

# Introduction

---

Clément Romac (Hugging Face & Inria)

[clement.romac@gmail.com](mailto:clement.romac@gmail.com)

[https://github.com/ClementRomac/Teaching/tree/main/ENSC3A\\_LLMs\\_2025-2026](https://github.com/ClementRomac/Teaching/tree/main/ENSC3A_LLMs_2025-2026)

# Pourquoi moi ?

- 3e année de doctorat Inria / Hugging Face
  - *Towards autonomous LLM agents with curiosity-driven RL*
  - => beaucoup de (petits) LLM finetunés
  - => beaucoup de RL appliqué au texte
  - => beaucoup d'ingénierie...

# Pourquoi moi ?

- 3e année de doctorat Inria / Hugging Face
  - *Towards autonomous LLM agents with curiosity-driven RL*
  - => beaucoup de (petits) LLM finetunés
  - => beaucoup de RL appliqué au texte
  - => beaucoup d'ingénierie...
- Avant ?
  - Formation initiale en informatique (puis Maths/Info)
  - Data Scientist/ML engineer quelques temps (industrie)
  - Ingénieur de recherche quelques temps (académie)

---

# Pourquoi moi ?

- 3e année de doctorat Inria / Hugging Face
  - *Towards autonomous LLM agents with curiosity-driven RL*
  - => beaucoup de (petits) LLM finetunés
  - => beaucoup de RL appliqué au texte
  - => beaucoup d'ingénierie...
- Avant ?
  - Formation initiale en informatique (puis Maths/Info)
  - Data Scientist/ML engineer quelques temps (industrie)
  - Ingénieur de recherche quelques temps (académie)
- Ce que je ne suis pas:
  - un expert en prompting
  - un utilisateur régulier de LLMs
  - un expert dans tous les détails des LLMs de 2025

# Pourquoi moi ?

- 3e année de doctorat Inria / Hugging Face
  - *Towards autonomous LLM agents with curiosity-driven RL*
  - => beaucoup de (petits) LLM finetunés
  - => beaucoup de RL appliqué au texte
  - => beaucoup d'ingénierie...
- Avant ?
  - Formation initiale en informatique (puis Maths/Info)
  - Data Scientist/ML engineer quelques temps (industrie)
  - Ingénieur de recherche quelques temps (académie)
- Ce que je ne suis pas:
  - un expert en prompting
  - un utilisateur régulier de LLMs
  - un expert dans tous les détails des LLMs de 2025

**Objectif:** vous transmettre des informations qui vont durer, dans un domaine où tout change très (très) rapidement

# Pourquoi vous ?



---

# Pourquoi ce cours ?

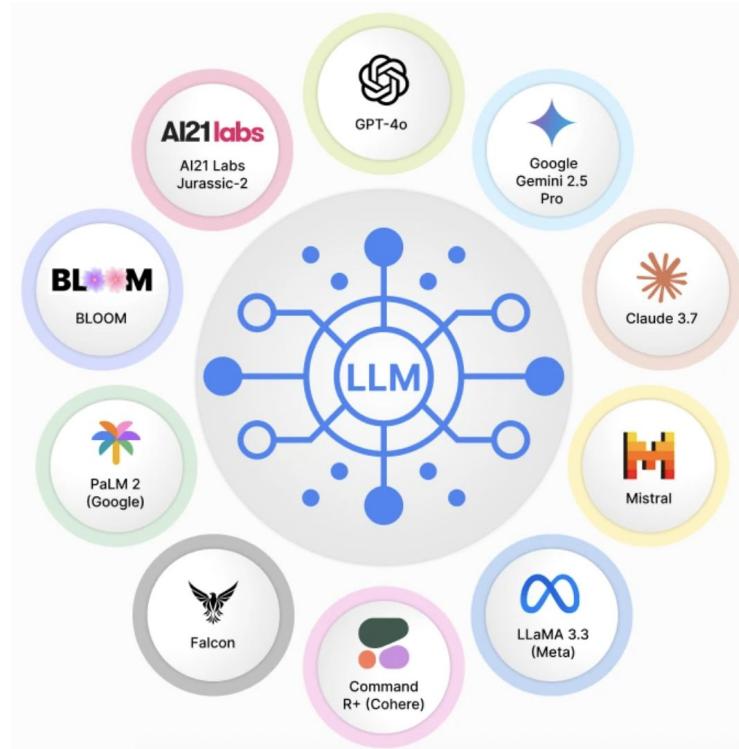
# A cause de ChatGPT bien sûr !



