The Future of Sustainable Data Preparation

Barbara Pernici[1,*], Cinzia Cappiello[1], Edoardo Ramalli[1], Matteo Palmonari[2], Federico Belotti[2], Flavio De Paoli[2], Angelo Mozzillo[3], Luca Zecchini[3], Giovanni Simonini[3], Sonia Bergamaschi[3], Tiziana Catarci[4], Matteo Filosa[4], Marco Angelini[4] and Dario Benvenuti[4]

[1]*Politecnico di Milano - DEIB, Milano, Italy*
[2]*Università di Milano-Bicocca, Milano, Italy*
[3]*Università degli Studi di Modena e Reggio Emilia, Modena, Italy*
[4]*Sapienza Università di Roma, Roma, Italy*

Abstract

Data preparation has an important role in data analysis, and it is time and resource-consuming, both in terms of human and computational resources. The "Discount quality for responsible data science" project aims to focus on data-quality-based data preparation, analyzing the main characteristics of related tasks, and proposing methods for improving the sustainability of the data preparation tasks, considering also new emerging techniques based on generative AI. The paper discusses the main challenges that emerged in the initial research work in the project, as well as possible strategies for developing more sustainable data preparation frameworks.

1. Introduction

The technological boost in the capability of analyzing data and reusing it is enormous. The attempt to build data spaces, or data ecosystems, that support the publication and reuse of data for feeding data science pipelines has inspired several initiatives worldwide and in Europe in several application domains. Data scientists specify and then execute pipelines to transform, enrich, and analyze data, passing through exploratory analyses and refinement cycles to control the quality of data and improve the final model. Completely automated pipelines, e.g., AutoML, have shown significant weaknesses in data science life cycles and are often not appreciated by data scientists, because of the difficulty of controlling the results in terms of quality, uncertainty, and explainability. On the other hand, assessing the quality of data and results can be very expensive in terms of computational and human costs. Recently emerging technologies, such as Large Language Models (LLM) are starting to show promising directions to support data analysis and manipulation operations, often triggering a demand based on a "wow effect"; however, applications of LLMs for processing data at moderate or large scales are associated with costs that make their fitness for use still unclear.

In this scenario, the PRIN 2022 project "Discount quality for responsible data science: Human-in-the-Loop (HITL) for quality data" focuses on making the whole process sustainable, both

SEBD 2024: 32nd Symposium on Advanced Database Systems, June 23-26, 2024, Villasimius, Sardinia, Italy
*Corresponding author.
barbara.pernici@polimi.it (B. Pernici)

computationally and in terms of human effort, in all the different phases of data analysis, from data preparation to data analysis and model building, and data exploitation.

Inspired by [1], the project challenges and approaches focus on sustainability aspects both concerning human effort in HILT approaches and on computational aspects when considering task automation, in the direction of effectively using limited resources in the process. Taking inspiration from the successful proposals for a "discount" usability evaluation proposed for usability assessment[1], we propose a "discount" quality evaluation and data preparation approach, based on methods and theories to reduce the annotations and assessment space and to control and decrease both the human computing effort and the use of computational resources. Two main goals will be pursued towards sustainability: i) reducing the computational effort needed to analyze tabular data and knowledge graphs; ii) introducing HITL in a sustainable way, to make human contributions effective, keeping them limited in time and size.

The paper is structured as follows. In Section 2, we discuss the state of the art. Data preparation pipelines are discussed in Section 3, illustrating the main challenges to make them sustainable, while in Section 4 we examine the principal sustainability strategies proposed in the project.

2. Related work

In [2], a systematic approach to developing data science projects is advocated. Several proposals for scientific data ecosystems are emerging, including the European Open Science Cloud EOSC[2], and the scientific debate focuses on the need to include humans in the loop in scientific data analysis while balancing the effort needed to achieve good-quality results.

The project research aims at improving the state of the art in different directions, as follows.

Data ecosystems and data spaces are widely used infrastructures enabling different stakeholders to interact and resolve interoperability issues among shared data [3]. The design of these data collaboratives has posed many sustainability challenges investigated first at the business and organizational level [4], and then translated into technological practices [5]. In this context, prior work has discussed the role of knowledge-driven approaches and related research challenges [6], metadata representation for data science pipelines [7], and requirements to make data FAIR (Findable, Accessible, Interoperable, Reusable) [8]. However, as of today, limited support is provided to help user develop data preparation and analysis pipelines to be integrated into data collaboratives.

