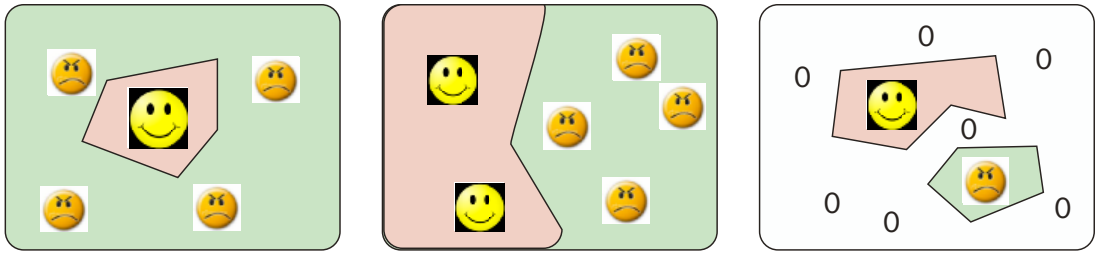
Le no-free-lunch theorem

Possible

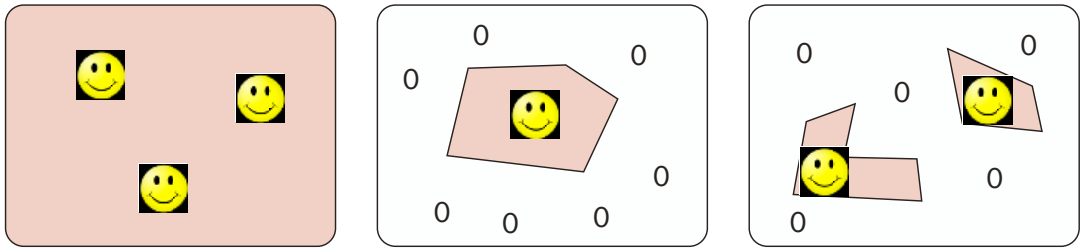


Le no-free-lunch theorem

Possible



Impossible



Il faut **choisir** le **bon algorithme** pour la **classe de problèmes** étudiée

- Très mathématisé et sophistiqué
- Conduit à l'invention de nouveaux algorithmes

Est-ce que cela permet de répondre aux **nouveaux défis** ?

Est-ce que l'on a **toujours été** dans ce paradigme ?

- **Paradigme** :

*Manière jugée cohérente de **percevoir** le monde et de le **représenter**,*

*ensemble de **valeurs** et **techniques** qui sont **partagées** par les membres d'une*

*communauté scientifique, au cours d'une **période de consensus** théorique*

Plan

1. L'IA est **attendue partout**
2. Quelles **garanties** sur l'induction
3. Un peu d'**histoire**
4. Et **demain** ? Prémisses de changement de paradigme
5. Retour sur les **défis**
6. Conclusions … et ouverture

Objectifs et méthode

- Question

  Le paradigme : centré sur **l'espérance de coût d'usage**

  1. De **quand** ça date ?
  2. Y a-t-il eu un **avant** ? Et, si oui, lequel ?
  3. Peut-on détecter des **changements de paradigmes** ?

- Méthode
  - Examiner les publications depuis ~1950
    - Western Joint Computer Conference (1955)
    - IJCAI 1969-2016
    - ICML
  - Que cherche t-on à **prouver** ?
  - Quels **critères d'évaluation** ?

Perspective historique : apprentissage automatique

1950s

- Expériences de pensée sous contrainte de réalisabilité computationnelle
  - **Opérateurs** sur des représentation
  - Recherche dans un **espace d'états**. Buts / sous-buts
  - **Apprentissage** par mutations aléatoires (mais guidée par ressemblance) ~ AG
  - Ivresse : comprendre la pensée
  - « *Our problem, our joint problem, is to discover what transformations must be made on the available data in order to preserve intact the significant features and to discard the irrelevant details* ».

1960s

- Principes, théorèmes et démonstrations (Checker. Problèmes « jouets »)
  - Reconnaissance des formes. Plutôt numérique (bayésien, perceptron)

1970s

Perspective historique : apprentissage automatique

1950s

- Expériences de pensée sous contrainte de réalisabilité computationnelle
  - **Opérateurs** sur des représentation
  - Recherche dans un **espace d'états**. Buts / sous-buts
  - **Apprentissage** par mutations aléatoires (mais guidée par ressemblance) ~ AG
  - Ivresse : comprendre la pensée
  - « *Our problem, our joint problem, is to discover what transformations must be made on the available data in order to preserve intact the significant features and to discard the irrelevant details* ».

