Opportunities for AI/ML in Streaming Readout

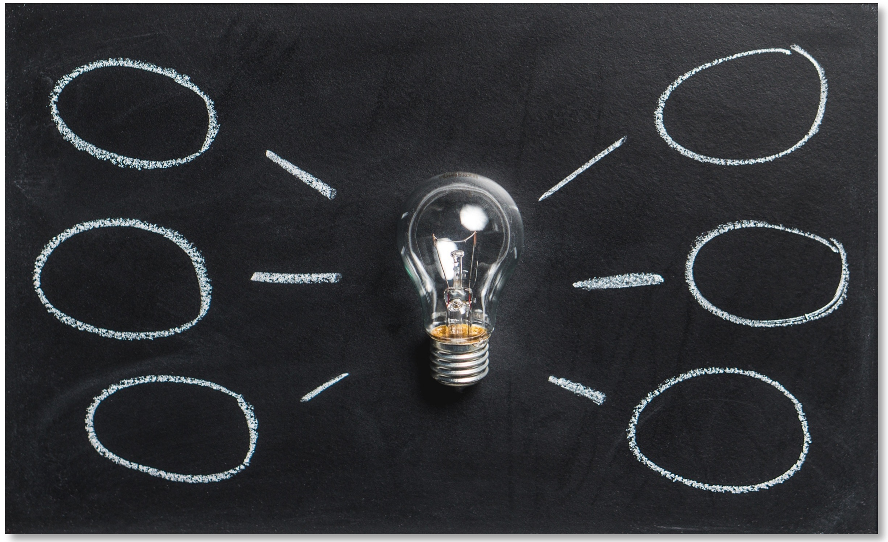
Introduction to Streaming Readout

Streaming Readout and AI/ML for rapid turnaround of data and starting the work on publications

Opportunities at the Electron-Ion Collider

Markus Diefenthaler

Towards the next-generation research model in Nuclear Physics



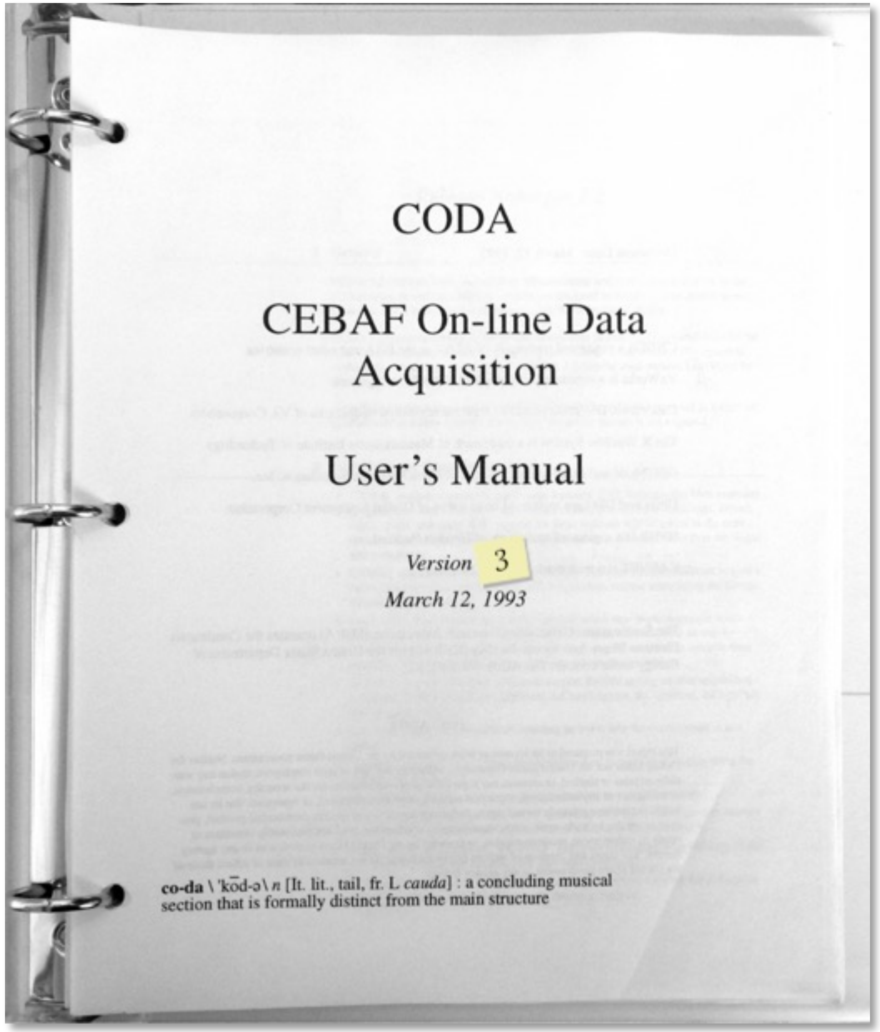
Science & Industry remarkable advances in electronics, computing, and software over last decade

Evolve & develop **Nuclear Physics research model** based on these advances

Roles of computing Data processing from data acquisition (DAQ) to analysis largely shaped by kinds of computing that has been available
Example Trigger-based readout systems

Advances in electronics, computing, and software Unique opportunity to think about new possibilities and paradigms
Example Streaming readout systems

CODA: Trigger-based readout system



Based upon assumptions in traditional DAQ design

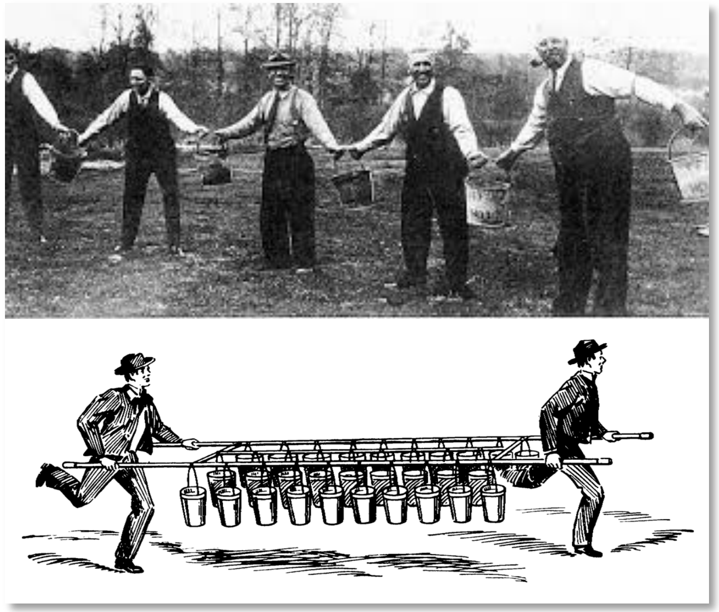
- The data rate from a detector is impossible to capture with an affordable data acquisition system without a trigger to reduce event rates.
- Even if the untriggered data rate could be captured, it would be impossible to store.
- Even if it could be stored the full dataset would represent a data volume that would require impractically large computing resources to process.

With computing advances Assumptions no longer valid

Limitation in trigger-based readout systems

- bias to low-energy particles
- do not deal well with event-pileup
- not an ideal for complex, general-purpose detectors

Alternative readout mode: Streaming



Traditional trigger-based readout

- data is digitized into buffers
- trigger starts readout
- parts of events are transported to an event builder where they are assembled into events
- at each stage the flow of data is controlled by *back pressure*
- data is organized sequentially by events

Streaming readout

- data is read continuously from all channels
- validation checks at source reject noise and suppress empty channels
- data then flows unimpeded in parallel channels to storage or a local compute resource
- data flow is controlled at source
- data is organized in multiple dimensions by channel and time

Streaming Readout: Trigger-less data acquisition

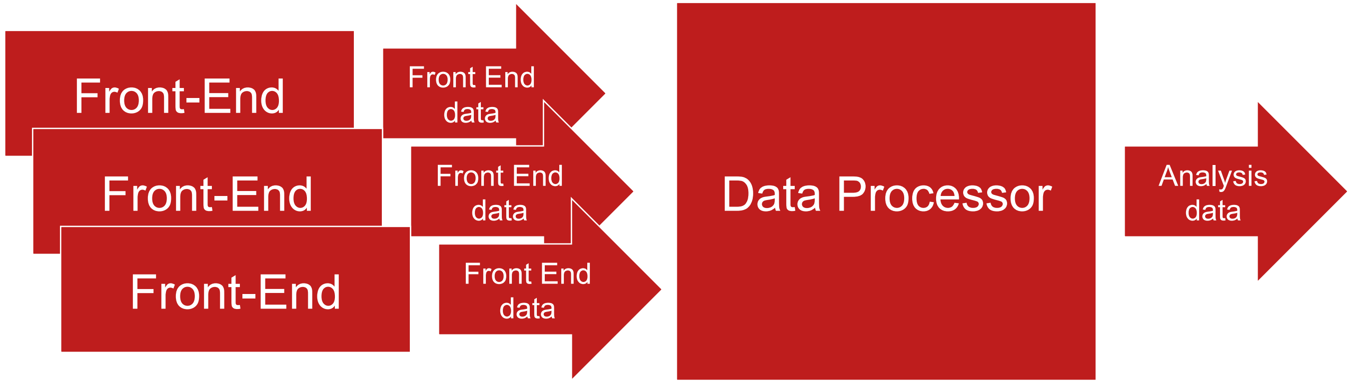
Definition of Streaming Readout

- Data is digitized at a fixed rate with thresholds and zero suppression applied locally.
- Data is read out in continuous parallel streams that are encoded with information about when and where the data was taken.
- Event building, filtering, monitoring, and other processing is deferred until the data is at rest in tiered storage.

