

# Analyse des données

Redha Moulla

12-15 décembre 2023

# Plan de la présentation



Introduction au machine learning



Machine learning : apprentissage supervisé



Machine learning : apprentissage non supervisé



Éléments de deep learning



Analyse des données textuelles

Qu'est-ce que l'intelligence  
artificielle ?

# Qu'est-ce que l'intelligence artificielle ?

1

## **Intelligence**

Ensemble des fonctions mentales ayant pour objet la connaissance conceptuelle et relationnelle.

*Larousse*

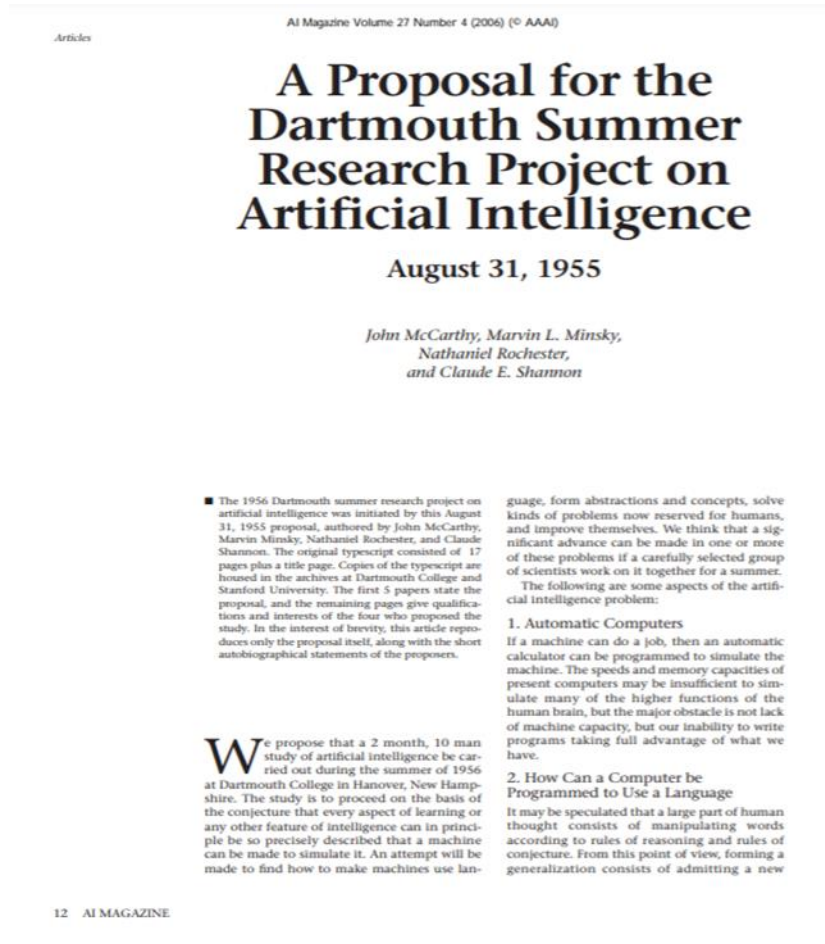
2

## **Artificielle**

Qui est produit de l'activité humaine (opposé à la nature).

*Larousse*

# Conférence de Dartmouth 1956



*“We propose that a 2-month, 10-man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.”*

# L'intelligence artificielle selon John McCarthy

“

*It is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable.*

”

John McCarthy

# L'intelligence artificielle selon Alan Turing

## Test de Turing

Le test de Turing est un test d'intelligence artificielle. Il consiste à évaluer la capacité d'une machine à imiter une conversation humaine.

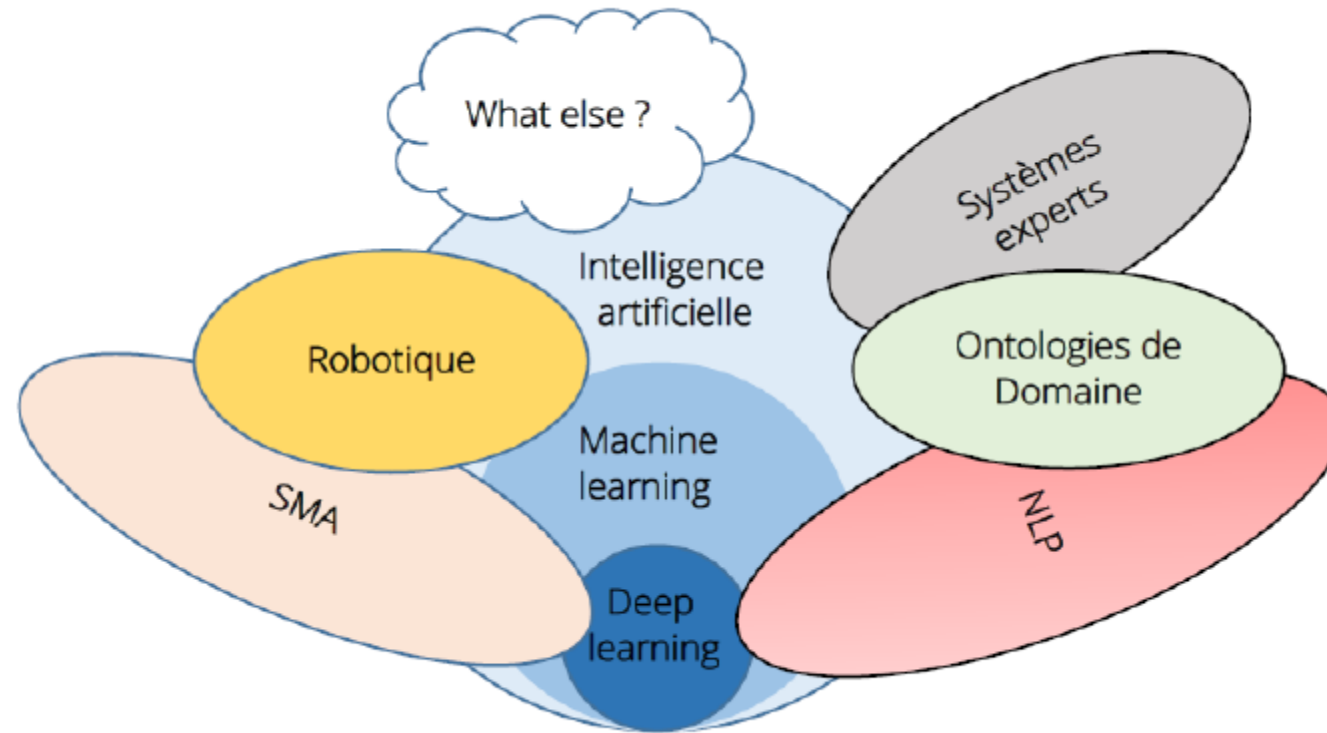
# L'intelligence artificielle selon Luc Julia

“*L'intelligence artificielle n'existe pas.*”

Luc Julia



# L'intelligence artificielle : une diversité d'approches



# IA connexionniste vs IA symbolique

Intelligence artificielle  
symbolique

Intelligence artificielle  
connexionniste

Logique



Probabiliste

Ensemble de règles



Apprentissage machine

Orientée connaissance  
*Knowledge driven*



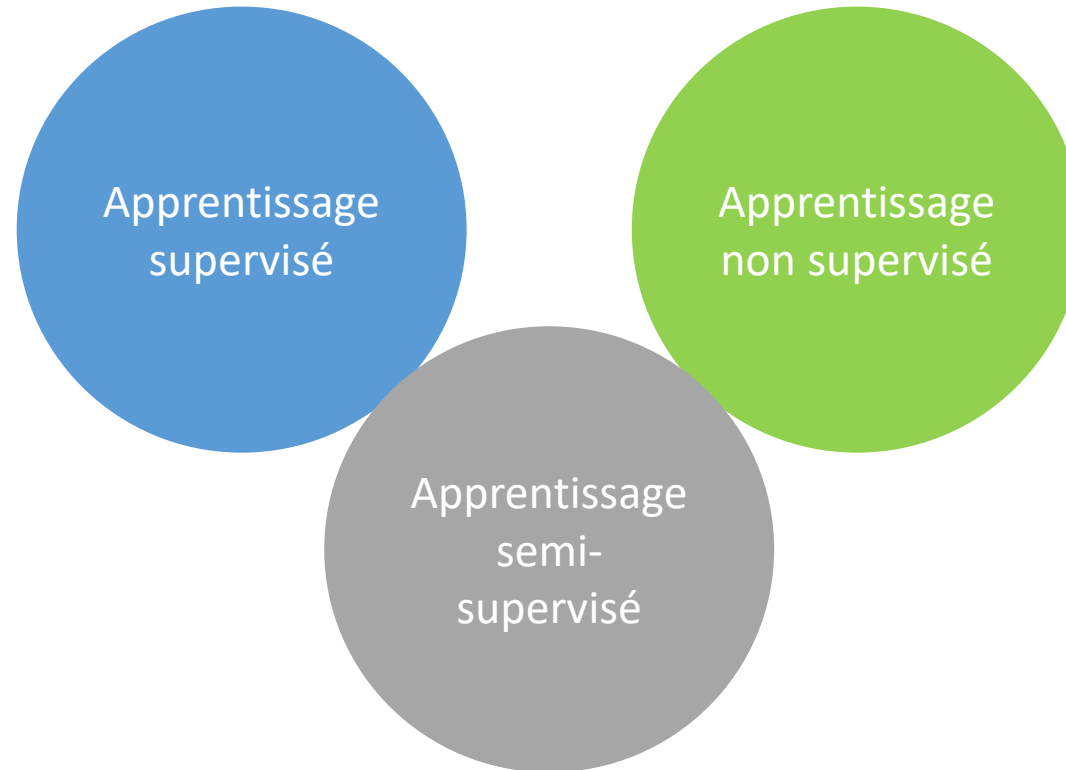
Orientée données  
*Data driven*

# Machine learning

## Introduction

# Définition de l'apprentissage machine

Le machine learning permet à la machine d'apprendre à partir des données un modèle de décision implicite (qui n'est pas nécessairement explicité pour l'humain).



# L'apprentissage supervisé

L'apprentissage supervisé consiste à apprendre un modèle qui associe une étiquette (label) à un ensemble de caractéristiques (features).

- Inputs : un jeu de données **annotées** pour entraîner le modèle.  
Exemple : des textes (tweets, etc.) avec les sentiment associés, positifs ou négatifs.
- Output : une étiquette pour un point de donnée inconnu par le modèle.

L'apprentissage supervisé se décline lui-même en deux grandes familles :

- **La classification** : prédire une catégorie ou une classe.  
Exemple : prédire l'étiquette d'une image (chat, chien, etc.), le sentiment associé à un texte, le centre d'intérêt d'un client à partir de ses commentaires, etc.
- **La régression** : prédire une valeur continue (un nombre réel typiquement).  
Exemple : prédire le prix d'un appartement, la life time value d'un client, etc.

# Classification

## Entraînement



Chien



Chien



Chien



Chat

## Test



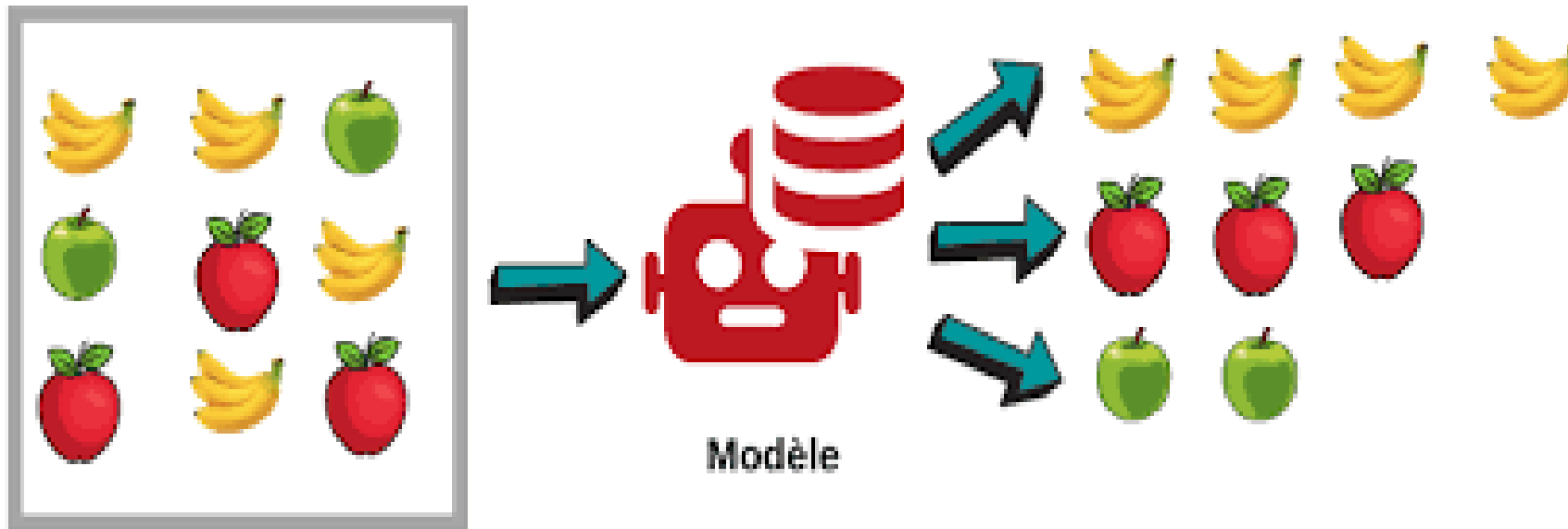
?



# Régression



# Apprentissage non supervisé





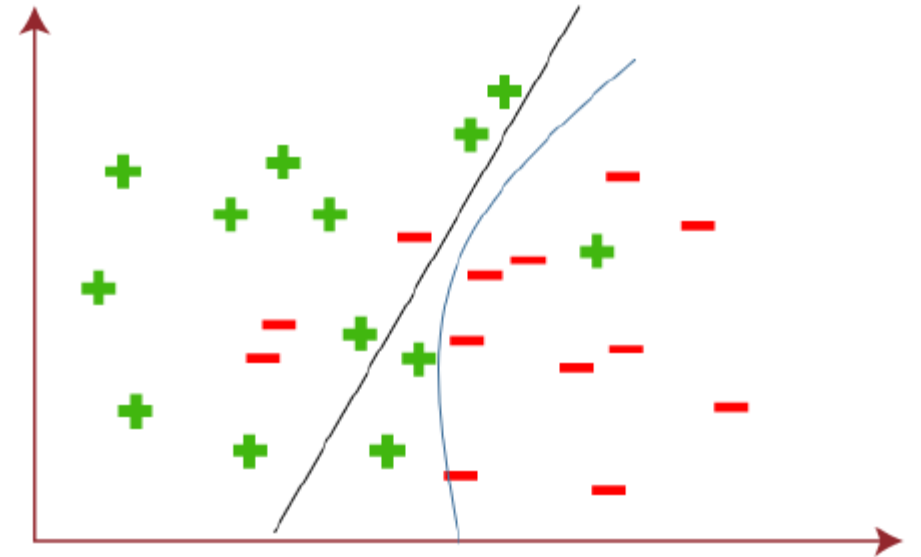
# Machine learning

Apprentissage supervisé

# Principes : espace des solutions

Comment choisir un modèle parmi tous les modèles possibles ?

- Formuler un a priori sur la classe à laquelle doit appartenir le modèle (expertise métier, contraintes imposées, etc.) ?
- Minimiser le risque empirique (l'erreur de prédiction sur les données d'entraînement)



# Principe : biais inductif

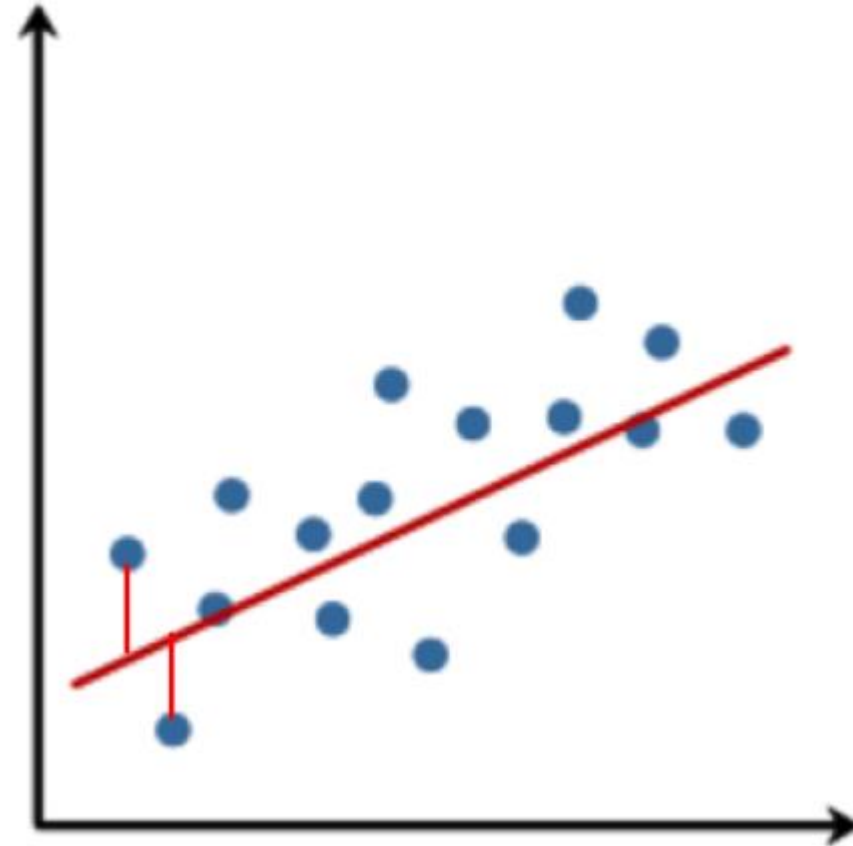
Le biais inductif est une connaissance ou une contrainte qui permet de réduire l'espace des solutions à une classe particulière de solutions (régression linéaire, random forest, etc.).

Exemples de biais inductifs :

- Linéarité de la relation entre la variable à prédire et les variables explicatives.
- Proche voisinage entre les données.
- Maximum de marge entre les classes.
- Localité dans les données.
- Invariance par translation des données.
- Etc.

## Principe : minimization du risque empirique

Une fois la classe des modèles possibles est déterminé, le meilleur modèle dans cette classe est a priori celui minimise l'erreur de prédiction sur les données d'entraînement (risque empirique).



# Sélection de modèle

Dans la pratique, pour sélectionner le modèle le plus pertinent par rapport à une métrique de performance donnée, on applique la méthodologie suivante :

- On partitionne le jeu de données disponible en trois parties : un jeu d'entraînement, un jeu de validation et un jeu de test.
- On entraîne un certain nombre de modèle sur le jeu de données d'entraînement.
- On évalue les performances respectives de ces modèles sur le jeu de données de validation et on sélectionne le modèle ayant la meilleure performance.
- On évalue la performance du modèle sélectionné sur le jeu de données de test. Il est important de noter ici que, idéalement, le jeu de données de test ne doit être utilisé qu'une seule fois.

# Quelques techniques de selection de modèle

Il existe plusieurs approches pour sélectionner un modèle. Les plus utilisées sont :

- La validation croisée k-fold.
- La validation croisée Leave-One-Out (LOO).
- Le bootstrap.
- Métriques spécifiques à certains modèles linéaires : R carré, AIC, BIC, etc.

# Métriques de performance

## Métriques associées à la régression

- L'erreur quadratique moyenne (MSE) : elle est définie comme la moyenne des carrées des écarts entre les prédictions et les valeurs observées.

$$MSE = \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2$$

- La racine carrée de l'erreur quadratique moyenne (RMSE) :

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2}$$

# Métriques de performance

## Métriques associées à la classification

- Précision : elle est définie comme la proportion des prédictions correctes parmi toutes les prédictions positives :

$$Précision = \frac{TP}{TP + FP}$$

Rappel (recall) : il représente la proportion des vrais positifs correctement Prédits par le modèle.

$$Rappel = \frac{TP}{TP + FN}$$

Score F1 (F1-score) : Le score F1 est défini comme la moyenne harmonique de la précision et du rappel.

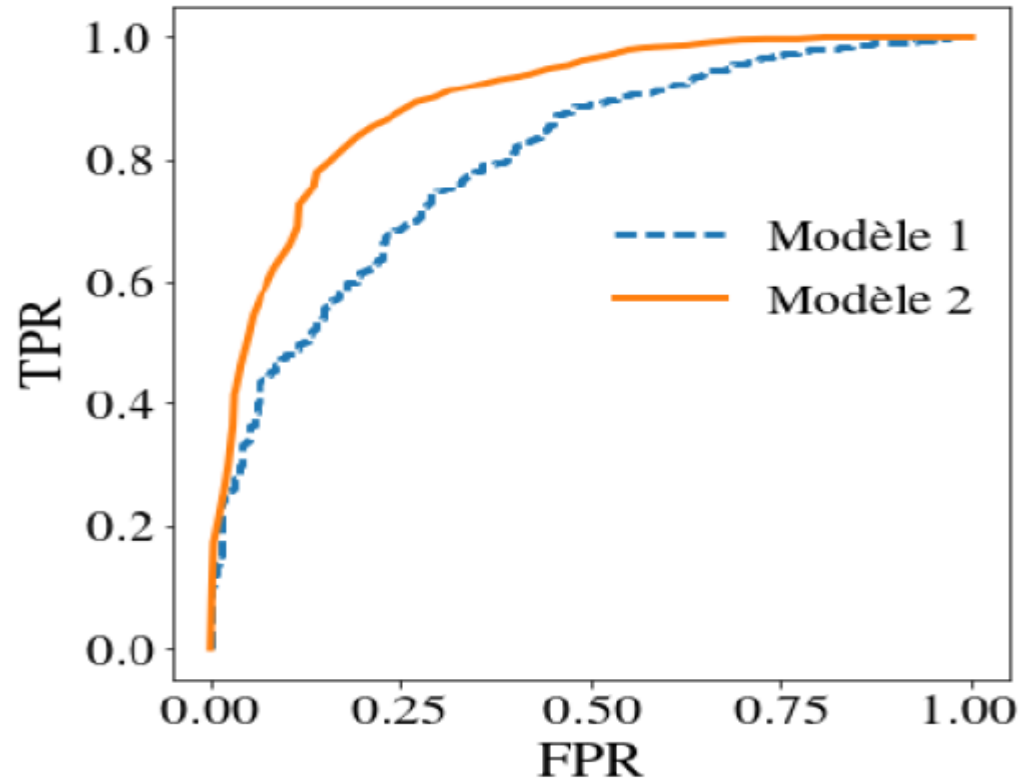
$$ScoreF1 = 2 \frac{Précision \times Rappel}{Précision + Rappel}$$



# Métriques de performance

## Métriques associées à la classification

Courbe ROC (Receiver-Operator Characteristic) : elle décrit l'évolution de la proportion des vrais positifs en fonction de celle des faux positifs.



