

# Le contexte sémantique et la dimension temporelle dans l'analyse de données complexes

Marian-Andrei RIZOIU

*Laboratoire ERIC, Lyon*

Avignon, France

13 Mars 2014

# Contexte

## Laboratoire ERIC

## Université Lumière

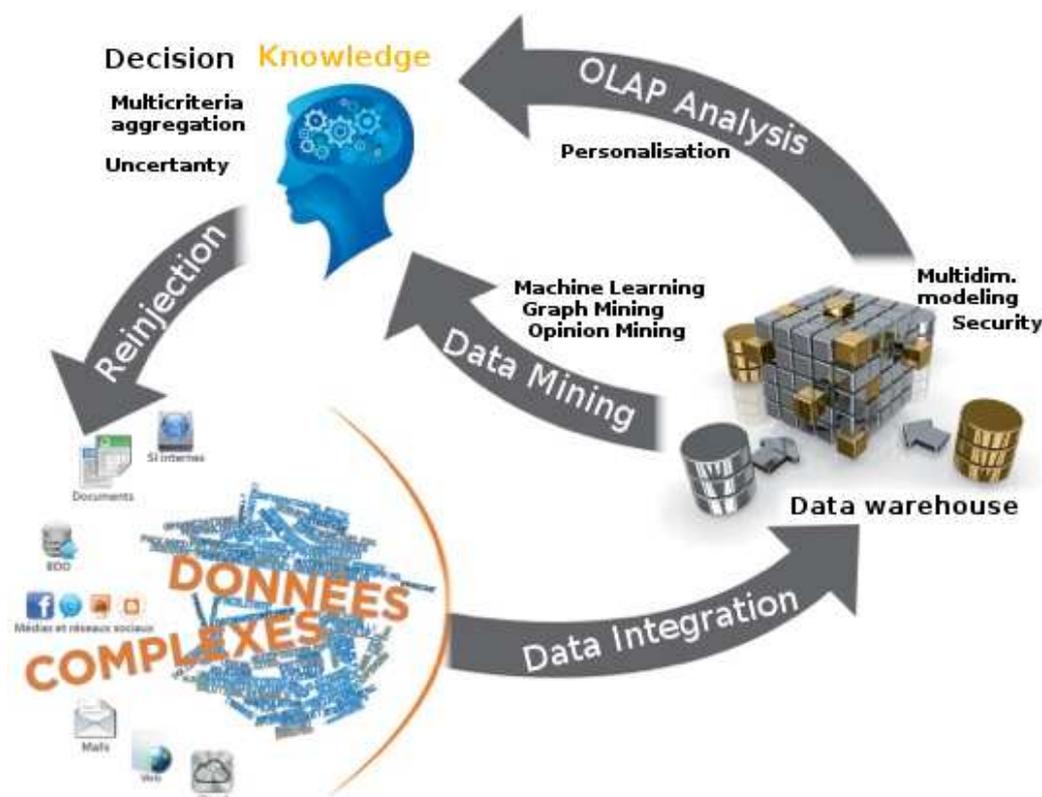
Sciences Humaines et Sociales

*(Sociologie, Psychologie, Linguistique, Histoire, etc.)*



### Projets de recherche multidisciplinaire

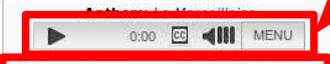
- Identifier des patrons à partir des textes mortuaires (*historiens*)
- Analyser des débats publics : forums en ligne et média traditionnel (*sciences sociales*)
- Discours sur la médecine nucléaire: évolution diachronique et diastratique (*linguistes*)
- Evolution de l'images des politiciens et des entreprises à travers le média social (*sciences politiques*)
- Détecter des rôles sociaux dans des réseaux sociaux enferred à partir des discussions sur des forum en ligne (*Technicolor*)



# Le contexte du travail - les données complexes

**Titre (structure)** → **France**

**Drapeau et emblème (image)** → 

**Hymne (audio)** → 

**Description (texte)** → **France** (English) /fræns or /frɑːns /frɑːns; French: [fʁɑ̃s]   ), officially the **French Republic** (French: *République française* French pronunciation: [ʁɛpyblik fʁɑ̃sɛz]), is a unitary semi-presidential republic located mostly in Western Europe,<sup>[note 12]</sup> with several overseas regions and territories. Metropolitan France extends from the Mediterranean Sea to the English Channel and the North Sea, and from the Rhine to the Atlantic Ocean. From its shape, it is often referred to in French as *l'Hexagone* ("The Hexagon").

France is the largest country in Western Europe and the third-largest in Europe as a whole. It possesses the second-largest exclusive economic zone in the world. France has been a major power with strong cultural, economic, military, and political influence in Europe and around the world.<sup>[6]</sup> France has its main ideals expressed in the 18th-century *Declaration of the Rights of Man and of the Citizen*. From the 17th to the early 20th century, France built the second-largest colonial empire of the time, ruling large portions of first North America and India and then Northwest and Central Africa; Madagascar; Indochina and southeast China; and many Caribbean and Pacific Islands.

France is a developed country,<sup>[7]</sup> possessing the world's fifth-largest and Europe's second-largest economy by nominal GDP. It is also the world's ninth-largest by GDP at purchasing power parity.<sup>[8]</sup> France is the wealthiest nation in Europe – and the fourth-wealthiest in the world – in aggregate household wealth.<sup>[9]</sup> French citizens enjoy a high standard of living, high public education level, and one of the world's longest life expectancies.<sup>[10]</sup> France has been listed as the world's "best overall health care" provider by the World Health Organization.<sup>[11]</sup> It is the most-visited country in the world, receiving 79.5 million foreign tourists annually.<sup>[12]</sup>

France has the world's fifth-largest nominal military budget,<sup>[13]</sup> as well as (in terms of personnel) the largest military in the EU,<sup>[citation needed]</sup> the third-largest deployable force in NATO, and the 26th-largest military in the world. France also possesses the third-largest stockpile of nuclear weapons in the world<sup>[14]</sup> – with around 300 active warheads as of 25 May 2010 – and the world's second-largest diplomatic corps behind the United States.<sup>[15]</sup> France is a founding member of the United Nations, one of the five permanent members of the UN Security Council, and a member of the Francophonie, the G8, G20, NATO, OECD, WTO, and the Latin Union. It is also a founding and leading member state of the European Union and the largest EU state by area.<sup>[16]</sup> In 2013, France was listed 20th on the Human Development Index and, in 2010, 24th on the Corruption Perceptions Index.

**Hyperlien (sources externes des connaissances)** → [World's second-largest diplomatic corps](#)

**Carte (image)** → 

Area	
- Total <sup>[note 2]</sup>	674,843 km <sup>2</sup> (41st) 260,558 sq mi
- Metropolitan France	
- IGN <sup>[note 3]</sup>	551,695 km <sup>2</sup> (47th) 213,010 sq mi
- Cadastre <sup>[note 4]</sup>	543,965 km <sup>2</sup> (47th) 210,026 sq mi
Population (2012)	
- Total <sup>[note 2]</sup>	65,350,000 <sup>[2]</sup> (19th)
- Metropolitan France	63,460,000 <sup>[1]</sup> (22nd)
- Density <sup>[note 5]</sup>	116/km <sup>2</sup> (89th) 301/sq mi

Indicateurs (format numérique)

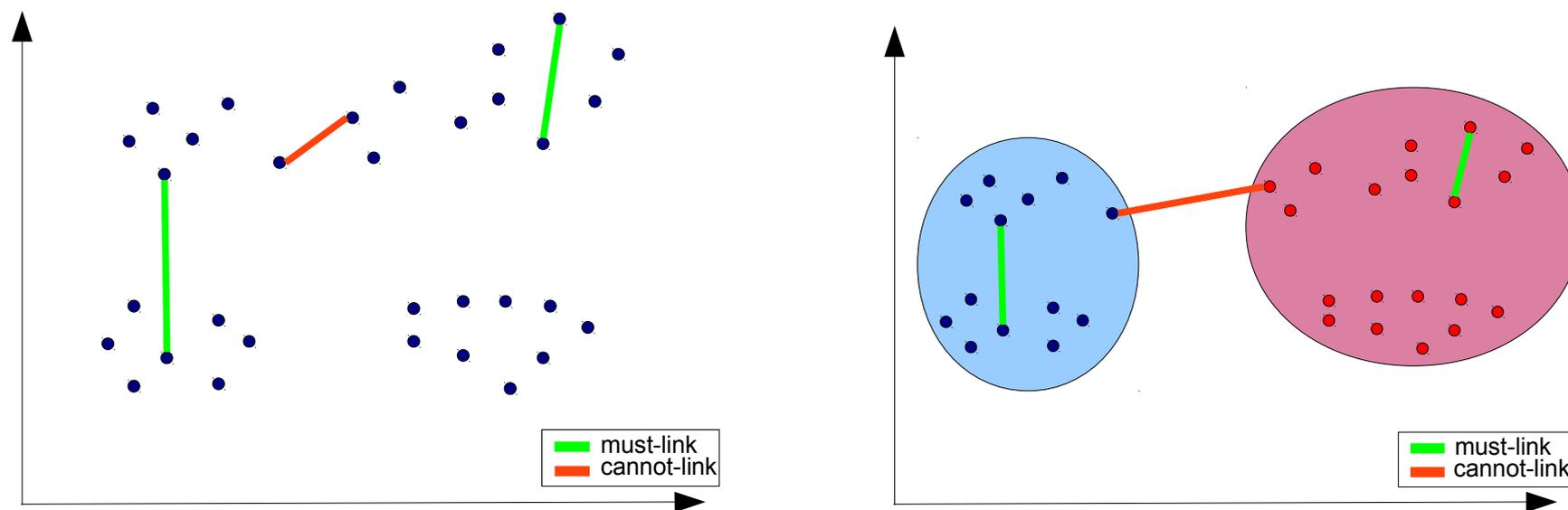
## Spécificités :

- Différentes types des données
- Information additionnelle
- Dimension temporelle/dynamique
- Grande dimensionnalité
- Sources diverses et distribuées

# Notre approche

- Objectif général :**
- extraire des connaissances à partir de données complexes, souvent dans un contexte non-supervisé ;  
(*e.g., construction des profils utilisateurs, campagne marketing*)
  - ajouter de la sémantique dans l'analyse de données ;  
(*e.g., améliorer l'interprétabilité des résultats*)
  - utiliser les connaissances (supervision) disponibles.

## Intégrer les connaissances : clustering semi-supervisé [WAG00]



## Enjeux de recherche

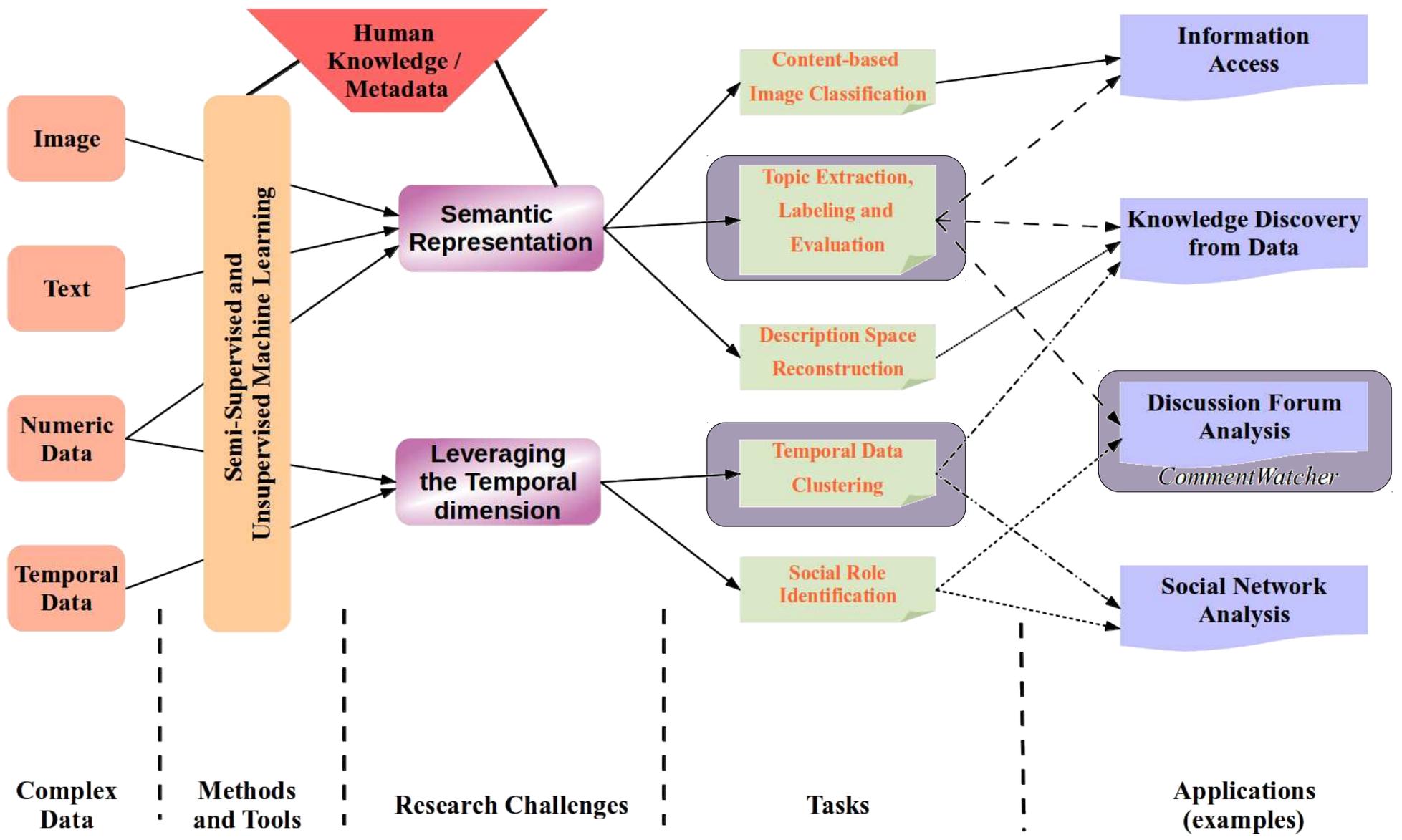
Utiliser la sémantique pour analyser les données complexes.

- plonger les données dans un espace de représentation capable de capturer la sémantique sous-jacente aux données
- injecter des connaissances externes dans les algorithmes d'apprentissage automatique

Prendre en compte la dimension temporelle des données complexes.