Advantages of Streaming Readout

- simplification of readout (no custom trigger hardware and firmware)
- trigger-less readout:
  - beneficial for experiments that are limited by event-pileup or overlapping signals from different events
  - beam time is expensive so data mining or taking generic datasets shared between experiments is becoming popular: loosen triggers to store as much as possible
- opportunity to streamline workflows
- take advantage of other emerging technologies

Streaming Readout and (near) real-time processing



Data Processor

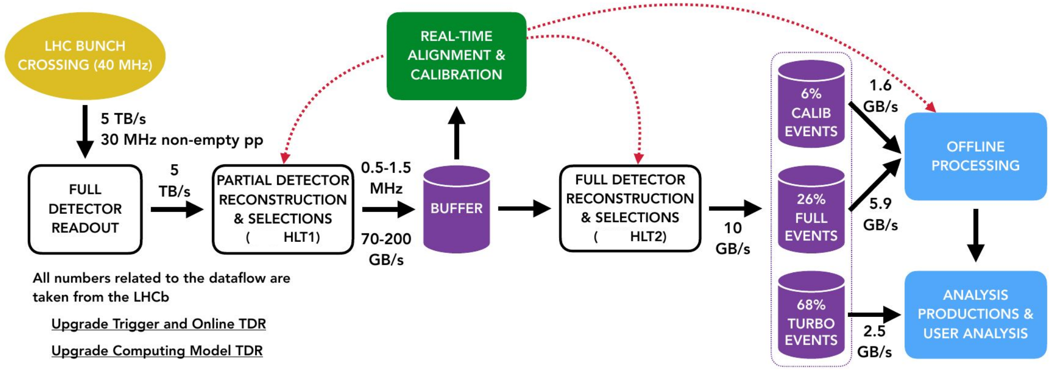
- assembles the data into events
- outputs data suitable for final analysis (**Analysis data**)

Features

- ideal for AI
- autonomous calibration in near real time
- autonomous alignment in near real time
- reconstruction in (near) real time
- event filtering into analysis streams based on full event information
- autonomous anomaly detection
- responsive detectors (conscious experiment)

LHCb Example

LHCb Upgrade Dataflow

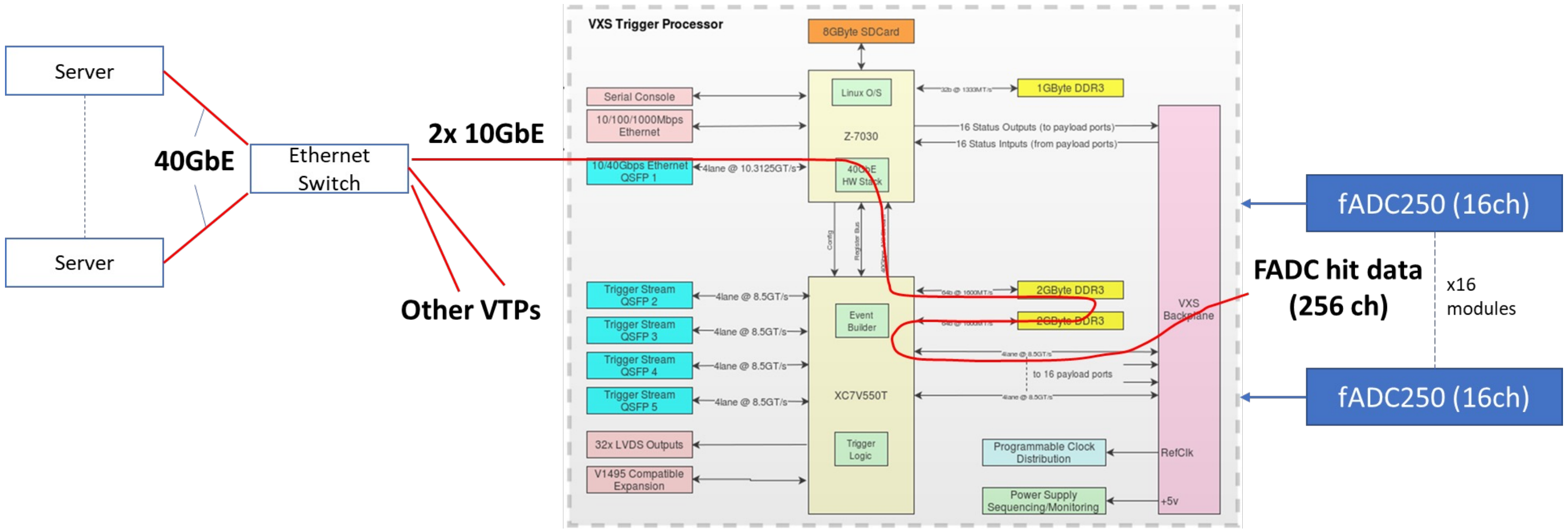


HLT1 challenge: reduce 5 TB/s to 70-200 GB/s in real-time with high physics efficiency

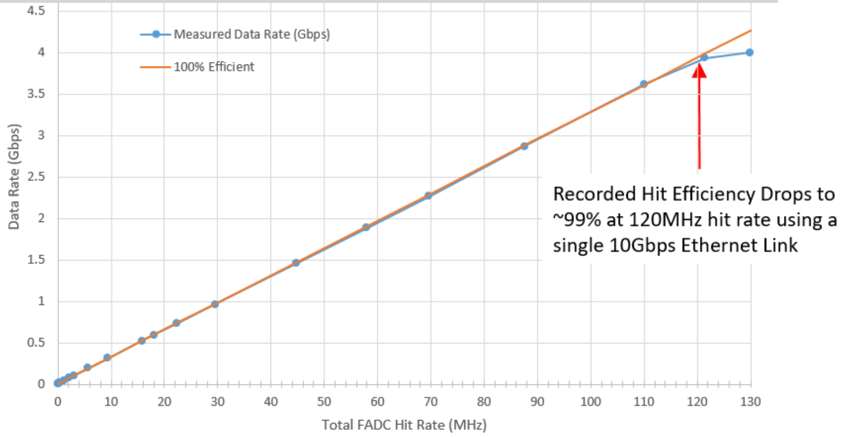
Streaming Readout test using the CLAS12 forward calorimeter (arXiv:2202.03085)

Front-end setup 2 VXS crates, each with VTP w/ 2x 10GbE optical links, 11 FADC250 modules, 336 PbWO crystals w/APD

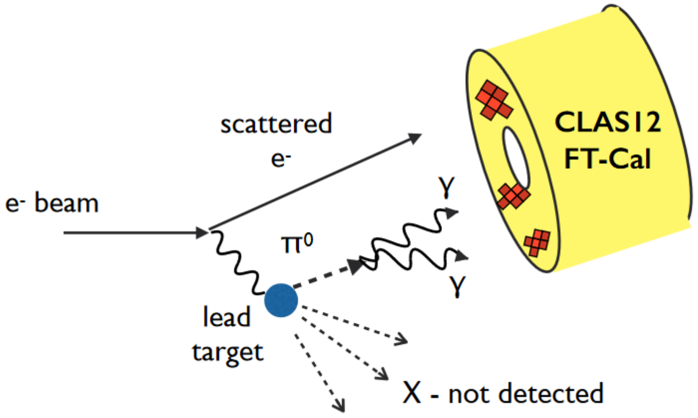
Backend setup servers connected to front-end ethernet switch by 40GbE. Combination of CODA, TRIDAS (by INFN), JANA2, ROOT for configuration, event selection, event reconstruction, and online monitoring



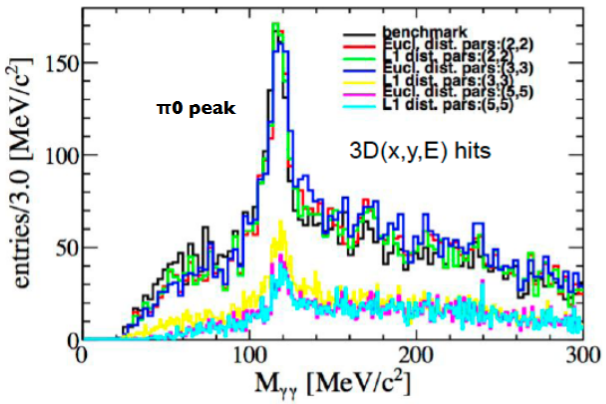
Rate capability Close to 1MHz rate per FADC channel before suffering efficiency loss (nearly 2x better after recent bandwidth improvements)



π^0 reconstructed signal from streaming readout test



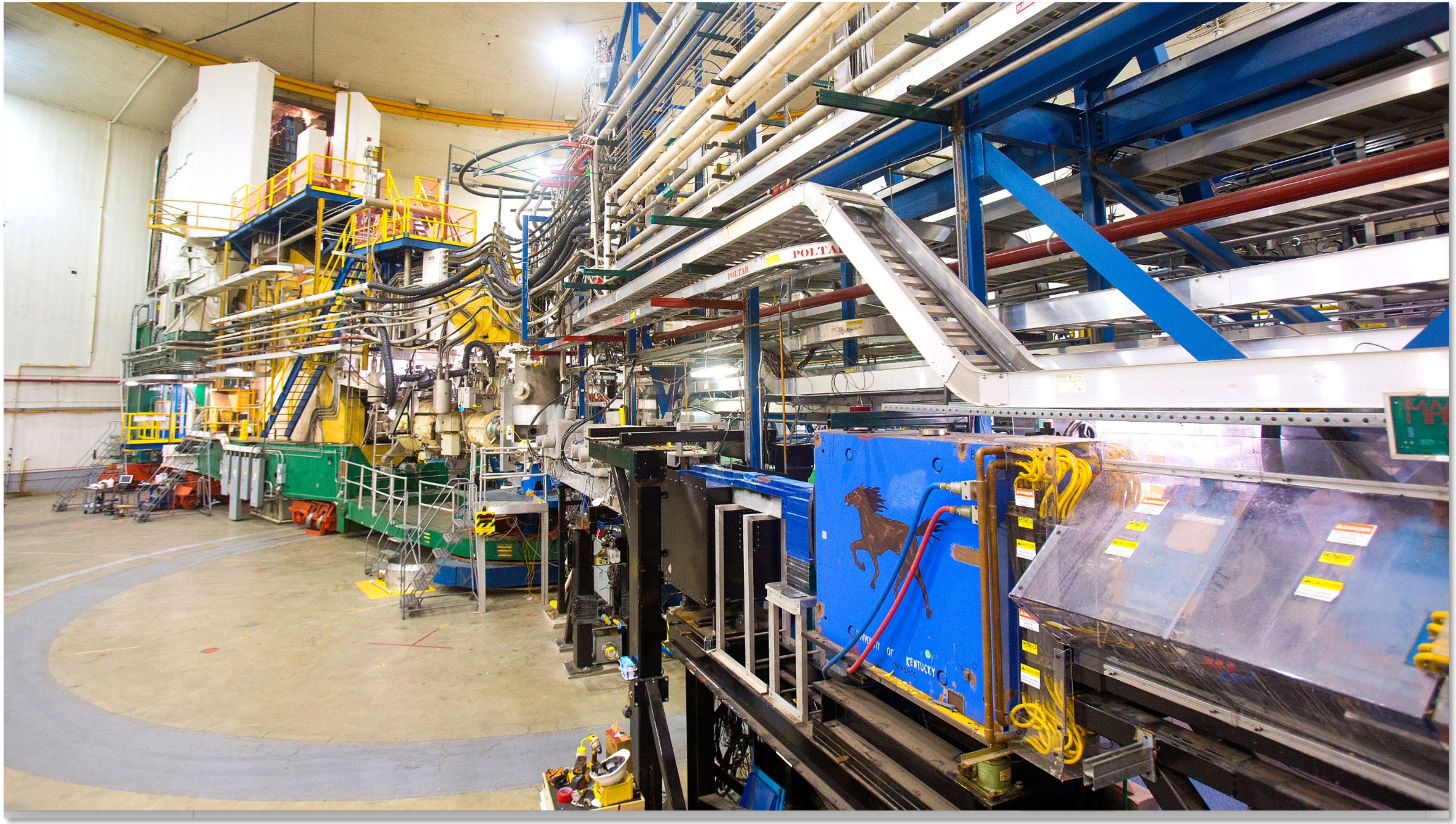
Double cluster π0 mass as obtained by an unsupervised hierarchical clustering algorithm implemented in JANA framework by C.Fanelli