# A cause de ChatGPT bien sûr !

Plus généralement:

- Nous sommes passés d'outils réservés aux experts à des **outils accessibles à tout le monde**
- Notamment grâce à l'utilisation du langage !



# A cause de ChatGPT bien sûr !

Plus généralement:

- Nous sommes passés d'outils réservés aux experts à des **outils accessibles à tout le monde**
- Notamment grâce à l'utilisation du langage !
- A ouvert la porte à plus que des chatbots avec l'idée de **prompt**



# Parce que ça a changé le domaine

- L'arrivée des **Transformers (2017)** a fait aux benchmarks de NLP ce que les CNNs (2012) avaient fait à ImageNet...
- Il y a désormais des Transformers (presque) partout en NLP

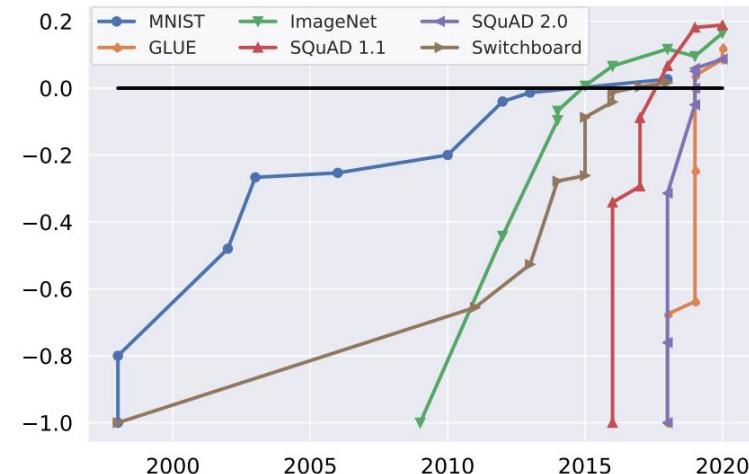


Figure 1: Benchmark saturation over time for popular benchmarks, normalized with initial performance at minus one and human performance at zero.

(Kiela et al., 2021)

# Parce que ça a changé le domaine

- L'arrivée des **Transformers (2017)** a fait aux benchmarks de NLP ce que les CNNs (2012) avaient fait à ImageNet...
- Il y a désormais des Transformers (presque) partout en NLP



Figure 1: Benchmark saturation over time for popular benchmarks, normalized with initial performance at minus one and human performance at zero.

(Kiela et al., 2021)

# Parce que vous risquez de vous en servir...

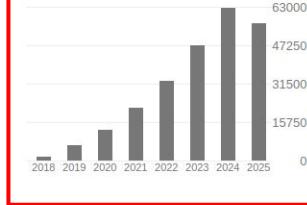
≡ Google Scholar



**Ashish Vaswani**  
Startup  
Verified email at fastmail.com  
Deep Learning

[FOLLOW](#)

Cited by	All	Since 2020
Citations	244090	234050
h-index	47	47
i10-index	70	62

 A bar chart showing the number of citations per year from 2018 to 2025. The y-axis ranges from 0 to 63,000. The chart shows a steady increase in citations over time.

TITLE	CITED BY	YEAR
<a href="#">Attention is all you need</a> A Vaswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, AN Gomez, ... Advances in neural information processing systems 30	213624	2017
<a href="#">Relational inductive biases, deep learning, and graph networks</a> PW Battaglia, JB Hamrick, V Bapst, A Sanchez-Gonzalez, V Zambaldi, ... arXiv preprint arXiv:1806.01261	4688	2018
<a href="#">Attention is all you need. arXiv 2017</a> A Vaswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, AN Gomez, ... arXiv preprint arXiv:1706.03762 30	4442	2017
<a href="#">Self-attention with relative position representations</a> P Shaw, J Uszkoreit, A Vaswani arXiv preprint arXiv:1803.02155	3464	2018
<a href="#">Image transformer</a> N Parmar, A Vaswani, J Uszkoreit, L Kaiser, N Shazeer, A Ku, D Tran International conference on machine learning, 4055-4064	2441	2018
<a href="#">Stand-alone self-attention in vision models</a> P Ramachandran, N Parmar, A Vaswani, I Bello, A Levskaya, J Shlens Advances in neural information processing systems 32	1631	2019
<a href="#">Attention augmented convolutional networks</a> I Ballas, R Zouh, A Vaswani, J Shlens, OV LP	1599	2019

