Cours 2012:

Le cerveau statisticien:
La révolution Bayésienne en sciences cognitives

Stanislas Dehaene
Chaire de Psychologie Cognitive Expérimentale

Cours n°2

Les mécanismes Bayésiens de l'induction

Le « scandale de l'induction » éclairé par la théorie Bayésienne

Un exemple d'induction rapide (Tenenbaum, *Science*, 2011):

Les objects rouges sont des « tufa ».

L’induction Bayésienne dans l’apprentissage du langage:

- si les hypothèses sur le sens des mots sont des branches d’un arbre des catégories de sens envisageables
- alors la règle Bayésienne va automatiquement choisir la catégorie la plus petite, compatible avec les observations (une des versions du « rasoir d’Ockham automatique »)

Supposons que toutes les hypothèses aient la même probabilité a priori (*contra* Rosch et al, 1976).

$P(H|D_1,D_2,D_3)$ est proportionnelle au produit des vraisemblances $P(D_i|H)$ (en supposant que les observations sont conditionnellement indépendantes)

Les hypothèses **H** qui correspondent à des branches « trop petites » sont immédiatement éliminées : leur vraisemblance est nulle pour au moins l'un des mots: $P(D_i|\mathbf{H}) = 0$

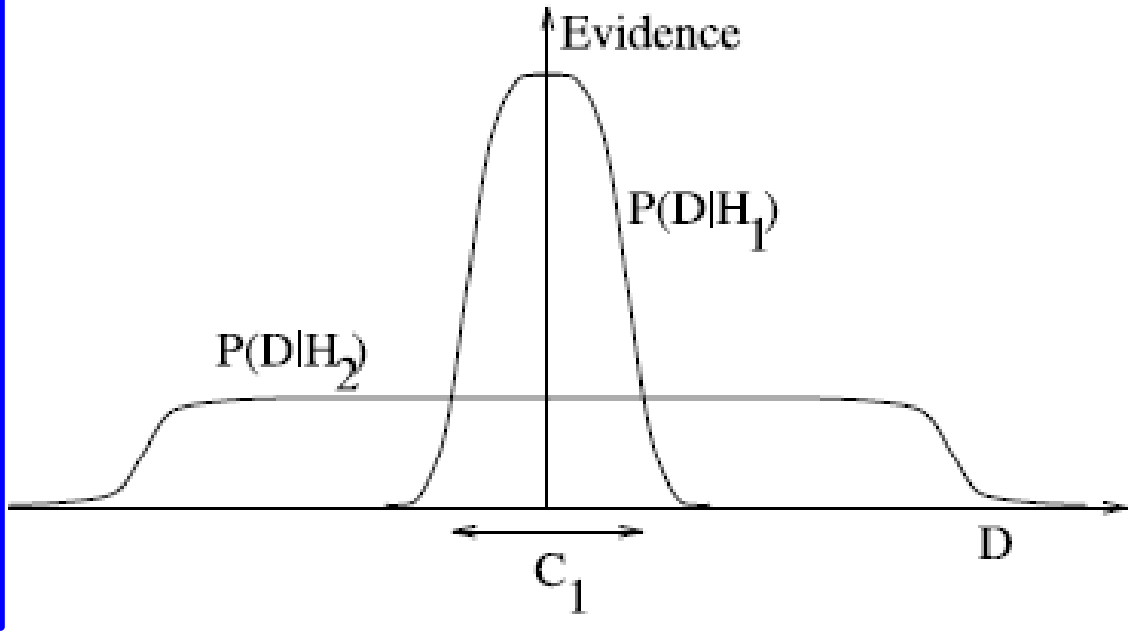
Pour les autres catégories : $P(D_i|H) = 1/n$ où n est le nombre d'éléments de la catégorie

Le mécanisme de Bayes attribue automatiquement une vraisemblance plus faible aux catégories les plus grandes:

$P(D_i|\mathbf{H}) = 1/8$

$P(D_i|\mathbf{H}) = 1/14$

C'est l'une des versions du rasoir d'Ockham:



Supposons que toutes les hypothèses ont la même probabilité a priori (*contra* Rosch et al, 1976).

$P(H|D_1,D_2,D_3)$ est proportionnelle au produit des vraisemblances $P(D_i|H)$ (en supposant que les observations sont conditionnellement indépendantes)

Les hypothèses **H** qui correspondent à des branches « trop petites » sont immédiatement éliminées : leur vraisemblance est nulle pour au moins l'un des mots: $P(D_i|\mathbf{H}) = 0$

Pour les autres catégories : $P(D_i|H) = 1/n$ où n est le nombre d'éléments de la catégorie

Le mécanisme de Bayes attribue automatiquement une vraisemblance plus faible aux catégories les plus grandes:

$P(D_i|\mathbf{H}) = 1/8$

$P(D_i|\mathbf{H}) = 1/14$

Pour un mot, le facteur de Bayes qui sépare les hypothèses **H** et **H** est 14/8 = 1.75

soit 2.4 décibans ($10 \log_{10}(1.66)$) « guère significatif »

La sélection de modèles à l'aide de la règle de Bayes

Le **facteur de Bayes** qui sépare deux hypothèses ou deux « modèles » M_1 et M_2, est une mesure de leur mérite relatif, le rapport de leurs vraisemblances.

$$K = \frac{\Pr(D|M_1)}{\Pr(D|M_2)}$$

Le **logarithme** de cette valeur est une mesure souvent plus intelligible.

On appelle « évidence » (*weight of evidence* [WOE] ou *log odds* [Turing]) la valeur log(p/(1-p)), qui quantifie la vraisemblance d'une hypothèse par rapport aux autres.

Harold Jeffreys propose une interprétation des valeurs de K (Log K est mesuré en *décibans*)

K	dB	bits	Strength of evidence
< 1:1	< 0		Negative (supports M_2)
1:1 to 3:1	0 to 5	0 to 1.6	Barely worth mentioning
3:1 to 10:1	5 to 10	1.6 to 3.3	Substantial
10:1 to 30:1	10 to 15	3.3 to 5.0	Strong
30:1 to 100:1	15 to 20	5.0 to 6.6	Very strong
> 100:1	> 20	> 6.6	Decisive

Supposons que toutes les hypothèses ont la même probabilité a priori (*contra* Rosch et al, 1976).

$P(H|D_1,D_2,D_3)$ est proportionnelle au produit des vraisemblances $P(D_i|H)$ (en supposant que les observations sont conditionnellement indépendantes)

Les hypothèses **H** qui correspondent à des branches « trop petites » sont immédiatement éliminées : leur vraisemblance est nulle pour au moins l'un des mots: $P(D_i|\mathbf{H}) = 0$

Pour les autres catégories : $P(D_i|H) = 1/n$ où n est le nombre d'éléments de la catégorie

Le mécanisme de Bayes attribue automatiquement une vraisemblance plus faible aux catégories les plus grandes:

$P(D_i|\mathbf{H}) = 1/8$

$P(D_i|\mathbf{H}) = 1/14$

Pour un mot, le facteur de Bayes qui sépare les hypothèses **H** et **H** est 14/8 = 1.75

soit 2.4 décibans ($10 \log_{10}(1.66)$) « guère significatif »

Mais… à chaque observation nouvelle, les vraisemblances **se multiplient**:

$P(H|D_1,D_2…D_w)$ est proportionnelle à $P(D_i|H)^w$ où w est le nombre de mots observés.

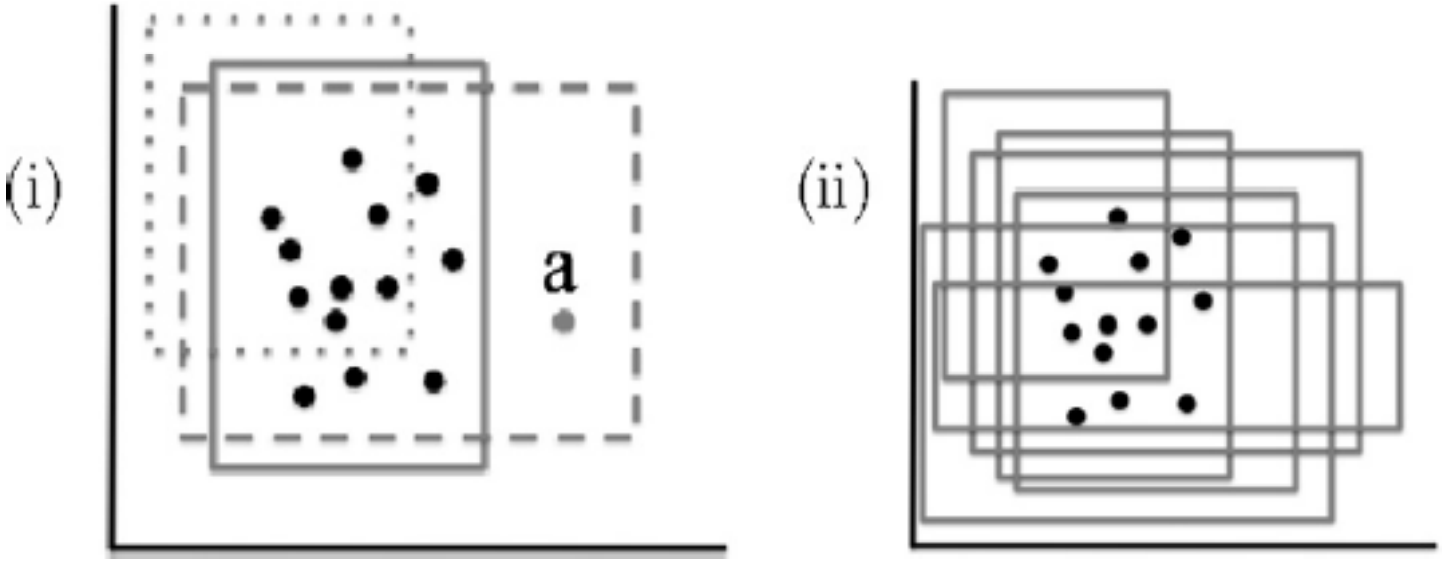
L'évidence, mesurée en décibans, **s'additionne**.

Très peu d'observations suffisent à conclure sur le sens d'un mot.

Et ce, même si les *a prioris* sont défavorables: l'évidence finit toujours par s'imposer!

L’espace des données et l’ensemble des hypothèses

Perfors, A., Tenenbaum, J. B., Griffiths, T. L., & Xu, F. (2011). A tutorial introduction to Bayesian models of cognitive development. *Cognition, 120(3), 302-321*.



L’espace des observations peut être très vaste et multi-dimensionnel: chaque observation consiste en de multiples traits (couleur, forme, nom, déplacement, etc.)

