Proposal: Eighth International BuildSys Workshop on DataFM: Data Acquisition & Analysis with Foundational Models

Organization

Co-Chairs & TPC Chairs

- **Shiwei Fang (Assistant Professor, Augusta University, shfang@augusta.edu):** Shiwei Fang is an assistant professor of School of Computer and Cyber Sciences at Augusta University. He has previously chaired the SenSys+BuildSys DATA workshop for three years. His research interest falls under the broad umbrella of Cyber-Physical Systems (CPS).

- **Yasra Chandio (PhD Candidate, graduating in Summer 2025):** Yasra Chandio is a final-year Ph.D. candidate in Electrical and Computer Engineering at the University of Massachusetts Amherst. Her research in Human-Centered Computing focuses on building adaptive Mixed Reality systems that optimize user experience through real-time sensing and system adjustment. Her work spans HCI, ML, and Systems, and has been recognized by CPS Rising Stars, the Heidelberg Laureate Forum, and media outlets like BBC and ScienceDaily. She has served as an editor, reviewer, and TPC chair of Sensors S&P and TPC member for top venues including ACM VRST, IEEE VR, ACM CHI, XRsecurity, and BuildSys.

Workshop Format & Logistical needs

We propose that the DataFM workshop be a full-day event in hybrid mode, and will go fully remote if in-person is not feasible. For hybrid mode, a projector, wireless internet access, and Zoom access are needed. We are prepared to run DataFM ‘25 fully remotely if needed.

Planned Publicity Procedure

We will send the CfP to the relevant email lists as well as prior attendees of DataFM. We will also send personal invitations to authors who have published work that has a germane sensor data interest in BuildSys, SenSys, IPSN, MobiSys, ISMAR or MobiCom in the last two years. In addition, we will contact the BuildSys TPC chairs to send personal invitations to authors who submitted to BuildSys 2025. Other than emailing CfP, we will post the workshop information on social media platforms including LinkedIn, X (Twitter), Facebook, WeChat groups, etc.

Workshop History

This will be the Eighth BuildSys DataFM workshop, previous as DATA workshop.

- DATA 24: 8 papers accepted, ~15 attendees
- DATA 23: Canceled

- DATA 22: 15/20 papers accepted; ~40 attendees
- DATA 21: 12/15 papers accepted; ~30 attendees
- DATA 20: 9/11 papers accepted, ~20 attendees
- DATA 19: 16/21 papers accepted
- DATA 18: 14/15 papers accepted

All held in conjunction with ACM SenSys/BuildSys

Abstract

As enthusiasm for and success in the Internet of Things (IoT), Cyber-Physical Systems (CPS), and Smart Buildings continues to grow, so too does the volume and variety of data generated by these systems. This raises important questions: How can we ensure high-quality data collection? And how can we maximize the utility of this data so that multiple projects can benefit from the time, cost, and effort invested in deployments?

With the rise of Foundational Models—particularly Large Language Models (LLMs)—we now have new tools that can potentially transform how we work with cyber-physical data. Yet, real-world data presents notable challenges, including diverse modalities, limited dataset sizes, and unstructured formats. Recent advances in large AI models, especially those based on transformer architectures, offer promise for improving how data is acquired, analyzed, manipulated, and consumed.

The DataFM: Data Acquisition & Analysis with Foundational Models workshop aims to look broadly at interesting data from interesting sensing systems and/or how such data can be adapted to Foundational Models. The workshop considers problems, solutions, and results from all across the real-world data pipeline. We solicit submissions on unexpected challenges and solutions in the collection of datasets, on new and novel datasets of interest to the community, on experiences and results, explicitly including negative results, in using prior datasets to develop new insights, and on discussions of impact and newfound opportunities with large AI foundational models.

Foundational Models could enhance data quality through sophisticated data cleaning, preprocessing, and augmentation techniques. They can also facilitate the analysis of data streams while identifying anomalies, inconsistencies, and potential biases. Generative AI can also create synthetic datasets that maintain the essential characteristics of real-world data while expanding the available training samples. This may be valuable when real data is challenging due to privacy concerns or logistical constraints. Transformer models can integrate multi-modal data, such as blending textual inputs from sensor logs with quantitative data from measurements. This new flavor of AI-driven analysis can factor in more contextual information, opening new areas of research in enhancing the predictive and diagnostic capabilities of data-driven AI systems deployed in smart environments.

