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.*

2011–2017 **Bachelor and Master of Science**, *Computer Science, 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 **Effective Table Ingestion for Interactive Analytics**.

Driven by the real-world TIPPERS testbed at UC Irvine, which stores sensor streams from over 40 buildings (about 30k logs per minute), we built systems for accurate table ingestion to compute indoor locations and occupancy counts, and developed query processing techniques to enable interactive analytics for downstream location-based applications.

- LOCATER and T-COVE, tools to enrich location and occupancy from WiFi logs, 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**.

2024-present **Effective and Efficient Document Analytics**.

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.

- 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.