Human In the Loop for Data Mining and Machine Learning

Thi Bich Hanh Dao and Arnaud Soulet

Groupe de travail du RTR DIAMS

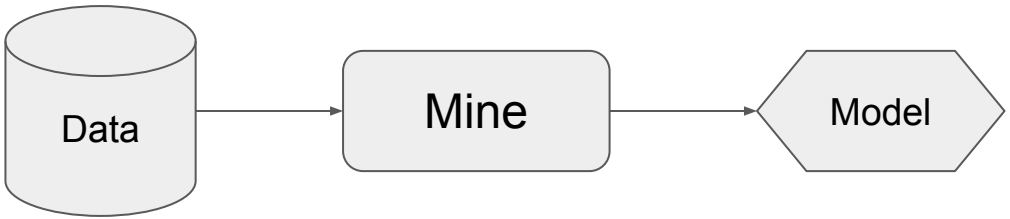
Axe 3 : Apprentissage, Optimisation et Aide à la décision

Mots clefs : apprentissage semi-supervisé, fouille de données, aide au diagnostic, aide à la conception de médicament

Equipes partenaires :

- Imagerie et cerveau (UMR1253), Université de Tours
- Institut de Chimie Organique et Analytique (UMR7311), Université d'Orléans
- Laboratoire d'Informatique Fondamentale et Appliquée de Tours, Université de Tours
- Laboratoire d'Informatique Fondamentale d'Orléans, Université d'Orléans

Data mining and Machine Learning



Amyotrophic lateral sclerosis (ALS)

Incidence : 2.5/100 000 in France (M/F 1,3-2)

