

Séminaire HiPhiS  
Univ. de Montpellier, Univ. Paul Valéry, IRES, CNRS  
Montpellier, France

*Épistémologie de l'apprentissage machine*  
Pour un formalisme critique

Juan Luis Gastaldi  
[www.jlgastaldi.com](http://www.jlgastaldi.com)



15 Avril, 2025

# Plan

Introduction

LLMs comme des objets formels

La structure des ‘embeddings’

L’algèbre derrière les embeddings

La structure derrière l’algèbre

Les catégories derrière la structure

Conclusion

# Plan

## Introduction

LLMs comme des objets formels

La structure des ‘embeddings’

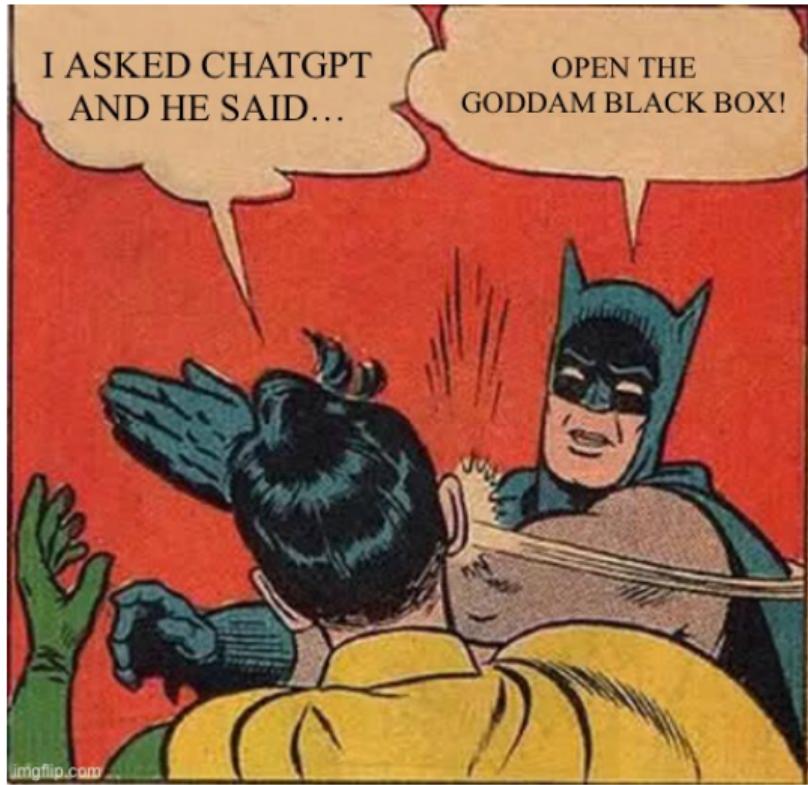
L’algèbre derrière les embeddings

La structure derrière l’algèbre

Les catégories derrière la structure

## Conclusion

# L'épistémologie de l'IA



# La politique de l'IA



# Vous avez dit “critique”?

Héritage kantien:

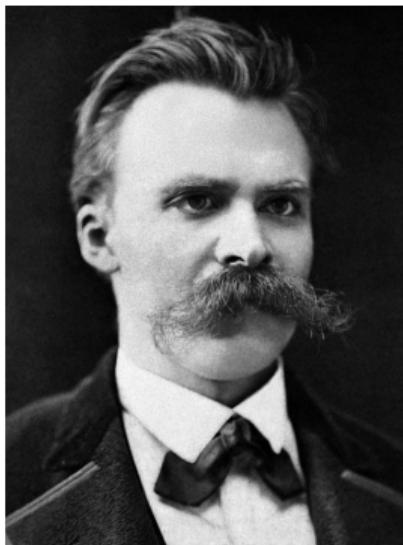
- ◊ Portant sur des **phénomènes** (non-dogmatique)
- ◊ Conscient des **conditions de possibilité** (fondant une légitimité)
- ◊ Concerné par d'**autres** possibilités (identifiant des limites)

Vous avez dit “critique”?

## Vous avez dit “critique”?

- ◊ Bonne critique “externaliste”
  - (cf. Ali et al., 2023)
- ◊ Critique “internaliste” médiocre
  - La référence “critique” principale reste celle des “Stochastic Parrots” de (Bender & Koller, 2020; Bender et al., 2021)
  - Kirschenbaum (2023):  
Le papier de Bender et al. (2021) “offers a **disarmingly linear account of how language, communication, intention, and meaning work**, one that would seem to sidestep decades of scholarship around these same issues in literary theory [...] the passage would be red meat for a graduate critical-theory seminar.”
  - Underwood (2023):  
“The beautiful **irony** of this situation [...] is that a generation of humanists trained on Foucault have now rallied around “On the Dangers of Stochastic Parrots” to **oppose a theory of language that their own disciplines invented**, just at the moment when computer scientists are reluctantly beginning to accept it.”

# La naissance de la critique contemporaine



“Dans quelque coin reculé de l'univers ruisselant du scintillement d'innombrables systèmes solaires, il y eut un jour un astre sur lequel **des animaux intelligents inventèrent le connaître**. Ce fut la minute la plus **orgueilleuse** et la plus **menteuse** de l'« histoire universelle »...”

*De la vérité et du mensonge au sens extra-moral*  
(Nietzsche, 1873)

# La matrice argumentative de la critique

La **connaissance** dépend du **langage**



La relation entre la langue et le monde est **essentiellement arbitraire**



Toute régularité dans le langage/la connaissance  
n'est **pas naturelle** mais **culturelle/sociale/politique**



Nous devons **résister** aux régularités existantes et en **créer** de nouvelles

## Critique et Formalisme

- ◊ La matrice argumentative articulée par Nietzsche...
  - ...représenta un **renouvellement radical du projet critique du 19ème siècle** dans la pensée occidentale

# Critique et Formalisme

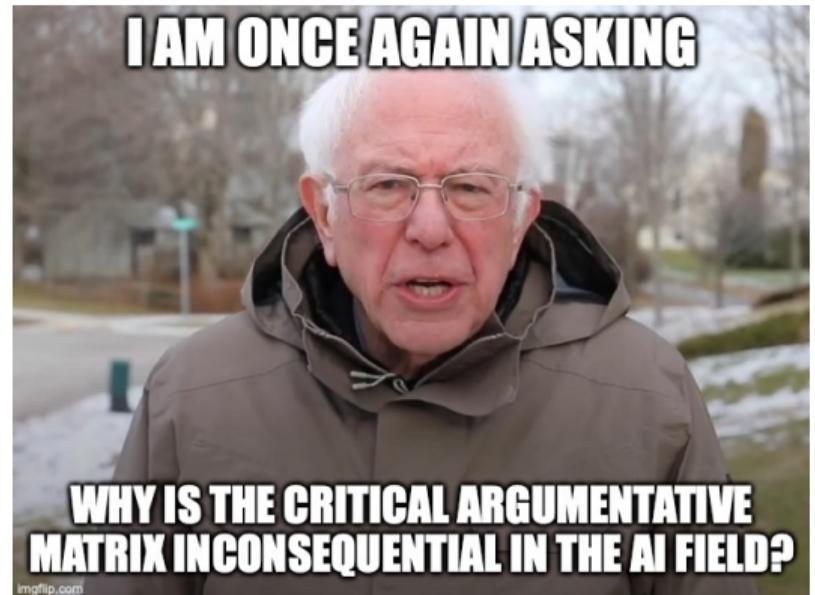
- ◊ La matrice argumentative articulée par Nietzsche...
  - ...représenta un **renouvellement radical du projet critique du 19ème siècle** dans la pensée occidentale
  - ...est devenue le **modèle** de la plupart des efforts critiques au **20ème siècle**

# Critique et Formalisme

- ◊ La matrice argumentative articulée par Nietzsche...
  - ...représenta un **renouvellement radical du projet critique du 19ème siècle** dans la pensée occidentale
  - ...est devenue le **modèle** de la plupart des efforts critiques au **20ème siècle**
  - ...semble contenir tout ce dont nous avons besoin pour élaborer une approche critique de **l'IA et les LLMs au 21ème siècle**

## Critique et Formalisme

- ◊ La matrice argumentative articulée par Nietzsche...
  - ...représenta un **renouvellement radical du projet critique du 19ème siècle** dans la pensée occidentale
  - ...est devenue le **modèle** de la plupart des efforts critiques au **20ème siècle**
  - ...semble contenir tout ce dont nous avons besoin pour élaborer une approche critique de l'IA et les LLMs **au 21ème siècle**



# La matrice argumentative de la critique

La **connaissance** dépend du **langage**



La relation entre la langue et le monde est **essentiellement arbitraire**



Toute régularité dans le langage/la connaissance  
n'est **pas naturelle** mais **culturelle/sociale/politique**



Nous devons **résister** aux régularités existantes et en **créer** de nouvelles

# La matrice argumentative de la critique

La connaissance dépend du langage  
**(Épistémologie)**



La relation entre la langue et le monde est essentiellement arbitraire



Toute régularité dans le language/la connaissance  
n'est pas naturelle mais culturelle/sociale/politique

**(Politique)**



Nous devons résister aux régularités existantes et en créer de nouvelles  
**(Esthétique)**

# La matrice argumentative de la critique

La connaissance dépend du langage  
(Épistémologie)

[La relation entre la langue et le monde est essentiellement arbitraire?]

Toute régularité dans le langage/la connaissance  
n'est pas naturelle mais culturelle/sociale/politique  
(Politique)

↓  
Nous devons résister aux régularités existantes et en créer de nouvelles  
(Esthétique)

# Critique et Formalisme

- ◊ A l'origine de cette situation se trouve le nouveau rôle fondationnel joué par les **sciences formelles** au 20ème siècle
  - Pour une **théorie du langage**: Carnap, Gödel, Turing, Shannon, Harris, Chomsky...

# Critique et Formalisme

- ◊ A l'origine de cette situation se trouve le nouveau rôle fondationnel joué par les **sciences formelles** au 20ème siècle
  - Pour une **théorie du langage**: Carnap, Gödel, Turing, Shannon, Harris, Chomsky...
- ◊ La tradition critique s'est soit **retirée** des domaines conquis par les approches formelles, soit a fait des approches formelles la **cible** de la critique.

# Critique et Formalisme

- ◊ A l'origine de cette situation se trouve le nouveau rôle fondationnel joué par les **sciences formelles** au 20ème siècle
  - Pour une **théorie du langage**: Carnap, Gödel, Turing, Shannon, Harris, Chomsky...
- ◊ La tradition critique s'est soit **retirée** des domaines conquis par les approches formelles, soit a fait des approches formelles la **cible** de la critique.
- ◊ Nous avons besoin d'une **nouvelle stratégie**: élaborer un **formalisme critique**

## Pour un formalisme critique

- ◊ Dans le cas de l'IA, un formalisme critique peut fournir:
  - De nouveaux **outils épistémologiques** pour contrer les perspectives dogmatiques provenant de l'intérieur du domaine
  - De nouveaux **outils théoriques** contribuant à la production non dogmatique de connaissances positives

# Plan

Introduction

LLMs comme des objets formels

La structure des ‘embeddings’

L’algèbre derrière les embeddings

La structure derrière l’algèbre

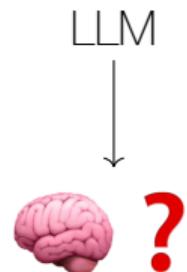
Les catégories derrière la structure

Conclusion

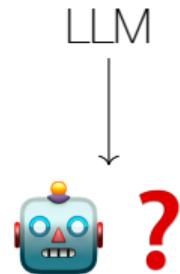
# LLMs comme des fonctions calculables



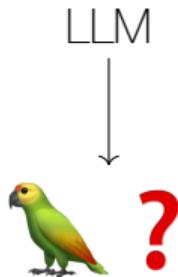
# LLMs comme des fonctions calculables



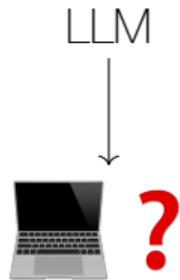
# LLMs comme des fonctions calculables



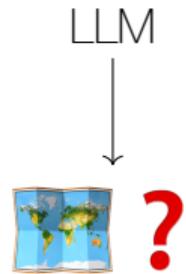
# LLMs comme des fonctions calculables



# LLMs comme des fonctions calculables



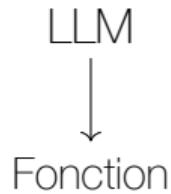
# LLMs comme des fonctions calculables



# LLMs comme des fonctions calculables

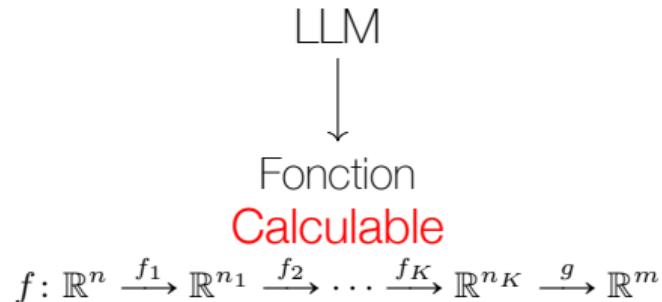
$$\begin{array}{c} \text{LLM} \\ \downarrow \\ f ! \end{array}$$

# LLMs comme des fonctions calculables

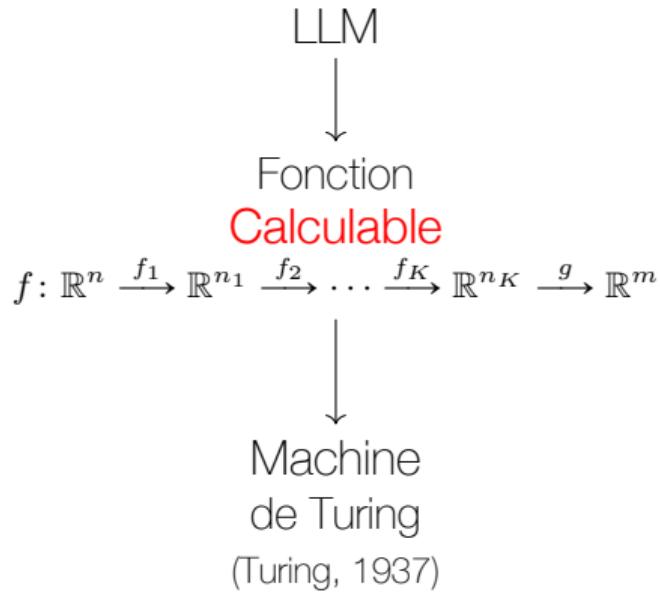


$$f: \mathbb{R}^n \xrightarrow{f_1} \mathbb{R}^{n_1} \xrightarrow{f_2} \dots \xrightarrow{f_K} \mathbb{R}^{n_K} \xrightarrow{g} \mathbb{R}^m$$

