Towards a Holistic Data Preparation Tool

Conference Paper · March 2022

CITATION
1

READS
67

3 authors:

Valerie Restat
FernUniversität in Hagen
2 PUBLICATIONS **1** CITATION
SEE PROFILE

Meike Klettke
University of Regensburg Germany
78 PUBLICATIONS **748** CITATIONS
SEE PROFILE

Uta Störl
University of Hagen
75 PUBLICATIONS **433** CITATIONS
SEE PROFILE

Towards a Holistic Data Preparation Tool

Valerie Restat[1], Meike Klettke[2] and Uta Störl[1]

[1]University of Hagen, Germany
[2]University of Rostock, Germany

Abstract

Data-driven systems and machine learning-based decisions are becoming increasingly important and are having an impact on our everyday lives. The prerequisite for this is good data quality, which must be ensured by preprocessing the data. However, a number of challenges arise in the process. These include the results of the process in terms of data quality, e.g., combating bias and ensuring fairness, and the preprocessing process itself. Here, human involvement and the lack of intelligent solutions and applications for domain experts without in-depth IT knowledge play a major role. This paper summarizes these challenges and provides an overview of the current state of the art. It proposes the design of a holistic tool, along with the necessary tasks to overcome these challenges and to support data preprocessing.

Keywords

data preparation, data quality, data preprocessing, data wrangling, data cleaning

1. Introduction

With the increasing amount of data, the quality of data is decreasing [1]. However, data-driven systems increasingly influence our everyday lives and support us in making decisions [2]. This ranges from search and recommendation services to medical diagnosis, hiring and loans decisions [1, 3]. Thereby, the quality of the decisions depends on the quality of the data [4], which makes ensuring data quality a key issue in big data management and an important aspect of almost every data-driven project [5, 6]. In order to ensure the quality of the data, it is necessary to preprocess them. This preprocessing possesses a number of challenges that will have to be solved by the data management community in the future.

In this paper, we highlight the challenges currently encountered in data preprocessing, with regard to the requirements for the results of the process and the process itself. To address these challenges, we propose the design of a holistic tool and present several tasks that need to be addressed for this purpose in future studies.

The rest of the paper is structured as follows. Section 2 describes the challenges facing data preprocessing. Section 3 depicts the current state of the art. Section 4 presents the necessary tasks that must be addressed in future work to achieve a holistic tool to support data preprocessing. Finally, section 5 summarizes the paper.

Published in the Workshop Proceedings of the EDBT/ICDT 2022 Joint Conference (March 29-April 1, 2022), Edinburgh, UK
valerie.restat@fernuni-hagen.de (V. Restat);
meike.klettke@uni-rostock.de (M. Klettke);
uta.stoerl@fernuni-hagen.de (U. Störl)
0000-0003-0551-8389 (M. Klettke); 0000-0003-2771-142X (U. Störl)

2. Challenges of Data Preprocessing

Data preprocessing, also referred to as data wrangling, data engineering or data preparation, is necessary to ensure data quality. This includes among other steps data profiling, data cleaning, data transformation [7], and data integration. Especially data cleaning is important to improve machine learning based solutions, as shown in [8] and [9]. Thus, we focus mostly on this task in the paper. Data cleaning consists mainly of two components: Error detection and error repairing [10], whereby a distinction can be made between classical approaches and those based on machine learning [7]. Error detection is done e.g. by integrity constraints like functional dependencies or denial constraints, error repairing is done e.g. by domain experts or by using reference data sets [10]. Error repairing is often more difficult because the use of domain experts is time-consuming and expensive, and reference data sets are usually not available in sufficient quality [10]. Hence, data cleaning is a continuous process, the evaluation of which is essential for good quality data [10]. The aspects and challenges of data quality are described in the following section 2.1. Subsequently, the challenges related to the preprocessing process itself are presented in section 2.2.

2.1. Data Quality

Data of good quality is a prerequisite for data analysis and machine learning, which is why the results of data preprocessing must be ensured in terms of quality [11]. Although the definition of data quality varies in the literature, it is undisputed that data quality depends on many different factors and does not only concern accuracy.

