Algorithmes de classification, data mining et text mining

- M1 SID
- 2019-2020
- J. G. Moreno et Y. Pitarch

Renseignements généraux

- Enseignants:
  - J. Moreno (jose.moreno@irit.fr)
  - Y. Pitarch (pitarch@irit.fr)
- Moodle:
  - Cours EMMAB2C1
- Évaluation:
  - 30% projet (CC) – kaggle inclass competition
  - 70% contrôle terminal

Sommaire

- Fouille de textes
  - Introduction
  - Exploration et analyse des associations de mots
  - Exploration et analyse des topiques
  - Regroupement et classification
- Network Mining
  - Network generalities
  - Nodes and ties
  - Network partitioning and community detection
  - Diffusion of information
  - Link prediction

Fouille et analyse de textes

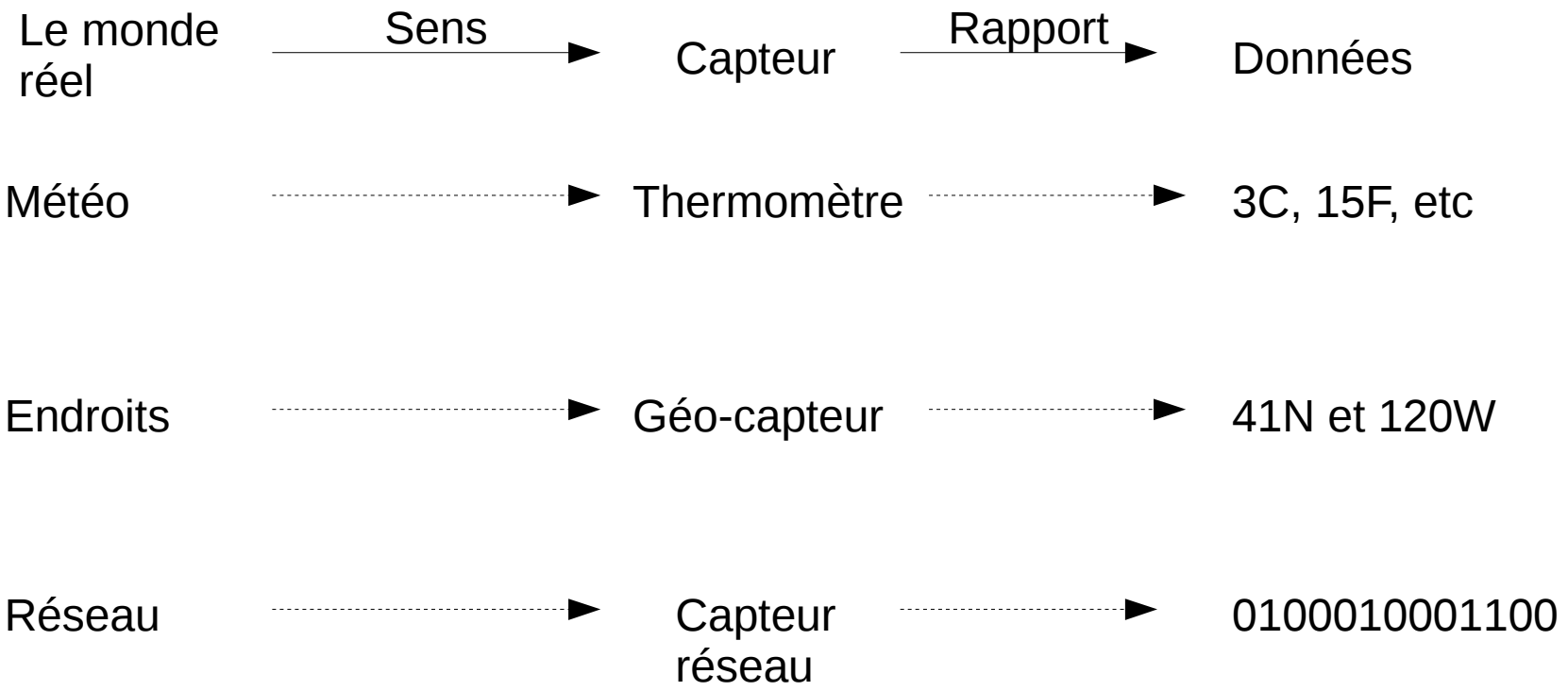
- Fouille de textes ~ Analyse de textes
- Transformer les données textuelles en informations de haute qualité ou en connaissances exploitables.
  - Minimise l'effort humain (sur la consommation de données texte)
  - Fournies de connaissances pour une prise de décision optimale
- Lié à la recherche d’information, qui est un élément essentiel dans tout système de text mining.
  - La recherche d’information peut être un préprocesseur pour le text mining
  - La recherche d’information est nécessaire pour retrouver des connaissances

Pourquoi la fouille de textes est-il difficile ?

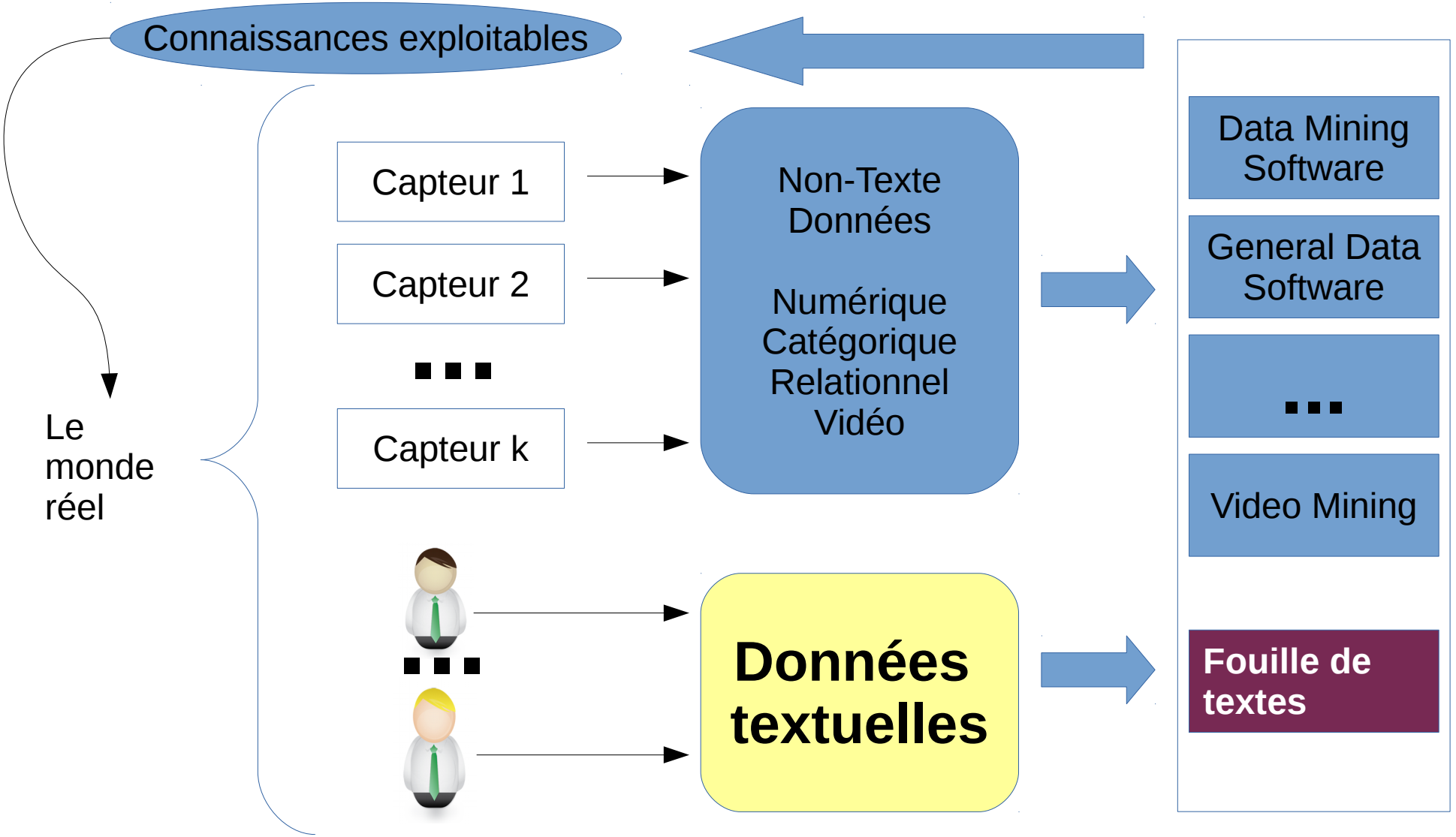
- J'ai mangé de la pizza avec des amis = J'ai mangé de la pizza avec des olives ?
  - Un seul mot différent
  - Amis = Olives ?
- J'ai mangé de la pizza avec des amis = Des amis et moi, on a partagé de la pizza ?
  - 2 mots similaires
  - Verbe différent
  - 8/9 vs 5/9 mots
  - Sens sémantique similaire

Données textuelles et non textuelles

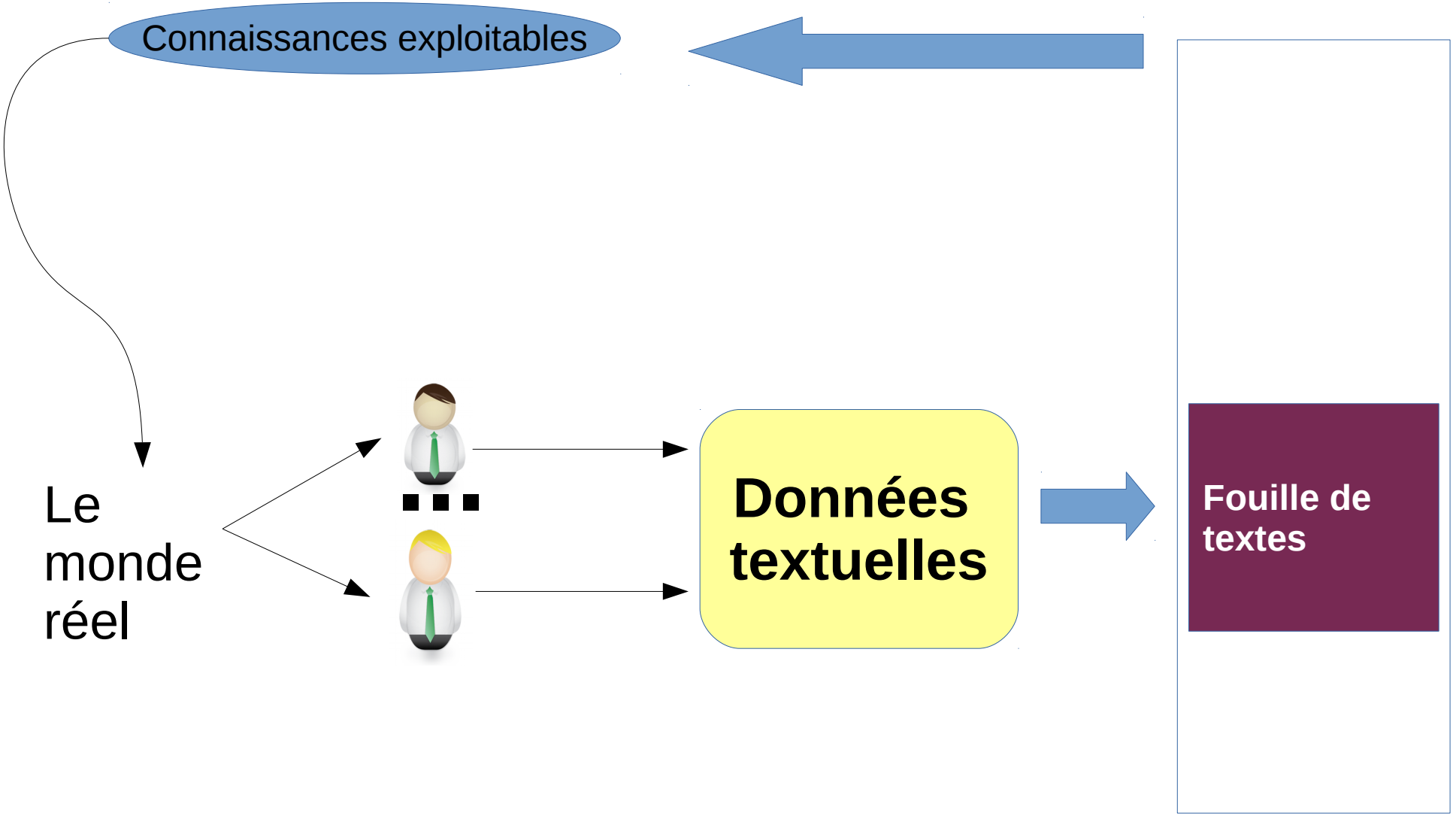
- L'humain en tant que "capteur" subjectif



