# 3ST
# *Surligneur Sémantique de Textes Scientifiques*

**Séminaire technique**
**« Chantiers d'usage » d'ISTEX**
**7 juin 2017**

Coordinateur du projet : Fabrice MUHLENBACH
courriel : fabrice.muhlenbach@univ-st-etienne.fr

# Plan de la présentation

- Contexte :
  motivations, équipe et moyens financiers
- Objectifs initiaux du projet 3ST
- Cas d'application :
  domaine → les sciences du sport
  sujet → la rotation mentale
- Approche, application et résultats
- Bilan du projet en cours

# Plan de la présentation

- Contexte :
  motivations, équipe et moyens financiers
- Objectifs initiaux du projet 3ST
- Cas d'application :
  domaine → les sciences du sport
  sujet → la rotation mentale
- Approche, application et résultats
- Bilan du projet en cours

# Contexte

## Motivation principale : favoriser l'accès au savoir

- renforcement des collaborations locales entre les chercheurs de Lyon et de Saint-Etienne, en particulier en sciences humaines et sociales (ISH de Lyon) :
  PRES (2007) → COMUE Université de Lyon (UdL en 2015)
  → préparation de l'IDEXLYON (label obtenu en fév. 2017)
- volonté de favoriser le partage et l'exploitation des connaissances produites par les chercheurs dans les différents domaines scientifiques (initialement en SHS) :
  des archives locales aux projets nationaux (HAL SHS)
  → bibliothèques numériques scientifiques internationales

# Contexte : l'**infonomie**

## ***Infonomie*** ?

- projet d'une nouvelle science s'intéressant à la manière de produire, de diffuser et d'utiliser des contenus immatériels : données, informations, connaissances, savoir-faire…
- étude des conditions dans lesquelles l'information, en tant qu'objet et produit socio-économique, scientifique et culturel, se crée, se transforme, se diffuse et agit à son tour pour créer d'autres informations

  → question de la valeur économique de l'information

  → question de la mesure du sens

# Contexte : l'**infonomie**

## *Infonomie* ?

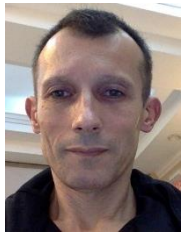
- l'**infonomie** se situe à la confluence des Arts et Humanités numériques, des Sciences sociales numériques et des Sciences et Technologies de l'Information
- **problème** : les disciplines scientifiques, et particulièrement en sciences humaines et sociales, mais aussi dans le domaine des arts et des lettres, sont le plus souvent déconnectées des autres disciplines
- pourtant : richesse des collaborations pluridisciplinaires

# • Équipe

Fabrice MUHLENBACH

Hussein AL-NATSHEH

Fabien RICO



Patrick FARGIER

Raphaël MASSARELLI

Lucie MARTINET

Djamel A. ZIGHED

# • Soutiens financiers

Ce travail a été réalisé grâce au soutien financier du *Programme Avenir Lyon Saint-Etienne* de l'Université de Lyon dans le cadre du programme « Investissements d'Avenir » (ANR-11-007)

La thèse d'Hussein AL-NATSHEH est soutenue par une allocation doctorale de recherche de la Région Auvergne-Rhône-Alpes

Le travail de post-doctorante de Lucie MARTINET a été financé dans le cadre des chantiers d'usage d'ISTEX (programme « Investissements d'Avenir » initié par le MESR) : **3ST**

# Plan de la présentation

- Contexte :
  motivations, équipe et moyens financiers
- Objectifs initiaux du projet 3ST
- Cas d'application :
  domaine → les sciences du sport
  sujet → la rotation mentale
- Approche, application et résultats
- Bilan du projet en cours

# Objectifs initiaux

## Constat

- accès des chercheurs à des masses d'informations (bibliothèques numériques d'articles scientifiques en ligne)
- exploration des documents scientifiques limitée à la communauté d'appartenance de chaque chercheur

## Proposition

- extension de l'exploration bibliographique au-delà de la communauté d'appartenance
- → contexte pluri- et trans-disciplinaire

# Objectifs initiaux

## Problèmes

- grande taille des bibliothèques numériques
- hétérogénéité des données
- complexité du langage naturel
- limitations cognitives et manque de temps :
  → incapacité à pouvoir embrasser des concepts issus :
  - d'articles anciens pourtant pertinents
  - d'articles venant de disciplines complémentaires (recherche le plus souvent limitée à la communauté scientifique d'appartenance)

# Objectifs initiaux

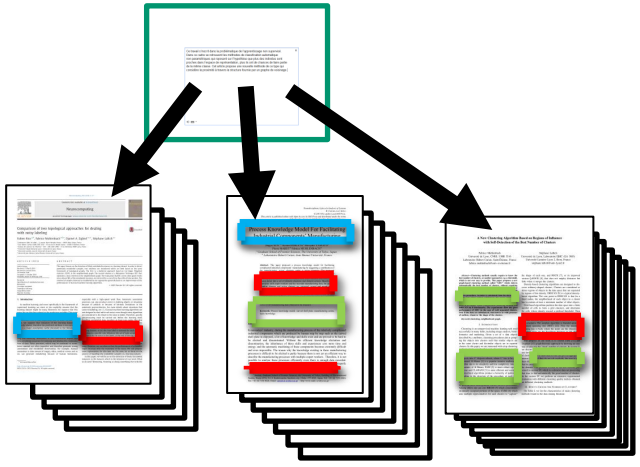
## Conséquence

→ le saut **quantitatif** en masse d'information apportée par les bibliothèques numériques ne se traduit pas vraiment en saut **qualitatif** pour le chercheur qui souhaite exploiter ces documents

