# Nelly Barret

ASSISTANT PROFESSOR AT INSA LYON AND LIRIS

**Blaise Pascal building, 7 Avenue Jean Capelle, 69100 Villeurbanne, France**

**Office 502.3.21 | nelly.barret@insa-lyon.fr | https://orcid.org/0000-0002-3469-4149**

## Research interests

I work in the broad field of data management, and more precisely Imy research interests focus on heterogeneous and/or multimodal data, data integration systems, and data exploration. Recently, I also added Federated Learning to my research interests. The projects that I worked on have always been part of multidisciplinary contexts, including urban planning, journalism, and health, and have been exemplified on large, complex, and often real-world data.

## Academic positions

**Assistant Professor | INSA Lyon & LIRIS (FR)** **Sept. 2025 – now**

- Research: part of the DRIM team, focusing on document engineering and distributed systems.
- Teaching: 1st and 2nd year courses at the engineer school INSA Lyon.

**Postdoctoral researcher | Politecnico di Milano (IT)** **Apr. 2024 – July 2025**

- Horizon Europe project: 'Better real-world health-data distributed analytics research platform'.
- Close collaborators: Pietro Pinoli (associate professor, Politecnico di Milano, IT), Anna Bernasconi (assistant professor, Politecnico di Milano, IT), Boris Bikbov (medical doctor, Politecnico di Milano, IT).

**PhD student | Institut Polytechnique de Paris & Inria Saclay (FR)** **Jan. 2021 – March 2024**

- PhD thesis title: 'User-oriented exploration of semi-structured datasets'
- PhD defense jury: Fatiha Sais (présidente – professor, Univ. Paris-Saclay and LISN), Jean-Marc Petit (rapporteur – professor, at INSA Lyon and LIRIS), Olivier Teste (rapporteur – professor, Univ. Toulouse Jean Jaurès and IRIT), Katja Hose (examinatrice – professor, TU Wien, AT), Stefano Ceri (examinateur – professor, Politecnico di Milano, IT), Fatemeh Nargesian (examinatrice – associate professor, Univ. Rochester, USA), Ioana Manolescu (advisor – research director, Ecole Polytechnique and Inria Saclay), Karen Bastien (co-advisor – WeDoData CEO).

**Master intern | LIRIS (FR)** **Feb. 2020 – July 2020**

- Master thesis title (in French): 'Prédiction de l'environnement d'un quartier'.
- Advisors: Fabien Duchateau (associate professor, Univ. Lyon 1 and LIRIS) and Franck Favetta (associate professor, Univ. de Lyon and LIRIS).
- Other collaborators: Nelly Duong (CEO, Home in Love), Behnaz Jullien (intern in psychology, Univ. Lyon 2), Wissame Laddada (post-doctoral researcher, LIRIS), and Ludovic Moncla (associate professor, computer science at INSA Lyon).

**Bachelor intern | LIRIS (FR)** **May 2018 – July 2018**

- Bachelor thesis title (in French): 'Intégration de données géographiques pour la recommandation de quartiers'.

- Advisors: Fabien Duchateau (associate professor, Univ. Lyon 1 and LIRIS) and Franck Favetta (associate professor, Univ. de Lyon and LIRIS).
- Other collaborators: Nelly Duong (CEO, Home in Love), Loic Bonneval (associate professor in sociology, Univ. Lyon 2) and Aurélien Gentil (PhD student in sociology, Univ. Lyon 2).

## Education

**PhD degree | Institut Polytechnique de Paris** **Jan. 2021 – March 2024**

*Major:* Computer Science, Data and Artificial Intelligence

**Master's degree | Université Claude Bernard Lyon 1** **Sept. 2018 – July 2020**

*Major:* Computer Science | *Minor:* Artificial Intelligence

**Bachelor's degree | Université Claude Bernard Lyon 1** **Sept. 2015 – July 2018**

*Major:* Computer Science

## Research visits

**Institut de Recherche en Informatique de Toulouse** **Oct. 2025 (2 days)**

## Awards

**PhD thesis award | BDA conference** **Oct. 2024**

*Website:* https://bda2024.sciencesconf.org/resource/page/id/10

- My PhD work won the 2nd prize for the PhD award delivered by the BDA conference, awarding one or two PhD with significant contributions in the French data management community.