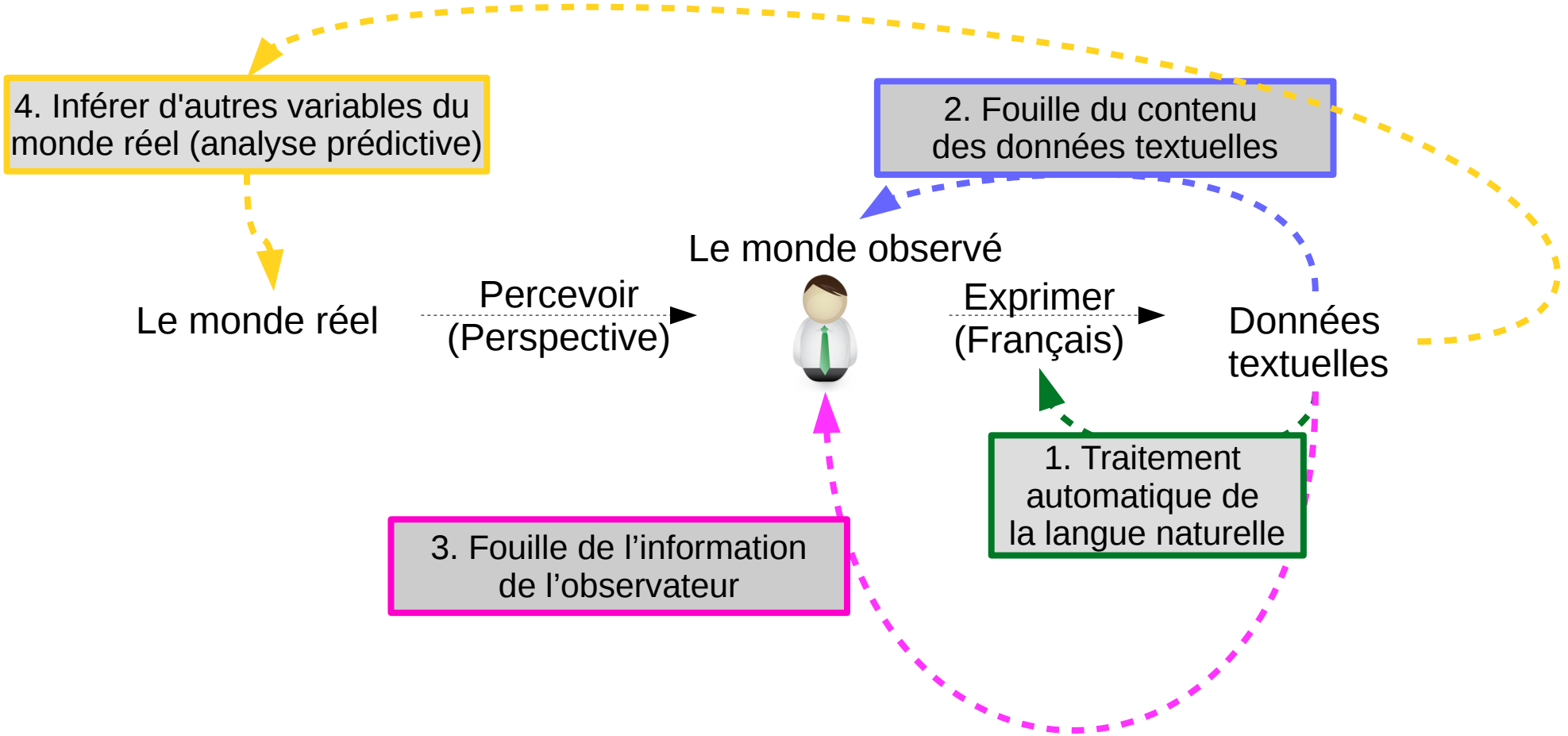
Le problème général de la fouille de données



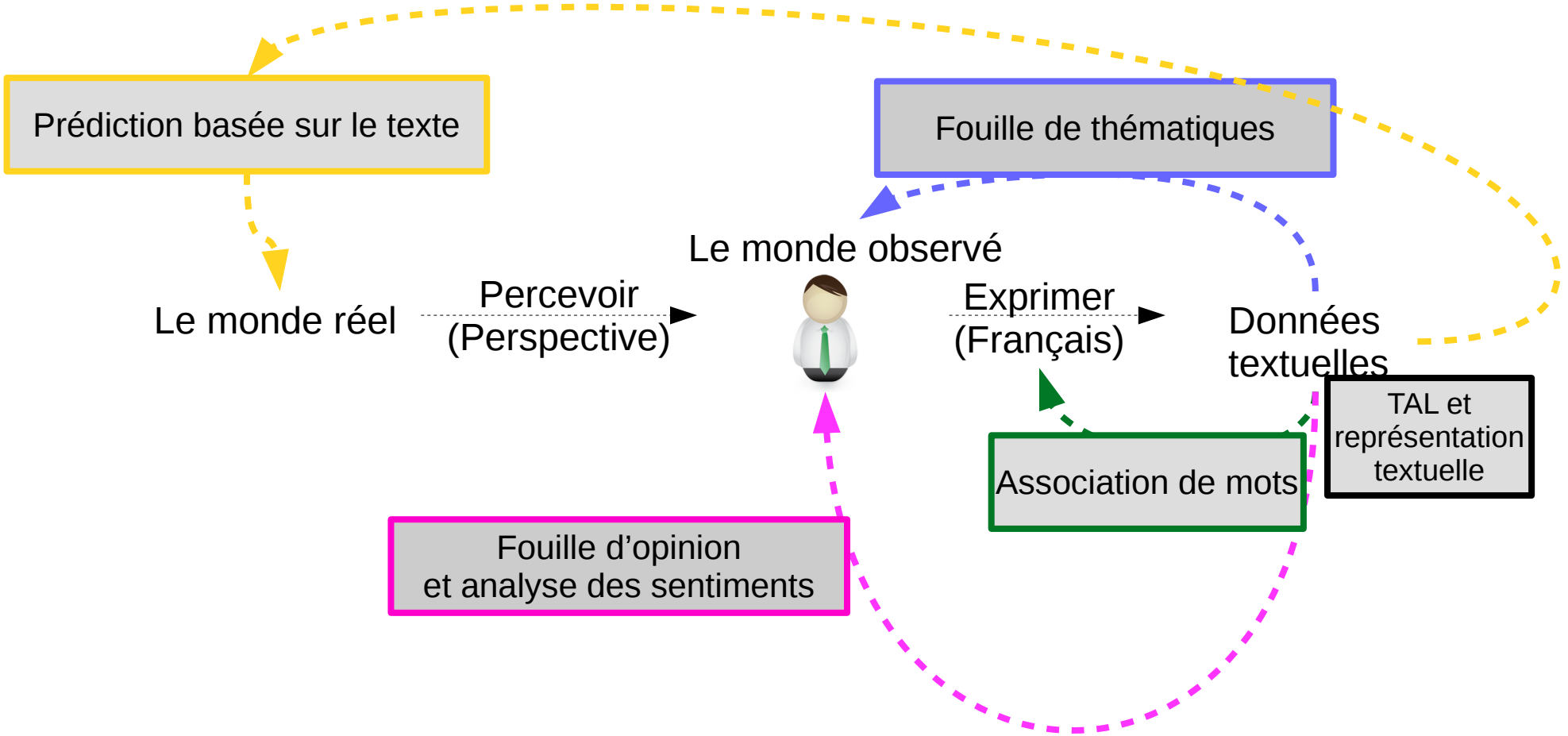
Le problème général de la fouille de textes



Paysage de la fouille et de l'analyse de textes



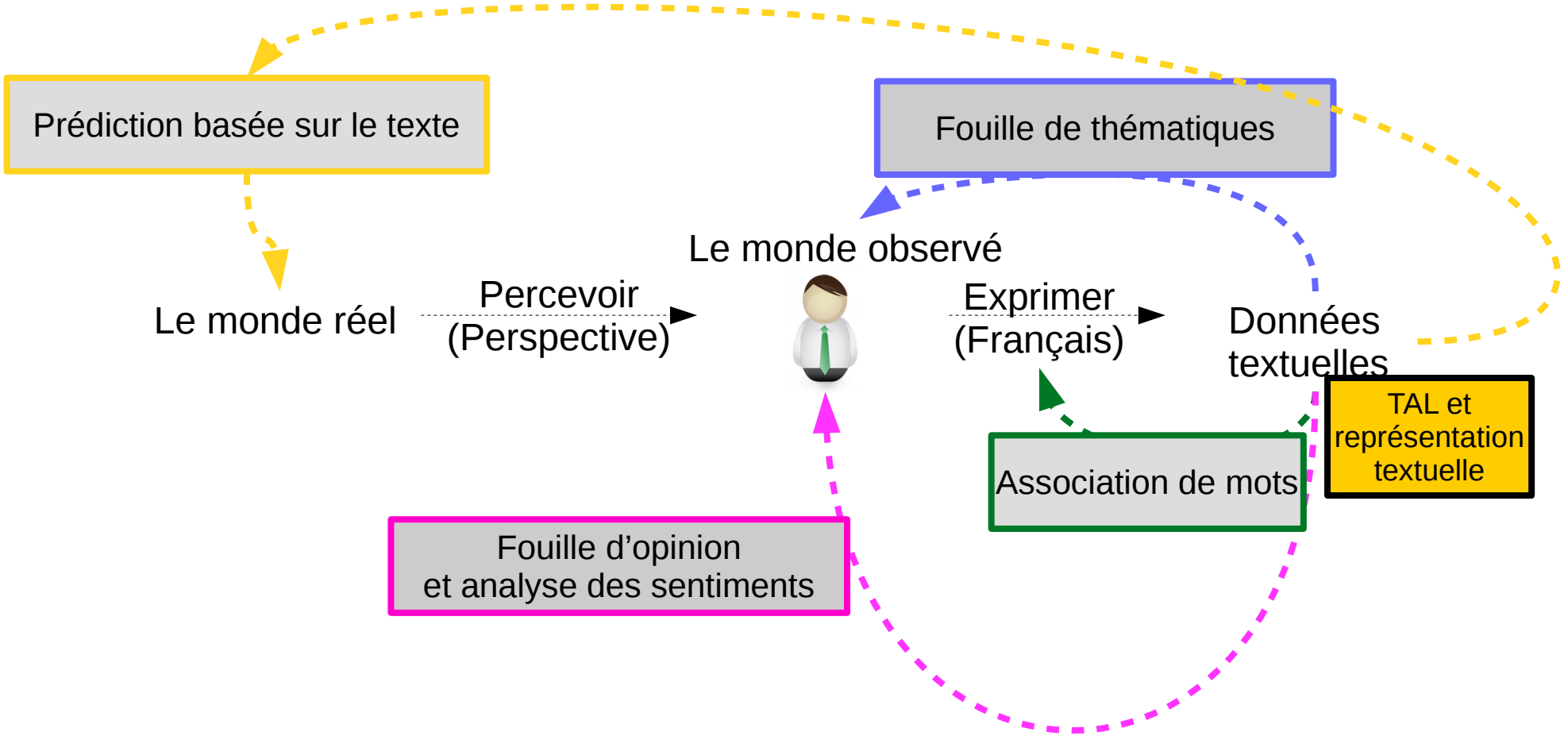
Paysage de la fouille et de l'analyse de textes



