*Last updated: May 25, 2022*

# Andreas Kipf

kipf@mit.edu | https://people.csail.mit.edu/kipf/

**Research Interests**

I am interested in building instance-optimized database systems with the help of machine learning, focusing on index structures, storage layouts, and query optimization.

**Education**

**Technical University of Munich**
Ph.D. (Dr. rer. nat) in Computer Science — January 2020
Thesis: "Advancing Analytical Database Systems"
Advisor: Prof. Alfons Kemper

**Technical University of Munich**
*Thesis work done at University of California, Berkeley*
M.Sc. with honors in Software Engineering (Elite Graduate Program) — January 2015
Thesis: "A Cloud-Based Data Management Platform for Impact Analysis"
Advisor: Prof. Eric Brewer
German grading system: 1.05

**University of Applied Sciences, Nuremberg**
B.Sc. in Information Systems and Management — March 2012
Thesis: "Large Data Transfers in Service-Oriented Architectures"
Advisor: Prof. Hans Delfs
Graduated first of class, German grading system: 1.41

**Academic Experience**

**Massachusetts Institute of Technology**, Cambridge, MA, USA
Postdoctoral Associate — 04/20 - Present
- Working with Prof. Tim Kraska.
- Collaborating with Google, Microsoft Research, and Intel Labs as part of MIT's DSAIL effort.
- Research on instance-optimized database systems.

**Technical University of Munich**, Munich, Germany
Research and Teaching Assistant — 02/15 - 03/20
- Worked with Prof. Alfons Kemper and Prof. Thomas Neumann.
- Research on in-memory geospatial data processing and learned cardinality estimation.
- Taught a Master's class on database systems and advised students.

**University of California, Berkeley**, Berkeley, CA, USA
Visiting Student Researcher — 04/14 - 09/14
- Worked with Prof. Eric Brewer.
- Research on a cloud-based data management platform for impact analysis.
- This project led to a journal publication and my Master's thesis.

**Industry Experience**

**Google, Data Infrastructure**, Munich, Germany
Student Researcher (Part-time) — 04/19 - 01/20
- Research on lightweight secondary indexing.
- Published in VLDB 2020: "Cuckoo Index: A Lightweight Secondary Index Structure".
- Cuckoo Index is up to 10x smaller than Bloom filters for low cardinality columns.

**Google, Data Infrastructure**, Zurich, Switzerland
Software Engineering Intern — 07/18 - 10/18
- Worked on lightweight secondary indexing for a new columnar storage format.
- Attended Google PhD Intern Research Conference in Mountain View.

**Google, Data Infrastructure**, Mountain View, CA, USA
Software Engineering Intern — 07/17 - 10/17
- Worked on query-driven materialization for Mesa, a scalable data warehouse.
- Collaborated with the F1 query optimizer team.
- Received two peer bonus awards.

**IBM Germany Research and Development GmbH**, Böblingen, Germany
Extreme Blue Intern — 07/13 - 09/13
- Integrated sensors and actuators into the IBM WebSphere Portal.
- Presented our project at IBM Hursley, UK.

**Avid Technology, Inc.**, Mountain View, CA, USA
Software Quality Assurance Intern — 09/10 - 03/11
- Designed and implemented automated tests for a video editing software.
- Worked with software engineering team to resolve issues.

**Honors and Awards**

VLDB Journal Best of PVLDB 2021 Invitation, 2022.
ACM TODS Best of EDBT 2017 Invitation, 2019.
Winner of the Programming Contest, SIGMOD 2017.
Best Demonstration Award, SIGMOD 2016.
Software Campus Fellow, €100,000 in funding, German Federal Ministry of Education and Research, 2015.
DAAD "FITweltweit" Scholarship for research stay at University of California, Berkeley, 2014.
Best Graduation in Information Systems, University of Applied Sciences, Nuremberg, 2012.
Max Weber Program Fellow, Elite Network of Bavaria, 2010.

**Service**

**Program Committee:**
SIGMOD 2021, 2022, 2023.
SIGMOD Demo 2022.
SIGMOD Reproducibility 2021.
ICDE 2021.
aiDM 2021, 2022.
SMDB 2020, 2021, 2022.
AIDB 2019, 2020, 2021, 2022.