# Objectifs initiaux

## Proposition pratique

- projet de recherche appliquée
- construction d'un outil de lecture assistée par ordinateur
- surlignage sémantique des textes scientifiques



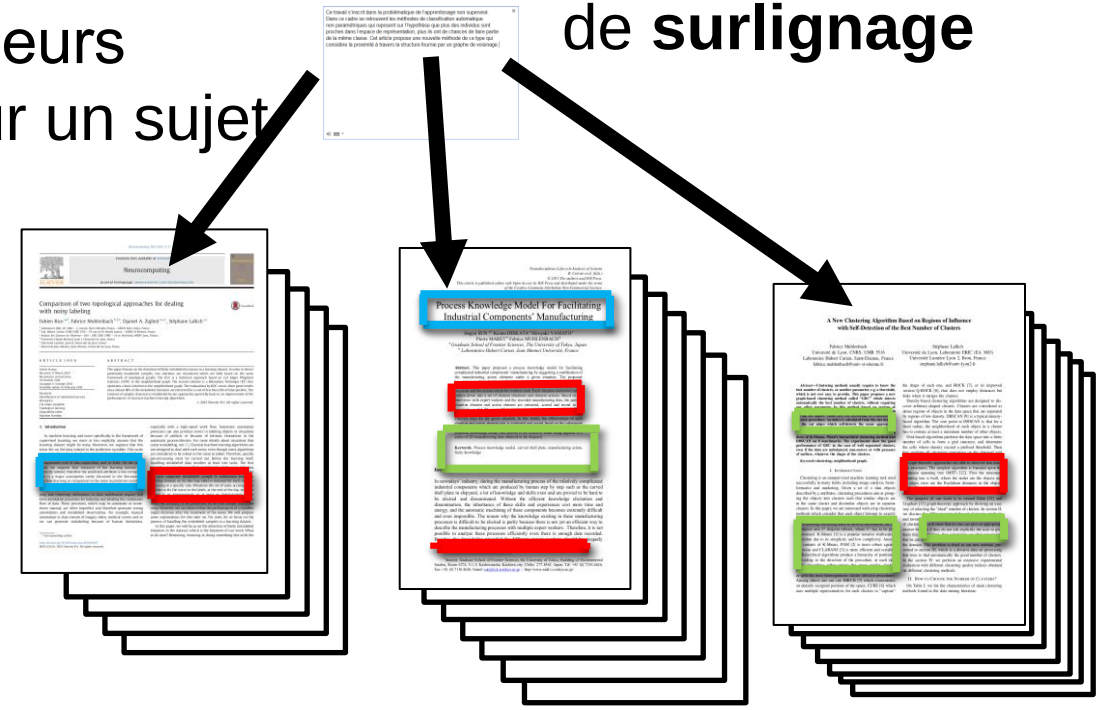
# Objectifs initiaux

## Le surligneur sémantique 3ST

en **entrée** : une requête composée d'une ou plusieurs phrases cibles portant sur un sujet d'intérêt de l'utilisateur

en **sortie** : présentation d'un ensemble d'articles du même domaine d'appartenance que cet utilisateur mais aussi d'autres disciplines

aide apportée à travers un système de **surlignage**



# Plan de la présentation

- Contexte :
  motivations, équipe et moyens financiers
- Objectifs initiaux du projet 3ST
- Cas d'application :
  domaine → les sciences du sport
  sujet → la rotation mentale
- Approche, application et résultats
- Bilan du projet en cours

