

Chunwei Liu

Postdoctoral Associate
Computer Science and Artificial Intelligence Laboratory (CSAIL)
Massachusetts Institute of Technology (MIT)

chunwei@csail.mit.edu -- 872.904.8134
Room 32-G930, 32 Vassar St.
Cambridge, MA 02139

Research Interests

I am interested in research spanning compound AI systems, database systems, cloud/edge computing, and database benchmarking. My focus includes integrating AI-powered analytics with database optimization, developing privacy-preserving techniques for cloud database benchmarking, and inventing novel data compression methods. I also implement ML-based adaptive compression in both standard and resource-constrained environments and conduct high-dimensional data analysis with a focus on time series applications.

Keywords: Compound AI Systems, LLM, Data Compression, Time Series, Edge, IoT

Professional Experience

MIT CSAIL

Advisors: Michael Cafarella (Sam Madden and Tim Kraska)

Postdoctoral Associate

Oct. 2022-Now

- Led the Palimpzest project, optimizing AI-powered analytics with declarative query processing and database techniques.
- Led the ClouDBench project, evaluating cloud databases with realistic benchmarks via industry collaborations.
- Conducted research on AI for scientific data discovery, information extraction, and model reuse under DARPA ASKEM and ARPA-H BDF projects.

The University of Chicago

Advisors: Aaron J. Elmore

Research Assistant

Oct. 2016-Sep. 2022

- Devised novel compression techniques with enhanced in-situ query support and compression performance.
- Developed adaptive compression selection and query execution frameworks for both server and edge hardware.
- Proposed new embeddings and indices for time series distance measurement and benchmarking.

Microsoft GSL

Mentors: Brandon Haynes, Matteo Interlandi | Manager: Alekh Jindal Microsoft Research Gray Systems Lab, Redmond, WA

Research Intern

Jun. 2021-Sep. 2021

- Applied encoding techniques to develop an encoding-aware caching system, accelerating query processing and minimizing space overhead in analytical data systems.

Education

University of Chicago, Chicago, IL

Thesis: *Fast and Effective Compression for IoT Systems*

Ph.D. in Computer Science; Systems Concentration

Degree Received: Oct. 2022

Advisor: Aaron J. Elmore

University of Chinese Academy of Sciences, Beijing, China

M.Sc.Eng. in Computer Science; Minors in Management of Technology

Degree Received: Jul. 2016

Advisors: Jizhou Tong, Yi Sun

Beihang University, Beijing, China

B.Eng. in Software Engineering;

Degree Received: Jul. 2013

Advisor: Jian Huang

Selected Refereed Publications

- **Chunwei Liu**, Matthew Russo, Michael Cafarella, Lei Cao, Peter Baile Chen, Zui Chen, Michael Franklin, Tim Kraska, Samuel Madden, Rana Shahout, Gerardo Vitagliano, Palimpzest: Optimizing AI-Powered Analytics with Declarative Query Processing, CIDR 2025.
- Rana Shahout, Eran Malach, **Chunwei Liu**, Weifan Jiang, Minlan Yu, Michael Mitzenmacher. Don't Stop Me Now: Embedding Based Scheduling for LLMs. ICLR 2025.