## Research projects

**BETTER: real-world health-data distributed analytics research platform** **Apr. 2024 – July 2025**

*Role:* contributor (tasks 4.1, 4.2) | *Grant:* Horizon Europe (2022 – 2027) | *Partners:* 7 European hospitals | *Website:* https://better-health-project.eu

- Main scientific contributions: (1) two conceptual models for data and metadata respectively, both general, able to represent various multi-modal healthcare data, allowing the usage of ontologies, and extensible to various healthcare scenarios; (2) an ETL algorithm to fully automatically convert existing data and metadata to instances of our models, resulting in an interoperable database; (3) a metadata and data catalogue for exploration and ease federated learning algorithm design.
- Main practical achievements: (1) each hospital has an interoperable database, part of the global BETTER network; (2) federated learning algorithms are co-created and implemented by researchers and practitioners; (3) the catalogue implementation for the 3 use-cases (genetic rare diseases).

**CONNECTIONSTUDIO: user data exploration of heterogeneous data** **March 2023 – March 2024**

*Role:* contributor | *Grant:* ANR SourcesSay project (2020 – 2024) | *Website:* https://connectionstudio.inria.fr

- Main scientific contribution: a user-oriented methodology for querying a relational data lake by selecting data pieces of interest instead of formulating SQL queries.
- Main practical achievement: a software suite for the data exploration of the underlying data lake.

**PATHWAYS: find interesting entity-to-entity paths in heterogeneous data** **May 2022 – March 2024**

*Role:* co-leader | *Grant:* DIM RFSI PHD 2020-01 | *Website:* https://team.inria.fr/cedar/projects/pathways

- Main scientific contributions: (1) efficiently enumerate all paths connecting named entities appearing in multi-modal and heterogeneous data by relying on an intermediate summary graph and a view-based algorithm; (2) the ranking of paths based on their interestingness, a quantitative measure based on entity confidence and information dilution.
- Main practical achievements: (1) a novel ChatGPT-based module for entity extraction; (2) the efficient evaluation of paths by using a dedicated multi-query optimization algorithm.

**ABSTRA: Entity-Relationship summaries from semi-structured data** **Jan. 2021 – March 2024**

*Role:* leader | *Grant:* DIM RFSI PHD 2020-01 | *Partners:* WeDoData start-up | *Website:* https://team.inria.fr/cedar/projects/abstra

- Main scientific contributions: (1) a novel model-agnostic data summarization method relying on data kinds instead of specific model features; (2) the selection of most representative entities, their attributes and relationships from the data summary by relying on a graph representation and PageRank node scores.
- Main practical achievement: (1) an end-to-end pipeline to summarize any (set of) heterogeneous datasets; (2) an interface to visualize summaries as Entity-Relationship diagrams.
- Partner: WeDoData is a French SME specializing in data journalism and data visualization.

**PREDIHOOD: supervised prediction of the environment of neighborhoods** **Feb. 2020 – July 2020**

*Role:* leader | *Grant:* LabEx IMU (Intelligences des Mondes Urbains) | *Partners:* HomeInLove start-up

- Main scientific contributions: (1) an algorithm to automatically select the most useful subset of indicators among the hundreds collected ones; (2) an algorithm to automatically predict in a supervised manner the environment of any French neighborhood based on several subsets of indicators.
- Main practical achievements: (1) collect and process cartographic and urban data from public data such as INSEE; (2) an interface for predicting the environment of any French neighborhood; (3) a general interface for testing, evaluating and comparing AI algorithms.
- Partner: HomeInLove is a French start-up helping people in case of professional mobilities to find a new place to live based on their criteria and needs (neighborhood, family, hobbies, ...).

**GEOALIGN: matching and merging geographic entities** **Jan. 2019 – June 2019**

*Role:* leader | *Grant:* Master 1 project

- Main scientific contributions: (1) a customizable similarity formula for estimating the matching similarity between geographic entities; (2) the computation of the estimated quality of the matching without ground truth.
- Main practical achievements: (1) a map interface for find, matching and merging similar geographic entities.

**VIZLIRIS: recommending and clustering French neighborhoods** **May 2018 – July 2018**

*Role:* leader | *Grant:* LabEx IMU (Intelligences des Mondes Urbains) | *Partners:* HomeInLove start-up | *Website:* https://perso.liris.cnrs.fr/fabien.duchateau/?page=VizLIRIS/

- Main scientific contribution: a data integration pipeline for integrating cartographic and INSEE data.
- Main practical achievements: (1) a map interface for recommending and clustering French neighborhoods based on well-known ML algorithms; (2) a use-case validation with HomeInLove data.