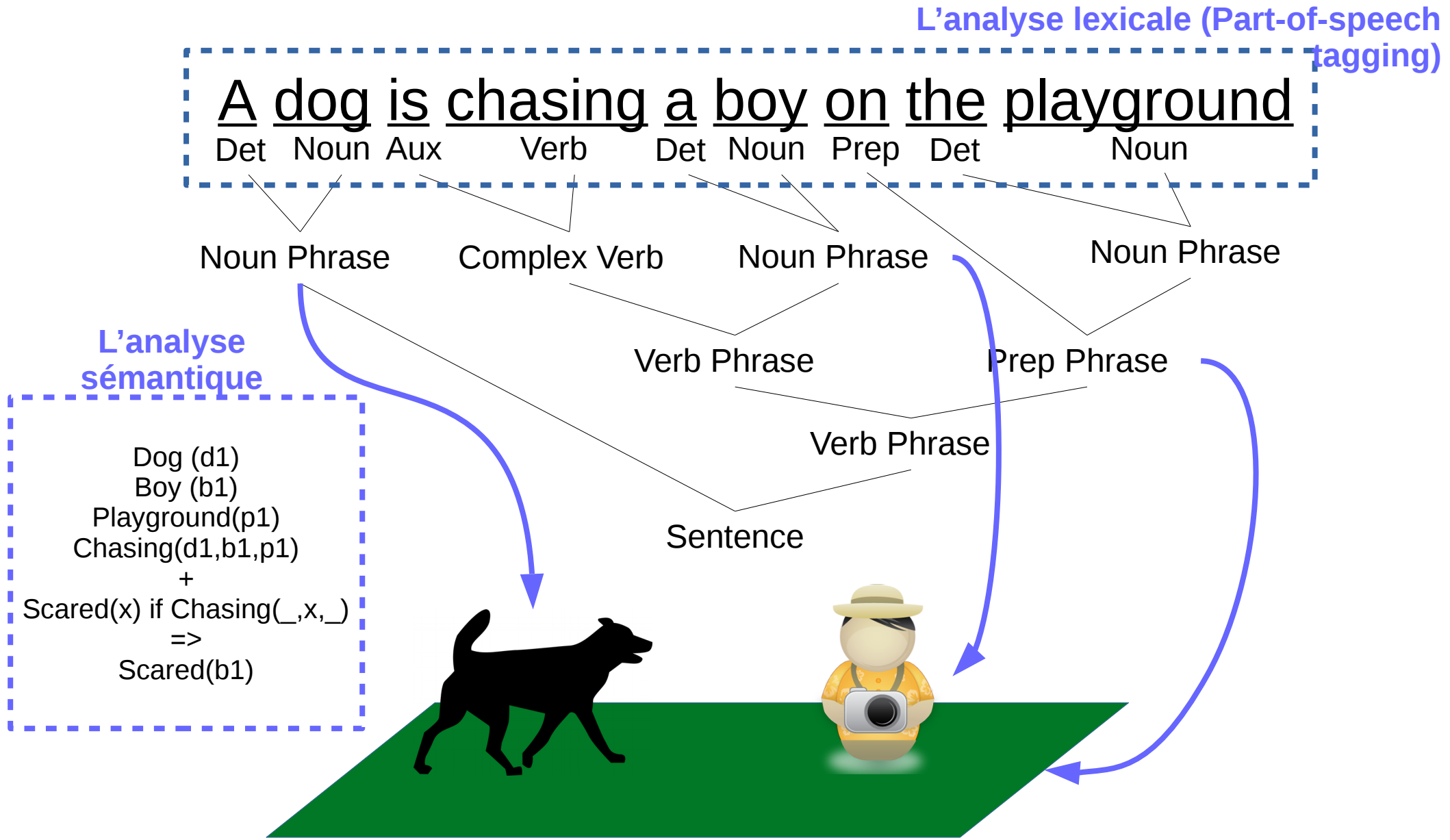
Notions de la fouille de textes

- Collections
  - Les documents sont compressés
  - Formats peu courants
  - Parfois, ils n'existent tout simplement pas
- Les documents
  - Beaucoup de prétraitement (encodage, nettoyage, découpage, etc.)
  - Niveau de granularité (document entier, paragraphe, phrase, etc.)
- Les mots
  - Racinisation, majuscules/minuscules, mots fréquents (mots vides) et peu fréquents, caractères spéciaux, dates, prix, noms, e-mails, etc.
- Général
  - Langue : les tâches principales ont été abordées pour l'anglais (sans performance à 100%), mais beaucoup d'autres langues n’ont pas autant de ressources comme l'anglais et beaucoup d'entre elles sont complètement sans ressources (langues régionales).
  - Beaucoup de formules et de concepts
  - La tâche est le facteur principal
  - Outils : python (NLKT, spacy, gensim), java (OpenNLP, CoreNLP, GATE), etc.

Paysage de la fouille et de l'analyse de textes



Concepts de base en TAL



Le TAL est difficile !

- Le langage naturel est conçu pour rendre la communication humaine efficace. En conséquence,
  - Nous omettons beaucoup de connaissances de bon sens, que nous supposons que l'auditeur/lecteur possède.
  - Nous gardons beaucoup d'ambiguïtés, que nous supposons que l'auditeur/lecteur sait comment résoudre.
- Cela rend CHAQUE étape du TAL difficile
  - L'ambiguïté est une tueuse !
  - Il faut d'abord faire preuve de bon sens.

Exemples de défis

- Ambiguïté au niveau des mots :
  - “copie” peut être un nom ou un verbe (POS ambigu)
  - “orange” a des significations multiples (sens ambigu)
- Ambiguïté syntaxique :
  - “Traitement automatique de la langue” (modification)
  - “Un homme a vu un garçon avec un télescope” (PP annexe)
  - Résolution d’anaphora “Mon voisin a adopté un gros chien. ***Cet animal*** n'est pas très sympathique” (Cet animal ?)
  - Présupposition : "Il a arreté la cigarette" implique qu'il a déjà fumé avant..

Ce que nous ne pouvons pas faire

- Etiquetage POS à 100 %
  - "Paul copie **(verbe)** sur sa voisine.” vs “Il ne devra pas s'étonner s'il trouve un zéro sur sa copie **(nom)**."
- Analyse générale
  - "Un homme a vu un garçon avec un télescope"
- Analyse sémantique profonde (avec interaction humaine) et précise
  - Serons-nous un jour capables de définir précisément le sens de "posséder" dans "Jean possède un restaurant" ?

Le TAL robust et général a tendance à être peu profonde, alors que la compréhension profonde ne passe pas à l'échelle.

Attention !!! À ne pas confondre

Le TAL robust et général a tendance à être peu profonde,
alors que la compréhension profonde ne passe pas à l'échelle.

“Deep learning” recent

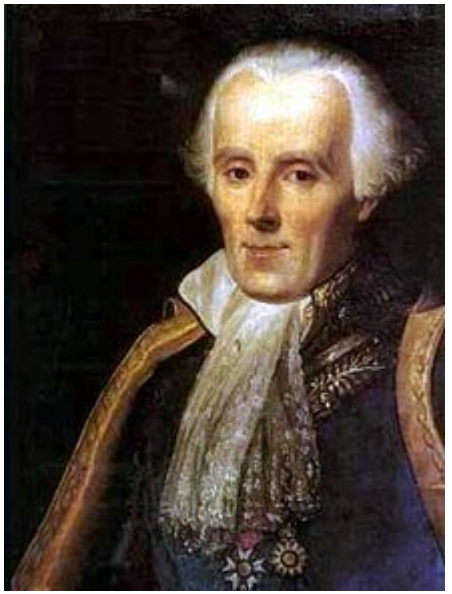
Message à emporter

- Le TAL est la base de la fouille de textes
- Les ordinateurs sont loin de comprendre le langage naturel
  - Le TAL “manuel" exige des connaissances et des inférences de bon sens, et ne fonctionne donc que dans des domaines très limités.
  - Le TAL peu “manuel" basée sur des méthodes statistiques peut être réalisée à grande échelle et est donc plus largement applicable.
- Dans la pratique : le TAL statistique est la base, tandis que l'humain apporte son aide en cas de besoin.

