Une **Introduction**
à l'apprentissage artificiel

Antoine Cornuéjols

AgroParisTech – INRA MIA 518

antoine.cornuejols@agroparistech.fr

AgroParisTech

Une science … son objet

« **How** can we **build** ***computer systems*** that automatically improve with experience,

and

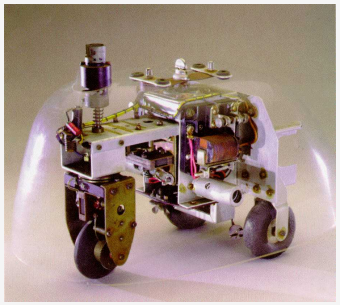
what are **the** **fundamental laws** that govern ***all learning processes***? »

Tom Mitchell, 2006

Apprendre pour quoi faire ?

- Science de la **modélisation**
  - Recherche de **régularités sous-jacentes** aux *données d'observation*
    - Pour **comprendre** et **expliquer**
    - Pour **prédire** et **décider**

- Science de l'adaptation
  - Recherche de schémas situation -> action par *interaction avec le monde*
    - Pour réagir et anticiper
    - Apprentissage par renforcement / Évolution simulée



Plan

1. Des données aux régularités : **l'induction** selon l'Apprentissage Artificiel
2. L'**espace des hypothèses** (modèle)
3. Le Graal : trouver le bon **changement de représentation**
4. Quelques **extensions** mais toujours dans le paradigme
5. **Conclusions** … et **nouvelles directions ?**

Des données aux régularités

L'induction

Apprentissage à partir d'exemples

- Concepts compliqués à programmer
  - *Mouvements appropriés pour un robot*
  - *Personne à recruter / ne pas recruter*
  - *Caractéristiques prédisposant à un certains types de cancer*

→ Apprentissage **à partir d'exemples**

Les données

- Vectorielles
- Séquences
- Structurés (ensemble de graphes ?)
- Temporelles
- Spatiales

Les données

- **Vectorielles**
- Séquences
- Structurés
- Temporelles
- Spatiales

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

Exemple
(*example, instance*)

Descripteur
Attribut
(*feature*)

Étiquette
(*label*)

Les données

- Vectorielles
- **Séquences**
- Structurés
- Temporelles
- Spatiales

La protéine « sp|P00004|CYC_HORSE » est activée par …

```
 1   ttcagttgtg aatgaatgga cgtgccaaat agacgtgccg ccgccgctcg attcgcactt
61   tgctttcggt tttgccgtcg tttcacgcgt ttagttccgt tcggttcatt cccagttctt
121 aaataccgga cgtaaaaata cactctaacg gtcccgcgaa gaaaaagata aagacatctc
181 gtagaaatat taaaataaat tcctaaagtc gttggtttct cgttcacttt cgctgcctgc
 …
4021 agaacacgcc gaggctccat tcatagcacc acttcgtcgt cttaatcccc tccctcatcc
4081 gccatggcgg tgcaaaaaat aaaaagaact c
```

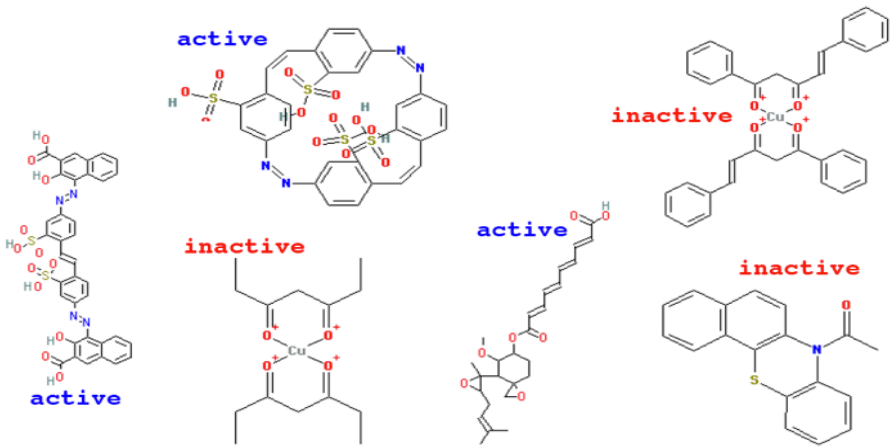
```
DEVICE=eth0
BOOTPROTO=none
ONBOOT=yes
IPADDR=192.168.0.X
NETMASK=255.255.255.O
GETEWAY=192.168.0.254
search exemple.com namserver
192.168.0.254
```

Les données

- Vectorielles
- Séquences
- **Structurés**
- Temporelles
- Spatiales

Logique du 1er ordre :

bloc(B1) & surtable(B2) & au-dessus(B1,B2) & …

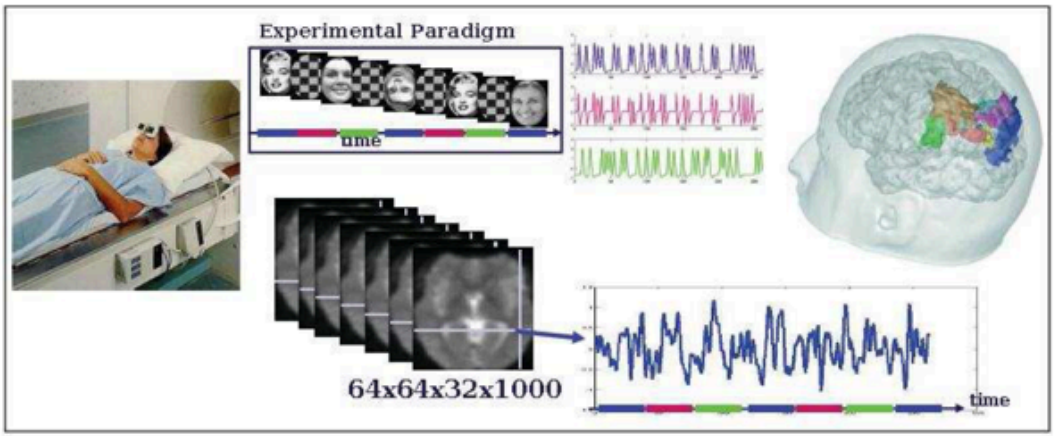


NCI AIDS screen results (from http://cactus.nci.nih.gov).

Les données

- Vectorielles
- Séquences
- Structurés
- **Temporelles**
- Spatiales

• **Apprentissage supervisé : interprétation d'IRMf**



Trouble de la reconnaissance de visage ou *non*

Les données

- Vectorielles
- Séquences
- Structurés
- Temporelles
- **Spatiales**

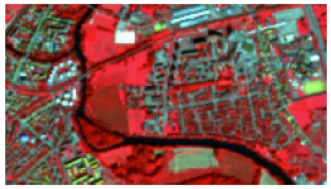
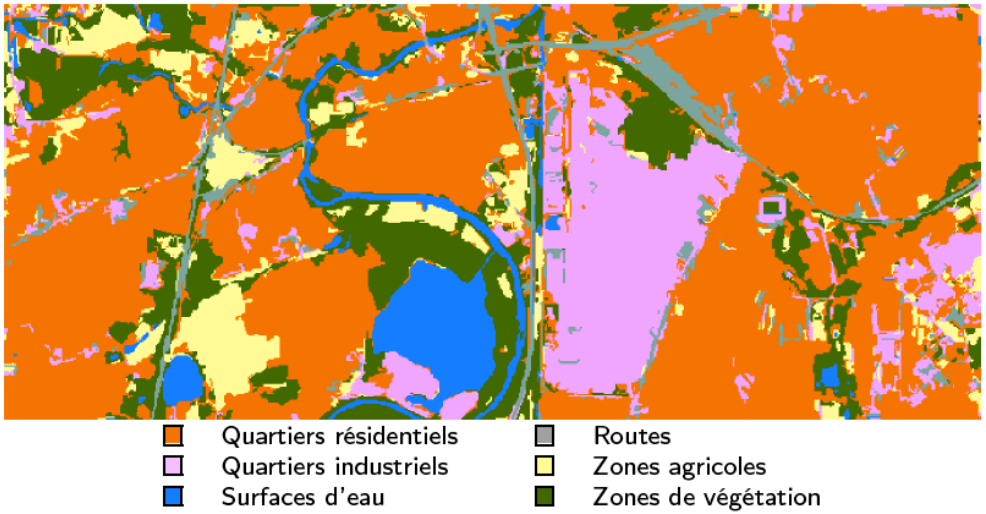


Image MRS

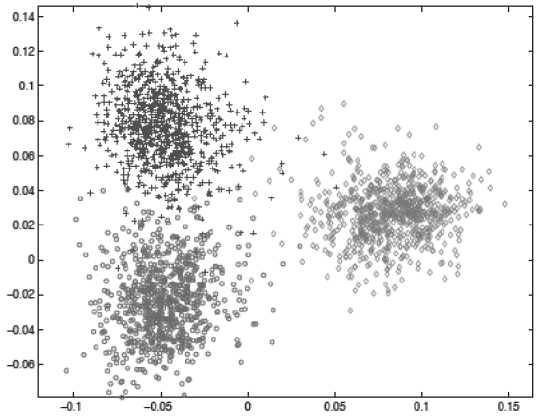


Cadre **non supervisé**

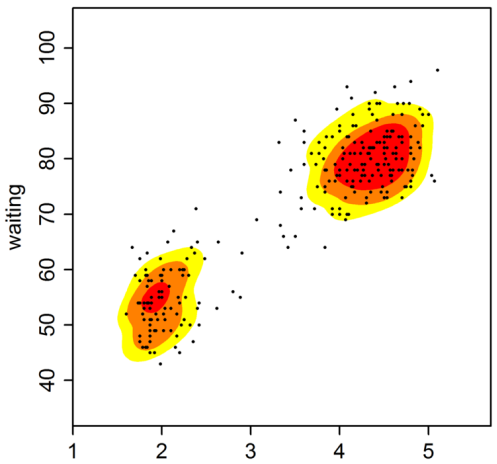
Les données **non supervisées**

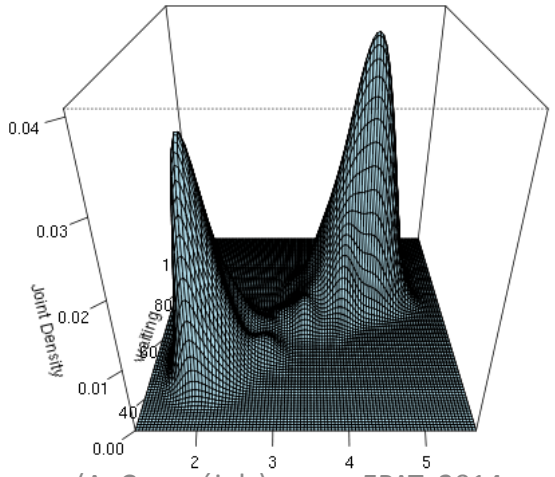
- **Sans sorties associées**

  - Clustering



  - Estimation de densité





L'apprentissage non supervisé

$$\mathcal{S} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$$

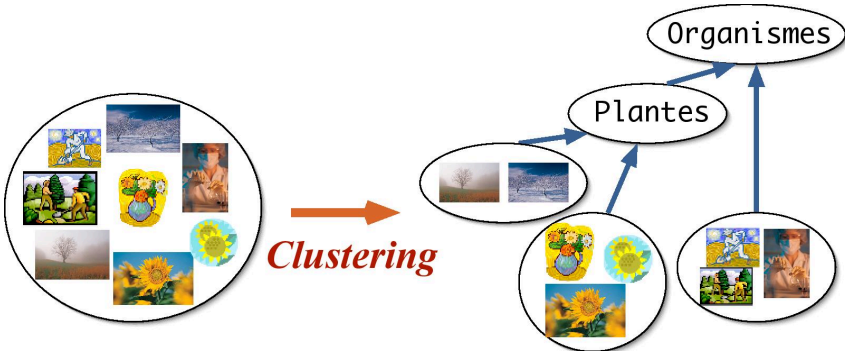
- *L'univers numérique*
  - En 2007 : ~ 281 exabytes (281 milliards de gigabytes = 2,81 . 10^{20})
  - En 2011 : ~3 . 10^{21}
  - Surtout des images et des videos

→ ***Comment organiser cette masse énorme de données ?***

Le clustering (catégorisation)

- **Organiser** un ensemble de « formes » en **groupes contrastés**
- **Comprimer** les données en y découvrant une **structure**

- Afin de
  - « comprendre » les données
  - Et de faire des prédictions



Le clustering (catégorisation)

Angkor Wat

Hindu temple built by a Khmer king ~1,150AD; Khmer kingdom declined in the 15th century; French explorers discovered the hidden ruins in late 1800's

Le clustering (catégorisation)

Apsaras of Angkor Wat

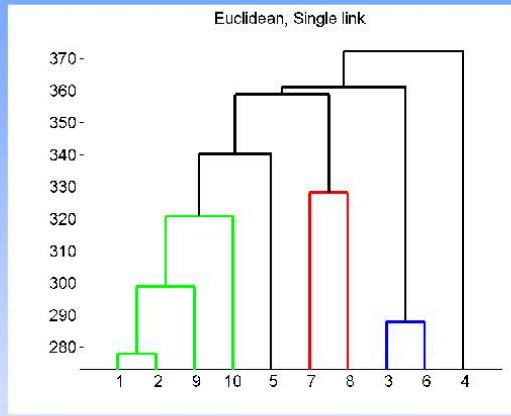
- Angkor Wat contains the most unique gallery of ~2,000 women depicted by detailed full body portraits
- What facial types are represented in these portraits?

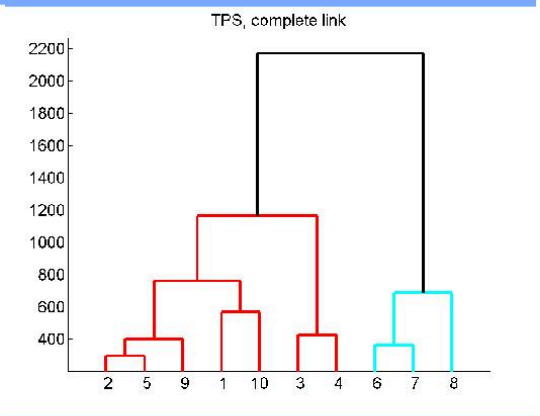
Kent Davis, "Biometrics of the Godedess", DatAsia, Aug 2008
S. Marchal, "Costumes et Parures Khmers: D'apres les devata D'Angkor-Vat", 1927

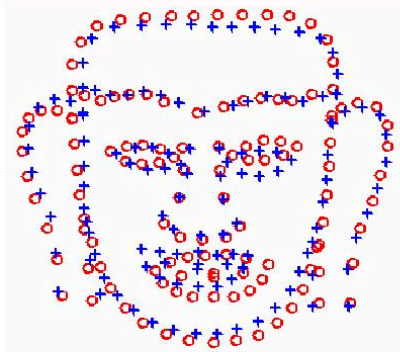
Le clustering (catégorisation)

Clustering of Apsara Faces

127 landmarks







Shape alignment

Single Link clusters

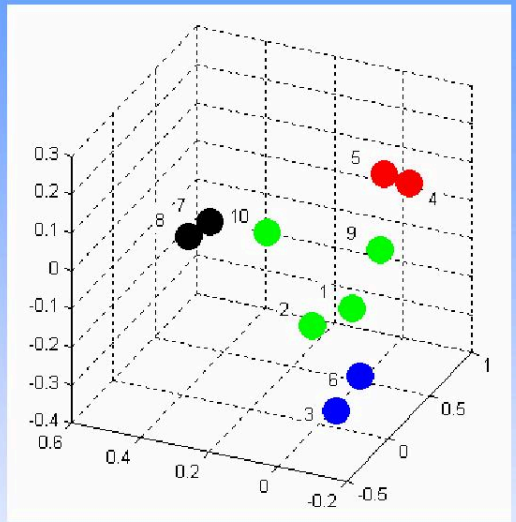
An ethnologist needs to validate the groups

Le clustering (catégorisation)

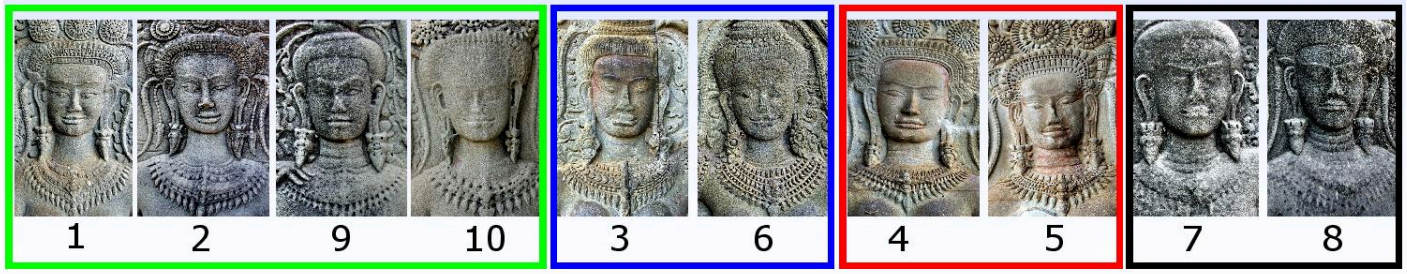
Clustering of Apsara Faces

.51	.72	.96	.70	.73	.89	.78	.61	.61
	.67	1	.72	.70	.80	.66	.55	.59
		.87	.78	.53	.73	.77	.71	.83
			.69	.78	.87	.94	.88	.97
				.72	.74	.75	.62	.71
					.72	.76	.69	.78
						.60	.75	.86
							.72	.67
		0						.68

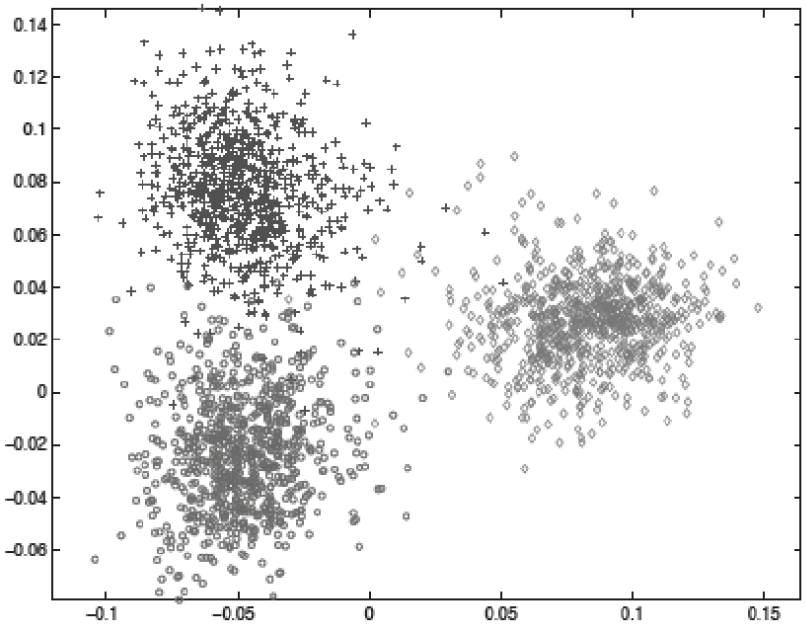
Dissimilarity matrix

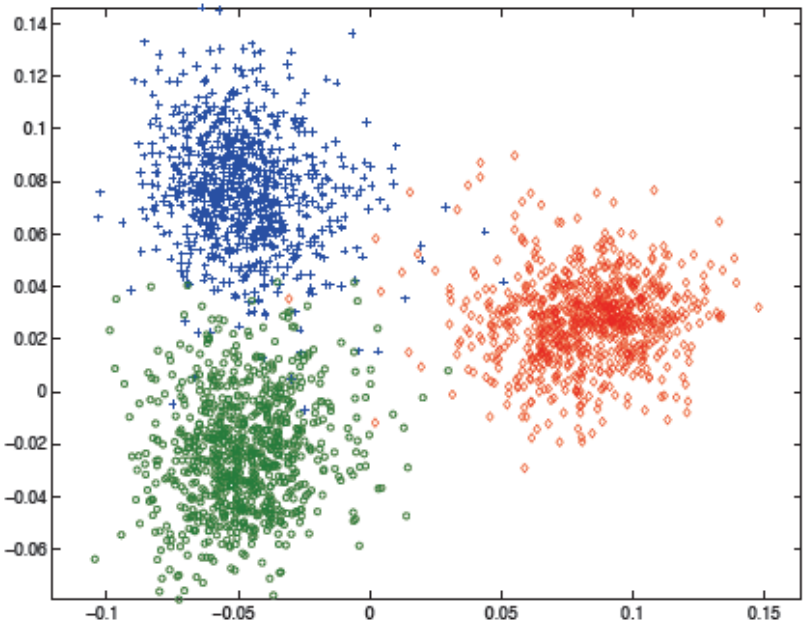


4 clusters with K-means in 3D feature space

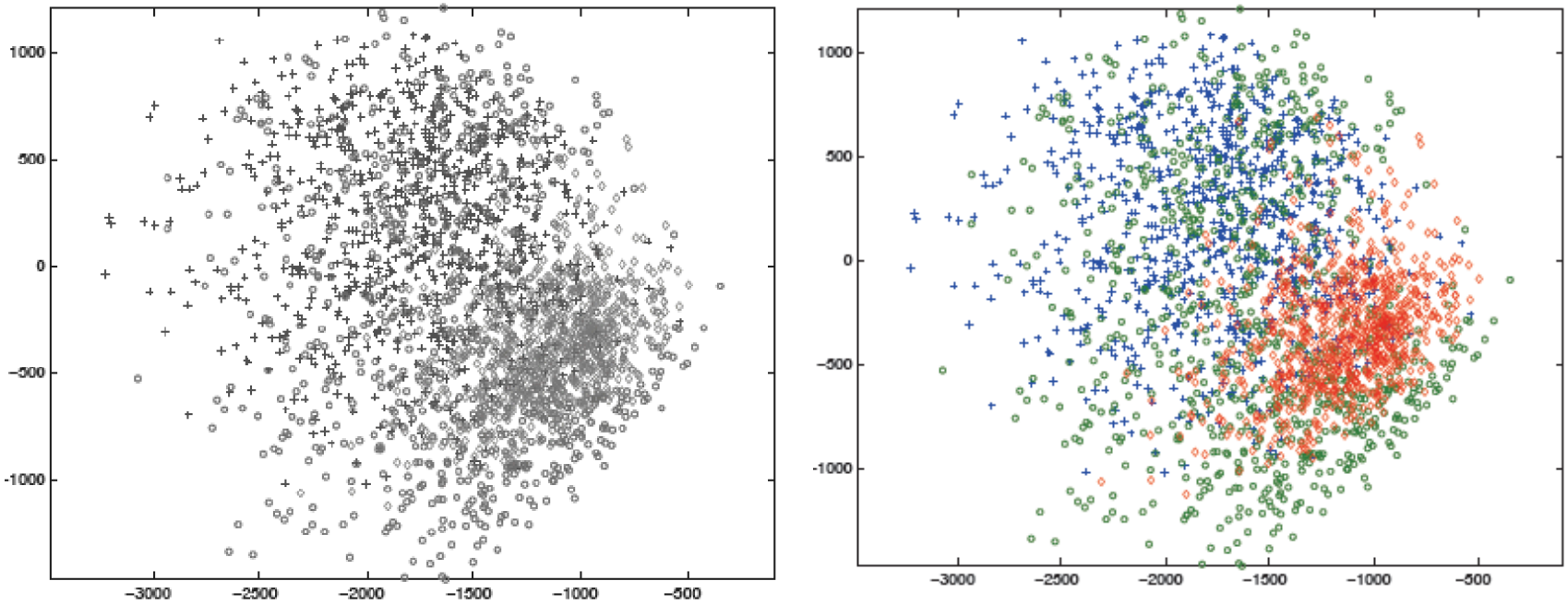


Clustering


0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
−0.02
−0.04
−0.06
−0.1
−0.05
0
0.05
0.1
0.15

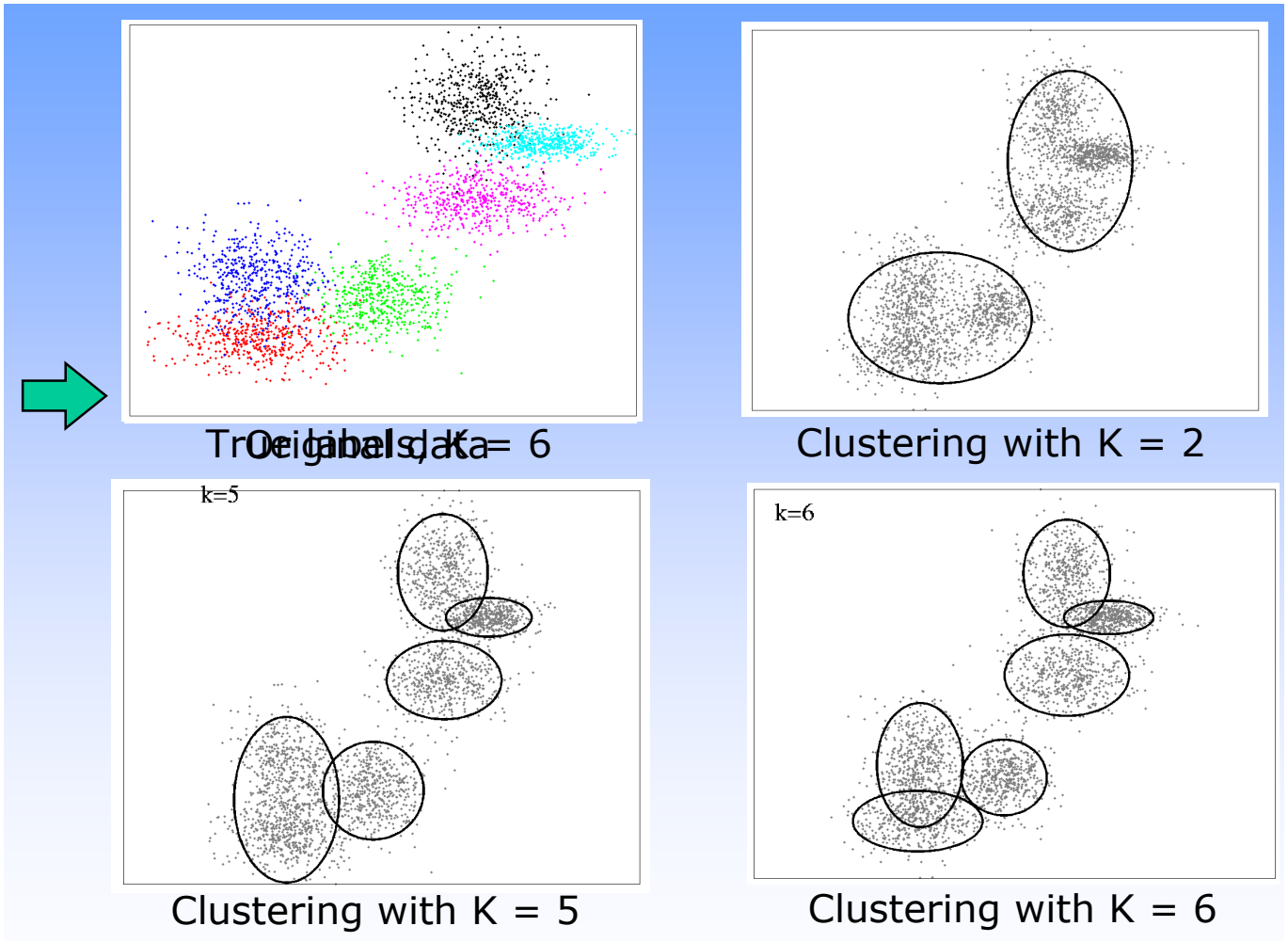

0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
−0.02
−0.04
−0.06
−0.1
−0.05
0
0.05
0.1
0.15

Clustering



Le clustering (catégorisation)

Importance du **choix du nombre de catégories attendues**

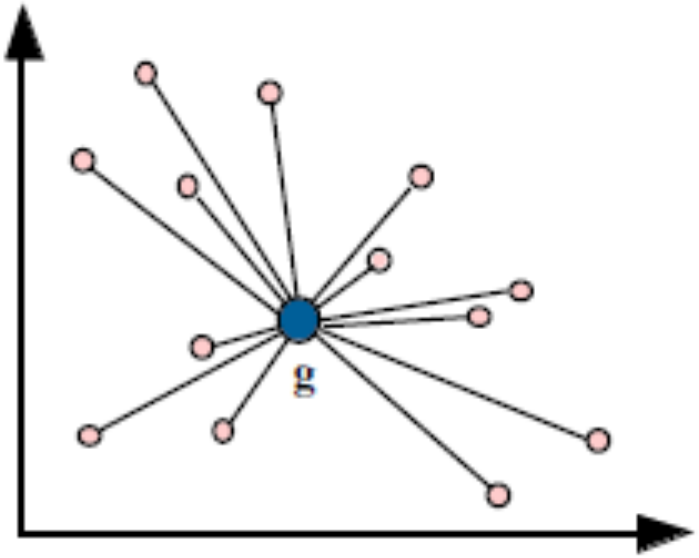


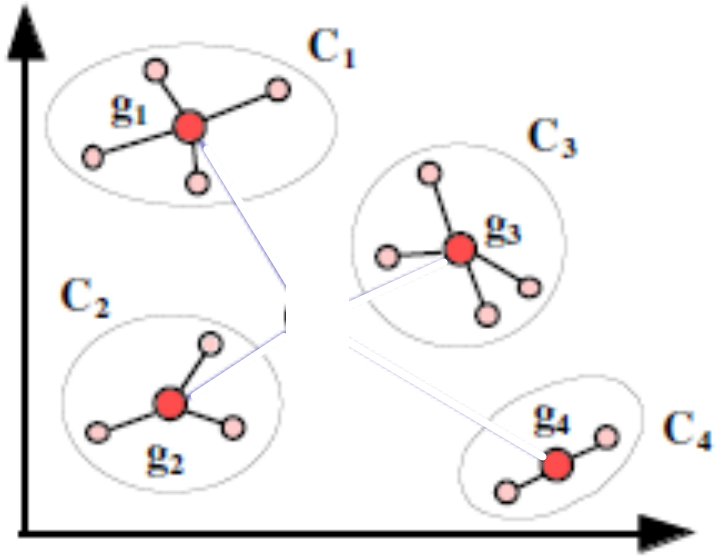
Le clustering (catégorisation)

Importance du **choix de la distance utilisée**

Le clustering (catégorisation)