As an effective **data preparation** is dependent on the users' goals, some interesting, although task-specific, proposals were recently developed to support HITL for data preparation pipelines and data-centric AI[3], and addressing data quality has become a prerequisite for data analysis, machine learning and crowdsourcing techniques. As discussed in [9, 10] data quality in Big Data presents additional issues for assessing the quality and privacy aspects of data, considering multiple data sources. Ontology-based data management (OBDM) paradigms and automatic annotation evaluation of uncertainty (e.g., [11]) are still open problems, in particular for un-

[1] https://www.nngroup.com/articles/discount-usability-20-years/

[2] https://ec.europa.eu/info/research-and-innovation/strategy/strategy-2020-2024/our-digital-future/open-science_en

[3] https://datacentricai.org/

derstanding unstructured data ([12, 13], and generating latent representation and comparing values in different datasets [14]. In [15], the issue of deciding the type and needed amount of cleaning has been raised focusing textual documents for information retrieval, issue that can be generalized to different types of textual data, including tabular data. In [16], the authors proposed a hybrid human-machine data integration framework for the entity-matching problem. JedAI [17] provides semi-automated pipelines involving data integration and data cleaning, where each component obtains feedback to refine the results of automated analyses. Yet, JedAI focuses on the narrow problem of deduplicating records in databases. BrewER [18, 19] has been proposed to clean only the portion of data useful to satisfy a user's need expressed through a SQL query. Data Civilizer [20] provides an end-to-end big data management system to support data discovery and preparation considering the user's end goal, providing primitives for performing data debugging and workflow visualization. In addition, the need to reduce the amount of computational resources is being emphasized (e.g., Green AI [21]). The emergence of the term "crowd science" [22] shows the need to study all aspects related to human intervention, including the management of scarce resources and user motivation in repetitive tasks such as labeling and manual data quality evaluation. In such activities, a way to estimate and assess human efforts is needed as a basis for reducing them in the development of datasets or during the analysis, still retaining meaningful results. In [23], the challenges of exploratory data analysis and data quality for AI steps are discussed, and a framework for selecting rows and columns, and identifying overlaps is proposed.

Information Visualization and Visual Analytics [24] support the real-time data analysis process, enabling a user to explore the data, parametrize models, investigate results, and hypothesize conclusions [25]. Concerning data preparation, few works coped with it through visual means and human intervention, focusing specifically on data quality aspects. DataPilot [26] is a recent contribution that focuses on visually supporting data preparation activities. The analysis of data quality and performance models has been covered by several works [27, 28] with proposals to allow human intervention through steering by Liu et al. [29]. To make this effort effective, fluency in data analysis and data quality visual exploration by a human user is a must [30]. For this reason, several works have explored how to keep human interaction fluid using visualizations [31], while some others focused on analyzing user traces to inform the system behavior on user intent, using different techniques to model it [32, 33]. The area of data preparation was less subjected to these studies, presenting a gap to fill in the literature.

3. Data preparation pipelines

3.1. General concepts

In the project, we focus on data preparation pipelines, defined as sets of tasks that are applied to a dataset or a data stream to explore the data's potential or improve its quality in the data preparation phase. As the main objective of the project is to improve the sustainability of the data preparation process, in this section, we focus first on the main relevant aspects we plan to consider in the project; then we examine some relevant challenges we are planning to address to provide support to data scientists developing pipelines in a sustainable way.

In data preparation pipelines, starting from the classical approach of achieving a data quality

that is "fit for use", we need to address two main important aspects: the goal of the data preparation and the tasks that can be performed. The preparation can be performed on several **types of data sources**: in the following, we consider data originating from one or more data sources and structured as textual content either as *tabular data* or as a *knowledge graph*.

Concerning the **goal**, several targets can be considered: i) preparation of a *dataset for further reuse*, where the goal is to improve the quality of the dataset in general, considering usual data quality dimensions (e.g., as described in [34]); ii) preparation *of datasets for data analysis*, with either an exploratory data analysis or a well-defined analysis goal; iii) preparation of *datasets for machine learning*, for training, validation, fine-tuning, and testing phases, to improve the quality of the learned model and/or its results. We advocate goal-oriented quality improvement to make data preparation activities more sustainable, i.e., consider the final goal when tailoring data preparation activities.

While pre-processing input data to improve data quality several **data preparation tasks** are considered in this research. We distinguish among: i) *data profiling* tasks, to analyze the characteristics of the data; ii) *data transformation* tasks (including data cleaning, normalization and standardization, merging, splitting, dropping data, data imputation); iii) *data matching* tasks at the instance and schema level (including deduplication, entity matching and/or linking, annotations of columns and column pairs), and *data augmentation* tasks to extend data with data from third-party sources.

