

Plan

1. L'IA est attendue partout
2. Quelles garanties sur l'induction
3. Un peu d'histoire
4. Et **demain** ? Prémises de changement de paradigme
5. Retour sur les défis
6. Conclusions ... et ouverture

Sait-on finalement **expliquer**
les **capacités de généralisation** ?

Quelque chose de troublant

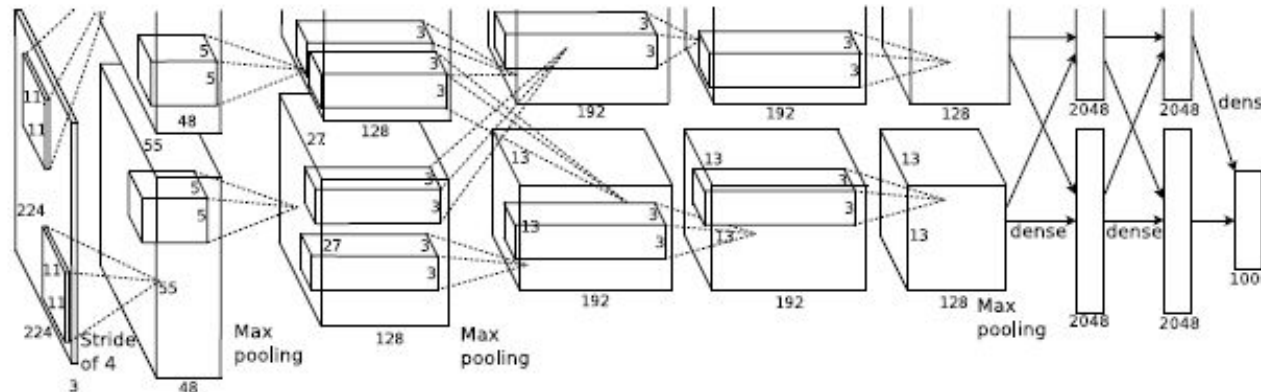
- C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals (ICLR, **May 2017**).
“Understanding deep learning requires rethinking generalization”

Quelque chose de troublant

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

Extensive experiments on the classification of images

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



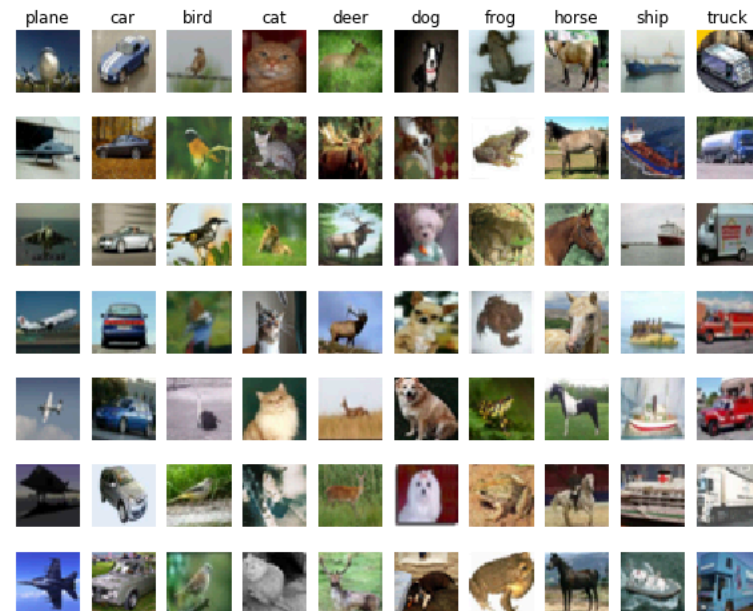
- The **CIFAR-10 data set**:
 - **60,000** images categorized in **10 classes** (50,000 for training and 10,000 for testing)
 - Images: 32x32 pixels in 3 color channels

Quelque chose de troublant

Experiments

1. Original dataset without modification

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



Quelque chose de troublant

Experiments

1. Original dataset without modification

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

Expected behavior if the capacity of the hypothesis space is limited

i.e. the system cannot fit any (arbitrary) training data

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

Troubling findings

Experiments

1. Original dataset without modification

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

2. Random labels

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

!!!



