

# CADRES D'APPRENTISSAGE POUR L'EXTRACTION D'INFORMATION

Vincent Guigue & Rémy Découpes  
[vincent.guigue@agroparistech.fr](mailto:vincent.guigue@agroparistech.fr)



# Plusieurs focus thématiques

- 1** La normalisation des entités nommées
- 2** La construction de corpus étiqueté
- 3** La modélisation des graphes de connaissances

# NORMALISATION



# Normalisation: entités + relations $\Rightarrow$ insuffisant

## Normalisation = Entity Linking

Mention extraite  $\Leftrightarrow$  entrée unique dans une base de connaissances (KB)

Le **premier ministre** termine sa déclaration à **Paris**

- "Paris"  $\rightarrow$  **Paris**, France (Wikidata: Q90), plutôt que **Paris**, Texas.
- "premier ministre"  $\rightarrow$  ...

### Dictionnaire, ontologie, thésaurus

MeSH, UMLS, Wikidata

[tâche critique en médecine, biologie, ...]

++ ambiguïté

### Similarité

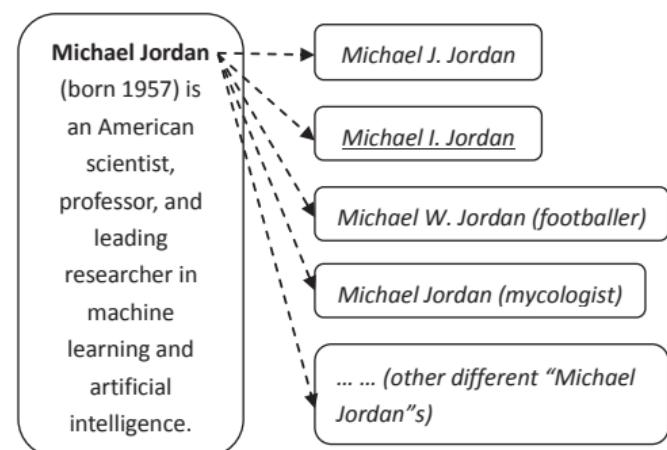
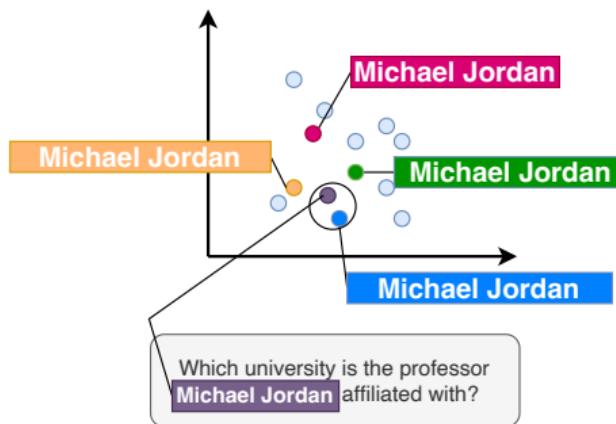
- Lexicale (souvent suffisant )
- Contexte (TF-IDF, BM25)
- Moteur de recherche (autorité)
- Embeddings

$\Rightarrow$  Représenter l'entité + la référence  $\Rightarrow$  + Distance minimale = match

# Normalisation: entités + relations $\Rightarrow$ insuffisant

Normalisation = Entity Linking

Mention extraite  $\Leftrightarrow$  entrée unique dans une base de connaissances (KB)



Répondre à la question:

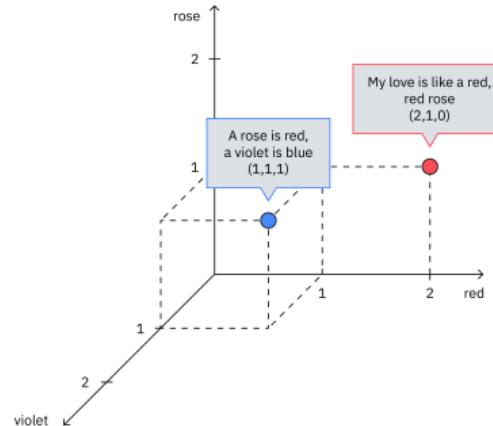
*Which university is the professor Michael Jordan affiliated with?*

$\Rightarrow$  Trouver le bon *Michael Jordan*  $\Rightarrow$  désambiguisation de la question

# Similarité tf-idf

- Entité : vecteur mono-valeur ou phrase
- Référence : mot, mot + synonymes, mot + description/définition  
(e.g. définition d'un ontologie, page wikipedia d'un terme)

Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions, [IEEE Trans. KDE 2025](#), Shen et al.



tf-idf

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

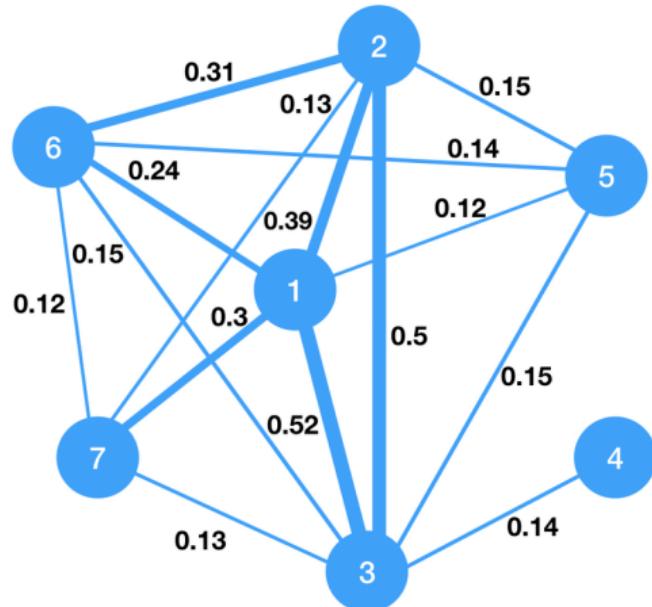
$$IDF(t, D) = \log_e \frac{\text{(Total number of documents in the corpus)}}{\text{(Number of documents with term } t \text{ in them)}}$$

$$TF-IDF(t, d, D) = TF(t, d) \times IDF(t, D)$$

# Lissage sur graphe

Le score tf-idf permet de construire un graphe :

- Entre les termes dans le **Dictionnaire**
  - Entre les phrases dans le **document**
- ⇒ Possibilité de lisser les représentations