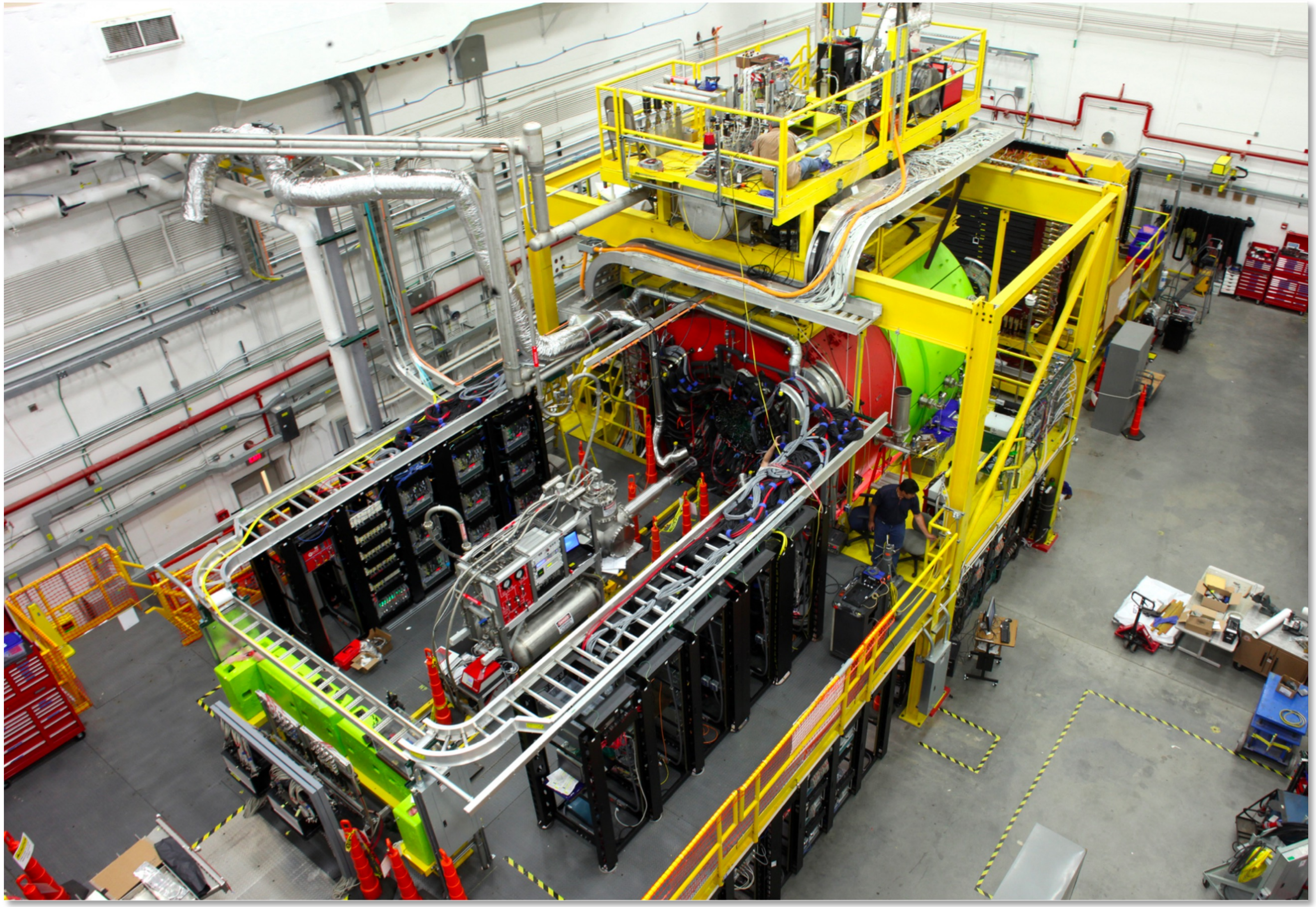
Streaming Readout

The Case For Automation

In Nuclear Physics we build cathedrals for the proton.



Building and understanding a large-scale experiment is a multi-year project.



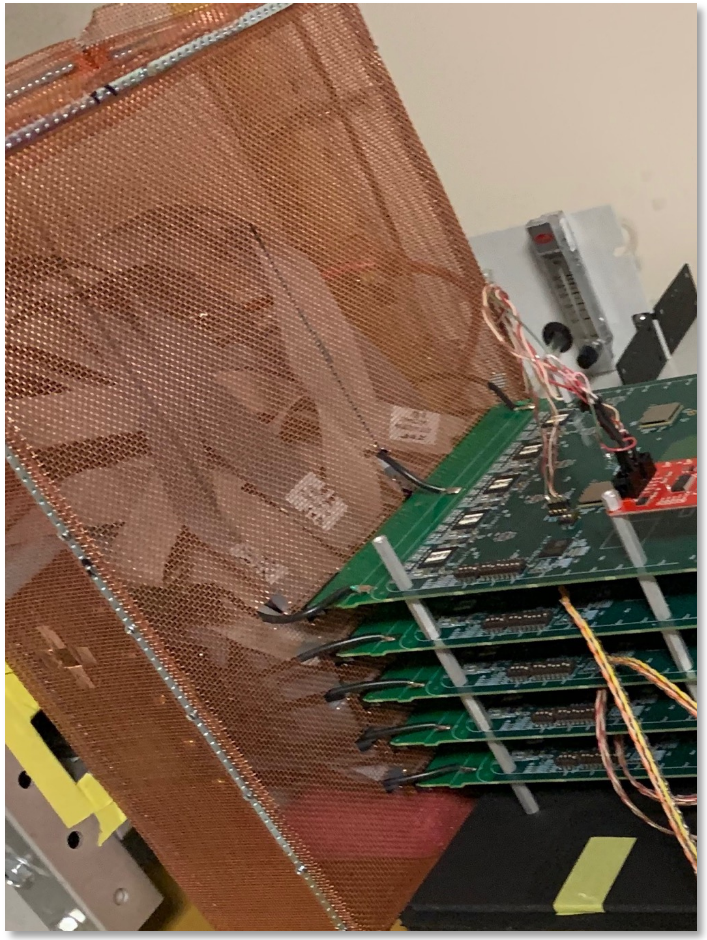
Unfortunately, it can take also a year or longer to align and calibrate the detector and reconstruct the events.

The important role of alignment and calibration

- Complexity of detectors means that it can take a year or even longer for us to "see" the events. Possible problems we find in the data a year or later, we cannot fix anymore.
- Also, it can take a year or longer to have the data for the physics analysis ready and start the work on publications.
- **Why is this?** The experts needed for alignment and calibration are the experts working on keeping the detector running.
- **The way out**

Automation
If we have AI/ML to align and calibrate detectors, we can find problems in data while data taking and start the physics analysis for the publications now.

Address Challenges of Autonomous Control and Experimentation



INDRA-ASTRA

Develop a prototype for a fully automated, responsive detector system as a first step towards a fully automated, self-conscious experiment.

R&D integrated with streaming readout and AI/ML efforts at Jefferson Lab

Team

Jefferson Lab

- **ENP** M. Diefenthaler, E. Jastrzembski, H. Szumila-Vance
- **CST** D. Lawrence, V. Gyurjyan

Old Dominion University

- **Applied Numerical Mathematics** R. Fang, A. Farhat, Y. Xu

Databricks

- S. Rajamohan

Automated data-quality monitoring and calibrations

"In most challenging data analysis applications, data evolve over time and must be analyzed in near real time. Patterns and relations in such data often evolve over time, thus, models built for analyzing such data quickly become obsolete over time. In machine learning and data mining this phenomenon is referred to as **concept drift**." (I. Žliobaitė, M. Pechenizkiy, J. Gama , An Overview of Concept Drift Applications)

To deal with time-changing data, one needs strategies, at least, for the following

- detecting when a change occurs
- determining which examples to keep and which to drop
- updating models when significant change is detected

OUR APPROACH

1. **Identify different data-taking periods** Use ADWIN2 or multi scale method to identify the start of distinct data-taking periods based on changes in the mean of the data stream.
2. **Calibrate different data-taking periods to a baseline** Use Hoeffding's inequality to estimate the mean of each data-taking period and apply a constant shift to each data taking period by the difference between the means of a baseline period and each subsequent period.

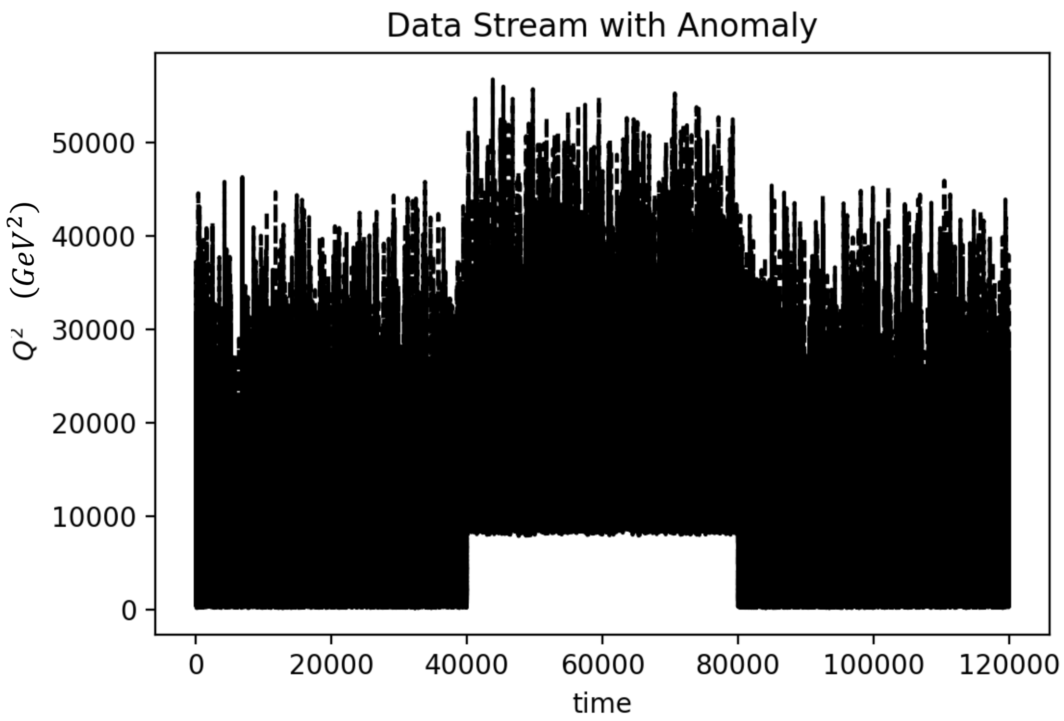
An example data stream

To represent the data stream we use a sample of 120,000 Inclusive Deep Inelastic Scattering Monte Carlo events

- generated in the context of the ZEUS experiments
- Includes full detector simulation
- Reconstructed kinematics with all detector effects.