Citation

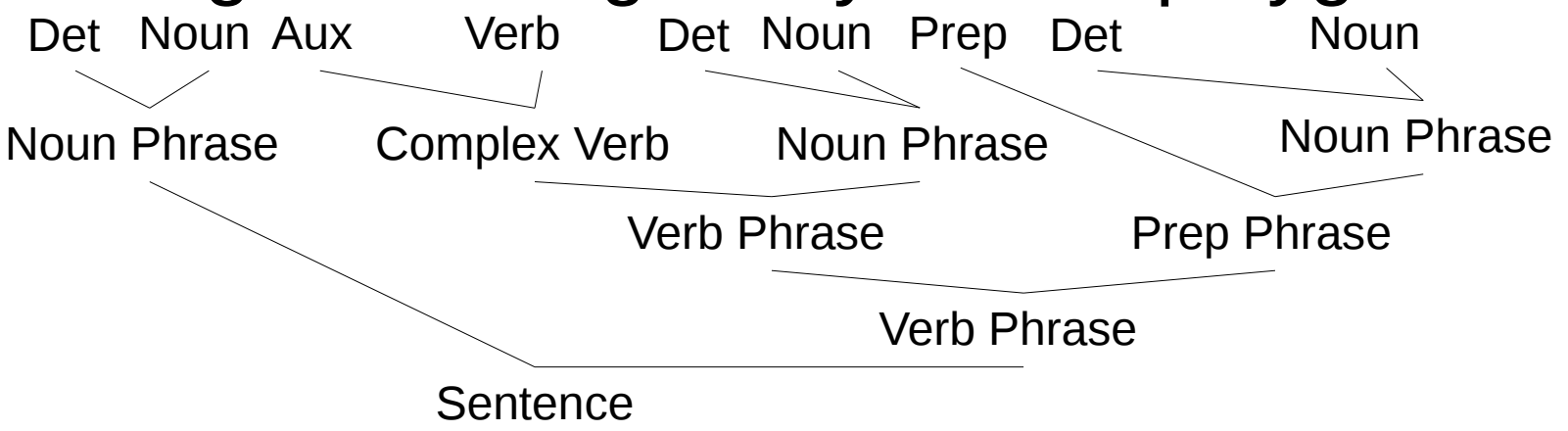
- « One sees, from this Essay, that the theory of probabilities is basically just common sense reduced to calculus »
- « On voit, par cet Essai, que la théorie des probabilités n'est, au fond, que le bon sens réduit au calcul; »

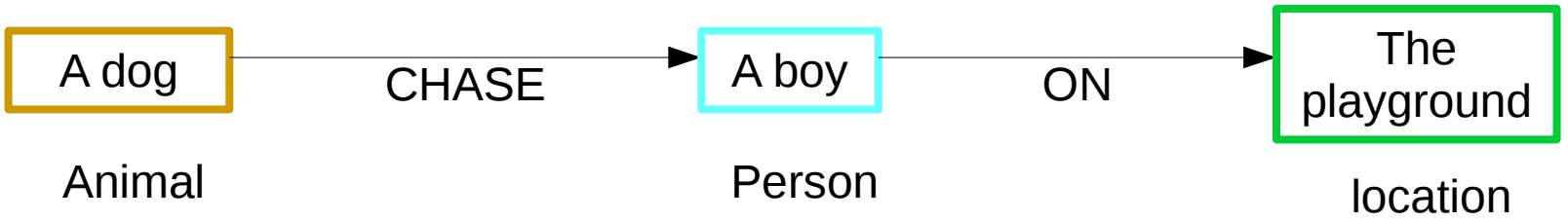
Laplace, 1814



A dog is chasing a boy on the playground

A dog is chasing a boy on the playground





Dog (d1), Boy (b1),Playground(p1), Chasing(d1,b1,p1)

Chaîne de caractères

Séquence de mots

POS tags

Structures syntaxiques

Entités et relations

Prédicats logiques

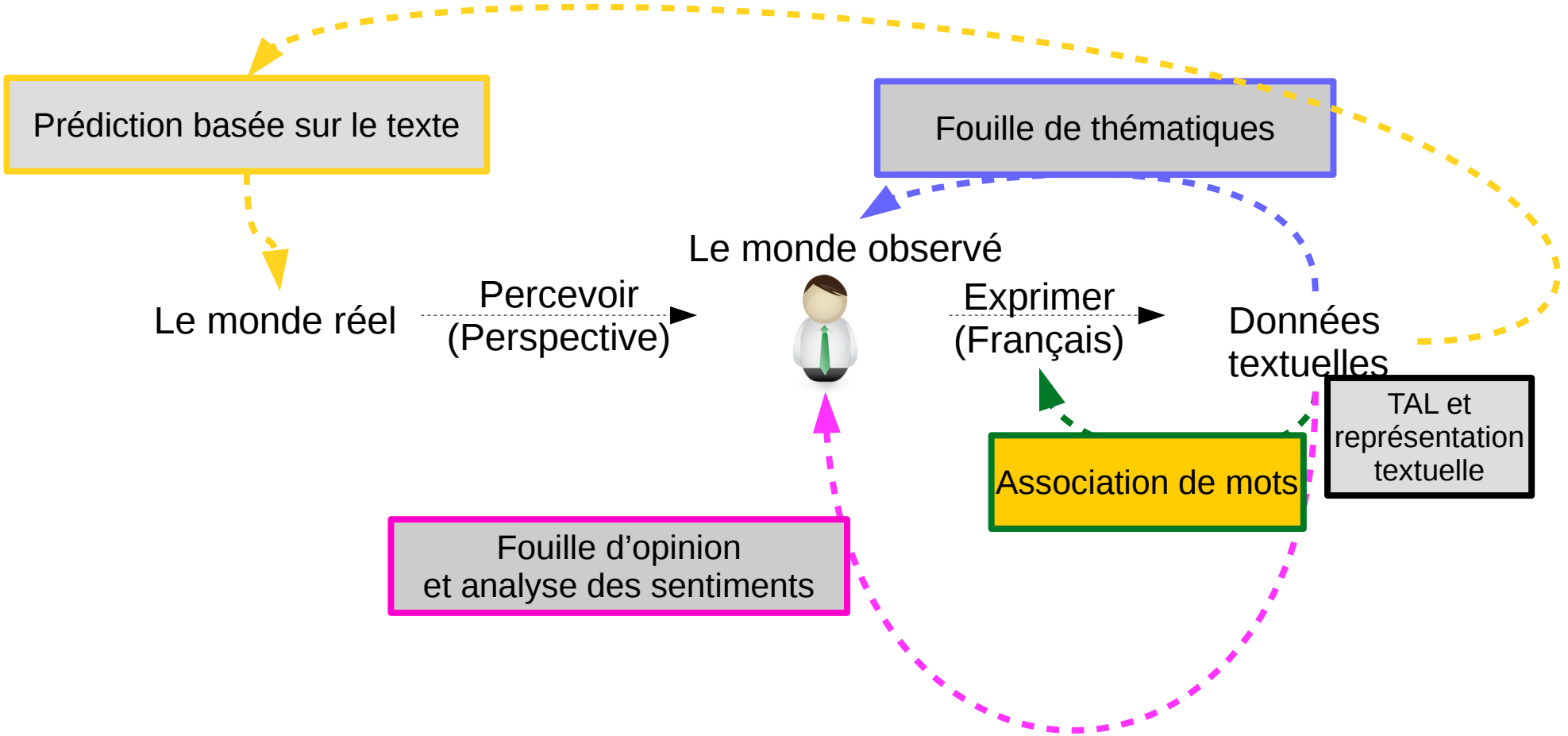
Représentation textuelle et analyse possible

Représentation	Généralité	Analyse possible	Exemples d'application
Chaîne de caractères	******	Traitement des chaînes	Compression
Morceaux de mots	*****	Composition de mots	Les mêmes que pour les mots
Mots	****	Analyse des relations entre les mots ; analyse des topiques ; analyse des sentiments	Découverte de thésaurus ; applications liées aux topiques et aux opinions
Structures syntaxiques	***	Analyse syntaxique des graphes	Analyse stylistique ; extraction de caractéristiques basée sur la structure
Entités et relations	**	Analyse des graphes de connaissances ; analyse des réseaux d'information	Découverte de connaissances et d'opinions sur des entités spécifiques ; population de bases de connaissances
Prédicats logiques	*	inférence logique	Assistant de connaissances pour un médecin

Message à emporter

- Les représentations textuelles déterminent le type d'algorithmes qui peuvent être appliqués.
  - Plusieurs façons de représenter le texte sont possibles
  - Chaîne, mots, structures syntaxiques, graphes d'entités-relations, prédicats, etc.
- Peut/devrait être combiné dans des applications réelles
  - Ce cours se concentre principalement sur la représentation basée sur les mots.
  - Généralités et robustesse : applicable à tout langage naturel
  - Pas/peu d'effort manuel
  - "Étonnamment" puissant pour de nombreuses applications (pas toutes !)
- Peut être combiné avec des représentations plus sophistiquées
  - Par exemple avec les morceaux de mots
- Outils
  - Python (NLTK, spacy, gensim, etc.)

Paysage de la fouille et de l'analyse de textes



Les relations fondamentales entre les mots

- Paradigmatiques : A et B ont une relation paradigmatique s'ils peuvent être substitués l'un à l'autre (c-à-d. A et B sont dans la même classe).
  - Par exemple, "chat" et "chien" ; "lundi" et "mardi".
- Syntagmatique : A et B ont une relation syntagmatique s'ils peuvent être combinés entre eux (c-à-d. A et B sont reliés sémantiquement).
  - Par exemple, "chat" et "dort" ; "voiture" et "conduite".
- Ces deux relations fondamentales et complémentaires peuvent être généralisées pour décrire les relations de n'importe quel élément d'une langue

Pourquoi fouiller les associations de mots ?

- Ils sont utiles pour améliorer la précision de nombreuses tâches du TAL
  - Etiquetage POS, analyse, reconnaissance d'entités, extension d'acronymes, etc.
  - Apprentissage de la grammaire
- Ils sont directement utiles pour de nombreuses applications dans le domaine de la recherche d'information et la foille de textes.
  - Recherche d'information (par exemple, utiliser des associations de mots pour suggérer une requête)
  - Construction automatique d'une carte thématique pour la navigation : les mots en tant que nœuds et les associations en tant que liens
  - Comparer et résumer les opinions (par exemple, quels mots sont les plus fortement associés à "batterie" dans les critiques positives et négatives sur iPhone7, respectivement ?)

Fouille d'associations de mots : Intuitions

My cat eats fish on Satuyday

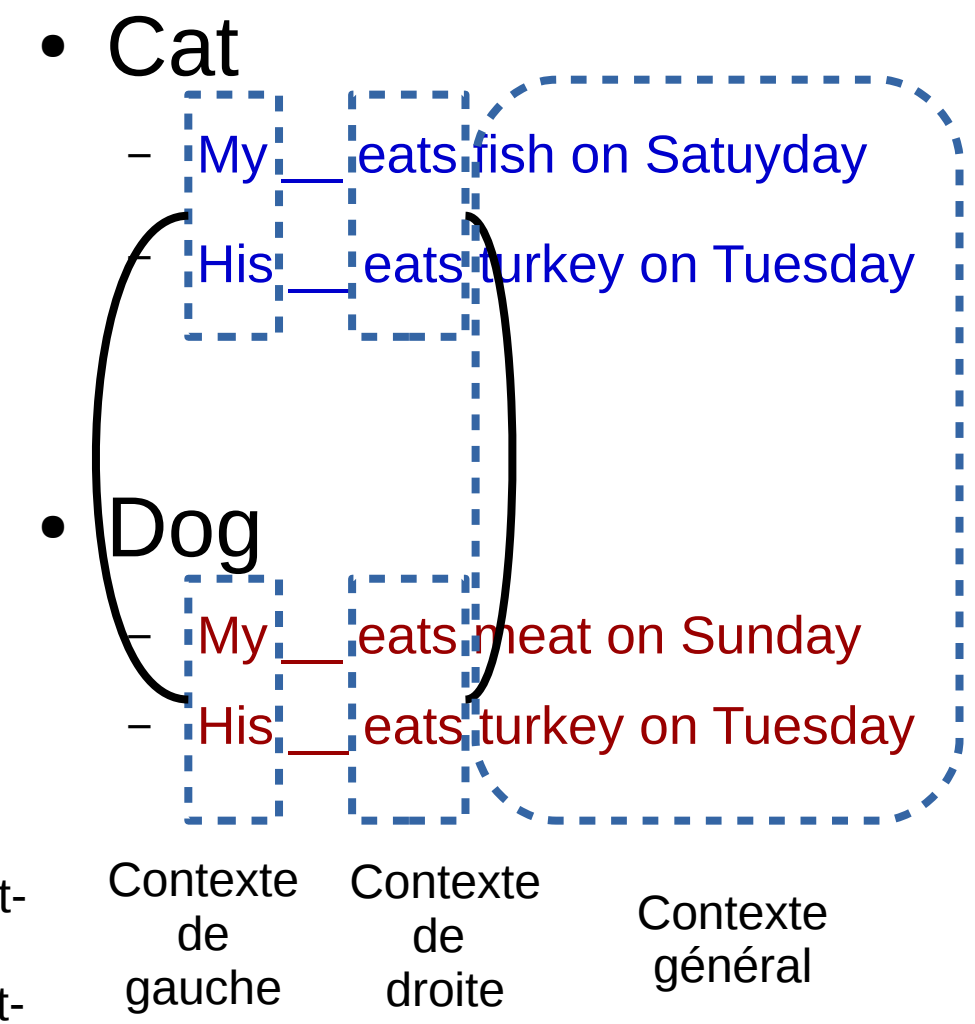
His cat eats turkey on Tuesday

My dog eats meat on Sunday

His dog eats turkey on Tuesday

Paradigmatique :
Dans quelle mesure le contexte de "cat" et le contexte de "dog" sont-ils similaires ?
Dans quelle mesure le contexte de "cat" et le contexte de "computer" sont-ils similaires ?