L’ensemble des hypothèses est donc également très vaste, voire infini (des procédures approximatives existent pour n’examiner qu’un sous-ensemble d’hypothèses).

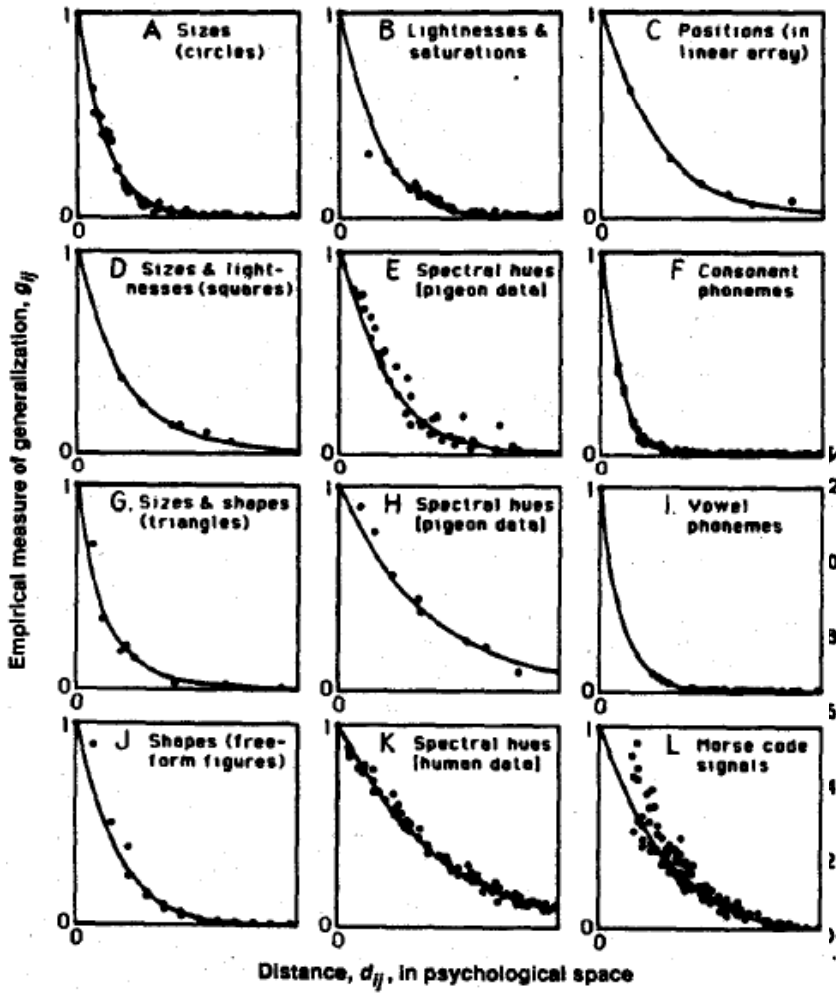
Mais il n’est pas nécessaire d’attendre qu’une seule hypothèse ait été retenue pour généraliser convenablement.

Vers une théorie universelle de la généralisation après apprentissage

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237(4820), 1317-1323.*

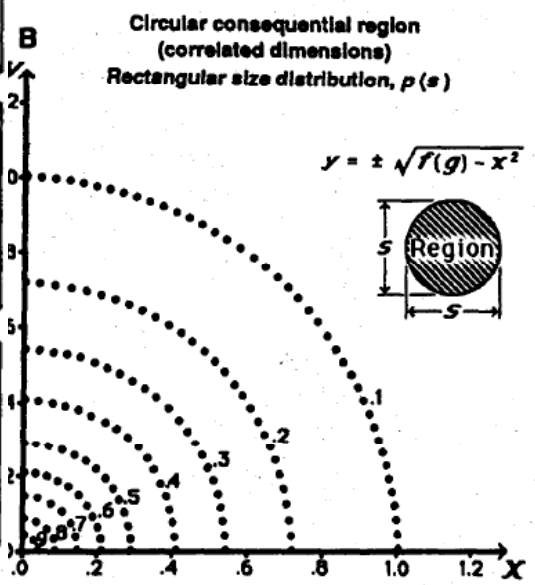
Après un apprentissage, la généralisation est parfois complexe, voire non-monotone.

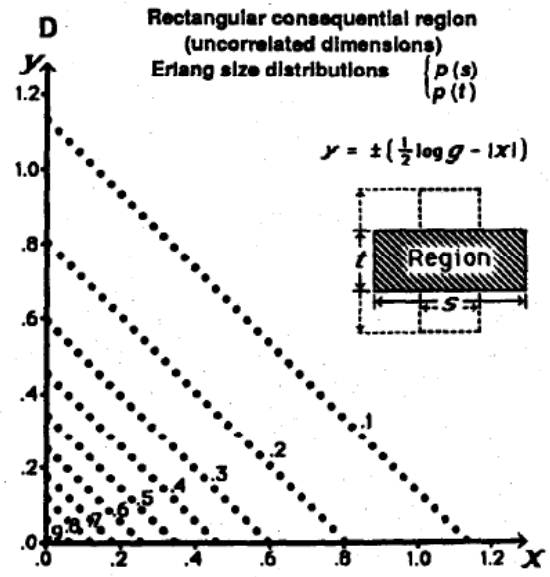
Shepard (1987) montre qu'elle peut *toujours* être ramenée à une fonction exponentielle sur un « espace psychologique » bien défini par « *multidimensional scaling* ».



Généraliser = identifier quelle est la « région conséquentielle » de l'espace psychologique.

Shepard déduit la loi exponentielle en supposant que l'espace des hypothèses ne comprend que des régions connexes, convexes et symétriques, de position quelconque.





Apprendre à apprendre: L'apprentissage Bayésien hiérarchique

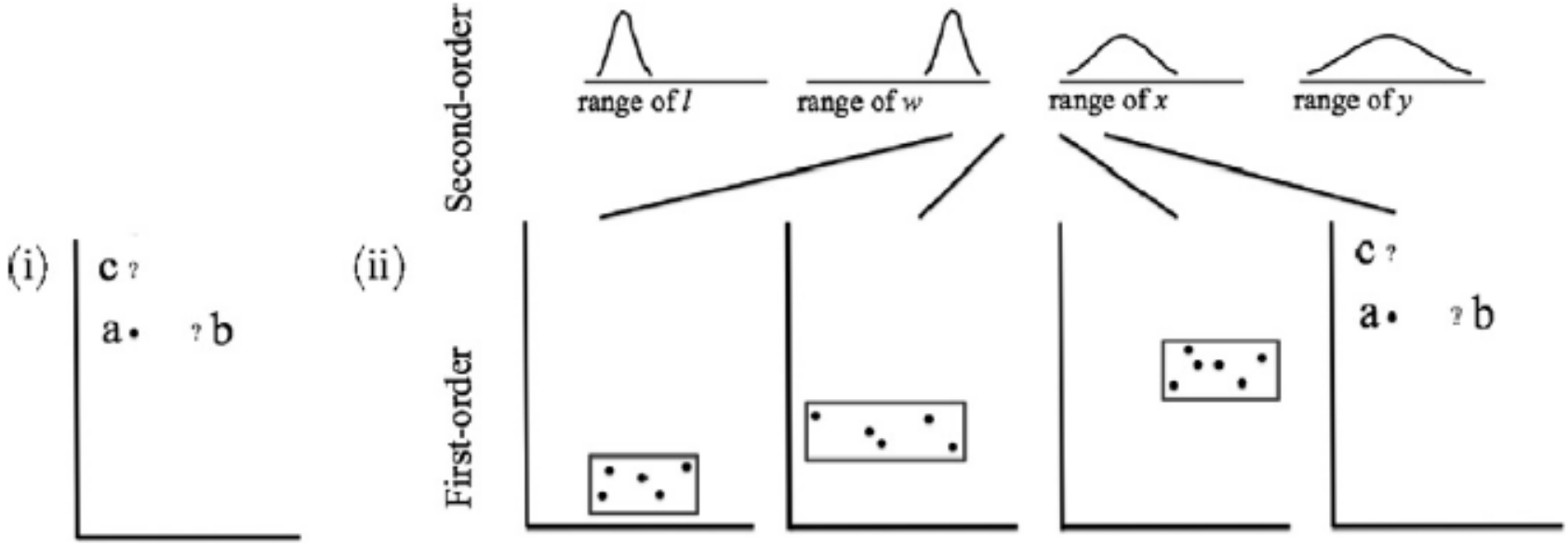
Au premier abord, la théorie Bayésienne attribue une place bien précise à l'inné:

espace des hypothèses, probabilité *a priori*, et fonction de vraisemblance

Cependant, l'espace des hypothèses peut être, à son tour, engendré par une fonction d'ordre supérieur, avec une probabilité *« a priori »* dont les **hyper-paramètres** sont susceptibles d'être appris.

Concept fondamental de **modèle Bayésien hiérarchique**.

- étant donnée une observation **a** , peut-on généraliser aux points b et c?



- si d'autres catégories ont été apprises auparavant, cela permet de fixer certains hyper-paramètres sur la taille typique des régions de l'espace de premier ordre.
- l'apprentissage va alors généraliser très rapidement à de nouvelles instances.

Le transfert de l'apprentissage à l'aide de modèles Bayésiens hiérarchiques

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: statistics, structure, and abstraction. *Science, 331(6022), 1279-1285. See supplementary materials*

Les modèles Bayésiens hiérarchiques possèdent une propriété intéressante: la « bénédiction de l'abstraction » (*blessing of abstraction*):

L'apprentissage est très rapide au niveau le plus élevé, celui qui concerne les principes mêmes d'organisation d'un domaine, car chaque bit de données contribue à sélectionner le modèle pertinent, et celui-ci se généralise à toutes les observations nouvelles.

Cette propriété s'apparente à ce que l'on appelle « le transfert d'apprentissage » ou « apprendre à apprendre ».

Exemple (dû à Nelson Goodman, 1955):

Supposons que l'on tire d'une urne une balle bleue. Que peut-on en conclure sur les tirages suivants?

Peu dans l'absolu.

