PDF Download
3534678.3542604.pdf
02 March 2026
Total Citations: 17
Total Downloads: 893

Published: 14 August 2022

Citation in BibTeX format

KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining
August 14 - 18, 2022
Washington DC, USA

Conference Sponsors:
SIGMOD
SIGKDD

ABSTRACT

Advances in Exploratory Data Analysis, Visualisation and Quality for Data Centric AI Systems

HIMA PATEL, IBM Research, Yorktown Heights, NY, United States

SHANMUKHA GUTTULA, IBM Research, Yorktown Heights, NY, United States

RUHI SHARMA MITTAL, IBM Research, Yorktown Heights, NY, United States

NARESH MANWANI, International Institute of Information Technology, Hyderabad, Hyderabad, TG, India

LAURE BERTI-EQUILLE, Research Institute for Development, Marseille, France

ABHIJIT MANATKAR, International Institute of Information Technology, Hyderabad, Hyderabad, TG, India

Open Access Support provided by:

IBM Research

International Institute of Information Technology, Hyderabad

Research Institute for Development

KDD '22: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (August 2022)
https://doi.org/10.1145/3534678.3542604
ISBN: 9781450393850

Advances in Exploratory Data Analysis, Visualisation and Quality for Data Centric AI Systems

Hima Patel
himapatel@in.ibm.com
IBM Research India

Shanmukha Guttula*
shagutt1@in.ibm.com
IBM Research India

Ruhi Sharma Mittal*
ruhi.sharma@in.ibm.com
IBM Research India

Naresh Manwani
naresh.manwani@iiit.ac.in
International Institute of Information Technology, Hyderabad

Laure Berti-Equille
laure.berti@ird.fr
Institut de Recherche pour le Développement, France

Abhijit Manatkar
abhijit.manatkar@students.iiit.ac.in
International Institute of Information Technology, Hyderabad

ABSTRACT

It is widely accepted that data preparation is one of the most time-consuming steps of the machine learning (ML) lifecycle. It is also one of the most important steps, as the quality of data directly influences the quality of a model. In this tutorial, we will discuss the importance and the role of exploratory data analysis (EDA) and data visualisation techniques to find data quality issues and for data preparation, relevant to building ML pipelines. We will also discuss the latest advances in these fields and bring out areas that need innovation. To make the tutorial actionable for practitioners, we will also discuss the most popular open-source packages that one can get started with along with their strengths and weaknesses. Finally, we will discuss on the challenges posed by industry workloads and the gaps to be addressed to make data-centric AI real in industry settings.

ACM Reference Format:
Hima Patel, Shanmukha Guttula, Ruhi Sharma Mittal, Naresh Manwani, Laure Berti-Equille, and Abhijit Manatkar. 2022. Advances in Exploratory Data Analysis, Visualisation and Quality for Data Centric AI Systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3534678.3542604

1 INTENDED AUDIENCE

The tutorial is of relevance to data scientists, ML researchers, practitioners who are facing data quality issues and need principled methods and tools to prepare their data.

2 OUTLINE

In the first two sections, we will discuss the importance and need of exploratory data analysis and data visualisations respectively for data quality analysis and data preparation for ML. While doing so, we will describe some of the state-of-the-art techniques in these areas and also highlight their limitations with a series of illustrative examples. In particular, we review works which describe systems that can play an assistive role to a data analyst by, for e.g., automating the processes like EDA and data visualization which can help with the discovery data quality issues, or by generating recommendations to users of specific actions to take. In the last section, we will focus on the challenges posed by industry workloads, efforts in overcoming the challenges and open areas.

2.1 Role of EDA in Data Understanding and Data Quality

EDA techniques are used by data scientists to analyze and get key insights into the data. This is also a field where data scientists' skills and domain knowledge play a huge role in performing effective EDA. In this tutorial, we will first look at how EDA helps in measuring and improving data quality as mentioned in [13]. We will discuss how EDA can be a useful tool not only for identifying data quality issues in a dataset but also for obtaining a general understanding of the dataset which can aid in making decisions related to algorithmic/architectural choices for models that can be fit on the data to perform meaningful inference at different tasks, as well as guiding pre-processing operations to execute on the data.

Next, we will look at recent advances made in automating EDA processes either partially or fully to help data scientists perform effective EDA irrespective of their skill and experience. We look at three approaches to EDA automation:

- Guided/Interactive Data Exploration systems which help the user by suggesting interesting data components based on the user's interaction with the system [17] [8].
- Next Step Recommendation Systems which suggest user what next exploratory action to apply based on the user's current EDA session, data being explored and previous EDA sessions [18] [3].
- End-End Automated EDA systems which produce full EDA sessions based on previous EDA sessions and dataset characteristics using techniques like Reinforcement Learning [9] [20].