**External Reviewer:**
EDBT 2022.
TKDE 2020.

**Teaching Experience**

Lecturer for Database Systems (Elite Graduate Program), TUM, Winter 2019.
4x Teaching Assistant for Database Systems (Elite Graduate Program), TUM, Winter 2015-2018.
Teaching Assistant for Database Systems on Modern CPU Architectures (Elite Graduate Program), TUM, Summer 2015.
Tutor for Mathematics I (Algebra), University of Applied Sciences, Nuremberg, Winter 2009.

**Student Advising**

**Master's Thesis:**
Bhavik Nagda (MIT), CHuff: Conditional Huffman String Compression, 2021. (Now at Covariant)
Lujing Cen (MIT), Learned Encodings in SageDB, 2021. (Now at Plaid)
Andreas Zimmerer (MIT), Dynamic Data Partitioning for Evolving Analytical Workloads, 2021. (Now at Snowflake)
Jonas Müller, Improving Cardinality Estimation with Deep Learning, 2019. (Now at Palantir)
Christoph Anneser, Evaluation of Succinct Trie Data Structures for Prefix Lookups, 2019. (Now PhD at TUM)
Florian Gratzer, Adaptive Optimization and Processing of Continuous Queries, 2018. (Now at Oracle Labs)
Maximilian Bandle, Efficient Spatio-Textual Joins, 2018. (Now PhD at TUM)
Philipp Heuer, Instant Loading of Geospatial Data, 2018. (Now at StudySmarter)
Christian Winter, An Adaptive Storage Layout for Spatio-Temporal Data, 2018. (Now PhD at TUM)
David Becher, An Efficient Nearest Neighbor Join Algorithm in Main Memory, 2017. (Now at Celonis)
Raul Alexandru Persa, Efficient Geospatial Joins Using Specialized Radix Trees, 2017. (Now at Tableau)
Jan Böttcher, Analytics on Fast Data Using Modern Stream Processing Systems, 2016. (Now PhD at TUM)

**Bachelor's Thesis:**
Jakob Meggendorfer, Geospatial Query Processing on Compressed Points, 2018. (Now at Uni Kiel)
Frederic Sauer, An Automated Weather Data Model Evaluation System, 2017. (Now at ?)
Philipp Schlieker, Mapping GPS Traces to Trajectories on Road Networks, 2017. (Now at ShopVibes)

**Guided Research:**
David Werner, Distributed High-Performance Geospatial Joins, 2018. (Now PhD at TUM)
Nikita Tselousov, Evaluation of the Google Cloud Spanner Database Service, 2018. (Now at Joyn)
Matthias Adams, Design and Optimization of a Streaming K-Means Algorithm, 2017. (Now at Snowflake)

**Conference Publications**

Adaptive Hybrid Indexes.
Christoph Anneser, **Andreas Kipf**, Huanchen Zhang, Thomas Neumann, and Alfons Kemper.
*SIGMOD 2022.*

Flow-Loss: Learning Cardinality Estimates That Matter.
Parimarjan Negi, Ryan Marcus, **Andreas Kipf**, Hongzi Mao, Nesime Tatbul, Tim Kraska, and Mohammad Alizadeh.
*VLDB 2021.*

Benchmarking Learned Indexes.
Ryan Marcus, **Andreas Kipf**, Alexander van Renen, Mihail Stoian, Sanchit Misra, Alfons Kemper, Thomas Neumann, and Tim Kraska.
*VLDB 2021.*

GeoBlocks: A Query-Cache Accelerated Data Structure for Spatial Aggregation over Polygons.
Christian Winter, **Andreas Kipf**, Christoph Anneser, Eleni Tzirita Zacharatou, Thomas Neumann, and Alfons Kemper.
*EDBT 2021.*

The Case for Distance-Bounded Spatial Approximations.
Eleni Tzirita Zacharatou, **Andreas Kipf**, Ibrahim Sabek, Varun Pandey, Harish Doraiswamy, and Volker Markl.
*CIDR 2021*.

An Evaluation of Modern Spatial Libraries.
Varun Pandey, Alexander van Renen, **Andreas Kipf**, and Alfons Kemper.
*DASFAA 2020*.

