Does the brain plays a role in Artificial Intelligence

Antoine **Cornuéjols**

AgroParisTech – INRAé

antoine.cornuejols@agroparistech.fr

AgroParisTech
Talents d'une planète soutenable

INRAE

Outline

1. Has AI been **bio-inspired**?

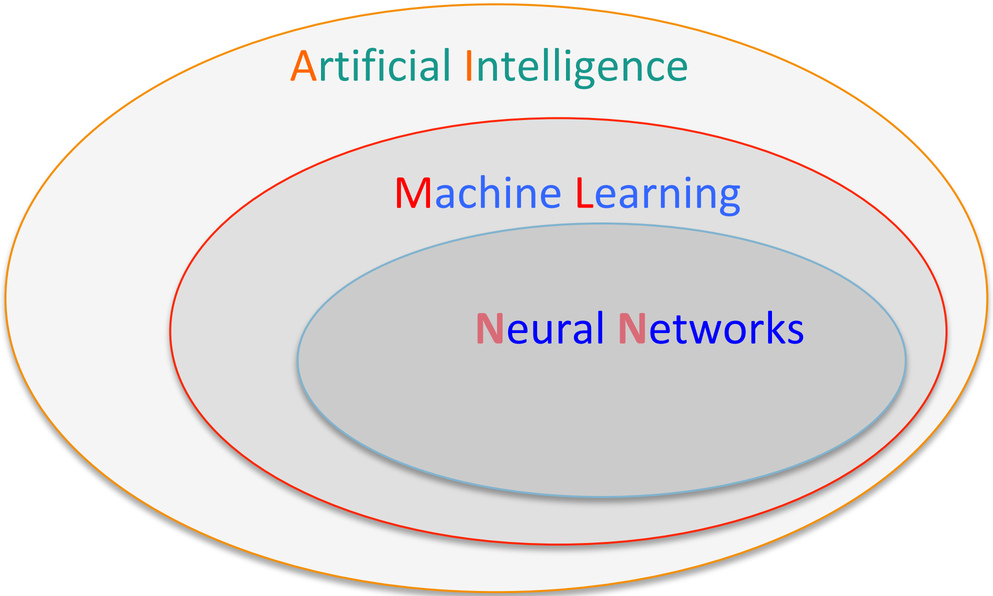
2. Interfacing AI with Humans

3. Interfacing AI with AI

4. Conclusion

The central place of **neurons** in AI! ??

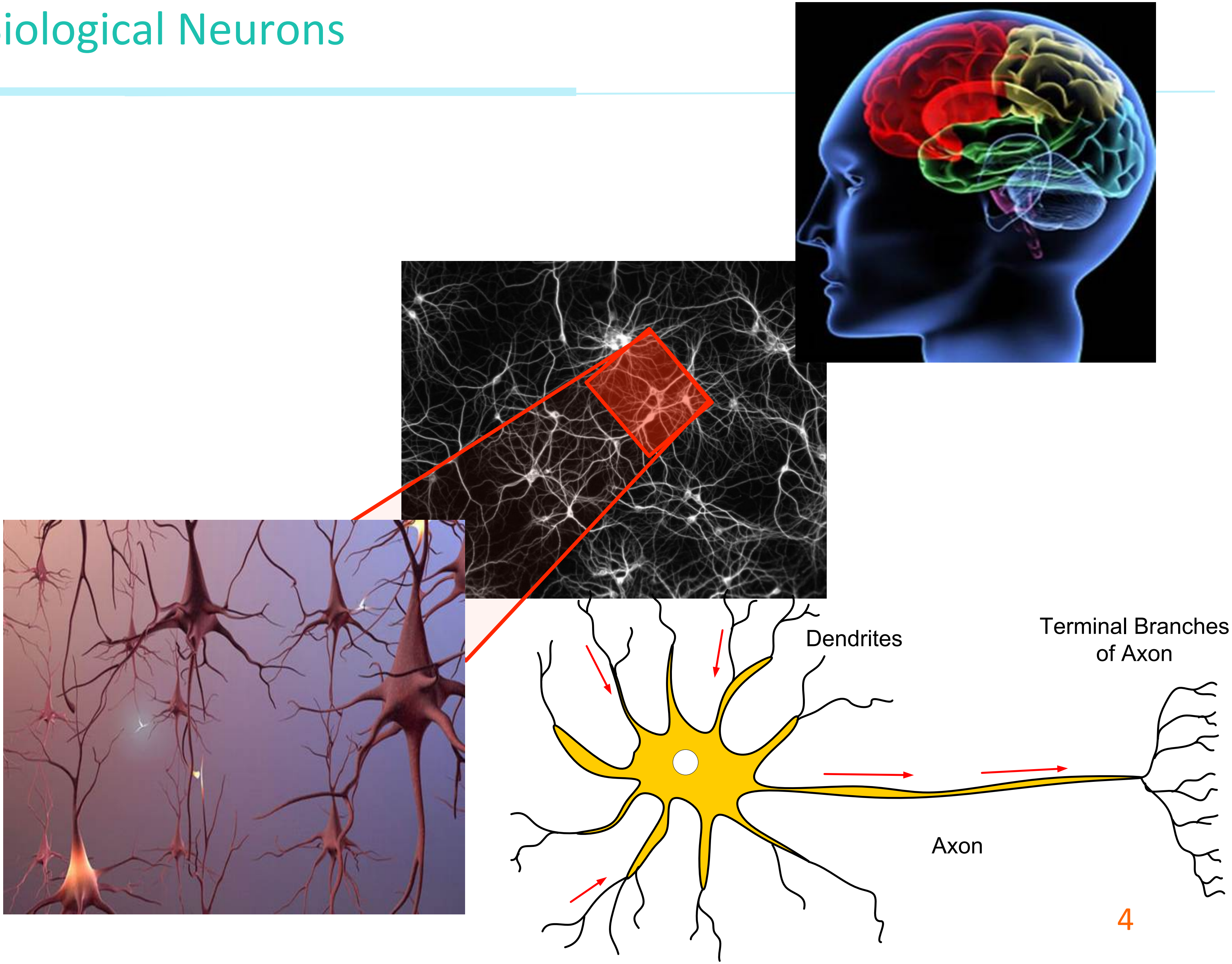
Artificial Intelligence = Machine Learning = **Deep** learning (**N**eural **N**etworks)



In "Artificial Neural Networks"

there is "**Neural Networks**"

Biological Neurons



Neural Networks?

- 1943 = the crucial year for AI

Neural Networks?

- 1943 = the crucial year for AI

A LOGICAL CALCULUS OF THE IDEAS IMMANENT IN NERVOUS ACTIVITY*

■ WARREN S. MCCULLOCH AND WALTER PITTS
University of Illinois, College of Medicine,
Department of Psychiatry at the Illinois Neuropsychiatric Institute,
University of Chicago, Chicago, U.S.A.

Neural Networks?

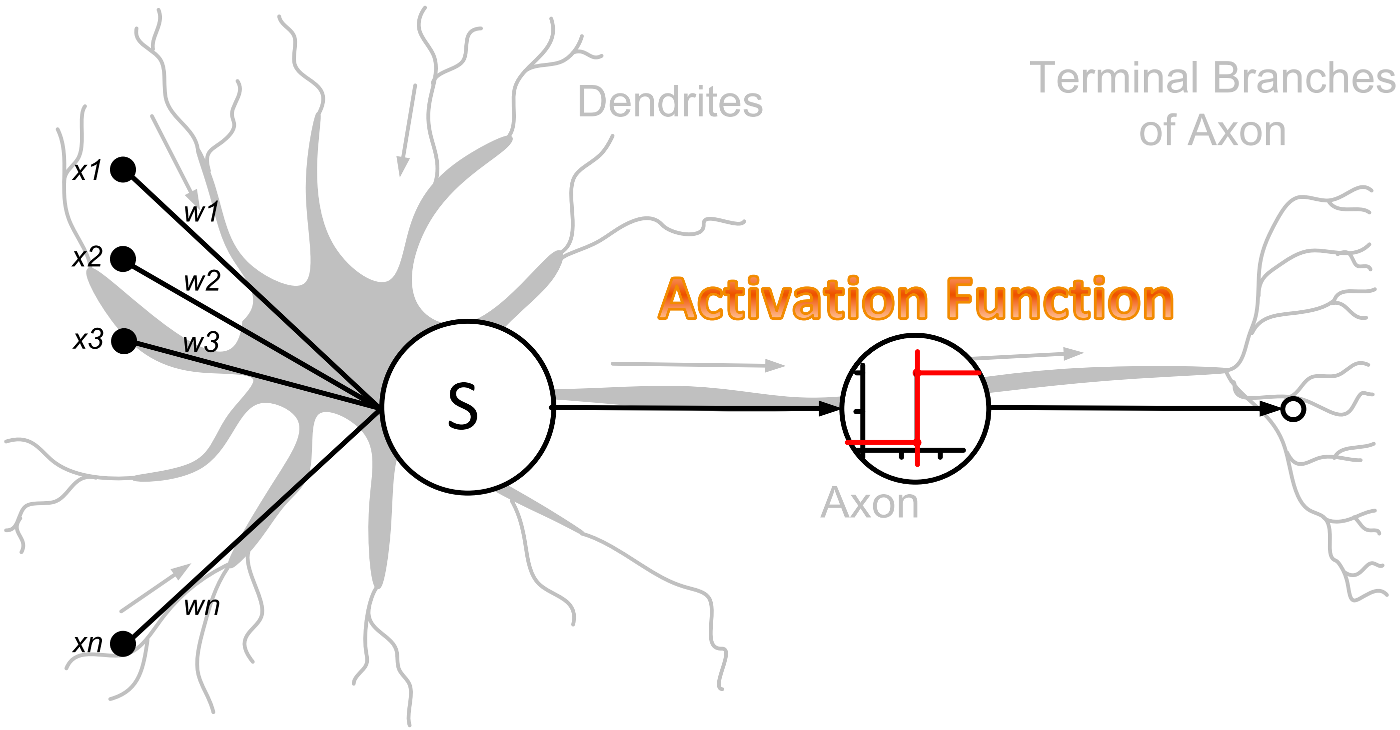
- 1943 = the crucial year for AI

A LOGICAL CALCULUS OF THE IDEAS IMMANENT IN NERVOUS ACTIVITY*

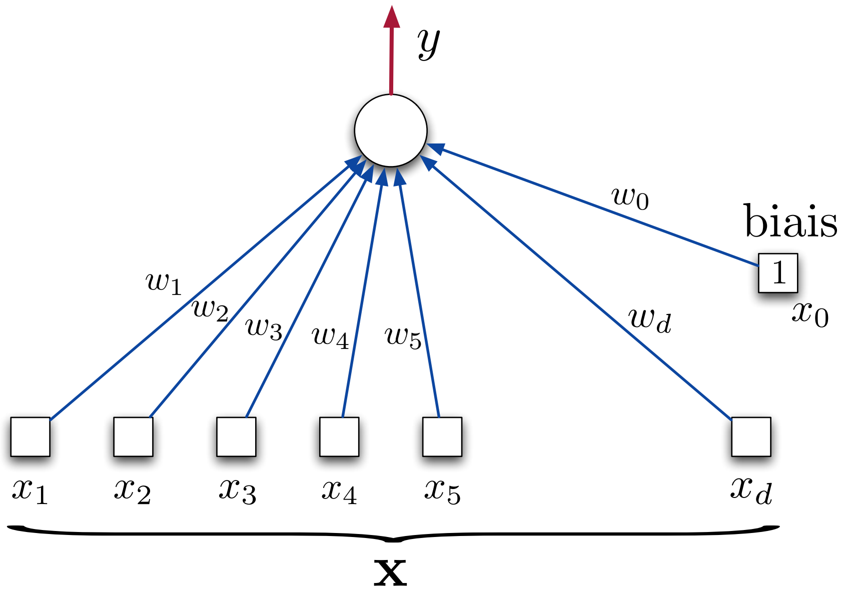
■ WARREN S. MCCULLOCH AND WALTER PITTS
University of Illinois, College of Medicine,
Department of Psychiatry at the Illinois Neuropsychiatric Institute,
University of Chicago, Chicago, U.S.A.

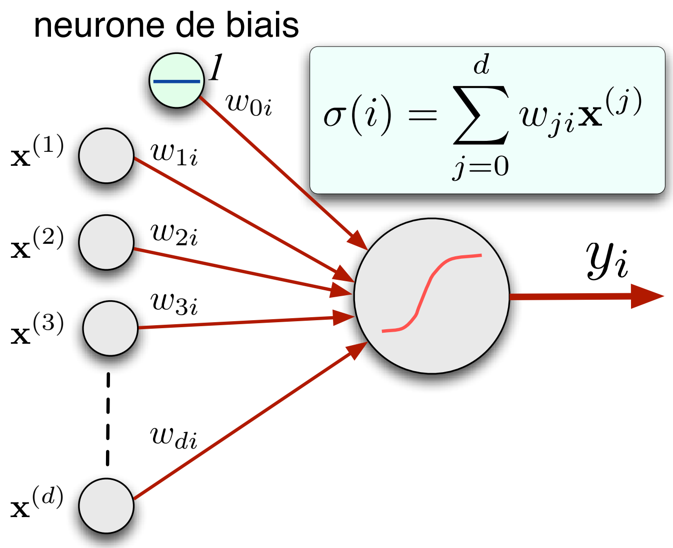
Many years ago one of us, by considerations impertinent to this argument, was led to conceive of the response of any neuron as factually equivalnt to a proposition which proposed its adequate stimulus. He therefore attempted to record the behavior of complicated nets in the notation of the symbolic logic of propositions. The "all-or-none" law of nervous activity is sufficient to insure that the activity of any neuron may be represented as a proposition.