# Machine learning

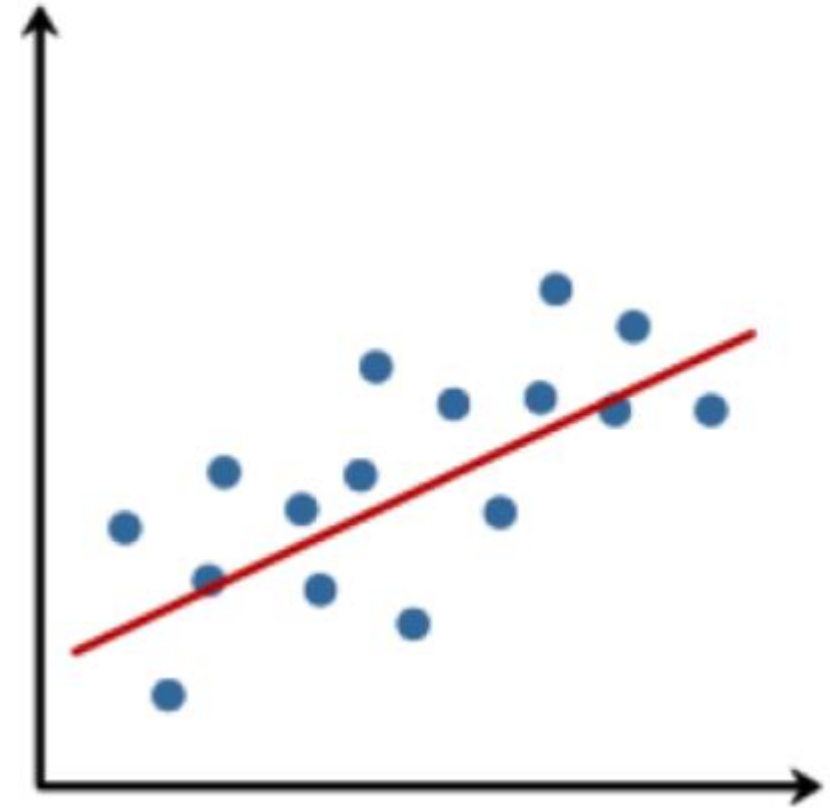
Apprentissage supervisé

# Régression linéaire simple

La régression linéaire essaie d'ajuster une droite aux données.

Dans le cas d'une seule variable explicative, l'équation de la droite est donnée par :

$$\hat{y} = \beta_0 + \beta_1 x$$



# Régression linéaire multiple

En présence de plusieurs variables explicatives, la régression linéaire s'écrit :

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

Les coefficients  $\beta_0, \beta_1, \dots, \beta_p$  sont déterminés par la méthode des moindres carrés.

$$\arg \min_{\beta_0, \beta_1, \dots, \beta_p} \sum_{i=1}^n \left( y^i - \left( \beta_0 + \sum_{j=1}^p \beta_j x_j^i \right) \right)^2$$

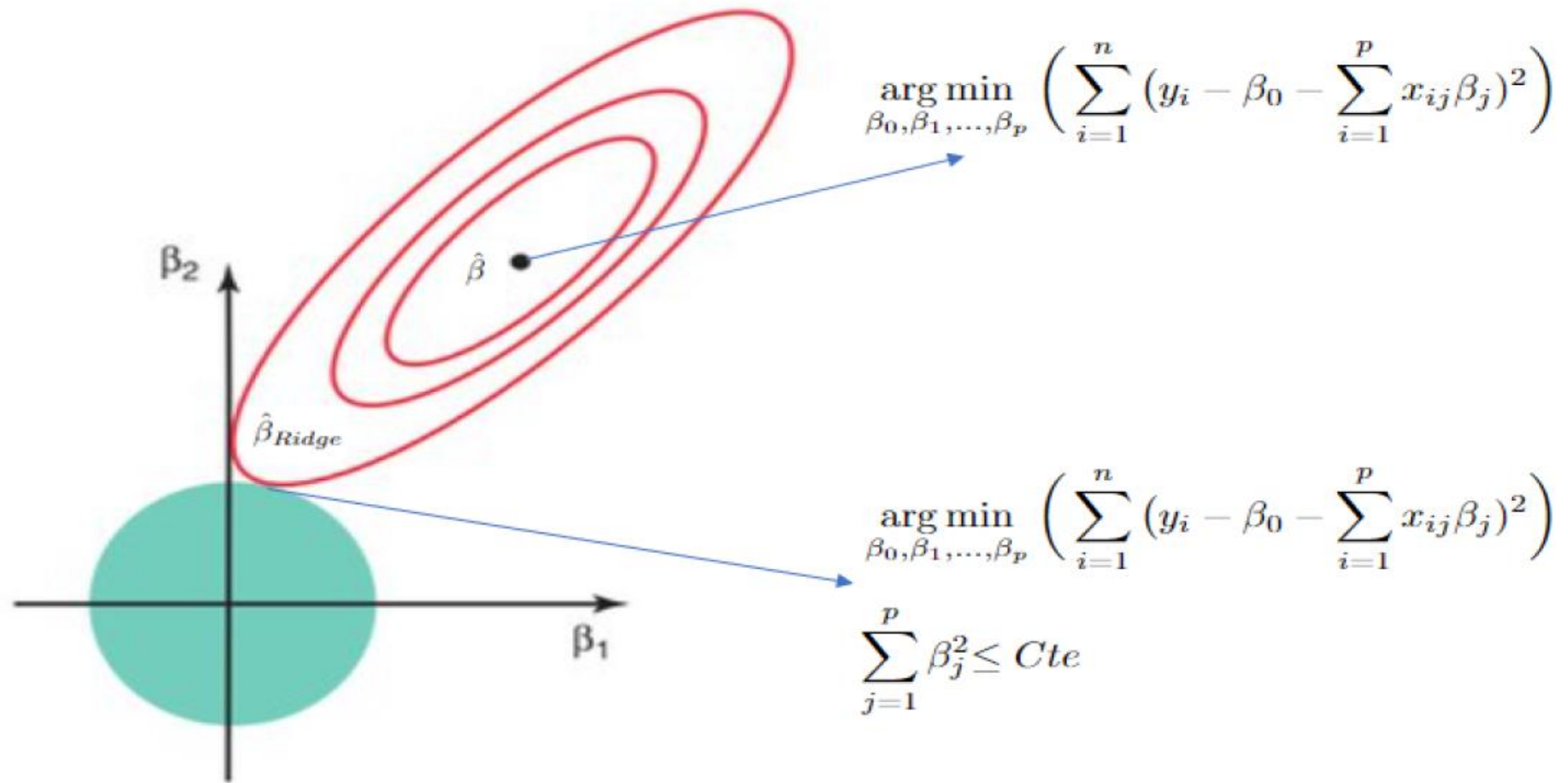
# Régression polynomiale

La régression linéaire peut prendre en compte les dépendances non linéaires entre les variables explicatives et la variable expliquée.

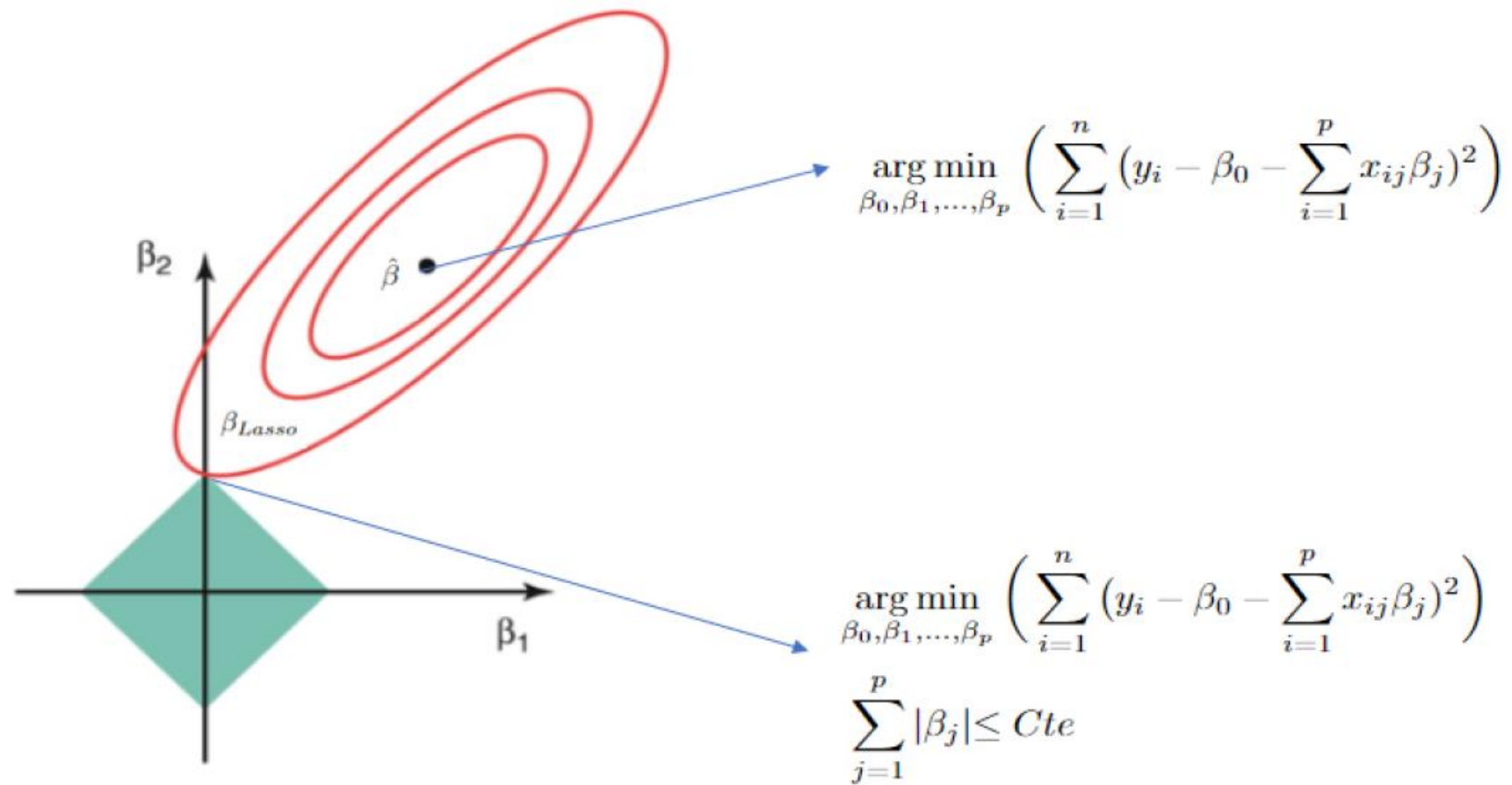
$$\hat{y} = \beta_{00} + \sum_{j=1}^p \beta_{1j} x_j + \sum_{j=1}^p \beta_{2j} x_j^2 + \dots + \sum_{j=1}^p \beta_{dj} x_j^d$$

Les coefficients  $\beta_{ij}$  peuvent être de la même manière, avec la méthode des moindres carrés.

# Régression Ridge



# Régression Lasso



# Régression linéaire

## Avantages :

- Modèle simple et facile à expliquer.
- Le risque de surapprentissage est relativement limité.
- Disponibilité de métriques de diagnostic permettant d'évaluer la qualité de l'ajustement.

## Inconvénients :

- Modèle trop simple si les données sont complexes.
- Très peu robuste pour les données déséquilibrées.



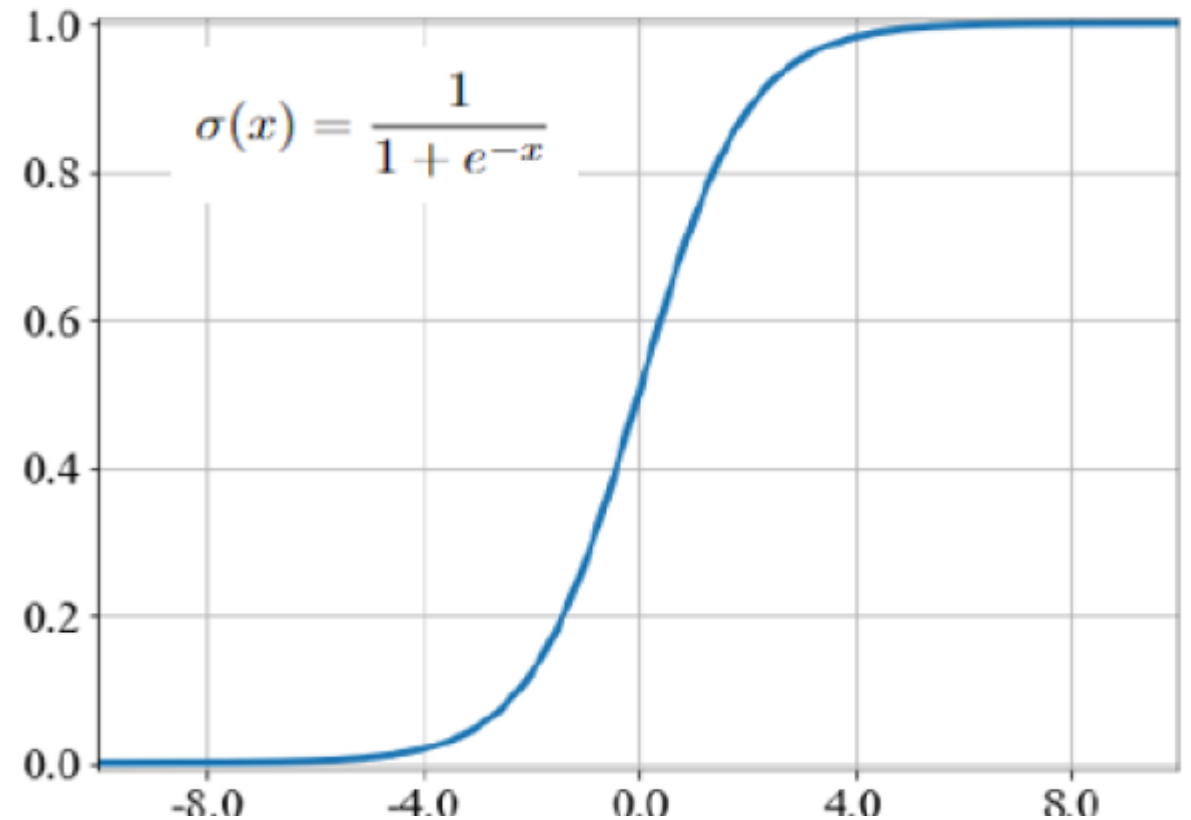
# Régression logistique

La régression essaie de séparer les données d'une manière linéaire.  
Elle s'écrit de la manière suivante :

$$\ln \frac{p(\vec{x})}{1 - p(\vec{x})} = \beta_0 + \vec{\beta} \vec{x}$$

Après quelques simplifications :

$$p(\vec{x}) = \frac{1}{1 + e^{-(\beta_0 + \vec{\beta}^T \vec{x})}}$$



# Régression logistique

Avantages :

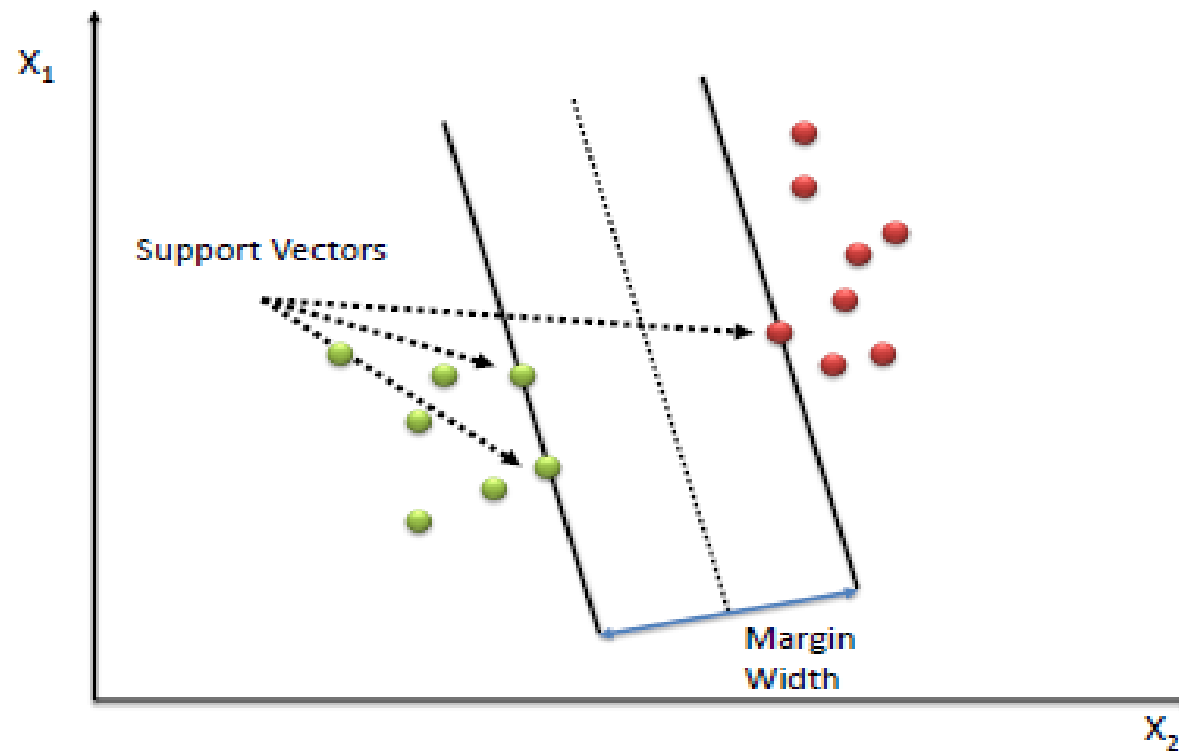
- Modèle simple et facile à expliquer.
- Le risque de surapprentissage est relativement limité.

Inconvénients :

- Modèle trop simple si les données sont complexes.
- Très peu robuste pour les données déséquilibrées.

# Machines à vecteurs de support (SVM)

Les SVMs tentent de séparer les données en construisant la surface de séparation ayant la plus vaste marge entre deux classes.



# Machines à vecteurs de support (SVM)

## Avantages :

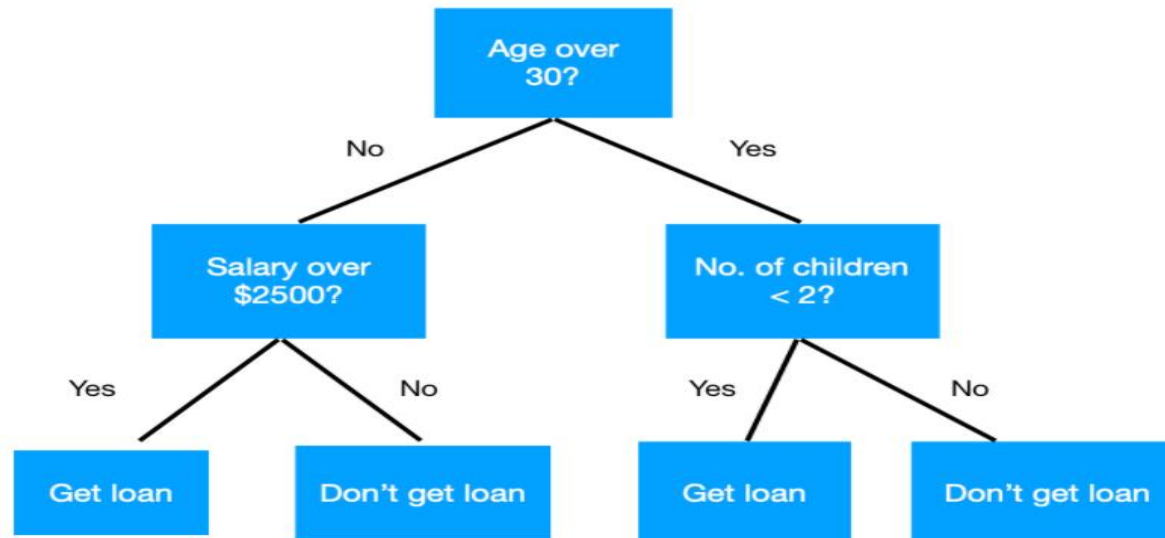
- Modèle assez robuste en grande dimension.
- Modèle nativement régularisé.
- Modèle paramétrable avec une variété de noyaux pour les cas non linéaires.

## Inconvénients :

- L'entraînement est relativement coûteux en termes de temps.
- Très peu robustes pour les données déséquilibrées.

# Arbres de décision

Les arbres de décisions essaient de séparer les données en se basant sur un ensemble de règles inférées.



# Arbres de décision

Avantages :

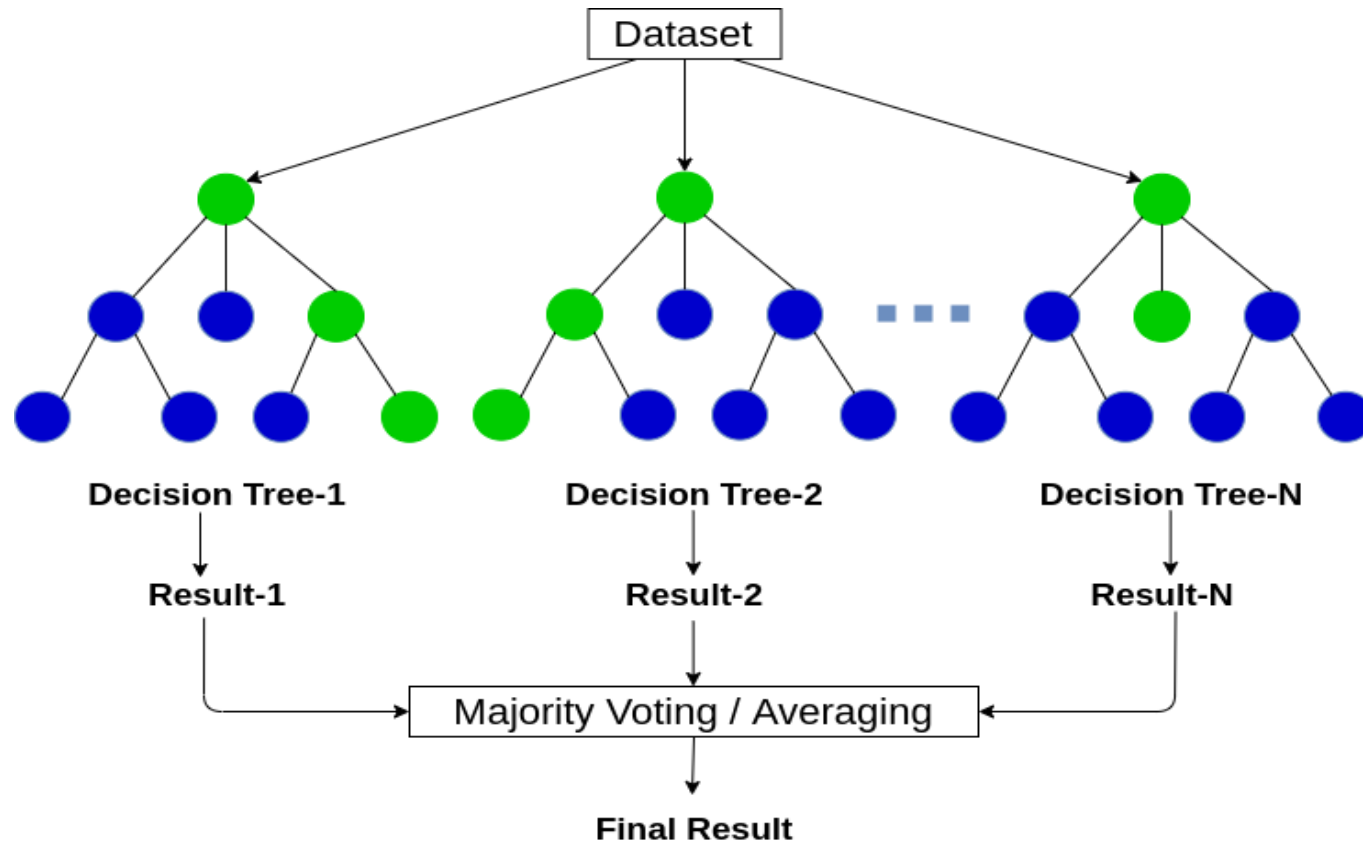
- Modèle simple et explicable.
- Modèle flexible et paramétrable.

