Yongjoo Park

Computer Science and Engineering
University of Michigan
2260 Hayward St. 4957 Beyster
Ann Arbor, MI 48109-2121

Website: **https://yongjoopark.com**
Email: pyongjoo@umich.edu
Voice: +1 (734) 707-9206

RESEARCH INTEREST

Database Systems, Big Data, Data Analytics, Machine Learning

My research interest lies in building **systems for interactive-speed data analytics and machine learning (ML)**, with a special focus on exploiting *quality-performance tradeoffs* for substantial gains in performance. I combine rigorous statistical theories and large-scale data-intensive systems to build *fast, quality-guaranteed* data analytics and ML systems.

EDUCATION & TRAINING

Research Fellow — September 2017–Present
University of Michigan, Ann Arbor
- *Principal Investigator:* Barzan Mozafari

Ph.D., Computer Science and Engineering — August 2017
University of Michigan, Ann Arbor
- *Thesis Title*: Fast Data Analytics by Learning
- *Advisors:* Michael Cafarella and Barzan Mozafari

M.S., Computer Science — June 2013
University of Michigan, Ann Arbor

B.S., Electrical Engineering — February 2009
Seoul National University (SNU), South Korea

RESEARCH

My research has pursued the idea of *quality-performance tradeoffs* in several directions.

Quality-Performance Tradeoffs for Autonomous Systems

- Database Learning: DBL is the first approximate query processing system that can produce increasingly more accurate answers as it processes more queries. DBL relies on a non-parametric probabilistic model of the underlying data, constructed using the answers to past queries. For a new query, DBL uses this probabilistic model to produce a more accurate answer, instead of reading larger volumes of raw data.
- Selectivity Learning: QuickSel is a selectivity estimation algorithm that becomes more accurate as it processes more queries. By newly introducing *uniform mixture models* for selectivity estimation, QuickSel achieves significant improvements in speed and accuracy.

Quality-Performance Tradeoffs for Exploratory Analytics

- SQL Analytics: VerdictDB is the first approximate query processing system that *can run on top of any SQL engines*. VerdictDB addresses the challenge of *platform-independent accuracy-guarantees* by proposing a highly-efficient, SQL-based error-estimation technique.
- Visualization: In visualizations, including a data point in a plot matters only insomuch as it makes a visual difference. VAS uses this observation to produce an extremely small sample of the original dataset but yet enables high-fidelity visualizations.
- Image Search: NSH is a hashcode-based k-nearest neighbor (kNN) algorithm that relies on a counter-intuitive idea. Unlike the previous work, NSH shows that by placing similar items far apart than usual in the hashcode space, we can achieve higher accuracy.

Quality-Performance Tradeoffs for Machine Learning

- BlinkML is a *fast* ML system with *probabilistic quality guarantees*. BlinkML automatically and efficiently determines the minimum sample size such that the model trained on that sample produces the identical predictions as the full model (trained on the entire data) with high probability.

IMPACT

ACADEMIC IMPACT

I have first-authored five research papers in premier database conferences (e.g., SIGMOD, VLDB). Part of this work was awarded **2018 ACM SIGMOD Jim Gray Dissertation Award runner-up.** Most of my systems and algorithms are open-sourced.

INDUSTRY IMPACT

VerdictDB has been **adopted by several companies: Walmart, Digital2Go, a Canadian media company.** In particular, Walmart and Digital2Go shared with us some details:

- Walmart analyzes a massive amount of its sales data for email marketing; VerdictDB speeds up their ad-hoc analytic queries.
- Digital2Go uses the check-in data for its mobile advertising. VerdictDB offers the core ability to quickly estimate the number of *impressions* that satisfy arbitrary selection predicates.

VerdictDB is also tested by Alibaba and Tableau.

AWARDS

- **2018 ACM SIGMOD Jim Gray Dissertation Award runner-up** — June 2018
- ACM SIGMOD Student Travel Award, USD 900 — May 2017
- Rackham Travel Grant, USD 800 — Jan 2017
- **Graduate Study Fellowship** (for Ph.D.), USD 100,000 — 2013
  Kwanjeong Education Foundation
  The biggest scholarship foundation in Korea
- **Graduate Study Fellowship** (for Masters), USD 55,000 — 2011
  Jeongsong Cultural Foundation
  One of *only eight* recipients in 2011
- **Korean National Science Scholarship**, USD 20,000 — 2004

