Plan

1. Pourquoi toute cette excitation ?
2. Grands types d'apprentissage
3. Apprentissage prédictif par réseaux de neurones
4. Quelles garanties ?
5. Le no-free-lunch theorem
6. Les réseaux de neurones profonds
7. Ce que l'on sait faire et les défis à relever

What do you want?

- Better **understand** your data

What do you want?

- Better **understand** your data

- Be able to make **prediction**

What do you want?

- Better **understand** your data
- Be able to make **prediction**
- Be able to make **prescription**

What do you want?

- Better **understand** your data
- Be able to make **prediction**
- Be able to make **prescription**

Apprentissage **descriptif**

non supervisé

Apprentissage descriptif

À propos d'un *échantillon d'apprentissage* $s = \{(x_i)\}_{1,m}$

identifier des régularités rendant compte de *S*

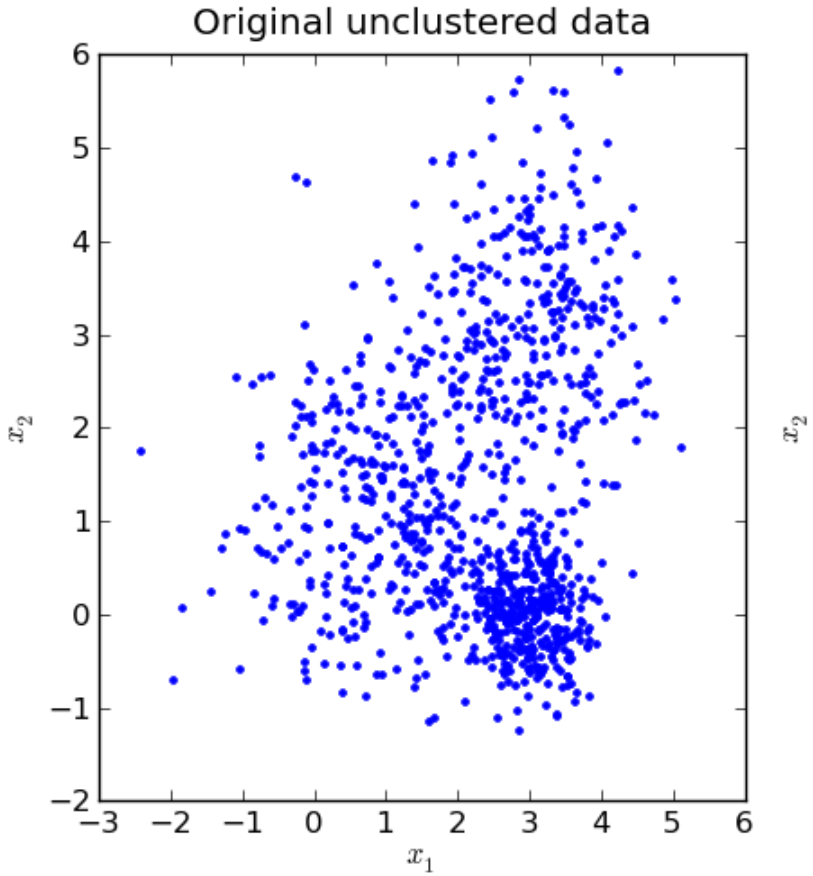
- E.g. sous la forme de **clusters** (e.g. *mélange de Gaussiennes*)
  - CLUSTERING
- E.g. sous la forme de **motifs fréquents** (fouille de données)

pour **résumer**, suggérer des **régularités**, **comprendre** …

Clustering / Catégorisation

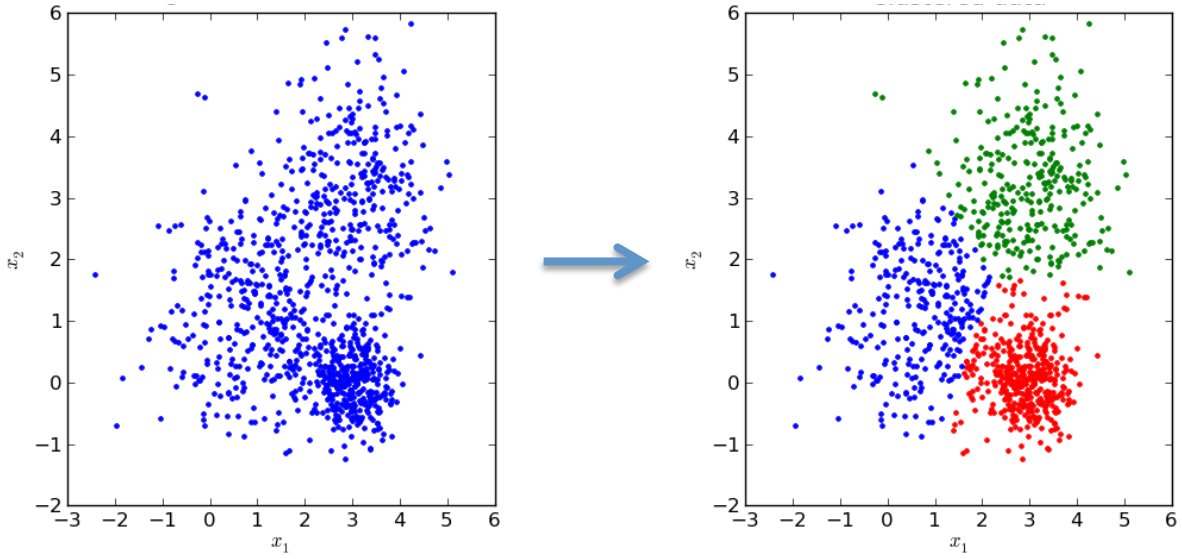
Les **grands types** d'apprentissage

- Apprentissage « **descriptif** » (non supervisé)



(1) **Understand** your data

- **Re-express** it
  - In a concise way
  - To be interpretable by an expert of the domain

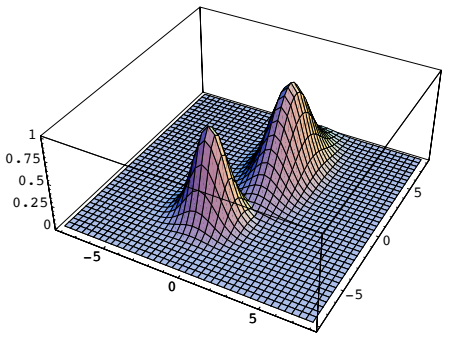


Original data

Clustered data

Three groups of customers with such and such **characteristics** ...

- Catégorisation de consommateurs
  - Base de données sur les répondants de la base Nutrinet
    - ~ 280 000
    - Données sur *âge*, *nb de personnes dans la famille*, *catégorie socio-professionnelle*, …
    - Données sur consommations alimentaires sur une certaine durée
  - Y a-t-il émergence de **groupes** distincts ?



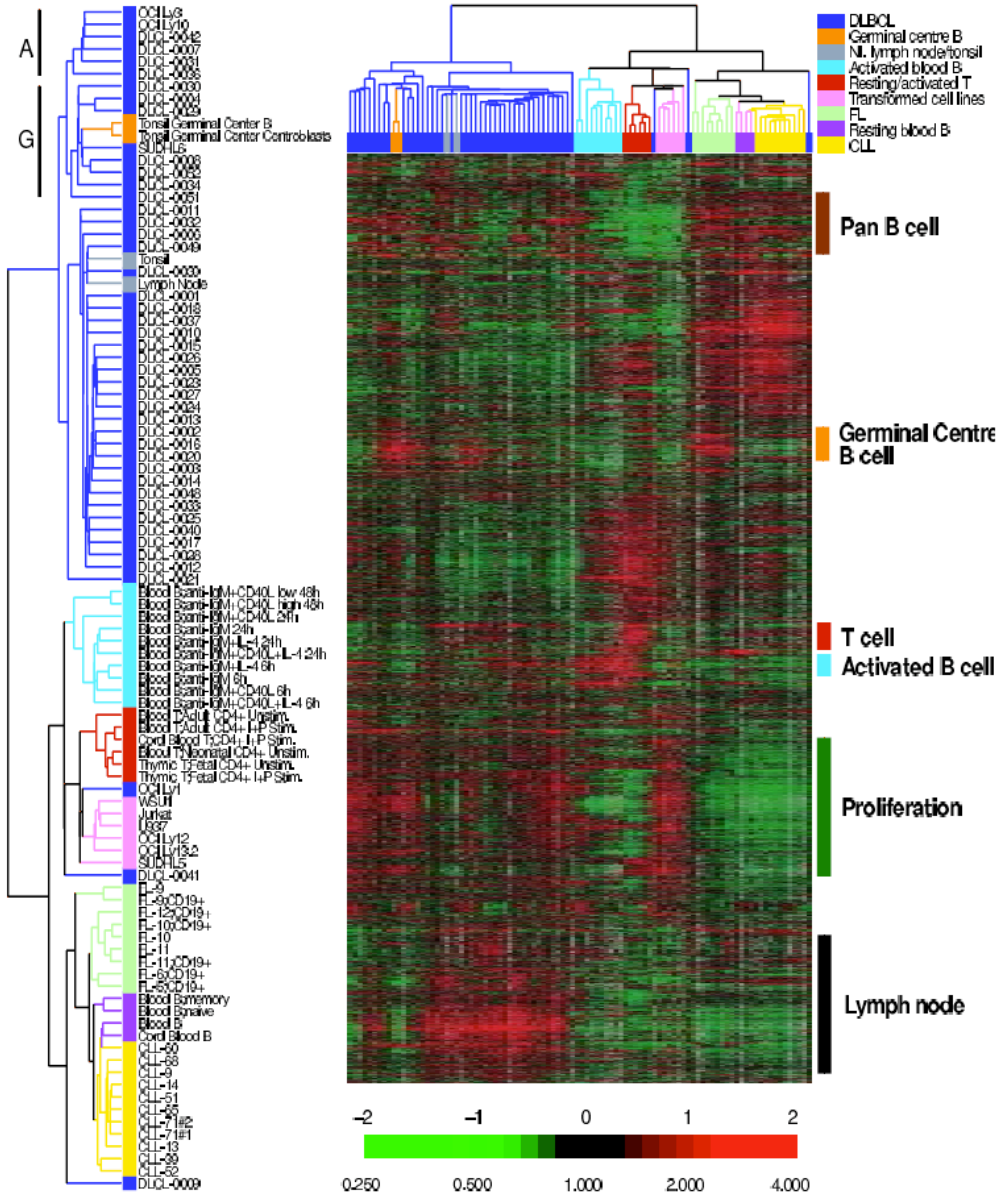
Apprentissage

Non supervisé

Clustering

Bi-clustering

gènes - patients



Recherche de motifs fréquents

Frequent Item Sets

Recherche de règles d'association

- Reconnaissance d'animaux malades ou en chaleur

  - Mesures en continu sur leur comportement

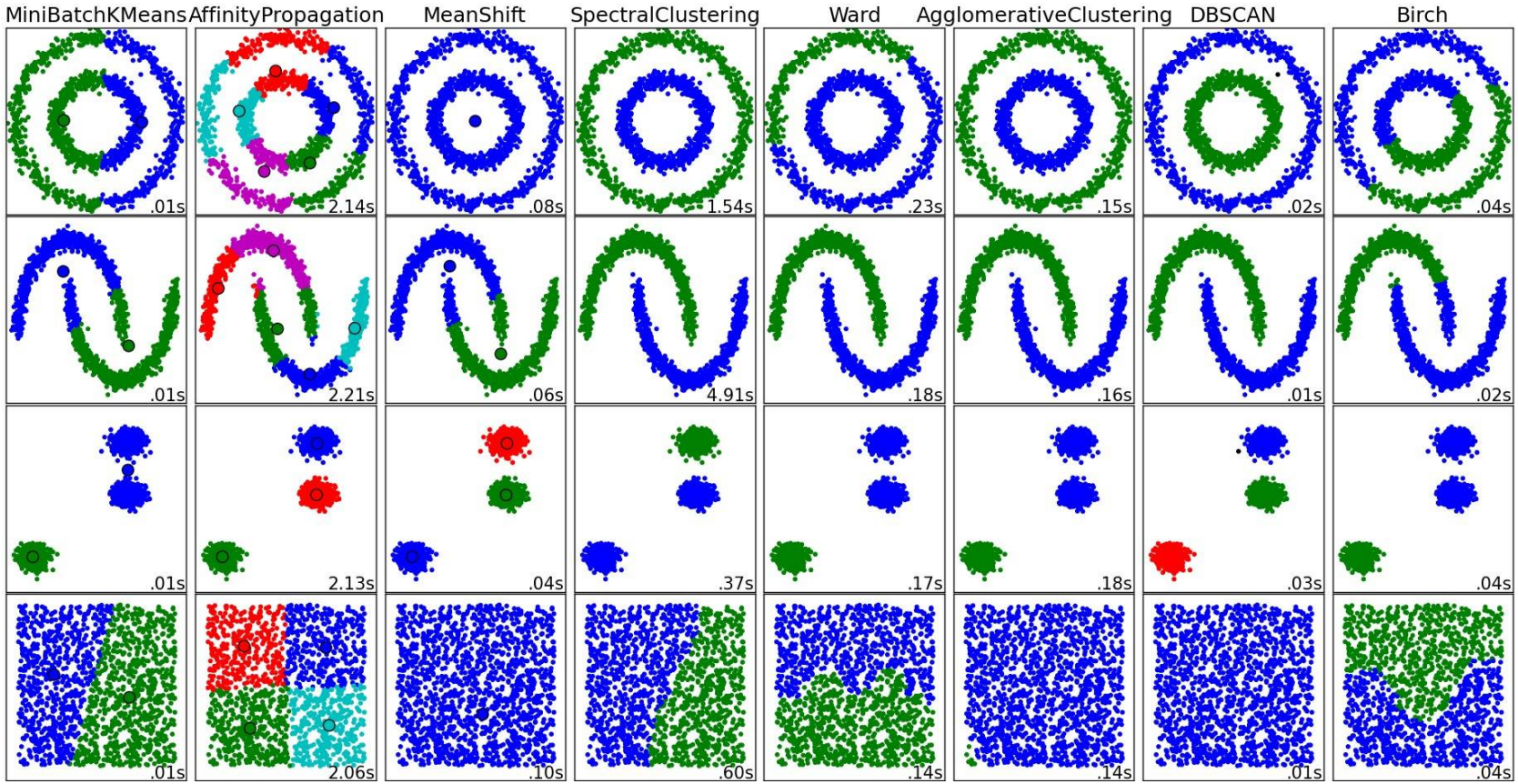
    - Vidéos
    - Capteurs « embarqués »

    - Mobilité (nb de pas / minute ; distance parcourue à l'heure)
    - Lieux visités
    - …

    ➢ *Reconnaissance de **comportements types***

Clustering

Dépend beaucoup des biais a priori



Apprentissage **prédictif**

supervisé

Apprentissage prédictif (*supervisé*)

- Extrapolate your data to find **predictive correlations**
  - From a **training** set $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i), \ldots, (x_m, y_m)\}$

(2) Make **predictions**

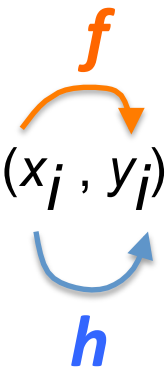
- Extrapolate your data to find predictive correlations
  - From a **training** set $S = \{(x_1, y_1), (x_2, y_2), \dots, (x_i, y_i), \dots, (x_m, y_m)\}$

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

(2) Make **predictions**

- Extrapolate your data to find predictive correlations
  - From a **training** set $S = \{(x_1, y_1), (x_2, y_2), \dots, (x_i, y_i), \dots, (x_m, y_m)\}$