In [12], different data quality aspects and definitions

from 1985 to 2009 were studied and 40 dimensions were identified, including timeliness, currency, accuracy and completeness, to name the most referenced.

These are also reflected in [11], in which a hierarchical data quality framework was formulated from the perspective of data users. They identified the following five dimensions:

- *Availability:* This refers to timeliness, also mentioned in [12], and accessibility (also a part of the FAIR principles *Findable*, *Accessible*, *Interoperable* and *Re-usable* [13]).
- *Usability:* This includes documentation, credibility and metadata.
- *Reliability:* Elements of this dimension include the aforementioned aspects of accuracy and completeness, as well as integrity, consistency and auditability.
- *Relevance:* This refers to the fitness of data, which plays a particularly important role in terms of fairness, which will be discussed later in greater detail.
- *Presentation Quality:* The last dimension includes readability and structure. These are crucial to arrive at a valid description of the data to enhance users' understanding of these data.

Another important aspect, which has also gained increasing attention in recent years is the evaluation of data quality in terms of bias and fairness. If a prediction uses a data set that is not representative of the entire population for which predictions are being made, there is a bias in the data [2]. This goes hand in hand with the aforementioned fitness of the data. In addition, prejudices, preconceptions, and various historical perceptions may be contained in data [2]. This problem is amplified when biased data are used for algorithms whose consequently biased output is in turn used for further predictions [2].

Bias can also be introduced into the data through preprocessing. This can be caused by several variants during the data cleaning or data filtering step, such as an accidental introduction of bias by methods for missing value imputation, as shown in [3]. The example presented here considers a form that offers a binary gender choice, and the option of not specifying gender. Suppose about half of the respondents were women and half men, but women were more likely than men not to report their gender and some respondents would also identify themselves as non-binary. If mode imputation were applied now, all unspecified values would be set as male, thus skewing the distribution and even excluding non-binary individuals. There are various other examples of this, making it clear that ways to improve data quality and control bias are an important aspect of data preprocessing [3].

2.2. Workflow Properties

In addition to the challenges posed by data quality, practitioners face many challenges in the data preprocessing workflow itself. In [14], the results of a user survey of data analysts and infrastructure engineers show that data cleaning is time-consuming and needs *human involvement*. Participants described data cleaning as an *iterative* approach in which evaluation is not automated but merely ad-hoc. In addition, a discrepancy emerged between the data analysts and the infrastructure engineers, who define data quality differently. While the engineers tended to look at errors in data at the syntactic level, e.g., incorrect schemas or inconsistencies, that require a precise definition of the errors, analysts tended to investigate semantic errors through domain expertise, which is more difficult to translate into clear rules.

A variety of tools exist to automate data cleaning. In [15], the authors examined multiple tools using real-world data sets. The results of the examination show that there is *no single dominant tool*, as most tools are only suitable for a certain type of error, while a variety of different errors occur in real-world data. A holistic multi-tool strategy improved the results, but still failed to achieve acceptable error coverage. The authors thus also emphasize the need for human involvement. Further, they highlight the need for real-world data sets for the development and testing of new approaches as well as advances in combination of data cleaning tools.

Data cleaning tools were also investigated in [16] and a survey was conducted for this purpose. The results show that most tools already require a preprocessed and cleaned data set as input, e.g. with uniform delimiters or the same number of fields per row. It is also shown here that human involvement is necessary, both *domain knowledge and IT-knowledge are required* for the usability of the tools. As in [15], the *lack of intelligent solutions* is addressed and the need for automated data preparation tasks and pipelining is highlighted.

As mentioned earlier, the increasing amount of data generates additional challenges to data preprocessing, especially in terms of *volume* and *variety*, also shown in [1]. It is noted here that due to the ever-increasing amount of data, quality is compromised and also many data cleaning tools do not scale sufficiently. In terms of variety, the challenges not only arise from diverse data formats but also from the variety of different errors. Moreover, repairing the data is difficult due to different constraints and can also possibly lead to new errors. The need for domain experts is underlined here as well.

