GDR Groupement de recherche
MaDICS Masses de données, informations et connaissances en sciences

GDR 3708 - MaDICS

Masses de Données, Informations et Connaissances en Sciences

Dossier de renouvellement pour la période 2025-2029

Équipe Projet MaDICS

Bernd Amann	LIP6 UMR 7606, Sorbonne Université
Khalid Belhajjame	LAMSADE UMR 7243, Université Paris-Dauphine
Frédéric Bimbot	IRISA UMR 6074, Rennes
Christophe Bobineau	LIG UMR 5217, Institut polytechnique de Grenoble
Sarah Cohen-Boulakia	LISN UMR 9015, Université Paris-Saclay
Bruno Crémilleux	GREYC UMR 6072, Université de Caen Normandie
François Goasdoué	IRISA UMR 6074, Université de Rennes
Nathalie Hernandez	IRIT UMR 5505, Université de Toulouse
Myriam Maumy-Bertrand	LIST3N, Université de Technologie de Troyes
Nathalie Pernelle	LIPN UMR 7030, Université Sorbonne Paris Nord
Farouk Toumani	LIMOS UMR CNRS 6158, Université Clermont Auvergne

SCIENCES INFORMATIQUES

Table des matières

Préambule

La Science des Données est un champ d'étude axé sur l'analyse et l'exploration de grands volumes de données, souvent multimodales et multi-échelles, pour résoudre des problèmes dans différents domaines scientifiques. En tant que champ intrinsèquement interdisciplinaire, elle combine des méthodes informatiques et mathématiques avec une expertise spécifique à un domaine d'application afin d'extraire des connaissances permettant de mieux comprendre des phénomènes, d'éclairer la prise de décision et d'optimiser des stratégies de résolution de problèmes.

Confrontée à des évolutions constantes, la Science des Données est devenue aujourd'hui un domaine incontournable dans le paysage de la recherche scientifique. Le GDR MaDICS, créé en 2015 par le CNRS dans le cadre de sa stratégie de recherche fondamentale et interdisciplinaire sur les masses de données, a été conçu comme un outil d'animation et de prospective pour répondre aux défis croissants et aux opportunités offertes par cette discipline. Sa mission est de structurer et animer la communauté scientifique et de réaliser des prospectives sur un domaine en évolution rapide.

Depuis sa création, MaDICS a connu une évolution significative, renforçant progressivement son positionnement unique dans le paysage de recherche national. MaDICS se distingue par son approche interdisciplinaire, couvrant l'intégralité du cycle de vie de la Science des Données, de la collecte et la gestion des données jusqu'à leur analyse et interprétation dans divers contextes applicatifs. Cette interdisciplinarité, ancrée dans son ADN, permet de relever des défis complexes en mobilisant des expertises variées, allant de l'informatique, des mathématiques et des statistiques aux autres disciplines scientifiques fortement concernées par la Science des Données comme la physique, la biologie, l'écologie, la santé, l'environnement, et les sciences humaines et sociales.

Ce document est composé de deux parties distinctes. La Partie A est dédiée à la présentation du domaine de la Science des Données, incluant une présentation des défis actuels ainsi que des objets et questions de recherche clés. Cette section vise à mettre en lumière les enjeux scientifiques et technologiques majeurs de la Science des Données. La Partie B est consacrée au bilan du GDR MaDICS sur la période 2020-2024, offrant une vue d'ensemble des activités du GDR. Elle présente également les objectifs et le projet du GDR pour la prochaine mandature, incluant les évolutions prévues pour renforcer l'interdisciplinarité, développer de nouvelles collaborations/orientations, et consolider l'animation scientifique ainsi que la réflexion prospective dans le domaine de la Science des Données. Le document est complété par trois annexes : l'Annexe A qui apporte des compléments au bilan 2020-2024 du GDR (listes des Actions, des Ateliers et des manifestations labellisées et/ou soutenues pendant la période de référence) ; l'Annexe C qui décrit la méthode adoptée pour préparer le projet 2025-2029 ; et l'Annexe D, qui contient la liste des membres du GDR [1]

1. La liste actualisée des membres du GDR sera disponible le 1^{er} octobre 2024.

Première partie

La Science des Données : un domaine de recherche à l'interface de plusieurs disciplines

Chapitre 1

Émergence de la Science des Données en tant que discipline scientifique

1.1 Qu'est-ce que la Science des Données ?

La Science des Données est apparue comme une nouvelle thématique de recherche il y a plusieurs décennies [GLNS+05, HTTG09, Özsu23, Sto20] ; elle a évolué avec le temps, adoptant différentes définitions en fonction des communautés qui s'y sont investies. Aujourd'hui, la Science des Données est un domaine de recherche actif, à l'interface de plusieurs disciplines, ce qui lui permet de jouer un rôle fécond dans l'émergence de nouveaux objets scientifiques et l'élaboration de nouvelles pratiques de recherche. Ceci conduit à considérer la Science des Données selon trois facettes complémentaires :

- ▷ *un domaine de recherche scientifique* à l'interface de l'informatique et des mathématiques dont l'objectif est d'extraire des connaissances à partir de collections de données volumineuses, structurées ou non, statiques ou dynamiques. Ces connaissances peuvent être représentées sous forme d'artefacts divers (modèles d'apprentissage statistiques, modèles de décision, graphes de connaissances, ontologies, régularités ou associations dans les données, ...).
- ▷ *un nouveau paradigme de recherche scientifique*, popularisé par Jim Gray [1] [GLNS+05], axé sur l'exploitation intensive des données et des algorithmes (data intensive computing) pour stimuler des avancées dans différentes disciplines scientifiques telles que l'astro-physique, la biologie, la santé, la chimie ou les sciences humaines et sociales. Dans ce contexte, la Science des Données vise à créer et théoriser de nouveaux objets de recherche en réinterprétant les concepts du domaine sous forme d'artefacts informatiques basés sur les données et les algorithmes.
- ▷ *un processus de valorisation des données et d'exploitation de leur valeur dans les domaines économique et industriel.* Il implique la collecte, l'organisation et l'analyse des données pour

1. Historiquement, la science a évolué de l'observation empirique des phénomènes naturels vers une approche théorique utilisant des modèles et des généralisations. Plus récemment, l'avènement de la *science computationnelle* a permis l'utilisation de simulations pour étudier des phénomènes complexes. Aujourd'hui, la Science des Données permet une intégration plus étroite de la théorie, de l'expérience et de la simulation à travers l'exploration des données. Les grandes masses de données, collectées par des instruments ou générées par des simulateurs, sont stockées dans des infrastructures modernes de gestion de données et traitées par des processus complexes de curation, d'analyse et de visualisation des données.

permettre aux organisations de comprendre les tendances, les modèles et les comportements cachés dans les données. Ce processus permet d'éclairer les décisions, d'optimiser les processus industriels et commerciaux, d'identifier de nouvelles opportunités économiques ou d'améliorer les performances globales des entreprises.

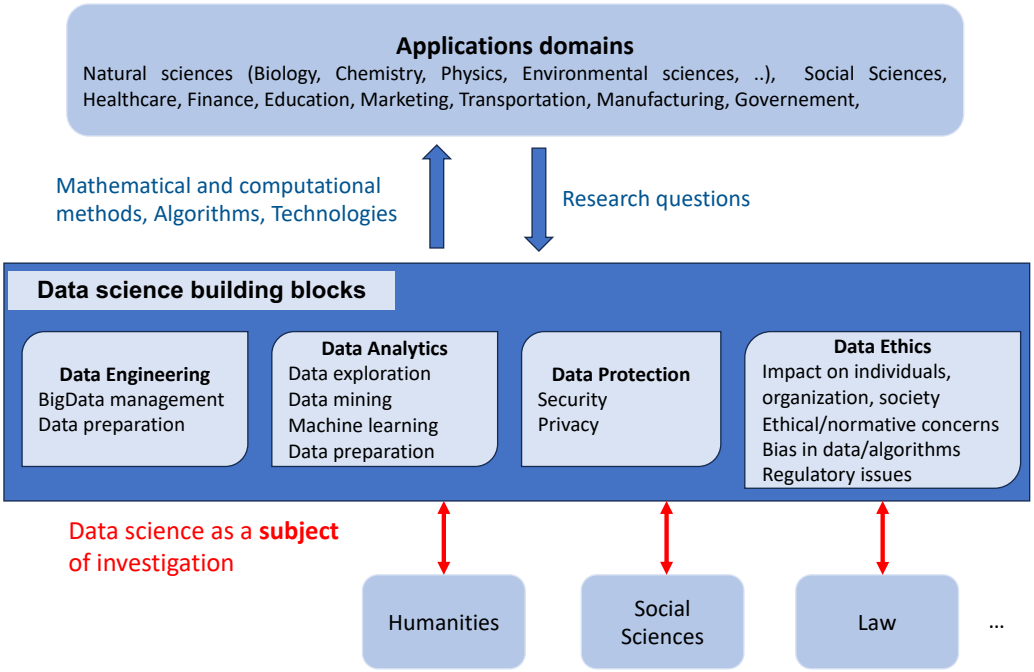

FIGURE 1.1 – Fondements de la Science des Données (version inspirée de [Özsu23])

Comme illustré par la Figure 1.1, en tant que domaine de recherche interdisciplinaire, la Science des Données repose sur un socle de compétences fondamentales constitué de quatre piliers [Özsu23] :

- **L'ingénierie des données** en science des données se concentre sur deux aspects majeurs : *la gestion des données*, qui englobe les méthodes, systèmes et plateformes modernes pour le stockage et la gestion de grands volumes de données, et *la préparation des données*, qui couvre l'ensemble du processus de sélection, d'acquisition, d'intégration des données et l'application de normes de qualité (cohérence et fiabilité des données, gestion des métadonnées et de la provenance, nettoyage des données, ...).
- **L'analyse des données** consiste en l'application de techniques statistiques, de fouille de données et d'apprentissage automatique pour extraire des connaissances à partir des données ou pour prédire le comportement du système étudié.
- **La protection des données** inclut à la fois la *sécurité des données*, qui vise à protéger les informations contre l'accès non autorisé ou les attaques malveillantes, et la *confidentialité des données*, qui assure les droits des utilisateurs en ce qui concerne leurs données (politiques de confidentialité et de réglementation, politiques de rétention et de suppression des données, accès aux données, gestion de l'utilisation des données par des tiers, et consentement des utilisateurs).
- **L'éthique des données et de leurs usages** aborde plusieurs dimensions : l'*éthique des données*, qui examine les enjeux éthiques liés à la collecte, l'analyse et l'utilisation des données

dans divers contextes ; l'*éthique des algorithmes*, qui se penche sur les préoccupations liées à la complexité croissante et à l'autonomie des algorithmes, notamment leur équité, impartialité, justice, validité et fiabilité ; et l'*éthique des pratiques*, qui traite des responsabilités et obligations des individus et organisations en charge des processus, stratégies et politiques de gestion des données.

Le socle fondamental de la Science des Données est en interaction étroite avec les domaines d'application qui ont une double fonction : exploiter les technologies, algorithmes et méthodologies de la Science des Données pour résoudre des problèmes spécifiques tout en nourrissant la discipline de la Science des Données avec des questions de recherche qui contribuent à sa fertilisation. De plus, dans une perspective réflexive, les sciences humaines étudient les aspects sociaux, politiques et sociétaux de la Science des Données, permettant ainsi une compréhension plus holistique de son impact et de ses implications.

Nous citons ci-dessous des exemples de problèmes considérés dans différents domaines scientifiques :

- En biologie, les chercheurs traitent des ensembles de données massifs issus de séquençages génomiques pour identifier des marqueurs de maladies [pro23, CP19].
- En astronomie, l'analyse de grandes quantités de données provenant de télescopes permet de découvrir de nouveaux objets célestes [pro23].
- En chimie, l'utilisation des graphes moléculaires aide à prédire les propriétés des molécules et à accélérer la découverte de nouveaux médicaments [SZN+20, ZXGZ24].
- En santé, l'analyse des données des dossiers médicaux électroniques et des capteurs de surveillance permet d'améliorer les diagnostics et de personnaliser les traitements pour les patients [SSP+21, SCKV22].
- En sciences sociales, l'exploration des données démographiques et comportementales aide à comprendre les dynamiques sociétales [MLC+15, LLL20, WLI+24].
- En archéologie, les techniques de fouille numérique permettent de reconstituer des sites anciens à partir de données de scan 3D [MSR+17].

Il convient de souligner que les volumes de données disponibles et le degré de maturité des processus d'acquisition, de préparation et d'analyse des données varient selon les disciplines. En astronomie ou en biologie, une longue tradition de gestion des grands volumes de données, avec des processus bien établis et des centres de données dédiés à l'intégration et au partage des informations, facilite cette tâche. En revanche, dans des disciplines comme les sciences humaines, la collecte de données via des enquêtes peut être fragmentée, difficile à standardiser, et poser des problèmes de qualité des données. En archéologie, la conservation et la numérisation d'artefacts fragiles compliquent la création de bases de données fiables. En santé, le processus de collecte est fortement réglementé pour assurer la confidentialité et la sécurité des informations des patients, ce qui peut ralentir l'accès et l'intégration des données.

La science des données est aujourd'hui un domaine de recherche en pleine effervescence, où de nouveaux défis et des problématiques réactualisées émergent en réponse aux caractéristiques particulières des données, à la diversité des sources, et à la multiplicité des usages découlant des questions variées posées dans les différents domaines scientifiques. En préambule, nous illustrons ci-dessous, à travers l'exemple d'AlphaFold, l'interaction féconde entre plusieurs domaines scientifiques, montrant comment ces synergies peuvent conduire à des avancées significatives. Un aperçu plus détaillé des sujets de recherche en science des données est proposé dans le Chapitre 2.

Exemple : la Science des Données et l'interdisciplinarité comme source de questions de recherche en informatique L'apport de la Science des Données à des domaines scientifiques part souvent de méthodes informatiques existantes qu'il faut adapter tout en donnant lieu à de nouvelles questions de recherche propres à l'informatique comme l'illustre l'exemple d'AlphaFold envers l'apprentissage profond. Pour pouvoir comprendre le mode d'action régissant l'interaction entre un ligand (un médicament) et une protéine, il est nécessaire d'avoir une idée précise de la conformation active de la protéine. Pour de nombreuses protéines, cette conformation correspond à la conformation par défaut. Aussi, l'un des principaux objectifs de la biologie et de la pharmacie depuis plus d'un demi-siècle est de prédire la conformation active d'une protéine à partir de la suite des acides aminés qui la composent [Ano74]. L'importance du défi, sa crédibilité, et les premiers résultats dans ce sens ont été reconnus via l'attribution du prix Nobel de chimie 1972 à Christian Boehmer Anfinsen. D'énormes progrès ont été réalisés dans ce sens grâce au modèle AlphaFold [JEP+21] et à ses évolutions AlphaFold2 [BPE22] et AlphaFold3 [AAD+24]. Ces modèles s'appuient sur la présence de données massives exploitables et une importante évolution des méthodes informatiques issues de l'apprentissage profond [BDA+21]. D'autre part, les performances de ce type de modèle reposent aussi sur un travail précis d'intégration de méthodes issues des connaissances de la physique [BMD24, DKS+23]. On peut donc dire que la résolution de ce défi posé par la biologie et à fort impact applicatif reposera sur une avancée issue d'un effort conjoint de plusieurs disciplines : la biologie évidemment, mais aussi l'informatique pour une exploitation optimale des données et la physique pour l'intégration de principes fondamentaux.

1.2 Cadre méthodologique de la Science des Données

Une vision holistique de la science des données est présentée à travers une approche orientée processus, qui permet de structurer les différentes étapes de traitement des données, depuis leur production ou collecte jusqu'à leur valorisation. Ce processus est communément appelé le cycle de vie de la Science de Données [Sto20, Özsu23]. La littérature propose plusieurs variations de ce cycle de vie, souvent dépendantes du contexte applicatif, mais convergeant globalement sur les grandes étapes du traitement des données. À chaque étape, les données sont enrichies, nettoyées, et transformées pour en améliorer la qualité et l'utilité.

La Figure 1.2 présente un exemple générique de cycle de vie de la Science de Données repris de [Özsu23]. Le processus commence par la définition de la question de recherche issue du domaine scientifique, suivie de la préparation des données, qui inclut la sélection des ensembles de données nécessaires à l'étude, leur ingestion, ainsi que la résolution des problèmes de qualité tels que la vérification de la cohérence des données, le nettoyage et la complétion des données. Ensuite, les problèmes de stockage et de gestion des données et des métadonnées sont traités. Une fois ces étapes accomplies, les données sont prêtes pour l'analyse, qui consiste à choisir ou développer des modèles statistiques, de fouille de données ou de machine learning, et à effectuer des validations pour en vérifier la pertinence. Si le modèle est validé avec succès, il passe à la phase de déploiement et de diffusion.

À titre d'exemple, la Figure 1.3, tirée de [BCP+23], illustre une chaine d'analyse de données provenant des réseaux sociaux. Cette figure instancie partiellement le processus générique de la Figure 1.2 en le déclinant en deux grandes étapes : la préparation des données et l'analyse des données, tout en omettant l'étape de gestion des données (stockage, indexation, interrogation, ...).

Un cycle de vie plus détaillé est présenté dans [Sto20], offrant une vue plus explicite des contributions des diverses communautés scientifiques à la discipline de la Science de Données. Ce point

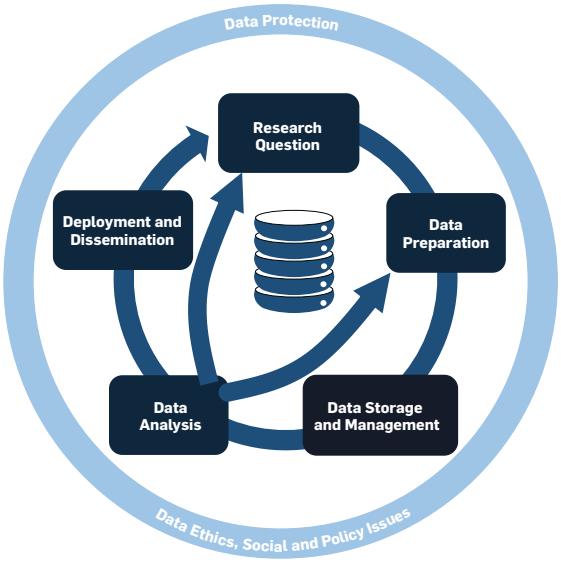

FIGURE 1.2 – Un processus de Science de Données itératif et guidé par [Özsu23]

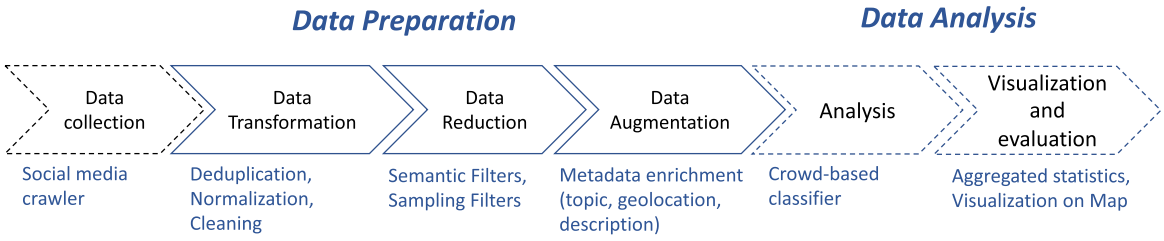

FIGURE 1.3 – Exemple d'un processus d'analyse des médias sociaux [BCP+23]

de vue permet une approche systématique pour aborder les questions méthodologiques, computationnelles ou spécifiques au domaine d'application, ainsi que des sujets telles que l'éthique, la reproductibilité et la cyberinfrastructure pour la Science des Données [2].

La Figure 1.4 illustre un cycle de vie de la Science des Données décomposé en quatre niveaux [Sto20] :

- Le *niveau de l'application ou du domaine*, qui fait référence à l'application scientifique ou au domaine de recherche, décrit les étapes d'un projet de recherche au niveau du domaine scientifique considéré : conception expérimentale ; obtention/génération/collecte de données ; exploration des données et génération d'hypothèses ; nettoyage, fusion et organisation des données ; sélection des caractéristiques et préparation des données ; estimation du modèle et inférence statistique ; simulation et validation croisée, visualisation ; publication et préservation/archivage des artefacts.
- Le niveau *infrastructure logicielle*, qui décrit la couche logicielle support, regroupe une variété de technologies informatiques dont chacune représente un domaine de recherche et de déve-

2. Le terme cyberinfrastructure désigne l'ensemble des nouveaux environnements de recherche qui intègrent des fonctions avancées d'acquisition (data acquisition), de stockage (data storage), de gestion (data management), d'intégration (data integration), de fouille (data mining), de visualisation (data visualization) des données, ou d'autres services de traitement informatique ou informationnel. Dans le domaine scientifique, la notion de cyberinfrastructure fait référence à toute solution technique destinée à établir des connexions efficientes entre des données, des ordinateurs et/ou des personnes, dans le but de favoriser la transmission des nouvelles théories scientifiques et du savoir (wikipédia).

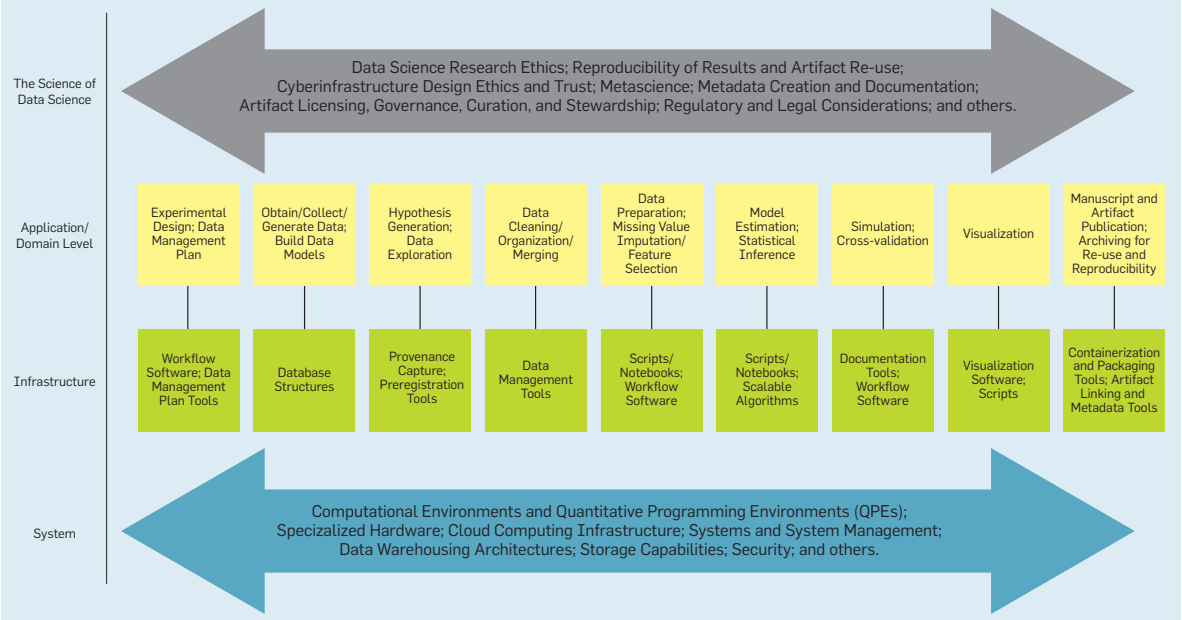

FIGURE 1.4 – Exemple de cycle de vie de la Science des Données [Sto20]

loppement à part entière. Cela inclut des notebooks et des logiciels de workflow, des outils de visualisation, des langages d'inférence statistique, des outils de gestion des données, ainsi que des outils d'intégration et d'archivage d'artefacts.

- Le niveau *système* qui englobe les environnements informatiques (acquisition, ingestion, stockage) et environnements de programmation quantitative (QPE), le matériel spécialisé ; les infrastructures de cloud computing, les outils d'administration des systèmes, les entrepôts de données, etc.
- Le niveau *méta-scientifique* "La Science de la Science des Données" qui couvre des sujets supplémentaires, notamment les aspects d'éthique (consentement, transparence, équité, protection de données personnelles) ; la documentation de la recherche et la création de métadonnées (traçabilité) ; l'explicabilité et la reproductibilité ; les aspects politiques et légaux, y compris la gouvernance, la confidentialité et les considérations de propriété intellectuelle (protection des sources) ; les aspects environnementales etc.

1.3 Le défi d'automatisation des processus de Science des Données

Comme illustré dans la section précédente, les processus de Science de Données sont exploratoires et itératifs, et souvent guidés par la nature des données analysées. L'automatisation est par conséquent un enjeux central pour construire des processus efficaces, fiables et évolutifs. Elle permet d'accélérer les étapes répétitives et laborieuses, réduire les erreurs humaines, faciliter l'application des principes FAIR (Findable, Accessible, Interoperable, Reusable) [ABB+16], augmenter la reproductibilité et la traçabilité et offrir des opportunités d'optimisation de ressources de stockage et de calcul.

L'automatisation peut revêtir plusieurs formes [DBDRHO+22], telles que des algorithmes permettant de mécaniser des tâches spécifiques (comme par exemple la complétion des valeurs manquantes ou la classification automatique d'objets ou d'items), la composition ou l'intégration automatique de tâches (par exemple dans le cadre d'un workflow), ou encore le développe-

ment d'assistants qui soutiennent l'activité humaine à travers des dispositifs de visualisation de motifs, ainsi que la production d'explications, de conseils ou de recommandations.

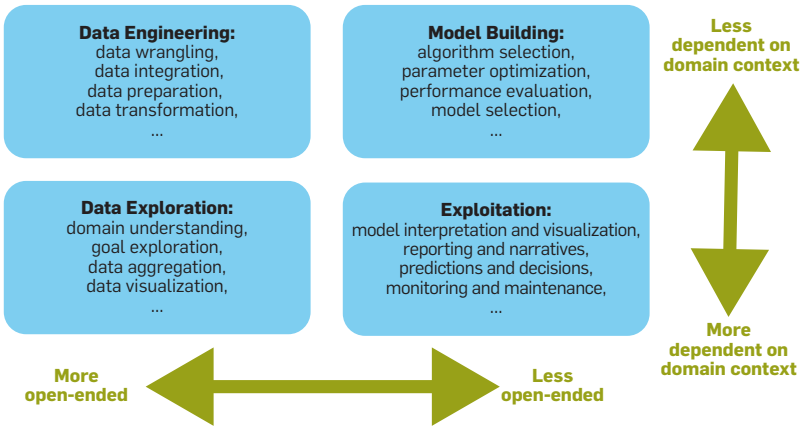

FIGURE 1.5 – Les deux dimensions d'un processus Science des Données [DBDRHO+22]

La Figure 1.5 [DBDRHO+22], présente un cadre conceptuel pour identifier les défis liés à l'automatisation de processus Science des Données en décomposant les principales tâches en quatre quadrants. La dimension verticale reflète la dépendance au contexte d'application, qui inclut non seulement les connaissances spécifiques au domaine, mais aussi les facteurs humains et sociaux, notamment les questions de sécurité, d'éthique et d'environnement. Les quadrants inférieurs, *Exploration* et *Exploitation des données*, sont généralement étroitement liés au domaine d'application, tandis que les quadrants supérieurs, *Ingénierie des données* et *Construction de modèles*, sont souvent plus indépendants du domaine. L'axe horizontal représente le degré de spécification des différentes activités, allant d'un faible niveau de précision à une spécification plus détaillée, incluant par exemple des objectifs bien définis, des tâches de modélisation claires, et des indicateurs de performance mesurables. L'*Ingénierie des données* et l'*Exploration des données* ne sont souvent pas précisément spécifiées et ont un caractère assez itératif, tandis que la *Construction de modèles* et l'*Exploitation* sont généralement définies de manière plus précise. Les deux dimensions offrent un cadre pour évaluer les possibilités d'automatisation des différentes activités en science des données. La *Construction de modèles* est l'activité où l'on peut anticiper l'impact le plus direct de l'automatisation, comme illustré par l'apprentissage automatique automatisé (AutoML). En revanche, les tâches d'*Ingénierie des données*, qui représentent souvent 80% de l'effort humain dans un projet de Science de Données, sont plus complexes à automatiser. De même, l'*Exploration des données* exige des connaissances préalables et une expertise humaine, ce qui en fait un défi majeur pour l'automatisation. L'*Exploitation*, centrée sur l'interprétation des résultats et la prise de décision, nécessite également une supervision humaine. Toutefois, les avancées récentes en IA, telles que l'IA générative ou les approches neuro-symboliques, ouvrent de nouvelles perspectives pour automatiser la rédaction de rapports et améliorer l'explicabilité des résultats.

Le chapitre suivant examine les questions de recherche en Science des Données, dont une grande partie se concentre sur l'automatisation des différentes tâches au sein du processus de Science de Données.

Chapitre 2

La recherche en Science des Données

Ce chapitre explore les enjeux de la recherche en science des données. Il ne s'agit pas de fournir un état de l'art complet, mais plutôt d'illustrer la diversité des questions scientifiques qui se posent dans ce domaine. La Section 2.1 décrit les différents objets de recherche, tandis que la Section 2.2 présente des exemples de sujets de recherche en Science des Données.

2.1 Les objets de recherche de la Science des Données

Les objets de recherche en Science des Données couvrent un large éventail de concepts et de techniques essentiels pour extraire des connaissances à partir des données. Ces objets incluent *les données* elles-mêmes, qui sont la matière première de toutes les analyses ; *les modèles*, qui permettent de représenter et de comprendre les relations entre les données ; *les opérateurs et algorithmes*, qui sont les outils et méthodes utilisés pour traiter et analyser les données ; *les chaînes de traitements (workflows)*, qui orchestrent les différentes étapes du processus d'analyse des données ; et *les infrastructures et systèmes*, qui fournissent le support matériel et logiciel nécessaire pour gérer et traiter les données de manière efficace et sécurisée. La suite de cette section, décrit succinctement chacun de ces objets de recherche, en explorant ses caractéristiques et son rôle dans le domaine de la Science des Données

2.1.1 Les données

La notion de "donnée" est centrale en Science des Données et englobe un large éventail de représentations d'informations, depuis les données brutes initiales extraites des sources de données (capteurs, bases de données, fichiers, réseaux sociaux, et autres) jusqu'aux données consolidées utilisées pour l'analyse et la prise de décision. Les données peuvent se présenter sous une multitude de formats et modèles, chacun ayant ses propres caractéristiques, avantages et défis en termes de traitement et d'analyse.

Nous pouvons distinguer différentes caractéristiques liées aux données :

- *Le volume de données :* le volume de données se réfère à la quantité de données à traiter dans un processus Science des Données. Dans de nombreux domaines scientifiques, les grandes quantités de données disponibles permettent l'utilisation de méthodes récentes pour la construction de modèles mais posent également des défis en termes de stockage, traitement, analyse et optimisation des performances. En même temps, dans certains domaines, par exemple en santé, les données sont nettement moins disponibles et nécessitent des procédures de collecte complexe pour produire des données exploitables par des méthodes statis-

tiques récentes. La collecte de données expérimentales est souvent strictement réglementées et nécessite des mesures sur le terrain avec des besoins humains et des équipements coûteux, ce qui en fait un processus souvent cher tant en ressources financières qu'en temps [RP24]. La protection des données produites et les réglementations sur les données impliquent aussi souvent que les expériences sont uniques et difficilement reproductibles.

- *La qualité des données :* elle se définit comme l'état d'une information qui est représentative de la réalité qu'elle décrit et qui est *appropriée pour les usages prévus*, tels que la prédiction, la prise de décision ou la planification. La qualité de données se mesure à travers plusieurs dimensions telles que l'exactitude, la consistance, la complétude, l'unicité (absence de redondance), l'objectivité (absence de biais), etc.
- *La diversité des données :* les données peuvent représenter des informations sous forme de textes, images, son, vidéo, réseaux, séries temporelles etc. Elles peuvent avoir une dimension spatiale et/ou temporelle qui rend leur modélisation, l'optimisation des traitements et l'analyse plus complexes et qui nécessitent des solutions spécifiques. Cette multi-modalité pose des défis importants de modélisation, transformation, d'analyse et d'interprétation.
- *Le mode de production des données* : les données peuvent être produites de diverses manières, notamment par des observations humaine, des enquêtes, des capteurs, des calculs dérivés, de transformations ou des simulations. Le mode de production des données a plusieurs impacts significatifs sur les caractéristiques des données, comme par exemple le volume de données généré, le coût de production, la qualité, la granularité et la temporalité incluant la fréquence de génération et la fréquence de mise-à-jour (obsolescence).
- *La granularité des données :* les données peuvent représenter des phénomènes à différentes échelles structurelles, temporelles et spatiales. La collecte, l'intégration et l'analyse de données multi-échelles introduisent des défis spécifiques par leur hétérogénéité sémantique et structurelle. Par exemple, les données provenant de différentes échelles peuvent avoir des représentations ou des interprétations différentes, ce qui rend l'intégration plus complexe.
- *Les métadonnées :* les métadonnées fournissent des informations sur la structure, le contenu, le format, la provenance, et la qualité des données, facilitant ainsi leur gestion, leur recherche, et leur utilisation. Elles jouent un rôle essentiel dans la compréhension, la traçabilité, la réproductibilité et la gouvernance des données en général. La création de métadonnées de bonne qualité est souvent coûteuses, en particulier quand il s'agit de décrire la qualité (complétude, précision), la provenance (source, traitements subis, ...) et le contenu des données.

