Zhongjun (Mark) Jin

CONTACT INFORMATION

4945 Bob and Betty Beyster Building
2260 Hayward Street
Ann Arbor, MI 48109, USA

Phone: (765) 421-5014
E-mail: markjin@umich.edu, markjin1990@gmail.com
Website: `https://markjin1990.github.io/`

OBJECTIVES

Applying for an industrial research scientist / applied scientist / research engineer / software engineer position starting in summer 2020.

RESEARCH INTERESTS

Build interactive data (like cleaning, integration, etc.) systems for data scientists/analysts, programmers, and non-experts using a combination of AI, HCI and PL techniques.

EDUCATION

University of Michigan, Ann Arbor, MI, USA **Aug. 2014 - May 2020**

Ph.D. Candidate, Computer Science and Engineering

- Advisor: Prof. Michael Cafarella and Prof. H. V. Jagadish

Purdue University, West Lafayette, IN, USA **Aug. 2011 - May 2014**

B.S. in Computer Science, Mathematics

Tianjin University, Tianjin, China **Aug. 2009 - Jul. 2011**

Electrical and Electronics Engineering

SELECTED PROFESSIONAL EXPERIENCE

Microsoft Research, Redmond, WA **Feb 2019 - May 2019**

Research Intern (Mentored by Yeye He)

Designed and implemented a system (like the seminal **FLASHFILL** system) recommending string data transformation programs which standardize or normalize regular-expression-like data patterns for a given set of string data (like phone numbers, or dates) with heterogeneous data patterns. Unlike most existing systems, which synthesize programs in real time, the recommended programs in **CLX** are learned offline from a large corpus of web data and of higher quality.

Trifacta, San Francisco, CA **May 2017 - Sep. 2017**

Software Engineering Intern (Mentored by Sean Kandel, Michael Minar, and Prof. Joe Hellerstein)

Designed and implemented **CLX**, an interactive data cleaning system. **CLX** 1) automatically identifies regular-expression-like data patterns for a given set of string data with heterogeneous data patterns for non-expert users to understand, and 2) suggests pattern-based transformation programs to unify various data patterns. The work was integrated to Trifacta Cloud Wrangler as a main feature in Aug 2018 and available at `https://cloud.trifacta.com/`.

Qualcomm, San Diego, CA **May 2013 - Aug. 2013**

Software Engineering Intern

SELECTED RESEARCH PROJECTS

FOOFAH - Programming-By-Example System for Synthesizing Data Transformation Programs.

FOOFAH performs data transformation/cleaning through programming by examples (PBE), which requires little domain knowledge from non-expert users. It efficiently discovers a sequence of data wrangling actions which guarantee to transform the raw data into the example form provided by the end user using a greedy search algorithm guided by the proposed distance metric customized for spreadsheets. The user interaction time is reduced by $\sim 60\%$ compared to the seminal **WRANGLER** system. The system is open-sourced at `https://github.com/umich-dbgroup/foofah/`.

MithraCoverage - System for Investigating Population Bias for Intersectional Fairness.

The system efficiently discovers under-represented/under-covered intersectional subgroups in a given dataset (e.g., a medical dataset may lack data records from a subgroup of "Hispanic women"), which may cause the problem of population bias. **MithraCoverage** also suggests a ranked list of subgroups in which the user could collect more data entries to remedy the above issue and ensure data fairness.

Prism - Example-based SQL Query Synthesizer.

The system infers SQL queries using imprecise and/or incomplete user examples from the target table the user desires from a relational database. The query discovery uses a bottom-up search-based algorithm and a filter-based validation process driven by a Bayesian network which reduces the overall number of query executions on the source database by $\sim 70\%$.

DeepWrangler - Data Transformation Program Synthesizer Guided by a Neural Network.

Many existing Programming by Example data transformation systems require many examples from end users to find the transformation programs they actually desire. **DeepWrangler** leverages a neural network to guide the program synthesis algorithm to reduce the number of examples needed and therefore saves the user interaction time.

Selected Publications

1. Christopher Baik, **Zhongjun Jin**, Michael Cafarella, and H. V. Jagadish, "Constructing Expressive Relational Queries with Dual-Specification Synthesis", in *CIDR* 2020.
2. **Zhongjun Jin**, Michael Cafarella, H. V. Jagadish, Sean Kandel, Michael Minar, and Joseph M. Hellerstein, "CLX: Towards verifiable PBE data transformation", in *EDBT* 2019.
3. **Zhongjun Jin**, Christopher Baik, Michael Cafarella, H. V. Jagadish, and Yuze Lou, "Demonstration of a Schema Mapping System Using Multiresolution Constraints", in *CIDR* 2019.
4. Abolfazl Asudeh, **Zhongjun Jin**, and H. V. Jagadish, "Assessing and Remedying Coverage for a Given Dataset", in *ICDE* 2019.
5. **Zhongjun Jin**, Michael R Anderson, Michael Cafarella, and H. V. Jagadish, "Foofah: Data Transformation By Example", in *SIGMOD* 2017.

Honors and Awards

- 1st Prize in "Systems, Software Engineering and Computer Science" session in *Michigan Engineering Graduate Symposium 2017 (EGS 2017) Graduate Research Contest*, 2017.
- Selected as "Best of Demos" at SIGMOD 2017.
- Sigmod Travel Award, 2017.
- University of Michigan Departmental PhD Fellowship, 2014.
- Outstanding Undergraduate Research Endeavor Award, Purdue Computer Science Dept, 2014
- Purdue Computer Science Neel Memorial Scholarship, 2013
- Purdue Computer Science Departmental Scholarship, 2012

Invited Talks

- "Intelligent Self-service Data Preparation: Problems and Solutions", 11/15/2018, Llamasoft Inc., USA.

Service

- External Reviewer: SoCC'19