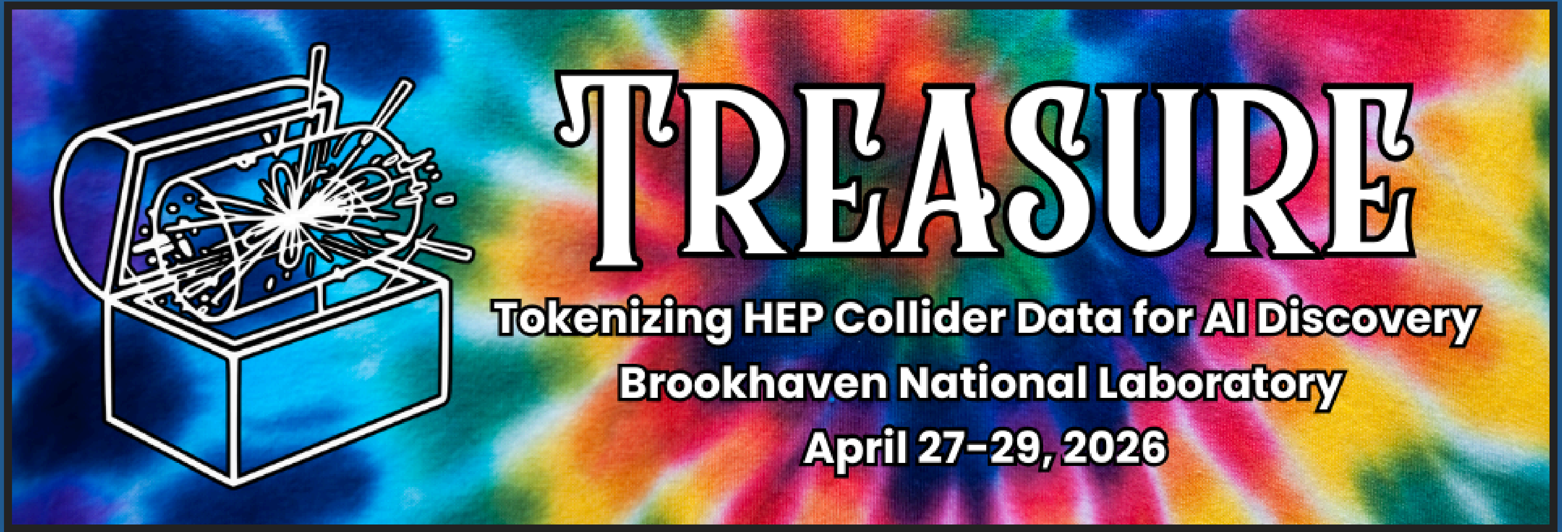
A Cross Experimental AI toolkit for HEP Experimental Operations



Walter Hopkins on behalf of HAAI: Next-Gen DQM with AI team
Amy Bender, Gabriele Benelli, Jijun Chen,
Simon Corrodi, Katie Harrington,
Aron Soha, Azton Wells, Junjie Zhu

