Editorial: Special Issue on Deep Learning for Data Quality

DONATELLO SANTORO, Universitá della Basilicata, Italy
SARAVANAN THIRUMURUGANATHAN, Qatar Computing Research Institute, HBKU, Qatar
PAOLO PAPOTTI, EURECOM, France

This editorial summarizes the content of the Special Issue on Deep Learning for Data Quality of the **Journal of Data and Information Quality (JDIQ)**.

CCS Concepts: • **Information systems** → **Data cleaning**; **Data management systems**; • **Computing methodologies** → **Machine learning;**

Additional Key Words and Phrases: Deep learning, schema matching, data labeling

ACM Reference format:
Donatello Santoro, Saravanan Thirumuruganathan, and Paolo Papotti. 2022. Editorial: Special Issue on Deep Learning for Data Quality. *J. Data and Information Quality* 14, 3, Article 14 (August 2022), 3 pages.
https://doi.org/10.1145/3513135

1 INTRODUCTION

As the guest editors, it is our pleasure to introduce this special issue of the Journal of Data and Information Quality on *Deep Learning for Data Quality*. This issue includes two novel research articles that make meaningful progress on important problems of data quality – schema matching and data labeling.

Organizations are drowning in big data and are increasingly turning to **deep learning (DL)** for integrating and analyzing data for decision making. In domains such as text, image, and speech that exhibit hidden structures, the contributions of DL has been nothing short of prodigious. It has obviated decades of hand engineered features and achieved state-of-the-art-results. Over the recent years, the data quality community has successfully adapted these DL techniques for data integration [2]. The first generation of these efforts, focused on problems such as entity matching [3, 4] that can be easily formulated as a binary classification problem. The solution often involved adaptation of techniques from natural language processing (such as embeddings and language models) for entity matching [6]. The next generation of these efforts, as exemplified in this special issue, focus on addressing and exploiting some of the characteristics that are unique to data integration.

An oft-quoted statistic is that 80% of work in any data analytics/**machine learning (ML)** task involves data preparation. This work is often time consuming, frustrating and least enjoyed by the data scientists. Not surprisingly, using DL/ML to ensure data quality has become of paramount interest within the data quality community [3]. However, blindly applying deep learning often produces sub-optimal results. Data integration is especially challenging given the heterogeneity

Authors' addresses: D. Santoro, Università della Basilicata, Via dell'Ateneo Lucano, 10 - 85100 Potenza, Italy; email: donatello.santoro@unibas.it; S. Thirumuruganathan, Qatar Computing Research Institute - HBKU, 34110 Education City, Doha, Qatar; email: sthirumuruganathan@hbku.edu.qa; P. Papotti, EURECOM, 450 route des Chappes, 06410 Biot, France. email: paolo.papotti@eurecom.fr.

1936-1955/2022/08-ART14
https://doi.org/10.1145/3513135

of the data, the substantial use of domain knowledge, a non trivial need for "common sense", the use of data dependencies to model errors, and so on. However, it is important to make progress in this important domain. The emerging field of 'data-centric AI' seeks to shift the focus from 'ML models' to all aspects of data quality. This involves the entire lifecycle of data collection, labeling, preprocessing, integration, augmentation, etc. [1, 5]. It is incumbent upon our community to provide leadership on this rapidly maturing field.

2 ARTICLES INCLUDED IN THE SPECIAL ISSUE

The two articles presented in this special issue tackle two non-trivial challenges in data integration. Instead of designing an ML/DL model for a specific problem, they address some meta-issues that are inherent in many other problems. Specifically, the first work carefully describes a mechanism to calibrate and improve human assessments for schema matching task. The second work proposes a novel and intuitive method for data labeling.

