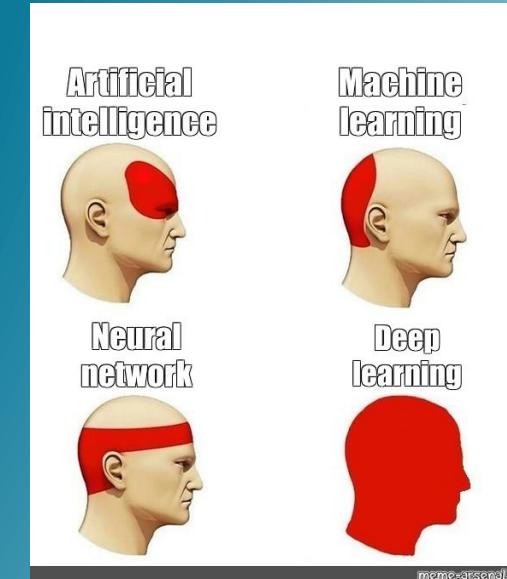


# Méthodes Numériques et Modélisation

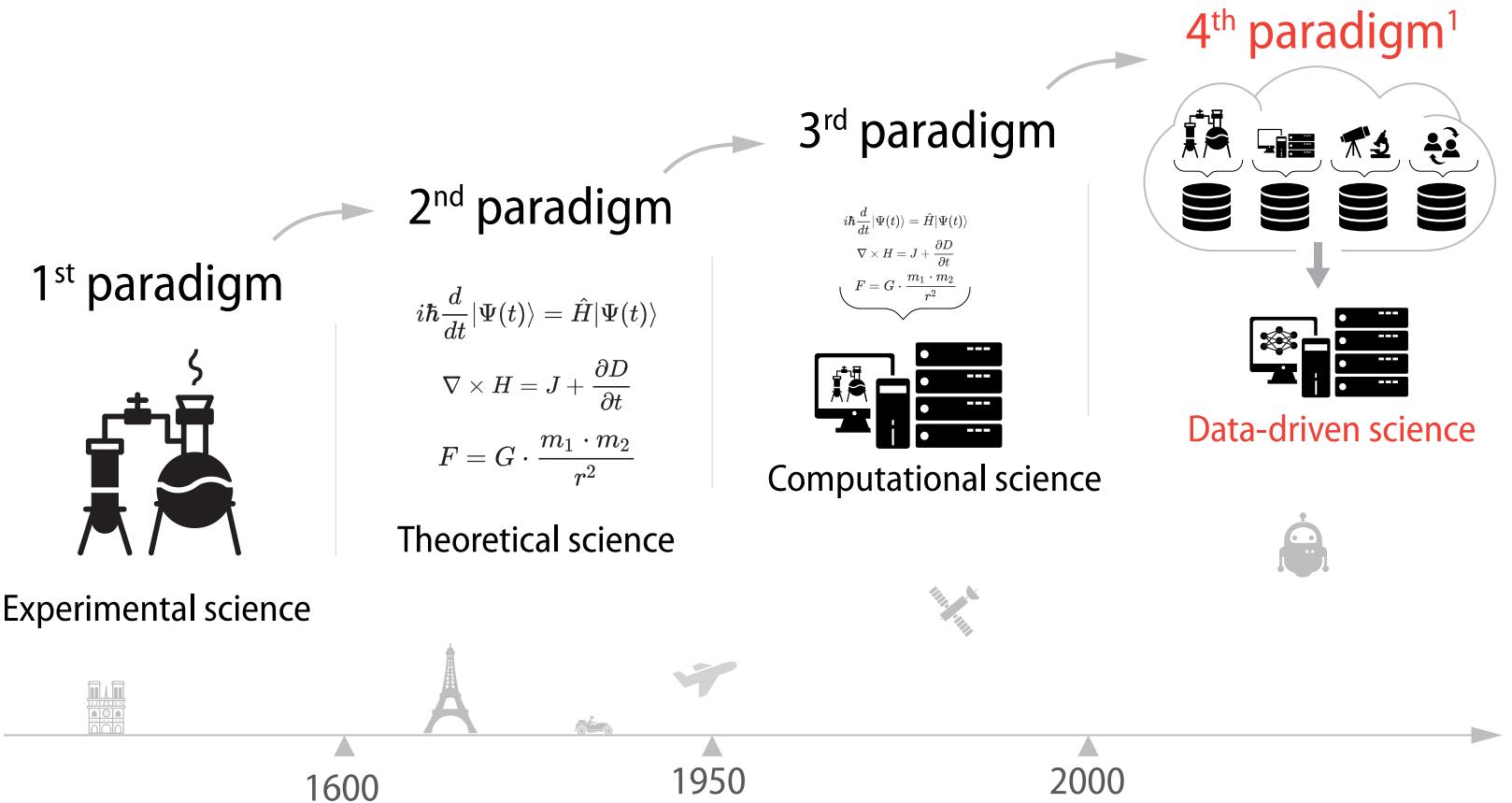
Fondements de l'intelligence artificielle et  
applications aux sciences atmosphériques



MINERVA



# INTRODUCTION – LA RÉVOLUTION IA

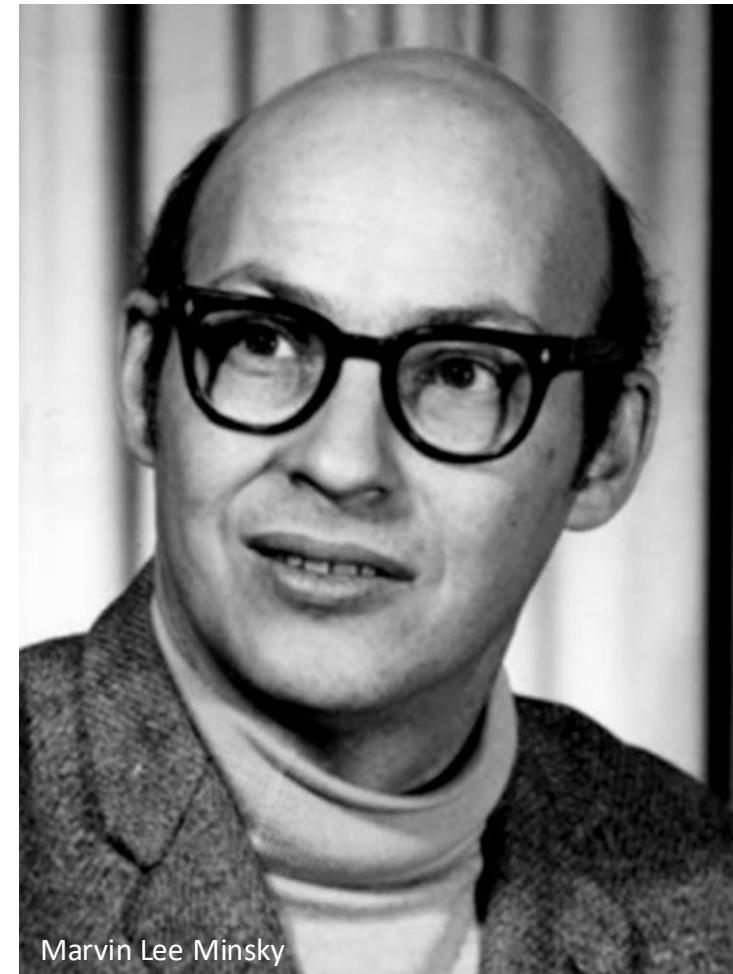


<sup>1</sup> Jim Gray, 2007

# L'INTELLIGENCE ARTIFICIELLE

- construction de programmes informatiques capables de prendre en charge des tâches habituellement effectuées par des humains
- L'objectif est de parvenir à transmettre à une machine des fonctions propres au vivant : rationalité, raisonnement, mémoire et perception.

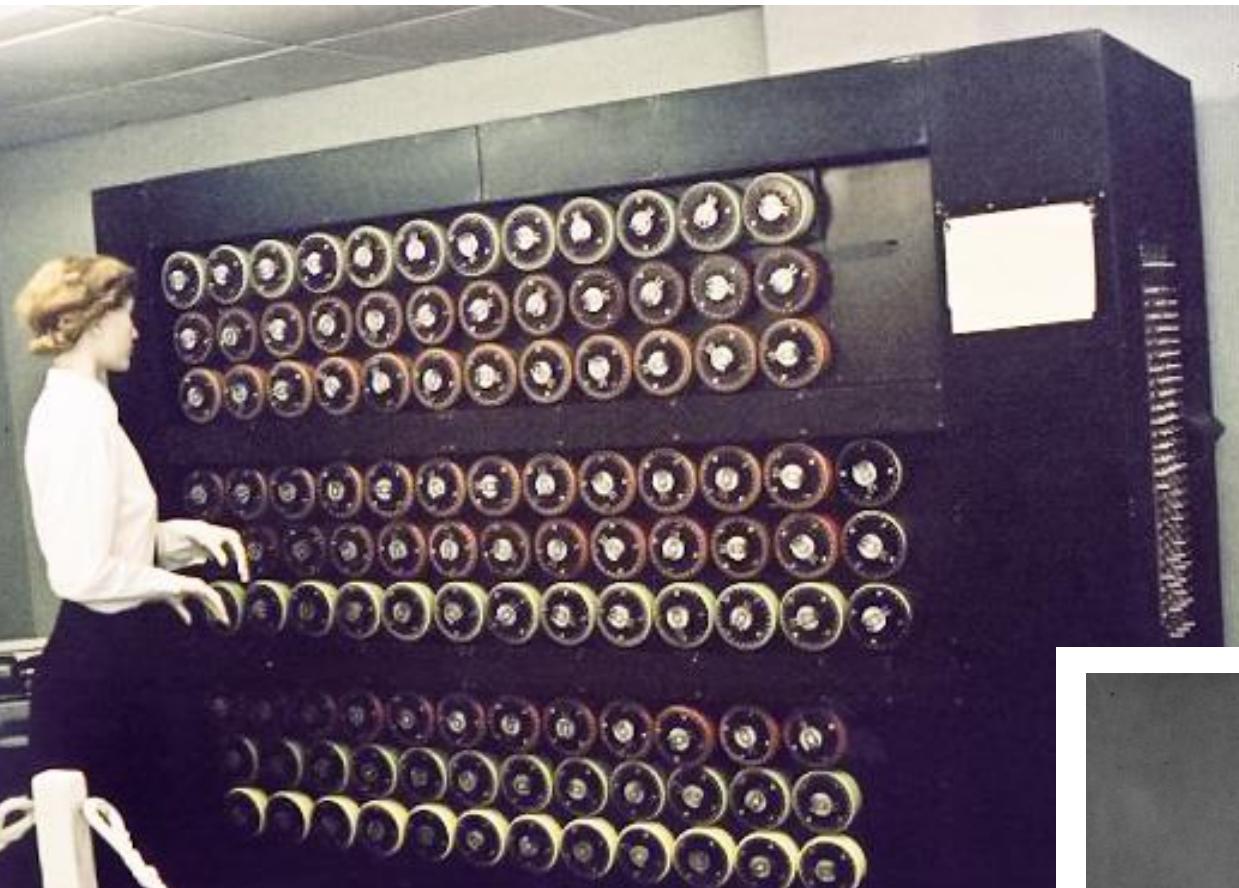
Raisonner / Comprendre / Interagir



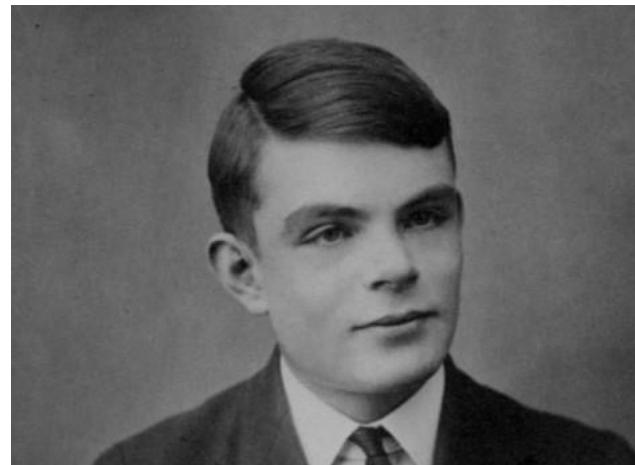
Marvin Lee Minsky

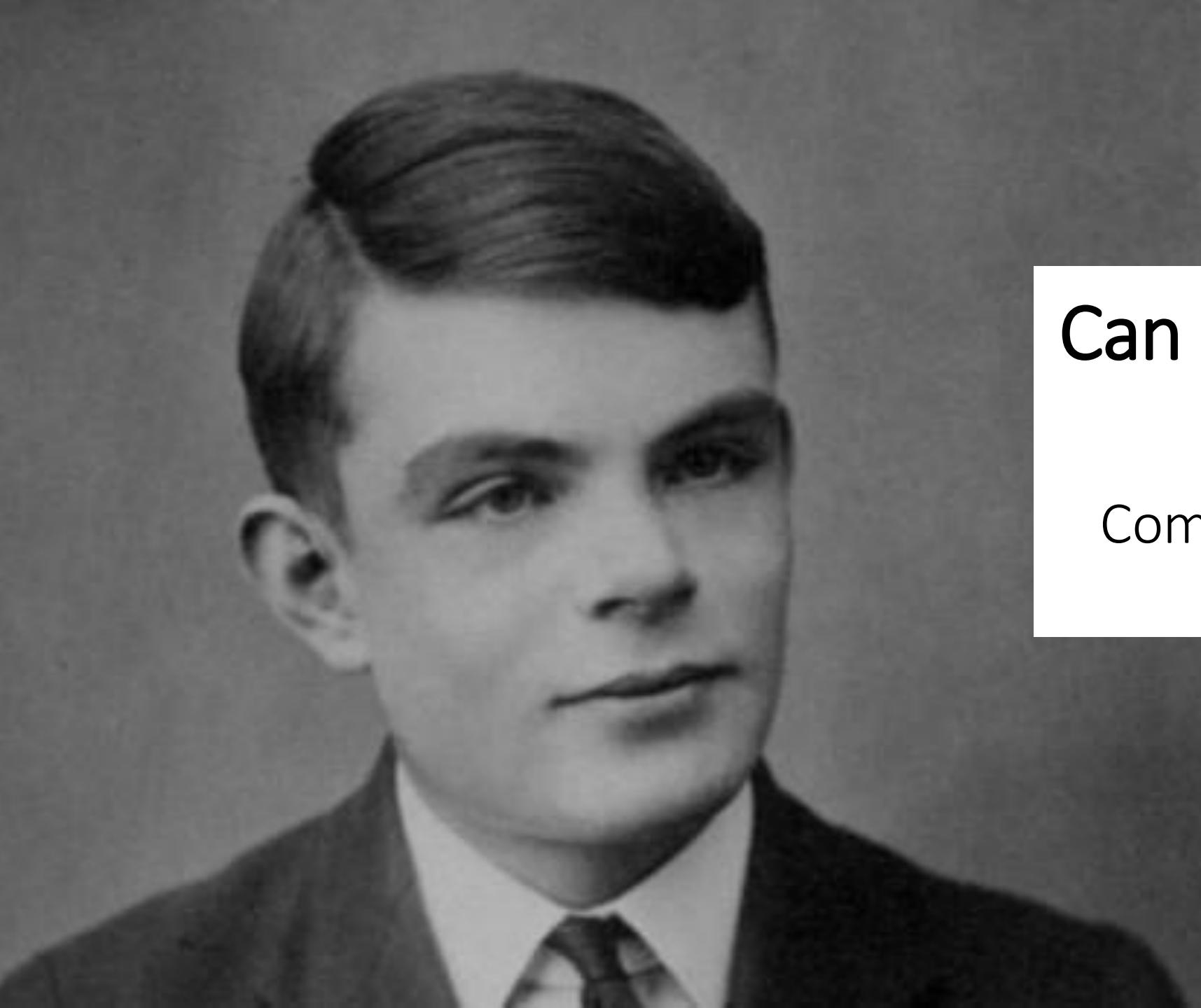
Un ordinateur peut-il  
devenir intelligent ?

