

# Software developments towards a high luminosity, medium energy 2nd interaction region at the EIC

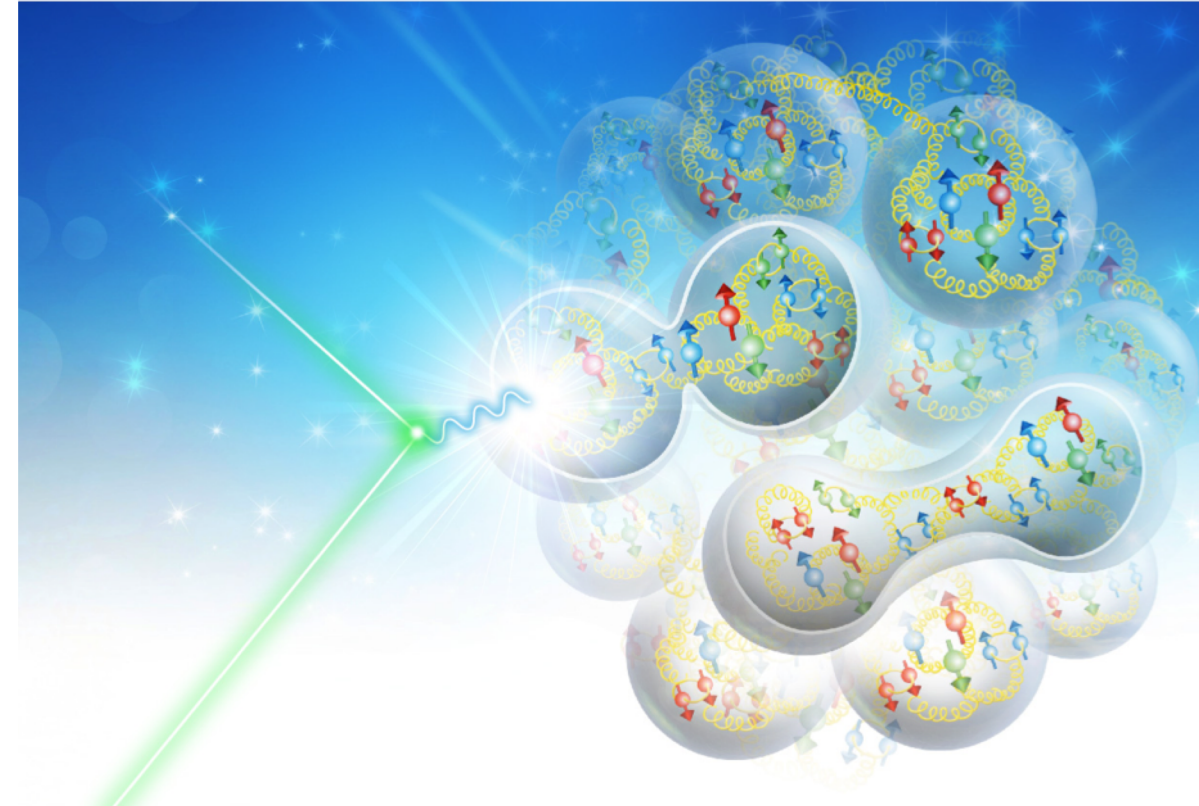
## Community activities

- Software Working Group
- Community Building

## Software priorities

- Machine-Detector-Analysis Interface
- Considerations for 2<sup>nd</sup> interaction region

Markus Diefenthaler with input from Andrea Bressan (Trieste) and Torre Wenaus (BNL)

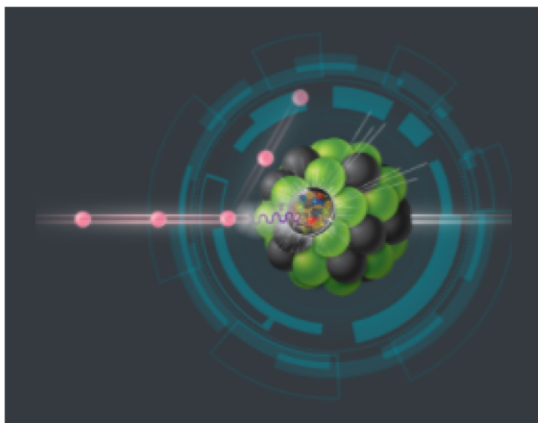
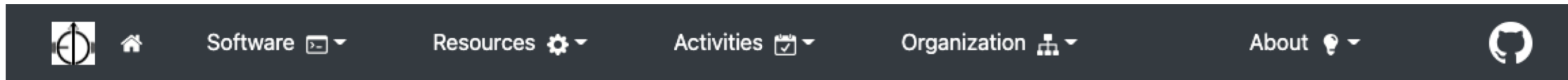


---

## **Section**

# **Activities by the Software Working Group**

# EICUG Software Working Group (<https://eic.github.io>)



## Purpose of this site

This is the main portal to the EIC software, repositories, documentation and resources. It is developed and maintained by the EIC Software Group.