The first article, *PoWareMatch: a Quality-aware Deep Learning Approach to Improve Human Schema Matching* by Roee Shraga and Avigdor Gal, introduces a novel angle of investigation for schema matching task. Despite the importance of schema matching for various downstream tasks such as data enrichment, aligning knowledge graphs and relational tables, semantic web, etc., there has been limited success in designing accurate DL-based matchers. While human-based matching of the schemata is often treated as the gold standard, it has numerous issues in practice due to the wide spread use of non-experts lacking the relevant domain expertise such as in crowdsourcing. The authors propose an innovative approach to improve the quality of schema matching through deep learning. They define a monotonic evaluation measure and its probabilistic derivative and propose a step-wise matching algorithm that uses the confidence of human labelers to obtain higher quality matches.

The second article, *A Cluster-then-label Approach for Few-shot Learning with Application to Automatic Image Data Labeling* by Renzhi Wu, Nilaksh Das, Sanya Chaba, Sakshi Gandhi, Duen Horng Chau and Xu Chu, focuses on an orthogonal problem of data labeling. A large number of labeled data often underlies the spectacular results of deep learning models. However, in a number of tasks in data integration, such large hand-labeled datasets often do not exist. Creating a high quality dataset is often expensive as it would require the use of domain experts. The 'data programming' paradigm reduces human effort in obtaining data labels through the use of labeling functions. However, it is not a panacea and could require significant effort for each task. In this paper, the authors introduce an intuitive and effective alternate approach. Specifically, they propose a domain agnostic paradigm that could be used for labeling without requiring any domain specific functions. While the authors focus on data labeling for images, it would be exciting to extend the general idea for other modalities.

We thank the Editor-in-Chief Tiziana Catarci and the senior editor Paolo Missier, for their guidance throughout this process. We also thank the administrative assistant Andrea Marrella for his help in coordinating this special issue. Finally, we express our thanks to the authors and reviewers, without whose input the issue would not have been possible.

REFERENCES

[1] Riccardo Cappuzzo, Paolo Papotti, and Saravanan Thirumuruganathan. 2020. Creating embeddings of heterogeneous relational datasets for data integration tasks. 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, 1335–1349. https://doi.org/10.1145/3318464.3389742

[2] Xin Luna Dong and Theodoros Rekatsinas. 2018. Data integration and machine learning: A natural synergy. In *Proceedings of the 2018 International Conference on Management of Data, SIGMOD Conference 2018*, Houston, TX,

USA, June 10-15, 2018, Gautam Das, Christopher M. Jermaine, and Philip A. Bernstein (Eds.). ACM, 1645–1650. https://doi.org/10.1145/3183713.3197387

[3] Yuliang Li, Jinfeng Li, Yoshihiko Suhara, AnHai Doan, and Wang-Chiew Tan. 2020. Deep entity matching with pre-trained language models. *Proc. VLDB Endow.* 14, 1 (2020), 50–60. https://doi.org/10.14778/3421424.3421431

[4] Sidharth Mudgal, Han Li, Theodoros Rekatsinas, AnHai Doan, Youngchoon Park, Ganesh Krishnan, Rohit Deep, Esteban Arcaute, and Vijay Raghavendra. 2018. Deep learning for entity matching: A design space exploration. In *Proceedings of the 2018 International Conference on Management of Data, SIGMOD Conference 2018*, Houston, TX, USA, June 10-15, 2018, Gautam Das, Christopher M. Jermaine, and Philip A. Bernstein (Eds.). ACM, 19–34. https://doi.org/10.1145/3183713.3196926

[5] Saravanan Thirumuruganathan, Nan Tang, Mourad Ouzzani, and AnHai Doan. 2020. Data curation with deep learning. In *Proceedings of the 23rd International Conference on Extending Database Technology, EDBT 2020*, Copenhagen, Denmark, March 30 - April 2, 2020, Angela Bonifati, Yongluan Zhou, Marcos Antonio Vaz Salles, Alexander Böhm, Dan Olteanu, George H. L. Fletcher, Arijit Khan, and Bin Yang (Eds.). OpenProceedings.org, 277–286. https://doi.org/10.5441/002/edbt.2020.25

[6] Joseph P. Turian, Lev-Arie Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, July 11-16, 2010, Uppsala, Sweden, Jan Hajic, Sandra Carberry, and Stephen Clark (Eds.). The Association for Computer Linguistics, 384–394. https://aclanthology.org/P10-1040/.