A pipeline can be interpreted as a sequence or a more complex workflow of operations to be executed on the data. While in many application scenarios pipelines must be executed by data engineers on large data sets, this large-scale execution is just a final step of a more complex task, which, inspired by requirements collected for data enrichment pipelines, we conceptualize as composed of three phases (see Figure 1).

In the first *exploration phase*, the goal is to 1) understand the fitness for usage in downstream tasks and 2) identify operations that can improve their quality based on the intended usage. Typical actors involved in this phase are *data scientists*, or other professional figures with domain expertise. The second phase consists of the *design* of the engineered pipeline, typically performed by *data engineers*, while the third phase consists of the actual *execution*, which also involves monitoring by operators.

Concerning the tasks in the exploratory phase, examples of questions to be answered are: what are the characteristics of the data? How can the data be transformed? Is it possible to enrich the data via integration with other data (matching and augmentation)? In this phase, users typically consider data samples and need some direct feedback on the results of the operations they explore to understand their effect, making the usage of proper interfaces very valuable. The output of this phase is the definition of a preparation pipeline and the specification of key elements, such as configurations for specific algorithms used within it. Suppose the pipeline needs to be applied to a large amount of data and/or replicated recurrently. In that case, it must usually be engineered to be efficient and, therefore, compliant with big data processing platforms (e.g., using distributed computation), which is the objective of the design phase. While the output of the design phase is the specification of the engineered pipeline, the output of the execution phase is the enriched data.

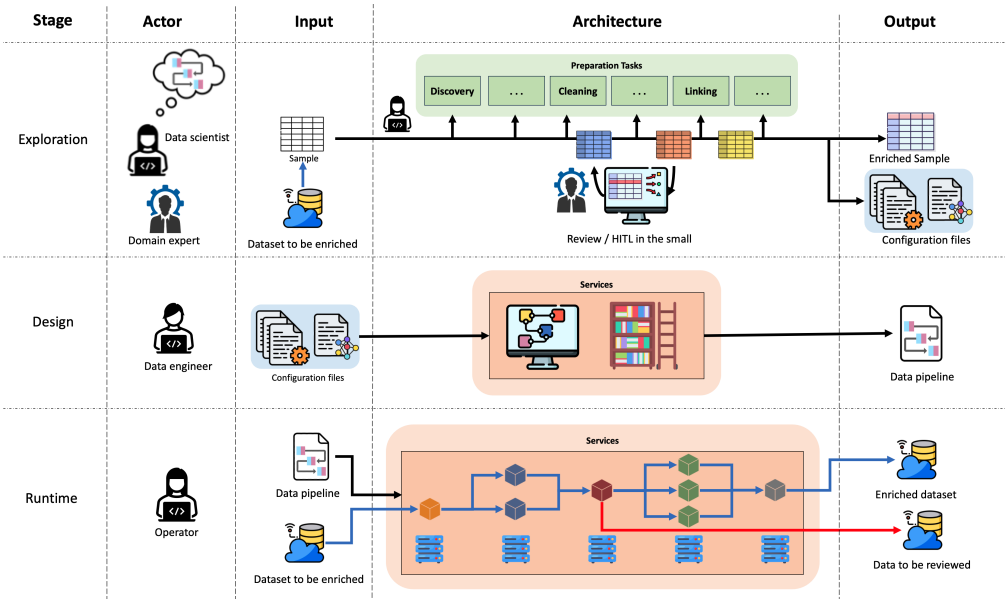

Figure 1: Exploration and data preparation pipelines

3.2. Challenges

Data exploration and the design and execution of data preparation pipelines present several challenges related to their sustainability. Understanding data preparation tasks under the lenses of their sustainability is particularly interesting today, considering the role of ML in downstream analytical modeling and the impact of LLMs (and similar models) on task automation. In the project, we have identified the following challenges for driving further research questions:

Preparation gain: *estimating the likelihood to improve the quality of the goal with data improvement actions with a given approach.*

For instance, systematically applying data cleaning on a dataset is likely to improve the quality of a machine learning model; however, the improvement ratio is difficult to assess and may depend on the selected features or parameters.
A clear understanding of the impact of a data preparation action (possibly on a selected portion of the data) can improve the sustainability of the result, both on the computational side and on the side of human involvement in the process, as some tasks may require human intervention. This understanding also presents other problems to be addressed, such as clearly defining goals and assessing the context.