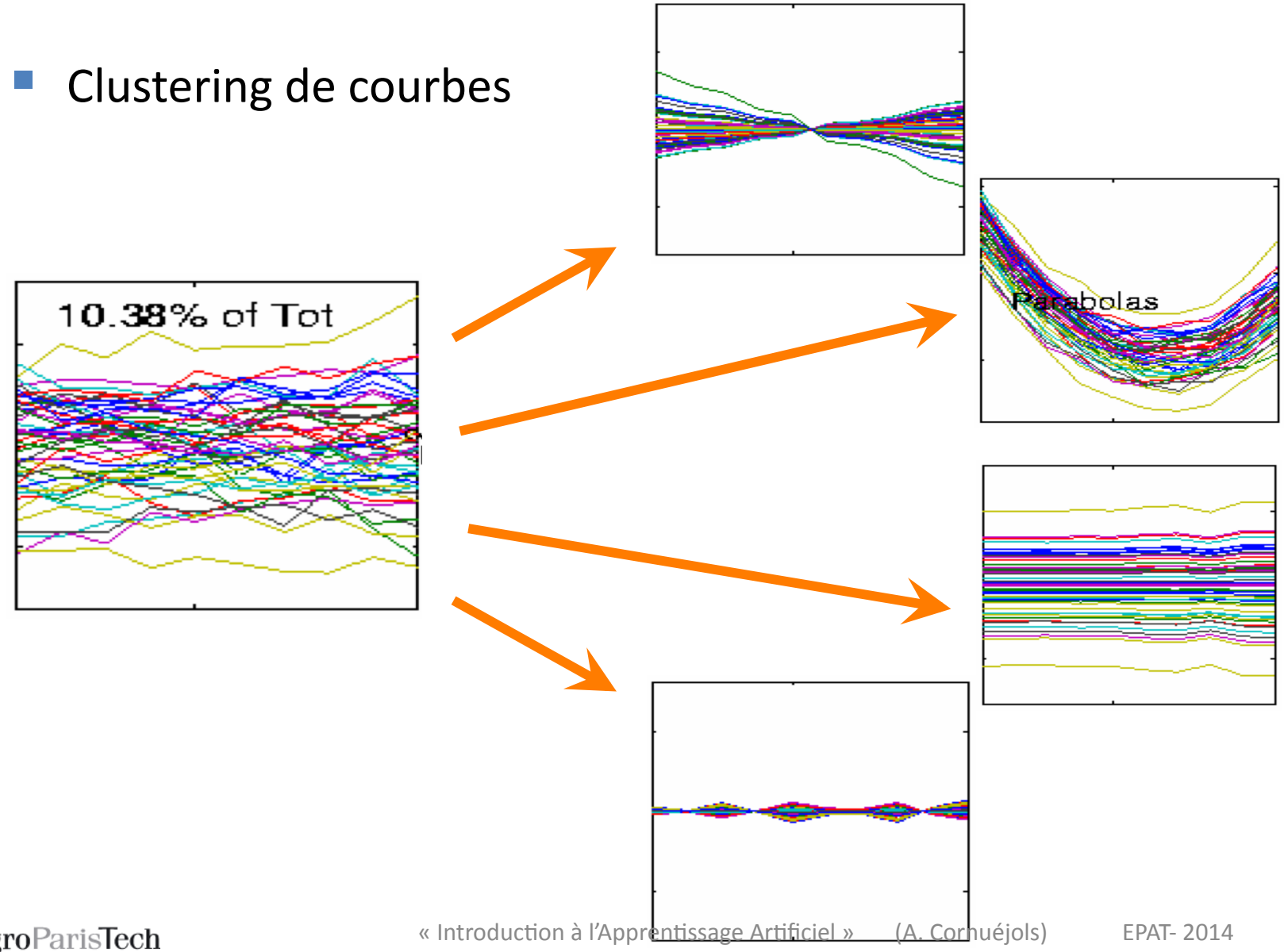
Importance du **choix de la distance utilisée**





Apprentissage **non supervisé**

- Clustering de courbes



Séries temporelles : Questions

1. **Représentation** des séquences

2. **Alignement** entre séquences

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

Séries temporelles : Questions

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

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

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

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

La représentation des séquences

1. Brute

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

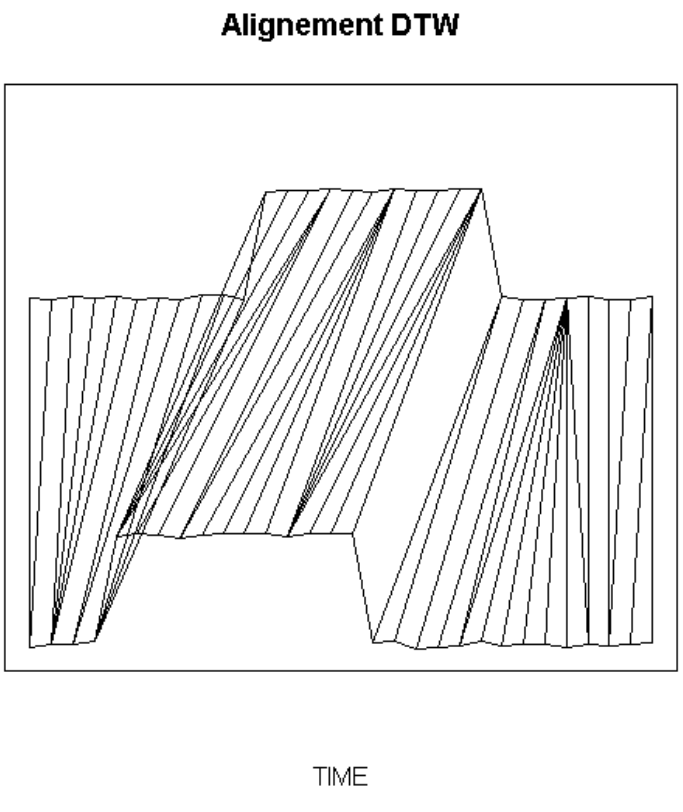
3. Régularités supposées : représentations spécifiques
   - Grammaires
   - Chaînes de Markov
   - Réseaux de neurones
   - …
   - Problème de la comparaison

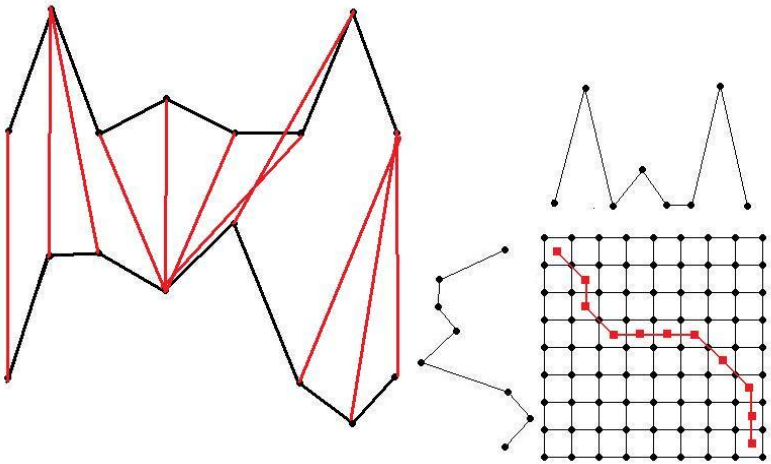
L'alignement

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

Alignement par DTW

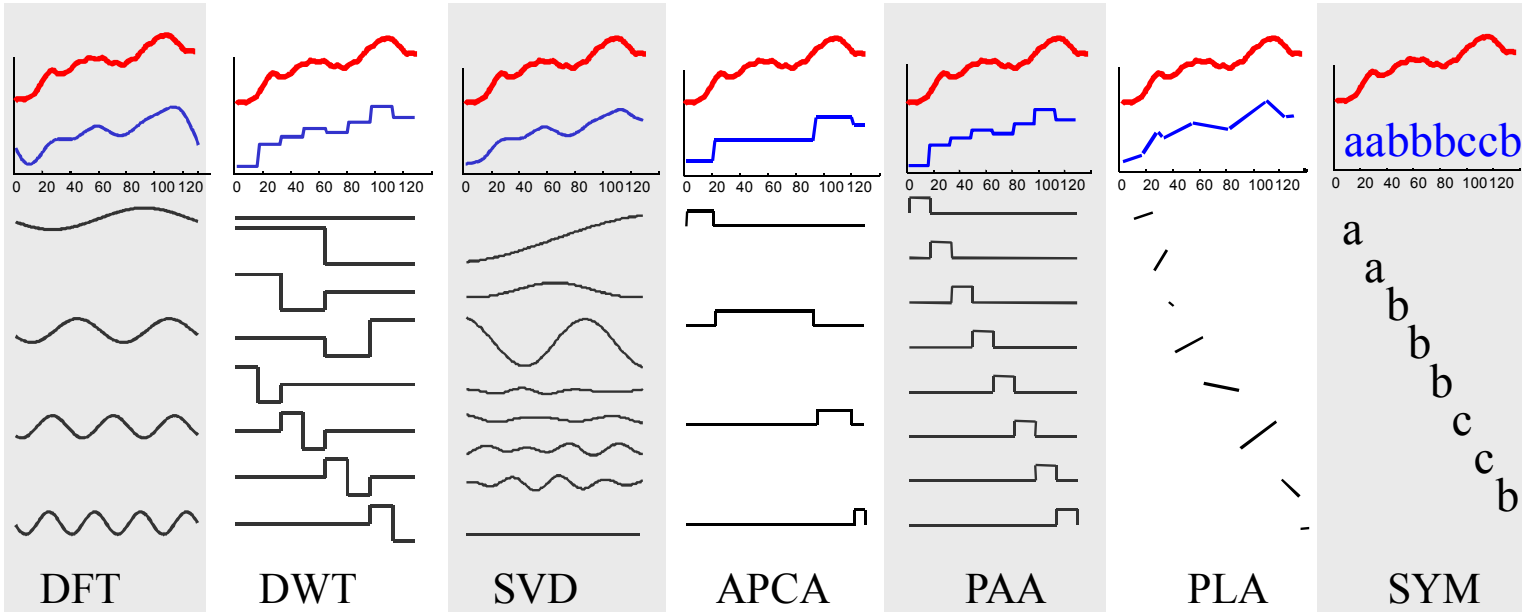
Sankoff, D. et Kruskal, J. (1983). Time warps, string edits, and macromolecules : the theory and practice of sequence comparison. *Reading : Addison-Wesley Publication, 1983, edited by Sankoff, David ; Kruskal, Joseph B., 1*





Représentations analytiques

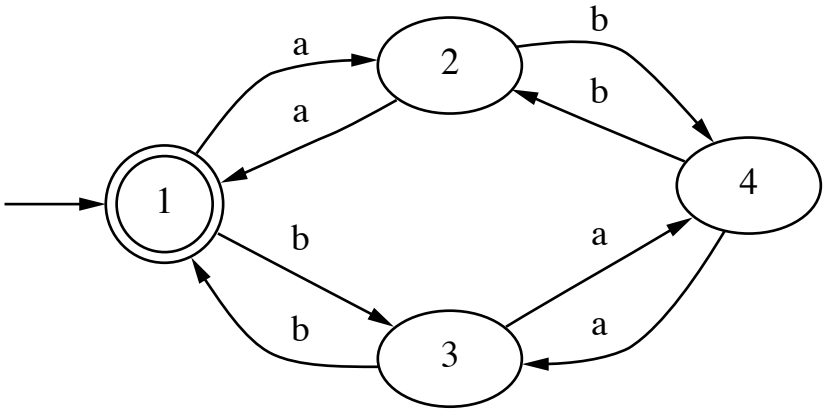
- Transformée de Fourier discrète
- Transformée en ondelettes discrète
- Approximation agrégée par morceaux
- …



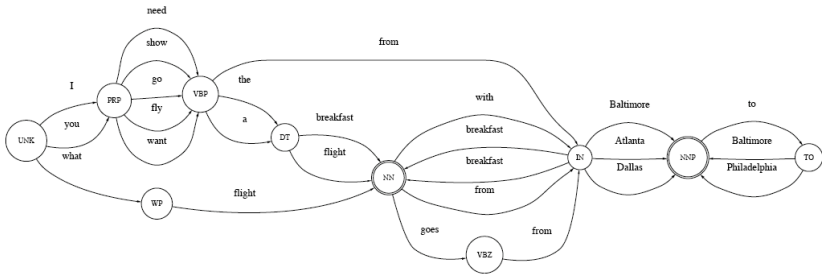
Représentation vectorielle

Représentations particulières (e.g. automates)

a b a a a b b a a b b a b b a b b a …



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



Clustering

Clustering

Le clustering (catégorisation)

- Ingrédients essentiels
  - **Nombre** de clusters
  - **Distance** employée
  - **Algorithme**
    - K-moyennes
    - Clustering Hiérarchique Ascendant (AHC)
    - Clustering descendant (e.g. Cobweb [Doug Fisher, 1987])
    - …

- *Peut-on les **déterminer automatiquement** ?*
  - *Nombre* de clusters : OUI
  - *Distance* employée : Vous verrez

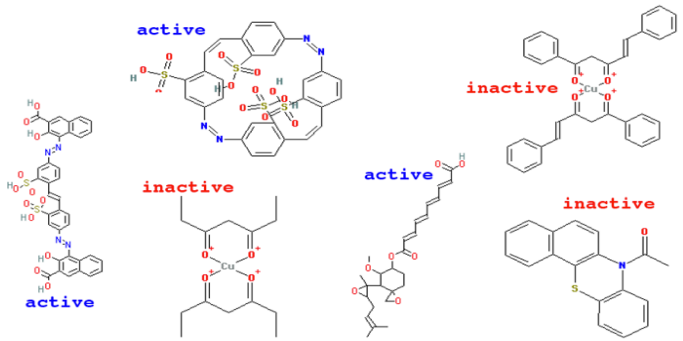
Cadre **supervisé**

Les données **supervisées**

- **Avec sorties associées**

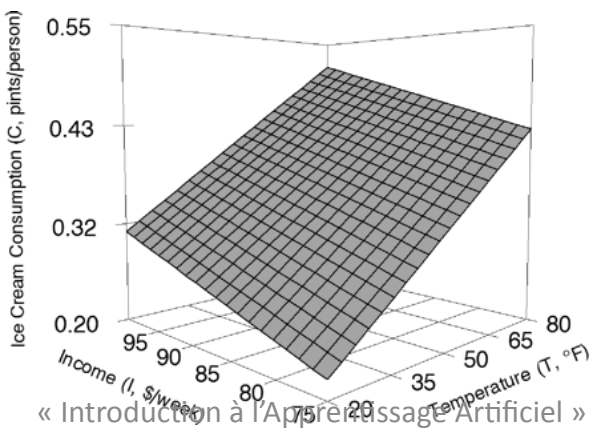
$$S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i), \ldots, (x_m, y_m)\}$$

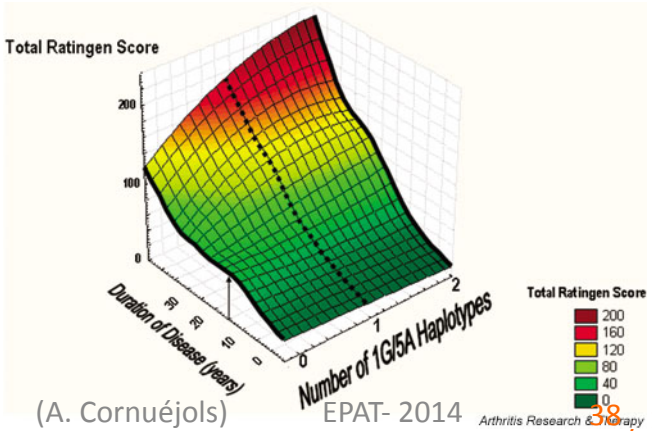
– Classification



– Régression

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





Plus formellement

- Echantillon d'apprentissage
  - **Données** d'observation (supposées représentatives)

$$S = \{(\boldsymbol{x}_1), (\boldsymbol{x}_2), \ldots, (\boldsymbol{x}_i), \ldots, (\boldsymbol{x}_m)\}$$

$$S = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_i, y_i), \ldots, (\boldsymbol{x}_m, y_m)\}$$

  - **Dépendance** entre $\mathcal{X}$ et $\mathcal{Y}$
    - **Densité de probabilité jointe** : $\mathbf{P}_{\mathcal{X}\times\mathcal{Y}}$
    - **Fonction cible** : f

Le point de vue **statistique**

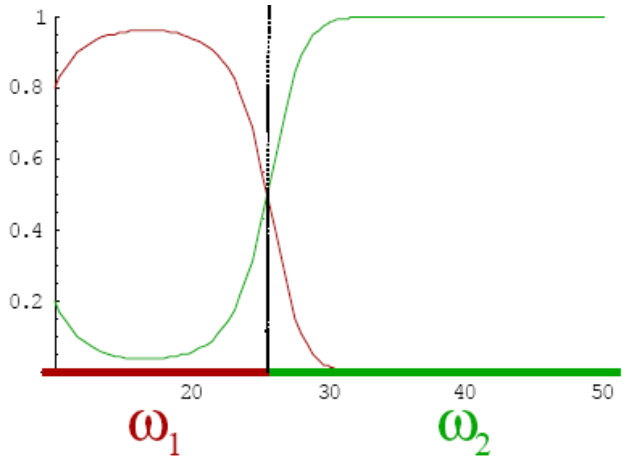
Reconnaissance des formes

Dépendance = loi de distribution jointe

Approche bayésienne

- Soit un ensemble fini de valeurs de sortie : $\{y_i\}_{1 \le i \le K}$
- **Si** l'on connaissait les distributions $P_{y|\mathcal{X}}(y_i|\boldsymbol{x})$:

$$y^\star = \underset{y_i}{\text{Argmax}}\ \mathbf{p}_{\mathcal{Y}|\mathcal{X}}(y_i|\mathbf{x})$$



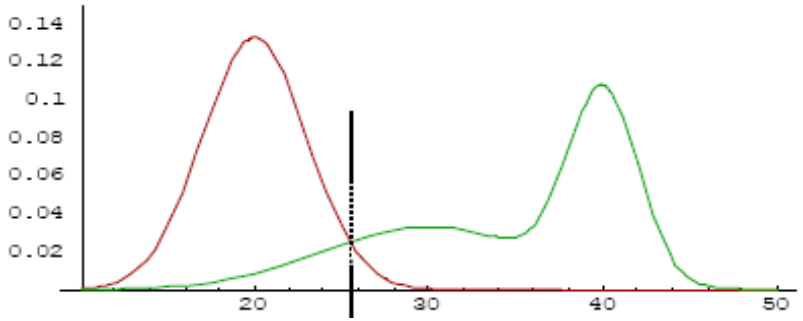
Approche « discriminative »

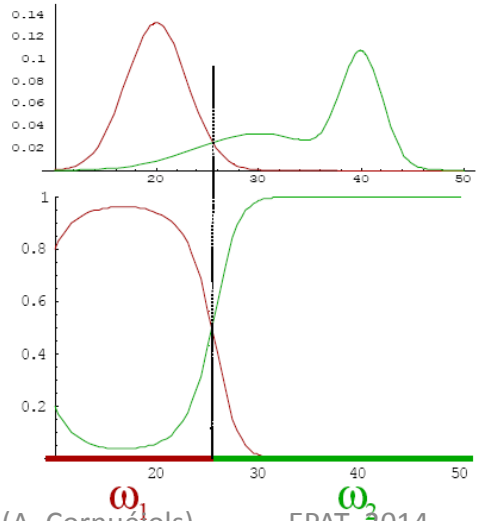
Dépendance = loi de distribution jointe

Approche bayésienne « générative »

$$y^\star = \text{Argmax}_{y_i} \; \mathbf{p}_{\mathcal{Y}|\mathcal{X}}(y_i|\mathbf{x})$$

$$= \text{Argmax}_{y_i} \; \mathbf{p}_{\mathcal{X}|\mathcal{Y}}(\mathbf{x}|y_i)\, \mathbf{p}_{\mathcal{Y}}(y_i)$$





Dépendance = loi de distribution jointe

- L'**apprentissage** dans l'approche bayésienne

  – On considère des **familles paramétrées de distributions**

  $$\mathbf{p}_{\theta_{\mathcal{X}|\mathcal{Y}}}(\mathbf{x}|y_i)$$

  – On estime θ à partir des données :

  - **MAP**

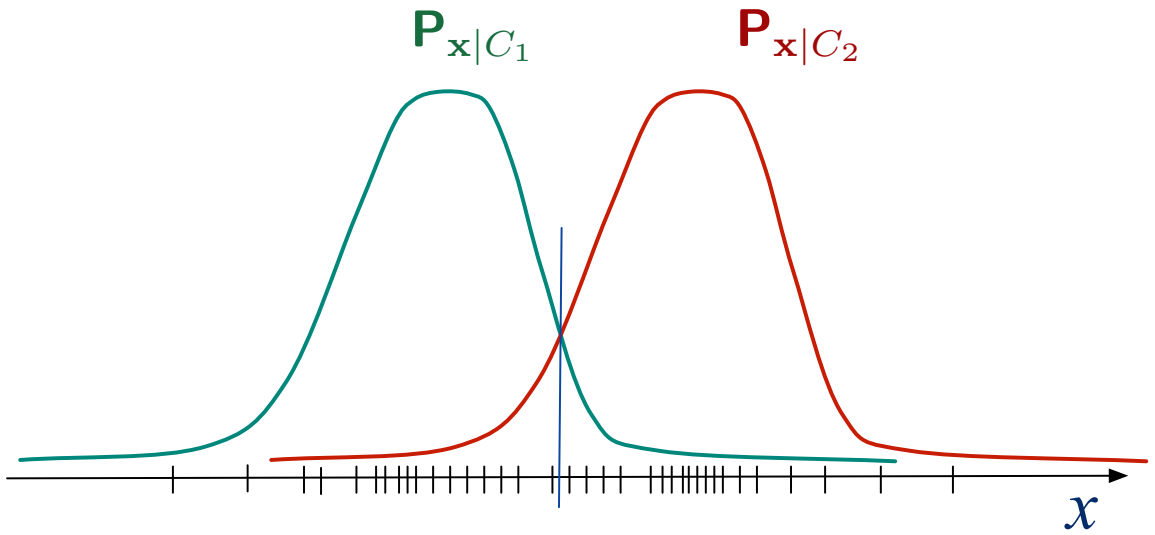
  $$\begin{aligned} \hat{\theta}_{\text{MAP}} &= \text{Argmax}_\theta \ \mathbf{p}(\theta|\mathcal{S}) \\ &= \text{Argmax}_\theta \ \mathbf{p}(\mathcal{S}|\theta)\ \mathbf{p}(\theta) \end{aligned}$$

  - **MLE**

  $$\begin{aligned} \hat{\theta}_{\text{MLE}} &= \text{Argmax}_\theta \ \mathcal{L}(\theta|\mathcal{S}) \\ &= \text{Argmax}_\theta \ \mathbf{p}(\mathcal{S}|\theta) \end{aligned}$$

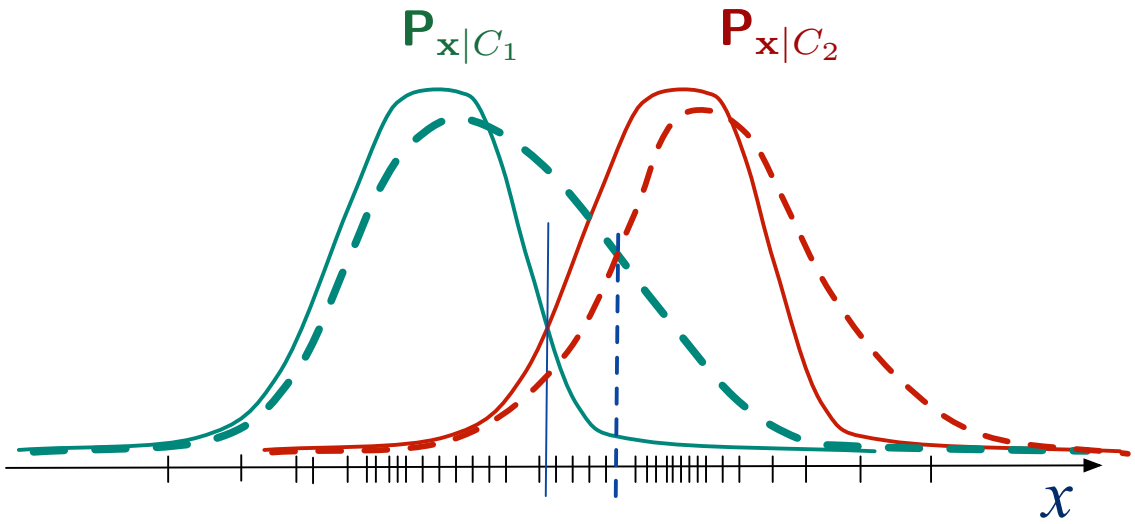
L'approche statistique : bilan

- Importance cruciale du choix des distributions de probabilité
  - Sinon …



L'approche statistique : bilan

- Importance cruciale du choix des distributions de probabilité
  - Sinon …



- En savoir plus que nécessaire :
  - un péché de gourmandise selon Vapnik : qui peut être mortel

L'approche statistique : bilan

- Importance cruciale du **choix des distributions de probabilité**

- Chercher à **en savoir trop peut être dangereux**

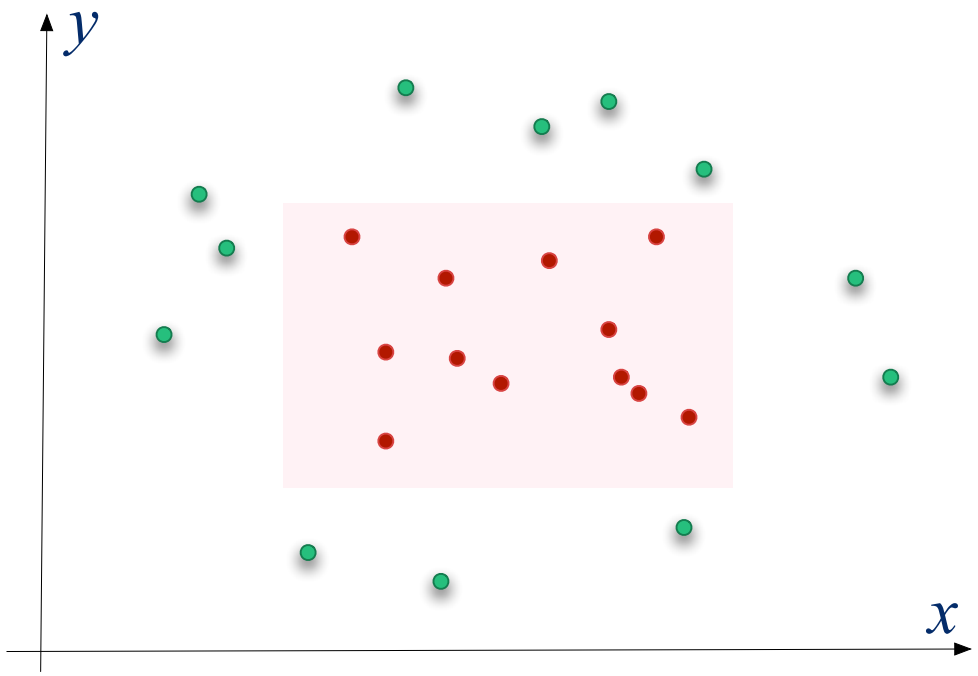
- **Difficultés techniques**
  - La ***malédiction de la dimensionalité***
  - Problèmes d'***optimisation difficiles***
    - Algorithmes EM
    - Algorithmes variationnels
    - MCMC (Markov Chains Monte Carlo)
    - …

Un **petit exemple**

Le point de vue de l'**Apprentissage Artificiel**

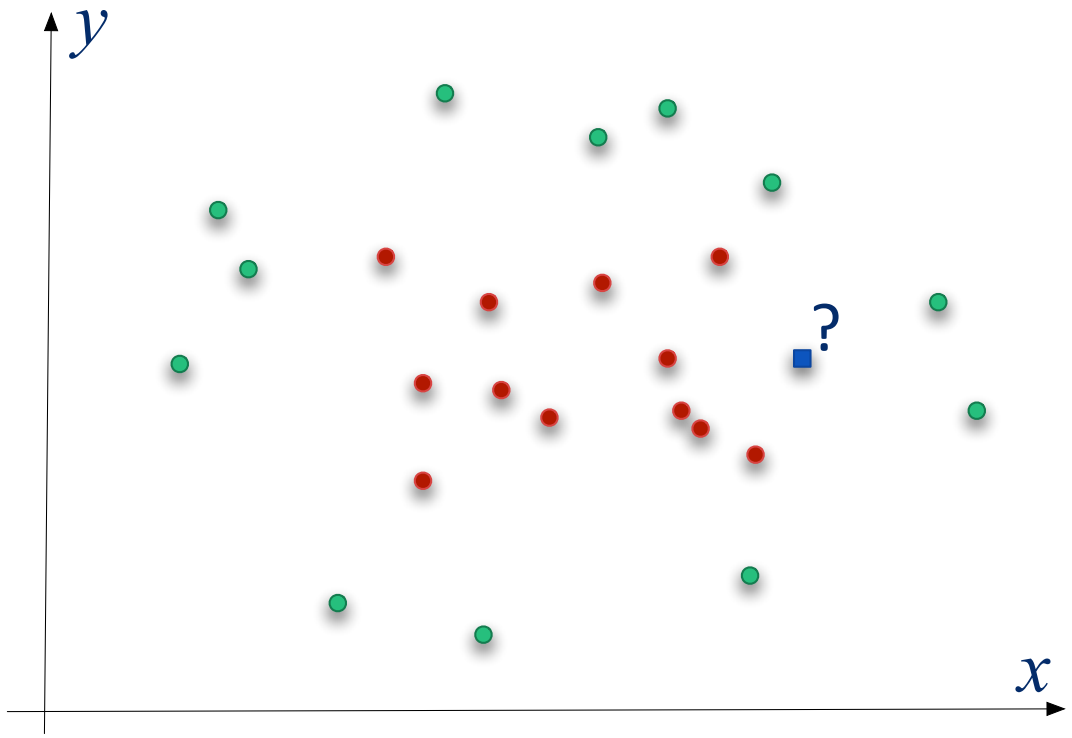
Apprentissage de rectangle

- Échantillon
  - D'exemples positifs $\mathbf{P}_{\mathcal{X}}^{+}$
  - D'exemples négatifs $\mathbf{P}_{\mathcal{X}}^{-}$



Apprentissage de rectangle

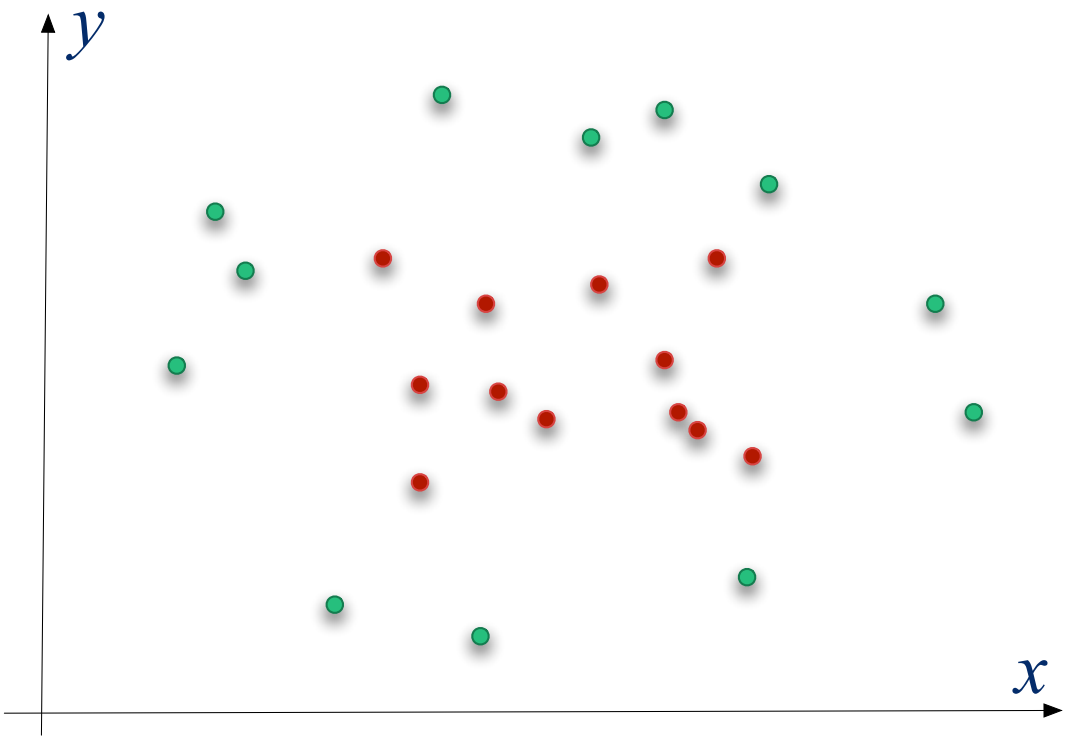
- Que cherche-t-on à apprendre ?



