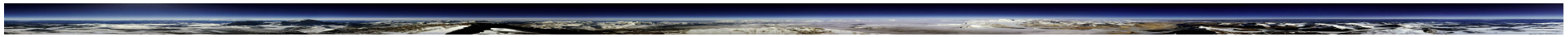
L'induction hier, aujourd'hui, demain :

Que **décide-t-on** de chercher à prouver ?



Antoine Cornuéjols

AgroParisTech – INRA MIA 518

antoine.cornuejols@agroparistech.fr

Le rôle de l'induction

- [Leslie Valiant, « *Probably Approximately Correct. Nature's Algorithms for Learning and Prospering in a Complex World* », Basic Books, 2013]

« From this, we have to conclude that **generalization** or **induction** is a **pervasive phenomenon** (…). It is as routine and reproducible a phenomenon as objects falling under gravity.

It is **reasonable to expect a quantitative scientific explanation** of this highly reproducible phenomenon. »

Le rôle de l'induction

- [Edwin T. Jaynes, « *Probability theory. The logic of science* », Cambridge U. Press, 2003], p.3

« We are hardly able to get through one waking hour without facing some situation (e.g. *will it rain or won't it?*) where we **do not have enough information to permit deductive reasoning**; but still we must decide immediately.

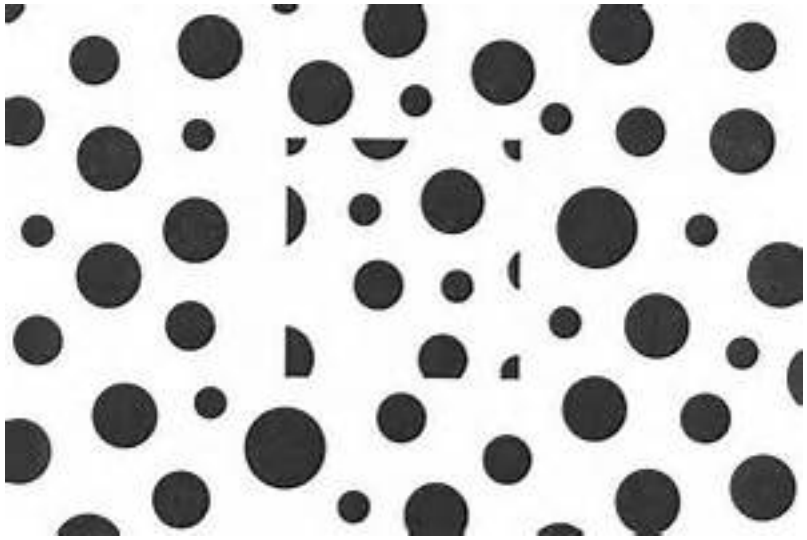
In spite of its familiarity, the formation of plausible conclusions is a **very subtle process**. »

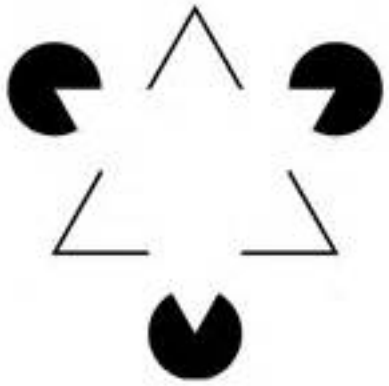
Trame

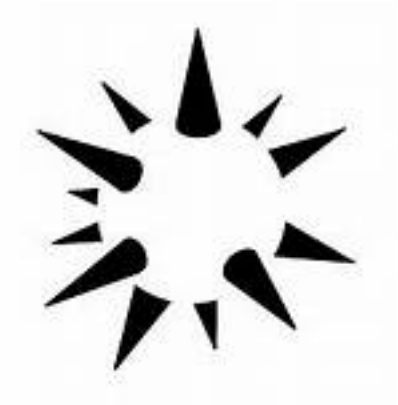
1. L'**induction**: omniprésence et faillibilité
2. Le **no-free-lunch theorem**
3. **Approches** de l'induction (hier et aujourd'hui)
   - Le **Perceptron**
   - La **théorie statistique** de l'apprentissage
   - Le **paradigme** dominant
   - Un point de vue indépassable ? Le **cas de l'EBL**
4. Quelle perspective pour **l'avenir** ?
5. Conclusion

