

List of Seminar Topics

Last update: Apr 19, 2024

- Papers in the fields of **systems** for **data engineering**, **data management**, and **machine learning**
- This Semester's **Umbrella Topic**: **“Efficiently Combining DB and ML Workloads”**
 - **Motivation**
 - Database **query processing** and **ML training and scoring** normally executed in **dedicated systems**
 - But: Trend towards **integrated data analysis pipelines** involving both query processing and ML
 - Orchestrations of existing DB and ML systems yields **inefficiencies** due to expensive data transfer and missed global optimization potential
 - **Ideas to address these challenges**
 - **Improve the data transfer** between DB and ML systems
 - Run **one kind of workload on existing software/hardware designed for the other kind of workload**
 - Entirely **new systems supporting both** query processing and ML at the same time
 - Affects **all levels of the system stack**, from query languages over optimization and compilation techniques as well as local/distributed runtime techniques to the use of multi-core CPUs and hardware accelerators.

▪ Efficient Data Transfer

- 1) Prasad et al.: **Large-scale Predictive Analytics in Vertica: Fast Data Transfer, Distributed Model Creation, and In-database Prediction** (SIGMOD, 2015) [[link](#)]
- 2) Raasveldt et al.: **Don't Hold My Data Hostage – A Case For Client Protocol Redesign** (PVLDB, 2017) [[link](#)]
- 3) Wang et al.: **ConnectorX: Accelerating Data Loading From Databases to Dataframes** (PVLDB, 2022) [[link](#)]

▪ ML Through DBMS Extensibility

- 4) Feng et al.: **Towards a Unified Architecture for In-RDBMS Analytics** (SIGMOD, 2012) [[link](#)]
- 5) Wolf et al.: **Extending Database Task Schedulers for Multi-threaded Application Code** (SSDBM, 2015) [[link](#)]
- 6) Sichert et al.: **User-Defined Operators: Efficiently Integrating Custom Algorithms into Modern Databases** (PVLDB, 2022) [[link](#)]

■ ML expressed in SQL

- 7) Hellerstein et al.: **The MADlib Analytics Library or MAD Skills, the SQL** (PVLDB, 2012) [[link](#)]
- 8) Luo et al.: **Scalable Linear Algebra on a Relational Database System** (ICDE, 2017) [[link](#)]
- 9) Gao et al.: **The BUDS Language for Distributed Bayesian Machine Learning** (SIGMOD, 2017) [[link](#)]
- 10) Schüle et al.: **In-Database Machine Learning with SQL on GPUs** (SIGMOD, 2021) [[link](#)]
- 11) Luo et al.: **Automatic Optimization of Matrix Implementations for Distributed Machine Learning and Linear Algebra** (SIGMOD, 2021) [[link](#)]
- 12) Tang et al.: **Auto-Differentiation of Relational Computations for Very Large Scale Machine Learning** (ICML, 2023) [[link](#)]
- 13) Paganelli et al.: **Pushing ML Predictions into DBMSs** (IEEE Trans. Know. Data Eng.) [[link](#)]

Seminar Topics (3/5)



▪ Learning Over Joins

- 14) Kumar et al.: **Learning Generalized Linear Models Over Normalized Data** (SIGMOD, 2015) [[link](#)]
- 15) Schleich et al.: **Learning Linear Regression Models over Factorized Joins** (SIGMOD, 2016) [[link](#)]
- 16) Chen et al.: **Towards Linear Algebra over Normalized Data** (PVLDB, 2017) [[link](#)]
- 17) Schleich et al.: **A Layered Aggregate Engine for Analytics Workloads** (SIGMOD, 2019) [[link](#)]

▪ Hybrid DB+ML Systems

18) Kernert et al.: **SLACID - Sparse Linear Algebra in a Column-Oriented In-Memory Database System**

(SSDBM, 2014) [[link](#)]

19) Kunft et al.: **BlockJoin: Efficient Matrix Partitioning Through Joins** (PVLDB, 2017) [[link](#)]

20) Aberger et al.: **LevelHeaded: A Unified Engine for Business Intelligence and Linear Algebra Querying**

(ICDE, 2018) [[link](#)]

21) Kunft et al.: **An Intermediate Representation for Optimizing Machine Learning Pipelines** (PVLDB, 2019) [[link](#)]

22) Jasny et al.: **DB4ML - An In-memory Database Kernel with Machine Learning Support** (SIGMOD, 2020) [[link](#)]

23) Jungmair et al.: **Designing an Open Framework for Query Optimization and Compilation** (PVLDB, 2022) [[link](#)]

Seminar Topics (5/5)



- **Query Processing on Tensors**

- 24) He et al.: **Query Processing on Tensor Computation Runtimes** (PVLDB, 2022) [[link](#)]

- 25) Hu et al.: **TCUDB: Accelerating Database with Tensor Processors** (SIGMOD, 2022) [[link](#)]

- **Various Approaches**

- 26) Kläbe et al.: **Exploration of Approaches for In-database ML** (EDBT, 2023) [[link](#)]