Datenbank-Spektrum (2024) 24:183–186
https://doi.org/10.1007/s13222-024-00493-2

EDITORIAL

Editorial

Maximilian E. Schüle[1] · Meike Klettke[2] · Uta Störl[3]

Angenommen: 30. Oktober 2024 / Online publiziert: 6. November 2024

1 Schwerpunktthema: „ML for Systems and Systems for ML“

Methoden des maschinellen Lernens (ML) benötigen große Mengen an gültigen, korrekten, vollständigen und aktuellen Daten, auf denen sie trainiert werden können. Um diese Daten bereitzustellen, werden Data-Engineering-Komponenten eingesetzt, die die Aufgaben Datengenerierung (Erfassen und Speichern von Daten), Data-Understanding (Exploration von Datensätzen, Schemaextraktion, Data-Type-Inference und das Ableiten semantischer Constraints), Datenauswahl (Selektion von Daten, Datenaggregation), Data-Cleaning (Kontrollieren und Korrigieren von Daten, Null-Value-Imputation, Ausreißer-Erkennung und Korrektur, Duplikateliminierung, Bias-Detection und Concept-Shift-Erkennung), Datentransformation und -integration (Transformation zwischen verschiedenen Datenmodellen und Schemata sowie Integration von heterogenen Daten in ein einheitliches Datenformat) durchführen. Alle diese Verfahren werden eingesetzt, um Daten für das (erneute) Trainieren eines Modells und dessen Bereitstellung (in einer verteilten Umgebung) umzusetzen. In der Regel werden für eine ML-Aufgabe mehrere Data-Engineering-Algorithmen eingesetzt, die zu sogenannten Data-Engineering-Pipelines kombiniert werden. Dadurch bereitet Data-Engineering nicht nur die Verfahren des maschinellen Lernens vor, sondern ist integraler Bestandteil jeder ML-Pipeline. In diesem Heft können wir Ihnen zwei Beiträge präsentieren, die Tools für Data-Engineering-Pipelines entwickeln bzw. diese anwenden: Ein Beitrag von Sebastian Schelter zu dem System MLINSPECT, sowie ein Beitrag zu einer Anwendung von Data-Engineering-Verfahren für pharmazeutische Fragestellungen von Tobias Schreier.

ML for Systems. Die Datenbankcommunity entwickelt nicht nur Lösungen für ML-Verfahren sondern setzt diese auch für die Aufgaben innerhalb von Datenbankmanagementsystemen ein. So wurden in den letzten Jahren ML-Verfahren für verschiedene Optimierungsaufgaben in Systemen entwickelt wie zum Beispiel die Optimierung der Anfrageausführung durch lernende Verfahren und die Verbesserung der Kardinalitätsabschätzung. Auch das Konfigurieren von Systemen und das Steuern von Workloads kann durch gelernte Modelle erfolgen, die sich schrittweise weiter optimieren. Gleiches gilt für Data-Engineering-Pipelines: Die darin ausgeführten Verfahren können ebenfalls mit Hilfe von ML-Methoden implementiert werden. So wurden ML-basierte Verfahren beispielsweise für Missing-Value-Imputation, Outlier-Detection und Automatisierung von Datentransformationen entwickelt. Bei vielen Data-Engineering-Aufgaben wird zudem aktuell mit LLMs getestet, wie diese genutzt werden können, um zusätzliches Wissen in die Verfahren einzubringen.

Stellvertretend für diese Klassen von Verfahren haben wir im vorliegenden Heft zwei Beiträge, einen zur Annotation und Verwendung multimodaler Daten für Routenplaner von Paul Walther sowie einen Artikel aus der Gruppe von Marina Tropmann-Frick zur Bewertung verschiedener Verfahren der Missing-Value-Imputation mit LLMs.