Median age at onset: 55 years

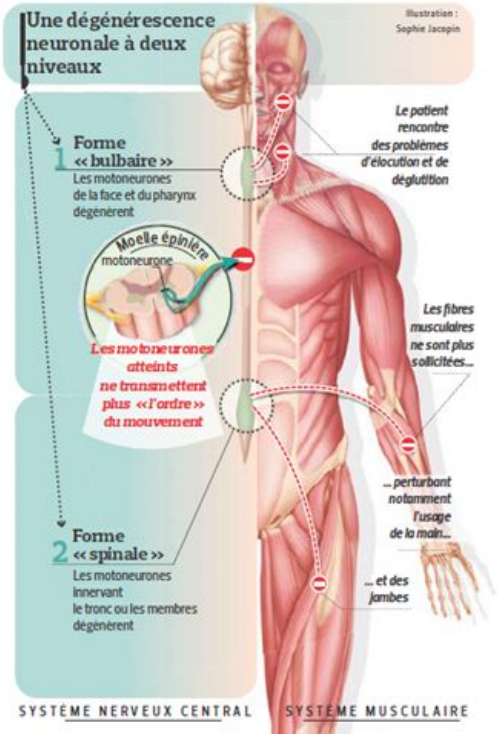
Median of disease duration: 36 months

Heterogeneous disease

Diagnostic delay : 9-12 months

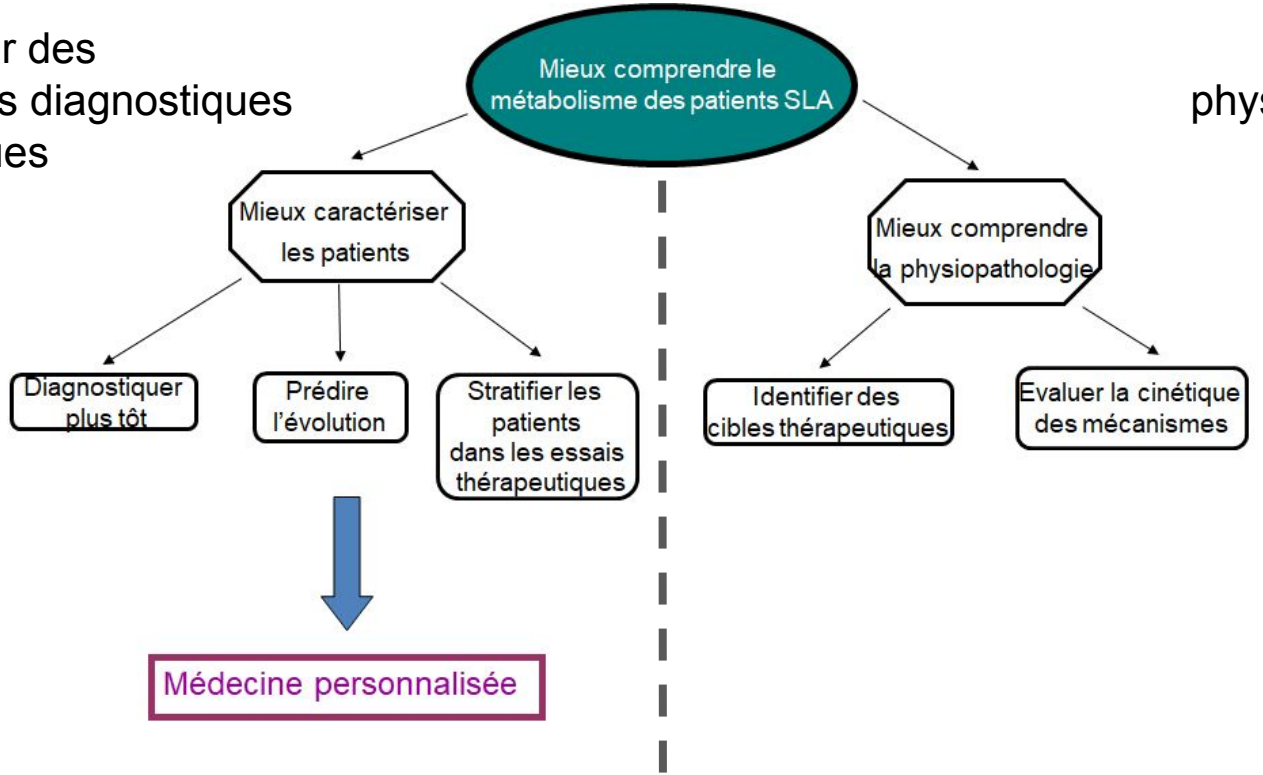
Only one drug

source: Hélène Blasco



Amyotrophic lateral sclerosis (ALS)

1-Rechercher des biomarqueurs diagnostiques et pronostiques

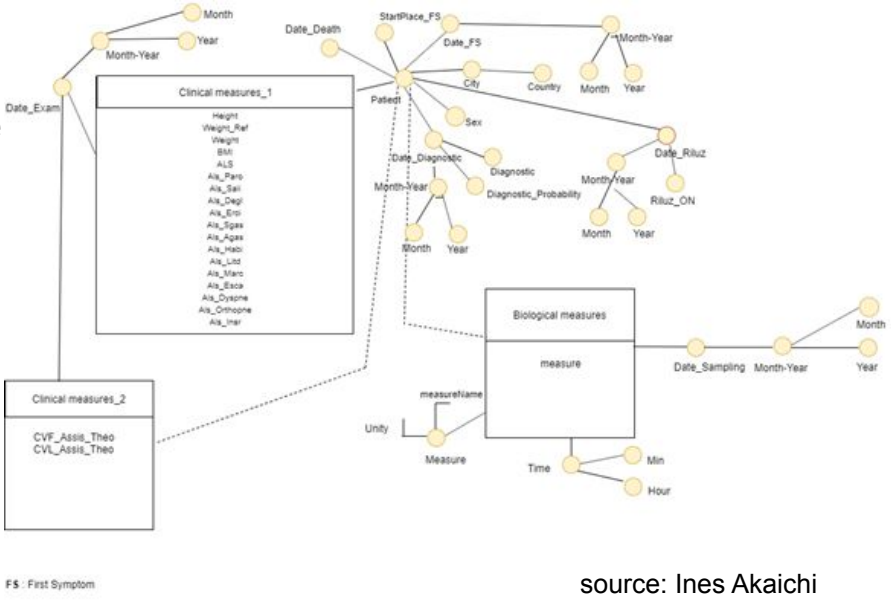


2-Explorer la physiopathologie

source: Hélène Blasco

Amyotrophic lateral sclerosis (ALS)