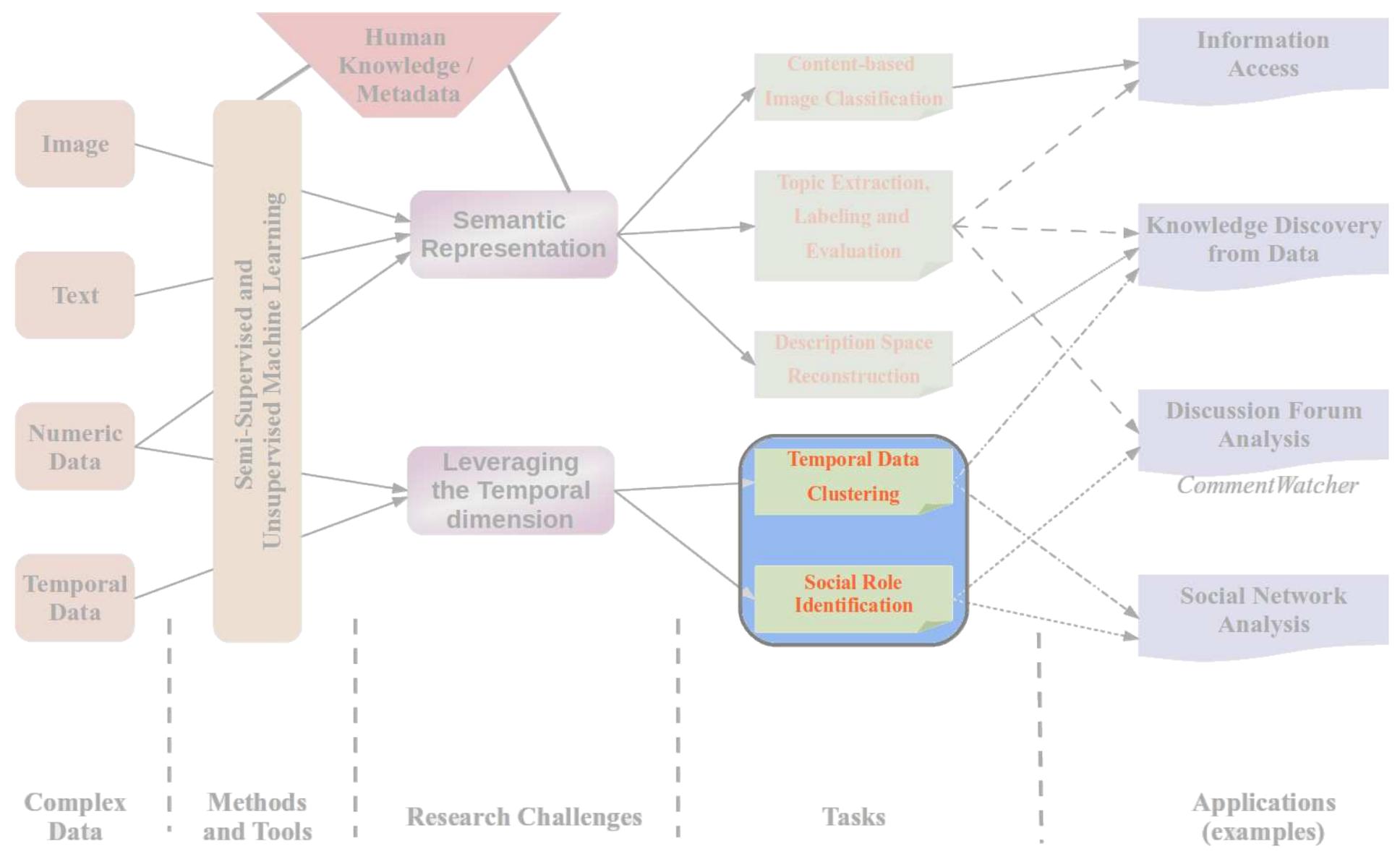
- Tâches spécifiques :
- *détection des évolutions typiques*
  - *reconstruction sémantique de l'espace de représentation des données*
  - *intégration des informations externes dans la description numérique des images.*

# Schéma des travaux



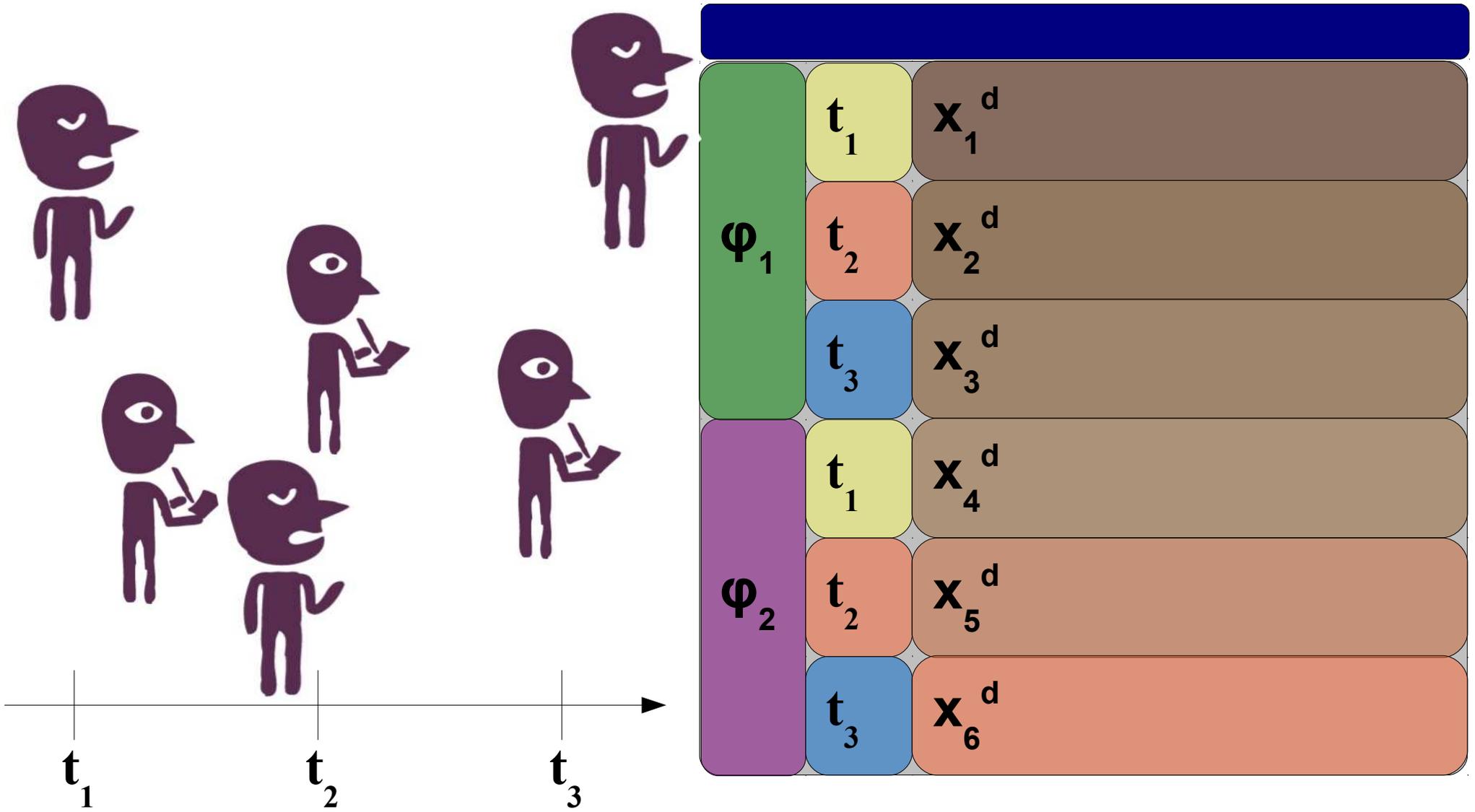
# Détection des évolutions typiques :

Prendre en compte l'effet temporel



# Jeux de données:

Les valeurs des attributs descriptifs ( $x^d$ ) pour plusieurs entités ( $\varphi$ ) sont enregistrées à plusieurs dates ( $t$ )



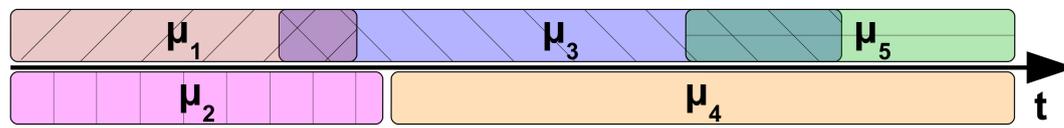
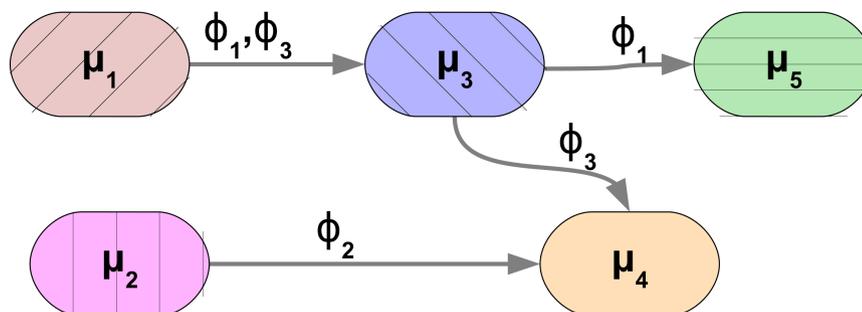
**Enjeu de recherche :**

Utiliser la dimension temporelle dans l'analyse de données complexes

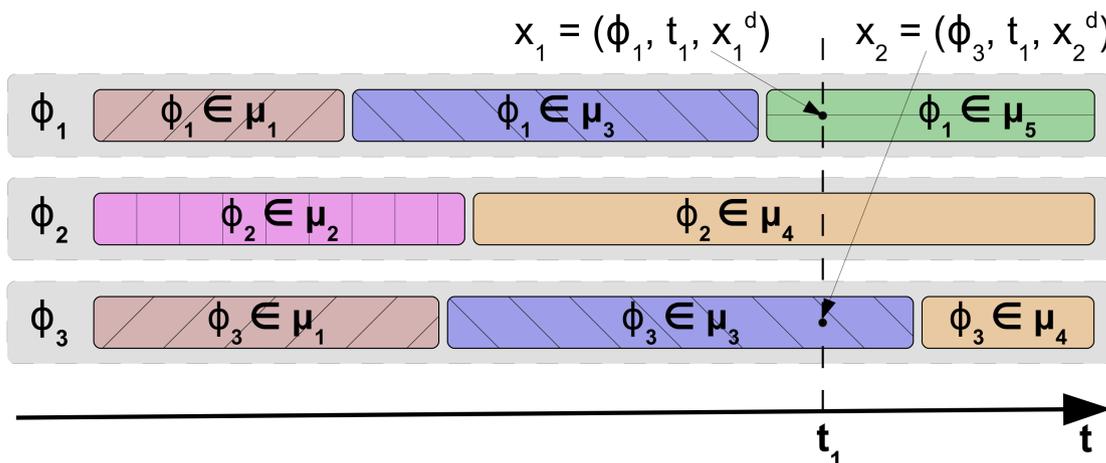
**Tâche d'apprentissage :**

Détecter les motifs d'évolutions typiques pour les entités du jeu de données

a) phases parmi lesquelles passent les entités durant le temps



b) trajectoires des entités parmi les phases



**Enjeu de recherche :** Utiliser la dimension temporelle dans l'analyse de données complexes

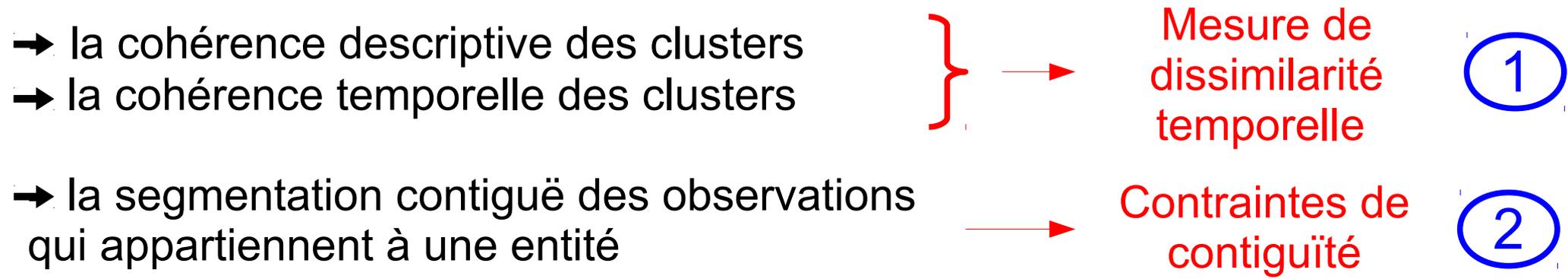
**Tâche d'apprentissage :** Détecter les motifs d'évolutions typiques pour les entités du jeu de données

## Difficultés :

- Les phases d'évolution ou les critères qui définissent les phases ne sont pas connus à l'avance
- Les phases doivent regrouper des observations similaires du point de vue descriptif et temporel
- Comment modéliser le temps dans l'algorithme de découverte des phases ?
- Les algorithmes de clustering temporel typiques [KIS10] généralement traitent des séries temporelles entières, nous travaillons au niveau d'observations.

**Solution proposée (1)**      Un algorithme de clustering temporel avec contraintes, les clusters obtenues servent comme phases d'évolution.

