

KYOUNGMIN KIM

kyoung-min.kim@epfl.ch ldis-kaist.github.io Google Scholar

Last updated: Aug. 2026

SUMMARY

I am joining the School of Computing at KAIST as an Assistant Professor, leading the Lab of Data-Intensive Systems (LDIS). Before joining KAIST, I was a postdoc at the Data-Intensive Applications and Systems (DIAS) Lab @ EPFL, led by Professor Anastasia Ailamaki, and received my Ph.D. from the Data Systems Lab @ POSTECH under Professor Wook-Shin Han. My research interests are in data systems and AI, covering heterogeneous data (graphs, relations, vectors, KVs in LLMs), workload (streaming processing, query optimization, worst-case optimal algorithms, semantic operators using LLMs), and hardware (CPU, GPU, memory, disk, and emerging architectures such as CXL, compute-storage disaggregation, and distributed clouds). My recent focus lies in deeper integrations of AI/LLMs and DBs (learned query optimization, NL2SQL, ELT pipeline synthesis, and data management in LLM inference), vector-relational databases, and unified and declarative query optimization for clouds, considering the large potential in AI+DB and dominant data transfer costs between hardware devices in modern data systems.

EDUCATION

- Ph.D. Feb. 2017 - Aug. 2023 POSTECH, South Korea
Dept. of Convergence IT Engineering (Major: CSE)
(Dissertation: Boosting Cardinality Estimation for Graph/Relational Databases using Sampling, Synopsis, and Deep Learning)
(Supervisor: Professor Wook-Shin Han)
- B.S. Mar. 2013 - Feb. 2017 POSTECH, South Korea
Dept. of Creative IT Engineering (Magna Cum Laude)

PUBLICATIONS

1. **Kim, K.** and Ailamaki, A., “Confining Nondeterminism: AI-Driven Research Systems as DBMSs for Reliable, Non-Wasteful, Transparent, and Collaborative Research [Vision]” [preprint]
2. Göçer, A., Tsakalidou, I., Nicholson, H., **Kim, K.**, and Ailamaki, A., “QVCache: A Query-Aware Vector Cache” [preprint]
3. **Kim, K.**, Catheland, M., Balcioglu, E., and Ailamaki, A., “Fast LLM-Based Semantic Filtering: From a Unified Framework to an Adaptive Two-Phase Method” [preprint]
4. **Kim, K.**, Roth, L., Liang, L., and Ailamaki, A., “Fast Approximate Vector Joins via Offline-Online Co-Design” [preprint]
5. Das, P., Boissier, M., **Kim, K.**, Mehrotra, S., and Rabl, T., “SmartRabbit: An Interactive Query Processor” (ACM SIGMOD 2026) [pdf]
6. **Kim, K.**, Li, J., Hong, K., Liu, Q., Huang, D., and Ailamaki, A., “Saving GPU Hours in LLM Inference System Development and Online Workloads with Simulation and DBMS-Inspired Cache Replacement Policies” [preprint]
7. **Kim, K.** and Ailamaki, A., “Trustworthy and Efficient LLMs Meet Databases” [preprint]
8. Hong, K., **Kim, K.**, Lee, Y., Moon, Y., Bhowmick, S., and Han, W., “Themis: A GPU-accelerated Relational Query Execution Engine” (VLDB 2025) [pdf]
9. Lee, S., **Kim, K.**, and Han, W., “ASM in Action: Fast and Practical Learned Cardinality Estimation” (ACM SIGMOD Demo 2024) [pdf]
10. Lee, Y., **Kim, K.**, Lee, W., and Han, W., “In-depth Analysis of Continuous Subgraph Matching in a Common Delta Query Compilation Framework” (ACM SIGMOD 2024) [pdf]

11. **Kim, K.**, Lee, S., Kim, I., and Han, W., “ASM: Harmonizing Autoregressive Model, Sampling, and Multi-dimensional Statistics Merging for Cardinality Estimation” (ACM SIGMOD 2024) [pdf]
12. **Kim, K.**, Ha, J., Fletcher, G., and Han, W., “Guaranteeing the $O(\text{AGM}/\text{OUT})$ Runtime for Uniform Sampling and Size Estimation over Joins” (ACM PODS 2023, first PODS paper from Korea) [pdf]
13. **Kim, K.**, Jung, J., Seo, I., Han, W., Choi, K., and Chong, J., “Learned Cardinality Estimation: An In-depth Study” (ACM SIGMOD 2022) [pdf]
14. **Kim, K.**, Kim, H., Fletcher, G., and Han, W., “Combining Sampling and Synopses with Worst-Case Optimal Runtime and Quality Guarantees for Graph Pattern Cardinality Estimation” (ACM SIGMOD 2021) [pdf]
15. Park, Y., Ko, S., Bhowmick, S., **Kim, K.**, Hong, K., and Han, W., “G-CARE: A Framework for Performance Benchmarking of Cardinality Estimation Techniques for Subgraph Matching” (ACM SIGMOD 2020) [pdf]
16. **Kim, K.**, Seo, I., Han, W., Lee, J., Hong, S., Chafi, H., Shin, H., and Jeong, G., “TurboFlux: A Fast Continuous Subgraph Matching System for Streaming Graph Data” (ACM SIGMOD 2018) [pdf]