1960s

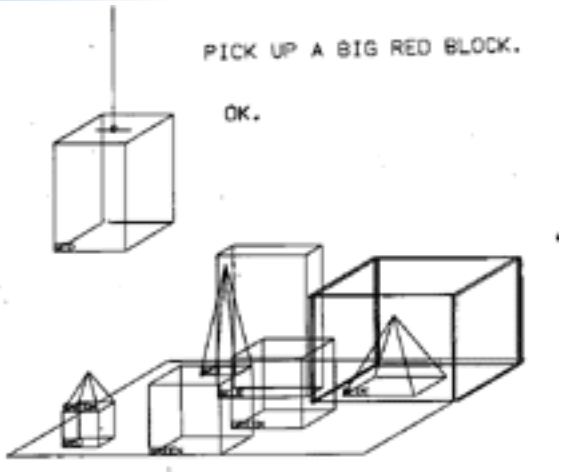
- Principes, théorèmes et démonstrations (Checker. Problèmes « jouets »)
  - Reconnaissance des formes. Plutôt numérique (bayésien, perceptron)

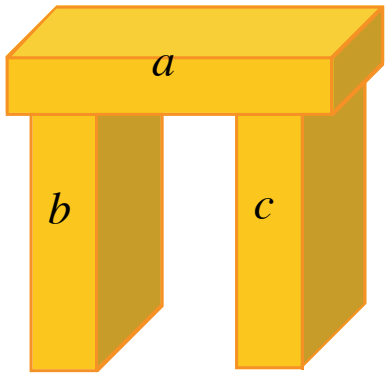
1970s

- Expertise. Sciences cognitives : plausibilité. Intégration dans le raisonnement
  - **Modèles de mémoire**. Réseaux sémantiques. Représentation des connaissances
  - **Règles** de production. Moteur d'inférence.
  - **Mécanismes** d'apprentissage et de **généralisation**
  - **Apprentissage et raisonnement** : heuristiques, macro-opérateurs, chunking

ARCH [Winston, 1970]

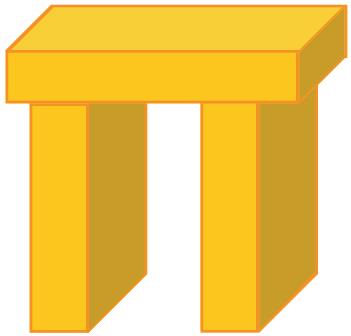
- **Apprentissage** de **concept** (e.g. arche) dans un monde de blocs

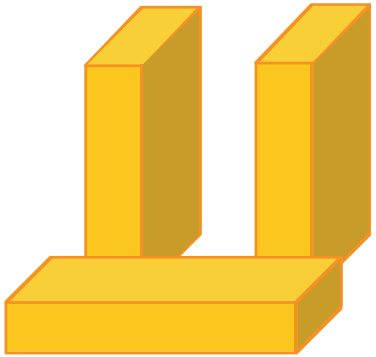




ARCH [Winston, 1970]

- Les exemples ne sont **pas choisis au hasard**



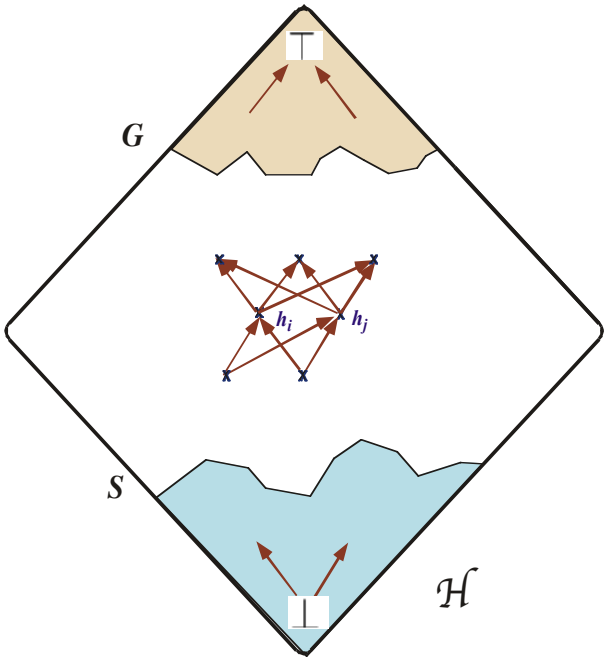


Apprentissage de l'espace des versions [Tom Mitchell, 1979]

Observation fondamentale :