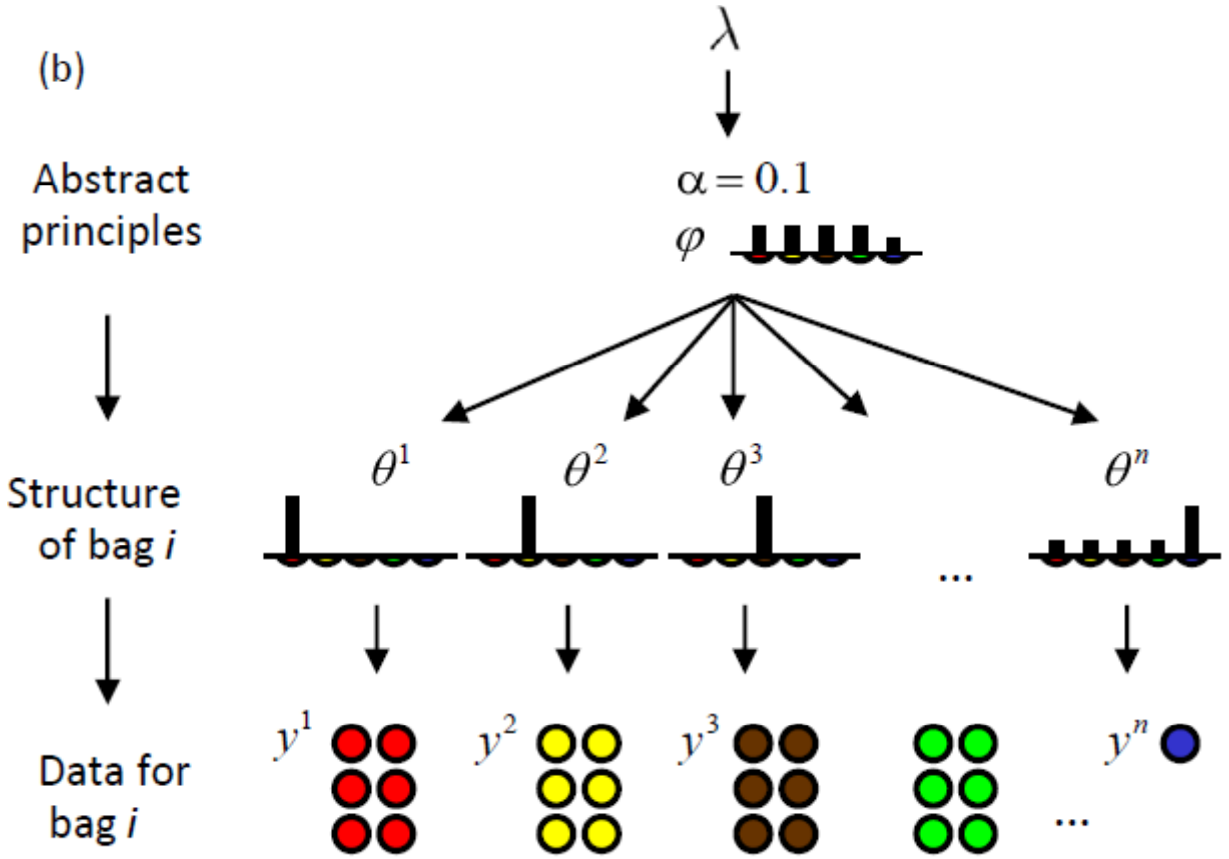
Mais beaucoup si, par le passé, on a fait l'expérience que pour un sac donné, toutes les billes sont toujours de la même couleur.

Le transfert de l'apprentissage à l'aide de modèles Bayésiens hiérarchiques

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: statistics, structure, and abstraction. *Science, 331(6022), 1279-1285. See supplementary materials*

Un modèle Bayésien hiérarchique (modèle Dirichlet-multinomial):

Le paramètre alpha contrôle à quel point chaque sac est unique: une valeur petite de alpha indique que chaque sac possède sa structure particulière de probabilité.

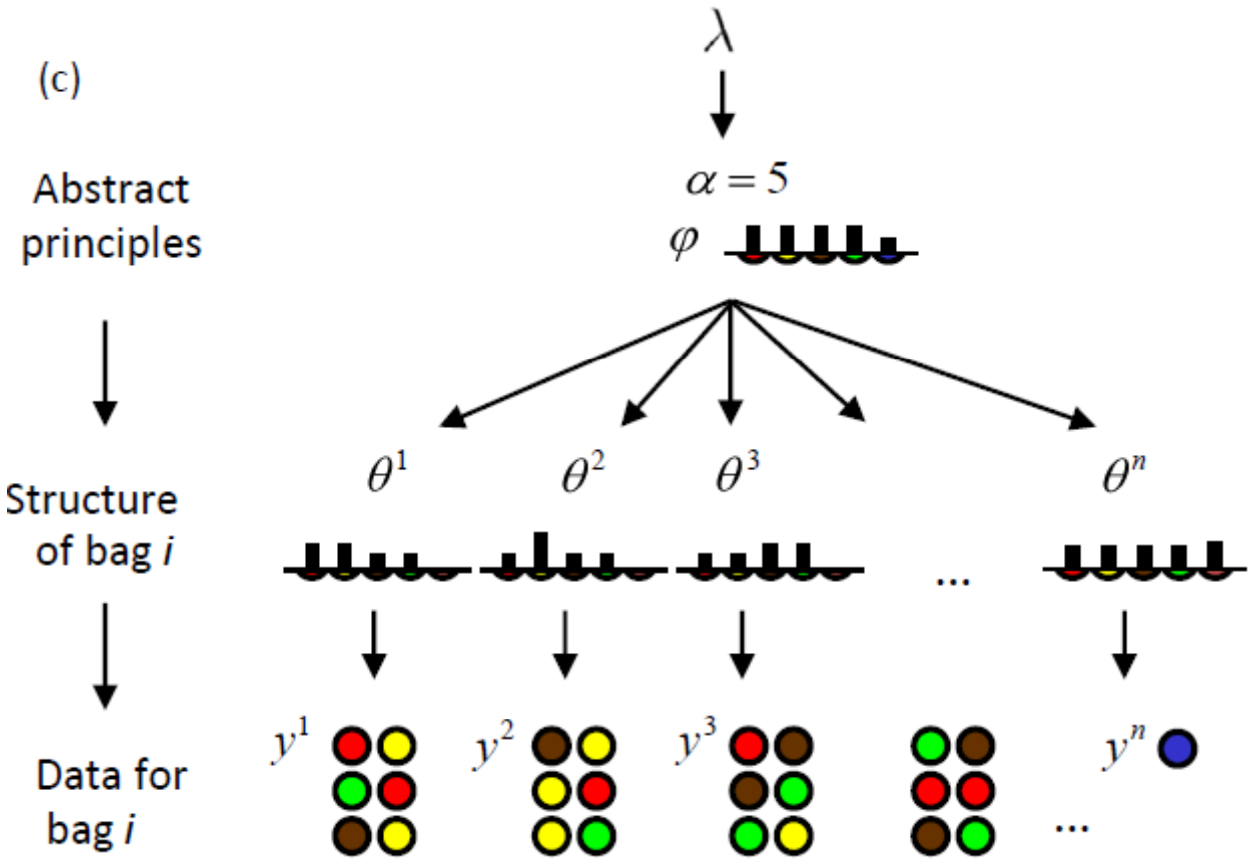


Le transfert de l'apprentissage à l'aide de modèles Bayésiens hiérarchiques

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: statistics, structure, and abstraction. *Science, 331(6022), 1279-1285. See supplementary materials*

Un modèle Bayésien hiérarchique (modèle Dirichlet-multinomial):

Le paramètre alpha contrôle à quel point chaque sac est unique: une valeur petite de alpha indique que chaque sac possède sa structure particulière de probabilité.



Un modèle hiérarchique pour l'apprentissage du langage

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object Name Learning Provides On-the-Job Training for Attention. *Psychological Science (Wiley-Blackwell), 13(1), 13.*

Etude longitudinale de l'apprentissage des mots chez l'enfant de 17 à 19 mois.

Apprentissage de mots associés à des formes, puis test de la généralisation (1) à des formes nouvelles; (2) à l'apprentissage de mots nouveaux.

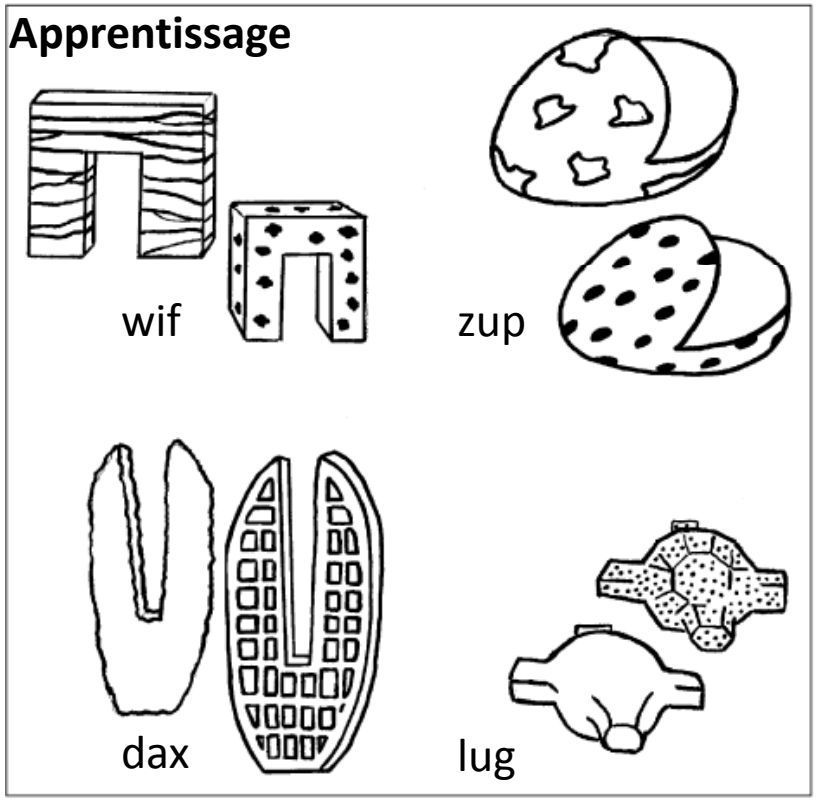
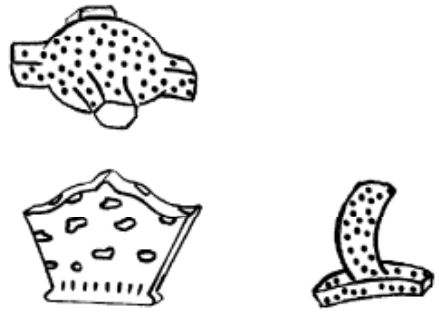


Fig. 2. Illustration of the training stimuli in Experiment 1. The stimuli included two exemplars for each of four novel object categories with novel names. Exemplars of the same category had the same shape, but differed in size, texture, and color.