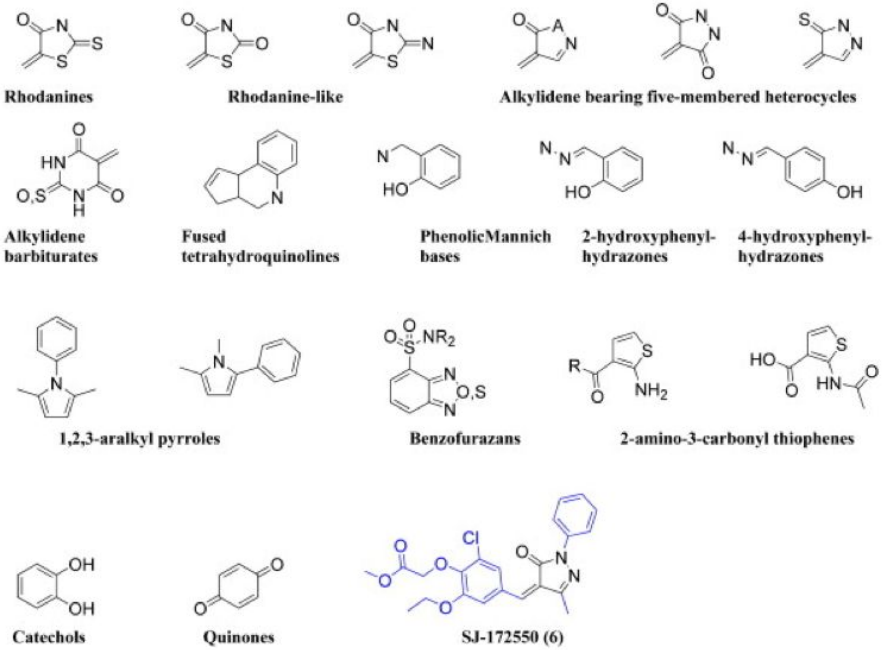
- Données
  - Données sur le patient : sexe, date du diagnostic, etc
  - Données cliniques : poids, hauteur, score ALSFRS, mesures respiratoires, etc
  - Données biologiques
- Défis
  - Données longitudinales
  - Censure à droite des données



source: Ines Akaichi

Chemical data

Pan-assay interference compounds (PAINS)



From Davis, B.J. & Erlanson, D. A., Bioorg. Med. Chem. Lett., 2013

Chemical compounds often giving false positive results in high-throughput screens.

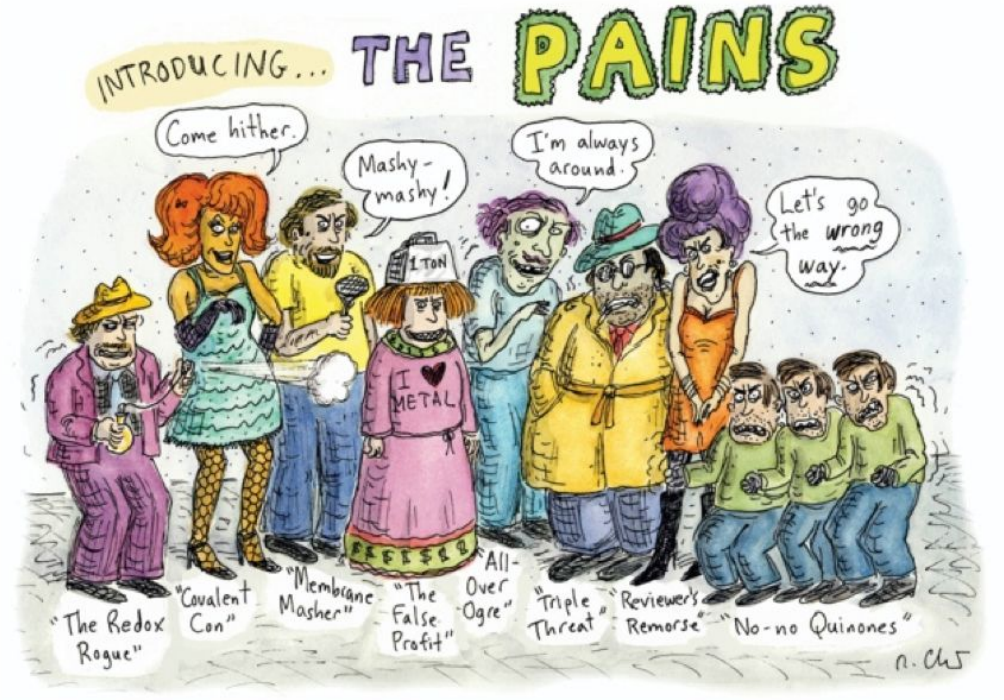
Substructures and compounds classes identified in the literature are most probable to provide PAINS. (1)

