

LES CHATBOTS VUS DE L'INTÉRIEUR

LE FONCTIONNEMENT DES RÉSEAUX DE NEURONES

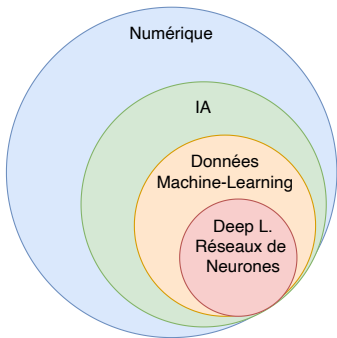
Vendredi 27 Mars 2026
Maths en Jeans, AgroParisTech

Vincent Guigue
<https://vguigue.github.io>

INTRODUCTION



Artificial Intelligence & Machine Learning



Input ($\mathbf{x}$)	Output ($\mathbf{Y}$)	Application
email	→ spam? (0/1)	spam filtering
audio	→ text transcript	speech recognition
English	→ Chinese	machine translation
ad, user info	→ click? (0/1)	online advertising
image, radar info	→ position of other cars	self-driving car
image of phone	→ defect? (0/1)	visual inspection

IA : programmes informatiques qui s'adonnent à des tâches qui sont, pour l'instant, accomplies de façon plus satisfaisante par des êtres humains car elles demandent des processus mentaux de haut niveau.

Marvin Lee Minsky, 1956

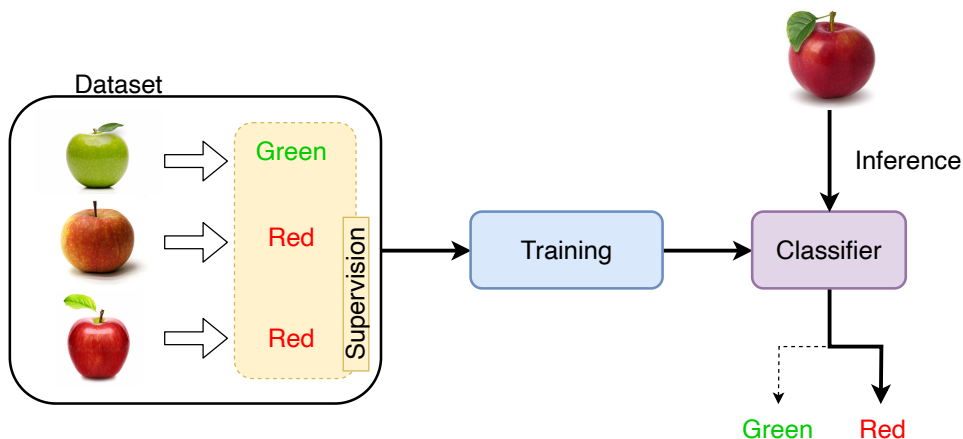
N-AI (Narrow Artificial Intelligence), dédiée à une tâche unique

≠ IA-G (IA Générale), qui remplace l'humain dans les systèmes complexes.

Andrew Ng, 2015



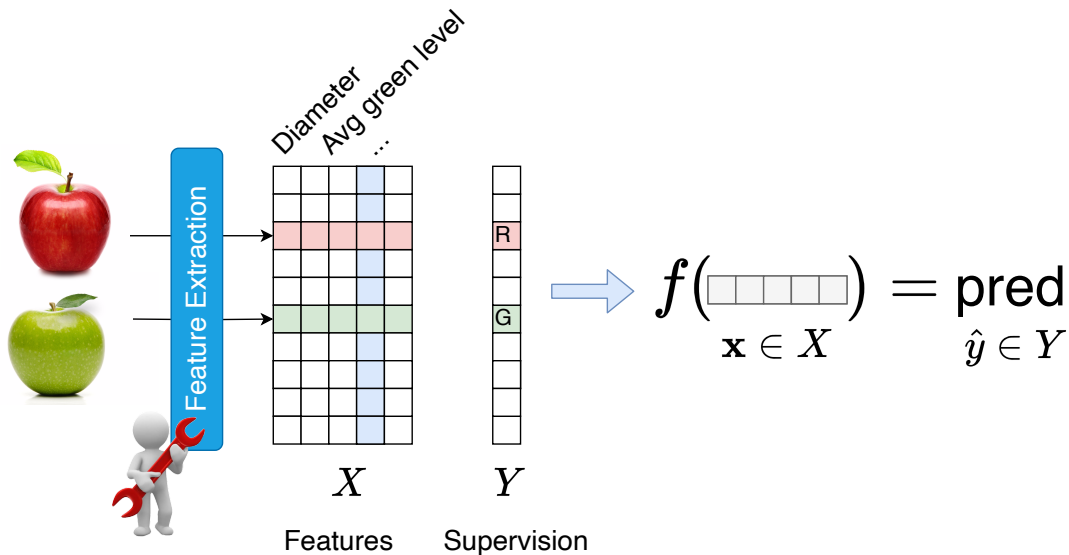
Chaîne de Traitement Supervisé & Modèles



- Promesse = construire un modèle *uniquement* à partir d'observations

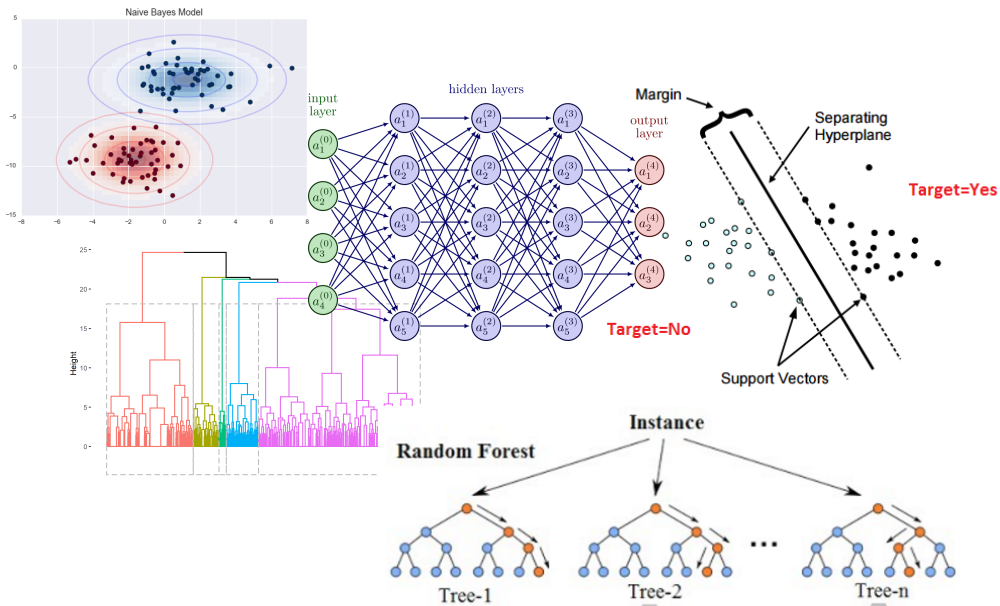


Chaîne de Traitement Supervisé & Modèles



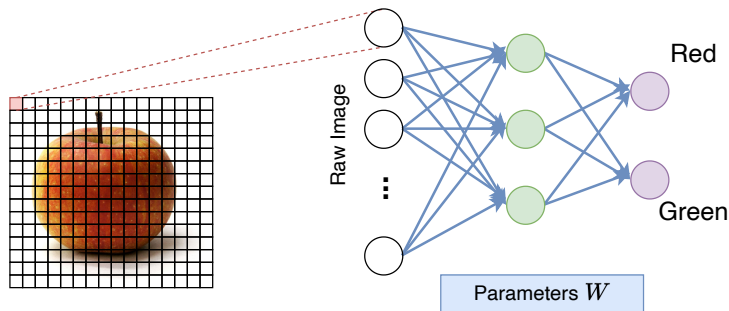


Chaîne de Traitement Supervisé & Modèles





Chaîne de Traitement Supervisé & Modèles

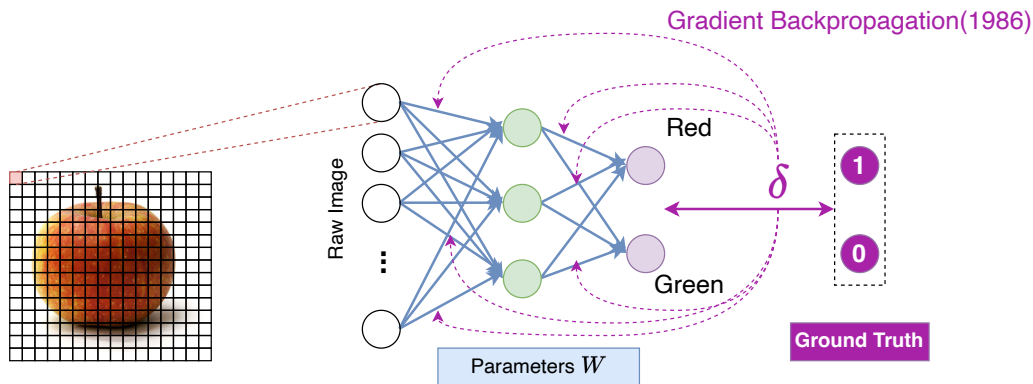


■ Initialisation aléatoire...

Et prise de décision aléatoire (au début !)



