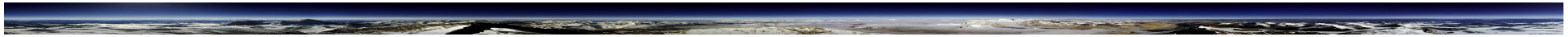
Qu'est-ce qu'un apprentissage réussi ?

Ce qui est évident aujourd'hui est différent d'hier. Et demain ?



Antoine Cornuéjols

AgroParisTech – INRA MIA 518

AgroParisTech

INRA SCIENCE & IMPACT

IA : deux directions

- Est-ce qu'une machine peut penser ?
  - Comme un humain adulte -> **difficile**

- Est-ce qu'une machine peut apprendre ?
  - Construisons un « enfant » (ou un organisme élémentaire)
  - **Faisons le apprendre**
  - Nous aurons un adulte pensant

Turing (1950) « *Computing Machinery and Intelligence* »,
Mind, vol. LIX, N° 236, pp. 433-460.

Question

Que signifie

Une bonne théorie de l'apprentissage ?

Ce qui guide la méthodologie et la pratique

Question iconoclaste

Que signifie

Un bon travail ?

Quelque chose de publiable

- **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. 1960s : Apprendre c'est **s'adapter**
2. 1970-1985 : apprendre **et raisonner**
3. 1985 - … : La **théorie statistique** de l'apprentissage
4. Et **maintenant** ?
5. Et **demain** ?

Plan

1. 1960s : Apprendre c'est **s'adapter**
2. 1970-1985 : apprendre et **raisonner**
3. 1985 - … : La **théorie statistique** de l'apprentissage
4. Et **maintenant** ?
5. Et **demain** ?

Apprendre c’est s’adapter

Connexionnisme : La règle de Widrow-Hoff

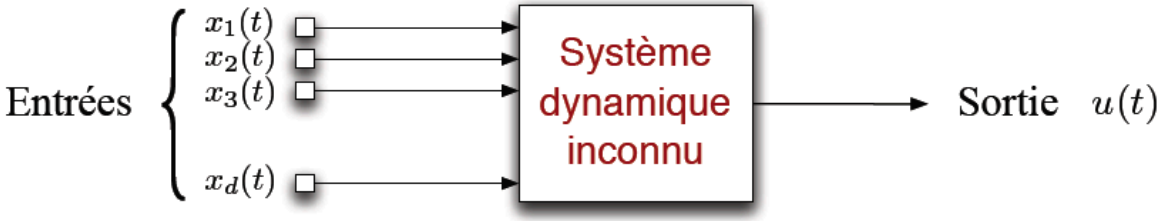
Conçue dans le cadre du **filtrage adaptatif**.

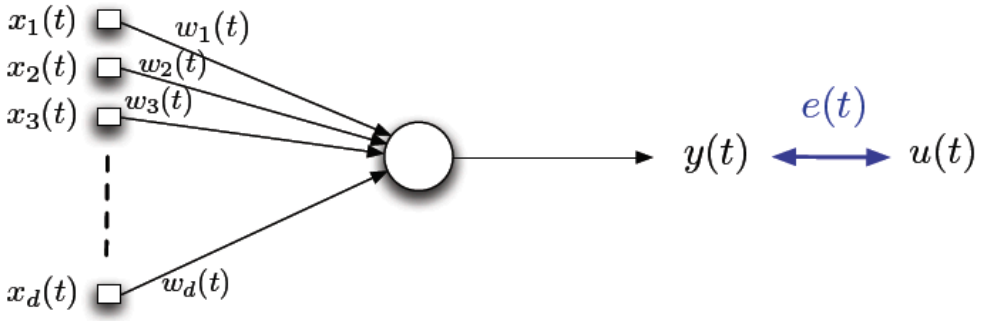
Chercher un **modèle linéaire d'un signal temporel** : $y(t) = \sum_{k=1}^{M} w_k(t)x_k(t)$

Connexionnisme : La règle de Widrow-Hoff

Conçue dans le cadre du **filtrage adaptatif**.

Chercher un **modèle linéaire d'un signal temporel** : $y(t) = \sum_{k=1}^{M} w_k(t) x_k(t)$





La règle de Widrow-Hoff

B. Widrow and M. Hoff. *Adaptive Switching Circuits.* IER WESCON Conv. Rec. Pt.4, pp;96-104, 1960

The problem of adjusting the h's is not trivial, because their effects upon performance interact. Suppose that the predictor has only two impulses in its impulse response, h_1 and h_2. The mean square error for any setting of h_1 and h_2 can be readily derived:

$$\epsilon(m) = f(m) - h_1 f(m-1) - h_2 f(m-2)$$

$$\overline{\epsilon^2}(m) = \phi_{ff}(0)h_1^2 + \phi_{ff}(0)h_1^2 - 2\phi_{ff}(1)h_1 - 2\phi_{ff}(2)h_2 + 2\phi_{ff}(1)h_1h_2 + \phi_{ff}(0) \quad (1)$$

The discrete autocorrelation function of the input is $\phi_{ff}(j)$.

The mean square error given by equations (1) is what the mean square meter would read if it were to average over very large sampled size. The mean square error is a parabolic function of the predictor adjustments h_1 and h_2, and, in general, can easily be shown to be a quadratic function of such adjustments, regardless of how many there are.

The optimum n-impulse predictor can be derived analytically by setting the partial derivatives of $\overline{\epsilon^2}$ of equation (1) equal to zero. This is the discrete analogue of Wiener's optimization[7] of continuous filters. Finding the optimum system experimentally is the same as finding a minimum of a paraboloid in n dimensions. This could be done manually by having a human operator read the meter and set the adjustment, or it could be done automatically

Dérivation par optimisation : règle de Widrow-Hoff

$$\boxed{\ell(\mathbf{w}) = \frac{1}{2} e^2(t)}$$

Méthode de gradient :

$$\frac{\partial \ell(\mathbf{w})}{\partial \mathbf{w}} = e(t) \frac{\partial e(t)}{\partial \mathbf{w}}$$

$$e(t) = u(t) - \mathbf{x}^\top(t)\,\mathbf{w}(t) \quad \text{d'où :} \quad \frac{\partial e(t)}{\partial \mathbf{w}(t)} = -\mathbf{x}(t)$$

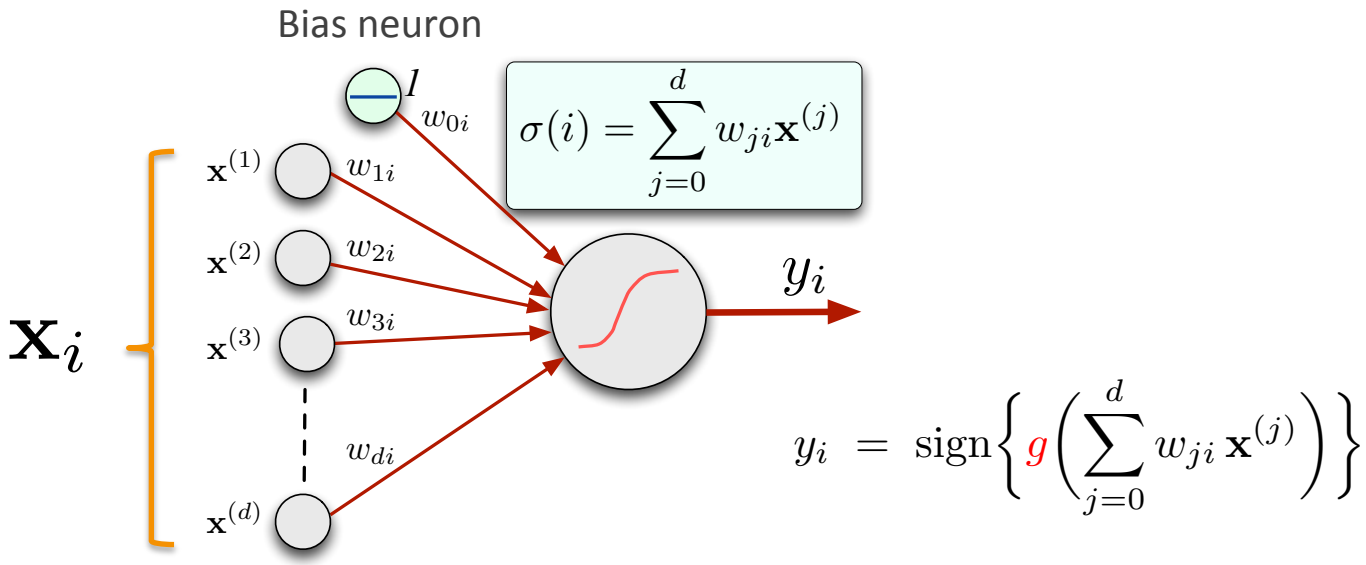
$$\frac{\partial \ell(\mathbf{w})}{\partial \mathbf{w}(t)} = -\mathbf{x}(t)\, e(t)$$

$$\boxed{\mathbf{w}(t+1) \;=\; \mathbf{w}(t) + \eta\, \mathbf{x}(t)\, e(t)}$$

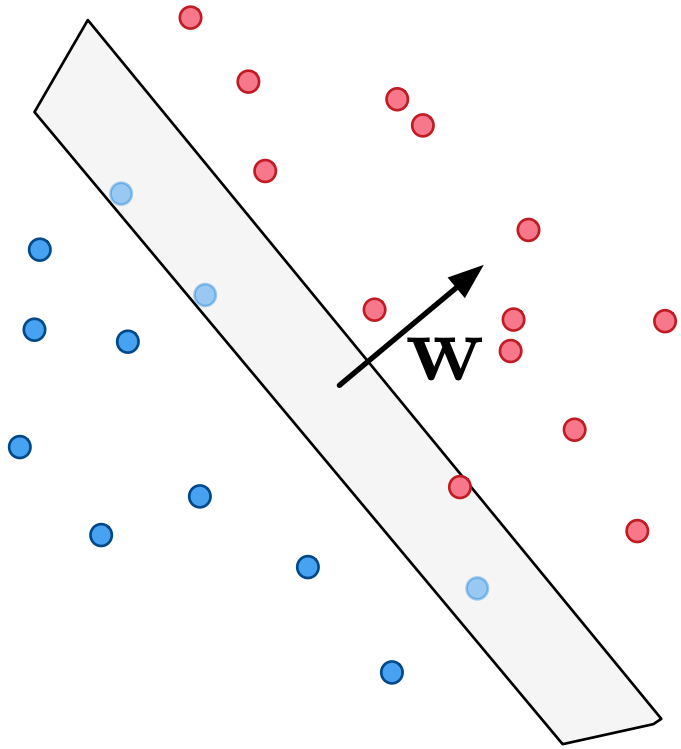
[Widrow-Hoff:60] B. Widrow and M. Hoff. *Adaptive Switching Circuits.* IRE WESCON Conv. Rec. Pt.4, pp.96-104.

The perceptron

- Rosenblatt (1958-1962)



The perceptron: a linear discriminant



Connexionnisme : 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

Qu'est-ce qu'**un bon apprentissage** ?