Sustainable LLM: *Leveraging on LLM in data preparation, combining the power of the latest generation models, e.g., LLMs, with efficiency, scalability, and environmental awareness.*
LLMs or similar models targeting structured and semi-structured data are showing promising performance on several data preparation tasks (e.g., [35, 36, 37, 38]). Careful prompt engineering [39], larger context sizes [40] and orchestration strategies seem to deliver interesting capabili-

ties related to tasks that can be mapped to language generation (e.g., code generation for data transformation, query generation and classification for data augmentation) or even decision/-classification (e.g., deduplication and disambiguation), yet they are extremely expensive and hard to scale. It is still unclear if these recently proposed solutions are still advantageous for, or *even compatible with*, large-scale processing when we consider speed (execution times), costs (infrastructure), and environmental sustainability (carbon emissions) at training and inference time. In general, enhancing the quality of the data consumed by LLMs improves the performances of models fine-tuned with those LLMs [41]. With structured and semi-structured data coming from large corpora employed to train the LLMs, it is unpractical to clean/prepare the entire data, thus the efforts could focus on the portion/tasks that yield the highest benefit for the LLMs. A first challenge is better characterizing the trade-offs between exploiting the power of LLMs' implicit knowledge and preserving efficiency, scalability, and environmental sustainability. A second challenge is finding sweet spots that make LLM usage valuable considering benefit-costs trade-offs, e.g., application to specific data samples.

User understanding: *providing the users with the capability to understand, control, and improve the outcomes of algorithmic decisions in a human-in-the-loop fashion, even when using the latest generation models.*
For example, in entity linking, users aware of the uncertainty associated with links selected by the algorithms can get insights into the quality of the results and revise these results faster [42]. However, solutions to learn from users' actions are still under-explored in several data preparation tasks. Latest models, e.g., those based on LLMs proposed for matching-related tasks [35, 36, 37, 38], do not natively support the interpretation of their decisions in terms of confidence and are very difficult to adapt with a limited number of user feedback. Adequate management of the human-in-the-loop approach represents a challenge, efficiently involving the human user without generating cognitive overload due to too much data to analyze, too broad and unfocused areas of intervention, or not well-supported decisions to make (human-driven versus human-as-reviewer). A challenge also arises in identifying the correct degree of control to provide to the human user, efficiently exploiting human and machine different capabilities.

User Experience Interaction-Driven Optimization: *while users interact with systems involving big data, such as big data visualization systems, their interactions can be recorded and stored in the form of interaction logs; such logs can then be used to capture characteristics of the user intent and to optimize the visualization systems used during the data preparation pipeline.*
Frequently, such logs work like black boxes due to their low-level nature (i.e., the log is explorable, but it contains just low-level atomic interactions like a mouse click or mouse move; it is not straightforward with state-of-the-art techniques to relate it to high-level user actions with reasonable accuracy) and does not provide information on the decision process regarding the user's interactions to the data preparation expert. When analyzing them, the data preparation expert may be overwhelmed and misled, since the info grasped from the logs is too low-level and does not give any insight into the user's intentions. By providing techniques for the extraction of the user's intent, it will be possible to know in advance in which portion of the interaction space and in which phase of the data preparation pipeline it is appropriate to apply

optimizations. Finally, such logs can be exploited to understand at which layer (e.g., data, rendering or interaction) the visualization systems used during the data preparation pipeline fail in maintaining response times low enough to keep the user experience optimal [43].

4. Sustainable strategies

Sustainability can be achieved by reducing time and resources. This can be achieved by reducing the computational complexity of the task execution (e.g., reduction of the volume of the input dataset) or avoiding redundant actions (e.g., reuse of components/information). In the project, we are examining several possible strategies, as follows.

Sustainable Data Preparation Components. Sustainable data preparation ensures that data-driven processes are efficient, effective, and environmentally friendly. Implementing such a strategy involves processing data efficiently from the data collection to their analysis. This implies starting from minimizing the acquired unnecessary data to properly selecting the data preparation components. Improving data quality can already be considered a sustainable action since poor data quality can lead to inefficiencies and resource waste. However, the data preparation components have different characteristics and, therefore, different impacts on the efficiency/effectiveness of the process. The volume and variety of components to consider are high. In [44], it is possible to find a classification of the tasks included in the data preparation pipeline: data discovery, data validation, data structuring, data enrichment, data filtering, and data cleaning. Each category contains a plethora of different functionalities and techniques, and their selection is not easy. The components can differ from different perspectives: scope, execution time, complexity, energy consumption, autonomy level, and effectiveness. The idea is to consider such properties to find the adequate combinations of components able to guarantee the right balance between sustainability and quality of the results.