Test 1

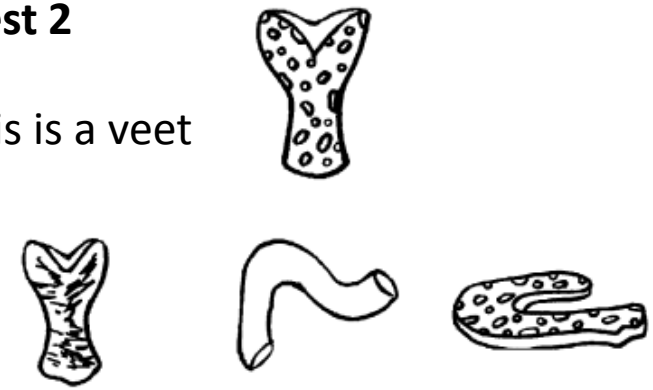
this is a lug



Where is the lug? Get the lug!

Test 2

this is a veet



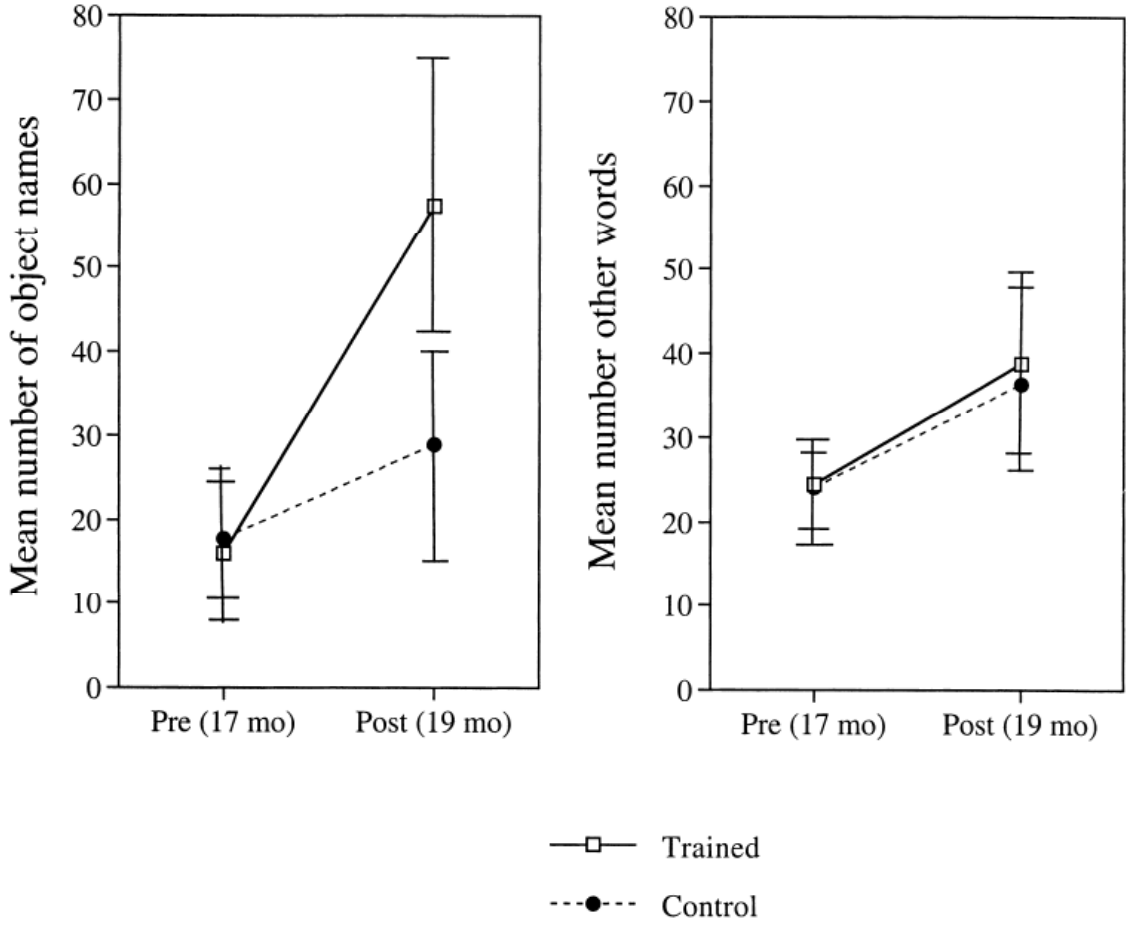
Where is the veet? Get the veet!

Un modèle de méta-apprentissage dans l'apprentissage du langage

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object Name Learning Provides On-the-Job Training for Attention. *Psychological Science (Wiley-Blackwell), 13(1), 13.*

Résultats: Il existe effectivement une généralisation

- d'ordre 1 (à des formes nouvelles pour des mots appris): 88% correct, hasard = 33%
- d'ordre 2 (à des mots entièrement nouveaux): 70% correct, hasard = 33%

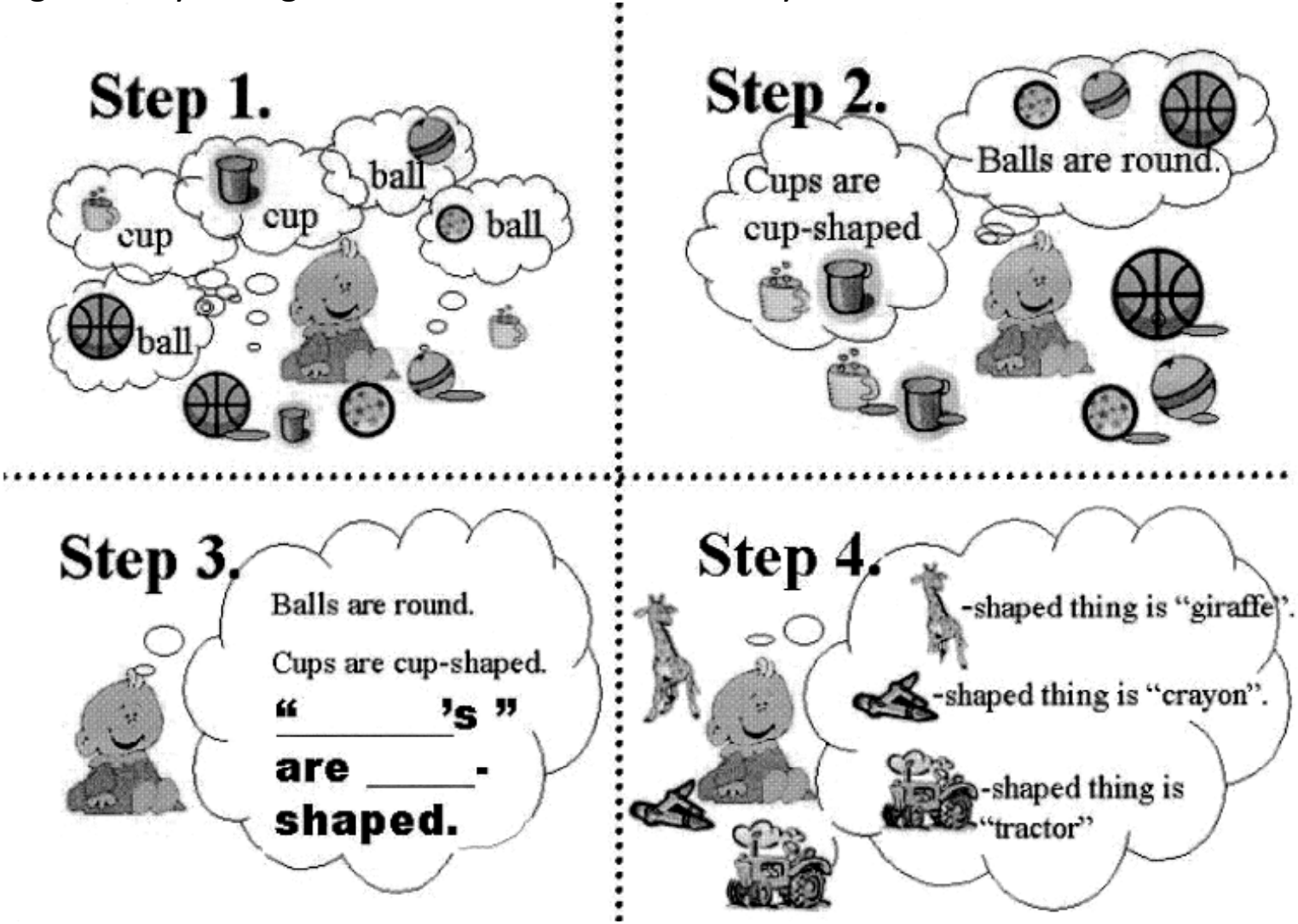


Et même une accélération remarquable du nombre de noms d'objets appris dans la même période de deux mois!

Un modèle de méta-apprentissage dans l'apprentissage du langage

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object Name Learning Provides On-the-Job Training for Attention. *Psychological Science (Wiley-Blackwell), 13(1), 13.*

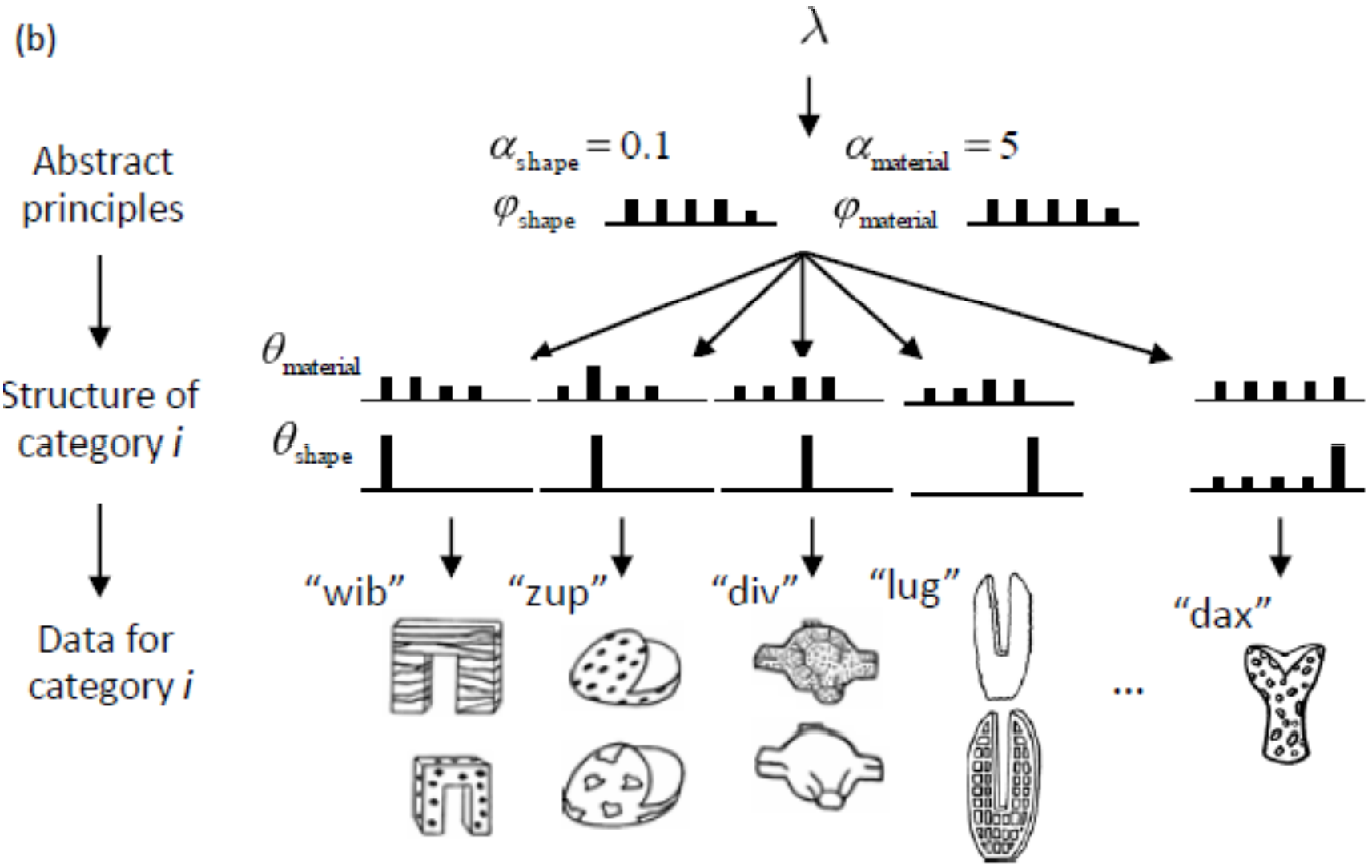
Apprendre à apprendre: "Each bit of individual learning changes the learner, and thus progressively changes what the learner finds easy to learn. »



