

# Introduction aux réseaux neuronaux

Olivier Ricou

2018

# Cas d'utilisations



Pub. ciblée



Recommendations



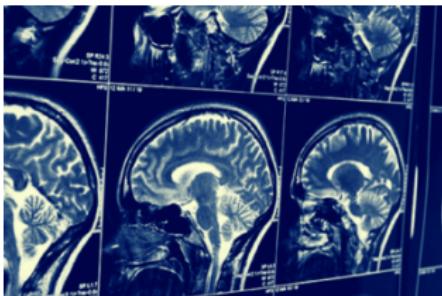
Description



Jeux



Sécurité



Diagnostique



Majordome

# Historique

Three AI waves... and two AI winters



# Les hivers passés

## Winters explained

1



Only able to capture explicit knowledge

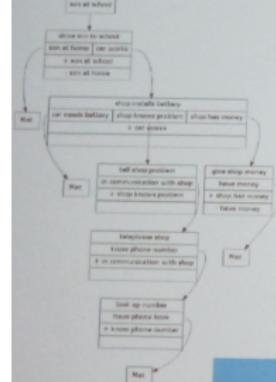
2



Lack of data

Lack of robustness

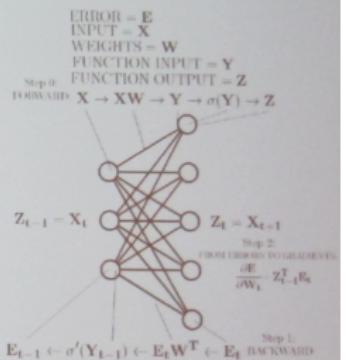
Computational burden



1959

Lack of interpretability

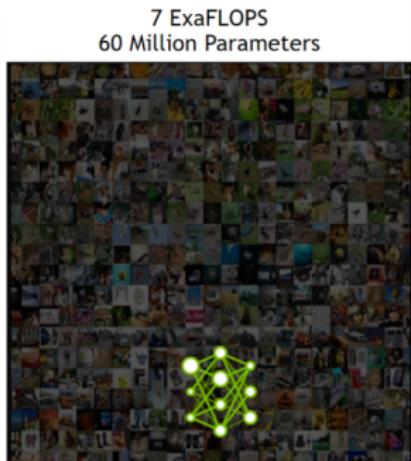
Hard to maintain



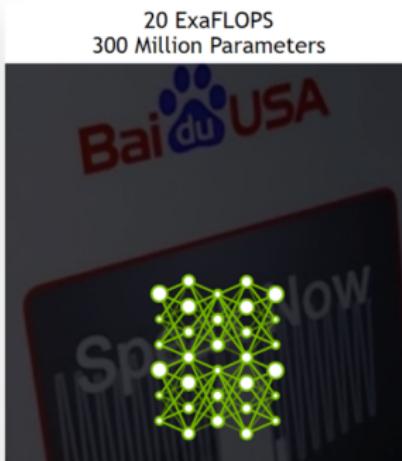
1986

La renaissance est due au triptique **données, hardware, théorie**.