Chaîne de Traitement Supervisé & Modèles

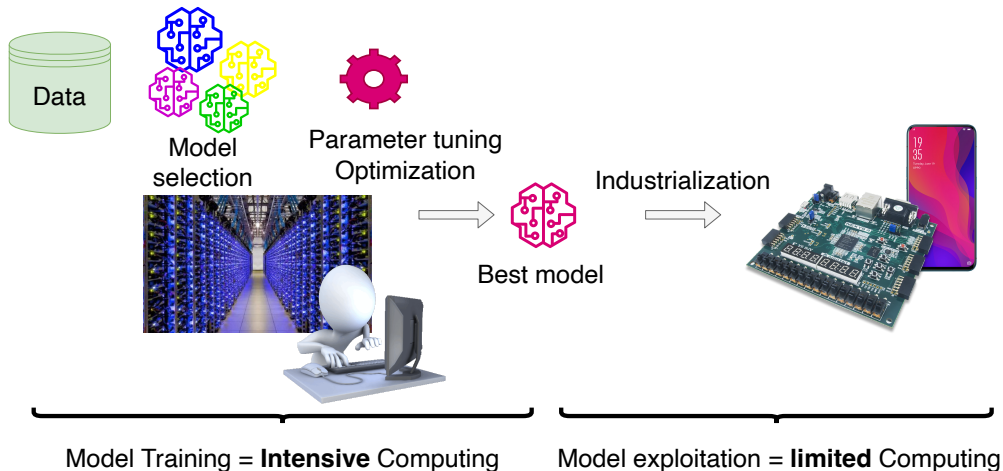


- Mise à jour des poids
- Pas de taille epsilon, nombreuses itérations sur les données



Apprentissage/Construction vs Déploiement/Exploitation

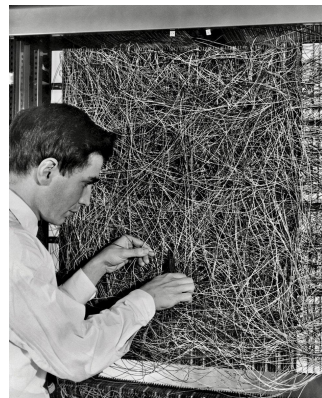
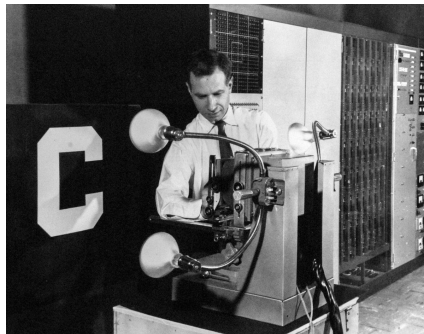
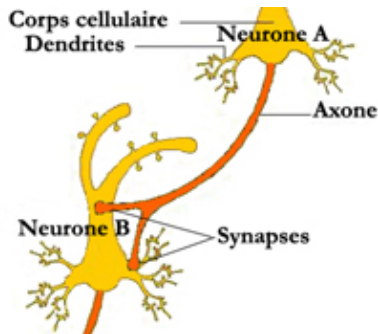
Différentes étapes en apprentissage automatique



FONCTIONNEMENT DES RÉSEAUX DE NEURONES



Inspiration biologique [plus ou moins lointaine]



- Feature
- Fusion de message = addition
- Activation = signe (=décision)

1958 Perceptron de Rosenblatt

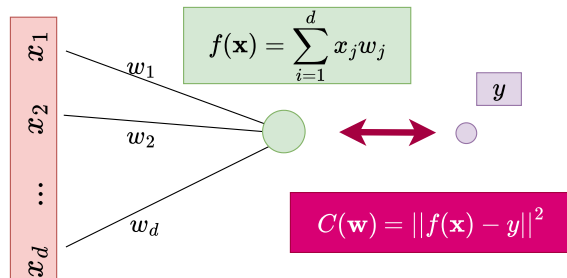
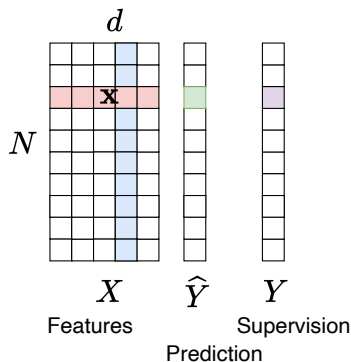
1979 Neo-Cognitron de Fukushima $\approx$ CNN

1986 Rétro-propagation du gradient
Rumelhart, Hinton, Williams



Les origines de l'apprentissage profond

$$\mathbf{x} = [x_1, x_2, \dots, x_d]$$

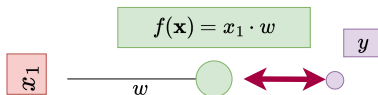
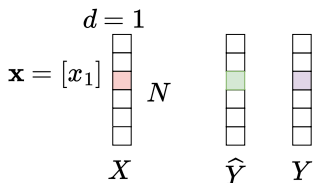
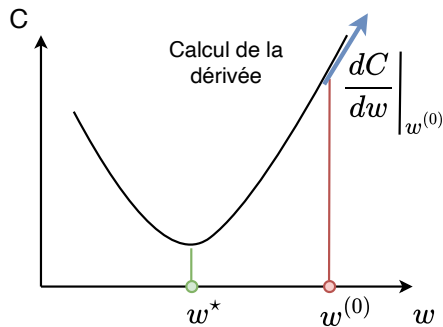
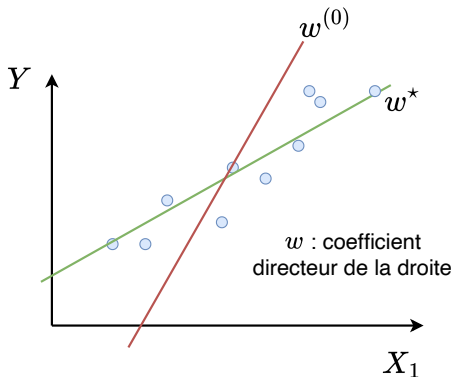


- Optimiser les paramètres $\mathbf{w} = [w_1, \dots, w_d]$
- ... Pour que $\hat{y}$ devienne proche de y
- au sens d'une métrique :

$$\mathcal{C}(\mathbf{w}) = \sum_i (f(\mathbf{x}_i) - y_i)^2$$

Comment trouver les paramètres w optimaux?

Dans le cas 1 dimension : $d = 1$

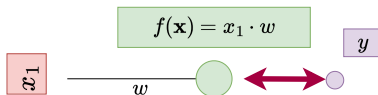
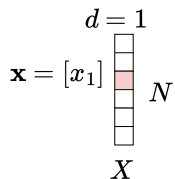
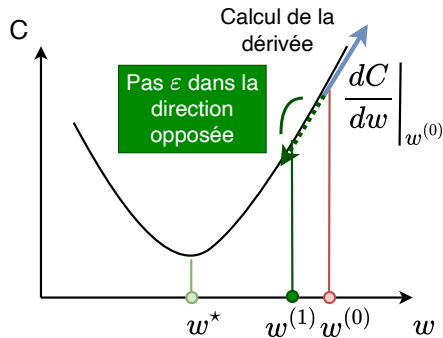
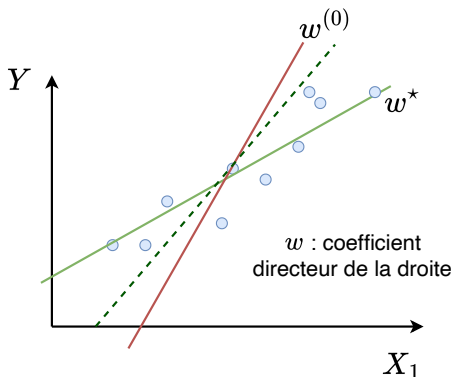


$$C(\mathbf{w}) = \sum_i (f(\mathbf{x}_i) - y_i)^2$$



Descente de gradient = exploitation de la dérivée

Dans le cas 1 dimension : $d = 1$

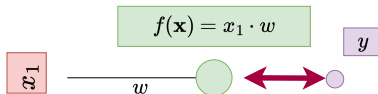
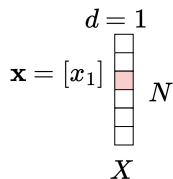
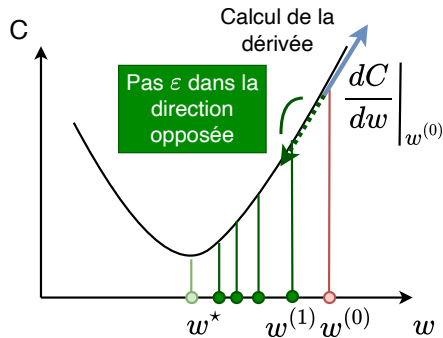
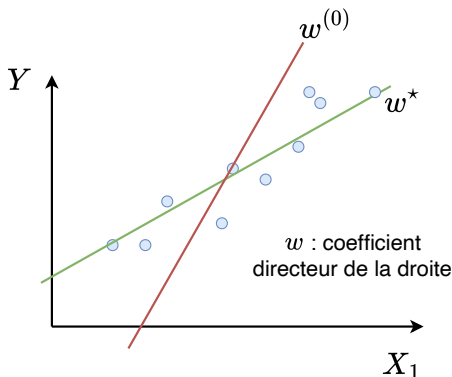