Fast Mapping onto Census Blocks.
Jeremy Kepner, **Andreas Kipf**, Darren Engwirda, Navin Vembar, Michael Jones, Lauren Milechin, Vijay Gadepally, Chris Hill, Tim Kraska, William Arcand, David Bestor, et al.
*HPEC 2020*.

Cuckoo Index: A Lightweight Secondary Index Structure.
**Andreas Kipf**, Damian Chromejko, Alexander Hall, Peter Boncz, and David G. Andersen.
*VLDB 2020*.

Low-Latency Communication for Fast DBMS Using RDMA and Shared Memory.
Philipp Fent, Alexander van Renen, **Andreas Kipf**, Viktor Leis, Thomas Neumann, and Alfons Kemper.
*ICDE 2020*.

The Case for Hybrid Succinct Data Structures.
Christoph Anneser, **Andreas Kipf**, Harald Lang, Thomas Neumann, and Alfons Kemper.
*EDBT 2020*.

Adaptive Main-Memory Indexing for High-Performance Point-Polygon Joins.
**Andreas Kipf**, Harald Lang, Varun Pandey, Raul Alexandru Persa, Christoph Anneser, Eleni Tzirita Zacharatou, Harish Doraiswamy, Peter Boncz, Thomas Neumann, and Alfons Kemper.
*EDBT 2020*.

DeepSPACE: Approximate Geospatial Query Processing with Deep Learning.
Dimitri Vorona, **Andreas Kipf**, Thomas Neumann, and Alfons Kemper.
*SIGSPATIAL 2019*.

Learned Cardinalities: Estimating Correlated Joins with Deep Learning.
**Andreas Kipf**, Thomas Kipf, Bernhard Radke, Viktor Leis, Peter Boncz, and Alfons Kemper.
*CIDR 2019*.

How Good Are Modern Spatial Analytics Systems?.
Varun Pandey, **Andreas Kipf**, Thomas Neumann, and Alfons Kemper.
*VLDB 2018*.

Approximate Geospatial Joins with Precision Guarantees.
**Andreas Kipf**, Harald Lang, Varun Pandey, Raul Alexandru Persa, Peter Boncz, Thomas Neumann, and Alfons Kemper.
*ICDE 2018*.

Analytics on Fast Data: Main-Memory Database Systems versus Modern Streaming Systems.
**Andreas Kipf**, Varun Pandey, Jan Böttcher, Lucas Braun, Thomas Neumann, and Alfons Kemper.
*EDBT 2017*.

**Journal Publications**

How Good Are Modern Spatial Libraries?.
Varun Pandey, Alexander van Renen, **Andreas Kipf**, and Alfons Kemper.
*Data Science and Engineering 2020, Special Issue of DASFAA 2020*.

Make the Most out of Your SIMD Investments: Counter Control Flow Divergence in Compiled Query Pipelines.
Harald Lang, Linnea Passing, **Andreas Kipf**, Peter Boncz, Thomas Neumann, and Alfons Kemper.
*VLDBJ 2019, Best of DaMoN 2018*.

Scalable Analytics on Fast Data.
**Andreas Kipf**, Varun Pandey, Jan Böttcher, Lucas Braun, Thomas Neumann, and Alfons Kemper.
*TODS 2019, Best of EDBT 2017*.

A Proposed Integrated Data Collection, Analysis and Sharing Platform for Impact Evaluation.
**Andreas Kipf**, Waylon Brunette, Jordan Kellerstrass, Matthew Podolsky, Javier Rosa, Mitchell Sundt, Daniel Wilson, Gaetano Borriello, Eric Brewer, and Evan Thomas.
*Development Engineering 2016*.

**Workshop Publications**

LSI: A Learned Secondary Index Structure.
**Andreas Kipf**, Dominik Horn, Pascal Pfeil, Ryan Marcus, and Tim Kraska.
*aiDM @ SIGMOD 2022*.

Bounding the Last Mile: Efficient Learned String Indexing.
Benjamin Spector, **Andreas Kipf**, Kapil Vaidya, Chi Wang, Umar Farooq Minhas, and Tim Kraska.
*AIDB @ VLDB 2021*.

Towards Practical Learned Indexing.
Mihail Stoian, **Andreas Kipf**, Ryan Marcus, and Tim Kraska.
*AIDB @ VLDB 2021*.