Public access

[VIEW ALL](#)

0 articles	3 articles
not available	available
Based on funding mandates	

---

# Déroulé du cours

# Déroulé du cours

- **Module 1 (03/11 - 04/11): Des RNNs aux Transformers**
  - Tokenization & Embeddings (NLP)
  - Rappels RNNs
  - Attention mechanism
  - Self-Attention and Transformers
- **Module 2 (07/11): LLMs**
  - Quizz/Rappels Transformers
  - Encoder-only (e.g. BERT)
  - Decoder-only (e.g. GPT)
  - Prompting
  - Chat models
- TP colab / Jupyter tout le long

---

# Déroulé du cours

- Objectif du cours: vous donner les clés pour comprendre comment ça marche, à vous d'aller plus loin si vous voulez

---

## Evaluation:

- ~~Quelques questions au QCM~~

- Feedback apprécié !
  - Formulaire anonyme à la fin du cours
- Tout est sur mon site GH:  
[https://github.com/ClementRomac/Teaching/tree/main/ENSC3A\\_LLMs\\_2025-2026](https://github.com/ClementRomac/Teaching/tree/main/ENSC3A_LLMs_2025-2026)

---

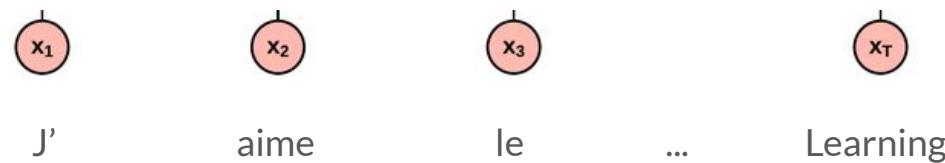
# Un peu de Natural Language Processing

---

# Tokenization

Objectif global: utiliser du texte comme entrée

Etape 1: Passer d'une séquence de mots à une séquence de symboles connus

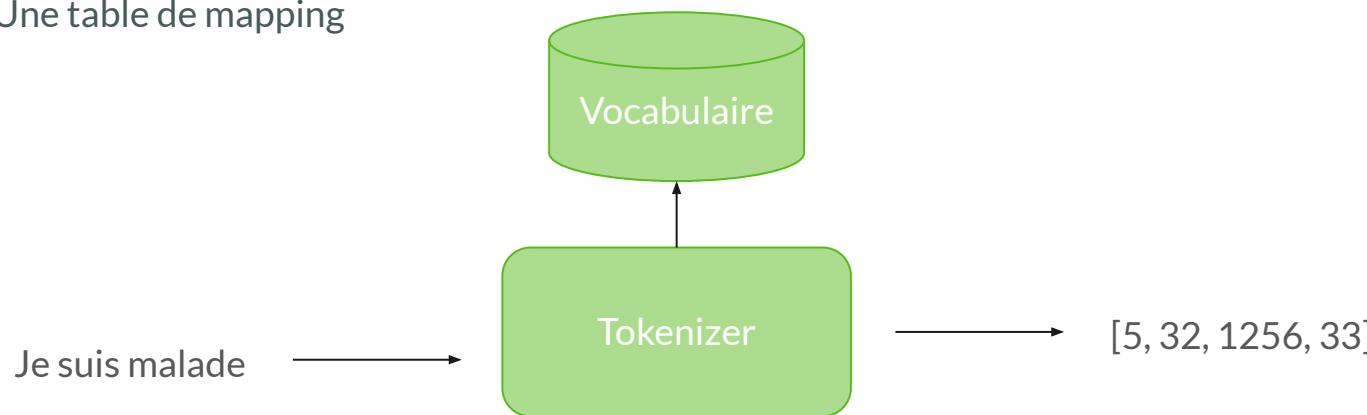


---

# Tokenization

## Tokenizer:

- Un “vocabulaire” de symboles
- Une table de mapping



---

# Tokenization

Word-level mapping:

Je suis malade → ["Je": 5, "suis": 32, "malade": 1256]

=> Vocabulaire relativement petit, aucun partage de racine

---

# Tokenization

## Word-level mapping:

Je suis malade → ["Je": 5, "suis": 32, "malade": 1256]

=> Vocabulaire relativement petit, aucun partage de racine

## Character-level mapping:

Je suis malade → ["J": 10, "e": 5, "s": 18, ...]