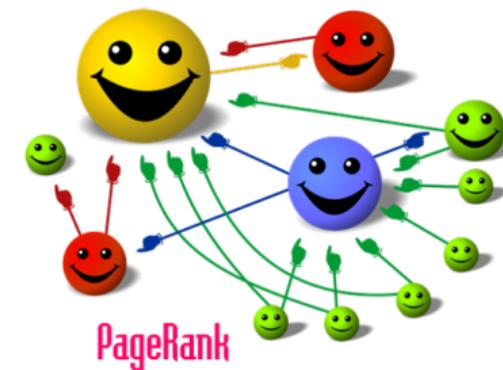
# Information Retrieval

- Corpus = Dictionnaire / thésaurus
- Requête = entité

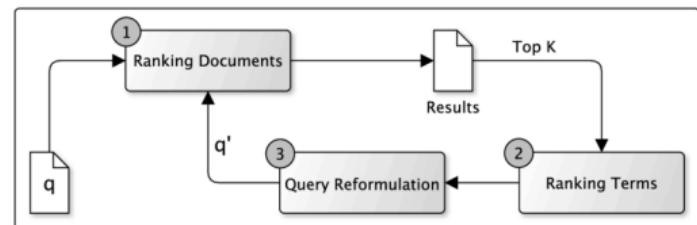
Recherche lexicale + Métrique BM25 (proche de tf-idf)

Il est possible d'ajouter un score d'autorité:

- Fréquence d'apparition des entités (e.g. Paris, France vs Paris, Texas)
- PageRank

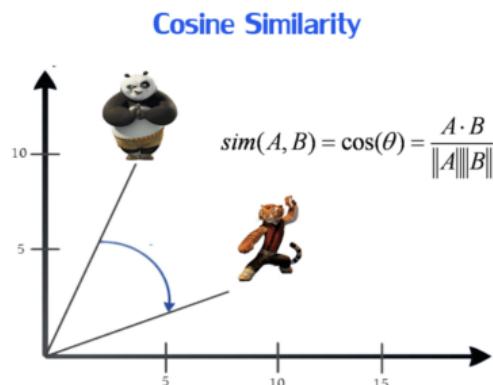


Historiquement: stratégie d'élargissement des requêtes type *pseudo-relevance-feedback*



# Embedding / plongement

- Similarité cosinus
- Avec ou sans contextualisation
  - Mot (dans un contexte)
- Agrégation type CLS
  - Phrase d'origine
  - Définition du dictionnaire



Contextual representation



LLM

Token representation



**Aspirin** : medication that relieves pain, reduces fever and inflammation

# Graph Neural Network & normalisation

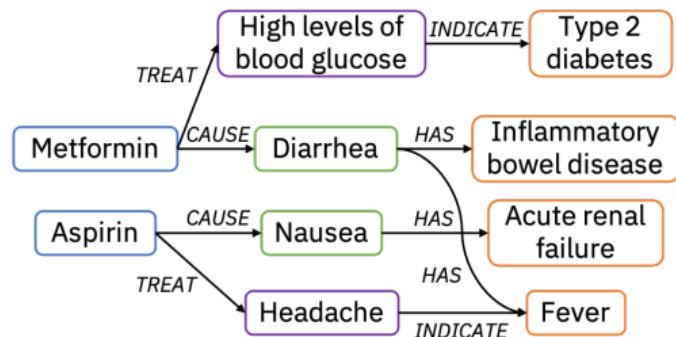
GNN = représentation de nœuds, repr. d'arc, repr. de graphe entier

- Message passing, local/global pooling

Application sur un graphe de connaissances

*Hyp:* la représentation de chaque noeud sera meilleure en prenant en compte le contexte du noeud

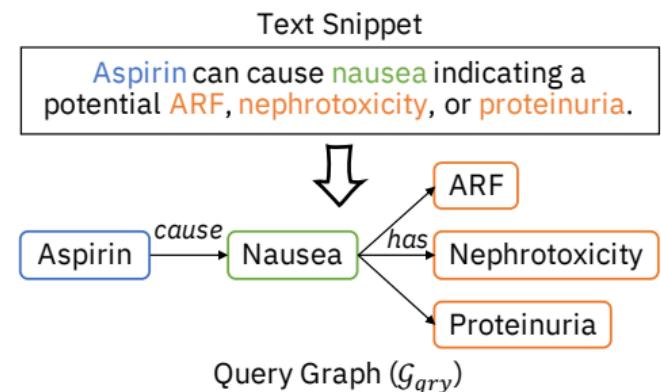
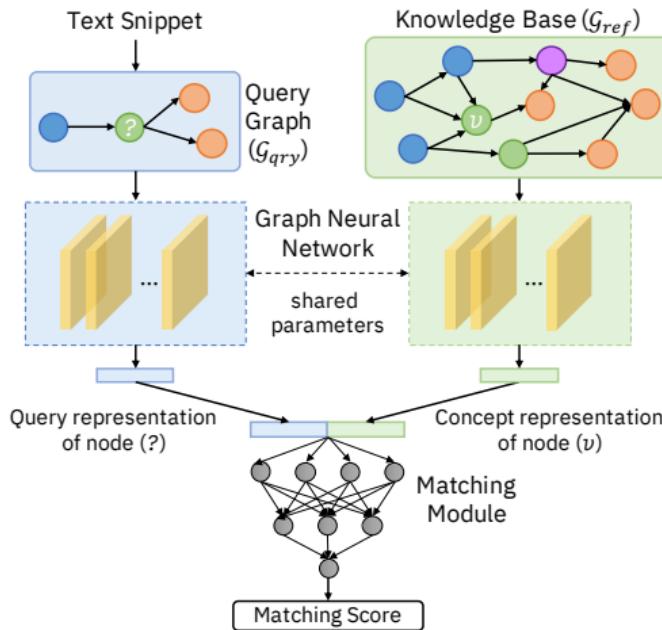
Medical Entity Disambiguation Using Graph Neural Networks, [ACM SIGMOD 2021, Vretinaris et al.](#)



# Graph Neural Network & normalisation

GNN = représentation de nœuds, repr. d'arc, repr. de graphe entier

- Message passing, local/global pooling

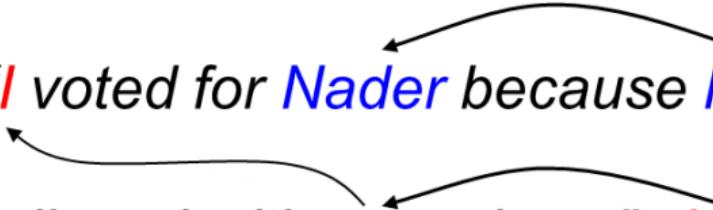


⇒ Quel apport / coût du matching module ?

