Dossier de candidature aux fonctions de maître de conférences

Section 27, Poste N°404

Alexandre Bazin

48 Rue Henri Poincaré
54000 Nancy
contact@alexandrebazin.com

Contents

Curriculum Vitae

Données personnelles

- Né le 29 mai 1987 à Auxerre
- Nationalité française
- Site professionnel : www.alexandrebazin.com
- Adresse professionnelle : 615, rue du Jardin Botanique, 54600 Villers-lès-Nancy
- Adresse électronique : contact@alexandrebazin.com

Domaines de recherche

- Découverte de connaissances
- Représentation des connaissances
- Fouille de données
- Analyse formelle de concepts

Formation

- **2010 - 2014 : Doctorat d'informatique** de l'Université Pierre et Marie Curie (UPMC), Paris :

  On the Enumeration of Pseudo-Intents : Choosing the Order and Extending to Partial Implications.

  avec mention très honorable, soutenue le 30/09/2014 devant le jury composé de

Mme Annick Valibouze	Professeure à l'UPMC	Présidente
Mme Karell Bertet	Mcf Hdr à l'Université de La Rochelle	Rapporteur
M. Henry Soldano	Mcf Hdr à l'Université Paris-Nord	Rapporteur
M. Alain Gély	Mcf à l'Université de Lorraine	Examinateur
M. Jean-Gabriel Ganascia	Professeur à l'UPMC	Directeur de thèse

- **2008 - 2010 : Master d'informatique** spécialité Intelligence Artificielle et Décision de l'Université Pierre et Marie Curie (UPMC), Paris. Mention Bien.
- **2005 - 2008 : Licence d'informatique** de l'Université de Bourgogne, Dijon. Mention Assez Bien.

Parcours professionnel

- **Janvier 2021 - Présent :** Post-doctorant à l'Institut national de recherche en sciences et technologies du numérique (Inria), Nancy, France. Projet ANR : "ELKER – Enhancing Link Keys: Extraction and Reasoning".
- **Janvier 2019 - Décembre 2020 :** Post-doctorant rattaché au Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA), Nancy, France. Projet LUE : "Functional Genomic, Epigenomic and ENvironnent interplay to IMPACT the Understanding, diagnosis and management of healthy and pathological AGEing (GEENAGE)".

- **Octobre 2017 - Décembre 2018 :** Post-doctorant rattaché au Laboratoire Electronique, Informatique et Image (Le2I), Dijon, France. Projet Eurostar : "Development of a predictive smart data platform for real time intelligence".
- **Octobre 2015 - Mars 2017 :** Post-doctorant rattaché au Laboratoire d'Informatique, de Modélisation et d'Optimisation des Systèmes (LIMOS), Clermont-Ferrand, France. Projet financé par la région Auvergne : "Étude de la biosphère rare microbienne par une approche in silico : nouvelle méthode de classification ensembliste et modélisation".
- **Septembre 2013 - Août 2014 :** ATER à temps plein (194h), rattaché au Laboratoire d'Informatique de Paris 6 (LIP6) à l'Université Pierre et Marie Curie (UPMC), Paris, France.
- **Octobre 2010 - Août 2013 :** Doctorant contractuel au Laboratoire d'Informatique de Paris 6 (LIP6) à l'Université Pierre et Marie Curie (UPMC), Paris, France. Thèse : "*On the Enumeration of Pseudo-Intents : Choosing the Order and Extending to Partial Implications*".

Activités pédagogiques et projet d'enseignement

Enseignements

J'ai effectué un total de 262 heures d'enseignement. En qualité d'ATER à l'Université Pierre et Marie Curie, entre septembre 2013 et mai 2014, j'ai enseigné dans les matières suivantes (194 heures) :

- **Introduction à la programmation** (38 heures de TP)

  Public : Etudiants de première année de licence Mathématiques, Physique et Informatique à l'Université Pierre et Marie Curie (UPMC)

  Contenu : Introduction à l'algorithmique et à la programmation à l'aide du langage fonctionnel Scheme. Correction des TP notés.

- **Atelier de recherche encadré** (40 heures de TD/TP)

  Public : Etudiants de première année de licence d'informatique à l'Université Pierre et Marie Curie (UPMC)

  Contenu : Enseignement du langage Python (20 heures) puis encadrement de projets en équipe (4 groupes de 4 étudiants). Participation au jury de soutenance.

- **Langage C** (60 heures de TD/TP)

  Public : Etudiants de première année à l'Institut de Statistiques de l'Université de Paris (ISUP)

  Contenu : Enseignement de l'algorithmique et de la programmation en langage C à un public hétéroclite. Préparation des feuilles d'exercice.