1. Réduit l'erreur de classification

… certes

2. Propriétés de l'algorithme

convergence en un nombre fini d'étapes

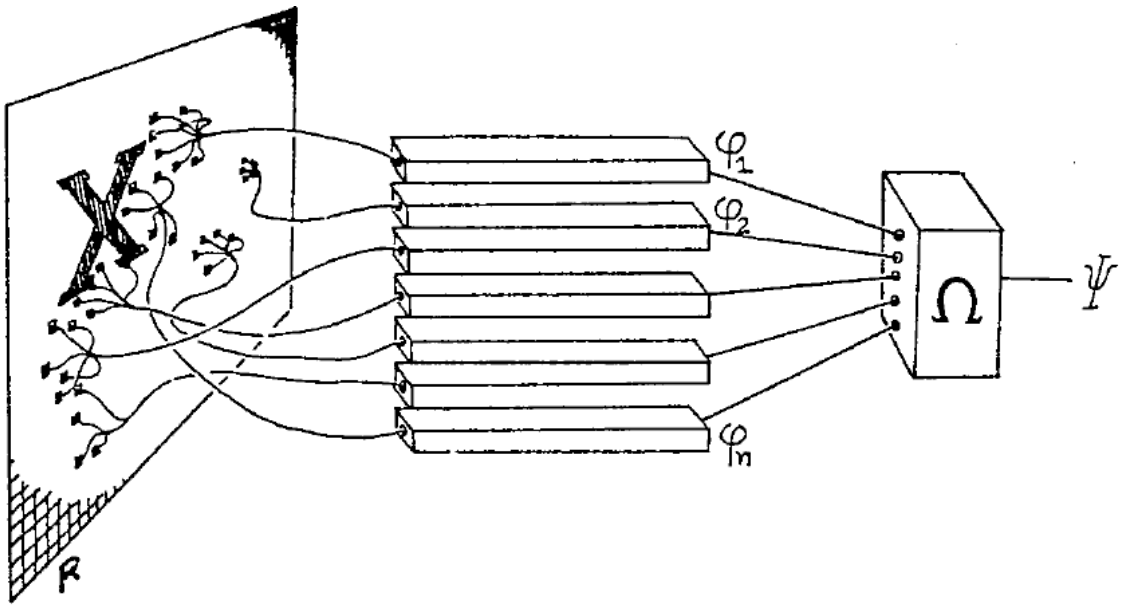
- Indépendamment du **nombre** d'exemples
- Indépendamment de la **distribution** des exemples
- Indépendamment de la **dimension** de l'espace d'entrée

!!!

S'il existe un **séparateur linéaire** des exemples d'apprentissage

Connexionnisme : le perceptron

- Frank Rosenblatt (1958 – 1962)

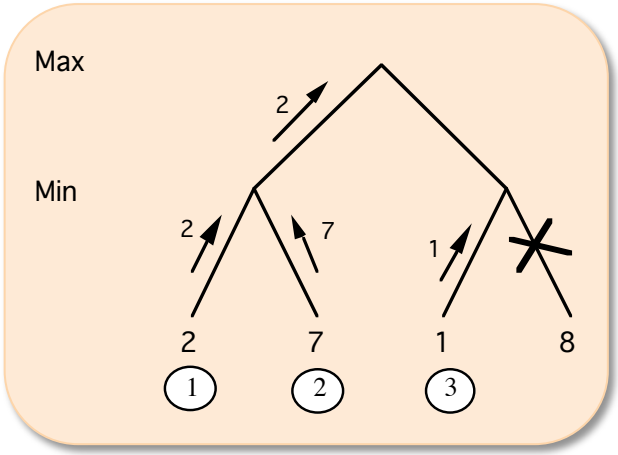


$$\Psi(\mathbf{x}) = \sum_{i=1}^{n} w_i \; \phi_i(\mathbf{x})$$

L'exemple de CHECKER

- **Combinaison de descripteurs** et **attribution de mérite**
  - Arthur Samuel. IBM, 1952 (IBM-701), 1954 (IBM-704), avec apprentissage : 1956 …
  - Modélisation MinMax du jeu
  - Apprentissage de la **fonction d'évaluation**

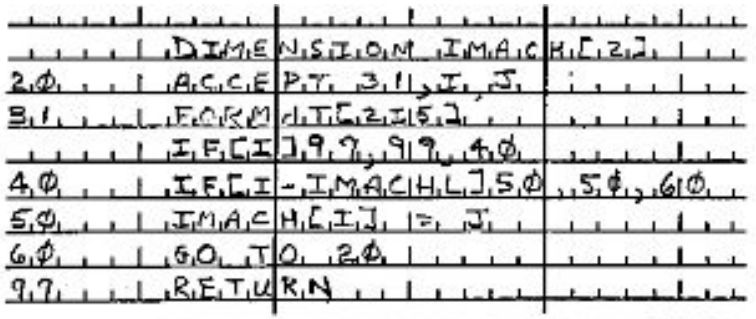
$$\text{valeur}(position) = \sum_{i=1}^{n} w_i \ \phi_i$$



- Deux problèmes
  1. Sélectionner de bonnes fonctions de base : ϕ_i
  2. Pondérer l'importance de ces fonctions : w_i

Des problèmes jouets ?

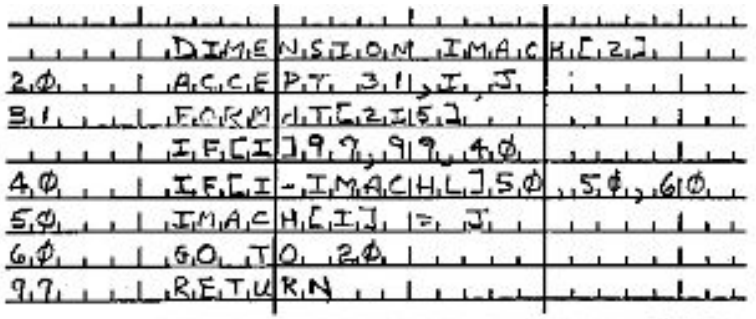
- Le perceptron
  - Taux de reconnaissance de 98%



```
      DIMENSION IMACM[2]
20    ACCEPT 31,I,J
31    FORMAT[215]
      IF[I]79,99,40
40    IF[I-IMACHL]50,50,60
50    IMACH[I]=J
60    GO TO 20
99    RETURN
```

Des problèmes jouets ?

- Le perceptron
  - Taux de reconnaissance de 98%
- CHECKER
  - Vice-champion du monde au jeu de dames anglo-saxon



```
      DIMENSION IMACM[2]
20    ACCEPT 31,I,J
31    FORMAT[215]
      IF[I]79,99,40
40    IF[I-IMACHL]50,50,60
50    IMACH[I]=J
60    GO TO 20
99    RETURN
```

Bilan (1) : aspects théoriques

1. Un bon système d'apprentissage minimise le nombre d'erreurs sur l'ensemble d'apprentissage

2. Garantie recherchée : il converge

Grosse question : comment trouver les bons descripteurs ?

Bilan (1) : publiable ?

1. Si ça marche sur un problème intéressant

2. Caractérisation de l'algorithme

Plan

1. 1960s : Apprendre c'est **s'adapter**
2. 1970-1985 : Apprendre et **raisonner**
3. 1985 - … : La **théorie statistique** de l'apprentissage
4. Et **maintenant** ?
5. Et **demain** ?

IA = méthodes générales de **raisonnement**

1956 - ~ 1970

- Raisonnement = manipulation de **représentations** discrètes des connaissances

  - Démonstrateurs de **théorèmes**
  - Résolveurs universels de **problèmes** (GPS)
  - CHECKER : vice-champion du monde au **jeu** de dames (1962)

IA = méthodes générales de **raisonnement**

1956 - ~ 1970

- Raisonnement = manipulation de **représentations** discrètes des connaissances
  - Démonstrateurs de théorèmes
  - Résolveurs universels de problèmes (GPS)
  - CHECKER : vice-champion du monde au jeu de dames

- Mais l'intelligence **n'est pas que ça**

- Les experts possèdent un répertoire énorme de connaissances

Apprendre c’est savoir représenter

et raisonner

Learning …

… as

a means to improve the efficiency of a problem solver

E.g. The PRODIGY system

ACM SIGART Bulletin, 1991, vol. 2, no 4, p. 51-55

PRODIGY: An Integrated Architecture for Planning and Learning

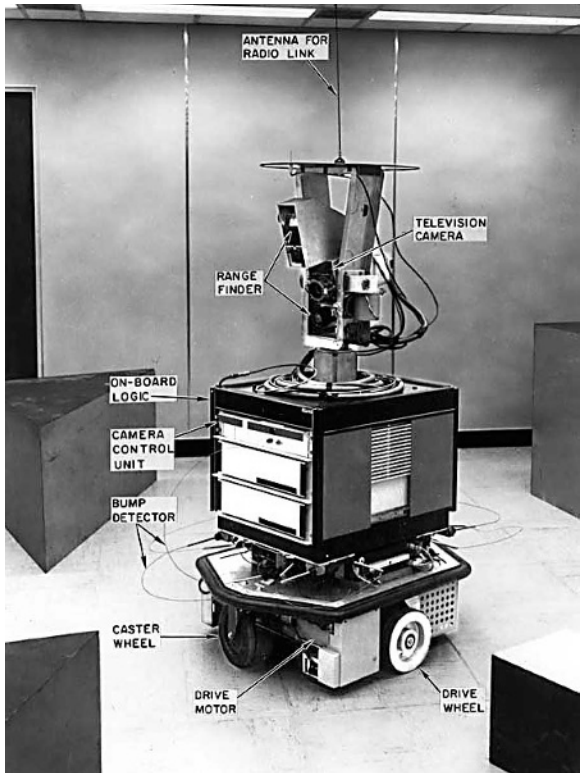
Jaime Carbonell, Oren Etzioni*, Yolanda Gil, Robert Joseph
Craig Knoblock, Steve Minton†, and Manuela Veloso

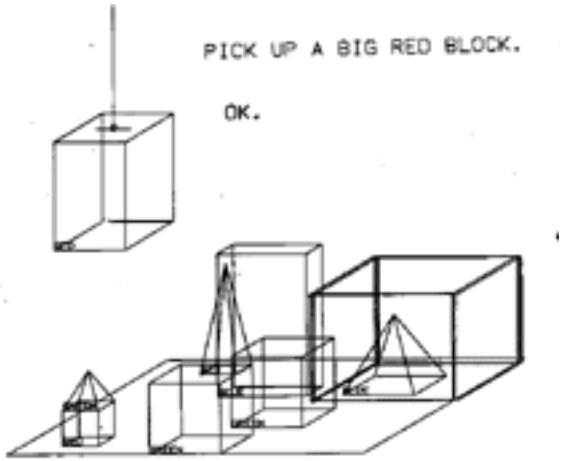
PRODIGY's basic reasoning engine is a general-purpose problem solver and planner [10] that searches for sequences of operators (i.e., plans) to accomplish a set of goals from a specified initial state description. Search in PRODIGY is guided by a set of *control rules* that apply at each decision point.

