

GRADUATION :

PROGRESSIVE RELATIONAL DATA AUGMENTATION

Research Laboratory: LISN

Team: AVIZ - Lahdak

Advisor: Benoît Groz groz@lisn.fr

Advisor: Francesca Bugiotti francesca.bugiotti@centralesupelec.fr

Advisor: Jean-Daniel Fekete Jean-Daniel.Fekete@inria.fr

Affiliations: LISN - Paris-Saclay - CentraleSupélec - Inria - CNRS

1 Context

The web provides access to many large datasets. For instance, open data platforms such as gouv.data.fr and NYC Open Data store data about a miscellany of topics such as loans, elections, traffic, economy, health, etc. When we process a given dataset, data analytics and machine learning models can sometimes be enhanced by joining some information from additional datasets. However, when the candidate external datasets reach billions of records, it is impractical to download the full candidate datasets in order to identify whether they add relevant information. The aim of this internship is to survey, experiment, and design techniques to progressively identify the relevant external datasets that can be joined to augment a given dataset [1].

2 The PDA Approach

The novel Progressive Data Analysis paradigm [2] (PDA) has been designed to overcome the latency limitation. Similar to Online Aggregation [3], PDA splits long computations into several batches of short computations arriving in sequence. Each new batch improves upon the previous one until the entire computation is complete, but intermediate results can be shown or visualized along the way. Analysts can also interact with the process to steer it if needed. With PDA, exploratory analysis can begin quickly, even with very large datasets, and results are improved iteratively. However, to go beyond monitoring the computation, a PDA system should also provide a quality assessment of the ongoing results, allowing the analyst to decide whether the quality is good enough to make a decision with the current approximate result or if more time is needed.

3 Challenges

The intern will first survey the state of the art to identify approaches amenable to progressive data analysis.

- In a first step, we plan to focus on relational pairwise (left) equijoins: identifying if a candidate table can be joined with the table of interest on some attribute(s).
- Based on this state-of-the-art, the work will then implement and benchmark one or a few progressive relational data augmentation algorithms.

- For this, one will identify quality measures for the progressive algorithm, such as information-theoretic measures, based on related work in relational data augmentation [5] and progressive data analysis (Poliuha 25). While the most relevant additional data sources can be defined independently of any application in a first step, augmentation tailored for specific end goals could then be considered. For example, to add features to improve machine learning training.

Multiple extensions could then be considered for follow-up work after this internship. Richer types of joins may then be considered, such as semantic joins that allow approximate matches, and multi-way inner joins that join records from multiple tables at once and drop records from the original tables that do not have matches. Another extension that could be considered would be Table union search [4], where the objective is to augment a candidate table with other tables using the union operator rather than joins. Beyond identifying relevant sources for augmentation, it would be interesting to progressively serve those sources alongside the original data, as part of a broader data processing platform. However, we believe this would exceed the scope of this internship.

References

- [1] Yuyang Dong, Chuan Xiao, Takuma Nozawa, Masafumi Enomoto, and Masafumi Oyamada. Deepjoin: Joinable table discovery with pre-trained language models, 2023.
- [2] Jean-Daniel Fekete, Danyel Fisher, and Michael Sedlmair. *Progressive Data Analysis*. Eurographics, November 2024.
- [3] Joseph M. Hellerstein, Peter J. Haas, and Helen J. Wang. Online aggregation. In *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data, SIGMOD '97*, page 171–182, New York, NY, USA, 1997. Association for Computing Machinery.
- [4] Aamod Khatiwada, Roe Shraga, and Renée J Miller. Diverse unionable tuple search: Novelty-driven discovery in data lakes. in edbt 2026.
- [5] Aécio Santos, Flip Korn, and Juliana Freire. Efficiently estimating mutual information between attributes across tables. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pages 193–206, 2024.