$$\mathcal{C}(\mathbf{w}) = \sum_i (f(\mathbf{x}_i) - y_i)^2$$



Descente de gradient = exploitation de la dérivée

Dans le cas 1 dimension : $d = 1$

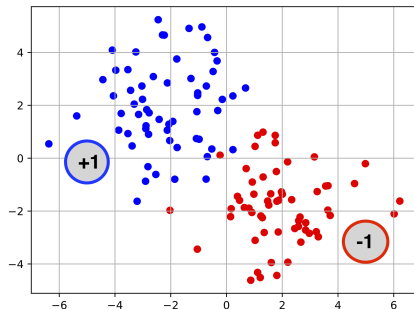


$$\mathcal{C}(\mathbf{w}) = \sum_i (f(\mathbf{x}_i) - y_i)^2$$

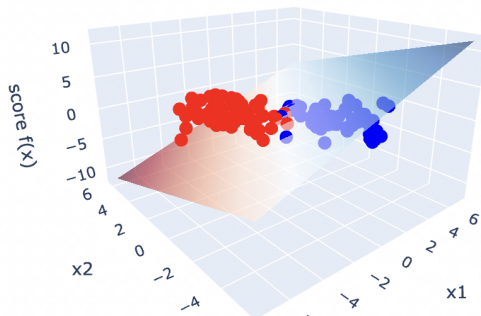


Passage en deux dimensions

Données 2D



Forme de la fonction de décision $f(\mathbf{x})$



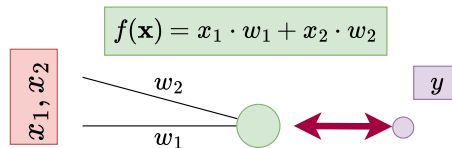
$$\mathbf{x} = [x_1, x_2]$$

$$d = 2$$

$$\begin{matrix} \begin{matrix} \square & \square \\ \square & \square \\ \textcolor{red}{\square} & \textcolor{red}{\square} \\ \square & \square \\ \square & \square \\ \square & \square \end{matrix} & N \\ X & \end{matrix}$$

$$\begin{matrix} \begin{matrix} \square \\ \square \\ \textcolor{green}{\square} \\ \square \\ \square \\ \square \end{matrix} & \hat{Y} \\ \end{matrix}$$

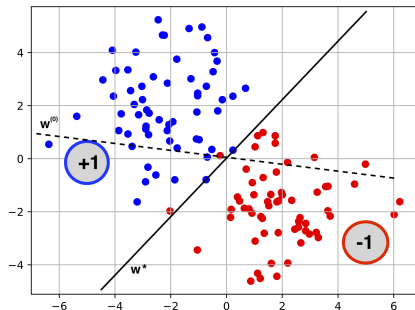
$$\begin{matrix} \begin{matrix} \square \\ \square \\ \square \\ \square \\ \square \\ \square \end{matrix} & Y \\ \end{matrix}$$





Passage en deux dimensions... Fonction de coût

Données 2D



$$d = 2$$

$$\mathbf{x} = [x_1, x_2]$$

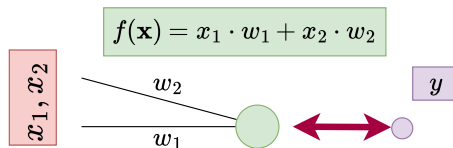
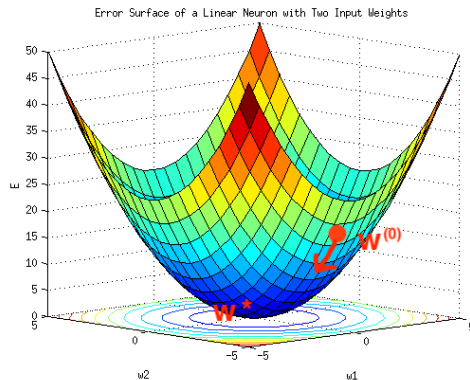
$$N$$

$$\hat{Y}$$

$$Y$$

X $\hat{Y}$ Y

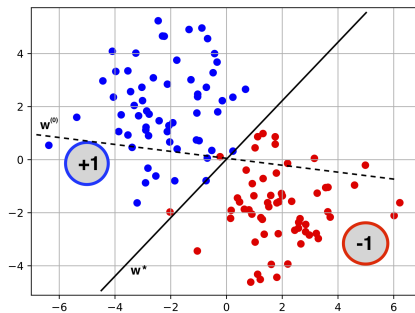
Forme de la fonction de coût $f(\mathbf{x})$



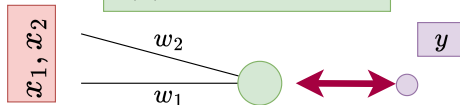


Passage en deux dimensions... Fonction de coût

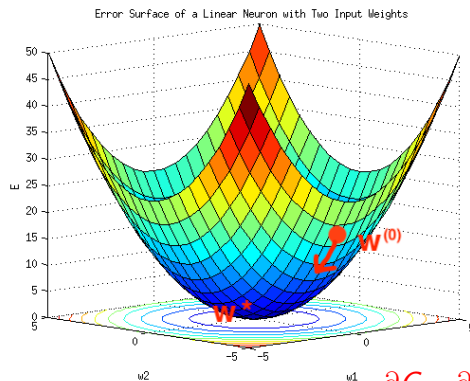
Données 2D



$$f(\mathbf{x}) = x_1 \cdot w_1 + x_2 \cdot w_2$$



Forme de la fonction de coût $f(\mathbf{x})$



Il faut calculer 2 dérivées : $\frac{\partial C}{\partial w_1}$, $\frac{\partial C}{\partial w_2}$

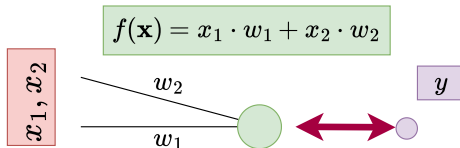
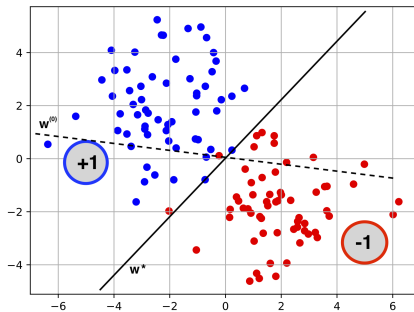
⇒ Vecteur 2D des dérivées =
direction dans l'espace des $\mathbf{w}$

Ce vecteur s'appelle le **gradient**

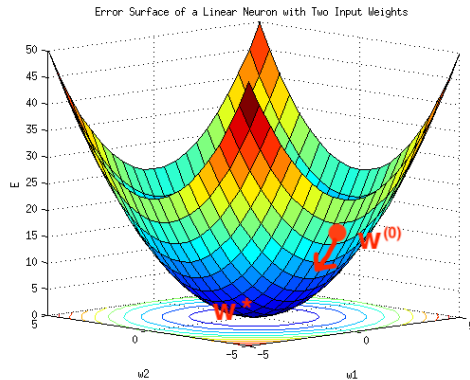


Passage en deux dimensions... Fonction de coût

Données 2D



Forme de la fonction de coût $f(\mathbf{x})$



$$f(\mathbf{x}) = w_1 x_1 + w_2 x_2$$

$$\mathcal{C}(f(\mathbf{x}), y) = (f(\mathbf{x}) - y)^2$$

$$\frac{\partial \mathcal{C}}{\partial w_1} = 2x_1(f(\mathbf{x}) - y), \quad \frac{\partial \mathcal{C}}{\partial w_2} = 2x_2(f(\mathbf{x}) - y)$$



Passage en deux dimensions... Fonction de coût

Algorithme **Descente de gradient**

Tant que:

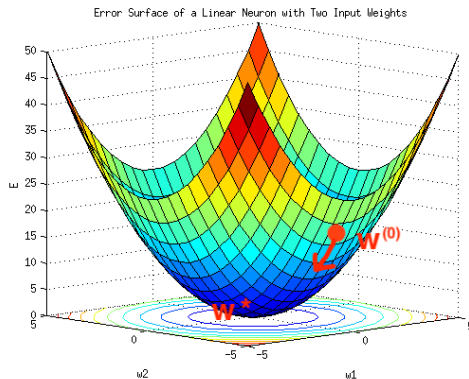
$$\left| \frac{\partial C(f(x), y)}{\partial w_j} \right| > \text{seuil}$$

$$w_1 \leftarrow w_1 - \varepsilon \frac{\partial C(f(x), y)}{\partial w_1}$$

$$w_2 \leftarrow w_2 - \varepsilon \frac{\partial C(f(x), y)}{\partial w_2}$$

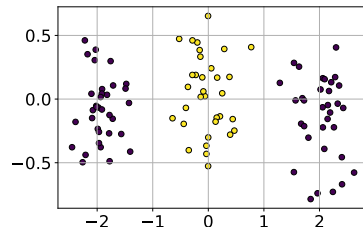
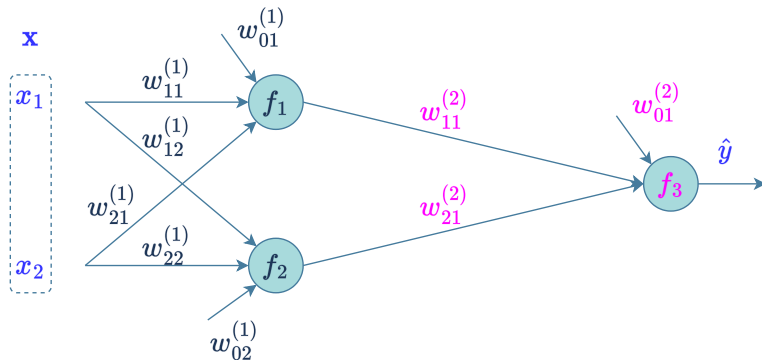
■ ε : Longueur du pas de descente

Forme de la fonction de coût $f(x)$





Combinons deux neurones



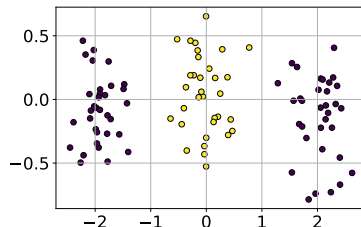
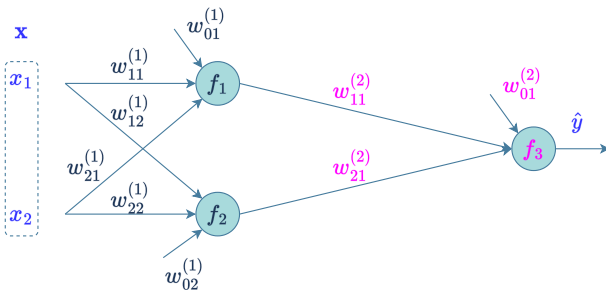
$$f_1(\mathbf{x}) = w_{11}^{(1)}x_1 + w_{21}^{(1)}x_2 + w_{01}^{(1)}, \quad f_2(\mathbf{x}) = w_{12}^{(1)}x_1 + w_{22}^{(1)}x_2 + w_{02}^{(1)}$$

$$f_3(\mathbf{x}) = w_{11}^{(2)}f_1(\mathbf{x}) + w_{21}^{(2)}f_2(\mathbf{x}) + w_{01}^{(2)}$$

Combiner des neurones $\Rightarrow$ suffisant ?