Syntagmatique :
Dans quelle mesure la présence de "eats" est-elle utile pour prédire la présence de "meat"?
Dans quelle mesure la présence de "eats" est-elle utile pour prédire la présence de "text"?



Fouille d’associations de mots : Idées Générales

- Paradigmatique
  - Représenter chaque mot par son contexte
  - Calculer la similarité du contexte
  - Les mots ayant une grande similarité contextuelle ont probablement une relation paradigmatique.
- Syntagmatique
  - Comptez combien de fois deux mots apparaissent ensemble dans un contexte (p. ex., phrase ou paragraphe).
  - Comparer leurs cooccurrences avec leurs occurrences individuelles.
  - Les mots ayant des cooccurrences élevées mais des occurrences individuelles relativement faibles ont probablement une relation syntagmatique.

Sémantique distributionnelle

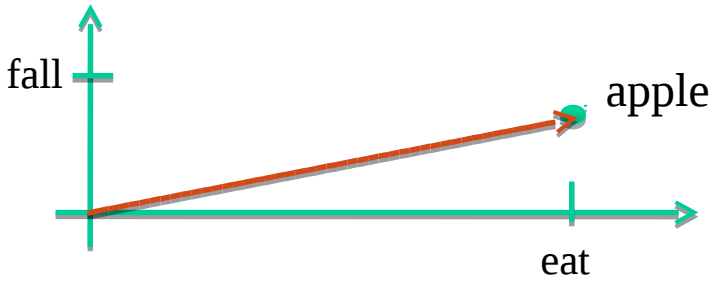
- Comparer deux mots :
  - Regarder tous les mots du contexte pour mot1
  - Regarder tous les mots du contexte pour mot2
  - Dans quelle mesure ces deux collections contextuelles sont-elles semblables dans leur intégralité ?
- Comparer les représentations distributionnelles de deux mots

Comment comparer deux collections de contextes dans leur intégralité ?

Compter combien de fois "apple" se produit à proximité d'autres mots dans une grande collection de textes (corpus) :

eat	fall	ripe	slice	peel	tree	throw	fruit	pie	bite	crab
794	244	47	221	208	160	145	156	109	104	88

Interpréter le compte comme coordonnées :



Chaque mot du contexte devient une dimension.

Comment comparer deux collections de contextes dans leur intégralité ?

Compter combien de fois "apple" se produit à proximité d'autres mots dans une grande collection de textes (corpus) :

eat	fall	ripe	slice	peel	tree	throw	fruit	pie	bite	crab
794	244	47	221	208	160	145	156	109	104	88

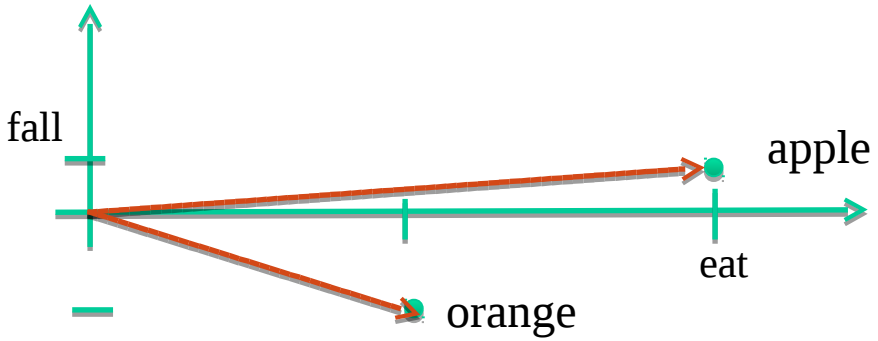
Appliquer la même procedure pour “orange”:

eat	fall	ripe	slice	peel	tree	throw	fruit	pie	bite	crab
265	22	25	62	220	64	74	111	4	4	8

Comment comparer deux collections de contextes dans leur intégralité ?

Puis visualiser les deux comptages comme vecteurs dans le même espace :

eat	fall	ripe	slice	peel	tree	throw	fruit	pie	bite	crab
794	244	47	221	208	160	145	156	109	104	88
eat	**fall**	**ripe**	**slice**	**peel**	**tree**	**throw**	**fruit**	**pie**	**bite**	**crab**
265	22	25	62	220	64	74	111	4	4	8



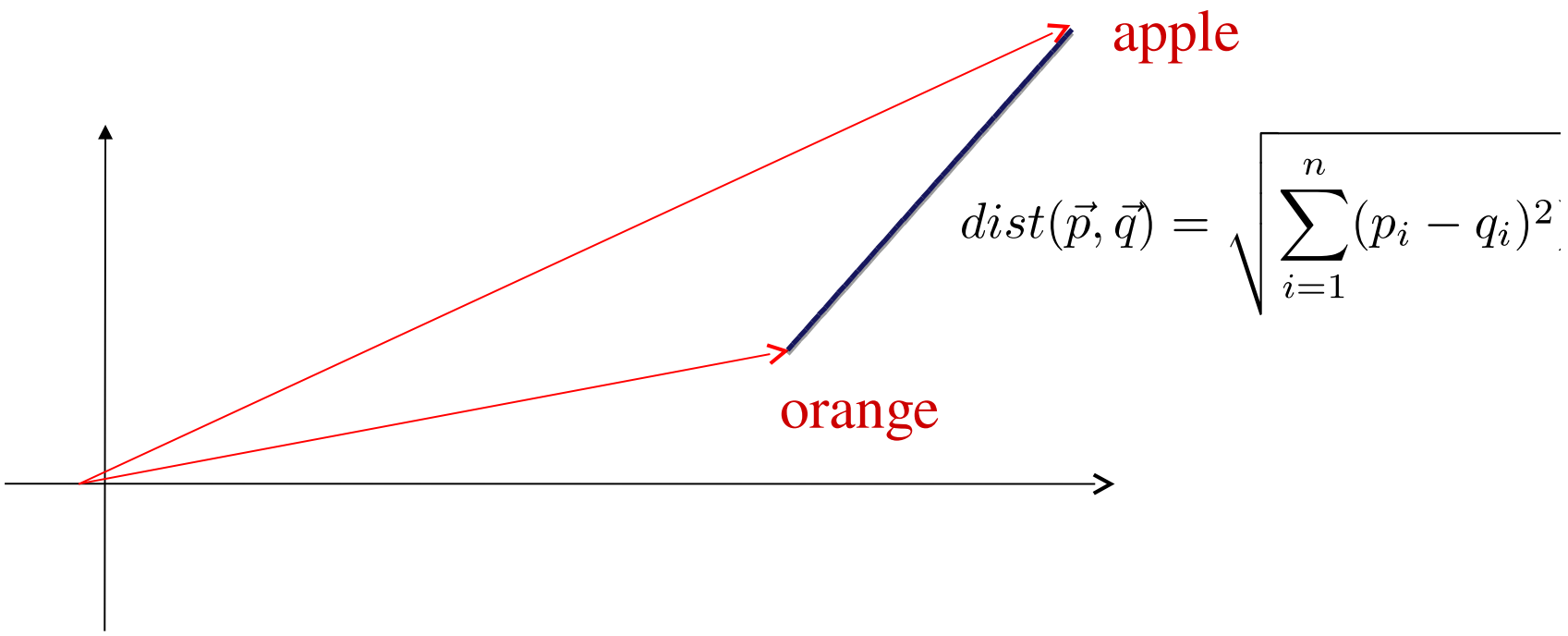
La ressemblance entre deux mots est la proximité dans l'espace

Où peut-on trouver des textes à utiliser pour faire un modèle distributionnel ?

- Texte sous forme numérique !
- Articles de journaux
- Projet Gutenberg :
  - Anciens livres disponibles gratuitement (pas tous tous)
- **Wikipédia**
- Collections de textes préparées pour l'analyse linguistique :
  - Corpus « balanced »
  - WaC : Pages Web d'un domaine particulier
  - ELRA, LDC, etc. (collections de corpus)
  - Google n-grams, Google books

Qu'entend-on par "similarité" des vecteurs ?

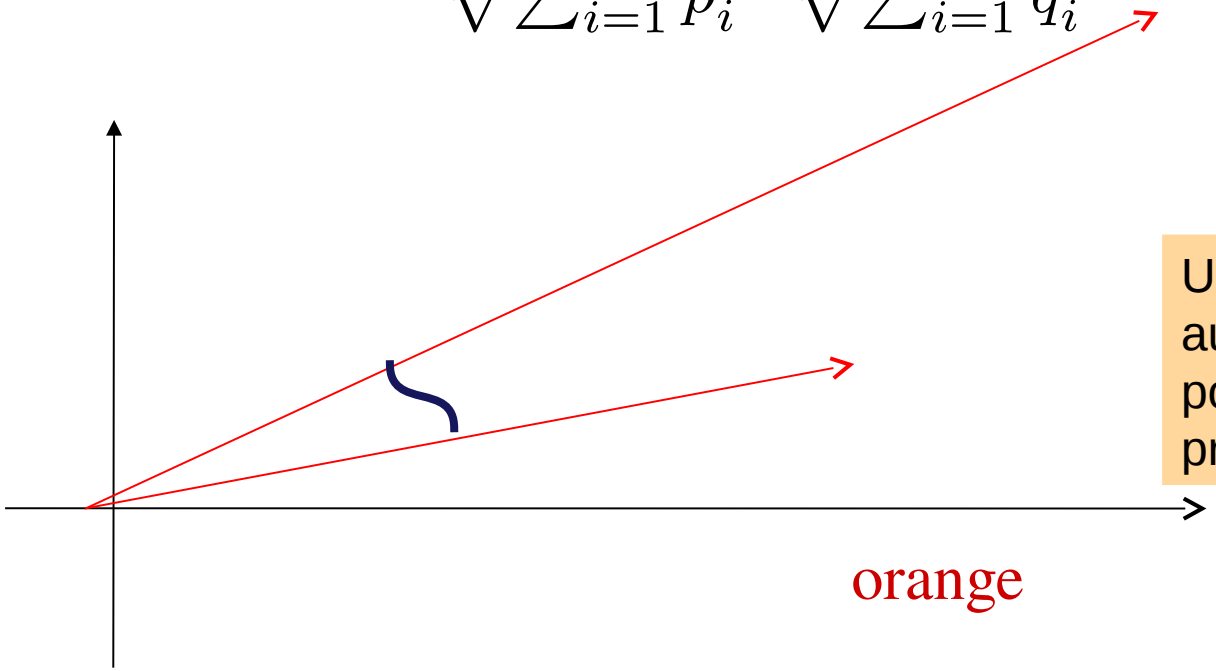
La distance euclidienne comme mesure de dissimilarité



Le problème de la distance euclidienne : très sensible à la fréquence des mots !