**Work in progress**

Questions? Contact the EICUG Software Working Group Conveners:  
[eicug-software-conveners@eicug.org](mailto:eicug-software-conveners@eicug.org)

## News

- [Software News July](#) 2020-07-31
- [Software News June](#) 2020-06-24
- [Software News April](#) 2020-04-07

**Behind on our communication**

©2020 EICUG Software Working Group

Site built at 2020-11-30 16:23:42 -0500

# Online tutorials



EIC User Group  
60 subscribers

<https://www.youtube.com/channel/UCXc9WfDKdILXoZMGrotkf7w>

SUBSCRIBE

HOME

VIDEOS

PLAYLISTS

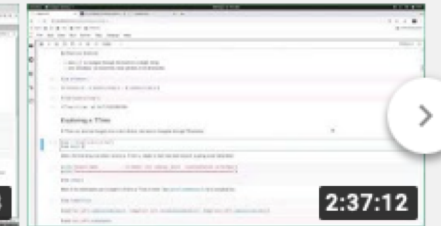
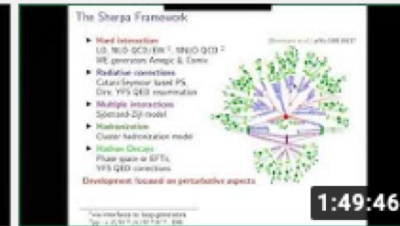
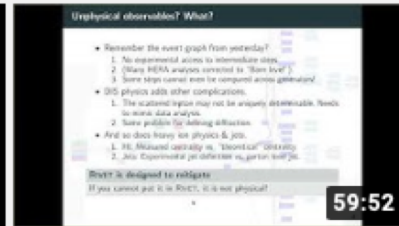
CHANNELS

DISCUSSION

ABOUT



Uploads ▶ PLAY ALL



EIC Software Tutorial: Pythia 8  
112 views • 1 month ago

EIC Software Tutorial: Herwig  
53 views • 3 months ago

EIC Software Tutorial: MC-Data Comparisons in Rivet  
53 views • 3 months ago

EIC Software Tutorial: Sherpa  
51 views • 3 months ago

EIC Software Tutorial: Advanced Fast Simulations...  
145 views • 6 months ago

Jim Pivarski Tutorial: uproot and Awkward Array  
298 views • 8 months ago

## Recordings from tutorials

- **Advanced Fast Simulation Tutorial** Fast simulations on the command line and in JupyterLab, singularity
- **Detector Full Simulation Tutorials** Geant4 for EIC, how to modify existing detector concepts, and how to integrate a new detector into one of the existing detector concepts.
- **Jim Pivarski Tutorial: uproot and Awkward Array** process and analyze Root files with pure Python libraries
- **MCEG Tutorial** Herwig, Pythia, Sherpa, Rivet, more to follow



# Validation of simulation tools

सॉफ्टवेयर Involvement



from EIC-India

**Full simulations**

**ESCalate**

IIT Indore, RKMRC  
Kolkata

**Fun4All**

IIT Indore, IIT Madras,  
Panjab

**Fast simulations**

**eic-smear**

IIT Delhi, IIT Patna,  
Karnataka, MNIT Jaipur

**MCEG validation**

Akal, DAV, Goa, IIT  
Bombay, Delhi, Indore,  
Madras, and Patna;  
MNIT Jaipur, Panjab

# Collecting, organizing, and documenting EIC Software

**new GitHub organization** for the EIC community <https://github.com/eic>

- Please help us to make your software available on the GitHub organization.



**Involvement from EICUG**

**EIC(UG) repositories for software, scripts,  
run cards, etc.: <https://github.com/eic>**

**Guidelines on <https://eic.github.io/github/>**

# Why is this important?

**SWG charge** “ (...) simulations of physics processes and detector response (...) will be pursued in a manner that is accessible, consistent, and **reproducible** to the EICUG as a whole (...)”