Another important aspect is to create metadata during preprocessing and to track and document data provenance [17]. The latter is especially important for *reproducibility*. The terminology is not standardized [18], so we adopt ACM's definition [19]. It states that an exper-

iment is reproducible if a different team with the same measurement procedure and setup can obtain the measurements under the same operational conditions.

In [20], the iterative nature of data preprocessing is mentioned as a particular challenge for the reproducibility of such pipelines. Furthermore, it is described that it is important to understand *data lineage* in order to ensure reproducibility. This also includes the identification of errors and the possibilities of a rollback, using data lineage.

The importance of reproducibility is also emphasized in [21], and with it the need for new algorithms to be comparable. It also states that the reproducibility of data preprocessing in particular has so far received little attention, compared to other areas. So the importance of a well-defined data preprocessing process is highlighted.

3. State of the Art

In this chapter, we describe the current state of the art in data preprocessing tools. A number of tools already exist, some of which have been studied in [16]: Altair Monarch Data Preparation[1], Paxata Self Service Data Preparation[2], SAP Agile Data Preparation[3], SAS Data Preparation[4], Tableau Prep[5], Talend Data Preparation[6] and Trifacta Wrangler[7]. Initially, 42 commercial tools were selected, of which these seven were chosen based on following criteria: Tools specific to the data preprocessing task, comprehensive coverage of 40 features, guides and documentation, availability of a trial version, a GUI, and customer support. The evaluation of the seven tools based on three data sets and 40 features, classified into the following six categories: data discovery, data validation, data structuring, data enrichment, data filtering, and data cleaning. The results show that there is no tool that can cover all these features. As described in section 2.2, all tools require a pre-preprocessing as well, and the authors describe data preprocessing as a mainly manual task performed by domain experts with data engineering knowledge.

Other systems for data cleaning are described also in literature, some of which are mentioned in [22]. BoostClean [23] detects and repairs domain value violations. As input, it requires a relational table, libraries of functions for detecting and repairing errors, and a user-specified classifier training procedure. The system relies on boosting to maximize the performance of a downstream machine learning model. HoloClean [24] is a system for automatic error repair that uses, in addition to the data set to be cleaned, integrity constraints, external data, and matching dependencies as input to create a probabilistic model, which suggests a cleaned data set based on statistical learning and probabilistic inference. HoloDetect [25], an error detection system, is related to this. In addition to the data set to be cleaned, a training data set and optional denial constraints are required as input. Through data augmentation, the training data set is expanded and a machine learning model is trained that classifies whether a cell is faulty or not. Raha [26] and Baran [27] are systems for error detection and error repairing, respectively. They do not require configuration and thus reduce human involvement. Instead, the configuration is done automatically. Raha is based on clustering to train a binary classifier that predicts for each cell whether it is clean or dirty. Baran can subsequently be used for error repairing. In an optional offline phase, external sources with value-based corrections can be used to pre-train error corrector models. In the online phase, the error corrector models are updated and used to generate potential corrections where a binary classifier predicts if it is an actual correction. TFX [28] is a machine learning platform with data analysis and data validation capabilities. It relies on a schema to detect errors and suggest possible fixes. Further, the schema and its different versions can be used to analyze the evolution of the data.

However, even if especially Raha and Baran reduce human involvement, it is mostly needed in the other tools and IT knowledge is partly necessary in addition to domain knowledge. Moreover, none of these systems can be considered a holistic tool for data preprocessing. They focus only on a certain aspect of data cleaning (e.g., only error detection) and only on certain types of errors (e.g., domain value violations). The evaluation is mostly done on only one or a few aspects of data quality (mostly accuracy), fairness for example is not considered in any of the systems, neither is reproducibility. Moreover, they are only suitable for tabular data; semi-structured or unstructured data cannot be used. Yet there is an increasing number of formats like JSON or text data. The data quality of semi-structured and unstructured data need to be better explored in the future [29].

4. Holistic Data Preparation Tool

In line with the challenges described in section 2, we now outline tasks that the community will have to address in the future. We propose the design of a holistic tool to support domain experts in data preprocessing, which must overcome the challenges mentioned.