=> Vocabulaire très petit mais peu informatif

# Tokenization

## Word-level mapping :

Je suis malade → ["Je": 5, "suis": 32, "malade": 1256]

=> Vocabulaire relativement petit, aucun partage de racine

## Character-level mapping :

Je suis malade → ["J": 10, "e": 5, "s": 18, ...]

=> Vocabulaire très petit mais peu informatif

## Tokenizers aujourd'hui utilisés :

- WordPiece (*Schuster et al., 2012*)
  - BERT
- Byte Pair Encoding (*Sennrich et al., 2018*)
  - GPT
- SentencePiece (*Kudo et al., 2018*)
  - T5, Llama

Note: GPT-2 -> 50k tokens

---

# Tokenization

## Tokens spéciaux:

Je suis malade → [“</s>”: 34, “Je”: 5, “suis”: 32, “malade”: 1256, “<s>”]

- <pad> => pad (+ mask) pour avoir des batchs de même taille
- </s> => début de séquence
- <s> => fin de séquence
- <unk> => symbole inconnu (hors de la table)
- ...

=> Dépend du tokenizer

# Tokenization



TP: Partie 1

## Tokens spéciaux:

Je suis malade → [“</s>”: 34, “Je”: 5, “suis”: 32, “malade”: 1256, “<s>”]

- <pad> => pad (+ mask) pour avoir des batchs de même taille
- </s> => début de séquence
- <s> => fin de séquence
- <unk> => symbole inconnu (hors de la table)
- ...