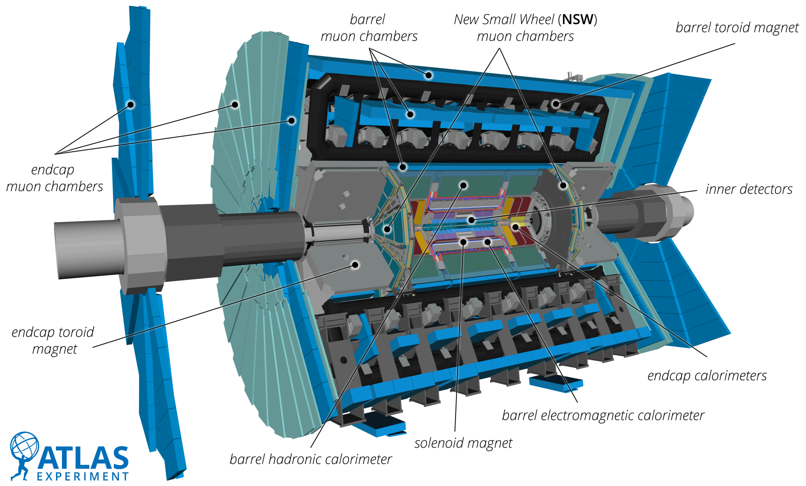
Argonne NATIONAL LABORATORY

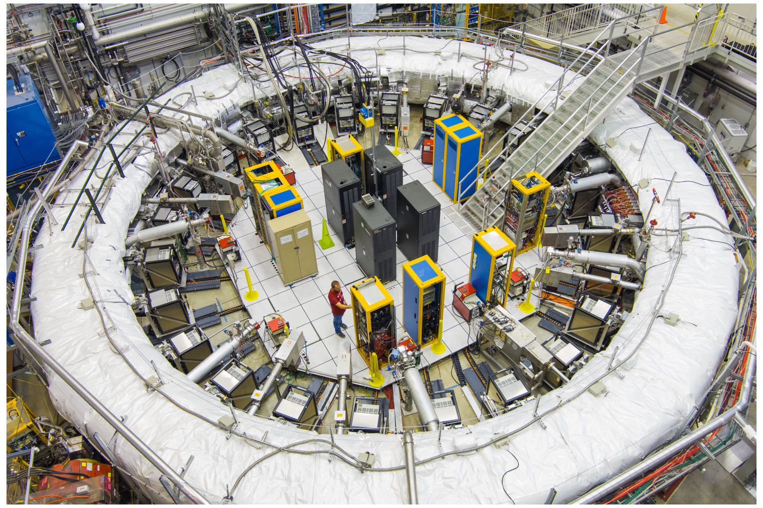
MOTIVATION: BETTER FAULT DETECTION, LESS HUMAN EFFORT

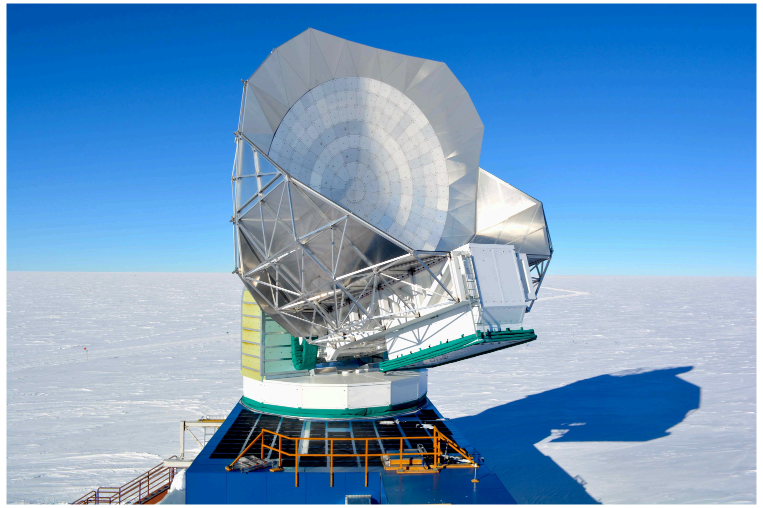
- Modern experiments produce enormous amounts of monitoring data
- Current monitoring systems include automation but require significant human effort
- Recent relevant AI/ML advances
  - Anomaly detection: we have lots of good data and don't know how future faults will look
  - Transformers: have long-term context awareness (i.e., they take past data into account)
  - Large-Language Models (LLMs): can process large bodies of documentation and logs

Combine different fault detection models and LLMs to find subtle patterns in high-dimensional data and to suggest corrective actions