When Are Learned Models Better Than Hash Functions?.
Ibrahim Sabek*, Kapil Vaidya*, Dominik Horn, **Andreas Kipf**, and Tim Kraska.
*AIDB @ VLDB 2021*.

LEA: A Learned Encoding Advisor for Column Stores.
Lujing Cen, **Andreas Kipf**, Ryan Marcus, and Tim Kraska.
*aiDM @ SIGMOD 2021*.

The Case for Learned Spatial Indexes.
Varun Pandey, Alexander van Renen, **Andreas Kipf**, Ibrahim Sabek, Jialin Ding, and Alfons Kemper.
*AIDB @ VLDB 2020*.

RadixSpline: A Single-Pass Learned Index.
**Andreas Kipf***, Ryan Marcus*, Alexander van Renen*, Mihail Stoian, Alfons Kemper, Tim Kraska, and Thomas Neumann.
*aiDM @ SIGMOD 2020*.

START – Self-Tuning Adaptive Radix Tree.
Philipp Fent*, Michael Jungmair*, **Andreas Kipf**, and Thomas Neumann.
*SMDB @ ICDE 2020*.

SOSD: A Benchmark for Learned Indexes.
**Andreas Kipf***, Ryan Marcus*, Alexander van Renen*, Mihail Stoian, Alfons Kemper, Tim Kraska, and Thomas Neumann.
*ML For Systems @ NeurIPS 2019*.

Estimating Filtered Group-By Queries is Hard: Deep Learning to the Rescue.
**Andreas Kipf**, Michael Freitag, Dimitri Vorona, Peter Boncz, Thomas Neumann, and Alfons Kemper.
*AIDB @ VLDB 2019*.

Make the Most out of Your SIMD Investments: Counter Control Flow Divergence in Compiled Query Pipelines.
Harald Lang, **Andreas Kipf**, Linnea Passing, Peter Boncz, Thomas Neumann, and Alfons Kemper.
*DaMoN @ SIGMOD 2018*.

High-Performance Main-Memory Database Systems and Modern Virtualization: Friends or Foes?.
Tobias Mühlbauer, Wolf Rödiger, **Andreas Kipf**, Alfons Kemper, and Thomas Neumann.
*DanaC @ SIGMOD 2015*.

**Demonstrations**

Estimating Cardinalities with Deep Sketches.
**Andreas Kipf**, Dimitri Vorona, Jonas Müller, Thomas Kipf, Bernhard Radke, Viktor Leis, Peter Boncz, Thomas Neumann, and Alfons Kemper.
*SIGMOD 2019*.

High-Performance Geospatial Analytics in HyPerSpace.
Varun Pandey, **Andreas Kipf**, Dimitri Vorona, Tobias Mühlbauer, Thomas Neumann, and Alfons Kemper.
*SIGMOD 2016. Received Best Demo Award.*

**Invited Talks**

Cardinality Estimation in the Learned Systems Era, Sigma Computing, January 2022.
Cuckoo Index: A Lightweight Secondary Index Structure, FAU Erlangen, August 2021.
Cardinality Estimation in the Learned Systems Era, LADSIOS @ VLDB, August 2021.
A Learned Storage System for Modern SSDs, DSAIL Retreat (remote), January 2021.
Cardinality Estimation in the Learned Systems Era, TU Berlin (remote), July 2020.
RadixSpline: A Single-Pass Learned Index, CWI Amsterdam (remote), July 2020.
RadixSpline: A Single-Pass Learned Index, RelationalAI (remote), June 2020.
Cardinality Estimation in the Learned Systems Era, FG DB Symposium, TU Darmstadt, March 2020.
Learned Cardinalities, IBM Research Almaden, January 2019.
High-Performance Geospatial Data Processing, EPFL, July 2018.
Instantly Joining 1B points with Hundreds of Polygons, Google Zurich, August 2018.
A Main-Memory Database for Future Connected Mobility Workloads, HPTS Asilomar, October 2017.
High-Performance Geospatial Data Processing, Google Mountain View, September 2017.

**Technical Skills**

C, C++, Java, Python, R, Rust, SQL, LaTeX, and Linux.