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

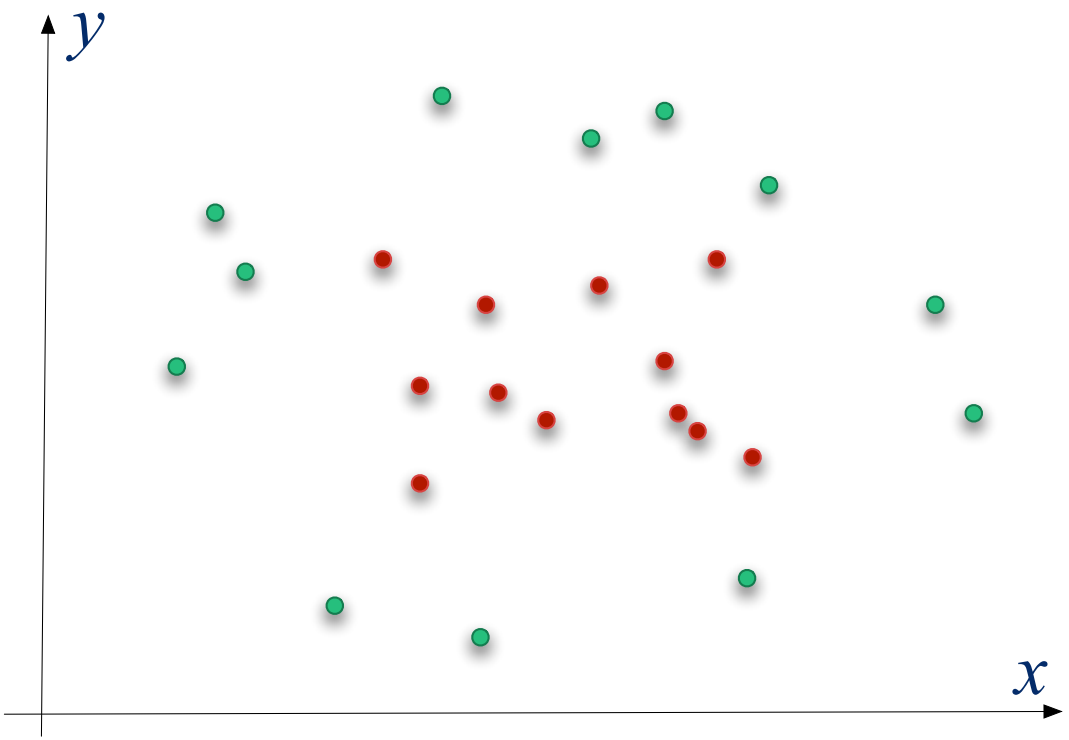
Apprentissage de rectangle

- Comment apprendre ?



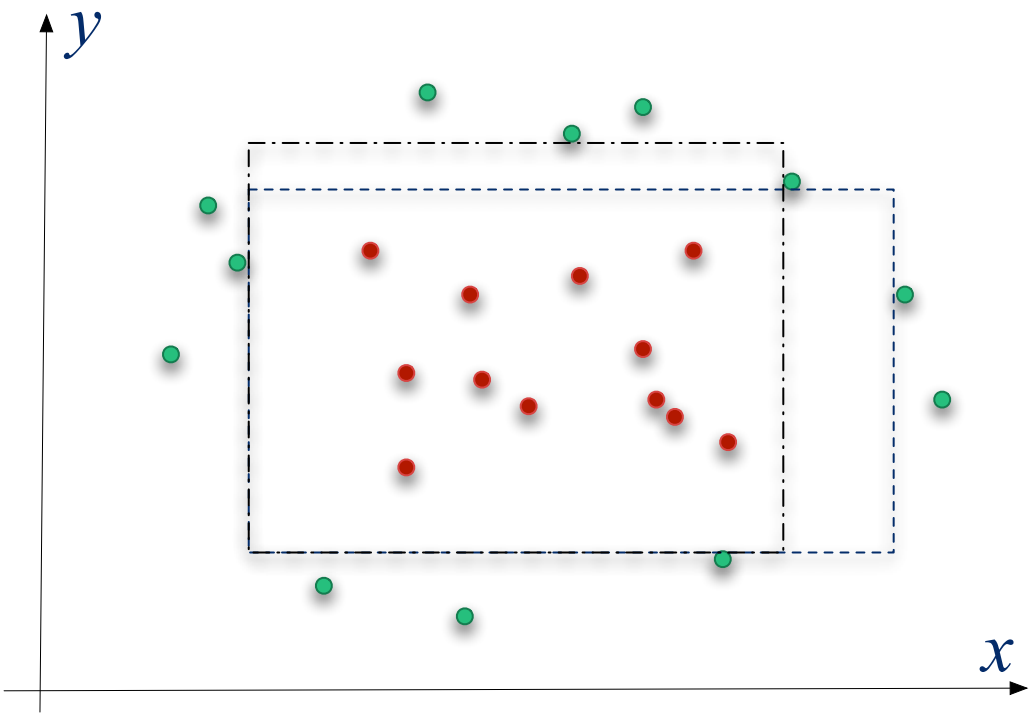
Apprentissage de rectangle

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



Apprentissage de rectangle

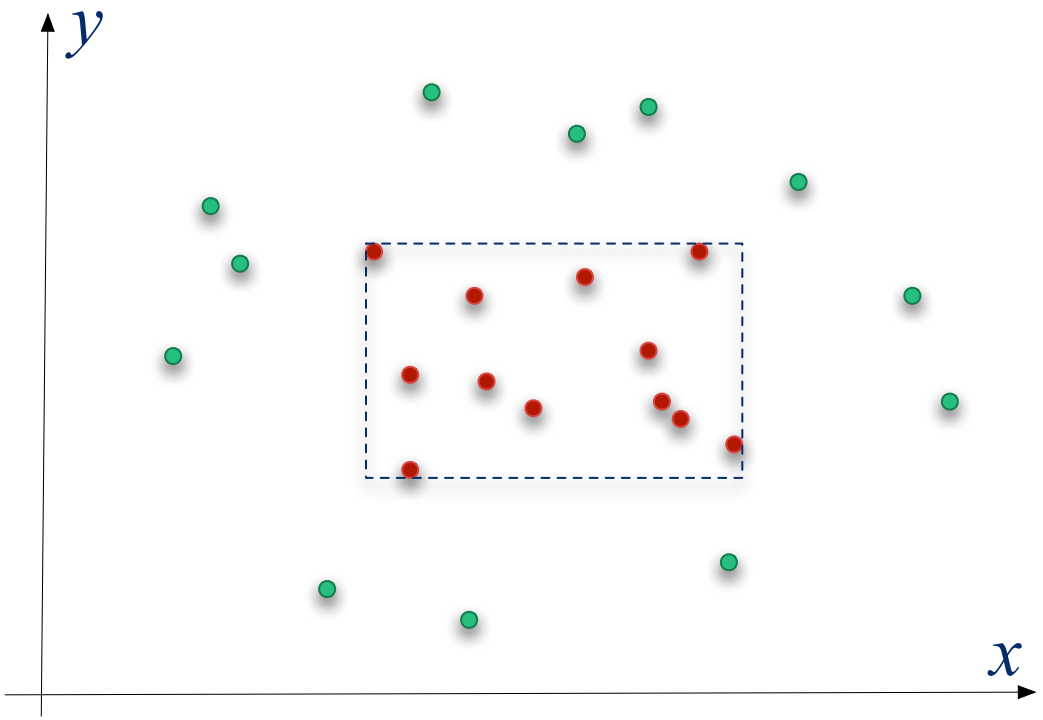
- Comment apprendre ?
  - Si **je sais que le concept cible est un rectangle**



Hypothèses les plus **générales**

Apprentissage de rectangle

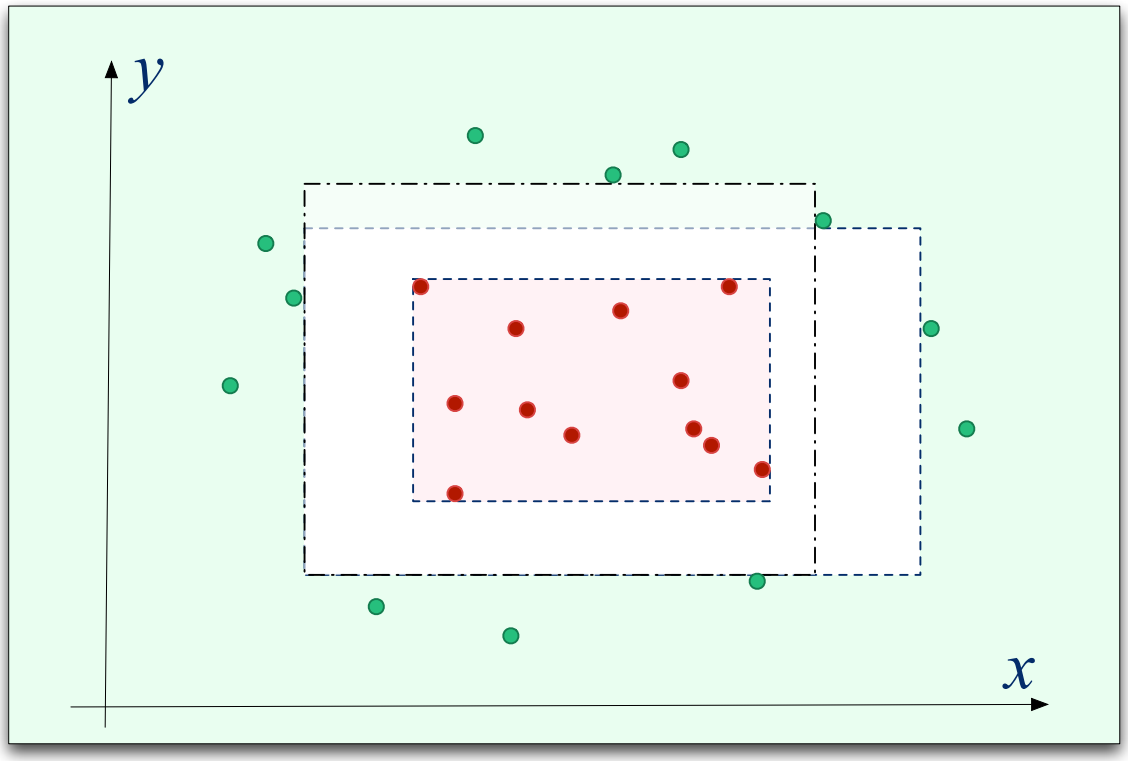
- Comment apprendre ?
  - Si **je sais que le concept cible est un rectangle**



Hypothèses les plus **spécifiques**

Apprentissage de rectangle

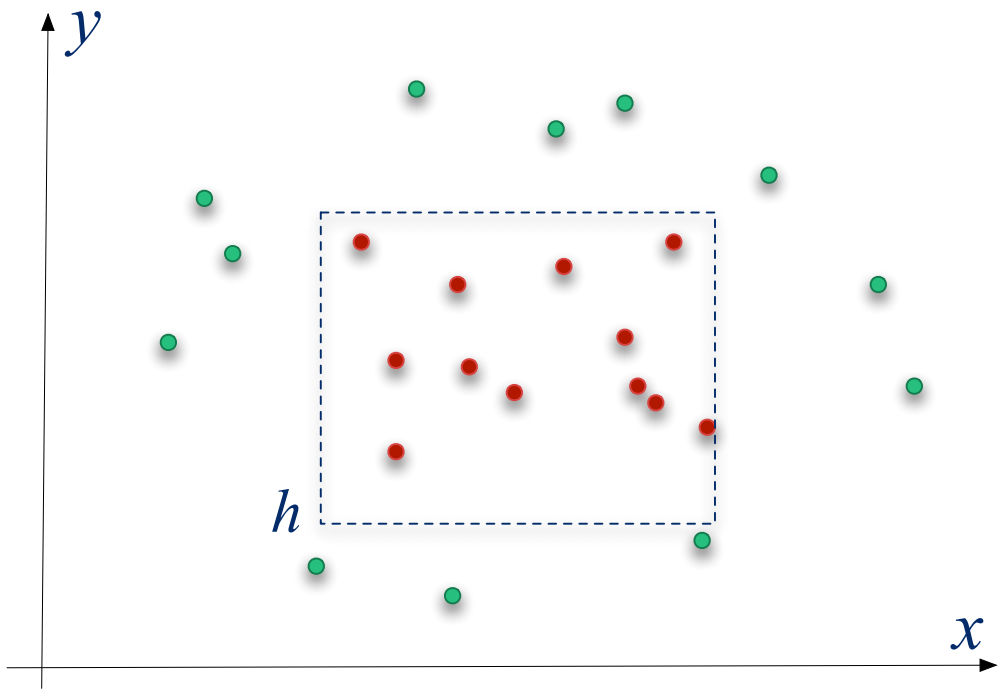
- Comment apprendre ?
  - Choix d'une hypothèse *h*



Espace des versions

Apprentissage de rectangle

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



1ère **étude statistique** de l'induction

Quelle performance ?

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

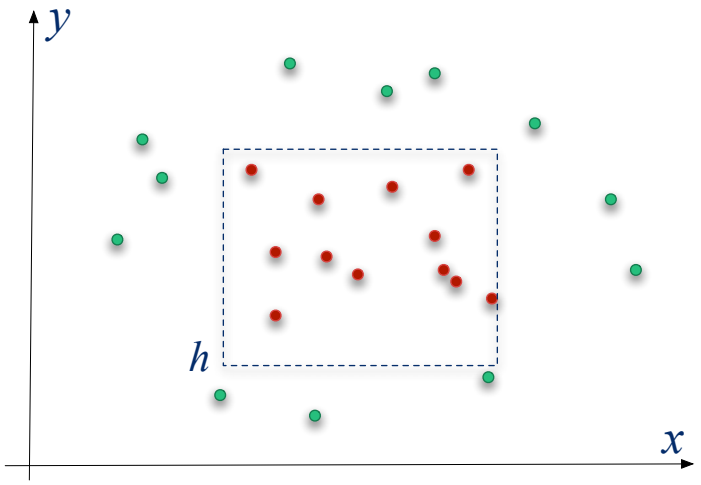
$$\ell\big(h(\mathbf{x}), y\big)$$

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

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

1ère **étude statistique** de l'induction

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

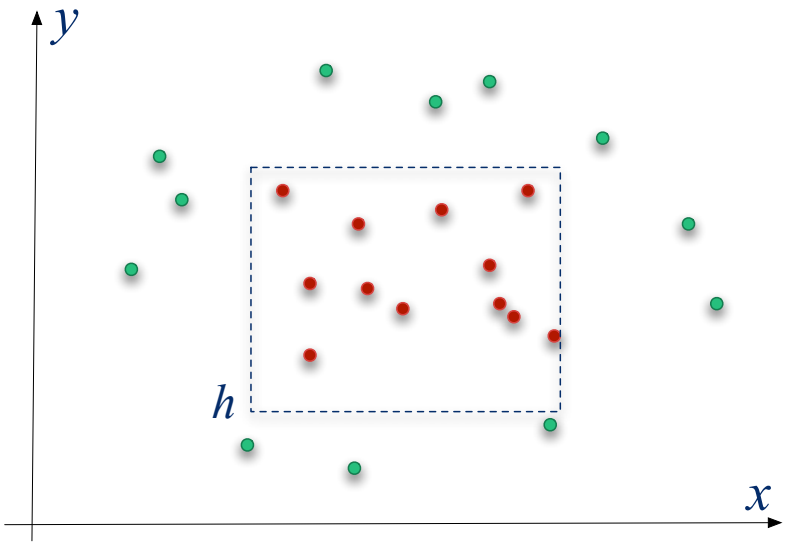


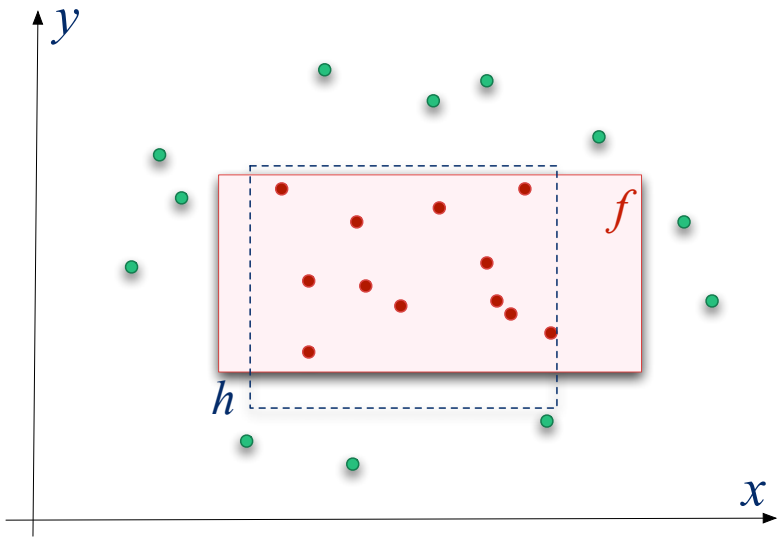
Le « risque empirique »

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

1ère **étude statistique** de l'induction

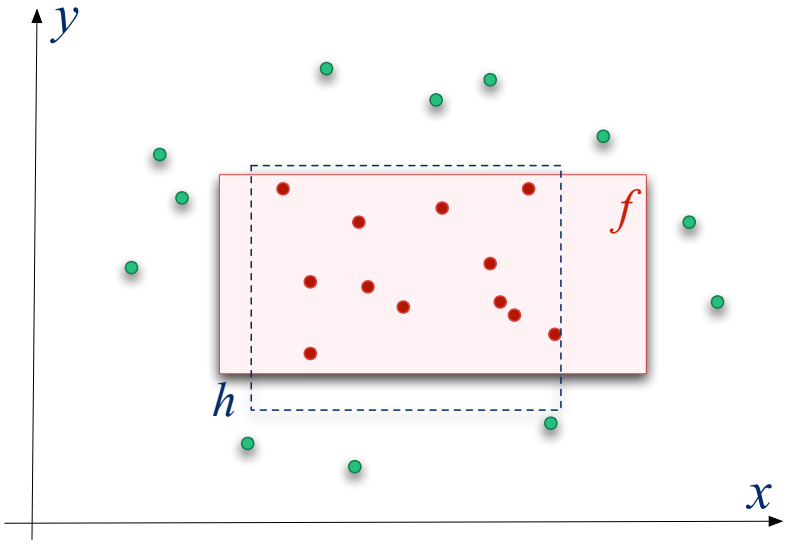
- Stratégie d'apprentissage :
  - choix d'une hypothèse de risque empirique nul (pas d'erreur sur l'échantillon d'apprentissage *S*)
  - Quelle performance attendue pour *h* ?

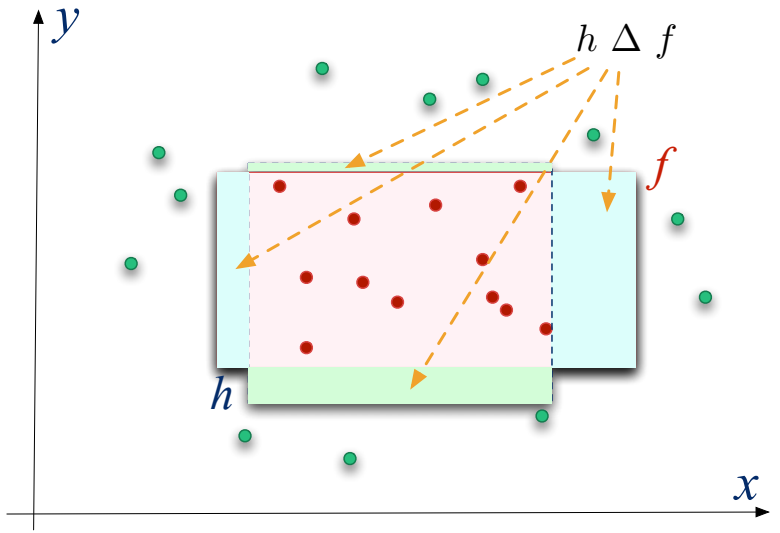




1ère **étude statistique** de l'induction

- 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 $R(h) > \varepsilon$?





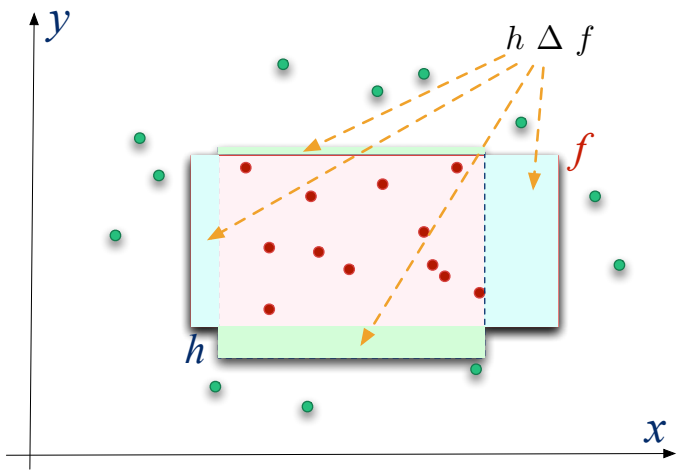
1ère **étude statistique** de l'induction

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

1ère **étude statistique** de l'induction

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

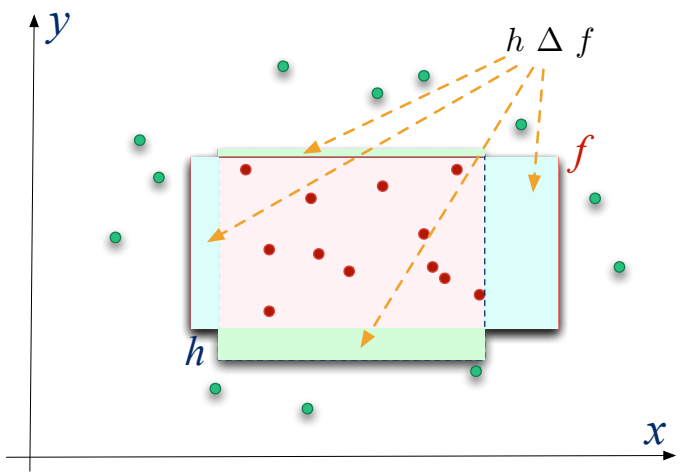
Soit :

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

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

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

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



2ème étude statistique de l'induction

- L'hypothèse est choisie sur la base de *S*

« ***cas réalisable*** »

- On veut donc en fait :

$$\forall\, \varepsilon, \delta \in [0,1] : \quad p^m\left(\exists h \ : \ \ R(h) \geq \varepsilon\right) \ \leq \ \delta$$

- On suppose : $|\mathcal{H}| \ < \ \infty$

Alors :

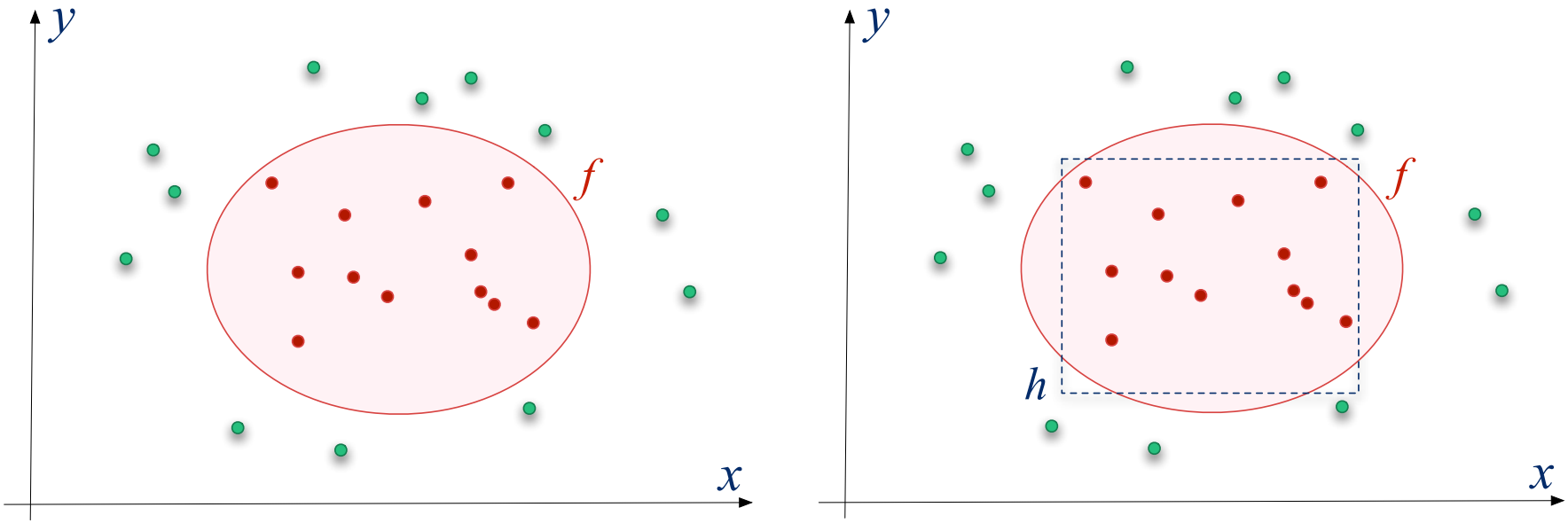
$$|\mathcal{H}|\,(1 \ - \ \varepsilon)^m \ \leq \ |\mathcal{H}|\,e^{-\varepsilon\, m} \ = \ \delta$$

$$-\varepsilon\, m \ \leq \ \ln(\delta) \ - \ \ln(|\mathcal{H}|)$$

$$m \ \geq \ \frac{1}{\varepsilon} \ln \frac{|\mathcal{H}|}{\delta}$$

Cas « non réalisable »

- $\mathcal{H} \neq \mathcal{F}$



Il n'existe plus d'hypothèse de **risque réel** nul.

Pas de garantie non plus de trouver une hypothèse de **risque empirique** nul.

3ème étude statistique de l'induction

Théorème 1 (Inégalité de Hoeffding). *Si les ξ_i sont des variables aléatoires, tirées* ***indépendamment*** *et selon une* ***même distribution*** *et prenant leur valeur dans l'intervalle $[a, b]$, alors :*

$$P\left(\left|\frac{1}{m}\sum_{i=1}^{m}\xi_i - \mathbb{E}(\xi)\right| \geq \varepsilon\right) \leq 2\exp\left(-\frac{2\,m\,\varepsilon^2}{(b-a)^2}\right)$$

Appliquée au risque empirique et au risque réel, cette inégalité nous donne :

$$P\big(|R_{\text{Emp}}(h) - R_{\text{Réel}}(h)| \geq \varepsilon\big) \leq 2\exp\big(-\frac{2\,m\,\varepsilon^2}{(b-a)^2}\big) \qquad (1)$$

si la fonction de perte ℓ est définie sur l'intervalle $[a, b]$.

« $\mathcal{H}$ fini »

$$P^m[\exists h \in \mathcal{H} : R_{\text{Réel}}(h) - R_{\text{Emp}}(h) > \varepsilon] \leq \sum_{i=1}^{|\mathcal{H}|} P^m[R_{\text{Réel}}(h^i) - R_{\text{Emp}}(h^i) > \varepsilon]$$

$$\leq |\mathcal{H}|\exp(-2\,m\,\varepsilon^2) = \delta$$

en supposant ici que la fonction de perte ℓ prend ses valeurs dans l'intervalle $[0, 1]$.

3ème **étude statistique** de l'induction

- On en tire :

$$\varepsilon = \sqrt{\frac{\log |\mathcal{H}| + \log \frac{1}{\delta}}{2\,m}} \qquad \text{et} \qquad m \geq \frac{\log |\mathcal{H}| + \log \frac{1}{\delta}}{2\,\varepsilon^2}$$

- Au lieu de (« cas réalisable ») :

$$\varepsilon = \frac{\log |\mathcal{H}| + \log \frac{1}{\delta}}{m} \qquad \text{et} \qquad m \geq \frac{\log |\mathcal{H}| + \log \frac{1}{\delta}}{\varepsilon}$$

Lien entre risque réel et risque empirique

- $\mathcal{H}$ fini, cas réalisable

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

- $\mathcal{H}$ fini, cas **non** réalisable

$$\forall h \in \mathcal{H}, \forall \delta \leq 1 : \quad P^m \left[R_{\text{Réel}}(h) \leq R_{\text{Emp}}(h) + \sqrt{\frac{\log |\mathcal{H}| + \log \frac{1}{\delta}}{2\,m}} \right] > 1 - \delta$$

4ème **étude statistique** de l'induction

- Cas non réalisable et $\mathcal{H}$ non finie

Comment faire ?

– Principe général :

1. **Réduire** l'étude du cas **infini** à celui de l'analyse d'un **ensemble fini d'hypothèses**
2. **Mesurer** à quel point, **pour n'importe quel échantillon** S de points étiquetés, on peut trouver une **hypothèse** de $\mathcal{H}$ **pouvant s'adapter** à S

4ème étude statistique de l'induction

- La ***complexité de Rademacher***

  – Soit : $\mathcal{S} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\} = \{z_1, \ldots, z_m\}$

  – Mesure de la corrélation entre les prédictions et les étiquettes :

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

  – L'hypothèse maximisant cette corrélation :

$$h^* = \underset{h \in \mathcal{H}}{\text{ArgMax}} \frac{1}{m} \sum_{i=1}^{m} y_i \, h(\mathbf{x}_i)$$

  – Mesure caractérisant l'adéquation de $\mathcal{H}$ avec S.

