Apprentissage et circulation de l'information

Antoine Cornuéjols

Laboratoire de Recherche en Informatique
Université de Paris-Sud, Orsay

Journée de la complexité du 2 février 2006

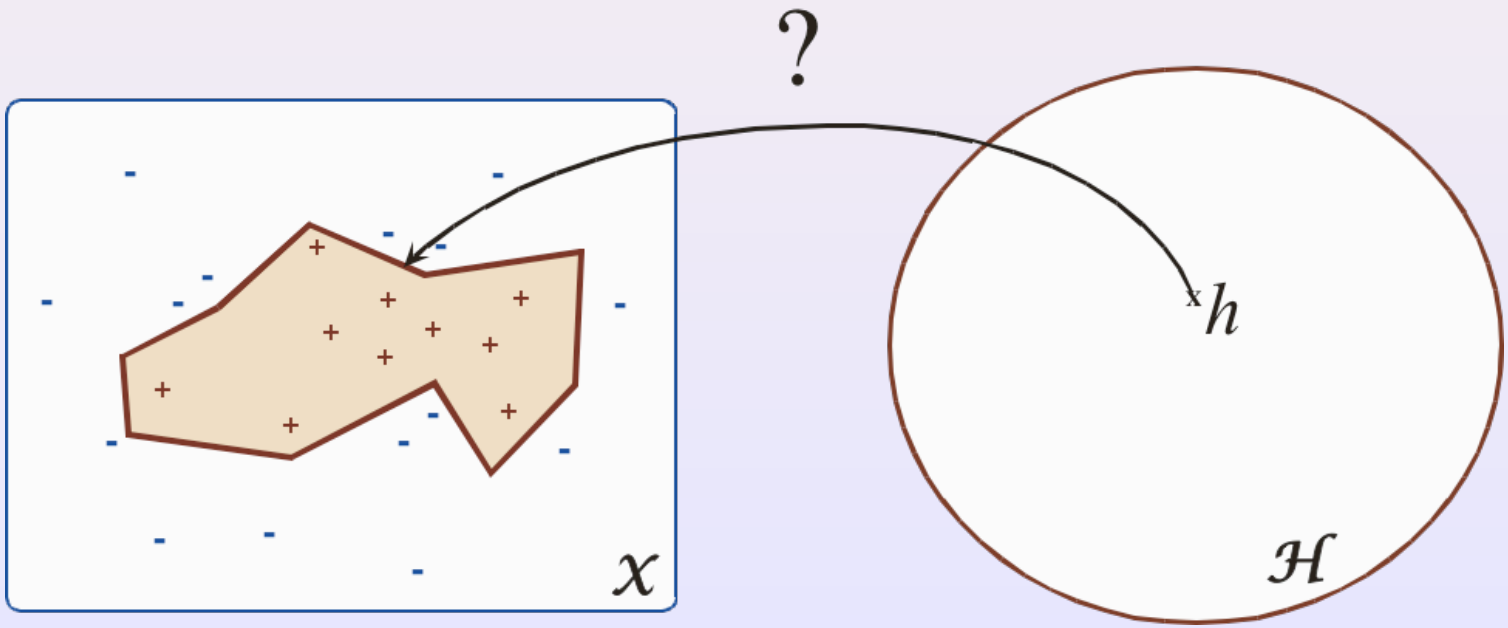
Plan

L'essence du problème

Définition [Mitchell, 82]

Apprentissage = Recherche dans un espace d'hypothèses

- Sous la contrainte des exemples d'apprentissage



Le paradigme

Définition (*Apprentissage*)

Apprentissage = Problème inverse mal posé

- À partir d'observations, trouver la loi f à laquelle obéissent ces observations

Le paradigme

Définition (*Apprentissage*)

Apprentissage = Problème inverse mal posé

- À partir d'observations, trouver la loi f à laquelle obéissent ces observations

Hypothèses

- Les observations sont des réalisations (i.i.d.) d'une variable aléatoire de loi f
- On cherche un estimateur $\hat{h}$ aussi proche que possible de la loi f

Le paradigme

Apprentissage = Problème inverse mal posé

- ... chercher $\hat{h}$ aussi proche que possible de la loi f

Proximité : Espérance de risque

$$R(h) = \mathbb{E}_{D_{\mathcal{X}\times\mathcal{Y}}}[h] = \int_{\mathcal{X}\times\mathcal{Y}} \underbrace{\ell(h(\boldsymbol{x}), f(\boldsymbol{x}))}_{\text{coût pour une observation}} \; d\boldsymbol{x}dy$$

Le paradigme

Apprentissage = Problème inverse mal posé

- ... chercher $\hat{h}$ aussi proche que possible de la loi f

Proximité : Espérance de risque

$$R(h) = \mathbb{E}_{D_{\mathcal{X}\times\mathcal{Y}}}[h] = \int_{\mathcal{X}\times\mathcal{Y}} \underbrace{\ell(h(\boldsymbol{x}), f(\boldsymbol{x}))}_{\text{coût pour une observation}} \, d\boldsymbol{x}dy$$

MRE

Choisir l'hypothèse $\hat{h}$ telle que $\hat{h} = \text{ArgMin}_{h\in\mathcal{H}}\left[R_{Emp}(h)\right]$

$$R_{\text{Emp}}(h) = \frac{1}{m}\sum_{(\boldsymbol{x}_i,u_i)\in S} \ell(h(\boldsymbol{x}_i), u_i)$$

La théorie statistique de l'apprentissage

Consistance du MRE

Conditions sous lesquelles le critère de MRE est correct ?

La théorie statistique de l'apprentissage

Consistance du MRE

Conditions sous lesquelles le critère de MRE est correct ?

$\rightarrow$ *Diversité* de l'espace des hypothèses $\mathcal{H}$ limitée

La théorie statistique de l'apprentissage

Consistance du MRE

Conditions sous lesquelles le critère de MRE est correct ?

$\rightarrow$ *Diversité* de l'espace des hypothèses $\mathcal{H}$ limitée

Qualité de l'estimation

$$|R(h) - R_{\mathrm{Emp}}(h)| \leq_P fct(\texttt{diversité}_{\mathcal{H}}, m)$$

La théorie statistique de l'apprentissage : actualités

Théorie

- Estimations plus fines de la diversité (capacité)
- Apprentissage actif : modification de la distribution en apprentissage

[Mar05] Jérémie Mary
Étude de l'apprentissage actif. Applications à la conduite d'expériences
Thèse, LRI, Orsay, Déc. 2005.

La théorie statistique de l'apprentissage : actualités

Théorie

- Estimations plus fines de la diversité (capacité)
- Apprentissage actif : modification de la distribution en apprentissage

... de nouvelles méthodes

- Minimisation du ϕ-risque empirique
  - Méthodes de votes (bagging, **boosting**, ...)
  - Méthodes à noyaux (**SVM**, ...)

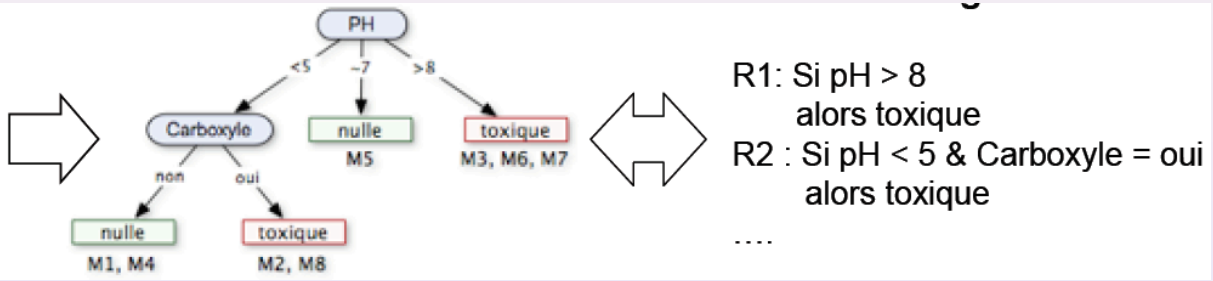
[Mar05] Jérémie Mary
Étude de l'apprentissage actif. Applications à la conduite d'expériences
Thèse, LRI, Orsay, Déc. 2005.

Le paradigme est adapté à ...

... l'analyse de données

BD $\Longrightarrow$ *régularités / prédictions*

	#Cycles	Masse	PH	Carboxyle	Activité
M1	1	faible	<5	non	nulle
M2	2	moyen	<5	oui	toxique
M3	0	moyen	>8	oui	toxique
M4	0	moyen	<5	non	nulle
M5	1	lourd	~7	non	nulle
M6	2	lourd	>8	non	toxique
M7	1	lourd	>8	non	toxique
M8	0	faible	<5	oui	toxique



[CM02] A. Cornuéjols and L. Miclet.
Apprentissage Artificiel. Concepts et Méthodes.
Eyrolles, 2002.

Le paradigme est adapté si ...

... peu de connaissances *a priori*

Seul critère : fidélité aux données

1. Prise en compte de la structure de $\mathcal{H}$ très pauvre
   - Relations de généralité
   - Niveaux d'abstraction
   - ...
2. Pas d'articulation à ce qui est déjà connu
   - Incrémentalité / Révision de théorie / Transfert
   - Critère de compréhensibilité

... monde supposé statique

Cadre i.i.d.

- Centralité du théorème central limite (et variantes)

Des problèmes difficiles

Évolution (dépendances) dans le temps

«Nouveaux »apprentissages

- **Dérive** de la dépendance cible
- **agents autonomes**
- par «démonstration »/ guidé par **un professeur**
- à **long terme** (*long-life learning*)
  - Articulation
  - Transfert

Plan

Transition de phase en induction

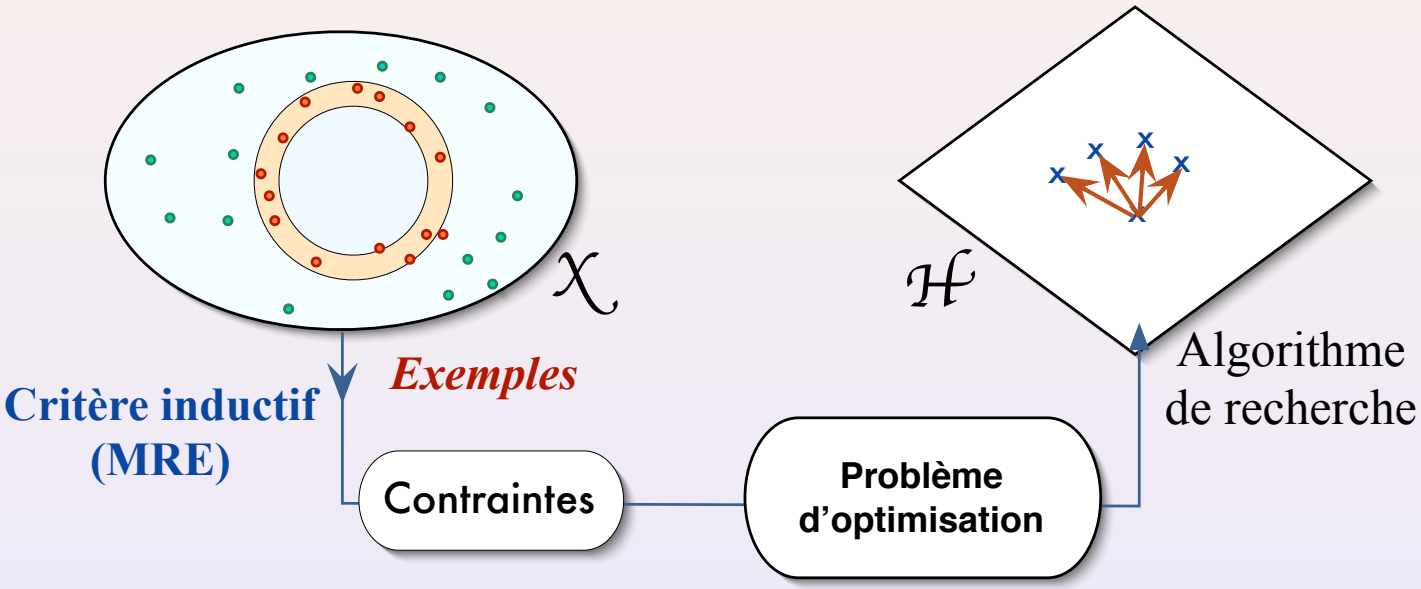
Collaborations

- Nicolas Baskiotis (*Doctorant*)
- Jérôme Maloberti (*Thèse*)
- Nicolas Pernot (*stage DEA*)
- Sandra Pinto (*stage DEA*)
- Raymond Ros (*Doctorant*)
- Michèle Sebag

- Mario Botta
- Attilio Giordana
- Lorenza Saitta

Sous quelles conditions l'induction est-elle possible ?

Des conditions sur le gain d'information

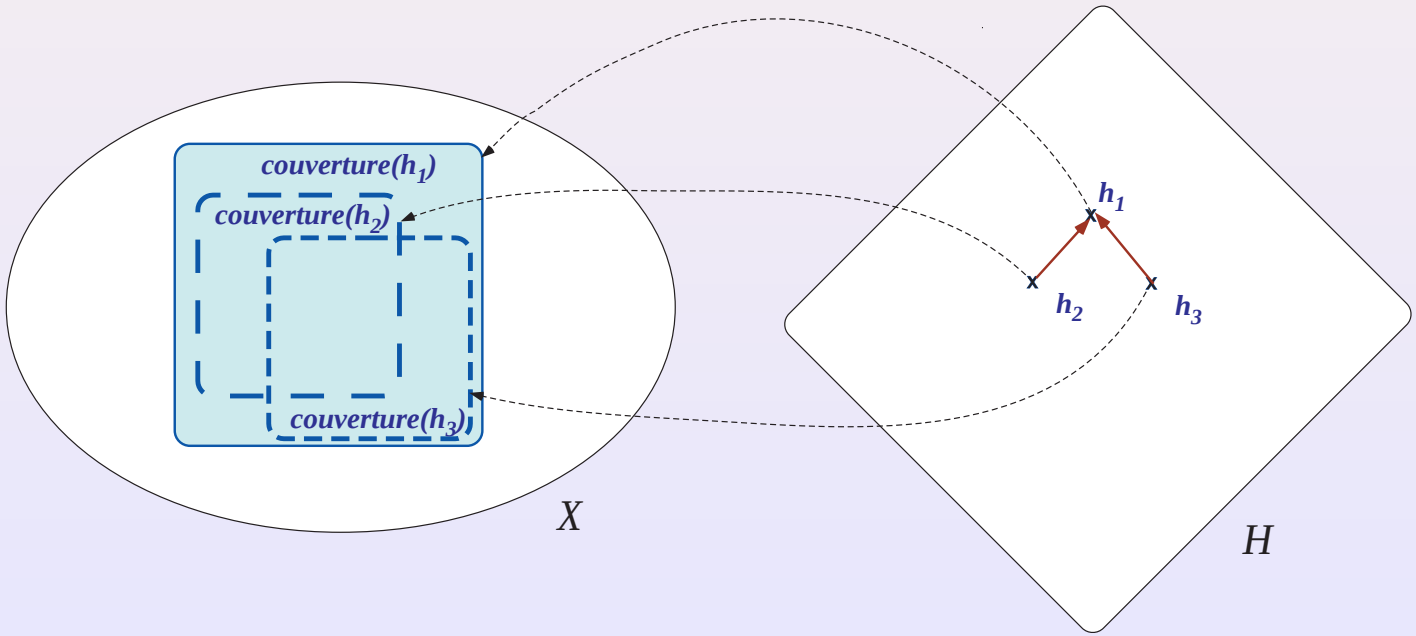


Les entrées doivent ***se traduire en «différences»*** *sur les hypothèses*

Des conditions sur le gain d'information

Gradient et taux de couverture

- **Le gradient** est lié aux **variations du taux de couverture**



Des conditions sur le gain d'information

Variations du taux de couverture

La mesure du taux de couverture apporte-t-elle de l'information ?

Des conditions sur le gain d'information

Variations du taux de couverture

La mesure du taux de couverture apporte-t-elle de l'information ?

Definition (Taux de couverture)

$$\tau(h) = P_{\mathcal{D}_{\mathcal{X}}}(h)$$

Des conditions sur le gain d'information

Variations du taux de couverture

La mesure du taux de couverture apporte-t-elle de l'information ?

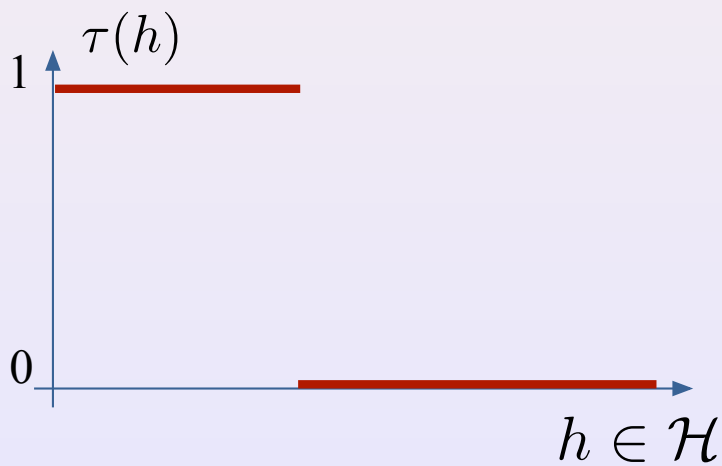
Definition (Taux de couverture)

$$\tau(h) = P_{\mathcal{D}_{\mathcal{X}}}(h)$$

*Étude des **variations de** $\tau(h)$*
*en fonction des **variations de** h (partie de $\mathcal{X}$)*

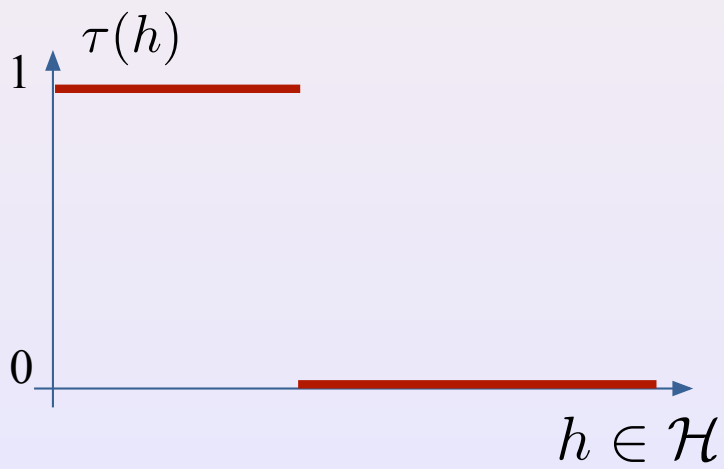
Des conditions sur le gain d'information

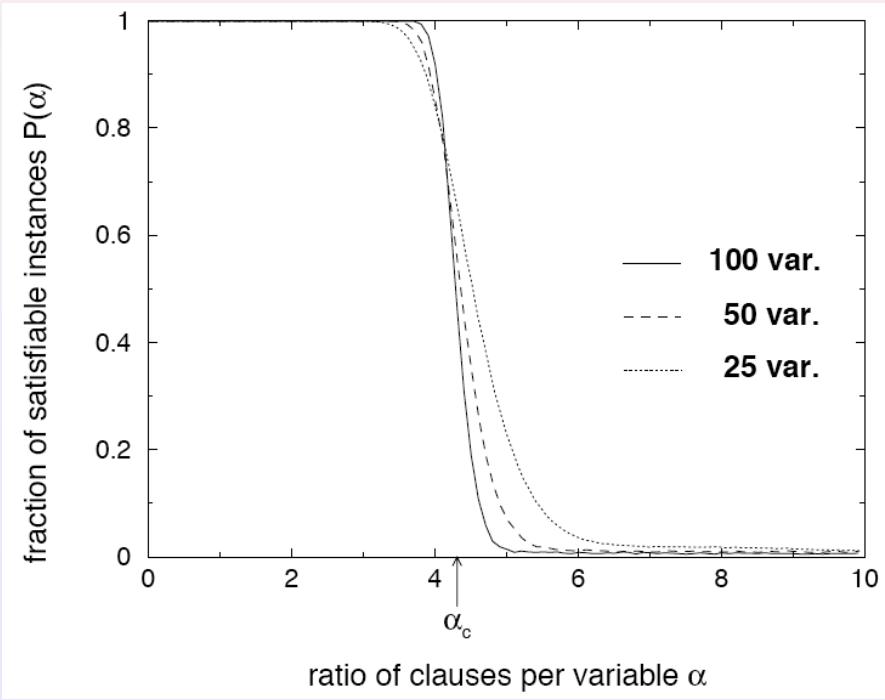
Un cas limite ... mais ...



Des conditions sur le gain d'information

Un cas limite ... mais ...





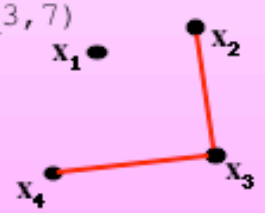
L'induction de connaissances structurelles

Programmation Logique Inductive (ILP)

INDUCE (Dietterich & Michalski, 1983)
SMART+ (Botta & Giordana, 1988, 1993)
FOIL (Quinlan, 1990)
PROGOL (Muggleton, 1994)
STILL (Sebag, 1998)
……..

```
active(d1)
lumo(d1, -1.246)
logp(d1, 4.23)
benzene(d1,[d1_6,d1_1,d1_2,d1_3,d1_4,d1_5])
atm(d1,d1_1,c,22,-0.117)
atm(d1,d1_2,c,22,-0.117)
atm(d1,d1_3,c,22,-0.117)
atm(d1,d1_4,c,195,-0.087)
atm(d1,d1_5,c,195,0.013)
bond(d1,d1_1,d1_2,7)
bond(d1,d1_2,d1_3,7)
bond(d1,d1_3,d1_4,7)
bond(d1,d1_4,d1_5,7)
bond(d1,d1_5,d1_6,7)
```

```
Nonactive(d167)
lumo(d167, -1.246)
logp(d167, 4.23)
atm(d167,d167_1,n,22,-0.117)
atm(d167,d167_2,c,22,-0.117)
atm(d167,d167_3,n,22,-0.117)
atm(d167,d167_4,c,195,-0.087)
atm(d167,d167_5,n,195,0.013)
bond(d167,d167_1,d167_2,7)
bond(d167,d167_2,d167_3,7)
....
```



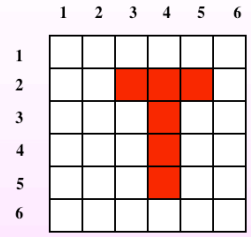
Relational Learner

$$\texttt{active}(M) :- \neg\, \text{chrg}(x_1, [-0.2]) \wedge \neg\, \text{type}(x_2, [N]) \wedge \neg\, \text{anm}(x_3, [22]) \wedge \neg\, \text{chrg}(x_3, [-0.6, -0.4]) \wedge \neg\, \text{type}(x_4, [H, N, O]) \wedge \text{bound}(x_2, x_3) \wedge \text{bound}(x_3, x_4) \wedge \text{atm}(M, x_1) \wedge \text{atm}(M, x_2) \wedge \text{atm}(M, x_3) \wedge \text{atm}(M, x_4)$$

L'induction de connaissances structurelles

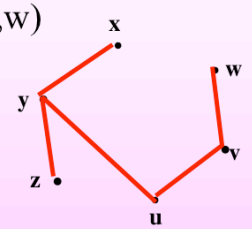
Programmation Logique Inductive (ILP)

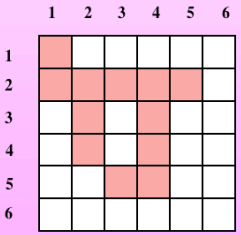
Matching Problem



Formula φ

$\varphi(x,y,z,u,v,w) = red(x) \wedge red(y) \wedge red(z) \wedge red(u) \wedge red(v) \wedge red(w) \wedge west(x,y) \wedge adj(x,y) \wedge west(y,z) \wedge adj(y,z) \wedge north(y,u) \wedge adj(y,u) \wedge north(u,v) \wedge adj(u,v) \wedge north(v,w) \wedge adj(v,w)$





1 solution

red(x)

x
a_{11}
a_{21}
…
a_{54}

west(x,y)

x	y
a_{21}	a_{22}
a_{21}	a_{23}
a_{21}	a_{24}
….	…
a_{32}	a_{34}
…	…
a_{53}	a_{54}

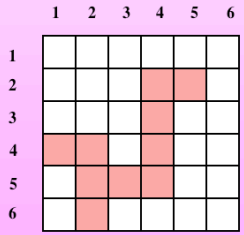
north(x,y)

x	y
a_{11}	a_{21}
a_{22}	a_{32}
a_{22}	a_{42}
a_{22}	a_{32}
a_{32}	a_{42}
…	…
a_{44}	a_{54}

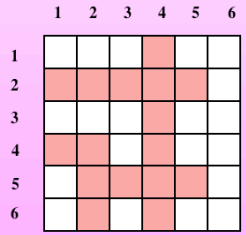
adj(x,y)

x	y
a_{11}	a_{21}
a_{21}	a_{11}
a_{21}	a_{22}
a_{22}	a_{21}
a_{22}	a_{23}
a_{22}	a_{32}
…	…
a_{53}	a_{54}