We observe a stream of x and Q^2, reconstructed by the electron method [3] based on the measurement of the (x, y, z) position and energy E of the outgoing lepton in the calorimeter.

We subdivide the stream into 3 data-taking periods of equal parts and apply a constant shift of two standard deviations to each (x, y, z) position and energy E measurements in the second data taking period.

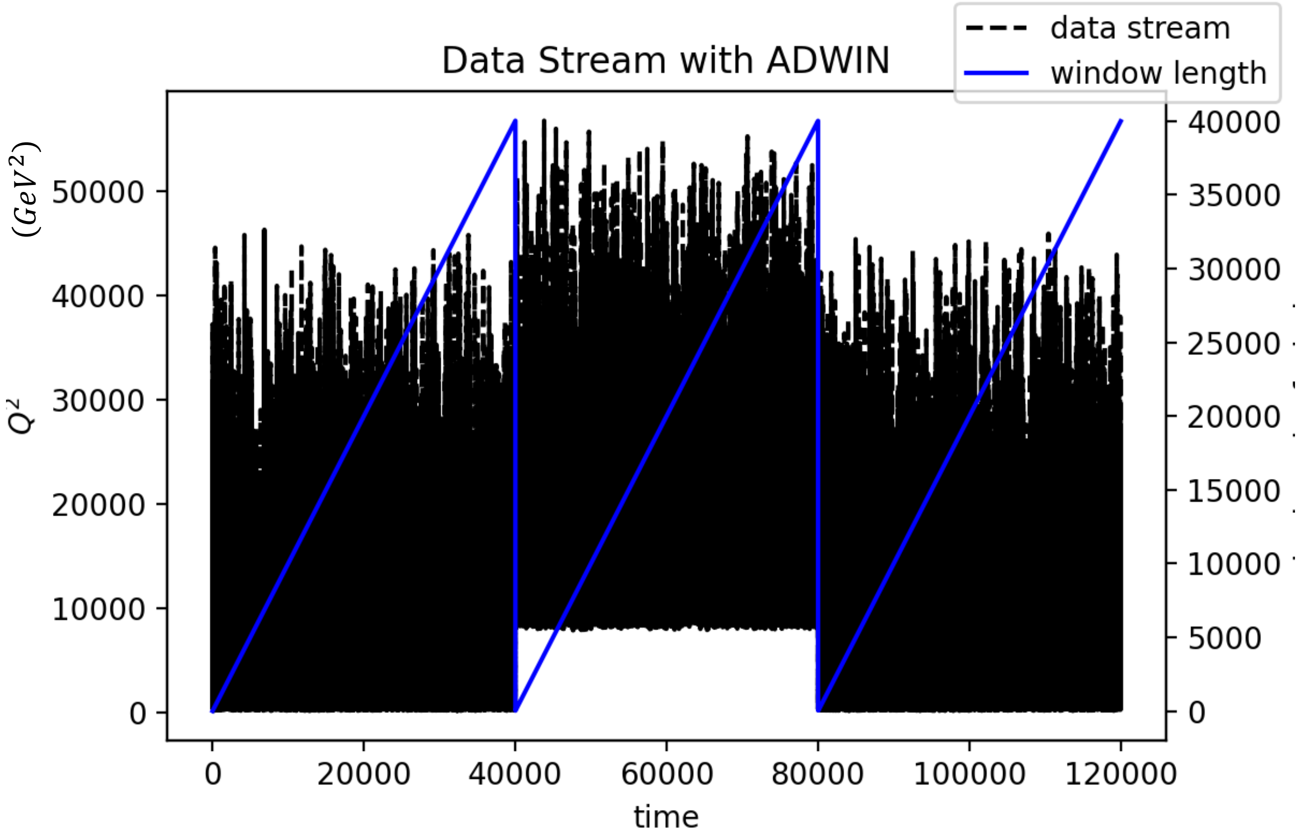


An example data stream

ADWIN is an **ADaptive WINdowing technique** used for detecting distribution changes, concept drift, or anomalies in data streams with established guarantees on the rates of false positives and false negatives

(A. Bifet and R. Gavalda, *Learning from time-changing data with adaptive windowing*, in Proceedings of the 2007 SIAM international conference on data mining, SIAM, 2007, pp. 443–448)

Data Period	Start Time	Time ADWIN Detects Change
2	40000	40020
3	80000	80012

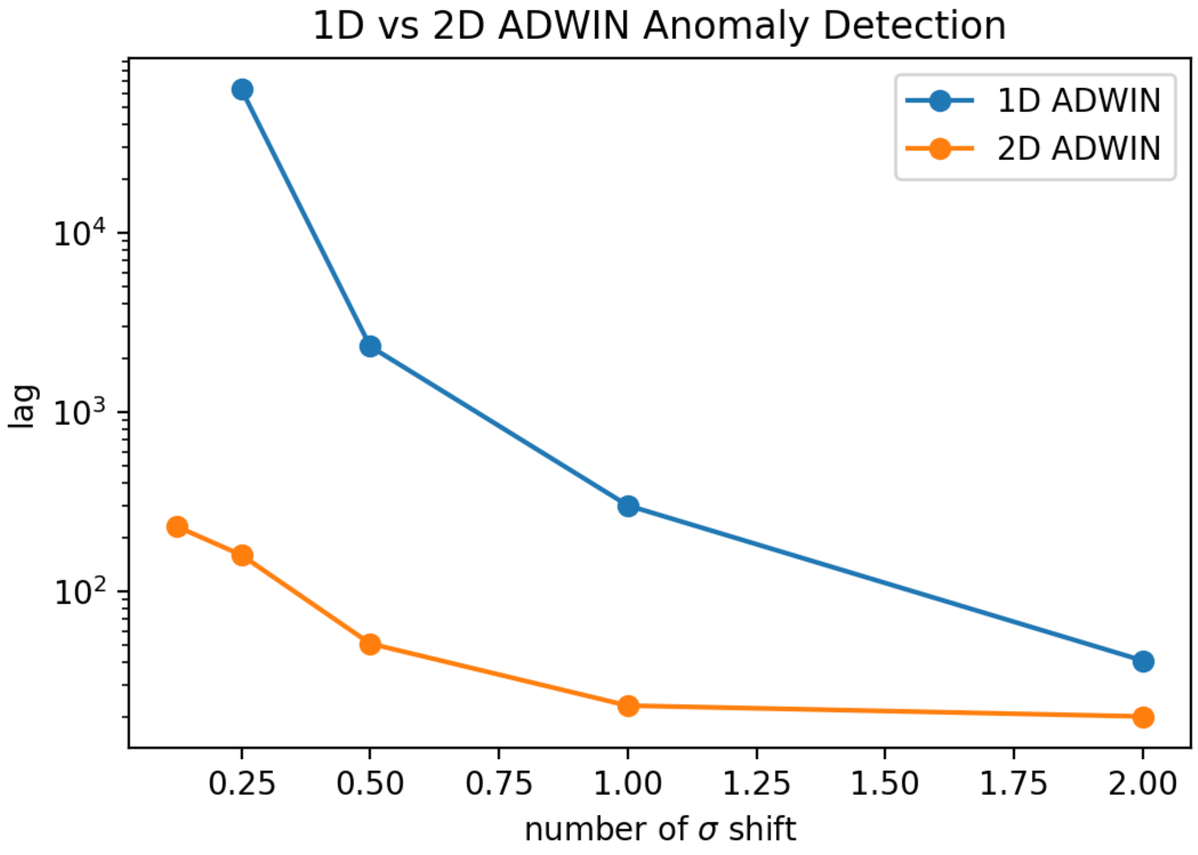


An example data stream

A higher-dimensional extension of ADWIN improves its ability to find changes in the data distribution.

Two cases:

- 1D: only use information from Q^2
- 2D: use information from (x, Q^2)

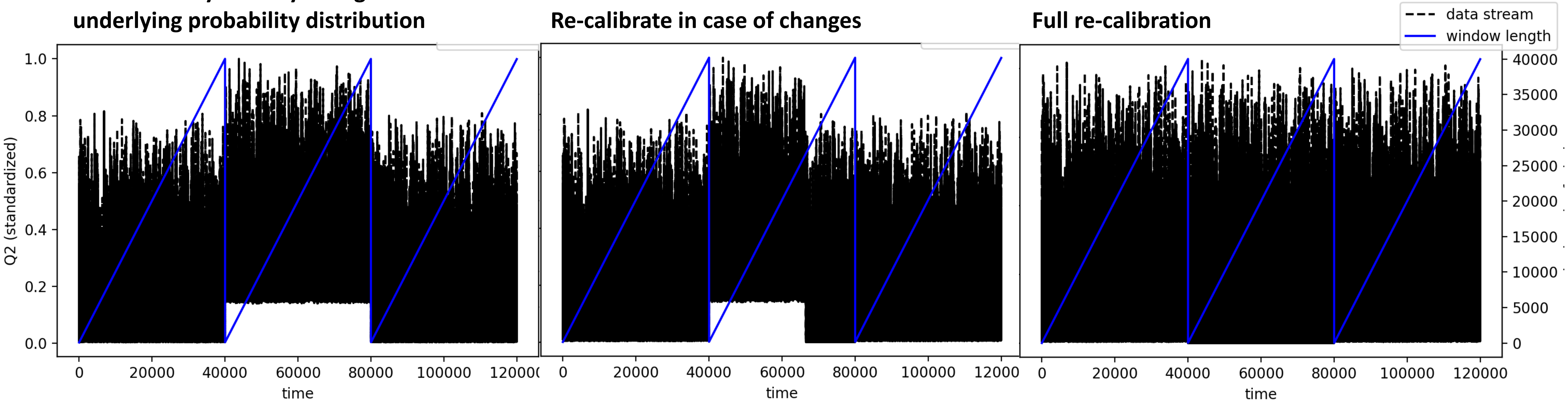


Calibrating each data-taking period to baseline period

Automatically identify changes in the underlying probability distribution

Re-calibrate in case of changes

Full re-calibration



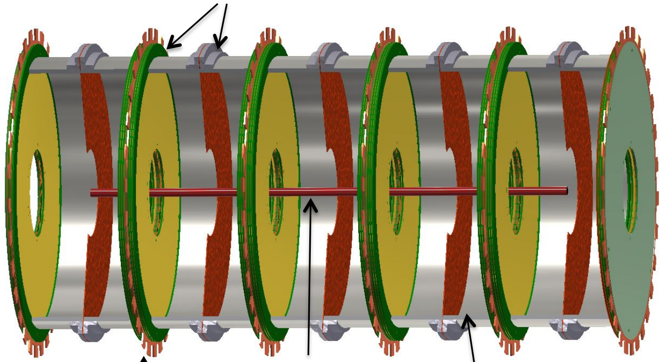
Hoeffding’s Inequality For a confidence level of 0.01 and a margin of error of 0.01, a minimum sample of 26492 observations is needed to estimate of the mean in each data-taking period.