OPERATIONS ACROSS HEP EXPERIMENTS



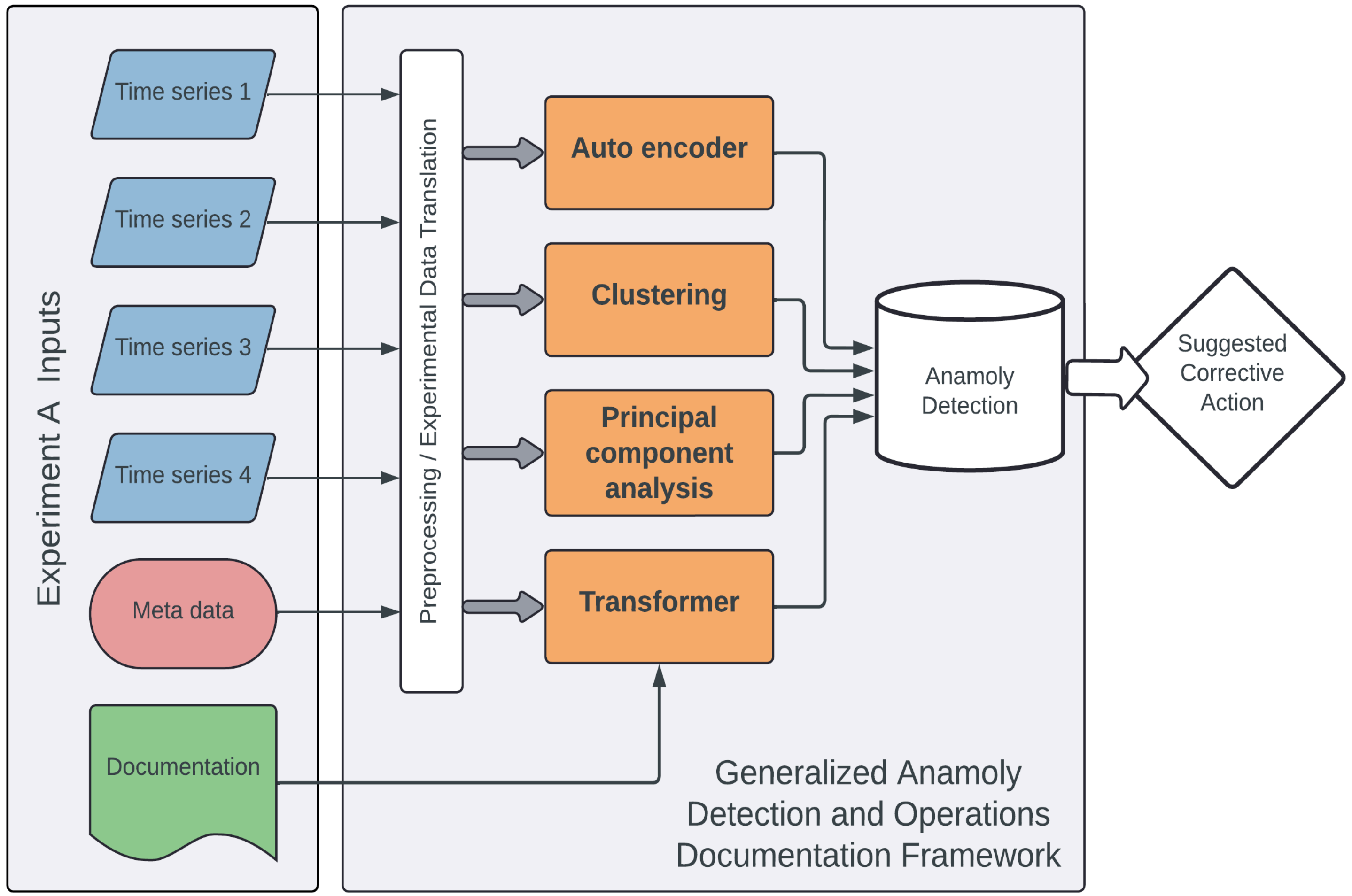




- Many HEP experiments have similar monitoring needs
- Similar data types (time series, histograms/heat maps, logs)
- Similar workflows:
  - Detect fault
  - Find and follow troubleshooting documentation
  - Log problem
  - Potentially notify and summarize problem to expert

Our goal: develop cross-experiment AI toolkit that leverages commonalities to reduce development effort and increase robustness

