Ph.D. in Information Technology
Thesis Defense

March 6th, 2026
at 14:30 pm
Emilio Gatti Conference Room – building 20

Camilla SANCRICCA – XXXVIII Cycle

Toward Knowledge-Driven Design of Data Preparation Pipelines

Supervisor: Prof. Cinzia Cappiello

Abstract:

Data-centric Artificial Intelligence (AI) has recently emerged as a new paradigm that emphasizes the importance of high-quality data for obtaining reliable results in data science pipelines. This new perspective has recently shifted the focus of the AI research field from model-centric approaches, which were primarily aimed at refining models to optimize their performance, to a data-centric view, which promotes the dynamic improvement of data using an iterative and systematic approach. From this novel perspective, the data preparation phase becomes a critical and even the most important phase of the data science pipeline. A typical data preparation pipeline involves exploration, profiling, and cleaning activities aimed at ensuring that the dataset quality is aligned with the analysis requirements. Despite its fundamental role, data preparation remains one of the most complex, time-consuming, and demanding phases of data science pipelines. Users often struggle to properly recognize a variety of different data quality errors and address them by selecting the right data preparation techniques from the plethora of available ones. This thesis addresses three major research gaps that limit the effectiveness of current approaches that support the data preparation phase of AI pipelines. First, most existing tools adopt fully automated solutions that offer limited transparency, user control, and interactivity. This reduces the interpretability and trustworthiness of the employed tools, especially for non-expert users. Second, current systems are often designed for static data, offering poor support for alternative or heterogeneous data sources increasingly employed in real-world applications (e.g., real-time data). Third, there is a need to consider ethical implications when data science pipelines are used in high-stakes domains, but a unified solution to address both data quality and fairness issues does not exist. For instance, a data scientist developing an AI-based model on a real-world dataset affected by data quality errors (such as missing values, outliers, potential biases, etc.) requires effective support to select and validate an appropriate data preparation pipeline, while maintaining transparency and user control over the process. To overcome these limitations, this thesis presents DIANA, a knowledge-driven, human-centered framework

for adaptive, data-centric AI that supports users in the design and validation of data preparation pipelines. The core component of DIANA is a knowledge base that collects evidence on the impact of data quality errors and the effectiveness of data preparation techniques across diverse datasets and analytics applications. A set of machine learning-based predictors leverages this knowledge to recommend a suitable sequence of data preparation actions for a new dataset. The suggested pipeline is tailored to the input dataset, with its own profile, and the machine-learning model to be executed on that dataset. The system has been designed with a human-centered approach and allows users to interact at all stages of the pipeline and be supported by explainability techniques. User feedback is continuously collected and exploited to improve future recommendations. Moreover, DIANA adopts a sliding autonomy approach, in which the level of provided support and detail of the explanations varies on the basis of users' expertise and preferences. Beyond its user-centric design, DIANA has been extended to operate across heterogeneous data sources. In this thesis, the framework has been adapted to include two representative non-standard data sources: time series and knowledge graphs. For each data source type, the knowledge base is enriched with domain-specific evidence on the impact of data issues and the effectiveness of cleaning strategies. Finally, the framework incorporates ethical considerations by including bias detection and mitigation strategies. A new set of metrics is introduced to measure bias in tabular datasets, and an experimental evaluation on the impact of bias mitigation on fairness and data quality has been conducted to suggest suitable mitigations. Based on these findings, the system can suggest strategies to satisfy the user's analysis goals (e.g., favoring fairness over data quality or vice versa, depending on the application domain). The proposed framework has been validated through experiments concerning both synthetic and real-world datasets, which demonstrate that DIANA's pipeline recommendations consistently improve model performance and pipeline efficiency by reducing the computational effort while ensuring results reliability. Summarizing, this thesis contributes to the data-centric AI field with a knowledge-driven, human-centered, and adaptive solution for the development of a reliable, transparent, and trustworthy system.

PhD Committee

Prof. Pierluigi Plebani, Politecnico di Milano

Prof. Sandra Geisler, RWTH Aachen University

Prof. Matteo Lissandrini, Università degli Studi di Verona