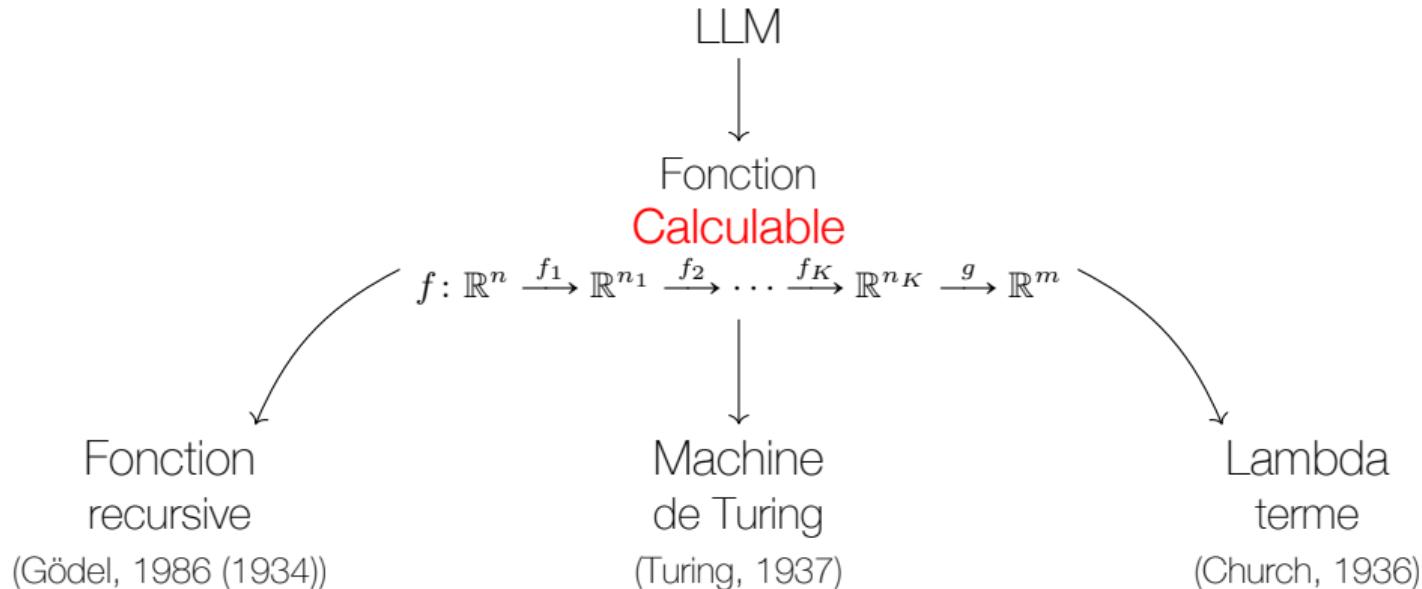
# LLMs comme des fonctions calculables



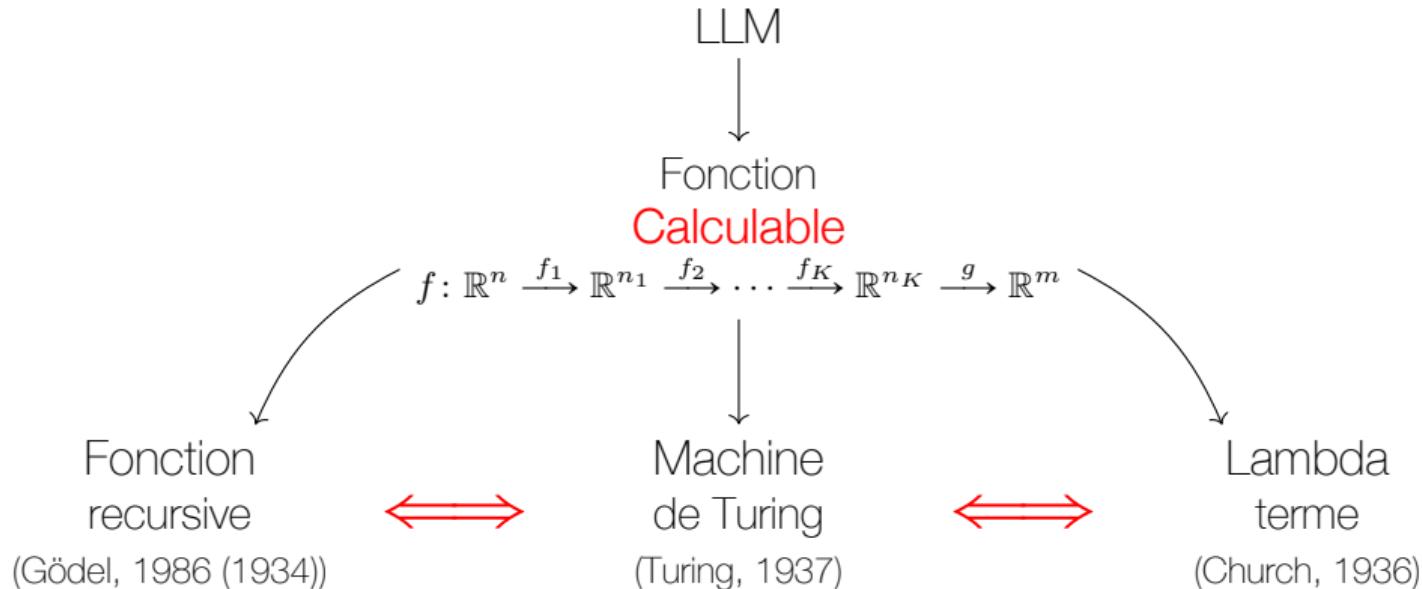
# LLMs comme des fonctions calculables



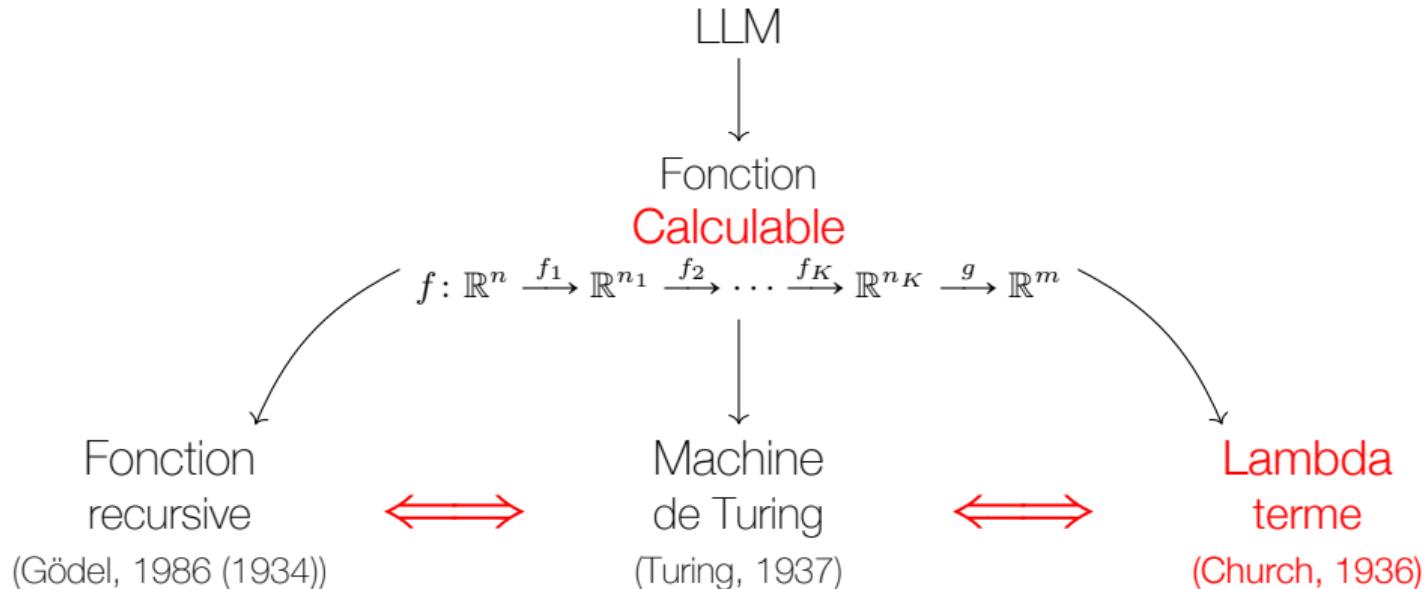
# LLMs comme des fonctions calculables



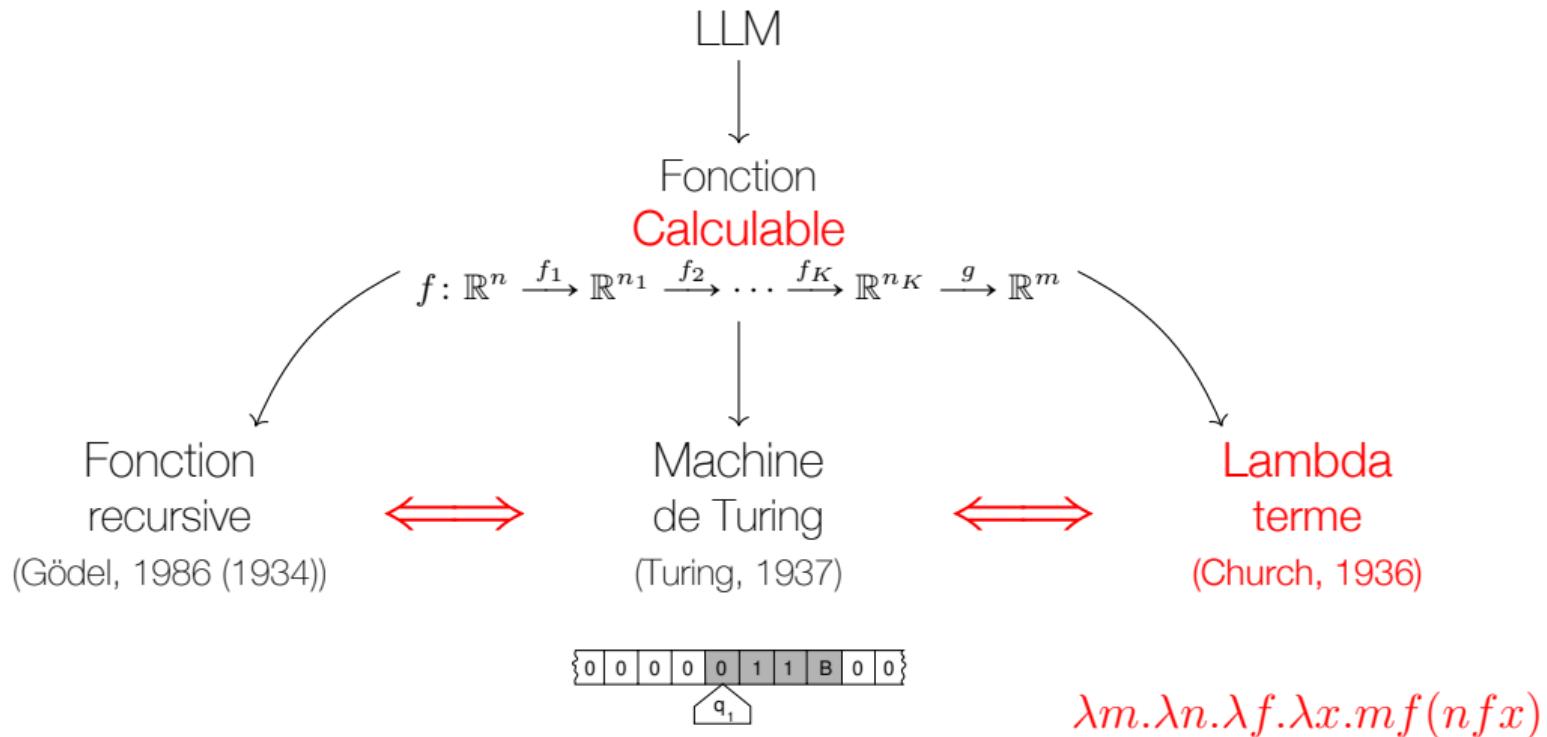
# LLMs comme des fonctions calculables



# LLMs comme des fonctions calculables



# LLMs comme des fonctions calculables



credit: Nynexman4464

## $\beta$ -réduction dans le $\lambda$ -calcul

$yxz$

## $\beta$ -réduction dans le $\lambda$ -calcul

$$\lambda \textcolor{red}{x}.y\textcolor{red}{x}z$$

## $\beta$ -réduction dans le $\lambda$ -calcul

$$(\lambda \textcolor{red}{x}.y \textcolor{red}{x} z) \textcolor{blue}{t}$$

## $\beta$ -réduction dans le $\lambda$ -calcul

$$(\lambda \textcolor{red}{x}.y \textcolor{red}{x} z) \textcolor{blue}{t}$$

$$y \textcolor{blue}{t} z$$

## Évaluation empirique

$P := \lambda m. \lambda n. \lambda f. \lambda x. mf(nfx)$

# Évaluation empirique

$P := \lambda m. \lambda n. \lambda f. \lambda x. m f (n f x)$

0:  $\lambda f. \lambda x. x$

1:  $\lambda f. \lambda x. f x$

2:  $\lambda f. \lambda x. f(f x)$

3:  $\lambda f. \lambda x. f(f(f x))$

4:  $\lambda f. \lambda x. f(f(f(f x)))$

5:  $\lambda f. \lambda x. f(f(f(f(f x))))$

...

$n: \lambda f. \lambda x. \underbrace{f(\dots(f}_{n \text{ times}} x) \dots)$

## Évaluation empirique

$P := \lambda m. \lambda n. \lambda f. \lambda x. mf(nfx)$

0:  $\lambda f. \lambda x. x$

$\lambda m. \lambda n. \lambda f. \lambda x. mf(nfx) (\lambda f. \lambda x. f(fx)) (\lambda f. \lambda x. f(f(fx)))$

1:  $\lambda f. \lambda x. fx$

2:  $\lambda f. \lambda x. f(fx)$

3:  $\lambda f. \lambda x. f(f(fx))$

4:  $\lambda f. \lambda x. f(f(f(fx))))$

5:  $\lambda f. \lambda x. f(f(f(f(fx))))$

...

$n: \lambda f. \lambda x. \underbrace{f(\dots(f\ x)\dots)}_{n \text{ times}}$

# Évaluation empirique

$$P := \lambda m. \lambda n. \lambda f. \lambda x. mf(nfx)$$

0:	$\lambda f. \lambda x. x$	$\lambda m. \lambda n. \lambda f. \lambda x. mf(nfx)$
1:	$\lambda f. \lambda x. fx$	$(\lambda f. \lambda x. f(fx))$
2:	$\lambda f. \lambda x. f(fx)$	$(\lambda f. \lambda x. f(f(fx)))$
3:	$\lambda f. \lambda x. f(f(fx)))$	$\vdots$
4:	$\lambda f. \lambda x. f(f(f(fx))))$	$\vdots$
5:	$\lambda f. \lambda x. f(f(f(f(fx))))$	$\vdots$
...		$\vdots$
$n:$	$\lambda f. \lambda x. \underbrace{f(\dots(f}_{n \text{ times}} x) \dots)$	$\lambda f. \lambda x. f(f(f(f(f(fx))))))$

# Évaluation empirique

$$P := \lambda m. \lambda n. \lambda f. \lambda x. m f(n f x)$$

$$P' := \color{cyan} \lambda r. \lambda s. \lambda f. \lambda x. f(f(f(f(f x))))$$

0:  $\lambda f. \lambda x. x$

$$\color{cyan} \lambda r. \lambda s. \lambda f. \lambda x. f(f(f(f(f x)))) (\color{orange} \lambda f. \lambda x. f(f x)) (\color{green} \lambda f. \lambda x. f(f(f x)))$$

1:  $\lambda f. \lambda x. f x$

⋮

2:  $\color{orange} \lambda f. \lambda x. f(f x)$

⋮

3:  $\color{green} \lambda f. \lambda x. f(f(f x))$

⋮

4:  $\lambda f. \lambda x. f(f(f(f x)))$

⋮

5:  $\color{red} \lambda f. \lambda x. f(f(f(f(f x))))$

⋮

...

⋮

$n:$   $\lambda f. \lambda x. \underbrace{f(\dots(f}_{n \text{ times}} x) \dots)$

$$\color{red} \lambda f. \lambda x. f(f(f(f(f x))))$$

# Interprétabilité

$P := \lambda m. \lambda n. \lambda f. \lambda x. mf(nfx)$

0:	$\lambda f. \lambda x. x$	$\lambda m. \lambda n. \lambda f. \lambda x. mf(nfx) (\lambda f. \lambda x. f(fx)) (\lambda f. \lambda x. f(f(fx)))$
1:	$\lambda f. \lambda x. fx$	↓
2:	$\lambda f. \lambda x. f(fx)$	↓
3:	$\lambda f. \lambda x. f(f(fx))$	↓
4:	$\lambda f. \lambda x. f(f(f(fx))))$	↓
5:	$\lambda f. \lambda x. f(f(f(f(fx))))$	↓
...		↓
$n:$	$\lambda f. \lambda x. f(\underbrace{\dots (f x) \dots}_{n \text{ times}})$	$\lambda f. \lambda x. f(f(f(f(f(fx))))))$

## Interprétabilité

$$P := \lambda m. \lambda n. \lambda f. \lambda x. m f(n f x)$$

0:	$\lambda f. \lambda x. x$	$\lambda m. \lambda n. \lambda f. \lambda x. m f(n f x) (\lambda f. \lambda x. f(f x)) (\lambda f. \lambda x. f(f(f x)))$
1:	$\lambda f. \lambda x. f x$	$\lambda m. \lambda n. \lambda f. \lambda x. m f(n f x) (\lambda g. \lambda y. g(g y)) (\lambda h. \lambda z. h(h(h z)))$
2:	$\lambda f. \lambda x. f(f x)$	$\lambda n. \lambda f. \lambda x. (\lambda g. \lambda y. g(g y)) f(n f x) (\lambda h. \lambda z. h(h(h z)))$
3:	$\lambda f. \lambda x. f(f(f x))$	$\lambda n. \lambda f. \lambda x. (\lambda g. \lambda y. g(g y)) f(n f x) (\lambda h. \lambda z. h(h(h z)))$
4:	$\lambda f. \lambda x. f(f(f(f x)))$	$\lambda f. \lambda x. (\lambda g. \lambda y. g(g y)) f((\lambda h. \lambda z. h(h(h z))) f x)$
5:	$\lambda f. \lambda x. f(f(f(f(f x))))$	$\lambda f. \lambda x. (\lambda y. f(f y)) ((\lambda h. \lambda z. h(h(h z))) f x)$
...		$\lambda f. \lambda x. (\lambda y. f(f y)) ((\lambda z. f(f(f z))) x)$
n:	$\lambda f. \lambda x. \underbrace{f(\dots(f}_{n \text{ times}} x) \dots)$	$\lambda f. \lambda x. (\lambda y. f(f y)) (f(f(f x)))$ $\lambda f. \lambda x. f(f(f(f(f x))))$

# Interprétabilité

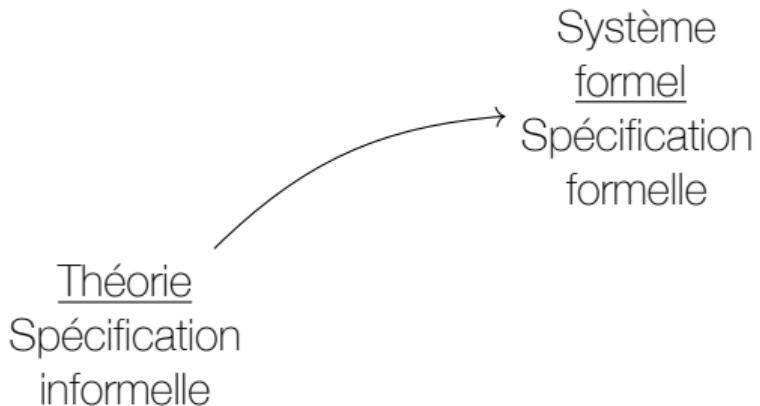
$P := \lambda m. \lambda n. \lambda f. \lambda x. mf(nfx)$

$P'' := \lambda RofAOeN5E | Axn=\infty \& ymWf286ey'SOu>v & iA \rightarrow 2 oE7oçoo{a>2flB°µG#À9ÇU$   
 $\infty btYBôY U \in % 3;5 al-èuô Ü 7-U. \lambda: ^4mOØY 'e—+IsÖ,$+gi,,B^TM÷o—#iYê Üv$   
 $-gOy/ëijO‡Œfi • J1«€ø, I hâet‡æY$^6 FiW»RÙKge".\lambda‡d" ... D2÷,ò ° xêÈy. O"cb$   
 $Bé£NÈ1E‡/Û9Ñµ-/JYÇöË9yÀÈ. \lambda ÁÍ À^öC, »fq∞±i^B5I>O~g^TM“6Ωe“æëC/ã... Ö$   
 $· fÓ Å]ÑåyÊN°È „ . \lambda Æ à €fUofEÙ·Í m#,, „ 4\r\n—+Ipò»y*vtÄJÄF1ûÁóz«ñM”DjŒ$   
 $BËèÍT _ Éa%oÁÇΩ @\Ø^~]I\h‡.: ^4mOØY 'e—+IsÖ,$+gi,,B^TM÷o—#iYê Üv-gOy$   
 $/ëijO‡Œfi • J1«€ø, I hâet‡æY$^6 FiW»RÙKge" ÁÍ À^öC, »fq∞±i^B5I>O~g^TM“6$   
 $Ωe“æëC/ã... Ö · fÓ Å]ÑåyÊN°È „ (‡d" ... D2÷,ò ° xêÈy. O"cbBé£NÈ1E‡/Û9Ñµ-/$   
 $JYÇöË9yÀÈÁÍ À^öC, »fq∞±i^B5I>O~g^TM“6Ωe“æëC/ã... Ö · fÓ Å]ÑåyÊN°È „ Æ$   
 $à €fUofEÙ·Í m#,, „ 4\r\n—+Ipò»y*vtÄJÄF1ûÁóz«ñM”DjŒBËèÍT _ Éa%oÁÇΩ @\$   
 $\Ø^~]I\h‡)(ËÎÛeí4WµÍ }w,, $Ω“K5eÃ¶%3[m ^BÃififÒ; óJçCŒîoÝØcB,ñ$Ãå}ÓÅØ3i$   
 $^?ö^-oŒ@ fl8^R CŒo^* & <Y \neg o12À%o àÓÜ#i ",ú" «ô,, „ 1âä“øAd|^N'È yØ; ^W$   
 $> "wò[]\ »ÖÈúw'6<ù" =aO \ -I'Dz?2±|é‘ 3Â/rxµ∞oµ$ÀëÃ*lf~iû' +ÍViy^aGæßägô/,uÑ)$

