Course

Learning Theory and Advanced Machine Learning

Antoine Cornuéjols

AgroParisTech – INRAE UMR MIA Paris-Saclay

antoine.cornuejols@agroparistech.fr

The course

- Documents
  - The book

    "L'apprentissage artificiel. Concepts et algorithmes. De Bayes et Hume au Deep Learning"

    V. Barra & A. Cornuéjols & L. Miclet

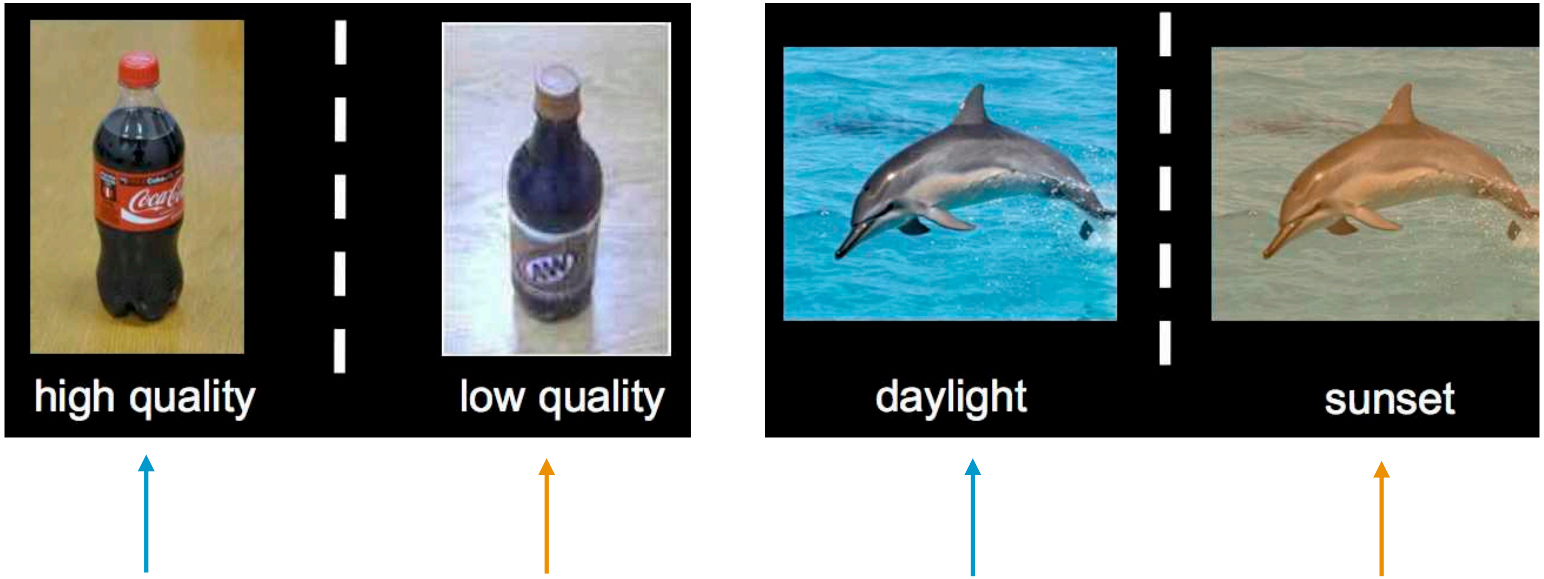
    Eyrolles. 4th éd. 2021

  - The slides + information on:

    https://antoinecornuejols.github.io/teaching/Master-AIC/M2-AIC-advanced-ML.html

The focus of the course

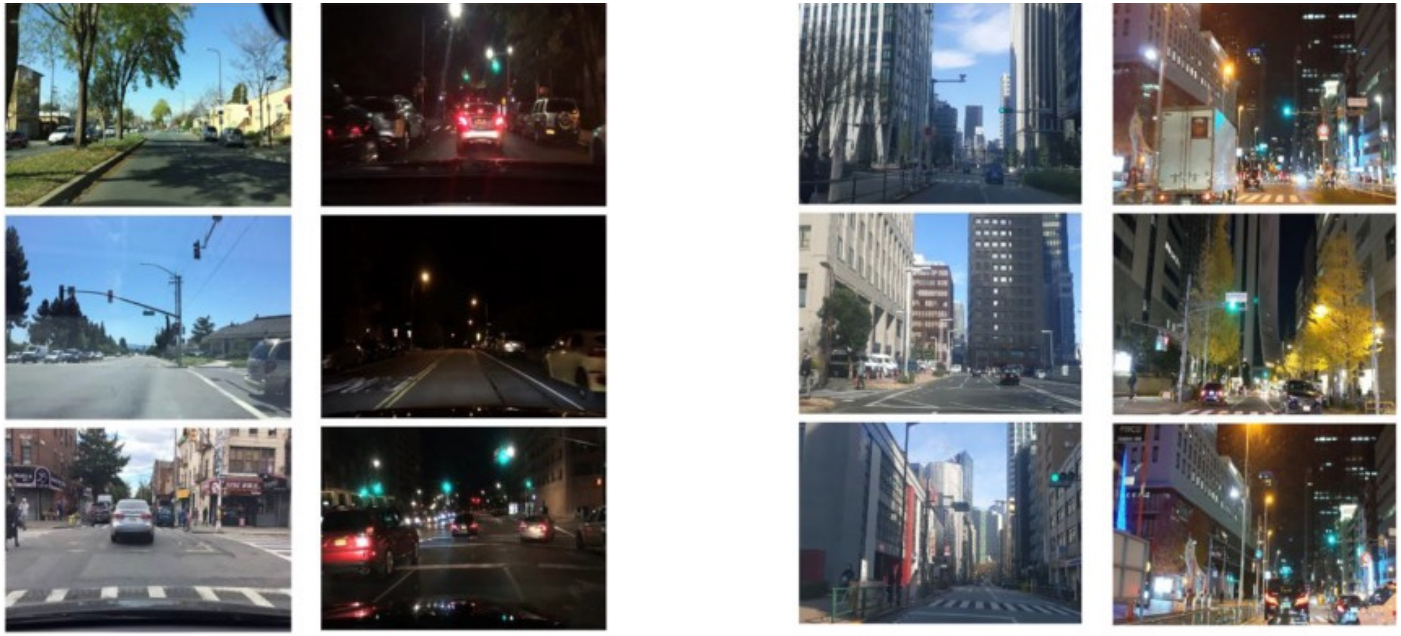
- Out-Of Distribution learning (OOD)
  - **Change of distribution** between learning and testing



Learning Testing Learning Testing

The focus of the course

- Out-Of Distribution learning (OOD)
  - **Change of distribution** between learning and testing



BDD: Daytime Images

BDD: Nighttime Images

Tokyo: Daytime Images

Tokyo: Nighttime Images

Learning

Testing

Learning

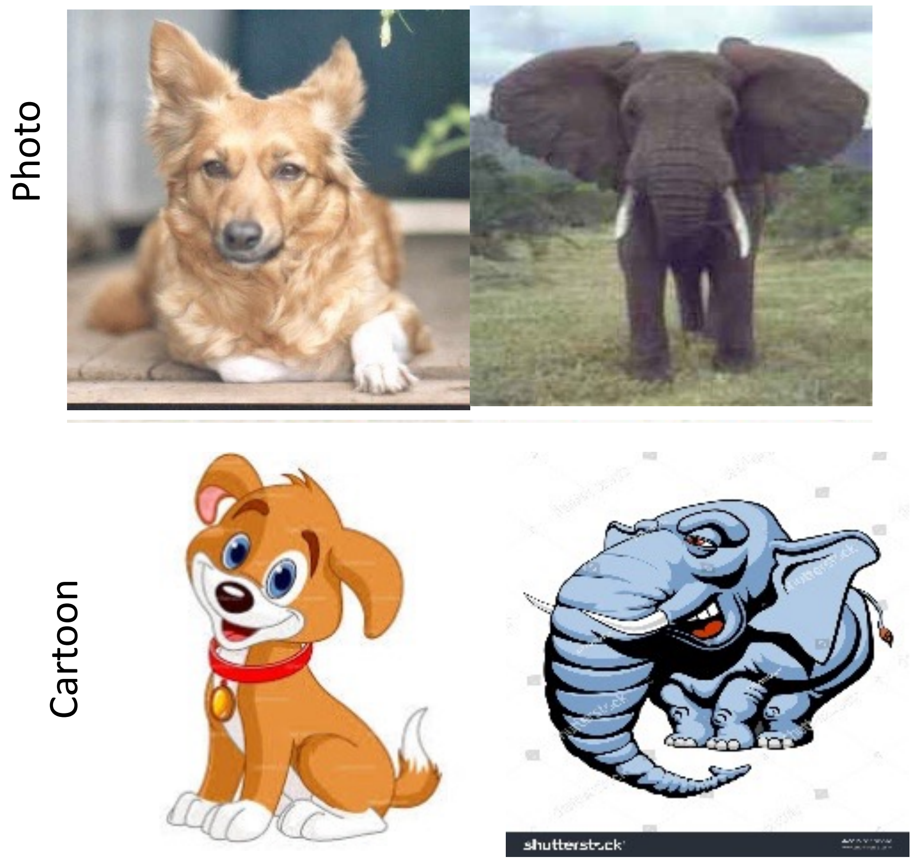
Testing

The focus of the course

- Out-Of Distribution learning (OOD)
  - Change of **domain** between learning and testing: **Transfer** Learning

Learning →

Testing →



The focus of the course

- Out-Of Distribution learning (OOD)
  - Change of **domain** between learning and testing: Transfer Learning

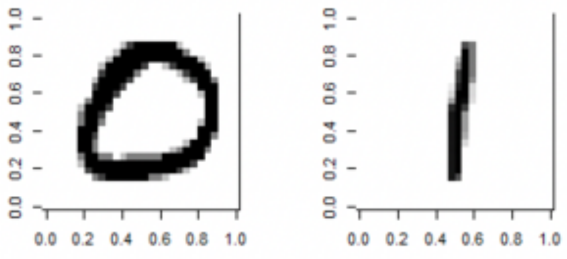
Learning →

Testing →

The focus of the course

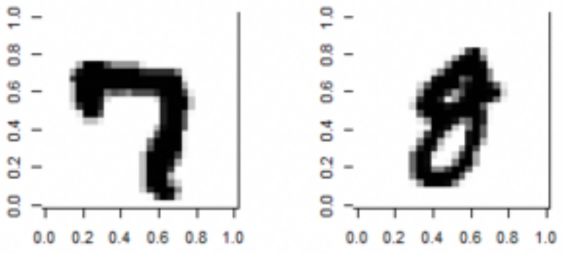
- Out-Of Distribution learning (OOD)
  - Change of **domain** between learning and testing: **Transfer** Learning

Learning →



(a) Is it a zero or a one ?

Testing →

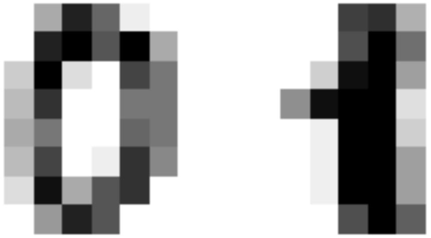


(b) Is it an eight or a seven ?

The focus of the course

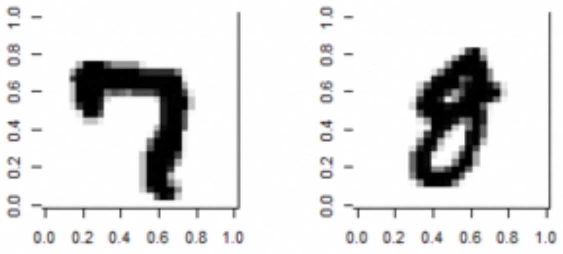
- Out-Of Distribution learning (OOD)
  - Change of **domain** between learning and testing: **Transfer** Learning

Learning →



(a) Is it a zero or a one ?

Testing →

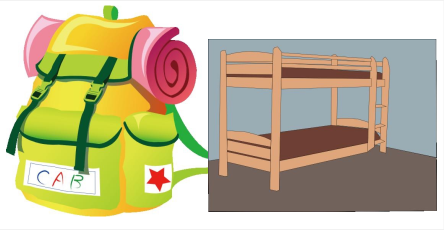


(b) Is it an eight or a seven ?

The focus of the course

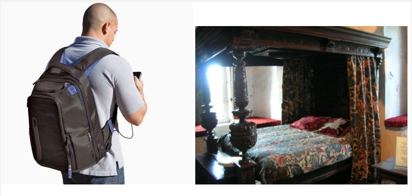
- Out-Of Distribution learning (OOD)
  - Change of **tasks**: Long Life Learning

Learning and testing (1) →



Learning and testing (2) →

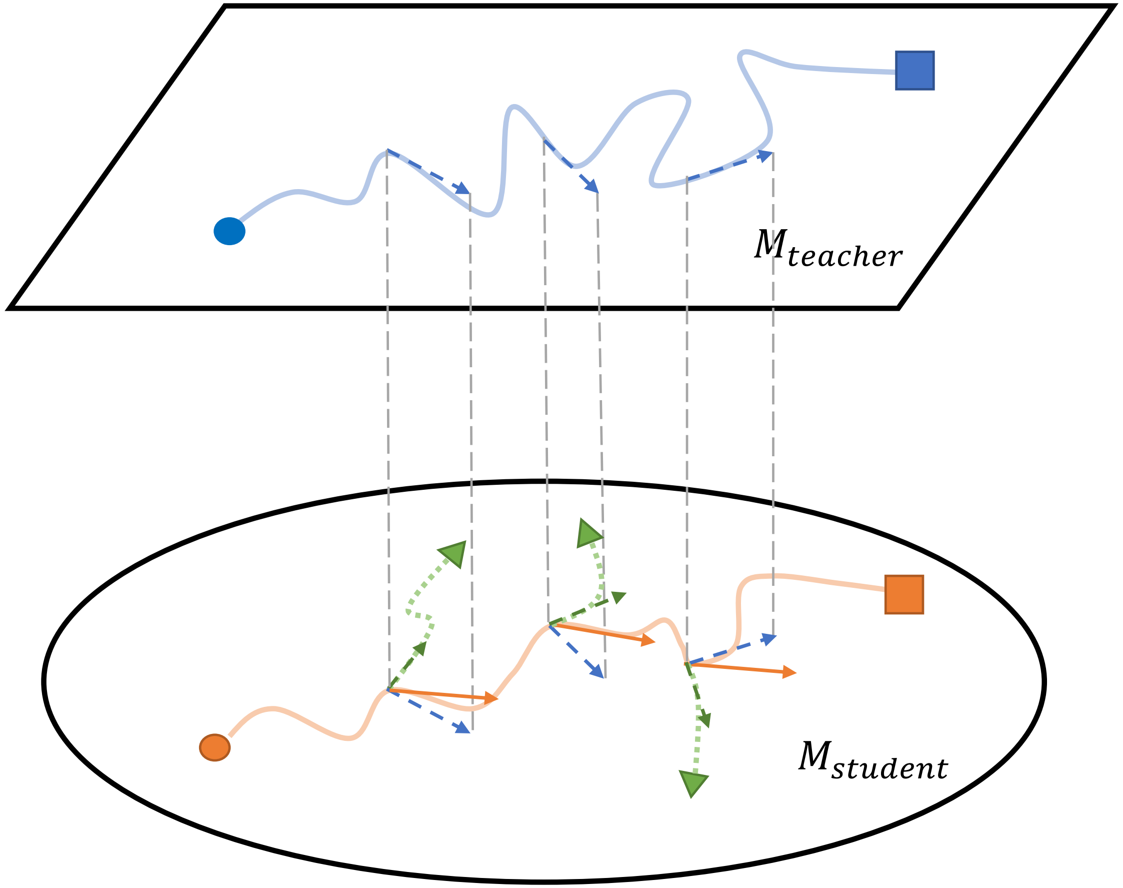
Learning and testing (3) →



Curriculum

learning

The focus of the course



Curriculum

and

on-line learning

…

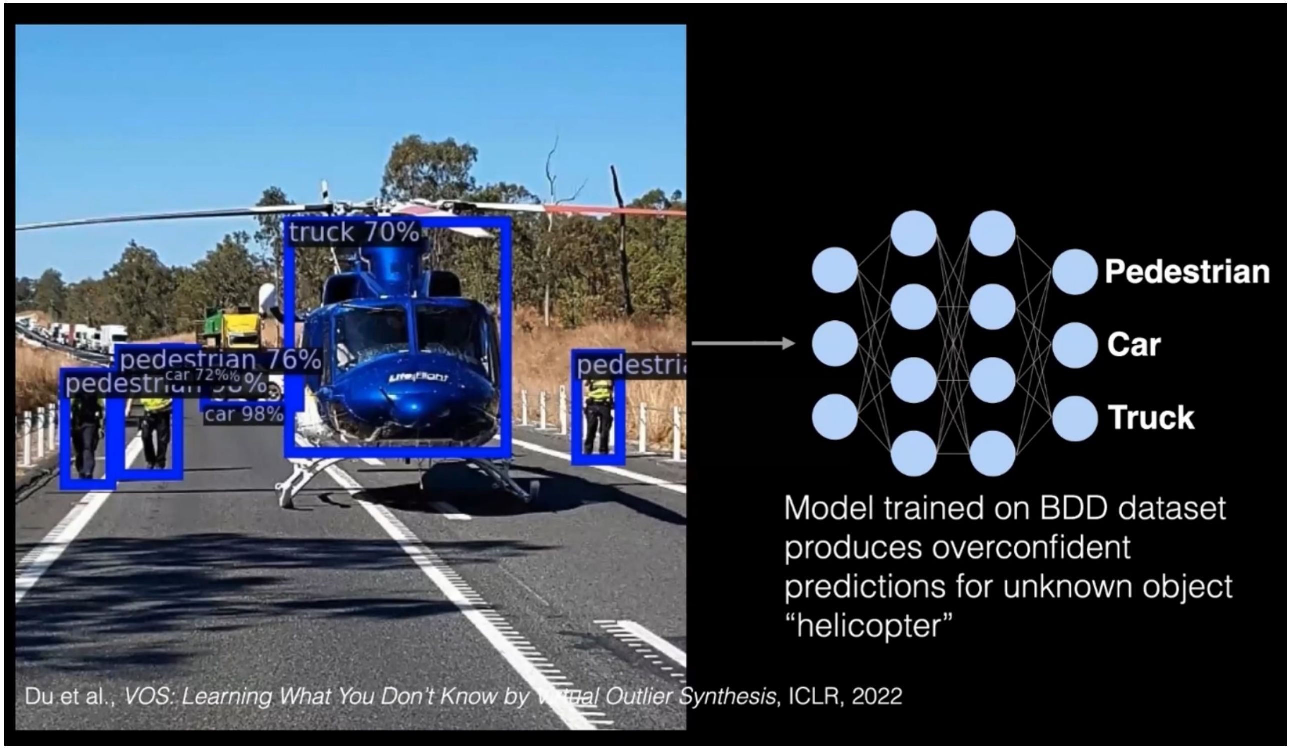
The focus of the course

- Out-Of Distribution learning (OOD)
  - **Zero-shot** learning

The focus of the course

- Out-Of Distribution learning (OOD)
  - **Zero-shot** learning

What you don't want

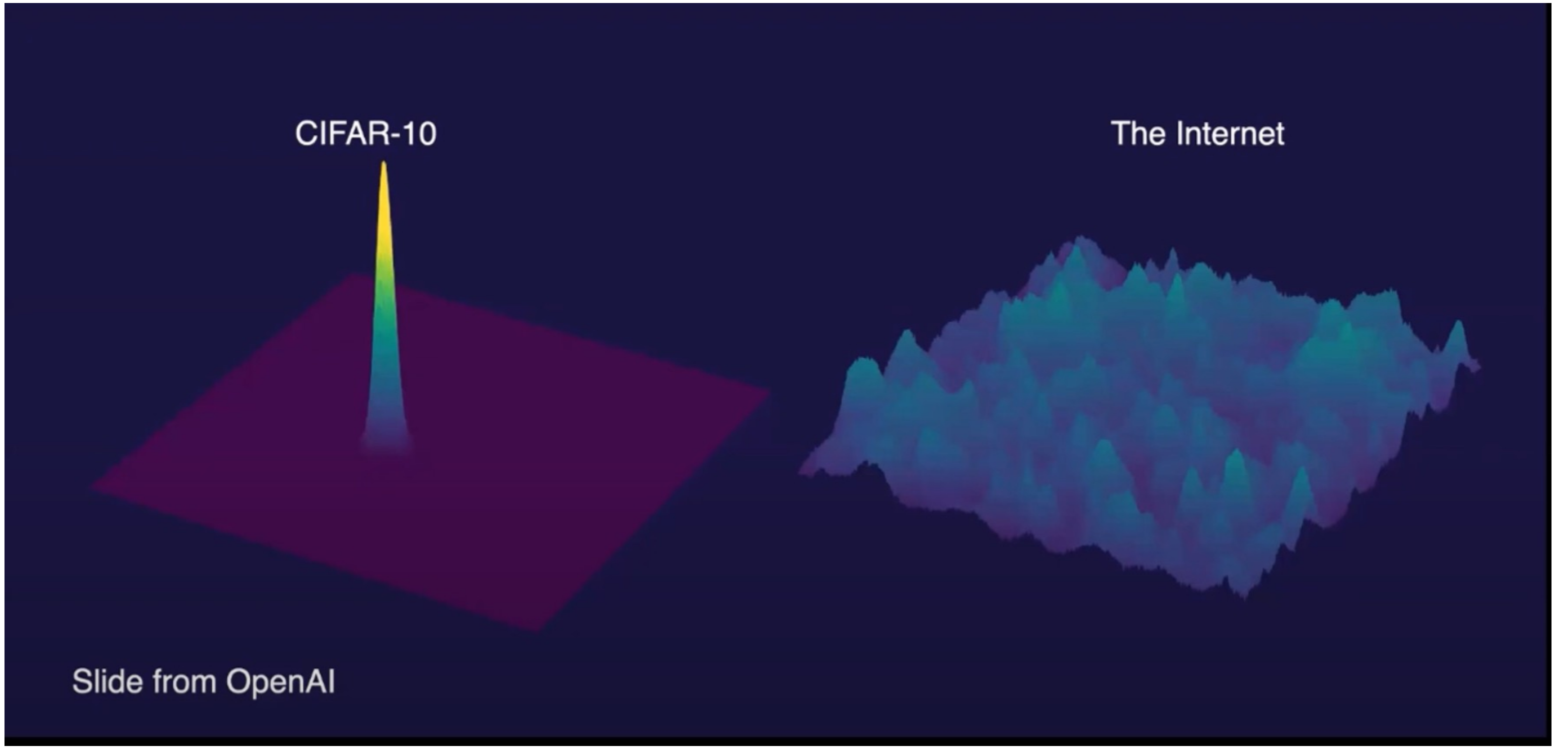


Model trained on BDD dataset produces overconfident predictions for unknown object "helicopter"

