Stephen Macke[†]

Databricks, Inc.
Bellevue, WA

Email: stephen.macke@gmail.com
Website: https://smacke.net

Interests

- Tools for scalable and productive data science, analytics, and machine learning.

Experience

- **Software Engineer**, Databricks, Inc. · 2023—
- **Research Scientist**, Meta Platforms, Inc. · 2020—2023
- **Graduate Researcher**, University of Illinois at Urbana-Champaign · 2015—2020
- **Visting Student Researcher**, University of California, Berkeley · 2019—2020
- **Software Engineering Intern**, Facebook (now Meta) · Summer 2019
- **Software Engineering Intern**, Google Brain · Summer 2018
- **Software Engineering Intern**, Alation · Summer 2016
- **Software Engineering Intern**, Palantir Technologies · Summer 2014
- **Software Engineering Intern**, Palantir Technologies · Summer 2013
- **NSF REU Researcher**, University of Illinois at Urbana-Champaign · Summer 2012

Education

- **University of Illinois at Urbana-Champaign**, Urbana, IL
  Ph.D. in Computer Science · 2015—2021
  · Advisor: Aditya Parameswaran
  · Thesis: Leveraging Distributional Context for Safe and Interactive Data Science at Scale
- **Stanford University**, Stanford, CA
  M.S. in Computer Science (AI concentration) · 2013—2015
- **University of Tulsa**, Tulsa, OK
  B.S. in Computer Science, Applied Math Major · 2009—2013
  · Summa Cum Laude, University Honors
  · Departmental Honors in Computer Science, Mathematics

Honors and Awards

2023 **Winner**, Databricks Q3 2023 Hackathon "Simple Yet Powerful" category, for my project *Notebook Dataflow*
2021 **1st Prize**, CIDR Gong Show, for my talk titled *Automating State Management in Computational Notebooks*
2019 **Awardee**, State Farm Companies Foundation Doctoral Scholarship
2019 **Honorable Mention**, HackIllinois open source hackathon
2015 **Awardee**, Andrew and Shana Laursen Fellowship
2015 **Awardee**, Diffenbaugh Graduate Fellowship
2014 **Awardee**, NSF Graduate Research Fellowship[*]
2014 **Awardee**, Sourcegraph Open Source Fellowship
2013 **1st Place**, DWR Governor's Cup Business Plan Competition, Oklahoma division ($22,000 prize)
2013 **Top-150 Score**, Putnam Competition (exact rank: 142.5, highest in Oklahoma/Arkansas)[*]
2012 **Finalist**, SignalFire University Hacker Olympics in San Francisco
2012 **Awardee**, Goldwater Scholarship[*]
2012 **World Finalist**, ACM ICPC World Finals in Warsaw, Poland
2009 **Awardee**, University of Tulsa Presidential Scholarship (covering all tuition and living expenses)

[†] he/him/his
[*] Nationally competitive

Publications

◇ **Refereed Full-length Papers**

6. S. Shankar*, **S. Macke***, S. Chasins, A. Head, A. Parameswaran
   Bolt-on, Compact, and Rapid Program Slicing for Notebooks
   49th International Conference on Very Large Data Bases (VLDB), Vancouver, CA, Aug. 2023.
5. **S. Macke**, H. Gong, D. Lee, A. Head, D. Xin, A. Parameswaran
   Fine-Grained Lineage for Safer Notebook Interactions
   47th International Conference on Very Large Data Bases (VLDB), Copenhagen, Denmark, Aug. 2021.
4. **S. Macke**, M. Aliakbarpour, I. Diakonikolas, A. Parameswaran, R. Rubinfeld
   Rapid Approximate Aggregation with Distribution-Sensitive Interval Guarantees
   37th IEEE International Conference on Data Engineering, April 2021.
3. D. Petersohn, **S. Macke**, D. Xin, W. Ma, D. Lee, X. Mo, J. Gonzalez, A. Joseph, J. Hellerstein, A. Parameswaran
   Towards Scalable Dataframe Systems
   46th International Conference on Very Large Data Bases (VLDB), Tokyo, Japan, Sept. 2020.
2. D. Xin, **S. Macke**, L. Ma, J. Liu, S. Song, A. Parameswaran
   Helix: Holistic Optimization for Accelerating Iterative Machine Learning
   45th International Conference on Very Large Data Bases (VLDB), Los Angeles, USA, Aug. 2019.
1. **S. Macke**, Y. Zhang, S. Huang, A. Parameswaran
   Adaptive Sampling for Rapidly Matching Histograms
   44th International Conference on Very Large Data Bases (VLDB), Rio de Janeiro, Brazil, Aug. 2018.

