INDRA-ASTRA: Evaluation & Development of Algorithms & Techniques for Streaming Detector Readout

Hindu mythology

INDRA Deity of lightning, thunder, rains and river flows

INDRA-ASTRA Indra's weapon

Jefferson Lab

INDRA Facility for **I**nnovations in **N**uclear **D**ata **R**eadout and **A**nalysis

INDRA-ASTRA on streaming readout

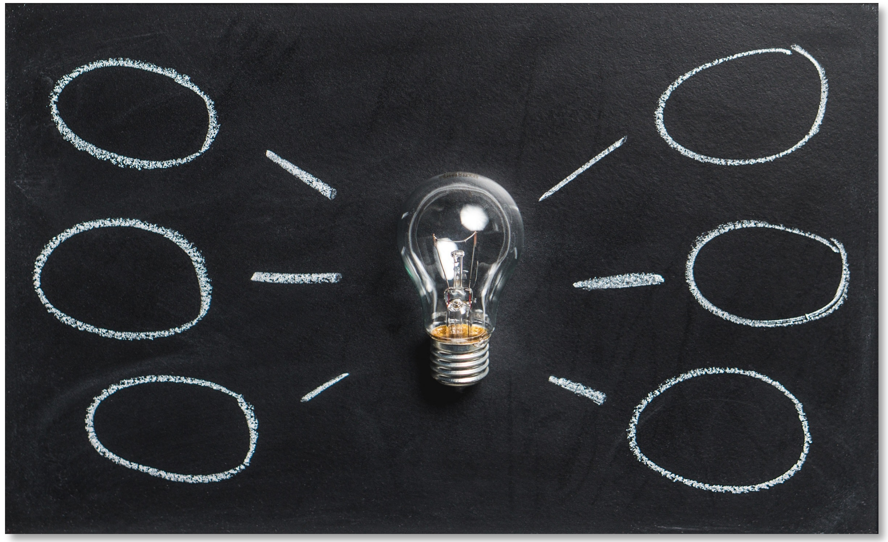
Abdullah Farhat

M. Diefenthaler (JLab), R. Fang (ODU), Y. Xu (ODU)

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



Role of computing Data processing from 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

Streaming readout and its opportunities

Definition of streaming readout

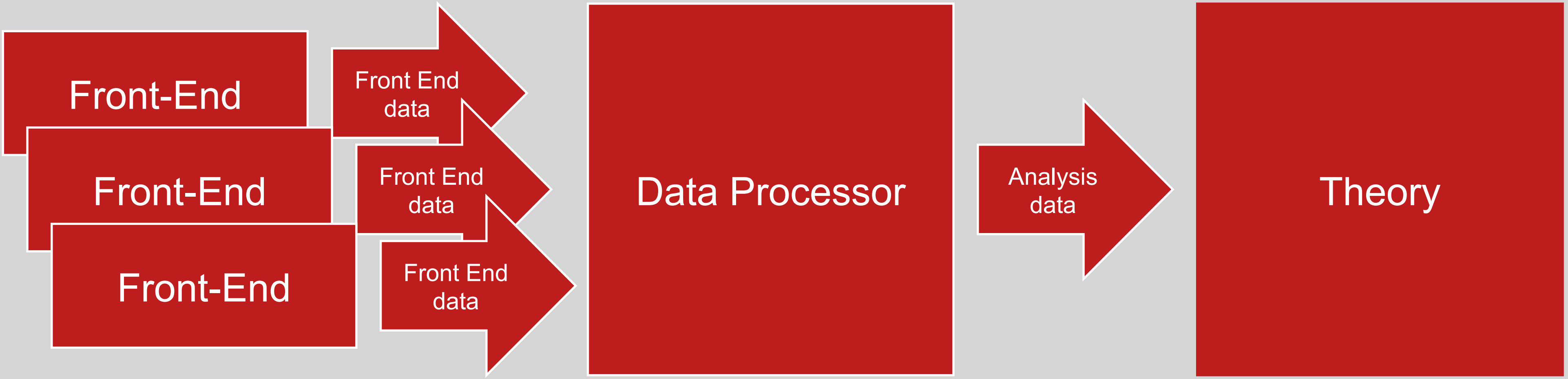
- data is read out in continuous parallel streams that are encoded with information about when and where the data was taken.

Advantages of streaming readout

- opportunity to streamline workflows
- take advantage of other emerging technologies, e.g. AI / ML