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

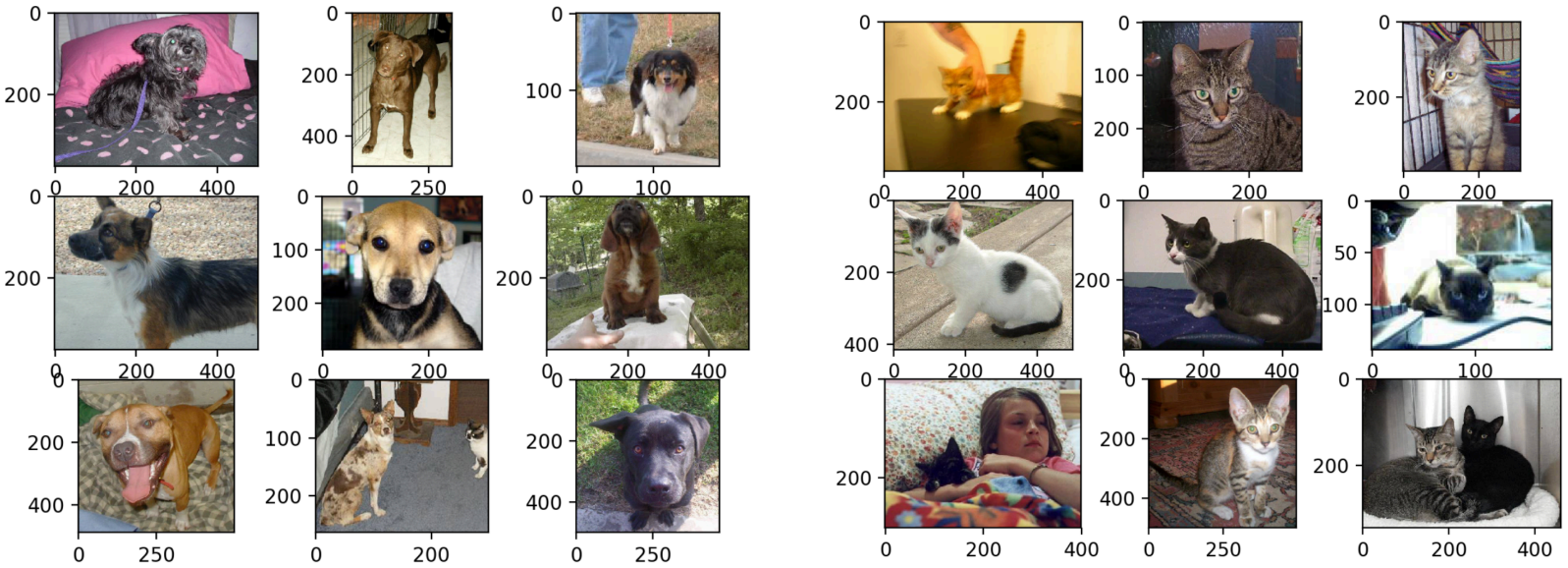


$$x - h \rightarrow y$$

(2) Supervised Learning as …

… Learning a function from an **input** space *X* to an **output** space *Y*

Cats vs. **dogs**

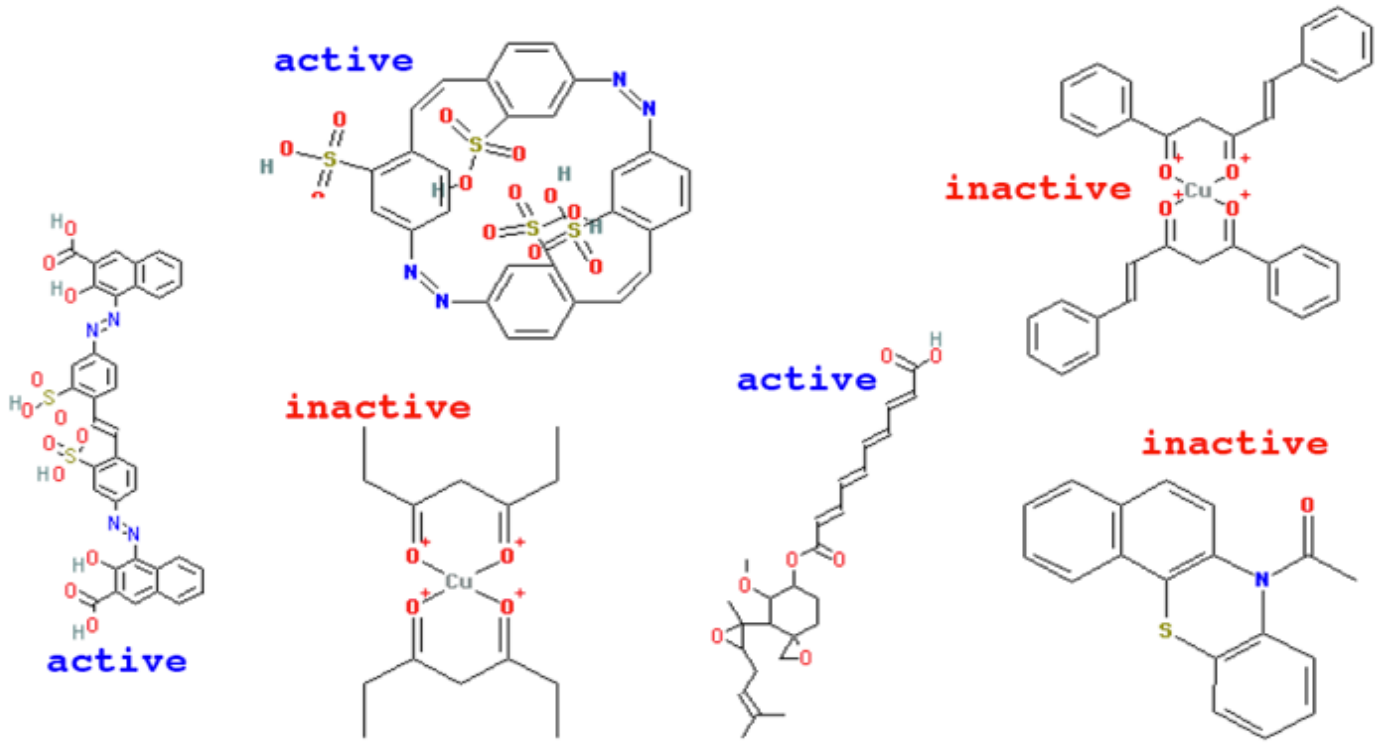


- Reconnaissance d'insectes ravageurs
  - Base d'images d'insectes dans des cuvettes
    - *Reconnaissance du type d'insectes*
    - *Comptage*

- **Spam** ou pas spam 

- Article portant sur la **politique** ou sur le **sport** 
- **Pathologie** dont souffre un patient 
- **Objet** présent dans une image 

Apprentissage prédictif



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

Apprentissage descriptif

À propos d'un *échantillon d'apprentissage* $s = \{(x_i)\}_{1,m}$

identifier des régularités rendant compte de *S*

- E.g. under the form of clusters (e.g. *mixture of Gaussians*)
  - CLUSTERING
- E.g. sous la forme de motifs fréquents (fouille de données)

pour résumer, suggérer des régularités, comprendre …

Apprentissage prescriptif

pour « intervenir »

Apprentissage prescriptif

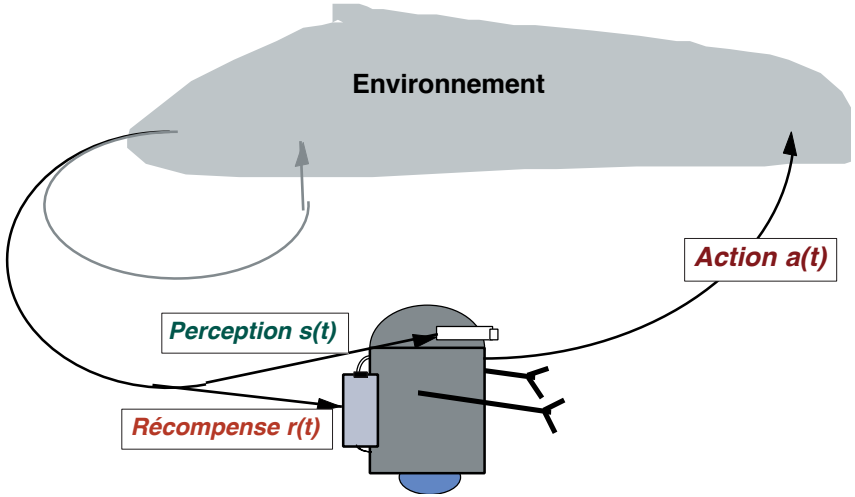
- Apprentissage « **prescriptif** » (recherche de ***causalités***)

1. J'observe que les gens qui mangent des glaces
   sont souvent en maillot de bain
2. Je voudrais vendre davantage de glaces

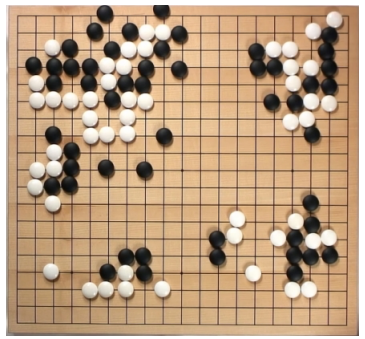
→ Je demande aux gens de se mettre en maillot de bain

Apprentissage « par **renforcement** »

- comment (ré)agir



1. Piloter un hélicoptère
2. Apprendre à jouer au tennis de table
3. Battre le champion de back-gammon (1992), de Go (2016)
4. Gérer un porte-feuille d'investissements
5. Contrôler une ferme



Plan

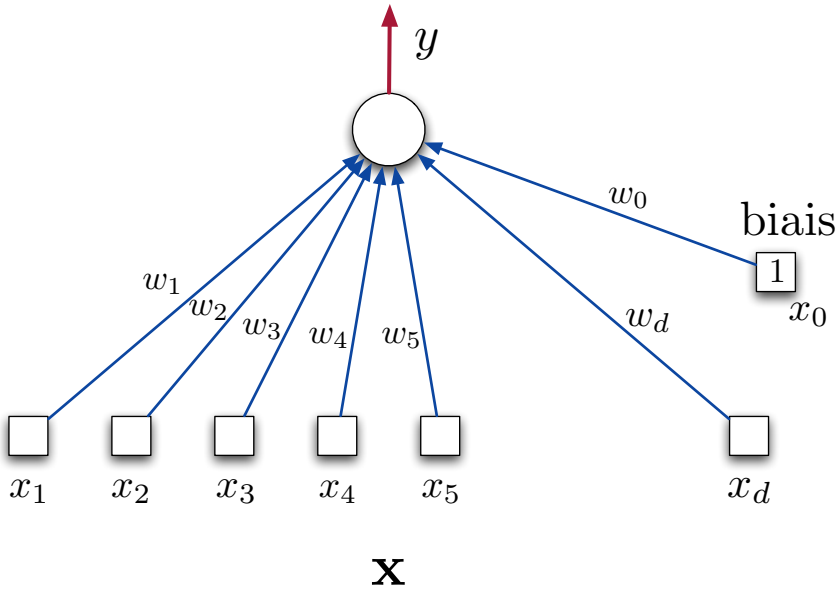
1. Pourquoi toute cette excitation ?
2. Grands types d'apprentissage
3. Apprentissage prédictif par réseaux de neurones
4. Quelles garanties ?
5. Le no-free-lunch theorem
6. Les réseaux de neurones profonds
7. Ce que l'on sait faire et les défis à relever

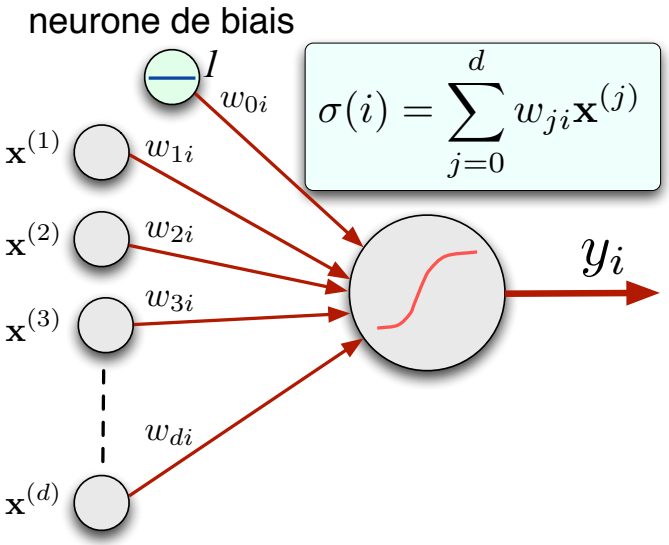
Où l'on va illustrer comment ça marche

Le cas du perceptron

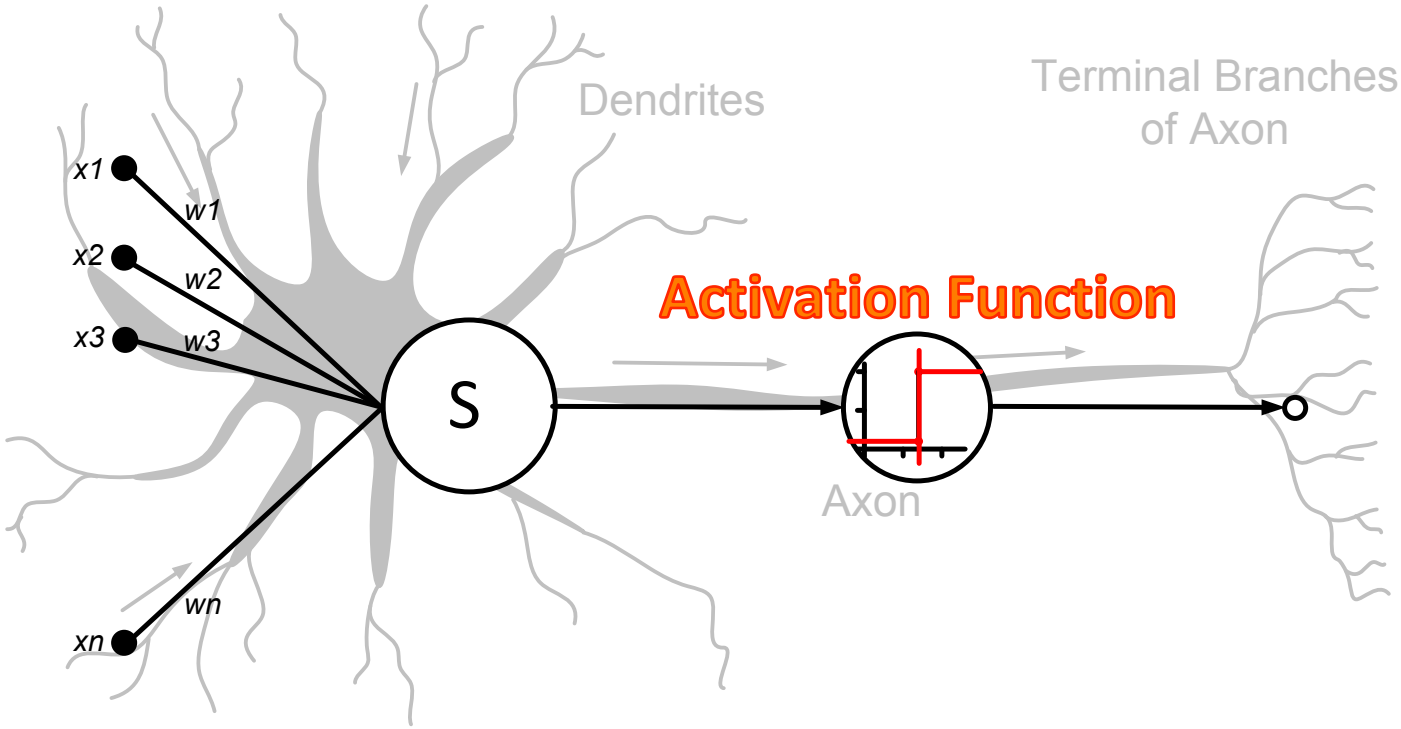
Le perceptron

– Rosenblatt (1958-1962)

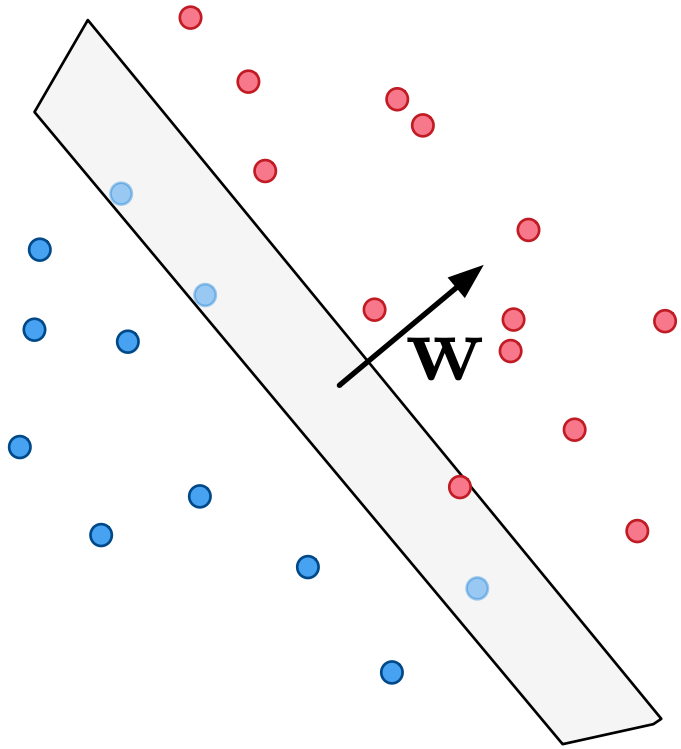




Artificial Neural Networks (ANN)