Systems for ML. Für Data-Engineering sind Systeme für die Verarbeitung von Daten von grundlegender Bedeutung. Deren Verbesserung ist entscheidend für das effiziente Trainieren und Anwenden von Modellen. Ein weiteres aktuelles Forschungsthema ist die Entwicklung von Systemen für maschinelles Lernen (*Systems for ML*) zur Unterstützung des Entwurfs ganzheitlicher ML-Lösungen. Systeme für maschinelles Lernen verbessern die Anwendung von Mo-

Maximilian E. Schüle
maximilian.schuele@uni-bamberg.de

Meike Klettke
meike.klettke@ur.de

✉ Uta Störl
uta.stoerl@fernuni-hagen.de

[1] University of Bamberg, Bamberg, Deutschland

[2] University of Regensburg, Regensburg, Deutschland

[3] FernUniversität in Hagen, 58084 Hagen, Deutschland

dellen durch die Bereitstellung von ML-Funktionalitäten in Datenverwaltungssysteme. Dadurch ergeben sich ganzheitliche Ansätze durch Haltung und Bearbeitung der Daten in einem System. Durch diese Integration werden zeitintensive Prozesse für die Extraktion, die Transformation und das Laden der Daten (ETL) in nur auf ML ausgelegte Systeme vermieden. Im vorliegenden Heft wird der Beitrag von Maximilian E. Schüle diese Kopplung vorstellen.

Die Forschungsgemeinschaft zu maschinellem Lernen für Systeme und Systemen für maschinelles Lernen ist ein stetig wachsender Teil des Fachbereiches. Deren Themen finden zunehmend Eingang in die Datenbank-Fachkonferenzen. Seit dem Jahr 2017 präsentiert der Workshop *Data Management for End-to-End Machine Learning (DEEM)* auf der *SIGMOD* ganzheitliche Ansätze, um maschinelles Lernen und Datenhaltung zu kombinieren. Im Jahr 2021 stand der Datenbanktrack auf der LWDA unter dem Titel *Data Engineering for Data Science* und parallel wurde ein Heft des Datenbankspektrums (Issue 3, 2021) zum gleichen Thema herausgegeben. Auf der BTW 2023 beleuchteten gleich zwei Workshops, *Data Engineering for Data Science* und *ML for Systems and Systems for ML* sowie ein Track im wissenschaftlichen Prgramm die Wechselwirkung zwischen maschinellem Lernen und Datenbanksystemen bzw. Data-Engineering. Das VLDB-Journal widmete diesem Forschungsbereich bereits eine Sonderedition (Juli 2024).

Zusammenfassend lässt sich festhalten, dass die Bereiche Datenbanken und maschinelles Lernen immer enger zusammenwachsen, sich gegenseitig bedingen und ergänzen. Die sich daraus ergebenden neuen Ansätze beleuchten wir in diesem Schwerpunktthema.

Aus den Einreichungen zum Heft konnten wir fünf Beiträge annehmen, die verschiedene Aspekte des Themas vorstellen. Zwei Beiträge sind Erweiterungen vorheriger Konferenzbeiträge, die die neuen Forschungsergebnisse der Themen repräsentieren, die anderen drei Artikel beinhalten vollständig neue Arbeiten. Wir danken allen Autoren für ihre Beiträge und den Gutachterinnen und Gutachtern, die ihre Fachexpertise eingebracht haben, für ihre wertvolle Unterstützung.

Im Einzelnen finden Sie folgende Beiträge in diesem Heft:

Sebastian Schelter, Shuba Guha und Stefan Grafberger beschreiben in ihrem Beitrag *Automated Provenance-Based Screening of ML Data Preparation Pipelines* eine Erweiterung von MLINSPECT. MLINSPECT wird entwickelt, um in Data-Preprocessing-Pipelines Tabellenattribute hinsichtlich technischer Datenbiase (Verzerrungen durch Über- oder Unterrepräsentation eines Wertes) zu überwachen. Das Tool unterstützt mehrere Data-Science-Bibliotheken und stellt die Informationen bereit, um Data-Provenance entlang der Datenvorverarbeitung zu überwachen. Das Anwendungsspektrum des Verfahrens ist weit: Vom Überwachen der Dateneigenschaften (und Erkennen von Biasen), dem Garantieren von Fairness in ML und der Sicherstellung der Anforderungen der DSGVO. Die vorliegende Implementierung validiert die Machbarkeit des Ansatzes und wurde an mehreren Anwendungsfällen gezeigt.

Ein Beitrag aus der Gruppe von Marina Tropmann-Frick entwickelt eine neue Lösung für eine der Standardaufgabe des Data-Preprocessing, die Vorhersage von fehlenden Werten (Missing-Value-Imputation). Zusammen mit Michael Jacobsen untersucht sie in dem Beitrag *Imputation Strategies in Time Series based on Language Models* das Potenzial verschiedener Large-Language-Models (LLMs) für die Datenimputation in Zeitreihen. Dabei wird ein Prozess vorgeschlagen, der verschiedene LLMs für diese Aufgabe einsetzt und umfassend vergleicht. Dabei werden sowohl die Ergebnisse des Einsatzes verschiedener LLMs untereinander als auch die Ergebnisse der Basisverfahren verglichen.

