Seokki Lee

Website: researchdirectory.uc.edu/p/lee5sk
Email: lee5sk@ucmail.uc.edu

Address: 2851 Woodside Dr,
Rhodes 885,
Cincinnati,
OH 45221

Academic Appointments

University of Cincinnati — Cincinnati, US
Assistant Professor, Computer Science — 2020–present

Illinois Institute of Technology — Chicago, US
Research & Teaching Assistant, Database Group — 2014–2020

Education

Illinois Institute of Technology — Chicago, US
Ph.D. in Computer Science, Advisor: Dr. Boris Glavic — 2014–2020

Hanyang University — Seoul, KR
M.S. in Computer Science & Engineering — 2007–2009

Research Interests

Data Management, Data Provenance, Big Data, Explainability
My research area focuses on database systems, specifically data provenance over big data. I have two main research pillars: (i) Efficient provenance management for providing concise and meaningful explanations for complex queries and workflows over large amounts of data; (ii) Developing provenance models to enhance explainability across diverse domains, with a particular focus on applications in machine learning (ML), privacy-enhancing technologies (PETs), and data visualization.

Impact: My research aims to create a new systematic paradigm for explainable ML, PETs, and data visualization using rich provenance information. These areas currently lack systematic mechanisms to explain decisions/outcomes, and developing techniques that provide enriched explanations is extremely important. In the long term, my research will contribute to making the broad big data ecosystem more explainable and trustworthy.

Awards

- **Research Launch Awards** — 2021
  Office of Research, University of Cincinnati
- **Research Professional Development Award** — 2020
  Office of Research, University of Cincinnati
- **Ph.D. Dissertation Fellowship** — 2019–2020
  Illinois Institute of Technology
- **Student Travel Grant** — 2017
  IEEE International Conference on Data Engineering (ICDE)

List of Advisees

Current Ph.D. Students

- N. Akwari (CSE) — Explainable Machine Learning using Provenance — 2022–present
- S. Rawat (CSE) — Exploring Provenance for Explainable Information Gain — 2021–present

Current Master Student Theses

- A. Margi (MSCS) — Efficiently Measuring Information Gain using Provenance — 2023–present
- S. Chouhan (MSCS) — Efficiently Sampling Big Provenance — 2023–present

Current Undergraduate Student Projects

- J. Turnau (CS) — Explainable Machine Learning using Provenance — 2022–present
- B. Ju (CS) — Hybrid Explanations and Repairs — 2023–present

Supervised Master Theses

- S. Moshtaghi Largani (CSE) — Provenance Summaries for Big Data — 2021–2023
- B. Su (CMPE) — Hybrid Explanations — 2021–2022

Supervised Student Projects

- B. Tyagi (MEng) — Efficiently Measuring Information Gain using Provenance — 2023
- P. Amezcua (MEng/CCHMC) — Effect of Hemostatic Proteins on Eczematic Microbiota in Mice — 2021–2022
- C. Lu (EE) — Provenance Support for Aggregation — 2021–2022
- D. Ma (EE) — Provenance Support for Aggregation — 2021–2022
- A. Liu (EE) — Provenance Support for Aggregation — 2021–2022
- B. He (EE) — Provenance Support for Aggregation — 2021–2022
- R. Strohm (CS) — Developing A Simplified ERP System using Postgres — 2021–2022
- D. Hosford (CS) — Developing A Simplified ERP System using Postgres — 2021–2022
- D. Rajput (BANA) — Efficient Evaluation of Machine Learning Model using Provenance — 2021
- N. Quynh (BS.Chem.Eng) — Data Analysis using big data systems — 2021
- S. Jayaraj (MEng) — Integration and Analysis of User's interests — 2021
- P. Kathan Hitesh (MEng) — Analysis of Data for User trend — 2021

Publications

Peer-reviewed Conference Articles (Acceptance Rate: SIGMOD=17.4%, VLDB=18.6%, ICDE=19.1%)

[1] **S. Rawat**, **M. Amin**, and **S. Lee**, "Exploring provenance for explainable information gain", in *ICDE*, 2024 (under review).