Integration of DAQ, analysis and theory to optimize physics reach

- seamless data processing from DAQ to analysis using streaming readout



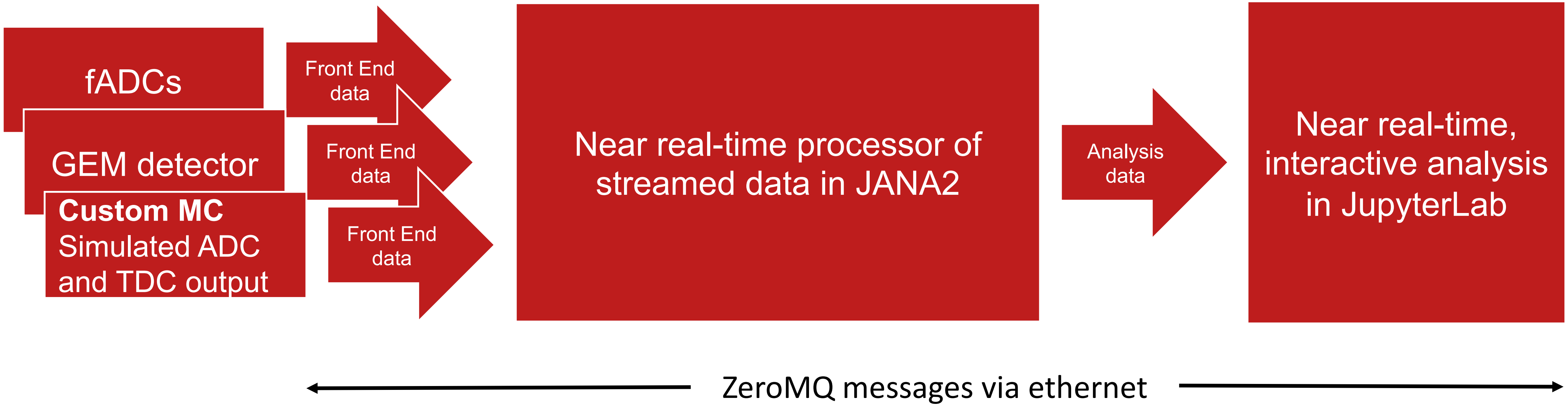
- opportunity for near real-time analysis using AI / ML (alignment, **calibration**, reconstruction)
- opportunity to accelerate science (significantly faster access to physics results)

Seamless integration of DAQ and analysis using AI/ML

GOAL

prototype components of streaming readout at NP experiments
→ integrated start to end system from detector read out through analysis
→ comprehensive view: no problems pushed into the interfaces

prototype near real-time analysis of NP data
→ inform design of new NP experiments



Streaming readout tests

fADCs

GEM detector

Custom MC
Simulated ADC and TDC output

Front End data

Front End data

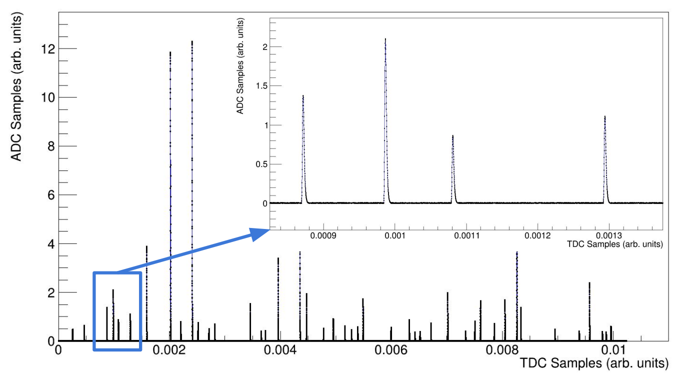
Front End data

Near real-time processor of streamed data in JANA2

Analysis data

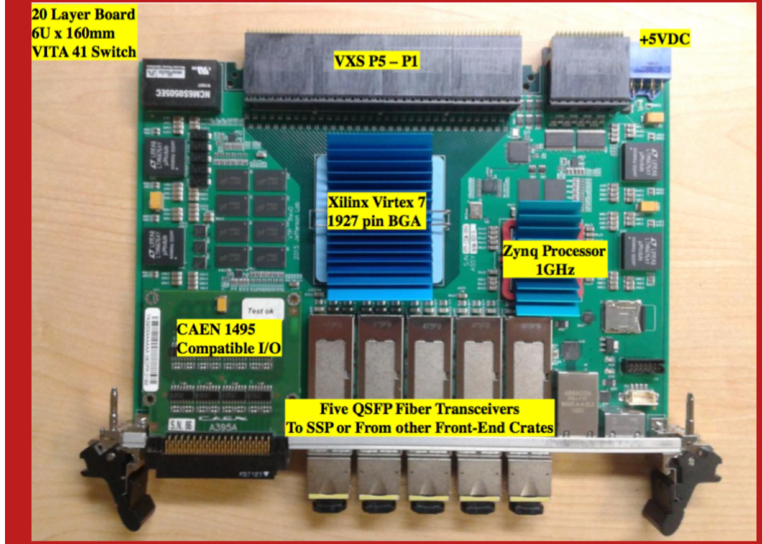
Near real-time, interactive analysis in JupyterLab