Le perceptron : un discriminant linéaire



Le perceptron

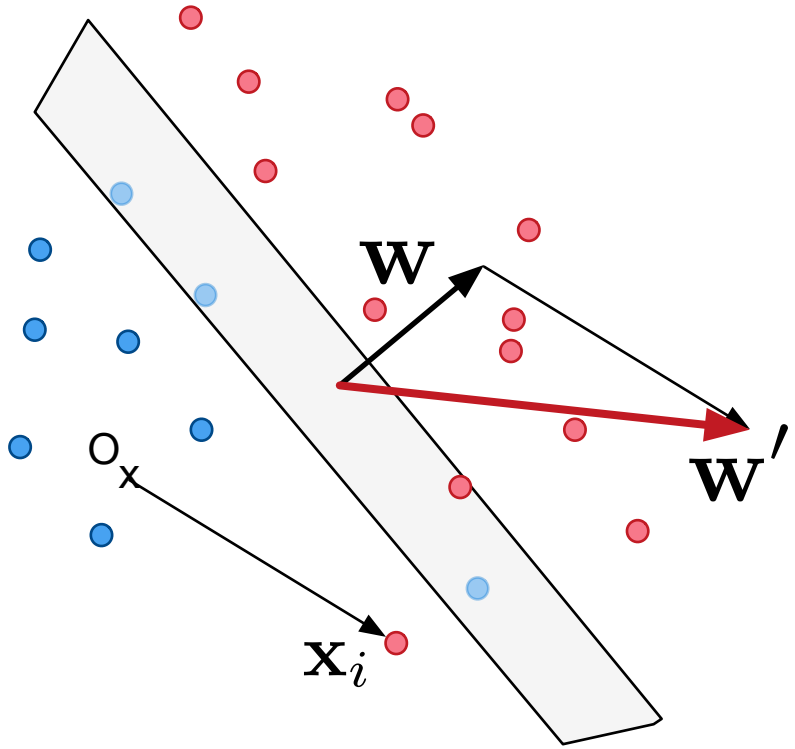
- **Apprentissage des poids** w_i
  - Principe (*règle de Hebb*) : en cas de succès, ajouter à chaque connexion quelque chose de proportionnel à l'entrée et à la sortie

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
    $\mathbf{w}(t+1) = \mathbf{w}(t) \stackrel{+}{-} \eta\, \mathbf{x}_i\, y_i$
  fin
  Passer à la forme d'apprentissage suivante
fin

The perceptron learning algorithm: intuition



$$\mathbf{w}' = \mathbf{w} + \eta\, y_i\, \mathbf{x}_i$$

Des propriétés remarquables !!

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

!!!

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

Garantie de généralisation ??

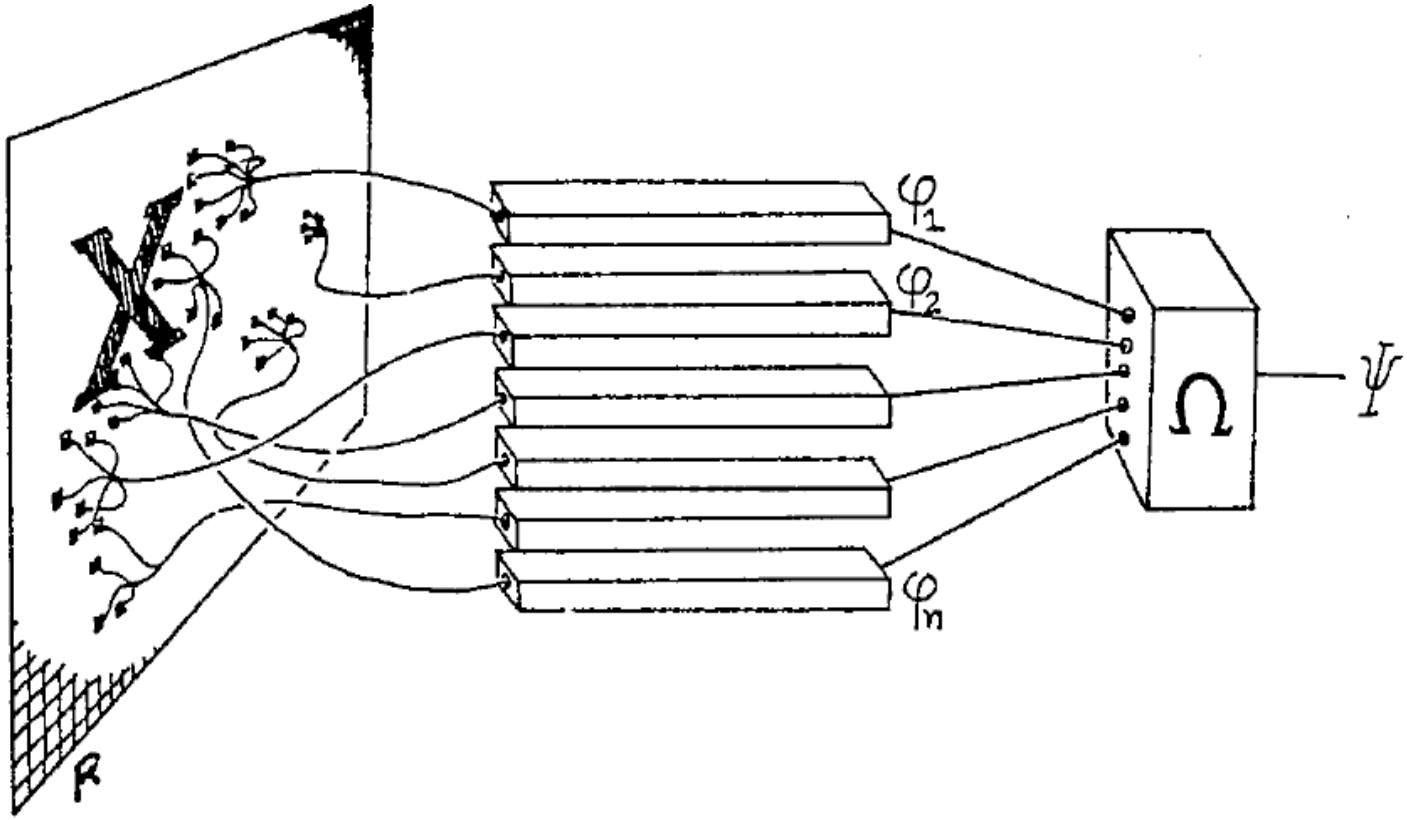
- Théorèmes sur la performance

  par rapport à l'échantillon d'apprentissage

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

Le Perceptron

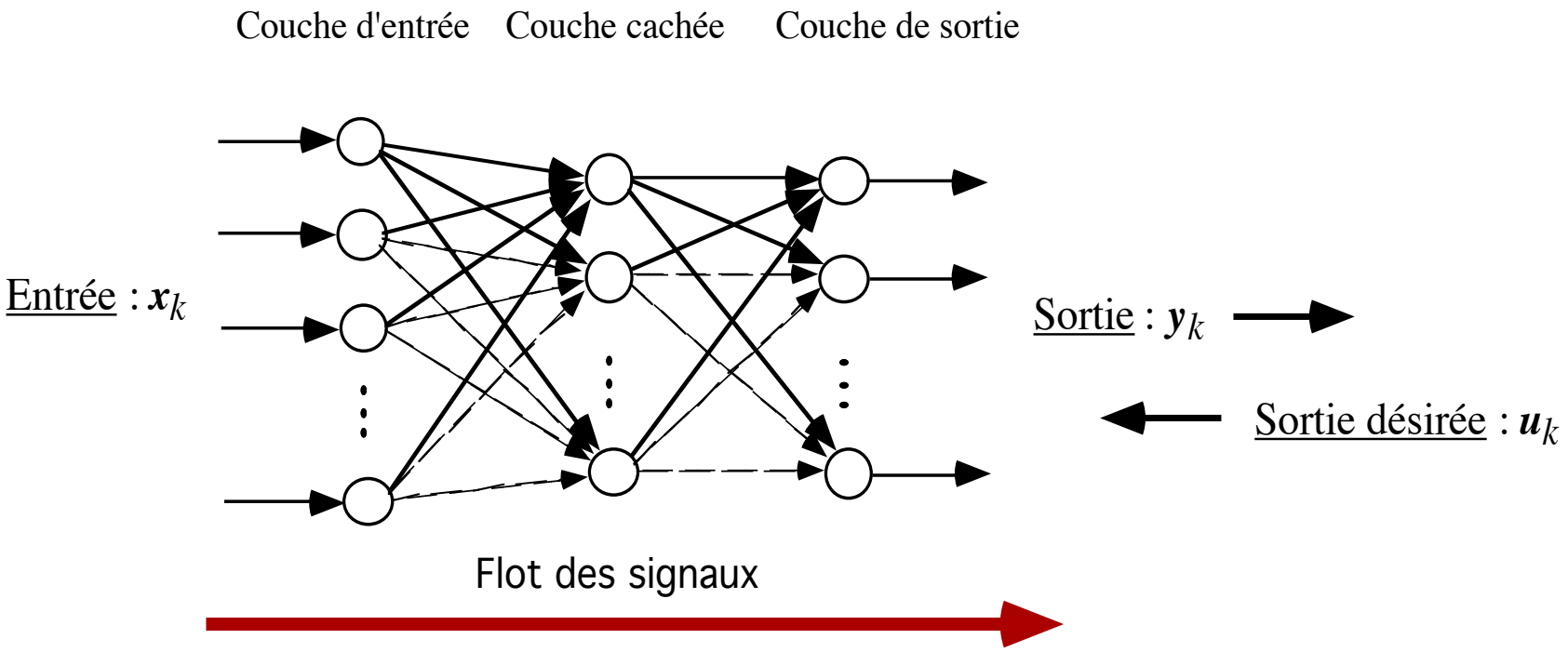
- Rosenblatt (1958-1962)



Les perceptrons multi-couches

Le perceptron multi-couche

- Topologie typique



Le Perceptron Multi-Couches : propagation

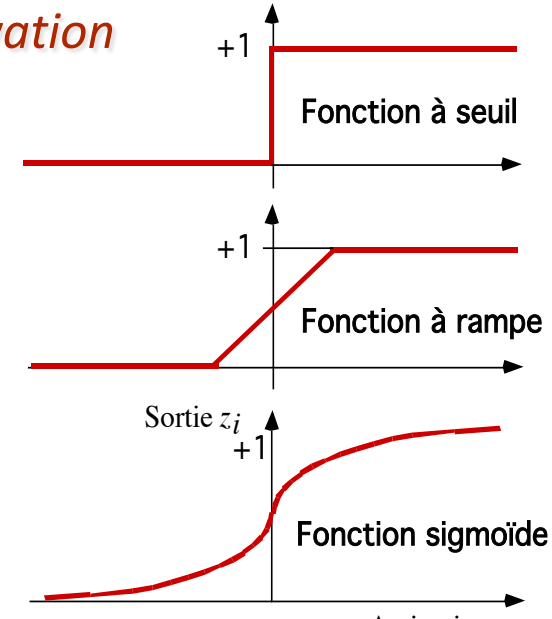
- Pour chaque neurone :

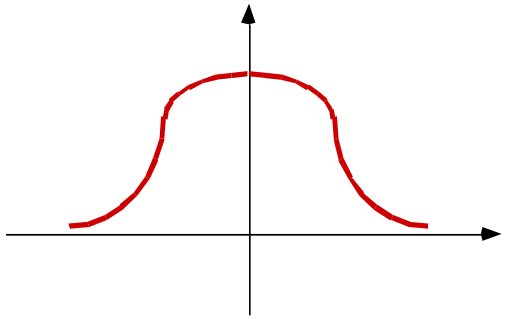
$$y_l = g\left(\sum_{j=0,d} w_{jk}\ \phi_j\right) = g(a_k)$$

  - w_{jk} : *poids* de la connexion de la cellule *j* à la cellule *k*
  - a_k : *activation* de la cellule *k*
  - g : *fonction d'activation*

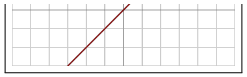
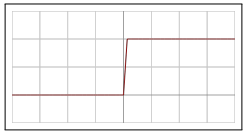
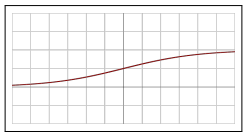
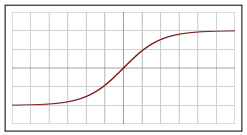
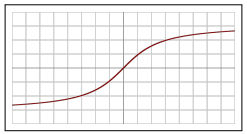
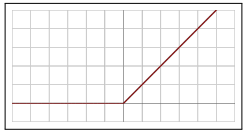
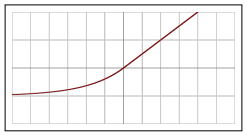
$$g(a) = \frac{1}{1 + e^{-a}}$$

g'(a) = g(a)(1-g(a))