PRODIGY's reliance on explicit control rules, which can be learned for specific domains, distinguishes it from most domain independent problem solvers. Instead of using a least-commitment search strategy, for example, PRODIGY expects that any important decisions will be guided by the presence of appropriate control knowledge. If no control rules are relevant to a decision, then PRODIGY makes a quick, arbitrary choice. If in fact the wrong choice is made, and costly backtracking proves necessary, an attempt will be made to learn the control knowledge that must be missing.

SHAKEY [SRI, ~1968 - 1975]

- Le robot SHAKEY

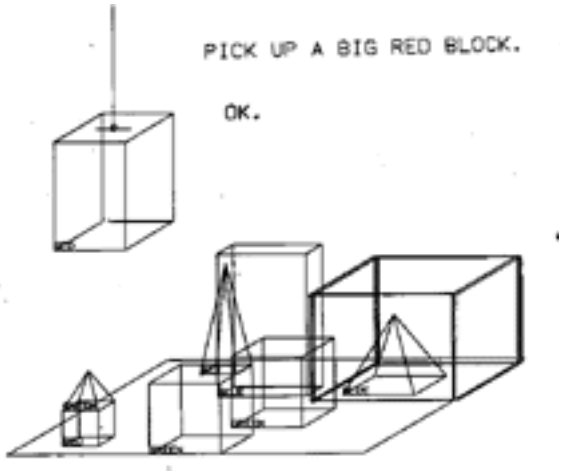


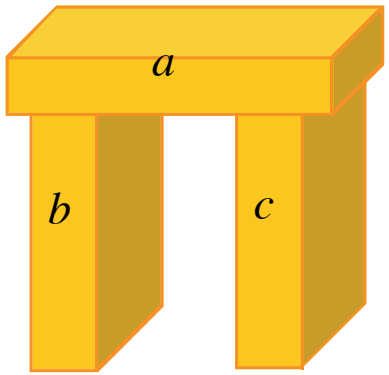


SHRLDU

ARCH [Winston, 1970]

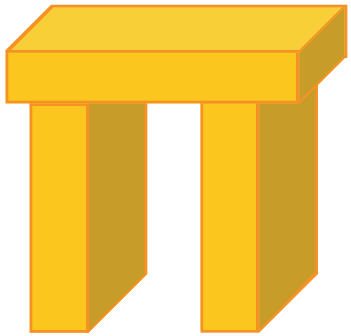
- **Apprentissage** de **concept** (e.g. arche) dans un monde de blocs

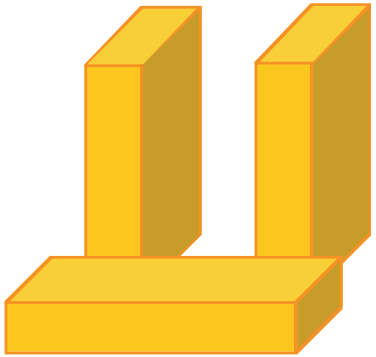




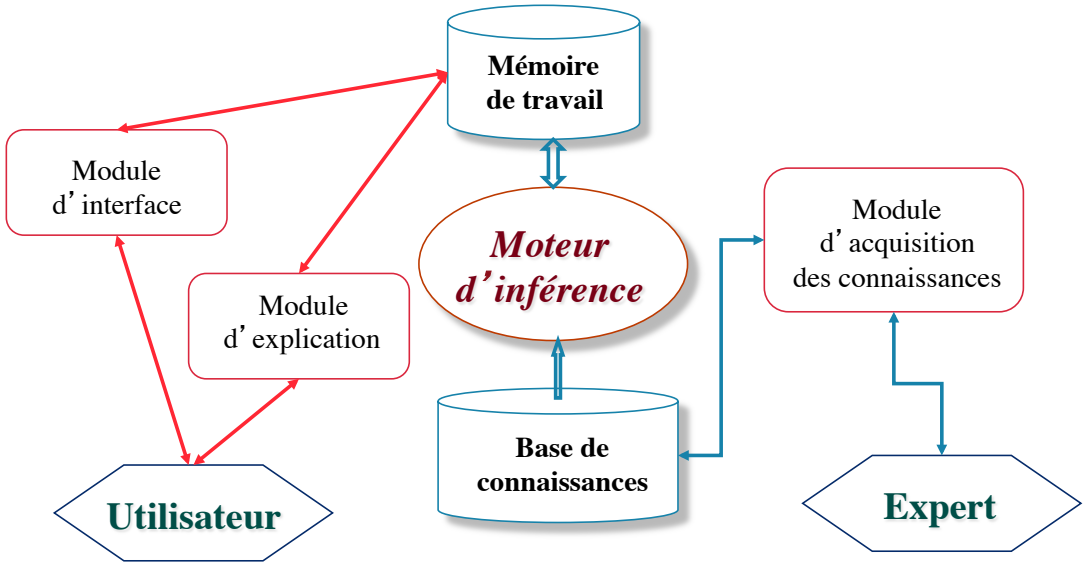
ARCH [Winston, 1970]

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





Les **systèmes experts**



- Règle :

  Si le spectre de la molécule présente deux pics x1 et x2 tels que :

  1. x1 - x2 = M + 28
  2. x1 - 28 est un pic élevé
  3. x2 - 28 est un pic élevé
  4. au moins l'un des pics x1 et x2 est élevé

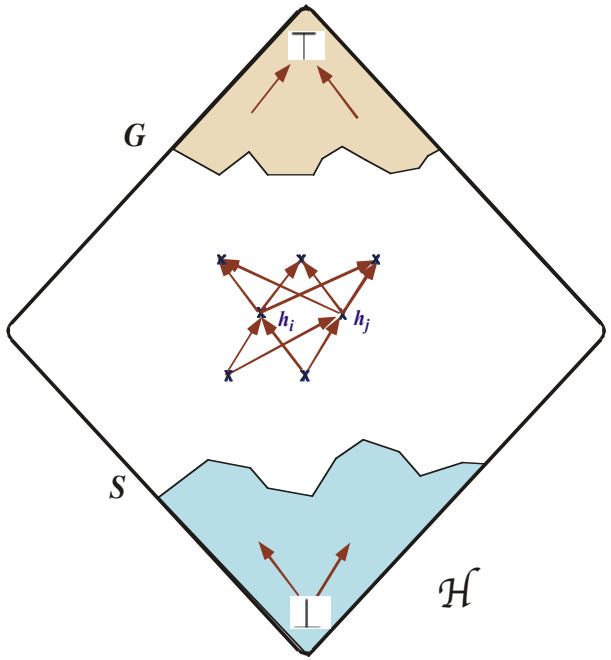
  Alors la molécule contient un groupe cétone

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*

Illustration: LEX (Tom Mitchell)

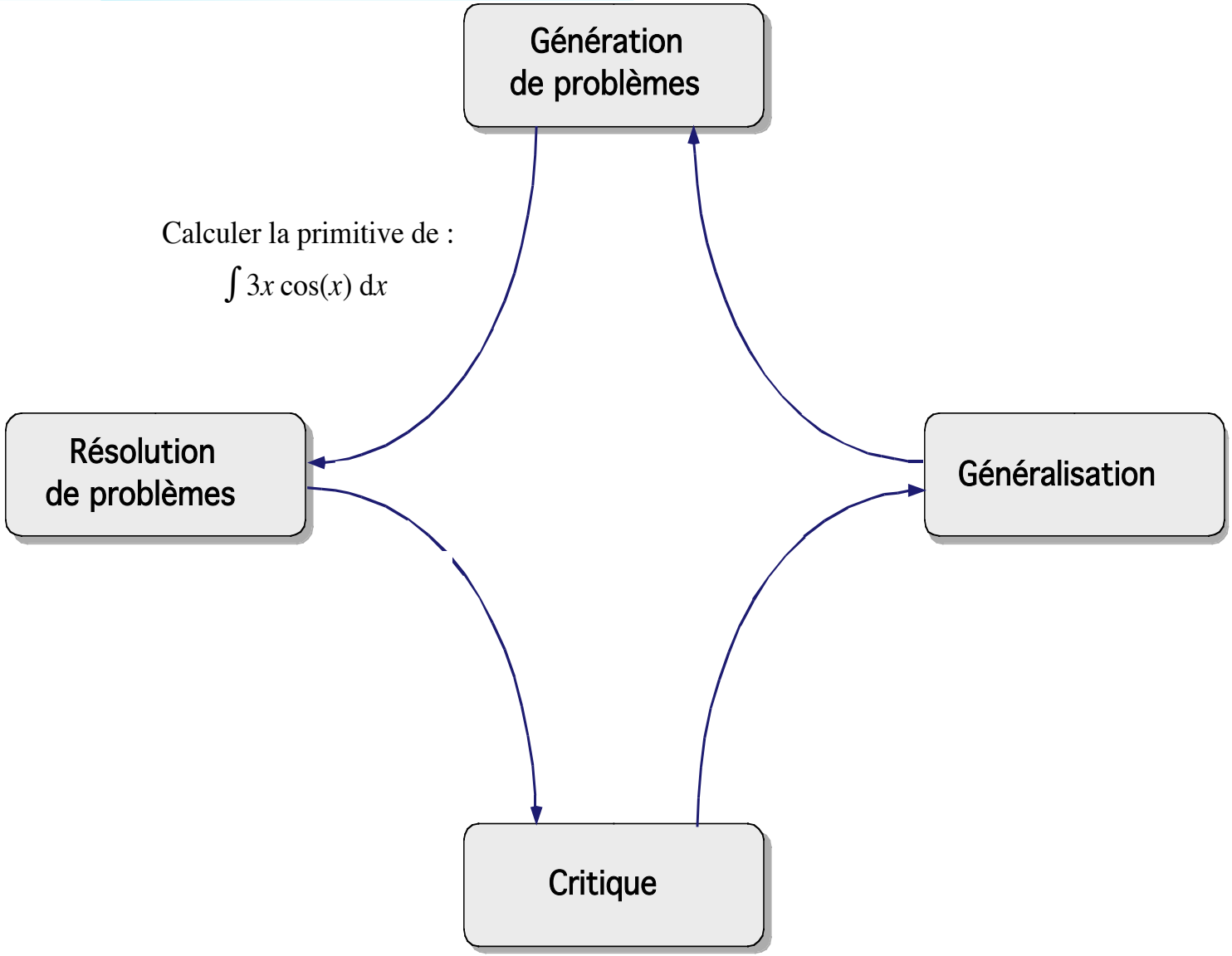


Illustration: LEX (Tom Mitchell)