Universe U

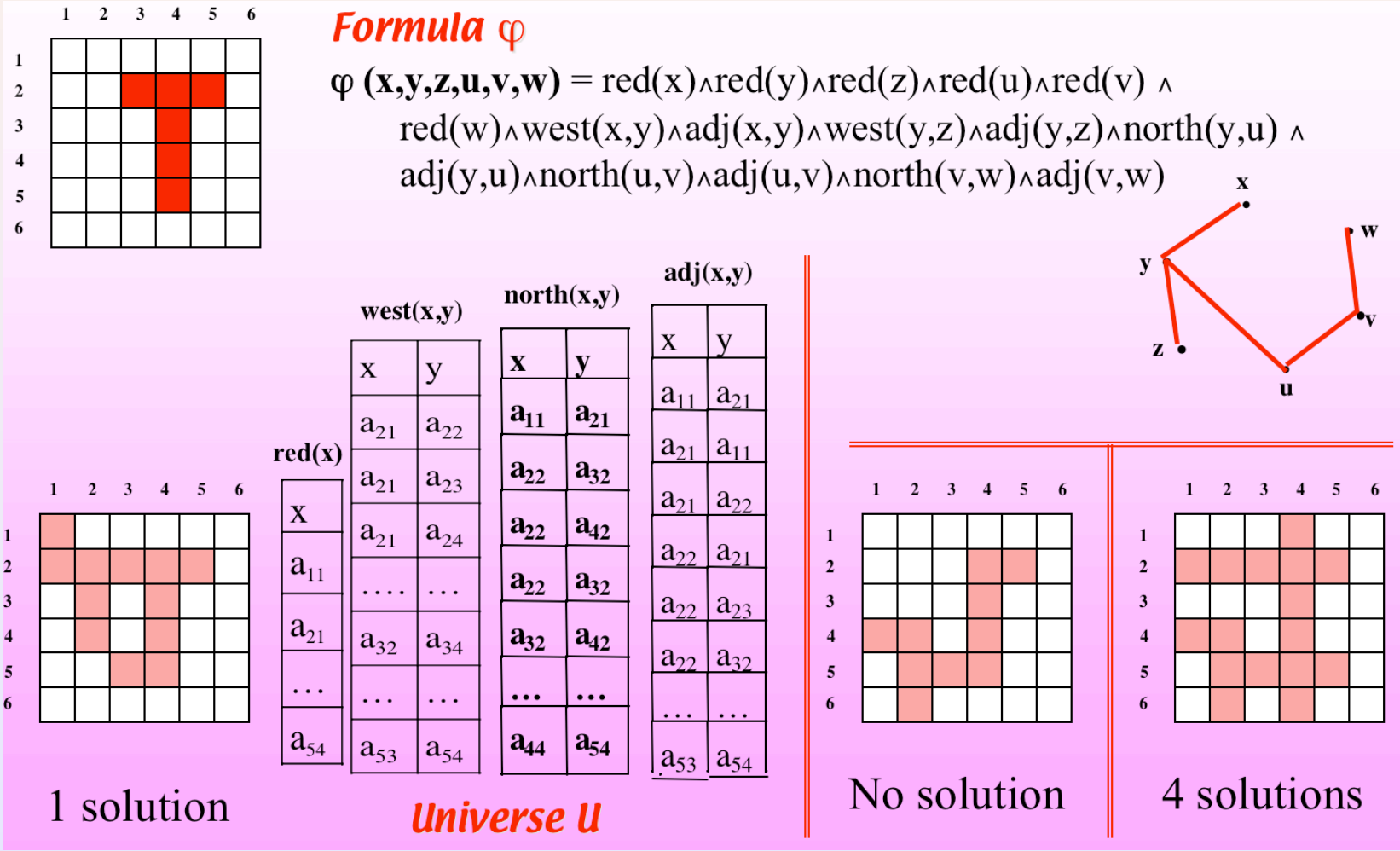


No solution

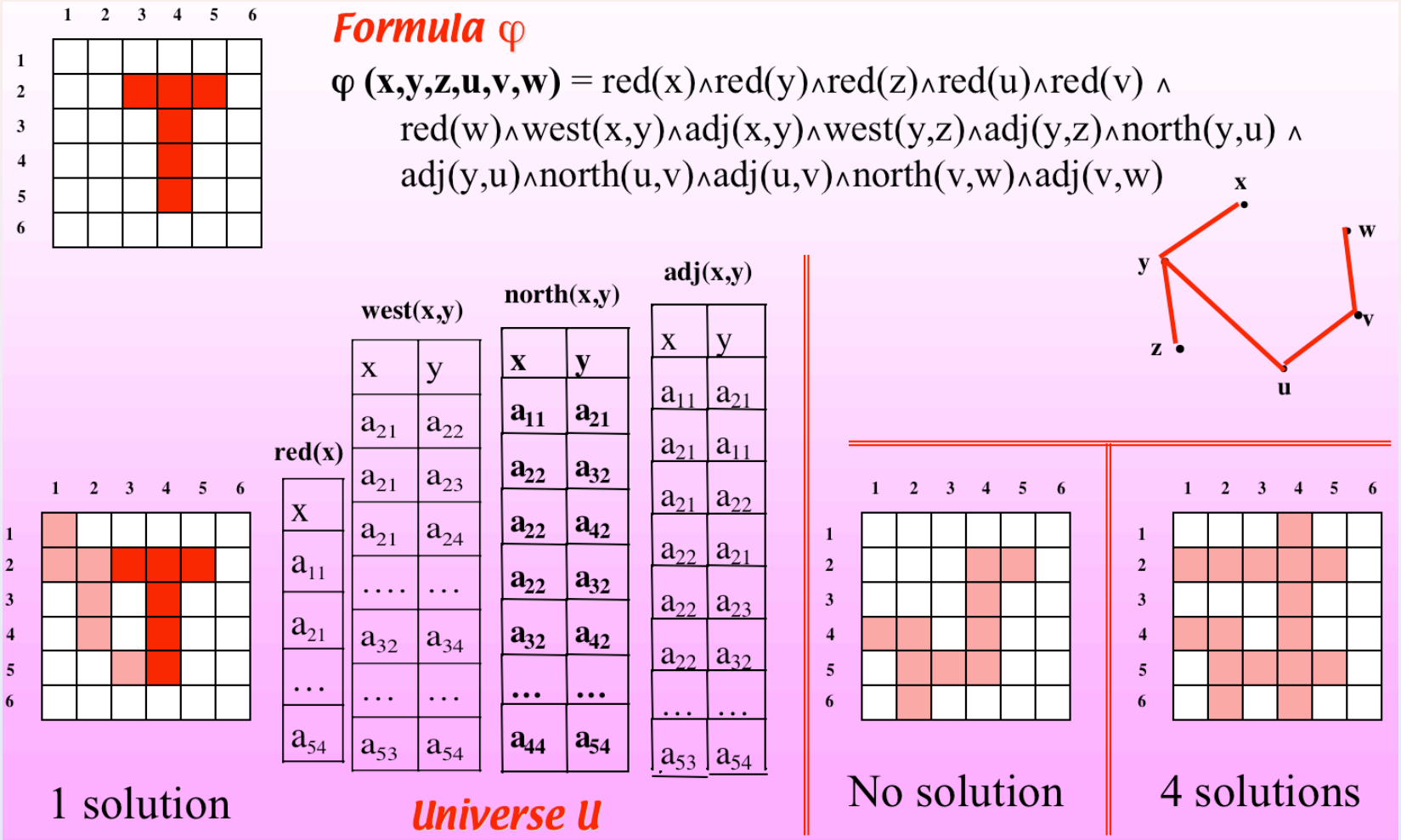


4 solutions

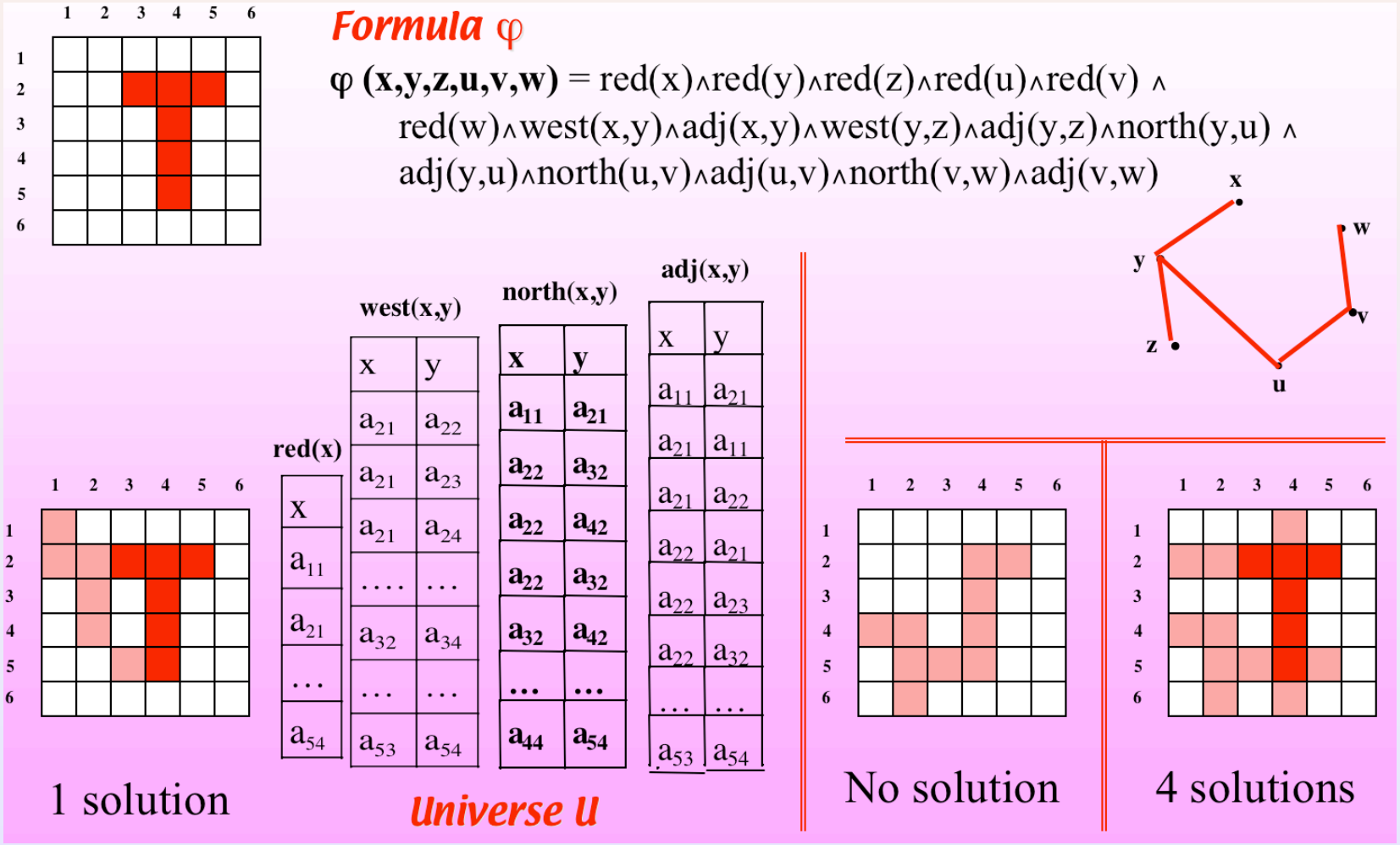
ILP et satisfaction de contraintes



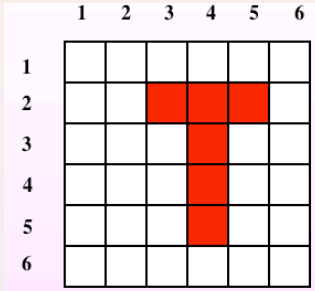
ILP et satisfaction de contraintes



ILP et satisfaction de contraintes

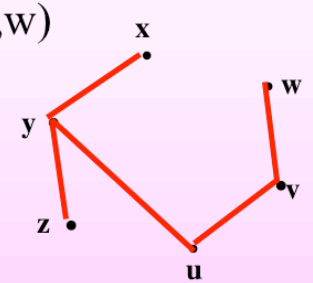


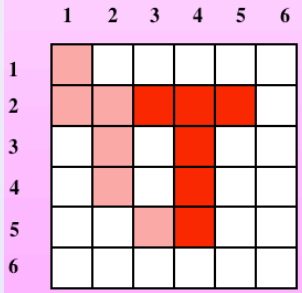
ILP et satisfaction de contraintes



Formula φ

φ **(x,y,z,u,v,w)** = red(x)∧red(y)∧red(z)∧red(u)∧red(v) ∧
red(w)∧west(x,y)∧adj(x,y)∧west(y,z)∧adj(y,z)∧north(y,u) ∧
adj(y,u)∧north(u,v)∧adj(u,v)∧north(v,w)∧adj(v,w)





1 solution

red(x)

x
a_{11}
a_{21}
…
a_{54}

west(x,y)

x	y
a_{21}	a_{22}
a_{21}	a_{23}
a_{21}	a_{24}
….	…
a_{32}	a_{34}
…	…
a_{53}	a_{54}

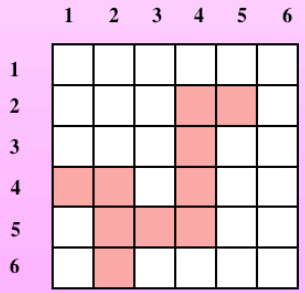
north(x,y)

x	**y**
$\mathbf{a_{11}}$	$\mathbf{a_{21}}$
$\mathbf{a_{22}}$	$\mathbf{a_{32}}$
$\mathbf{a_{22}}$	$\mathbf{a_{42}}$
$\mathbf{a_{22}}$	$\mathbf{a_{32}}$
$\mathbf{a_{32}}$	$\mathbf{a_{42}}$
…	…
$\mathbf{a_{44}}$	$\mathbf{a_{54}}$

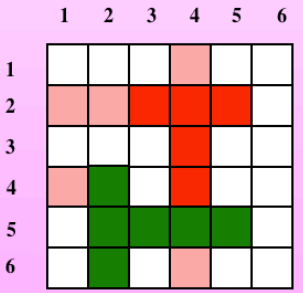
adj(x,y)

x	y
a_{11}	a_{21}
a_{21}	a_{11}
a_{21}	a_{22}
a_{22}	a_{21}
a_{22}	a_{23}
a_{22}	a_{32}
…	…
a_{53}	a_{54}

Universe U

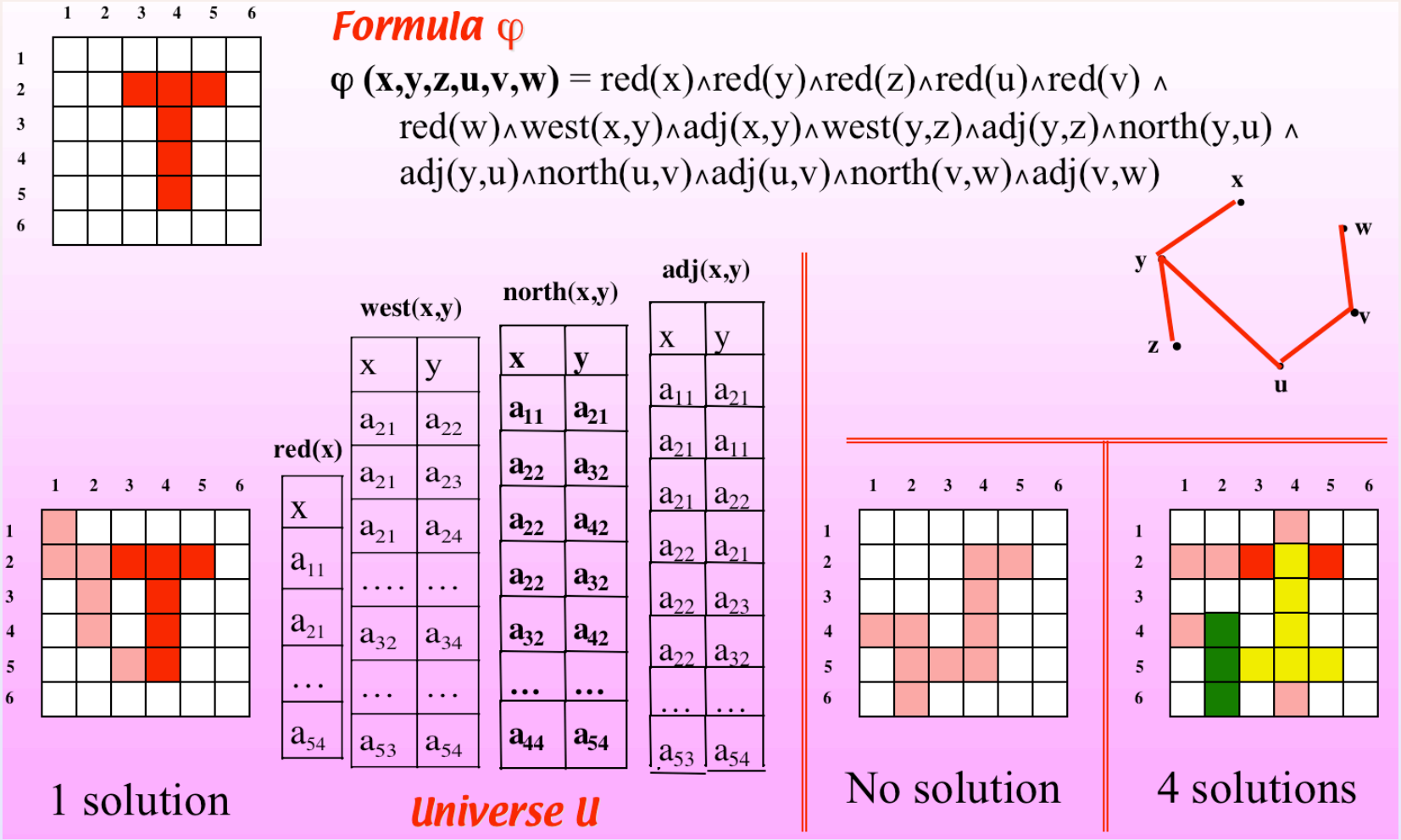


No solution

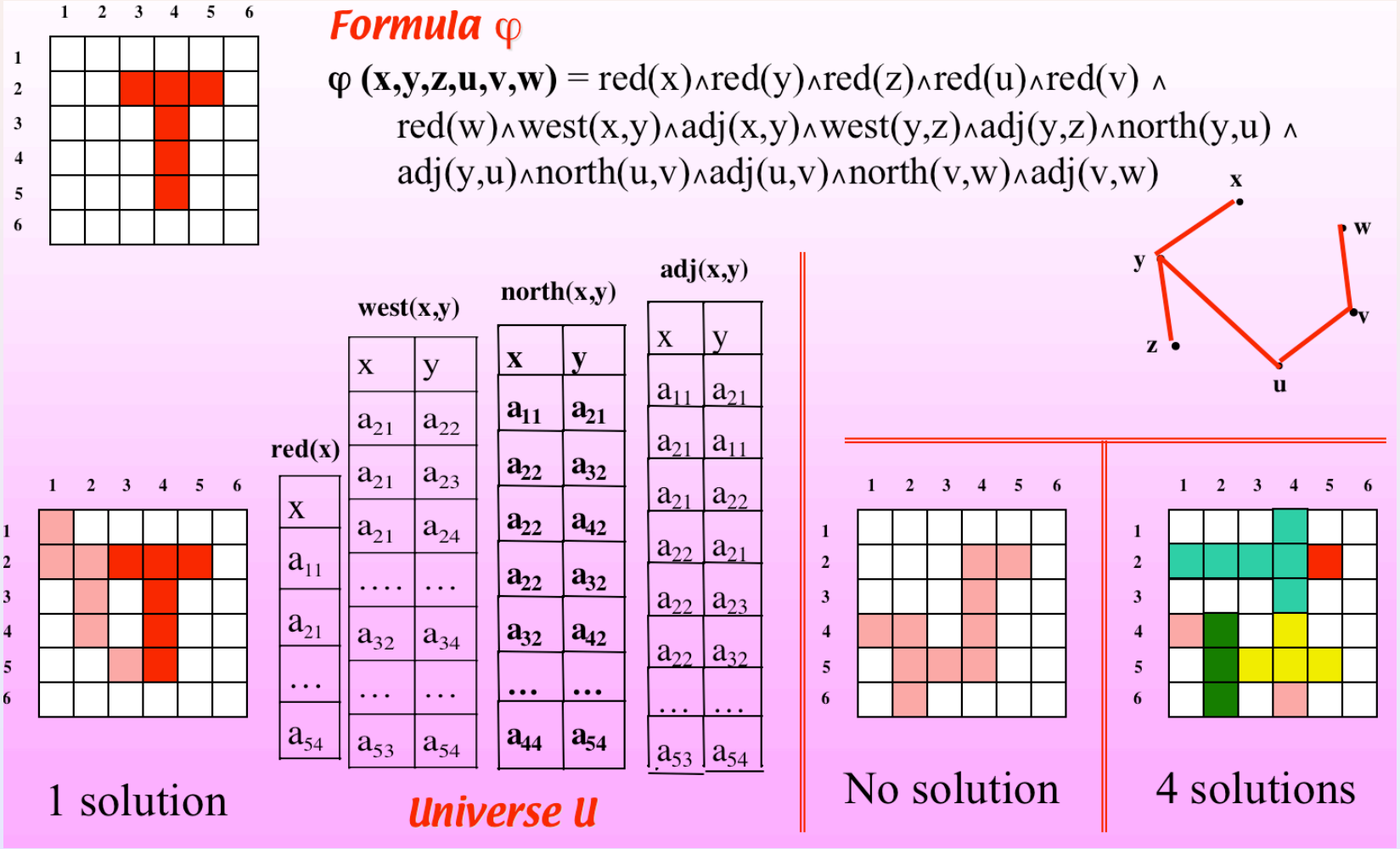


4 solutions

ILP et satisfaction de contraintes

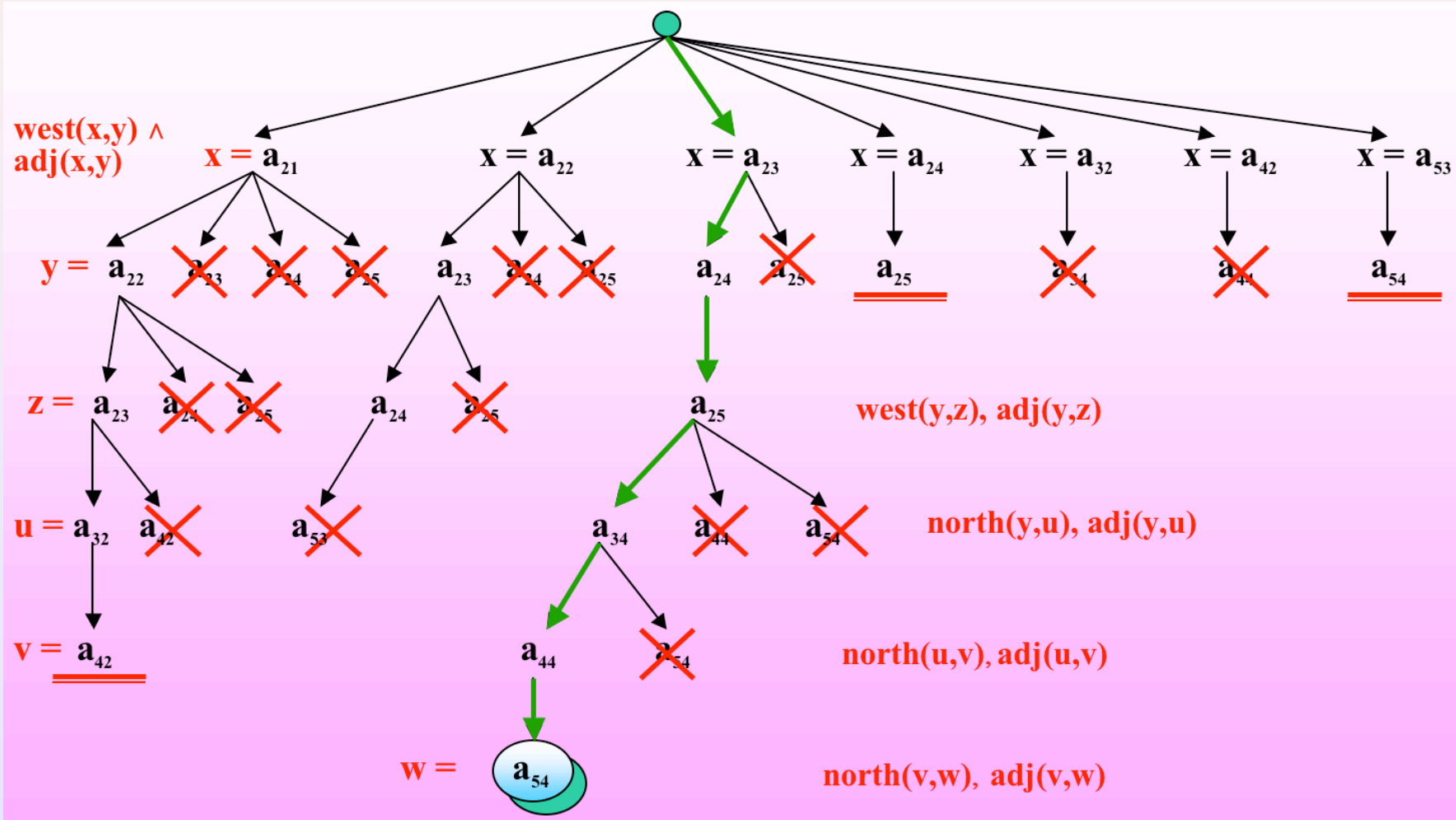


ILP et satisfaction de contraintes



ILP et satisfaction de contraintes

Complexité de la recherche



Principe des expériences

Étude en cas moyen

Génération aléatoire de problèmes

Hypothèses

- ***n*** : nombre de *variables* dans l'hypothèse $\underline{h}$ testée,
- ***m*** : nombre de *symboles de prédicats* dans $\underline{h}$,

Exemples

- ***L*** : nombre total de *constantes* dans l'exemple $\underline{e}$,
- ***N*** : nombre de *littéraux* construits sur chaque symbole de prédicat dans $\underline{e}$.