## Research software and tools

**I-ETL** **Apr. 2024 – July 2025**

*Role:* developer | *Time:* 1 year | *Languages:* Language: Python | *LOC:* 6k | *Repository:* https://github.com/DEIB-GECO/i-etl

**ConnectionStudio** **March 2023 – March 2024**

*Role:* contributor | *Time:* 1 year | *Languages:* Languages: Java, Javascript | *LOC:* 25k | *Repository:* https://gitlab.inria.fr/cedar/connection-studio

**PathWays** **May 2022 – March 2024**

*Role:* main developer | *Time:* 2 years | *Languages:* Language: Java | *LOC:* 4k | *Repository:* https://gitlab.inria.fr/cedar/pathways

**Abstra** **Jan. 2021 – March 2024**

*Role:* developer | *Time:* 3 years | *Languages:* Language: Java | *LOC:* 10k | *Repository:* https://gitlab.inria.fr/cedar/abstra

**Predihood** **Feb. 2020 – July 2020**

*Role:* main developer | *Time:* 6 months | *Languages:* Languages: Python, JavaScript | *LOC:* 3k | *Repository:* https://gitlab.com/fduchate/predihood

**GeoAlign** **Jan. 2019 – June 2019**

*Role:* main developer | *Time:* 6 months | *Languages:* Languages: PHP, JavaScript | *LOC:* 4k | *Repository:* private

**VizLIRIS** **May 2018 – July 2018**

*Role:* main developer | *Time:* 3 months | *Languages:* Languages: Python, JavaScript | *LOC:* 1k | *Repository:* https://forge.univ-lyon1.fr/stage-Nelly/HiL-recommender

## Working groups

**Commission égalité & parité femme-homme** **Nov. 2025 – now**

*Role:* permanent and active member | *Website:* https://vargas-solar.com/aequitas

- The aim of this commission is to work towards building a more equitable scientific community by: (a) increasing knowledge about the scientific contributions of the female community by promoting the visibility of women at LIRIS, (b) increase communication in the lab to raise awareness on F/M equality questions, and (c) studying women's participation in the data science and artificial intelligence workforce.

**Diversity, Equity and Inclusion** **Apr. 2023 – now**

*Role:* active member | *Website:* https://dbdni.github.io/

- I co-lead the SCOUT action with Madhulika Mohanty (researcher, Inria Saclay) and Sujaya Maiyya (assistant professor, Univ. of Waterloo, USA). It aims at facilitating DEI efforts in the database community. My main proposal is a checklist to easily comply with inclusive submission guidelines; it will be integrated into EasyChair and CMT.

**FAIRification of genomic annotations** **Dec. 2024 – now**

*Role:* active member | *Website:* https://www.rd-alliance.org/groups/fairification-genomic-annotations-wg/activity/

- We aim at defining novel methods to make genomic annotations and data more interoperable. I co-lead two actions: (1) define a strategy to publicly and permanently store harmonized genomic metadata with Sveinung Gundersen (computer scientist, ELIXIR) and Evan Christensen (PhD in bio-informatics, Univ. d'Utah, CA); and (2) define new models for representing genomic annotations with Sveinung Gundersen and Adam Wright (researcher, Cancer institute of Ontario, USA).

## Publications

**Peer-reviewed international journals**

1. Nelly Barret, Sourav Bhowmick, Angela Bonifati, Barbara Catania, Stratos Idreos, Ekaterini Ioannou, Madhulika Mohanty, Sana Sellami, Roee Shraga, Utku Sirin, Juno Steegmans, Pınar Tözün, Soror Sahri, Genoveva Vargas-Solar. Diversity, Equity and Inclusion Activities in DatabaseConferences: A 2024 Report. *ACM SIGMOD Record.* 2025.
2. Nelly Barret, Anna Bernasconi, Boris Bikbov, Pietro Pinoli. I-ETL: an interoperability-aware health (meta)data pipeline to enable federated analyses. *BMC Medical Informatics and Decision Making.* 2025.
3. Nelly Barret, Antoine Gauquier, Jia-Jean Law, Ioana Manolescu. Finding meaningful paths in heterogeneous graphs with PathWays. *Information Systems.* 2025.
4. Nelly Barret, Fabien Duchateau, Franck Favetta. Predihood: an open-source tool for predicting neighbourhoods' information. *Journal of Open Source Software.* 2021.
5. Nelly Barret, Fabien Duchateau, Franck Favetta, Aurélien Gentil, Loïc Bonneval. An Environmental Study of French Neighbourhood.. *Communications in Computer and Information Science.* 2021.

