AG du département MIA

CT1 : Bioinformatique et modélisation pour la biologie des systèmes et de synthèse

Représentations vectorielles et apprentissage automatique pour l'alignement d'entités textuelles et de concepts d'ontologie : application à la biologie

Arnaud Ferré

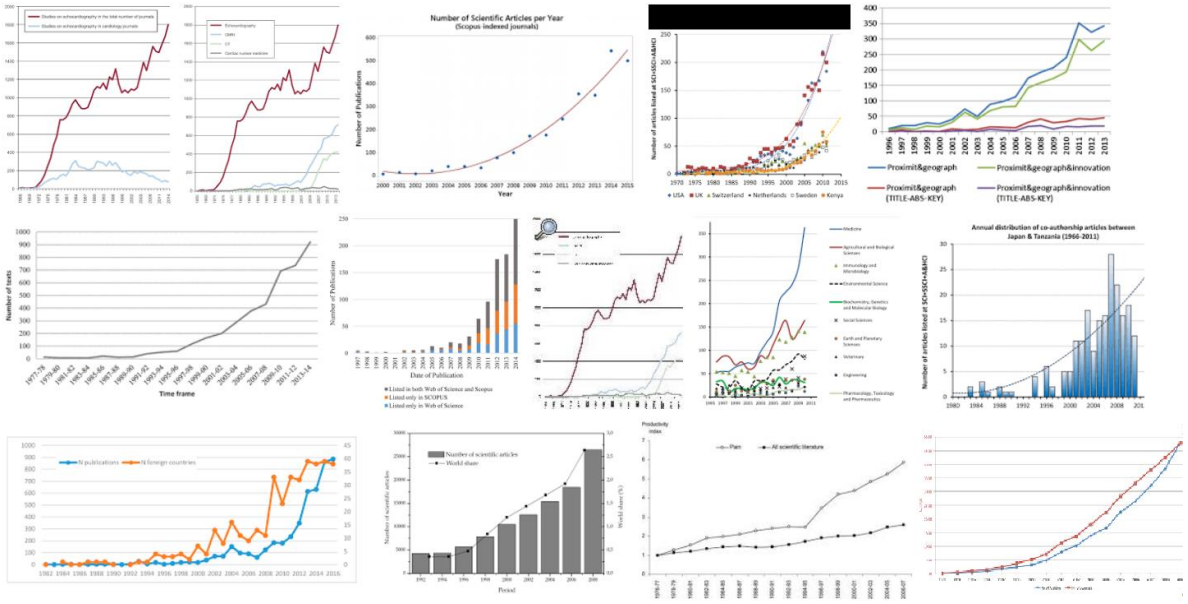
Équipe BIBLIOME, MaIAGE
INRA, Université Paris-Saclay

Directrice de thèse :
Claire Nédellec

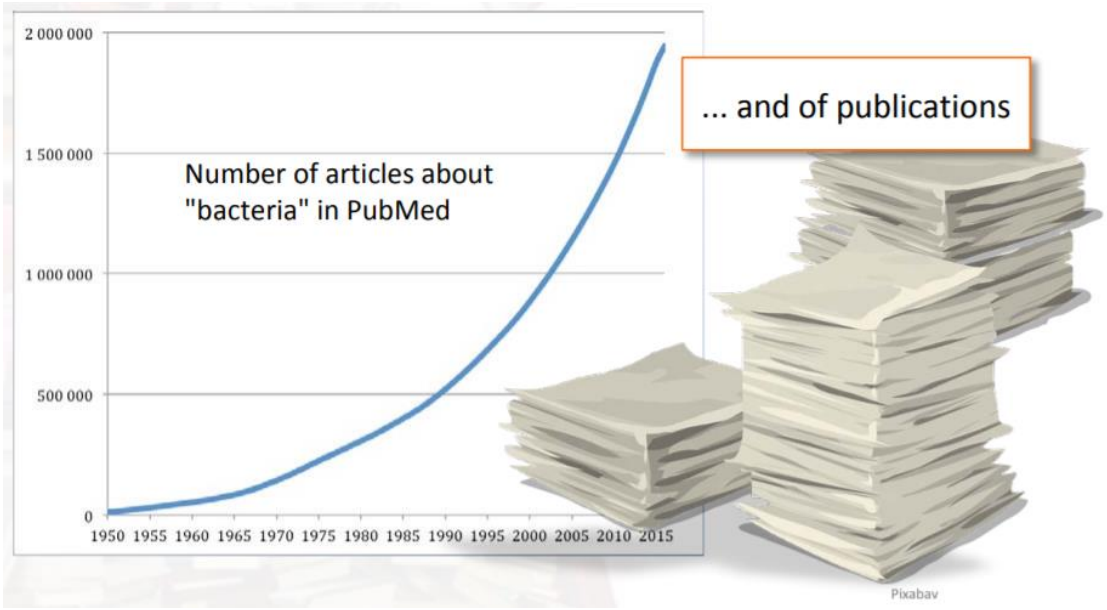
CONTEXTE : Problématique

Entre 31% à 85% des données numérisées seraient des données non-structurées (dont données textuelles)

Source : *Philip Russom. 2007. BI search and text analytics*

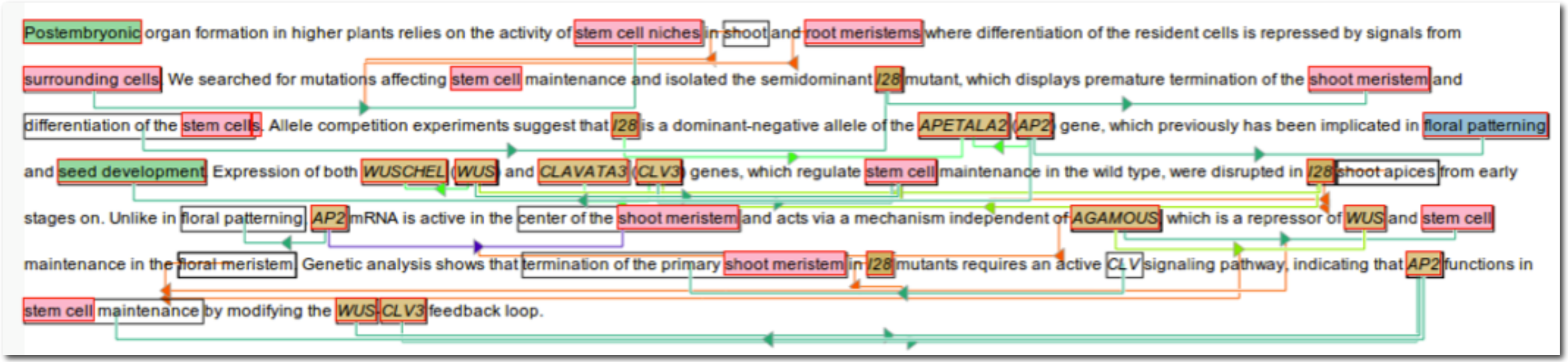


Recherche sur Google Images : "*Scientific publications number evolution*"

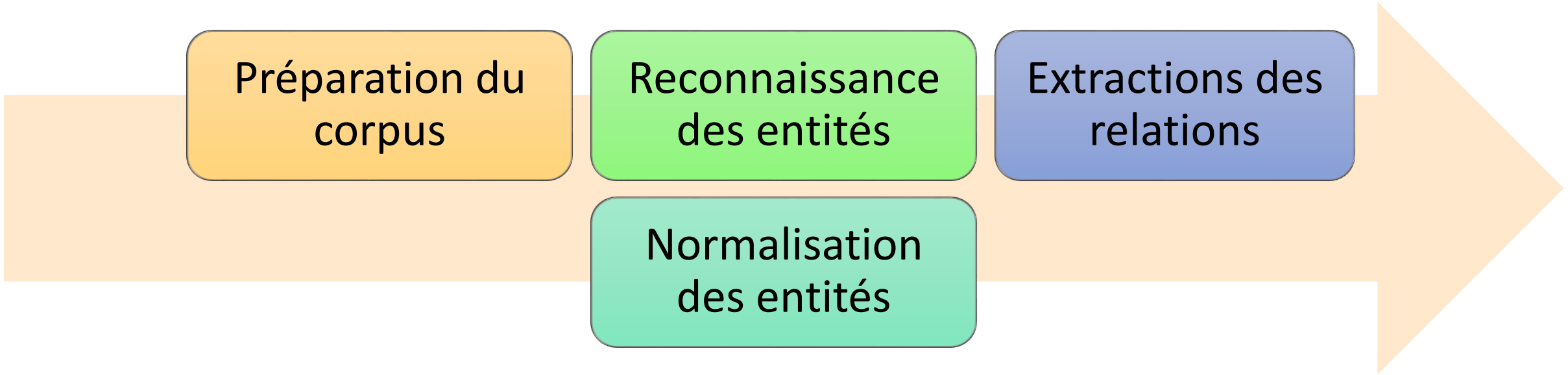


CONTEXTE : Problématique

Besoin en extraction d'information : plus que de la fouille de documents.
Différence entre récupérer les documents qui contiennent des informations d'intérêt et récupérer directement ces informations structurées.



INTRODUCTION : L'extraction d'information



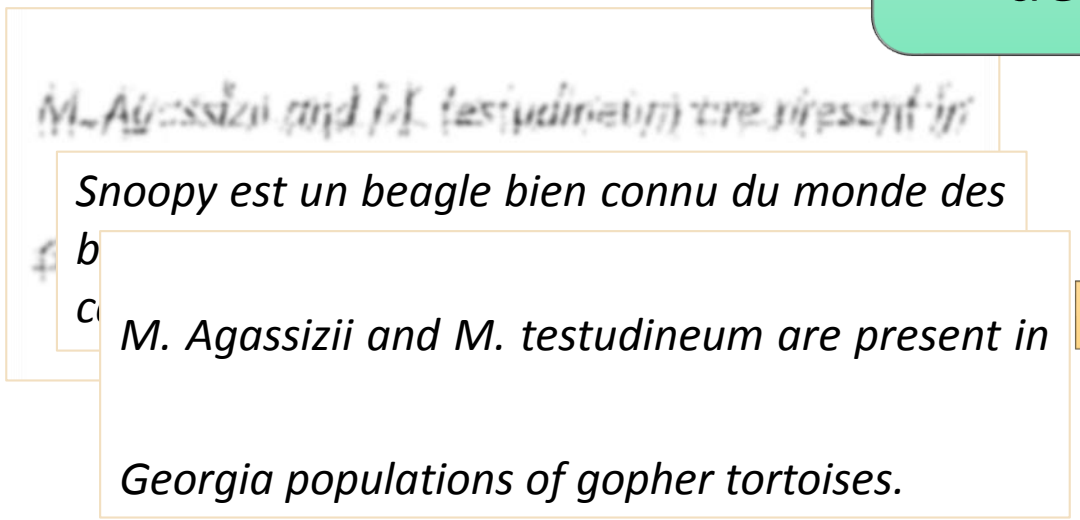
INTRODUCTION : L'extraction d'information

Préparation du corpus

Reconnaissance des entités

Extractions des relations

Normalisation des entités



M. Agassizii and M. testudineum are present in

Georgia populations of gopher tortoises.

INTRODUCTION : L'extraction d'information

Préparation du corpus

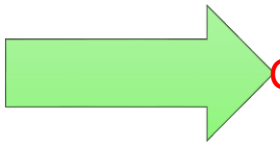
Reconnaissance des entités

Extractions des relations

Normalisation des entités

M. Agassizii and M. testudineum are present in

Georgia populations of gopher tortoises.



Bactérie — Bactérie

M. Agassizii and *M. testudineum* are present in

Geograph. — Habitat bactérien

Georgia *populations of gopher tortoises*.

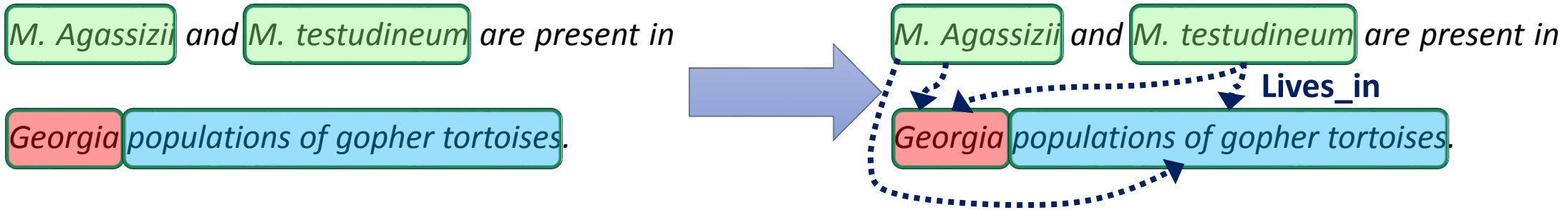
INTRODUCTION : L'extraction d'information

Préparation du corpus

Reconnaissance des entités

Extractions des relations

Normalisation des entités



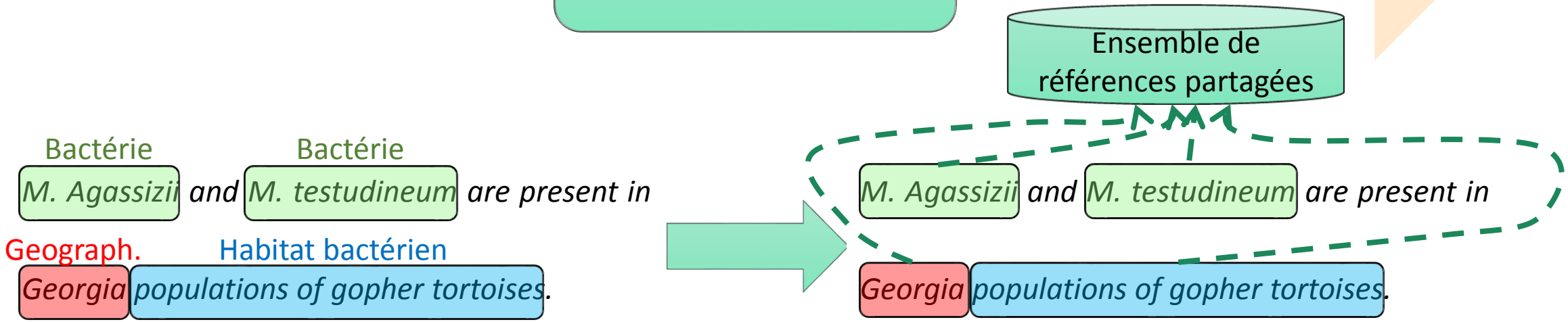
INTRODUCTION : L'extraction d'information

Préparation du corpus

Reconnaissance des entités

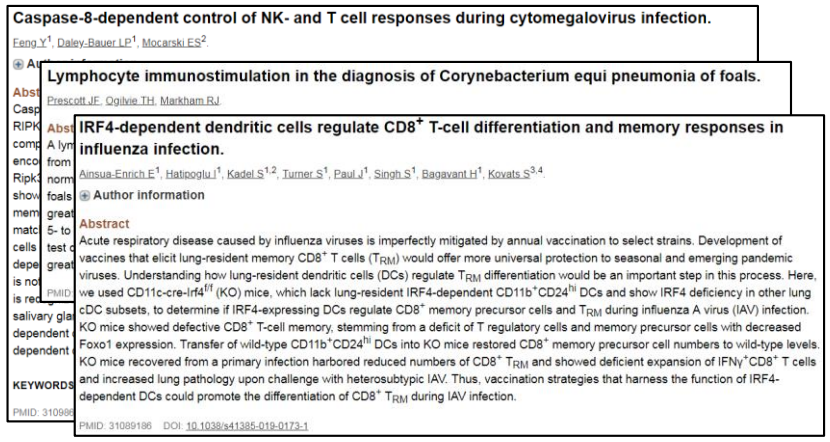
Extractions des relations

Normalisation des entités



INTRODUCTION : intérêts de la normalisation d'entités

Reconnaissance d'entités + extraction de relations :



Caspase-8-dependent control of NK- and T cell responses during cytomegalovirus infection.

