RÉPUBLIQUE FRANÇAISE

Liberté
Égalité
Fraternité

INERIS

maîtriser le risque
pour un développement durable

L'IA POUR RECONNAÎTRE DES EMPREINTES CHIMIQUES ?

09/06/2023

Ineris - 166139 - 2772091 0.1

Sommaire

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

INERIS
maîtriser le risque pour un développement durable

1. Position du problème

Des substances chimiques, encore des substances

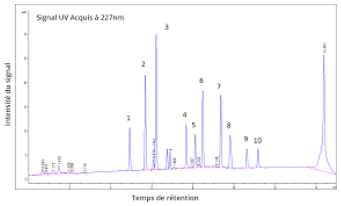
Quelques chiffres

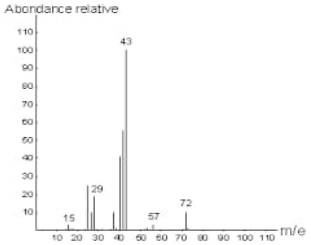
- 100 Millions de substances enregistrées dans le CAS*
- 4000 de plus … chaque jour !
- 30 000 à 50 000 substances pourraient se trouver dans les produits du quotidien

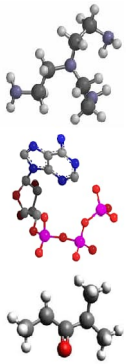
Des substances et des méthodes

Pour caractériser des substances, il faut

- Des prélèvements d'échantillons,
- Des méthodes d'analyses,
- Des exploitations de résultats







Tout ça prend beaucoup de temps et parfois …

- Il y a urgence (Lubrizol, Notre Dame de Paris…)
- On aimerait prévenir le risque plutôt que subir les conséquences

*Chemical Abstract Service

Des données, encore des données

Sources des données

- Données de mesures en chromatographie et spectrométrie de masse (Ineris, Massbank, ACSM…)
- Données de constructeurs de matériels métrologiques
- Données de contexte environnemental
- Données de la littérature (rapports Ineris, US EPA…)

Tout n'est pas sous Excel…

- Des formats complexes et difficiles à traiter par l'humain (MzML)
- Des données nichées dans des rapports et même dans des images
- Des données dans des formats propriétaires ou peu standards

Exemple de fichier au format mzML

L'IA pour nous aider à caractériser les substances

Les méthodes d'IA semblent intéressantes, ça …

- Ressemble à un problème d'apprentissage supervisé
- Traite des volumes importants de données
- Nécessite des temps limités de développement d'outils de calcul (vs. des modélisations physiques)
- Commence à être à la mode

Quelques faits à prendre en compte

- Les équipes expertes de la chimie de l'environnement ne sont pas datascientists
- Les équipes sont déjà bien chargées
- Avant d'investir dans l'IA, il faut tester

Que recherchions-nous ?

- Automatiser le processus de caractérisation des substances chimiques ou de leurs sources de contamination
- Augmenter et objectiver nos capacités d'expertise

Et au passage…

- Structurer & valoriser nos données
- Apprendre à exploiter l'IA

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

INERIS
maîtriser le risque pour un développement durable

2. Le POC « Empreintes environnementales »

Fonctionnalités de l'outil développé

Besoin #1

Obtention d'un score sur la probabilité d'identification d'une substance dans un échantillon

- Onglet **base de données** : affichage des échantillons ayant été intégrés dans la base, avec la possibilité d'intégrer de nouveaux échantillons et de réaliser une analyse
- Onglet **analyses** : recensement des analyses ayant été lancées, avec les paramètres associés, ainsi que les couples (échantillons, analyses)
- Onglet **résultats** : visualisation des résultats obtenus avec un accès aux scores définis par l'algorithme
- Onglet **dashboard** : visualisation des résultats détaillés par couple (échantillon, molécule)

Besoin #2

Obtention d'éléments d'analyse et de comparaison objectifs qui permettent d'aider dans l'analyse et l'identification des sources de pollution dans un échantillon