PROFESSIONAL EXPERIENCES

Assistant Professor	Aug. 2026 - Present	KAIST, South Korea School of Computing, Lab of Data-Intensive Systems (LDIS)
Post Doctoral Fellow	Mar. 2024 - Aug. 2026	EPFL, Switzerland Data-Intensive Applications and Systems Lab
Post Doctoral Fellow	Aug. 2023 - Feb. 2024	POSTECH, South Korea Data Systems Lab

PROFESSIONAL ACTIVITIES

PC member of SIGMOD	2026, 2027
PC member of VLDB	2024, 2026
PC member of ICDE	2025
PC member of VLDB Demo	2025, 2026
PC member of BTW Demo	2025
Consultant @ SAP Labs Korea	Mar. 2023 - Feb. 2024
Organization Co-chair of Korea-Japan Workshop on Management of Data	Oct. 2023

INVITED TALKS

Invited Seminar Talks	Dec. 2022, Oct. 2023, Oct. 2023, May. 2024, Dec. 2024, May. 2025, Dec. 2025, Feb. 2026
Seminar talks @ MIT, University of Washington, EPFL, TU/e, UNIST, Seoul National University, Sungkyunkwan University, and Microsoft about my research.	
Invited Talk for the Top Conference Session @ KSC	2018, 2020, 2021
Invited to present my SIGMOD papers @ Korea Software Congress (KSC)	

GRANT PROPOSALS WRITTEN DURING POSTDOC

Heterogeneous Resource Management for Efficient Systems	Oct. 2025
Swiss National Science Foundation (SNSF), 1 million CHF for 4 years, accepted in Oct. 2025. Contributed as a co-writer.	

Energy-efficient Large Foundation Models for Databases

Jul. 2025

Huawei Zurich, 0.2 million CHF for 1 year, accepted in Jul. 2025.
Raised the topic and led the proposal writing.

Principled Approach to Resource Optimization through Declarative Data Systems

European Research Council (ERC) Advanced Grant, 2.5 million EUR for 5 years, accepted in Jun. 2025.
Contributed as a co-writer.

EXAMPLES OF RESEARCH SUPERVISION

As a postdoc at EPFL, I supervised 40+ undergraduate and master students (including 7 master theses) on topics spanning (cloud-based) OLAP query optimization/execution, cardinality estimation, semantic operator optimization, NL2SQL, vector search, and RAG. Example projects below.

Multi-modal Few-Shot Adaptation for Databases

Elif Sema Balcıoğlu

Master Thesis, Co-supervised with Oracle, Nominated for Fritz-Kutter Award, Scored 6.0/6.0 Fall 2024

Scaling HNSW with Data Locality and Caching

André Loureiro Espírito Santo

Master Thesis, Co-supervised with Oracle, Scored 6.0/6.0

Spring 2025

Adaptive Graph Optimization for Hybrid RAG Systems

Hugues Louis Christophe

Master Thesis, Co-supervising with TELEMAQUE, Scored 6.0/6.0

Fall 2025

Efficient Data Collector for Legaltech Platforms

Jérôme Jean-Martin Raymond Ceccaldi

Master Thesis, Co-supervising with LegalPass SA, Scored 6.0/6.0

Fall 2025

Two-Stage Agentic Evidence Retrieval for Financial Question Answering over Company Filings

Hanwen Zhang

Master Thesis, Co-supervised with Sterwen SA, Scored 6.0/6.0

Spring 2026

Machine Learning Approximation for Asian Options Pricing

Mechanical Engineering Students

Bachelor Thesis, EPFL

Spring 2026

Efficient Disk-based Approximate Vector Join

Lennart Roth

Master's Semester Project, Exchange Student from KTH

Fall 2025

ELT Pipeline Synthesis using LLM Agents

Camille Emma Lannoye, Quentin Austin Sandoz

Master's Semester Project, EPFL

Fall 2025

Adapting Join Filtering and Partition Pruning Optimizations to FlexPushdownDB

Brandon Tang, Nikita Pozharskyi

Master's Semester Project, EPFL

Spring 2026

Evaluating True Cardinality Injection in DuckDB Query Optimization

Vincentas Antanas

Danys

Semester Project, EPFL

Spring 2026

NL2SQL using LLM Agents and Query Planning/Execution

Soyun Kim, Hyunjun Kim

Semester Project, Exchange Students from KAIST

Spring 2025, Fall 2025

Adaptive Translation: State-Based Incremental Reasoning and Multi-Path Search for NL2SQL

Reinatt Hansel Wijaya

Master's Semester Project, EPFL

Spring 2026

COVER-SQL: Contract-Oriented Verification and Evidence-Guided Resampling for NL2SQL