PUBLICATION

Referred Conference Papers

1. **Yongjoo Park**, Jingyi Qing, Xiaoyang Shen, Barzan Mozafari
   BlinkML: Efficient Maximum Likelihood Estimation with Probabilistic Guarantees
   SIGMOD'19 (research): ACM SIGMOD/PODS International Conference on Management of Data, Amsterdam, The Netherlands, 2019.
2. **Yongjoo Park**, Barzan Mozafari, Joseph Sorenson, Junhao Wang
   VerdictDB: Universalizing Approximate Query Processing
   SIGMOD'18 (research): ACM SIGMOD/PODS International Conference on Management of Data, Houston, TX, USA, 2018.
3. Wen He, **Yongjoo Park**, Idris Hanafi, Jacob Yatvitskiy, Barzan Mozafari
   Demonstration of VerdictDB, the Platform-Independent AQP System
   SIGMOD'18 (demo): ACM SIGMOD/PODS International Conference on Management of Data, Houston, TX, USA, 2018.

4. **Yongjoo Park**, Amhad Shahab Tajik, Michael Cafarella, Barzan Mozafari
   Database Learning: Toward a Database System that Becomes Smarter Over Time
   SIGMOD'17 (research): ACM SIGMOD/PODS International Conference on Management of Data, Chicago, IL, USA, 2017.
   SIGMOD Travel Award

5. **Yongjoo Park**
   Active Database Learning
   CIDR'17 (abstract): The biennial Conference on Innovative Data Systems Research, Chaminade, CA, USA, 2017.

6. **Yongjoo Park**, Michael Cafarella, Barzan Mozafari
   Visualization-Aware Sampling for Very Large Databases
   ICDE'16 (research): IEEE 32nd International Conference on Data Engineering, Helsinki, Finland, 2016.

7. **Yongjoo Park**, Michael Cafarella, Barzan Mozafari
   Neighbor-Sensitive Hashing
   VLDB'16 (research): 42nd International Conference on Very Large Data Bases, New Delhi, India, 2016.

8. Michael Anderson, Dolan Antenucci, Victor Bittorf, Matthew Burgess, Michael Cafarella, Arun Kumar, Feng Niu, **Yongjoo Park**, Christopher Ré, Ce Zhang
   Brainwash: A Data System for Feature Engineering
   CIDR'13 (vision): The biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, 2013.

In Submission

9. **Yongjoo Park**, Shucheng Zhang, Barzan Mozafari
   QuickSel: Quick Selectivity Learning with Mixture Models
   In Submission (research)

Workshop Presentations (without Proceedings)

10. Yongjoo Park, Amhad Shahab Tajik, Michael Cafarella, Barzan Mozafari
    Building Databases that Become Smarter over Time
    MBDOC'16: Midwest Big Data Opportunities and Challenges Workshop, Chicago, IL, USA, 2016

11. Yongjoo Park, Amhad Shahab Tajik, Michael Cafarella, Barzan Mozafari
    Database Learning: Toward a Database System that Becomes Smarter Over Time
    NEDB'16: North East Database Day 2016 (oral), Boston, MA, USA, 2016.

12. Yongjoo Park, Michael Cafarella, Barzan Mozafari
    Neighbor-Sensitive Hashing
    VSM'16: 3rd Workshop on Web-scale Vision and Social Media at ICCV (Extended Abstract), Santiago, Chile, 2015.

Thesis

13. Yongjoo Park
    Fast Data Analytics by Learning
    Ph.D. Dissertation
    Awarded 2018 ACM SIGMOD Jim Gray Dissertation Award runner-up

Non-Referred Technical Reports (full versions to published papers)

14. Yongjoo Park, Barzan Mozafari, Joseph Sorenson, Junhao Wang
    VerdictDB: Universalizing Approximate Query Processing
15. Yongjoo Park, Amhad Shahab Tajik, Michael Cafarella, Barzan Mozafari
    Database Learning: Toward a Database System that Becomes Smarter Over Time
16. Yongjoo Park, Michael Cafarella, Barzan Mozafari
    Neighbor-Sensitive Hashing
17. Yongjoo Park, Michael Cafarella, Barzan Mozafari
    Visualization-Aware Sampling for Very Large Databases

TALKS

BlinkML

1. AVL (www.avl.com), Ann Arbor, April 2018

VerdictDB

2. Oracle BI Group, Redwood City, December 2017
3. ACAIA workshop, San Jose, November 2017
4. Oracle Database Group, Redwood City, November 2017
5. Cloudera Impala Team, Palo Alto, November 2017
6. Big Data Innovation Summit, Boston, Septempber 2017
7. New Tech Meetup, Ann Arbor, July 2017
8. SIGMOD, Chicago, May 2017
9. University of Michigan Software Group, Ann Arbor, May 2017