Artificial Neural Networks (ANN)

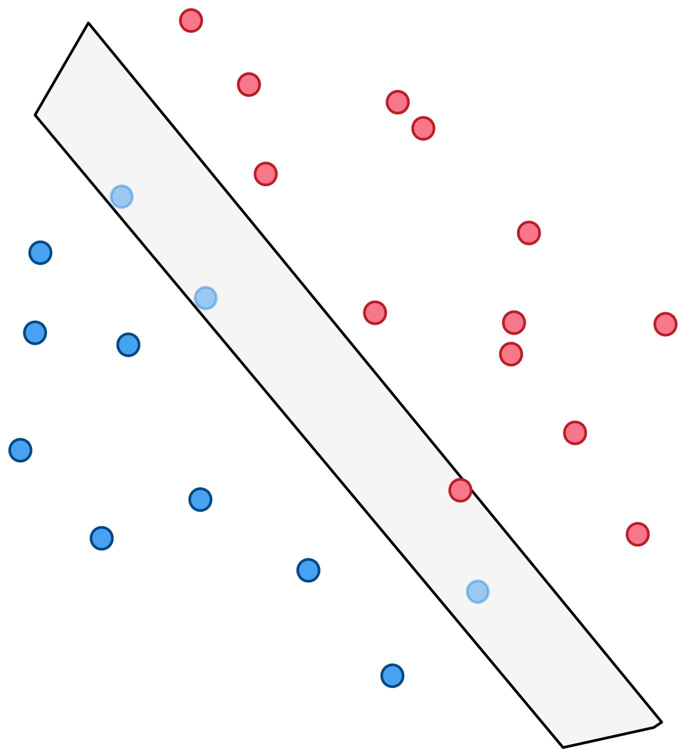


The perceptron

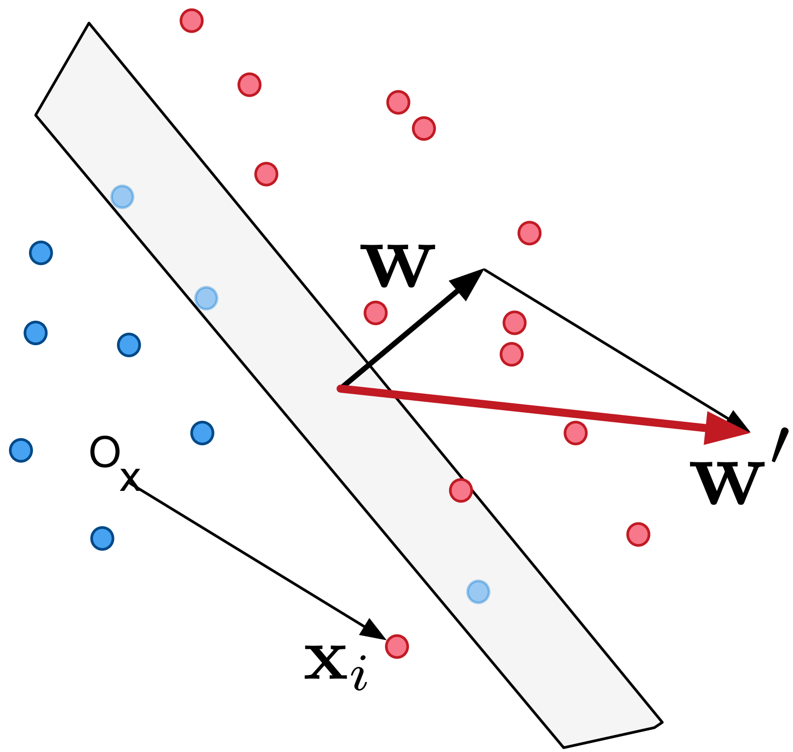




The perceptron: a **linear discriminant**



The perceptron **learning** algorithm: intuition

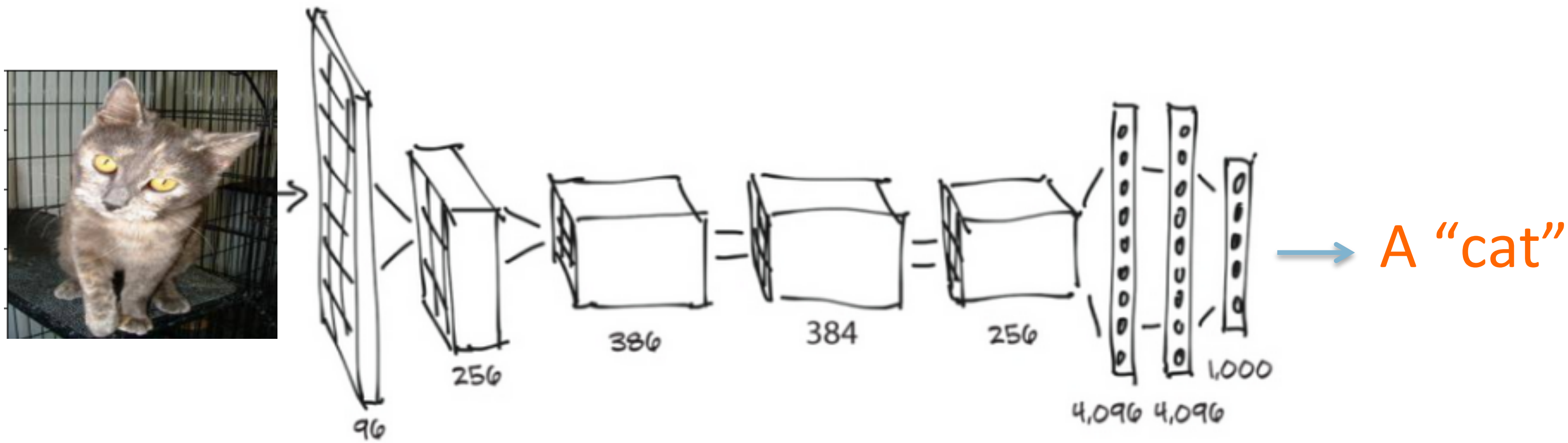


$$\mathbf{w}' = \mathbf{w} + \eta\, y_i\, \mathbf{x}_i$$

Deep Neural Networks

- AlexNet *(a rather small network by today's standard)*

  (2012)



62,378,344 parameters (connections)

How did we get there?

... In three stages

The **assumption**

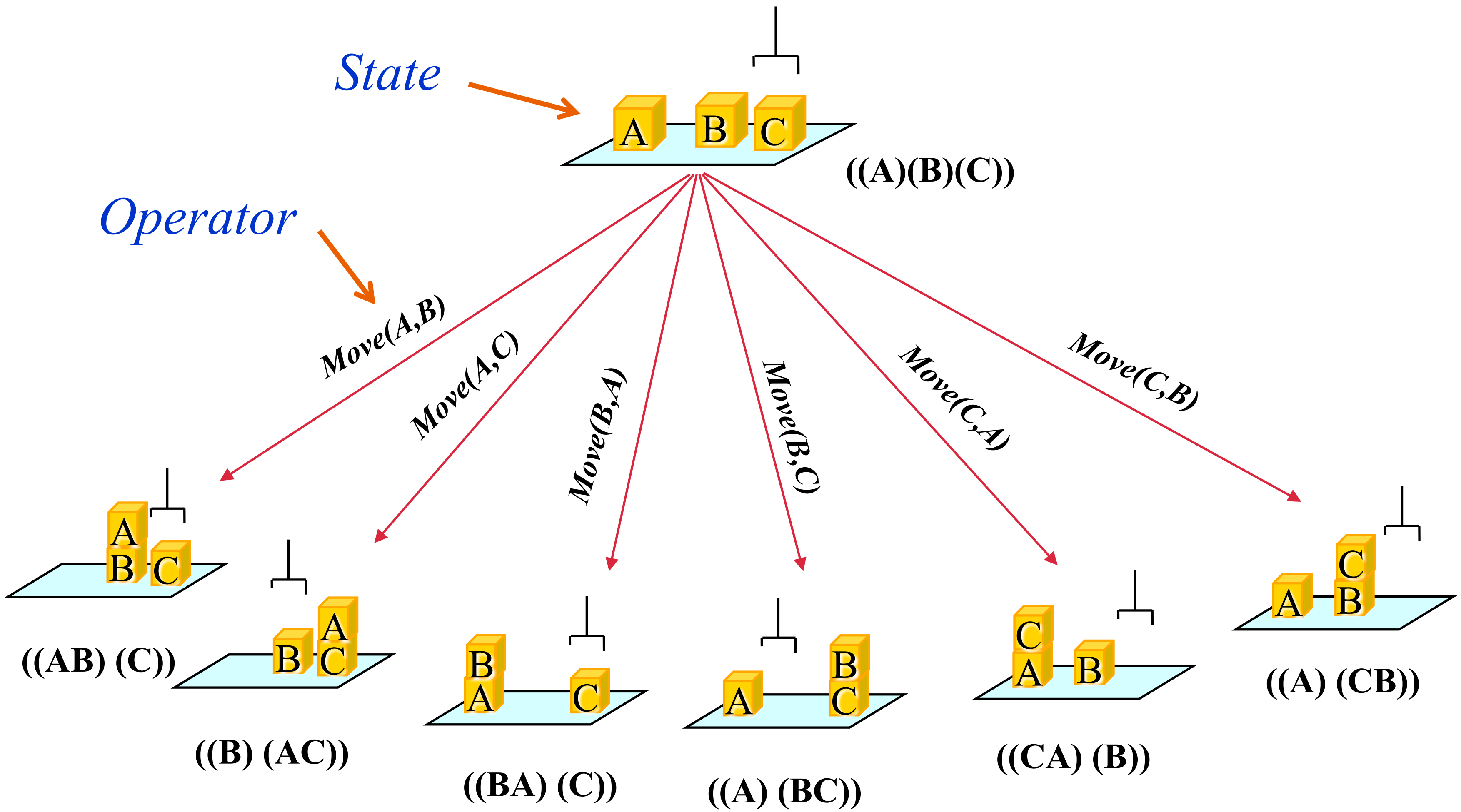
1

Intelligence is

general reasoning processes

(~1956 – ~1969)

Reasoning / problem solving



- **Theorem** proving
- General **Problem Solver**
- The first **world level champion** in the game checker
- **Planning**
- (Attempts at) automatic **translation**
- ...

Second assumption

2

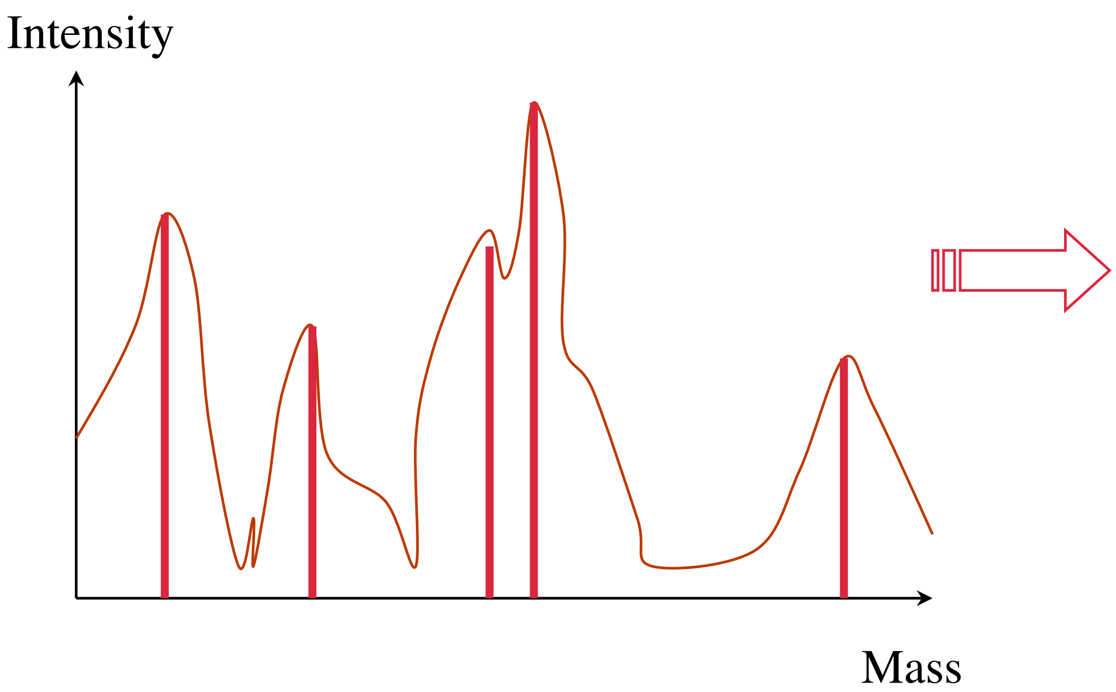
Knowledge is power

(~1970 – ~1985)

Expert Systems: DENDRAL

- A project of the NASA:
- Is there life on Mars?
- Mass spectrography

How does an expert performs this?



