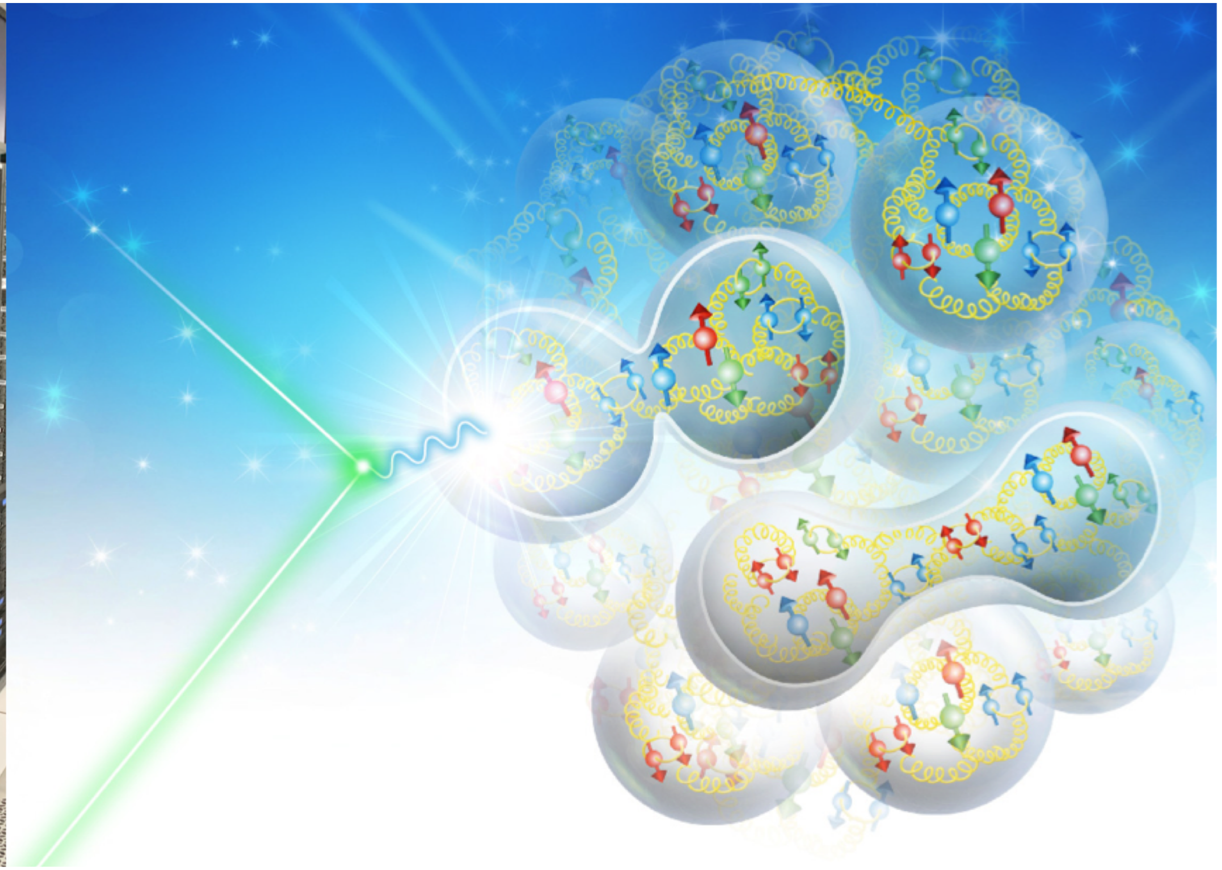
Software & Computing for the Electron-Ion Collider