# Résolution de co-références

Des problèmes connexes impactent directement la détection et la normalisation des entités

*“I voted for Nader because he was most aligned with my values,” she said.*



Approches historiques:

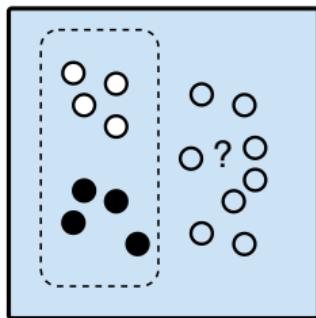
- Systèmes de règles
- HMM / CRF (Markov / Conditional Random Field) pour le machine learning sur des séquences
- LLM : modèles génératifs très forts sur la tâches (considérée comme résolue par le *Stanford NLP Group*)

# ETIQUETAGE DES DONNÉES

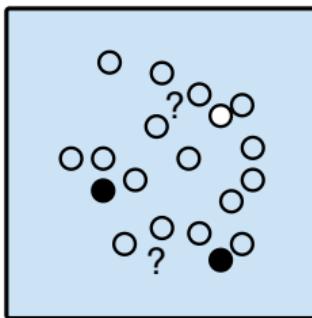


# Apprentissage supervisé

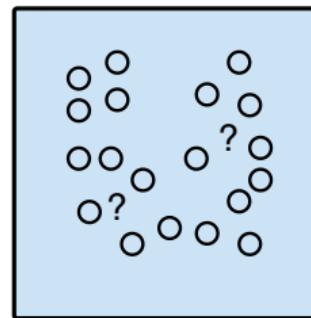
## Cadres historiques d'apprentissage



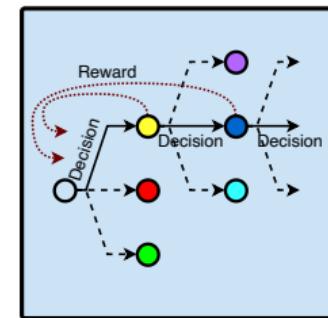
Supervised  
Learning



Semi-Supervised  
Learning

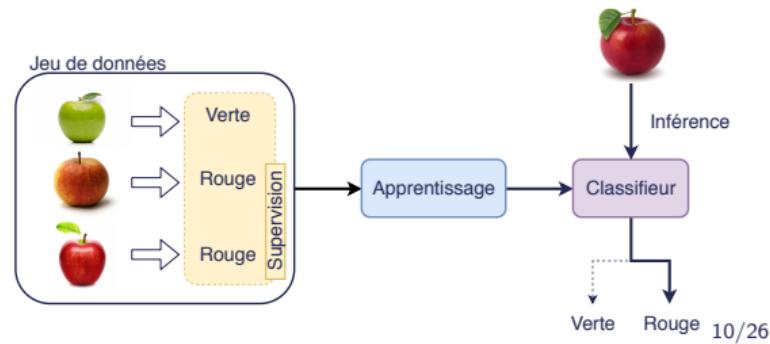


Unsupervised  
Learning



Reinforcement  
Learning

Deep learning: supervisé... Ou  
auto-supervisé  
⇒ Quid des étiquettes?





# Historique rapide en vision

VOC 2007, caltech256

- milliers d'images, dizaines de catégories

ImageNet 2009:

- 3.5M d'images annotées, 1000 catégories (moins de 0.3% d'erreur)

Construction sur plusieurs années : objectif 50M images

Localisation pour 1.2M d'objets (2012)

Google car

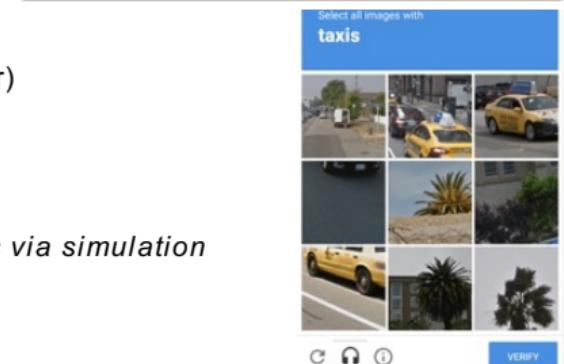
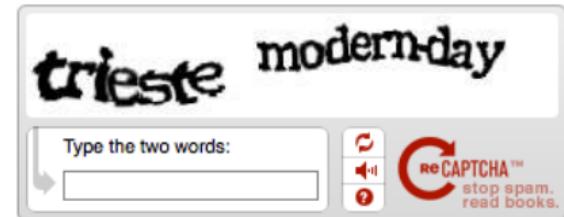
- 2018: *5 million miles on public roads and more than 5 billion miles via simulation*

Mobileye

- 500 personnes au Sri-Lanka: étiquetage scènes/objets

- Réduire les modèles pour les embarquer dans une caméra

(Google) Captcha



L'augmentation de la puissance de calcul et des tailles de bases de données expliquent les succès actuels

# Les sources déjà disponibles

- Page de désambiguisation de wikipedia + lien vers les sources

## Michael Jordan (homonymie)

[Article](#)[Discussion](#)[Lire](#)[Modifier](#)[Modifier le code](#)[Voir l'historique](#)[Outils](#)[丈 A 8 langues](#)

- *Cette page d'homonymie répertorie différentes personnes portant le même nom et le même prénom.*
- *Pour les articles homonymes, voir [Jordan](#).*

**Michael Jordan** est un nom de personne

notamment porté par :

- [Michael Jordan \(1963-\)](#), un joueur américain de basket-ball ;
- [Michael Jordan \(1986-\)](#), un joueur anglais de football ;
- [Michael Jordan](#), un homme politique irlandais.

**Mike** est un nom de personne notamment porté par :

- [Mike Jordan \(1958-\)](#), un pilote automobile anglais.

## Voir aussi

[\[ modifier \]](#) [\[ modifier le code \]](#)

- [Michael B. Jordan \(1987-\)](#), un acteur américain
- [Michael I. Jordan \(1956-\)](#), un chercheur américain
- [Michael-Hakim Jordan \(1977-\)](#), un joueur américain de basket-ball

Sur les autres projets Wikimedia :

• [Michael Jordan](#), sur Wikiquote

Hypothèse: sur chacune des pages cibles, les entités *Michael Jordan* correspondent à la page en question

Exemple de corpus disponible basé sur ces pages:

Interactive Query Clarification and Refinement via User Simulation, [SIGIR 2022, Erbacher et al.](#)

# Les sources déjà disponibles

## ■ Classification d'opinion: les sites d'avis en ligne



### Bazza! Aussie IPA

American IPA | 6% ABV

Lucky Envelope Brewing in Seattle, Washington

Reviewed by mactrall from Washington

**3.7/5** rDev +2.8% | Average: 3.6

look: 3.75 | smell: 3.75 | taste: 3.75 | feel: 3.75 | overall: 3.5

Slight haze on the pale amber brew and plenty of foam in the Stella Artois goblet. Pleasant aroma of berries and green apricot. Taste is like a fresh hop brew with that resinous and sharp flavor. There is some fruity flavor, like blueberry, but also a green, unripe taste. Bitterness is moderate but the finish is on the mineral side. Overall this is an interesting hoppy brew with some rewards as well as challenges. From the 16 oz can purchased at Elizabeth Station. Dated 8/19/25.