La partition obtenue doit assurer :



Algorithme inspiré des K-Means. La fonction objective à minimiser est :

$$J = \sum_{\mu_j \in M} \sum_{x_i \in C_j} \left( \underbrace{\|x_i - \mu_j\|_{TA}}_{(1)} + \sum_{(x_k \notin C_j) \wedge (x_k^\varphi = x_i^\varphi)} \underbrace{w(x_i, x_k)}_{(2)} \right)$$

## Solution proposée (2)

### Les contributions

Mesure de dissimilarité temporelle



Distance dans l'espace descriptif et temporel

Segmentation contiguë

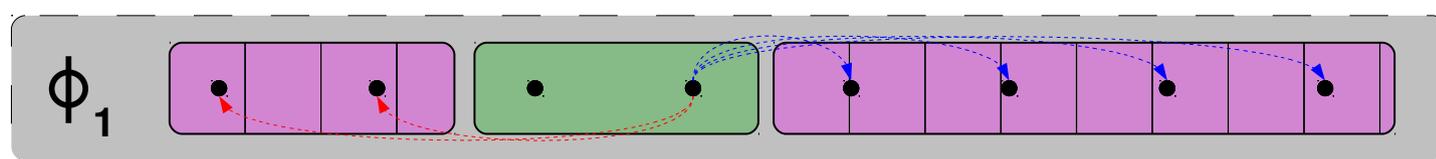


Contraintes semi-supervisées MUST-LINK



Fonction de pénalisation dépendante du temps

## Fonction de pénalisation



## Solution proposée (2)

## Les contributions

Mesure de dissimilarité temporelle



Distance dans l'espace descriptif et temporel

Segmentation contiguë



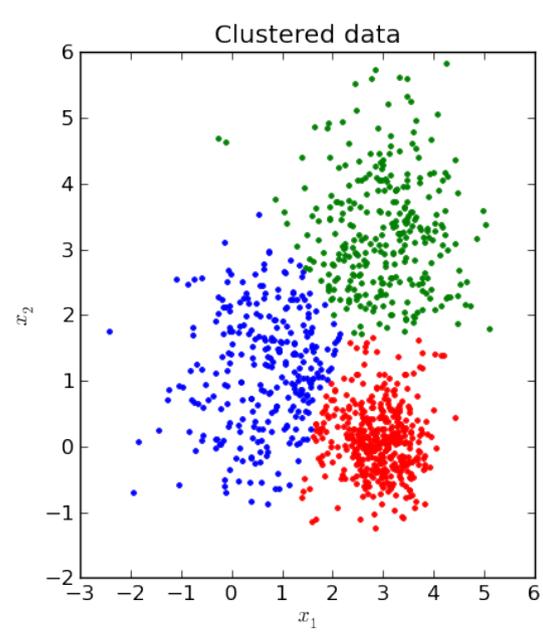
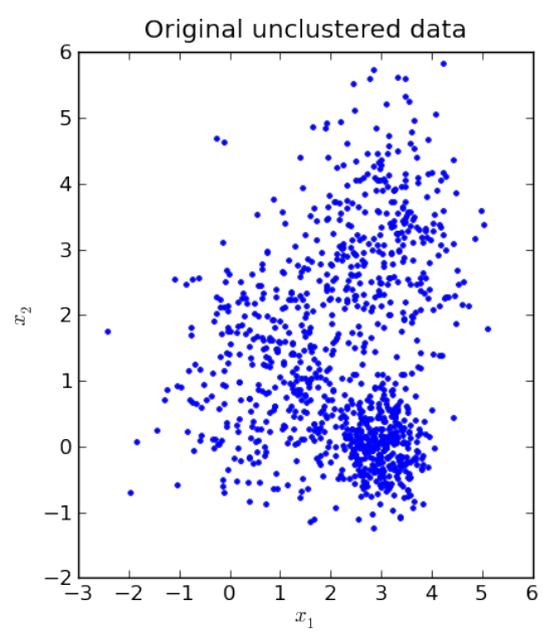
Contraintes semi-supervisées MUST-LINK



Fonction de pénalisation dépendante du temps

## L'algorithme TDCK-Means

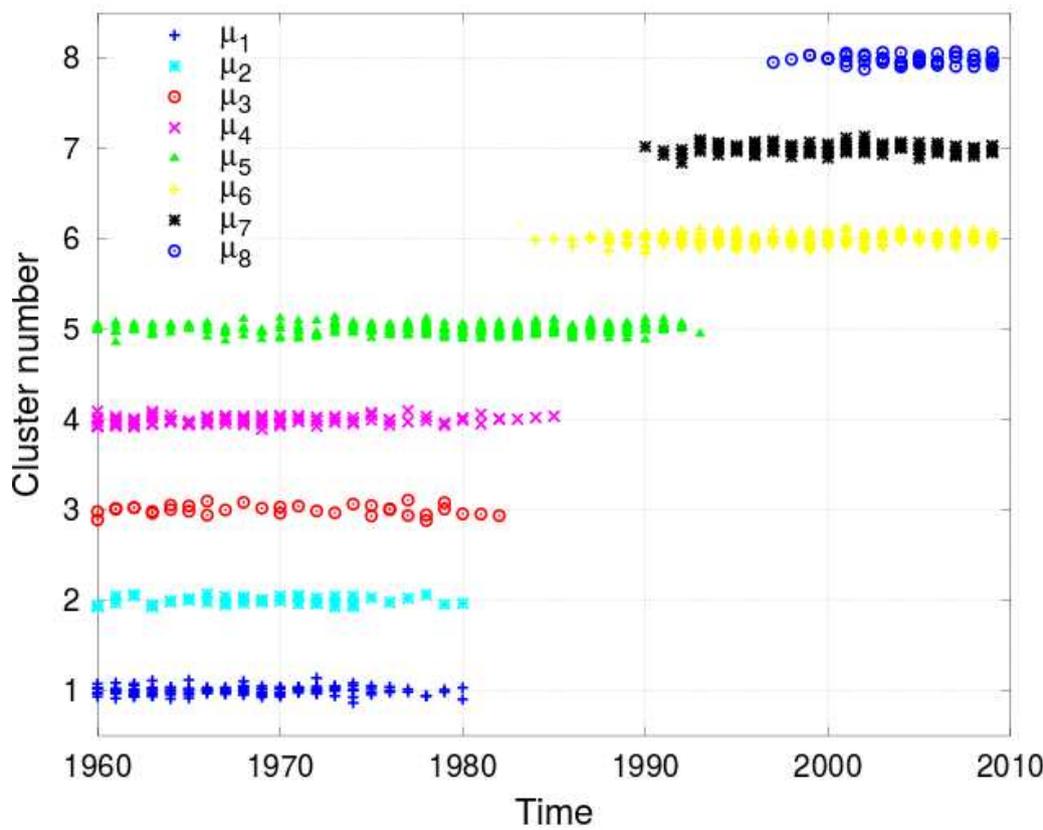
- ✓ Inspiré des K-Means.
- ✓ Mesure de *dissimilarité temporelle* et la *fonction de pénalisation*



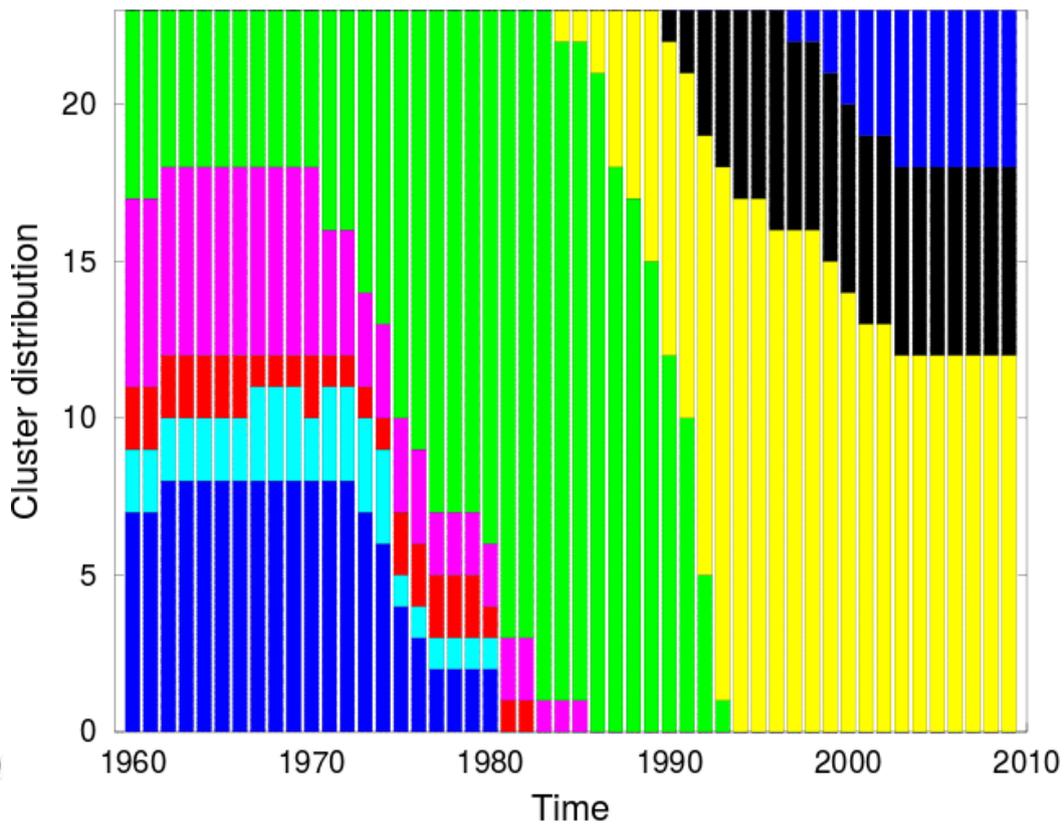
# Expérimentations et résultats

**Compared Political Dataset I [ARM11] :**  
23 pays, 60 années, 207 variables politiques, démographiques, sociales et économiques.

Observations in clusters over time

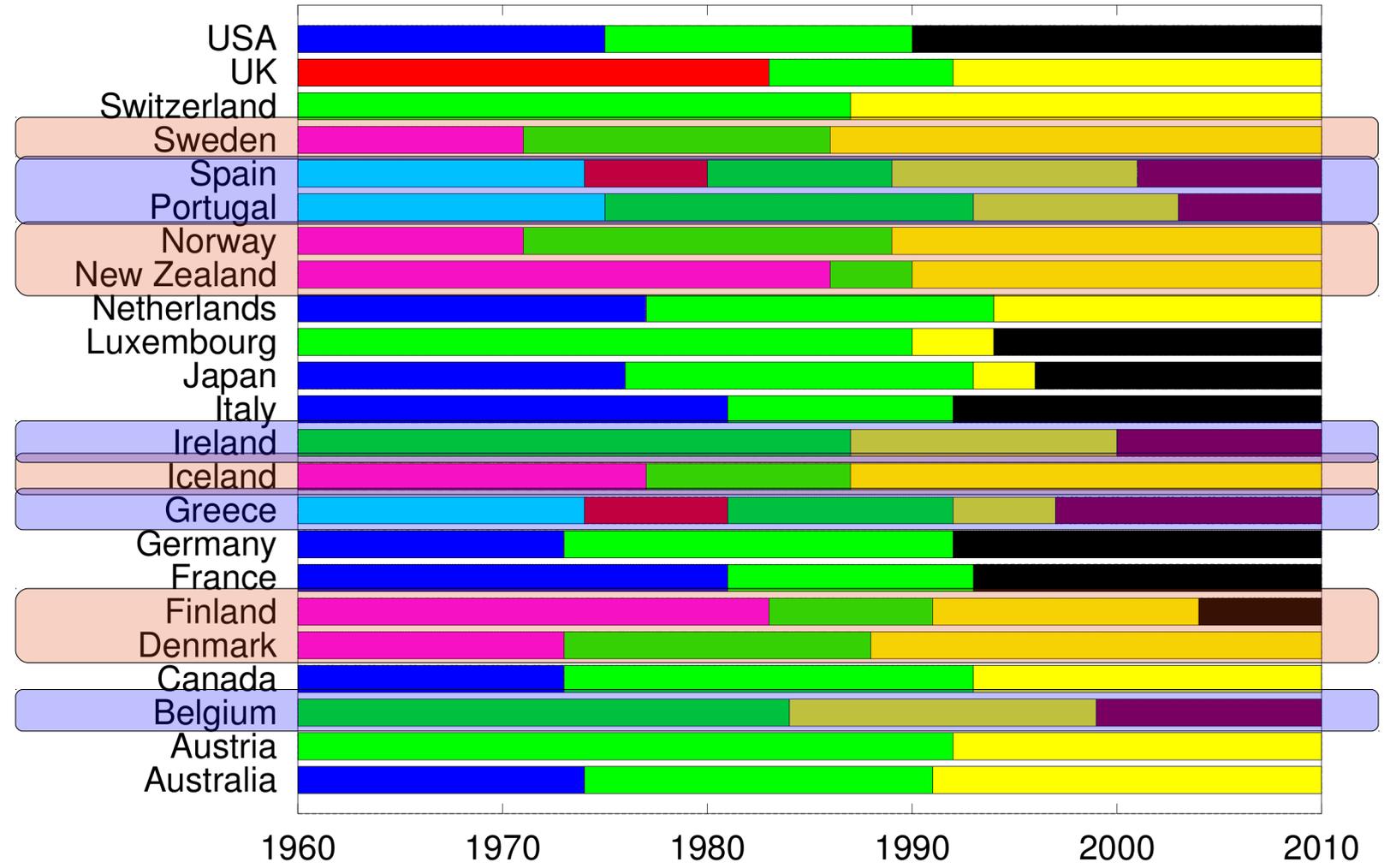


Cluster distribution over time



# Expérimentations et résultats

**Compared Political Dataset I [ARM11] :**  
23 pays, 60 années, 207 variables politiques, démographiques, sociales et économiques.