# THE BOMB



- Alan Turing
- Enigma
- The bomb





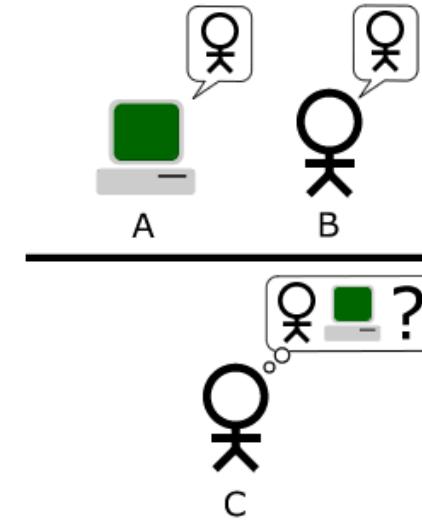
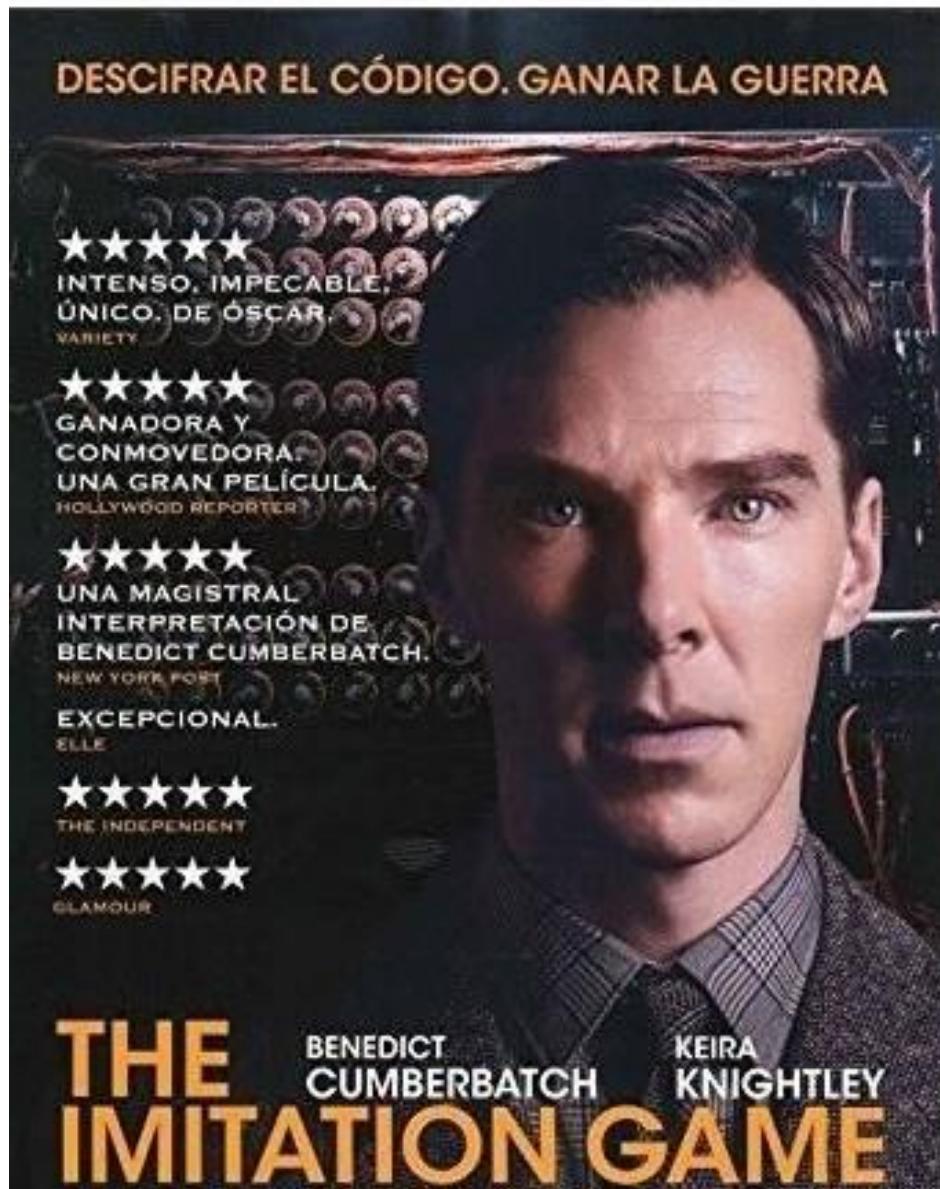
# Can machines think ?

A. Turing

Computing machinery and  
intelligence, 1950

Mind (en), Oxford University Press,  
vol. 59, no 236, octobre 1950 (DOI  
[10.1093/mind/LIX.236.433](https://doi.org/10.1093/mind/LIX.236.433))

# The imitation game

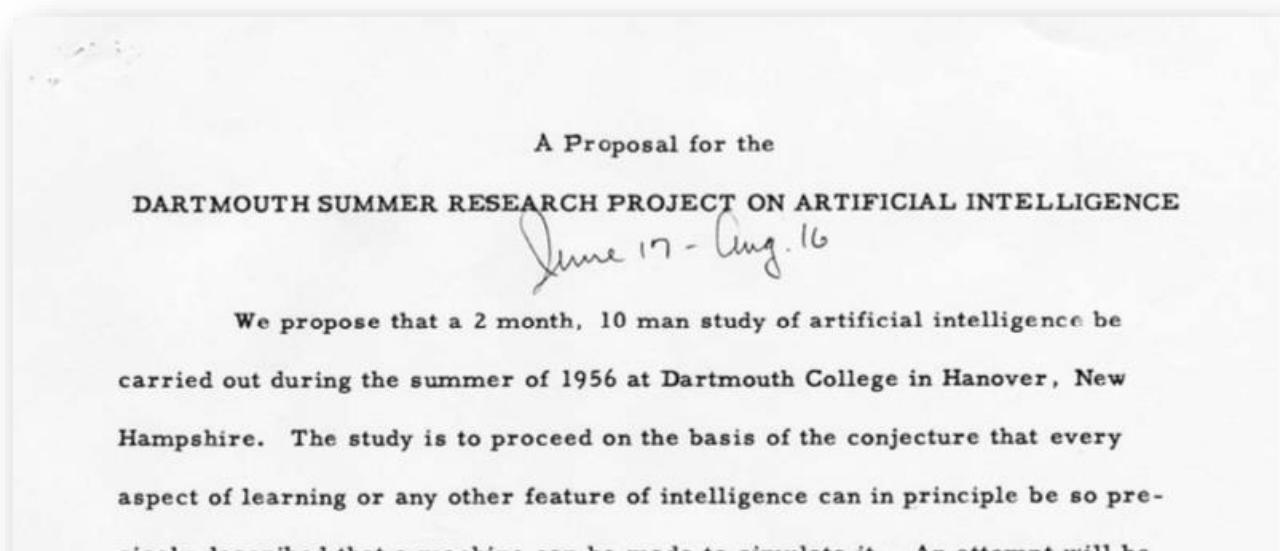


- Test de Turing
- « Un ordinateur peut-il penser »
- Intelligence multiple
- Faire semblant de faire des erreurs
- Machine expérimentale (pas de preuve ou d'algo)
- Ordinateur digital : disponibilité, **puissance**
- Ordinateur à variable discrète : mémoire ?

# Historique de l'IA

- 1956 : L'intelligence artificielle devient un véritable domaine scientifique
  - Dartmouth Summer Research Project on Artificial Intelligence
  - Organisée par John McCarthy

“ find how to make machines [...] solve kinds of problems now reserved for humans ”



Nathaniel Rochester

Marvin L. Minsky

John McCarthy

Oliver G. Selfridge Ray Solomonoff

Trenchard More

Claude E. Shannon

# Historique de l'IA

- 1956 : L'intelligence artificielle devient un véritable domaine scientifique
- 60' : financements, laboratoires USA & UK
  - « des machines seront capables, d'ici 20 ans, de faire le travail que toute personne peut faire »



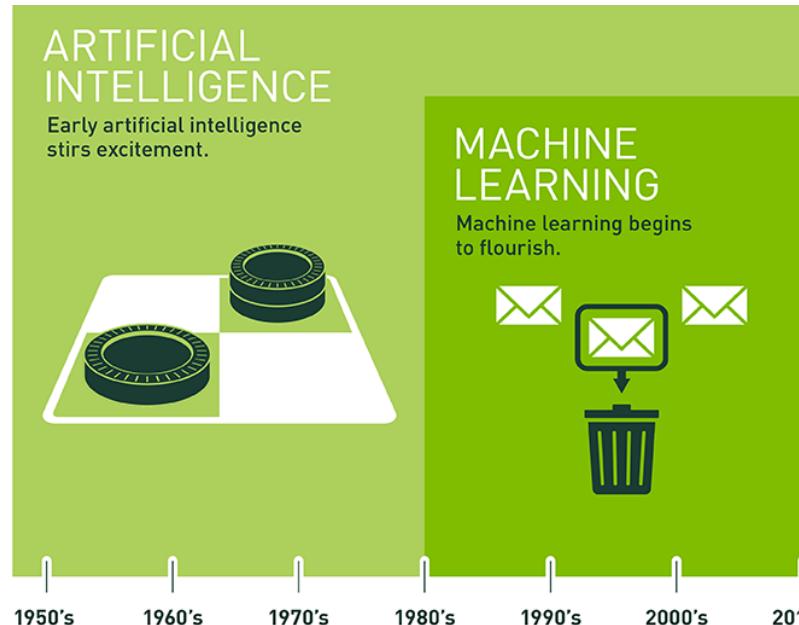
# Historique de l'IA

- 1956 : L'intelligence artificielle devient un véritable domaine scientifique
- 60' : financements, laboratoires USA & UK
  - « des machines seront capables, d'ici 20 ans, de faire le travail que toute personne peut faire »
- 1974 : AI Winter : les projets n'aboutissent pas,  
les ordinateurs ne sont pas assez puissants,  
les financements sont réduits



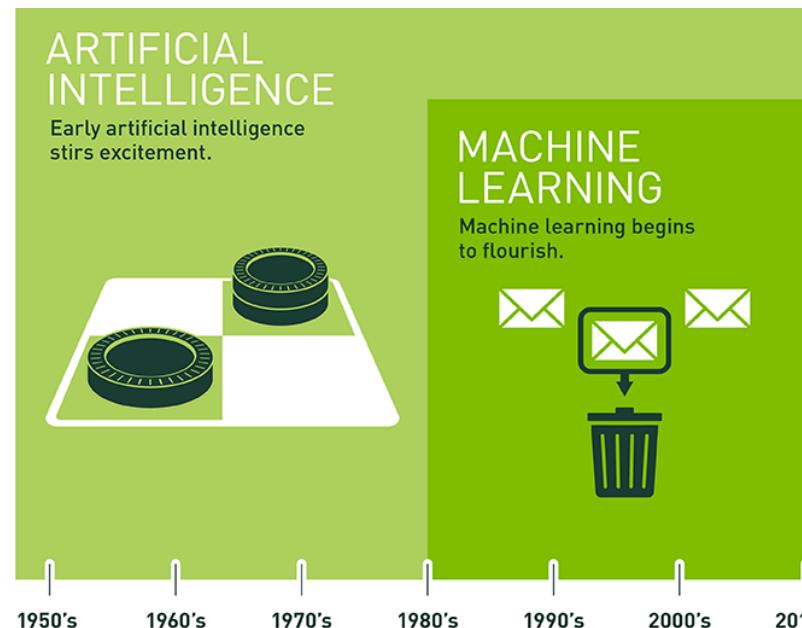
# Historique de l'IA

- 1956 : L'intelligence artificielle devient un véritable domaine scientifique
- 60' : financements, laboratoires USA & UK
  - « des machines seront capables, d'ici 20 ans, de faire le travail que toute personne peut faire »
- 1974 : AI Winter : les projets n'aboutissent pas, les ordinateurs ne sont pas assez puissants, les financements sont réduits
- 80' : Systèmes experts (une seule tâche)



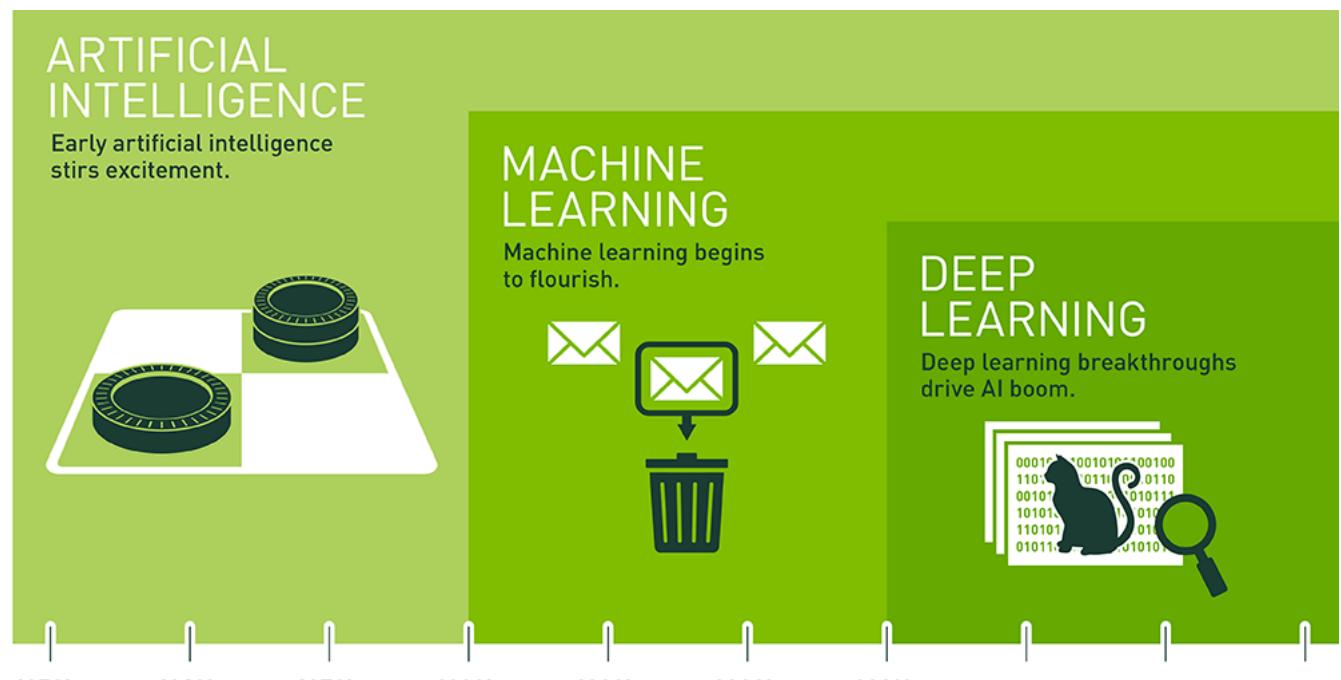
# Historique de l'IA

- 1956 : L'intelligence artificielle devient un véritable domaine scientifique
- 60' : financements, laboratoires USA & UK
  - « des machines seront capables, d'ici 20 ans, de faire le travail que toute personne peut faire »
- 1974 : AI Winter : les projets n'aboutissent pas, les ordinateurs ne sont pas assez puissants, les financements sont réduits
- 80' : Systèmes experts (une seule tâche)
- 90' : Ordinateurs puissants, nouveaux domaines

