

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

DATAI 2024

## Abstract

The advent of artificial intelligence (AI), particularly through deep learning (DL) and large language models (LLMs), has marked a significant milestone in technological advancement, attributing to its unparalleled accuracy and generalization abilities. The rapid evolution of AI model structures to achieve superior performance underscores the dynamic progression and potential of AI technologies. However, the cornerstone of any AI's success lies not just in its algorithmic prowess but in the quality of data it is trained on. High-quality, accurate, consistent, and representative data sets are imperative for enhancing AI models' learning efficacy, thereby optimizing their generalization capabilities and reducing computational demands.

Beyond just leveraging quality data, AI technology itself plays a pivotal role in enhancing data quality through its powerful tools for data management. From cleaning, labeling, and validation to sophisticated feature engineering, AI ensures data accuracy, integrity, consistency, and reliability. This creates a symbiotic relationship between AI technology and high-quality data, highlighting their mutual dependence and the complementary nature of their interaction. It is this synergy that the 1st International Workshop on Data-driven AI (DATAI) aims to explore, delving into the latest research breakthroughs and presenting innovative techniques and methodologies at the forefront of data-driven AI.

This workshop is dedicated to fostering a comprehensive understanding of the intricate relationship between AI technologies and the data they depend on, focusing on the development of high-quality data specifically tailored for AI technologies, with a particular emphasis on large-scale models. Through engaging researchers, developers, and practitioners in rigorous discussions, the workshop seeks to explore sustained advancements, design innovations, and practical applications of data construction techniques that propel the progress of AI technologies forward.

## Topics of Interest:

- Data discovery for ML
- Data cleaning & integration for ML
- Labeling quality and ML performance
- Data-efficient solutions for ML training
- LLM-based data cleaning & integration
- Multi-modal data lakes (retrieval-)augmented large language models

By fostering a collaborative environment, DATAI aims to inspire a diverse audience of participants from the realms of AI and data quality management, facilitating an exchange of ideas that propels the field toward groundbreaking developments.

# Important Dates

- Submission Deadline for Research Papers: June 1, 2024
- Notification of Authors: June 15, 2024
- Camera-ready Version of Accepted Papers: June 27, 2024

## Organizing Team

### General Chairs:

**Hongzhi Wang**, Professor and PhD supervisor at Harbin Institute of Technology. Secretary General of ACM SIGMOD China and an outstanding member of CCF. Research focuses include big data management, analysis, and data quality.

**Nan Tang**, Associate Professor at The Hong Kong University of Science and Technology (Guangzhou).

### PC Chairs:

**Lei Cao**, Assistant Professor at University of Arizona and MIT CSAIL. Research interests in data systems and science.

**Chengliang Chai**, Associate Professor in Computer Science and Technology at Beijing Institute of Technology. Focuses on leveraging data management techniques for AI.

**Xiaou Ding**, Assistant Professor in Computer Science and Technology, Harbin Institute of Technology. Specializes in data cleaning, integration, and temporal data quality management.

## Publication

All accepted papers are required to strictly follow the formatting guidelines as per the Camera Ready instructions outlined by VLDB. These guidelines are crucial for maintaining the integrity and uniformity of the workshop proceedings. Authors must ensure their submissions are prepared according to the instructions available at [VLDB Volume 17 Formatting Guidelines](#). Compliance with these formatting requirements is essential, as the template cannot enforce them. The final papers will be rigorously reviewed by the workshop PC members to ensure adherence to these standards before forwarding them to the VLDB workshop proceedings chairs for publication.

## Contact Information:

For further inquiries, please contact the chairs through the provided email addresses in the official document.

- **Hongzhi Wang**: [wangzh@hit.edu.cn](mailto:wangzh@hit.edu.cn)
- **Nan Tang**: [nantang@hkust-gz.edu.cn](mailto:nantang@hkust-gz.edu.cn)