ANALYSE DE DONNÉES TEXTUELLES À LA SNCF : CAS D'USAGE ET DIFFICULTÉS RENCONTRÉES

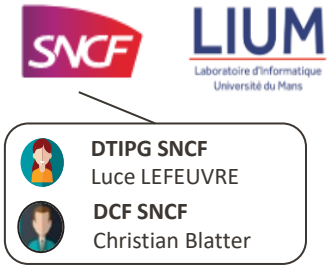
Coralie REUTENAUER – SNCF
Journée de lancement TextMine – 21/10/2022

SNCF SÉCURITÉ

SNCF

SOMMAIRE

- **PARTIE 1 : Expérimentations TAL - cas d'usage**
- **Recherche d'information dans les textes de prescription**
- **Analyse de Retours d'Expérience**
- **Standards de rédaction technique**
- **Analyse du web pour des opérations à l'interface client / agent**



- **PARTIE 2 : Limites récurrentes**

RECHERCHE D'INFORMATION DANS LES TEXTES DE PRESCRIPTION

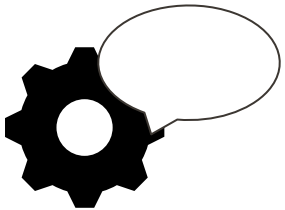
CONTEXTE ET OBJECTIFS

- **Programme sécurité de transformation de la documentation technique**

90 000 textes

- **Expérimentation de technologies sémantiques**

~ 7 000 textes
Corps de texte + thème



Apprentissage artificiel (LIUM, Lincoln)
Ontologies (Viseo)

Prototypage et évaluation

RECHERCHE D'INFORMATION DANS LES TEXTES DE PRESCRIPTION

COLLABORATION LIUM / SNCF : APPRENTISSAGE ARTIFICIEL POUR LA RI

LIUM Laboratoire d'Informatique Université du Mans

LINCOLN

- **Apprentissage artificiel pour améliorer la recherche d'information et la qualification de l'information**

Représentation du vocabulaire
Dont : plongements lexicaux (word embeddings)

Classification pour prédire les thèmes des documents
Random Forest & Naïve bayes

Clustering pour regrouper les résultats de recherche
K-Means et Spectral

SNCF SÉCURITÉ

SNCF

RECHERCHE D'INFORMATION DANS LES TEXTES DE PRESCRIPTION

REPRÉSENTATION DES TEXTES ET MODÉLISATION DU VOCABULAIRE

LIUM
Laboratoire d'Informatique
Université du Mans

- **Spécificités du corpus**
- Petit corpus (7 000 textes)
- Langue française, vocabulaire de spécialité

- **Prétraitements textuels et linguistiques**
- Nettoyage des données, enrichissement linguistique, filtrage du lexique
- Sortie : corpus de 18 000 mots et 10 millions occurrences

- **Représentations vectorielles**
- Classiques : tf-idf, rake
- **Plongements lexicaux ➔ choix d'une méthode efficace sur de petits corpus**

SNCF SÉCURITÉ

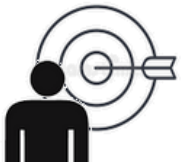
SNCF

RECHERCHE D'INFORMATION DANS LES TEXTES DE PRESCRIPTION

REPRÉSENTATION DES TEXTES ET MODÉLISATION DU VOCABULAIRE

LIUM
Laboratoire d'Informatique
Université du Mans

- **Evaluation de la modélisation du vocabulaire par les plongements lexicaux**



Valider la pertinence d'association de 2 mots

9 experts SNCF
160 mots (960 paires) par expert

Interface avec formulaire web

- **Complexe à évaluer :** accord inter-annotateurs bas, associations inconnues ; faux positifs et faux négatifs
- **Limites des plongements lexicaux** pour modéliser le vocabulaire spécialisé, notamment les acronymes