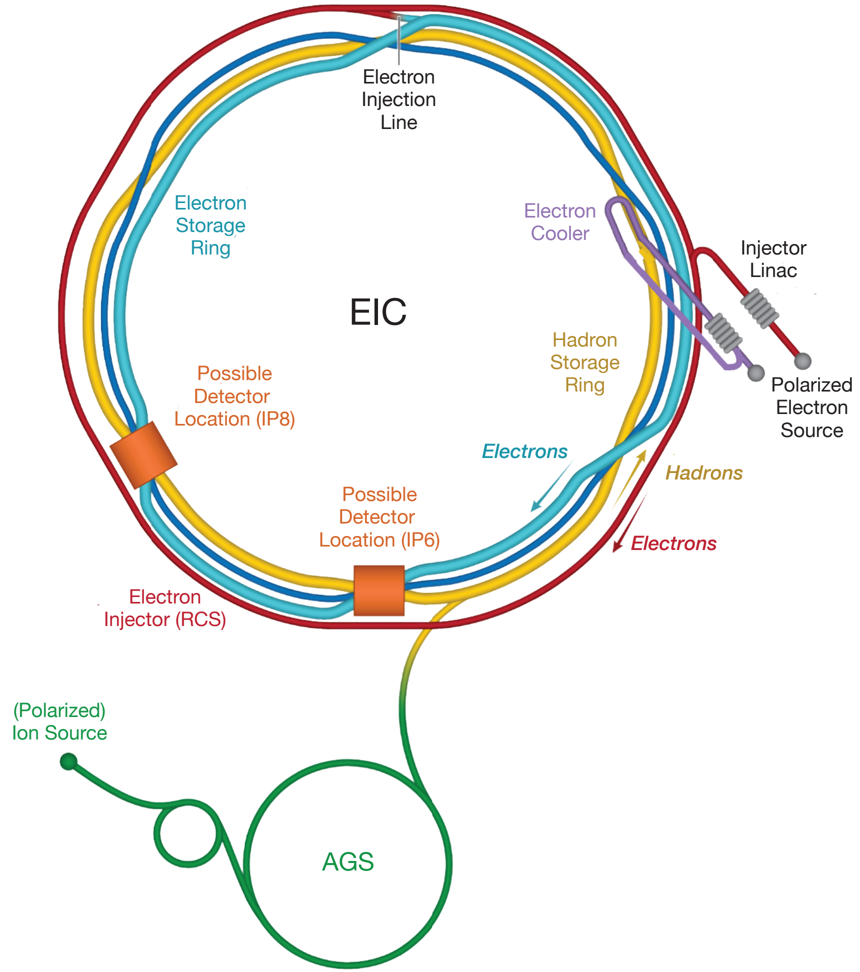


Markus Diefenthaler

EIC^2 EIC Center at Jefferson Lab

Jefferson Lab

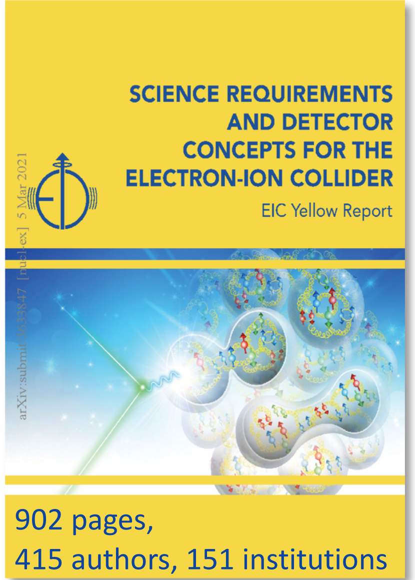
The Electron-Ion Collider (EIC)

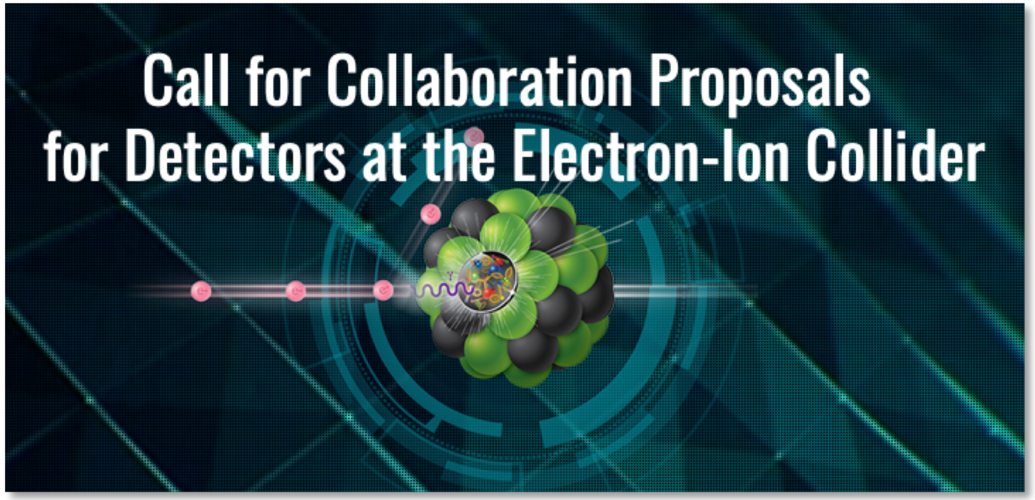


Frontier accelerator facility in the U.S.

- **World's first collider of**:
  - Polarized electrons and polarized protons,
  - Polarized electrons and light ions (d, ^{3}He),
  - Electrons and heavy ions (up to Uranium).
- The EIC will enable us to embark on a **precision study of the nucleon and the nucleus at the scale of sea quarks and gluons**, over all of the kinematic range that are relevant.
- Jefferson Lab and BNL will be host laboratories for the EIC Experimental Program. Leadership roles in the EIC project are shared.