- **Modélisation et représentation des connaissances** (16 heures de TD/TP)

  Public : Etudiants de première année de master à l'Université Pierre et Marie Curie (UPMC)

  Contenu : Enseignement de différentes techniques de modélisation des connaissances : logiques de description, graphes de Sowa, réseaux sémantiques et logique du premier ordre. Correction des examens.

- **Méthodes symboliques pour l'intelligence artificielle** (40 heures de TD/TP)

  Public : Etudiants de première année de master à l'Université Pierre et Marie Curie (UPMC)

  Contenu : Enseignement de diverses techniques fondamentales de l'intelligence artificielle symbolique : systèmes à base de règles de production, raisonnement dans l'incertain, programmation par ensembles-réponses, raisonnement qualitatif, arbres de jeu et planification. Correction des examens.

En qualité de post-doctorant à l'Université Clermont Auvergne, entre février 2017 et mars 2017, j'ai enseigné dans la matière suivante (14 heures) :

- **Apprentissage artificiel** (14 heures de TP)

  Public : Etudiants de troisième année de licence Informatique à l'Université Clermont Auvergne (UCA)

  Contenu : Enseignement de techniques d'apprentissage et classification à l'aide du logiciel Weka et du langage R : arbres de décision, k plus proches voisins, règles d'association. Correction des compte-rendus de TP.

En qualité de post-doctorant à l'Université de Lorraine, entre octobre 2019 et janvier 2020, j'ai enseigné dans les matières suivantes (54 heures) :

- **Algorithmique pour le génie industriel** (14 heures de TD)

  Public : Etudiants de première année à l'Ecole des Mines de Nancy

  Contenu : Enseignement des bases de l'algorithmique et des structures associées : complexité algorithmique, algorithmes de tri, expression régulières, automates, arbres. Correction de TP notés.

- **Structures de données** (40 heures de cours et TP)

  Public : Etudiants de première année à l'IUT Nancy-Charlemagne

  Contenu : Enseignement des structures de données les plus utilisées (listes et tables) sous forme logique et les stratégies possibles pour leur implémentation. Séances intégrant cours et travaux pratiques.

Encadrement de stages

- **Romain Dalbard**, Stagiaire de juin à août 2019 (2 mois), 3ème année à l'Ecole des Mines de Nancy.
  - **Titre** : Fouille de données pour la découverte de connaissances dans des données de métabolomique
  - **Contenu** : Romain Dalbard avait pour tâche de reproduire et vérifier les résultats expérimentaux d'une publication sur l'identification de variables prédictives et discriminantes dans un jeu de données biomédicales. Il était chargé d'étudier les performances de classifieurs de type *random forests* et *support vector machines* aussi utilisés dans la publication.
  - **Résultats** : Il a pu montrer que les informations de la publication de référence n'étaient pas suffisantes pour reproduire les résultats avec les mêmes performances de classification. Il a aussi pu identifier les ensembles de variables à présélectionner pour optimiser ces performances.
  - **Co-supervision** : Amedeo Napoli.
- **Murat Koşak**, stagiaire de juillet à octobre 2019 (3 mois), Bachelor à l'Université de Galatasaray.
  - **Titre** : Supervised exploratory data mining and knowledge discovery in metabolomic data
  - **Contenu** : Murat Koşak avait lui aussi pour tâche de reproduire les résultats expérimentaux d'une publication sur l'identification de variables prédictives et discriminantes dans un jeu de données biomédicales. Il était chargé d'étudier les performances de classifieurs de type *Naive Bayes* et *k plus proches voisins* qui n'étaient pas utilisés dans la publication.
  - **Résultats** : Il a identifié les méthodes de sélection de variables et les paramètres optimaux pour optimiser les performances des classifieurs et obtenir des résultats rivalisant avec ceux de la publication de référence.
  - **Co-supervision** : Amedeo Napoli.
- **Eduardo Calò**, stagiaire de mars à août 2021 (6 mois), Master 2 TAL à l'Université de Lorraine.
  - **Titre** : Building a learning space for Mandarin language acquisition
  - **Contenu** : Eduardo Calò a pour tâche de développer un outil d'aide à l'enseignement et à l'apprentissage du mandarin. Pour cela, il doit collecter des données puis les utiliser pour construire un espace d'apprentissage représentant les différents états de connaissances qu'un étudiant en mandarin peut atteindre et étudier différentes manières de le parcourir pour aider les enseignants et les étudiants.
  - **Résultats** : Eduardo est parvenu à créer un espace de connaissances à partir de données qu'il a recueillies. L'outil permettant de parcourir cet espace est en cours d'intégration dans les outils de la startup NAIMROD.
  - **Co-supervision** : Miguel Couceiro, Benedict O'Donnell.