Feng Y[1], Daley-Bauer LP[1], Mocarski ES[2].

Lymphocyte immunostimulation in the diagnosis of Corynebacterium equi pneumonia of foals.

Prescott JF, Ogilvie TH, Markham RJ.

IRF4-dependent dendritic cells regulate CD8$^+$ T-cell differentiation and memory responses in influenza infection.

Ainsua-Enrich E[1], Hatipoglu I[1], Kadel S[1,2], Turner S[1], Paul J[1], Singh S[1], Bagavant H[1], Kovats S[3,4].

Author information

Abstract

Acute respiratory disease caused by influenza viruses is imperfectly mitigated by annual vaccination to select strains. Development of vaccines that elicit lung-resident memory CD8$^+$ T cells (T_{RM}) would offer more universal protection to seasonal and emerging pandemic viruses. Understanding how lung-resident dendritic cells (DCs) regulate T_{RM} differentiation would be an important step in this process. Here, we used CD11c-cre-Irf4$^{f/f}$ (KO) mice, which lack lung-resident IRF4-dependent CD11b$^+$CD24^{hi} DCs and show IRF4 deficiency in other lung cDC subsets, to determine if IRF4-expressing DCs regulate CD8$^+$ memory precursor cells and T_{RM} during influenza A virus (IAV) infection. KO mice showed defective CD8$^+$ T-cell memory, stemming from a deficit of T regulatory cells and memory precursor cells with decreased Foxo1 expression. Transfer of wild-type CD11b$^+$CD24^{hi} DCs into KO mice restored CD8$^+$ memory precursor cell numbers to wild-type levels. KO mice recovered from a primary infection harbored reduced numbers of CD8$^+$ T_{RM} and showed deficient expansion of IFNγ$^+$CD8$^+$ T cells and increased lung pathology upon challenge with heterosubtypic IAV. Thus, vaccination strategies that harness the function of IRF4-dependent DCs could promote the differentiation of CD8$^+$ T_{RM} during IAV infection.

PMID: 31089186 DOI: 10.1038/s41385-019-0173-1

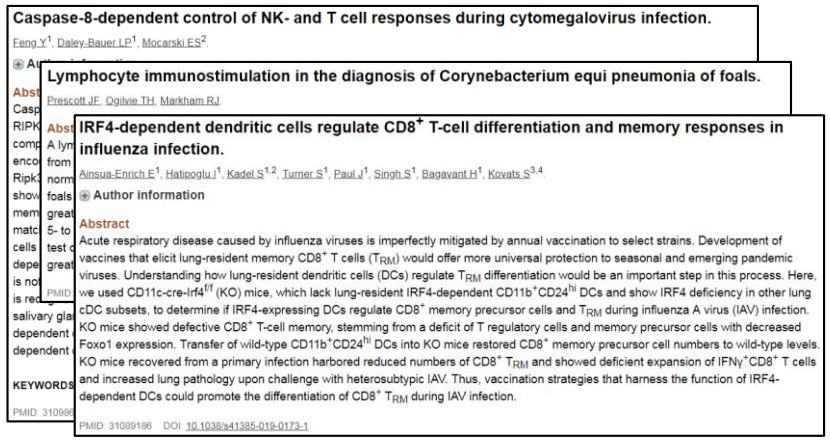
Lives_in

Bactérie	Habitat (mention)
Listeria	*"CD20"*
Pseudomonas	*"CD20-positive cells"*
Corynebacterium	*"monoclonal B cells"*
Pseudomonas	*"T cells"*
Listeria	*"Lymphocytic"*

Requête : Bactéries infectieuses "*lymphocite*" ? ⟹ 0 ou 1 réponse

INTRODUCTION : intérêts de la normalisation d'entités

Dès qu'il est nécessaire de comparer des expressions textuelles entre elles ou avec des concepts :



Caspase-8-dependent control of NK- and T cell responses during cytomegalovirus infection.

Feng Y[1], Daley-Bauer LP[1], Mocarski ES[2].

Lymphocyte immunostimulation in the diagnosis of Corynebacterium equi pneumonia of foals.

Prescott JF, Ogilvie TH, Markham RJ.

IRF4-dependent dendritic cells regulate CD8$^+$ T-cell differentiation and memory responses in influenza infection.

Ainsua-Enrich E[1], Hatipoglu I[1], Kadel S[1,2], Turner S[1], Paul J[1], Singh S[1], Bagavant H[1], Kovats S[3,4].

Author information

Abstract

Acute respiratory disease caused by influenza viruses is imperfectly mitigated by annual vaccination to select strains. Development of vaccines that elicit lung-resident memory CD8$^+$ T cells (T_{RM}) would offer more universal protection to seasonal and emerging pandemic viruses. Understanding how lung-resident dendritic cells (DCs) regulate T_{RM} differentiation would be an important step in this process. Here, we used CD11c-cre-Irf4$^{f/f}$ (KO) mice, which lack lung-resident IRF4-dependent CD11b$^+$CD24^{hi} DCs and show IRF4 deficiency in other lung cDC subsets, to determine if IRF4-expressing DCs regulate CD8$^+$ memory precursor cells and T_{RM} during influenza A virus (IAV) infection. KO mice showed defective CD8$^+$ T-cell memory, stemming from a deficit of T regulatory cells and memory precursor cells with decreased Foxo1 expression. Transfer of wild-type CD11b$^+$CD24^{hi} DCs into KO mice restored CD8$^+$ memory precursor cell numbers to wild-type levels. KO mice recovered from a primary infection harbored reduced numbers of CD8$^+$ T_{RM} and showed deficient expansion of IFNγ$^+$CD8$^+$ T cells and increased lung pathology upon challenge with heterosubtypic IAV. Thus, vaccination strategies that harness the function of IRF4-dependent DCs could promote the differentiation of CD8$^+$ T_{RM} during IAV infection.

PMID: 31089186 DOI: 10.1038/s41385-019-0173-1

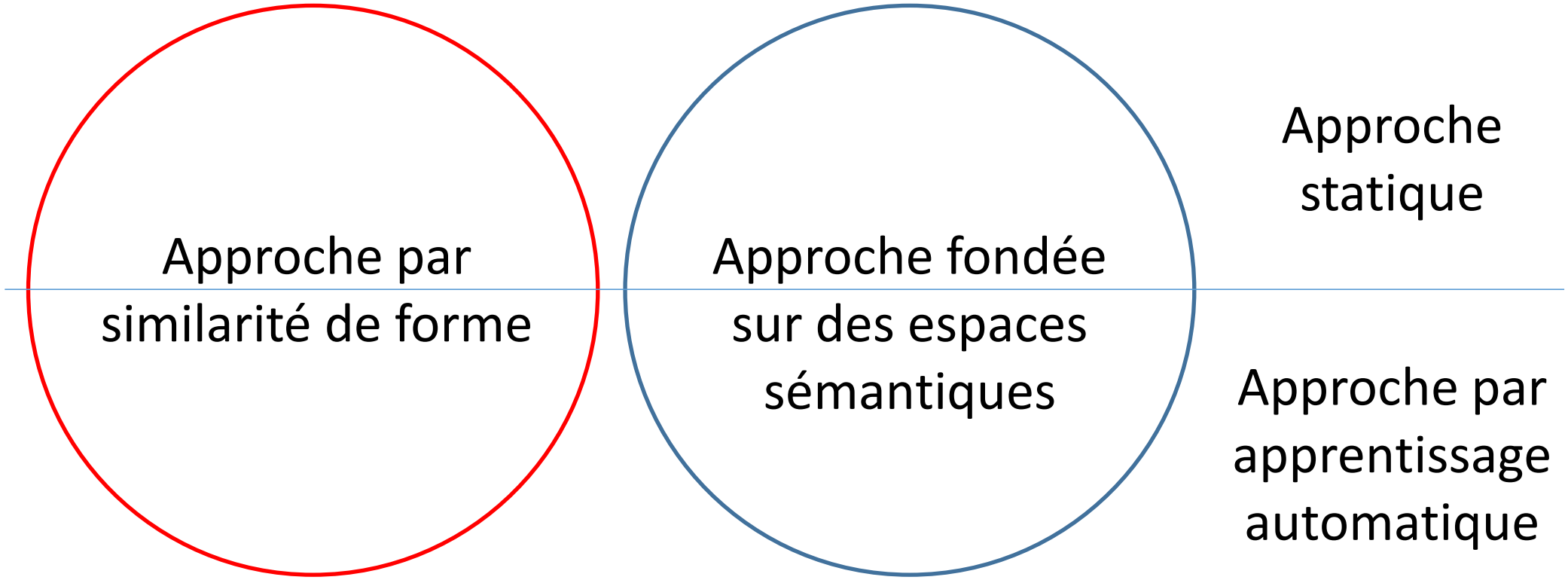
Lives_in

Bactérie	Habitat (mention)	Habitat (concept)
Listeria	*"CD20"*	OBT:001342: lymphocyte
Pseudomonas	*"CD20-positive cells"*	OBT:001342: lymphocyte
Corynebacterium	*"monoclonal B cells"*	OBT:001342: lymphocyte
Pseudomonas	*"T cells"*	OBT:001342: lymphocyte
Listeria	*"Lymphocytic"*	OBT:001342: lymphocyte

Requête : Bactéries infectieuses de lymphocyte ? ⟹ 5 réponses

ÉTAT DE L'ART :

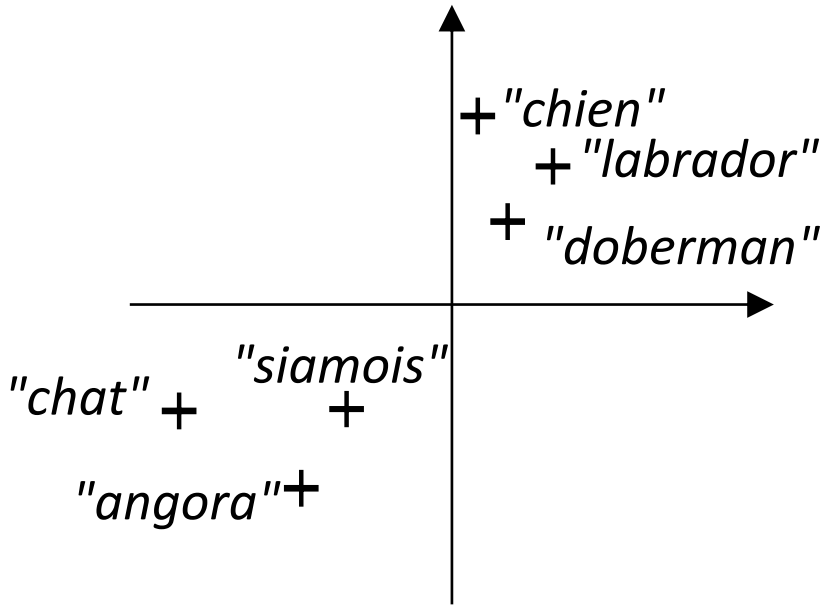
ÉTAT DE L'ART : 2 approches



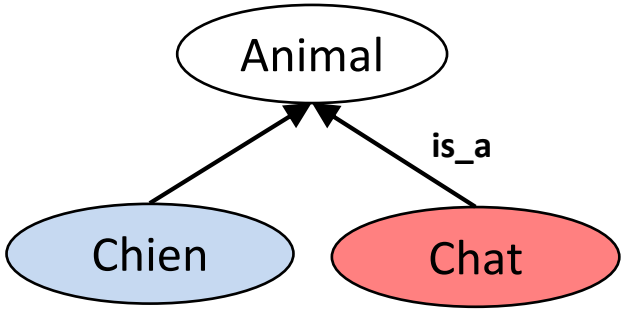
Exemple : "*monoclonal B cells*" et "*lymphocyte*"
Pour Bacteria Biotope : > 60% de non-similarités entre mentions et étiquettes associés