# Expérimentations et résultats

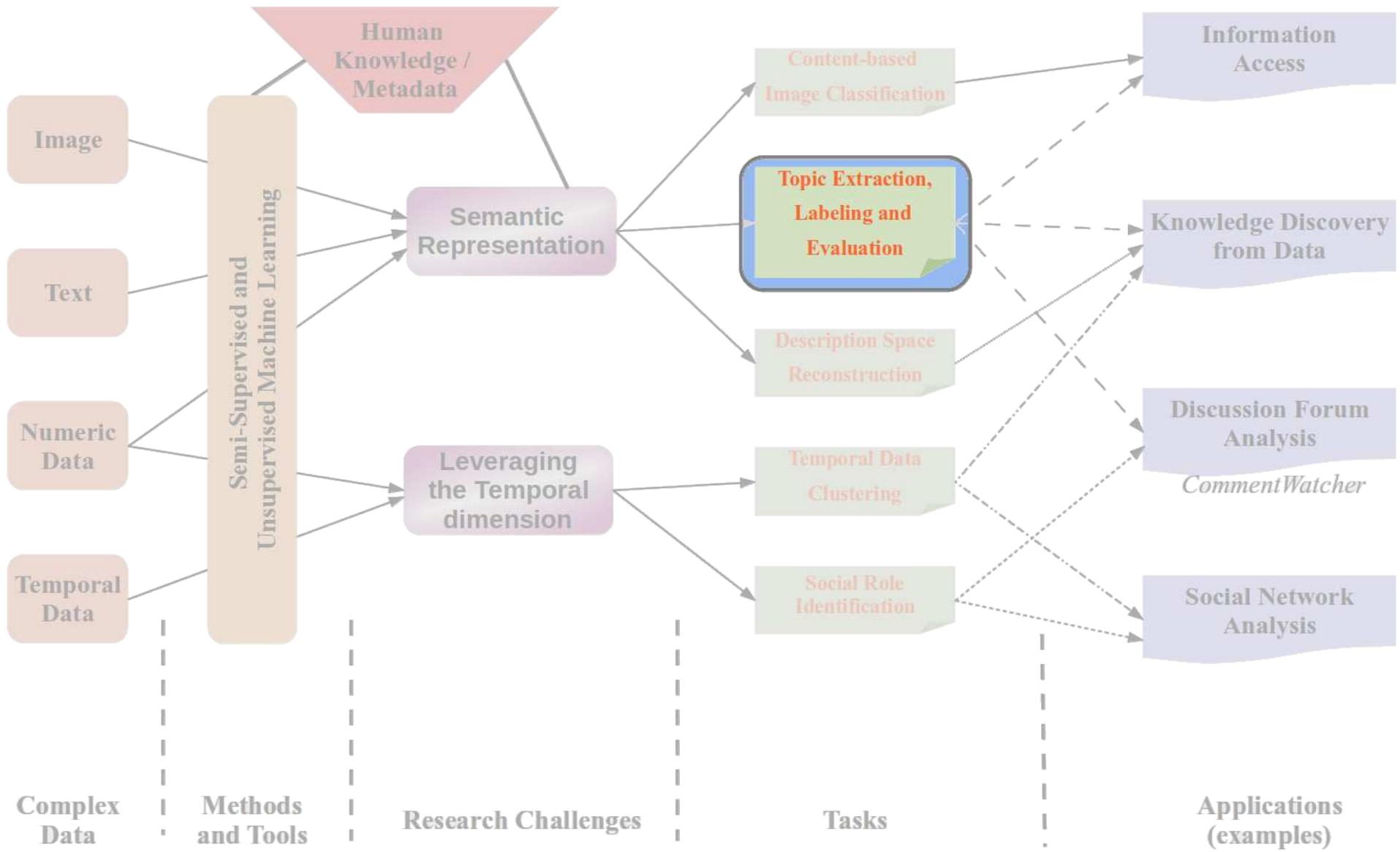
**Compared Political Dataset I** [ARM11] :  
23 pays, 60 années, 207 variables politiques, démographiques, sociales et économiques.

Un exemple de graphe d'évolutions :  
(*construction postérieure au clustering*)



# Analyse de données textuelles :

Extraction, étiquetage et évaluation des *thématiques*



**Les données :** Une collection de textes en langage naturel, souvent issus de l'internet.

*Projets avec i) un startup, ii) des sociologues, iii) des linguistes*

**Les défis :**

- grands volumes de données ;
- besoin de résumer les « idées » principales : *les thématiques*
- la plupart de la littérature évalue les thématiques à l'aide des mesures statistiques, sans prendre en compte la sémantique

*ex. indice de perplexité [WAL09]*

### Nouvelle hausse du prix du tabac en juillet, jusqu'à 7 euros le paquet de cigarettes

Le HuffPost/AFP | Publication: 12/06/2013 09h09 CEST | Mis à jour: 12/06/2013 10h33 CEST



30 4 0

partager tweeter envoyer

SUIVRE: Smoking, Video, Marisol Touraine, Ac Prix, Santé, Santé, Tabac, Actualités

SANTÉ - Le prix des paquets de ciga juillet, a déclaré mercredi la ministre intervenir début juillet" et se fera "a p itélé.

L'hypothèse d'une hausse en deux te octobre- est donc abandonnée. Prévu sociale 2012 cette hausse ferait passe et celui des plus vendus à 7 euros.

**SUPER UTILISATEUR DU HUFFPOST**  
dieu  
295 Fans

il y a 19 minutes (11h09)  
Pourquoi taxer un fumeur ? pourquoi ne pas détruire les plantations et éliminer les buralistes ?

**LUMINET**  
17 Fans

il y a 12 minutes (11h16)  
Pourquoi subventionner les producteurs de tabac???

**cpamafaute**  
2 Fans

il y a 39 minutes (10h49)  
Continuons d'appauvrir les Français par des taxes imbéciles qui ne changeront rien aux comportements des fumeurs ! Entre toutes les taxes et les impôts comment allons nous faire ? On ne cesse de nous parler de relance économique et on détruit le pouvoir d'achat des Français ! Vous pensez que lorsque toutes les entreprises seront fermées l'argent rentrera dans les caisses de l'état ????? Cette politique est un desastre pour notre pays.

## Tâches d'apprentissage :

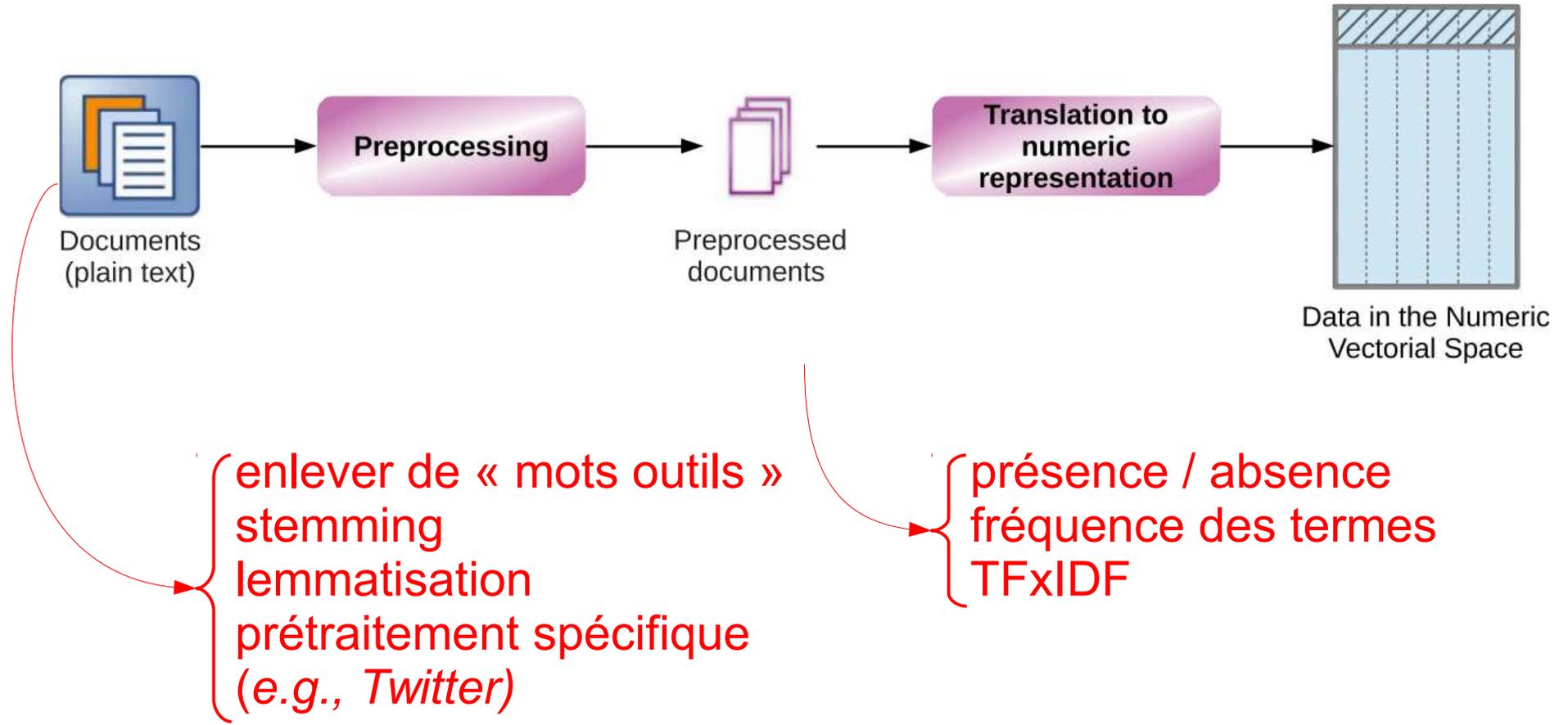
- extraction des thématiques
- étiquetage des thématiques avec des noms compréhensibles pour un être humain
- utilisation des connaissances sémantiques dans l'évaluation de ces thématiques.

## Dimension appliquée :

- Demande forte de la part des scientifiques des **Sciences Humaines et Sociales** (*Sociologie, Psychologie, Linguistics, Histoire, etc.*)
- implémentation dans le logiciel **CommentWatcher** ;

**Solution proposée (1) :** Une alternative aux modèles graphiques (e.g., LDA [BLE03]): le clustering textuel

**Prérequis:** Plonger les textes dans un espace numérique : « sac-de-mots »



## Solution proposée (2) :      Extraction et étiquetage des thématiques

### I. Extraction des thématiques à l'aide du clustering recouvrant

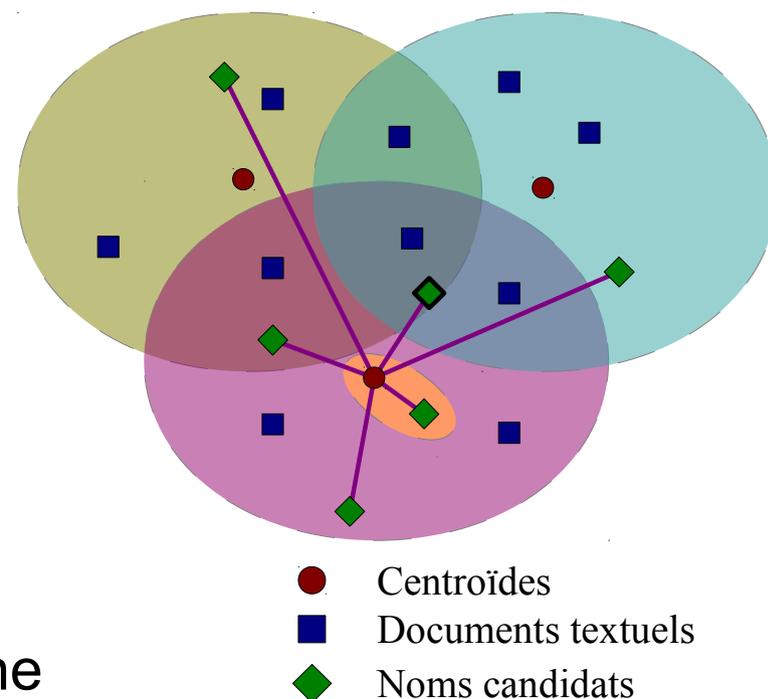
#### OKM [CLE08]

Une extension des KMeans qui autorise un document à appartenir à plusieurs clusters

- construire une partition avec recouvrement
- les centroids sont des abstractions de leur clusters: *les thématiques*

### II. Étiquetage des thématiques

- extraire des expressions complètes fréquentes à partir du texte original  
*tableaux de suffixes* [MAN93].
- injecter les expressions comme des pseudo-documents
- calculer la similarité et choisir le plus proche candidat



**Solution proposée (3) :** Évaluer la cohésion sémantique des thématiques

**Hypothèse sous-jacente :**

Les mesures statistiques ne parviennent pas totalement à émuler le jugement humain [CHA09]

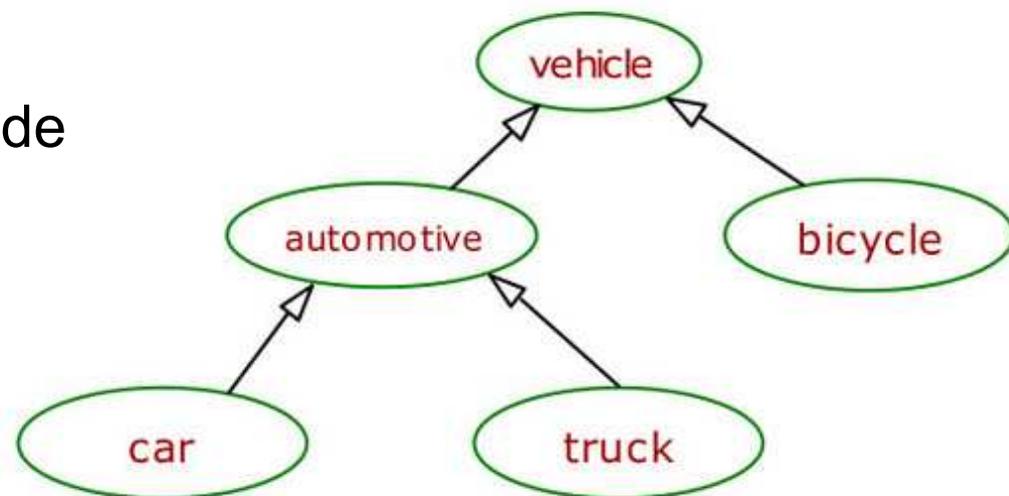
**Idée :**

Relier une distribution statistique de fréquences à une structure sémantique

Utiliser les termes les plus pertinents attachés à une thématique

### WordNet [MIL95]

- ressource linguistique, hiérarchie de concepts
- un concept regroupe ensemble un set de synonymes
- **polysemie**: un mot peut avoir plusieurs significations, dépendent du contexte



