

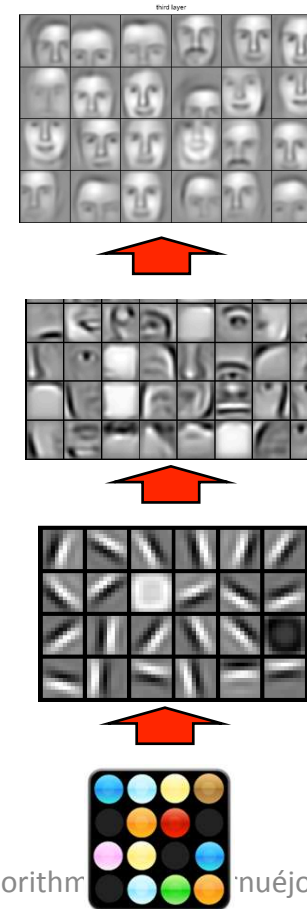
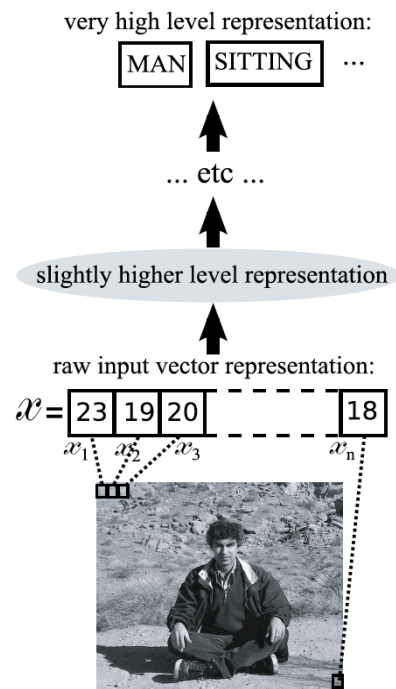
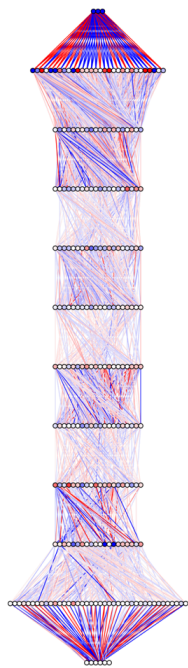
« Deep learning »
as THE universal solution (2006– ...)

"The paper focuses on a subject that might be of limited importance at ICML, *given the current trend towards neural networks.*"

« Deep Neural Networks »

Artificial Neural Networks

- With numerous hidden layers (possibly > 100s)
- And a **very large number of parameters** ($\sim 10^7 - 10^8$ parameters)
- Learn **hierarchical** and **compositional representations**



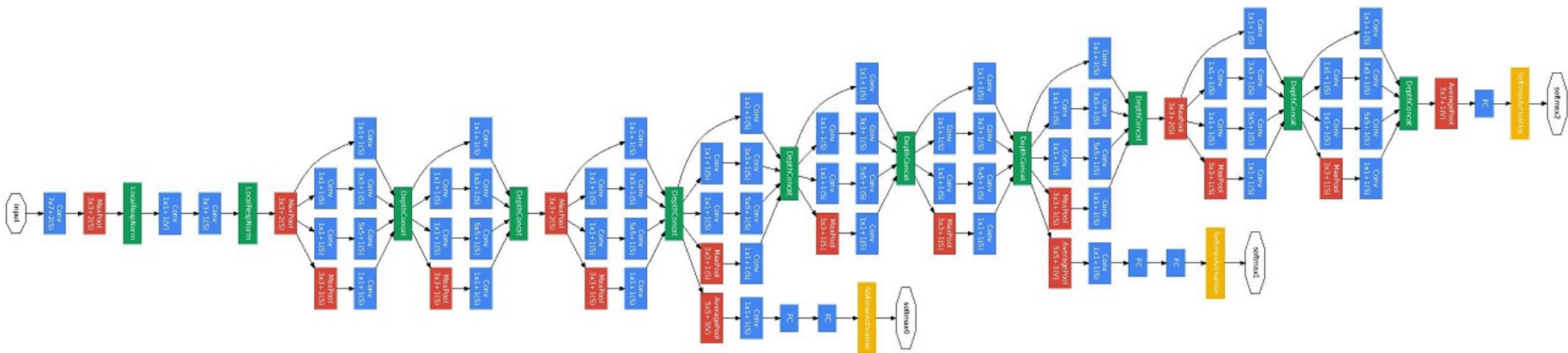
object models

object parts
(combination
of edges)

edges

GoogleNet

- A **mécano** of neural networks



BUT ... does deep learning
bring big trouble (for the theory of induction)?