Die Arbeit von Maximilian E. Schüle, Thomas Neumann und Alfons Kemper mit dem Titel *The Duck's Brain: Training and Inference of Neural Networks within Database Engines* zeigt, wie ein einfaches neuronales Netzwerk in SQL implementiert werden kann. Die Arbeit erweitert ein auf der BTW 2023 erschienenes Papier um die Inferenz von Modellen in SQL und DuckDB als zusätzliches Backend. Der Beitrag spannt den großen Bogen von den ML-Technologie, deren formaler Beschreibung bis hin zum Python-Code. Dieser wird dann in eine relationale Darstellung in SQL übersetzt. Das kann entweder in Standard-SQL-92, SQL mit Array-Erweiterungen oder Fensterfunktionen erfolgen.

Im Beitrag *Multi-Modal Contextualization of Trajectory Data for Advanced Analysis* von Paul Walther, Fabian Deuser und Martin Werner wird eine neuartige Methode zur Analyse von Bewegungsdaten vorgestellt. Dabei werden Verfahren entwickelt, um multi-modale Daten zu verbinden; Bewegungsdaten werden dabei mit Rasterbildern, wie Luftbildern oder Vektorkarten, überlagert. Aus dieser Kombination ergeben sich bessere Interpretationsmöglichkeiten durch ML-Modelle sowie für Human-in-the-Loop-Verfahren in geographischen Anwendungen.

Tobias Schreier, Marina Tropmann-Frick und Ruwen Böhm untersuchen in ihrem Beitrag *Integration of FAERS, DrugBank and SIDER Data for Machine Learning-based Detection of Adverse Drug Reactions*, ob Modelle des maschinellen Lernens die Signalerkennung in der Pharmakovigilanz im Vergleich zur traditionell verwendeten Disproportionalitätsanalyse (DPA) verbessern können. Der Beitrag beschreibt einen Ansatz zur ML-basierten Erkennung statistisch signifikanter unerwünschter Wirkungen von Arzneimitteln, wobei Einzelfallberichte (single-case events reports, SER) und öffentlich verfügbare, arzneimittelbezogene Informationen verwendet werden.

Das vorliegende Heft bietet also vielfältige Blickwinkel auf das Thema *ML for Systems and Systems for ML.*

2 Kurz erklärt

In der Rubrik „Kurz erklärt“ führen Maximilian Plazotta und Meike Klettke in das Thema *Data Architectures in Cloud Environments* ein. Sie geben einen Überblick darüber, wie Cloud Computing im Bereich des Datenmanagements funktioniert. Das Zusammenspiel von Cloud Computing und Datenbankarchitekturen wird durch die Einführung von Design-Prinzipien, Tools und Services für Cloud-Datenarchitekturen erläutert. Dieser Beitrag ist ein erster Ausblick auf das nächste Themenheft „Cloud-Native Database Management Systems“ (DASP 1-2025).

3 Community-Beiträge

In der Rubrik „Datenbankgruppen vorgestellt“, wird der Lehrstuhl „Datenbanktechnologien und Datenanalytik“ von Lena Wiese an der Goethe-Universität Frankfurt vorgestellt. Die Gruppe forscht und lehrt insbesondere in den Bereichen NoSQL-Datenbanken, Graphdatenmodellierung, Sportanalytik und maschinelles Lernen im Gesundheitsbereich. Es freut uns besonders, in dieser Rubrik seit längerer Zeit wieder einen Beitrag präsentieren zu können. Einreichungen für diese Rubrik – gerade auch von neueren bzw. neu besetzten (aber natürlich auch von etablierten) Lehrstühlen und Arbeitsgruppen aus den Bereichen Datenbanken, Data Engineering, Data Analytics und Informatik Retrieval sind sehr willkommen.

Die Rubrik „Dissertationen“ enthält in diesem Heft sechzehn Kurzfassungen von Dissertationen aus der deutschsprachigen DBIS-Community, welche im letzten halben Jahr erfolgreich verteidigt wurden.

Im März 2025 findet an der Universität Bamberg die 21. BTW-Tagung statt. Nachdem in den letzten Heften bereits die Calls dieser Tagung veröffentlicht wurden, enthält dieses Heft in den „News“ einen Ausblick auf die Keynotes der BTW.