Qu'entend-on par "similarité" des vecteurs ?

$$cos(\vec{p}, \vec{q}) = \frac{\sum_{i=1}^{n} p_i \cdot q_i}{\sqrt{\sum_{i=1}^{n} p_i^2} \cdot \sqrt{\sum_{i=1}^{n} q_i^2}}$$



Utilisez l'angle entre les vecteurs au lieu de la distance entre les points pour contourner les problèmes de fréquence des mots.

Certains comptes par "lettre" dans "Pride and Prejudice". Qu'est-ce que vous remarquez ?

the	to	of	and	a	her	she	his	is	was	in	that
102	75	72	56	52	50	41	36	35	34	34	33

had	i	from	you	as	this	mr	for	not	on	be	he
32	28	28	25	23	23	22	21	21	20	18	17

but	elizabeth	with	him	which	by	when	jane
17	17	16	16	16	15	14	12

Certains comptes par "lettre" dans "Pride and Prejudice". Qu'est-ce que vous remarquez ?

the	to	of	and	a	her	she	his	is	was	in	that
102	75	72	56	52	50	41	36	35	34	34	33

had	i	from	you	as	this	mr	for	not	on	be	he
32	28	28	25	23	23	22	21	21	20	18	17

but	elizabeth	with	him	which	by	when	jane
17	17	16	16	16	15	14	12

Tous les mots cooccurrents les plus fréquament sont des mots de fonction (ou mots vides).

Certains mots sont plus informatifs que d'autres

- Les mots de fonction apparaissent fréquemment en même temps que tous les mots.
  - Ce qui les rend moins informatifs
- Ils ont un nombre de cooccurrences beaucoup plus élevé que les mots de contenu.
  - Ils peuvent "étouffer" des contextes plus informatifs.
- Fréquence
  - Sélectionner les bigrams les plus fréquents

$$max_{x,y}\, p(x,y) \qquad max_{x,y} \log\left(p(x,y)+1\right)$$

  - Un simple filtre POS améliore considérablement les résultats. Filtrer les étiquetes POS de couples de mots tels que "Aux Noun" traite des résultats tels que "Premier ministre".

Quelques mesures statistiques

- Information mutuelle par points (théorie de l'information)
  - Le PMI nous indique la quantité d'informations qui est fournie par l'occurrence d'un mot sur l'occurrence de l'autre mot.

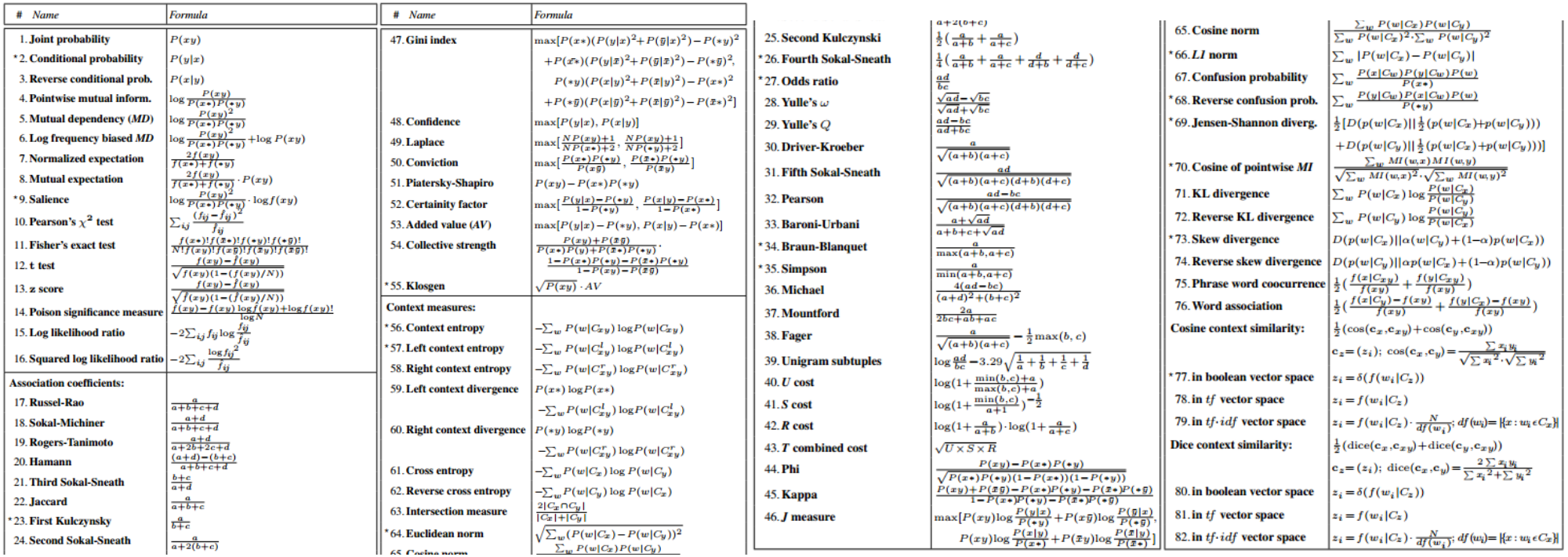
$$PMI(x, y) = \log_2 \frac{p(x, y)}{p(x)\,p(y)}$$

- Test Phi-Square de Pearson (hypothèse de test)
  - L'essence du test est de comparer les fréquences observées avec les fréquences attendues pour l'indépendance dans un tableau de contingence.
  - Essayez de réfuter l'hypothèse nulle. Plus le score est élevé, plus l'hypothèse Ho peut être rejetée avec confiance.

$$Ho : p(x, y) = p(x)\,p(y)$$

Autres mesures

- Il existe de nombreuses mesures d'association



#	Name	Formula
1.	Joint probability	$P(xy)$
*2.	Conditional probability	$P(y\|x)$
3.	Reverse conditional prob.	$P(x\|y)$
4.	Pointwise mutual inform.	$\log \frac{P(xy)}{P(x*)P(*y)}$
5.	Mutual dependency (*MD*)	$\log \frac{P(xy)^2}{P(x*)P(*y)}$
6.	Log frequency biased *MD*	$\log \frac{P(xy)^2}{P(x*)P(*y)} + \log P(xy)$
7.	Normalized expectation	$\frac{2f(xy)}{f(x*)+f(*y)}$
8.	Mutual expectation	$\frac{2f(xy)}{f(x*)+f(*y)} \cdot P(xy)$
*9.	Salience	$\log \frac{P(xy)^2}{P(x*)P(*y)} \cdot \log f(xy)$
10.	Pearson's χ^2 test	$\sum_{ij} \frac{(f_{ij}-\hat{f}_{ij})^2}{\hat{f}_{ij}}$
11.	Fisher's exact test	$\frac{f(x*)!f(\bar{x}*)!f(*y)!f(*\bar{y})!}{N!f(xy)!f(x\bar{y})!f(\bar{x}y)!f(\bar{x}\bar{y})!}$
12.	t test	$\frac{f(xy)-\hat{f}(xy)}{\sqrt{f(xy)(1-(f(xy)/N))}}$
13.	z score	$\frac{f(xy)-\hat{f}(xy)}{\sqrt{\hat{f}(xy)(1-(\hat{f}(xy)/N))}}$
14.	Poison significance measure	$\frac{f(xy)-\hat{f}(xy)\log f(xy)+\log f(xy)!}{\log N}$
15.	Log likelihood ratio	$-2\sum_{ij} f_{ij} \log \frac{f_{ij}}{\hat{f}_{ij}}$
16.	Squared log likelihood ratio	$-2\sum_{ij} \frac{\log f_{ij}^2}{\hat{f}_{ij}}$
	Association coefficients:	
17.	Russel-Rao	$\frac{a}{a+b+c+d}$
18.	Sokal-Michiner	$\frac{a+d}{a+b+c+d}$
19.	Rogers-Tanimoto	$\frac{a+d}{a+2b+2c+d}$
20.	Hamann	$\frac{(a+d)-(b+c)}{a+b+c+d}$
21.	Third Sokal-Sneath	$\frac{b+c}{a+d}$
22.	Jaccard	$\frac{a}{a+b+c}$
*23.	First Kulczynsky	$\frac{a}{b+c}$
24.	Second Sokal-Sneath	$\frac{a}{a+2(b+c)}$

#	Name	Formula
47.	Gini index	$\max[P(x*)(P(y\|x)^2+P(\bar{y}\|x)^2)-P(*y)^2$ $+P(\bar{x}*)(P(y\|\bar{x})^2+P(\bar{y}\|\bar{x})^2)-P(*\bar{y})^2,$ $P(*y)(P(x\|y)^2+P(\bar{x}\|y)^2)-P(x*)^2$ $+P(*\bar{y})(P(x\|\bar{y})^2+P(\bar{x}\|\bar{y})^2)-P(\bar{x}*)^2]$
48.	Confidence	$\max[P(y\|x), P(x\|y)]$
49.	Laplace	$\max[\frac{NP(xy)+1}{NP(x*)+2}, \frac{NP(xy)+1}{NP(*y)+2}]$
50.	Conviction	$\max[\frac{P(x*)P(*y)}{P(x\bar{y})}, \frac{P(\bar{x}*)P(*y)}{P(\bar{x}y)}]$
51.	Piatersky-Shapiro	$P(xy)-P(x*)P(*y)$
52.	Certainity factor	$\max[\frac{P(y\|x)-P(*y)}{1-P(*y)}, \frac{P(x\|y)-P(x*)}{1-P(x*)}]$
53.	Added value (*AV*)	$\max[P(y\|x)-P(*y), P(x\|y)-P(x*)]$
54.	Collective strength	$\frac{P(xy)+P(\bar{x}\bar{y})}{P(x*)P(y)+P(\bar{x}*)P(*y)} \cdot \frac{1-P(x*)P(*y)-P(\bar{x}*)P(*y)}{1-P(x*)-P(\bar{x}\bar{y})}$
*55.	Klosgen	$\sqrt{P(xy)} \cdot AV$
	Context measures:	
*56.	Context entropy	$-\sum_w P(w\|C_{xy}) \log P(w\|C_{xy})$
*57.	Left context entropy	$-\sum_w P(w\|C^l_{xy}) \log P(w\|C^l_{xy})$
58.	Right context entropy	$-\sum_w P(w\|C^r_{xy}) \log P(w\|C^r_{xy})$
59.	Left context divergence	$P(x*) \log P(x*)$ $-\sum_w P(w\|C^l_{xy}) \log P(w\|C^l_{xy})$
60.	Right context divergence	$P(*y) \log P(*y)$ $-\sum_w P(w\|C^r_{xy}) \log P(w\|C^r_{xy})$
61.	Cross entropy	$-\sum_w P(w\|C_x) \log P(w\|C_y)$
62.	Reverse cross entropy	$-\sum_w P(w\|C_y) \log P(w\|C_x)$
63.	Intersection measure	$\frac{2\|C_x \cap C_y\|}{\|C_x\|+\|C_y\|}$
*64.	Euclidean norm	$\sqrt{\sum_w (P(w\|C_x)-P(w\|C_y))^2}$
65.	Cosine norm	$\frac{\sum_w P(w\|C_x)P(w\|C_y)}{\sum_w P(w\|C_x)^2 \cdot \sum_w P(w\|C_y)^2}$