Du et al., *VOS: Learning What You Don't Know by Virtual Outlier Synthesis*, ICLR, 2022

The focus of the course

- Out-Of Distribution learning (OOD)

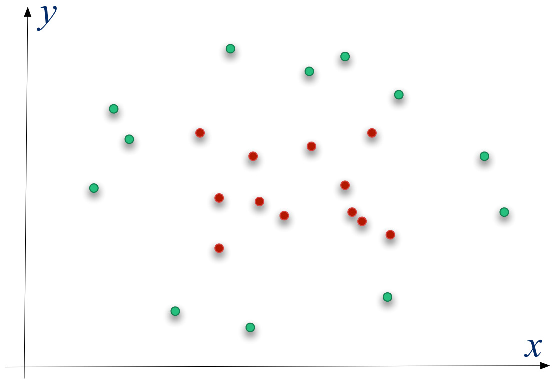


Slide from OpenAI

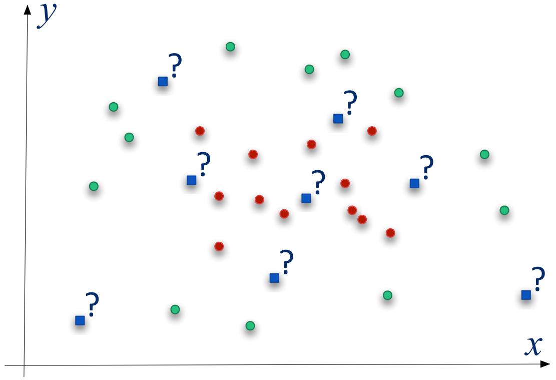
The focus of the course

- In-Distribution learning (I.I.D. setting)
  - **Same domain** and **distribution** between *learning* and *testing*

Learning



Testing



Is there any **difference** with Out-Of Distribution?

Why?

Issues that are the focus of the class

- Learning is about **extrapolating** predictions and regularities **from limited data**
  - **How** to achieve this?
  - **What** kind of **guarantees** can we hope?
  - **How** can we **obtain** them? Under which **assumptions**?

Issues that are the focus of the class

- In the case of **non stationary environments**, as in *domain adaptation*, *transfer learning* or *online learning*. (**O**ut-**O**f-**D**istribution learning)
  - How to **benefit (?) from learning** in a different environment?
  - Are there ways to **order the tasks** in the most beneficial way?
  - Can we still hope to have **guarantees**?
  - Under which **assumptions**? What are we ready to assume?

Issues that are the focus of the class

- Learning is about **extrapolating** predictions and regularities **from limited data**

  - **How** to achieve this?
  - **What** kind of **guarantees** can we hope?
  - **How** can we **obtain** them? Under which **assumptions**?

- In the case of **non stationary environments**, as in *domain adaptation*, *transfer learning* or *online learning*. (**O**ut-**O**f-**D**istribution learning)

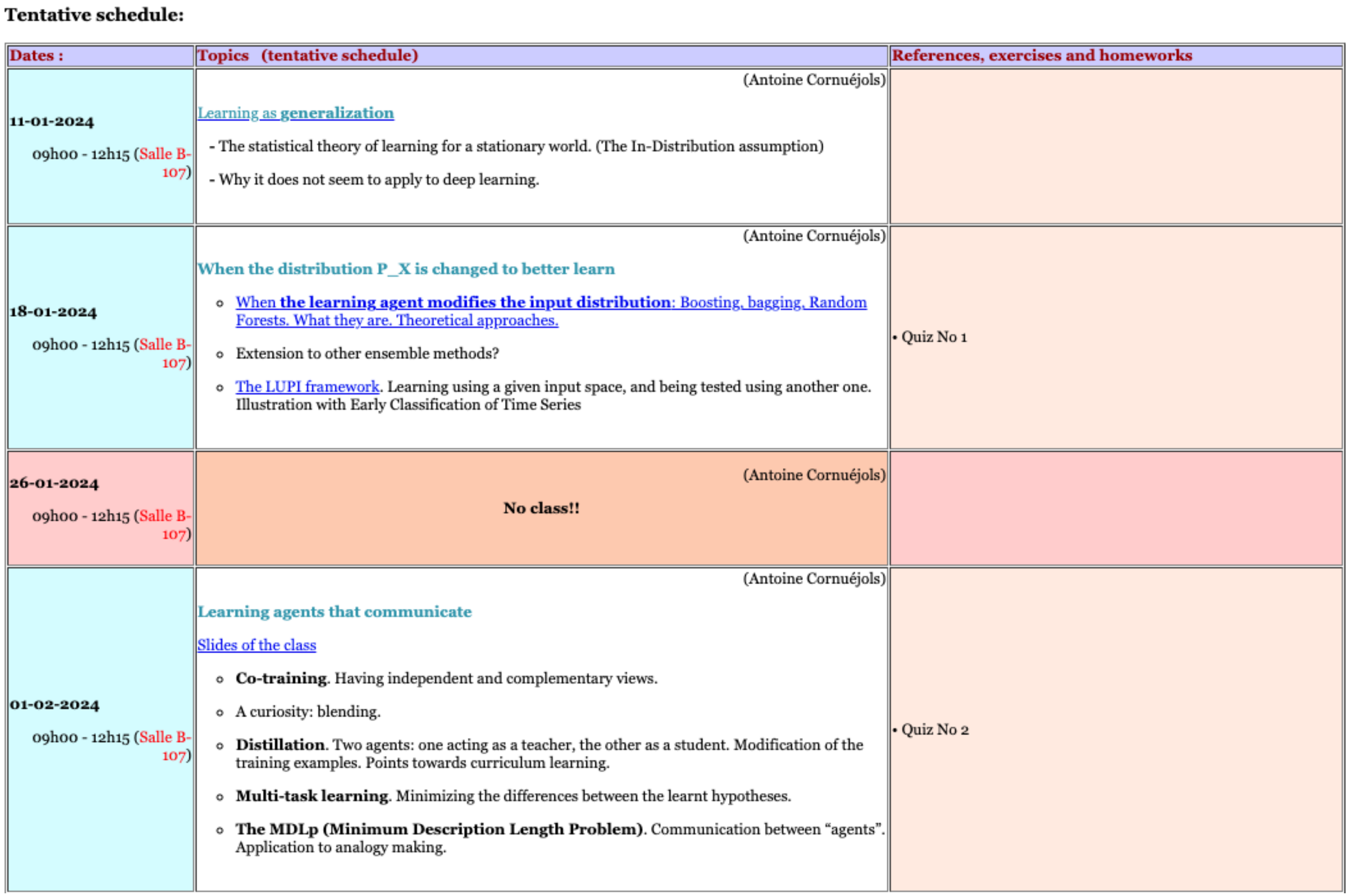
  - How to **benefit** (?) **from learning** in a different environment?
  - Are there ways to **order the tasks** in the most beneficial way?
  - Can we still hope to have **guarantees**?
  - Under which **assumptions**? What are we ready to assume?

Is it **trivial** to perform Out-Of-Distribution?

https://www.youtube.com/watch?v=QPSgM13hTK8&t=117

Outline of the course

https://antoinecornuejols.github.io/teaching/Master-AIC/M2-AIC-advanced-ML.html



Tentative schedule:

Dates :	Topics (tentative schedule)	References, exercises and homeworks
11-01-2024 09h00 - 12h15 (Salle B-107)	(Antoine Cornuéjols) Learning as **generalization** - The statistical theory of learning for a stationary world. (The In-Distribution assumption) - Why it does not seem to apply to deep learning.	
18-01-2024 09h00 - 12h15 (Salle B-107)	(Antoine Cornuéjols) **When the distribution P_X is changed to better learn** ○ When **the learning agent modifies the input distribution**: Boosting, bagging, Random Forests. What they are. Theoretical approaches. ○ Extension to other ensemble methods? ○ The LUPI framework. Learning using a given input space, and being tested using another one. Illustration with Early Classification of Time Series	• Quiz No 1
26-01-2024 09h00 - 12h15 (Salle B-107)	(Antoine Cornuéjols) **No class!!**	
01-02-2024 09h00 - 12h15 (Salle B-107)	(Antoine Cornuéjols) **Learning agents that communicate** Slides of the class ○ **Co-training**. Having independent and complementary views. ○ A curiosity: blending. ○ **Distillation**. Two agents: one acting as a teacher, the other as a student. Modification of the training examples. Points towards curriculum learning. ○ **Multi-task learning**. Minimizing the differences between the learnt hypotheses. ○ **The MDLp (Minimum Description Length Problem)**. Communication between "agents". Application to analogy making.	• Quiz No 2

Course's organization

6 **Courses:** 11/01 ; 18/01 ; 25/01 (no class!) ;
01/02 ; 08/02 ; 15/02 ; 29/02

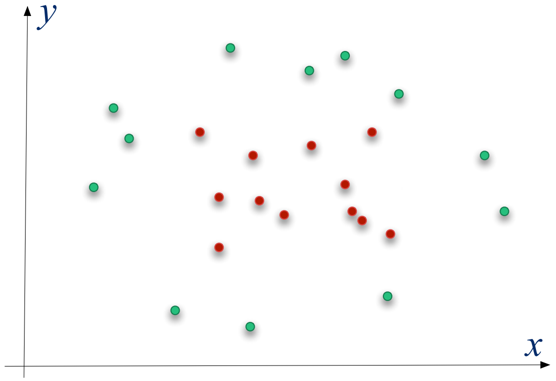
- 5 **quizz** (5 x 6 = **30** %)
- **Project**: Trying to **replicate** the experiments of a scientific paper : **50** %
  - 12/01/2021 : chosen project + team members (email)
  - 23/02/2021 : **final report** (10 pages strict. Format article ICML)
- **Critical review** of the paper by same groups : **20**%

Questions?

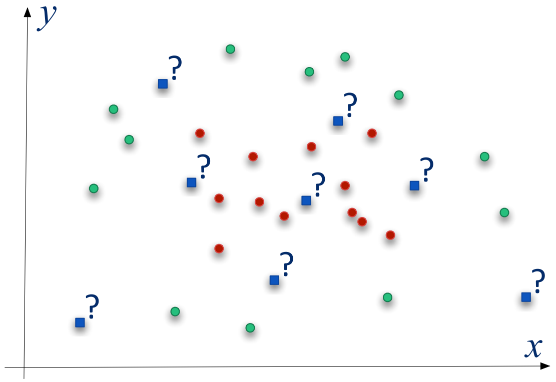
In-Distribution learning (I.I.D. setting)

…

Learning →



Testing →



Is there any **difference** with Out-Of Distribution?

Why?

Outline of today's class

1. The **mystery** of in-distribution learning (standard induction)

2. A 101 course on the statistical learning **theory**

3. **Why does it fail** to account for deep neural networks?

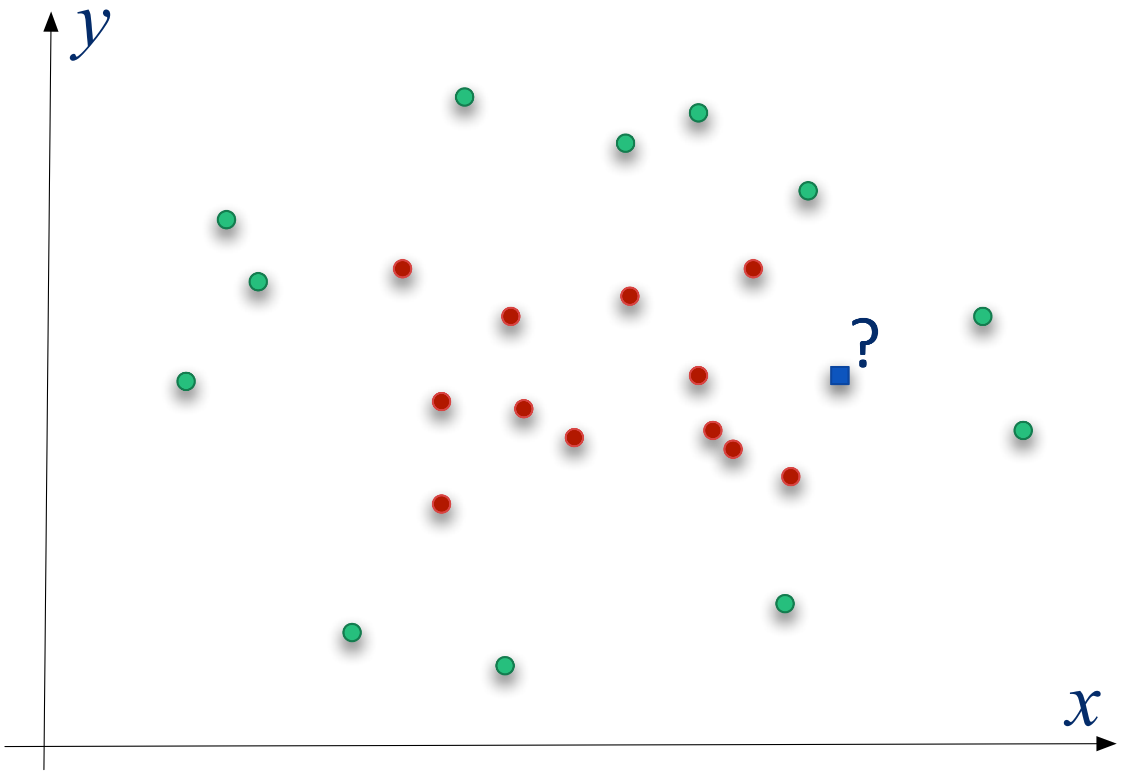
4. The **no-free-lunch** theorem

In-Distribution Supervised learning:

Obvious **really**?

Supervised induction

- We want to be able to predict the class of unseen examples



→ A **decision function**

One example that tells a lot …

- Examples described using:

  Number (1 or 2); ***size*** (small or large); ***shape*** (circle or square); ***color*** (red or green)

- They belong either to class ‘**+**’ or to class ‘**-**’

- Examples described using:

  Number (1 or 2); ***size*** (small or large); ***shape*** (circle or square); ***color*** (red or green)

- They belong either to class '**+**' or to class '**-**'

Description	Your answer	True answer
1 large red square		

- Examples described using:

  Number (1 or 2); ***size*** (small or large); ***shape*** (circle or square); ***color*** (red or green)

- They belong either to class '**+**' or to class '**-**'

Description	Your answer	True answer
1 large red square		-

- Examples described using:

  Number (1 or 2); ***size*** (small or large); ***shape*** (circle or square); ***color*** (red or green)

- They belong either to class '**+**' or to class '**-**'

Description	Your answer	True answer
1 large red square		-
1 large green square		

AgroParisTech

- Examples described using:

  Number (1 or 2); ***size*** (small or large); ***shape*** (circle or square); ***color*** (red or green)

- They belong either to class '**+**' or to class '**-**'

Description	Your answer	True answer
1 large red square		-
1 large green square		+

- Examples described using:

  Number (1 or 2); ***size*** (small or large); ***shape*** (circle or square); ***color*** (red or green)

- They belong either to class '**+**' or to class '**-**'

Description	Your answer	True answer
1 large red square		-
1 large green square		+
2 small red squares		

- Examples described using:

  Number (1 or 2); ***size*** (small or large); ***shape*** (circle or square); ***color*** (red or green)

- They belong either to class '**+**' or to class '**-**'

Description	Your answer	True answer
1 large red square		-
1 large green square		+
2 small red squares		+

- Examples described using:

  Number (1 or 2); ***size*** (small or large); ***shape*** (circle or square); ***color*** (red or green)

- They belong either to class '**+**' or to class '**-**'

Description	Your answer	True answer
1 large red square		-
1 large green square		+
2 small red squares		+
2 large red circles		

- Examples described using:

  Number (1 or 2); ***size*** (small or large); ***shape*** (circle or square); ***color*** (red or green)

- They belong either to class '**+**' or to class '**-**'

Description	Your answer	True answer
1 large red square		-
1 large green square		+
2 small red squares		+
2 large red circles		-

- Examples described using:

  Number (1 or 2); ***size*** (small or large); ***shape*** (circle or square); ***color*** (red or green)

- They belong either to class '**+**' or to class '**-**'

Description	Your answer	True answer
1 large red square		-
1 large green square		+
2 small red squares		+
2 large red circles		-
1 large green circle		

- Examples described using:

  Number (1 or 2); ***size*** (small or large); ***shape*** (circle or square); ***color*** (red or green)

- They belong either to class '**+**' or to class '**-**'

Description	Your answer	True answer
1 large red square		-
1 large green square		+
2 small red squares		+
2 large red circles		-
1 large green circle		+

- Examples described using:

  Number (1 or 2); ***size*** (small or large); ***shape*** (circle or square); ***color*** (red or green)

- They belong either to class '**+**' or to class '**-**'

Description	Your answer	True answer
1 large red square		-
1 large green square		+
2 small red squares		+
2 large red circles		-
1 large green circle		+
1 small red circle		

- Examples described using:

  Number (1 or 2); ***size*** (small or large); ***shape*** (circle or square); ***color*** (red or green)

- They belong either to class '**+**' or to class '**-**'

Description	Your answer	True answer
1 large red square		-
1 large green square		+
2 small red squares		+
2 large red circles		-
1 large green circle		+
1 small red circle		+

- When would you be **certain** about your guess?

- What **assumption** are you making?

