INDRA-ASTRA Seamless data processing from DAQ to data analysis

Hindu mythology

INDRA Deity of lightning, thunder, rains and river flows

INDRA-ASTRA Indra's weapon

Nuclear Physics

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

INDRA-ASTRA LDRD on streaming readout

Markus Diefenthaler

Streaming readout and its opportunities

Definition of streaming readout

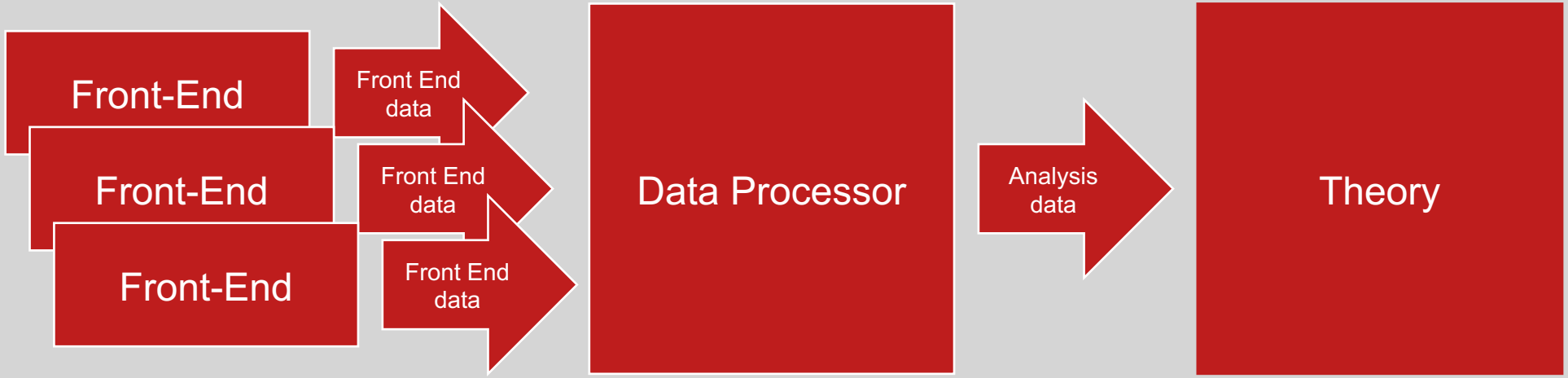
- data is read out in continuous parallel streams that are encoded with information about when and where the data was taken.

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



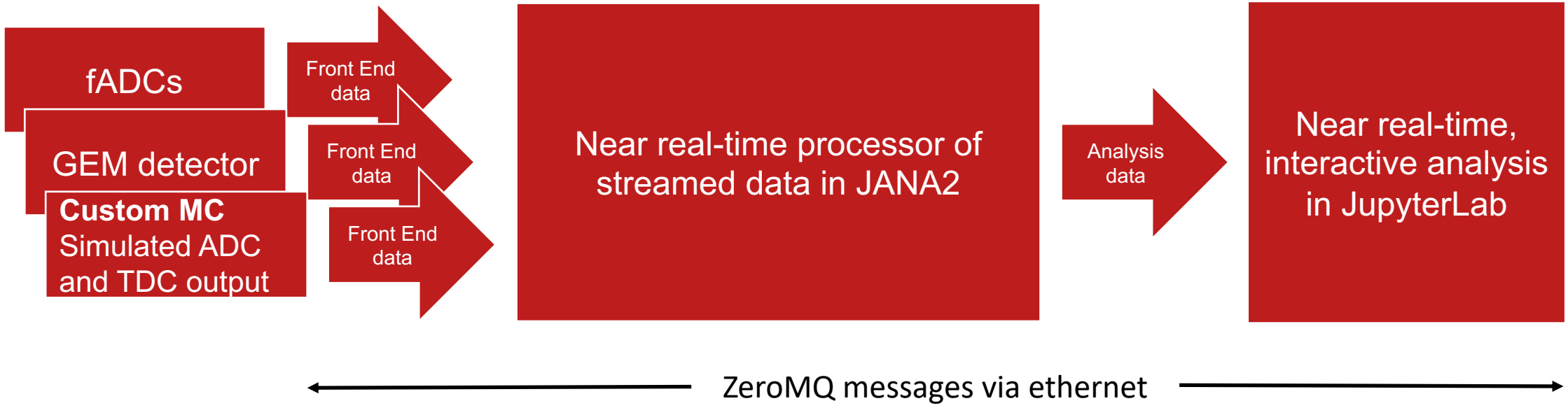
- opportunity for near real-time analysis using AI / ML
- opportunity to accelerate science (significantly faster access to physics results)

INDRA-ASTRA: 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



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 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.