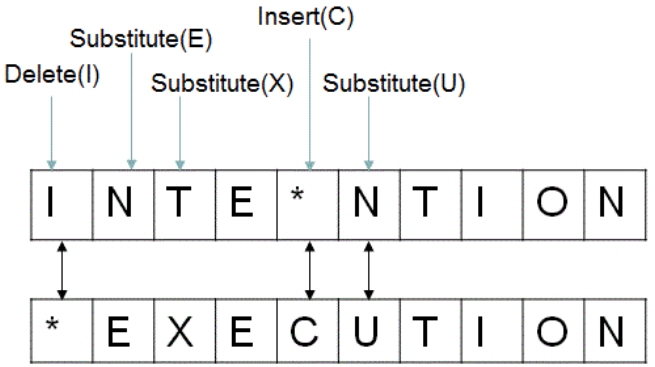
		$a+2(b+c)$
25.	Second Kulczynski	$\frac{1}{2}(\frac{a}{a+b}+\frac{a}{a+c})$
*26.	Fourth Sokal-Sneath	$\frac{1}{4}(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{d+b}+\frac{d}{d+c})$
*27.	Odds ratio	$\frac{ad}{bc}$
28.	Yulle's ω	$\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$
29.	Yulle's Q	$\frac{ad-bc}{ad+bc}$
30.	Driver-Kroeber	$\frac{a}{\sqrt{(a+b)(a+c)}}$
31.	Fifth Sokal-Sneath	$\frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$
32.	Pearson	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$
33.	Baroni-Urbani	$\frac{a+\sqrt{ad}}{a+b+c+\sqrt{ad}}$
*34.	Braun-Blanquet	$\frac{a}{\max(a+b,a+c)}$
*35.	Simpson	$\frac{a}{\min(a+b,a+c)}$
36.	Michael	$\frac{4(ad-bc)}{(a+d)^2+(b+c)^2}$
37.	Mountford	$\frac{2a}{2bc+ab+ac}$
38.	Fager	$\frac{a}{\sqrt{(a+b)(a+c)}} - \frac{1}{2}\max(b,c)$
39.	Unigram subtuples	$\log \frac{ad}{bc} - 3.29\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$
40.	*U* cost	$\log(1+\frac{\min(b,c)+a}{\max(b,c)+a})$
41.	*S* cost	$\log(1+\frac{\min(b,c)}{a+1})^{-\frac{1}{2}}$
42.	*R* cost	$\log(1+\frac{a}{a+b}) \cdot \log(1+\frac{a}{a+c})$
43.	*T* combined cost	$\sqrt{U \times S \times R}$
44.	Phi	$\frac{P(xy)-P(x*)P(*y)}{\sqrt{P(x*)P(*y)(1-P(x*))(1-P(*y))}}$
45.	Kappa	$\frac{P(xy)+P(\bar{x}\bar{y})-P(x*)P(*y)-P(\bar{x}*)P(*\bar{y})}{1-P(x*)P(*y)-P(\bar{x}*)P(*\bar{y})}$
46.	*J* measure	$\max[P(xy)\log\frac{P(y\|x)}{P(*y)}+P(x\bar{y})\log\frac{P(\bar{y}\|x)}{P(*\bar{y})},$ $P(xy)\log\frac{P(x\|y)}{P(x*)}+P(\bar{x}y)\log\frac{P(\bar{x}\|y)}{P(\bar{x}*)}]$

*66.	*L1* norm	$\sum_w \|P(w\|C_x)-P(w\|C_y)\|$
67.	Confusion probability	$\sum_w \frac{P(x\|C_w)P(y\|C_w)P(w)}{P(x*)}$
*68.	Reverse confusion prob.	$\sum_w \frac{P(y\|C_w)P(x\|C_w)P(w)}{P(*y)}$
*69.	Jensen-Shannon diverg.	$\frac{1}{2}[D(p(w\|C_x)\|\|\frac{1}{2}(p(w\|C_x)+p(w\|C_y)))$ $+D(p(w\|C_y)\|\|\frac{1}{2}(p(w\|C_x)+p(w\|C_y)))]$
*70.	Cosine of pointwise *MI*	$\frac{\sum_w MI(w,x)\,MI(w,y)}{\sqrt{\sum_w MI(w,x)^2} \cdot \sqrt{\sum_w MI(w,y)^2}}$
71.	KL divergence	$\sum_w P(w\|C_x) \log \frac{P(w\|C_x)}{P(w\|C_y)}$
72.	Reverse KL divergence	$\sum_w P(w\|C_y) \log \frac{P(w\|C_y)}{P(w\|C_x)}$
*73.	Skew divergence	$D(p(w\|C_x)\|\|\alpha(w\|C_y)+(1-\alpha)p(w\|C_x))$
74.	Reverse skew divergence	$D(p(w\|C_y)\|\|\alpha p(w\|C_x)+(1-\alpha)p(w\|C_y))$
75.	Phrase word coocurrence	$\frac{1}{2}(\frac{f(x\|C_{xy})}{f(xy)}+\frac{f(y\|C_{xy})}{f(xy)})$
76.	Word association	$\frac{1}{2}(\frac{f(x\|C_y)-f(xy)}{f(xy)}+\frac{f(y\|C_x)-f(xy)}{f(xy)})$
	Cosine context similarity:	$\frac{1}{2}(\cos(\mathbf{c}_x,\mathbf{c}_{xy})+\cos(\mathbf{c}_y,\mathbf{c}_{xy}))$ $\mathbf{c}_z=(z_i);\ \cos(\mathbf{c}_x,\mathbf{c}_y)=\frac{\sum x_i y_i}{\sqrt{\sum {x_i}^2} \cdot \sqrt{\sum {y_i}^2}}$
*77.	in boolean vector space	$z_i=\delta(f(w_i\|C_z))$
78.	in *tf* vector space	$z_i=f(w_i\|C_z)$
79.	in *tf·idf* vector space	$z_i=f(w_i\|C_z) \cdot \frac{N}{df(w_i)};\ df(w_i)=\|\{x : w_i \epsilon C_x\}\|$
	Dice context similarity:	$\frac{1}{2}(\mathrm{dice}(\mathbf{c}_x,\mathbf{c}_{xy})+\mathrm{dice}(\mathbf{c}_y,\mathbf{c}_{xy}))$ $\mathbf{c}_z=(z_i);\ \mathrm{dice}(\mathbf{c}_x,\mathbf{c}_y)=\frac{2\sum x_i y_i}{\sum {x_i}^2+\sum {y_i}^2}$
80.	in boolean vector space	$z_i=\delta(f(w_i\|C_z))$
81.	in *tf* vector space	$z_i=f(w_i\|C_z)$
82.	in *tf·idf* vector space	$z_i=f(w_i\|C_z) \cdot \frac{N}{df(w_i)};\ df(w_i)=\|\{x : w_i \epsilon C_x\}\|$

et beaucoup d'autres proposés après 2006 (non inclus ici)

Mesures pour les chaînes de caractères

- Les mesures pour les chaîne de caractères
  - sont de mesures qui calculent la distance entre deux chaînes de texte
  - cherchent une correspondance ou une comparaison approximative
  - sont utiles pour la recherche approximative de chaînes de texte
- Par exemple, les chaînes de caractères "Sam" et "Samuel" peuvent être considérées comme proches
- Une mesure fournit un nombre indiquant une distance spécifique à l'algorithme



Mesures pour les chaînes de caractères

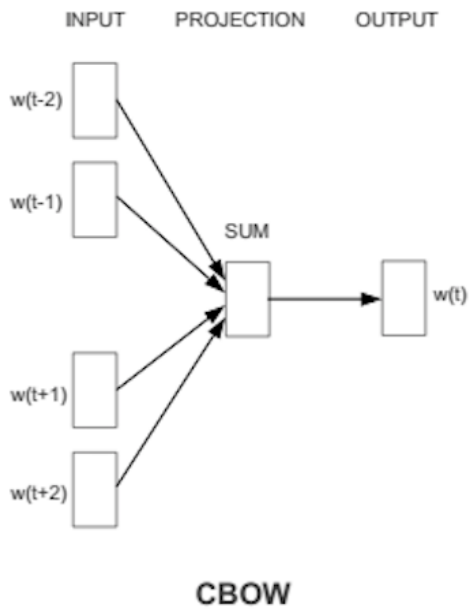
- Liste de mesures pour les chaînes de caractères
  - Sørensen–Dice coefficient
  - Block distance or L1 distance or City block distance
  - Jaro–Winkler distance
  - Simple matching coefficient (SMC)
  - Jaccard similarity or Jaccard coefficient or Tanimoto coefficient
  - Tversky index
  - Overlap coefficient
  - Variational distance
  - Hellinger distance or Bhattacharyya distance
  - Information radius (Jensen–Shannon divergence)
  - Skew divergence
  - Confusion probability
  - Tau metric, an approximation of the Kullback–Leibler divergence
  - Fellegi and Sunters metric (SFS)
  - Maximal matches
  - Grammar-based distance

Exemples

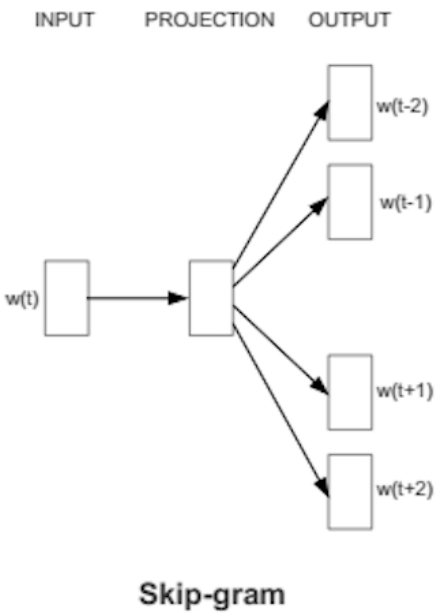
Name	Example
Hamming distance	"**ka**r**o**l**in**" and "**ka**t**hr**i**n**" is 3.
Levenshtein distance and Damerau–Levenshtein distance	**k**i**t**t**e**n and **s**i**t**t**i**n**g** have a distance of 3. 1. **k**itten → **s**itten (substitution of "s" for "k") 2. sitt**e**n → sitt**i**n (substitution of "i" for "e") 3. sittin → sitting**g** (insertion of "g" at the end).
Jaro–Winkler distance	JaroWinklerDist("MARTHA","MARHTA") = $$d_j = \frac{1}{3}\left(\frac{m}{\lvert s_1\rvert} + \frac{m}{\lvert s_2\rvert} + \frac{m-t}{m}\right) = \frac{1}{3}\left(\frac{6}{6} + \frac{6}{6} + \frac{6-\frac{2}{2}}{6}\right) = 0.944$$ • m is the number of *matching characters*; • t is half the number of *transpositions*(`"MARTHA"[3]!=H, "MARHTA"[3]!=T`).
Most frequent k characters	MostFreqKeySimilarity('research', 'seeking', 2) = 2

Word2vec: une technique de l'état de l'art

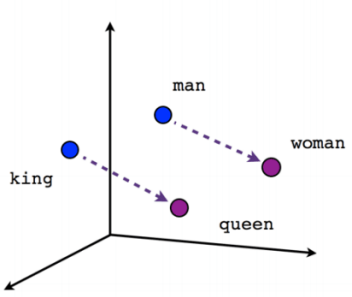
- Word2Vec est une technique intéressante qui pourrait être utilisée pour résoudre ce problème.
  - Proposé en 2013 (il y a 6 ans)
  - Des tonnes d'implémentations disponibles et des vecteurs prêts à l'emploi
  - Chaque mot est mappé à un vecteur (idée pas vraiment nouvelle)
  - L'idée principale est de prédire la relation entre un mot et un contexte en utilisant un réseau neuronal.



