Second Workshop on ML for Systems and Systems for ML

Matthias Boehm [1], Andreas Kipf [2], and Manisha Luthra [3]

Abstract: After the first instance at BTW 2023, this second instance of the Workshop on ML for Systems and Systems for ML was held on March 04, 2025 in conjunction with BTW 2025 in Bamberg. The workshop aimed to bring together researchers and practitioners working on applying machine learning (ML) to a spectrum of systems problems as well as building systems for ML pipelines and applications. As a new format, we solicited one-page abstract submissions of already published or ongoing work as well as invited a diverse set of complementary speakers to present prior work. Finally, the workshop featured 12 talks in three sessions and was generally well attended and received.

1 Introduction

The rapid advances in machine learning (ML) have significantly increased its adoption across various fields including data systems, both in academia and industry. These advancements have not only enhanced existing data systems but, in some cases, have completely transformed their internal components, leading to the development of an important field of "learned data system components" in the ML for Systems area. Similarly, a well-structured systems approach has also played a crucial role in advancing current ML techniques and systems, forming the basis for the Systems for the ML area.

Workshop Goals. The workshop aimed to bring together renowned researchers and practitioners through open abstract submissions and invited talks to discuss intriguing topics at the core of the two focus areas. The open format for submitting abstracts from anything like a position paper to technical experiences allowed for lively discussions as well as fostered collaborations among participants. The speakers also shared insights into their ongoing projects and the open research challenges they are currently addressing.

Topics of Interest. In detail, the topics of interest included:

- ML for Systems
    - Learned query processing and optimization
    - Learned index structures and storage layouts
    - Learned algorithms for sorting, compressing, and encoding data
    - Learned data exploration, discovery, and integration
    - Self-tuning and instance-optimized database systems

[1] TU Berlin & BIFOLD, matthias.boehm@tu-berlin.de, https://orcid.org/0000-0003-1344-3663
[2] University of Technology Nuremberg, andreas.kipf@utn.de, https://orcid.org/0000-0003-3463-0564
[3] TU Darmstadt & DFKI, manisha.luthra@dfki.de, https://orcid.org/0000-0002-3788-6664

- Learned data systems on emerging hardware and cloud platforms
- New datasets, benchmarks, and evaluation methods for learned databases
- Novel use of ML techniques in big data applications
- Novel use of natural language models and interfaces in data management
- Building and managing large-scale knowledge bases

- Systems for ML
  - Data and model management in ML applications and complex ML pipelines
  - Data integration, alignment, and preparation of multi-modal training datasets
  - Data cleaning/debugging techniques and data quality management
  - Data augmentation techniques, pipelines, and algorithm integration
  - Data flow optimizations in ML systems (e.g., rewrites, operator fusion)
  - Data- and task-parallel execution strategies for ML pipelines
  - Data access methods in ML systems (e.g., indexing, compression, partitioning)
  - New data system infrastructures and tools for applied ML

2 Organization Committee

Workshop Co-chairs. The workshop was co-organized by the following workshop chairs drawing from their previous experience and papers in this area.

- Manisha Luthra (TU Darmstadt & DFKI)
- Andreas Kipf (University of Technology Nuremberg)
- Matthias Boehm (TU Berlin & BIFOLD)

3 Workshop Format and Program

Workshop Format. Compared to the first instance of the workshop [LKB23], we established a new format of one-page abstract submissions of already published or ongoing work. We received seven of such abstracts, accepted the six of them that matched the focus of the workshop, and additionally invited five complementary talks.

Program. Together, we assembled the following well-balanced program of 12 talks, where submitted abstracts and invited talks got 15+5min and 25+5min speaking time, respectively:

- **Session 1: Systems for ML** (chaired by Matthias)
  - Matthias and Andreas: Opening Remarks
  - Maximilian Schüle (Univeristy of Bamberg): Blue Elephants Inspecting Pandas: Inspection and Execution of Machine Learning Pipelines in SQL
  - Maximilian Böther (ETH Zurich): Modyn: Data-Centric Machine Learning Pipeline Orchestration
  - Stefan Grafberger (TU Berlin): mlwhatif: Data-Centric What-If Analysis for Native Machine Learning Pipelines

- **Session 2: Applications & Benchmarks** (chaired by Matthias)
  - Stefan Hagedorn, Steffen Kläbe (Actian): Experiences of Implementing In-Database TPCx-AI
  - Thaleia-Dimitra Doudali (IMDEA): Keep it Simple, Sustainable! When Is ML Necessary in Cloud Resource Management?
  - Jan-Micha Bodensohn, Liane Vogel (TU Darmstadt & DFKI): Large Language Models for Enterprise Data Engineering
  - Akanksha Vijayvergiya (University Passau): Time-Series Analysis for Life-Science Data

- **Session 3: ML for Systems** (chaired by Andreas)
  - Silvan Reiner (University of Konstanz): ML4DB: Don't Learn What You Already Know
  - Johannes Wehrstein (TU Darmstadt): GRACEFUL: A Learned Cost Estimator For UDFs
  - Giorgio Vinciguerra (University of Pisa): Learned Compression of Nonlinear Time Series With Random Access
  - Immanuel Trummer (Cornell University): CheaPT: Using Language Models Without Breaking the Bank

References

[LKB23]   Luthra, M.; Kipf, A.; Boehm, M.: A Tutorial Workshop on ML for Systems and Systems for ML. In: BTW. Pp. 707–708, 2023, DOI: 10.18420/BTW2023-43.