Stampfenbachstrasse 114
8092 Zürich, Switzerland
maxi@boether.de
www.mboether.com
MaxiBoether

Maximilian Böther

I work on the intersection of systems and data-centric AI. My research interests include large-scale LLM/VLM training, data management for machine learning, and machine learning pipelines.

Education

Nov 2022 – *present* **Ph.D. in Computer Science**, *ETH Zurich*, Zurich, Switzerland
- Working on data-centric machine learning systems, supervising students, supporting lectures
- Supervisors: Ana Klimovic and Gustavo Alonso (Systems Group)

Oct 2020 – Sep 2022 **M.Sc. in IT-Systems Engineering**, *Hasso Plattner Institute*, Potsdam, Germany
- GPA: 1.0/1.0, Focus on machine learning and data processing
- Thesis: Designing a CPU-aware Hash Table for Hash Joins (published at VLDB'23)

Oct 2017 – Sep 2020 **B.Sc. in IT-Systems Engineering**, *Hasso Plattner Institute*, Potsdam, Germany
- GPA: 1.1/1.0, Thesis: Heuristic Optimization of Strategic Alternative Routes in Traffic Networks Using Evolutionary Algorithms (published at GECCO'21, **Best Paper Award**)

Aug 2009 – Jun 2017 **Abitur**, *Goethegymnasium*, Hildesheim, Germany
- GPA: 1.0/1.0 (884/900 points, fourth-best student of entire state)

Experience

Nov 2022 – *present* **Doctoral Research Assistant**, *ETH Zurich*, Zurich, Switzerland

Sep 2025 – *present* **Research Intern**, *DatologyAI*, Zurich, Switzerland and Redwood City, CA, USA

May 2025 – Aug 2025 **Machine Learning Intern**, *Apple*, Seattle, WA, USA
- Ported verl (Volcano Engine Reinforcement Learning Engine) to Apple infrastructure
- Extended verl to support flow/diffusion models via the FlowGRPO algorithm and conducted research on improving text generation using VLMs as reward models

Jun 2023 – Aug 2023, Nov 2023 – Feb 2024 **Student Researcher**, *Google*, Sunnyvale, CA, USA, and Zurich, Switzerland
- Designed, implemented, and optimized an distributed submodular greedy optimization algorithm for finding representative subsets of billion-scale datasets. Published paper at MLSys'25.
- Built daff, a tool for DAta eFFiciency, integrating internal services for data selection

Jul 2018 – Aug 2022 **Teaching and Research Assistant**, *Hasso Plattner Institute*, Potsdam, Germany
- TA for Mathematics 1/2, RA in Data Engineering Systems Group, system administrator for HPC servers

Aug 2020 - Oct 2020 **Entrepreneur in Residence Intern**, *EMIL Group GmbH*, Berlin, Germany
- Used QuickSight, R, and Python to create insurance benchmark, used immediately by two major insurers

Dec 2016 – Aug 2020 **Member of Federal Board**, *Junge Liberale e.V.*, Berlin, Germany
- Position in the board of a political youth organization with > 10 000 members
- Responsible for coordinating the IT team and mediation between board and IT

Jul 2017 – Jul 2020 **Software Engineer (part-time)**, *Universum AG*, Berlin, Germany
- Developed social-media software used in a German federal election campaign (Javascript, Elastic)
- Implementation of e-learning platform for major German textbook publisher (PHP, Drupal)

Scholarships & Awards

May 2025 **MLCommons ML and Systems Rising Stars Award**

Sep 2023 **HPI Graduate Award**, *Award for my academic achievements during my master studies at HPI*

Oct 2020 – Sep 2022 **Hasso Plattner Scholar**, *Scholarship for graduating top of the year of my bachelor class at HPI*

Nov 2017 – Sep 2022 **Scholar of the German Academic Scholarship Foundation (Studienstiftung)**, *Germany's largest, oldest and most prestigious scholarship foundation*