ÉTAT DE L'ART : cadre général

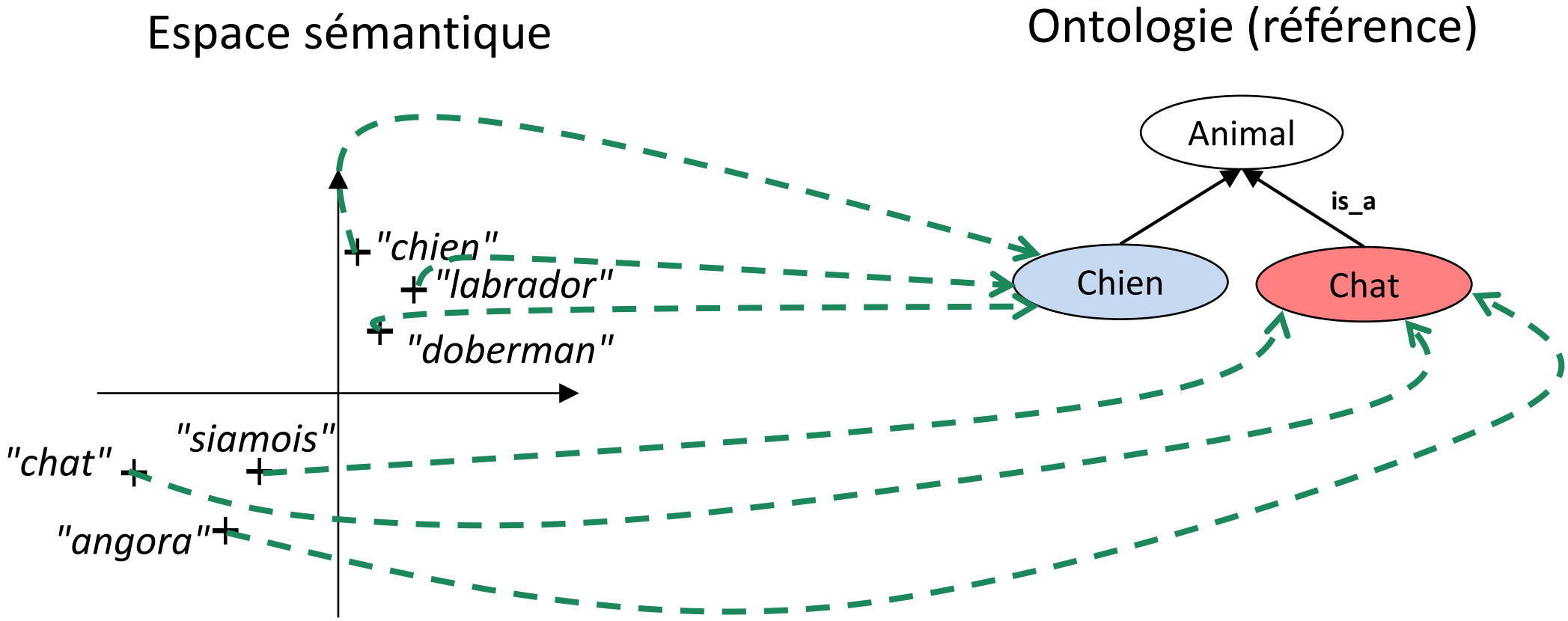
Espace sémantique



Ontologie (référence)



ÉTAT DE L'ART : cadre général



ÉTAT DE L'ART : sémantique distributionnelle

Idée 1 : si ses contextes d’apparition sont connus, on peut appréhender le sens d’un mot.

Exemple : "*Un chien ronge un _____.*"

ÉTAT DE L'ART : sémantique distributionnelle

Idée 1 : si ses contextes d'apparition sont connus, on peut appréhender le sens d'un mot.

Exemple : "*Un chien ronge un _____.*"

Idée 2 : si deux mots partagent les mêmes contextes, ils auront un sens proche.

Exemple : "*Un chien ronge un bâton*" $\Longrightarrow$ "*bâton"* et *"os"* désignent des entités qui peuvent être rongées par un chien, donc similaires.

(Harris, 1954)
(Firth, 1957)

ÉTAT DE L'ART : sémantique distributionnelle

Méthodes par sacs-de-mots distributionnels

"Un chien ronge un bâton."

"Un chien ronge un os."

Vocabulaire (sans mots-outils) :
{"chien", "ronge", "bâton", "os"}

	"chien"	*"ronge"*	*"bâton"*	*"os"*
"chien"	0	2	1	1
"ronge"	2	0	1	1
"bâton"	1	1	0	0
"os"	1	1	0	0

ÉTAT DE L'ART : sémantique distributionnelle

Méthodes par sacs-de-mots distributionnels

"Un chien ronge un bâton."

"Un chien ronge un os."

Vocabulaire (sans mots-outils) :
{"chien", "ronge", "bâton", "os"}

	"chien"	*"ronge"*	*"bâton"*	*"os"*
"chien"	0	2	1	1
"ronge"	2	0	1	1
"bâton"	1	1	0	0
"os"	1	1	0	0

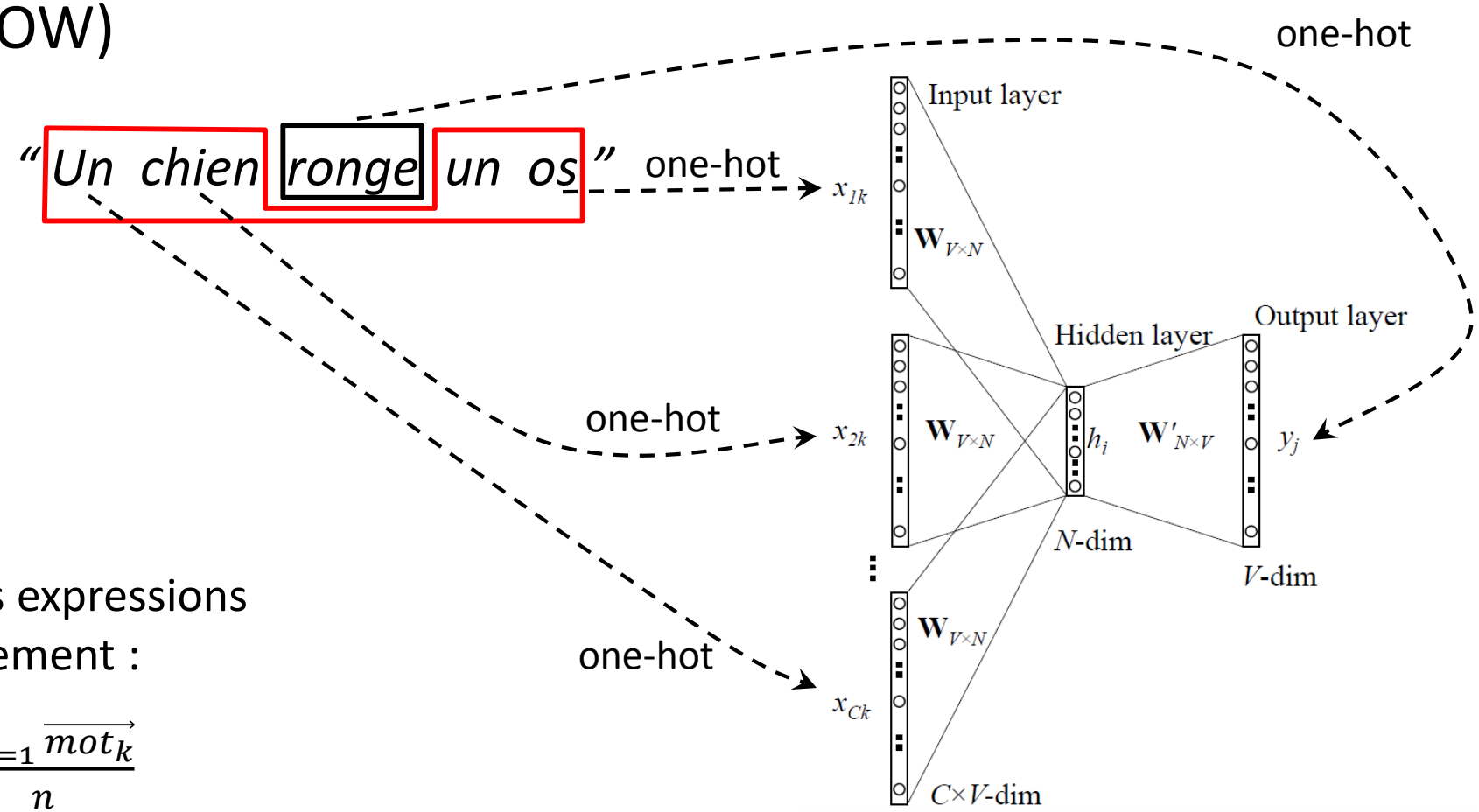
Distance euclidienne :

	"chien"	*"ronge"*	*"bâton"*	*"os"*
"chien"	0	2,83	2,45	2,45
"ronge"	2,83	0	2,45	2,45
"bâton"	2,45	2,45	0	**0**
"os"	2,45	2,45	**0**	0

(Hinton et al., 1986)
(Pollack, 1990)
(Deerwester et al., 1990)
(Elman, 1991)

ÉTAT DE L'ART : sémantique distributionnelle

Word2Vec (CBOW)

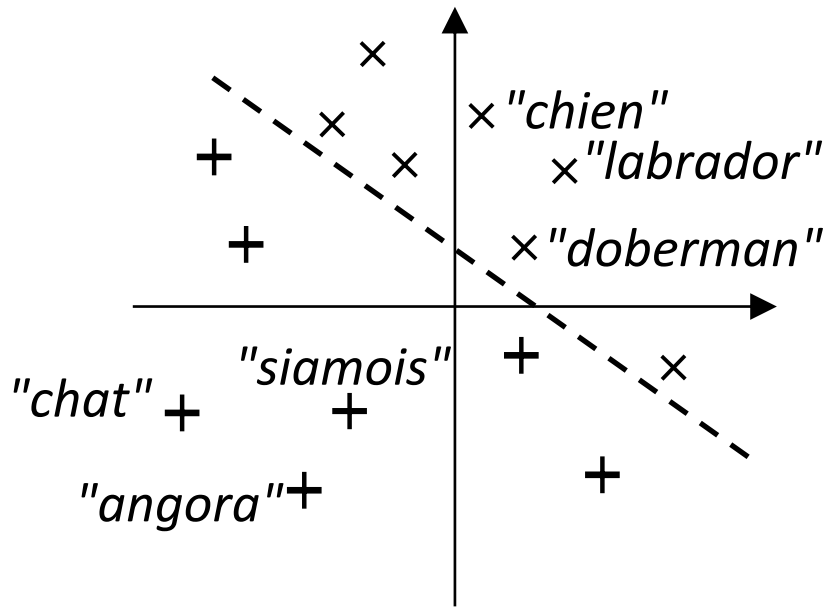


Pour représenter des expressions multi-mots, classiquement :

$$\overrightarrow{expression} = \frac{\sum_{k=1}^{n} \overrightarrow{mot_k}}{n}$$

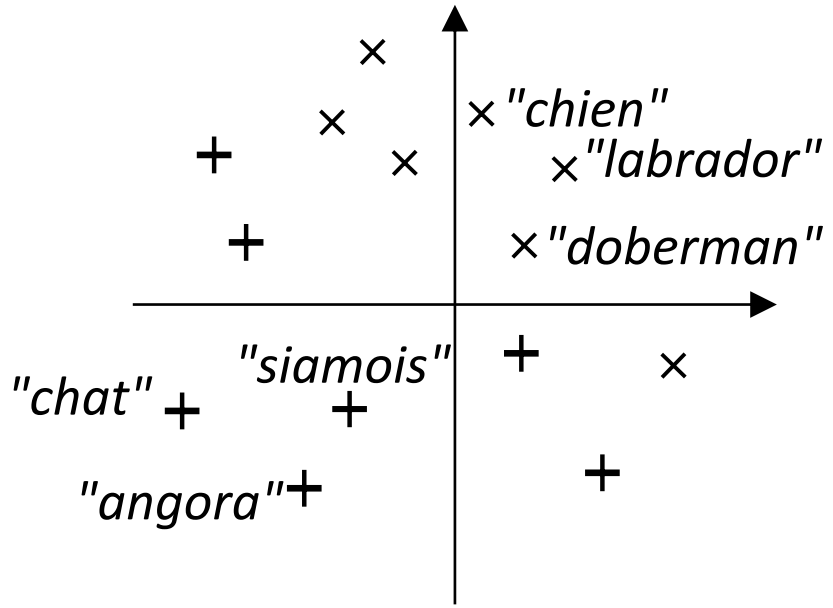
(Mikolov et al. 2013)

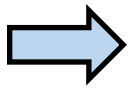
ÉTAT DE L'ART : normalisation distributionnelle