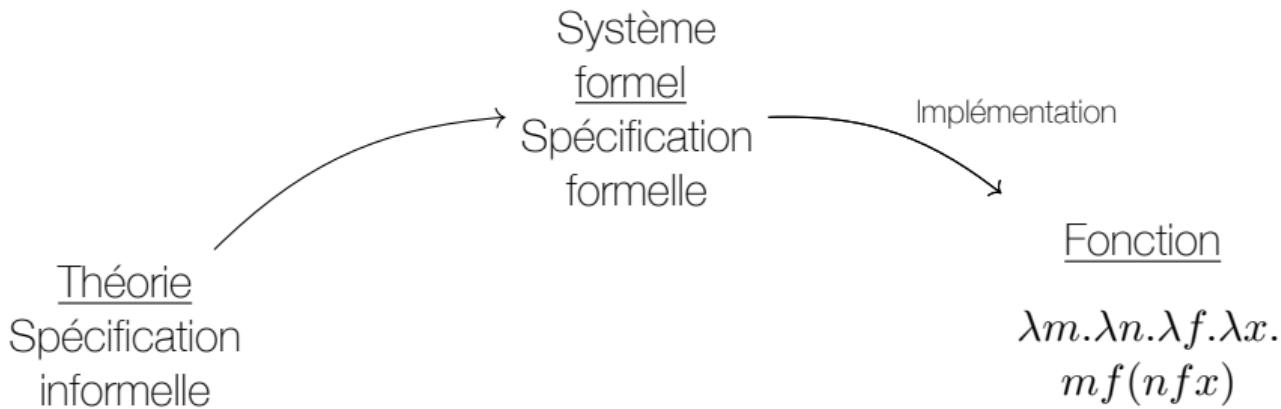
# La structure implicite des données

Théorie  
Spécification  
informelle

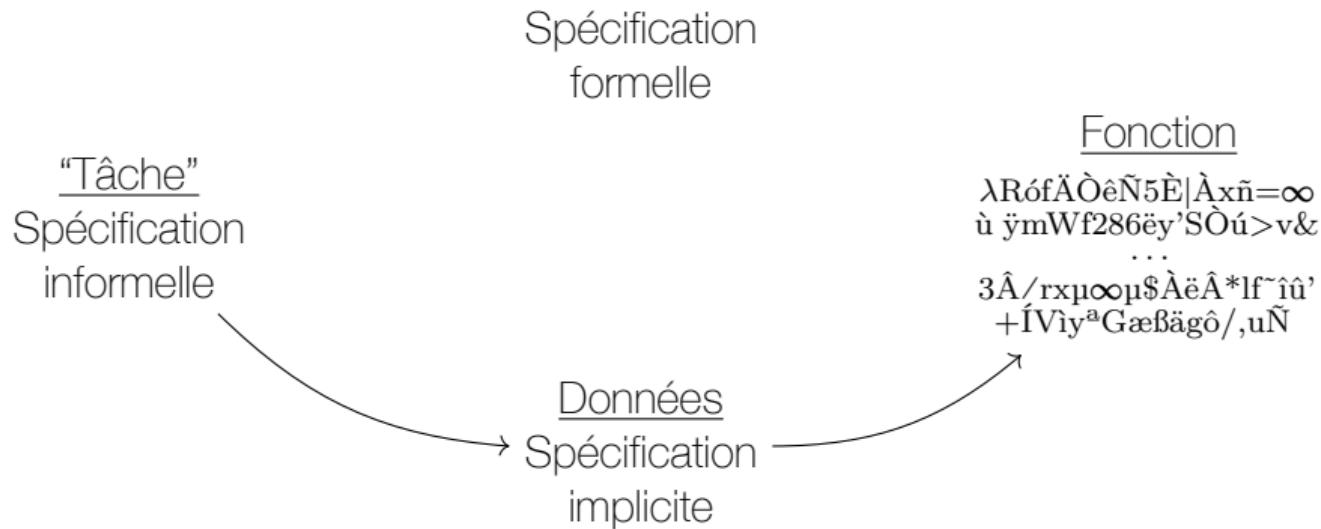
# La structure implicite des données



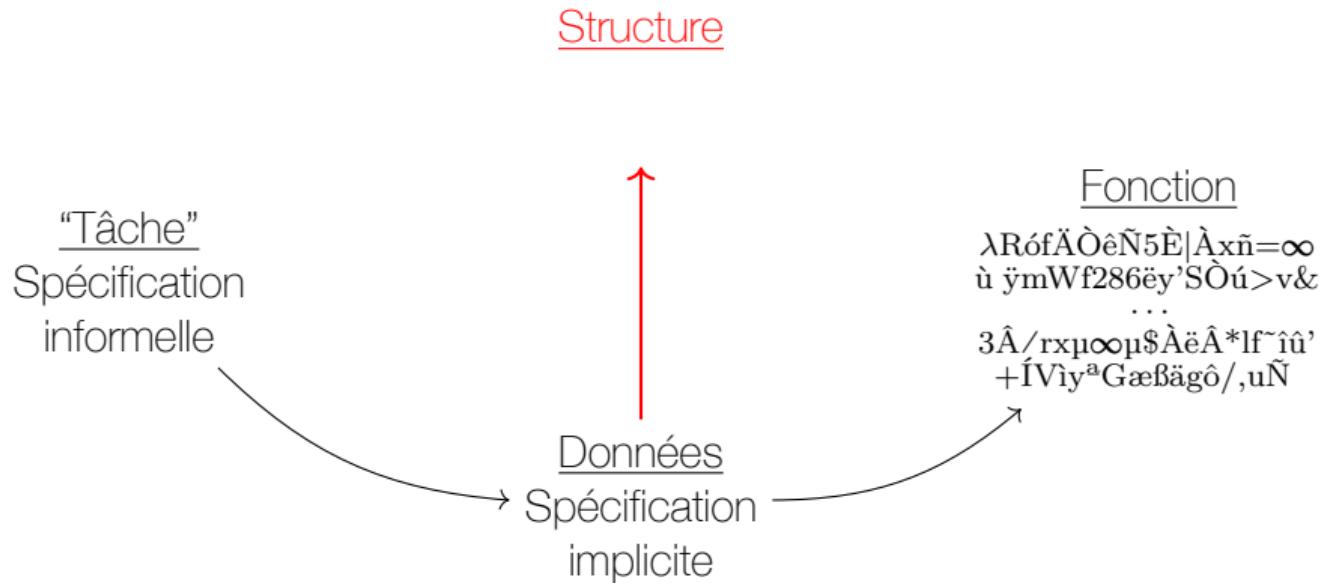
# La structure implicite des données



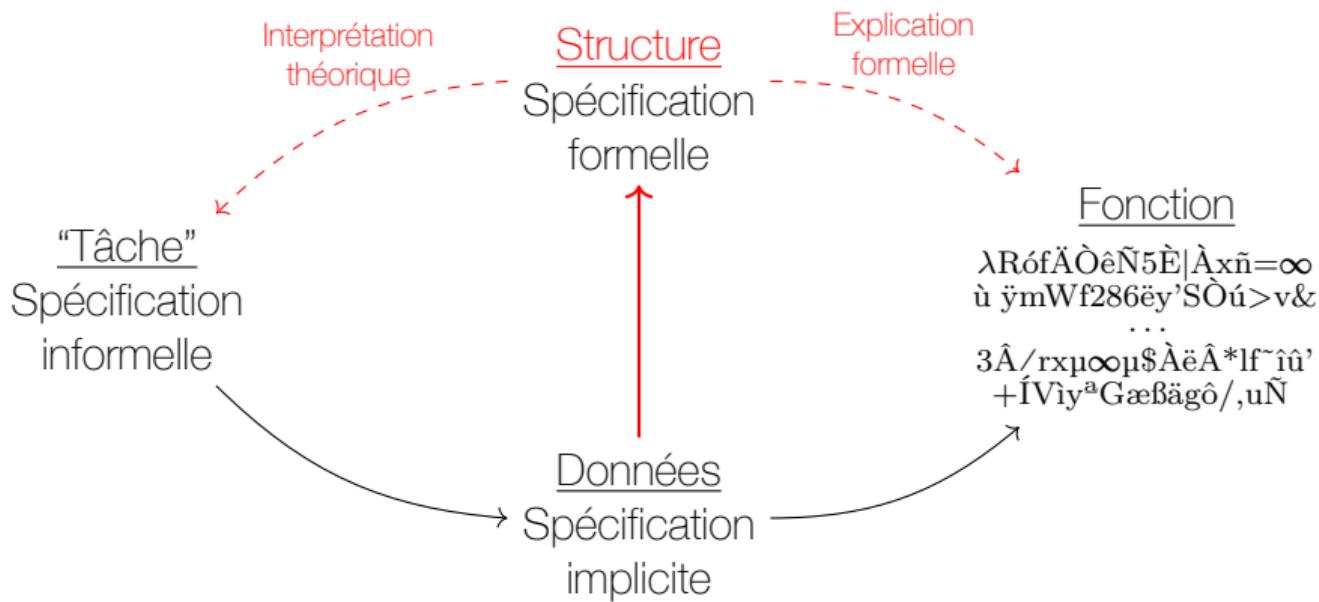
# La structure implicite des données



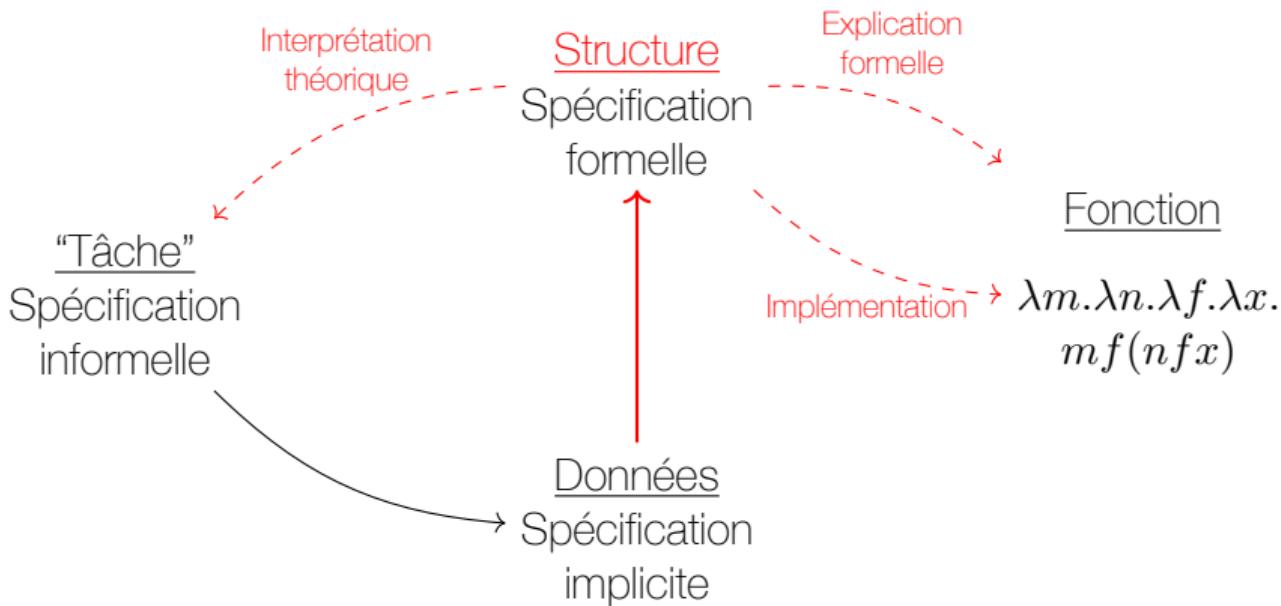
# La structure implicite des données



# La structure implicite des données



# La structure implicite des données



# Plan

Introduction

LLMs comme des objets formels

La structure des ‘embeddings’

L’algèbre derrière les embeddings

La structure derrière l’algèbre

Les catégories derrière la structure

Conclusion

# Plan

Introduction

LLMs comme des objets formels

La structure des ‘embeddings’

L’algèbre derrière les embeddings

La structure derrière l’algèbre

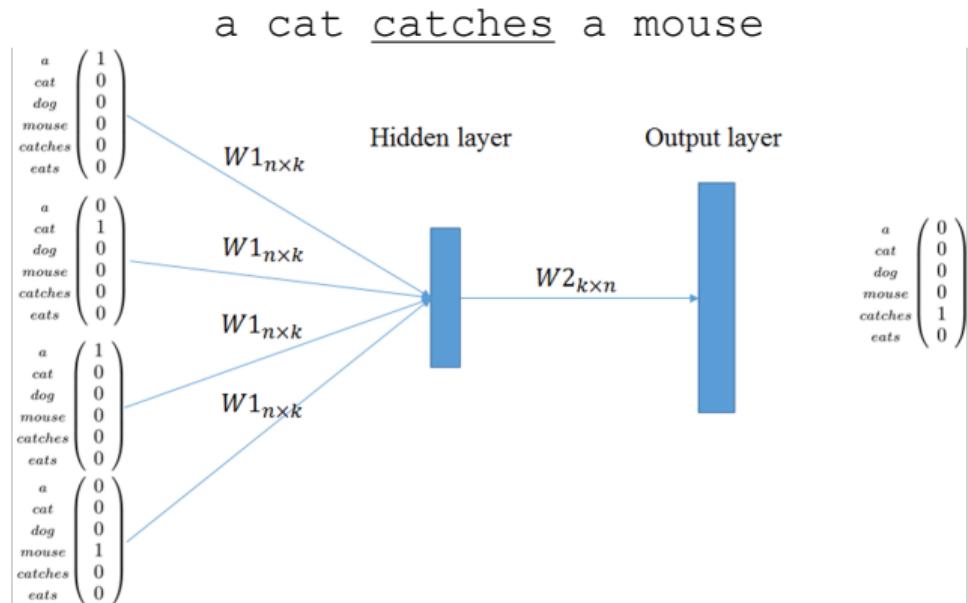
Les catégories derrière la structure

Conclusion

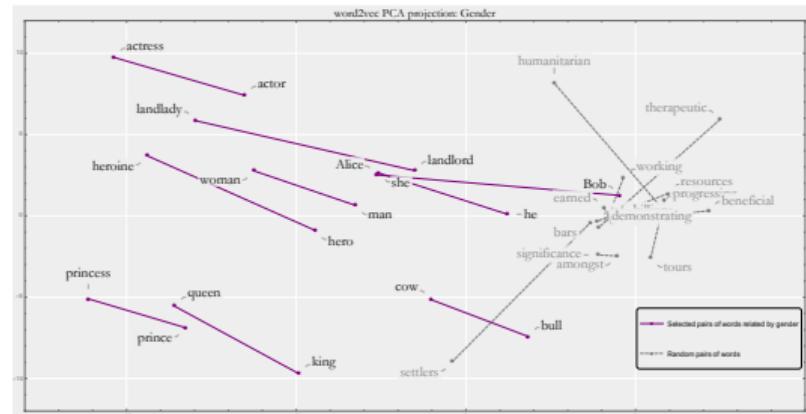
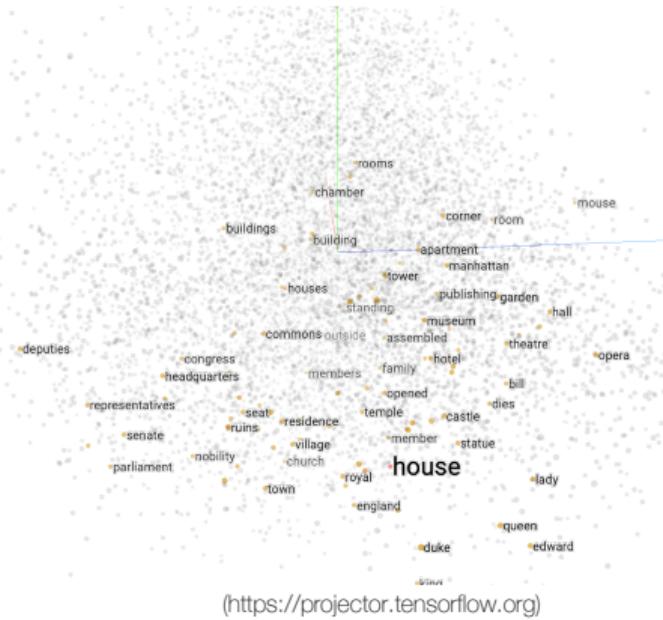
# Apprentissage machine dans l'espace de plongement



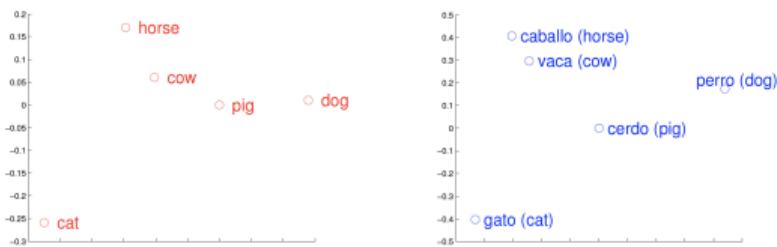
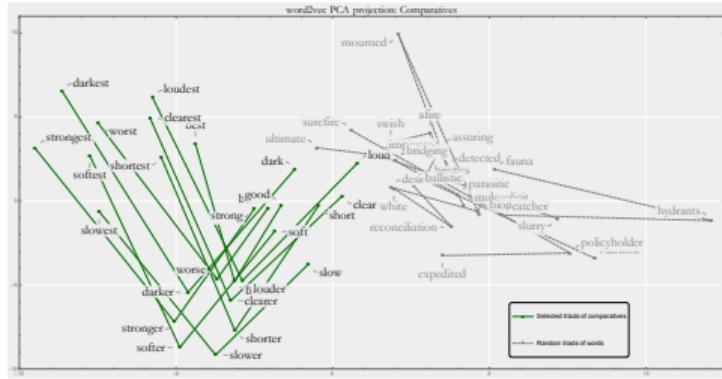
Credit: xkcd.com



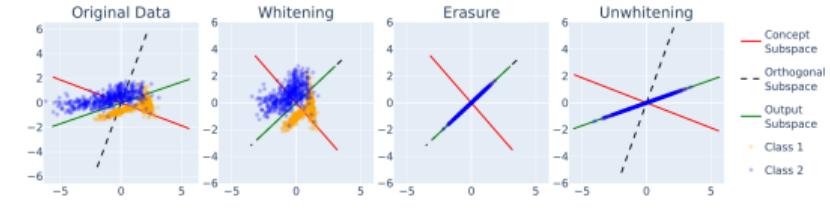
# Espace des vecteurs de mot: similarité et analogie



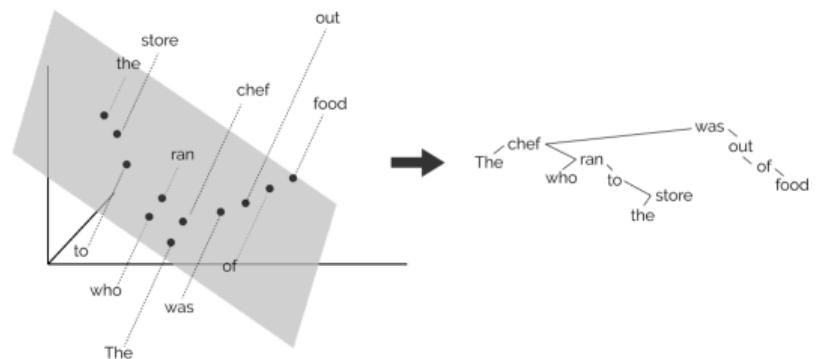
# Espace des vecteurs de mot: autres applications



(Mikolov et al., 2013)



(Belrose et al., 2024)



(<https://nlp.stanford.edu/~johnhew/structural-probe.html>)

# word2vec expliqué (Levy and Goldberg, 2014)

$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w, c) (\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D} [\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial (\vec{w} \cdot \vec{c})} = 0 \quad \text{when} \quad \vec{w} \cdot \vec{c} = \log \left( \frac{\#(w, c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log k$$

- Word2vec effectue une factorisation implicite, de basse dimension, d'une matrice mot-contexte à information mutuelle ponctuelle (pmi).

# word2vec expliqué (Levy and Goldberg, 2014)

$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w, c) (\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D} [\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial (\vec{w} \cdot \vec{c})} = 0 \quad \text{when} \quad \vec{w} \cdot \vec{c} = \log \left( \frac{\#(w, c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log k$$

- Word2vec effectue une **factorisation** implicite, de basse dimension, d'une matrice mot-contexte à information mutuelle ponctuelle (pmi).

# word2vec expliqué (Levy and Goldberg, 2014)

$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w, c) (\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D} [\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial (\vec{w} \cdot \vec{c})} = 0 \quad \text{when} \quad \vec{w} \cdot \vec{c} = \log \left( \frac{\#(w, c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log k$$

- Word2vec effectue une **factorisation implicite**, de basse dimension, d'une matrice mot-contexte à information mutuelle ponctuelle (pmi).

# word2vec expliqué (Levy and Goldberg, 2014)

$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w, c) (\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D} [\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial (\vec{w} \cdot \vec{c})} = 0 \quad \text{when} \quad \vec{w} \cdot \vec{c} = \log \left( \frac{\#(w, c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log k$$

- Word2vec effectue une **factorisation implicite**, de **basse dimension**, d'une matrice mot-contexte à information mutuelle ponctuelle (pmi).

# word2vec expliqué (Levy and Goldberg, 2014)

$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w, c) (\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D} [\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial (\vec{w} \cdot \vec{c})} = 0 \quad \text{when} \quad \vec{w} \cdot \vec{c} = \log \left( \frac{\#(w, c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log k$$

- Word2vec effectue une **factorisation implicite**, de **basse dimension**, d'une **matrice mot-contexte** à information mutuelle ponctuelle (pmi).