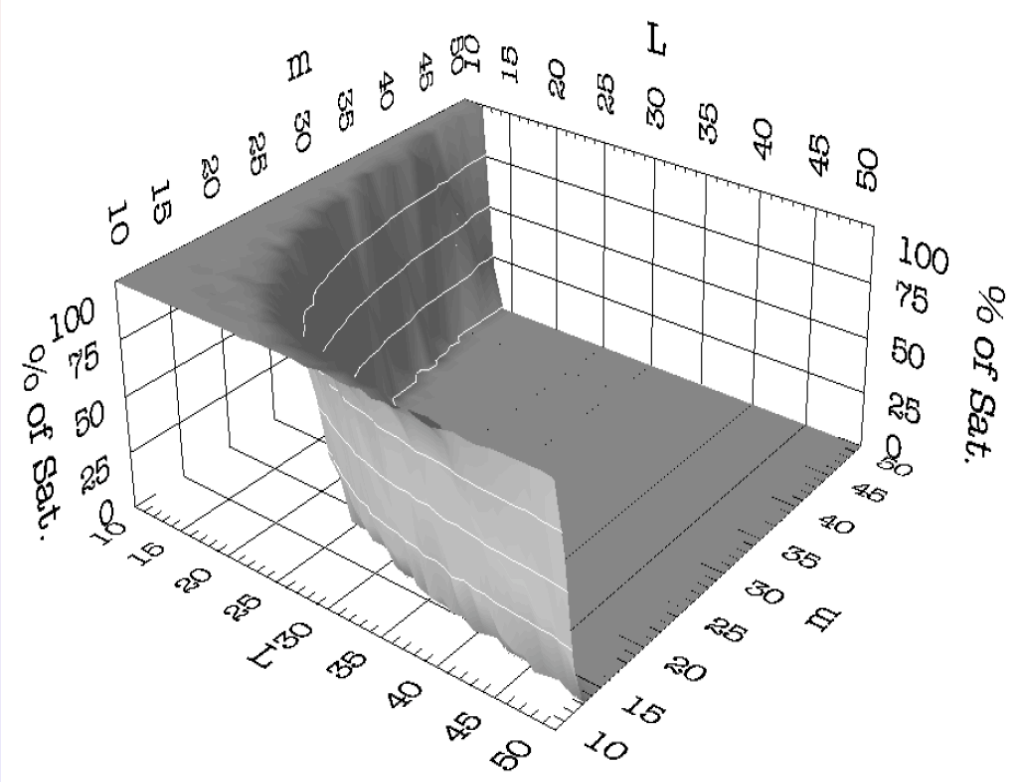
→ ***Étude dans le cas moyen***

Des conditions sur le gain d'information

Distribution uniforme suivant des paramètres de contrôle : le cas de l'ILP

Paramètres de contrôle :

- ***n*** : nombre de *variables* dans l'hypothèse *h* testée,
- ***m*** : nombre de *symboles de prédicats* dans *h*,
- ***L*** : nombre total de *constantes* dans l'exemple *e*,
- ***N*** : nombre de *littéraux* construits sur chaque symbole de prédicat dans *e*.



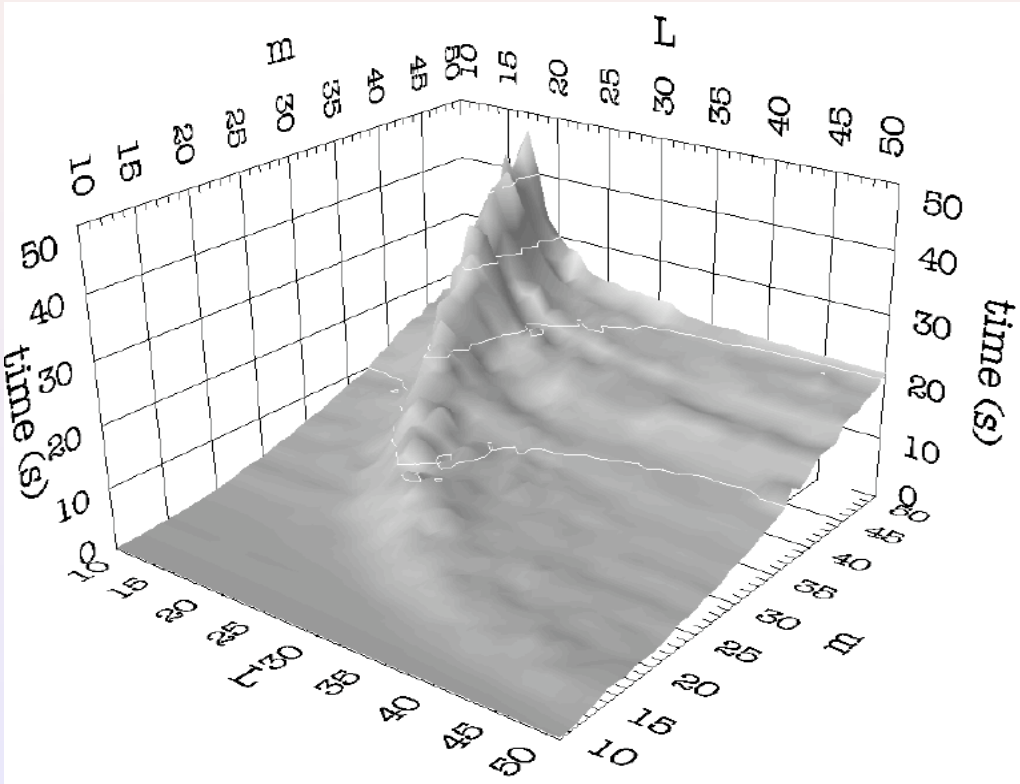
[bot03] Botta, M., A. Giordana, L. Saitta, and M. Sebag
Relational learning as search in a critical region.
Journal of Machine Learning Research, 4, 431-463, 2003.

Coût du test de couverture

Distribution uniforme suivant des paramètres de contrôle : le cas de l'ILP

Paramètres de contrôle :

- ***n*** : nombre de *variables* dans l'hypothèse *h* testée,
- ***m*** : nombre de *symboles de prédicats* dans *h*,
- ***L*** : nombre total de *constantes* dans l'exemple *e*,
- ***N*** : nombre de *littéraux* construits sur chaque symbole de prédicat dans *e*.

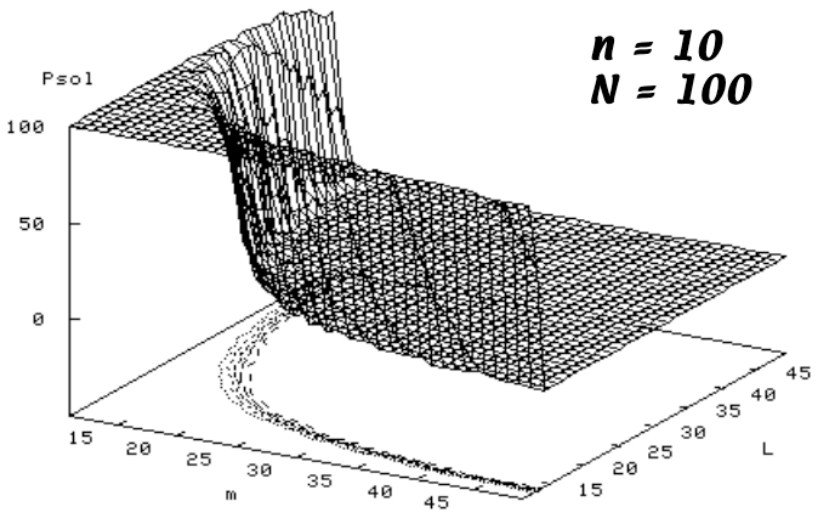


[mal04] Maloberti, J. and M. Sebag
Fast Theta-Subsumption with Constraint Satisfaction Algorithms.
Machine Learning Journal, 55, 137-174, 2004.

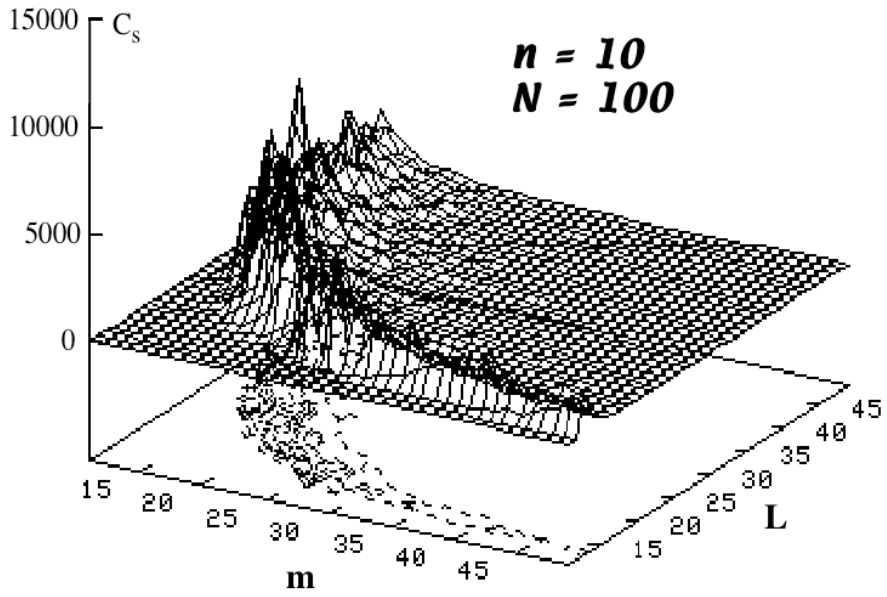
ILP et satisfaction de contraintes

Probabilité de couverture et complexité du test

L = Number of constant in the universe
m = Number of (binary) predicates in a formula



n = 10
N = 100



n = 10
N = 100

n = 4, 6, 10, 12, 14
N = 30, 50, 100, 130

100 problems for each pair (m,L)

Set of 900,000 matching problems

Conséquences sur l'apprentissage de concept

Paramètre :

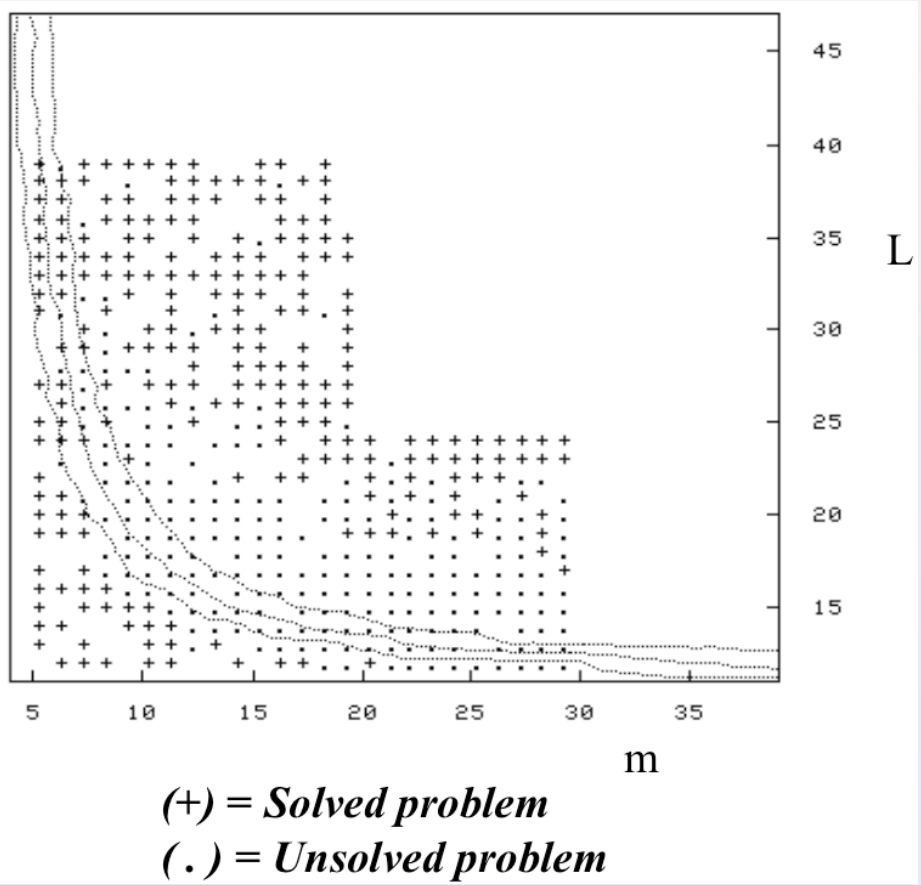
- ***n*** = 4
- ***N*** = 100

Contours :

- $P_{sol} = 0.9$
- $P_{sol} = 0.5$
- $P_{sol} = 0.1$

Succès :

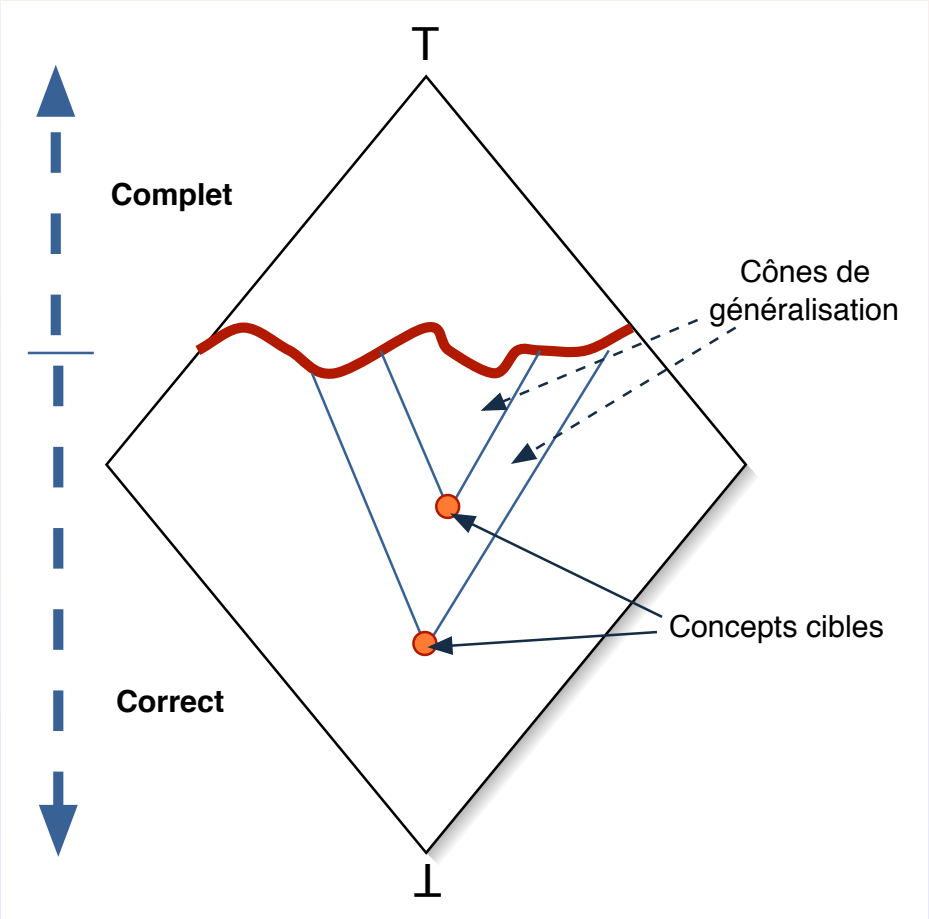
- \+ : précision $>$ 80%
- . : précision $<$ 80%



Conséquences sur l'apprentissage de concept

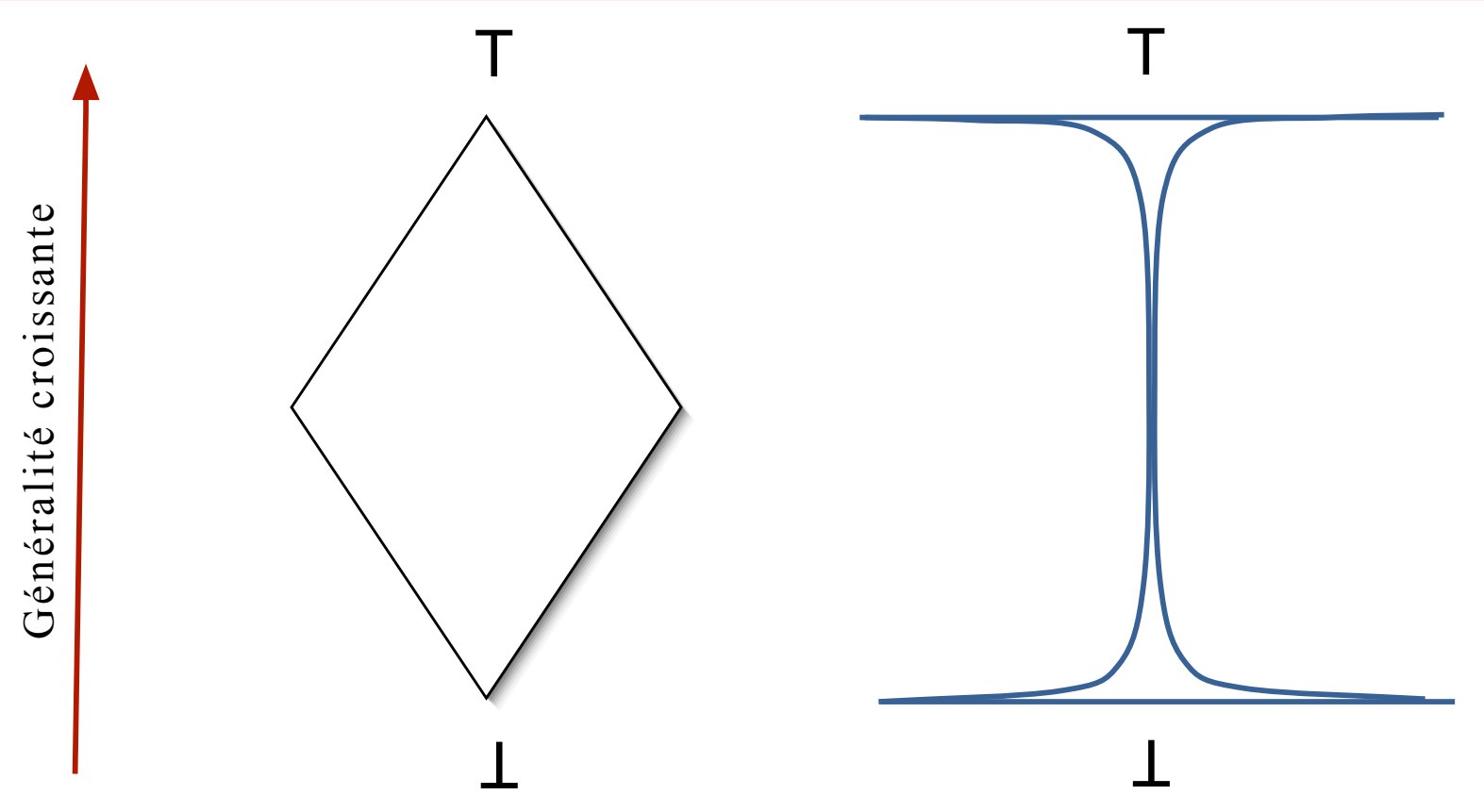
	Probl.	#Nof Clauses	Complexity	Class.Rate (LS) [%]	Class. Rate (TS) [%]	CPU Time [sec]	Avg. #N of Models
	LP_1*	10	<7-13> 8.9	88	50	398.2	1.7
	LP_2*	11	<6-11> 8.6	92	53	624.7	2.0
NO	LP_3*	15	<7-11> 8.9	98.5	52	513.9	4.9
	LP_4*	1	6	100	100	43.3	2.1
	LP_5*	1	6	99.9	100	132.6	1.25
	LP_6	12	<1-12> 6	81	58	825.4	10.7
Approx	LP_7	1	6	100	96	73.4	34.6
	LP_8	6	<1-11> 5	98.5	75.3	723.8	1.4
	LP_9	1	9	100	99.6	620.1	1.0
YES	LP_{10}	1	6	100	99.6	36.9	4.2
	LP_{11}	1	6	100	99.6	72.2	9.1

Analyse


T
Complet
Cônes de
généralisation
Concepts cibles
Correct
⊥

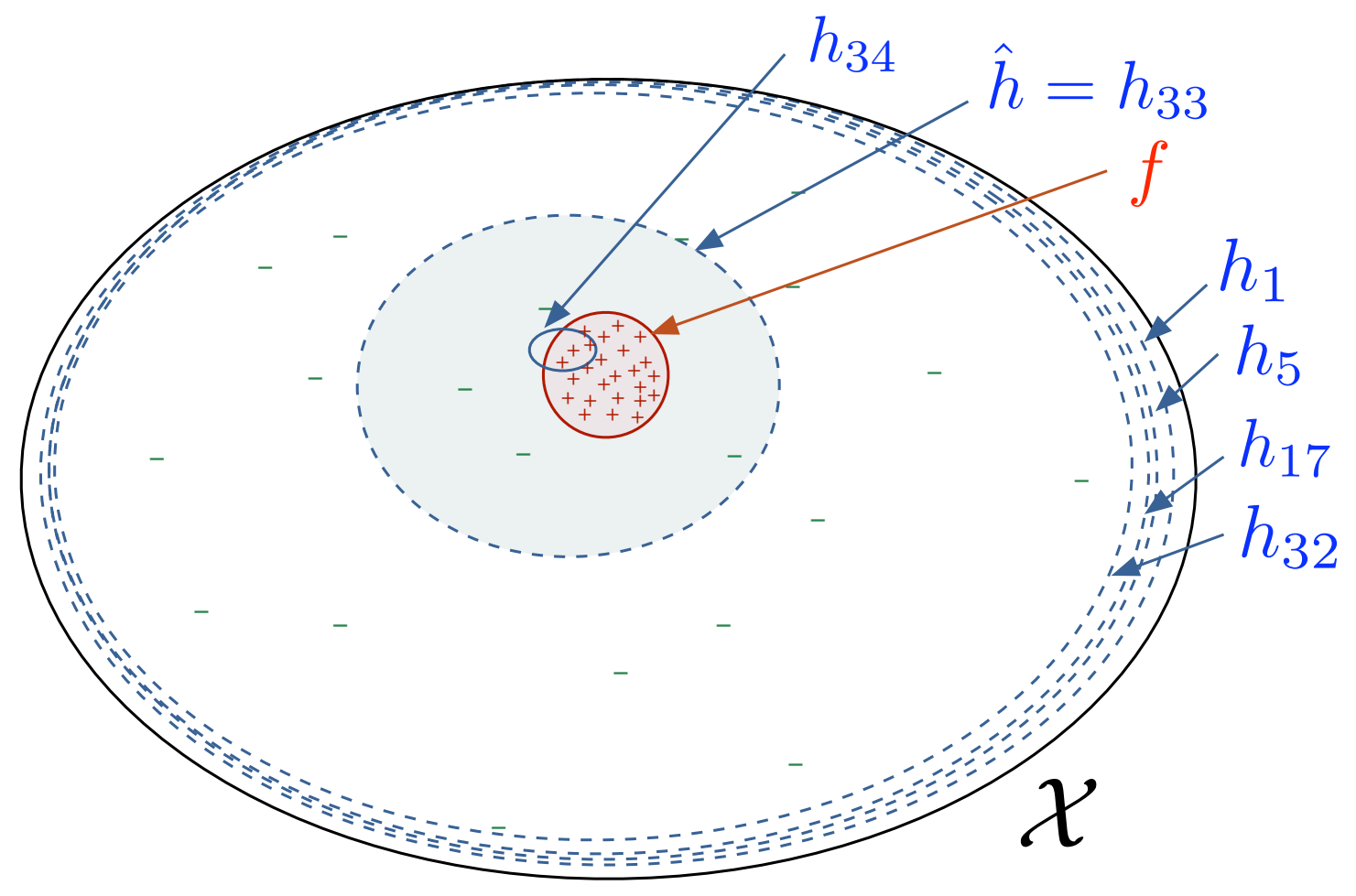
Des conditions sur le gain d'information

Transition de phase et espace des versions



Des conditions sur le gain d'information

Une distribution uniforme ... suspecte ?



Des conditions sur le gain d'information

Bilan

- $\exists$ ***transition de phase*** dans les variations de taux de couverture
  - observé en ILP
- ***Impact considérable*** sur les performances
- ***Non prévu*** par l'analyse statistique

Ce phénomène dépend de :

1. $\mathcal{L}_{\mathcal{H}}$: langage des hypothèses
2. $\mathcal{L}_{\mathcal{X}}$: langage des exemples

Des conditions sur le gain d'information

Questions ouvertes

1. ***Quels sont les langages affectés ?***

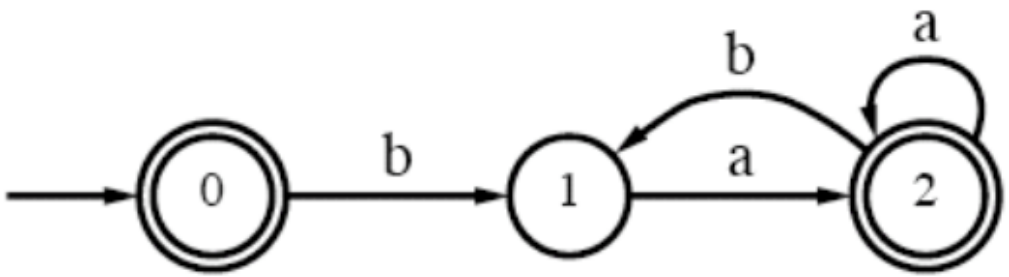
2. ***Peut-on contourner le problème ?***

Inférence grammaticale : rappels

- **Entrée** : *chaînes* sur un alphabet Σ, de longueur ℓ
- **Sortie** : *Automate fini* (langage régulier)
  - DFA : *Deterministic Finite Automata*
  - NFA : *Non deterministic Finite Automata*

Échantillon positif : $S^+ = \{\texttt{ba}, \texttt{baa}, \texttt{baba}, \lambda\}$

FSA (DFA) couvrant S^+



Gain d'information en **inférence grammaticale**

Distribution uniforme avec paramètres de contrôle

Paramètres de contrôle :

- Q états
- B d'arcs sortants / état
- L lettres / arc
- Fraction $a \in [0, 1]$ d'états acceptants
- Taille $|\Sigma|$ de l'alphabet
- Longueur ℓ des exemples testés.

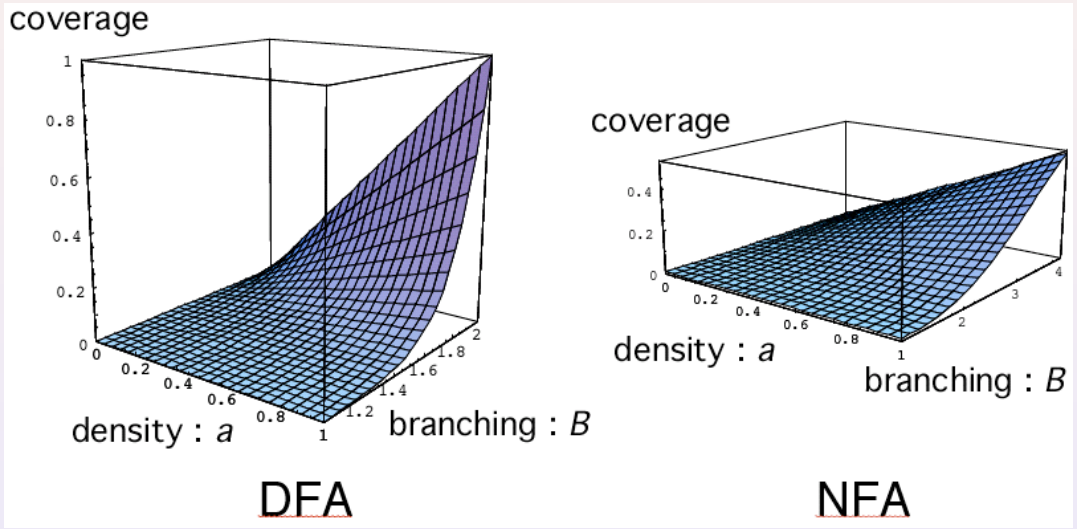
[Pin01] S. Pinto.
Etude du phénomène de transition de phase dans l'induction supervisée.
Rapport de DEA (LRI, Univ. Paris-Sud, Orsay), 2001.