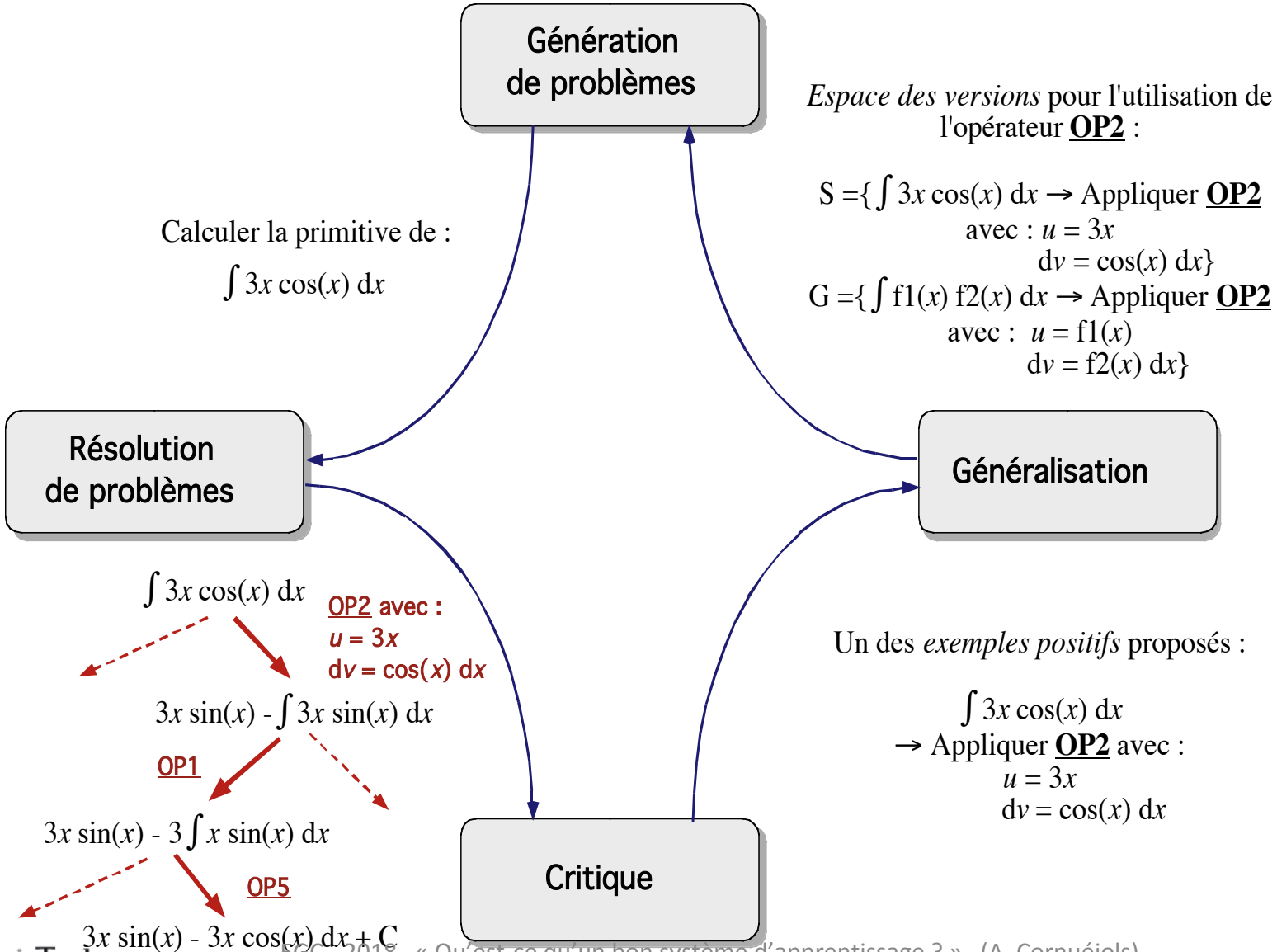
L'espace des versions structuré par une relation d'ordre partiel peut être représenté par :

