Université Claude Bernard Lyon 1

Advancement Report 2024 (D1)

Amedeo Pachera
supervised by Angela Bonifati (UCBL) and Andrea Mauri (UCBL)

Research Subject

This thesis focuses on the integration of artificial and human intelligence to understand how to design data-intensive systems more scalable, efficient, effective, and sustainable (in terms of impact on society). This will include the investigation of how to integrate different kinds of data (from social media, sensors, lab studies, clinical trials, interviews, etc..), how to embed technology in a usually human-driven process, and how to provide trustworthy human-machine interactions in data-intensive applications. The developed methods will combine and integrate different HCI and data management areas including crowdsourcing, gamification, user-generated content analysis on one side, and algorithms for data quality, integration and querying on the other.

Consequently, the thesis develops across two research directions :

- Enhancing algorithm performances using Human-in-the-loop techniques.
- Studying and proposing quantitative metrics to evaluate the human aspect/presence in human-machine applications.

Human-In-The-Loop

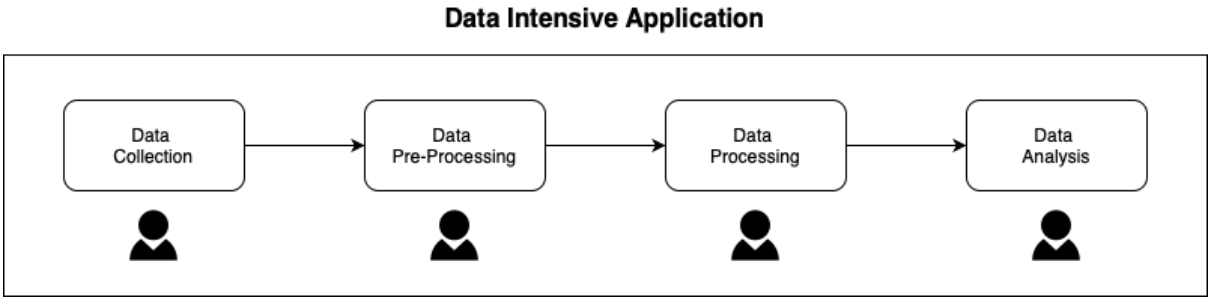
Human-in-the-loop (HIL) methods integrate human insight or judgments into machine-driven processes, particularly when algorithms alone may struggle due to the necessity of specific domain knowledge [1,2] or when existing algorithms fall short [3,4]. While involving humans can be resource intensive in terms of time and effort, recent efforts have focused on optimizing the use of crowdsourcing [5,6,7], developing robust experimental methodologies [8], applying it across various domains such as image classification [9], transcription [2,10], natural language processing [11], and healthcare [12].

In the data management community, crowdsourcing has proven valuable for answering queries beyond the capabilities of computers alone, with a focus on minimizing the cost of obtaining such answers [13,14,15,16]. In the realm of graph data, Cong et al. [17] addressed the Interactive Graph Search (IGS) problem, involving human intelligence to locate a target

node and developing algorithms that minimize the number of questions posed to users. Human-in-the-loop approaches have been applied to data cleaning [18,19] to ensure better quality fixes with respect to automated repairs.

In recent years, HIL also found applications in machine learning [20] not only aiming towards more accurate algorithms or to obtain the desired accuracy faster, but also to make humans more effective and more efficient. Depending on who is in control of the learning process, we can identify different approaches : Active learning (AL) [21], in which the system remains in control of the learning process and treats humans as oracles to annotate unlabeled data, Interactive machine learning (IML) [22], in which there is a closer interaction between users and learning systems, with people interactively supplying information in a more focused, frequent, and incremental way compared to traditional machine learning, Machine teaching (MT) [23,24], where human domain experts have control over the learning process by delimiting the knowledge that they intend to transfer to the machine learning model.

In this thesis, we focus on exploiting HIL techniques for developing Data-Intensive applications. In particular, we identified four main components in the data pipeline shown in Figure 1 in which the human could enhance the existing techniques resulting in an overall improvement for the application.



Evaluating the human aspects in the Human-in-the-loop applications.