Gain d'information en **inférence grammaticale**

Distribution uniforme avec paramètres de contrôle

Paramètres de contrôle :

- ***Q*** états
- ***B*** d'arcs sortants / état
- ***L*** lettres / arc
- Fraction ***a*** $\in [0, 1]$ d'états acceptants
- Taille $|\Sigma|$ de l'alphabet
- Longueur ℓ des exemples testés.

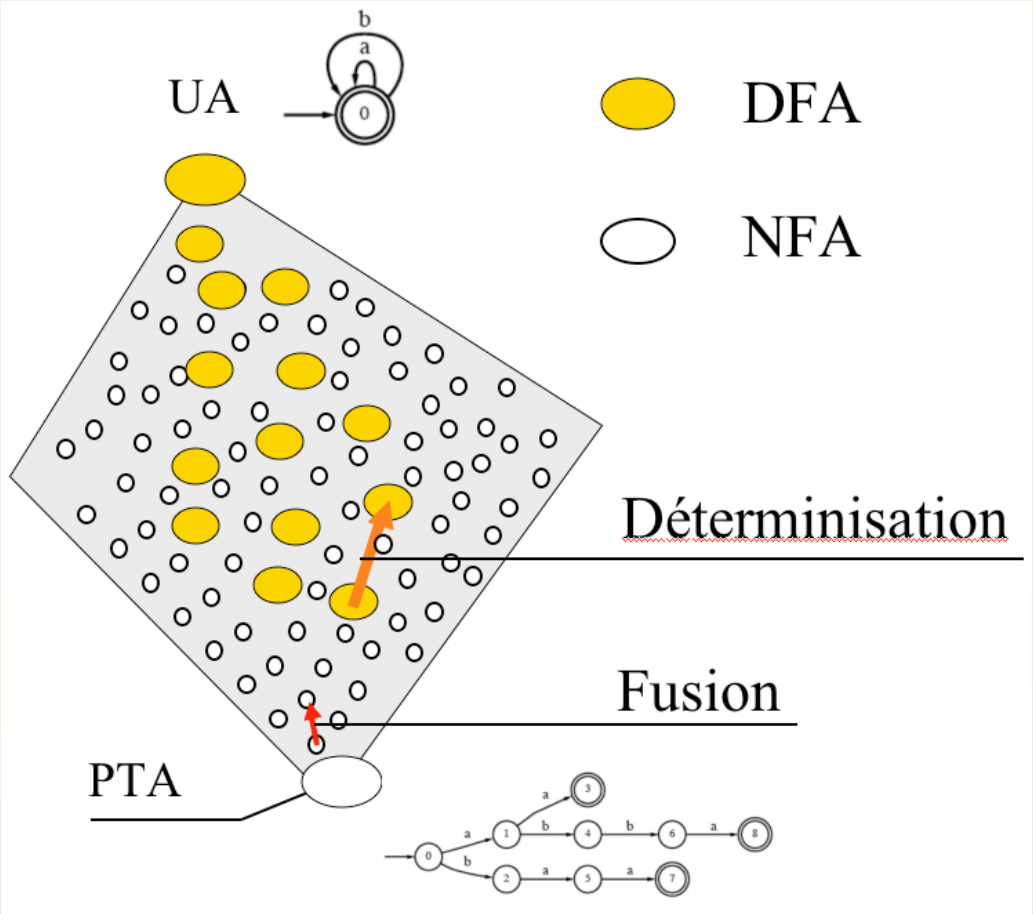


$$P(\text{accept}) = \begin{cases} a \cdot (\frac{B \cdot L}{|\Sigma|})^{\ell} & \text{pour un DFA} \\ a \cdot [1 - (1 - \frac{L}{|\Sigma|})^{B}]^{\ell} & \text{pour un NFA} \end{cases}$$

[Pin01] S. Pinto.
Etude du phénomène de transition de phase dans l'induction supervisée.
Rapport de DEA (LRI, Univ. Paris-Sud, Orsay), 2001.

Gain d'information en **inférence grammaticale**

Principe des algorithmes d'apprentissage



Gain d'information en **inférence grammaticale**

Étude sur l'espace d'hypothèses **effectivement** exploré

Protocole expérimental

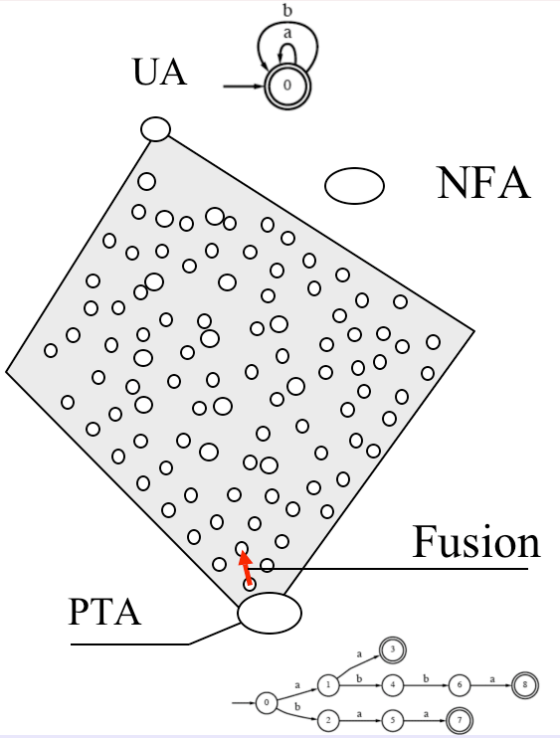
1. Génération aléatoire d'un ***échantillon d'apprentissage*** : $|S^+|$ (= 200) chaînes de taille ℓ
2. Construction du ***PTA*** pour chaque échantillon S^+
3. ***Calcul de chemins de généralisation partant du PTA*** :
   - **Fusions aléatoires**
   - **Couverture** calculée **pour chaque automate engendré** (sur un *ensemble test* : 1000 chaînes aléatoires $\notin$ ens. d'apprentissage)

Expériences

- $|\Sigma| = \{2, 4, 8\}$
- $\ell = \{4, 8, 16, 32\}$
- 50 PTAs $\times$ 20 trajectoires aléatoires = 1000 trajectoires ($\approx$ 270 000 automates)

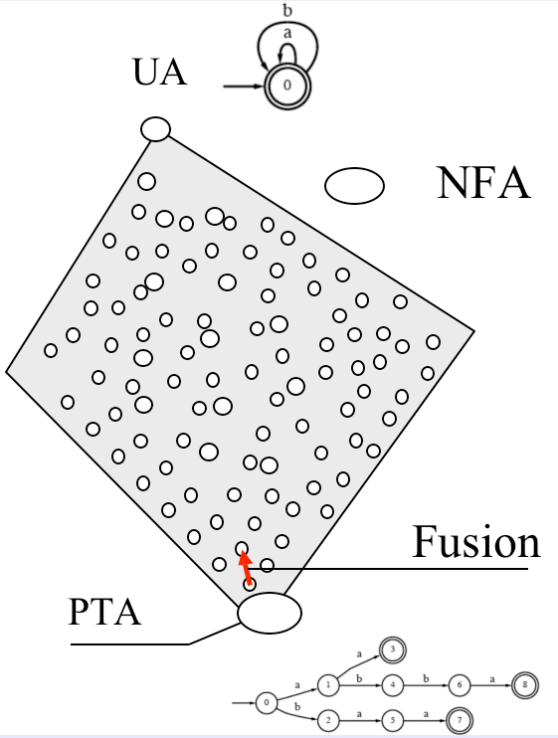
Gain d'information en **inférence grammaticale**

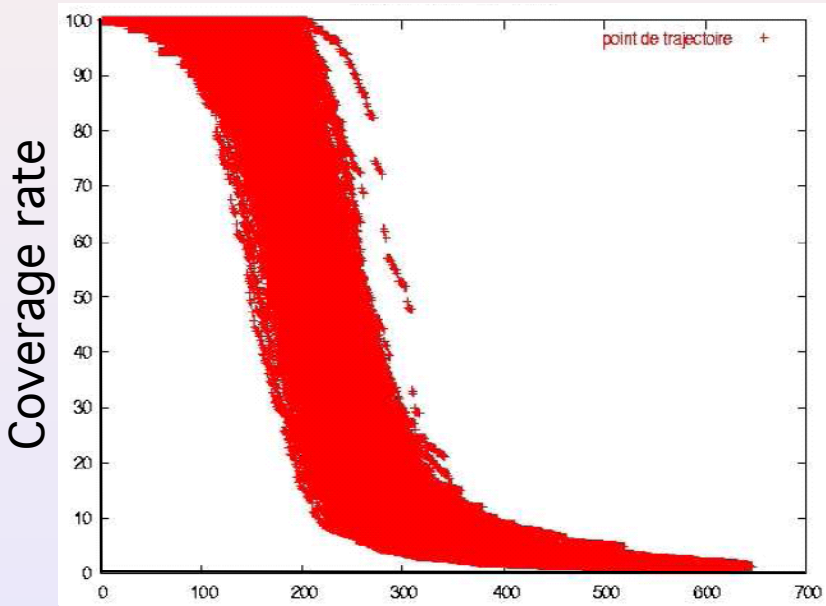
Cas non-déterministe : **NFA**



Gain d'information en **inférence grammaticale**

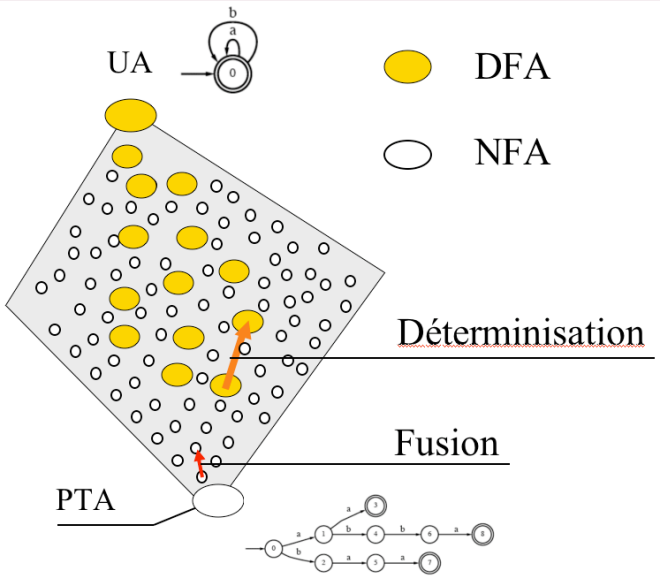
Cas non-déterministe : **NFA**





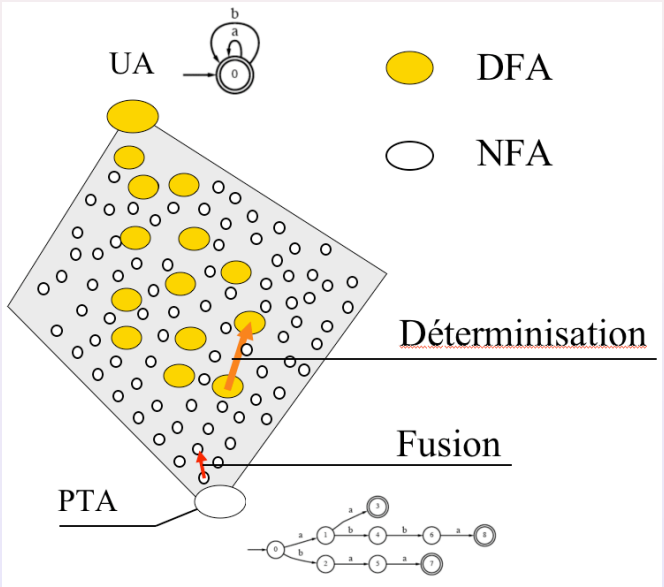
Gain d'information en **inférence grammaticale**

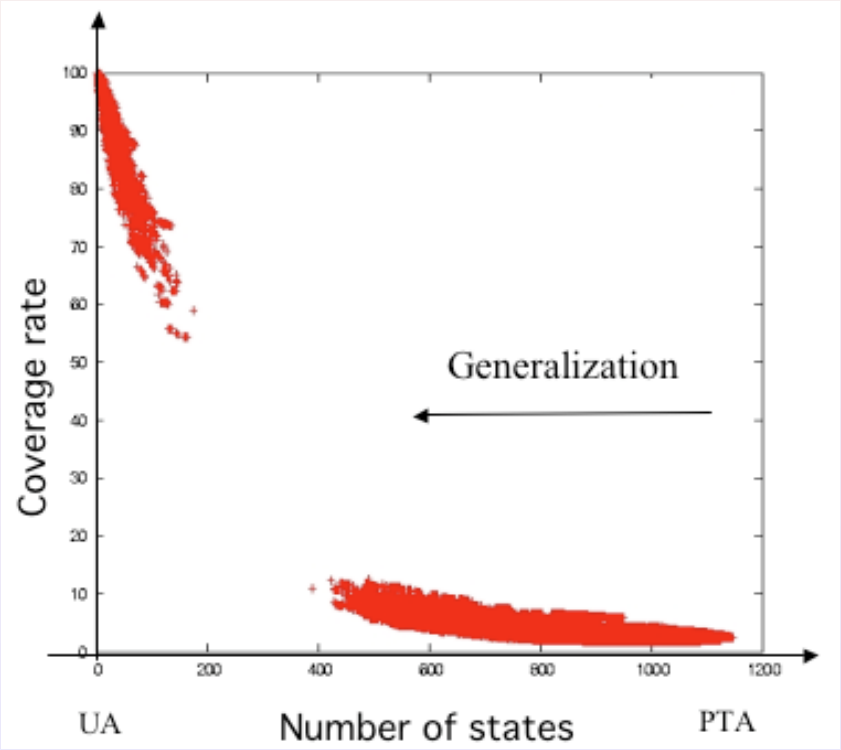
Cas déterministe : **DFA**



Gain d'information en **inférence grammaticale**

Cas déterministe : **DFA**





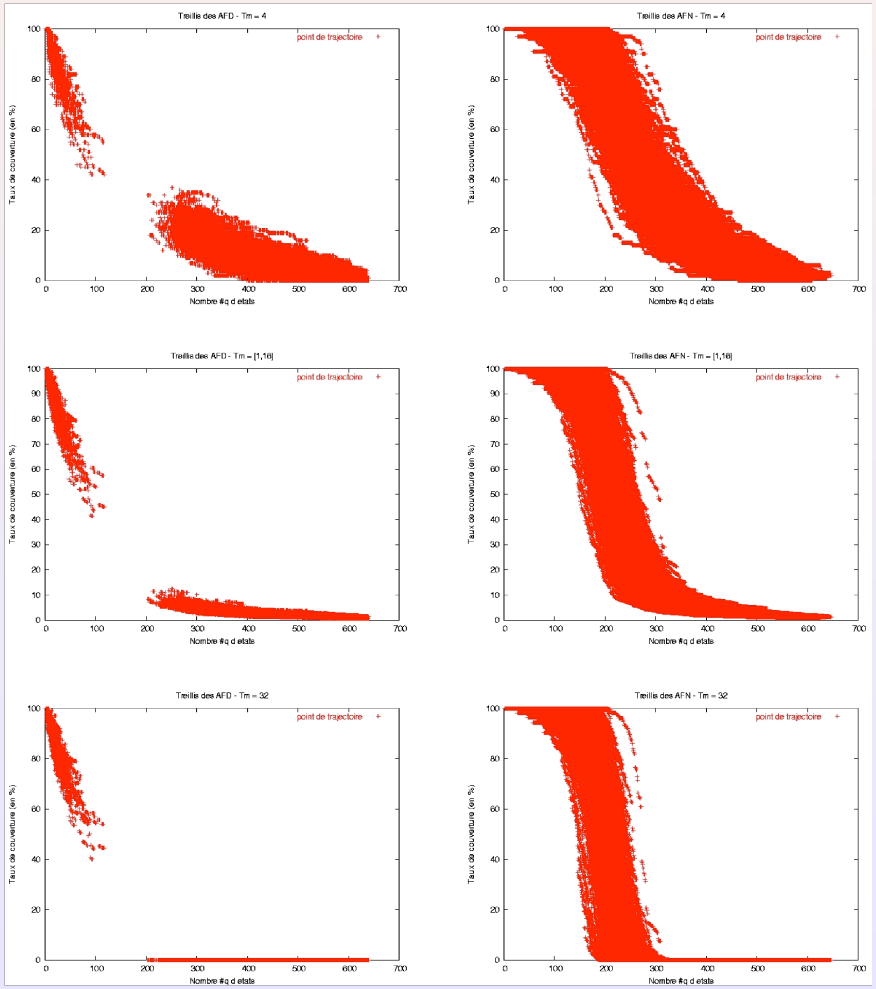
$$|\Sigma| = 8\,;\,\ell = 8$$

Gain d'information en **inférence grammaticale**

Variété de situations

$$
\begin{aligned}
|\Sigma| &= 4 \\
\ell &= 16 \\
|S^+| &= 100
\end{aligned}
$$

Test sur 1000 chaînes
de tailles : 4, 16 et 32

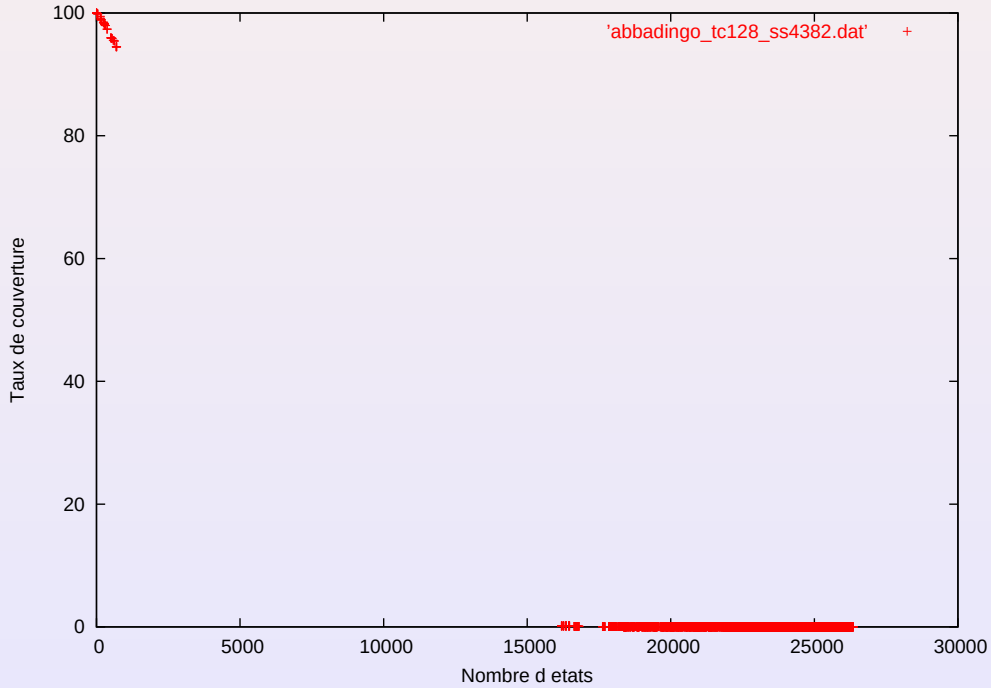


Gain d'information en **inférence grammaticale**

Le défi ***Abbadingo***

$$
\begin{aligned}
|\Sigma| &= 2 \\
\ell &= 17 \\
|S^+| &= 4382
\end{aligned}
$$

Test sur 1000 chaînes
de taille : 17

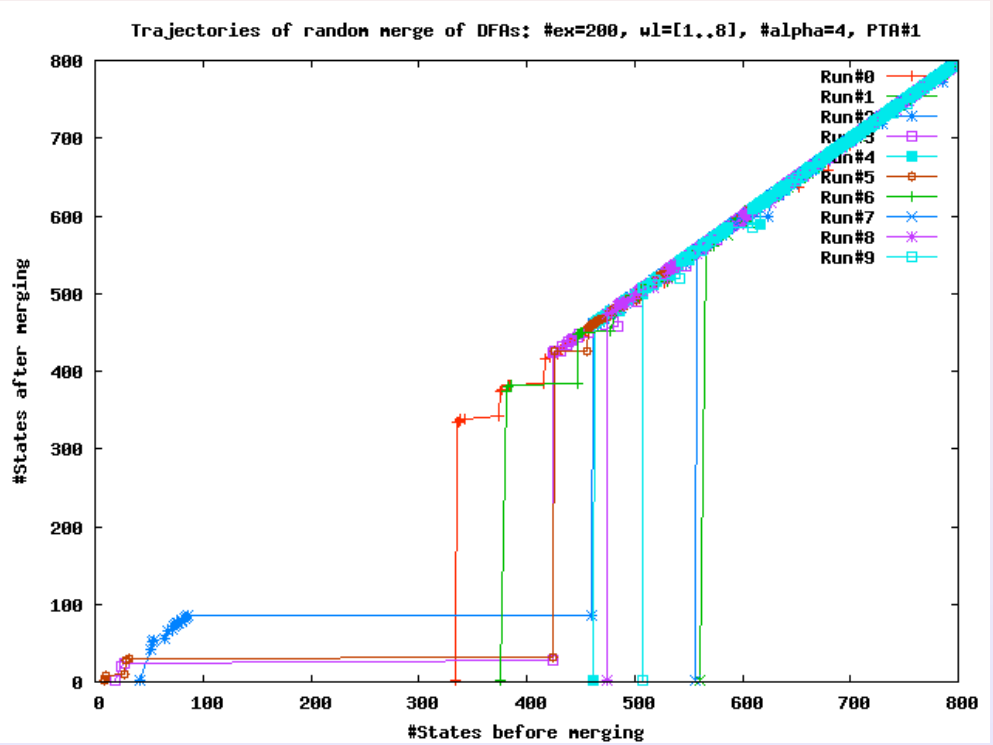


Analyse pour les DFA

Saut du taux de couverture et saut du nombre d'états

$$
\begin{aligned}
|\Sigma| &= 4 \\
\ell &\in [1, .., 8] \\
|S^+| &= 200
\end{aligned}
$$

8 trajectoires aléatoires

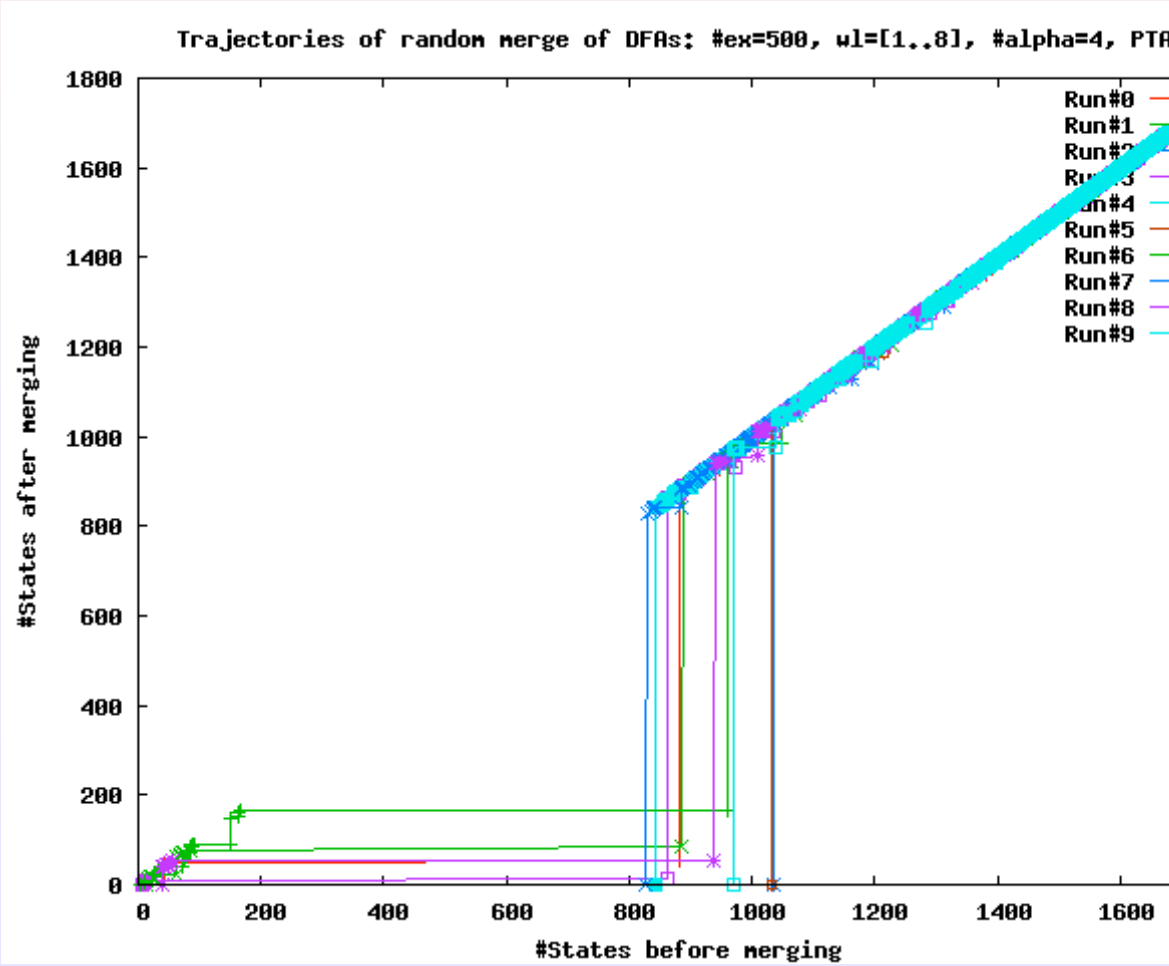


Analyse pour les DFA

Saut du taux de couverture et saut du nombre d'états

$$
\begin{aligned}
|\Sigma| &= 4 \\
\ell &\in [1,..,8] \\
|S^+| &= 500
\end{aligned}
$$