[2] R. Diestelkämper, **S. Lee**, B. Glavic, and M. Herschel, "Debugging missing answers for spark queries over nested data with breadcrumb", *Proceedings of the VLDB Endowment*, pp. 2731–2734, 2021.

[3] R. Diestelkämper, **S. Lee**, M. Herschel, and B. Glavic, "To not miss the forest for the trees-a holistic approach for explaining missing answers over nested data", in *SIGMOD*, 2021, pp. 405–417.

[4] **S. Lee**, B. Ludäscher, and B. Glavic, "Approximate summaries for why and why-not provenance", *Proceedings of the VLDB Endowment*, 2020.

[5] **S. Lee**, B. Ludäscher, and B. Glavic, "Provenance summaries for answers and non-answers", *PVLDB*, pp. 1954–1957, 2018.

[6] **S. Lee**, S. Köhler, B. Ludäscher, and B. Glavic, "A sql-middleware unifying why and why-not provenance for first-order queries", in *ICDE*, 2017, pp. 485–496.

[7] X. Niu, B. S. Arab, **S. Lee**, S. Feng, X. Zou, D. Gawlick, V. Krishnaswamy, Z. H. Liu, and B. Glavic, "Debugging transactions and tracking their provenance with reenactment", *Proceedings of the VLDB Endowment*, 2017.

[8] W. Spoth, B. S. Arab, E. S. Chan, D. Gawlick, A. Ghoneimy, B. Glavic, B. Hammerschmidt, O. Kennedy, **S. Lee**, Z. H. Liu, *et al.*, "Adaptive schema databases", in *CIDR*, 2017.

Journal Articles

[9] **S. Lee**, B. Ludäscher, and B. Glavic, "Pug: A framework and practical implementation for why and why-not provenance", *VLDBJ*, pp. 47–71, 2019.

[10] B. S. Arab, S. Feng, B. Glavic, **S. Lee**, X. Niu, and Q. Zeng, "Gprom-a swiss army knife for your provenance needs", *A Quarterly bulletin of the Computer Society of the IEEE Technical Committee on Data Engineering*, 2018.

Peer-reviewed Workshops

[11] **S. Lee**, B. Glavic, A. Chapman, and B. Ludäscher, "Hybrid query and instance explanations and repairs", in *Companion Proceedings of the ACM Web Conference 2023*, 2023, pp. 1559–1562.

[12] **S. Moshtaghi Largani** and **S. Lee**, "Efficient sampling for big provenance", in *Companion Proceedings of the ACM Web Conference 2023*, 2023, pp. 1508–1511.

[13] **J. Turnau**, **N. Akwari**, **S. Lee**, and D. Rajput, "Provenance-based explanations for machine learning (ml) models", in *2023 IEEE 39th International Conference on Data Engineering Workshops (ICDEW)*, IEEE, 2023, pp. 40–43.

[14] **S. Rawat**, **S. Lee**, and T. Jung, "Measuring information gain using provenance", in *Proceedings of the 14th International Workshop on the Theory and Practice of Provenance*, 2022, pp. 1–4.

[15] T. Jung, **S. Lee**, and W. Tang, "Using provenance to evaluate risk and benefit of data sharing", in *13th International Workshop on Theory and Practice of Provenance (TaPP 2021)*, 2021.

[16] R. Diestelkämper, B. Glavic, M. Herschel, and **S. Lee**, "Query-based why-not explanations for nested data", in *TaPP*, 2019.

[17] **S. Lee**, X. Niu, B. Ludäscher, and B. Glavic, "Integrating approximate summarization with provenance capture", in *TaPP*, 2017.