The developped formula of the molecules

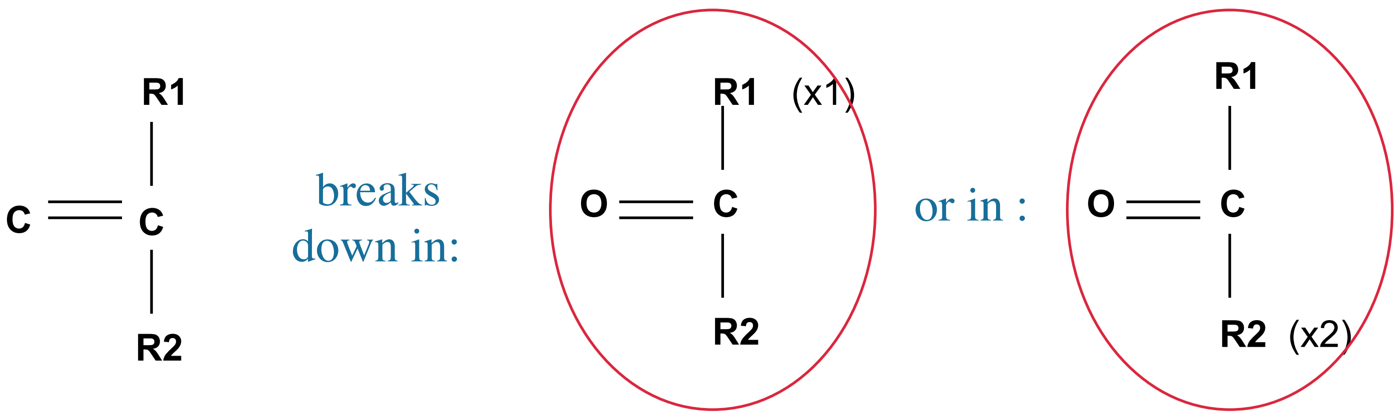
Expert Systems: DENDRAL

- Examples of a piece of knowlege
  - Rule:

    If the spectrum of the molecule has two peaks x1 et x2 such that:

    1. x1 - x2 = M + 28
    2. x1 - 28 is a high peak
    3. x2 - 28 is a high peak
    4. At least one of the peaks x1 et x2 is high

    Then the molecule contains a cetone group



Third assumption (~1985 - ...)

3

Intelligence involves a lot of **knowledge**

that is **difficult** to acquire and to maintain

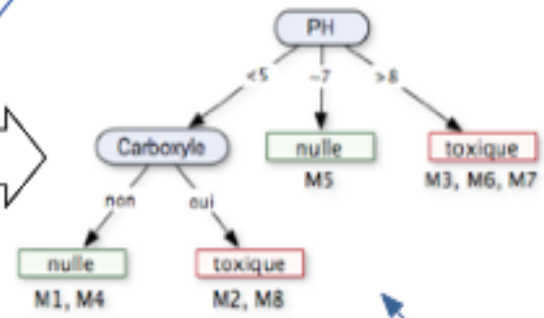
Why not learn everything from data?

through **general learning processes**

Supervised Induction

Descriptors　　　　　　Labels

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



R1: Si pH > 8
alors toxique
R2 : Si pH < 5 & Carboxyle = oui
alors toxique
....

Example

Values of the descriptors

Concept or hypothesis

Two **characteristics**

1. We want to acquire **knowledge** automatically

2. The choice of the **descriptors** (features) is crucial

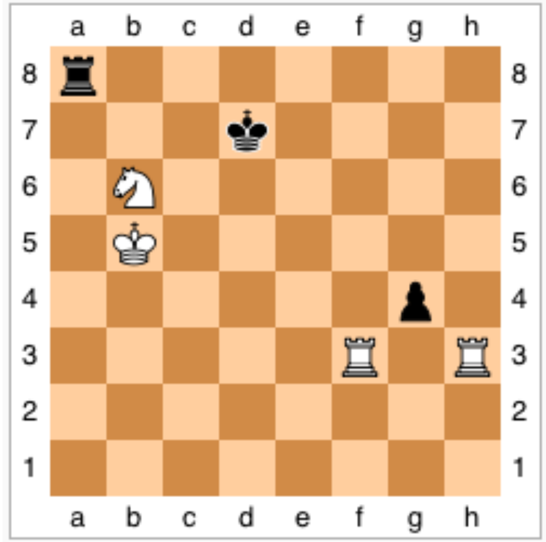
Learning **from a single example**

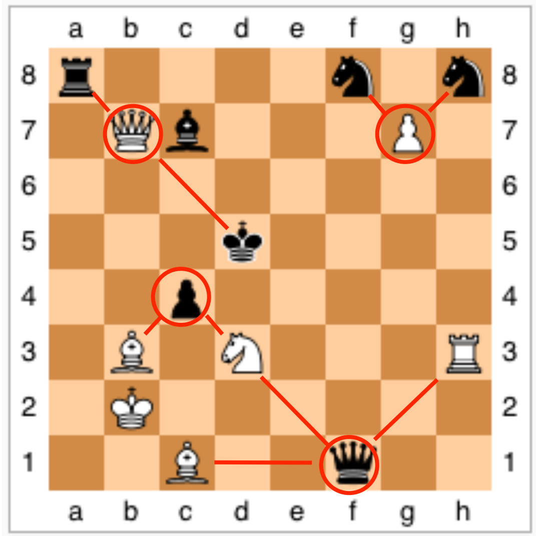
Explanation-Based Learning

1. A **single** example

2. **Search for a proof** of a « fork »

3. **Generalization**





An empirical fact

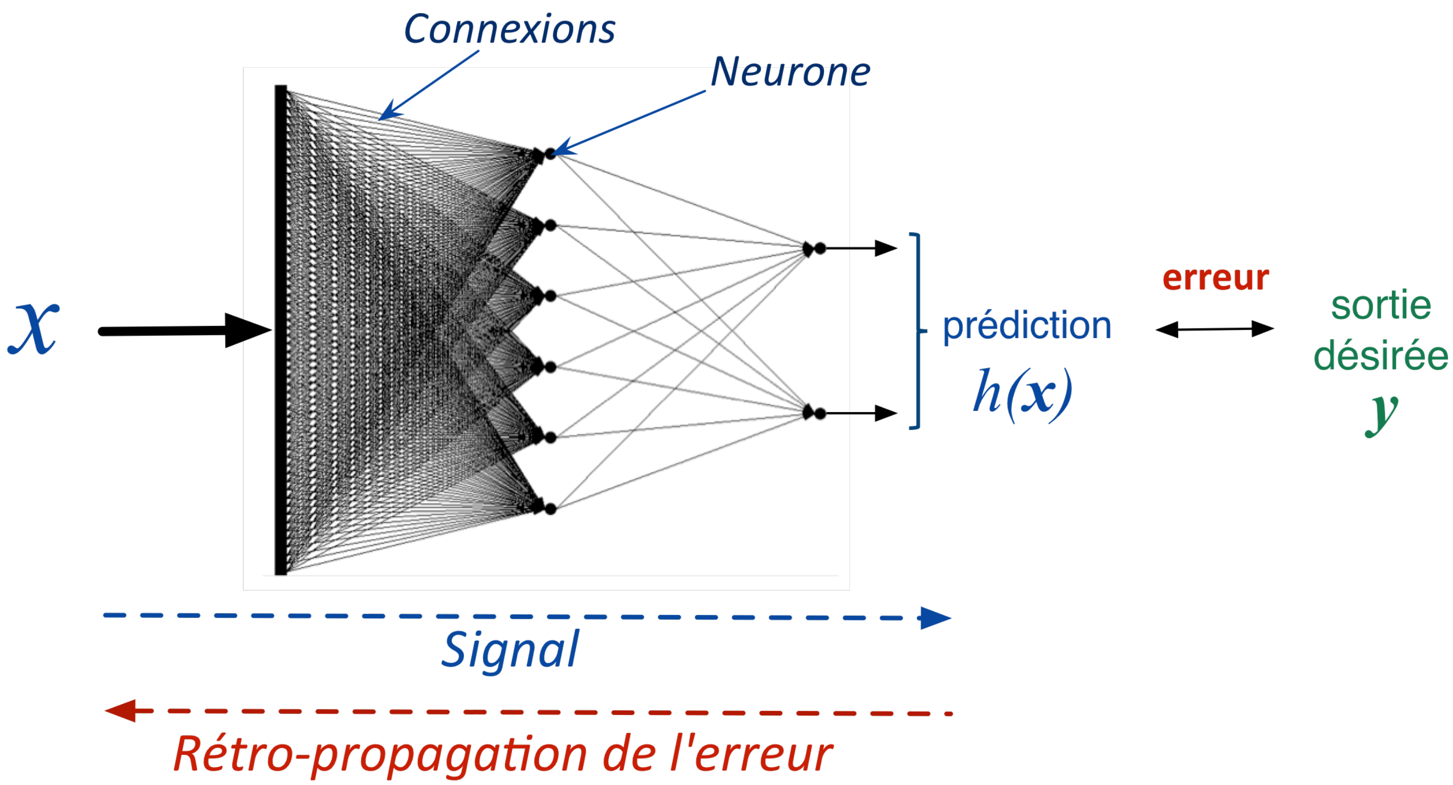
- Powerful **symbolic machine learning** methods

- Are brittle when the data is **imperfect**
  - Noisy
  - Missing values
  - Uncertainties

Learning with Multi-Layer Perceptrons

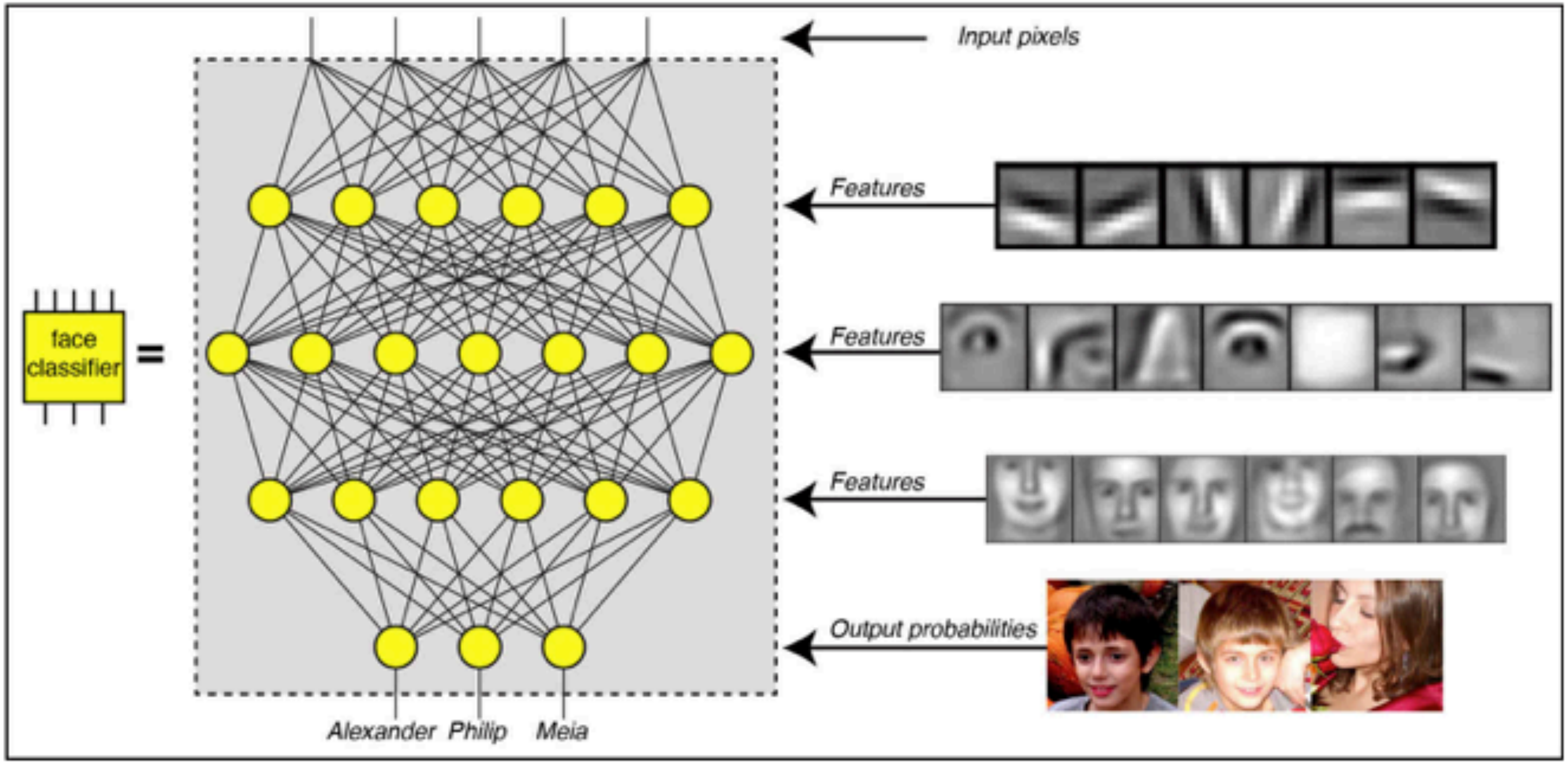
Performs magic!