2.1.2 Les modèles

De nombreux domaines scientifiques font aujourd'hui un usage intensif de la modélisation, avec des définitions souvent distinctes de la notion de modèle. Le domaine de la Science des Données confère un sens large à la notion de modèle, en distinguant notamment quatre catégories :

- *Les modèles de données* correspondent à une abstraction du réel sous forme de structures de données appelées schémas. Ils fournissent des descriptions abstraites de la structure, des relations, des contraintes, des règles d'intégrité des données, ainsi que des primitives de manipulation des données.
- *Les modèles dynamiques* décrivent des objets et des relations dans un système qui évoluent avec le temps. Ceci inclut la modélisation des changements d'états d'une entité ou d'un système (données d'un système de contrôle, données de simulation) et des flux de données générées par un système (e.g., flux de transactions financières, données de capteurs).
- *Les modèles analytiques et prédictifs*, qui sont des représentations mathématiques, statis-

tiques, logiques ou algorithmiques qui permettent d'extraire de l'information des données et de capturer/prédire les relations entre ces informations. Ceci inclut les modèles d'apprentissage supervisé et non-supervisés pour identifier des motifs cachés dans les données, prédire des données et automatiser des décisions complexes.

- ▷ *Les modèles d'interaction* formalisent les liens sémantiques entre modèles de données (e.g., modèles de mapping ETL - Extract, Transform, Load), les liens de dépendance entre modèles de comportement ou les liens entre les données et les comportements (ressources, contraintes). Ces modèles sont essentiels pour comprendre les interdépendances dans les données et entre les systèmes, notamment dans le cadre de l'intégration de données.

Quelques caractéristiques principales des modèles sont décrits ci-dessous :

- ▷ *L'hétérogénéité des modèles :* la diversité des données et de leur propriétés implique la définition de modèles spécialisés pour leur représentation et traitement. La diversité des représentations (tables, graphes, séries temporelles, multi-dimensionnelles, statistiques, matrices, ...) nécessite des solutions avancées pour assurer leur interopérabilité et une exploitation efficaces des données modélisées.
- ▷ *Flexibilité* : cela inclut la capacité à supporter différents types de schémas, qu'ils soient stricts ou flexibles, ainsi que des schémas de validation qui s'adaptent aux besoins spécifiques des données.
- ▷ *Interprétabilité* : la facilité de compréhension par l'utilisateur est cruciale, ce qui inclut également la capacité d'expliquer les résultats produits par le modèle. L'interprétabilité permet aux utilisateurs de mieux appréhender les décisions prises par le modèle.
- ▷ *Abstraction* : cette caractéristique fait référence à la capacité de modéliser des informations en fonction d'un objectif donné, en simplifiant les détails superflus pour se concentrer sur l'essentiel.
- ▷ *Fiabilité* : un exemple classique est SQL, qui est fondé sur la logique, assurant ainsi un haut niveau de fiabilité dans les requêtes et les transactions.
- ▷ *Robustesse* : la capacité du modèle à fournir des représentations fiables même en présence de données erronées ou imprécises. Un modèle robuste doit pouvoir gérer les erreurs et les incohérences sans compromettre les résultats.
- ▷ *Expressivité* : cette caractéristique se rapporte à la capacité du modèle à exprimer des contraintes d'intégrité et des relations complexes entre les données, offrant ainsi une large gamme de possibilités pour la modélisation.
- ▷ *Complexité* : la complexité d'un modèle influence ses performances et son optimisation. Les modèles plus complexes peuvent offrir plus de fonctionnalités, mais peuvent aussi nécessiter plus de ressources pour être exécutés efficacement.
- ▷ *Généralisation* : la capacité du modèle à s'appliquer à un large éventail de cas d'usage, au-delà des scénarios spécifiques pour lesquels il a été initialement conçu.
- ▷ *Manipulation de données :* les modèles peuvent permettre la manipulation des données à travers des langages déclaratifs ou APIs pour la définition (conception), l'interrogation et/ou la mise à jour.

2.1.3 Les opérateurs/algorithmes

Dans un processus de Science des Données, les *opérateurs* réalisent les diverses tâches sur les données telles que l'extraction, l'intégration, la curation, l'analyse et l'apprentissage automatique.

Ces opérateurs jouent un rôle essentiel avec un impact direct sur l'efficacité et la performance globale des processus. Les questions de recherche sous-jacentes concernent la conception et l'implémentation des algorithmes nécessaires pour résoudre les problèmes induits par ces tâches.

Quelques caractéristiques principales des opérateurs sont décrites ci-dessous :

- ▷ *La complexité algorithmique et les performances théoriques :* l'évaluation de la complexité algorithmique (en temps et en espace) et des performances théoriques des opérateurs est essentielle pour évaluer leur scalabilité sur des jeux de données de tailles variables et garantir leur efficacité en fonction des contraintes de performance.
- ▷ *Les exigences d'utilisation :* les opérateurs peuvent être conçus pour fonctionner en mode batch ou en mode continue dépendant des données à traiter, des contraintes (par exemple, temps-réel) et des objectifs. Chaque mode présente des défis spécifiques en termes de latence, volume de données, et de gestion des ressources.
- ▷ *L'environnements de mise en œuvre :* la mise en œuvre des opérateurs peut varier en fonction des environnements de calcul (cluster de données, cloud, HPC). La parallélisation et la distribution sont des stratégies couramment utilisées dans ces environnement pour améliorer les performances des opérateurs qui posent des défis de parallélisation des algorithmes, la distribution optimale des données et des traitements (parallélisme de données), de gestion optimale de ressources de stockage et de calcul.

2.1.4 Les chaînes de traitements

Les chaînes de traitement ou workflows sont des séquences de tâches automatisées pour extraire, transformer, analyser et visualiser des données. Elles sont obtenues par composition des opérateurs (e.g., nettoyage des données, intégration de données de données, analyse et agrégation de données, exploration et visualisation de données...), caractérisées par :

- ▷ *L'hétérogénéité des traitements :* les workflows sont conçus pour intégrer et orchestrer une variété de traitements pour l'extraction, l'exploration, l'analyse et l'agrégation des données, l'extraction des connaissances et la visualisation.
- ▷ *Le degré d'automatisation :* cette caractéristique mesure le degré d'automatisation et les besoins d'intervention humaines des différentes étapes d'un workflow. Cela inclut les étapes de sélection et de configuration des méthodes, des opérateurs, des algorithmes, des ressources requises pour l'exécution et la maintenance du workflow.
- ▷ *La gestion et l'optimisation des opérations et des ressources :* cette caractéristique concerne l'exécution des différentes tâches du workflow, ainsi que la gestion efficace des données et des ressources de stockage et de calcul. Bien que chaque opérateur puisse être optimisé individuellement, la performance globale d'une chaîne de traitements dépend des stratégies d'exécution (séquentielle ou parallèle) et des environnements utilisés.
- ▷ *La traçabilité d'exécution :* cette caractéristique décrit la capacité de suivre et documenter les étapes et transformations qu'un jeu de données subit au sein d'un workflow. La traçabilité d'éxecution est essentielle pour analyser l'origine des données et les transformations (provenance), ainsi que pour garantir la transparence, l'explicabilité, la reproductibilité et la fiabilité des résultats des analyses.

2.1.5 Les infrastructures d'implémention et d'exécution

Les infrastructures et systèmes constituent des objets de recherche fondamentaux en Science des Données, englobant l'ensemble des technologies, architectures et plateformes nécessaires pour stocker et gérer de grandes quantités de données et exécuter efficacement des workflows complexes.

Les infrastructures sont caractérisées par :

- *Scalabilité et élasticité :* cette caractéristique réfère à la capacité de l'infrastructure à gérer des volumes croissants de données et à s'adapter à l'augmentation des besoins de traitement, en limitant les coûts (matériels et humains) et en garantissant des performances acceptables, même en cas de forte croissance des données.
- *Performance :* cette caractéristique évalue l'efficacité de l'infrastructure dans le déploiement et l'exécution des workflows sur des données massives. Les infrastructures modernes reposent sur des systèmes massivement parallèles et distribués. Une question importante porte également sur l'optimisation de la consommation d'énergie, afin de réduire l'empreinte carbone et les coûts opérationnels.
- *Fiabilité et disponibilité :* cette caractéristique mesure la robustesse du système pour assurer une continuité des services avec une tolérance aux pannes élevée. Cela inclut la mise en place de mécanismes de sauvegarde et de récupération afin de protéger les données et les opérations contre les défaillances, particulièrement dans des environnements distribués et massivement parallèles.
- *Flexibilité et extensibilité :* cette caractéristique évalue la capacité du système à intégrer des nouvelles fonctionnalités et à interagir avec divers types de systèmes et nouvelles technologies au fil du temps. Cela inclut la compatibilité avec divers outils, frameworks, et modèles de déploiement (cloud, on-premise, hybrid).
- *L'innovation technologique :* nécessite l'incorporation de technologies nouvelles ou émergentes telles que l'intelligence artificielle, l'apprentissage automatique, le calcul quantique et les systèmes de calcul avancés pour pousser les limites actuelles des infrastructures et des systèmes.

2.2 Les sujets de recherche en Science des Données

Cette section présente les sujets de recherche actuels en science des données. Étant donné l'étendue thématique de ce domaine, il est difficile d'en offrir une vue d'ensemble exhaustive. L'objectif est donc de mettre en lumière quelques exemples de questions de recherche et de travaux récents pertinents.

La section Section 2.2.1 présente des exemples de sujets de recherche liés aux piliers fondamentaux de la Science des Données, tandis que la Section 2.2.2 illustre des problématiques spécifiques provenant des domaines scientifiques.

2.2.1 Sujets de recherche liés aux piliers fondamentaux de la Science des Données

Cette section présente divers sujets de recherche en Science des Données, organisés par grandes thématiques. Certaines thématiques, comme par exemple la qualité des données, bien établies depuis de nombreuses années au sein de la communauté scientifique, connaissent un regain d'intérêt en raison de l'évolution des problématiques sous-jacentes pour répondre aux exigences

spécifiques à la Science des Données. Nous fournissons des exemples pour illustrer ces nouvelles questions.

2.2.1.1 Préparation et protection des données

Cette section explore les problèmes de recherche associés à la préparation et la protection des données dans le cycle de vie de la Science des Données. Ces étapes comprennent la *découverte et l'exploration des données*, la *qualité des données*, le *profilage des données*, l'*intégration des données*, la *protection des données*, la *gestion des métadonnées* et la *traçabilite des données*. Nous examinerons des exemples de questions de recherche actuelles et les approches développées pour répondre aux exigences spécifiques de ces tâches.

La *découverte et l'exploration des données* est une étape importante dans le cycle de vie de la Science des Données. La découverte de données vise à identifier les ensembles de données les plus pertinents pour répondre à des questions spécifiques, tandis que l'exploration permet de saisir leurs propriétés et les relations entre elles [PCW23]. Ces processus nécessitent des approches méthodologiques et des outils performants pour naviguer dans la diversité et la complexité des données. Les thèmes de recherche associés sont variés et incluent la recherche d'ensembles de données, la navigation dans les données, l'annotation des données, la qualité des données, et l'inférence de schéma (voir [CSK+19, PCW23] pour une revue des problèmes de recherche dans ce domaine). La *découverte de données* reste principalement basée sur des mots-clés à partir des métadonnées [CSK+19], une approche qui peut être améliorée en utilisant des techniques de similarité issues de la recherche d'information [CHK09], de matching sémantique [ZB18], ou en exploitant les grands modèles de languages [CTH+20]. Des travaux récents ont abordé divers défis liés à l'identification des liens entre les ensembles de données, soit en adoptant une approche basée sur les opérateurs relationnels classiques, jointure dans les lacs de données massifs [ZDNM19, BFPK20] ou l'union [NZPM18, BFPK20], soit en exploitant des relations de causalité entre attributs [LSN24]. Par ailleurs, l'utilisation des grands modèles de langage pour la découverte et l'exploration des données est un domaine de recherche en plein essor [KLF+24]. L'*inférence de schémas* est le processus qui consiste à identifier les ensembles de données partageant des éléments structurels communs, afin de générer un schéma global capable de les représenter. Une riche littérature traite de ce sujet, avec une majorité de travaux se concentrant sur des modèles de données spécifiques, comme par exemple les données relationnelles [YPS09], XML [BNV07], graphes RDF [GGM20] ou données JSON [BCGS22, Kle23]. Des états de l'art récents dans ce domaine sont présentés dans [KMKT+21, PCW23]. Les recherches récentes ont également élargi le champ d'application à l'extraction de schémas à partir de textes en langage naturel. Par exemple, [SCM18] propose une revue des méthodes d'extraction de relations qui utilisent des données structurées ou semi-structurées pour orienter le processus d'extraction, tandis que [GSeOC+21] se concentre sur les approches dédiées à l'extraction de relations temporelles à partir de textes cliniques. [DDW+22, XCP+24] explorent également l'extraction d'informations structurées à partir de textes en langages naturels en utilisant des modèles de langage de grande taille (LLMs).

La *qualité des données* est un sujet récurrent de la recherche en gestion des données [Fan15], qui reste ouvert en raison de la diversifications des données (données non structurées, absence de méta données, plus grandes hétérogénéité des sources, ...) et des usages (prédiction par des algorithmes apprentissage automatique, ...). Par exemple, une tâche centrale dans le nettoyage des données concerne la détection et la correction d'erreurs. Les questions de recherche sous-jacentes sont variées et concernent par exemple la complexité des algorithmes de détection d'erreurs en fonction du type de contraintes considérées [FTWY21, LDF+24], la conception d'algorithmes parallèles [FTWY21] ou qui s'adaptent aux caractéristiques des données [PdAN21] ainsi que

la prise en compte de conflits dans un cadre unifié [FTWY21]. Les approches quantitatives de la qualité des données pour l'apprentissage automatique font également l'objet d'une attention croissante. Une question sous-jacente concerne l'établissement de correspondances entre les problèmes de qualité de la prédiction (modèles ML) et les problèmes de qualité de données [WFWW20, LRB+21, RRK+20]. Par ailleurs, l'utilisation de l'apprentissage automatique pour le nettoyage des données connaît un intérêt croissant. Des travaux récents incluent le nettoyage de données en utilisant les GAN (Generative Adversarial Networks) [HCL20] ou l'apprentissage profond [PSCH21], le nettoyage de données auto-supervisé et interprétable [PST+22] ou l'imputation des données manquantes en exploitant différentes techniques d'apprentissage automatique [PN24, HB21, ZSS24].

Le *profilage des données* [AGN15] vise à extraire des métadonnées à partir des données [FTR+21a, FTM+21]. Les approches traditionnelles se sont principalement concentrées sur le calcul de statistiques sur des attributs individuels, ainsi que sur l'analyse des corrélations et des règles d'association entre plusieurs attributs. Récemment, l'intérêt s'est tourné par exemple vers le traitement des attributs non catégoriels [FTM+21], la découverte de différentes formes de dépendances, conditionnelles [FGJK08, JTW+23], partielles ou approximatives [CDNP21, XTWM22] ou differentielles [KYTM24], la découverte incrémentale de contraintes lors de mise à jour des données [QLT+23, CCDP22] ou la proposition de nouvelles primitives de profilage de données, comme par exemple les contraintes de conformité qui permettent de mesurer la confiance dans les données [FTR+21b] ou les dépendances fonctionnelles de motifs (pattern DF) pour le nettoyage des données [QTO+20].

L'*intégration des données* était définie comme le problème de la combinaison de données provenant de différentes sources afin de fournir à l'utilisateur une vue unifiée de ces données [Len02]. Aujourd'hui, cette vision a évolué en raison de l'explosion des volumes de données arrivant à grande vitesse, provenant de sources dynamiques, avec des niveaux de granularité et de qualité variés et ayant une grande variété de formats, y compris des formats personnalisés avec des structures et des délimiteurs spécifiques ou des séquences d'attributs non documentées [FB23]. Ce nouveau contexte ouvre la voie à un nouvel agenda de recherche [GHMT17, Mil18, NAJ23]. Des travaux récents portent notamment sur la génération automatique de structures (schémas, matrices, frames, etc.) à partir de données brutes textuelles aux formats variés (JSON, XML, CSV, LibSVM, MatrixMarket, etc.) ou provenant de secteurs divers [ABC+22, FB23], ainsi que sur l'intégration des tables des data lakes [KSGM22]. Par ailleurs, l'utilisation de l'apprentissage automatique pour automatiser différentes tâches d'intégration est une direction de recherche très active. Les travaux récents incluent l'utilisation des réseaux de neurones [BG21], des GNN (graph neural network) [DR07], ou des techniques de co-apprentissage [WWD+23] pour la résolution d'entités (entity matching), des modèles de fondation (foundation models) pour le nettoyage et l'intégration des données [NCOR22], l'apprentissage de représentations contrastives (contrastive representation learning) pour automatiser différentes tâches de préparation et d'intégration des données [WLW22] ou pour l'annotation de colonnes [MW23], ainsi que l'utilisation de la classification automatique basée sur les forêt aléatoires [DURG+18] ou bien l'apprentissage profond pour améliorer l'automatisation du matching de schémas [SGR20].

La gestion des métadonnées est devenue une question centrale dans les environnements modernes tels que les lacs de données, qui manipulent des données brutes structurées, semi-structurées et non structurées, dans des formats et à des échelles variés. Les métadonnées sont utilisées pour compenser l'absence de schéma structurant préétabli, facilitant ainsi la recherche, le catalogage et la compréhension des données stockées dans un lac de données. Les recherches récentes se concentrent sur l'architecture des lacs de données et les mécanismes sous-jacents de gestion des métadonnées [SD21, MRZ21], le développement de modèles génériques de métadonnées [HSG+17, EGG+21], ainsi que sur les services de gestion des catalogues de données

[SD21, SML+23]. La conception de ces catalogues soulève des questions complexes liées à l'extraction des métadonnées [HKN+16], aux politiques de gouvernance des données [KB10], et à la représentation des informations concernant la provenance des ensembles de données [MBC13], le profilage des données [AGN15] ainsi que les contextes d'utilisation des données [KB10].

La *traçabilite des données* est utile pour garantir des processus Science des Données fiables et reproductibles, en particulier dans les domaines scientifiques [HZB+24, MTF+23]. Elle englobe les travaux sur les modèles de provenance [MBC13], qui posent de nombreux défis dans les environnements multiplateformes [HZB+24], ainsi que sur l'extraction de la provenance, par exemple à partir des journaux de requêtes [PAS+23]. L'annotation des données constitue également une problématique importante qui fait l'objet d'études principalement issues des communautés des bases de données, du web sémantique, du traitement du langage naturel et de l'apprentissage automatique [PCW23].

La *protection des données* représente un défi majeur en science de données [RVF24] en raison des risques de fuite ou de falsification des données ou d'exploitation malveillante de ces dernières. Des travaux récents se sont intéressés à la confidentialité différentielle (*differential privacy*) [DR14], qui offre des garanties de protection utile pour l'analyse statistique (requêtes d'agrégation), à la protection des données dans le cadre d'une chaîne d'opérations (workflow) [Bel21] ainsi qu'à l'équilibre entre l'anonymisation des données et leur utilité [RKIW18], en prenant en compte des facteurs tels que l'origine et la provenance des données. La protection de données devient plus sensible dès lors qu'il s'agit de partager des données pour permettre à des tiers de les utiliser, les explorer et les analyser [FWCY10]. À cet égard, la cryptographie homomorphe [BBB+22] connaît un regain d'intérêt. Cette technique permet d'effectuer des calculs, par exemple évaluation des requêtes, sur des données chiffrées sans avoir besoin de les déchiffrer au préalable.

2.2.1.2 Modélisation des données et des connaissances

Les modèles de données et de connaissances structurent les informations pour les rendre compréhensibles (avec une sémantique précise et bien fondée) et exploitables (à travers des syntaxes/langages standards) par les humains et/ou les algorithmes. Leur rôle est de faciliter le partage de données et de connaissances ainsi que l'implémentation d'outils génériques pour la manipulation des données et des métadonnées dans un processus de Science des Données.

Données hétérogènes et complexes (noSQL) : le besoin de gérer et d'analyser des données de plus en plus complexes et massives est accompagné par l'apparition de nouveaux modèles et langages pour la représentation et la gestion de données, de métadonnées et de connaissances. Ces modèles sont de plus en plus divers et spécialisés pour représenter des données spatiales et temporelles [LZW+19, BBDM+22, EM17], des graphes de données [AG08, AAB+17, BGP+23], des flux de données [BBD+02, BT24], etc. Ce phénomène est caractérisé par le terme noSQL (not only SQL) pour indiquer les limites du modèle relationnel. Néanmoins, la flexibilité et les performances de la technologie SQL (stockage, optimisation, transactions...) et la robustesse et la généricité de ses fondements théoriques (algèbre et calcul relationnel, ...) restent d'actualité et continue à influencer le développement de ces nouveaux modèles [PA16]. Par ailleurs, pour répondre à la diversité croissante des types de données, des systèmes de gestion de bases de données multi-modèles ont émergé, offrant une plateforme unique pour gérer des données aux formats variés [LH19].

Les *Graphes de connaissances* sont exploités en Science des Données [HBC+21, PXNO23] pour modéliser des connaissances structurées et introduire des capacités de raisonnement sémantique/logique dans les différentes étapes du cycle de vie de la Science des Données. Les graphes

de connaissances comme DBPedia, Yago et Wikidata sont largement utilisés comme sources de connaissances structurées et universelles dans les processus Science de Données pour annoter, valider et expliquer des résultats et aider à la prise de décision [JPC+21]. Le pouvoir expressif des langages pour la représentation des données complexes et de connaissances formelles (logiques de description, OWL) combiné aux langages de requêtes spécialisées (SPARQL, GraphQL) pose de nombreux défis de recherche dans l'analyse théorique (sémantique, compléxité, validation) et dans l'implémentation (scalabilité, optimisation) de ces modèles.

Les modèles de données vectoriels. Les représentations vectorielles des données permettent l'application de techniques avancées d'analyse de données et d'apprentissage automatique. Par conséquent, une littérature récente propose un large éventail de méthodes d'encodage vectoriel (embeddings) adaptées à divers types de données et de connaissances, notamment les données structurées [DSL+22], les graphes de connaissances [WMWG17], et les ontologies [CMZC+24]. Les fondements théoriques des approches d'embedding sont explorés dans [Gro20]. De plus, l'omniprésence des données vectorielles a catalysé une recherche intense autour des bases de données vectorielles, visant à surmonter les limites des systèmes de gestion de bases de données traditionnels face aux défis posés par le stockage et l'interrogation de vecteurs (voir [PWL24a, PWL24b] pour un état de l'art sur les bases de données vectorielles).

2.2.1.3 Analyse et fouille de données, apprentissage et algorithmes

Une tâche centrale en fouille de données est la *découverte de motifs*, c'est-à-dire d'associations de caractéristiques décrivant des données et apportant une information. Par exemple, un motif associant des fragments moléculaires peut caractériser un sous-ensemble de molécules ayant un comportement spécifique par rapport à une cible thérapeutique. Dépendant des structures analysées, cette tâche peut devenir très complexe. Les chercheurs en fouille de données ont d'abord effectué d'importants efforts sur les aspects algorithmiques afin de diminuer les temps de calcul tout en privilégiant des approches complètes [CRB04, GZ04].

Malheureusement, la masse de motifs produits est peu exploitable et de nombreuses directions de recherche se sont développées afin de cibler les meilleurs résultats selon des attentes de l'analyste : introduction de **méthodes d'optimisation** [Bie11, NDGN13], évaluation de l'**intérêt d'un motif** ou d'un ensemble de motifs [VT14, CGS18], échantillonnage dans l'espace des solutions qui peut être biaisé selon un intérêt de l'analyste [BLPG11, GS18]. En parallèle, de nouvelles méthodes sont apparues pour prendre en compte des données hétérogènes (e.g. texte, graphes [DTG+19, YH02], séquences [GCMG13], flux de données [BBG+21]) ou encore intégrer des préférences ou des rankings [dSDA+18].

Plus récemment, la communauté a élaboré des approches de *fouille interactive* dont une idée forte est d'introduire explicitement l'analyste dans le processus de découverte de connaissance [vL14, BH16]. La fouille interactive repose sur des cycles rapides d'extraction d'information où, à chaque cycle, l'utilisateur formule des retours sur ce qui l'intéresse ou pas dans l'information produite afin de converger vers une information utile. Ce champ de recherche ne s'oppose pas à l'Automatic Data Science (AutoML) qui regarde comment fixer des (hyper-) paramètres, comment automatiser des étapes de préparation des données. En effet, dans un processus de fouille, il est intéressant de déterminer quelles parties du processus sont automatisables (cf. section Section 1.3) et mettre en place pour celles-ci des méthodes automatiques et, d'autre part, comment intégrer et bénéficier de retours utilisateur. [DBDRHS19] présente un ensemble argumenté de défis.

La recherche dans le domaine de *l'apprentissage automatique* est actuellement en pleine effervescence, avec de nombreux travaux sur des thématiques clés telles que l'apprentissage gé-

néral (apprentissage actif, clustering, apprentissage par renforcement, apprentissage supervisé, semi-supervisé et auto-supervisé, analyse des séries temporelles, etc.), l'apprentissage profond (architectures, modèles génératifs, apprentissage par renforcement profond, etc.), la théorie de l'apprentissage (bandits, théorie des jeux, théorie de l'apprentissage statistique, etc.), l'optimisation (optimisation convexe et non convexe, méthodes matricielles/tensorielles, stochastiques, en ligne, non lisses, composites, etc.), l'inférence probabiliste (méthodes bayésiennes, modèles graphiques, méthodes de Monte Carlo, etc.), l'apprentissage automatique de confiance (responsabilité, causalité, équité, confidentialité, robustesse, etc.), l'intelligence artificielle explicable (XAI), ainsi que sur les différentes applications (biologie, santé, neurosciences, sciences sociales, science du climat, etc). Pour une vue d'ensemble des recherches actuelles, plusieurs états de l'art sont disponibles, notamment sur l'apprentissage automatique en général [NG20], les algorithmes d'apprentissage profond [PSY+18], l'apprentissage de représentations [ZGQ+24], l'équité en apprentissage automatique [CH24], ses responsabilités sociales [CVL21], l'IA frugale [SDSE20, TTH+23], les implications de l'apprentissage automatique en matière de sécurité et de confidentialité [RG23], l'IA explicable [DDN+23, BDEQEH+24, MTVF24], les approches neuro-symboliques [YYL+23, Mon22, BWE+23], ainsi que l'application de l'apprentissage automatique dans des domaines spécifiques, comme par exemple en santé [Kir22], en physique [SBO23], pour les systèmes de recommandations [CZW+24], pour la vérification automatique de la conception matérielle [WLY+24], l'utilisation du deep learning pour la télédétection de l'environnement géologique [HZW+23], pour mener des analyses en Single-Cell en Biologie [MDT+24] ou l'utilisation de l'IA pour les systèmes critiques pour la sécurité dans les domaines industriel et des transports [PCAB+24].

2.2.1.4 Systèmes et infrastructures

Ces dernières années, la transformation des infrastructures et des systèmes de gestion des données s'est fortement accélérée entrainée par plusieurs facteurs, notamment la généralisation de la gestion de données sur le Cloud et l'emergence de nouveaux environnements tels que l'Edge et le Fog, les avancées en machine learning et en intelligence artificielle, l'expansion de l'IoT, ainsi que le développement de nouveaux accélérateurs matériel et les avancées en matière de technologies du disque, de la mémoire et du réseau [AAA+22]. Dans le reste de cette section, nous présentons quelques directions de recherche actuelles relatives aux infrastructures et aux systèmes de gestion de données.

La *gestion des données dans les nouveaux environnements*, tels que le Cloud, l'Edge et le Fog, est un domaine de recherche actif [MFCP24, ZHZ+21, ZRL+21, WDW+23, DZLZ24, ZCR+24]. Ces environnements, dotés de caractéristiques spécifiques, introduisent de nouveaux paradigmes de calcul qui imposent une réévaluation approfondie des architectures et des principes fondamentaux des systèmes de gestion de données. Après une première vague de déploiement de bases de données sur le Cloud, qui a notamment conduit au développement du concept de bases de données en tant que service (DBaaS), une nouvelle génération de systèmes, connue sous le nom de bases de données natives du Cloud, est en train d'émerger [DZLZ24]. Ces bases de données cherchent à offrir une plus grande élasticité et des coûts réduits en introduisant des nouvelles techniques telles que la désagrégation du calcul, de la mémoire et du stockage [CZY+21, LDZ22, ZRL+21]. Parallèlement, une nouvelle tendance se développe autour du traitement de données sans serveur, c'est-à-dire l'utilisation des fonctions Cloud en tant que service (FaaS). Le FaaS se distingue comme un nouveau paradigme de calcul, influençant significativement la conception des systèmes de gestion de données [KHA+23, WT24b, SSZW24]. Par ailleurs, l'extension de la gestion des données dans des environnements volatiles tels que l'Edge et le Fog est une problématique émergente qui pose de nouveaux défis (voir [WLS+23, MFCP24] pour un état de l'art dans ce domaine). La sécurisation des données ex-

ternalisées sur un Cloud est également un sujet important qui suscite de nombreux travaux de recherche [ZHZ+21, WDW+23].

La recherche sur les *moteurs des Systèmes de Gestion de Bases de Données* (SGBDs) demeure extrêmement dynamique. Parmi les axes de recherche actifs, on trouve l'optimisation des requêtes [Nau17, Neu24, WT24a, KL24], l'optimisation automatique des bases de données (self-tuning) [MDGS24], ainsi que la gestion et l'analyse des données dans des environnements distribués et massivement parallèles [MRT+21, ABB+22, BGJ+24]. En outre, l'optimisation des procédures définies par l'utilisateur (UDFs) [FL15, HGD20, FS23, ZF24] et les approches modernes de compilation des requêtes, qui unifient compilation et interprétation [GLN+24], sont également des axes de recherche très explorés. Par ailleurs, l'impact des nouvelles architectures matérielles, telles que les GPU ou la combinaison de GPU et CPU, continue de susciter un vif intérêt, comme en témoignent les études récentes [HNB+22, CTL24, ZZC+23] (voir [RBM22] pour une synthèse). De même, les avancées dans l'utilisation des FPGA [DRF23, MTA09, JKA23] et du RDMA (Remote Direct Memory Access) [ZRL+21] offrent de nouvelles perspectives pour optimiser les performances des SGBDs. La conception de moteurs de SGBDs spécialement adaptés à l'apprentissage automatique, connue sous le nom de In-Database Machine Learning, est un domaine de recherche d'actualité [SCAK+19, WW22, XCC+24]. Parallèlement, l'utilisation de l'apprentissage automatique pour l'optimisation des performances des SGBDs, par exemple à travers l'exploitation de différentes techniques de deep learning pour l'optimisation des requêtes [MNM+21, CGCT23, LSS24, ZWDZ24], suscite également un vif intérêt.

La science des données a un impact de plus en plus important sur la consommation d'énergie en raison des volumes croissants de données traitées et des exigences de calcul de plus en plus élevées. Malgré l'importance de cette problématique pour la durabilité environnementale, les recherches sur l'optimisation énergétique dans le domaine de l'analyse et de la gestion des données restent relativement limitées. Comme exemples de travaux de recherche, [Sir17] explore des approches logicielles et matérielles visant à améliorer l'efficacité énergétique des bases de données, tandis que [Gra08] aborde les défis et les techniques pour rendre les systèmes de gestion de bases de données plus économes en énergie. [GYY+22] propose un état de l'art sur l'efficacité énergétique dans la gestion des données, mettant en lumière deux aspects critiques : les modèles permettant de quantifier précisément la consommation énergétique des opérations de gestion des données, et les techniques d'optimisation visant à gérer efficacement cette consommation.

2.2.2 Sujets de recherche liés à des domaines scientifiques spécifiques

Comme évoqué à la Section 1.1, les différents domaines scientifiques exploitent les techniques de la Science des Données pour résoudre des problématiques spécifiques, tout en apportant de nouvelles questions de recherche à cette discipline. Cette section illustre des exemples de défis posés par ces domaines et les questions de recherche qui en découlent.

Chaque discipline scientifique aborde les questions fondamentales discutées dans la section précédente de manière spécifique, en fonction de trois éléments clés : la *nature des sources de données*, la *nature des données* et la *nature des usages*. Cette section présente des exemples concrets des défis spécifiques rencontrés dans différents domaines scientifiques, notamment :

- ▷ les difficultés d'acquisition et d'intégration des données en raison de leur nature, du processus et du contexte de collecte (par exemple, dans le domaine de la santé, des géosciences et des SHS),
- ▷ les problèmes d'interprétation des données (par exemple, les données à sémantiques multiples ou multi-vues en géosciences ou les interprétations subjectives ou contextuelles en SHS),

- ▷ les problèmes de représentation des concepts du domaine scientifique sous forme d'artefacts informatiques (par exemple, la représentation des molécules sous forme de graphes en pharmacologie),
- ▷ l'enrichissement des algorithmes d'apprentissage par les connaissances et les modèles du domaine scientifique,
- ▷ les problèmes du choix des techniques d'apprentissage et de l'adaptation des algorithmes au contexte applicatif.