- Markos Markakis, Brit Youngmann, Trinity Gao, Ziyu Zhang, Rana Shahout, Peter Baile Chen, **Chunwei Liu**, Ibrahim Sabek, Michael Cafarella, From Logs to Causal Inference: Diagnosing Large Systems, VLDB 2025.
- Peter Baile Chen, Yi Zhang, **Chunwei Liu**, Sejal Gupta, Yoon Kim, Michael Cafarella, MDCR: A Dataset for Multi-Document Conditional Reasoning, EMNLP 2024.
- **Chunwei Liu**, John Paparrizos, Aaron J. Elmore, AdaEdge: A Dynamic Compression Selection Framework for Resource Constrained Devices, ICDE 2024.
- Markos Markakis, Ziyu Zhang, Rana Shahout, Trinity Gao, **Chunwei Liu**, Ibrahim Sabek, Michael Cafarella, Press ECCS to Doubt (Your Causal Graph), Proceedings of the Conference on Governance, GUIDE-AI@SIGMOD 2024. **Best Paper Award**
- Markos Markakis, An Bo Chen, Brit Youngmann, Trinity Gao, Ziyu Zhang, Rana Shahout, Peter Baile Chen, **Chunwei Liu**, Ibrahim Sabek, Michael J. Cafarella, Sawmill: From Logs to Causal Diagnosis of Large Systems, SIGMOD Conference Companion 2024.
- **Chunwei Liu**, Anna Pavlenko, Matteo Interlandi, Brandon Haynes, A Deep Dive into Common Open Formats for Analytical DBMSs, Proceedings of the VLDB Endowment, 16(11), 3044-3056, 2023. **Best Paper Runner-Up**
- John Paparrizos, **Chunwei Liu**, Aaron J. Elmore, Michael J. Franklin, Querying Time-Series Data: A Comprehensive Comparison of Distance Measures, Data Engineering, 69-88, 2023.
- John Paparrizos, Ikraduya Edian, **Chunwei Liu**, Aaron Elmore, Michael J. Franklin, Fast Adaptive Similarity Search through Variance-Aware Quantization, IEEE 38th International Conference on Data Engineering (ICDE), 2969-2983, 2022.
- **Chunwei Liu**, Hao Jiang, John Paparrizos, Aaron J. Elmore, Decomposed Bounded Floats for Fast Compression and Queries, VLDB, 14(11), 2586-2598, 2021.
- John Paparrizos, **Chunwei Liu**, Bruno Barbarioli, Johnny Hwang, Ikraduya Edian, Aaron J. Elmore, Michael J. Franklin, Sanjay Krishnan, VergeDB: A Database for IoT Analytics on Edge Devices, CIDR 2021.
- Hao Jiang, **Chunwei Liu**, John Paparrizos, Andrew A. Chien, Jihong Ma, Aaron J. Elmore, Good to the Last Bit: Data-Driven Encoding with CodecDB, Proceedings of the 2021 ACM SIGMOD, 843-856, 2021.
- John Paparrizos, **Chunwei Liu**, Aaron J. Elmore, Michael J. Franklin, Debunking Four Long-Standing Misconceptions of Time-Series Distance Measures, Proceedings of the 2020 ACM SIGMOD, 1887-1905, 2020.
- Hao Jiang, **Chunwei Liu**, Qi Jin, John Paparrizos, Aaron J. Elmore, PIDS: Attribute Decomposition for Improved Compression and Query Performance in Columnar Storage, Proceedings of VLDB Endowment, 13(6), 925-938, 2020.
- **Chunwei Liu**, McKade Umbenhowe, Hao Jiang, Pranav Subramaniam, Jihong Ma, Aaron J. Elmore, Mostly Order Preserving Dictionaries, 2019 IEEE 35th International Conference on Data Engineering (ICDE), 1214-1225, 2019.

Preprint and In Submission

- **Chunwei Liu**, Anna Pavlenko, Matteo Interlandi, Brandon Haynes. Data Formats in Analytical DBMSs: Performance Trade-offs and Future Directions. Under Review VLDBJ, 2024. Invited to the “**Best of VLDB 2023**” special issue of VLDB Journal. arXiv:2411.14331.
- **Chunwei Liu**, Gerardo Vitagliano, Brandon Rose, Matt Prinz, David Andrew Samson, Michael Cafarella. (2025). PalimpChat: Declarative and Interactive AI Analytics. arXiv preprint arXiv:2502.03368.
- **Chunwei Liu**, Enrique Noriega-Atala, Adarsh Pyarelal, Clayton T Morrison, Mike Cafarella. Variable Extraction for Model Recovery in Scientific Literature. Under Round 2 Review, AAAI, 2025. arXiv:2411.14569.
- **Chunwei Liu**, Matthew Russo, Gerardo Vitagliano, Samuel Madden, Michael Cafarella. Enhancing Document Extraction with LLMs: Fine-Grained Indexing and Attention-Driven Token Reduction, In Preparation, 2024.
- **Chunwei Liu**, Matthew Russo, Michael Cafarella, Lei Cao, Peter Baile Chen, Zui Chen, Michael Franklin, Tim Kraska, Samuel Madden, Gerardo Vitagliano. (2024). A Declarative System for Optimizing AI Workloads. arXiv preprint arXiv:2405.14696.
- **Chunwei Liu**, Yan Zhou, Samuel Madden, Tim Kraska, Ju Fan. PBench: Workload Synthesizer with Real Statistics for Cloud Analytics Benchmarking. Under Review, 2024.
- Ziniu Wu, Markos Markakis, **Chunwei Liu**, Peter Baile Chen, Balakrishnan Narayanaswamy, Tim Kraska, Samuel Madden. Improving DBMS Scheduling Decisions with Fine-grained Performance Prediction on Concurrent Queries. Under Review, SIGMOD 2025.
- Kaisei Hishida, **Chunwei Liu**, John Paparrizos, Aaron Elmore. Beyond Compression: A Comprehensive Evaluation of Lossless Floating-Point Compression. Under Review, SIGMOD 2025.

