Zheguang Samuel Zhao

Brown University
Department of Computer Science
115 Waterman St
Providence, RI 02912
United States of America

Email: zheguang.zhao@gmail.com
Homepage: zheguang.github.io
LinkedIn: www.linkedin.com/in/samuelzhao
Github: github.com/zheguang
Google Scholar: goo.gl/DR8pSa

Education

Ph.D. Candidate in Computer Science, Brown University, expected 2019.
Advisor: Prof. Stan Zdonik, Prof. Seny Kamara

M.S. in Computer Science, Brown University, 2016.
Advisor: Prof. Stan Zdonik

B.S. in Computer Science, University of Wisconsin at Madison, 2012.
Advisor: Prof. Jignesh Patel

Experiences

Sifr Systems, RI, Database Scientist, 2018 – present.

Develop provably-secure end-to-end encrypted big data systems including PostgreSQL and Apache Spark

Brown University, RI

Research Assistant, 2014 – present.

Teaching Assistant, 2015, 2018

Microsoft AI & Research, WA, Research Intern, 2017.

Research on constraint learning for automatic puzzle solving AI

Intel Labs, CA, Research Intern, 2015.

Research on the efficiency of machine learning algorithms in Apache Spark

Research on in-memory transactional database VoltDB using non-volatile memory

Hadapt (Acquired by Teradata), MA, Software Engineer, 2013 – 2014.

Develop the enterprise SQL-on-Hadoop system including query execution, storage engine, high availability and analytics toolkit. Use Agile methodology and continuous integration.

Kosmix (Acquired by @WalmartLabs), CA, Software Engineer Intern, 2012.

Develop an in-memory distributed queue system for the in-house distributed stream processing system in support for data analytics and machine learning

Great Lakes Bioenergy Research Center, WI, Software Engineer Intern, 2010 – 2012.

Develop biological data management system using .NET and Oracle database

Honors

Eta Kappa Nu

Upsilon Pi Epsilon

Golden Key International Honour Society

Open-source Projects

Searchable encryption for mobile messaging in Signal
https://github.com/encryptedsystems/Searchable-Signal-Android

Macau: statistical hypothesis testing based on resampling
https://github.com/zheguang/macau

Machine learning algorithms in Spark
https://github.com/zheguang/spark-study/tree/master/study/src/main/scala/edu/brown/cs/sparkstudy

Consistency control for machine learning algorithms
https://github.com/zheguang/babel

R-tree in Rust
https://github.com/zheguang/rtree

Spark performance analysis tool
https://github.com/zheguang/spark-perftool

VoltDB on non-volatile memory
https://github.com/zheguang/voltdb

Research

I am interested in the theories and designs of big data systems that are intelligent and safe. My research spans a broad area covering cryptography, data science/machine learning, and big data systems.
In this spirit I have dabbled in:

Constraint learning for puzzle-solving AI

> Can kids teach an AI system to solve their favorite games like Rubik's cube or Sudoku, or even a new game they invent? This project explores the architecture for such a general-purpose AI puzzle-solving system, from natural language interface, programming by demonstration, knowledge representation, and finally learning the optimal winning strategy of the game.

False discovery control in data science

> Recommendation engines and human analysts are not very capable of distinguishing true relationships from random noises that are inherent in the data. What is the way to automatically detect and control the risk of false insights? We build the first system to systematically guide the data exploration process away from noises.

Approximate data structures for visualization

How do we visualize a distribution on a cloud-scale dataset within interactive time? This project investigates how to augment the B-tree index with information to help approximate the underlying data distribution and refine the answer progressively.

Data system design on hybrid memory

This project studies the design of data systems if the memory hierarchy consists of both a volatile and fast component, and a non-volatile but slightly slower component.

Consistency control for stochastic machine learning algorithms

Many machine learning algorithms are shown to converge faster when models are updated stochastically. Stochasticity leads to parallelism. This leads to a fundamental question about to what degree of stochasticity can trade off model consistency for shorter training time. This project studies the consistency levels for machine learning algorithms.