Les disciplines choisies pour l'illustration de ces problématiques — à savoir les géosciences, la santé, les sciences humaines et sociales, et la pharmacologie — sont **sélectionnées uniquement à des fins d'exemple, sans implication de priorité ou d'importance particulière** par rapport à d'autres domaines scientifiques.

Exemples de problèmes de Science de Données en géosciences En géosciences, la collecte, la structuration et le partage des données géologiques posent des défis complexes [1]. Dans certaines spécialités, les données sont rares car difficiles à collecter compte tenu de conditions extrêmes (e.g., environnements des plateformes glaciaires). De plus, les données de terrain sont souvent recueillies par des chercheurs à des endroits et à des moments variés, avec des résultats stockés dans des dépôts séparés et souvent déconnectés. Cette fragmentation rend les analyses à grande échelle extrêmement difficiles et complique leur reproduction [GPB+18, CCP23]. En outre, une donnée géologique peut revêtir plusieurs significations, qu'il s'agisse d'un constat géologique, d'une observation de terrain ou d'un résultat d'analyse, ce qui crée des problèmes d'interprétation en raison de la coexistence de multiples sémantiques. Les données géologiques englobent également une large gamme de spécialités (géologie, climatologie, ...), qui se manifestent par des vues distinctes sur les données, des localisations distinctes, des formats et types hétérogènes (images, mesures, textes) ainsi qu'une nature multi-résolution et multi-échelle. Pour relever ces défis, des efforts importants sont réalisés afin d'homogénéiser, intégrer et partager les données géologiques de manière cohérente. Des initiatives notables incluent l'infrastructure de recherche DATA TERRA [2], l'entrepôt de données longue traîne [3], EaSy Data [4], la directive Européenne INSPIRE [5], l'initiative GEO OSE [6] de la NSF, ainsi que le portail Européen OneGeology-Europe [7], et son versant international OneGeology [8]. Parallèlement, des recherches sont menées sur la découverte et l'accessibilité des données [TL16, Ma21], la définition et l'exploitation d'ontologies [NF15, DH24, WTL+24], ainsi que sur la traçabilité et l'interopérabilité des entrepôts de données. Les principes FAIR Data et les jumeaux numériques sont également explorés pour améliorer la gestion et l'exploitation des données géologiques [MRZ+23]. Les usages des données en géosciences sont variés et visent principalement à approfondir notre compréhension des phénomènes complexes des systèmes terrestres, caractérisés par des processus

1. c.f. conférence invitée au Symposium MaDICS 2024 (https://www.madics.fr/event/symposium-madics-6/)
2. Data Terra est une E-Infrastructure de recherche dédiée au domaine de la Terre et de l'environnement, visant à développer un dispositif global pour l'accès, le traitement et le partage des données liées à l'observation de la Terre https://www.data-terra.org/
3. Données issues de projets ou publications de recherche dont la préservation et la diffusion ne sont pas organisées de façon pérenne ou communautaire. Une partie de ces données sont dites « orphelines », n'ayant pas de service d'entrepôt dédié (https://www.data-terra.org/donnees-services/entrepot-de-donnees-data-terra/)
4. https://www.easydata.earth/
5. La Directive INSPIRE a pour but de créer une Infrastructure Européenne de Données Spatiales pour les politiques environnementales de l'UE et celles ayant un impact environnemental (https://knowledge-base.inspire.ec.europa.eu/index_en)
6. https://new.nsf.gov/funding/opportunities/geosciences-open-science-ecosystem-geo-ose
7. https://eurogeosurveys.org/projects/onegeology-europe/
8. https://onegeology.org/

non linéaires, multirésolution, multi-échelles, hétérogènes et hautement dynamiques. Les données seules ne suffisent pas pour modéliser ces phénomènes complexes. Pour relever ces défis, un agenda de recherche présenté dans [GPB+18] propose de développer une nouvelle génération de systèmes qui intègrent non seulement les données disponibles, mais aussi des connaissances spécifiques aux phénomènes étudiés. Cet agenda se concentre sur cinq axes de recherche : la représentation et la capture des processus, modèles et hypothèses scientifiques ; la collecte de données guidée par les connaissances scientifiques ; l'intégration des informations en tant que *système de systèmes* ; l'enrichissement des algorithmes d'apprentissage automatique avec des connaissances et des modèles en géosciences ; et le développement d'interfaces et de systèmes interactifs tenant compte du contexte des utilisateurs à l'aide de connaissances interconnectées.

Exemples de problèmes de Science de données en santé Une grande variété de données est disponible dans le domaine de la santé. Ces données proviennent de différentes sources et peuvent être classées en plusieurs catégories, en fonction de leur nature, leur origine, et leur format :

- ▷ données cliniques (DME [9], données de laboratroires ou de consultations et d'hospitalisation, ...),
- ▷ données d'imagerie médicale (IRM, tomodensitométrie [10], radiographies, échographies, imagerie par médecine nucléaire [11], ...),
- ▷ données génomiques et génétiques,
- ▷ données de capteurs/wearables (montres intelligentes, données de capteurs médicaux, ...),
- ▷ données de registres de santé (registres de vaccination ou de maladies spécifiques, ...),
- ▷ données administratives et de facturation, données issues des études cliniques (résultats d'essais cliniques, données d'enquêtes et de questionnaires, ...),
- ▷ données de santé publique (données démographiques, surveillance épidémiologique, ...),
- ▷ données de pharmacie (données de prescription, données de ventes de médicaments, ...),
- ▷ données issues des réseaux sociaux et des forums médicaux (sentiments des patients, données de recherche sur le web, ...).

Ces différents types de données présentent une grande diversité de formats, allant de données structurées, comme les dossiers médicaux électroniques dans les hôpitaux, à des données semi-structurées, telles que les résultats d'examens médicaux provenant d'équipements connectés, en passant par des données non structurées, comme les notes cliniques ou les images médicales. Ces données sont souvent volumineuses dans des secteurs tels que l'imagerie médicale, où les scans et les IRM génèrent d'importantes quantités d'informations, tandis qu'elles peuvent être plus rares et difficiles à collecter dans d'autres domaines, comme la recherche sur les maladies rares, les essais cliniques en phase précoce ou les études longitudinales. De plus, ces données sont souvent produites en silos, c'est-à-dire qu'elles sont générées et stockées de manière isolée au sein de différents laboratoires, équipes de recherche, institutions ou systèmes informatiques, sans interopérabilité entre eux. Cette fragmentation des données, ajoutée à la diversité des formats et les exigences en matière de confidentialité, complique leur intégration et rend difficile la création d'une vue d'ensemble cohérente, limitant ainsi leur potentiel pour des analyses globales et à

9. Dossiers Médicaux Électroniques.
10. Images en coupe transversale du corps (scanner, ...).
11. images obtenues grâce à des traceurs radioactifs pour observer la fonction des organes.

grande échelle.

Différentes questions liées à l'ingénierie des données en santé sont traitées dans la littérature, comme par exemple la gestion des grandes masses de données de santé [DSSK19, CLW20], la gouvernance des données médicales [PAVB23], l'intégration et l'interopérabilité des systèmes de santé [LTF+14, Dog12, BSB17], la gestion de la provenance [SMPJ+23], ainsi que L'utilisation des ontologies [YLW+23] et des techniques du Web Sémantique [KSR21]. Une revue des problématiques de sécurité et de confidentialité dans le domaine de la santé est présentée dans [LMGPRM23], et une analyse des stratégies basées sur la blockchain est détaillée dans [DAFKU20]. En matière d'analyse de données, une revue des techniques récentes d'IA utilisées pour diagnostiquer et prédire de nombreuses maladies telles que les cancers, les maladies cardiaques, pulmonaires, cutanées, génétiques et neurologiques est présentée dans [NKN23]. Une direction de recherche importante ces dernières années a porté sur l'application de l'apprentissage profond à une variété de processus médicaux, notamment le diagnostic automatique de maladies [SBJS21, SCKV22], la classification, l'analyse de données biomédicales, les systèmes de question-réponse dans le domaine médical [PUS22], la segmentation d'images [AFK23, KTT23] et la radiomique [12] [Bad21].

Cependant, l'utilisation de la science des données en santé présente plusieurs défis, parmi lesquels la nécessité de disposer de grandes quantités de données, notamment pour les phases d'entraînement des modèles d'apprentissage, le coût élevé et la complexité de l'étiquetage des données, la qualité des données et des techniques de curation, en particulier en présence de biais dans les jeux de données, la confidentialité des données ainsi que la complexité, le coût et la transparence des modèles d'apprentissage. De plus, l'analyse des images médicales à faible contraste constitue également un défi important. Ces défis conduisent à des travaux de recherche visant à surmonter ces obstacles, tels que l'utilisation de techniques comme l'augmentation de données, qui consiste à générer des données synthétiques à partir des données existantes pour augmenter artificiellement la taille des jeux de données et compléter les valeurs manquantes tout en protégeant la vie privée des patients [MWZY22, HEA+22] (voir également [XWW+24] pour la synthèse de données médicales non-image [XWW+24] et [GLZ24] pour une revue de l'utilisation des GAN pour la synthèse des dossiers médicaux électroniques structurés), le développement de techniques de compression de modèles telles que l'élagage, la quantification et la factorisation de rang faible [MPMF23, SYYZ22, JHT+23] pour réduire le nombre de paramètres et les ressources informatiques nécessaires tout en maintenant de bonnes performances, ou encore l'utilisation de l'apprentissage automatique pour améliorer le contraste des images médicales [CMA+19]. L'explicabilité des algorithmes appliqués aux données médicales est une problématique de recherche émergente, qui est importante pour assurer la transparence des modèles et renforcer ainsi l'acceptabilité et la confiance dans les applications de santé basées sur les techniques d'apprentissage automatique. Une revue des méthodes d'explicabilité appliquées à divers types de données médicales (images, textes, données tabulaires et multimodales) est présentée dans [HZM+23].

L'apprentissage fédéré suscite un intérêt particulier dans le domaine de la santé en raison de ses nombreux avantages. Il permet en effet de construire des modèles d'apprentissage à partir de données distribuées tout en garantissant la confidentialité des informations patients en ne partageant que les paramètres mathématiques et les métadonnées. Cette approche a pour objectif de favoriser la collaboration entre divers acteurs de la santé pour améliorer la précision des diagnostics et diminuer les biais résultant des spécificités liées aux équipements, aux populations ou aux protocoles d'acquisition associés à une seule source de données. [JPS22] présente

12. L'imagerie médicale computationnelle (ou radiomique) est une discipline récente qui consiste en l'analyse informatique d'images médicales pour les traduire en données quantitatives complexes (extraction de caractéristiques ou de mesures à partir d'images médicales).

un état de l'art des travaux sur l'apprentissage fédéré dans le secteur de la santé ainsi que ses applications en pronostic, diagnostic et workflow clinique.

L'application des modèles de langage de grande taille (LLMs) dans le domaine de la santé connaît un intérêt croissant, en particulier pour la génération de résumés médicaux, l'interprétation des dossiers médicaux, ainsi que pour l'assistance au diagnostic et à la prise de décision clinique. Par exemple, le système Med-PaLM 2, qui combine des améliorations du modèle de langage de base PaLM 2, d'un réglage (finetuning) spécifique au domaine médical et de stratégies de *prompts* originales, atteint un score record de 86,5 % sur le jeu de données MedQA [13], lui permettant ainsi d'égaler le niveau d'un expert médical [STG+23]. Nous renvoyons le lecteur intéressé vers [CKM+23] qui analyse le potentiel et les limites des LLMs dans la pratique clinique, la recherche et l'éducation médicales ainsi qu'à [XZL+24] pour un état de l'art plus complet sur les grands modèles de langage et les grands modèles de langage multimodaux en médecine. Ce dernier identifie plusieurs défis liés à l'application des LLMs dans le domaine de la santé, parmis lesquels on retrouvent ceux mentionnés précédemment et qui portent sur la nécessité d'une quantité importante de données médicales ainsi que le coût et la complexité des modèles. De plus, les LLMs médicaux doivent non seulement démontrer une expertise médicale, mais aussi une capacité à suivre les instructions tout en répondant aux exigences de sécurité et d'éthique. Cela nécessite des stratégies d'entraînement spécifiques pour améliorer leurs performances dans ce domaine. Par ailleurs, l'évaluation des LLMs en santé est complexe, car elle exige de mesurer leur précision sur des ensembles de données de référence tout en évaluant leur performance éthique et leur biais.

Exemples de problèmes de Science de données dans les sciences humaines et sociales Les questions de recherche liées aux processus de Science des Données appliqués aux sciences humaines et sociales (SHS) touchent divers aspects. Tout d'abord, les caractéristiques des données, qui en plus de leur diversité et hétérogénéité, les données SHS sont souvent riches en significations *contextuelles*, dans le sens où leur signification et leur utilité dépendent fortement du contexte dans lequel elles ont été collectées ainsi que de la source de données, et *interprétatives*, dans le sens où leur analyse nécessite une compréhension approfondie et souvent subjective. La gestion, l'analyse et l'interprétation de telles données requiert des approches spécifiques basées sur une analyse qualitative pour interpréter les significations sous-jacentes, tenant compte de la subjectivité et des perspectives diverses des participants. Par ailleurs, les données peuvent parfois être présentées sous forme de récits, ce qui exige une extraction des thèmes et concepts pertinents dans leurs contextes personnels et sociaux. La qualité des données peut varier largement, avec des informations parfois incomplètes ou biaisées, ce qui exige des méthodes robustes pour assurer leur fiabilité. En outre, les données peuvent couvrir de longues périodes historiques, nécessitant des techniques adaptées pour gérer les évolutions temporelles du contexte et des significations. La sensibilité et la confidentialité des données sont des enjeux majeurs, particulièrement pour celles qui contiennent des informations personnelles. Les données en SHS sont souvent multidimensionnelle, avec des aspects spatiaux, temporels, sociaux, économiques et politiques, et incluent à la fois des données qualitatives et quantitatives, nécessitant des méthodes adaptées pour chaque type. Enfin, les relations entre ces données peuvent être complexes et multidimensionnelles, nécessitant des techniques avancées pour leur analyse, et l'accès à ces données peut être limité par des droits d'auteur, des restrictions d'accès, ou des considérations éthiques.

En sciences humaines et sociales, une tension existe entre la confidentialité des données et la protection de la vie privée, d'une part, et l'exploitation et la publication des données, d'autre

13. Cela représente une amélioration de 19% par rapport à Med-PaLM, qui, avec un score de 67,2%, avait déjà dépassé le seuil de réussite de l'examen de licence médicale américain (USMLE).

part. Un sujet de recherche actuel concerne le développement de méthodes permettant de concilier la confidentialité des données avec leur exploitation dans des analyses, sans compromettre les résultats finaux [DRW+14].

Exemples de problèmes de Science de données en pharmacologie La prédiction des propriétés moléculaires est une tâche importante dans le domaine de la découverte de médicaments. En utilisant des méthodes dirigées par les données, cette tâche peut être réalisée avec une grande efficacité, réduisant ainsi le temps et les coûts associés à l'identification des médicaments candidats. Dans ce contexte, les *questions de représentation* sont essentielles. Les approches reposant sur des représentations basées sur les descripteurs [NJW+07] supposent que toutes les informations pertinentes pour les prédictions sont capturées par l'ensemble de descripteurs choisi, ce qui limite la capacité des modèles à aller au-delà des connaissances chimiques préexistantes [SZN+20]. En revanche, de nouvelles méthodes utilisant les techniques d'apprentissage profond sur graphes, en particulier les réseaux neuronaux de graphes (GNNs), se révèlent particulièrement prometteuses pour la prédiction des propriétés moléculaires [SZN+20, ZXGZ24]. En général, une structure chimique peut être modélisée comme un graphe, où les nœuds représentent les atomes et les arêtes représentent les liaisons. Des travaux pionniers proposent d'utiliser des réseaux de neurones convolutifs pour opérer directement sur de telles structures de graphes [DMI+15]. Ces dernières années, les méthodes d'apprentissage profond sur graphes, notamment les différents réseaux neuronaux de graphes (GNNs), ont été appliquées dans ce domaine, offrant des représentations efficaces des graphes moléculaires pour une meilleure prédiction des propriétés moléculaires (MPP). Parmi ces approches, l'architecture générale des réseaux neuronaux à passage de message (MPNNs) [GSR+17] se distingue. Plusieurs extension de cette architecture ont été proposées, comme le Direct MPNN (D-MPNN) [YSJ+19], qui intègre les attributs des liaisons tout en évitant les boucles inutiles dans le trajet de passage de message pour obtenir des informations sans redondance, les CMPNN (Communicative Message Passing Neural Network [SZN+20], qui permettent d'améliorer l'encodage des graphes moléculaires en renforçant les interactions entre les arêtes (liaisons) et les nœuds (atomes), et GSL (Graph Structure Learning) [ZXGZ24], qui permet de combiner à la fois les informations intra-moléculaires et inter-moléculaires.

Deuxième partie

MaDICS : Un GDR d'animation et de prospectives scientifiques autour de la Science des Données

Chapitre 3

Historique et bilan du GDR MaDICS

3.1 Historique de MaDICS 2015-2024

3.1.1 La création du GDR MaDICS

Le Groupement de Recherche (GDR) MaDICS a été créé sous l'impulsion du CNRS dans le cadre de sa stratégie de recherche fondamentale et interdisciplinaire sur les masses de données. Cette initiative découle du défi MASTODONS, lancé en 2011 par la Mission pour l'interdisciplinarité du CNRS, afin de pérenniser la dynamique d'échanges entre les chercheurs et favoriser l'animation scientifique d'une communauté interdisciplinaire. Après une phase de préparation, le CNRS a mandaté en janvier 2014 Christine Collet, Professeure au LIG, Grenoble, pour former un groupe projet chargé de préparer le dossier de création du GDR. Ce dossier a été présenté au comité national du CNRS au printemps 2014, et sa version révisée a fait l'objet d'un échange à l'automne 2014. Le GDR MaDICS a été officiellement créé en 2015, et son assemblée constitutive s'est tenue en juin 2015 à Lyon.

3.1.2 La période 2015-2019 : un GDR pionnier

3.1.2.1 Positionnement Scientifique

La première mandature (2015 à 2019) a été marquée par une activité de recherche autour de la gestion des grandes masses de données. Dès le départ, les données ont été reconnues comme une opportunité pour favoriser la *recherche interdisciplinaire* qui est au cœur du GDR MaDICS. Ainsi, la création de MaDICS s'inscrivait dans les premières initiatives du CNRS visant à établir la Science des Données comme un domaine de recherche, à l'instar des premiers instituts en Science des Données aux États-Unis, tels que le Center for Data Science à l'Université de New York créé en 2012 et le Berkeley Institute for Data Science créé en 2013.

3.1.2.2 Fonctionnement et Structuration

L'objectif général du GDR MaDICS est depuis ses débuts de fédérer les efforts de recherche et de stimuler la collaboration entre chercheurs de différents domaines scientifiques autour des problématiques spécifiques liées à l'exploitation et à la valorisation des données. Pour atteindre cet objectif, MaDICS a mis en place un instrument d'animation flexible pour ses activités

scientifiques, appelé les Actions. Une *Action* MaDICS permet de structurer et d'animer des activités interdisciplinaires autour de thématiques de recherche et des données spécifiques à une application et/ou un domaine scientifique, pour une durée limitée de deux ans, renouvelable une fois. Cet élément structurant, favorisant l'émergence et le dynamisme des activités, a été accompagné par d'autres outils comme la *labellisation et le soutien de manifestations* et les *réseaux de formation et d'innovation.*

3.1.2.3 Gouvernance

Le fonctionnement du GDR MaDICS est supervisé par une *Direction* (composée, durant la période 2015-2019, de Christine Collet, Vincent Claveau et Aurélien Garivier), chargée de porter le projet du GDR et de gérer les aspects administratifs et scientifiques auprès du CNRS. En parallèle, un *comité de direction* (ComDir) élargi organise le fonctionnement quotidien du GDR, incluant la création et le suivi des Actions, le soutien et la labellisation de manifestations, ainsi que les interactions avec le CNRS. Une *Assemblée des Responsables des Actions* (ARA) regroupe les responsables d'Actions pour discuter des enjeux scientifiques, établir des bilans, élaborer des perspectives et organiser des événements inter-Actions. Enfin, le GDR propose également une infrastructure informatique (site web, espace de travail partagé, liste mails) comme support à ses activités et à sa communication.

3.1.3 La période 2020-2024 : du Big Data à la Science des Données

3.1.3.1 Positionnement scientifique

Le projet scientifique du GDR sur la période suivante (2020-2024) a été marqué par la consolidation de la *Science des Données* comme enjeu majeur pour la recherche scientifique. Pendant cette période, le progrès des technologies de collecte, de production et de gestion de données, la disponibilité croissante des données dans divers domaines scientifiques, et le développement des méthodes en intelligence artificielle (IA) basées sur l'exploitation de grandes quantités de données pour générer des modèles d'apprentissage plus performants, ont renforcé l'importance des données pour la recherche scientifique.

3.1.3.2 Évolution du fonctionnement du GDR

Deux évolutions importantes concernant le fonctionnement du GDR ont marqué la période 2020-2024. Premièrement, le GDR a introduit la notion d'*Atelier* comme un outil pour faciliter la création d'Actions en permettant aux porteurs potentiels d'Actions, moyennant un soutien financier de la part du GDR, de proposer d'abord un Atelier d'une durée d'un an afin d'organiser et de construire une première communauté autour d'une Action future. Deuxièmement, après le succès du premier Symposium MaDICS en 2019 (à la fin de la première période), l'organisation d'un *Symposium MaDICS* annuel a été instaurée, réunissant la communauté MaDICS autour de conférences scientifiques sur des sujets d'actualité et de sessions dédiées aux Actions et Ateliers du GDR. Une demi-journée est spécialement destinée aux doctorants et jeunes chercheurs, incluant une session de mise en lumière de leurs travaux sous forme de posters et la présentation de quelques travaux de thèse primés par un prix.

3.1.3.3 Évolution de la gouvernance

La première période s'est achevée avec le décès tragique de la directrice du GDR MaDICS, Christine Collet et une année 2019 de transition pour la préparation du nouveau projet. Pendant cette année de transition, Pierre Gançarski a pris la direction du GDR, soutenu par Vincent

Claveau et Aurélien Garivier en tant que directeurs adjoints. L'objectif principal de cette direction intérimaire était d'assurer la stabilité et la continuité des activités du GDR (Symposium, suivi des Actions) tout en préparant le terrain pour le nouveau projet. Le comité de direction a été renouvelé pour inclure de nouveaux membres, apportant des nouvelles perspectives au GDR. En 2019, Sarah Cohen a accepté de porter le projet 2020-2024, assistée par Bruno Crémilleux en tant que directeur adjoint.

3.2 Bilan scientifique 2020-2024

Cette section fournit un bilan synthétique des activités du GDR MaDICS pour la période 2020-2024, en reprenant les différents dispositifs d'animation mis en place : Actions et Ateliers, Symposium MaDICS, SEEDS@MaDICS, Labellisation et Soutien. Elle est complétée par les listes des Actions (c.f., Annexe A.1), des Ateliers (c.f., Annexe A.2) ainsi que des manifestations labellisées et/ou soutenues (c.f., Annexe A.3) actifs pendant la période.

3.2.1 Les Actions et les Ateliers MaDICS

Pour soutenir la dynamique scientifiques et fédérer les efforts de recherche autour des Sciences des Donnes et de l'interdisciplinarité qu'elle induit, le GDR a choisi de créer deux instruments, les *Ateliers* et les *Actions*. Les *Ateliers* sont utilisés comme des phases de préparation à la création d'Actions dans le but de clarifier des problématiques émergentes et de former une communauté prête à s'y investir. Un Atelier est soutenu financièrement par le GDR et se déroule sur une période d'un an. À l'issue de cette année, l'Atelier se transforme, après une évaluation par le comité de direction du GDR, en une Action. Les *Actions*, quant à elles, sont des initiatives plus formalisées et de plus longue durée, permettant de structurer et d'animer des activités interdisciplinaires autour de thématiques précises, pour une durée initiale de deux ans, renouvelable une fois.

Depuis sa création en 2014, le GDR a mis en place 21 Actions, dont 6 sont en cours, et 8 Ateliers, dont 3 sont en cours. Le Tableau 3.1 donne la liste des Actions de MaDICS sur la période 2017-2024. Un Atelier devenu Action n'apparaît qu'une seule fois dans la liste, la période indiquée est alors celle couvrant l'Atelier et l'Action. Les Actions indiquées en italique sont susceptibles d'êtres renouvelées (après l'évaluation de la demande de renouvellement par le ComDir) pour deux ans couvrant la période 2025-2026. Le Tableau 3.2 donne la liste des Ateliers de MaDICS actifs en 2024.

Des fiches présentant les Actions et Ateliers du GDR réalisées en 2021 dans le cadre de notre communication scientifique sont disponibles à l'adresse : `http://www.madics.fr/actions/fiches-action/`. Chaque fiche est une synthèse de ce qu'est une Action MaDICS par le biais de 6 rubriques : "où ?", "quoi ?", "pour qui ?", "comment ?", "pour quoi ?" et "avec qui ?"

La Tableau 3.3 répertorie les mots-clés associés aux 21 Actions MaDICS qui ont été mises en place depuis la création du GDR.

3.2.2 Le Symposium MaDICS

Le symposium MaDICS est devenu un événement annuel majeur du GDR, offrant un cadre d'échange entre les membres de MaDICS pour présenter les travaux réalisés dans les Actions et les Ateliers. Souvent précédé par une réunion entre le comité de direction et les responsables des Actions et des Ateliers, le symposium permet de faire le bilan de l'année écoulée et de mener un travail de prospective. Ainsi, l'édition 2024 du symposium a permis de consulter la communauté

Période	**Action**
2016-2017	ADOC : Entrepôts et analyse de documents
2016-2017 et 2019-2020	ATLAS : Apprentissage, opTimisation Large-échelle et cAlculs distribuéS
2016-2019	EADM : Environmental Acoustic Data Mining
2016-2019	Imhyp : Imagerie Hyperspectrale
2016-2019	MAESTRO : MAsses de données En aSTROnomie et astrophysique
2016-2019	ReProVirtuFlow : Reproductibilité des expériences d'analyse de données scientifiques : enjeux et défis
2016-2019	PREDON : Preservation des données scientifiques
2017-2018	GRAMINEES : GRaph data Mining in Natural, Ecological and Environnemental Sciences
2017-2020	RoD : Raisonner sur les données (Reasoning on Data)
2018-2019	ARQUADS : Action de Recherche sur la Qualité des Données Scientifiques
2018-2022	LEMON : anaLysE et dynaMique des messages et cONversations radicales sur Internet
2018-2022	MADONA : Maîtriser l'Analyse interactive de DOnnées pour la NArration journalistique
2019-2022	MACLEAN : MAChine LEarning for EArth observatioN
2020-2021	PLATFORM : Impact Sociétal des Algorithmes Décisionnels
2020-2024	BigData4Astro
2020-2024	DOING : Données Intelligentes : transformer l'information en connaissance
2021-2023	ROCED - Reasoning on Complex and Evolving Data
2021-2025	HELP - Human Explainable machine Learning Pipeline
2021-2024	*SimpleText – Simplification et Adaptation de Textes*
2022-2024	*DSChem - Data Science in Chemistry*
2022-2025	*Musiscale - Modélisation multi-échelle de masses de données musicales*

Table 3.1 – Liste des Actions MaDICS (éventuellement précédées d'un Atelier) sur la période 2017-2024

Année	Atelier
2024	SaD-HN : Des Sources aux Données en Humanités Numériques
2024	DFEa : Prévention et détection des anomalies et fraudes en Agroalimentaire et dans l'Environnement
2024	TIDS : Traitement Informatique des Données de Santé

TABLE 3.2 – Liste des Ateliers MaDICS en cours en 2024

pour définir les orientations futures et recueillir des perspectives sur l'avenir de MaDICS pour le prochain mandat.

Le Tableau 3.4 résume les cinq dernières éditions du symposium MaDICS, organisées entre 2020 et 2024. Les pages web contiennent les programmes détaillés de chaque édition. Il est à noter que les éditions 2020 et 2021 se sont tenues en distanciel en raison de la crise de la COVID-19, entraînant ainsi un pic de participation pour ces deux années. Habituellement, le symposium accueille environ une centaine de participants, un niveau de fréquentation retrouvé dès 2022 avec la reprise des éditions en présentiel. L'édition 2024 a d'ailleurs établi un nouveau record, rassemblant 129 participants.

Le symposium MaDICS constitue également un lieu d'échange pour les jeunes chercheurs travaillant dans le domaine de la Science des Données. Chaque édition met en lumière le travail de deux à trois jeunes chercheurs, issus de laboratoires français, qui ont été récompensés par des prix nationaux ou internationaux pour leurs recherches. Par ailleurs, le symposium offre aux jeunes chercheurs l'occasion de présenter leurs travaux à travers des posters. Au cours des trois dernières éditions (2022, 2023 et 2024), respectivement 29, 32 et 36 jeunes chercheurs ont ainsi pu exposer leurs travaux.

3.2.3 SEEDS@MaDICS 2023

Les Semaines Études Entreprises en Data Sciences du GDR CNRS MaDICS (SEEDS@MaDICS) visent à créer des échanges entre les milieux industriels et le monde académique par le biais d'une semaine de travail sur des problèmes posés par des industriels et nécessitant des approches informatiques et mathématiques innovantes.

La première édition SEEDS@MaDICS s'est tenue du 17 au 31 mars 2023 et a réuni 20 doctorants, dont 7 femmes. Les doctorants ont travaillé sur 5 problèmes posés par 5 entreprises et organisations :

- YourData Consulting,
- BrainCube, http://fr.braincube.com
- Association Spot
- Caisse Primaire d'Assurance Maladie de Meurthe-et-Moselle
- Labcom Ditex-IFTH

Les problèmes posés concernaient : (1) la conception d'algorithmes permettant de prescrire des réglages pour un processus de fabrication industriel, (2) l'identification de la bonne approche en Sciences des Données pour extraire des informations dans des données textuelles de débat, (3) la conception d'une méthode d'extraction des mensurations à partir de scan 3D du corps humain en s'inspirant de méthodes géométriques, de vision artificielle, et de méthodes d'apprentissage, et (4) la conception de méthodes pour le repérage de factures suspectes en matière de soins den-

	Apprent./fouille	Bases de données/Big Data	Biologie	Ecologie/environ.	Histoire	Ing. données et logiciels	Santé	Représentation des connaissances	Sciences Humaines et Sociales	Sciences de la terre et de l'univers	Statistiques et maths appli.	Traitement de la langue/texte	Traitement d'images	Traitement du signal	Visualisation des données	Chimie	Données musicales, jounralisme
ADOC		++			++			+	++			++					
ATLAS	++		+				++				++						
EADM	++			++		+					+			++	+		
Imhyp	++		+							+	++		++	+	+		
MAESTRO	++	++								++					+		
Predon		+	+					+		+	++						
ReProVirtuFlow		++	++	++		++	++	+									
Graminees				++		++		+							+		
RoD		++						+									
ARQUADS	++	++	+	+			+				++						
LEMON	++							+				++					
MADONA	++											+					++
MACLEAN	++									++			++	++			
PLATFORM	+								+								
BigData4Astro	++	++				+				++	++				+		
DOING	+	++										++					
ROCED		+				++		++									
HELP	++			+		+	+										
SimpleText	+	++										++					+
DSChem	++	++					+									++	
Musiscale	++										++			++			++

TABLE 3.3 – Mots-clés des Actions MaDICS. Pour chaque mot-clé, le signe *"++"* indique une implication forte, tandis que le signe *"+"* représente une implication moyenne

Éd	Année	Lieu	Programme
2	2020	Distanciel	10 sessions scientifiques organisées par les Actions et les Ateliers, et une session bilan organisée par le GDR. Lien
3	2021	Distanciel	10 sessions scientifiques organisées par les Actions et les Ateliers. Lien
4	2022	Lyon	2 keynotes féminines, 15 sessions dont une dédiée aux jeunes chercheurs, présentation du prix de thèse et 21 exposés (2 minutes) avec posters associés. Lien
5	2023	Troyes	10 sessions scientifiques organisées par les Actions et les Ateliers, exposés invités par Guillaume Levavasseur (IPSL Climate Modeling Center) et Marc Chemillier (Centre d'Analyse et de Mathématique Sociales, EHESS), session "Mise en lumière de travaux de Jeunes Chercheuses et Chercheurs" avec doctorants, postdoctorants et jeunes recrutés CR (CID 55), et un gong show de 38 posters. Lien
6	2024	Blois	9 sessions scientifiques organisées par les Actions et les Ateliers, session spéciale avec une présentation de la politique scientifique de l'Institut CNRS Sciences Informatiques par Anne Siegel (Directrice adjointe scientifique Interdisciplinarité et Interfaces, CNRS Sciences Informatiques) et du projet MaDICS 2025-2029 par la direction du GDR, deux conférences invitées par Vincent Guigue (AgroParisTech) et Christelle Loiselet (BRGM), une session "Mise en lumière de travaux de Jeunes Chercheuses et Chercheurs" par Lucas Simonne (Ekimetrics) et Edwige Cyffers (INRIA), et Gong Show – Flash talk précédant la session posters des doctorants. Lien

TABLE 3.4 – Éditions du Symposium MaDICS sur le période 2020-2024

taires. Une description détaillée des sujets proposés est disponible sur la page SEEDS@MaDICS 2023 : https://www.madics.fr/manifestations/seedsmadics-2023/

3.2.4 Regards croisés : qu'est-ce que la Science des Données pour vous ?

En 2022, le GDR a entrepris un travail d'explication et de promotion de l'interdisciplinarité en tant qu'élément central en Science des Données. Le but était d'expliquer et promouvoir la place prépondérante de l'interdisciplinarité en sciences des données en l'illustrant avec des cas concrets. Ce travail s'est concrétisé par des interviews entre le comité de direction et chaque Action et Atelier, interview fondée sur la technique de l'incident critique et articulé, pour un cas concret de travail interdisciplinaire, autour de questions comme "quel était le problème ?", "comment étaient les données ?", "quelles étaient les difficultés ?", "comment a été faite la validation ?"... Une mise en commun des résultats entre les responsables d'Ateliers et d'Actions et le comité de direction a eu lieu le 23 mai 2023, lors d'une journée précédent le symposium MaDICS.