EIC Initiatives in 2019–2022



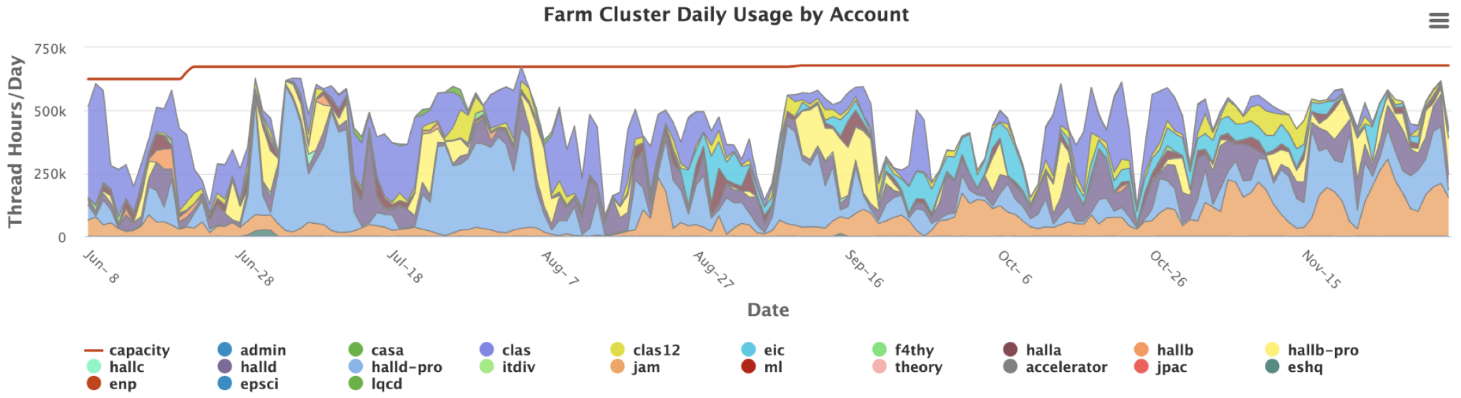


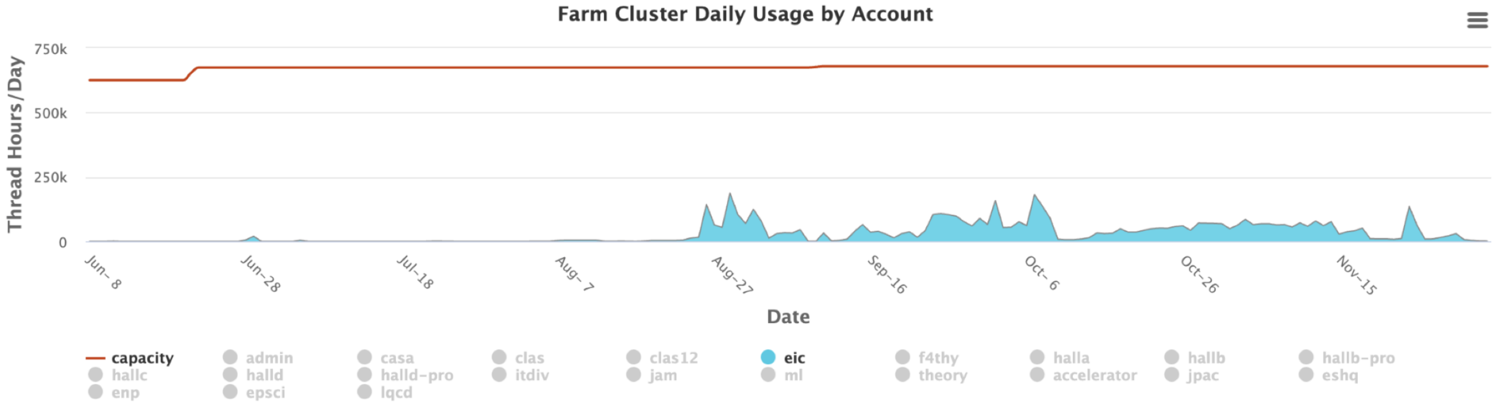
- The **EIC Yellow Report** describes the physics case, the resulting detector requirements, and the evolving detector concepts for the experimental program at the EIC.
- The studies leading to the EIC Yellow Report were commissioned and organized by the **EIC User Group**.
- Jefferson Lab and BNL issued a **Call for Collaboration Proposals for Detectors**, with three proposals being submitted end of 2021 (ATHENA, CORE, ECCE).
- A scientific-technical committee of renowned and independent subject matter experts advised Jefferson Lab and BNL in early 2022 on how to realize the EIC Project Detector.
- The international EIC community is now forming a scientific collaboration to support the realization of the EIC Project Detector.