Supervised learning

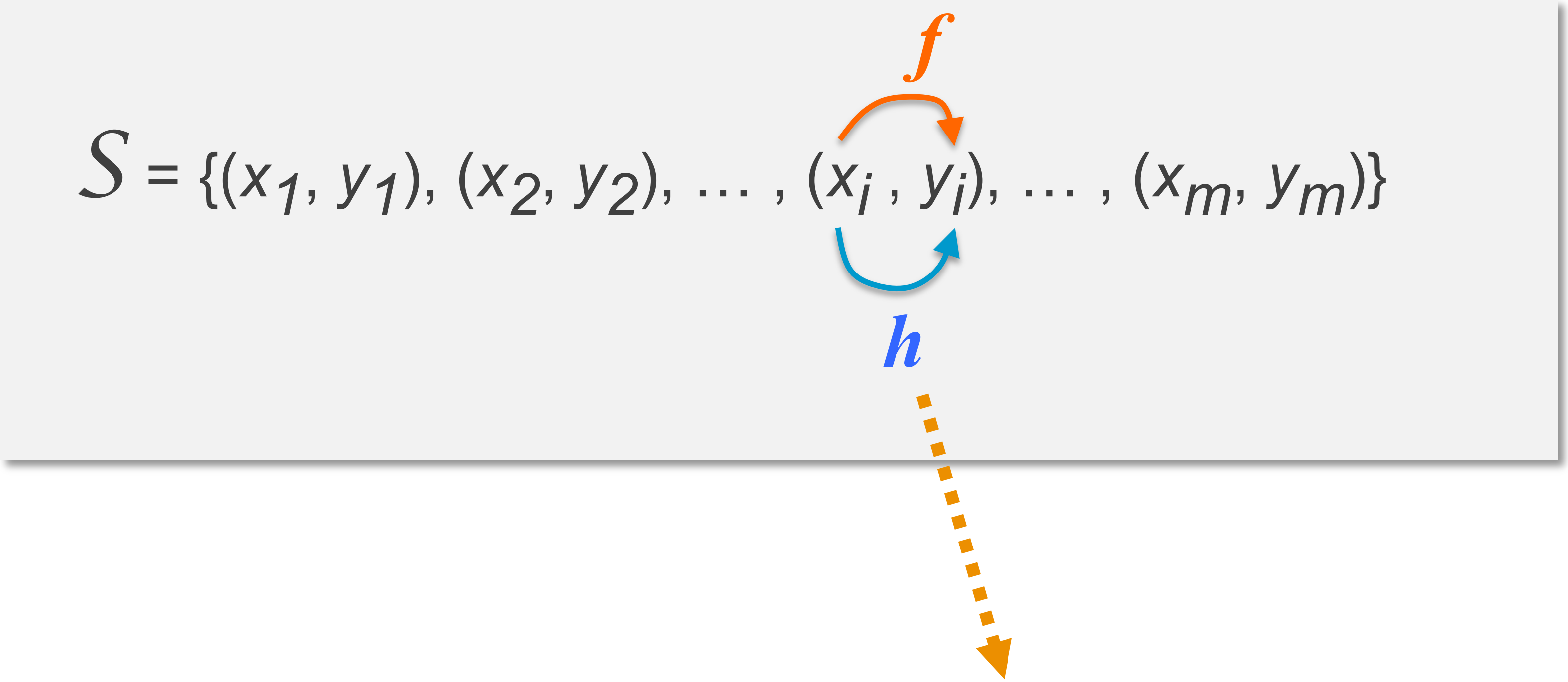
- A *training set*

$$\begin{tikzcd}[column sep=small] S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_i \arrow[r, bend left, "f"] \arrow[r, bend right, "h"'] & y_i), \ldots, (x_m, y_m)\} \end{tikzcd}$$

Prediction for new examples $x \xrightarrow{h} y$?

Supervised learning

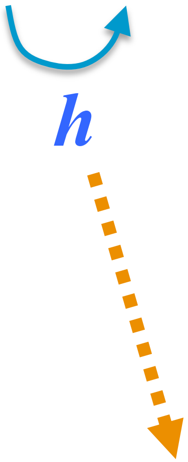
- A *training set*



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

Prediction for new examples x –h-> y ?

- What **assumption** are you making?



h

x –*h*-> *y* ?

Is this assumption reasonable?

Is it sufficient?

- Examples described using:

  Number (1 or 2); ***size*** (small or large); ***shape*** (circle or square); ***color*** (red or green)

- They belong either to class '**+**' or to class '**-**'

Description	Your answer	True answer
1 large red square		-
1 large green square		+
2 small red squares		+
2 large red circles		-
1 large green circle		+
1 small red circle		+
1 small green square		

- Examples described using:

  Number (1 or 2); ***size*** (small or large); ***shape*** (circle or square); ***color*** (red or green)

- They belong either to class '**+**' or to class '**-**'

Description	Your answer	True answer
1 large red square		-
1 large green square		+
2 small red squares		+
2 large red circles		-
1 large green circle		+
1 small red circle		+
1 small green square		-

- Examples described using:

  Number (1 or 2); ***size*** (small or large); ***shape*** (circle or square); ***color*** (red or green)

- They belong either to class '**+**' or to class '**-**'

Description	Your answer	True answer
1 large red square		-
1 large green square		+
2 small red squares		+
2 large red circles		-
1 large green circle		+
1 small red circle		+
1 small green square		-
1 small red square		

- Examples described using:

  Number (1 or 2); ***size*** (small or large); ***shape*** (circle or square); ***color*** (red or green)

- They belong either to class '**+**' or to class '**-**'

Description	Your answer	True answer
1 large red square		-
1 large green square		+
2 small red squares		+
2 large red circles		-
1 large green circle		+
1 small red circle		+
1 small green square		-
1 small red square		+

AgroParisTech

- Examples described using:

  Number (1 or 2); ***size*** (small or large); ***shape*** (circle or square); ***color*** (red or green)

- They belong either to class '**+**' or to class '**-**'

Description	Your answer	True answer
1 large red square		-
1 large green square		+
2 small red squares		+
2 large red circles		-
1 large green circle		+
1 small red circle		+
1 small green square		-
1 small red square		+
2 large green squares		

- Examples described using:

  Number (1 or 2); ***size*** (small or large); ***shape*** (circle or square); ***color*** (red or green)

- They belong either to class '**+**' or to class '**-**'

Description	Your answer	True answer
1 large red square		-
1 large green square		+
2 small red squares		+
2 large red circles		-
1 large green circle		+
1 small red circle		+
1 small green square		-
1 small red square		+
2 large green squares		+

One example that tells a lot …

- Examples described using:

  Number (1 or 2); ***size*** (small or large); ***shape*** (circle or square); ***color*** (red or green)

Description	**Your** prediction	**True** class
1 large red square		-
1 large green square		+
2 small red squares		+
2 large red circles		-
1 large green circle		+
1 small red circle		+

How many possible functions altogether from *X* to *Y* ? $2^{2^4} = 2^{16} = 65,536$

How many functions do remain after 9 training examples? $2^5 = 512$

- Are you not **worried**?

One example that tells a lot …

- Examples described using:

  Number (1 or 2); ***size*** (small or large); ***shape*** (circle or square); ***color*** (red or green)

15

Description	**Your** prediction	**True** class
1 large red square		-
1 large green square		+
2 small red squares		+
2 large red circles		-
1 large green circle		+
1 small red circle		+
1 small green square		-
1 small red square		+
2 large green squares		+
2 small green squares		+
2 small red circles		+
1 small green circle		-
2 large green circles		-
2 small green circles		+
1 large red circle		-
2 large red squares	?	

How many **remaining functions?**

← ?

One example that tells a lot …

- Examples described using:

  ~~Number~~ (1 or 2); ***size*** (small or large); ***~~shape~~*** (circle or square); ***color*** (red or green)

Description	Your prediction	True class
~~1~~ large red ~~square~~		-
~~1~~ large green ~~square~~		+
~~2~~ small red ~~squares~~		+
~~2~~ large red ~~circles~~		-
~~1~~ large green ~~circle~~		+
~~1~~ small red ~~circle~~		+

How many possible functions with 2 descriptors from *X* to *Y* ? $2^{2^2} = 2^4 = 16$

How many functions do remain after 3 ≠ training examples? $2^1 = 2$

small green → ?

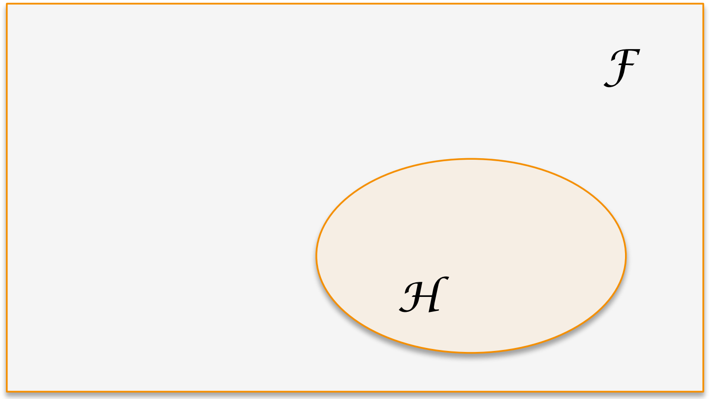
Induction: an impossible game?

- **A bias is need**

Induction: an impossible game?

- **A bias is need**

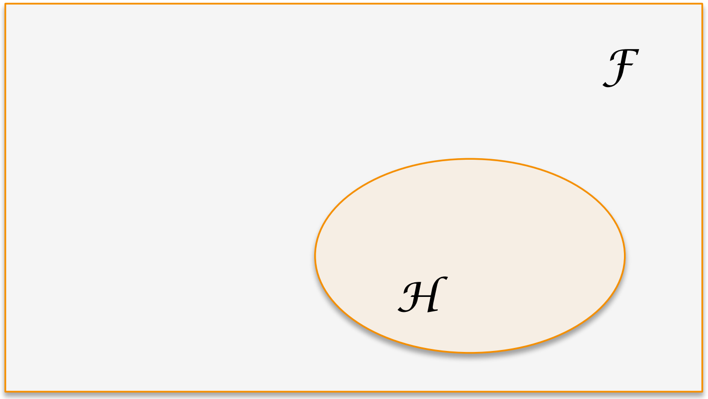
- **Types** of bias
  - **Representation** bias (declarative)

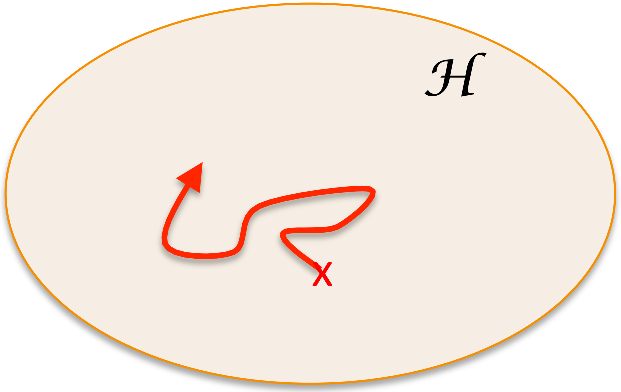


Induction: an impossible game?

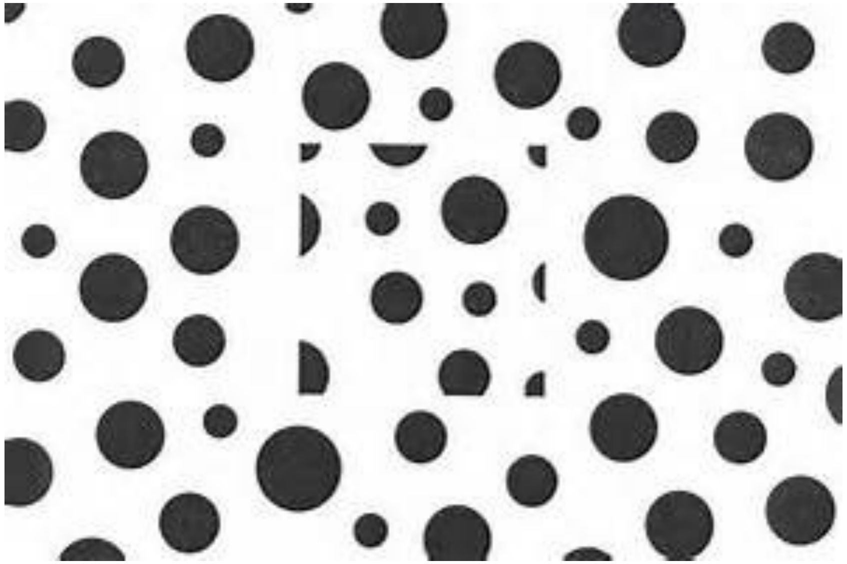
- **A bias is need**

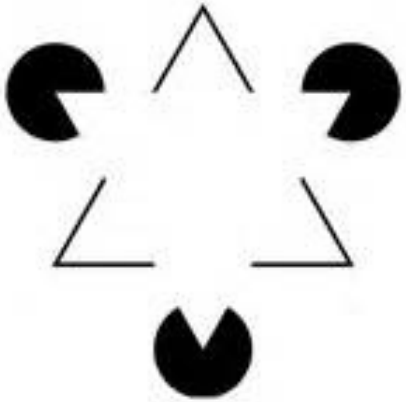
- **Types** of bias
  - **Representation** bias (declarative)
  - **Research** bias (procedural)