8 trajectoires aléatoires

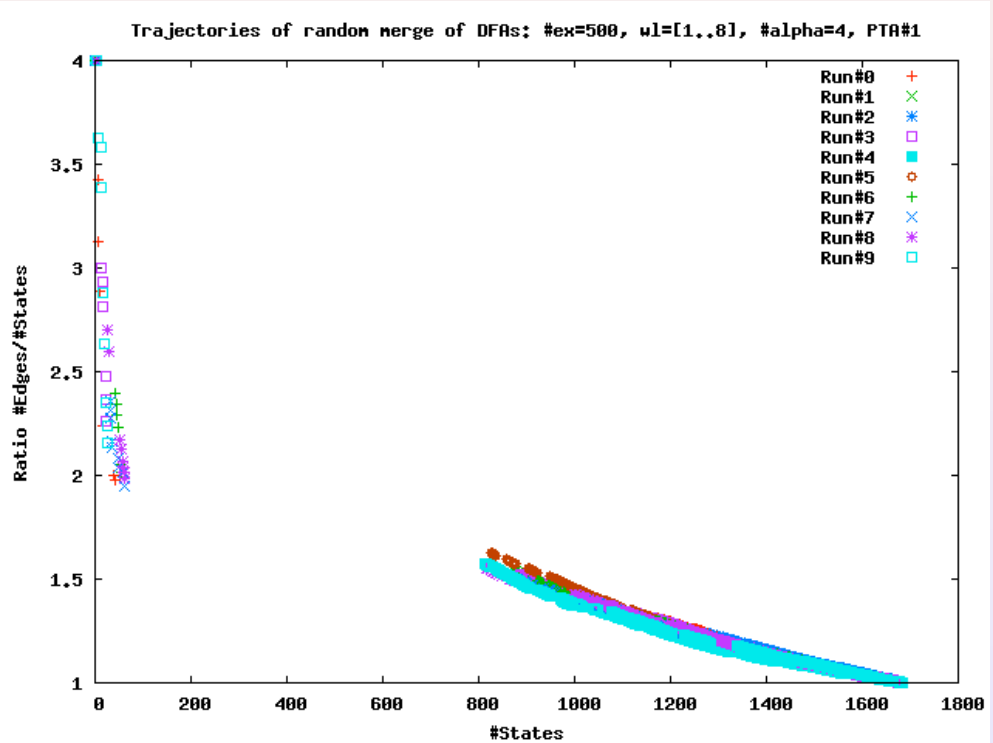


Analyse pour les DFA

Saut du taux de couverture et nombre d'arcs / états

$$
\begin{aligned}
|\Sigma| &= 4 \\
\ell &\in [1, .., 8] \\
|S^+| &= 500
\end{aligned}
$$

8 trajectoires aléatoires

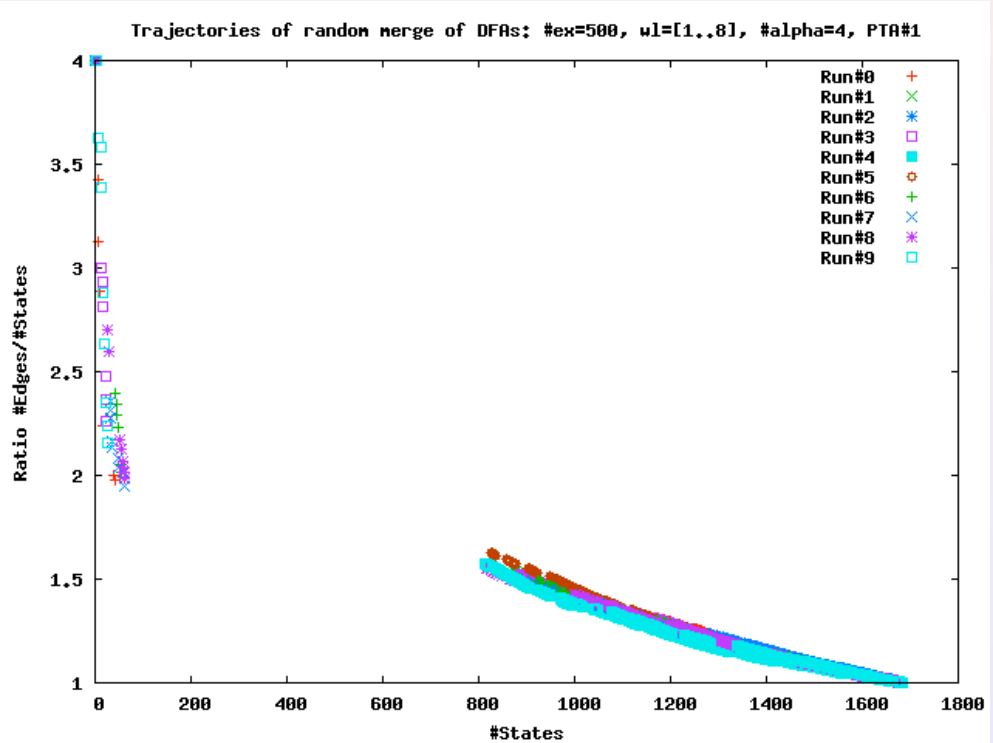


Analyse pour les DFA

Saut du taux de couverture et nombre d'arcs / états

$$
\begin{aligned}
|\Sigma| &= 4 \\
\ell &\in [1, .., 8] \\
|S^+| &= 500
\end{aligned}
$$

8 trajectoires aléatoires



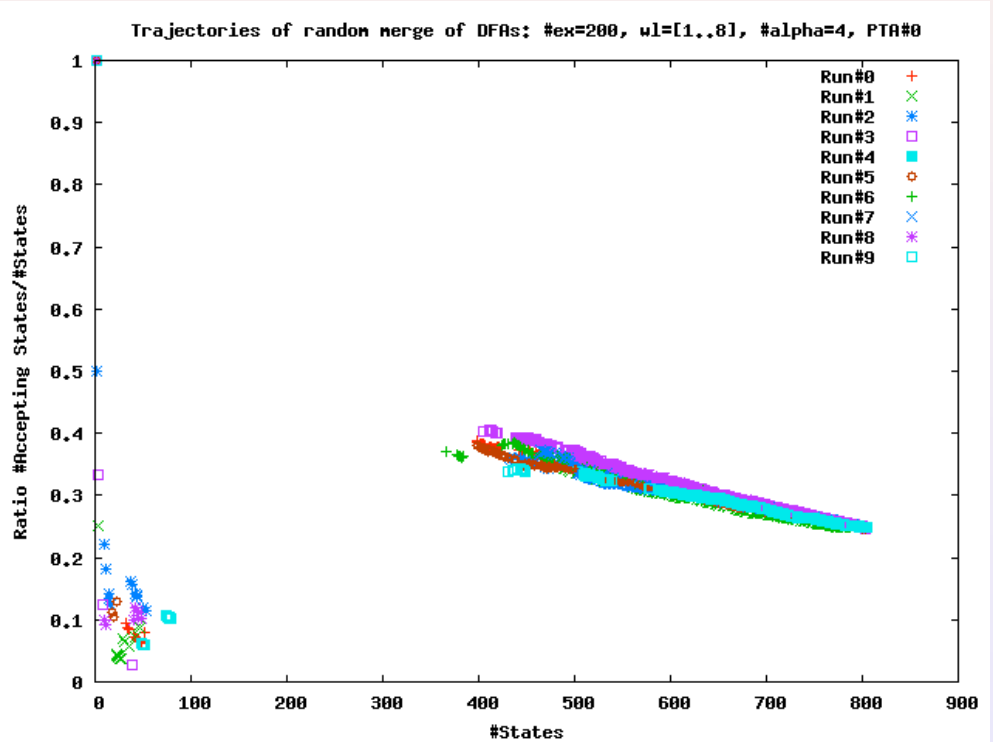
Pas de saut ! !

Analyse pour les DFA

Saut du taux d'états acceptants

$$
\begin{aligned}
|\Sigma| &= 4 \\
\ell &\in [1, .., 8] \\
|S^+| &= 200
\end{aligned}
$$

8 trajectoires aléatoires

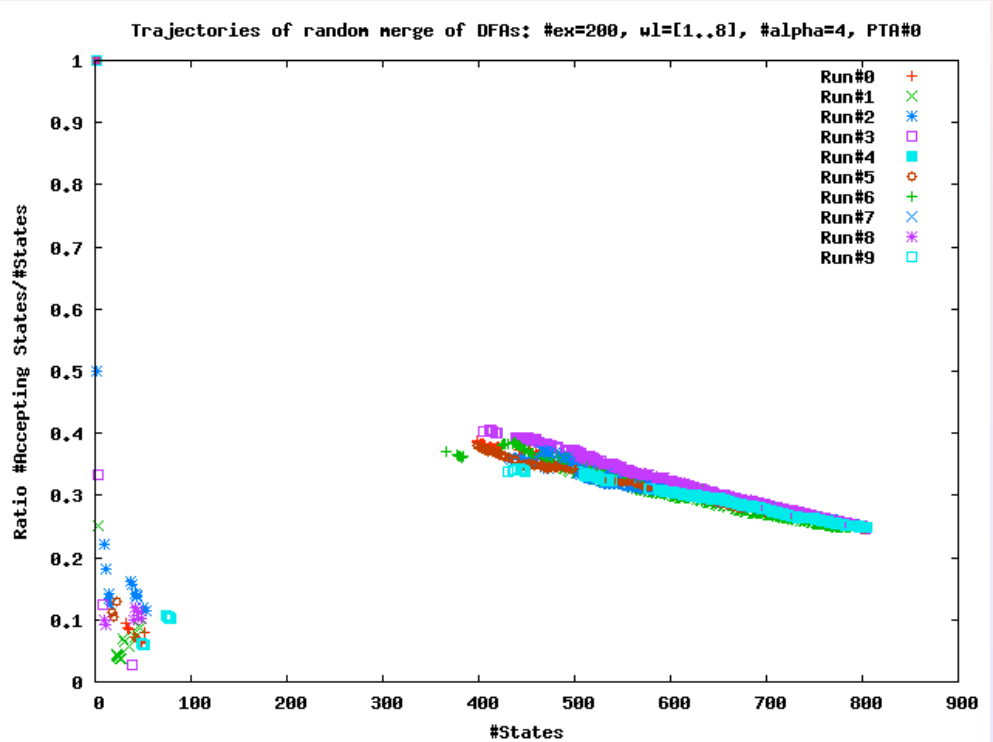


Analyse pour les DFA

Saut du taux d'états acceptants

$$
\begin{aligned}
|\Sigma| &= 4 \\
\ell &\in [1, .., 8] \\
|S^+| &= 200
\end{aligned}
$$

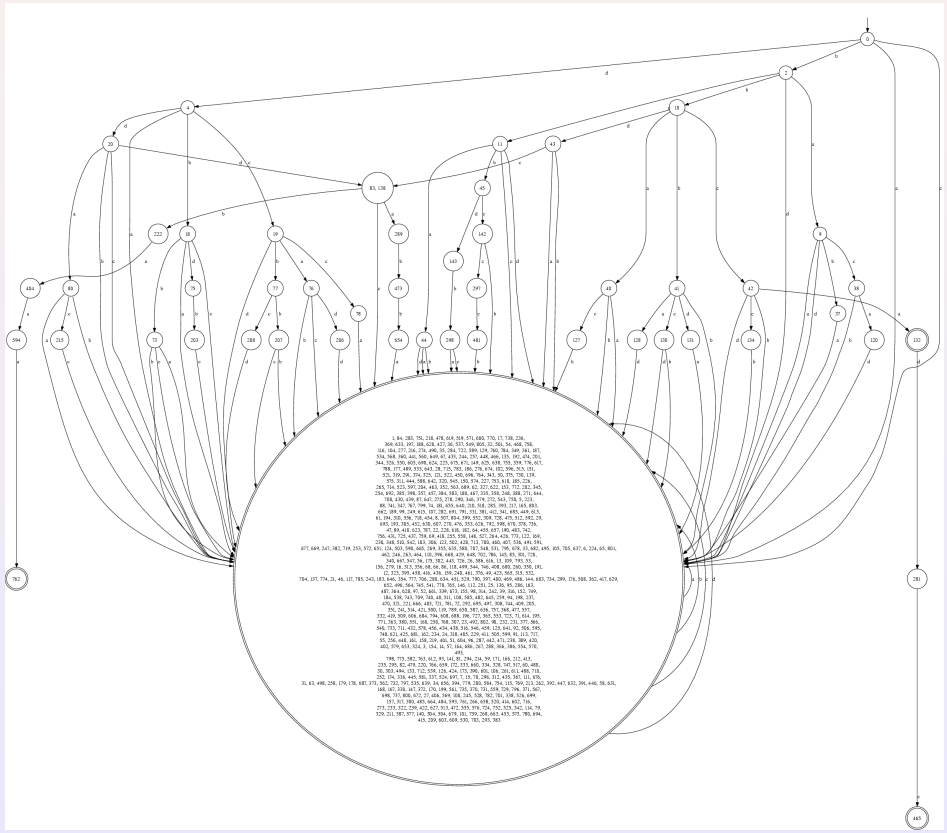
8 trajectoires aléatoires



Saut opposé aux attentes

Analyse pour les DFA

Saut du taux d'états acceptants



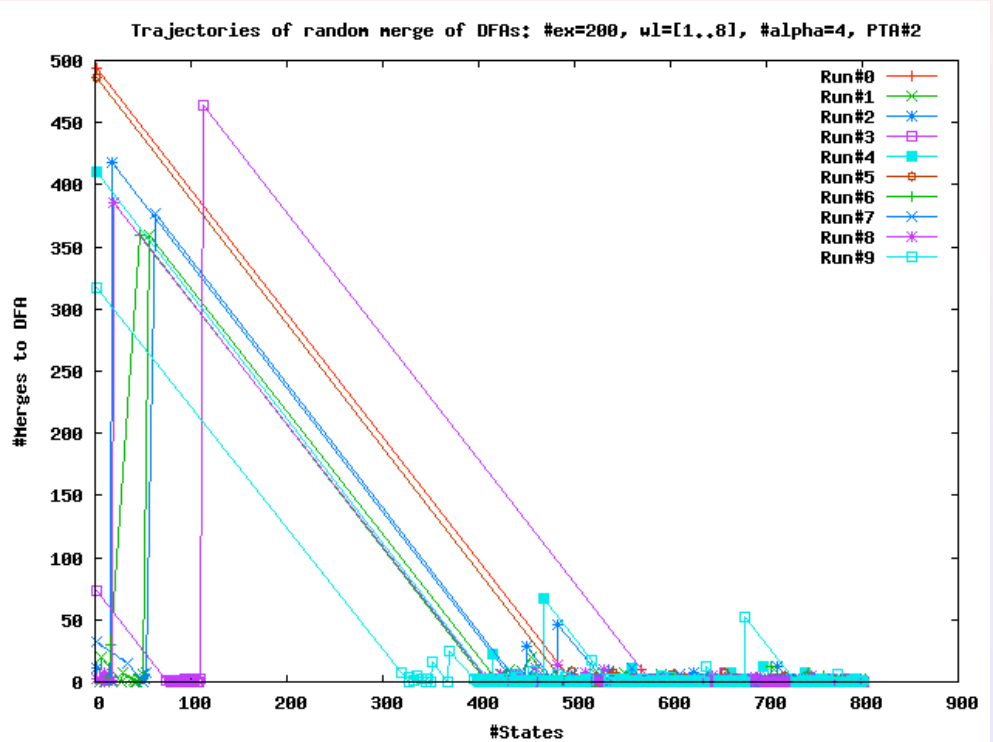
Les états acceptants sont davantage fusionnés ! !

Analyse pour les DFA

L'explication : avalanches de fusions pour déterminisation

$$
\begin{aligned}
|\Sigma| &= 4 \\
\ell &\in [1,..,8] \\
|S^+| &= 200
\end{aligned}
$$

8 trajectoires aléatoires

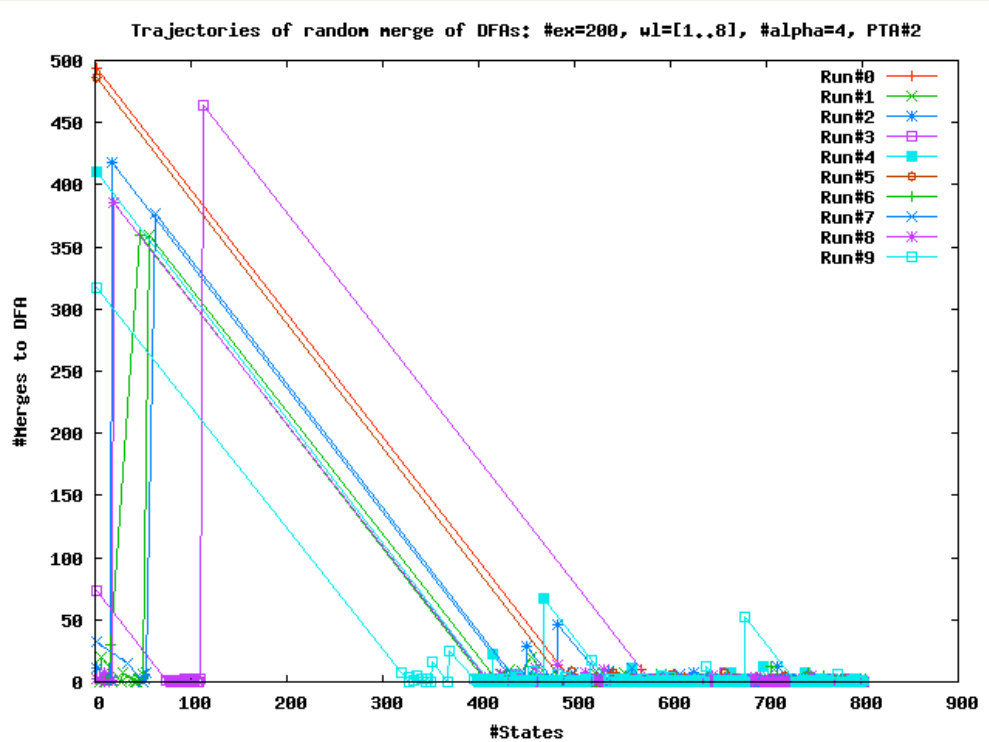


Analyse pour les DFA

L'explication : avalanches de fusions pour déterminisation

$$
\begin{aligned}
|\Sigma| &= 4 \\
\ell &\in [1,..,8] \\
|S^+| &= 200
\end{aligned}
$$

8 trajectoires aléatoires



Phénomène d'avalanche (ou de réactions en chaînes)

Analyse pour les DFA
Remèdes ?

Lutte contre l'avalanche de déterminisations

- *Heuristiques de choix des états à fusionner*

Autres opérateurs de généralisation

- *Fusion des nœuds feuilles*
- *Ajout d'arcs*
- *Ajout d'états acceptants*

Autres types d'apprentissage : choix des exemples

- *Apprentissage actif*
- *Apprentissage guidé*

Analyse pour les DFA

Algorithmes avec heuristiques de recherche

- Par fusion d'états
- Jusqu'à la couverture d'exemples négatifs

RPNI [OG92][Lan92]

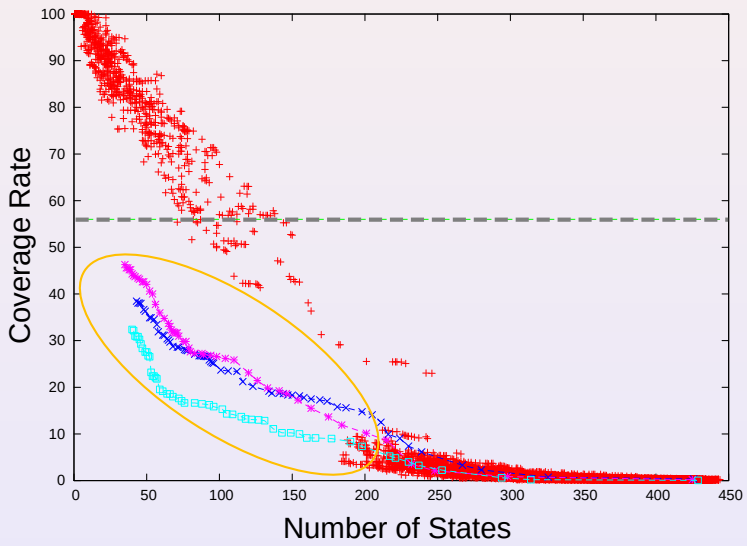
- Choix des états *en largeur d'abord*

EDSM [Lan et al.98][Lan98]

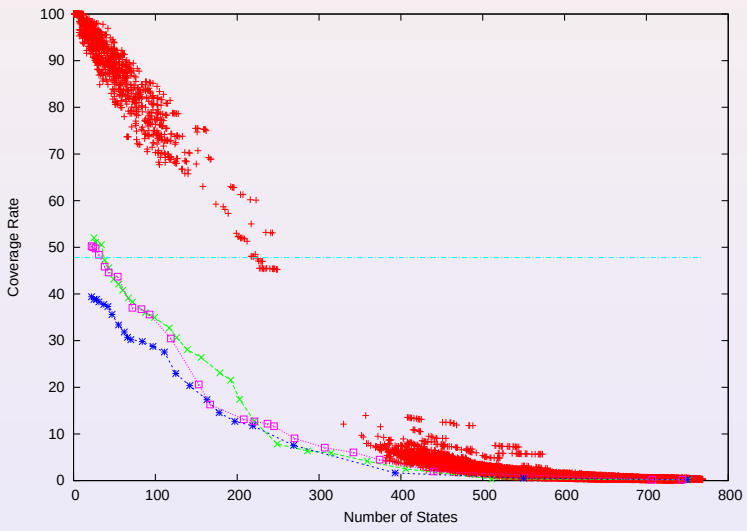
- Choix des états *dont la fusion conduit au maximum de fusions pour déterminisation*

Analyse pour les DFA

Algorithmes avec stratégies de recherche : **résultats**



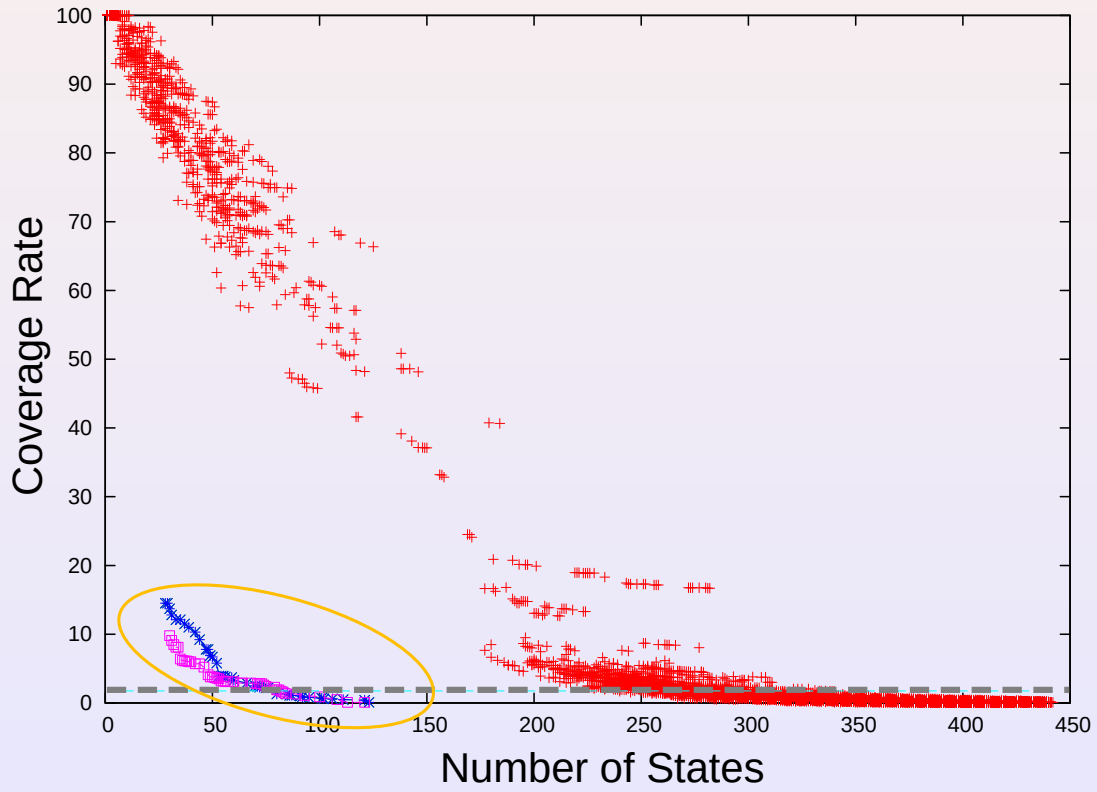
RPNI



EDSM

Analyse pour les DFA

Algorithmes avec stratégies de recherche : **résultats**



Analyse pour les DFA

Algorithmes avec stratégies de recherche : **résultats**

	Target automata		*Learned automata*			
Algo.	Q_c	$ucov_c$	Q_f	$ucov_f$	$\%cov+$	$\%cov-$
RB	15	5.97	10.38	33.81	60.93	34.69
RB	25	4.88	12.77	40.35	62.68	37.87
RB	50	4.2	14.23	45.38	66.14	42.23
RB	100	3.39	13.13	30.35	42.81	28.69
RPNI	15	5.95	5.14	22.9	57.51	26.99
RPNI	25	4.7	7.56	23.07	56.38	25.98
RPNI	50	3.87	14.08	23.45	51.89	24.42
RPNI	100	3.12	26.41	23.151	50.12	24.40

Gain d'information et induction

Conclusions pour l'inférence grammaticale

Perspectives

- Autres ***opérateurs ou heuristiques***
  pour modifier le paysage apparent

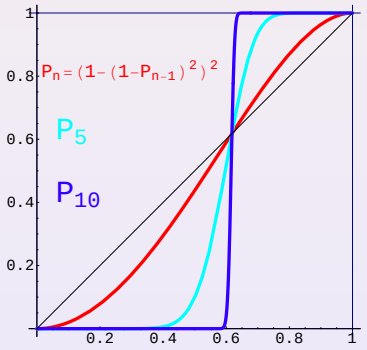
- Possibilités d'***apprentissage guidé ou actif***
  - Modification de l'espace des exemples
    - longueur des chaînes (e.g. les longues avant les courtes !)
    - ignorer des lettres de l'alphabet (?)
  - dynamiquement

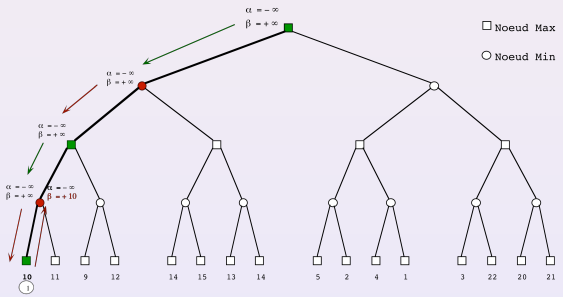
Transition de phase ou pas ?