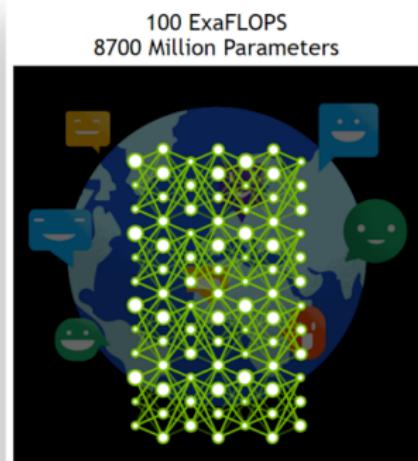
# Plus gros c'est mieux



2015 - Microsoft ResNet  
Superhuman Image Recognition



2016 - Baidu Deep Speech 2  
Superhuman Voice Recognition



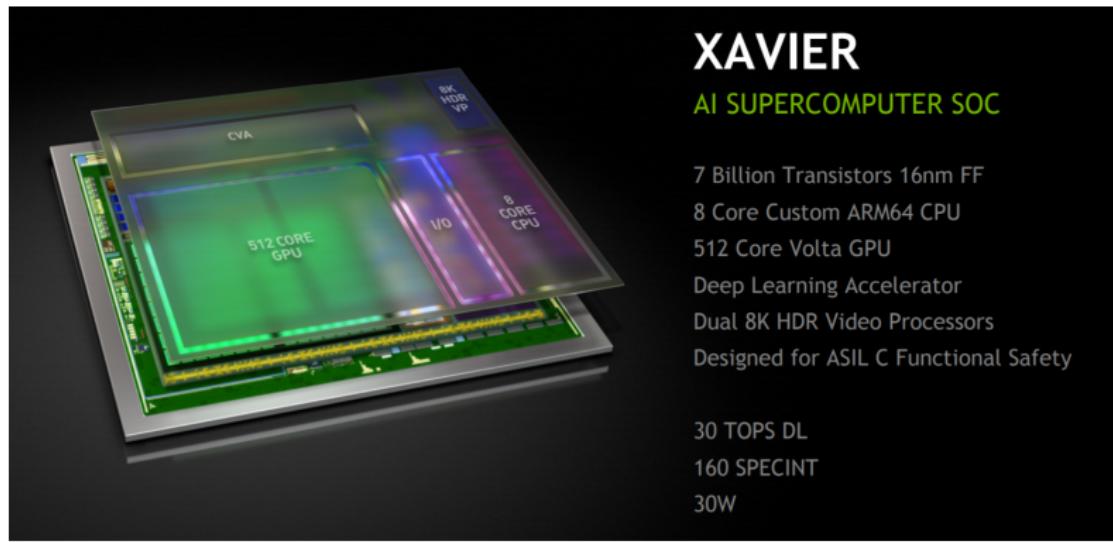
2017 - Google Neural Machine Translation  
Near Human Language Translation

Besoin de calculs pour l'entraînement et taille des réseaux

# Exemple d'une voiture autonome

ResNet-50 à besoin de 7,72 G opérations pour traiter une image 255x255.

- 230 Gops pour 30 fps
- 9,4 Tops pour du HD
- 338 Tops pour 12 caméras et 3 réseaux par caméra



# De la puissance pour suivre le rythme



## CONVERGENCE HPC - IA

A venir à IDRIS : 1<sup>ère</sup> machine convergée en 2019

Contexte

- Fin Mars 2018 : Parution du rapport Villani et conférence « *AI For Humanity* » à Paris
- Demande du MESRI à GENCI d'intégrer composante dédiée recherche IA dans AO en cours IDRIS

Approche

- Groupe de travail mixte experts HPC et IA → besoin, mode d'accès, ...
- Définition nœud convergé
  - Nœud de calcul hybride capable tourner travaux HPC et IA
  - Intérêt GPU type nVIDIA V100 (HBM2, nVLINK, tensorcores, ...) 16 ou 32 Go
  - Au moins 4 GPU par nœud et 192 Go/nœud, nVLINK 2.0 inter GPU, ...
- Piles logicielles IA containerisées
- Unité allocation minimale *scheduler* = 1 GPU
  - 1<sup>ère</sup> étape = throughput mono GPU
  - Puis multi GPU et multi nœud visé
- Stockage global full flash 1 Po > 300 Go/s, réflexion sur pool NVMe over fabric
- Quelques nœuds larges hybrides additionnels



## Les leaders les plus visibles sont

- Google (Tensorflow, Keras, DeepMind)
- Facebook (Torch, PyTorch)
- Microsoft (CNTK)
- IBM (Watson)
- Baidu