Pipeline configurations for sustainable data quality. The design of data preparation pipelines is challenging: the data analyst must choose the appropriate operations accounting for several factors. Trial-and-error approaches only sometimes lead to the most effective solution. Instead, a systematic and automatic strategy supported by provenance information can optimize this procedure and lead faster to the desired solution while constantly having feedback from the user [45]. However, while the methodology to build a cost-effective data preparation pipeline is clear, a sustainable method to reuse these pipelines is still missing. This research work aims to propose a strategy that defines pipeline embeddings based on the components' characteristics that add context-aware capabilities. Such an approach can be used for reusing data profiling activities, which is fundamental to be applied for LLM and knowledge graphs.

Data Preparation On-Demand. since the paradigm for data integration is increasingly moving from ETL to ELT, novel *on-demand* solutions are required to efficiently perform data preparation and integration on large amounts of raw data (usually stored in data lakes), cleaning only the portion of data relevant to the downstream task at hand. We aim therefore to work in this direction to provide practitioners with novel tools, as previously done with BrewER [18, 19],

which runs SQL SP queries directly on dirty data through entity resolution on-demand, and Sloth [46], designed to detect duplicate and possibly inconsistent versions of the same table on the Web or in data lakes.

User Experience Sustainable Analysis. Logs collected during the usage of big data visualization systems can be exploited by leveraging generative AI-based techniques on them to extract the user intent. By relating low-level traces to high-level visualization tasks taxonomies [47] it will be possible to capture characteristics of the users' intent during the data preparation pipeline, optimizing the steps requiring the human intervention due to improved and more efficient interaction. In this way, such a pipeline could be refined and optimized, considering the users' intent, allowing the creation of a broader design space for optimization strategies in the other layers (due to the more semantic nature of the user intent with respect to low-level traces). By leveraging these data, it is possible to fine-tune LLMs to support optimizations of the data preparation pipelines the user can select during her work. This strategy, which is linked to the *User Understanding* challenge, asks for providing *explainability* to each user's choice, which will be investigated [48] to enable the inspection and understanding of the process for their derivation and their expected costs and outcomes. Finally, by optimizing the user experience in the visualization systems used during the data preparation pipeline, we can cascade into more favorable outcomes for each of its phases. State-of-the-art approaches [49, 50] tend to mitigate the factor which negatively impacts the most user experience in such systems - high response time [30] - by looking only at the database level. We can exploit log analysis to pinpoint which layer of the visualization system (e.g., data, rendering, interaction) is causing the failure, to highlight which portion of the interaction space tends to lead the system into troubles, and to suggest appropriate optimization techniques.

To conclude, in Table 1, we summarize the main directions that are being explored in the project, individually or in combination, proposing solutions for the named challenges to achieve the different strategies.

	Preparation Gain	*Sustainable LLMs*	*User Understanding*	*User Experience Interaction-Driven Optimization*
Sustainable Data Preparation Components	X	X		
Pipeline configurations for sustainable data quality	X	X		
Data Preparation On-Demand	X	X		X
User Experience Sustainable Analysis	X		X	X

Table 1
Challenges defined for each of the proposed strategies

5. Concluding remarks

Data preparation pipelines make an important contribution both to the required quality of data in different contexts and to the reduction of the amount of resources needed for their use. In the project, we are studying the main challenges to be addressed to achieve the proposed set of strategies to increase the sustainability of data preparation tasks. In particular, we are exploring

several research directions, namely: exploiting the reuse of sustainable pipelines; concentrating on LLMs, both as a case study of resource-hunger application and as a tool for data preparation; increasing the user-in-the-loop role by leveraging on usable visual information exploration approaches.

Acknowledgments

This work has been supported by the PRIN 2022 Project "Discount quality for responsible data science: Human-in-the-Loop for quality data" and by the PNRR-PE-AI "FAIR" project funded by the NextGenerationEU program.