Außerdem findet sich in dieser Rubrik ein Bericht über die traditionelle LWDA-Konferenz, welche im September in Würzburg stattfand.

Der erste Platz im „Graduate Track“ der „ACM Student Research Competition Grand Finals“ ging in diesem Jahr erstmalig an einen Teilnehmer einer europäischen Universität – an Stefan Klessinger von der Universität Passau. Über diesen Erfolg berichten wir ebenfalls in den „News“

4 Künftige Schwerpunktthemen

4.1 Using LLMs for Data Management—Chances and Risks for Research and Education

The emergence of Large Language Models (LLMs) offers transformative opportunities for enhancing data management education and research, enabling personalized learning experiences and facilitating more intuitive data management practices. Educators can leverage LLMs to create engaging curricula that adapt to individual student needs, while researchers can utilize these models to explore complex datasets with unprecedented efficiency. However, as we harness the potential of LLMs, it is crucial to remain vigilant about the associated risks, particularly the tendency for models to generate inaccurate or misleading information, commonly referred to as hallucinations. These inaccuracies raise important challenges that must be addressed to ensure the reliability of LLM outputs in both teaching and research contexts.

This special issue seeks to explore the dynamic interplay between the benefits and risks of LLM integration, highlighting innovative applications, effective pedagogical approaches, and the necessary safeguards that can help mitigate potential downsides while maximizing the potential of LLMs in the field of data management and information retrieval.

Topics of interest include, but are not limited to:

1. Applications of LLMs in Data Management Teaching
   - Designing intelligent tutoring systems for data management courses.
   - Utilizing LLMs to create interactive learning environments for database concepts.
   - Case studies: Experiences with implementations of LLMs in teaching databases.
   - Facilitating collaborative projects and peer-to-peer learning using LLM support.
   - The use of LLMs in fostering communication among database students.
   - Resource generation for database courses using LLMs.
   - Developing adaptive learning materials with LLM assistance.
   - Creating assessments and quizzes powered by LLMs.
   - Evaluation of LLM performance for teaching
2. LLMs in Research Data Management
   - Automating data preparation, cleaning, and transformation using LLMs.
   - Enhancing data quality through LLM-driven suggestions and corrections.
   - Case studies: LLM applications in managing large datasets in academia.

3. Natural Language Querying of Databases
   - Exploring LLM capabilities for natural language processing of SQL commands.
   - Enhancing database querying with natural language interfaces.
   - User studies: Effectiveness of LLMs in assisting non-technical users.
4. Advanced Research Applications
   - Innovative uses of LLMs in database research methodologies.
   - Exploring LLMs in developing new database technologies and architectures.
   - The impact of LLMs on database optimization and performance research.
   - LLM support for schema design and model learning.
5. Ethical and Practical Considerations
   - Addressing biases in LLMs within database applications.
   - Ensuring data privacy and security when using LLMs in educational contexts.
   - Discussing the implications of LLM use on traditional teaching methods.
6. Future Directions and Challenges
   - Emerging trends in LLMs relevant to database education and research.
   - Challenges of integrating LLMs seamlessly into existing educational frameworks.
   - Predictions for the future of LLMs in the field of databases.

We welcome traditional research articles, experience and application reports, as well as system overviews, surveys, and experimental studies. These can either be in the form of full submissions (8-10 pages) as well as short papers/ extended abstracts (not more than 4 pages) for this issue.

Deadline for Submissions: April 1st, 2025.
Publication of special issue: DASP-2-2025 (September 2025).

Guest editors:
Richard Lenz, FAU Erlangen, richard.lenz@fau.de
Uta Störl, FernUni in Hagen, uta.stoerl@fernuni-hagen.de
Andreas Thor, HTWK Leipzig, andreas.thor@htwk-leipzig.de

4.2 Best Workshop Papers of BTW 2025

Deadline for Submissions: July 1st, 2025
Publication of special issue: DASP-3-2025 (December 2025)

Guest editor:
Uta Störl, FernUni in Hagen, uta.stoerl@fernuni-hagen.de

4.3 Bias in Information Systems

Deadline for Submissions: October 1st, 2025
Publication of special issue: DASP-1-2026 (March 2026)

Guest editors:
Philipp Schaer, TH Köln, philipp.schaer@th-koeln.de
Timo Spinde, Uni Göttingen, timo.spinde@uni-goettingen.de

Funding Open Access funding enabled and organized by Projekt DEAL.