**Peer-reviewed international conferences**

6. Nelly Barret, Anna Bernasconi, Cinzia Cappiello, Giaccomo Palu, Pietro Pinoli. Leveraging profiling to bridge healthcare silos for federated analyses. *International Conference on Advanced Information Systems Engineering (Forum track).* 2025.
7. Nelly Barret, Ioana Manolescu, Prajna Upadhyay. Computing generic abstractions from application datasets. *International Conference on Extending Database Technology.* 2024.
8. Nelly Barret, Simon Ebel, Théo Galizzi, Ioana Manolescu, Madhulika Mohanty. User-friendly exploration of highly heterogeneous data lakes. *International Conference on Cooperative Information Systems.* 2023.
9. Nelly Barret, Antoine Gauquier, Jia-Jean Law, Ioana Manolescu. Exploring heterogeneous data graphs through their entity paths. *European Conference on Advances in Databases and Information Systems.* 2023.
10. Nelly Barret, Ioana Manolescu, Prajna Upadhyay. Abstra: toward generic abstractions for data of any model. *ACM International Conference on Information & Knowledge Management.* 2022.
11. Nelly Barret, Fabien Duchateau, Franck Favetta, Loïc Bonneval. Predicting the environment of a neighbourhood: a use case for France. *International Conference on Data Science, Technology and Applications.* 2020.

**Peer-reviewed international workshops**

12. Nelly Barret, Tudor Enache, Ioana Manolescu, Madhulika Mohanty. Finding the PG schema of any (semi) structured dataset: a tale of graphs and abstraction. *SEAGRAPH workshop in International Conference on Data Engineering.* 2024.
13. Oana Balalau, Nelly Barret, Simon Ebel, Théo Galizzi, Ioana Manolescu, Madhulika Mohanty. Graph lenses over any data: the ConnectionLens experience. *SEAGRAPH workshop in International Conference on Data Engineering.* 2024.

**Peer-reviewed national conferences**

14. Nelly Barret, Simon Ebel, Théo Galizzi, Ioana Manolescu, Madhulika Mohanty. Exploration utilisateur de lacs de données très hétérogènes. *Conférence Francophone sur l'Extraction et la Gestion des Connaissances.* 2024.
15. Nelly Barret, Fabien Duchateau, Franck Favetta, Maryvonne Miquel, Aurélien Gentil, Loïc Bonneval. À la recherche du quartier idéal. *Conférence Francophone sur l'Extraction et la Gestion des Connaissances.* 2019.

**Demonstrations**

16. Nelly Barret, Antoine Gauquier, Jia-Jean Law, Ioana Manolescu. PathWays: entity-focused exploration of heterogeneous data graphs. *The Semantic Web: ESWC Satellite Events.* 2023.
17. Nelly Barret. Facilitating heterogeneous dataset understanding. *Conférence sur la Gestion de Données – Principes, Technologies et Applications.* 2021.
18. Nelly Barret, Fabien Duchateau, Franck Favetta, Ludovic Moncla. Spatial entity matching with GeoAlign. *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems.* 2019.

**Manuscripts**

19. Nelly Barret. User-oriented exploration of semi-structured datasets. *Institut Polytechnique de Paris.* 2024.

## Professional service

**Member of organization committees**

- 2025: Conceptual Modeling for Life Sciences [CMLS – A] (workshop co-chair)

**Member of program committees**

- 2025: International Conference on Advances in Databases, Knowledge, and Data Applications [DBKDA – N/A], European Conference on Advances in Databases and Information Systems [ADBIS – C], BDA Gestion de Données – Principes, Technologies et Applications [BDA – N/A] (demonstrations only)
- 2025: Information Systems and AI for Life Sciences [iSAILS – A]

## Reviewing activities

**Journal reviewer**

- 2025: GigaScience [Q1], Journal of Data and Information Quality [JDIQ – Q2], Computers, Materials and Continua [CMC – Q2], Journal of Open Source Software [JOSS – N/A]
- 2024: PLoS One [Q1], BMC Medical informatics and Decision Making [Q1], Scientific Reports [Q1]

**Conference reviewer**

- 2025: International Conference on Data Engineering* [ICDE – A*]
- 2023: Conference on Innovative Data systems Research* [CIDR – A], Extending DataBase Technology* [EDBT – A]
- 2021: International Conference on Web Engineering* [ICWE – B]

## Institutional responsibilities

**DRIM webpage maintainer** **Sept. 2025 - now**

*Role:* main maintainer | *Website:* https://liris.cnrs.fr/equipe/drim

- Manage the institutional webpage of the DRIM team at the LIRIS lab.