L'exemple de l'algorithme de Min-Max

Probabilité que la **position racine** soit **gagnante** en fonction de la probabilité qu'une **position feuille** soit **gagnante**

$$P_n = 1 - (1 - P_{n-1}^b)^b$$



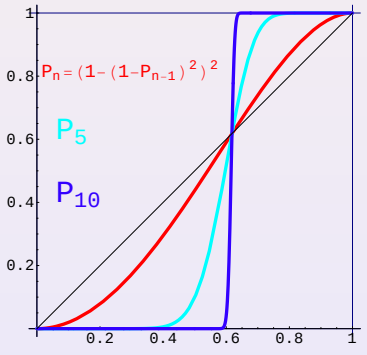


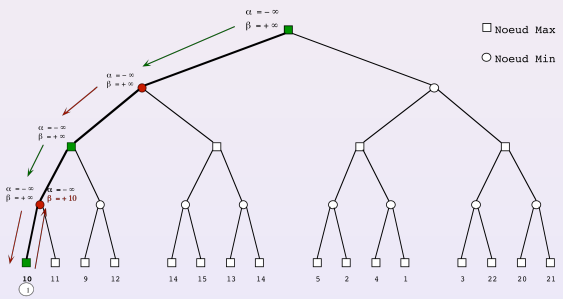
Transition de phase ou pas ?

L'exemple de l'algorithme de Min-Max

Probabilité que la **position racine** soit **gagnante** en fonction de la probabilité qu'une **position feuille** soit **gagnante**

$$P_n = 1 - (1 - P_{n-1}^b)^b$$



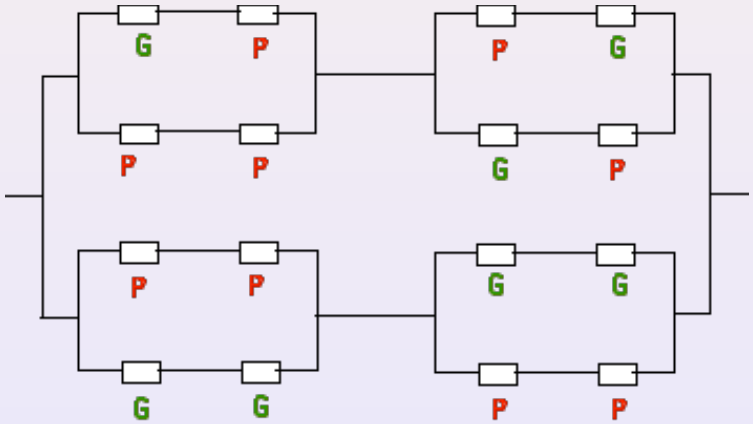


En informatique :

La récurrence peut conduire à un phénomène de TP

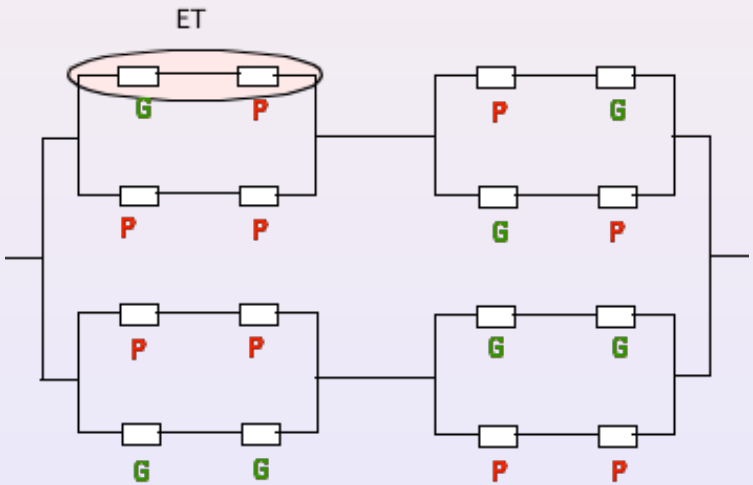
Transition de phase ou pas ?

Le cas de l'inférence grammaticale



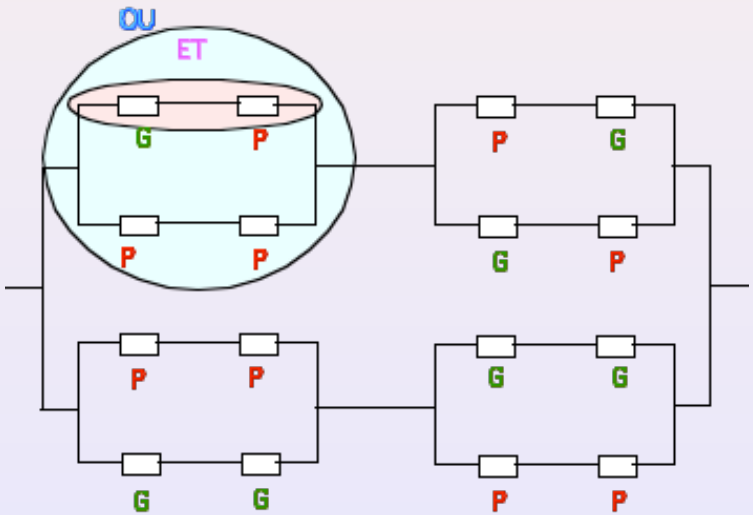
Transition de phase ou pas ?

Le cas de l'inférence grammaticale



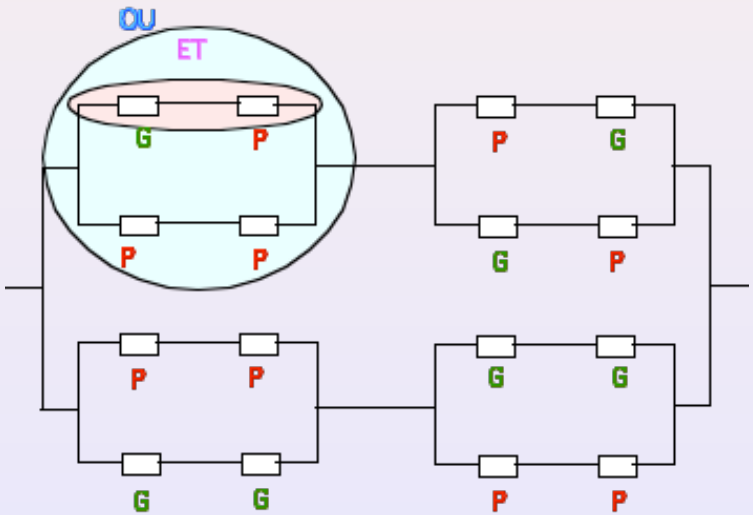
Transition de phase ou pas ?

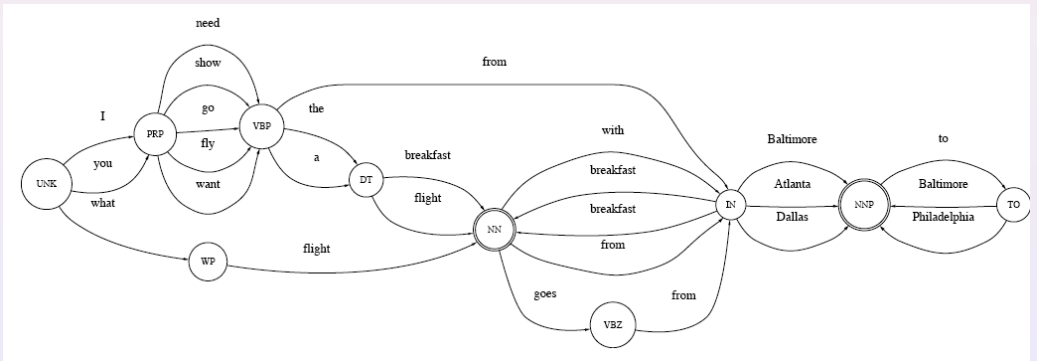
Le cas de l'inférence grammaticale



Transition de phase ou pas ?

Le cas de l'inférence grammaticale

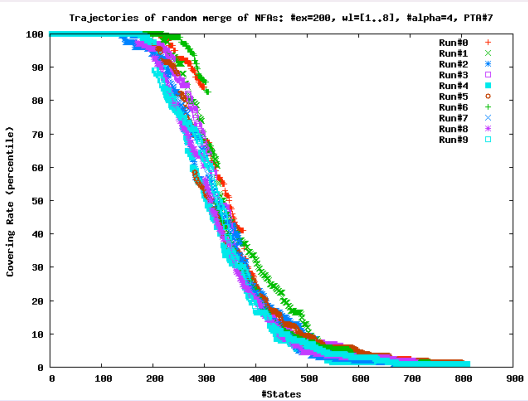




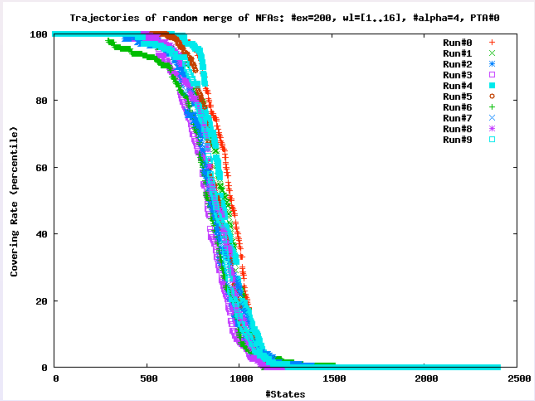
Transition de phase ou pas ?

Retour sur les NFA

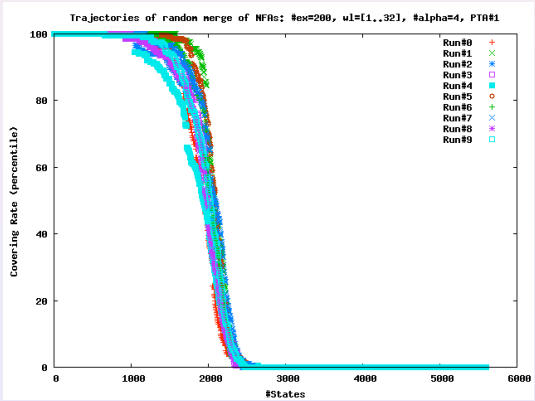
La modélisation en réseau récurrent semble s'appliquer



$l = 8$



$l = 16$



$l = 32$

Transition de phase ou pas ?

Le cas de la programmation logique inductive

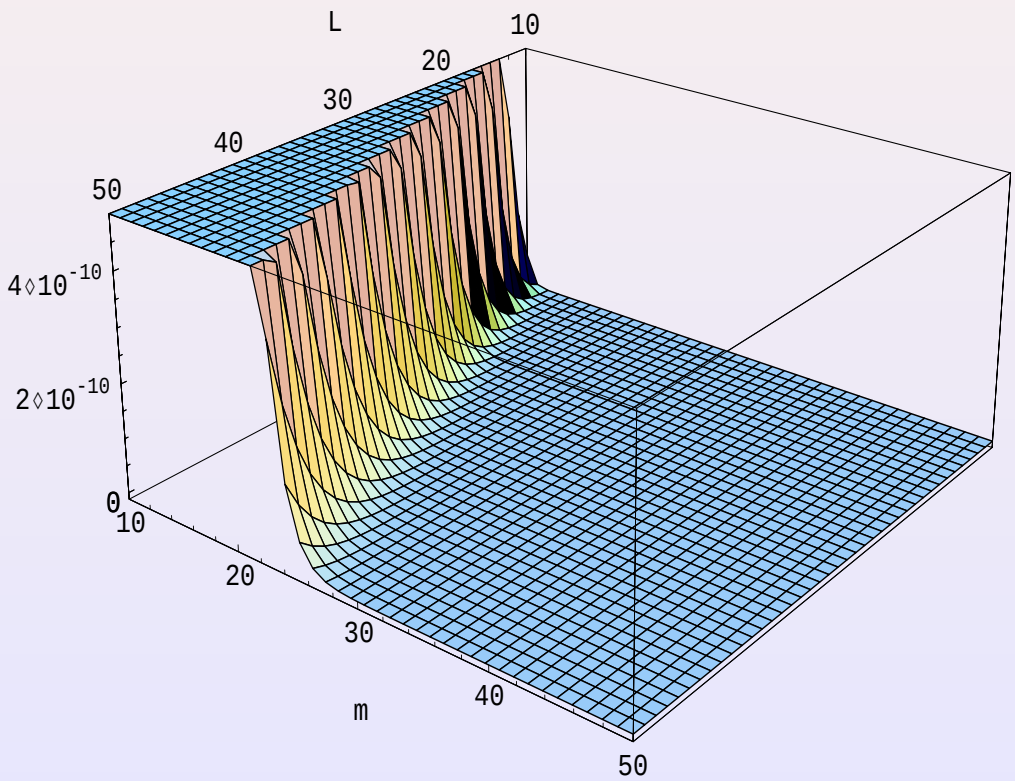
$$h(\boldsymbol{x}) = \bigwedge_{i=1}^{n} p_i(x_i, y_i)$$

Probabilité de couverture :

$$\left(1 - (\textit{une probabilité})^{L}\right)^{m}$$

Rappel :

- ***m*** : nombre de *symboles de prédicats* dans $\underline{h}$
- ***L*** : nombre total de *constantes* dans l'exemple e



Analyse

Alors ... pour quels langages ?

- Programmation logique inductive
- Inférence d'automates à états finis
- ... ?

Importance du concept de récurrence

- structure (e.g. inférence grammaticale
- test de couverture

Encore du domaine de la recherche

Gain d'information et induction

Importance de l'analyse du taux de couverture

Bilan

1. Importance de l'étude des variations du taux de couverture
2. ... par rapport à l'espace de recherche effectif

- Permet d'expliquer des comportements
- Orienter les recherches

[PCS05] N. Pernot, A. Cornuéjols and M. Sebag.
Phase transition within grammatical inference.
Int. Joint Conf. on Artificial Intelligence (IJCAI-05), Edinburgh, UK, 2005, (Ed. L. P. Kaelbling), pp.811-816.

[CS05] A. Cornuéjols and M. Sebag.
A Note on Phase Transitions and Computational Pitfalls of Learning from Sequences.
Second Franco-Japanese Workshop on Information Search, Integration and Personalization (ISIP-05), Lyon, France, 2005. (en soumission à Int. J. of Intelligent Information Systems (JIIS))

Gain d'information et induction

Conclusions générales

Par rapport à l'analyse statistique de l'induction

- **Étude plus fine de l'induction**
- prenant en compte le **gain d'information** ($\frac{\partial \tau_S(h)}{\partial h}$)
- et l'**espace effectivement exploré**

Gain d'information et induction

Conclusions générales

Par rapport à l'analyse statistique de l'induction

- **Étude plus fine de l'induction**
- prenant en compte le **gain d'information** $(\frac{\partial \tau_S(h)}{\partial h})$
- et l'**espace effectivement exploré**

Mais reste dans le cadre i.i.d. ! !

Plan

Apprentissage en-ligne
Les différentes approches

- Apprentissage incrémental : approches heuristiques
- Prédiction
  - Prédiction universelle
  - Suivi de porte-feuille
  - Dérive de concept
- Modèles d'apprentissage en-ligne
  - Identification à la limite
  - Apprentissage par requêtes
  - Dérive de concept
- Approche de la physique statistique
- Teachability
- Apprentissage actif
- Heuristiques d'apprentissage guidé
  - Learning one sub-procedure per lesson
  - Apprentissage hiérarchique
  - Apprentissage à partir d'exemples simples
  - Mistake-bound learning
  - Learning from expert advice
- Théorie du contrôle
- Apprentissage par renforcement
- Théorie des jeux itérés
- ...

Apprentissage en-ligne
Définition

État mis à jour après observation de chaque exemple de la séquence

Éventuellement des **contraintes** sur la capacité mémoire

Éventuellement des **contraintes** sur la capacité calcul

Apprentissage en-ligne

Les deux approches en théorie de l'apprentissage en-ligne

Même objectif : minimiser l'espérance de coût

- Comment régler la fonction de coût instantanée (gradient stochastique)
- Comment converger le plus vite possible vers la même performance que apprentissage hors-ligne (notion de ***regret***)

Autre objectif : comparaison à ensemble d'experts

Apprentissage de type classique

- Mais fonction de coût en nombre d'erreurs de prédiction (*mistake-bound learning*)

$\longrightarrow$ Nouveaux algorithmes (e.g. Winnow)
$\longrightarrow$ Nouvelles bornes sur les convergence

Apprentissage à partir de conseil d'experts

$\longrightarrow$ Nouveaux algorithmes

Apprentissage en-ligne

Comparaison au cas hors-ligne

Pas de véritable étude du cas non i.i.d.

Le cadre i.i.d.

Definition (Le cadre i.i.d.)

Pour faire une prédiction sur la prochaine entrée, il suffit de connaître la distribution génératrice sous-jacente.

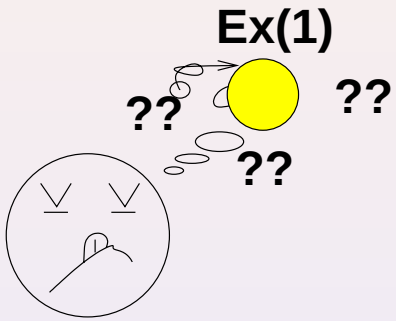
Corollaire

La connaissance d'autres entrées n'apporte aucune information supplémentaire.

$\Longrightarrow$ ***Efface la notion d'histoire***

Les effets de séquences

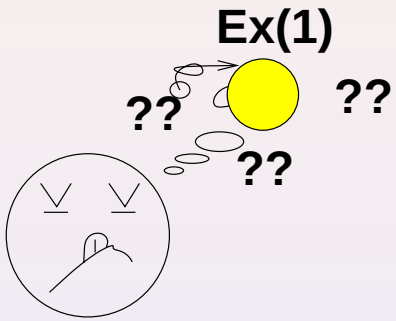
1er exemple (du à Laurent Chaudron [HDR,2005])



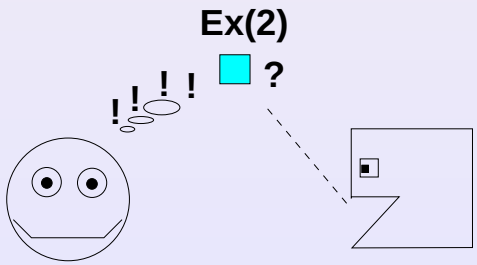
Exercice(1) *Soit* $\left(\frac{p_n}{q_n}\right)_{n\in N}$ *une séquence de nombres rationnels convergeant vers x irrationnel. Prouver que* (p_n) *et* (q_n) *convergent tous les deux vers l'infini.*

Les effets de séquences

1er exemple (du à Laurent Chaudron [HDR,2005])



Exercice(1) *Soit* $\left(\frac{p_n}{q_n}\right)_{n\in N}$ *une séquence de nombres rationnels convergeant vers x irrationnel. Prouver que* (p_n) *et* (q_n) *convergent tous les deux vers l'infini.*



Exercice(2) *Prouver que l'image de n'importe quelle séquence finie de nombres naturels est un ensemble fini.*

Les effets de séquences

2ème exemple

[Sur 24 étudiants de DEA, 1996]

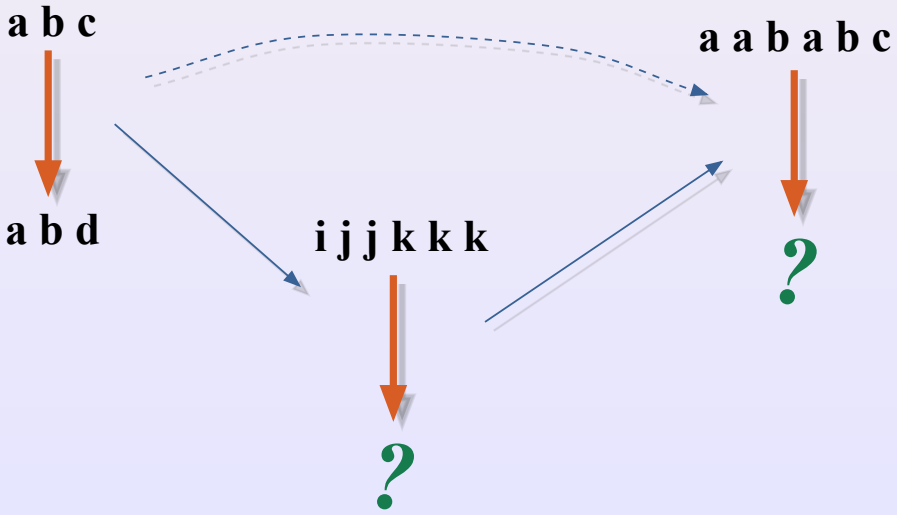
- Long et difficile
- Grande variété de réponses

Les effets de séquences

2ème exemple

[Sur 24 étudiants de DEA, 1996]

- Long et difficile
- Grande variété de réponses



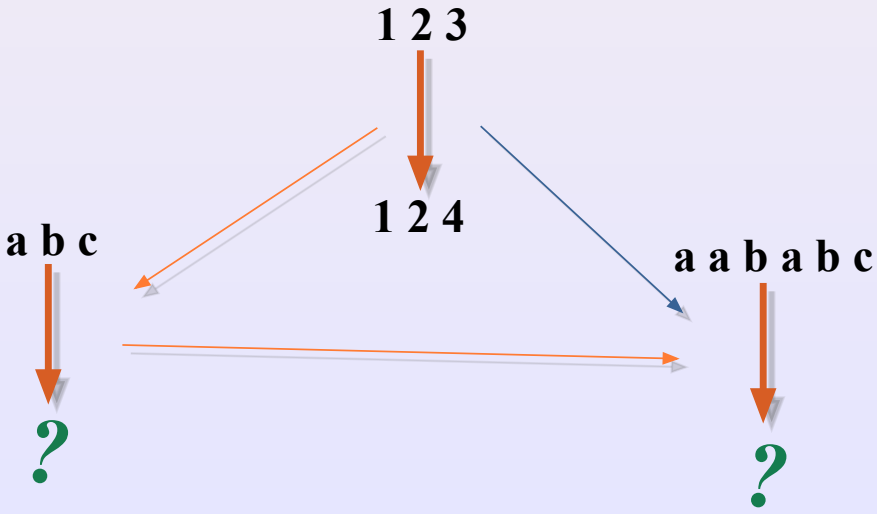
- Beaucoup plus rapide
- Spectre de réponses beaucoup plus serré

Les effets de séquences

2ème exemple

[Sur 24 étudiants de DEA, 1996]

- Long et difficile
- Grande variété de réponses



- Chemin rouge : plus difficile et réponses plus confuses

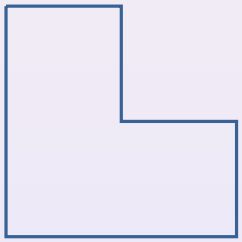
Les effets de séquences

3ème exemple

[Sur quelques étudiants de Polytechnique, 1994]

Consigne : découper la figure suivante en n parties superposables.

En **2** :



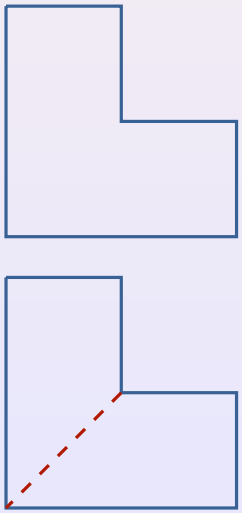
Les effets de séquences

3ème exemple

[Sur quelques étudiants de Polytechnique, 1994]

Consigne : découper la figure suivante en n parties superposables.

En **2** :



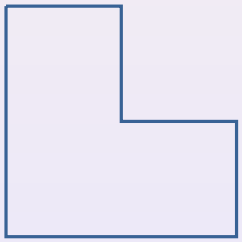
Les effets de séquences

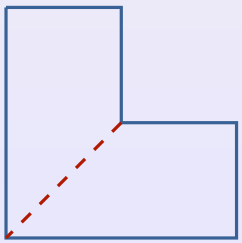
3ème exemple

[Sur quelques étudiants de Polytechnique, 1994]

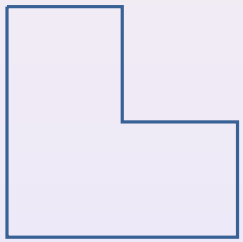
Consigne : découper la figure suivante en n parties superposables.

En **2** :





En **3** :



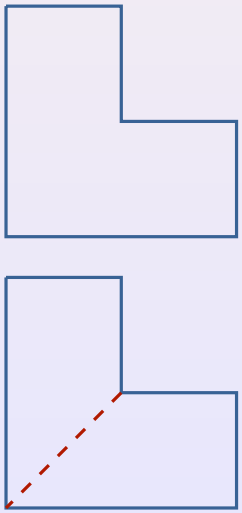
Les effets de séquences

3ème exemple

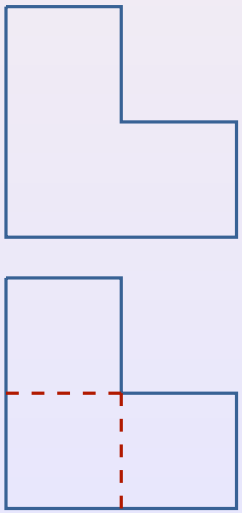
[Sur quelques étudiants de Polytechnique, 1994]

Consigne : découper la figure suivante en n parties superposables.

En **2** :



En **3** :



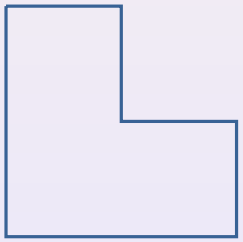
Les effets de séquences

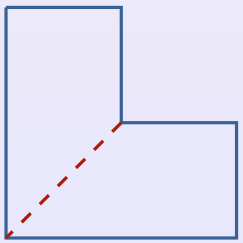
3ème exemple

[Sur quelques étudiants de Polytechnique, 1994]

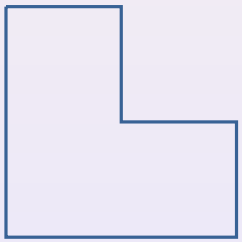
Consigne : découper la figure suivante en n parties superposables.