Inconvénients :

- Modèle trop simple si les règles sous-jacentes sont complexes.
- Risque de surapprentissage relativement élevé.

# Random forest

Random forest est une technique ensembliste qui fait appel à plusieurs arbre de décisions simultanément. Le résultat final est obtenu à l'aide d'un « vote » de tous les arbres.



# Random forest

## Avantages :

- Modèle très efficace pour les données tabulaires.
- Modèle flexible et paramétrable.
- Robustesse et variance relativement limitée.

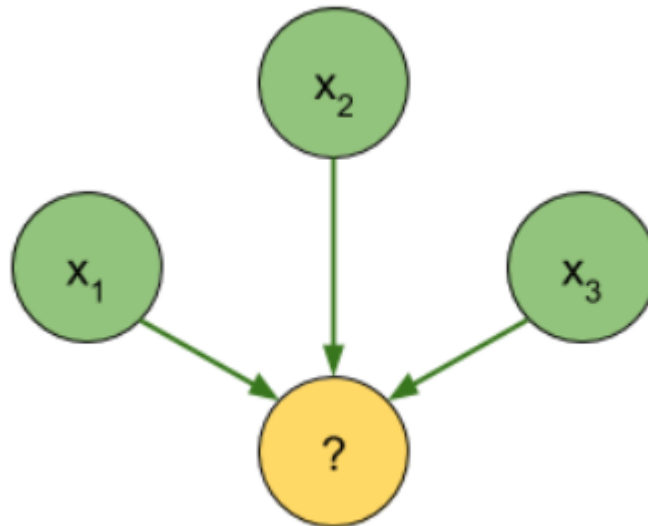
## Inconvénients :

- Modèle peu explicable.
- Modèle peu adapté aux données en grande dimension.



# Naïve Bayes

Naïve Bayes (ou bayésien naïf) est un classifieur qui tente de séparer les données d'une manière linéaire en faisant l'hypothèse, très forte, que les caractéristiques (features) sont indépendantes.



# Naïve Bayes

Avantages :

- Modèle simple à implémenter.
- Modèle très efficace notamment pour les données textuelles.

Inconvénients :

- L'hypothèse d'indépendance est rarement vérifiée.
- Très peu efficaces pour les données non linéaires.

# Machine learning

Apprentissage non supervisé

# L'apprentissage non supervisé

L'apprentissage non supervisé consiste à laisser le modèle apprendre des patterns et régularités sous-jacentes aux données sans supervision humaine.

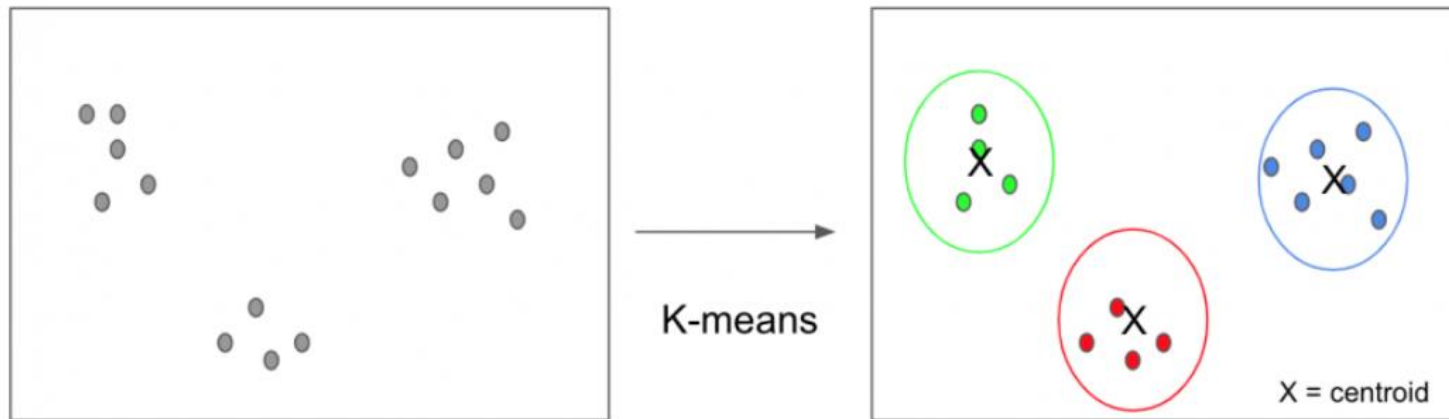
- Inputs : un jeu de données **non annotées**.  
Exemple : des documents, des caractéristiques clients, etc.
- Output : des clusters de points, des outliers, etc.  
Exemple : segments de clients, catégories de documents, etc.

Il y a deux grandes familles d'apprentissage non supervisé :

- Le clustering.
- La réduction de dimensionalité.

# K-means

K-means est un algorithme qui tente de regrouper les données en clusters en minimisant les distances entre les points d'un même cluster tout en maximisant les distances entre points appartenant à différents clusters.



# K-means

Avantages :

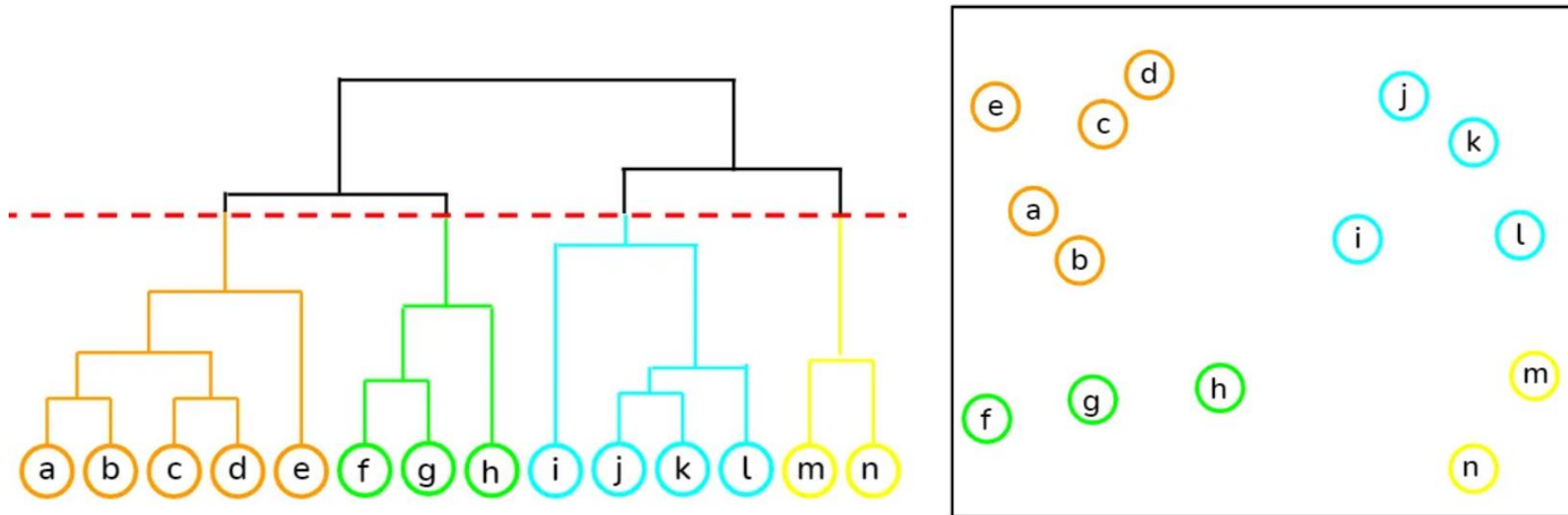
- Un algorithme efficace et simple à mettre en œuvre.
- Il converge toujours.

Inconvénients :

- Il n'est pas stable du fait de l'initialisation aléatoire (voir k-means++).
- Il n'est pas adapté aux données en grande dimension.

# Clustering hiérarchique

K-means est un algorithme qui tente de regrouper les données en clusters selon une structure arborescente (hiérarchique), qui prend en compte les relations entre les données.



# Clustering hiérarchique

Avantages :

- Algorithme flexible, qui ne requiert pas de connaître a priori le nombre de clusters.
- Clusters stables, qui ne changent pas d'une exécution à une autre.

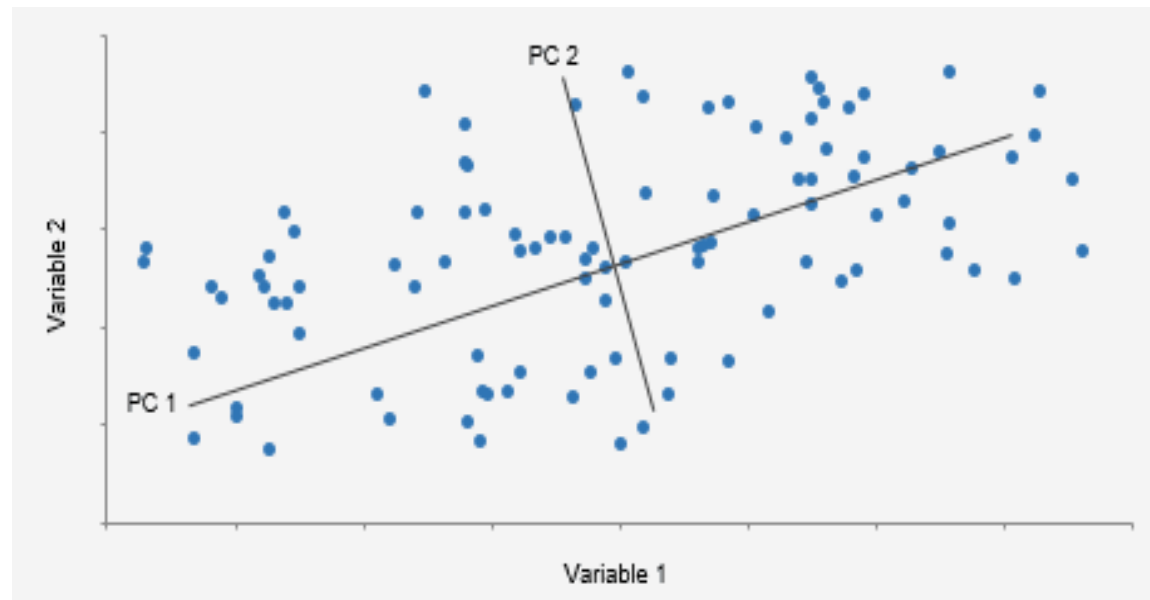
Inconvénients :

- Algorithme assez coûteux en temps de calcul.



# Analyse en composantes principales (ACP)

L'analyse par composantes principales consiste à projeter des données en grande dimension sur un espace de plus petite dimension, généralement deux ou trois, tout en gardant un maximum d'information (de variance). Elle est notamment utilisée à des fins de visualisation des données.



# Analyse en composantes principaux (ACP)

Avantages :

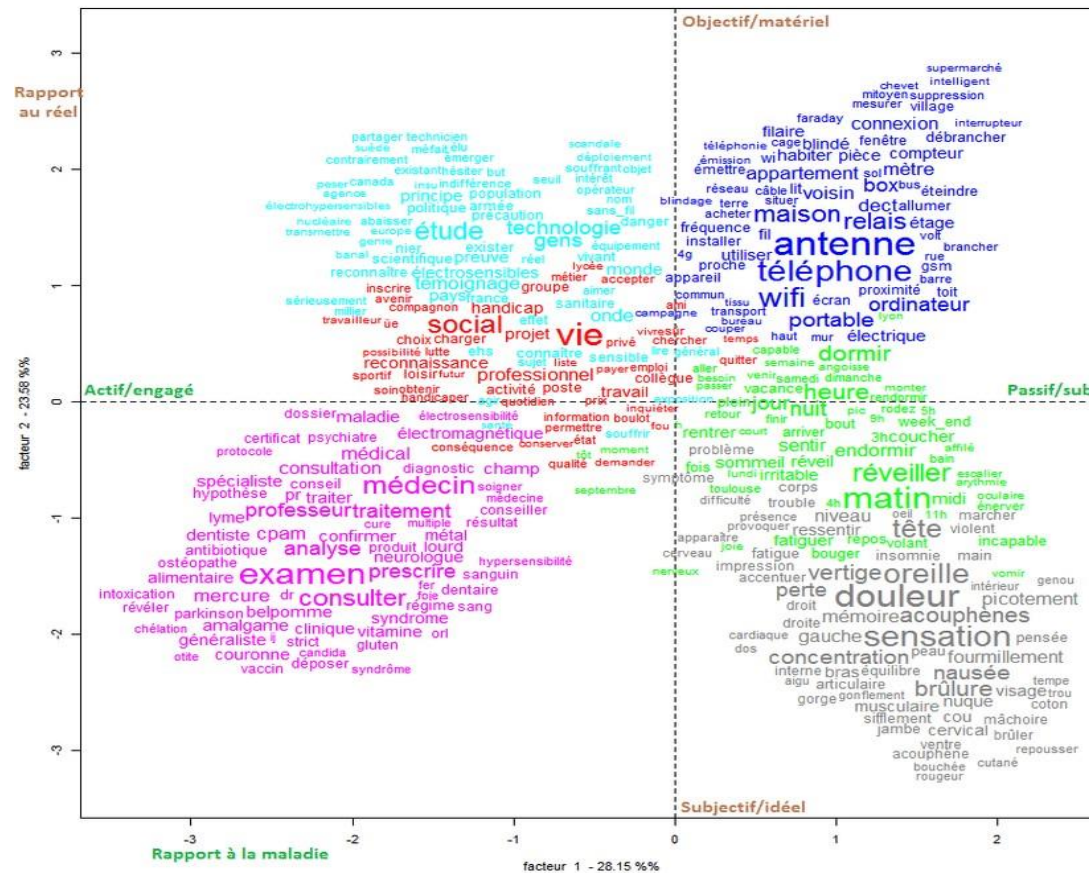
- Un algorithme simple à mettre en œuvre.
- Très adapté aux données en grande dimension.

Inconvénients :

- Les composantes principales sont relativement difficiles à interpréter.
- Peu adaptée aux données non linéaires (voir ACP à noyau).

## Analyse factorielle des correspondances

L'analyse factorielle des correspondances permet de représenter les informations contenues dans les données quand celles-ci impliquent des variables qualitatives.

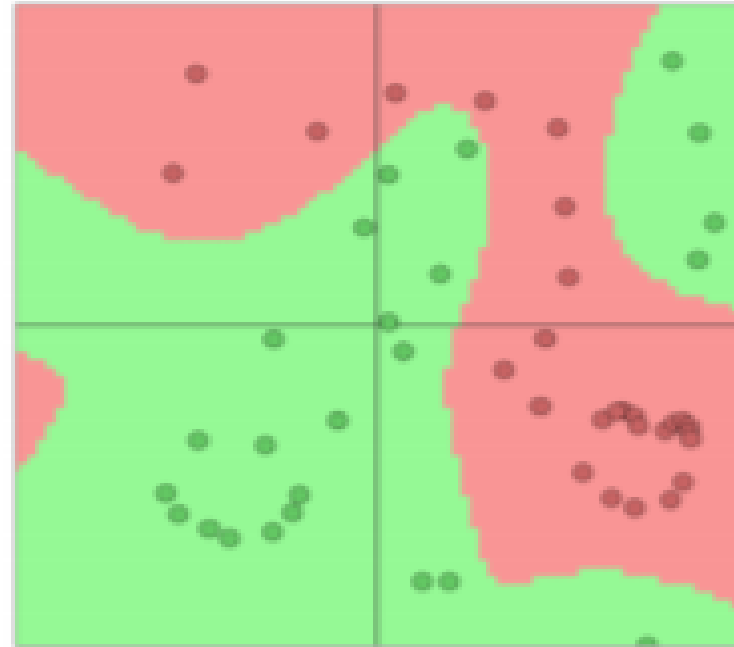


# Deep learning

# Éléments sur les réseaux de neurones

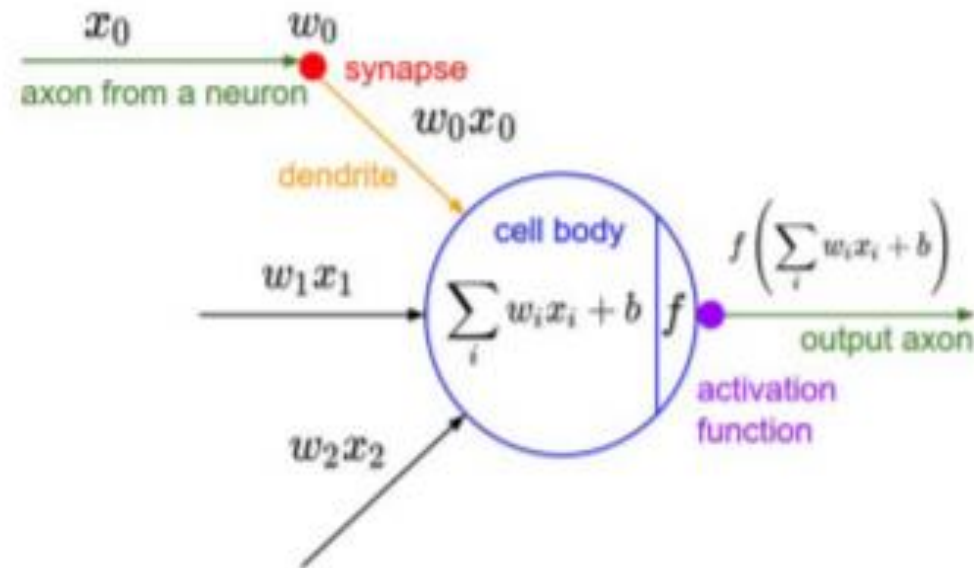
Les réseaux de neurones sont des modèles très efficace pour les données non structurés (images, texte, etc.).

Ils sont particulièrement adaptés pour les données non linéaires et en grande dimension



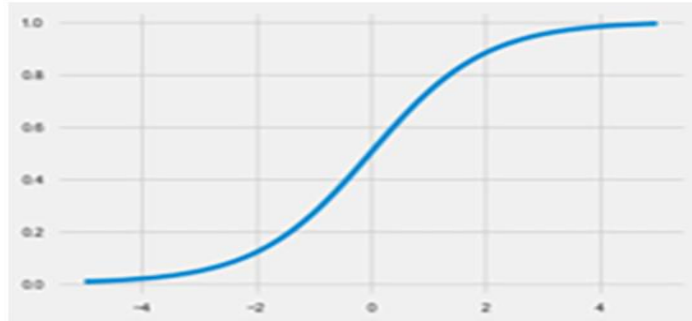
# Concept des réseaux de neurones

Les réseaux de neurones sont composées de cellules élémentaires, les neurones, permettant de réaliser des opérations simples (addition, multiplication).

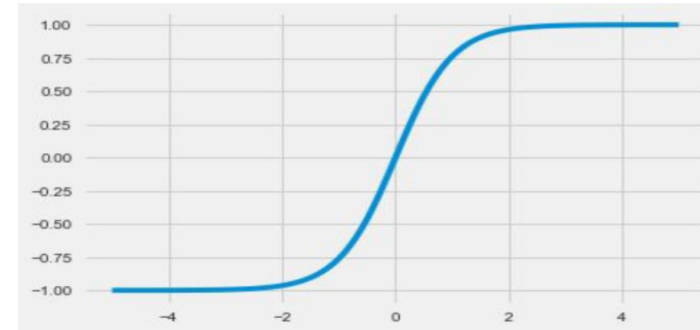


# Fonctions d'activation

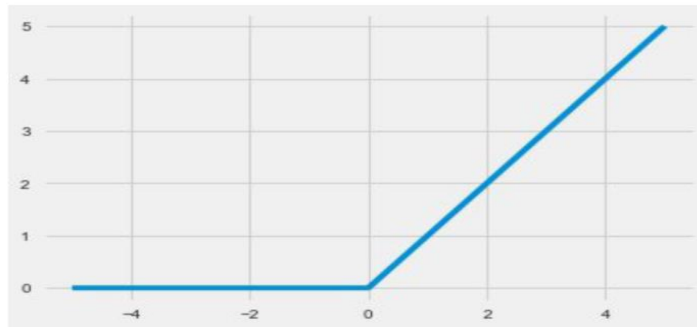
Sigmoïde



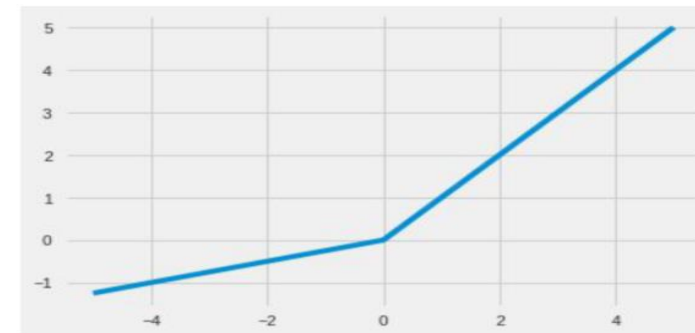
Tanh



ReLU



ReLU paramétrique

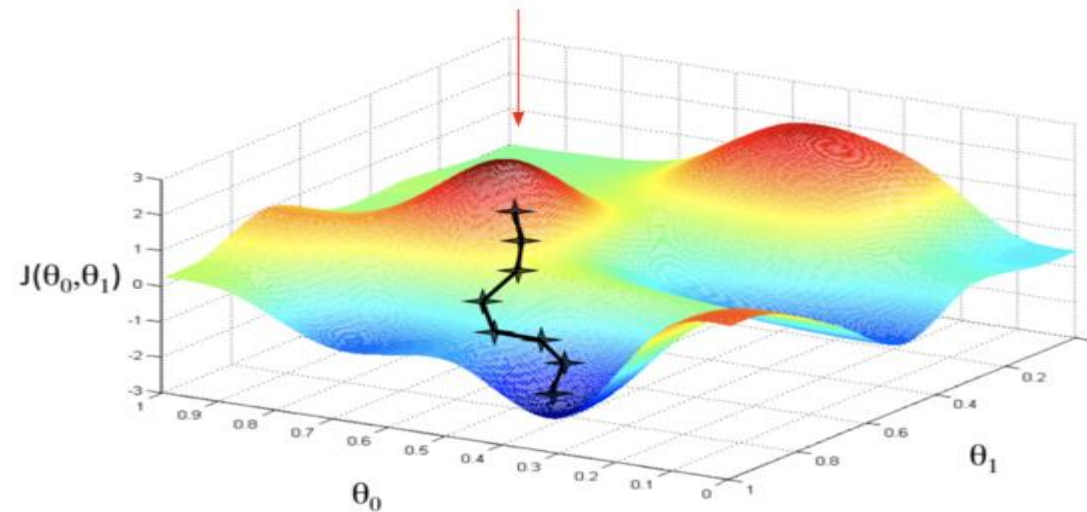


$$\text{softmax}(x_i) = \frac{e^{x_i}}{\sum_{j=1}^k e^{x_j}}$$

# Entraînement des réseaux de neurones : descente du gradient

$$\theta^{new} = \theta^{old} - \alpha \nabla_{\theta} J(\theta) \quad \nabla_{\theta} J(\theta) \text{ est calculé par backpropagation}$$