- Automatically **self-adapt** from the data
- And **resistant** to **noisy data**

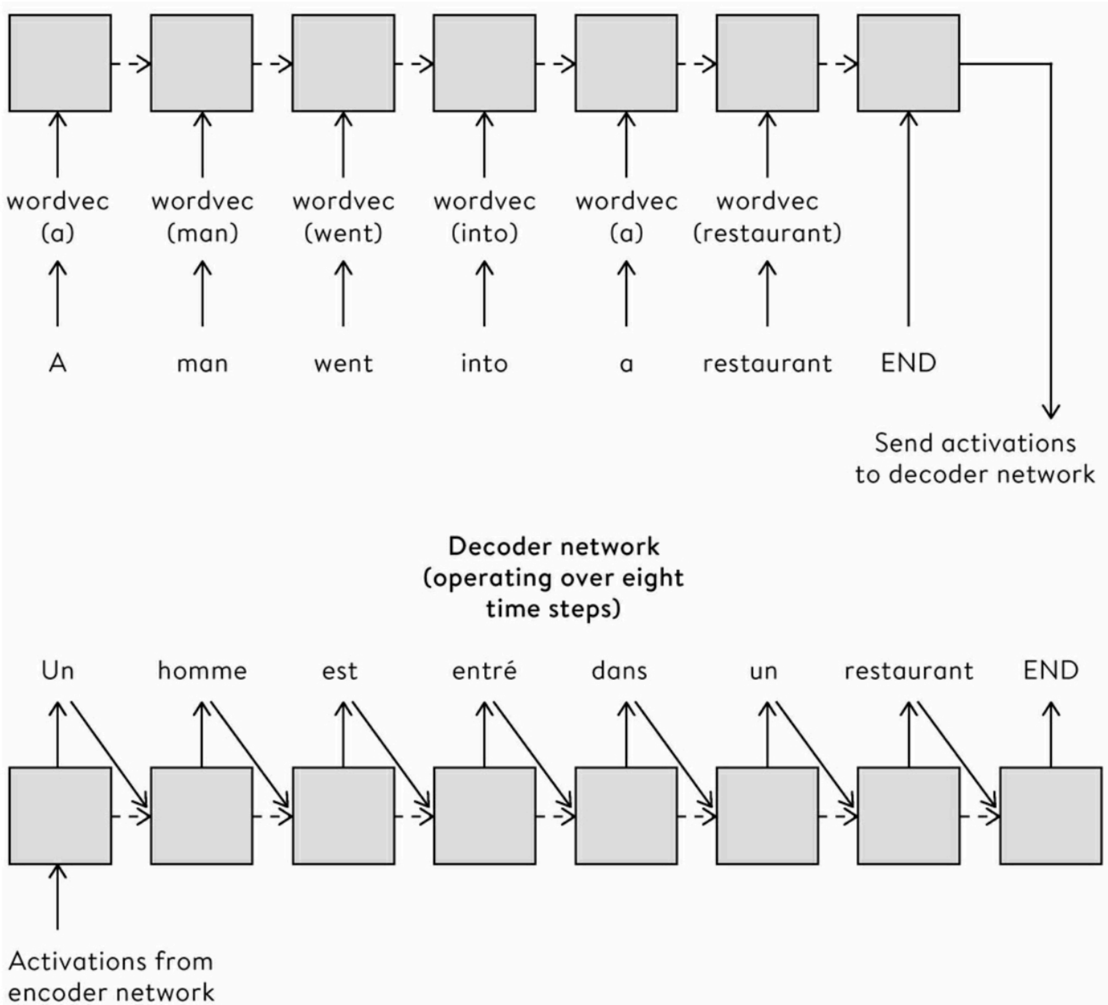


Deep learning => automatic feature construction

...

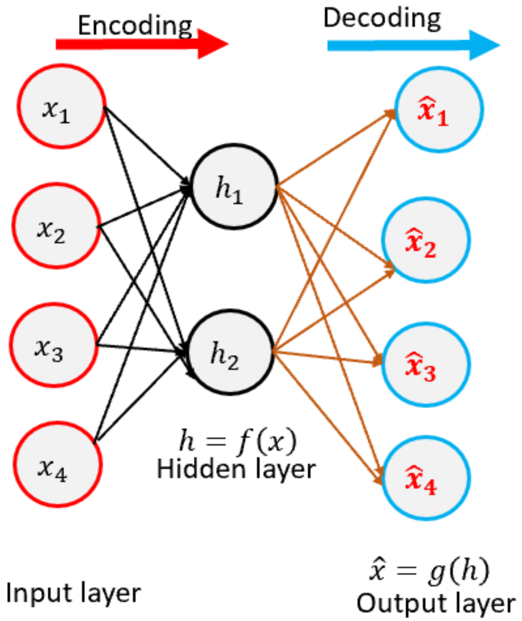


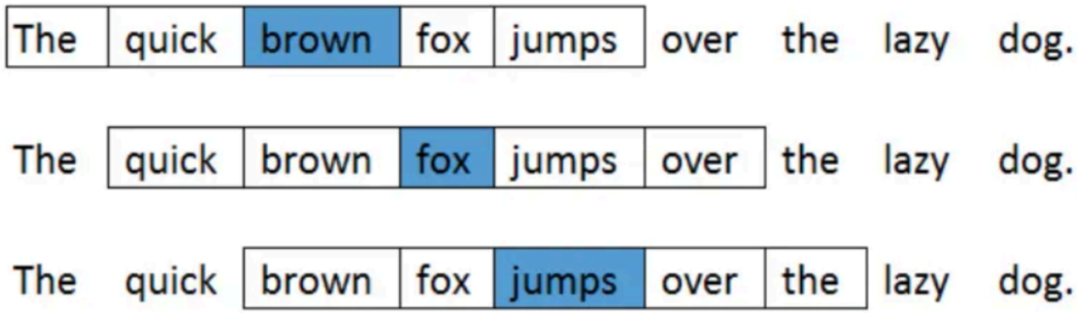
Automated **translation**



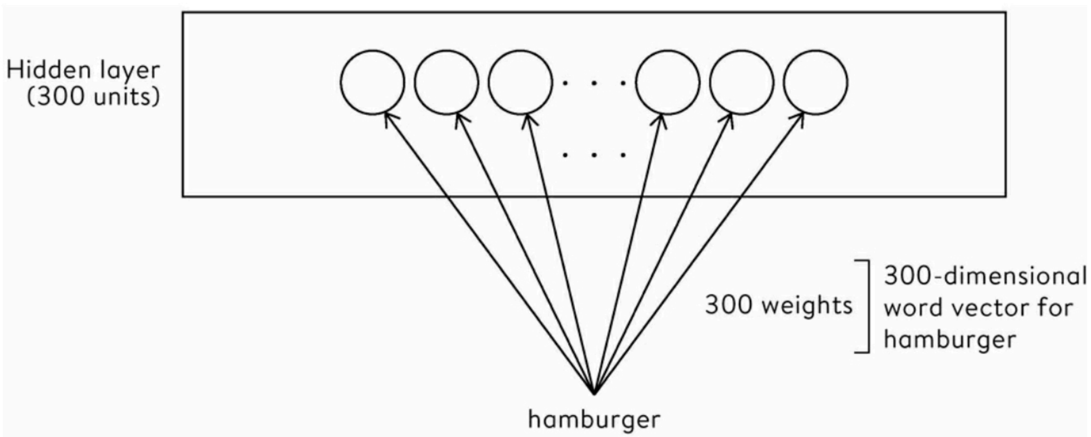
From [Melanie Mitchell "*Artificial Intelligence: A Guide for Thinking Humans*" (2021)]

Learning a "semantic space"

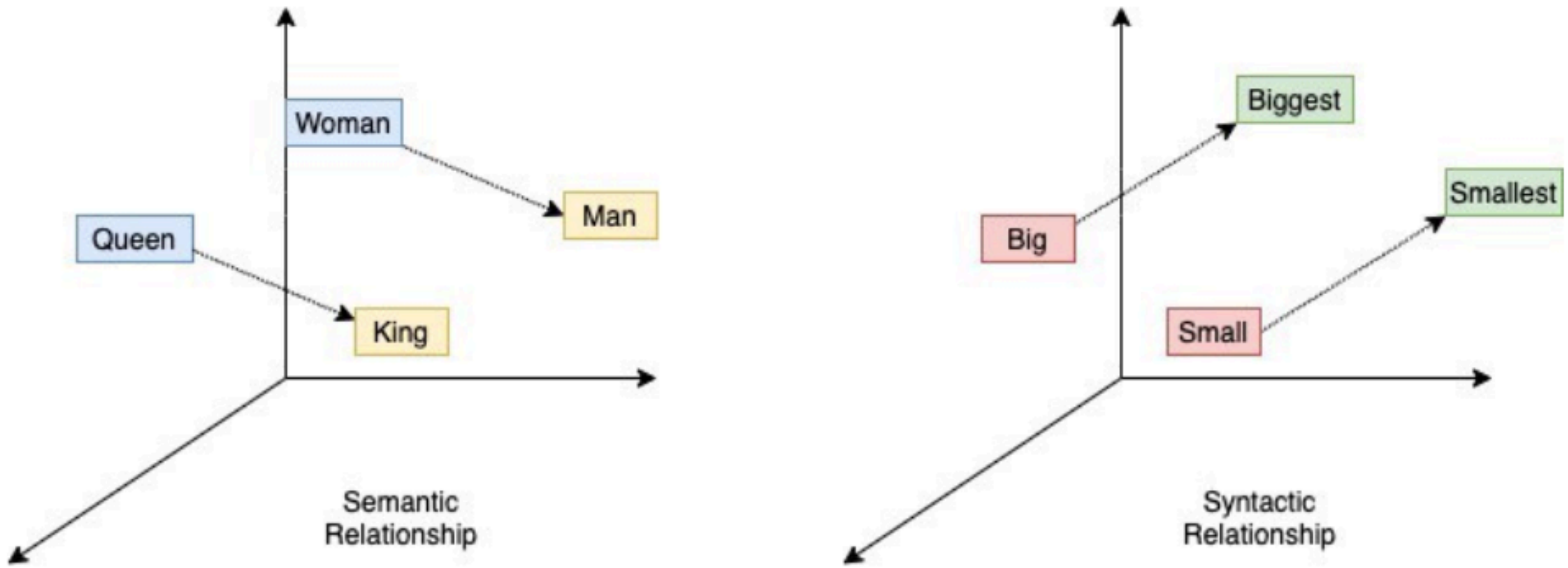




Contexte				Mot Cible
The	Quick	Fox	Jump	**Brown**
Quick	Brown	Jumps	Over	**Fox**
Brown	Fox	Over	The	**Jumps**



Learning a "semantic space"