TDIS Streaming Readout Prototype

Ed Jastrzembski et al.

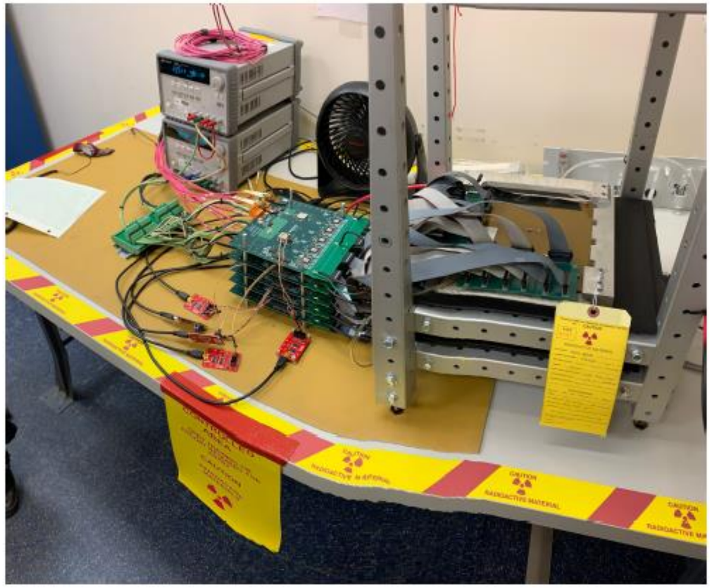
Tagged Deep Inelastic Scattering (TDIS)

- **Hall A** Super Big-Bite (SBS)
- **measurement** tagging for meson structure via the Sullivan process
- **science goals** meson structure functions and PDF
- **detection of low momentum spectators** GEM based multiple TPC (mTPC), reduced drift time in mTPC allows for triggered or streaming readout

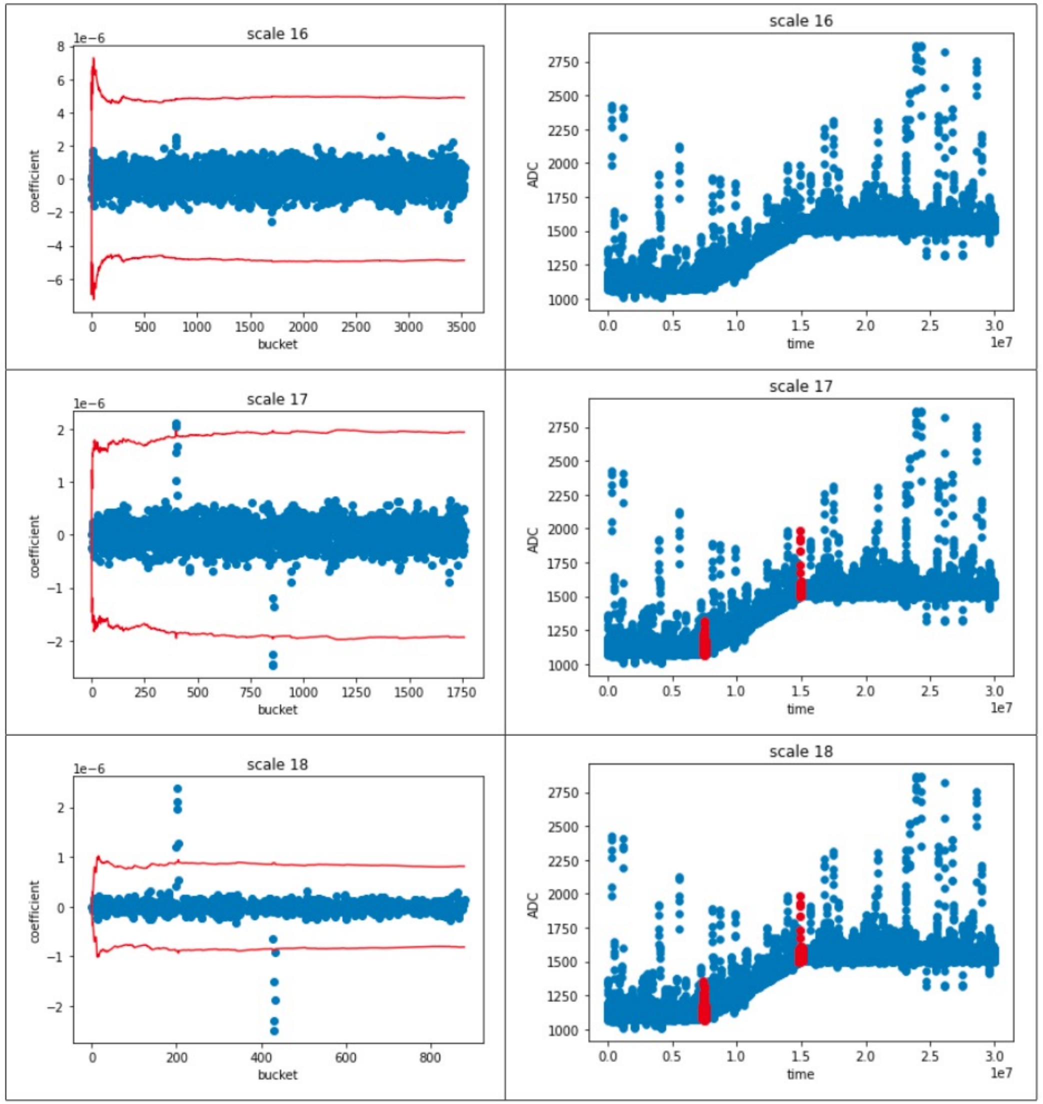


TDIS Streaming Readout Prototype

- **SAMPA** novel front-end ASIC developed for streaming readout of GEM based ALICE TPC
- **ongoing tests** study GEM pulse data and stream continuously
- **preliminary results** stream trigger-less GEM data (768 channels) in DAS and DSP modes at 45 Gb/s via 5 ALICE front-end cards (FECs)
- **next steps**
    - using FELIX hardware and software for read out GEM data
    - integrate FELIX hardware and software into CODA

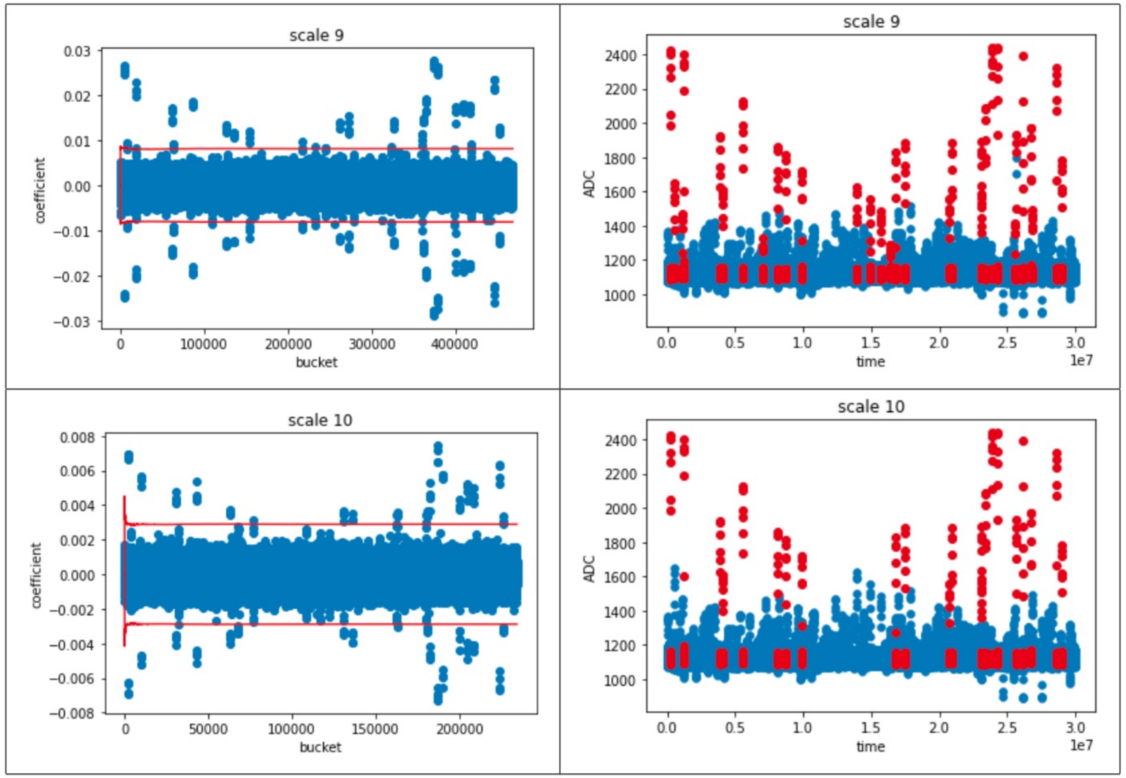


Monitoring the TDIS Streaming Readout Prototype



Multi Scale Method: Various test functions for various changes

- Represent data in multiscale basis:
  - Increase of base coefficients → Change.
- Transform to **coefficient space**:
  - Outliers in the distribution → Change.
- Detect Changes → **Detect outliers using IQR**, symbolized in **red**.