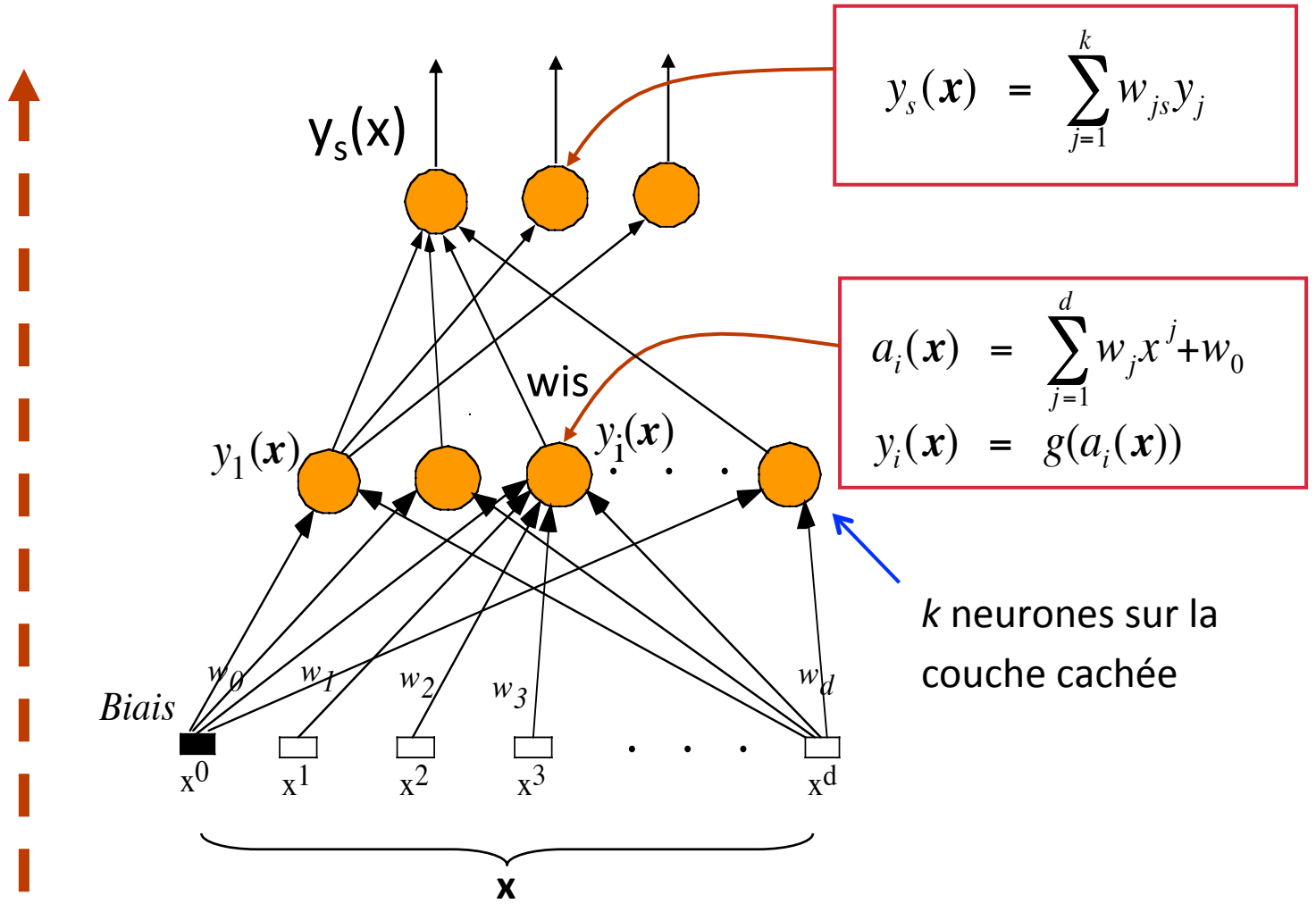




Fonction à base radiale

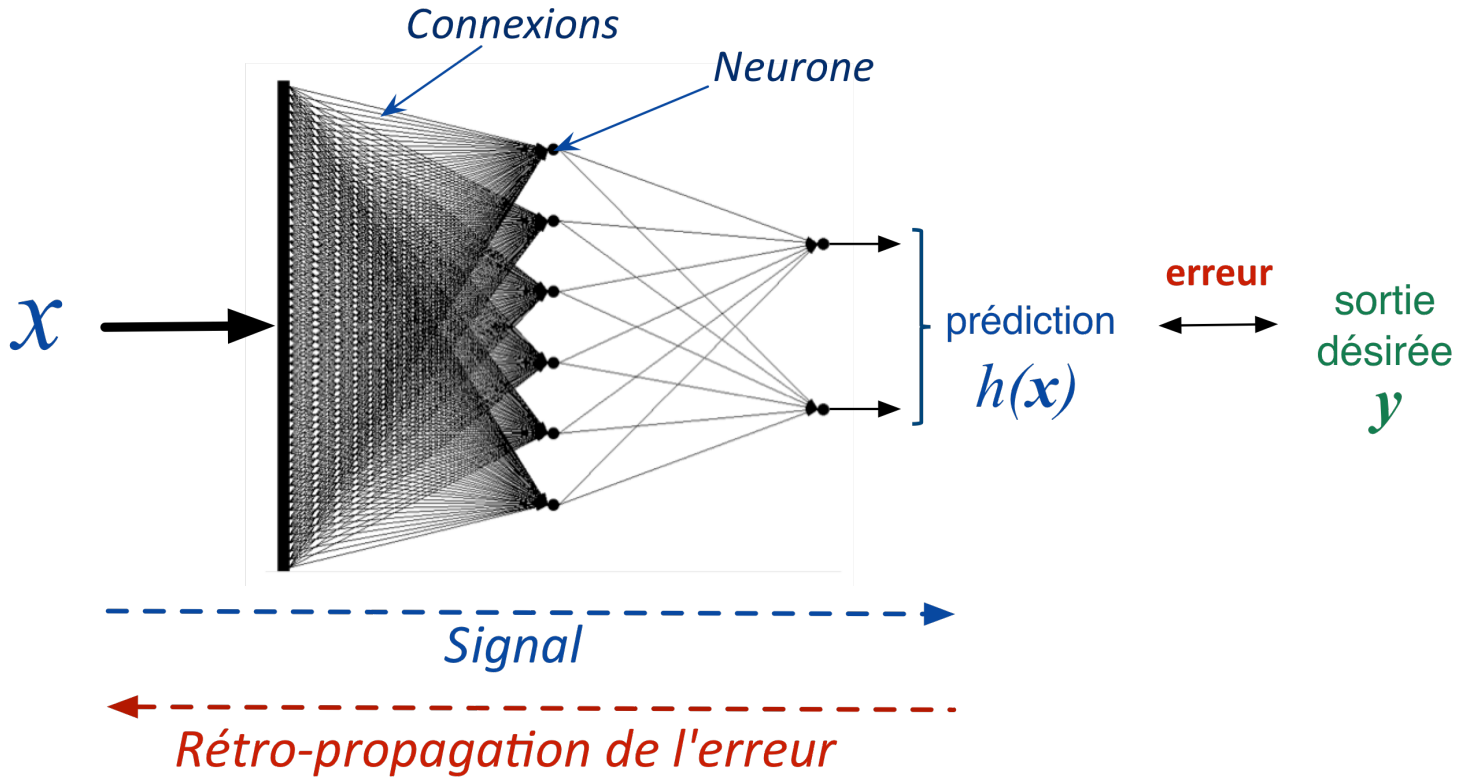
Nom	Graphe	Équation	Dérivée
Rampe		$f(x) = x$	$f'(x) = 1$
Heaviside		$f(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1 & \text{si } x \geq 0 \end{cases}$	$f'(x) = \begin{cases} 0 & \text{si } x \neq 0 \\ ? & \text{si } x = 0 \end{cases}$
Logistique ou sigmoïde		$f(x) = \dfrac{1}{1+e^{-x}}$	$f'(x) = f(x)\left(1 - f(x)\right)$
Tangente hyperbolique		$f(x) = \tanh(x) = \dfrac{2}{1+e^{-2x}} - 1$	$f'(x) = 1 - f^2(x)$
Arc Tangente		$f(x) = \tan^{-1}(x)$	$f'(x) = \dfrac{1}{x^2+1}$
Unité ReLU		$f(x) = \begin{cases} 0 & \text{si } x < 0 \\ x & \text{si } x \geq 0 \end{cases}$	$f'(x) = \begin{cases} 0 & \text{si } x \neq 0 \\ 1 & \text{si } x = 0 \end{cases}$
Unité Exponentielle Linéaire		$f(x) = \begin{cases} \alpha\left(e^x - 1\right) & \text{si } x < 0 \\ x & \text{si } x \geq 0 \end{cases}$	$f'(x) = \begin{cases} f(x) + \alpha & \text{si } x < 0 \\ 1 & \text{si } x \geq 0 \end{cases}$

Le PMC : passes avant et arrière (résumé)



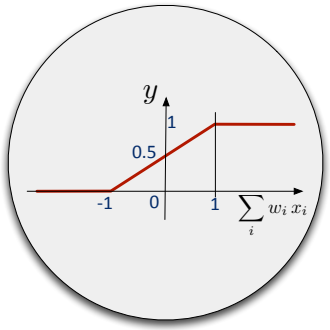
Learning with Multi-Layer Perceptrons

- Questions:
  - How to learn the **parameters** (weights of the connections) ?
  - How to set the **architecture** of the network?



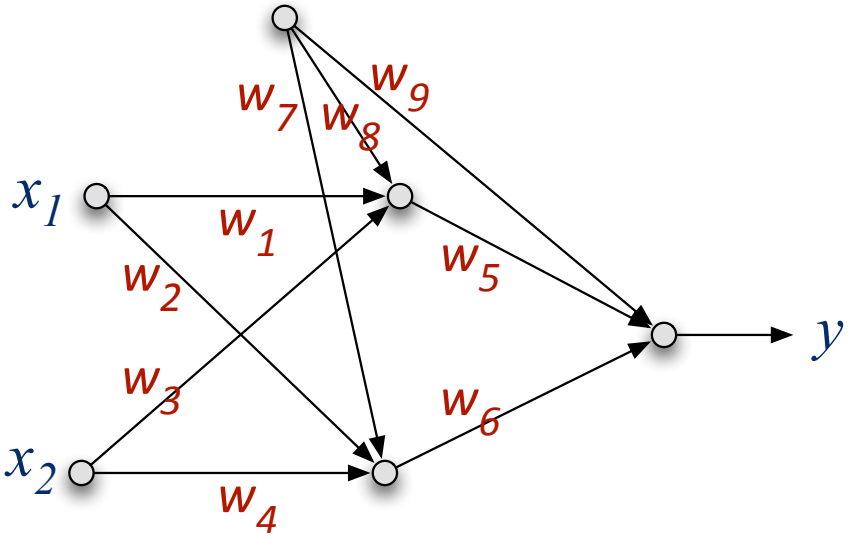
Compute the weights such that …

Calcul de la fonction XOR



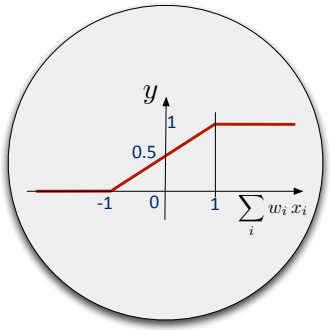
Entrée x_1	Entrée x_2	Sortie y
0	0	0
0	1	1
1	0	1
1	1	0

W_1 =
W_2 =
W_3 =
W_4 =
W_5 =
W_6 =
W_7 =
W_8 =
W_9 =



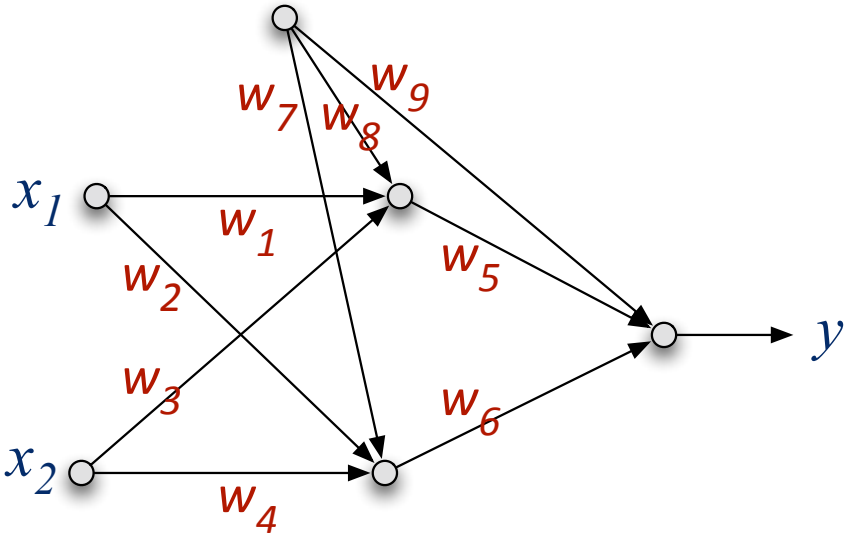
Compute the weights such that …

Calcul de la fonction XOR

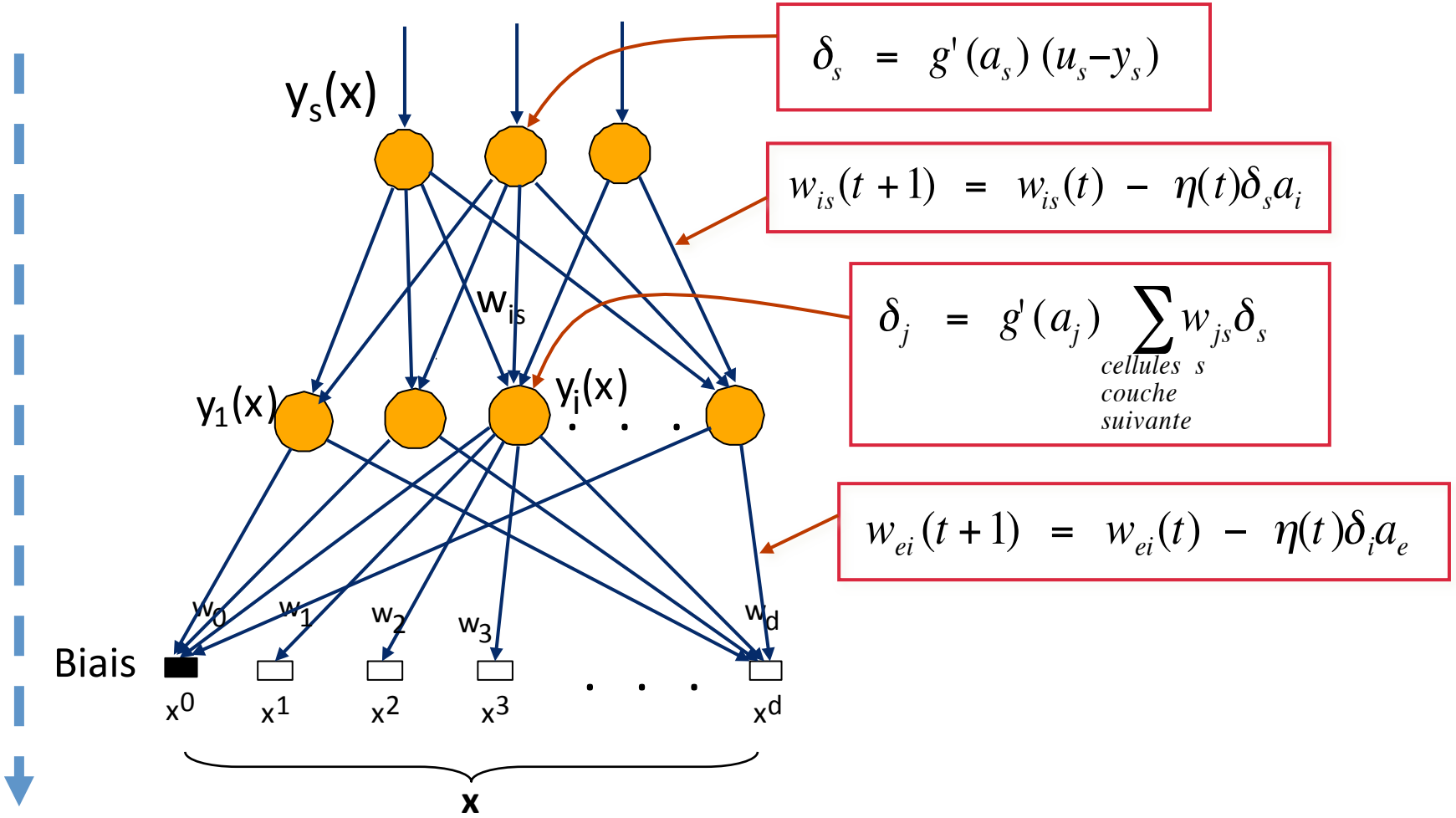


Entrée x_1	Entrée x_2	Sortie y
0	0	0
0	1	1
1	0	1
1	1	0

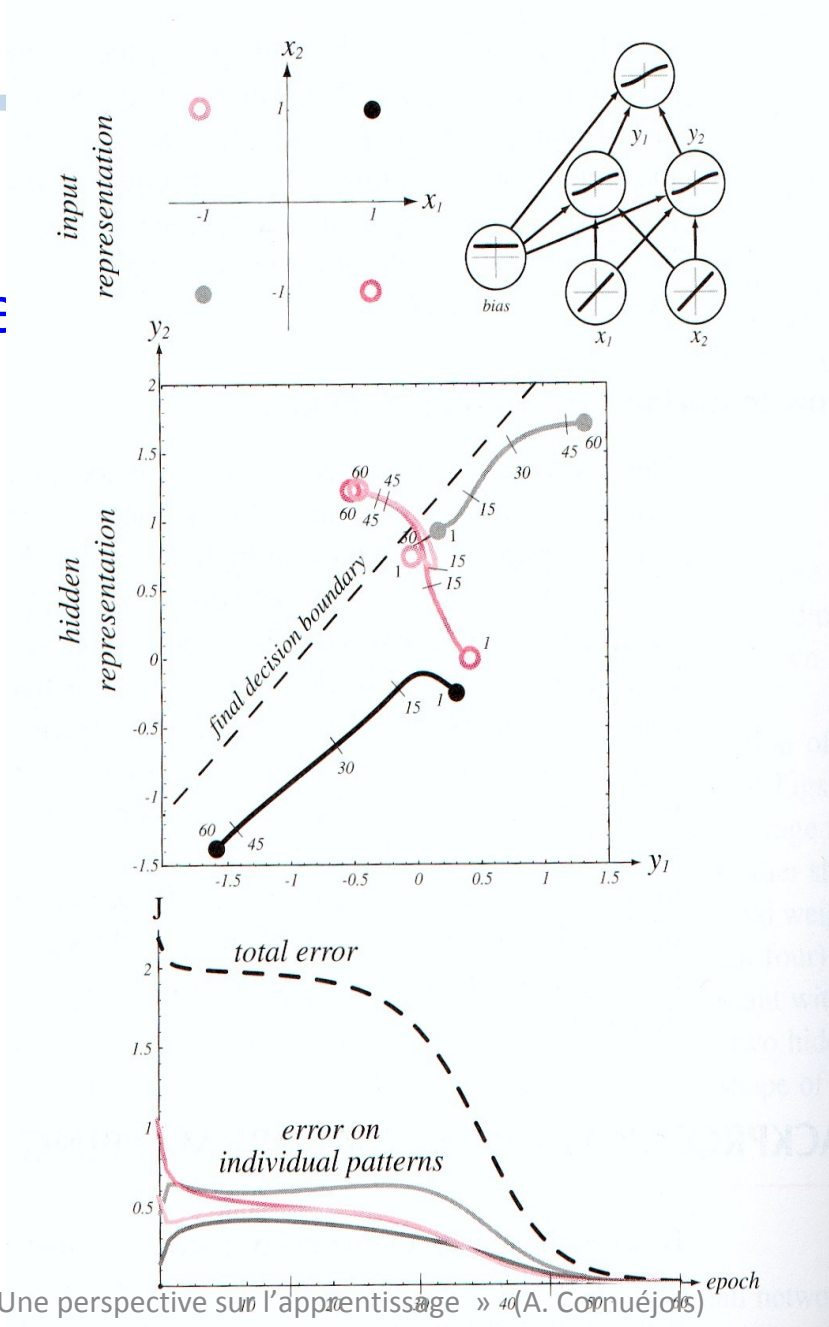
$W_1 = 1$
$W_2 = -1$
$W_3 = -1$
$W_4 = 1$
$W_5 = 4$
$W_6 = 4$
$W_7 = -1$
$W_8 = -1$
$W_9 = -1$