SNCF SÉCURITÉ

SNCF

RECHERCHE D'INFORMATION DANS LES TEXTES DE PRESCRIPTION

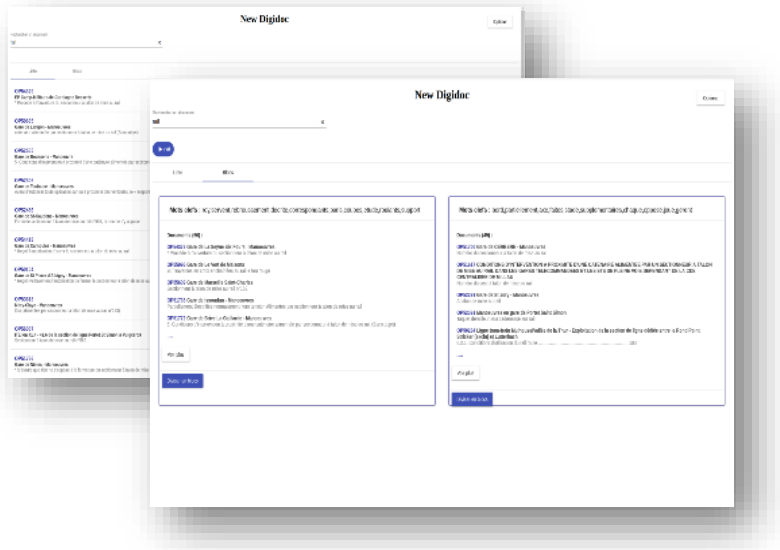
REGROUPEMENT SÉMANTIQUE ITÉRATIF DES RÉSULTATS DE RECHERCHE

LIUM
Laboratoire d'Informatique
Université du Mans

LINCOLN

- **Objectif :** organiser les résultats de recherche d'une requête utilisateur

- **Clustering :** K-means, recherche parallélisée, choix de la meilleure partition (indice EC) ; plusieurs représentations



- **Résultats**
- Beaucoup de **mots non porteurs de sens**, représentation optimale difficile à trouver
- **Evaluation** de la cohérence sémantique difficile
- **Maturité** insuffisante pour une industrialisation en l'état

SNCF SÉCURITÉ

SNCF

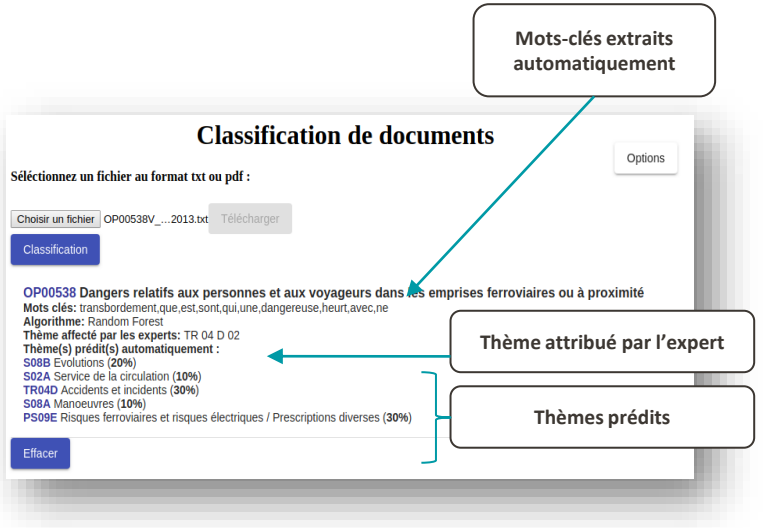
RECHERCHE D'INFORMATION DANS LES TEXTES DE PRESCRIPTION

PRÉDICTION THÉMATIQUE

LIUM Laboratoire d'informatique Université du Mans

LINCOLN

- **Objectif :** classifier les documents par thème
- **Thème :** arborescence à 5 niveaux, répartition déséquilibrée
- **Classification (apprentissage supervisé) :** représentation tf-idf, test de plusieurs classifieurs, validation croisée 5 plis