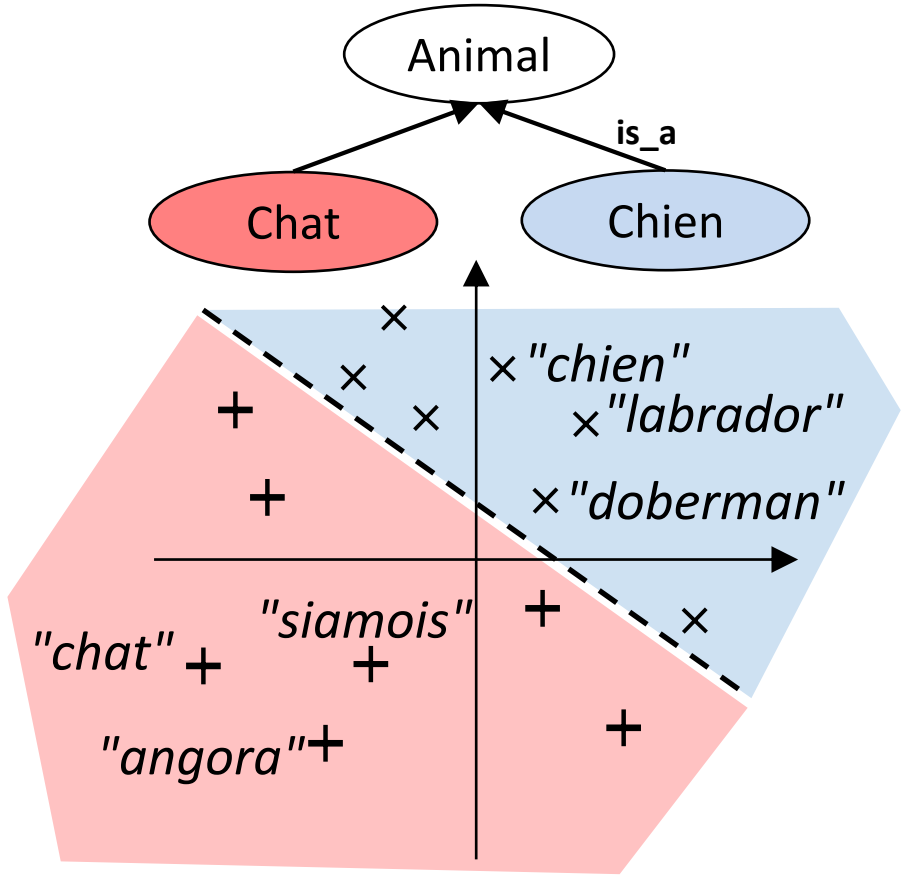


ÉTAT DE L'ART : normalisation distributionnelle

Approche par classification supervisée



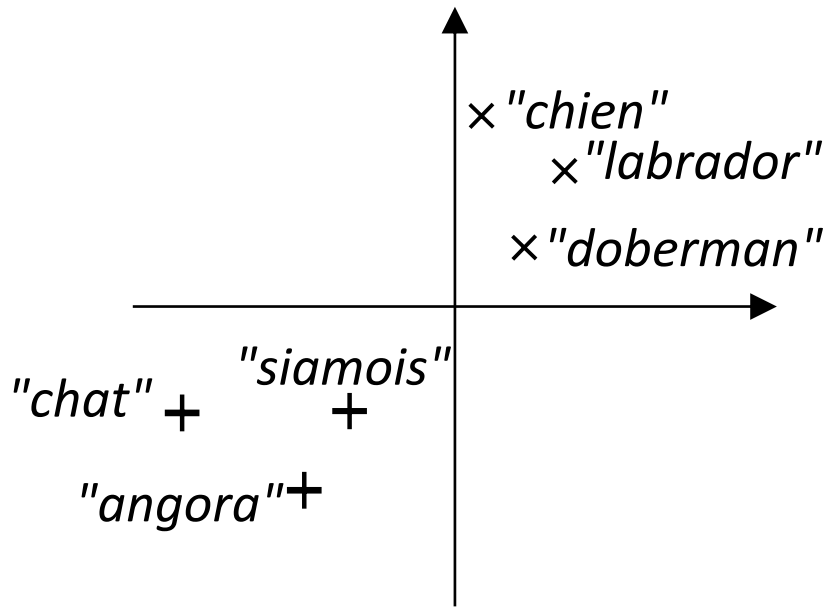


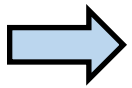


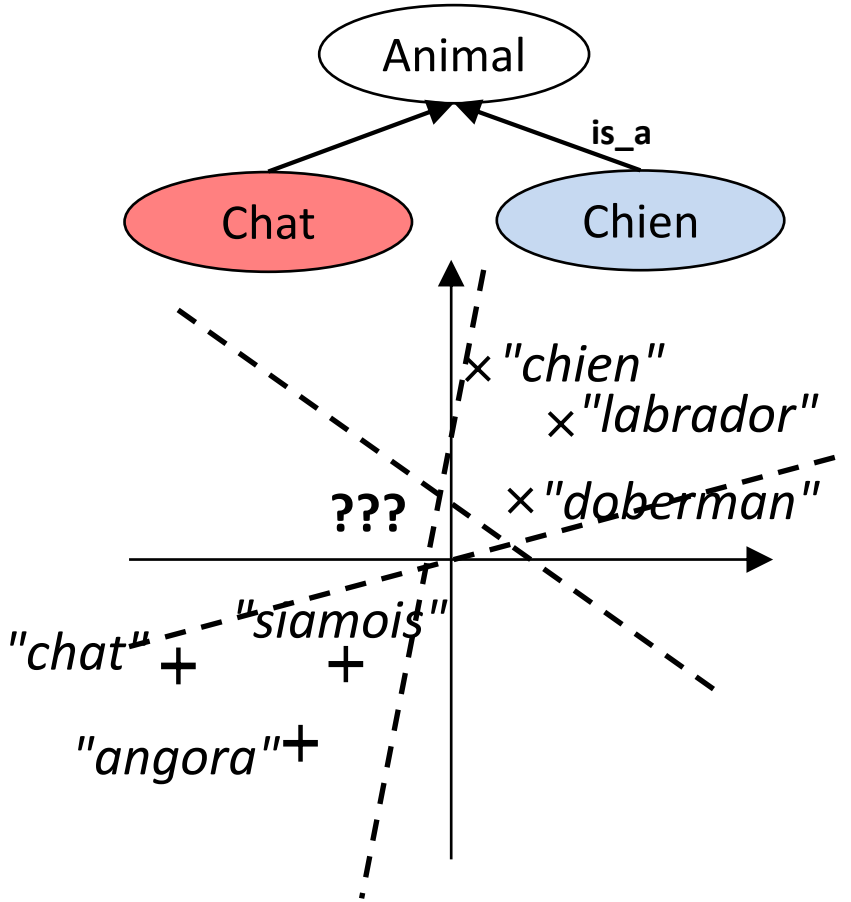
(Limsopatham and Collier, 2016)

ÉTAT DE L'ART : verrous scientifiques

Approche par classification supervisée

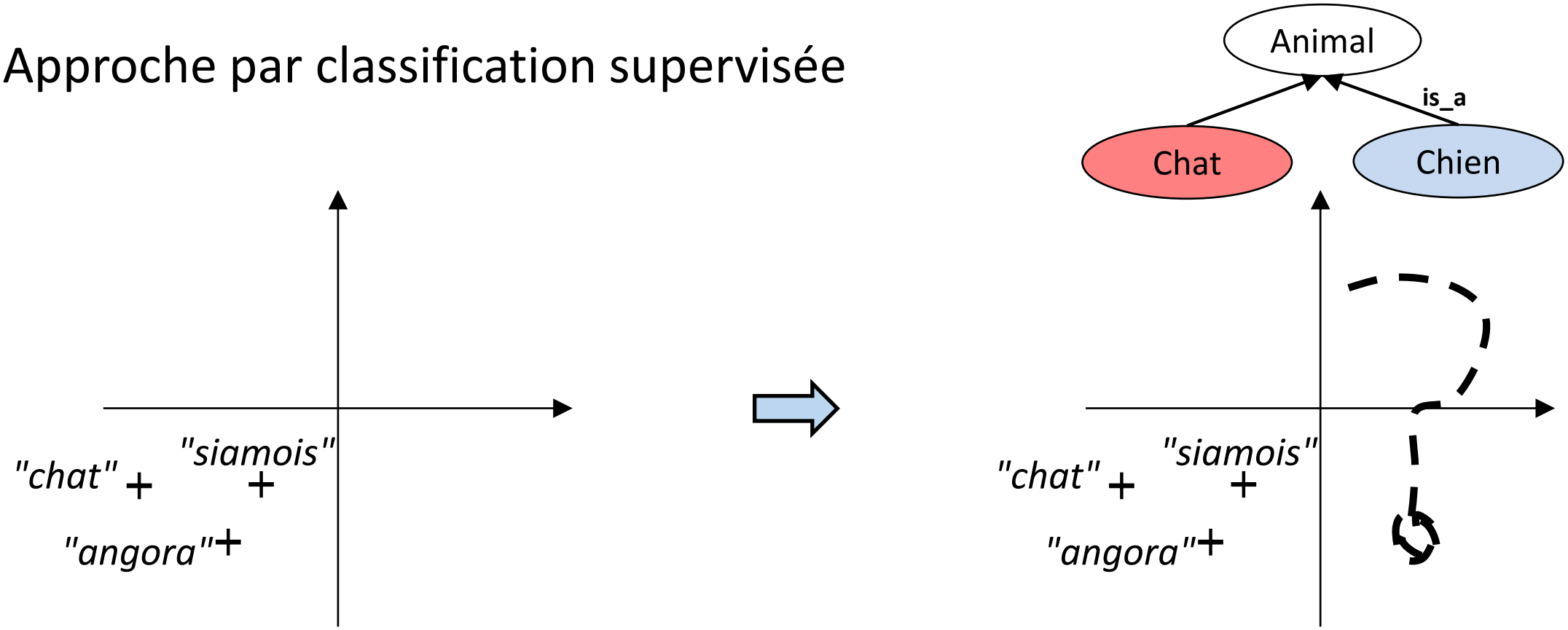






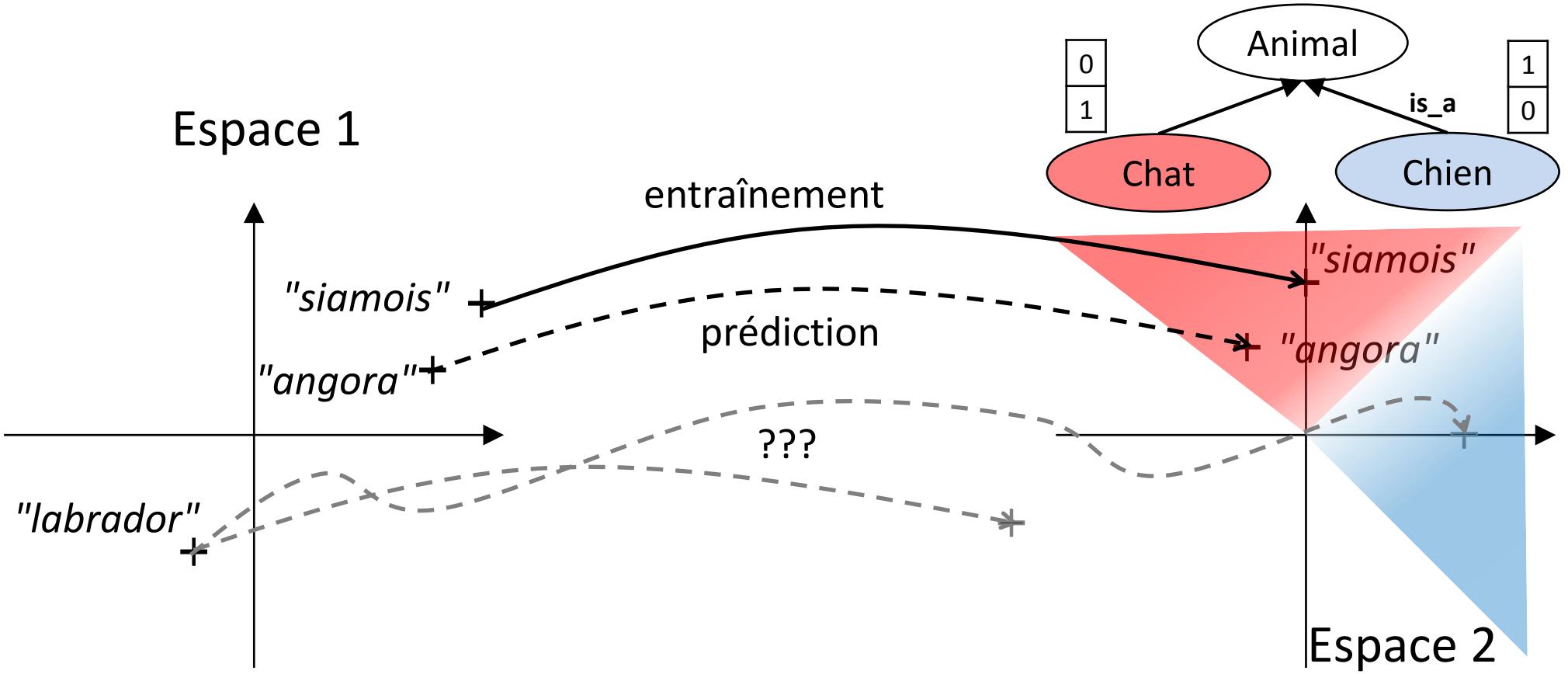
ÉTAT DE L'ART : verrous scientifiques

Approche par classification supervisée



MÉTHODE :

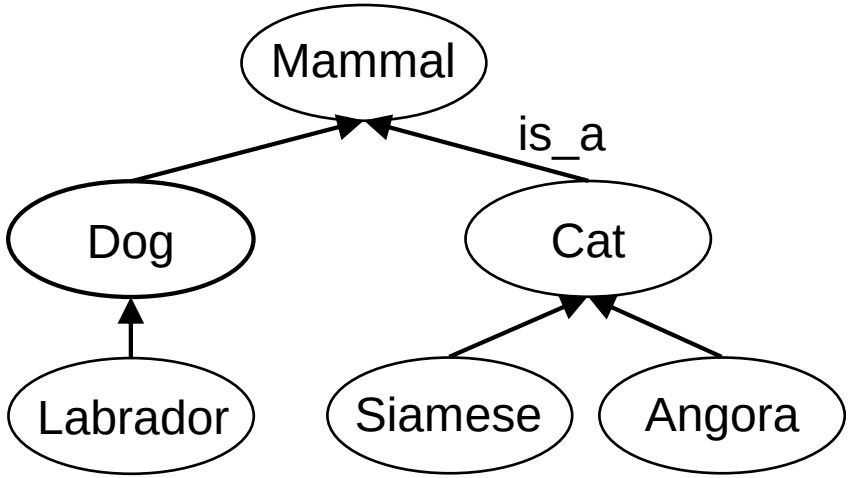
MÉTHODE : difficulté abordée



Problème de classification avec des classes n'apparaissant pas dans les exemples d'entraînement

(Hugo Larochelle et al., 2008, "**Zero-data learning** of new tasks.")

