# The 1st International Workshop on Data-Centric AI (DATAI)

## 1 CHAIRS AND ORGANIZERS

### 1.1 General Chairs

**Hongzhi Wang** is a professor and PhD supervisor of Harbin Institute of Technology, the secretary general of ACM SIGMOD China, CCF outstanding member. Research Fields include big data management and analysis, and data quality. He has been PI for more than 10 national or international projects including NSFC key project, NSFC projects and National Technical support project. He has won MOE technological First award, Microsoft Fellowship, IBM PHD Fellowship and Chinese excellent database engineer. He has published over 200 papers including VLDB, SIGMOD, and 6 books. He severs as the reviewer of more than 20 international journal including IEEE TKDE and PC member of over 30 internal conference. His papers were cited more than 6000 times.

**Nan Tang** an associate professor in University of Science and Technology (Guangzhou). Before joining HKUST(GZ), He worked as a senior scientist at Qatar Computing Research Institute, a research fellow at University of Edinburgh, and a scientific staff member at CWI (national research institute for mathematics and computer science in the Netherlands).

He has published over 100 research papers, mainly in SIGMOD, VLDB and ICDE. He has won many awards in database conferences, including 2023 ACM SIGMOD Research Highlight Award, Best papers of SIGMOD 2023, ICDE 2023 Distinguished Reviewer Award, VLDB 2021 Distinguished Reviewer Award, ACM SIGMOD 2020 Reproducibility Award, Best papers of ICDE 2018, Best papers of VLDB 2015, Best papers of ICDE 2012, The Best Paper Award of VLDB 2010 and Best papers of ICDE 2009. He also serves as the VLDB 2025 area chair and ICDE 2024 area chair.

### 1.2 PC Chairs

**Lei Cao**, is an Assistant Professor at the Computer Science department of University of Arizona. He also holds a research affiliation at MIT CSAIL where he spent several years as a Postdoc Associate and then a Research Scientist, actively collaborating with Prof. Samuel Madden, Prof. Michael Stonebraker, and Prof. Tim Kraska. He has conducted research in the broad areas of data systems and data science ranging from the low-level core database performance optimization to designing the high level, application specific machine learning techniques. His recent research falls in the emerging area of "Systems for AI and AI for Systems", focused on building data management and analytics tools that satisfy the SAUL properties: Scalable, Automatic, Human-in-the-loop.

**Chengliang Chai** is an associated professor in the department of Computer Science and Technology, at Beijing Institute of Technology, China. He received his PhD from Tsinghua Universitity in 2020. He received the ACM China Doctoral Dissertation Award and Best of SIGMOD Papers 2023. His research interests include leveraging data management techniques to benefit artificial intelligence, including data cleaning, data discovery and labeling and utilizing artificial intelligence to improve the database performance, including the learned optimizer, learned index and database tunner.

**Xiaoou Ding** is an assistant professor in School of Computer Science and Technology, Harbin Institute of Technology, China. She received her PhD from Harbin Institute of Technology in 2021. Her research work covers data cleaning and integration, temporal data quality management. She has published more than 15 academic papers in various international conferences and journals in database community, including TKDE, VLDB, ICDE, CIKM, DASFAA, and published 5 papers in the top Chinese journals. She has won the excellent academic paper award from China Association for Science and Technology (CAST) in 2020.

## 2 CONTACT INFORMATION

**Hongzhi Wang**, wangzh@hit.edu.cn
**Nan Tang**, nantang@hkust-gz.edu.cn
**Lei Cao**, lcao@csail.mit.edu
**Chengliang Chai**, ccl@bit.edu.cn
**Xiaoou Ding**, dingxiaoou@hit.edu.cn

## 3 THE WORKSHOP'S TOPIC AND ITS GOALS

In recent years, AI technology, especially deep learning (DL) and large language model (LLM), has garnered significant attention due to its remarkable accuracy and generalization capabilities. With the deepening of research, the structures of AI models are continuously evolving to become more intricate and sophisticated in pursuit of higher performance. This exemplifies the vibrant development and boundless potential of AI technology.

While model design is paramount, high-quality data is equally crucial for its success, which provides accurate, consistent, and representative training material for AI models, thereby enhancing their generalization capabilities. It also contributes to more efficient model training, reducing computational resources and time costs. The absence of high-quality data can expose AI algorithms to noise, biases, and erroneous information, leading to degraded model performance and potentially misleading outcomes. On the other hand, AI technology itself offers powerful tools for processing and enhancing data quality. It provides robust support across various stages of data management, from cleaning, labeling, and validation to feature engineering, ensuring data accuracy, integrity, consistency, and reliability. The symbiotic relationship between AI technology and good data underscores their mutual dependence and complementarity.

This workshop is committed to delving profoundly into the most recent research breakthroughs in data-centric AI, disseminating innovative techniques and methodologies at the forefront of the field. Additionally, it will undertake a comprehensive exploration of the emerging trends and future orientations in the evolution of high-quality data construction techniques, specifically tailored for AI technology, with a focus on large-scale models. The goal of this workshop is to engage and inspire researchers in the domains of AI and data quality management, as well as researchers who harbor a profound interest in these areas. It strives to cultivate rigorous discussions and exchanges among researchers, developers, and practitioners, collaboratively exploring the sustained advancement, design innovations, and practical applications of optimal data construction techniques that propel the progress of AI technologies.

To be specific, the topics of particular interest for the workshop include, but are not limited to:

- Data discovery for ML.
- Data cleaning & integration for ML.
- Data management during the lifecycle of ML models.
- Labeling quality *v.s.* ML performance.
- Data-efficient solutions for ML training.
- Data quality& scale *v.s.* LLM performance
- LLM-based data cleaning & integration.