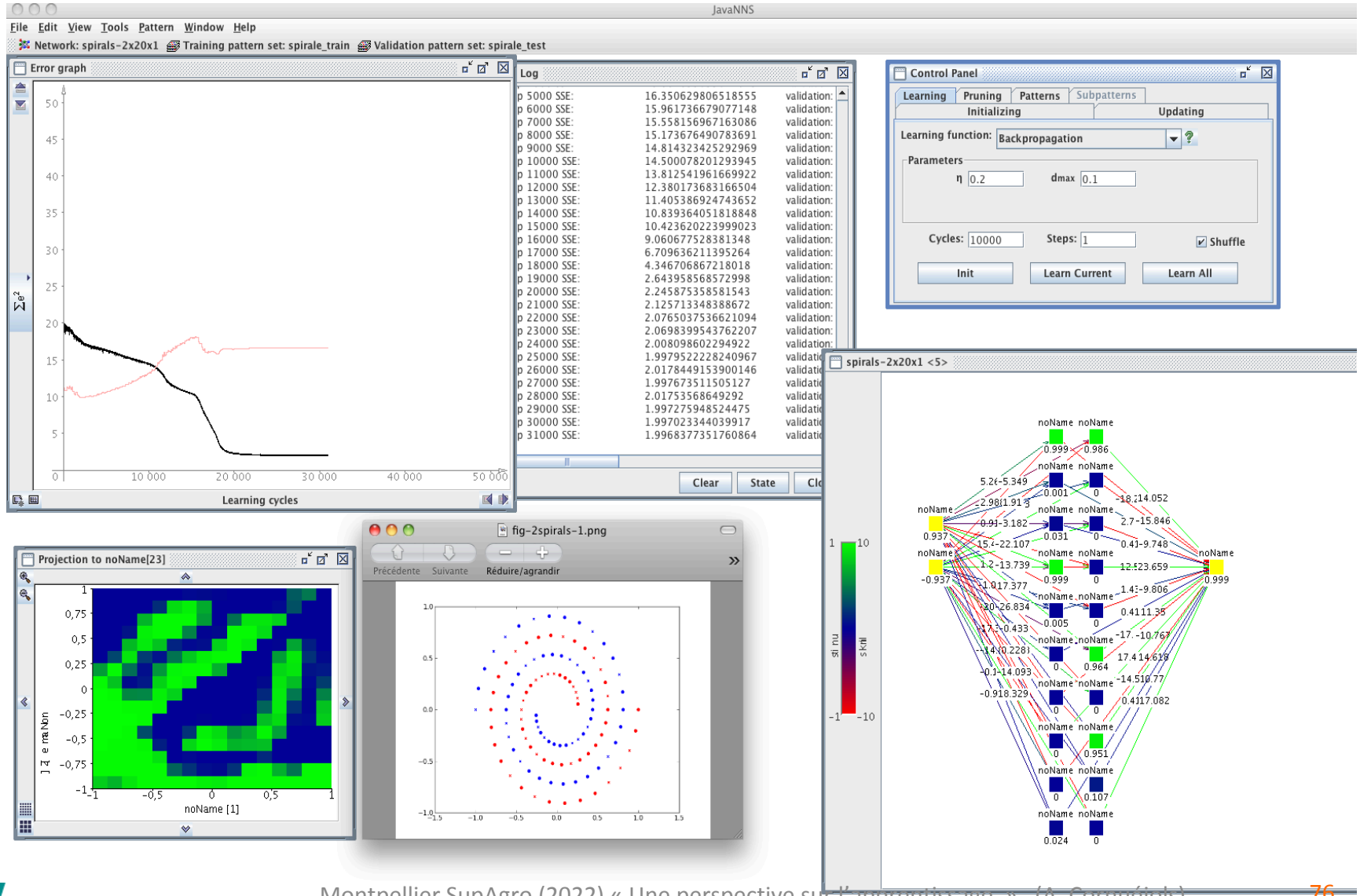
Le PMC : passes avant et arrière (résumé)