Combinons deux neurones



$$f_1(\mathbf{x}) = w_{11}^{(1)} x_1 + w_{21}^{(1)} x_2 + w_{01}^{(1)}, \quad f_2(\mathbf{x}) = w_{12}^{(1)} x_1 + w_{22}^{(1)} x_2 + w_{02}^{(1)}$$

$$f_3(\mathbf{x}) = w_{11}^{(2)} f_1(\mathbf{x}) + w_{21}^{(2)} f_2(\mathbf{x}) + w_{01}^{(2)}$$

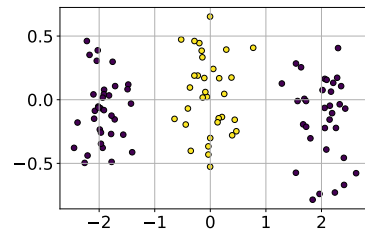
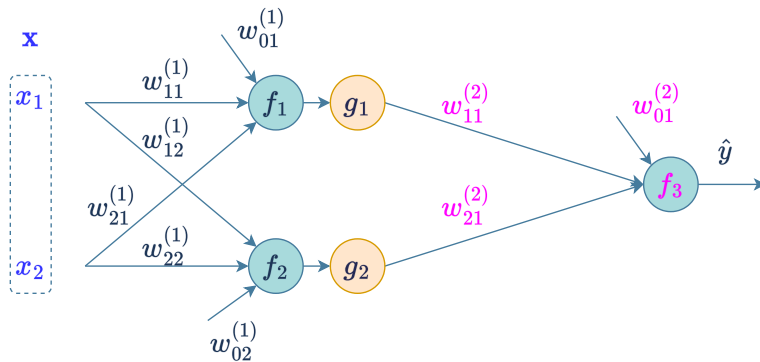
$$f_3(\mathbf{x}) = w_{11}^{(2)} (w_{11}^{(1)} x_1 + w_{21}^{(1)} x_2 + w_{01}^{(1)}) + w_{21}^{(2)} (w_{12}^{(1)} x_1 + w_{22}^{(1)} x_2 + w_{02}^{(1)}) + w_{01}^{(2)}$$

$$\Leftrightarrow f_3(\mathbf{x}) = x_1 (w_{11}^{(2)} w_{11}^{(1)} + w_{21}^{(2)} w_{12}^{(1)}) + x_2 (w_{11}^{(2)} w_{21}^{(1)} + w_{21}^{(2)} w_{22}^{(1)}) + w_{01}^{(2)} + w_{11}^{(2)} w_{01}^{(1)} + w_{21}^{(2)} w_{02}^{(1)}$$

Non ! il faut introduire de la non linéarité, sinon équivalent à un perceptron ...



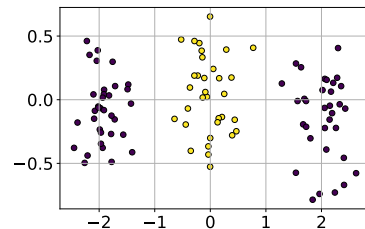
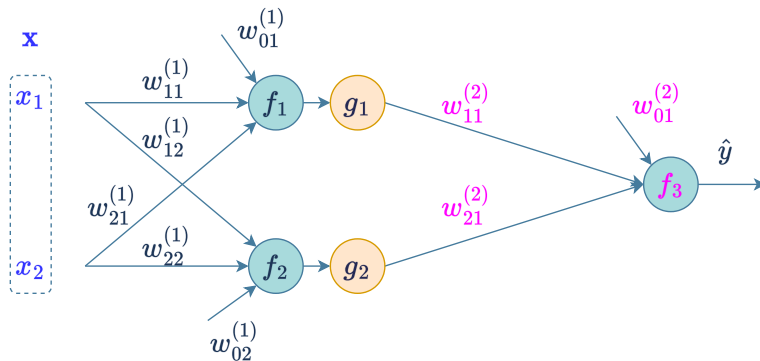
Non-linéarité



■ Quelle non-linéarité ?



Non-linéarité



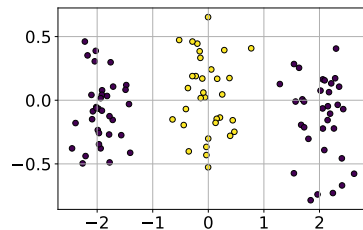
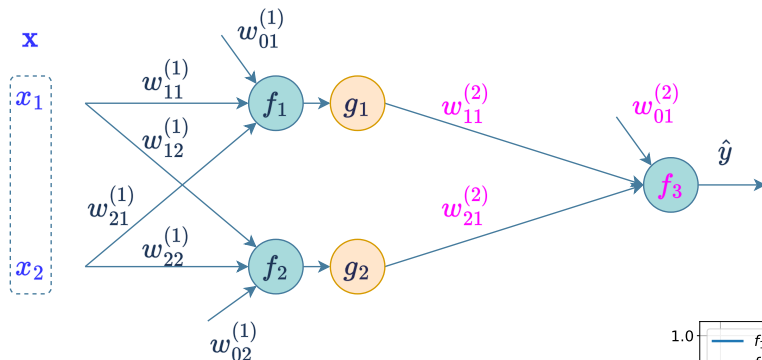
■ Quelle non-linéarité ?

■ Fonction *signe* ?

⇒ dérivée problématique ...



Non-linéarité



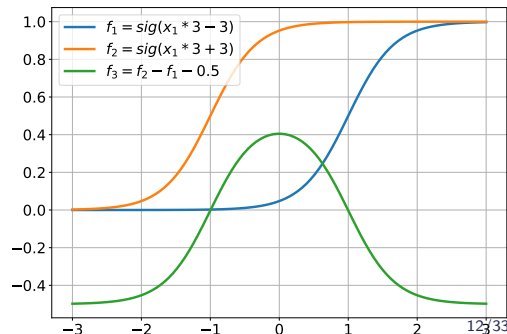
■ Quelle non-linéarité ?

■ Fonction *signe* ?

⇒ dérivée problématique ...

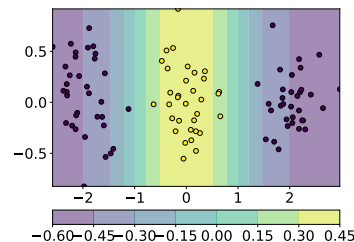
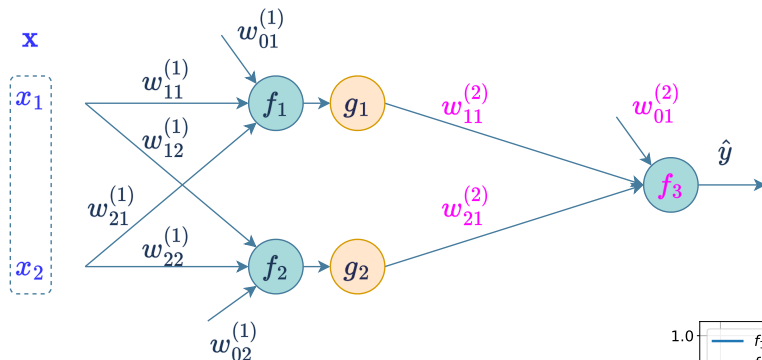
■ Fonctions *tanh*, *sigmoïde*, ... + biais

$$g(x) = \frac{1}{1 + \exp(-x)}$$





Non-linéarité

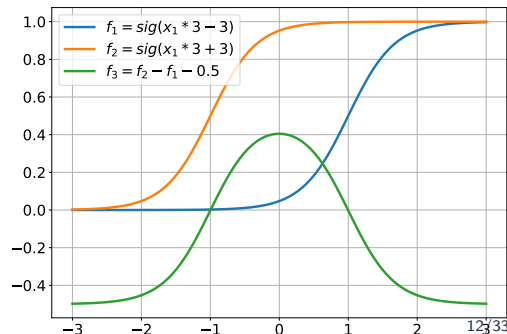


■ Quelle non-linéarité ?

■ Fonction *signe* ?