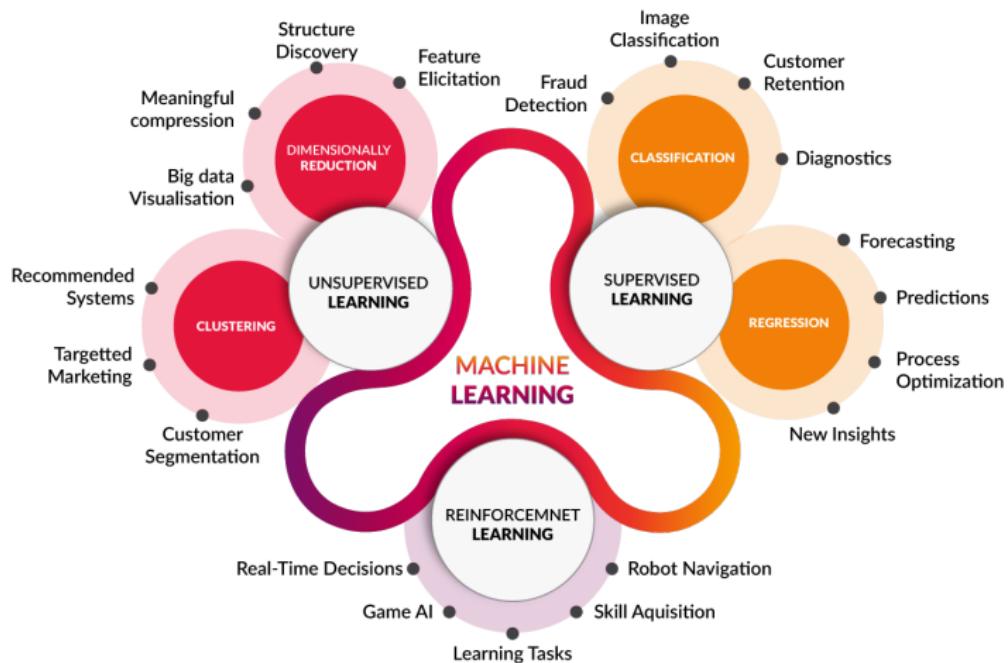
et bien sûr le principal fabriquant : NVidia (Cuda, CuDNN)

## Ceux qu'on voit moins

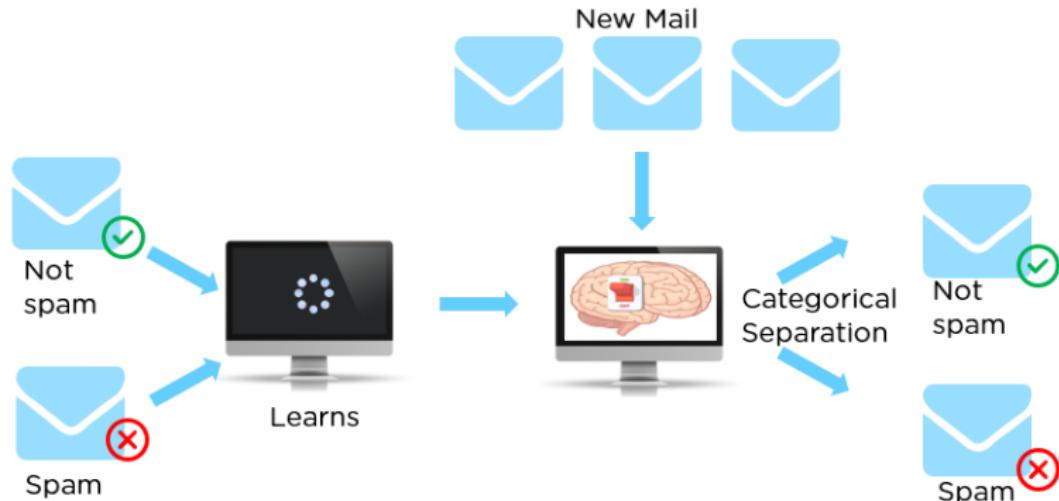
A coté de ceux qui participent activement à la recherche et au développement des outils, il y a ceux qui l'utilisent en interne.