Scientific Computing Resources for the EIC in 2021

CPU

Total for year 5.3M core hours
Pledged 4.3M core hours

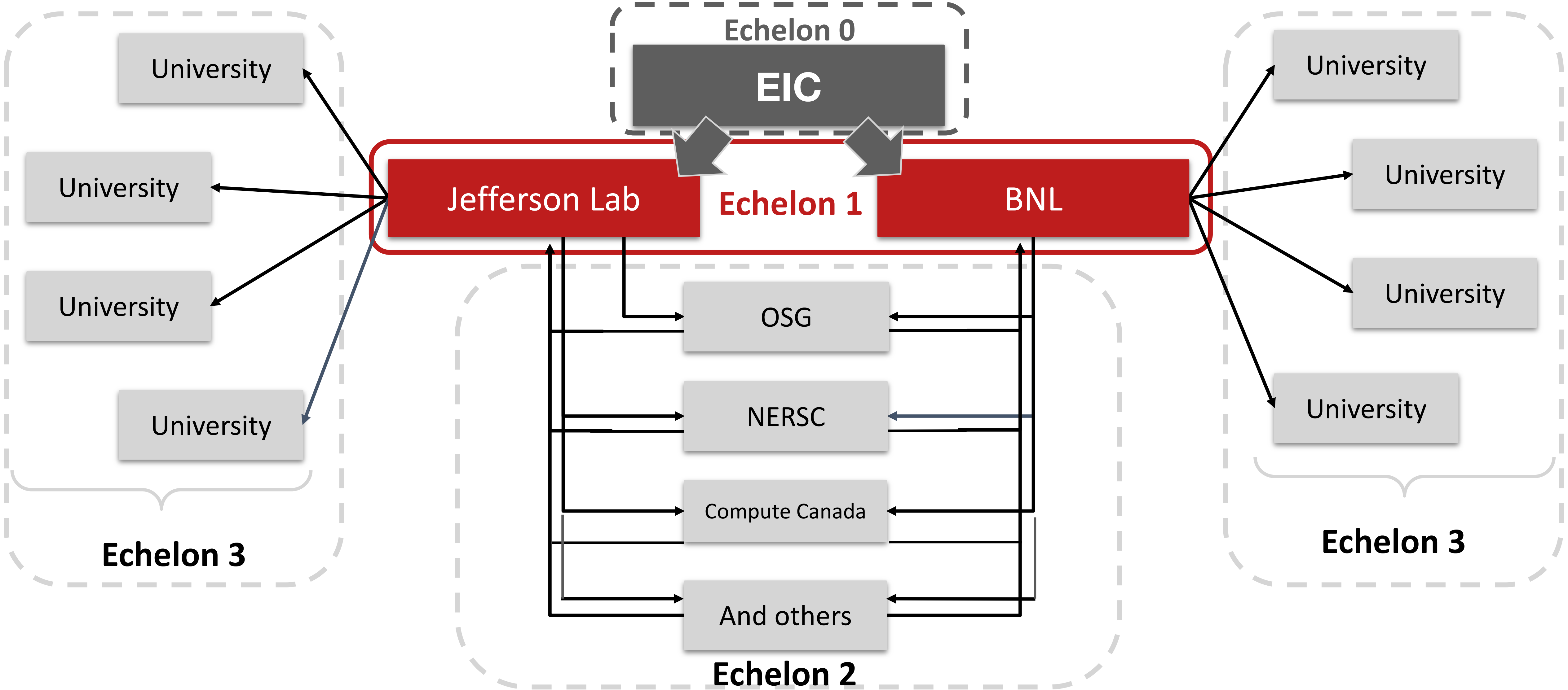




Storage

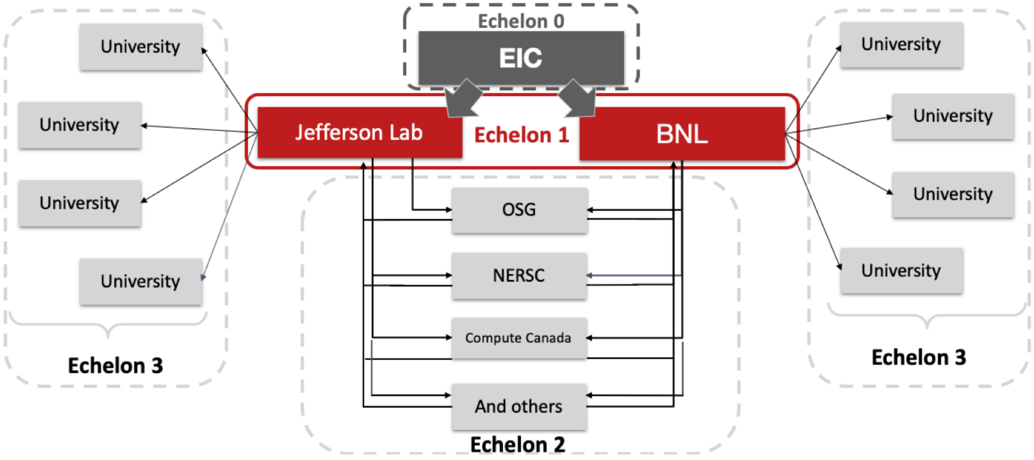
1PB available
10% used

Distributed Computing Model



Nearly all storage (raw data, reconstructed data, simulated data) is stored across **Echelon 1** sites.

Federated Approach and Data Management



Federated approach

- Computing coordination effort between Jefferson Lab and BNL.
- Containerization and cvmfs share allowed to run software on various sites.
- Progress towards federated login for access to shared computing resources.

XRootD

RUCIO

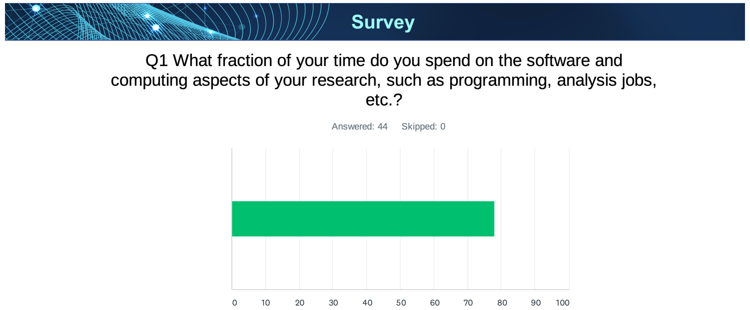
Data transfer and management

- We are building up **XRootD** as high performance data system for the EIC and our experimental program:
  - XRootD development is driven by the physics community. This will allow us to work with the developers in case of issues or feature requests.
- We are considering **Rucio** as scientific data management system that is compatible with a heterogenous computing environment.
- **Data and analysis preservation** will be intrinsic part of the workflow tools.

Our Vision for Software & Computing at the EIC

"The purpose of computing is insight, not numbers." Richard Hamming (1962)

Software & computing are an integral part of our research:



Survey among NP Ph.D. students and postdocs in preparation of "Future Trends in NP Computing"

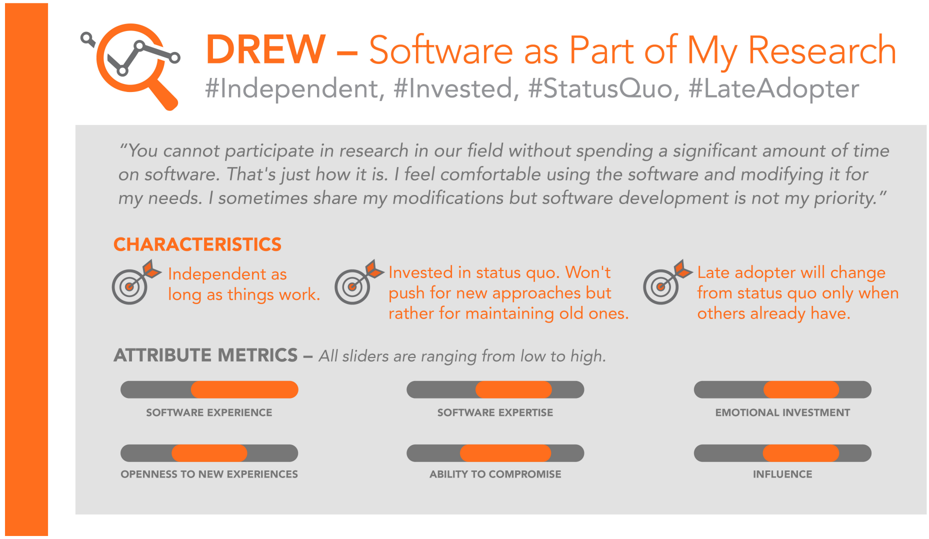
- **Goal** We work with our large Users Organization (over 1600 scientists from over 275 institutions) on data-intensive challenges and AI/ML and would like to ensure that also for the EIC scientists of all levels worldwide can participate in EIC analysis actively.
- **User-Centered Design**: To achieve this goal, we must develop simulation and analysis software using modern and advanced technologies while hiding that complexity and engage the wider community in the development.