This thesis focuses on every aspect of the data pipeline and aims towards developing a framework for evaluating HIL applications in both the algorithm and human aspects.
While in the Data Collection, Data cleaning and Machine learning domains researchers use quantitative metrics that evaluate the goodness of their proposed models (e.g. Accuracy, F1, etc.), in the Human Centered AI (HCAI) domain the focus is towards qualitative metrics such as explainability, contestability, thrust, interpretability and so on. Big companies and institutions like Google, Microsoft and the European Union have developed guidelines for HCAI evaluations [25] but most of the research focuses only on qualitative analysis with respect to quantitative analysis [26].
In this thesis, we will investigate metrics for the quantitative analysis of the human aspects of Human-Machine applications and applying the principles and guidelines coming from the HCAI domain to the HIL domain.

Published Articles

- [Extended Abstract] Amedeo Pachera, Angela Bonifati, Andrea Mauri. **Towards User-Centric Graph Repairs SEAGraph 2024 (ICDE 2024 workshop).**

Paper Under Submission

- Amedeo Pachera, Angela Bonifati, Andrea Mauri. **Collaborative Graph Repair Under Denial Constraints.**
  Our first work focused on the data pre-processing component of the data pipeline described above. In particular, we studied interactive graph repair. In a prior work [27], HIL techniques have been applied to repair labeled graphs under neighborhood constraints. We extended this work using property graphs and denial constraints. With more expressiveness in the data model, the challenge of repairing a graph with automated repair becomes more tough because. A violation indeed, may be solved with multiple transformations and Human Knowledge is essential to apply the right repair and to ensure the quality of the data. We propose a Question-Answer-Repair framework that allows multiple users to repair a property graph together with an assignment algorithm that allows collaborations among multiple violations. The approach is based on a violation dependency graph, an hypergraph with violations as hypernodes and edges representing dependencies (overlaps between the subgraphs included in the hypernodes).

Current Work

- Amedeo Pachera, Angela Bonifati, Stefania Dumbrava, Andrea Mauri.
  Understanding user errors in graph query formulation.
  In this study, we engaged 60 students in an exercise to write 15 Cypher queries on a pangenome dataset. Our objective was to identify common errors and understand their underlying causes. To achieve this, we propose an extended taxonomy with respect to [28] specifically designed to classify the types of errors made in Cypher queries. This taxonomy provides a comprehensive framework for categorizing and analyzing the errors, enabling us to pinpoint specific areas where students struggle. To further investigate the causes of these errors, we conducted surveys and live interviews with the participants. The surveys included questions designed to probe the students' understanding of Cypher and their approach to query writing. The live interviews provided an opportunity for a deeper dialogue, allowing us to explore the students' thought processes and identify any gaps in their knowledge or misconceptions.

Future Work

- **Graph Repair With LLMs** : Our research on interactive graph repairs has identified a significant limitation: users often lack the IT domain knowledge required to work effectively with graphs. This presents a major challenge: how can we enable users to repair graphs without requiring specialized knowledge? To address this, we propose investigating the use of Large Language Models (LLMs) to facilitate graph repairs. LLMs have demonstrated remarkable capabilities in understanding and generating human-like text. We aim to leverage these capabilities in two aspects: converting graphs to text and vice versa. By translating graph structures into natural language descriptions and subsequently converting these descriptions back into graph structures, we can enable users to interact with and repair graphs using plain language. To evaluate the effectiveness of LLMs in this context, we propose an ablation study comparing the performance of LLMs and users in three key tasks:
  Graph-to-Text Translation: Assessing the accuracy and comprehensibility of the text generated from graph data by both LLMs and human users.
  Text Repair: Comparing the quality and precision of repairs made by users directly on the graph versus those made on the text representation.
  Text-to-Graph Translation: Evaluating the fidelity of the graph structure reconstructed from the repaired text by LLMs and users.
  This ablation study will help us understand the strengths and limitations of using LLMs for graph repairs and identify areas where human intervention remains crucial.