Filters encoding PAINS used to detect potential problematic compounds but not really efficient.(2)

New methodologies need to be developed for a better prediction.

1) J. B. Baell; G. A. Holloway. J. Med. Chem. 2010, 53, 2719−2740
2) J. B. Baell; J. W. M. Nissink. ACS Chem. Biol. 2018, 13, 36−44

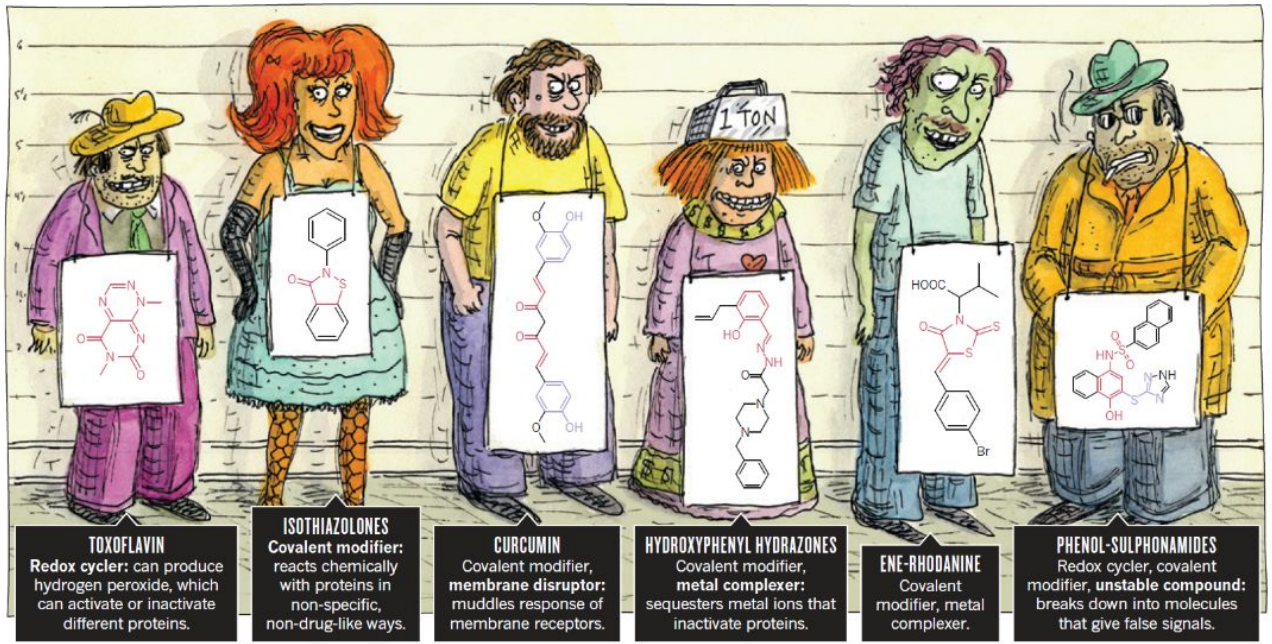
PAINS (A Pan Assay Interference Compounds)



Nature Comments 2014, 513, pp 481

Samia Aci-Sèche

PAINS (A Pan Assay Interference Compounds)



Nature Comments 2014, 513, pp 481

Samia Aci-Sèche

PAINS (A Pan Assay Interference Compounds)