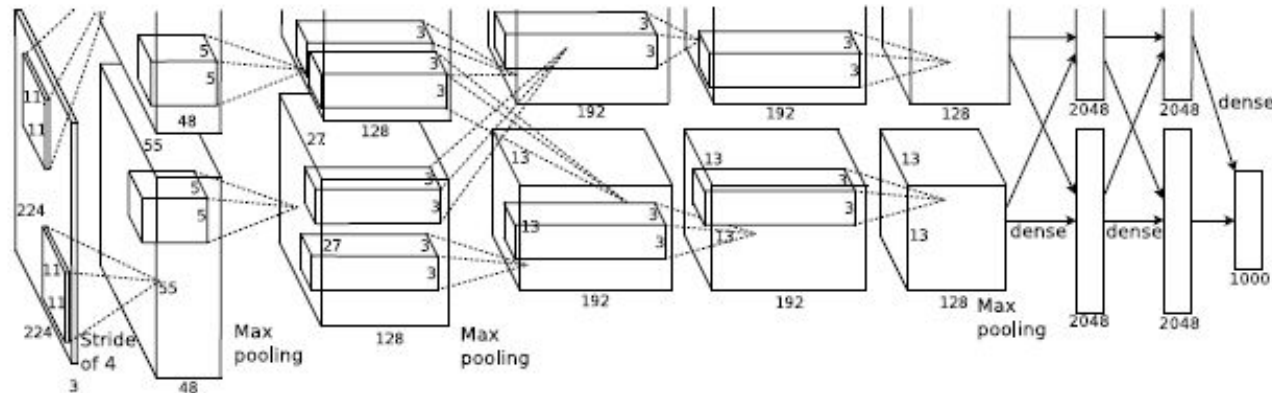
Troubling findings

A paper

- 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

Troubling findings

Experiments

1. Original dataset without modification

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

Troubling findings

Experiments

1. Original dataset without modification

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

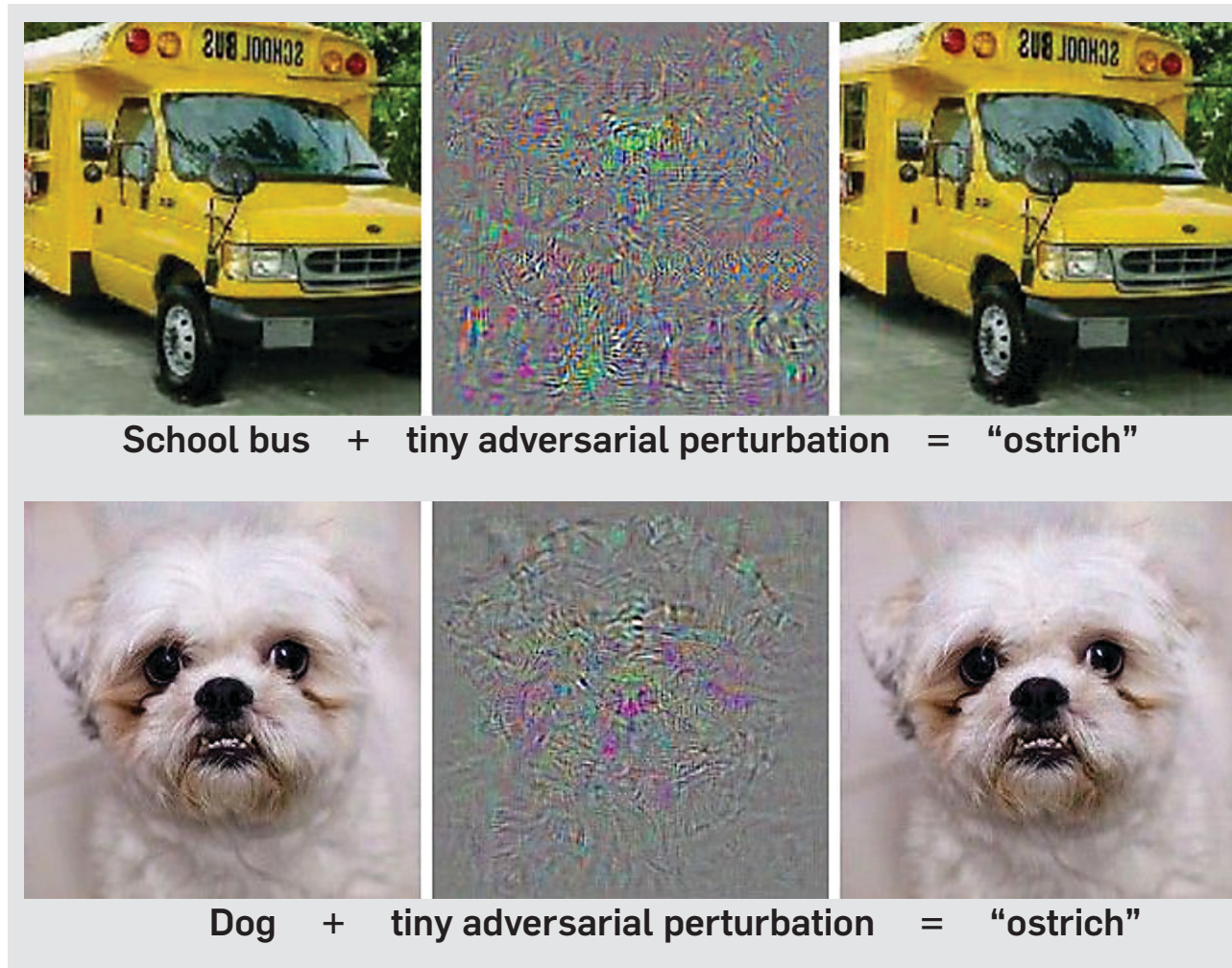
Can grow without limit!!

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

But then,

why are deep NNs so good on image classification tasks?

Adversarial learning



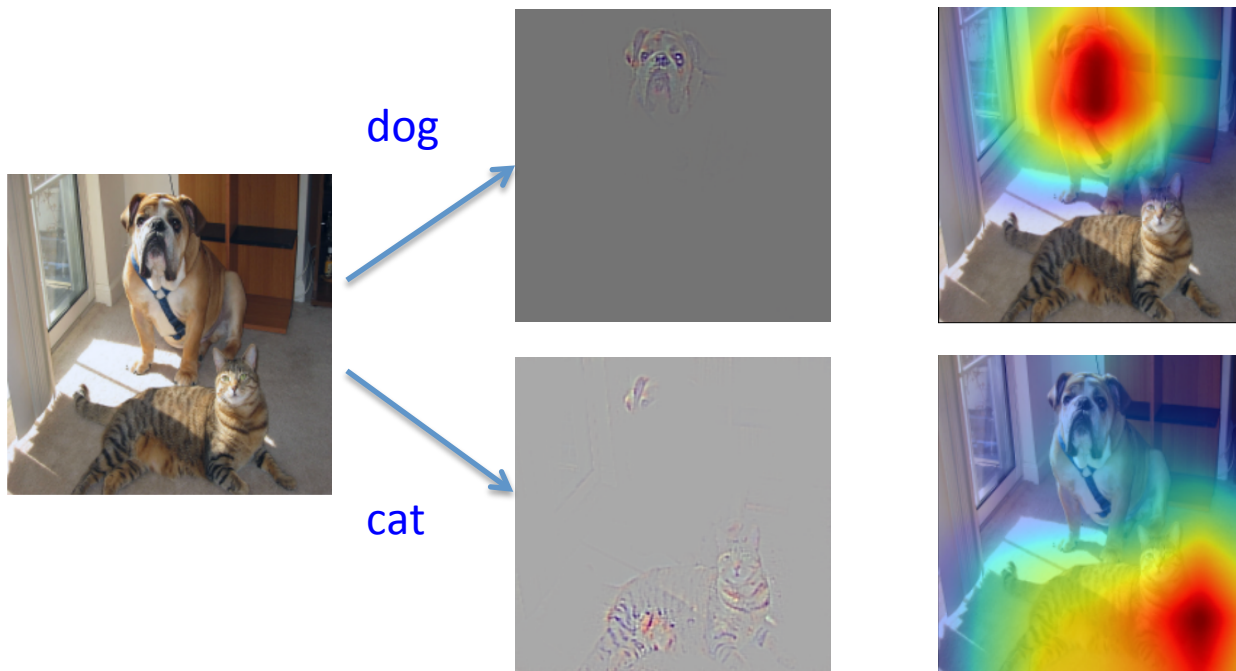
Adversarial input can fool a machine-learning algorithm into misperceiving images.

Illustration

Explanations and deep neural networks

Identification object categories in an image

- Here, two classes : « **dog** » and « **tiger cat** »



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


Explanations and deep neural networks

Evaluation protocol: comparison between explanations


- Which robot do you trust most?

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?