# Cas d'application : la rotation mentale

## Pourquoi les sciences du sport ?

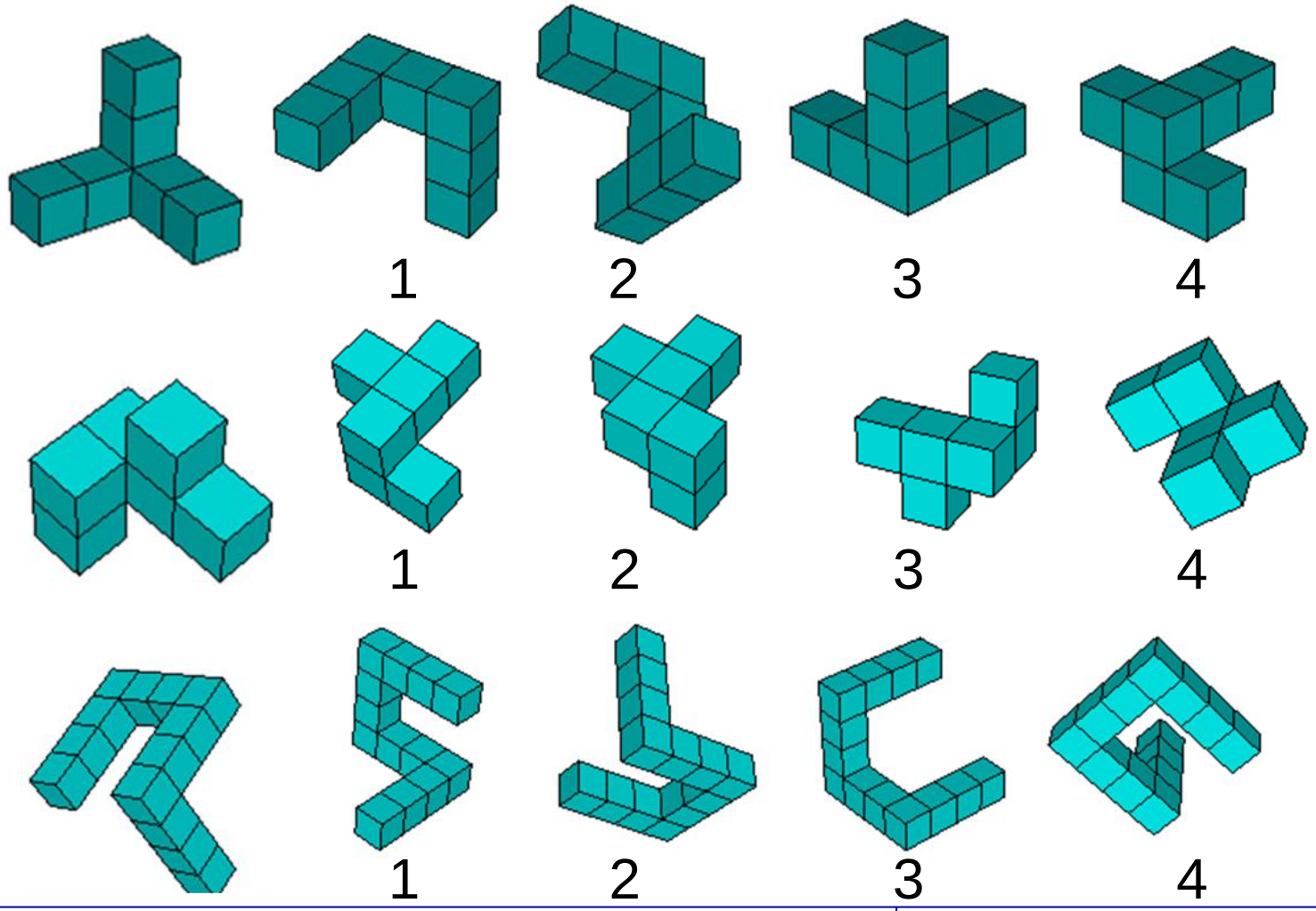
- par définition, les sciences du sport sont l'ensemble des sciences qui ont pour but la connaissance des différents aspects des pratiques sportives
- elles ont pour objet de déterminer des théories concernant la pratique physique (reconnaissance de lois et constantes, et repérer des principes généralisables à un ensemble de phénomènes)
- domaine par nature **pluridisciplinaire** : les pratiques sportives constituent un objet d'étude qui peut être abordé par différentes disciplines telles que la physiologie, la psychologie ou la psycho-sociologie

# Cas d'application : la rotation mentale

## Qu'est-ce que la rotation mentale ?

- la rotation mentale consiste en la capacité à faire tourner mentalement l'image d'un objet en 2 ou en 3 dimensions
- forme particulière d'imagerie mentale ou d'imagerie motrice nécessitant une structuration de l'espace
- implication des processus moteurs : une bonne représentation mentale requiert la capacité à travailler l'image mentale visuelle et à la faire tourner mentalement
- tâche de rotation mentale classique : indiquer le plus rapidement possible si deux images en 2 ou 3 dimensions, présentées sous différents angles, sont identiques ou différentes

# Cas d'application : la rotation mentale



# Cas d'application : la rotation mentale

## Caractéristiques de la rotation mentale

- opération mentale 1 : rotation d'au moins une des figures dans un des plans de l'espace pour la superposer avec l'autre afin de pouvoir juger de leur similitude ou différence
- opération mentale 2 : s'imaginer se déplacer soi-même en tournant autour de l'objet afin de le visualiser sous un angle différent pour pouvoir effectuer le jugement de similitude
- études : recherche des mécanismes sous-jacents de la rotation mentale et ses liens avec la performance et l'expérience motrice ; évolution des capacités individuelles à la suite d'un entraînement et transfert vers d'autres domaines d'expertise (capacités motrices et intellectuelles)

# Cas d'application : la rotation mentale

## Recherches sur la rotation mentale

- étude des bases neuro-fonctionnelles de la rotation mentale à l'aide de l'imagerie cérébrale (IRMf, TEP, MEG, EEG)
- liens entre les capacités de traitement d'une image mentale (construction, transformation et manipulation d'une image visuelle) et les processus moteurs permettant de mettre en mouvement et de faire tourner cette image mentale
- la rotation mentale est un phénomène complexe :