Ali El Azdi

Semester Project, EPFL

Spring 2026

Optimizing Semantic Filtering: Active Learning and a Bayes-Error View <i>Master's Semester Project, EPFL</i>	Martin Catheland <i>Spring 2026</i>
Effective KV Cache Management for Faster LLM Inference <i>Summer Intern from NUS, Joining our lab in 2026</i>	Jiacheng Li <i>Summer 2025</i>
Optimizing Semantic-Operator Pipelines <i>Summer Intern from University of Chicago</i>	Hanchen Li <i>Summer 2024</i>

SELECTED PROJECTS BEFORE PROFESSORSHIP

LLM for Databases: NL2SQL and Beyond	Mar. 2025 - Present
- Novel approaches to harness LLMs and databases for NL2SQL and ELT pipeline synthesis using LLM agents, adaptive translation, and new intermediate representations for queries. - NL2SQL project funded by Huawei Zurich for 1 year (led writing the proposal).	
Hybrid Vector-Relational Data System	Feb. 2025 - Present
- Novel approaches to make vectors as first-class citizens (vector search/join/aggregation) inside data systems. - Cardinality estimation and adaptive strategies to select between indexes and brute-force methods. 20x faster approximate vector join than the previous SOTA method, preparing for a submission. - New vector-relational benchmark, showing that existing vector-relational DBMSs suffer from slow vector processing (spending more than 95% of the execution time). - Developed into a joint funding proposal with ETH for 4 years (lead writing the proposal).	
Data Management and Simulation for LLM Inference Serving	Aug. 2024 - Present
- Modeling LLM inference costs in LLM inference serving systems and formulating optimal schedules of LLM inference requests as a constraint satisfaction problem (CSP) for the first time. - Applying Jim Gray's five-minute rule for LLM inference. - Counter-intuitive findings on inference scheduling, a new easy-to-implement policy that can reduce LLM inference costs by 1.2x where an estimated optimal reduction is 1.4x.	
Trustworthy and Efficient LLMs Meet Databases	Mar. 2024 - Nov. 2024
- A tutorial, comprehensive survey of studies in making LLMs more trustworthy and efficient. - Also discussing the intersection of LLMs and databases, including the mixed relational-LLM workloads, cost-based holistic optimization, and computation pushdown under storage-aggregation.	
Sampling, Synopsis, and Deep Learning for Cardinality Estimation	Dec. 2022 - Jun. 2023
- Built a novel and practical cardinality estimator on top of recent deep generative models, estimating thousands of sub-queries in seconds and producing near-optimal query plans. - Resolved the limitation of learned estimators that could support very simple queries only. - Also was the topic of my dissertation. - Published at SIGMOD 2024. Developed further ideas with even better results.	
Sampling-based Estimation with Hypertree Decomposition	Oct. 2021 - Mar. 2023
- A theoretical work that has a potential to outperform existing work in practice. - Proposed a novel method for arbitrary join size estimation with the best time complexity bound and further extended it using hypertree decomposition. - Published at PODS 2023 (the first PODS paper from Korea in its 42-year conference history).	
Learned Cardinality Estimation over Tabular Data	Jan. 2021 - Jun. 2022
- Compared and analyzed the learned cardinality estimators for SPJ queries on small to large databases. - Demystified when and why their errors occur, proposed better methods by fixing these sources of errors, and improved the plan quality of PostgreSQL by using the learned cardinalities.	

- Published at SIGMOD 2022.

Join Size Estimation for Graph Databases

Dec. 2019 - Jun. 2021

- Compared and analyzed the join size estimation methods over graph databases, including sampling/synopsis-based methods and methods for relational databases.
- Figured out that a table-at-a-time sampling-based method for relational databases, outperforms all others.
- Implemented an estimator that outperforms the above one, with an attribute-at-a-time sampling to reduce sampling variance and an optional synopsis that precomputes the sample spaces for the patterns that are inherently hard to reduce the variance.
- Published at SIGMOD 2020 and 2021.

GPU-based Subgraph Matching

Sep. 2019 - Dec. 2019

- 70-1000x faster implementation compared to a single-threaded CPU-based subgraph matching algorithm, and 50% improvements over a GPU-based baseline.
- Implemented optimized intersections using signatures, assigned one warp per partial embedding, balanced loads from estimated costs, and postponed the heavy Cartesian products of enumeration.

Worst-case Optimal Joins

Jun. 2018 - Jun. 2019

- Developed a worst-case optimal join algorithm to reduce the runtime bound, starting from a practical but unbounded algorithm.
- Built knowledge of theoretical works along with SIMD-based optimizations.

Continuous Subgraph Matching

Jan. 2016 - Jun. 2018

- Implemented an efficient continuous subgraph matching system over dynamic graphs that outperforms previous SOTA by several orders of magnitude.
- Proposed a data-centric representation of intermediate results for the first time and an edge transition rule that efficiently update these results.
- Published at SIGMOD 2018.