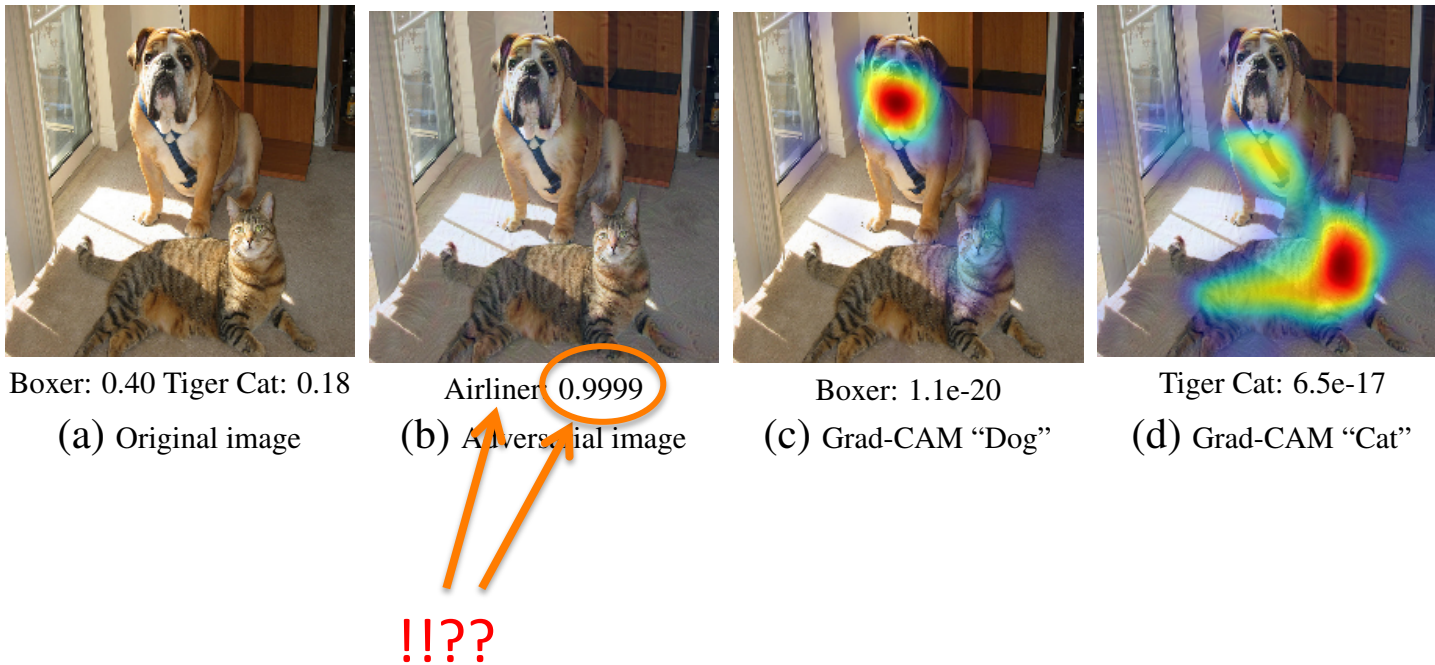
- Robot A** seems clearly more reasonable than **robot B**
- Robot A** seems slightly more reasonable than **robot B**
- Both robots seem equally reasonable
- Robot B** seems slightly more reasonable than **robot A**
- Robot B** seems clearly more reasonable than **robot A**

54 subjects on Amazon Turk -> robot B evaluated 1.27 (between -2 et +2)

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

Explanations and deep neural networks

Optical illusions: how to explain them?



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

Car in a swimming pool

- ... or **no car** ...?



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

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

Assessment

1. A theory

- for **stationary** environments and **i.i.d. data and queries!!**
- focused on the expectation of the **cost of errors**
 - Prior knowledge must be encoded in the cost
- that **can produce learning algorithms** when combined with optimization techniques

2. Deep NNs

- **depart** from this framework
 - Demand at least a **reworking of the theory**
 - **Prior knowledge** encoded in the architecture
- **Still**
 - require **enormous amount of data**
 - Focused on **error rates**
 - Based on **correlations**

Outline

1. What does work
2. Limitations
3. Learning comes with **which guarantees?**
 - Induction: how to win this game?
 - The statistical learning theory
 - A **closed case?** Not so sure
4. Other paradigms? An **historical perspective**
5. Is there a **paradigmatic change in sight?**
6. Conclusions

Are there other paradigms?

An historical perspective on ML

Learning ...

... as

a means to **improve the efficiency** of a **problem solver**

E.g. The PRODIGY system

ACM SIGART Bulletin, 1991, vol. 2, no 4, p. 51-55

PRODIGY: An Integrated Architecture for Planning and Learning

Jaime Carbonell, Oren Etzioni*, Yolanda Gil, Robert Joseph
Craig Knoblock, Steve Minton†, and Manuela Veloso

PRODIGY's basic reasoning engine is a general-purpose problem solver and planner [10] that searches for sequences of operators (i.e., plans) to accomplish a set of goals from a specified initial state description. Search in PRODIGY is guided by a set of control rules that apply at each decision point.

PRODIGY's reliance on explicit control rules, which can be learned for specific domains, distinguishes it from most domain independent problem solvers. Instead of using a least-commitment search strategy, for example, PRODIGY expects that any important decisions will be guided by the presence of appropriate control knowledge. If no control rules are relevant to a decision, then PRODIGY makes a quick, arbitrary choice. If in fact the wrong choice is made, and costly backtracking proves necessary, an attempt will be made to learn the control knowledge that must be missing.

Illustration: LEX (Tom Mitchell)