Developed streaming readout simulations

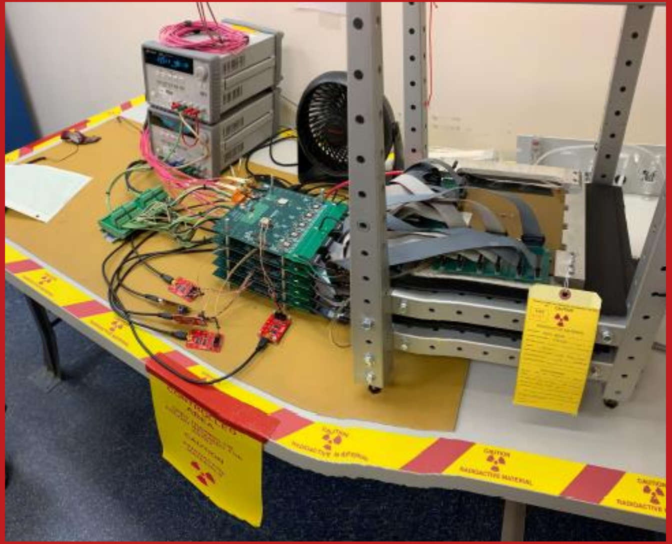


Demonstrated how to integrate any MCEG into streaming readout

Streaming readout of fADC250



TDIS Streaming Readout Prototype



Jefferson Lab

Streaming readout software

fADCs

GEM detector

Custom MC
Simulated ADC and TDC output

Front End data

Front End data

Front End data

Near real-time processor of streamed data in JANA2

Analysis data

Near real-time, interactive analysis in JupyterLab

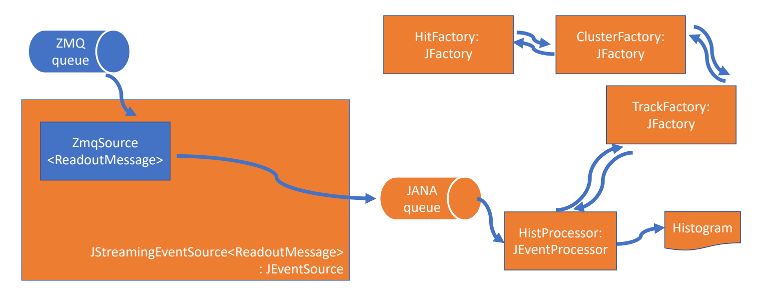
ZeroMQ messages via ethernet

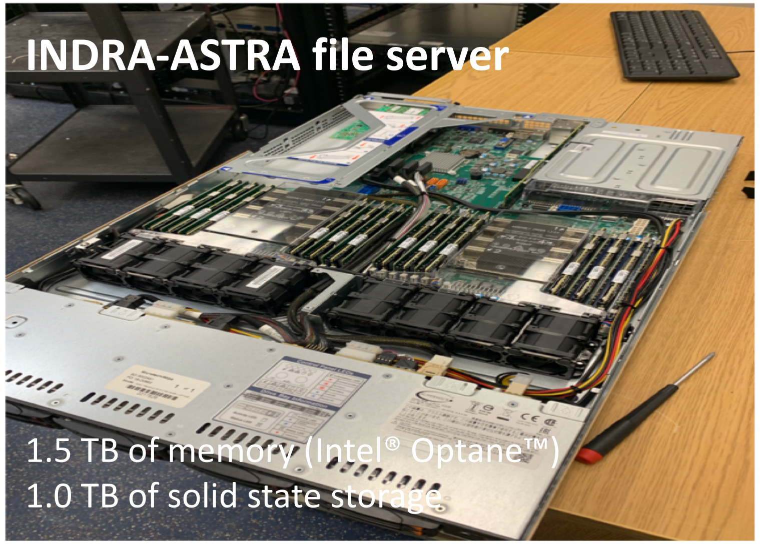
Developed messaging library

- receiving data from streaming data source
- sending it efficiently over the network (TCP)
- re-broadcasting the data using the ZeroMQ messaging protocol
- subscribing to a stream of data