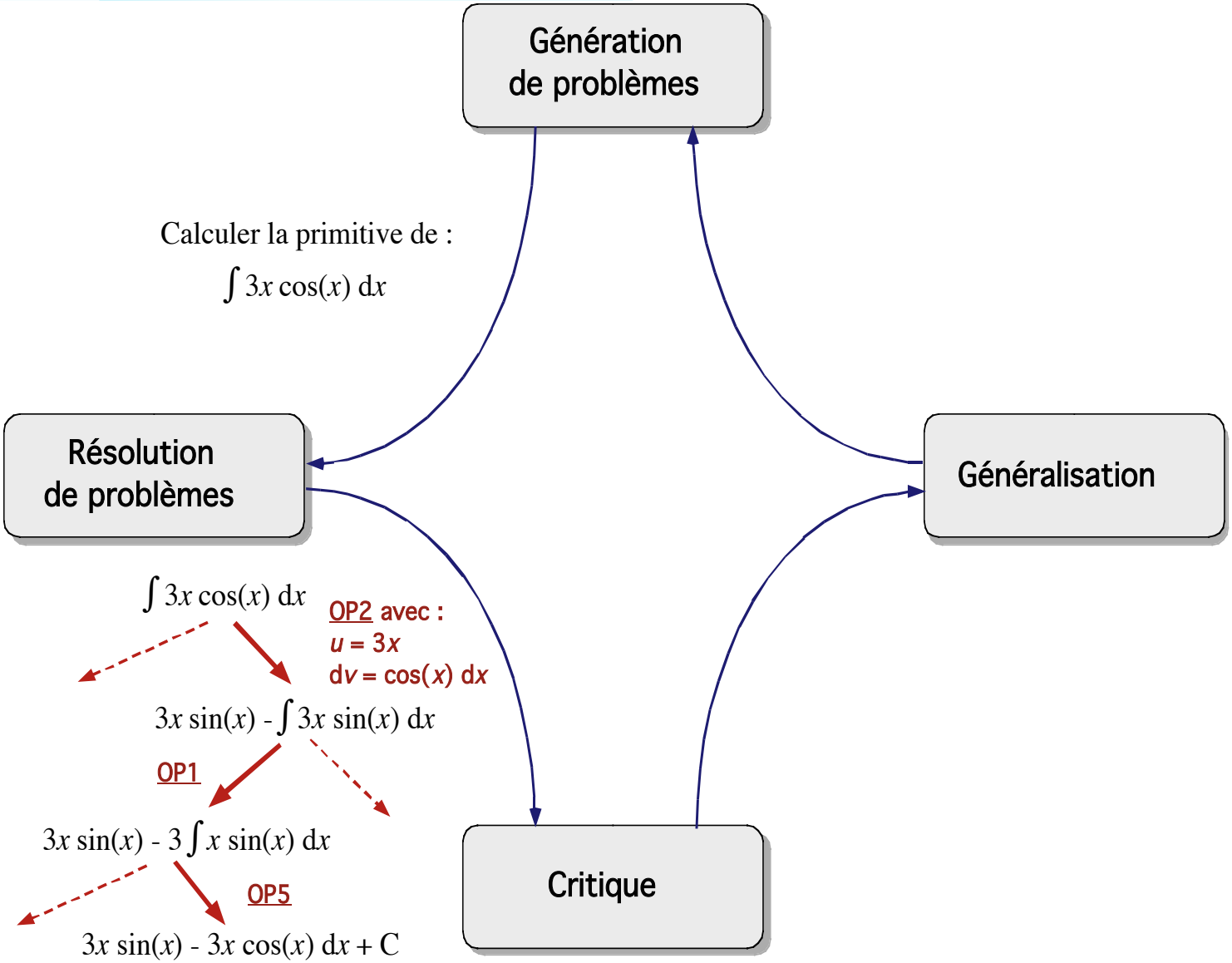


Illustration: LEX (Tom Mitchell)

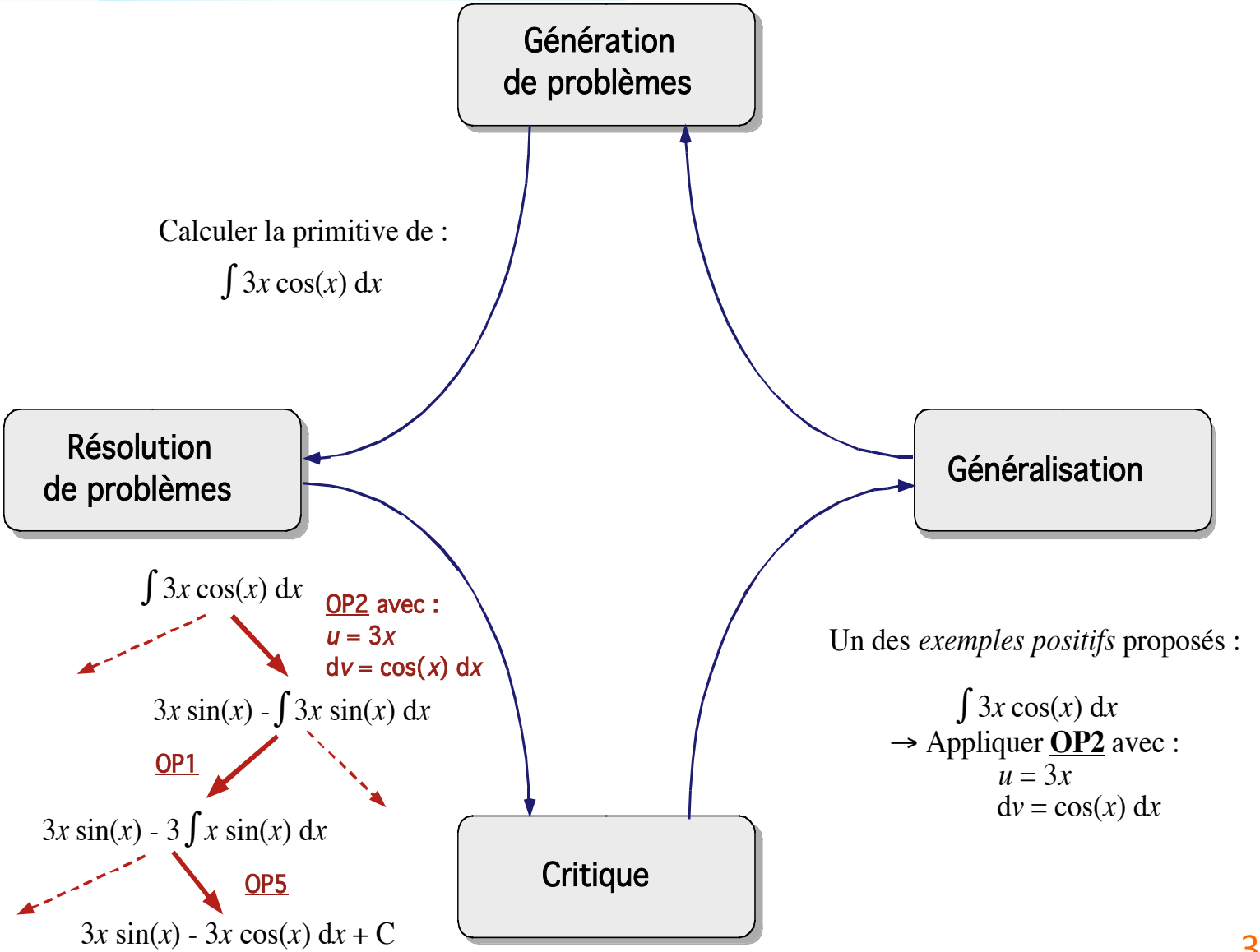
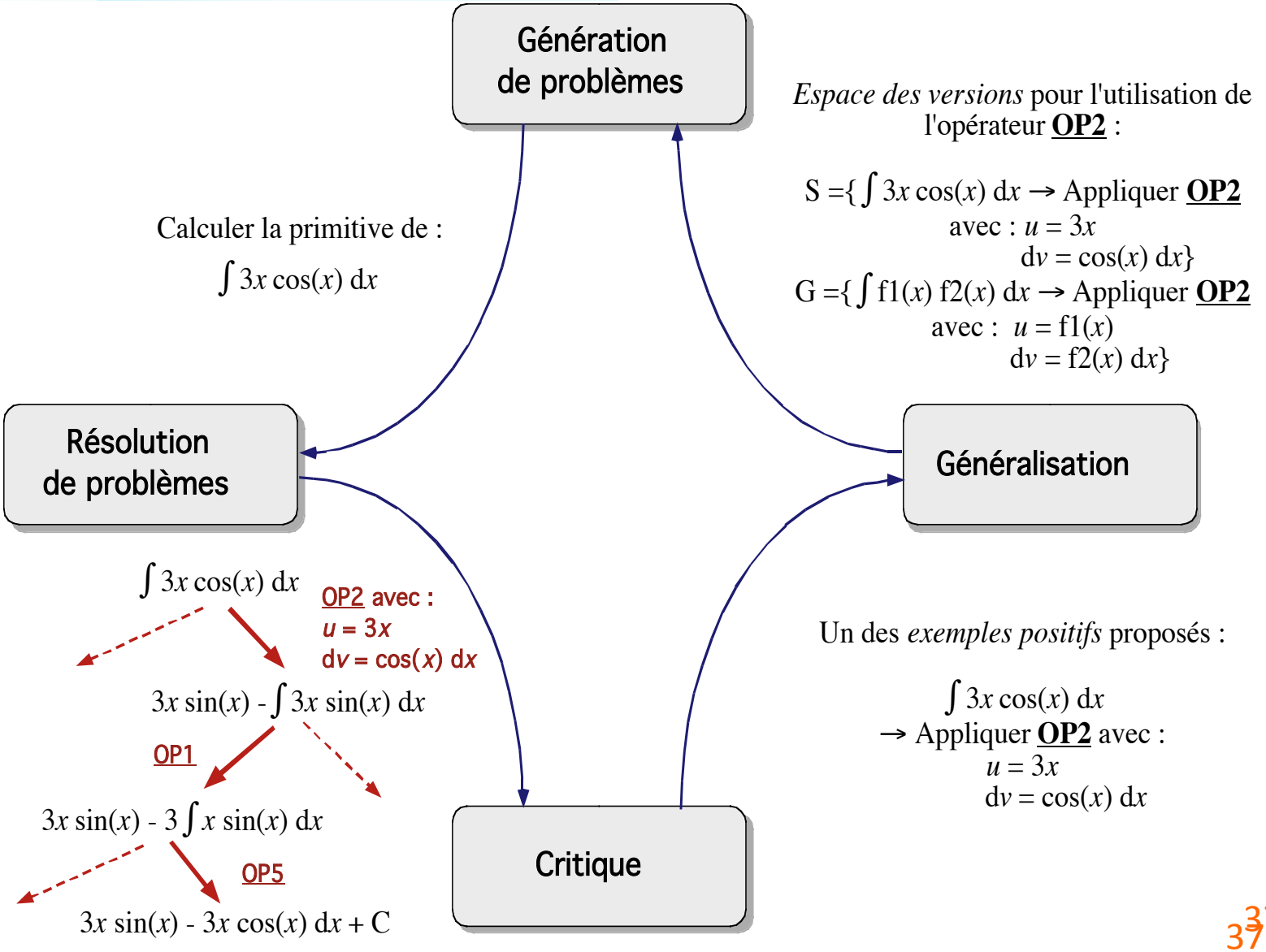