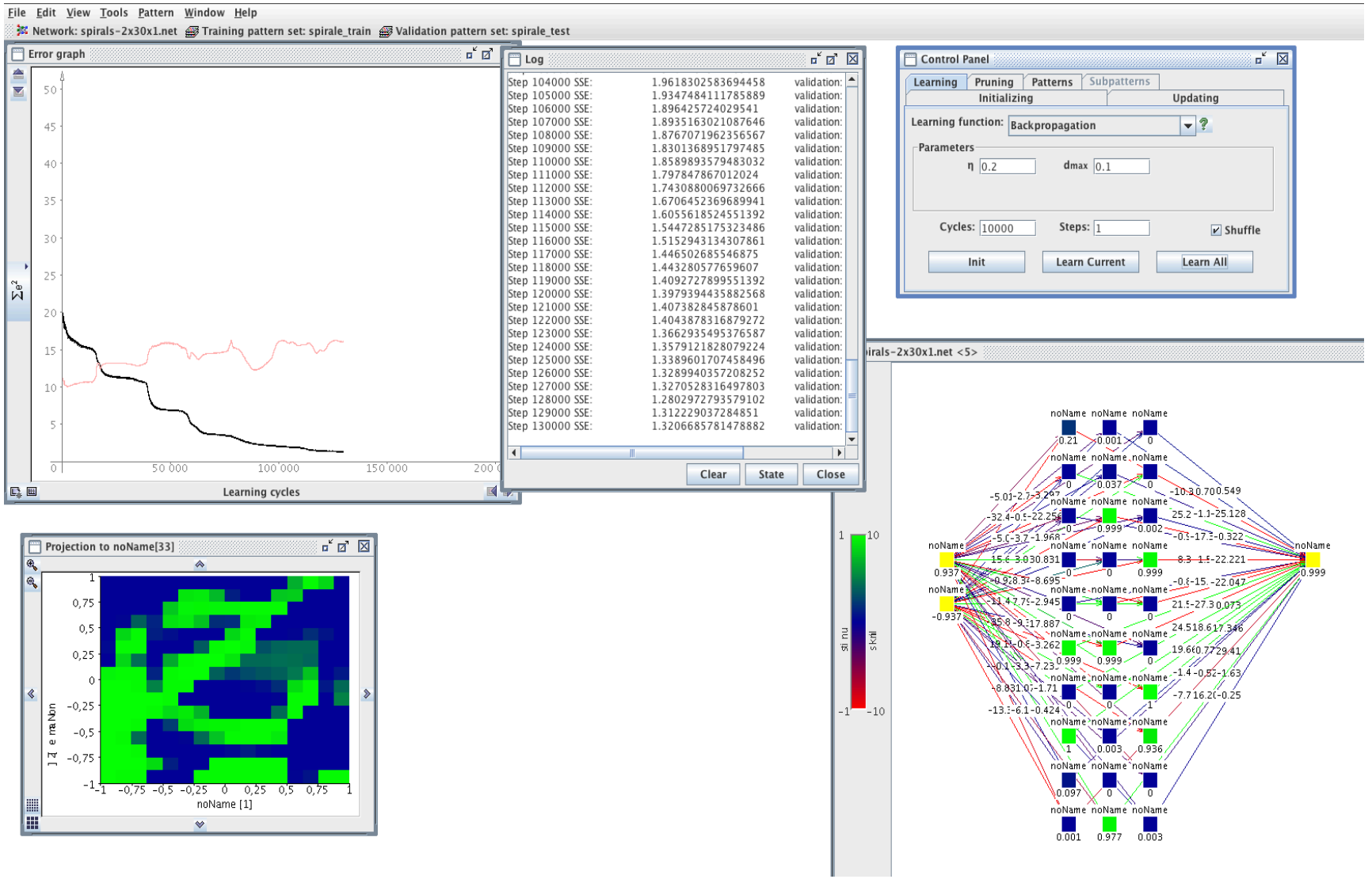
Rôle de la couche cachée



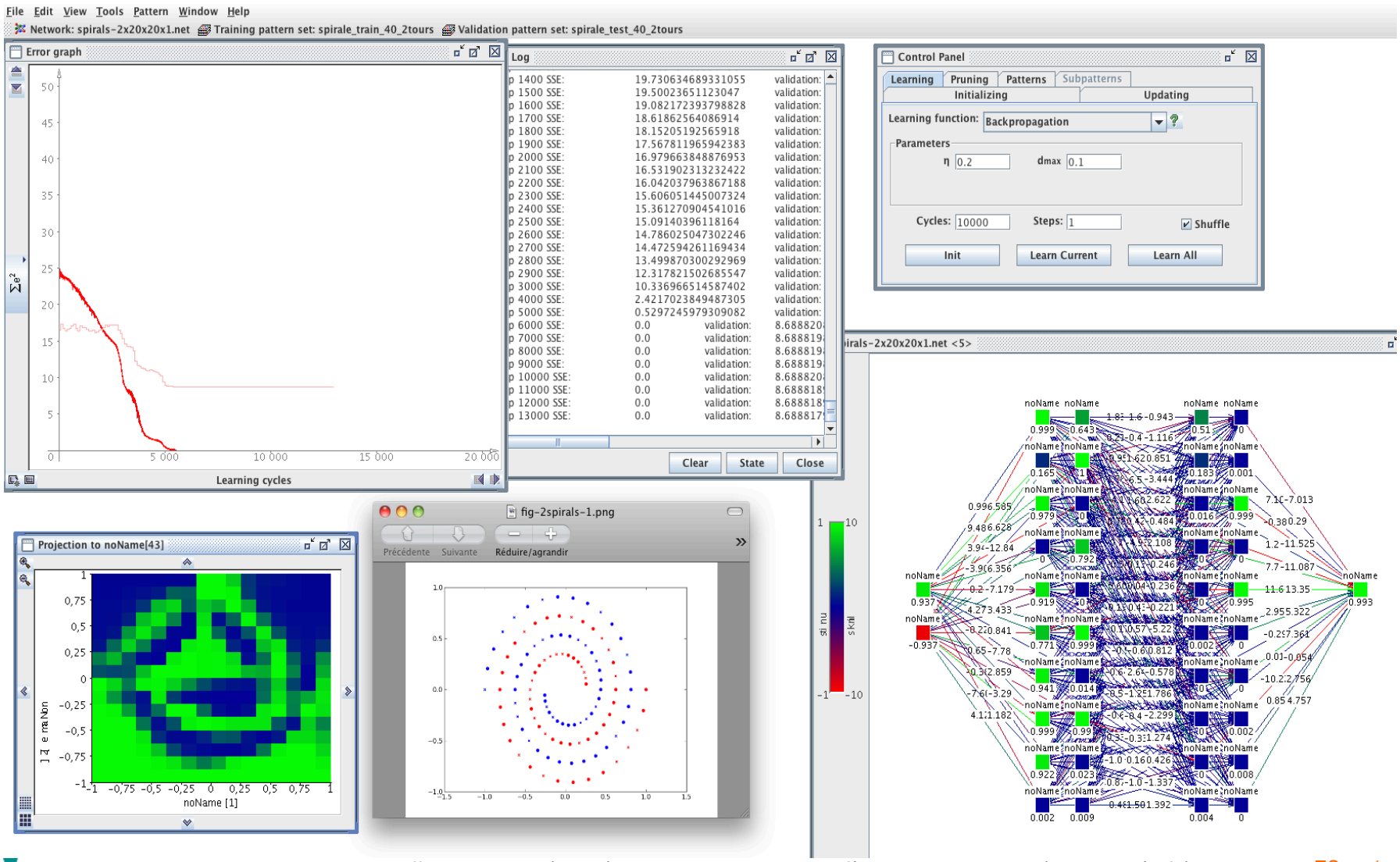
Illustration



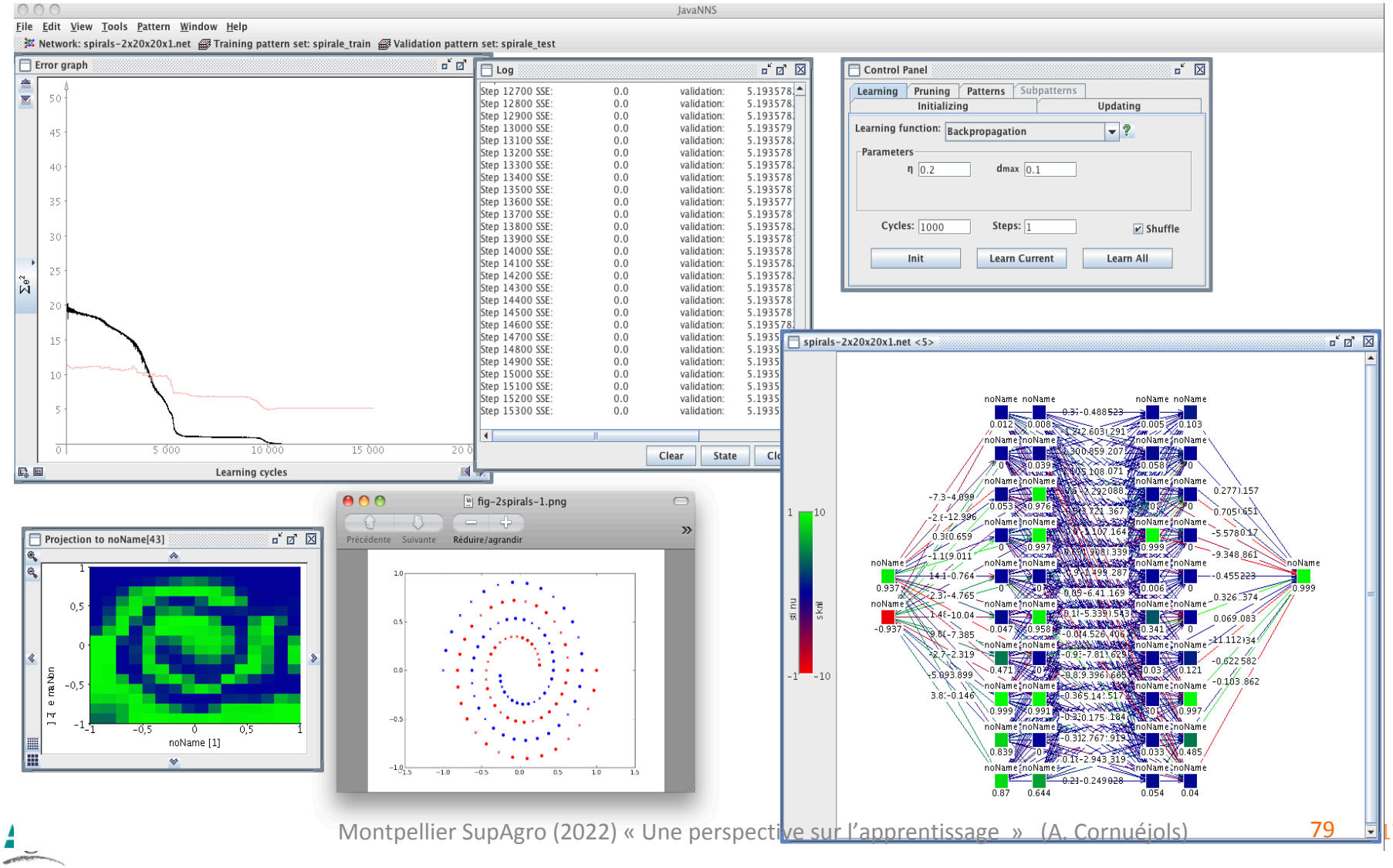
Illustration



Illustration

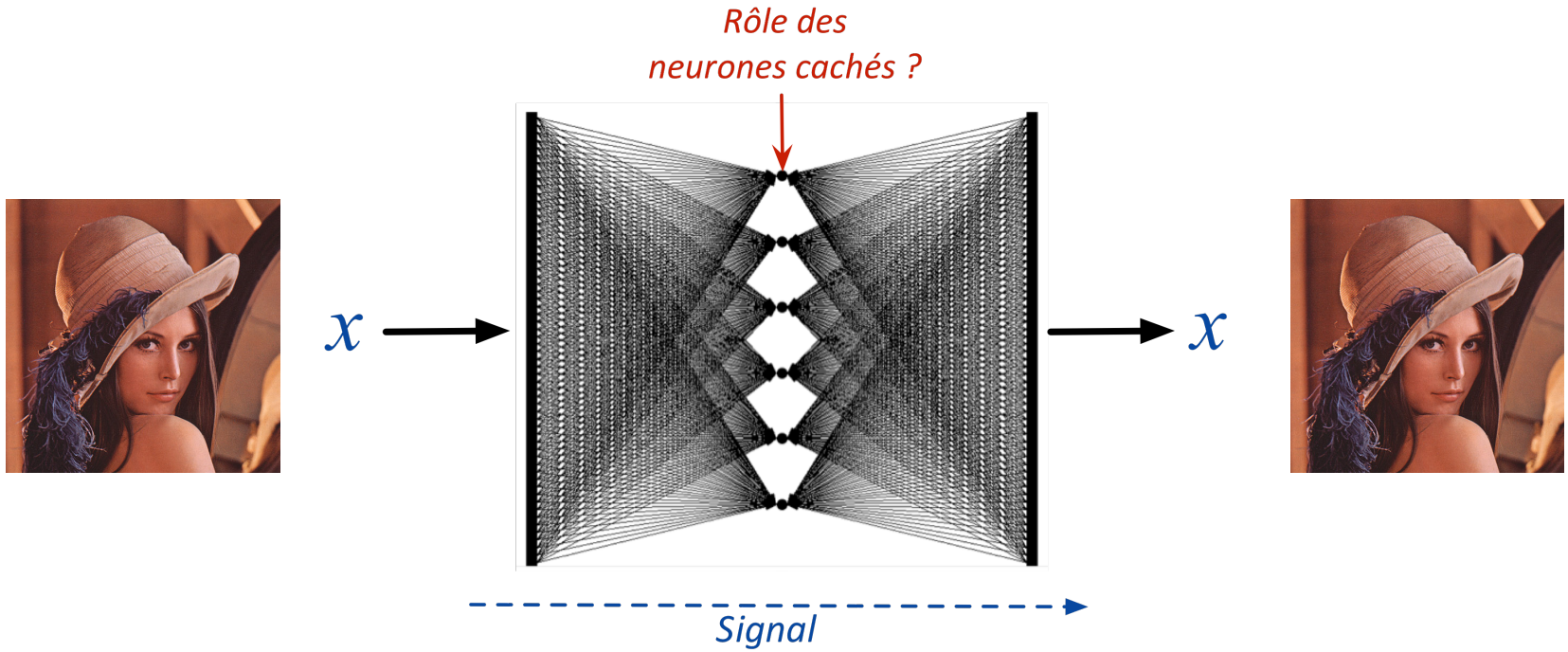


Illustration



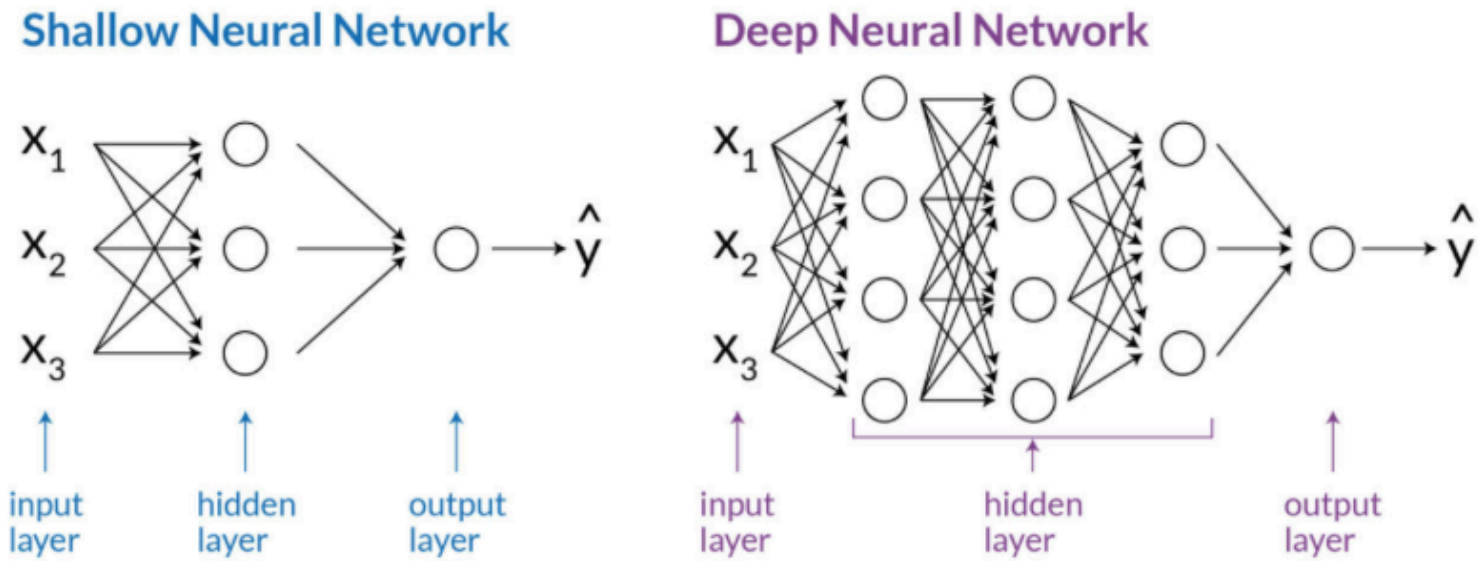
Change of representation: the role of hidden layer(s)

- **Which new representation** (latent variables)?
- How to choose the architecture?



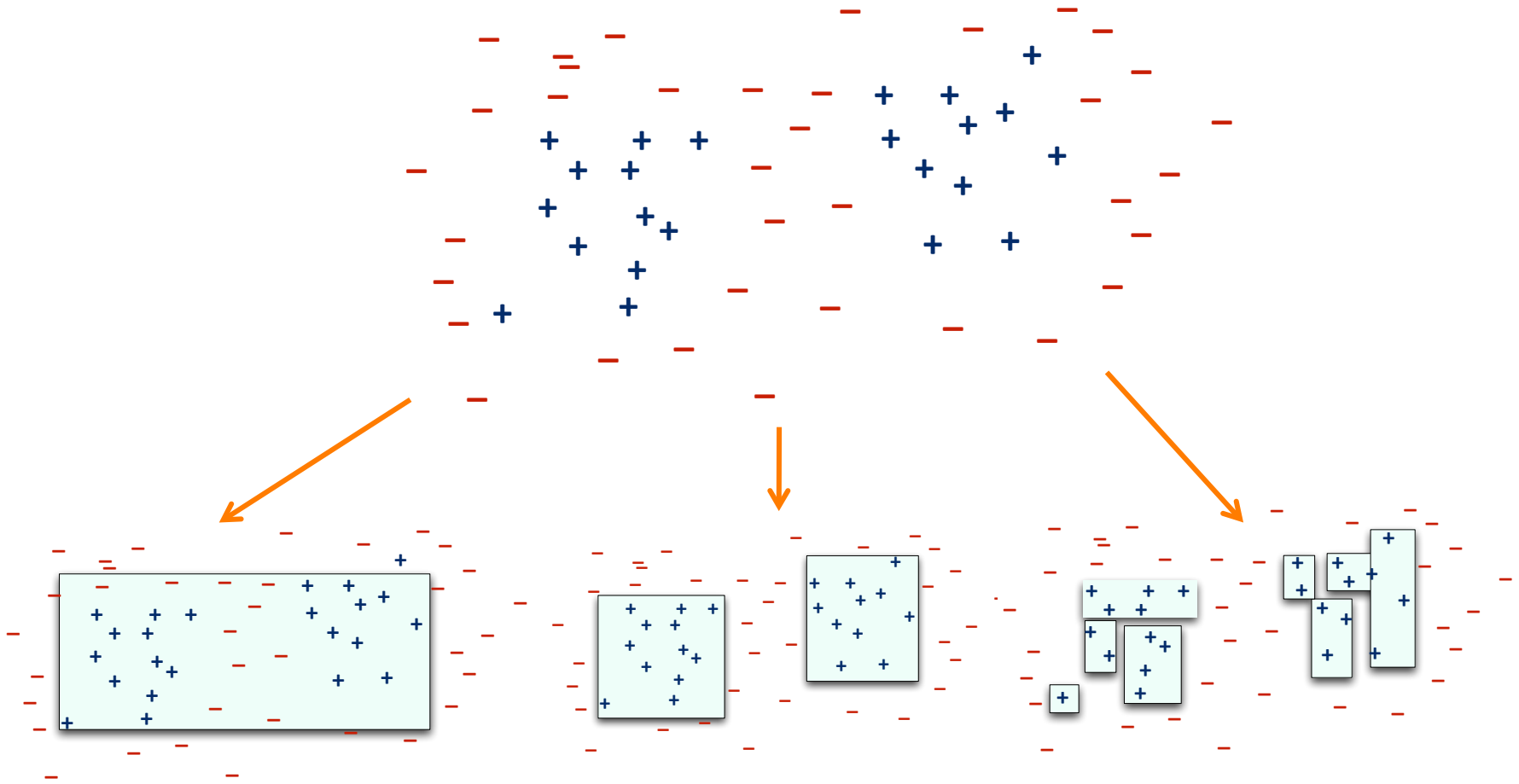
Comment choisir l'architecture du RN ?

- Contrôle la performance en généralisation



Shallow and Deep Neural Networks.

Illustration

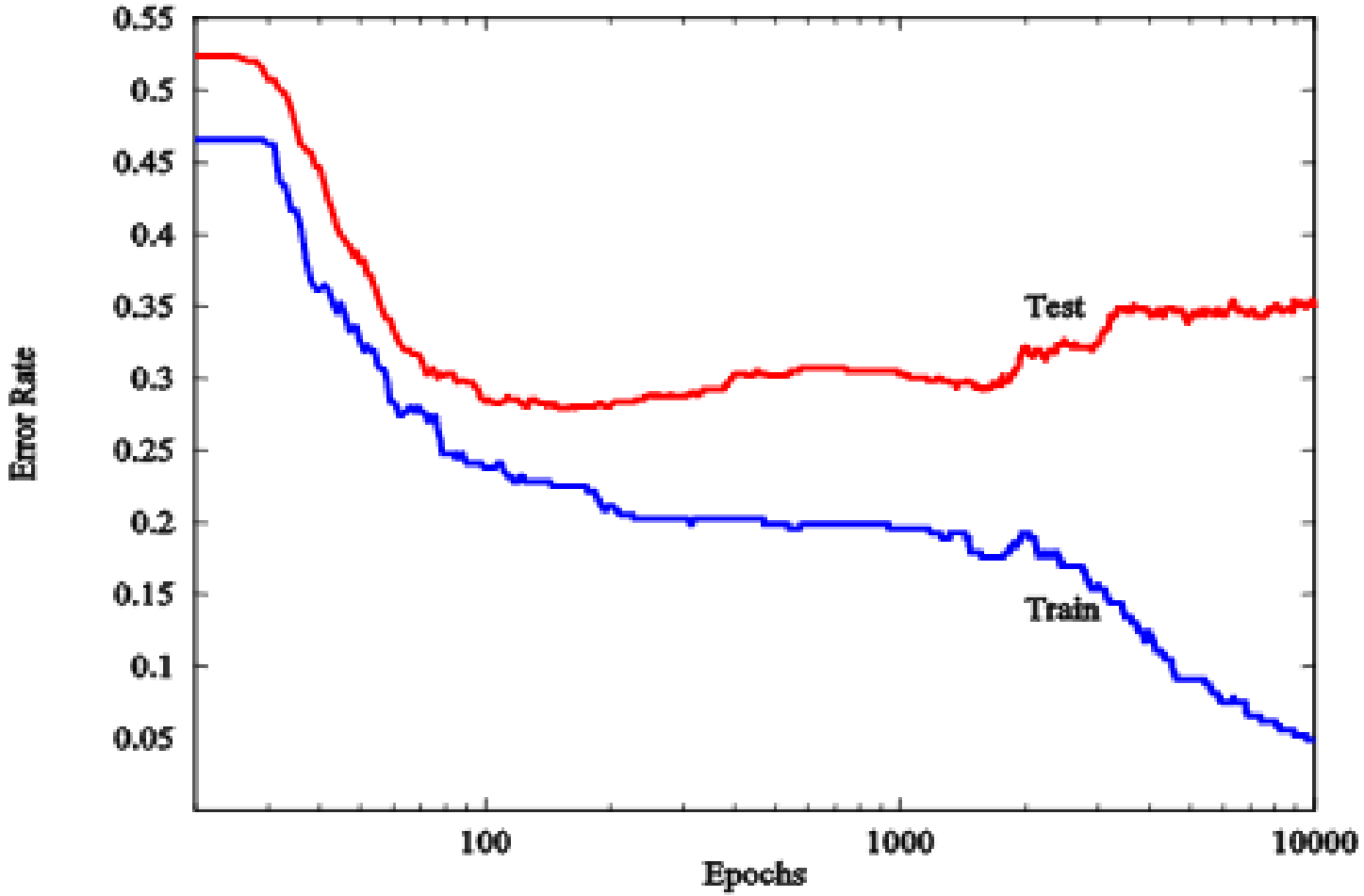


Sous-apprentissage

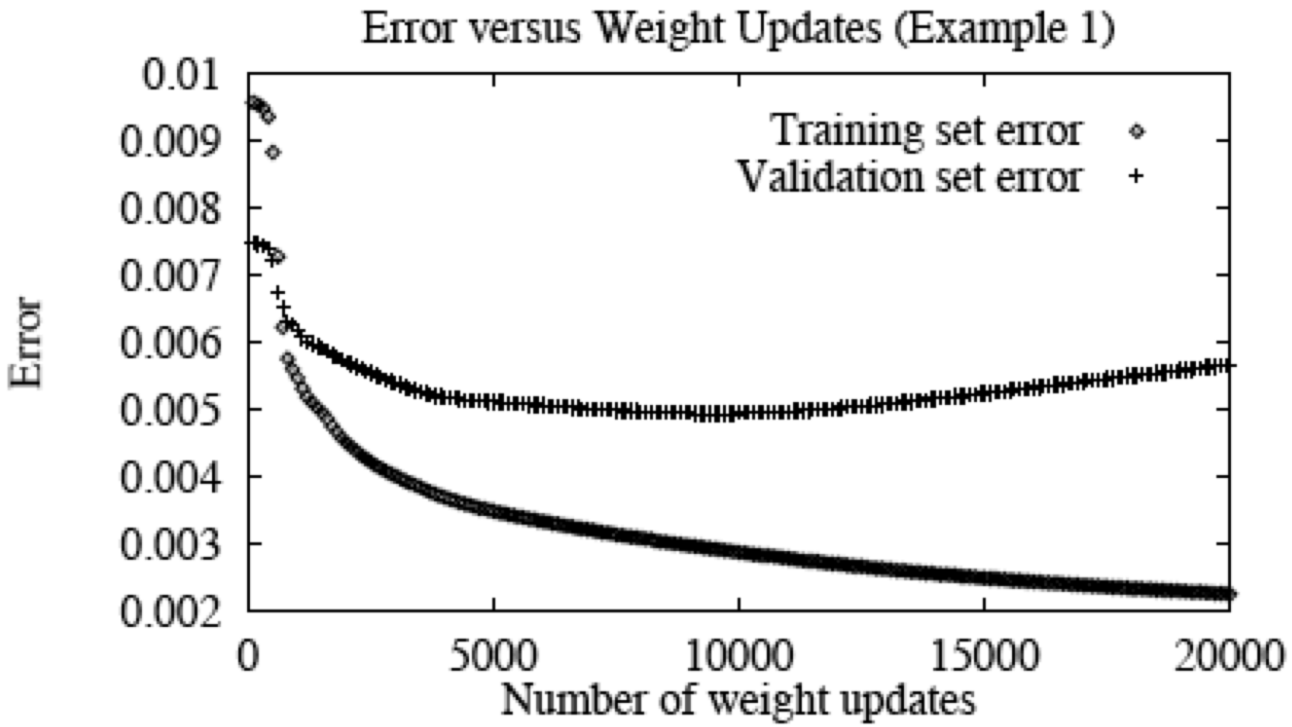
Modèle « correct »

Sur-apprentissage

Sur-apprentissage



Sur-apprentissage (RN)



- Courbes pour 1 000 exemples

- *Quelles courbes si on avait 2 000 exemples ?*

Les trois ingrédients

de l'apprentissage artificiel

Trois ingrédients

1. Le choix de l'**espace des hypothèses** H
   - Généralement H ≠ F

2. Le **critère inductif**
   - Comment évaluer chaque hypothèse en fonction de S

3. La méthode d'**exploration** de H
   - Comment trouver une bonne (optimale ?) hypothèse

Plan

1. Pourquoi toute cette excitation ?
2. Grands types d'apprentissage
3. Apprentissage prédictif par réseaux de neurones
4. Quelles garanties ?
5. Le no-free-lunch theorem
6. Les réseaux de neurones profonds
7. Ce que l'on sait faire et les défis à relever

Il est temps de se poser la question

Mot d'ordre : **pas de pensée magique !**

Comment fonder l'induction ?