MÉTHODE : Ancestry

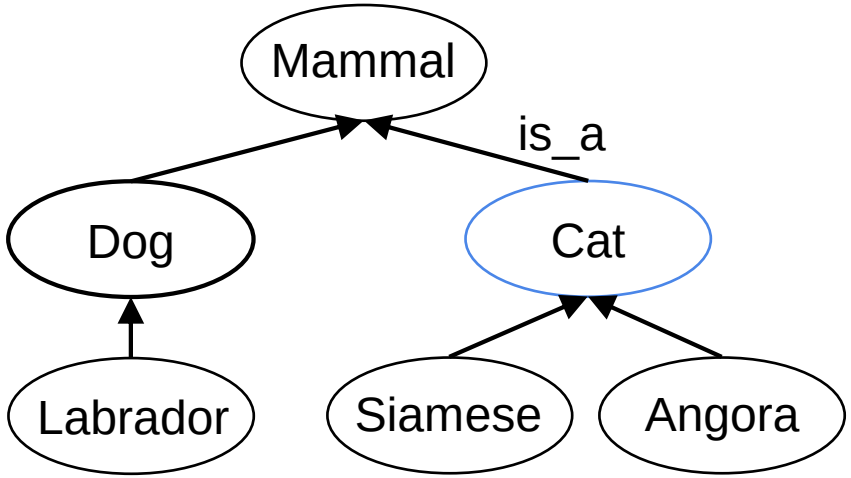


$$\forall k \in [\![1, n]\!], v_{c_k} = \left(w_{c_k}^0, \ldots, w_{c_k}^i, \ldots, w_{c_k}^n \right)$$

$$w_{c_k}^i = \begin{cases} 1 & si\ i = k \\ 1 & si\ c_i\ ancêtre\ de\ c_k \\ 0 & sinon \end{cases}$$

(Arnaud Ferré et al., 2017, "Representation of complex terms in a vector space structured by an ontology for a normalization task. ")

MÉTHODE : Ancestry

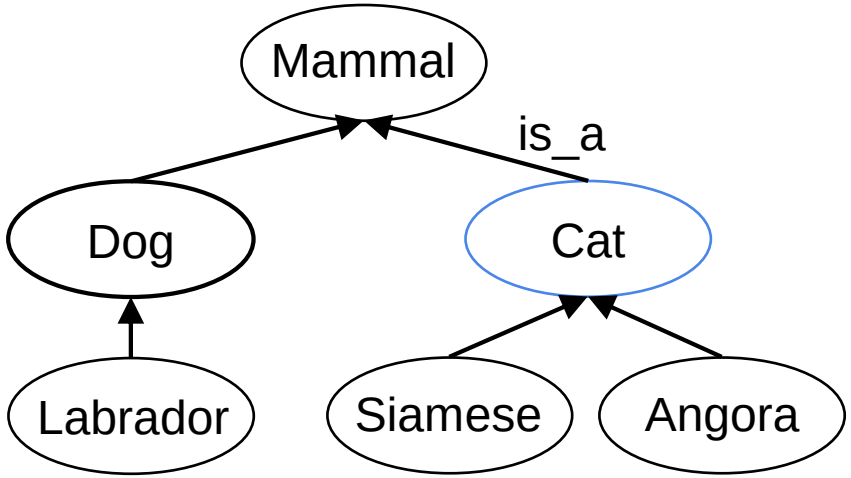


$$\forall k \in [\![1, n]\!], v_{c_k} = \left(w_{c_k}^0, \ldots, w_{c_k}^i, \ldots, w_{c_k}^n \right)$$

$$w_{c_k}^i = \begin{cases} 1 & si\ i = k \\ 1 & si\ c_i\ ancêtre\ de\ c_k \\ 0 & sinon \end{cases}$$

One-Hot	Mammal	Dog	Cat	Labrador	Siamois	Angora
$\overrightarrow{Mammal}$	1	0	0	0	0	0
$\overrightarrow{Dog}$	0	1	0	0	0	0
$\overrightarrow{Cat}$	0	0	1	0	0	0
$\overrightarrow{Labrador}$	0	0	0	1	0	0
$\overrightarrow{Siamese}$	0	0	0	0	1	0
$\overrightarrow{Angora}$	0	0	0	0	0	1

MÉTHODE : Ancestry



$$\forall k \in [\![1,n]\!], v_{c_k} = \left(w_{c_k}^0, \ldots, w_{c_k}^i, \ldots, w_{c_k}^n \right)$$

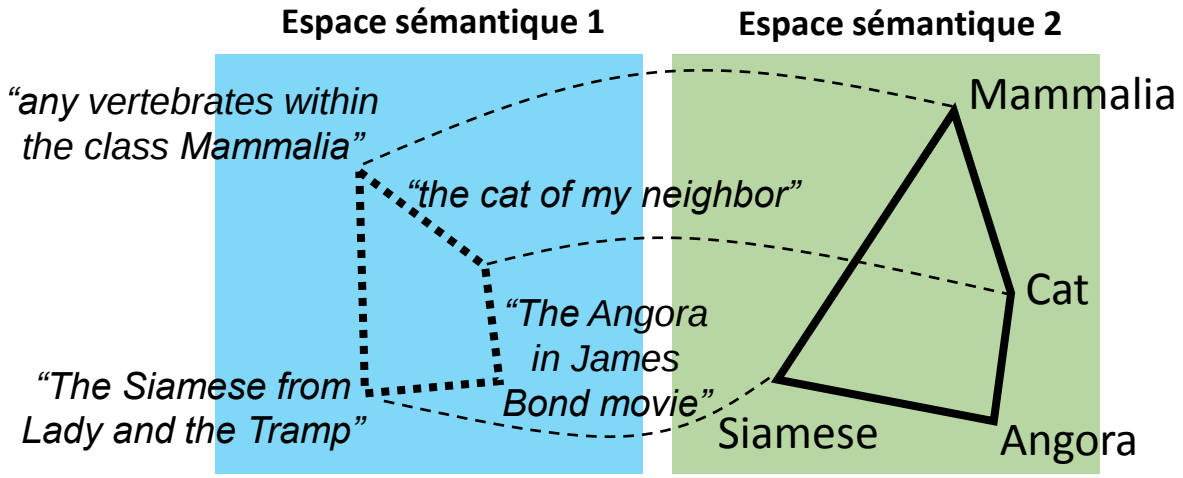
$$w_{c_k}^i = \begin{cases} 1 & si\ i = k \\ 1 & si\ c_i\ ancêtre\ de\ c_k \\ 0 & sinon \end{cases}$$

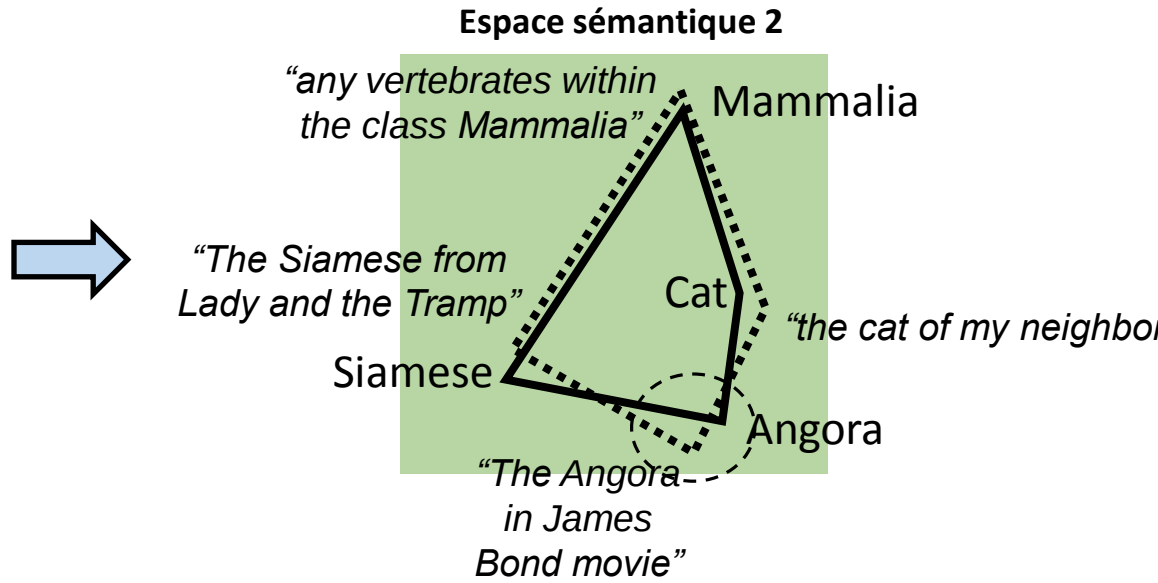
Ancestry	Mammal	Dog	Cat	Labrador	Siamois	Angora
$\overrightarrow{Mammal}$	1	0	0	0	0	0
$\overrightarrow{Dog}$	**1**	1	0	0	0	0
$\overrightarrow{Cat}$	**1**	0	1	0	0	0
$\overrightarrow{Labrador}$	**1**	**1**	0	1	0	0
$\overrightarrow{Siamese}$	**1**	0	**1**	0	1	0
$\overrightarrow{Angora}$	**1**	0	**1**	0	0	1

Similarité à Cat	**Similarité cosinus**
Cat	1,00
Siamese	0.82
Angora	0.82
Mammal	0.71
Dog	0.50
Labrador	0.41

MÉTHODE : CONTES

Hypothèse : Si les deux espaces sont suffisamment similaires, alors une projection qui permettra d'obtenir des prédictions pertinentes peut être déterminée.





MÉTHODE : CONTES

Comment obtenir cette fonction de projection à partir d'exemples ?

Algorithme d'apprentissage :

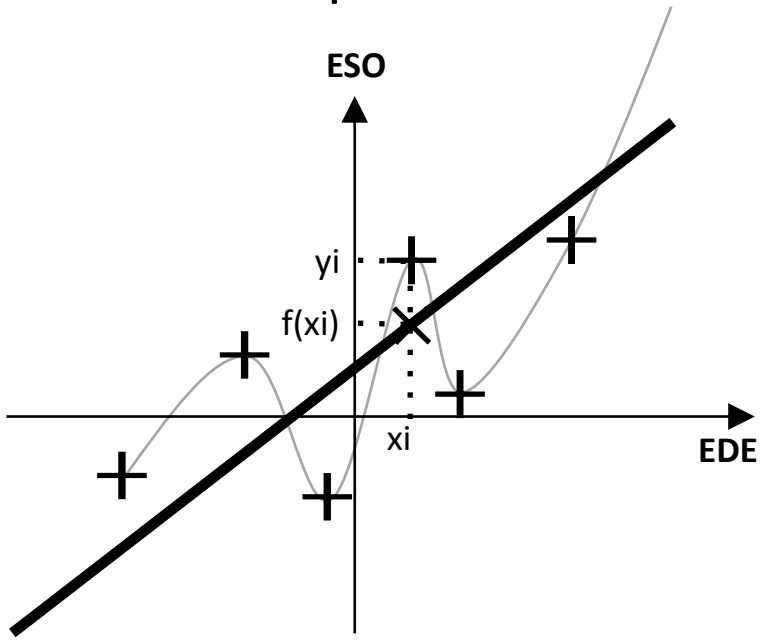
- Régression linéaire multivariée
- Définition :

Soit $n, m \in \mathbb{N}^+$, tels que :

$$\forall x \in \mathbb{R}^n, \forall A \in \mathcal{M}_{m,n}(\mathbb{R}), f_A(x) = A.x$$

Soit N $\in \mathbb{N}^+, \forall i \in [\![1, N]\!], (x_i, y_i) \in \mathbb{R}^n \times \mathbb{R}^m$, alors :

$$A^* \approx \underset{A \in \mathcal{M}_{n,m}(\mathbb{R})}{\arg\min} \sum_{i=1}^{i=N} dist(y_i - f_A(x_i))$$



Intérêts :

- Limitation du sur-apprentissage
- La convergence ne nécessite pas de grand jeu de données
- Conservation relative de la structure de l'espace projeté

MÉTHODE : Résultats

Challenge Bacteria Biotope
de BioNLP Shared Task 2016

Nécessite des exemples annotés	**Méthode**	**Score**
	CONTES (Ferré et al., 2017)	0,61
	Full-CONTES	**0,72**
	HONOR (Ferré et al., 2018)	0,74
	Full-HONOR	**0,76**
Ne nécessite pas d'exemples annotés		
Fondée sur des espaces sémantiques		
	WSEP-CONTES	0,59
	WSOT-CONTES	0,63
	WSOT-HONOR	**0,73**
	BOUNEL (Karadeniz et al., 2019)	0,66
	Turku (Mehryary et al., 2017)	0,63
	BOUN (Tiftikci et al., 2016)	0,62
Non-fondée sur des espaces sémantiques		
	ToMap (avec règles spécifiques au domaine)	0,66
	ToMap	0,61
	Méthode de référence	0,54

CONCLUSION

Nouvelle approche de normalisation par les concepts d'une ontologie performante :

- Malgré peu (voire pas) de données d’entraînement
- Répond au problème de variabilité de forme
- Relativement adaptable à d'autres tâches de normalisation

Valorisation :

- Distribution sur GitHub, intégrée à la suite AlvisNLP/ML
- En cours d'implémentation pour des projets applicatifs (Florilège)

Merci pour votre attention

Données :

Description :

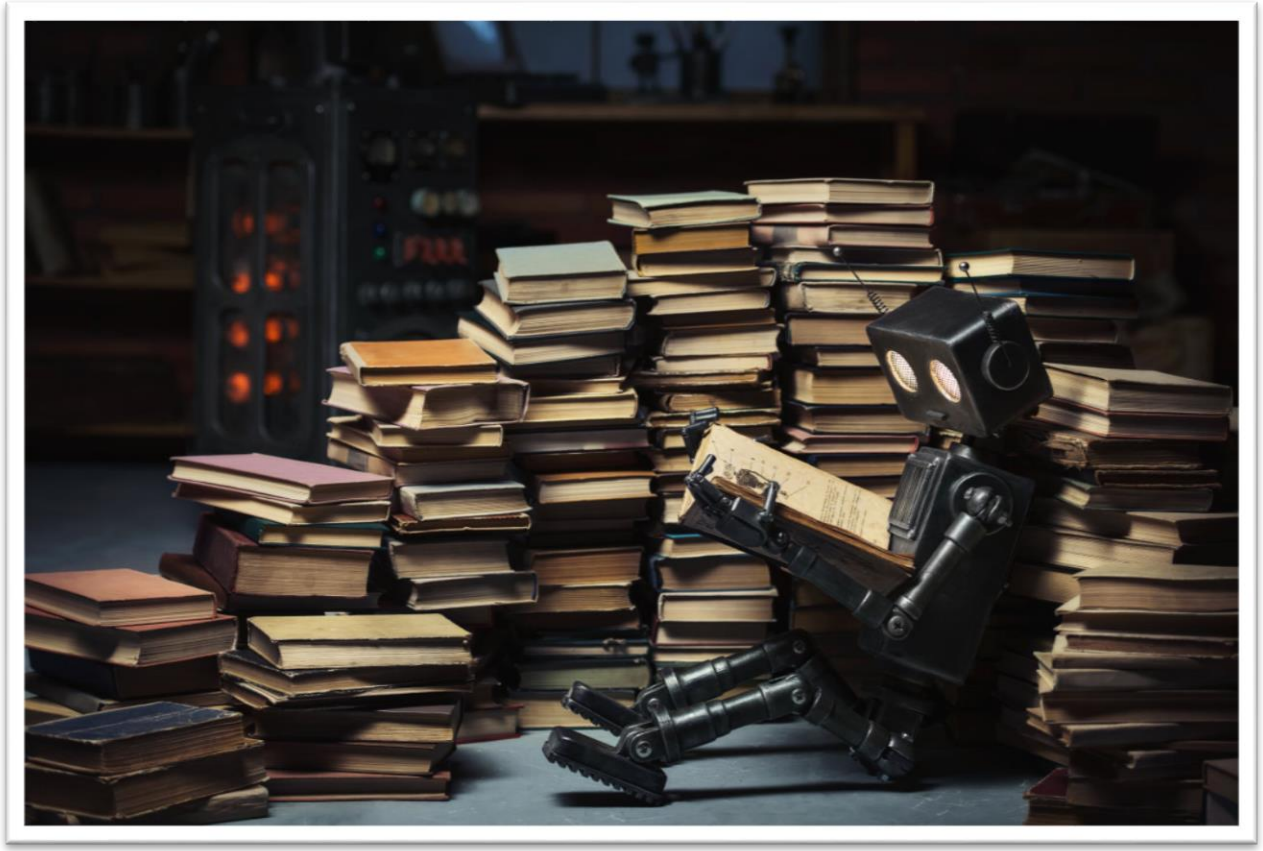
- Challenge Bacteria Biotope de BioNLP Shared Task 2016
- Objectif : Normalisation de mentions d'habitat bactérien
- Données : Sur des titres + résumés d'articles scientifiques en biologie