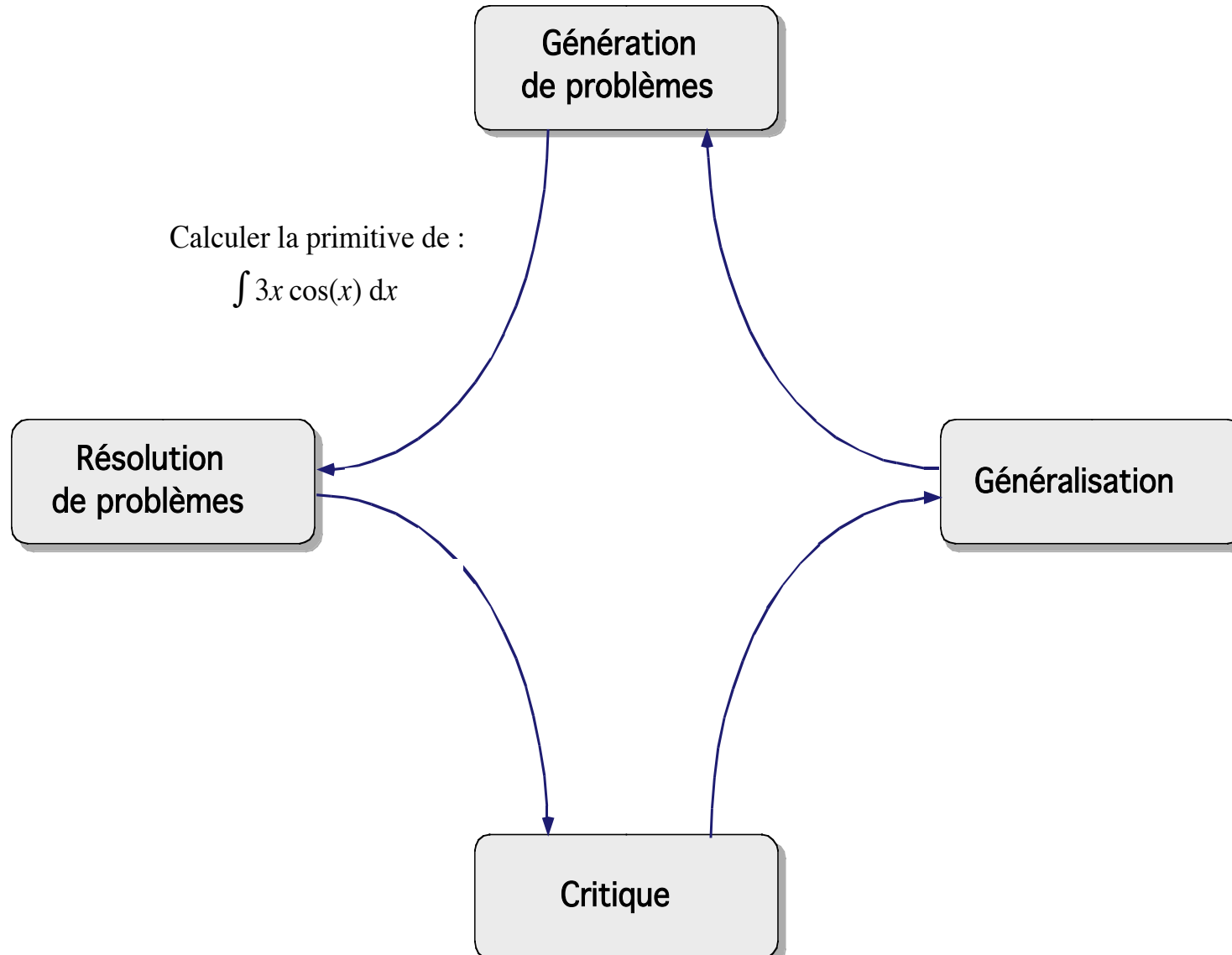


Illustration: LEX (Tom Mitchell)

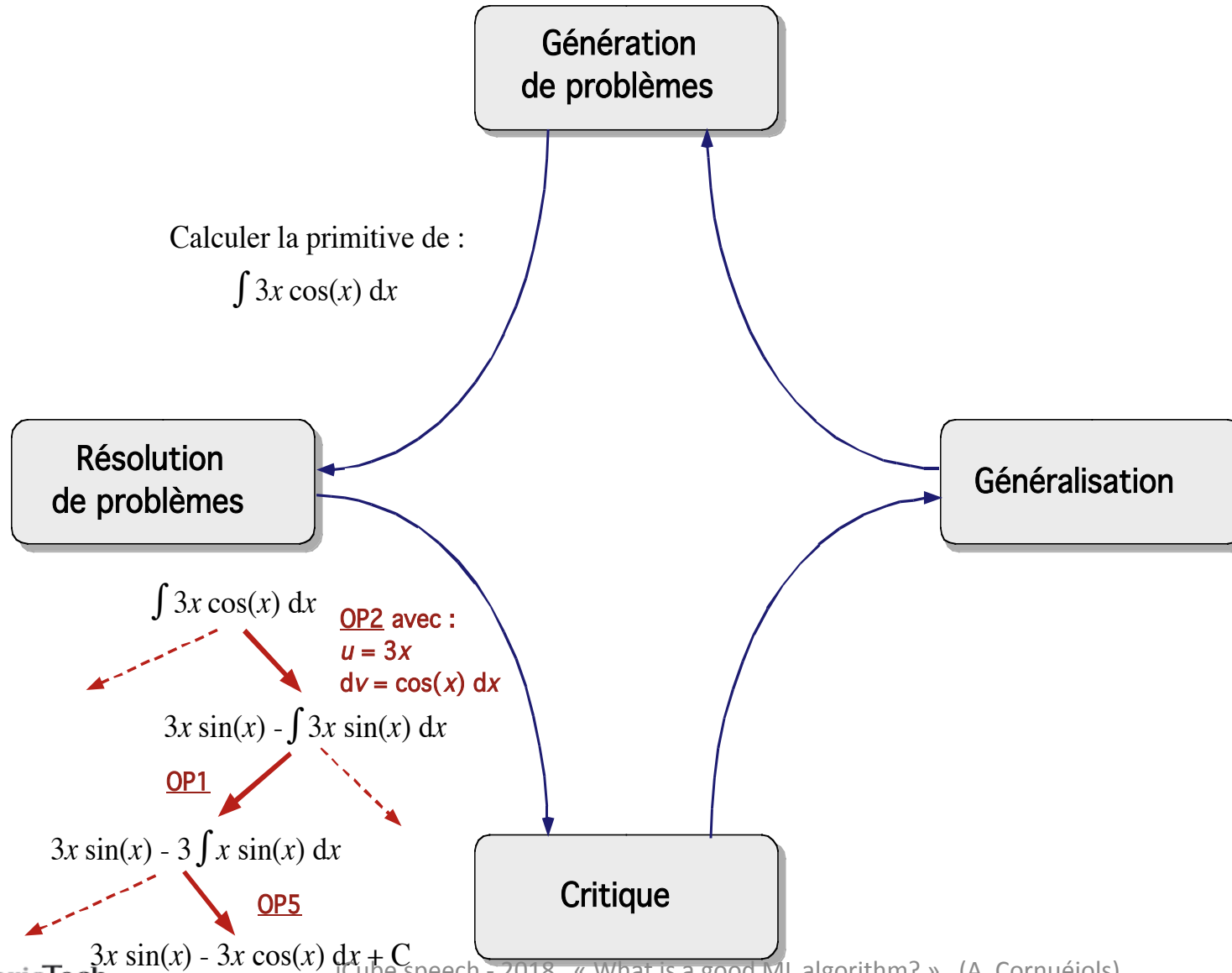


Illustration: LEX (Tom Mitchell)