Un modèle de méta-apprentissage dans l'apprentissage du langage

(Tenenbaum, *Science*, 2011, supplementary materials)

Tenenbaum montre qu'un modèle Bayésien « découvre » le principe que le nom de l'objet dépend de sa forme, mais pas du matériau dont il est fait.

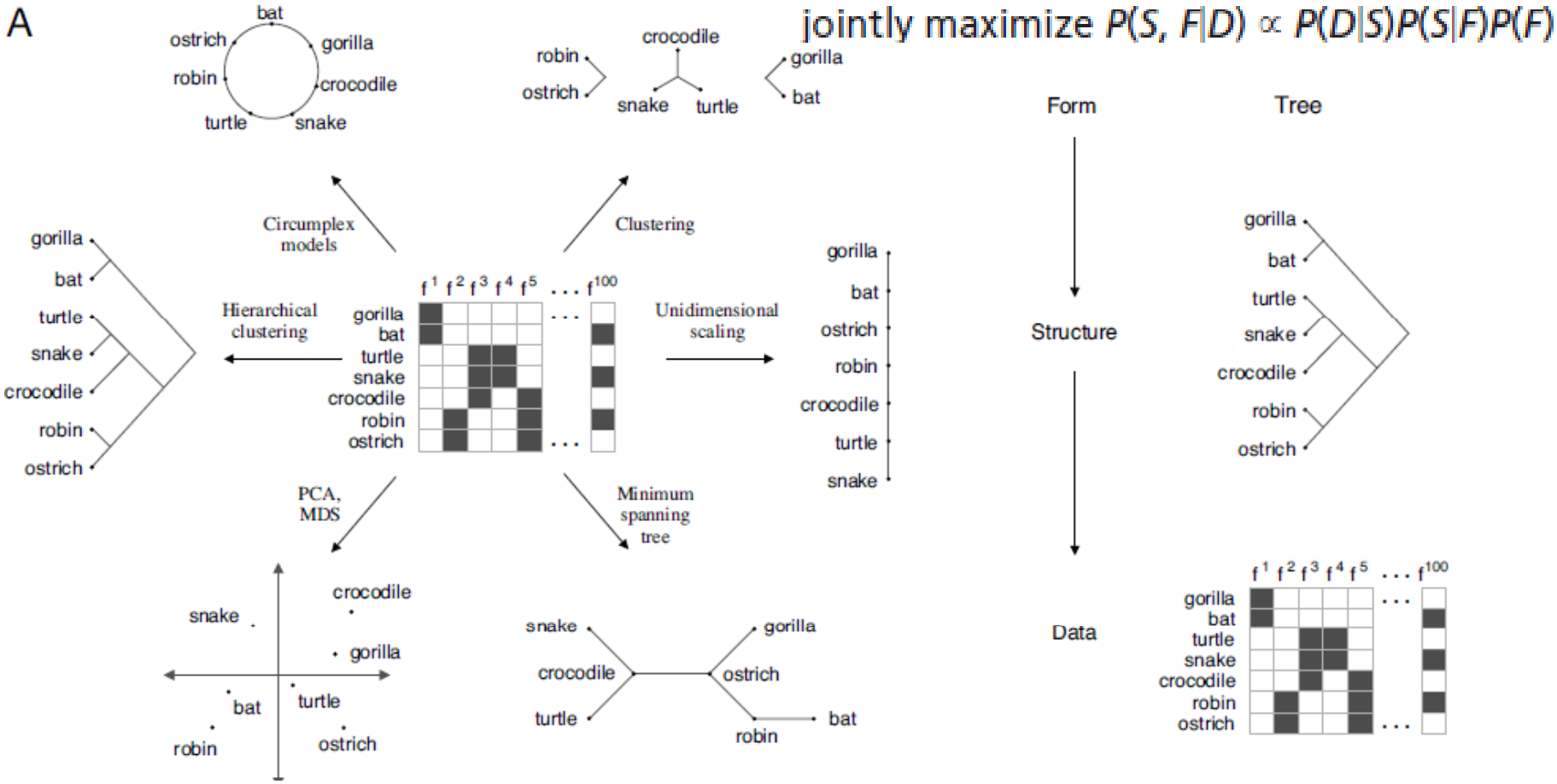


La forme des connaissances abstraites

Le raisonnement Bayésien peut s'appuyer sur plusieurs « **modèles génératifs** »:

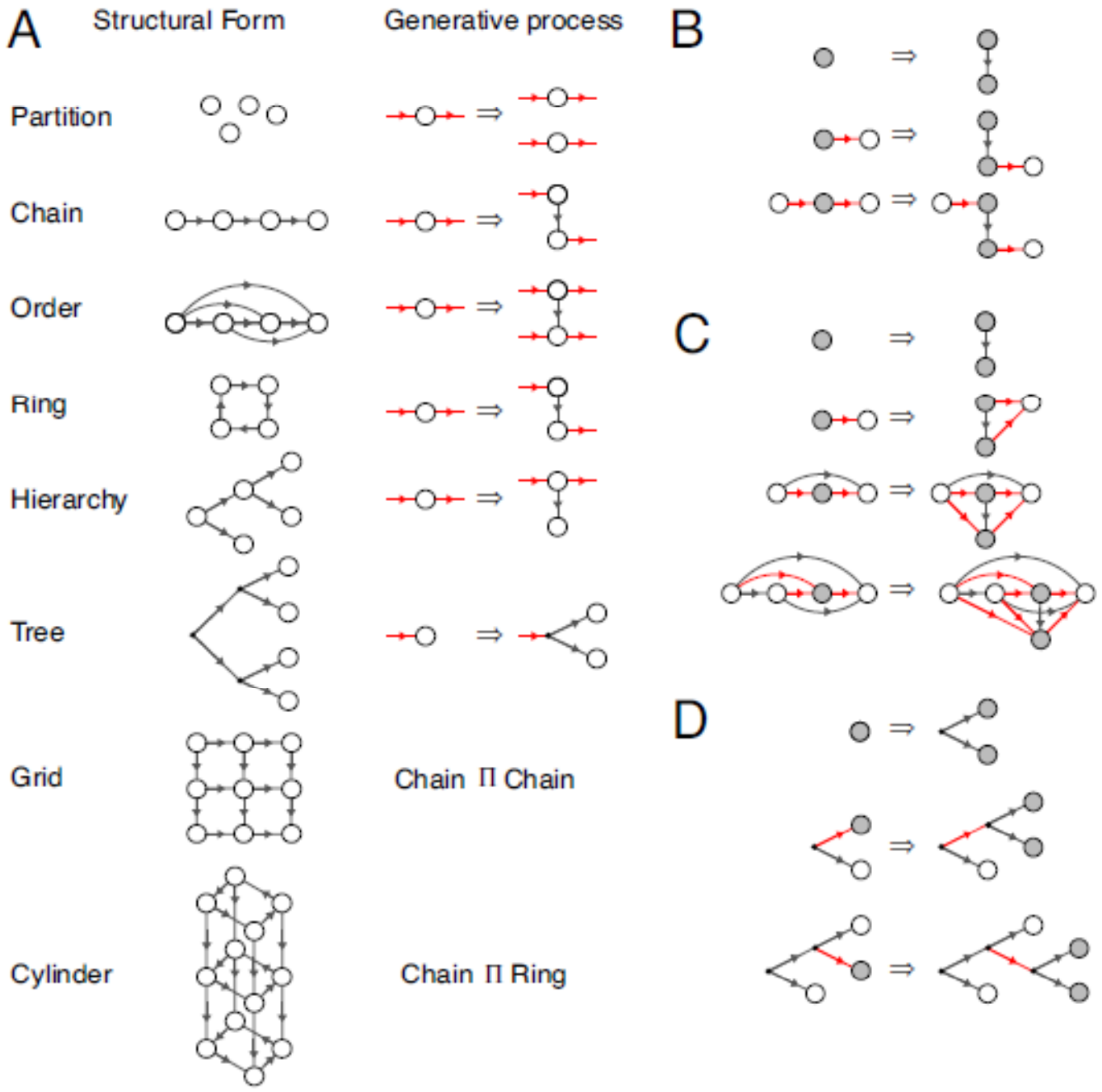
des structures **hiérarchiques**, **abstraites** et **probabilistes** qui constituent des classes d'hypothèses sur la manière dont les exemples observés ont pu être engendrés.

Kemp et Tenenbaum (PNAS, 2008) montrent comment la règle de Bayes permet de choisir automatiquement parmi des arbres, des projections, des graphes orientés… la meilleure manière de représenter des données.