Troubling findings

Experiments

1. Original dataset without modification

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

2. Random labels

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

3. Random pixels

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

Now, we
are in
trouble!!

Troubling findings

- Deep NNs can accommodate ANY training set

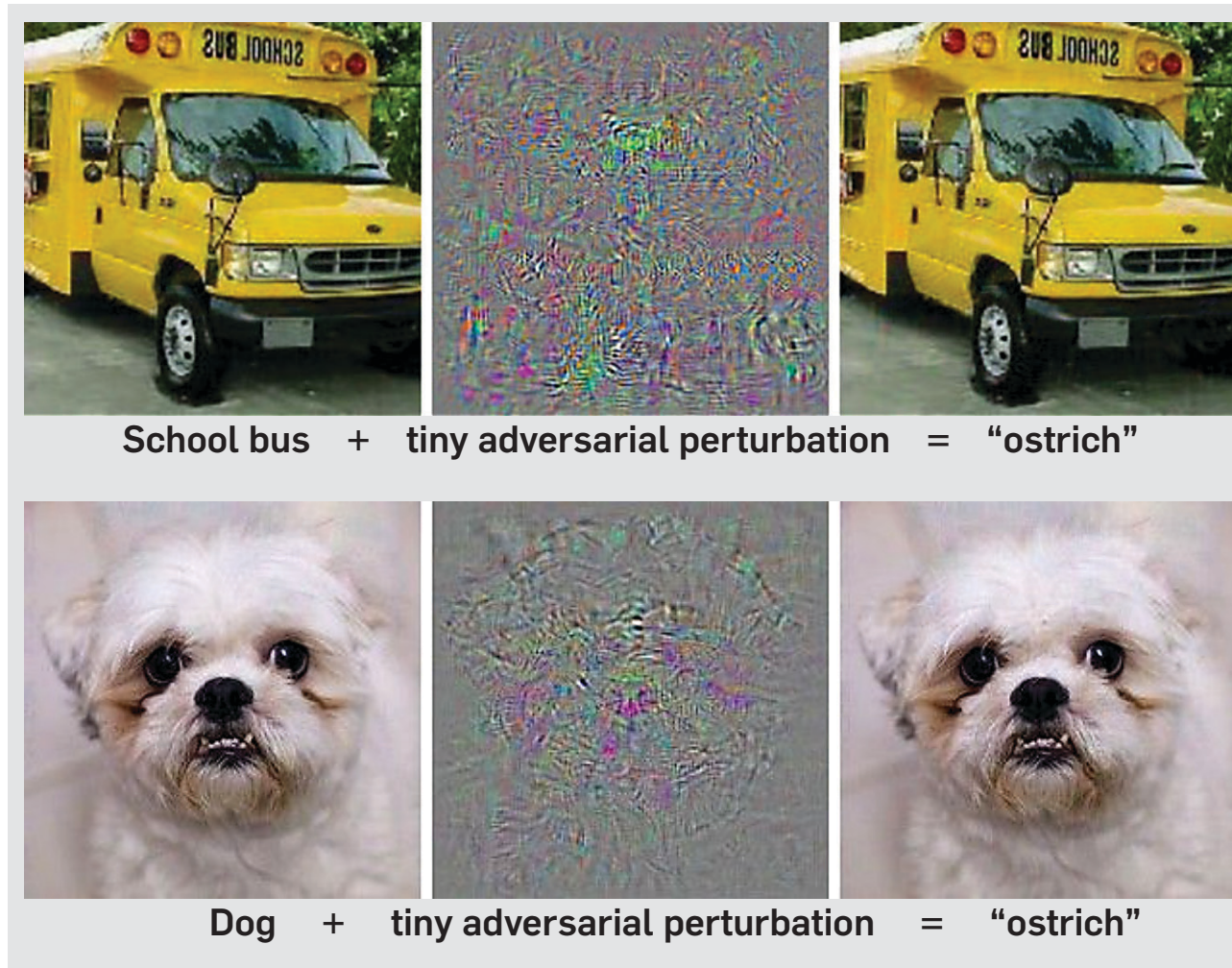
Can grow without limit!!