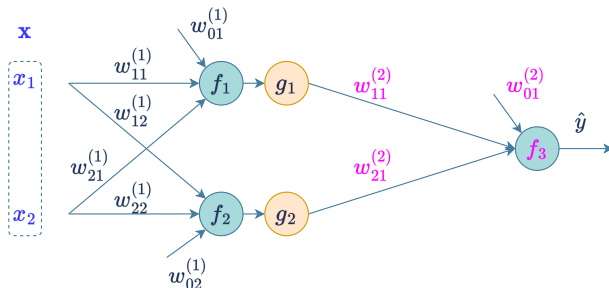
⇒ dérivée problématique ...

$$g(x) = \frac{1}{1 + \exp(-x)}$$





Calcul des dérivées



1 Initialisation aléatoire

2 Quelles dérivées? $\forall i, j, \frac{\partial C(f(\mathbf{x}), y)}{\partial w_{ij}}$

3 Comment les calculer ?

Forward: calcul de $f(\mathbf{x})$,

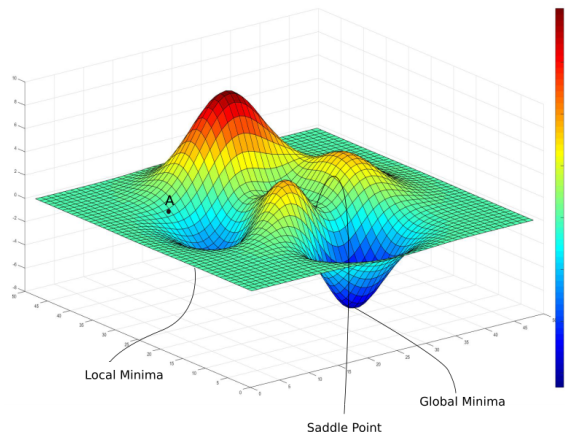
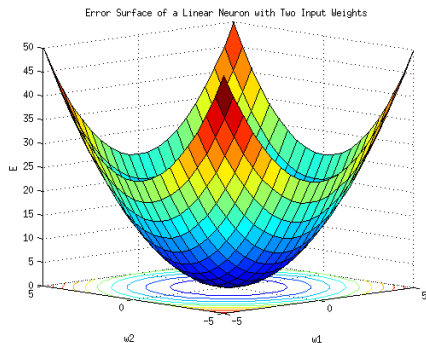
Backward: calcul des gradients

4 Comment les exploiter ? $\Rightarrow$ Optimisation: descente de gradient

$$w_{ij} \leftarrow w_{ij} - \underbrace{\varepsilon}_{\text{Pas de gradient}} \frac{\partial C(f(\mathbf{x}), y)}{\partial w_{ij}}$$



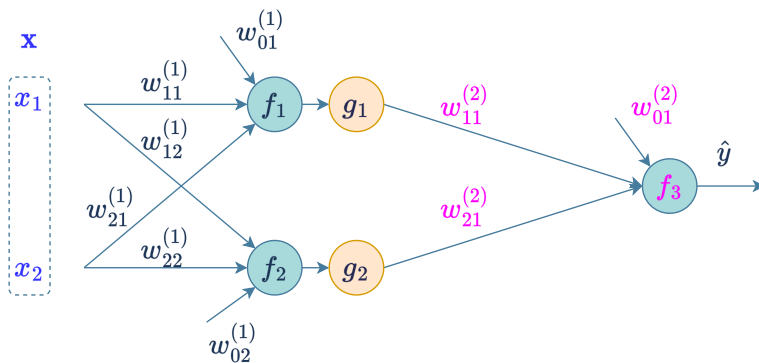
L'optimisation des réseaux de neurones non-linéaires



Le passage à des architectures complexes et non-linéaires implique la perte de la convexité du problème d'optimisation



Vers la retro-propagation du gradient



Les grandes étapes:

- 1 Initialisation aléatoire des poids w_{ij}
- 2 Inférence / *forward*: on met les données dans le modèle
- 3 Optimisation / *backward* : on corrige les poids

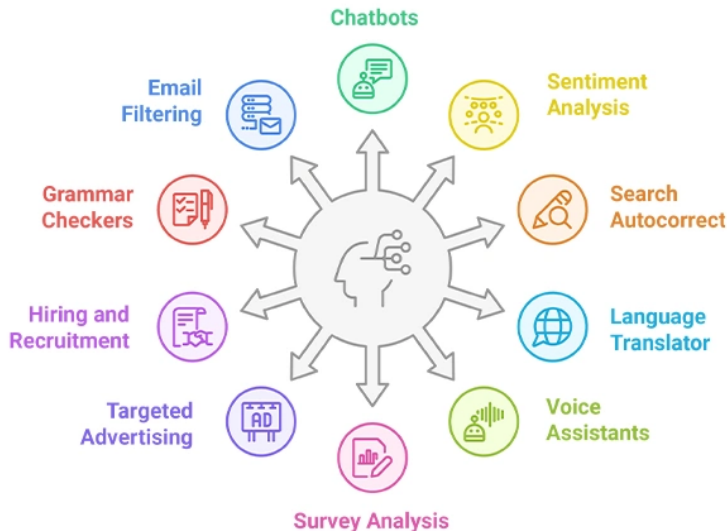
⇒ Démo dynamique

DES RÉSEAUX DE NEURONES À L'ANALYSE DU TEXTE



Texte + réseaux de neurones: Pourquoi? Comment?

- **Pourquoi** faire rentrer du texte dans des réseaux de neurones?



- **Comment** faire rentrer du texte dans des réseaux de neurones?



Texte + réseaux de neurones: Pourquoi? Comment?

- **Pourquoi** faire rentrer du texte dans des réseaux de neurones?
- **Comment** faire rentrer du texte dans des réseaux de neurones?

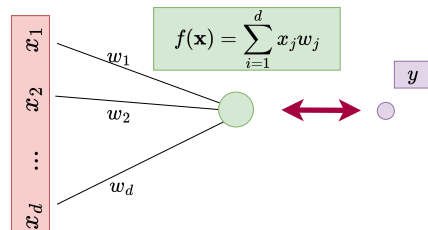
This new iPhone, what a marvel

An iPhone, What a scam!

Half the price is for the logo

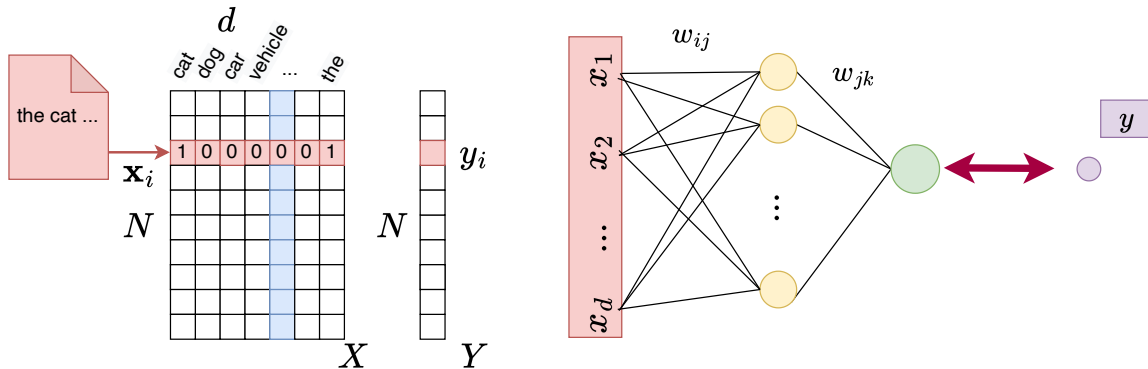
Apple once again proves that perfection can be sold

Comment rentrer
dans un
tableau?





Comment gérer les mots dans un réseau de neurones?



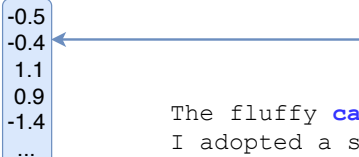
- Vision déstructurée des textes
- Un problème de gestion des synonymes
 - $\text{cat} \neq \text{car} \neq \text{vehicle}$



Apprendre les représentations de mots

Représentation/plongement (fr) = *Embedding* (eng)

-0.5
-0.4
1.1
0.9
-1.4
...



The fluffy **cat** napped lazily in the sunbeam.
I adopted a stray **cat** from the shelter last week.
My **cat** loves to chase after toy mice.
The black **cat** stealthily crept through the dark alley.
I often find my **cat** perched on the windowsill, watching birds.

Chaque mot devient un vecteur de nombres...

Les mots peuvent être proches ou lointains les uns des autres

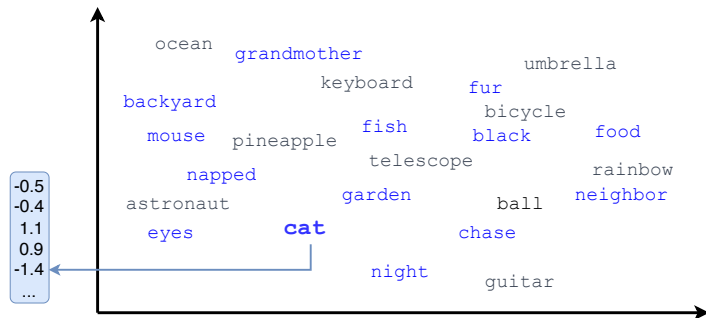


Apprendre les représentations de mots (2)

Toujours la même idée:

- 1 Initialisation aléatoire
- 2 Critère d'optimisation

The fluffy **cat** napped lazily in the sunbeam.
I adopted a stray **cat** from the shelter last week.
My **cat** loves to chase after toy mice.
The black **cat** stealthily crept through the dark alley.
I often find my **cat** perched on the windowsill, watching birds.





Apprendre les représentations de mots (3)

The fluffy **cat** napped lazily in the sunbeam.

I adopted a stray **cat** from the shelter last week.

My **cat** loves to chase after toy mice.