**Solution proposée (3) :** Évaluer la cohésion sémantique des thématiques

**Alignement des thématiques :**

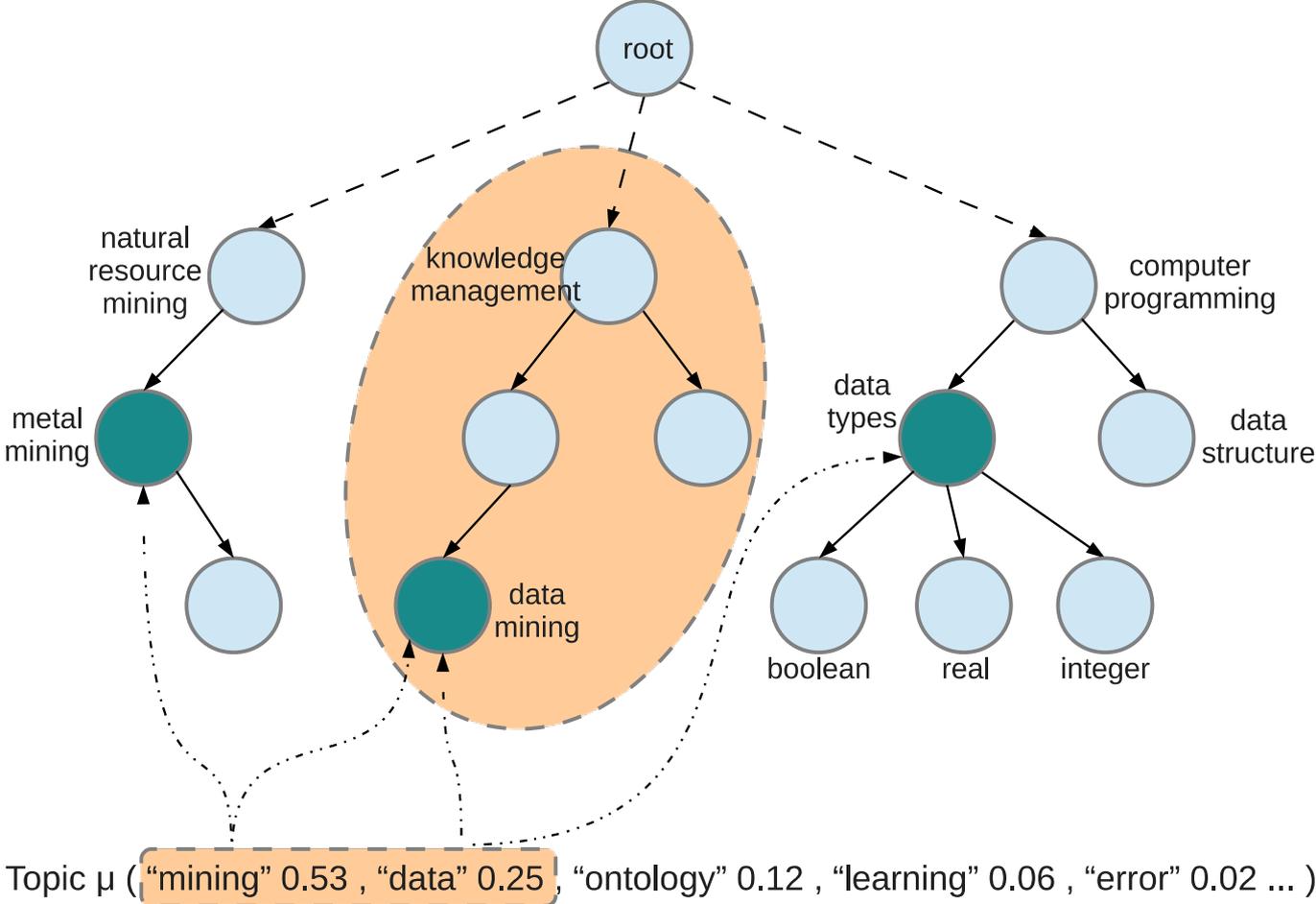
Déterminer le sous-arbre le plus spécifique qui contient au moins un sens pour chacun des mots les plus représentatifs de la thématique.

**Mesures :**

- couverture
- spécificité

**Évaluation d'une thématique :**

$$\varphi(\mu, c) = \omega_{spec} spec(\mu, c) + \omega_{cov} cov(\mu, c)$$



## Expérimentations et résultats

### Reuters, Suall11

Le forum « Y a-t-il trop de commémorations en France? », sur [www.liberation.fr](http://www.liberation.fr)

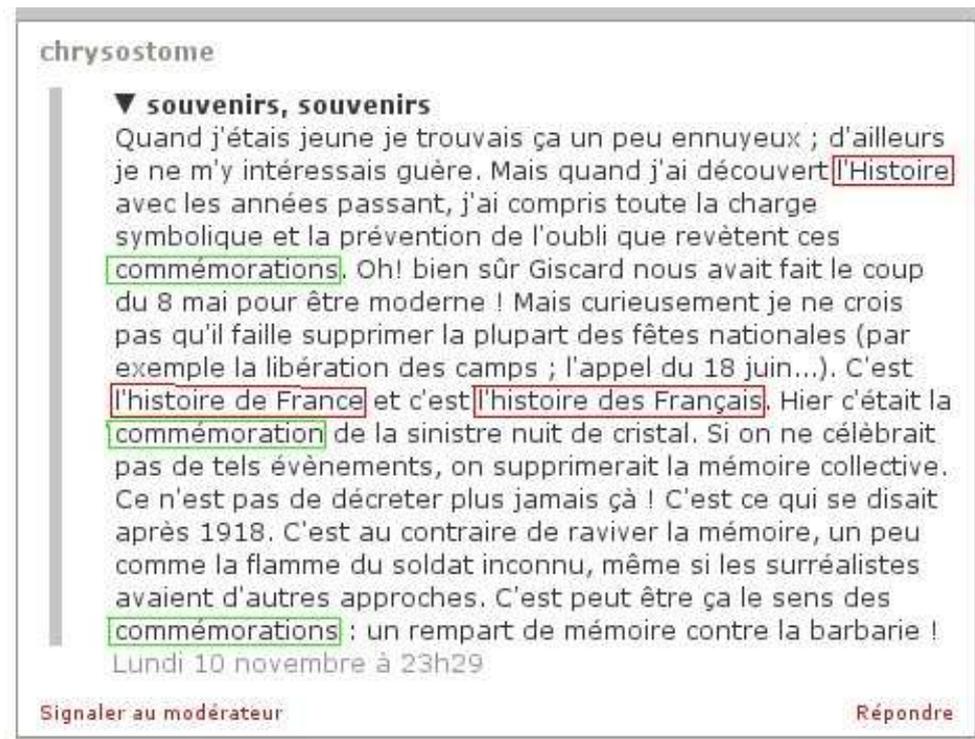
Corpus économique extrait à partir du site d'[Associated Press](http://Associated Press)

--> Iteration no 11:  
 ---> Objective function value: 189.154  
 ---> Partitions:  
 ----> Cluster 0 [101]: .....  
 ----> Cluster 1 [90]: .....  
 ----> Cluster 2 [128]: ..... **texte\_81** .....  
 ----> Cluster 3 [192]: ..... **texte\_81** .....

#### Result - Cluster description:

-> Centroid[0]: "jours fériés"  
 -> Centroid[1]: "travailler plus pour gagner"  
 -> Centroid[2]: "**commémoration**"  
 -> Centroid[3]: "**histoire de france**"

Exemple de sortie du logiciel d'extraction de thématiques, inclus dans **CommentWatcher**



Document « texte\_81 » sur le site du forum

# Expérimentations et résultats

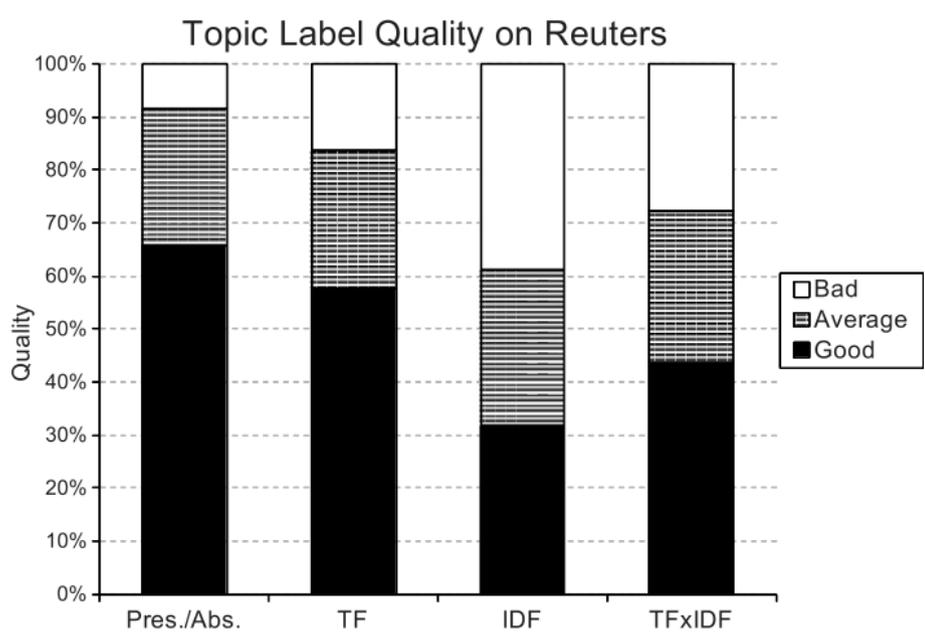
## Reuters, Suall11

Le forum « Y a-t-il trop de commémorations en France? », sur [www.liberation.fr](http://www.liberation.fr)

Corpus économique extrait à partir du site d'[Associated Press](http://Associated Press)

### Protocole :

Basé sur des experts, inspiré de la littérature [CHA09]



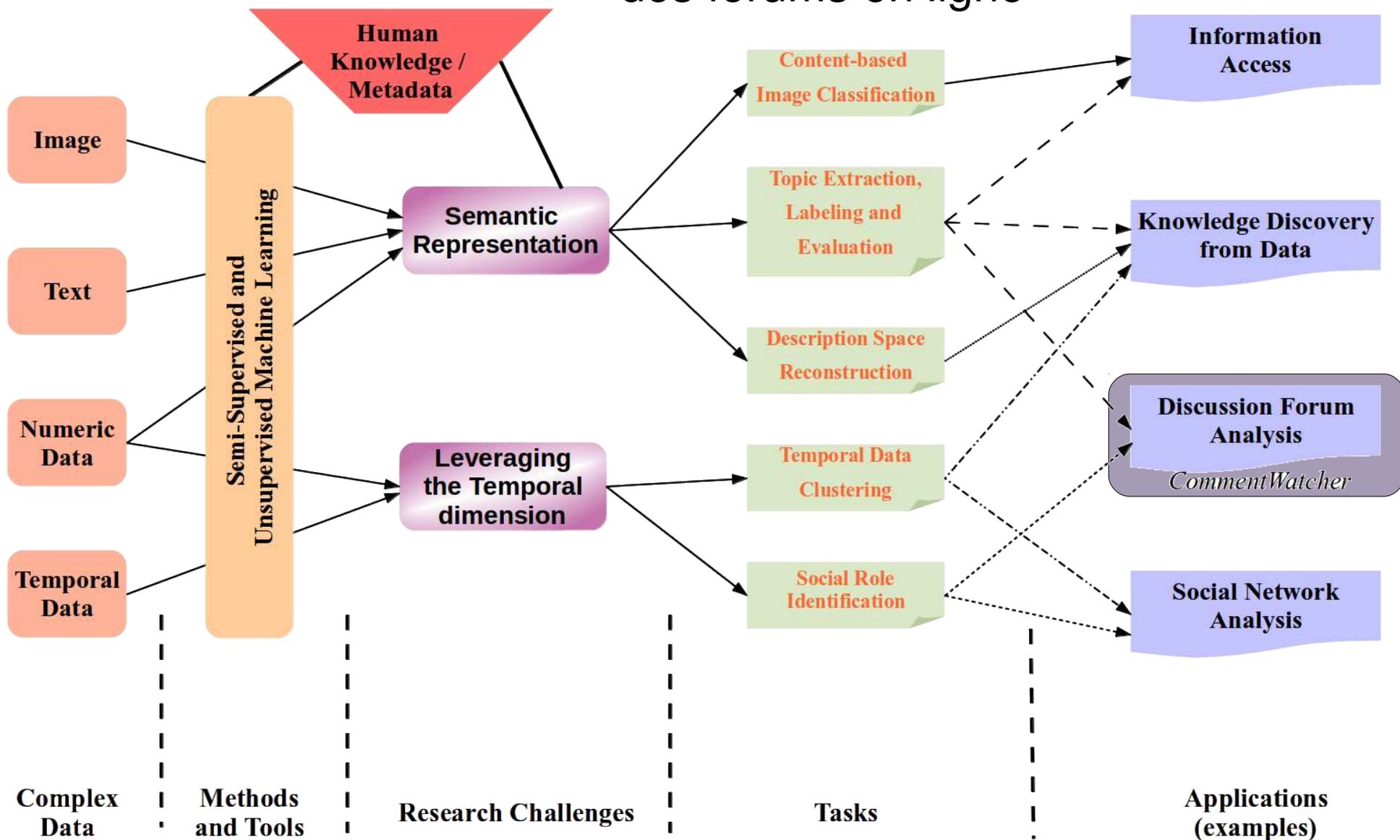
Évaluation de noms des thématiques

Dataset	$\overline{hit}_+$	$\overline{hit}_-$	Avantage rel. $\overline{hit}$
AP	<b>0,69</b>	0,65	6,93 %
Suall11	<b>0,75</b>	0,59	28,55 %

Évaluation de l'alignement des thématiques aux sous-arbres des concepts

# Dimension appliqué :

Logiciel d'analyse de discussions sur des forums en ligne



# Le contexte du travail - les forums de discussion

## Difficultés :

- La plupart des outils ne traitent pas l'aspect réseau social des données forum [AME12, GUI13]
- Manque de jeux de données issues de forums
- Structure des sites qui change constamment
- Problème de licence sur le contenu des forums

The image shows a screenshot of a forum thread with several posts. Red arrows and boxes highlight specific elements:

- Nom d'utilisateur:** Points to the name 'nicolas22' in the first post.
- Date du message:** Points to the timestamp 'il y a 54 minutes (11h47)' in the first post.
- Popularité (infos supplémentaires):** Points to the number of fans '40 Fans' for the user 'XenoPhil' in the third post.
- Relation structurelle (réponds à):** Points to the 'Répondre' button in the second post, which is linked to the first post.