Ce travail s'est appuyé aussi sur une réflexion amont avec les responsables d'Ateliers et d'Actions organisée autour des questions suivantes qui incluent un aspect prospectives :

- ▷ qu'est-ce que la Science des Données pour vous ?
- ▷ en tant que data scientist, ou en tant qu'expert d'un domaine d'application de la science de données, qu'est-ce qui a changé pour vous au cours des 10 dernières années ?
- ▷ selon vous, qu'est-ce qui va changer dans votre discipline au cours des 10 prochaines années ?
- ▷ quelles sont les difficultés de compréhension mutuelle que vous avez rencontrées, en tant que data scientist avec des collaborateurs d'un autre domaine, ou en tant qu'expert d'un domaine avec des informaticiens avec qui vous avez collaboré.
- ▷ (pour les data scientists) : est-ce que votre collaboration avec des chercheurs d'un autre domaine a mené à des solutions nouvelles (sur le plan méthodologique ou algorithmique) ou bien adapter des techniques existantes en sciences des données s'est-il avéré suffisant ?
- ▷ (pour les experts d'un domaine disciplinaire) : est-ce que votre collaboration avec des chercheurs en Science des Données a débouché sur des méthodes nouvelles ou des résultats originaux dans votre domaine ?

Ces réflexions ont notamment permis d'alimenter le projet MaDICS pour 2025-2029 que nous présentons dans ce document.

3.2.5 Manifestions labelisées et soutenues

La liste des évènements labellisés et/ou ayant bénéficié d'un soutien financier de MaDICS sur la période 2020-2024 est présentée dans l'annexe Annexe A.3.

3.2.6 Les interactions avec les autres GDR

Le paysage scientifique est structuré par plusieurs GDR qui sont soit pluridisciplinaires mais ciblant des thématiques et des types de données spécifiques (traitement des langues, bioinformatique moléculaire, chémoinformatique...), soit plus centrés sur les disciplines informatique et mathématiques appliqués et leurs sous-domaines comme c'est le cas en intelligence artificielle ou en traitement du signal. Rappelons que l'originalité de MaDICS est de placer son ancrage sur les données et son envergure dans l'interdisciplinarité (à la fois entre disciplines et dans les

sous-domaines à l'intérieur de l'informatique et des mathématiques appliquées).

MaDICS a développé plusieurs liens avec d'autres GDR par le biais de collaborations formelles, telles qu'une Action MaDICS qui a également servi de groupe de travail conjoint avec le GDR IA, ainsi que par le biais d'interactions grâce à des chercheurs et chercheuses engagés dans plusieurs GDR. Le Tableau 5.2, page 52, donne une vue globale des interactions de MaDICS avec les autres GDR. Ci-dessous, nous listons les collaborations construites en précisant la thématique et, lorsque cela est applicable, les Ateliers/Actions concernés.

GDR RADIA (Raisonnement, Apprentissage, et Décision en Intelligence Artificielle) : https://gdr-radia.cnrs.fr/

- ▷ les Actions RoD et RoCED ont été communes avec le GDR IA puis le GDR RADIA.
- ▷ l'Action HELP et le GT Explicon du GDR RADIA ont des activités associées sur l'explication et la transparence des méthodes.
- ▷ SaD-HN s'intéresse aux problématiques abordées par RADIA notamment sur le raisonnement et l'apprentissage automatique.
- ▷ le GT Caviar du GDR RADIA rassemble notamment les approches déclaratives de la fouille de données aussi présentes dans MaDICS

GDR IASIS (Information, Apprentissage, Signal, Image et ViSion) : https://gdr-iasis.cnrs.fr/ :

- ▷ organisation d'une journée commune avec l'Action Musiscale.

GDR MAGIS (Méthodes et Applications pour la Géomatique et l'Information Spatiale) : https://gdr-magis.cnrs.fr/

- ▷ l'Action MACLEAN était fortement liée au GDR MAGIS avec l'organisation de 2 journées communes avec MACLEAN en 2019 et 2021.
- ▷ l'atelier SaD-HN a des liens étroits avec le GDR MAGIS qui s'intéresse notamment à la spatialisation d'information et à la gestion de données spatialisés avec un intérêt particulier pour la géomatique et la géographie historique. Deux actions de recherche notables de MAGIS : Graphes de Connaissances Géohistoriques et Humanités Numériques Spatialisées (plutôt orientée NLP).

GDR BIMMM (BioInformatique Moléculaire : Modélisation et Méthodologie), anciennement BIM : http://www.gdr-bim.cnrs.fr :

- ▷ labélisation de plusieurs journées BIM

GDR BigDataChim (Big data en chimie) : http://gdr-bigdatachim.cn.cnrs.fr/web/accueil :

- ▷ les problématiques scientifiques de l'Action DSChem comme les small data et le machine learning relèvent aussi du GDR BigDataChim et un nombre conséquent de chercheurs de DSChem sont aussi membres du GDR BigDataChim.

GDR IG-RV (Informatique Géométrique et Graphique, la Réalité Virtuelle et la Visualisation.) : https://gdr-igrv.fr :

▷ *Observation 3D : outils et verrous*, Journées inter-GdR CNRS MAGIS-MADICS-IGRV, 24 et 25 novembre 2021

Par ailleurs, MaDICS a co-organisé les journées GreenDays 2023 et 2024 avec le **GDS EcoInfo** [1] et les GDR suivants :

▷ **GDR ROD** Recherche Opérationnelle et Décision (http://gdrro.lip6.fr)

▷ **GDR RSD** Réseaux et Systèmes Distribués (https://gdr-rsd.fr)

▷ **GDR SOC2** System On Chip, Systèmes embarqués et Objets Connectés (https://www.gdr-soc.cnrs.fr)

▷ **GPL** Génie de la Programmation et du Logiciel (https://gdr-gpl-ie.lacl.fr

▷ **GDR IASIS** [2] (Information, Apprentissage, Signal, Image et ViSion) https://gdr-iasis.cnrs.fr/

3.2.7 Les membres du GDR MaDICS

La base de données de MaDICS répertorie un total de 1788 membres, dont 1076 nouvelles adhésions depuis janvier 2020. On note une parité relativement équilibrée, avec 55% d'hommes et 45% de femmes. Un fait marquant à noter est que quatre Décodeuses du Numérique (2022) [3] sont des membres actives du GDR MaDICS : Sarah Cohen-Boulakia, Françoise Conil, Anne-Cécile Orgerie et Marie-Christine Rousset.

La Figure 3.1 illustre la répartition des membres de MaDICS selon leur grade. Une part significative des membres du GDR MaDICS, environ 97%, est composée d'académiques, avec une bonne participation des jeunes chercheurs (24% de doctorants et 7% de post-doctorants).

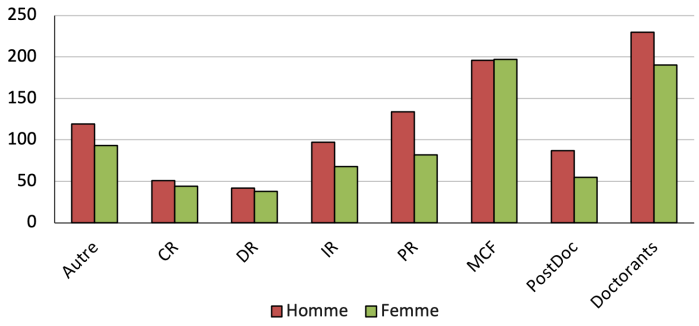

FIGURE 3.1 – Grades des membres de MaDICS.

Les membres de MaDICS assurent une bonne couverture des instituts du CNRS, avec un total de 10 instituts représentés, comme illustré par la Figure 3.2. Cette diversité témoigne de la forte dimension interdisciplinaire du GDR. Il convient de souligner une bonne représentation de l'INS2I (actuellement Institut des Sciences Informatiques), l'informatique étant une discipline centrale en Science des Données.

La Figure 3.3 montre les intérêts de recherche des membres du GDR, lesquels couvrent l'intégralité du cycle de vie de la Science des Données.

1. https ://ecoinfo.cnrs.fr
2. Le GDR IASIS est co-organisateur uniquement de l'édition 2024 de ces journées.
3. https://www.ins2i.cnrs.fr/fr/les-decodeuses-du-numerique

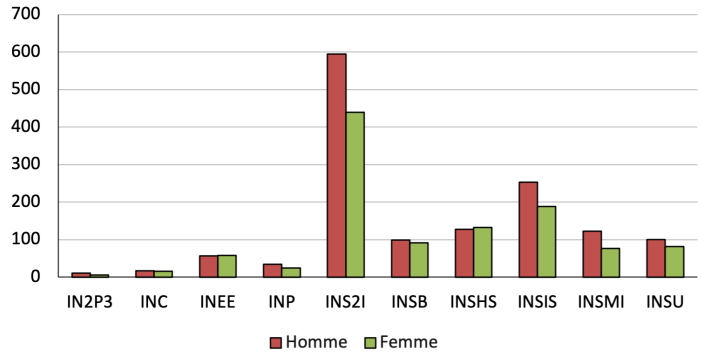

FIGURE 3.2 – Instituts CNRS impliqués dans MaDICS

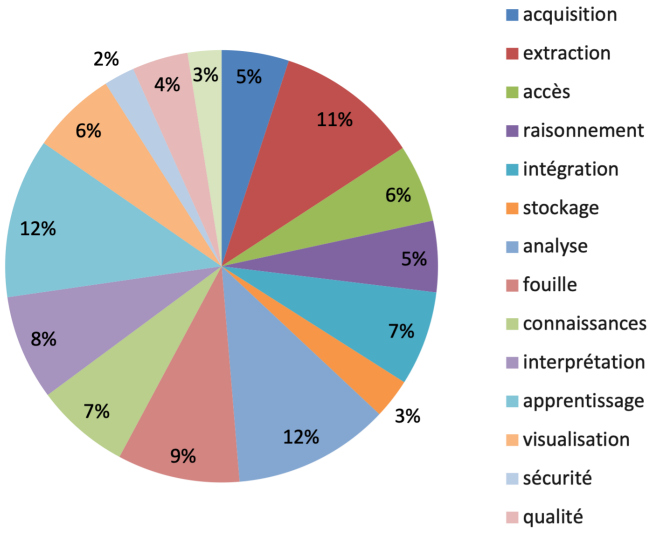

FIGURE 3.3 – Thèmes de recherche

La Figure 3.4 montre la participation des membres du GDR dans les différentes Actions et Ateliers. Pour les Actions, la participation se situe entre 36 et 144 membres, tandis que pour les Ateliers, elle varie entre 13 et 18 membres. Cela démontre bien le rôle des Ateliers comme un outil de gestation et de maturation des Actions.

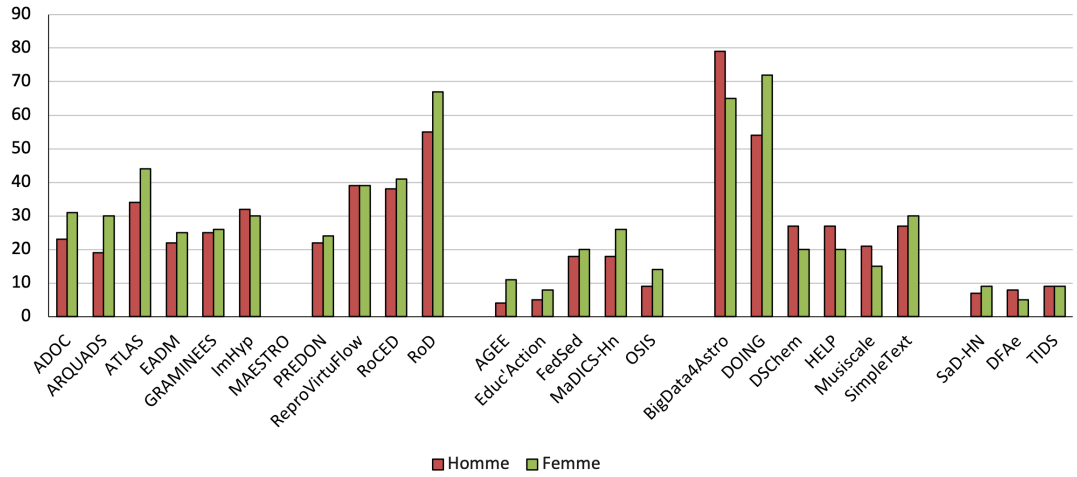

FIGURE 3.4 – Ateliers et Actions de MaDICS depuis 2014.

3.3 Analyse SWOT

	Impacts positifs	Impacts négatifs
Interne	**Forces** ▷ Couverture d'un large spectre de disciplines scientifiques ▷ Outils de structuration scientifique fédérateurs et agiles (Ateliers, Actions, Symposium, ...) ▷ Très bonne participation des laboratoires à l'échelle nationale ▷ Activité de co-labellisation : Actions et évènements avec d'autres GDR	**Faiblesses** ▷ Pérennisation des thématiques à la fin des Actions ▷ Manque d'interactions structurées et à plus long terme avec les autres GDR ▷ Interactions entre les Actions/Ateliers ▷ Difficultés à impliquer certains domaines scientifiques et les acteurs du monde industriel
Externe	**Opportunités** ▷ Rôle central des données et un domaine de la Science des Données en pleine expansion ▷ Multidisciplinarité et interdisciplinarité de plus en plus reconnues comme besoin par les domaines scientifiques ▷ Spécificités des problématiques data science dans les domaines scientifiques non couverts par les outils sur étagères	**Menaces** ▷ Nombreuses sollicitations avec un risque de dispersion des activités ▷ Fort besoin en ingénierie avec un risque d'affaiblissement des objectifs et des domaines d'expertise scientifiques du GDR ▷ Forte concurrence industrielle qui entraîne des évolutions technologiques profondes et rapides avec un risque de perte de compétitivité de la recherche académique

Chapitre 4

Évolutions de MaDICS pour la mandature 2025-2029

4.1 Objectifs visés sur la mandature

Fort de son expérience et de sa communauté scientifique interdisciplinaire bien établie, le GDR se fixe plusieurs objectifs stratégiques pour la nouvelle mandature :

- ▷ renforcer le travail de prospectives pour anticiper les évolutions technologiques et méthodologiques, identifier les nouvelles tendances et défis dans le domaine de la Science des Données, et orienter la recherche vers des problématiques émergentes (Section 4.2).
- ▷ poursuivre et renforcer la dynamique interdisciplinaire en consolidant les collaborations existantes, en explorant de nouveaux domaines scientifiques et en renforçant les interactions avec les autres GDR par la mise en place d'*actions structurantes* (Section 4.3) ;
- ▷ élargir ses structures d'animation, par la mise en place de *groupes de travail thématiques*, pour permettre le traitement de questions fondamentales en Science des Données et impliquer les communautés scientifiques concernées (Section 4.3) ;
- ▷ intégrer les questions d'éthique [1] ainsi que les objectifs de développement durable dans ses activités (Section 4.3).

4.2 Évolutions sur le volet des prospectives

Pour la nouvelle mandature, le GDR MaDICS renforcera son rôle de veille scientifique et d'élaboration de prospectives de recherche en Sciences des Données. L'objectif principal est d'anticiper les évolutions technologiques et méthodologiques, ainsi que d'identifier les tendances émergentes et les défis futurs. Pour cela, le GDR mettra en place les initiatives suivantes :

- ▷ Un *conseil scientifique* composé d'experts pour identifier les thématiques émergentes, et dont les membres pourront également être sollicités pour contribuer aux réflexions sur les prospectives.
- ▷ Des *cahiers de la prospective en Science des Données*, sous forme de publications interactives et dynamiques, permettront une exploration approfondie des tendances et des défis futurs, avec la participation d'experts sur des sujets spécifiques. Les publications pourront prendre

1. L'éthique peut concerner les données, les algorithmes ou les pratiques.

différentes formes telles que des articles courts thématiques, des entretiens, des études de cas illustrant des applications concrètes de technologies émergentes, etc. et seront publiées sur le site web du GDR.

- ▷ Des *journées prospectives*[2], réunissant des chercheurs pour discuter les orientations futures de la recherche en Science des Données autour de thèmes spécifiques.

Enfin, le GDR encouragera activement la publication de livres blancs et de rapports stratégiques pour diffuser les connaissances acquises et influencer les politiques de recherche et de développement dans le domaine de la Science des Données.

4.3 Évolutions sur le volet de l'animation scientifique

Le GDR MaDICS conservera ses instruments interdisciplinaires, les Actions et les Ateliers, qui ont montré leur intérêt durant les mandatures précédentes. À ces deux instruments s'ajoutent désormais les *Groupes de Travail (GT) thématiques*. Ces derniers se focalisent sur les thématiques clés de MaDICS et complètent ainsi les actions et les ateliers, qui restent étroitement liés à l'application de ces thèmes dans des domaines spécifiques. Ils permettent d'impliquer la communauté scientifique s'intéressant aux questions fondamentales dans le domaine de la Science des Données. Les GT serviront de socle pour fournir des expertises scientifiques aux Actions et Ateliers et favoriseront également les interactions avec d'autres GDR axés sur les questions fondamentales, tels que les GDR RADIA, Sécurité Informatique, Éco-Info, etc.

Les GT s'inscriront dans les quatre axes scientifiques présentés ci-dessous et qui correspondent aux quatre piliers de la Science de Données auxquels le GDR MaDICS et sa communauté scientifique peuvent apporter des contributions fondamentales :

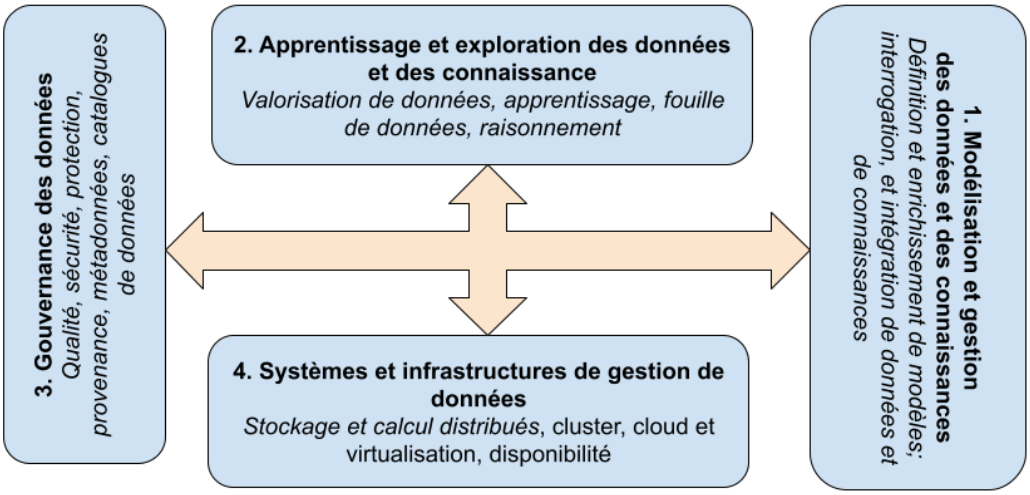

FIGURE 4.1 – Axes Scientifiques de MaDICS

2. Inspiré des séminaires Schloss Dagstuhl.

Axe 1 : Modélisation et gestion des données et des connaissances Cet axe se focalise sur les objets de recherche que sont les données, les modèles pour les représenter et les opérateurs pour les gérer. Il a pour objectif d'héberger les travaux sur la représentation (symbolique et/ou numérique) et la manipulation des données. Les travaux en représentation de données peuvent par exemple concerner la *définition de modèles* permettant de capturer les dimensions d'intérêt des données hétérogènes incluant les informations sur leur provenance ou qualité, la *comparaison de modèles* en termes d'expressivité pour déterminer leur capacité à représenter des données, ou encore l'*enrichissement de modèles* avec des connaissances ou des métadonnées. Les travaux en manipulation de données peuvent par exemple concerner l'interrogation, l'intégration ou la fusion de données, ou encore la transformation, l'échange ou l'interfaçage de données entre différents modèles de représentation.

Axe 2 : Apprentissage et exploration des données et des connaissances Cet axe s'intéresse essentiellement aux objets de recherche de type opérateur pour mettre en œuvre des tâches de valorisation des données de haut niveau. Il a pour objectif d'héberger les travaux sur les nouvelles méthodes statistiques et logiques adaptées à l'analyse et l'exploration de données et de connaissances hétérogènes comme par exemple l'apprentissage automatique pour le développement et l'utilisation de modèles capables d'apprendre à partir de données, la fouille de données pour l'analyse et l'exploration des données hétérogènes ou encore le raisonnement à l'aide de connaissances pour enrichir les analyses et améliorer la pertinence des résultats.

Axe 3 : Gouvernance des données Cet axe de recherche se concentre sur les aspects fondamentaux de la qualité, de la sécurité, de la confidentialité et de la conformité des données tout au long de leur cycle de vie. L'objectif est de développer des cadres et des outils pour organiser, protéger et gérer les données, en veillant à leur utilisation éthique et responsable. Les défis spécifiques incluent la gestion des métadonnées (annotation automatique, inférence de types, évolution de schéma), la provenance des données (passage à l'échelle), le nettoyage des données, la sécurité et l'intégrité des données, la gestion des données incomplètes et inconsistantes ainsi que l'intégration des principes de la Science des Données FATE (Fairness, Accountability, Transparency, Ethics) et des données FAIR (Findable, Accessible, Interoperable, Reusable).

Axe 4 : Systèmes et infrastructures de gestion de données Cet axe de recherche se concentre sur les défis liés au développement et à l'optimisation des systèmes et infrastructures pour une gestion efficace des données, en répondant aux besoins croissants de volume, de diversité et de vitesse. Il aborde les défis techniques liés à l'architecture, au déploiement, à l'évolutivité et à la performance. Les travaux visent à concevoir des solutions robustes et évolutives pour le stockage, la récupération, le traitement et la distribution des données, tout en garantissant leur intégrité et disponibilité. Les problématiques spécifiques incluent la gestion des données sur le Cloud (SaaS, lac de données, Cloud hybride, ...), les moteurs de gestion de données (couplage matériel-logiciel, SGBDs pour l'apprentissage automatique, ...) pour la sciences des données ainsi que les systèmes de gestion des données responsables (nouveaux modèles de coûts, ...).

Ces quatre axes ne sont pas exclusifs les uns des autres. Ils offrent un cadre structuré pour organiser et donner de la visibilité aux contributions scientifiques, en fonction des principaux défis (ou verrous) ciblés par chaque axe. Ils ont vocation à accueillir, dans leur thématique respective, une diversité de contributions théoriques (par exemple : résultats de décidabilité, complexité ou d'expressivité), algorithmiques (par exemple : résultats d'approximation, justesse, correction, complétude ou de complexité) et expérimentales (par exemple : évaluations comparatives, benchmarks à diffuser à la communauté).

4.4 Évolutions sur le volet de la gouvernance du GDR

Pour la prochaine mandature, la structure de gouvernance du GDR MaDICS fera l'objet de transformations, incluant la suppression du bureau du GDR et une redéfinition des rôles au sein du comité de direction. Ces ajustements ont pour but de renforcer la coordination et l'efficacité en simplifiant les processus décisionnels, en introduisant de nouveaux rôles et en augmentant le nombre de membres du comité de direction, afin de mieux répondre aux besoins émergents et aux sollicitations croissantes.

Parallèlement, le GDR MaDICS procédera à un renouvellement partiel de son comité de direction, avec l'intégration de trois nouveaux membres et le départ de deux anciens. La direction sera assurée par l'équipe actuelle, mise en place en janvier 2024, pour préparer le projet.

Chapitre 5

Organisation et pilotage du GDR

5.1 Animation scientifique

Cette section présente les instruments d'animation qui seront mis en place par le GDR, les modalités de constitution, ainsi que le programme prévu pour l'année 2024-2025.

5.1.1 Les instruments de l'animation scientifique

Pour la prochaine mandature, les activités d'animation du GDR seront structurées autour des Axes, des Actions, des Ateliers et des Groupes de Travail (GT). À ces instruments s'ajoutent deux événements récurrents, le symposium MaDICS annuel et une journée industrielle. Les échanges entre secteur industriel et monde académique sont complétés par la Semaine d'Études Entreprises en Data Sciences (SEEDS@MaDICS). Ces instruments et événements sont détaillés ci-dessous.

Axes scientifiques Les axes scientifiques, décrits dans la Section 4.3, sont de nouveaux outils de structuration qui seront introduits lors de la prochaine mandature. Ils sont destinés à jouer un rôle central dans l'organisation des activités du GDR, en structurant les Ateliers, Actions et Groupes de Travail (GT) autour des thématiques fondamentales. Les axes servent de grille de lecture et d'outil de pilotage, permettant d'évaluer la couverture thématique et d'identifier les éventuelles lacunes pour orienter la création de nouveaux Ateliers ou GT et de guider la réflexion prospective. Les axes contribuent également à décloisonner les activités des Actions et Ateliers en les alignant avec les questions fondamentales traitées dans les GT, favorisant ainsi une meilleure transversalité au sein du GDR. En conséquence, l'organisation du symposium évoluera avec la mise en place de sessions d'axes, regroupant les GT, Actions et Ateliers d'un même axe autour d'un programme scientifique. Ces sessions seront organisées par des comités d'axe, composés des animateurs des Ateliers, des Actions et des GT rattachés à l'axe, sous la coordination d'un membre du comité de direction du GDR.

Action Une Action au sein du GDR MaDICS est un espace d'animation scientifique qui réunit divers acteurs, incluant des producteurs et consommateurs de données ainsi que des scientifiques issus de multiples disciplines. Sous la responsabilité d'animateurs issus de différents domaines (STIC, SHS, Vie et Santé...) et à travers diverses activités (journées thématiques, écoles d'été, études prospectives, défis scientifiques, etc.), une Action vise à promouvoir et amplifier les échanges autour de problématiques de recherche interdisciplinaires, couvrant un continuum des données aux connaissances et à la prise de décision. Construire une Action MaDICS requiert une

bonne connaissance de l'écosystème interdisciplinaire du GDR pour déterminer un programme de travail et identifier les acteurs pertinents. Chaque Action est initialement établie pour une durée de deux ans et peut être renouvelée une fois.

Atelier Préfigurant une Action MaDICS, un Atelier traite une problématique de recherche commune à des partenaires issus de disciplines variées, en s'appuyant sur des données scientifiques identifiées. Comme une Action, un Atelier est un lieu d'animation avec l'objectif de constituer une première communauté de chercheurs autour des thématiques d'une Action à venir et de définir, en collaboration avec les responsables du GDR, les contours et les attentes de cette Action. Un Atelier est mis en place pour une durée d'un an, sans possibilité de renouvellement. À l'issue de cette période, une Action de deux ans, renouvelable une fois, pourra être proposée en fonction des résultats et des besoins identifiés.

Groupes de Travail (GT) Les GT sont des outils d'animation nouvellement introduits par le GDR MaDICS pour la prochaine mandature. Conçus pour fonctionner sur toute la durée de la mandature, soit cinq ans, les GT se concentrent sur des thématiques clés, structurées autour des quatre axes scientifiques. Chaque GT est affilié à un axe principal et peut être rattaché à d'autres axes secondaires. La mission d'un GT est d'approfondir ses thématiques clés et de mener une réflexion prospective sur les évolutions futures du domaine. En outre, les GT apportent leur expertise aux Actions et Ateliers en lien avec ces thématiques fondamentales.

Symposium MaDICS Le GDR MaDICS poursuivra l'organisation d'un symposium annuel, qui vise à réunir la communauté scientifique autour de conférences sur des sujets d'actualité, de sessions dédiées aux activités du GDR, et d'une demi-journée spécialement consacrée aux doctorants et jeunes chercheurs. Le symposium, qui se déroule sur deux jours, est un moment clé dans la vie du GDR. Il permet de renforcer les liens au sein de la communauté, partager les avancées scientifiques, et explorer de nouvelles perspectives de recherche.

Une évolution dans l'organisation de ce symposium consiste en la mise en place de *sessions d'axes*, organisées par les comités d'axes. Ces sessions favoriseront les synergies entre les GT, les Actions et les Ateliers, en offrant un cadre structuré pour l'échange d'idées sur les thématiques des axes et l'exploration de leurs interactions avec les problèmes interdisciplinaires abordés dans les Ateliers et les Actions.

Journées industrielles Les journées industrielles, organisées une à deux fois par an, visent à renforcer les échanges entre le monde académique et les entreprises. Elles permettent aux chercheurs de mieux cerner les besoins et enjeux des industriels, tout en offrant aux entreprises l'opportunité de s'informer sur des avancées scientifiques et technologiques pertinentes pour leurs activités. En outre, ces évènements visent à favoriser la création de partenariats et à stimuler des projets de recherche appliquée. Ces journées peuvent être co-organisées avec d'autres partenaires tels que des GDR ou des sociétés savantes.

Journées SEEDS Les Semaines Études Entreprises en Data Sciences (SEEDS@MaDICS) visent à faciliter les échanges entre le secteur industriel et le monde académique en organisant une semaine de travail intensif sur des problématiques apportées par des entreprises et nécessitant des approches informatiques et mathématiques innovantes.

Les défis industriels sont présentés et discutés le premier jour (lundi matin) et sont ensuite abordés par des groupes de 4 à 5 jeunes chercheurs (doctorants et post-doctorants). Les avancées sont présentées oralement le dernier jour (vendredi). Chaque groupe peut bénéficier du soutien de chercheurs plus expérimentés, tout en conservant une totale liberté dans l'orientation de ses

travaux. Une synthèse écrite des résultats est remise à l'entreprise en fin de semaine, et un rapport détaillé est rédigé et mis à disposition du public dans les mois qui suivent.

Les SEEDS@MaDICS, inspirées des SEME de l'AMIES et du modèle des European Study Groups with Industry initiés au Royaume-Uni dans les années 1960, constituent une initiative importante du GDR. La première édition s'est tenue du 27 au 31 mars 2023 à l'Université de Technologie de Troyes (UTT) (cf. section Section 3.2.3), et une deuxième édition est prévue pour décembre 2024.

5.1.2 Modalités de constitution

Création d'Ateliers/Actions/GT Le comité de direction procédera selon deux approches pour la création des Actions, Ateliers et Groupes de Travail (GT) au sein du GDR MaDICS : une approche top-down, en sollicitant directement des porteurs sur des thématiques jugées prioritaires pour le GDR, et une approche bottom-up, à travers un appel à projets annuel ouvert à l'ensemble de la communauté scientifique. L'appel à projets est suivi d'une phase de sélection, au cours de laquelle les propositions sont évaluées par le comité de direction selon des critères définis et connus des porteurs des propositions. Les propositions font l'objet d'un dialogue avec leurs porteurs, afin d'affiner les contenus et d'élaborer des demandes d'Ateliers, d'Actions et de GT correspondant aux objectifs et priorités du GDR.

Les Actions sont créées à la demande des porteurs d'Ateliers. Pour transformer un Atelier en Action, les porteurs doivent soumettre une demande de création d'Action à la fin de l'Atelier (en novembre), qui doit inclure une réponse aux recommandations éventuelles du comité de direction du GDR formulées au cours de l'accompagnement de l'Atelier. Cette demande est ensuite évaluée par le comité de direction à la fin de l'année. L'évaluation prend en compte la pertinence des travaux menés durant l'Atelier, l'alignement avec les objectifs stratégiques du GDR, et la capacité du projet à évoluer en une Action plus structurée et durable. Si l'évaluation est positive, l'Action démarre officiellement au 1^{er} janvier de l'année suivante, avec une durée initiale de deux ans, renouvelable une fois.

Modalités d'organisation du Symposium et des Journées SEEDS@MaDICS L'organisation de chaque édition du Symposium MaDICS est confiée à un comité d'organisation local, sous la coordination d'un correspondant membre du comité de direction. Le comité de direction prospecte les organisateurs potentiels, qui sont désignés l'année N-1 pour préparer l'édition de l'année N.

Les journées industrielles et les journées SEEDS@MaDICS sont, quant à elles, organisées par les responsables du réseau innovation du comité de direction.

5.1.3 Programme d'animation scientifique envisagé en 2024-2025

Le programme envisagé pour 2024-2025 prévoit la poursuite des Actions suivantes [1] :

- ▷ HELP : Human Explainable machine Learning Pipeline
- ▷ Musiscale : Modélisation multi-échelle de masses de données
- ▷ DSChem (Data Science in Chemistry) et
- ▷ SimpleText (Simplification et Vulgarisation des Textes Scientifiques)

1. Les Actions DSChem et SimpleText sont destinées à être renouvelées pour deux années supplémentaires après évaluation à mi-parcours.