tested at rates up to ~50 Gbit/s

JANA2 for parallel processing of streamed data





Streaming readout analysis

fADCs

GEM detector

Custom MC
Simulated ADC and TDC output

Front End data

Front End data

Front End data

Near real-time processor of streamed data in JANA2

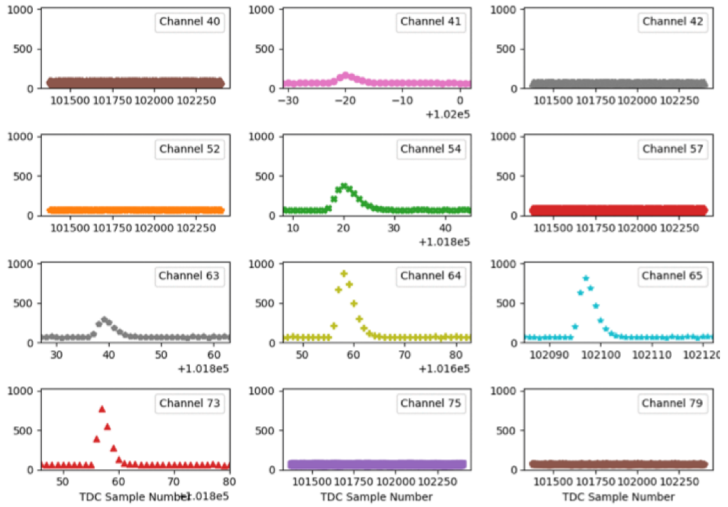
Analysis data

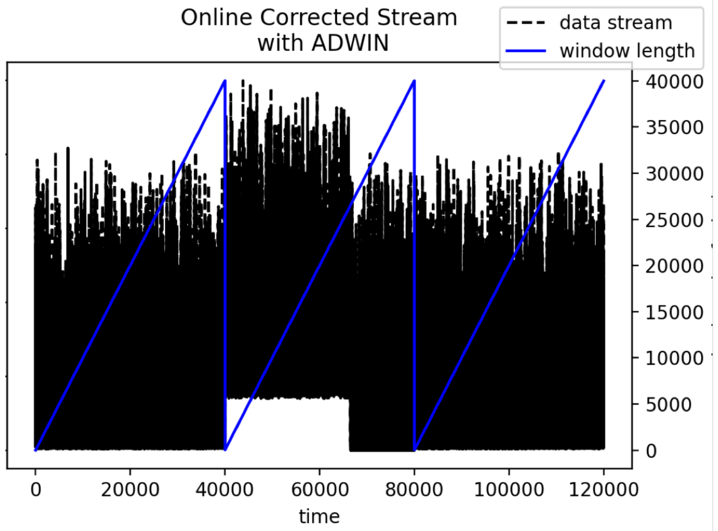
Near real-time, interactive analysis in JupyterLab

ZeroMQ messages via ethernet

Streaming plugins (data, MC)

- decoding of streamed data
- visualization of streamed data
- automated data-quality monitoring
- online calibrations
- fully extensible in JupyterLab





Jefferson Lab

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**." [1]

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

1. detecting when a change occurs
2. determining which examples to keep and which to drop
3. updating models when significant change is detected

Automated data-quality monitoring and calibrations

OUR APPROACH

1. **Identify different data-taking periods**
Use ADWIN 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.

ADWIN Algorithm

- 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 [2].

- ADWIN Inputs:
  - confidence value $\delta \in (0,1)$
  - data stream $\{x_1, x_2, \ldots, x_t, \ldots\}$ where each x_t is available at time t drawn from some distribution with expected value μ_t

- ADWIN keeps a sliding window W with the most recently read x_i

MAIN IDEA: whenever two sufficiently large subwindows of W have sufficiently different means, then it is likely the corresponding expected values are different, and the older portion of the window is dropped.

- Moreover, the window size is expected to stay large while μ_t remains constant in W, and becomes small when μ_t changes

ADWIN Algorithm

Partion W into subwindows W_0 and W_1.

Let $|W_0| = n_0$, $|W_1| = n_1$, and $|W| = n$.

Define:

$$m = \frac{1}{1/n_0 + 1/n_1}$$

$$\delta' = \frac{\delta}{n}$$

$$\epsilon_{cut} = \sqrt{\frac{1}{2m} \ln \frac{4}{\delta'}}$$

The probability for both false positive and false negative is at most δ.

ADWIN: Adaptive Windowing Algorithm

```
Initialize Window W
for each t > 0
      do W ← W ∪ {x_t} (i.e., add x_t to the head of W)
         repeat Drop elements from the tail of W
           until |μ̂_W0 − μ̂_W1| ≥ ε_cut holds
             for every split of W into W = W0 · W1
         output μ̂_W
```

$$\underbrace{\overbrace{x_i, x_{i+1}, \dots, x_{i+n_0}}^{W_0}, \overbrace{x_{i+n_0+1}, \dots, x_{i+n}}^{W_1}}_{W}$$