Other Activities

Attended Talks

- “Building Data Management Systems for Precision Medicine: Lessons Learnt from Five National Flagship Projects.” Irini Fundulaki (Institute of Computer Science, FORTH)
- “ Data Science for Social Goods: STAR Lab’s Experience” . Reynold Cheng (Hong Kong University)
- “AI-assisted Knowledge Navigation“. Akhil Arora (EPFL)

Conferences, Workshops and Summer School

- BDA 2023, Communauté Francophone en Gestion de données : Principes, Technologies et Applications, Montpellier, 26-29/10/2023
- ICDE 2024, IEEE International Conference on Data Engineering, Utrecht, 13-17/10/2024
- CIX’24 : 8th Summer School on Computational Interaction, ACM Europe School, 3-7/07/2024

Courses (Doctoral School)

- MOOC / Intégrité scientifique dans les métiers de la recherche (15 h)
- [In Progress] MOOC / Cours de Français / French course : Vivre en France B1 (90 h)

References

[1] : Jasper Oosterman, Archana Nottamkandath, Chris Dijkshoorn, Alessandro Bozzon, Geert-Jan Houben, and Lora Aroyo. 2014. Crowdsourcing knowledge-intensive tasks in cultural heritage. In Proceedings of the 2014 ACM conference on Web science. 267–268.
[2] : IP Samiotis, S Qiu, A Mauri, CCS Liem, C Lofi, and A Bozzon. 2020. Micro- task crowdsourcing for music score Transcriptions: an experiment with error detection. In 21st International Society for Music Information Retrieval Conference.
[3] : Akansha Bhardwaj, Jie Yang, and Philippe Cudré-Mauroux. 2022. Human-in- the-Loop Rule Discovery for Micropost Event Detection. IEEE Transactions on Knowledge and Data Engineering (2022).
[4] : A Mauri and A Bozzon. 2021. Towards a human in the loop approach to preserve privacy in images. In CEUR Workshop Proceedings, Vol. 2947.
[5] : Shahzad Sarwar Bhatti, Xiaofeng Gao, and Guihai Chen. 2020. General frame- work, opportunities and challenges for crowdsourcing techniques: A Compre- hensive survey. Journal of Systems and Software 167 (2020), 110611. https: //doi.org/10.1016/j.jss.2020.110611
[6] : David R. Karger, Sewoong Oh, and Devavrat Shah. 2013. Efficient Crowdsourcing for Multi-Class Labeling. SIGMETRICS Perform. Eval. Rev. 41, 1 (jun 2013), 81–92. https://doi.org/10.1145/2494232.2465761
[7] : Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E Kraut. 2011. Crowd- forge: Crowdsourcing complex work. In Proceedings of the 24th annual ACM symposium on User interface software and technology. 43–52.
[8] : Jorge Ramírez, Burcu Sayin, Marcos Baez, Fabio Casati, Luca Cernuzzi, Boualem Benatallah, and Gianluca Demartini. 2021. On the State of Reporting in Crowd- sourcing Experiments and a Checklist to Aid Current Practices. Proc. ACM Hum.-Comput. Interact. 5, CSCW2, Article 387 (oct 2021), 34 pages. https: //doi.org/10.1145/3479531
[9] : Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2017. Imagenet classifi- cation with deep convolutional neural networks. Commun. ACM 60, 6 (2017), 84–90.
[10] : Luis Von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum. 2008. recaptcha: Human-based character recognition via web security measures. Science 321, 5895 (2008), 1465–1468.
[11] : Oana Inel, Khalid Khamkham, Tatiana Cristea, Anca Dumitrache, Arne Rut- jes, Jelle van der Ploeg, Lukasz Romaszko, Lora Aroyo, and Robert-Jan Sips. 2014. Crowdtruth: Machine-human computation framework for harnessing dis- agreement in gathering annotated data. In International semantic web conference. Springer, 486–504.
[12] : Kerri Wazny. 2018. Applications of crowdsourcing in health: an overview. Journal of global health 8, 1 (2018).
[13] : Susan B. Davidson, Sanjeev Khanna, Tova Milo, and Sudeepa Roy. 2013. Us- ing the Crowd for Top-k and Group-by Queries. In Proceedings of the 16th International Conference on Database Theory (Genoa, Italy) (ICDT '13). As- sociation for Computing Machinery, New York, NY, USA, 225–236. https: //doi.org/10.1145/2448496.2448524
[14] : Michael J. Franklin, Donald Kossmann, Tim Kraska, Sukriti Ramesh, and Reynold Xin. 2011. CrowdDB: Answering Queries with Crowdsourcing. In Proceedings of the 2011 ACM SIGMOD