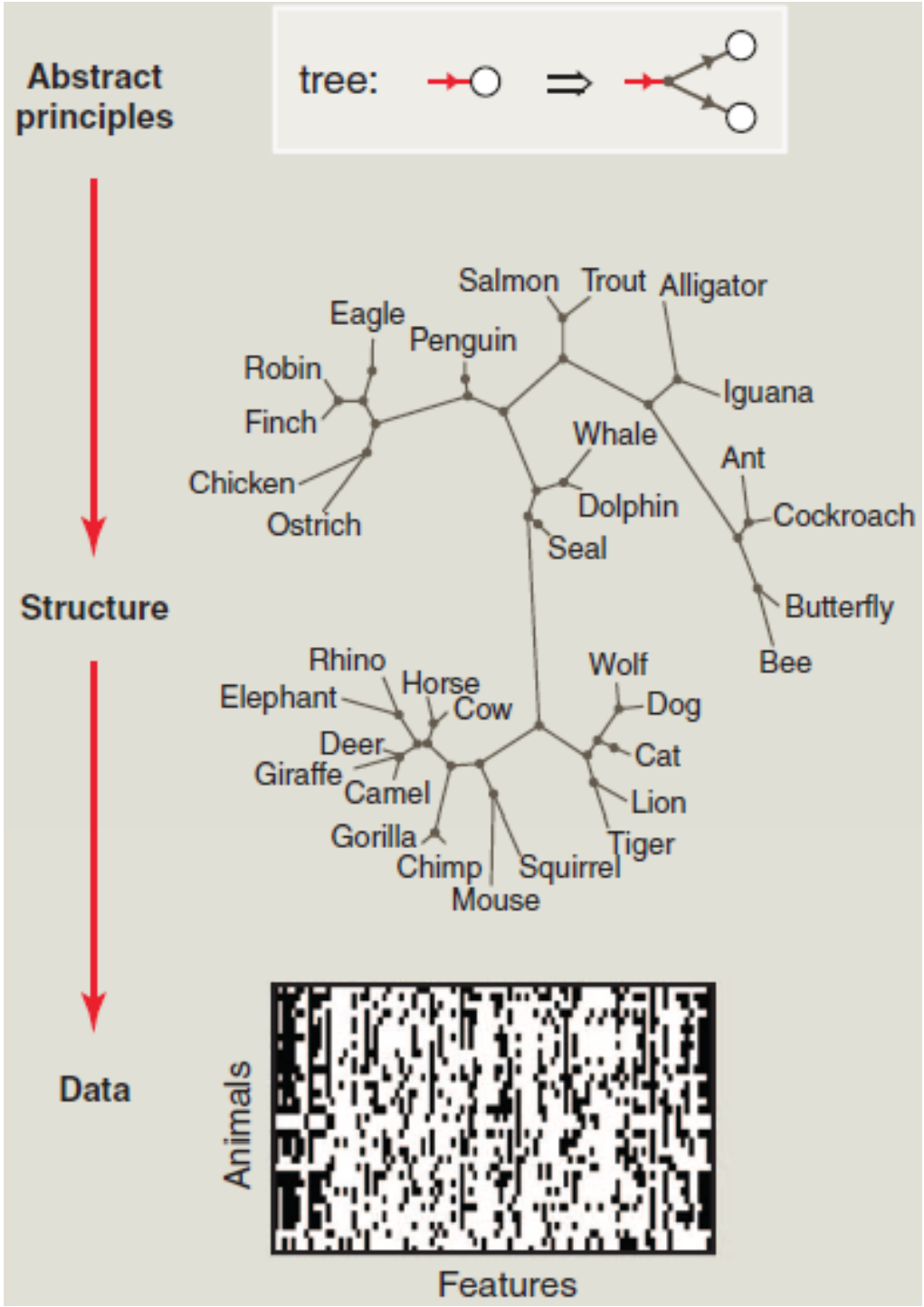
Une grammaire de processus génératifs

(Kemp et Tenenbaum, PNAS 2008)



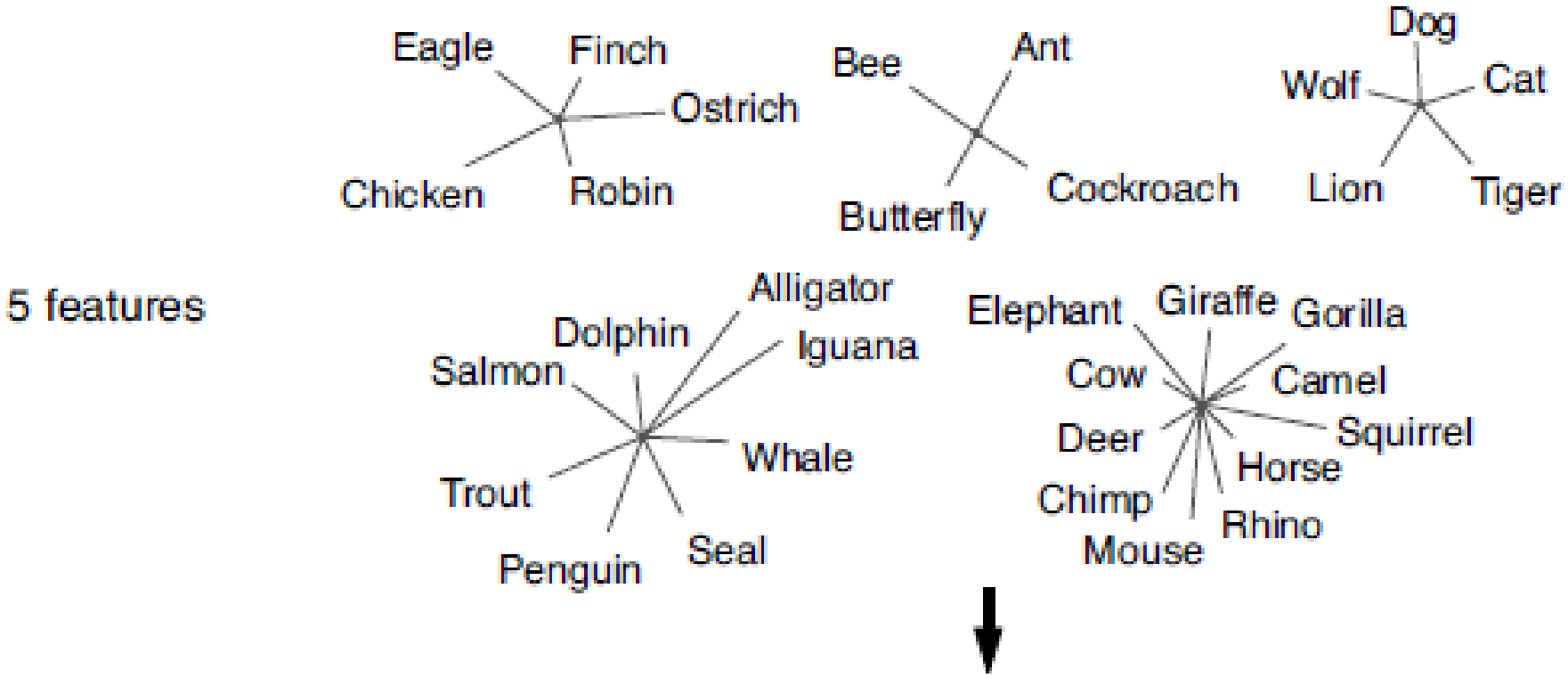
Découverte automatique de la structure arborescente de la famille des êtres vivants

(Kemp et Tenenbaum, PNAS 2008)



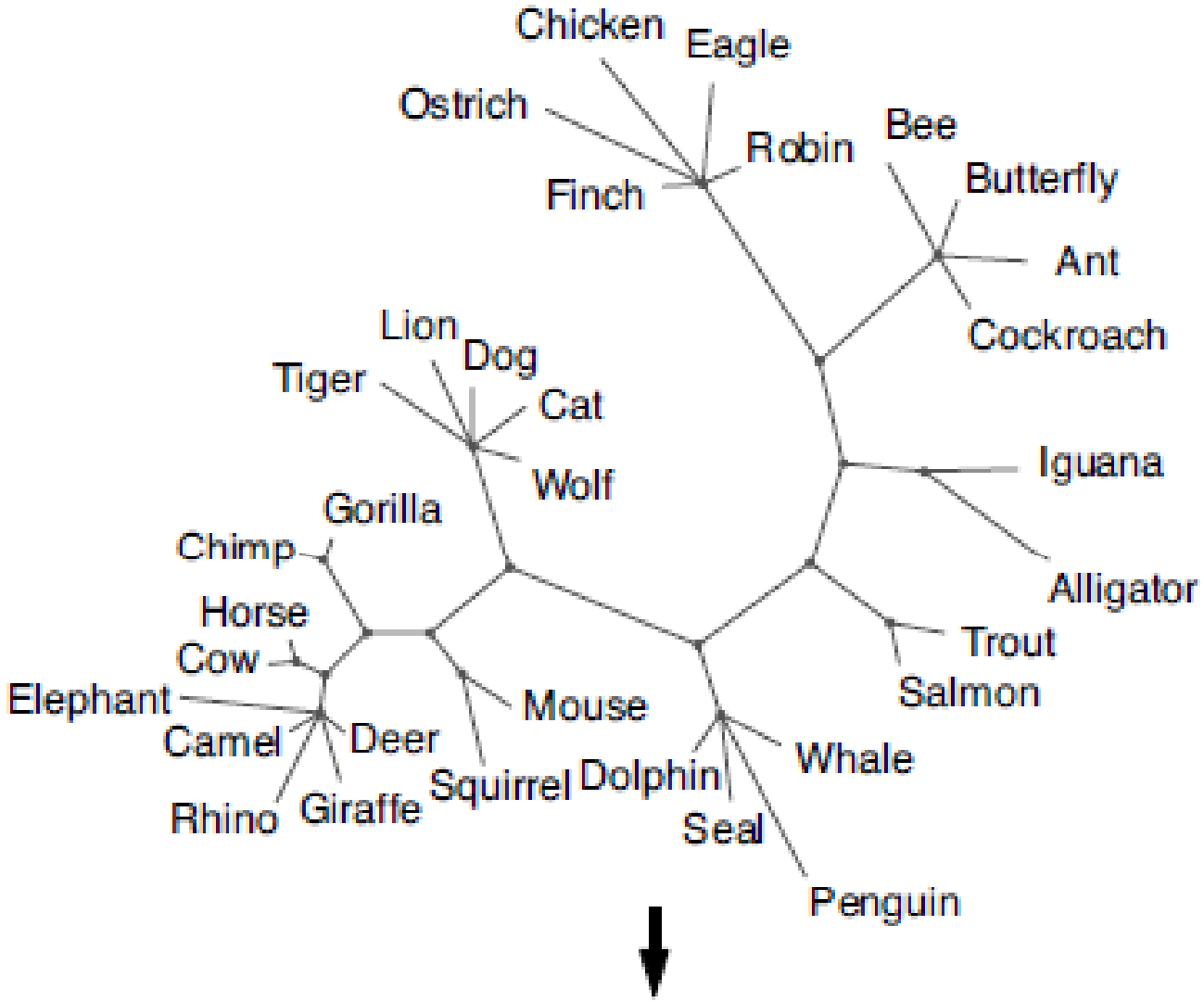
La découverte de la structure est progressive et dépend du nombre de traits disponibles pour chaque animal