Intégration à l'IUT de Montpellier-Sète

Les différents cours, TD et TP dont j'ai été chargé étaient de deux types : les bases de l'informatique et l'intelligence artificielle. En bases de l'informatique, j'ai enseigné la programmation (Python, C et Scheme), l'algorithmique et les structures de données. En intelligence artificielle, j'ai enseigné la représentation des connaissances, les méthodes symboliques (raisonnements, théorie des jeux...) et l'apprentissage automatique. Les publics devant lesquels j'ai enseigné étaient eux aussi très variés : informaticiens (IUT, licence et master), statisticiens (ISUP), ingénieurs civils (Ecole des mines de Nancy), biologistes et physiciens ($1^{ère}$ année de licence). J'ai pu expérimenter différents formats pédagogiques : TD, TP, enseignements intégrés, enseignement par projet et classe inversée, que j'apprécie particulièrement. J'ai une expérience de l'encadrement de projets et de stages ainsi que de la conception de fiches d'exercices. Je suis aussi capable d'enseigner en anglais.

Le département informatique de l'IUT Montpellier-Sète recherche un enseignant pour intervenir dans le nouveau BUT. Il s'agira principalement d'enseigner dans les domaines du développement informatique et du génie logiciel. Mon expérience me permettra d'organiser et de dispenser directement la plupart de ces enseignements et je suis prêt à me mettre à niveau sur tout autre sujet, comme je l'ai déjà fait par le passé. Je serai aussi, selon les besoins, en mesure d'assurer des enseignements de mathématiques liés à l'informatique comme les mathématiques discrètes (graphes, hypergraphes, treillis...) et la logique (propositionnelle, du premier ordre, de description...), que je manipule dans mes activités de recherche.

Je pense qu'il est important, dans une formation professionnalisante telle que le BUT, de mettre l'accent sur l'autonomisation des étudiants. En effet, tout informaticien doit être capable d'apprendre de nouvelles notions et de se maintenir à niveau par lui-même à partir de ressources externes de qualité variable. Je prévois donc d'utiliser autant que possible des approches pédagogiques favorisant cette autonomie. En première année, il s'agira de cours intégrés dans lesquels les étudiants devront participer activement aux corrections d'exercices afin de mettre en place une dynamique de groupe et d'éviter qu'ils attendent passivement les réponses. En seconde et troisième année, il s'agira de formes de classes inversées dans lesquelles les étudiants apprendront par eux-mêmes les notions à partir de documents, fournis ou publics, avant de les mettre en pratique.

La réalisation de projets sera bien entendu au centre des enseignements que je proposerai. La meilleure façon d'intéresser les étudiants, surtout en première année, est à mon sens de leur proposer des réalisations qu'ils pourront utiliser ou montrer à leurs proches. J'illustrerai donc autant que possible les différentes notions par le développement de jeux (algorithmique, IHM, réseau...) ou de sites web.

Je suis bien entendu prêt à prendre des responsabilités au sein de l'IUT, que ce soit pour l'organisation de cours et d'évènements, le recrutement d'étudiants et d'enseignants ou la gestion des emplois du temps. Je suis notamment particulièrement enthousiaste à l'idée de participer à la création d'un nouveau diplôme et du nouveau site de Sète.

Publications et activités de recherche

Publications

Les publications mentionnées ici peuvent aussi être consultées sur www.alexandrebazin.com.

- **Journaux internationaux avec comité de lecture**

1. Alexandre Bazin. "*On Implication Bases in n-Lattices*." Discrete Applied Mathematics (DAM) Volume 273, p. 21-29, 2020. (pdf)
2. Alexandre Bazin, Didier Debroas, Engelbert Mephu Nguifo. "*A De Novo Robust Clustering Approach for Amplicon-Based Sequence Data*." Journal of Computation Biology (JCB) Volume 26, p. 618-624, 2018. (pdf)
3. Alexandre Bazin. "*A Depth-first Search Algorithm for Computing Pseudo-closed Sets*." Discrete Applied Mathematics (DAM) Volume 249, p. 28-35, 2018. (pdf)
4. Alexandre Bazin et Jean-Gabriel Ganascia. "*Computing the Duquenne-Guigues Basis : An Algorithm for Choosing the Order*." International Journal of General Systems (IJGS) Volume 45, Issue 2, p. 57-85, 2016. (pdf)

