Unlocking the Power of Data Dependencies in Data Pipelines

Location	LAMSADE - PSL Research University - Université Paris-Dauphine, France
Supervisors	Khalid Belhajjame and Maude Manouvrier
Funding	4 to 6 months paid 560 euros per month, from March/April to August 2023.

Context

Data dependencies [1, 3] refer to relationships or connections between different variables in a dataset. Understanding these dependencies is crucial and has a number of applications.
Data profiling for Machine Learning: Understanding data dependencies is critical for creating accurate and effective machine learning models. The quality of the input data has a direct impact on the accuracy of the model, and understanding data dependencies helps ensure that the data is suitable for use in machine learning.
Data mining: Data dependencies can help you identify patterns and relationships in the data that may not be immediately obvious. These patterns can be used to make predictions and classify data, making it useful in various data mining tasks such as association rule mining and clustering.

Objectives

This internship will build upon the recent research in data dependency mining in dynamic settings [2]. As a member of a dynamic team, the student will be exploring innovative ways to compute data dependencies in situations where the data is transformed through a data preparation pipeline. The goal is to assess the impact of this preparation process on the dependencies within the data, as well as its overall quality [4].

Opportunities

The subject of data dependencies is a critical and fascinating aspect of machine learning and AI, providing students with the opportunity to gain practical skills and explore cutting-edge technologies that are shaping the future of the field. The demand for professionals with skills in machine learning and AI is growing rapidly, and understanding data dependencies is a valuable skill for anyone looking to build a career in this field in both academia and industry. On this point, it is worth noting that **the internship is likely to lead to a PhD on a related topic**.

How to apply

We seek for excellent and highly motivated student with a background in Computer Science having good knowledge of database theory and good programming skills (Python or Java).
Please send the following material in a single PDF document before February 20th, 2023: - fully detailed CV, - academic records (master's degree or equivalent), - recommendation(s) and supporting letter(s).

References

[1] Abedjan, Z., Golab, L., Naumann, F., & Papenbrock, T. (2018). Data profiling. Synthesis Lectures on Data Management, 10(4), 1-154.

[2] Belhajjame, K., (2023) Efficient Maintenance of Agree-Sets Against Dynamic Datasets. EDBT 2023: 14-26

[3] Comignani, U., Berti-Équille, L., Novelli, N., & Bonifati, A. (2022) Provenance-aware Discovery of Functional Dependencies on Integrated Views. ICDE 2022: 621-633

[4] Zheng, Z., Zheng, L., Alipourlangouri, M., Chiang, F., Golab, L., Szlichta, J., & Baskaran, S. (2022). Contextual Data Cleaning with Ontology Functional Dependencies. ACM Journal of Data and Information Quality (JDIQ), 14(3), 1-26.