Yiming Lin

✉ *yiminglin@berkeley.edu*

Education and Employment

2023–Present **Postdoctoral Researcher**, *EECS, University of California, Berkeley*.
2022 summer **Research Intern**, *Microsoft Research*.
2021 summer **Applied Scientist Intern**, *Amazon*.
2017–2023 **Ph.D**, *Dept. of Computer Science, University of Calfornia Irvine*.
2015–2017 **Master of Science**, *Computer Science and Technology, Harbin Institute of Technology*.
2011–2015 **Bachelor of Science**, *Computer Science and Technology, Harbin Institute of Technology*.

Areas & Technical Skills

Areas **Unstructured Data Analytics, Query Processing and Optimization, Data cleaning**.
Advisor Prof. Sharad Mehrotra at UC, Irvine, and Prof. Aditya Parameswaran at UC, Berkeley.

Experiences

2026 **Reviewer** at **SIGMOD** 2025, 2026, 2027, **ICDE** 2024, 2026.
2024 Travel Award for **VLDB** 2024, **ICDE** 2023.
2023 **Winner of the First Place** in **ASTRIDE@ICDE 2023** workshop competition.
2022 **Research Intern** at **Microsoft Research**.
2021 **Applied Scientist Intern** at **Amazon**.
2020-2023 Recipient of **Hasso-Plattner-Institute**(**HPI**) Fellowship, 2020,2021,2022,2023.
Nov. 2016 **National Scholarship** (**TOP 1%**)

Publications

[1] **Yiming Lin**, Sepanta Zeighami, Yash Jain, HC Moore, Aditya G. Parameswaran. Bolt-on, Verifiable Provenance for LLM-Powered Data Processing. Under review in **VLDB 2026**.

[2] **Yiming Lin**, Mawil Hasan, Rohan Kosalge, Alvin Cheung, Aditya G. Parameswaran. TWIX: Automatically Reconstructing Structured Data from Templatized Documents. In **SIGMOD 2026**.

[3] **Yiming Lin**, Madelon Hulsebos, Ruiying Ma, Shreya Shankar, Sepanta Zeighami, Aditya G. Parameswaran, Eugene Wu. Towards Accurate and Efficient Document Analytics with Large Language Models. In **ICDE 2025**.

[4] Zeighami, Sepanta, **Yiming Lin**, Shreya Shankar, and Aditya Parameswaran. LLM-Powered Proactive Data Systems. In **IEEE Bulletin 2025**.

[5] Shreya Shankar, Haotian Li, Parth Asawa, Madelon Hulsebos, **Yiming Lin**, J.D. Zamfirescu-Pereira, Harrison Chase, Will Fu-Hinthorn, Aditya G. Parameswaran, Eugene Wu. SPADE: Synthesizing Data Quality Assertions for Large Language Model Pipelines. In **PVLDB 2024** (Industry).

[6] **Yiming Lin**, Sharad Mehrotra. Towards Automatic Predicate Learning at Query Time. In **SIGMOD 2024**.

[7] **Yiming Lin**, Sharad Mehrotra. ZIP: Lazy Imputation during Query Processing. In **PVLDB 2023**.

[8] **Yiming Lin**, Yeye He, Surajit Chaudhuri. Auto-BI: Automatically Build BI-Models Leveraging Local Join Prediction and Global Schema Graph. In **PVLDB 2023**.

[9] Rithwik Kerur, **Yiming Lin**. Robust Occupancy Computation Based on WiFi Connectivity Events. **ICDE@ASTRIDE 2023**, **Winner of the First Place** in Workshop Competition.

[10] **Yiming Lin**, Daokun Jiang, Roberto Yus, Andrew Chio, Georgios Bouloukakis, Sharad Mehrotra, Nalini Venkatasubramanian: LOCATER: Cleaning WiFi Connectivity Datasets for Semantic Localization. **PVLDB 2021**.

