2nd International Workshop on Tabular Data Analysis (TaDA)

Vasilis Efthymiou
Harokopio University of Athens &
FORTH-ICS
Greece
vefthym@hua.gr

Sainyam Galhotra
Cornell University
USA
sg@cs.cornell.edu

Oktie Hassanzadeh
IBM Research
USA
hassanzadeh@us.ibm.com

Chuan Lei
Amazon Web Services
USA
chuanlei@amazon.com

Kavitha Srinivas
IBM Research
USA
kavitha.srinivas@ibm.com

ABSTRACT

With the advent of data lakes and open data repositories containing heterogeneous collections of structured datasets, there is an increasing need for automated methods to analyze tabular data collections for a wide range of applications in data management, data science, and decision support. Performing such tasks over large and heterogeneous collections of tabular data is extremely challenging and an attractive research topic. The goal of this workshop is to provide a venue for the growing number of researchers in data management, AI, and Semantic Web communities working on problems relevant to tabular data analysis. TaDA 2024 was the second edition of this workshop and it included a keynote talk, a research track comprising presentations and posters, and virtual talks of the work done in these communities.

VLDB Workshop Reference Format:
Vasilis Efthymiou, Sainyam Galhotra, Oktie Hassanzadeh, Chuan Lei, and Kavitha Srinivas. 2nd International Workshop on Tabular Data Analysis (TaDA). VLDB 2024 Workshop: Tabular Data Analysis Workshop (TaDA).

1 INTRODUCTION

Data Analysis, as a crucial process in various domains, involves examining, cleaning, transforming, and modeling data to extract valuable insights, make informed conclusions, and facilitate decision-making [2]. However, performing such data analysis tasks becomes exceedingly complex when dealing with vast and diverse collections of tabular data, commonly found in enterprise data lakes and on the Web. Consequently, this challenge has piqued the interest of researchers and practitioners in data management, AI, and related communities [5–7, 11, 12].

To address the fundamental research challenges posed by tabular data analysis and foster the development of automated solutions, Tabular Data Analysis (TaDA) workshop (https://tabular-data-analysis.github.io/tada2024/) is organized with the primary goal of bringing together experts from diverse communities [1]. This workshop aims to create a collaborative environment for researchers and practitioners in data management and AI fields, enabling them to share insights, methodologies, and advancements in tackling the complexities of analyzing large and heterogeneous collections of tabular data. The Tabular Data Analysis workshop provides a forum for:

- Exchange of ideas between two communities: 1) an active community of data management researchers working on data integration, schema and data matching problems over tabular data, and 2) a vibrant community of researchers in AI and Semantic Web working on matching tabular data to Knowledge Graphs as a part of the ISWC SemTab Challenge [3, 4, 8–10].
- Presentation of late-breaking results related to several emerging research areas such as table representation learning and its applications, automation of data science pipelines, and data lake and data lakehouse solutions.
- Discussion of real-world data management challenges related to implementing industrial scale tabular data analysis solutions.

2 OVERVIEW OF THE PROGRAM

The workshop received several interesting submissions, almost twice as many as it received in its first edition in 2023 [1], on the different aspects of tabular data analysis, and each submission was reviewed by at least three reviewers. The accepted papers encompassed a wide range of topics, including data discovery, semantic table understanding, table union search, schema matching, benchmarking, and responsible AI. The workshop program consisted of a keynote talk from two researchers at Microsoft Research: Shi Han and Haoyu Dong.

Shi and Haoyu discussed cutting-edge technologies designed to tackle the major challenges in spreadsheet intelligence, encompassing areas such as detecting table ranges, analyzing table structures and sheet layouts, understanding data semantics, and recommending data presentations. Based on spreadsheet intelligence, the talk also highlighted their research and engineering efforts in boosting automation of data analytics to help Microsoft build technical leadership in the Business Intelligence market. In the trend of Large

Proceedings of the VLDB Endowment. ISSN 2150-8097.

Language Models (LLMs), they also presented their latest explorations into integrating LLMs with spreadsheet intelligence and data analytics.