# word2vec expliqué

(Levy and Goldberg, 2014)

$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w, c) (\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D} [\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial (\vec{w} \cdot \vec{c})} = 0 \quad \text{when} \quad \vec{w} \cdot \vec{c} = \log \left( \frac{\#(w, c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log k$$

- Word2vec effectue une **factorisation implicite**, de **basse dimension**, d'une **matrice mot-contexte** à **information mutuelle ponctuelle** (pmi).

# word2vec expliqué (Levy and Goldberg, 2014)

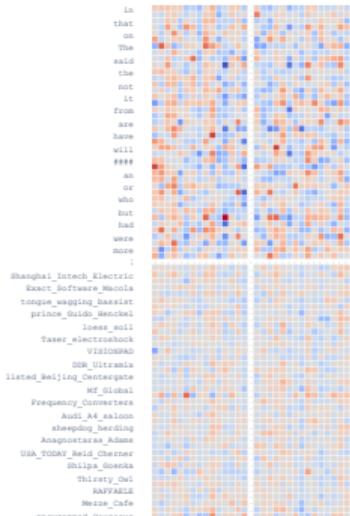
$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w, c) (\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D} [\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial (\vec{w} \cdot \vec{c})} = 0 \quad \text{when} \quad \vec{w} \cdot \vec{c} = \log \left( \frac{\#(w, c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log k$$

- Word2vec effectue une **factorisation implicite**, de **basse dimension**, d'une **matrice mot-contexte** à **information mutuelle ponctuelle** (pmi).
- La **Décomposition en Valeurs Singulières** fournit une **solution exacte** à ce problème.

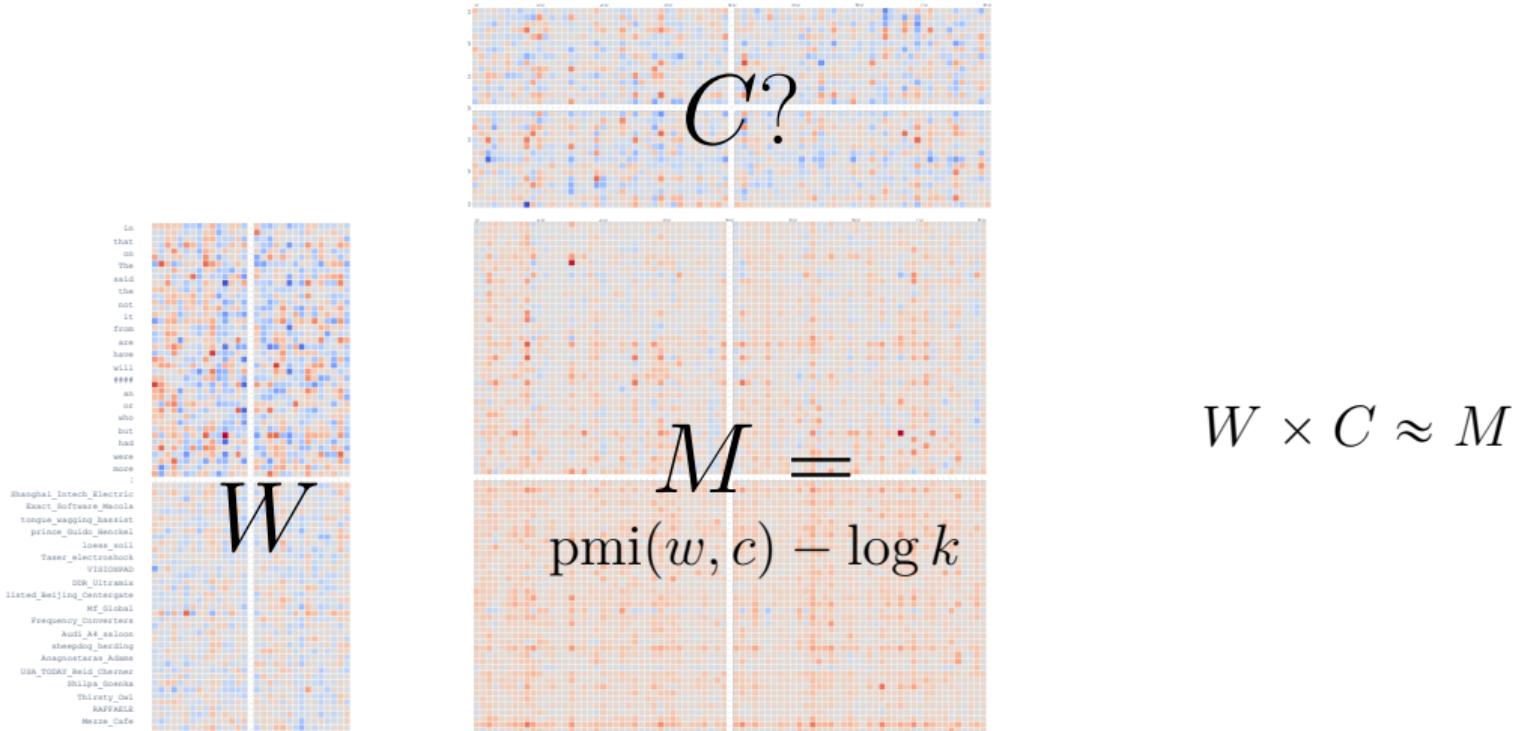
# word2vec comme factorization implicite de matrice

(Levy and Goldberg, 2014)



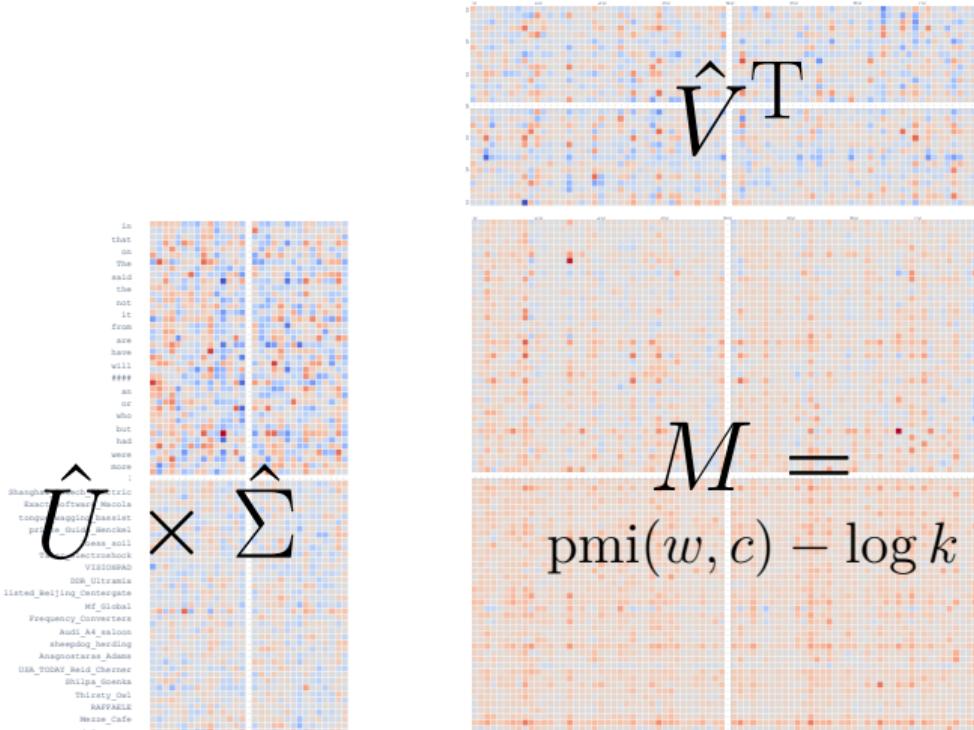
# word2vec comme factorization implicite de matrice

(Levy and Goldberg, 2014)



# Vecteurs de mot comme SVD

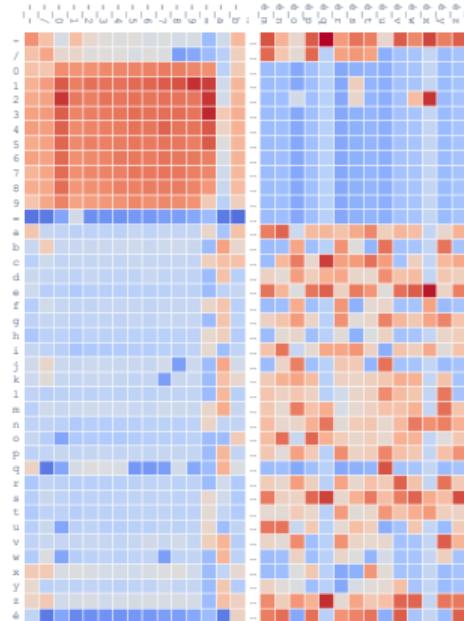
(Levy and Goldberg, 2014)



## Exemple: Caractères dans Wikipédia

$$W = \{-, /, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, =, a, b, c, \dots, w, x, y, z, é\}$$

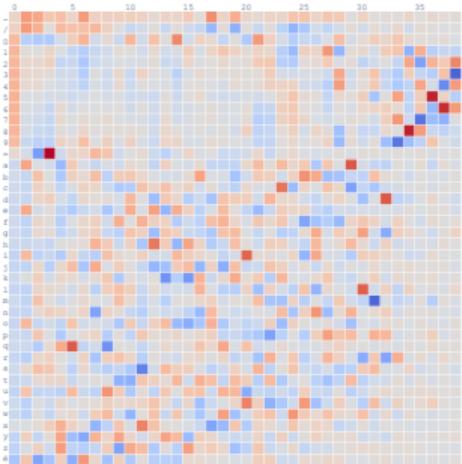
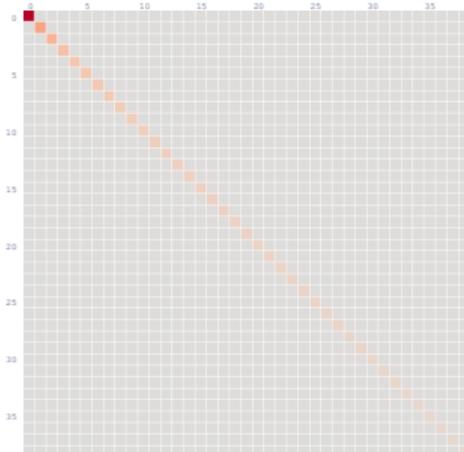
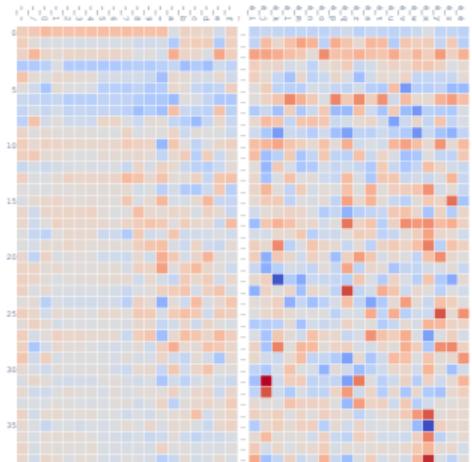
$$C = X \times X = \{(-, -), (-, /), (-, 0), \dots, (é, z), (é, é)\}$$



$$M_{wc} = \text{pmi}(w, c)$$

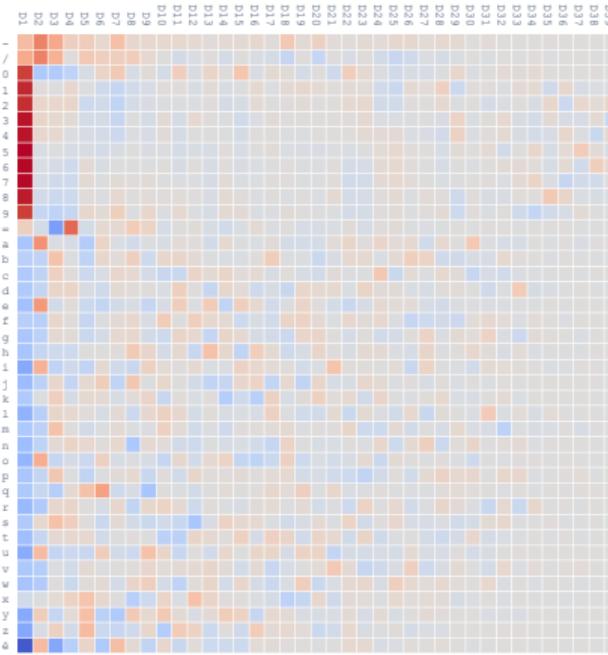
$$= \log \frac{p(w, c)}{p(w)p(c)}$$

# SVD d'une matrice pmi des caractères dans Wikipédia

 $U$  $\Sigma$  $V^T$ 

# Tronquer

$$U \times \Sigma$$



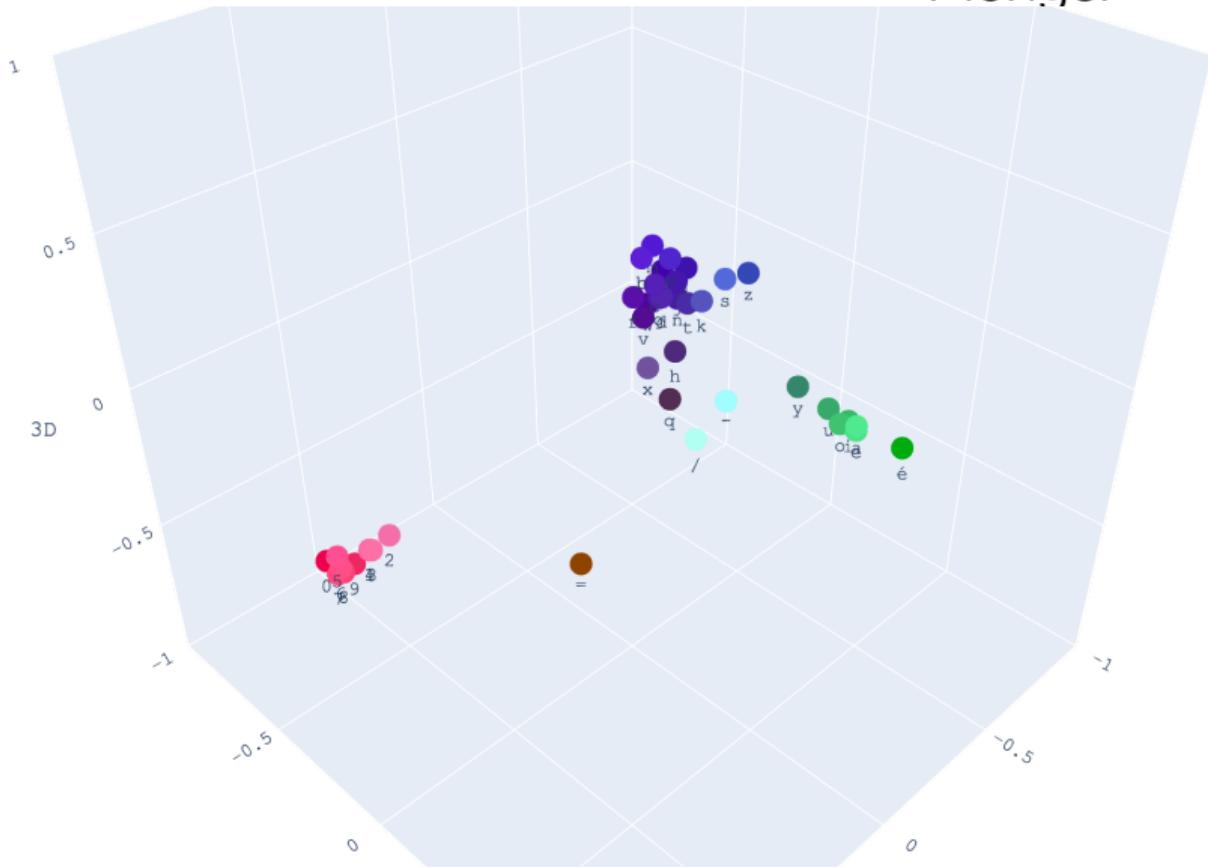
Tronquer

$\hat{U} \times \hat{\Sigma}$

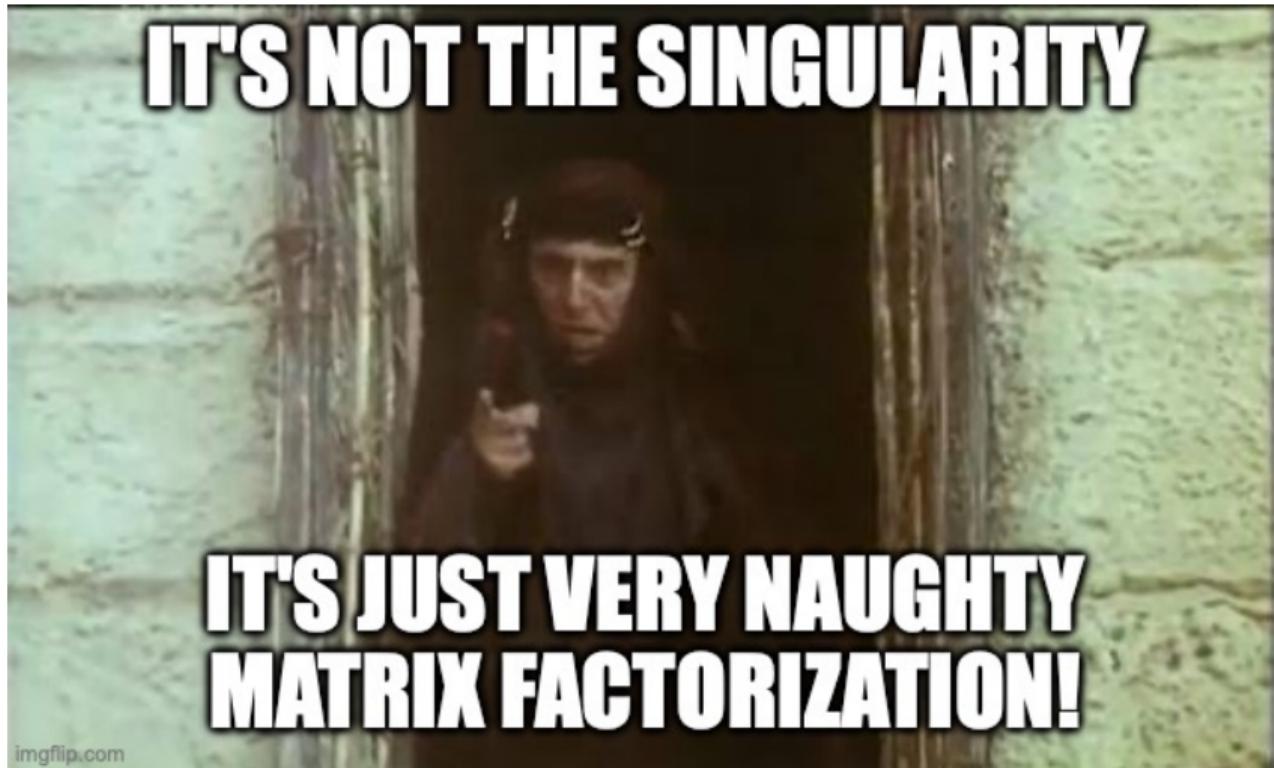


Plonger

$\hat{U} \times \hat{\Sigma}$



Que conclure?



imgflip.com

## 4 Why does this produce good word representations?

Good question. We don't really know.

The distributional hypothesis states that words in similar contexts have similar meanings. The objective above clearly tries to increase the quantity  $v_w \cdot v_c$  for good word-context pairs, and decrease it for bad ones. Intuitively, this means that words that share many contexts will be similar to each other (note also that contexts sharing many words will also be similar to each other). This is, however, very hand-wavy.