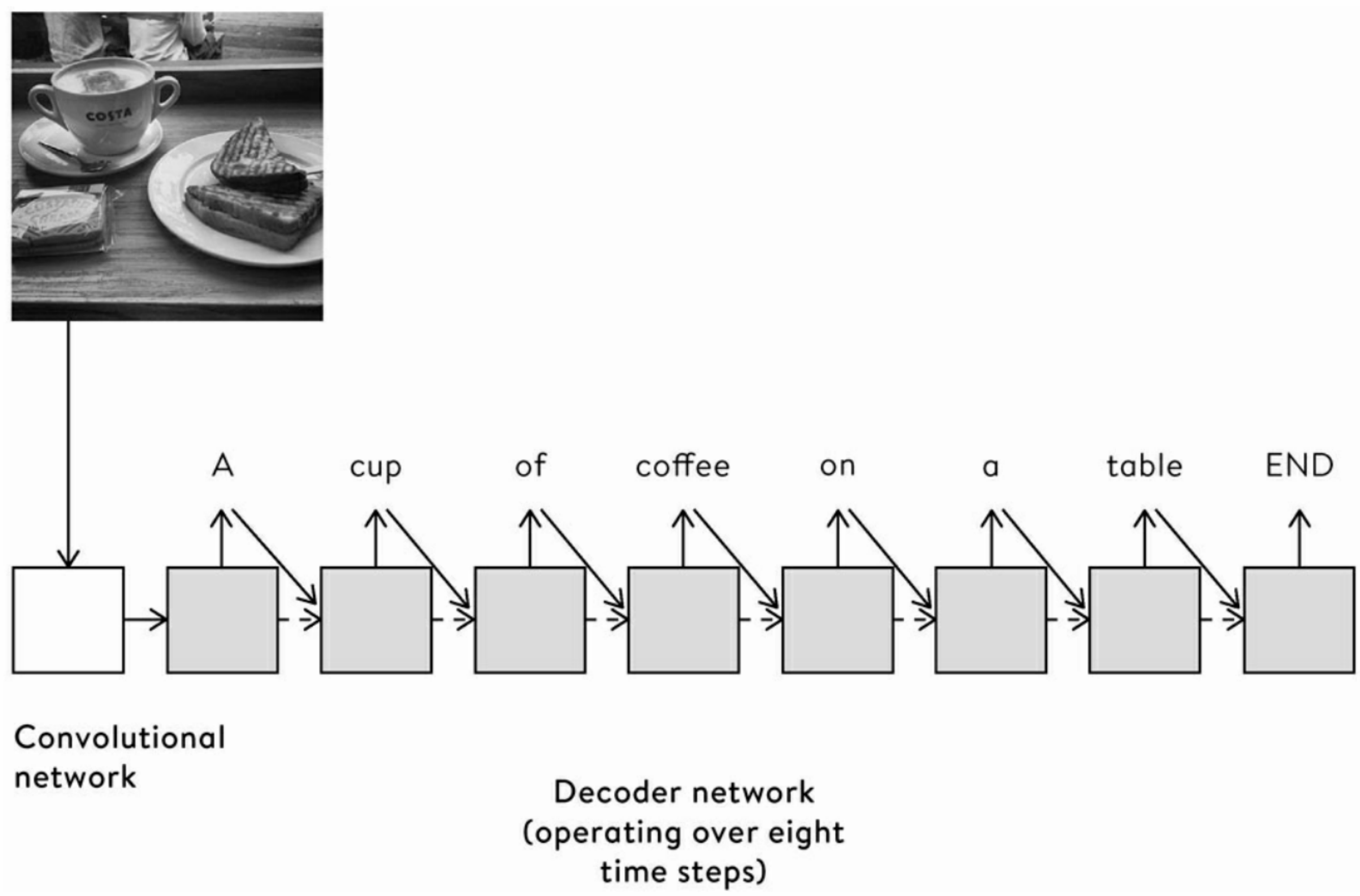


...

Automated image-captioning

A group of young people playing a game of frisbee

Automated **image-captioning**



From [Melanie Mitchell "*Artificial Intelligence: A Guide for Thinking Humans*" (2021)]

The deep learning revolution

– brings **unexpected** levels of performance

– **solves** new problems

- Automatic translation
- Autonomous vehicles
- Discovery of protein foldings
- …

Shall/should we reason in terms of **neural networks units**?

Outline

1. Has AI been bio-inspired?

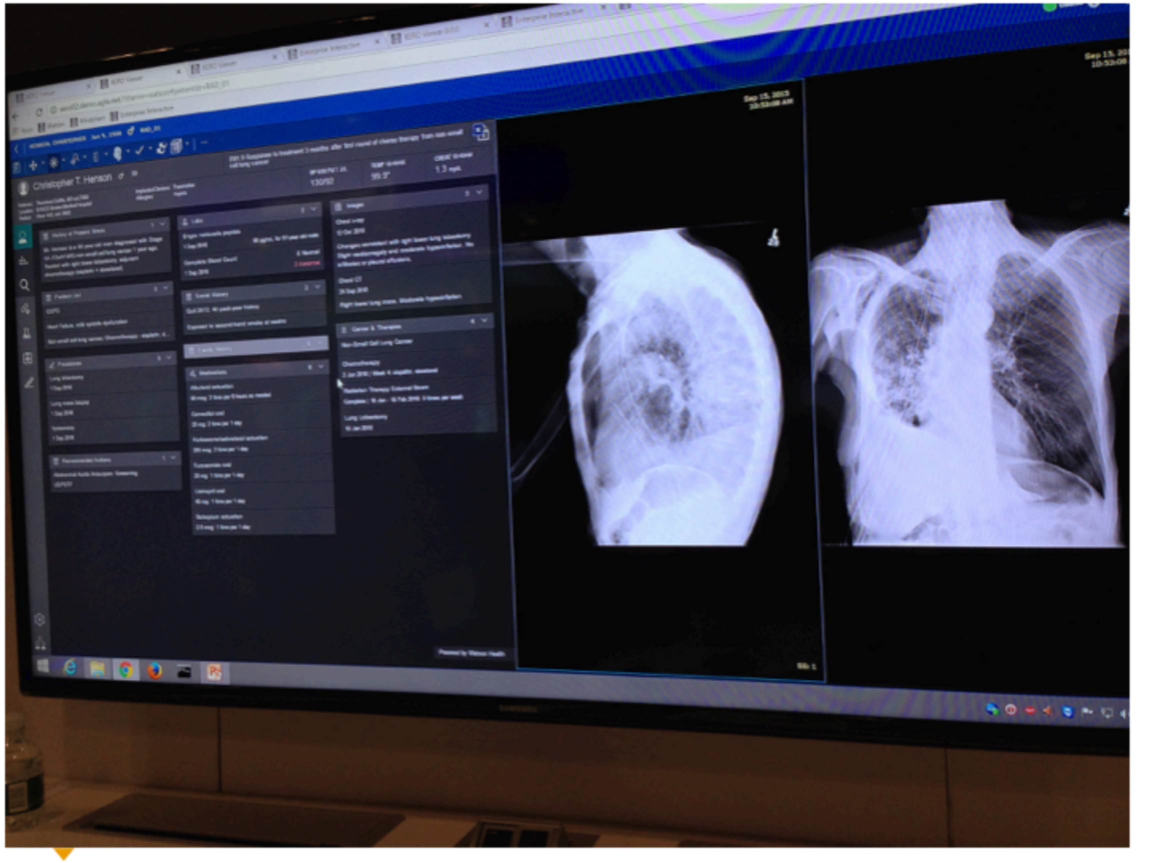
2. Interfacing **AI with Humans**

3. Interfacing AI with AI

4. Conclusion

How Artificial Intelligence will change Medical Imaging

Machine learning software will serve as a very experienced clinical assistant, augmenting the doctor and making workflow more efficient in radiology



An example of how Agfa is integrating IBM Watson into its radiology workflow. Watson reviewed the X-ray images and the image order and determined the patient had lung cancer and a cardiac history and pulled in the relevant prior exams, sections of the patient history, cardiology and oncology department information. It also pulled in recent lab values, current drugs being taken. This allows for a more complete view of the patient's condition and may aid in diagnosis or determining the next step in care.

…

Automated image-captioning

- Not always so good!

A dog is jumping to catch a frisbee

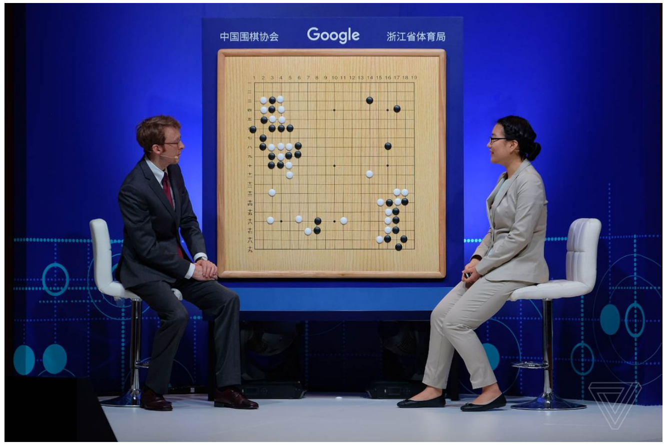
The AlphaGo case

- Plays like an "alien"
- An amazing game
- Revolutionizes the way we play
- Effervescence in 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 of exegesis. Explanations a posteriori
- Necessary for
  - Communication
  - teaching



And even AlphaGo might err

Error in medicine

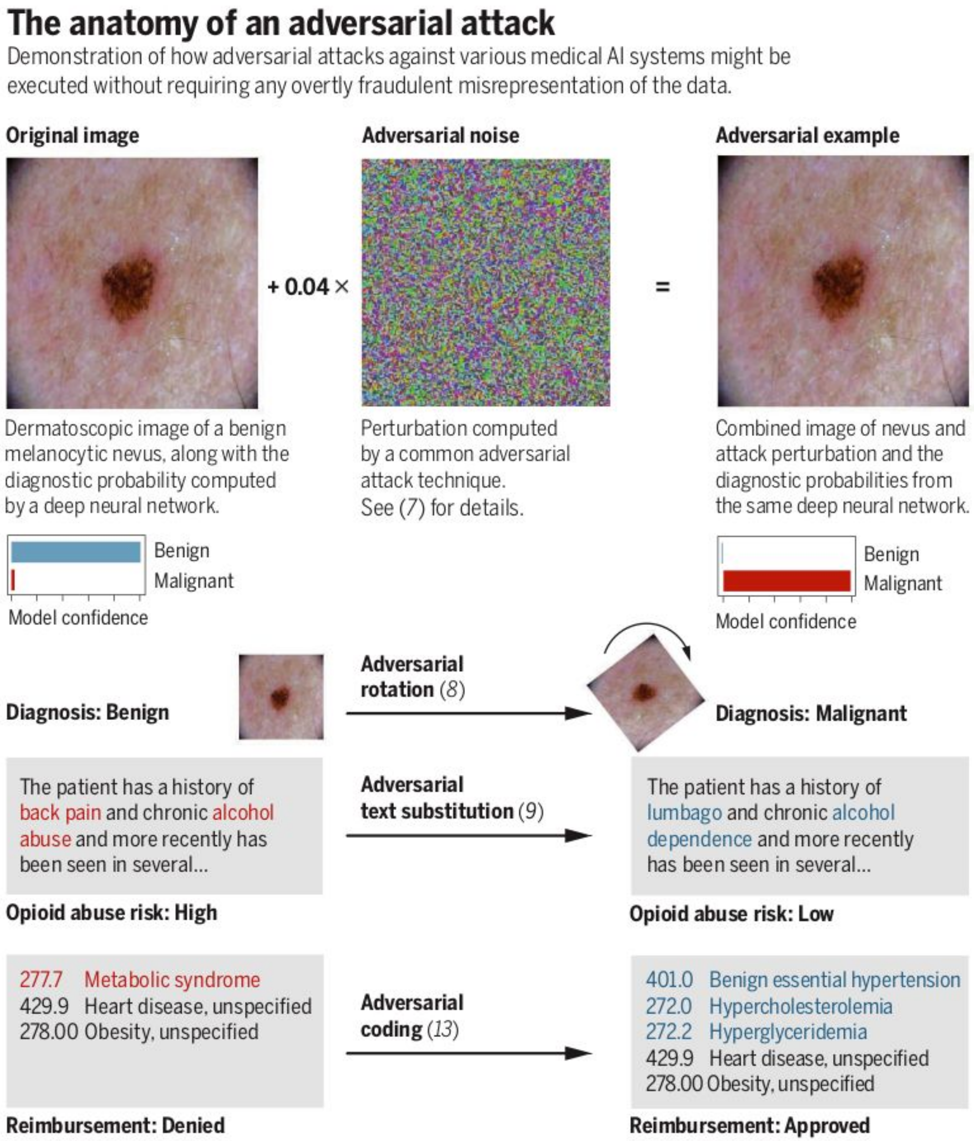
MACHINE LEARNING

Science

Adversarial attacks on medical machine learning

Emerging vulnerabilities demand new conversations

22 March 2019



Problem

- So far efficient predictors are often black boxes

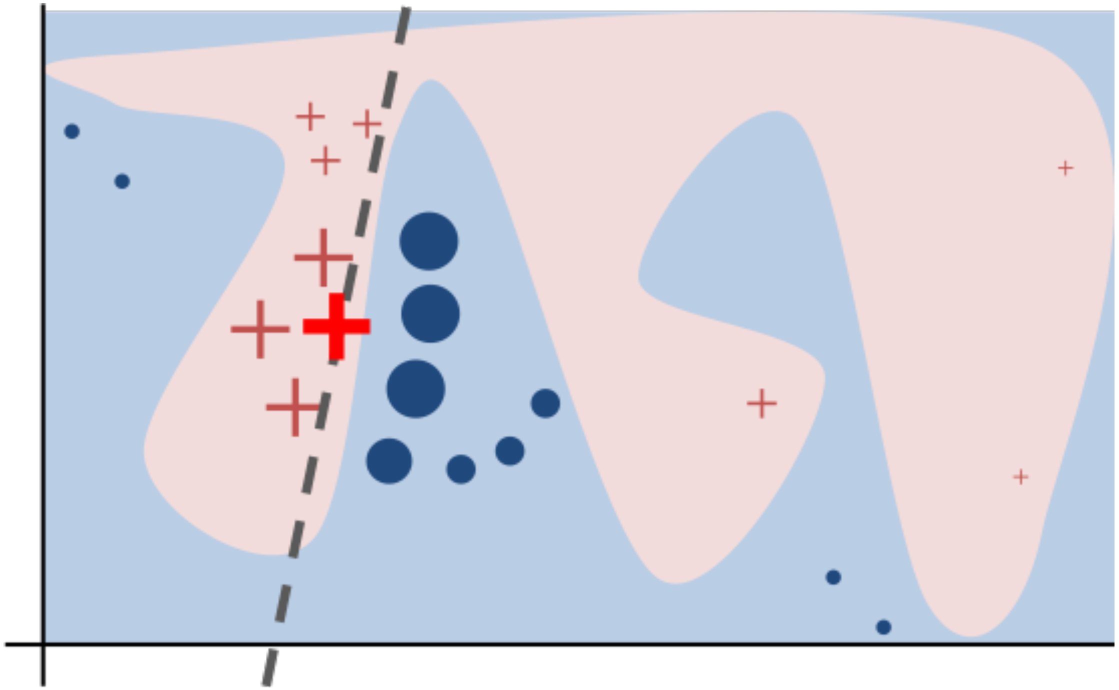
- This is an issue for a number of applications (e.g. in medicine)
  - We want to be able to be **confident** in the system
  - It can justify its **decisions**
  - It can justify its **reasoning**