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.

User-Centered Design

- <u>State of Software Survey</u>: Collected information on software tools and practices during the Yellow Report Initiative.
- As part of the State of Software Survey, we asked for volunteers for focus-group discussions:
  - Students (2f, 2m), Junior Postdocs (2f, 3m), Senior Postdocs (2f, 3m), Professors (5m), Staff Scientists (2f, 3m), Industry (2f, 2m)
- **Results from the six focus-group discussions**:
  - Extremely valuable feedback, documented many suggestions and ideas.
  - Developed user archetypes with Communication Office at Jefferson Lab and UX Design Consultant:



User Archetypes: Input to software developers as to which users they are writing software for:

- Software is not my strong suit.
- Software as a necessary tool.
- **Software as part of my research.**
- Software is a social activity.
- Software emperors.

- Will repeat State of Software Survey now after detector collaboration proposals:
  - The regular software census will be essential to better understand and quantify software usage throughout the EIC community. During the next survey, we will also ask on feedback on the user archetypes.

User-Centered Design: Listen to Users, and/then Develop Software

User-Centered Design

- Software census
- Focus groups and user archetypes
- Develop testing community

Discoverable Software

- Single point of entry
- Feasible option for >80% of EIC simulations and analyses
- Spack as package manager

Workflows

- Template repositories for key analyses
- Template repositories for validation workflows