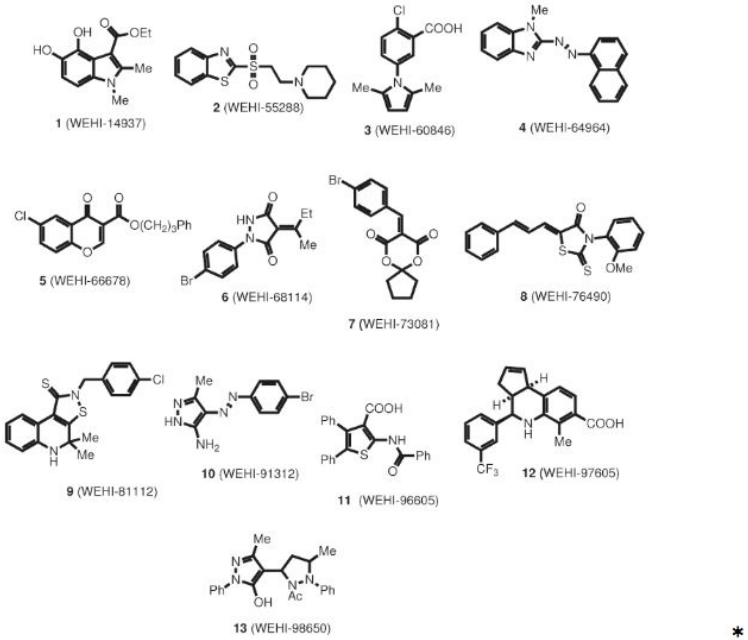


Figure 2. Problematic cul de sac compounds that have incurred wasted resources through being followed up to varying degrees at our Institute. We have found chromones such as **5** to be highly susceptible to nucleophilic attack at the 2-position, while β-amino sulfones (and ketones) such as **2** readily form reactive retro Michael alkenes. Compounds **6**−**9** are also susceptible to attack by biologically relevant nucelophiles. The other compounds are problematic for reasons that are either discussed in the text or remain unknown.

*Pfizer

J. Med. Chem., 2010, 53 (7), pp 2719–2740

Samia Aci-Sèche

Chemical data

Dataset

~23,000 extensively tested compounds containing 270 PAINS substructures.[1,2]
(https://www.zenodo.org/record/557207)

Descriptors:

- Molecular descriptors
- Molecular fingerprints
- Molecular graph
- Other?

Human intervention

Suspicious PAINS need to be submitted to the chemist expert eye to orient the decision depending on the biological test realized and the chosen target

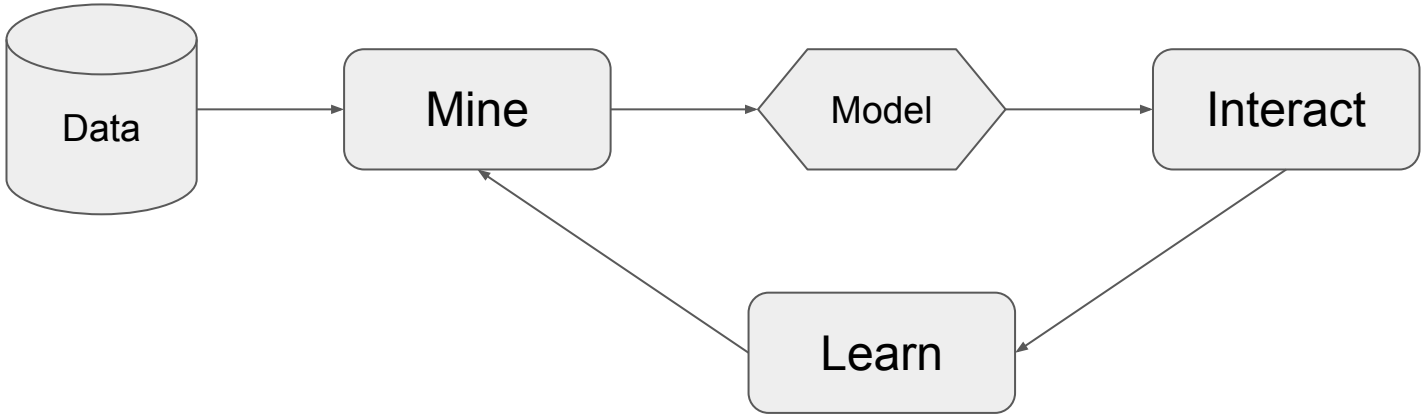
1) S. J. Capuzzi; E. N. Muratov; A. Tropsha. J. Chem. Inf. Model. 57 (2017) 417-427.
2) S. Jasial; Y. Hu; J. Bajorath. J. Med. Chem. 60 (2017) 3879-3886.