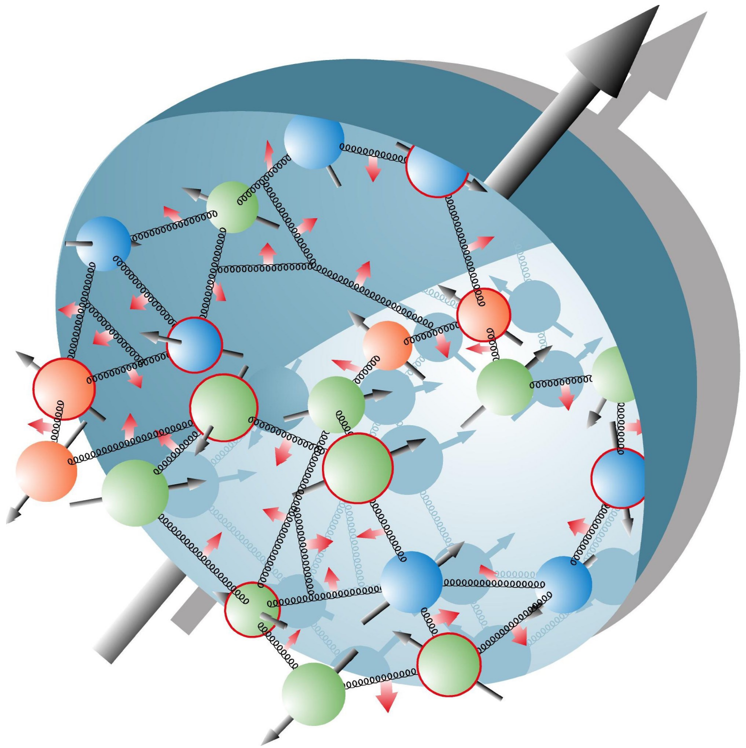


Streaming Readout

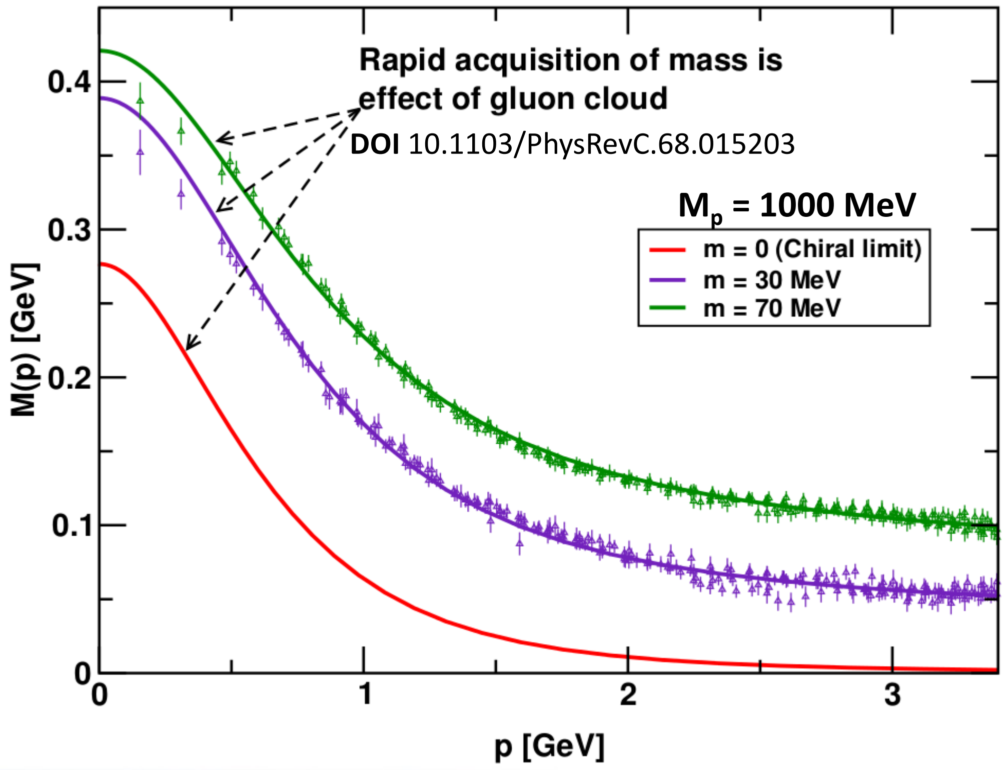
Opportunities at the Electron-Ion Collider

The dynamical nature of nuclear matter

Nuclear Matter Interactions and structures are inextricably mixed up



Observed properties such as mass and spin emerge out of the complex system



Ultimate goal Understand how matter at its most fundamental level is made

To reach goal precisely image quarks and gluons and their interactions

Advances in Nuclear Physics

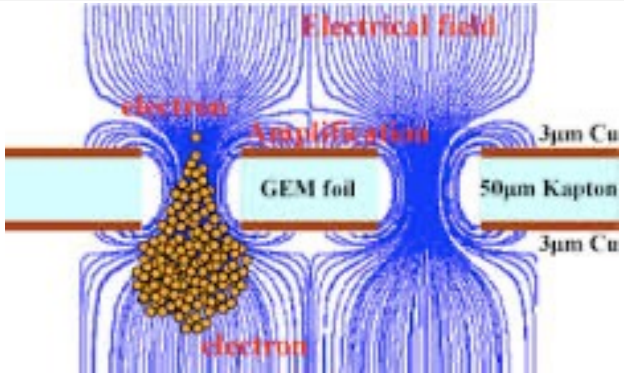
Theory of the strong interaction

$$
\begin{aligned}
\frac{\mathrm{d}\sigma}{\mathrm{d}Q^2\,\mathrm{d}y\,\mathrm{d}q_{\mathrm{T}}^2} = \frac{4\pi^2\alpha^2}{9Q^2 s} \sum_{j,j_A,j_B} e_j^2 \int \frac{\mathrm{d}^2\boldsymbol{b}_{\mathrm{T}}}{(2\pi)^2} e^{i\boldsymbol{q}_{\mathrm{T}}\cdot\boldsymbol{b}_{\mathrm{T}}} \\
\times \int_{x_A}^{1} \frac{\mathrm{d}\xi_A}{\xi_A} f_{j_A/A}(\xi_A;\mu_{b_*})\, \tilde{C}^{\mathrm{CSS1,\ DY}}_{j/j_A}\left(\frac{x_A}{\xi_A}, b_*; \mu_{b_*}^2, \mu_{b_*}, C_2, a_s(\mu_{b_*})\right) \\
\times \int_{x_B}^{1} \frac{\mathrm{d}\xi_B}{\xi_B} f_{j_B/B}(\xi_B;\mu_{b_*})\, \tilde{C}^{\mathrm{CSS1,\ DY}}_{\bar{j}/j_B}\left(\frac{x_B}{\xi_B}, b_*; \mu_{b_*}^2, \mu_{b_*}, C_2, a_s(\mu_{b_*})\right) \\
\times \exp\left\{ -\int_{\mu_{b_*}^2}^{\mu_Q^2} \frac{\mathrm{d}\mu'^2}{\mu'^2} \left[A_{\mathrm{CSS1}}(a_s(\mu');C_1) \ln\left(\frac{\mu_Q^2}{\mu'^2}\right) + B_{\mathrm{CSS1,\ DY}}(a_s(\mu');C_1,C_2) \right] \right\} \\
\times \exp\left[-g^{\mathrm{CSS1}}_{j/A}(x_A, b_{\mathrm{T}}; b_{\max}) - g^{\mathrm{CSS1}}_{\bar{j}/B}(x_B, b_{\mathrm{T}}; b_{\max}) - g^{\mathrm{CSS1}}_{K}(b_{\mathrm{T}}; b_{\max}) \ln(Q^2/Q_0^2) \right] \\
+ \text{suppressed corrections.}
\end{aligned}
$$

Quantumchromo-
dynamics (QCD)

Accelerator technologies

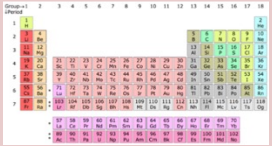
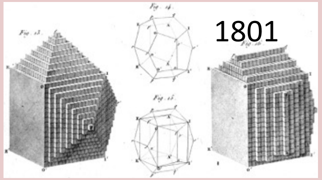
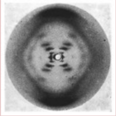
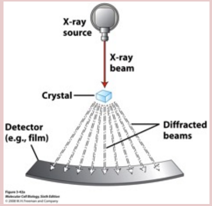
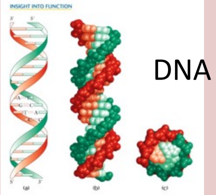
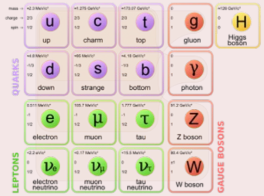
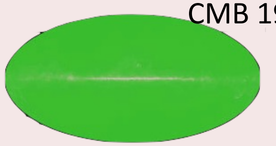
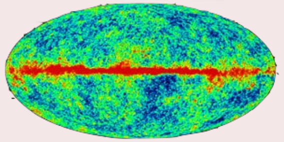
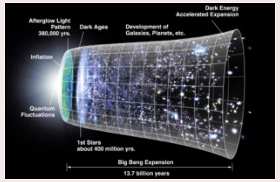
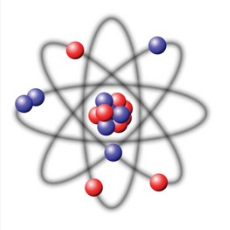
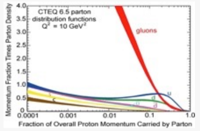
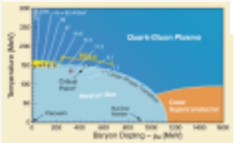
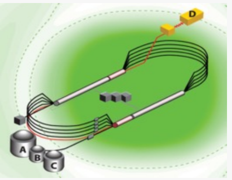
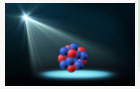
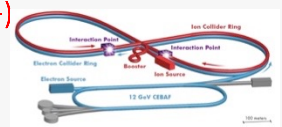
Detector technologies



Computer technologies

Steady advances in all of these areas mean that →

EIC: A new frontier in science

Dynamical System	Fundamental Knowns	Unknowns	Breakthrough Structure Probes (Date)	New Sciences, New Frontiers
Solids 	Electromagnetism Atoms 	Structure 1801 	X-ray Diffraction (~1920)  	Solid state physics Molecular biology DNA 
Universe 	General Relativity Standard Model 	Quantum Gravity, Dark matter, Dark energy. Structure CMB 1965 	Large Scale Surveys CMB Probes (~2000)  	Precision Observational Cosmology 
Nuclei and Nucleons 	Perturbative QCD Quarks and Gluons $\mathcal{L}_{QCD} = \overline{\psi}(i\not\partial - g\not A)\psi - \frac{1}{2}\mathrm{tr}\, F_{\mu\nu}F^{\mu\nu}$	Non-perturbative QCD Structure 2017   	CEBAF12 (2018)   Electron-Ion Collider (2025+) 	Structure & Dynamics in QCD 

Our Vision for Software & Computing at the EIC

Rapid turnaround of data for the physics analysis and to start the work on publications:

- **Goal**: Analysis-ready data from the DAQ system.
- **Compute-detector integration** with AI at the DAQ and analysis level.