We further discuss limitations of these current approaches. Lastly, we look at some open source tools (e.g., [19]) that help users in doing effective EDA focused on data quality.

[*] Equal Contribution

KDD '22, August 14–18, 2022, Washington, DC, USA
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9385-0/22/08.
https://doi.org/10.1145/3534678.3542604

2.2 Data Visualisation and its Role in Data Quality Analysis

Data visualisation is a key EDA technique which makes the analysis easy and streamlined with the help of visual components like charts and graphs. Visual EDA is particularly relevant in the context of data quality profiling. In this section, we first discuss the importance of data visualisation in EDA in general and for data quality in particular, as mentioned in [2], [12].

We discuss the advancement in the field of visualisation recommendations mentioned in [15], [7], [11]. We also specifically discuss advancements in interactive visualisation mentioned in [24]. We discuss these advancements as rules of thumb to avoid misleading information that may be conveyed by visual artefacts. Further, we discuss the impact of data quality on visualisation as mentioned in [16]. Lastly, we review a few open-source visualisation tools (e.g., [14] and [1]) used for EDA focused on data quality.

2.3 Challenges to make Data Centric AI Algorithms Suitable for Industry Workloads

The field of data quality or data centric AI has been gaining traction in the last couple of years. In this section, we focus on the challenges exposed by industry workloads. Real world datasets often run into gigabytes or terabytes, and thus there is a need for algorithms that can address the scale of the data. In this section, we will cover the scalable data validation work covered in [23], [5], [22], [25], [21], [6]. We will also discuss the challenges associated with running advanced data quality metrics for ML as described in [10], automated sequencing [4] as well as performing EDA and visualisation tasks on large scale datasets. We will end this section with the discussion on open source toolkits that enable data validation on large datasets.

REFERENCES

[1] 2019. Facets. https://github.com/pair-code/facets.

[2] Shazia Afzal, Arunima Chaudhary, Nitin Gupta, Hima Patel, Carolina Spina, and Dakuo Wang. 2021. Data-Debugging Through Interactive Visual Explanations. In *Trends and Applications in Knowledge Discovery and Data Mining*, Manish Gupta and Ganesh Ramakrishnan (Eds.). Springer International Publishing, Cham, 133–142.

[3] Julien Aligon, Enrico Gallinucci, Matteo Golfarelli, Patrick Marcel, and Stefano Rizzi. 2015. A collaborative filtering approach for recommending OLAP sessions. *Decision Support Systems* 69 (01 2015), 20–30. https://doi.org/10.1016/j.dss.2014.11.003

[4] Laure Berti-Equille. 2019. Learn2clean: Optimizing the sequence of tasks for web data preparation. In *The World Wide Web Conference*. 2580–2586.

[5] Eric Breck, Neoklis Polyzotis, Sudip Roy, Steven Whang, and Martin Zinkevich. 2019. Data Validation for Machine Learning. In *Conference on Systems and Machine Learning (SysML)*.

[6] Ugo Comignani, Noël Novelli, and Laure Berti-Équille. 2020. Data quality checking for machine learning with mesqual. In *Advances in Database Technology-EDBT 2020, 23rd International Conference on Extending Database Technology,*.

[7] Victor Dibia and Çagatay Demiralp. 2018. Data2Vis: Automatic Generation of Data Visualizations Using Sequence to Sequence Recurrent Neural Networks. *CoRR* abs/1804.03126 (2018). arXiv:1804.03126 http://arxiv.org/abs/1804.03126

[8] Kyriaki Dimitriadou, Olga Papaemmanouil, and Yanlei Diao. 2016. AIDE: An Active Learning-Based Approach for Interactive Data Exploration. *IEEE Transactions on Knowledge and Data Engineering* 28, 11 (2016), 2842–2856. https://doi.org/10.1109/TKDE.2016.2599168

[9] Ori Bar El, Tova Milo, and Amit Somech. 2019. ATENA: An Autonomous System for Data Exploration Based on Deep Reinforcement Learning. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (2019).

[10] Nitin Gupta, Hima Patel, Shazia Afzal, Naveen Panwar, Ruhi Sharma Mittal, Shanmukha Guttula, Abhinav Jain, Lokesh Nagalapatti, Sameep Mehta, Sandeep Hans, et al. 2021. Data Quality Toolkit: Automatic assessment of data quality and remediation for machine learning datasets. *arXiv preprint arXiv:2108.05935* (2021).

