DIANA: a Knowledge-driven Framework for Data-centric AI

Camilla Sancricca

Supervised by: Cinzia Cappiello
Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Politecnico di Milano, Italy

Abstract
Data analysis plays a key role in companies that adopt machine learning models to support their decision-making processes. Among the phases of a machine learning pipeline, data preparation is essential to obtain high-quality data. Data-centric AI shifted the focus of such processes on the quality of data rather than on the machine learning model performance. Users from different application fields face data preparation, and they frequently encounter difficulties in designing effective data preparation pipelines when dealing with a multitude of data quality errors and data quality improvement techniques; this highlights the necessity for approaches to simplify the process of defining an effective data preparation pipeline. The main goal of my Ph.D. project is to design a framework to support users in selecting the data preparation tasks to perform in a machine learning pipeline. Using a knowledge-driven approach, we aim to guide *(more and less experienced)* users through an interactive process in which recommendations, explanations, and different levels of autonomy can simplify the design of an effective data preparation pipeline.

Keywords
Data Quality, Data Preparation, Knowledge-driven Approach

1. Introduction

With the widespread diffusion of a data-driven culture, data analysis is becoming crucial for organizations to gain competitive advantages. The volume and variety of the available data have enabled enterprises to perform data analysis pipelines, employing their results to support their decision-making processes.

Data analysis pipelines include multiple stages: data acquisition, preparation, modeling and analysis, and evaluation. Among them, the most challenging phase is data preparation, which is essential to obtain good pipeline outputs. Indeed, the goal of data preparation is to ensure that the data have a good level of quality, guaranteeing the dependability of the analysis results.

Designing an effective data preparation pipeline has become extremely difficult for users due to various errors and the plethora of available data preparation techniques. For a data scientist, it has been demonstrated that data preparation could take up to 80% of the total data analysis time [1]. Moreover, the majority of the data preparation actions are based on approximate methods; if not performed well, data preparation can introduce a piece of uncertainty in the data. In turn, this uncertainty can also propagate in the final results of the analysis.

Currently, some approaches exist to assist users in designing these pipelines. However, they aim to completely automate the data preparation process without considering *(i)* the benefits of the interaction with users and that *(ii)* the optimal pipeline may change according to different analysis goals, types of data, and user needs.

An emerging concept in this domain is Data-centric AI [2], which is based on the idea that data and their quality are the most important aspects to consider in defining new AI systems. Data-centric AI shifted the primary focus of these systems from the goodness of the model to the quality of the data.

Moreover, DQ is not the only facet to be considered when decision-making processes are employed in contexts that use sensitive data. The outputs of data analysis, in that case, will support decisions that could impact people's lives. The ethical aspect also comes into play: we must ensure that data do not contain biases and that the models we are using are fair.

Another important aspect arising from the adoption of data analysis in many domains is that even less experienced users have started to face problems similar to the ones mentioned before. We also conducted a study interviewing users from different backgrounds and found that the less experienced ones have no idea of the type of analysis to perform once they have data. The need has emerged to enable even non-expert end-users to perform effective data analysis processes [3]. The use of explainability techniques is currently demonstrating its effectiveness in helping less experienced users to have a better understanding of these systems. However, more experienced users probably need to use them in a way that is more self-service than non-experts. For this reason, the concept of sliding or adjustable autonomy is emerging in many research areas, *i.e.,* the ability of a system to involve humans when needed or to proceed autonomously. To achieve these goals, new systems need

Published in the Proceedings of the Workshops of the EDBT/ICDT 2024 Joint Conference (March 25-28, 2024), Paestum, Italy
camilla.sancricca@polimi.it (C. Sancricca)
0000-0002-3820-7870 (C. Sancricca)

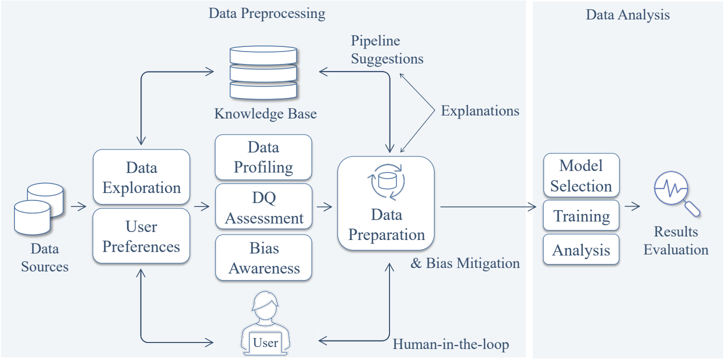
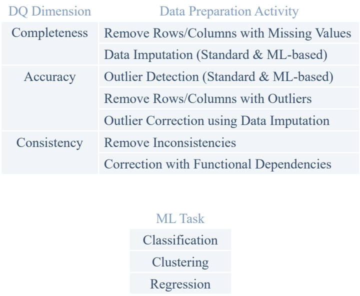