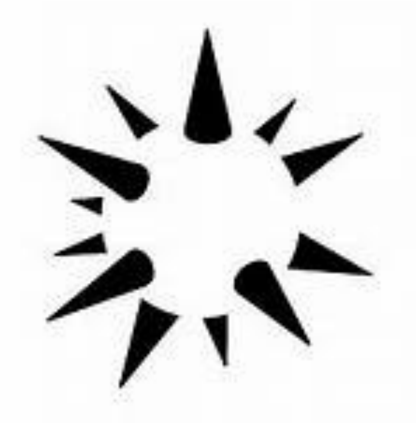




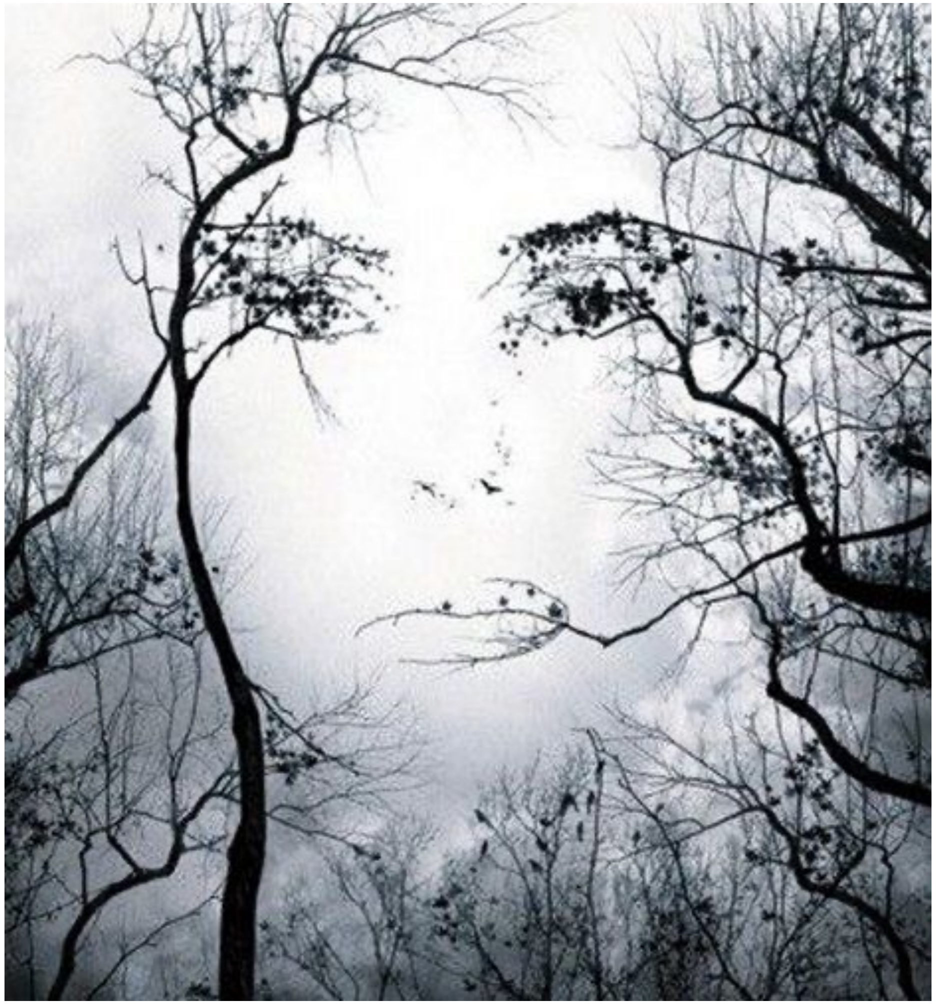
Interpretation – completion of percepts



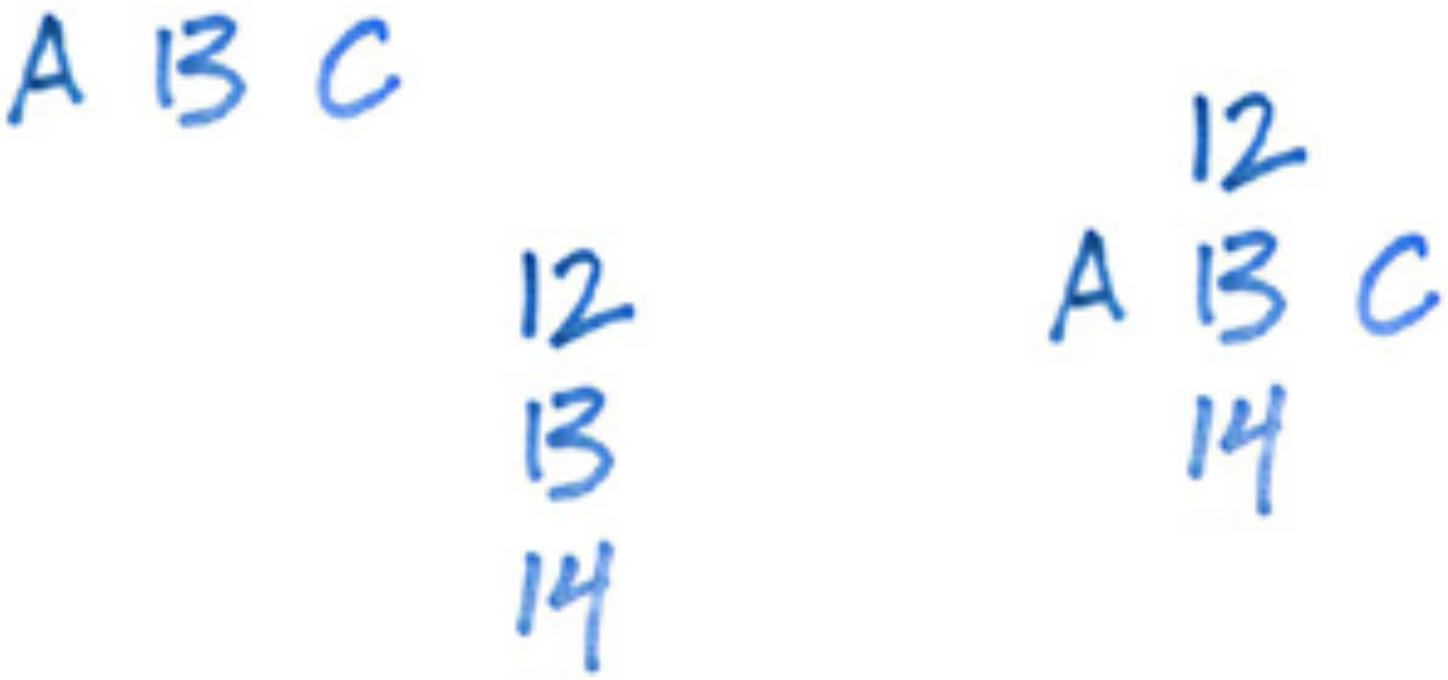




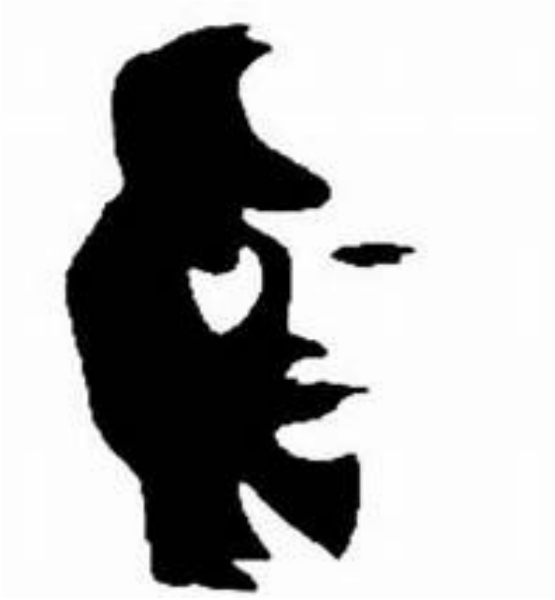
Interpretation – completion of percepts

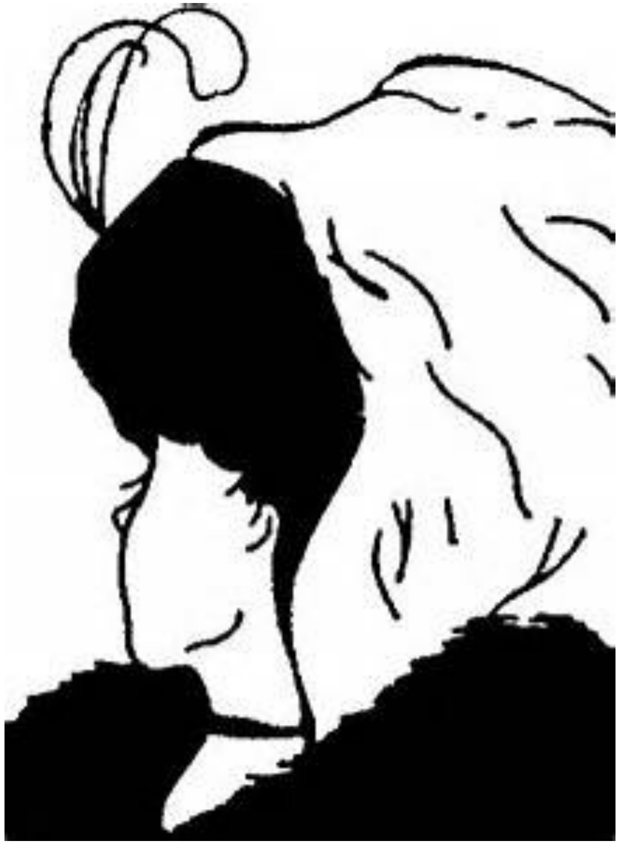


Interpretation – completion of percepts

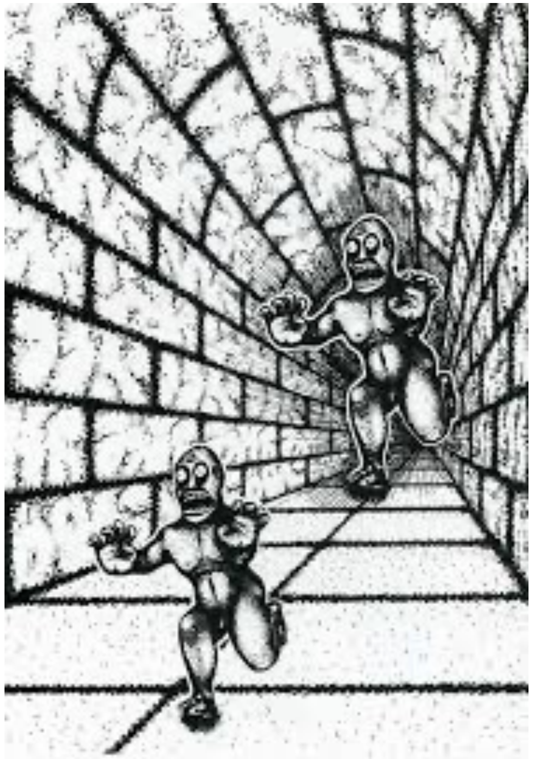


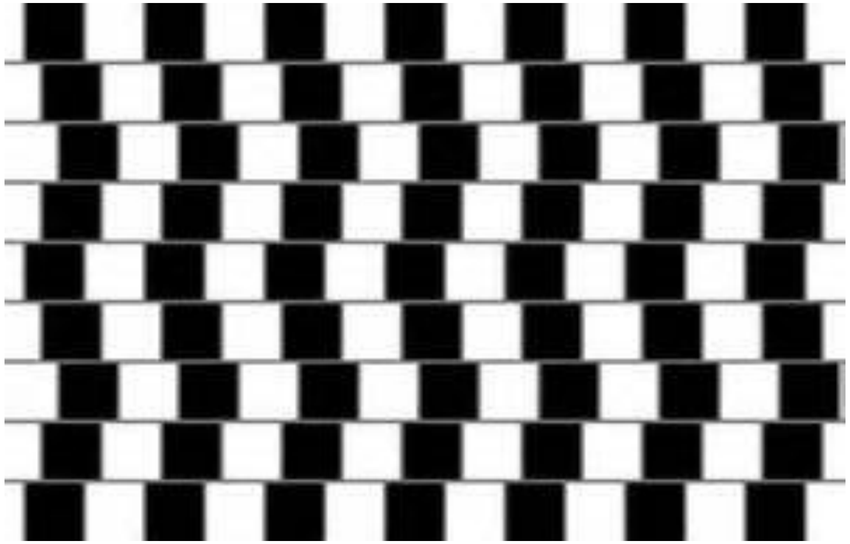
Interprétation – complétion de percepts



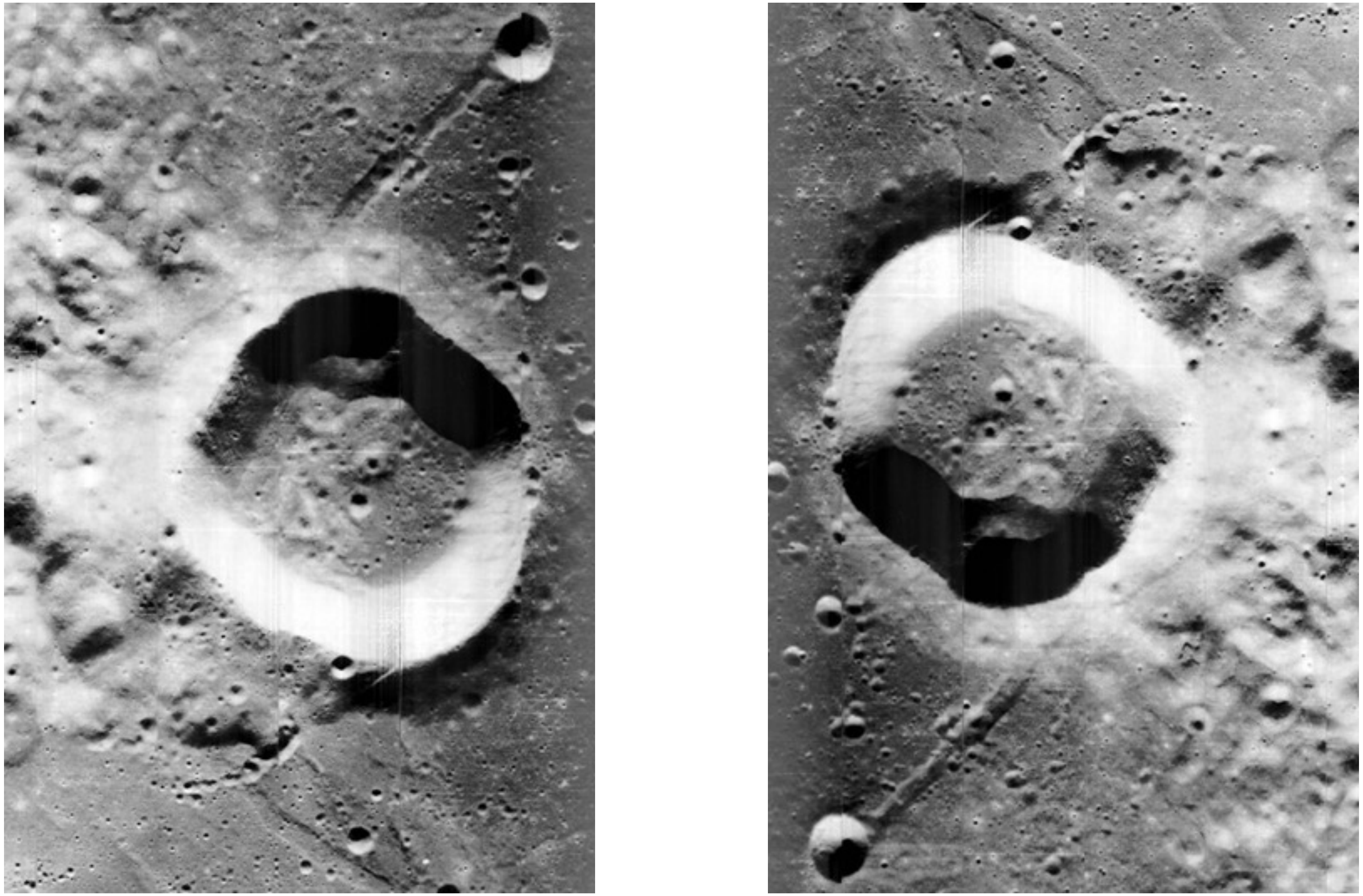


Optical illusions



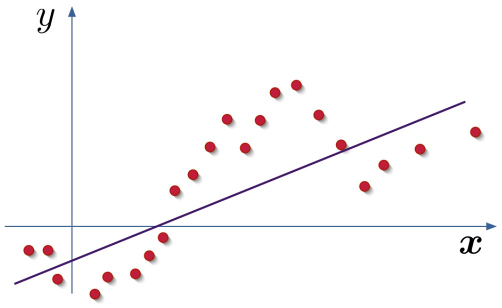


Induction and its illusions

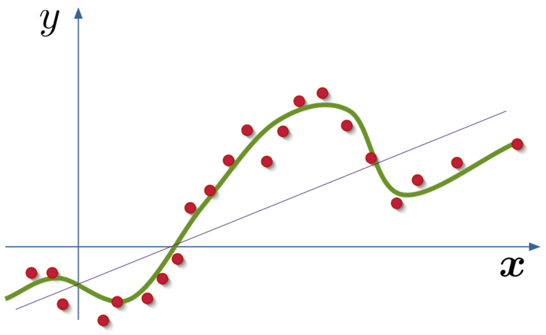


Illustration

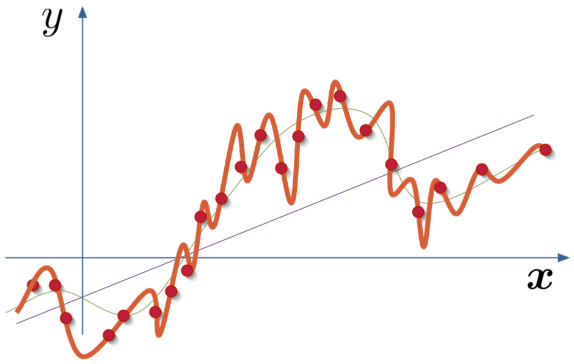
Bias is what make you prefer some hypotheses over other



Under-fitting

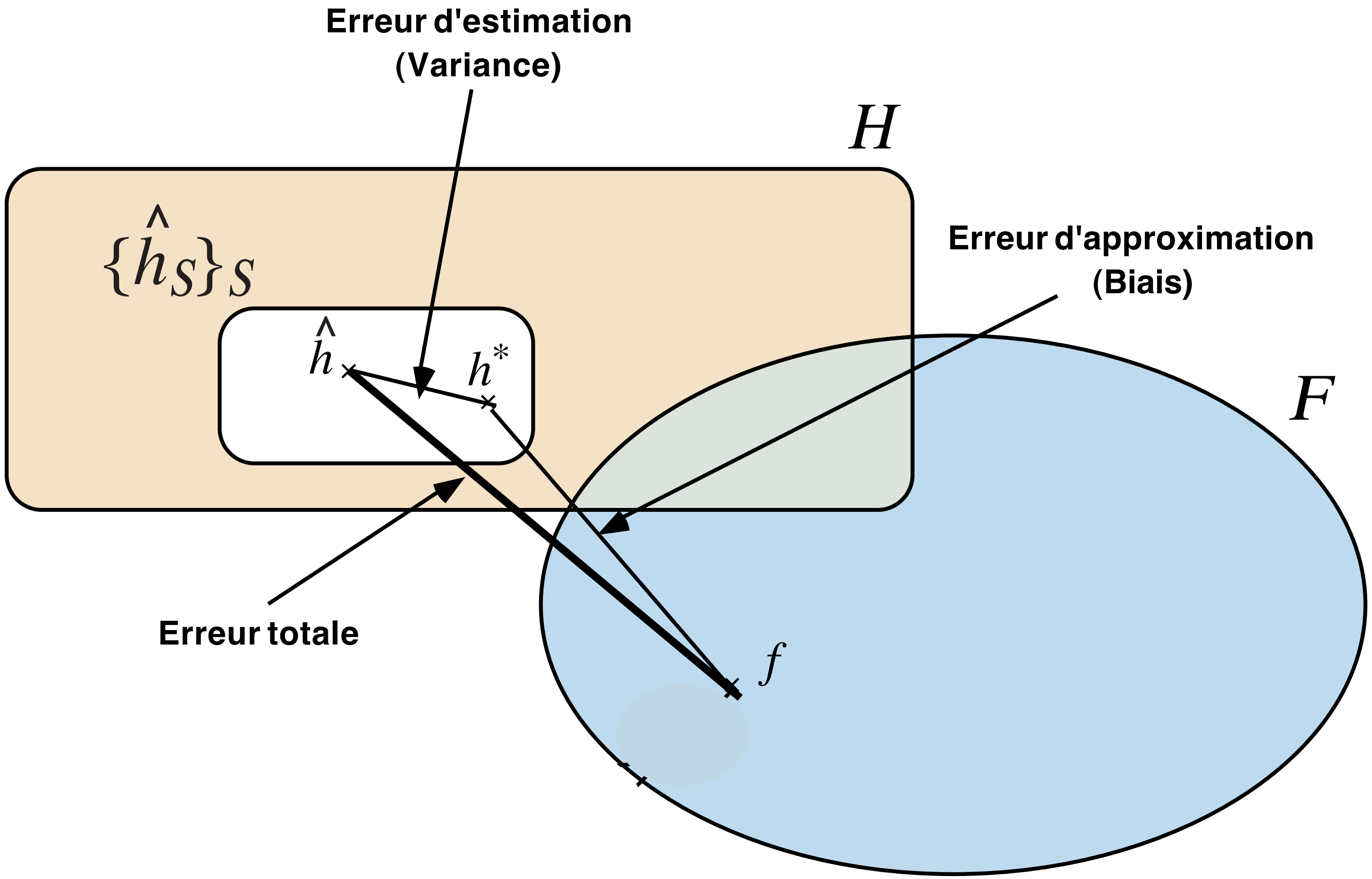


Appropriate-fitting



Over-fitting

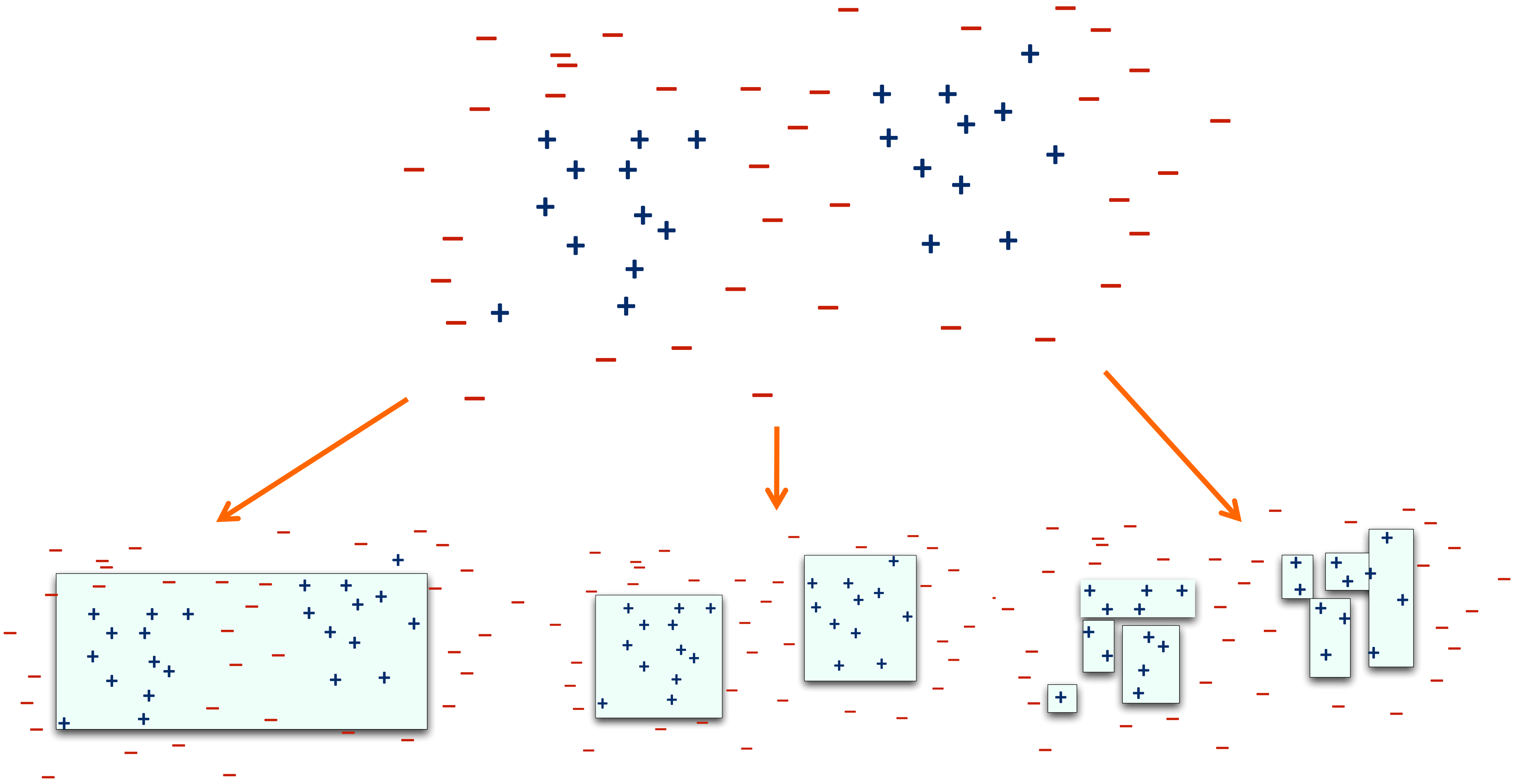
The bias-variance tradeoff



...

Illustration

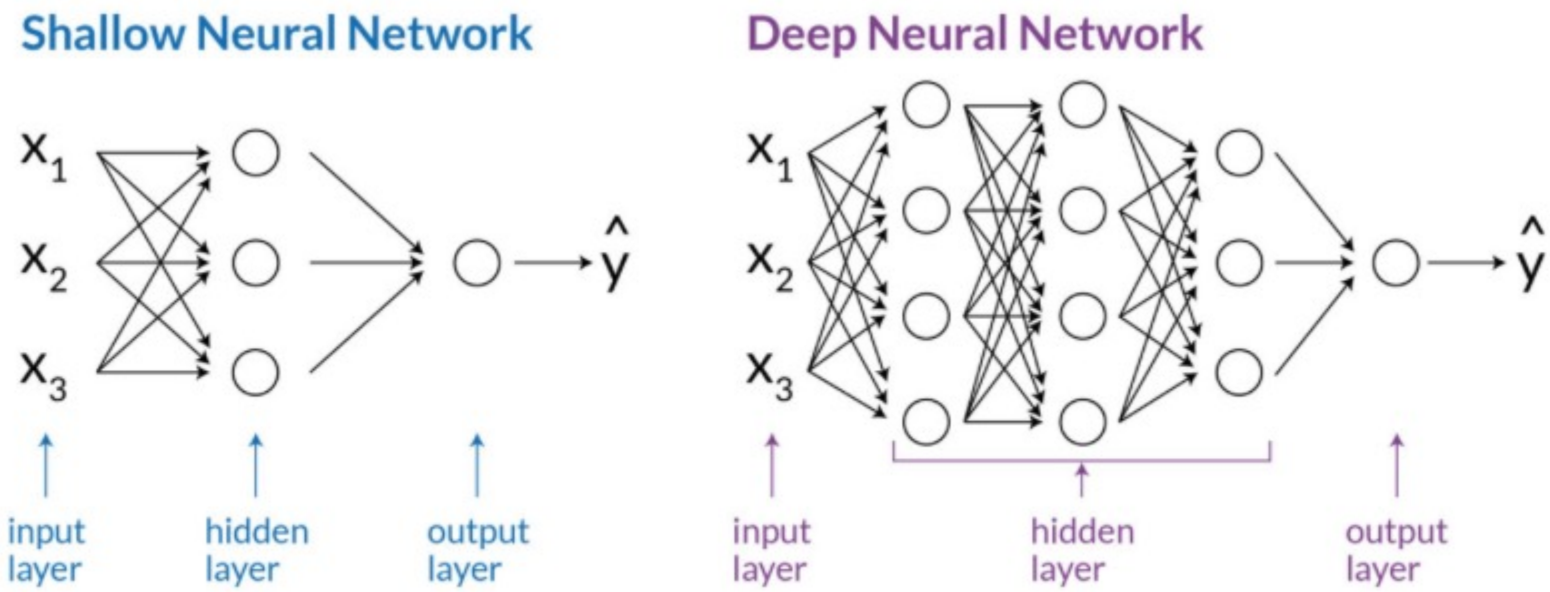
...



