DEEM'22: Data Management for End-to-End Machine Learning

Matthias Boehm
Graz University of Technology
Graz, Austria

Paroma Varma
Snorkel AI
Palo Alto, USA

Doris Xin
UC Berkeley / Linea
Berkeley / San Francisco, USA

ABSTRACT

The DEEM'22 workshop (Data Management for End-to-End Machine Learning) is held on Sunday June 12th, in conjunction with SIGMOD/PODS 2022. DEEM brings together researchers and practitioners at the intersection of applied machine learning, data management and systems research, with the goal to discuss the arising data management issues in ML application scenarios. The workshop solicits regular research papers (10 pages) describing preliminary and ongoing research results, including industrial experience reports of end-to-end ML deployments, related to DEEM topics. In addition, DEEM 2022 establishes a new paper category for reports on applications and tools (4 pages) as a forum for sharing interesting use cases, problems, datasets, benchmarks, visionary ideas, system designs, and descriptions of system components and tools related to end-to-end ML pipelines. DEEM 2022 received 13 high-quality submissions from Africa, Asia, Europe, and North America, with 5 regular research papers, and 8 reports on applications and tools.

ACM Reference Format:
Matthias Boehm, Paroma Varma, and Doris Xin. 2022. DEEM'22: Data Management for End-to-End Machine Learning. In *Proceedings of the 2022 International Conference on Management of Data (SIGMOD '22), June 12–17, 2022, Philadelphia, PA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3514221.3524075

1 INTRODUCTION

Applying Machine Learning (ML) in real-world scenarios is a challenging task. In recent years, the main focus of the data management community has been on creating systems and abstractions for the efficient training of ML models on large datasets. However, model training is only one of many steps in an end-to-end ML application, and a number of orthogonal data management problems arise from the large-scale use of ML.

For example, data preprocessing and feature extraction workloads may be complicated and require simultaneous execution of relational and linear algebraic operations. Next, model selection may involve searching many combinations of model architectures, features, and hyper-parameters to find the best-performing model. After model training, the resulting model may have to be deployed and integrated into business workflows and require lifecycle management using metadata and lineage. As a further complication, the resulting system may have to take into account a heterogeneous audience, ranging from domain experts without programming skills to data engineers and statisticians who develop custom algorithms.

Additionally, the importance of incorporating ethics and legal compliance into machine-assisted decision-making is being broadly recognized. Critical opportunities for improving data quality and representativeness, controlling for bias, and allowing humans to oversee and impact computational processes are missed if we do not consider the lifecycle stages upstream from model training and deployment. DEEM welcomes research on providing system-level support to data scientists who wish to develop and deploy responsible machine learning methods.

DEEM [1–5] aims to bring together researchers and practitioners at the intersection of applied machine learning, data management and systems research, with the goal to discuss the arising data management issues in ML application scenarios. The workshop solicits regular research papers (10 pages plus references) describing preliminary or completed research results, as well as reports on applications and tools (4 pages). With this new paper category on applications and tools, the DEEM workshop aims—complementary to the recently introduced, scalable data science and engineering tracks at SIGMOD and PVLDB—to establish a broader forum for sharing interesting use cases, problems, datasets, benchmarks, visionary ideas, system designs, and descriptions of system components and tools related to end-to-end ML pipelines.

2 TOPICS OF INTEREST

Over the last years the topics of interest naturally expanded as ML pipelines become increasingly complex, and aspects of data-centric AI become increasingly important for building end-to-end ML systems in practice. The DEEM'22 call for papers outlined the following areas of particular interest for the workshop:

- Data Management in Machine Learning Applications
- Definition, Execution and Optimization of Complex Machine Learning Pipelines
- Systems for Managing the Lifecycle of ML Models
- Systems for Efficient Hyper-parameter Search and Feature Engineering and Selection
- Machine Learning Services in the Cloud
- Modeling, Storage, and Provenance of ML Artifacts
- Integration of Machine Learning and Dataflow Systems
- Integration of Machine Learning and ETL Processing
- Definition and Execution of Complex Ensemble Predictors
- Sourcing, Labeling, Integrating, and Cleaning Data for Machine Learning
- Data Validation and Model Debugging Techniques
- Privacy-preserving Machine Learning
- Benchmarking of Machine Learning Applications
- Responsible Data Management
- Transparency and Accountability of Machine-Assisted Decision Making
- Impact of Data Quality and Data Pre-processing on the Fairness of ML Predictions

SIGMOD '22, June 12–17, 2022, Philadelphia, PA, USA
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9249-5/22/06.
https://doi.org/10.1145/3514221.3524075

3 ORGANIZATION

Since its inception in 2017, the DEEM workshop is organized by changing teams of workshop co-chairs, governed by a steering committee, and supported by expert program committees.