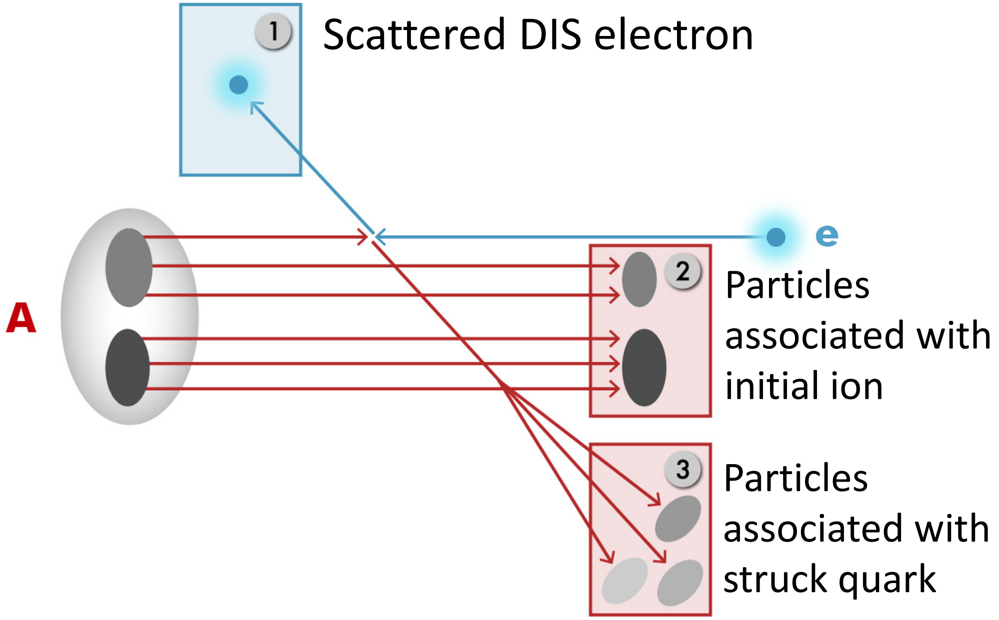
Data and Analysis Preservations

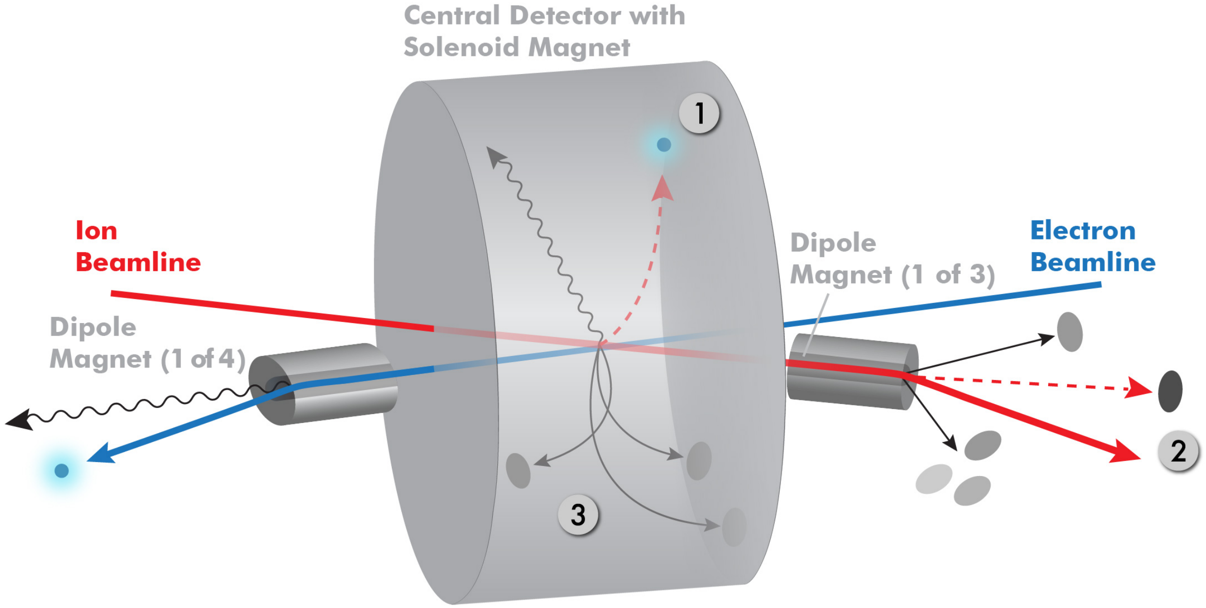
- User analysis code/software registry
- Tutorials on reproducible analyses

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

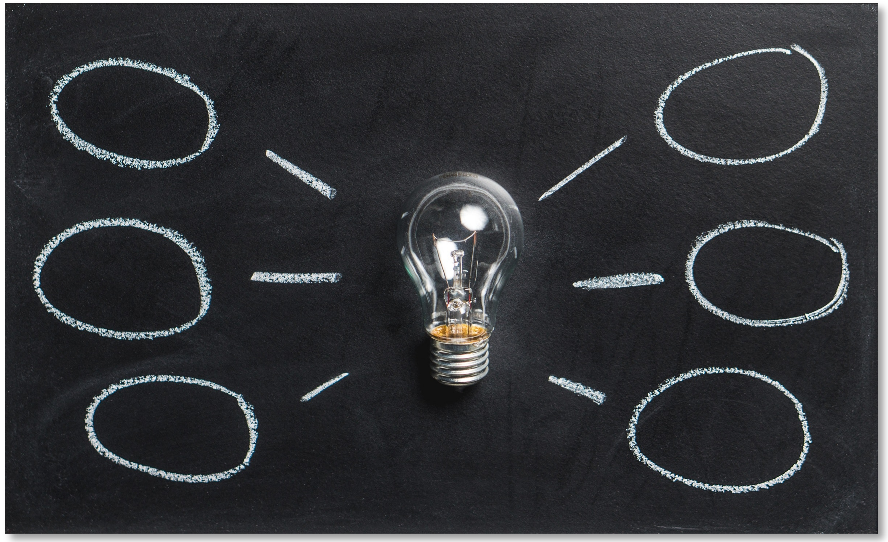
Integration of DAQ, analysis and theory

- 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.

Towards the next-generation research model in Nuclear Physics

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

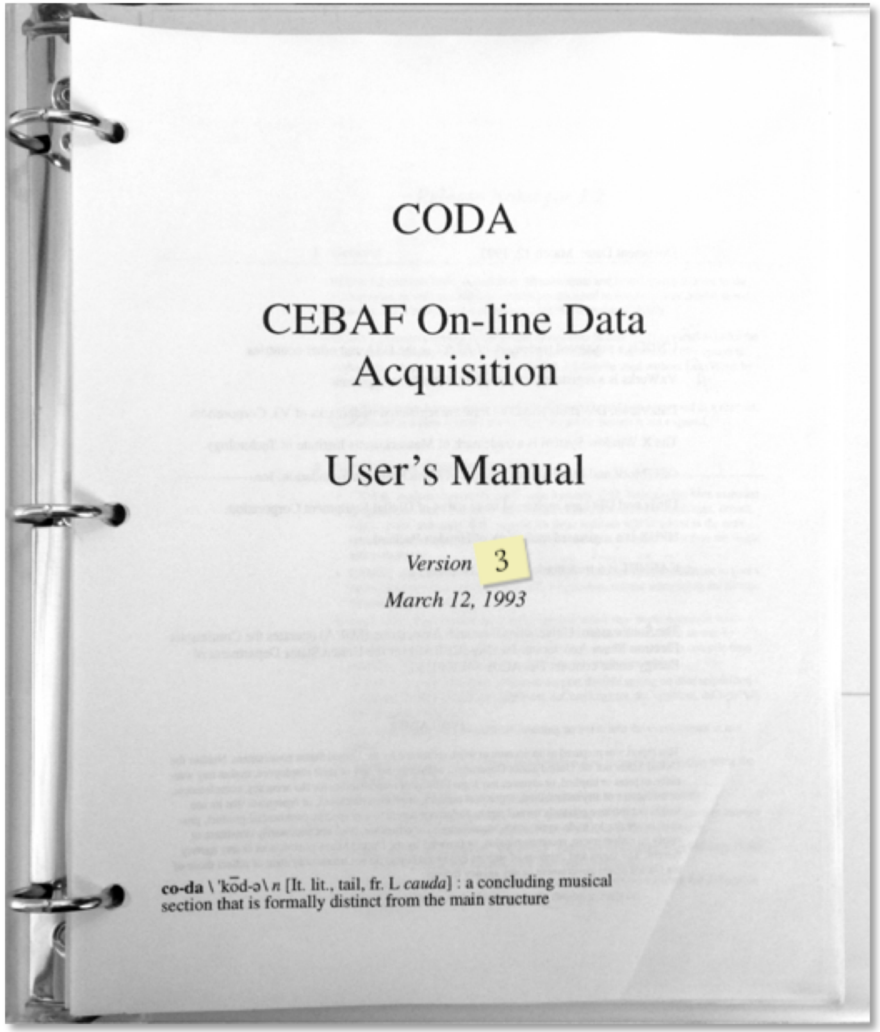
evolve & develop **Nuclear Physics research model** based on these advances