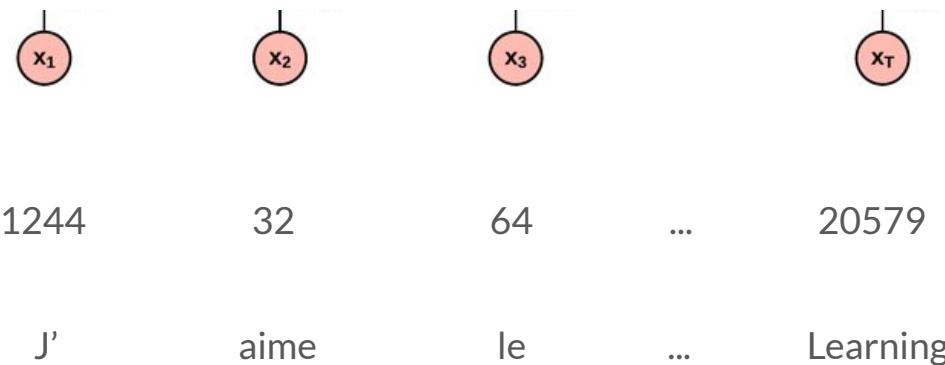
=> Dépend du tokenizer

---

# Word Embeddings

Objectif global: utiliser du texte comme entrée

Etape 2: Passer d'une séquence de tokens à une séquence de vecteurs



# Word Embeddings

Objectif global: utiliser du texte comme entrée

=> Vecteurs sparses

=> Aucun intérêt sémantique

Etape 2: Passer d'une séquence de tokens à une séquence de vecteurs

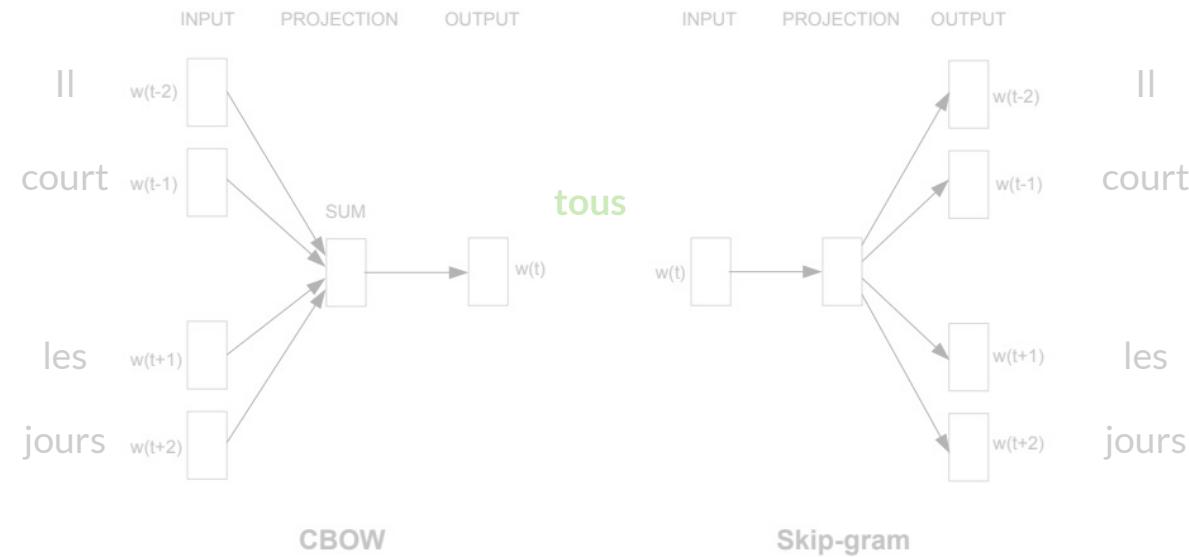
Solution naïve:  
OneHotEncoding

				
[0, 0, ..., 1, 0, ...]	[0, 0, ..., 1, 0, ...]	[0, 0, ..., 1, 0, ...]	[0, 0, ..., 1, 0, ...]	[0, 0, ..., 1, 0, ...]
1244	32	64	...	20579
J'	aime	le	...	Learning

# Word2Vec (*Mikolov et al., 2013*)

Intuition: Apprendre à représenter un mot à partir de son contexte

- 1) On utilise un vecteur **OneHotEncoding** pour chaque mot
- 2) Deux approches:

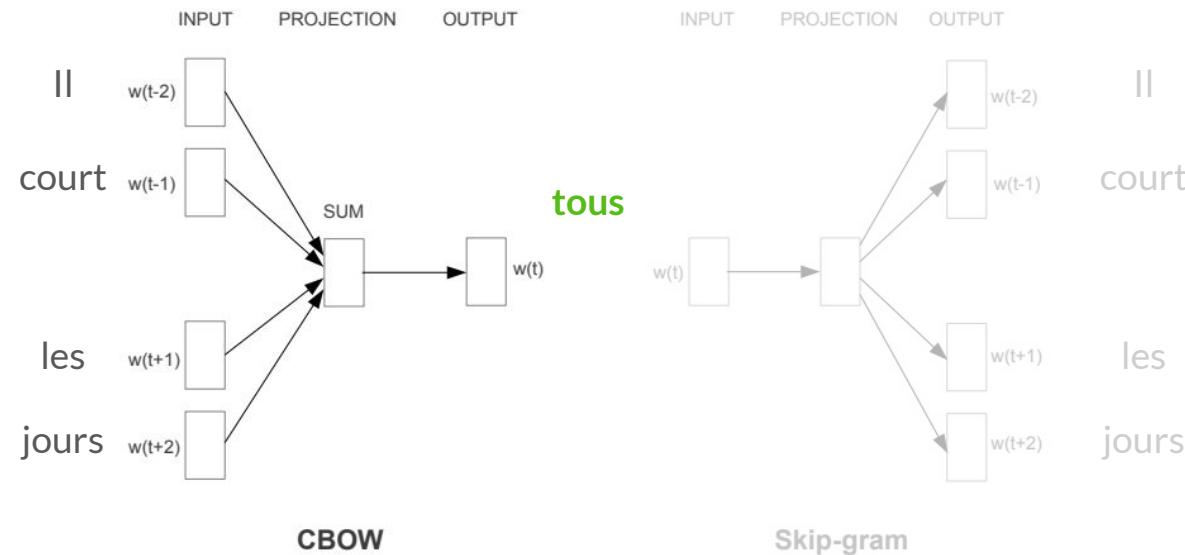


# Word2Vec (*Mikolov et al., 2013*)

Intuition: Apprendre à représenter un mot à partir de son contexte

- 1) On utilise un vecteur **OneHotEncoding** pour chaque mot
- 2) Deux approches:

**CBOW:** On passe chaque mot du contexte dans une **couche linéaire partagée**, on fait la moyenne de tous les vecteurs et on **prédit** le mot attendu

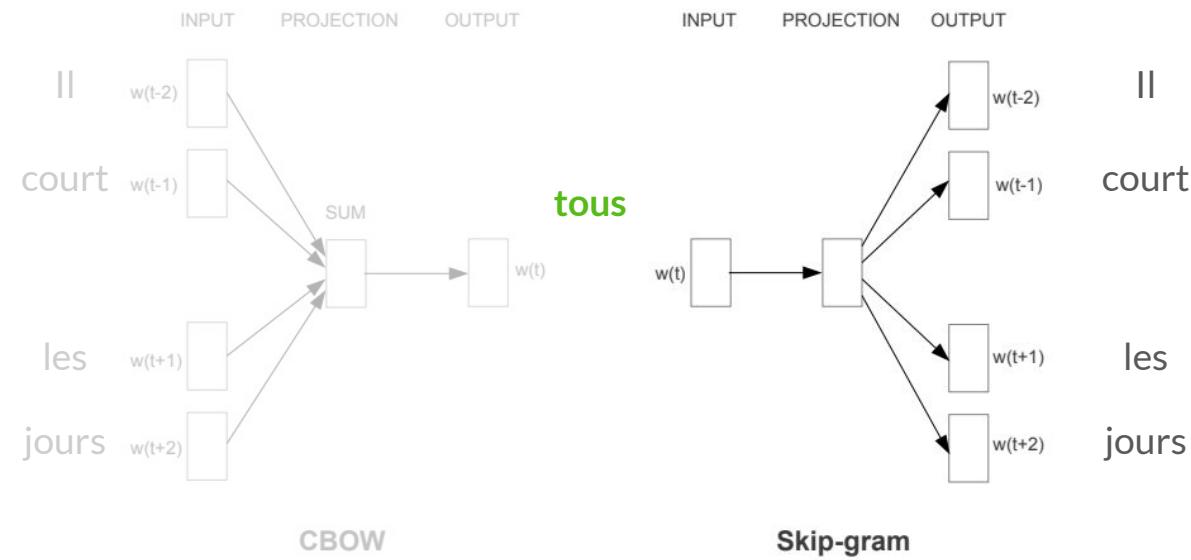


# Word2Vec (*Mikolov et al., 2013*)

Intuition: Apprendre à représenter un mot à partir de son contexte

- 1) On utilise un vecteur **OneHotEncoding** pour chaque mot
- 2) Deux approches:

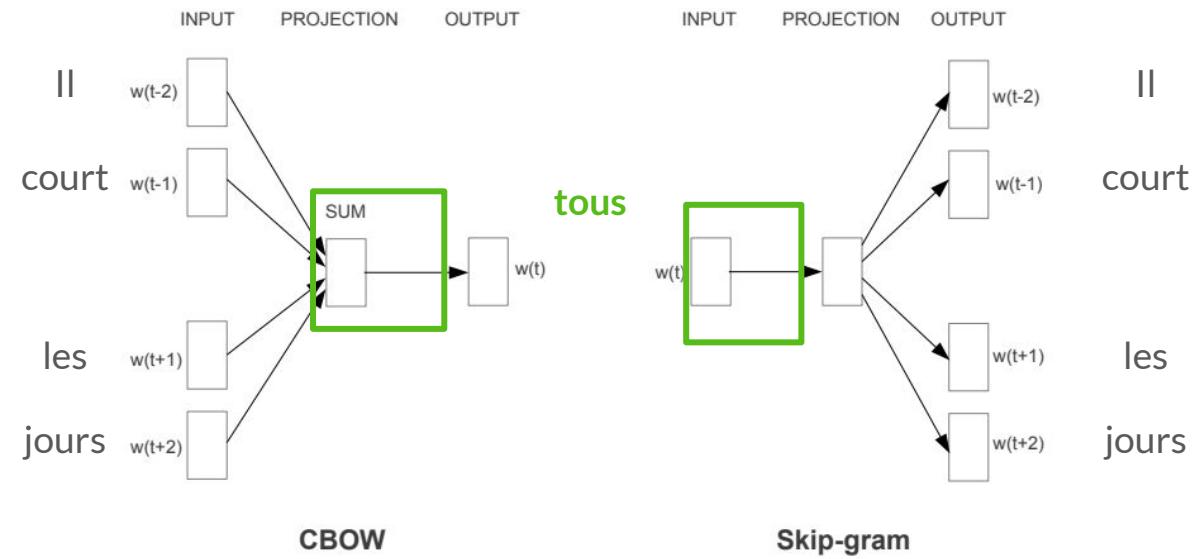
**Skip-gram:** On passe le mot principal dans une couche linéaire, on essaie de prédire chacun des mots du contexte