Parallèlement, les Actions suivantes seront conclues fin 2024 :

- BigData4Astro (Big Data for Astronomy),
- DOING (Données Intelligentes : transformer l'information en connaissance),

Les ateliers ci-dessous devront se poursuivre sous forme d'Actions, après une évaluation par le comité de direction du GDR à l'automne 2024 :

- SaD-HN : Des Sources aux Données en Humanités Numériques
- DFAe : Prévention et détection des anomalies et fraudes en Agroalimentaire et dans l'Environnement
- TIDS : Traitement Informatique des Données de Santé

Un appel à Ateliers, diffusé en juin 2024 (http://www.madics.fr/ateliers/actions-appel/), donnera lieu à la création de nouveaux Ateliers. Dans le cadre de sa démarche proactive, et s'appuyant sur l'intérêt confirmé lors d'une conférence invitée au dernier symposium MaDICS, le comité de direction proposera la création d'un Atelier consacré à la *Science des Données en Géosciences*. Par ailleurs, le comité de direction a amorcé des échanges avec des chercheurs visant à la création d'Ateliers sur les thématiques suivantes : la standardisation d'un entrepôt de données de données d'imagerie afin de permettre une exploitation multicentrique des données ; l'apport des méthodes d'exploration des données (comme par exemple des motifs de co-évolution) en paléoécologie ; le traitement des données ouvertes dans un contexte de construction d'infrastructures de "science citoyenne" pour des données de santé.

Parallèlement, il sollicitera des équipes pour former des GT alignés avec les axes thématiques. La création d'un premier GT commun avec le GDR RADIA sur la *Représentation des connaissances et le raisonnement pour les données* est prévue en 2025.

La prochaine édition de SEEDS@MaDICS se déroulera en janvier 2025 à Troyes, tandis que l'édition 2025 du Symposium MaDICS est prévue à Toulouse au printemps prochain.

5.2 Pilotage du GDR

5.2.1 Comité de direction

Le comité de direction est l'organe de pilotage du GDR MaDICS. Il élabore les appels à projets, évalue les propositions soumises par la communauté scientifique et instruit les demandes de labellisation et de soutien aux manifestations scientifiques, garantissant ainsi leur qualité et leur alignement avec les objectifs du GDR. Chargé d'assurer le bon fonctionnement du GDR, il assure les échanges avec l'Assemblée des Ateliers et des Actions (ARA) et veille à la bonne organisation du symposium annuel. Le comité est également responsable de l'organisation des réunions du conseil scientifique. Il est l'interlocuteur principal du CNRS Sciences Informatiques pour les activités du GDR et assure les relations avec les autres GDR et partenaires, facilitant ainsi les collaborations et les échanges d'informations. Le comité se réunit régulièrement, à raison d'une réunion toutes les trois semaines, pour assurer une gestion efficace et réactive.

5.2.2 Composition du comité de direction

Le comité de direction proposé est composé comme suit :

- Direction
  - Directeur : Farouk Toumani, Prof., LIMOS UMR CNRS 6158, Université Clermont Auvergne, Clermont-Ferrand
  - Co-Directeur : Bernd Amann, Prof., LIP6 UMR 7606, Sorbonne Université, Paris
- Membres (ordre alphabétique)
  - Khalid Belhajjame, MdC HDR, LAMSADE UMR 7243, Université Paris-Dauphine
  - Frédéric Bimbot, DR CNRS IRISA UMR 6074, Rennes
  - Christophe Bobineau, MdC, LIG UMR 5217, Institut polytechnique de Grenoble
  - Bruno Crémilleux, Prof., GREYC UMR 6072, Université de Caen Normandie
  - François Goasdoué, Prof., IRISA UMR 6074, Université de Rennes
  - Nathalie Hernandez, Prof., IRIT UMR 5505, Université de Toulouse
  - Myriam Maumy-Bertrand, MdC HDR, LIST3N, Université de Technologie de Troyes
  - Nathalie Pernelle, Prof., LIPN UMR 7030, Université Sorbonne Paris Nord, Paris

Chaque membre du comité de direction assure des responsabilités spécifiques, comme indiqué dans le Tableau 5.1.

Rôle	**Membres du comdir**
Relations avec l'ARA	Bruno Crémilleux, François Goasdoué
Chargé des manifestations et des écoles (labélisation, soutien, animation)	Frédéric Bimbot, Farouk Toumani
Responsable(s) du réseau Innovation	Myriam Maumy-Bertrand, Nathalie Pernelle
Site web	Christophe Bobineau, Myriam Maumy-Bertrand
Responsable de symposium	Nathalie Hernandez
Relation avec les GDR/Conférences	Nathalie Hernandez
Correspondant Europe	Khalid Belhajjame
Correspondant Science Ouverte	Bernd Amann
Référent.es Parité	Nathalie Hernandez, Frédéric Bimbot
Communication	Christophe Bobineau
Prospective	Nathalie Pernelle, Farouk Toumani

TABLE 5.1 – Rôles des membres du comité de direction

5.2.3 Assemblée des Actions, Ateliers et Groupes de travail

La direction du GDR s'appuie sur l'Assemblée des Responsables d'Actions (ARA), composée des animateurs d'Actions, d'Ateliers et des GT. L'ARA est régulièrement consultée sur les principales décisions et les choix de politique scientifique et de gestion. Elle est sollicitée pour des enquêtes, des réflexions prospectives et d'autres activités. L'ARA constitue un lieu privilégié pour le partage d'informations externes et sur les actions communes du GDR (communication, valorisation, diffusion et animation scientifique). Deux réunions régulières sont organisées avec

l'ARA : en janvier, au démarrage des Actions et Ateliers, et la veille du symposium pour finaliser l'organisation et le programme. Des rencontres supplémentaires peuvent être organisées en fonction de l'actualité.

5.2.4 Conseil scientifique

Un Conseil Scientifique (CS) fournit au GDR une perspective sur la réalisation de son projet, son fonctionnement, son organisation, ainsi que sur ses orientations scientifiques et organisationnelles futures. Il contribue également à la prospective scientifique en identifiant les domaines émergents d'intérêt scientifique et sociétal pour le développement du GDR. Le CS se réunit une fois par an, avec la possibilité d'organiser des réunions extraordinaires en fonction de l'actualité.

La composition du CS est proposée par la direction du GDR, en concertation avec CNRS Sciences Informatiques, et comprend des membres dont les activités sont liées aux thématiques et disciplines du GDR, ainsi qu'au monde académique, sociétal et industriel. La taille du CS peut évoluer en fonction des avancées scientifiques du GDR. Les représentants des GDR ayant des relations existantes ou potentielles avec le GDR MaDICS sont invités aux réunions du CS.

5.3 Les budgets alloués

Le budget du GDR MaDICS est structuré pour soutenir diverses activités clés. Chaque année, 3000 € sont alloués aux Actions et aux Groupes de Travail (GT) et 1500 € aux Ateliers. En outre, des fonds sont dédiés à l'organisation du symposium annuel, y compris des bourses pour les doctorants, ainsi qu'à l'événement SEEDS@MaDICS. Un financement est également consacré au soutien des manifestations scientifiques, permettant de promouvoir et de valoriser les avancées de la communauté. Ces allocations budgétaires visent à encourager la participation des doctorants et jeunes chercheurs et à garantir le bon déroulement des initiatives scientifiques et collaboratives du GDR.

5.4 Communication

La communication du GDR est coordonnée par un membre du comité de direction. Elle repose sur plusieurs dispositifs : les listes de diffusion, le site Web, un Intranet et les réseaux sociaux.

5.4.1 Les listes de diffusion

- ▷ Liste ARA (ara@madics.fr), qui regroupe les responsables Actions, Ateliers et Groupes de travail.
- ▷ Liste CS (cs@madics.fr), nouvelle liste qui sera mise en place pour assurer l'échange d'information avec les membres du conseil scientifique du GDR MaDICS
- ▷ Liste du comité de direction (comdir@madics.fr), qui regroupe l'ensemble des membres du comité de direction de MaDICS
- ▷ Liste Annonces du GDR (annonces@madics.fr), qui regroupe l'ensemble des membres du GDR. Cette liste est utilisée pour la diffusion des différents types d'annonces (postes, offres de thèse, d'ingénieur ou de postdoc, évènement à venir, ...). Les membres ont la possibilité de se désabonner de cette liste.
- ▷ Liste de Diffusion (academique@madics.fr), qui regroupe l'ensemble des membres inscrits sur le site du GDR sans possibilité de désabonnement. Cette liste est exclusivement réservée à l'usage de la direction pour la diffusion de messages liés à la gestion du GDR (par exemple,

pour l'actualisation de la liste des membres, etc).

5.4.2 Site Web (https://www.madics.fr)

Le site web du GDR constitue un outil de communication dynamique avec des informations permanentes sur la vie et les activités du GDR. Le site MaDICS permet de diffuser des informations diverses (évènements, offres d'emplois, proposition de thèses, ...) liées aux thématiques de recherche du GDR. Ces informations sont envoyées à tous les abonnés de la liste de diffusion MaDICS et publiées dans un Calendrier public (évènements). Le site Web de MaDICS propose plusieurs outils de support et de communication ouverts à la communauté concernée par les Sciences des Données :

- une présentation du GDR et de son organisation
- des informations régulières sur les activités du GDR
- des pages dédiées aux Actions, Ateliers et GT actifs
- un calendrier des évènements
- un système d'adhésion au GDR
- un système d'inscription aux évènements
- un espace des Doctorants qui présente les différentes aides (mobilité, participation à des manifestations, ...) proposées aux doctorants ainsi que le formulaire de demande d'aide
- une page sur le réseau de formation qui présente des listes de formations, d'écoles d'été ou de ressources pédagogiques dans le domaine de la Science des Données
- une page d'offres d'emplois qui permet de proposer ou de consulter les annonces de postes, de thèses ou d'ingénieurs

5.4.3 Intranet

L'intranet du GDR MaDICS permet aux membres du GDR de réaliser les opérations suivantes :

- répondre aux appels à Actions, Ateliers et GT
- consulter les pages des Actions, Ateliers et GT en cours
- soumettre des demandes de soutien avec ou sans financement
- publier une offre de poste (Prof, MdC, autre) / postDoc / Doctorant
- diffuser une annonce (de conférence, atelier, école, webinars, hackatlon, etc.)
- avoir accès aux services réservés

5.5 Unité d'adossement du GDR MaDICS

- Code unité : UMR 6158
- Intitulé unité : Le Laboratoire d'Informatique, de Modélisation et d'Optimisation des Systèmes (LIMOS)
- Adresse principale de l'unité : 1 Rue de la Chebarde, 63178 Aubière
- Délégation régionale : Rhône Auvergne
- Nom et prénom du responsable d'unité : Mourad Baiou

▷ Adresse électronique : mourad.baiou@isima.fr

5.6 Interaction avec les autres GDR et les conférences nationales

5.6.1 Positionnement thématique et relations avec les autres GDR

Le GDR MaDICS est le seul GDR spécifiquement positionné sur la Science des Données, couvrant l'intégralité du cycle de vie des données, depuis leur collecte et gestion jusqu'à leur analyse et interprétation dans divers contextes applicatifs. La résolution des défis associés à ce domaine complexe nécessite une synergie étroite entre différentes communautés scientifiques : l'informatique, les mathématiques et les statistiques d'un côté, et les domaines d'application tels que la biologie, la physique, l'écologie, la santé, l'environnement, les sciences de l'ingénieur ainsi que les sciences humaines et sociales de l'autre.

MaDICS développe une approche interdisciplinaire qui fait partie intégrante de son ADN, constituant un pilier de son identité. Cette interdisciplinarité est essentielle pour relever les défis contemporains de la Science des Données.

Le positionnement spécifique de MaDICS ne limite pas ses interactions avec d'autres GDR ; au contraire, ses thématiques présentent des intersections avec celles d'autres disciplines, ce qui facilite des collaborations et stimule les interactions. La Tableau 5.2 montre les interactions avec les autres GDR, qu'elles soient existantes (lignes en gras), envisagées (lignes simples) ou potentielles (lignes en pointillés).

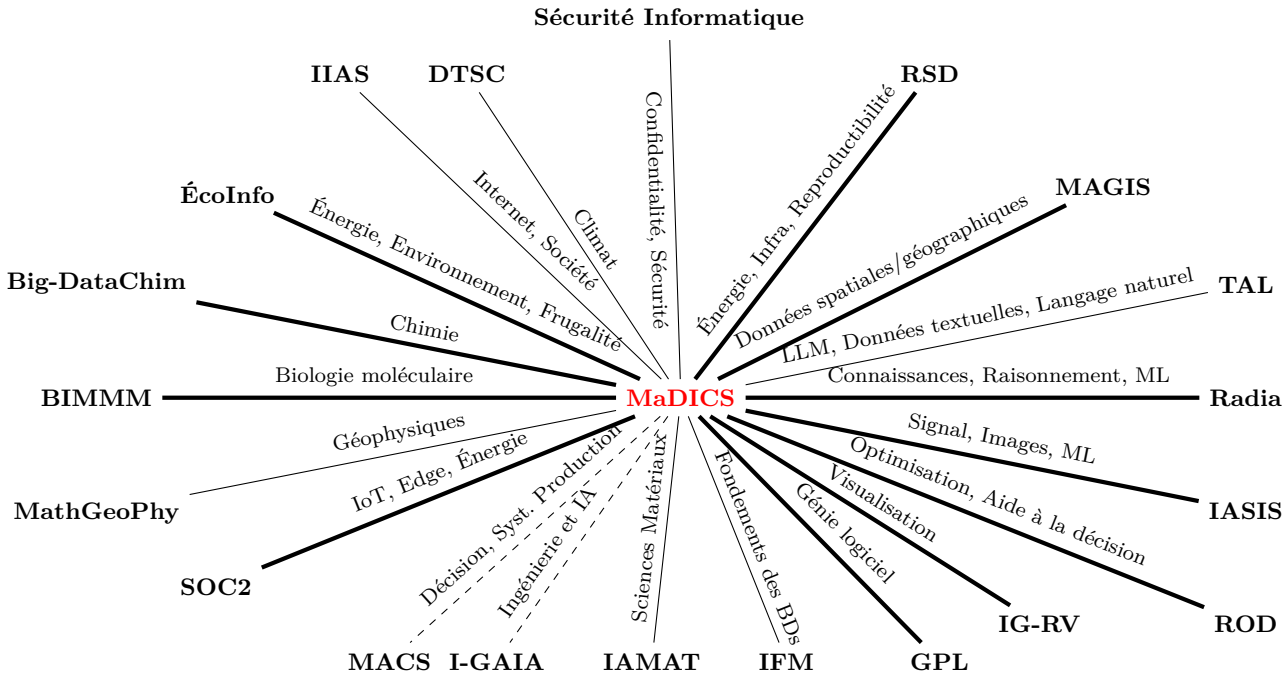

TABLE 5.2 – Relations avec les autres GDR : existantes (lignes en gras), envisagées (lignes simples) ou potentielles (lignes en pointillés)

Parmi les GDR avec lesquels MaDICS a déjà établi des collaborations, on trouve RADIA (Raisonnement, Apprentissage, et Décision en Intelligence Artificielle), IASIS (Information, Apprentissage, Signal, Image et Vision), MAGIS (Méthodes et Applications pour la Géomatique

et l'Information Spatiale), IG-RV (Informatique Géométrique et Graphique, Réalité Virtuelle et Visualisation), ROD (Recherche Opérationnelle et Décision), GPL (Génie de la Programmation et du Logiciel), RSD (Réseaux et Systèmes Distribués) et SOC2 (System On Chip, Systèmes embarqués et Objets Connectés). MaDICS collabore également avec des GDR disciplinaires tels que BigDataChim (Big Data en Chimie) et BIMMM (BioInformatique Moléculaire : Modélisation et Méthodologie). MaDICS prévoit de maintenir et de consolider les collaborations existantes avec ces GDR, afin de renforcer les synergies établies.

Concernant le GDR RADIA, et les autres GDR couvrant des thématiques de l'IA comme ROD, il est important de noter qu'il existe une intersection entre la Science des Données et l'Intelligence Artificielle (IA), mais ces deux domaines ne se recouvrent pas entièrement et ne sont pas inclusifs l'un de l'autre. Les centres d'intérêt communs entre MaDICS et ces GDR concernent principalement les thématiques autour de la représentation des connaissances et le raisonnement pour la gestion des données couverts par le GDR RADIA. Cependant, l'intérêt de MaDICS se limite à l'application de ces principes à la gestion des données, tandis que RADIA adopte une perspective plus générale sur les modèles permettant de manipuler les connaissances et de raisonner face à des enjeux tels que l'imprécision, l'incomplétude ou l'inconsistance des connaissances. Un autre domaine d'intérêt commun entre la Science des Données et l'IA réside dans l'analyse des données et l'apprentissage automatique. Certaines thématiques de MaDICS utilise des méthodes d'apprentissage en se concentrant sur les spécificités requises pour ces méthodes dans un contexte interdisciplinaire. Enfin, la qualité des données étant cruciale pour l'efficacité des méthodes d'apprentissage, le travail mené dans MaDICS sur la qualité des données peut contribuer significativement à améliorer ces méthodes.

Pour la prochaine mandature, MaDICS prévoit d'explorer les opportunités de collaboration avec les GDR disciplinaires suivants, en vue par d'organiser des événements conjoints et de créer ou de mettre en place des Ateliers communs :

- GDR Internet, IA et Société (https://cis.cnrs.fr/presentation-gdr/) : Internet, Intelligence artificielle et Société.
- GDR Défis Théoriques pour les Sciences du Climat (https://defi-theo-climat.ipsl.fr) : Sciences du climat.
- GDR MathGeoPhy (https://mathgeophy.math.cnrs.fr/) : Mathématiques, Énergie et Environnement.
- GDR IAMAT (https://iamat.cnrs.fr) : Intelligence Artificielle en Sciences des Matériaux

Pour ce qui concerne les thématiques fondamentales, MaDICS prévoit de nouer des collaborations avec les GDR suivants afin de renforcer les activités de ses Groupes de Travail (GT) :

- GDR Sécurité Informatique (https://gdr-securite.irisa.fr)
- GDR ROD : Recherche Opérationnelle et Décision (http://gdrro.lip6.fr)
- GDR RSD : Réseaux et Systèmes Distribués (https://gdr-rsd.fr)
- GDR SOC2 : System On Chip, Systèmes embarqués et Objets Connectés (https://www.gdr-soc.cnrs.fr)
- GDR IFM : Informatique Fondamentale et ses Mathématiques (https://www.gdr-ifm.fr)
- GDR TAL : Traitement Automatique des Langues (https://gdr-tal.ls2n.fr)

Enfin, MaDICS envisage également de participer au GDS EcoInfo (https://ecoinfo.cnrs.fr/), qui se concentre sur des enjeux interdisciplinaires liés à la sobriété numérique.

5.6.2 Mise en place d'actions structurées et à long terme

MaDICS continuera à promouvoir le développement d'actions communes avec d'autres GDR à travers ses appels à projets. Parallèlement, une réflexion sera initiée par le comité de direction, en étroite collaboration avec le conseil scientifique, pour élaborer et mettre en œuvre des actions structurées et à long terme avec des GDR ciblés. Cette démarche vise à identifier et concrétiser des synergies en réponse aux défis contemporains de la Science des Données. Une première initiative dans cette direction sera la création, dès 2025, d'un GT conjoint avec le GDR RADIA, centré sur la thématique de la "représentation des connaissances et raisonnement pour les données".

5.6.3 Renforcement de l'interdisciplinarité

MaDICS poursuivra sa politique proactive d'appels à projets pour la mise en place d'Actions et d'Ateliers interdisciplinaires, exigeant que les responsables proviennent de deux instituts différents du CNRS. En parallèle, MaDICS renforcera ses interactions avec les GDR disciplinaires portés par d'autres instituts du CNRS, afin de tirer parti des expertises diverses et de créer des ponts entre les domaines scientifiques. De plus, MaDICS renforcera ses liens avec la MITI (Mission pour les Initiatives Transverses et Interdisciplinaires) du CNRS et sera particulièrement attentif aux appels à soutien pour les projets interdisciplinaires lancés par cette mission, afin de mobiliser efficacement les réponses de la communauté.

5.6.4 Liens avec les conférences nationales et les sociétés savantes

MaDICS s'attachera à établir des relations régulières avec les conférences nationales positionnées sur des thématiques d'intérêt pour le GDR, telles que :

- **BDA** (Bases de Données Avancées), https://bdav.irisa.fr
- **CORIA** (Conférence en Recherche d'Information et Applications), https://www.asso-aria.org/coria/
- **EDA** (Business Intelligence & Big Data), https://eric.univ-lyon2.fr/eda/
- **EGC** (Extraction et Gestion des Connaissances), https://www.egc.asso.fr
- **GRETSI** (Groupe de Recherche et d'Etudes de Traitement du Signal et des Images), https://www.gretsi.fr
- **IC** (Ingénierie des Connaissances), https://pfia2024.univ-lr.fr/ConfÃ¨rences/IC/
- **INFORSID** (INFormatique des ORganisations et Systèmes d'Information et de Décision), http://inforsid.fr
- **JEP** (Journées d'Etudes sur la Parole), https://www.afcp-parole.org/
- **JIM** (Journées d'Informatique Musicale), http://www.afim-asso.org
- **JOBIM** (Journées Ouvertes en Biologie, Informatique, et Mathématiques), https://www.sfbi.fr/blog/article/jobim
- **SFCi** (Journées de la Société Française de Chémoinformatique)
- **TALN-RECITAL** (Traitement Automatique des Langues Naturelles), https://www.atala.org/-Conference-TALN-RECITAL

Ces relations peuvent prendre différentes formes :

- Organisation conjointe d'évènements. MaDICS peut collaborer avec les organisateurs de conférences nationales pour co-organiser différents types d'évènements tels que des tables rondes, des ateliers ou des tutoriels sur des sujets interdisciplinaires. L'objectif est de renforcer l'engagement de ces communautés au sein de MaDICS, en les sensibilisant aux enjeux interdisciplinaires et en favorisant des échanges sur les défis posés par les domaines scientifiques.
- Échanges et réseautage. Les conférences nationales offrent un espace privilégié pour les échanges informels et le réseautage entre les membres du GDR et la communauté scientifique plus large. Cela permet de créer des synergies pour faire émerger de nouvelles thématiques d'Ateliers et de trouver, par exemple, des partenaires potentiels pour des projets d'Ateliers.
- Réflexion prospective et stratégie. Les conférences nationales seront sollicitées par le GDR pour participer à des réflexions prospectives sur l'évolution du domaine. Cela peut inclure des contributions aux *cahiers de la prospective*, l'organisation de panels de discussion ou de sessions spécifiques lors de ces conférences pour discuter des orientations de recherche et des défis futurs.
- Diffusion et communication. Les conférences nationales sont un vecteur important pour diffuser les activités de MaDICS et renforcer sa visibilité. MaDICS utilise activement les listes de diffusion de ces conférences, telles que bull-i3, liste-egc, liste-ic et gazettebd3, pour toucher les communautés scientifiques sous-jacentes.

Par ailleurs, en matière de relations avec les sociétés savantes, MaDICS participe au Conseil des Associations de la SIF (Société Informatique de France) en tant que membre observateur.

Chapitre 6

Bibliographie

[AAA+22] Daniel Abadi, Anastasia Ailamaki, David Andersen, Peter Bailis, Magdalena Balazinska, Philip A. Bernstein, Peter Boncz, Surajit Chaudhuri, Alvin Cheung, Anhai Doan, Luna Dong, Michael J. Franklin, Juliana Freire, Alon Halevy, Joseph M. Hellerstein, Stratos Idreos, Donald Kossmann, Tim Kraska, Sailesh Krishnamurthy, Volker Markl, Sergey Melnik, Tova Milo, C. Mohan, Thomas Neumann, Beng Chin Ooi, Fatma Ozcan, Jignesh Patel, Andrew Pavlo, Raluca Popa, Raghu Ramakrishnan, Christopher Re, Michael Stonebraker, and Dan Suciu. The seattle report on database research. *Commun. ACM*, 65(8) :72–79, jul 2022.

[AAB+17] Renzo Angles, Marcelo Arenas, Pablo Barceló, Aidan Hogan, Juan Reutter, and Domagoj Vrgoč. Foundations of modern query languages for graph databases. *ACM Comput. Surv.*, 50(5), sep 2017.

[AAD+24] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pages 1–3, 2024.

[ABB+16] Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, Jildau Bouwman, Anthony J Brookes, Tim Clark, M Crosas, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1), 2016.

[ABB+22] Nikos Armenatzoglou, Sanuj Basu, Naga Bhanoori, Mengchu Cai, Naresh Chainani, Kiran Chinta, Venkatraman Govindaraju, Todd J. Green, Monish Gupta, Sebastian Hillig, Eric Hotinger, Yan Leshinksy, Jintian Liang, Michael McCreedy, Fabian Nagel, Ippokratis Pandis, Panos Parchas, Rahul Pathak, Orestis Polychroniou, Foyzur Rahman, Gaurav Saxena, Gokul Soundararajan, Sriram Subramanian, and Doug Terry. Amazon redshift re-invented. In *Proceedings of the 2022 International Conference on Management of Data*, SIGMOD '22, page 2205–2217, New York, NY, USA, 2022. Association for Computing Machinery.

[ABC+22] Lyes Attouche, Mohamed-Amine Baazizi, Dario Colazzo, Giorgio Ghelli, Carlo Sartiani, and Stefanie Scherzinger. Witness generation for json schema. *Proc. VLDB Endow.*, 15(13) :4002–4014, sep 2022.

[AFK23] 2 Ali F. Khalifa, Eman Badr. Deep learning for image segmentation : A focus on medical imaging. *Computers, Materials & Continua*, 75(1) :1995–2024, 2023.

[AG08] Renzo Angles and Claudio Gutierrez. Survey of graph database models. *ACM Comput. Surv.*, 40(1), feb 2008.

[AGN15] Ziawasch Abedjan, Lukasz Golab, and Felix Naumann. Profiling relational data : a survey. *The VLDB Journal*, 24(4) :557–581, aug 2015.

[Ano74] Anonymous. Predicting how proteins fold up. *NATURE*, 248(5450) :636, 1974.

[Bad21] Eman Badr. Images in space and time : Real big data in healthcare. *ACM Comput. Surv.*, 54(6), jul 2021.

[BBB+22] Ahmad Al Badawi, Jack Bates, Flávio Bergamaschi, David Bruce Cousins, Saroja Erabelli, Nicholas Genise, Shai Halevi, Hamish Hunt, Andrey Kim, Yongwoo Lee, Zeyu Liu, Daniele Micciancio, Ian Quah, Yuriy Polyakov, R. V. Saraswathy, Kurt Rohloff, Jonathan Saylor, Dmitriy Suponitsky, Matthew Triplett, Vinod Vaikuntanathan, and Vincent Zucca. Openfhe : Open-source fully homomorphic encryption library. In Michael Brenner, Anamaria Costache, and Kurt Rohloff, editors, *Proceedings of the 10th Workshop on Encrypted Computing & Applied Homomorphic Cryptography, Los Angeles, CA, USA, 7 November 2022*, pages 53–63. ACM, 2022.

[BBD+02] Brian Babcock, Shivnath Babu, Mayur Datar, Rajeev Motwani, and Jennifer Widom. Models and issues in data stream systems. In *Proceedings of the Twenty-First ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '02, page 1–16, New York, NY, USA, 2002. Association for Computing Machinery.

[BBDM+22] Antoon Bronselaer, Christophe Billiet, Robin De Mol, Joachim Nielandt, and Guy De Tré. Compact representations of temporal databases. *The VLDB Journal*, 28(4) :473–496, mar 2022.

[BBG+21] Maroua Bahri, Albert Bifet, João Gama, Heitor Murilo Gomes, and Silviu Maniu. Data stream analysis : Foundations, major tasks and tools. *WIREs Data Mining Knowl. Discov.*, 11(3), 2021.

[BCGS22] Mohamed-Amine Baazizi, Dario Colazzo, Giorgio Ghelli, and Carlo Sartiani. Parametric schema inference for massive json datasets. *The VLDB Journal*, 28(4) :497–521, mar 2022.

[BCP+23] Carlo A. Bono, Cinzia Cappiello, Barbara Pernici, Edoardo Ramalli, and Monica Vitali. Pipeline design for data preparation for social media analysis. *J. Data and Information Quality*, 15(4), nov 2023.

[BDA+21] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557) :871–876, 2021.

[BDEQEH+24] Adrien Bennetot, Ivan Donadello, Ayoub El Qadi El Haouari, Mauro Dragoni, Thomas Frossard, Benedikt Wagner, Anna Sarranti, Silvia Tulli, Maria Trocan, Raja Chatila, Andreas Holzinger, Artur d'Avila Garcez, and Natalia Díaz-Rodríguez. A practical tutorial on explainable ai techniques. *ACM Comput. Surv.*, jun 2024. Just Accepted.

[Bel21] Khalid Belhajjame. On the Anonymization of Workflow Provenance without Compromising the Transparency of Lineage. *Journal of Data and Information Quality*, 14(1) :4 :1–4 :27, December 2021.

[BFPK20] Alex Bogatu, Alvaro A. A. Fernandes, Norman W. Paton, and Nikolaos Konstantinou. Dataset discovery in data lakes. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 709–720, 2020.

[BG21] Nils Barlaug and Jon Atle Gulla. Neural networks for entity matching : A survey. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(3) :1–37, 2021.

[BGJ+24] Nicolas Bruno, César A. Galindo-Legaria, Milind Joshi, Esteban Calvo Vargas, Kabita Mahapatra, Sharon Ravindran, Guoheng Chen, Ernesto Cervantes Juárez, and Beysim Sezgin. Unified query optimization in the fabric data warehouse. In Pablo Barceló, Nayat Sánchez Pi, Alexandra Meliou, and S. Sudarshan, editors, *Companion of the 2024 International Conference on Management of Data, SIGMOD/PODS 2024, Santiago AA, Chile, June 9-15, 2024*, pages 18–30. ACM, 2024.

[BGP+23] Maciej Besta, Robert Gerstenberger, Emanuel Peter, Marc Fischer, Michał Podstawski, Claude Barthels, Gustavo Alonso, and Torsten Hoefler. Demystifying graph databases : Analysis and taxonomy of data organization, system designs, and graph queries. *ACM Comput. Surv.*, 56(2), sep 2023.

[BH16] Mansurul Bhuiyan and Mohammad Al Hasan. Interactive knowledge discovery from hidden data through sampling of frequent patterns. *Stat. Anal. Data Min.*, 9(4) :205–229, 2016.

[Bie11] Tijl De Bie. Maximum entropy models and subjective interestingness : an application to tiles in binary databases. *Data Min. Knowl. Discov.*, 23(3) :407–446, 2011.

[BLPG11] Mario Boley, Claudio Lucchese, Daniel Paurat, and Thomas Gärtner. Direct local pattern sampling by efficient two-step random procedures. In Chid Apté, Joydeep Ghosh, and Padhraic Smyth, editors, *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 21-24, 2011*, pages 582–590. ACM, 2011.

[BMD24] Martin Buttenschoen, Garrett M Morris, and Charlotte M Deane. Posebusters : Ai-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 15(9) :3130–3139, 2024.

[BNV07] Geert Jan Bex, Frank Neven, and Stijn Vansummeren. Inferring xml schema definitions from xml data. In *Proceedings of the 33rd International Conference on Very Large Data Bases*, VLDB '07, page 998–1009. VLDB Endowment, 2007.

[BPE22] P Bryant, G Pozzati, and A Elofsson. Improved prediction of protein-protein interactions using alphafold2. nat. commun. 13, 1265, 2022.

[BSB17] Subhash Bhalla, Shelly Sachdeva, and Shivani Batra. Semantic interoperability in electronic health record databases : Standards, architecture and e-health systems. In *Big Data Analytics : 5th International Conference, BDA 2017, Hyderabad, India, December 12-15, 2017, Proceedings*, page 235–242, Berlin, Heidelberg, 2017. Springer-Verlag.

[BT24] Angela Bonifati and Riccardo Tommasini. An overview of continuous querying in (modern) data systems. In *Companion of the 2024 International Conference on Management of Data*, SIGMOD/PODS '24, page 605–612, New York, NY, USA, 2024. Association for Computing Machinery.

[BWE+23] Anna Breit, Laura Waltersdorfer, Fajar J. Ekaputra, Marta Sabou, Andreas Ekelhart, Andreea Iana, Heiko Paulheim, Jan Portisch, Artem Revenko, Annette Ten Teije, and Frank Van Harmelen. Combining machine learning and semantic web : A systematic mapping study. *ACM Comput. Surv.*, 55(14s), jul 2023.

[CCDP22] Loredana Caruccio, Stefano Cirillo, Vincenzo Deufemia, and Giuseppe Polese. Efficient discovery of functional dependencies from incremental databases. In *The 23rd International Conference on Information Integration and Web Intelligence*, iiWAS2021, page 400–409, New York, NY, USA, 2022. Association for Computing Machinery.

[CCP23] Guoxiong Chen, Qiuming Cheng, and Steve Puetz. Special Issue : Data-Driven Discovery in Geosciences : Opportunities and Challenges. *Mathematical Geosciences*, 55(3) :287–293, April 2023.

[CDNP21] Loredana Caruccio, Vincenzo Deufemia, Felix Naumann, and Giuseppe Polese. Discovering relaxed functional dependencies based on multi-attribute dominance. *IEEE Trans. on Knowl. and Data Eng.*, 33(9) :3212–3228, sep 2021.