➢ pas de spécialisations exclusives des aires cérébrales
➢ lien avec la réussite à l'école
➢ existence de différences garçons/filles sur les performances
➢ la pratique du sport d'équipe augmente les performances

# Plan de la présentation

- Contexte :
  motivations, équipe et moyens financiers
- Objectifs initiaux du projet 3ST
- Cas d'application :
  domaine → les sciences du sport
  sujet → la rotation mentale
- Approche, application et résultats
- Bilan du projet en cours

# Approche, expérimentations, résultats

## Approche

- système d'extension du corpus :
  - ➢ en entrée : des articles scientifiques portant sur un sujet
  - ➢ en sortie : des articles scientifiques associés au sujet

- contraintes :
  - ➢ articles issus de disciplines scientifiques différentes
  - ➢ articles présentant des « pépites » (la fouille de données, telle que définie par D. J. Hand en 2000, est la découverte de structures **intéressantes**, **inattendues** ou **précieuses** dans les grands ensembles de données)
  - ➢ meilleurs résultats que ceux obtenus par l'approche de la RI

# Approche, expérimentations, résultats

## Processus général

ISTEX L'excellence documentaire pour tous
bibliothèque numérique

**corpus de base**
articles sources

portant sur
un sujet donné
(exemples fournis
par l'utilisateur)