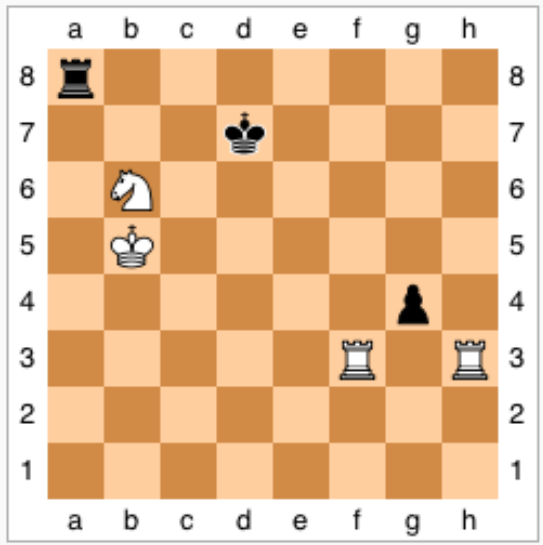
Illustration: LEX (Tom Mitchell)

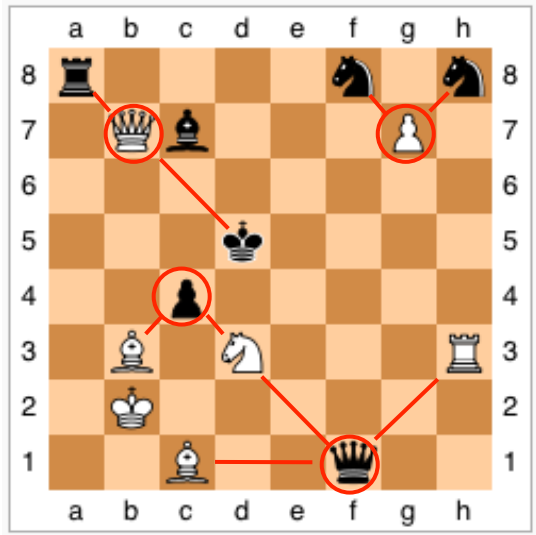


Apprendre à partir d'un **exemple**

Explanation-Based Learning

1. À partir d'un **seul exemple**
2. **Chercher à prouver** la "fourchette"
3. **Généraliser la preuve**





Explanation-Based Learning

Ex : **learn the concept** `stackable(Object1, Object2)`

- **Domain theory** :

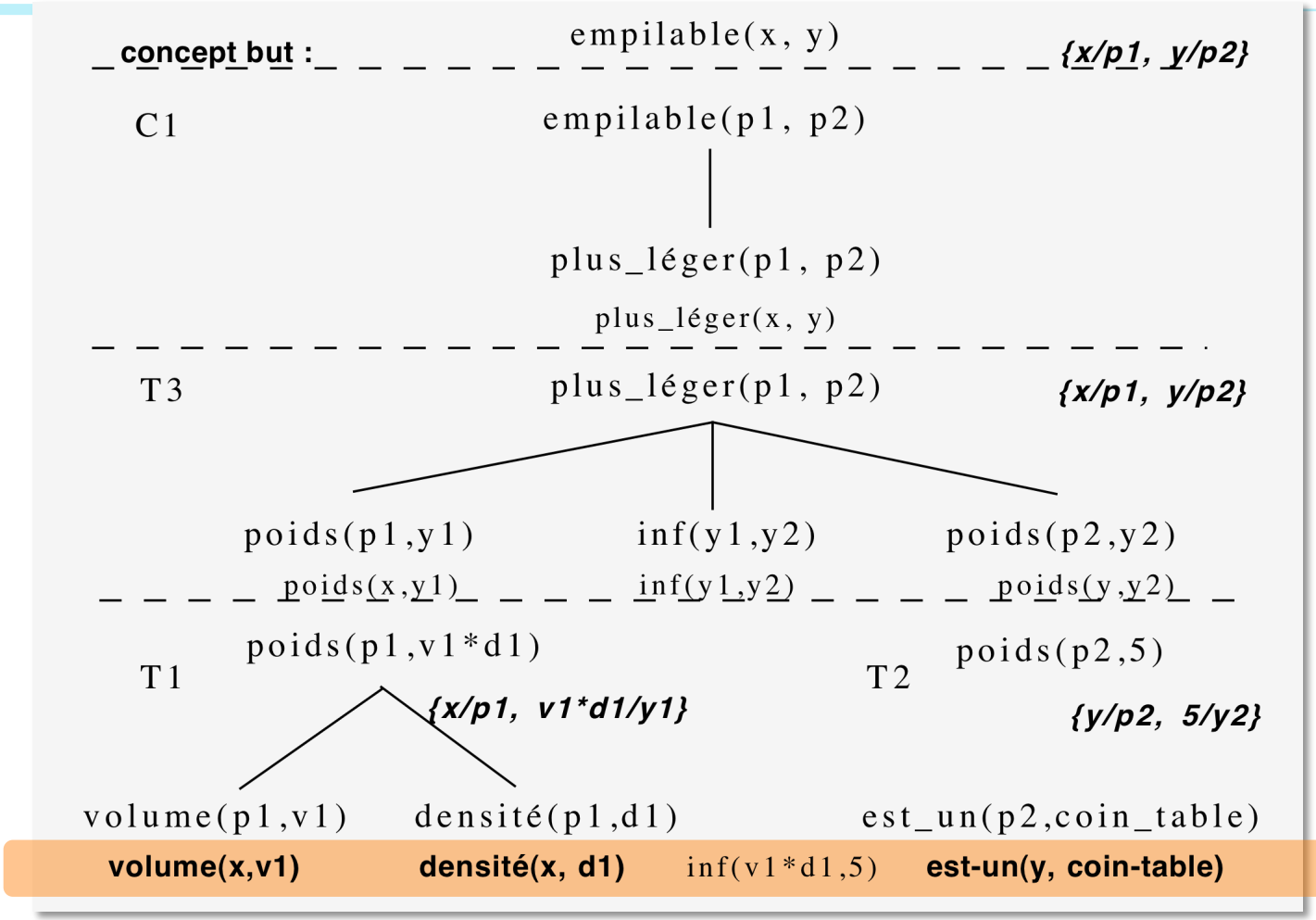
```
(T1) : weight(X, W) :- volume(X, V), density(X, D), W is V*D.

(T2) : weight(X, 50) :- is_a(X, table).

(T3) : lighter_than(X, Y) :- weight(X, W1), weight(X, W2), W1 < W2.
```

- **Operationality constraint**:
  - Concept should be expressible using *volume, density, color, …*
- **Positive example** (solution) :

```
on(obj1, obj2).
is_a(object1, box).
is_a(object2, table).
color(object1, red).
color(object2, blue).
made_of(object2, wood).
```

```
volume(object1, 1).
volume(object2, 0.1).
owner(object1, frederic).
density(object1, 0.3).
Made_of(object1, cardboard).
owner(object2, marc).
```

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.

Explanation-Based Learning

- Induction **from a single example**
  - … plus a strong domain theory
- Based on
  - **Logic-based** knowledge representation
  - **Reasoning Operators** (deduction, goal regression in a proof tree, …)

Now used in SAT "solvers"

Explanation-Based Learning

- What was the **aim** of learning?
- What was a **good theory/ method** of learning ?

Explanation-Based Learning

- What was the **aim** of learning?
- What was a **good method** of learning ?

1. Method **improving** the **problem solving performances**
   - [Steve Minton (1990) « *Quantitative results concerning the **utility** of Explanation-Based Learning* »]
2. Method that **simulates** the performances (and limits) of a **natural cognitive agent** (human or animal)
   - [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* »]

Learning and reasoning

o

But

No measure of generalization performance **independent of the problem-solver**

Difficulties to **scale up** and to face **noisy data**

… when data started to pour down

Bilan (2) : publiable

1. Un bon système d'apprentissage permet
   une résolution de problème plus efficace (rapide)
   - Apprentissage d'**heuristique**s

2. Permet d'accroître la base de connaissances
   (e.g. base de règles)
   - Apprentissage de **concepts**

Bilan (2) : aspects théoriques

1. Contrôle de la recherche
    1. Converge
    2. Elagage raisonné

Espace d'hypothèses
structuré par une
relation d'ordre partiel
de généralisation

2. **Critère d'opérationalité** sur la théorie produite (applicable)

Plan

1. 1960s : Apprendre c'est **s'adapter**
2. 1970-1985 : Apprendre et **raisonner**
3. 1985 - … : La **théorie statistique** de l'apprentissage
4. Et **maintenant** ?
5. Et **demain** ?

Comment fonder l'induction ?

Le PAC learning : un tournant

- Motivations de Leslie Valiant (1984)
  1. Montrer que la **classe des concepts apprenables** (correspondant à des classes de représentations logiques (e.g. *k*-DNF)) est **non vide mais limitée**
  2. D'où la **nécessité** d'un apprentissage **cumulatif, hiérarchique** et **guidé**

[L. Valiant (1984) « A theory of the learnable », Com. of the ACM, 27(11): 1134-1142]

Le PAC learning : un tournant

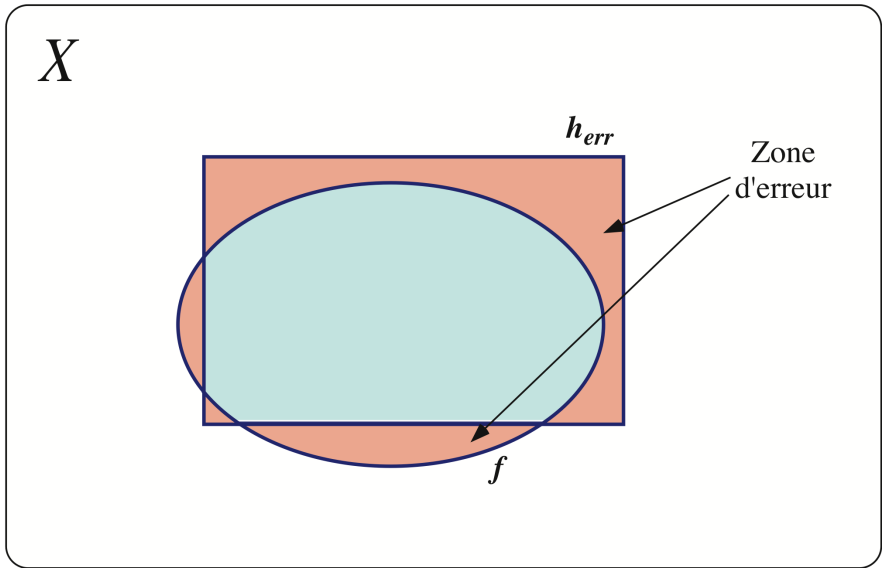
- Motivations de Leslie Valiant (1984)
  1. Montrer que la **classe des concepts apprenables** (correspondant à des classes de représentations logiques (e.g. *k*-DNF)) est **non vide mais limitée**
  2. D'où la **nécessité** d'un apprentissage **cumulatif, hiérarchique** et **guidé**

- Une classe d'hypothèses $\mathcal{H}$ est PAC apprenable si :
  - $\forall \varepsilon, \delta \in [0,1]^2$, **pour toute distribution** $\mathcal{D}$ sur $\mathcal{X}$ et **pour toute fonction cible** f : $\mathcal{X}$ -> {0,1} et $f \in \mathcal{H}$, il existe un algorithme d'apprentissage qui, utilisant $m(\mathcal{H}, \varepsilon, \delta)$ exemples tirés selon $\mathcal{D}$ retourne une hypothèse *h* telle que *h* est d'erreur < ε par rapport à *f* sur au moins 1-δ échantillons d'apprentissage.