◇ **Refereed Full-length Journal Papers**

1. D. Lee*, **S. Macke***, D. Xin*, A. Lee, S. Huang, A. Parameswaran
   A Human-in-the-loop Perspective on AutoML: Milestones and the Road Ahead
   IEEE Data Engineering Bulletin, June 2019.

◇ **Refereed Short, Demo, and Vision Papers + Extended Abstracts**

5. **S. Macke**
   Automating State Management in Computational Notebooks (Extended Abstract)
   Conference on Innovative Data Systems Research (CIDR), Jan. 2021.
4. **S. Macke**, A. Beutel, T. Kraska, M. Sathiamoorthy, D. Cheng, E. Chi
   Lifting the Curse of Multidimensional Data with Learned Existence Indexes
   ML for Systems Workshop at NeurIPS, Montreal, CA, Dec. 2018.
3. D. Xin, L. Ma, J. Liu, **S. Macke**, S. Song, A. Parameswaran
   Helix: Accelerating Human-in-the-loop Machine Learning (Demo)
   44th International Conference on Very Large Data Bases (VLDB), Rio de Janeiro, Brazil, Aug. 2018.
2. D. Xin, L. Ma, J. Liu, **S. Macke**, S. Song, A. Parameswaran
   Accelerating Human-in-the-loop Machine Learning: Challenges and Opportunities
   DEEM Workshop at SIGMOD Int'l Conf. on Management of Data, Houston, USA, June 2018.
1. A. El-Kishky, F. Xu, A. Zhang, **S. Macke**, J. Han
   Entropy-Based Subword Mining for Word Embeddings
   SCLeM Workshop at 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), New Orleans, USA, June 2018.

◇ **Preprints**

1. T. Wattanawaroon, **S. Macke**, A. Parameswaran
   Towards a Theory of Data-Diff: Optimal Synthesis of Succinct Data Modification Scripts
   ArXiv preprint.

*Equal contribution

Teaching

◇ **Teaching or Course Assistant for the following courses:**
· UIUC CS589-CCC (Cloud Computing Capstone) · Spring 2019
· Stanford CS149 (Parallel Computing) · Winter 2013 and Winter 2014
· Stanford CS103 (Mathematical Foundations of Computing) · Fall 2013 and Fall 2014
· Stanford CS101 (Introduction to Computing Principles) · Spring 2014

Software

◇ **Reactive Python Kernel for Jupyter Notebooks (ipyflow)**
· Drop-in replacement for Jupyter's Python 3 kernel that tracks data dependencies across variables, thereby providing reactive execution and other usability improvements.

◇ **Automatic Subtitle Synchronizer (ffsubsync)**
· Project synchronizes subtitles to video using voice audio detection to detect and fix constant drift.
· **Honorable Mention** at HackIllinois 2019 (among the top 5 submissions that did not win company awards).
· 5000 GitHub stars and thousands of monthly downloads, as of February 2022.

◇ **Python Framework for Declarative Instrumentation (pyccolo)**
· Pyccolo is a framework that abstracts away details such as AST transformations and system tracing in Python in order to monitor, alter, and augment running Python code.
· Supports instrumentation use cases such as: code coverage, syntactic macros, augmented Python syntax, dynamic dataflow analysis, and more.

◇ **Java Direct I/O (jaydio)**
· Project provides direct I/O functionality for Java (bypassing FS cache)
· Development funded by a Sourcegraph Open Source Fellowship ($1000 stipend)

Service

· Databricks intern mentor, summer 2024
· External reviewer for TKDD, SDM, SIGMOD Record, EACL, SIGMOD Student Research Competition
· Organizer for UC Berkeley's Spring 2020 Database Seminar
· Organized reading group for my lab to discuss recent DB research papers
· CS grad ambassador (host for students admitted to UIUC's CS PhD program) in 2018
· Served as mentor to undergraduate and master's students