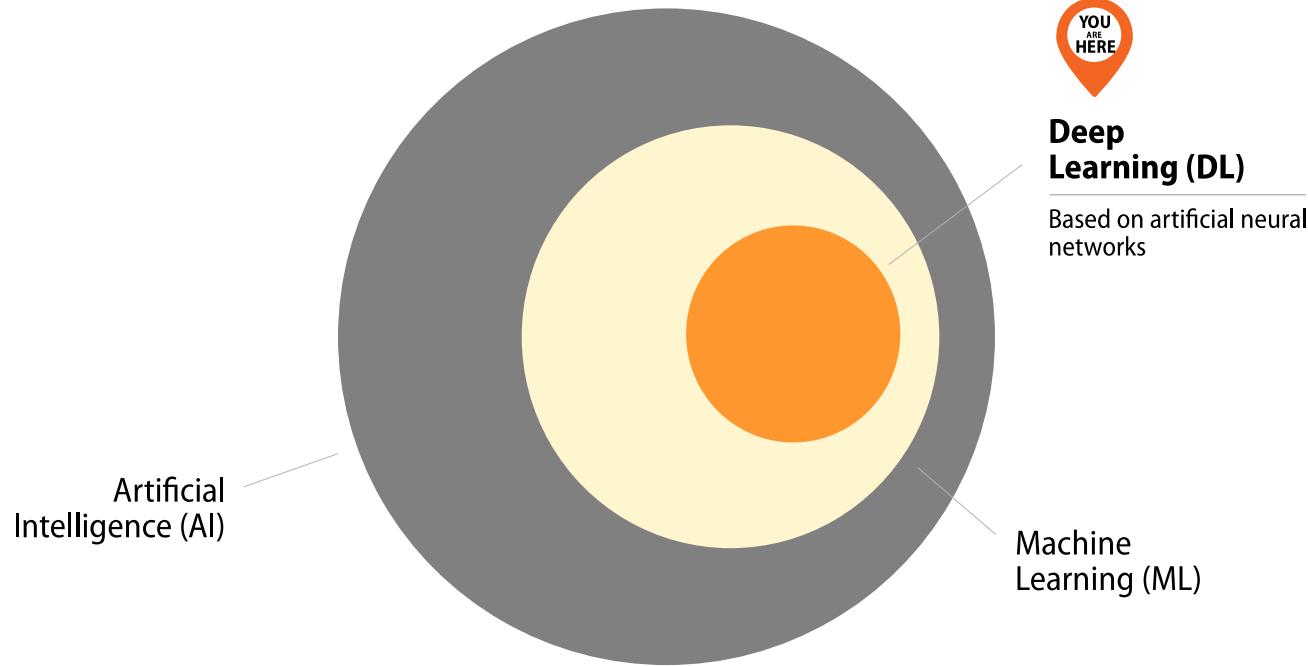


# Historique de l'IA

- 2000' : l'avènement des données
  - Les nouvelles algorithmes
  - Les puissances de calcul phénoménales
  - Apprentissage profond : Deep Learning



# TOUTE IA EST PAREILLE ALORS ?

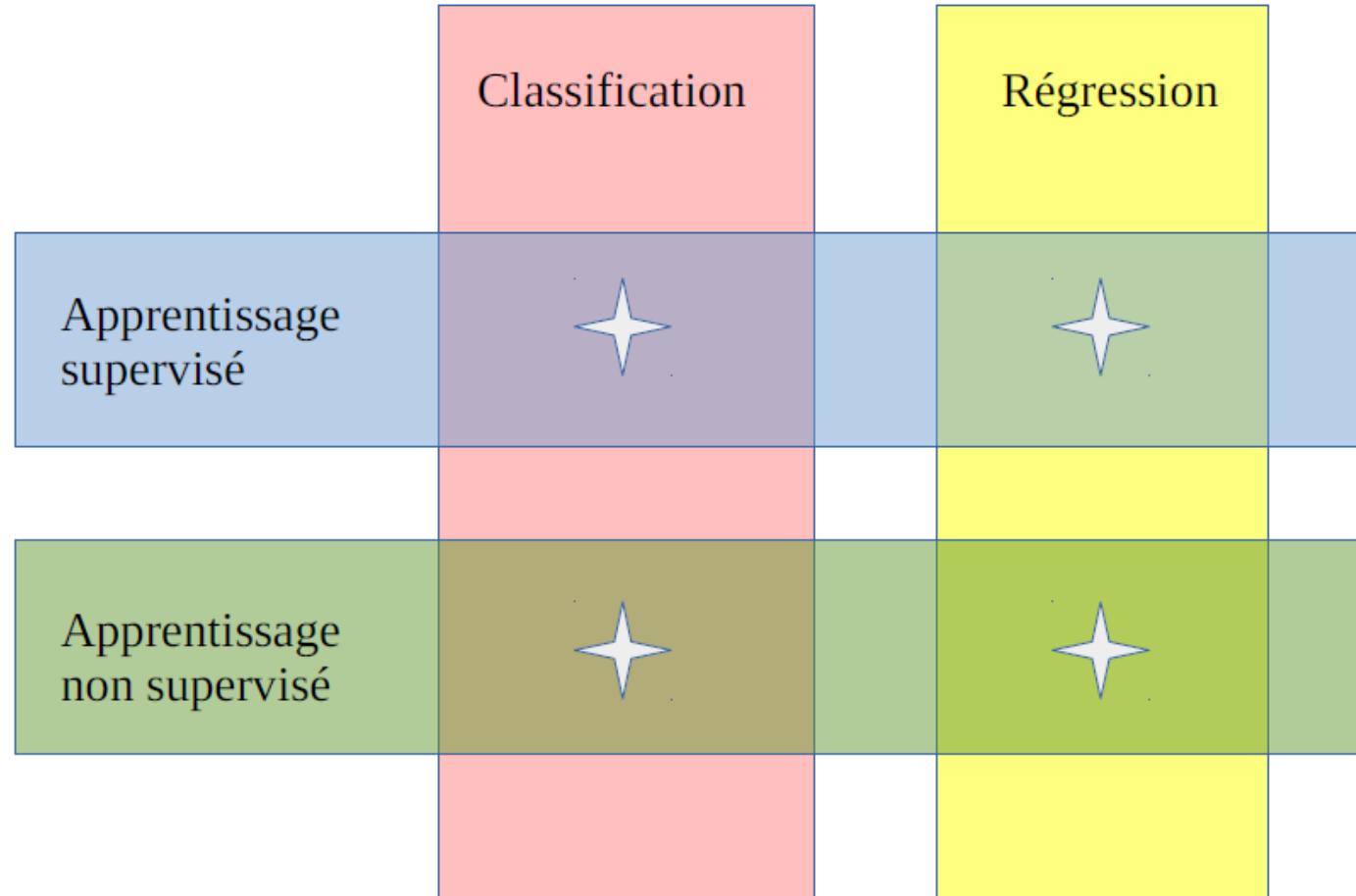


**Intelligence Artificielle** = ensemble des techniques permettant à des machines d'accomplir des tâches et de résoudre des problèmes normalement réservés aux humains

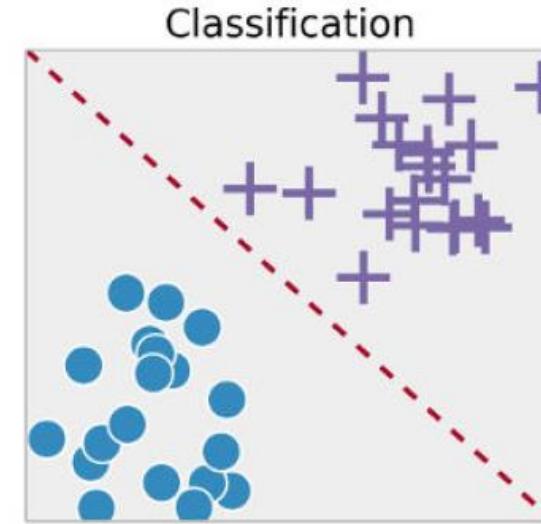
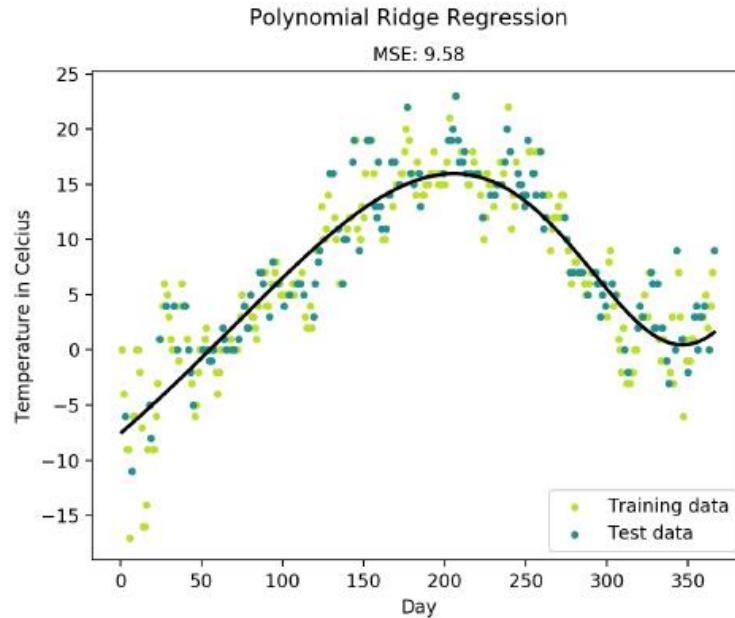
**Machine Learning** = apprentissage automatique  
« champ d'étude de l'intelligence artificielle qui se base sur des approches pour donner aux ordinateurs la capacité d'apprendre à partir de données, c à d d'améliorer leurs performances à résoudre des tâches sans être explicitement programmés pour chacune »

**Deep learning** = Branche du Machine Learning spécialisée dans l'utilisation de réseaux de neurones profonds

# LES GRANDES FAMILLES



# CLASSIFICATION VS REGRESSION

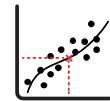


## Régression

Prédire une variable quantitative



Tell me,  
what's the  
price ?



## Classification

Prédire une classe (qualitative, discrète)



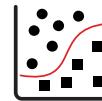
This is a cat



This is a rabbit

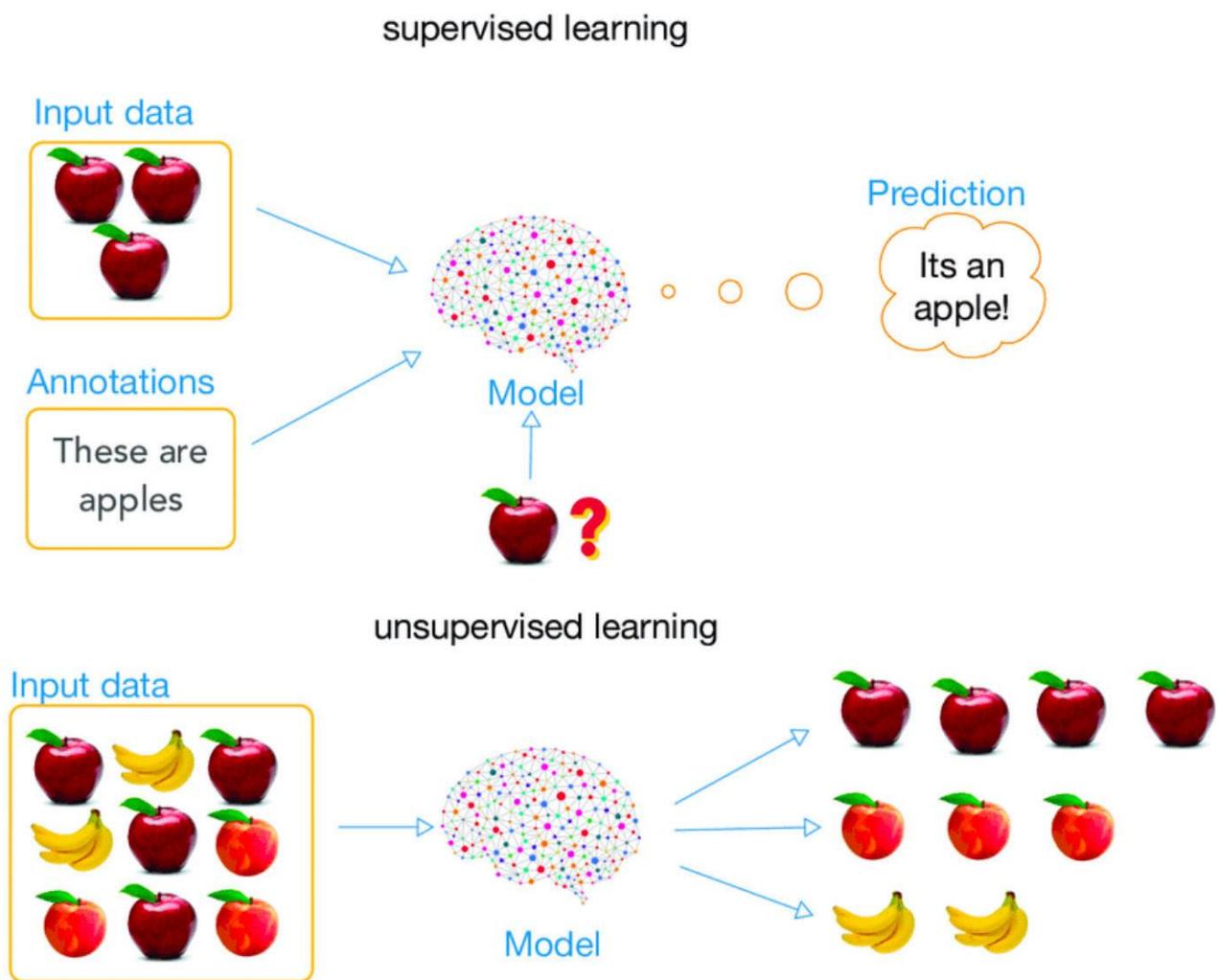


Tell me,  
what is it ?