4ème étude statistique de l'induction

- La ***complexité de Rademacher*** (suite)

  – Supposons que les étiquettes soient choisies au hasard

    • Chaque y_i est remplacée par une variable aléatoire σ_i = -1 ou +1

  – On peut mesurer comment $\mathcal{H}$ peut s'ajuster à ce bruit par l'espérance :

$$R_{\mathcal{S}}(\mathcal{H}) \;=\; \mathbb{E}_{\sigma}\left[\operatorname*{Max}_{h\in\mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \, h(\mathbf{x}_i)\right]$$

On en tire la borne :

$$\forall h \in \mathcal{H}, \forall \delta \leq 1 : \quad P^m\left[R_{\text{Réel}}(h) \;\leq\; R_{\text{Emp}}(h) \;+\; R_{\mathcal{S}}(\mathcal{H}) \;+\; 3\sqrt{\frac{\log \frac{2}{\delta}}{2\,m}}\right] \;>\; 1-\delta$$

4ème étude statistique de l'induction

- Mesure par la ***fonction de croissance***
  - **Critère** purement **combinatoire**, ne dépendant pas de la distribution $\mathbf{P}_{XY}$
  - Nombre maximal de manière distinctes d'étiqueter *m* points de $\mathcal{X}$ en utilisant une hypothèse de $\mathcal{H}$

$$\forall m \in \mathbb{N}, \quad \Pi_{\mathcal{H}}(m) \;=\; \max_{\{\mathbf{x}_1,\ldots,\mathbf{x}_m\}\subseteq\mathcal{X}} \left| \left\{ \big(h(\mathbf{x}_1),\ldots,h(\mathbf{x}_m)\big) \;:\; h \in \mathcal{H} \right\} \right|$$

On en tire la borne :

$$\forall h \in \mathcal{H}, \forall \delta \leq 1 : \quad P^m \left[R_{\text{Réel}}(h) \;\leq\; R_{\text{Emp}}(h) \;+\; \sqrt{\frac{2\log \Pi_{\mathcal{H}}(m)}{m}} \;+\; \sqrt{\frac{\log \frac{1}{\delta}}{2\,m}} \right] \;>\; 1-\delta$$

4ème étude statistique de l'induction

- Mesure par la ***dimension de Vapnik-Chervonenkis***

  – **Critère** purement **combinatoire**, ne dépendant pas du nombre d'exemples

  – Taille du plus grand ensemble de points pouvant être étiquetés de n'importe quelle manière par les hypothèses tirées de H

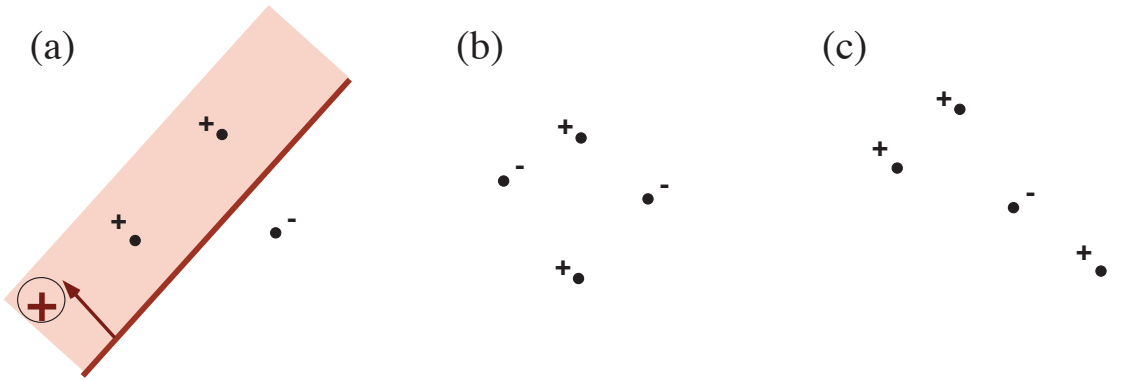
$$d_{VC}(\mathcal{H}) \;=\; \max\{m \;:\; \Pi_{\mathcal{H}}(m) \;=\; 2^m\}$$

On en tire la borne :

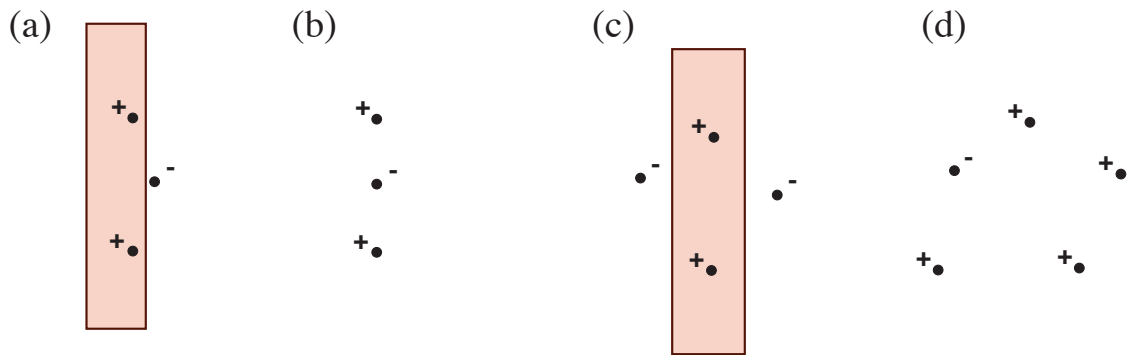
$$\forall h \in \mathcal{H}, \forall \delta \leq 1 : \quad P^m \left[R_{\text{Réel}}(h) \;\leq\; R_{\text{Emp}}(h) \;+\; \sqrt{\frac{8\, d_{VC}(\mathcal{H}) \log \frac{2\,e\,m}{d_{VC}(\mathcal{H})} + 8 \log \frac{4}{\delta}}{m}} \right] \;>\; 1 - \delta$$

VC dim : illustration

- d_{VC}(séparateurs linéraires) = ?



- d_{VC}(rectangles) = ?



Étude statistique de l'induction

1. Ces mesures de capacité **ne dépendent pas de la dimension** de $\mathcal{X}$!!

2. La complexité de Rademacher **de l'enveloppe convexe** d'espaces $\mathcal{H}$ n'est **pas plus grande** que celle de $\mathcal{H}$!

   – Intéressant pour les méthodes d'ensemble et les méthodes collaboratives

Quelques remarques

1. On n'a pas vraiment parlé d'***algorithme d'apprentissage*** !?
   - Où est l'apprentissage ?

2. Importance cruciale de l'hypothèse de ***stationnarité*** et d'***indépendance*** : tirage i.i.d.

3. Importance du **choix de** $\mathcal{H}$
   - Mais étude « contre toute distribution »

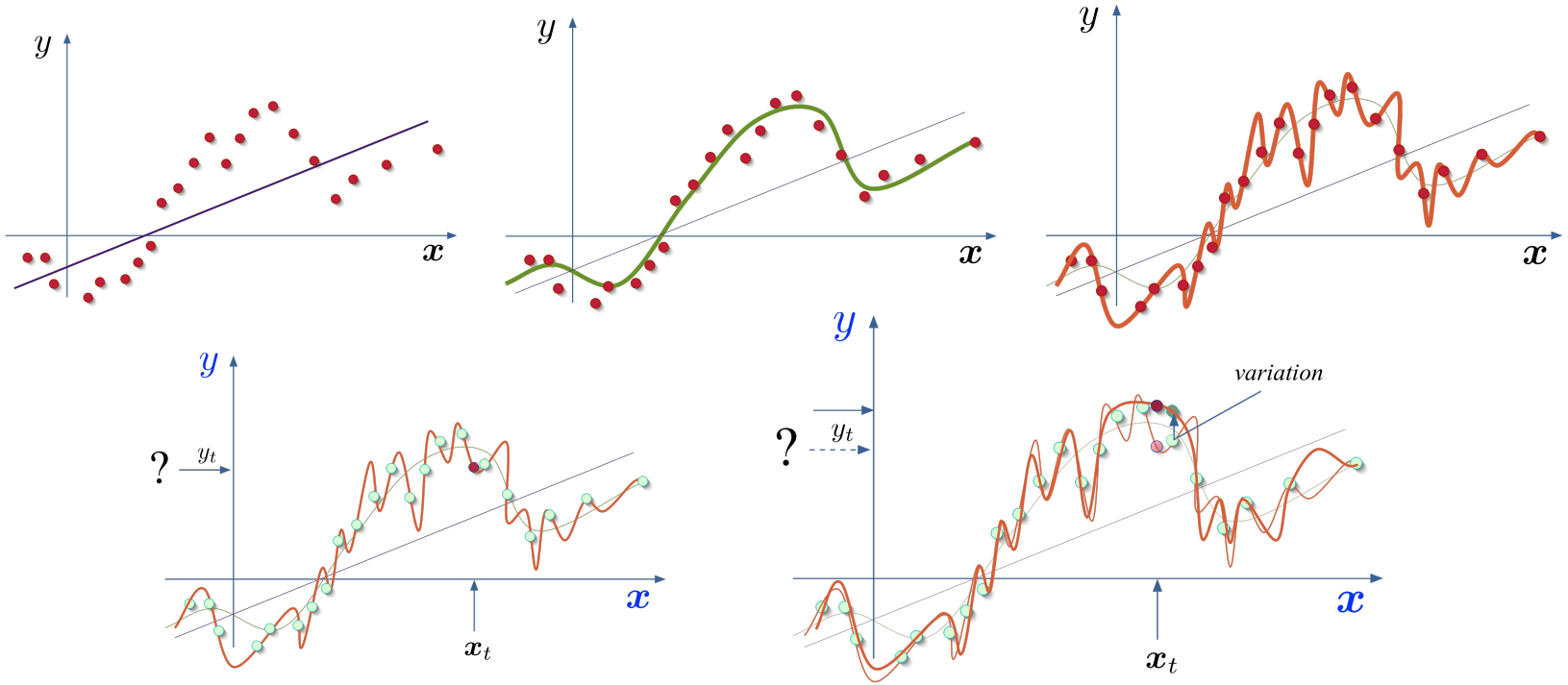
     Ne présuppose pas un certain type de distribution

Fondamentalement


J. S. Hadamard, 1865-1963

1. Apprentissage = problème inverse mal-posé

- $f \xrightarrow{\text{tirage i.i.d.}} \mathcal{S}$
- Induction $\equiv$ trouver f à partir de $\mathcal{S}$
- mais $\mathcal{S} \underset{\delta}{\longrightarrow} \mathcal{S}_\delta$ peut conduire à f très différent de f_δ



Fondamentalement

1. **Apprentissage = problème inverse mal-posé**
   - $f \xrightarrow{\text{tirage i.i.d.}} \mathcal{S}$
   - Induction $\equiv$ trouver f à partir de $\mathcal{S}$
   - mais $\mathcal{S} \underset{\delta}{\longrightarrow} \mathcal{S}_\delta$ peut conduire à f très différent de f_δ
2. Le principe de minimisation du risque empirique est **naïf**
3. À remplacer par un **principe de minimisation d'un risque régularisé**

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

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

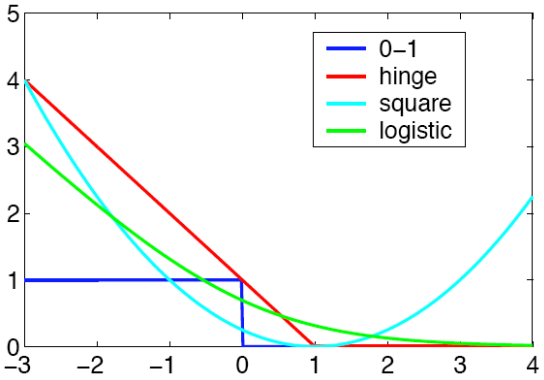
Le critère inductif

- **Critère de substitution** à la place du risque réel

1. **Risque empirique régularisé**

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

2. **Fonction de perte de substitution**
   (surrogate function)



3. Contrainte sur l'**espace des hypothèses** considéré $\mathcal{H}$

L'induction supervisée

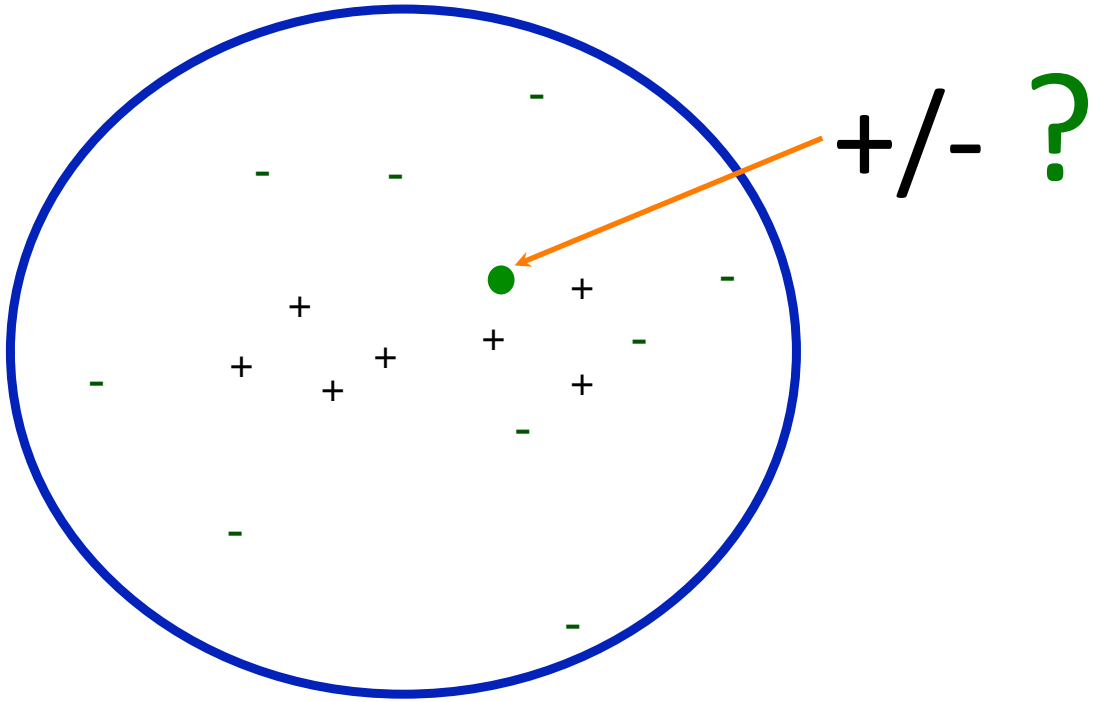
en trois questions

Trois ingrédients essentiels

1. Choix de ***l'espace des hypothèses*** $\mathcal{H}$
   - Contrôler sa « capacité »
2. Choix du ***critère à optimiser*** $R(h)$
   - Risque empirique régularisé
3. Choix de la ***méthode d'exploration*** de $\mathcal{H}$
   - Plus facile si $R(.)$ convexe

Panorama de quelques espaces d'hypothèses

Apprendre **sans** espace d'hypothèses



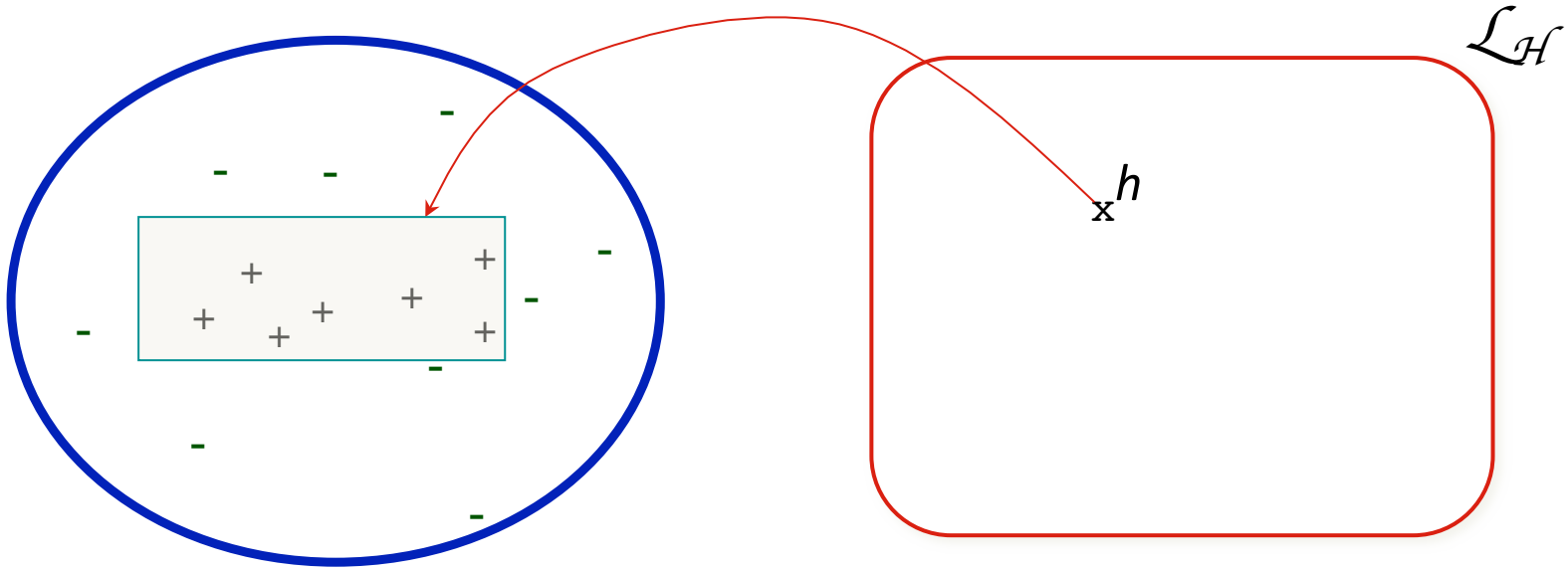
Espace des exemples : $\mathcal{X}$

- Méthodes par *plus proches voisins*
- Nécessité d'une *notion de distance*

➠ Hypothèse de continuité dans $\mathcal{X}$

Le rôle de l'espace des hypothèses

Cas particulier de l'*apprentissage de concepts*

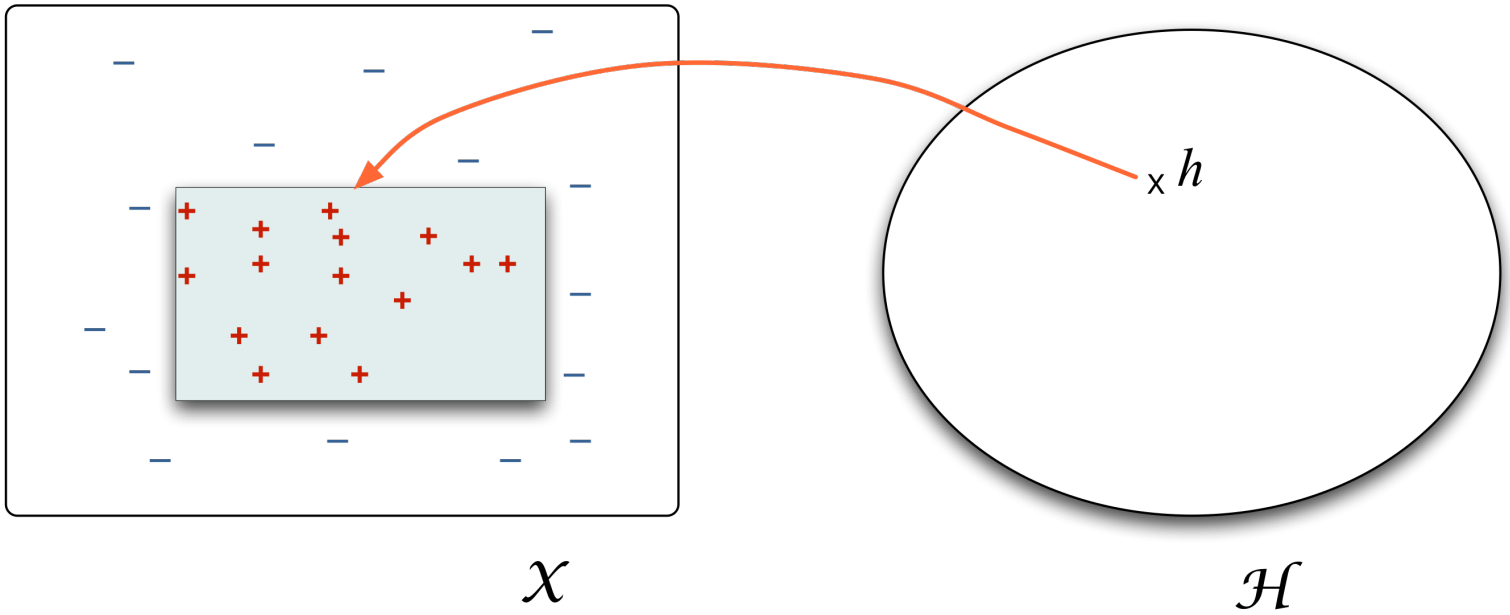


Espace des exemples : $\mathcal{X}$ Espace des hypothèses : $\mathcal{H}$

➥ Comment choisir l'espace des hypothèses (i.e. le langage $\mathcal{L}_{\mathcal{H}}$) ?

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

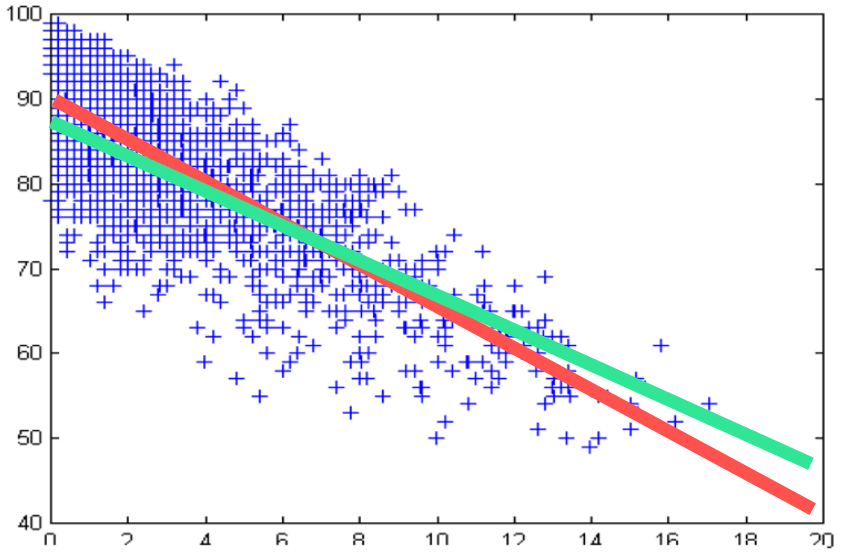
- Introduit explicitement l'idée de **recherche dans un espace d'hypothèses**

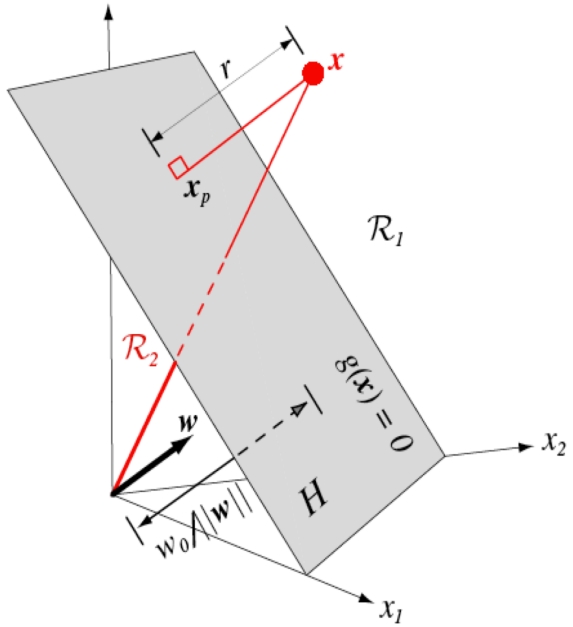


Types d' espaces d'hypothèses

$\mathcal{H}$ = combinaison de fonctions de base

Combinaisons linéaires de fonctions de base

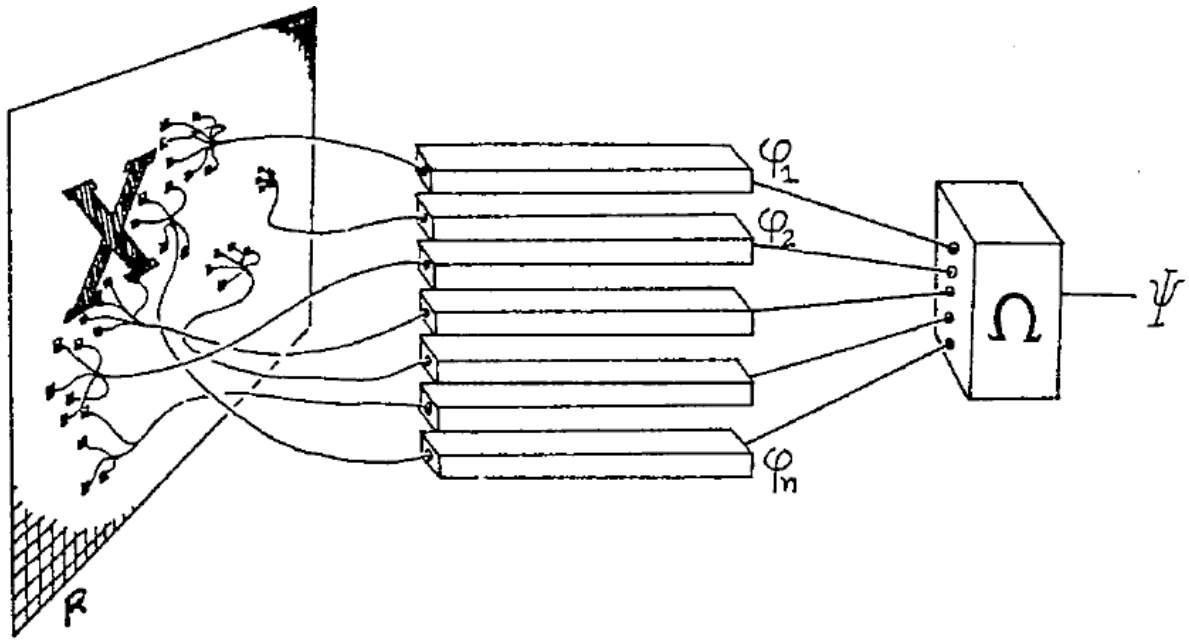




$$h(\mathbf{x}) = \sum_{i=1}^{n} w_i \ \phi_i(\mathbf{x})$$

Combinaisons linéaires de fonctions de base

- Le **Perceptron** : Frank Rosenblatt (1958 – 1962)



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

Le perceptron : algorithme de gradient

- Méthode d'exploration de $\mathcal{H}$
  - Recherche par gradient
    - Minimisation de la fonction d'erreur

$$\mathbf{w}(t+1) = \mathbf{w}(t) - \eta \, \nabla_{\mathbf{w}} E(\mathbf{w}(t))$$

  - Algorithme :

si la forme est correctement classée : ne rien faire

sinon : $\boldsymbol{w}(t+1) = \boldsymbol{w}(t) - \eta \; \boldsymbol{x}_i \; u_i$

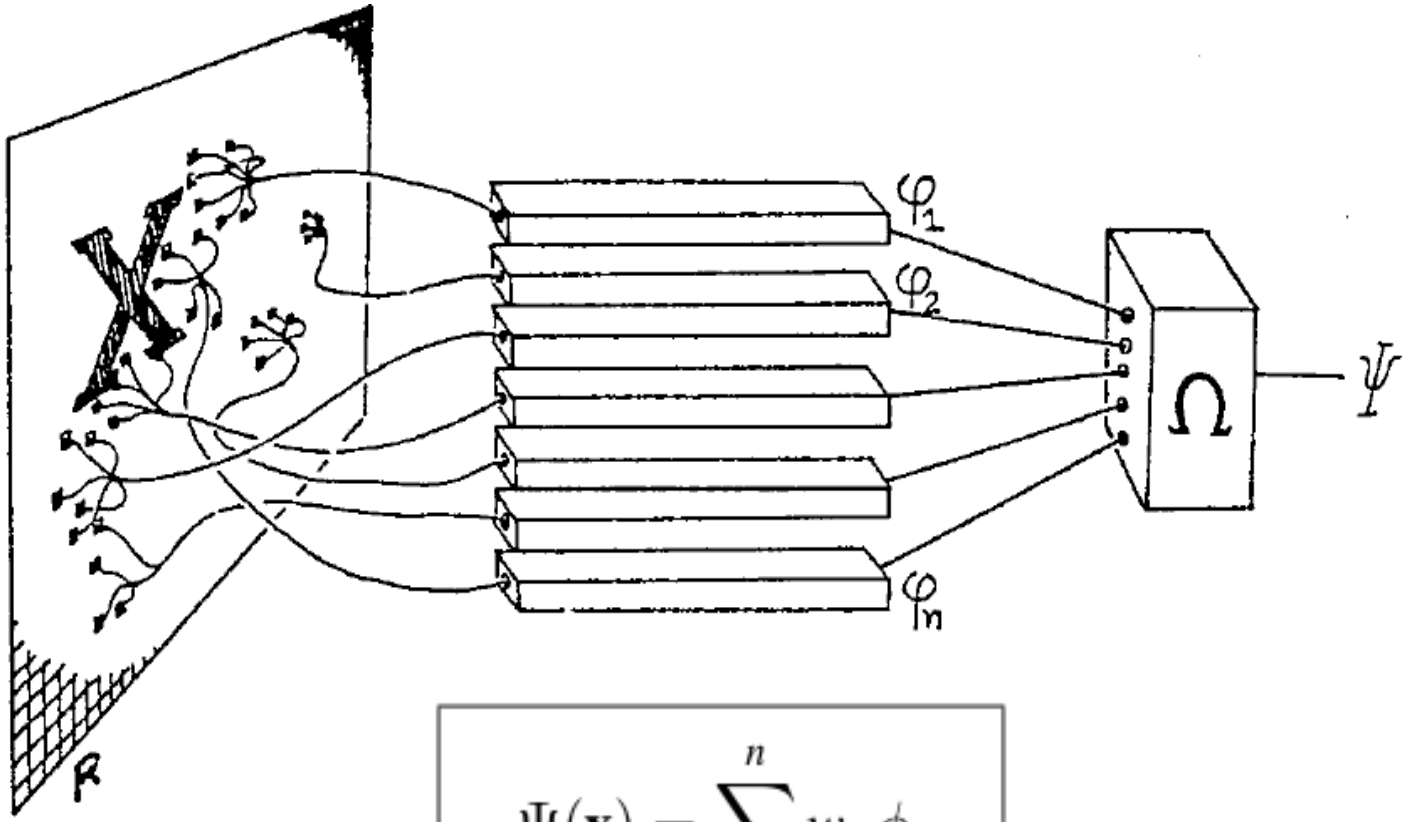
boucler sur les formes d'apprentissage jusqu'à critère d'arrêt

Combinaisons linéaires de fonctions de base

- Souvent possible d'avoir un **optimum unique**
  - Critère d'**optimisation convexe**
- **Interprétabilité** facilitée
  - Critère d'optimisation favorisant la **parcimonie**
    - Norme L1 (e.g. LASSO)
- **Algorithmes**
  - **Gradient**
  - Itérativement pour **chaque terme**
    - E.g. gradient conjugué, boosting, …

Le Perceptron

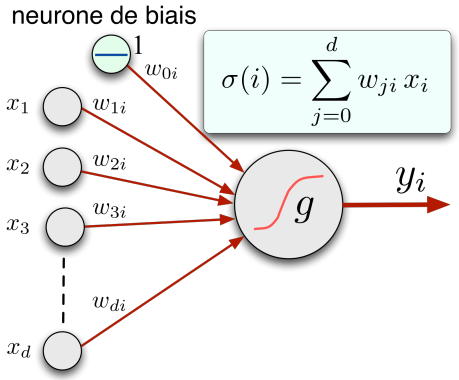
– Rosenblatt (1958-1962)