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

But then,

why are deep NNs so good on image classification tasks?

Adversarial learning



Adversarial input can fool a machine-learning algorithm into misperceiving images.

Sait-on expliquer une conclusion ?

Voiture dans une piscine

- ... ou pas de voiture ... ?



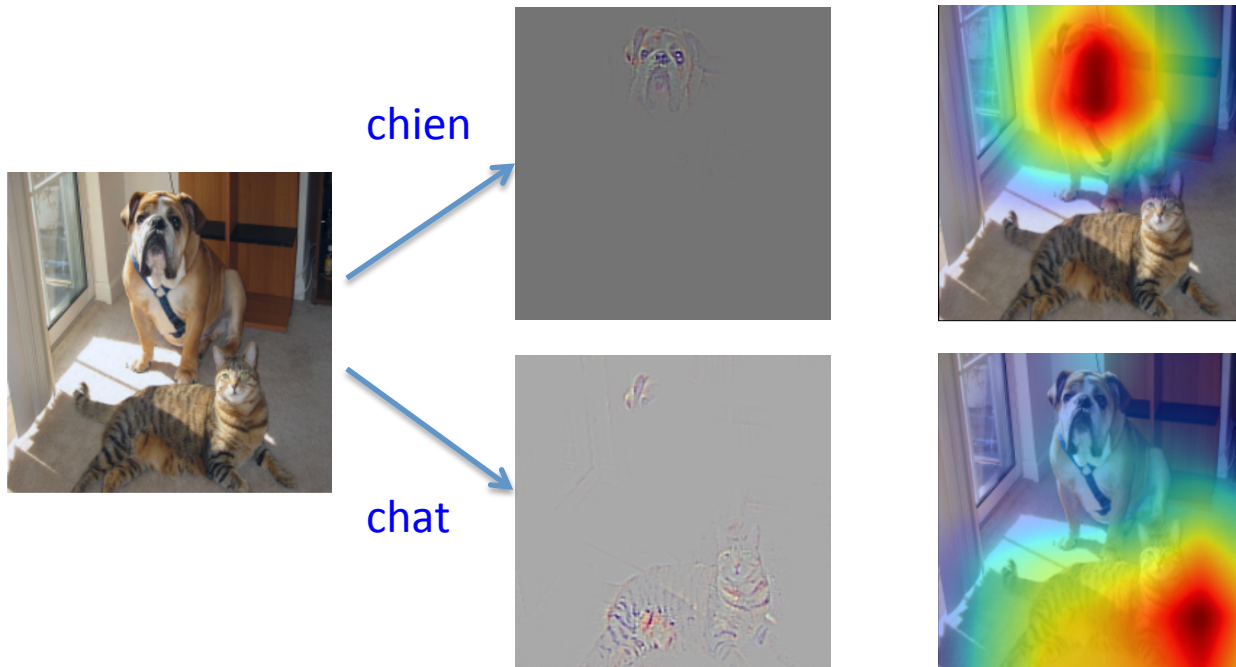
Is this less of a car
because the context is wrong?

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

Explication et réseaux de neurones profonds

Identification de classes d'objets dans une image

- Ici deux classes : « **chien** » et « **chat tigré** »



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


Explication et réseaux de neurones profonds

Protocole d'évaluation : comparaison entre explications


- À quel robot faites-vous le plus confiance ?

Both robots predicted: Person


What do you see?



Robot A based it's decision on



Robot B based it's decision on



Your options:

- Horse
- Person

Which robot is more reasonable?