- **Mesures d'évaluation classiques en RI**
- **Résultats**
- bonnes performances sur le 1er niveau, baisse sur les niveaux suivants
- prometteur, mais à évaluer au regard de la plue-value métier de l'information selon le niveau

SNCF SÉCURITÉ

SNCF

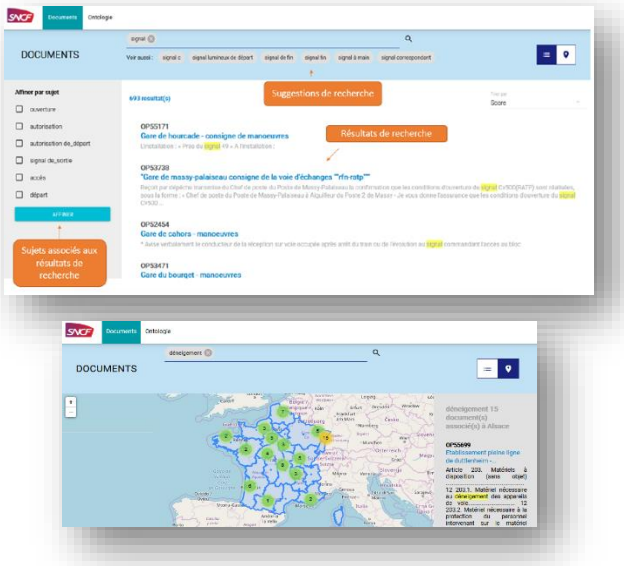
RECHERCHE D'INFORMATION DANS LES TEXTES DE PRESCRIPTION

APPROCHE ONTOLOGIQUE POUR LA RECHERCHE SÉMANTIQUE, LA VISUALISATION CARTOGRAPHIQUE ET LA NAVIGATION PAR GRAPHE DE CONCEPTS (VISEO RECHERCHE ET INNOVATION)

VISEO | POSITIVE DIGITAL MAKERS

- **Création de ressources lexicales :** 3 ontologies

- **Plusieurs applications**
- **Autocomplétion** sémantique
- **Tri par pertinence** : score BM25
- **Suggestions de recherche** issues des relations entre concepts
- **Suggestions de recherche** algorithmiques (score BM25)
- **Recherche cartographique**, avec extraction de lieux



SNCF SÉCURITÉ

SNCF

RECHERCHE D'INFORMATION DANS LES TEXTES DE PRESCRIPTION

APPROCHE ONTOLOGIQUE POUR LA RECHERCHE SÉMANTIQUE, LA VISUALISATION CARTOGRAPHIQUE ET LA NAVIGATION PAR GRAPHE DE CONCEPTS (VISEO RECHERCHE ET INNOVATION)

VISEO | POSITIVE DIGITAL MAKERS

- **Evaluation de l'utilité et de l'utilisabilité**
- 12 testeurs, 40'-1h
- Observations, questionnaire, débriefing

- **Résultats**

± **Suggestions issues de l'ontologie** : utiles, intéressantes, des réserves sur la pertinence

× **Suggestions algorithmiques** : non pertinentes (trop génériques)

✓ **IHM et navigation** très appréciées, intérêt opérationnel pour la navigation cartographique

× **Projection sur l'industrialisation** > réserve : maintenance / maintenabilité de l'ontologie ?