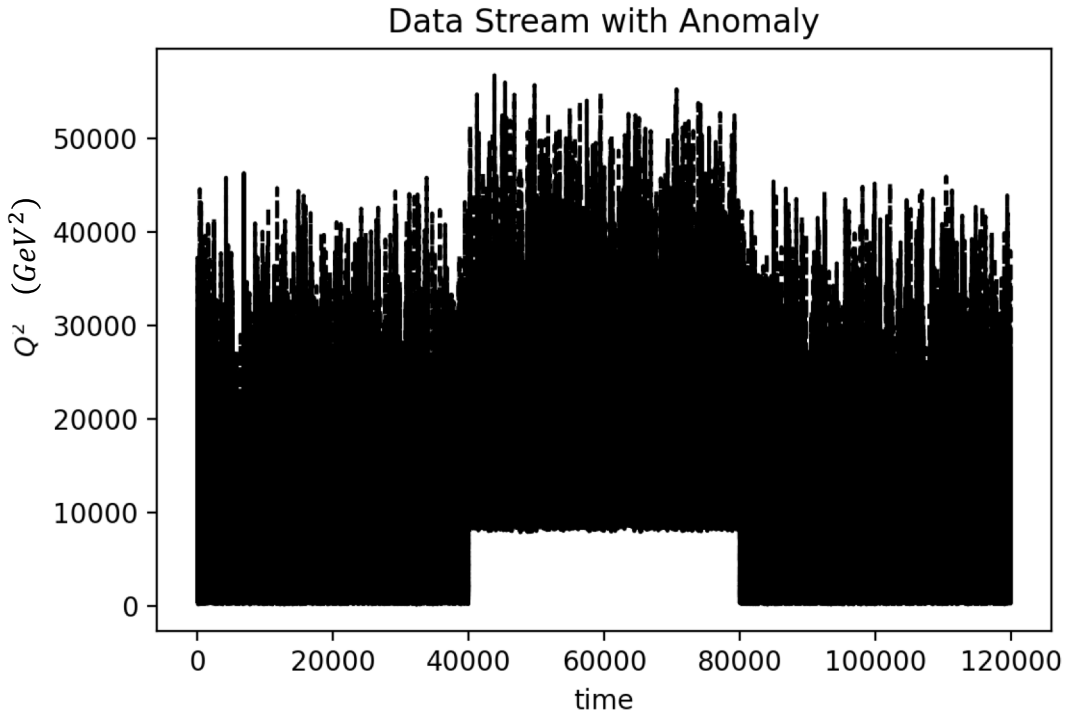
An example data stream

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.