The ability of providing explanations is **required in Europe** since May 2018 (GDRP, Recital 71)

XAI: Explainable Artificial Intelligence

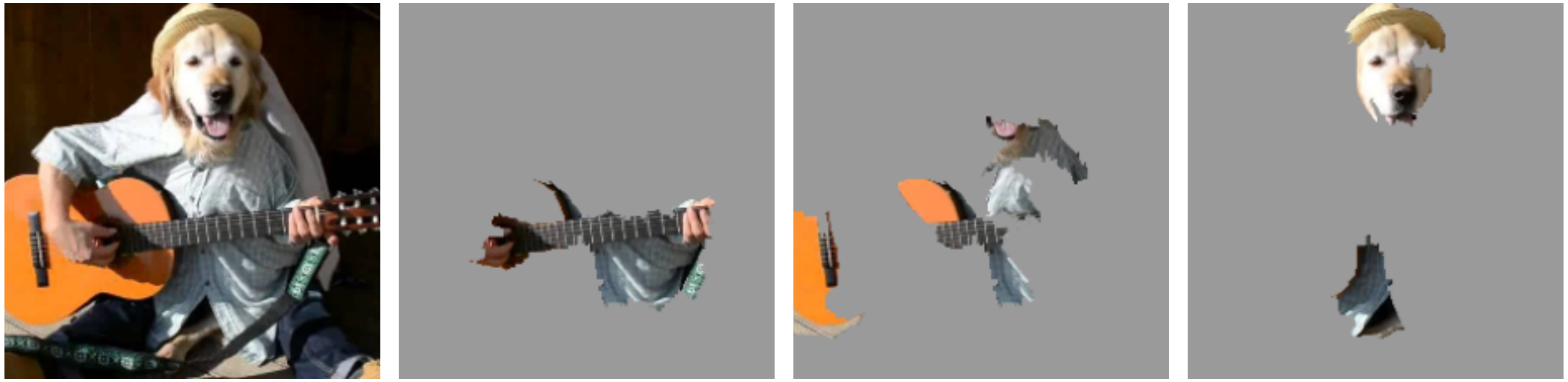
What is a (good) explanation?

Local simplification



- LIME

Sensitivity analysis

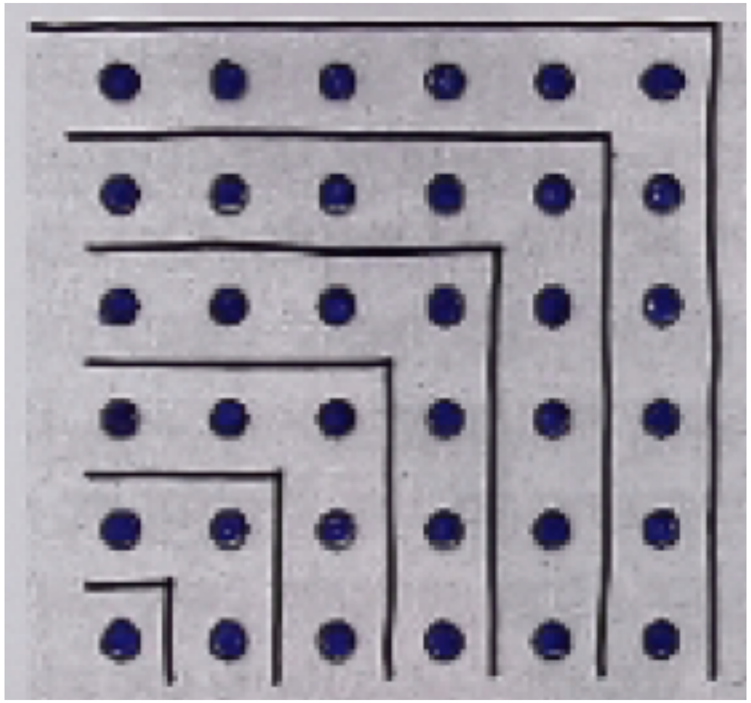


- The pixels that best "explain"
  - The recognition of a **electric guitar**
  - The recognition of an **acoustic guitar**
  - The recognition of a **dog**

- Still very rudimentary

Many types of representations and of operations

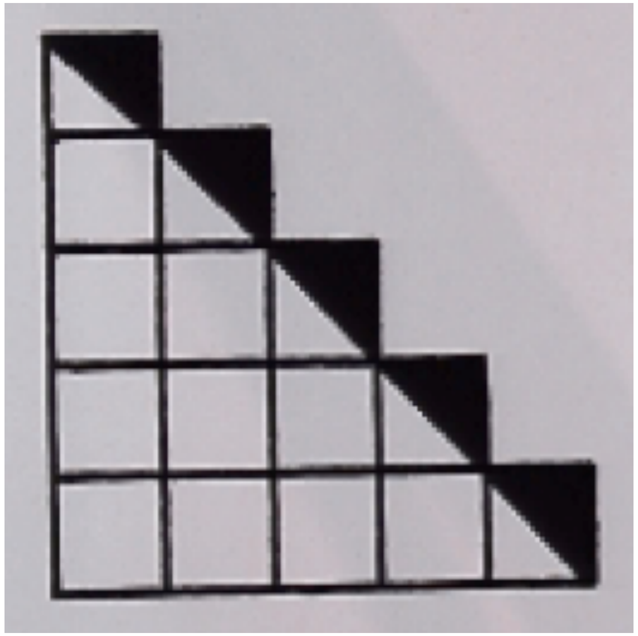
$$1 + 3 + 5 + \ldots + (2n-1) \overset{?}{=} n^2$$



...

Lots of types of “explanations”

$$1 + 2 + 3 + \ldots + n \overset{?}{=} \frac{n^2}{2} + \frac{n}{2}$$



...

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 contactée lors du séjour à l'hôpital
	alors il est presque certain (0,8) que l'organisme est de la catégorie des enterobactoriaceae; de plus, il se peut (0,1 ; règle 37) que ce soit la pseudomonia 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

- How old is the patient?
- ** Why
- In order to determine if tetracycline can be prescribed
  - If the patient is less than 8 years old
  - Then, it is not possible to prescribe tetracycline
  - [Rule 122]
- ** Why?
- …

Why is it **not possible** to prescribe tetracycline to a child **less than 8 years old**?

Explanations in Expert Systems

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

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

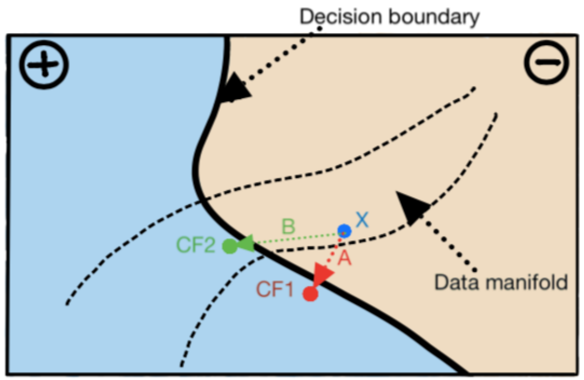
Counterfactuals

- **If** James Dean had **taken the train** the day of his car accident, he would not have died
- **If** you could **increase your savings** by 5000€ each year, you would get this loan

Counterfactuals

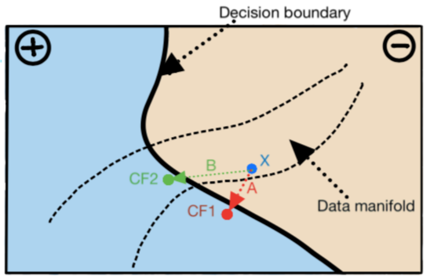
- **If** James Dean had **taken the train** the day of his car accident, he would not have died
- **If** you could **increase your savings** by 5000€ each year, you would get this loan

Local explanation for a given prediction



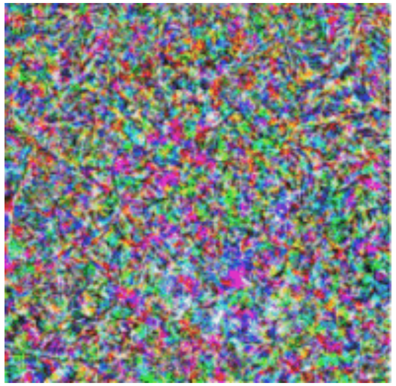
Two possible **counterfactuals**: CF1 is closest to **x** than CF2

Counterfactuals



- Oh yes. But what is the difference with **adversarial examples?!**

And the **difference** is



This is not "toilet paper" because this is "dog"

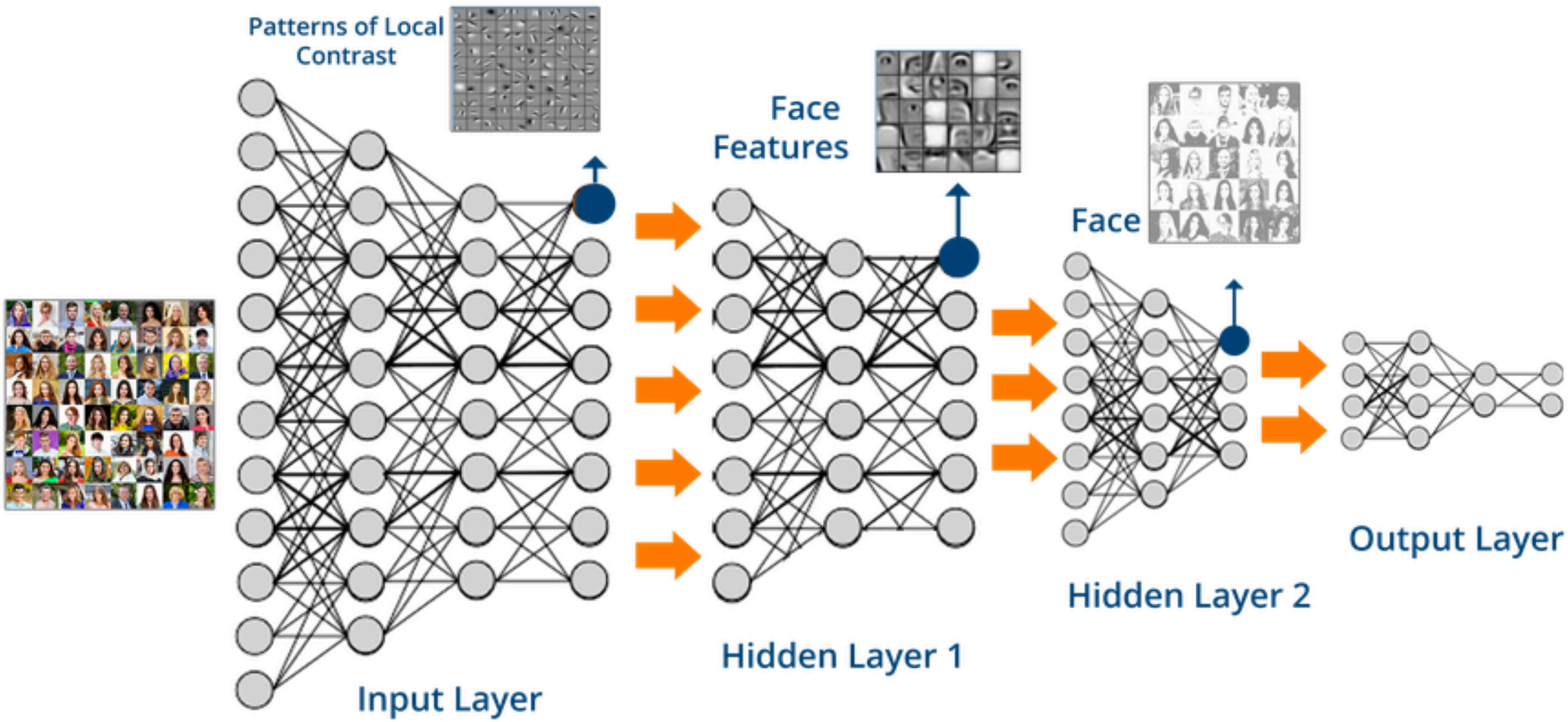
What is a **good level** of communication?