ANALYSE DE RETOURS D'EXPERIENCE PAR APPROCHE ONTOLOGIQUE

CONTEXTE ET OBJECTIFS

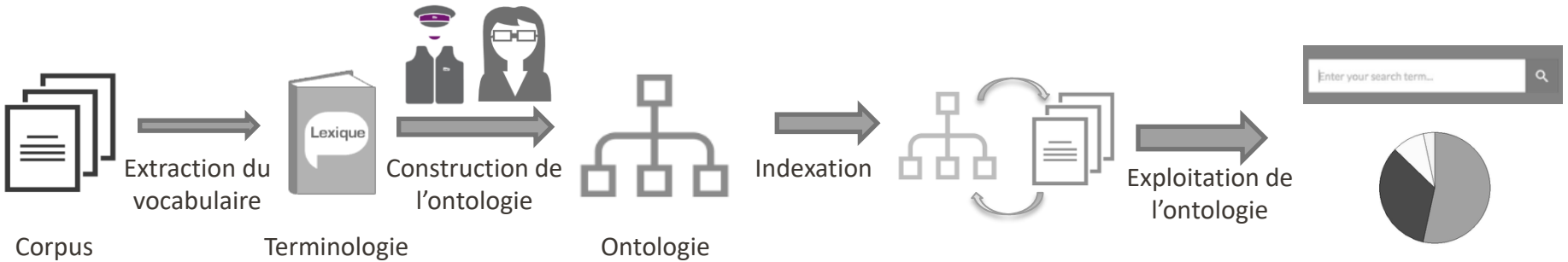
- **Analyse des REX (retours d'expériences) d'événements sécurité SNCF**
- 115 00 événements sécurité depuis 1992, ~100 facteurs (FOH)
- Refonte d'un des outils de REX de SNCF

- **Preuve de concept : améliorer la recherche d'information grâce à des ontologies sur les facteurs d'influence/causes**
- Rechercher et identifier des facteurs d'influence
- Faire émerger des relations causales
- Catégoriser les causes

ANALYSE DE RETOURS D'EXPERIENCE PAR APPROCHE ONTOLOGIQUE

DÉMARCHE

- **Démarche collaborative Direction de la Circulation Ferroviaire SNCF / société Aboutgoods**



ANALYSE DE RETOURS D'EXPERIENCE PAR APPROCHE ONTOLOGIQUE

DÉMARCHE

- **Interface avec fonctionnalités de recherche et analyse**
- Sémantiques
  - Saisie multi-concepts
  - Recherche sémantique
  - Suggestion de termes et affinage de la recherche
- Statistiques
  - Co-occurrences
  - Nuages de tags
- Schémas causaux

ANALYSE DE RETOURS D'EXPERIENCE PAR APPROCHE ONTOLOGIQUE

RÉSULTATS

ABOUT GOODS

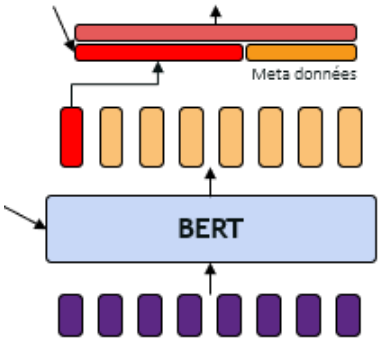
- **Principaux apports**
- Création d'une **ressource sémantique réutilisable**
- Capacité à capter les concepts, **pertinence et précision des résultats** (mais rappel à améliorer)
- **Recherche d'information et interprétation facilitées**
- **Limites / difficultés**
- **Analyse sémantique incomplète** (ex : négation) : complémentarité avec les approches syntaxiques
- Rappel et précision : **pas d'évaluation stricte** (pas de ressource annotée de référence)
- Manque de **fiabilité des résultats statistiques**
- Plue-value du **nuage de tags** faible (**interprétation** des résultats + termes les plus saillants trop génériques)
- **Maintenance** de la base : délicate
- **Expertise et disponibilité métier insuffisantes** pour pousser les approches ontologiques

SNCF SÉCURITÉ

SNCF

INTERPRÉTATION DE REX : ÉVALUATION DE MÉTHODES TAL RÉCENTES

- **Evaluation de méthodes TAL actuelles pour l'analyse de Retours d'Expérience** : Transformers (plongements lexicaux de textes contextualisés)
- **Etat de l'art, analyse des cas d'usage, expérimentation sur un jeu de données réel (réduit)**
- **Comparaison de méthodes standards (tf-idf, Word2Vec) à des modèles basés sur l'architecture Transformers**