Induction(s) : Illustrations

Interprétation – complétion de percepts

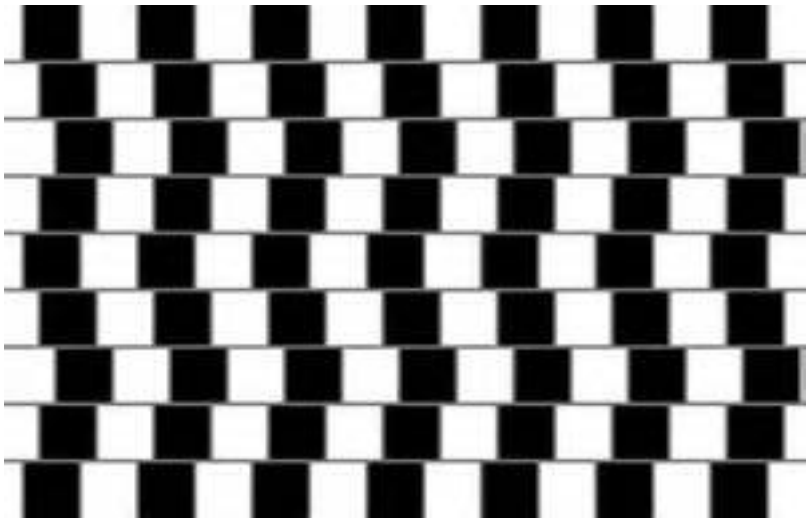




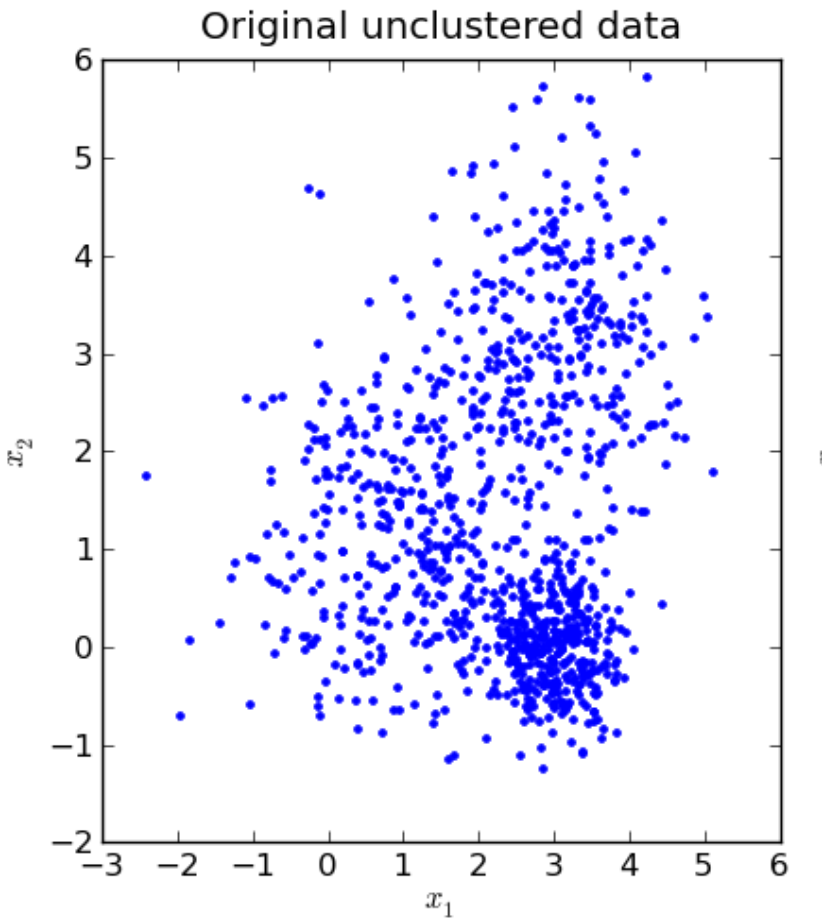


Interprétation – complétion de percepts

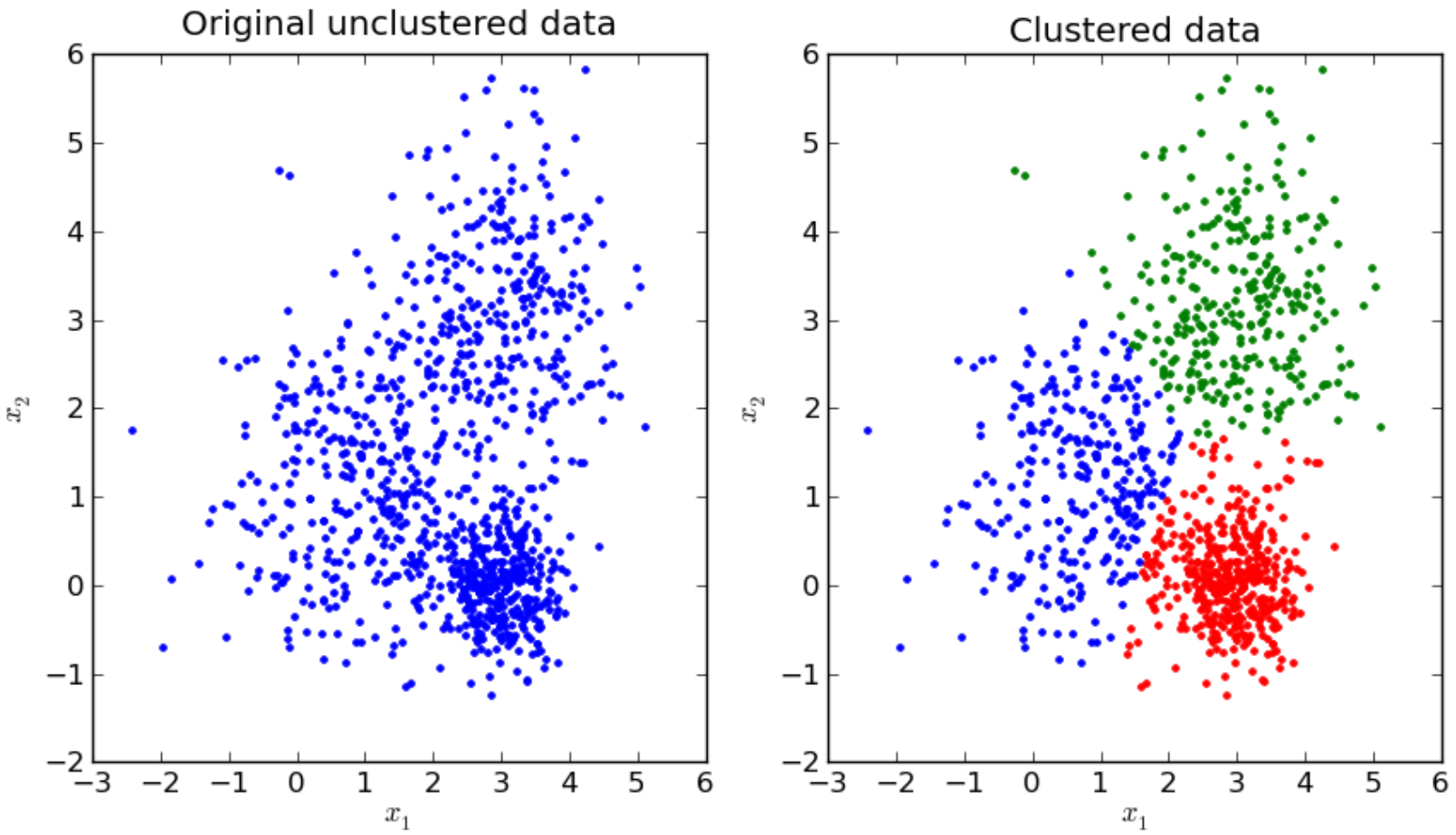
Illusions d'optique



Clustering

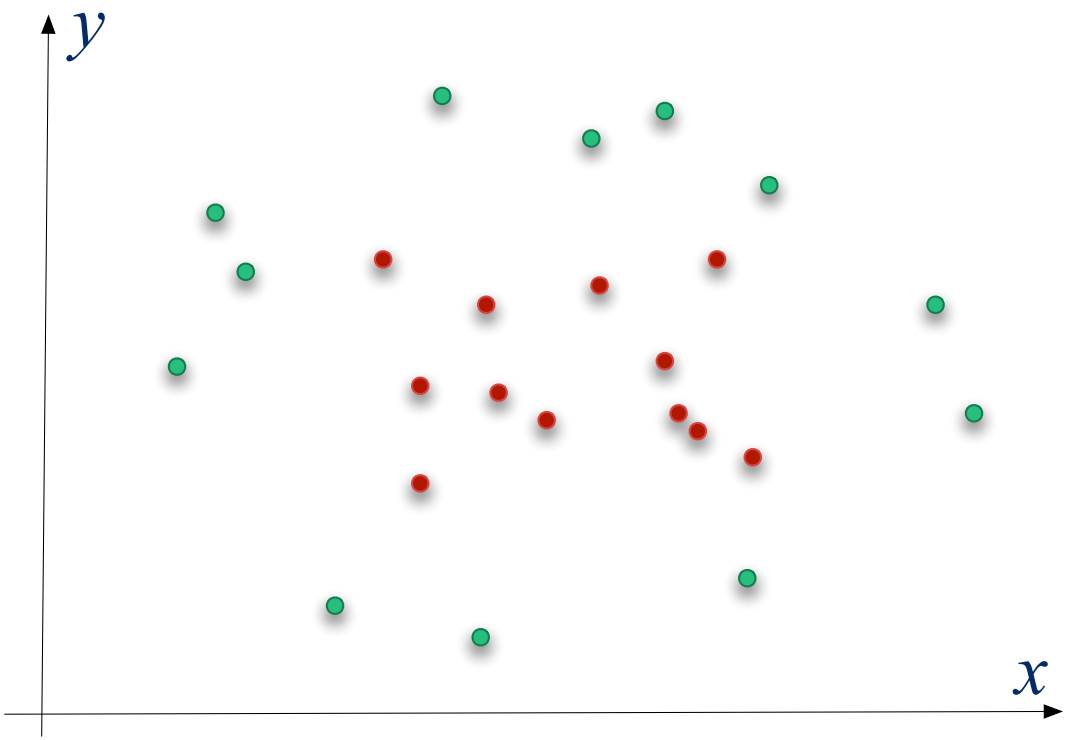


Clustering

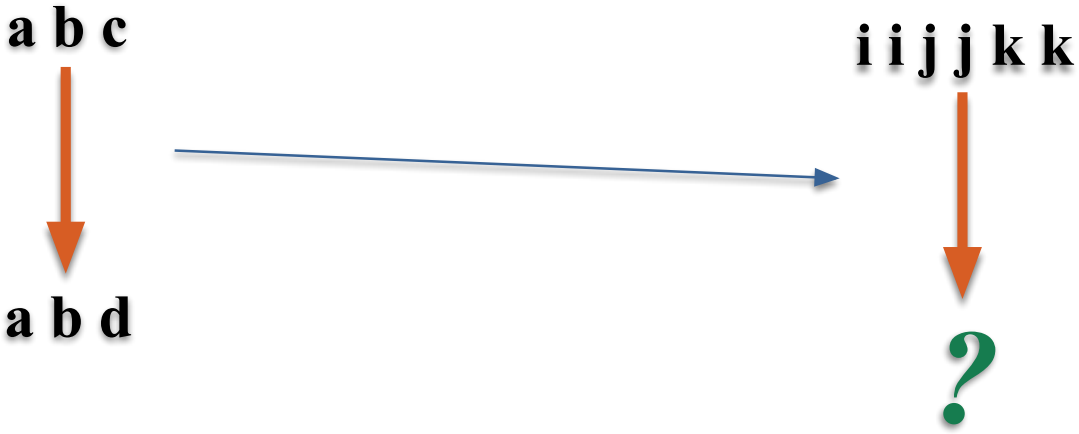


Induction supervisée

- Comment choisir la fonction de décision ?



Transfert et analogie



- a b d
- i i j j k d
- i i j j k l
- i i j j k k
- ?

Séquences

- 1 1 2 3 5 8 13 21 …
- 1 2 3 5 …
- 1 1 1 2 1 1 2 1 1 1 1 1 2 2 1 3 1 2 2 1 1 …

Interrogations

À chaque fois :

Cas particuliers => **loi** générale ou adaptation à **nouveau cas**

1. Qu'est-ce qui **autorise** ce passage ?

2. Est-ce que l'on peut **garantir quelque chose ?**

Trame

1. L'**induction**: omniprésence et faillibilité
2. Le **no-free-lunch theorem**
3. **Approches** de l'induction (hier et aujourd'hui)
   - Le **Perceptron**
   - La **théorie statistique** de l'apprentissage
   - Le **paradigme** dominant
   - Un point de vue indépassable ? Le **cas de l'EBL**
4. Quelle perspective pour **l'avenir** ?
5. Conclusion

Le no-free-lunch theorem

Le no-free-lunch theorem

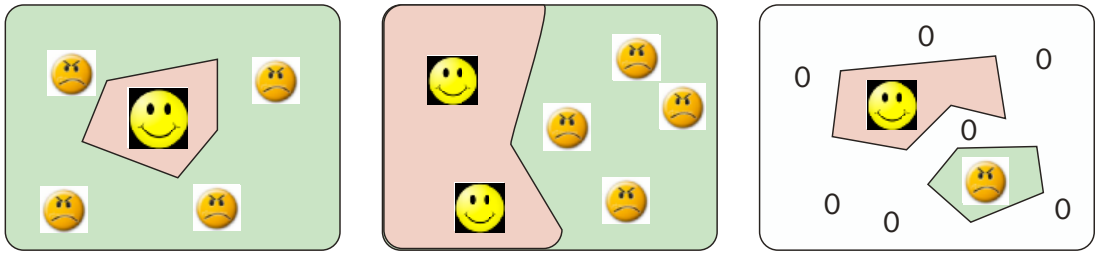
Théorème 2.1 (No-free-lunch theorem (Wolpert, 1992))

Pour tout couple d'algorithmes d'apprentissage $\mathcal{A}_1$ *et* $\mathcal{A}_2$*, caractérisés par leur distribution de probabilité* a posteriori $\mathbf{p}_1(h|\mathcal{S})$ *et* $\mathbf{p}_2(h|\mathcal{S})$*, et pour toute distribution* $d_{\mathcal{X}}$ *des formes d'entrées* $\mathbf{x}$ *et tout nombre* m *d'exemples d'apprentissage, les propositions suivantes sont vraies :*

1. *En moyenne uniforme sur toutes les fonctions cible* f *dans* $\mathcal{F}$ *:*
   $\mathbb{E}_1[R_{\text{Réel}}|f, m] - \mathbb{E}_2[R_{\text{Réel}}|f, m] = 0.$
2. *Pour tout échantillon d'apprentissage* $\mathcal{S}$ *donné, en moyenne uniforme sur toutes les fonctions cible* f *dans* $\mathcal{F}$ *:* $\mathbb{E}_1[R_{\text{Réel}}|f, \mathcal{S}] - \mathbb{E}_2[R_{\text{Réel}}|f, \mathcal{S}] = 0.$
3. *En moyenne uniforme sur toutes les distributions possibles* $\mathbf{P}(f)$ *:*
   $\mathbb{E}_1[R_{\text{Réel}}|m] - \mathbb{E}_2[R_{\text{Réel}}|m] = 0.$
4. *Pour tout échantillon d'apprentissage* $\mathcal{S}$ *donné, en moyenne uniforme sur toutes les distributions possibles* $\mathbf{p}(f)$ *:* $\mathbb{E}_1[R_{\text{Réel}}|\mathcal{S}] - \mathbb{E}_2[R_{\text{Réel}}|\mathcal{S}] = 0.$

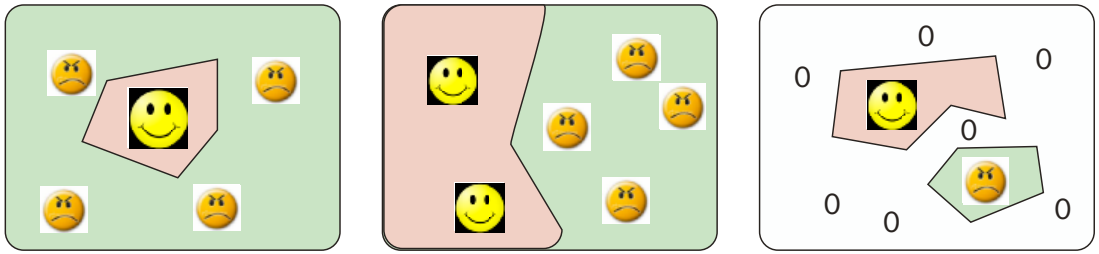
Le no-free-lunch theorem

Possible

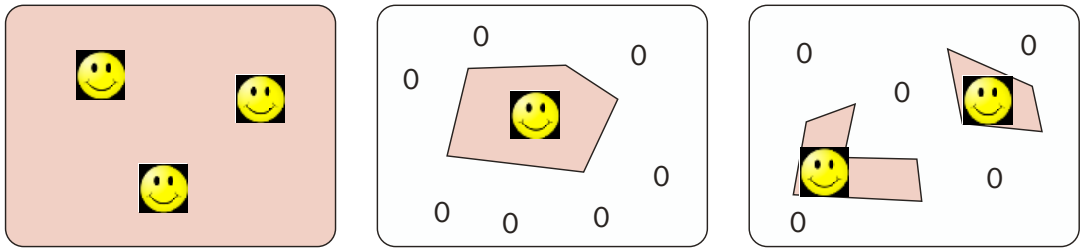


Le no-free-lunch theorem

Possible



Impossible



Déduction !

1. **Tous les algorithmes** inductifs **se valent**

2. Il ne peut y avoir **aucune garantie sur les inductions** réalisées

Allons à la plage !!

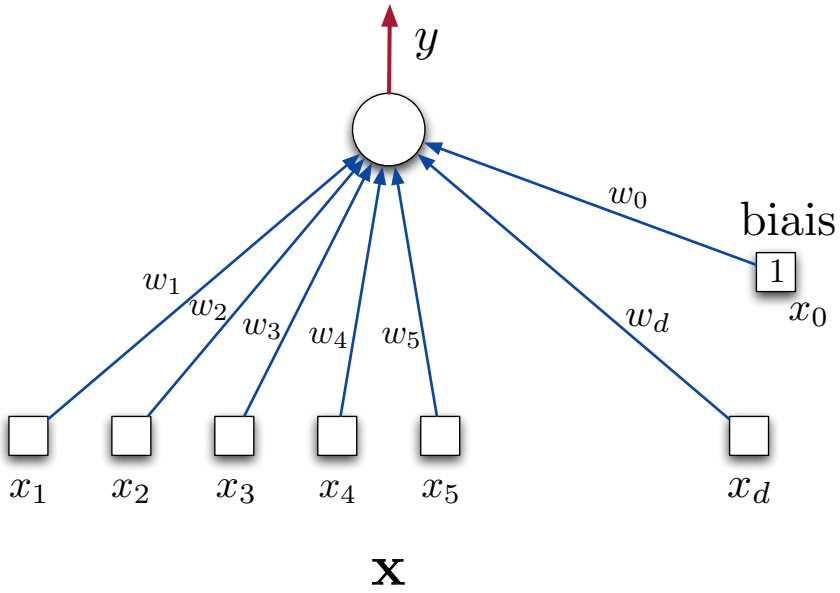
Trame

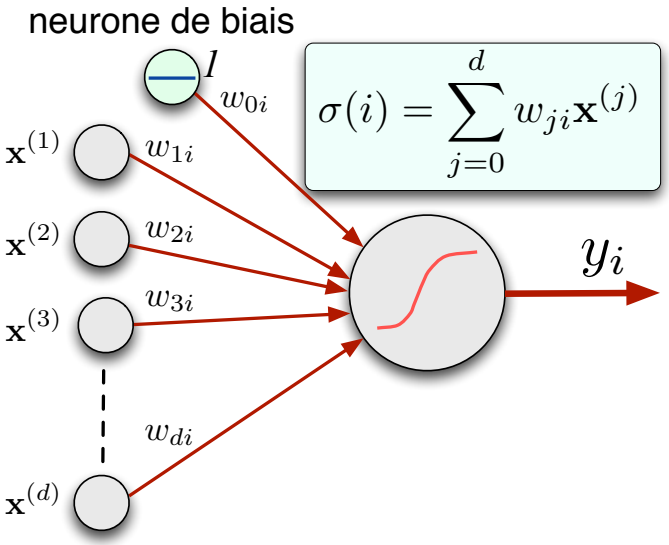
1. L'**induction**: omniprésence et faillibilité
2. Le **no-free-lunch theorem**
3. **Approches** de l'induction (hier et aujourd'hui)
   - Le **Perceptron**
   - La **théorie statistique** de l'apprentissage
   - Le **paradigme** dominant
   - Un point de vue indépassable ? Le **cas de l'EBL**
4. Quelle perspective pour **l'avenir** ?
5. Conclusion