- **Conférences internationales avec comité de lecture**

5. Alexandre Bazin, Miguel Couceiro, Marie-Dominique Devignes, Amedeo Napoli. "Explaining Multicriteria Decision Making with Formal Concept Analysis." Concept Lattices and Applications (CLA), p. 119-130, 2020. (pdf)
6. Alexandre Bazin, Jessie Carbonnel, Marianne Huchard, Giacomo Kahn, Priscilla Keip, Amirouche Ouzerdine. "*On-demand Relational Concept Analysis*." International Conference on Formal Concept Analysis (ICFCA), p. 155-172, 2019. (pdf)
7. Alexandre Bazin et Giacomo Kahn. "*Reduction and Introducers in d-contexts*." International Conference on Formal Concept Analysis (ICFCA), p. 73-88, 2019. (pdf)
8. Alexandre Bazin et Aurélie Bertaux. "*k-Partite Graphs as Contexts*." Concept Lattices and Applications (CLA), p. 59-67, 2018. (pdf)
9. Giacomo Kahn et Alexandre Bazin. "*Average Size of Implicational Bases*." Concept Lattices and Applications (CLA), p. 37-46, 2018. (pdf)
10. Alexandre Bazin, Jessie Carbonnel, Giacomo Kahn. "*On-demand Generation of AOC-posets: Reducing the Complexity of Conceptual Navigation*." International Symposium on Methodologies for Intelligent Systems (ISMIS), p. 611-621, 2017. (pdf)
11. Alexandre Bazin. "*Comparing Algorithms for Computing Lower Covers of Implication-closed Sets*." Concept Lattices and Applications (CLA), p. 21-31, 2016. (pdf)
12. Alexandre Bazin. "*Depth-first Search for Pseudo-intents through Minimal Transversals*." International Conference on Formal Concept Analysis (ICFCA), actes secondaires "FCA&A", p. 1-16, 2015. (pdf)
13. Alexandre Bazin et Jean-Gabriel Ganascia. "*Enumerating Pseudo-Intents in a Partial Order*." Concept Lattices and Applications (CLA), p. 45-56, 2013. (pdf)
14. Alexandre Bazin et Jean-Gabriel Ganascia. "*Completing Terminological Axioms with Formal Concept Analysis*." International Conference on Formal Concept Analysis (ICFCA), actes secondaires, p. 29-40, 2012. (pdf)

- **Workshops internationaux**

15. Alexandre Bazin, Miguel Couceiro, Marie-Dominique Devignes, Amedeo Napoli. "*An Approach to Identifying the Most Predictive and Discriminant Features in Supervised Classification Problems*." International Conference on Conceptual Structures 2021, short paper. (pdf)

16. Nacira Abbas, Alexandre Bazin, Jérôme David, Amedeo Napoli. "Sandwich: an Algorithm for Discovering Relevant Link Keys in an LKPS Concept Lattice." International Conference on Formal Concept Analysis (ICFCA) 2021, short paper.
17. Alexandre Bazin, Laurent Beaudou, Giacomo Kahn, Kaveh Khoshkhah. "*Bounding the Number of Minimal Transversals in Tripartite 3-Uniform Hypergraphs*." International Colloquium on Graph Theory and combinatorics (ICGT) 2018. (pdf (version journal))
18. Alexandre Bazin, Didier Debroas, Engelbert Mephu Nguifo. "*A De Novo Robust Clustering Approach for Amplicon-Based Sequence Data*". bai@IJCAI2017, p. 21-25, 2017. (pdf)

- **Conférences nationales**

19. Alexandre Bazin, Aurélie Bertaux, Christophe Nicolle. "*Représentation condensée de règles d'association multidimensionnelles*." Extraction et Gestion des Connaissances (EGC), p. 225-236, 2019. (pdf)
20. Alexandre Bazin et Giacomo Kahn. "*Du nombre maximum d'ensembles fermés en 3 dimensions*." Extraction et Gestion des Connaissances (EGC), p. 345-350, 2019. (pdf)
21. Alexandre Bazin, Didier Debroas, Engelbert Mephu Nguifo. "*A De Novo Robust Clustering Approach for Amplicon-Based Sequence Data*". CNIA+RJCIA, p. 113-115, 2017. (pdf)