- Amazon (Alexa, Amazon Go)
- Apple
- les constructeurs automobiles (Tesla, Uber, tous)
- tout ceux qui font du conseil (Netflix, Expedia...), de la pub (Critéo)
- plein de startups

# Types d'apprentissage



# L'apprentissage supervisé



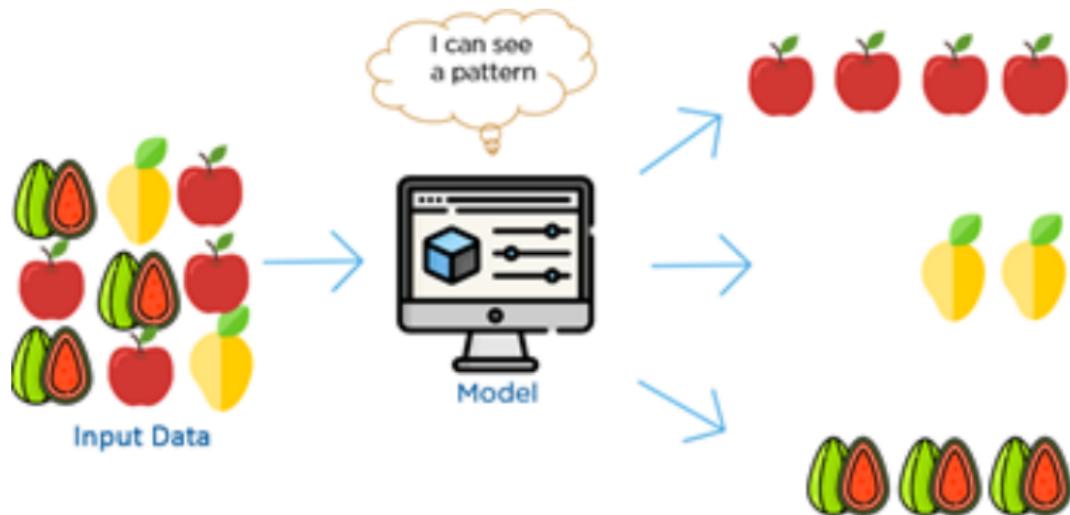
# L'apprentissage supervisé

Régression	Classification
Moindres carrés	SVM
Régression polynomiale	Regression logistique
Réseau neuronal	Arbre de décision Réseau neuronal

La révolution vient des réseaux neuronaux :

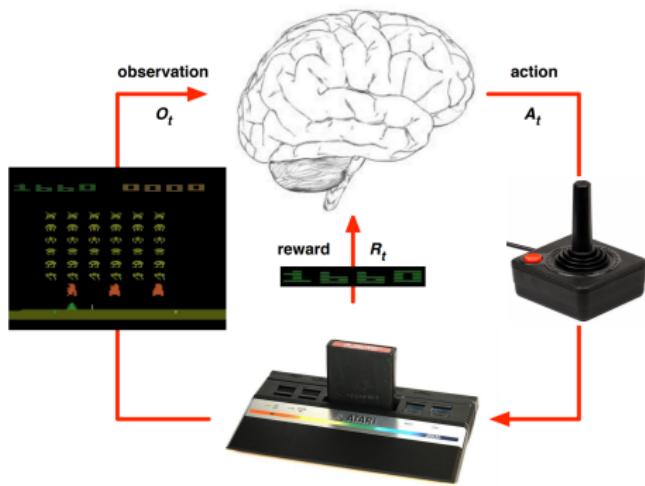
- Mûr
- Demande des quantités énormes de données étiquetées
- Pas toujours simple à faire marcher
- De plus en plus complexe
- Produit des résultats remarquables en
  - ▶ traitement d'image
  - ▶ traitement de la parole