Le perceptron

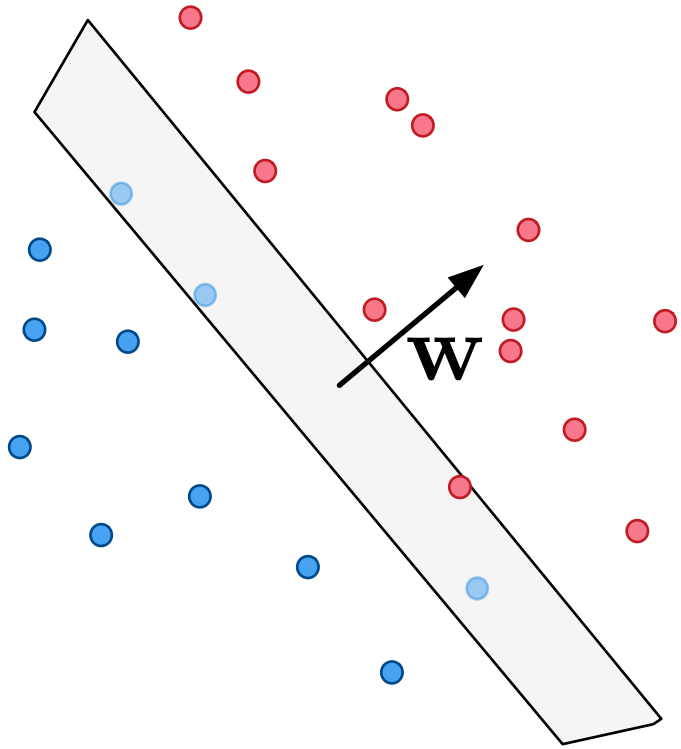
Le perceptron

– Rosenblatt (1958-1962)





Le perceptron : un discriminant linéaire



Le perceptron

- **Apprentissage des poids** w_i
  - Principe (*règle de Hebb*) : en cas de succès, ajouter à chaque connexion quelque chose de proportionnel à l'entrée et à la sortie

Règle du perceptron : apprendre seulement en cas d'échec

Algorithme 1 : Algorithme d'apprentissage du perceptron

tant que *non convergence* **faire**
  si *la forme d'entrée est correctement classée* **alors**
    ne rien faire
  sinon
    $\mathbf{w}(t+1) = \mathbf{w}(t) + \eta\, \mathbf{x}_i\, y_i$
  fin
  Passer à la forme d'apprentissage suivante
fin

Des propriétés remarquables !!

- **Convergence** en un nombre fini d'étapes
  - Indépendamment du **nombre** d'exemples
  - Indépendamment de la **distribution** des exemples
  - Indépendamment de la **dimension** de l'espace d'entrée

!!!

Si il existe au moins *une séparatrice linéaire des exemples*

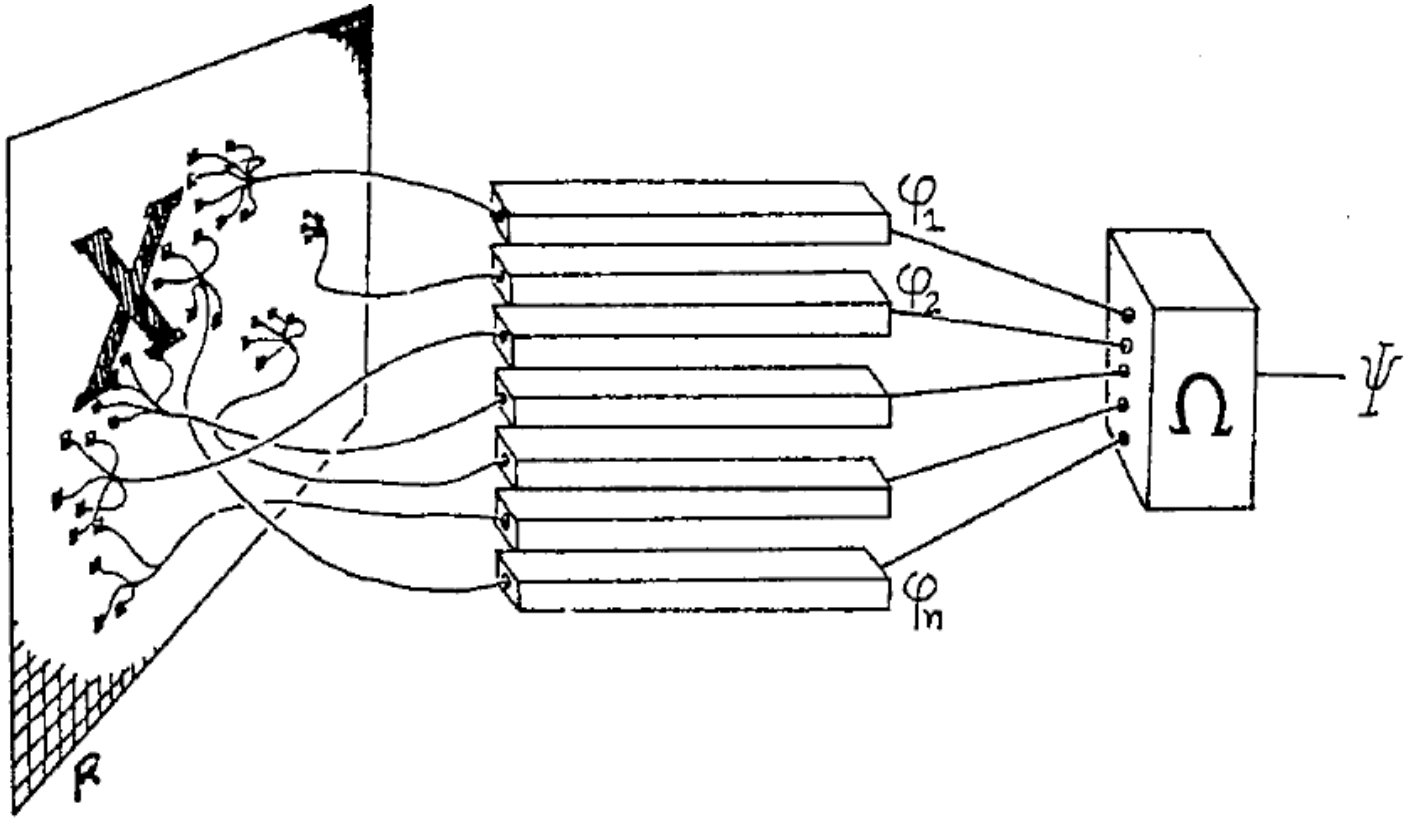
Garantie de généralisation ??

- Théorèmes sur la performance
  par rapport à l'échantillon d'apprentissage

- Mais qu'en est-il pour des **exemples à venir** ?

Le Perceptron

– Rosenblatt (1958-1962)



Théorie statistique

de l'apprentissage

(illustration)

Encore un autre exemple

- Exemples décrits par :
  - *nombre* (1 ou 2); *taille* (petit ou grand); *forme* (cercle ou carré); *couleur* (rouge ou vert)
- Les objets appartiennent soit à la classe + soit à la classe -

Description	Votre réponse	Vraie réponse
1 grand carré rouge		-
1 grand carré vert		+
2 petits carrés rouges		+
2 grands cercles rouges		-
1 grand cercle vert		+
1 petit cercle rouge		+
1 petit carré vert		-
1 petit carré rouge		+
2 grands carrés verts		+

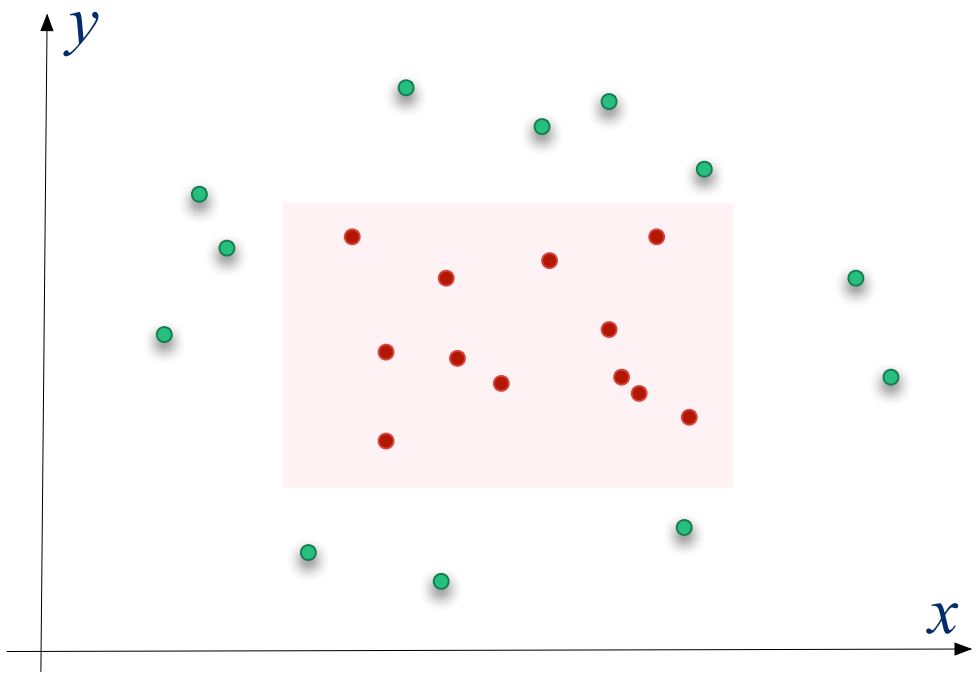
L'induction : un jeu impossible ?

- **Nécessité** d'un biais

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

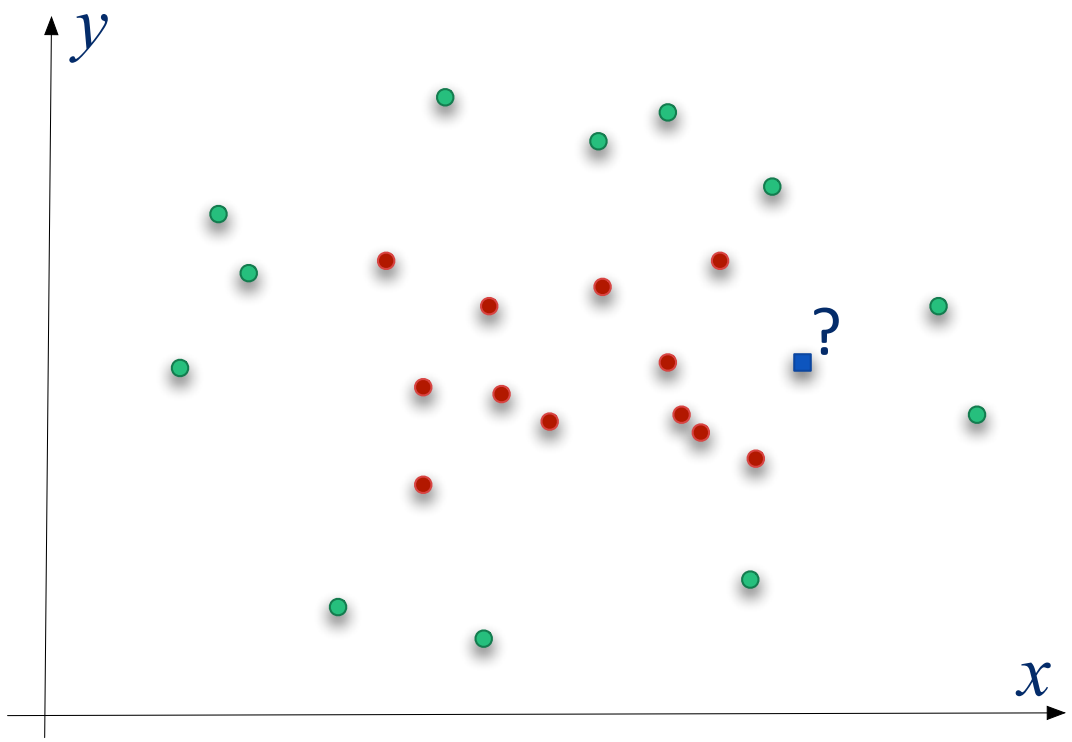
Apprentissage de rectangle

- Échantillon
  - D'exemples positifs $\mathbf{P}_x^+$
  - D'exemples négatifs $\mathbf{P}_x^-$



Apprentissage de rectangle

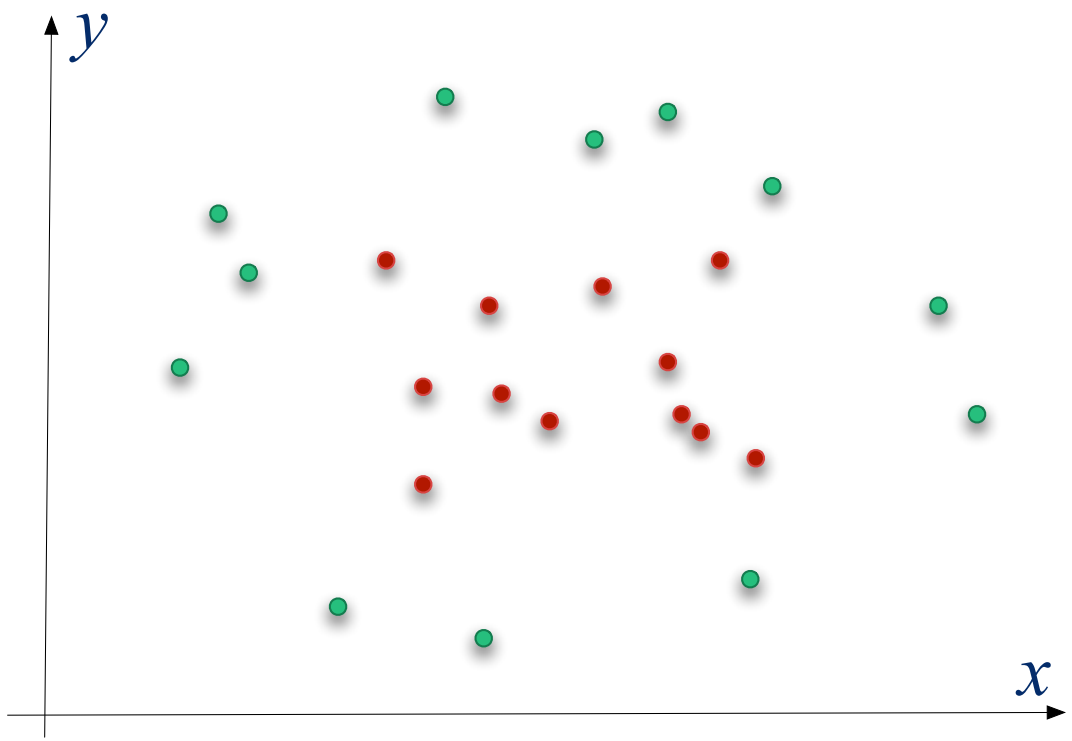
- Que cherche-t-on à apprendre ?