- **Pré-publications et rapports techniques**

22. Alexandre Bazin, Nicolas Gros, Aurélie Bertaux, Christophe Nicolle. "*Condensed Representations of Association Rules in n-ary Relations*." Soumis à IEEE Transactions on Knowledge and Data Engineering (TKDE) (major revision).
23. Alexandre Bazin, Laurent Beaudou, Giacomo Kahn, Kaveh Khoshkhah. "*Bounding the Number of Minimal Transversals in Tripartite 3-Uniform Hypergraphs*." Soumis à Discrete Mathematics & Theoretical Computer Science (DMTCS). (pdf)
24. Alexandre Bazin, Miguel Couceiro, Marie-Dominique Devignes, Amedeo Napoli. "*Steps Towards Causal Formal Concept Analysis*." Soumis à International Journal of Approximate Reasoning (IJAR). (pdf)
25. Nacira Abbas, Alexandre Bazin, Jérôme David, Amedeo Napoli. "*The Discovery of Link Keys Between two RDF Datasets Based on Partition Pattern Structures*." Soumis à International Conference on Information and Knowledge Management (CIKM) 2021.
26. Nacira Abbas, Alexandre Bazin, Jérôme David, Amedeo Napoli. "*From Pattern Structures to Partition Pattern Structures for Link Key Discovery*." Soumis à FCA4AI 2021.

- **Thèse de doctorat**

27. "*On the Enumeration of Pseudo-Intents : Choosing the Order and Extending to Partial Implications*." Alexandre Bazin. Manuscrit de thèse, 2014. (pdf)

Réalisation de logiciels

Les modules et logiciels décrits ici sont accessibles depuis `www.alexandrebazin.com`.

- SCLUST, un logiciel de clustering flou non-supervisé de séquences d'ADN
- PCA, un module python contenant des algorithmes pour le calcul d'implications, de règles d'associations et de concepts formels dans les données binaires multidimensionnelles
- exMCDM, un module python pour l'explication symbolique de la présence d'alternatives dans un front de Pareto
- GEENAGE, un module python pour la sélection de variables particulièrement discriminantes et prédictives d'une classe dans un jeu de données pour la classification supervisée

Présentations lors de rencontres et séminaires

- **Février 2021, Groupe de travail CODAG** : "*Explication de décisions multicritères avec l'analyse formelle de concepts*".
- **Décembre 2020, Séminaire MALOTEC** : "*On Algorithmic Approaches to Recognising Causal Relations*".
- **Février 2020, Séminaire à TalTech** : "*Association Rules in Multidimensionnal Data*". Tallinn, Estonie.
- **Mai 2019, Séminaire MALOTEC** : "*Rules in Symbolic Data*". Nancy.
- **Mars 2019, Séminaire doctorants Orpailleur** : "*Rules in Binary Data (And Why They are More Interesting in the Multidimensional Case)*". Nancy.
- **Juin 2017, Conférence ICFCA** : "*Local Generation of AOC-posets: Reducing the Complexity of Conceptual Navigation for SPLE Product Selection*" (Poster). Rennes.
- **Décembre 2016, Journée Intelligence Artificielle et Big Data** : "*Une nouvelle méthode de clustering avec incertitude pour de grands volumes de séquences*" (Poster). Lyon.
- **Décembre 2016, PEPS HYDRATA** : Présentation sur les définitions des ensembles fermés, quasi-fermés et pseudo-fermés. Caen.
- **Juin 2016, Conférence JOBIM** : "*Une nouvelle méthode de clustering avec incertitude de données de séquençage haut-débit*" (Poster). Lyon.
- **Juin 2016, PEPS HYDRATA** : Présentation sur les ensembles pseudo-fermés et leur calcul grâce aux traverses minimales. Paris.
- **Octobre 2015, Séminaire LIMOS** : "*Calculer la base canonique d'implications - influence de l'ordre*". Clermont-Ferrand.
- **Janvier 2014, Séminaire interne au LIP6** : Présentation sur l'énumération d'ensembles pseudo-fermés dans des ordres non-lectiques et extension à d'autres types de bases d'implications. Paris.
- **Mars 2013, Journées des treillis** : Présentation sur l'énumération d'ensembles pseudo-fermés dans des ordres non-lectiques. Metz.
- **Décembre 2011, Journée des treillis** : "*Apprentissage d'ontologies en logiques de descriptions par l'analyse formelle de concepts*". Rennes.

Responsabilités collectives

- Membre du comité de programme de la conférence Concept Lattices and their Applications (CLA)
- Organisation des séances de travail pour doctorants du séminaire MALOTEC au LORIA en 2019

Participation à des groupes de recherche

- Participation au PEPS HYDRATA (HYpergraphes et Datamining : AlgoriThmes et Analyses probabilistes)

Séjours de recherche

- Une semaine au LIRMM (Montpellier), invité par la Professeure Marianne Huchard en janvier 2018
- Une semaine à Taltech (Tallinn, Estonie), invité par le Professeur Sadok Ben Yahia en février 2020