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

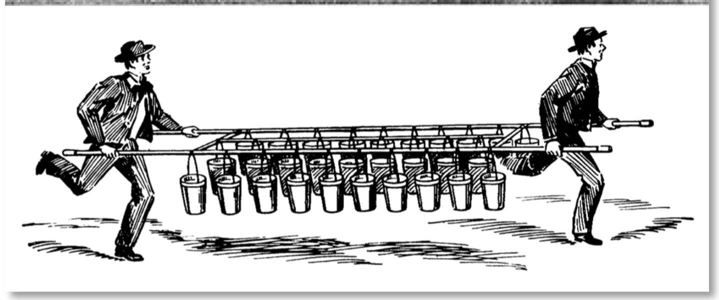
- The data rate from a detector is impossible to capture with an affordable data acquisition system without a trigger to reduce event rates.
- Even if the untriggered data rate could be captured, it would be impossible to store.
- Even if it could be stored the full dataset would represent a data volume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

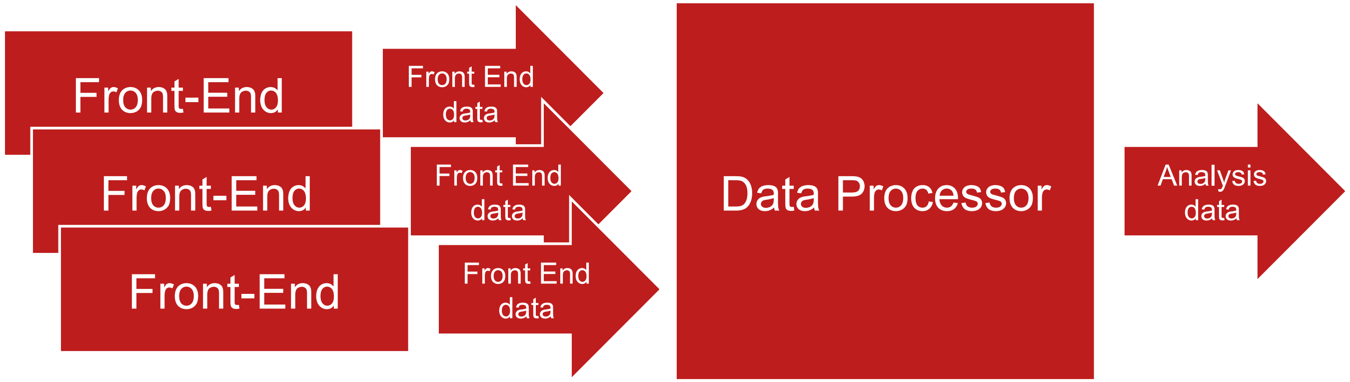
Definition of Streaming Readout

- 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



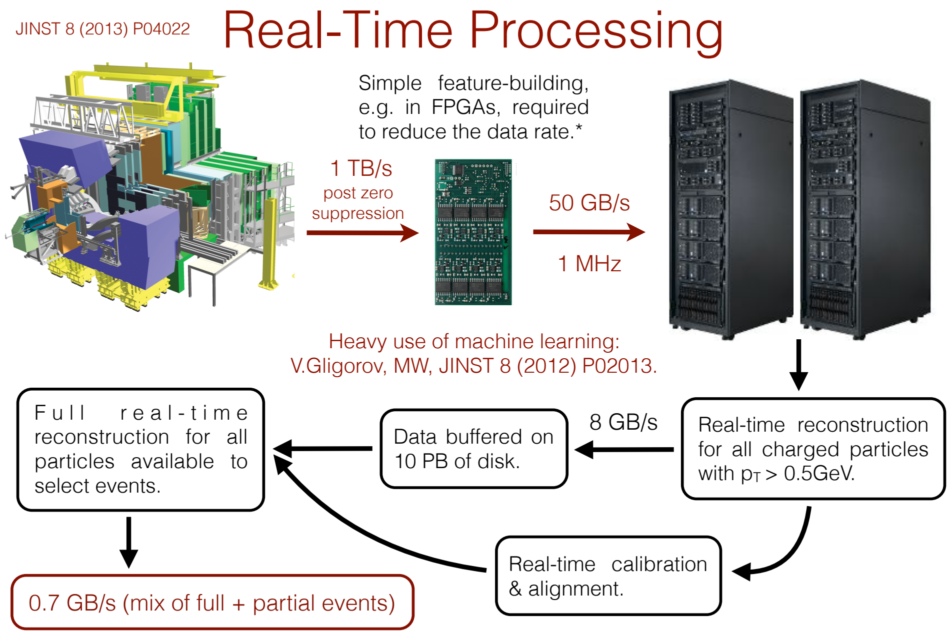
Data Processor

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

Features

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