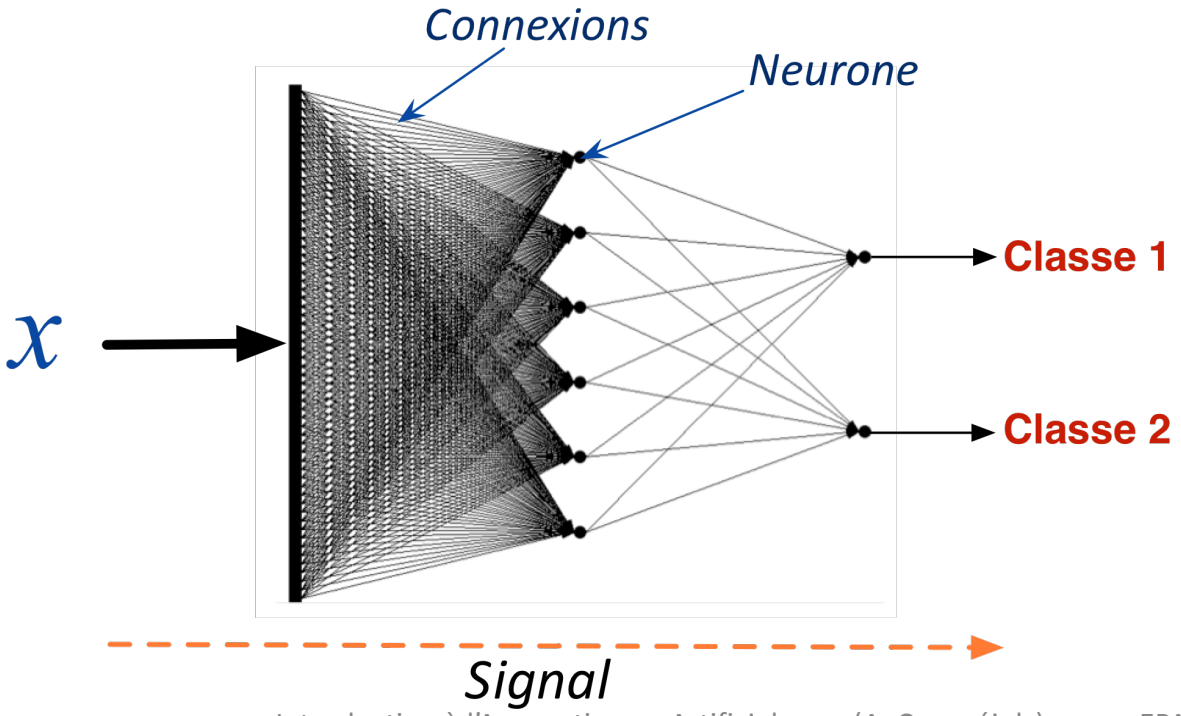


$$\Psi(\mathbf{x}) = \sum_{i=1}^{n} w_i \, \phi_i$$

Combinaisons **non** linéaires de fonctions de base

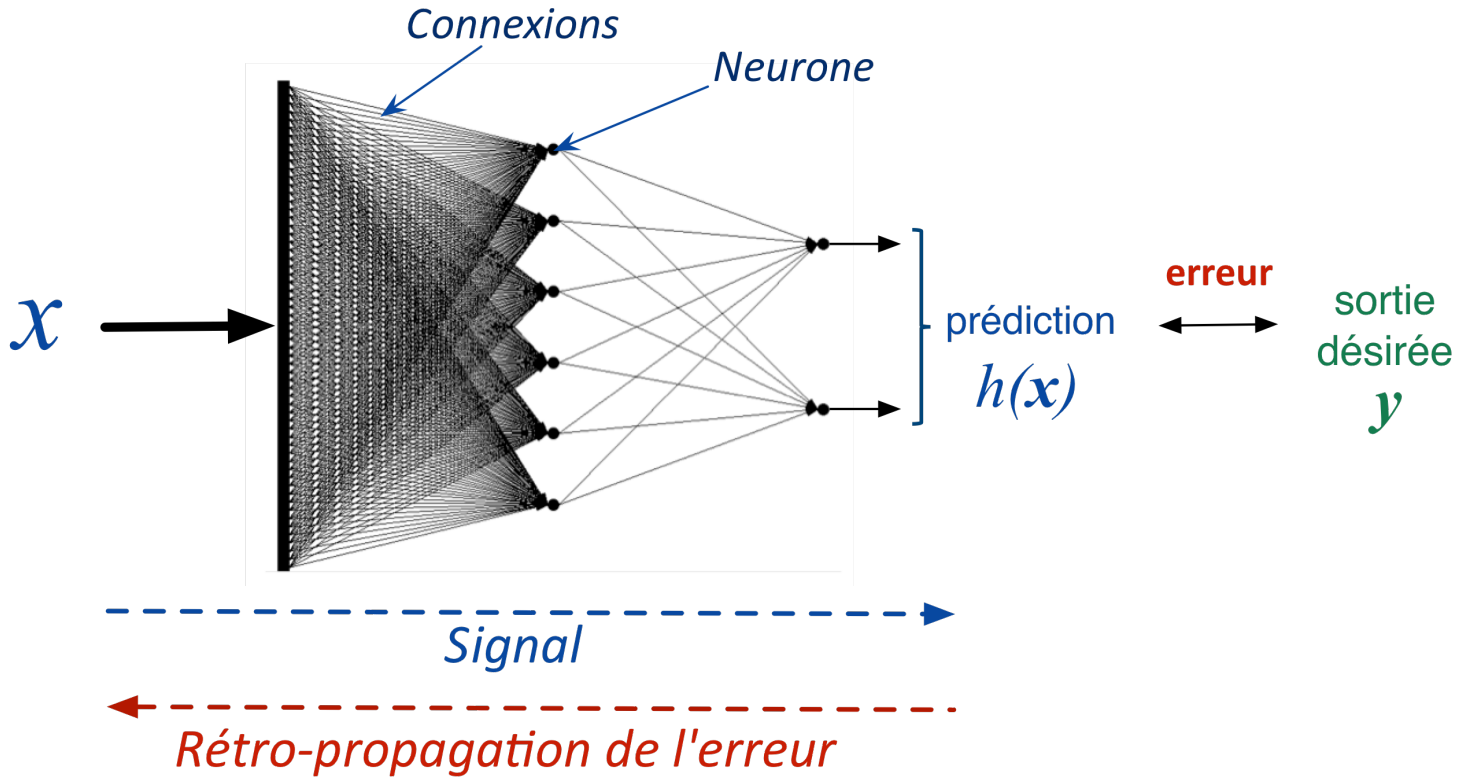
- Une réponse :
  - À la recherche d'une bonne représentation



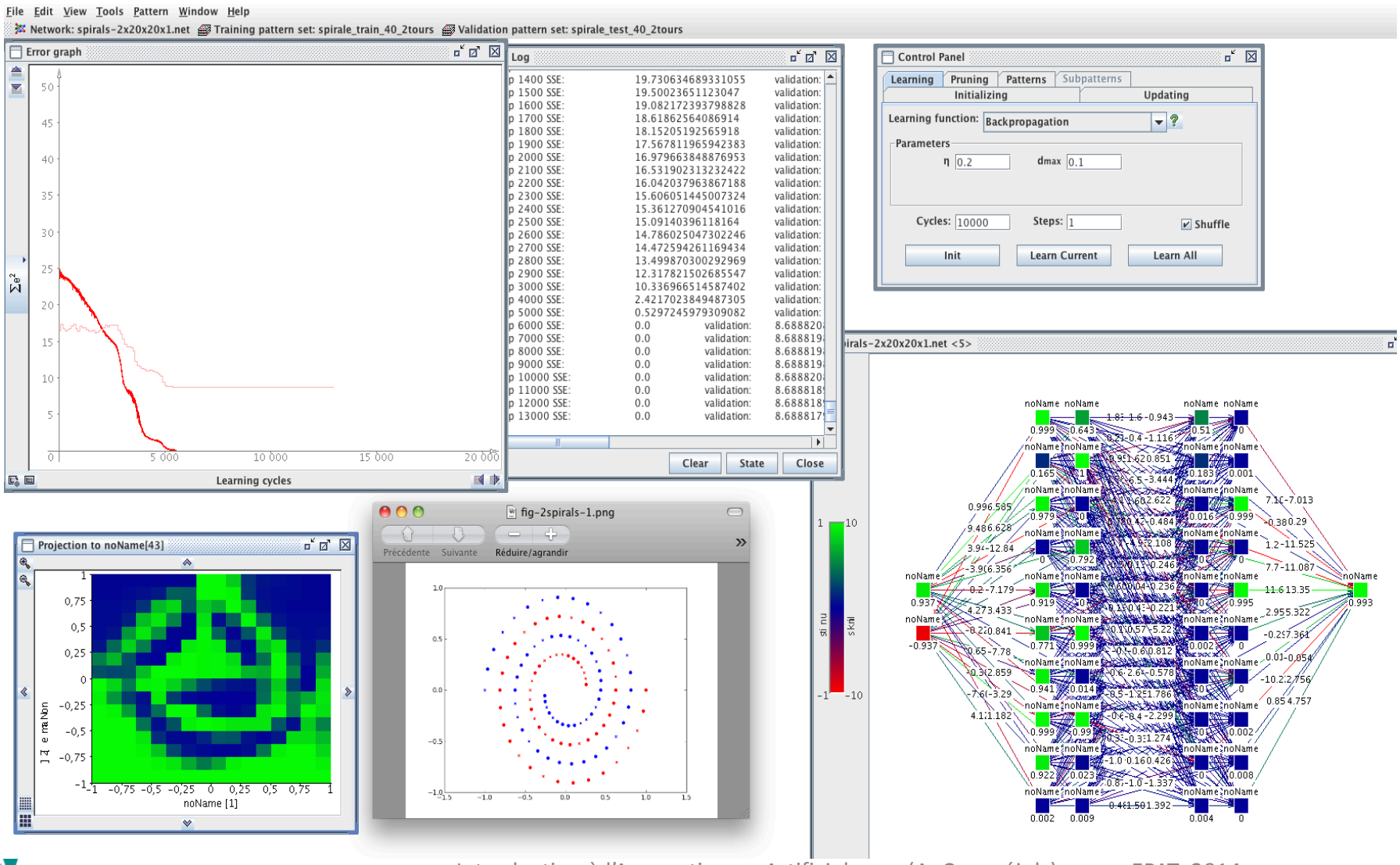


Combinaisons **non** linéaires de fonctions de base

- Questions :
  - Comment apprendre les **paramètres** (poids des connexions) ?
  - Comment déterminer l'**architecture** du réseau ?



Illustration



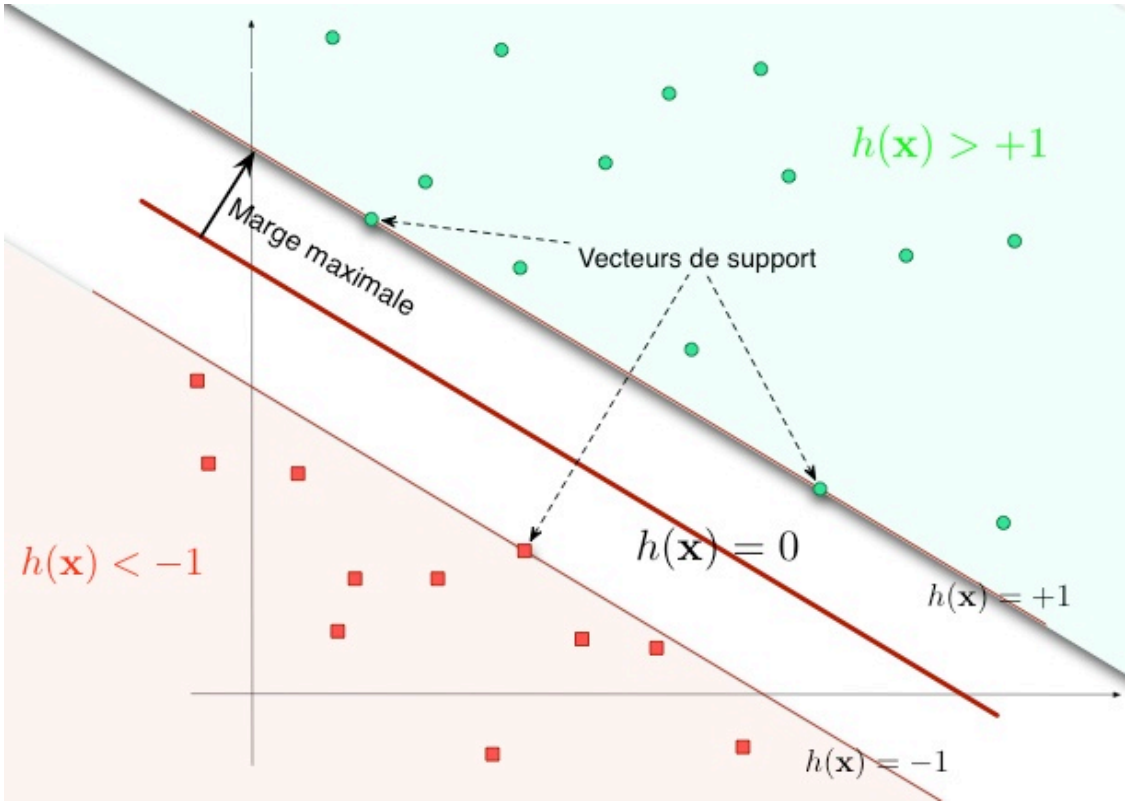
Combinaisons **non** linéaires de fonctions de base

- Souvent des optima multiples
  - Optimisation difficile
- **Interprétabilité** pas immédiate
  - Interactions multiples
- **Algorithmes**
  - **Gradient**

$\mathcal{H}$ = modèles non paramétriques

définis à partir des exemples d'apprentissage

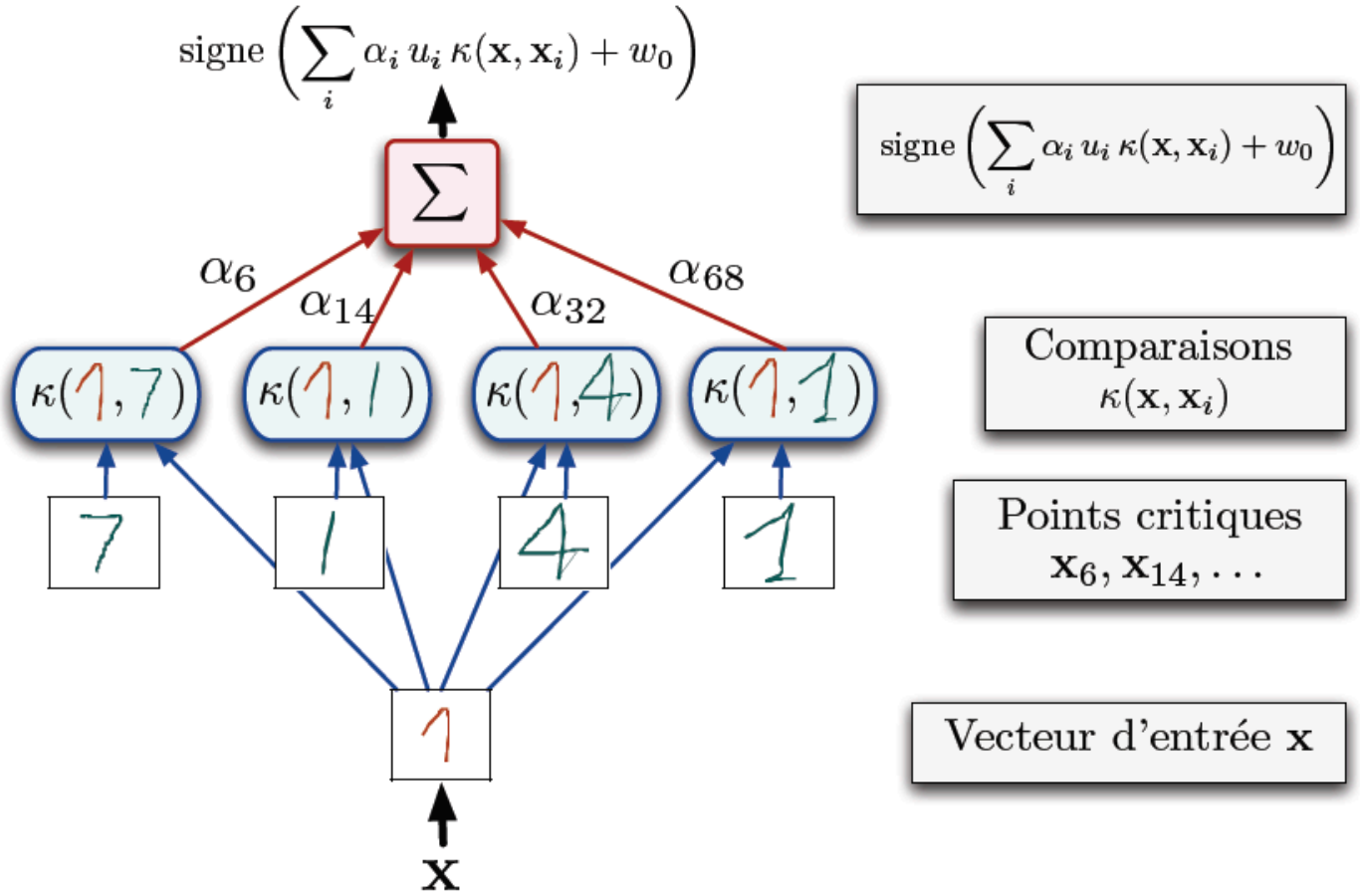
Les séparateurs à Vastes Marges (SVM)



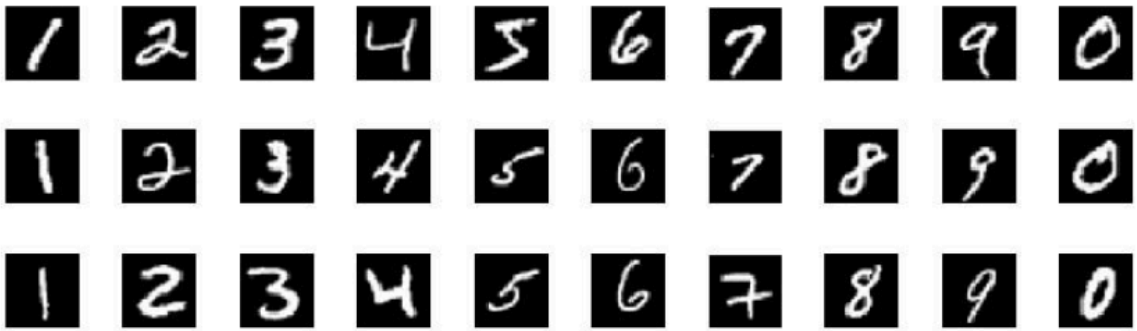
$$h^\star(\mathbf{x}) \;=\; \text{sign}\left\{ \sum_{i \in \mathcal{P}_S} \alpha_i^\star \, y_i \, \langle \phi(\mathbf{x}), \phi(\mathbf{x}_i) \rangle \;+\; w_0^\star \right\}$$

Hypothèses exprimées à partir des exemples

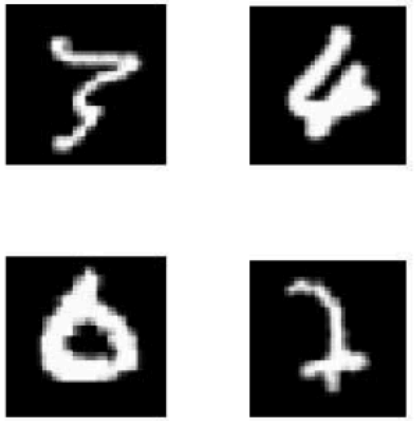
- **Séparateurs à Vastes Marges** (SVM)



Hypothèses exprimées à partir des exemples



Data



Support vectors

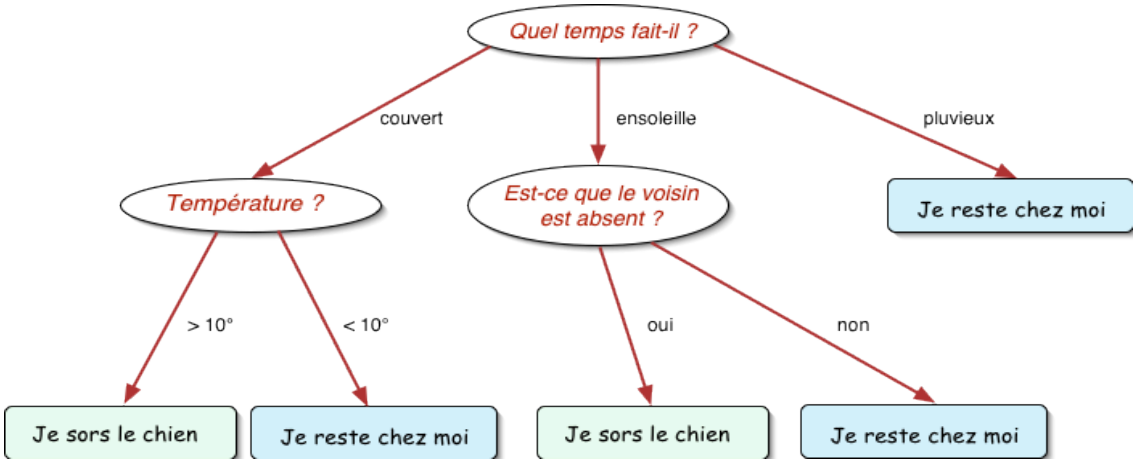
Hypothèses exprimées à partir des exemples

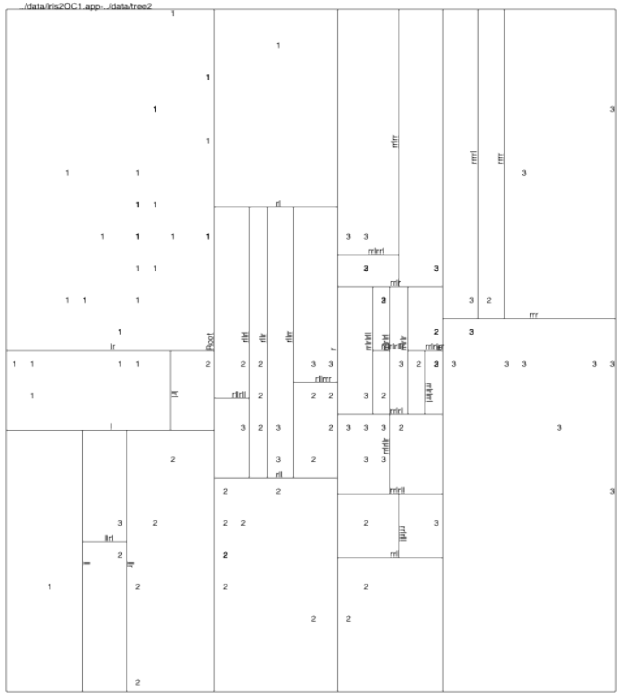
- Recherche de critères convexes
  - Optimisation « facile »
- **Interprétabilité** pas immédiate
  - À partir de cas non typiques (mais « near-misses »)
- **Algorithmes**
  - **Gradient sophistiqué**

$\mathcal{H}$ = par apprentissage constructif

Modèles constructifs

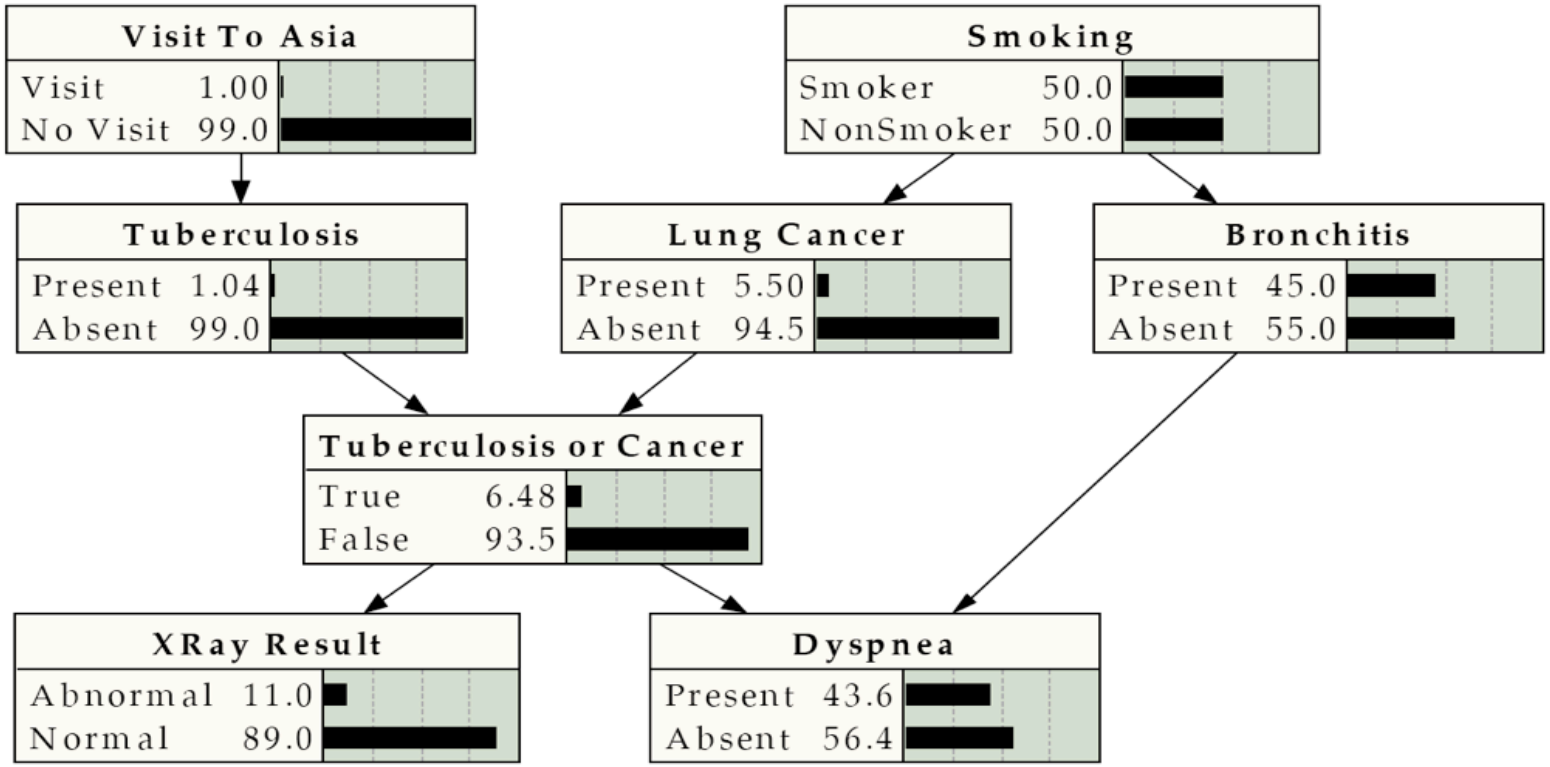
- Arbre de décision



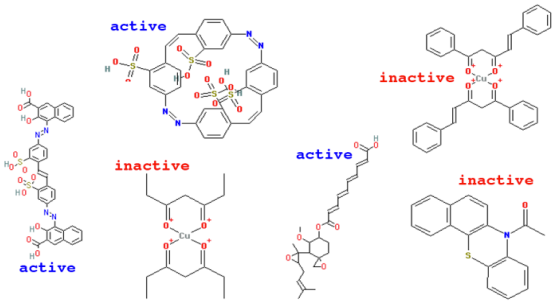


Modèles constructifs

- Modèles graphiques



Modèles « symboliques »



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

$$\varphi_1 : \; anm(x_3, [195, 22, 3, 27, 38, 40, 92]) \wedge \neg chrg(x_3, [-0.2, 0.2]) \wedge$$
$$anm(x_4, [195, 22, 3, 38, 40, 29, 92]) \wedge \neg type(x_4, [O]) \wedge \neg chrg(x_4, [-0.2]) \wedge$$
$$(x_1 < x_2) \wedge (x_1 < x_3) \wedge (x_1 < x_4) \wedge (x_2 < x_3) \wedge (x_2 < x_4) \wedge (x_3 < x_4) \wedge$$
$$bound(x_3, x_4) \;\rightarrow\; mutagenic,$$

$$\varphi_2 : \; \neg chrg(x_1, [-0.2]) \wedge \neg type(x_2, [N]) \wedge \neg anm(x_3, [22]) \wedge \neg chrg(x_3, [-0.6, -0.4]) \wedge$$
$$\neg type(x_4, [H, N, O]) \wedge (x_1 < x_2) \wedge (x_1 < x_3) \wedge (x_1 < x_4) \wedge (x_2 < x_3) \wedge$$
$$bound(x_2, x_3) \wedge (x_2 < x_4) \wedge (x_3 < x_4) \wedge bound(x_3, x_4) \;\rightarrow\; mutagenic,$$

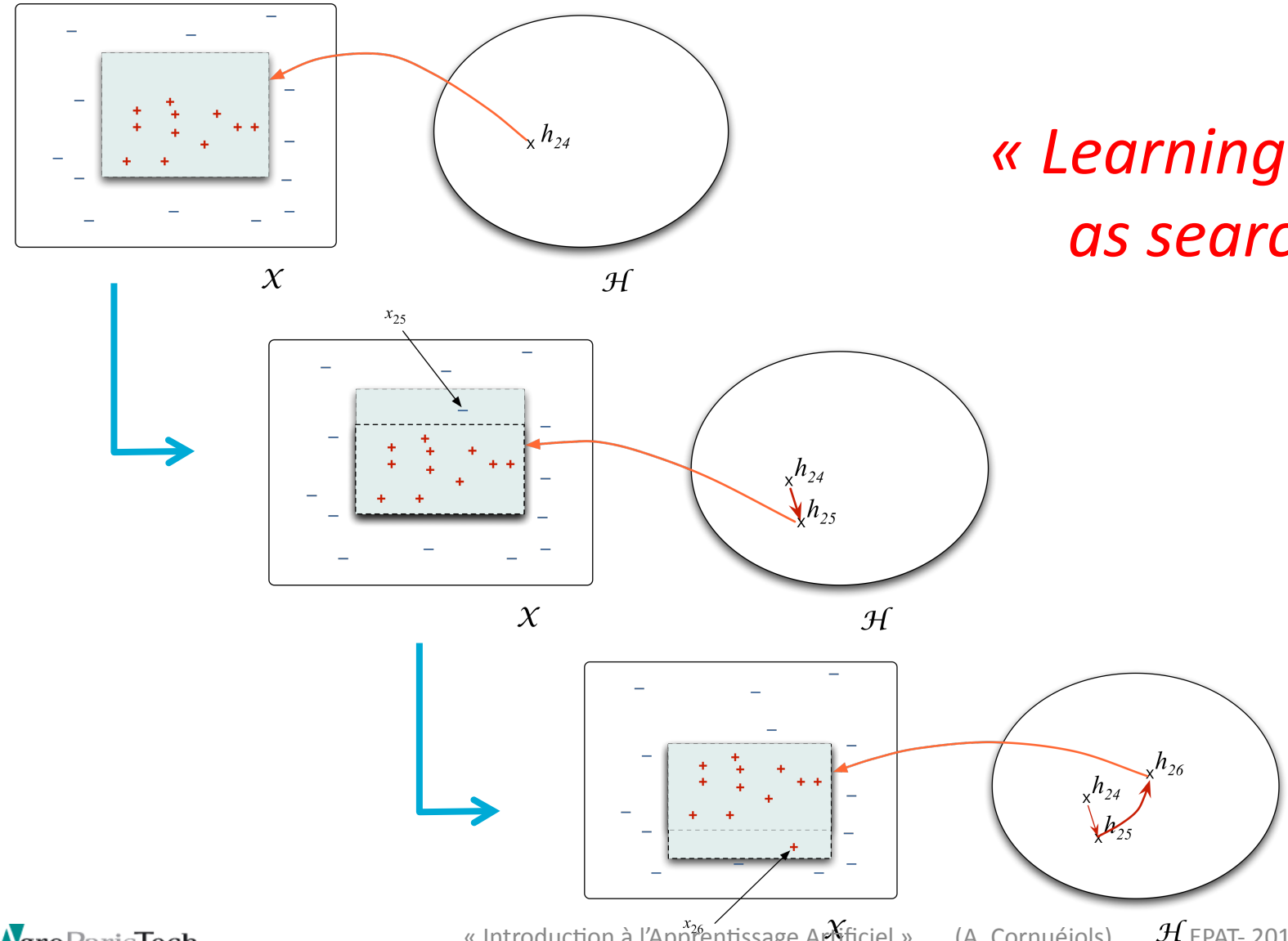
$$\varphi_3 : \; anm(x_1, [195, 38, 29, 92]) \wedge chrg(x_1, [-0.8 \div 0.6]) \wedge \neg type(x_3, [C]) \wedge \neg chrg(x_3, [0.0]) \wedge$$
$$anm(x_4, [195, 22, 3, 27, 38, 29, 92]) \wedge \neg type(x_4, [N]) \wedge (x_1 < x_2) \wedge (x_1 < x_3) \wedge$$
$$(x_1 < x_4) \wedge (x_2 < x_3) \wedge (x_2 < x_4) \wedge (x_3 < x_4) \;\rightarrow\; mutagenic,$$

$$\varphi_4 : \; anm(x\ 1, [195, 3, 27, 38, 40, 29, 92]) \wedge \neg type(x_1, [H]) \wedge \neg chrg(x_1, [-0.2])$$
$$\neg anm(x_3, [40]) \wedge anm(x_4, [195, 22, 27, 38, 40, 29, 92]) \wedge \neg type(x_4, [H, N])$$
$$(x_1 < x_2) \wedge \neg bound(x_1, x2) \wedge (x_1 < x_3) \wedge (x_1 < x_4) \wedge (x_2 < x_3) \wedge (x_2 < x_4) \wedge$$
$$bound(x_{,}x_{)} \wedge (x3 < x4) \;\rightarrow\; mutagenic.$$