The black **cat** stealthily crept through the dark alley.

I often find my **cat** perched on the windowsill, watching birds.

fluffy

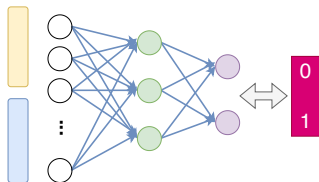
0.1
-1.3
-0.6
1.9
0.3
...

cat

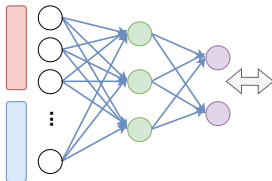
-0.5
-0.4
1.1
0.9
-1.4
...

vehicle

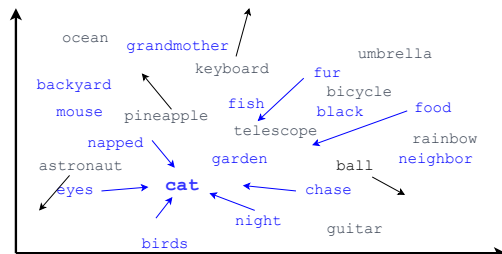
0.1
-1.3
-0.6
1.9
0.3
...



Same
Sentence

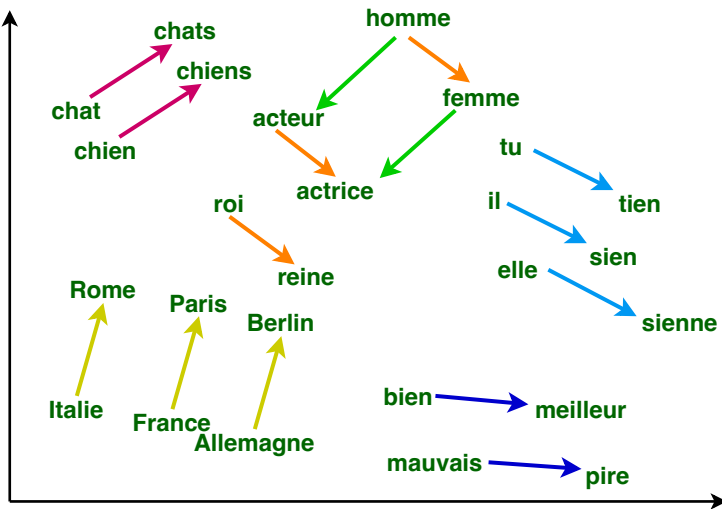


NOT
Same
Sentence





Représentation des mots: ce que l'on a appris



- Espace sémantique:
même signification
 $\Leftrightarrow$
même position
- Espace structuré:
régularités grammaticales,
connaissances élémentaires
...

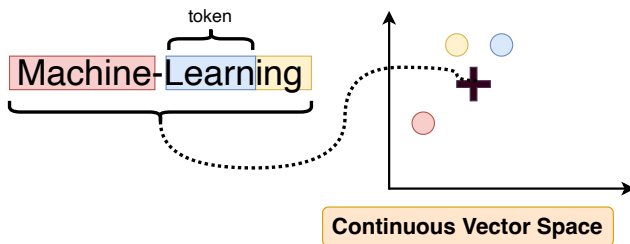
Distributed representations of words and phrases and their compositionality, [Mikolov et al. NeurIPS 2013](#)



Des mots aux tokens

Décomposition des mots en groupes de lettres fréquents...
= mots, radicaux, préfixes, suffixes, ...

Word Piece statistical split



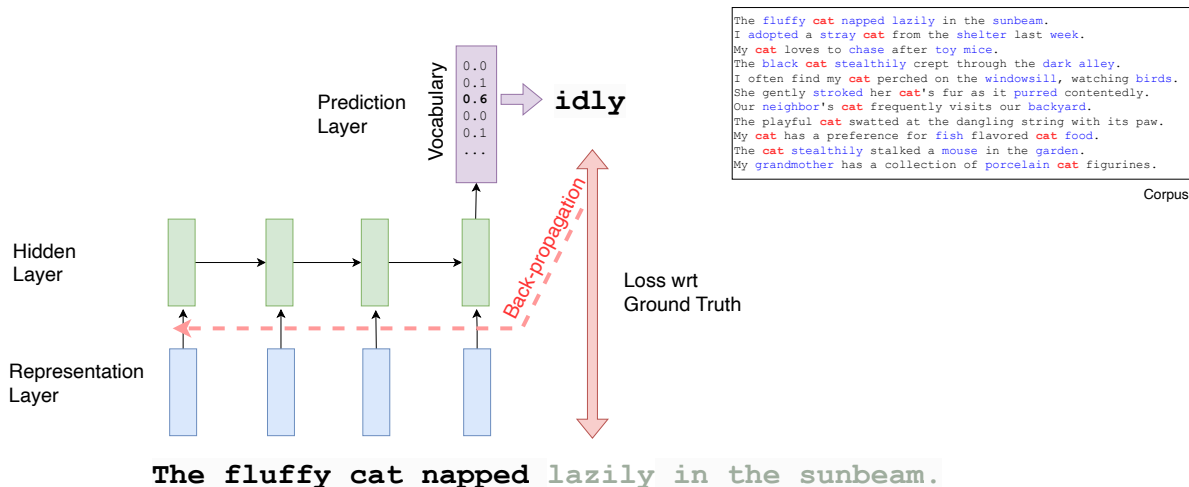
- Représentations des mots inconnus
- des mots techniques
- robustesse aux fautes de frappe

Enriching word vectors with subword information. [Bojanowski et al. TACL 2017.](#)



Agréger les représentations de mots: vers l'IA générative

■ Apprendre la génération et la représentation en même temps



CHATGPT

30 NOVEMBRE 2022

1 MILLION D'UTILISATEURS EN 5 JOURS

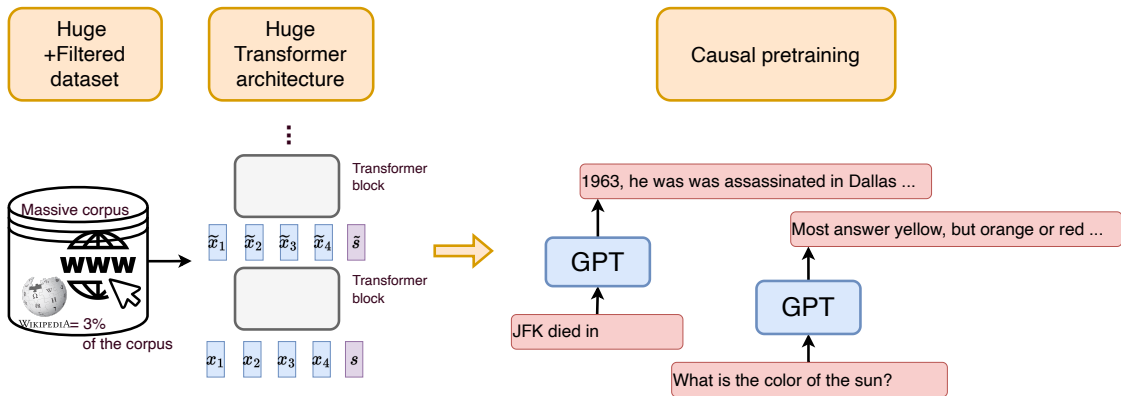
100 MILLIONS À LA FIN JANVIER 2023

1.16 MILLIARD EN MARS 2023



Les ingrédients de chatGPT

0. Transformer + données massives (GPT)



- Grammaire : accord singulier/pluriel, concordance des temps
- Connaissances : entités, nom, lieux, dates, ...



Les ingrédients de chatGPT

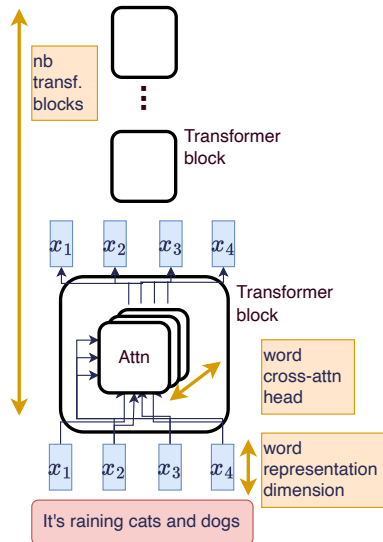
1. Plus, c'est mieux ! (GPT)

- + plus de mots en entrée [500 $\Rightarrow$ 2k, 32k, 100k]
- + plus de dim. dans l'espace des mots [500-2k $\Rightarrow$ 12k]
- + plus de têtes d'attention [12 $\Rightarrow$ 96]
- + plus de blocs/couches [5-12 $\Rightarrow$ 96]

175 milliards de paramètres...

Qu'est-ce que cela signifie ?