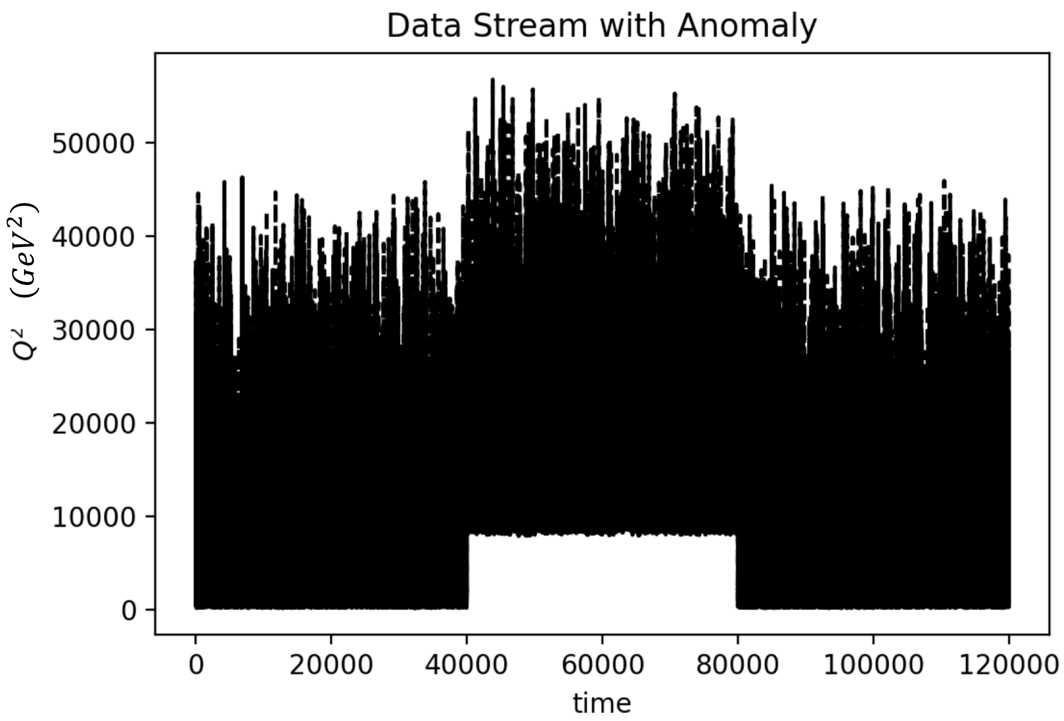
An example data stream

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.

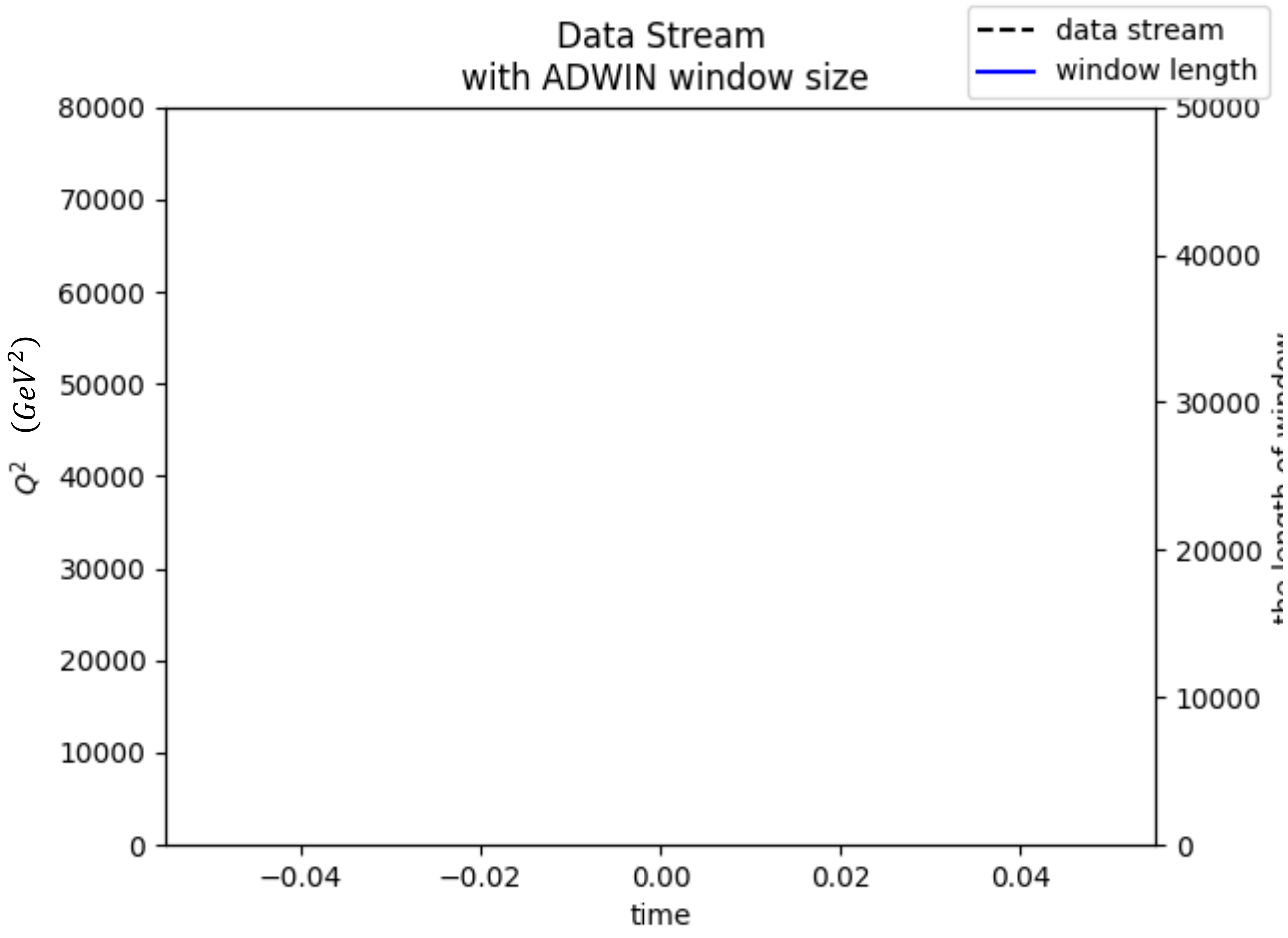
We observe a stream of x and Q^2, reconstructed by the electron method [3] based on the measurement of the (x, y, z) position and energy E of the outgoing lepton in the calorimeter.