The forum posts include:

- Post 1: User 'nicolas22' (0 Fans) posted 'il y a 54 minutes (11h47)'. Content: 'pour avoir des réponses concernant l'espionnage de l'Europe il suffit de demander a nos amis anglais, ils sont copain comme cochon .Il faut que les anglais sortent de l'Europe'.
- Post 2: User 'Isabelle Forger' (27 Fans) posted 'il y a 35 minutes (12h06)'. Content: 'ou qu'ils choisissent leur camp...'.
- Post 3: User 'SUPER UTILISATEUR DU HUFFPOST XenoPhil' (40 Fans) posted 'il y a 1 heure (11h23)'. Content: 'Video Not Available This video has not been made available in your country by the owner' Ils devraient prendre exemple sur les "joumaux", à la NSA ??? (au moins eux ils savent garder les infos secrètes)'. Includes a 'Répondre' button.
- Post 4: User 'pablico' (168 Fans) posted 'il y a 2 heures (11h01)'. Content: 'Écouter aux portes, et regarder par les trous de serrures est ce du terrorisme?'.
- Post 5: User 'SUPER UTILISATEUR DU HUFFPOST XenoPhil' (40 Fans) posted 'il y a 1 heure (11h30)'. Content: 'Lorsque c'est les "ncains", juste derrière il y a les drones qui assassinent !!!! Alors OUI, dans ce cas ce sont bien des terroristes ... D'ailleurs, rien d'étonnant à cela, les USA sont ou ont été derrière quasiment tous les terroristes comme BenLaden, les rebelles syriens, le Mossad, Jundollah en Iran etc etc'.

## Objectif général :

deux types d'utilisateurs

**l'analyste des forums** : comprendre les discussions entre les utilisateurs et leurs thématiques

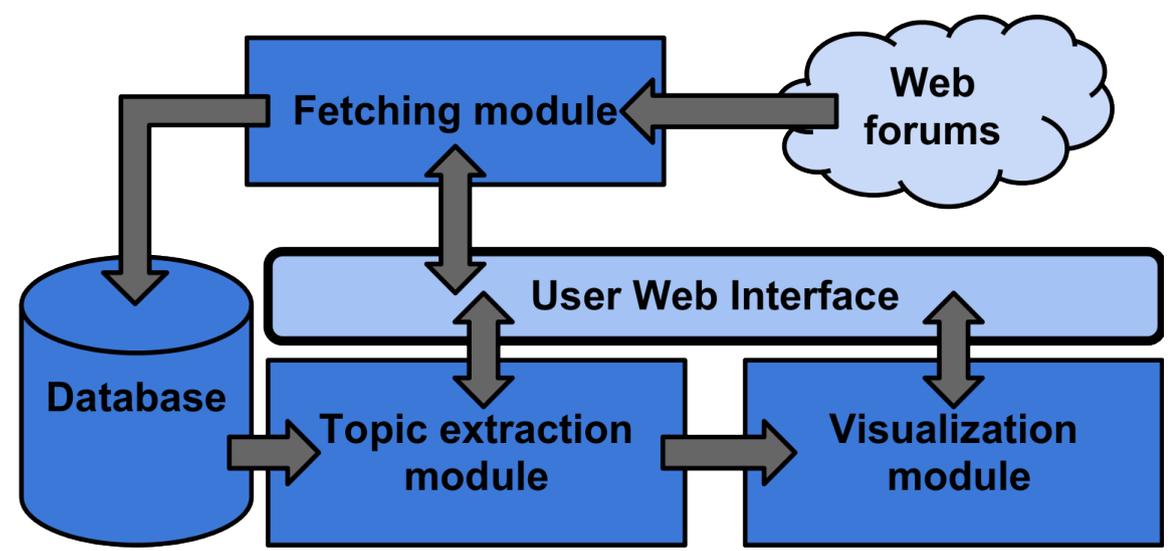
**le chercheur** : construire des jeux de données forums, analyser les évolutions des thématiques de discussion

## Notre proposition : CommentWatcher

Plateforme Web opensource (GPLv3)

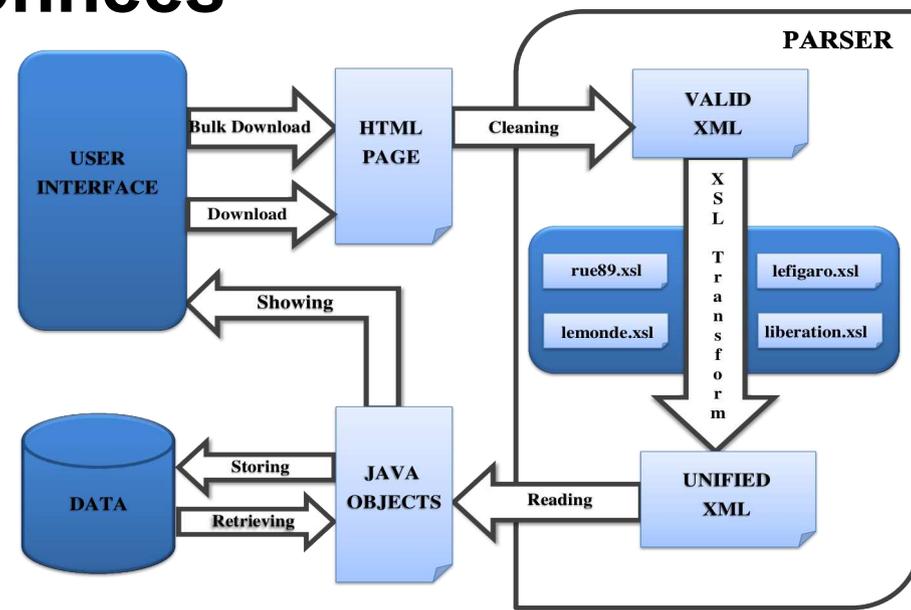
### 4 tâches :

- Récupération des données à partir d'Internet
- Extraction de thématiques
- Visualisation de thématiques comme un nuage d'expressions et l'évolution temporelle
- Visualisation du réseau social sous-jacent



## Module I. Récupération des données

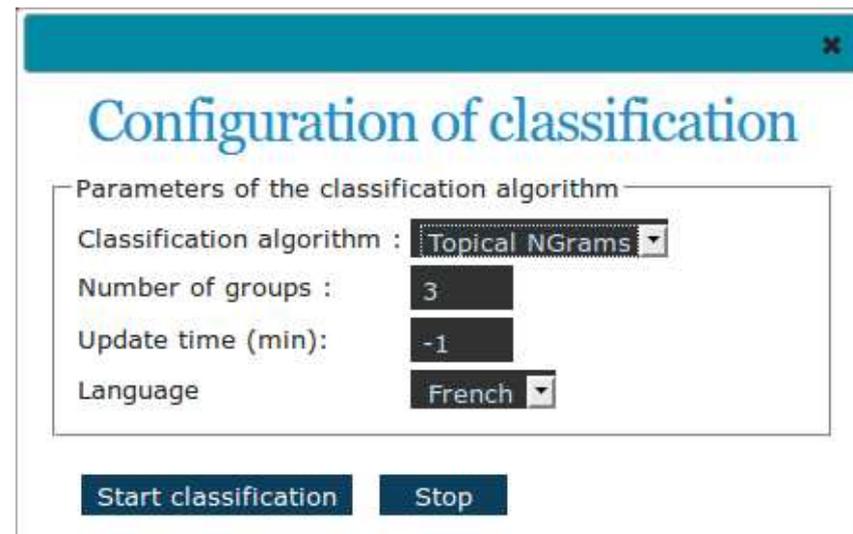
- Méta-parseur, indépendant de la structure des pages web
- Support pour de nouveaux sites via des fichiers de définition
- Recherche des forums supportés via une requête, en utilisant l'API Bing
- Téléchargement « en masse » des forums



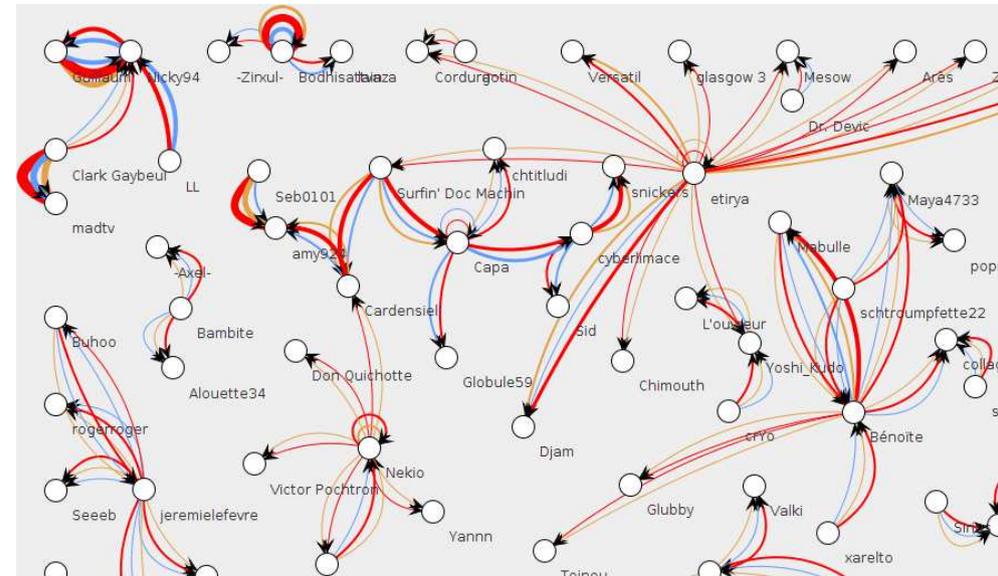
## Module II. Extraction des thématiques

3 algorithmes supportés :

- Topical Ngrams (suite Mallet [\[MCC02\]](#))
- CKP [\[RIZ10\]](#)
- Dynamic Topic Models [\[BLE06\]](#) (en développement)



# Module III. Visualiseurs



- ➔ Nuage d'expressions pour chaque thématique
- ➔ Évolution temporelle par forum et par site
- ➔ Évolution de la popularité d'une thématique

- ➔ Réseau social modélisé comme un multigraphe
- ➔ **Nœuds** : les utilisateurs ; **Arcs** : les messages associés à des thématiques
- ➔ Basé sur la relation de citation

# Travaux en cours

Faire évoluer automatiquement un ensemble de cibles

## Projet ImagiWeb

**Objectif du projet :** analyser les images (représentations) qui circulent sur Internet

### Partenaires



éditeur de logiciel pour la veille



sociologues spécialistes sciences politiques



utilisateurs, étude sémiologique



datamining, apprentissage automatique



fouille de textes et d'opinion, recherche d'information



TAL, extraction d'information

**Problématique d'apprentissage :** Faire évoluer d'une façon automatique les cibles utilisées pour annoter

## Motivation :

Actuellement, les textes (tweets, blogs) sont annotés en utilisant des cibles statiques, définies par des expertes.

*Ex :* L'ensemble des cibles politiques semble particulièrement adapté pour la période avant les élections

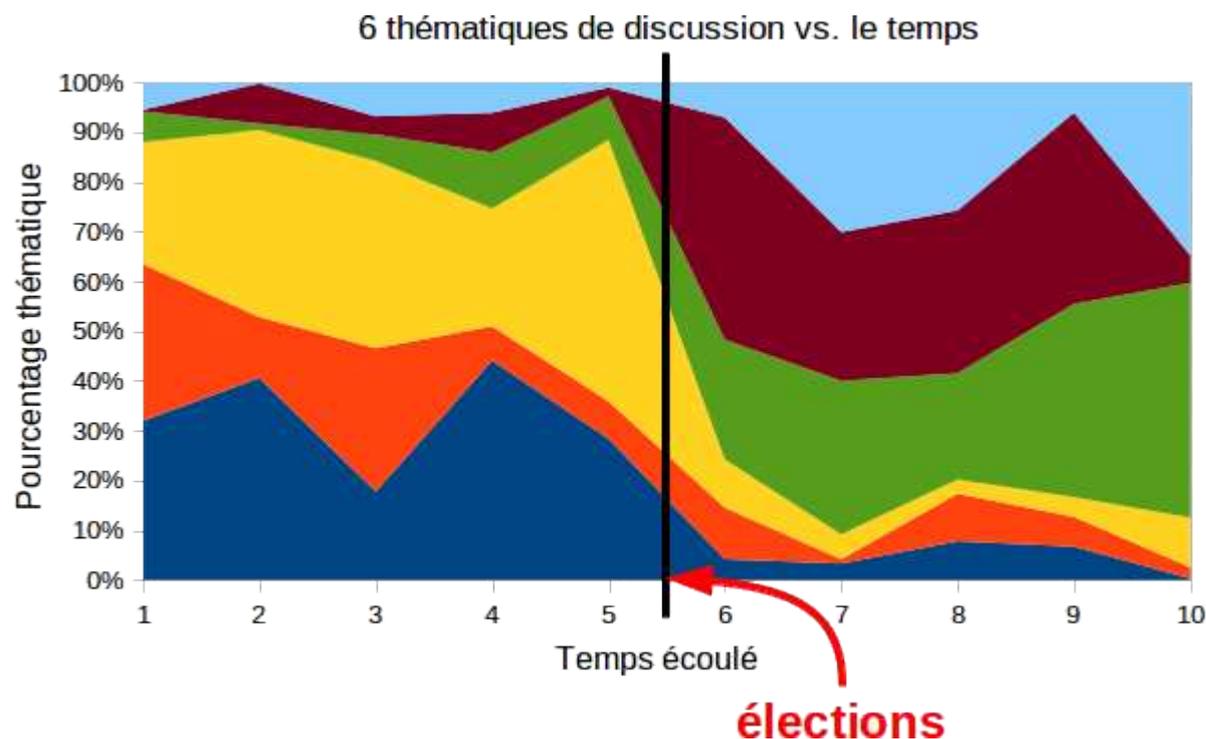
*Cibles :* Injonction / Soutien, Attribut / Sondage

**Problématique d'apprentissage :** Faire évoluer d'une façon automatique les cibles utilisées pour annoter

**Hypothèse :**

Avec l'écoulement de temps, les sujets de discussion entre internautes changent. Par conséquent, les cibles d'annotation doivent évoluer.