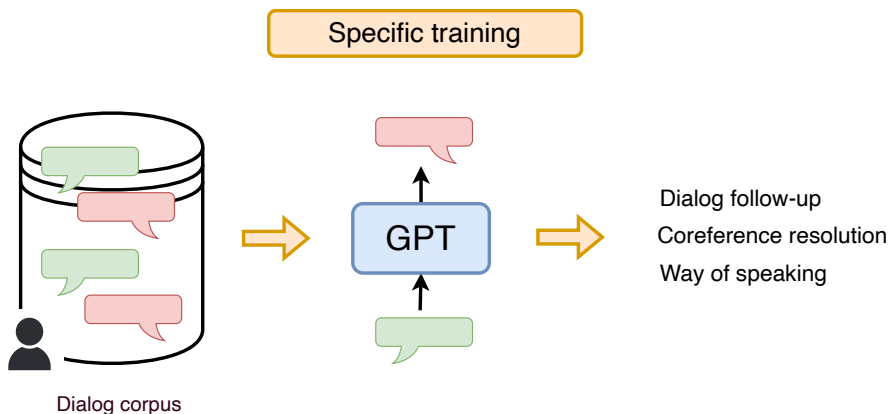
- $1,75 \cdot 10^{11} \Rightarrow 300 \text{ Go} + 100 \text{ Go}$ (stockage pour l'inférence) $\approx 400 \text{ Go}$
- GPU NVidia A100 = 80 Go de mémoire (=20k€)
- Coût de (1) entraînement : 4,6 millions €





Les ingrédients de chatGPT

2. Suivi du dialogue

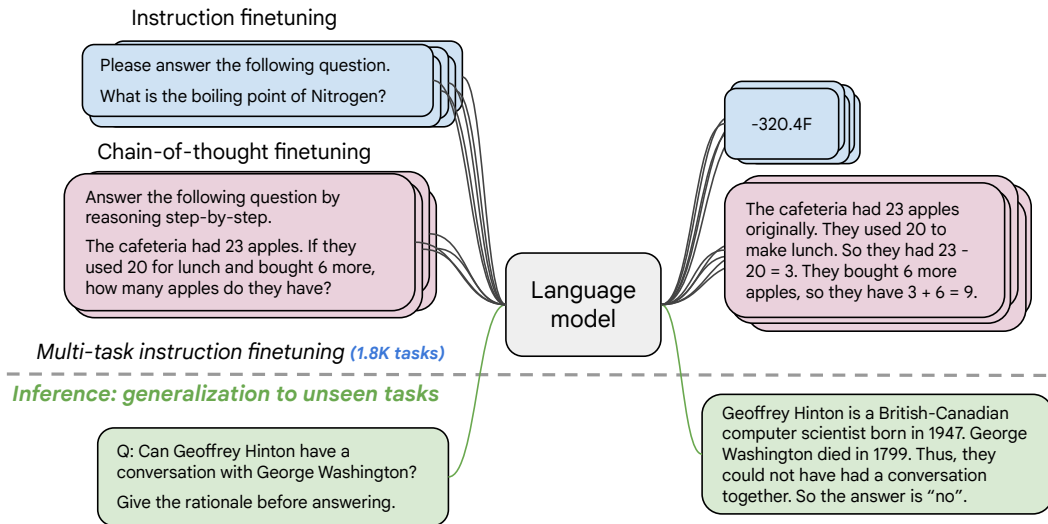


- **Données très propres** Données générées/validées/classées par des humains



Les ingrédients de chatGPT

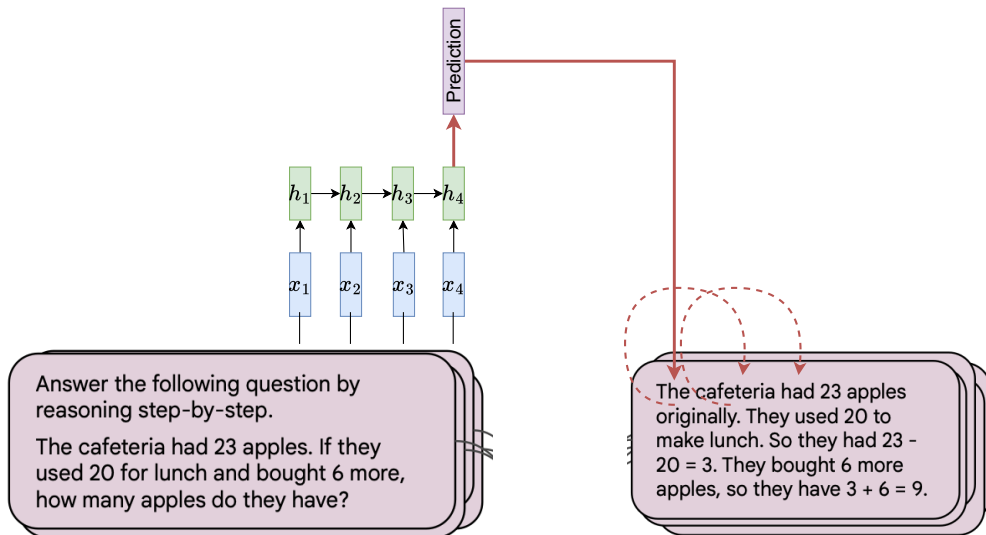
3. Ajustement fin sur des tâches de raisonnement ($\pm$) complexes





Les ingrédients de chatGPT

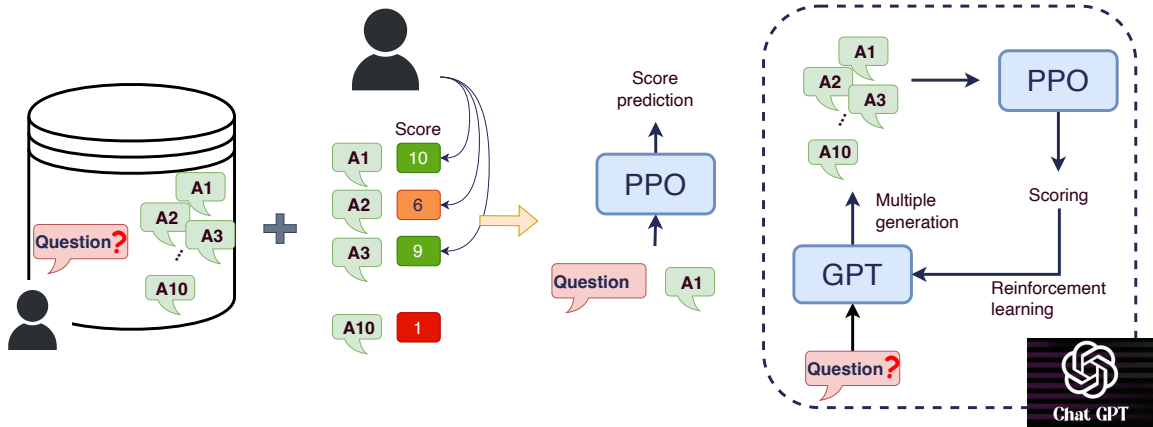
3. Ajustement fin sur des tâches de raisonnement ($\pm$) complexes





Les ingrédients de chatGPT

4. Instructions + classement des réponses



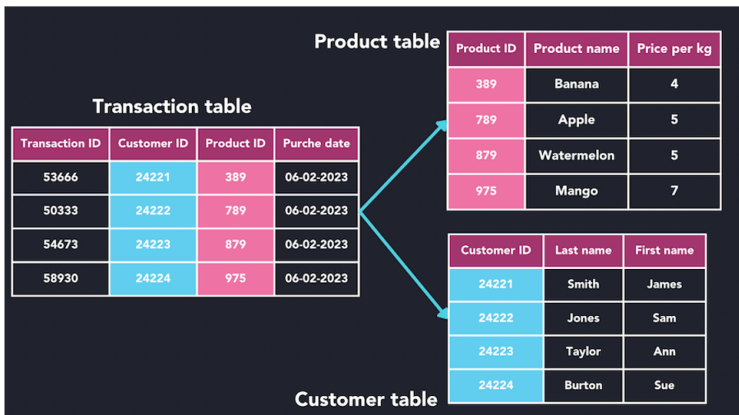
- Base de données créée par des humains
- Amélioration des réponses

- ... Aussi un moyen d'éviter les sujets sensibles = censure



Un nouveau rapport aux connaissances

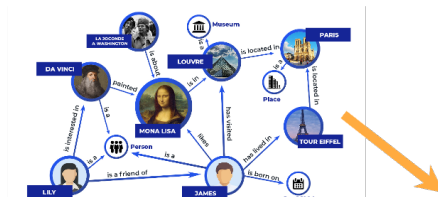
■ Bases de données + accès en SQL





Un nouveau rapport aux connaissances

- Bases de données + accès en SQL
- Knowledge Graphs



Barack Obama

44th President of the United States

In office
January 20, 2009 – January 20, 2017

Vice President Joe Biden

Preceded by George W. Bush

Succeeded by Donald Trump

United States Senator from Illinois

Google

quel est l'âge de barack obama

Tous Actualités Images Vidéos courtes Web Finance Livres Plus Outils

Barack Obama > Âge :

64 ans

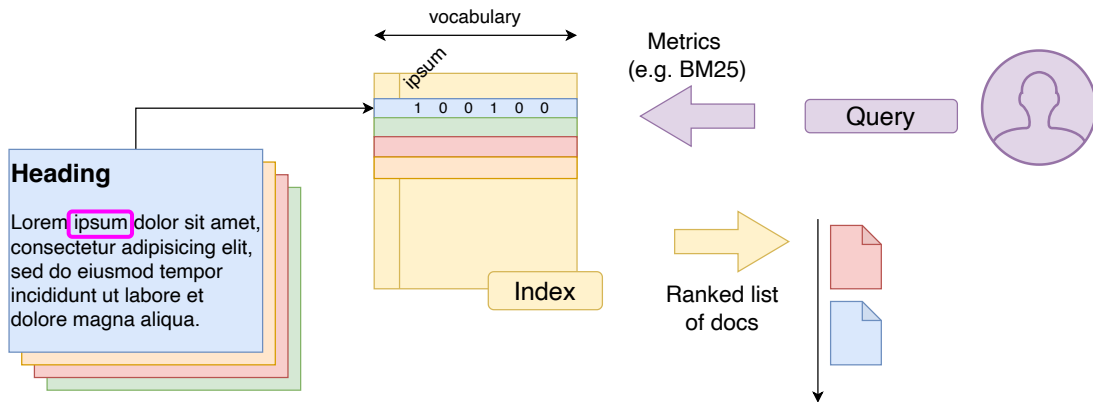
4 août 1961

Autres questions :



Un nouveau rapport aux connaissances

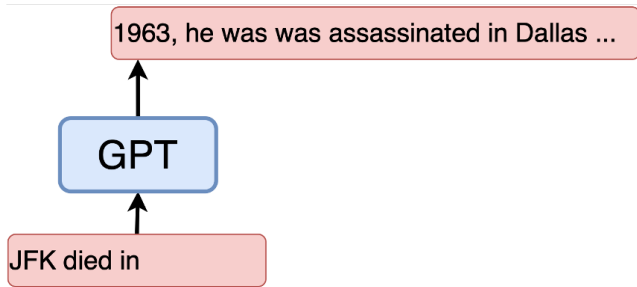
- Bases de données + accès en SQL
- Knowledge Graphs
- Indexation du web / moteurs de recherche





Un nouveau rapport aux connaissances

- Bases de données + accès en SQL
- Knowledge Graphs
- Indexation du web / moteurs de recherche
- ... Loi de probabilité !