“No computation can get around the semantic problem”

K. Browne & B. Swift (2020). “**Semantics and explanation: why counterfactual explanations produce adversarial examples in deep neural networks**”. *arXiv preprint arXiv:2012.10076*.

What is a **good level** of communication?

- Should we look at **intermediate layers** in deep NNs?

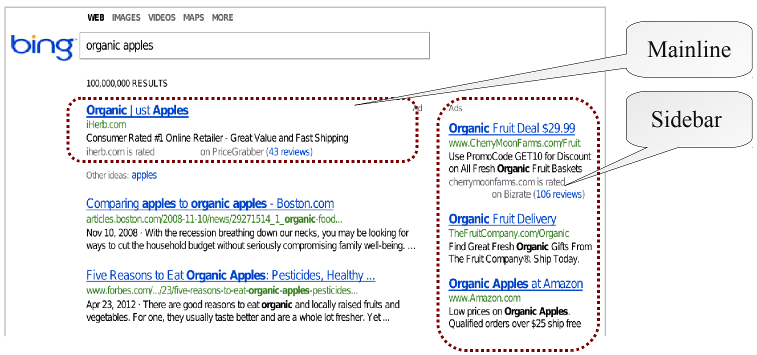


Outline

1. Has AI been bio-inspired?
2. Interfacing AI with Humans
3. Interfacing **AI with AI**
4. Conclusion

What should **learning sub-systems** exchange?

- Two **sub-systems**
  1. One **locating** the **ads links**
  2. The other **choosing** the **adds to present**



- That **influence each other**
  - Each takes into account the **clicks**
  - Which **depend** in part from the actions of the other sub-system
  - In addition of other **uncontrolled factors** (price, user's queries, …)

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

What should **learning sub-systems** exchange?

- The subsystem **locating** the adds gathers the following statistics

	Overall
Add placed in **mainline**	**0.78**% (273/35000)
Add placed on **sideline**	**0.83**% (289/35000)

What is the best choice?

What should **learning sub-systems** exchange?

- The subsystem **locating** the adds gathers the following statistics

	Overall	Add ranked **1**st	Add ranked **2**nd
Add placed in **mainline**	**0.78**% (273/35000)	**0.93**% (81/8700)	**0.73**% (192/26300)
Add placed on **sideline**	**0.83**% (289/35000)	**0.87**% (234/27000)	**0.69**% (55/8000)

What is the best choice?

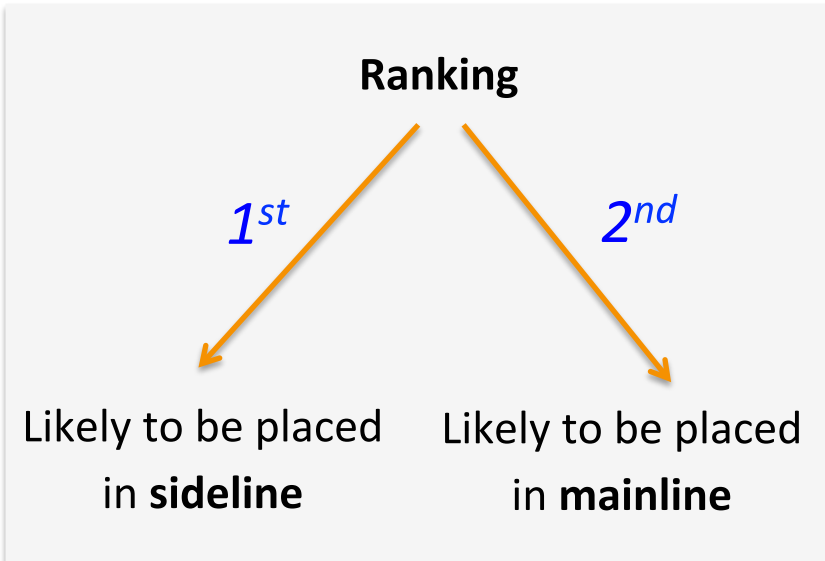
What should **learning sub-systems** exchange?

…

	Overall	Add ranked **1st**	Add ranked **2nd**
Add placed in **mainline**	**0.78%** (273/35,000)	**0.93%** (81/8700)	**0.73%** (192/26,300)
Add placed on **sideline**	**0.83%** (289/35,000)	**0.87%** (234/27,000)	**0.69%** (55/8000)

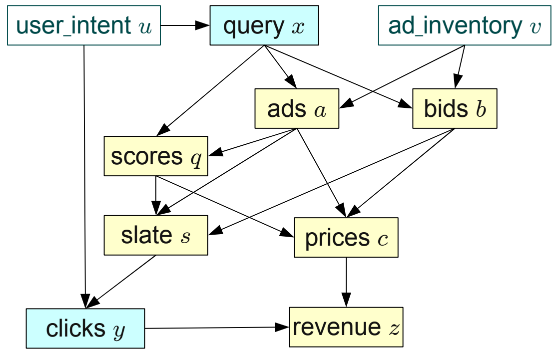
- Influencing factor

  The **choice of the placement** was *function of* the **ranking** of the add (by the other subsystem)



What should **learning sub-systems** exchange?

- The subsystems should **communicate** on
  - the **influencing factors**
  - and **causality** relationships



***Conceptual** level*

- Ok. But what about

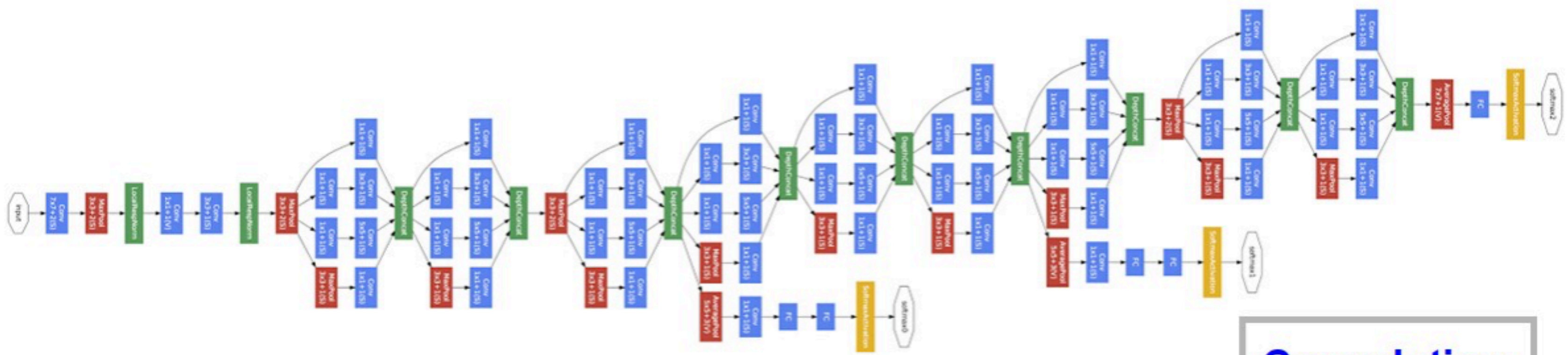
  a **neural network** learning from **another one**?

“Distillation”

One neural network **teaching** another one

- Why?
  - A “**master**” **complex** neural network
  - A “**student**” neural network with **limited capacity**

Example: A sophisticated learning technique - GoogLeNet



Quite a costly machine to train

AND to use for prediction

“Distillation”

Many ways to do it

1. Changing the training examples (x_i, y_i) by **modifying** the **targets** y_i

"Distillation"

Many ways to do it

1. Changing the training examples (x_i, y_i) by **modifying** the **targets** y_i

2. **Changing** the training **inputs** x_i

“Distillation”

Many ways to do it

1. Changing the training examples (x_i, y_i) by **modifying** the **targets** y_i

2. **Changing** the training **inputs** x_i

3. Changing **the learning task** through a “**curriculum**”: *sequence of intermediate tasks*

How to measure the difficulty

of examples?

"Prediction depth"

- The number of hidden layers after which the network's final prediction **is** already **determined**

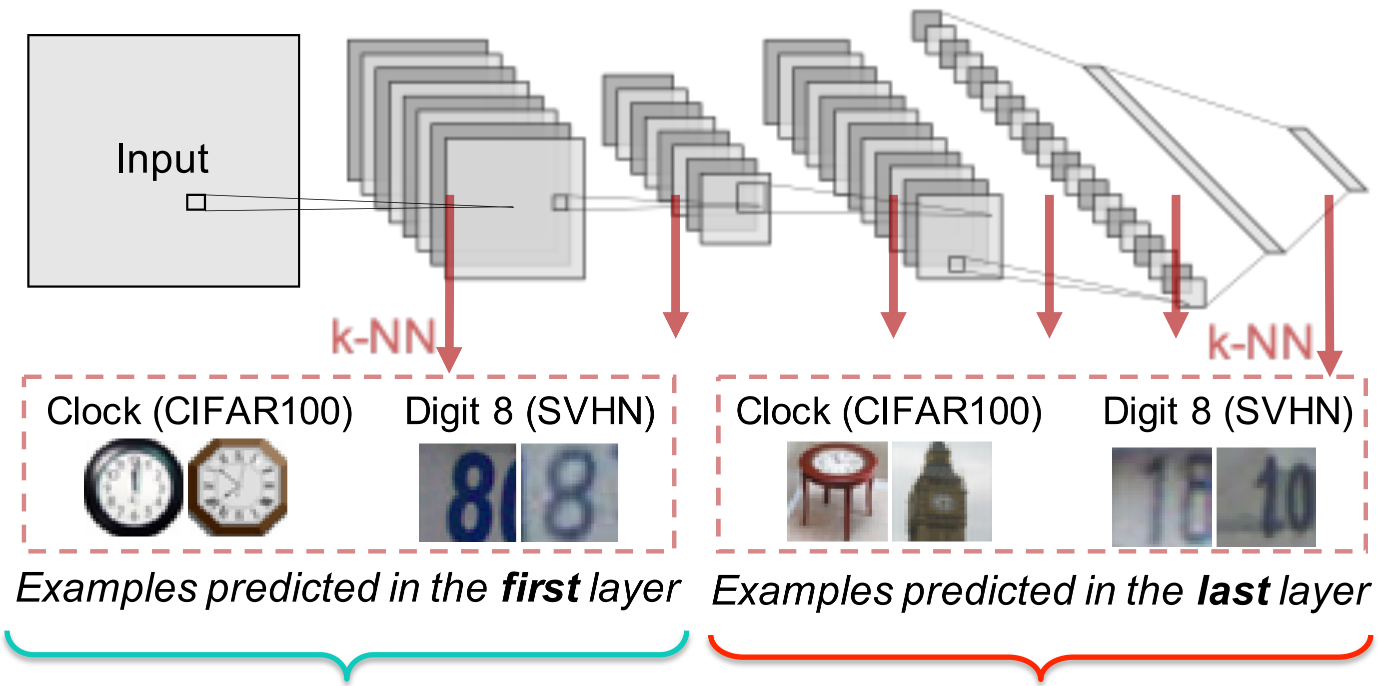
Deep neural networks use

- fewer layers to determine the prediction for **easy** examples
- and more layers for **hard** examples

Baldock, R., Maennel, H., & Neyshabur, B. (2021). **Deep learning through the lens of example difficulty**. *Advances in Neural Information Processing Systems, 34*.

“Prediction depth”

- The number of hidden layers after which the network’s final prediction **is** already **determined**



Easy examples

Hard examples

How to measure the prediction depth?