Du texte, des notes, parfois même des notes sur des aspects spécifiques  
⇒ une source intéressante pour l'analyse d'opinion et de thématiques

## ■ Les news / forum déjà classés par thématiques

## ■ ...

# Pattern/motifs: l'expertise pour amorcer le système

Expert  $\Rightarrow$  motif lexical/grammatical + variantes  $\Rightarrow$  étiquetage

- L'expert formule des règles pour identifier ce qui l'intéresse

ID	Pattern Synset & Support Sets
$P_1$	$\langle \text{Politician} \rangle$ was governor of $\langle \text{State} \rangle$ A,80 B,75 C,70
$P_2$	$\langle \text{Politician} \rangle$ politician from $\langle \text{State} \rangle$ A,80 B,75 C,70 D,66 E,64
$P_3$	$\langle \text{Person} \rangle$ daughter of $\langle \text{Person} \rangle$ F,78 G,75 H,66
$P_4$	$\langle \text{Person} \rangle$ child of $\langle \text{Person} \rangle$ I,88 J,87 F,78 G,75 K,64

Table 1: Pattern Synsets and their Support Sets

- Permet un premier étiquetage
- Utilisation de patrons (regex, dépendances syntaxiques) combinés avec la KB pour annoter automatiquement.

PATTY: A Taxonomy of Relational Patterns with Semantic Types, [EMNLP 2012](#), Nakashole et al.

# Supervision distante: exploitation des bases de connaissances

- Entités :  
Base de connaissances  $\Rightarrow$  termes  $\Rightarrow$  Recherche & annotation dans le texte
- Relation : triplet  $(e_h, R, e_t) \Rightarrow$  si  $(e_h, e_t)$  dans un paragraphe alors  $R$

**Risque de bruit** : si la KB indique que *Barack Obama* est né à *Hawaï*, tout texte contenant ces deux entités est étiqueté comme instance de la relation *BornIn*.

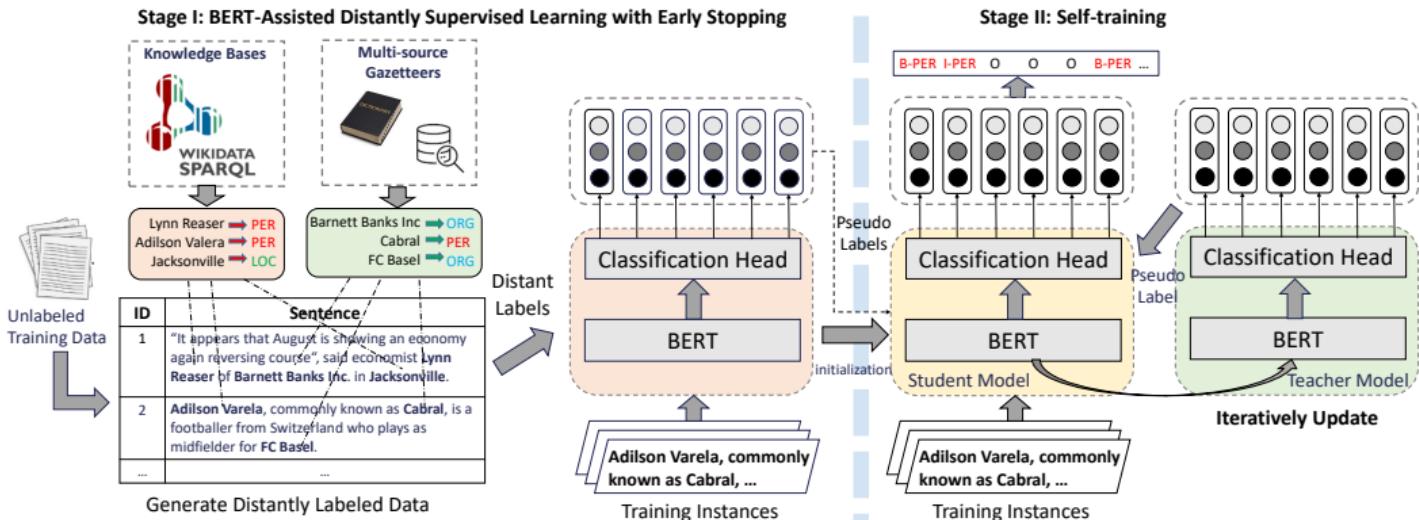
**Risque de silence** : toutes les variations d'écriture d'une entité sont oubliées (+ problème des co-références)

$\Rightarrow$  Il faut des approches/détecteurs spécifiques pour traiter des corpus étiquetés de la sorte.



# Bond : une méthode populaire pour la supervision distante

- Coopération d'un modèle *Teacher* et d'un second *Student* pour intégrer les conjectures du premier

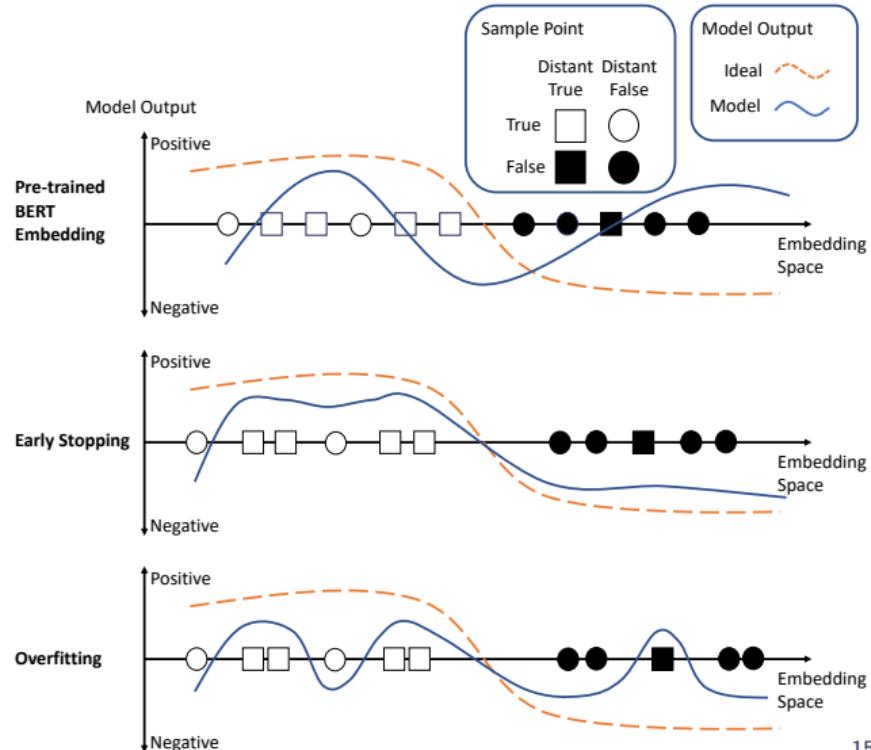




# Bond : une méthode populaire pour la supervision distante

- Coopération d'un modèle *Teacher* et d'un second *Student* pour intégrer les conjectures du premier