Apprentissage constructif

- Souvent des optima multiples
  - E.g. *Espace des versions*
- **Interprétabilité** aisée
  - « Fait pour »
  - Plus proche de la cognition humaine
- **Algorithmes**
  - Mise en jeu d'**opérateurs spécifiques**
    - E.g. *Spécialisation / généralisation*
  - Algorithmes d'**exploration** +/- heuristiques +/- ad hoc

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



« Learning as search »

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

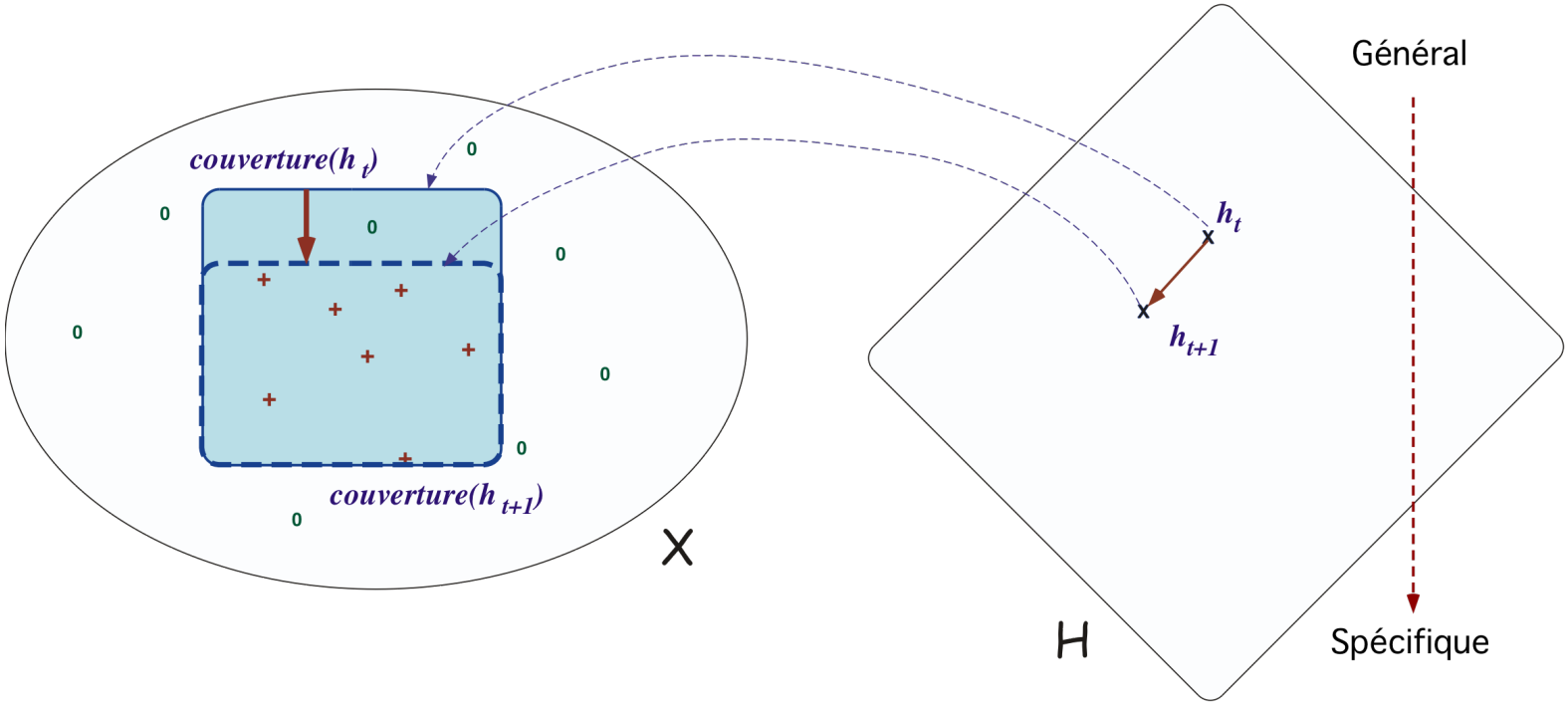


Figure: *La relation d'inclusion dans $\mathcal{X}$ induit la relation de généralisation dans $\mathcal{H}$. Ici, $h_{t+1} \preceq h_t$.*

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

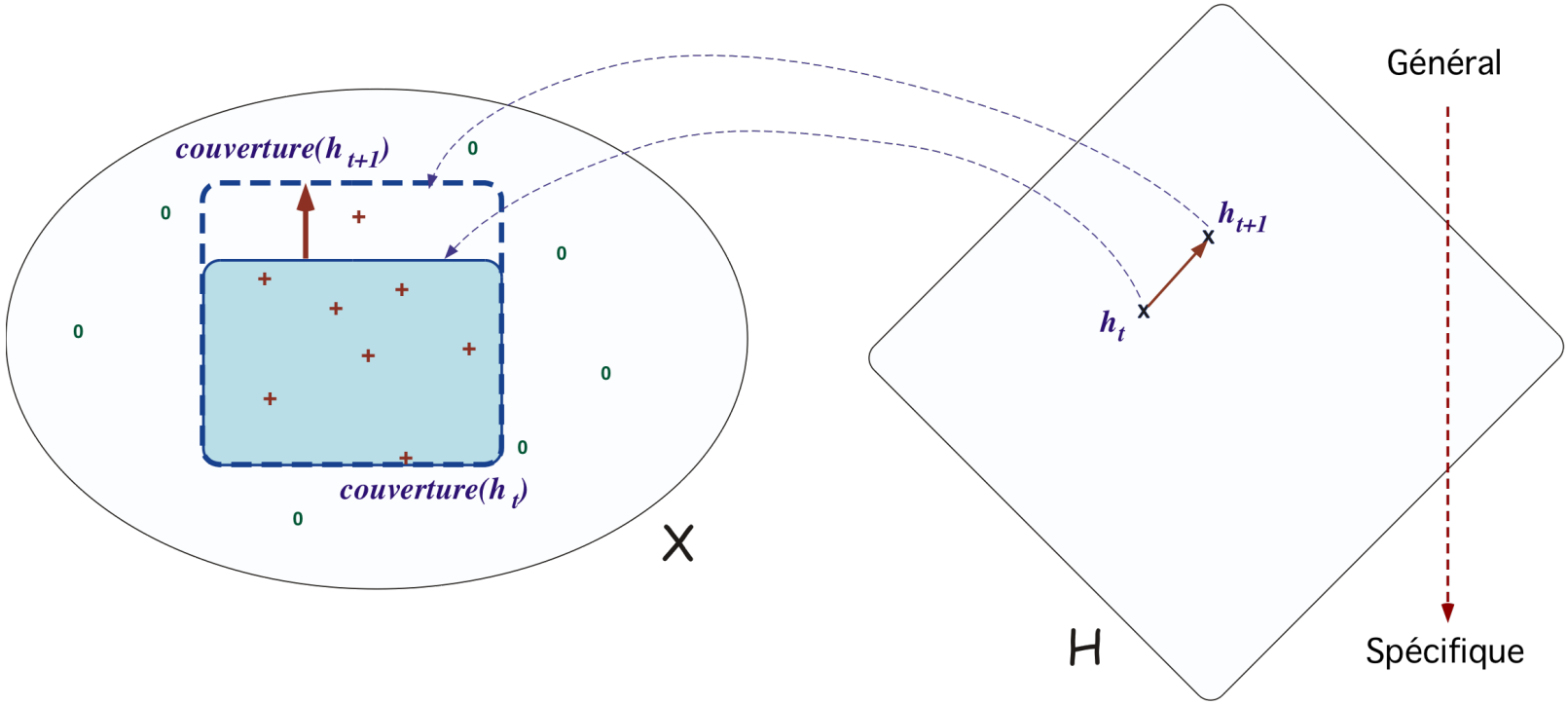


Figure: *La relation d'inclusion dans $\mathcal{X}$ induit la relation de généralisation dans $\mathcal{H}$. Ici, $h_{t+1} \succeq h_t$.*

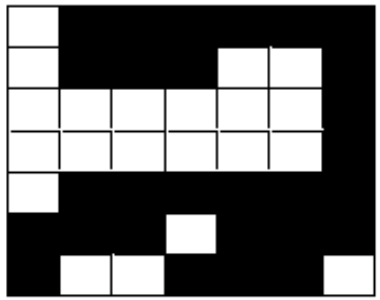
Le Graal :

Trouver **le bon changement de représentation**

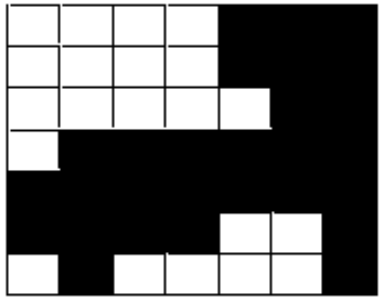
Prédéfinition d'un espace fonctionnel

Les méthodes paramétriques

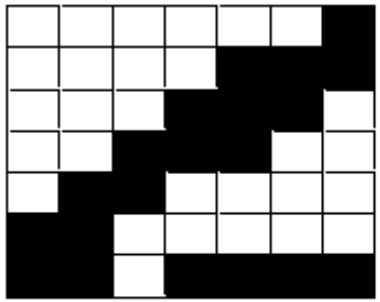
Identifier le bon espace H



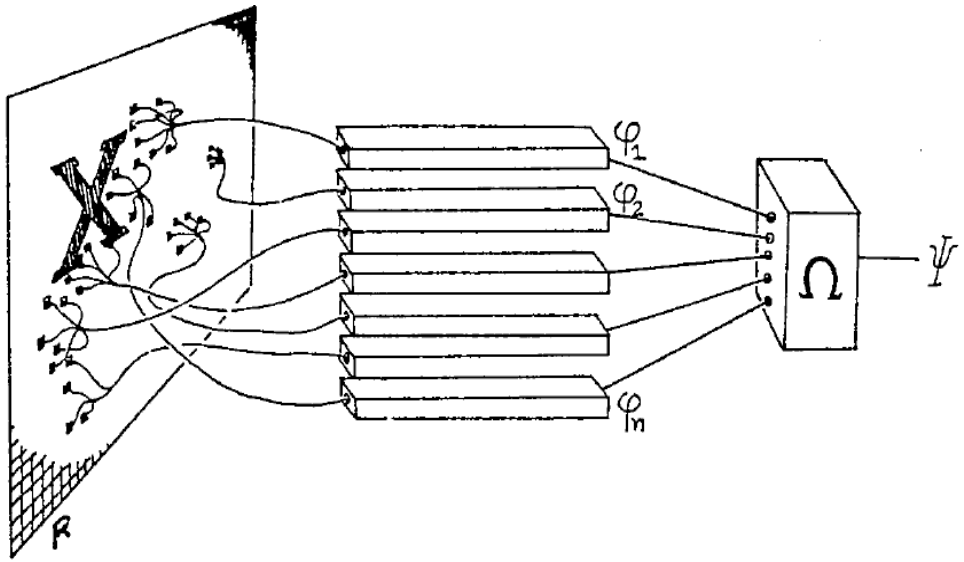
- Oui



- Oui



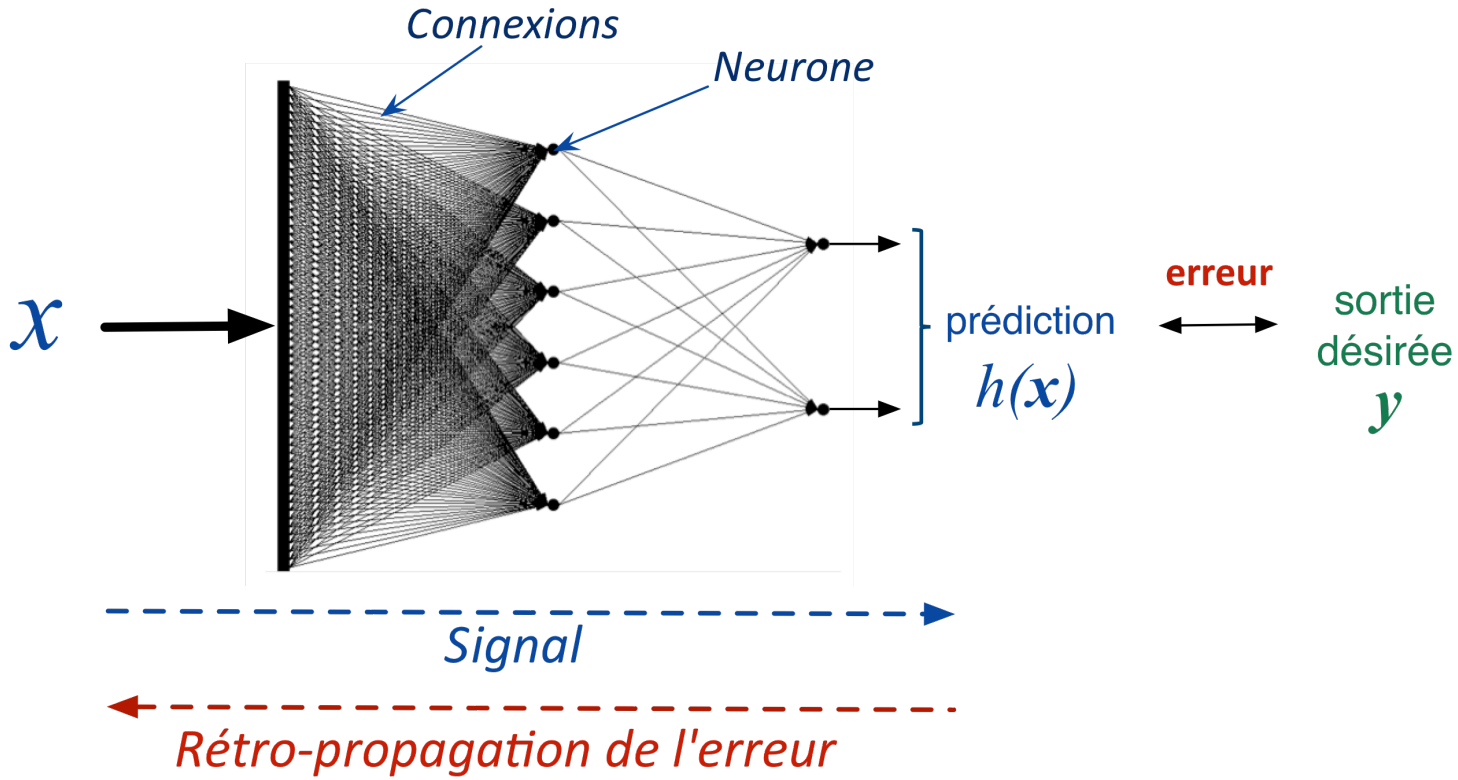
- Non



0 1 1 1 1 1 1 0 1 1 1 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 0 0 1 1 1 0

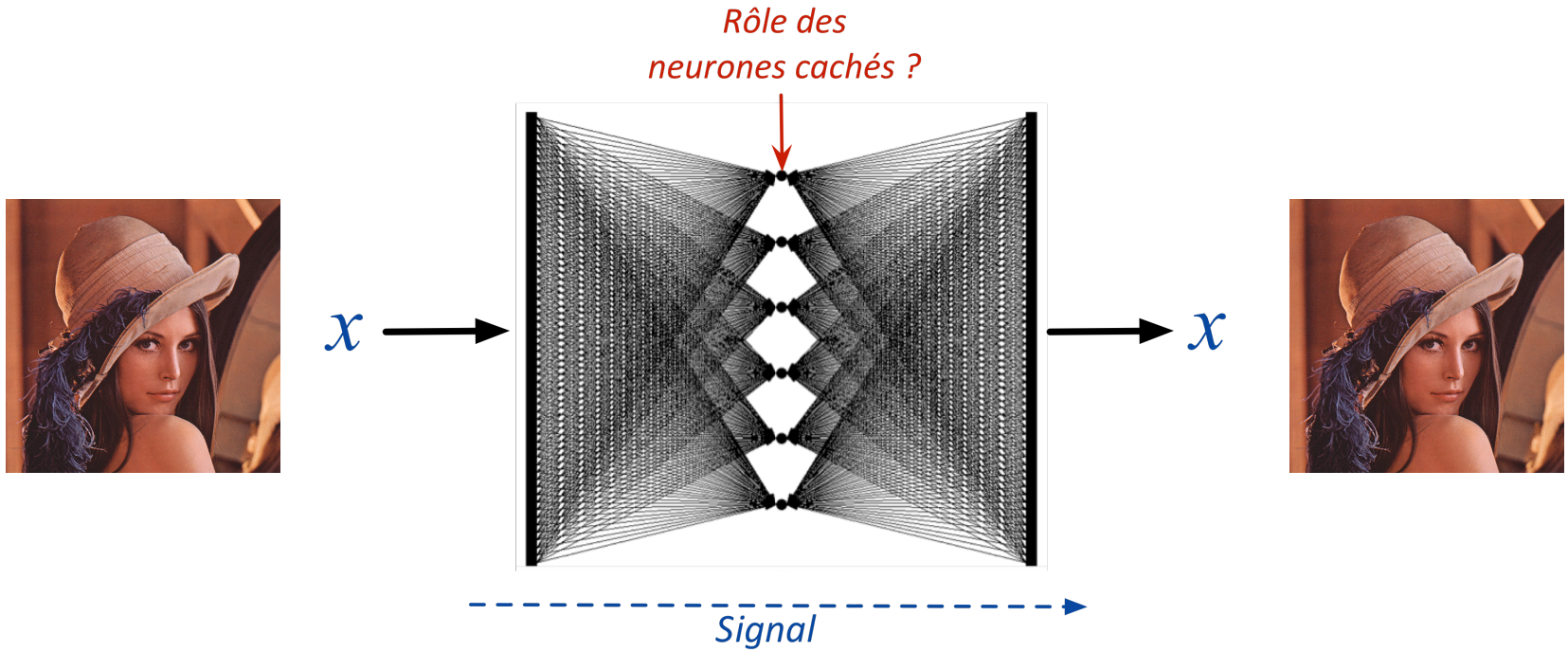
Modèles non linéaires : les réseaux connectionnistes

- Questions :
  - Comment apprendre les **paramètres** (poids des connexions) ?
  - Comment déterminer l'**architecture** du réseau ?



Le changement de représentation : l'architecture

- **Quelle représentation** (variables latentes) ?
- Comment choisir l'architecture ?



Représentation dépendante des données

Méthodes non paramétriques

Changer de représentation

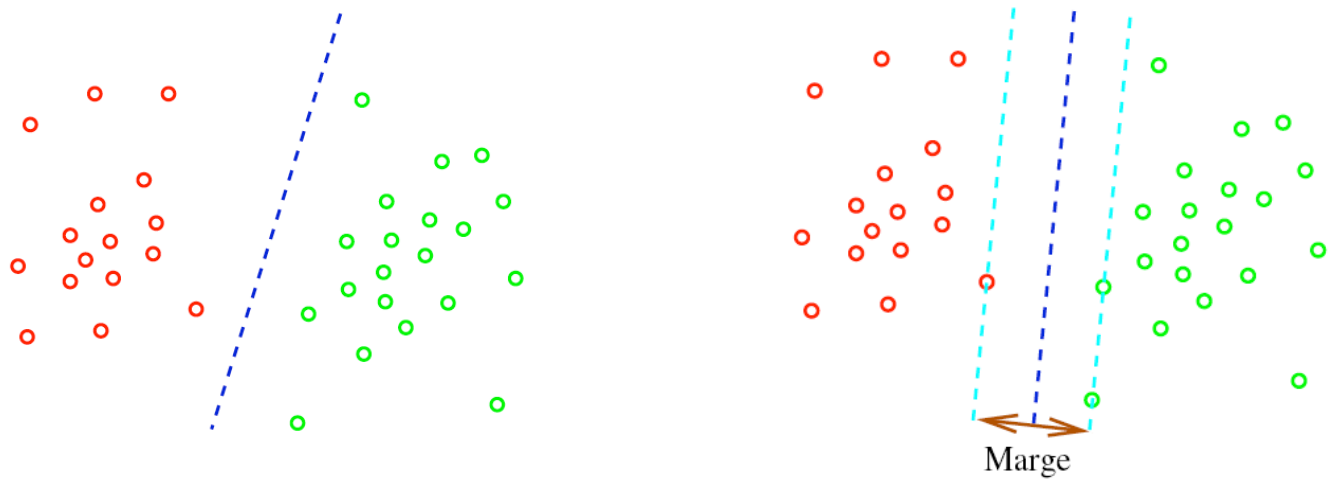
- Changement dépendant des données :

méthodes non paramétriques

  - Méthodes à **noyaux**
  - Méthodes **d'ensemble**
  - Méthodes **constructives** de $\mathcal{H}$
    - Programmation Logique Inductive, …

SVM et **méthodes à noyaux**

- Séparateur linéaire à plus Vaste Marge



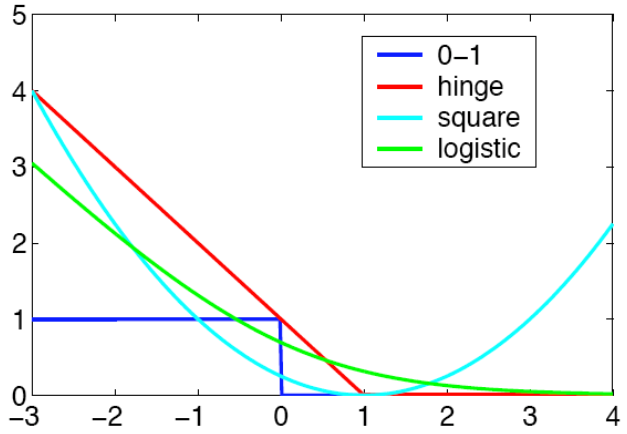
  - Plus robuste à variations de l'échantillon d'apprentissage
  - Validé par analyse théorique
    - bornes de convergence fonction de la marge

SVM et **méthodes à noyaux**

- La recherche de la marge maximale conduit au **critère** :

$$\mathbf{w}^\star = \underset{\mathbf{w}}{\text{ArgMin}} \Big[\underbrace{\sum_{i=1}^{m} |1 - y_i h(\mathbf{x}_i)|_+}_{\text{Risque empirique}} + \underbrace{\frac{1}{2}\mathbf{w}^\top \mathbf{w}}_{\text{Marge}} \Big]$$

Fonction de *perte de substitution* (surrogate loss)



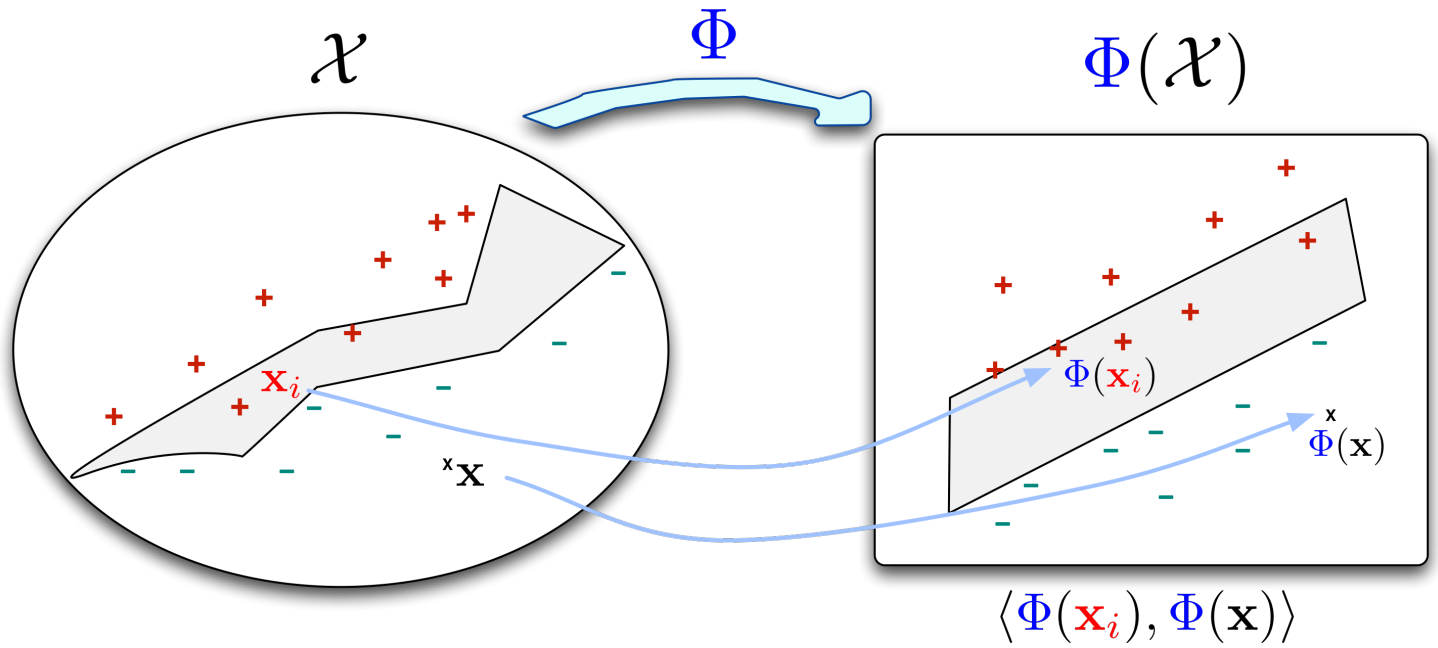
SVM et **méthodes à noyaux**

- Expression de l'hypothèse (fameuse « forme duale »)

$$h^*(\mathbf{x}) = \text{sign}\left\{ \sum_{i \in \mathcal{P}_S} \alpha_i^\star \; y_i \; \langle \mathbf{x}_i \, , \, \mathbf{x} \rangle + w_0^\star \right\}$$

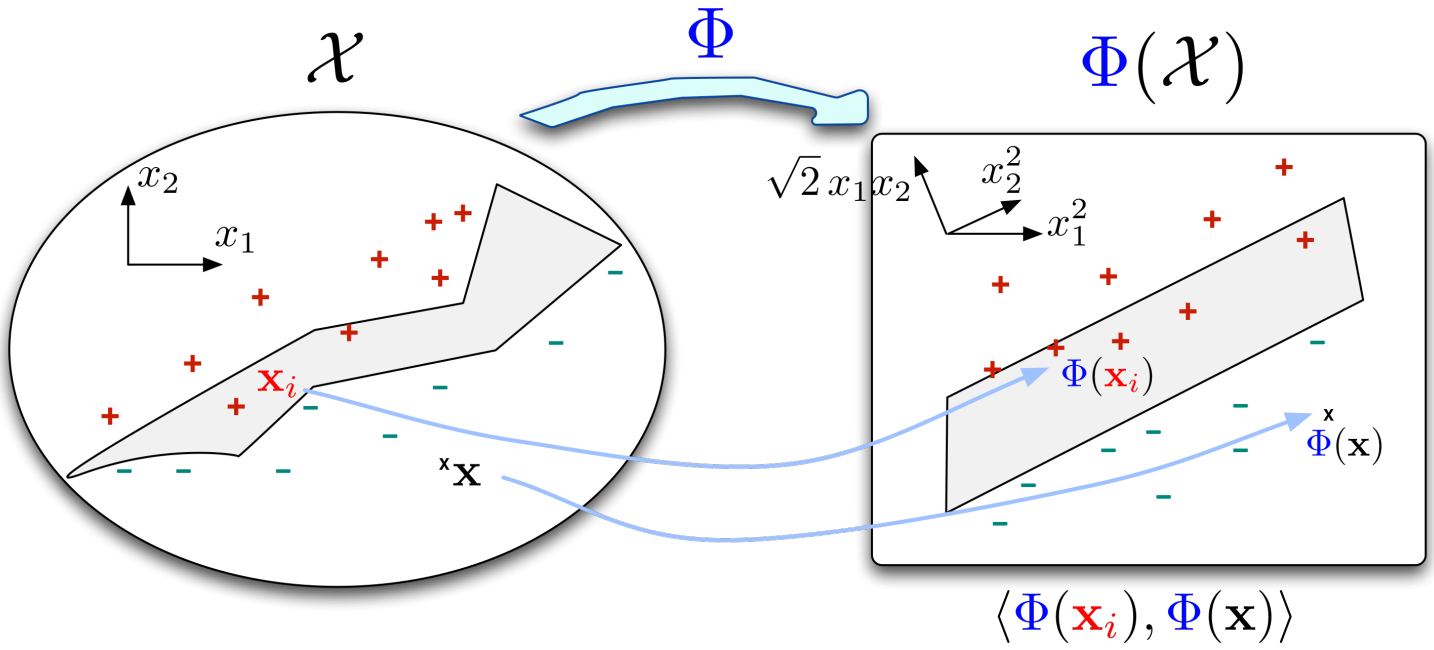
- Trois idées
  - Hypothèses comme combinaison **linéaire**
  - Directement fonction des exemples (*exemples support*)
  - Minimise un risque régularisé dans lequel **la marge** mesure la versatilité de l'hypothèse

SVM et méthodes à noyaux



$$h^*(\mathbf{x}) = \text{sign}\left\{ \sum_{i \in \mathcal{P}_S} \alpha_i^\star \; y_i \; \kappa(\mathbf{x}_i \, , \, \mathbf{x}) + w_0^\star \right\}$$

SVM et méthodes à noyaux

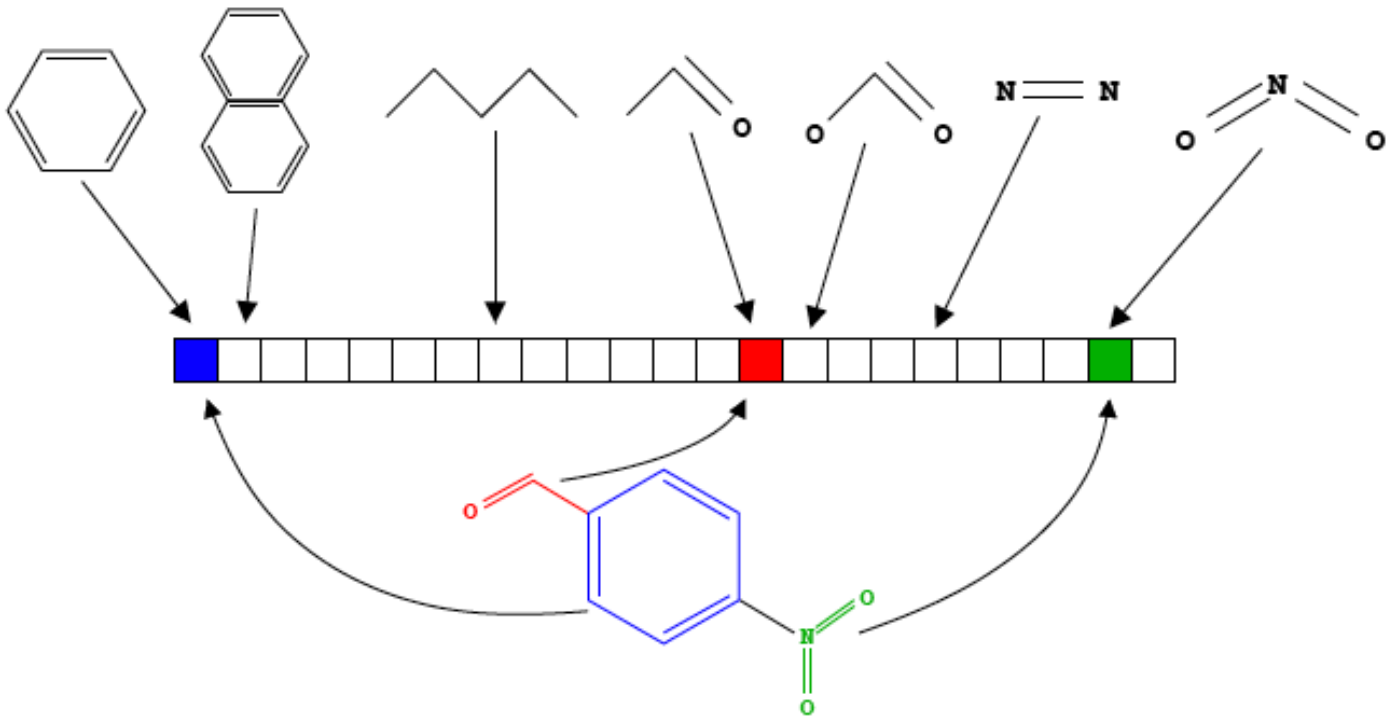


$$h^*(\mathbf{x}) = \text{sign}\left\{ \sum_{i \in \mathcal{P}_S} \alpha_i^\star \; y_i \; \langle \mathbf{x}_i \, , \, \mathbf{x} \rangle^2 + w_0^\star \right\}$$

- Le choix de la fonction noyau **induit** implicitement **un changement de représentation**

Méthodes à noyaux : dans quel espace ?