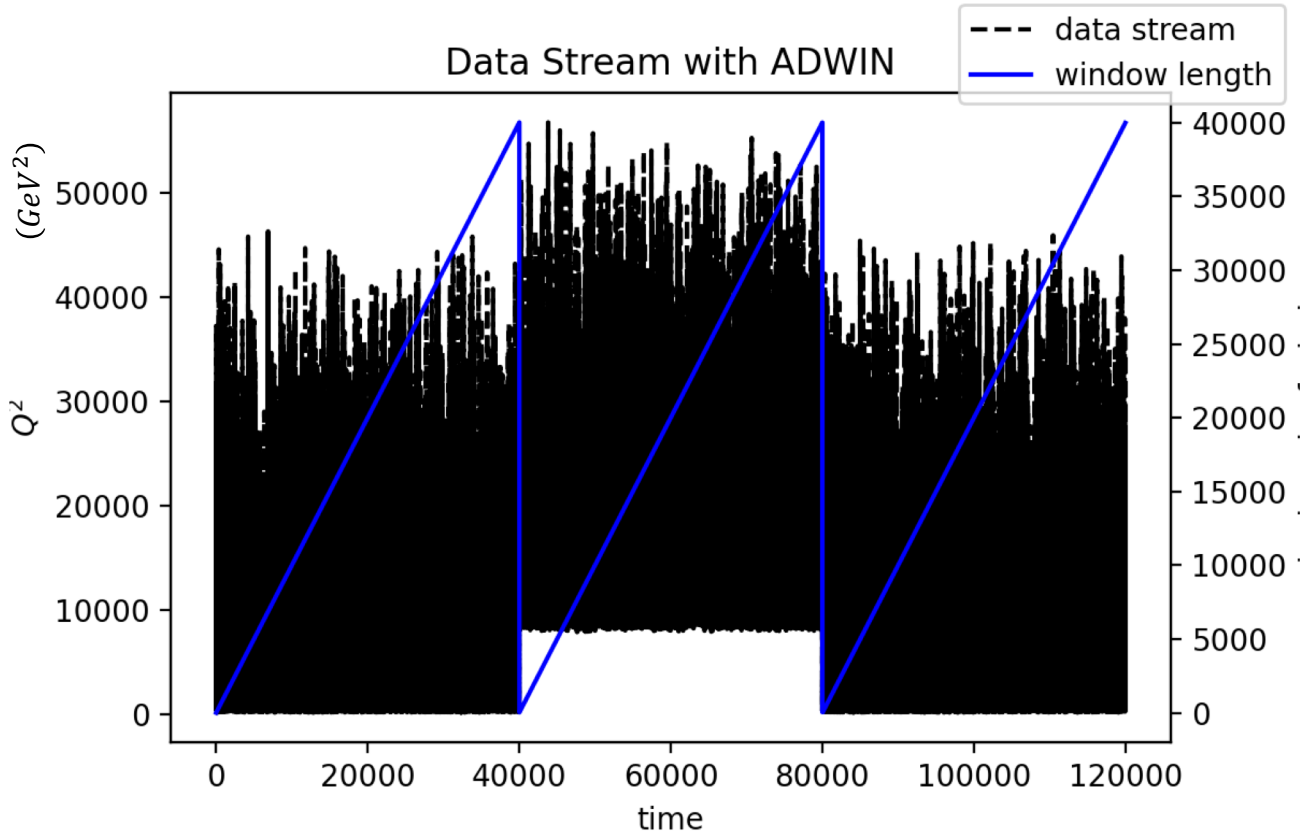


An example data stream

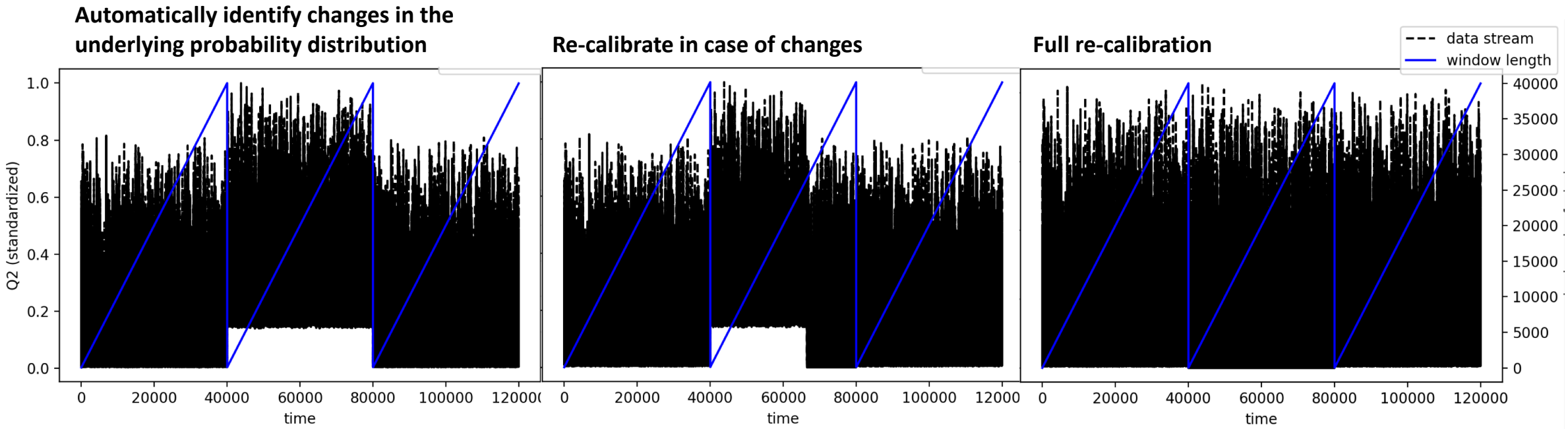
ADWIN is an **ADaptive WINdowing technique** used for detecting distribution changes, concept drift, or anomalies in data streams with established guarantees on the rates of false positives and false negatives

(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



Calibrating each data-taking period to baseline period



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.

Agenda for today's meeting

INDRA-ASTRA

Wednesday 13 Jan 2021, 10:00 → 12:00 US/Eastern

Description Streaming readout gives opportunity to streamline workflows and to take advantage of other emerging technologies, e.g., artificial intelligence (AI) or machine learning (ML). In the INDRA-ASTRA project, we have explored the possibility for automated calibrations using AI / ML which would allow a rapid turnaround from data taking to physics results.

We will use BlueJeans for the remote meeting: https://bluejeans.com/305428987

10:00 → 10:15 **INDRA-ASTRA: Introduction** 15m
Speaker: Markus Diefenthaler (Jefferson Lab)

10:15 → 10:45 **ADWIN2: Algorithm and Evaluation** 30m
Speaker: Abdullah Farhat (ODU)

10:45 → 11:15 **Beyond ADWIN2: Algorithms and Evaluation** 30m
Speaker: Ronglong Fang (ODU)

11:15 → 12:00 **Discussion: Next steps (Further Evaluation, Tests with detector data)** 45m

Summary

Markus Diefenthaler
mdiefent@jlab.org

- **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
- **Many opportunities to get involved!**