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

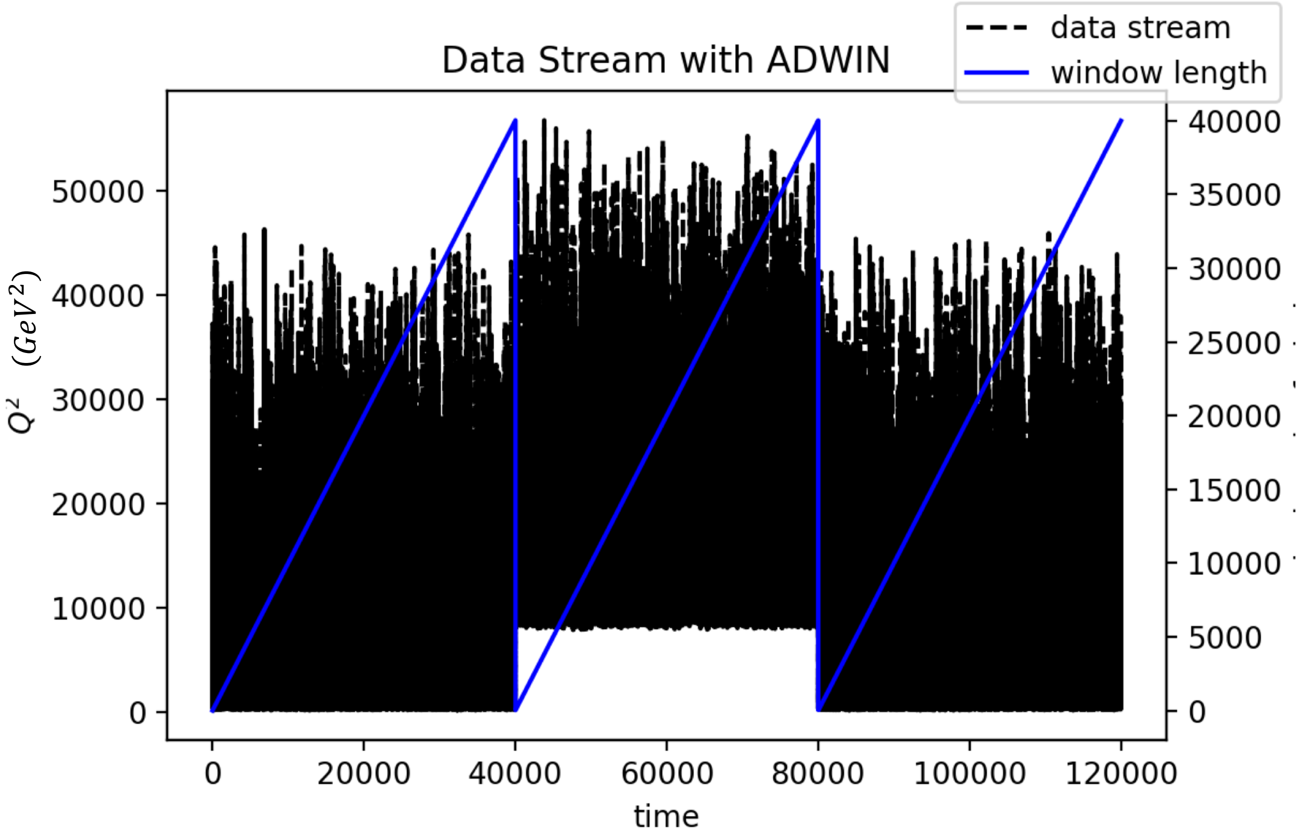
Data Period	Start Time	Time ADWIN Detects Change
2	40000	40020
3	80000	80012