Illustration : apprendre à **classer des exemples**

- Comment faire ?

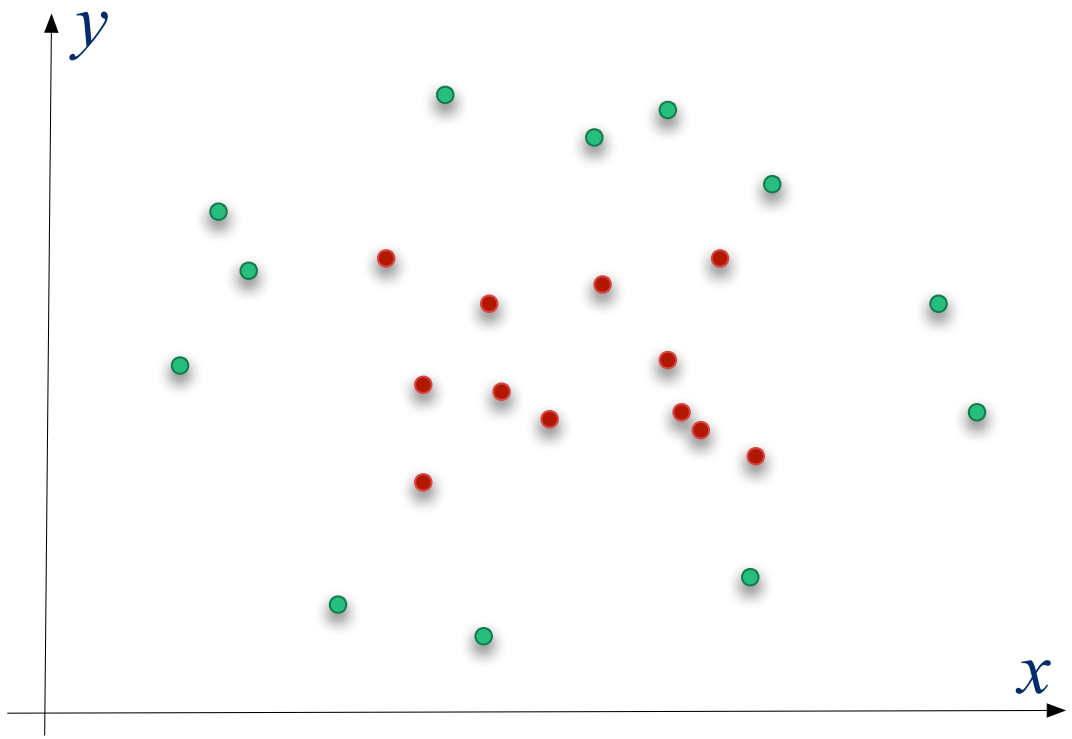


Illustration : apprendre à **classer des exemples**

- Comment faire ?

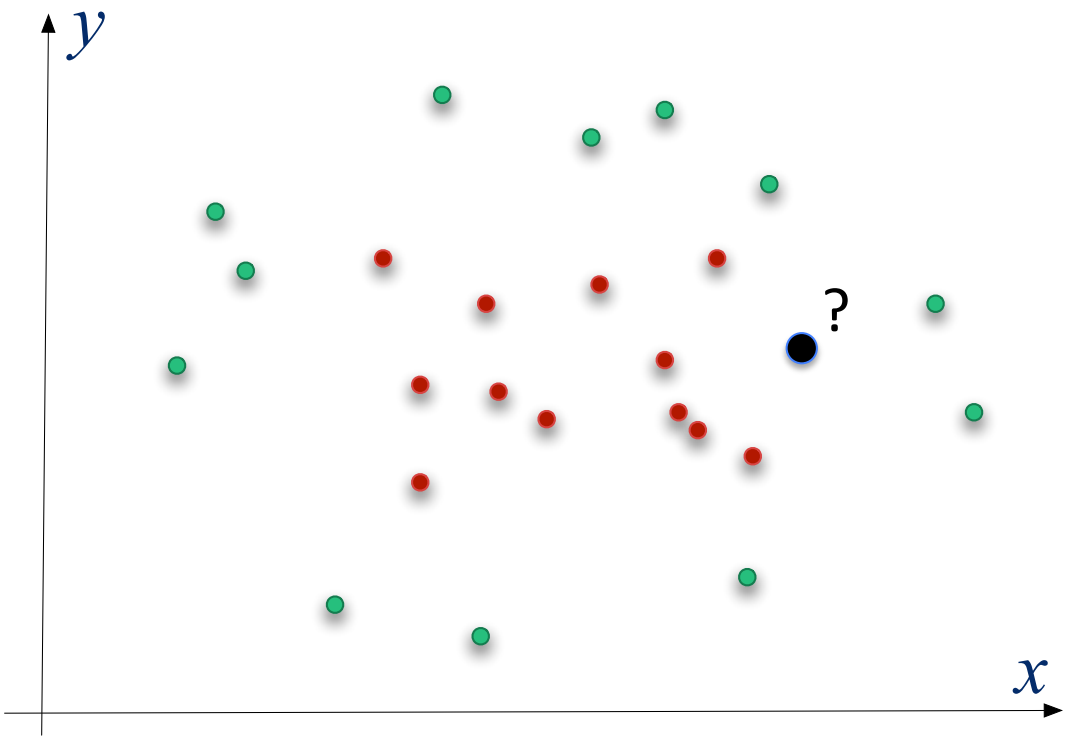
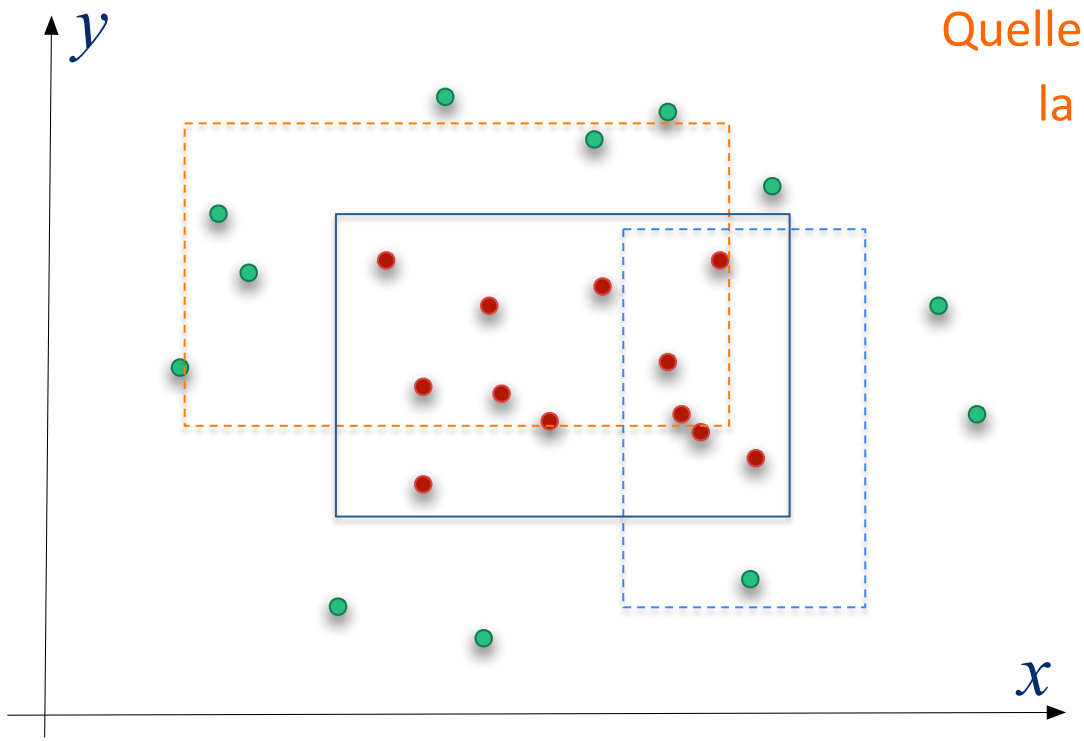


Illustration : apprendre à classer des exemples en 2D

- Comment faire ?



Quelle est l'hypothèse la meilleure ?

Quelle hypothèse choisir ?

Quelle qualité pour chaque hypothèse candidate ?

Le « critère inductif »

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 ?

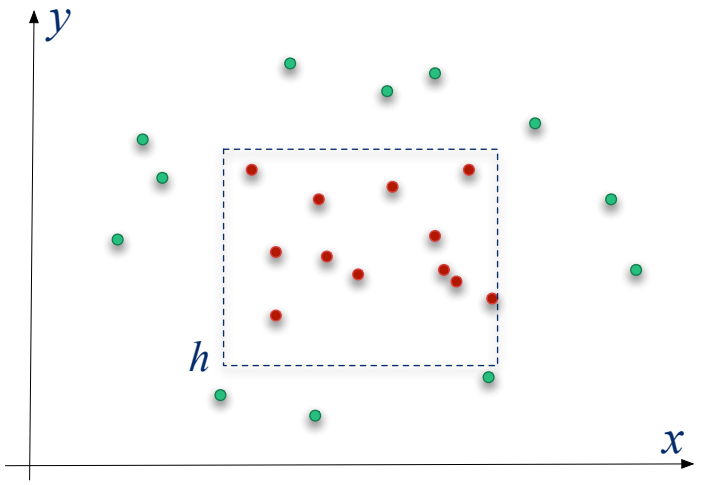
Comment trouver *h** (ou une bonne hypothèse) alors que l'on n'a accès qu'à un **échantillon d'apprentissage limité** ?

- 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 en 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		

Un exemple qui en 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		

Un exemple qui en 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		

Un exemple qui en 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		

Un exemple qui en 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		

Un exemple qui en 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		+

Un exemple qui en 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? $2^{10} = 1024$

Un exemple qui en dit beaucoup …

- Examples described using:

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

15

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		+
1 small green square		-
1 small red square		+
2 large green squares		+
2 small green squares		+
2 small red circles		+
1 small green circle		-
2 large green circles		-
2 small green circles		+
1 large red circle		-
2 large red squares	?	

How many remaining functions?

← ?

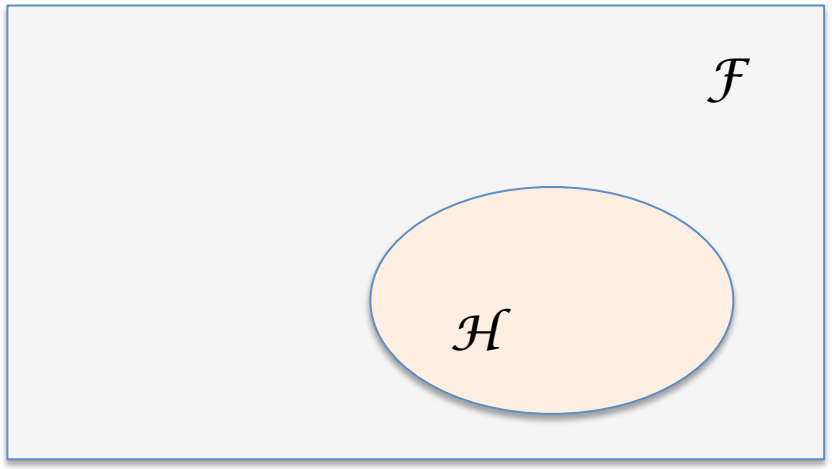
Induction: impossible de gagner ?

- **Un biais est nécessaire**

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

Induction: impossible de gagner ?

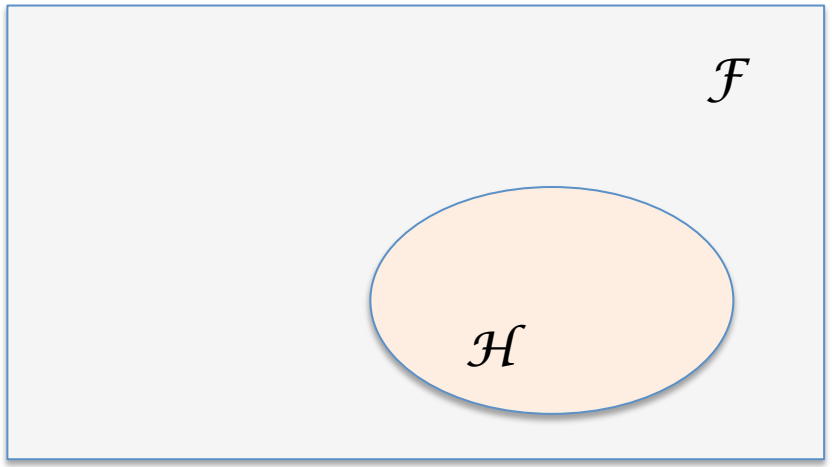
- **Un biais est nécessaire**



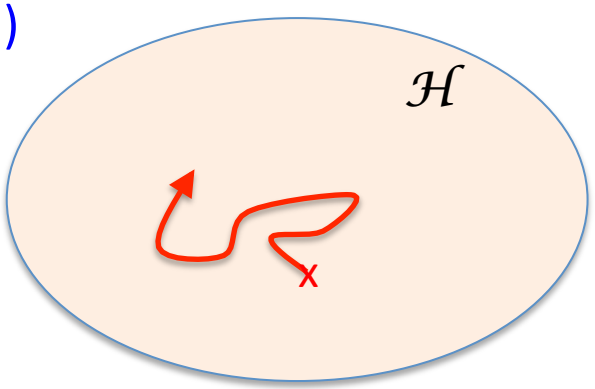
- **Types** de biais
  - **De représentation** (déclaratif)
  - **De recherche** (procédural)

Induction: impossible de gagner ?