[11] **Yiming Lin**, Pramod Khargonekar, Sharad Mehrotra, Nalini Venkatasubramanian: T-Cove: An exposure tracing System based on Cleaning Wi-Fi Events on Organizational Premises. **PVLDB** (Demo), **2021**.

[12] **Yiming Lin**, Hongzhi Wang, Jianzhong Li, Hong Gao: Efficient entity resolution on heterogeneous records. **ICDE 2020**.

Teaching Assistant Experiences

2018-19 Discrete Mathematics, Information and Computer Science Department. (3 Quarters)

Fall-2017 Introduction to Database, Information and Computer Science Department.

Fall-2016 Computational Complexity Theory. CSTI.

2014-15 C/C++ programming. (2 Quarters)

Projects and Impact

2018-2023 **Efficient and Low-cost Table Analytics using AI/ML**.

Driven by a real-world testbed, TIPPERS, at UC Irvine—which stores massive sensor streams from over 40 buildings—we built systems for accurate table analysis to support location-based services by developing LOCATER and T-COVE using AI and ML. We also built database systems with efficient query processing techniques (ZIP and PLAQUE) to bring down the large latency and cost introduced by AI operations across to support multiple downstream applications.

- **LOCATER** and **T-COVE**, tools to infer location and occupancy from WiFi logs using AI, have been deployed across **five sites** in **two countries over six years**, including **over 40 buildings** across three universities (UCI, BSU, and Plaksha University in India), a living facility (Walnut Village in Orange County), and the U.S. Navy.
- **T-COVE** has been adopted as **Covid-19 protection strategy** over **30 buildings at UC Irvine** to display the real-time occupancy in campus building for **over 3 years**.
- **PLAQUE** learns predicates that are not present in a given SQL query, speeding up query execution by up to **33×**, and achieving up to **100× speedups** when AI operations are involved. PLAQUE is being adopted in two commercial systems: **Couchbase** and **Amazon Redshift**.
- **ZIP** performs only the AI operations necessary to answer a given SQL query on tables, leading to a **19,607× speedup** compared to the strategy that executes all AI operations offline.

2024-present **Effective and Confident Document Analytics Powered by AI**.

Driven by real-world documents from our colloborators from industry, journalists, etc, we worked on developing accurate document analytics by exploring varied document structures. We identified three major types of document structures built specialized systems—TWIX, ZenDB, SHED, and LSF—to extract and exploit them for accurate, efficient, and theoretically grounded analytics. We built a document ingestion pipeline by clustering documents by structure and processes them with the corresponding tools.

Impact
- Open-source project (such as TWIX, SHED): **over 200 GitHub stars** within a month of release.
- TWIX helps our collaborators to process over **9,600 pages** of court dispositions with CA Public Defenders, **6,500+ pages** of civic reports and invoices with journalists at Stanford's Big Local News, and **4,300+ pages** of police use-of-force and employee records with the CA Police Records CLEAN team, transforming heterogeneous documents into structured data.

Project Details
- **TWIX** focuses on **Form-Like Templatized Documents**, which are programmatically generated from structured data using the same visual template, such as invoices and tax documents, etc. TWIX infers the underlying structure and further extracts structured data from documents accurately at no cost, **520× faster and 3,786× cheaper** than the most competitive compared tool, gpt-4-vision.
- **ZenDB** and **SHED** focuses on **Hierarchically Structured Documents**, with semantic hierarchical structures like scientific and legal documents. They infer this structure with a theoretical correctness guarantees coupled with effective tree search algorithms for downstream analysis, enabling highly accurate retrieval: **up to 61% higher precision and 81% higher recall** than RAG at a marginally higher cost.
- We built **BLIP** to help gain trust **how answers are produced and whether those answers are correct.** BLIP is guaranteed to return a **minimal verifiable provenance**, a subset of the document that can reproduce the same answer as the one obtained from the full document when queried using the same LLM, **30+ higher accuracy** than the best-performing baseline.