Travaux et projet de recherche

Travaux de recherche

Depuis la formalisation des principes de **représentation des connaissances** dans les années 1990 et les débuts du web sémantique autour des années 2000, les ontologies ont pris une place de premier plan dans la représentation des connaissances et le raisonnement. Bien que les graphes de connaissance soient actuellement au centre de l'attention, construire des ontologies reste une tâche de toute première importance. Liées à la représentation des connaissances mais développées en parallèle, les méthodes de **découverte de connaissances** visent à extraire automatiquement de données des motifs pouvant être considérés comme des connaissances. La représentation et la découverte de connaissances peuvent et doivent interagir mais leur interfaçage n'a pas encore été réalisé de manière satisfaisante. C'est dans ce cadre que mes travaux de recherche ont débuté : comment la découverte de connaissances peut-elle être utilisée pour construire des ontologies semi-automatiquement ?

Mes travaux ont principalement porté sur la découverte de connaissances prenant la forme de **classes** et de **règles**. Les classes sont des paires composées d'un ensemble d'objets (l'extension) possédant des propriétés en commun (l'intension). Ainsi, l'ensemble des clients ayant acheté du sucre, du beurre et de la farine forme une classe pouvant être interprété comme celle des gens qui préparent un gâteau. Les règles sont des régularités dans les descriptions des objets : si un appartement consomme moins de 100kWh/an, alors il est inoccupé. Elles permettent de hiérarchiser les classes. Les classes et les règles extraites des données peuvent être utilisées pour construire automatiquement les concepts et axiomes d'ontologies [14,27]. Le coeur de mes travaux concerne l'identification et le calcul de **représentations condensées**, c'est à dire de sous-ensembles de classes et de règles résumant sans perte toute l'information contenue dans les autres. Dans le cadre de mes travaux de thèse, nous avons étudié l'extraction de classes et de règles dans le type de données le plus simple, les données tabulaires binaires pouvant être représentées par un tableau de croix. Par la suite, lors de mes séjours post-doctoraux, nous avons abordé les mêmes problématiques dans des formes de données plus complexes, les **données multidimensionnelles** et les **données d'origine biologique**. Plus récemment, nous avons appliqué ces travaux à l'**explication d'approches d'intelligence artificielle** [5] et la **découverte et la représentation de relations causales** [24].

Les règles pouvant être extraites d'un jeu de données sont trop nombreuses pour être utiles à un analyste ou même manipulées par une machine. Un problème classique est la recherche d'un sous-ensemble de règles, appelé *base de règles*, qui résume l'information contenue dans toutes les autres. Pour les règles certaines, les implications, dont la confiance est égale à 1, la plus petite de ces bases est la *base canonique*. Bien qu'elle ait été identifiée en 1987, toutes les questions liées à son calcul n'avaient pas encore trouvé de réponse en 2012. La principale contribution de mes travaux de thèse prend la forme de deux algorithmes [3,4,12,13] pour l'extraction de la base canonique et son utilisation lors de la construction des axiomes d'une ontologie représentant les connaissances contenues dans le jeu de données.

a	b	c	d	e	
×	×				$\{a\} \to \{b\}$
	×	×	×		$\{b,c\} \to \{b,c,d\}$
	×		×	×	$\{c,d\} \to \{b,c,d\}$
		×		×	$\{b,e\} \to \{b,d,e\}$
			×	×	$\{a,b,d\} \to \{a,b,c,d,e\}$
					$\{b,c,d,e\} \to \{a,b,c,d,e\}$

Figure 1: Un jeu de données binaires bidimensionnelles et sa base canonique.

Les données multidimensionnelles mettent en relation $n \geq 3$ entités de natures différentes et permettent de décrire la réalité de façon plus riche : des clients ayant acheté des produits dans des magasins, la consommation énergétique d'appartements durant différentes saisons... Elles prennent généralement la forme d'une relation n-aire et font l'objet de plus en plus d'attention. Elles sont notamment liées à internet. Ainsi, la base de données de films IMDb implique, a minima, des films, des acteurs, des producteurs, des scénaristes, des genres et des notes. Les données RDF entrent aussi dans cette catégorie. Les données multidimensionnelles comportent encore plus de classes et de règles que les données

bidimensionnelles et nous avons donc d'autant plus besoin de bases de règles et de méthodes efficaces pour l'extraction des classes. Durant mes séjours post-doctoraux, mon expérience de thèse a été mise à profit pour identifier les premières bases de règles certaines (implications) [1] et approximatives (règles d'association) [19,22] dans les données multidimensionnelles afin de pouvoir exploiter les connaissances que ces données contiennent. De la même façon, nous avons étudié le nombre maximal de classes pouvant coexister dans un tel jeu de données [17,23] ainsi que la généralisation multidimensionnelle de la notion de classe introductrice [7] qui permet de condenser l'information contenue dans les classes et donc de ne pas avoir à toutes les extraire pour analyser le contenu des données.