[L. Valiant (1984) « A theory of the learnable », Com. of the ACM, 27(11): 1134-1142]

L'analyse « PAC learning »

- Supposons $f \in \mathcal{H}$. À tout instant il existe **au moins une hypothèse** dans l'espace des versions (**d'erreur nulle**)
  - Je choisis(*) une hypothèse apparemment sans erreur : h_{err}
  - La probabilité d'erreur de h_{err} est égale à la probabilité de tirer un exemple dans la zone d'erreur (différence entre f et h_{err})



L'analyse « PAC learning »

- 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

Le problème de l'induction

- Exemples décrits par :

  Nombre (1 ou 2); ***taille*** (petit ou grand); ***forme*** (cercle ou carré); ***couleur*** (rouge ou vert)

- De classe + ou -

Description	Votre prédiction	**Vraie** classe

Combien de fonctions possibles de *X* vers *Y* ? $2^{2^4} = 2^{16} = 65{,}536$

Combien de fonctions restent après 6 exemples d'apprentissage? $2^{10} = 1024$

Le problème de l'induction

- Exemples décrits par :

  Nombre (1 ou 2); ***taille*** (petit ou grand); ***forme*** (cercle ou carré); ***couleur*** (rouge ou vert)

15

Description	**Votre** prédiction	**Vraie** classe
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 r carré ouge		+
2 grands carrés verts		+
2 petits carrés verts		+
2 petits cercles rouges		+
1 petit cercle vert		-
2 grands cercles verts		-
2 petits cercles verts		+
1 grand cercle rouge		-
2 grands carrés rouges	?	

Combien de fonctions restantes ?

← ?

Le problème de l'induction

- Examples described using:

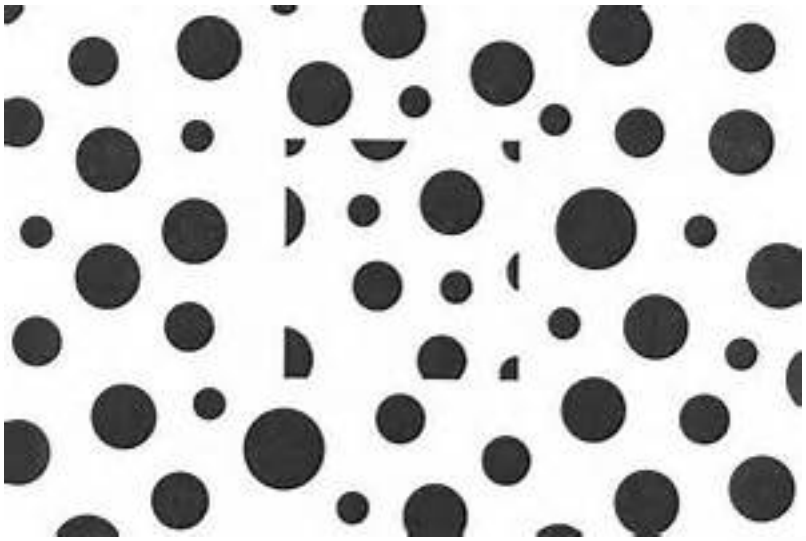
  ~~***Number***~~ (1 or 2); ***size*** (small or large); ~~***shape***~~ (circle or square); ***color*** (red or green)

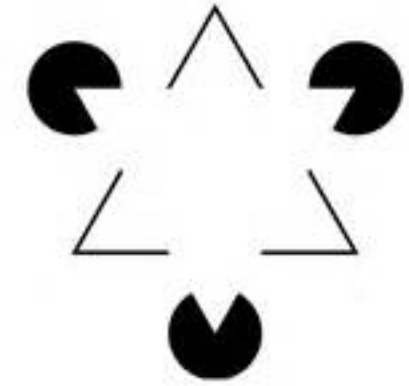
Description	**Your** prediction	**True** class
~~1~~ grand ~~carré~~ rouge		-
~~1~~ grand ~~carré~~ vert		+
~~2~~ petits ~~carrés~~ rouges		+
~~2~~ grands ~~cercles~~ rouges		-
~~1~~ grand ~~cercle~~ vert		+
~~1~~ petit ~~cercle~~ rouge		+

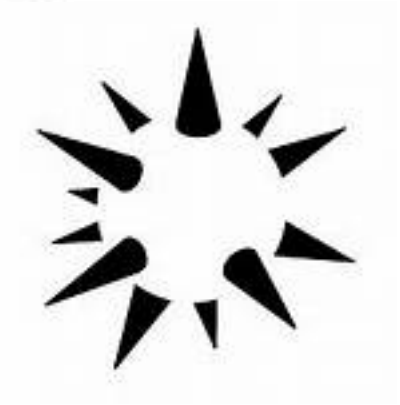
Combien de fonctions possibles avec 2 descripteurs de *X* à *Y* ? $2^{2^2} = 2^4 = 16$

Combien de fonctions restent après 3 exemples différents ? $2^1 = 2$

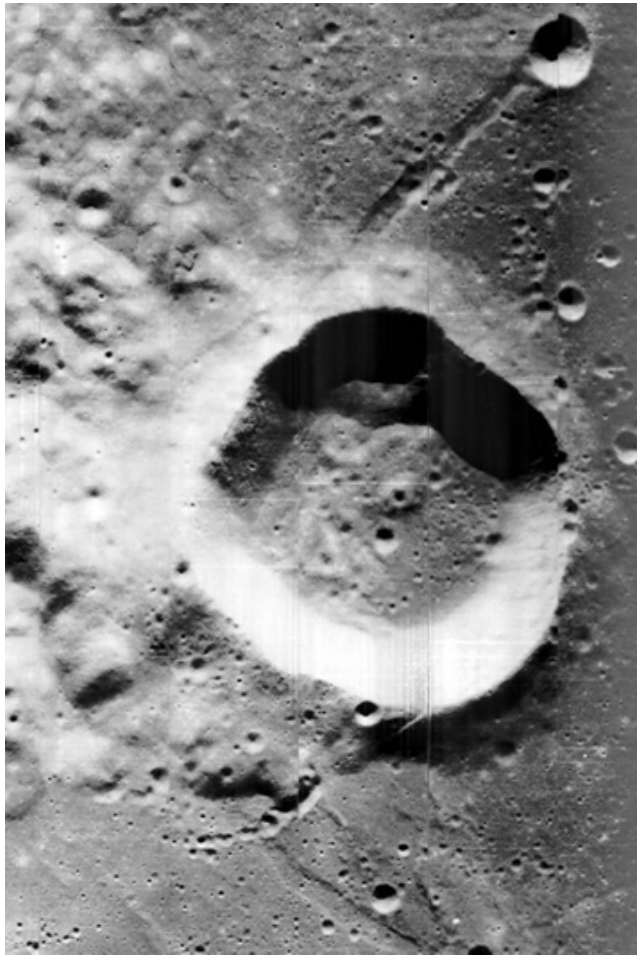
L’apprentissage – une extrapolation nécessitant des a apriori

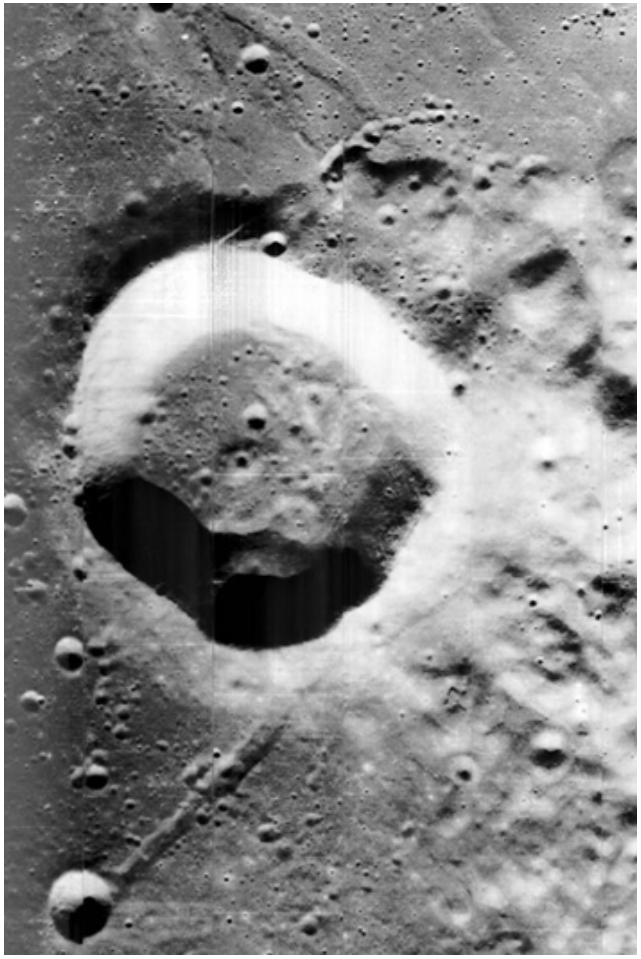






Induction et illusions





Cratère ou colline ?

L'analyse « PAC learning »

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

- *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*)

La question est résolue sans notion d'algorithme !!!

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

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

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 :

Risque empirique régularisé

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

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 :

Expression de **substitution** du risque régularisé

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

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

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

Optimisation

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 :

Bilan (3) : l'empire des normes

- Une **démarche générique** et générale
  - Définition d'un **risque régularisé**
    - Traduisant des attentes sur les régularités d'intérêt
    - Assurant problème convexe
  - Algorithme d'**apprentissage** = algorithme d'**optimisation**

Bilan (3) : l'empire des normes

- Une **démarche générique** et générale
  - Définition d'un **risque régularisé**
    - Traduisant des attentes sur les régularités d'intérêt
    - Assurant problème convexe
  - Algorithme d'**apprentissage** = algorithme d'**optimisation**
- Un **certificat d'excellence**
  - Bornes en généralisation, bien mathématiques

Bilan (3) : l'empire des normes

- Une **démarche générique** et générale
  - Définition d'un **risque régularisé**
    - Traduisant des attentes sur les régularités d'intérêt
    - Assurant problème convexe
  - Algorithme d'**apprentissage** = algorithme d'**optimisation**
- Un **certificat d'excellence**
  - Bornes en généralisation, bien mathématiques
- Des **présupposés** supposés **modestes**
  - Et adaptés au « big data »

Données et questions i.i.d.

