FICHE PROJET EUROPEEN			
ACRONYME : SCHISM			
NOM COMPLET DU PROJET	Supporting chemoinformatics via interactive unsupervised and semi-supervised data mining (SCHISM)		
NUMERO DE CONVENTION	20^{E}05838		
DATE DE DEBUT	01/01/2021		
DATE DE FIN	31/12/2022		
COORDINATEURS	Laurent HEUTTE		
• *Etablissement(s)*	• *Laboratoire(s)*	• *Responsable(s)*	• *Partenaire(s)*
URN	LITIS		GREYC
			CERMN
CONTACT	laurent.heutte@univ-rouen.fr		
SITE INTERNET DU LABORATOIRE ET PROJET			

DESCRIPTION DU PROJET	
RESUME	L'objectif du projet SCHISM est de développer une approche robuste de fouille de données interactive et de fournir un prototype permettant aux utilisateurs de lancer des algorithmes d'extraction ou de regroupement de motifs, de visualiser les résultats, de donner un retour d'information et de relancer les opérations d'extraction en tenant compte de ce retour d'information. Notre hypothèse de recherche peut être exprimée sans détours : en remettant l'utilisateur dans la boucle, on peut obtenir des résultats mieux compris et des processus d'extraction qui s'exécutent en fait plus rapidement que dans le cas non interactif. Les premières recherches sur l'apprentissage automatique (ML) et l'exploration de données (DM) ont tenté d'automatiser entièrement les processus de découverte des connaissances et de réduire les interventions humaines. Pour de bonnes raisons : nous avons du mal à gérer de grandes quantités de données (de haute dimension), une tendance à voir des modèles partout et le progrès technique nous a toujours soulagé des tâches chronophages. Cela motive également les travaux actuels sur le réglage automatique des paramètres [26]. Bien que cela puisse bien fonctionner dans des environnements supervisés où l'étiquette à prédire donne des commentaires et un modèle de boîte noire qui fonctionne bien pourrait être tout ce qui est nécessaire, ce consensus est de plus en plus remis en question dans la DM aujourd'hui. Alors qu'en ML, les systèmes basés sur la logique utilisent les connaissances de base depuis plus de 20 ans [27], et le clustering contraint ou semi-supervisé, qui utilise des connaissances supplémentaires pour guider le processus de clustering, a été proposé il y a une quinzaine d'années [14], Les méthodes utilisant l'intérêt subjectif dans l'exploration de modèles, c'est-à-dire les mesures de l'intérêt faisant intervenir les hypothèses des utilisateurs [1], sont relativement récentes. Les recherches sur les méthodes d'exploration de données interactives, qui permettent à l'utilisateur de donner son avis pendant le processus d'exploration - pas seulement avant et après - sont également récentes, sont également récentes pour modifier les stratégies d'exploration, réduire ou élargir les espaces de recherche, etc. [3]. Les raisons de ce changement sont multiples :

	1. Dans des contextes non supervisés tels que le clustering et l'exploration de modèles, les étiquettes sont par définition absentes et l'automatisation basée sur les informations d'étiquette est donc impossible. Pourtant, même dans de nombreux contextes de problèmes « supervisés » de la vie réelle, une grande partie des données peut ne pas être étiquetée et les étiquettes existantes peuvent ne pas être fiables, de sorte qu'un utilisateur se trouve au mieux dans un environnement semi-supervisé. 2. Dans un environnement non supervisé ou semi-supervisé, il est presque impossible pour les utilisateurs de préciser a priori leurs hypothèses, leurs attentes et leurs objectifs. S'ils réussissent, il est difficile de les traduire dans des langues de contraintes disponibles, quelque peu limitées. Le cadre actuel, dans lequel les utilisateurs définissent les paramètres avant l'extraction, filtrent et interprètent la sortie après, gaspille du temps (et de l'argent) et est contre-intuitif par rapport à la façon dont nous traitons les informations. Les utilisateurs peuvent cependant réagir aux résultats (partiels) et indiquer si ceux-ci sont d'accord avec leur intuition, semblent intéressants, etc. 3. Souvent, les experts doivent comprendre pourquoi les algorithmes produisent les résultats qu'ils produisent, car de grandes quantités d'argent ou de ressources sont en jeu, par exemple. Dans le développement de médicaments ou le déploiement d'infrastructures, ou même des vies en jeu, par ex. en médecine, en préparation aux catastrophes ou en milieu militaire. Ou ils veulent les comprendre parce que le DM non supervisé sert de générateur d'hypothèses : l'observation des résultats d'une opération d'exploration de modèles ou d'un clustering produit peut déclencher de nouvelles perspectives et éclairer de nouvelles directions de recherche - la dernière étape du processus de « découverte des connaissances » [15].