Under-fitting Appropriate-fitting Over-fitting

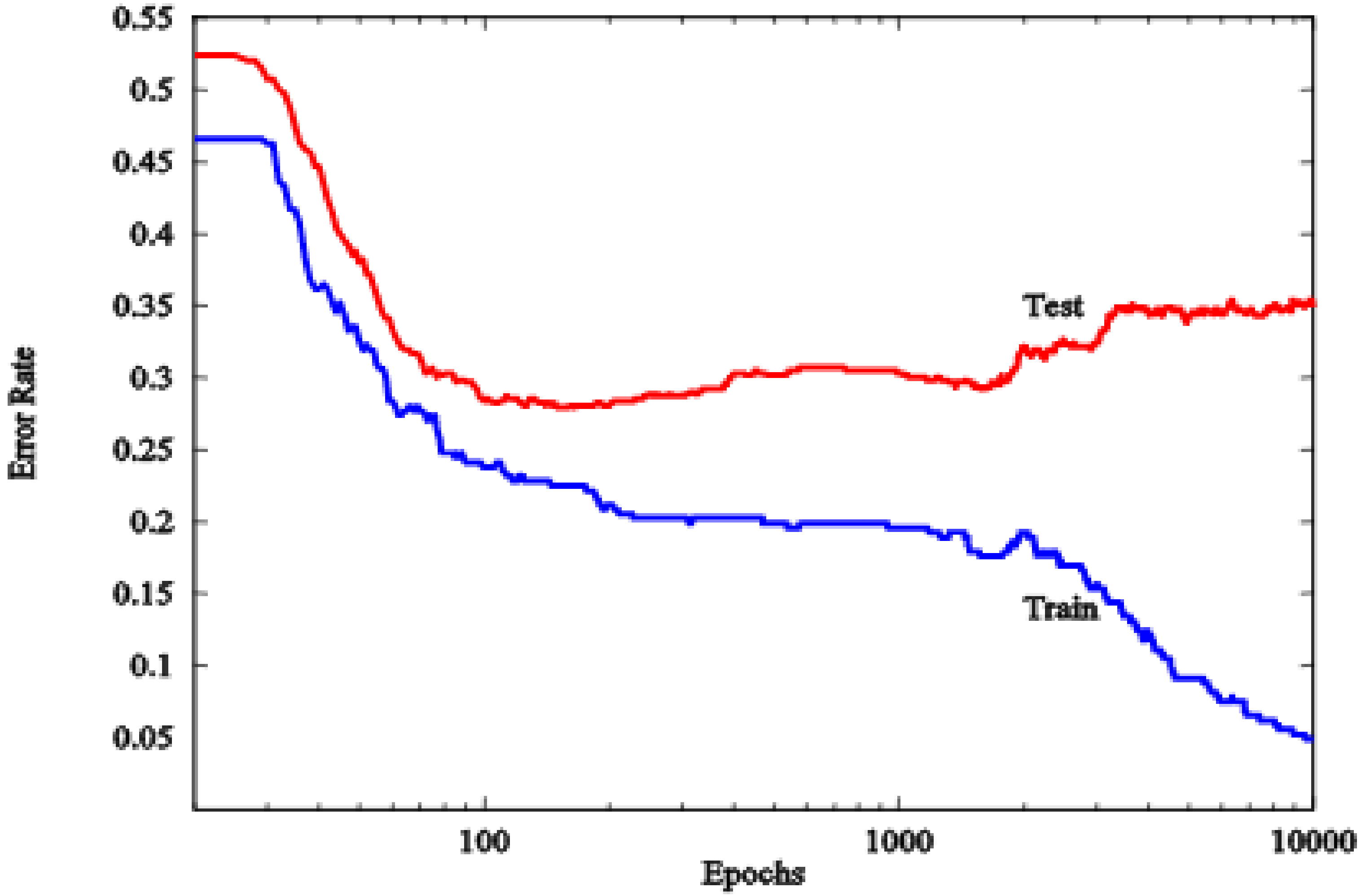
How to chose the architecture of a NN?

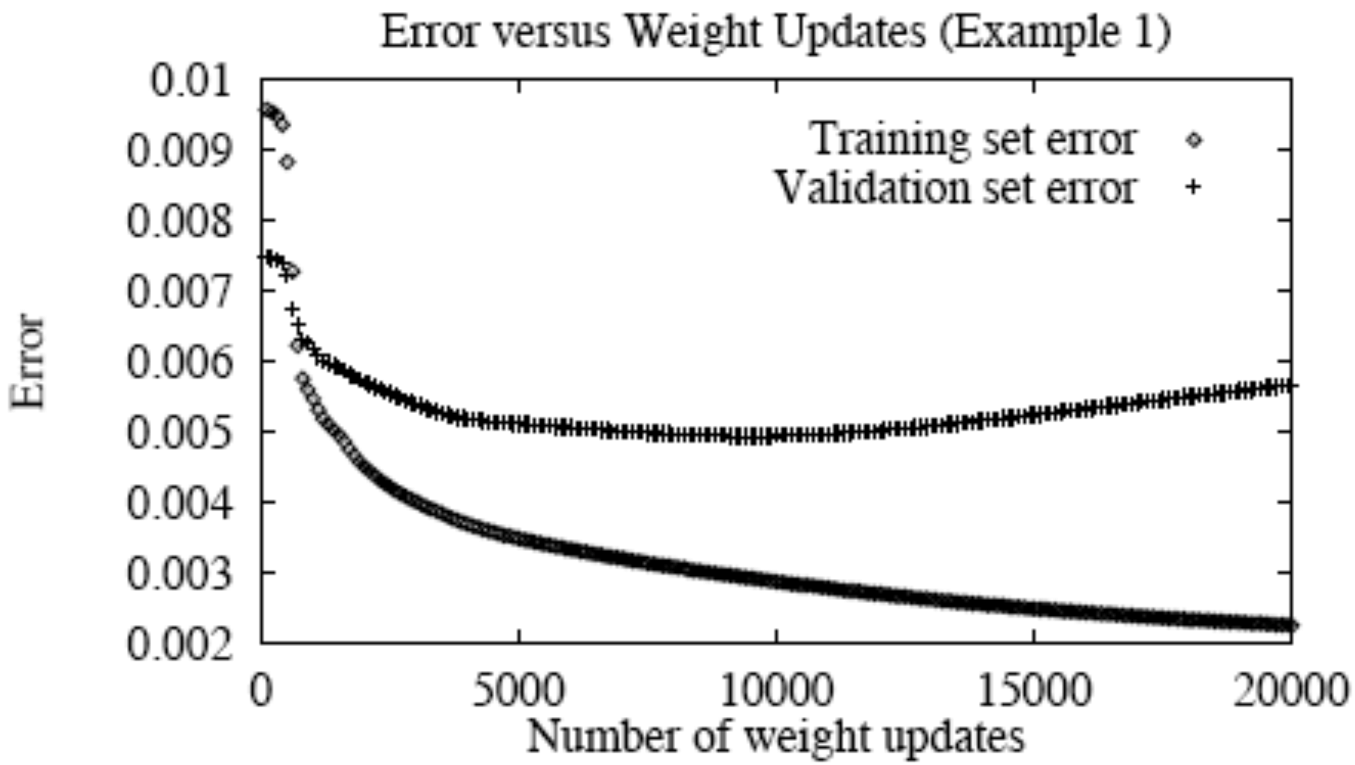
...



Shallow and Deep Neural Networks.

Over-fitting when learning

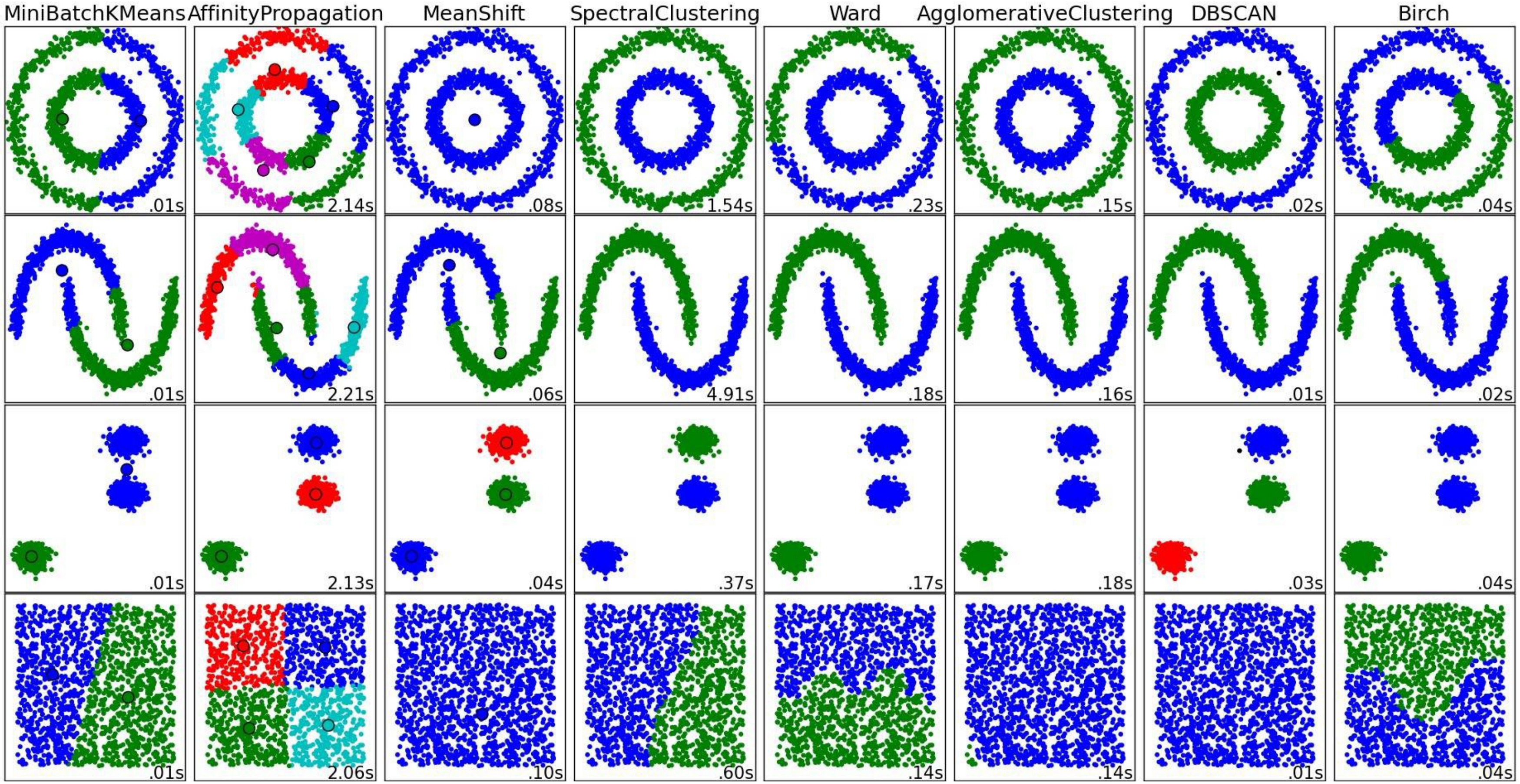




- Curves for **1 000** examples
- *and for* ***2 000 examples ?***

Clustering

Effects of the a priori bias



Induction everywhere

The role of induction

- [Leslie Valiant, « *Probably Approximately Correct. Nature's Algorithms for Learning and Prospering in a Complex World* », Basic Books, 2013]

« From this, we have to conclude that **generalization** or **induction** is a **pervasive phenomenon** (…). It is as routine and reproducible a phenomenon as objects falling under gravity. It is **reasonable to expect a quantitative scientific explanation** of this highly reproducible phenomenon. »

The role of induction

- [Edwin T. Jaynes, « *Probability theory. The logic of science* », Cambridge U. Press, 2003], p.3

« We are hardly able to get through one waking hour without facing some situation (e.g. *will it rain or won't it?*) where we **do not have enough information to permit deductive reasoning**; but still we must decide immediately.

In spite of its familiarity, the formation of plausible conclusions is a **very subtle process**. »

Sequences

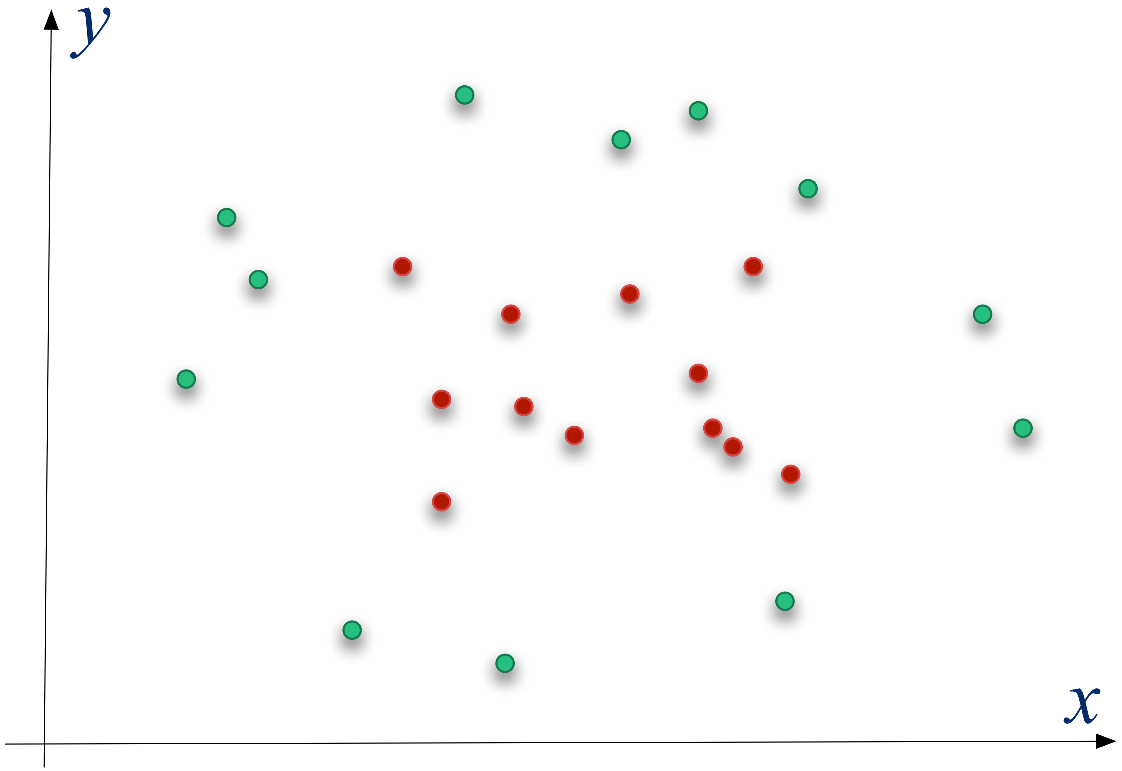
- 1 1 2 3 5 8 13 21 …
- 1 2 3 5 …
- 1 1 1 2 1 1 2 1 1 1 1 1 2 2 1 3 1 2 2 1 1 …

Sequences

- 1 1 1 2 1 1 2 1 1 1 1 1 2 2 1 3 1 2 2 1 1 …
- 1
- 1 1
- 2 1
- 1 2 & 1 1
- 1 1 & 1 2 & 2 1
- 1 1 1 2 1 1 2 1 1 1 1 1 2 2 1 3 1 2 2 1 1 …
  - **Comment** ?
  - **Pourquoi** serait-il possible de faire de l'induction ?
  - Est-ce qu'**un exemple supplémentaire** doit augmenter la confiance dans la règle induite ?
  - **Combien faut-il d'exemples** ?

Supervised induction

- How to chose the decision function?



Interrogations

Each time:

Specific cases => general **law** or adaptation to a **new case**

1. **How** this generalization **is allowed**?

2. Can we **guarantee something?**

Outline of today's class

1. The mystery of in-distribution learning (standard induction)

2. A 101 course on the statistical learning **theory**

3. **Why does it fail** to account for deep neural networks?

4. The **no-free-lunch** theorem

What kind of theoretical guarantees

on induction can we get?

A centuries-old question

A centuries-old question

- How do we know that the chosen hypothesis is **correct**?
- **How many** examples do you need to get a good result?
- Which **hypothesis space to** explore?
- If the hypothesis space is very complex, can we expect to find the **global optimum**? Or only a **local optimum**?
- How to **avoid over-fitting**?

A centuries-old question

- The razor of **Ockham** (1288 – 1348)
  - The MDLp (Minimum Description Length principle)

- The **bayésian** analysis

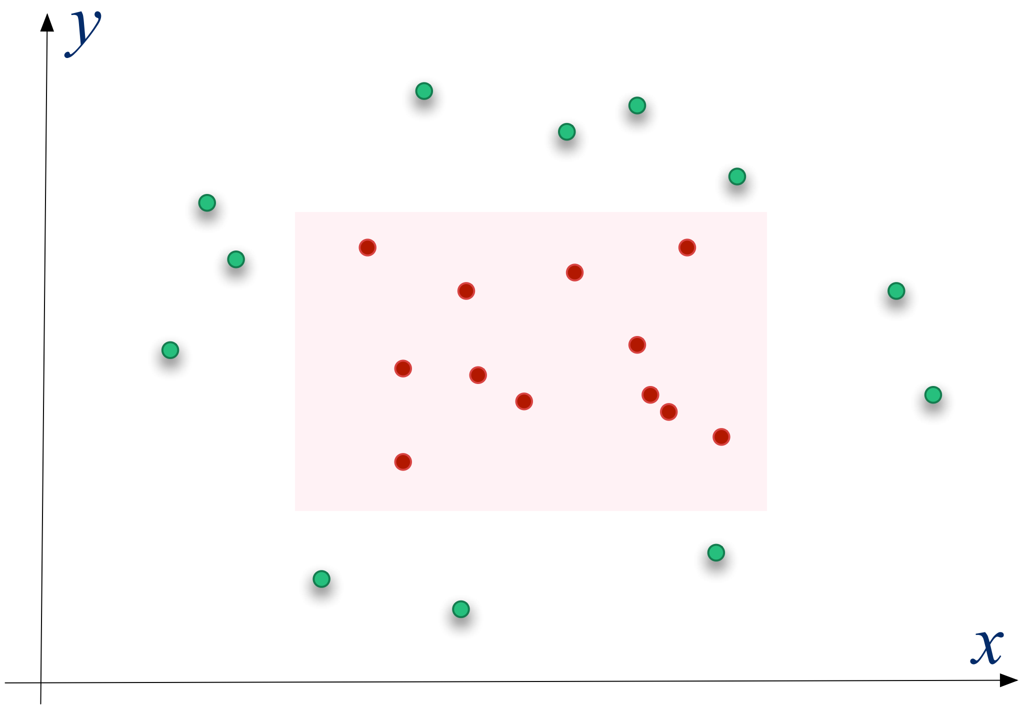
- The **E**mpirical **R**isk **M**inimization (ERM)
  - Minimization of a regularized empirical risk**régularisé**

PAC learning

Probably Approximatively Correct

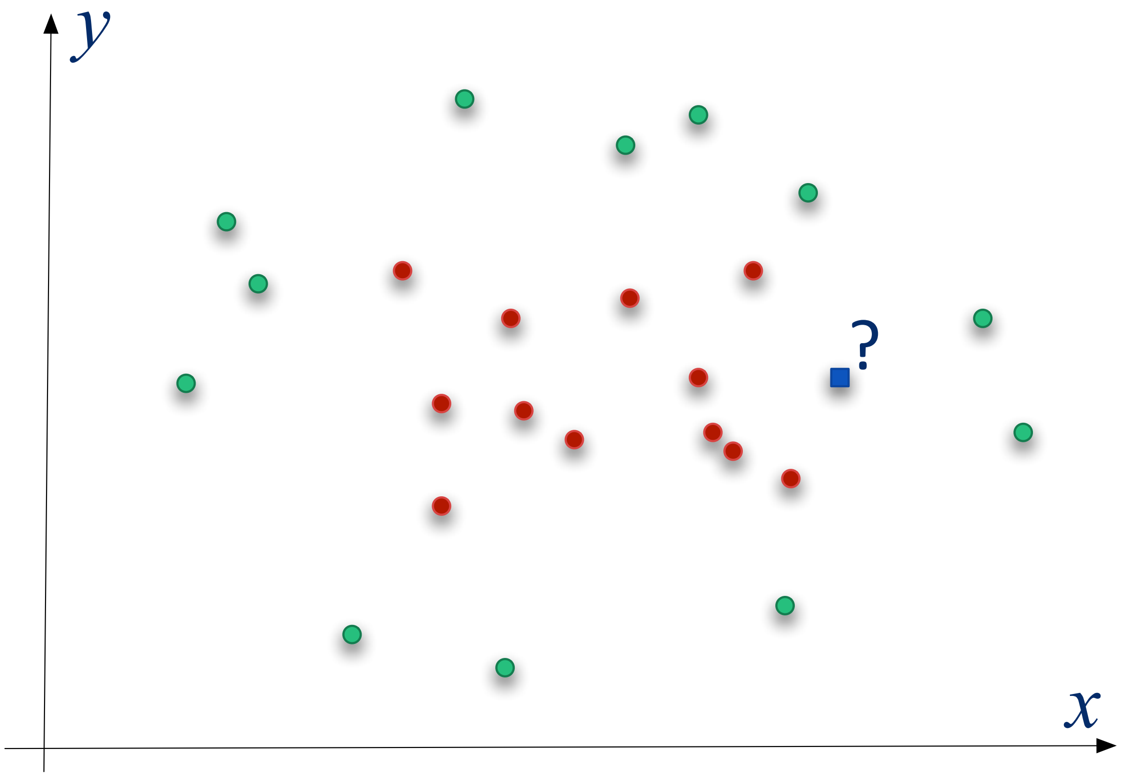
Target class: rectangles in R^2

- Sample
  - Positive instances $\mathbf{P}^{+}_{\mathcal{X}}$
  - Negative instances $\mathbf{P}^{-}_{\mathcal{X}}$