Bilan (3) : publiable

1. Applique les préceptes de la **théorie statistique de l'apprentissage**
   - Algorithme d'optimisation d'un risque régularisé (e.g. SVM, Boosting, semi-supervisé, LASSO, …)

2. Contient des garanties sous forme de **bornes**
   - Entre risque **empirique** et risque **réel**

Bilan (2) : aspects théoriques

1. Le no-free-lunch theorem
    1. À chaque problème, ses présupposés
    2. Les traduire sous forme de biais

Bilan (2) : aspects théoriques

1. Le no-free-lunch theorem
    1. À chaque problème, ses présupposés
    2. Les traduire sous forme de biais

2. Si le biais est adapté au monde

    -> garantie sur le résultat de l'algorithme

Bilan (2) : aspects théoriques

1. Le no-free-lunch theorem
    1. À chaque problème, ses présupposés
    2. Les traduire sous forme de biais

2. Si le biais est adapté au monde

    -> garantie sur le résultat de l'algorithme

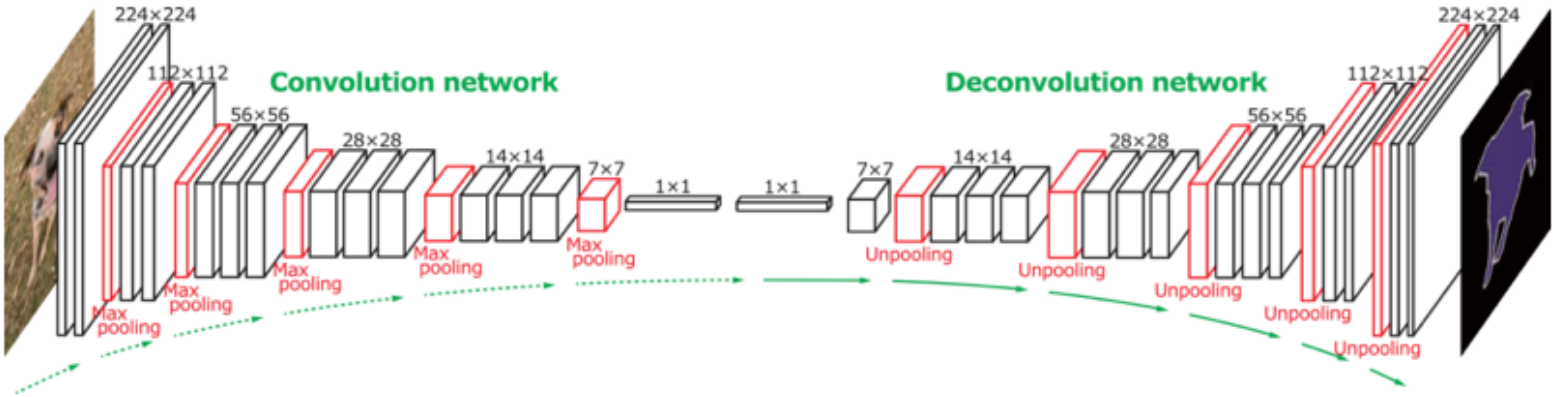
Garanties de « lampadaire »

Plan

1. 1960s : Apprendre c'est **s'adapter**
2. 1970-1985 : Apprendre et **raisonner**
3. 1985 - … : La **théorie statistique** de l'apprentissage
4. Et **maintenant** ?
5. Et **demain** ?

Fin de l’histoire ?

Les réseaux de neurones **profonds**



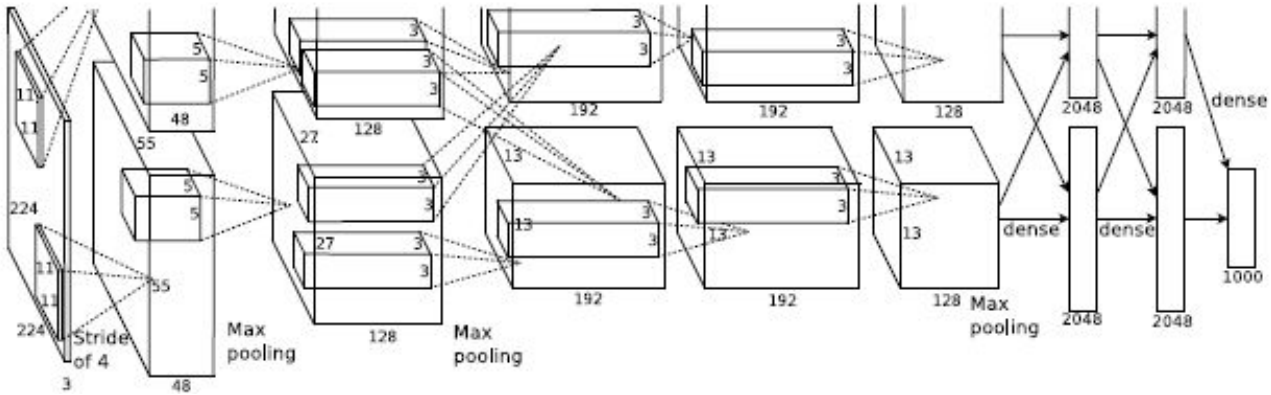
- On s'amuse à nouveau à construire des réseaux (et des algorithmes) qui résolvent des problèmes (plutôt que seulement réduire un taux d'erreur)

Quelque chose de **troublant**

– C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals (ICLR, **May 2017**).
"Understanding deep learning requires rethinking generalization"

Extensive experiments on the classification of images

– The AlexNet (**> 1,000,000 parameters**) + 2 other architectures



– The **CIFAR-10 data set**:

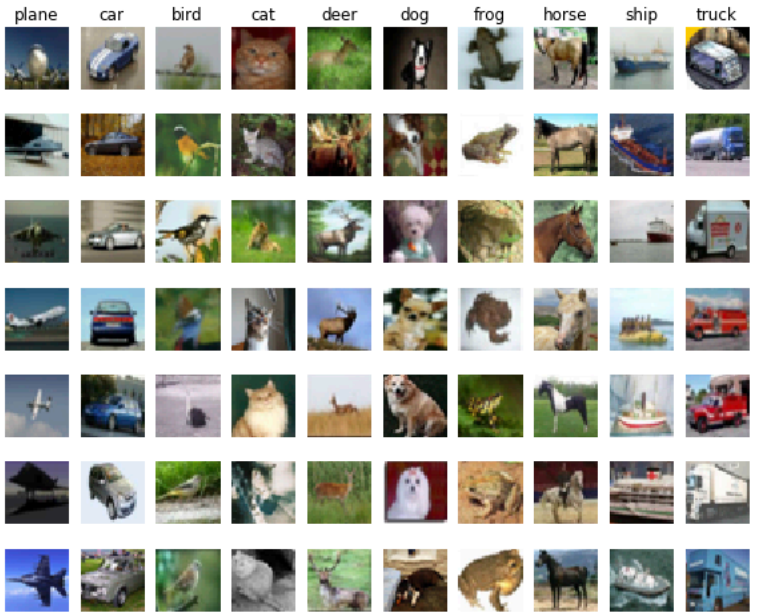
- **60,000** images categorized in **10 classes** (50,000 for training and 10,000 for testing)
- Images: 32x32 pixels in 3 color channels

Quelque chose de troublant

Experiments

1. **Original dataset** without modification

- Results ?
  - **Training** accuracy = 100% ; **Test** accuracy = 89%
  - Speed of convergence ~ 5,000 steps



Troubling findings

Experiments

1. **Original dataset** without modification
   - Results ?
     - **Training** accuracy = 100% ; **Test** accuracy = 89%
     - Speed of convergence ~ 5,000 steps

2. Random **labels**
   - **Training** accuracy = 100% !!?? ; **Test** accuracy = 9.8%
   - Speed of convergence = similar behavior (~ 10,000 steps)

3. Random **pixels**
   - **Training** accuracy = 100% !!?? ; **Test** accuracy ~ 10%
   - Speed of convergence = similar behavior (~ 10,000 steps)

Now, we are in trouble!!

Troubling findings

- Deep NNs can accommodate ANY training set

Can grow without limit!!

$$\forall h \in \mathcal{H}, \forall \delta \leq 1: \quad P^m \left[R(h) \leq \widehat{R}(h) + 2\,\widehat{Rad}_m(\mathcal{H}) + 3\sqrt{\frac{\ln(2/\delta)}{m}} \right] > 1 - \delta$$

But then,

why are deep NNs ***so good on image classification tasks****?*

Le **taux d'erreur** …

… est une mesure insuffisante

- Recommandation
- Écrire la légende d'une image
- Sorties structurées

La **qualité** de la **solution** dépend de plusieurs facteurs

Image annotating

Figure 2.11: "A group of young people playing a game of frisbee"—that caption was written by a computer with no understanding of people, games or frisbees.

Adversarial learning

Question centrale : quelles garanties ?

Boxer: 0.40 Tiger Cat: 0.18

(a) Original image

Airliner: 0.9999

(b) Adversarial image

!!??

[Selvaraju et al. (2017) « *Grad-CAM: Visual explanations from deep networks via gradient-based localization* »]

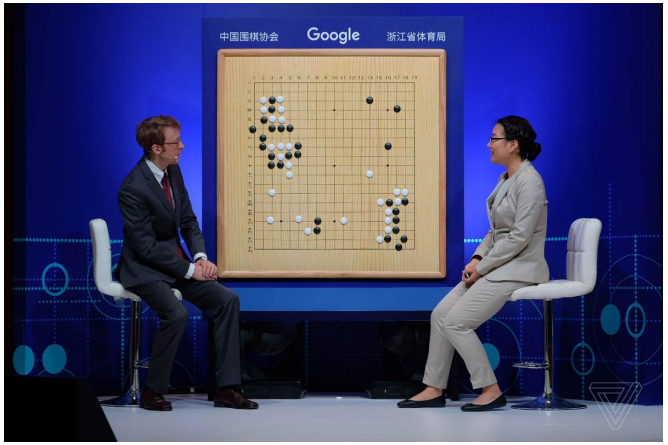
Le cas AlphaGo

- Un joueur « extraterrestre »
- Un jeu stupéfiant
- Révolutionne la manière de jouer
- Effervescence dans les écoles de go

Le cas AlphaGo : comprendre

Fan Hui, Gu Li, Zhou Ruyang (très forts joueurs de Go) se reconvertissent dans l'analyse des parties jouées par AlphaGo

- Sorte d'exégèse. Explications a posteriori
- Nécessaire pour
  - La communication
  - L'enseignement



Et même AlphaGo peut se tromper

Les explications des SEs

Pourquoi ne faut-il pas prescrire de tétracycline à un enfant de moins de 8 ans ?

Connaissances justificatives

Dépôt de la drogue sur les **os en développement**

→ **Noircissement** définitif des dents

→ Coloration socialement **indésirable**

→ **Ne pas administrer** de tétracycline aux enfants de moins de 8 ans

Notion d'**effets secondaires** indésirables

Relations de **causalité**

Les explications des SEs

Pourquoi ne faut-il pas prescrire de tétracycline à un enfant de moins de 8 ans ?

Les explications des SEs

Pourquoi ne faut-il pas prescrire de tétracycline à un enfant de moins de 8 ans ?