Furthermore, new areas of future work may emerge from exploring the ethical implications of deploying Foundational Models within these domains, ensuring that the benefits of AI are equitably distributed while safeguarding user privacy. The workshop's focus on privacy

challenges and solutions becomes increasingly relevant in the era of AI, where the capacity to analyze vast amounts of sensitive data poses significant risks.

The workshop aims to bring together a community of application researchers and algorithm researchers in the sensing systems and building domains to promote breakthroughs from the integration of the generators and users of datasets. The workshop will foster cross-domain understanding by enabling both the understanding of application needs and data collection limitations.

The workshop seeks contributions across two major thrusts, but is open to a broad view of interesting questions around the collection, dissemination, and use of data as well as interesting datasets:

- The collection, evaluation, analysis, and use of data
    - Role of cyber-physical or similar data and metadata for informing training and inference of foundational models and applications, such as LLMs.
    - Insights on generative AI to synthesize data
    - Usage of multi-modal data within a single AI model
    - Pitfalls on AI models with cyber-physical or similar data and metadata
    - Potential applications of large AI models within cyber-physical space
    - Challenges and solutions in privacy protection with large AI models
    - Cyber-physical data embedding techniques for existing foundational models
    - Challenges and solutions in data collection, especially around security and privacy
    - Challenges and solutions in hardware/system design of data collection devices.
    - Expectations and norms for data collection from sensor networks, especially those that involve human factors
    - Novel insights from existing datasets
    - Metadata management for complex datasets
    - Synthetic data, including its generation, application, and utility
    - Success stories, key properties of useful datasets and how to generalize these
    - Preprocessing, cleaning, and fusing datasets
    - Preliminary analysis and visualization of the data
    - Shortcomings of prior datasets, and how to address these in the future
    - Position papers on policies and norms from experimental design through data management and use are explicitly welcomed.
- New and interesting datasets, including but not limited to:
    - Smart building, occupancy, motion data, energy, human comfort, vibration, BIM
    - Indoor localization, especially unprocessed/unfiltered physical layer measurements
    - Shopping-related sensing data
    - Animal-related data or sensed data
    - Anonymized health, or synthetic health-related data
    - Anonymized human-centric interaction and physiological data from applications such as Extended Reality

  - Vehicular, GPS, cellular, or wifi traces and remote sensing
  - Reproductions of prior work that validate, refute, or enhance results
  - Anonymized contact tracing, interaction, and exposure notification data

To enable the longevity of submitted datasets, we plan on providing a central location where a repository for the data, and information about the data can be archived for at least 5 years.

Requirements

Each accepted submission is required to have at least one author attend the workshop and present to the workshop attendees.

Full Papers

Submissions may range from 2-5 pages in PDF format, excluding references, using the standard ACM conference template. Submissions are strongly encouraged to use only as much space as needed to clearly convey the ideas, contributions and the significance of the work. We fully expect many submissions, especially datasets, to use only 2-4 pages but wish to allow those interested in fully elucidating positions on data collection and use or insights from reproducibility efforts ample space to do so.

Datasets

Dataset submissions should prefix paper titles with “Dataset: “ and must include a description of the dataset as well as a reasonable accompanying data sample. Once accepted, a fully described dataset must be shared to a public repository by the camera-ready deadline. Issues on licenses will be resolved by generally following the procedure similar to CRAWDAD (https://crawdad.org/joinup.html) and special treatments, if needed, will be discussed separately with the TPC chairs. **The dataset submission must submit a link to the dataset at the time of submission.**

Datasets will be reviewed by an artifact evaluation committee. To support this, dataset submissions must include:

- A link to the full dataset (not just a single sample) at the time of submission
- An example analysis or result from the dataset (what kind of insights might folks glean?)
- Steps to run an analysis on the dataset, e.g.
  - A graph and the steps (sample code) to generate the graph
  - A video demonstrating access and manipulation of the data or execution of queries and results on the data
  - Other evidence or demonstration of how the dataset can be accessed and used

The evaluation committee will work with submitters to ask clarifying questions, etc. The goal is not to be a barrier to submission, but instead to help make sure datasets are usable and useful for folks in the future.