Target class: unknown

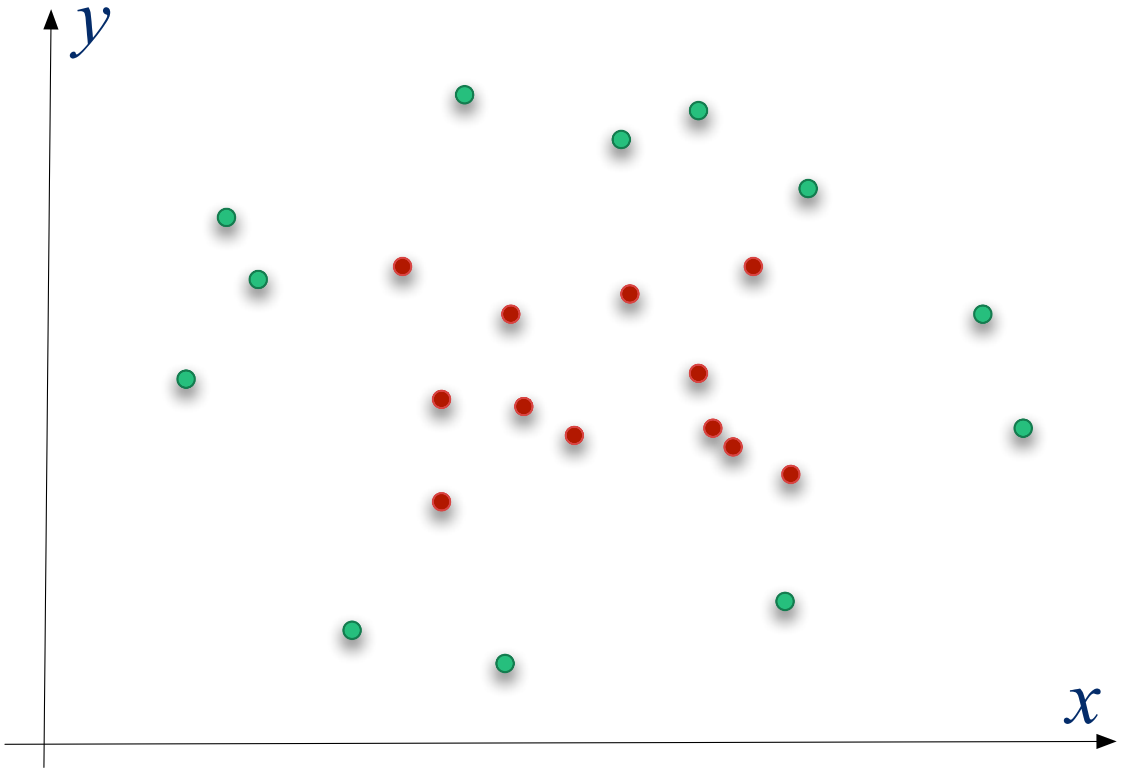
- What do we want to learn?



→ A **decision** fonction (**prediction**)

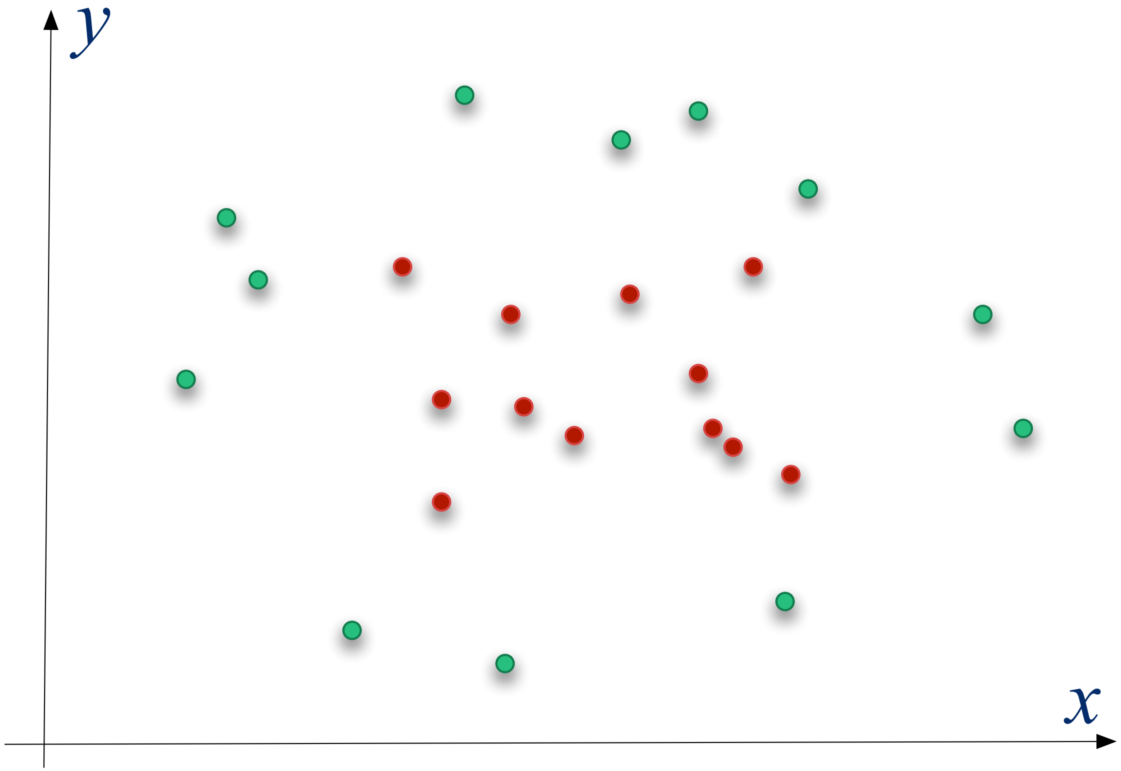
Target class: unknown

- How to learn?



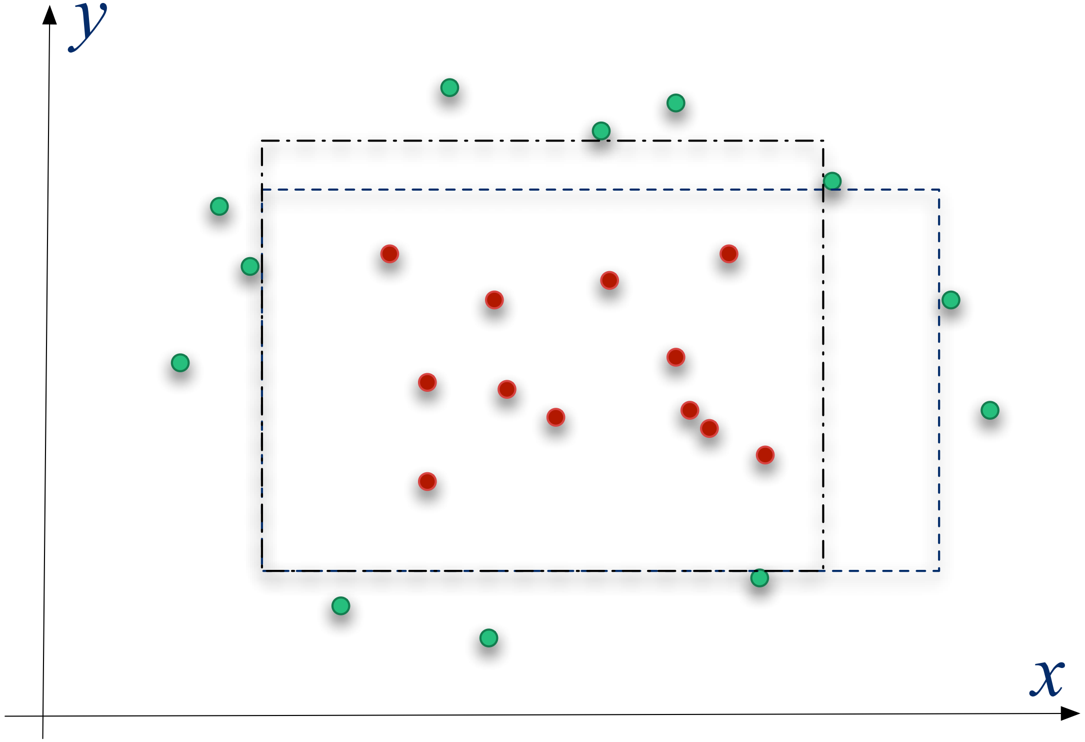
Target class: rectangles in R^2

- How to learn?
  - If **I know that the target concept is a rectangle**



Target class: rectangles in R^2

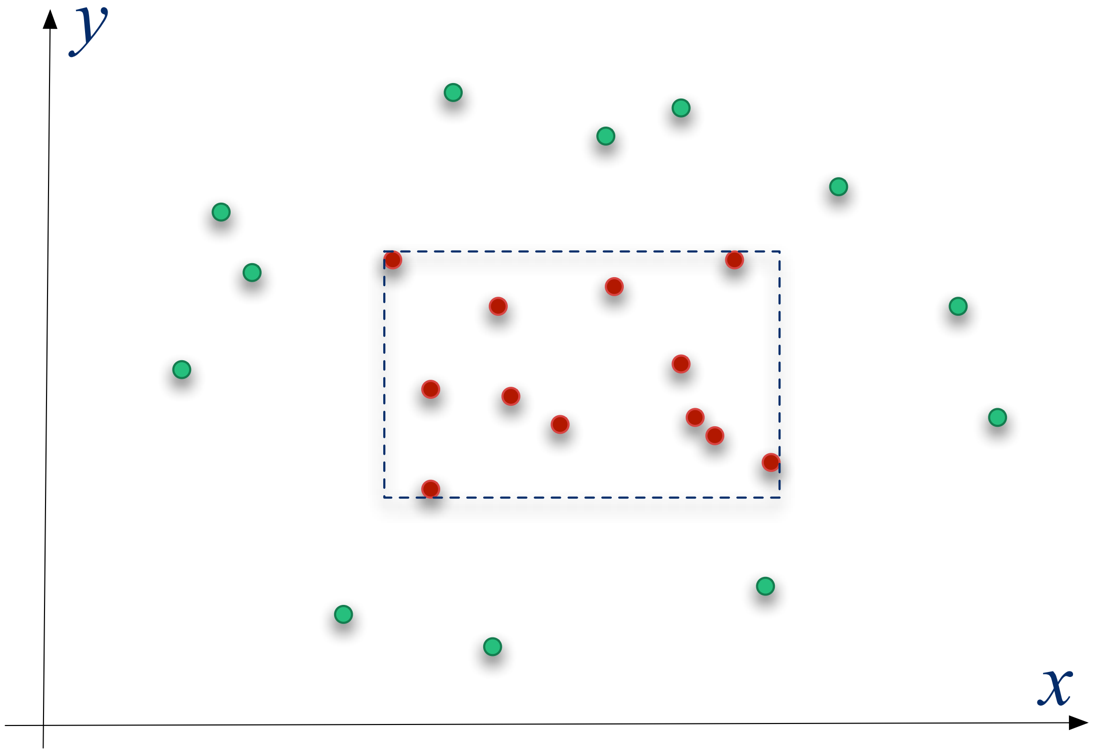
- How to learn?
  - If **I know that the target concept is a rectangle**



Most **general** hypotheses

Target class: rectangles in R^2

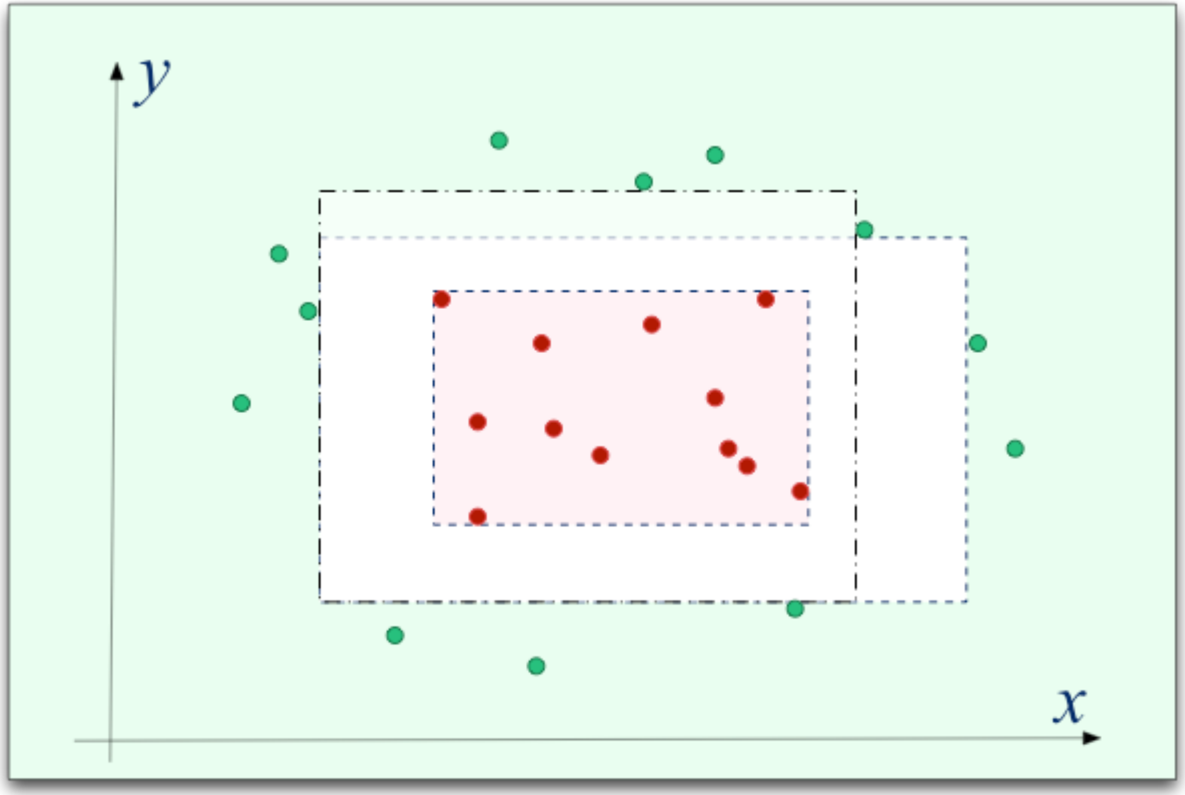
- How to learn?
  - If **I know that the target concept is a rectangle**



Most **specific** hypotheses

Target class: rectangles in R^2

- How to learn?
  - Choice of one hypothesis *h*

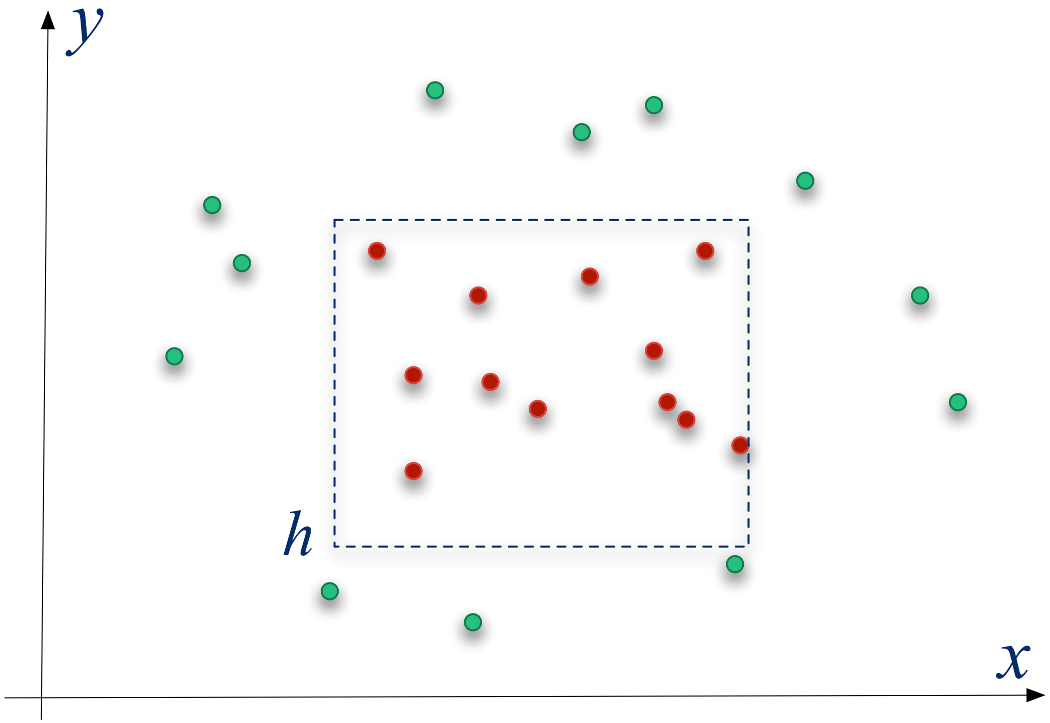


Version

space

Target class: rectangles in R^2

- Learning: choice de *h*
  - Which performance to expect?



The statistical theory of learning

Which performance ?

- Cost for a prediction error
  - The ***loss function***

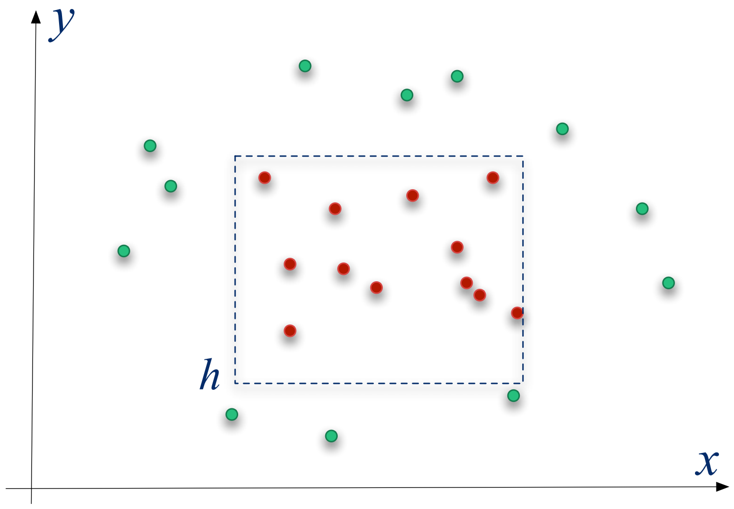
$$\ell\big(h(\mathbf{x}), y\big)$$

- Which **expected cost** if I choose *h*?
  - The « real ***risk*** » (or true risk)

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

The statistical theory of learning

- Which expected cost when *h* is chosen?
  - Assuming that there is no training error on *S*

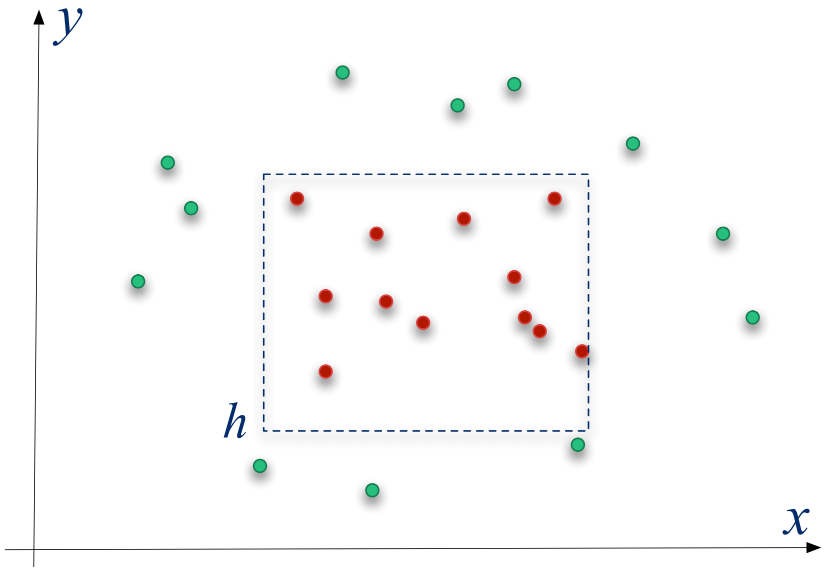


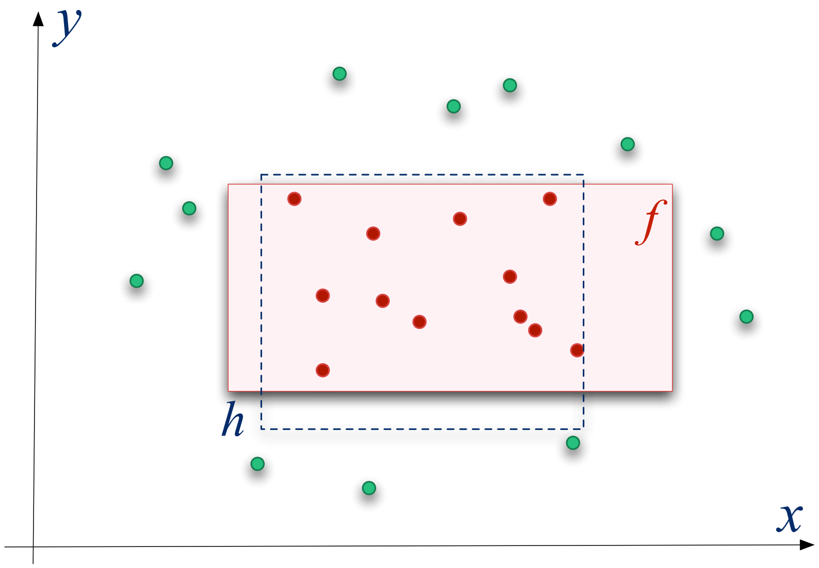
The « empirical risk »