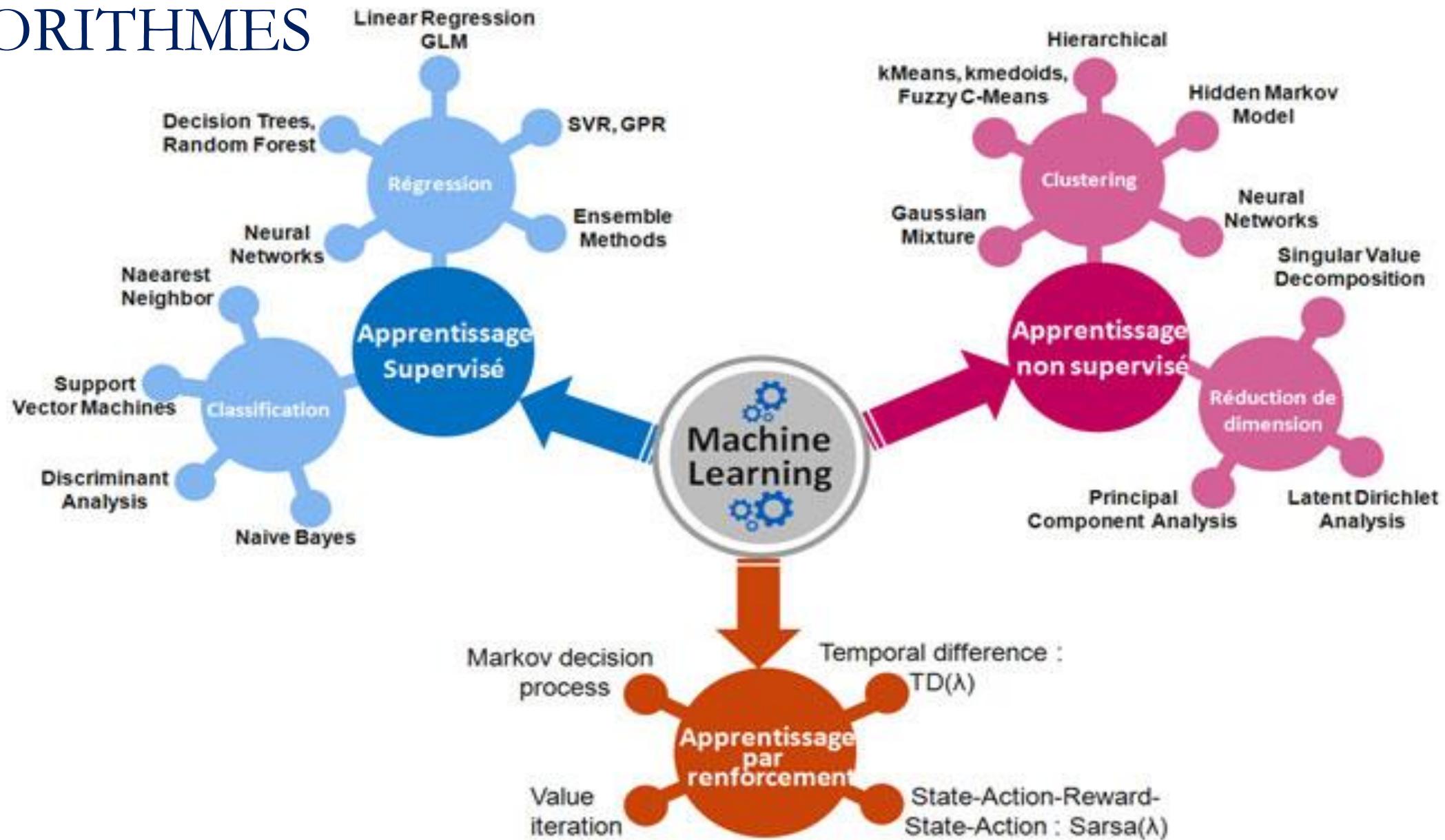


# APPRENTISSAGE SUPERVISÉ VS NON-SUPERVISÉ

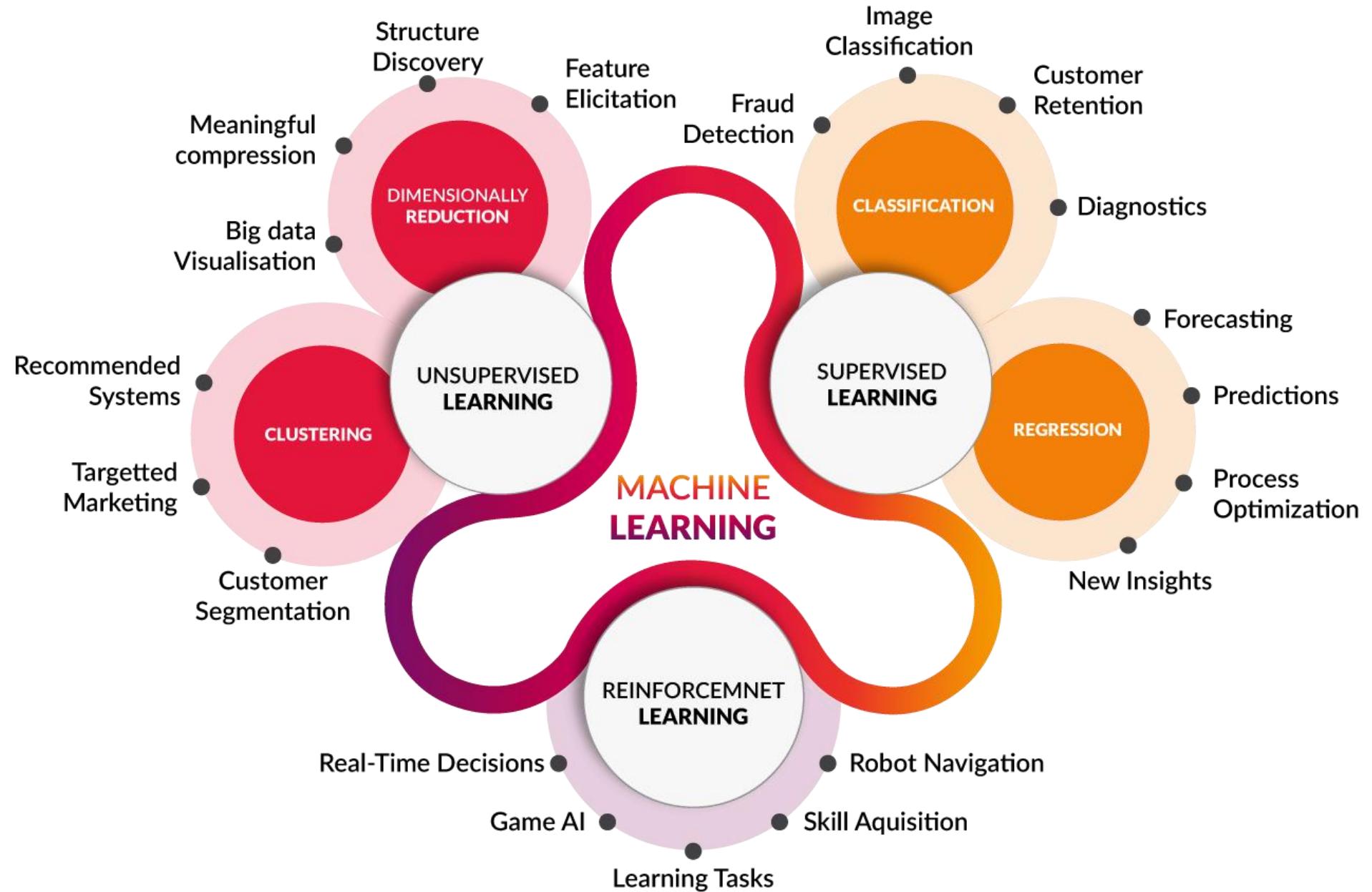
- Apprentissage supervisé :
  - Nécessite un jeu d'entraînement  $X, y$ 
    - $X$  : prédicteurs
    - $y$  : variable à prédire
- Apprentissage non supervisé :
  - Nécessite un jeu d'entraînement  $X$
  - L'algorithme "décide" quelles sont les caractéristiques importantes

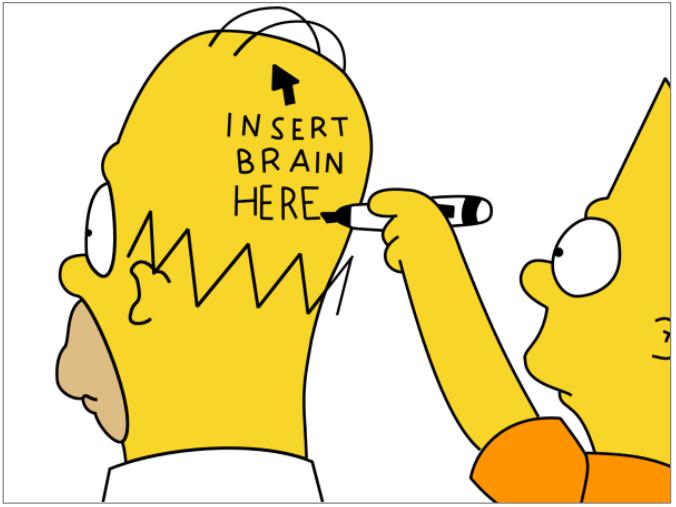


# ALGORITHMES



# USAGES



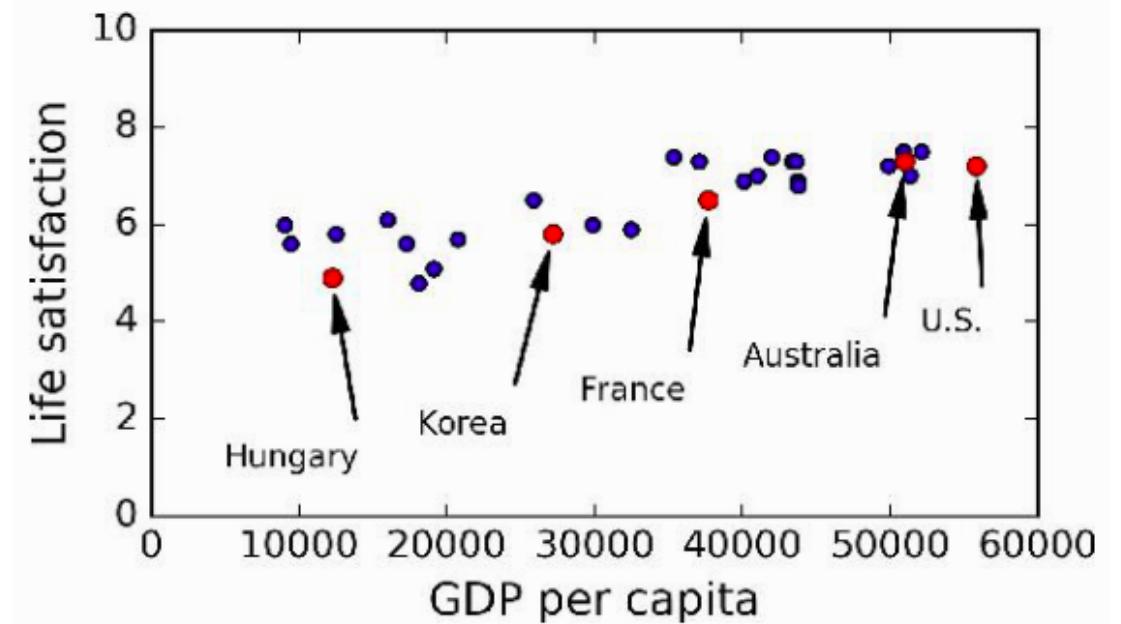


# Les principaux algorithmes

# EXEMPLE : EST-CE QUE L'ARGENT REND HEUREUX ?

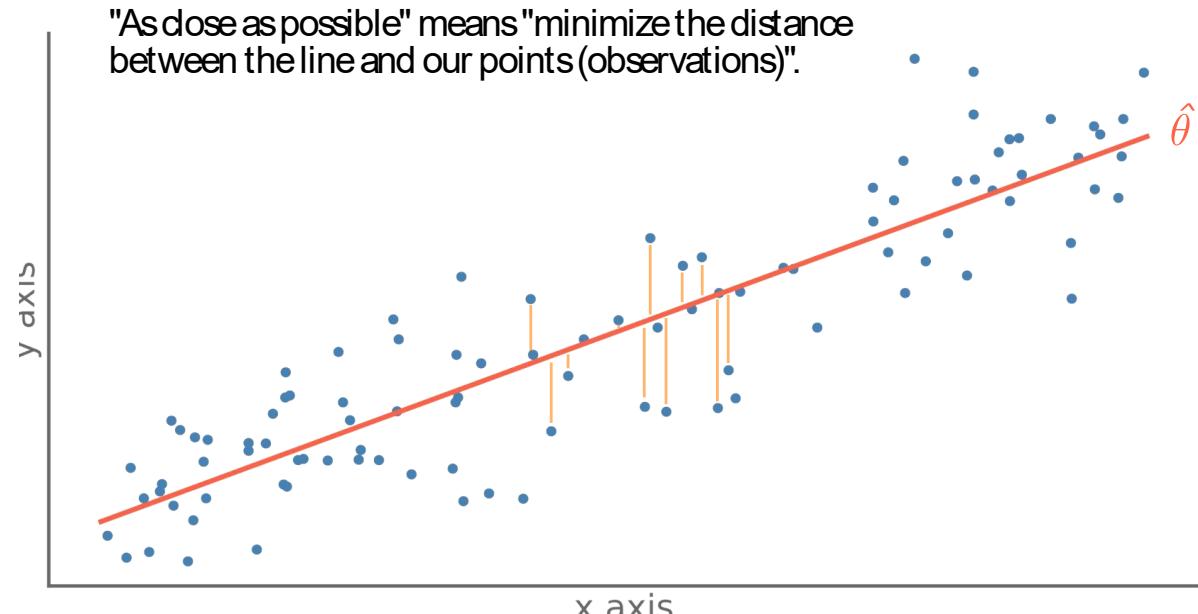
- Idée : croiser la rente per capita des pays et leur index de satisfaction
- Est-ce qu'on observe une tendance ?

Country	GDP per capita (USD)	Life satisfaction
Hungary	12,240	4.9
Korea	27,195	5.8
France	37,675	6.5
Australia	50,962	7.3
United States	55,805	7.2



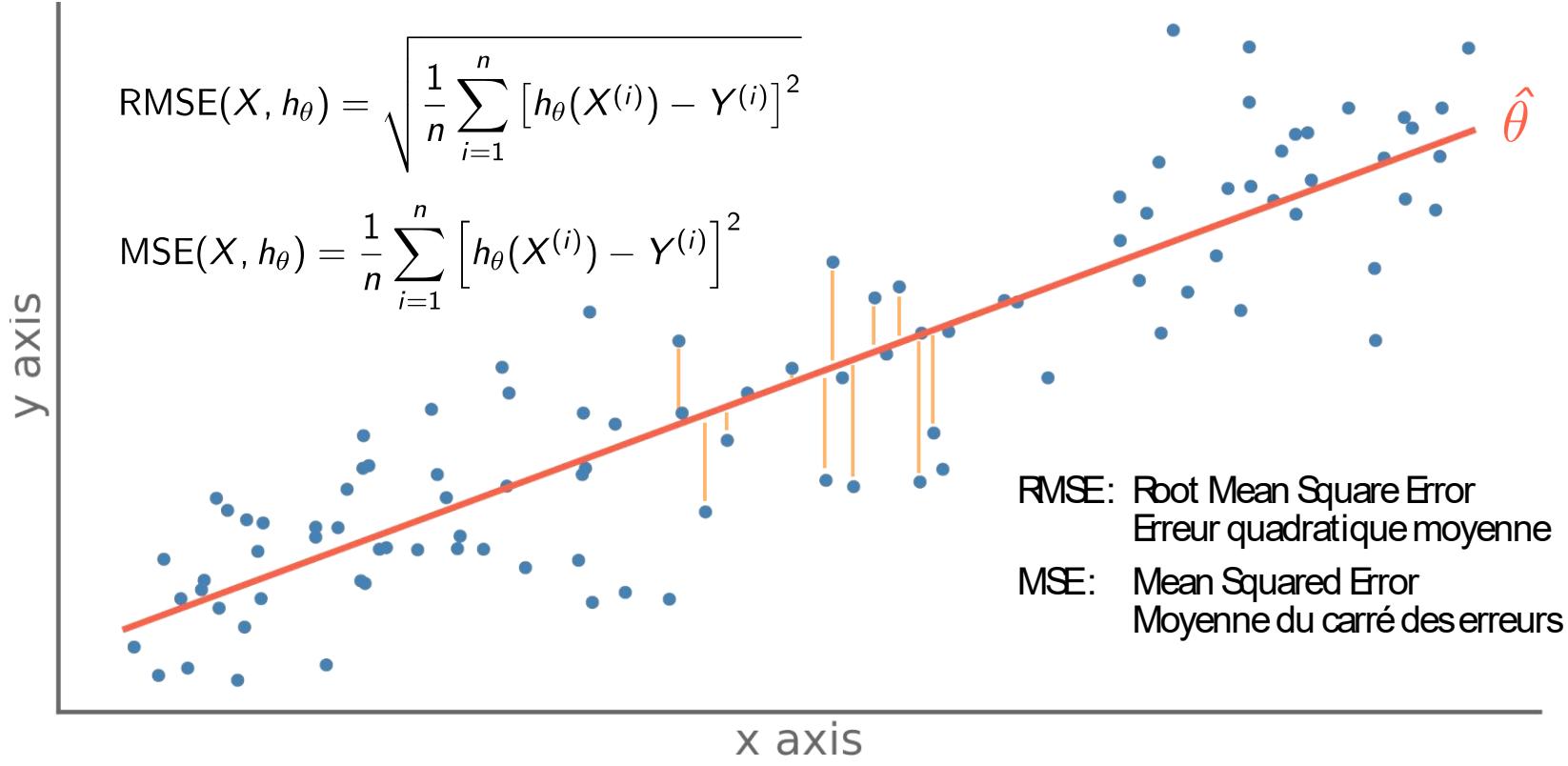
# LA RÉGRESSION LINÉAIRE

- Le modèle suppose une fonction de prédiction de forme
  - $f(x) = a_1x_1 + \dots + a_px_p + b = a \cdot x + b$
- L'apprentissage consiste à calculer les coefficients  $a$  et  $b$  qui minimisent l'erreur de prédiction (coût)
- Mais comment définir le coût ?