Figure 1: The High-Level Architecture of DIANA



DQ Dimension	Data Preparation Activity
Completeness	Remove Rows/Columns with Missing Values
	Data Imputation (Standard & ML-based)
Accuracy	Outlier Detection (Standard & ML-based)
	Remove Rows/Columns with Outliers
	Outlier Correction using Data Imputation
Consistency	Remove Inconsistencies
	Correction with Functional Dependencies

ML Task
Classification
Clustering
Regression

Figure 2: Setup of the Experiments

to be designed with human-centered approaches. My research is aimed to cover these open issues.

The challenges of my Ph.D. research focus on developing a self-service environment to help, support, and guide users in designing data preparation pipelines. Within this environment, our vision is to provide users with:
(a) recommendations on the best sequence of data preparation actions that best fit their analysis purposes, simplifying the design of a data preparation pipeline
(b) the detection/mitigation of potential biases
(c) explainability, with a human-centered design
(d) more or less support/autonomy *w.r.t.* their needs.

This paper presents the work done in my first Ph.D. year. It is structured as follows. Section 2 describes related work and challenges in data preparation; Section 3 describes the main contributions and presents the high-level architecture of a framework to address the described open challenges; Finally, Section 4 depicts conclusions and future work.

2. Related Work

Several literature contributions focused on data preparation and related challenges in the last few years. As a starting point of my work, I analyzed different frameworks for supporting the data preparation process. Methods that focus on supporting users in designing effective data analysis pipelines are proposed in [4, 5]. However, the type of Machine Learning (ML) algorithm to be run on the data is rarely considered in the recent literature [6]. Instead, Section 3 will show that the type of analysis has a crucial role in the selection of the most suitable pipeline. Some recent tools [6, 7, 8] are based on optimization algorithms that build from scratch the best pipeline, trying several combinations of data preparation tasks; however, extracting the recommendations in this manner can often be very time-consuming. Other contributions [9, 10] use knowledge-based approaches exploiting past users' actions to make recommendations. In my project, we aim to use a knowledge-driven approach, but the recommendations will be based on empirical evidence. Moreover, almost all the above-mentioned methods, like most of the AutoML approaches, aim to automate data preparation keeping the user unaware of how the data was prepared. However, it has been shown the integration of human factors in the data science process achieves more effective and trustworthy systems [11]. For these reasons, there is a need for more human-centered and transparent approaches.

3. Contributions

This section presents my Ph.D. contribution. After analyzing the research gaps related to data preparation and the existing tools in this domain, my research mainly focused on designing a system to address all the open challenges mentioned before. This section aims to present the approach and, for each component, highlight preliminary results obtained and ongoing works.

DIANA is a self-service environment for A**D**apt**I**ve D**A**ta-ce**N**tric **A**I. The main objective of DIANA is to facilitate users in designing a data preparation pipeline by recommending the list of the most appropriate activities to obtain reliable analysis results ***(a)***. The suggestions are based on the user's *analysis context* and are extracted by a Knowledge Base (KB). The *analysis context* is defined as the combination of *(i)* the data source profile, *i.e.,* all the pieces of information that we can extract through data profiling operations, and *(ii)* the type of analysis, *i.e.,* the ML algorithm that the user intends to run on the data.

The system warns users of potential biases, suggesting how to mitigate them ***(b)***. Moreover, it can adapt to expert and less-experienced users, providing more or less support according to their preferences ***(c)***. Human-in-the-loop techniques engage users through the process, and explanations are provided to support those in need

(d). Figure 1 depicts the high-level architecture of the proposed approach. The figure is divided into the two main phases of a data analysis pipeline: *Data Preprocessing* and *Data Analysis*. My Ph.D. research is focused on the *Data Preprocessing* phase.