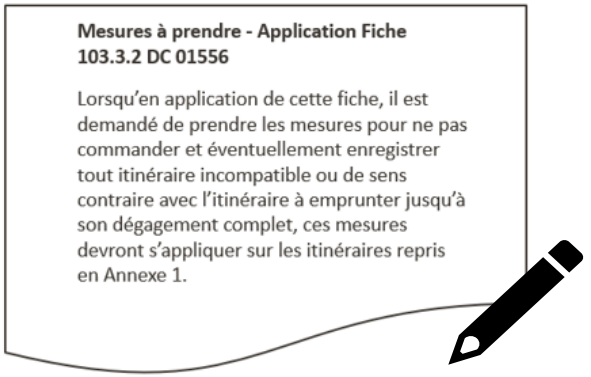


- **Evaluation difficile**
- **Remise en question des données** en entrée (justesse de la labellisation),
- Recommandation de **spécifier le modèle sur un corpus métier**, besoin d'un **Gold Standard**
- Des **conclusions fragiles**, plutôt en faveur des modèles de la famille de BERT, mais dans un cadre d'étude (durée, disponibilité…) qui n'a pu aboutir à des résultats fortement étayés
- Des besoins d'approfondissement, **pas de preuve**

AIDE À LA QUALITÉ RÉDACTIONNELLE

CONTEXTE ET OBJECTIFS

- **Programme de transformation documentaire**
- Améliorer la qualité de rédaction
- Documentation opérationnelle plus claire



- **Expérimentation d'un outil d'aide à la qualité rédactionnelle**
- Evaluation du potentiel technologique et d'usage
- Objectivation de nos standards et pratiques de rédaction à partir des analyse outil

isChecker
Instruction Sheets Checker

SAFETY DATA
UNE ENTITÉ DU GROUPE OmniContact

AIDE À LA QUALITÉ RÉDACTIONNELLE

DÉMARCHE

SAFETY DATA
UNE ENTITÉ DU GROUPE OmniContact

- **Outil isChecker de la société Safety Data (groupe Omnicontact)**
- Traitement de **documents techniques**
- Modélisation linguistiques de **règles de rédaction technique** (guides internationaux & client)
- **Moteur d'analyse** textuelle et de vérification
- Sorties : **documents annotés** en écarts linguistiques, **rapport** d'analyse (excel)
- **Démarche itérative et progressive avec la SNCF**
- 4 étapes, test progressif
- **Forte implication des utilisateurs** : experts linguistes et métiers, dialogue/ateliers récurrents, analyse linguistique et métier des résultats
- Ajustement des règles et du paramétrage, prise en compte des spécificités textuelles de SNCF

AIDE À LA QUALITÉ RÉDACTIONNELLE

RÉSULTATS

SAFETY DATA
UNE ENTITÉ DU GROUPE OmniContact

- **Evaluation de l'outil : positive**
- Pertinence des résultats
- Utilité et efficacité de l'outil confirmées
- Nécessaire complémentarité outil – humain – formation

- **Difficultés et points d'attention**
- **Contextes et utilisateurs variés** : la pertinence dépend du métier, du contexte d'usage, du profil utilisateur
- B**esoin d'accompagnement** important, **interprétabilité** des résultats essentielle, **usage de l'outil à encadrer** (effet psychologique, compétences, cadrage organisationnel)
- **Processus progressif**: itérations, inscription dans la durée et dans le dialogue experts métiers & linguistes

TERMINOLOGIE DE RÉFÉRENCE

- **Plusieurs démarches autour de la terminologie SNCF**
- Désambiguïsation d'acronymes (projet Polysemy – collaboration LIUM / DTIPG SNCF)
- Mise en qualité d'une base d'acronymes collaborative (travaux de stage de **Morgann Sabatier**)
- Normalisation : constitution d'une terminologie de référence