3 ORGANIZATION

Workshop Chairs:

- Vasilis Efthymiou (Harokopio University of Athens)
- Sainyam Galhotra (Cornell University)
- Oktie Hassanzadeh (IBM Research)
- Chuan Lei (AWS)
- Kavitha Srinivas (IBM Research)

Steering Committee:

- Madelon Hulsebos (UC Berkeley)
- Ernesto Jiménez-Ruiz (City, University of London)
- Fatemeh Nargesian (University of Rochester)
- Natasha Noy (Google)
- Horst Samulowitz (IBM Research)

Program Committee:

- Nora Abdelmageed (University of Jena)
- Omar Benjelloun (Google)
- Rafael Berlanga Llavori (University Jaume I)
- Carsten Binnig (TU Darmstadt)
- Christian Bizer (University of Mannheim)
- Anastasia Dimou (KU Leuven)
- Christos Diou (Harokopio University of Athens)
- Zezhou Huang (Columbia University)
- Madelon Hulsebos (UC Berkeley)
- Andra Ionescu (TU Delft)
- Ernesto Jiménez-Ruiz (City, University of London)
- Aamod Khatiwada (Northeastern University)
- Udayan Khurana (IBM Research)
- Haridimos Kondylakis (FORTH-ICS)
- Marco Mesiti (University of Milan)
- Renée Miller (Northeastern University)
- George Papadakis (University of Athens)
- Paolo Papotti (EURECOM)
- Nhan Pham (IBM Research)
- Horst Samulowitz (IBM Research)
- Ismael Sanz (Universitat Jaume I)
- Roee Shraga (WPI)
- Kostas Stefanidis (Tampere University)
- Gerhard Weikum (Max Planck Institute for Informatics)
- You Wu (Google)

ACKNOWLEDGMENTS

We would like to thank the steering committee, the program committee, the keynote speakers, and the authors for their contributions. Finally, we thank the workshop attendees for making TaDA a great venue to discuss the works in the area of tabular data analysis.

REFERENCES

[1] Rajesh Bordawekar, Cinzia Cappiello, Vasilis Efthymiou, Lisa Ehrlinger, Vijay Gadepally, Sainyam Galhotra, Sandra Geisler, Sven Groppe, Le Gruenwald, Alon Y. Halevy, Hazar Harmouch, Oktie Hassanzadeh, Ihab F. Ilyas, Ernesto Jiménez-Ruiz, Sanjay Krishnan, Tirthankar Lahiri, Guoliang Li, Jiaheng Lu, Wolfgang Mauerer, Umar Farooq Minhas, Felix Naumann, M. Tamer Özsu, El Kindi Rezig, Kavitha Srinivas, Michael Stonebraker, Satyanarayana R. Valluri, Maria-Esther Vidal, Haixun Wang, Jiannan Wang, Yingjun Wu, Xun Xue, Mohamed Zaït, and Kai Zeng (Eds.). 2023. *Joint Proceedings of Workshops at the 49th International Conference on Very Large Data Bases (VLDB 2023), Vancouver, Canada, August 28 - September 1, 2023.* CEUR Workshop Proceedings, Vol. 3462. CEUR-WS.org. https://ceur-ws.org/Vol-3462

[2] Meta Brown. 2014. *Transforming Unstructured Data into Useful Information.* 211–230. https://doi.org/10.1201/b16666-11

[3] Vasilis Efthymiou, Ernesto Jiménez-Ruiz, Jiaoyan Chen, Vincenzo Cutrona, Oktie Hassanzadeh, Juan Sequeda, Kavitha Srinivas, Nora Abdelmageed, and Madelon Hulsebos (Eds.). 2023. *Proceedings of the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching, SemTab 2022, co-located with the 21st International Semantic Web Conference, ISWC 2022, Virtual conference, October 23-27, 2022.* CEUR Workshop Proceedings, Vol. 3320. CEUR-WS.org.