Can we make this intuition more precise? We'd really like to see something more formal.

(Goldberg and Levy, 2014)

# Plan

Introduction

LLMs comme des objets formels

La structure des ‘embeddings’

L’algèbre derrière les embeddings

La structure derrière l’algèbre

Les catégories derrière la structure

Conclusion

## Vecteurs de mots comme fonctions sur des ensembles

$$\textcolor{red}{X} = \{-, /, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, =, \text{a}, \text{b}, \text{c}, \dots, \text{w}, \text{x}, \text{y}, \text{z}, \text{é}\}$$

$$\textcolor{blue}{Y} = X \times X = \{ (-, -), (-, /), (-, 0), \dots, (\text{é}, \text{z}), (\text{é}, \text{é}) \}$$

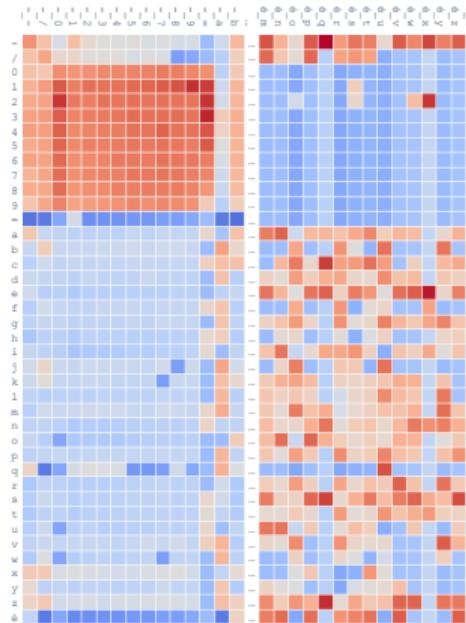
# Vecteurs de mots comme fonctions sur des ensembles

$$X = \{-, /, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, =, \text{a}, \text{b}, \text{c}, \dots, \text{w}, \text{x}, \text{y}, \text{z}, \text{é}\}$$

$$Y = X \times X = \{(-, -), (-, /), (-, 0), \dots, (\text{é}, \text{z}), (\text{é}, \text{é})\}$$

$$M: X \times Y \rightarrow \mathbb{R}$$

$$(\textcolor{red}{x}, \textcolor{blue}{y}) \mapsto \text{pmi}(\textcolor{red}{x}, \textcolor{blue}{y})$$



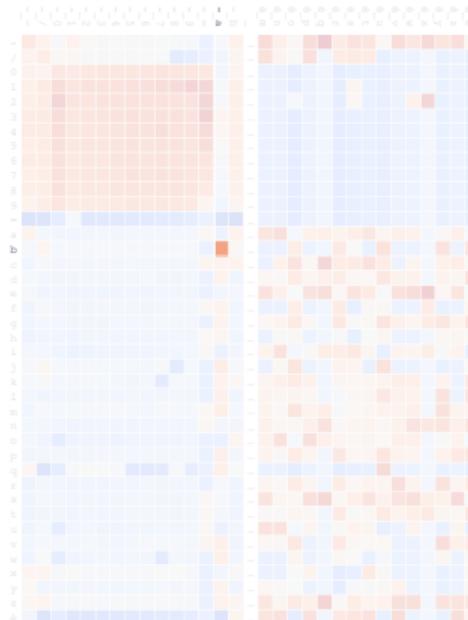
# Vecteurs de mots comme fonctions sur des ensembles

$$\textcolor{red}{X} = \{-, /, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, =, \text{a}, \text{b}, \text{c}, \dots, \text{w}, \text{x}, \text{y}, \text{z}, \text{é}\}$$

$$\textcolor{blue}{Y} = X \times X = \{ (-, -), (-, /), (-, 0), \dots, (\text{é}, \text{z}), (\text{é}, \text{é}) \}$$

$$M: \textcolor{red}{X} \times \textcolor{blue}{Y} \rightarrow \mathbb{R}$$

$$(\textcolor{red}{x}, \textcolor{blue}{y}) \mapsto \text{pmi}(\textcolor{red}{x}, \textcolor{blue}{y})$$



# Vecteurs de mots comme fonctions sur des ensembles

$$\textcolor{red}{X} = \{-, /, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, =, \text{a}, \text{b}, \text{c}, \dots, \text{w}, \text{x}, \text{y}, \text{z}, \text{é}\}$$

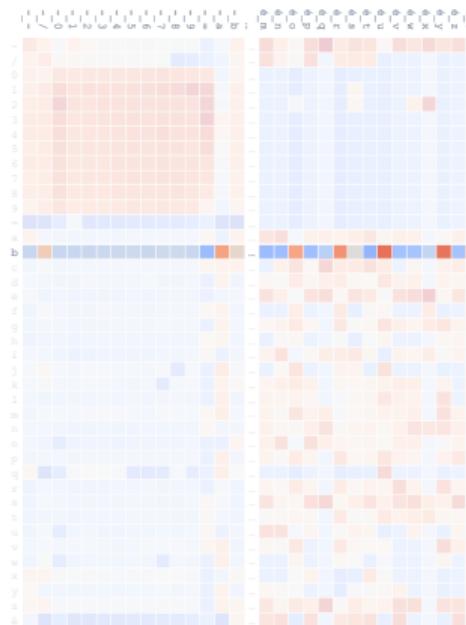
$$\textcolor{blue}{Y} = X \times X = \{(-, -), (-, /), (-, 0), \dots, (\text{é}, \text{z}), (\text{é}, \text{é})\}$$

$$M: \textcolor{red}{X} \times \textcolor{blue}{Y} \rightarrow \mathbb{R}$$

$$(\textcolor{red}{x}, \textcolor{blue}{y}) \mapsto \text{pmi}(\textcolor{red}{x}, \textcolor{blue}{y})$$

$$M_x: \textcolor{red}{X} \rightarrow \mathbb{R}^{\textcolor{blue}{Y}}$$

$$\textcolor{red}{x} \mapsto M(x, -)$$



# Vecteurs de mots comme fonctions sur des ensembles

$$\textcolor{red}{X} = \{-, /, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, =, \text{a}, \text{b}, \text{c}, \dots, \text{w}, \text{x}, \text{y}, \text{z}, \text{é}\}$$

$$\textcolor{blue}{Y} = X \times X = \{(-, -), (-, /), (-, 0), \dots, (\text{é}, \text{z}), (\text{é}, \text{é})\}$$

$$M: \textcolor{red}{X} \times \textcolor{blue}{Y} \rightarrow \mathbb{R}$$

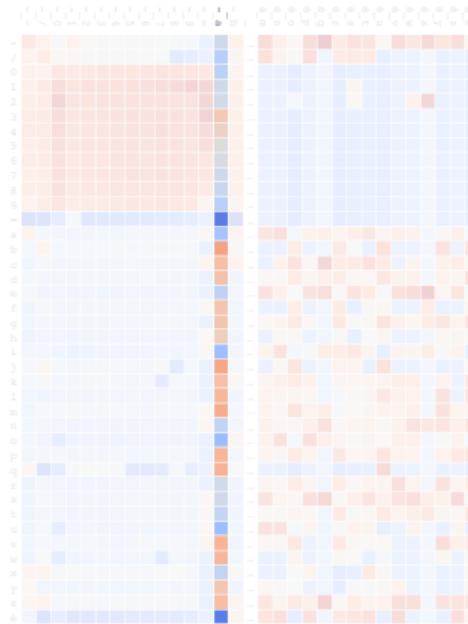
$$(\textcolor{red}{x}, \textcolor{blue}{y}) \mapsto \text{pmi}(\textcolor{red}{x}, \textcolor{blue}{y})$$

$$M_x: \textcolor{red}{X} \rightarrow \mathbb{R}^{\textcolor{blue}{Y}}$$

$$\textcolor{red}{x} \mapsto M(x, -)$$

$$M_y: \textcolor{blue}{Y} \rightarrow \mathbb{R}^{\textcolor{red}{X}}$$

$$y \mapsto M(-, y)$$



## Vecteurs de mots comme fonctions sur des ensembles

$$\textcolor{red}{X} = \{-, /, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, =, \text{a}, \text{b}, \text{c}, \dots, \text{w}, \text{x}, \text{y}, \text{z}, \text{é}\}$$

$$\textcolor{blue}{Y} = X \times X = \{(-, -), (-, /), (-, 0), \dots, (\text{é}, \text{z}), (\text{é}, \text{é})\}$$

$$M: \textcolor{red}{X} \times \textcolor{blue}{Y} \rightarrow \mathbb{R}$$

$$(\textcolor{red}{x}, \textcolor{blue}{y}) \mapsto \text{pmi}(\textcolor{red}{x}, \textcolor{blue}{y})$$

$$\textcolor{red}{X} \xrightarrow{M_x} \mathbb{R}^{\textcolor{blue}{Y}}$$

$$M_x: \textcolor{red}{X} \rightarrow \mathbb{R}^{\textcolor{blue}{Y}}$$

$$\textcolor{red}{x} \mapsto M(x, -)$$

$$\mathbb{R}^{\textcolor{red}{X}} \xleftarrow{M_y} \textcolor{blue}{Y}$$

$$M_y: \textcolor{blue}{Y} \rightarrow \mathbb{R}^{\textcolor{red}{X}}$$

$$\textcolor{blue}{y} \mapsto M(-, y)$$

# Vecteurs de mots comme fonctions sur des ensembles

$$\textcolor{red}{X} = \{-, /, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, =, \text{a}, \text{b}, \text{c}, \dots, \text{w}, \text{x}, \text{y}, \text{z}, \text{é}\}$$

$$\textcolor{blue}{Y} = X \times X = \{(-, -), (-, /), (-, 0), \dots, (\text{é}, \text{z}), (\text{é}, \text{é})\}$$

$$M: \textcolor{red}{X} \times \textcolor{blue}{Y} \rightarrow \mathbb{R}$$

$$(\textcolor{red}{x}, \textcolor{blue}{y}) \mapsto \text{pmi}(\textcolor{red}{x}, \textcolor{blue}{y})$$

$$M_x: \textcolor{red}{X} \rightarrow \mathbb{R}^{\textcolor{blue}{Y}}$$

$$\textcolor{red}{x} \mapsto \textcolor{blue}{M}(x, -)$$

$$M_y: \textcolor{blue}{Y} \rightarrow \mathbb{R}^{\textcolor{red}{X}}$$

$$\textcolor{blue}{y} \mapsto \textcolor{red}{M}(-, y)$$

$$\begin{array}{ccc} \textcolor{red}{X} & \xrightarrow{M_x} & \mathbb{R}^{\textcolor{blue}{Y}} \\ \downarrow & & \uparrow \\ \mathbb{R}^{\textcolor{red}{X}} & \xleftarrow{M_y} & \textcolor{blue}{Y} \end{array}$$

# Vecteurs de mots comme fonctions sur des ensembles

$$\textcolor{red}{X} = \{-, /, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, =, \text{a}, \text{b}, \text{c}, \dots, \text{w}, \text{x}, \text{y}, \text{z}, \text{é}\}$$

$$\textcolor{blue}{Y} = X \times X = \{(-, -), (-, /), (-, 0), \dots, (\text{é}, \text{z}), (\text{é}, \text{é})\}$$

$$M: \textcolor{red}{X} \times \textcolor{blue}{Y} \rightarrow \mathbb{R}$$

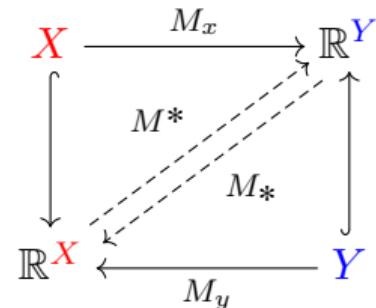
$$(\textcolor{red}{x}, \textcolor{blue}{y}) \mapsto \text{pmi}(\textcolor{red}{x}, \textcolor{blue}{y})$$

$$M_x: \textcolor{red}{X} \rightarrow \mathbb{R}^{\textcolor{blue}{Y}}$$

$$\textcolor{red}{x} \mapsto M(x, -)$$

$$M_y: \textcolor{blue}{Y} \rightarrow \mathbb{R}^{\textcolor{red}{X}}$$

$$y \mapsto M(-, y)$$

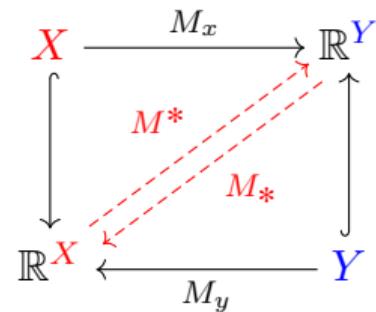


$$M^*: \mathbb{R}^{\textcolor{red}{X}} \rightarrow \mathbb{R}^{\textcolor{blue}{Y}}$$

$$M_*: \mathbb{R}^{\textcolor{blue}{Y}} \rightarrow \mathbb{R}^{\textcolor{red}{X}}$$

# Vecteurs de mots comme fonctions sur des ensembles

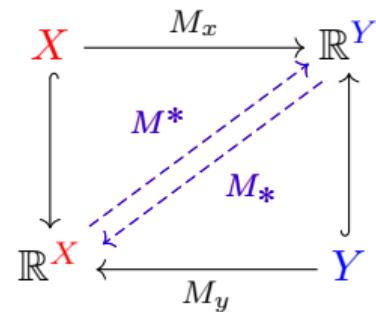
$$M_* M^* : \mathbb{R}^X \rightarrow \mathbb{R}^X$$



# Vecteurs de mots comme fonctions sur des ensembles

$$M_* M^* : \mathbb{R}^X \rightarrow \mathbb{R}^X$$

$$M^* M_* : \mathbb{R}^Y \rightarrow \mathbb{R}^Y$$



# Vecteurs de mots comme fonctions sur des ensembles

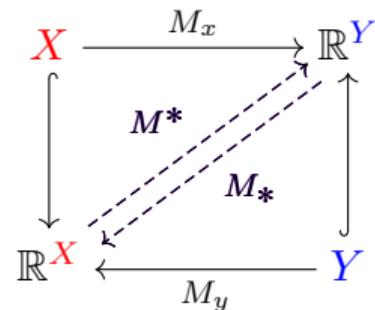
$$M_* M^* : \mathbb{R}^X \rightarrow \mathbb{R}^X$$

$$M^* M_* : \mathbb{R}^Y \rightarrow \mathbb{R}^Y$$

$$\{u_1, \dots, u_m\} \subset \mathbb{R}^X$$

$$\{v_1, \dots, v_n\} \subset \mathbb{R}^Y$$

$$\{\lambda_1, \dots, \lambda_{\min(m,n)}, 0, \dots, 0\}$$



# Vecteurs de mots comme fonctions sur des ensembles

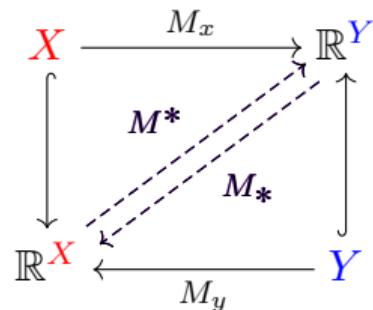
$$M_* M^* : \mathbb{R}^X \rightarrow \mathbb{R}^X$$

$$M^* M_* : \mathbb{R}^Y \rightarrow \mathbb{R}^Y$$

$$\{u_1, \dots, u_m\} \subset \mathbb{R}^X$$

$$\{v_1, \dots, v_n\} \subset \mathbb{R}^Y$$

$$\{\lambda_1, \dots, \lambda_{\min(m,n)}, 0, \dots, 0\}$$



$$U := [\textcolor{red}{u}_1, \dots, \textcolor{red}{u}_m]$$

$$M = U \Sigma V^T \quad V := [\textcolor{blue}{v}_1, \dots, \textcolor{blue}{v}_n]$$

$$\Sigma := \begin{bmatrix} \sqrt{\lambda_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{\lambda_r} \end{bmatrix}$$

# Vecteurs de mots comme fonctions sur des ensembles

$$M_* M^* : \mathbb{R}^X \rightarrow \mathbb{R}^X$$

$$M^* M_* : \mathbb{R}^Y \rightarrow \mathbb{R}^Y$$

$$\{u_1, \dots, u_m\} \subset \mathbb{R}^X$$

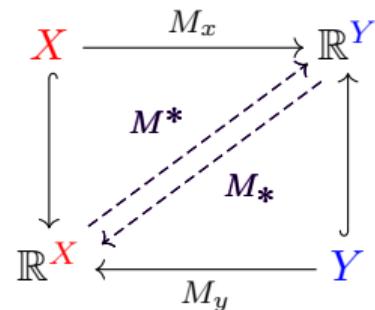
$$\{v_1, \dots, v_n\} \subset \mathbb{R}^Y$$

$$\{\lambda_1, \dots, \lambda_{\min(m,n)}, 0, \dots, 0\}$$

$$M_* M^* u_i = \lambda_i u_i$$

$$M^* M_* v_i = \lambda_i v_i$$

Les  $u_i$  and  $v_i$  sont des points fixes (linéaires)!

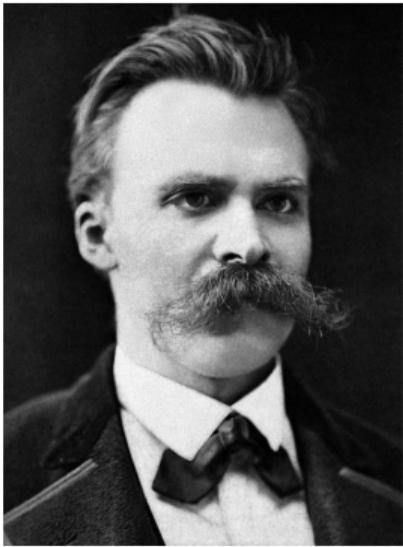


$$U := [\underline{u_1}, \dots, \underline{u_m}]$$

$$M = U \Sigma V^T \quad V := [\underline{v_1}, \dots, \underline{v_n}]$$

$$\Sigma := \begin{bmatrix} \sqrt{\lambda_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{\lambda_r} \end{bmatrix}$$

## Figures de Chladni



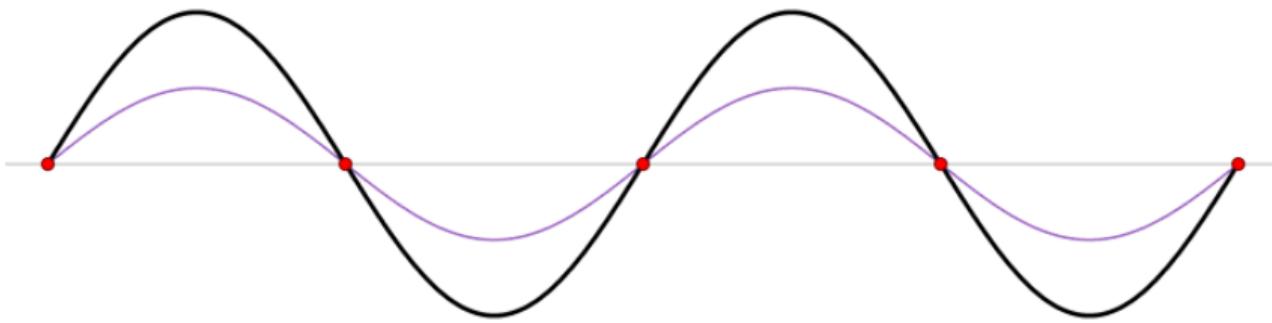
“On peut imaginer un homme qui soit totalement sourd, et n'ait jamais ressenti une sensation sonore et musicale: tout comme cet homme considérera avec stupéfaction **les figures acoustiques de Chladni dans le sable**, trouvera leur cause dans la vibration de la corde et jurera ensuite qu'**il sait nécessairement à présent ce que les hommes appellent le son**, c'est ainsi qu'il en va pour nous tous avec le **langage**.”