Learning rate





# Backpropagation

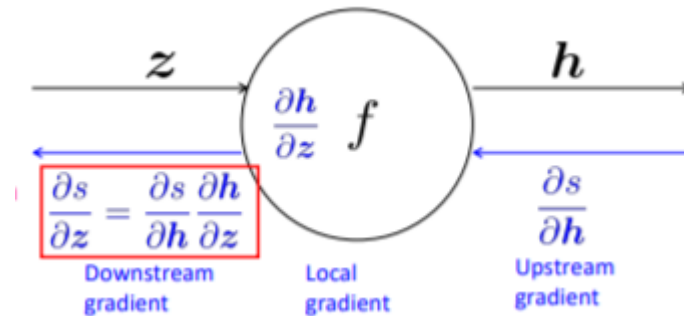
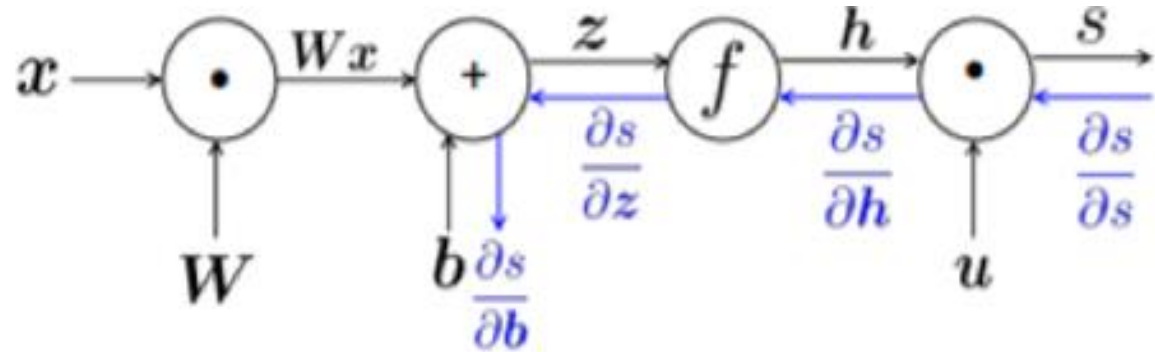
Les paramètres sont mis à jour (calculés) en partant de l'erreur et en remettant les différentes couches du réseau.

$$s = u^T h$$

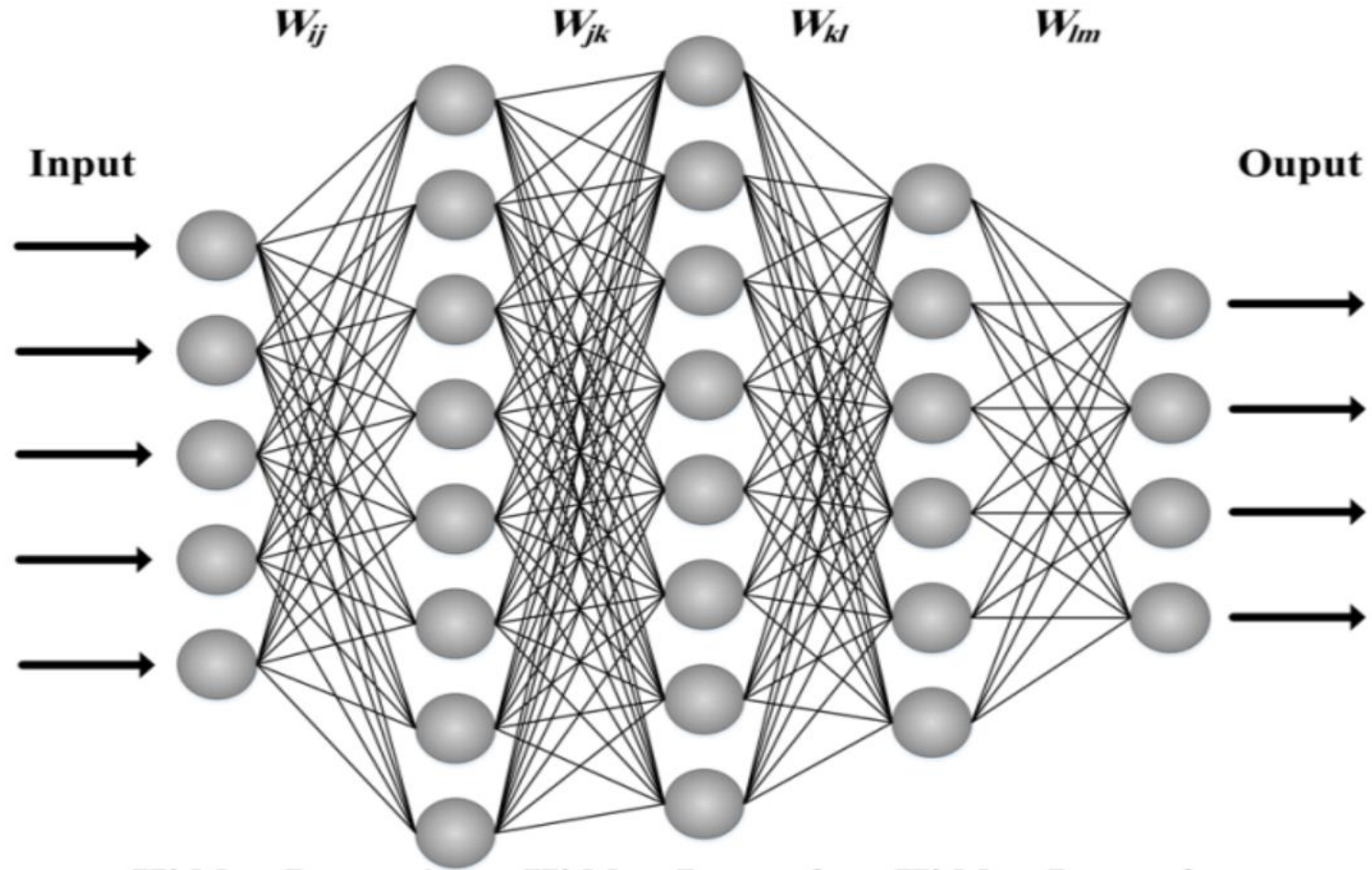
$$h = f(z)$$

$$z = Wx + b$$

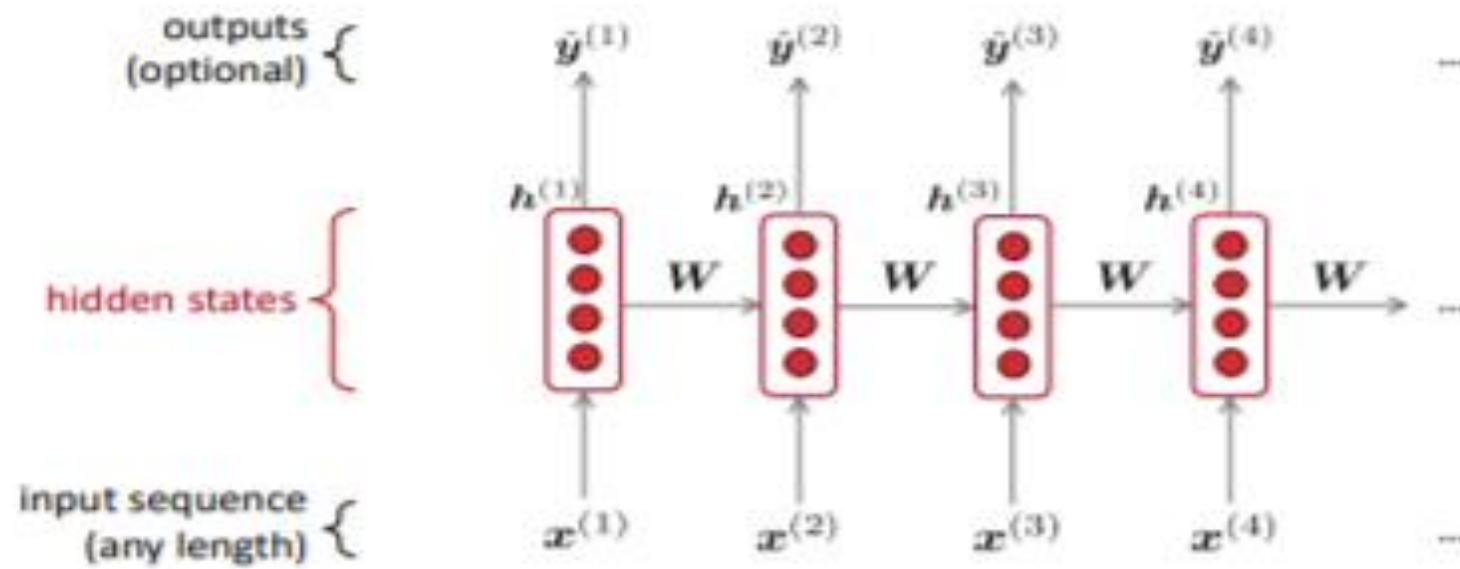
$x$  (input)



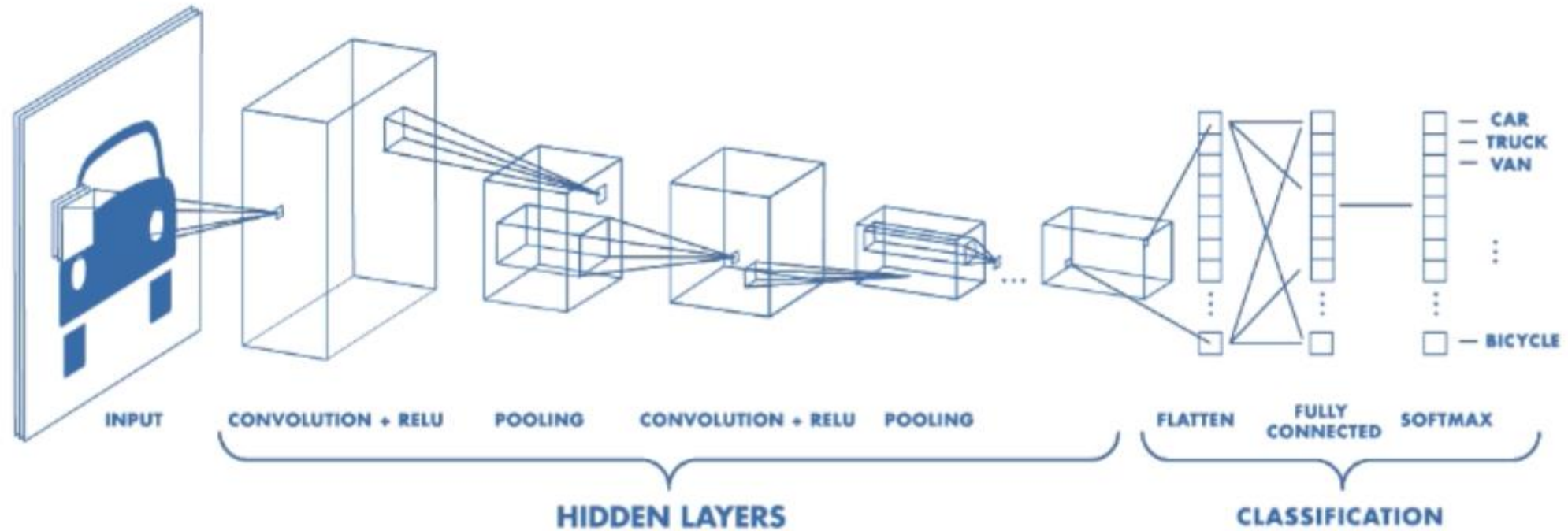
# Perceptron multi-couches



# Les réseaux de neurones récurrents (RNN)



# Les réseaux de neurones convolutifs (CNN)



# Les réseaux de neurones graphiques (GNN)

Les réseaux de neurones graphiques sont adaptés aux données qui se présentent sous forme de graphe : réseaux de transport, réseaux sociaux, plateforme e-commerce, molécules, etc.

Les tâches peuvent consister en :

- Classification des nœuds : classer les utilisateurs d'une plateforme e-commerce en fonction de leurs habitudes d'achat (liens avec les articles).
- Prédiction des liens : suggérer aux utilisateurs d'un réseau social de suivre d'autres utilisateurs.
- Classification des graphes : catégoriser les molécules en fonction de leur structure géométrique.



T1037 / 6vr4  
90.7 GDT  
(RNA polymerase domain)



T1049 / 6y4f  
93.3 GDT  
(adhesin tip)

# Points d'attention

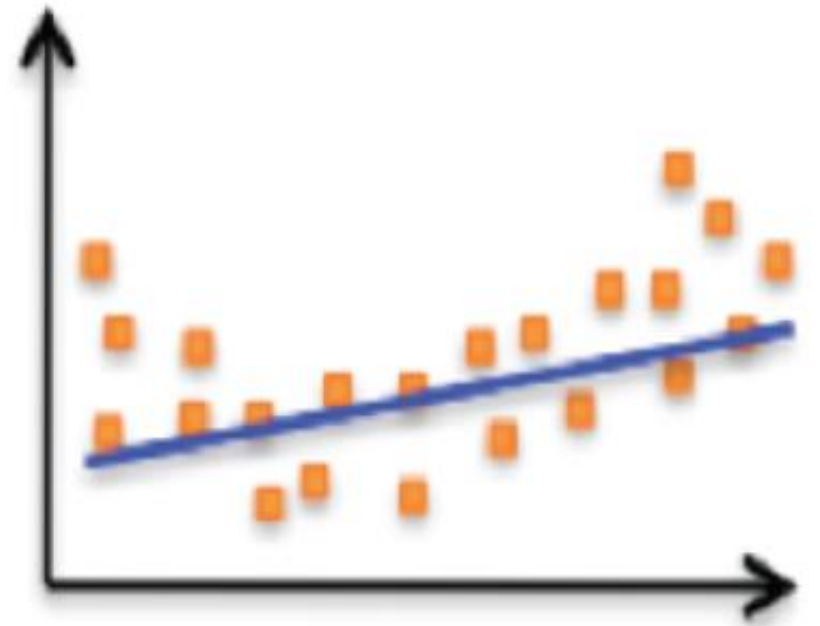


# Concept de sous-apprentissage

On dit qu'un modèle est en régime de sous-apprentissage lorsque celui-ci ne permet pas de capturer la complexité (l'information) des données.

Le sous-apprentissage apparaît généralement lorsque le modèle est trop simple (peu de paramètres, linéaire, réseau de neurones peu profond, etc.).

Pour remédier au problème de sous-apprentissage, il suffit de considérer un modèle plus complexe.

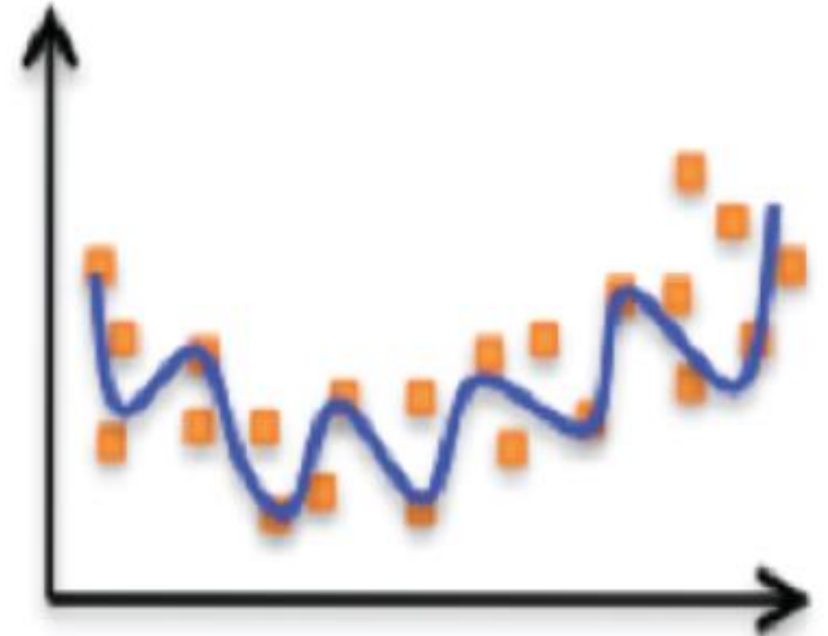


# Concept de sur-apprentissage

On dit qu'un modèle surapprend (ou overfit) quand il est trop adapté aux données d'entraînement, ce qui induit une mauvaise généralisation aux données de test, a fortiori celle du monde réel.

Le surapprentissage apparaît notamment quand :

- Le nombre de points de données est très faible (grande dimension, etc.);
- Le modèle est trop complexe par rapport aux données.





# Régularisation

Il y a plusieurs approches ou techniques permettant de réduire le problème de surapprentissage.

- Des techniques de régularisation de type L1, L2, dropout, etc.
- Construire un modèle plus simple
- Assembler plus de données pour l'entraînement.
- Etc.

# Données manquantes

Les données manquantes sont très fréquentes dans le monde réel ; elles peuvent avoir un impact significatif sur le modèle et, potentiellement, biaiser les résultats de l'analyse ou de la prédiction.

Les données manquantes sont dues à plusieurs facteurs :

- Erreurs de saisie.
- Rareté de certaines caractéristiques pour une certaine population.
- Problèmes ou difficultés de mesure de certaines grandeurs.
- Panne de capteurs.
- Désinscription, churn, etc.
- Etc. etc. etc.

# Solutions pour les données manquantes

Il y a plusieurs manières de traiter le problème des données manquantes, en fonction de la situation :

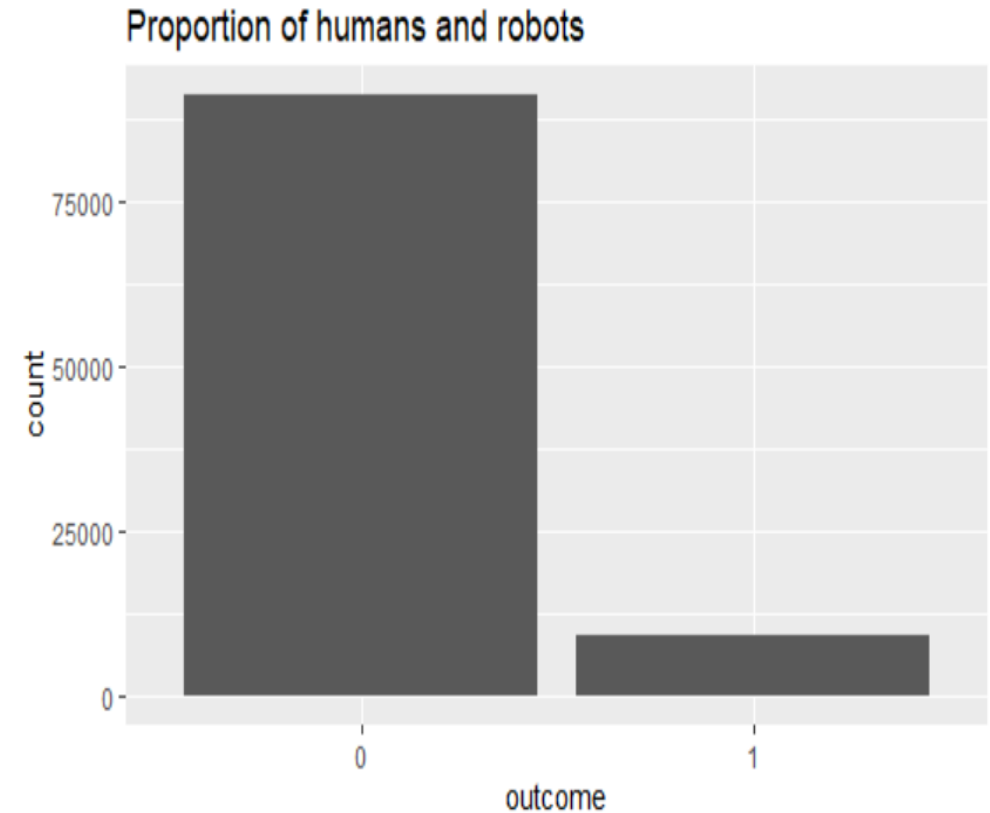
- Supprimer les individus correspondant aux données manquantes du dataset. Il faut cependant s'assurer qu'il reste suffisamment de données et que les données manquantes sont de type « missing at random ».
- Remplacer les données manquantes par la moyenne. Ceci n'est en général pas recommandé, sauf si la distribution des données a une forme gaussienne avec un écart-type pas trop large.
- Remplacer les données manquantes par la médiane. Il faut s'assurer que la distribution des données n'est pas trop concentrée loin de la médiane.
- Techniques d'imputation (random forest, imputation multiple, etc.).

# Données déséquilibrées

On dit que les données sont déséquilibrées lorsqu'une classe au moins est sur-représentée par rapport aux autres.

Quelques solutions pour les données déséquilibrées :

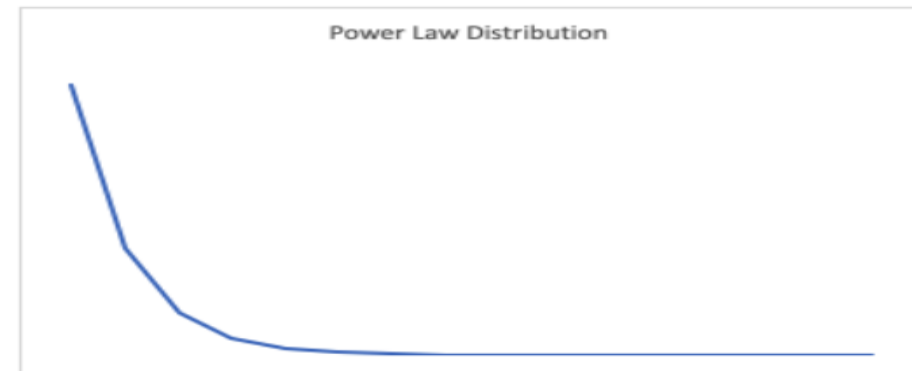
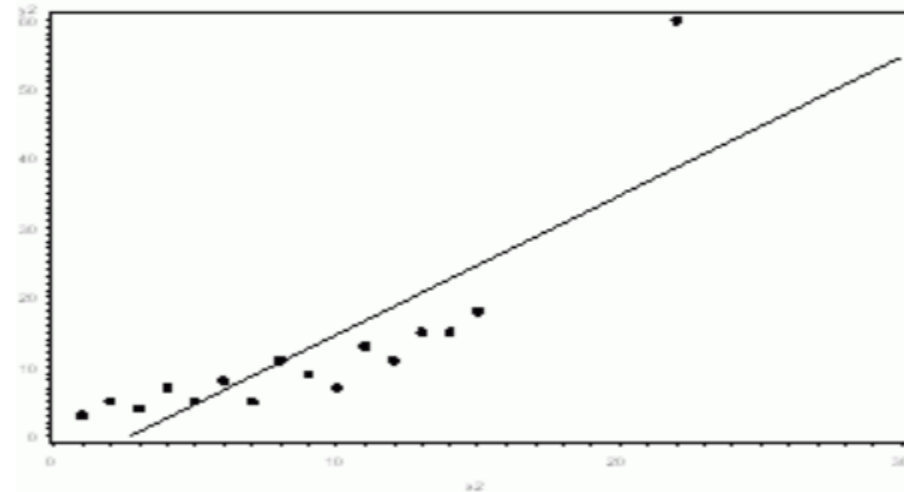
- Undersampling de la classe majoritaire s'il y a suffisamment de données.
- Oversampling de la classe minoritaire sinon.
- Modèles de type arbre de décision.