International Conference on Management of Data (Athens, Greece) (SIGMOD '11). Association for Computing Machinery, New York, NY, USA, 61–72. https://doi.org/10.1145/1989323.1989331
[15] : Adam Marcus, Eugene Wu, David R Karger, Samuel Madden, and Robert C Miller. 2011. Crowdsourced databases: Query processing with people. Cidr.
[16] : Petros Venetis, Hector Garcia-Molina, Kerui Huang, and Neoklis Polyzotis. 2012. Max Algorithms in Crowdsourcing Environments. In Proceedings of the 21st International Conference on World Wide Web (Lyon, France) (WWW '12). Association for Computing Machinery, New York, NY, USA, 989–998. https://doi.org/10.1145/2187836.2187969
[17] : Qianhao Cong, Jing Tang, Yuming Huang, Lei Chen, and Yeow Meng Chee. 2022. Cost-Effective Algorithms for Average-Case Interactive Graph Search. In 38th IEEE International Conference on Data Engineering, ICDE 2022, Kuala Lumpur, Malaysia,May9-12,2022.IEEE,1152–1165. https://doi.org/10.1109/ICDE53745. 2022.00091
[18] : Jian He, Enzo Veltri, Donatello Santoro, Guoliang Li, Giansalvatore Mecca, Paolo Papotti, and Nan Tang. 2016. Interactive and Deterministic Data Cleaning. In Proceedings of the 2016 International Conference on Management of Data (San Francisco, California, USA) (SIGMOD '16). Association for Computing Machinery, New York, NY, USA, 893–907. https://doi.org/10.1145/2882903.2915242
[19] : Mohamed Yakout, Ahmed K. Elmagarmid, Jennifer Neville, Mourad Ouzzani, and Ihab F. Ilyas. 2011. Guided data repair. Proc. VLDB Endow. 4, 5 (feb 2011), 279–289. https://doi.org/10.14778/1952376.1952378
[20] : Jiang, Liu, Shixia Liu, e Changjian Chen. «Recent Research Advances on Interactive Machine Learning». Journal of Visualization 22, fasc. 2 (aprile 2019): 401–17. https://doi.org/10.1007/s12650-018-0531-1.
[21] : Settles B (2009) Active learning literature survey. Tech. rep., University of Wisconsin-Madison. Department of Computer Sciences, https://minds.wisconsin.edu/handle/1793/60660
[22] : Amershi S, Cakmak M, Knox WB et al (2014) Power to the people: the role of humans in interactive machine learning. AI Magazine 35(4):105–120. https://doi.org/10.1609/aimag.v35i4.2513
[23] : Simard PY, Amershi S, Chickering DM et al (2017) Machine teaching: A new paradigm for building machine learning systems. arXiv e-prints arxiv:1707.06742
[24] : Ramos G, Meek C, Simard P et al (2020) Interactive machine teaching: a human-centered approach to building machine-learned models. Hum Comput Interact 35(5–6):413–451. https://doi.org/10.1080/ 07370024.2020.1734931
[25] : Bingley, William J., Caitlin Curtis, Steven Lockey, Alina Bialkowski, Nicole Gillespie, S. Alexander Haslam, Ryan K.L. Ko, Niklas Steffens, Janet Wiles, e Peter Worthy. «Where Is the Human in Human-Centered AI? Insights from Developer Priorities and User Experiences». Computers in Human Behavior 141 (aprile 2023): 107617. https://doi.org/10.1016/j.chb.2022.107617.
[26] : Sperrle, F., M. El-Assady, G. Guo, R. Borgo, D. Horng Chau, A. Endert, e D. Keim. «A Survey of Human-Centered Evaluations in Human-Centered Machine Learning». Computer Graphics Forum 40, fasc. 3 (giugno 2021): 543–68. https://doi.org/10.1111/cgf.14329.
[27] : Paul Juillard, Angela Bonifati, and Andrea Mauri. 2024. Interactive Graph Repairs for Neighborhood Constraints. In Proceedings 27th International Conference on Extending Database Technology (EDBT 2024) Paestum, Italy, March 25 - March 28. OpenProceedings.org, 2:175–2:187.
[28] : Toni Taipalus, Mikko Siponen, and Tero Vartiainen. 2018. Errors and Complications in SQL Query Formulation. ACM Trans. Comput. Educ. 18, 3, Article 15 (aug 2018), 29 pages. https://doi.org/10.1145/3231712