→ Une fonction de **décision** (de **prédiction**)

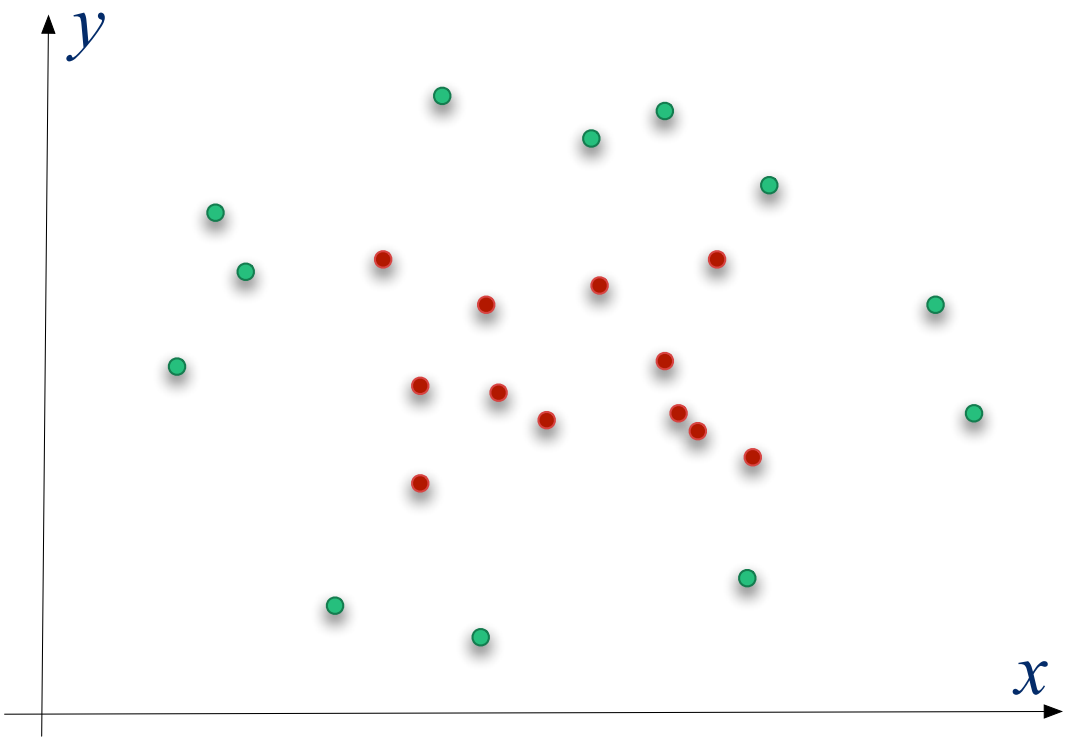
Apprentissage de rectangle

- Comment apprendre ?



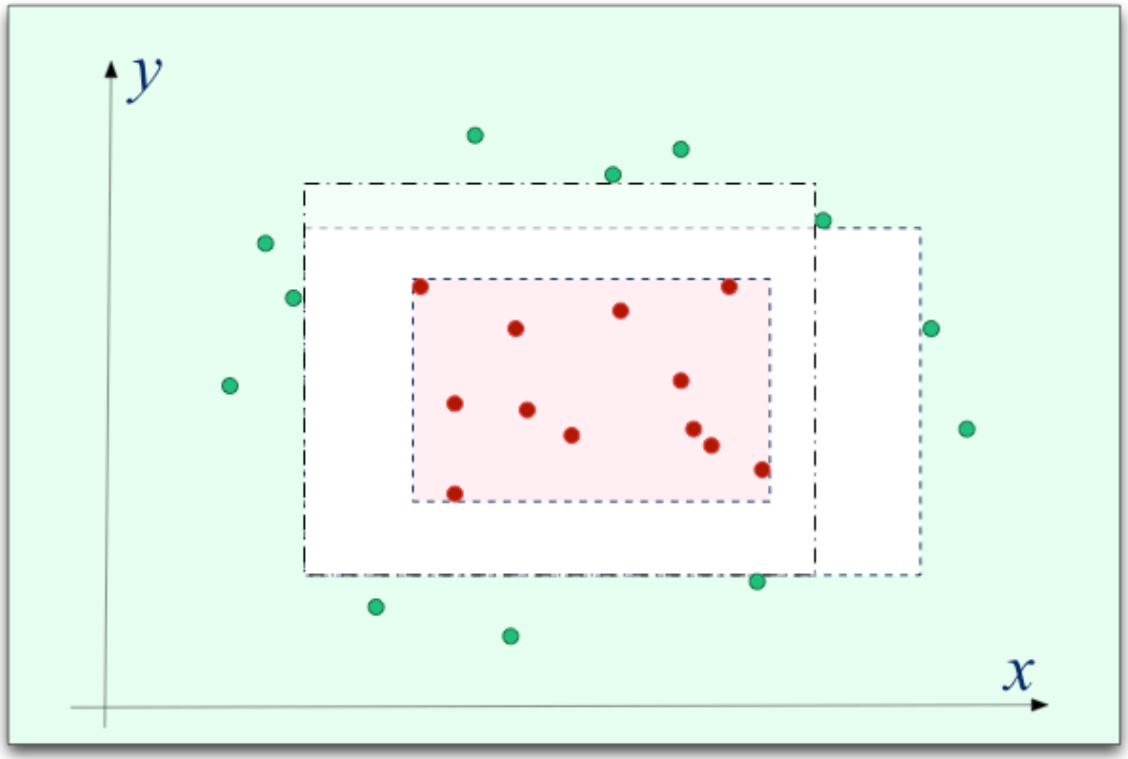
Apprentissage de rectangle

- Comment apprendre ?
  - Si **je sais que le concept cible est un rectangle**



Apprentissage de rectangle

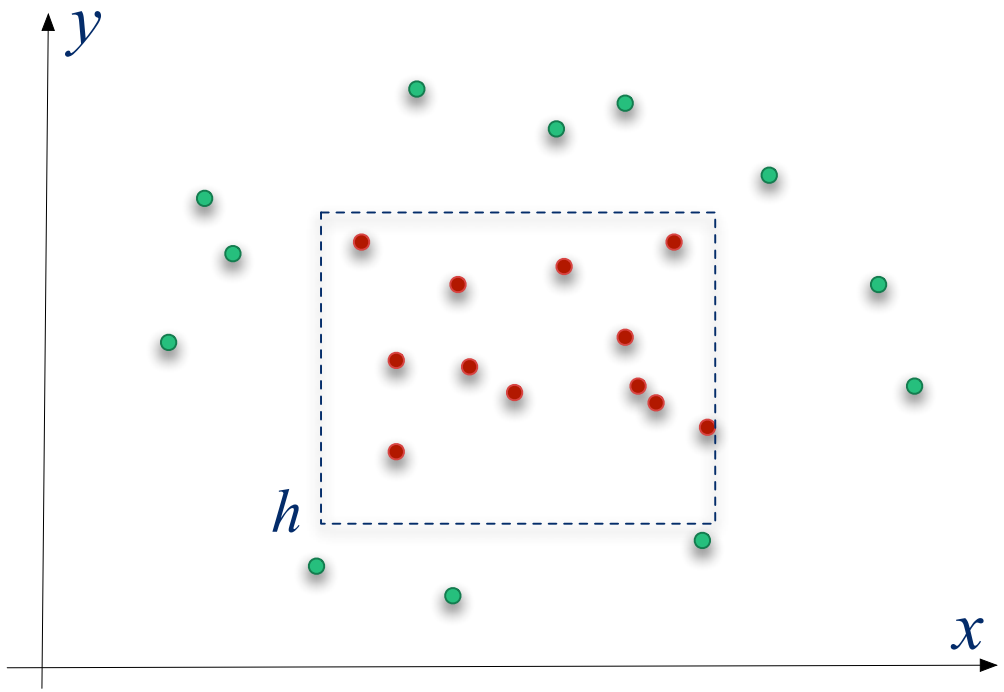
- Comment apprendre ?
  - Choix d'une hypothèse *h*



Espace des versions

Apprentissage de rectangle

- Apprentissage : choix de *h*
  - Quelle performance ?



Théorie statistique de l'apprentissage

Une valse à trois temps

Étude statistique de l'induction

Quelle performance ?

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

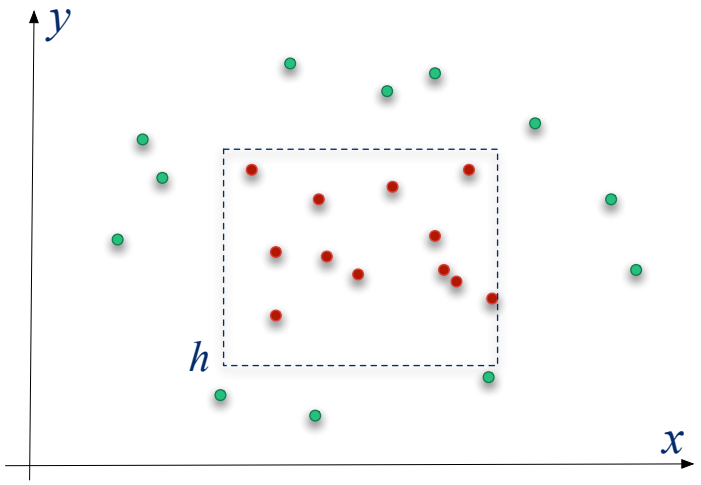
$$\ell\big(h(\mathbf{x}), y\big)$$

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

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

Étude statistique de l'induction

- La performance **empirique** de *h*
  - E.g. Pas d'erreur sur l'échantillon d'apprentissage *S*



Le « risque empirique »

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

Question centrale : le principe inductif

- Le principe de minimisation du risque empirique (ERM)

  … est-il sain ?