Système
d'extension
du corpus

articles associés
au sujet

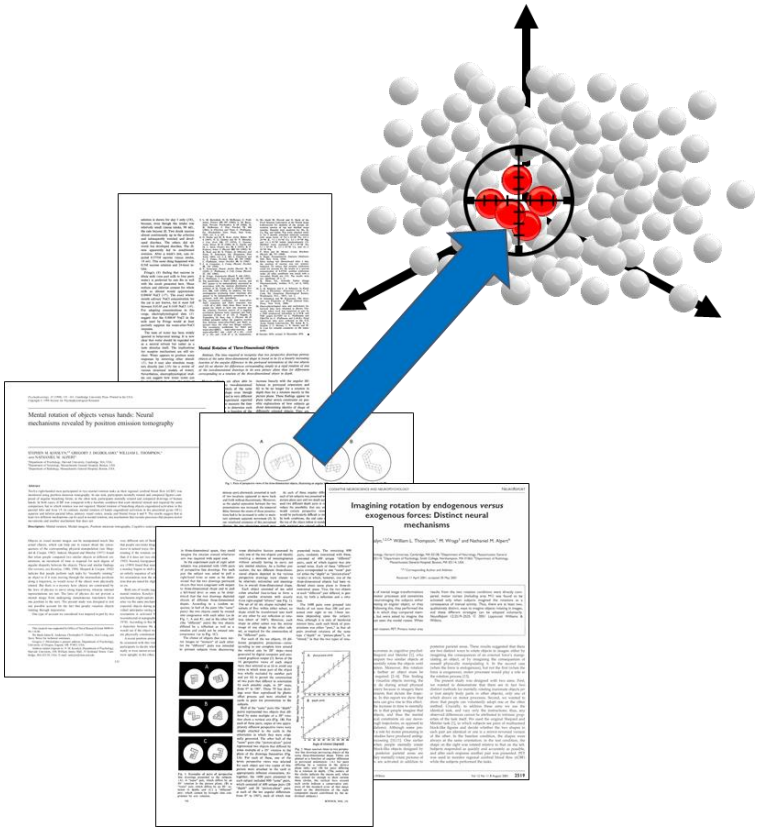
thème 1 … thème *i*

# Approche, expérimentations, résultats

## Construction de la représentation vectorielle

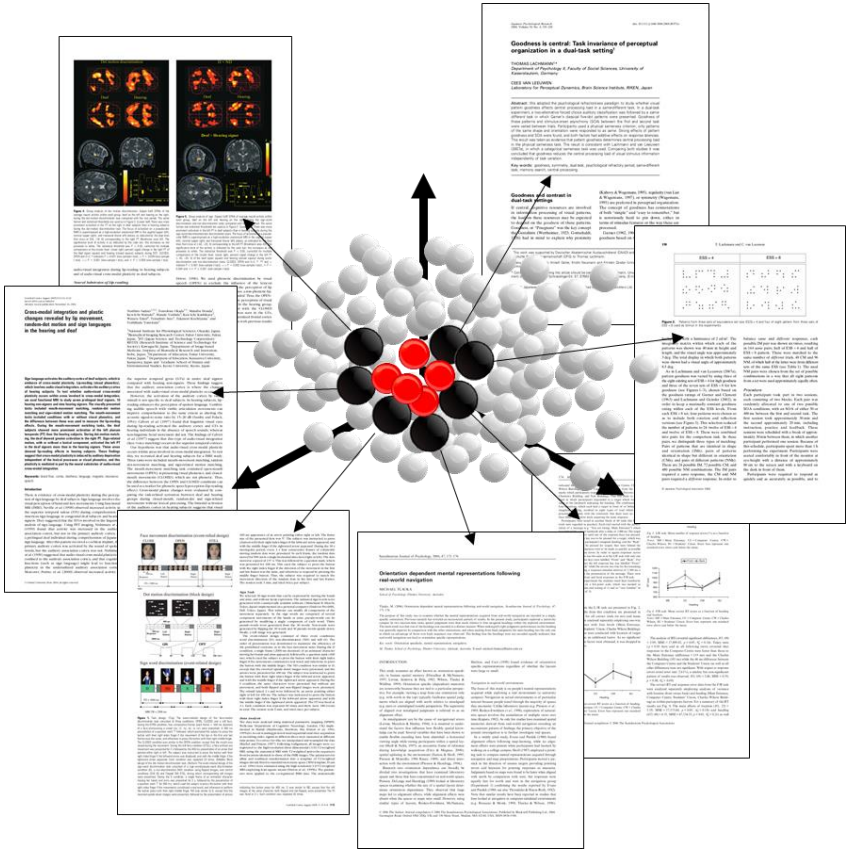
# Approche, expérimentations, résultats

## Projection de documents cibles dans cet espace



# Approche, expérimentations, résultats

## Activation d'articles sémantiquement proches



# Approche, expérimentations, résultats

## Fonctionnement du système d'extension du corpus

- représentation sémantique des documents par des vecteurs denses : décomposition en valeurs singulières, analyse sémantique latente
- apprentissage supervisé avec 2 classes (exemples positifs et exemples négatifs)
- prédiction effectuée sur tous les autres articles de la bibliothèque numérique
- tri par « top *k* » des articles prédits
- regroupement des articles par thème suivant l'allocation de Dirichlet latente (LDA)