https://eic.github.io/activities/principles.html

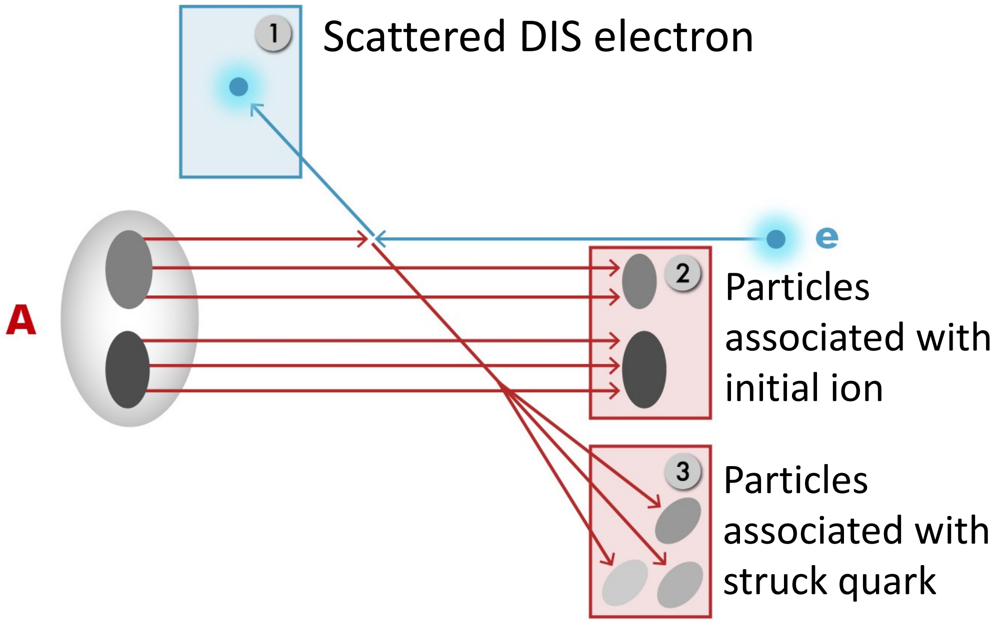
Principle 2. We will have an unprecedented compute-detector integration:

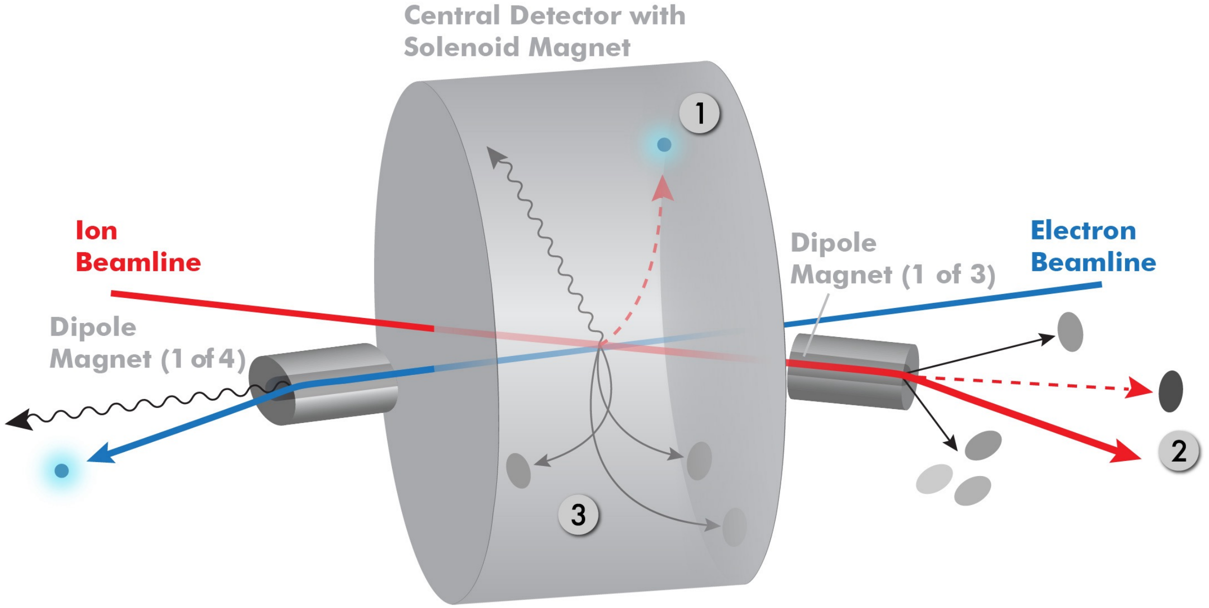
- We will have a common software stack for online and offline software, including the processing of streamed data and its time-ordered structure.
- We aim for autonomous alignment and calibration.
- We aim for a rapid, near-real-time turnaround of the raw data to online and offline productions.

Machine-Detector Interface

Integrated interaction region and detector design to optimize physics reach

The aim is to get **~100% acceptance** for all final state particles, and measure them with good resolution.





Experimental challenges:

- Beam elements limit forward acceptance.
- Central Solenoid not effective for forward.

Possible to get ~100% acceptance for the whole event:

- Beam crossing angle creates room for forward dipoles.
- Dipoles analyze the forward particles and create space for detectors in the forward ion and electron direction.

Extend our Vision beyond Machine-Detector Interface

Integration of DAQ, analysis and theory to optimize physics reach

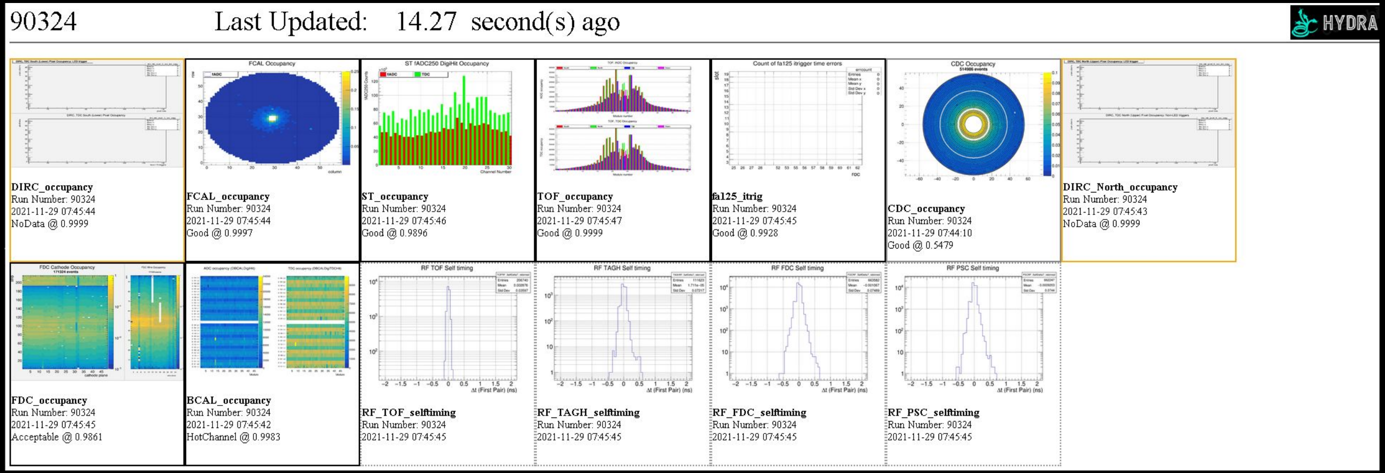
Integration of DAQ, analysis and theory

- Research model with seamless data processing from DAQ to data analysis:
  - Not about building the best detector,
  - But the best detector that fully supports streaming readout and fast algorithms for alignment, calibration, and reconstruction in near real time.
  - For rapid turnaround of data for the physics analysis and to start the work on publications.

Online Monitoring Tasks: Hydra

T. Britton, D. Lawrence, K. Rajput,
arXiv:2105.07948v1 [cs.CY]

- Take off-the-shelf ML technologies and deploy in near real-time monitoring tasks for GlueX in Hall D.
- It was the online monitoring coordinator's job to sift through hundreds of images produced in the previous 24 hours, looking for missed anomalies. This "human-in-the-loop" method was prone to errors.
- Hydra was created to tackle these challenges. Hydra is an AI system that leverages Google's Inception v3 for image classification.



It uses for training the collection of monitoring plots that GlueX had previously recorded.

A webpage was created to label the collected images and the entire system is driven by a database.

Hydra is able to spot problems missed by humans and has been shown to perform better than humans at diagnosing problems.

- Large network, ~70% of processing time spent on inference. Techniques are being tested to make Hydra models interpretable (e.g., Layerwise Relevance Propagation). Plans to deploy Hydra in other experimental halls.

7

See M. ito and D. Lawrence talks

Autonomous Calibrations

Autonomous Control and Experimentation

See M. Diefenthaler's talk

INDRA ASTRA

Approach:

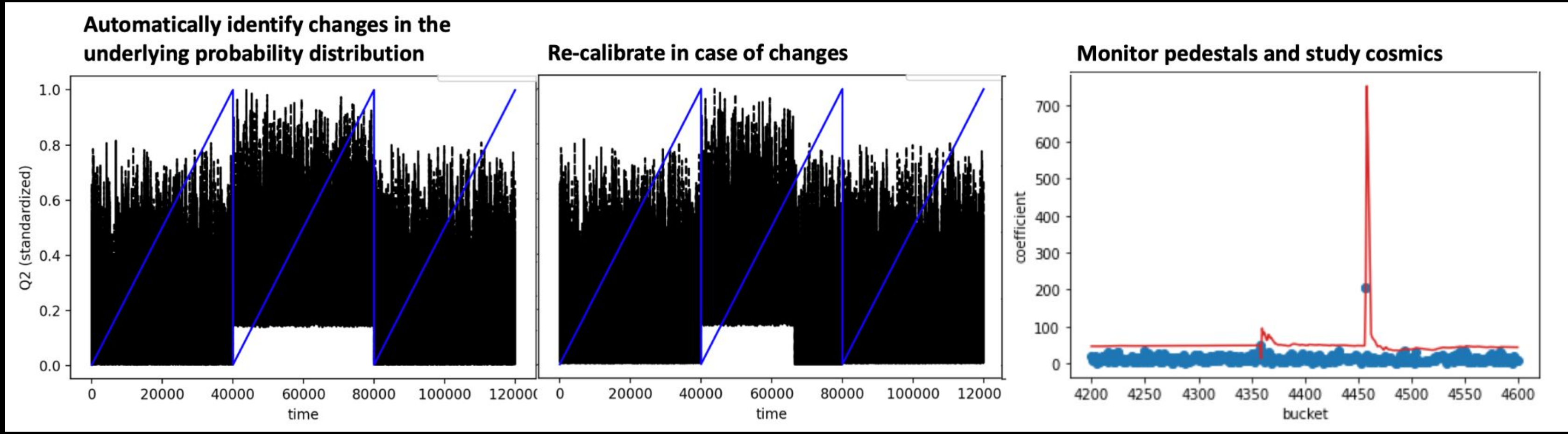
1. **Identify different data-taking periods** Use ML for a) online change detection and b) online data-quality monitoring
2. **Calibrate different data-taking periods to a baseline**

Learning how constant the data is within online adjustable thresholds

Developed **Multi Scale Method**:

- Represent data in multiscale basis: Increase of base coefficients → Change.
- Transform to coefficient space: Outliers in the distribution → Change.
- Detect Changes → Detect outliers using IQR