# Outliers

Les outliers sont des observations (des points de données situées loin de la moyenne (généralement plus de trois écarts-types). Pour certains modèles, notamment linéaires, les outliers peuvent avoir un effet levier très important et biaiser l'ajustement.

- Les outliers peuvent être supprimés du dataset pour améliorer la qualité de l'ajustement.
- Mais les outliers peuvent aussi être les points les plus intéressants dans le dataset.



Données suivant une loi de puissance.

# Normalisation des données

Certains modèles nécessitent une normalisation des données :

- Analyse en composantes principales.
- Régressions Ridge et Lasso.
- Machines à vecteurs de support.
- Réseaux de neurones.
- Régression linéaire quand il s'agit de comparer les coefficients.

Modèles ne nécessitant pas une normalisation :

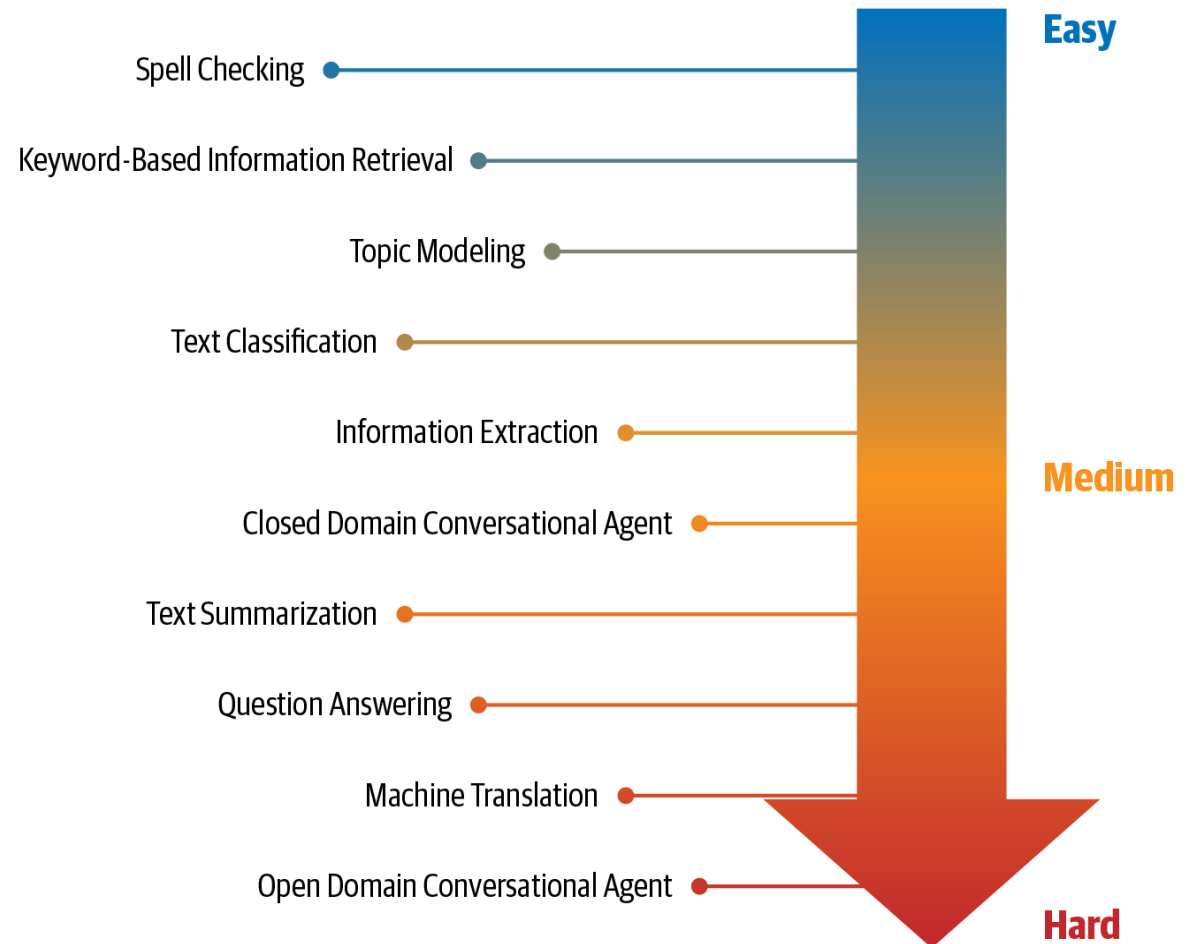
- Régression linéaire.
- Régression logistique.
- Arbres de décision.
- Random forests.

# Analyse des données textuelles

# Tâches classiques en NLP

On retrouve dans le NLP un certain nombre de tâches classiques, qui servent à la fois dans les applications pratiques que dans l'évaluation de l'état de l'art.

- Modélisation du langage (language modeling)
- Classification de texte
- Extraction d'informations
- Détection de thématiques (topic modeling)
- Résumé de texte
- Question/réponse
- Etc.





# Pourquoi le NLP est si difficile ?

- **Le langage naturel est ambigu**  
“Acheter un avocat”
- **Le langage naturel repose sur le sens commun**  
Il fait froid en hiver  
Le beurre fond dans le four
- **Le langage naturel n’est pas basé sur des règles**  
*En bleu adorable fleurit  
Le toit de métal du clocher. Alentour  
Plane un cri d’hirondelles, autour  
S’étend le bleu le plus touchant. Le soleil*  
Friedrich Hölderlin

# Pré-traitement des données textuelles

Les données textuelles sont non structurées. Les principales étapes de pré-traitement impliquent :

1. Segmentation des phrases et tokénisation ;
2. Suppression de stop words, de la ponctuation, etc.
3. Stemming, lemmatisation, conversion des majuscules, etc.

Et

Détection du langage, POS tagging, etc.

# Tokénisation

La tokénisation consiste à segmenter une phrase en unités plus petites (tokens), qui sont généralement des mots, mais qui peuvent également être des caractères ou ensembles de caractères.

"Il fait beau aujourd'hui à Paris" → ["Il", "fait", "beau", "aujourd'hui", "à", "Paris"]

"Paris" → ["P", "a", "r", "i", "s"]

La tokénisation par subword est utilisée d'une manière assez fréquente en deep learning.

# Stemming and lemmatisation

- Le stemming est l'opération qui consiste à supprimer les suffixes et réduire les mots à leur forme de base.

```
marcher --> march  
marchait --> march  
marcha --> march  
marchassiez --> march  
marché --> march
```

- La lemmatisation est l'opération qui consiste à réduire un mot à sa racine.

```
nous --> nous  
sommes --> être  
venus --> venir  
faire --> faire  
les --> le  
marchés --> marché
```

# Vectorisation

- One-Hot encoding: Etant donné un vocabulaire  $V$ , chaque mot est représenté par un vecteur binaire (à 0 ou 1) de dimension  $V$ .

$V = \{\text{Tom, mange, pomme, chien, ciel}\},$  où  $\dim(V) = 5$

Tom = [1, 0, 0, 0, 0]

mange = [0, 1, 0, 0, 0]

pomme = [0, 0, 1, 0, 0]

etc.

- N-grams

“machine learning est très utilisé en NLP.”

1-gram: {machine, learning, est, très, utilisé, en, NLP}

2-gram: {machine learning, learning est, est utilisé, utilisé en, en NLP}

# Entités nommées (NER)

L'extraction d'entités nommées consiste à détecter des informations clés qui sont des noms propres (noms de personnes, de lieux, etc.)

Established in 1896, the site was taken over by **Askham Bryan** in 2011 and it has 536 **students**, including apprentices. A group set up to keep it open had tried to find another college to take it over, after two previous bids were deemed unsuitable. Tim Whitaker, chief executive officer and principal at Askham Bryan College, said he regretted "the upset" the closure and job losses will cause.

"Whilst it was very disappointing that the strategic review didn't receive a sustainable option for Newton Rigg campus, we welcome the plans for the preservation of land-based provision in Cumbria.

"We will support and work with those involved in these plans, to ensure that current students and future applicants interested in land-based courses have a smooth transition."

He added that the college had "always been clear" educational provision would not continue at Newton Rigg from July 2021, and students, staff and the local community were told in 2020.

# Extraction de relations

L'extraction de relations consiste à détecter des relations entre deux entités nommées.

Established in 1896, the site was taken over by Askham Bryan in 2011 and it has 536 students, including apprentices.

A group set up to keep it open had tried to find another college to take it over, after two previous bids were deemed unsuitable.

**Tim Whitaker, chief executive officer and principal at Askham Bryan College**, said he regretted "the upset" the closure and job losses will cause.

"Whilst it was very disappointing that the strategic review didn't receive a sustainable option for Newton Rigg campus, we welcome the plans for the preservation of land-based provision in Cumbria.

"We will support and work with those involved in these plans, to ensure that current students and future applicants interested in land-based courses have a smooth transition."

He added that the college had "always been clear" educational provision would not continue at Newton Rigg from July 2021, and students, staff and the local community were told in 2020.

# Extraction de mots clés

L'extraction de mots clés consiste à extraire les mots clés les plus importants, permettant de capturer le contenu d'un texte.

Established in 1896, the site was taken over by Askham Bryan in 2011 and it has 536 **students**, including apprentices. A group set up to keep it open had tried to find another college to take it over, after two previous bids were deemed unsuitable. Tim Whitaker, chief executive officer and principal at Askham Bryan College, said he regretted "the upset" the closure and job losses will cause.

"Whilst it was very disappointing that the strategic review didn't receive a sustainable option for Newton Rigg campus, we welcome the plans for the preservation of land-based provision in Cumbria.

"We will support and work with those involved in these plans, to ensure that current students and future applicants interested in land-based courses have a smooth transition."

He added that the college had "always been clear" educational provision would not continue at Newton Rigg from July 2021, and students, staff and the local community were told in 2020.



# Part of Speech tagging

Part of speech consiste à détecter la catégorie grammaticale d'un mot au sein d'un texte.

Exemple :

*« this is a very simple example »*

```
[('this', 'DT'), ('is', 'VBZ'), ('a', 'DT'), ('very', 'RB'), ('simple', 'JJ'), ('example', 'NN')]
```

# Term Frequency

- Term Frequency (TF) : la fréquence avec laquelle un terme  $t$  apparaît dans un document  $d$ .

$$\text{TF}(t, d) = \frac{(\text{Number of occurrences of term } t \text{ in document } d)}{(\text{Total number of terms in the document } d)}$$

Exemple : étant donné 1000 mots.

Terme	Fréquence	TF score
chat	15	0.015
lait	4	0.004
mange	6	0.006
feu	3	0.003
...	...	...

# Inverse Document Frequency

- Inverse Document Frequency (IDF) : permet de pénaliser les termes les plus fréquents et de donner plus de poids aux termes les moins fréquents (plus parlants).

$$\text{IDF}(t) = \log_e \frac{(\text{Total number of documents in the corpus})}{(\text{Number of documents with term } t \text{ in them})}$$

- Example: given 100 documents

Terme	Nombre de documents	IDF score
chat	70	0.36
lait	9	2.41
manger	80	0.22
feu	5	3.00
...	...	...

# Term Frequency – Inverse Document Frequency

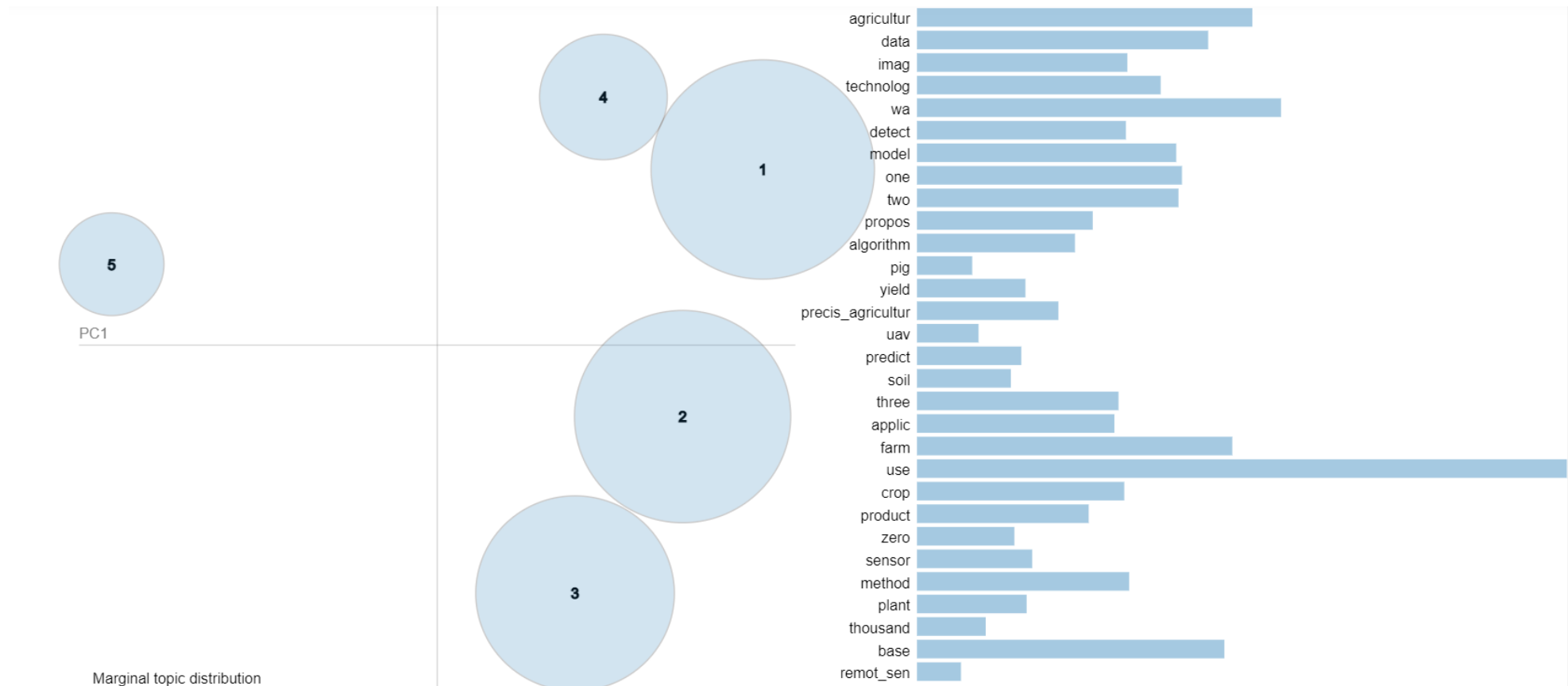
- TF-IDF score est le produit de Tf et IDF.

$$TF\text{-}IDF\text{ Score} = TF\text{ Score} \times IDF\text{ score}$$

Terme	TF score	IDF score	TF-IDF score
chat	0.015	0.36	0.0054
lait	0.004	2.41	0.0096
manger	0.006	0.22	0.0013
feu	0.003	3.00	0.0090
...	...	...	

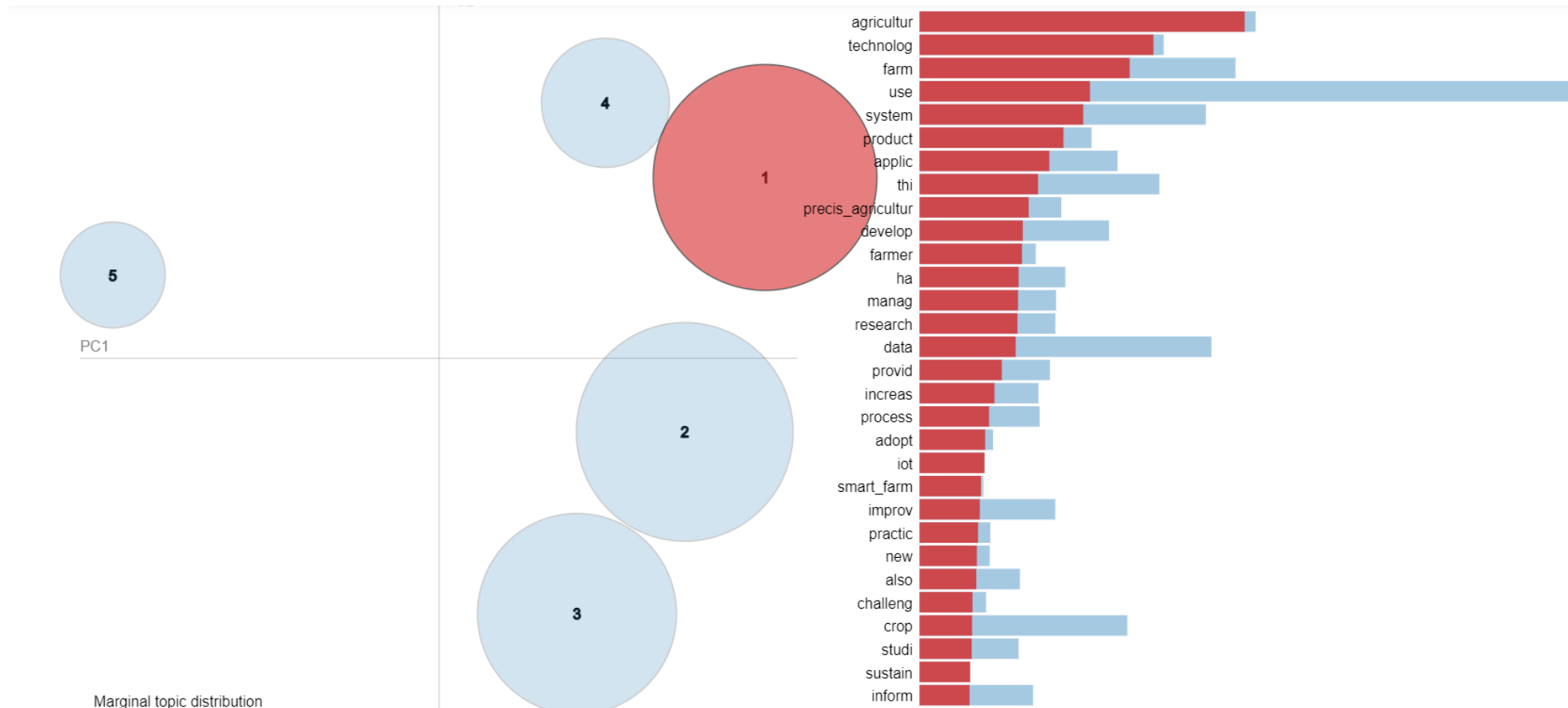
# Topic Modeling : LDA

Extraction de thématiques d'un ensemble de documents traitant de l'agriculture l'aide d'une technique LDA (Latent Dirichlet Allocation).



# Topic Modeling

La thématique extraite est définie par un ensemble de mots clés.



# NLP avec deep learning

# One-Hot encoding

Dans les approches traditionnelles du NLP, les mots sont représentés par des vecteurs binaires dont les composantes sont 0 ou 1.

Par exemple, pour un vocabulaire de dimension  $\sim 100\,000$

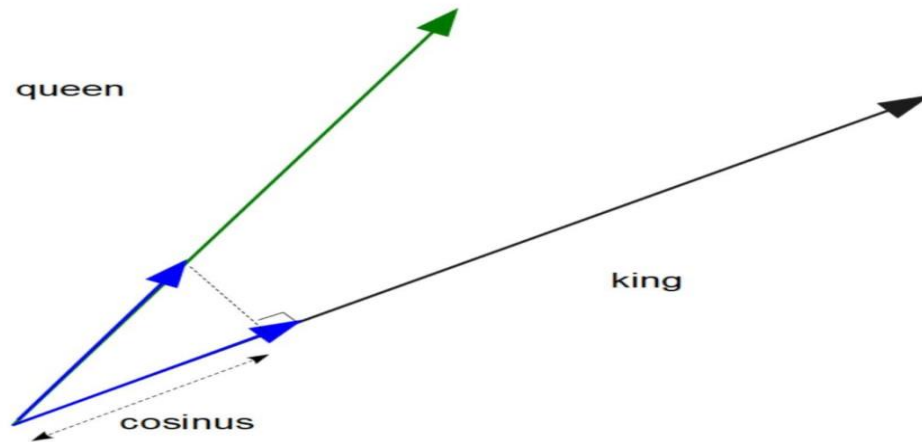
- Véhicule =  $[0, 0, 1, 0, 0, 0, 0, 0, 0, \dots, 0, 0]$
- Voiture =  $[0, 0, 0, 0, 0, 1, 0, 0, 0, \dots, 0, 0]$

Cette representation ne permet pas de prendre en compte la dimension sémantique.



# Similarités sémantiques

Nous sommes intéressés par des représentations de mots qui capturent la distance sémantique, sous forme de produit scalaire par exemple (ou distance cosine).



# Représentations distribuées (word embeddings)

*« You shall know a word by the company it keeps »  
J.R. Firth (1957)*

Les mots sont similaires s'ils apparaissent fréquemment dans le même contexte.

- Il conduit son **véhicule** pour rentrer à la maison.
- Il conduit sa **voiture** pour rentrer chez lui.

# Construction d'une représentation distribuée

Considérons ce corpus à titre d'exemple :

*cnn in crop analysis*

*cnn and svm are widely used*

*linear\_regression performed along with svm*

*linear regression for crop and farm*

*svm being used for farm monitoring*

*Do cnn, svm and linear\_regression appear in the same context?*

Le vocabulaire est alors :

[cnn, in, crop, analysis, and, svm, are, widely, used, linear\_regression, performed, along, with, for, farm, being, monitoring, do, appear, the, same, context]

**dim(vocabulaire) = 22**

# Matrice de co-occurrences

La matrice de co-occurrence représente la fréquence à laquelle les mots apparaissent ensemble deux à deux.

```

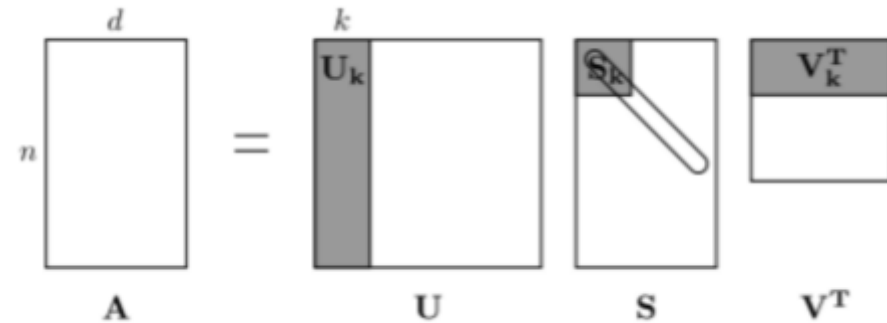
      cnn in crop ...
cnn  [[0, 2, 1, 1, 1, 2, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1],
in    [2, 0, 1, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1],
crop  [1, 1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0],
...   [1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
      [1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0],
      [2, 1, 0, 0, 1, 0, 1, 1, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1],
      [1, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
      [1, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
      [1, 0, 0, 0, 1, 2, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0],
      [1, 1, 1, 0, 1, 2, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1],
      [0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0],
      [0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0],
      [0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0],
      [0, 0, 1, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0, 0, 2, 1, 1, 0, 0, 0, 0],
      [0, 0, 1, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0, 2, 0, 1, 1, 0, 0, 0, 0],
      [0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0],
      [0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0],
      [1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1],
      [1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1],
      [1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1],
      [1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0]]
```

# Réduction de la dimension

La décomposition en valeurs singulières est une technique permettant de réduire la dimension des données (similaire à l'ACP).