## 4 HOW THE WORKSHOP COMPLEMENTS OR RELATES TO OTHER EVENTS

This workshop complements and relates to other events at VLDB in several key ways. Firstly, it offers a dedicated platform for in-depth discussions on the critical aspect of "Data-centric AI", which may not be the primary focus of other events. Secondly, the workshop bridges the gap between theory and practice by facilitating the exchange of real-world experiences and lessons learned in both data preparation and AI. Additionally, it fosters a more intimate and interactive environment for participants to engage in lively discussions and share perspectives. Furthermore, the workshop explores the intersections between data preparation and other database research areas, enriching the overall content of the conference. Finally, by emphasizing the importance of high-quality data for reliable and effective database systems, the workshop aligns with the overall theme and objectives of VLDB.

## 5 DESIRED WORKSHOP FORMAT, INCLUDING PREFERRED DATE AND DURATION

We are planning to host a half-day single-track workshop, anticipated to be attended by 40 to 60 participants. We have meticulously designed a series of activities, adhering to a tentative schedule as follows:

- A brief Introduction by one of the general chairs (5 min).
- Invited keynote presentations by 3-4 experts in their respective fields (Each 30 min +10 min Q &A).
- 5-8 paper presentations (Each 10 min + 2 min Q &A).
- Panel discussion.

## 6 THE SUBMISSION REVIEW PROCESS, INCLUDING KEY DATES AND COVERAGE OF HOW CONFLICTS OF INTEREST ARE HANDLED

Submissions must be original (not previously published and not under review in other forums). Authors are advised to interpret these limitations strictly and to contact the PC chairs in case of doubt. Paper submission must be in English. The page limits for regular papers are 8 pages, including all figures, tables, and references. All accepted papers MUST follow the formatting guidelines according to the Camera Ready formatting instructions on https://vldb.org/pvldb/volumes/17/formatting which cannot be enforced by the template.

The papers will be uploaded as PDF files to the review system. Each paper will be reviewed by at least three reviewers from the program committee (PC). We will apply the same principles for handling conflicts of interests with PC members or workshop chairs as the VLDB conference. We plan to recruit a diverse world-class program committee of about 10-15 experts in the domains of data management and analytics, data science, machine learning.

The timeline for the workshop is planned as follows:
**February 20, 2024**: The workshop website will be available. Call for Papers will be published.
**April 1, 2024**: Submission deadline for research papers.
**April 15, 2024**: Notification of authors.
**April 27, 2024**: Camera-ready version of accepted papers.

## 7 PLANS FOR PUBLICITY

We will commence the promotion of the workshop at the earliest possible time and continue until its opening. The primary forms of publicity will include the following:

- Website publicity: We are in the final stages of designing the workshop's official website, which will be hosted on GitHub. This website will feature introductory themes, calls for papers, report solicitations, schedules, and other pertinent information, which will be updated regularly according to predefined timelines.
- Social Media publicity: We will utilize the official social media platforms of our research center and laboratory to disseminate news and previews related to the workshop.
- Academic Group publicity: We will advertise the workshop within relevant academic groups to garner attention and participation from a broader scholarly audience.
- Poster publicity: Subject to authorization, we will prepare meticulously designed posters and display boards to be posted at the venue of the workshop prior to its commencement. This will serve to attract the attention of VLDB attendees on the day of the workshop.

## 8 PLANS FOR DISSEMINATING THE RESULTS OF THE WORKSHOP (E.G., PROCEEDINGS DETAILS)

After the acceptance of papers has been confirmed, we will require authors to submit camera-ready materials and signed copyright documents according to the scheduled timeline. The workshop PC members will rigorously examine whether the format of the papers meets the established requirements and will then forward them to the VLDB workshop proceeding chairs. During this process, we will collaborate closely with the VLDB workshop proceeding chairs to ensure effective management of the publication process.

## 9 POTENTIAL (OR ACCEPTED) PROGRAM COMMITTEE MEMBERS

**Dong Deng**, Rutgers University
**Raul Fernandez**, University of Chicago, USA
**El Kindi Rezig**, University of Utah, USA

**Ju Fan**, Remin University
**Jia Zou**, ASU, USA
**Jiannan Wang**, Simon Fraser University
**Meihui Zhang**, Beijing Institute of Technology
**Paolo Papotti**, Eurocom, France
**Mourad Ouzzani**, QCRI, Qatar
**Shaoxu Song**, Tsinghua University
**Xu Chu**, Georgia Institute of Technology, USA
**Ziawasch Abedjan**, Leibniz Universität Hannover, Germany
**Oscar Moll**, MIT, USA
**Peter Chen**, MIT, USA

## 10 POTENTIAL (OR ACCEPTED) SPEAKERS OR KEYNOTES

**Sam Madden**, CSAIL, MIT
**Zachary Ives**, CIS, University of Pennsylvania
**Felix Naumann**, Hasso Plattner Institute, University of Potsdam, Germany
**Ihab Ilyas**, Apple and Universityf of Waterloo (ACM Fellow, IEEE Fellow), USA and Canada
**Guoliang Li**, Tsinghua University
**Yeye He**, MSR, USA

## 11 PLANS FOR SEEKING SPONSORSHIP

We will plan the budget for the workshop based on actual requirements and subsequently identify target sponsors who have a strong connection to the content of our workshop and demonstrate a keen interest in its topic. Dedicated members of our program committee will be assigned to liaise closely with the sponsors and offer tailored sponsorship packages aimed at fulfilling their aspirations and promotional objectives. By leveraging the support of our sponsors, we aim to enhance the visibility and ensure the successful delivery of our workshop.