[1] https://www.altair.com/monarch/
[2] https://www.paxata.com/self-service-data-prep/
[3] https://www.sap.com/germany/products/database-data-management.html
[4] https://www.sas.com/en_us/software/data-preparation.html
[5] https://www.tableau.com/products/prep
[6] https://www.talend.com/products/data-preparation/
[7] https://www.trifacta.com/products/why-trifacta/

4.1. Data Quality

In section 2.1, we described the challenges that arise in connection with data quality. Here we see the following tasks:

Ensuring fairness and explainability As mentioned in section 2.1, data quality plays an essential role. Combating bias and ensuring fairness are an important aspect of preprocessing. This especially applies to automated solutions [3]. Therefore, we would particularly like to highlight the need for such a tool to help data scientists identify biases and unfairness present in the data or arising during preprocessing. For the former, there should be tool support for extensive testing; for the latter, the data changes need to be measured as described in [4]. To date, there are no standard methods for measuring data changes. Developing appropriate metrics and descriptions are important research topics in future studies [4].

The same applies to explainability, an issue often associated with fairness. Even though many studies consider the explainability of machine learning models themselves, many decisions that affect the behavior of the models are made in preprocessing [30]. Logging and measuring data changes through benchmarks and analysis of algorithm characteristics can help establish explainability and examine preprocessing steps in terms of introducing bias [4]. Moreover, consumer labels such as those envisioned for machine learning models in [31] could also support combating bias and ensuring fairness and explainability [4].

Nevertheless, the topic of fairness is a very complex one and difficult to grasp. For example, in [32], almost twenty different types of bias were presented. Moreover, most fairness metrics do not consider the social consequences of decisions based on predictions by machine learning models [2]. In [33], it is argued that current fairness metrics and research in ethical AI are not sufficient. Many practical issues need to be considered, several of which have been presented here. Nevertheless, the proposed approaches will provide initial help to address this complex issue, since social-minded measures are crucial for data quality.

Comprehensive integrated evaluation As described in section 2, the evaluation of preprocessing is usually difficult, especially due to its iterative nature. According to the before mentioned aspect to ensure fairness and measure data changes, we propose a comprehensive evaluation related to all dimensions of data quality, described in [11] and presented in this paper in section 2.1.

- *Availability:* Data accessibility comes into play before preprocessing and is therefore not considered here. With regard to timelessness, checks can be made according to the arrival time and intervals of the data.
- *Usability:* To ensure credibility, in addition to verification by domain experts, automated checks could be made according to the range of data or accepted values.
- *Reliability:* At this point, an evaluation according to precision and recall could take place as well as the examination of integrity constraints and functional dependencies and the adherence to formats. For recall and precision ground truth is required, but often not available.
- *Relevance:* Whether data is suitable for a use case primarily depends on the goal of the analysis or prediction and must be assessed by a domain expert. This expert can be supported in the decision-making process by providing automated information about the distribution of the data, warning of protected features and fairness metrics. As explained above, an audit based on fairness metrics is not sufficient for an assessment according to ethical AI, but it would be a starting point to counteract unfairness and bias. Measuring changes in data can further support this, as well as the consumer labels mentioned.
- *Presentation Quality:* This also requires human involvement and can be supported by automated checks based of certain standards and specifications.

In order to take a first step in this direction and to create a basis for working on the aspects mentioned above, we are extending the data generator implemented as part of the EvoBench project [34]. Since data sets are essential for testing new approaches, we aim to generate *test data for evaluating data preparation pipelines*. Therefore, we generate targeted data sets with specific error types that are as close to reality as possible. These can then be used to run tests and evaluate approaches.

4.2. Workflow Properties

In section 2.2, we described the challenges currently facing the preparation process itself. In this context, we see the following tasks:

High-level application and abstraction from IT-knowledge One of the most frequently cited challenges relates to human involvement and required user expertise. Human involvement by domain experts is indispensable in data preprocessing. To make their work as easy as possible and to overcome the aforementioned discrepancy between data analysts and infrastructure engineers, domain experts must be able to use tools without in-depth IT-knowledge.