References

[1] J. Nielsen, Applying discount usability engineering, IEEE Software 12 (1995) 98–100.
[2] V. Stodden, The data science life cycle: a disciplined approach to advancing data science as a science, Communications of the ACM 63 (2020) 58–66.
[3] M. I. S. Oliveira, B. F. Lóscio, What is a data ecosystem?, in: Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age, 2018, pp. 1–9.
[4] E. Ruijer, Designing and implementing data collaboratives: A governance perspective, Government Information Quarterly 38 (2021) 101612.
[5] E. Ramalli, B. Pernici, Sustainability and governance of data ecosystems, in: 2023 IEEE International Conference on Web Services (ICWS), IEEE, 2023, pp. 740–745.
[6] S. Geisler, M. Vidal, C. Cappiello, B. F. Lóscio, A. Gal, M. Jarke, M. Lenzerini, P. Missier, B. Otto, E. Paja, B. Pernici, J. Rehof, Knowledge-driven data ecosystems toward data transparency, ACM Journal of Data and Information Quality 14 (2022) 3:1–3:12.
[7] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. M. Wallach, H. D. III, K. Crawford, Datasheets for datasets, Commun. ACM 64 (2021) 86–92. URL: https://doi.org/10.1145/3458723. doi:10.1145/3458723.
[8] A. Jacobsen, R. de Miranda Azevedo, N. S. Juty, D. Batista, S. J. Coles, R. Cornet, M. Courtot, M. Crosas, M. Dumontier, C. T. A. Evelo, C. A. Goble, G. Guizzardi, K. K. Hansen, A. Hasnain, K. M. Hettne, J. Heringa, R. W. W. Hooft, M. Imming, K. G. Jeffery, R. Kaliyaperumal, M. G. Kersloot, C. R. Kirkpatrick, T. Kuhn, I. Labastida, B. Magagna, P. McQuilton, N. Meyers, A. Montesanti, M. van Reisen, P. Rocca-Serra, R. Pergl, S. Sansone, L. O. B. da Silva Santos, J. Schneider, G. O. Strawn, M. Thompson, A. Waagmeester, T. Weigel, M. D. Wilkinson, E. L. Willighagen, P. Wittenburg, M. Roos, B. Mons, E. Schultes, FAIR principles: Interpretations and implementation considerations, Data Intell. 2 (2020) 10–29. URL: https://doi.org/10.1162/dint_r_00024. doi:10.1162/DINT_R_00024.
[9] T. Catarci, M. Scannapieco, M. Console, C. Demetrescu, My (fair) big data, in: J. Nie, Z. Obradovic, T. Suzumura, R. Ghosh, R. Nambiar, C. Wang, H. Zang, R. Baeza-Yates, X. Hu, J. Kepner, A. Cuzzocrea, J. Tang, M. Toyoda (Eds.), 2017 IEEE International Conference on Big Data (IEEE BigData 2017), Boston, MA, USA, December 11-14, 2017, IEEE Computer

Society, 2017, pp. 2974–2979. URL: https://doi.org/10.1109/BigData.2017.8258267. doi:10.1109/BIGDATA.2017.8258267.

[10] D. Ardagna, C. Cappiello, W. Samá, M. Vitali, Context-aware data quality assessment for big data, Future Generation Computer Systems 89 (2018) 548–562. URL: https://doi.org/10.1016/j.future.2018.07.014. doi:10.1016/J.FUTURE.2018.07.014.

[11] G. Scalia, C. A. Grambow, B. Pernici, Y. Li, W. H. G. Jr., Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction, Journal of Chemical Information and Modeling 60 (2020) 2697–2717. URL: https://doi.org/10.1021/acs.jcim.9b00975. doi:10.1021/ACS.JCIM.9B00975.

[12] D. Ritze, O. Lehmberg, C. Bizer, Matching html tables to dbpedia, in: Proceedings of the 5th international conference on web intelligence, mining and semantics, 2015, pp. 1–6.

[13] M. Cremaschi, F. De Paoli, A. Rula, B. Spahiu, A fully automated approach to a complete semantic table interpretation, Future Generation Computer Systems 112 (2020) 478–500.

[14] V. Cutrona, M. Ciavotta, F. De Paoli, M. Palmonari, et al., ASIA: A tool for assisted semantic interpretation and annotation of tabular data, in: CEUR WORKSHOP PROCEEDINGS, volume 2456, CEUR-WS, 2019, pp. 209–212.

[15] D. Roy, M. Mitra, D. Ganguly, To clean or not to clean: Document preprocessing and reproducibility, Journal of Data and Information Quality (JDIQ) 10 (2018) 1–25.

[16] G. Li, Human-in-the-loop data integration, Proceedings of the VLDB Endowment 10 (2017) 2006–2017.

[17] G. Papadakis, G. Mandilaras, L. Gagliardelli, G. Simonini, E. Thanos, G. Giannakopoulos, S. Bergamaschi, T. Palpanas, M. Koubarakis, Three-dimensional entity resolution with jedai, Information Systems 93 (2020) 101565.