Semi-supervision avec apprentissage actif

“Je ne sais pas ce que je cherche, mais je pourrais le savoir si je le vois”

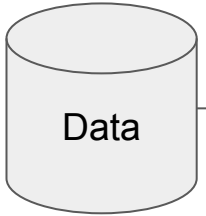
- L’utilisateur peut corriger le résultat de la fouille ou de l’apprentissage.
  - “Ce patient devrait être dans le groupe B plutôt que le groupe A”
  - “Ce composant moléculaire ne devrait pas faire partie du PAINS”
  - …
- L’utilisateur peut noter le résultat de la fouille ou de l’apprentissage
  - “Ce composant moléculaire n’est pas intéressant pour notre problème” (retour binaire)
  - “Le composant moléculaire A est plus intéressant que le composant moléculaire B” (retour gradué)
  - ...

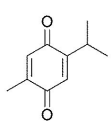
Human in the loop

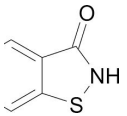


Pattern mining

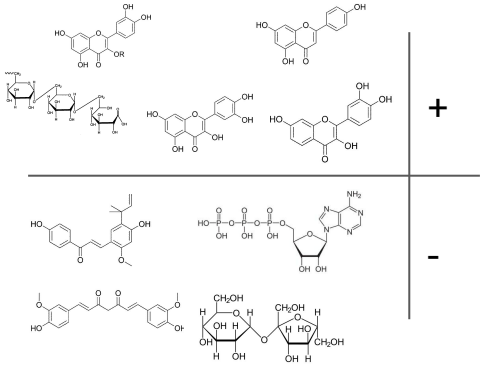
Frequent pattern mining [Agrawal et al 1993]
Emerging minng pattern mining [Dong et Li 1999]
...