- **Difficultés récurrentes**
- **Evaluation complexe** : connaissance experte partielle, disponibilité des experts, polysémie des acronymes
- Normes linguistiques vs normes législatives ; sujets de **gouvernance**
- **Evolutivité de la terminologie** : importante (contexte de réorganisation)

ANALYSE DU WEB POUR DES OPÉRATIONS À L'INTERFACE AGENT/CLIENTS

CADRE

- **Objectif :** exploiter les messages du web pour mieux comprendre certaines pratiques client

- **Sortie :** plateforme avec modules de collecte web et analyse sémantique, évaluation de la solution

- **Verrous scientifiques**
- Contextualiser les messages (lieux, dates…)
- Interpréter intelligemment les données

ANALYSE DU WEB POUR DES OPÉRATIONS À L'INTERFACE AGENT/CLIENTS

DÉMARCHE

synomia

- **Collecte web :** sources Twitter, forums d'actualité, forums, 1 site dédié => 200 000 messages

- **Analyse en lieux et modes de transport**
- par mots-clés sur les stations, gares, villes, lignes, objets liés à un territoire (ex : Navigo)
- Découpage IDF / Bretagne, puis affinement sur l'IDF (détection de lignes)
- Modes identifiés grâce à des termes génériques (« train », « bus ») ou désignations spécifiques (« RER C », « la 14 »)

SNCF SÉCURITÉ

SNCF

ANALYSE DU WEB POUR DES OPÉRATIONS À L'INTERFACE AGENT/CLIENTS

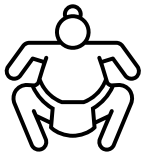
RÉSULTATS

- **Bilan global**
- Web largement balayé
- Contenus **pertinents** mais **peu informatifs**
- Des **résultats** à confirmer par un **retour au texte**

- **Analyse en modes et lieux : une tâche difficile**
- Noms ambigus et erreurs d'affectation *Ex : République*
- Identification des lignes, stations ou lieux d'Ile-de-France : **notions liées**, besoin de **règles de déduction** *e(x : RER direction Saint-Denis ➔ RER B ou RER D)*

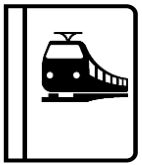
LIMITES RÉCURRENTES

LIMITES DES DONNÉES MÉTIERS



Volumétrie

Qualité des données



Spécificités métiers

Disponibilité et confidentialité des données

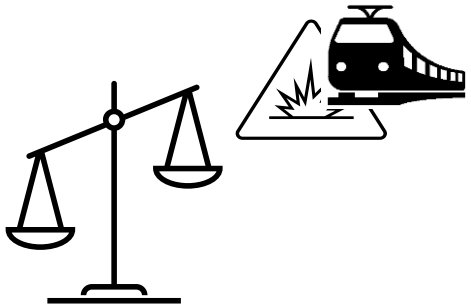
- **Gouvernance et agrégation des sources métiers : en cours, processus long**
- **Les données sont… données en entrée, faire avec : analyse de corpus & adaptation du choix des algorithmes**

LIMITES DES RESSOURCES LEXICALES

- **Des ressources contextuelles dans l'entreprise**
- **Une disponibilité et un partage interne des ressources lexicales limités**
- **Une qualité des ressources lexicales variable**
- **Des initiatives en cours et des ressources en construction**
- **Ressources vivantes : quelle tenue à jour, quelle maintenance ? Compétences & organisation nécessaires**

EXIGENCES DE SÉCURITÉ ET PERFORMANCES DU TAL

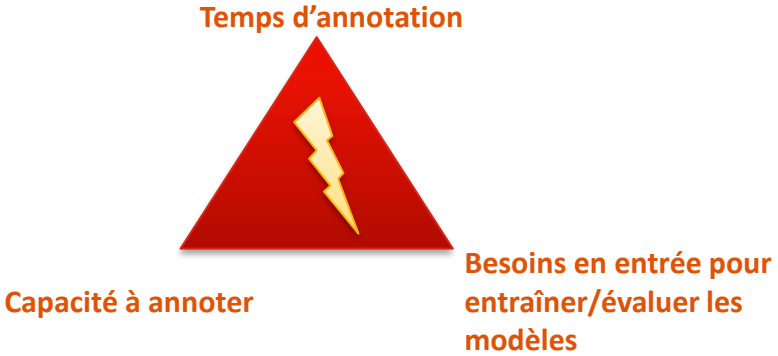
F-mesure



« L'intelligence artificielle n'est pas sécuritaire »