Searchable encryption on mobile text messaging

We bring the world's first encrypted search to the secure mobile messaging app, Signal, which is widely used by government agencies, journalists, activists and people who are concerned about security and privacy.

Articles

Behavior of Large Random Graph.
Z. Zhao, supervised by Prof. Paul Dupius,
Randomized Algorithms for Counting, Integration and Optimization, Brown University, April 2017.

Investigating the Effect of the Multiple Comparisons Problem in Visual Analysis.
E. Zgraggen, Z. Zhao, R. Zeleznik, and T. Kraska,
CHI, April 2018.

Signal Search.
J. Engelman, S. Kamara, T. Moataz and S. Zhao,
Software release: `http://github.com/encryptedsystems/Searchable-Signal-Android`.
Press release: `http://esl.cs.brown.edu/blog/signal`, April 2017.

Controlling False Discoveries During Interactive Data Exploration.
Z. Zhao, L. De Stefani, E. Zgraggen, C. Binnig, E. Upfal and T. Kraska,
SIGMOD, May 2017.

Safe Visual Data Exploration.
Z. Zhao, E. Zgraggen, L. De Stefani, C. Binnig, E. Upfal and T. Kraska,
SIGMOD Demo, May 2017.

Bridging the Gap between HPC and Big Data frameworks.
M. Anderson, S. Smith, N. Sundaram, M. Capota, Z. Zhao, S. Dulloor, N. Satish and T. Willke,
VLDB, 2017.

Towards Sustainable Insights.
C. Binnig, L. De Stefani, T. Kraska, E. Upfal, E. Zgraggen and Z. Zhao,
CIDR, January 2017.

Towards a Benchmark for Interactive Data Exploration.
P. Eichmann, E. Zgraggen, Z. Zhao, C. Binnig, T. Kraska.
IEEE Data Engineering Bulletin, 2016.

Larger-than-memory Data Management on Modern Storage Hardware for In-memory OLTP Database Systems.
L. Ma, J. Arulraj, S. Zhao, A. Pavlo, S. Dulloor, M. Giardino, J. Parkhurst, J. Gardner, K. Doshi and S. Zdonik,
SIGMOD DaMoN, June 2016.

VisTrees: Fast Indexes for Interactive Data Exploration.
M. El-Hindi, Z. Zhao, C. Binnig and T. Kraska,
SIGMOD HILDA, June 2016.

Data Tiering in Heterogeneous Memory Systems.
S. Dulloor, A. Roy, Z. Zhao, N. Sundaram, N. Satish, R. Sankaran, J. Jackson and K. Schwan,
EuroSys, April 2016.

Selected Coursework

Abstract Algebra, Prof. Rich Schwartz

Calculus, Prof. Donald Passman, Gheorghe Craciun

Randomized Algorithms for Counting, Integration and Optimization, Prof. Paul G. Dupuis

Cryptography, Prof. Seny Kamara, Joseph Silverman

Probability, Prof. Erik Sudderth, Samuel S. Watson

Computational Linguistics, Prof. Eugene Charniak

Computer Architecture, Prof. Sherief Reda, Mark D. Hill

Distributed Computing through Combinatorial Topology, Prof. Maurice Herlihy

Database Management, Prof. Stan Zdonik, Jignesh Patel, Chris Ré

Microprocessor Synchronization, Prof. Maurice Herlihy

Algorithms and Data Structures, Prof. Eric Vigoda, Ben Liblit

Operating Systems, Prof. Michael Swift

Computer Networks, Prof. Aditya Akella

Physics, Prof. Peter Timbie, Daniel Chung, Ellen Zweibel

Reference

Prof. Stanley Zdonik, Professor at Brown University, *sbz@cs.brown.edu*

Prof. Seny Kamara, Professor at Brown University, *seny@cs.brown.edu*

Dr. Emanuel Zgraggen, Postdoctoral associate at MIT, *emanuel.zgraggen@gmail.com*

Dr. Subramanya Dulloor, Intel Labs, *dulloor@gmail.com*

Dr. Wang Lam, WalmartLabs, *wlam@cs.stanford.edu*