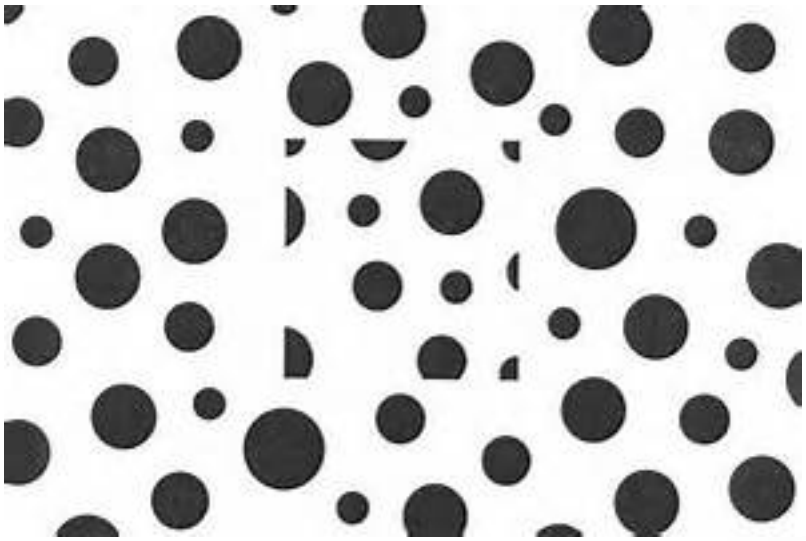
- **Un biais est nécessaire**

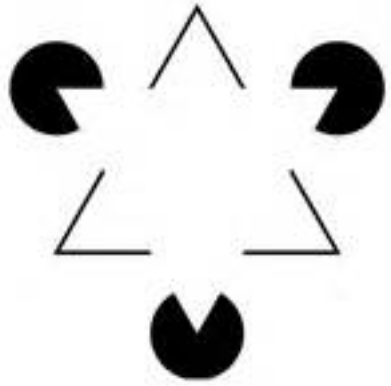


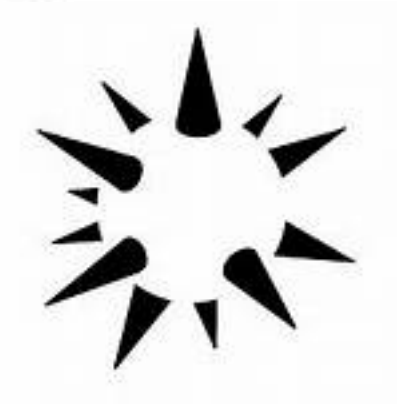
- **Types** de biais
  - **De représentation** (déclaratif)
  - **De recherche** (procédural)



Learning – completing data with necessary a aprioris

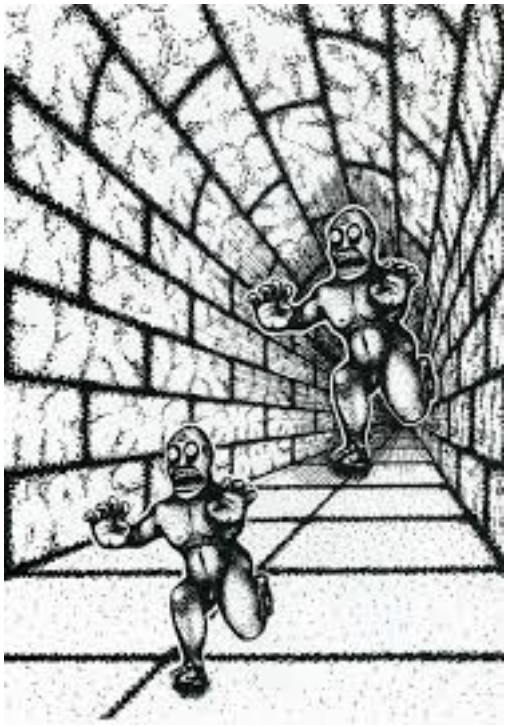


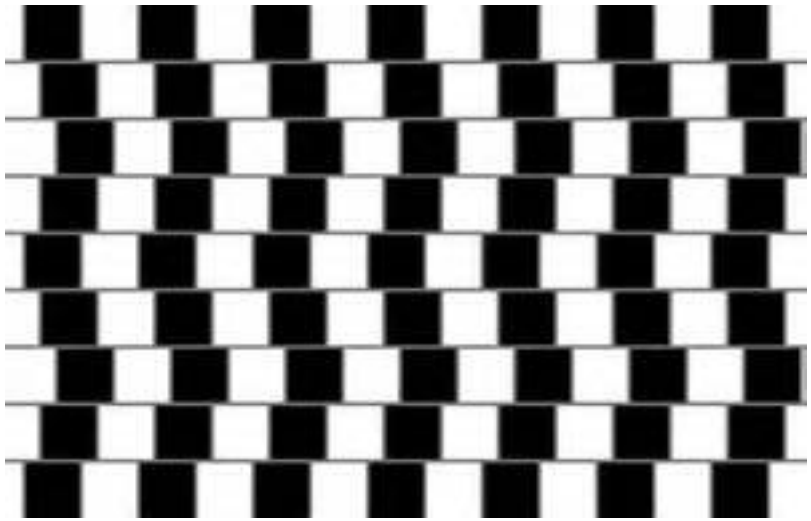




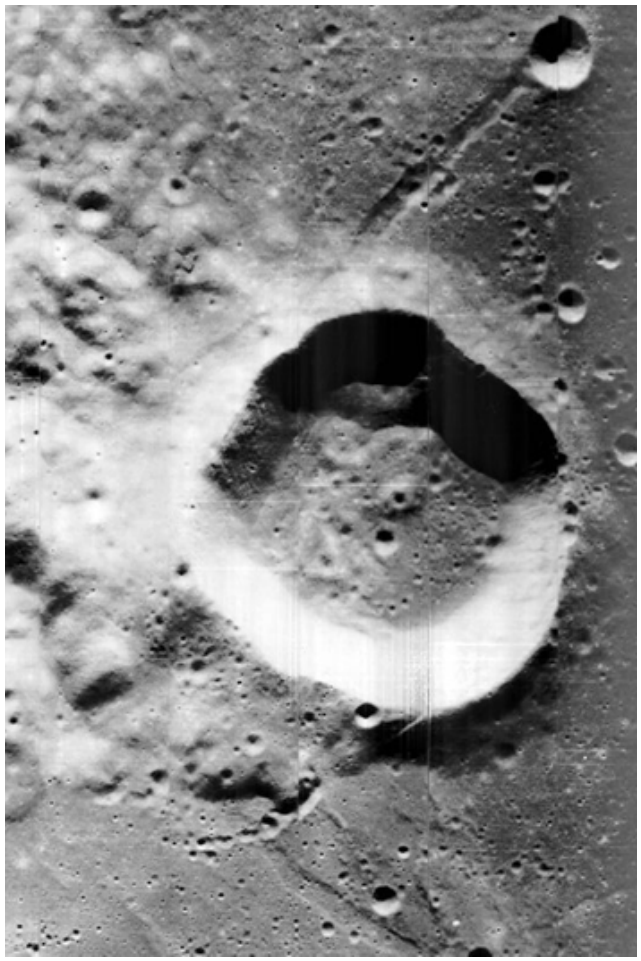
Interpreting – completion of percepts

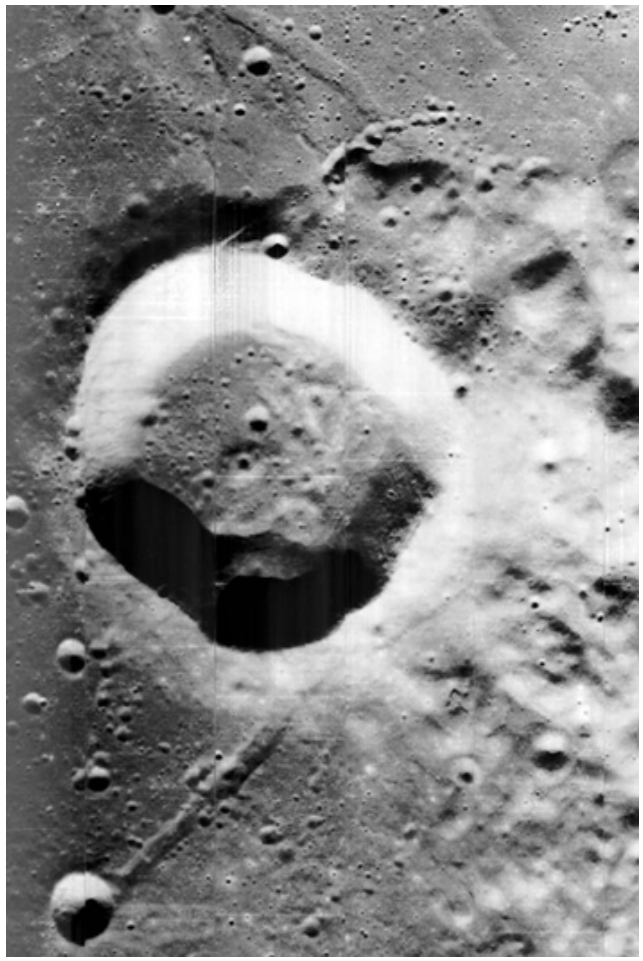
Induction and its illusions





Induction and illusions





Crater *or* hill?

Un exemple qui en dit beaucoup …

- Examples described using:

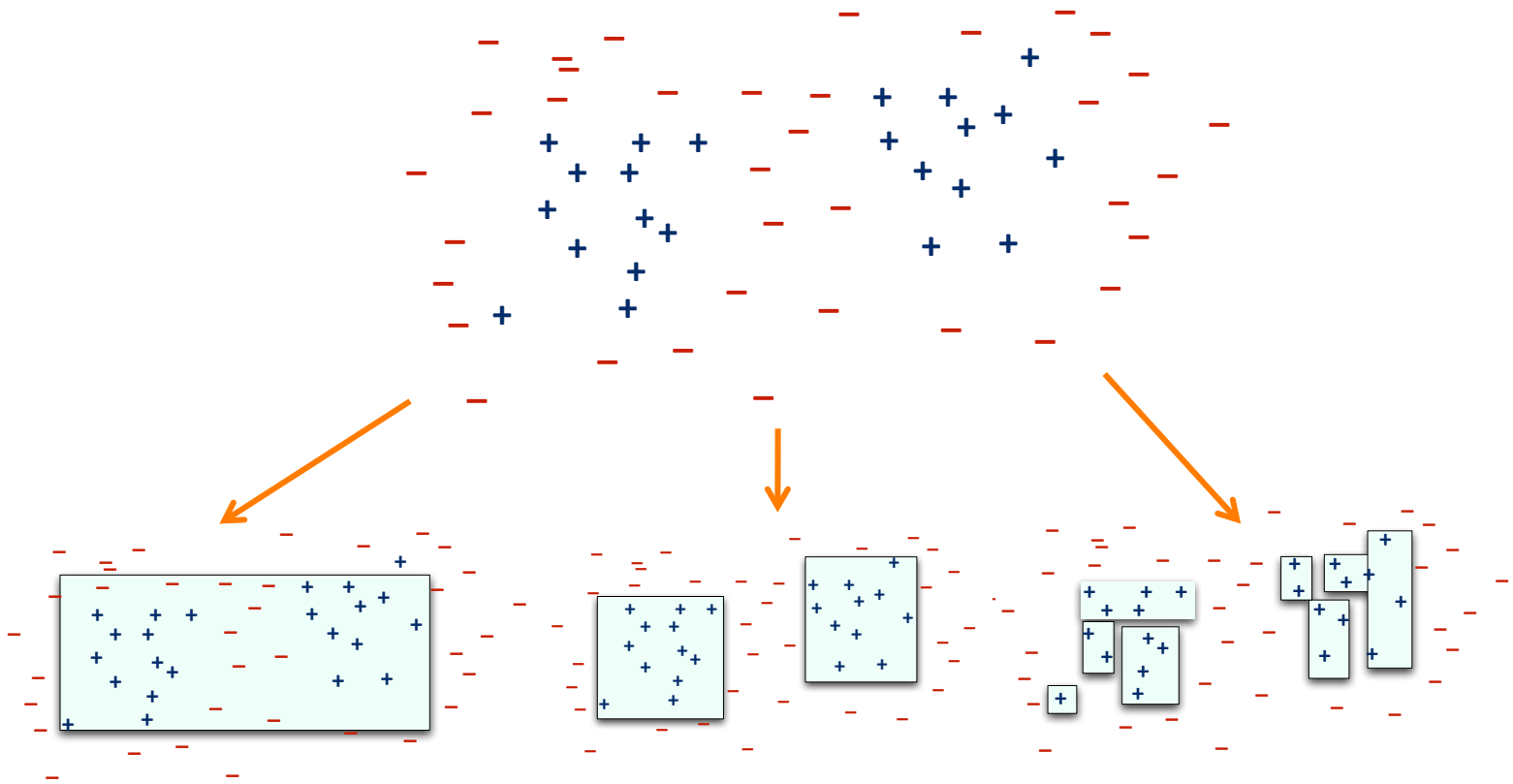
  ~~***Number***~~ (1 or 2); ***size*** (small or large); ~~***shape***~~ (circle or square); ***color*** (red or green)

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? $2^1 = 2$

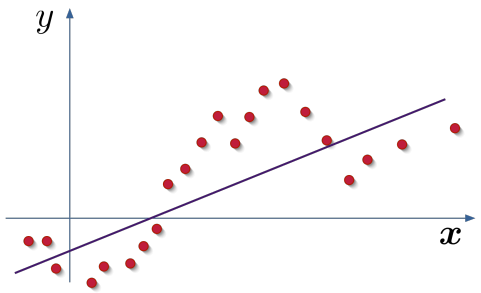
Comment garantir un niveau de performance ?



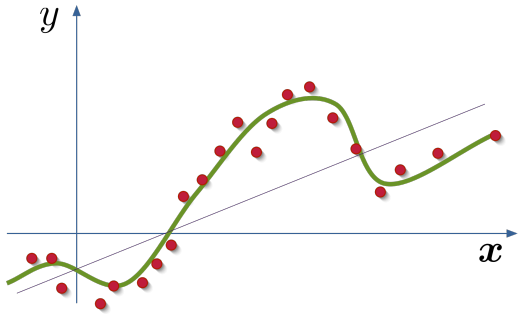
Sous-apprentissage **Bon**-apprentissage Sur-apprentissage

x

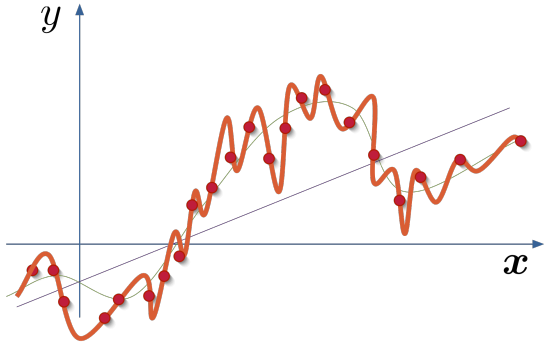
Comment garantir un niveau de performance ?



Sous-apprentissage



Bon-apprentissage



Sur-apprentissage

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 ? $h^* = \underset{h \in \mathcal{H}}{\text{ArgMin}} \ R(h)$

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

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

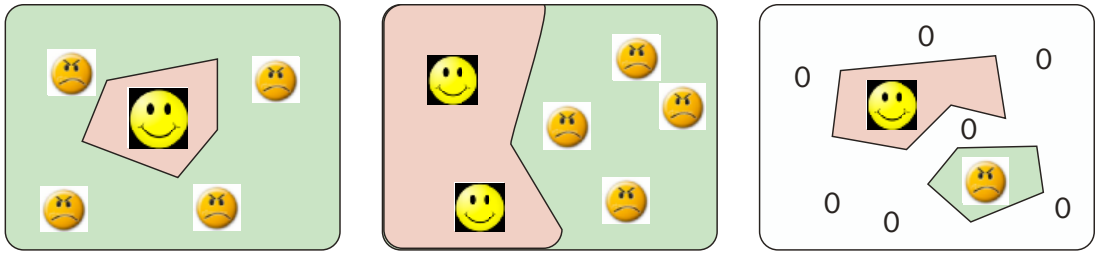
n'est sain **que si** il y a des contraintes sur l'espace des hypothèses

Plan

1. Pourquoi toute cette excitation ?
2. Grands types d'apprentissage
3. Apprentissage prédictif par réseaux de neurones
4. Quelles garanties ?
5. Le no-free-lunch theorem
6. Les réseaux de neurones profonds
7. Ce que l'on sait faire et les défis à relever

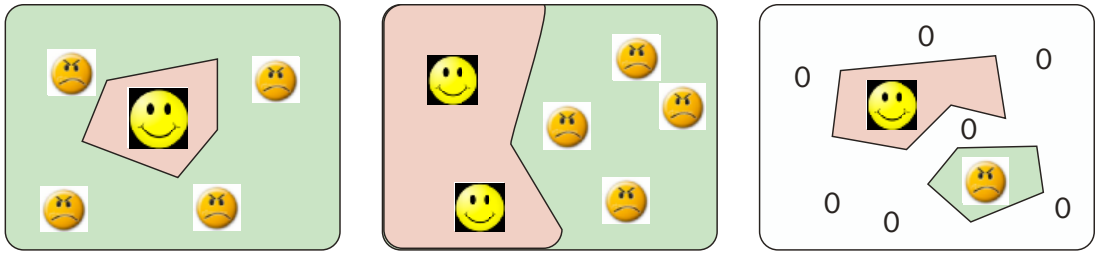
Le no-free-lunch theorem

Possible

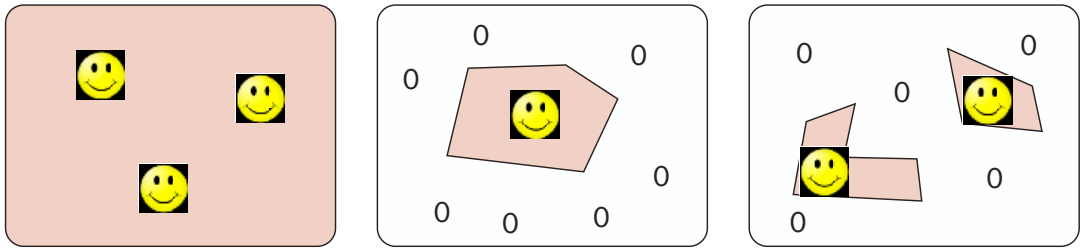


Le no-free-lunch theorem

Possible



Impossible



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