# Approche, expérimentations, résultats

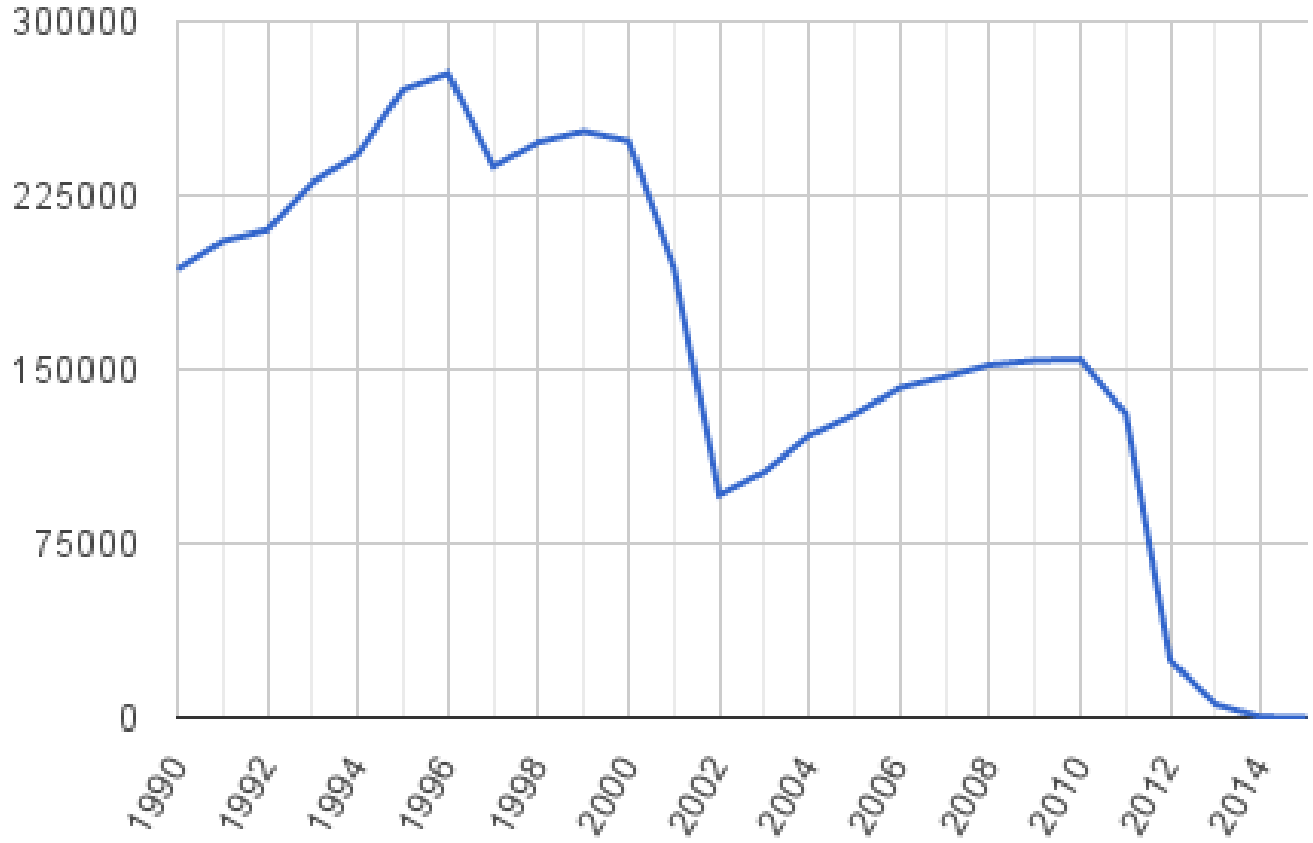
## Préparation : sélection des articles

- articles avec méta-données complètes (titre, résumé de 35 à 500 mots, auteurs…), assez récents et en anglais
- articles avec un nombre de pages compris entre 3 et 60
- articles de recherche, de journaux ou d'actes de conférence (pas de table des matières ou d'index, pas de posters…)
- utilisation de l'interface de programmation (API) d'ISTEX :

```
$istex-api-harvester -q "publicationDate:[1990 2016]
AND language:("eng" OR "unknown")
AND pdfPageCount:[3 60] AND abstractWordCount:[35 500]
AND genre:("research_article" OR "conference[eBooks]" OR "article")"
```

# Approche, expérimentations, résultats

## Articles issus de la bibliothèque numérique (en nb/an)



# Approche, expérimentations, résultats

## Détail des opérations du processus

- transformation des articles sélectionnés en représentation par sac de mots (matrice creuse)
- extraction des caractéristiques sémantiques descriptives des documents par modèles de vectorisation sémantique (Doc2Vec, décomposition en valeurs singulières, analyse sémantique latente)
- construction d'un modèle d'apprentissage supervisé (ici, classement par forêts aléatoires)
- utilisation du modèle pour retrouver des documents pertinents sans les mots clés du sujet (ici, "mental rotation")
- tri des documents par pertinence, tests par des experts

# Approche, expérimentations, résultats

## Expérimentations

- sujet : « rotation mentale », un centre d'intérêt de la recherche des sciences du sport, combinant les disciplines :
  - ➢ neurosciences (aires cérébrales impliquées)
  - ➢ psychologie (compétences cognitives, développement…)
  - ➢ physiologie (habiletés motrices)
- comparaison de la méthode proposée avec la méthode "more_like_this" d'*Elasticsearch* (indexation et RI)
- affichage des résultats combinant les retours des deux approches (notre méthode et *Elasticsearch*)
- présentation des résultats aux experts du domaine (vérité terrain permettant de quantifier les résultats)

# Approche, expérimentations, résultats

## Articles utilisés comme exemples positifs

- articles fournis par les utilisateurs experts du domaine (sciences du sport) sur le thème "mental rotation"
- ajout d'articles issus de la bibliothèque numérique avec les mots clés "mental rotation" :

```
wget https://api.istex.fr/document/?q="mental rotation"
AND publicationDate:[1990 2016]
AND language:("eng" OR "unknown")
AND pdfPageCount:[3 60]
AND abstractWordCount:[35 500]
AND genre:("research-article" OR "conference [eBooks]"
            OR "article" )&size=2000
```