Hence, we suggest that the tool should have high-level functions that can be applied without IT-knowledge. First approaches in this direction could be:

- Visual and interactive approaches
- Approaches based on natural language processing
- Example-oriented approaches

These are intended to help select the appropriate preprocessing method according to the data and the objective and provide an intelligent guidance for data cleaning. To support users in choosing the appropriate preprocessing methods, the aforementioned consumer labels [31] are conceivable, for example.

Minimizing human involvement Even though human involvement is indispensable, it is very expensive. Therefore, the manual effort of data preprocessing must be reduced as much as possible. This underlines the need for intelligent and automated solutions that combine individual tools in a pipeline and optimize it, as referred to in [15] and [16].

For this purpose, the usage of machine learning techniques is conceivable, for example. This could be applied to automatically predict the default configuration settings of tools and pipelines, similar to the automated tuning of database systems, e.g. in [35]. Raha [26] and Baran [27] are, as described, first such approaches in this direction. Alternatively, the use of dictionaries or knowledge stores could be a possible solution.

Data Lineage In the challenges mentioned in section 2.2 the need for data lineage and the associated reproducibility of data preprocessing was emphasized. For this reason, the proposed tool should provide data lineage tracking capabilities, as well as rollback capabilities. Despite the iterative nature of data preprocessing, in addition to being measured, changes must be tracked and, if necessary, reversed. Furthermore, results of the preprocessing pipeline ought to be *reproducible*. This once again highlights the need for tracking data lineage, as the raw data and ready preprocessed data alone are not sufficient for *reproducibility*, as shown in [21].

This means that in addition to these data, the preprocessing process itself also needs to be documented and a description must be provided which algorithms were used to modify data and which parameters were applied. One of the next research questions we want to explore will be how best to accomplish this.

Scalability In section 2.2, the challenges posed by growing data volumes have been described. To cope with the increasing amount of data and the resulting challenges mentioned, the proposed tool and algorithms used need to be scalable. As described in [1], this is not the case with most current tools.

Support of semi-structured and unstructured data As described in section 3, most tools are only suitable for relational data. The data quality of semi-structured and unstructured data needs to be better explored as well as tools developed for such data formats.

This will also be one of the research questions that we want to investigate next.

5. Conclusion and Future Work

In this paper, we have identified the challenges that arise in data preprocessing and, depending on them, presented a series of tasks that need to be addressed in future work. For maximum benefit, the different aspects must be combined, which is why we have proposed the concept of a holistic tool.

Furthermore, to produce a holistic data preparation tool, we intend to investigate the following research questions in future work:

- Which preprocessing algorithms are suitable for which data?
- Which of these algorithms are covered by which existing tools?
- What is the most efficient way to combine different data preprocessing tools in a pipeline in terms of accuracy and to minimize false positives?
- How can the pipeline be evaluated automatically, with respect to all the aspects of data quality mentioned above?
- How can the data quality of semi-structured and unstructured data be ensured?
- How can data changes be reliably measured?
- What measures need to be taken to detect bias in the data and counteract unfairness?
- How can data lineage, especially in iterative processes, be reliably tracked and reproducibility ensured?
- How can machine learning be used to automatically preprocess data or generate suggestions for domain experts?

References

[1] F. Ridzuan, W. M. N. Wan Zainon, A Review on Data Cleansing Methods for Big Data, Procedia Computer Science (2019) 731–738. URL: https://www.sciencedirect.com/science/article/pii/S1877050919318885, the Fifth Information Systems International Conference, 23-24 July 2019, Surabaya, Indonesia.

[2] E. Pitoura, Social-minded Measures of Data Quality: Fairness, Diversity, and Lack of Bias, ACM J. Data Inf. Qual. (2020) 12:1–12:8. URL: https://dl.acm.org/doi/10.1145/3404193.

[3] S. Schelter, J. Stoyanovich, Taming technical bias in machine learning pipelines, IEEE Data Engineering Bulletin (Special Issue on Interdisciplinary Perspectives on Fairness and Artificial Intelligence Systems) (2020) 39–50.

[4] M. Klettke, A. Lutsch, U. Störl, Kurz erklärt: Measuring data changes in data engineering and their impact on explainability and algorithm fairness, Datenbank-Spektrum (2021). URL: https://doi.org/10.1007/s13222-021-00392-w.