Data Collection and Exploration The targeted user selects and loads into the system the *Data Sources*. First, the system provides a *Data Exploration* engine with interactive visualizations and allows users to enter some *User Preferences*, *e.g.*, the subset of the most relevant features, the type of analysis to be performed on the data *(if the user already knows it)*, and the needed level of support in the next phases. Depending on the users' level of expertise, they can specify if they want support in the choice of *(1)* possible analysis to perform or *(2)* the data preparation tasks to apply. An expert user could also prefer to perform all the steps autonomously. In the former case, a list of possible analyses will be generated.

Ongoing We are currently working on a methodology that, given a dataset, provides suggestions on suitable analyses by exploiting Large Language Models (LLM).

In the latter case, a suggested pipeline of data preparation actions that best fits the selected analysis will be displayed during the *Data Preparation* phase.

Advanced Data Profiling Once all the data have been collected, they are inspected via the *Data Profiling* and the *DQ Assessment*. The former extracts metadata and visualizations, while the latter assesses the level of DQ. Finally, the *Bias Awareness* phase provides insights into possible biases that could affect the data. The results of these phases should help users understand the datasets' content and their initial suitability for the task at hand.

The Knowledge Base The recommendations about the optimal data preparation pipeline in our framework will be extracted by a KB. We start from the intuition that different error types could impact the final results differently depending on the selected *analysis context*. We envisioned that the best sequence of data preparation actions to apply should depend on such an impact.

Results To verify this approach, we investigate the impact of data errors *(related to different DQ dimensions)* on the results quality of different ML models. We found that issues related to DQ dimensions can have a different impact on the outcome performance of a ML analysis depending on both the ML algorithm used and the characteristics of the data, which defines our *analysis context*. Thus, we perform experiments creating rankings of DQ dimensions for different combinations of datasets and ML models, showing that improving the DQ dimensions in order of importance for that specific *analysis context* gives better final results [12]. Once the DQ dimensions that need to be prioritized have been identified, we focused on extracting, for each combination of dataset profile, ML method, and DQ dimension to improve, the corresponding top-k data preparation actions. We demonstrate that, again, the goodness of the preparation techniques depends on the specific context.

Given the evidence that we can extract suggestions based on such an impact, we defined the final structure of the KB, which contains a static and a dynamic module. The *static KB* mainly contains descriptive and experimental data. Descriptive data concern: *(i)* the DQ dimensions, associated with *(ii)* the data preparation methods which improve them, *(iii)* the list of considered ML models, and *(iv)* profiles of the data sources analyzed in previous experiments. Experimental data are the results of such experiments fundamental to support the generation of suggestions; they contain: *(i)* the sequence of the most impacting DQ dimension and *(ii)* the most suitable data preparation tasks to apply in a multitude of different *analysis context*. The above-mentioned results have been added to the experimental data and will have a fundamental role during the extraction of the suggestions.

Ongoing We are developing the KB conceptual model as a graph database. In addition, we are currently feeding the KB with the empirical knowledge acquired from the experiments, considering a heterogeneous set of datasets taken from several open repositories. Figure 2 shows the initial setup we planned to enrich the KB.

Generating Suggestions The *dynamic KB* takes an unexplored *analysis context* as input and considers all the previously analyzed contexts to extract the suggestions. Our final goal is to identify the previously analyzed context in the KB closer to the one chosen by the user and, accordingly, to extract the ranking of the most impacting DQ dimensions in such a context. This ranking and the data provided by the *Data Profiling* and the *DQ Assessment* engines are the input for identifying the suggested ranking of DQ dimensions. For each DQ dimension, we make use of a classifier that takes as input the results stored in the KB and extracts the best data preparation actions, building the suggested pipeline.

Ongoing We trained a classifier to extract the best data imputation method for improving the Completeness dimension, and we are currently validating it.

Data Preparation The results of the last phase are sent to the *Data Preparation* engine, which has two main goals: showing the most appropriate task to perform and executing the DQ improvement methods. During this phase, depending on what the user has specified at the beginning, a suggested pipeline of preparation actions extracted by the KB could be shown or not to the user. In the envisioned approach, we assume that the users are free to follow the suggestions or not, letting them change the order of the suggested actions, or to substitute them by selecting from all the available ones.

Results We conducted experiments to understand the effect of data preparation techniques on the uncertainty of ML models. We found that the amount of uncertainty introduced is again context-dependent [13].

Bias Mitigation In addition to traditional DQ dimensions and data preparation techniques, we aim to offer a set of bias mitigation techniques to be applied within the *Data Preparation* engine.

Results We had the intuition that a trade-off could exist between the concept of DQ and data ethics. We demonstrate the existence of such a trade-off [14], and we defined a preliminary set of guidelines to balance it [15].