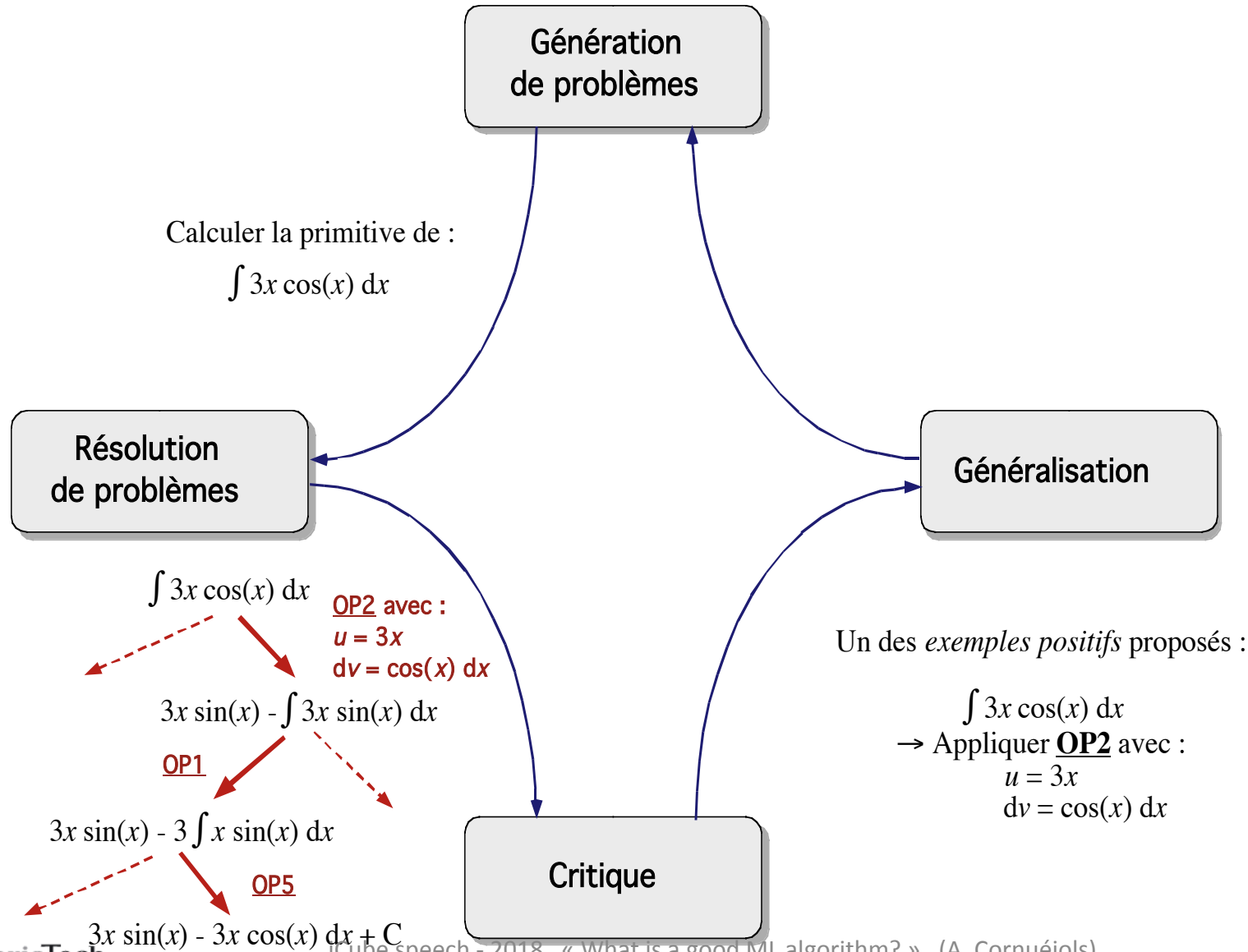
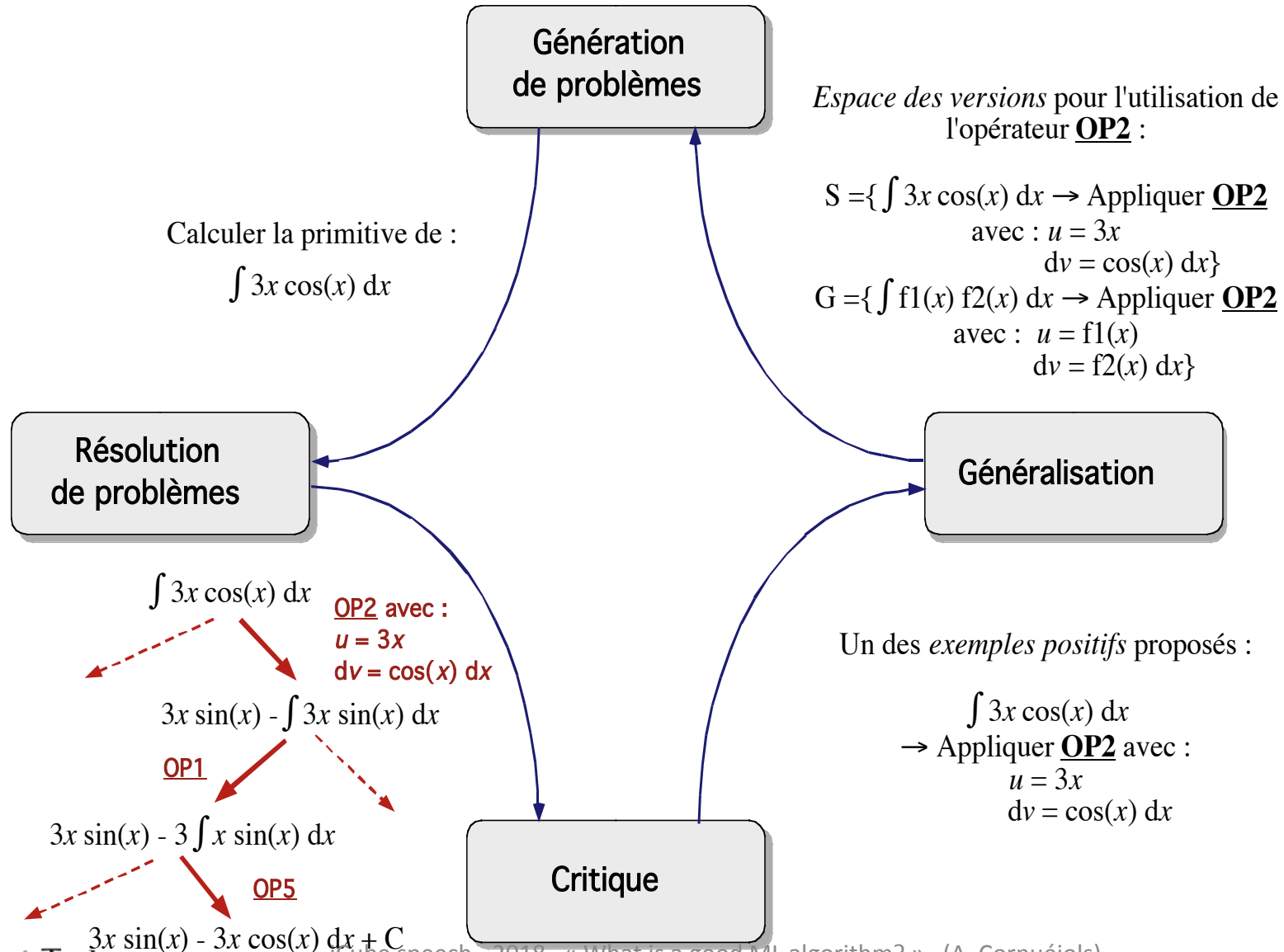


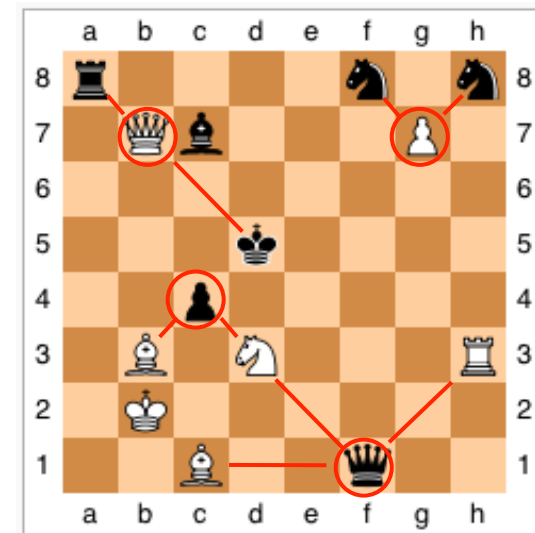
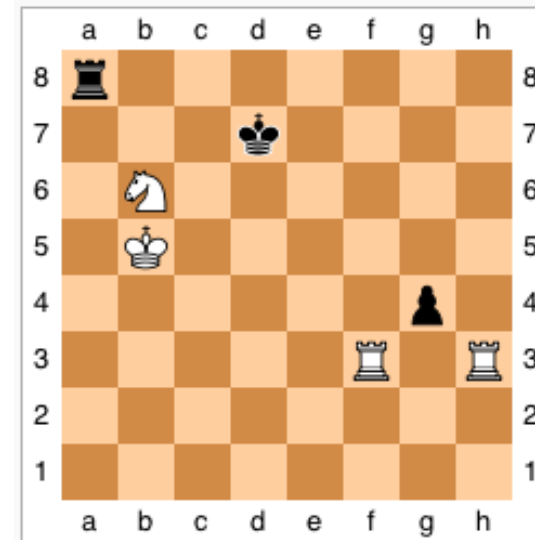
Illustration: LEX (Tom Mitchell)