[CGCT23] Tianyi Chen, Jun Gao, Hedui Chen, and Yaofeng Tu. Loger : A learned optimizer towards generating efficient and robust query execution plans. *Proc. VLDB Endow.*, 16(7) :1777–1789, mar 2023.

[CGS18] Bruno Crémilleux, Arnaud Giacometti, and Arnaud Soulet. How your supporters and opponents define your interestingness. In Michele Berlingerio, Francesco Bonchi, Thomas Gärtner, Neil Hurley, and Georgiana Ifrim, editors, *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2018, Dublin, Ireland, September 10-14, 2018, Proceedings, Part I*, volume 11051 of *Lecture Notes in Computer Science*, pages 373–389. Springer, 2018.

[CH24] Simon Caton and Christian Haas. Fairness in machine learning : A survey. *ACM Comput. Surv.*, 56(7), apr 2024.

[CHK09] Michael J. Cafarella, Alon Halevy, and Nodira Khoussainova. Data integration for the relational web. *Proc. VLDB Endow.*, 2(1) :1090–1101, aug 2009.

[CKM+23] Jan Clusmann, Fiona Kolbinger, Hannah Muti, Zunamys Carrero, Jan-Niklas Eckardt, Narmin Laleh, Chiara Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory Veldhuizen, Sophia Wagner, and Jakob Kather. The future landscape of large language models in medicine. *Communications medicine*, 3 :141, 10 2023.

[CLW20] Peng-Ting Chen, Chia-Li Lin, and Wan-Ning Wu. Big data management in healthcare : Adoption challenges and implications. *International Journal of Information Management*, 53 :102078, 2020.

[CMA+19] Gabriel Fillipe Centini Campos, Saulo Martiello Mastelini, Gabriel Jonas Aguiar, Rafael Gomes Mantovani, Leonimer Flavio de Melo, and Sylvio Barbon Junior. Machine learning hyperparameter selection for contrast limited adaptive histogram equalization. *EURASIP J. Image Video Process.*, 2019 :59, 2019.

[CMZC+24] Jiaoyan Chen, Olga Mashkova, Fernando Zhapa-Camacho, Robert Hoehndorf, Yuan He, and Ian Horrocks. Ontology embedding : A survey of methods, applications and resources, 2024.

[CP19] Stefano Ceri and Pietro Pinoli. Data science for genomic data management : Challenges, resources, experiences. *SN Comput. Sci.*, 1(1), jun 2019.

[CRB04] Toon Calders, Christophe Rigotti, and Jean-François Boulicaut. A survey on condensed representations for frequent sets. In Jean-François Boulicaut, Luc De Raedt, and Heikki Mannila, editors, *Constraint-Based Mining and Inductive Databases, European Workshop on Inductive Databases and Constraint Based Mining, Hinterzarten, Germany, March 11-13, 2004, Revised Selected Papers*, volume 3848 of *Lecture Notes in Computer Science*, pages 64–80. Springer, 2004.

[CSK+19] Adriane Chapman, Elena Simperl, Laura Koesten, George Konstantinidis, Luis-Daniel Ibáñez, Emilia Kacprzak, and Paul Groth. Dataset search : a survey. *The VLDB Journal*, 29(1) :251–272, aug 2019.

[CTH+20] Zhiyu Chen, Mohamed Trabelsi, Jeff Heflin, Yinan Xu, and Brian D. Davison. Table search using a deep contextualized language model. In *Proceedings of the*

43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '20, page 589–598, New York, NY, USA, 2020. Association for Computing Machinery.

[CTL24] Wei-Mei Chen, Hsin-Hung Tsai, and Joon Fong Ling. Parallel computation of dominance scores for multidimensional datasets on gpus. *IEEE Transactions on Parallel and Distributed Systems*, 35(6) :919–931, 2024.

[CVL21] Lu Cheng, Kush R. Varshney, and Huan Liu. Socially responsible ai algorithms : Issues, purposes, and challenges. *J. Artif. Int. Res.*, 71 :1137–1181, sep 2021.

[CZW+24] Bo Chen, Xiangyu Zhao, Yejing Wang, Wenqi Fan, Huifeng Guo, and Ruiming Tang. A comprehensive survey on automated machine learning for recommendations. *ACM Trans. Recomm. Syst.*, 2(2), apr 2024.

[CZY+21] Wei Cao, Yingqiang Zhang, Xinjun Yang, Feifei Li, Sheng Wang, Qingda Hu, Xuntao Cheng, Zongzhi Chen, Zhenjun Liu, Jing Fang, et al. Polardb serverless : A cloud native database for disaggregated data centers. In *Proceedings of the 2021 International Conference on Management of Data*, pages 2477–2489, 2021.

[DAFKU20] Erikson Júlio De Aguiar, Bruno S. Faiçal, Bhaskar Krishnamachari, and Jó Ueyama. A survey of blockchain-based strategies for healthcare. *ACM Comput. Surv.*, 53(2), mar 2020.

[DBDRHO+22] Tijl De Bie, Luc De Raedt, José Hernández-Orallo, Holger H. Hoos, Padhraic Smyth, and Christopher K. I. Williams. Automating data science. *Commun. ACM*, 65(3) :76–87, feb 2022.

[DBDRHS19] Tijl De Bie, Luc De Raedt, Holger H. Hoos, and Padhraic Smyth. Automating Data Science (Dagstuhl Seminar 18401). *Dagstuhl Reports*, 8(9) :154–181, 2019.

[DDN+23] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, and Rajiv Ranjan. Explainable ai (xai) : Core ideas, techniques, and solutions. *ACM Comput. Surv.*, 55(9), jan 2023.

[DDW+22] Alexander Dunn, John Dagdelen, Nicholas Walker, Sanghoon Lee, Andrew S. Rosen, Gerbrand Ceder, Kristin Persson, and Anubhav Jain. Structured information extraction from complex scientific text with fine-tuned large language models, 2022.

[DH24] William Davis and Cassandra R. Hunt. Knowledge graphs for seismic data and metadata. *Applied Computing and Geosciences*, 21 :100151, 2024.

[DKS+23] Rhiju Das, Rachael C Kretsch, Adam J Simpkin, Thomas Mulvaney, Phillip Pham, Ramya Rangan, Fan Bu, Ronan M Keegan, Maya Topf, Daniel J Rigden, et al. Assessment of three-dimensional rna structure prediction in casp15. *Proteins : Structure, Function, and Bioinformatics*, 91(12) :1747–1770, 2023.

[DMI+15] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. *Advances in neural information processing systems*, 28, 2015.

[Dog12] Asuman Dogac. Interoperability in ehealth systems. *Proc. VLDB Endow.*, 5(12) :2026–2027, aug 2012.

[DR07] Hong-Hai Do and Erhard Rahm. Matching large schemas : Approaches and evaluation. *Inf. Syst.*, 32(6) :857–885, sep 2007.

[DR14] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4) :211–407, 2014.

[DRF23] Jonas Dann, Daniel Ritter, and Holger Fröning. Non-relational databases on fpgas : Survey, design decisions, challenges. *ACM Comput. Surv.*, 55(11), feb 2023.

[DRW+14] Jon P. Daries, Justin Reich, Jim Waldo, Elise M. Young, Jonathan Whittinghill, Andrew Dean Ho, Daniel Thomas Seaton, and Isaac Chuang. Privacy, anonymity, and big data in the social sciences. *Commun. ACM*, 57(9) :56–63, sep 2014.

[dSDA+18] Cláudio Rebelo de Sá, Wouter Duivesteijn, Paulo J. Azevedo, Alípio Mário Jorge, Carlos Soares, and Arno J. Knobbe. Discovering a taste for the unusual : exceptional models for preference mining. *Mach. Learn.*, 107(11) :1775–1807, 2018.

[DSL+22] Xiang Deng, Huan Sun, Alyssa Lees, You Wu, and Cong Yu. Turl : Table understanding through representation learning. *ACM SIGMOD Record*, 51(1) :33–40, 2022.

[DSSK19] Sabyasachi Dash, Sushil Shakyawar, Mohit Sharma, and Sandeep Kaushik. Big data in healthcare : management, analysis and future prospects. *Journal of Big Data*, 6, 06 2019.

[DTG+19] Vinicius Dias, Carlos HC Teixeira, Dorgival Guedes, Wagner Meira, and Srinivasan Parthasarathy. Fractal : A general-purpose graph pattern mining system. In *Proceedings of the 2019 International Conference on Management of Data*, pages 1357–1374, 2019.

[DURG+18] Diego De Uña, Nataliia Rümmele, Graeme Gange, Peter Schachte, and Peter J Stuckey. Machine learning and constraint programming for relational-to-ontology schema mapping. In *International Joint Conference on Artificial Intelligence 2018*, pages 1277–1283. Association for the Advancement of Artificial Intelligence (AAAI), 2018.

[DZLZ24] Haowen Dong, Chao Zhang, Guoliang Li, and Huanchen Zhang. Cloud-native databases : A survey. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–20, 2024.

[EGG+21] Rebecca Eichler, Corinna Giebler, Christoph Gröger, Holger Schwarz, and Bernhard Mitschang. Modeling metadata in data lakes—a generic model. *Data Knowl. Eng.*, 136(C), nov 2021.

[EM17] Ahmed Eldawy and Mohamed F. Mokbel. The era of big spatial data. *Proc. VLDB Endow.*, 10(12) :1992–1995, aug 2017.

[Fan15] Wenfei Fan. Data quality : From theory to practice. *SIGMOD Rec.*, 44(3) :7–18, dec 2015.

[FB23] Saeed Fathollahzadeh and Matthias Boehm. Gio : Generating efficient matrix and frame readers for custom data formats by example. *Proc. ACM Manag. Data*, 1(2), jun 2023.

[FGJK08] Wenfei Fan, Floris Geerts, Xibei Jia, and Anastasios Kementsietsidis. Conditional functional dependencies for capturing data inconsistencies. *ACM Trans. Database Syst.*, 33(2), jun 2008.

[FL15] Ziqiang Feng and Eric Lo. Accelerating aggregation using intra-cycle parallelism. In *2015 IEEE 31st International Conference on Data Engineering*, pages 291–302. IEEE, 2015.

[FS23] Yannis Foufoulas and Alkis Simitsis. Efficient execution of user-defined functions in sql queries. *Proceedings of the VLDB Endowment*, 16(12) :3874–3877, 2023.

[FTM+21] Anna Fariha, Ashish Tiwari, Alexandra Meliou, Arjun Radhakrishna, and Sumit Gulwani. Coco : Interactive exploration of conformance constraints for data understanding and data cleaning. In *Proceedings of the 2021 International Conference on Management of Data*, SIGMOD '21, page 2706–2710, New York, NY, USA, 2021. Association for Computing Machinery.

[FTR+21a] Anna Fariha, Ashish Tiwari, Arjun Radhakrishna, Sumit Gulwani, and Alexandra Meliou. Conformance constraint discovery : Measuring trust in data-driven systems. In *Proceedings of the 2021 International Conference on Management of Data*, SIGMOD '21, page 499–512, New York, NY, USA, 2021. Association for Computing Machinery.

[FTR+21b] Anna Fariha, Ashish Tiwari, Arjun Radhakrishna, Sumit Gulwani, and Alexandra Meliou. Conformance constraint discovery : Measuring trust in data-driven systems. In *Proceedings of the 2021 International Conference on Management of Data*, SIGMOD '21, page 499–512, New York, NY, USA, 2021. Association for Computing Machinery.

[FTWY21] Wenfei Fan, Chao Tian, Yanghao Wang, and Qiang Yin. Parallel discrepancy detection and incremental detection. *Proc. VLDB Endow.*, 14(8) :1351–1364, apr 2021.

[FWCY10] Benjamin C. M. Fung, Ke Wang, Rui Chen, and Philip S. Yu. Privacy-preserving data publishing : A survey of recent developments. *ACM Comput. Surv.*, 42(4), jun 2010.

[GCMG13] Antonio Gomariz, Manuel Campos, Roque Marín, and Bart Goethals. Clasp : An efficient algorithm for mining frequent closed sequences. In Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu, editors, *Advances in Knowledge Discovery and Data Mining, 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Proceedings, Part I*, volume 7818 of *Lecture Notes in Computer Science*, pages 50–61. Springer, 2013.

[GGM20] François Goasdoué, Pawel Guzewicz, and Ioana Manolescu. Rdf graph summarization for first-sight structure discovery. *The VLDB Journal*, 29 :1191 – 1218, 2020.

[GHMT17] Behzad Golshan, Alon Halevy, George Mihaila, and Wang-Chiew Tan. Data integration : After the teenage years. In *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, PODS '17, page 101–106, New York, NY, USA, 2017. Association for Computing Machinery.

[GLN+24] Philipp M. Grulich, Aljoscha P. Lepping, Dwi P. A. Nugroho, Varun Pandey, Bonaventura Del Monte, Steffen Zeuch, and Volker Markl. Query compilation without regrets. *Proc. ACM Manag. Data*, 2(3), may 2024.

[GLNS+05] Jim Gray, David T. Liu, Maria Nieto-Santisteban, Alex Szalay, David J. DeWitt, and Gerd Heber. Scientific data management in the coming decade. *SIGMOD Rec.*, 34(4) :34–41, dec 2005.

[GLZ24] Ghadeer O. Ghosheh, Jin Li, and Tingting Zhu. A survey of generative adversarial networks for synthesizing structured electronic health records. *ACM Comput. Surv.*, 56(6), jan 2024.

[GPB+18] Yolanda Gil, Suzanne A. Pierce, Hassan Babaie, Arindam Banerjee, Kirk Borne, Gary Bust, Michelle Cheatham, Imme Ebert-Uphoff, Carla Gomes, Mary Hill, John Horel, Leslie Hsu, Jim Kinter, Craig Knoblock, David Krum, Vipin Kumar, Pierre Lermusiaux, Yan Liu, Chris North, Victor Pankratius, Shanan Peters, Beth Plale, Allen Pope, Sai Ravela, Juan Restrepo, Aaron Ridley, Hanan

Samet, Shashi Shekhar, Katie Skinner, Padhraic Smyth, Basil Tikoff, Lynn Yarmey, and Jia Zhang. Intelligent systems for geosciences : an essential research agenda. *Commun. ACM*, 62(1) :76–84, dec 2018.

[Gra08] Goetz Graefe. Database servers tailored to improve energy efficiency. In *Proceedings of the 2008 EDBT workshop on Software engineering for tailor-made data management*, pages 24–28, 2008.

[Gro20] Martin Grohe. word2vec, node2vec, graph2vec, x2vec : Towards a theory of vector embeddings of structured data. In *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pages 1–16, 2020.

[GS18] Arnaud Giacometti and Arnaud Soulet. Dense neighborhood pattern sampling in numerical data. In Martin Ester and Dino Pedreschi, editors, *Proceedings of the 2018 SIAM International Conference on Data Mining, SDM 2018, May 3-5, 2018, San Diego Marriott Mission Valley, San Diego, CA, USA*, pages 756–764. SIAM, 2018.

[GSeOC+21] Yohan Boneschi Gumiel, Lucas Emanuel Silva e Oliveira, Vincent Claveau, Natalia Grabar, Emerson Cabrera Paraiso, Claudia Moro, and Deborah Ribeiro Carvalho. Temporal relation extraction in clinical texts : A systematic review. *ACM Comput. Surv.*, 54(7), sep 2021.

[GSR+17] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1263–1272. JMLR.org, 2017.

[GYY+22] Binglei Guo, Jiong Yu, Dexian Yang, Hongyong Leng, and Bin Liao. Energy-efficient database systems : A systematic survey. *ACM Comput. Surv.*, 55(6), dec 2022.

[GZ04] Bart Goethals and Mohammed Javeed Zaki. Advances in frequent itemset mining implementations : report on fimi'03. *SIGKDD Explor.*, 6(1) :109–117, 2004.

[HB21] Benjamin Hilprecht and Carsten Binnig. Restore - neural data completion for relational databases. In *Proceedings of the 2021 International Conference on Management of Data*, SIGMOD '21, page 710–722, New York, NY, USA, 2021. Association for Computing Machinery.

[HBC+21] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. Knowledge graphs. *ACM Computing Surveys (Csur)*, 54(4) :1–37, 2021.

[HCL20] Xu Han, Xiaohui Chen, and Li-Ping Liu. Gan ensemble for anomaly detection, 2020.

[HEA+22] Mikel Hernandez, Gorka Epelde, Ane Alberdi, Rodrigo Cilla, and Debbie Rankin. Synthetic data generation for tabular health records : A systematic review. *Neurocomputing*, 493 :28–45, 2022.

[HGD20] Denis Hirn, Torsten Grust, and Christian Duta. Compiling pl/sql away. In *Proceedings of the 10th Conference on Innovative Data Systems Research (CIDR 2020)*, 2020.

[HKN+16] Alon Halevy, Flip Korn, Natalya F Noy, Christopher Olston, Neoklis Polyzotis, Sudip Roy, and Steven Euijong Whang. Goods : Organizing google's datasets. In *Proceedings of the 2016 International Conference on Management of Data*, pages 795–806, 2016.

[HNB+22] Dong He, Supun C Nakandala, Dalitso Banda, Rathijit Sen, Karla Saur, Kwanghyun Park, Carlo Curino, Jesús Camacho-Rodríguez, Konstantinos Karanasos, and Matteo Interlandi. Query processing on tensor computation runtimes. *Proc. VLDB Endow.*, 15(11) :2811–2825, jul 2022.

[HSG+17] Joseph M. Hellerstein, Vikram Sreekanti, Joseph E. Gonzalez, James Dalton, Akon Dey, Sreyashi Nag, Krishna Ramachandran, Sudhanshu Arora, Arka Bhattacharyya, Shirshanka Das, Mark Donsky, Gabriel Fierro, Chang She, Carl Steinbach, Venkat Ram Subramanian, and Eric Sun. Ground : A data context service. In *Conference on Innovative Data Systems Research*, 2017.

[HTTG09] Tony Hey, Stewart Tansley, Kristin Tolle, and Jim Gray. *The Fourth Paradigm : Data-Intensive Scientific Discovery.* Microsoft Research, October 2009.

[HZB+24] Runzhou Han, Mai Zheng, Suren Byna, Houjun Tang, Bin Dong, Dong Dai, Yong Chen, Dongkyun Kim, Joseph Hassoun, and David Thorsley. Prov-io+ : A cross-platform provenance framework for scientific data on hpc systems. *IEEE Transactions on Parallel and Distributed Systems*, 35(5) :844–861, 2024.

[HZM+23] Md Imran Hossain, Ghada Zamzmi, Peter R. Mouton, Md Sirajus Salekin, Yu Sun, and Dmitry Goldgof. Explainable AI for medical data : Current methods, limitations, and future directions. *ACM Comput. Surv.*, dec 2023. Just Accepted.

[HZW+23] Wei Han, Xiaohan Zhang, Yi Wang, Lizhe Wang, Xiaohui Huang, Jun Li, Sheng Wang, Weitao Chen, Xianju Li, Ruyi Feng, Runyu Fan, Xinyu Zhang, and Yuewei Wang. A survey of machine learning and deep learning in remote sensing of geological environment : Challenges, advances, and opportunities. *ISPRS Journal of Photogrammetry and Remote Sensing*, 202 :87–113, 2023.

[JEP+21] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873) :583–589, 2021.

[JHT+23] Praveen Joshi, Mohammed Hasanuzzaman, Chandra Thapa, Haithem Afli, and Ted Scully. Enabling all in-edge deep learning : A literature review. *IEEE Access*, 11 :3431–3460, 2023.

[JKA23] Wenqi Jiang, Dario Korolija, and Gustavo Alonso. Data processing with fpgas on modern architectures. In *Companion of the 2023 International Conference on Management of Data*, SIGMOD '23, page 77–82, New York, NY, USA, 2023. Association for Computing Machinery.

[JPC+21] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and S Yu Philip. A survey on knowledge graphs : Representation, acquisition, and applications. *IEEE transactions on neural networks and learning systems*, 33(2) :494–514, 2021.

[JPS22] Madhura Joshi, Ankit Pal, and Malaikannan Sankarasubbu. Federated learning for healthcare domain - pipeline, applications and challenges. *ACM Trans. Comput. Healthcare*, 3(4), nov 2022.

[JTW+23] Sijia Jiang, Zijing Tan, Jiawei Wang, Zhikang Wang, and Shuai Ma. Guided conditional functional dependency discovery. *Inf. Syst.*, 114(C), mar 2023.

[KB10] Vijay Khatri and Carol V. Brown. Designing data governance. *Commun. ACM*, 53(1) :148–152, jan 2010.

[KHA+23] Samuel Kounev, Nikolas Herbst, Cristina L. Abad, Alexandru Iosup, Ian Foster, Prashant Shenoy, Omer Rana, and Andrew A. Chien. Serverless computing : What it is, and what it is not ? *Commun. ACM*, 66(9) :80–92, aug 2023.

[Kir22] Keith Kirkpatrick. Artificial intelligence and mental health. *Commun. ACM*, 65(5) :32–34, apr 2022.

[KL24] Christoph Koch and Peter Lindner. Query optimization by quantifier elimination. *Proc. ACM Manag. Data*, 2(2), may 2024.

[Kle23] Stefan Klessinger. Capturing data-inherent dependencies in json schema extraction. In *Companion of the 2023 International Conference on Management of Data*, SIGMOD '23, page 295–297, New York, NY, USA, 2023. Association for Computing Machinery.

[KLF+24] Moe Kayali, Anton Lykov, Ilias Fountalis, Nikolaos Vasiloglou, Dan Olteanu, and Dan Suciu. Chorus : Foundation models for unified data discovery and exploration. *Proc. VLDB Endow.*, 17(8) :2104–2114, may 2024.

[KMKT+21] Kenza Kellou-Menouer, Nikolaos Kardoulakis, Georgia Troullinou, Zoubida Kedad, Dimitris Plexousakis, and Haridimos Kondylakis. A survey on semantic schema discovery. *The VLDB Journal*, 31(4) :675–710, nov 2021.

[KSGM22] Aamod Khatiwada, Roee Shraga, Wolfgang Gatterbauer, and Renée J. Miller. Integrating data lake tables. *Proc. VLDB Endow.*, 16(4) :932–945, dec 2022.

[KSR21] Haridimos Kondylakis, Kostas Stefanidis, and Praveen Rao. Report on the third international workshop on semantic web meets health data management (swh 2020). *SIGMOD Rec.*, 50(3) :32–35, dec 2021.

[KTT23] Devidas T. Kushnure, Shweta Tyagi, and Sanjay N. Talbar. Lim-net : Lightweight multi-level multiscale network with deep residual learning for automatic liver segmentation in CT images. *Biomed. Signal Process. Control.*, 80(Part) :104305, 2023.

[KYTM24] Shulei Kuang, Honghui Yang, Zijing Tan, and Shuai Ma. Efficient differential dependency discovery. *Proc. VLDB Endow.*, 17(7) :1552–1564, may 2024.

[LDF+24] Zifan Liu, Shaleen Deep, Anna Fariha, Fotis Psallidas, Ashish Tiwari, and Avrilia Floratou. Rapidash : Efficient detection of constraint violations. *Proc. VLDB Endow.*, 17(8) :2009–2021, may 2024.

[LDZ22] Guoliang Li, Haowen Dong, and Chao Zhang. Cloud databases : new techniques, challenges, and opportunities. *Proc. VLDB Endow.*, 15(12) :3758–3761, aug 2022.

[Len02] Maurizio Lenzerini. Data integration : a theoretical perspective. In *Proceedings of the Twenty-First ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '02, page 233–246, New York, NY, USA, 2002. Association for Computing Machinery.

[LH19] Jiaheng Lu and Irena Holubová. Multi-model databases : A new journey to handle the variety of data. *ACM Comput. Surv.*, 52(3), jun 2019.

[LLL20] Chih-Te Lai, Cheng-Te Li, and Shou-De Lin. Deep energy factorization model for demographic prediction. *ACM Trans. Intell. Syst. Technol.*, 12(1), nov 2020.

[LMGPRM23] Antonio López Martínez, Manuel Gil Pérez, and Antonio Ruiz-Martínez. A comprehensive review of the state-of-the-art on security and privacy issues in healthcare. *ACM Comput. Surv.*, 55(12), mar 2023.

[LRB+21] Peng Li, Xi Rao, Jennifer Blase, Yue Zhang, Xu Chu, and Ce Zhang. Cleanml : A study for evaluating the impact of data cleaning on ML classification tasks. In *37th IEEE International Conference on Data Engineering, ICDE 2021, Chania, Greece, April 19-22, 2021*, pages 13–24. IEEE, 2021.

[LSN24] Junfei Liu, Shaotong Sun, and Fatemeh Nargesian. Causal dataset discovery with large language models. In *Proceedings of the 2024 Workshop on Human-*

In-the-Loop Data Analytics, HILDA 24, page 1–8, New York, NY, USA, 2024. Association for Computing Machinery.

[LSS24] Claude Lehmann, Pavel Sulimov, and Kurt Stockinger. Is your learned query optimizer behaving as you expect ? a machine learning perspective. *Proc. VLDB Endow.*, 17(7) :1565–1577, may 2024.

[LTF+14] Zheng Jye Ling, Quoc Trung Tran, Ju Fan, Gerald C. H. Koh, Thi Nguyen, Chuen Seng Tan, James W. L. Yip, and Meihui Zhang. Gemini : an integrative healthcare analytics system. *Proc. VLDB Endow.*, 7(13) :1766–1771, aug 2014.

[LZW+19] Wei Lu, Zhanhao Zhao, Xiaoyu Wang, Haixiang Li, Zhenmiao Zhang, Zhiyu Shui, Sheng Ye, Anqun Pan, and Xiaoyong Du. A lightweight and efficient temporal database management system in tdsql. *Proc. VLDB Endow.*, 12(12) :2035–2046, aug 2019.

[Ma21] Xiaogang Ma. Data science for geoscience : Recent progress and future trends from the perspective of a data life cycle. 2021.

[MBC13] Paolo Missier, Khalid Belhajjame, and James Cheney. The w3c prov family of specifications for modelling provenance metadata. In *Proceedings of the 16th International Conference on Extending Database Technology*, EDBT '13, page 773–776, New York, NY, USA, 2013. Association for Computing Machinery.

[MDGS24] Maryam Mozaffari, Anton Dignös, Johann Gamper, and Uta Störl. Self-tuning database systems : A systematic literature review of automatic database schema design and tuning. *ACM Comput. Surv.*, 56(11), jun 2024.

[MDT+24] Dylan Molho, Jiayuan Ding, Wenzhuo Tang, Zhaoheng Li, Hongzhi Wen, Yixin Wang, Julian Venegas, Wei Jin, Renming Liu, Runze Su, Patrick Danaher, Robert Yang, Yu Leo Lei, Yuying Xie, and Jiliang Tang. Deep learning in single-cell analysis. *ACM Trans. Intell. Syst. Technol.*, 15(3), mar 2024.

[MFCP24] Luís Manuel Meruje Ferreira, Fabio Coelho, and José Pereira. Databases in edge and fog environments : A survey. *ACM Comput. Surv.*, 56(11), jul 2024.

[Mil18] Renée J. Miller. Open data integration. *Proc. VLDB Endow.*, 11(12) :2130–2139, aug 2018.

[MLC+15] Tyler Mccormick, Hedwig Lee, Nina Cesare, Ali Shojaie, and Emma Spiro. Using twitter for demographic and social science research : Tools for data collection and processing. *Sociological Methods Research*, 09 2015.

[MNM+21] Ryan Marcus, Parimarjan Negi, Hongzi Mao, Nesime Tatbul, Mohammad Alizadeh, and Tim Kraska. Bao : Making learned query optimization practical. In *Proceedings of the 2021 International Conference on Management of Data*, SIGMOD '21, page 1275–1288, New York, NY, USA, 2021. Association for Computing Machinery.

[Mon22] Don Monroe. Neurosymbolic ai. *Commun. ACM*, 65(10) :11–13, sep 2022.

[MPMF23] Giosué Cataldo Marinó, Alessandro Petrini, Dario Malchiodi, and Marco Frasca. Deep neural networks compression : A comparative survey and choice recommendations. *Neurocomputing*, 520 :152–170, 2023.

[MRT+21] Abhishek Modi, Kaushik Rajan, Srinivas Thimmaiah, Prakhar Jain, Swinky Mann, Ayushi Agarwal, Ajith Shetty, Shahid K I, Ashit Gosalia, and Partho Sarthi. New query optimization techniques in the spark engine of azure synapse. *Proc. VLDB Endow.*, 15(4) :936–948, dec 2021.

[MRZ21] Imen Megdiche, Franck Ravat, and Yan Zhao. Metadata management on data processing in data lakes. In *SOFSEM 2021 : Theory and Practice of Computer Science : 47th International Conference on Current Trends in Theory and Practice of Computer Science, SOFSEM 2021, Bolzano-Bozen, Italy, January*

25–29, 2021, Proceedings, page 553–562, Berlin, Heidelberg, 2021. Springer-Verlag.

[MRZ+23] Xiaogang Ma, Jolyon P. Ralph, Jiyin Zhang, Xiang Que, Anirudh Prabhu, Shaunna M. Morrison, Robert M. Hazen, Lesley Wyborn, and Kerstin A. Lehnert. Openmindat : Open and fair mineralogy data from the mindat database. *Geoscience Data Journal*, 11 :104 – 94, 2023.

[MSR+17] Carlo Meghini, Roberto Scopigno, Julian Richards, Holly Wright, Guntram Geser, Sebastian Cuy, Johan Fihn, Bruno Fanini, Hella Hollander, Franco Niccolucci, Achille Felicetti, Paola Ronzino, Federico Nurra, Christos Papatheodorou, Dimitris Gavrilis, Maria Theodoridou, Martin Doerr, Douglas Tudhope, Ceri Binding, and Andreas Vlachidis. Ariadne : A research infrastructure for archaeology. *J. Comput. Cult. Herit.*, 10(3), aug 2017.

[MTA09] Rene Mueller, Jens Teubner, and Gustavo Alonso. Data processing on fpgas. *Proc. VLDB Endow.*, 2(1) :910–921, aug 2009.

[MTF+23] Johns M, Meurers T, Wirth FN, Haber AC, Müller A, Halilovic M, Balzer F, and Prasser F. Data provenance in biomedical research : Scoping review. *J Med Internet Res*, 2023.

[MTVF24] Stephanie Milani, Nicholay Topin, Manuela Veloso, and Fei Fang. Explainable reinforcement learning : A survey and comparative review. *ACM Comput. Surv.*, 56(7), apr 2024.

[MW23] Zhengjie Miao and Jin Wang. Watchog : A light-weight contrastive learning based framework for column annotation. *Proc. ACM Manag. Data*, 1(4), dec 2023.

[MWZY22] Rui Min, Zhi Wang, Yingping Zhuang, and Xiaoping Yi. Application of semi-supervised convolutional neural network regression model based on data augmentation and process spectral labeling in raman predictive modeling of cell culture processes. *Biochemical Engineering Journal*, 191 :108774, 12 2022.

[NAJ23] Fatemeh Nargesian, Abolfazl Asudeh, and H. V. Jagadish. Next-generation challenges of responsible data integration. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23, page 1256–1259, New York, NY, USA, 2023. Association for Computing Machinery.

[Nau17] Jeffrey F. Naughton. Technical perspective : Broadening and deepening query optimization yet still making progress. *Commun. ACM*, 60(10) :80, sep 2017.

[NCOR22] Avanika Narayan, Ines Chami, Laurel J. Orr, and Christopher Ré. Can foundation models wrangle your data ? *Proc. VLDB Endow.*, 16(4) :738–746, 2022.

[NDGN13] Benjamin Négrevergne, Anton Dries, Tias Guns, and Siegfried Nijssen. Dominance programming for itemset mining. In Hui Xiong, George Karypis, Bhavani Thuraisingham, Diane J. Cook, and Xindong Wu, editors, *2013 IEEE 13th International Conference on Data Mining, Dallas, TX, USA, December 7-10, 2013*, pages 557–566. IEEE Computer Society, 2013.

[Neu24] Thomas Neumann. Closing the gap between theory and practice in query optimization. In *Companion of the 43rd Symposium on Principles of Database Systems*, PODS '24, page 4, New York, NY, USA, 2024. Association for Computing Machinery.

[NF15] Tom Narock and Peter Fox. The semantic web in earth and space science. current status and future directions. 2015.

[NG20] Thomas Rincy N and Roopam Gupta. A survey on machine learning approaches and its techniques :. In *2020 IEEE International Students' Conference on Electrical,Electronics and Computer Science (SCEECS)*, pages 1–6, 2020.

[NJW+07] James H. Nettles, Jeremy L. Jenkins, Chris Williams, Alex M. Clark, Andreas Bender, Zhan Deng, John W. Davies, and Meir Glick. Flexible 3d pharmacophores as descriptors of dynamic biological space. *Journal of Molecular Graphics and Modelling*, 26(3) :622–633, 2007. Graham Richards 67th Birthday Honour issue.

[NKN23] Nafiseh Ghaffar Nia, Erkan Kaplanoglu, and Ahad Nasab. Evaluation of artificial intelligence techniques in disease diagnosis and prediction. *Discov. Artif. Intell.*, 3(1), 2023.