Having reliable access to results previously obtained and to the arsenal of the software tools being developed by EICUG will create efficiencies as EIC is approaching the detector design stage.

Any study you share with the SWG can be used to **benchmark & validate** the EIC Software tools. If you send us analysis scripts and macros, the SWG can reproduce your studies and build up a **validation scheme and tools** on top of it, validation you can use yourself. But this can only work with your strong support.

## **Data and Analysis Preservation** (DAP)

- Importance of preserving metadata and code alongside raw data
- Importance of documentation and proper choice of tools
- Importance of building DAP into the infrastructure at an early stage



**Ongoing discussion on intellectual property.**

# Motivation for Expression of Interest for Software

---

- **Precept** EIC detector collaborations will determine for themselves what they do for software, but that will include common software elements.
- In January we were discussing 'greenfield framework' as a post Yellow Reports common effort, but a more flexible and perhaps more tenable discussion would be 'greenfield components':
  - Can we define common software components/projects now that we think will be of interest to one or more collaborations later
  - and make useful progress on them to inform collaboration choices later and promote common software choices?
- **Expression of Interest** process seemed a good mechanism, with appropriate timescale, to give context and visibility to this.

# Community input for Expression of Interest

---

## Software Needs

**Requirements** What software needs for EIC Software would you like to highlight now, in a few years, and for the completion of the EIC project?

**Technologies & Techniques** What software technologies and techniques should be considered for the EIC?

## Meeting Software Needs

What resources can your group contribute?

# Expression of Interest for Software

1

## Expression of Interest (EOI) for Software

Please indicate the name of the contact person for this submission:

Conveners of the Software Working Group:

- A. Bressan, M. Diefenthaler, and T. Wenaus
- [eicug-software-conveners@eicug.org](mailto:eicug-software-conveners@eicug.org)

Please indicate all institutions collectively involved in this submission of interest:

ANL	Argonne National Laboratory	<b>29 institutions</b>
BNL	Brookhaven National Laboratory	
CEA/Irfu	IRFU at CEA /Saclay institute	
EIC-India	Akal University, Central University of Karnataka, DAV College Chandigarh, Goa University, Indian Institute of Technology Bombay, Indian Institute of Technology Delhi, Indian Institute of Technology Indore, Indian Institute of Technology Patna, Indian Institute of Technology Madras, Malaviya National Institute of Technology Jaipur, Panjab University, Ramkrishna Mission Residential College Kolkata	
IMP-CAS	Institute of Modern Physics - Chinese Academy of Sciences	
INFN	Istituto Nazionale di Fisica Nucleare	
JLab	Thomas Jefferson National Accelerator Facility	
LANL	Los Alamos National Laboratory	
LBNL and UC Berkeley	Lawrence Berkeley National Laboratory and University of California, Berkeley	
NCBJ	National Centre for Nuclear Research	
OhioU	Ohio University	
ORNL	Oak Ridge National Laboratory	
SBU	Stony Brook University	
SLAC	SLAC National Accelerator Laboratory	
SU	Shandong University	

<https://indico.bnl.gov/event/8552/contributions/43221/>

## Common Projects

- **Software Tools for Simulations and Reconstruction**
  - Monte Carlo Event Generators
  - Detector Simulations
  - Reconstruction
- **Middleware and Preservation**
  - Workflows
  - Data and Analysis Preservation
- **Interaction with the Software Tools**
  - Explore User-Centered Design
  - Discoverable Software
  - Data Model

## Future Technologies

- Artificial Intelligence
- Heterogeneous computing
- New languages and tools
- Collaborative software

**Develop into work plan for the Software Working Group**



---

# **Section**

# **Community Building**

# Future Trends in Nuclear Physics Computing



**BROOKHAVEN** & **Jefferson Lab**  
NATIONAL LABORATORY

## FUTURE TRENDS IN NUCLEAR PHYSICS COMPUTING

SEPT. 29 - OCT. 1, 2020

The workshop focuses on the Nuclear Physics Software & Computing community. We will identify what is unique about our community and we will discuss how we can strengthen common efforts and chart a path for Software & Computing in Nuclear Physics for the next ten years.