This thesis focuses on combining artificial and human intelligence to make data-intensive systems more scalable, efficient, effective, and sustainable, considering their societal impact. This includes studying how to integrate different types of data (like from social media, sensors, lab studies, clinical trials, interviews, etc.), how to incorporate technology into processes typically driven by humans, and how to ensure trustworthy interactions between humans and machines in data-heavy applications. The methods developed will bring together different areas of human-computer interaction (HCI) and data management, including crowdsourcing, gamification, user-generated content analysis, and algorithms for data quality, integration, and querying.

In his first year, M. Pachera focused on the pre-processing and analysis stages. For pre-processing, he developed and tested a human-in-the-loop framework to make interactive graph repairs. This work led to a workshop publication and a submission to a top-tier conference, which is currently under review.

For the analysis part, he examined common errors and their root causes in Cypher queries. He created an extended taxonomy to classify these errors and conducted surveys and interviews to understand students' thought processes, identifying knowledge gaps and misconceptions.

Next year, M. Pachera will focus on user modeling for graph repair to assign repairs to the most suitable users. He will also explore other types of data beyond graphs, such as time series.

Villeurbanne 20/06/2024 Angela Bonifati

Version finale - juillet 2023

INFOMATHS
INFORMATIQUE
MATHÉMATIQUES
UNIVERSITÉ DE LYON

ÉCOLE DOCTORALE 512

Année universitaire
2024-2025

Convention Individuelle de Formation Doctorale

Vu l'arrêté du 25 mai 2016 modifié en août 2022 fixant le cadre national de la formation et les modalités conduisant à la délivrance du diplôme national du doctorat,
Vu la charte du doctorat commune aux établissements du site Lyon - St Etienne, membres et associés de la COMUE Université de Lyon.

Entre :

M. Pachera Amedeo, ci-après dénommé le doctorant ou la doctorante, d'une part

Et

Professeur Bonifati Angela, ci-après dénommé le directeur ou la directrice de thèse, d'autre part,

Considérant que :

- L'article 10 de l'arrêté du 25 mai 2016 modifié prévoit l'élaboration d'une convention individuelle de formation en application de la charte du doctorat ;
- L'établissement d'inscription est garant de sa mise en œuvre à travers les écoles doctorales.

Il est convenu ce qui suit :

Article 1. La doctorante ou le doctorant est inscrit :

Etablissement d'inscription : Université de Lyon 1 Claude Bernard
École doctorale : 512 INFOMATHS, dirigée par Monsieur Hamamache Kheddouci
Intitulé du doctorat : Human-Machine Intelligence Integration for Effective Healthcare Data-Intensive Applications
Sujet de la thèse : data management, human in the loop
Unité de recherche : UMR 5205, Laboratoire d'InfoRmatique en Image et Systèmes d'information (LIRIS), dirigée par Jean-Marc Petit

Directeur ou directrice de thèse : Professeur Angela Bonifati
Co-directeur ou co-directrice de thèse (le cas échéant) : Professeur Andrea Mauri
Co-directeur ou co-directrice de thèse en entreprise (le cas échéant) : [civilité, nom, prénom]
Adresse mail du doctorant ou de la doctorante : amedeopachera@gmail.com

Dans le cadre d'une cotutelle internationale de thèse
Etablissement partenaire : [Dénomination de l'établissement partenaire, ville, pays]
Unité de recherche : [libellé], dirigée par [nom + prénom du directeur ou directrice]
Directeur ou directrice de thèse à l'étranger : [civilité, nom, prénom]

Article 2. Statut du doctorant ou de la doctorante

Contractuel oui x non □	Si oui, type de contrat : x Contrat Doctoral □ Contrat Durée Déterminée □ Contrat CIFRE □ Contrat Doctoral de droit privé
Boursier (d'un organisme étranger) oui □ non x	
Exerçant une activité salariée (statut professionnel) oui □ non x	
Autre financement oui □ non x	
Sans financement oui □ non x	