[NZPM18] Fatemeh Nargesian, Erkang Zhu, Ken Q Pu, and Renée J Miller. Table union search on open data. *Proceedings of the VLDB Endowment*, 11(7) :813–825, 2018.

[Özsu23] M. Tamer Özsu. Data science–a systematic treatment. *Commun. ACM*, 66(7) :106–116, jun 2023.

[PA16] Andrew Pavlo and Matthew Aslett. What's really new with newsql ? *ACM Sigmod Record*, 45(2) :45–55, 2016.

[PAS+23] Fotis Psallidas, Ashvin Agrawal, Chandru Sugunan, Khaled Ibrahim, Konstantinos Karanasos, Jesús Camacho-Rodríguez, Avrilia Floratou, Carlo Curino, and Raghu Ramakrishnan. Oneprovenance : Efficient extraction of dynamic coarse-grained provenance from database query event logs. *Proc. VLDB Endow.*, 16(12) :3662–3675, aug 2023.

[PAVB23] Dragana Paparova, Margunn Aanestad, Polyxeni Vassilakopoulou, and Marianne Klungland Bahus. Data governance spaces : The case of a national digital service for personal health data. *Information and Organization*, 33(1) :100451, 2023.

[PCAB+24] Jon Perez-Cerrolaza, Jaume Abella, Markus Borg, Carlo Donzella, Jesús Cerquides, Francisco J. Cazorla, Cristofer Englund, Markus Tauber, George Nikolakopoulos, and Jose Luis Flores. Artificial intelligence for safety-critical systems in industrial and transportation domains : A survey. *ACM Comput. Surv.*, 56(7), apr 2024.

[PCW23] Norman W. Paton, Jiaoyan Chen, and Zhenyu Wu. Dataset discovery and exploration : A survey. *ACM Comput. Surv.*, 56(4), nov 2023.

[PdAN21] Eduardo H. M. Pena, Eduardo C. de Almeida, and Felix Naumann. Fast detection of denial constraint violations. *Proc. VLDB Endow.*, 15(4) :859–871, dec 2021.

[PN24] Massimo Perini and Milos Nikolic. In-database data imputation. *Proc. ACM Manag. Data*, 2(1), mar 2024.

[pro23] *Big Data Analytics in Astronomy, Science, and Engineering : 11th International Conference on Big Data Analytics, BDA 2023, Aizu, Japan, December 5–7, 2023, Proceedings*, Berlin, Heidelberg, 2023. Springer-Verlag.

[PSCH21] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection : A review. *ACM Comput. Surv.*, 54(2), mar 2021.

[PST+22] Jinfeng Peng, Derong Shen, Nan Tang, Tieying Liu, Yue Kou, Tiezheng Nie, Hang Cui, and Ge Yu. Self-supervised and interpretable data cleaning with sequence generative adversarial networks. *Proc. VLDB Endow.*, 16(3) :433–446, nov 2022.

[PSY+18] Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and S. S. Iyengar. A survey

on deep learning : Algorithms, techniques, and applications. *ACM Comput. Surv.*, 51(5), sep 2018.

[PUS22] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann, editors, *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR, 07–08 Apr 2022.

[PWL24a] James Jie Pan, Jianguo Wang, and Guoliang Li. Survey of vector database management systems. *The VLDB Journal*, pages 1–25, 2024.

[PWL24b] James Jie Pan, Jianguo Wang, and Guoliang Li. Vector database management techniques and systems. In *Companion of the 2024 International Conference on Management of Data*, SIGMOD/PODS '24, page 597–604, New York, NY, USA, 2024. Association for Computing Machinery.

[PXNO23] Ciyuan Peng, Feng Xia, Mehdi Naseriparsa, and Francesco Osborne. Knowledge graphs : Opportunities and challenges. *Artificial Intelligence Review*, 56(11) :13071–13102, 2023.

[QLT+23] Chaoqin Qian, Menglu Li, Zijing Tan, Ai Ran, and Shuai Ma. Incremental discovery of denial constraints. *The VLDB Journal*, 32(6) :1289–1313, mar 2023.

[QTO+20] Abdulhakim Qahtan, Nan Tang, Mourad Ouzzani, Yang Cao, and Michael Stonebraker. Pattern functional dependencies for data cleaning. *Proc. VLDB Endow.*, 13(5) :684–697, jan 2020.

[RBM22] Viktor Rosenfeld, Sebastian Breß, and Volker Markl. Query processing on heterogeneous cpu/gpu systems. *ACM Comput. Surv.*, 55(1), jan 2022.

[RG23] Maria Rigaki and Sebastian Garcia. A survey of privacy attacks in machine learning. *ACM Comput. Surv.*, 56(4), nov 2023.

[RKIW18] Eric D. Ragan, Hye-Chung Kum, Gurudev Ilangovan, and Han Wang. Balancing privacy and information disclosure in interactive record linkage with visual masking. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–12, New York, NY, USA, 2018. Association for Computing Machinery.

[RP24] Edoardo Ramalli and Barbara Pernici. Challenges of a data ecosystem for scientific data. *Data Knowl. Eng.*, 148(C), feb 2024.

[RRK+20] Cédric Renggli, Luka Rimanic, Luka Kolar, Wentao Wu, and Ce Zhang. Automatic feasibility study via data quality analysis for ml : A case-study on label noise. *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 218–231, 2020.

[RVF24] Andrey Rodrigues, Maria Lúcia Villela, and Eduardo Feitosa. A systematic mapping study on social network privacy : Threats and solutions. *ACM Comput. Surv.*, 56(7), apr 2024.

[SBJS21] Deepak Kumar Sharma, Saakshi Bhargava, Aashna Jha, and Pawan Singh. 7 - early detection and diagnosis using deep learning. In Valentina Emilia Balas, Brojo Kishore Mishra, and Raghvendra Kumar, editors, *Handbook of Deep Learning in Biomedical Engineering*, pages 191–217. Academic Press, 2021.

[SBO23] Azra Seyyedi, Mahdi Bohlouli, and Seyedehsan Nedaaee Oskoee. Machine learning and physics : A survey of integrated models. *ACM Comput. Surv.*, 56(5), nov 2023.

[SCAK+19] Sandeep Singh Sandha, Wellington Cabrera, Mohammed Al-Kateb, Sanjay Nair, and Mani Srivastava. In-database distributed machine learning : demonstration using teradata sql engine. *Proc. VLDB Endow.*, 12(12) :1854–1857, aug 2019.

[SCKV22] Deepak Kumar Sharma, Mayukh Chatterjee, Gurmehak Kaur, and Suchitra Vavilala. 3 - deep learning applications for disease diagnosis. In Deepak Gupta, Utku Kose, Ashish Khanna, and Valentina Emilia Balas, editors, *Deep Learning for Medical Applications with Unique Data*, pages 31–51. Academic Press, 2022.

[SCM18] Alisa Smirnova and Philippe Cudré-Mauroux. Relation extraction using distant supervision : A survey. *ACM Comput. Surv.*, 51(5), nov 2018.

[SD21] Pegdwendé Sawadogo and Jérôme Darmont. On data lake architectures and metadata management. *Journal of Intelligent Information Systems*, 56(1) :97–120, 2021.

[SDSE20] Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. Green ai. *Commun. ACM*, 63(12) :54–63, nov 2020.

[SGR20] Roee Shraga, Avigdor Gal, and Haggai Roitman. Adnev : cross-domain schema matching using deep similarity matrix adjustment and evaluation. *Proc. VLDB Endow.*, 13(9) :1401–1415, may 2020.

[Sir17] Utku Sirin. Energy-efficient database machines. In *Proceedings of the 2017 ACM International Conference on Management of Data*, SIGMOD '17, page 46–48, New York, NY, USA, 2017. Association for Computing Machinery.

[SML+23] Pranav Subramaniam, Yintong Ma, Chi Li, Ipsita Mohanty, and Raul Castro Fernandez. Comprehensive and comprehensible data catalogs : The what, who, where, when, why, and how of metadata management, 2023.

[SMPJ+23] Márcio Sembay, Douglas Macedo, Laercio Pioli Junior, Regina Braga, and Antonio Cabezuelo. Provenance data management in health information systems : A systematic literature review. *Journal of Personalized Medicine*, 13 :991, 06 2023.

[SSP+21] Sri Subrahmanya, Dasharathraj Shetty, Dr Patil, Bm Hameed, Rahul Paul, Komal Smriti, Nithesh Naik, and Bhaskar Somani. The role of data science in healthcare advancements : applications, benefits, and future prospects. *Irish Journal of Medical Science (1971 -)*, 191, 08 2021.

[SSZW24] Yongye Su, Yinqi Sun, Minjia Zhang, and Jianguo Wang. Vexless : A serverless vector data management system using cloud functions. *Proc. ACM Manag. Data*, 2(3), may 2024.

[STG+23] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Towards expert-level medical question answering with large language models, 2023.

[Sto20] Victoria Stodden. The data science life cycle : a disciplined approach to advancing data science as a science. *Commun. ACM*, 63(7) :58–66, jun 2020.

[SYYZ22] Yuanming Shi, Kai Yang, Zhanpeng Yang, and Yong Zhou. Chapter five - model compression for on-device inference. In Yuanming Shi, Kai Yang, Zhanpeng Yang, and Yong Zhou, editors, *Mobile Edge Artificial Intelligence*, pages 71–82. Academic Press, 2022.

[SZN+20] Ying Song, Shuangjia Zheng, Zhangming Niu, Zhang-Hua Fu, Yutong Lu, and Yuedong Yang. Communicative representation learning on attributed molecular graphs. In *IJCAI*, volume 2020, pages 2831–2838, 2020.

[TL16] Etienne Taffoureau and Christelle Loiselet. Metadata for 3d geological models : definition and implementation. In *Proceedings of the 2016 International Conference on Dublin Core and Metadata Applications*, DCMI'16, page 11. Dublin Core Metadata Initiative, 2016.

[TTH+23] Tanja Tornede, Alexander Tornede, Jonas Hanselle, Felix Mohr, Marcel Wever, and Eyke Hüllermeier. Towards green automated machine learning : Status quo and future directions. *J. Artif. Int. Res.*, 77, jun 2023.

[vL14] Matthijs van Leeuwen. Interactive data exploration using pattern mining. In Andreas Holzinger and Igor Jurisica, editors, *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics - State-of-the-Art and Future Challenges*, volume 8401 of *Lecture Notes in Computer Science*, pages 169–182. Springer, 2014.

[VT14] Jilles Vreeken and Nikolaj Tatti. Interesting patterns. In Charu C. Aggarwal and Jiawei Han, editors, *Frequent Pattern Mining*, pages 105–134. Springer, 2014.

[WDW+23] Haiqin Wu, Boris Düdder, Liangmin Wang, Zhenfu Cao, Jun Zhou, and Xia Feng. Survey on secure keyword search over outsourced data : From cloud to blockchain-assisted architecture. *ACM Comput. Surv.*, 56(3), oct 2023.

[WFWW20] Weiyuan Wu, Lampros Flokas, Eugene Wu, and Jiannan Wang. Complaint-driven training data debugging for query 2.0. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, SIGMOD '20, page 1317–1334, New York, NY, USA, 2020. Association for Computing Machinery.

[WLI+24] John Wamburu, Stephen Lee, Srinivasan Iyengar, David Irwin, and Prashant Shenoy. Analyzing the energy usage of a community and the benefits of energy storage. *ACM J. Comput. Sustain. Soc.*, 2(2), may 2024.

[WLS+23] Tian Wang, Yuzhu Liang, Xuewei Shen, Xi Zheng, Adnan Mahmood, and Quan Z. Sheng. Edge computing and sensor-cloud : Overview, solutions, and directions. *ACM Comput. Surv.*, 55(13s), jul 2023.

[WLW22] Runhui Wang, Yuliang Li, and Jin Wang. Sudowoodo : Contrastive self-supervised learning for multi-purpose data integration and preparation. *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 1502–1515, 2022.

[WLY+24] Nan Wu, Yingjie Li, Hang Yang, Hanqiu Chen, Steve Dai, Cong Hao, Cunxi Yu, and Yuan Xie. Survey of machine learning for software-assisted hardware design verification : Past, present, and prospect. *ACM Trans. Des. Autom. Electron. Syst.*, 29(4), jun 2024.

[WMWG17] Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. Knowledge graph embedding : A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12) :2724–2743, 2017.

[WT24a] Ziyun Wei and Immanuel Trummer. Rome : Robust query optimization via parallel multi-plan execution. *Proc. ACM Manag. Data*, 2(3), may 2024.

[WT24b] Sebastian Werner and Stefan Tai. A reference architecture for serverless big data processing. *Future Generation Computer Systems*, 155 :179–192, 2024.

[WTL+24] Chengbin Wang, Liangquan Tan, Yuanjun Li, Mingguo Wang, Xiaogang Ma, and Jianguo Chen. Ontology-driven relational data mapping for construc-

ting a knowledge graph of porphyry copper deposits. *Earth Sci. Informatics*, 17(3) :2649–2660, 2024.

[WW22] Yifan Wang and Daisy Zhe Wang. Extensible database simulator for fast prototyping in-database algorithms. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 5029–5033, New York, NY, USA, 2022. Association for Computing Machinery.

[WWD+23] Shiwen Wu, Qiyu Wu, Honghua Dong, Wen Hua, and Xiaofang Zhou. Blocker and matcher can mutually benefit : A co-learning framework for low-resource entity resolution. *Proc. VLDB Endow.*, 17(3) :292–304, nov 2023.

[XCC+24] Naili Xing, Shaofeng Cai, Gang Chen, Zhaojing Luo, Beng Chin Ooi, and Jian Pei. Database native model selection : Harnessing deep neural networks in database systems. *Proc. VLDB Endow.*, 17(5) :1020–1033, may 2024.

[XCP+24] Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. Large language models for generative information extraction : A survey, 2024.

[XTWM22] Renjie Xiao, Zijing Tan, Haojin Wang, and Shuai Ma. Fast approximate denial constraint discovery. *Proc. VLDB Endow.*, 16(2) :269–281, oct 2022.

[XWW+24] Xiaodan Xing, Huanjun Wu, Lichao Wang, Iain Stenson, May Yong, Javier Del Ser, Simon Walsh, and Guang Yang. Non-imaging medical data synthesis for trustworthy ai : A comprehensive survey. *ACM Comput. Surv.*, 56(7), apr 2024.

[XZL+24] Hanguang Xiao, Feizhong Zhou, Xingyue Liu, Tianqi Liu, Zhipeng Li, Xin Liu, and Xiaoxuan Huang. A comprehensive survey of large language models and multimodal large language models in medicine. *CoRR*, abs/2405.08603, 2024.

[YH02] Xifeng Yan and Jiawei Han. gspan : Graph-based substructure pattern mining. In *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM 2002), 9-12 December 2002, Maebashi City, Japan*, pages 721–724. IEEE Computer Society, 2002.

[YLW+23] Zijun Yao, Bin Liu, Fei Wang, Daby Sow, and Ying Li. Ontology-aware prescription recommendation in treatment pathways using multi-evidence healthcare data. *ACM Trans. Inf. Syst.*, 41(4), apr 2023.

[YPS09] Xiaoyan Yang, Cecilia M. Procopiuc, and Divesh Srivastava. Summarizing relational databases. *Proc. VLDB Endow.*, 2(1) :634–645, aug 2009.

[YSJ+19] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian P. Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi S Jaakkola, Klavs F. Jensen, and Regina Barzilay. Analyzing learned molecular representations for property prediction. *ChemRxiv*, 2019.

[YYL+23] Dongran Yu, Bo Yang, Dayou Liu, Hui Wang, and Shirui Pan. A survey on neural-symbolic learning systems. *Neural Netw.*, 166(C) :105–126, sep 2023.

[ZB18] Shuo Zhang and Krisztian Balog. Ad hoc table retrieval using semantic similarity. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 1553–1562, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee.

[ZCR+24] Steffen Zeuch, Ankit Chaudhary, Viktor Rosenfeld, Taha Tekdogan, Adrian Michalke, Matthis Gördel, Ariane Ziehn, and Volker Markl. Using and enhancing nebulastream - a tutorial. In *Proceedings of the 18th ACM International Conference on Distributed and Event-Based Systems*, DEBS '24, page 212–216, New York, NY, USA, 2024. Association for Computing Machinery.

[ZDNM19] Erkang Zhu, Dong Deng, Fatemeh Nargesian, and Renée J Miller. Josie : Overlap set similarity search for finding joinable tables in data lakes. In *Proceedings of the 2019 International Conference on Management of Data*, pages 847–864, 2019.

[ZF24] Chao Zhang and Toumani Farouk. Sharing queries with nonequivalent user-defined aggregate functions. *ACM Trans. Database Syst.*, 49(2), apr 2024.

[ZGQ+24] Ronghang Zhu, Dongliang Guo, Daiqing Qi, Zhixuan Chu, Xiang Yu, and Sheng Li. A survey of trustworthy representation learning across domains. *ACM Trans. Knowl. Discov. Data*, 18(7), jun 2024.

[ZHZ+21] Jinglin Zou, Debiao He, Sherali Zeadally, Neeraj Kumar, Huaqun Wang, and Kkwang Raymond Choo. Integrated blockchain and cloud computing systems : A systematic survey, solutions, and challenges. *ACM Comput. Surv.*, 54(8), oct 2021.

[ZRL+21] Yingqiang Zhang, Chaoyi Ruan, Cheng Li, Xinjun Yang, Wei Cao, Feifei Li, Bo Wang, Jing Fang, Yuhui Wang, Jingze Huo, and Chao Bi. Towards cost-effective and elastic cloud database deployment via memory disaggregation. *Proc. VLDB Endow.*, 14(10) :1900–1912, jun 2021.

[ZSS24] Sepanta Zeighami, Raghav Seshadri, and Cyrus Shahabi. A neural database for answering aggregate queries on incomplete relational data. *IEEE Transactions on Knowledge and Data Engineering*, 36(7) :2790–2802, 2024.

[ZWDZ24] Rong Zhu, Lianggui Weng, Bolin Ding, and Jingren Zhou. Learned query optimizer : What is new and what is next. In *Companion of the 2024 International Conference on Management of Data*, SIGMOD/PODS '24, page 561–569, New York, NY, USA, 2024. Association for Computing Machinery.

[ZXGZ24] Bangyi Zhao, Weixia Xu, Jihong Guan, and Shuigeng Zhou. Molecular property prediction based on graph structure learning. *Bioinformatics*, 40(5) :btae304, 05 2024.

[ZZC+23] Chenyang Zhang, Feng Zhang, Kuangyu Chen, Mingjun Chen, Bingsheng He, and Xiaoyong Du. Edgenn : Efficient neural network inference for cpu-gpu integrated edge devices. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 1193–1207, 2023.

Annexe A

Bilan 2020-2024

A.1 Listes des Actions actives pendant la période 2020-2024

- **Big Data for Astronomy** (2020-2024)
  - Acronyme : BigData4Astro
  - Responsables :
    - André SCHAAFF, CDS, INSU
    - Karine ZEITOUNI, DAVID, INS2i
    - Nicolas Lumineau, LIRIS, INS2i
  - Atelier en 2020 - Action 2021-2022 - Action 2023-2024
  - Résumé : En astronomie, comme dans d'autres domaines, les recherches s'appuient en partie sur des analyses fines de grandes masses de données et des simulations à très grande échelle présentant des exigences spécifiques. En astronomie, la réponse à certaines questions, nécessite un saut d'un ordre de grandeur dans la taille, e.g. des simulations numériques. Les environnements informatiques actuels s'appuient sur des architectures HPC confrontées à des difficultés de gestion des données massives. Notre objectif est d'organiser et de mettre en réseau une communauté de chercheurs et d'ingénieurs s'intéressant à ces problématiques, favorable à une synergie interdisciplinaire scientifique et technique, s'inscrivant dans les efforts de convergence entre HPC et Big Data pour le traitement de ce type de données.
- **Données Intelligentes : transformer l'information en connaissance** (2020-2024)
  - Acronyme : DOING
  - Responsables :
    - Mirian HALFELD FERRARI ALVES, LIFO, INS2i
    - Anne-Lyse MINARD-FORST, LLL, INSHS
    - Genoveva VARGAS-SOLAR, LIRIS, INS2i
  - Atelier en 2020 - Action 2021-2022 - Action 2023-2024
  - Résumé : L'action DOING se concentre sur la transformation des données en information puis en connaissance. L'idée est de mettre à contribution les expertises des chercheurs en TAL, BD et IA pour : (1) extraire des informations dans les données textuelles et les représenter pour peupler des bases de connaissances. Le choix des entités et des relations à extraire pourra être guidé par les requêtes et éventuellement la structure de la base,

avec l'idée d'extraire et stocker les données les plus intéressantes et importantes pour les consultations et les analyses des experts. DOING s'intéresse en particulier à la question de l'extraction de relations (pas uniquement binaire et pas uniquement inter-phrase) et à ce que l'interaction avec les bases peut apporter à cette tâche ; (2) proposer des méthodes intelligentes pour la manipulation et la maintenance de ces bases avec de nouvelles formes de requêtes englobant des mécanismes d'analyse efficaces, flexibles, sûrs, adaptés à l'utilisateur et qui respectent des contraintes de qualité et de vie privée. DOING s'intéresse en particulier aux graphes de propriétés, au langage d'interrogation style Cypher dans le cadre de Open Cypher et GQL et aux algorithmes d'analyse des graphes (centrality, community detection, similarity, link prediction, pathfinding).

- ▷ **Data Science in Chemistry** (2022-2024)
  - Acronyme : DSChem
  - Responsables :
    - Dominique Douguet, IPMC, INSB
    - Nicolas Blanchard, LIMA, INC
    - Bertrand Cuissart, GREYC, INS2i
  - Atelier en 2022 - Action 2023-2024
  - Résumé : DSChem a pour objectif de renforcer l'interface entre la chimie et l'informatique, au niveau national. En complément des moyens existants, DSChem vise à organiser un groupe de travail ouvert à tous et dont les échanges permettront d'identifier rapidement les besoins informatiques en traitement d'information chimique et les compétences à mobiliser pour y répondre. Ainsi, à court terme, DSChem favorisera l'identification d'une communauté de chercheurs autour de ce thème. A moyen terme, DSChem favorisera la création et la conduite de projets dans ce périmètre, stimulant ainsi la qualité de la production scientifique associée.
- ▷ **Human Explainable machine Learning Pipeline** (2021-2025)
  - Acronyme : HELP
  - Responsables :
    - Michael Baker, i3, INSHSi
    - Nicolas Labroche, LIFAT, INS2i
    - Julien Aligon, IRIT, INS2i
  - Atelier en 2021 (FENDER) - Action 2022-2023 - Action 2024-2025
  - Résumé : L'action vise à faire se rencontrer des chercheurs en informatique, spécialistes de la fouille de données et de manipulation de données, argumentation et traçabilité (INS2I) et des chercheurs en sciences humaines et sociales, spécialistes du discours explicatif et des mécanismes cognitifs en prise avec l'élaboration d'explications (INSHS) [Miller, 2019], en interaction avec des spécialistes industriels ayant des problématiques d'explication liée à l'exploitation automatique de leurs données à travers des pipelines de bout-en-bout allant des données brutes aux résultats d'analyse finaux. L'objectif de ces rencontres est : (i) De mieux comprendre les mécanismes sousjacents à une explication et à sa réception et de produire une typologie des problèmes d'explications en fonction des scénarios d'usage en informatique : en fonction des données et des prétraitements de données effectués dessus, de l'utilisateur et de la nature de la méthodes qui traite les données (clustering, classification,recommandation) ; (ii) De produire des algorithmes interactifs, suivant le principe UXAI (User Centric eXplainable AI), permettant d'accompagner un utilisateur dans la co-

construction d'une explication portant principalement sur les phases de prétraitement de données ; (iii) De définir des méthodes d'évaluation de la qualité d'une explication, en lien avec les utilisateur mais aussi les données sur lesquelles reposent le processus d'analyse.

- ▷ **Modélisation multi-échelle de masses de données musicales** (2022-2025)
  - Acronyme : Musiscale
  - Atelier en 2022, en 2023 - Action 2024-2025
  - Responsables :
    - Florence Levé, MIS, UPJV, CRISTAL INS2i
    - Sylvain Marchand, L3i, Univ. de La Rochelle
    - Jean-Marc Chouvel, IReMus, INSHS
  - Atelier en 2021, en 2022 - Action 2023-2024
  - Résumé : La masse de données musicales disponibles de nos jours représentent une richesse considérable d'un point de vue culturel et créatif, ainsi qu'en terme de capacité de valorisation éducative et industrielle. Le but de cet atelier est d'amorcer une réflexion collective sur des paradigmes de représentation et de traitement des données musicales visant à rendre compte de la structure de leur organisation à différentes échelles. On privilégiera les approches se rattachant à la théorie de l'information, pouvant s'appliquer aux données musicales. On espère ainsi pouvoir faire mieux converger les visions informatiques, musicologiques, cognitives et applicatives dans la façon d'appréhender ces données.
- ▷ **Simplification et Vulgarisation des Textes Scientifiques** (2021-2024)
  - Acronyme : SimpleText
  - Responsables :
    - Liana Ermakova, HCTI, Univ. Bretagne Occidentale
    - Olivier Augereau, Lab-STICC, INS2i
    - Eric Sanjuan, LIA, INS2i
  - Atelier en 2021, en 2022 - Action 2023-2024
  - Résumé : Les systèmes modernes d'accès à l'information promettent de donner aux citoyens un accès direct à des informations clés provenant de sources primaires faisant autorité. La littérature scientifique est concernée mais est en réalité difficilement accessible aux non-experts en raison de la complexité langagière, la structure, longueur, etc des documents scientifiques et de leur manque d'acculturation scientifique. La teneur du débat scientifique qui procède par confrontation d'une multiplicité d'études avant de parvenir à un consensus est aussi source de complexité. Les décisions individuelles ou politiques sont potentiellement impactées par une méconnaissance de l'ensemble des travaux et débats scientifiques. Cette difficulté de lecture de document scientifique existe également lorsque les scientifiques s'intéressent aux documents scientifiques des autres disciplines que celles dont ils sont experts. Les résultats contradictoires à l'intérieur même d'une discipline sont difficilement appréhendables par des non-spécialistes. Quid alors des résultats contradictoires entre disciplines ? La simplification de textes se donne pour objectif de réduire ces obstacles. L'atelier SimpleText abordera les opportunités et les défis des approches de simplification de textes scientifiques pour améliorer l'accès à l'information scientifique et l'acculturation scientifique. L'atelier SimpleText s'appuiera sur une communauté interdisciplinaire de chercheurs en TALN, en RI, de linguistes, de sociologues, de journalistes scientifiques et de vulgarisateurs scientifiques travaillant ensemble pour tenter de résoudre l'un des plus grands défis d'aujourd'hui.

▷ **Reasoning on Complex and Evolving Data – GDR IA** (2021-2023)

- Acronyme : ROCED
- Responsables :
  - Nathalie Hernandez, IRIT, INS2i
  - Catherine ROUSSEY, INRAE
  - Fatiha SAIS, LISN, INS2i
- Atelier en 2021 - Action 2022-2023- Pas de renouvellement demandé en 2024 mais demande d'Atelier SaF-Hn sur l'appel Ateliers 2024
- Résumé : Cette Action est inter-GDR (MaDICS et IA), avec deux thématiques centrales : la gestion de données hétérogènes de complexité et de qualité variables (sciences des données), et la représentation de connaissances et les raisonnements (intelligence artificielle). Une multitude de méthodes et de systèmes ont été développés pour répondre aux problématiques liées à l'acquisition, la publication et l'exploitation des données et des connaissances. Des avancées considérables ont été réalisées pour la construction de graphes de connaissances, l'alignement d'ontologies, le liage de données, la prédiction/l'invalidation de liens dans les graphes de connaissances, mais aussi dans le domaine de la représentation de connaissances et le raisonnement via des systèmes d'OBDA (Ontology-based Data Access). Cette Action se focalise notamment sur la complexité, l'hétérogénéité, l'incertitude et l'évolution des données et des connaissances.

▷ **Raisonner sur les données (Reasoning on Data) - GDR IA** (2017-2020)

- Acronyme : RoD
- Responsables :
  - Nathalie Hernandez, IRIT-Université de Toulouse Jean Jaurès, INS2i
  - Catherine Roussey, INRAE
  - Fatiha Sais, LRI-Université Paris-Saclay, INS2i
- Action en 2017 et 2018, renouvelée en 2019 et 2020 - Avec GDR IA
- Résumé :
  Cette action est inter-GDR (MaDICS et IA), avec deux thématiques centrales : la gestion de données hétérogènes (Sciences des Données), la représentation de connaissances et les raisonnements (Intelligence Artificielle). La question de l'exploitation de connaissances pour accéder à des données volumineuses et hétérogènes est très étudiée depuis quelques années. De nombreux travaux visent à prendre en compte des connaissances de nature ontologique (qui peuvent aller de la simple taxomomie à des connaissances décrites dans des langages logiques, comme les logiques de description, ou des langages à base de règles, organisés et étudiés en fragments d'expressivité variable) et à exploiter les inférences associées à ces connaissances dans tout le cycle de vie des données (accès, validation, enrichissement, etc.). Le problème emblématique est celui de l'interrogation (ou requêtage) des données, de nombreuses tâches complexes sur les données pouvant se reformuler en termes d'interrogation.

▷ **anaLysE et dynaMique des messages et cONversations radicales sur Internet** (2019-2022)

- Acronyme : LEMON
- Responsables :
  - Valentina Dragos, ONERA - valentina.dragos@onera.fr
  - Delphine Battistelli, MODYCO UMR (INSHS) - del.battistelli@gmail.com

— Farah Benamara, IRIT UMR (INS2I) - farah.benamara@irit.fr

- Atelier en 2018 - Action en 2019 et 2020 - Renouvellement pour 2021-2022
- Résumé : LEMON réunit des chercheurs en sociologie, linguistique, TAL et intelligence artificielle autour de plusieurs défis sociétaux, scientifiques et techniques concernant l'analyse des messages et conversations sur internet. L'action se focalise en particulier sur la propagation des contenus extrémistes ainsi que leurs connexions avec les contenus haineux. Les défis scientifiques sont nombreux et impactent différentes thématiques : pour le TAL, le défi est lié à la mise au point de méthodes capables de traiter des contenus hétérogènes (topics, structures et volumes diverses) et bruités (présence possible d'abréviations ou de textes en plusieurs langues). Pour les méthodes d'apprentissage, un défi majeur reste l'adaptation à un domaine en permanente évolution à la fois dans son contenu (exemple : topics émergents dans la rhétorique de la propagande) et sa forme. Un verrou transversal est la constitution d'un socle cohérent des connaissances soutenu par un modèle formel mettant en évidence à la fois les indices et facteurs de risque fournis par les modèles sociologiques ainsi que leur ancrage linguistique dans les données récupérées sur Internet.

▷ **MAChine LEarning for EArth observatioN** (2019-2022)

- Acronyme : MACLEAN
- Responsables :
  — Dino Ienco, TETIS, INS2i - dino.ienco@inrae.fr
  — Thomas Corpetti, LETG, (INEE) - thomas.corpetti@univ-rennes2.fr
  — Sébastien Lefèvre, IRISA, UMR (INS2i) - sebastien.lefevre@irisa.fr
- Action en 2019 et 2020 - Renouvellement pour 2021-2022
- Résumé : Avec le nombre croissant d'images et de données satellitaires disponibles associé à une augmentation des fréquences d'acquisition, l'interprétation automatique des données de télédétection pour l'observation de la Terre est un domaine très actif. Les capteurs sont aujourd'hui capables d'offrir des images à (très) haute résolution avec des fréquences d'acquisition jamais atteintes. Aussi, il est de plus en plus difficile de concevoir des méthodes capables de traiter efficacement ces données et de tirer partie des complémentarités des différentes sources de données (i.e. multi-modalités). Bien que des techniques avancées existent pour traiter chaque source de données, la combinaison de ces sources afin d'exploiter efficacement leur complémentarité pour des tâches spécifiques (i.e. estimation du rendement, occupation du sol, urbanisation, détection du parcellaire, surveillances des forêts, etc..) représente encore un verrou majeur dans la communauté de la télédétection et de l'observation de la Terre.

▷ **Maîtriser l'Analyse interactive de DOnnées pour la NArration journalistique** (2019-2022)

- Acronyme : MADONA
- Responsables :
  — Patrick Marcel, LIFAT EA 6300, INS2i - Patrick.Marcel@univ-tours.fr
  — Marie Chagnoux, CREM, INSHS - marie.chagnoux@univ-paris8.fr
- Atelier en 2018 - Action en 2019 et 2020 - Renouvellement pour 2021-2022
- Résumé : L'action MADONA s'inscrit dans la continuité d'un atelier MADICS (Atelier Cajole, CORIA/TALN 2018) qui avait fait émerger la nécessité de documenter et de modéliser les processus d'analyse des datajournalistes. Il s'agit de documenter et accompagner l'exploration et la mise en narration manuelle du journaliste, ou plus généralement d'un analyste

devant produire une narration. Cette action s'intéresse aux modèles et aux méthodes sur lesquels peut s'appuyer la construction d'un système interactif complet permettant d'accompagner la mise en narration de données. d'attacher une importance particulière aux aspects formels (forme discursive, schéma narratif) et sémantiques (intentions, argumentation) de la narration.