	Lait	Pain	Couches	Bière	Lait	Pain	Couches	Bière	Lait	Pain	Couches	Bière
c_1	×	×			×		×				×	
c_2			×	×	×	×			×		×	
c_3	×	×	×	×		×	×	×	×	×		
c_4	×	×		×	×			×			×	×
	Hiver				Printemps				Été			

Figure 2: Un jeu de données binaires tridimensionnelles dans lequel des *clients* (c_1 à c_4) achètent des *produits* (Lait, Pain, Couches, Bières) pendant différentes *saisons* (Hiver, Printemps, Été).

Les données d'origine biologique peuvent prendre toutes les formes mais présentent des défis particuliers : des données rares ou extrêmement volumineuses, avec de nombreuses variables et, surtout, nécessitant des connaissances du domaine pour être appréhendées. Le projet de mon premier post-doctorat, réalisé au LIMOS de l'Université Clermont Auvergne, traitait du regroupement par espèces de millions d'organismes en fonction des similarités de leurs séquences d'ARN. Les espèces peuvent être considérées comme des classes d'organismes extraites automatiquement des données. Nous avons conçu un algorithme [2] permettant le clustering de ces grands volumes de séquences tout en fournissant à l'utilisateur une évaluation de la confiance qu'il peut accorder aux résultats et implémenté l'approche au sein du logiciel SCLUST[1]. Le projet de mon post-doctorat dans l'équipe-projet Orpailleur, financé par le projet GEENAGE[2], concernait les données biomédicales. Dans un jeu de données décrivant des patients et leur état de santé, les médecins souhaiteraient découvrir parmi les variables du jeu de données quelles sont celles qui sont plutôt discriminantes et celles qui sont plutôt prédictives de l'apparition de la maladie. Nous avons développé une nouvelle approche [15] permettant la sélection et l'identification de variables prédictives de l'apparition des maladies et discriminantes entre maladie et bonne santé. Cette approche combine des classifieurs supervisés avec des algorithmes de fouille de données non supervisée et des méthodes d'aide à la décision multicritères.

Pendant mon séjour dans l'équipe Orpailleur, nous avons travaillé à mettre en relation la découverte de connaissances et deux thèmes qui sont aujourd'hui au coeur de la recherche en intelligence artificielle : l'explicabilité et la causalité. Les systèmes d'intelligence artificielle, tels que les réseaux de neurones, sont des boites noires dont le comportement est inaccessible à leurs utilisateurs qui doivent pourtant faire confiance à leurs résultats. L'explication soit du fonctionnement de ces boites noires, soit directement de leurs résultats, est un des sujets les plus étudiés actuellement. Ces explications se situent le plus souvent dans un cadre d'apprentissage supervisé et prennent la forme de scores numériques. Nous avons proposé une approche [5] qui produit une explication symbolique du résultat de processus de décision multicritères – le calcul de fronts de Pareto – sous la forme d'un ensemble de connaissances. La découverte de relations causales entre les variables de jeux de données est un autre grand problème qui fait l'objet de beaucoup d'attention. La plupart des travaux à ce sujet s'intéressent à la découverte et la représentation de relations causales univariées, c'est à dire entre deux variables x et y. Moins de travaux existent sur la découverte de relations causales multivariées, entre deux ensembles X et Y, et quasiment aucun sur la représentation de la structure que forment ces relations. Nous avons proposé l'utilisation des algorithmes et structures de l'analyse formelle de concepts pour découvrir et représenter l'ensemble de ces relations causales multivariées [24].

Je travaille actuellement toujours au sein de l'équipe Orpailleur sur le projet ANR ELKER[3]. Il s'agit de découvrir des clés de liage permettant d'identifier des entités, relations et classes communes à deux jeux de données RDF mais décrites différemment [16,25,26].

[1] http://fc.isima.fr/ bazina/Expe.html
[2] lue.univ-lorraine.fr/en/heatlh-and-fight-against-ageing/impact-geenage
[3] https://project.inria.fr/elker/

Projet de recherche et intégration au LIRMM

Les connaissances ne sont pas uniquement, pour l'informatique, des informations à découvrir, structurer et représenter. Elles sont des données à part entière qu'il convient d'exploiter, soit pour y découvrir encore plus de connaissances, soit pour guider la recherche d'information. Cependant, les connaissances forment des jeux de données complexes, qui ne peuvent pas encore être parfaitement traités par les algorithmes dont nous disposons.