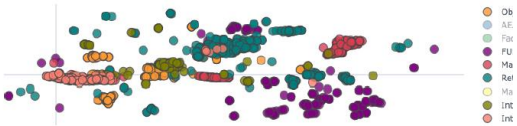
- **Maturité des algorithmes et traitements : à l'aune des enjeux sécurité et des cas d'usage associés**
- **TAL comme une aide à la décision, garder un contrôle : allier TAL et ergonomie du dialogue, complémentarité outil / humain**

EVALUATION DES DONNÉES



- **Annotation des données**
- Corpus annotés : coût, rares, pas la volumétrie.
- Tâches d'annotation : à cadrer, IHM / formulaire de saisie simple et clair
- Ressources dédiées : experts peu disponibles, manque de connaissance experte pour des annotateurs tiers
- **Connaissance experte : pas universelle, des divergences**
- Accords inter-annotateurs faible

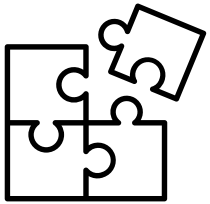
INTERPRÉTABILITÉ ET EXPLICABILITÉ



- **Modèles « boîtes noires » :** difficulté à expliquer les sorties (compréhension profonde des algorithmes et erreurs)
- **Représentations machine learning :** interprétations source d'ambiguïté, parfois erronées
- **Ecart entre l'expertise machine learning / TAL et l'expertise métier**

- **Internalisation des compétences : des initiatives en cours ; processus long**
- **Travailler ensemble : combiner les compétence (linguistique, data, métier), dialoguer, inscrire la collaboration dans le temps**

IA, LINGUISTIQUE, ETAT-DE-L'ART: QUELLE ANALYSE TEXTUELLE ?



- **Les algorithmes à l'état-de-l'art ne sont pas toujours les plus efficaces pour les cas métiers**
- **Une culture & maîtrise d'anciennes méthodes qui gagnerait à être développée (contexte de « code sur étagère »)**

- Combiner les approches linguistiques et d'intelligence artificielle
- Interprétation locale & globale, retour au texte
- Combiner les compétences, dialoguer (métier / expert TAL)
- Capitaliser sur les outils et méthodes passés

Merci pour votre attention.

Des questions ?

RÉFÉRENCES BIBLIOGRAPHIQUES

- Blatter, C., Tonnerre, P., Donnet, S., Reutenauer, C., & Million-Rousseau, C. (2018, October). TRAITEMENTS LINGUISTIQUES POUR LA RECHERCHE D'INFORMATION ET L'ANALYSE EN FOH DE REX FERROVIAIRES. In *Congrès Lambda Mu 21,«Maîtrise des risques et transformation numérique: opportunités et menaces»*.
- REUTENAUER, Coralie, LEFEUVRE, Luce, FOUQUERAY, Aurélie, *et al.* Technologies sémantiques et accès à l'information dans le prescrit SNCF. In : *Congrès Lambda Mu 22 «Les risques au cœur des transitions»(e-congrès)-22e Congrès de Maîtrise des Risques et de Sûreté de Fonctionnement, Institut pour la Maîtrise des Risques*. 2020.
- DUGUE, Nicolas, CAMELIN, Nathalie, LEFEUVRE, Luce, *et al.* Apprentissage et évaluation de plongements lexicaux sur un corpus SNCF en langue spécialisée. In EGC 2019, pp.279-284