TOPICS:

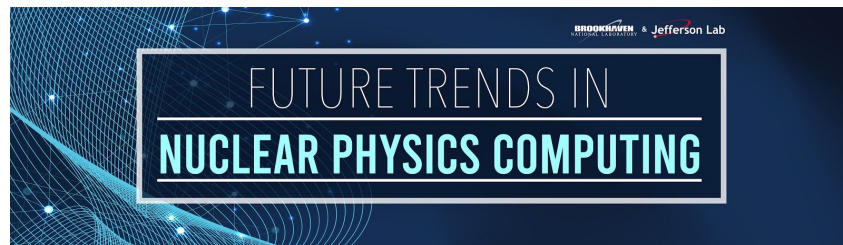
- Common Scientific Software
- The Role of Data Centers in Scientific Discovery
- Unique Software Challenges for Nuclear Physics

Focus on the **Nuclear Physics Software & Computing community**

- Identify what is unique about our community
- Discuss how we could strengthen common efforts



# Workshop discussion



## Future Trends in Nuclear Physics Computing Meeting Notes

[Timetable](#)

This is the live meeting notes document for the [Future Trends in Nuclear Physics Computing Workshop](#) held on September 29 - October 1, 2020. This workshop, the third of the series (previous editions were in [2017](#) and [2016](#)), focuses on the Nuclear Physics Software & Computing community itself. Goals for the workshop are to identify what is unique about our community, find ways to strengthen common efforts, and chart a path for Software & Computing in Nuclear Physics for the next ten years.

We meet for four hours each day in a time window chosen to be as inclusive as possible for participants around the world. Substantial discussion time is included in the agenda, and session conveners will keep speakers to time in order to preserve the discussion time. This google doc will be used in advance to give the discussions structure and focus, as well as during the workshop itself to moderate and record the discussion and gather input from all participants, and after the workshop as the basis for summarizing and report writing. Editing is on, and all participants are encouraged to contribute in all phases.

Each day has a theme. In advance of the workshop, questions and discussion points for each day will be gathered here to guide a moderated common discussion following the talks. A short discussion period will follow each talk to address questions specific to the talk. The content prepared in advance will be augmented during the presentations and discussions.

A brief synopsis of the previous day will be part of an intro talk on days two and three.

The workshop will conclude with a short summary, but summarizing and report writing proper will proceed after the workshop. All participants are welcome and encouraged to join the meeting

**Live Notes (26 (!) pages)**

**WORKSHOP REPORT**

FUTURE TRENDS IN  
**NUCLEAR PHYSICS  
COMPUTING**

SEPT. 29 - OCT. 1, 2020

**EDITORS**

Alexander Kiselev (BNL)	Markus Diefenthaler (JLAB)
Amber Boehnlein (JLAB)	Ofer Rind (BNL)
Graham Heyes (JLAB)	Paul Laycock (BNL)
Mark Ito (JLAB)	Torre Wenaus (BNL)

**Draft (28 pages)**

# Unique software challenges for Nuclear Physics

## Scientific Problem Space

- Focus on non-perturbative QCD phenomena
- MC event generators for spin-dependent measurement, including novel QCD phenomena (e.g., GPDs, TMDs, Wigner functions)
- **Analyses considering large number of signal events simultaneously** (or multiple times)
  - **Contrary** to separating a few events from a large number of background events
    - **Example** Search of rare events with novel topologies
  - **Example** complexity of multi-dimensional, strongly correlated relationships among data (e.g., GPDs, TMDs, Wigner functions)
  - **Example** high-precision results which require complex analyses to control systematic uncertainties
  - Require unique software and computing strategies
- Relatively smaller size of experiments goes along with shorter experimental life cycles and faster changes in scientific goals

## Small Group Size

- Collaboration size in average smaller in NP than in HEP
- Tendency for everyone *“doing their own thing”*
  - Larger experiments, individual analyses can be numerous and quite different from another, with a small team on each top.
- Non-unified approach has inhibited progress in the field in the past.
- Transition to experiments with larger data size and more complex analyses
- Old culture cannot effectively address problems of scale of future experiments
- Relatively smaller group size asks for careful planning and design of the software effort: mix of in-house development, adoption of outside packages, and the choice of appropriate scale throughout.
- Challenge in finding the right balance.

# Common scientific software

## Common Scientific Software – The keys to success

- **The team is the most important** Do not separate development and operations, both ACTS and Rucio benefited from experience with developing and operating a worse software package, crucial experience. Developers keen to use modern software paradigms, open-source and open-minded, proactively searching out best practice and adopting it.
- **The project** Clear, well-focused short-term goals are important, grounded in real-world deliverables. Aligned with the long-term plan of building something sustainable and designed to be used by outside collaborators.
- **The management** Accept that the long-view takes longer to deliver the short-term product, manage expectations of the collaboration and funders to ensure the team have sufficient time and space to succeed.