Jul 2021 **GECCO'21 Best Paper Award**

Jul 2021 **ISMB/ECCB'21 Best Poster Award**

Nov 2019 – Nov 2020 **IT-Talents Scholar**, *Sponsor: Robert Bosch GmbH*

Oct 2017 – Oct 2020 **Scholar of the Friedrich-Naumann-Foundation for Freedom**

Sep 2020 **MLP Scholarship**

Sep 2019 **Winner HackZurich 2019 Challenge 'LegalTech'**, *Europe's largest hackathon*

Dec 2018 – May 2019 **Kearney Scholar**

Sep 2017 – Sep 2018 **IT-Talents Scholar**, *Sponsor: EBP Deutschland GmbH*

2016 **1st Prize Jugend Forscht "Youth Researches" Regional Competition Hildesheim**

Skills

Languages	Python, C++
ML Stack	PyTorch, Torchtitan, Pandas, Numpy, scikit-learn
Tools/Libraries	OpenMP, OpenMPI, Ray, gRPC, Beam, Spark, Gurobi, LaTeX
Software	Docker, Apache HTTP Server, nginx, PostgreSQL, Postfix, VMware vSphere, Ansible

Selected Projects

Nov 2024 – *present* **Apertus: Switzerland's first large-scale open, multilingual language model**
I am involved in training the Apertus-8B and -70B LLMs, and support the *pretraining data* efforts.
Huggingface: `huggingface.co/swiss-ai`
Report: `github.com/swiss-ai/apertus-tech-report`

Jan 2024 – *present* **Mixtera: A Data Plane for Foundation Model Training**
Mixtera enables declarative specification and dynamic adjustment of training data mixtures across abitrary properties for LLM/VLM training.
GitHub: `github.com/eth-easl/mixtera`
Paper (preprint): `arxiv.org/abs/2502.19790`

Nov 2022 – *present* **Modyn: A Research Platform for Dynamic Datasets**
Modyn is a end-to-end platform for model training on datasets that grow over time, enabling exploration of triggering and data selection policies.
GitHub: `github.com/eth-easl/modyn`
Paper (SIGMOD'25): `arxiv.org/abs/2312.06254`

Publications

June 2025 **Modyn: Data-Centric Machine Learning Pipeline Orchestration**
M. Böther, T. Robroek, V. Gsteiger, R. Holzinger, X. Ma, P. Tözun, A. Klimovic. In Proceedings of the Conference on Management of Data (**SIGMOD**)'25.

May 2025 **On Distributed Larger-Than-Memory Subset Selection With Pairwise Submodular Functions**
M. Böther, A. Sebastian, P. Awasthi, A. Klimovic, S. Ramalingam. In Proceedings of the Conference on Machine Learning and Systems (**MLSys**)'25.

Mar 2025 **Mixtera: A Data Plane for Foundation Model Training**
M. Böther, X. Yao, T. Kerimoglu, A. Klimovic. arXiv preprint.

Nov 2024 **Decluttering the data mess in LLM training**
M. Böther, D. Graur, X. Yao, A. Klimovic. In Non-Archival Proceedings of the Workshop on Hot Topics in System Infrastructure (**HotInfra**) @ **SOSP**'24.

May 2024 **Deploying Data Selection Techniques on Dynamic Datasets**
M. Böther, A. Klimovic. In Non-Archival Proceedings of the **DMLR** Workshop @ **ICLR**'24.

Jun 2023 **Analyzing Vectorized Hash Tables Across CPU Architectures**
M. Böther, L. Benson, A. Klimovic, T. Rabl. In Proceedings of the **VLDB** Endowment 16 (11).

Apr 2023 **Towards A Platform and Benchmark Suite for Model Training on Dynamic Datasets**
M. Böther, F. Strati, V. Gsteiger, A. Klimovic. In Proceedings of the Workshop on Machine Learning and Systems (**EuroMLSys**)'23.

Jan 2023 **Efficiently Computing Directed Minimum Spanning Trees**
M. Böther, O. Kißig, C. Weyand. In Proceedings of the Symposium on Algorithm Engineering and Experiments (**ALENEX**)'23.

Jun 2022 **Law Smells - Defining and Detecting Problematic Patterns in Legal Drafting**
C. Coupette, D. Hartung, J. Beckedorf, *M. Böther*, D.M. Katz. In: Artificial Intelligence and Law 32 (2).