Jefferson Lab

An example data stream

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

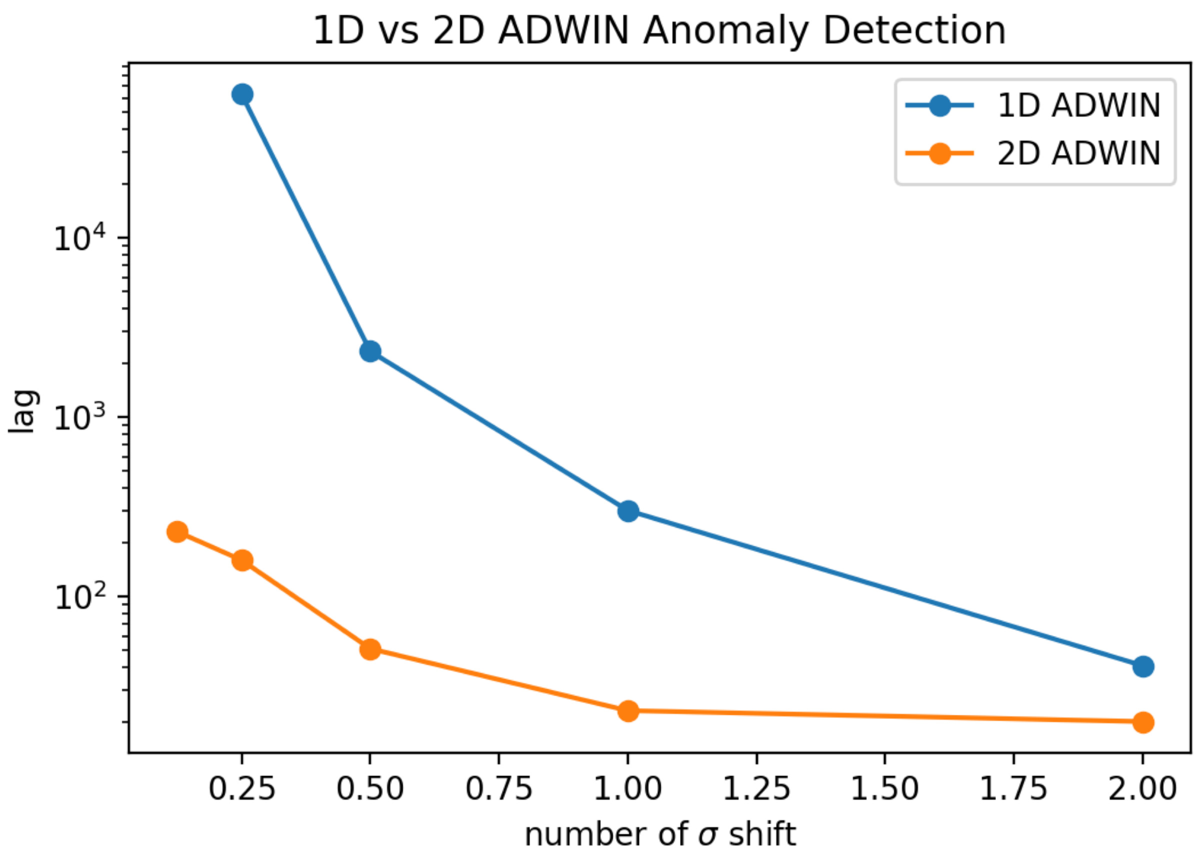


An example data stream

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)



Jefferson Lab

An example data stream

- After using ADWIN2 to detect different data-taking periods, each period is calibrated to the baseline period.

- The simple calibration we use is to shift each period by a constant value to force its mean to be equal to the baseline mean

An example data stream

Hoeffding's Inequality:
If $X_1, X_2, \dots, X_n$ are independent random variables bounded between [0,1] drawn from the same distribution with expected value μ, and define $\overline{X}$ to be the sample mean, then for any $t > 0$, $\mathbb{P}(\overline{X} - \mu > t) < e^{-2nt^2}$.

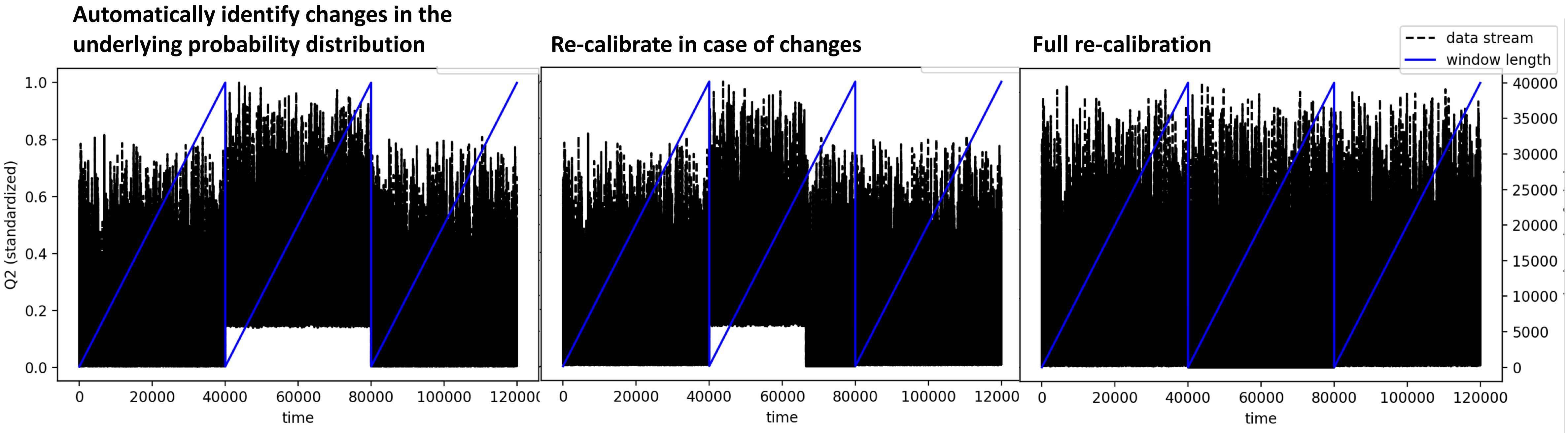
Consequently, to estimate the mean of a distribution with (1−α)%-confidence and a margin of error of t, we need at least n observations, where:

$$n = \frac{\log(2/\alpha)}{2t^2}$$

For a confidence level $\alpha = 0.01$ and a margin of error of $t = 0.01$:

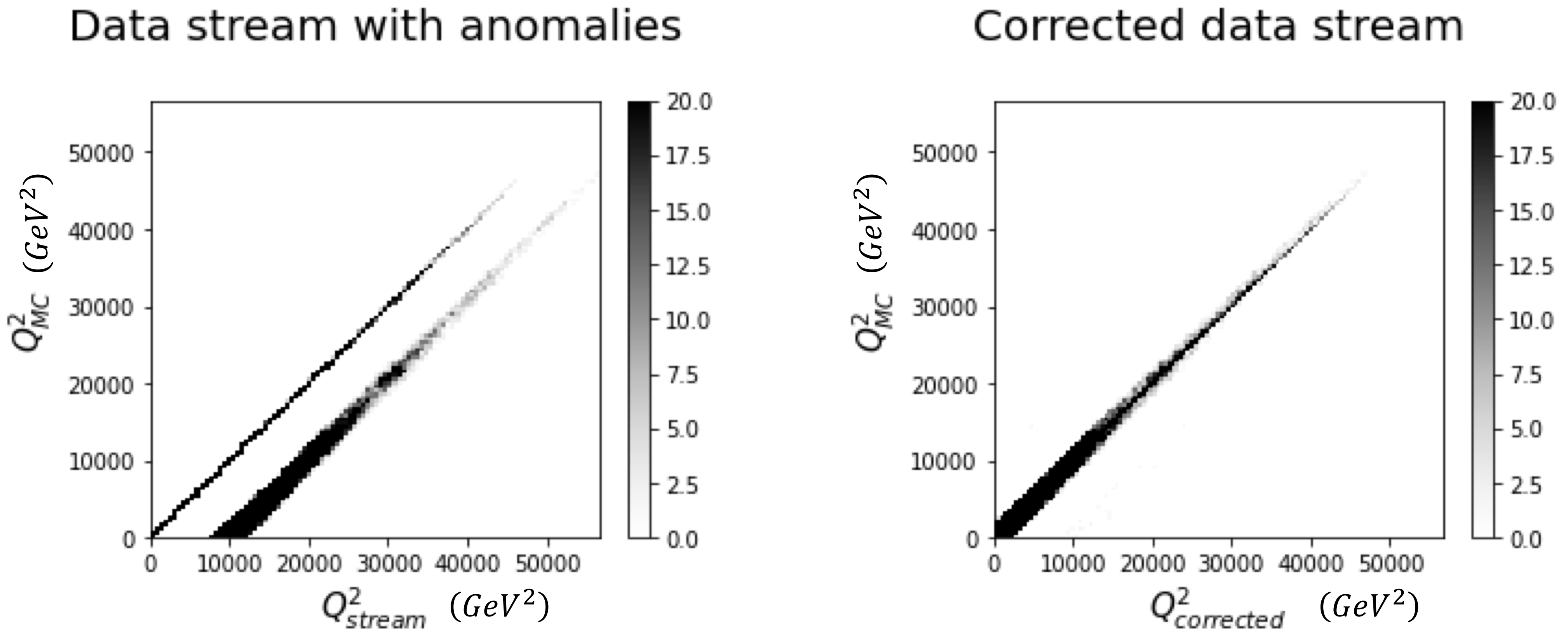
a minimum sample of 26492 observations is needed to estimate of the mean in each data-taking period.

An example data stream



Jefferson Lab

An example data stream



ADWIN Algorithm

ADVANTAGES:

- Fast algorithm
- No prior assumption on the underlying distribution of data samples
- No a priori determination of a fixed window size
- Can easily be extended to higher-dimensional anomaly detection
- Does not require any training on simulated data sets

DISADVANTAGES:

- Need to store a large window size when the data stream distribution is stable
- Uses only the mean to characterize changes

Abdullah Farhat
afarhat@odu.edu

ADWIN provides a general, automated data-quality monitoring technique that can be adapted to other data streams

- We assume that the auto-calibration method is mainly about the automated data-quality monitoring
- the calibration problem is a standard optimization problem and does not require ML. Moreover, we can re-use existing calibration methods. ADWIN provides this flexibility.

Upcoming Task: real detector tests

Ultimately, new possibilities and paradigms for NP

- seamless data processing from DAQ to analysis using streaming readout
- opportunity for near real-time analysis (auto-alignment, auto-calibration, near real-time reconstruction)
- opportunity to accelerate science

References

1. Žliobaitė I., Pechenizkiy M., Gama J. (2016) An Overview of Concept Drift Applications. In: Japkowicz N., Stefanowski J. (eds) Big Data Analysis: New Algorithms for a New Society. Studies in Big Data, vol 16. Springer, Cham. https://doi.org/10.1007/978-3-319-26989-4_4
2. 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.
3. S. Bentvelsen, J. Engelen and P. Kooijman, Reconstruction of (x, Q2) and extraction of structure functions in neutral current scattering at HERA, in Workshop on Physics at HERA Hamburg, Germany, October 29-30, 1991, 1992, pp. 23–42.