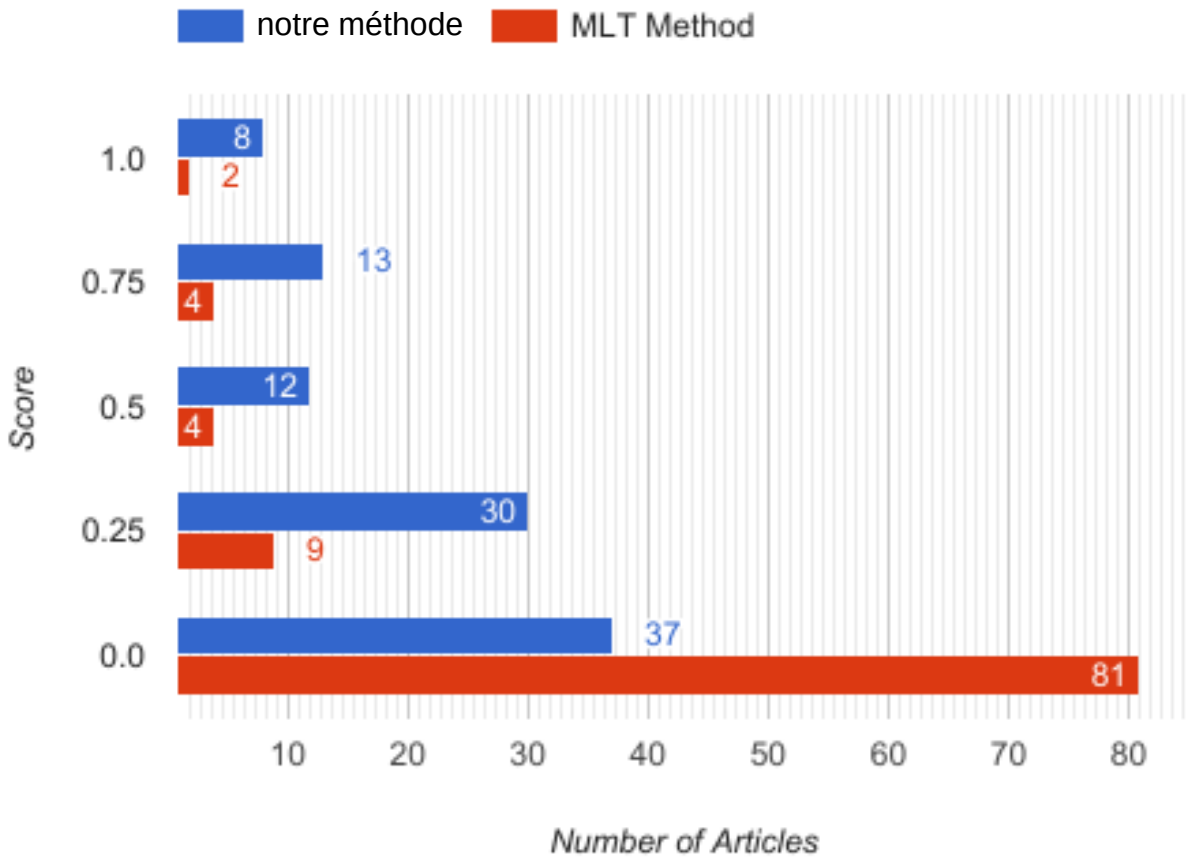
# Approche, expérimentations, résultats

## Résultats

- comparaison de notre approche avec l'état de l'art : *More-like-this*
- étude des similarités sémantiques entre les titres : avec 182 articles initiaux → 124 paires similaires avec MLT, 217 avec notre méthode et 382 après apprentissage actif
- étude de la pertinence des articles recommandés : évaluations manuelles par des experts (score de 0 à 1)
- étude de la diversité des articles présentés : modèle thématique, catégories de termes
- est-ce que le modèle favorise la découverte d'articles de recherche surprenants ? (sérendipité)

# Approche, expérimentations, résultats

## Pertinence des articles recommandés



# Approche, expérimentations, résultats

## Modèle thématique (*topic modeling*)

- sur les articles proposés par notre méthode, identification de thèmes considérés comme pertinents par les experts lorsque les articles sont regroupés en 2 thèmes :
  - ➢ un thème où sont présents les mots "motor", "task", "orientation", "stimuli", donc ayant plutôt trait à la partie expérimentale (psychologie cognitive) de la rotation mentale
  - ➢ un thème où apparaissent les mots "spatial ability", "visual", "mental rotation", "performance", "sex/age/profession differences", donc plutôt des éléments ayant trait aux aspects comportementaux ou sociaux de la rotation mentale