[4] Vasilis Efthymiou, Ernesto Jiménez-Ruiz, Jiaoyan Chen, Vincenzo Cutrona, Oktie Hassanzadeh, Juan Sequeda, Kavitha Srinivas, Nora Abdelmageed, Madelon Hulsebos, Aamod Khatiwada, Keti Korini, and Benno Kruit (Eds.). 2023. *Proceedings of the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching, SemTab 2023, co-located with the 22nd International Semantic Web Conference, ISWC 2023, Athens, Greece, November 6-10, 2023.* CEUR Workshop Proceedings, Vol. 3557. CEUR-WS.org. https://ceur-ws.org/Vol-3557

[5] Sainyam Galhotra, Anna Fariha, Raoni Lourenço, Juliana Freire, Alexandra Meliou, and Divesh Srivastava. 2022. DataPrism: Exposing Disconnect between Data and Systems. In *SIGMOD '22: International Conference on Management of Data, Philadelphia, PA, USA, June 12 - 17, 2022*, Zachary Ives, Angela Bonifati, and Amr El Abbadi (Eds.). ACM, 217–231. https://doi.org/10.1145/3514221.3517864

[6] Oktie Hassanzadeh, Anastasios Kementsietsidis, Benny Kimelfeld, Rajasekar Krishnamurthy, Fatma Ozcan, and Ippokratis Pandis. 2013. Next Generation Data Analytics at IBM Research. *Proc. VLDB Endow.* 6, 11 (2013), 1174–1175. https://doi.org/10.14778/2536222.2536246

[7] Mossad Helali, Essam Mansour, Ibrahim Abdelaziz, Julian Dolby, and Kavitha Srinivas. 2022. A Scalable AutoML Approach Based on Graph Neural Networks. *Proc. VLDB Endow.* 15, 11 (2022), 2428–2436. https://doi.org/10.14778/3551793.3551804

[8] Ernesto Jiménez-Ruiz, Vasilis Efthymiou, Jiaoyan Chen, Vincenzo Cutrona, Oktie Hassanzadeh, Juan Sequeda, Kavitha Srinivas, Nora Abdelmageed, Madelon Hulsebos, Daniela Oliveira, and Catia Pesquita (Eds.). 2022. *Proceedings of the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching co-located with the 20th International Semantic Web Conference (ISWC 2021), Virtual conference, October 27, 2021.* CEUR Workshop Proceedings, Vol. 3103. CEUR-WS.org.

[9] Ernesto Jiménez-Ruiz, Oktie Hassanzadeh, Vasilis Efthymiou, Jiaoyan Chen, Kavitha Srinivas, and Vincenzo Cutrona (Eds.). 2020. *Proceedings of the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab 2020) co-located with the 19th International Semantic Web Conference (ISWC 2020), Virtual conference, November 5, 2020.* CEUR Workshop Proceedings, Vol. 2775. CEUR-WS.org.

[10] Ernesto Jiménez-Ruiz, Oktie Hassanzadeh, Kavitha Srinivas, Vasilis Efthymiou, and Jiaoyan Chen (Eds.). 2020. *Proceedings of the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching co-located with the 18th International Semantic Web Conference, SemTab@ISWC 2019, Auckland, New Zealand, October 30, 2019.* CEUR Workshop Proceedings, Vol. 2553. CEUR-WS.org.

[11] Fatemeh Nargesian, Ken Q. Pu, Erkang Zhu, Bahar Ghadiri Bashardoost, and Renée J. Miller. 2020. Organizing Data Lakes for Navigation. In *Proceedings of the 2020 International Conference on Management of Data, SIGMOD Conference 2020, online conference [Portland, OR, USA], June 14-19, 2020*, David Maier, Rachel Pottinger, AnHai Doan, Wang-Chiew Tan, Abdussalam Alawini, and Hung Q. Ngo (Eds.). ACM, 1939–1950. https://doi.org/10.1145/3318464.3380605

[12] Fatma Özcan, Chuan Lei, Abdul Quamar, and Vasilis Efthymiou. 2021. Semantic enrichment of data for AI applications. In *Proceedings of the Fifth Workshop on Data Management for End-To-End Machine Learning, In conjunction with the 2021 ACM SIGMOD/PODS Conference, DEEM@SIGMOD 2021, Virtual Event, China, 20 June, 2021*, Matthias Boehm, Julia Stoyanovich, and Steven Whang (Eds.). ACM, 4:1–4:7. https://doi.org/10.1145/3462462.3468881