(Nietzsche, 1873)

## Figures de Chladni

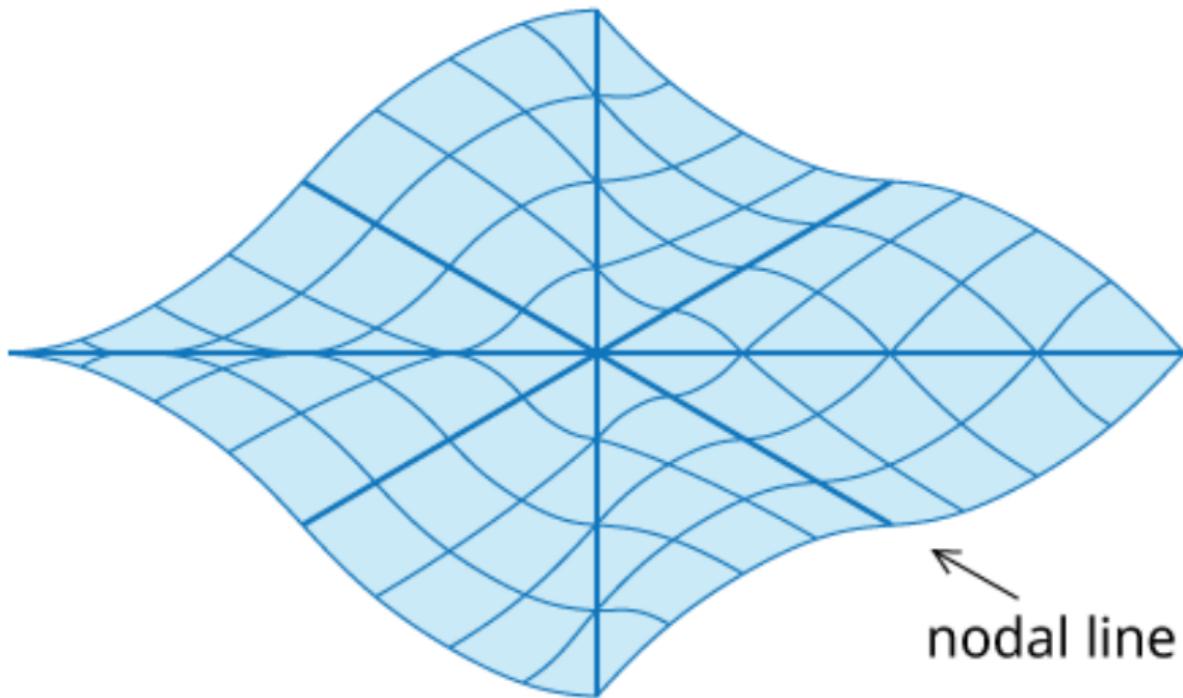


# Figures de Chladni

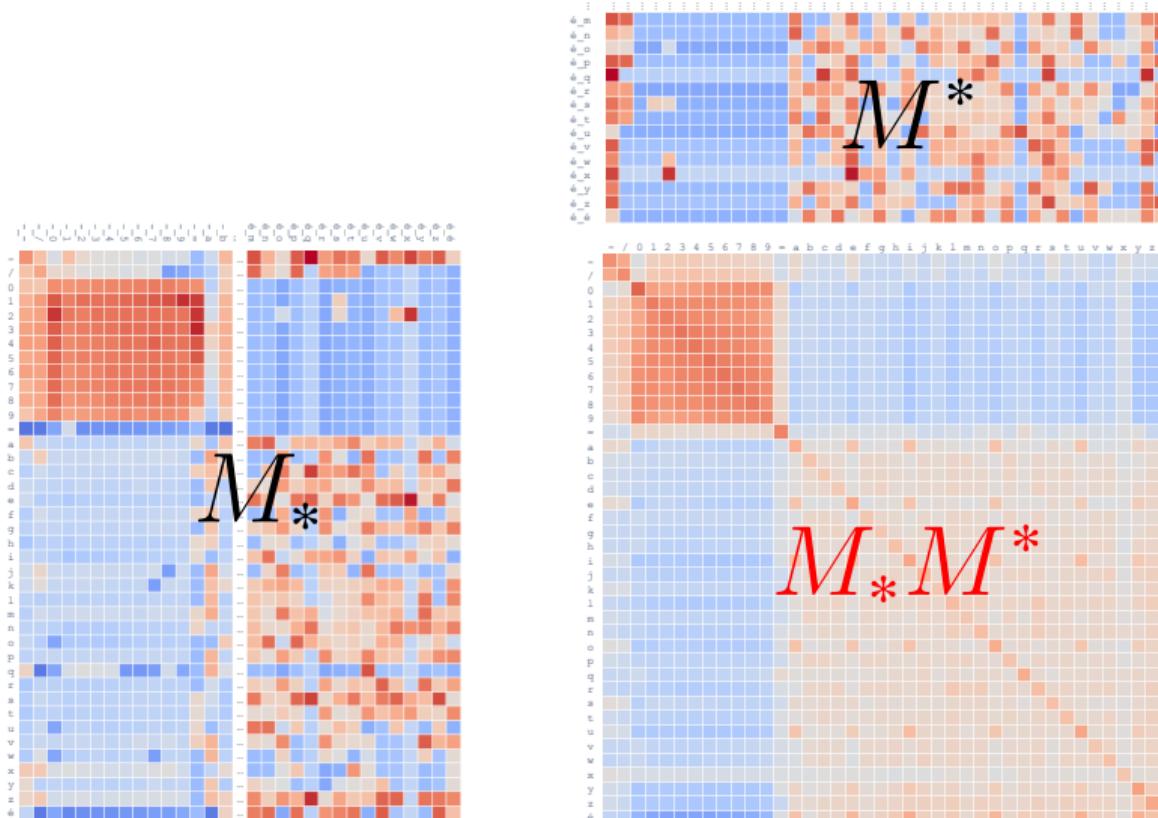


Figures de Chladni

## Chladni Plate

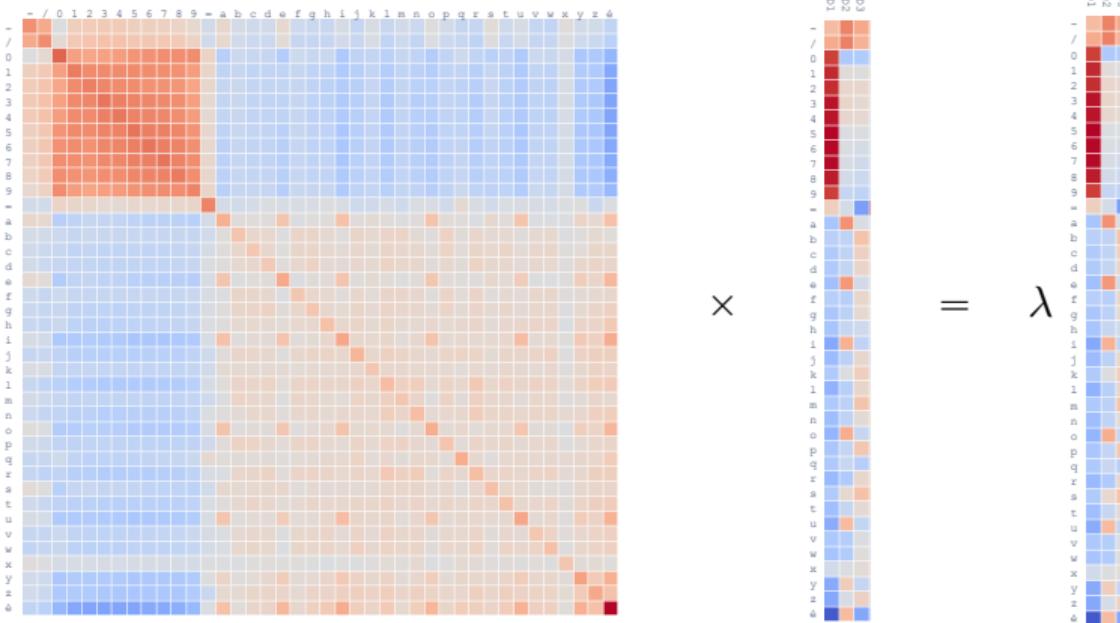


# $M_* M^*$ comme matrice de covariance



# Vecteurs propres comme points fixes

$$M_* M^* u = \lambda u$$

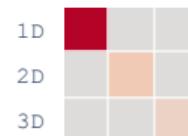


# Traits structuraux

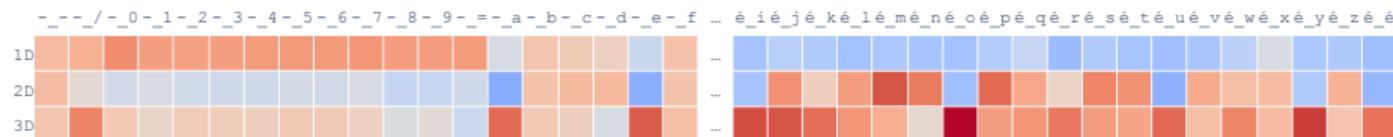
Eigenvectors of  $M_* M^*$ :



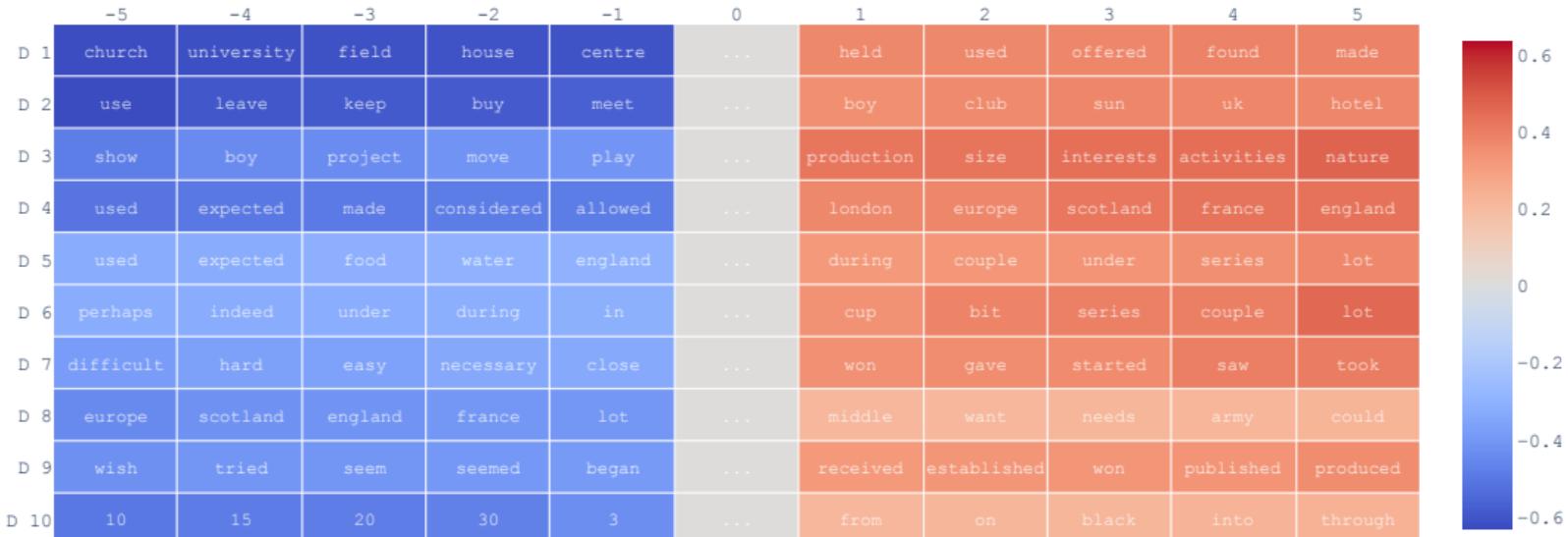
Eigenvalues of  $M_* M^*$  and  $M^* M_*$ :



Eigenvectors of  $M^* M_*$ :



# Mots



# Plan

Introduction

LLMs comme des objets formels

La structure des ‘embeddings’

L’algèbre derrière les embeddings

La structure derrière l’algèbre

Les catégories derrière la structure

Conclusion

# Vecteurs de mots comme foncteurs sur des catégories

$$\textcolor{red}{X} = \{-, /, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, =, \text{a}, \text{b}, \text{c}, \dots, \text{w}, \text{x}, \text{y}, \text{z}, \text{é}\}$$

$$\textcolor{blue}{Y} = X \times X = \{(-, -), (-, /), (-, 0), \dots, (\text{é}, z), (\text{é}, \text{é})\}$$

$$M: \textcolor{red}{X} \times \textcolor{blue}{Y} \rightarrow \mathbb{R}$$

$$(\textcolor{red}{x}, \textcolor{blue}{y}) \mapsto \text{pmi}(\textcolor{red}{x}, \textcolor{blue}{y})$$

$$M_x: \textcolor{red}{X} \rightarrow \mathbb{R}^{\textcolor{blue}{Y}}$$

$$\textcolor{red}{x} \mapsto M(x, -)$$

$$M_y: \textcolor{blue}{Y} \rightarrow \mathbb{R}^{\textcolor{red}{X}}$$

$$\textcolor{blue}{y} \mapsto M(-, y)$$

$$\begin{array}{ccc} \textcolor{red}{X} & \xrightarrow{M_x} & \mathbb{R}^{\textcolor{blue}{Y}} \\ \downarrow & \nearrow M^* & \downarrow \\ \mathbb{R}^{\textcolor{red}{X}} & \xleftarrow[M_y]{\quad} & \textcolor{blue}{Y} \end{array}$$

$$M^*: \mathbb{R}^{\textcolor{red}{X}} \rightarrow \mathbb{R}^{\textcolor{blue}{Y}}$$

$$M_*: \mathbb{R}^{\textcolor{blue}{Y}} \rightarrow \mathbb{R}^{\textcolor{red}{X}}$$

# Vecteurs de mots comme foncteurs sur des catégories

$$\textcolor{orange}{C} = \{-, /, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, =, \text{a}, \text{b}, \text{c}, \dots, \text{w}, \text{x}, \text{y}, \text{z}, \text{é}\}$$

$$\textcolor{green}{D} = \textcolor{orange}{C} = \{-, /, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, =, \text{a}, \text{b}, \text{c}, \dots, \text{w}, \text{x}, \text{y}, \text{z}, \text{é}\}$$

Profunctor

$$\mathcal{M}: \textcolor{orange}{C}^{\text{op}} \times \textcolor{green}{D} \rightarrow \text{Set}$$

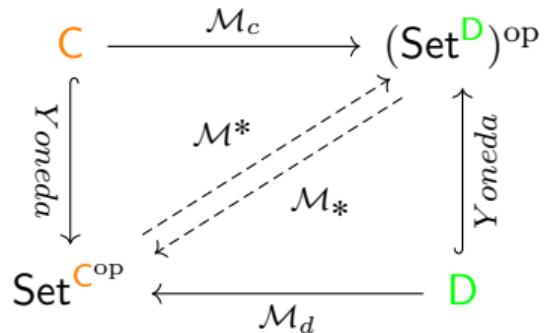
$$(\textcolor{orange}{c}, \textcolor{green}{d}) \mapsto \mathcal{M}(\textcolor{orange}{c}, \textcolor{green}{d})$$

$$\mathcal{M}_c: \textcolor{orange}{C} \rightarrow (\text{Set}^{\textcolor{green}{D}})^{\text{op}}$$

$$\textcolor{orange}{c} \mapsto \mathcal{M}(\textcolor{orange}{c}, -)$$

$$\mathcal{M}_d: \textcolor{green}{D} \rightarrow \text{Set}^{\textcolor{orange}{C}^{\text{op}}}$$

$$\textcolor{green}{d} \mapsto \mathcal{M}(-, \textcolor{green}{d})$$



$$\mathcal{M}^*: \text{Set}^{\textcolor{orange}{C}^{\text{op}}} \rightarrow (\text{Set}^{\textcolor{green}{D}})^{\text{op}}$$

$$\mathcal{M}_*: (\text{Set}^{\textcolor{green}{D}})^{\text{op}} \rightarrow \text{Set}^{\textcolor{orange}{C}^{\text{op}}}$$

# Vecteurs de mots comme foncteurs sur des catégories

Adjonction d'Isbell

$$\mathcal{M}^*: \text{Set}^{\text{C}^{\text{op}}} \leftrightarrows (\text{Set}^{\text{D}})^{\text{op}}: \mathcal{M}_*$$

$$\mathcal{M}_* \mathcal{M}^*: \text{Set}^{\text{C}^{\text{op}}} \rightarrow \text{Set}^{\text{C}^{\text{op}}}$$

$$\mathcal{M}^* \mathcal{M}_*: (\text{Set}^{\text{D}})^{\text{op}} \rightarrow (\text{Set}^{\text{D}})^{\text{op}}$$

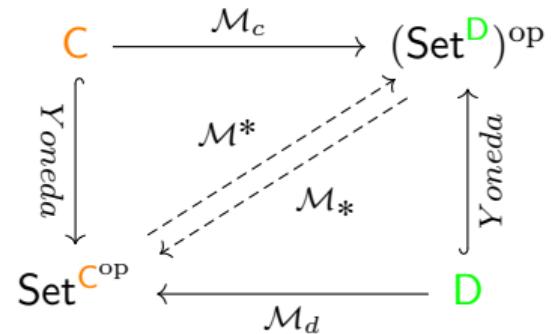
$$\text{Fix}(\mathcal{M}_* \mathcal{M}^*) := \{f \in \text{Set}^{\text{C}^{\text{op}}} \mid \mathcal{M}_* \mathcal{M}^*(f) \cong f\}$$

$$\text{Fix}(\mathcal{M}^* \mathcal{M}_*) := \{g \in (\text{Set}^{\text{D}})^{\text{op}} \mid \mathcal{M}^* \mathcal{M}_*(g) \cong g\}$$

Nucleus of  $\mathcal{M} = \{(f_i, g_i)\}$ , such that:

$$\mathcal{M}^* f_i \cong g_i \text{ and } \mathcal{M}_* g_i \cong f_i$$

Le **noyau** (nucleus) est une **catégorie complète et cocomplète**



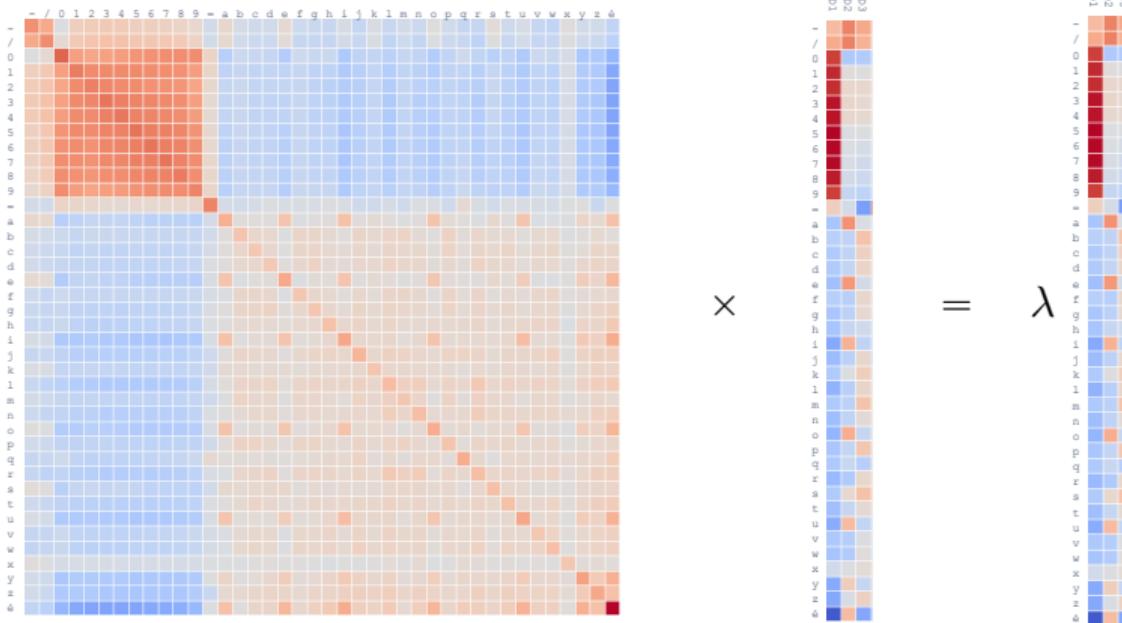
Les catégories **C** et **D** peuvent être enrichies!