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

Statistical theory of learning: the ERM

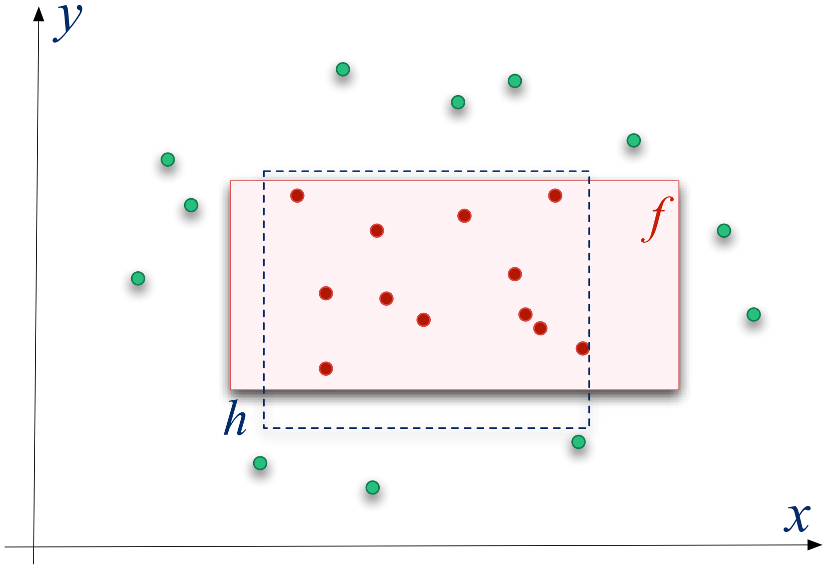
- **Learning strategy**:
  - Select an hypothesis with null empirical risk (no training error)
  - Which generalization performance to expect for *h* ?

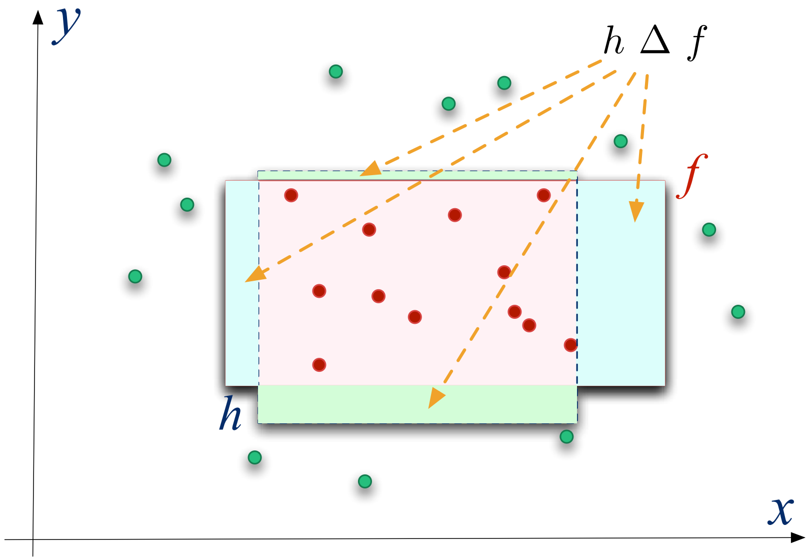




Statistical theory of learning: the ERM

- Select an hypothesis with null empirical risk (no training error)
- Which generalization performance to expect for *h* ?
- What is the **risk of getting error *R*(*h*) >** ε ?





Central interrogation: the inductive principle

- The empirical risk minimization principle (ERM)

  … is it sound?

  - If I chose *h* such that $$\hat{h} \; = \; \underset{h \in \mathcal{H}}{\text{ArgMin}} \; \hat{R}(h)$$

  - Is *h* good with respect to the real risk?

$$\hat{R}(\hat{h}) \; \overset{?}{\longleftrightarrow} \; R(\hat{h})$$

  - Could I have done much better?

$$h^* \; = \; \underset{h \in \mathcal{H}}{\text{ArgMin}} \; R(h)$$

$$R(h^*) \overset{?}{\longleftrightarrow} R(\hat{h})$$

AgroParisTech

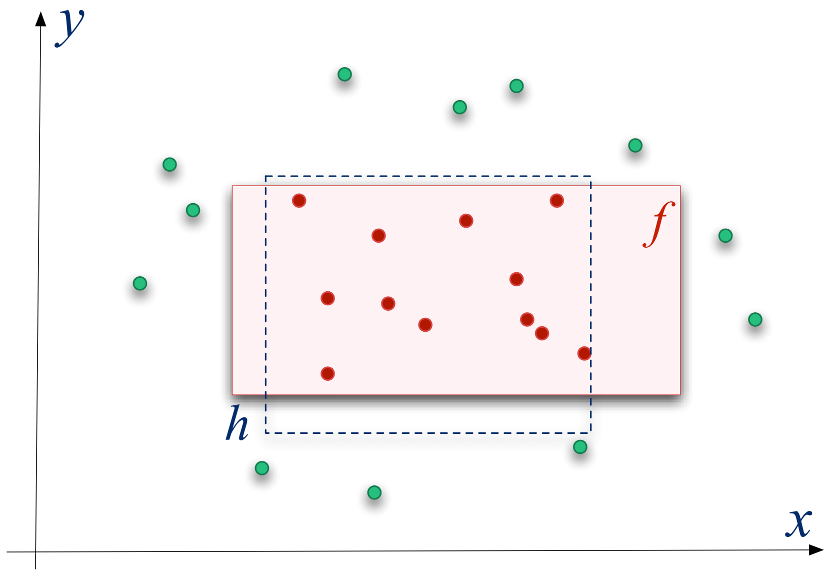
The **statistical theory** of learning

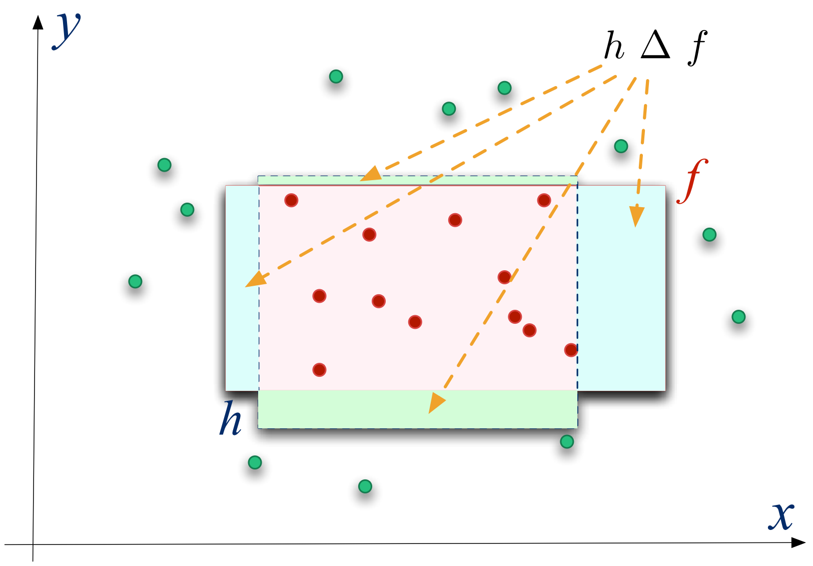
The 1er step

One hypothesis

Statistical study for ONE hypothesis

- Chose one hypothesis of nul empirical risk (no error on the training set *S*)
- Which performance can we expect for *h*?
- What is the risk of having *R*(*h*) > ε?





Statistical study for ONE hypothesis

- Assume that *h* st. $R(h) \geq \varepsilon$ (*h* is « bad »)
- What is the probability that nonetheless *h* have been selected?

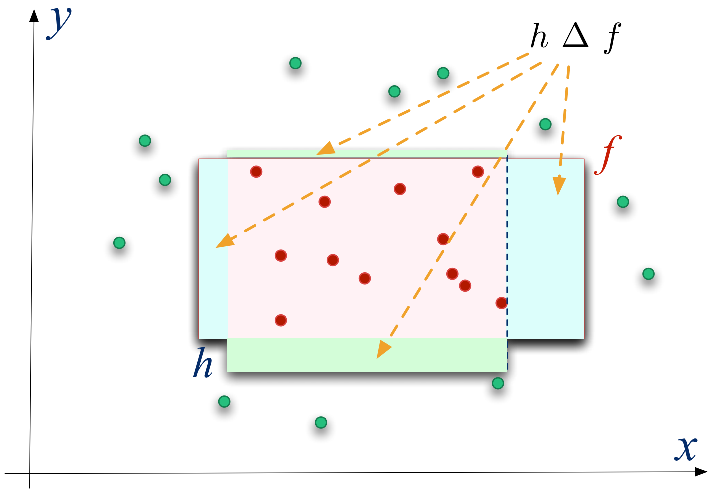
$$R(h) = \mathbf{p}_{\mathcal{X}}(h \,\Delta\, f)$$

After one example : $p\big(\hat{R}(h) = 0\big) \leq 1 - \varepsilon$

« falls » outside $h \,\Delta\, f$

After *m* examples (**i.i.d.**) :

$$p^m\big(\hat{R}(h) = 0\big) \leq (1 - \varepsilon)^m$$



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

AgroParisTech

Statistical study for ONE hypothesis

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

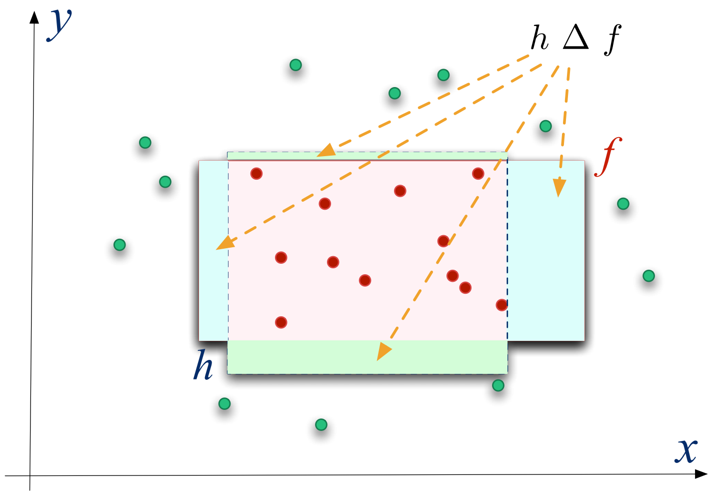
Or:

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

$$< \quad e^{-\varepsilon\, m} \;\leq\; \delta$$

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

Hence : $\boxed{m \;\geq\; \dfrac{\ln(1/\delta)}{\varepsilon}}$



The statistical theory of learning

- For **any hypothesis** chosen when observing *S*
- What we really want:

"Realizable case"

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

- Let's assume: $|\mathcal{H}| \;<\; \infty$

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

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

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

The **statistical theory** of learning

The 2nd step

Which hypothesis in the crowd

Statistical study for $|\mathcal{H}|$ hypotheses

- What is the probability that I chose one hypothesis h_{err} of real risk > ε **and that I do not realize it** after *m* examples?
- Probability of survival of ***h*~err~ after 1 example** : $(1-\varepsilon)$
- Probability of survival of ***h*~err~ after *m* examples** : $(1-\varepsilon)^m$
- Probability of survival of **at least one hypothesis in $\mathcal{H}$** : $|\mathcal{H}|\,(1-\varepsilon)^m$
  - We use the probability of the union $\mathbf{P}(A \cup B) \;\leq\; \mathbf{P}(A) + \mathbf{P}(B)$
- We want that the probability that there remains at least one hypothesis of real risk > ε in the version space be bounded by δ :

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

$$\log|\mathcal{H}| - \varepsilon\, m \;<\; \log\delta$$

$$\boxed{m \;>\; \frac{1}{\varepsilon}\,\log\frac{|\mathcal{H}|}{\delta}}$$

The « PAC learning » analysis

- We get:

$$\forall h \in \mathcal{H}, \forall \delta \leq 1 : \quad \mathbf{P}^m \left[R_{\text{Réel}}(h) \; \leq \; \underbrace{R_{\text{Emp}}(h)}_{= 0} \; + \; \overbrace{\frac{\log |\mathcal{H}| + \log \frac{1}{\delta}}{m}}^{\varepsilon} \right] \; > \; 1 - \delta$$

Realizable case: there exists at least one function *h* of risk 0

The Empirical Risk Minimization principle

is sound only if there are **constraints on the hypothesis space**

AgroParisTech

- ATTENTION :

  – This analysis makes a **big assumption** about the relation between the “past” and the “future”

- The world is **stationnary**
  - The training examples (“ past”)
    and the **test examples** (“future”) follow the same distribution
  - The training and test examples are i.i.d.

PAC learning: definition

[Valiant, 1984]

Given $0 < \delta, \varepsilon < 1$, a *concept class* C is *learnable* by a polynomial time algorithm A if,
for any distribution P of samples and any concept $c \in C$,
there exists a polynomial $p(\cdot, \cdot, \cdot)$ such that
A will produce with probability at least $1 - \delta$ a hypothesis $h \in C$ whose error is $\leq \varepsilon$
when given at least $p(m, 1/\delta, 1\varepsilon)$ independent random examples drawn according to P.

- **Worst case** analysis
  - Against **all distributions** *P*
  - For **any target hypothesis** in a class of hypotheses
- Notion of *computational complexity*

The statistical theory of learning

Uniform convergence bounds

(for the **unrealizable case**)

Generalizing the law of large numbers: uniform convergence

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

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

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

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

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

« $\mathcal{H}$ *finite* »

$$\begin{aligned} P^m[\exists h \in \mathcal{H} : R_{\text{Réel}}(h) - R_{\text{Emp}}(h) > \varepsilon] \;&\leq\; \sum_{i=1}^{|\mathcal{H}|} P^m[R_{\text{Réel}}(h^i) - R_{\text{Emp}}(h^i) > \varepsilon] \\ &\leq\; |\mathcal{H}|\exp(-2\,m\,\varepsilon^2) \;=\; \delta \end{aligned}$$

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

Bounding the true risk with the empirical risk + …

- $\mathcal{H}$ finite, realizable case

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

- $\mathcal{H}$ finite, **non** realizable case

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

To sum up: for $|\mathcal{H}|$ finite

- **Non** realizable case

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

- **Realizable** case

$$\varepsilon \;=\; \frac{\log|\mathcal{H}| + \log\frac{1}{\delta}}{m}$$ and $$m \;\geq\; \frac{\log|\mathcal{H}| + \log\frac{1}{\delta}}{\varepsilon}$$

$|\mathcal{H}|$ infinite !!

- Effective dimension of $\mathcal{H}$ = *the* ***Vapnik-Chervonenkis*** *dimension*
  - **Combinatorial criterion**
  - Size of the largest set of points (in general configuration) that can be labeled in any way by hypotheses drawn from $\mathcal{H}$

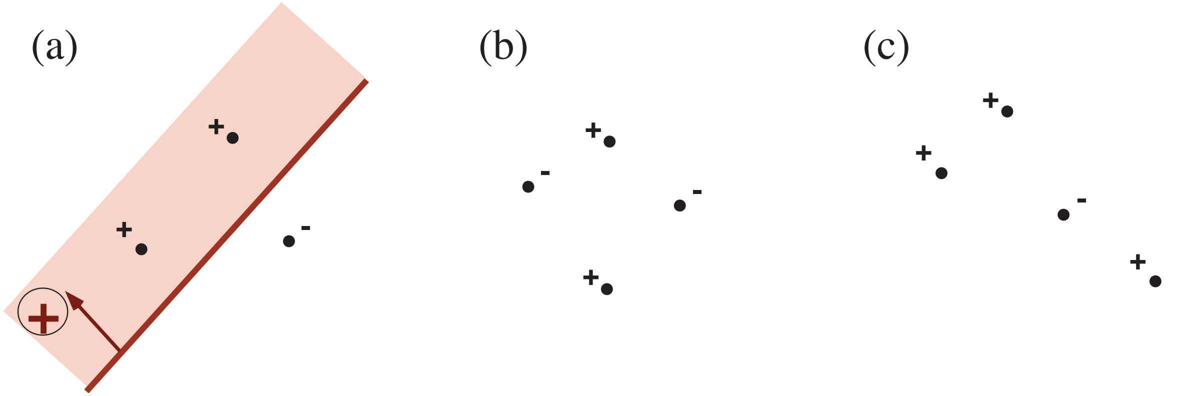
$$d_{VC}(\mathcal{H}) \;=\; \max\{m \;:\; \Pi_{\mathcal{H}}(m) \;=\; 2^m\}$$

Bound on the true risk

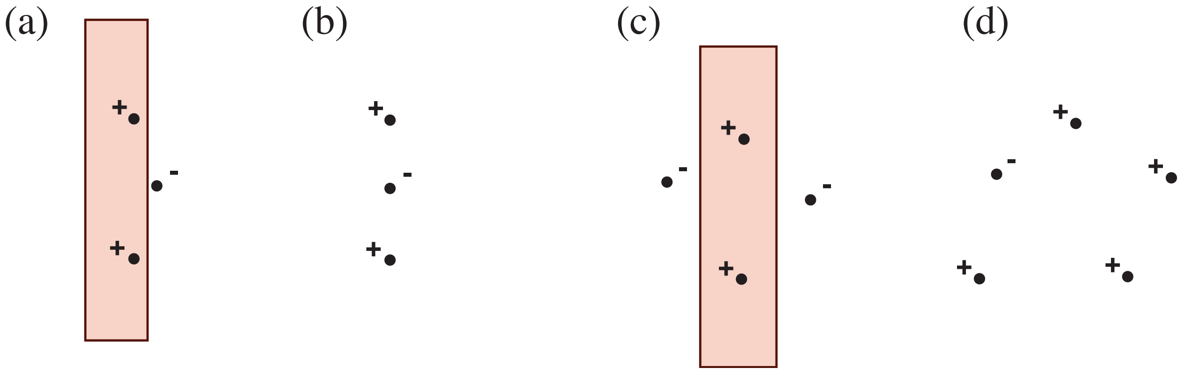
$$\forall h \in \mathcal{H}, \forall \delta \leq 1 : \quad P^m \left[R_{\text{Réel}}(h) \;\leq\; R_{\text{Emp}}(h) \;+\; \sqrt{\frac{8\, d_{VC}(\mathcal{H}) \log \frac{2\,e\,m}{d_{VC}(\mathcal{H})} + 8 \log \frac{4}{\delta}}{m}} \right] \;>\; 1 - \delta$$

VC dim: illustrations

- d_{VC}(linear separator) = ?



- d_{VC}(rectangles) = ?



Lesson

- You **cannot guarantee** anything about induction
- **Even if** you assume that the world is stationary

  and examples are i.i.d.
- Unless there are (severe) **constraints on the hypothesis space**

But wait … ?