## Scientific software careers need support

- Recognition, encouragement and reward: need to make software citations a priority
- Career paths of Research Software Engineers (RSE) need to be supported and not only at the labs

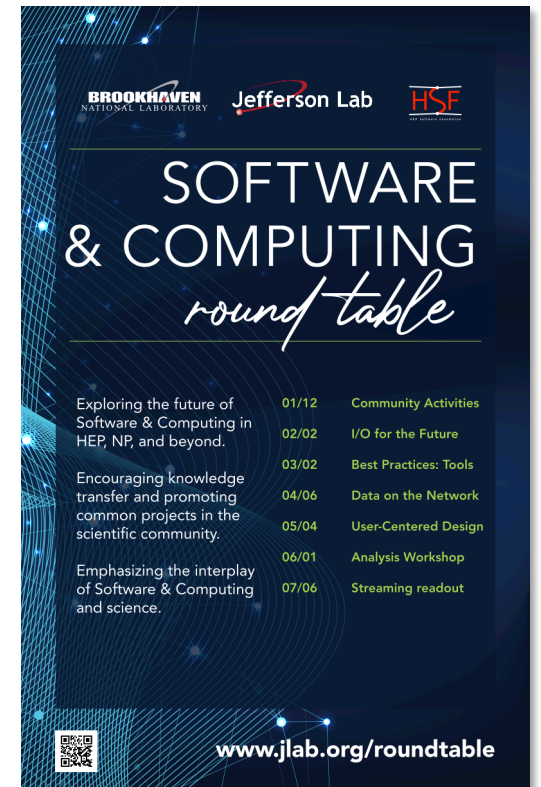
## NP software - should NP participate in HSF or build its own organization?

- Pros and cons, the balance of opinion favored NP participation in HSF. HSF is a do-ocracy, active participation will yield the biggest rewards.
- NP often has small groups developing solutions in-house, work with this reality.



# Community forum

- There was consensus on starting a new **community forum for NP** to discuss common projects, the role of data centers, unique challenges etc.
- Ongoing discussion on details and possible connection to HSF.
- Possible **goals of the community forum**
  - inform on building successful scientific software projects,
    - taking the unique challenges of NP in consideration
  - foster collaborative common software projects in NP
  - promote scientific software career support
- More details in Software & Computing Round table on January 12, 2021





---


# Section


## Software vision


# Towards the next-generation Nuclear Physics research model


FUTURE TRENDS IN  
**NUCLEAR PHYSICS  
COMPUTING**

SYMPOSIUM: MAY 2 • 1:00 p.m.  
Main Auditorium • Free Admission


 NUCLEAR PHYSICS IN A DECADE  
Donald Geesaman (ANL)

 NUCLEAR PHYSICS COMPUTING IN A DECADE  
Martin Savage (INT)

 MONTE-CARLO EVENT SIMULATION IN A DECADE  
Stefan Hoeche (SLAC)

 SYNERGY OF COMPUTING AND THE NEXT GENERATION  
OF NUCLEAR PHYSICS EXPERIMENTS  
Rolf Ent (JLAB)

RECEPTION TO FOLLOW

WWW.JLAB.ORG/CONFERENCES/TRENDS2017 



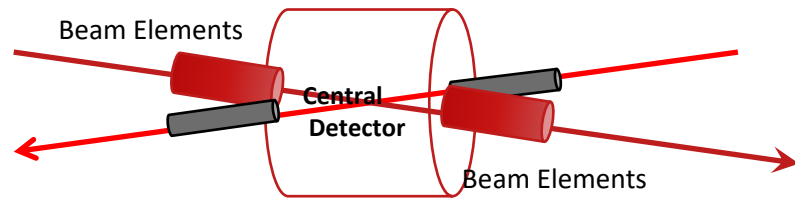
Donald Geesaman (ANL, former NSAC Chair) *“It will be **joint progress of theory and experiment** that moves us forward, not in one side alone”*

- All scientists of all levels, worldwide, should be enabled to actively participate in the NP data analysis.
- To achieve this goal, we must develop analysis toolkits using modern and advanced technologies while hiding that complexity (user-centered design).
- We must emphasize **data** as much as **analysis**. Experimental data must be open access, **readily accessible** and in self-describing formats.