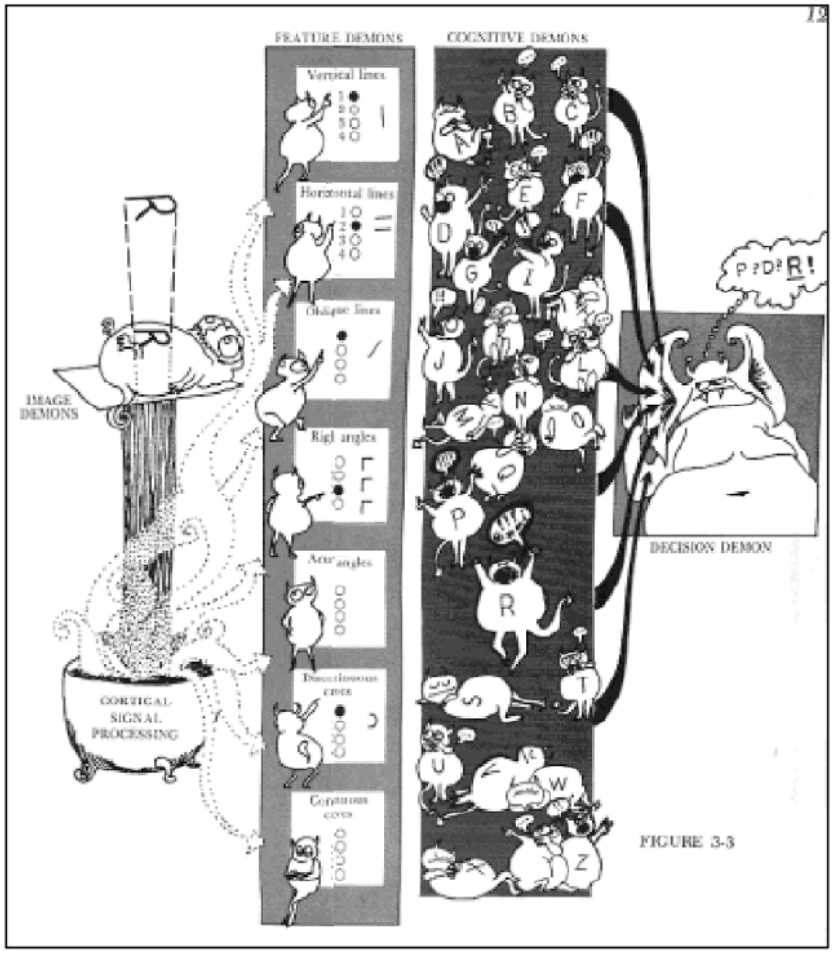
- Changement de représentation



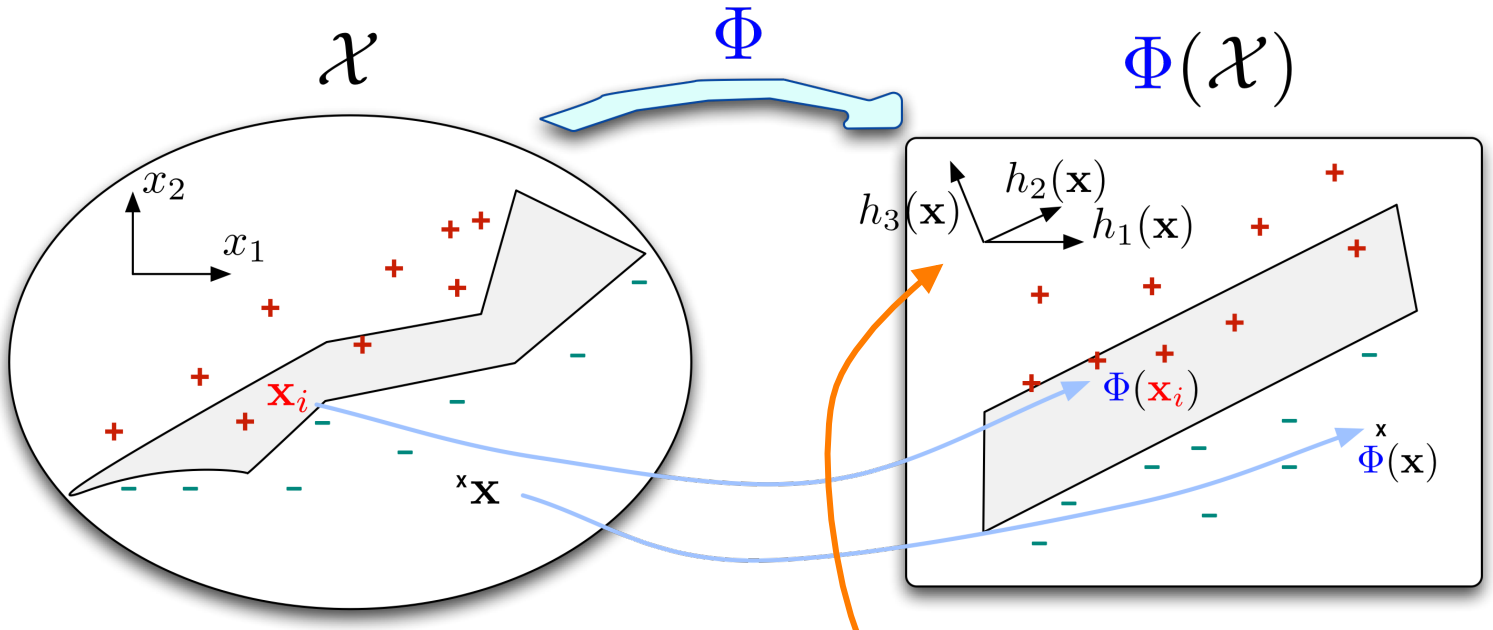
Méthodes collaboratives

Oliver Selfridge (1956) :

« Pandemonium: A Paradigm for Learning ».



Boosting et redescription

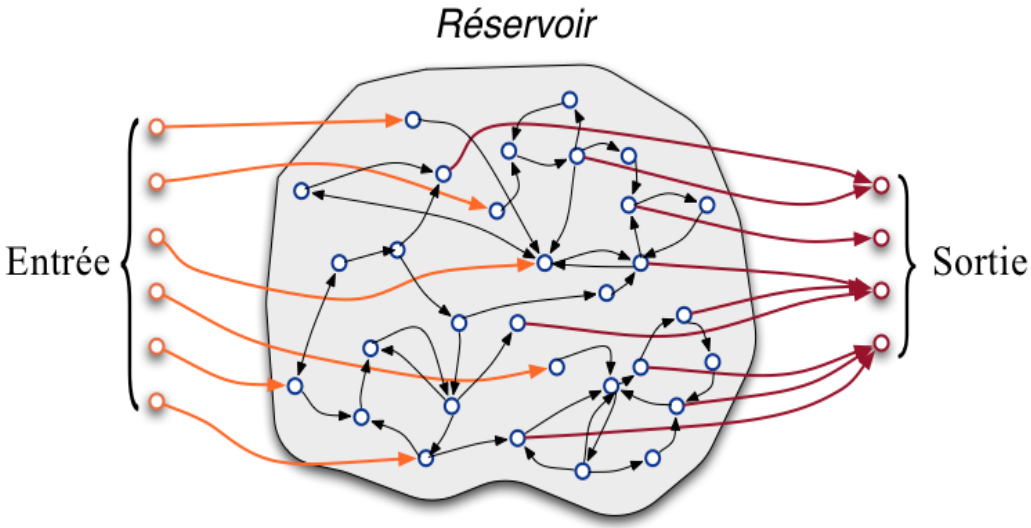


$$H(\mathbf{x}) = \text{sign}\left\{\sum_{i=1}^{T} \alpha_t \, h_t(\mathbf{x})\right\}$$

- Construction **itérative** de l'espace de redescription

Une idée intrigante : le « reservoir computing »

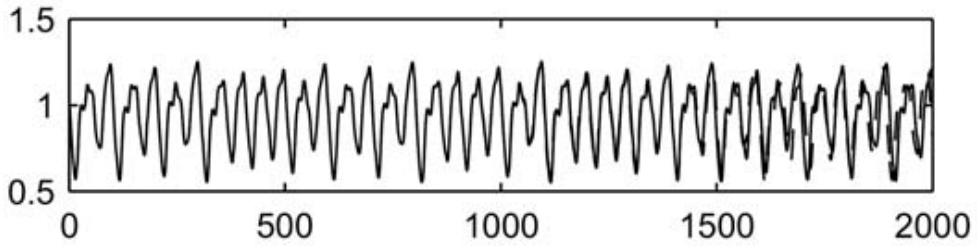
- Idée :
  - Utiliser un réseau récurrent sans l'entraîner explicitement
  - Entraîner une seule couche de sortie

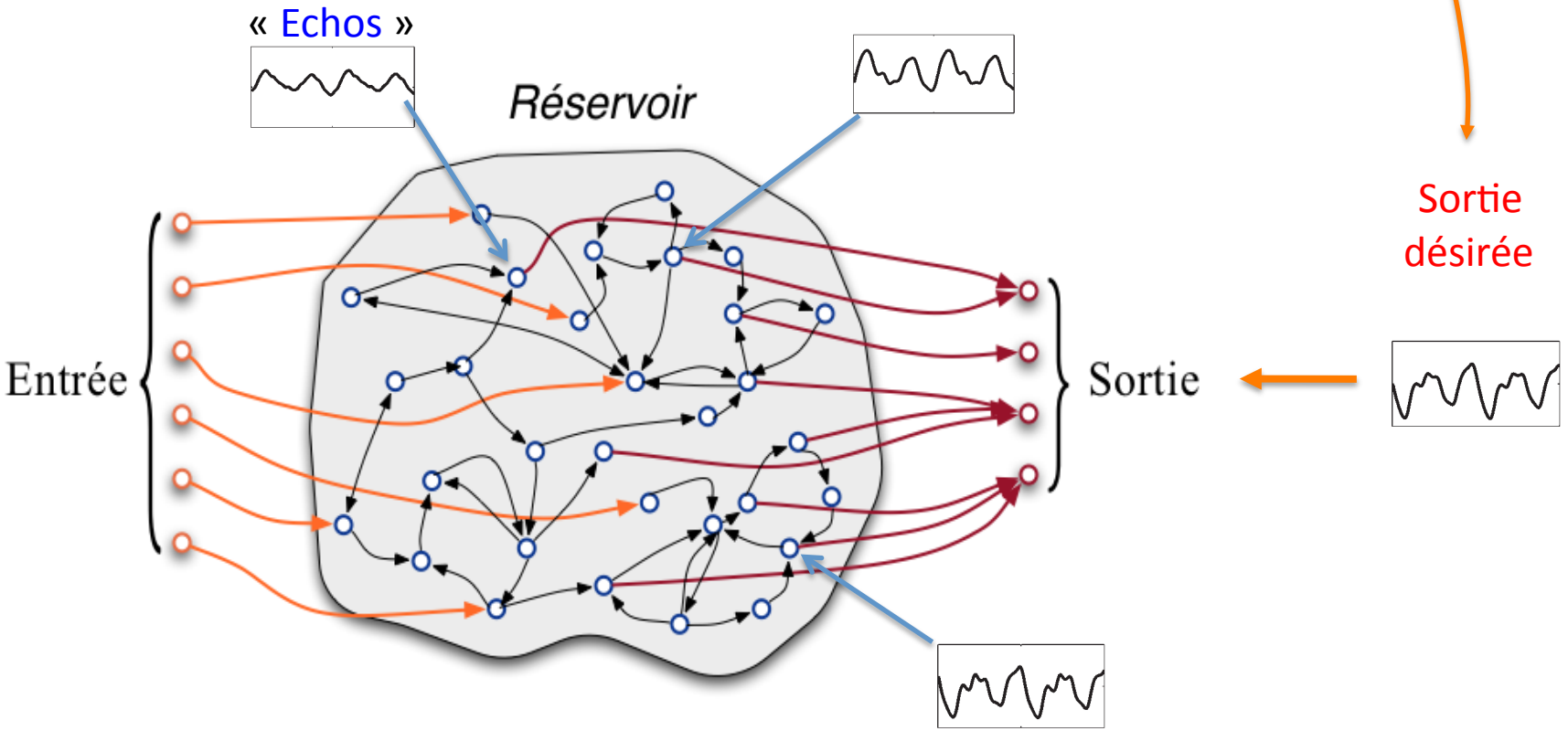


- Ré-introduit une « dynamique »
  - Séries temporelles

Prédiction : « reservoir computing »

Signal à « apprendre »





Les « deep belief networks »

- Motivation
  - Disposer d'un espace d'hypothèses adapté : langage de concepts
  - L'apprendre
- Moyen proposé
  - Réseaux de neurones « profonds »
  - Permettant de **décomposer** en descripteurs
  - Avec **interdépendances** complexes
- Approches
  - Réseaux à convolution (LeCun et al.)
  - « Deep belief networks » (Hinton et al.)

Les « deep beliefs networks »

- En un transparent

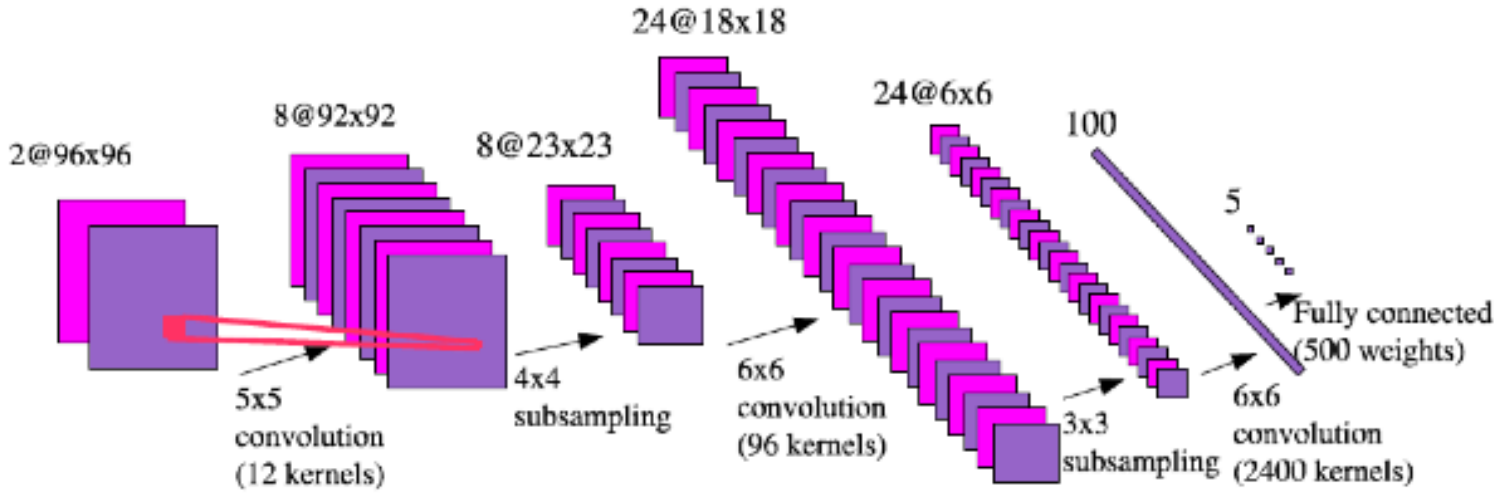
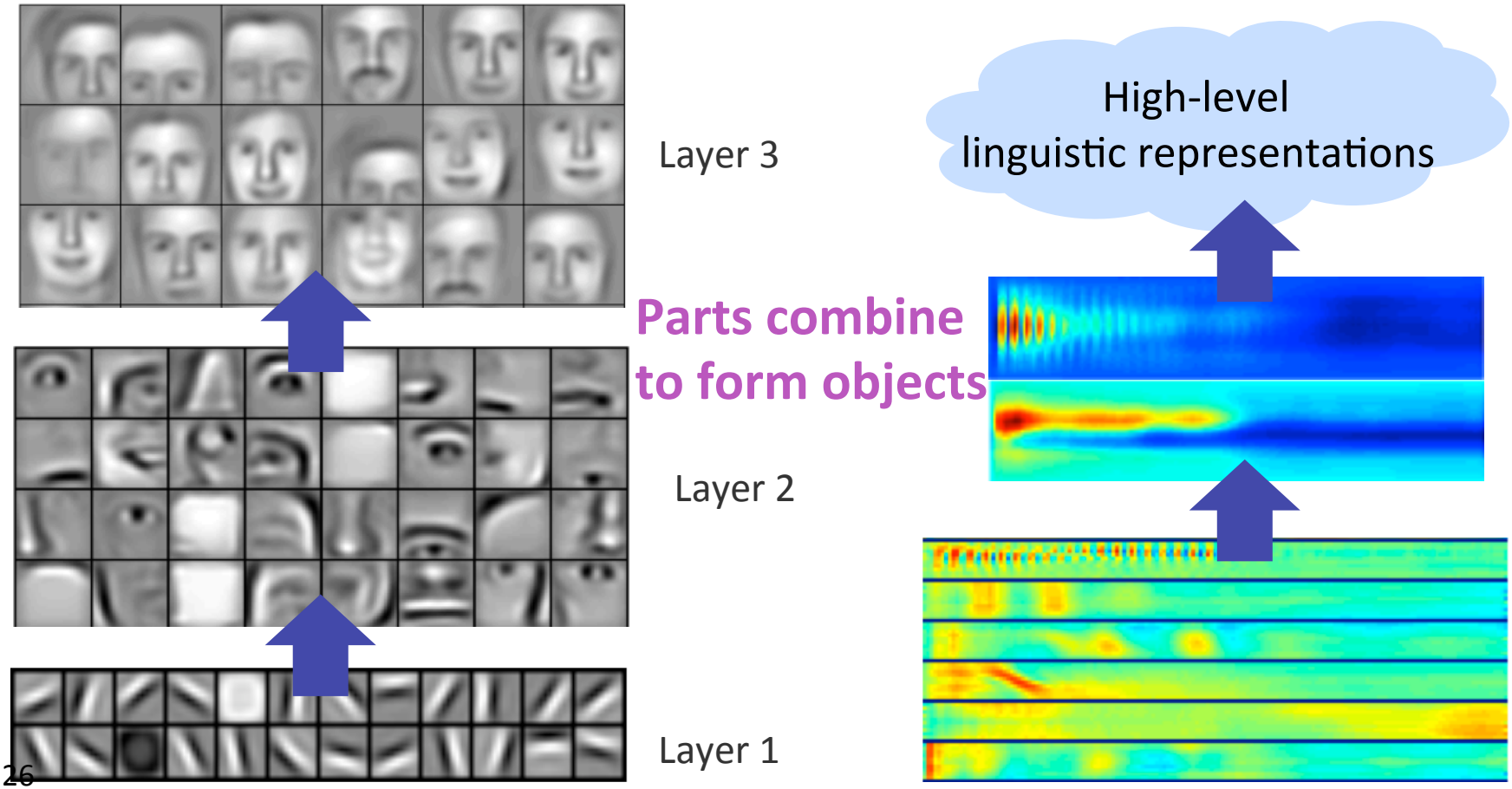


FIG. 10.15: *Architecture d'un réseau connexionniste utilisé pour traiter des images de la base NORB. L'entrée consiste en une paire d'images, dont le système extraie 8 descripteurs de taille* 92×92 *calculés sur des imagettes* 5×5. *Les sorties de ces descripteurs sont reprises par 8 descripteurs* 23×23, *24 descripteurs* 18×18, *24 descripteurs* 6×6 *et une couche de 100 neurones complètement connectés aux 5 neurones de sortie qui donnent la distance avec les vecteurs cibles. (Repris de [BL07].)*

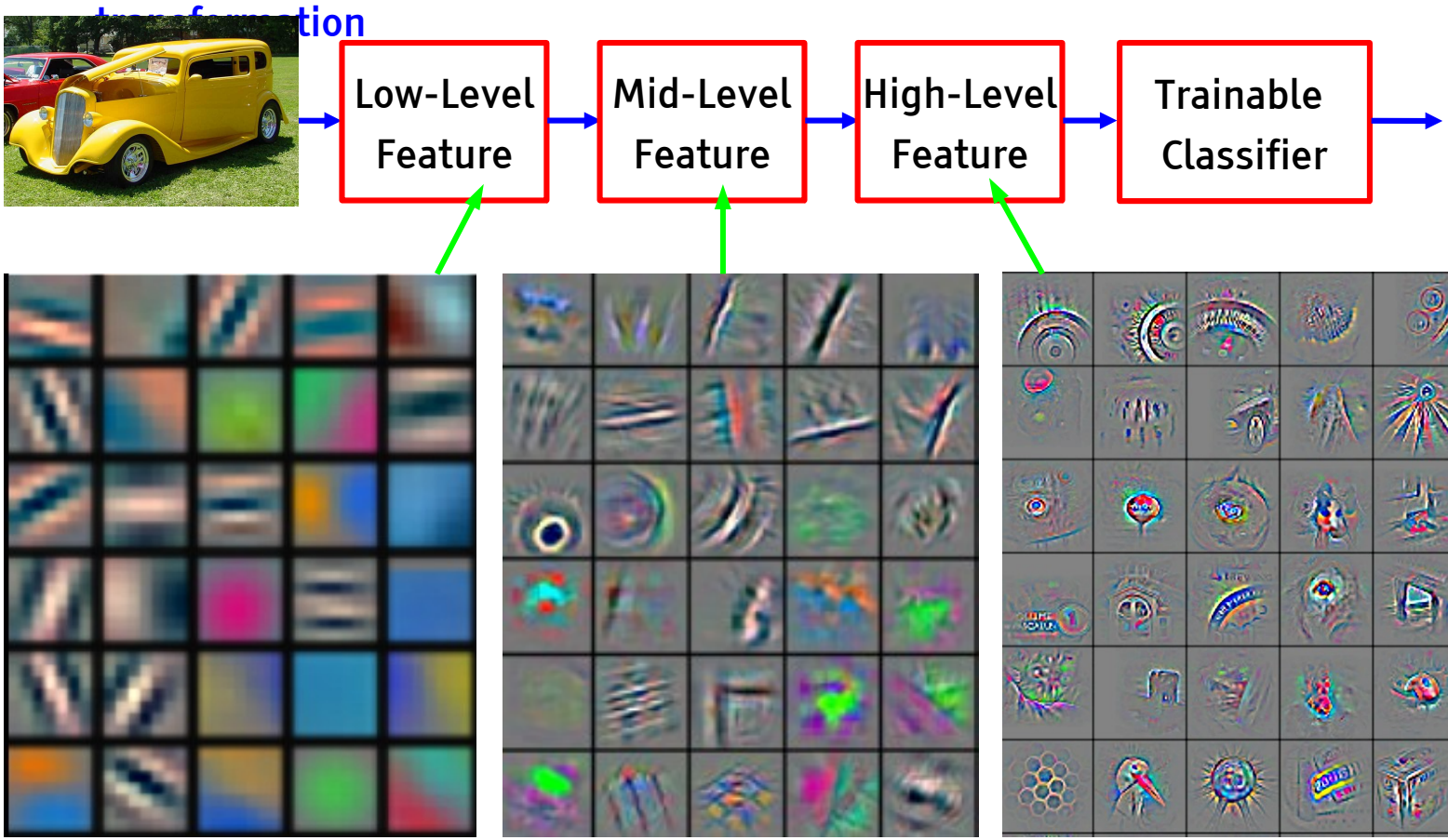
Apprentissage de représentations hiérarchiques

- Apprentissage de représentations hiérarchiques



26

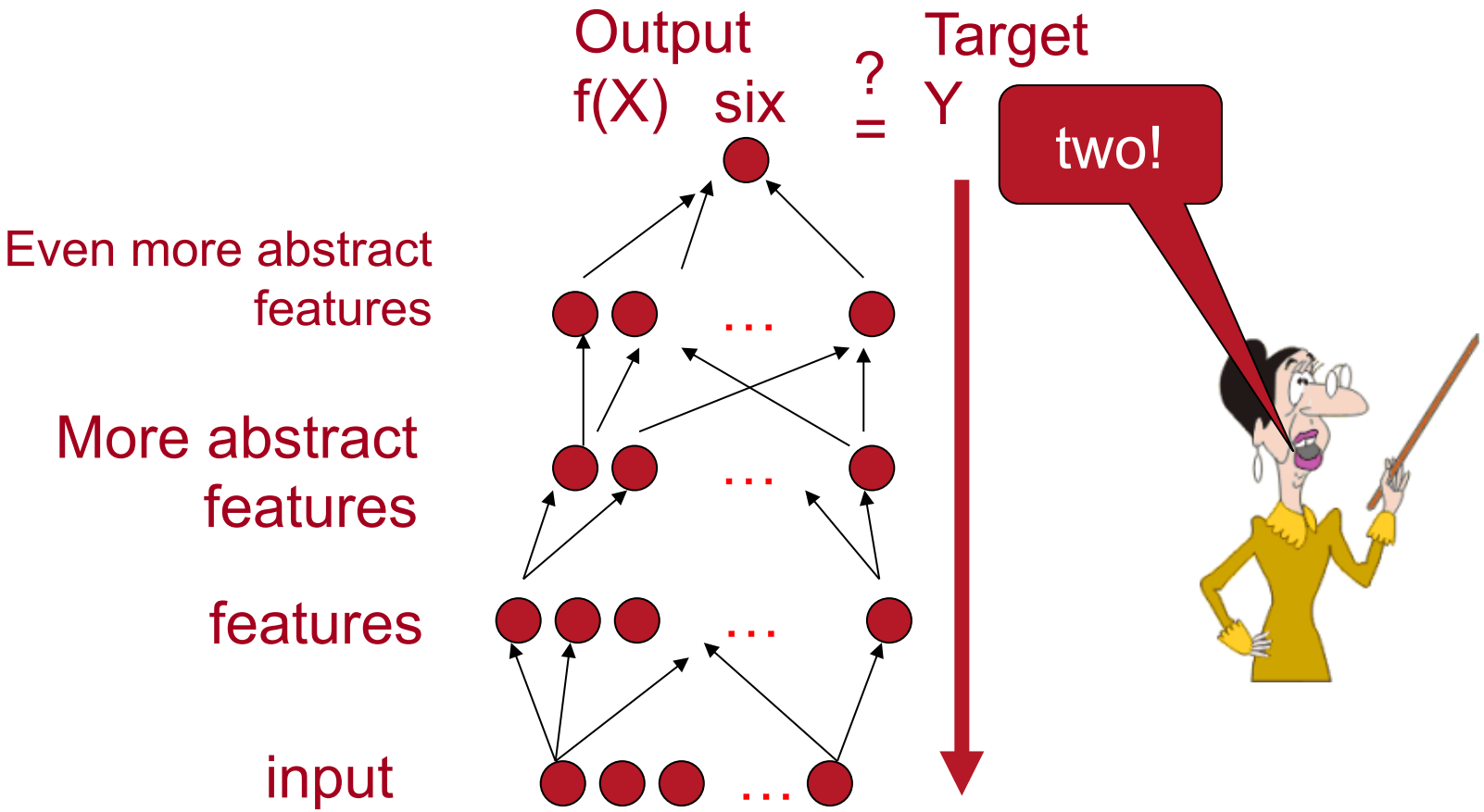
Apprentissage de représentations hiérarchiques



Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

Apprentissage

- Apprentissage supervisé final



- Additional hypothesis: features good for P(x) good for P(y|x)

Illustration

- Système développé par Google et U. de Stanford

- Reconnaissance de visages
  - Sous conditions de lumière diverses
  - Sous tout angle
- Apprentissage non supervisé
  - 9 couches ; 10^9 connexions
  - 10 millions d'images
  - 3 jours de calcul sur 16 000 processeurs
- Amélioration des performances de 70% / état de l'art

Apprentissage de dictionnaire : **Le problème**

- Étant donné un ensemble de signaux (exemples)

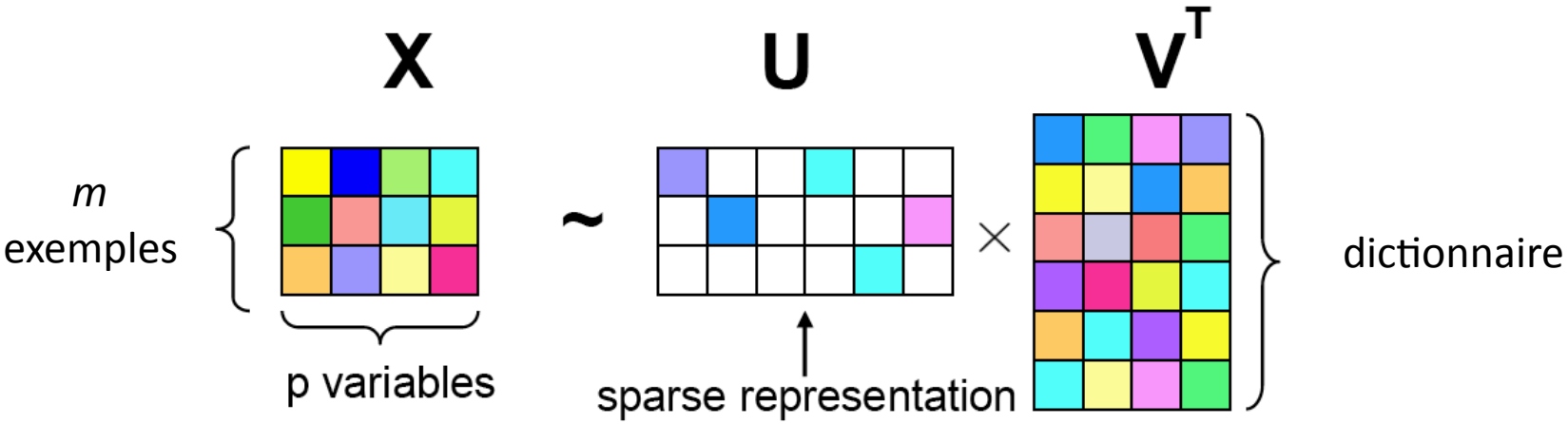
$$\mathcal{S} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$$

- Supposés avoir une représentation parcimonieuse dans un dictionnaire inconnu *D*

$$\mathbf{x}_i = \sum_k w_k \mathbf{d}_k$$

- Peut-on trouver *D* (ou une approximation) ?