TOOLKIT OVERVIEW



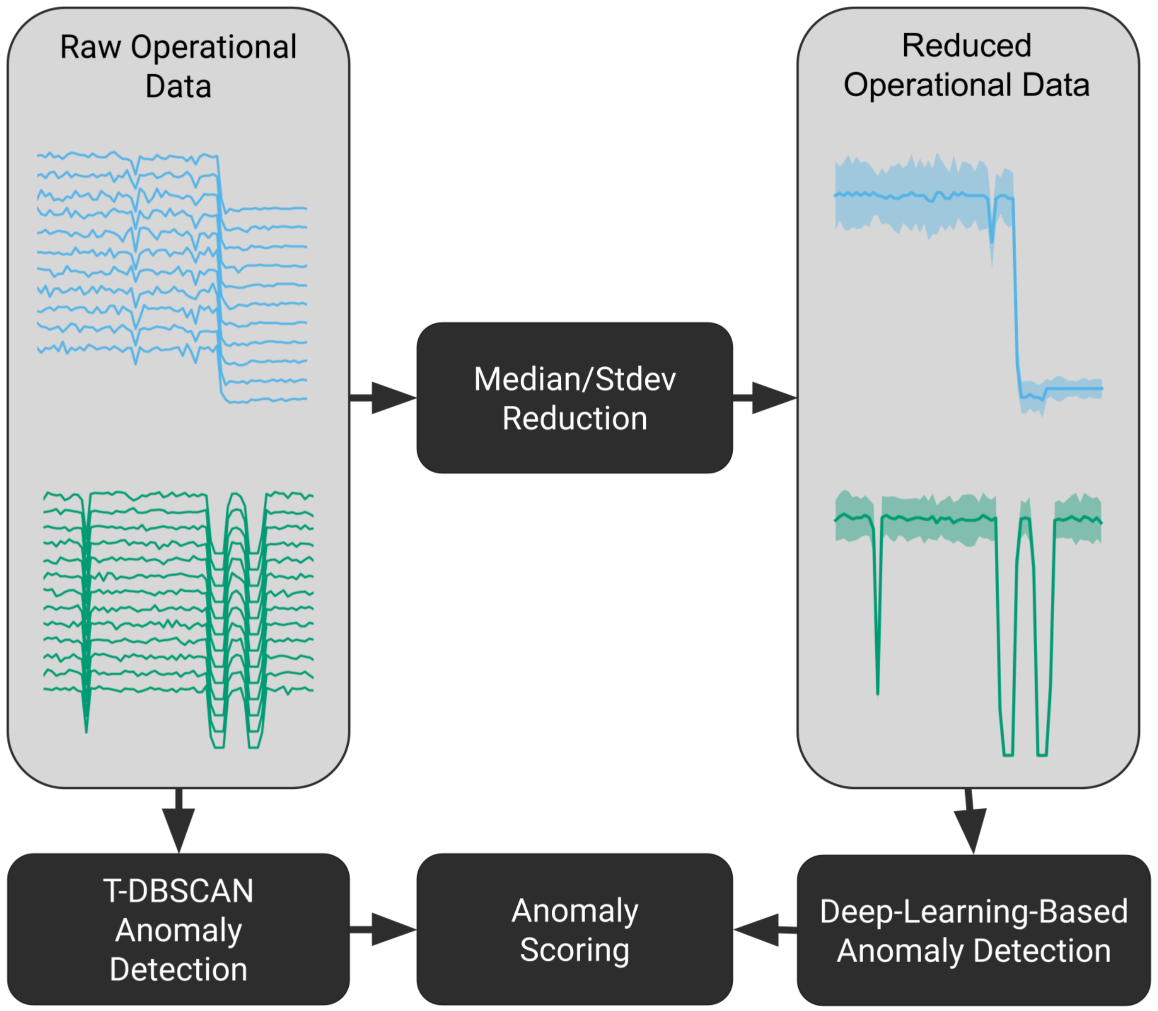
- Common data-preprocessing methods
- Modular AI/ML models for different tasks
  - Dimensionality reduction: clustering, PCA
  - Autoencoders
  - Transformer models
- Orchestration layer to combine model outputs
- LLM integration for documentation and log parsing

Toolkit is aimed at R&D: set of functions, examples, and documentation that experiments can fine tune and integrate into their operations framework

Will also build knowledge base for adapting models developed outside of HEP

INITIAL TOOL: DEEPHYDRA

- Developed for HLT computing farm
  - Analyzes DCM rate to detect problems in racks and nodes
- Combines
  - Clustering to find issues with individual nodes within a rack
  - Deep learning models (e.g., transformers) to find problems with racks
- Added value when only using simple algorithms
- See Joaquin's SCOPE updates for details on the improved HLT implementation
- Attempted to apply DeepHYDRA on data from Muon g-2 and South Pole Telescope 3G (SPT-3G)...



arXiv:2405.07749

DeepHYDRA works for DCM data but was too fine tuned to work with data from other experiments

GENERALIZATION FOR USE ON DIFFERENT EXPERIMENTS

Initially, applied existing non-HEP-specific models (e.g.,TranAD, transformer-based anomaly detection model) to different experiments' data with minimal changes to the model

Different experiments' data have different characteristics:

- ATLAS TDAQ data:
    - DCM rate as a test. Already used in DeepHYDRA, simple 1D data.
- Muon g-2: test simple multi-dimensional time-series (3 features)
- SPT-3G: large channel count (~10K) and natural grouping of channels

Goal: learn from applying models to different data and modify models for HEP-specific needs, not to train optimal model for all experiments

MODELS AND SCALING

- TranAD couldn't be trained on 10K channels of SPT-3G data due computational/model architecture constraints
- There are methods to reduce the input diminsionality
  - Clustering, PCA
  - Autoencoder-based models
  - Domain knowledge, e.g., using rack medians and standard deviations for DCM rate
- Dimensionality reduction forced by computing constraints is suboptimal

Goal of project is not to build models from scratch but to:

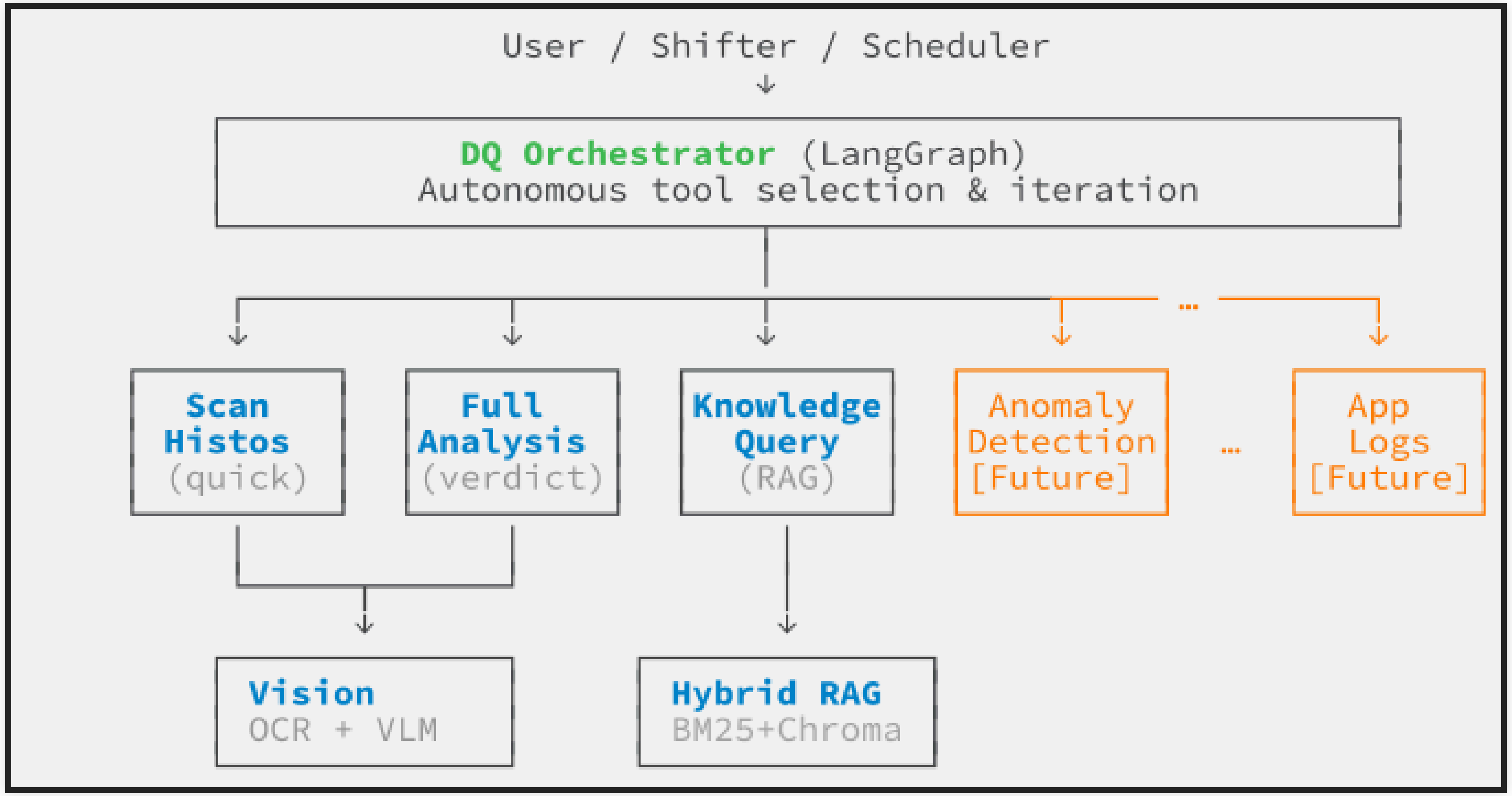
- Modify models for HEP-specific needs (although these can be similar across experiments)
- Leverage High-Performance Computers (HPCs) which are optimized for AI large-scale training

RECENT DEVELOPMENTS: ATLAS INTEGRATION

- ATLAS DCM-specific implementation was needed
  - Significant changes to TranAD model implementation
  - Lots of data preprocessing acrobatics
- Currently running on 2026 data
  - Anomalies (bit flaps) found that were missed by sys admins
- Will move to HAAI code after 2026 running

RECENT DEVELOPMENTS: ORCHESTRATION WITH AGENTS

- Prototype for agent-based orchestration was developed
- Promising results: pileup dependence inferred by agents
- Time-series anomaly detection not yet included



CONNECTION TO FOUNDATION MODELS

- Foundation model benefits:
  - Potential for transfer learning across experiments
  - Faster fine-tuning with less data during commissioning
- Can an operations-based foundation model be built? Maybe?
  - Commonalities in detector control systems data (e.g., cryogentics, voltages, etc)?
  - Commonalities in error reporting and diagnostics?

Foundation models are not a requirement for this project but is a future research direction