  - **Si** je choisis *h* telle que $\hat{h} = \underset{h \in \mathcal{H}}{\text{ArgMin}} \ \hat{R}(h)$
  - Est-ce que *h* est bonne relativement au risque réel ?

$$\hat{R}(\hat{h}) \overset{?}{\longleftrightarrow} R(\hat{h})$$

  - Est-ce que j'aurais pu faire beaucoup mieux ? $h^* = \underset{h \in \mathcal{H}}{\text{ArgMin}} \ R(h)$

$$R(h^*) \overset{?}{\longleftrightarrow} R(\hat{h})$$

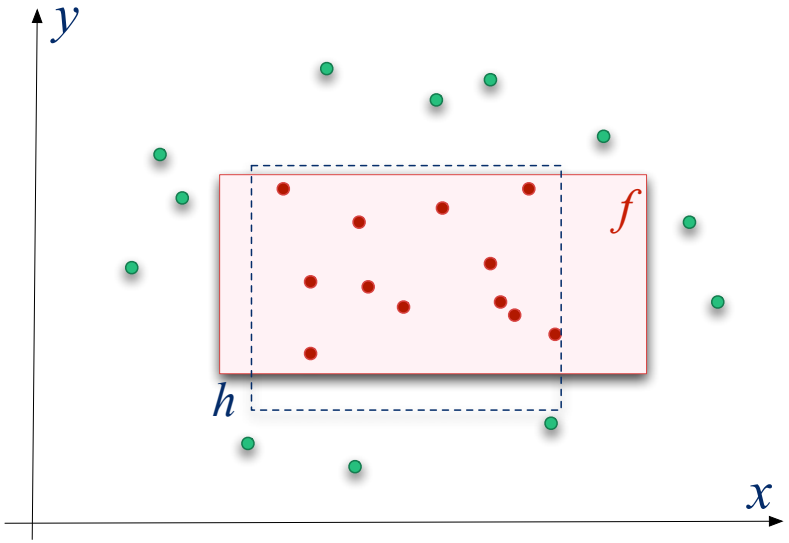
Théorie statistique de l'apprentissage

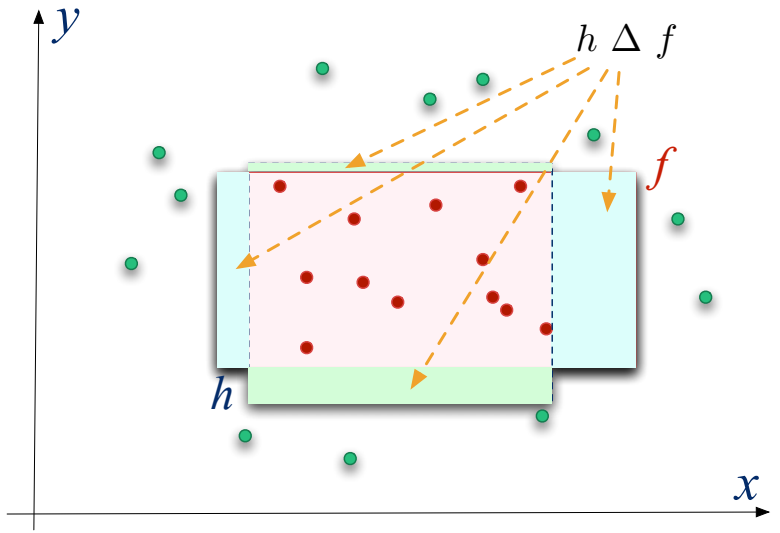
Le 1er temps

Un **individu**

Étude statistique pour UNE hypothèse

- choix d'une hypothèse de risque empirique nul (pas d'erreur sur l'échantillon d'apprentissage *S*)
- Quelle performance attendue pour *h* ?
- Quel est le risque d'avoir une erreur $R(h) > \varepsilon$?





Étude statistique pour UNE hypothèse

- Supposons *h* tq. $R(h) \geq \varepsilon$ (*h* « mauvaise »)
- Quelle est la probabilité que pourtant *h* ait été sélectionnée ?

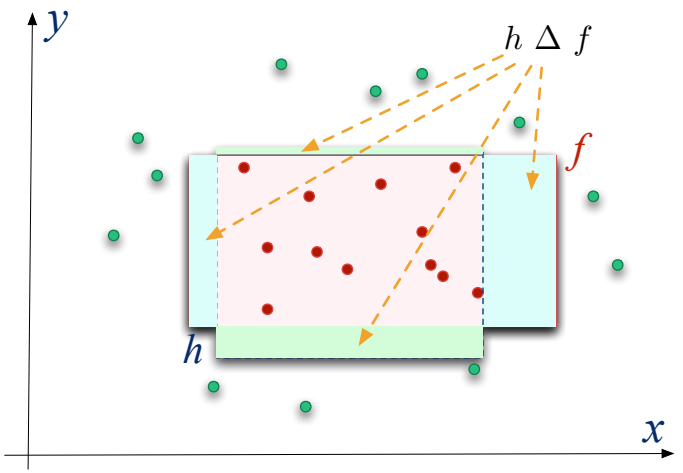
$$R(h) = \mathbf{p}_{\mathcal{X}}(h \,\Delta\, f)$$

Après un exemple : $p\big(\hat{R}(h) = 0\big) \leq 1 - \varepsilon$

« tombe » en dehors de $h \,\Delta\, f$

Après *m* exemple (**i.i.d.**) :

$$p^m\big(\hat{R}(h) = 0\big) \leq (1 - \varepsilon)^m$$



On veut : $\forall\, \varepsilon, \delta \in [0,1] : \quad p^m\big(R(h) \geq \varepsilon\big) \leq \delta$

Étude statistique pour UNE hypothèse

- On cherche : $\forall\, \varepsilon, \delta \in [0,1] : \quad p^m\big(R(h) \geq \varepsilon\big) \leq \delta$

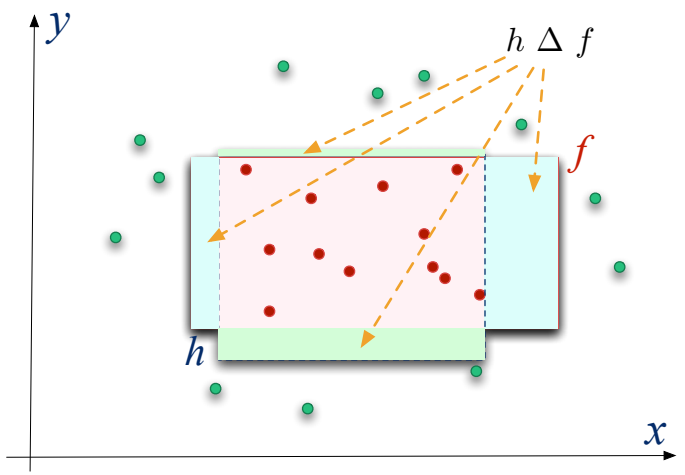
Soit :

$$(1-\varepsilon)^m \leq \delta$$

$$e^{-\varepsilon m} \leq \delta$$

$$-\varepsilon m \leq \ln(\delta)$$

D'où : $$m \geq \frac{\ln(1/\delta)}{\varepsilon}$$



Théorie statistique de l'apprentissage

Le 2ème temps

Quel individu dans la **Foule**

Étude statistique pour $|\mathcal{H}|$ hypothèses

- Quelle est la probabilité que je choisisse une hypothèse h_{err} de risque réel > ε **et que je ne m'en aperçoive pas** après l'observation de *m* exemples ?
- Probabilité de survie de h_{err} **après 1 exemple** : $(1-\varepsilon)$
- Probabilité de survie de h_{err} **après *m* exemples** : $(1-\varepsilon)^m$
- Probabilité de survie d'**au moins une hypothèse dans $\mathcal{H}$** : $|\mathcal{H}|\,(1-\varepsilon)^m$
  - On utilise la probabilité de l'union

$$\mathbf{P}(A \cup B) \leq \mathbf{P}(A) + \mathbf{P}(B)$$

- On veut que la probabilité qu'il reste au moins une hypothèse de risque réel > ε dans l'espace des versions soit bornée par δ :

$$|\mathcal{H}|\,(1-\varepsilon)^m < |\mathcal{H}|e^{(-\varepsilon\, m)} < \delta$$

$$\log|\mathcal{H}| - \varepsilon\, m < \log\delta$$

$$m > \frac{1}{\varepsilon}\log\frac{|\mathcal{H}|}{\delta}$$

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

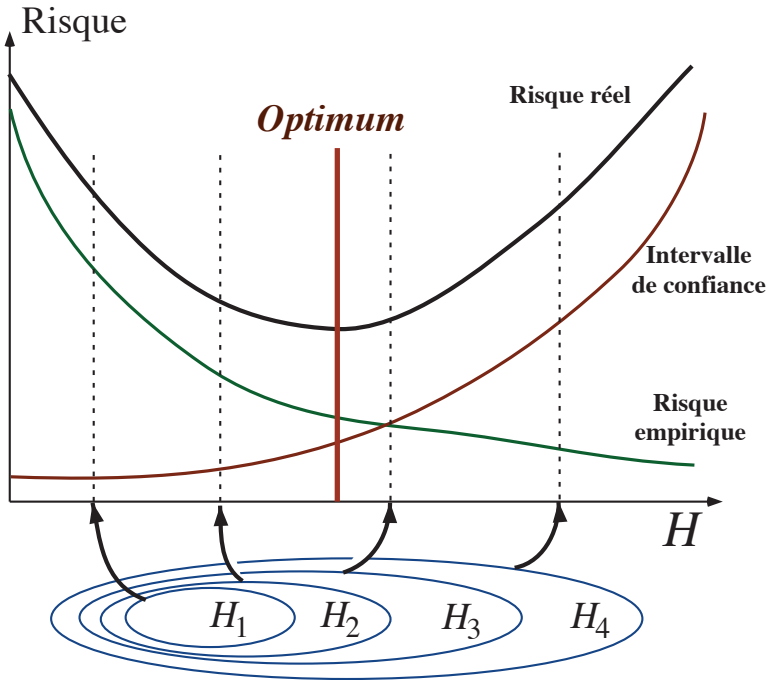
Théorie statistique de l'apprentissage

Le 3ème temps

Quelle **Foule** ?

SRM : Structural Risk Minimization

- **Stratification** des espaces d'hypothèses
  - Faite *a priori* (indépendam-ment des données)
  - Par exemple en utilisant la d_{VC}



L'analyse « PAC learning » ou statistique

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

- *Nouveau critère inductif* :
  - Le risque empirique régularisé
    1. Satisfaire les contraintes posées par les exemples
    2. Choisir le meilleur espace d'hypothèses (capacité de *H*)

L'apprentissage devient …

1. Le **choix de** l'espace des hypothèses ***H***
   - Nécessairement contraint

2. Le **choix** d'**un critère inductif**
   - Risque empirique nécessairement régularisé

3. Une **stratégie d'exploration de *H*** pour minimiser le risque empirique régularisé
   - Faire ce qu'il faut pour que l'exploration soit efficace
     - Rapide
     - Si possible un seul optimum

Nouvelle perspective

- **Poser un problème d'apprentissage**, c'est :

1. L'exprimer sous forme d'**un critère inductif** à optimiser
   - **Risque empirique**
     - avec une **fonction d'erreur** adéquate
   - Un **terme de régularisation**

$$h^\star = \underset{h \in \mathcal{H}}{\text{ArgMin}} \left[R_{\text{Emp}}(h) + \lambda \, \text{reg}(h) \right]$$

     - exprimant les contraintes
     - et **connaissances a priori**
     - si possible conduisant à problème convexe

2. Trouver un **algorithme d'optimisation** adapté

Cadre séduisant

- **Algorithme d'apprentissage**
  - Générique : ***minimisation du risque empirique régularisé***
  - Apprentissage **=** optimisation
- **Faible a priori** sur le monde
  - Suppose données (et questions) **i.i.d.**
  - $f \in H$ ou $f \notin H$
  - **Valable dans le pire cas** : contre toute distribution cible
- **Bornes** en généralisation
  - Formalisation mathématique supportant son bien-fondé

Un paradigme triomphant

Apprentissage = choix de normes + optimisation

(~ 1995 - ~20??)

Un paradigme général

- Boosting
- Arbres de décisions (random forests)
- Régression logistique
- Réseaux de neurones
- Séparateurs à Vastes Marges (SVM)
- …

« Traduction » : préférence pour les **hypothèses parcimonieuses**

- Recherche d'**hypothèse linéaire** parcimonieuse

$$h^\star = \underset{h \in \mathcal{H}}{\text{ArgMin}} \left[\frac{1}{m} \sum_{i=1}^{m} \ell(h(\mathbf{x}_i), y_i) + \lambda \, \text{reg}(h) \right]$$

$$h^\star = \underset{h \in \mathcal{H}}{\text{ArgMin}} \left[\frac{1}{m} \sum_{i=1}^{m} \ell(h(\mathbf{x}_i), y_i) + \lambda \, ||h||_1 \right]$$

Norme l_1 : $\quad ||\mathbf{w}||_1 = \sum_{j=1}^{p} |w_j|$

- Méthodes de type LASSO

« Traduction » : apprentissage multi-tâches

- *T* tâches de classification binaire définies sur $\mathcal{X}$ x $\mathcal{Y}$

$$\mathcal{S} = \{\{(\mathbf{x}_{11}, y_{11}), (\mathbf{x}_{21}, y_{21}), \ldots, (\mathbf{x}_{m1}, y_{m1})\}, \ldots, \{(\mathbf{x}_{1T}, y_{1T}), (\mathbf{x}_{2T}, y_{2T}), \ldots, (\mathbf{x}_{mT}, y_{mT})\}\}$$

$$h_j(\mathbf{x}) = \mathbf{w}_j \cdot \mathbf{x}$$ Hypothèses linéaires

Partage entre tâches $$\mathbf{w}_j = \mathbf{w}_0 + \mathbf{v}_j$$

$$h_1^\star, \ldots, h_T^\star = \underset{\mathbf{w}_0, \mathbf{v}_j, \xi_{ij}}{\text{Argmin}} \left\{ \sum_{j=1}^{T} \sum_{i=1}^{m} \xi_{ij} + \frac{\lambda_1}{T} \sum_{j=1}^{T} ||\mathbf{v}_j||^2 + \lambda_2 ||\mathbf{w}_0||^2 \right\}$$