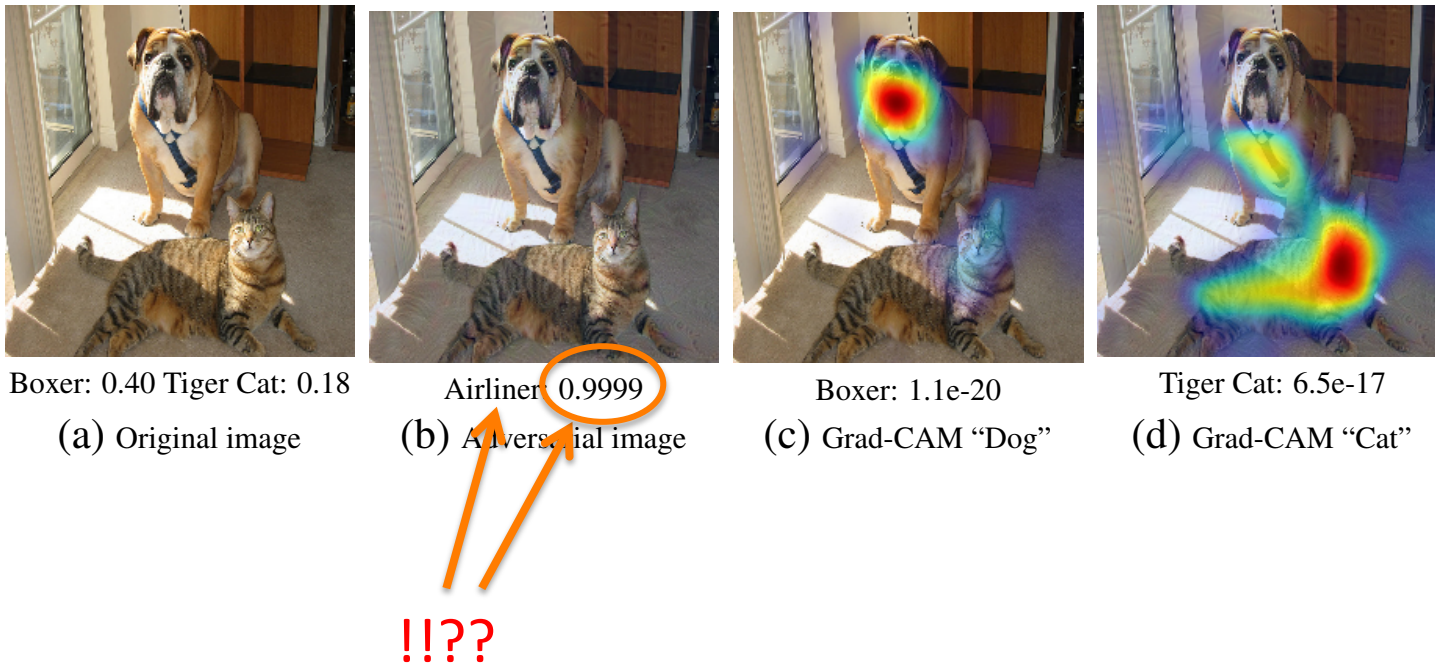
- Robot A** seems clearly more reasonable than **robot B**
- Robot A** seems slightly more reasonable than **robot B**
- Both robots seem equally reasonable
- Robot B** seems slightly more reasonable than **robot A**
- Robot B** seems clearly more reasonable than **robot A**

54 sujets sur Amazon Turk -> robot B évalué à 1.27 (entre -2 et +2)

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

Explication et réseaux de neurones profonds

Illusions d'optique : quelle explication ?