3 Corpus :

- Corpus d'entraînement : termes d'habitat bactérien + concepts associés (< 1500 exemples annotés)
- Corpus élargi non-annoté du domaine biomédical de 150 000 000 mots (source PubMed)
- Corpus de test : expressions d'habitat bactérien non-annotés (< 800 termes)

Ontologie : Ontobiotope (2320 concepts d'habitats bactériens)

Objectif du TDM : La compréhension par les machines des langues naturelles



Intelligence Artificielle (IA) : Ensemble de théories et de techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence.

Pourquoi est-ce difficile de comprendre que c’est difficile ?!

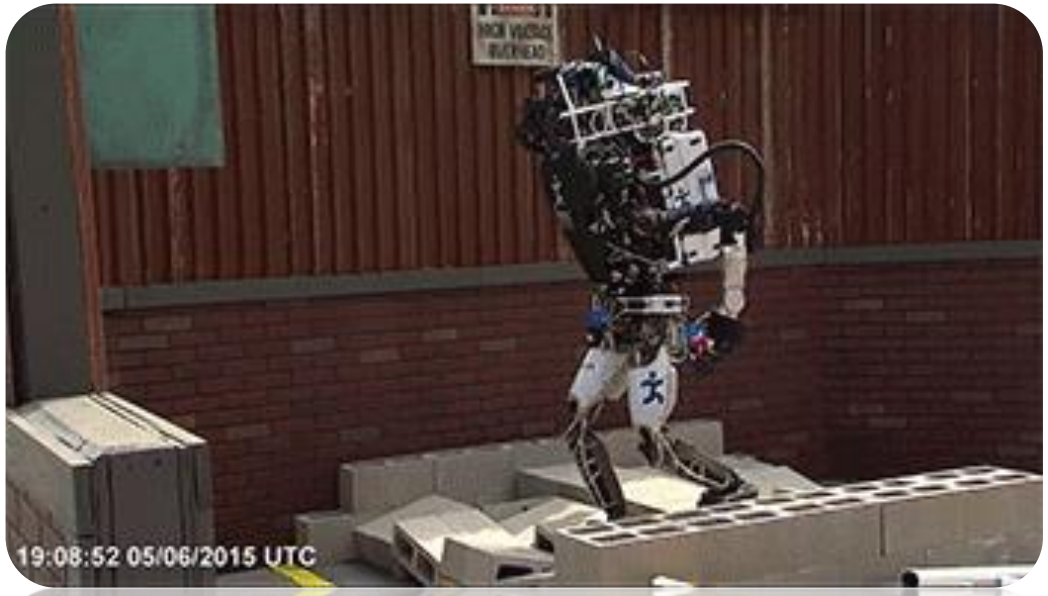
Nous ne comprenons pas consciemment le langage humain nous-mêmes, tout en l’utilisant instinctivement quotidiennement.

Autre exemple : Est-ce que marcher, c’est difficile ?

Pourquoi est-ce difficile de comprendre que c'est difficile ?!

Nous ne comprenons pas consciemment le langage humain nous-mêmes, tout en l'utilisant instinctivement quotidiennement.

Autre exemple : Est-ce que marcher, c'est difficile ?

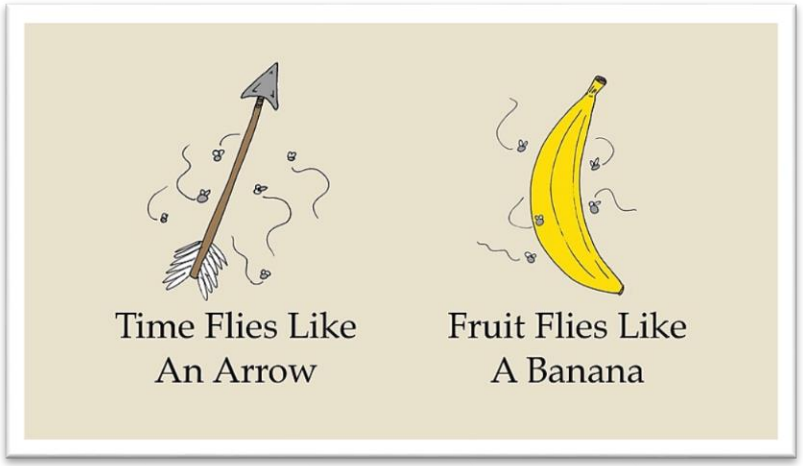


Source : 2015 DARPA Robotics Challenge

Pourquoi est-ce difficile ?

Quelques exemples d’ambiguïtés :

Amphibologies :



*“I shot a **photo**” / “I shot a **person**“*

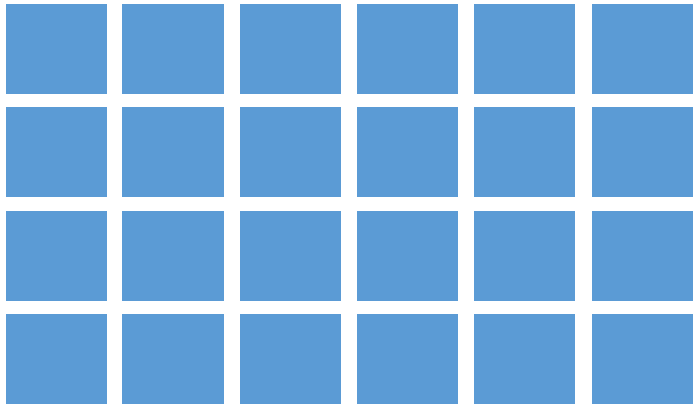
Segmentation en mot :

“Los Angeles”, “rock'n'roll”, “mother-in-law” /

“I'm”=“I”+”am”, “can’t”=“can”+”not”

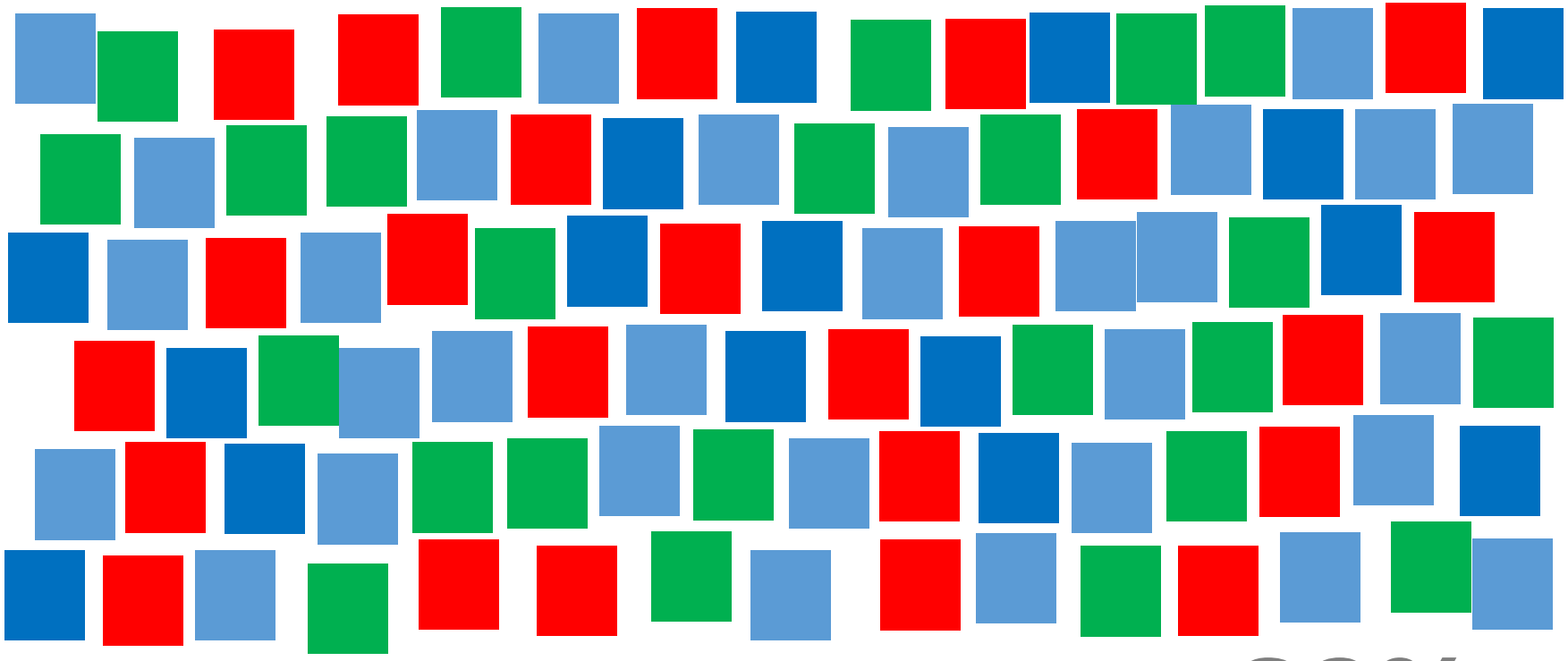
Le texte, une donnée pas comme les autres

Données structurées



Ce que l'on trouve typiquement dans des bases de données ou des ontologies

Données non-structurées



Ce que l'on trouve typiquement ailleurs (**texte**, image, audio, vidéo)

80%

Source : International Data Corporation

Objectif du TDM : extraire l'information

Ressources extérieures

(base de données, ontologie, corpus de textes, modèle algorithmique, lexique, …)

Données non-structurées

Snoopy est un beagle bien connu du monde des bandes dessinées, tout comme son homologue canin, Idefix, le fidèle compagnon d'Obélix.

Données structurées

Animal	
Name	**Specie**
Snoopy	chien
Idefix	chien

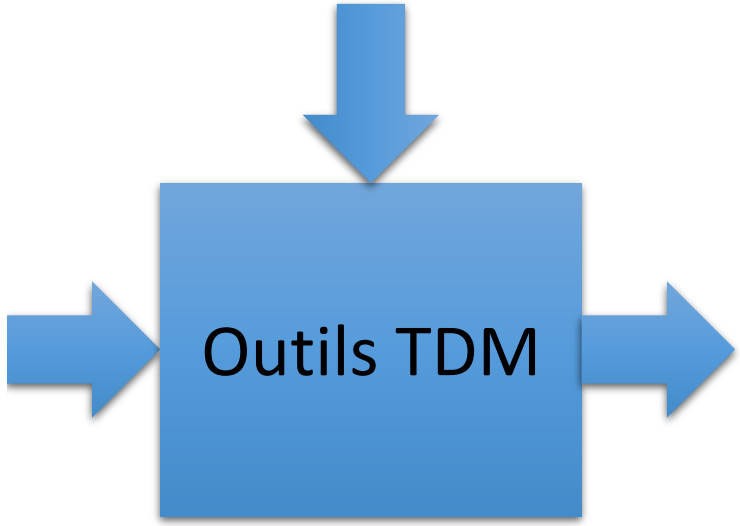
Objectif du TDM : extraire l'information

Données non-structurées

M. Agassizii and M. testudineum are present in Georgia populations of gopher tortoises.

Ressources extérieures

(base de données, ontologie, corpus de textes, modèle algorithmique, lexique, …)