Article 3. Rythme de la thèse

Aux termes de l'Article 14 de l'arrêté du 25 mai 2016, *la préparation du doctorat s'effectue en règle générale en 3 ans en équivalent temps plein consacré à la recherche. Dans les autres cas, la durée de préparation peut-être au plus de 6 ans*

Le doctorant ou la doctorante réalise sa thèse à :

X **Temps complet**	□ Activités complémentaires au contrat doctoral (le cas échéant)
□ **Temps partiel** *(au minimum, 50% du temps doit être consacré à la thèse)* **Quotité :** **Si temps partiel, préciser le statut professionnel du doctorant ou de la doctorante :**	

Article 4. Description du projet de thèse

Limité à 2 pages maximum, le descriptif du projet de thèse présentera le contexte scientifique et les principaux objectifs de la thèse. Il doit être joint en annexe de la présente convention.

Article 5. Encadrement et suivi de la thèse

Je suis logé dans le bureau 11.008 du bâtiment Nautibus. Je fais partie du laboratoire LIRIS dans l'équipe BD dirigée par Angela Bonifati. Mes superviseurs sont les Professeurs Angela Bonifati et Andrea Mauri. Nous suivons l'avancement de la thèse à raison d'une réunion par semaine où nous sommes tous les trois présents, tandis que chaque jour je rencontre Andrea Mauri pour des questions ou des doutes.
Dans mon bureau, j'ai le moniteur offert par le laboratoire et l'ordinateur portable offert par mon directeur de thèse. J'accède au bâtiment et au laboratoire grâce à mon badge de l'Université Claude Bernard Lyon 1.

Article 6. Calendrier prévisionnel du projet de recherche

Année scolaire 2024-2025 :
- poursuivre le développement de la thèse et la publication d'articles

Année scolaire 2025-2026 :
- poursuivre le développement de la thèse et la publication d'articles
- Mai - septembre 2026 : rédaction de la thèse
- Octobre - decembre 2026 : soutenance de thèse

Article 7. Conditions matérielles de réalisation du projet de recherche

Mon doctorat est financé par Data4Health ANR project CPJ Andrea Mauri. Grâce à ce financement, j'ai reçu l'équipement nécessaire à mon travail et j'ai été payée pour mes missions à la conférence ICDE 2024 à Utrecht et à l'école d'été CIX 2024 au Luxembourg.

Article 8. Projet professionnel

Mon projet professionnel est de continuer à travailler dans le domaine de la recherche, que ce soit dans un environnement public ou privé.

Mon projet professionnel est de poursuivre dans le domaine de la recherche, que ce soit dans un environnement public ou privé. Ma thèse me prépare à approfondir mes compétences dans le domaine de la recherche. Le thème de ma thèse est l'interaction entre l'homme et la machine, ce qui ouvre de nombreuses pistes de recherche et de projets futurs.

Article 9. Parcours individuel prévisionnel de formation en lien avec ce projet

■ Intégrité scientifique dans les métiers de la recherche - 21 avril 2024-30 avril 2024 - *15 heures*
La formation participe à l'objectif suivant :former à l'éthique de la recherche et à l'intégrité scientifique
■ Tutorial on LLMs: Principles and Practice - 14 mai 2024 - *1 heure*
La formation participe à l'objectif suivant :être directement utile pour la réalisation des travaux personnels de recherche
■ An Interactive Dive (Tutorial) into Time-Series Anomaly Detection - 17 mai 2024 - *1 heure*
La formation participe à l'objectif suivant :être directement utile pour la réalisation des travaux personnels de recherche
■ 8th Summer School on Computational Interaction, ACM Europe School - 03 juin 2024-7 juin 2024 - *25 heures*
La formation participe à l'objectif suivant :être directement utile pour la réalisation des travaux personnels de recherche

Article 10. Objectifs de valorisation des travaux de recherche

Amedeo Pachera, Angela Bonifati, Andrea Mauri 2024. Towards User-Centric Graph Repairs, *SEAGraph 2024* , 2 pages, **https://seagraph.day**

Article 11. Intégrité scientifique

Les parties s'engagent à respecter, tout au long des travaux de recherche, les principes et exigences de l'intégrité scientifique.