[18] G. Simonini, L. Zecchini, S. Bergamaschi, F. Naumann, Entity Resolution On-Demand, Proceedings of the VLDB Endowment (PVLDB) 15 (2022) 1506–1518. doi:10.14778/3523210.3523226.

[19] L. Zecchini, G. Simonini, S. Bergamaschi, F. Naumann, BrewER: Entity Resolution On-Demand, Proceedings of the VLDB Endowment (PVLDB) 16 (2023) 4026–4029. doi:10.14778/3611540.3611612.

[20] E. K. Rezig, L. Cao, M. Stonebraker, G. Simonini, W. Tao, S. Madden, M. Ouzzani, N. Tang, A. K. Elmagarmid, Data civilizer 2.0: A holistic framework for data preparation and analytics, Proc. VLDB Endow. 12 (2019) 1954–1957. URL: http://www.vldb.org/pvldb/vol12/p1954-rezig.pdf. doi:10.14778/3352063.3352108.

[21] R. Schwartz, J. Dodge, N. A. Smith, O. Etzioni, Green AI, Communications of the ACM 63 (2020) 54–63.

[22] D. Ustalov, F. Casati, A. Drutsa, D. Baidakova (Eds.), Proceedings of the Crowd Science Workshop: Remoteness, Fairness, and Mechanisms as Challenges of Data Supply by Humans for Automation co-located with 34th Conference on Neural Information Processing Systems, CSW@NeurIPS 2020, Vancouver, BC, Canada / Online Event, December 11, 2020, volume 2736 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: https://ceur-ws.org/Vol-2736.

[23] H. Patel, S. Guttula, N. Gupta, S. Hans, R. S. Mittal, L. N, A data centric ai framework for automating exploratory data analysis and data quality tasks, ACM Journal of Data and Information Quality (2023).

[24] D. A. Keim, F. Mansmann, J. Schneidewind, J. Thomas, H. Ziegler, Visual analytics: Scope and challenges, Springer, 2008.

[25] L. Battle, P. Eichmann, M. Angelini, T. Catarci, G. Santucci, Y. Zheng, C. Binnig, J.-D. Fekete, D. Moritz, Database benchmarking for supporting real-time interactive querying of large data, in: Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data, 2020, pp. 1571–1587.

[26] A. Narechania, F. Du, A. R. Sinha, R. Rossi, J. Hoffswell, S. Guo, E. Koh, S. B. Navathe, A. Endert, Datapilot: Utilizing quality and usage information for subset selection during visual data preparation, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23, Association for Computing Machinery, New York, NY, USA, 2023. URL: https://doi.org/10.1145/3544548.3581509. doi:10.1145/3544548.3581509.

[27] M. Angelini, C. Daraio, M. Lenzerini, F. Leotta, G. Santucci, Performance model's development: A novel approach encompassing ontology-based data access and visual analytics, Scientometrics 125 (2020) 865–892.

[28] T. Gschwandtner, W. Aigner, S. Miksch, J. Gärtner, S. Kriglstein, M. Pohl, N. Suchy, Timecleanser: a visual analytics approach for data cleansing of time-oriented data, in: Proceedings of the 14th International Conference on Knowledge Technologies and Data-Driven Business, i-KNOW '14, Association for Computing Machinery, New York, NY, USA, 2014. URL: https://doi.org/10.1145/2637748.2638423. doi:10.1145/2637748.2638423.

[29] S. Liu, G. Andrienko, Y. Wu, N. Cao, L. Jiang, C. Shi, Y.-S. Wang, S. Hong, Steering data quality with visual analytics: The complexity challenge, Visual Informatics 2 (2018) 191–197.

[30] Z. Liu, J. Heer, The effects of interactive latency on exploratory visual analysis, IEEE Transactions on Visualization and Computer Graphics 20 (2014) 2122–2131.

[31] A. Ulmer, M. Angelini, J.-D. Fekete, J. Kohlhammer, T. May, A survey on progressive visualization, IEEE Transactions on Visualization and Computer Graphics (2023) 1–18. doi:10.1109/TVCG.2023.3346641.

[32] J. S. Yi, Y. a. Kang, J. Stasko, J. Jacko, Toward a deeper understanding of the role of interaction in information visualization, IEEE Transactions on Visualization and Computer Graphics 13 (2007) 1224–1231. doi:10.1109/TVCG.2007.70515.