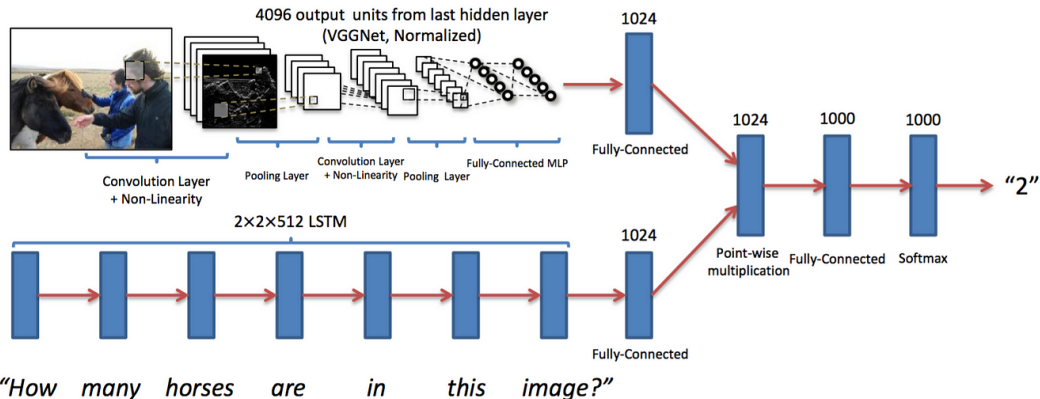
$$p(w_{t+1} | w_t, w_{t-1}, \dots, w_1)$$



GPT-4 & Multimodalité

Fusionner info. texte + image. **Apprendre** l'information conjointement

Exemple du VQA : Visual Question Answering (questions visuelles)



⇒ Rétropropager l'erreur ⇒ modifier les repr. des mots + l'analyse d'image

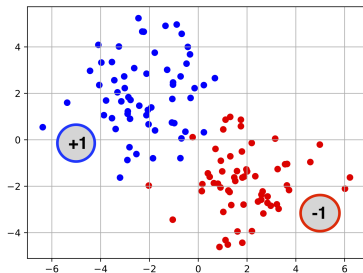


VQA : Visual Question Answering, arXiv, 2016 , A. Agrawal et al.

CALCUL MATRICIEL ET PARALLÉLISATION: L'ESSOR DES GPU

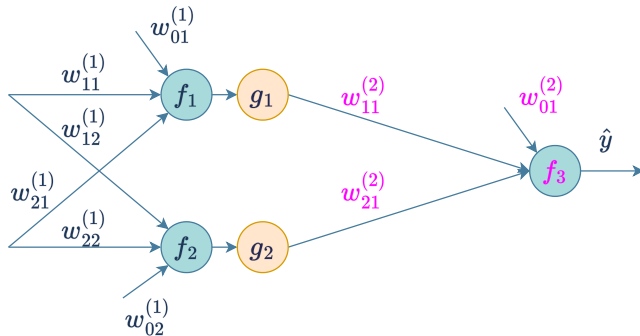


Quels sont les calculs effectués dans un réseau de neurones?



$\mathbf{x}$

x_1
 x_2

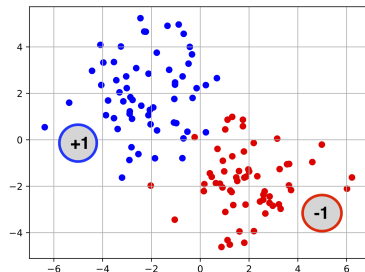
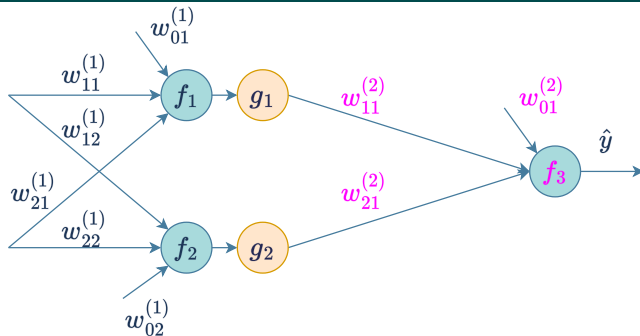


Besoin de calcul:

$$f_1(\mathbf{x}) = w_{11}^{(1)} x_1 + w_{21}^{(1)} x_2 + w_{01}^{(1)}, \quad f_2(\mathbf{x}) = w_{12}^{(1)} x_1 + w_{22}^{(1)} x_2 + w_{02}^{(1)}$$

$$f_3(\mathbf{x}) = w_{11}^{(2)} g_1(f_1(\mathbf{x})) + w_{21}^{(2)} g_2(f_2(\mathbf{x})) + w_{01}^{(2)}$$

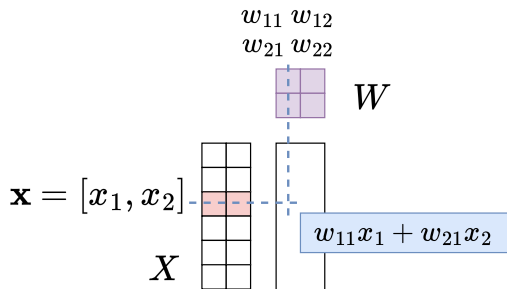
Quels sont les calculs effectués dans un réseau de neurones?


 $\mathbf{x}$
 x_1
 x_2


$\Rightarrow$ Calcul des valeurs de f_1 et f_2 pour tous les points de X !

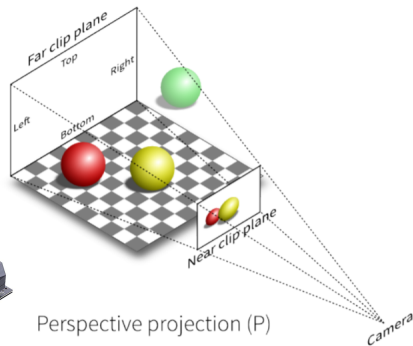
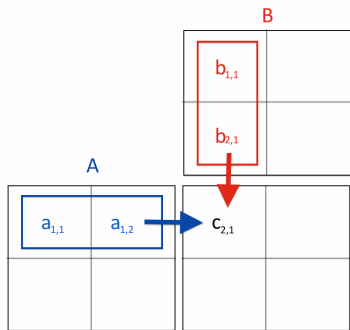
Ensemble de produit scalaire...

... = produit matriciel





Parallélisation des calculs matriciels / GPU



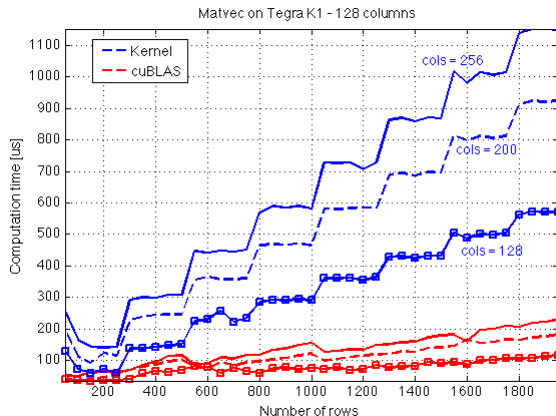
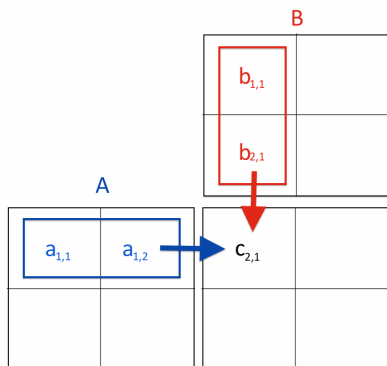
Perspective projection (P)

La carte graphique de l'ordinateur peut

- Calculer la projection sur l'écran d'une scène 3D...
- ... Et faire 1000 petits produits scalaires en parallèle !



Parallélisation des calculs matriciels / GPU



CONCLUSION



Pourquoi autant de controverse ?

- Nouvel outil [Décembre 2022]
- + Vitesse d'adoption sans précédent [1M d'utilisateurs en 5 jours]
- Forces et faiblesses... mal comprises par les utilisateurs
 - Gains de productivité significatifs
 - Usages surprenants / parfois absurdes
 - Biais / usages dangereux / risques
- Retours d'expérience mal interprétés
 - Anthropomorphisation de l'algorithme et de ses erreurs
- Quel modèle économique, pédagogique, écologique et sociétal ?