# Machine-Detector interface (MDI)

## Integrated interaction region and detector design to optimize physics reach

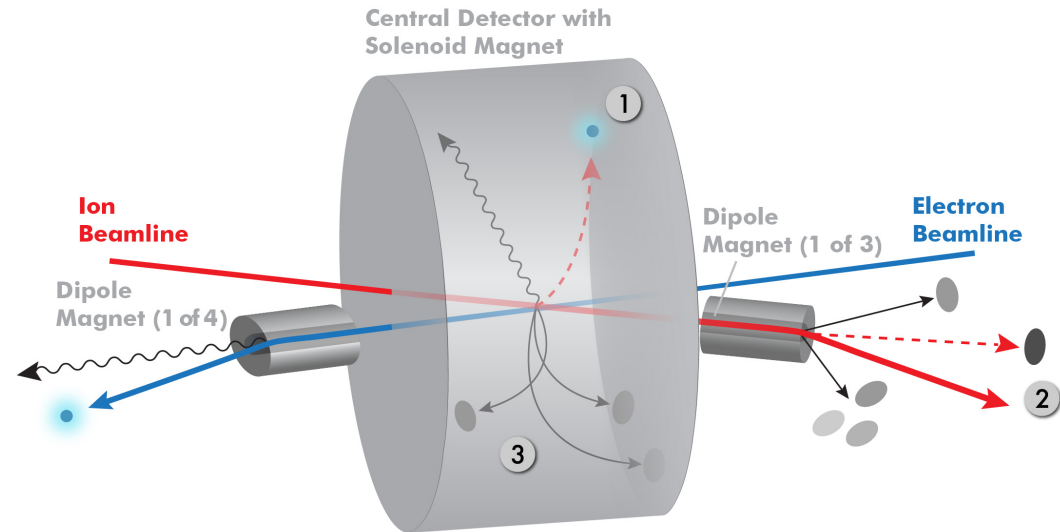
The aim is to get **~100% acceptance** for all final state particles, and measure them with good resolution.



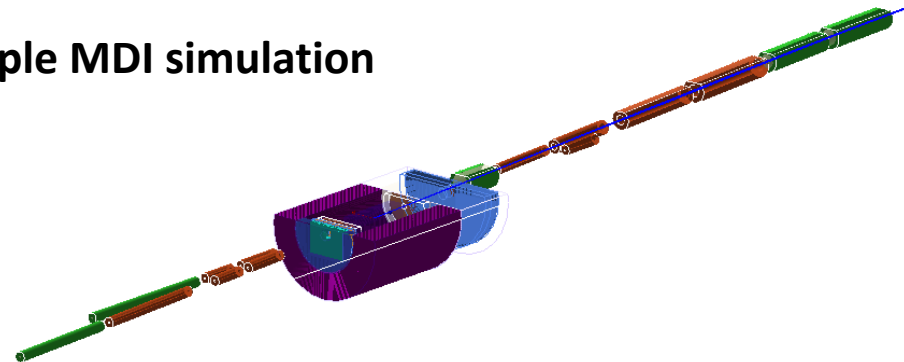
### Experimental challenges:

- beam elements limit forward acceptance
- central Solenoid not effective for forward

Possible to get **~100% acceptance** for the whole event.



### Example MDI simulation



# Beyond Machine-Detector Interface

## Integration of DAQ, analysis and theory to optimize physics reach

Compute-detector integration to deliver **analysis-ready data from the DAQ system**

- responsive alignment and calibrations in *real time / online*
- *real-time / online* event reconstruction and filtering
- *real time / online* physics analysis

**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, fast alignment and calibration, and reconstruction algorithms for near real-time analysis