Etant donné une matrice  $\mathbf{A} \in \mathbb{R}^{n \times d}$

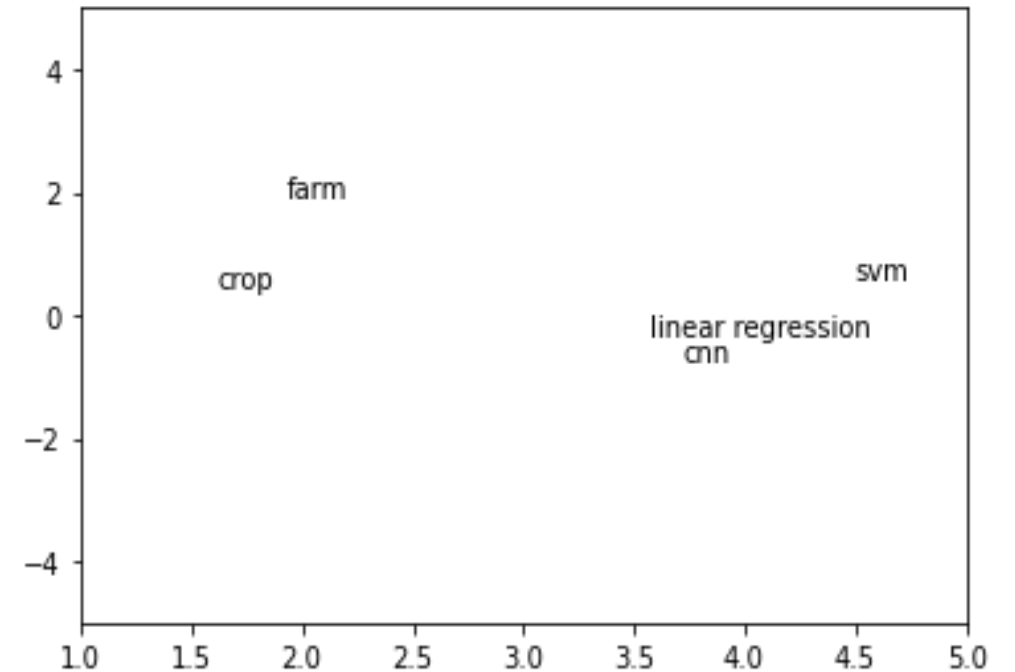
$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T \quad \text{où} \quad \mathbf{U} \in \mathbb{R}^{n \times r}, \mathbf{D} \in \mathbb{R}^{r \times r}, \mathbf{V} \in \mathbb{R}^{d \times r}.$$



$\mathbf{U}$  est la matrice contenant les représentations (vecteurs de mots)

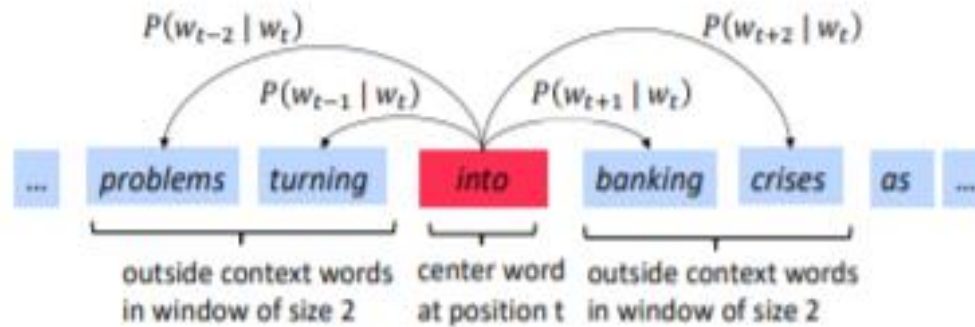
# Visualisation des représentations distribuées

```
cnr [ 3.72113518, -0.73233585],  
ln [ 2.7940387 , -1.40864784],  
crop [ 1.61178082, 0.44786021],  
... [ 0.77784994, -0.31230389],  
[ 2.12245048, 1.29571556],  
[ 4.49293777, 0.58090417],  
[ 1.33312748, 0.66758743],  
[ 1.33312748, 0.66758743],  
[ 2.25882809, 1.80738544],  
[ 3.56593605, -0.31006976],  
[ 0.95394179, 0.079159 ],  
[ 0.95394179, 0.079159 ],  
[ 0.95394179, 0.079159 ],  
[ 1.92921553, 1.94783497],  
[ 1.92921553, 1.94783497],  
[ 1.12301312, 1.42126096],  
[ 1.12301312, 1.42126096],  
[ 2.26025282, -1.31571175],  
[ 2.26025282, -1.31571175],  
[ 2.26025282, -1.31571175],  
[ 2.26025282, -1.31571175],  
[ 2.26025282, -1.31571175]]
```



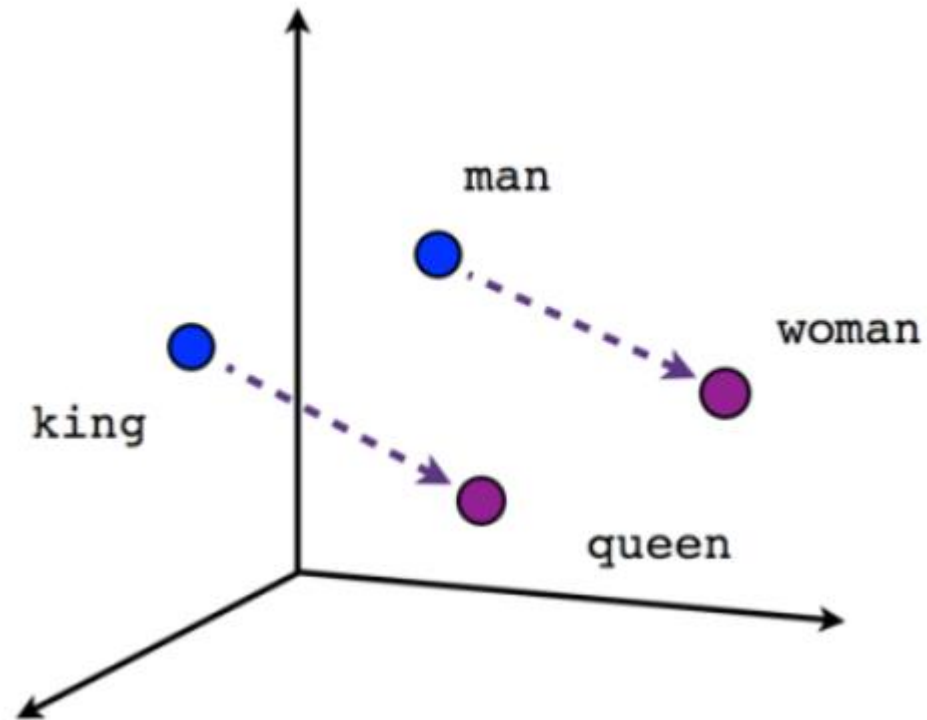
# Skip-gram et Continuous Bag of Words (CBOW)

Prédire un mot masqué dans une phrase sachant le contexte (les autres mots).



# Représentations distribuées : Word2Vec

Le modèle Word2Vec permet de prendre en compte la dimension sémantique entre les mots.





## Autres représentations distribuées

- Glove (2014) : proposé par une équipe de Stanford, il combine à la fois les techniques modernes de deep learning et les techniques statistiques (co-occurrences entre les mots). Il permet d'avoir des représentations des mots plus globales que Word2Vec.
- Fasttext (à partir de 2016) : la librairie a été créée par Facebook. Le modèle est entraîné sur des subwords (n-grams de caractères). Il est ainsi plus efficace pour traiter les mots inconnus (out of vocabulary).
- Représentations contextuelles (Transformers), etc.

# RNNs pour la génération de texte

Title: CHOCOLATE RANCH BARBECUE

Categories: Game, Casseroles, Cookies, Cookies

Yield: 6 Servings

2 tb Parmesan cheese -- chopped

1 c Coconut milk

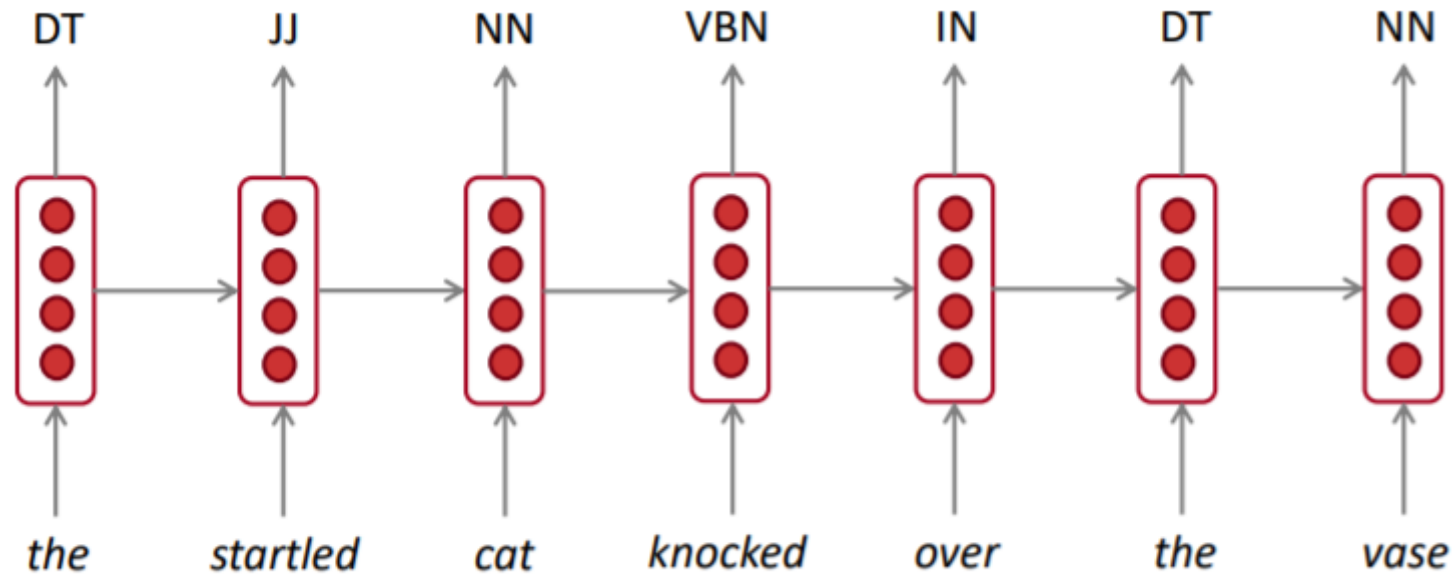
3 Eggs, beaten

Place each pasta over layers of lumps. Shape mixture into the moderate oven and simmer until firm. Serve hot in bodied fresh, mustard, orange and cheese.

Combine the cheese and salt together the dough in a large skillet; add the ingredients and stir in the chocolate and pepper.

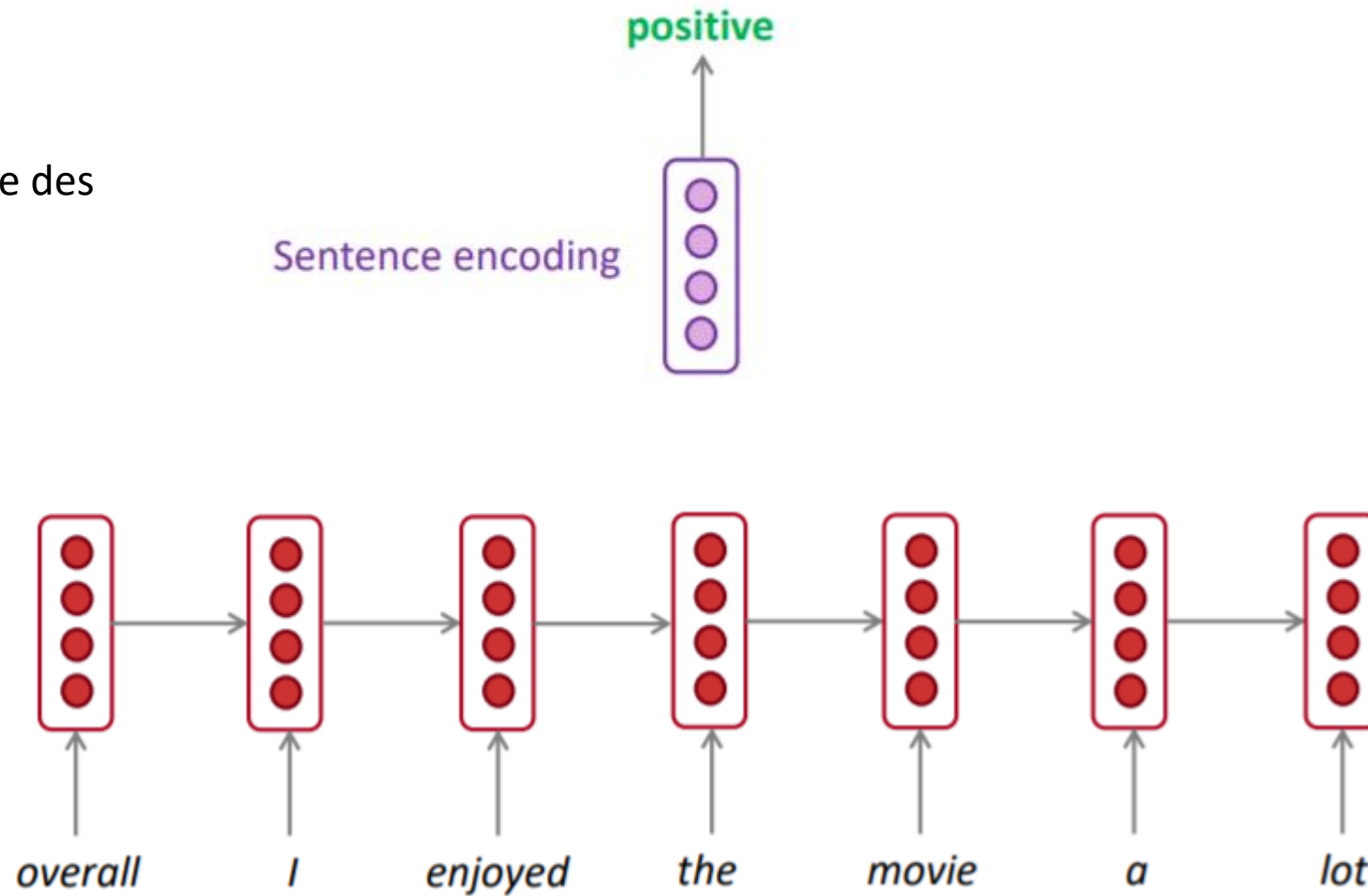
Source <https://gist.github.com/nylki/1efbaa36635956d35bcc>

## RNNs pour le part of speech tagging



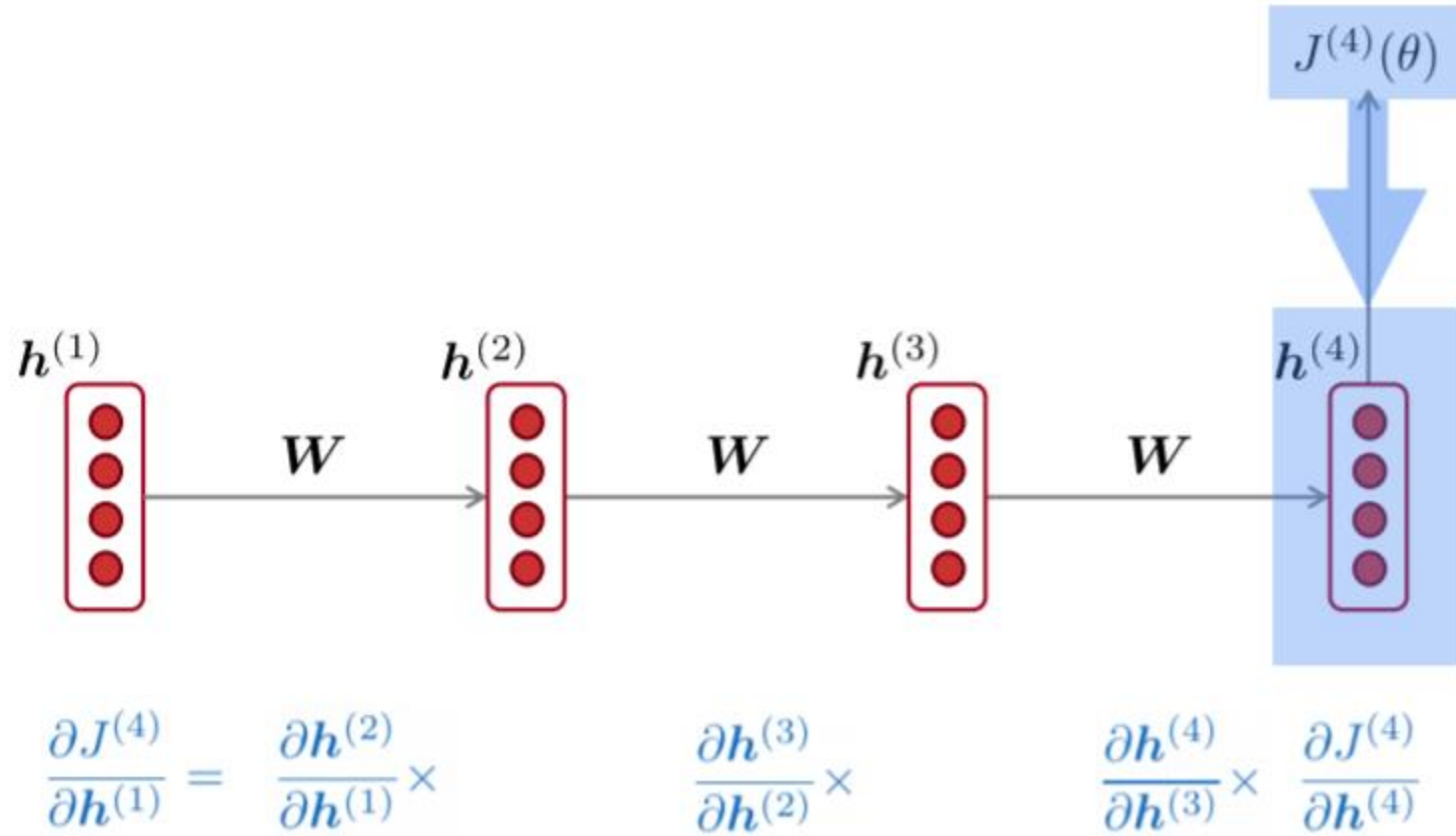
# RNNs pour la classification de texte

Exemple : analyse des sentiments

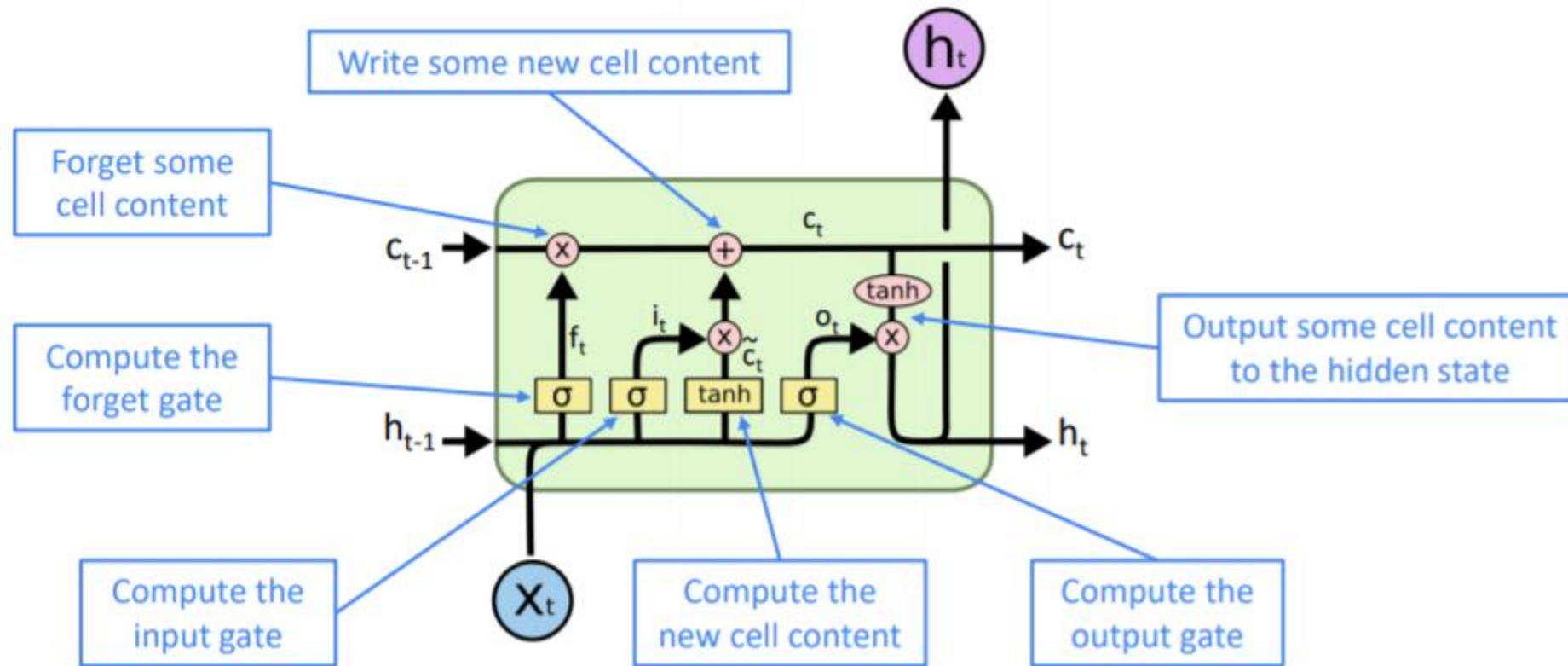


# Limitations des RNNs

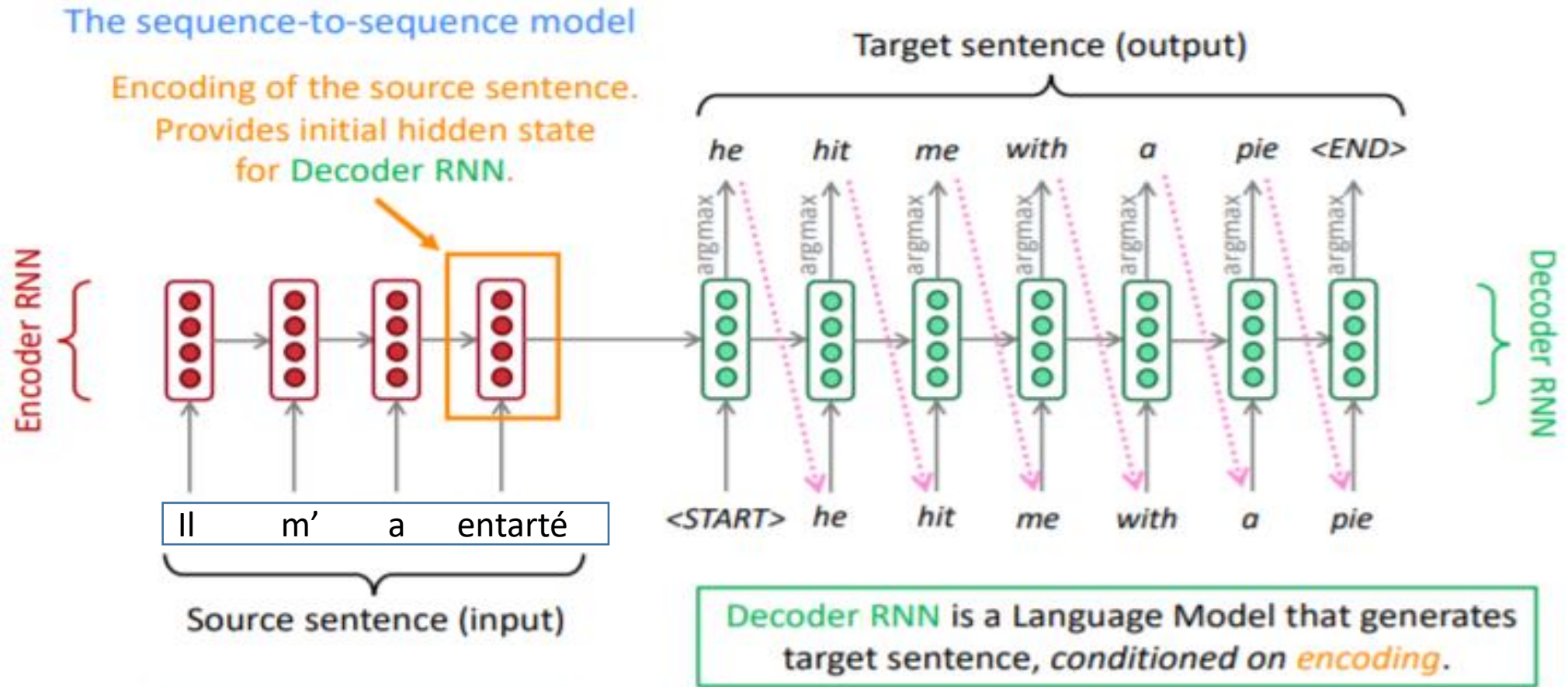
La principale limitation des RNNs est l'annulation et l'explosion du gradient.