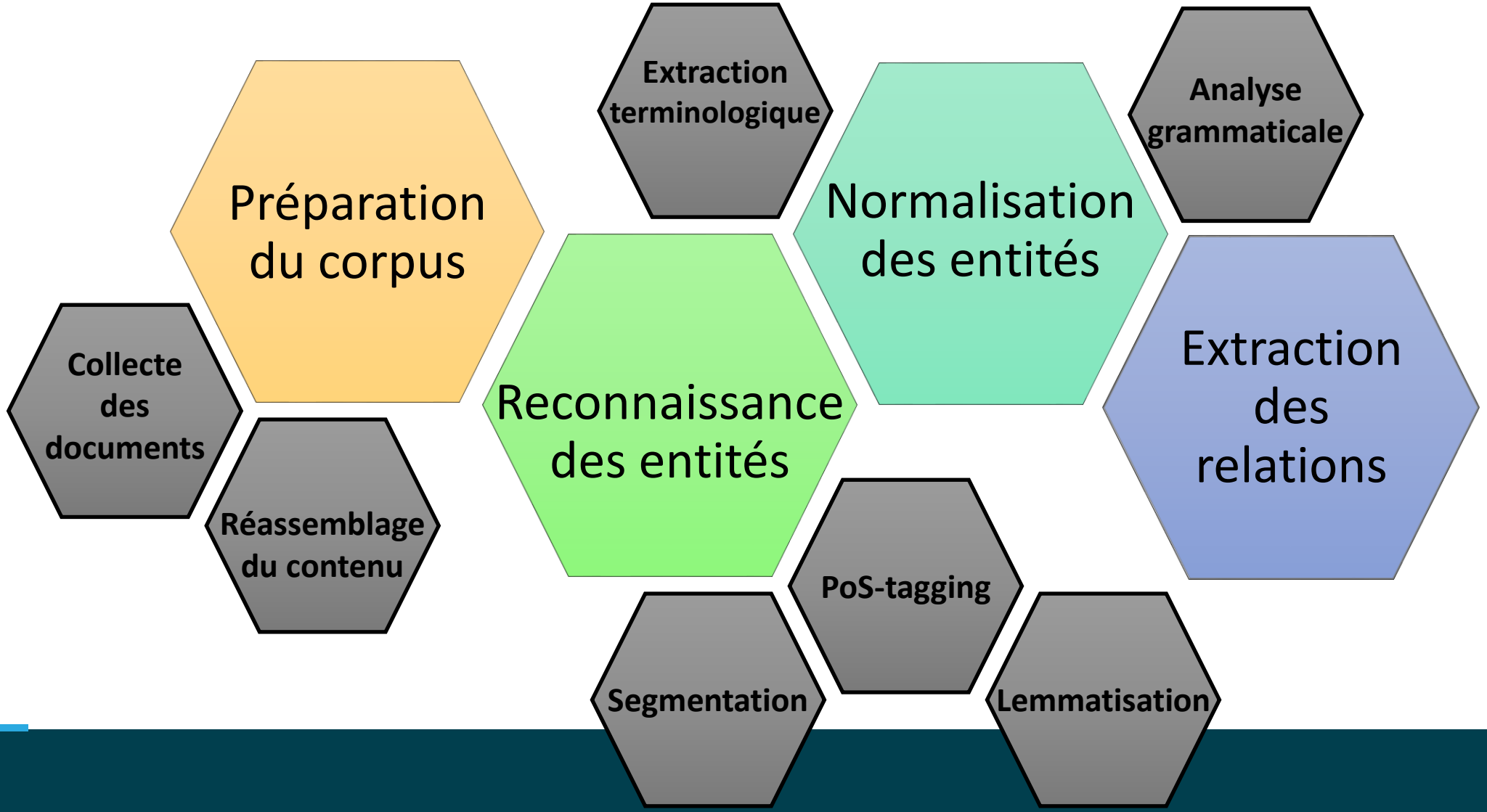


Données structurées

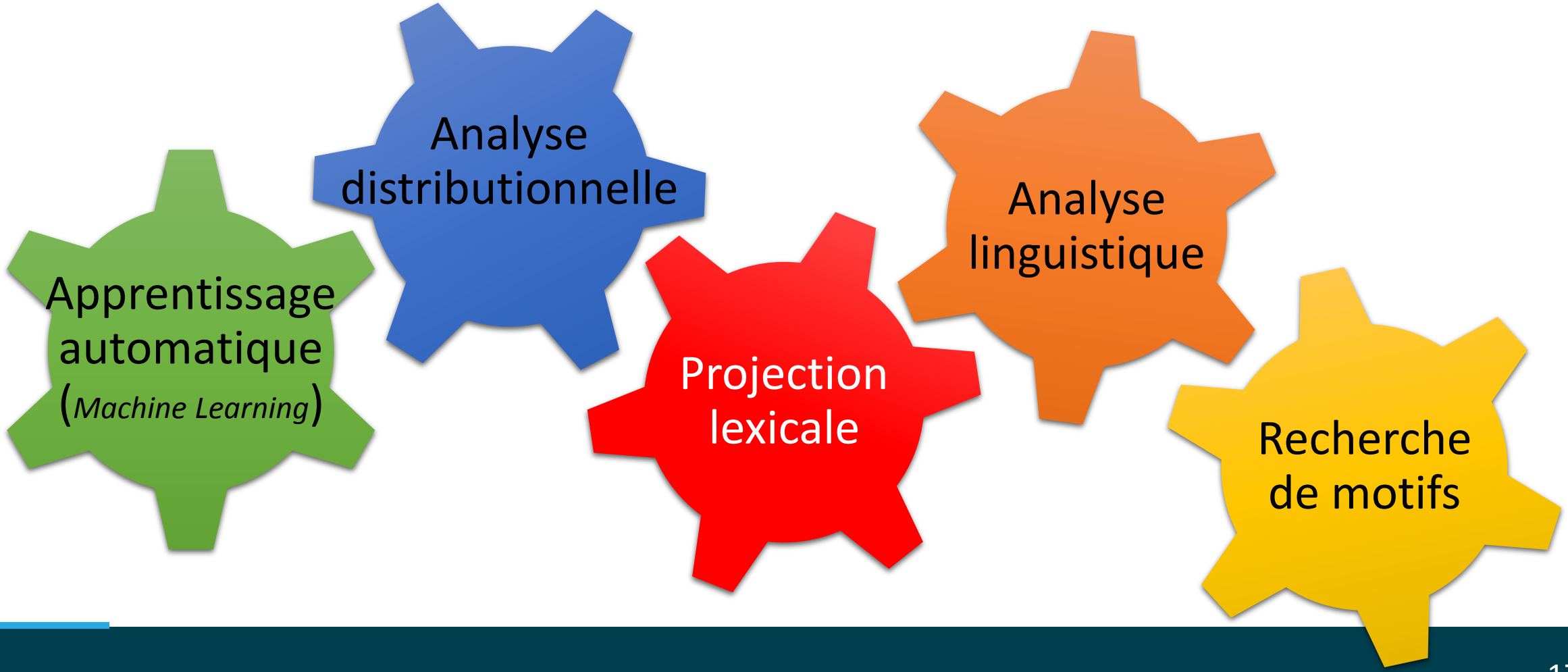
Lives_in

Entité	Type d'entité	Entité normalisée	Habitat
'M. Agassizii'	Bactérie	NCBI 33922	Georgia (GE), OBT 001351 Tortoise
'M. testudineum'	Bactérie	NCBI 244584	Georgia (GE), OBT 001351 Tortoise

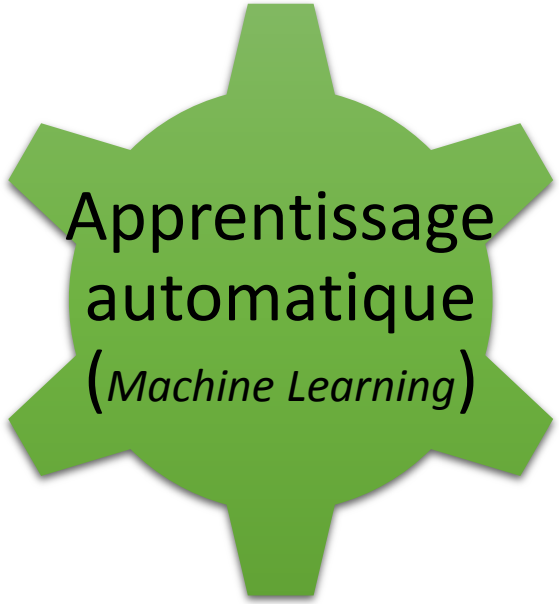
De multiples traitements intermédiaires



Les méthodes mobilisées



Intuition : Apprentissage automatique supervisé



2 étapes :
1) L'apprentissage
2) La prédiction

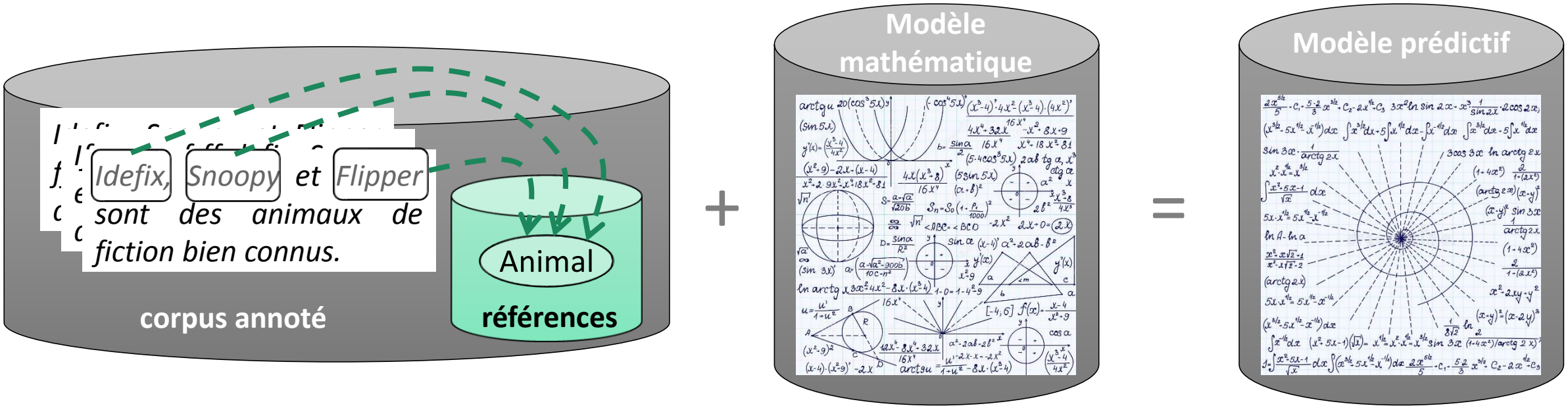
2 ressources :

- Le modèle mathématique 

- Un corpus avec données structurés 

Intuition : Apprentissage automatique supervisé

1) L'étape d'apprentissage :



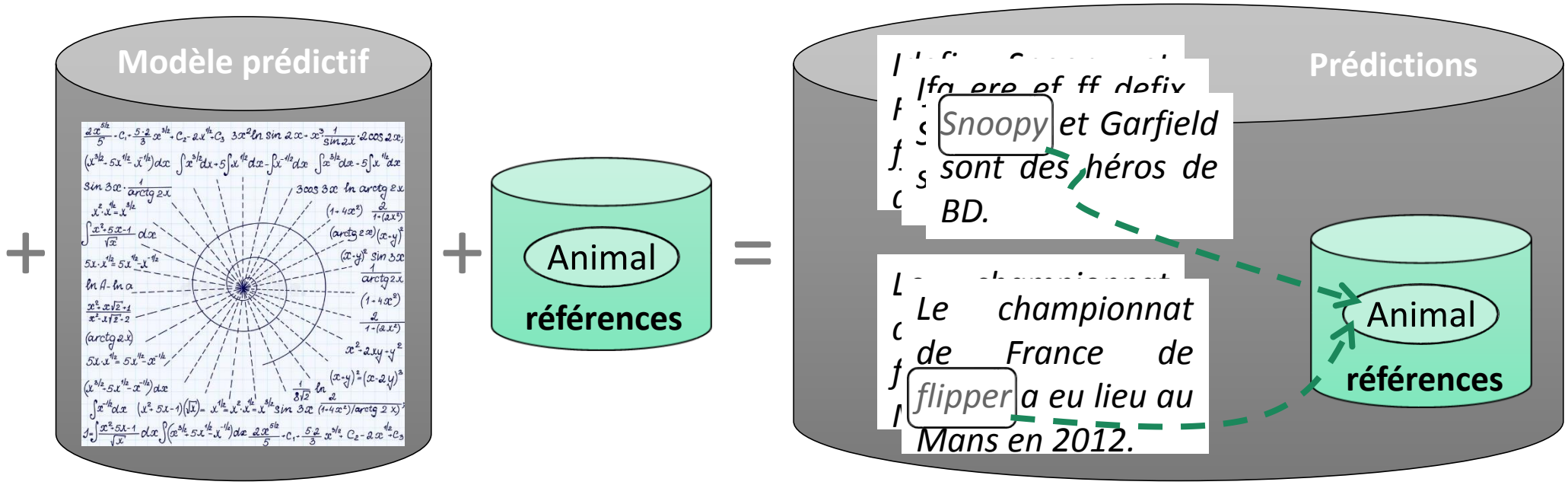
Intuition : Apprentissage automatique supervisé

2) L'étape de prédiction :

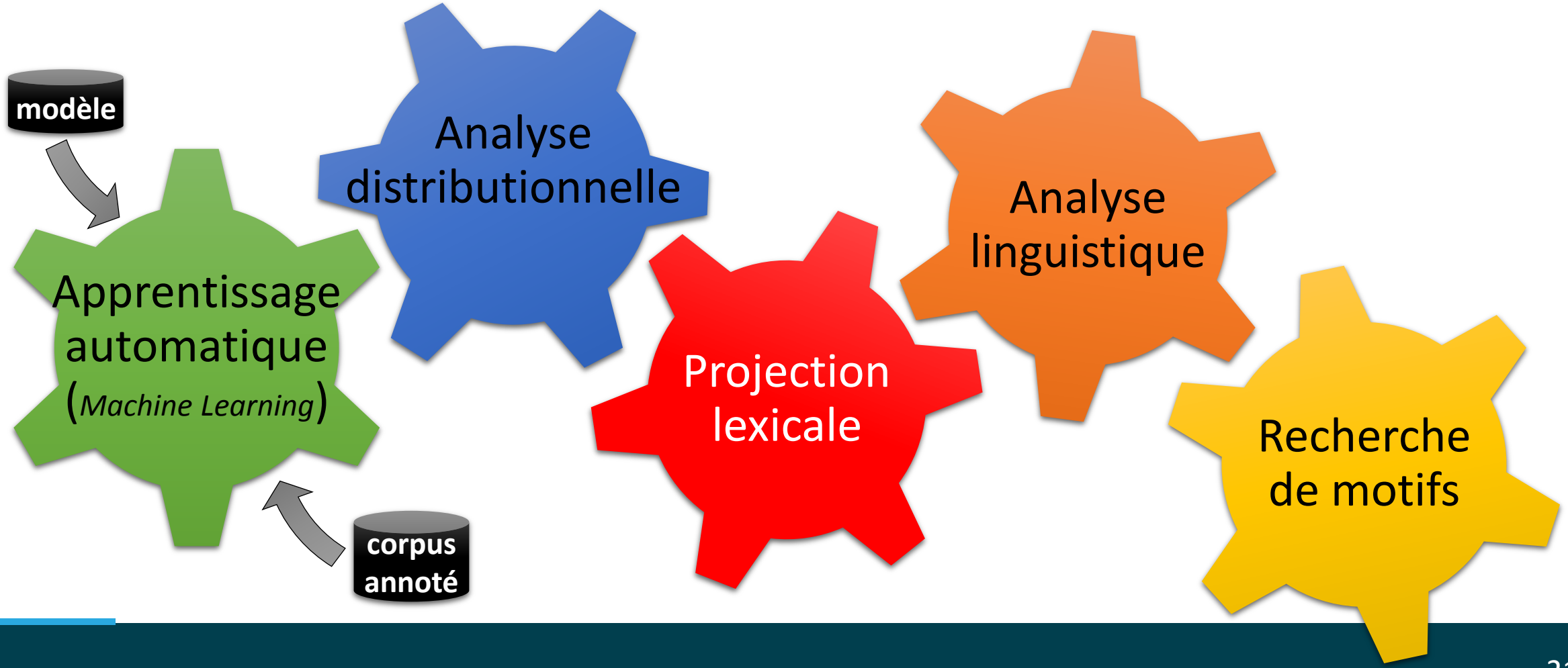
Corpus brut

Snoopy et Garfield sont des héros de BD.