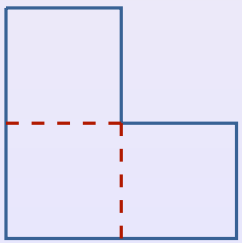
En **2** :



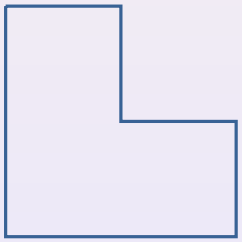


En **3** :





En **4** :



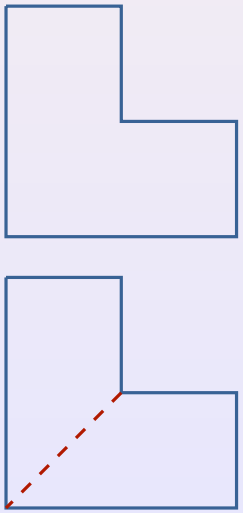
Les effets de séquences

3ème exemple

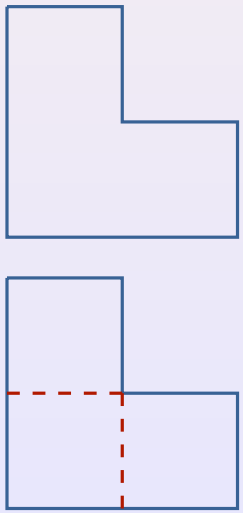
[Sur quelques étudiants de Polytechnique, 1994]

Consigne : découper la figure suivante en n parties superposables.

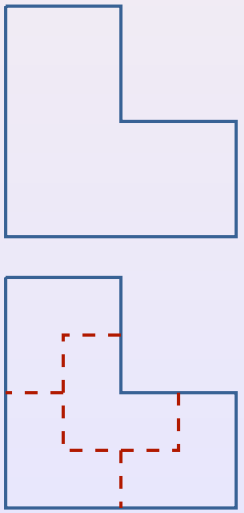
En **2** :



En **3** :



En **4** :



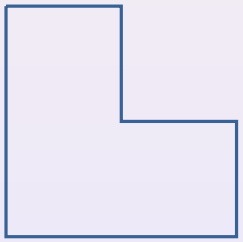
Les effets de séquences

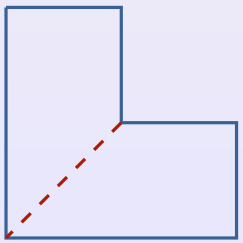
3ème exemple

[Sur quelques étudiants de Polytechnique, 1994]

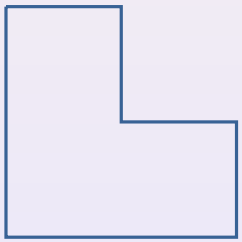
Consigne : découper la figure suivante en n parties superposables.

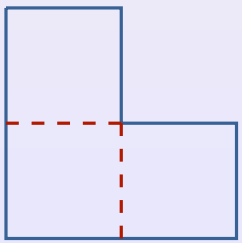
En **2** :



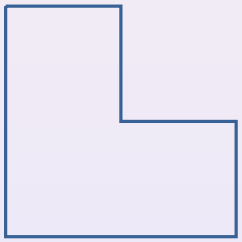


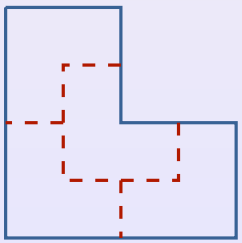
En **3** :



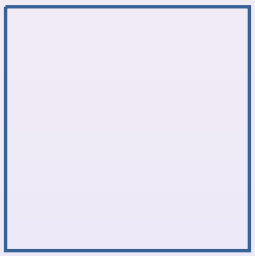


En **4** :





En **5** :



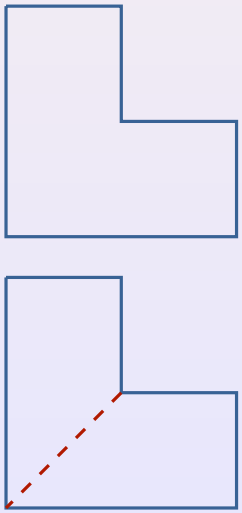
Les effets de séquences

3ème exemple

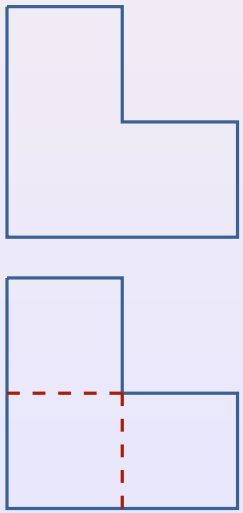
[Sur quelques étudiants de Polytechnique, 1994]

Consigne : découper la figure suivante en n parties superposables.

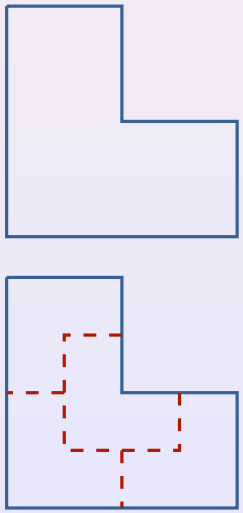
En **2** :



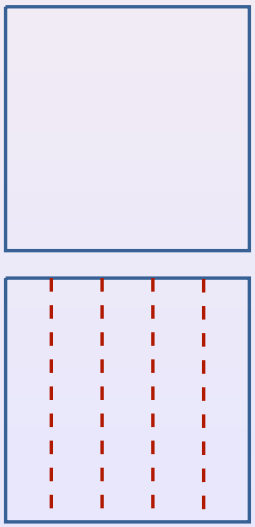
En **3** :



En **4** :



En **5** :



Effets de séquences : constatations

Dépendance sur l'histoire

- De la vitesse de résolution
- Du résultat

Effets de séquences : constatations

Dépendance sur l'histoire

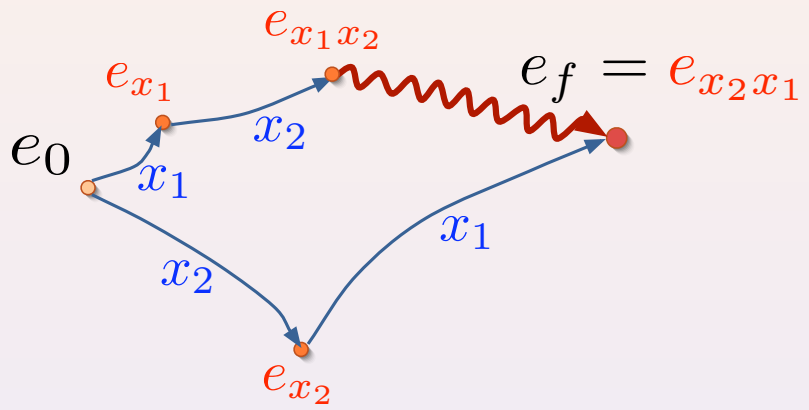
- De la vitesse de résolution
- Du résultat

Courbure de l'espace des états

$\longrightarrow$ ce n'est plus un espace euclidien

Mesure de courbure

Le «**crochet de Lie** »
$[x_1, x_2]_{\texttt{état}}$



Remarque

Le crochet de Lie (la courbure) s'annule si :

1. Ressources de ***calcul*** suffisantes
2. Ressources ***mémoire*** suffisantes

Hors-ligne vs. ***en-ligne***

Pour sortir du cadre i.i.d. ...

Il faut **aborder de front** ***les effets de séquences***

1. Quelles sont les **entrées utiles** (les plus utiles) ?
2. Quelles sont les entrées nuisibles ?

Nouvelles questions

1. ***Ordres de présentation les plus favorables*** ?
2. Quels sont les ***systèmes sensibles aux effets de séquence*** ?

Pour sortir du cadre i.i.d. ...

... Il faut de nouveaux outils

Outils nécessaires

1. Une ***métrique*** (entre programmes)
   - pour mesurer la *distance entre états*
   - pour mesurer la *corrélation entre des entrées*
2. Une ***mesure de courbure*** de l'espace

Pour sortir du cadre i.i.d. ...

... Il faut de nouveaux outils

Outils nécessaires

1. Une ***métrique*** (entre programmes)
   - pour mesurer la *distance entre états*
   - pour mesurer la *corrélation entre des entrées*
2. Une ***mesure de courbure*** de l'espace

Outils existants

- Entropie relative / information mutuelle / complexité algorithmique

Pour sortir du cadre i.i.d. ...

... Il faut de nouveaux outils

Outils nécessaires

1. Une ***métrique*** (entre programmes)
   - pour mesurer la *distance entre états*
   - pour mesurer la *corrélation entre des entrées*
2. Une ***mesure de courbure*** de l'espace

Outils existants

- Entropie relative / information mutuelle / complexité algorithmique

Limites

1. Ne permettent pas de rendre compte de corrélations négatives
2. Inadaptés à espaces courbes

Une nouvelle mesure de corrélation entre programmes

qui peut prendre une valeur négative

Une nouvelle mesure de corrélation entre programmes

qui peut prendre une valeur négative

Le *problème* : ***évaluation (tri) d'attributs***
(analyse du transcriptome)

Une nouvelle mesure de corrélation entre programmes
qui peut prendre une valeur négative

Le *problème* : ***évaluation (tri) d'attributs***
(analyse du transcriptome)

Inférence très précaire

1. Beaucoup plus d'attributs que de dimensions
2. Nombreuses sources de «bruit »

Une nouvelle mesure de corrélation entre programmes
qui peut prendre une valeur négative

Le *problème* : ***évaluation (tri) d'attributs***
(analyse du transcriptome)

Inférence très précaire

1. Beaucoup plus d'attributs que de dimensions
2. Nombreuses sources de «bruit »

Comment évaluer le résultat ?

Une nouvelle mesure de corrélation entre programmes
qui peut prendre une valeur négative

Le *problème* : ***évaluation (tri) d'attributs***
(analyse du transcriptome)

Inférence très précaire

1. Beaucoup plus d'attributs que de dimensions
2. Nombreuses sources de «bruit »

Comment évaluer le résultat ?

Deux méthodes (non supervisées)

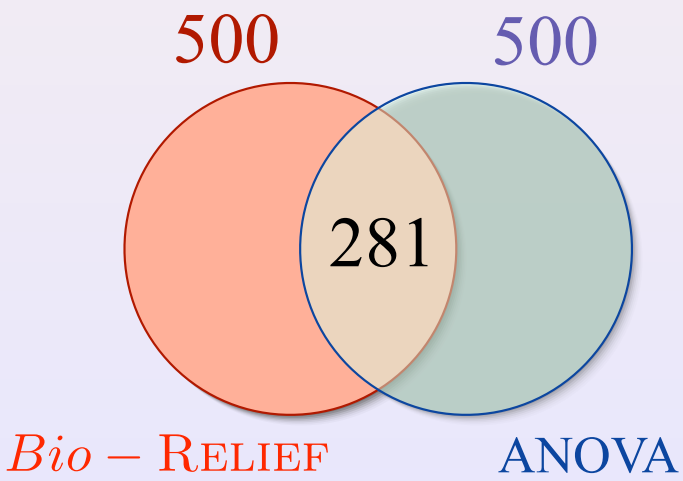
valent-elles mieux qu'une ?

Une nouvelle mesure de corrélation entre programmes

qui peut prendre une valeur négative

Illustration

1. 6135 gènes ; 18 exemples (6+, 12-)
2. Deux méthodes d'évaluation : ANOVA et *Bio*-RELIEF



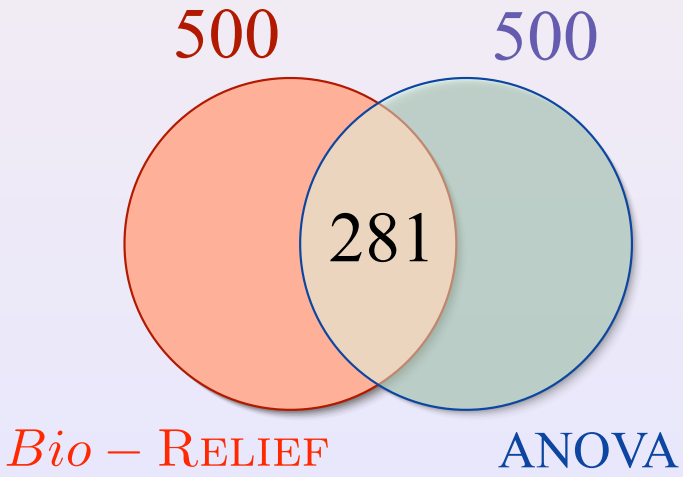
Comment juger ces 281 gènes en commun ?

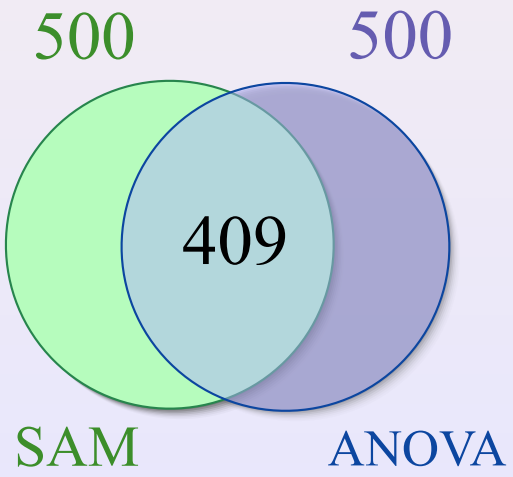
Une nouvelle mesure de corrélation entre programmes

qui peut prendre une valeur négative

Illustration

1. 6135 gènes ; 18 exemples (6+, 12-)
2. Deux méthodes d'évaluation : ANOVA et *Bio*-RELIEF





Comment juger ces 281 gènes en commun ?

Une nouvelle mesure de corrélation entre programmes
qui peut prendre une valeur négative

L'intersection est due :

1. au hasard (k) : $H(d,n,k) = \frac{\binom{n}{k}\cdot\binom{d-n}{n-k}}{\binom{d}{n}}$
2. à la **corrélation des méthodes *a priori***
3. aux régularités dans les données

Sorte d'hypothèse nulle :

$$\langle M1, M_2 \rangle = \mathrm{E}_{\mathcal{D}}(|\mathtt{top}_n(M_1) \cap \mathtt{top}_n(M_2)|)$$

Une nouvelle mesure de corrélation entre programmes

qui peut prendre une valeur négative

$$\langle M1, M_2 \rangle = \mathrm{E}_{\mathcal{D}}(|\mathtt{top}_n(M_1) \cap \mathtt{top}_n(M_2)|)$$

0 : ?

40 : ?

281 : ?

500 : ?

Une nouvelle mesure de corrélation entre programmes

qui peut prendre une valeur négative

$$\langle M1, M_2 \rangle = \mathrm{E}_{\mathcal{D}}(|\mathtt{top}_n(M_1) \cap \mathtt{top}_n(M_2)|)$$

0 : ? Anticorrélés

40 : ?

281 : ?

500 : ?

Une nouvelle mesure de corrélation entre programmes

qui peut prendre une valeur négative

$$\langle M1, M_2 \rangle = \mathrm{E}_{\mathcal{D}}(|\mathtt{top}_n(M_1) \cap \mathtt{top}_n(M_2)|)$$

0	: ?	Anticorrélés
40	: ?	Décorrélés
281	: ?	
500	: ?	

Une nouvelle mesure de corrélation entre programmes

qui peut prendre une valeur négative

$$\langle M1, M_2 \rangle = \mathrm{E}_{\mathcal{D}}(|\mathtt{top}_n(M_1) \cap \mathtt{top}_n(M_2)|)$$

0	: ?	Anticorrélés
40	: ?	Décorrélés
281	: ?	Pas de sur-représentation des régularités
500	: ?	

Une nouvelle mesure de corrélation entre programmes

qui peut prendre une valeur négative

$$\langle M1, M_2 \rangle = \mathrm{E}_{\mathcal{D}}(|\mathtt{top}_n(M_1) \cap \mathtt{top}_n(M_2)|)$$

0	:?	Anticorrélés
40	:?	Décorrélés
281	:?	Pas de sur-représentation des régularités
500	:?	Méthodes totalement corrélées

Une nouvelle mesure de corrélation entre programmes

qui peut prendre une valeur négative

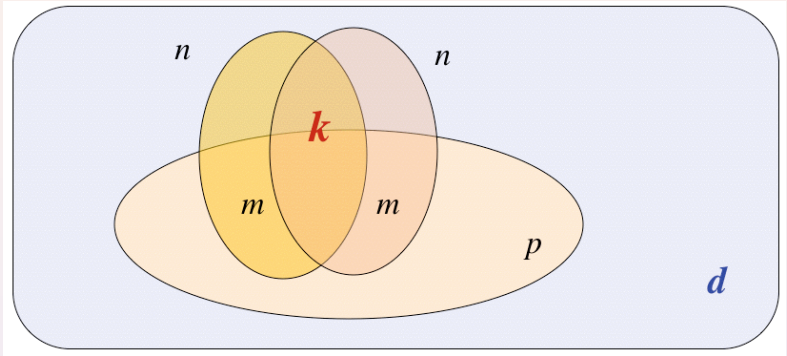
$$\langle M1, M_2 \rangle = \mathrm{E}_{\mathcal{D}}(|\mathtt{top}_n(M_1) \cap \mathtt{top}_n(M_2)|)$$

0	: ?	Anticorrélés
40	: ?	Décorrélés
281	: ?	Pas de sur-représentation des régularités
500	: ?	Méthodes totalement corrélées

Ici : 170 ± 40

Une nouvelle mesure de corrélation entre programmes

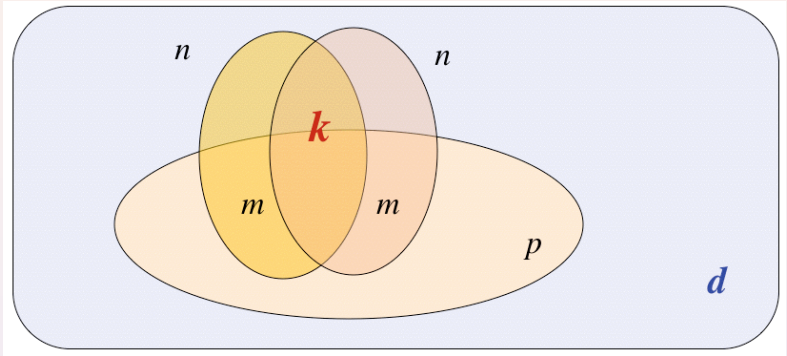
Application



$$p(\cap = k | d, p, n, m, \mu_{\mathcal{H}_0}) = \frac{\binom{p}{m}\binom{d-p}{n-m}\sum_{k^+=2m-p}^{m}\binom{m}{k^+}\binom{p-m}{m-k^+}\binom{n-m}{k-k^+}\binom{d-n-(p-m)}{n-m-(k-k^+)}}{\binom{d}{n}\cdot\binom{d}{n}} \Big/ c(\mu_{\mathcal{H}_0})$$

Une nouvelle mesure de corrélation entre programmes

Application



$$p(\cap = k | d, p, n, m, \mu_{\mathcal{H}_0}) = \frac{\binom{p}{m}\binom{d-p}{n-m}\sum_{k^+=2m-p}^{m}\binom{m}{k^+}\binom{p-m}{m-k^+}\binom{n-m}{k-k^+}\binom{d-n-(p-m)}{n-m-(k-k^+)}}{\binom{d}{n}\cdot\binom{d}{n}} \Big/ C(\mu_{\mathcal{H}_0})$$

Résultats

- $p = 420 \pm 20$
- $m = 340 \pm 20$
- ≈ 265 des 281 sont pertinents ! (*précision* = 0.94)

Une nouvelle mesure de corrélation entre programmes

Portée

- Permet de rendre compte de corrélations négatives
- Applicable aussi à des algorithmes d'apprentissage supervisé

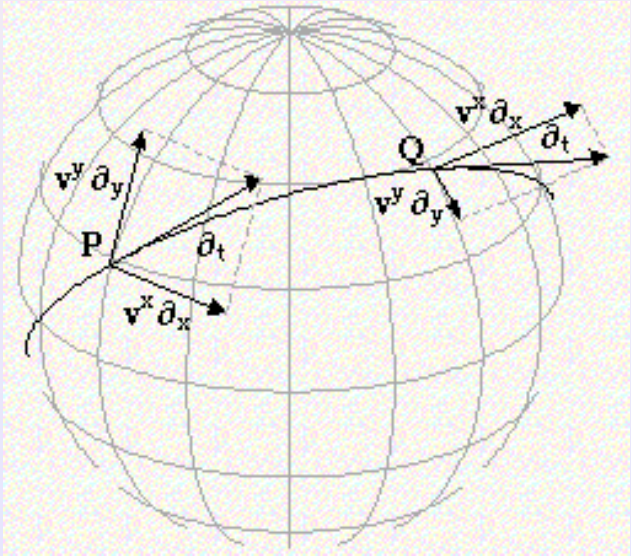
[CFM05] A. Cornuéjols, Ch. Froidevaux and J. Mary.
Comparing and combining feature estimation methods for the analysis of microarray data.
JOBIM-05 : Journées Ouvertes Biologie Informatique Mathématiques (poster), Lyon, France, 2005.

Changement de référentiel

Espace courbe

$\longrightarrow$ Pour comparer deux états du système en deux situations-problèmes différents

$\longrightarrow$ Notion de ***transport parallèle***



Propriétés des trajectoires d'apprentissage

Caractéristiques d'une trajectoire d'apprentissage fonction de :

- Propriétés de l'apprenant
- Caractéristiques de la séquence d'entrées

Quelle séquence idéale pour passer d'un état à un autre ?

Apprentissage et dynamique des systèmes

Caractérisation de l'évolution d'un système

Équation d'évolution locale

$$\frac{d^2x}{dt^2} = \frac{f}{m} \tag{1}$$

Caractérisation globale : formulation lagrangienne

"Principe de moindre action"

Apprentissage et dynamique des systèmes

Caractérisation de l'évolution d'un système

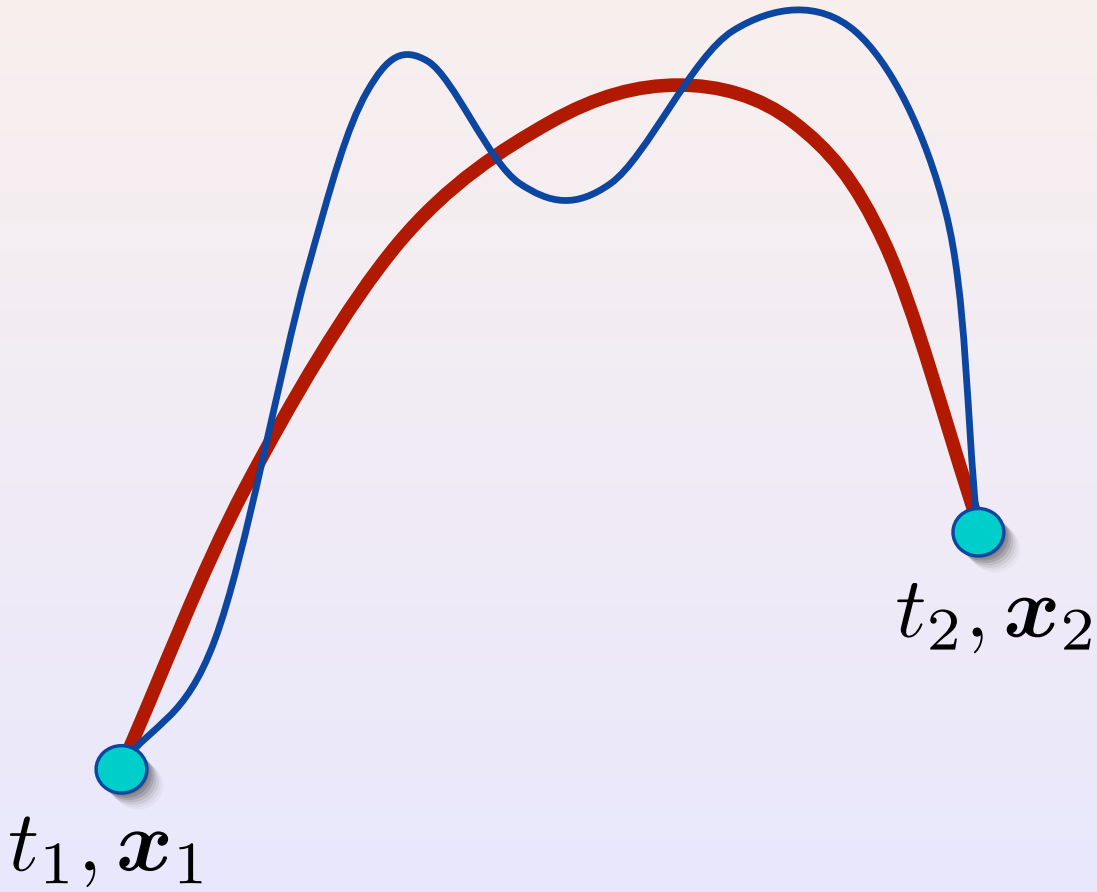
Équation d'évolution locale

$$\frac{d^2x}{dt^2} = \frac{f}{m}$$

Caractérisation globale : formulation lagrangienne

"Principe de moindre action"

Apprentissage et dynamique des systèmes



Apprentissage et dynamique des systèmes
Formulation lagrangienne

- ***Degrés de liberté*** (e.g. N)
- ***Espace de configuration*** (N coordonnées)
- ***Espace des phases*** (N coordonnées + N vitesses ou moments)
- ***Trajectoire*** : ligne parcourue dans l'espace des phases

Problème central : Détermination d'une trajectoire

- pour un système
- connaissant le point de départ et les forces agissantes
- ou entre deux points de l'espace des phases

Apprentissage et dynamique des systèmes

Intégrale d'action et principe de moindre action

Intégrale d'action

Fonctionnelle de la trajectoire

Principe de moindre action

Parmi toutes les trajectoire possibles,
la trajectoire effectivement suivie
rend minimale (extrémale) l'intégrale d'action

Intégrale d'action
Exemples

Particule dans un champ conservatif de forces

Lagrangien :

$$\mathcal{L} = E_c - E_p \ \left(= \frac{1}{2}mv^2 - mgh\right)$$

Action :

$$\mathcal{S} = \int_{t_0}^{t_1} \mathcal{L}\, dt$$

Particule chargée dans un champ magnétique (potentiel non conservatif)