(Kemp et Tenenbaum, PNAS 2008)



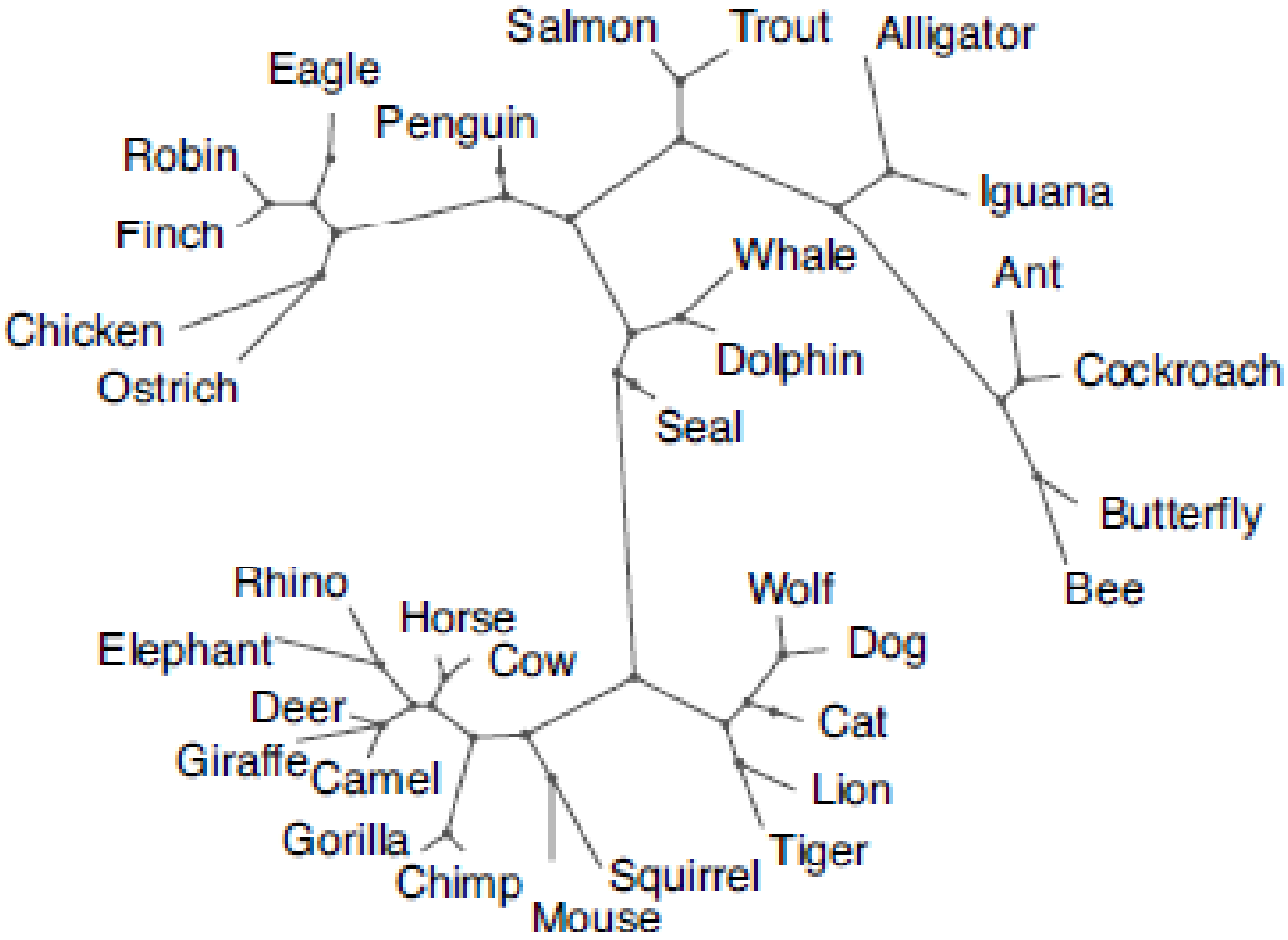
La découverte de la structure est progressive et dépend du nombre de traits disponibles pour chaque animal

20 features



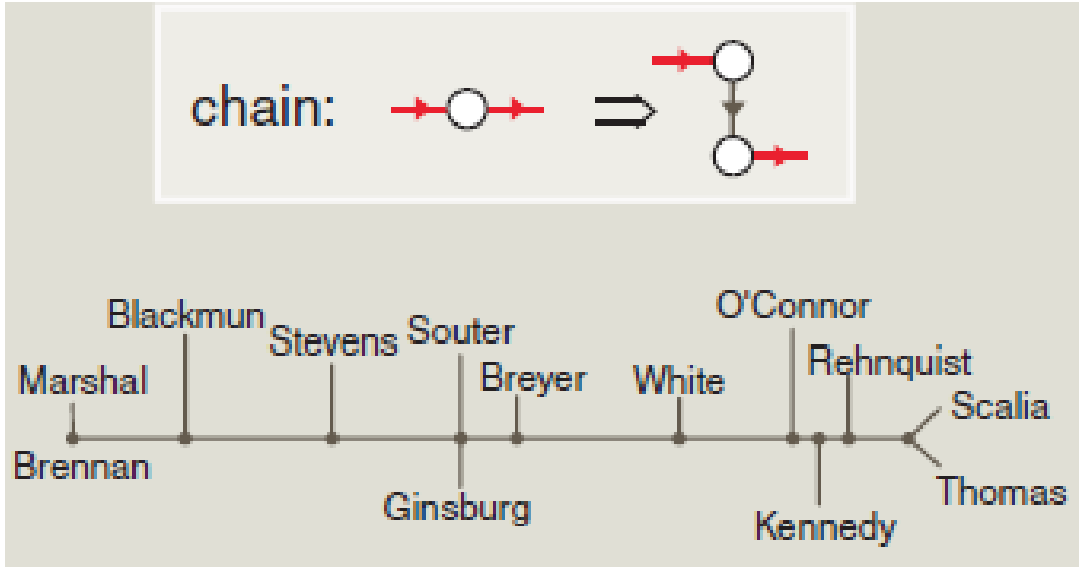
La découverte de la structure est progressive et dépend du nombre de traits disponibles pour chaque animal

110 features

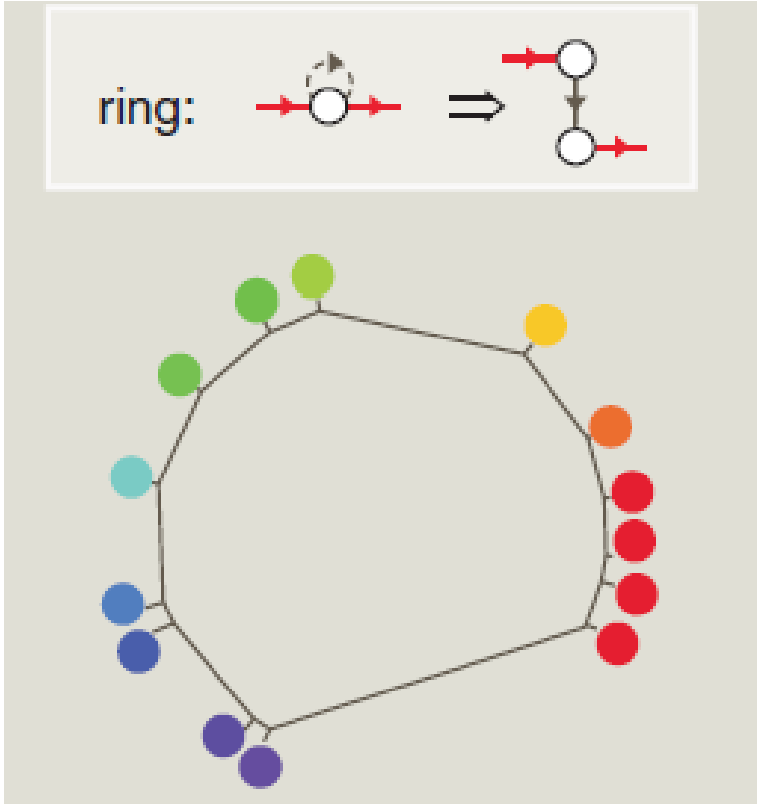


La politique obéit réellement à une structure linéaire gauche-droite!

(Cour suprême des Etats-Unis; Kemp et Tenenbaum, PNAS 2008)