[5] F. Endel, H. Piringer, Data Wrangling: Making data useful again, IFAC-PapersOnLine (2015) 111–112. URL: https://www.sciencedirect.com/science/article/pii/S2405896315001986, 8th Vienna International Conferenceon Mathematical Modelling.

[6] W. Fan, F. Geerts, Foundations of Data Quality Management, Morgan & Claypool Publishers, 2012. URL: https://doi.org/10.2200/S00439ED1V01Y201207DTM030.

[7] M. Klettke, U. Störl, Four Generations in Data Engineering for Data Science: The Past, Presence and Future of a Field of Science, Datenbank-Spektrum (2021). URL: https://doi.org/10.1007/s13222-021-00399-3.

[8] M. Mahdavi, et al., Towards Automated Data Cleaning Workflows, in: Proceedings of the Conference on "Lernen, Wissen, Daten, Analysen", Berlin, Germany, September 30 - October 2, 2019, CEUR-WS.org, 2019, pp. 10–19. URL: http://ceur-ws.org/Vol-2454/paper_8.pdf.

[9] P. Li, et al., CleanML: A Study for Evaluating the Impact of Data Cleaning on ML Classification Tasks, in: 37th IEEE International Conference on Data Engineering, ICDE 2021, Chania, Greece, April 19-22, 2021, IEEE, 2021, pp. 13–24. URL: https://doi.org/10.1109/ICDE51399.2021.00009.

[10] I. F. Ilyas, Effective Data Cleaning with Continuous Evaluation, IEEE Data Eng. Bull. (2016) 38–46. URL: http://sites.computer.org/debull/A16june/p38.pdf.

[11] L. Cai, Y. Zhu, The Challenges of Data Quality and Data Quality Assessment in the Big Data Era, Data Sci. J. (2015) 2. URL: https://doi.org/10.5334/dsj-2015-002.

[12] F. Sidi, et al., Data quality: A survey of data quality dimensions, in: 2012 International Conference on Information Retrieval & Knowledge Management, Kuala Lumpur, Malaysia, March 13-15, 2012, IEEE, 2012, pp. 300–304. URL: https://doi.org/10.1109/InfRKM.2012.6204995.

[13] M. D. Wilkinson, et al., The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data (2016) 160018. URL: https://doi.org/10.1038/sdata.2016.18.

[14] S. Krishnan, et al., Towards reliable interactive data cleaning: a user survey and recommendations, in: Proceedings of the Workshop on Human-In-the-Loop Data Analytics, HILDA@SIGMOD 2016, San Francisco, CA, USA, June 26 - July 01, 2016, ACM, 2016, p. 9. URL: https://doi.org/10.1145/2939502.2939511.

[15] Z. Abedjan, et al., Detecting Data Errors: Where are we and what needs to be done?, Proc. VLDB Endow. (2016) 993–1004. URL: http://www.vldb.org/pvldb/vol9/p993-abedjan.pdf.

[16] M. Hameed, F. Naumann, Data Preparation: A Survey of Commercial Tools, SIGMOD Rec. (2020) 18–29. URL: https://doi.org/10.1145/3444831.3444835.

[17] C. A. Goble, et al., FAIR Computational Workflows, Data Intell. (2020) 108–121. URL: https://doi.org/10.1162/dint_a_00033.

[18] W. Mauerer, S. Scherzinger, Nullius in Verba: Reproducibility for Database Systems Research, Revisited, in: 37th IEEE International Conference on Data Engineering, ICDE 2021, Chania, Greece, April 19-22, 2021, IEEE, 2021, pp. 2377–2380. URL: https://doi.org/10.1109/ICDE51399.2021.00270.

[19] ACM, Artifact review and badging – version 2.0, 2021. URL: https://www.acm.org/publications/policies/artifact-review-badging.

[20] L. Rupprecht, et al., Improving Reproducibility of Data Science Pipelines through Transparent Provenance Capture, Proc. VLDB Endow. (2020) 3354–3368. URL: http://www.vldb.org/pvldb/vol13/p3354-rupprecht.pdf.

