**Master’s M2 internship (DataGEMS)**

**Query Planning for Task-Driven Dataset Discovery and Data Exploration.**

The purpose of this internship is to formalize the boundary between dataset discovery and data exploration, and design appropriate task-driven optimizations.

**Co-advisors: Sihem Amer-Yahia (CNRS, Univ. Grenoble Alpes), Silviu Maniu (Univ. Grenoble Alpes)**

**LIG - Work context**

LIG is a 500-member laboratory with teaching faculty, full-time researchers, PhD students, administrative and technical staff. The mission of LIG is to contribute to the development of fundamental aspects of Computer Science (models, languages, methodologies, algorithms) and address conceptual, technological, and societal challenges. The 24 research teams in LIG aim to increase diversity and dynamism of data, services, interaction devices, and use cases influence the evolution of software and systems to guarantee the essential properties such as reliability, performance, autonomy, and adaptability. Research within LIG is organized into 5 focus areas: Intelligent Systems for Bridging Data, Knowledge and Humans, Software and Information System Engineering, Formal Methods, Models, and Languages, Interactive and Cognitive Systems, Distributed Systems, Parallel Computing, and Networks.

**European H2024-INFRA DataGEMS**

DataGEMS is a HORIZON-INFRA-2024-EOSC-01-05 - HORIZON-RIA HORIZON Research and Innovation Actions whose purpose is to build a fully operational and sustainable ecosystem of open-source tools for data FAIRness and provide an ecosystem of Free and Open-Source tools and a number of services covering  all phases of the data lifecycle dealing with storage and management, discovery, analysis, description, publication  and reuse. The project has 12 partners distributed in 8 European countries who will collaborate to develop novel tools and services  to access FAIR-by-design datasets more rapidly than previously. They facilitate the collection and analysis of  heterogeneous and/or large-scale data sets, provide automatic production of FAIR data at the research instruments (e.g., meteorological stations) and support infrastructures by metadata automation tools and techniques.

**Scientific context.** Data is an asset that spurs innovation, drives decision making, improves operations and impacts several domains, including science, environment, health, energy, education, industry, and society as a whole. A growing number of open datasets from governments, academic institutions, and companies bring new opportunities for innovation, economic growth, and societal benefits. From real-time to historical data, from structured data in tabular form to unstructured text, images or videos, data is highly heterogeneous. Moreover, its volume and complexity create a “needle-in-the-haystack” problem: *it is extremely challenging and time-consuming to discover, leverage and combine data within this expanding sea of data*. Data discovery systems, such as Google Datasets, and open data portals, such as the EOSC Portal promise to bring data closer to the users, but fall short for the following reasons: (a) *Limited data discovery capabilities,* (b) *Poor metadata*, (c) *Superficial query answering*, and (d) *Single-table datasets.* Existing tools allow searching for spreadsheets or data published in formats such as CSV or JSON but not complex datasets, e.g., collections of tables, text, or temporal data.

​​To address the above limitations, the DataGEMS project proposes a data discovery platform with Generalized Exploratory, Management, and Search capabilities. DataGEMS is built on the principles of data FAIRness, openness and re-use. It aims to seamlessly integrate data sharing, discovery and analysis into a system that addresses the whole data lifecycle, i.e., sharing, storing, managing, discovering, analyzing and reusing (data and/or metadata), bridging the gap between the data provider and the data consumer.

**Dataset discovery** is the process of identifying and collating datasets to create a new, (potentially virtual) dataset [5]. This may, for example, be done directly through a search, by navigating from related datasets, or by browsing the datasets with a specific annotation. For instance, in education, a teacher who seeks to build a course would use dataset discovery primitives to search for relevant teaching material on a given topic. **Data exploration** is the process of stepwise querying a given dataset. The most common approach for data exploration is the generation of pipelines of operators whose purpose is to transform data to achieve a task. Those operators include drill down and roll up actions but also LLM calls. In the case of the education example, a pipeline would serve a specific learning goal, i.e. a task, such as reducing the learning gap [4]. There are many reinforcement learning methods whose purpose is to generate an exploration policy that yields a pipeline [1,2,3].

**This internship will tie the task-driven need expressed in data exploration to dataset discovery.** Doing that will result in specific operations and optimizations being triggered for dataset discovery. The candidate will tackle the following tasks:

1. Define a set of data exploration tasks in the context of education in conjunction with the advisors and formalize them according to the work in [4]
2. Revisit previous work on exploratory data analysis [1,2,3] use the code to realize the scenarios: <https://github.com/apersonnaz/rl-guided-galaxy-exploration>
3. Review dataset discovery operators and formalize the connection between data exploration operators defined in [1,2,3] and dataset discovery operators defined in [5]
4. Implement the formalism and test it with datasets used in [4]

**This internship may lead to a PhD in the context of a European project.**

**Contact :** "SIHEM AMER YAHIA" <sihem.amer-yahia@univ-grenoble-alpes.fr>; "silviu maniu" <silviu.maniu@univ-grenoble-alpes.fr>;

**References**

[1] Brit Youngmann, Sihem Amer-Yahia, Aurélien Personnaz: Guided Exploration of Data Summaries. Proc. VLDB Endow. 15(9): 1798-1807 (2022)

[2] Behrooz Omidvar-Tehrani, Aurélien Personnaz, Sihem Amer-Yahia: Guided Text-based Item Exploration. CIKM 2022: 3410-3420

[3] Aurélien Personnaz, Sihem Amer-Yahia, Laure Berti-Équille, Maximilian Fabricius, Srividya Subramanian: Balancing Familiarity and Curiosity in Data Exploration with Deep Reinforcement Learning. aiDM@SIGMOD 2021: 16-23

[4] Nassim Bouarour, Idir Benouaret, Sihem Amer-Yahia: Multi-objective Test Recommendation for Adaptive Learning. Trans. Large Scale Data Knowl. Centered Syst.: 1-36 (2024)

[5] Norman W. Paton, Jiaoyan Chen, Zhenyu Wu: Dataset Discovery and Exploration: A Survey. ACM Comput. Surv. 56(4): 102:1-102:37 (2024)