ADWIN2 algorithm

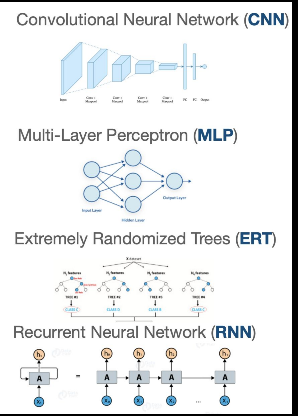


9

AI-based Tracking

G. Gavalian, et al. *arXiv preprint arXiv:2008.12860* (2020).

G. Gavalian. *arXiv preprint arXiv:2009.05144*(2020).

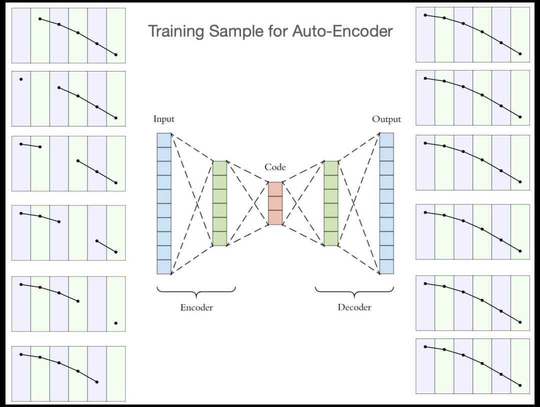


Different Network types were evaluated for accuracy and speed. MLP is chosen to be the best fit, due to implementation simplicity, accuracy and inference speed

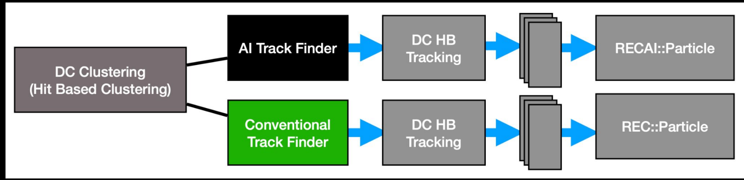
	Features	TP	FP	PA	TA	Time (ms)
ERT	6	100%	6.14%	100%	100%	0.36
MLP	6	99.96%	10.77%	98.88%	99.65%	0.12
CNN	36x112	96.11%	28.11%	94.26%	94.26%	1.2
RNN	36	88.40%	11.60%	-	-	-

TP - True Positive
FP - False Positive
TA - Training Accuracy
PA - Positive Accuracy

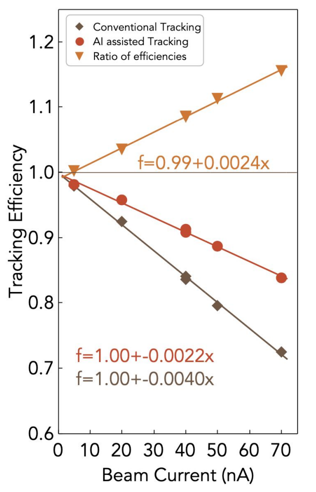
Autoencoders are typically used for de-noising, but can be used for fixing glitches



AI track classification and segment recovery network was implemented as a CLARA service. Tracking code was modified to separate clustering from track finding.



See N. Baltzell talk



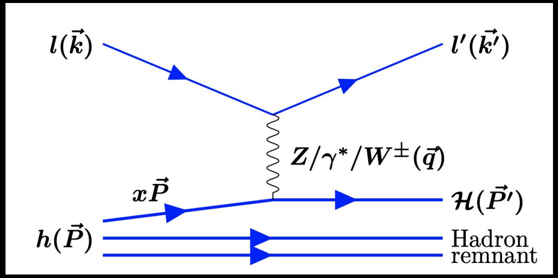
- The implementation of AI assisted tracking into the CLAS12 reconstruction workflow and provided a 6 times code speedup.
- Implemented neural network was able to reliably reconstruct missing segment positions with accuracy of ≈0.35 wires, and lead to recovery of missing tracks with accuracy of >99.8%.

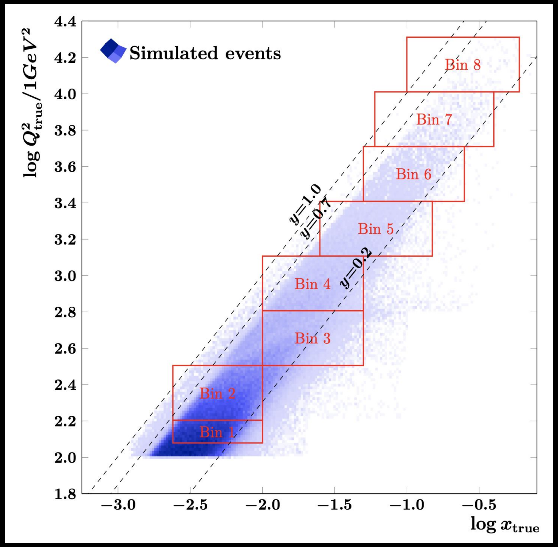
5

Reconstruction of DIS event

Deeply Learning Deep Inelastic Scattering

M. Diefenthaler, et al. "Deeply Learning Deep Inelastic Scattering Kinematics." *arXiv:2108.11638*(2021).





Training set

- Use of DNN to reconstruct the kinematic observables Q^2 and x in the study of neutral current DIS events at the ZEUS experiment at HERA.
- The performance of DNN-based reconstruction of DIS kinematics is compared to the performance of the electron method, the Jacquet-Blondel method, and the double-angle methods using data-sets independent from those used for the training
- Compared to the classical reconstruction methods, the DNN-based approach enables significant improvements in the resolution of Q^2 and x
- DIS measurements at upcoming EIC

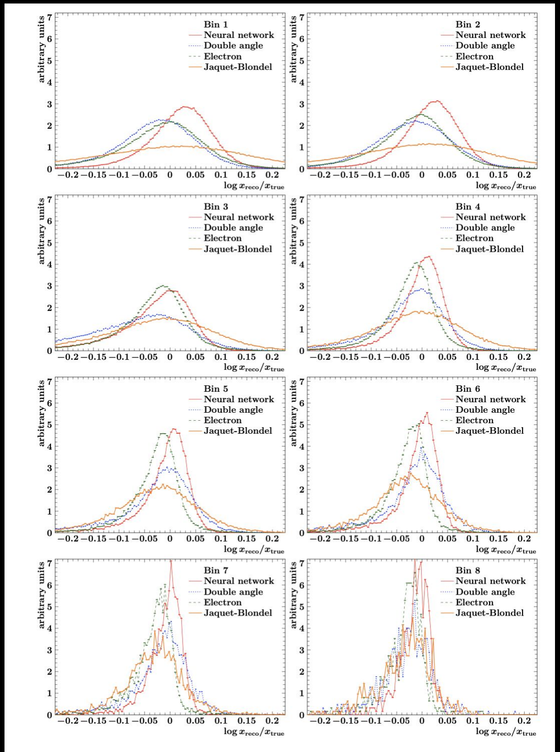
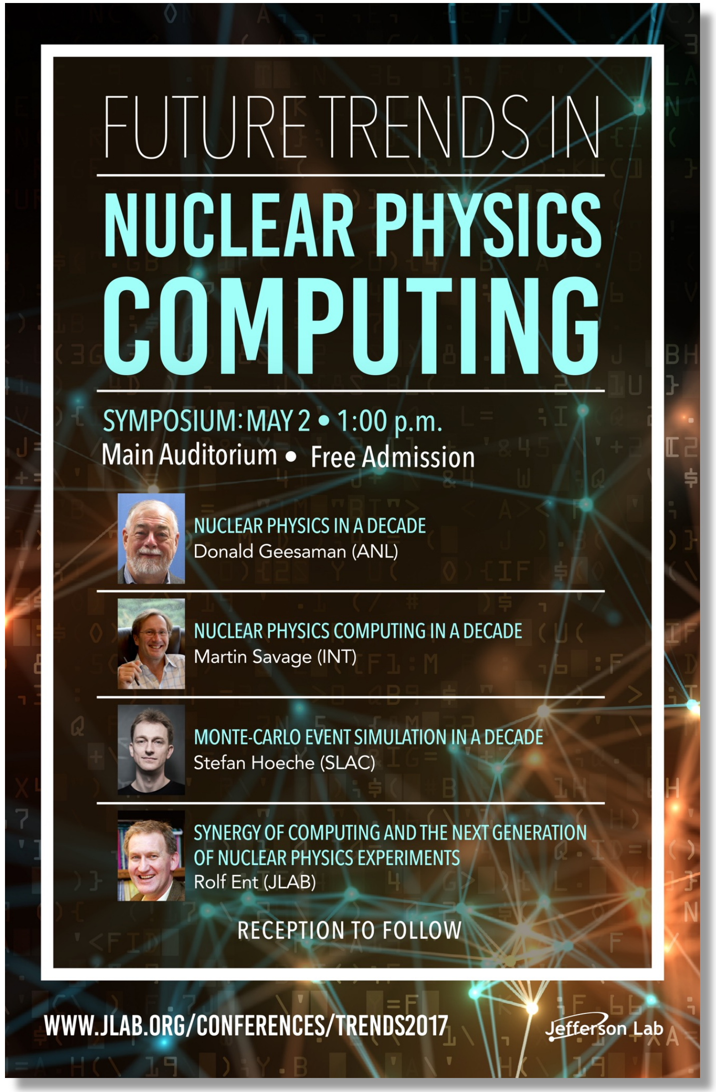


Figure 6: *Distributions of $\log x - \log x_{\mathrm{true}}$ for different reconstruction methods in individual analysis bins. For a better visibility, the centers of bins in each distribution connected with straight lines.*

19

Future Trends in Nuclear Physics Computing



Donald Geesaman (ANL, former NSAC Chair) *"It will be **joint progress of theory and experiment** that moves us forward, not in one side alone"*

One path: Sharing event-level data early, comparing experiment and theory at the event level

Martin Savage (INT) *"The next decade will be looked back upon as a **truly astonishing period in Nuclear Physics** and in our understanding of fundamental aspects of nature. This will be **made possible by advances in scientific computing** and in how the Nuclear Physics community organizes and collaborates, and how DOE and NSF supports this, to take full advantage of these advances."*

How the NP community organizes and collaborates: The AIWG / AI4EIC community is a vital part of that. AI/ML is a tremendous opportunity and we can make a difference.

Markus Diefenthaler

mdiefent@jlab.org

Over the last decade remarkable advances in electronics, computing, and software changed assumptions.

New possibilities and paradigms

- Streaming readout and AI/ML for rapid turnaround of data and starting the work on publications.

AI/ML for streaming readout

- Autonomous control and experimentation
  - Autonomous alignment and calibration in near real time
  - Autonomous anomaly detection in near real time
  - Reconstruction in near real time
  - Physics analysis in near real time
- Self-conscious detectors