- Onglet **exploration** : exploration de la base de données INERIS à travers ses différents degrés de liberté (projets, distance à la source, matrices et type d'exposition)
- Onglet **analyse** : mise en relief des données de la base INERIS pour obtenir des éléments de comparaison entre les différents échantillons qui composent cette base

Besoin #3

Détermination et visualisation en temps réel des sources de pollution à partir d'un échantillon d'air

- Onglet **dashboard** : visualisation de nombreux résultats graphiques et quantitatifs sur une période d'étude définie avec une modulation possible des graphes via les filtres
- Onglet **échantillon** : blacklistage des résultats aberrants et visualisation de l'état des tâches de chargement des données ACSM temps réel du SIRTA

Organisation des développements

2019
- AMI IA#2
- Cahier des charges
- Sélection du prestataire
- Lancement des travaux

2020
- Collecte données
- Ateliers experts métier / datascientists
- Sprints de développement
- Ateliers d'acculturation à l'IA

2021
- France Relance ITN5
- Expérimentation
- Présentations

5-6 SIAPARTNERS

6-8 Experts

3-4 Accompagnants

Ce qu'on en retire

Du point de vue métier

Des traitements de données plus fiables, plus exhaustifs et plus rapides

Un processus de caractérisation plus objectif : fixation de seuils, calculs de scores

Sensibilisation à la gestion des données en situation post-accidentelle notamment : structuration, gouvernance…

Du point de vue applicatif

Une application rend les résultats partageables entre les acteurs

Il faut toujours bien garder à l'esprit que l'application reste un démonstrateur (moins frustrant)

Une application d'IA doit embarquer des fonctions type :

- apprentissage constant *,
- traçabilité des modèles
- …

Du point de vue organisationnel

La mise en œuvre de l'IA permet :

- S'acculturer au métier pour le datascientist et réciproquement
- Catalyser l'ensemble du cycle de vie des données

*Active learning

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

INERIS
maîtriser le risque pour un développement durable

3. Et maintenant

Une application à industrialiser (besoin #1)

Développer une application dans les règles de l'art

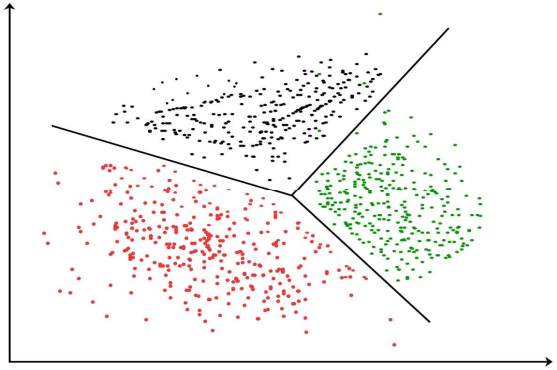
- Revoir le parcours utilisateur
- Traiter les questions de sécurité

Maintenir l'application en conditions opérationnelles

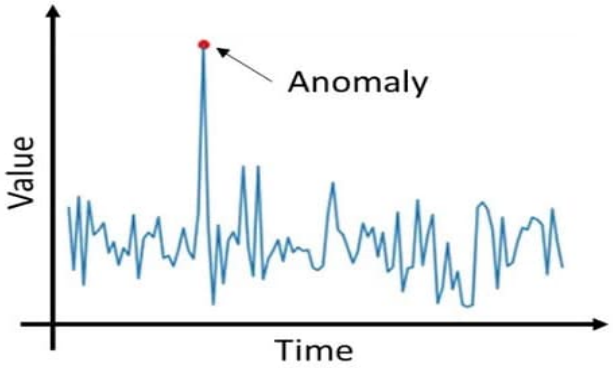
- Penser gouvernance des données : gestion des données d'apprentissage, active learning, traçabilité des résultats…
- Suivre les performances : temps de calcul, taux d'erreurs…
- Garantir son financement

On pourrait imaginer aller plus loin

Clustering / groupes de substances



Recherches d'anomalies / suivi de sources



Utilisation de modèles prédictifs

PubChem