Ninth Workshop on Data Management for End-to-End Machine Learning (DEEM)

Stefan Grafberger
BIFOLD & TU Berlin
Berlin, Germany

Madelon Hulsebos
CWI
Amsterdam, Netherlands

Matteo Interlandi
Microsoft
Los Angeles, USA

Shreya Shankar
UC Berkeley
Berkeley, USA

Abstract

The DEEM'25 workshop (Data Management for End-to-End Machine Learning) is held on Friday, June 27th, in conjunction with SIGMOD/PODS 2025. DEEM brings together researchers and practitioners at the intersection of applied machine learning, data management, and systems research, with the goal of discussing the arising data management issues in ML application scenarios. The workshop solicits regular research papers (8 pages) describing preliminary and ongoing research results, including industrial experience reports of end-to-end ML deployments, related to DEEM topics. In addition, DEEM 2025 has a category for short papers (4 pages) as a forum for sharing interesting use cases, problems, datasets, benchmarks, visionary ideas, system designs, preliminary results, and descriptions of system components and tools related to end-to-end ML pipelines. This year, the workshop received 18 high-quality submissions on diverse topics relevant to DEEM.

ACM Reference Format:
Stefan Grafberger, Madelon Hulsebos, Matteo Interlandi, and Shreya Shankar. 2025. Ninth Workshop on Data Management for End-to-End Machine Learning (DEEM). In *Companion of the 2025 International Conference on Management of Data (SIGMOD-Companion '25), June 22–27, 2025, Berlin, Germany.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3722212.3724483

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 and increased adoption of large language models (LLMs).

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. Many such challenges are human or engineer-centered (e.g., monitoring ML pipelines, leveraging LLMs for domain-specific tasks at scale), and DEEM uniquely encourages submissions on such topics.

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 [1–8] aims to bring together researchers and practitioners at the intersection of applied machine learning, data management, and systems research, with the goal of discussing the arising data management issues in ML and LLM application scenarios. The workshop solicits regular research and industry papers (8 pages plus references) describing preliminary or completed research results, as well as short papers (4 pages). With the latter paper category, 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 have naturally expanded as ML pipelines have become increasingly complex, and aspects of data-centric AI have become increasingly important for building end-to-end ML systems in practice. The DEEM'25 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 ML, e.g., for Managing the Lifecycle of ML Models, Efficient Hyper-parameter Search, or Feature Selection
- Machine Learning Services in the Cloud

SIGMOD-Companion '25, Berlin, Germany

ACM ISBN 979-8-4007-1564-8/2025/06
https://doi.org/10.1145/3722212.3724483

- Modeling, Storage, and Provenance of Machine Learning 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
- MLOps, 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 Preprocessing on the Fairness of ML Predictions
- Horror stories, Anecdotes, and Lessons Learned on data management for ML
- Data management for multimodal ML
- Vector Databases for Retrieval and Systems for Retrieval Augmented Generation
- ML for data management for ML
- Data Management challenges for LLMs

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'25 workshop is jointly organized by the following workshop chairs, drawing from the experience of the previous DEEM workshops and their chairs:

- Stefan Grafberger (BIFOLD & TU Berlin)
- Madelon Hulsebos (CWI)
- Matteo Interlandi (Microsoft)
- Shreya Shankar (UC Berkeley)

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)
- Sebastian Schelter (BIFOLD & TU Berlin)
- Stefan Seufert (Amazon Research)
- Markus Weimer (Microsoft AI)

Program Committee: Furthermore, we thank the DEEM'23 program committee—with diverse backgrounds and seniority levels—for reviewing the individual submissions and providing detailed and constructive feedback.

- Anna Pavlenko, Microsoft Gray Systems Lab
- Bojan Karlaš, Harvard University
- Gerardo Vitagliano, MIT CSAIL
- Haralampos Gavriilidis, Technische Universität Berlin
- Jacopo Tagliabue, Bauplan
- Joy Arulraj, Georgia Tech
- Konstantinos Kanellis, University of Wisconsin-Madison
- Manisha Luthra, TU Darmstadt and DFKI
- Matthias Boehm, Technische Universität Berlin
- Maximilian Böther, ETH Zurich
- Maximilian Schüle, University of Bamberg
- Pinar Tozun, IT University of Copenhagen
- Rainer Gemulla, Universität Mannheim
- Sebastian Schelter, BIFOLD & TU Berlin
- Sivaprasad Sudhir, MIT
- Ties Robroek, IT University of Copenhagen
- Till Döhmen, MotherDuck
- Xue Li, CWI
- Yiming Lin, UC Berkeley
- Zezhou Huang, Columbia University

4 Workshop Format

The workshop will be held in-person with the following half- or full-day schedule:

- 2-3 technical sessions, each featuring an invited speaker and several accepted papers, and
- Keynotes by Pinar Tozun (Technical University of Denmark) and Gaël Varoquaux (Inria, Probabl),
- 1-2 poster sessions.

At the time of writing, we are discussing sponsor opportunities with several companies.

References

[1] 2024. *DEEM '24: Proceedings of the Eighth Workshop on Data Management for End-to-End Machine Learning* (Santiago, AA, Chile). Association for Computing Machinery, New York, NY, USA.

[2] Matthias Boehm, Madelon Hulsebos, Shreya Shankar, and Paroma Varma (Eds.). 2023. *Proceedings of the Seventh Workshop on Data Management for End-to-End Machine Learning* (Seattle, WA, USA). Association for Computing Machinery, New York, NY, USA. http://deem-workshop.github.io/2023/index.html.

[3] 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.github.io/2021/index.html.

[4] 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* (Philadelphia, PA, USA) *(SIGMOD '22)*. Association for Computing Machinery, New York, NY, USA, 2548–2549. doi:10.1145/3514221.3524075

[5] 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.github.io/2019/index.html.

[6] 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.github.io/2018/index.html.

[7] 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.github.io/2020/index.html.

[8] 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.github.io/2017/index.html.