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 :

$$\min_{w} \; c\,m\,R_S(G_{\rho_w}) + a\,m\,\mathrm{dis}_{\rho_{w'}}(S_u, T_u) + \mathrm{KL}(\rho_w \| \pi_0), \tag{7.5}$$

où $\mathrm{dis}_{\rho_{w'}}(S_u, T_u) = \left| \mathop{\mathbb{E}}_{(h,h')\sim\rho_{w^2}} R_{S_u}(h,h') - \mathop{\mathbb{E}}_{(h,h')\sim\rho_{w^2}} R_{T_u}(h,h') \right|$ est le désaccord empirique entre S_u et T_u spécialisé à une distribution ρ_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 :

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

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

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

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 :

L'adaptation de domaine

- Une mise en théorie pionnière (prisonnière ?) [Ben-David et al., 2010]

Théorème classique [Ben-David et al., 2010, Mansour et al., 2009a]

Soit $\mathcal{H}$ un espace d'hypothèses. Si D_S et D_T sont deux distributions sur X, alors :

$$\forall h \in \mathcal{H}, \quad \overbrace{R_{P_T}(h)}^{\text{erreur cible}} \leq \underbrace{R_{P_S}(h)}_{\text{erreur source}} + \underbrace{\tfrac{1}{2} d_{\mathcal{H}}(D_S, D_T) + \nu}_{\text{divergences}}$$

$R_{P_S}(h)$: erreur classique sur le domaine source

Minimisable via une méthode de classification supervisée sans adaptation

$\frac{1}{2} d_{\mathcal{H}}(D_S, D_T)$: la $\mathcal{H}$-divergence entre D_S et D_T

$$\begin{aligned} \tfrac{1}{2} d_{\mathcal{H}}(D_S, D_T) &= \sup_{(h,h') \in \mathcal{H}^2} \left| R_{D_T}(h, h') - R_{D_S}(h, h') \right| \\ &= \sup_{(h,h') \in \mathcal{H}^2} \left| \mathop{\mathbf{E}}_{\mathbf{x}^t \sim D_T} \mathbf{I}\left[h(\mathbf{x}^t) \neq h'(\mathbf{x}^t)\right] - \mathop{\mathbf{E}}_{\mathbf{x}^s \sim D_S} \mathbf{I}\left[h(\mathbf{x}^s) \neq h'(\mathbf{x}^s)\right] \right| \end{aligned}$$

ν : divergence entre les étiquetages

$\nu = \inf_{h' \in \mathcal{H}} \left(R_{P_S}(h') + R_{P_T}(h') \right)$,
erreur jointe optimale [Ben-David et al., 2010]

ou $\nu = R_{P_T}(h_T^*) + R_{P_T}(h_T^*, h_S^*)$,
$h_{\mathcal{X}}^*$ est la meilleure hypothèse sur le domaine $\mathcal{X}$ [Mansour et al., 2009a]

Idée : **construire** un nouvel espace de projection dans laquelle les deux distributions sont proches, tout en gardant une bonne performance sur le domaine source

Quelles garanties exactement ?

Apprentissage statistique : quelles garanties ?

- Lien entre risque **empirique** et risque **réel**
  - Coût d'usage de *h* (e.g. taux d'erreur)

- **Seulement si**
  - Monde stationnaire
  - **Données** i.i.d.
  - **Questions** i.i.d. !!?

PAS :

- **Intelligibilité**
- **Fécondité**
- **Utilisation dans une théorie du domaine**

Limites

- Apprentissage **passif** et **données et questions i.i.d.**
  - Agents situés : le monde n'est pas i.i.d.
- Requiert **beaucoup** d'exemples
  - Nous sommes beaucoup plus efficaces
  - « Producteurs de théories », théories que nous testons ensuite
- Pas adapté à la recherche de **causalités**
- Pas **intégré** avec un **raisonnement**

Ces machines **apprenantes** ne sont pas des machines **pensantes**

Extensions … qui **ne sortent pas du paradigme**

Monde non stationnaire

- P_X -> $P_{X'}$: co-variate shift => Importance sampling
- Transfert / adaptation de domaine
  - La théorie suppose toujours données et questions i.i.d. dans chaque domaine
  - Pas d'« histoire »
- Apprentissage **en-ligne**
  - Contre toute séquence
  - Tellement trop peu exigeant
  - Que trop faible

Trame

1. L'**induction**: omniprésence et faillibilité
2. Le **no-free-lunch theorem**
3. **Approches** de l'induction (hier et aujourd'hui)
   - Le **Perceptron**
   - La **théorie statistique** de l'apprentissage
   - Le **paradigme** dominant
   - Un point de vue indépassable ? Le **cas de l'EBL**
4. Quelle perspective pour **l'avenir** ?
5. Conclusion

Un point de vue **indépassable** ?

Que faisait-on **avant** ?

Le cas de l'EBL

Un peu d'histoire

IA et **résolution automatique de problèmes**

- Arch [Winston, 1972]
  - Stratégie de **recherche guidée** dans un espace de descriptions structurées
- [Simon & Lea (1979) « *Problem-solving and rule induction: a unified view* »]
  - Se focalisent sur les **mécanismes de raisonnement** (generate_and_test, heuristic search, hypothesis_and_match)
  - Au lieu de chercher à résoudre un problème, on cherche à « couvrir » des exemples, mais **mêmes types de procédures**
  - GPS -> GRI (Generalized Rule Induction)
- [Tom Michell (1980, 1982) « *Generalization as Search* », « *The need for biases in learning generalizations* »]
  - **Comment organiser la recherche** d'une (bonne) hypothèse
  - Si pas de biais, l'apprentissage ne peut pas faire mieux que l'apprentissage par cœur
- [David Haussler (1988) « *Quantifying inductive bias: AI learning algorithms and Valiant's learning* »]
  - Quantification du biais (par la dimension de Vapnik-Cervonenkis) **de classes d'expressions logiques**

L'apprentissage …

… comme

l'amélioration de l'efficacité d'un résolveur de problème

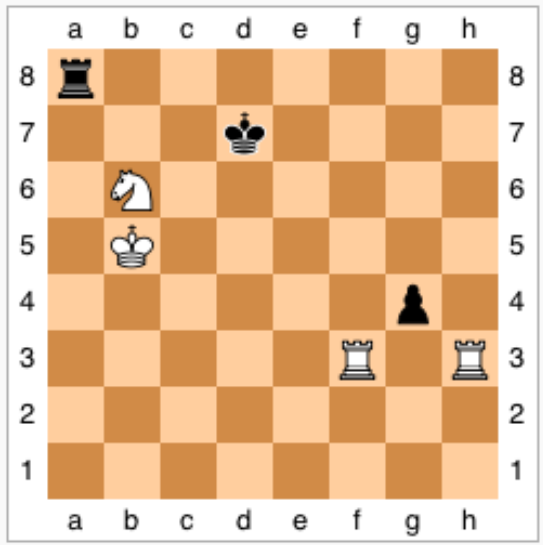
Apprendre **à partir d'un exemple**

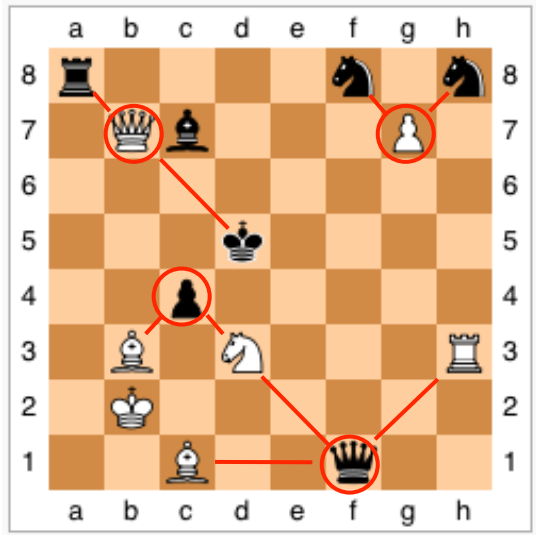
Explanation-Based Learning

1. Un **exemple unique**

2. **Recherche de la preuve** de la « fourchette »

3. **Généralisation**





Explanation-Based Learning

Ex : **apprendre le concept** `empilable(Objet1, Objet2)`

- **Théorie** :

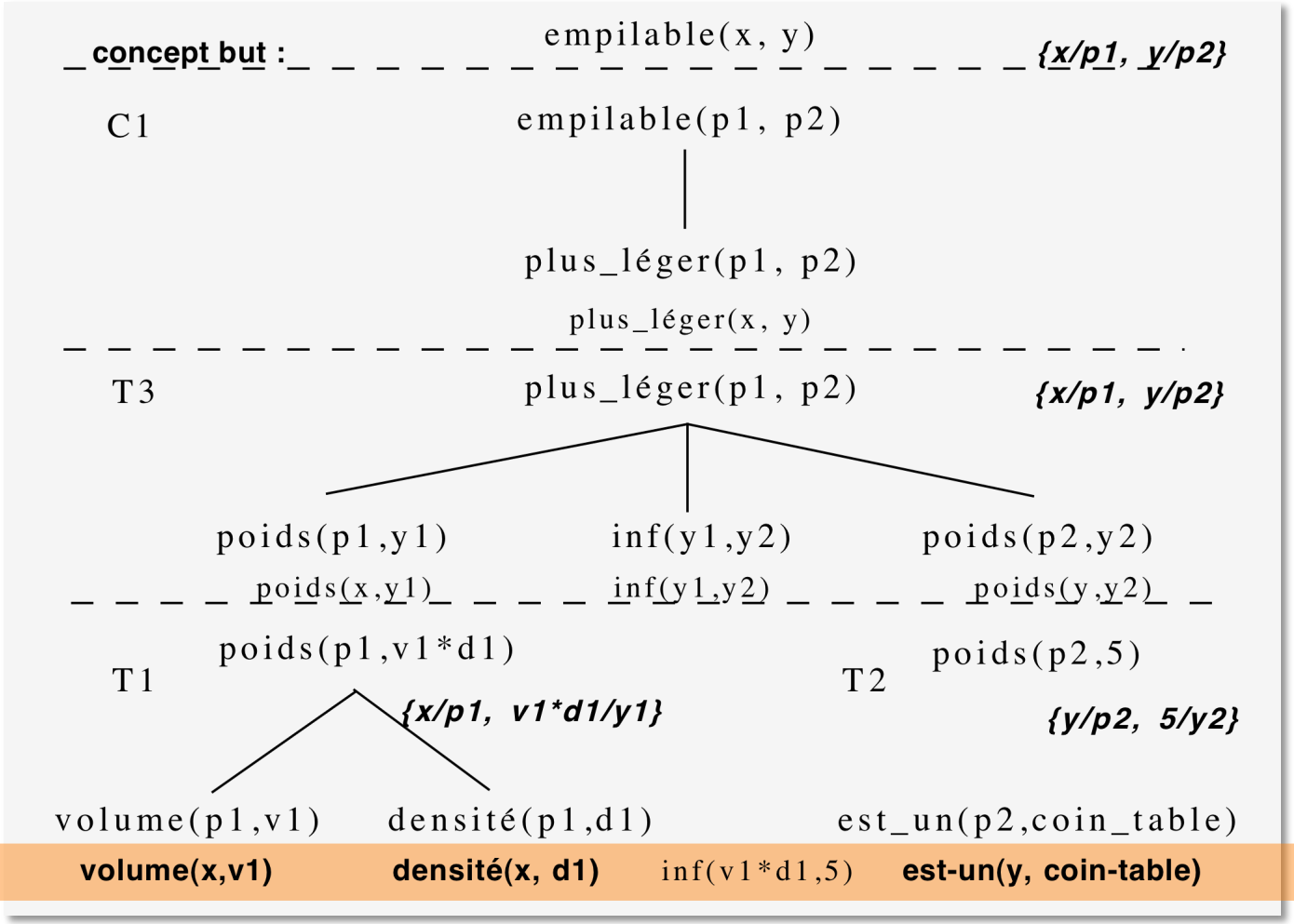
```
(T1) : poids(X, W) :- volume(X, V), densité(X, D), W is V*D.

(T2) : poids(X, 50) :- est-un(X, table).

(T3) : plus-léger(X, Y) :- poids(X, W1), poids(X, W2), W1 < W2.
```

- **Contrainte d'opérationalité** :
  - Concept à exprimer à l'aide des prédicats *volume, densité, couleur, …*
- **Exemple positif** (solution) :

```
sur(obj1, obj2).
est_un(objet1, boîte).
est_un(objet2, table).
couleur(objet1, rouge).
couleur(objet2, bleu).
matériau(objet2, bois).
```

```
volume(objet1, 1).
volume(objet2, 0.1).
propriétaire(objet1, frederic).
densité(objet1, 0.3).
matériau(objet1, carton).
propriétaire(objet2, marc).
```

Explanation-Based Learning



Arbre de preuve généralisé obtenu par régression du concept cible dans l'arbre de preuve en calculant à chaque étape les littéraux les plus généraux permettant cette étape.