# LA FONCTION DE COÛT (LOSS)

- C'est l'écart moyen entre les prédictions et la vérité terrain



RMSE: Root Mean Square Error  
Erreur quadratique moyenne

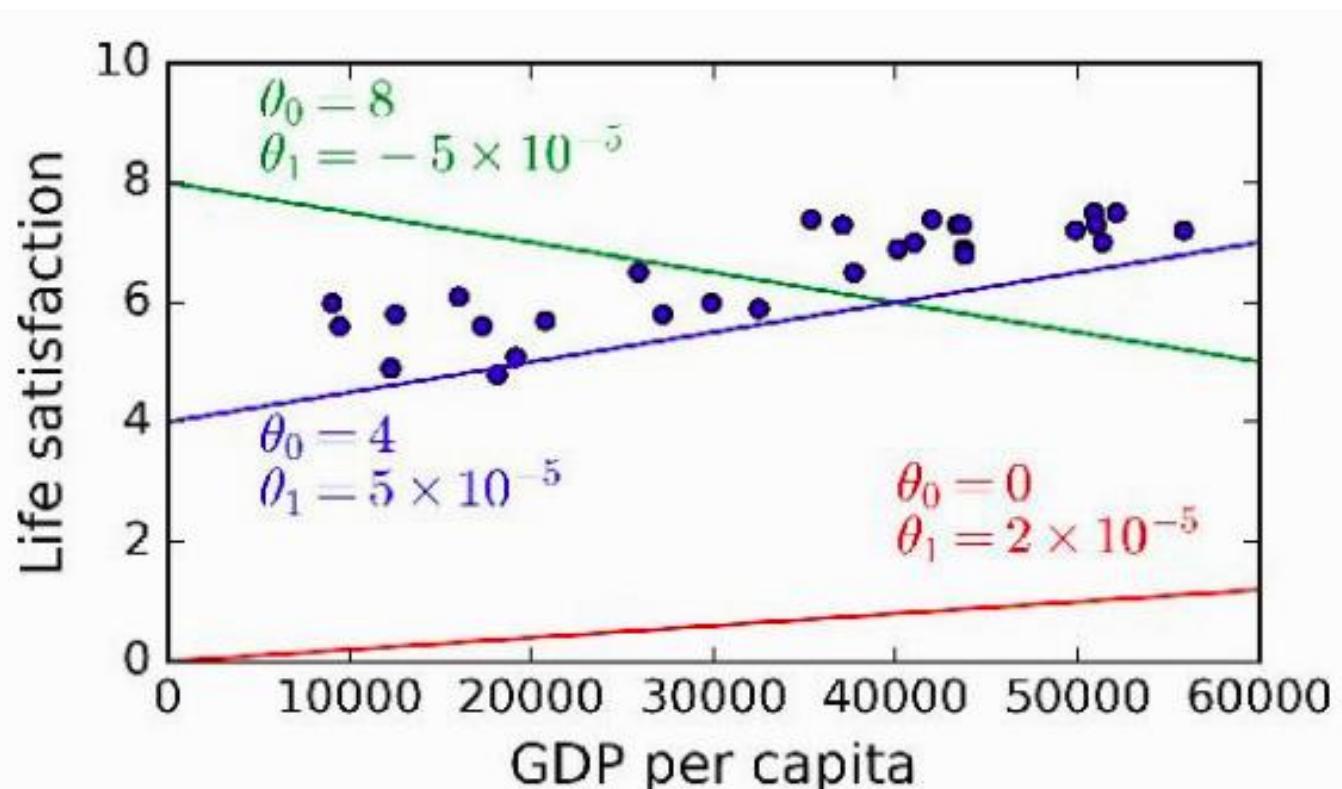
MSE: Mean Squared Error  
Moyenne du carré des erreurs

MAE : Mean Absolute Error  
Erreur absolue moyenne

$$MAE(X, h_{\theta}) = \frac{1}{n} \sum_{i=1}^n |h_{\theta}(X^{(i)}) - Y^{(i)}|$$

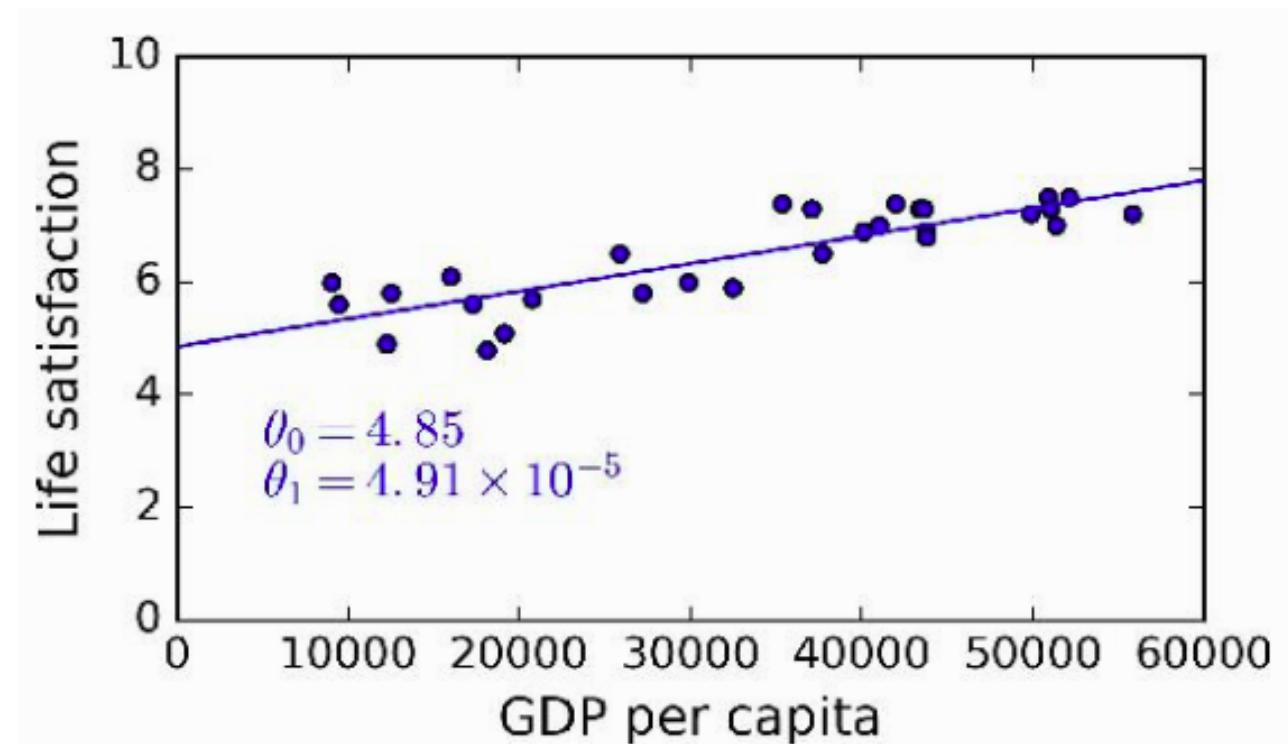
# EXEMPLE : EST-CE QUE L'ARGENT REND HEUREUX ?

- Ce modèle a deux paramètres,  $\theta_0$  et  $\theta_1$ .
- On peut créer une formule simple et essayer plusieurs combinaisons
  - $satisfaction = \theta_0 + \theta_1 \times GDP\_per\_capita$

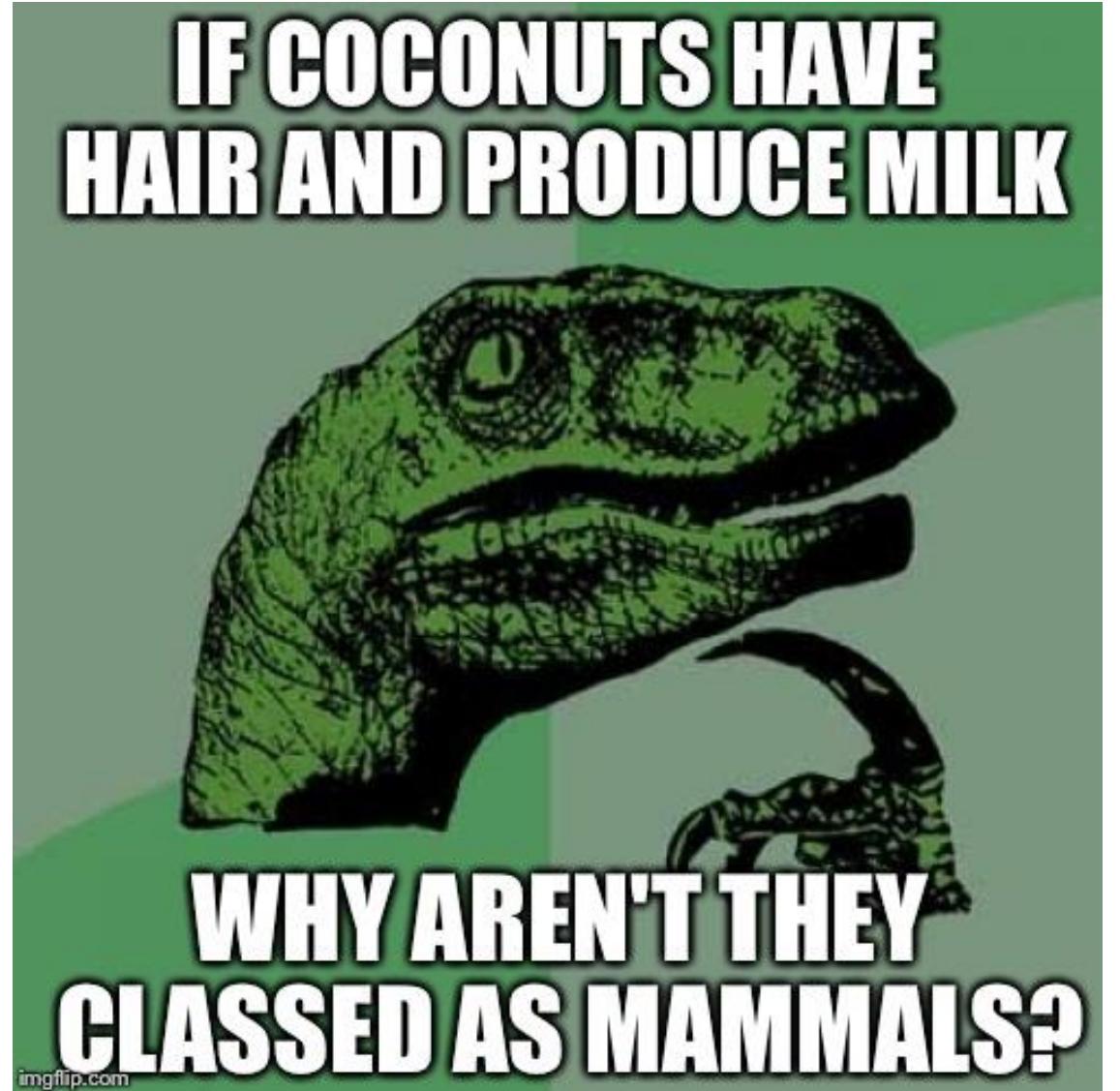


# EXEMPLE : EST-CE QUE L'ARGENT REND HEUREUX ?

- Le modèle qui mieux correspond aux données est
  - $\theta_0 = 4.85$
  - $\theta_1 = 4.91 \times 10^{-5}$
- Grâce à ce modèle, on peut essayer d'estimer la satisfaction de la population de Chypre
  - $\text{GDP\_per\_capita} = 22875 \text{ USD}$
  - $\text{Estimation} = 4.85 + 22875 * 4.91 * 10^{-5}$
  - $\text{Estimation} = 5.96$

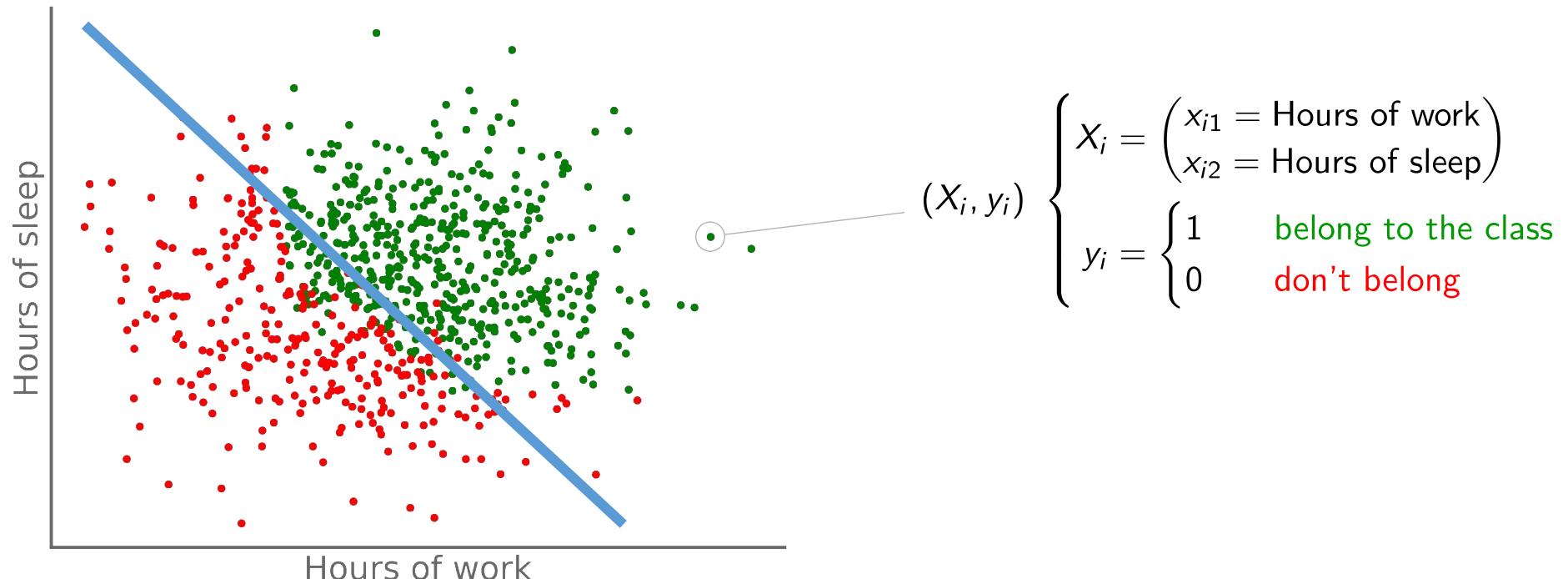


# De la Régression à la Classification



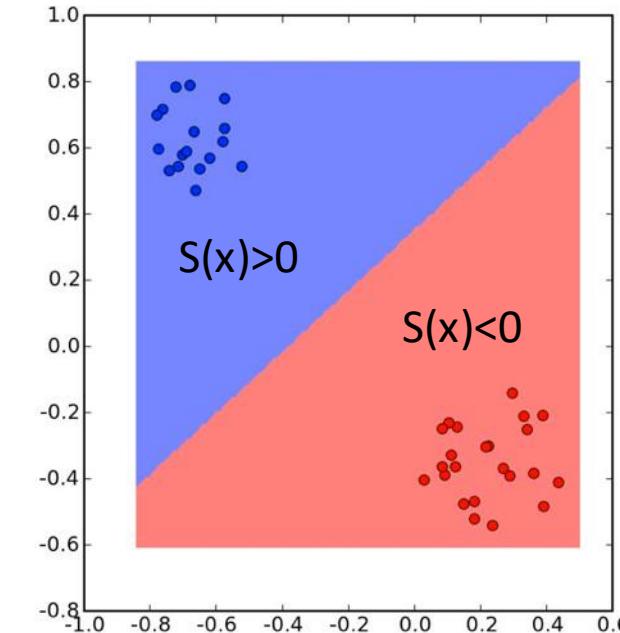
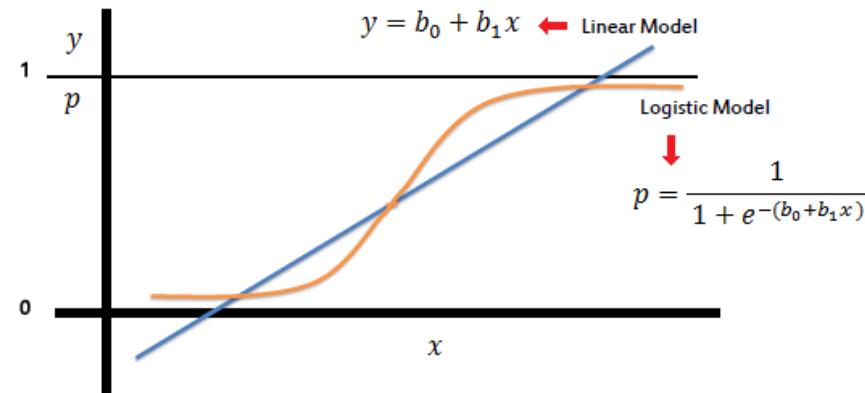
# LE CAS DE LA RÉGRESSION LOGISTIQUE

- Au lieu de trouver la ligne qui est la moins éloignée des points, **on cherche la ligne qui mieux les sépare**
  - Une Régression Logistique sert à donner une probabilité d'appartenance à une classe



# RÉGRESSION LOGISTIQUE

- Comme pour la régression linéaire, on cherche une fonction  $S(x) = a_1x_1 + \dots + a_px_p$  appelée **Score** qui doit délimiter deux groupes (classes) de données
- Le principe est de trouver des coefficients  $a$  de manière à ce que la valeur de  $S(x)$  soit positive lorsque les chances d'appartenir au groupe 1 sont grandes, et  $S(x)$  est négative si la probabilité est grande pour le groupe 0
- Une fonction d'interpolation  $\text{logit}(S) = 1/[1 + \exp(-S)]$  est souvent utilisée pour exprimer cette probabilité