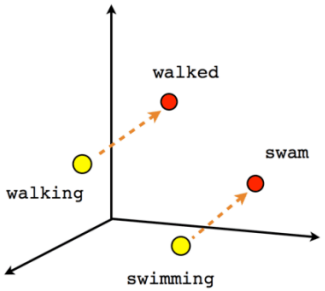
CBOW



Skip-gram



Male-Female



Verb tense

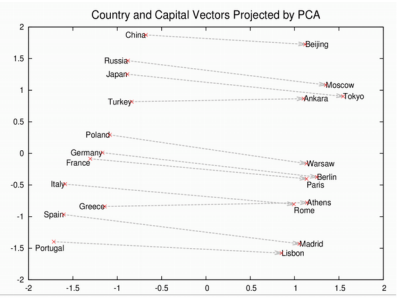
Word2Vec - clarifications

- Deep learning!= Word embeddings
- Word2Vec est une technique pour calculer les embeddings de mots, mais il y en a beaucoup plus
- Plusieurs implémentations publiques disponibles et des vecteurs déjà calculés sur le Web
- C'est cool et ça marche bien, mais pas si facile à entraîner.
- La sélection des paramètres est importante (tout comme le traitement computationel)

Évaluation de la représentation de mots

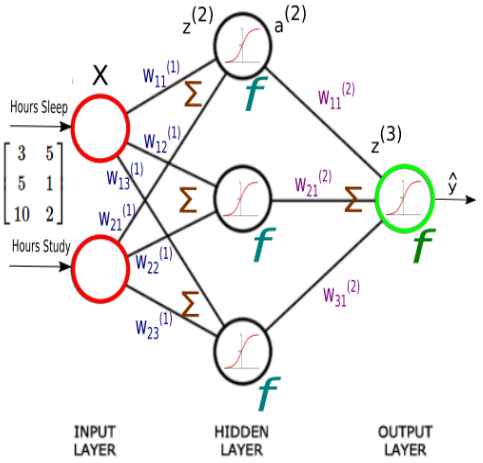
- Ce ne sont que des vecteurs, alors utilisez-les comme ça
- Ensemble de données publiques d'analogies

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

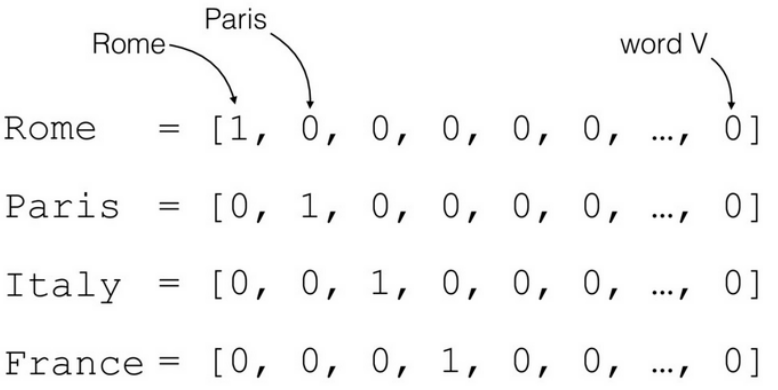


- Plus la précision est élevée, mieux c'est. N'oubliez pas pour quoi vous les voulez.

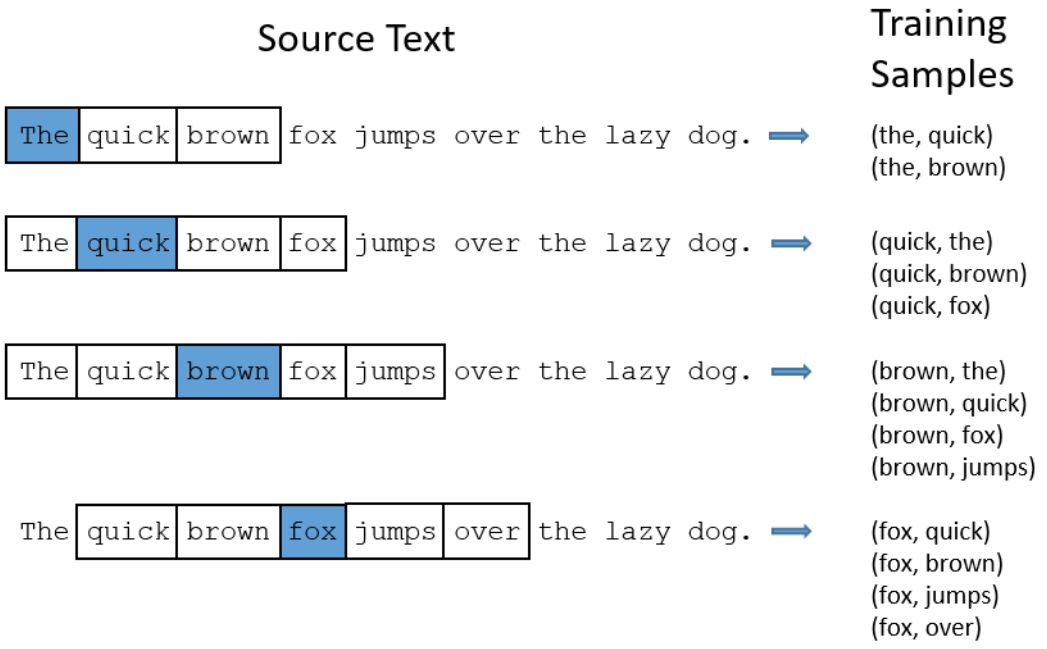
Word2Vec



Artificial Neural Networks

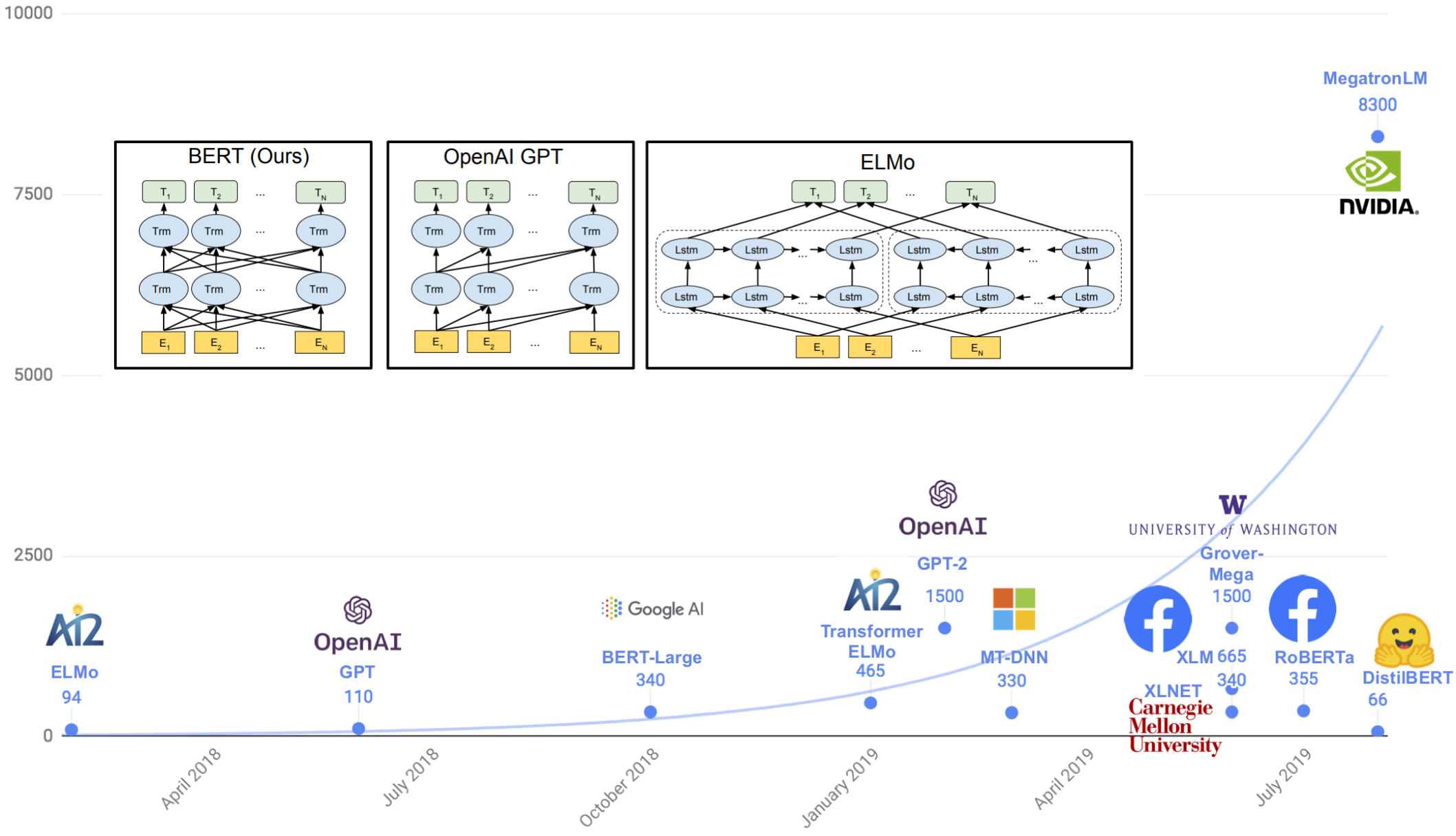


1-hot encoding



Large ‘Unsupervised’ Training

BERT et compagnie



Message à emporter

- Vous connaissez les formules....alors, vous pouvez les coder par vous-même !
- Selon la tâche, le prétraitement peut être une phase critique. (par ex. orange et oranges)
- Différents niveaux de granularité peuvent être utilisés : document, paragraphe, phrase, etc.
- Python
  - nltk.metrics.association module
- Comment savoir ce qui est mieux ?
  - Évaluation !!!!! De nombreux ensembles de données permettent d'évaluer la performance d'algorithmes existants (par ex. ensembles de données SemEval, ensembles de données de similarité sémantique et syntaxique, etc.).