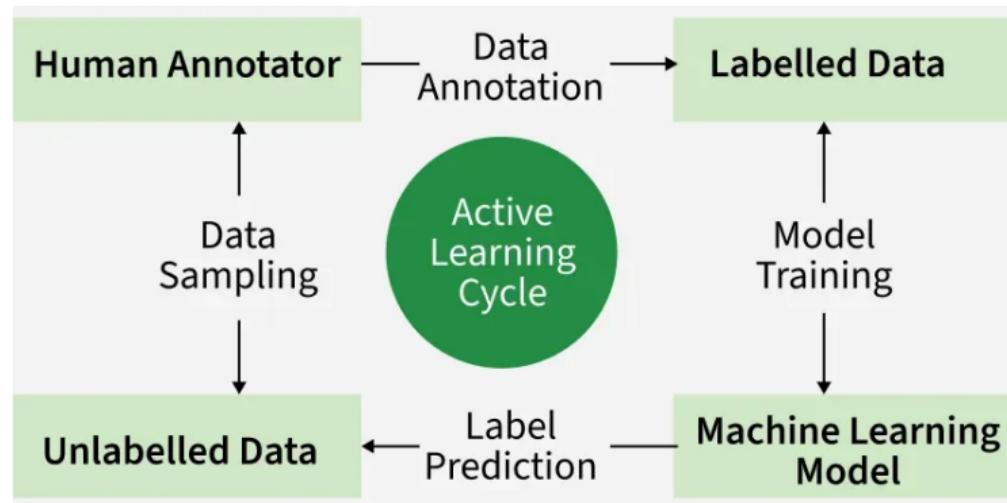
Tenter d'infléchir la courbe de décision sans tomber dans le sur-apprentissage





# Active learning

Garder l'humain dans la boucle pour améliorer le système en toute confiance



Quels exemples présenter à l'utilisateur?

- Les plus ambigus
- Les plus représentatif d'une classe
- Ceux maximisant un critère statistique



# Crowdsourcing

- L'IA moderne (post 2012) repose sur de vastes corpus annotés par des humains
  - ChatGPT et consort (post 2022) ont accéléré le mouvement!
- ⇒ On parle des *travailleurs du clic*

Amazon mechanical turk

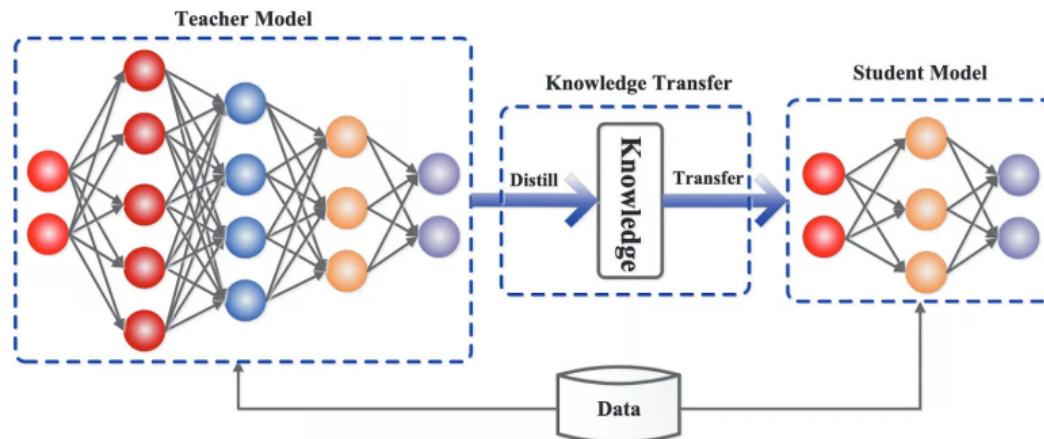


Possibilité d'annoter des corpus larges sur des tâches complexes avec de la redondance entre les annotateurs



# Distillation LLM

*Hypothèse:* les très gros LLM ont acquis une véritable expertise dans certains domaines... Ils seraient en mesure d'étiqueter des corpus



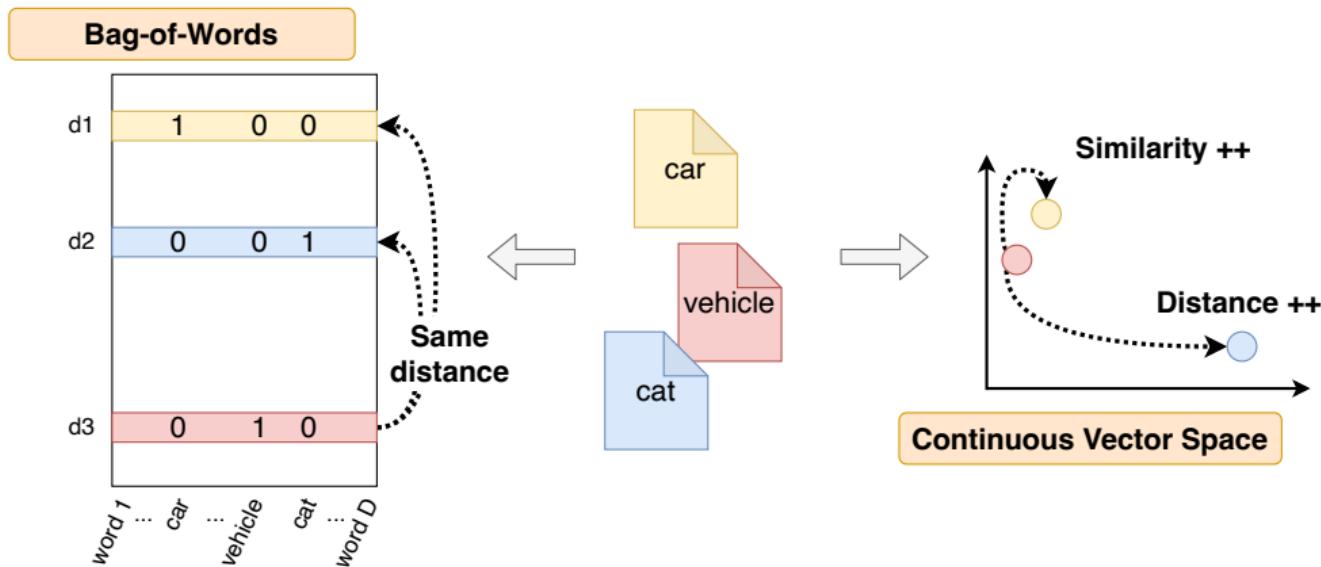
- Très peu coûteux
- Risque de propagation des erreurs, excès de confiance dans les LLM?

