Letter from the Special Issue Editor

The ever-growing abundance of data in diverse formats–tables, text, images, audio, video–has led to the rise of **multimodal data lakes**, which promise unified storage and querying capabilities across heterogeneous modalities. With the advent of large language models (LLMs), new opportunities and challenges emerge in managing, understanding, and extracting value from such multimodal ecosystems.

This special issue brings together a collection of papers that address these challenges from multiple angles: foundational architectures, LLM-powered querying, optimized storage, agent-based orchestration, and socio-technical integration.

We begin with the article by **Castro Fernandez** from University of Chicago, which frames ***data discovery*** as a socio-technical process. He presents a reference architecture that separates data identification and retrieval and argues for a broader agenda called *data ecology* to bridge organizational knowledge and automated systems. Next, **Binnig et al.** from the Technical University of Darmstadt introduce ***LLM-augmented databases***, extending the relational model with natural language interfaces, multimodal query plans, and RAG-style execution pipelines. Their systems, ***CAESURA*** and ***ELEET***, showcase how traditional DBMS optimizations can coexist with AI-native capabilities. Complementing system-level design, **Sirin et al.** from Harvard University propose the ***Image Calculator***, a novel storage system that automatically tailors image storage formats to inference and training workloads. By breaking images into frequency components, they achieve substantial improvements in performance and space efficiency for image AI tasks. The MIT Data Systems and AI Lab's contribution articulates a vision for ***AI-enabled data-to-insights systems***, covering interactive pipeline orchestration, model routing, metadata-aware retrieval, and explainability. Their ***Sunroom*** prototype and ***KRAMABENCH*** benchmark push the boundaries of collaborative human-AI systems for data analysis. The system ***Taiji***, presented by **Zhang et al**. from Renmin University, adopts a Model Context Protocol (MCP) to orchestrate multimodal queries across specialized servers. This architecture supports query decomposition, feedback-driven refinement, and machine unlearning to keep both data and LLMs fresh and trustworthy. Finally, **Sun et al.** from Tsinghua University propose the concept of the ***Data Agent***, a holistic multi-agent architecture for orchestrating Data+AI pipelines. Their system, ***iDataScience***, dynamically selects agents using data skill embeddings and benchmark profiling, enabling robust and extensible analytics over complex tasks.

Together, these papers represent the state of the art in building scalable, intelligent, and trustworthy systems for multimodal data lakes. We hope this special issue will inspire further research in creating AI-native infrastructures for data-centric discovery.

Nan Tang, Yuyu Luo
The Hong Kong University of Science and Technology (Guangzhou)