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

Détecter des biais



Ground-Truth: Nurse
(a) Original image



Predicted: Nurse
(b) Grad-CAM for biased model



Predicted: Nurse
(c) Grad-CAM for unbiased model



Ground-Truth: Doctor
(d) Original Image



Predicted: Nurse
(e) Grad-CAM for biased model



Predicted: Doctor
(f) Grad-CAM for unbiased model



Ground-Truth: Doctor
(g) Original Image



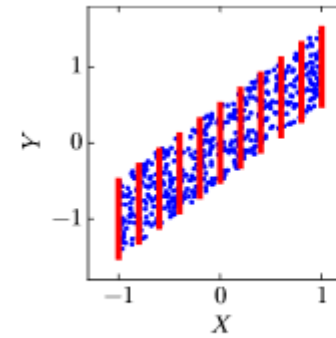
Predicted: Nurse
(h) Grad-CAM for biased model



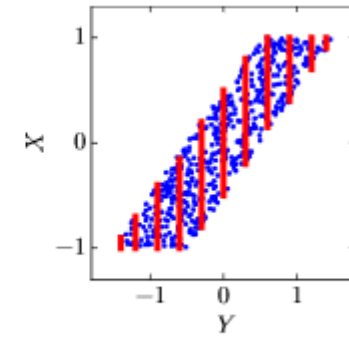
Predicted: Doctor
(i) Grad-CAM for unbiased model

L'identification de relations causales

- Dans les images



(a) ANM $X \rightarrow Y$.



(b) ANM $Y \rightarrow X$.



[David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Schölkopf, and Léon Bottou.
« *Discovering causal signals in images* ». *arXiv preprint arXiv :1605.08179*, 2016.]

Plan

1. L'IA est attendue partout
2. Quelles garanties sur l'induction
3. Un peu d'histoire
4. Et **demain** ? Prémises de changement de paradigme
5. Retour sur les **défis**
6. Conclusions ... et ouverture

Retour sur les défis

Est-ce que cela suffira ?

Le cas AlphaGo

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



Le cas AlphaGo : comprendre

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

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

Et même AlphaGo peut se tromper



Les explications des SEs

MYCIN : L'infection du malade a t-elle été contractée à l'hôpital ?

Utilisateur : pourquoi ?

TEIRESIAS : [pourquoi est-ce important de savoir cela ?]

Cela aidera à déterminer la catégorie à laquelle appartient l'organisme responsable de l'infection :

- il a déjà été établi que 1° l'organisme est à Gram négatif, 2° c'est un bâtonnet, 3° il se développe à l'air ou sans air,

- donc si l'infection par cet organisme n'a pas été contractée lors du séjour à l'hôpital alors il est presque certain (0,8) que l'organisme est de la catégorie des enterobacteriaceae; de plus, il se peut (0,1 ; règle 37) que ce soit la pseudomonias aeruginosa.

Utilisateur : pourquoi n'avez-vous pas envisagé la possibilité que l'organisme soit un streptocoque ?

TEIRESIAS : la règle 33 aurait du être utilisée pour identifier l'organisme comme un streptocoque, mais la seconde clause "la morphologie est de type coque" était déjà connue comme fausse et la règle 33 n'a jamais été essayée.

Les explications des SEs

- Quel est l'âge du patient ?
- ** Pourquoi
- Pour déterminer si on peut prescrire de la tétracycline
 - Si le patient a moins de 8 ans
 - Alors on ne peut pas prescrire de tétracycline
 - [Règle 122]
- ** Pourquoi ?
- ...

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

Les explications des SEs

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

Connaissances justificatives

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

→ **Noircissement** définitif des dents

→ Coloration socialement **indésirable**

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

Notion d'**effets secondaires** indésirables

Relations de **causalité**

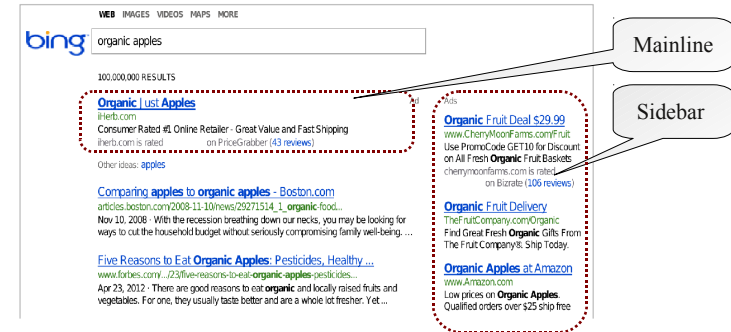
Système adaptatif de placement de publicité