- sa **borne supérieure** : le *G-set*
- sa **borne inférieure** : le *S-set*

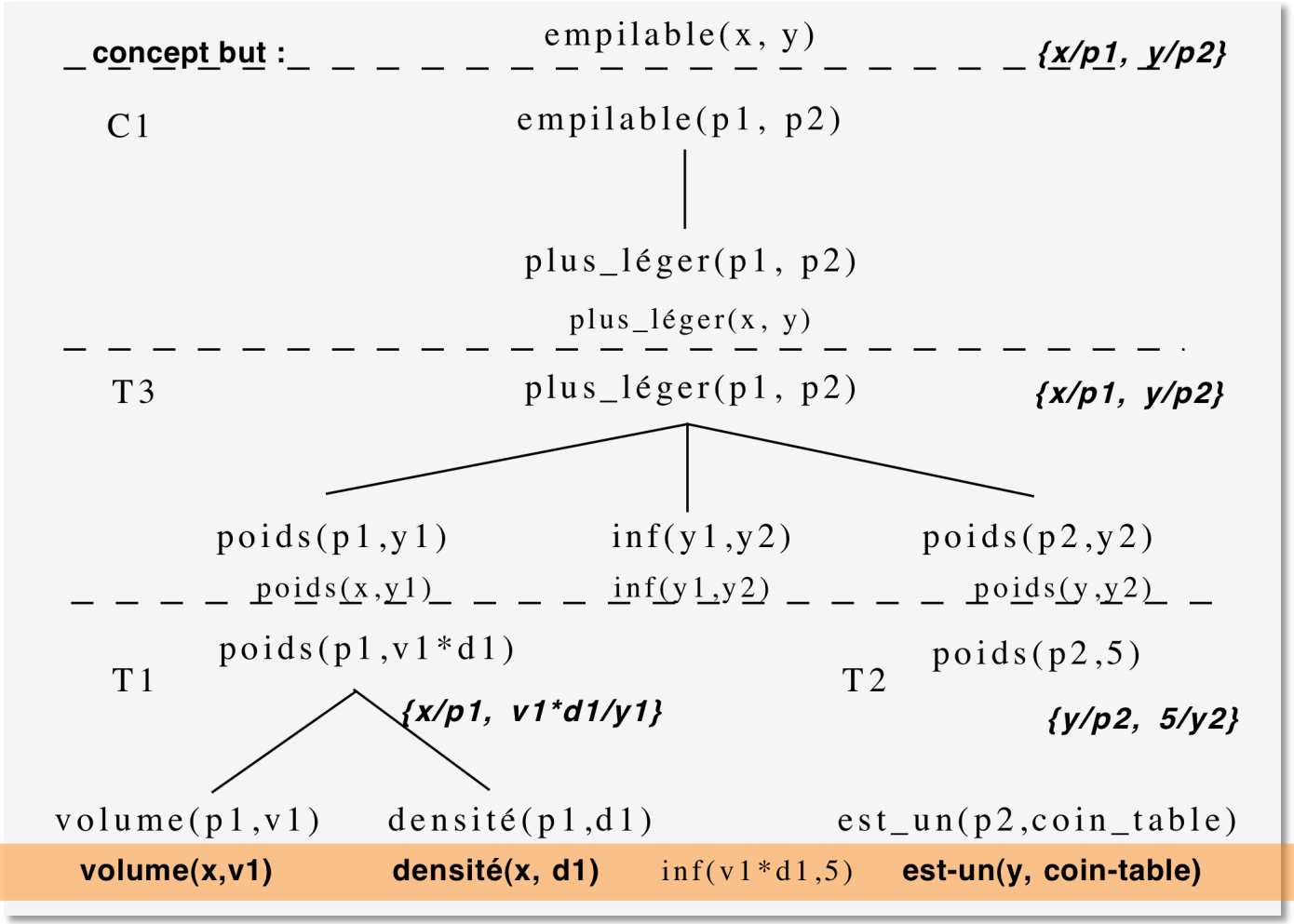


- *G-set = Ensemble de toutes les hypothèses **les plus** **générales** cohérentes avec les exemples connus*
- *S-set = Ensemble de toutes les hypothèses **les plus** **spécifiques** cohérentes avec les exemples connus*

Le système LEX [Tom Mitchell, 1983]



Explanation-Based Learning



Generalized search tree resulting from regression of the target concept in the proof tree by computing at each step the most general literals allowing this step.

Perspective historique : apprentissage automatique

1990s

- Espérance d'erreur de prédiction
  - Bases de données supposées i.i.d.
  - Données bruitées
  - Espérance du coût de prédiction
  - Théorie : quels principes inductifs ? Donc quel problème d'optimisation
  - Comment accélérer le processus d'optimisation : convexification
  - Nouvelles méthodes : SVM / Boosting / random forests / Lasso, …
  - Plus de raisonnement !!!

2000s

2010s

- Applications à des problèmes « réels »

2020s

Perspective historique : apprentissage automatique

1990s

- Espérance d'erreur de prédiction
  - Bases de données supposées i.i.d.
  - Données bruitées
  - Espérance du coût de prédiction
  - Théorie : quels principes inductifs ? Donc quel problème d'optimisation
  - Comment accélérer le processus d'optimisation : convexification
  - Nouvelles méthodes : SVM / Boosting / random forests / Lasso, …
  - Plus de raisonnement !!!

2000s

- Applications à des problèmes « réels »

2010s

- L'apprentissage « profond » : apprendre les (hiérarchies de) descripteurs
  - Performances en prédiction
  - Exhiber les descripteurs (et espaces « sémantiques »)
  - Transfert

2020s

- … ?

Changements de paradigmes

- Changements de **paradigme**
  - Très **progressifs** !!
  - Au début : signaux faibles

Changements de paradigmes

- Changements de **paradigme**
  - Très **progressifs** !!
  - Au début : signaux faibles

- Changements de **méthodes**
  - Très **rapides** (modes) !!
  - **Au sein** d'un paradigme
    - SVM ; Boosting
  - Facilitant une **charnière**
    - RNs profonds ???