Database Learning

10. Brown Database Group, Providence, March 2017
11. Stanford InfoLab, Palo Alto, February 2017
12. CIDR, Chaminade, California, January 2017
13. MBDOC, Chicago, September 2016
14. NEDB, Boston, January 2016

Visualization-Aware Sampling

15. ICDE, Helsinki, Finland, May 2016
16. AVL (www.avl.com), Ann Arbor, April 2016

Neighbor-Sensitive Hashing

17. VLDB, New Delhi, India, September 2016
18. VSM@ICCV, Santiago, Chile, December 2015

TEACHING

Guest Lecturer, Advanced Database Management Systems (EECS 584)
University of Michigan, Fall 2018
- Led the sections for streaming database systems

Guest Lecturer, Database Management Systems (EECS 484)
University of Michigan, Winter 2018
- On approximate database systems

Graduate Student Instructor, Web Databases and Information Systems (EECS 485)
University of Michigan, Winter 2012
- Designed programming assignments (interactive web using JavaScript, and PageRank computation of Wikipedia pages using Hadoop)
- Led weekly discussion sessions

MENTORING

Wen He (B.S.) Summer 2017–Winter 2018

Wen He and I implemented **VerdictDB**'s driver for Apache Spark SQL and tested the driver on top of various platforms, such as Cloudera cluster, MapR cluster, and Google's Dataproc. Wen authored **our demo paper** published in SIGMOD 2018. Wen He joined the master's program at Carnegie Mellon University.

Shucheng Zhong (B.S.) 2018

Shucheng Zhong and I implemented **VerdictDB**'s query parsing and planning logic. He also contributed, **QuickSel**, the selectivity estimation algorithm that becomes more accurate over time. He is the co-author on the corresponding research paper. He applies to graduate programs this year.

Jingyi Qing (B.S.) Fall 2017–Winter 2018

I and Jingyi Qing developed **BlinkML**, a system that can quickly train ML models with probabilistic accuracy guarantees. He is the co-author on the corresponding research paper. Jingyi Qing joined Amazon.com, Seattle.

Xiaoyang Shen (B.S.) Fall 2017–Winter 2018

I worked with Xiaoyang Shen on developing and testing **BlinkML**, which appears in SIGMOD 2019. Xiaoyang Shen applies to graduate programs this year.

Junhao Wang (B.S.) Summer 2017–Fall 2018

I worked with Junhao Wang to implement **VerdictDB**'s driver for Amazon Redshift. He is a co-author of the corresponding research paper published in SIGMOD 2018. Junhao Wang joined the master's program at McGill University.

EMPLOYMENT

Graduate Student Research Assistant Fall 2012–Winter 2017
University of Michigan, Ann Arbor

Software Engineer Intern — Summer 2014
Amazon.com, Seattle
- Developed a data center capacity prediction system for all Amazon data centers

Graduate Student Instructor — Winter 2012
University of Michigan, Ann Arbor
- EECS 485 Web Databases and Information Systems
  (see TEACHING for details)

Software Engineer — Dec 2008–May 2011
Webcash, Seoul
- Developed an online banking system for J.P. Morgan, Hong Kong
- Developed financial iPhone applications

Research Assistant — June 2007–Jan 2008
Seoul National University, Seoul
- Developed a power-efficient vehicle entertainment system that runs on embedded-processors (ARM)

SERVICE

Core PC, SoCC 2019

Reviewer, TKDE 2018

Program Committee, aiDM workshop at SIGMOD 2018 (http://www.aidm-conf.org/)

Reviewer, SIGMOD 2018

Publicity Chair, ACAIA workshop 2017 (http://dbgroup.eecs.umich.edu/acaia/)

Reviewer, VLDBJ 2017

External reviewer
- SIGMOD 2019
- VLDB 2018
- CIDR 2017
- VLDBJ, VLDB, SIGMOD 2016
- VLDB, ICDE, CIDR 2015

Organizer of
- University of Michigan Database Group meetings 2016, 2014
- MIDAS (Michigan Data Science) seminars 2014

REFERENCES

Michael J. Cafarella
Associate Professor
The University of Michigan, Ann Arbor
http://web.eecs.umich.edu/~michjc/
michjc@umich.edu

Barzan Mozafari
Assistant Professor
The University of Michigan, Ann Arbor
http://web.eecs.umich.edu/~mozafari/
mozafari@umich.edu

Jeffrey F. Naughton
Emeritus Professor
The University of Wisconsin, Madison
The head of Google Madison
http://pages.cs.wisc.edu/~naughton/
naughton@google.com

Srikanth Kandula
Principal Researcher
Microsoft Research
https://www.microsoft.com/en-us/research/people/srikanth/
srikanth@microsoft.com