- Deux sous-systèmes

- L'un plaçant les liens publicitaires
- L'autre choisissant les publicités

- Qui s'influencent mutuellement

- Chacun s'appuie sur les données de clicks
- Qui dépendent aussi de l'intervention de l'autre systèmes
- Et d'autres facteurs non contrôlés (prix, requête de l'utilisateur, ...)



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

Le paradoxe de Simpson

- Les médecins voudraient savoir si le **traitement A** est plus performant ou moins performant que le **traitement B**
- Deux **groupes** de **350 patients** sont sélectionnés, l'un recevant le **traitement A**, l'autre le **traitement B**

	Overall
Treatment A: Open surgery	78% (273/350)
Treatment B: Percutaneous nephrolithotomy	83% (289/350)

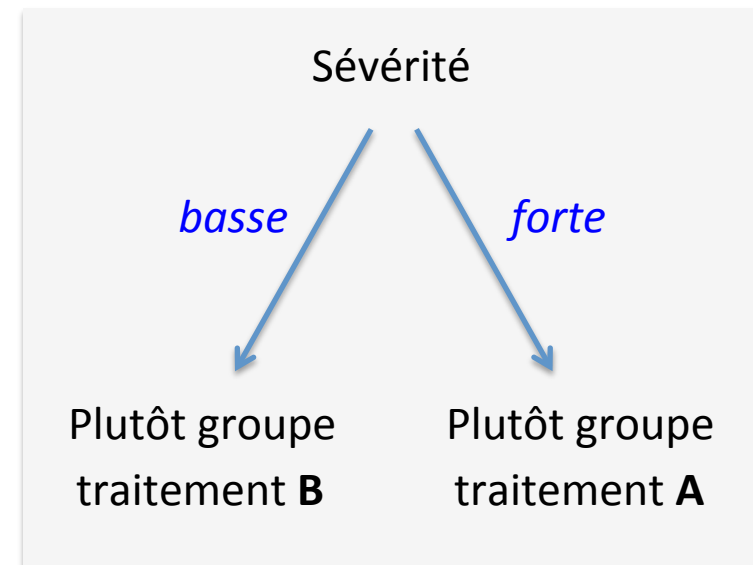
B est meilleur ?

Le paradoxe de Simpson

	Overall	Patients with small stones	Patients with large stones
Treatment A: Open surgery	78% (273/350)	93% (81/87)	73% (192/263)
Treatment B: Percutaneous nephrolithotomy	83% (289/350)	87% (234/270)	69% (55/80)

- Variable **influençante**

Le choix des patients entrant dans les deux groupes dépendait de la sévérité de la pathologie



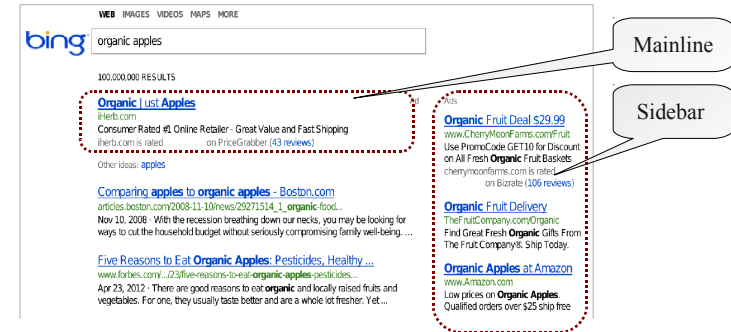
Système adaptatif de placement de publicité