# EXEMPLE : PROBABILITÉ DE PASSER LES EXAMENS

- Dans cet exemple issu de Wikipedia, on a un tableau avec 20 étudiants, indiquant combien d'heures ils ont étudié et s'ils ont passé les examens ou pas

Hours	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1

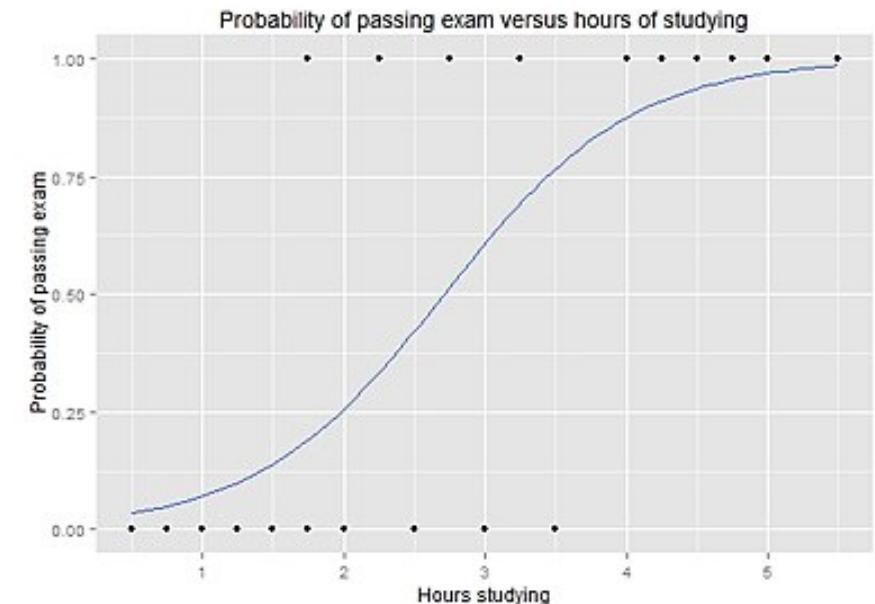
- L'analyse de ce graphique donne une fonction score

$$s(x) = 1.5046 \times \text{Hours} - 4.0777$$

- Ainsi, la probabilité de passer un examen est donné par

- $$\text{prob} = \frac{1}{1 + \exp(-(1.5046 \cdot \text{Hours} - 4.0777))}$$

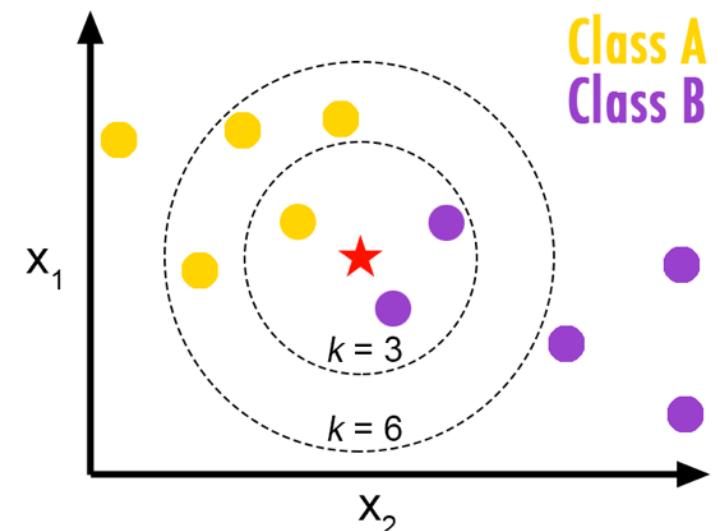
- Pour un étudiant qui s'est dédié uniquement 2h, la probabilité est de 26%
- Un étudiant qui a révisé 4h a 87% de probabilité de réussir



Graph of a logistic regression curve showing probability of passing an exam versus hours studying

# FAIRE UNE RÉGRESSION AVEC UNE CLASSIFICATION

- L'algorithme KNN (pour *K Nearest Neighbors*) est à la base un algorithme de classification supervisé
  - On représente les observations (données) dans un espace aux dimensions des variables prédictives
    - Ces données sont étiquetées
  - Pour une nouvelle donnée, on cherche l'étiquette qui est la plus fréquente dans un rayon autour de cette nouvelle entrée (les *K* premiers voisins)
    - Besoin d'un concept de distance
      - Euclidienne ?

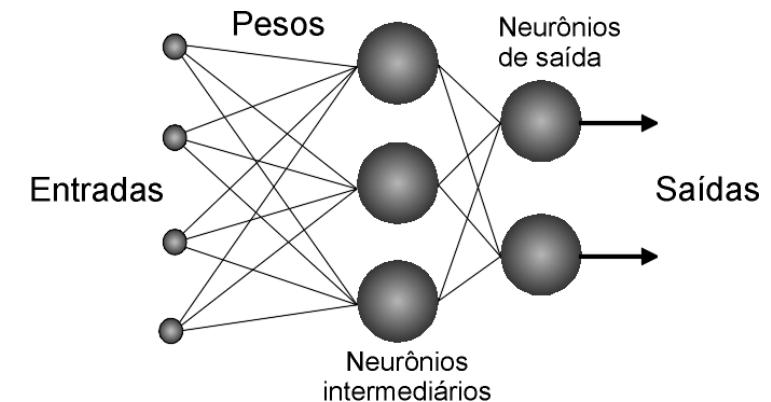
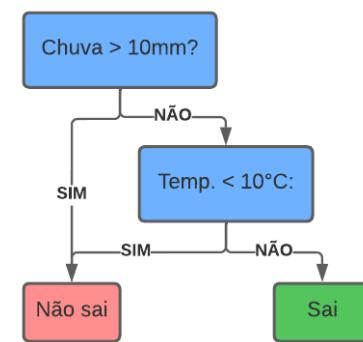
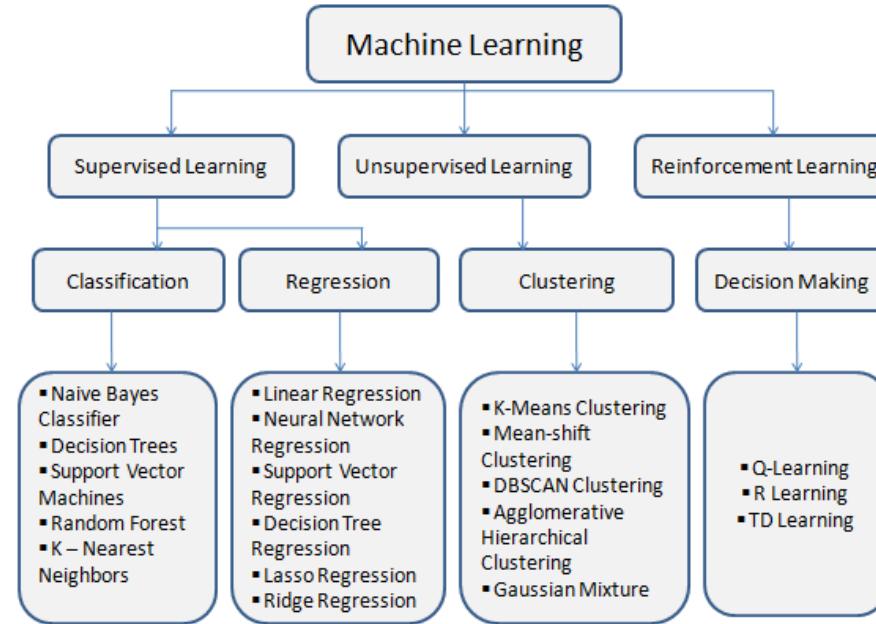


# EXEMPLE D'USAGE KNN

- On reprend l'exemple sur l'indice de satisfaction des pays
- Au lieu d'essayer d'estimer les paramètres  $\theta_0$  et  $\theta_1$ , on regarde les pays avec des rentes proches de celle de Chypre (22587 USD)
  - K=1
    - La Slovénie a la rente per capita la plus proche (20732 USD)
    - Son indice de satisfaction est de 5.7
  - K=3
    - En plus de la Slovénie, on trouve aussi le Portugal (19122 USD, sat=5.1) et l'Espagne (25865 USD, sat=6.5). Si on fait une moyenne, on obtient un indice de 5.77
    - Pas si loin de l'indice 5.96 obtenu par la régression linéaire

# MACHINE LEARNING "CLASSIQUE" VS DEEP LEARNING

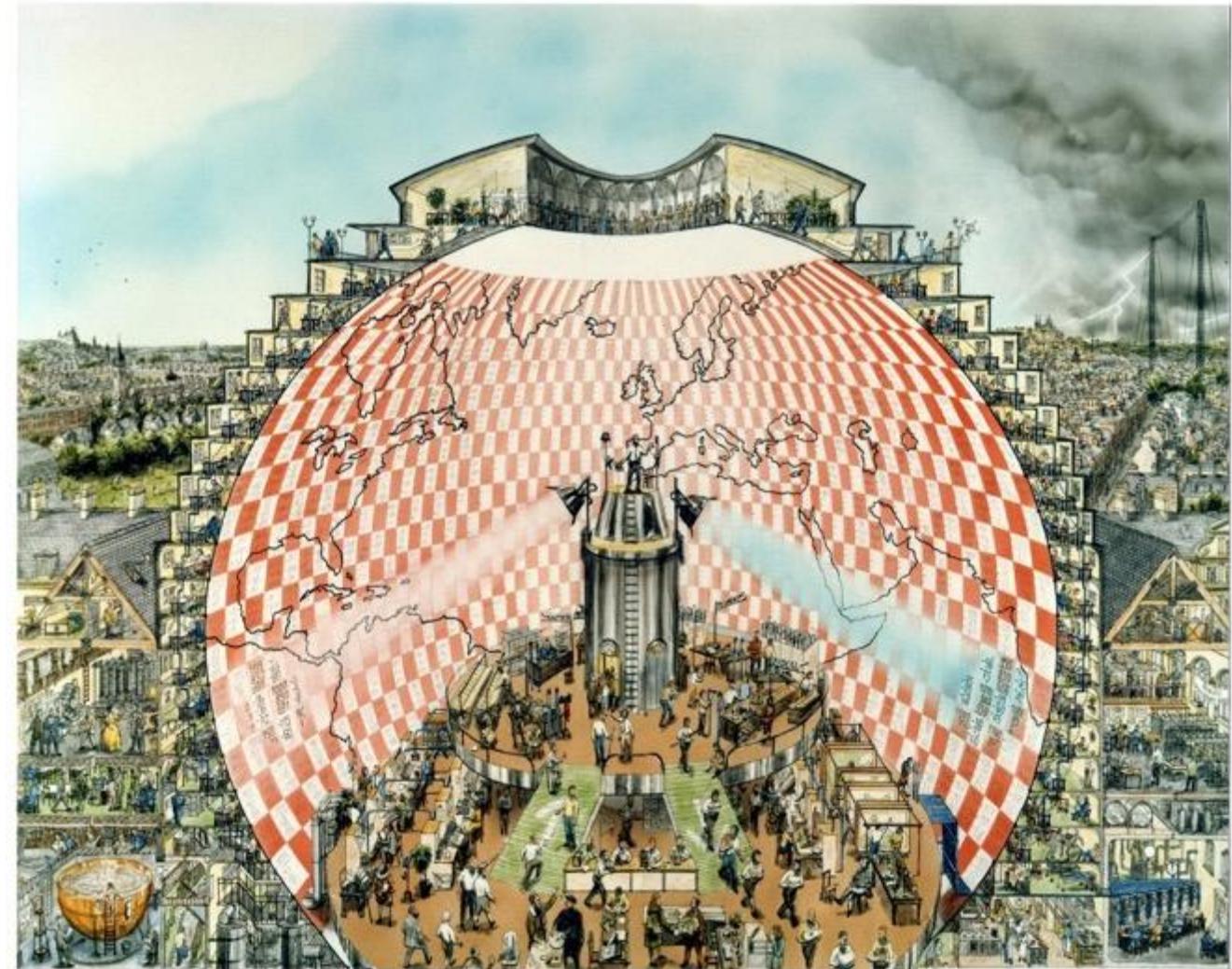
- Le Machine Learning classique est tout à fait pertinent pour plusieurs problèmes
- Ces dernières années nous avons vu l'émergence du Deep Learning
  - Des réseaux de neurones profonds
- Le Deep Learning est très puissant pour l'analyse d'images, séries temporelles
  - Encore, il faut savoir utiliser
- Dans ce cours, les prochaines séances vont se concentrer sur le DL
  - Les fondements
  - Analyse d'images
  - Séries temporelles



Et la météorologie dans tout ça ?

# MODÉLISATION NUMÉRIQUE

- Les principes de la modélisation mathématique du climat datent des années 20
  - Livre "Numerical Weather Prediction" de L.F. Richardson
  - Algorithmes et statistique pour résoudre des équations physiques
- 6 semaines pour calculer 6 heures de prévision
  - Erreur énorme, augmentation de 145 hPa en 6h
- Une constatation :
  - Grand besoin de calcul pour traiter la quantité de données

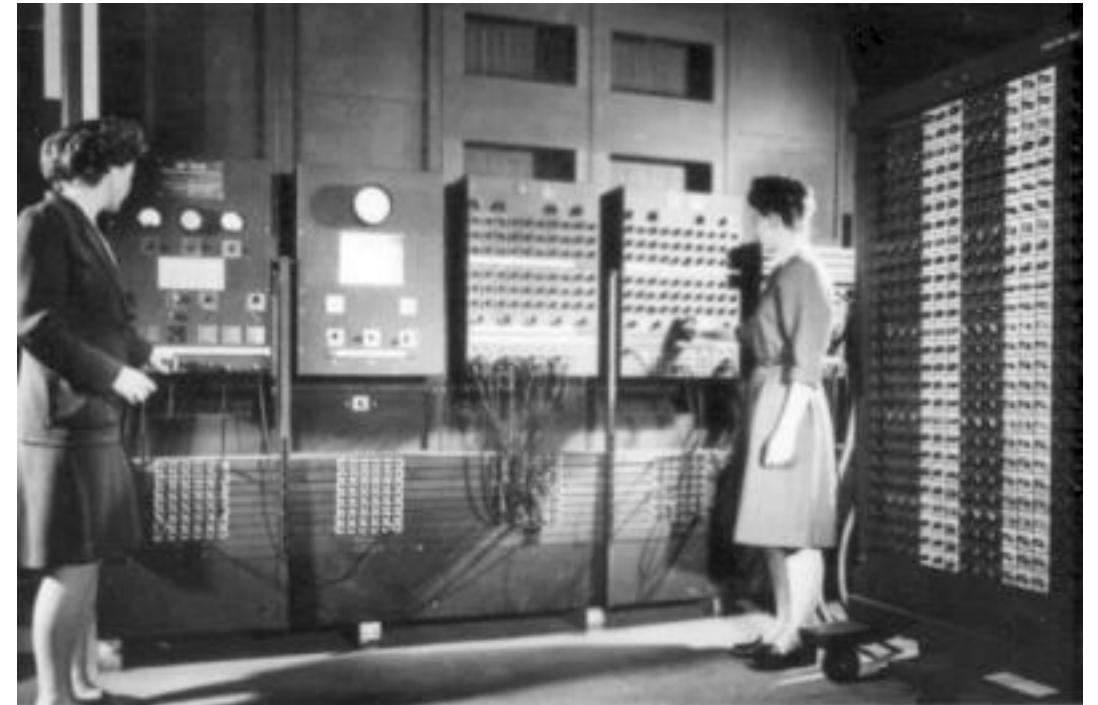


# UNE PREMIÈRE MODÉLISATION PAR ORDINATEUR

- Seulement en 1950 on a vu les premiers modèles informatiques
  - Jules Charney et John Von Neuman, avec l'ordinateur ENIAC
- Intégration de l'équation de vorticité barotropique

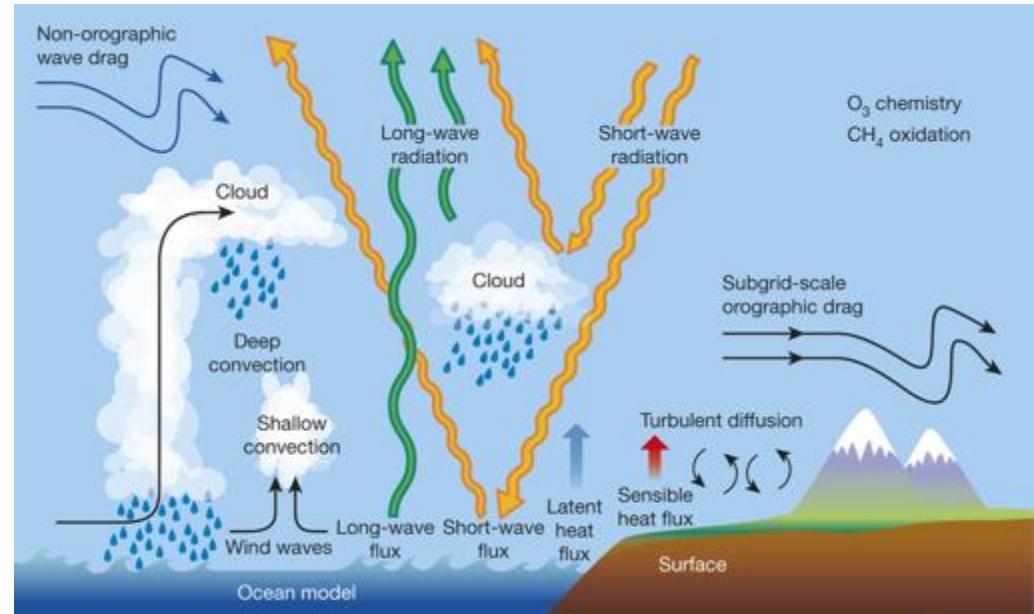
$$\frac{\partial \nabla^2 \psi}{\partial t} = \frac{1}{a^2} \left[ \frac{\partial \psi}{\partial \mu} \frac{\partial \nabla^2 \psi}{\partial \lambda} - \frac{\partial \psi}{\partial \lambda} \frac{\partial \nabla^2 \psi}{\partial \mu} \right] - \frac{2\Omega}{a^2} \frac{\partial \psi}{\partial \lambda}$$

- Modèle simplifié, une seule couche
  - Presque 24h de calcul pour 24h de prévision
- Aujourd'hui les machines sont plus performantes !!