Action :

$$\mathcal{S} = -m_0 c^2 \int_{t_0}^{t_1} \sqrt{1 - v^2/c^2}\, dt - q \int_{t_0}^{t_1} [\Phi(x,y,z,t) - v \cdot \boldsymbol{A}(x,y,z,t)]\, dt$$

(ϕ : *potentiel scalaire* ; $\boldsymbol{A}$: *potentiel vecteur*)

Intégrale d'action
Exemples

Particule dans un champ conservatif de forces

Lagrangien :

$$\mathcal{L} = E_c - E_p \ \ (= \frac{1}{2}mv^2 - mgh)$$

Action :

$$\mathcal{S} = \int_{t_0}^{t_1} \mathcal{L}\, dt$$

Particule chargée dans un champ magnétique (potentiel non conservatif)

Action :

$$\mathcal{S} = -m_0 c^2 \int_{t_0}^{t_1} \sqrt{1 - v^2/c^2}\, dt - q \int_{t_0}^{t_1} [\Phi(x, y, z, t) - v \cdot \boldsymbol{A}(x, y, z, t)]\, dt$$

(ϕ : *potentiel scalaire* ; $\boldsymbol{A}$: *potentiel vecteur*)

Intégrale d'action

Le cas du calcul de la moyenne

Exemple de système d'apprentissage (quasi-) minimal

Cas discret

$$\mu(t) = \frac{1}{t}\sum_{i=0}^{t} x(i) \;=\; \frac{(t-1)\cdot\mu(t-1) \;+\; x(t)}{t}$$

Cas continu

$$\mu(t) = \frac{1}{t}\int_{\tau=0}^{t} x(\tau)\,d\tau = \frac{(t-dt)\cdot\mu(t-dt) \;+\; dt\cdot x(t)}{t}$$

Intégrale d'action

Le cas du calcul de la moyenne

Exemple de système d'apprentissage (quasi-) minimal

Cas discret

$$\mu(t) = \frac{1}{t}\sum_{i=0}^{t} x(i) \;=\; \frac{(t-1)\cdot\mu(t-1) \;+\; x(t)}{t}$$

Cas continu

$$\mu(t) = \frac{1}{t}\int_{\tau=0}^{t} x(\tau)\,d\tau = \frac{(t-dt)\cdot\mu(t-dt) \;+\; dt\cdot x(t)}{t}$$

Question fondamentale

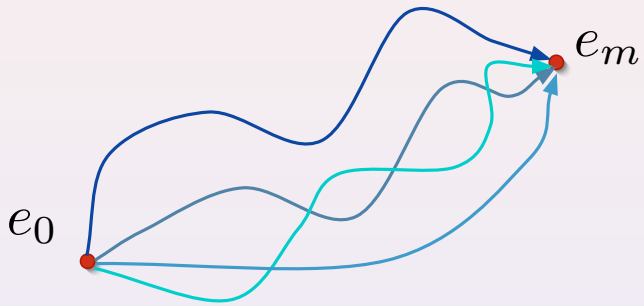
*Quelle **information** doit être **transmise***

d'un état de l'apprenant au suivant

*pour **réaliser un calcul ?***

Un cas particulièrement intéressant

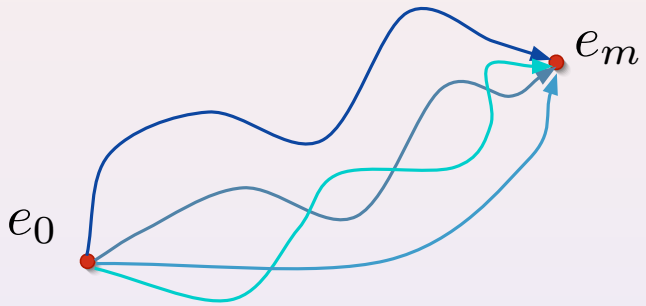
Systèmes insensibles à l'ordre



[Cor93b] A. Cornuéjols.
Getting Order Independence in Incremental Learning.
European Conference on Machine Lerarning (ECML-93), (Ed. P. Brazdil), Vienna, Austria (1993). Springer-Verlag, LNAI-667, pp. 196-212.

Un cas particulièrement intéressant

Systèmes insensibles à l'ordre



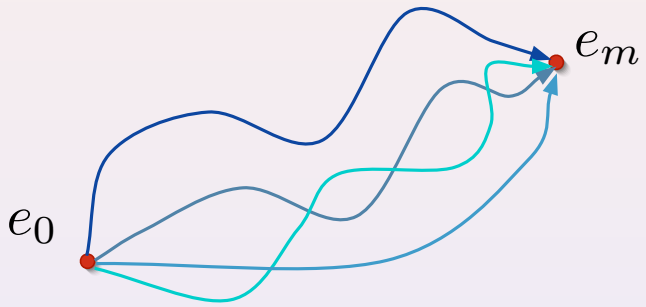
Espace mémoire :

- Calcul du Max :
- Calcul de la Moyenne :
- ID5R :
- Espace des versions :

[Cor93b] A. Cornuéjols.
Getting Order Independence in Incremental Learning.
European Conference on Machine Lerarning (ECML-93), (Ed. P. Brazdil), Vienna, Austria (1993). Springer-Verlag, LNAI-667, pp. 196-212.

Un cas particulièrement intéressant

Systèmes insensibles à l'ordre



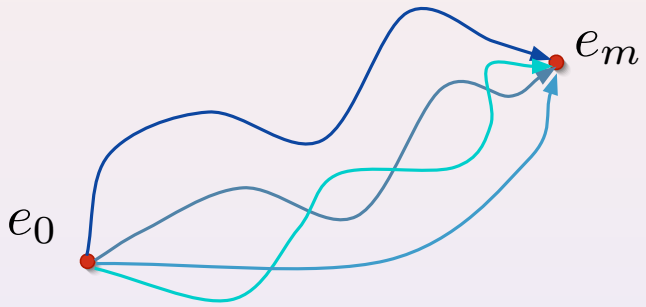
Espace mémoire :

- Calcul du Max : $\max(t)$
- Calcul de la Moyenne :
- ID5R :
- Espace des versions :

[Cor93b] A. Cornuéjols.
Getting Order Independence in Incremental Learning.
European Conference on Machine Lerarning (ECML-93), (Ed. P. Brazdil), Vienna, Austria (1993). Springer-Verlag, LNAI-667, pp. 196-212.

Un cas particulièrement intéressant

Systèmes insensibles à l'ordre



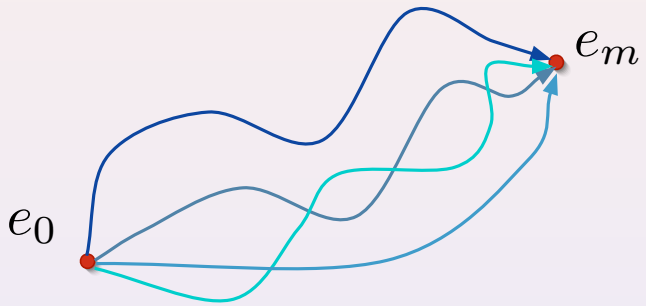
Espace mémoire :

- Calcul du Max : max(t)
- Calcul de la Moyenne : moy(t) et t
- ID5R :
- Espace des versions :

[Cor93b] A. Cornuéjols.
Getting Order Independence in Incremental Learning.
European Conference on Machine Lerarning (ECML-93), (Ed. P. Brazdil), Vienna, Austria (1993). Springer-Verlag, LNAI-667, pp. 196-212.

Un cas particulièrement intéressant

Systèmes insensibles à l'ordre



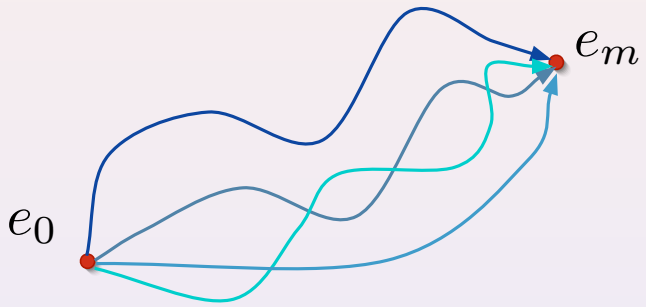
Espace mémoire :

- Calcul du Max : max(t)
- Calcul de la Moyenne : moy(t) et t
- ID5R : $\mathcal{O}(t)$
- Espace des versions :

[Cor93b] A. Cornuéjols.
Getting Order Independence in Incremental Learning.
European Conference on Machine Lerarning (ECML-93), (Ed. P. Brazdil), Vienna, Austria (1993). Springer-Verlag, LNAI-667, pp. 196-212.

Un cas particulièrement intéressant

Systèmes insensibles à l'ordre



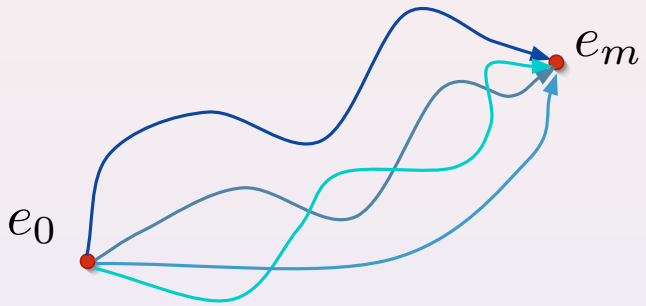
Espace mémoire :

- Calcul du Max : max(t)
- Calcul de la Moyenne : moy(t) et t
- ID5R : $\mathcal{O}(t)$
- Espace des versions : $\leq \mathcal{O}(t)$

[Cor93b] A. Cornuéjols.
Getting Order Independence in Incremental Learning.
European Conference on Machine Lerarning (ECML-93), (Ed. P. Brazdil), Vienna, Austria (1993).
Springer-Verlag, LNAI-667, pp. 196-212.

Un cas particulièrement intéressant

Systèmes insensibles à l'ordre



Espace mémoire :

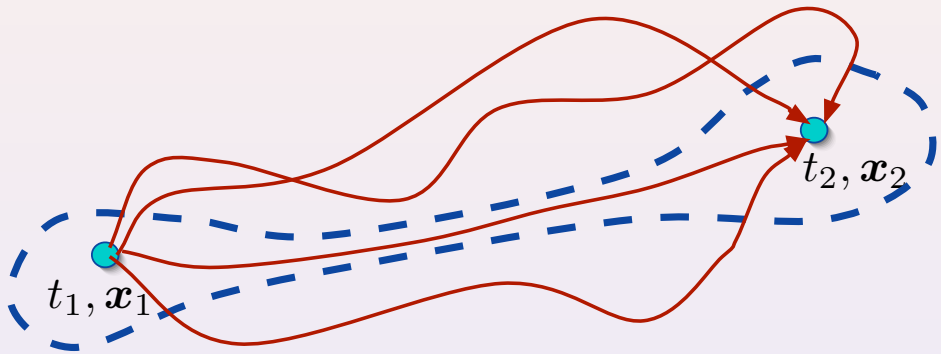
- Calcul du Max : max(t)
- Calcul de la Moyenne : moy(t) et t
- ID5R : $\mathcal{O}(t)$
- Espace des versions : $\leq \mathcal{O}(t)$

Oubli $\equiv$ exemples supplémentaires !

[Cor93b] A. Cornuéjols.
Getting Order Independence in Incremental Learning.
European Conference on Machine Lerarning (ECML-93), (Ed. P. Brazdil), Vienna, Austria (1993). Springer-Verlag, LNAI-667, pp. 196-212.

Systèmes insensibles à l'ordre

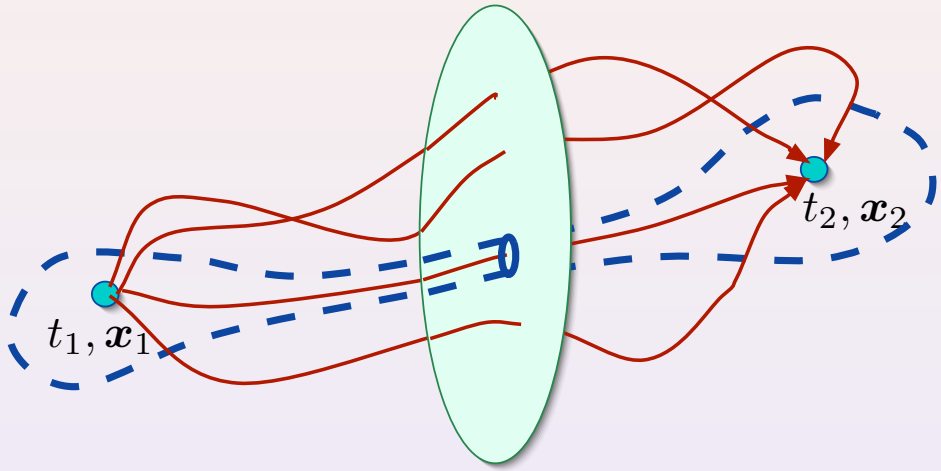
Un courant associé



Pour toute surface enfermant les points de départ et d'arrivée
le nombre de trajectoires sortantes = le nombre de trajectoires rentrantes.

Systèmes insensibles à l'ordre

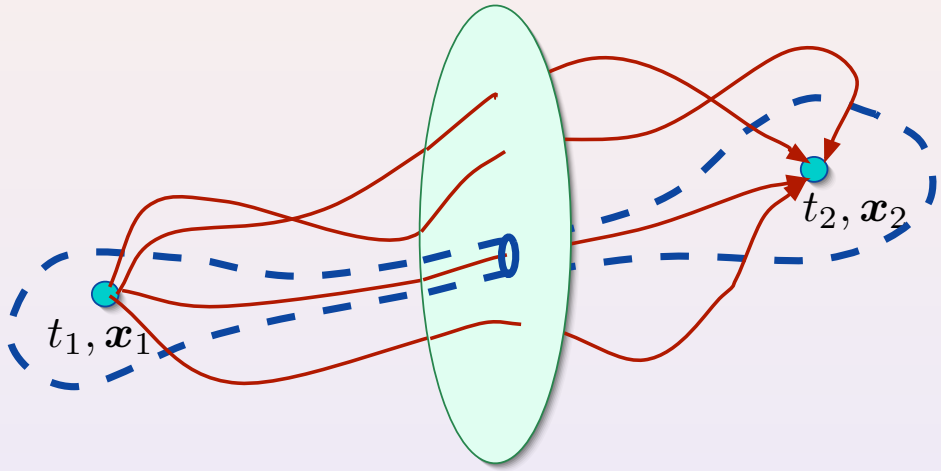
Un courant associé



Pour toute surface enfermant les points de départ et d'arrivée
le nombre de trajectoires sortantes = le nombre de trajectoires rentrantes.

Systèmes insensibles à l'ordre

Un courant associé



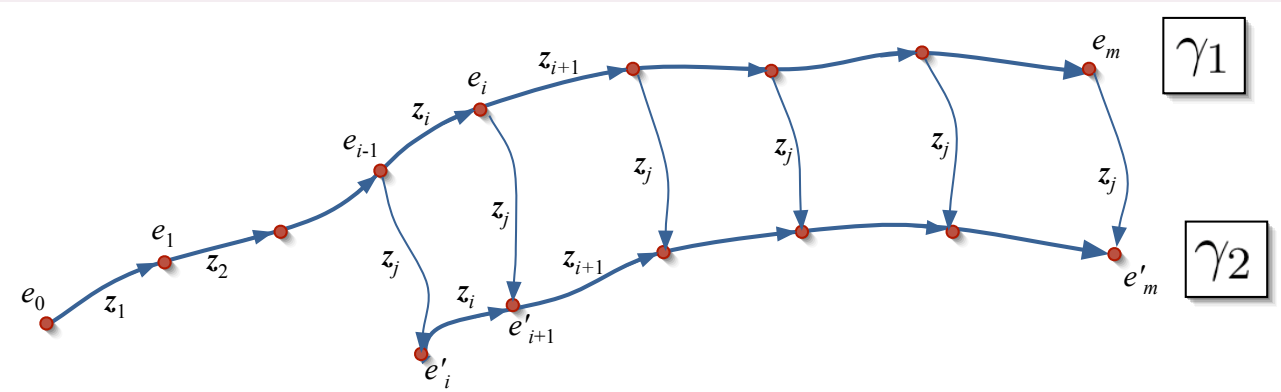
Pour toute surface enfermant les points de départ et d'arrivée
le nombre de trajectoires sortantes = le nombre de trajectoires rentrantes.

Notion de *divergence*

divergence nulle $\Longleftrightarrow$ un **courant conservé associé**

Systèmes insensibles à l'ordre

Un courant associé



Propriétés des trajectoires d'apprentissage

Trajectoires et systèmes dynamiques

Trajectoire $\Longleftrightarrow$ ***Lagrangien*** (moindre action)

Application

Symétrie / permutation des entrées $\Longleftrightarrow$ ***courant associé***

Théorème de Noether

[Cor93a] A. Cornuéjols
Training Issues in Incremental Learning.
AAAI Press, 1993.

[Cor06] A. Cornuéjols
Machine Learning : The Necessity of Order.
In Order to Learn : How ordering processes and sequencing effects in machines illuminate human learning and vice-versa, E. Lehtinen and F. Richter (Eds.), Cambridge University press, 2006.

Symétries par invariance de jauge

Le principe

1. Des grandeurs physiques mesurables (e.g., $\boldsymbol{E}$, $\boldsymbol{B}$)
2. Des grandeurs non observables liées aux précédentes de manière non univoque (e.g. phase Ψ de l'électron, ou V : *potentiel scalaire*, $\boldsymbol{A}$: *potentiel vecteur* associés)
3. Une transformation de jauge (locale) qui agit sur les grandeurs non mesurables
4. Pour compenser, il faut l'interaction avec un nouveau champ (de forces) (e.g. photons : un scalaire)
5. Calcul possible d'un courant (transmission d'information) associé (e.g. un nombre correspondant à la phase)

Toutes les forces connues en physique peuvent se déduire de ce principe

---

[Icke95] Vincent Icke
The force of symmetry.
Cambridge University Press, 1995.

[Ryd96] L. Ryder
Quantum field theory
Cambridge University press, 1996.

Symétries par invariance de jauge

Le cas de l'apprentissage incrémental

1. Grandeurs mesurables :
   état avant - état après
2. Grandeurs non observables :
   ordre de la séquence d'apprentissage
3. Transformation de jauge :
   permutation de l'ordre des entrées
4. Nouveau champ : ?
5. courant (transmission d'information) associé : ?

Données non i.i.d. : vers une dynamique de l'apprentissage

Inévitabilité / opportunité des effets de séquence

Étude de l'apprentissage comme un système dynamique

Développer des outils pour des espaces de programmes

- Produit scalaire
- Changement de référentiel
- Lagrangien

Propriétés déterminantes :

- **capacités de calcul et de mémoire limitées**
- ... la structure de la connaissance

Données non i.i.d. : vers une dynamique de l'apprentissage

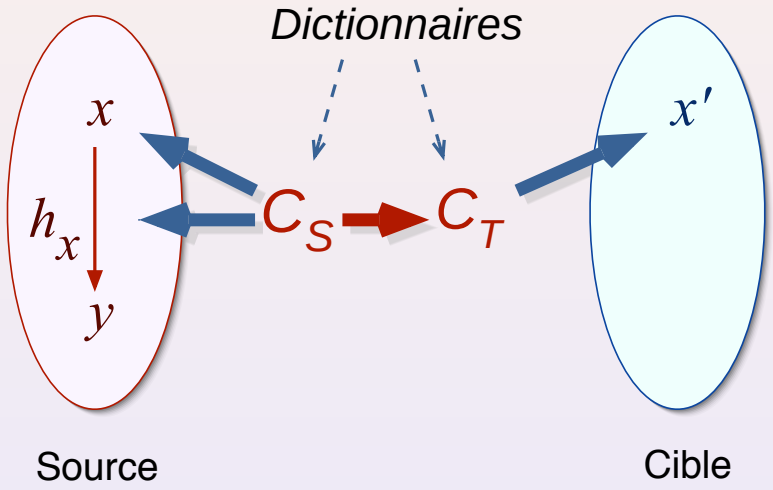
Mieux comprendre la nature de l'information

et de sa circulation

Plan

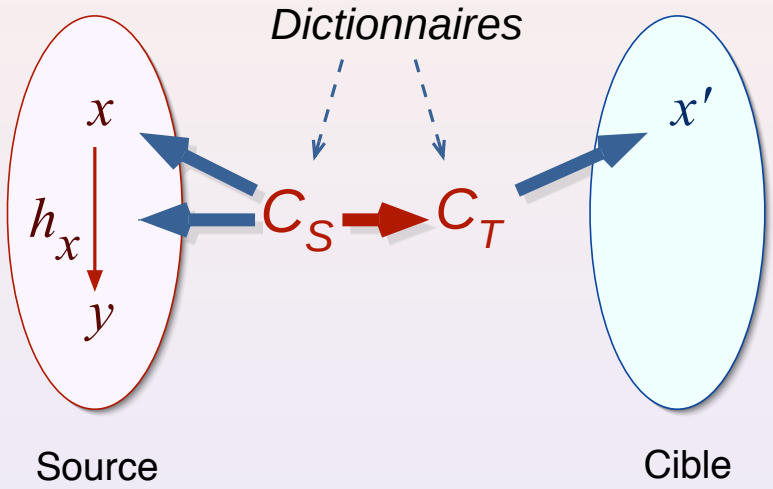
Table des Appendices

Changement de référentiel
et analogie



$$Coût((\mathrm{x} \rightarrow \mathrm{y}), \mathrm{x}') = K(C_S) + K(C_T|C_S) + K(\mathrm{x}|C_S) + K(h_x|C_S) + K(\mathrm{x}'|C_T)$$

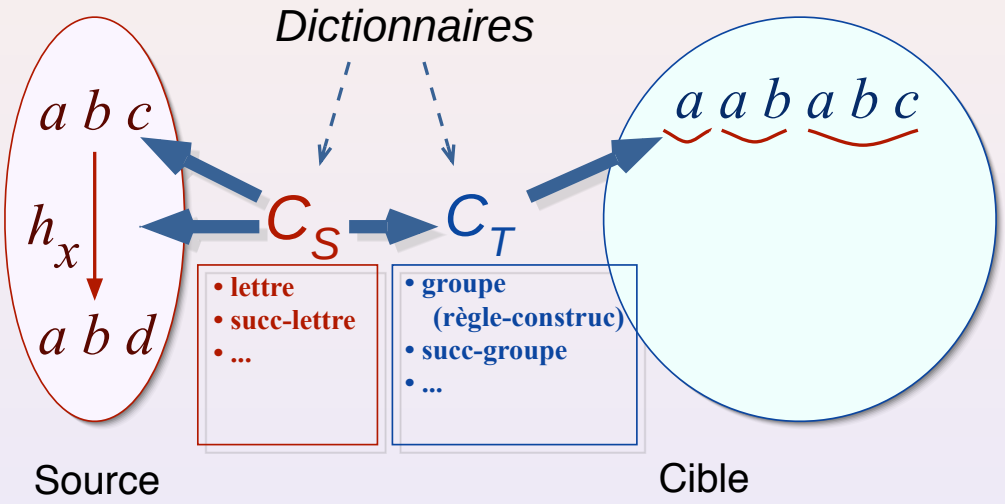
Changement de référentiel
et analogie



$$Coût((\mathrm{x} \to \mathrm{y}), \mathrm{x}') = K(C_S) + K(C_T|C_S) + K(\mathrm{x}|C_S) + K(h_x|C_S) + K(\mathrm{x}'|C_T)$$

Induction = cas particulier de l'analogie

Changement de référentiel
et analogie



[Cor96a] A. Cornuéjols.
Analogie, principe d'économie et complexité algorithmique.
Journées Francophones d'Apprentissage (JFA-96), Sètes, France, 1996, pp.233-247.

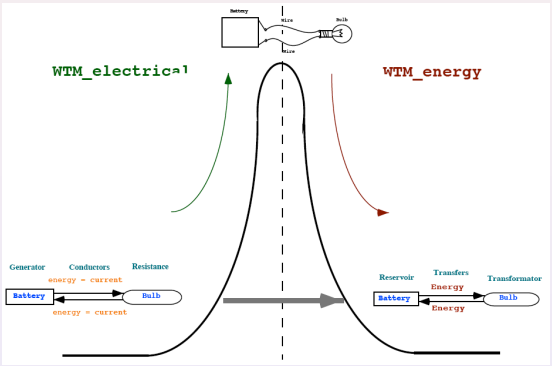
[CAB98] A. Cornuéjols and J. Ales-Bianchetti.
Analogy and Induction : which (missing) link ?
Workshop *Advances in Analogy research : Integration of theory ans data from cognitive, computational and neural sciences*, Sofia, Bulgaria, 1998. New Bulgarian University Series (Eds. K. Holyoak, D. Gentner and B. Kokinov), pp. 365-372.

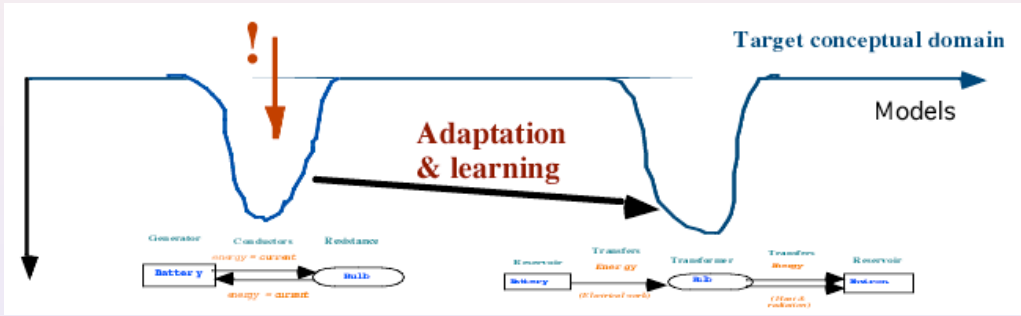
[ECML-94 (wkp)], [COLT-94 (wkp)], [Dagstuhl-94], [Book-chapter-96]

Changement de référentiel

Transferts entre domaines conceptuels

L'effet tunnel cognitif





Apprentissage du concept d'énergie chez des lycéens

[CTC00] A. Cornuéjols, A. Tiberghien and G. Collet
A new mechanism for transfer between conceptual domains in scientific discovery and education
Foundations of Science, vol.5, No.2, (2000), 129-155.

[ECCS'97], [MBR'98], [AISB'99], [CAP'99], [Book-chapter,02]