[18] **S. Lee**, S. Köhler, B. Ludäscher, and B. Glavic, "Implementing unified why-and why-not provenance through games", in *IPAW*, 2016, pp. 209–213.

Technical Reports

[19] R. Diestelkaemper, **S. Lee**, M. Herschel, and B. Glavic, "To not miss the forest for the trees–a holistic approach for explaining missing answers over nested data (extended version)", *arXiv preprint arXiv:2103.07561*, 2021.

[20] **S. Lee**, B. Ludäscher, and B. Glavic, "Approximate summaries for why and why-not provenance (extended version)", *arXiv preprint arXiv:2002.00084*, 2020.

[21] **S. Lee**, B. Ludäscher, and B. Glavic, "Pug: A framework and practical implementation for why & why-not provenance (extended version)", *arXiv preprint arXiv:1808.05752*, 2018.

[22] **S. Lee**, S. Koehler, B. Ludäscher, and B. Glavic, "Efficiently computing provenance graphs for queries with negation", *arXiv preprint arXiv:1701.05699*, 2017.

[23] **S. Lee**, Y. Tang, B. Ludäscher, and B. Glavic, "An efficient implementation of game provenance in dbms", Tech. Rep., 2015.

[24] **S. Lee**, Z. Wang, B. Glavic, and R. J. Miller, "Automatic generation and ranking of explanations for mapping errors", 2014.

Extended Abstracts and Posters

[25] **S. Lee** and B. Ludäscher, "Sharing reproducible research through dataone and whole tale", *Whole Tale Workshop (poster)*, 2018.

[26] **S. Lee**, S. Köhler, B. Ludäscher, and B. Glavic, "A sql-middleware unifying why and why-not provenance for first-order queries", *GCASR (poster)*, 2017.

[27] **S. Lee** and B. Glavic, "Automatic generation and ranking of explanations for mapping errors", *GCASR (poster)*, 2015.

Proposal Experience

Federal Agencies

- **Democratized Visualization Assistants for Data-Rich Collaboration Infrastructure** — Pending
  NSF CSSI Elements / Co-PI / Total: $586,263 (evenly budgeted)
- **A Data Visualization Assistant for Enabling Scientific Conversations** — Pending
  NSF III Medium / Co-PI / Total: $1,199,999 (evenly budgeted)
- **Trustworthy Data Sharing with Pre-appraisal and Negotiation** — Pending
  NSF SaTC Small / PI / Total: 594,140 / Lee's portion: $327,116
- **Provenance-based Explanations for Machine Learning (ML) Model Predictions** — Pending
  DARPA Young Faculty Award (Executive Summary) / PI
- **Efficient and Comprehensive Data Provenance Management** — 2023
  NSF CAREER / PI / Total: $470,150 — (Prepare resubmission)
- **Trustworthy Data Sharing with Pre-appraisal and Negotiation** — 2022
  NSF SaTC Small / PI / Total: $599,997 / Lee's portion: 300,000
  (Declined) The feedback was highly positive and the modified version is currently pending.
- **Robust Privacy Guarantee by Detecting Differential Privacy Misuse using Provenance** — 2020
  DARPA Young Faculty Award / PI — (Declined)

Industry

- **Explaining Machine Learning Model Prediction using Provenance** — 2022
  Google Research Scholar Program / Total: $60,000 — (Prepare resubmission)
- **Provenance for Reproducibility and Replicability of Geospatial Research** — 2020
  Google Research Scholar Program / Total: $60,000 — (Prepare resubmission)
- **Provenance-based Explanations for Machine Learning (ML) Models** — 2022
  Meta Research Awards / Total: $50,000 — (Prepare resubmission)