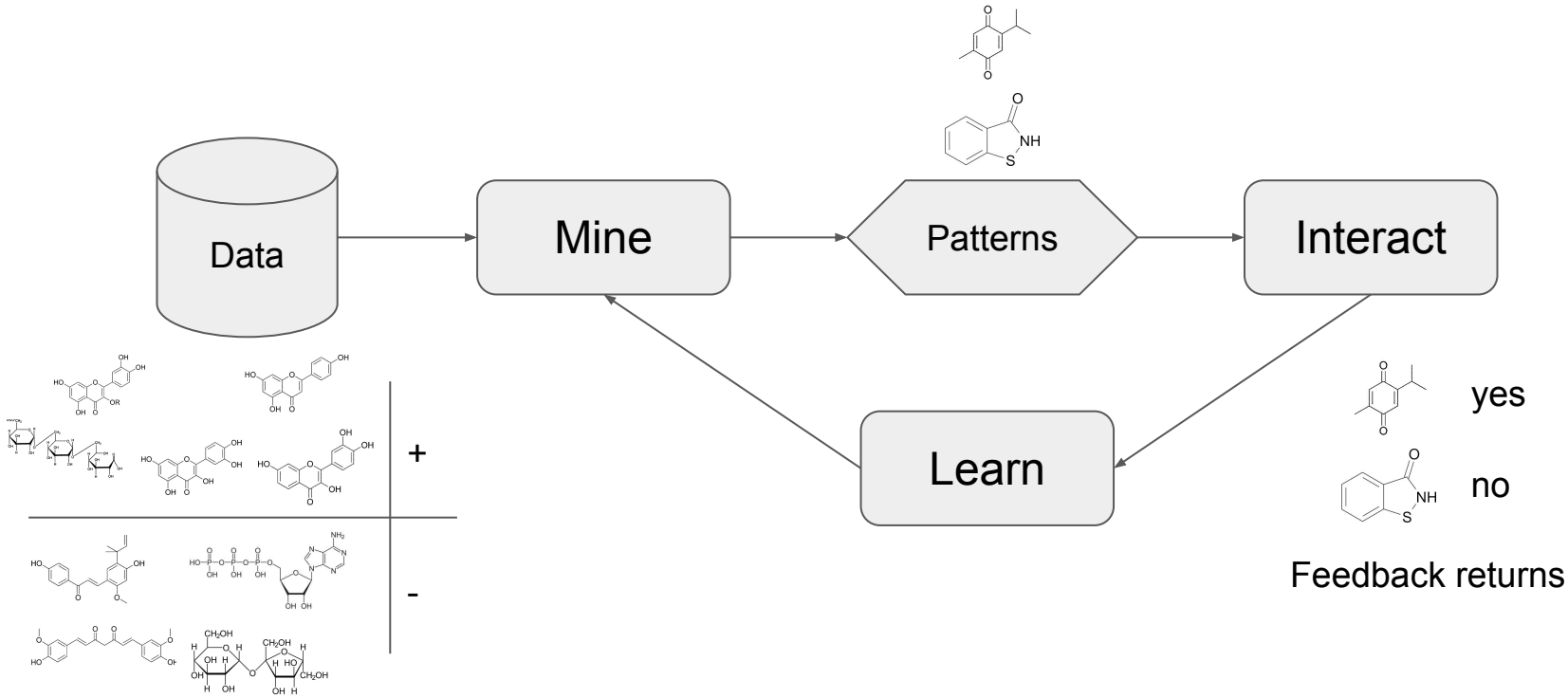


...

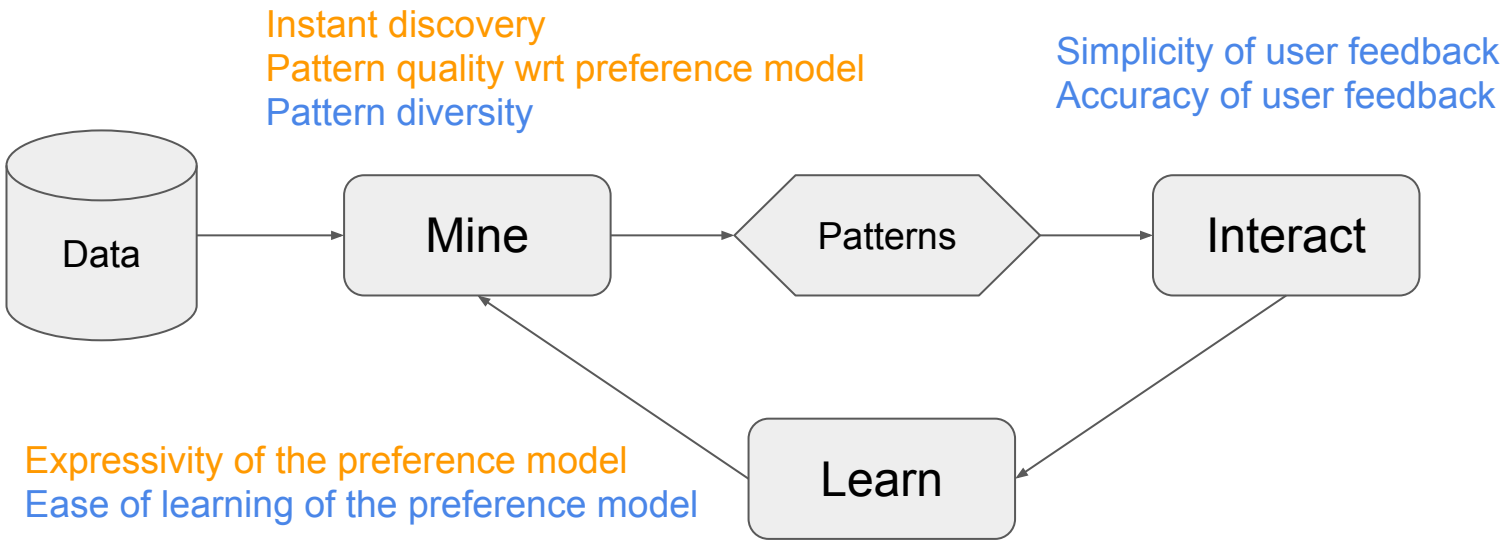


Issues: Too many patterns, trivial patterns, nonsense patterns, etc
⇒ Need of expert knowledge