- Zui Chen, Lei Cao, Samuel Madden, Ju Fan, Nan Tang, Zihui Gu, Zeyuan Shang, **Chunwei Liu**, Michael Cafarella, Tim Kraska. (2023). Seed: Simple, efficient, and effective data management via large language models. arXiv preprint arXiv:2310.00749.

Awards and Honors

GUIDEAI@SIGMOD Best Paper Award	2024
VLDB Best Paper Runner-Up, EA&B Track	2023
University Unrestricted (UU) Fellowship - The University of Chicago	2019
Outstanding Postgraduate Students Award	2015,2014
Excellent Student Leader Scholarship	2015,2014
Outstanding Graduates Award	2013

Service

- Program Committee of SIGMOD'2025; VLDB'2024, 2025; KDD'2023; SMDB'2023
- Reviewer of The International Journal on Very Large Data Bases (VLDBJ)
- SIGMOD Availability & Reproducibility Committee 2022, 2023

TA Experience

Introduction to Databases	<i>University of Chicago</i>	Winter 2019, 2020, 2021
Introduction to Computer Science II	<i>University of Chicago</i>	Spring 2017
Parallel Computing	<i>University of Chicago</i>	Winter 2017
Operating Systems	<i>University of Chicago</i>	Autumn 2016
Mining Massive Data	<i>University of Chinese Academy of Sciences</i>	Spring 2015
Computer Network	<i>University of Chinese Academy of Sciences</i>	Autumn 2014

Selected Mentoring Experience

McKade Umbenhower	Compression	University of Wyoming, B.S., 2019,	Next: M.S. at Carnegie Mellon University
Ikraduya Edian	Auto Migration	Bandung Institute of Technology, B.S., 2020,	Next: M.S. at University of Edinburgh
Pranav Subramaniam	Encoding Evaluation	University of Chicago B.S., 2020,	Next: Ph.D. at University of Chicago
Milka Piszczek	Storage Disaggregated System	M.S. at the MIT, 2022	Next: Datadog
Sarah Wang	Database Benchmarking	MIT B.S., 2023,	Next: M.S. at the MIT
Kaisei Hishida	Compression for ML	Keio University B.S., 2024,	Next: TBD
Sejal Gupta	PlotQA	M.S. at the MIT, 2024	Next: Sutter Hill Ventures
Sabiyah Ali	Database Benchmarking	M.S. at the MIT, 2024	Next: Ph.D Program TBD

Referees

Michael J. Cafarella

michjc@csail.mit.edu
Principal Research Scientist,
MIT CSAIL

Samuel Madden

madden@csail.mit.edu
CS Head, MIT EECS Department
MIT College of Computing,
Distinguished Professor, MIT CSAIL

Aaron J. Elmore

aelmore@cs.uchicago.edu
Associate Professor,
The University of Chicago

Michael J. Franklin

mjfranklin@uchicago.edu
Morton D. Hull Distinguished Service
Professor, The University of Chicago

John Paparrizos

paparrizos.1@osu.edu
Assistant Professor,
Director, DATUM Lab,
The Ohio State University