Nécessité de faire évoluer ou proposer des nouvelles cibles, d'un façon automatique



# Problématique d'apprentissage : Faire évoluer d'une façon automatique les cibles utilisées pour annoter

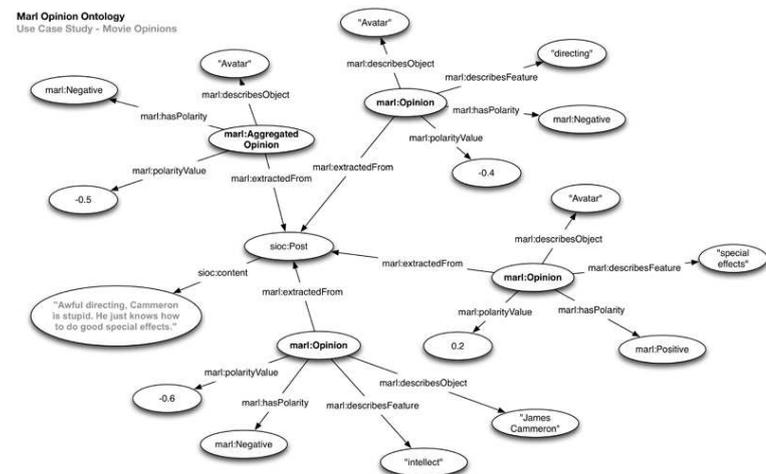
## Les données :

- Le texte de tweets et blogs, qui évolue temporellement
- L'ensemble original de cibles, proposées par des expertes

- Dernière ligne gauche! Pour virer la droite! Un seul vote @fhollande le seul qui ne promet pas ce qu'il ne pourra pas faire ! #FH2012 !
- RT @Vanneur: Francois Hollande demande a Aqmi de liberer les otages avant qu il soit trop tard... Traduction vous les tuez, on vous tue. #mali
- @reineroro<BR><BR>Le nouveau slogan avec @Hollande #larnaquecestmaintenant !<BR><BR>
- Hollande dit « dégage » à Boris Boillon, le « Sarko boy » de Tunis <http://t.co/2VJpTEAk>

Attribut / Sondage ; Attribut / Soutien/Non-soutien ; Bilan / Écologie ; Bilan / Économie ; Bilan / Societal ; ...

- Un ensemble de concepts dans une base de connaissances



# La base de connaissances :

**Dbpedia** – extraite automatiquement à partir de Wikipédia (en français)

Article Discussion

**Nicolas Sarkozy**

« Sarkozy » redirige ici. Pour les autres significations, voir Sarkozy (homonymie).

**Nicolas Sarkózy de Nagy-Bocsa**<sup>N 1</sup>, dit **Nicolas Sarkozy** [ni.kɔ.la.sak.kɔ.zi] <sup>( écouter)</sup><sup>N 2</sup>, né le 28 janvier à d'États français. Avocat de profession, Il occupe d'abord les fonctions de maire de Neuilly-sur-Seine, député des Hauts-de-Seine du gouvernement, ministre de la Communication ou encore de président par intérim du Rassemblement pour la France. Il est notamment ministre de l'Intérieur, ministre de l'Économie et des Finances et président du conseil général des Hauts-de-Seine les plus en vue de l'Union pour un mouvement populaire (UMP), dont il devient le président en 2004.

Il remporte l'élection présidentielle de 2007 avec 53,06 % des suffrages exprimés au second tour, face à la . Son mandat de président de la République française est marqué, entre autres, par une rupture de style par rapport à ses prédécesseurs comme celle des universités en 2007 ou des retraites en 2010, et par l'impact de grands événements internationaux tels que la dette dans la zone euro. Candidat à sa réélection à l'élection présidentielle de 2012, il recueille 48,36 % des votes exprimés par le candidat socialiste François Hollande.

Après son départ de la présidence, il est membre de droit et à vie du Conseil constitutionnel, où il siège pendant des conférences à l'étranger.

**Sommaire** [masquer]

- Famille et vie privée
- Études
- Carrière professionnelle
- Débuts en politique
- Premières responsabilités gouvernementales et « traversées du désert »
  - Ministre du Budget et porte-parole du gouvernement
  - Soutien à Édouard Balladur
  - Dirigeant du RPR
- Une influence grandissante au niveau national
  - Ministre de l'Intérieur, de la Sécurité intérieure et des Libertés locales
  - Ministre d'État, ministre de l'Économie, des Finances et de l'Industrie
  - Président de l'Union pour un mouvement populaire
  - Ministre d'État, ministre de l'Intérieur et de l'Aménagement du territoire
  - Campagne présidentielle de 2007
- Présidence de la République

dbpedia-owl:birthDate	1955-01-28 (xsd:date)
dbpedia-owl:birthName	Nicolas Paul Stéphane Sarkózy de Nagy-Bocsa
dbpedia-owl:birthPlace	<ul style="list-style-type: none"> <li>dbpedia-fr:17e_arrondissement_de_Paris</li> <li>dbpedia-fr:Paris</li> </ul>
dbpedia-owl:child	dbpedia-fr:Jean_Sarkozy
dbpedia-owl:nationality	dbpedia-fr:Nationalité_française
dbpedia-owl:occupation	<ul style="list-style-type: none"> <li>dbpedia-fr:Avocat_(métier)</li> <li>dbpedia-fr:Nicolas_Sarkozy_1</li> <li>dbpedia-fr:Nicolas_Sarkozy_2</li> <li>dbpedia-fr:Nicolas_Sarkozy_3</li> </ul>
dbpedia-owl:party	<ul style="list-style-type: none"> <li>dbpedia-fr:Rassemblement_pour_la_République</li> <li>dbpedia-fr:Union_pour_un_mouvement_populaire</li> <li>dbpedia-fr:Union_des_démocrates_pour_la_République</li> </ul>
dbpedia-owl:religion	dbpedia-fr:Catholicisme
dbpedia-owl:spouse	<ul style="list-style-type: none"> <li>dbpedia-fr:Cécilia_Attias</li> <li>dbpedia-fr:Carla_Bruni-Sarkozy</li> <li>dbpedia-fr:Marie-Dominique_Cuillio</li> </ul>
dbpedia-owl:thumbnail	http://upload.wikimedia.org/wikipedia/commons/thumb/d/d7/Flickr_-_europeanpeoplesparty_-_EPP_Summit_October_2011
dbpedia-owl:thumbnailCaption	Nicolas Sarkozy, au sommet du Parti populaire européen, le .
dbpedia-owl:university	dbpedia-fr:Université_Paris_Ouest_Nanterre-La_Défense
dbpedia-owl:wikiPageExternalLink	<ul style="list-style-type: none"> <li>http://infokiosques.net/implimersans2.php3?id_article=295</li> <li>http://www.404brain.info/NEWExpression/ExpressionEnginePB/images/uploads/Serge.Portelli.Ruptures.FRENCH.pdf</li> <li>http://www.editions-zones.fr/spip.php?id_article=116&amp;page=lyberplayer</li> <li>http://www.assemblee-nationale.fr/12/tribun/fiches_id/2680.asp</li> <li>http://www.ina.fr/recherche/recherche?search=nicolas+sarkozy&amp;vue=Video</li> <li>http://www.amisdenicolassarkozy.fr</li> <li>http://www.conseil-constitutionnel.fr/conseil-constitutionnel/francais/le-conseil-constitutionnel/les-membres-du-conseil</li> </ul>
dbpedia-owl:wikiPageID	675918 (xsd:integer)
dbpedia-owl:wikiPageRevisionID	95629613 (xsd:integer)
dbpedia-owl:wikiPageWikiLink	<ul style="list-style-type: none"> <li>dbpedia-fr:Catégorie:Naissance_dans_le_département_de_la_Seine</li> <li>dbpedia-fr:Catégorie:Ancien_conseiller_régional_d'Île-de-France</li> <li>dbpedia-fr:Catégorie:Député_européen_eû_en_France_1999-2004</li> <li>dbpedia-fr:Aïna_Madelin</li> <li>dbpedia-fr:Catégorie:Ministre_français_des_Finances</li> <li>dbpedia-fr:Catégorie:Personnalité_de_l'Union_pour_un_mouvement_populaire</li> <li>dbpedia-fr:Catégorie:Député_de_la_IXe_législature_de_la_Ve_République</li> <li>dbpedia-fr:Catégorie:Député_de_la_Xe_législature_de_la_Ve_République</li> </ul>

*L'article Wikipédia sur Nicolas Sarkozy*

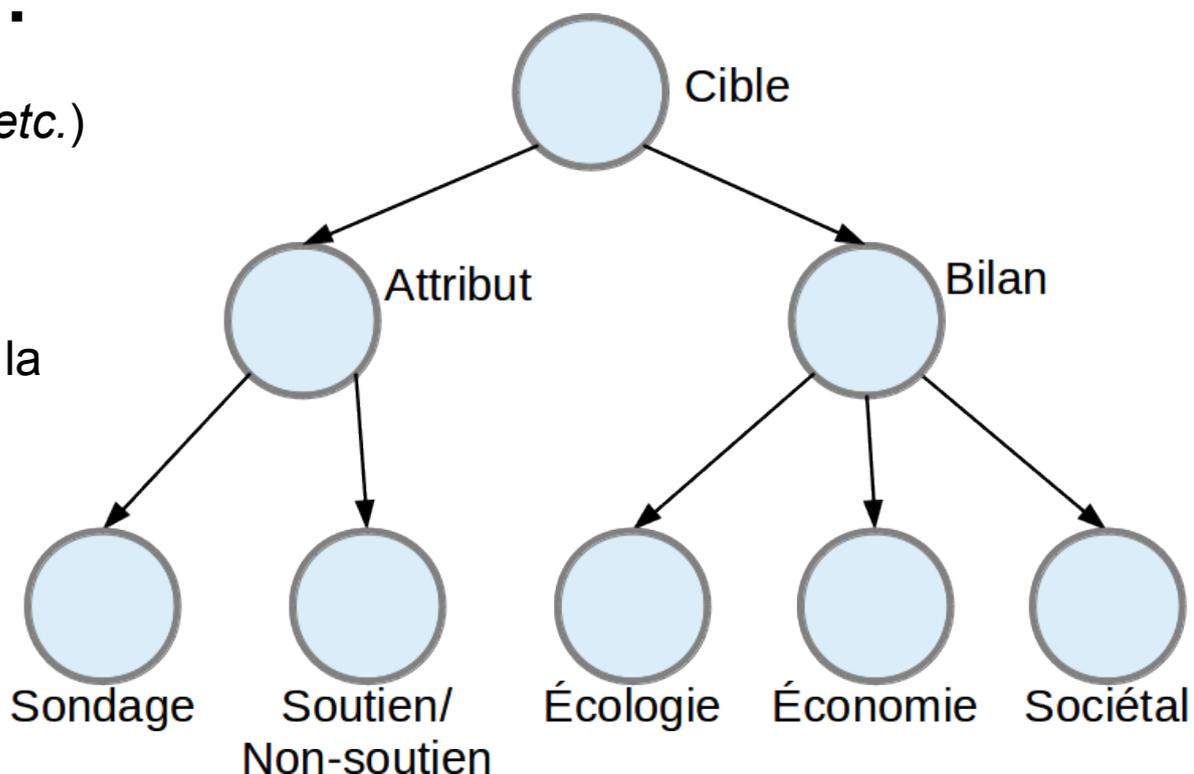
*L'entité Nicolas Sarkozy dans Dbpedia*

**La base de connaissances :**      **Dbpedia** – extraite automatiquement à partir de Wikipédia (en français)

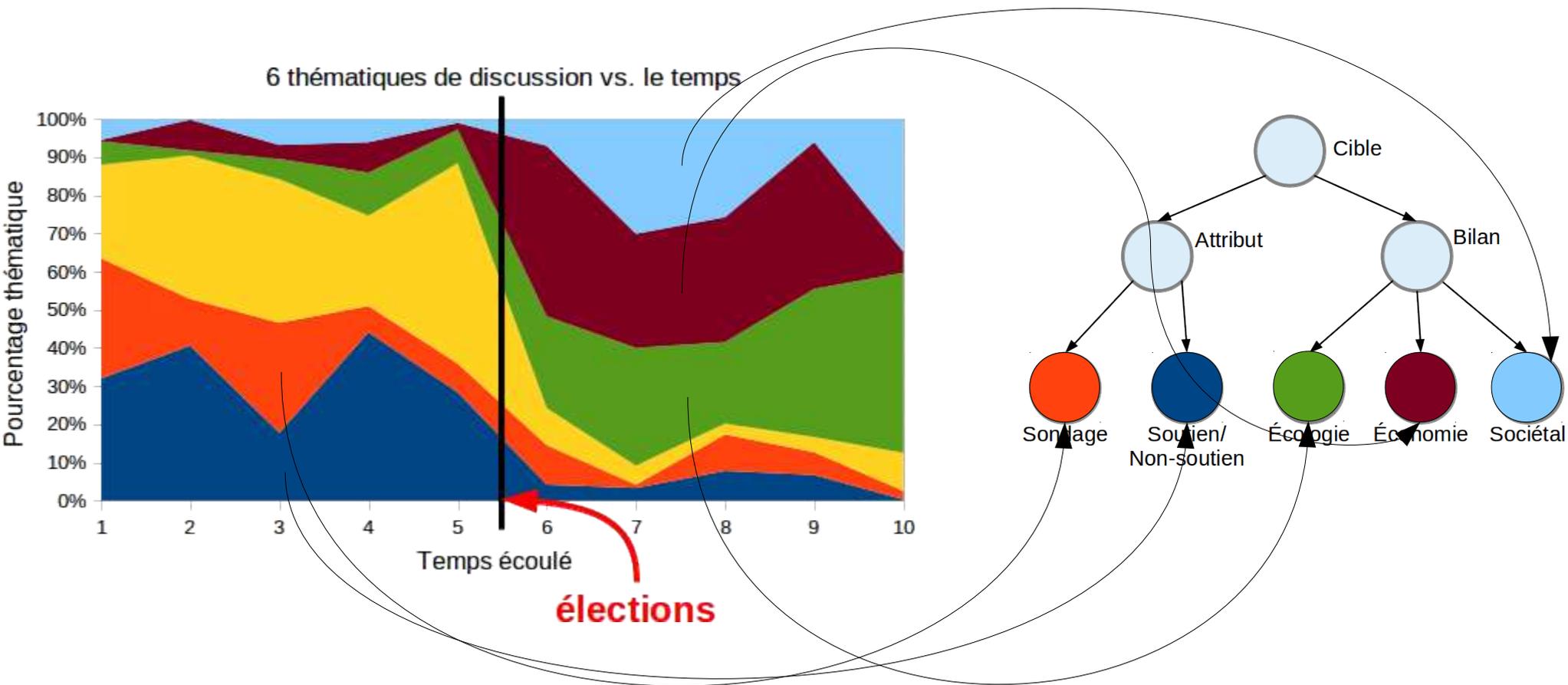
### Des connaissances sur :

- Des lieux (villes, pays, continents *etc.*)
- Des événements (*ex. les guerres mondiales, des attentats*)
- Des concepts (*ex. le bien et le mal, la philosophie etc.*)
- ...