E.g.:

$$\begin{aligned} \mathcal{M}^*: \mathbf{2}^{\text{C}^{\text{op}}} &\leftrightarrows (\mathbf{2}^{\text{D}})^{\text{op}}: \mathcal{M}_* \\ \mathcal{M}^*: \bar{\mathbb{R}}^{\text{C}^{\text{op}}} &\leftrightarrows (\bar{\mathbb{R}}^{\text{D}})^{\text{op}}: \mathcal{M}_* \end{aligned}$$

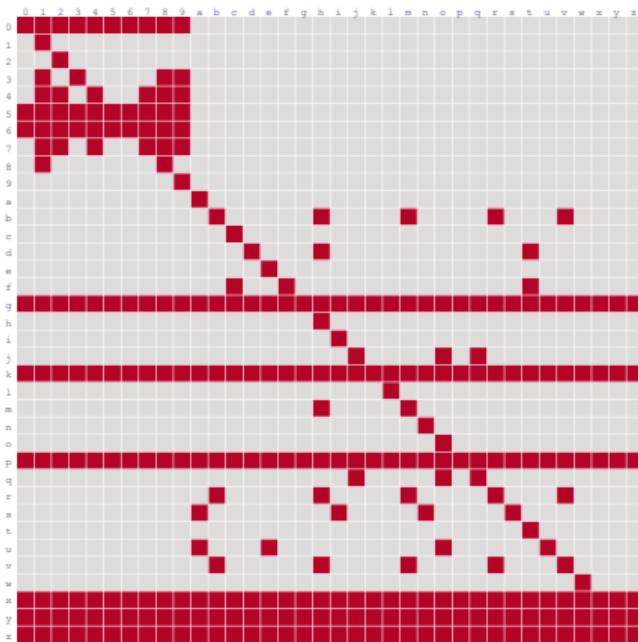
# Points fixes booléens

$$M_* M^* u = \lambda u$$

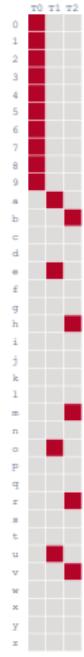


# Points fixes booléens

$$\mathcal{M}_*\mathcal{M}^*f = f$$



★

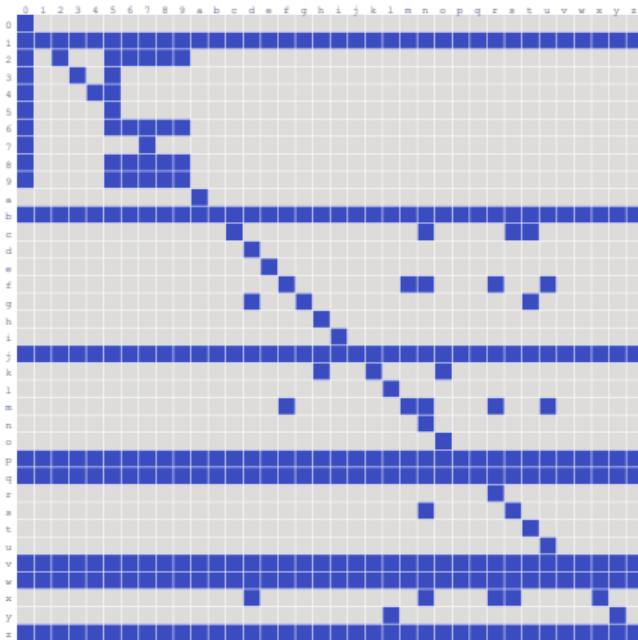


?



# Points fixes booléens

$$M_i^* M_*^i \textcolor{blue}{d} = d$$



★



?



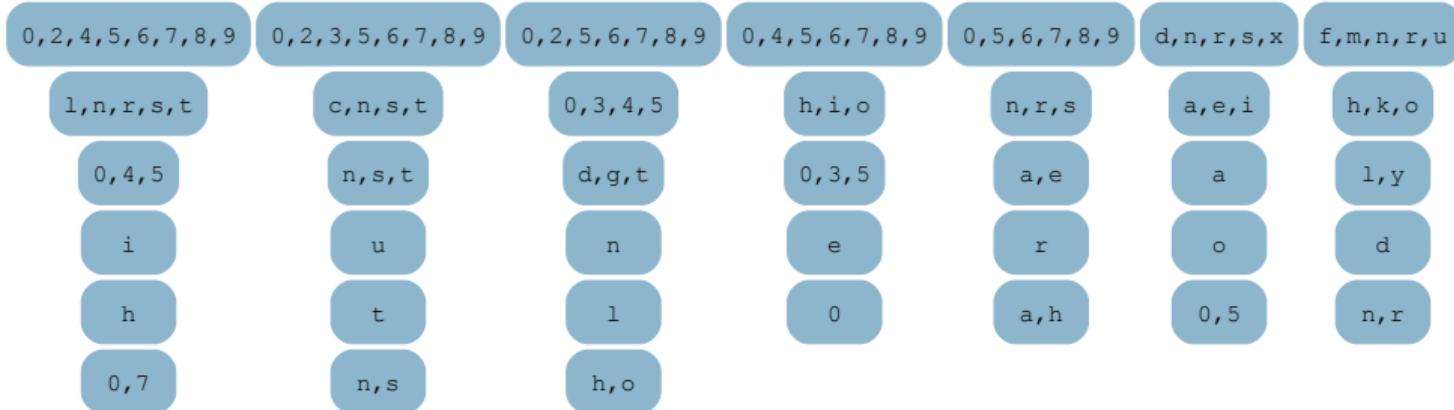
# “Eigensets”

$$\mathcal{M}_*\mathcal{M}^*f = f$$

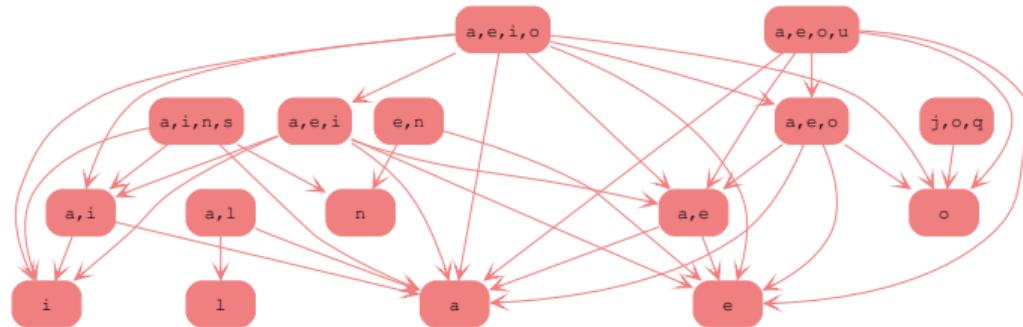
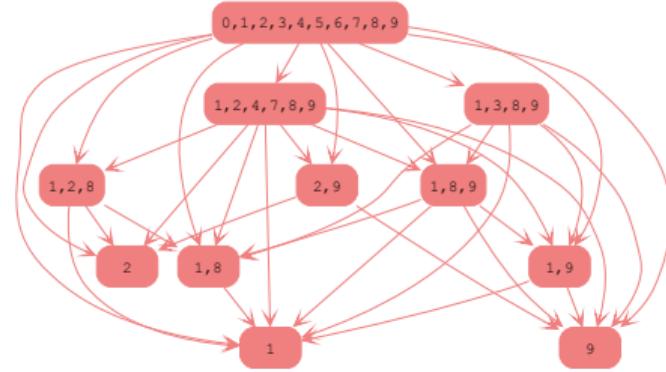
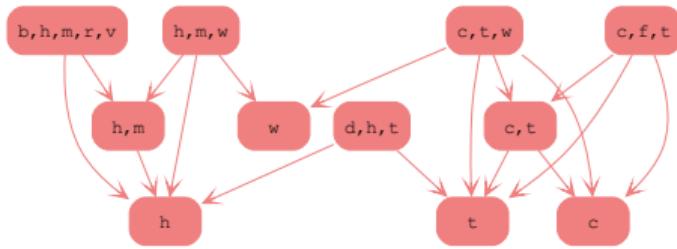
0,1,2,3,4,5,6,7,8,9	1,2,4,7,8,9	b,h,m,r,v	a,e,i,o	a,e,o,u	a,i,n,s	1,3,8,9
1,2,8	h,m,w	1,8,9	d,h,t	j,o,q	c,f,t	c,t,w
a,e,o	a,e,i	h,m	2,9	a,i	w	1,9
1,8	a,e	l	t	n	c	h
2	i	e	a	o	1	9
e,n	a,l	c,t				

# “Eigensets”

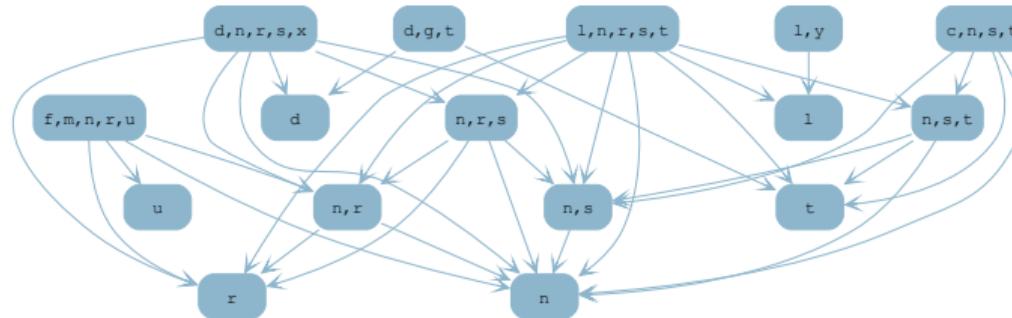
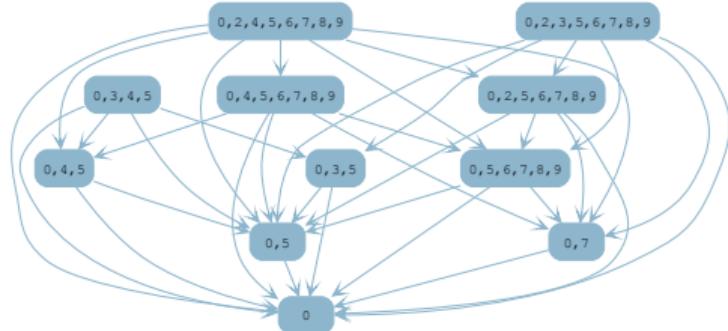
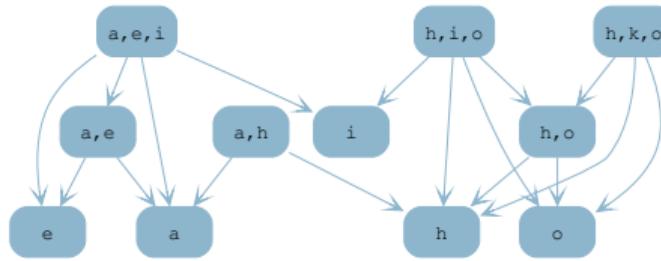
$$M_i^* M_*^i \textcolor{blue}{d} = d$$



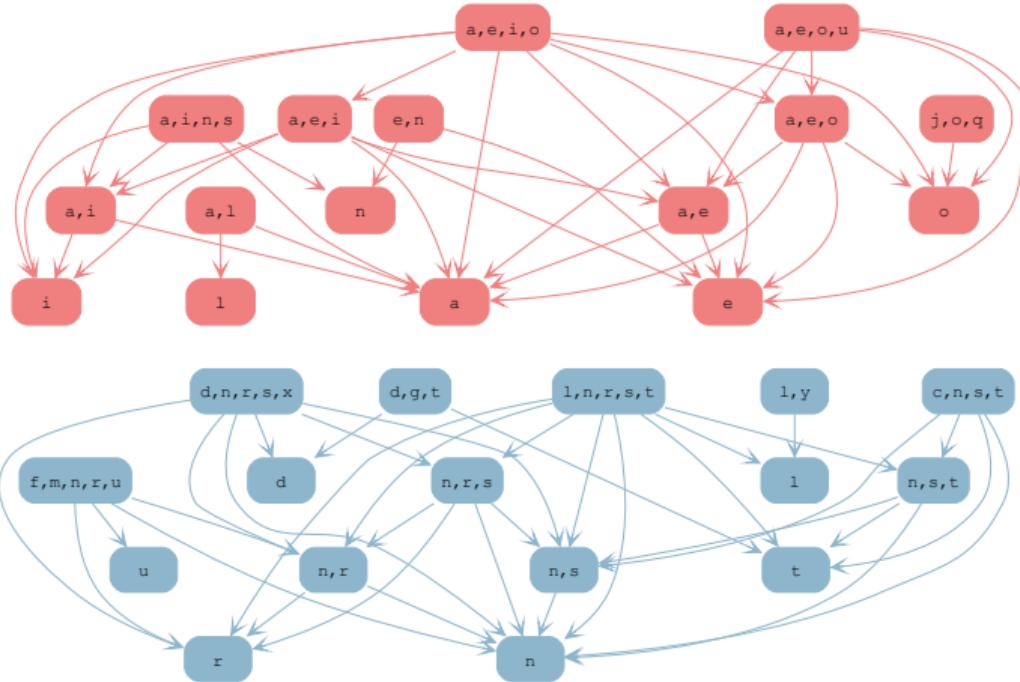
# Quelle Structure?



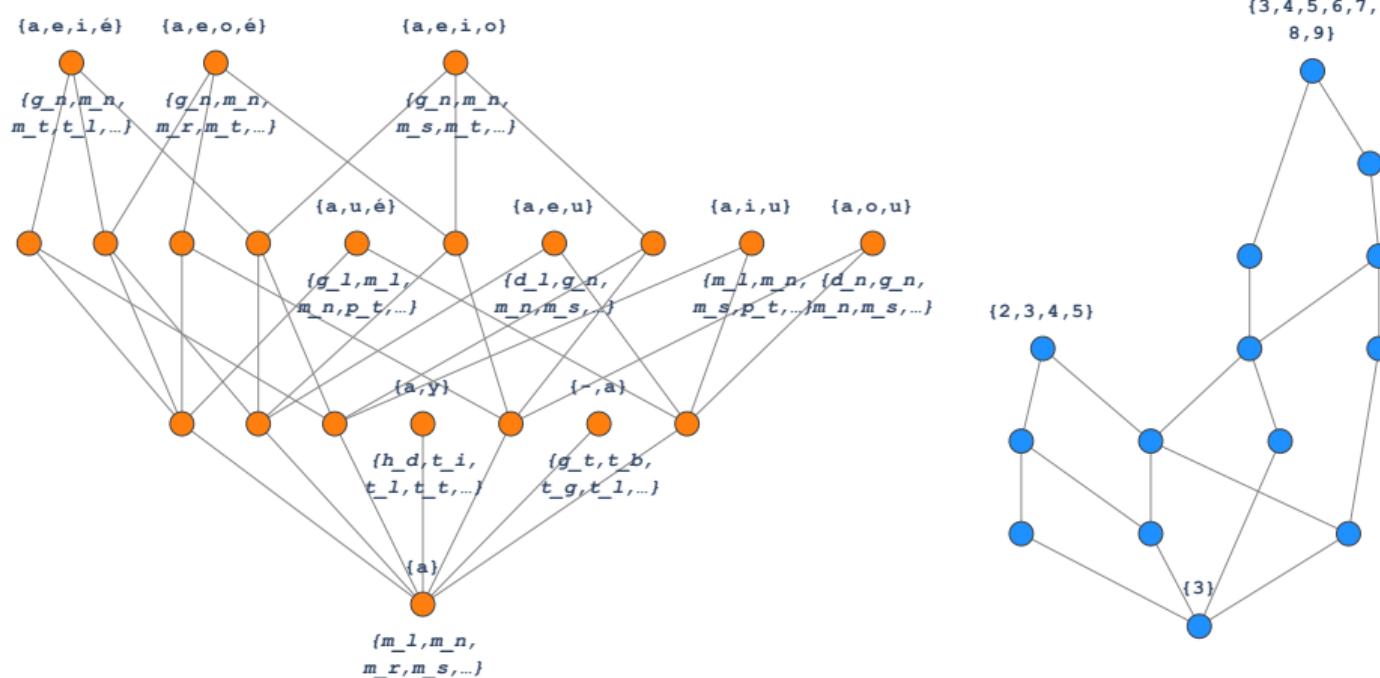
# Quelle Structure?



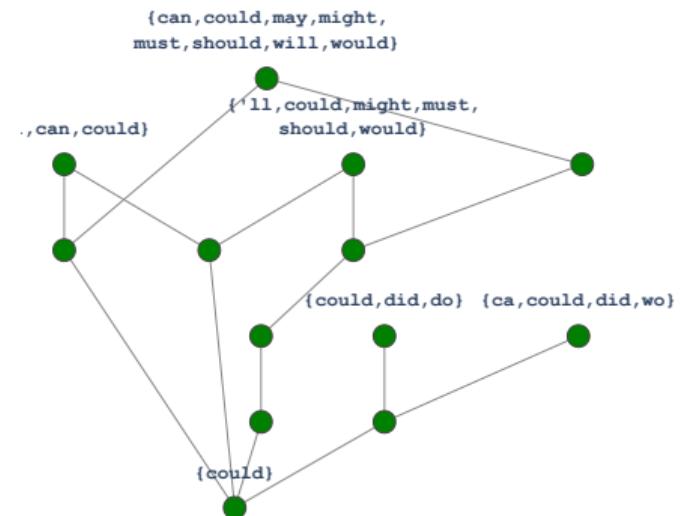
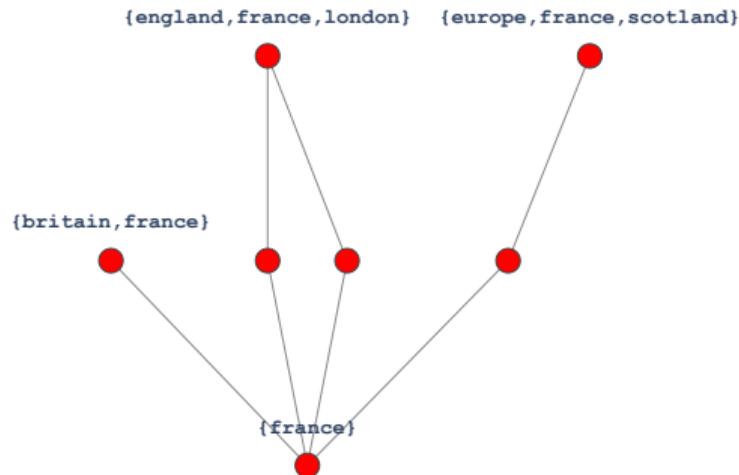
# Quelle Structure?