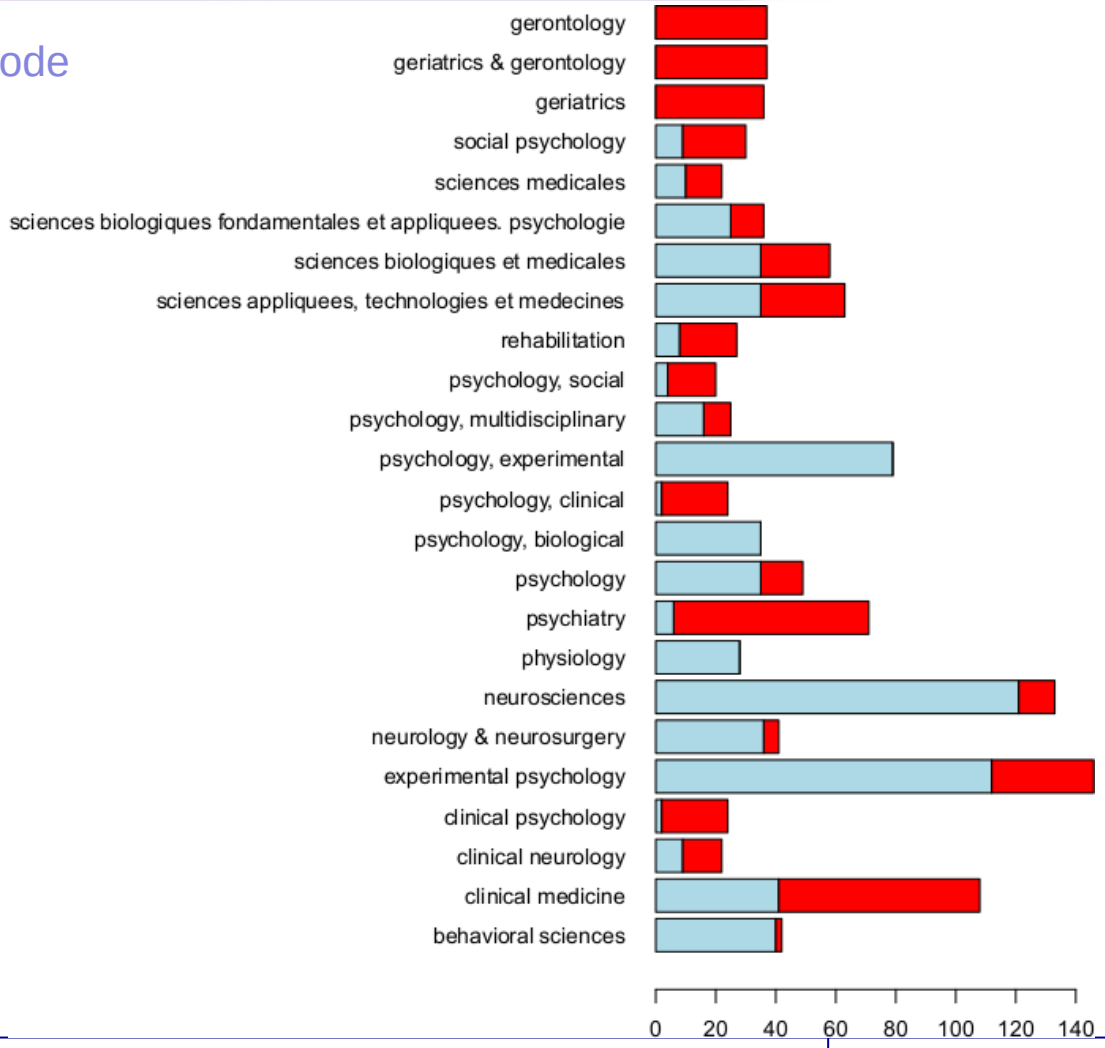
# Approche, expérimentations, résultats

## Modèle thématique (suite)

- des analyses plus poussées font apparaître des phrases clés dans ces thèmes, phrases qui sont liées :
  - ➢ soit aux approches expérimentales (p. ex. les potentiels évoqués ou la stimulation magnétique transcrânienne) ;
  - ➢ soit aux phénomènes enregistrés (p. ex. négativité de discordance) ;
  - ➢ soit aux conséquences comportementales (p. ex. le trouble du déficit de l'attention avec hyperactivité) ;
  - ➢ soit les aires cérébrales impliquées (p. ex. le lobule lingual ou le cortex périrhinal)...
- diversité thématique : notre approche > méthode MLT

# Approche, expérimentations, résultats

- notre méthode
- MLT



# Conclusions et perspectives

- travail encore en cours (l'évaluation manuelle de la pertinence des résultats par des experts prend du temps)
- résultats préliminaires encourageants
- découvertes d'articles « surprenants » : articles associés au sujet mais ne comportant pas l'expression "mental rotation" dans le texte, par exemple le langage des signes, lien entre des tâches de poursuites oculaires (qui sont un ensemble de tâches d'imagerie motrice), l'attention et la schizophrénie
- perspectives : améliorations possibles de la méthode

➢ test d'autres représentations sémantiques des documents par vecteurs denses (par exemple, par plongement lexical)

➢ autres méthodes d'apprentissage supervisé (p. ex., SVM)

➢ application à d'autres sujets de recherche pluridisciplinaire

# Plan de la présentation

- Contexte :
  motivations, équipe et moyens financiers
- Objectifs initiaux du projet 3ST
- Cas d'application :
  domaine → les sciences du sport
  sujet → la rotation mentale
- Approche, application et résultats
- Bilan du projet en cours

# Bilan du projet

## Projet encore en cours

- pas encore de « surligneur sémantique » disponible à l'heure actuelle
- ➢ travaux dans le domaine de l'IHM (visualisation, réseau) afin de faciliter l'interprétation des modèles thématiques

- valorisations scientifiques :
- ➢ similarité sémantique (article accepté à SemEval 2017)
- ➢ système de recommandation d'articles scientifiques
- ➢ apports en fouille de données, fouille de textes

# Bilan du projet

## Les « plus » indéniables apportés par ISTEX

- possibilité de travailler sur une vraie bibliothèque virtuelle d'articles scientifiques de niveau international (application réelle intéressante en fouille de textes)
- élément moteur pour établir des collaborations
- financement d'un poste de post-doctorant pendant 1 an
  → démarrage d'un projet concret, dégrossissage des problèmes techniques et scientifiques

# Bilan du projet

## Les difficultés du projet

- un an, c'est court…
- des données parfois difficilement exploitables en raison d'erreurs de catégorisation → focus sur les méta-données (résumé parfois absent du champ « abstract » mais présent dans le corps du texte, texte plein mal construit pour les articles en 2 colonnes, sujets à la place de mots-clés, etc.)
- pas (ou peu) d'accès aux articles les plus récents
- accès aux articles d'ISTEX limité aux universités partenaires (ou nécessité d'installer un VPN)
- besoins financiers (matériel, fonctionnement…)

UNIVERSITÉ DE LYON

LABORATOIRE HUBERT CURIEN
UMR • CNRS • 5516 • SAINT-ETIENNE

CONNECTED INTELLIGENCE

eric
ENTREPÔTS, REPRÉSENTATION & INGÉNIERIE des CONNAISSANCES

UNIVERSITÉ LUMIÈRE LYON 2
UNIVERSITÉ DE LYON

# 3ST
# *Surligneur Sémantique de Textes Scientifiques*

**Séminaire technique**
**« Chantiers d'usage » d'ISTEX**
**7 juin 2017**

ISTEX
L'excellence documentaire pour tous

Coordinateur du projet : Fabrice MUHLENBACH
courriel : fabrice.muhlenbach@univ-st-etienne.fr