# L'apprentissage non supervisé



- K-moyennes, ACP, des réseaux neuronaux
- Difficile d'en mesurer l'efficacité (besoin de juges humains)
- Usage limité

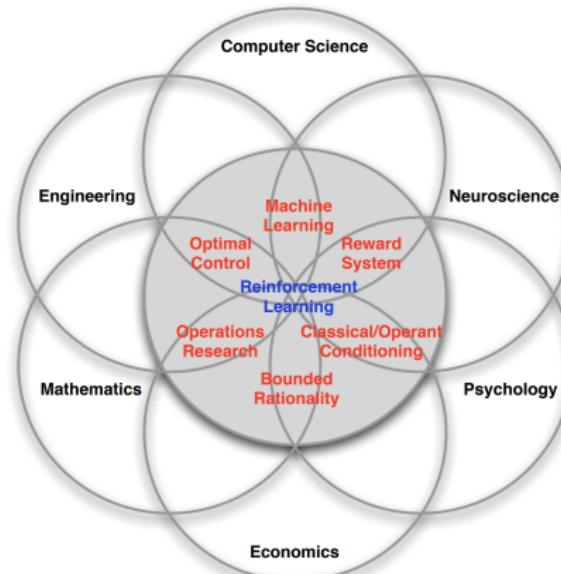
# L'apprentissage par renforcement



- Rules of the game are unknown
- Learn directly from interactive game-play
- Pick actions on joystick, see pixels and scores

# Points clefs du renforcement

- Pas de superviseur qui connaît la solution, seulement une note
- Le retour d'information est décalé (pas immédiat)
- La notion de temps est importante → Système dynamique
- L'agent qui note a un impact sur la suite des données qu'on va recevoir



# Test

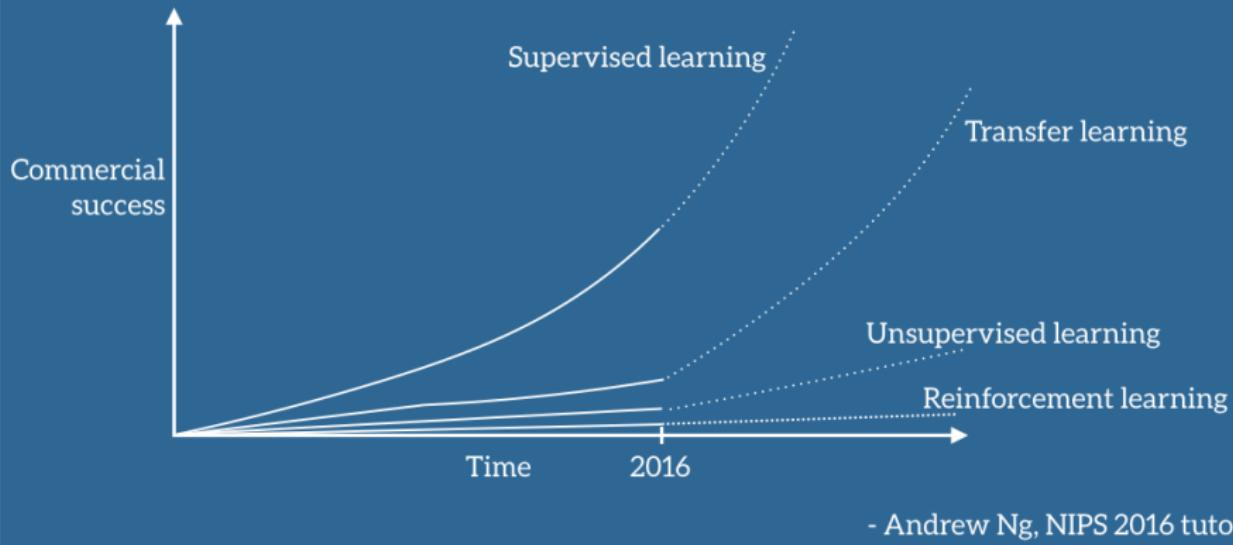
Quel type d'apprentissage ?