Interactive pattern mining [van Leeuwen 2014]



Interactive pattern mining: main challenges



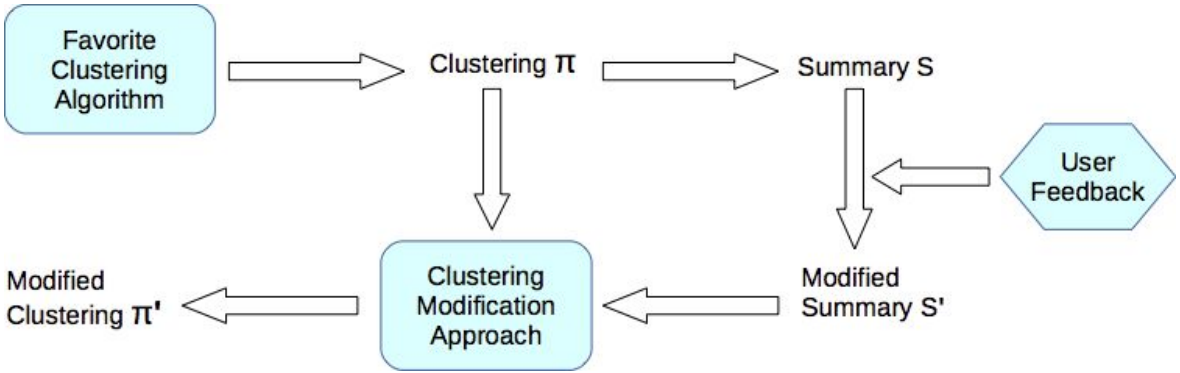
How to combine **pattern mining** and **active learning**?

Constrained clustering

- Clustering: partition the data into groups with respect to a distance/similarity metric
  - K-means, spectral clustering, density based clustering, …
- Constrained clustering: integrating prior knowledge into the clustering process by mean of constraints
- pairwise constraints must-link/cannot-link (semi-supervised clustering)
  - COP-Kmeans, LCVQE, ...
- constraints on clusters: diameter, size, density, …
  - actionable clustering
  - general and declarative frameworks, using SAT, ILP, CP
- Constraints are known before the clustering process

Active and interactive clustering

- Active clustering: identifying constraints during the clustering process
  - identify most informative instances / pairs of instances
  - query to determine must-link / cannot-link constraint
  - restrict to pair-wise constraints
- Interactive clustering: taking into account user feedback
  - splitting/merging clusters: user feedback on pair of instances
  - minimal modification clustering: user feedback can be on instances or on cluster



Défis scientifiques

- Combiner apprentissage semi-supervisé avec l’apprentissage actif
- Produire de nouveaux modèles en un temps raisonnable
- Identifier des contraintes plus générales que contraintes sur deux instances
- Identifier des sous-groupes et des marqueurs biologiques pour la SLA
- Identifier des Pan-assay interference compounds

Actions passées

- Stage de master (2019) : construction d'un entrepôt de données pour la SLA (iBrain/LIFAT/LIFO)
- Projets CNRS Mastodons (2017-2018) sur la détection de PAIN (ICOA/LIFAT/LIFO)
- Projet régional GIRAFON (2016-2018)
- ...

Objectifs du GT

rassembler des chercheurs autour de cette thématique innovante

Actions envisagées :

- Organisation de journées
- Montage de projet région ou ANR
- Montage de workshop d’une journée
- Co-encadrement de stages
- Co-encadrement de thèse

Human In the Loop for Data mining and Machine Learning

Prochaine réunion

Mercredi 11 décembre à 13h30 à Blois

Contact

- Thi Bich Hanh Dao thi-bich-hanh.dao@univ-orleans.fr
- Arnaud Soulet arnaud.soulet@univ-tours.fr