# REPRÉSENTATION DES CONNAISSANCES DANS UN ES- PACE LATENT



# Repésentation latente de texte / base de connaissances

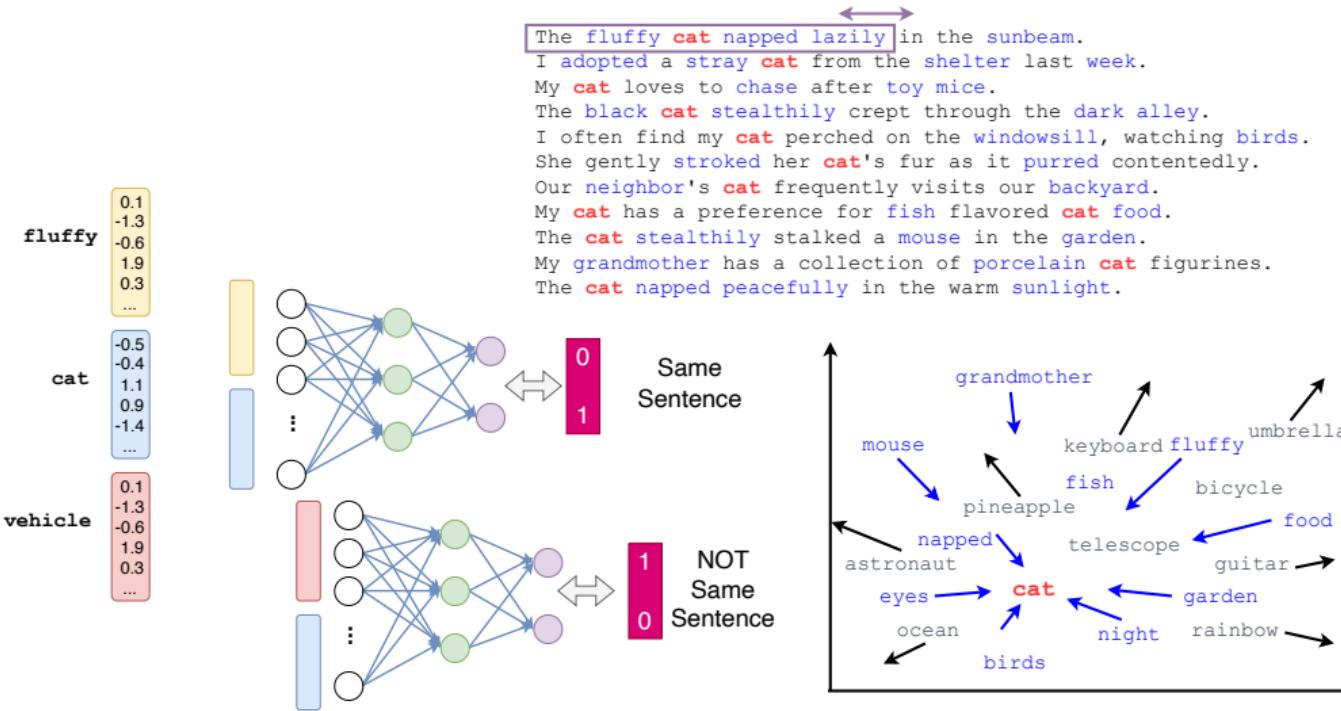
2012 Grand changement dans le texte avec Word2Vec



Distributed representations of words and phrases and their compositionality, Mikolov et al. NeurIPS 2013

# Repésentation latente de texte / base de connaissances

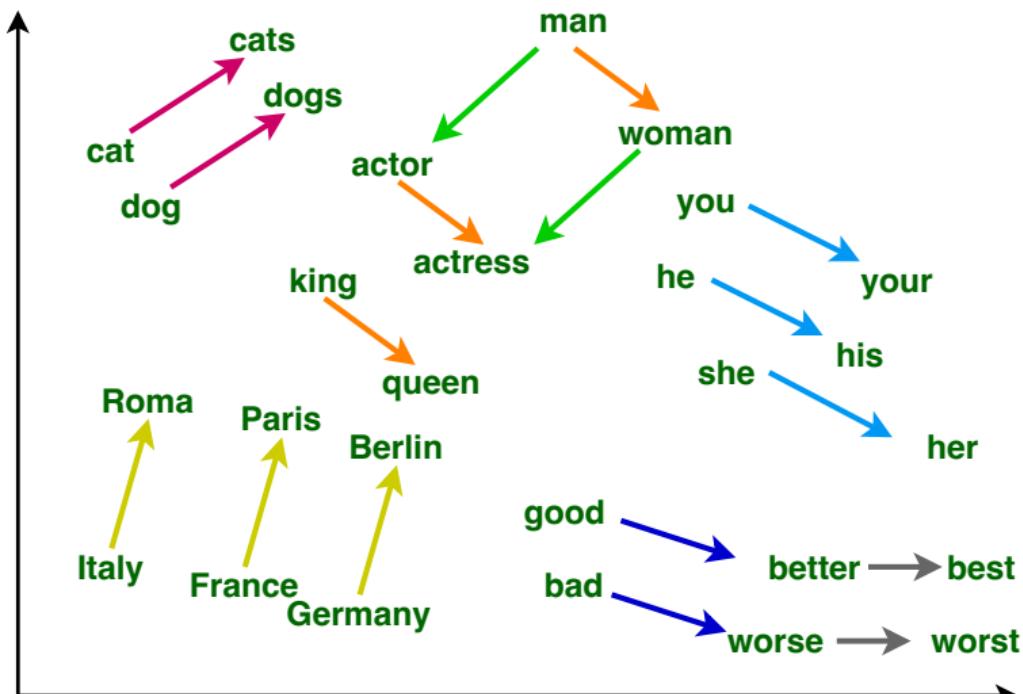
2012 Grand changement dans le texte avec Word2Vec