# Word2Vec (*Mikolov et al., 2013*)

Intuition: Apprendre à représenter un mot à partir de son contexte

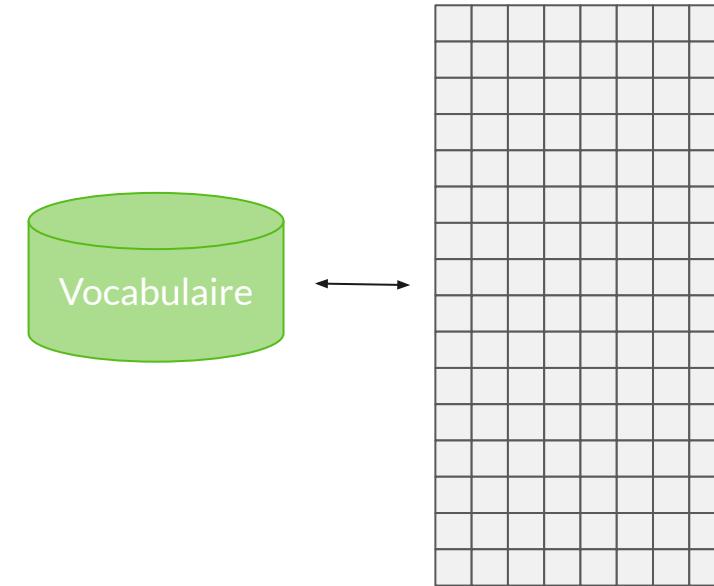
- 1) On utilise un vecteur **OneHotEncoding** pour chaque mot
- 2) Deux approches
- 3) On retient le vecteur obtenu avec le mot principal associé => **lookup table**



# Embedding lookup

- On a donc une **table** associant chaque **token** à son **embedding**
- On peut **utiliser ces embedding** en entrée pour entraîner notre modèle

Lookup Table



# Word2Vec + Language Model

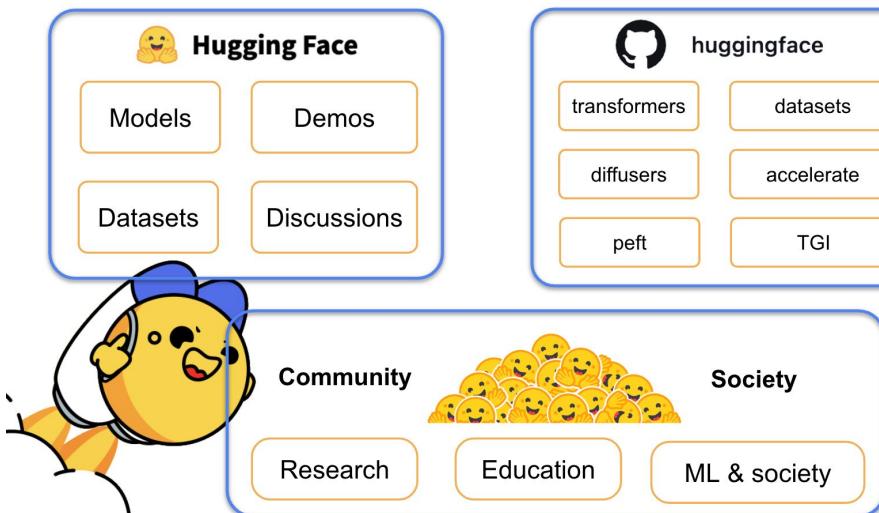
Dans ce cours (et généralement):

On initialise la table aléatoirement et les  
embeddings sont appris en même temps que le  
modèle

---

# Outils open-source

# Outils Open-Source *(non exhaustif)*



# Outils Open-Source *(non exhaustif)*



spaCy



SentencePiece