Learning from one example

Explanation-Based Learning

1. From a **single example**
2. Try to prove the “fork”
3. Generalize



Explanation-Based Learning

Ex : **learn the concept** `stackable(Object1, Object2)`

- **Domain theory :**

```
(T1) : weight(X, W) :- volume(X, V), density(X, D), W is V*D.
```

```
(T2) : weight(X, 50) :- is_a(X, table).
```

```
(T3) : lighter_than(X, Y) :- weight(X, W1), weight(X, W2), W1 < W2.
```

- **Operationality constraint:**

- Concept should be expressible using *volume, density, color, ...*

- **Positive example (solution) :**

```
on(obj1, obj2).
```

```
is_a(object1, box).
```

```
is_a(object2, table).
```

```
color(object1, red).
```

```
color(object2, blue).
```

```
made_of(object2, wood).
```

```
volume(object1, 1).
```

```
volume(object2, 0.1).
```

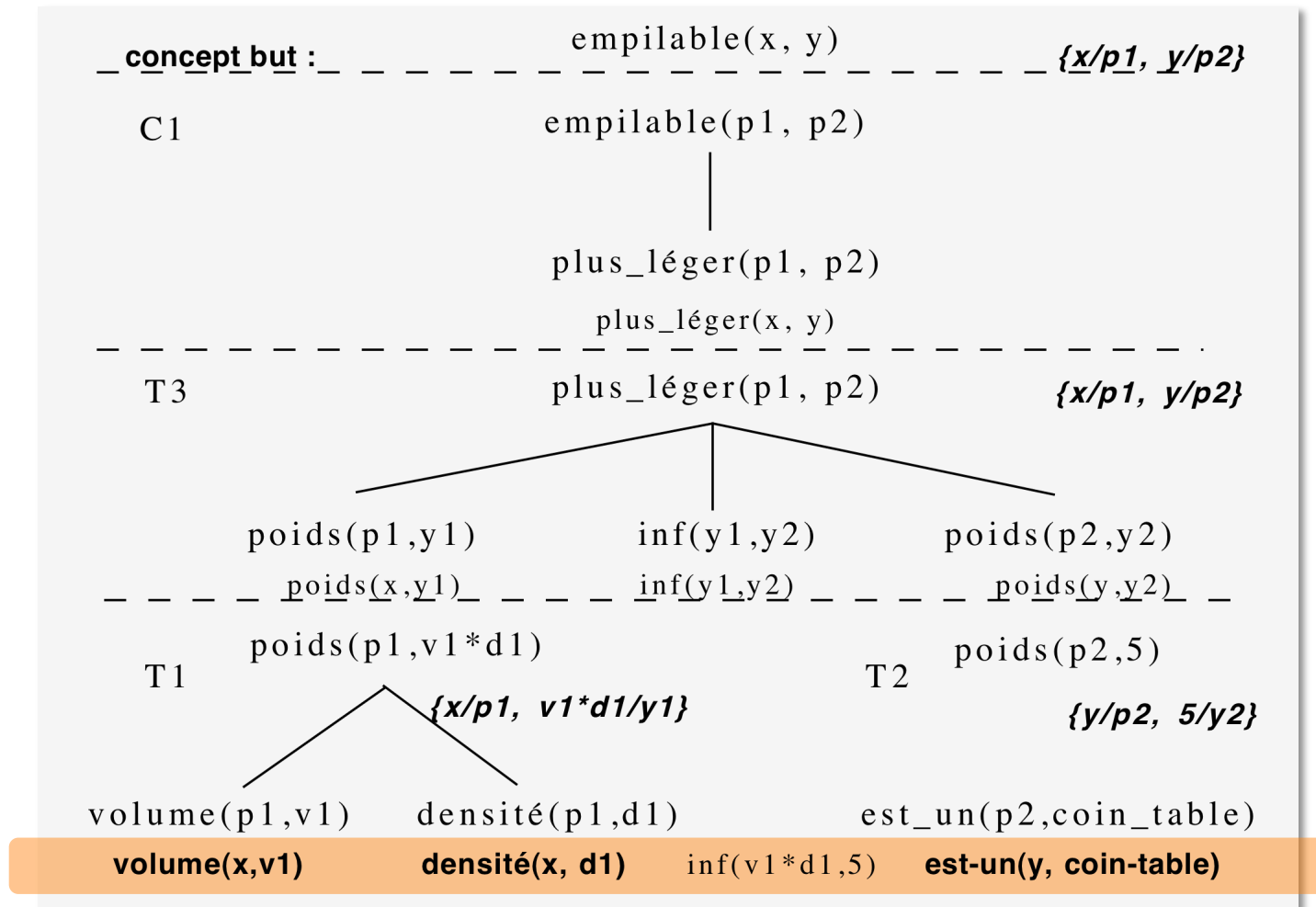
```
owner(object1, frederic).
```

```
density(object1, 0.3).
```

```
Made_of(object1, cardboard).
```

```
owner(object2, marc).
```

Explanation-Based Learning



Generalized search tree resulting from **regression of the target concept in the proof tree** by computing at each step the most general literals allowing this step.

Explanation-Based Learning

- Induction **from a single example**
 - ... plus a strong domain theory
- Based on
 - **Logic-based** knowledge representation
 - **Reasoning Operators** (deduction, goal regression in a proof tree, ...)

Now used in SAT “solvers”

Explanation-Based Learning

- What was the **aim** of learning?
- What was a **good theory/ method** of learning ?

Explanation-Based Learning

- What was the **aim** of learning?
- What was a **good method** of learning ?

1. Method **improving** the **problem solving performances**

- [Steve Minton (1990) « *Quantitative results concerning the **utility** of Explanation-Based Learning* »]

2. Method that **simulates** the performances (and limits) of a **natural cognitive agent** (human or animal)

- [Laird, Rosenbloom, Newell (1986) « *Chunking in SOAR: The anatomy of a general learning mechanism* »]
- [Anderson (1993) « *Rules of the mind* » ;
Taatgen (2003) « *Learning rules and productions* »]

Learning and reasoning

Papers like

- Stephen José Hanson (1990). **Conceptual clustering and categorization: bridging the gap between induction and causal models.**

Machine Learning journal, 1990, pp.235-268.