Apr 2022 **What's Wrong with Deep Learning in Tree Search for Combinatorial Optimization**
M. Böther, O. Kißig, M. Taraz, S. Cohen, K. Seidel, T. Friedrich. In Proceedings of the International Conference on Learning Representations (**ICLR**)'22.

Jul 2021 **Evolutionary Minimization of Traffic Congestion**
M. Böther, L. Schiller, P. Fischbeck, L. Molitor, M. Krejca, and T. Friedrich. In Proceedings of the Genetic and Evolutionary Computation Conference (**GECCO**)'21. **Best Paper Award**. Also in: IEEE TEVC 27 (6).

Jul 2021 **Learning Languages with Decidable Hypotheses**
J. Berger, *M. Böther*, V. Doskoč, J. GadeaHarder, N. Klodt, T. Kötzing, W. Lötzsch, J. Peters, L. Schiller, L. Seifert, A. Wells, and S. Wietheger. In Proceedings of the Conference on Computability in Europe (**CiE**)'21.

Jun 2021 **Drop It In Like It's Hot: An Analysis of Persistent Memory as a Drop-in Replacement for NVMe SSDs**
M. Böther, O. Kißig, L. Benson, and T. Rabl. In Proceedings of the International Workshop on Data Management on New Hardware (**DaMoN**) @ **SIGMOD**'21.

Jun 2021 **Scale-Down Experiments on TPCx-HS**
M. Böther and T. Rabl. In Proceedings of the Workshop on Big Data in Emergent Distributed Environments (**BiDEDE**) @ ACM **SIGMOD**'21.

Oct 2020 **Maps for Learning Indexable Classes**
J. Berger, *M. Böther*, V. Doskoč, J. GadeaHarder, N. Klodt, T. Kötzing, W. Lötzsch, J. Peters, L. Schiller, L. Seifert, A. Wells, and S. Wietheger. In: Computability 13 (3-4).

Sep 2020 **A Strategic Routing Framework and Algorithms for Computing Alternative Paths**
T. Bläsius, *M. Böther*, P. Fischbeck, T. Friedrich, A. Gries, F. Hüffner, O. Kißig, P. Lenzner, L. Molitor, L. Schiller, A. Wells, and S. Wietheger. In Proceedings of the Symposium on Algorithmic Approaches for Transportation Modelling, Optimizations, and Systems (**ATMOS**)'20.

Teaching

Lectures and Courses

Cloud Computing Architecture: Spring 2023, Spring 2024, Spring 2025

Systems Programming and Computer Architecture: Autumn 2023, Autumn 2024 (Head TA), Autumn 2025 (Head TA)

Distributed Systems Lab: Autumn 2022

Individual Supervision

Master Theses Francesco Deaglio (MA Data Science): Data Selection in Modyn, Jingyi Zhu (MA Computer Science): Implementation and comparison of model retraining triggering policies, Xianzhe Ma (MA Computer Science): Exploring data selection under distribution shift, Tolga Kerimoğlu (MA Computer Science): Multimodality in Mixtera, John Staib Matilla (MA Computer Science): LLM Finetuning in Modyn, George Manos (MA Computer Science): GRADiator: Efficient Gradient Logging for Gradient Analytics at Scale

Bachelor Theses Robin Oester (BA Informatik): Model Management and Evaluation in Modyn, Robin Holzinger (BA Informatik at TU Munich): An Analysis of Drift- and Cost-Aware ML Retraining Triggering Policies in Modyn

Individual Research Viktor Gsteiger: System Optimizations in Modyn, Beste Güney: Dynamic Mixture in Mixtera

Activities and Service

2025 **Program Commitee Member**, *DEEM Workshop @ SIGMOD 2025*

2025 **Program Commitee Member**, *aiDM Workshop @ SIGMOD 2025*

2024 **Reviewer**, *DMLR Workshop @ ICLR 2024*

2021 – 2022 **Member of Appointment Committee for Professorship Digital Technology, Governance and Policy**, *Hasso Plattner Institute, University of Potsdam*

2020 – 2022 **Member of Appointment Committee for Professorship Digital Health and AI**, *Hasso Plattner Institute, University of Potsdam*