Teaching Experience

- **Lecture** at University of Cincinnati — Fall 2023
  Database Theory (CS5151/6051)
- **Lecture** at University of Cincinnati — Fall 2022
  Intro to Computer Science (CS1100)
- **Lecture** at University of Cincinnati — Fall 2022
  Database Theory (CS5151/6051)
- **Lecture** at University of Cincinnati — Spring 2022
  Advanced Database Management (CS7071)
- **Lecture** at University of Cincinnati — Fall 2021
  Database Theory (CS6051)
- **Lecture** at University of Cincinnati — Spring 2021
  Database Theory (CS5151/6051)
- **Teaching Assistant** at Illinois Institute of Technology — Spring 2019
  Advanced Database Organization (CS525)
- **Teaching Assistant** at Illinois Institute of Technology — Fall 2017
  Advanced Database Organization (CS525)
- **Teaching Assistant** at Illinois Institute of Technology — Spring 2017
  Advanced Database Organization (CS525)
- **Teaching Assistant** at Illinois Institute of Technology — Spring 2016
  Data Integration, Warehouse, and Provenance (CS520)

Professional Service

Program Committee

- International Conference on Very Large Databases (**VLDB**) — 2022–present
- IEEE International Conference on Data Engineering (**ICDE**) — 2023
- International Conference on Scientific and Statistical Database Management (**SSDBM**) — 2022–2023
- ACM International Conference on Information and Knowledge Management (**CIKM**) — 2023
- ProvenanceWeek (**TaPP**/**IPAW**) — 2020–2023
- International Conference on Big Data Computing and Communications (**BigCom**) — 2020

Journal Reviews

- IEEE Transactions on Knowledge and Data Engineering (**TKDE**) — 2020–present
- Information Systems — 2021–present
- Distributed and Parallel Databases (**DAPD**) — 2021
- Journal of Data and Information Quality (**JDIQ**) — 2020

NSF Panels

- Information and Intelligent system (**CISE-IIS-III**) — 2023

Other Proposal Reviews

- US-Israel Binational Science Foundation (**BSF**) — 2023

External Reviewer

- International Conference on Very Large Databases (**VLDB**) 2020
- International Conference on Database Systems for Advanced Applications (**DASFAA**) 2020
- International Conference on Scientific and Statistical Database Management (**SSDBM**) 2020
- International Conference on Distributed Event-Based Systems (**DEBS**) 2020
- ACM SIGMOD International Conference on Management of Data (**SIGMOD**) 2019
- International Conference on Very Large Databases (**VLDB**) 2019
- IEEE International Conference on Data Engineering (**ICDE**) 2019
- International Conference on Extending Database Technology (**EDBT**) 2019
- ACM International Conference on Information and Knowledge Management (**CIKM**) 2019
- International Conference on Very Large Databases (**VLDB**) 2018
- IEEE International Conference on Data Engineering (**ICDE**) 2017
- ACM International Conference on Information and Knowledge Management (**CIKM**) 2017
- IEEE International Conference on Data Engineering (**ICDE**) 2015

NSF-supported Project Contribution

- Data Observation Network for Earth (**DataONE**) 2018

Invited Talks

- **Korean-American Scientists and Engineers Association (KSEA)** 2023
  Efficient and Concise Provenance Management
- **Cincinnati Children's Hospital (CCHMC)** 2022
  Introduction to Data Provenance
- **Postech** 2021
  Provenance Management
- **University of Notre Dame** 2019
  Why and Why-not Provenance for Queries with Negation

Other Experiences

University of Stuttgart Stuttgart, DE
Research Intern, Visualization Research Center Summer 2019

- Efficient computation and visualization of why-not explanations in big data analysis pipelines
- https://www.sfbtrr161.de/research/project_d03

Data Observation Network for Earth (DataOne) Chicago, US
Research Intern Summer 2018

- Sharing reproducible research through DataONE and Whole Tale
- https://www.dataone.org/intern/2018/sharing-reproducible-research-through-dataone-and-whole-tale

Compuware Korea Seoul, KR
Customer Care 2007–2010

Elitek Info & Communication Seoul, KR
Web Application Developer & Oracle DBA 2004–2006