Antoine Gauquier

PhD Student in Computer Science
Data Management, Machine Learning

French, born 21 March 2001.
📱 *(+33) 7 68 79 70 95*
✉ *antoine.gauquier@ens.psl.eu*
🌐 *https://antoinegauquier.github.io/*
🎓 *Google Scholar Profile*

Research Interests

My research focuses on **efficient and scalable systems for acquiring and searching large volumes of heterogeneous, real-world data**. My work sits at the intersection of **data management and applied machine learning**, with an emphasis on systems that are **efficient, scalable, and practically deployable**. During my PhD, I have designed and implemented : a **focused and RL-based Web crawler** for automated large-scale data acquisition (*EDBT* 2026 ; journal submission under review) ; a **modular statistical table retrieval system** (STAR) that searches at scale across heterogeneous tabular corpora (*ICDE* 2026 ; journal submission under review) ; and **TabAgree**, a platform for reliable spreadsheet table annotation leveraging inter-annotator agreement (under review). I am currently working on **automatic table structure detection and cell role classification** in spreadsheets, combining layout analysis and sequence labeling. Beyond these main research directions, I have contributed to collaborative work on **heterogeneous graph exploration and efficient entity path finding** (*ADBIS* 2023 ; *ESWC* 2023 ; *Information Systems* 2025), **multimodal information extraction from scientific documents** (*JCDL* 2024 ; *ECIR* 2025), and **uncertainty reasoning and cooperative game theory** (*ECSQARU* 2025).

Education

2023–Present **PhD in Computer Science**, *École normale supérieure–PSL*, Paris, France

2022–2023 **M2 in Data Management and Artificial Intelligence**, *Télécom Paris*, Palaiseau, France

2021 **Exchange Semester in Computer Science**, *Université du Québec à Montréal (UQAM)*, Montréal, Canada

2018–2023 **Engineering Diploma (equivalent to M.Sc.) in Computer Science**, *IMT Nord Europe*, Lille, France

Academic Positions

2023–Present **PhD Candidate and Teaching Assistant**, *École normale supérieure–PSL*, Paris, France
Advisors : Pierre Senellart (Full Professor, *École normale supérieure–PSL*) and Ioana Manolescu (Research Director, *Inria* ; part-time Professor, *École Polytechnique*).

Feb 2023–Jul 2023 **Research Intern**, *École normale supérieure–PSL*, Paris, France
Topic : Impact of document class on the automatic extraction of mathematical environments from scientific literature.
Advisor : Pierre Senellart.

May 2022–Sep 2022 **Research Intern**, *Inria and École Polytechnique*, Palaiseau, France
Topic : Path-based ConnectionLens graph exploration.
Advisor : Ioana Manolescu.

Sep 2021–Jan 2022 **Part-Time Data Science Assistant**, *Inserm*, Lille, France
Part-time position at the *Translational Research Institute for Diabetes*, carried out alongside academic studies.

May 2021–Sep 2021 **Research Intern in Statistics and Machine Learning**, *CHU de Lille*, Lille, France
Topic : Predictive modeling for remote monitoring and long-term outcomes after bariatric surgery.
Advisors : Violeta Raverdy (Clinical Researcher, *Translational Research Institute for Diabetes*) and Cristian Preda (Full Professor, *Université de Lille*).

Publications

Under Review

Antoine Gauquier, Pierre Senellart, and Ioana Manolescu. TabAgree : An Agreement-Aware Platform for Reliable Spreadsheet Table Annotation. 2026.

Antoine Gauquier, Pierre Senellart, and Ioana Manolescu. Reinforcement Learning–Based Focused Web Crawling for Robust and Scalable Data Acquisition. 2026.

Antoine Gauquier, Pierre Senellart, and Ioana Manolescu. STAR : Efficient, Scalable, and Modular Retrieval of Statistical Tables. 2026.

Peer-Reviewed International Conferences

Rankings are based on CORE data up to early 2026.

Antoine Gauquier, Simon Ebel, Helena Galhardas, Théo Galizzi, Ioana Manolescu, Aurélien Peden, and Pierre Senellart. Efficient and Scalable Search for Statistics. *Proceedings of ICDE*, 2026. **[A*]**

Antoine Gauquier, Pierre Senellart, and Ioana Manolescu. Efficient Crawling for Scalable Web Data Acquisition. *Proceedings of EDBT*, 2026. **[A]**

Pratik Karmakar, Antoine Gauquier, and Pierre Senellart. Expected Shapley Value is Shapley Value for Expected Utility Game. *Proceedings of ECSQARU*, 2025. **[C]**

Shrey Mishra, Antoine Gauquier, and Pierre Senellart. Modular Multimodal Machine Learning for Extraction of Theorems and Proofs in Long Scientific Documents. *Proceedings of JCDL*, 2024. **[A*]** *Ranked A* until 2018 (removed—not considered a pure CS conference).*

Antoine Gauquier and Pierre Senellart. Automatically Inferring the Document Class of a Scientific Article. *Proceedings of DocEng*, 2023. **Runner-up for Best Paper Award [B]**

Nelly Barret, Antoine Gauquier, Jia-Jean Law, and Ioana Manolescu. Exploring Heterogeneous Data Graphs Through Their Entity Paths. *Proceedings of ADBIS*, 2023. **[C]**