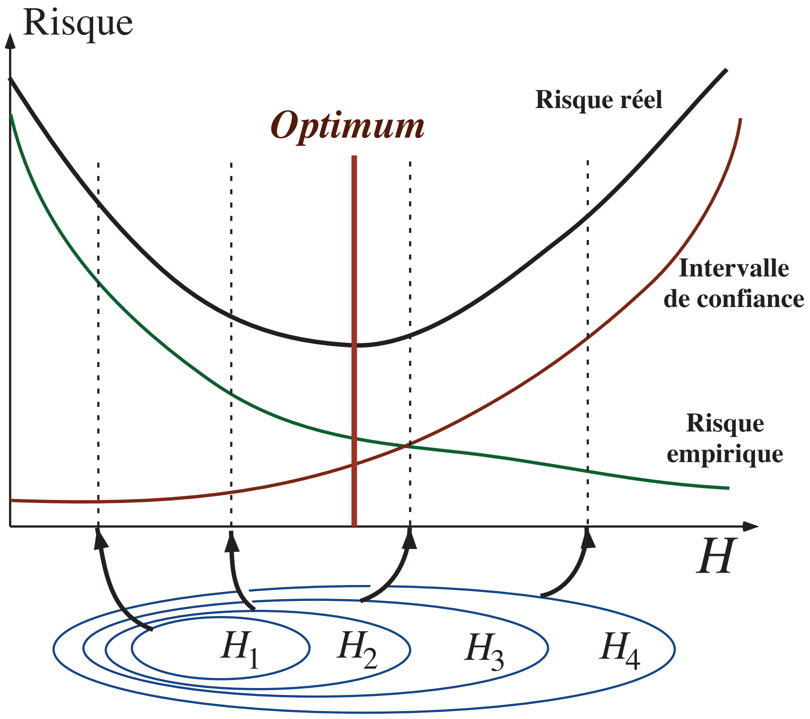
The **statistical theory** of learning

The 3rd step

Which hypothesis space?

SRM : Structural Risk Minimization

- **Stratification** of the hypotheses spaces
  - Determined *a priori* (independently of the data)
  - Using for instance the d_{VC}

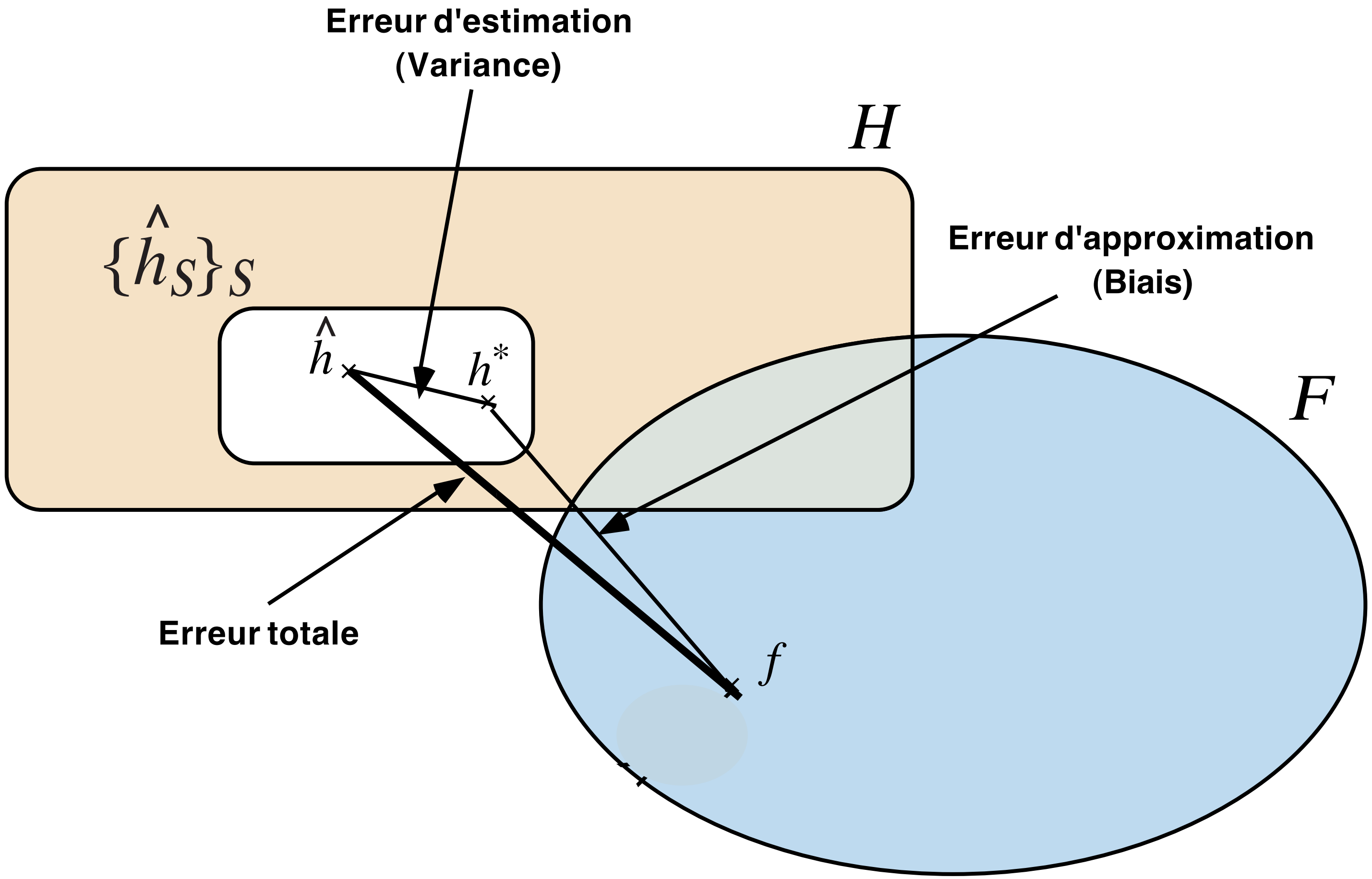


The « PAC learning » or statistical analysis

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

- *New inductive criteria*:
  - The **regularized empirical risk**
    1. Satisfy as well as possible the constraints imposed by the **training examples**
    2. Choose the best **hypothesis space** (capacity of *H*)

The bias-variance tradeoff



...

Learning becomes …

1. The **choice** of the hypothesis space ***H***
   - Which is constrained by necessity

2. The **choice** of an **inductive criterion**
   - Empirical Risk which must be regularized

3. An **exploration strategy for *H*** in order to minimize the regularized empirical risk
   - It must be efficient
     - Fast
     - With only one optimum if possible (e.g. convex problem)

Outline of today's class

1. The mystery of in-distribution learning (standard induction)

2. A 101 course on the statistical learning theory

3. Why does it **fail** to account for **deep neural networks**?

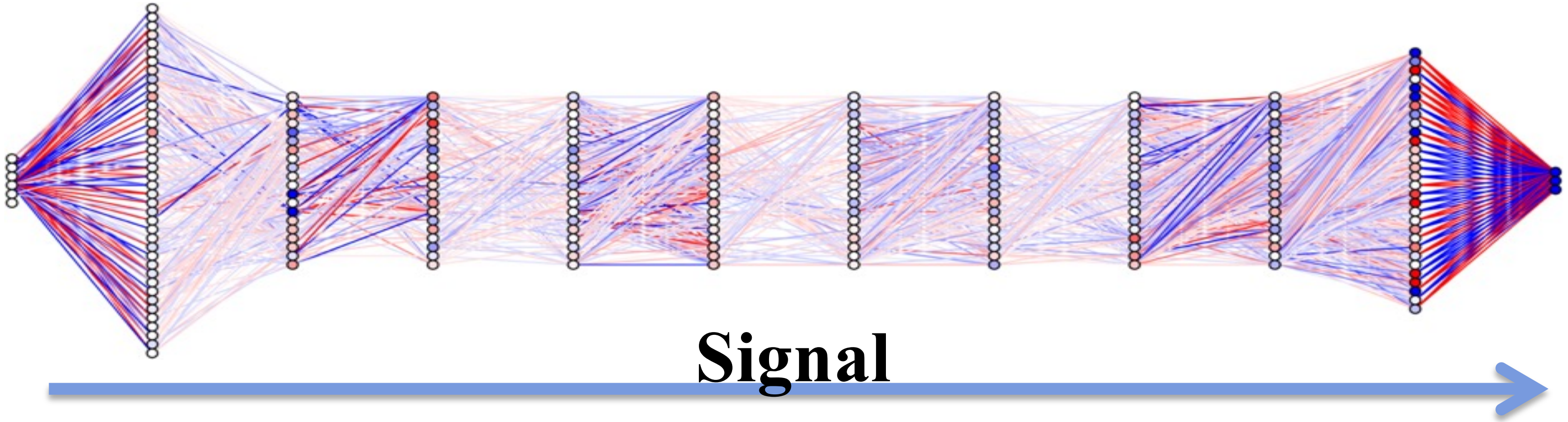
4. The **no-free-lunch** theorem

The SuperVision network

Image classification with deep convolutional neural networks

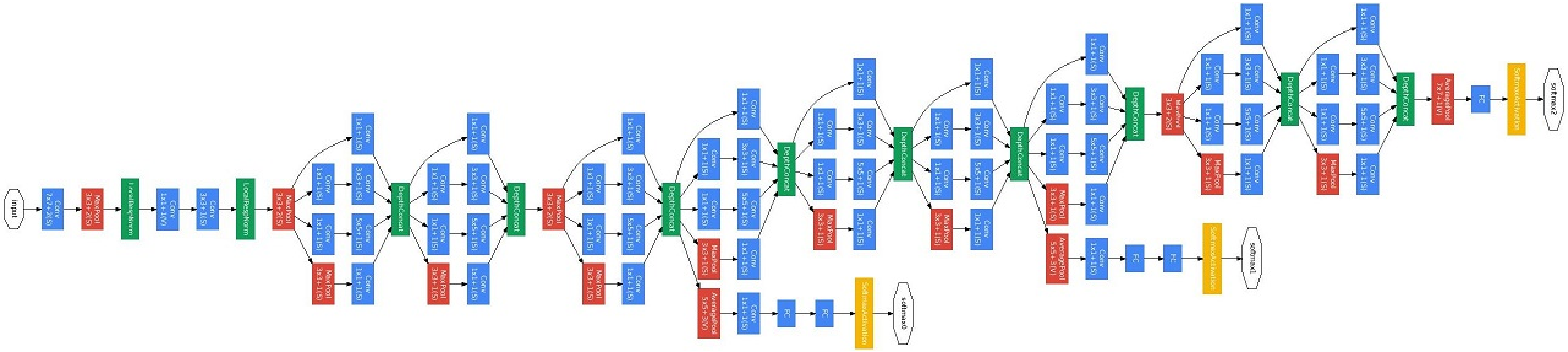
http://image-net.org/challenges/LSVRC/2012/supervision.pdf

- 7 hidden “weight” layers
- 650K neurons
- **60M** parameters
- 630M connections



GoogleNet

- A **mécano** of neural networks



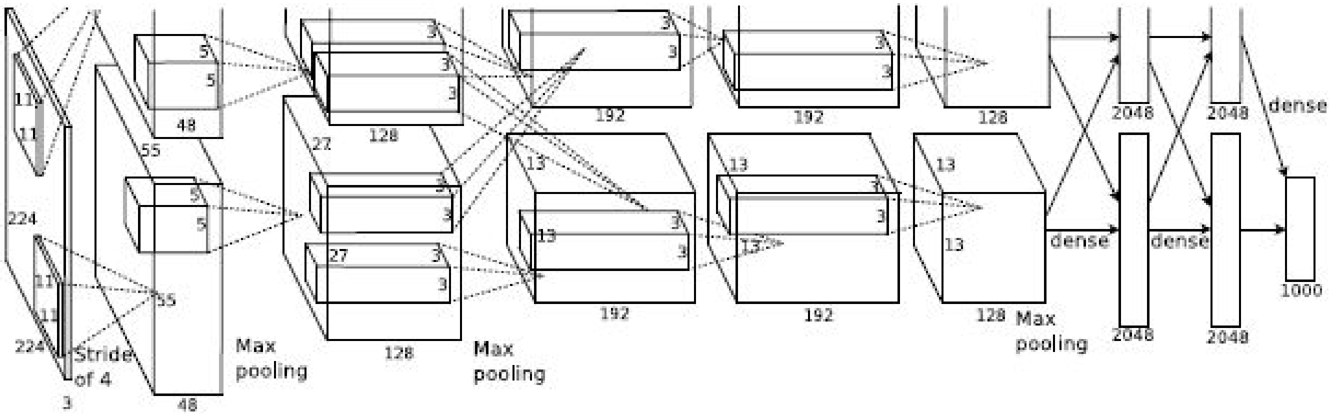
Troubling findings

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

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

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!!

AgroParisTech

Troubling findings

- Deep NNs can accommodate ANY training set

Can grow without limit!!

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

But then,

why are deep NNs so good on image classification tasks?

Alternative explanations?

- See for example Nati Srebro

https://www.youtube.com/playlist?list=PLGJm1x3XQeK0gmqfRkP-VmrEf4UYx5IDW&pbjreload=101

- The search bias would conduct the algorithm to first explore simple (?) hypotheses

Alternative explanations?

- See also explanations that stem from the information bottleneck principle (Naftali Tishby et al.)

  (several papers in ICLR-2020)

Which garantees exactly?

Statistical learning: which garantees?

- Link between **empirique** risk and **real** risk
  - Cost of using *h* (e.g. error rate)

Says **nothing** on:

- Intelligibility
- Fruitfulness
- Place in a domain theory

- **Valid only if**
  - Stationary environment
  - **Examples** i.i.d.
  - **Questions** i.i.d. !!?

Limits

- **Passive** learning and **data and questions** supposedly **i.i.d.**
  - Situated agents: the world is not i.i.d. when you are acting in it
- Needs **a lot of** traning examples
  - We are far more efficient
  - We cannot help but « produce theories » constantly, testing them afterwards
- Not adapted to the search for **causality** relationships
- Not **integrated** with **reasoning**

Those **learning** machines are not **thinking** machines

Outline of today's class

1. The mystery of in-distribution learning (standard induction)

2. A 101 course on the statistical learning theory

3. Why does it fail to account for deep neural networks?

4. The **no-free-lunch** theorem

The no-free-lunch theorem

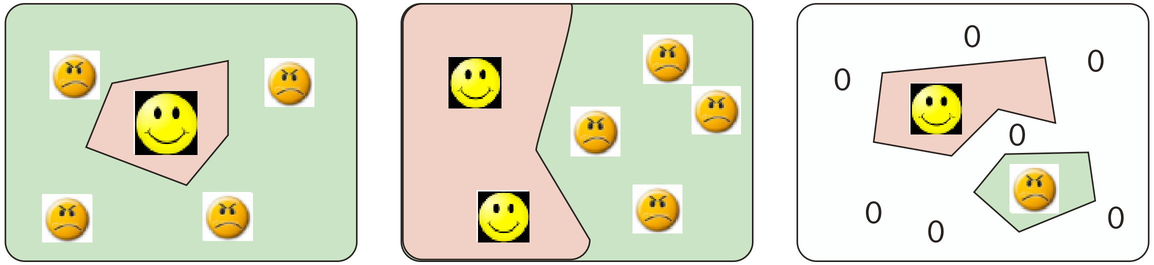
Théorème 2.1 (No-free-lunch theorem (Wolpert, 1992))

Pour tout couple d'algorithmes d'apprentissage $\mathcal{A}_1$ et $\mathcal{A}_2$, caractérisés par leur distribution de probabilité a posteriori $\mathbf{p}_1(h|\mathcal{S})$ *et* $\mathbf{p}_2(h|\mathcal{S})$*, et pour toute distribution $d_{\mathcal{X}}$ des formes d'entrées* $\mathbf{x}$ *et tout nombre m d'exemples d'apprentissage, les propositions suivantes sont vraies :*

1. *En moyenne uniforme sur toutes les fonctions cible f dans $\mathcal{F}$:* $\mathbb{E}_1[R_{\text{Réel}}|f,m] \; - \; \mathbb{E}_2[R_{\text{Réel}}|f,m] \; = \; 0.$
2. *Pour tout échantillon d'apprentissage $\mathcal{S}$ donné, en moyenne uniforme sur toutes les fonctions cible f dans $\mathcal{F}$:* $\mathbb{E}_1[R_{\text{Réel}}|f,\mathcal{S}] \; - \; \mathbb{E}_2[R_{\text{Réel}}|f,\mathcal{S}] \; = \; 0.$
3. *En moyenne uniforme sur toutes les distributions possibles* $\mathbf{P}(f)$ *:* $\mathbb{E}_1[R_{\text{Réel}}|m] \; - \; \mathbb{E}_2[R_{\text{Réel}}|m] \; = \; 0.$
4. *Pour tout échantillon d'apprentissage $\mathcal{S}$ donné, en moyenne uniforme sur toutes les distributions possibles* $\mathbf{p}(f)$ *:* $\mathbb{E}_1[R_{\text{Réel}}|\mathcal{S}] \; - \; \mathbb{E}_2[R_{\text{Réel}}|\mathcal{S}] \; = \; 0.$

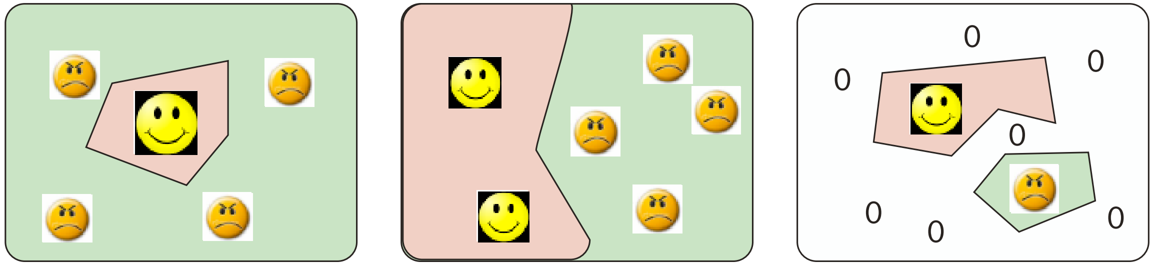
The no-free-lunch theorem

Possible

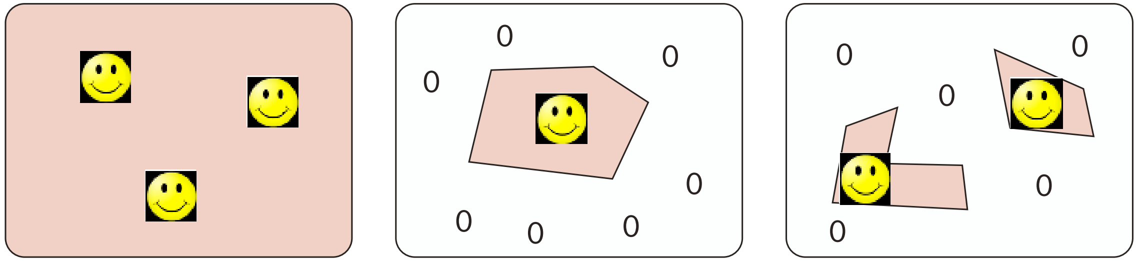


The no-free-lunch theorem

Possible



Impossible



Deduction!

1. **All** inductive **learning algorithms are equivalent** (to a random guessing one)

2. There cannot be any **guarantees** on the **inductions** made

Let’s go to the beach or skying!!

Lesson

- (Quasi) guarantees about the results
  - **If** the signal actually presents the properties **assumed a priori**
  - **Then** the method ensures that learning **using this bias**

    will converge to the target function

    if enough (i.i.d.) data is available

« Lampost » theorems

Conclusions

How can we prove the **validity**

of a new inductive principle?

Conclusions

1. Induction, which is at the **center** of learning is a **under-constrained problem**.

2. There **cannot be** a validation of induction unrelated to the domain

3. Guarantees cannot **only** be obtained **by making assumptions** about the world
   - E.g. i.i.d. data and queries and a bias

- A **theory** of **induction** aims at
  - Proposing reasonable **meta-assumptions**
    - *E.g. the world is **stationary** and the **data** and **queries** are **i.i.d.***
  - Providing a **formal framework** where “lampost theorems” can be obtained
    - ***If*** *the data obeys the assumptions about the world*
      Then *it is possible to PAC guarantee that …*

Conclusions: the statistical theory of learning

Performance measured :

the expectation of the cost of using the learned hypothesis

(i.e. but **no** concern for causality, intelligibility, the articulation with reasoning, …)

Only valid if **stationary environment** + i.i.d. **data** + i.i.d. **queries**

How to select a bias?

Conclusions: “new scenarios” are out of the statistical box

- Very **few data points**
  - Very often, we learn with **very little data**

- **Past history** plays a role: education (curriculum)
  - **Sequence** effects

- We learn in order to and because we (constantly) **construct theories**
  - Both at the **micro** and the **macro** level

References

V. Barra & A. Cornuéjols & L. Miclet

Apprentissage Artificiel. Concepts et algorithmes. De Bayes et Hume au deep learning

Eyrolles (4° éd.), 2021

M. Mohri & A. Rostamizadeh & A. Talwalkar

Foundations of Machine Learning

MIT Press, 2012

S. Shalev-Shartz & S. Ben-David

Understanding Machine Learning

Cambridge University Press, 2044