**LinkedIn page maintainer** **April. 2024 – July 2025**

*Role:* main maintainer | *Website:* https://linkedin.com/company/polimi-data-science-group

- Manage the LinkedIn page of the Data Science group at DEIB, Politecnico di Milano.

**PhD representative** **Jan. 2023 – March 2024**

*Role:* main representative

- Represent Inria Saclay PhD students in Institut Polytechnique de Paris.

**Team seminar organizer** **Jan. 2021 – March 2024**

*Role:* main organizer | *Website:* https://team.inria.fr/cedar/category/seminar

- Organize and disseminate the CEDAR team seminars at Inria Saclay.

**Internship application website contributor** **Jan. 2021 – Dec. 2021**

*Role:* active contributor | *Website:* https://stages.dix.polytechnique.fr/home

- Co-develop and deploy the website for internships at École Polytechnique, in collaboration with École Polytechnique IT team.

## Talks

**Scientific talks (seminars)**

- Multimodal data without borders: integration and exploration to the rescue. *IRIT, Toulouse, FR.* Oct. 2025.
- Leveraging profiling to bridge healthcare silos. *FIL, Lyon, FR.* June 2025.
- Entrepôts de santé : de l'intégration à l'analyse fédérée. *GDR MaDICS, Toulouse, FR.* May 2025.
- Integrating and exploring heterogeneous datasets. *Politecnico di Milano, Milan, IT.* April 2024.
- Heterogeneous datasets: a tale of integration and exploration. *LIRIS, Lyon, FR.* Jan. 2024.
- User-oriented exploration of semi-structured datasets. *LIB, Dijon, FR.* Jan. 2024.
- User-oriented exploration of semi-structured datasets. *LISN, Paris, FR.* Oct. 2023.

**Panels**

- Quand l'intelligence artificielle hérite de nos préjugés (video).. *Fête de la Science, Lyon, FR.* Oct. 2025.

**Vulgarization**

- From data to journalism (interactive course). *Lycée international de Palaiseau, Palaiseau, FR.* Jan. 2024.
- Intelligence artificielle : un outil pour le journalisme (seminar). *FI forum, Paris, FR.* July 2023.
- ConnectionStudio : des données au journalisme (forum). *DataJournos, Paris, FR.* May 2023.
- Recherche en integration de données : le cas de ConnectionLens (seminar). *Online.* March 2022.

**Female empowerment talks**

- **Femmes scientifiques (round table for middle school students).** *Lyon, FR.* Nov. 2025.
- **Journée Filles, Mathématiques, Informatique (speed meetings).** *Lyon, FR.* Nov. 2025.
- Girls@CSE (talks). *Politecnico di Milano, IT.* Sept. 2024.

## Teaching responsibilities

**Recurring courses** **Sept. 2025 – now**

*Role:* magistral and lab courses

- ISN1: informatique et société numérique (fall, INSA Lyon, engineer 1st year, in French)
- SOL: systèmes et outils logiciels (fall, INSA Lyon, engineer 1st year, in French)

**Involvement in teaching teams** **Sept. 2025 – now**

*Role:* contributor

- ISN2: new project topic
- ISN2: network module redesign
- CGA: member of the Animation and Management Committee (Comité de Gestion et d'Animation)

**Previous courses** **Feb. 2022 – May 2024**

- CSE203: CS project (spring 2024, Ecole Polytechnique, bachelor 2nd year, in English)
- CSE204: Machine Learning (spring 2023, Ecole Polytechnique, bachelor 2nd year, in English)
- INF411: bases de programmation et d'algorithmique (fall 2022, Ecole Polytechnique, engineer 2nd year, in French)
- INF371: mécanismes de la programmation orientée objet (spring 2022, Ecole Polytechnique, engineer 1st year, in French)

## Advising

**Master students**

- Ahmed Bel Hadj Youssef. 'FADE : vers des lacs de données plus FAIR' (fall 2025, INSA Lyon, engineer 5th year)
- Antoine Gauquier. 'Path exploration of ConnectionLens graphs' (summer 2022, IMT Nord-Europe, master 1st year)

**Bachelor students**

- Shay Pripstein. 'Target data acquisition in open repositories' (winter 2023, École Polytechnique, bachelor 3rd year)
- Tudor Enache. 'From simple to property graphs' (winter 2023, École Polytechnique, bachelor 3rd year)
- Nikola Dobricic. 'Graph queries for abstractions' (summer 2023, Ecole Polytechnique, bachelor 2nd year)
- Jia-Jean Law. 'Optimization of entity-to-entity data paths' (winter 2022, Ecole Polytechnique, bachelor 3rd year)