# LE DILEMME DE LA MODÉLISATION

- Tout modèle est régi par des **équations fondamentales**
  - Conservation de la masse, énergie, momentum
  - Dynamique des fluides
  - ...
- Ces équations sont accompagnées de multiples **paramétrages**
  - Approximation numérique -> précision, stabilité, coût
  - Échelle de résolution -> taille de la grille
  - Influence de la surface terrestre
  - Couverture des nuages, précipitation
  - Facteurs chimiques et biologiques
- La plupart de ce **paramétrage** est **empirique**
  - Tentative et erreur
- Opportunités pour l'automation / auto-correction ?



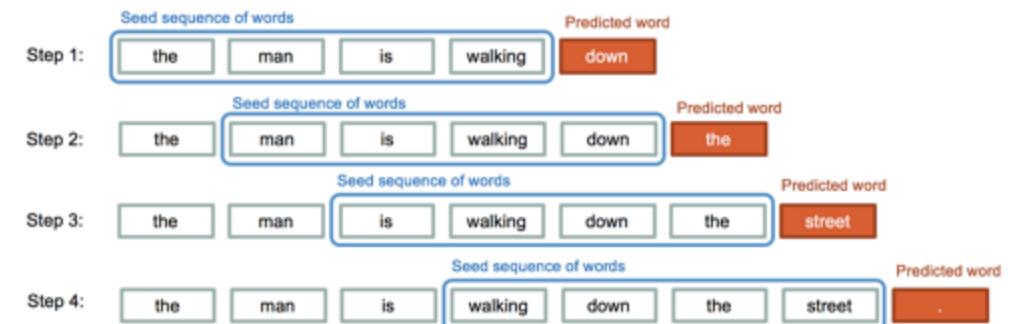
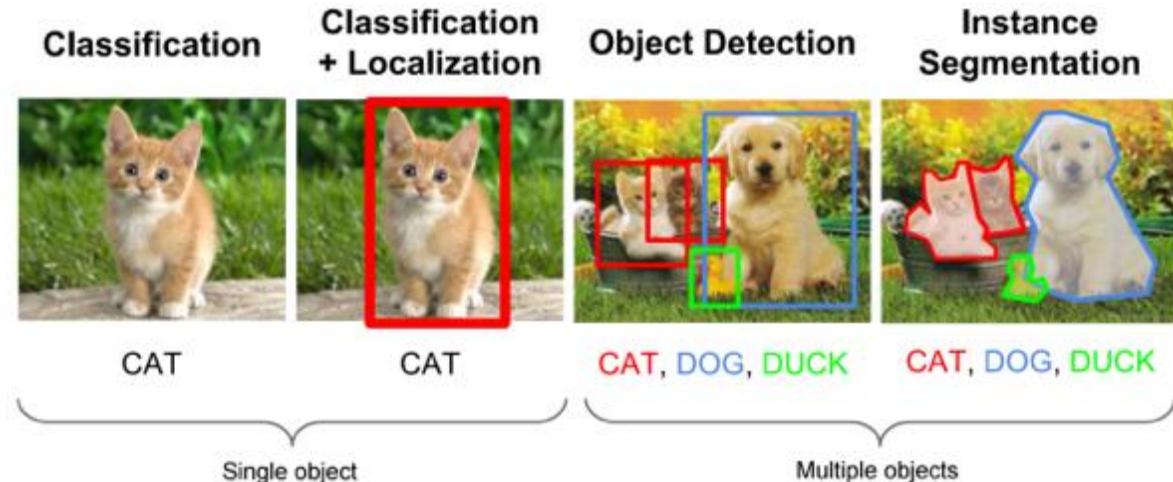
# DEEP LEARNING EVERYWHERE

- Les réseaux convolutionnels (CNN) sont le moteur de la révolution Deep Learning

- Classification d'images
- Analyse de textes/traductions
- Reconnaissance vocale
- Séries temporelles

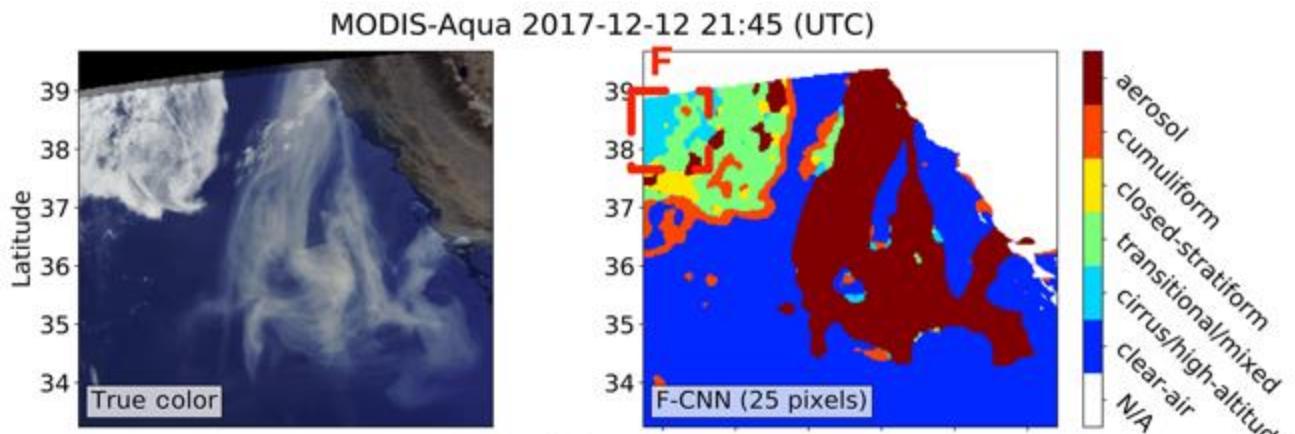


- Et les applications en météorologie ?



# ANALYSE ET CLASSIFICATION D'IMAGES

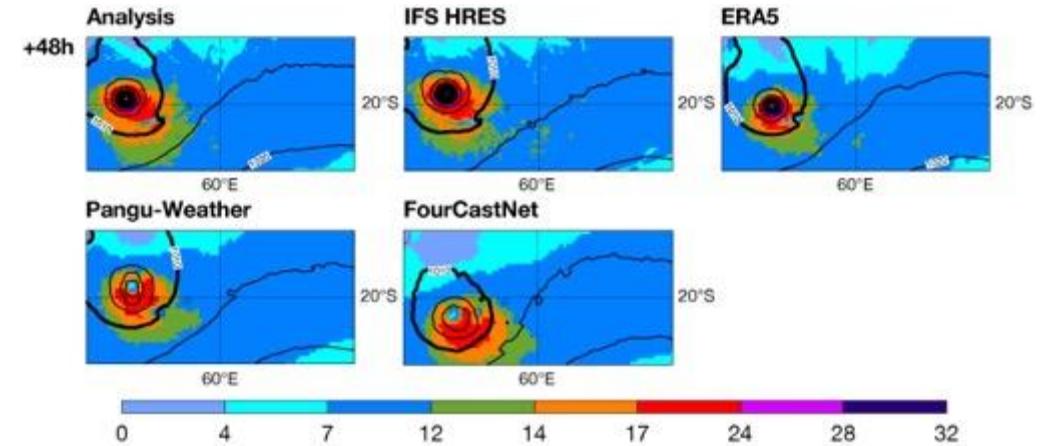
- Identification d'évènements extrêmes
  - Ouragans, typhons...
  - Vagues de chaleur
- Segmentation des types de nuages



<https://amt.copernicus.org/articles/13/5459/2020/>

- Mais aussi une vue sur la surface

<http://ieee-dataport.org/open-access/benchmark-dataset-automatic-damaged-building-detection-post-hurricane-remotely-sensed>

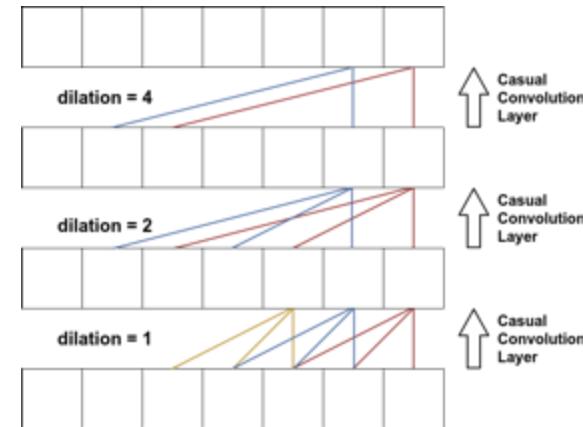
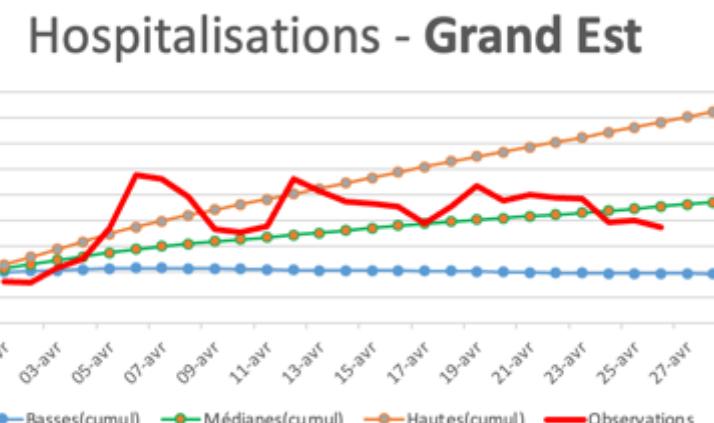
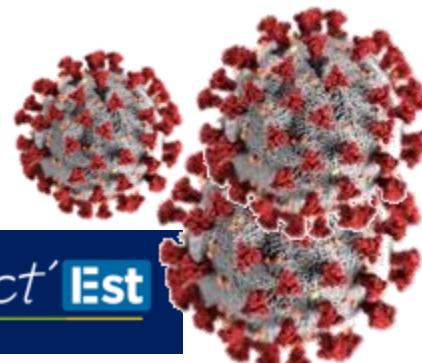


'[www.ecmwf.int/en/about/media-centre/science-blog/2023/rise-machine-learning-weather-forecasting](https://www.ecmwf.int/en/about/media-centre/science-blog/2023/rise-machine-learning-weather-forecasting)



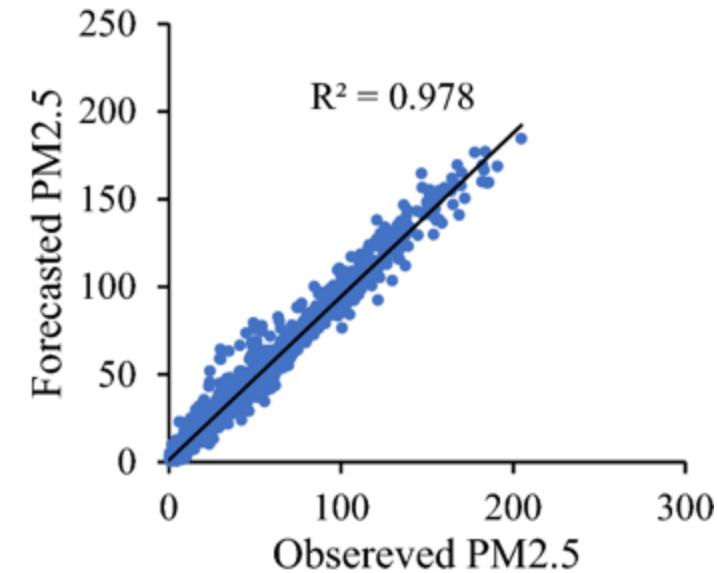
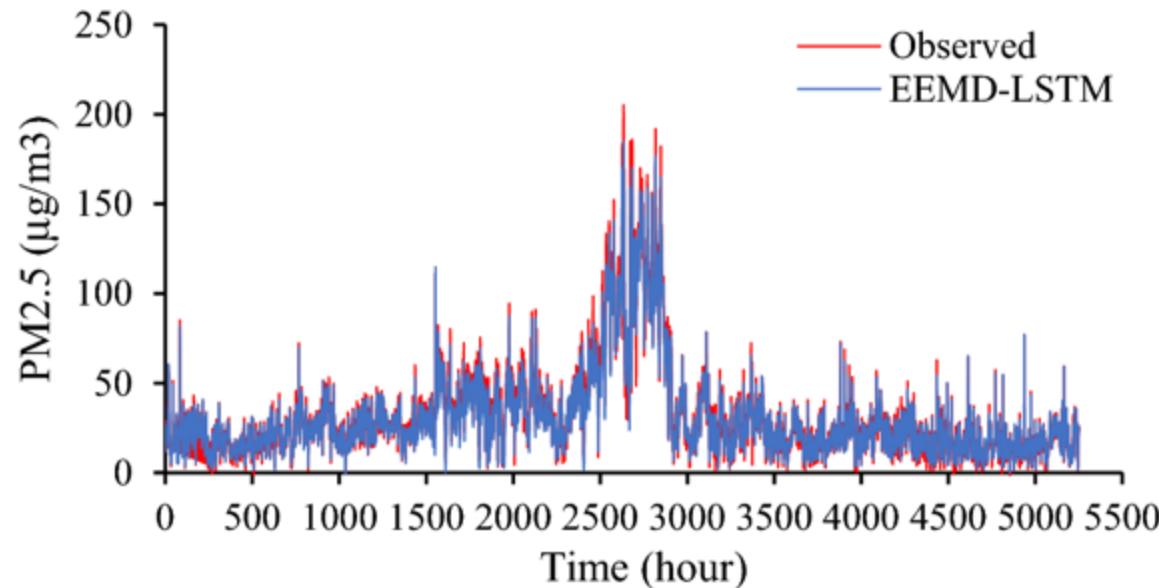
# SÉRIES TEMPORELLES MULTIPARAMÉTRIQUES

- Un réseau neural "simples" ne sait pas garder des dépendances temporelles
- Il existe des modèles avec mémoire des évènements précédents
  - RNN (Réseaux récursifs), LSTM (Long-Short term memory), GRU (Gated Recurrent Unit)
  - Encoder-Decoder, TCN (Temporal Convolution Networks)
- Avantages
  - Données multivariate (et possibilité de sorties multiples)
  - Différents types de génération (1-step, multi-step)
- Largement utilisés pendant COVID-19
  - Les modèles traditionnels étaient insuffisants
  - Pas assez de données