[33] D. Benvenuti, M. Filosa, T. Catarci, M. Angelini, Modeling and assessing user interaction in big data visualization systems, in: J. Abdelnour Nocera, M. Kristín Lárusdóttir, H. Petrie, A. Piccinno, M. Winckler (Eds.), Human-Computer Interaction – INTERACT 2023, Springer Nature Switzerland, Cham, 2023, pp. 86–109.

[34] C. Batini, M. Scannapieco, Data and Information Quality - Dimensions, Principles and Techniques, Data-Centric Systems and Applications, Springer, 2016. URL: https://doi.org/10.1007/978-3-319-24106-7. doi:10.1007/978-3-319-24106-7.

[35] X. Deng, H. Sun, A. Lees, Y. Wu, C. Yu, TURL: Table understanding through representation learning, ACM SIGMOD Record 51 (2022) 33–40.

[36] J. Tu, J. Fan, N. Tang, P. Wang, G. Li, X. Du, X. Jia, S. Gao, Unicorn: A unified multi-tasking model for supporting matching tasks in data integration, Proceedings of the ACM on Management of Data 1 (2023) 1–26.

[37] T. Zhang, X. Yue, Y. Li, H. Sun, Tablellama: Towards open large generalist models for tables, arXiv preprint arXiv:2311.09206 (2023).

[38] M. Trabelsi, Z. Chen, S. Zhang, B. D. Davison, J. Heflin, StruBERT: Structure-aware bert for table search and matching, in: Proceedings of the ACM Web Conference, WWW '22, 2022.

[39] M. Mosbach, T. Pimentel, S. Ravfogel, D. Klakow, Y. Elazar, Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 12284–12314. URL: https://aclanthology.org/2023.findings-acl.779. doi:10.18653/v1/2023.findings-acl.779.

[40] Y. Chen, S. Qian, H. Tang, X. Lai, Z. Liu, S. Han, J. Jia, LongloRA: Efficient fine-tuning of long-context large language models, in: The Twelfth International Conference on Learning Representations, 2024. URL: https://openreview.net/forum?id=6PmJoRfdaK.

[41] S. Gunasekar, Y. Zhang, J. Aneja, C. C. T. Mendes, A. Del Giorno, S. Gopi, M. Javaheripi, P. Kauffmann, G. de Rosa, O. Saarikivi, et al., Textbooks are all you need, arXiv preprint arXiv:2306.11644 (2023).

[42] R. Avogadro, M. Ciavotta, F. De Paoli, M. Palmonari, D. Roman, Estimating link confidence for human-in-the-loop table annotation, in: 2023 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2023, pp. 142–149. doi:10.1109/WI-IAT59888.2023.00025.

[43] Z. Liu, J. Heer, The effects of interactive latency on Exploratory Visual Analysis, IEEE Transactions on Visualization and Computer Graphics 20 (2014) 2122–2131. doi:10.1109/TVCG.2014.2346452.

[44] M. Hameed, F. Naumann, Data preparation: A survey of commercial tools, ACM SIGMOD Record 49 (2020) 18–29.

[45] C. A. Bono, C. Cappiello, B. Pernici, E. Ramalli, M. Vitali, Pipeline design for data preparation for social media analysis, ACM Journal of Data and Information Quality 15 (2023) 1–25.

[46] L. Zecchini, T. Bleifuß, G. Simonini, S. Bergamaschi, F. Naumann, Determining the Largest Overlap between Tables, Proceedings of the ACM on Management of Data (PACMMOD) 2 (2024) 48:1–48:26. doi:10.1145/3639303.

[47] M. Brehmer, T. Munzner, A multi-level typology of abstract visualization tasks, IEEE Transactions on Visualization and Computer Graphics 19 (2013) 2376–2385.

[48] B. La Rosa, G. Blasilli, R. Bourqui, D. Auber, G. Santucci, R. Capobianco, E. Bertini, R. Giot, M. Angelini, State of the art of visual analytics for explainable deep learning, in: Computer Graphics Forum, volume 42, Wiley Online Library, 2023, pp. 319–355.

[49] M. Livny, R. Ramakrishnan, K. Beyer, G. Chen, D. Donjerkovic, S. Lawande, J. Myllymaki, K. Wenger, Devise: Integrated querying and visual exploration of large datasets, SIGMOD '97, Association for Computing Machinery, New York, NY, USA, 1997, p. 301–312.

[50] T. Zhang, R. Ramakrishnan, M. Livny, Birch: An efficient data clustering method for very large databases, SIGMOD '96, Association for Computing Machinery, New York, NY, USA, 1996, p. 103–114.