Connaissances justificatives

Dépôt de la drogue sur les **os en développement**

→ **Noircissement** définitif des dents

→ Coloration socialement **indésirable**

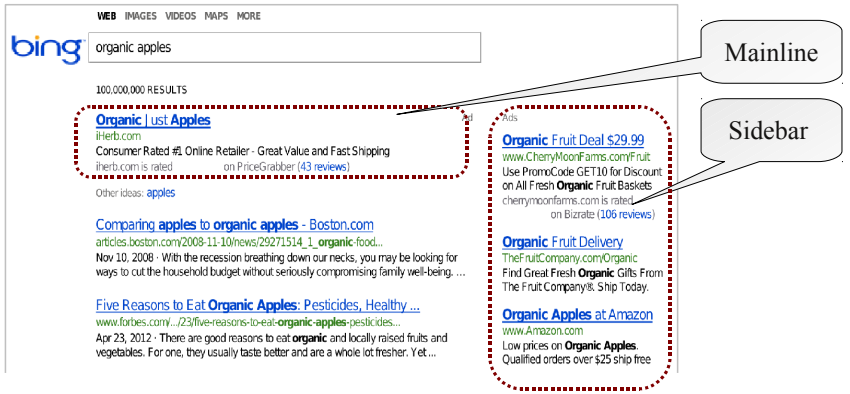
→ **Ne pas administrer** de tétracycline aux enfants de moins de 8 ans

Notion d'**effets secondaires** indésirables

Relations de **causalité**

Interactions entre systèmes apprenants

Système adaptatif de placement de publicité



- Deux **sous-systèmes**
  - L'un plaçant les **liens publicitaires**
  - L'autre choisissant les **publicités**
- Qui **s'influencent mutuellement**
  - Chacun s'appuie sur les données de clicks
  - Qui dépendent aussi de l'intervention de l'autre systèmes
  - Et d'autres facteurs non contrôlés (prix, requête de l'utilisateur, …)

[L. Bottou et al. «*Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising* », JMLR, 14, (2013), 3207-3260]

Bilan (4) : publiable

1. Contient le mot « **deep** »
2. Attaque un problème avec **beaucoup de données**
3. Est lié à un problème « **difficile** » et **utile pour l'industrie**
4. Garanties en **contexte adversarial**

Plan

1. 1960s : Apprendre c'est **s'adapter**
2. 1970-1985 : Apprendre et **raisonner**
3. 1985 - … : La **théorie statistique** de l'apprentissage
4. Et **maintenant** ?
5. Et **demain** ?

Voiture dans une piscine

… ou pas de voiture … ?

Is this less of a car
because the context is wrong?

[Léon Bottou (ICML-2015, invited talk) « *Two big challenges in Machine Learning* »]

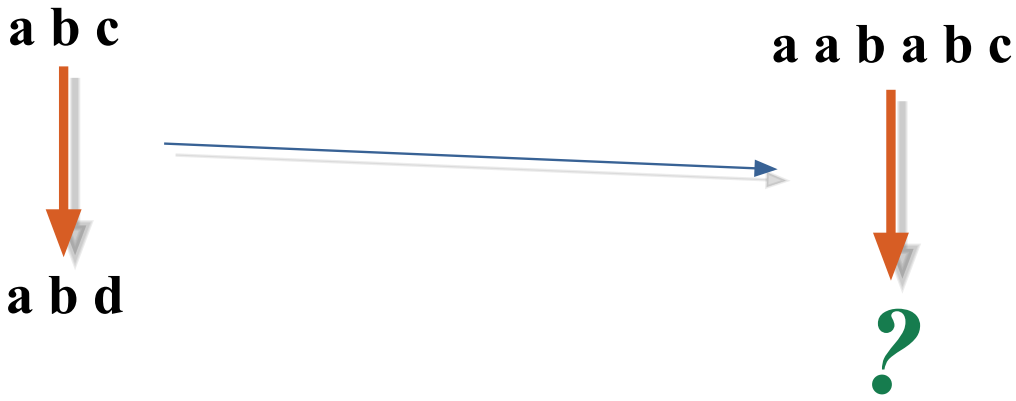
Autres scénarios

- Apprendre à partir de (très) peu d'exemples
  - Ce que nous faisons sans cesse
  - Nous sommes des **constructeurs de théories** (que nous testons ensuite)

Ré-introduire le **raisonnement**

Autres scénarios

- Apprentissage par analogie



Why should ‘**a a b a b c d**’ be any better than ‘**a b d**’?

Transfert learning: questions

- What can be **the basis** of transfer learning?

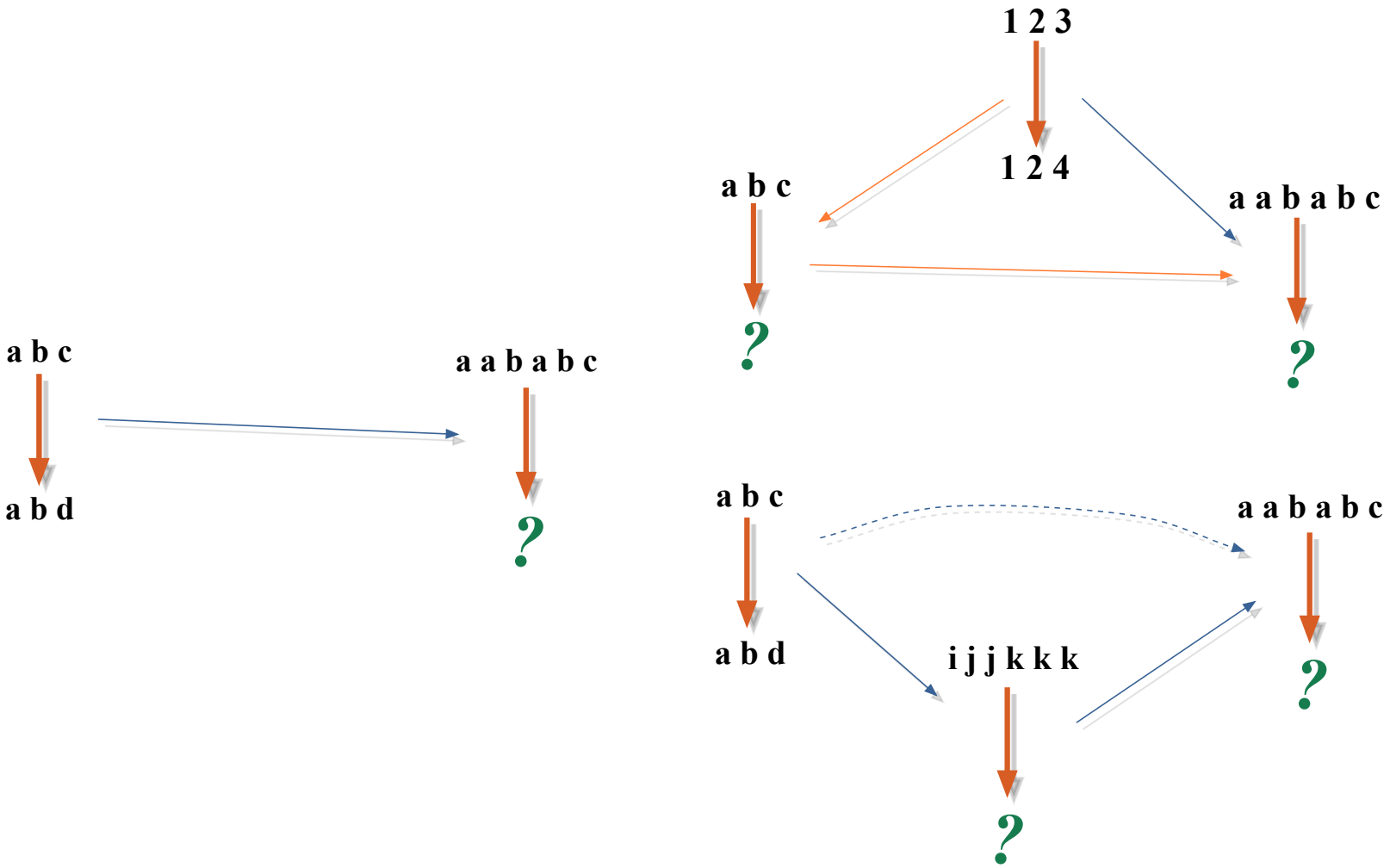
  How to translate formally :

  *"the target domain **is like** the source domain"*?

Not i.i.d. anymore

- What **determine a good transfer**?
  - A "good source"?
  - A high "similarity" between source and target?
- What **formal guarantees** can we have on the transferred hypothesis?

Transfer and sequence effects

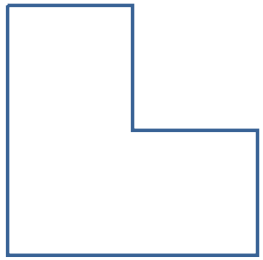


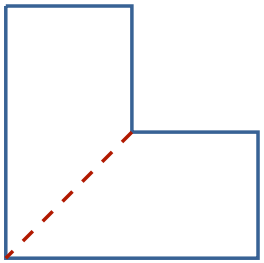
t

Effets de séquence

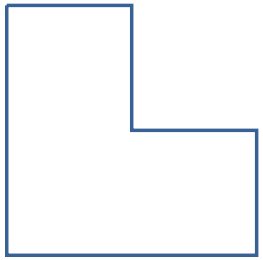
Instruction : découper la figure géométrique suivante en *n* parties **superposables**.

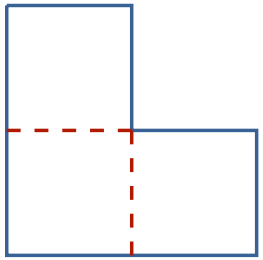
En 2:



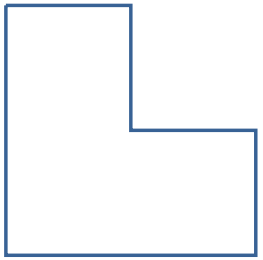


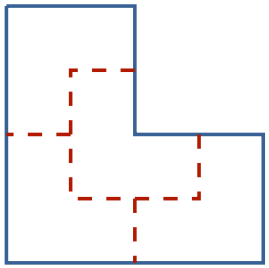
En 3:



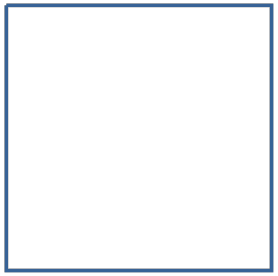


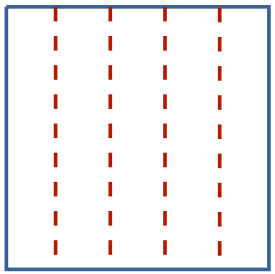
En 4:





En 5:





Perspectives

- **Nouveaux scénarios** d'apprentissage
  - Long-life learning et transfert
    - Transfert **constructif** et transfert **négatif**
  - Construction de curriculum
    - [Foglino et al. « ***Curriculum learning for cumulative return maximization*** », IJCAI-2019]
      - *A central aspect of curriculum learning is task sequencing, since the order in which the tasks are executed is a major factor in the quality of a curriculum.*
      - *Curriculum learning leverages transfer learning to transfer knowledge through the curriculum, in order to benefit a final task.*

Un **pari**

Vers des systèmes qui savent comment enseigner

1. **Expliquer** un cas
2. **Synthétiser**
3. Organiser un **curriculum**

➢ **Évaluer** les systèmes par **la performance de leurs élèves ?**

Un espoir

*"Another great strength of this book is the way that **ideas build upon one another**. The author has masterfully written a book in which your intuition about **early concepts pave the way for understanding later concepts** even though missing some ideas in the beginning will not cripple you in later chapters."*

À propos de [Steven Strogatz "Non linear dynamics and chaos", 2014]

t

Suppléments

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**

Les machines apprenantes ne sont pas des machines pensantes