Explanation-Based Learning

- Induction **à partir d'un seul exemple**
  - … et d'une **théorie forte du domaine**
- Langage de la logique
- **Opérateurs** de raisonnement (déduction, …)
- *Maintenant utilisées dans les « solveurs » de problèmes SAT.*

Explanation-Based Learning

- Que cherche-t-on à **prouver** ?
- Qu'est-ce qui est une **bonne** (moins bonne) **théorie / méthode** ?

Explanation-Based Learning

- Que cherche-t-on à **prouver** ?
- Qu'est-ce qui est une **bonne** (moins bonne) **théorie / méthode** ?

1. Méthode **améliorant** les performances de **résolution de problème**
   - [Steve Minton (1990) « *Quantitative results concerning the **utility** of Explanation-Based Learning* »]
2. Méthode « **reproduisant** » les performances (et limites) d'un **agent cognitif naturel** (animal ou humain)
   - [Laird, Rosenbloom, Newell (1986) « *Chunking in SOAR: The anatomy of a general learning mechanism* »]
   - [Anderson (1993) « *Rules of the mind* » ; Taatgen (2003) « *Learning rules and productions* »]

Explanation-Based Learning

1. On ne s'interroge pas directement sur **la validité des hypothèses** induites (i.e. espérance de coût)

2. « Utility » ~ **espérance d'utilité**
   en termes de situations de **résolution de problèmes**

Explanation-Based Learning

- Questions traitées dans les publications
  - Quel type d'induction en fonction de la **notion de conséquence logique** utilisée ?
  - Comment **utiliser la théorie** du domaine ?
  - Que faire si la théorie du domaine est **incomplète** ou **erronée** ?
  - Comment utiliser **des contre-exemples** ?
  - Quel est le rôle du **critère d'opérationnalité** ?
  - Que faire si on obtient **plusieurs arbres de preuves** ?

Explanation-Based Learning

- Est-ce de l'induction ?

Déduction guidée par des critères d'opérationnalité

Trame

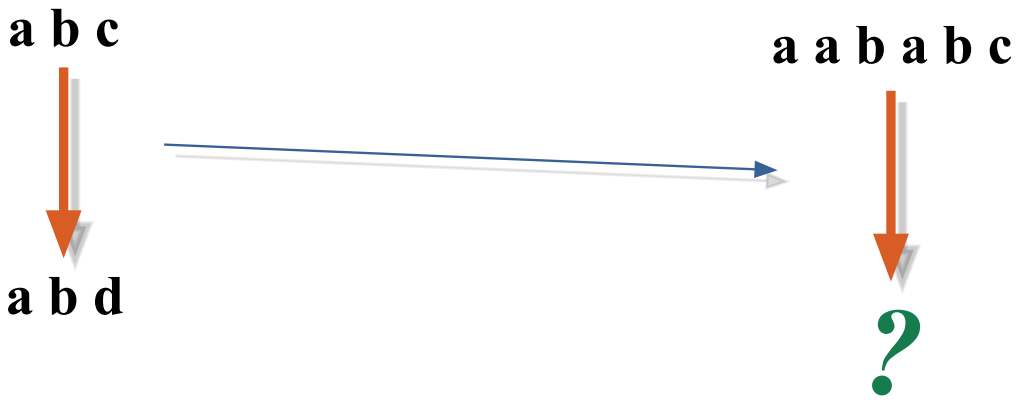
1. L'**induction**: omniprésence et faillibilité
2. Le **no-free-lunch theorem**
3. **Approches** de l'induction (hier et aujourd'hui)
   - Le **Perceptron**
   - La **théorie statistique** de l'apprentissage
   - Le **paradigme** dominant
   - Un point de vue indépassable ? Le **cas de l'EBL**
4. Quelles perspectives pour **l'avenir** ?
5. Conclusion

Quelles perspectives

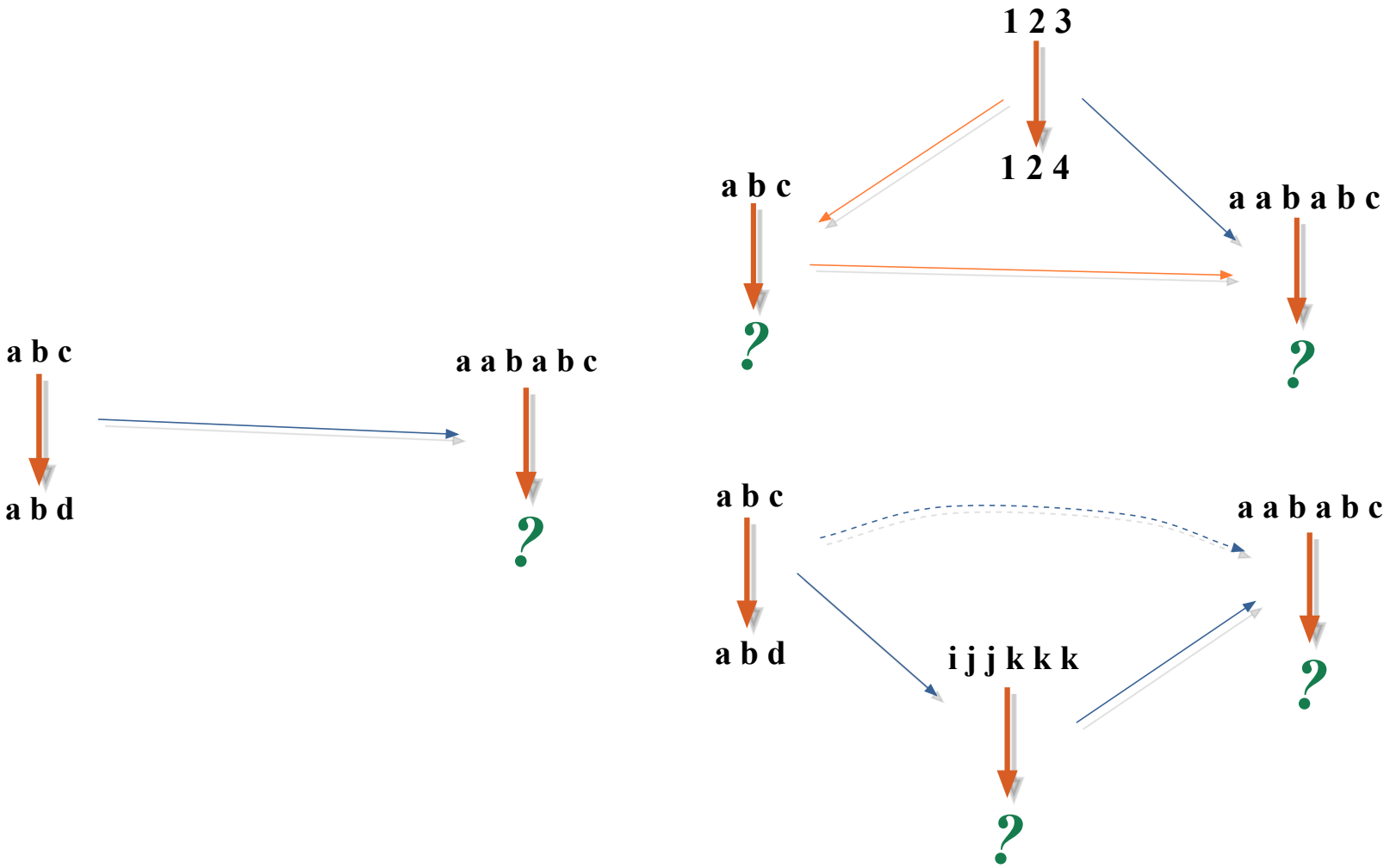
Transfert, analogie, éducation :

quel principe inductif ?

Transfert et analogie



Transfer and sequence effects

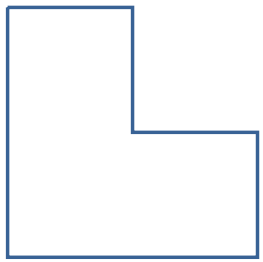


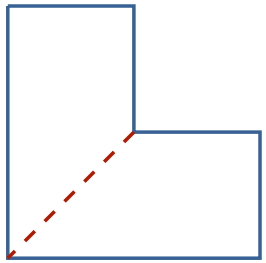
- t

Effets de séquences

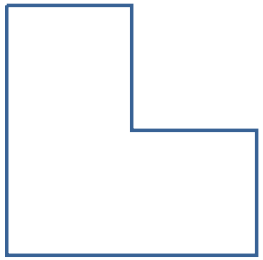
- *Consigne* : découper la figure suivante en *n* parties superposables

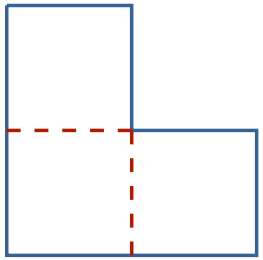
En 2 :



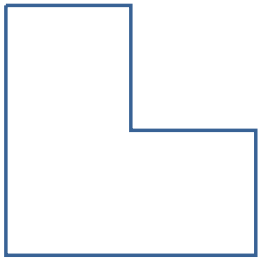


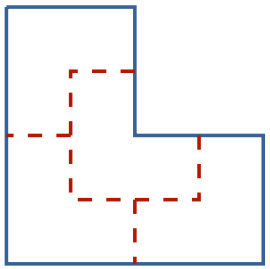
En 3 :



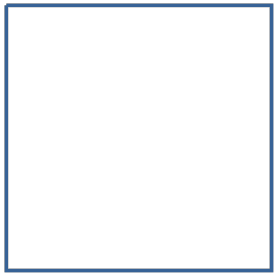


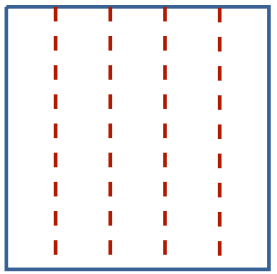
En 4 :



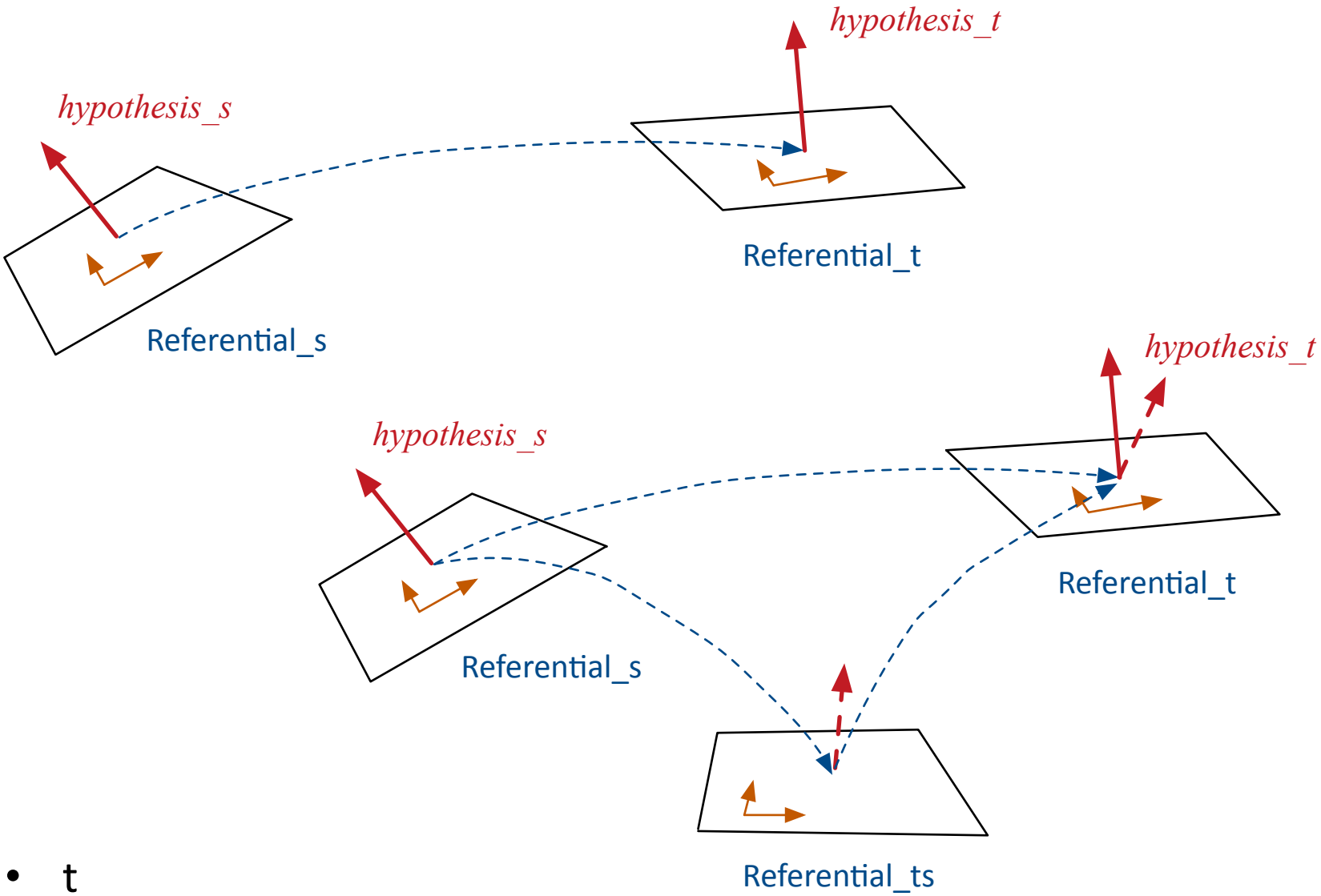


En 5 :