Apprentissage de dictionnaire



Problème associé

- On veut résoudre le problème :

$$\min_{\mathbf{D},\mathbf{w}\in\mathbb{R}^p} \underbrace{\left\{\frac{1}{2}||\mathbf{x}-\mathbf{D}\mathbf{w}||_2^2\right\}}_{\text{data fitting term}} + \underbrace{\lambda\,\psi(\mathbf{w})}_{\text{sparsity-inducing regularization}}$$

- Mais ici on veut apprendre aussi le dictionnaire *D*
- en même temps que le codage parcimonieux optimal

Comment résoudre le problème

- Problème non convexe car bilinéaire en deux objectifs

- Démarche proposée : itérations de deux étapes

  1. Recherche de **codage** parcimonieux : les w_k étant donné un dictionnaire

  2. Mise à jour du **dictionnaire *D*** en fixant ***W***

Bilan

- Savoir choisir une bonne représentation est essentiel
  - Pour **apprendre**
  - Pour aider à l'**interprétation**
  - Pour aider au **raisonnement**
    - [Bengio & Le Cun, 07] « *Scaling Learning Towards AI* »
    - [Bengio et al., 09] « *Curriculum Learning* »
    - [Bottou, 11] « *From Machine Learning to Machine Reasoning* »
    - [Valiant, 00] « *A Neuroidal Architecture for Cognitive Computation* »
  - Pour aider au **transfert**

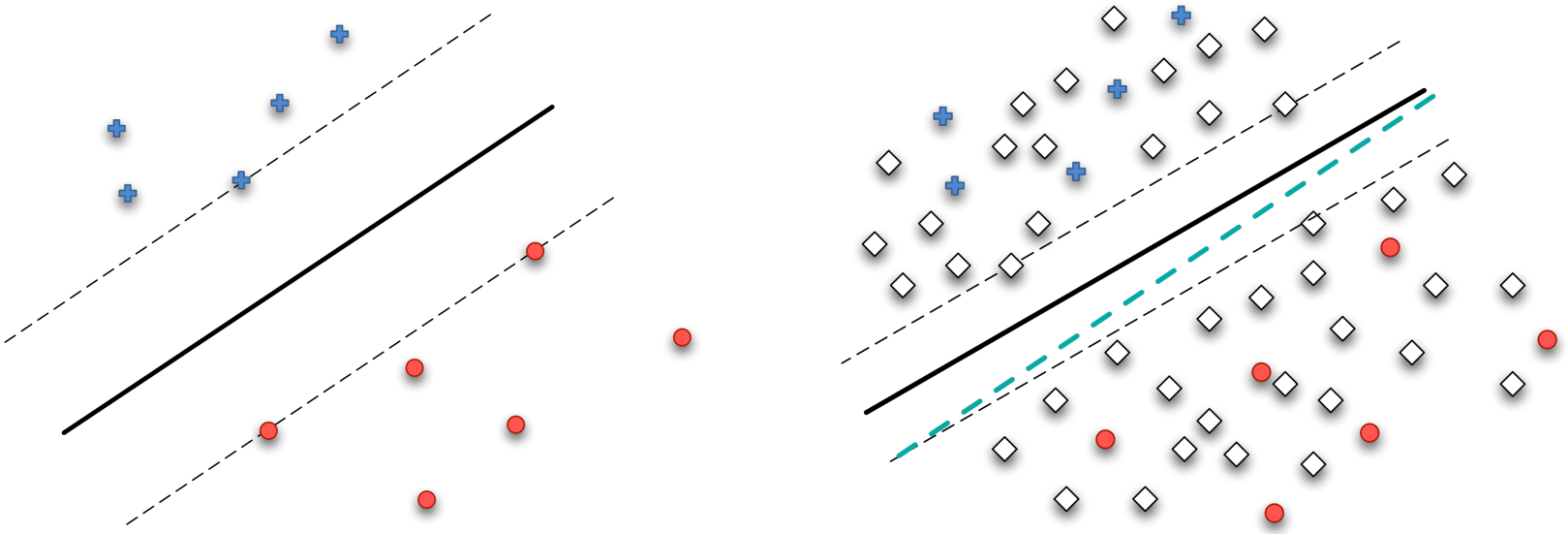
De **nouvelles questions**

De nouveaux défis

- Nouveaux scénarios
  - Semi-supervisé
  - Apprentissage actif
- Espaces **non vectoriels**
  - Entrées multi-formes
  - Sorties structurées
- **Flux** de données
  - Apprentissage en-ligne
    - Nouveaux critères
  - Environnements non stationnaires
- Apprentissage **multi-tâches**
- Apprentissage **multi-domaines** (transfert)
- Apprentissage **incrémental**
- Apprentissage **collaboratif**
- **Enseigner** à un système par l'exemple ?

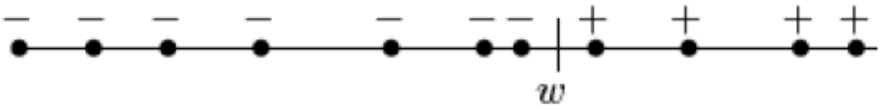
Apprentissage **semi-supervisé**

- $S = \{(x_1, y_1), \ldots, (x_l, y_l), x_{l+1}, \ldots x_m\}$



Apprentissage **actif**

$$h_w(x) = \begin{cases} 1 & \text{if } x \geq w \\ 0 & \text{if } x < w \end{cases}$$



Tirage au hasard des points : $m = \mathcal{O}(\frac{1}{\varepsilon})$ (en supposant $\mathcal{H} = \mathcal{F}$)

Sélection active : $m = \mathcal{O}(\log \frac{1}{\varepsilon})$

Amélioration exponentielle en terme d'échantillonnage !!

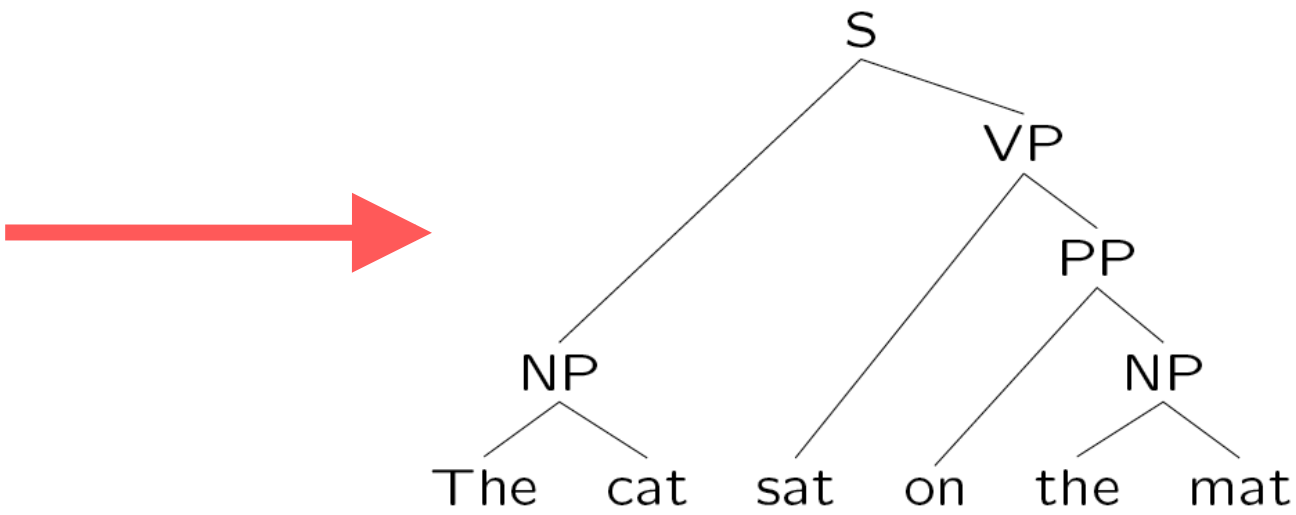
Données et sorties non vectorielles

- Entrées
  - Documents
  - Vidéos
  - Graphes
  - …

- Sorties
  - Structure
    - Arbre ; Grammaire ; …
    - Graphe
    - …

Données et sorties **structurées**

The
cat
sat
on
the
map



Masses et flux de données

- $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i), \ldots, (x_m, y_m)\}$
  - Très très grand
  - Sans fin

- Si stationnaire
  - Faut-il encore régulariser ?

- Si **non** stationnaire
  - Nouveaux algorithmes
    - Compromis stabilité / plasticité : comment et quoi oublier ?
  - Nouveaux critères

Données massives et flux de données

- **Contraintes**
  - Impossible de stocker les données et de les traiter toutes d'un coup
  - Traitement à la volée : complexité en O(1)
  - Apprentissage incrémental

- Si environnement **non stationnaire**
  - $\mathbf{P}_{\mathcal{X}}$ change
    - Échantillon d'apprentissage ≠ exemples à traiter
    - Co-variate shift (e.g. saisons)
  - $\mathbf{P}_{y|\mathcal{X}}$ change : dérive de concept
  - **On-line learning**

Apprentissage multi-objectifs ou multi-tâches

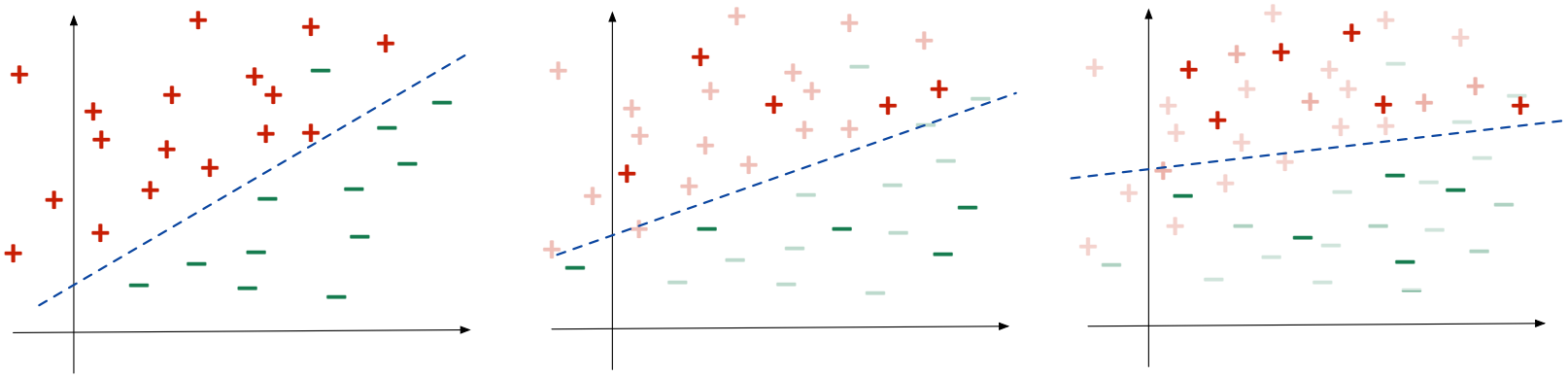
- Apprendre à répondre **simultanément** à plusieurs questions
  - E.g. maladie du rein ou non ET symptômes de diabète ou non
- **Bénéfique** dans de nombreux cas
  - E.g. [Tom Mitchell, 2005]
    - Apprendre à reconnaître la couleur et la forme
- **Pourquoi** ?
  - Notion de similarité entre tâches
    - Définition ?
    - Mesure ?

Apprentissage multi-domaines et transfert

- Utilisation d'une (ou plusieurs) **tâche**(s) **source** pour résoudre une **tâche cible**
  - Exemples :
    - apprendre Java après C++ ?
    - Apprendre à configurer un trieur de courriels pour un utilisateur après qu'il ait été configuré pour un ou plusieurs autres utilisateurs
  - ***Séquentiellement*** ? Oui et non
  - Ce qui est ***partagé / transféré***
    - **mémoire des exemples** sources
    - seulement **mémoire des hypothèses** apprises sur la source

Apprentissage incrémental … et transfert

- La séquence devient primordiale



Apprentissage collaboratif … et transfert

- Un même problème
- Des problèmes locaux mais reliés

Conclusions

… ou pas ?

Sortir du cadre …

Les acquis

- **Théorie** de l'induction
  - Les **ingrédients**
    - Espace d'hypothèses
    - Critère à optimiser : un risque régularisé
    - Algorithme d'exploration de $\mathcal{H}$: un problème d'optimisation
  - Grande puissance pour en **dériver des algorithmes**
  - Apparemment approche **très générique**

Même distribution en **apprentissage** et en **test**

- Cas particuliers
  - *Apprentissage descriptif (non supervisé)*
  - *Apprentissage par renforcement*

Le cadre devenu « paradigmatique »

- Données i.i.d.
  - Environnement **stationnaire**
  - Tirage **indépendant** et représentatif
  - Traitement « **batch** »
- Association **1** entrée -> **1** sortie

- Approche toute puissante
  - Définition d'un **risque empirique régularisé** (contrôle de $\mathcal{H}$)

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

Le paradigme actuel

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

1. L'exprimer sous forme d'**un critère inductif** à optimiser (si possible *convexe*)
   - **Risque empirique**
     - avec une **fonction d'erreur** adéquate
   - Un **terme de régularisation**
     - exprimant les contraintes
     - et connaissances a priori

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

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

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 $\mathrm{dis}_{\rho_{w'}}(S_u, T_u) = \left| \mathop{\mathbf{E}}_{(h,h')\sim\rho_{w^2}} R_{S_u}(h,h') - \mathop{\mathbf{E}}_{(h,h')\sim\rho_{w^2}} R_{T_u}(h,h') \right|$ 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 :

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

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 :

« Traduction » : sélection de descripteurs

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

- Méthodes de type LASSO

« Traduction » : classification **semi-supervisée**

- *l* données **étiquetées**, *u* données **non étiquetées**

$$\mathcal{S} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_l, y_l), \mathbf{x}_{l+1}, \ldots, \mathbf{x}_{l+u}\}$$

$$\mathbf{h} = \left[h(\mathbf{x}_1), h(\mathbf{x}_2), \ldots, h(\mathbf{x}_{l+u})\right]$$

Mesure de régularité sur les données $\quad \mathbf{h}^\top \mathcal{L}\, \mathbf{h} = \frac{1}{2} \sum_{i,j=1}^{l+u} W_{ij} \left(h(\mathbf{x}_i) - h(\mathbf{x}_i)\right)^2$

$$h^\star = \underset{h \in \mathcal{H}}{\text{Argmin}} \left\{ \frac{1}{l} \sum_{i+1}^{l} \left(y_i - h(\mathbf{x})\right)^2 + \lambda_1 \, ||h||_2 + \lambda_2 \, \mathbf{h}^\top \mathcal{L}\, \mathbf{h} \right\}$$

« 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 $$\mathbf{w}_j = \mathbf{w}_0 + \mathbf{v}_j$$

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

… sortir du cadre ?

Présupposés sous-jacents … fondamentaux

- Tirage i.i.d.
  - Des données
  - Des tâches
  - Pas de dépendance temporelle ou causale

- Notion fondamentale de distribution
  - Même si « agnostique » (contre toute distribution)

Des défis … qui obligent à sortir du cadre ?

- Masses de données
  - NON : Le cadre semble parfait

- Quand il n'y a **pas de distribution** …
  - Apprentissage à partir de (très) **peu d'exemples**
  - Apprentissage à partir d'**exemples choisis** : ***enseignement***

- Qu'est-ce qui peut être transféré ?
  - Entre tâches
  - Entre agents

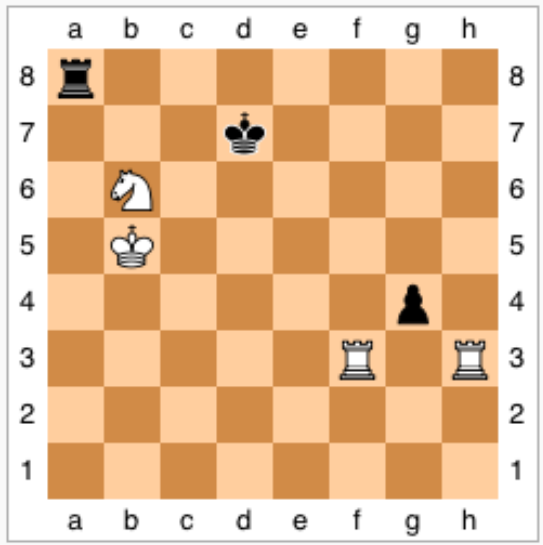
Apprendre **à partir d'un exemple**

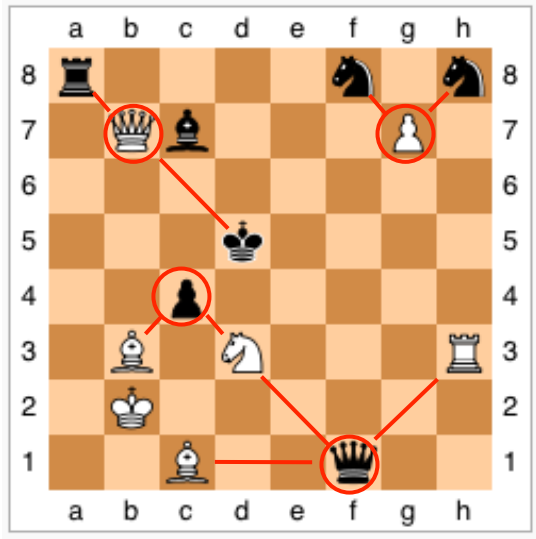
Explanation-Based Learning

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** :

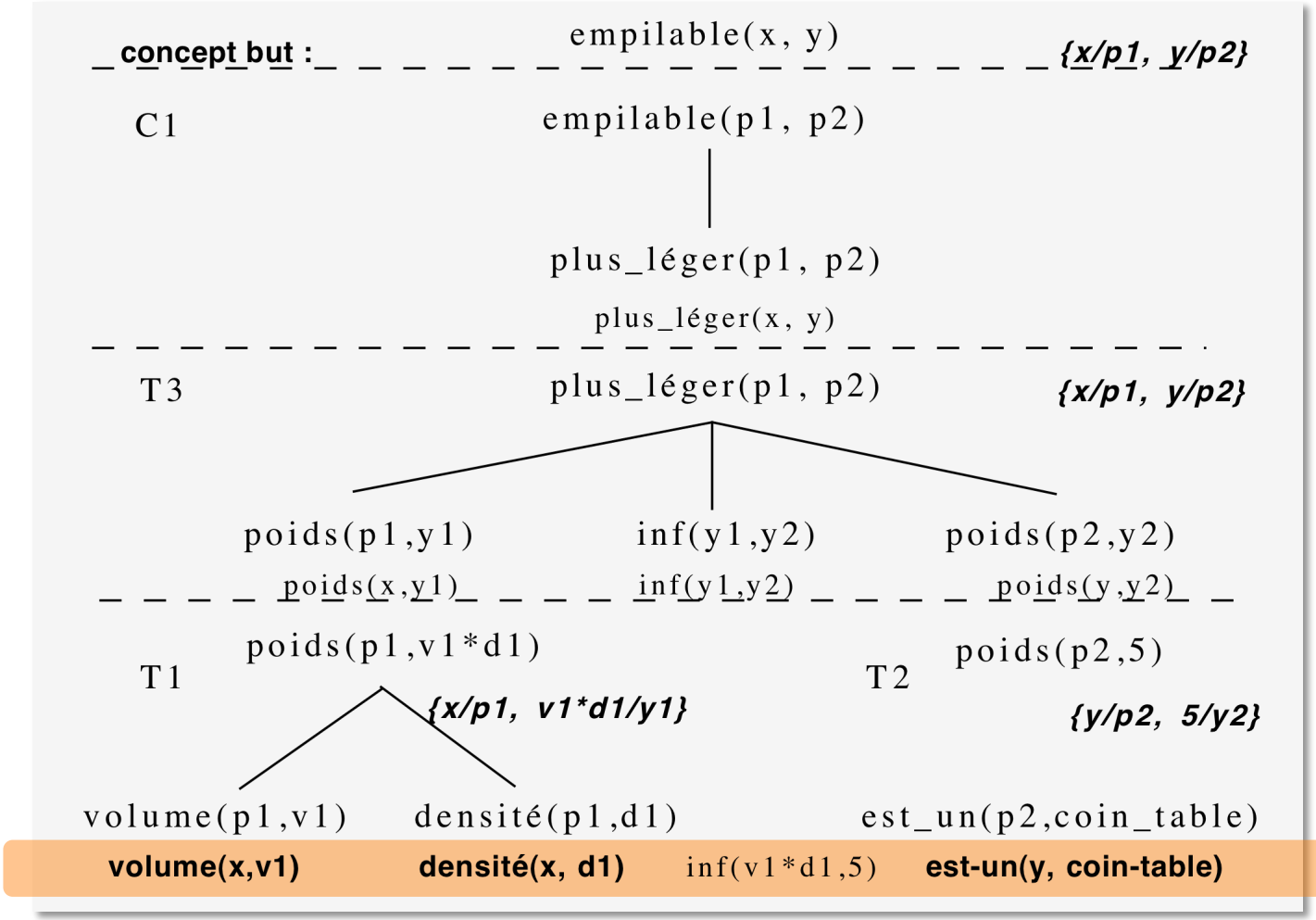
```
(T1) : poids(X, W) :- volume(X, V), densité(X, D), W is V*D.

(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*.

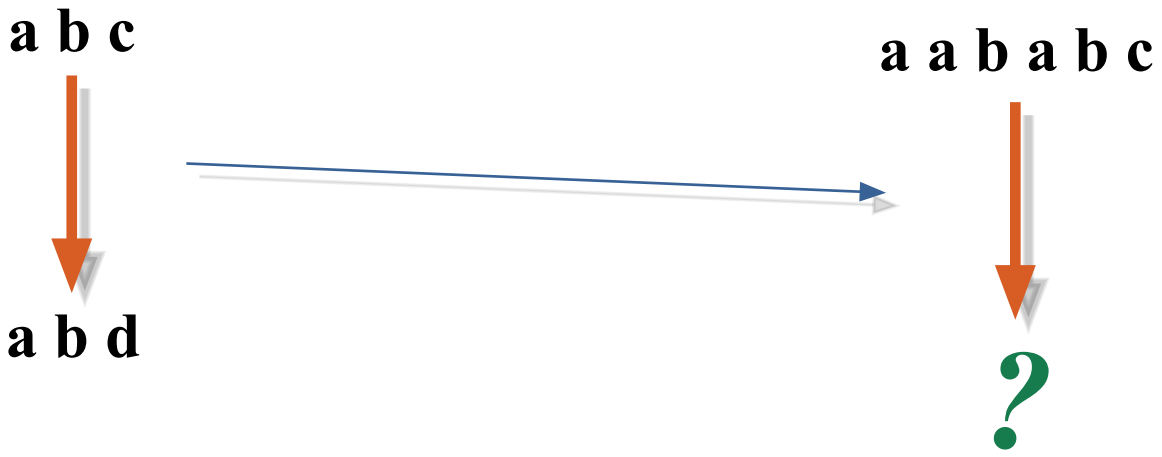
Apprentissage de **causalités**

- **Est-ce que** A cause B ?
  - *Est-ce que telle altération génétique **cause** telle malformation ?*
  - *Est-ce que fumer **cause** le cancer ?*
  - *Est-ce que l'inscription du baromètre cause la tempête ?*
- A **quel degré** A cause B ?
  - *A quel degré l'ingestion d'OGM cause X ?*

- Identifier des **corrélations stables**
  - Nous sommes mauvais à identifier des structures de causalité à partir de corrélations
- Considérer **l'ordre**
- Emploi de **contre-factuelles**
- …

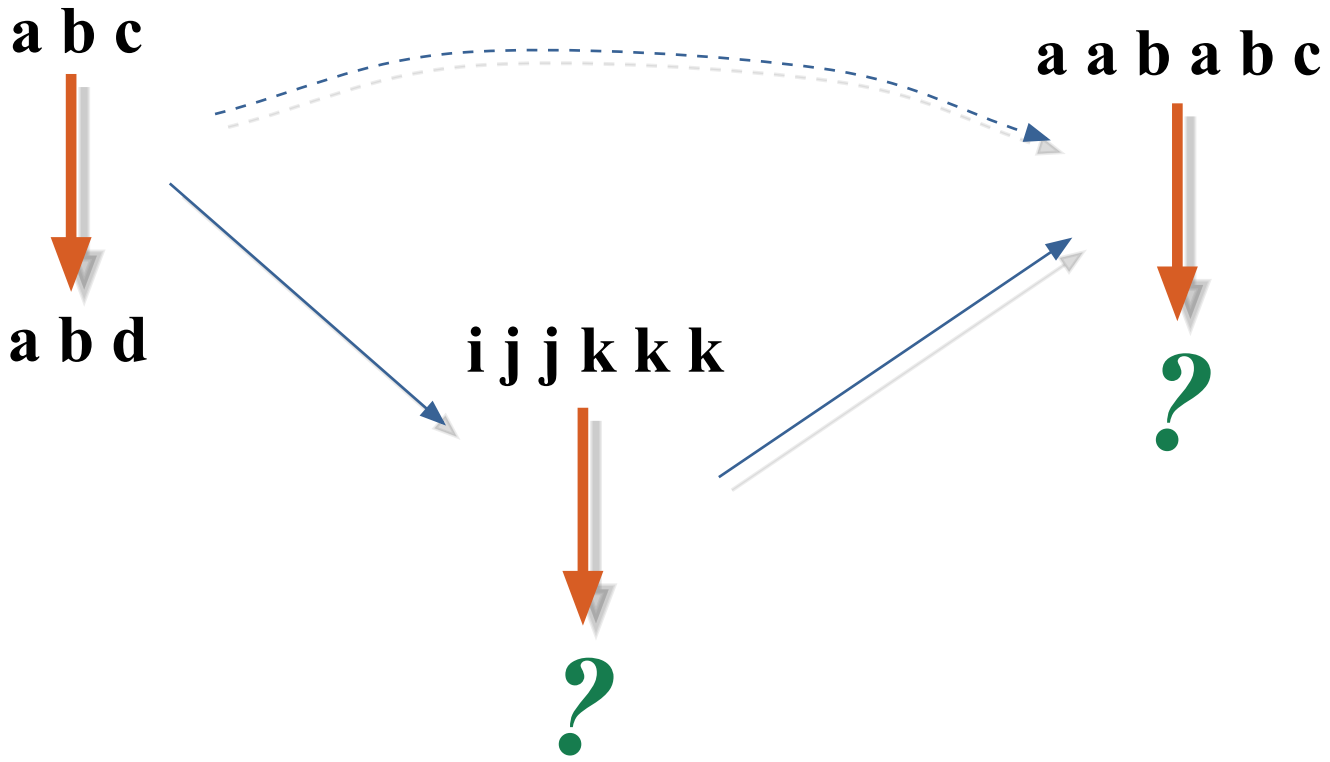
Analogie

- Transfert



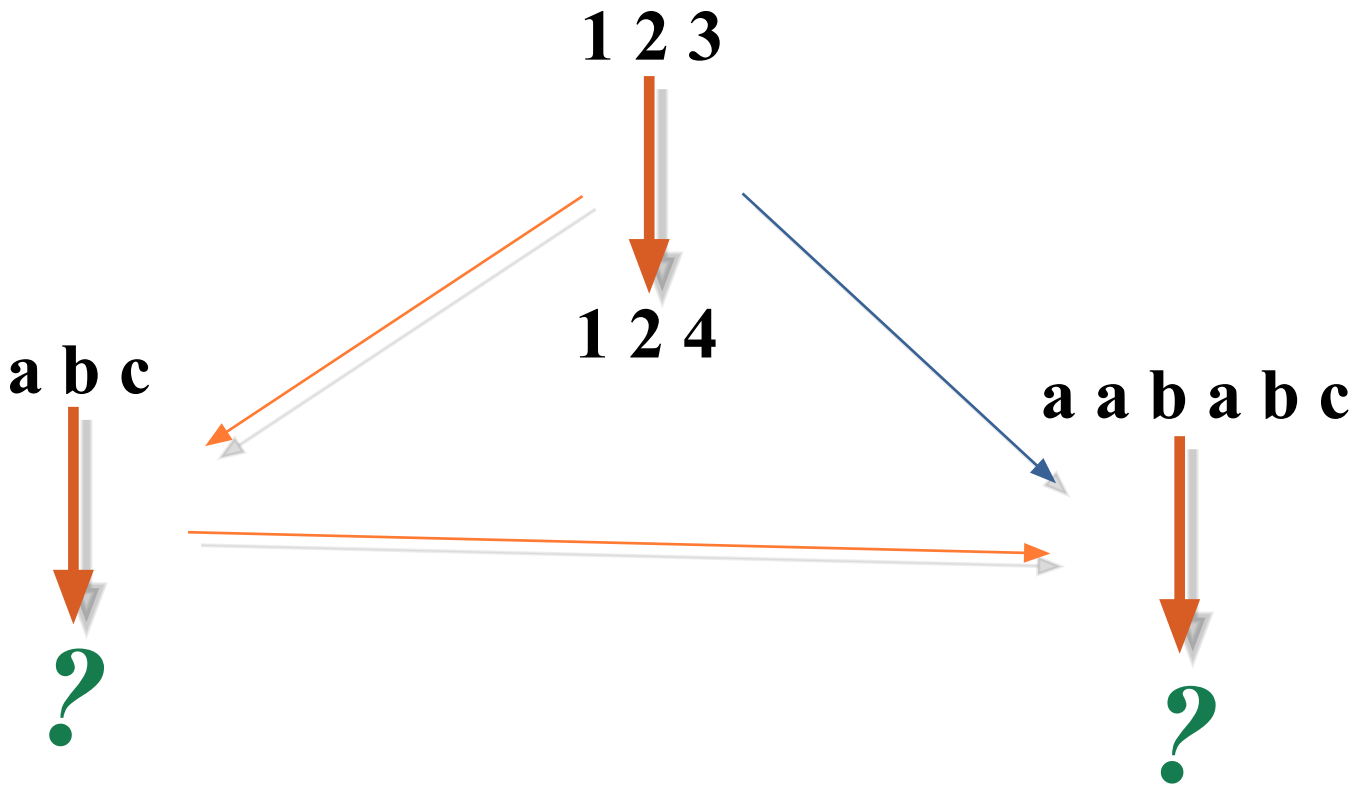
Analogie

- Transfert et séquence



Analogie

- Transfert et séquence : mieux, moins bien ?



Effets de séquences

Sortir du cadre ?

- Masses de données :
  - **NON** (?)
- Nouveaux défis :
  - **Pas clair**
- Pas de distribution, pas i.i.d. :
  - **Sans doute**

Apprendre et raisonner

À vous de jouer !