Mes travaux de recherche se sont toujours dirigés vers la découverte de connaissances dans des données toujours plus complexes. Je souhaite poursuivre sur cette trajectoire en m'intéressant à la découverte de connaissances dans, avec et pour les graphes de connaissances. Ces graphes sont centraux dans le web sémantique et sont notamment utilisés par les grands moteurs de recherche, Google, Yahoo, et Bing. Leur structure leur permet d'être stockés, visualisés et utilisés pour raisonner efficacement. Ce projet de recherche vise à traiter les graphes de connaissances à la fois comme des *données* (entrée), des *connaissances à découvrir* (sortie) et des *connaissances à exploiter* (connaissances du domaine) dans des processus de découverte de connaissances.

Les graphes de connaissances prennent le plus souvent la forme de données au format RDF, c'est à dire un ensemble de triplets $(Sujet, Prédicat, Objet)$. Ces données peuvent être vues comme des données binaires tridimensionnelles. Je m'appuyerai donc sur mes travaux passés [1,8,21] sur les données multidimensionnelles pour aborder la fouille de motifs dans les graphes de connaissances classiques avant d'évoluer vers les graphes plus riches (informations supplémentaires sur les arêtes, incomplétude, évolution temporelle...). Les graphes de connaissances peuvent aussi être utilisés pour raisonner et donc être vus comme un ensemble de règles. J'étendrai donc mes travaux sur le calcul de règles pour apprendre ou compléter automatiquement des graphes de connaissances, ou encore utiliser les graphes comme connaissances du domaine. Une telle utilisation pourrait être la génération d'explications symboliques, sous la forme de sous-graphes de connaissances, du contenu de jeux de données ou de résultats d'approches d'intelligence artificielle.

Au LIRMM, l'équipe FADO s'intéresse aux données du web des données. Elle travaille notamment sur des problématiques de liage de données RDF[4], d'alignement d'ontologies[5] et de fouille de motifs dans les graphes[6]. Les algorithmes que je développerai pour la découverte de connaissance dans les données RDF pourront s'appliquer à l'alignement d'ontologies et au liage de données, ce que j'ai déjà commencé à faire. De plus, je pourrai étendre ces travaux aux graphes de propriétés[7] qui sont traités dans l'équipe. J'estime donc pouvoir m'intégrer à tous les axes de recherche de l'équipe FADO. Je viendrai les renforcer en y apportant ma propre expertise de la fouille de données binaires multidimensionnelles et, en particulier, de l'analyse formelle de concepts qui a déjà été utilisée dans certains travaux de l'équipe.

Je pourrai de plus maintenir des collaborations avec certaines des autres équipes du LIRMM telles qu'ADVANSE, sur la fouille de motifs dans des données complexes, ou GraphiK, sur la représentation des connaissances. J'ai aussi eu la plaisir d'une collaboration avec Marianne Huchard de l'équipe MAREL, que je pourrai poursuivre. Enfin, j'ai un intérêt pour les structures combinatoires, ce qui me permettra de discuter avec l'équipe AlGCo.

[4]Manel Achichi, Zohra Bellahsene, Mohamed Ben Ellefi, Konstantin Todorov. Linking and disambiguating entities across heterogeneous RDF graphs. Journal of Web Semantics 55 (2019): 108-121.

[5]Konstantin Todorov, Céline Hudelot, Adrian Popescu, Peter Geibel. Fuzzy ontology alignment using background knowledge. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 22.01 (2014): 75-112.

[6]Shah, Faaiz, Arnaud Castelltort, and Anne Laurent. "Handling missing values for mining gradual patterns from NoSQL graph databases." Future Generation Computer Systems 111 (2020): 523-538.

[7]Faaiz Shah, Arnaud Castelltort, Anne Laurent. Extracting Fuzzy Gradual Patterns from Property Graphs. 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). IEEE, 2019.

Liste des travaux adressés en cas de convocation à l'audition

---

- Alexandre Bazin. "*On Implication Bases in n-Lattices.*" Discrete Applied Mathematics (DAM) Volume 273, p. 21-29, 2020. (voir)
- Alexandre Bazin, Jessie Carbonnel, Marianne Huchard, Giacomo Kahn, Priscilla Keip, Amirouche Ouzerdine. "*On-demand Relational Concept Analysis.*" International Conference on Formal Concept Analysis (ICFCA), p. 155-172, 2019. (voir)
- Alexandre Bazin et Aurélie Bertaux. "*k-Partite Graphs as Contexts.*" Concept Lattices and Applications (CLA), p. 59-67, 2018. (pdf)