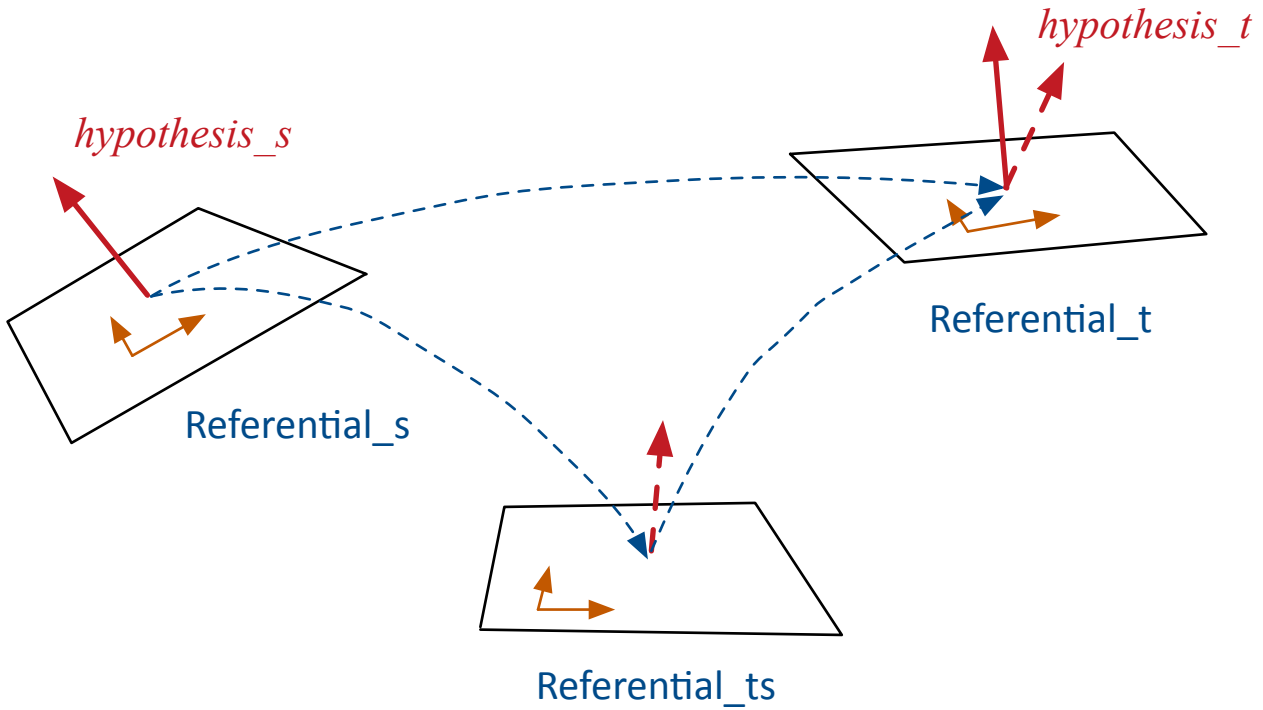


Transfer and sequence effects



- t

Transfer and sequence effects



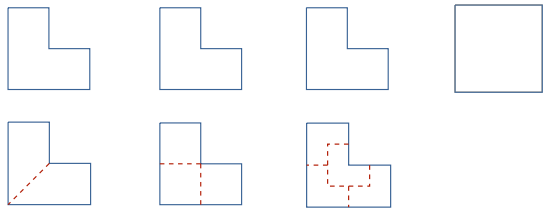
1. **Quelles équations** pour les changements de référentiels et le transfert d'hypothèses ?

2. Comment montrer que ces **équations** sont **optimales** ?

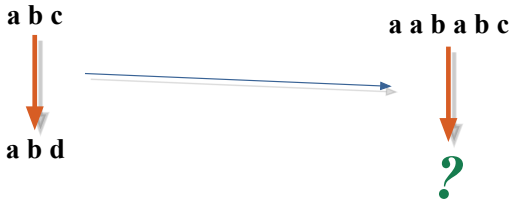
Nouveau scénario …

… Nouveaux **principes inductifs**

1. Principe de **pertinence maximale**
   - des **situations rencontrées** juste avant
   - des **connaissances mobilisées** juste avant

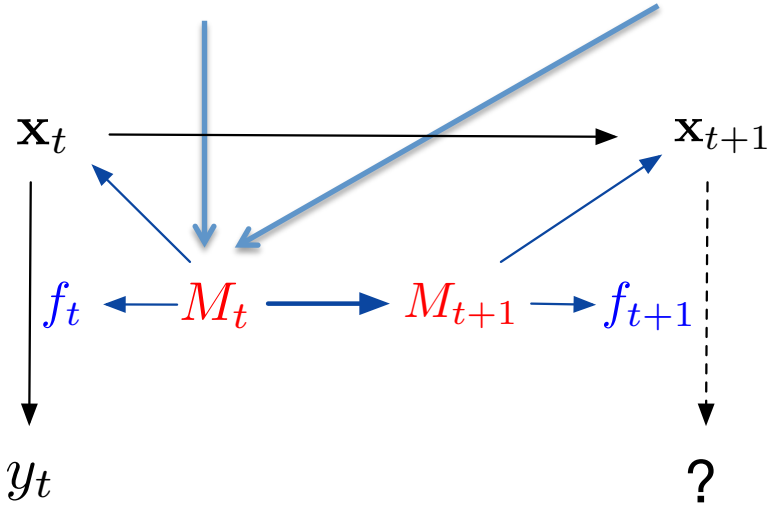


2. Principe de **non indifférence** à la « question à venir »



Une formalisation

- Complexité de Kolmogorov
  - Repose sur un **codage**
  - Qui **dépend de la connaissance a priori** et de l'**utilisation passée**



$$K(M_t) + K(\mathbf{x}_t|M_t) + K(y_t|M_t) + K(M_{t+1}|M_t) + K(\mathbf{x}_{t+1}|M_{t+1}) + K(f_{t+1}|M_{t+1})$$

[A. Cornuéjols (1996) « Analogie, principe d'économie et complexité algorithmique »]

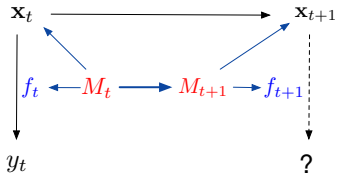
Questions

1. Comment **construire une théorie** de l'analogie ?

- Le choix du **critère inductif**
- Sa **déclinaison** en problème d'optimisation
- Algorithme

Cadre de l'**analogie proportionnelle**

Principe de **pertinence maximale**
+ Principe de **non indifférence à la question**

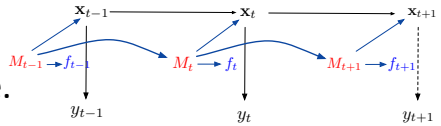


2. Comment **valider ?**

Pas de validation absolue

- *Pourquoi une analogie serait jugée meilleure qu'une autre ?*

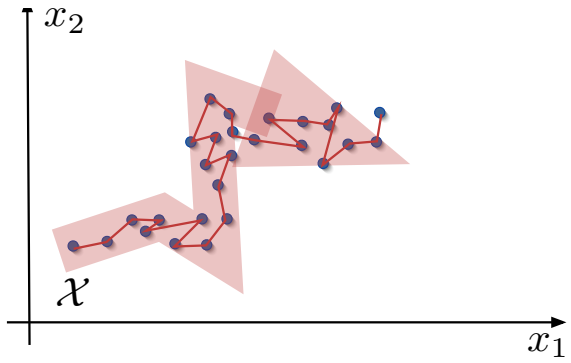
- **Compatible à la limite** i.i.d. avec l'apprentissage statistique
- Permet d'**obtenir de « meilleurs résultats »** quand dérive de concept (i.e. conforme à nos attentes) ?
- Produit des **conséquences inattendues** (e.g. sur l'éducation) ?

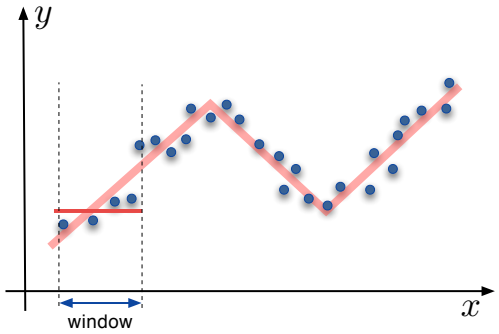


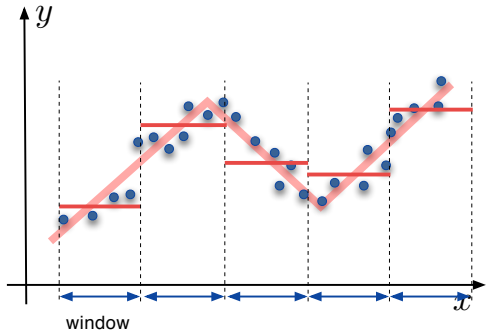
[Miclet & Prade (2009) « *Handling analogical proportions in classical logic and fuzzy logic settings* ».]

[Murena & Cornuéjols (2016) « *Minimum Description Length Principle applied to structure adaptation for classification under concept drift* ». IJCNN-2016.]

Le tracking [Sutton et al., 2007]







In **tracking**, the learning agent receives inputs that are driven by a time dependent process. It therefore encounters different parts of the environment at different times.

Even though the world involves a piecewise linear law, the learning agent may perform well by maintaining a very simple model, a constant, over its local environment.

Sutton, Koop & Silver (2007) « *On the role of tracking in stationary environments* ». ICML-07.

Le tracking [Sutton et al., 2007]

Temporal Consistency	⟷	*i.i.d. data*
Small memory Simple $\mathcal{H}$		Large memory "Complex" $\mathcal{H}$

t

Conclusions

Conclusions

1. L'induction est au **centre** de l'apprentissage et est un problème **sous-contraint**.
2. Il ne peut y avoir de validation absolue de l'induction
3. On ne peut garantir une performance en induction qu'en faisant des **méta-présupposés** sur le monde
   - E.g. données i.i.d.

- Une **théorie de l'induction** vise à
  - Proposer **des méta-présupposés** « raisonnables »
  - Offrir un **cadre formel** dans lequel obtenir des « **théorèmes du lampadaire** »
    - ***Si*** *les pré-supposés sont vérifiés par les données* ***alors*** *je peux garantir que …*

Un guide :
La **compatibilité**
des théories
à leurs **interfaces**

Conclusions : la théorie statistique de l'apprentissage

- **Performance** visée : l'espérance de coût d'usage
  (i.e. pas causalité, pas intelligibilité, pas articulation à raisonnement, …)
- Valable si **monde stationnaire** + **données** i.i.d. + **questions** i.i.d.

- Même le "big data" va présenter des **défis sortant du cadre**
  - Même si on stocke tout : il faut **indexer la mémoire** => choix (pb de l'utilité)
  - Objectif : aider à la **décision** => il faut articuler au raisonnement
  - **Systèmes** d'apprentissage **collaborant** entre eux
    => gros problèmes de spécification des entrées et sorties
    [Léon Bottou, ICML-2015]

Conclusions : de « nouveaux » scénarios

- Assez **peu de données**
  - on apprend (très souvent) avec très peu
- L'**histoire passée** compte : éducation
  - Effets de séquence
- Apprendre pour **construire des théories**
  - Nous construisons constamment des théories micro et macro

Votre question à vous ?

Conclusion (fin)

Comment faire ?

Conclusion (fin)

Comment faire ?

La construction de nouveaux paradigmes est difficile

Surtout quand **le paradigme dominant** Apprentissage statistique

- Semble très **bien fonctionner**
- Semble être parfaitement **adapté aux besoins** (e.g. « big data »)
- Fait appel à des **mathématiques sophistiquées**
  (valorisées, intimidantes et forcément objectives)

Intuition des bons problèmes + **intrépidité** + **rigueur**