- k-NN classifier probes (with k = 30)
  - Compare the the **hidden embedding of an input** to those of the **training set** (what is the class of the *k* nearest neighbors in the embedding considered)
- A prediction is defined to be made at a depth L = *l* if
  - The k-NN classification **after** layer l = *l* − 1 is **different** from the network's final classification,
  - but the classification of k-NN probes **after** every layer L ≥ *l* are all **equal** to the final classification of the network

What **they claim** to show

1. The **prediction depth is larger** for examples that visually appear to be **more difficult**
   – And this is consistent between NN's architectures and random seeds

What **they claim** to show

1. The **prediction depth is larger** for examples that visually appear to be **more difficult**
   - And this is consistent between NN's architectures and random seeds

2. Predictions are on average **more accurate** for validation points with **small prediction depths**

What **they claim** to show

1. The **prediction depth is larger** for examples that visually appear to be **more difficult**
   - And this is consistent between NN's architectures and random seeds

2. Predictions are on average **more accurate** for validation points with **small prediction depths**

3. Final predictions for data points that **converge earlier** during training are typically determined in **earlier layers**

What **they claim** to show

1. The **prediction depth is larger** for examples that visually appear to be **more difficult**
   – And this is consistent between NN's architectures and random seeds

2. Predictions are on average **more accurate** for validation points with **small prediction depths**

3. Final predictions for data points that **converge earlier** during training are typically determined in **earlier layers**

4. Both the adversarial input margin and output margin are **larger** for examples with **smaller prediction depths**
   – Intervention to **reduce the output margin** leads to **predictions** being made only **in the latest hidden layers**

What they claim to show

1. Early layers **generalize** while later layers **memorize**

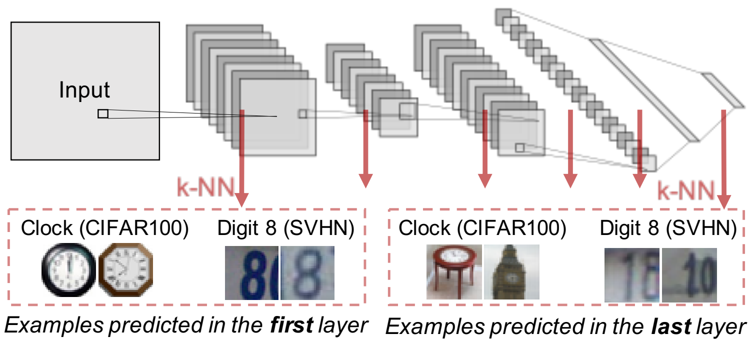
2. Networks converge **from** input layers **towards** output layers

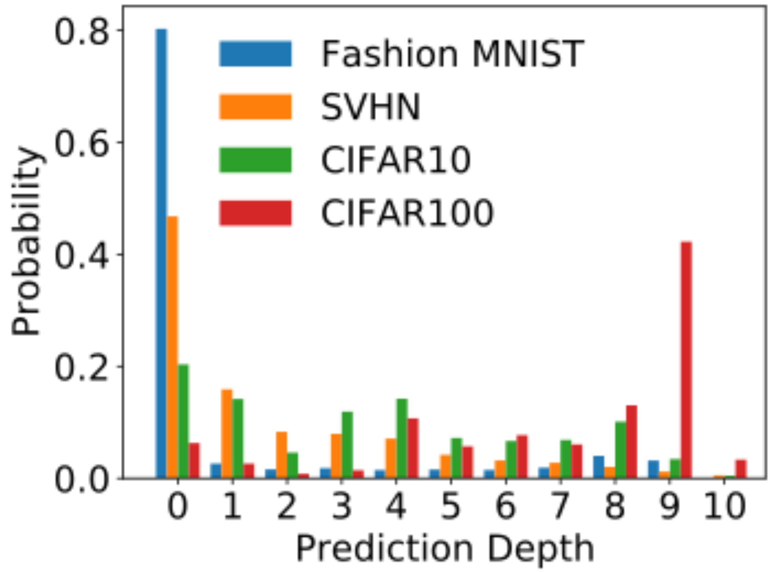
3. **Easy** examples are learned **first**

4. Networks present **simpler functions earlier** in the training

What they claim to show

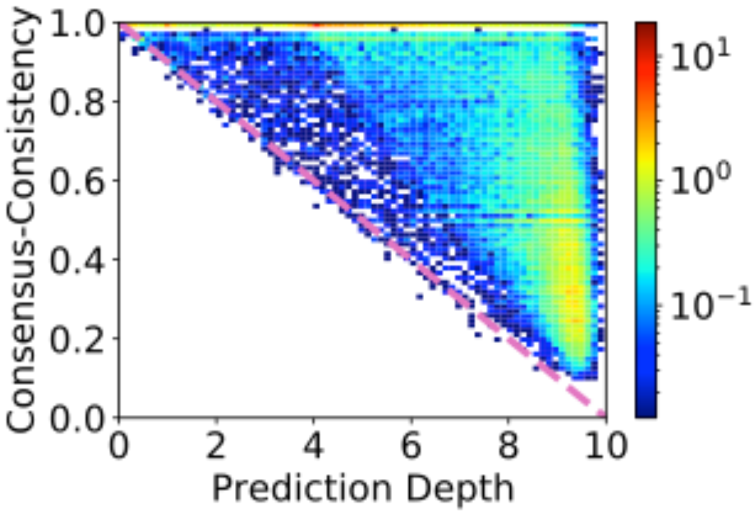
- The **prediction depth is larger** for examples that visually appear to be **more difficult**

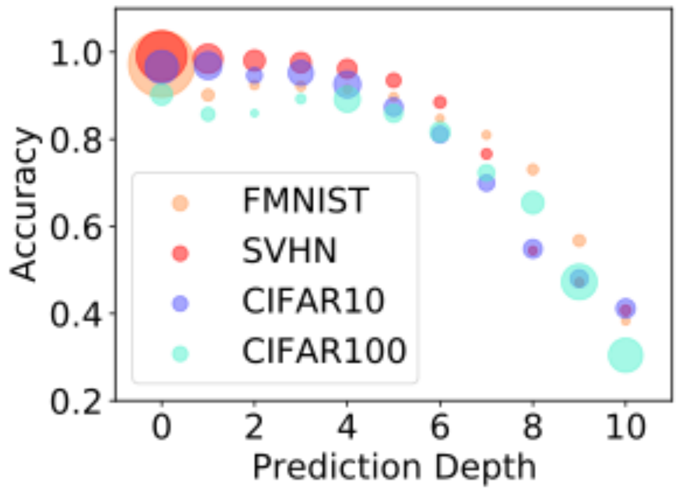




What they claim to show

- Predictions are on average **more accurate** for validation points with **small prediction depths**





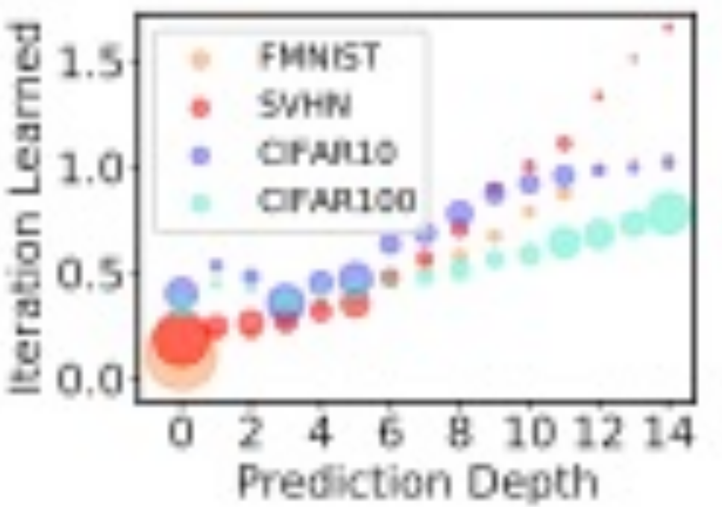
250 ResNet18 were trained on CIFAR100 (90:10% random train:validation splits). Comparison of the average **prediction depth** of a point to the **consensus-consistency** of the corresponding prediction.

For each dataset, **250** ResNet18 were trained on CIFAR100 (90:10% random train:validation splits). Each time a point appears in the validation split, its **prediction depth** and whether the **prediction was correct** was recorded.

Consensus-consistency: the fraction of NNs that predict the ensemble's consensus class

What they claim to show

- Final predictions for data points that **converge earlier** during training are typically determined in **earlier layers**
  - Measure of the difficulty of learning an example by the **speed at which the model's prediction converges** for that input during training
  - **Iteration learned**. A data point is said to be learned by a classifier at training iteration $t = \tau$ if the predicted class at iteration $t = \tau - 1$ is different from the final prediction of the converged NN and the predictions at all iterations $t \geq \tau$ are equal to the final prediction of the converged NN.

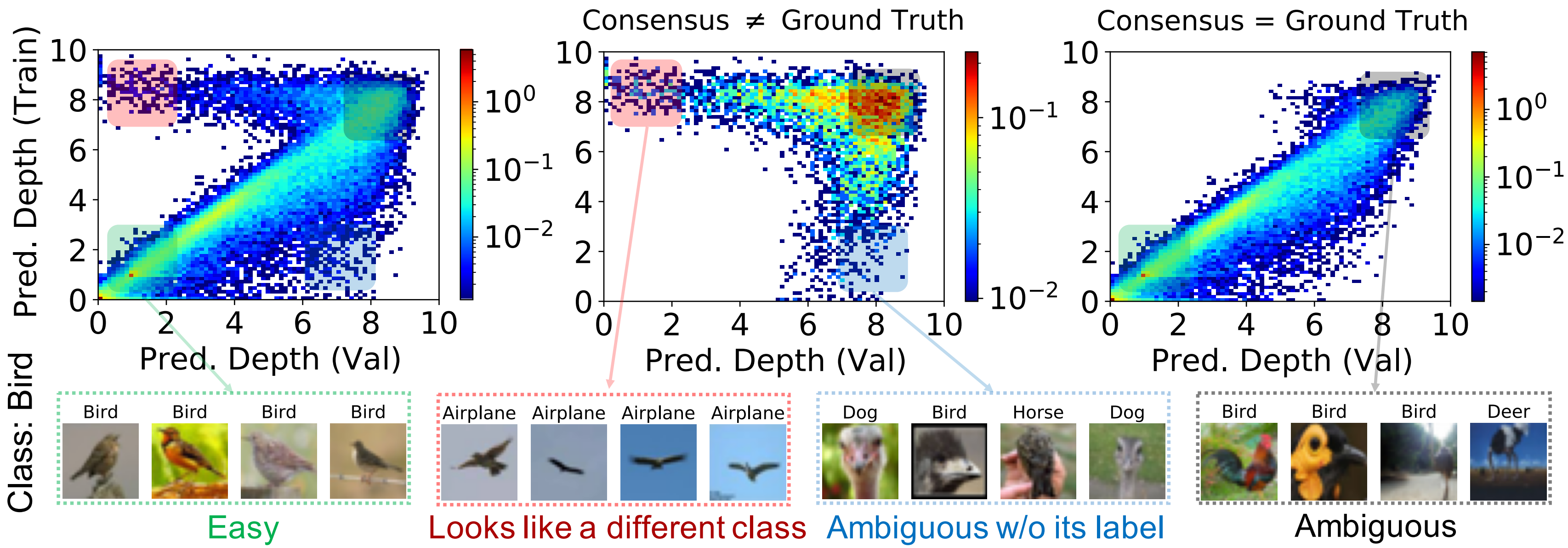


What they claim to show

- **Different forms** of example difficulty
    - **Validation**: points with low prediction depth are “clear” and “ambiguous” otherwise
    - **Training**: idem
  - **Easy examples** (Low PD_{val} and low Pd_{train})
  - **Look like a different class** (Low PD_{val} and high Pd_{train}).
    - E.g. mislabeled examples
  - **Ambiguous unless the label is given** (High PD_{val} and low Pd_{train}).
    - E.g. ressemble both their own class and another class. Likely to be misclassified
  - **Ambiguous** (High PD_{val} and high Pd_{train}).
    - Examples that may be corrupted or of a rare sub-class.

What they claim to show

- Illustration



ressemble both their own class
and another class. Likely to be misclassified

Collaborative learning

- Exchanges **between learning agents** (methods)
  - E.g. **supervised** and **unsupervised** methods

    [ANR “Herelles” on satellite image processing. Pierre Gançarski (PI). ICube]

- Exchange of **parameters**
  - **Number** of clusters
  - **Prototypes**
  - **Labels** for training examples

- **NOT** the specifics of the methods (e.g. neuron activations)

Outline

1. Has AI been bio-inspired?
2. Interfacing AI with Humans
3. Interfacing AI with AI
4. Conclusion

Conclusions

- The brain metaphor **has not played** a determining role in AI
- The **cognitive level** is important (inescapable?)
  - Explaining results and "reasoning"
- **Exchanges between AIs** is (currently) done at the level of training **examples** and the organization of **curricula**
- We are **far from**
  - **being able to "read"** the brain / mind
  - **being able to interact** at the level of **neurons** for conceptual exchanges

BUT ...

World Psychiatry
OFFICIAL JOURNAL OF THE WORLD PSYCHIATRIC ASSOCIATION (WPA)

World Psychiatry. 2019 Jun; 18(2): 119–129. PMCID: PMC6502424
Published online 2019 May 6. doi: 10.1002/wps.20617 PMID: 31059635

The "online brain": how the Internet may be changing our cognition

Joseph Firth,[1,2,3] John Torous,[4] Brendon Stubbs,[5,6] Josh A. Firth,[7,8] Genevieve Z. Steiner,[1,9] Lee Smith,[10] Mario Alvarez-Jimenez,[3,11] John Gleeson,[3,12] Davy Vancampfort,[13,14] Christopher J. Armitage,[2,15,16] and Jerome Sarris[1,17]

- Does AI plays a role in our brain?

BUT ...

Does the use of **computers**, the **Internet**, and

"intelligent" assistants

change our brain?