- ▷ **Impact Sociétal des Algorithmes Décisionnels** (2020-2021)
  - Acronyme : PLATFORM
  - Responsables :
    - Béatrice Roussillon, GAEL, UMR CNRS 5313 / UMR INRAE 1215, Université Grenoble Alpes
    - Juliette Senechal Centre Droits et perspectives du droit (CRDP, EA 4487), Université de Lille
    - Oana Goga, CNRS, Univ. Grenoble Alpes (LIG) - INS2i
  - Action en 2020 et 2021
  - Résumé : La réalisation de notre action nécessite l'établissement d'un dialogue entre experts en sciences et technologies de l'information et en sciences de la consommation. D'une part, les informaticiens devront développer des algorithmes permettant de révéler les préférences et ainsi produire des recommandations. D'autre part, les comportementalistes devront concevoir des déploiements d'expériences contrôlées selon des principes établis en psychologie cognitive et en économie comportementale. Ces dernières permettront de théoriser les choix assistés par les algorithmes décisionnels et de mesurer leur acceptabilité. Ces deux parties devront se nourrir l'une de l'autre pour au final déployer des tests à grande échelle grâce au développement de plateformes et de plug-ins. Cette collecte de données et méta-données permettra une mesure robuste de l'impact des décisions algorithmiques sur les individus et sur la société. Pour y parvenir, l'action nécessite la conception d'expériences pour l'observation de comportements de consommation en milieu contrôlé, la définition de mesures de satisfaction et d'adoption, et bien sûr le développement d'algorithmes et d'outils de collecte et d'analyse de données.
- ▷ **Apprentissage, opTimisation Large-échelle et cAlculs distribuéS** (2016-2020)
  - Acronyme : ATLAS
  - Responsables :
    - Emilie Devijver, Laboratoire d'Informatique de Grenoble, INSMI
    - Massih-Reza Amini, Laboratoire d'Informatique de Grenoble, INS2i
    - Charlotte Laclau, Laboratoire Hubert Curien, INS2i
  - Action en 2016 et 2017, atelier en 2018, redevient action en 2019 et 2020
  - Résumé : Avec l'augmentation du volume et de la complexité des données relatives aux soins, il existe un vrai défi pour la pratique clinique et la recherche médicale, où les approches conventionnelles ne peuvent pas exploiter les informations disponibles. Les hôpitaux stockent des centaines de millions de fichiers, nombre qui a augmenté de manière exponentielle au cours de la dernière décennie. Les fournisseurs de soins de santé se tournent vers les dossiers de santé électroniques, les lames de laboratoire numérisé et les images et vidéos de radiologie à haute résolution. Ajoutons à cela les pétaoctets de données stockées dans les bases de données des réclamations des compagnies d'assurance maladie et les archives de la recherche universitaire et pharmaceutique, ainsi que les milliards de données transmises en continu par des capteurs portables – suivis d'activité, dispositifs de surveillance

continue du glucose et défibrillateurs implantables. En outre, ces données sont de plus en plus complexes, par exemple hétérogènes et/ou présentant une structure de dépendance qu'il convient d'être capable de modéliser. De toute évidence, il faudra des décennies à un être humain pour analyser cette quantité de données et en extraire des informations utiles. C'est dans ce cadre que se place l'action ATLAS, en plein dans les thématiques du GdR.

A.2 Liste des Ateliers actifs dans la période 2020-2024

Nous listons ici uniquement les ateliers sont en cours et les ateliers qui n'ont pas mené à la création d'une action pendant la période 2020-2024.

- **Prévention et détection des anomalies en Agroalimentaire et dans l'Environnement**
  - Acronyme : DFAe
  - Responsables :
    - Maximilien Servajean - LIRMM - servajean@lirmm.fr
    - Lylia Abrouk - LIB Univ. Bourgogne / INRAE - lylia.abrouk@u-bourgogne.fr
    - Pascal Neuveu - INRAE Montpellier - Pascal.Neveu@inrae.fr
    - Pierre Labadie - CNRS, UMR 5805 EPOC - pierre.labadie@u-bordeaux.fr
  - Atelier en 2024
  - Résumé : L'objectif principal de l'atelier DFAe est d'explorer et de débattre des approches et des méthodologies permettant de détecter les anomalies et comportements atypiques qui impactent l'agroalimentaire, l'environnement et les ressources en eau. Cela inclut l'analyse des polluants, la surveillance de la qualité de l'eau, et la détection des pratiques non conformes dans la gestion des ressources naturelles. Un accent particulier est mis sur l'application des dernières innovations en intelligence artificielle, en analyse de données, ainsi qu'en représentation et extraction de connaissances, dans le but de développer des solutions novatrices et durables.
- **Des Sources Aux Données en Humanités Numérique**
  - Acronyme : SaD-HN
  - Responsables :
    - Nathalie Abadie LASTIG/IGN - nathalie-f.abadie@ign.fr
    - Nathalie Hernandez IRIT UMR (INS2I), nathalie.hernandez@irit.fr
    - Bertrand Duménieu (Centre de Recherches Historiques, EHESS) - bertrand.dumenieu@ehess.fr
    - Sébastien Poublanc (FRAMESPA, INSHS) - sebastien.poublanc@univ-tlse2.fr (à vérifier)
  - Atelier en 2024
  - Résumé : Les humanités numériques constituent un domaine à l'interface des arts, lettres, sciences humaines et sociales et des sciences du numérique. L'objectif de SaD-HN est de s'intéresser aux approches permettant de structurer les données et les connaissances qu'elles contiennent à des fins d'accès, de réutilisation et d'analyses. La finalité des approches proposées est de mettre en place des solutions opérationnelles offertes par les sciences du numériques pour soutenir et amplifier l'exploitation des données manipulées et produites dans les domaines relevant des humanités.
- **Traitement Informatique des Données de Santé**
  - Acronyme : TIDS
  - Responsables :
    - Nicolas Lachiche - ICUBE UMR 7357 Strasbourg - INS2i- nicolas.lachiche@unistra.fr

  - — Natalia Grabar - STL UMR8163 Lille - INSHS - natalia.grabar@univ-lille.fr
  - — Christine Sinoquet - LS2N UMR 6004 Nantes - INS2I - christine.sinoquet@ls2n.fr
  - Atelier en 2024
  - Résumé : L'objectif de l'Atelier TIDS consiste à échanger sur les problématiques liées à l'exploitation secondaire des données de santé, au niveau national. L'objectif est de regrouper la communauté qui s'intéresse aux données de santé (chercheurs provenant de différentes disciplines, cliniciens, industriels, experts des institutions gouvernementales, associations de patients) pour aborder différents aspects des données de santé. Il s'agit en particulier de savoir comment lever les verrous lors de l'accès aux données cliniques, y compris dans le cadre des collaborations scientifiques et projets communs.

- ▷ **EducAction**
  - Acronyme : EducAction
  - Responsables :
    - — Sihem Amer Yahia LIG UMR 5217 sihem.amer-yahia@univ-grenoble-alpes.fr
    - — Emilie Hoareau emilie.hoareau@univ-grenoble-alpes.fr CERAG/Grenoble IAE
    - — Philippe Dessus Philippe.Dessus@univ-grenoble-alpes.fr LaRAC Grenoble
  - Atelier en 2023 (pas d'action en 2024)
  - Résumé : La crise sanitaire liée à la Covid-19 a vu s'accentuer l'utilisation des plateformes de travail numériques (i.e. visioconférence, plateformes éducatives de type MOOC, Learning Management Systems, crowdsourcing ou encore freelancing). L'utilisation des technologies éducatives (EdTech) fondées sur les avancées récentes des SHS et de la science des données est une voie prometteuse pour aider les acteurs de l'éducation. Cependant, ces technologies restent encore limitées en termes de représentation de la richesse des interactions sociales et de l'apprentissage, et leur accès est inégalitaire et sujet à espionnage des données. L'atelier EDUC'ACTION s'inscrit dans la perspective des travaux sur l'AIED (AI for education) et propose d'aborder la question de la montée en compétences pour tous sous les angles informatique et SHS, en se fondant sur une réflexion éthique permettant de mettre en balance les bienfaits et les préjudices de l'IA.

- ▷ **Federated Learning for Sensitive Data**
  - Acronyme : FedSeD
  - Responsables :
    - — Imen Megdiche, IRIT, imen.megdiche@irit.fr
    - — Marco Lorenzi, INRIA Côte d'Azur, marco.lorenzi@inria.fr
    - — Laetitia Kameni, Accenture Labs SophiaAntipolis - laetitia.kameni@accenture.com
  - Atelier en 2022, en 2023
  - Résumé : Malgré la quantité de données disponibles aujourd'hui dans les systèmes d'information des entreprises ou générée par des objets connectées, ces données ne participent pas systématiquement et simplement à des tâches d'apprentissage machine ciblant l'aide à la prise de décision. Cette complexité est souvent reliée à la nature de la donnée qui peut être sensible ou à l'effort non négligeable qui doit être déployé pour rendre la donnée compatible avec d'autres sources données. Depuis quelques années des approches dites de 'federated learning' ont émergé. Le principe de ces approches est de faire collaborer sur des tâches d'apprentissages plusieurs clients détenant leurs données sans les sortir de leurs silos. Cet atelier propose de réunir la communauté scientifique et des industriels pour mettre en

évidence des uses cases et prioriser les difficultés posées par ce type d'approche.

- ▷ **Federate ExplaiNability Definitions and Research**
  - ○ Acronyme : FENDER
  - ○ Responsables :
    - — Julien Aligon, IRIT, julien.aligon@ut-capitole.fr
    - — Michael Baker, Institut Interdiscilplinaire de l'Innovation (INSHS)
    - — Nicolas Labroche, LIFAT, informatique
  - ○ atelier en 2022 et 2023, action HELP en 2024
  - ○ Résumé : L'atelier FENDER s'intéresse à l'explicabilité et l'interprétabilité du point de vue de l'informatique et du point de vue des sciences humaines et a pour objectif de voir ce que chaque champ disciplinaire peut apporter à l'autre dans la manière de représenter, définir et même consommer ou évaluer des explications et le rôle des données et leurs propriétés dans ces explications. Les travaux amèneront notamment à s'interroger sur la forme d'une explication et en proposer de nouvelles permettant de voir l'explication d'algorithmes ou de données comme un processus exploratoire continu et co-construit avec l'utilisateur cible de l'explication.
- ▷ **De la gestion à l'analyse de données pour l'agriculture, l'écologie et l'environnement**
  - ○ Acronyme : AGEE
  - ○ Responsables :
    - — Sandro Bimonte, INRAE, Clermont-Ferrand
    - — Florence Le Ber, ICUBE
    - — Yvan Le Bras, Muséum national d'Histoire naturelle
    - — François Brun, ACTA, Toulouse
    - — Tassadit Bouadi, IRISA
  - ○ Atelier en 2019 et 2020
  - ○ Résumé : L'agriculture, l'écologie et l'environnement de manière générale sont aujourd'hui de gros producteurs de données, et ceci à partir de différentes sources : capteurs physiques (sondes, satellites, GPS, etc.) ou humains. Les techniques et méthodologies du Big Data et des sciences de données permettent aujourd'hui de pouvoir stocker, gérer et analyser des données complexes et volumineuses. Elles ont déjà montré leurs atouts dans les champs de l'agriculture, de l'écologie et de l'environnement. L'atelier a pour objectif de partager et tester des méthodes, et de permettre aux participants de partager des jeux de données intéressants, données ouvertes, données issues de projets académiques, données d'observation et données participatives.
- ▷ **Humanités Numériques**
  - ○ Acronyme : HN
  - ○ Responsables :
    - — Cyril Grouin, LIMSI, INS2I
    - — Natalia Grabar, STL, INSHS
    - — Fatiha Idmhand, CRLA, INSHS
    - — Sabine Loudcher, ERIC, INS2I

- Atelier en 2020
- Résumé : L'atelier HN a pour objectif d'incorporer des éléments méthodologiques utilisés en SHS (e.g. processus de constitution de corpus et de mise en relation de données, méthodologies d'analyse de corpus ...) dans les processus informatiques, et vice versa. Il propose d'interroger en particulier l'enjeu que les métadonnées des masses ou des collections de données représentent pour les Humanités à l'heure du Web sémantique.

▷ **Outils Statistiques pour l'Imagerie hyperSpectrale du milieu interstellaire**

- Acronyme : OSIS
- Responsables :
  - Emeric Bron, LERMA, INSU
  - Antoine Roueff, Institut Fresnel, INSIS
- Atelier en 2020
- Résumé : L'objectif de l'atelier OSIS est de développer des collaborations entre d'une part des astrophysiciens (spécialistes des observations et des modèles physico-chimiques) et d'autre part des spécialistes en traitement statistique, apprentissage et problèmes inverses, afin de trouver des approches adaptées pour exploiter pleinement les jeux de données concernant l'Imagerie hyperSpectrale du milieu interstellaire.

A.3 Liste des évènements labellisés ou soutenus dans la période 2020-2024

Type	lab/Sout	Titre	Action	Date	Part
Ecole	Label	JFMS 2024 : Journées Francophones de la Modélisation et de la Simulation		04/11/24	52
Journée	Label	Workshop MACLEAN : MAChine Learning for EArth ObservatioN		09/09/24	30
Ecole	Label	École d'été Deep Learning for Medical Imaging (DLMI) 2024		08/07/24	69
Ecole	Label	Ecole thématique Masses de Données Distribuées		23/06/24	42
Journée	Label	ExCH-24 : Journées sur l'explicabilité et la gestion d'informations géographiques et spatiales	HELP	17/06/24	30
Journée	Soutien	Sixième édition du Symposium MaDICS		29/05/24	129
Journée	Soutien	Sparsification de grands Graphes		28/05/24	40
Journée	Label	Econom'IA		06/05/24	30
Journée	Label	Journée IA et Humanités Numériques	SaD-HN	03/05/24	56
Journée	Soutien	GreenDays 2024		27/03/24	80
Journée	Label	Structures multi-échelles dans les données musicales	Musiscale	14/02/24	14
Journée	Label	Machine Learning for EArth Observation Data (Workshop @ECML/PKDD2023)		18/09/23	80
Journée	Label	DOING@ADBIS 2023 : 4th International Workshop on Intelligent Data From Data to Knowledge	DOING	04/09/23	18
Ecole	Soutien	Advanced computational analysis for behavioral and neurophysiological recordings		06/08/23	20
Journée	Label	Stockage de données numériques dans de l'ADN synthétique - Grandes avancées et défis à relever		03/07/23	63
Journée	Soutien	Cinquième Symposium du GDR CNRS MaDICS		24/05/23	92
Journée	Label	Journées de travail DOING [Appel à présentations]	DOING	13/04/23	
Journée	Label	Mots/Machines 5 : Terminologie	SimpleText	17/03/23	57
Journée	Soutien	Journée Recherche Reproductible		08/03/23	167
Journée	Label	Journées d'étude Musiscale	Musiscale	06/10/22	20

Type	Lab/Sout	Titre	Action	Date	Part
Journée	Label	DOING@ADBIS 2022 : 3rd workshop on Intelligent Data from data to knowledge	DOING	05/09/22	23
Journée	Soutien	Quatrième édition du Symposium MaDICS		11/07/22	96
Journée	Soutien	Mini symposium JOBIM 2022 GIDAPE	RoCED	07/07/22	52
Journée	Label	Journée MACLEAN @ CAp/RFIAP (Vannes - 5 juillet 2022)	MACLEAN	05/07/22	120
Journée	Label	Atelier ROCED @ PFIA 2022	RoCED	27/06/22	42
Journée	Soutien	Application of Reasoning on Complex and Evolving Data : Methods and Use-Cases	RoCED	02/05/22	43
Ecole	Label	JFMS2022 : Journées francophones de la modélisation et de la simulation	DOING	28/03/22	34
Ecole	Label	AstroInformatique 2021	BigData4Astro	29/11/21	50
Journée	Label	SimpleText@CLEF	SimpleText	21/09/21	34
Journée	Label	2nd Workshop DOING : Intelligent Data - From Data to Knowledge (ADBIS 2021)	DOING	24/08/21	22
Ecole	Label	InnEO Summer School	MACLEAN	19/07/21	40
Journée	Label	Raconter Rennes par la donnée	MADONA	07/07/21	9
Journée	Label	Troisième édition du Symposium MaDICS		05/07/21	273
Journée	Label	2ème Webinar Action DOING MADICS	DOING	17/06/21	34
Journée	Label	SimpleText@INFORSID	SimpleText	01/06/21	21
Journée	Label	ALIAS	LEMON	04/12/20	
Journée	Label	36ème Conférence sur la Gestion de Données Principes - Technologies et Applications		27/10/20	150
Journée	Label	Action PLATFORM dans GdR Internet et Société	PLATFORM	25/09/20	44
Journée	Label	DOING 2020 : International Workshop on Intelligent Data From Data to Knowledge	DOING	25/08/20	21
Journée	Label	Second Symposium GDR MaDICS		06/07/20	320
Ecole	Label	Ecole d'été ETAL		29/06/20	30
Ecole	Label	Ecole BDA Masses de Données 2020		20/06/20	54
Ecole	Label	(JC)2BIM		08/06/20	30
Ecole	Soutien	European VIVA SUMMER SCHOOL on Artificial Intelligence and Software Verification and Validation		08/06/20	42
Journée	Label	Journées Francophones de la Modélisation et de la Simulation : JFMS 2018		13/04/20	29
Journée	Label	Hands-on Workshop on Machine Learning Applied to Medical Imaging	ATLAS	09/03/20	30

A.4 Liste des annonces d'évènements co-organisés par les actions pendant la période 2020-2024

Titre	Action	Date	Lieu
Appel à communication - EGC 2025	TIDS	27/01/2025	Strasbourg
5th International Workshop on Intelligent Data - From Data to Knowledge	DOING	28/08/2024	Bayonne
2024 Summer School on AI Technologies for Trust Interoperability Autonomy and Resilience in Industry 4.0'	DFAe	22/07/2024	Saint-Étienne
DEBS 2024 - ACM International Conference on Distributed and Event-Based Systems	DOING	25/06/2024	Lyon
1st Edition of Econom'IA Conference	DOING	06/05/2024	Bordeaux
Intelligence Artificielle et Humanités Numériques	SaD-HN	03/05/2024	BNF Paris
Journée Thématique : Gestion et Analyse des données Aériennes et Satellitaires (G2AS' 24)	DOING	17/04/2024	EPITA - Le Kremlin-Bicêtre
COnférence en Recherche d"Information et Applications CORIA	SimpleText	03/04/2024	La Rochelle
Atelier : La place des usagères et usagers dans les outils de fouille et d'exploration de données (PAUL) @ EGC	DOING	23/01/2024	Dijon
Workshop AI for Biological Imaging	DOING	08/01/2024	Sorbonne Université Paris
CFP : Workshop on Artificial Intelligence for Predictive Maintenance and IIoT (AI4PMI) @IEEE AICCSA	DOING	04/12/2023	Le Caire (Egypte)
11 :47 Debating the potential of machine learning for astronomical surveys (2)	BigData4Astro	27/11/2023	Institut Astrophysique de Paris
Intelligence Artificielle et Gestion des Informations et des Donnees Imparfaites et Hétérogènes	RoCED	30/10/2023	34200 Sète
Appel à contributions/participations : 11èmes Journées de la Société Française de Chémoinformatique	DSChem	05/10/2023	Caen
SimpleText Shared Tasks [DL 28/04/23] : Automatic Simplification of Scientific Texts	SimpleText	18/09/2023	Thessaloniki (Greece)
DOING@ADBIS2023	DOING	04/09/2023	Barcelona
13th International Conference on Formal Ontology in Information Systems (FOIS 2023)	RoCED	17/07/2023	Sherbrooke (Canada)
SOEM : Science Ouverte et Sémantique associé à IC@PFIA 2023	RoCED	07/07/2023	Strasbourg
École Thématique AstroInformatique 2023 et Hackathon AstroInfo AISSAI	BigData4Astro	26/06/2023	Sud-Est de la France
Appel à participation en visio pour la journée commune AFIA-EGC le 11 mai 2023 à EPITA - Paris	FedSed	11/05/2023	EPITA Le Kremlin-Bicêtre

titre	action	date	lieu
Journées de travail DOING [Appel à présentations]	DOING	13/04/2023	Orléans
1st International Conference on Data & Digital Humanities | Text Mining and Multimodal Storytelling	SimpleText	08/03/2023	Universidade do Minho (Portugal)
Appel à communication pour l'atelier GAST/EGC'2023 (Gestion et Analyse des données Spatiales et Temporelles)	RoCED	17/01/2023	Lyon
Appel à communication pour l'atelier GAST/EGC'2023 (Gestion et Analyse des données Spatiales et Temporelles)	MACLEAN	17/01/2023	Lyon
Appel à communication - Atelier EXPLAIN'AI 2023 (2ème édition) @ EGC 2023	HELP	17/01/2023	Lyon
PhD Defense : Explainable Classification of Uncertain Time Series	BigData4Astro	13/12/2022	ISIMA
ASNUM2022 : Conférence Action Spécifique Numérique Astrophysique	BigData4Astro	12/12/2022	ENS Lyon
Appel à communications – Atelier ALIAS 2022	LEMON	07/12/2022	Université Paris Nanterre
IEEE GRSS second workshop on Remote Sensing Data Management Technologies in GeoScience 2022	MACLEAN	02/12/2022	Paris-Dauphine University
Prochain séminaire du GDR IA : Marie-Christine Rousset	RoCED	01/12/2022	en ligne
Atelier@EGC'23 : Mécanismes d'Attention et Apprentissage Automatique : avancées récentes et perspectives	MACLEAN	28/10/2022	Lyon
BDA'2022	DOING	24/10/2022	LIMOS
Advanced course on Deep Learning and Geophysical Dynamics	MACLEAN	19/10/2022	Brest
AALTD@ECML : 7th Workshop on Advanced Analytics and Learning on Temporal Data	SimpleText	23/09/2022	Grenoble
Machine Learning for Earth Observation data (MACLEAN22) Workshop @ECML/PKDD2022	MACLEAN	19/09/2022	Grenoble
Summer school : IA for EO For PhD or young researchers - no registration fees	MACLEAN	25/07/2022	en ligne
International Federation of Classification Societies	MACLEAN	19/07/2022	Porto - Portugal
1er atelier pluridisciplinaire : narration de données datajournalisme et engagement(s)	MADONA	12/07/2022	Lyon
Apprentissage automatique et vision par ordinateur : résultats scientifiques face aux besoins industriels	MACLEAN	05/07/2022	Vannes
Machine learning and computer vision in earth observation : scientific results versus industrial needs	MACLEAN	05/07/2022	Vannes

Titre	Action	Date	Lieu
CIRCLE 2022 | 2nd edition of the Joint Conference of the Information Retrieval Communities in Europe	LEMON	04/07/2022	Samatan - Gers
Atelier ROCED à PFIA 2022	RoCED	27/06/2022	Saint Etienne
summer school on Point clouds and change detection in the geosciences	MACLEAN	22/06/2022	en ligne
Session 4.8 – Remote Sensing Methods and Applications @MELECON 2022	MACLEAN	14/06/2022	Palermo – Italie
Training on Entrepreneurship	MACLEAN	30/05/2022	en ligne
[CFP] SimpleText@CLEF : Automatic Simplification of Scientific Texts	SimpleText	27/05/2022	
new EXTENDED DEADLINE : DOING@ADBIS 2022	DOING	26/05/2022	Torino (Italie)
CfP - RCIS 2022	DOING	17/05/2022	Barcelona (Espagne)
Special issue Text Complexity and Simplification in Frontiers in Artificial Intelligence / Natural Language Processing	SimpleText	03/05/2022	
1-st Call for Participation - SimpleText Track @ CLEF-2022	SimpleText	22/04/2022	
Les journées Francophones de la modélisation et de la simulation 2022	DOING	28/03/2022	IES de Cargèse (Corse)
Journée d'étude Mots/Machines 4 : Simplification et adaptation du texte	SimpleText	25/03/2022	Brest Cedex 3
EXTENSION 8ème École d"Hiver é-EGC L"humain dans la boucle de l"exploration des données et de l"apprentissage	SimpleText	24/01/2022	Blois
Ecole Thématique AstroInformatique 2021	BigData4Astro	29/11/2021	Barcelonnette
Observation 3D : outils et verrous	MACLEAN	24/11/2021	Paris
publication du numéro spécial Agriculture Numérique dans ROIA	RoCED	18/11/2021	
TAL pour la détection du discours de haine	LEMON	16/11/2021	IRIT – Toulouse
ASI 11-2021 - 11ème Colloque international sur l'Analyse Statistique Implicative	SimpleText	03/11/2021	IUT de Belfort-Montbéliard
7th International Conference on Algorithmic Decision Theory	PLATFORM	28/10/2021	IRIT - Toulouse
Journée du GDR TAL — 5 octobre 2021	SimpleText	05/10/2021	Toulouse
Machine Learning for Earth Observation data (MACLEAN21) Workshop @ECML/PKDD2021	MACLEAN	13/09/2021	Bilbao – Espagne
Advanced course on Deep Learning and Geophysical Dynamics	MACLEAN	13/09/2021	Brest
Demi-journée MACLEAN dans le cadre du Symposium MADICS	MACLEAN	05/07/2021	
Colloque Droit IA et Santé (DRIAS)	PLATFORM	21/06/2021	en ligne
2ème Webinar Action DOING – MADICS	DOING	17/06/2021	en ligne

titre	action	date	lieu
SimpleText@INFORSID Simplification et Vulgarisation des Textes Scientifiques	SimpleText	01/06/2021	Dijon
Atelier SimpleText@INFORSID — extension de la date limite	SimpleText	01/06/2021	Dijon
SAGEO 2021 à les 5-6 et 7 mai 2021	RoD	05/05/2021	La Rochelle
SAGEO 2021 à les 5-6 et 7 mai 2021 en distanciel	RoCED	05/05/2021	en ligne
Enquête-Invitation Premier Webinar Action DOING – MADICS	DOING	10/03/2021	en ligne
Kick-off DOING - MADICS ACTION : Keynote G. Fletcher	DOING	10/03/2021	en ligne
[Kind Reminder] Webinar DOING : G. Fletcher TUE	DOING	10/03/2021	en ligne
Second Appel à contribution pour le numéro spécial Intelligence Artificielle et Agriculture Numérique	RoCED	30/01/2021	
Webinaire INRAE : Les systèmes d'information agro-environnementaux à l'ère du Big Data	AGEE	29/01/2021	en ligne
Sondage : Vulgarisation scientifique	SimpleText	14/01/2021	
ICDAR2021 : Historical Map Segmentation challenge	MaDICS-HN	08/12/2020	Lausanne
Quels problèmes posent l"hétérogénéité des données en informatique et en Humanités numériques ? Reg	MaDICS-HN	25/11/2020	en ligne
BDA 2020 - 3e appel à communications	RoD	27/10/2020	en ligne
[SFC - Workshop virtuel (25 septembre 2020)] - Nouvelles méthodes pour l'analyse descriptive et pré	MACLEAN	25/09/2020	en ligne
Machine Learning for Earth Observation data Workshop @ECML/PKDD2020	MACLEAN	14/09/2020	Ghent (Belgium)
5th ECML/PKDD Workshop on Advanced Analytics and Learning on Temporal Data	MACLEAN	14/09/2020	
DOING 2020 : International Workshop on Intelligent Data – From Data to Knowledge	DOING	25/08/2020	Lyon
Second international workshop on Qualitative Aspects of User-Centered Analytics	MADONA	25/08/2020	Lyon
Appel à participation aux Journées Rod	RoD	06/07/2020	en ligne
Journée Thématique Agronomie et IA à PFIA 2020 – Appel à présentation	AGEE	29/06/2020	Angers
Deuxième Webinar DOING – MADICS	DOING	17/06/2020	en ligne
Premier Webinar DOING – MADICS	DOING	05/06/2020	en ligne
[Rappel] : Premier Webinar DOING – MADICS	DOING	05/06/2020	en ligne
CfP : MACLEAN2 : MAChine Learning for EArth ObservatioN (workshop @ECML/PKDD2020)	MACLEAN	08/05/2020	Ghent (Belgium)

Annexe B

Autres Sections CNRS dans le périmètre de MaDICS

Table B.1: Autres sections du CNRS dont les mots clés sont liés au périmètre scientifique de MaDICS

Section	**Description**
01	▷ Science et traitement massif des données, recherche d'événements rares
17	▷ Outils numériques et calcul haute performance ▷ Analyse de données haute performance ▷ Bases de données, catalogues
34	▷ Linguistique computationnelle : traitement automatique des langues, traitement du langage naturel
41	▷ Systèmes dynamiques et équations différentielles ordinaires, théorie ergodique, équations aux dérivées partielles, physique mathématique, probabilités, statistiques, apprentissage automatique, modèles stochastiques, traitement de données, aspects mathématiques du traitement du signal et de l'image, analyse numérique et calcul scientifique

Annexe C

Méthodologie adoptée pour la préparation du nouveau projet

Pour préparer le nouveau projet, une méthodologie structurée en plusieurs étapes a été mise en place. Tout d'abord, une équipe projet a été formée, servant de préfiguration pour le futur comité de direction du GDR. Un calendrier de travail a été élaboré (voir section suivante), comprenant des échanges réguliers avec l'institut CNRS Sciences Informatiques. Une consultation des Actions et Ateliers a été réalisée par le biais d'un questionnaire visant à recueillir leurs avis sur les enjeux de MaDICS ainsi qu'à dresser un bilan du fonctionnement et de la gouvernance. Le projet MaDICS 2025-2029 a été présenté et discuté avec l'ARA lors de la réunion annuelle précédant le symposium. Enfin, la politique scientifique de l'Institut CNRS Sciences Informatiques et le projet MaDICS 2025-2029 ont été exposés à la communauté MaDICS lors du symposium, par Anne Siegel (Directrice adjointe scientifique à l'Interdisciplinarité et aux Interfaces, CNRS Sciences Informatiques) et le comité de direction de MaDICS, permettant une large consultation avec la communauté.

C.1 Composition de l'équipe projet

Bernd Amann	LIP6 UMR 7606, Sorbonne Université
Khalid Belhajjame	LAMSADE UMR 7243, Université Paris-Dauphine
Frédéric Bimbot	IRISA UMR 6074, Rennes
Christophe Bobineau	LIG UMR 5217, Institut polytechnique de Grenoble
Sarah Cohen-Boulakia	LISN UMR 9015, Université Paris-Saclay
Bruno Crémilleux	GREYC UMR 6072, Université de Caen Normandie
François Goasdoué	IRISA UMR 6074, Université de Rennes
Nathalie Hernandez	IRIT UMR 5505, Université de Toulouse
Myriam Maumy-Bertrand	LIST3N, Université de Technologie de Troyes
Nathalie Pernelle	LIPN UMR 7030, Université Sorbonne Paris Nord
Farouk Toumani	LIMOS UMR CNRS 6158, Université Clermont Auvergne

C.2 Calendrier

- 1er février 2024 : Sollicitation de l'ARA pour une contribution en deux temps :
  - Volet 1 : identité et positionnement du GDR MaDICS/objectifs du GDR
  - Volet 2 : organisation et gouvernance du GDR

- ▷ 14 février 2024 : réunion projet
- ▷ 1er mars 2024 : retours des contributions de l'ARA sur le volet 1
- ▷ 14 mars 2024 : réunion projet
- ▷ 21 mars 2024 : réunion de cadrage avec l'institut CNRS Sciences Informatiques (3-5 enjeux majeurs, ébauche de gouvernance)
- ▷ 15 avril 2024 : Retours des contributions de l'ARA sur le volet 2
- ▷ 18 avril 2024 : réunion projet
- ▷ 15 mai 2024 : réunion projet (discussion sur la V0)
- ▷ 17 mai 2024 : V0 du volet 1 (transmission à l'ARA)
- ▷ 23 mai 2024 : réunion de cadrage avec l'institut CNRS Sciences Informatiques (discussion sur V0)
- ▷ 24 mai 2024 : réunion avec le GDR RADIA
- ▷ 28 mai 2024 : réunion projet (Symposium) échanges autour de la V0 du volet 1
  - ○ ARA : échanges autour de la V0 du volet 1
  - ○ échange autour de l'organisation et de la gouvernance du GDR
- ▷ 4 juin 2024 : réunion avec le GDR TAL
- ▷ 17 juin 2024 : version V1 du projet (transmission à l'ARA)
- ▷ 18 juin 2024 : journée préparation du projet, finalisation V1
- ▷ 1 juillet 2024 : retour ARA sur la V1
- ▷ 9 juillet 2024 : réunion projet
- ▷ 26 juillet 2024 : version V0 du dossier de renouvellement (transmission à l'Institut CNRS Sciences Informatiques)
- ▷ 18 août 2024 : version V1 du dossier de renouvellement (transmission à l'Institut Sciences Informatiques)
- ▷ 22 août 2024 : retours de l'Institut CNRS Sciences Informatiques sur la V1
- ▷ 29 août 2024 : réunion projet
- ▷ 1^{er} septembre 2024 : dépôt de la version finale du dossier de renouvellement du GDR MaDICS à l'institut CNRS Sciences Informatiques

Annexe D

Liste des membres du GDR MaDICS

La base de données de MaDICS recense un total de 1 788 membres, dont 1 076 nouvelles adhésions depuis janvier 2020. Nous avons initié un processus de renouvellement des adhésions auprès de l'ensemble des membres du GDR dans la perspective du nouveau projet 2025-2029. La liste actualisée des membres sera disponible le 1^{er} octobre 2024.