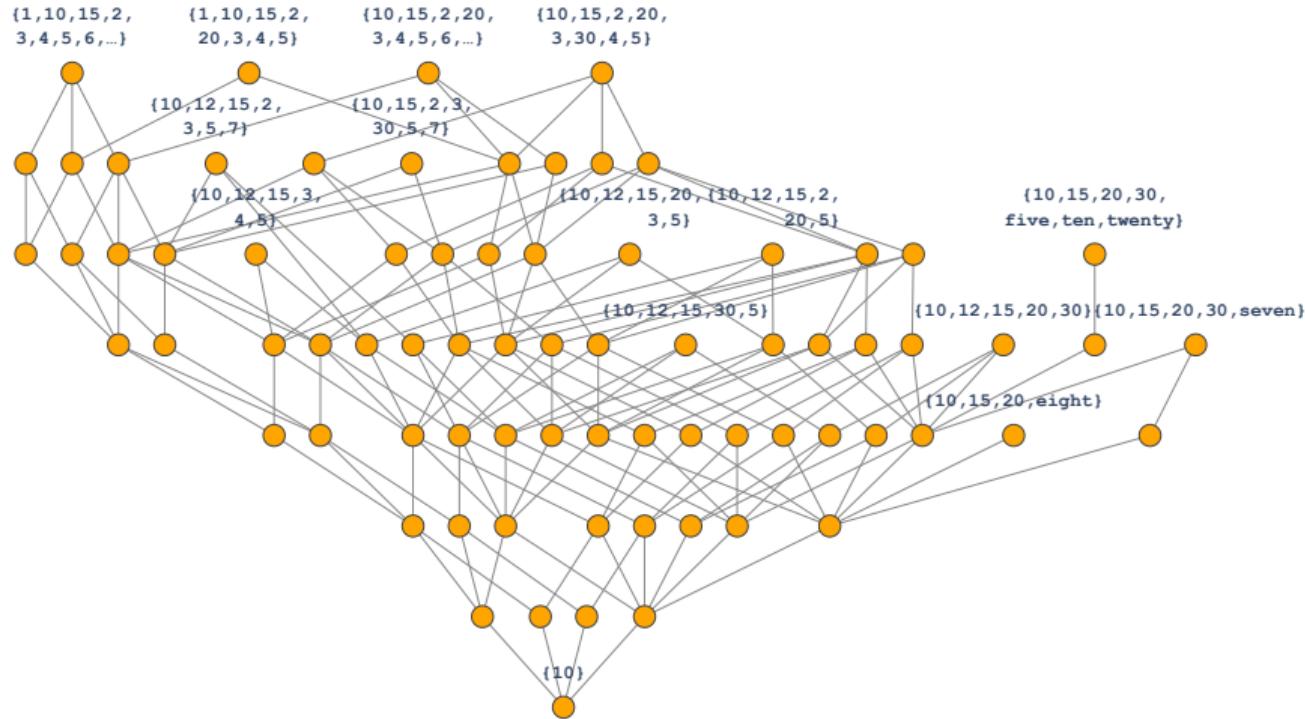
# Concepts formels



# Concepts formels (mots)



# Concepts formels (mots)

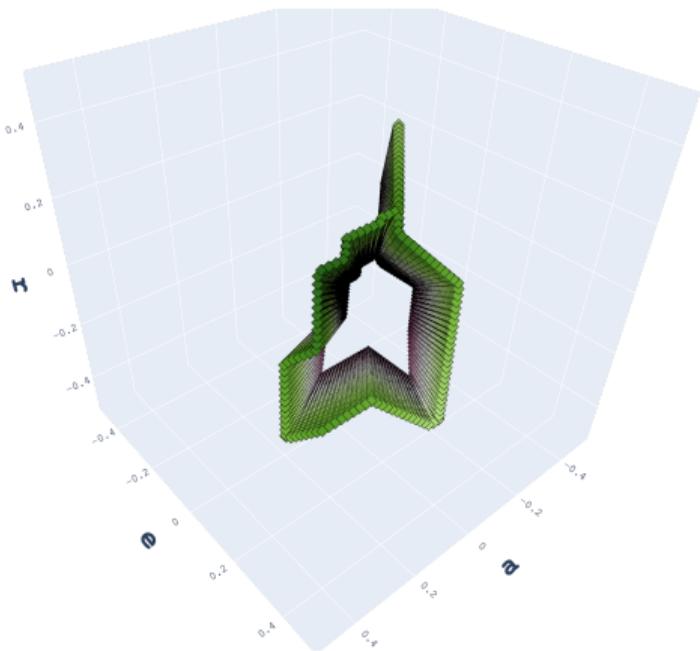
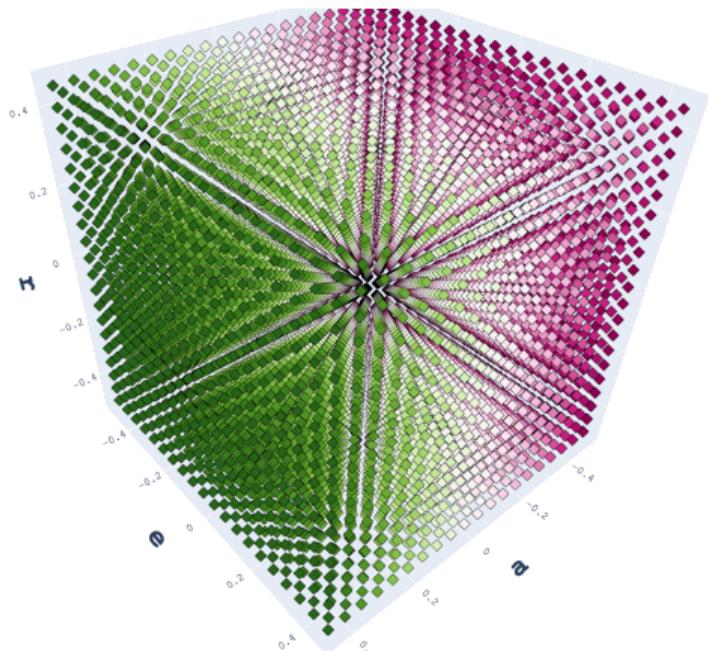


Chladni

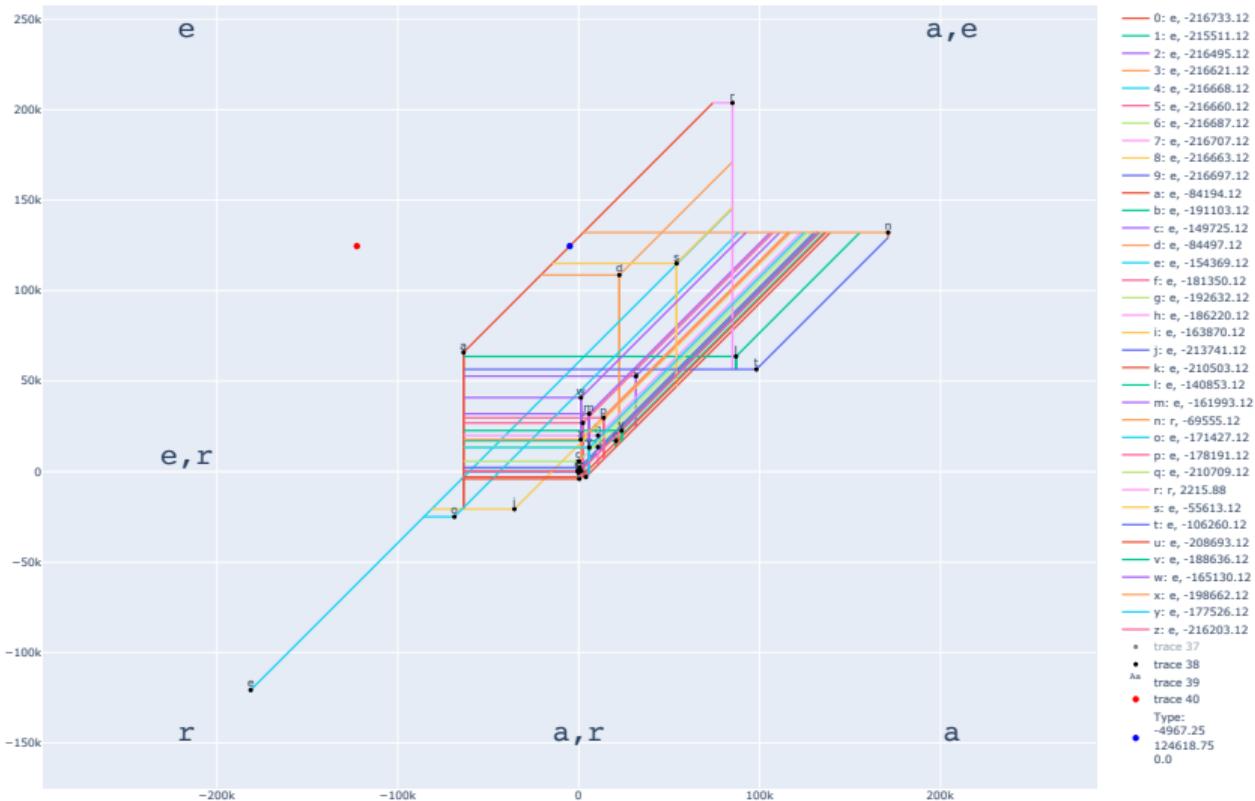


## Noyau (nucleus)

$$\bar{\mathbb{R}}^{\{a,e,r\}} \xrightarrow{\mathcal{M}_*\mathcal{M}^*} \bar{\mathbb{R}}^{\{a,e,r\}}$$



# Structure interne du noyau



# Théorie des types computationnels

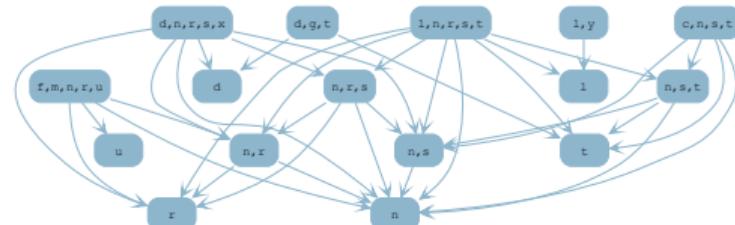
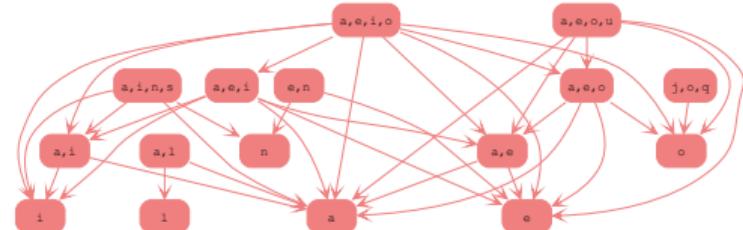
Definition (Polaire/Orthogonal - Girard, 2006)

[É]tant donnée une fonction binaire  $a, b \rightsquigarrow \langle a|b \rangle : A \times B \rightarrow C$  et un sous-ensemble  $P \subset C$  (le « pôle »), on peut définir le *polaire*  $X^\perp \subset B$  d'un sous-ensemble  $X \subset A$  (resp.  $Y^\perp \subset A$  d'un sous-ensemble  $Y \subset B$ ) par :

$$X^\perp := \{y \in B : \forall x \in X, \langle a|b \rangle \in P\}$$

$$Y^\perp := \{x \in A : \forall y \in Y, \langle a|b \rangle \in P\}$$

- ◊ L'application « polaire » est décroissante:  
 $X \subset X' \Rightarrow X'^\perp \subset X^\perp$ .
- ◊ L'ensemble  $\text{Pol}(A) \subset \mathcal{P}(A)$  des ensembles *polaires*, i.e., de la forme  $Y^\perp$ , est stable par intersections arbitraires. En particulier,  $A$  est polaire et  $X^{\perp\perp}$  est le plus petit ensemble polaire contenant  $X$ .
- ◊ En conséquence,  $X^{\perp\perp\perp} = X^\perp$ .



# Théorie des types computationnels

Definition (Polaire/Orthogonal - Girard, 2006)

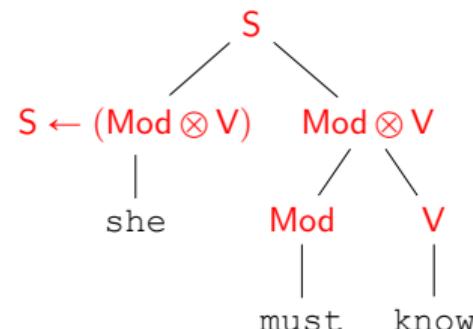
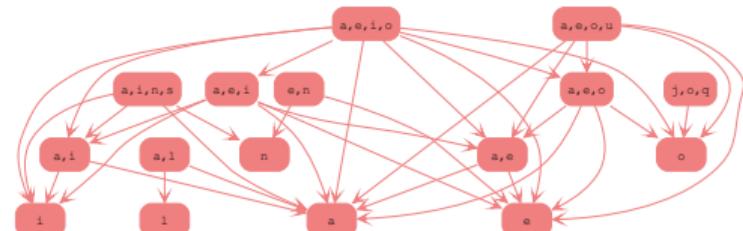
[É]tant donnée une fonction binaire

$a, b \rightsquigarrow \langle a|b \rangle : A \times B \rightarrow C$  et un sous-ensemble  $P \subset C$  (le « pôle »), on peut définir le *polaire*  $X^\perp \subset B$  d'un sous-ensemble  $X \subset A$  (resp.  $Y^\perp \subset A$  d'un sous-ensemble  $Y \subset B$ ) par :

$$X^\perp := \{y \in B : \forall x \in X, \langle a|b \rangle \in P\}$$

$$Y^\perp := \{x \in A : \forall y \in Y, \langle a|b \rangle \in P\}$$

- ◊ L'application « polaire » est décroissante:  
 $X \subset X' \Rightarrow X'^\perp \subset X^\perp$ .
- ◊ L'ensemble  $\text{Pol}(A) \subset \mathcal{P}(A)$  des ensembles *polaires*, i.e., de la forme  $Y^\perp$ , est stable par intersections arbitraires. En particulier,  $A$  est polaire et  $X^{\perp\perp}$  est le plus petit ensemble polaire contenant  $X$ .
- ◊ En conséquence,  $X^{\perp\perp\perp} = X^\perp$ .



(Gastaldi and Pellissier, 2021)

# Plan

Introduction

LLMs comme des objets formels

La structure des ‘embeddings’

L’algèbre derrière les embeddings

La structure derrière l’algèbre

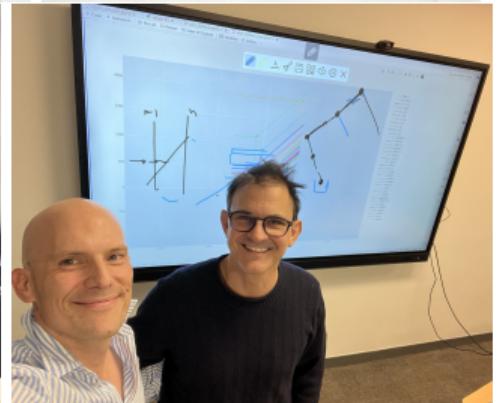
Les catégories derrière la structure

Conclusion

# Conclusion: Pour un formalisme critique

- ◊ Il est urgent d'aborder la dimension **épistémologique** du projet critique pour pouvoir aborder l'IA critiquement
- ◊ Cela nécessite de développer une **approche critique au sein des sciences formelles** où la formalisation n'est pas supposée conduire à une **naturalisation**.
  - Le nouveau rôle des **données** au sein des sciences formelles est crucial en ce sens
- ◊ Un **formalisme critique** sera incomplet s'il reste déconnecté de la dimension **politique** (voire **esthétique**) associée aux données.
  - Nous avons besoin d'une **nouvelle alliance** entre les **sciences formelles** et les **sciences humaines et sociales**.

# Collaborations



J. Terilla (CUNY), T.-D. Bradley (SandboxAQ), L. Pellissier (Paris-Est Créteil), Th. Seiller (CNRS), S. Jarvis (CUNY)

## Papiers de référence

- ◊ Gastaldi, J. L. (2021). Why Can Computers Understand Natural Language? *Philosophy & Technology*, 34(1), 149–214. <https://doi.org/10.1007/s13347-020-00393-9>
- ◊ Gastaldi, J. L., & Pellissier, L. (2021). The calculus of language: Explicit representation of emergent linguistic structure through type-theoretical paradigms. *Interdisciplinary Science Reviews*. <https://doi.org/10.1080/03080188.2021.1890484>
- ◊ Bradley, T.-D., Gastaldi, J. L., & Terilla, J. (2024). The structure of meaning in language: Parallel narratives in linear algebra and category theory. *Notices of the American Mathematical Society*. <https://api.semanticscholar.org/CorpusID:263613625>

# Références |

- Ali, S. M., Dick, S., Dillon, S., Jones, M. L., Penn, J., & Staley, R. (2023). Histories of artificial intelligence: A genealogy of power. *BJHS Themes*, 8, 1–18. <https://doi.org/10.1017/bjt.2023.15>
- Belrose, N., Schneider-Joseph, D., Ravfogel, S., Cotterell, R., Raff, E., & Biderman, S. (2024). Leace: Perfect linear concept erasure in closed form. *Proceedings of the 37th International Conference on Neural Information Processing Systems*.
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. <https://doi.org/10.1145/3442188.3445922>
- Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. <https://doi.org/10.18653/v1/2020.acl-main.463>
- Bradley, T.-D., Gastaldi, J. L., & Terilla, J. (2024). The structure of meaning in language: Parallel narratives in linear algebra and category theory. *Notices of the American Mathematical Society*. <https://api.semanticscholar.org/CorpusID:263613625>
- Church, A. (1936). An unsolvable problem of elementary number theory. *American Journal of Mathematics*, 58(2), 345–363.
- Gastaldi, J. L. (2021). Why Can Computers Understand Natural Language? *Philosophy & Technology*, 34(1), 149–214. <https://doi.org/10.1007/s13347-020-00393-9>
- Gastaldi, J. L., & Pellissier, L. (2021). The calculus of language: Explicit representation of emergent linguistic structure through type-theoretical paradigms. *Interdisciplinary Science Reviews*. <https://doi.org/10.1080/03080188.2021.1890484>

## Références II

- Girard, J.-Y. (2006). *Le point aveugle: Cours de logique. vers la perfection.* Editions Hermann.
- Gödel, K. (1986 (1934)). On undecidable propositions of formal mathematical systems. In *Collected works* (pp. 346–371). Clarendon Press Oxford University Press.
- Goldberg, Y., & Levy, O. (2014). Word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method. *CoRR*, abs/1402.3722.
- Kirschenbaum, M. (2023). *Again theory: A forum on language, meaning, and intent in the time of stochastic parrot.* <https://critinq.wordpress.com/2023/06/26/again-theory-a-forum-on-language-meaning-and-intent-in-the-time-of-stochastic-parrots/>
- Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, 2177–2185.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., Le, Q., & Strohmann, T. (2013). *Learning representations of text using neural networks. NIPS deep learning workshop 2013 slides.*
- Nietzsche, F. (1873). *De la vérité et du mensonge au sens extra-moral* (P. Wotling, Trans.). Flammarion.
- Turing, A. (1937). On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, s2-42(1), 230–265. <https://doi.org/10.1112/plms/s2-42.1.230>
- Underwood, T. (2023, October 15). *The empirical triumph of theory* [Accessed: 2023-10-15].  
<https://critinq.wordpress.com/2023/06/29/the-empirical-triumph-of-theory/>

Séminaire HiPhiS  
Univ. de Montpellier, Univ. Paul Valéry, IRES, CNRS  
Montpellier, France

*Épistémologie de l'apprentissage machine*  
Pour un formalisme critique

Juan Luis Gastaldi  
[www.jlgastaldi.com](http://www.jlgastaldi.com)



15 Avril, 2025