# 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.

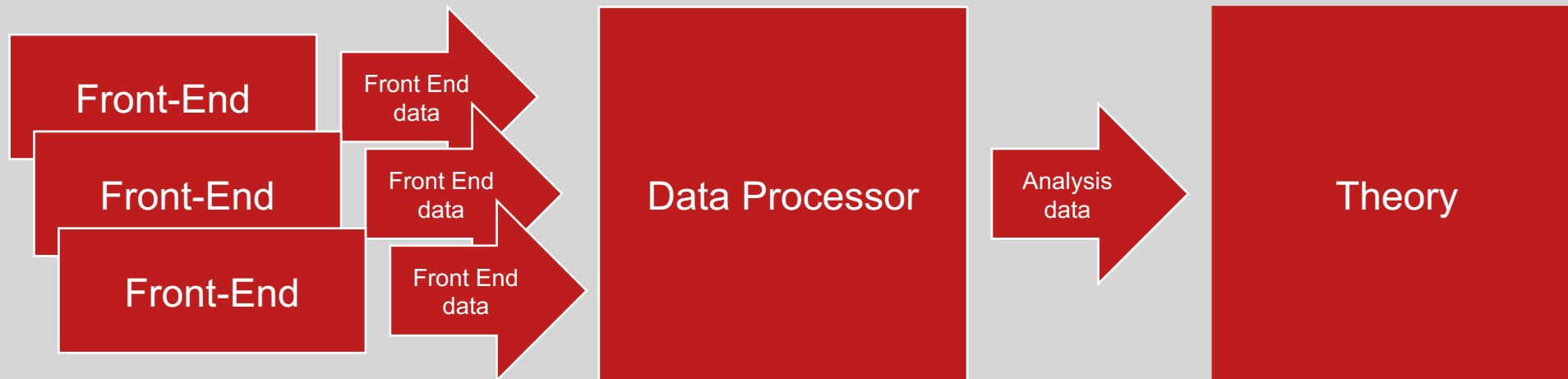
## Advantages of streaming readout

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



## 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)



# Example: 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.

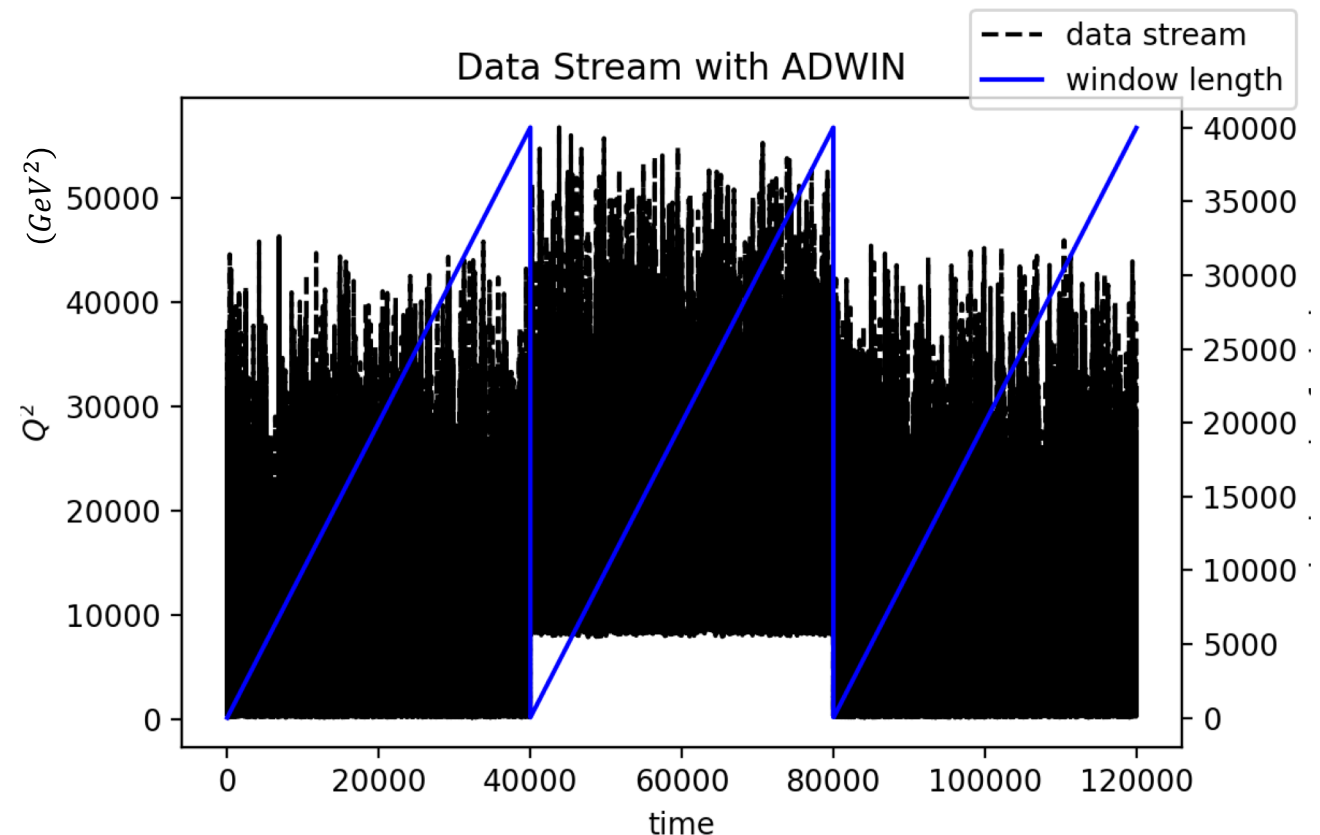


# Example data stream

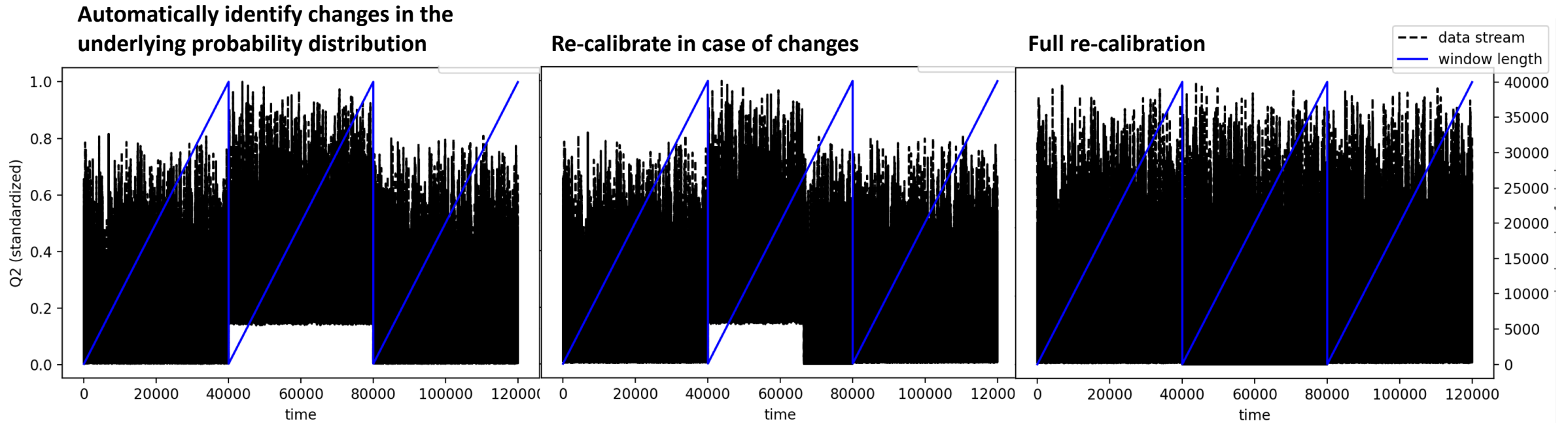
ADWIN is an **ADaptive WINdowing technique** used for detecting distribution changes, concept drift, or anomalies in data streams with established guarantees on the rates of false positives and false negatives

(A. Bifet and R. Gavaldà, *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.

# Considerations for 2<sup>nd</sup> interaction region

- **Discussion about high luminosity, medium energy 2nd interaction region**
  - IR optimized for nuclear imaging (GPDs, TMDs, Wigner functions, etc.)
  - Emphasizes the **unique challenges of Software & Computing in Nuclear Physics (slide 14)**
    - MC event generators for spin-dependent measurement, including nuclear imaging
    - complexity of multi-dimensional, strongly correlated relationships among data
  - require unique software and computing strategies
  - effort on 2<sup>nd</sup> IR will push for R&D on these challenges
- **Considerations for common software for both IR**
  - use simulation tools that allow to switch between IRs
  - **maximize common efforts**
    - modularity in the detector components and reconstruction software
    - work towards common formats
    - work towards common validation
    - prepare/optimize compatible binning so that what is used in one experiment may be mapped easily to what in use in the other (same edges)
    - work towards common data and analysis preparation

# Summary

Markus Diefenthaler

[mdiefent@jlab.org](mailto:mdiefent@jlab.org)

- Developments towards a high luminosity, medium energy 2nd interaction region at the EIC emphasize **unique challenges of Software & Computing in Nuclear Physics**
- **Common projects will help to build a sustainable software efforts**
  - Work plan for the Software Working Group (EoI)
  - Building a software community
- **Everyone invited to join common efforts.**