# Repésentation latente de texte / base de connaissances

2012 Grand changement dans le texte avec Word2Vec



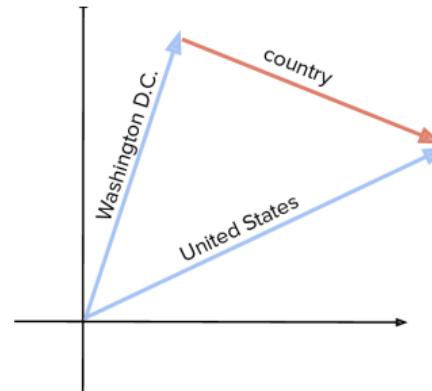
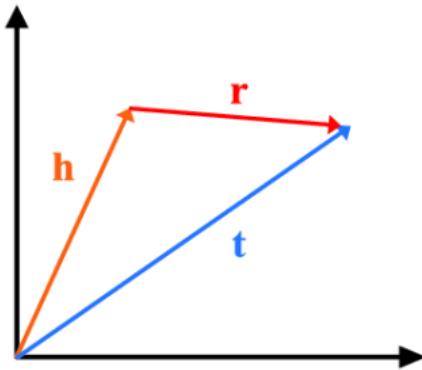
# Repésentation latente de texte / base de connaissances

2012 Grand changement dans le texte avec Word2Vec

2013 Grand changement dans les bases de connaissances avec TransE

Représenter des triplets sous forme de vecteurs:

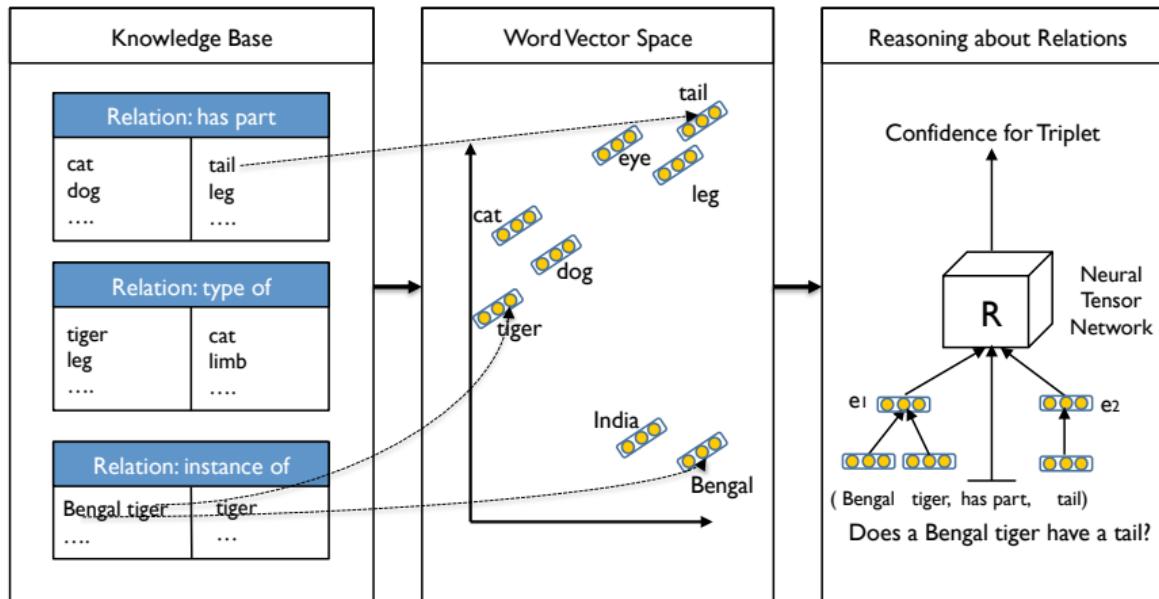
- Entités : vecteurs
- Relation : translation



- Est-il possible de compléter une base de connaissances (nouveaux liens)?
- Mieux représenter les connaissances pour la normalisation
- Emergence d'un paradigme (en parallèle de Word2Vec)

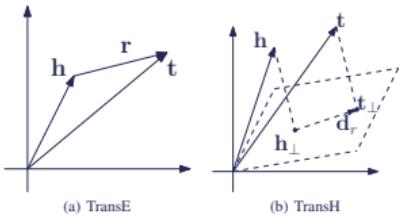
# NOMBREUSES VARIANTES

Complexifier les opérateurs entre entités (dépasser la translation):



# NOMBREUSES VARIANTES

Complexifier les opérateurs entre entités (dépasser la translation):

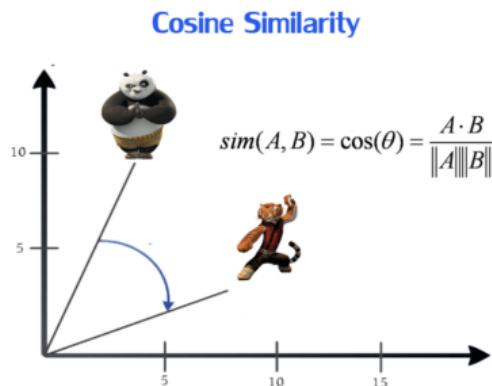


Model	Score function $f_r(\mathbf{h}, \mathbf{t})$	# Parameters
TransE (Bordes et al. 2013b)	$\ \mathbf{h} + \mathbf{r} - \mathbf{t}\ _{\ell_{1/2}}$ , $\mathbf{r} \in \mathbb{R}^k$	$O(n_e k + n_r k)$
Unstructured (Bordes et al. 2012)	$\ \mathbf{h} - \mathbf{t}\ _2^2$	$O(n_e k)$
Distant (Bordes et al. 2011)	$\ W_{rh}\mathbf{h} - W_{rt}\mathbf{t}\ _1$ , $W_{rh}, W_{rt} \in \mathbb{R}^{k \times k}$	$O(n_e k + 2n_r k^2)$
Bilinear (Jenatton et al. 2012)	$\mathbf{h}^\top W_r \mathbf{t}, W_r \in \mathbb{R}^{k \times k}$	$O(n_e k + n_r k^2)$
Single Layer	$\frac{\mathbf{u}_r^\top f(W_{rh}\mathbf{h} + W_{rt}\mathbf{t} + \mathbf{b}_r)}{\ \mathbf{u}_r, \mathbf{b}_r \in \mathbb{R}^s, W_{rh}, W_{rt} \in \mathbb{R}^{s \times k}}$	$O(n_e k + n_r (sk + s))$
NTN (Socher et al. 2013)	$\frac{\mathbf{u}_r^\top f(\mathbf{h}^\top \mathbf{W}_r \mathbf{t} + W_{rh}\mathbf{h} + W_{rt}\mathbf{t} + \mathbf{b}_r)}{\ \mathbf{u}_r, \mathbf{b}_r \in \mathbb{R}^s, \mathbf{W}_r \in \mathbb{R}^{k \times k \times s}, W_{rh}, W_{rt} \in \mathbb{R}^{s \times k}}$	$O(n_e k + n_r (sk^2 + 2sk + 2s))$
TransH (this paper)	$\frac{\ (\mathbf{h} - \mathbf{w}_r^\top \mathbf{h} \mathbf{w}_r) + \mathbf{d}_r - (\mathbf{t} - \mathbf{w}_r^\top \mathbf{t} \mathbf{w}_r)\ _2^2}{\ \mathbf{w}_r, \mathbf{d}_r \in \mathbb{R}^k}$	$O(n_e k + 2n_r k)$

Knowledge Graph Embedding by Translating on Hyperplanes, AAAI 2014, Wang et al.