Peer-Reviewed International Journals

Nelly Barret, Antoine Gauquier, Jia-Jean Law, and Ioana Manolescu. Finding meaningful paths in heterogeneous graphs with PathWays. *Information Systems*, 2025. **[A*]** *CORE 2022 ; Q1 in SJR.*

Peer-Reviewed International Demonstrations

Shrey Mishra, Neil Sharma, Antoine Gauquier, and Pierre Senellart. TheoremView: A Framework for Extracting Theorem-Like Environments from Raw PDFs. *Proceedings of ECIR*, 2025. **[A]**

Nelly Barret, Antoine Gauquier, Jia-Jean Law, and Ioana Manolescu. PathWays: Entity-Focused Exploration of Heterogeneous Data Graphs. *Proceedings of ESWC*, 2023. **[B]**

Peer-Reviewed International Workshops

Antoine Gauquier. Towards Efficient Construction of a Traceable, Multimodal, and Heterogeneous Data Warehouse. *Proceedings of VLDB PhD Workshop*, 2024.

Shrey Mishra, Yacine Brihmouche, Théo Delemazure, Antoine Gauquier, and Pierre Senellart. First Steps in Building a Knowledge Base of Mathematical Results. *Proceedings of the 4th Scholarly Document Processing workshop (co-located with ACL)*, 2024.

Peer-Reviewed National Conferences

Shrey Mishra, Antoine Gauquier, and Pierre Senellart. Apprentissage multimodal modulaire pour l'extraction de théorèmes et de preuves dans des documents scientifiques longs. *Proceedings of EGC*, 2025.

Reports and pre-prints

Shrey Mishra, Antoine Gauquier, and Pierre Senellart. Modular Multimodal Machine Learning for Extraction of Theorems and Proofs in Long Scientific Documents (Extended Version). *arXiv abs/2307.09047*, 2024.

Antoine Gauquier. Impact of the document class in the automatic extraction of mathematical environments in the scientific literature. *IMT Nord Europe and École normale supérieure–PSL*, 2023.

Antoine Gauquier. Path-based ConnectionLens graph exploration. *IMT Nord Europe and Inria*, 2022.

Talks and Presentations

Nov 2025 Efficient Crawling for Scalable Web Data Acquisition, *Valda Team Seminar*, École normale supérieure–PSL, Paris, France

Oct 2025 Efficient and Scalable Search for Statistics, *41st Conference on Databases (BDA) « Gestion de Données – Principes, Technologies et Applications »*, IRIT, Toulouse, France

Sep 2025 Expected Shapley Value is Shapley Value for Expected Utility Game, *18th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU)*, University of Hagen, Hagen, Germany

Jan 2025 Modular Multimodal Machine Learning for Extraction of Theorems and Proofs in Long Scientific Documents (Poster), *25th Conference on « Extraction et Gestion des Connaissances » (EGC)*, INSA Strasbourg, Strasbourg, France

Oct 2024 Efficient Crawler for Scalable Web Data Acquisition, *40th Conference on Databases (BDA) « Gestion de Données – Principes, Technologies et Applications »*, CIUR, Orléans, France

Aug 2024 Towards Efficient Construction of a Traceable, Multimodal, and Heterogeneous Data Warehouse, *Prof. Stéphane Bressan's Team Seminar*, NUS School of Computing, Singapore, Singapore

Aug 2024 Towards Efficient Construction of a Traceable, Multimodal, and Heterogeneous Data Warehouse, *50th International Conference on Very Large Databases – PhD Workshop (VLDB PhD Workshop)*, Langham Place, Guangzhou, China

Apr 2024 Efficient and Focused Web Crawling for Statistical Data Sources Retrieval, *2nd Workshop « Infox sur Seine »*, Académie du climat, Paris, France

Aug 2023 Automatically Inferring the Document Class of a Scientific Article, *23rd ACM Symposium on Document Engineering (DocEng)*, University of Limerick, Limerick, Ireland

Research Visits

Jul–Aug 2025 **Visiting PhD Student**, *National University of Singapore and CNRS@CREATE*, Singapore, Singapore

Teaching Experience

2023–2026 **Teaching Assistant**, *PSL University and Lycée Louis-le-Grand*, Paris, France

- **Introduction to Algorithms** and **Python Programming** (L1) : 80 hours. Laboratory sessions and selected lectures.
- **Differential Calculus** (L2, *PSL University*) : 32 hours in 2024 and 2025. Tutorials.
- **Bibliographic Research** (L2, *PSL University*) : 16 hours in 2023 and 2024. Workshop-style lectures.

Supervision (Undergraduate Students)

2024 **Léo Boullot**, *PSL University*, Paris, France
Construction of a dataset for the automatic extraction of tabular data from PDF documents.

2024 **Paul Sevestre**, *PSL University*, Paris, France
Automatic semantic interpretation of tabular data.

Professional Service

Scientific Community

2025 **Program Committee Member**, *Workshop on Artificial Intelligence for Scientific Publications (WASP)*
Co-located with IJCNLP–AACL 2025.

Education Community

2024–Present **Selection Committee Member**, *CPES "Data Science, Arts and Cultures"*
Joint program between *PSL University* and *Lycée Louis-le-Grand*. Committee responsible for the selection of incoming students.

Institutional Responsibilities

2025–Present **PhD Student Representative**, *Board of Directors (Conseil d'Administration), École normale supérieure*

2024–Present **Representative of Non-Permanent Members**, *Computer Science Laboratory, École normale supérieure*