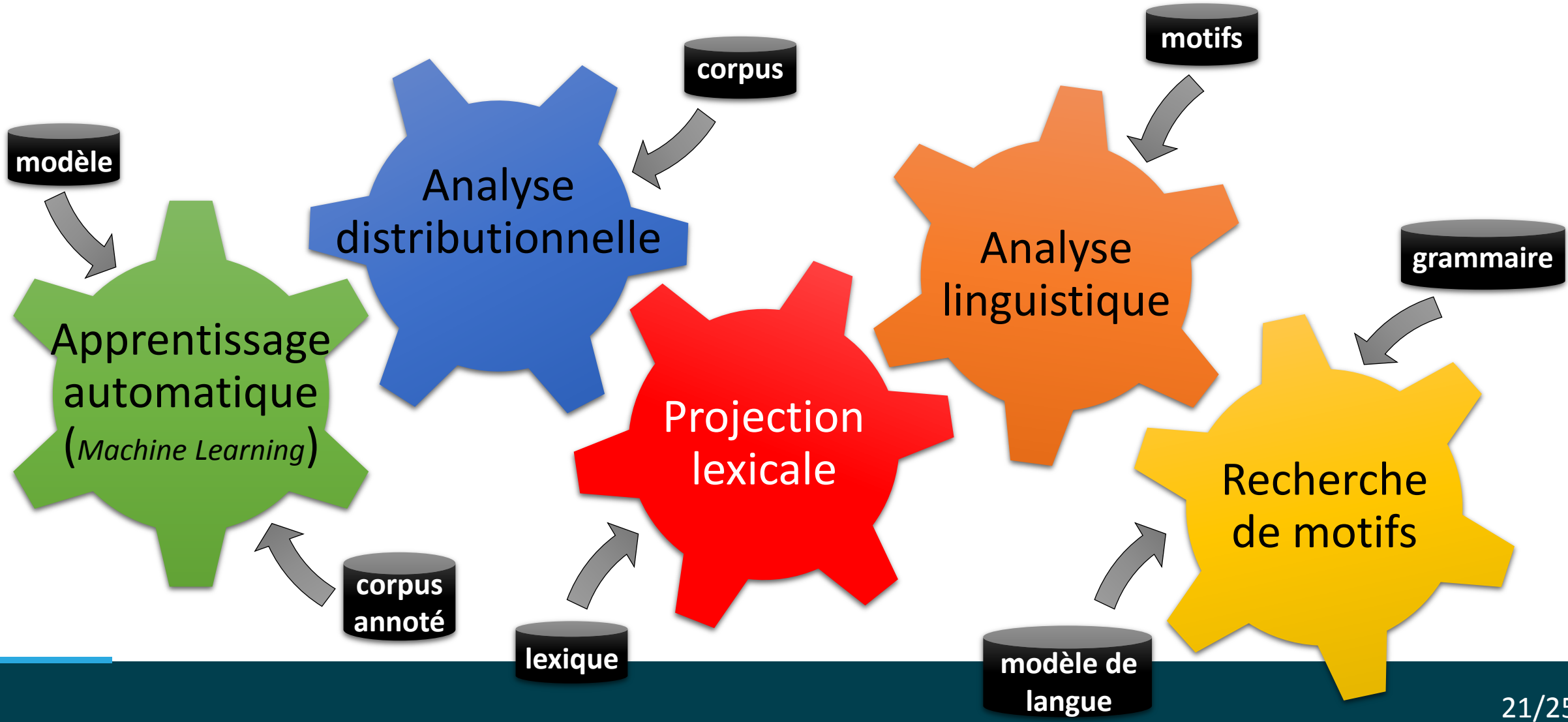
Le championnat de France de flipper a eu lieu au Mans en 2012.



Les ressources mobilisées

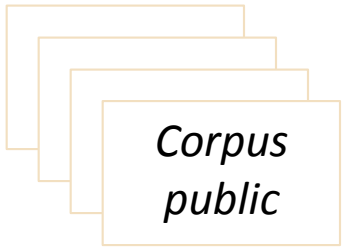


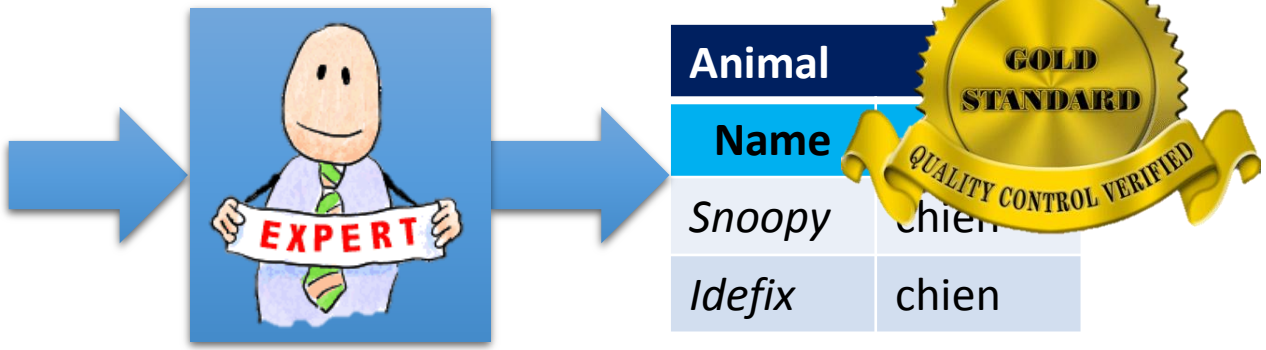
Les ressources mobilisées



Qualité des méthodes

Données non-structurées





Données structurées

Animal	
Name	
Snoopy	chien
Idefix	chien

Qualité des méthodes

Données non-structurées

Données structurées

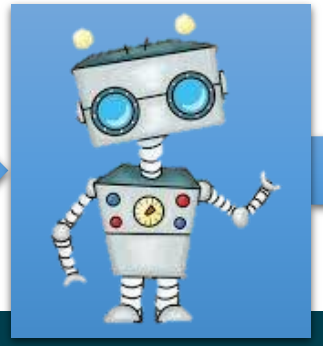
Corpus public

Animal	
Name	
Snoopy	chien
Idefix	chien

Méthode publiée

Prédictions publiées

Corpus public



Animal	
Name	**Specie**
Snoopy	chien
Idefix	chat

Qualité des méthodes

Données non-structurées

Corpus public

Données structurées

Animal	
Name	
Snoopy	chien
Idefix	chien

Méthode publiée

Corpus public

Prédictions publiées

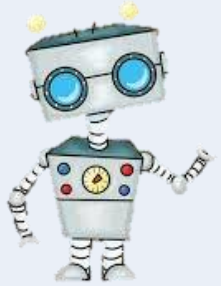
Animal	
Name	**Specie**
Snoopy	chien
Idefix	chat

Evaluation indépendante

(mesures spécifiques : rappel, précision, F-score, taux d'erreur, …)

Qualité des méthodes

Estimation de la performance des méthodes

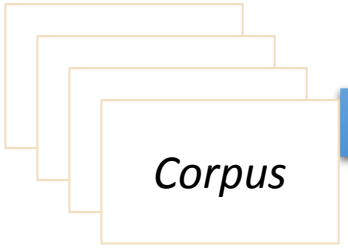
Méthode	Score
	73%
	63%
	61%

Evaluation indépendante

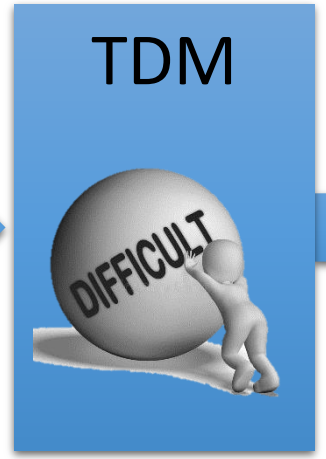
(mesures spécifiques : rappel, précision, F-score, taux d'erreur, …)

Conclusion

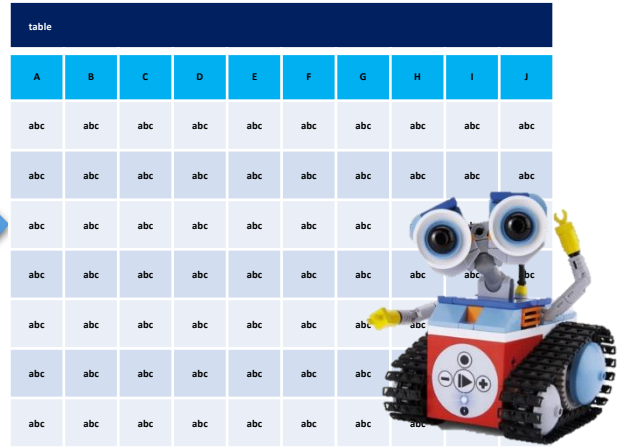
Données non-structurées



Ambigües

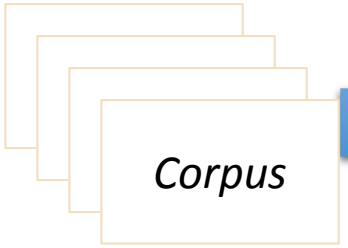


Données structurées

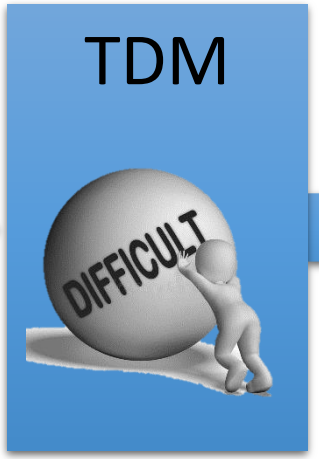


Conclusion

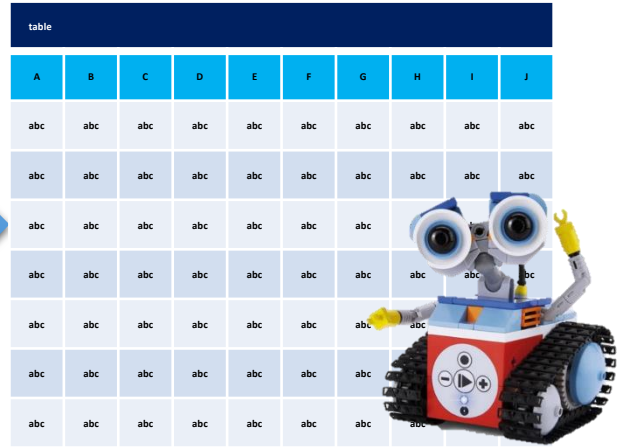
Données non-structurées



Ambigües



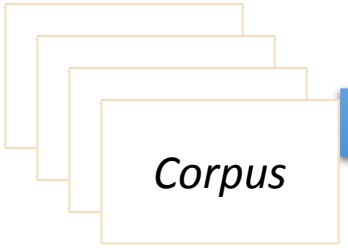
Données structurées



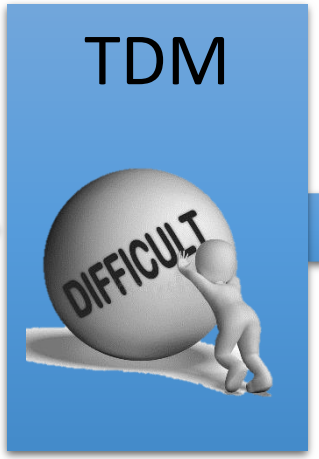
S’appuient sur des méthodes sophistiquées,

Conclusion

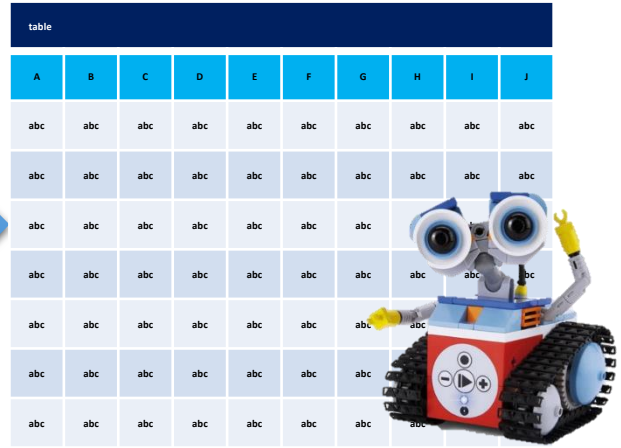
Données non-structurées

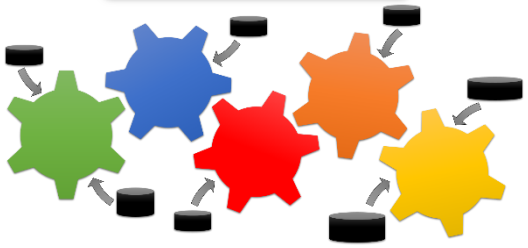


Ambigües



Données structurées





S’appuient sur des méthodes sophistiquées,

Qui nécessitent des ressources très diverses