- Deux sous-systèmes

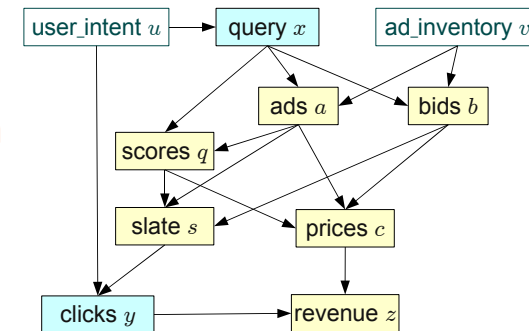
- L'un plaçant les liens publicitaires
- L'autre choisissant les publicités

- Qui s'influencent mutuellement

- Chacun s'appuie sur les données de clicks
- Qui dépendent aussi de l'intervention de l'autre systèmes
- Et d'autres facteurs non contrôlés (prix, requête de l'utilisateur, ...)



Importance de l'identification
du graphe causal



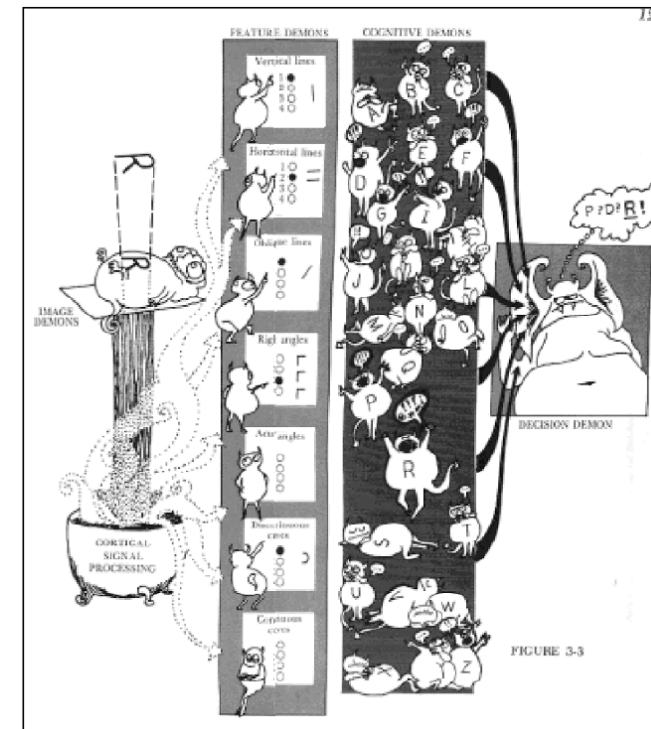
[L. Bottou et al. «Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising », JMLR, 14, (2013), 3207-3260]

Dartmouth Summer School on Artificial Intelligence (1956)

- Oliver Selfridge : « Pandemonium: A Paradigm for Learning ».

Une architecture hiérarchique de « démons » pour résoudre des problèmes + la suggestion d'un mécanisme d'apprentissage

- L'architecture « blackboard » (1975)



Plan

1. L'IA est attendue partout
2. Quelles garanties sur l'induction
3. Un peu d'histoire
4. Et demain ? Prémises de changement de paradigme
5. Retour sur les défis
6. Conclusions ... et ouverture

Pour généraliser l'usage de systèmes apprenants et adaptatifs

1. Savoir **expliquer** et **justifier** des décisions pour chaque cas particulier
2. **Sous-systèmes en interaction** ni chaotiques, ni s'égarant

- **Aller au-delà** de l'apprentissage statistique
 - Les **explications** ... entre nouveauté et revisite ?
 - Sous-systèmes **en interaction**
 - Graphes de dépendances causales et conditions d'indépendance conditionnelle
 - Contrefactuelles
 - La **causalité** : fondamental

Question

- **Quels critères de performance ?**

Un pari

Aller vers des systèmes **capables d'enseigner**

1. **Expliquer** un cas
 2. **Synthétiser**
 3. Organiser un **curriculum**
- Vers une **évaluation** des systèmes **par la performance de leurs élèves ?**