But

No measure of generalization
performance **independent of**
the problem-solver

Difficulties to scale up and to face noisy data

... when data started to pour down

Outline

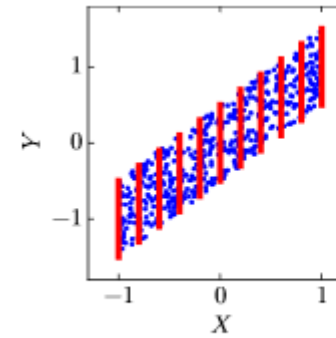
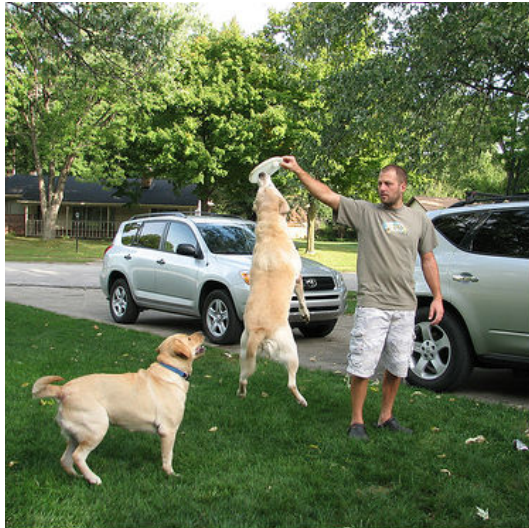
1. What **does work**
2. **Limitations**
3. Learning comes with **which guarantees?**
 - Induction: how to win this game?
 - The statistical learning theory
 - A **closed case?** Not so sure
4. Other paradigms? An **historical perspective**
5. Is there a **paradigmatic change in sight?**
6. **Conclusions**

New learning scenarios

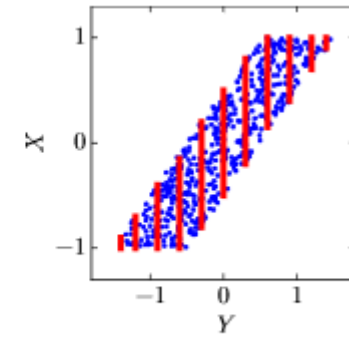
=> Is there a paradigmatic change in sight?

Identification of causal relationships

- In images
- With unsupervised learning!!



(a) ANM $X \rightarrow Y$.



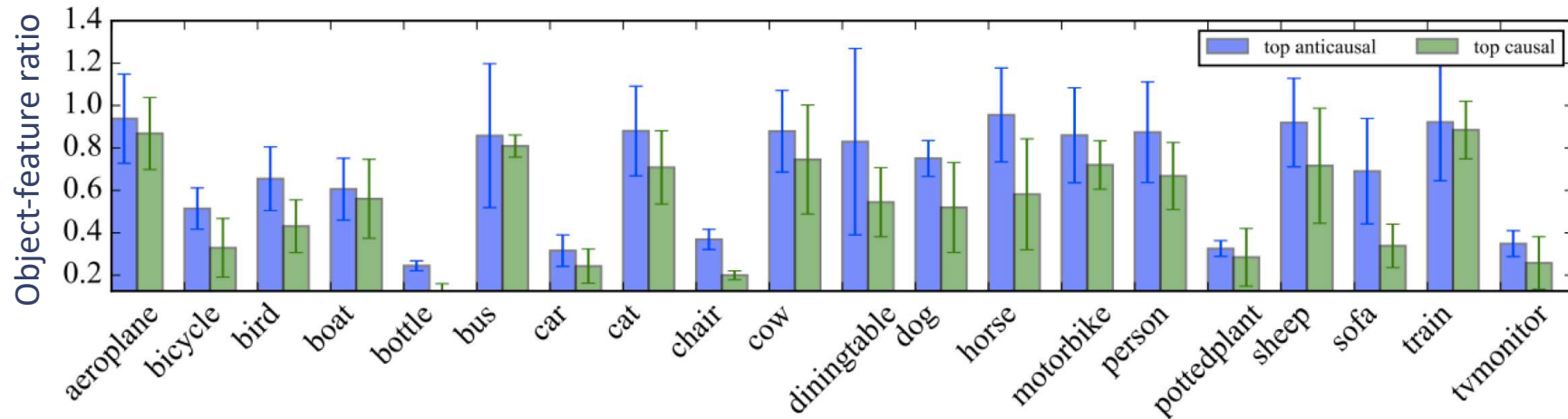
(b) ANM $Y \rightarrow X$



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

Identification of causal relationships

- Results



- Top anticausal features have higher object scores for all twenty categories.
- The probability that this happens for all 20 classes out of chance is $2^{-20} \approx 10^{-6}$.

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

The AlphaGo case

- Plays like an « alien »
- Out of the blue moves
- A revolution in the theory of the game
- Excitement in the go schools



The AlphaGo case: understanding

Fan Hui, Gu Li, Zhou Ruyang (very strong Go players) turn to the activity of analyzing the games played by AlphaGo

- Kind on of exegesis. Explanations a posteriori
- Necessary for
 - communication
 - teaching

And even AlphaGo might err



Explanations in Expert Systems

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.

Explanations in Expert Systems

- 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 ?

Explanations in Expert Systems

Why should we not prescribe tetracycline to a child under the age of 8?

Expert justifications

Drug depot on developing bones

→ Definitive **blackening** of the teeth

→ **Socially unwanted** coloration

→ **Do not administer** tetracycline to children under the age of

Notion of undesirable **side effects**

Causality relationships

Transfer learning

Definition [Pan, TL-IJCAI'13 tutorial]

- Ability of a system to **recognize** and **apply** knowledge and skills learned in **previous domains/tasks** to **novel domains/tasks**

Example

- We have **labeled images** (person / no person) from a **web corpus**
- Novel task: **is there a person** in unlabeled images from a **video corpus**?

Person no Person ? Is there a Person?

Web corpus Video corpus

Transfert learning: questions

- What can be **the basis** of transfer learning?

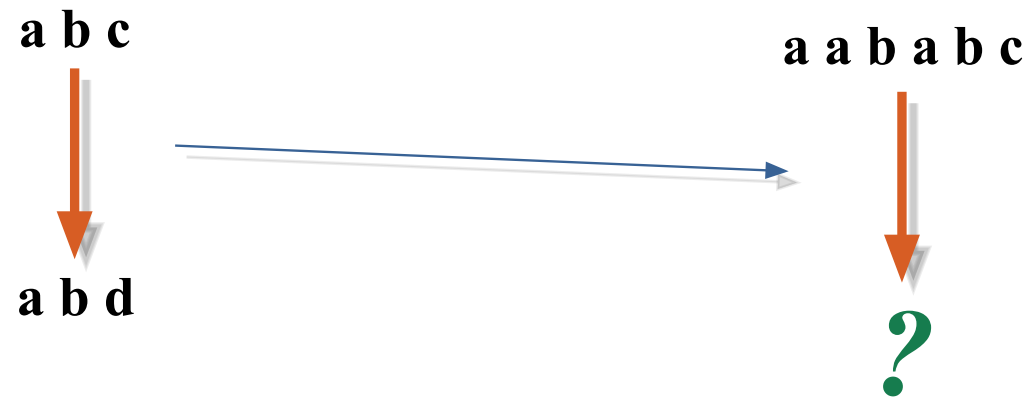
How to translate formally :

“the target domain is like the source domain”?