Workshop Chairs: The DEEM'22 workshop is jointly organized by the following workshop chairs, drawing from the experience of the previous five DEEM workshops and their chairs:

- Matthias Boehm (Graz University of Technology)
- Paroma Varma (Snorkel AI)
- Doris Xin (UC Berkeley & Linea)

Steering Committee: A steering committee governs the DEEM workshop and ensures its continuation with a healthy rolling handover of workshop co-chairs. This committee includes:

- Juliana Freire (New York University)
- Bill Howe (University of Washington)
- H.V. Jagadish (University of Michigan)
- Volker Markl (TU Berlin)
- Stefan Seufert (Amazon Research)
- Markus Weimer (Microsoft AI)

As an honorary member of the steering committee, we would like to mention Sebastian Schelter (University of Amsterdam) who established the DEEM workshop but retired from the committee this year to avoid conflicts of interest in a clear and transparent manner.

Program Committee: Furthermore, we thank the DEEM'22 program committee—with diverse backgrounds and seniority levels—for reviewing the individual submissions and providing detailed and constructive feedback.

- Khaled Ammar (Univ. of Waterloo, Thomson Reuters Labs)
- Abolfazl Asudeh (University of Illinois at Chicago)
- Srikanta Bedathur (IIT Delhi)
- Renata Borovica-Gajic (University of Melbourne)
- Patrick Damme (TU Graz, Know-Center GmbH)
- Ahmed Elgohary (Microsoft)
- Edward Gan (Stanford University)
- Rainer Gemulla (University of Mannheim)
- Chris Jermaine (Rice University)
- Zoi Kaoudi (TU Berlin)
- Sanjay Krishnan (University of Chicago)
- Arun Kumar (UC San Diego)
- Milos Nikolic (University of Edinburgh)
- Tilmann Rabl (HPI, University of Potsdam)
- Berthold Reinwald (IBM Research - Almaden)
- Maximilian Schleich (University of Washington)
- Christin Seifert (University of Duisburg-Essen)
- Vraj Shah (UC San Diego)
- Nesime Tatbul (Intel Labs and MIT)
- Shirish Tatikonda (Target Corporation)
- Ce Zhang (ETH Zurich)

4 WORKSHOP FORMAT

The workshop will be held in hybrid form (in-person and virtually) with the following half- or full-day schedule:

- Opening and closing keynotes, by industrial and academic speakers of diverse backgrounds,
- 2-3 technical sessions, each featuring an invited speaker and several accepted papers, and
- Optionally, a panel on specific topics such as Teaching Data Science & ML Systems.

REFERENCES

[1] Matthias Boehm, Julia Stoyanovich, and Steven Whang (Eds.). 2021. *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.* ACM. https://doi.org/10.1145/3462462 http://deem-workshop.org/2021/index.html.

[2] Sebastian Schelter, Neoklis Polyzotis, Stephan Seufert, and Manasi Vartak (Eds.). 2019. *Proceedings of the 3rd International Workshop on Data Management for End-to-End Machine Learning, DEEM@SIGMOD 2019, Amsterdam, The Netherlands, June 30, 2019.* ACM. https://doi.org/10.1145/3329486 http://deem-workshop.org/2019/index.html.

[3] Sebastian Schelter, Stephan Seufert, and Arun Kumar (Eds.). 2018. *Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning, DEEM@SIGMOD 2018, Houston, TX, USA, June 15, 2018.* ACM. http://dl.acm.org/citation.cfm?id=3209889 http://deem-workshop.org/2018/index.html.

[4] Sebastian Schelter, Steven Whang, and Julia Stoyanovich (Eds.). 2020. *Proceedings of the Fourth Workshop on Data Management for End-To-End Machine Learning, In conjunction with the 2020 ACM SIGMOD/PODS Conference, DEEM@SIGMOD 2020, Portland, OR, USA, June 14, 2020.* ACM. https://doi.org/10.1145/3399579 http://deem-workshop.org/2020/index.html.

[5] Sebastian Schelter and Reza Zadeh (Eds.). 2017. *Proceedings of the 1st Workshop on Data Management for End-to-End Machine Learning, DEEM@SIGMOD 2017, Chicago, IL, USA, May 14, 2017.* ACM. https://doi.org/10.1145/3076246 http://deem-workshop.org/2017/index.html.