Explainability and Human-in-the-loop Our framework is enriched with explanations tailored to the users' expertise that help them understand the data, the reason behind the suggestions, and the results of the data preparation actions. Moreover, human-in-the-loop techniques involve users in various steps of the process.

Ongoing We are working on enriching the environment to guarantee explainability by extracting and formulating explanations through the support of LLM tools.

4. Conclusion and Future Developments

This paper aims to present the main objectives of my Ph.D. research project. I described the work related to my project, its main challenges, and the high-level architecture of the designed approach. As soon as a preliminary version of the system is finalized, we plan to evaluate it in comparison with similar tools such as [6] and with real users. Future work will focus on exploiting past users' experiences and feedback to improve the recommendations, letting the system evolve and learn. We aim to extend the KB model to include users' profiles, goals, and past actions (*i.e.,* provenance).

References

[1] M. Hameed, F. Naumann, Data preparation: A survey of commercial tools, SIGMOD Rec. (2020).

[2] M. H. Jarrahi, A. Memariani, S. Guha, The principles of data-centric AI, Commun. ACM (2023).

[3] J. M. Hellerstein, J. Heer, S. Kandel, Self-service data preparation: Research to practice, IEEE Data Eng. Bull. (2018).

[4] S. Shrivastava, D. Patel, A. Bhamidipaty, W. M. Gifford, S. A. Siegel, V. S. Ganapavarapu, J. R. Kalagnanam, DQA: scalable, automated and interactive data quality advisor, in: IEEE International Conference on Big Data, Los Angeles, USA, 2019.

[5] L. A. Melgar, D. Dao, S. Gan, N. M. Gürel, N. Hollenstein, J. Jiang, B. Karlas, T. Lemmin, T. Li, Y. Li, S. X. Rao, J. Rausch, C. Renggli, L. Rimanic, M. Weber, S. Zhang, Z. Zhao, K. Schawinski, W. Wu, C. Zhang, Ease.ml: A lifecycle management system for machine learning, in: 11th Conference on Innovative Data Systems Research, CIDR, 2021.

[6] L. Berti-Équille, Learn2clean: Optimizing the sequence of tasks for web data preparation, in: The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, 2019.

[7] F. Neutatz, B. Chen, Y. Alkhatib, J. Ye, Z. Abedjan, Data cleaning and automl: Would an optimizer choose to clean?, Datenbank-Spektrum (2022).

[8] Q. Cui, W. Zheng, W. Hou, M. Sheng, P. Ren, W. Chang, X. Li, Holocleanx: A multi-source heterogeneous data cleaning solution based on lakehouse, in: Health Information Science - 11th International Conference, HIS, 2022.

[9] M. Mahdavi, Z. Abedjan, Semi-supervised data cleaning with raha and baran, in: 11th Conference on Innovative Data Systems Research, CIDR, 2021.

[10] C. Yan, Y. He, Auto-suggest: Learning-to-recommend data preparation steps using data science notebooks, in: Proceedings of the 2020 International Conference on Management of Data, SIGMOD, Portland, OR, USA, 2020.

[11] Ö. Ö. Garibay, B. Winslow, S. Andolina, als., Six human-centered artificial intelligence grand challenges, Int. J. Hum. Comput. Interact. 39 (2023) 391–437.

[12] C. Sancricca, C. Cappiello, Supporting the design of data preparation pipelines, in: Proceedings of the 30th Italian Symposium on Advanced Database Systems, SEBD 2022, Tirrenia (PI), Italy, 2022.

[13] C. Cappiello, F. Cerutti, C. Sancricca, R. Zanelli, About the effects of data imputation techniques on ML uncertainty, in: Joint Proceedings of Workshops at the 49th International Conference on Very Large Data Bases (VLDB 2023), Vancouver, Canada, 2023.

[14] F. Azzalini, C. Cappiello, C. Criscuolo, S. Cuzzucoli, A. Dangelo, C. Sancricca, L. Tanca, Data quality and fairness: Rivals or friends?, in: Proceedings of the 31st Symposium of Advanced Database Systems, Galzingano Terme, Italy, 2023.

[15] F. Azzalini, C. Cappiello, C. Criscuolo, C. Sancricca, L. Tanca, Data quality and data ethics: Towards a trade-off evaluation, in: Joint Proceedings of Workshops at the 49th International Conference on Very Large Data Bases (VLDB 2023), Vancouver, Canada, 2023.