# Long Short-Term Memory (LSTM)

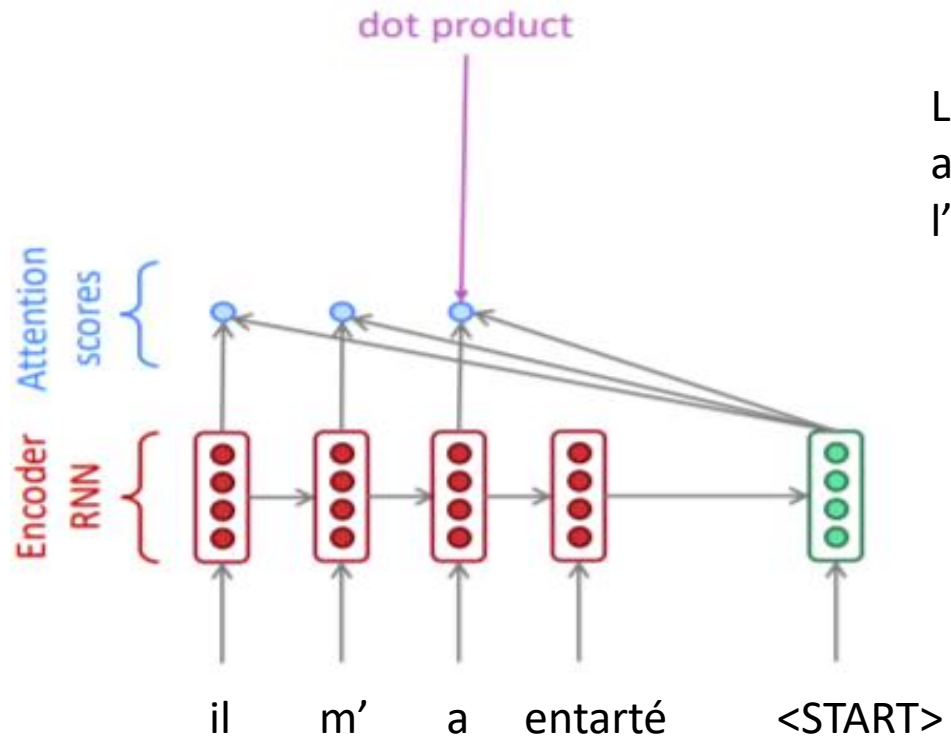


# Les modèles sequence-to-sequence models (seq2seq)





# Le mécanisme d'attention



Le mécanisme d'attention offre la possibilité au décodeur d'accéder à plusieurs mots dans l'encodeur.



## Attention Is All You Need

Ashish Vaswani\*  
Google Brain  
avaswani@google.com

Noam Shazeer\*  
Google Brain  
noam@google.com

Niki Parmar\*  
Google Research  
nikip@google.com

Jakob Uszkoreit\*  
Google Research  
usz@google.com

Llion Jones\*  
Google Research  
llion@google.com

Aidan N. Gomez\*<sup>†</sup>  
University of Toronto  
aidan@cs.toronto.edu

Łukasz Kaiser\*  
Google Brain  
lukaszkaiser@google.com

Illia Polosukhin\*<sup>‡</sup>  
illia.polosukhin@gmail.com

### Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

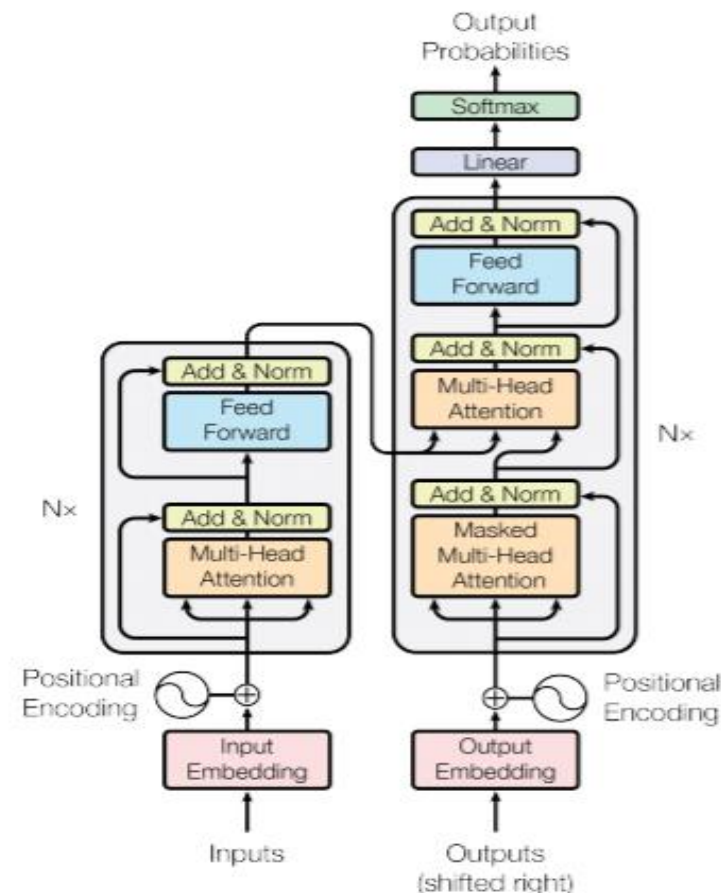
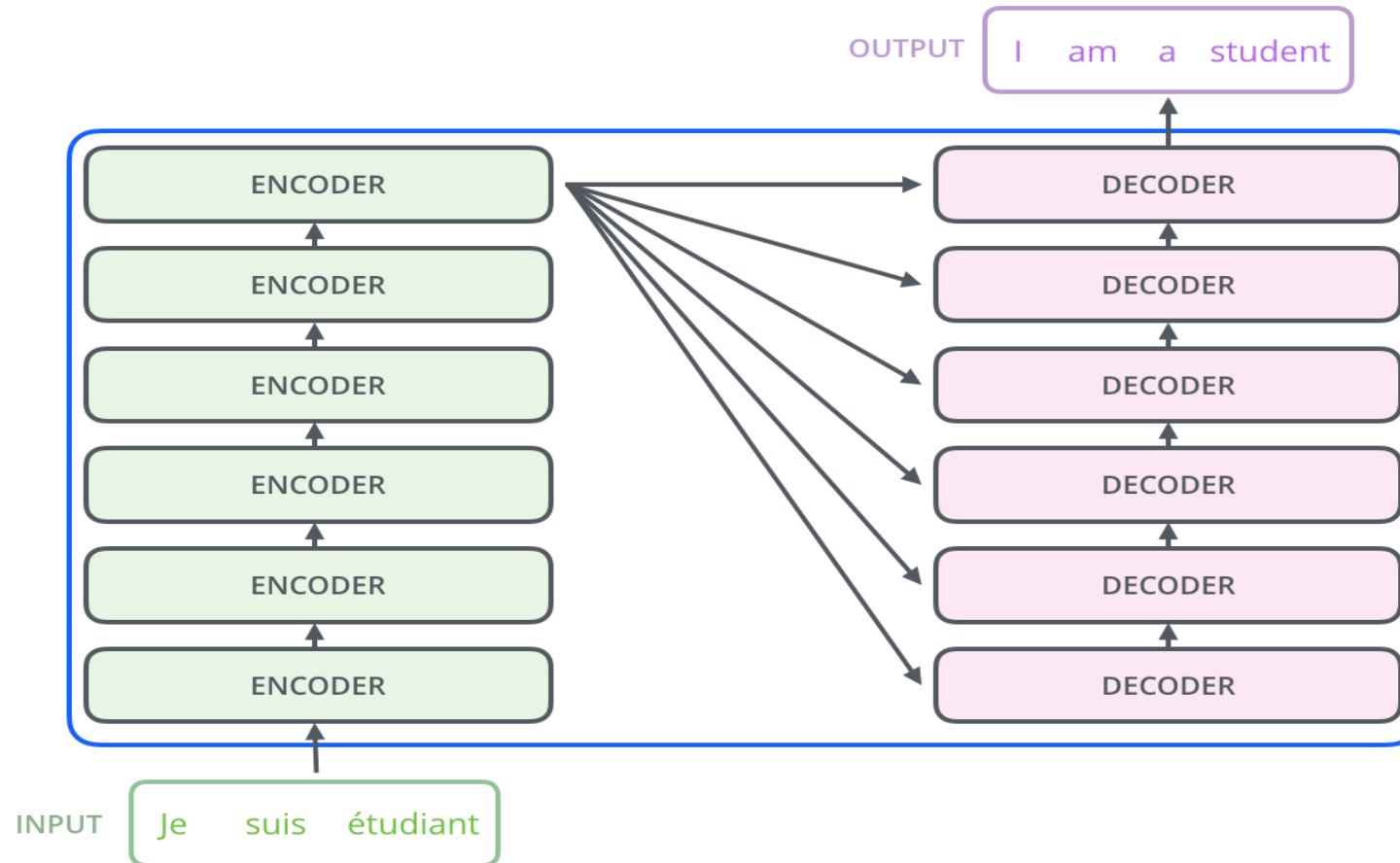
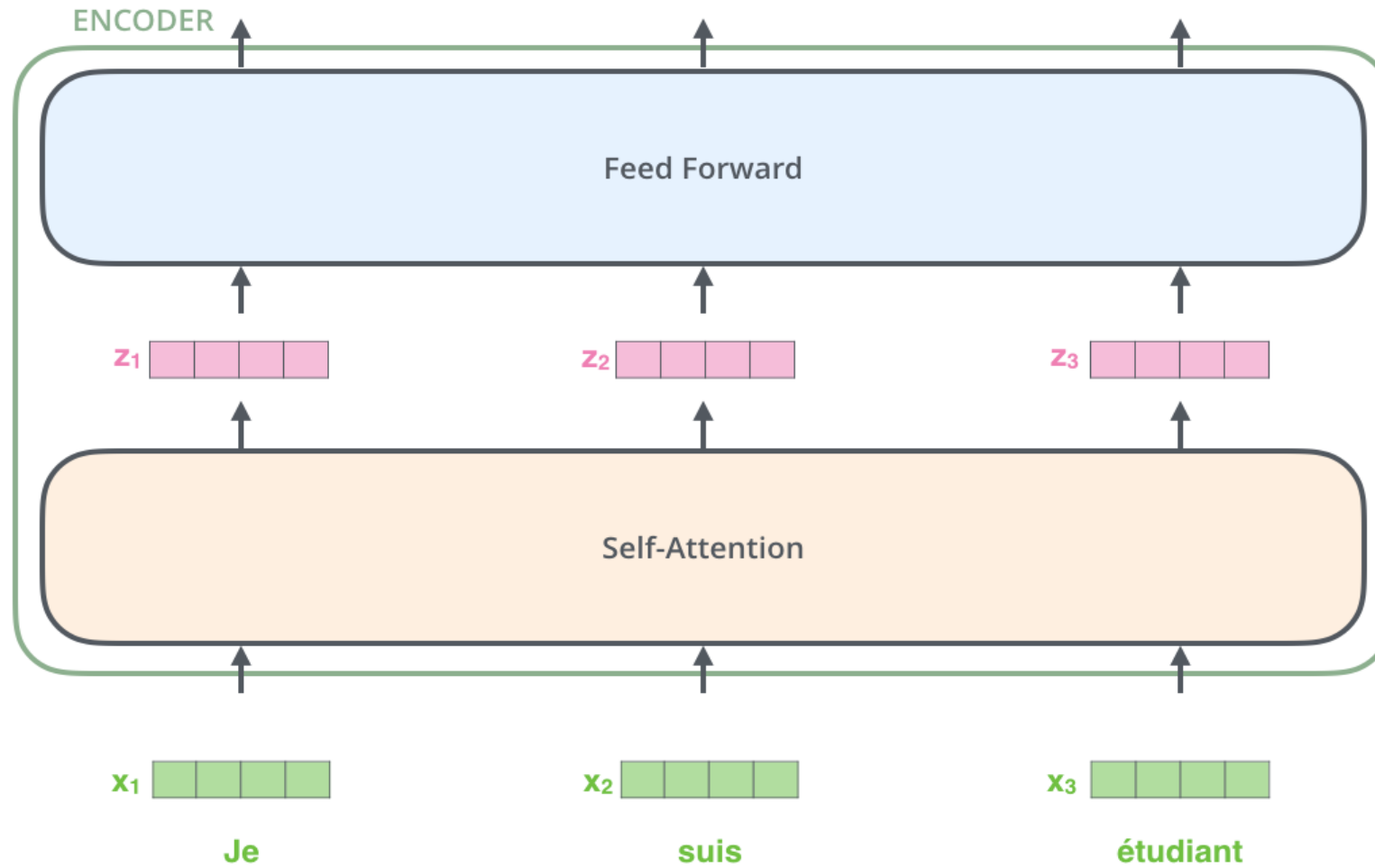


Figure 1: The Transformer - model architecture.