Fait à Villeurbanne le 12/06/2024

en 1 exemplaire original,

Signatures (avec visa éventuel du Directeur ou de la directrice de laboratoire) :	
Directeur ou directrice de thèse 	Doctorant ou doctorante 

Co-encadrant ou co-directeur	Référent en entreprise

Article 12. Durée et résiliation

La convention est conclue pour la durée de la thèse. Elle peut être modifiée en tant que de besoin, lors des réinscriptions par accord signé entre les parties.

Révisée le :	
Signatures (avec visa éventuel du Directeur ou Directrice de laboratoire) :	
Directeur ou directrice de thèse	Doctorant ou doctorante
Co-encadrant ou co-directeur	Référent en entreprise

IMPORTANT : *Le document, signé dans les 6 mois suivants la 1ère inscription, doit être intégré au dossier ADUM du doctorant ou de la doctorante.*

Description du projet de thèse

Amedeo Pachera - 2024

This thesis focuses on the integration of artificial and human intelligence to understand how to design data-intensive systems more scalable, efficient, effective, and sustainable (in terms of impact on society). This will include the investigation of how to integrate different kinds of data (from social media, sensors, lab studies, clinical trials, interviews, etc..), how to embed technology in a usually human-driven process, and how to provide trustworthy human-machine interactions in data-intensive applications. The developed methods will combine and integrate different HCI and data management areas including crowdsourcing, gamification, user-generated content analysis on one side, and algorithms for data quality, integration and querying on the other.

Consequently, the thesis develops across two research directions :

- Enhancing algorithm performances using Human-in-the-loop techniques.
- Studying and proposing quantitative metrics to evaluate the human aspect/presence in human-machine applications.

Human-in-the-loop (HIL) methods integrate human insight or judgments into machine-driven processes, particularly when algorithms alone may struggle due to the necessity of specific domain knowledge or when existing algorithms fall short. While involving humans can be resource intensive in terms of time and effort, recent efforts have focused on optimizing the use of crowdsourcing, developing robust experimental methodologies, applying it across various domains such as image classification, transcription, natural language processing, and healthcare.

In the data management community, crowdsourcing has proven valuable for answering queries beyond the capabilities of computers alone, with a focus on minimizing the cost of obtaining such answers.In the realm of graph data, the Interactive Graph Search (IGS) problem, involving human intelligence to locate a target node and developing algorithms that minimize the number of questions posed to users. Human-in-the-loop approaches have been applied to data cleaning to ensure better quality fixes with respect to automated repairs.

In recent years, HIL also found applications in machine learning not only aiming towards more accurate algorithms or to obtain the desired accuracy faster, but also to make humans more effective and more efficient. Depending on who is in control of the learning process, we can identify different approaches : Active learning (AL), in which the system remains in control of the learning process and treats humans as oracles to annotate unlabeled data, Interactive machine learning (IML), in which there is a closer interaction between users and learning systems, with people interactively supplying information in a more focused, frequent, and incremental way compared to traditional machine learning, Machine teaching (MT), where human domain experts have control over the learning process by delimiting the knowledge that they intend to transfer to the machine learning model.

In this thesis, we focus on exploiting HIL techniques for developing Data-Intensive applications.

This thesis focuses on every aspect of the data pipeline and aims towards developing a framework for evaluating HIL applications in both the algorithm and human aspects.
While in the Data Collection, Data cleaning and Machine learning domains researchers use quantitative metrics that evaluate the goodness of their proposed models (e.g. Accuracy, F1, etc.), in the Human Centered AI (HCAI) domain the focus is towards qualitative metrics such as explainability, contestability, thrust, interpretability and so on. Big companies and institutions like Google, Microsoft and the European Union have developed guidelines for HCAI evaluations but most of the research focuses only on qualitative analysis with respect to quantitative analysis.
In this thesis, we will investigate metrics for the quantitative analysis of the human aspects of Human-Machine applications and applying the principles and guidelines coming from the HCAI domain to the HIL domain.