OBJECTIFS	Résultats attendus et livrables : Concrètement, nous étudierons et développerons les différents composants nécessaires à la construction d'un système minier interactif dans les différents composants nécessaires à la construction d'un système minier interactif dans les work packages 1 à 3. Ces résultats seront implémentés dans un prototype dont une partie sera basée sur l'outil Norns développé par le CERMN [13]. Les options de visualisation et de rétroaction seront séparées mais intégrées aux Norns, et les commentaires des utilisateurs seront traduits en modifications du flux de travail des Norns tel que défini à ce moment-là. L'interface pour l'interaction utilisateur impliquera les nouvelles options de visualisation et de rétroaction à développer dans WP 3. Les manipulations de l'utilisateur sur cette interface alimenteront les méthodes développées dans WP 1 et 2 pour être traduites en nœuds de flux de travail supplémentaires et / ou modifier les paramètres des paramètres.
IMPACTS ATTENDUS ET FINALITE	Les récents succès de l'apprentissage profond ont renouvelé l'intérêt pour le ML, et les secteurs public et privé prévoient d'utiliser les techniques de ML et de DM. Dans le même temps, les obstacles des périodes passées d'adoption du ML refont surface : les utilisateurs de domaines sensibles tardent à faire confiance aux nouvelles méthodes, en particulier si la vie humaine est en danger, et de mauvaises décisions peuvent rapidement éroder la confiance durement acquise. De plus, les législateurs envisagent de plus en plus la décision algorithmique : le RGPD européen, entré en vigueur cette année, donne aux citoyens concernés par de telles décisions le droit à une explication, et le FCRA américain exige des créanciers qu'ils offrent jusqu'à quatre critères clés utilisés dans les décisions de demande de prêt. SCHISM vise à faire un pas vers les outils et les connaissances nécessaires pour que les méthodes de décision algorithmiques jouent leur rôle dans la société de l'information, permettant aux chercheurs de tirer des connaissances exploitables, aux décideurs de baser les tâches de gestion sur des méthodes de confiance, et aux responsables de la conformité pour mettre les l'architecture en place. Le SCHISM étant un projet de recherche fondamentale, nous mesurerons son succès par l'influence de nos travaux sur les communautés de recherche concernées. L'expérience montre que la conception de méthodes pour des problèmes très complexes introduit de nouveaux défis, par ex. cela pourrait conduire à de nouvelles options de rétroaction, des contraintes ou des stratégies de recherche mieux adaptées, qui pourraient être adoptées par la communauté dans son ensemble. Les résultats seront diffusés à travers des publications dans des IA générales de haut niveau (IJCAI, ECAI...) Et des conférences spécialisées (CP, ECML / PKDD, ICDM, KDD...) Et des revues. Les résultats dérivés de données biologiques ou chimiques seront soumis à des revues à haute visibilité (J. of Chem. Info. And Modeling, J. of Cheminformatics, J. Med. Chem., Mol. Inf...) Et présentés à établi rencontres internationales axées sur la chimio-informatique. Nous organiserons deux ateliers en 2020 et 2021, sur les contraintes d'apprentissage à partir des commentaires des utilisateurs, l'analyse exploratoire des données et la chimio-informatique, et un atelier final de 2 jours. Généralisabilité des solutions développées Bien que nous ayons choisi un cas d'utilisation de la chimio-informatique, nous nous attendons à ce que nos travaux soient applicables à d'autres domaines. En clair, d'autres données présentées sous forme structurée bénéficieront des solutions que nous développerons et l'exploration de données non structurée pourra exploiter les résultats concernant les contraintes et les retours utilisateurs. Il est possible que des choix de visualisation particuliers soient plus spécifiques à l'intuition des experts en chimio-informatique, mais même dans ce cas, le processus de leur développement devrait guider les développeurs dans d'autres domaines.

RESULTATS	
MODALITES DE FINANCEMENT	BUDGET TOTAL : 116 000 €
Niveau de soutien FEDER / FSE / FAEDER	116 000 €
Niveau de soutien région	
Nombre de personnes travaillant sur le projet	1 ingénieur
L'Europe s'engage en Normandie avec le Fonds Européen de Développement Régional	