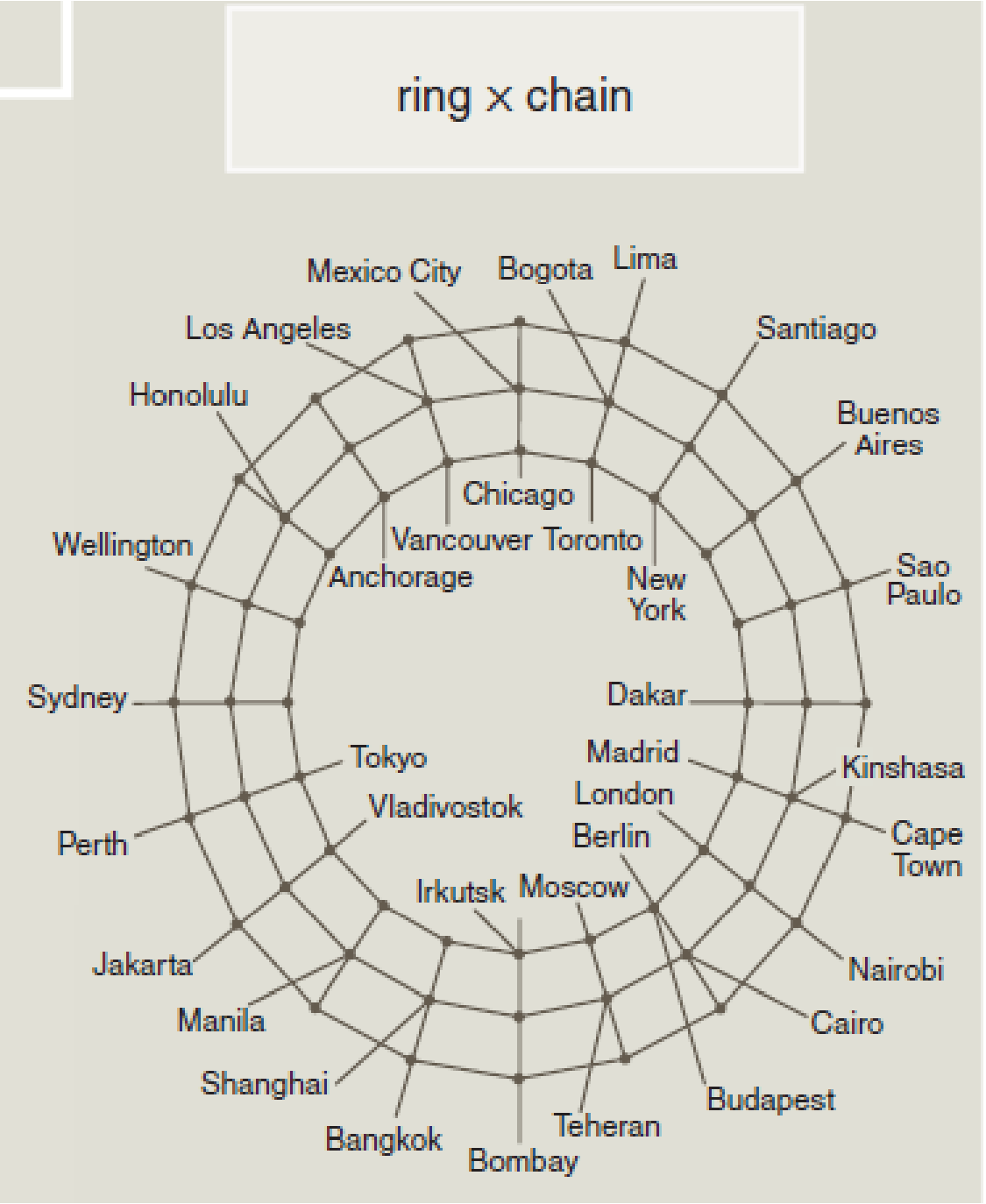


Découverte automatique du cercle des couleurs



Découverte automatique de la structure du globe par produit Cartésien de structures en chaîne et en anneau

(Kemp et Tenenbaum, PNAS 2008)



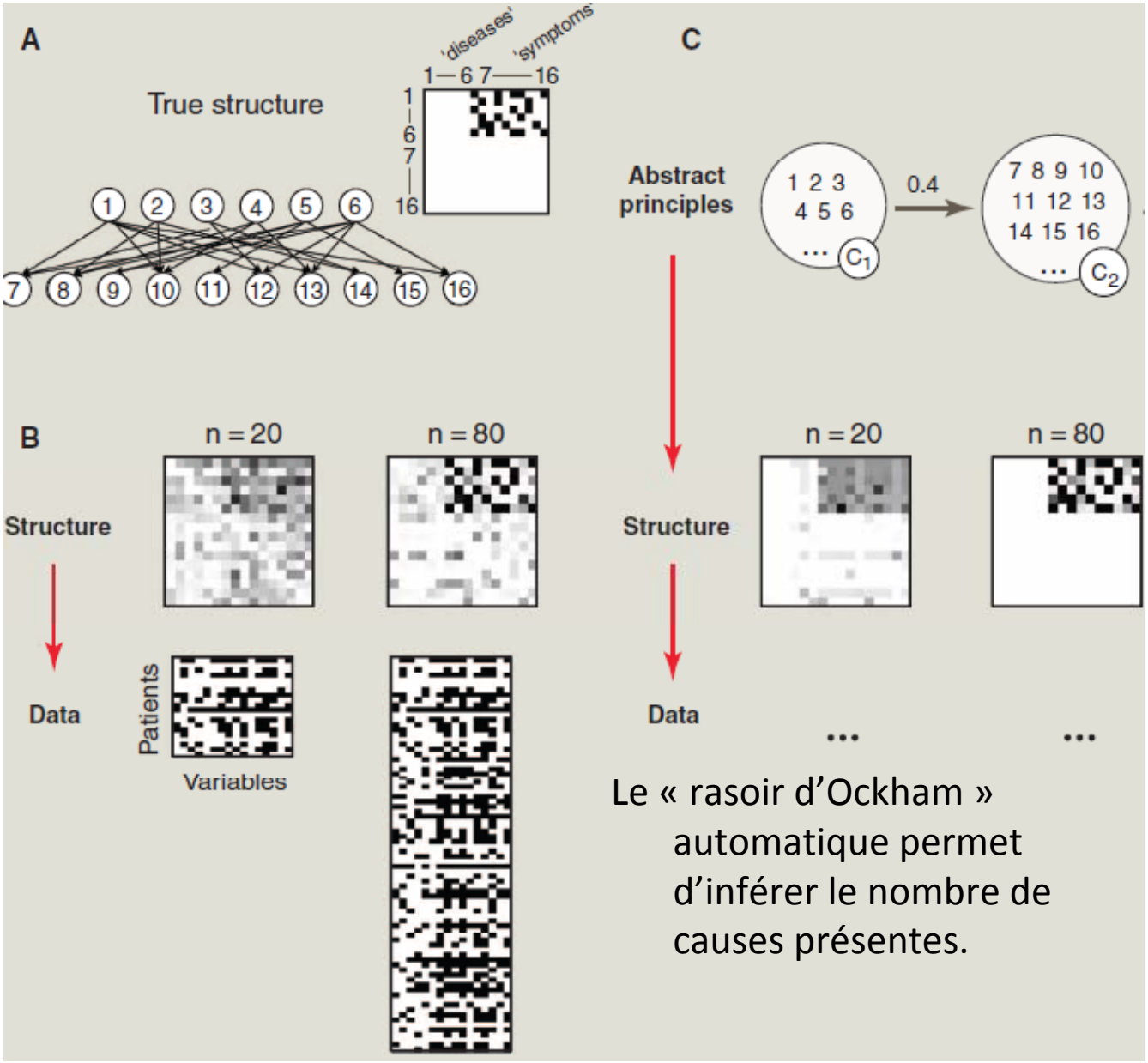
Un principe abstrait facilite l'apprentissage de relations de causalité

Kemp et al, *Cognition*, 2010

A. Un domaine médical est décrit par un réseau de relations entre 16 variables: 6 causes et 10 effets.

B. Un modèle à deux niveaux apprend ces relations en observant les 16 variables chez *n* patients. La matrice représente la probabilité *a posteriori.*

C. L'apprentissage dans un modèle à trois niveaux, guidé par un principe abstrait: il existe des causes et des effets.



Le « rasoir d'Ockham » automatique permet d'inférer le nombre de causes présentes.

Le « scandale de l'induction » éclairé par la théorie Bayésienne

Les modèles Bayésiens incluent à la fois

- des hypothèses a priori très riches (tant sur sur l'espace des représentations mentales accessibles que sur la probabilité relative de chacune d'elles)
- un mécanisme à la fois très simple et très puissant d'apprentissage: l'inférence statistique

Jusqu'à récemment, on ne comprenait pas comment combiner ces deux aspects dans un seul et même modèle computationnel. Nous disposions

- soit, de mécanismes sophistiqués d'apprentissage, mais fondés sur des représentations non-structurées des connaissances (matrices d'associations, connectionnisme)
- soit, de représentations symboliques et structurées des connaissances, mais sans mécanismes d'apprentissage autre que la cohérence logique.

L'approche Bayésienne permet de résoudre un débat classique qui a pris des formes diverses:

- inné – acquis
- nativisme versus empirisme (associationnisme ou connectionnisme)
- Le débat Chomsky versus Skinner ou Piaget

Conclusion:
Pourquoi parler de « révolution Bayésienne » en sciences cognitives?

La théorie Bayésienne est **normative**: c'est tout simplement la manière optimale et rationnelle de tirer des conclusions logiques (**raisonnement plausible**).

L'approche Bayésienne fournit **des algorithmes d'apprentissage** très puissants.

Il est donc naturel de se demander si ce cadre théorique s'applique en sciences cognitives:

- perception, décision et action
- théorie générale de l'état initial et de l'apprentissage, modèle générique du cortex
- apprentissage du langage et accès au lexique
- théorie de l'esprit et psychiatrie (*delusions as inferences from bad data*)

Le cadre Bayésien conduit à poser de nouvelles questions, et notamment:

- quel est l'algorithme par lequel nous approximons des inférences Bayésiennes? (EM [Expectation-Maximization], échantillonnage…)
- quel sont les mécanismes neuronaux d'inférence Bayésienne et de représentation symbolique ?
- Le cerveau du bébé contient-il déjà des mécanismes Bayésiens quasi optimaux?
- Quel est l'espace d'hypothèses accessibles au bébé?
- Pourquoi nos décisions s'écartent-elles parfois de l'optimalité Bayésienne?

Quelques références utilisées pour préparer ce cours

Braun, D. A., Nagengast, A. J., & Wolpert, D. M. (2011). Risk-sensitivity in sensorimotor control. *Front Hum Neurosci, 5, 1.*

Coltheart, M., Menzies, P., & Sutton, J. (2010). Abductive inference and delusional belief. *Cogn Neuropsychiatry, 15(1), 261-287.*

Fletcher, P. C., & Frith, C. D. (2009). Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nat Rev Neurosci, 10(1), 48-58.*

Friston, K. (2005). A theory of cortical responses. *Philos Trans R Soc Lond B Biol Sci, 360(1456), 815-836.*

Glimcher, P. (2003). *Decisions, uncertainty, and the brain: the science of neuroeconomics. Cambridge: MIT Press.*

Glimcher, P., Camerer, C. F., Poldrack, R. A., Rangel, A., & Fehr, E. (Eds.). (2008). *Neuroeconomics: Decision making and the brain. New York: Academic Press.*

Jaynes, E. T. (2003). *Probability theory: the logic of science. Cambridge: Cambridge University Press.*

Kahneman, D. (2011). *Thinking, fast and slow. New York: Farrar Straus Giroux.*

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annu Rev Psychol, 55, 271-304.*

Kruschke, J. K. (2006). Locally Bayesian learning with applications to retrospective revaluation and highlighting. *Psychol Rev, 113(4), 677-699.*

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nat Neurosci, 9(11), 1432-1438.*

Maloney, L. T., & Zhang, H. (2010). Decision-theoretic models of visual perception and action. *Vision Res, 50(23), 2362-2374.*

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: statistics, structure, and abstraction. *Science, 331(6022), 1279-1285.*

Perfors, A., Tenenbaum, J. B., Griffiths, T. L., & Xu, F. (2011). A tutorial introduction to Bayesian models of cognitive development. *Cognition, 120(3), 302-321.*

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychol Rev, 114(2), 245-272.*