# SÉRIES TEMPORELLES MULTIPARAMÉTRIQUES

- Ex : plusieurs travaux avec LSTM pour la prédition en environnement urbain
  - Ex : prévision 1h pour PM2.5 en Malaisie

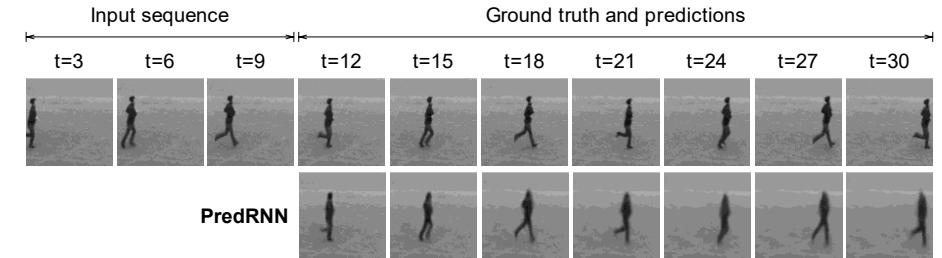


<https://www.nature.com/articles/s41598-022-21769-1>

# GÉNÉRATION DE VIDÉOS (NEXT FRAME)

- Recherche importante sur le sujet "next frame prediction"

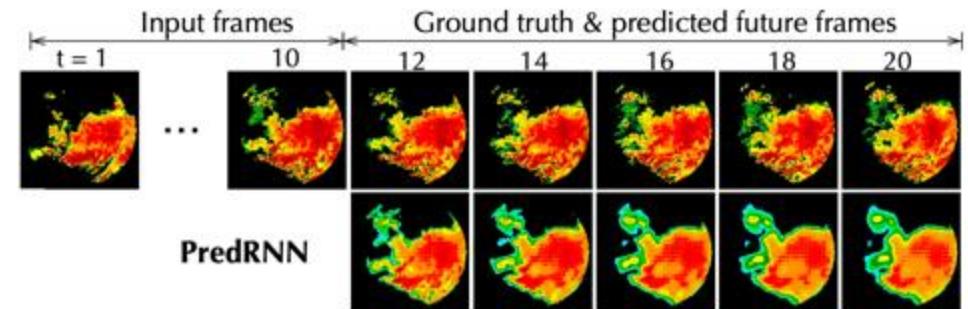
- Quelle sera la prochaine action
    - Trajectoire d'une voiture ou personne
  - Compléter une séquence de données



- Next Frame a besoin d'un apprentissage espace-temporel

- Ne pas déformer l'objet (états "connus")
  - Garder une trajectoire cohérente
  - Estimer les interactions entre objets

- Très vite, ce type d'application a été détourné pour les prévisions météorologiques, notamment le "nowcast"



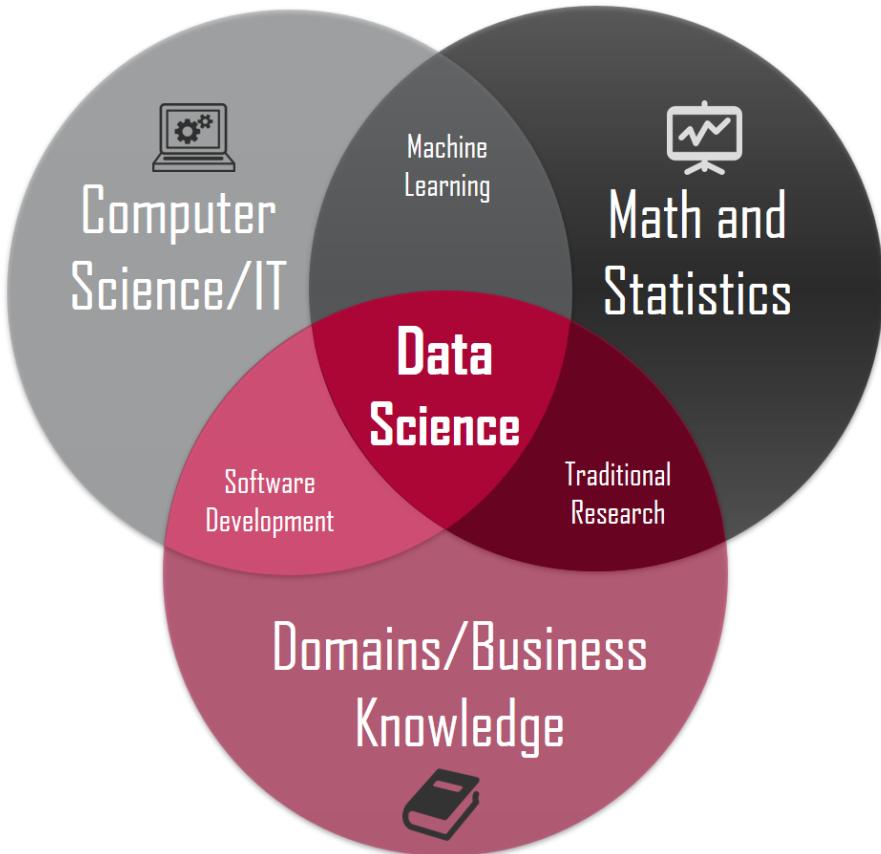
PredRNN++: Towards a Resolution of the Deep-in-Time Dilemma in Spatiotemporal Predictive Learning. Wang, Yunbo and Gao, zhifeng and Long, Mingsheng and Wang, Jianmin and Yu, Philip S., ICML, 2018

Avant tout, les données

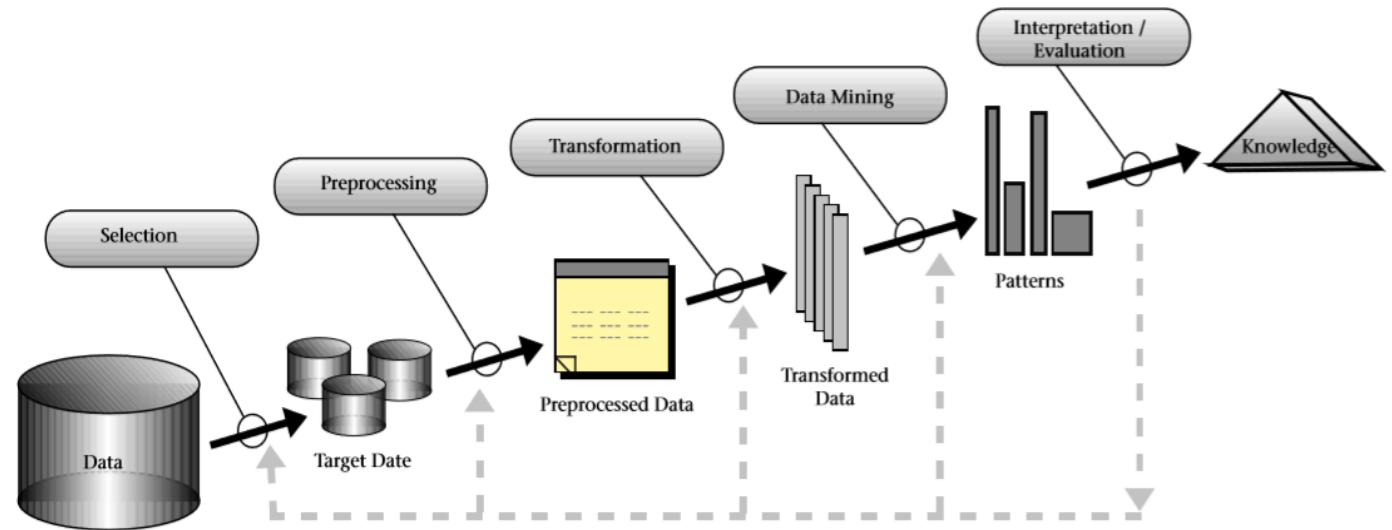
# LES PREMIERS 80% DU TRAVAIL

- Souvent il est dit que la préparation des données représente les premiers 80% du travail
- Les outils Big Data ou même les bases de données ne sont que la partie ETL
  - Extraction
  - Transformation
  - Load (chargement)
- Une fois les données prêtes, il faut savoir les utiliser/analyser (les autres 80%)
  - Utilisation directe (bases de données, requêtes SQL)
  - Analyse des données (Machine Learning, Data Mining)
- Il faut aussi les visualiser (les troisièmes 80% ?)
- Dans le cas du Machine Learning, un langage s'est imposé : Python

# DATA SCIENCE / DATA MINING



La fouille de données est un ensemble de techniques et méthodes permettant l'analyse et l'exploitation de données afin d'en extraire des connaissances.



Processus de découverte de connaissance KDD

Source : Fayyad et al., 1996

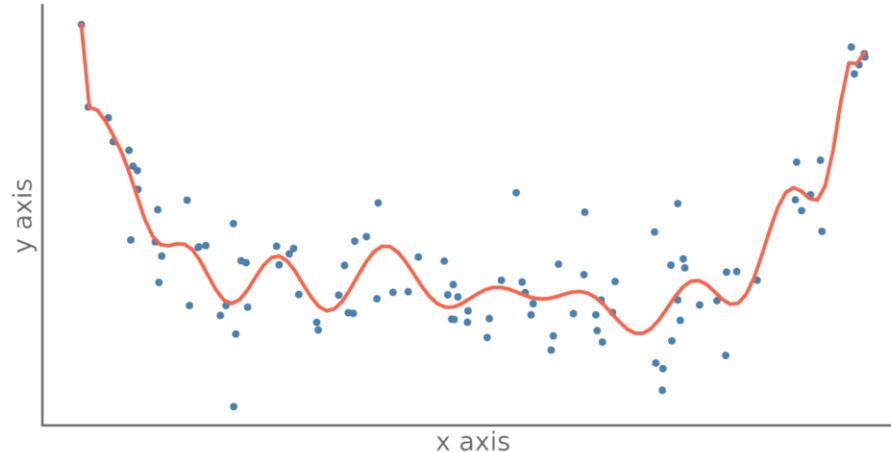
# DIFFÉRENTES ÉTAPES AVANT LE MODÈLE IA

- Obtention des données
- Nettoyage et sélection
- Formatage
- Exploratory Data Analysis (EDA)
- Séparation des données
  - Groupe d'entraînement (train)
  - Groupe de validation (val)
  - Groupe de test (test)

# POURQUOI DÉCOUPER SES DONNÉES ?

- Première impulse : utiliser toutes les données !!!
  - Pas bon, on risque de surentraîner (apprendre par cœur) et le modèle ne pourra pas généraliser

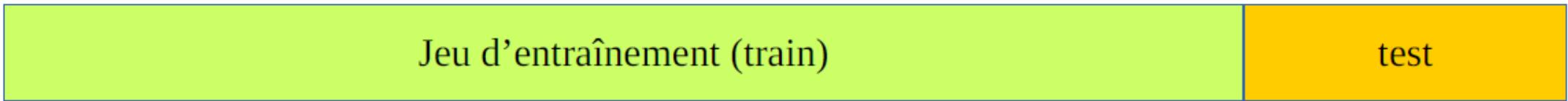
Jeu de données d'entraînement



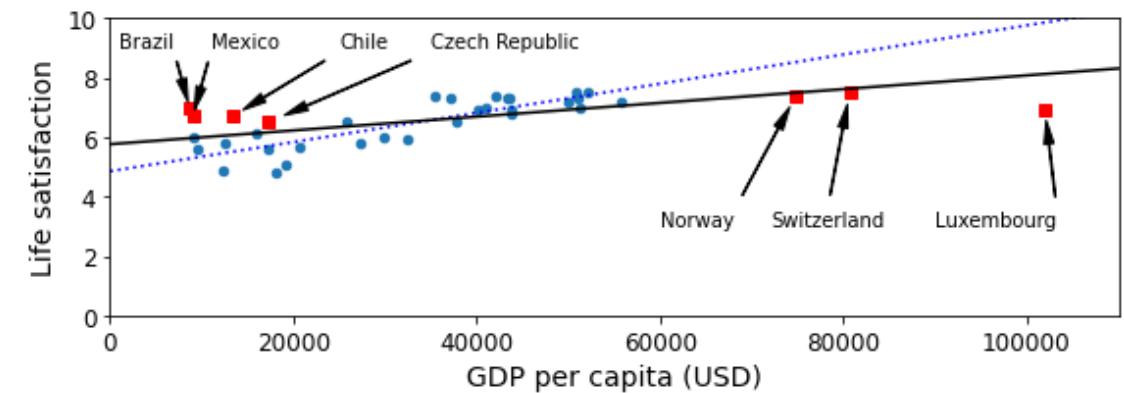
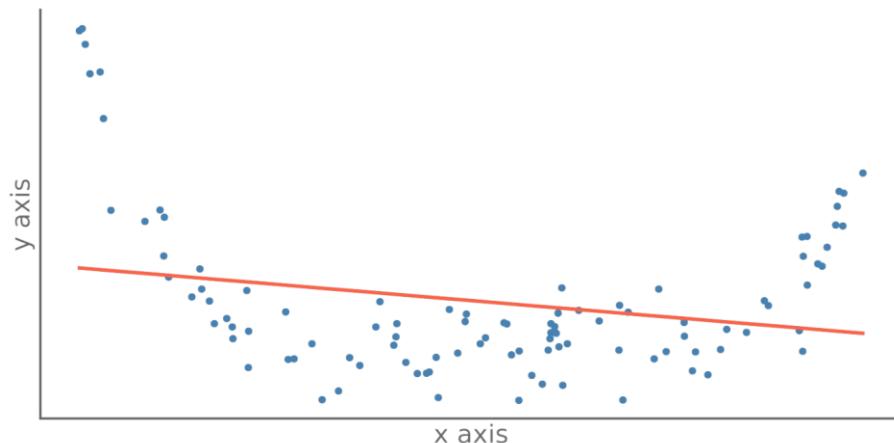
Overfitting

# DÉCOUPER SES DONNÉES

- Séparer un groupe pour l'entraînement et autre pour la validation
  - \*\*\* CONFUSION : "val" est souvent appelé "test" si on découpe en 2 groupes

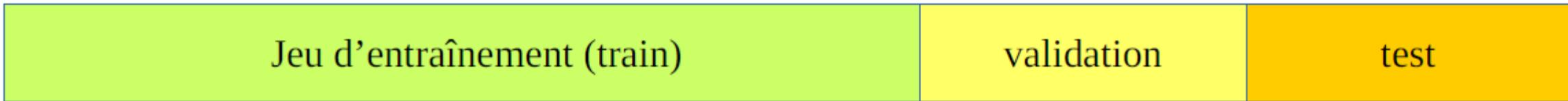


- Aucune garantie que l'algorithme fonctionnera bien avec de nouvelles données
  - Underfitting : modèle trop simple pour expliquer la variance



# DÉCOUPER SES DONNÉES

- Approche recommandée (pas tjs suivie) :
  1. entraîner sur le jeu d'entraînement (train),
  2. vérifier si le modèle marche sur un jeu de validation (valid),
  3. Répéter 1-2 jusqu'à ce que le modèle soit acceptable
    - 1. Ou alors entraîner plusieurs modèles différents pour les comparer
  4. Une fois le modèle entraîné, l'évaluer sur un jeu de test (test)



# LES POINTS D'ATTENTION : LES BIAIS

l'IA amplifie  
les biais

## Raciste et détestable, l'intelligence artificielle tente de progresser

Depuis ses premiers pas, l'outil GPT-3 – une IA capable d'écrire des textes originaux – a fait l'admiration de tout un chacun. OpenAI, avec cette troisième évolution du projet parvenait à faire rédiger poésie, articles de presse ou même code de programmation. Cependant, les dérives sont rapidement arrivées, avec une Intelligence qui virait à la grossièreté, voire aux propos toxiques. Que faire ? Une bonne correction, tout simplement.



écolier



écolière



# ATTENTION : DONNÉES PERSONNELLES

Contexte : RGPD

- **Transparence** : « Que fait-on de vos données »
- **Évaluation des risques** : évaluer et atténuer les risques de confidentialité à l'avance
- **Audits** : mieux éclairer la position de l'entreprise sur le plan de l'IA et la confidentialité. Difficulté de l'audit des algorithmes et de l'
- **Explicabilité**