**Peuvent être organisées comme une hiérarchie**

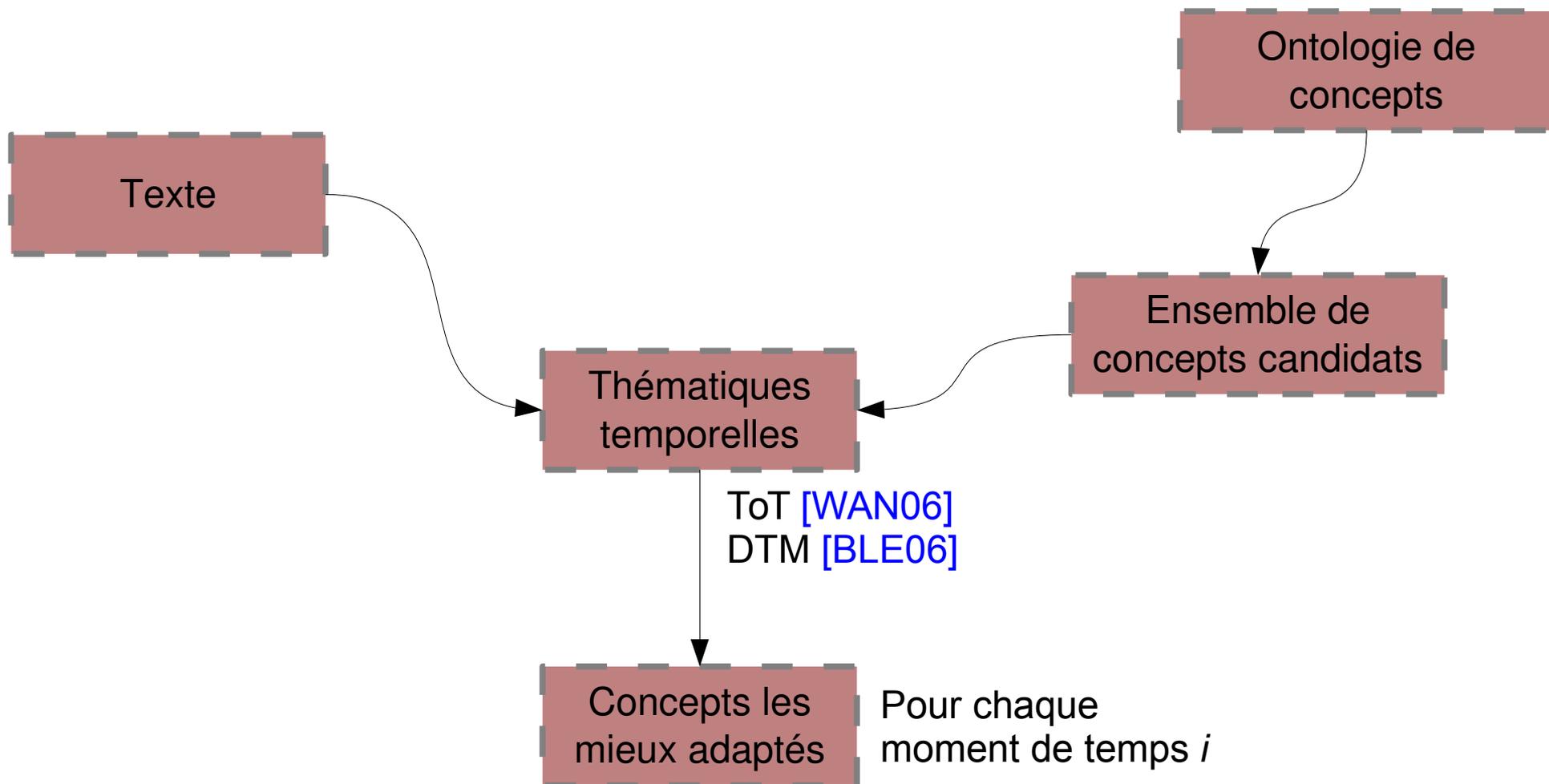


**L'idée :**      Faire évoluer automatiquement les cibles en utilisant les concepts de **Dbpedia**.



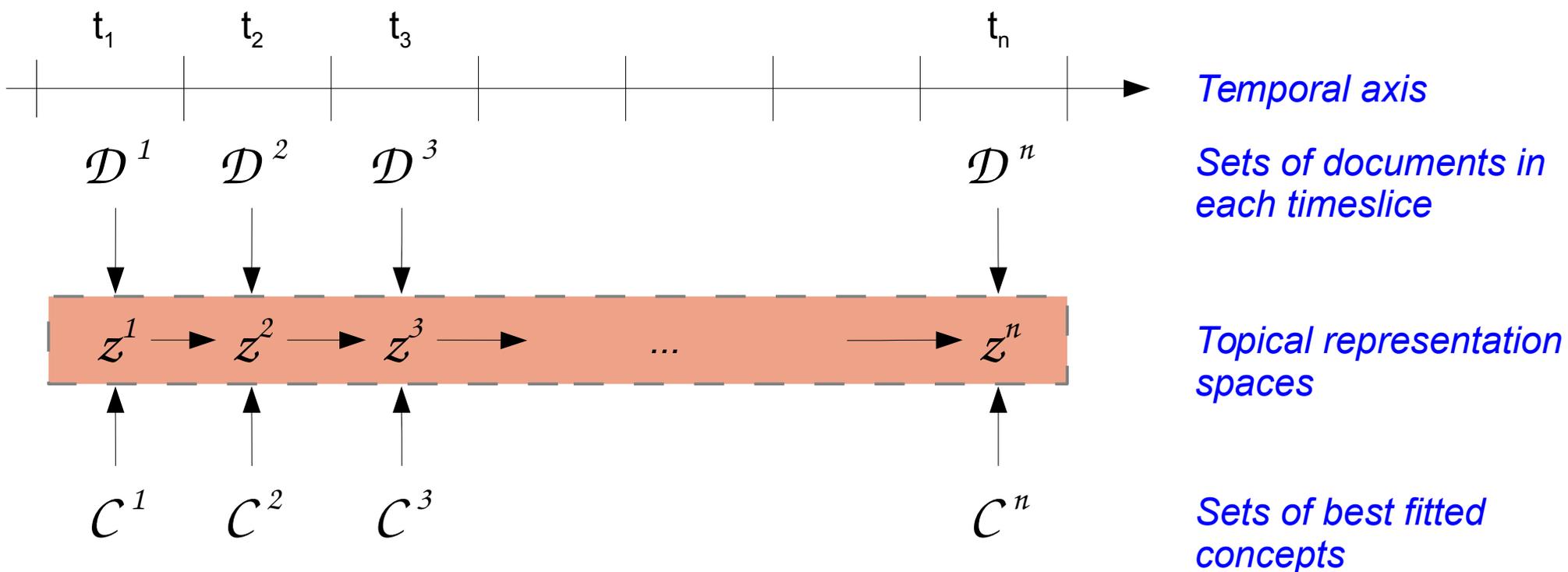
**L'idée :**      Faire évoluer automatiquement les cibles en utilisant les concepts de **Dbpedia**.

Faire le lien entre le texte et les concepts à travers l'extraction temporelle de thématiques :



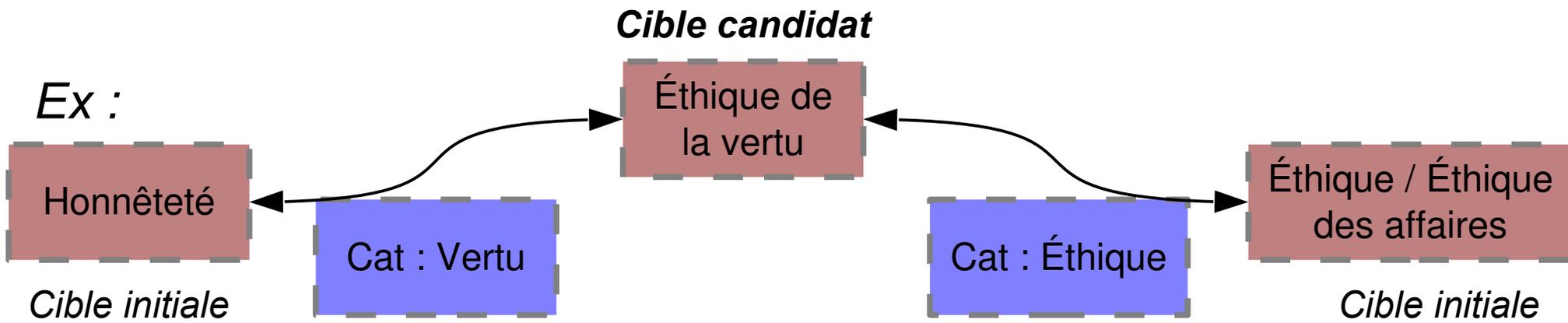
**L'idée :**      Faire évoluer automatiquement les cibles en utilisant les concepts de **Dbpedia**.

Faire le lien entre le texte et les concepts à travers l'extraction temporelle de thématiques :



**L'idée :**      Faire évoluer automatiquement les cibles en utilisant les concepts de **Dbpedia**.

Peupler l'ensemble de concepts candidats – liaison sémantique basée sur les connexions dans **Dbpedia** et la similarité textuelle



# Conclusions

- exemples de travaux en liaison avec la **dimension temporelle** de données complexes (e.g., détection d'évolutions typiques) et l'intégration des connaissances sémantiques dans la représentation des données
- exemples de travaux qui adressent en partie le **caractère multimodale** de données (image et texte) issue du Web (e.g., réseaux sociaux en ligne, microblogging, articles de presse)
- des informations externes, souvent en liaison avec le **Web Sémantique**, peuvent être utilisées pour améliorer l'analyse de données et l'extraction des connaissances (e.g., construction des profils utilisateur)
- des autres travaux en cours sur des **problématiques de vie privé** : la perte inhérente d'intimité en ligne et l'inférence de traits privés à partir des comportements publics.

## Bibliographie

**[WAG00]** Kiri Wagstaff and Claire Cardie. Clustering with Instance-level Constraints. In International Conference on Machine Learning, Proceedings of the Seventeenth, pages 1103–1110, 2000.

**[ZHE98]** Zijian Zheng. Constructing conjunctions using systematic search on decision trees. Knowledge-Based Systems, vol. 10, no. 7, pages 421–430, 1998.

**[SAW85]** Y. Sawaragi, H. Nakayama and T. Tanino. Theory of multiobjective optimization, volume 176. Academic Press New York, 1985.

**[RUS08]** Bryan C. Russell, Antonio Torralba, Kevin P. Murphy and William T. Freeman. LabelMe: a database and web-based tool for image annotation. International Journal of Computer Vision, vol. 77, no. 1, pages 157–173, 2008.

**[MIK04]** Krystian Mikolajczyk and Cordelia Schmid. Scale & affine invariant interest point detectors. International Journal of Computer Vision, vol. 60, no. 1, pages 63–86, 2004.

**[FRI10]** Simone Frintrop, Erich Rome and Henrik I. Christensen. Computational visual attention systems and their cognitive foundations: A survey. ACM Transactions on Applied Perception (TAP), vol. 7, no. 1, page 6, 2010.

**[KIS10]** Slava Kisilevich, Florian Mansmann, Mirco Nanni and Salvatore Rinzivillo. Spatio-temporal clustering. Data mining and knowledge discovery handbook, pages 855–874, 2010.

**[SIV03]** Josef Sivic and Andrew Zisserman. Video Google: A text retrieval approach to object matching in videos. In Computer Vision, Proceedings of the Ninth IEEE International Conference on, ICCV 2003, pages 1470–1477. IEEE, 2003.

**[LOW04]** David G. Lowe. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, vol. 60, no. 2, pages 91–110, 2004.

**[BAY06]** Herbert Bay, Tinne Tuytelaars and Luc Van Gool. Surf: Speeded up robust features. Computer Vision–ECCV 2006, pages 404–417, 2006.

**[LAZ06]** Svetlana Lazebnik, Cordelia Schmid and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, volume 2, pages 2169–2178. IEEE, 2006.

**[KIN10]** Teemu Kinnunen, Joni Kristian Kamarainen, Lasse Lensu, Jukka Lankinen and Heikki Kälviäinen. Making Visual Object Categorization More Challenging: Randomized Caltech-101 Data Set. In 2010 International Conference on Pattern Recognition, pages 476–479. 2010.

**[FEI07]** Li Fei-Fei, Rob Fergus and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. Computer Vision and Image Understanding, vol. 106, no. 1, pages 59–70, 2007.

**[ARM11]** Klaus Armingeon, David Weisstanner, Sarah Engler, Panajotis Potolidis, Marlène Gerber and Philipp Leimgruber. Comparative Political Data Set 1960-2009. Institute of Political Science, University of Berne., 2011.

**[WAL09]** Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov and David Mimno. Evaluation methods for topic models. In International Conference on Machine Learning, Proceedings of the 26th Annual, pages 1105–1112. ACM, 2009.

**[MAN93]** Udi Manber and Gene Myers. Suffix arrays: A new method for on-line string searches. SIAM Journal on Computing, vol. 22, no. 5, pages 935–948, 1993.

**[CLE08]** Guillaume Cleuziou. An extended version of the k-means method for overlapping clustering. In Pattern Recognition, 2008. ICPR 2008. 19th International Conference on, pages 1–4. IEEE, 2008.

**[MIL95]** George A. Miller. WordNet: a lexical database for English. Communications of the ACM, vol. 38, no. 11, pages 39–41, 1995.

**[ZHA07]** Qingfu Zhang and Hui Li. MOEA/D: A multiobjective evolutionary algorithm based on decomposition. Evolutionary Computation, IEEE Transactions on, vol. 11, no. 6, pages 712–731, 2007.

**[CHA09]** Jonathan Chang, Jonathan Boyd-Graber, Sean Gerrish, Chong Wang and David M. Blei. Reading Tea Leaves: How Humans Interpret Topic Models. In Advances in Neural Information Processing Systems, Proceedings of the 23rd Annual Conference on, volume 31 of NIPS 2009, 2009.

**[PHA08]** Pham, N.K., Morin, A., Gros, P., Le, Q.T.: Factorial correspondence analysis for image retrieval. In: Research, Innovation and Vision for the Future, 2008. RIVF 2008. IEEE International Conference on. pp. 269–275. IEEE (2008)