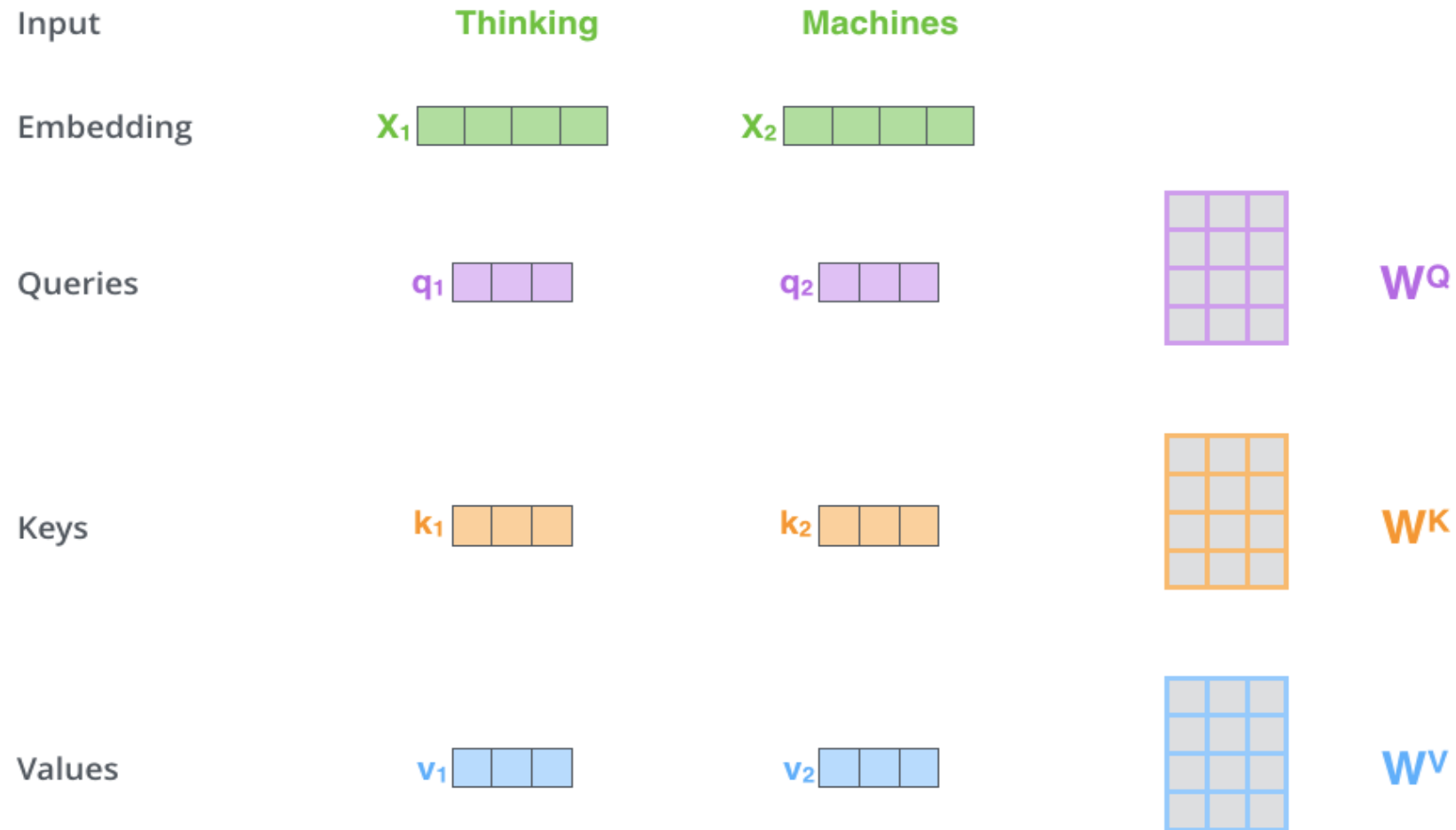
# Architecture du Transformer originel



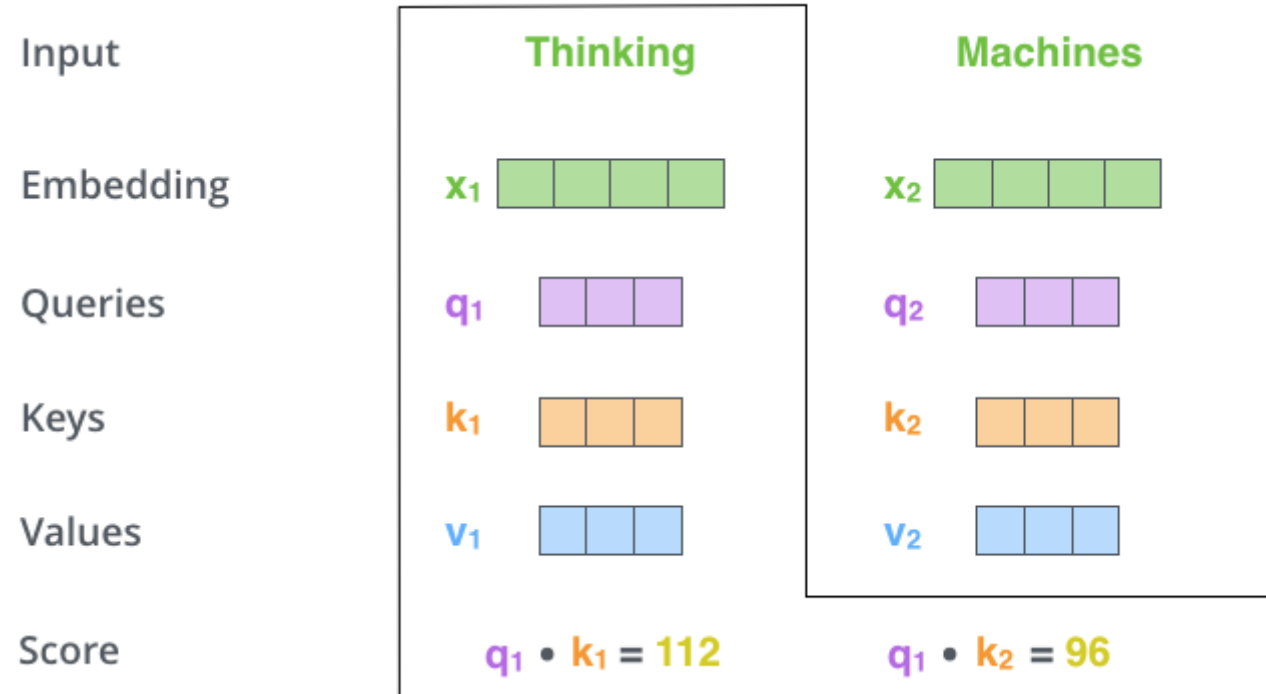
# Mécanisme de self-attention



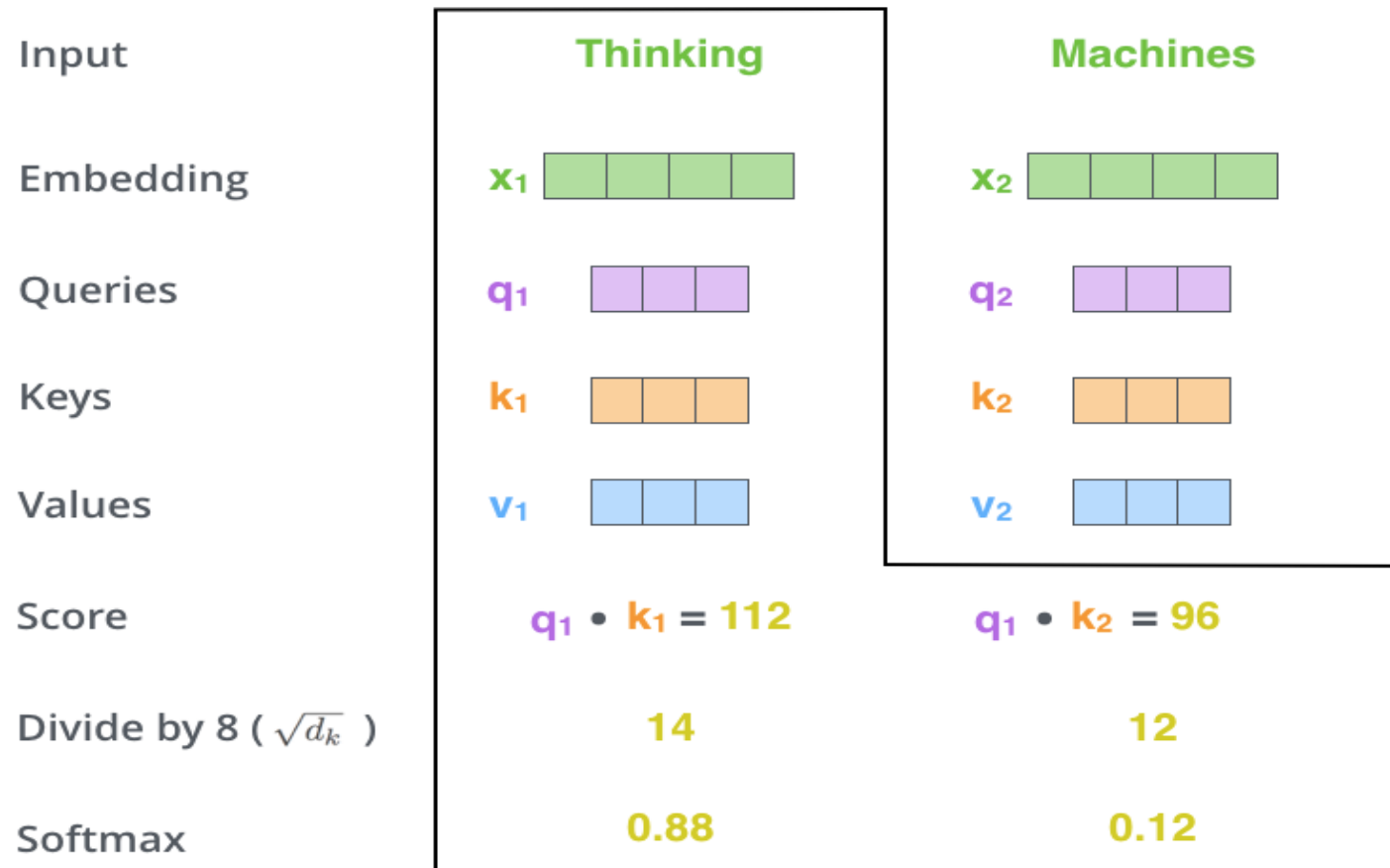
# Mécanisme de self-attention



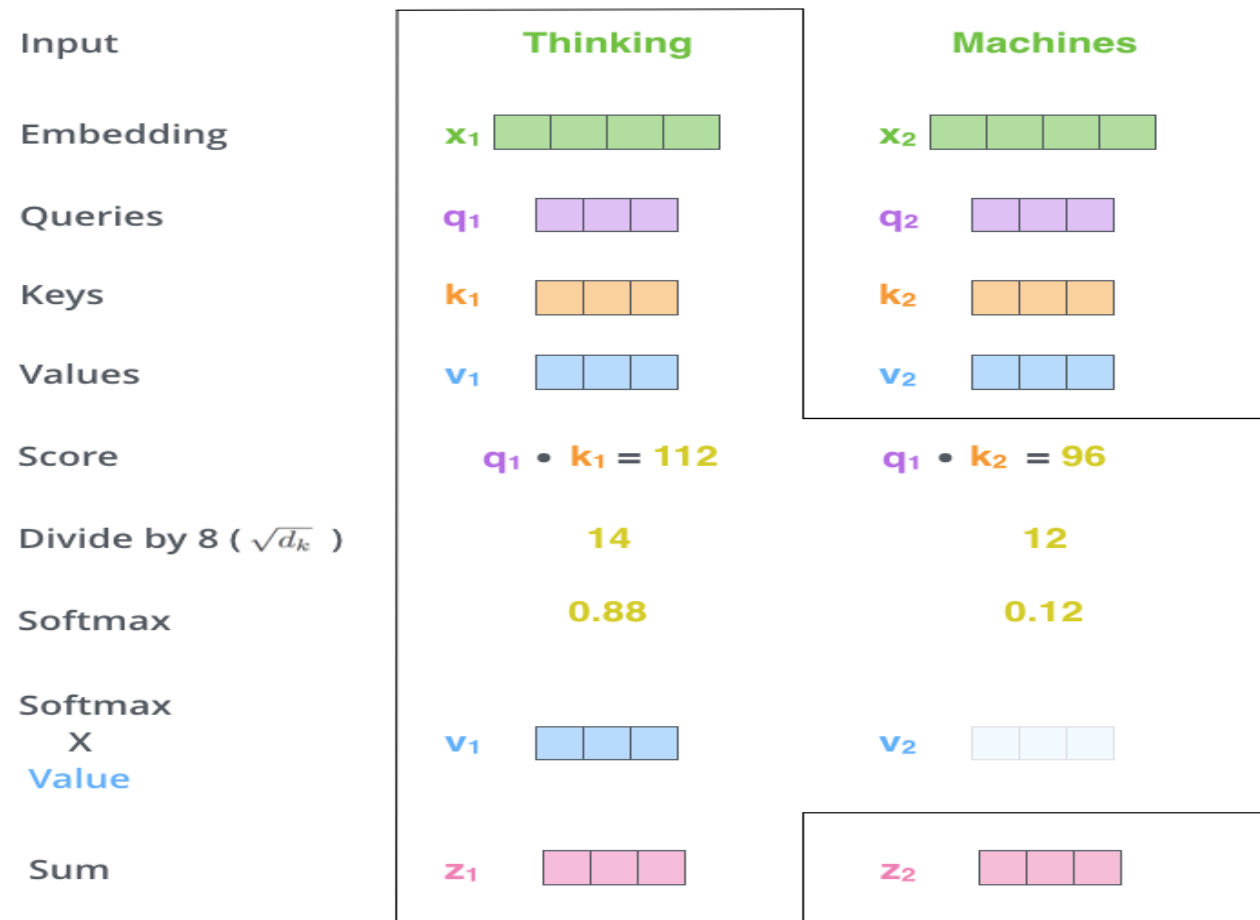
# Mécanisme de self-attention



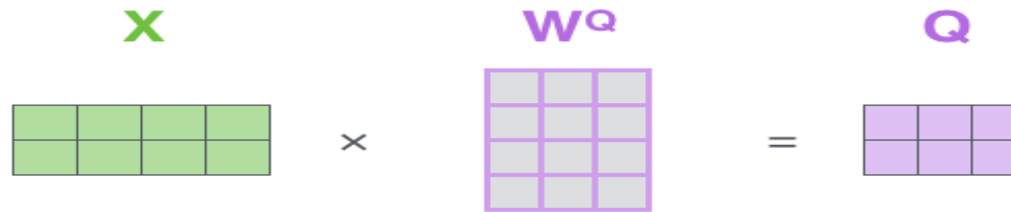
# Mécanisme de self-attention



# Mécanisme de self-attention



# Mécanisme de self-attention : formulation matricielle

$$\mathbf{X} \times \mathbf{W}^Q = \mathbf{Q}$$


$$\mathbf{X} \times \mathbf{W}^K = \mathbf{K}$$


$$\mathbf{X} \times \mathbf{W}^V = \mathbf{V}$$




# Mécanisme de self-attention : formulation matricielle

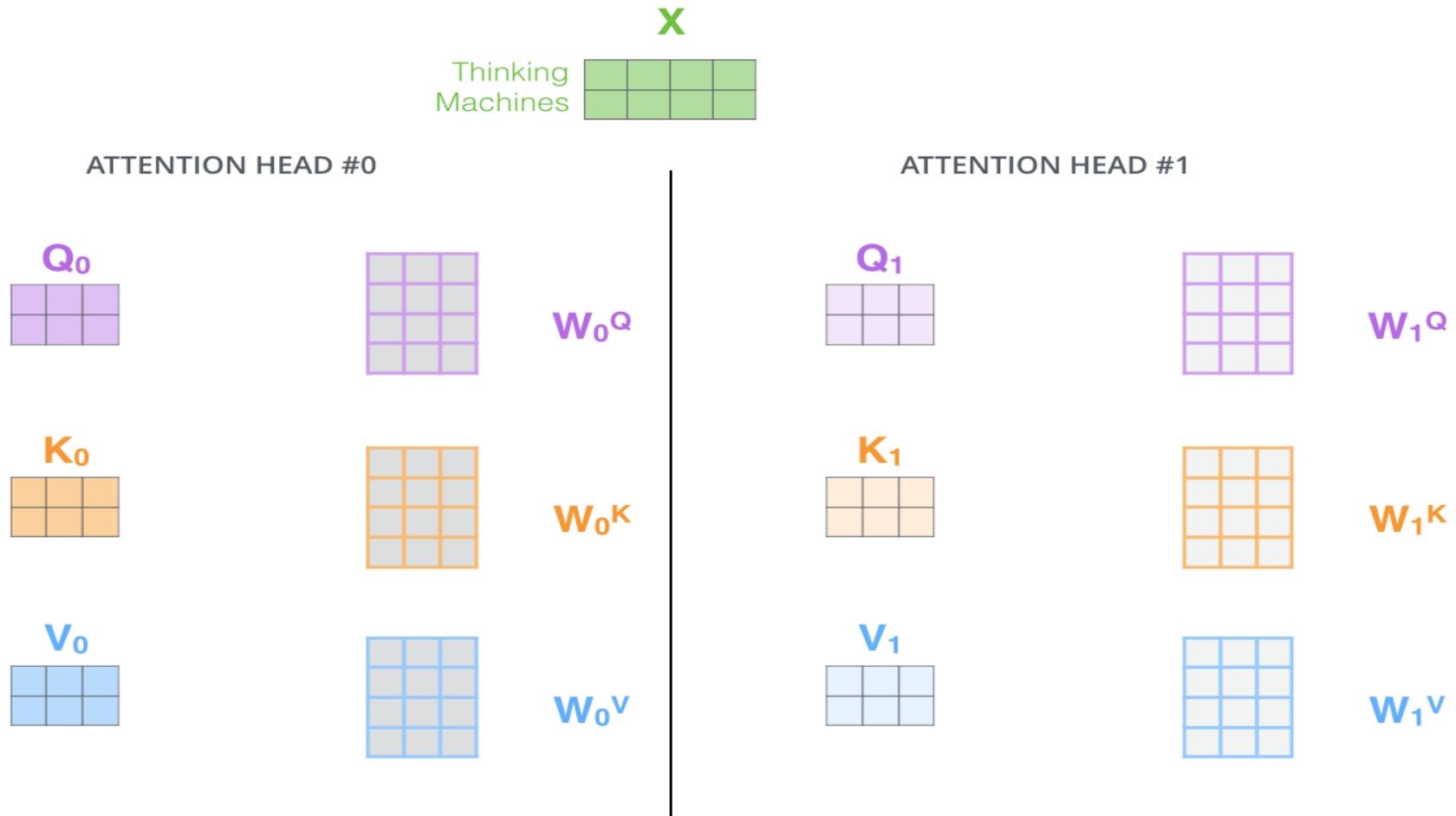
$$\text{softmax}\left(\frac{\begin{matrix} \text{Q} \\ \begin{array}{|c|c|c|} \hline & & \\ \hline & & \\ \hline \end{array} \end{matrix} \times \begin{matrix} \text{K}^T \\ \begin{array}{|c|c|} \hline & \\ \hline & \\ \hline & \\ \hline \end{array} \end{matrix}\right) \begin{matrix} \text{V} \\ \begin{array}{|c|c|c|} \hline & & \\ \hline & & \\ \hline \end{array} \end{matrix}$$

=

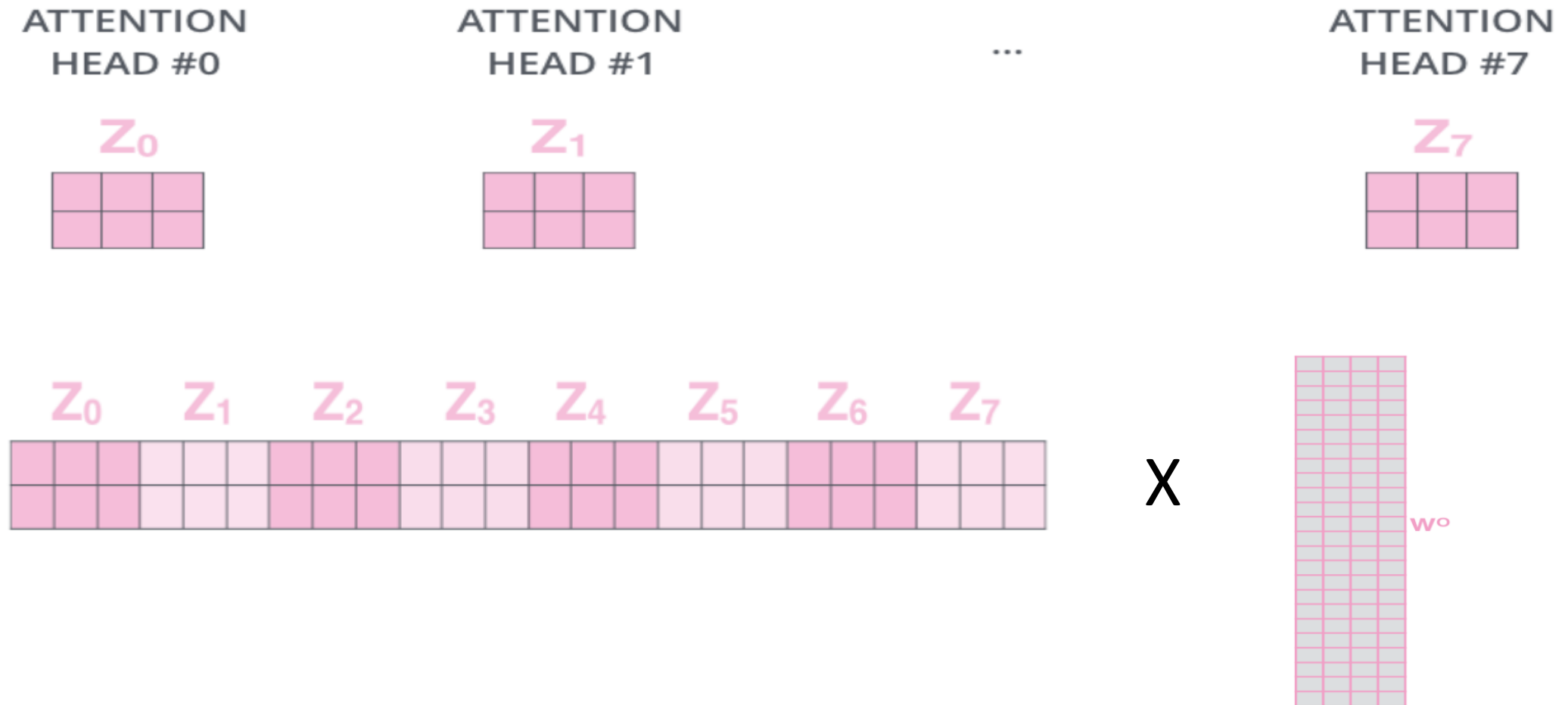
$\text{Z}$

$\begin{array}{|c|c|c|} \hline & & \\ \hline & & \\ \hline \end{array}$

# Multihead Attention



# Multihead Attention



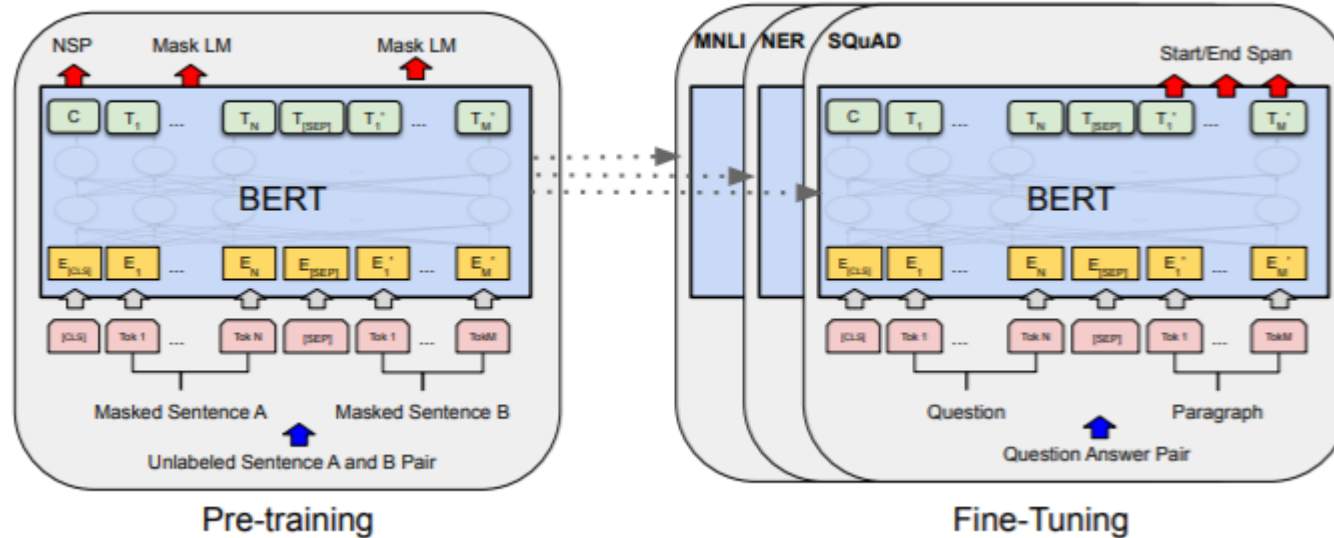
# Apprentissage auto-supervisé

- Limites de l'apprentissage supervisé : le deep learning est très gourmand en données pour l'entraînement des modèles. L'annotation des données est rédhibitoire dans certains cas.
- L'apprentissage supervisé exploite « l'annotation naturelle » des données.

## L'étudiant a ouvert son []

- Le modèle est entraîné en prédisant des mots masqués aléatoires dans le corpus. Cet apprentissage est particulièrement adapté aux données textuelles.

# BERT : pre-training et fine-tuning



BERT a été pré-entraîné sur le BookCorpus (800 millions de mots) et English Wikipedia (2,500 millions de mots).

Deux modèles :

- BERT Base: 12 couches avec 110 millions de paramètres.
- BERT Large: 24 couches avec 340 millions de paramètres.

# GPT-3 : few shot learner

## The three settings we explore for in-context learning

### Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1 Translate English to French: ← task description
2 cheese => ..... ← prompt
```

### One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1 Translate English to French: ← task description
2 sea otter => loutre de mer ← example
3 cheese => ..... ← prompt
```

### Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1 Translate English to French: ← task description
2 sea otter => loutre de mer ← examples
3 peppermint => menthe poivrée ←
4 plush girafe => girafe peluche ←
5 cheese => ..... ← prompt
```

## Traditional fine-tuning (not used for GPT-3)

### Fine-tuning

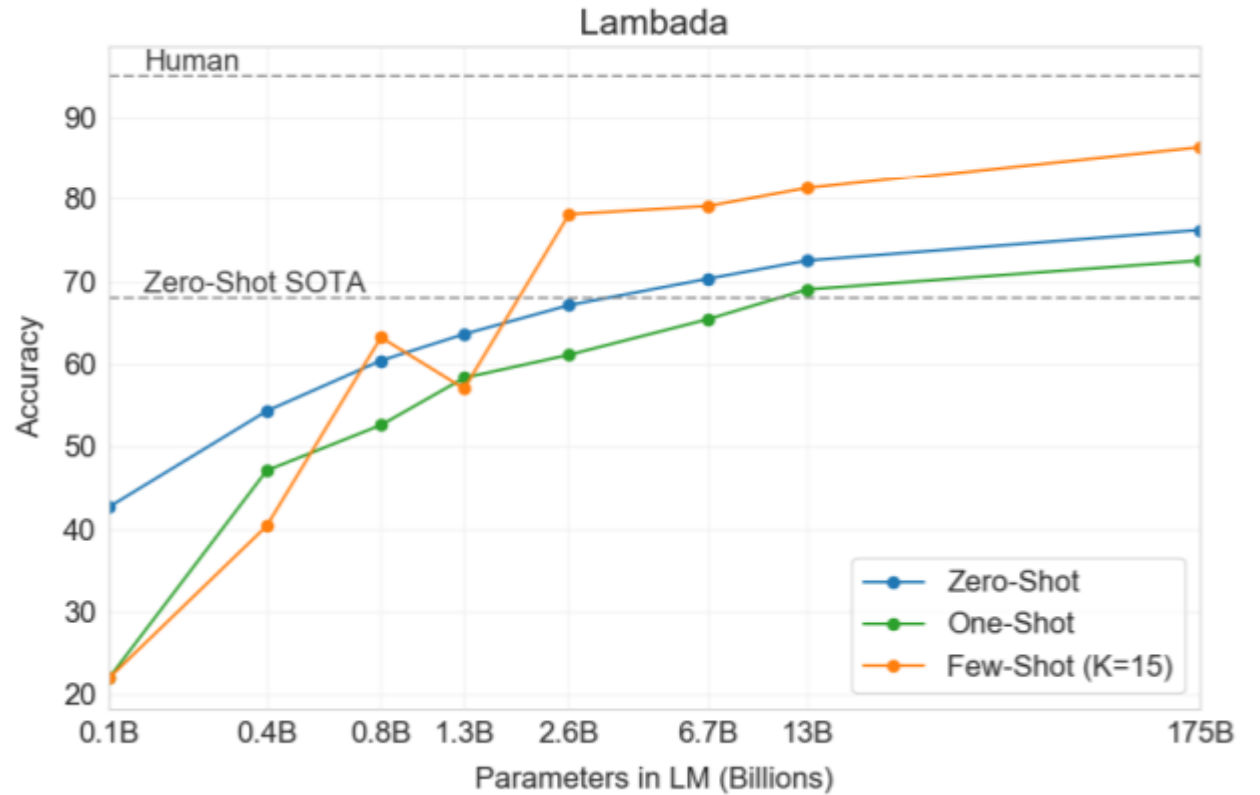
The model is trained via repeated gradient updates using a large corpus of example tasks.



# GPT-3 : architecture et performance

Le modèle a été entraîné sur Common Crawl dataset et WebText2, Books1, Books2 et Wikipedia, pour un total de près de 500 milliards de tokens.

Le modèle GPT-3 le plus large possède **96 couches** et **175 billion paramètres**.



## GPT-3 : exemple de génération de texte

----- Generated Poem 1 -----

I must have shadows on the way  
If I am to walk I must have  
Each step taken slowly and alone  
To have it ready made

And I must think in lines of grey  
To have dim thoughts to be my guide  
Must look on blue and green  
And never let my eye forget  
That color is my friend  
And purple must surround me too

The yellow of the sun is no more  
Intrusive than the bluish snow  
That falls on all of us. I must have  
Grey thoughts and blue thoughts walk with me  
If I am to go away at all.



Merci de votre attention