Not i.i.d.
anymore

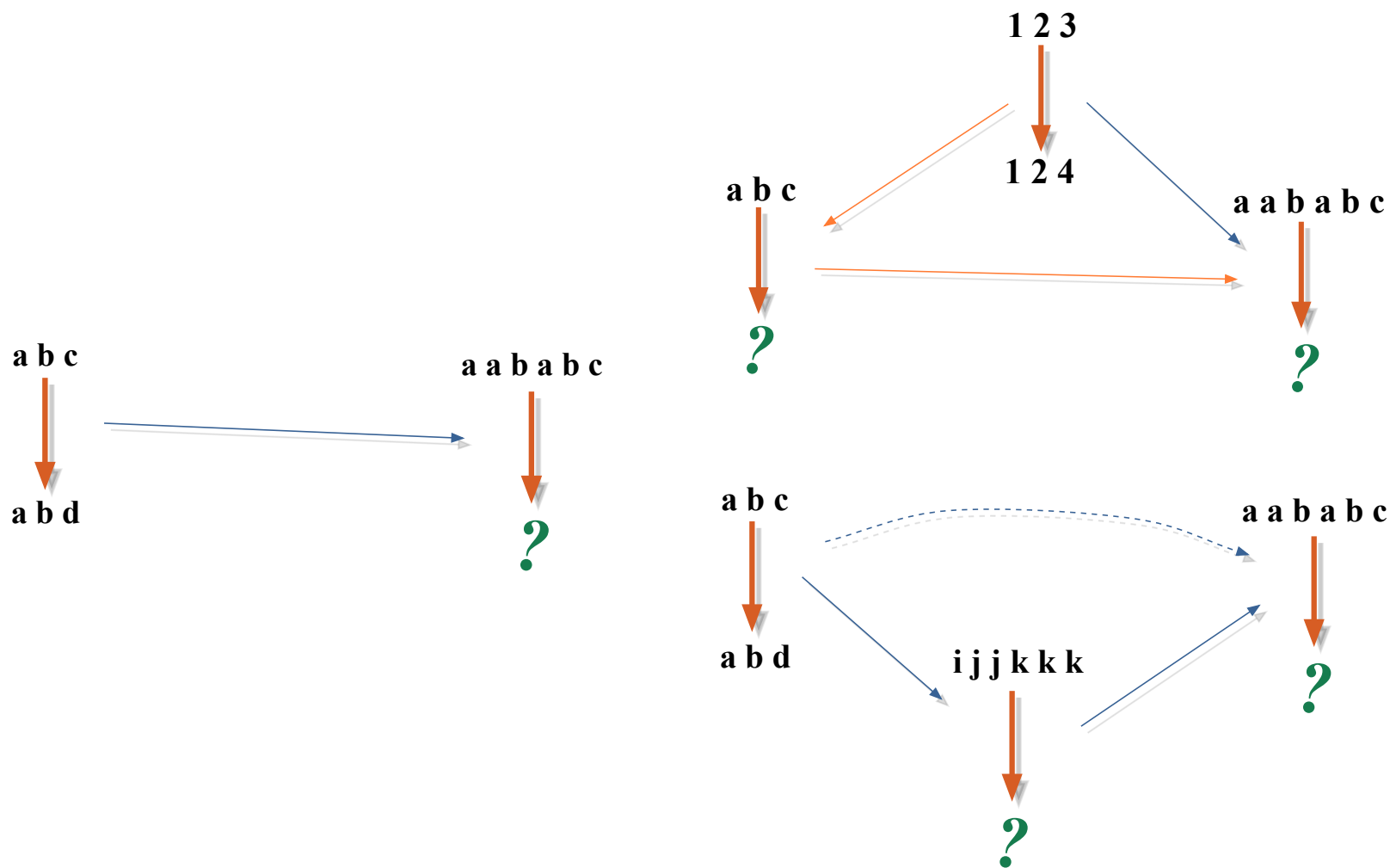
- What **determine a good transfer**?
 - A “good source”?
 - A high “similarity” between source and target?
- What **formal guarantees** can we have on the transferred hypothesis?

Transfer and analogy



Why should 'a a b a b c d' be any better than 'a b d'?

Transfer and sequence effects



- t

Long-life learning

- Learning organized in a **sequence of tasks**
 - **Very far from the i.i.d. scenario**
- Learning will be affected by the **history of the system**
- We **need a theory of the dynamics of learning**
 1. Which **sequence effects** can we expect?
 2. How to **best organize the curriculum** of a learning system?

Conclusions

The current situation

- Inductive learning **needs biases**
 - No objective bias-free results
- The **theory**
 - Is focused entirely on the **error rate**
 - Assumes stationary environment and random inputs (**i.i.d.**)
 - **Requires large enough data sets** w.r.t. to the capacity of \mathcal{H}
- We **do not understand** well deep neural networks
- Correlations **≠** structures, semantics, causation

We start to pay attention to **new demands**

1. The need for **explanations**

- Structures
- **Causal** reasoning
- No more only error rate

We start to pay attention to **new demands**

1. The need for **explanations**

- Structures
- **Causal** reasoning
- No more only error rate

2. The need for **transfer learning**

- **What** should be transferred?
- **Conditions** for positive / negative transfer?

We start to pay attention to **new demands**

1. The need for **explanations**

- Structures
- **Causal** reasoning
- No more only error rate

2. The need for **transfer learning**

- **What** should be transferred?
- **Conditions** for positive / negative transfer?

3. Scenarios **away from the i.i.d. assumption**

- Online learning / **changing environments**
- **Curriculum** learning
- Long-life learning

Conclusions: “new” scenarios

- **Limited data sources**
 - We often learn from (very) few examples
- The past **history of learning** affects learning: **Education**
 - Sequence effects
- We learn in order to **build “theories”**
 - All the time: small and large theories

For instance, what would you like to ask?

A bet

Towards systems **that know how to teach**

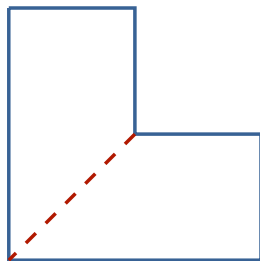
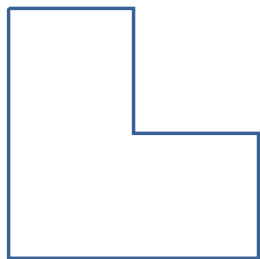
1. **Explain** a case
 2. **Synthesizing**
 3. Organize a **curriculum**
- **Evaluating the systems by the performance of their pupils?**

Suppléments

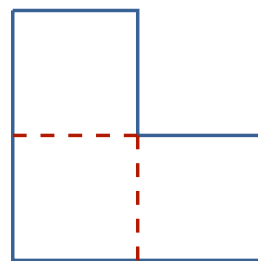
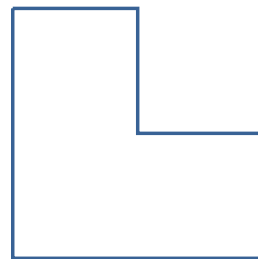
Sequence effects

Instructions: cut the following geometrical figure into n parts that **can be superposed**

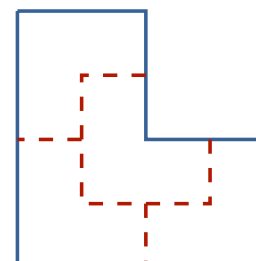
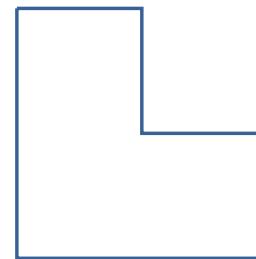
Into 2:



Into 3:



Into 4:



Into 5:

