Faire de la science à partir des données ?

L'avènement du deep learning

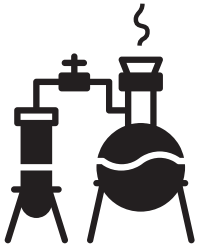
Journée Science Ouverte CNRS 2022 - Mercredi 30 novembre 2022

La science ouverte et les données de la recherche

Jean-Luc.Parouty@cnrs.fr

Scientific paradigms

1st paradigm



Experimental science

2nd paradigm

$$i\hbar\frac{d}{dt}|\Psi(t)\rangle = \hat{H}|\Psi(t)\rangle$$

$$\nabla \times H = J + \frac{\partial D}{\partial t}$$

$$F = G \cdot \frac{m_1 \cdot m_2}{r^2}$$

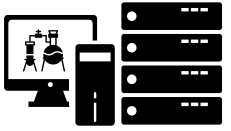
Theoretical science

3rd paradigm

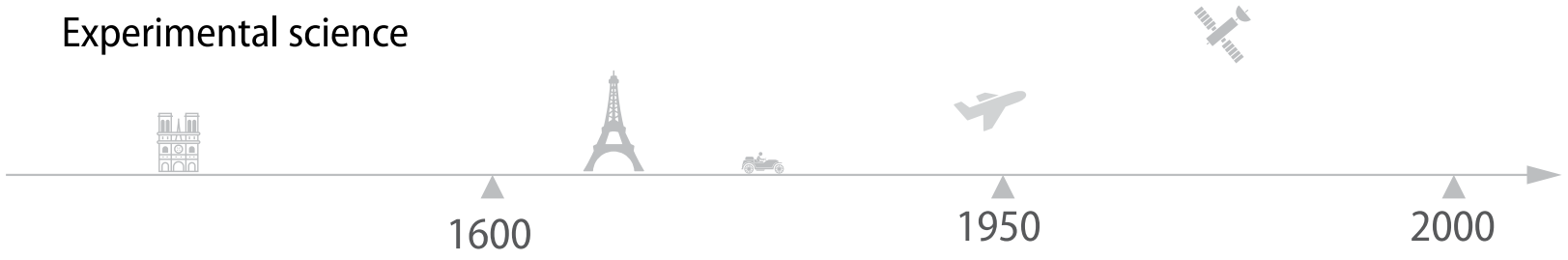
$$i\hbar\frac{d}{dt}|\Psi(t)\rangle = \hat{H}|\Psi(t)\rangle$$

$$\nabla \times H = J + \frac{\partial D}{\partial t}$$

$$F = G \cdot \frac{m_1 \cdot m_2}{r^2}$$

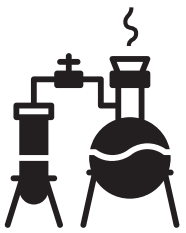


Computational science



Scientific paradigms

1st paradigm



Experimental science

2nd paradigm

$$i\hbar\frac{d}{dt}|\Psi(t)\rangle = \hat{H}|\Psi(t)\rangle$$

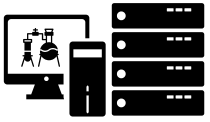
$$\nabla \times H = J + \frac{\partial D}{\partial t}$$

$$F = G \cdot \frac{m_1 \cdot m_2}{r^2}$$

Theoretical science

3rd paradigm

$$i\hbar\frac{d}{dt}|\Psi(t)\rangle = \hat{H}|\Psi(t)\rangle$$
$$\nabla \times H = J + \frac{\partial D}{\partial t}$$
$$F = G \cdot \frac{m_1 \cdot m_2}{r^2}$$



Computational science

4th paradigm[1]

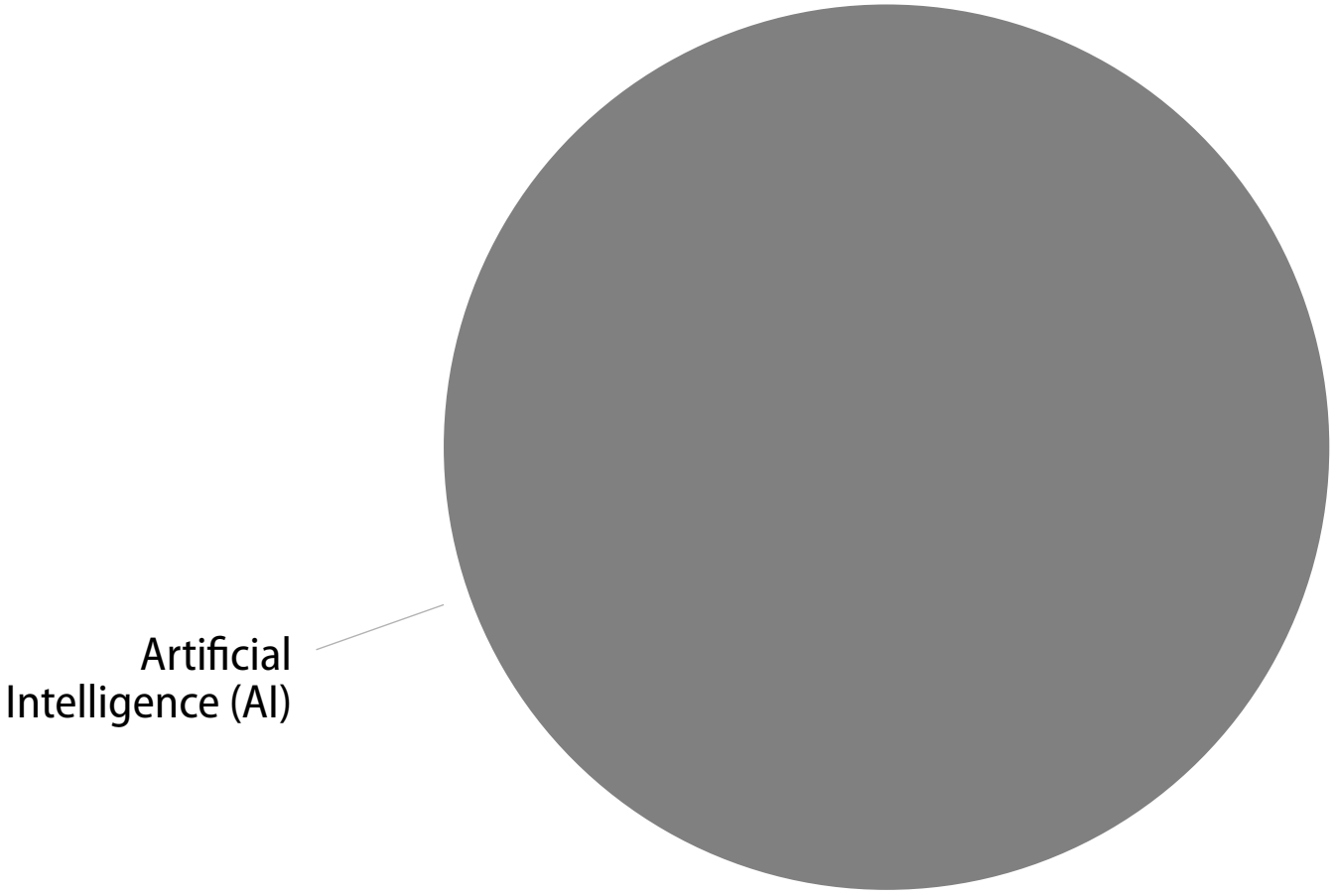
Data-driven science

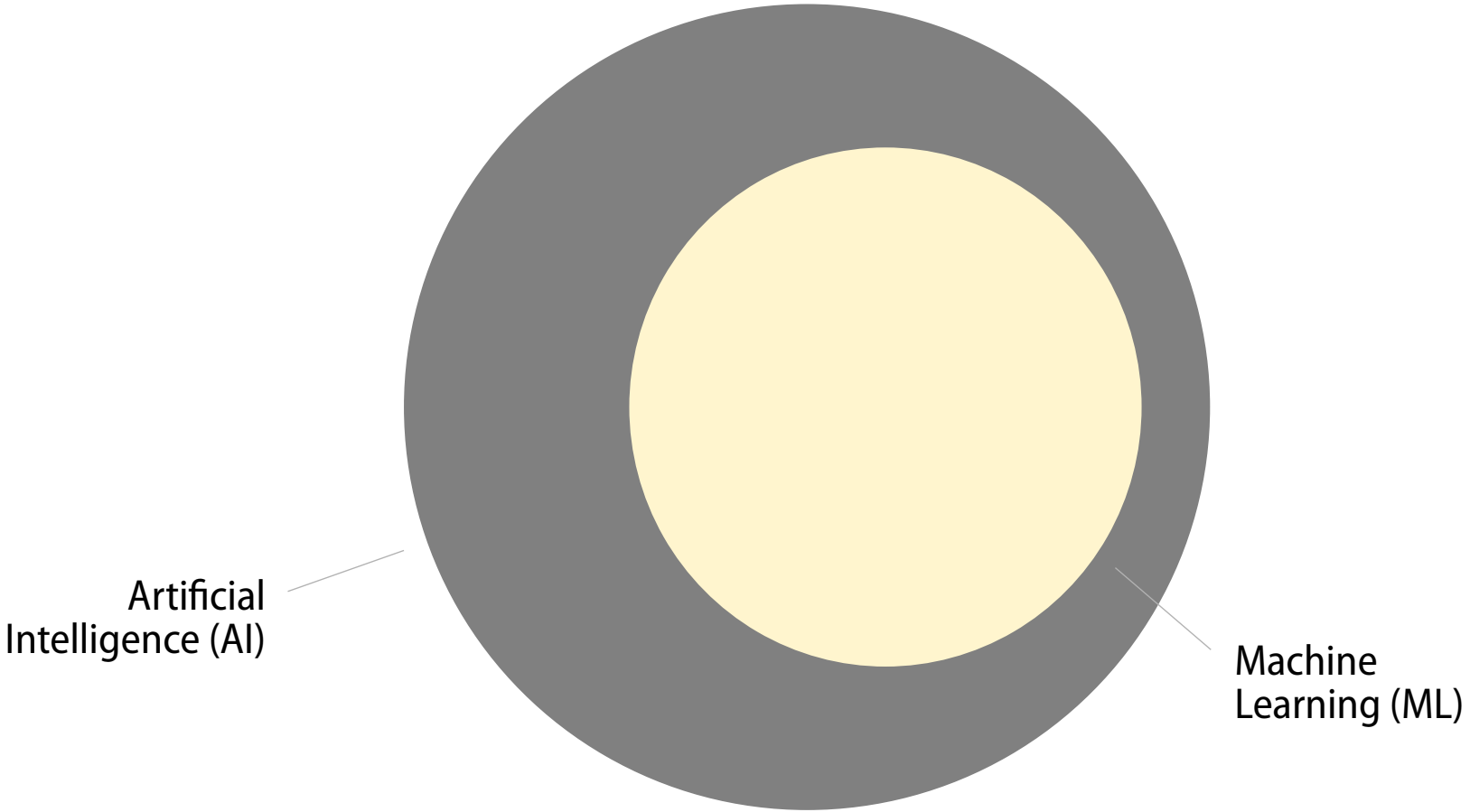
1600

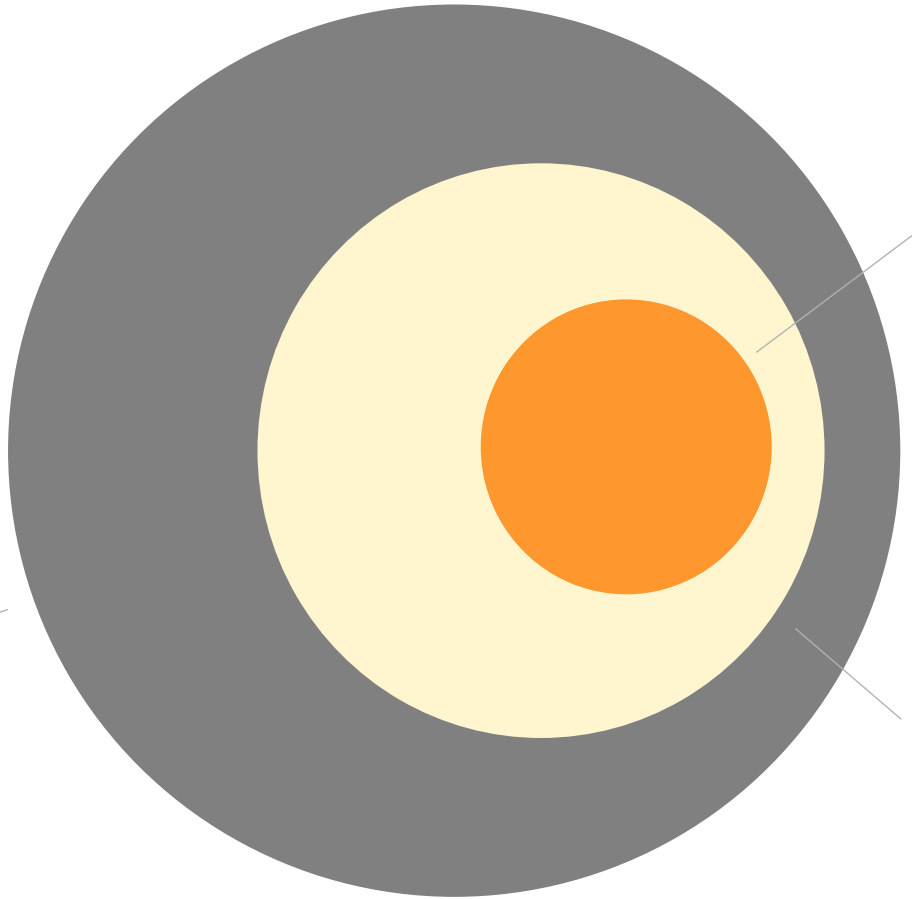
1950

2000

[1] Jim Gray, 2007 [GRAY]


Artificial
Intelligence (AI)


Artificial
Intelligence (AI)
Machine
Learning (ML)



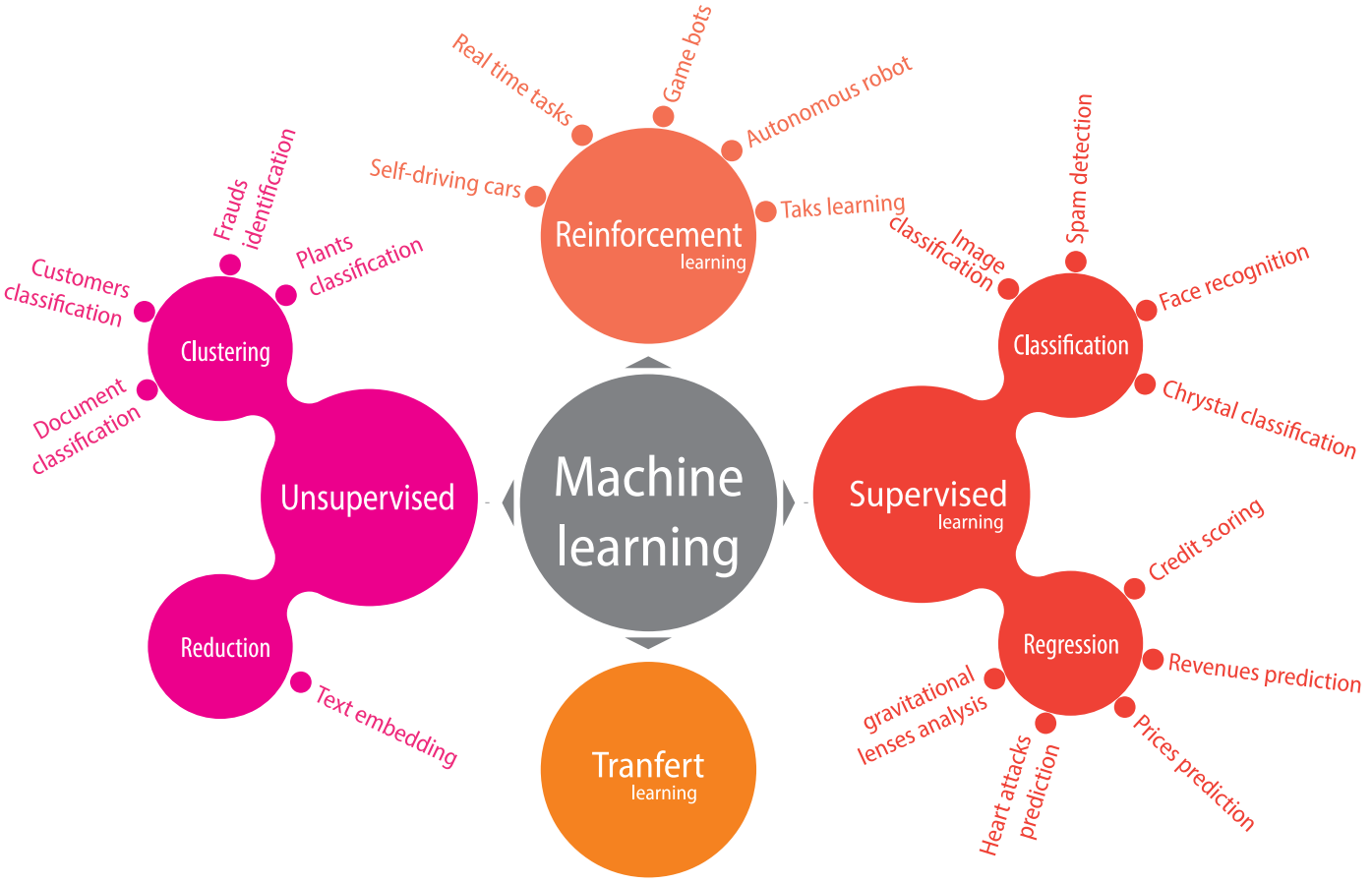
Deep Learning (DL)

Based on artificial neural networks

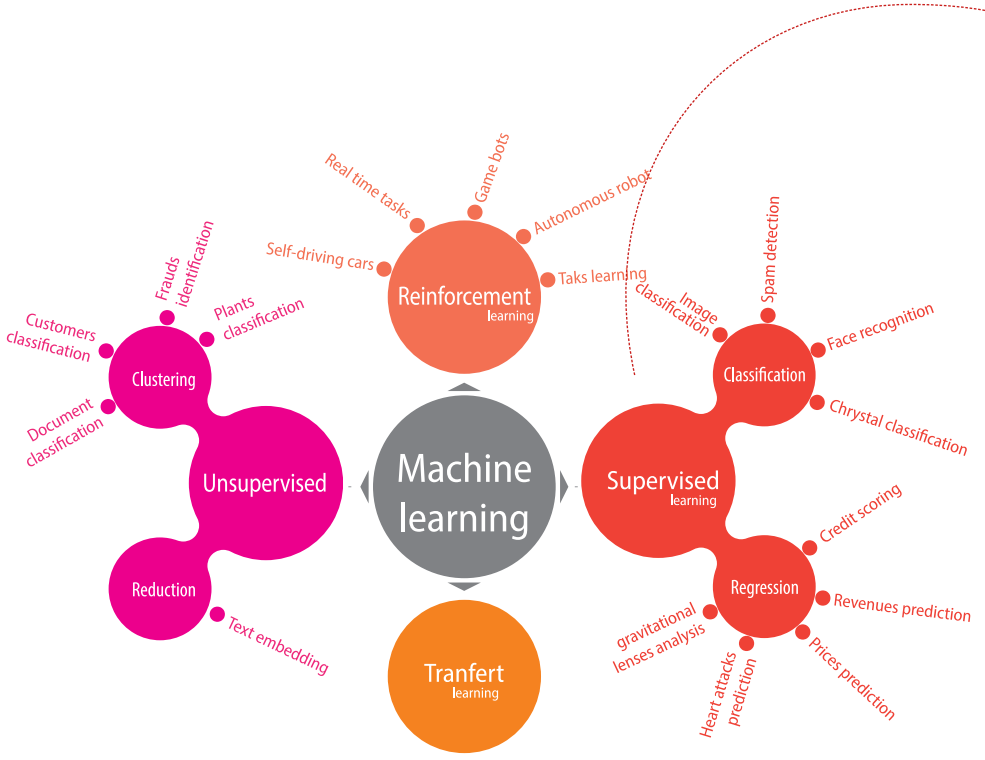
Artificial Intelligence (AI)

Machine Learning (ML)

[*-learning]

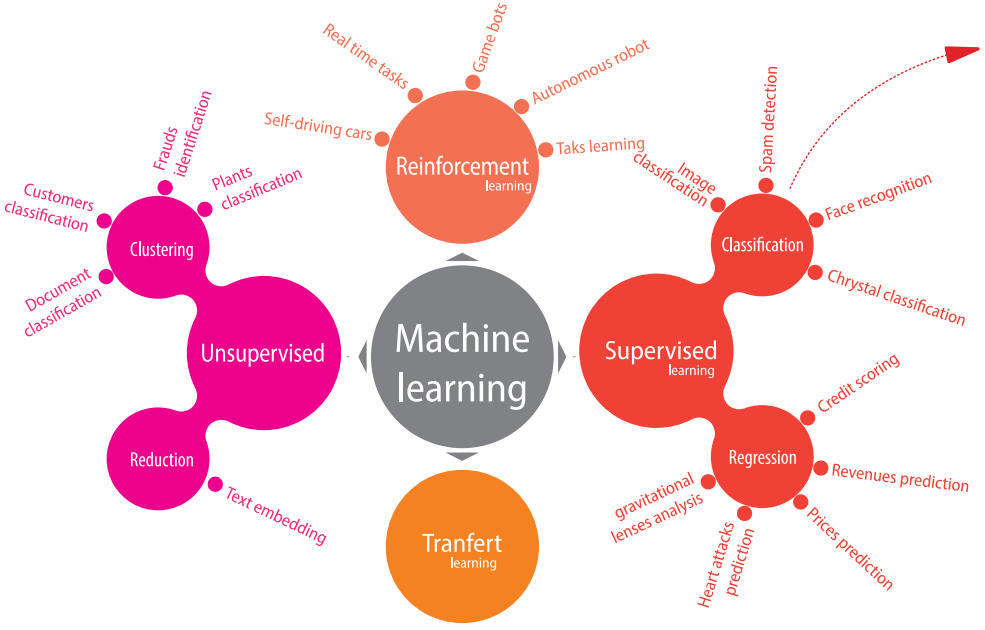

Machine learning
Supervised learning
Classification
Face recognition
Chrystal classification
Spam detection
Image classification
Regression
Credit scoring
Revenues prediction
Prices prediction
Heart attacks prediction
gravitational lenses analysis
Unsupervised
Clustering
Frauds identification
Plants classification
Customers classification
Document classification
Reduction
Text embedding
Reinforcement learning
Real time tasks
Game bots
Autonomous robot
Taks learning
Self-driving cars
Tranfert learning

Supervised learning



Learning
from **examples**

Supervised learning



Classification :

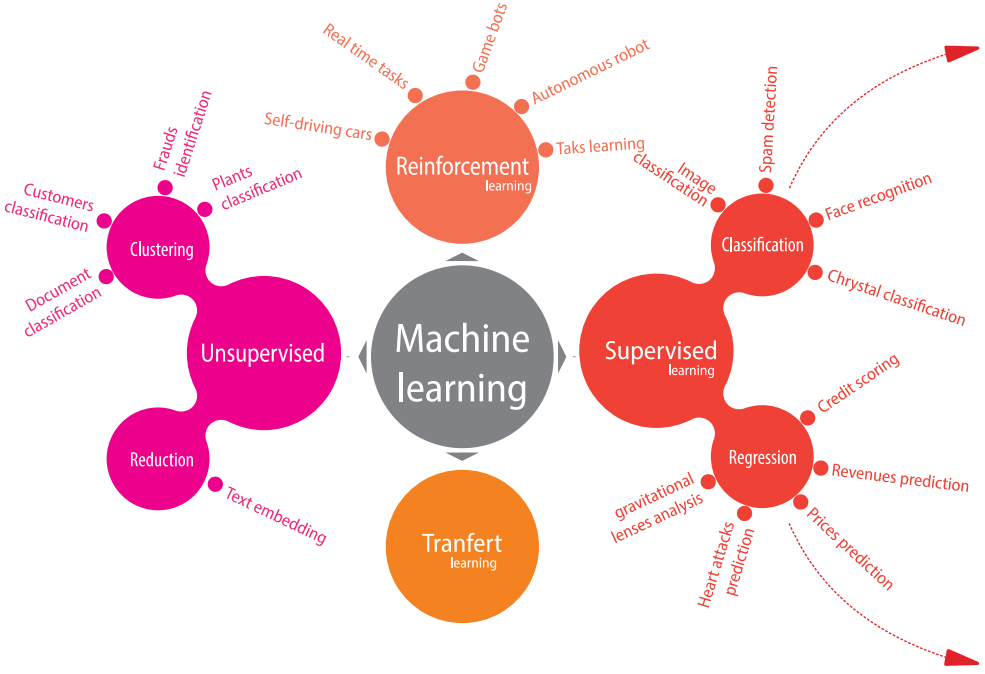
Predict qualitative informations

This is a cat

This is a rabbit

Tell me,
what is it ?

Supervised learning



Classification :

Predict qualitative informations

This is a cat

This is a rabbit

Tell me, what is it ?

Régression :

Predict quantitative informations

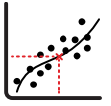
150 K€

400 K€

120 K€

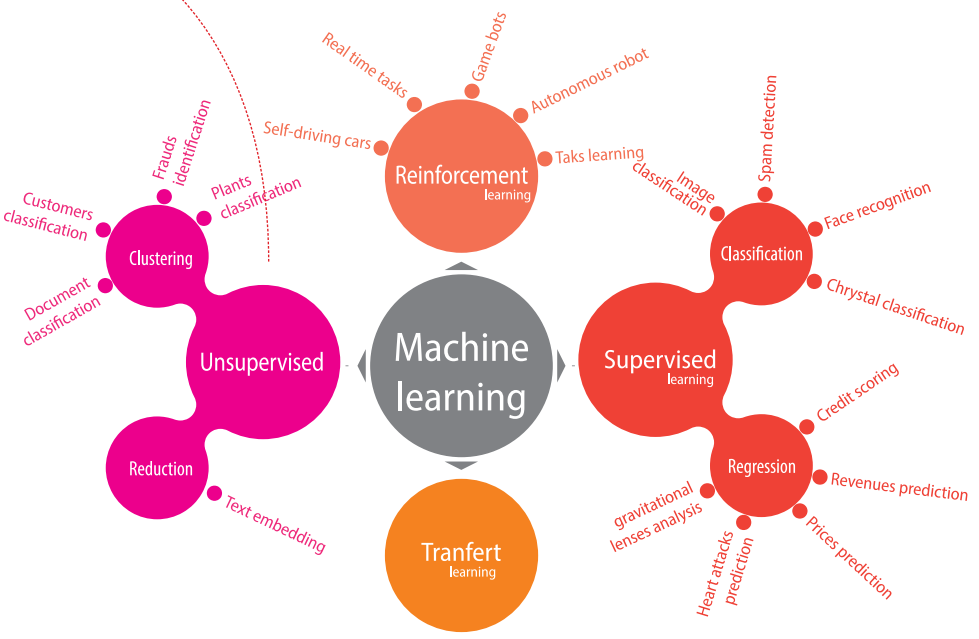
100 K€

Tell me, what's the price ?



Unsupervised learning

Learning from data **alone**

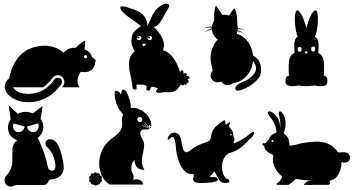


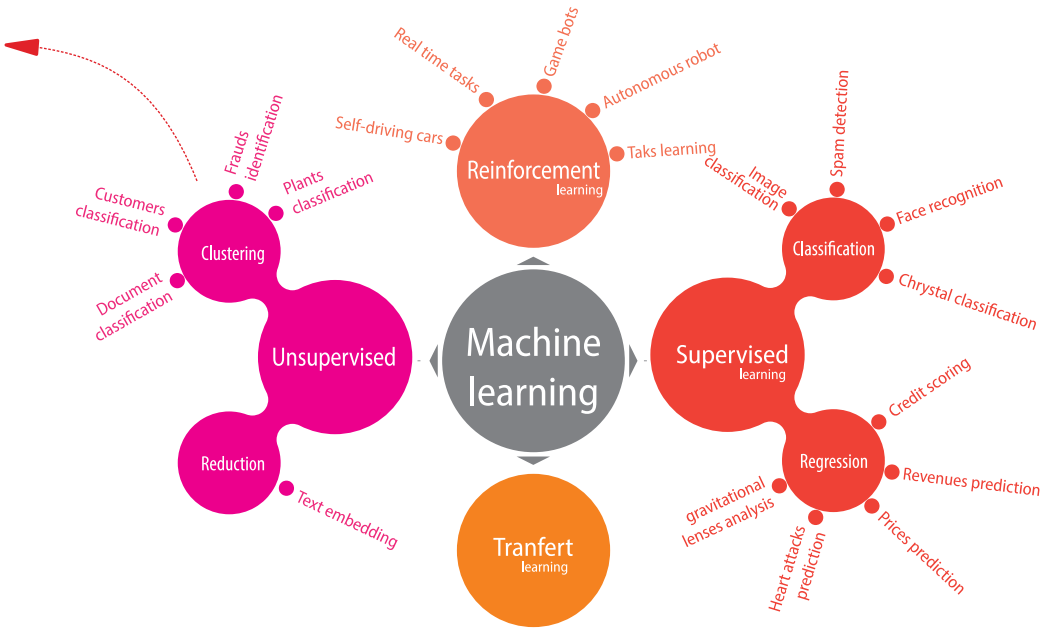
Unsupervised learning

Learning from data alone

Clustering :
Finding Common Relationships

What is the relationship between these data ?



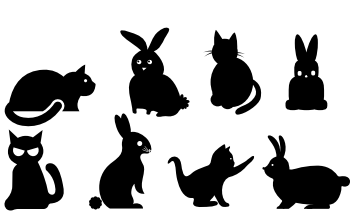


Unsupervised learning

Learning from data alone

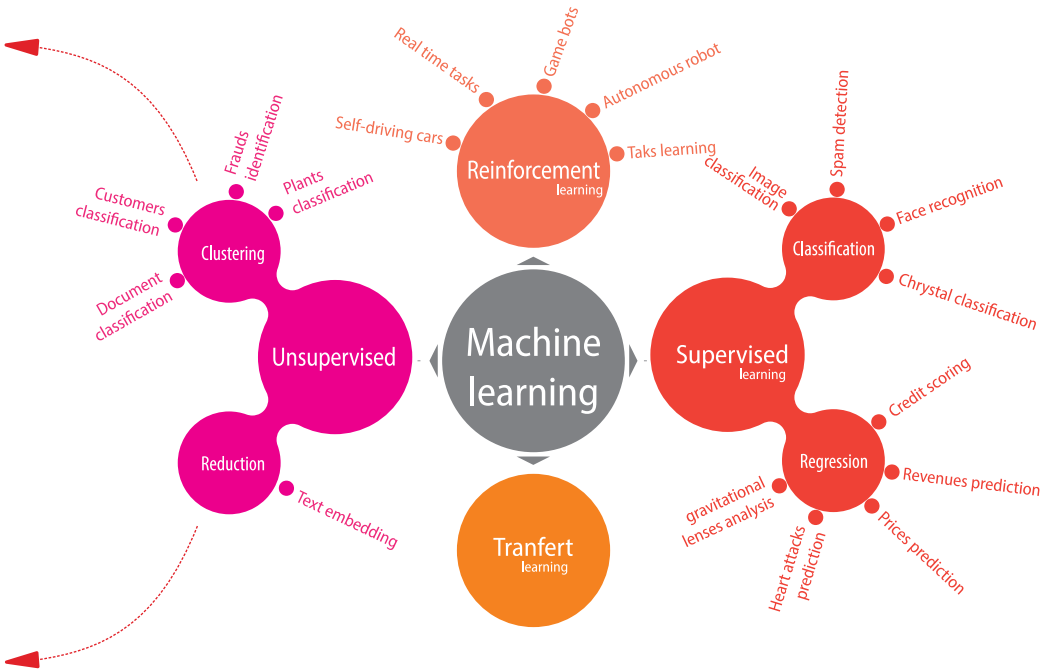
Clustering :
Finding Common Relationships

What is the relationship between these data ?

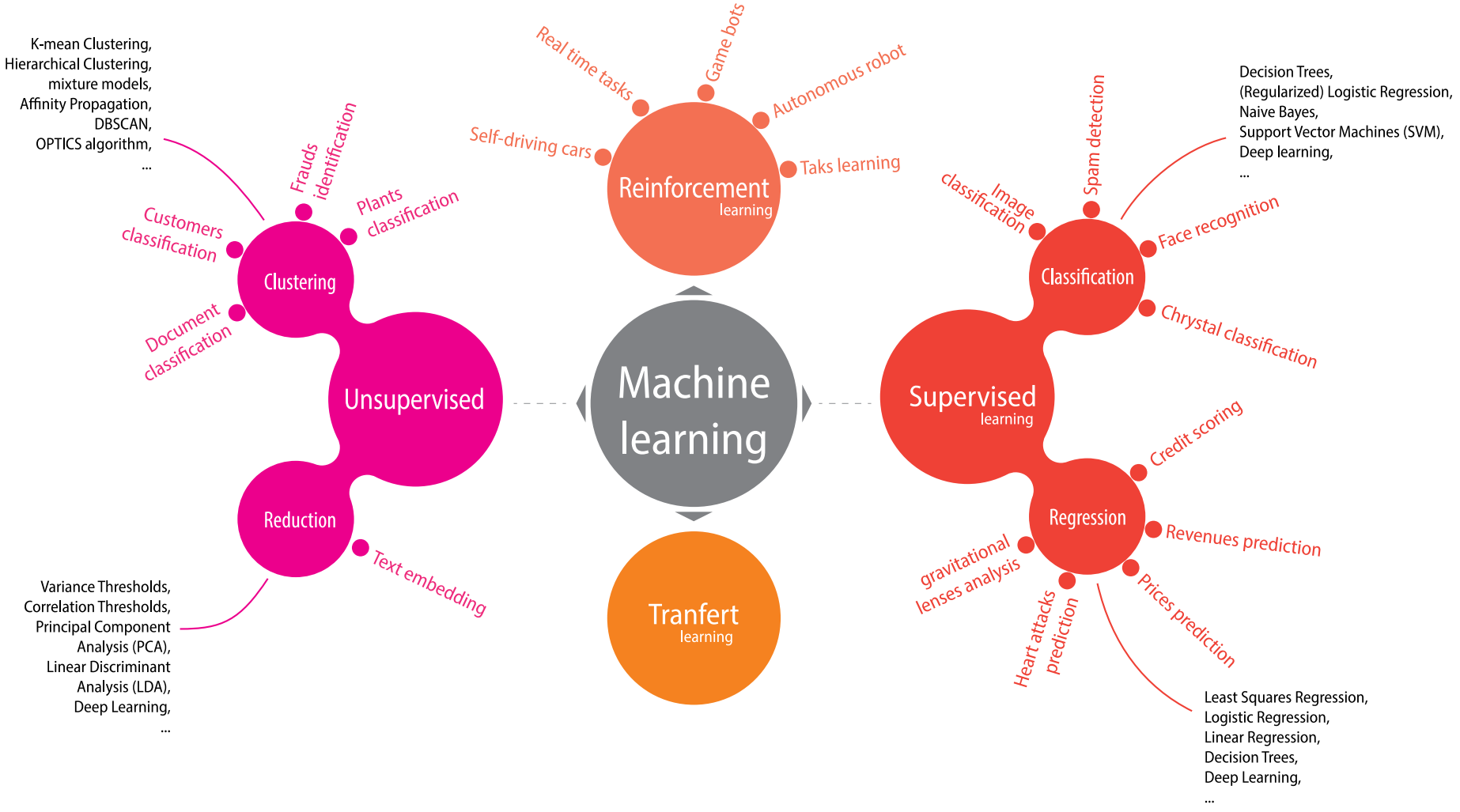


Reduction :
Reduce the number of dimensions

Simplify while keeping meaning



[*-learning]


Machine learning
Unsupervised
Clustering
K-mean Clustering,
Hierarchical Clustering,
mixture models,
Affinity Propagation,
DBSCAN,
OPTICS algorithm,
...
Frauds identification
Plants classification
Customers classification
Document classification
Reduction
Variance Thresholds,
Correlation Thresholds,
Principal Component
Analysis (PCA),
Linear Discriminant
Analysis (LDA),
Deep Learning,
...
Text embedding
Reinforcement learning
Real time tasks
Game bots
Autonomous robot
Self-driving cars
Taks learning
Tranfert learning
Supervised learning
Classification
Decision Trees,
(Regularized) Logistic Regression,
Naive Bayes,
Support Vector Machines (SVM),
Deep learning,
...
Spam detection
Image classification
Face recognition
Chrystal classification
Regression
Credit scoring
Revenues prediction
Prices prediction
Least Squares Regression,
Logistic Regression,
Linear Regression,
Decision Trees,
Deep Learning,
...
gravitational lenses analysis
Heart attacks prediction

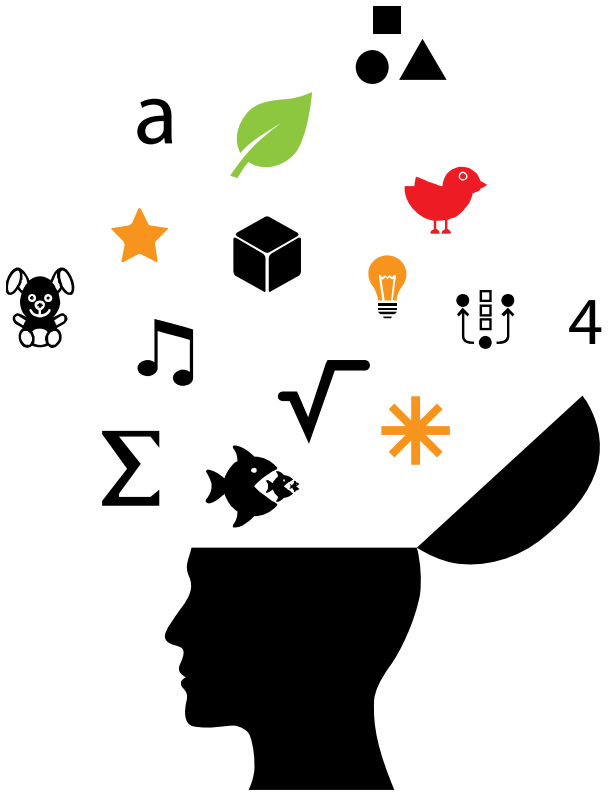
[intelligence]

[intelligence]

« Capacité de percevoir ou d'inférer l'information, et de la conserver comme une connaissance à appliquer à des comportements adaptatifs dans un environnement ou un contexte donné »

« Ability to perceive or infer information, and to retain it as knowledge to be appli towards adaptive behaviors within an environment or context »[*]

[*] Wikipedia - Illustration : [POTATO]



[intelligence]

« Ensemble des fonctions mentales ayant pour objet la connaissance conceptuelle et rationnelle »*

« Set of mental functions aimed at conceptual and rational knowledge »

* https://www.larousse.fr

The big Controversy

Modelling the brain :

« Penser s'apparente
à un calcul massivement parallèle de
fonctions élémentaires.

L'information est un **signal** avant
d'être un code »[1]

Making a mind :

« Penser, c'est calculer des **symboles** qui
ont à la fois une réalité matérielle et une
valeur sémantique de représentation »[1]

L'information est une donnée
symbolique de **haut niveau**.

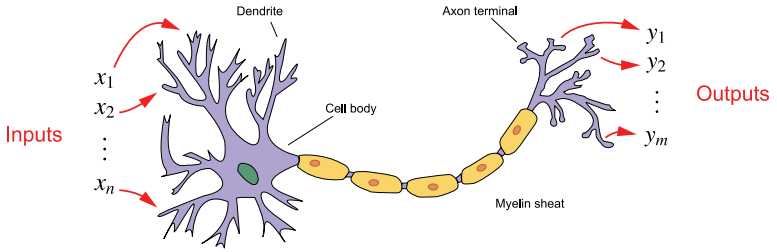
Connectionnism

Modelling the brain
Modéliser le cerveau

vs

Symbolic

Making a mind
Forger une opinion




```
Tout [homme] est [mortel]
[Socrate] est un [homme]

Donc [Socrate] est [mortel]
```

[1] Dominique Cardon, Jean-Philippe Cointet, Antoine Mazieres (2018) [LRDN]

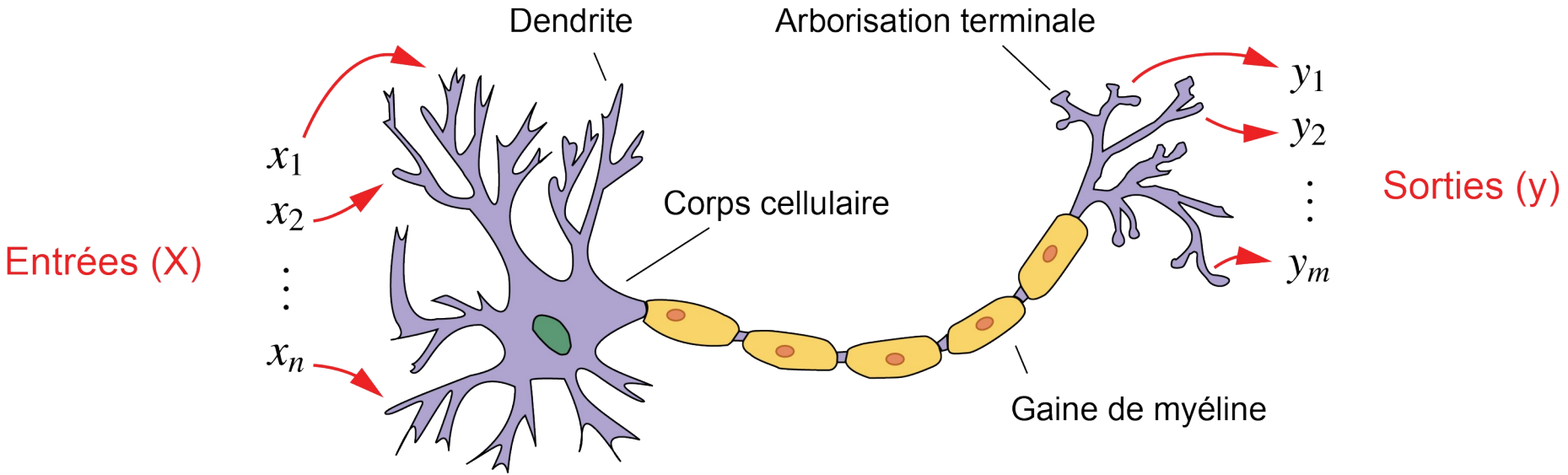
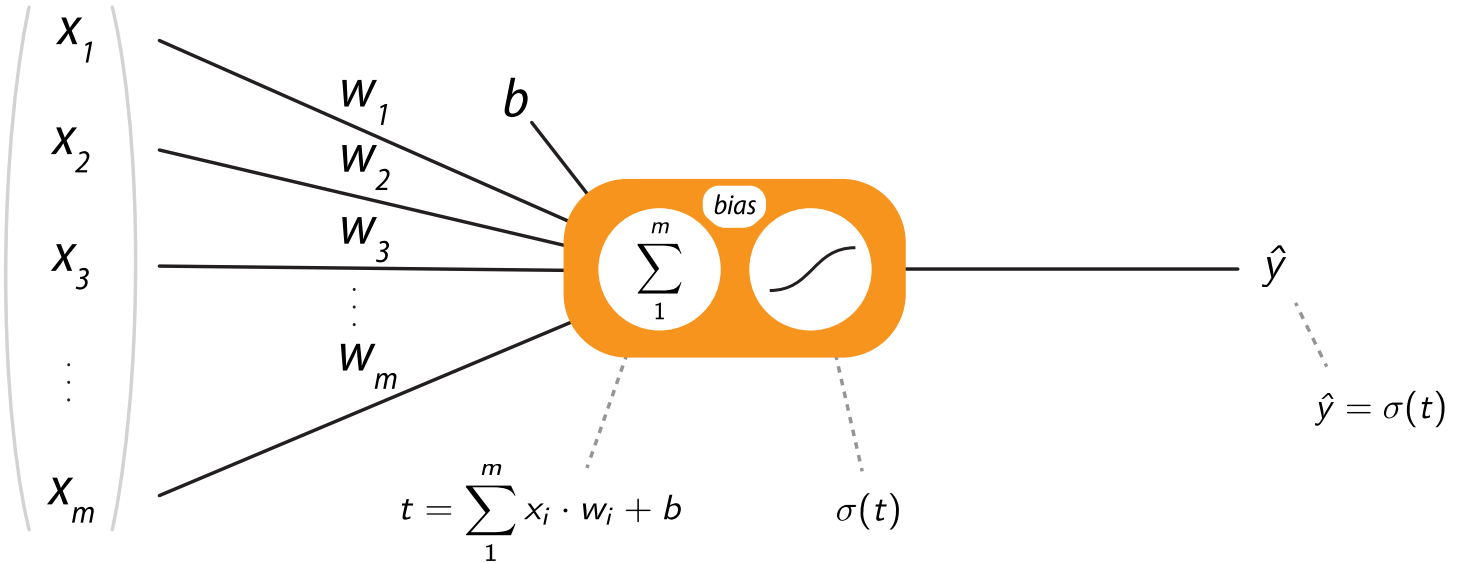


Illustration : [NEURON]

$$\hat{y} = \sigma(\Theta^T \cdot X + b)$$



Input	Bias / Weight	Activation function	Output
X	Θ, b	$\sigma(t)$	$\hat{y}$

The big Controversy

Modelling the brain :

« Penser s'apparente
à un calcul massivement parallèle de
fonctions élémentaires.

L'information est un **signal** avant
d'être un code »[1]

Making a mind :

« Penser, c'est calculer des **symboles** qui
ont à la fois une réalité matérielle et une
valeur sémantique de représentation »[1]

L'information est une donnée
symbolique de **haut niveau**.

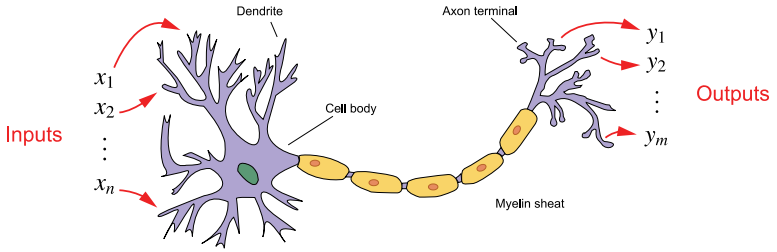
Connectionnism

Modelling the brain
Modéliser le cerveau

vs

Symbolic

Making a mind
Forger une opinion



Tout [homme] est [mortel]
[Socrate] est un [homme]

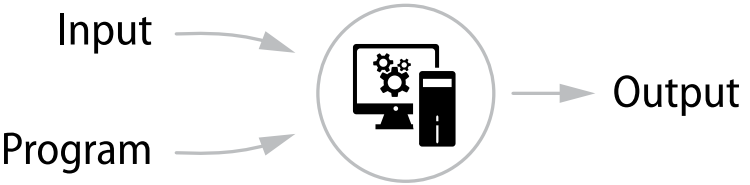
Donc [Socrate] est [mortel]

[1] D Cardon, JP Cointet, A Mazieres, 2018 [LRDN]

The big Controversy

Inductive approach

Deductive approach



Connectionnism vs **Symbolic**

Facts ▶ Rules and laws

Model

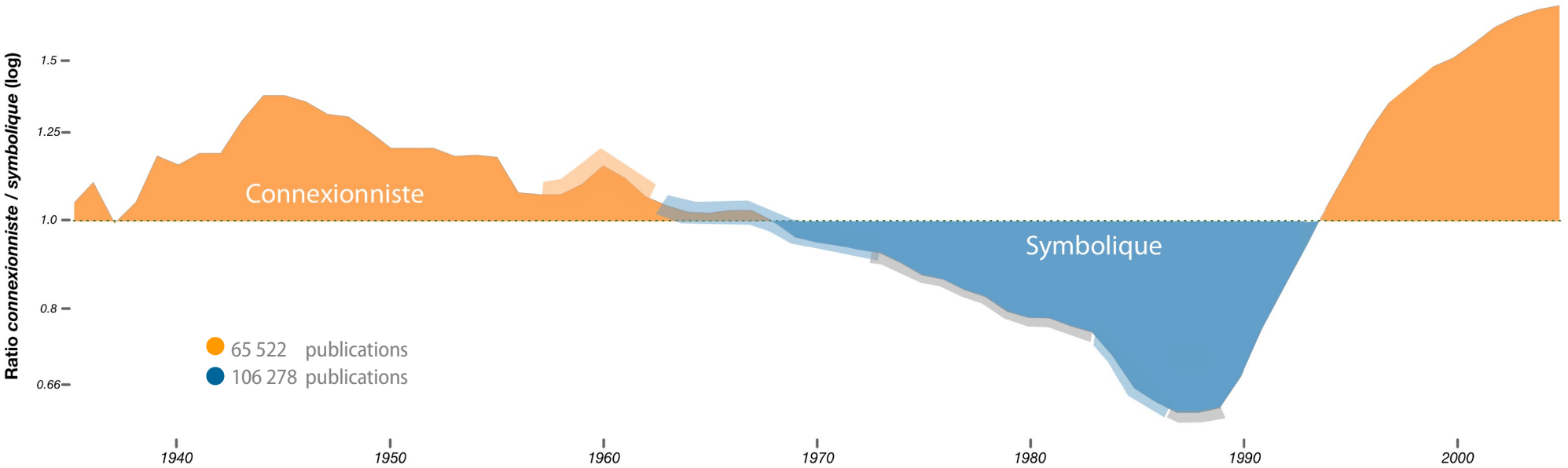
Expert

Rules and laws ▶ Special case

The big Controversy

Evolution of the academic influence of connexionist and symbolic approaches[1]

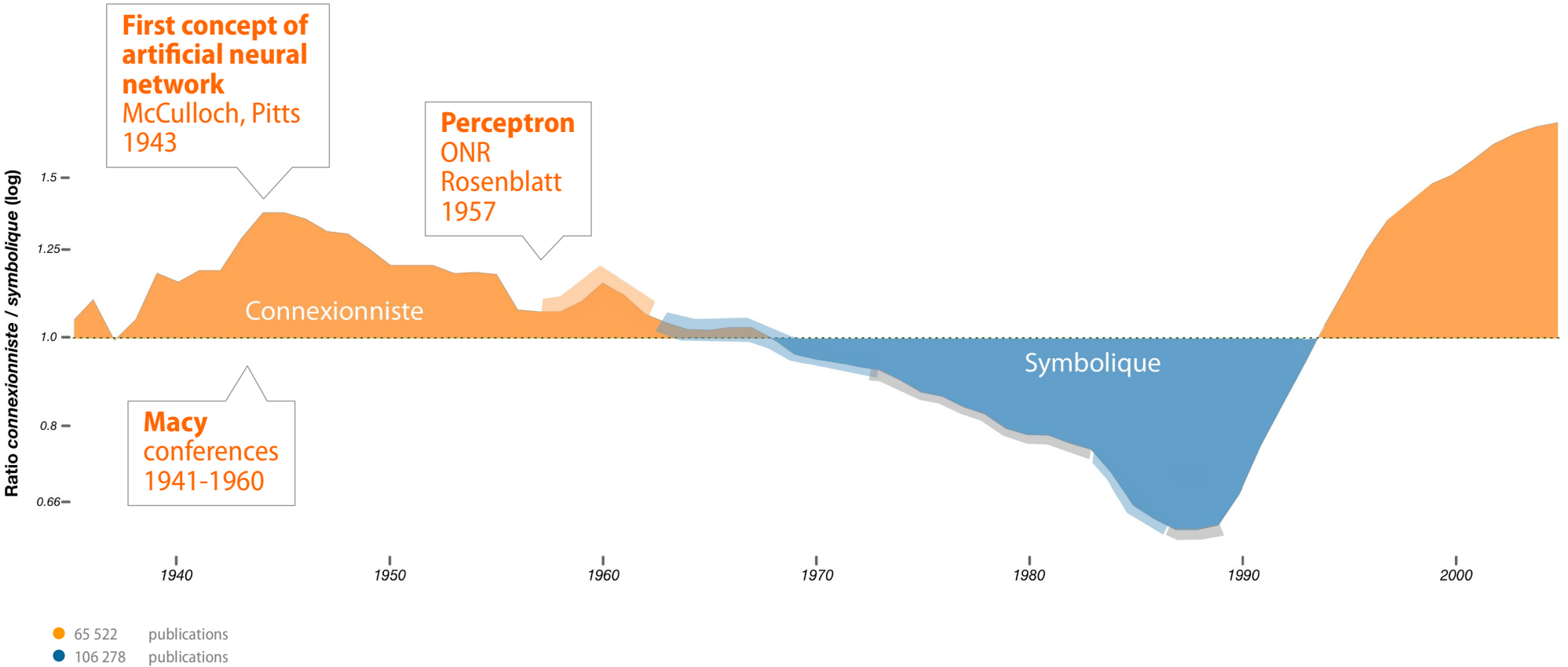
Ration of publications between connexionists and symbolists



[1] D Cardon, JP Cointet, A Mazieres, 2018 [LRDN]

The big Controversy

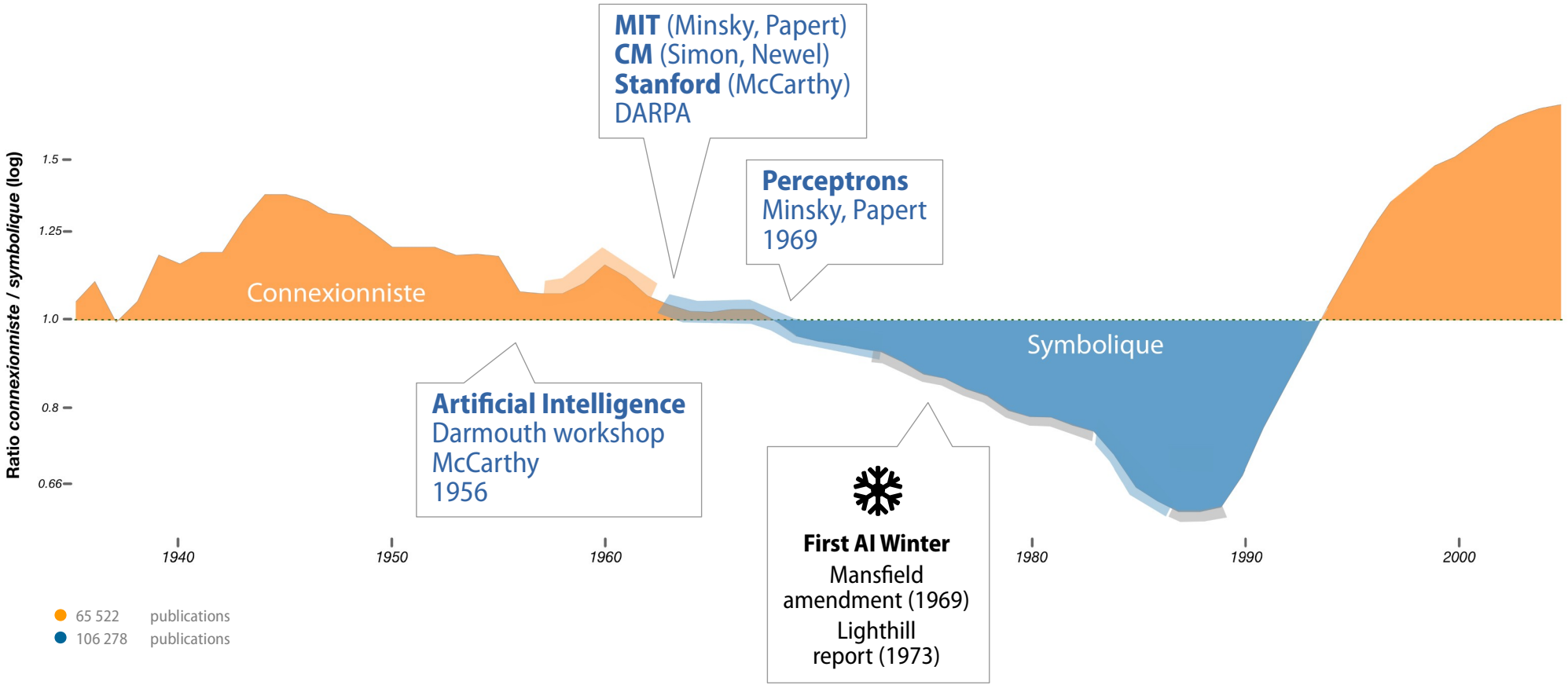
Evolution of the academic influence of connexionist and symbolic approaches[1]



[1] D Cardon, JP Cointet, A Mazieres, 2018 [LRDN]

The big Controversy

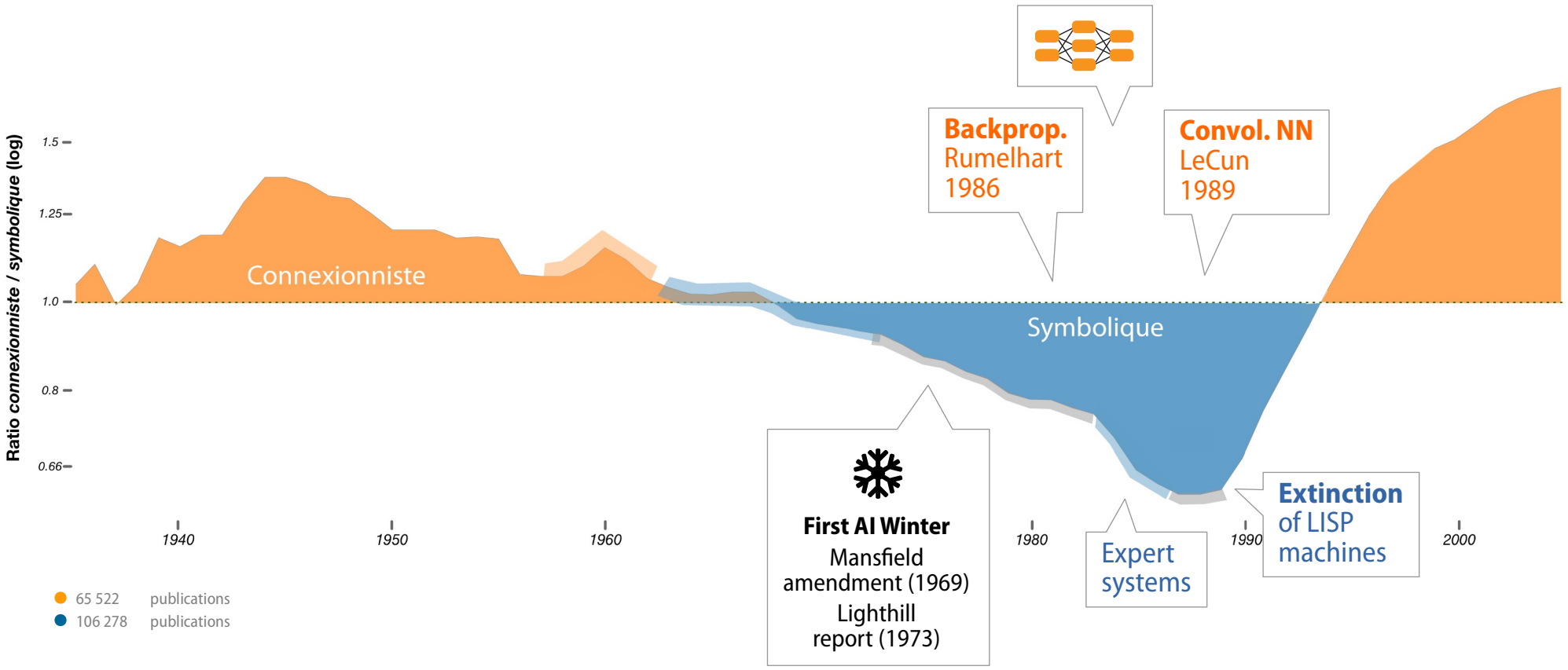
Evolution of the academic influence of connexionist and symbolic approaches[1]



[1] D Cardon, JP Cointet, A Mazieres, 2018 [LRDN]

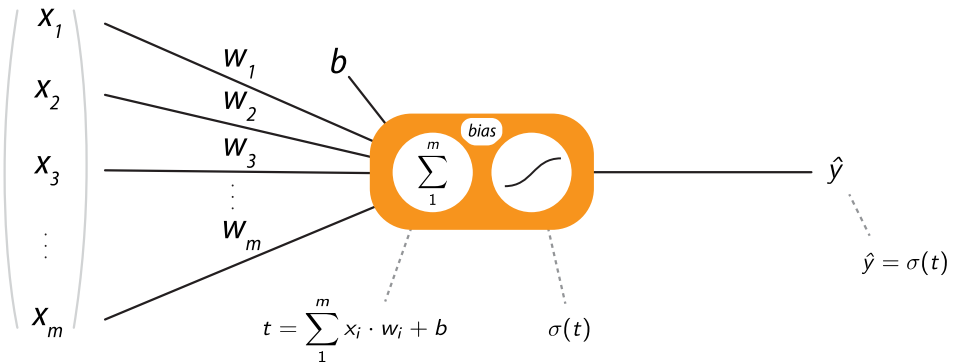
The big Controversy

Evolution of the academic influence of connexionist and symbolic approaches[1]

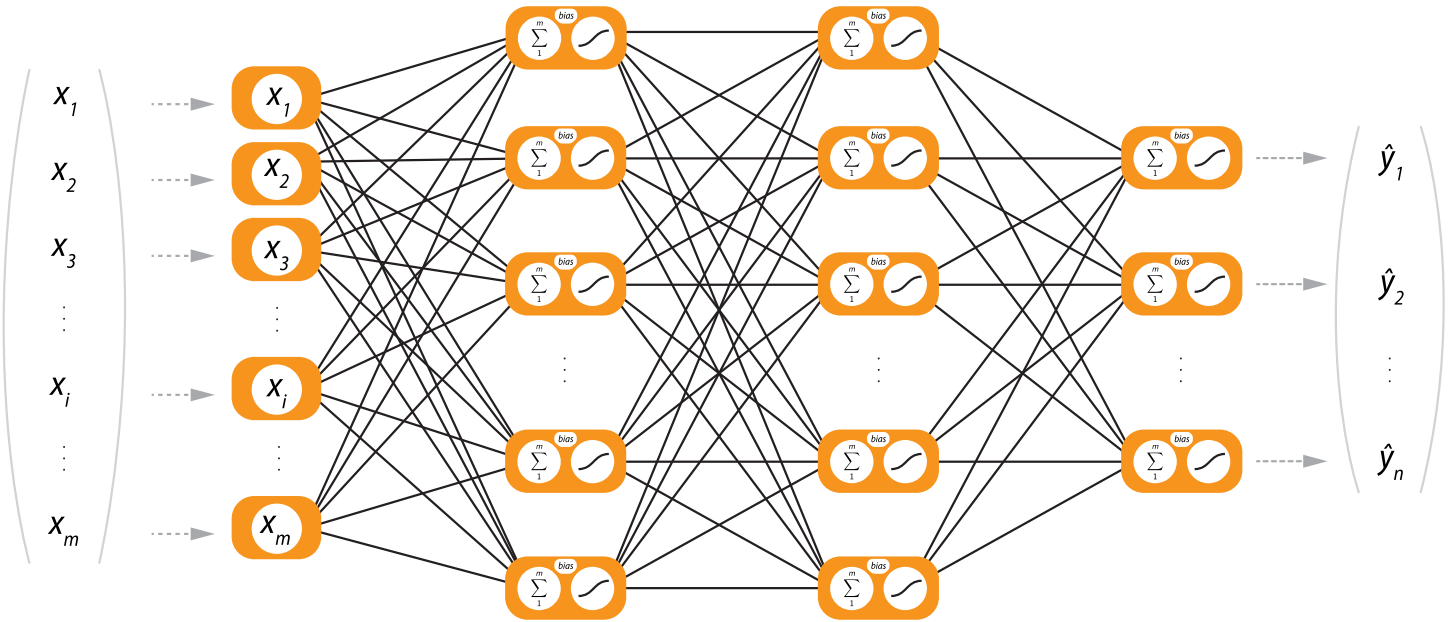


[1] D Cardon, JP Cointet, A Mazieres, 2018 [LRDN]

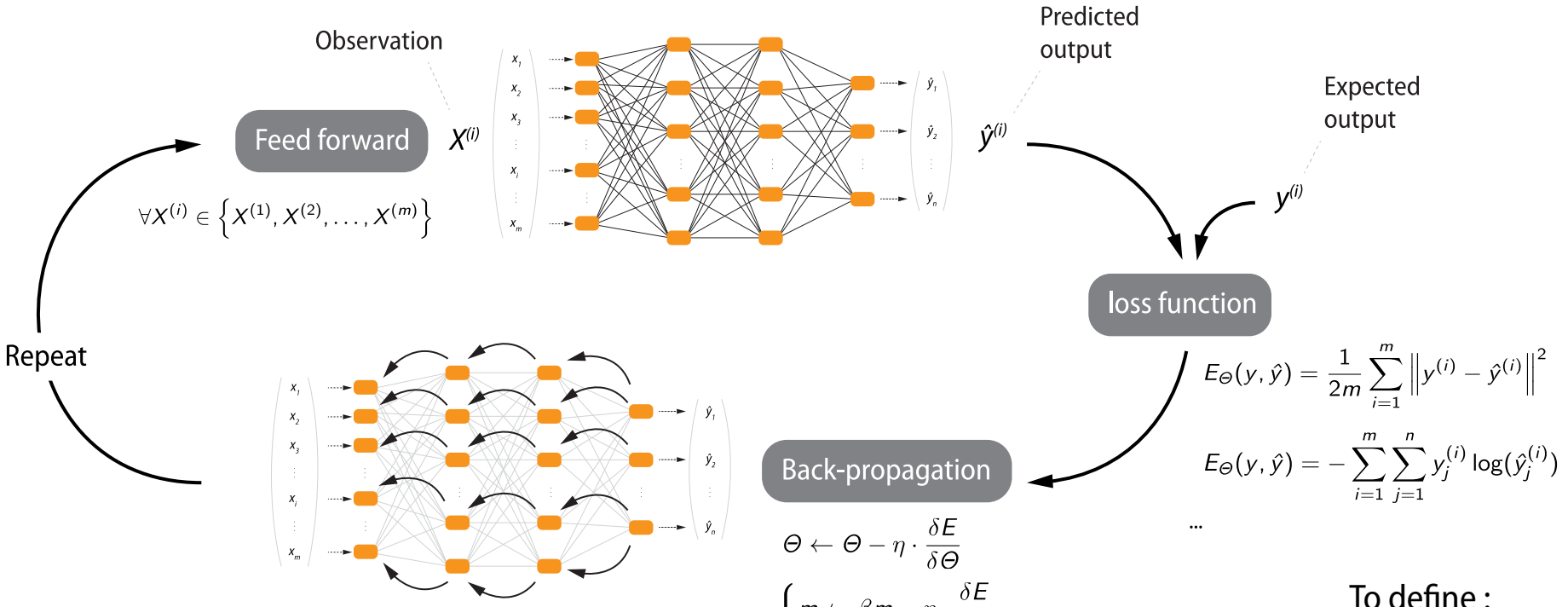
Deep Neural Networks



Deep Neural Networks



Deep Neural Networks

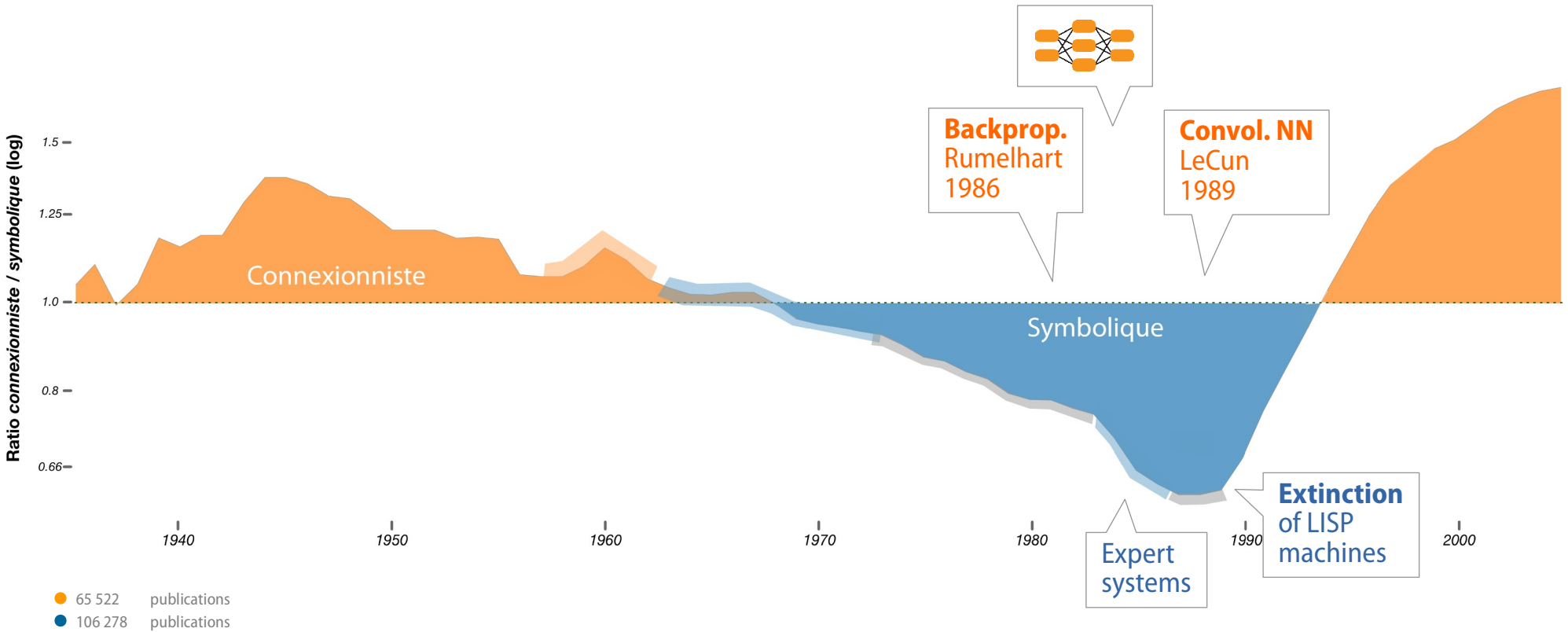


Back-propagation

Learning process

The big Controversy

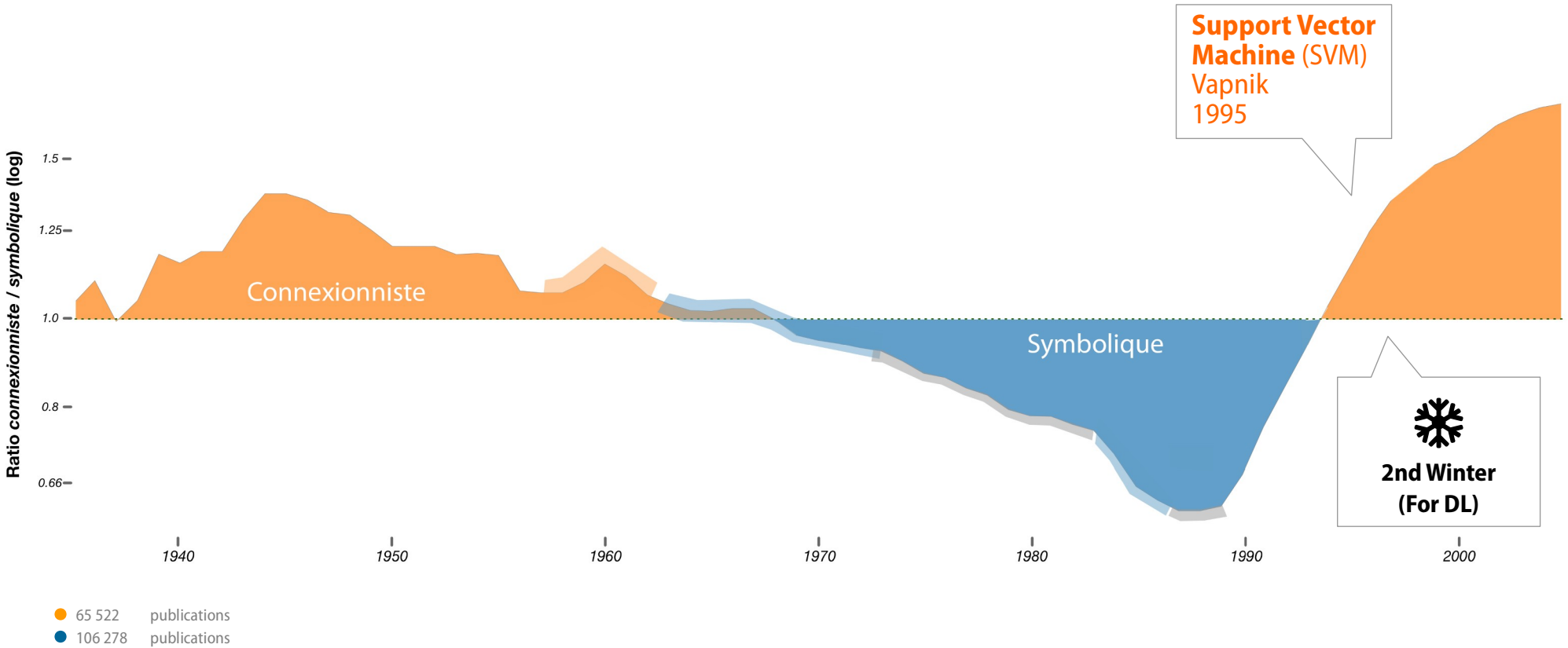
Evolution of the academic influence of connexionist and symbolic approaches[1]



[1] D Cardon, JP Cointet, A Mazieres, 2018 [LRDN]

The big Controversy

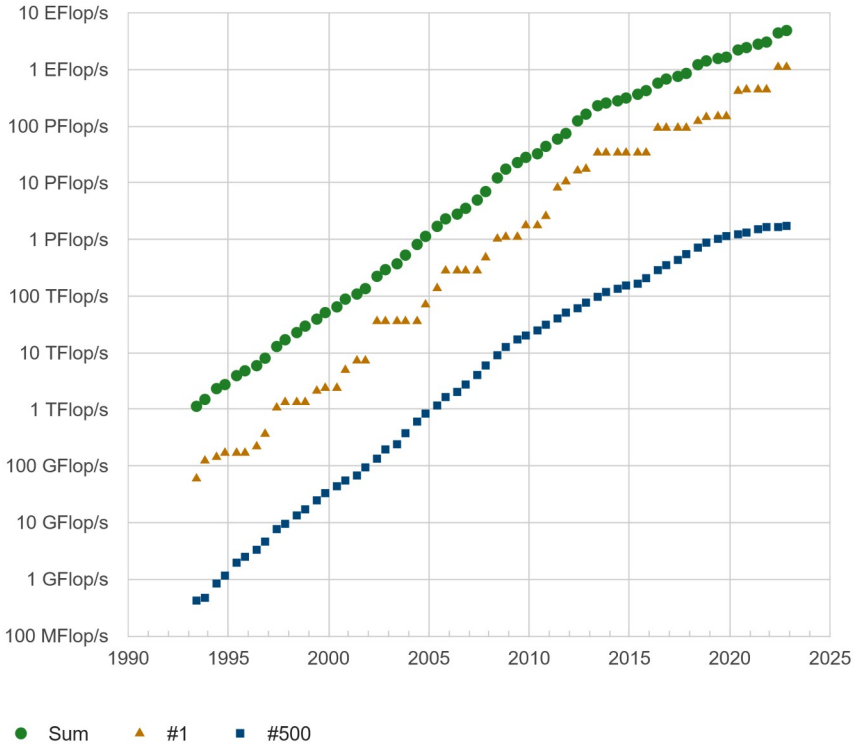
Evolution of the academic influence of connexionist and symbolic approaches[1]



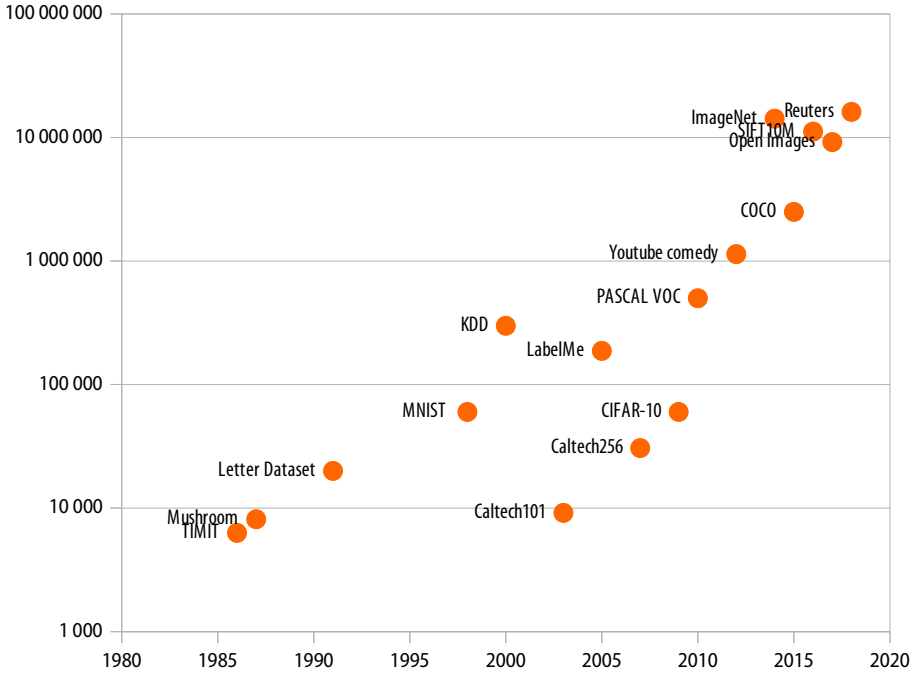
[1] D Cardon, JP Cointet, A Mazieres, 2018 [LRDN]

The big Controversy

Performance Development[1]



Datasets for machine-learning[2]



Laboratoire
Cas particulier

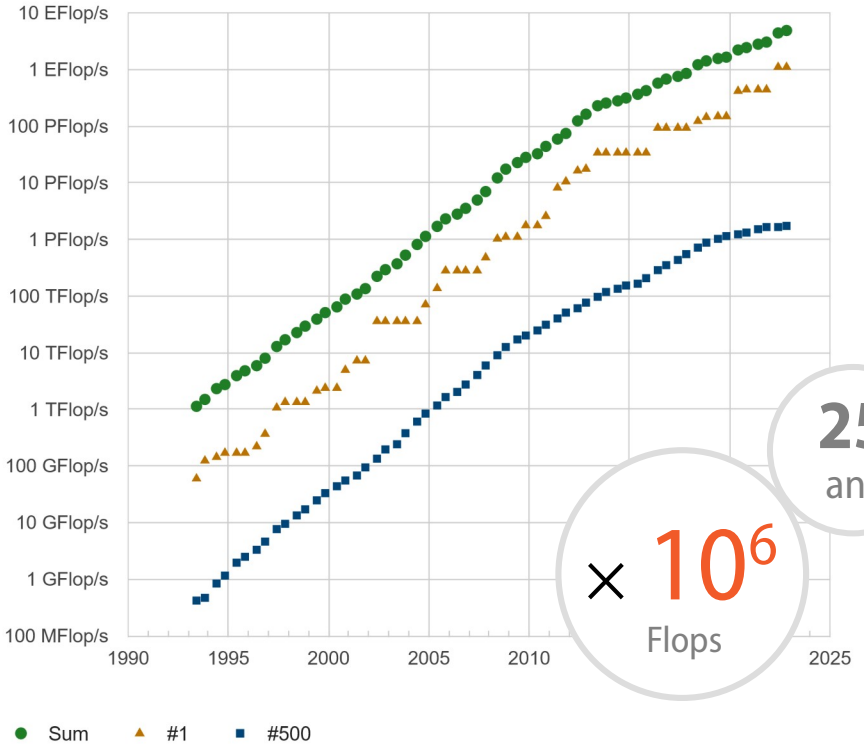
Monde réel

[1] TOP500 List [TOP500]

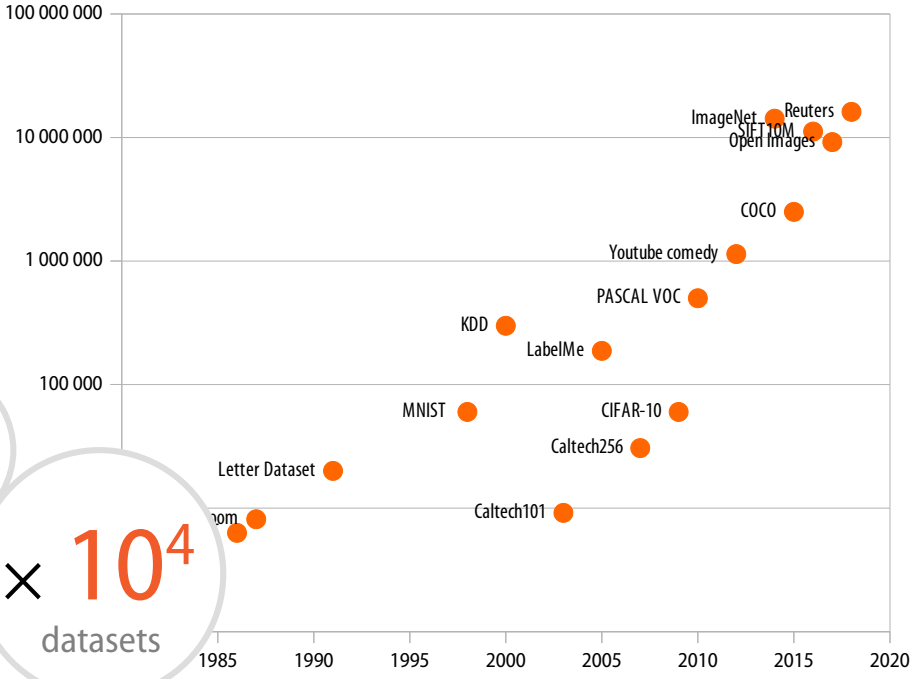
[2] Wikipedia [WKP1]

The big Controversy

Performance Development[1]



Datasets for machine-learning[2]



25 ans

× 10^6 Flops

× 10^4 datasets

Laboratoire
Cas particulier

→

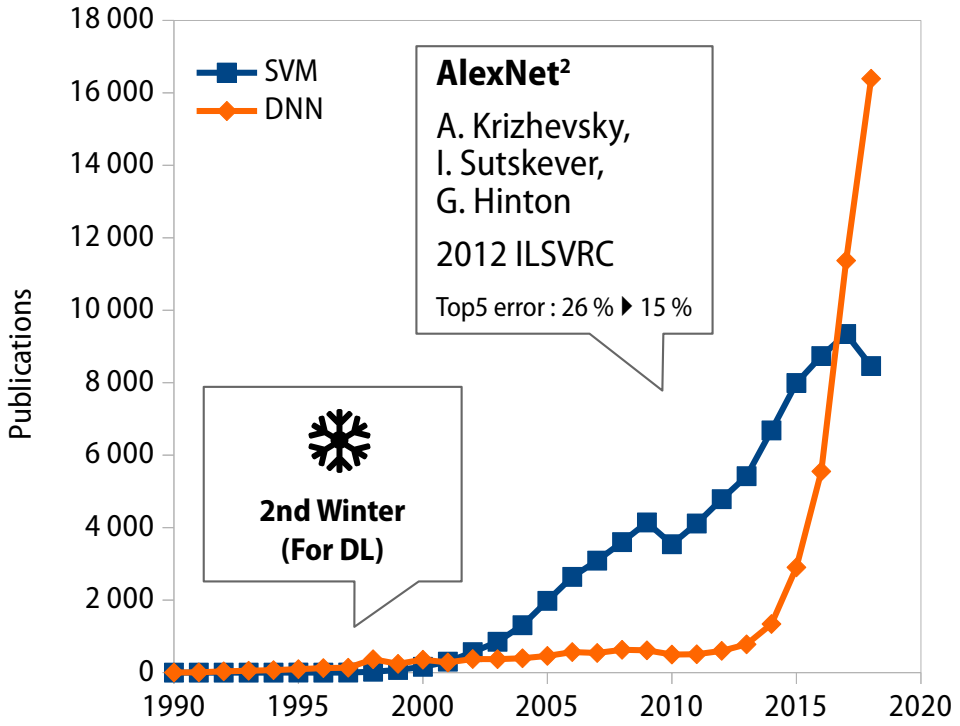
Monde réel

[1] TOP500 List [TOP500]

[2] Wikipedia [WKP1]

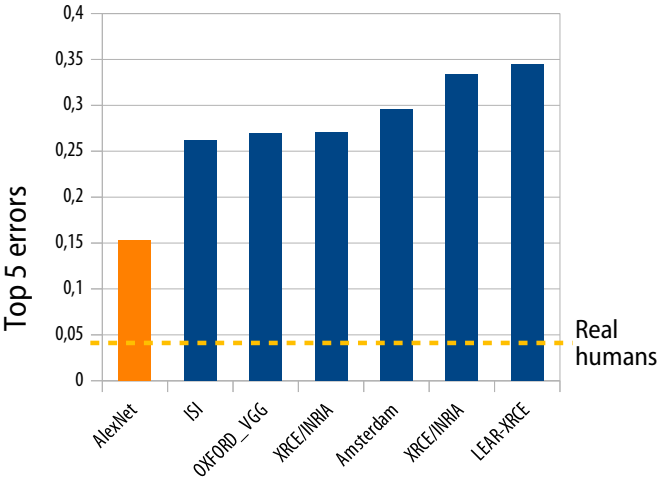
The big Controversy

Publications SVM vs DNN[1]



DNN

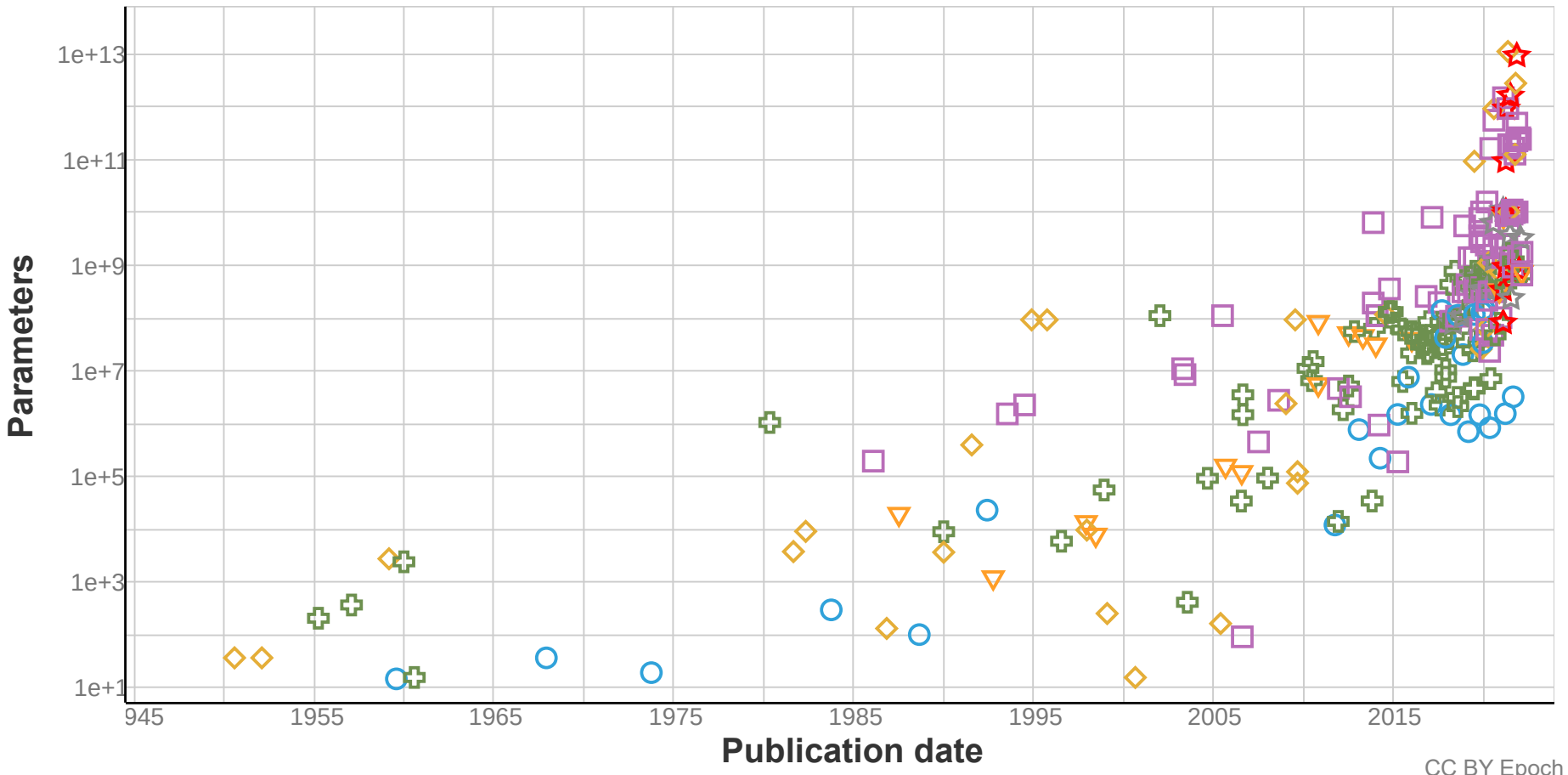
Images classification
Top 5 error at ILSVRC 2012[3,4]



Without mathematical guarantee, DNN have proven to be more effective in the face of the complexity of the real world !

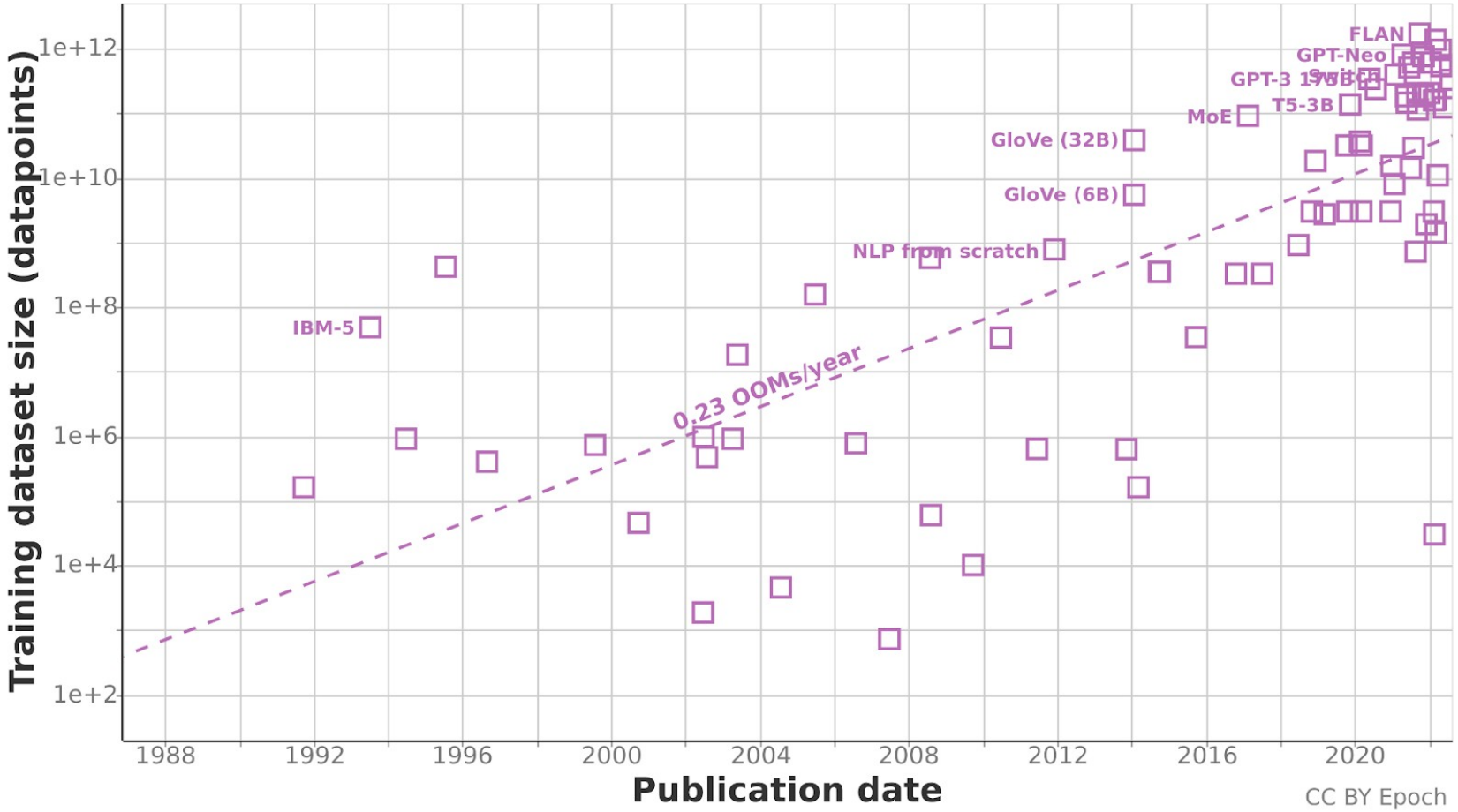
[1] Web of Science [WOS1][WOS2]
[2] AlexNet [ALEX]
[3] ImageNet Large Scale Visual Recognition [ILSVRC]
[4] Similar evolution in Natural language processing, translation, board games, etc. See : DeepL.com, AlphaGo, AlphaZero, ...

Machine Learning Model Sizes



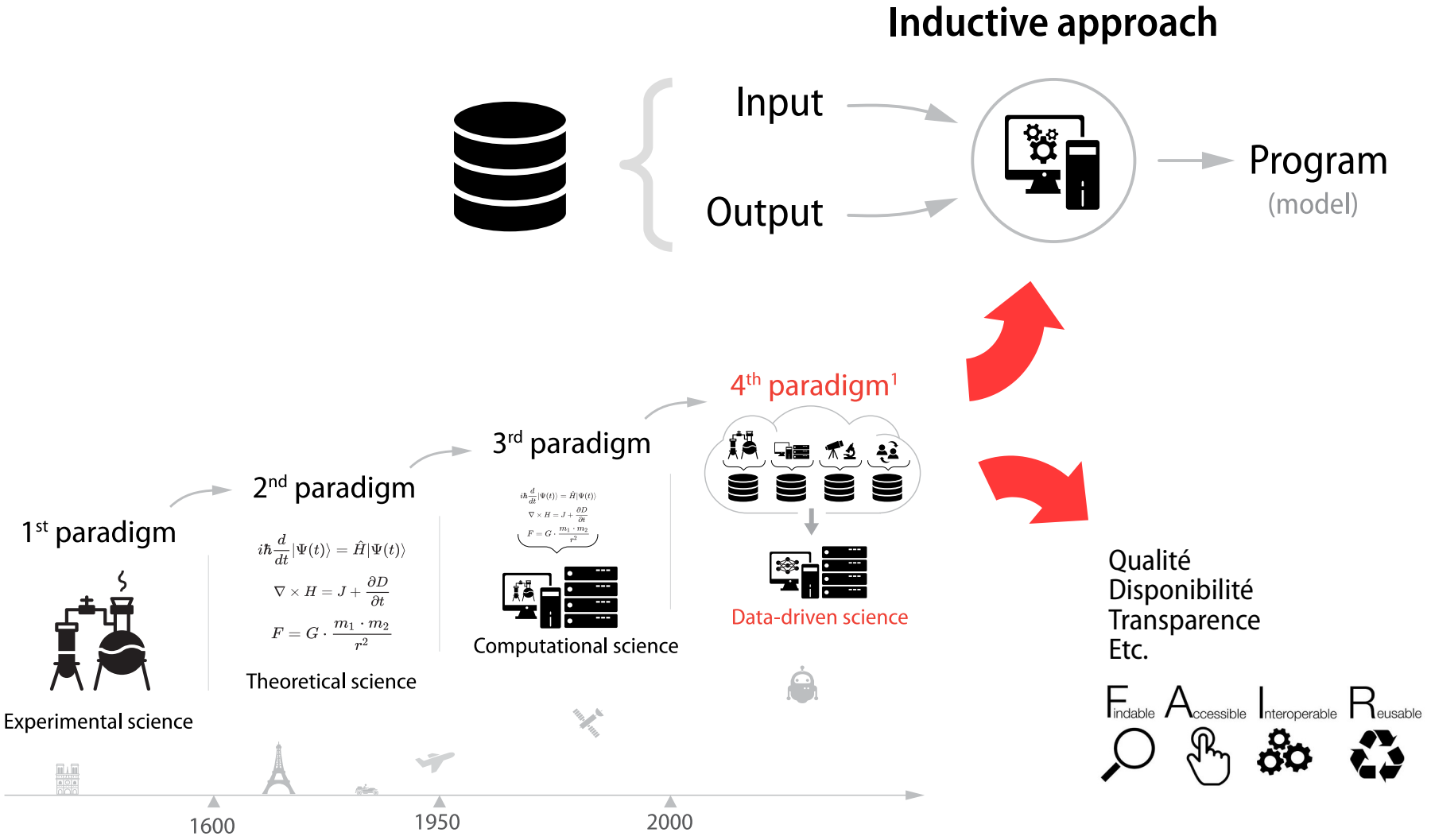
Machine Learning Model Sizes and the Parameter Gap
Pablo Villalobos, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, Anson Ho, Marius Hobbhahn (2022)
https://doi.org/10.48550/arXiv.2207.02852
https://epochai.org/blog/machine-learning-model-sizes-and-the-parameter-gap

Datasets Sizes



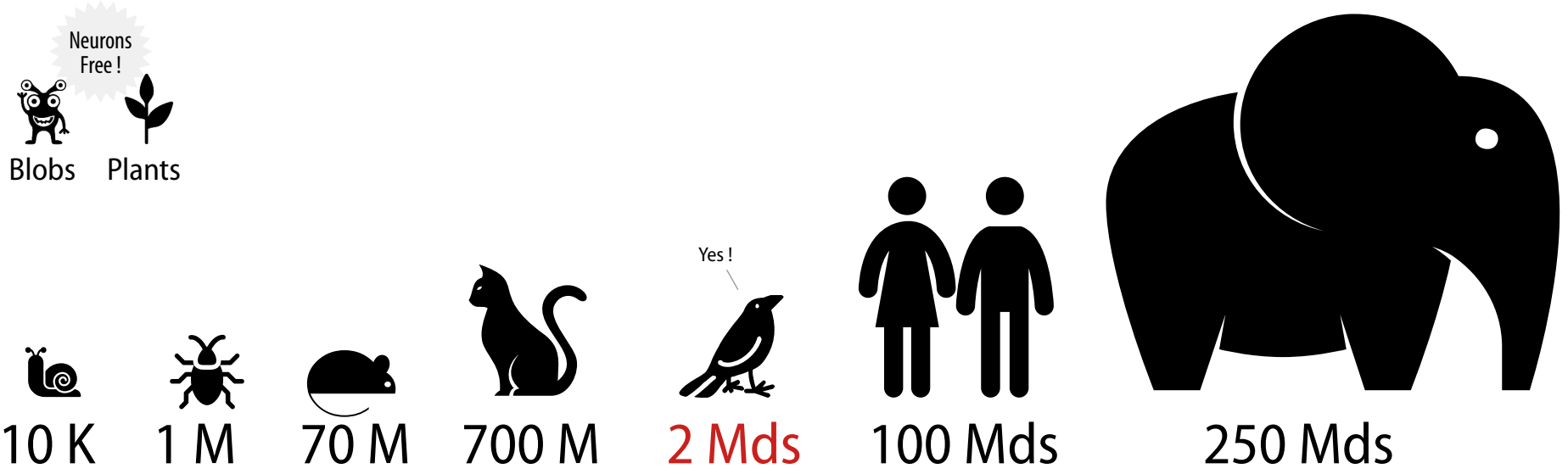
Trends in Training Dataset Sizes
Pablo Villalobos and Anson Ho (2022)
https://epochai.org/blog/trends-in-training-dataset-sizes

In one image


Inductive approach
Input
Output
Program
(model)
4th paradigm[1]
3rd paradigm
2nd paradigm
1st paradigm
Data-driven science
Computational science
Theoretical science
Experimental science
Qualité
Disponibilité
Transparence
Etc.
Findable Accessible Interoperable Reusable
1600
1950
2000

Brain size ?

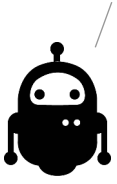
Some brain sizes, in number of neurons...



Short Disgression !

Cette fois, c'est bien fini ;-)

Merci beaucoup !



Références

[JGRAY] Gray, J. (2001), from « The Fourth Paradigm: Data-Intensive Scientific Discovery » Tony Hey, Stewart Tansley, Kristin Tolle (2009). Published by Microsoft Research.
ISBN: 978-0-9825442-0-4

[FROS] Rosenblatt, Frank. (1958). « The perceptron: A probabilistic model for information storage and organization in the brain. » Psychological Review, 65(6), 386-408.

[MIPA] Minsky, Marvin; Papert, Seymour. (1969). « Perceptrons : An Introduction to Computational Geometry », MIT Press

[DRUM] Rumelhart, David E.; Hinton, Geoffrey E.; Williams, Ronald J. (1986). « Learning representations by back-propagating errors ». Nature. 323 (6088): 533–536. doi:10.1038/323533a0.

[YLEC1] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel, « Backpropagation Applied to Handwritten Zip Code Recognition », AT&T Bell Laboratories

[LRDN] Dominique Cardon, Jean-Philippe Cointet, Antoine Mazieres. (2018). « La revanche des neurones », Réseaux, La Découverte, 5 (211), <10.3917/res.211.0173>. <hal-01925644>

[TOP500] Statistics on top 500 high-performance computers. (2018) « Exponential growth of supercomputing power as recorded by the TOP500 list ». https://www.top500.org

[WKP1] Wikipedia/en. (2018) « List of datasets for machine-learning research ». https://en.wikipedia.org

[WOS1] Core database : TS=("support vector machine*" OR ("SVM" AND "classification") OR ("SVM" AND "regression") OR ("SVM" AND "classifier") OR "support vector network*" OR ("SVM" AND "kernel trick*"))

[WOS2] Core database : TS=("deep learning" OR "deep neural network*" OR ("DNN" AND "neural network*") OR "convolutional neural network*" OR ("CNN" AND "neural network*") OR "recurrent neural network*" OR ("LSTM" AND "neural network*") OR ("RNN*" AND "neural network*"))

Illustrations :

Potato plant From *Die Giftpflanzen Deutschlands,* Peter Esser, 1910, via iconspng.com

Neuron Wikimedia Commons, the free media repository.

Icons thenounproject.com