[11] Kevin Hu, Michiel A. Bakker, Stephen Li, Tim Kraska, and César Hidalgo. 2019. VizML: A Machine Learning Approach to Visualization Recommendation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300358

[12] Sean Kandel, Ravi Parikh, Andreas Paepcke, Joseph M. Hellerstein, and Jeffrey Heer. 2012. Profiler: integrated statistical analysis and visualization for data quality assessment. In *AVI*.

[13] Alan F. Karr, Ashish P. Sanil, and David L. Banks. 2006. Data quality: A statistical perspective. *Statistical Methodology* 3, 2 (2006), 137–173. https://doi.org/10.1016/j.stamet.2005.08.005

[14] Doris Jung-Lin Lee, Dixin Tang, Kunal Agarwal, Thyne Boonmark, Caitlyn Chen, Jake Kang, Ujjaini Mukhopadhyay, Jerry Song, Micah Yong, Marti A. Hearst, and Aditya G. Parameswaran. 2021. Lux: Always-on Visualization Recommendations for Exploratory Dataframe Workflows. *Proc. VLDB Endow.* 15, 3 (nov 2021), 727–738. https://doi.org/10.14778/3494124.3494151

[15] Yuyu Luo, Xuedi Qin, Nan Tang, and Guoliang Li. 2018. DeepEye: Towards Automatic Data Visualization. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*. 101–112. https://doi.org/10.1109/ICDE.2018.00019

[16] Rischan Mafrur, Mohamed A. Sharaf, and G. Zuccon. 2020. Quality Matters: Understanding the Impact of Incomplete Data on Visualization Recommendation. In *DEXA*.

[17] Patrick Marcel, Nicolas Labroche, and Panos Vassiliadis. 2019. Towards a benefit-based optimizer for Interactive Data Analysis. In *DOLAP 2019*. Lisboa, France. https://hal.archives-ouvertes.fr/hal-02375855

[18] Tova Milo and Amit Somech. 2016. REACT: Context-Sensitive Recommendations for Data Analysis. 2137–2140. https://doi.org/10.1145/2882903.2899392

[19] Jinglin Peng, Weiyuan Wu, Brandon Lockhart, Song Bian, Jing Nathan Yan, Linghao Xu, Zhixuan Chi, Jeffrey M. Rzeszotarski, and Jiannan Wang. 2021. DataPrep.EDA: Task-Centric Exploratory Data Analysis for Statistical Modeling in Python. In *Proceedings of the 2021 International Conference on Management of Data (SIGMOD '21), June 20–25, 2021, Virtual Event, China.*

[20] A. Personnaz, S. Amer-Yahia, Laure Berti-Équille, M. Fabricius, and S. Subramanian. 2021. Balancing familiarity and curiosity in data exploration with deep reinforcement learning. In *Fourth workshop in exploiting AI techniques for data management (aiDM'21)*, R. (ed.) Bordawekar, Y. (ed.) Amsterdamer, O. (ed.) Shmueli, and N. (ed.) Tatbul (Eds.). ACM, 16–23. https://hal.archives-ouvertes.fr/hal-03278966 SIGMOD/PODS '21 : International Conference on Management of Data, En ligne, CHN, 12-/12/2025 - 12/12/2030.

[21] Sergey Redyuk, Zoi Kaoudi, Volker Markl, and Sebastian Schelter. 2021. Automating Data Quality Validation for Dynamic Data Ingestion.. In *EDBT*. 61–72.

[22] Sebastian Schelter, Stefan Grafberger, Philipp Schmidt, Tammo Rukat, Mario Kiessling, Andrey Taptunov, Felix Biessmann, and Dustin Lange. 2019. Differential data quality verification on partitioned data. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE, 1940–1945.

[23] Sebastian Schelter, Dustin Lange, Philipp Schmidt, Meltem Celikel, Felix Biessmann, and Andreas Grafberger. 2018. Automating large-scale data quality verification. *Proceedings of the VLDB Endowment* 11, 12 (2018), 1781–1794.

[24] L. Shen, E. Shen, Y. Luo, X. Yang, X. Hu, X. Zhang, Z. Tai, and J. Wang. 5555. Towards Natural Language Interfaces for Data Visualization: A Survey. *IEEE Transactions on Visualization Computer Graphics* 01 (jan 5555), 1–1. https://doi.org/10.1109/TVCG.2022.3148007

[25] Arun Swami, Sriram Vasudevan, and Joojay Huyn. 2020. Data sentinel: A declarative production-scale data validation platform. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE, 1579–1590.