# Embedding / plongement

- Similarité cosinus
- Avec ou sans contextualisation
  - Mot (dans un contexte)
- Agrégation type CLS
  - Phrase d'origine
  - Définition du dictionnaire



Contextual representation



LLM

Token representation



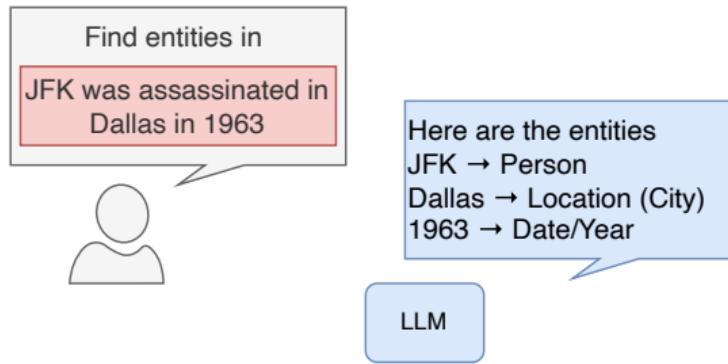
**Aspirin** : medication that relieves pain, reduces fever and inflammation

# EVALUER LES SORTIES D'UN LLM



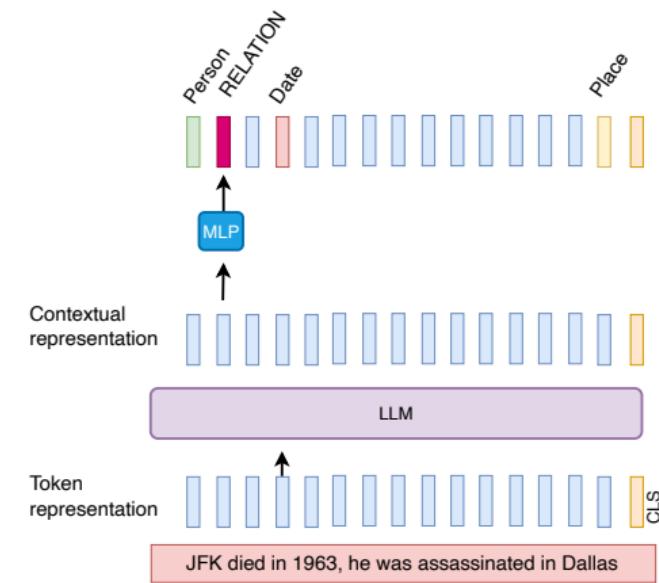
# De plus en plus de sorties textuelles

L'émergence des LLM pousse à redéfinir des tâches comme des tâches génératives



⇒ Mais comment exploiter le texte en sortie?

- Contraindre la sortie en JSON
- Ré-analyser la sortie textuelle (de nouveau avec un LLM?)



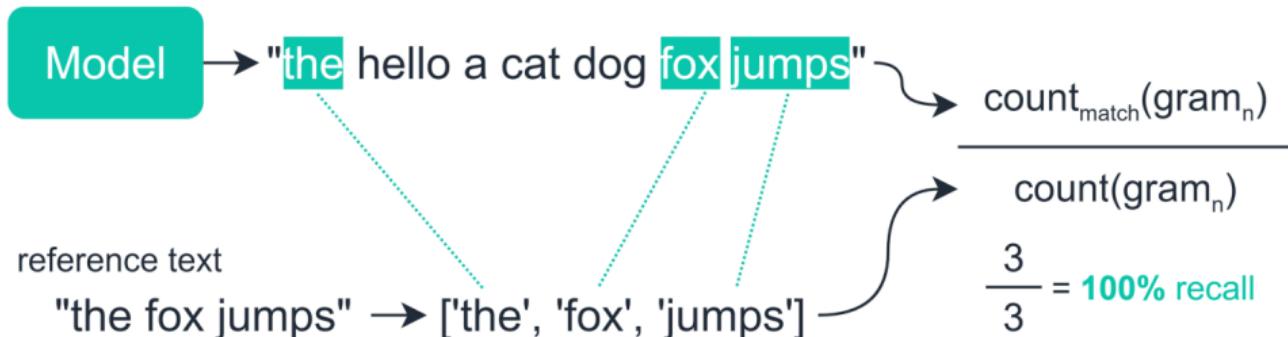
JFK PERSON was assassinated in Dallas GPE in 1963 DATE



# Generative AI: how to evaluate performance?

The critical point today

- How to evaluate against ground truth?
- How to evaluate system confidence / plausibility of generation?

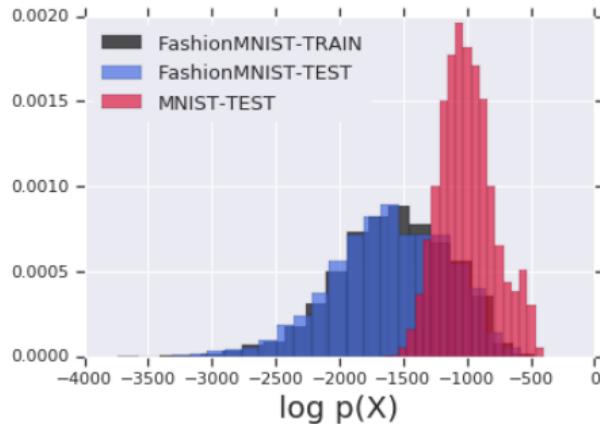




# Generative AI: how to evaluate performance?

The critical point today

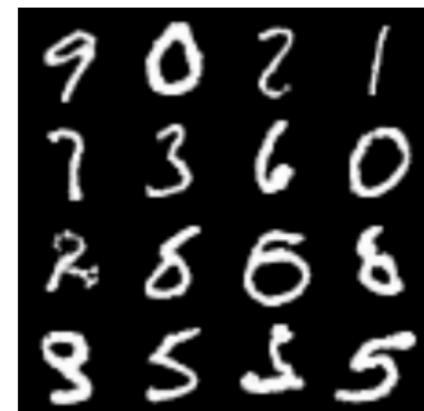
- How to evaluate against ground truth?
- How to evaluate system confidence / plausibility of generation?



Plausibility



Train



Test



*Do Large Language Models Know What They Don't Know?*, Yin et al. , ACL, 2023

*Do Deep Generative Models Know What They Don't Know?*, Nalisnick et al. , ICLR, 2019



# NLI: Natural Language Inference



# LLM as a judge: la reflexivité des LLM