- Comparaison de CNN pour la vision sur route – 2018
- DeepMind StarCraft II combat et explications – 2019
- Appel au téléphone - Google – 2018
- Helicopter - Stanford Univ. – 2008
- Mélodie travaillée - Music VAE – 2018
- Robot Atlas - Boston Dynamics – 2017
- Débat : L'État doit-il financer les écoles *pre-maternelle* ? (3 à 4 ans)  
Non – Harish Natarajan    Oui – IBM Debater – 2019

VLC runs the video, F fullscreen, Crtl-Q quit

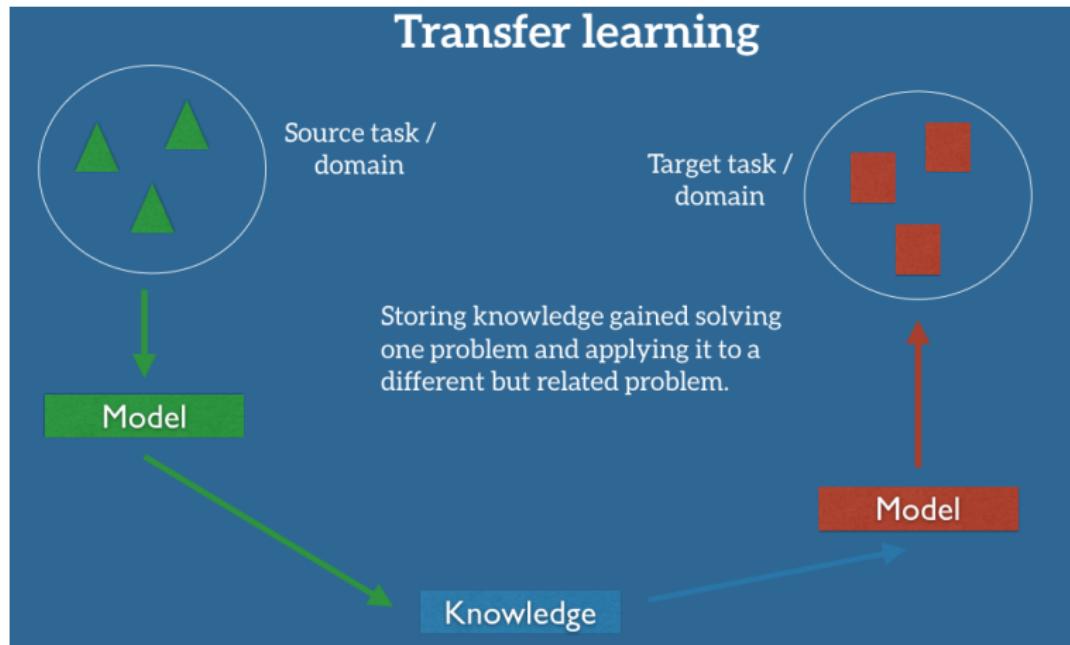
# Usage futur des différents types d'apprentissage

## Drivers of ML success in industry



Le monde académique/internet et industriel sont différents.

# Transfert ML



Ainsi il est tout à fait possible d'utiliser un réseau neuronal entraîné pour une tâche A pour initier l'entraînement du réseau d'une tâche B proche.

# IBM IA pour l'industrie

- IBM Watson Recruitement une aide à l'embauche pour les entreprises
- Watson solution pour la vente
- Watson Assistant pour le marketing
- Watson Decision Plateform pour l'agriculture
- IBM Equipment Maintenance Assistant pour améliorer la qualité et réduire la maintenance
- IBM Watson Supply Chain Insights

<https://www.ibm.com/watson/ai-for-industries/>