[21] M. Pawlik, et al., A Link is not Enough - Reproducibility of Data, Datenbank-Spektrum (2019) 107–115. URL: https://doi.org/10.1007/s13222-019-00317-8.

[22] M. Boehm, A. Kumar, J. Yang, Data Management in Machine Learning Systems, Morgan & Claypool Publishers, 2019. URL: https://doi.org/10.2200/S00895ED1V01Y201901DTM057.

[23] S. Krishnan, et al., BoostClean: Automated Error Detection and Repair for Machine Learning, CoRR (2017). URL: http://arxiv.org/abs/1711.01299. arXiv:1711.01299.

[24] T. Rekatsinas, et al., HoloClean: Holistic Data Repairs with Probabilistic Inference, CoRR (2017). URL: http://arxiv.org/abs/1702.00820. arXiv:1702.00820.

[25] A. Heidari, et al., HoloDetect: Few-Shot Learning for Error Detection, in: Proceedings of the 2019 International Conference on Management of Data, SIGMOD Conference 2019, Amsterdam, The Netherlands, June 30 - July 5, 2019, ACM, 2019, pp. 829–846. URL: https://doi.org/10.1145/3299869.

3319888.

[26] M. Mahdavi, et al., Raha: A Configuration-Free Error Detection System, in: Proceedings of the 2019 International Conference on Management of Data, SIGMOD Conference 2019, Amsterdam, The Netherlands, June 30 - July 5, 2019, ACM, 2019, pp. 865–882. URL: https://doi.org/10.1145/3299869.3324956.

[27] M. Mahdavi, Z. Abedjan, Baran: Effective Error Correction via a Unified Context Representation and Transfer Learning, Proc. VLDB Endow. (2020) 1948–1961. URL: http://www.vldb.org/pvldb/vol13/p1948-mahdavi.pdf.

[28] D. Baylor, et al., TFX: A TensorFlow-Based Production-Scale Machine Learning Platform, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017, ACM, 2017, pp. 1387–1395. URL: https://doi.org/10.1145/3097983.3098021.

[29] X. Chu, et al., Data Cleaning: Overview and Emerging Challenges, in: Proceedings of the 2016 International Conference on Management of Data, SIGMOD Conference 2016, San Francisco, CA, USA, June 26 - July 01, 2016, ACM, 2016, pp. 2201–2206. URL: https://doi.org/10.1145/2882903.2912574.

[30] C. V. G. Zelaya, Towards Explaining the Effects of Data Preprocessing on Machine Learning, in: 35th IEEE International Conference on Data Engineering, ICDE 2019, Macao, China, April 8-11, 2019, IEEE, 2019, pp. 2086–2090. URL: https://doi.org/10.1109/ICDE.2019.00245.

[31] C. Seifert, S. Scherzinger, L. Wiese, Towards Generating Consumer Labels for Machine Learning Models, in: 2019 IEEE First International Conference on Cognitive Machine Intelligence (CogMI), Los Angeles, CA, USA, December 12-14, 2019, IEEE, 2019, pp. 173–179. URL: https://doi.org/10.1109/CogMI48466.2019.00033.

[32] N. Mehrabi, et al., A Survey on Bias and Fairness in Machine Learning, ACM Comput. Surv. (2021) 115:1–115:35. URL: https://doi.org/10.1145/3457607.

[33] J. Chen, V. Storchan, E. Kurshan, Beyond Fairness Metrics: Roadblocks and Challenges for Ethical AI in Practice, CoRR (2021). URL: https://arxiv.org/abs/2108.06217. arXiv:2108.06217.

[34] A. Conrad, et al., EvoBench: Benchmarking Schema Evolution in NoSQL, in: Performance Evaluation and Benchmarking - 13th TPC Technology Conference, TPCTC 2021, Copenhagen, Denmark, August, 2021, Springer, 2021, pp. 33–49. URL: https://doi.org/10.1007/978-3-030-94437-7_3.

[35] D. V. Aken, et al., An Inquiry into Machine Learning-based Automatic Configuration Tuning Services on Real-World Database Management Systems, Proc. VLDB Endow. (2021) 1241–1253. URL: http://www.vldb.org/pvldb/vol14/p1241-aken.pdf.