LHCb Example

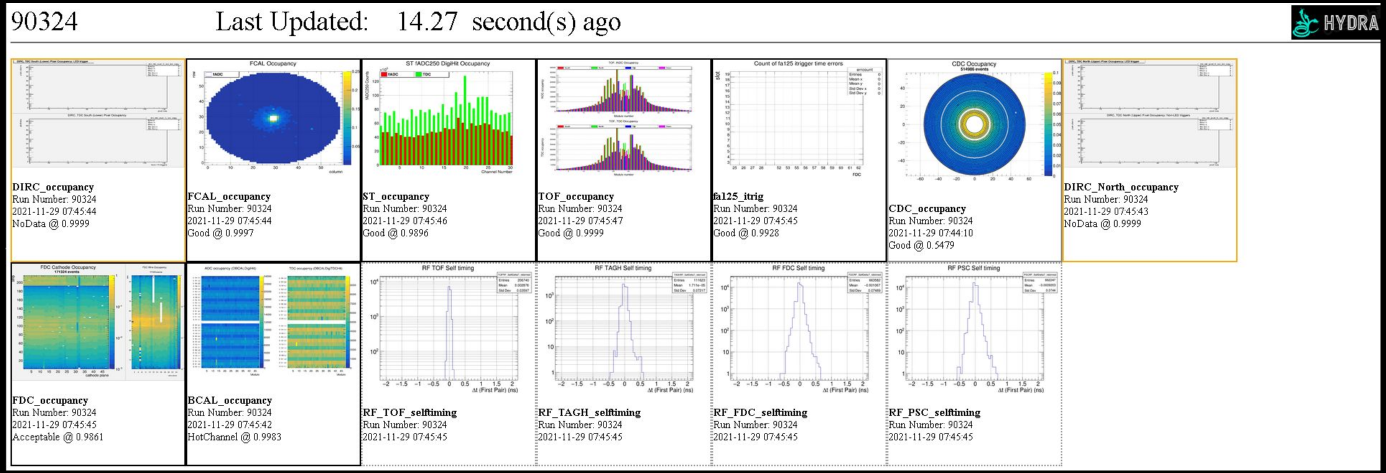


*LHCb will move to a triggerless-readout system for LHC Run 3 (2021-2023), and process 5 TB/s in real time on the CPU farm.

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

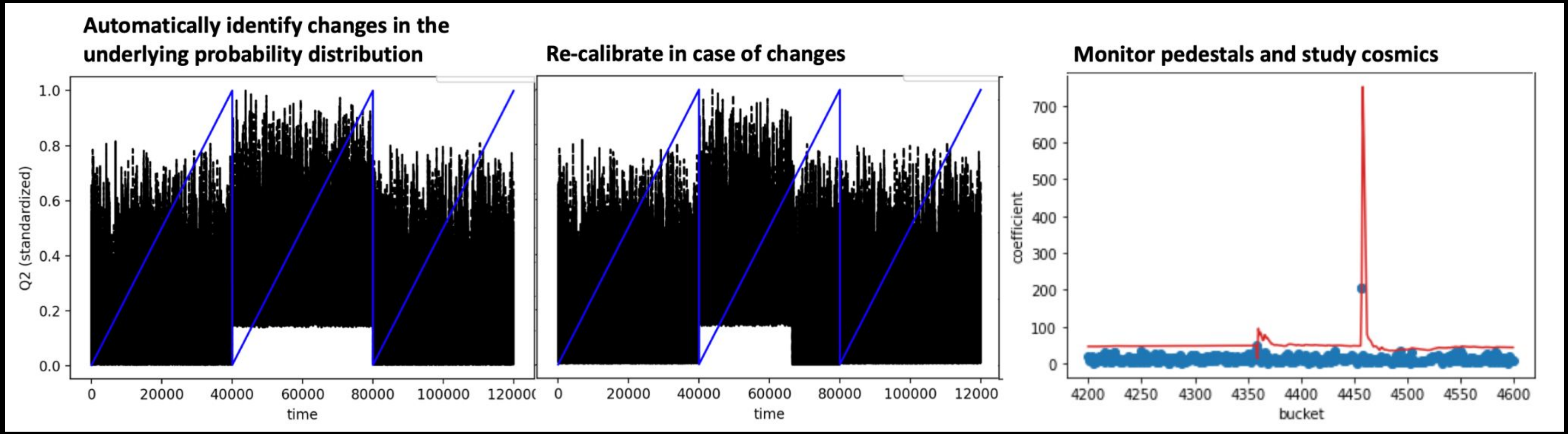
Approach:

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



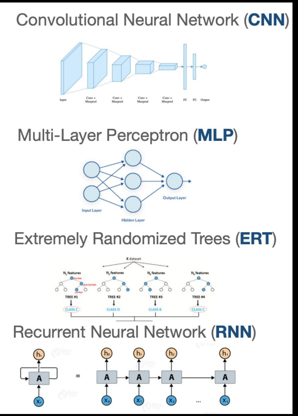
9

Event Reconstruction

AI-based Tracking

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

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

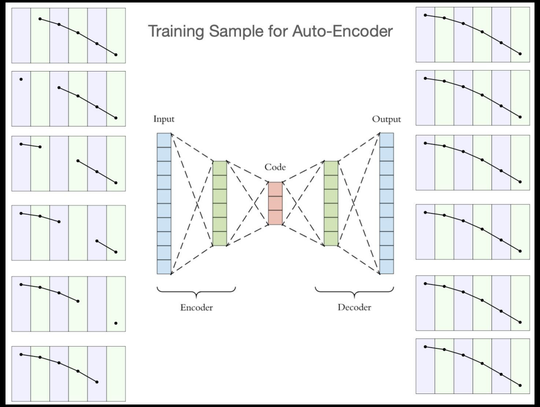


Different Network types were evaluated for accuracy and speed. MLP is chosen to be the best fit, due to implementation simplicity, accuracy and inference speed

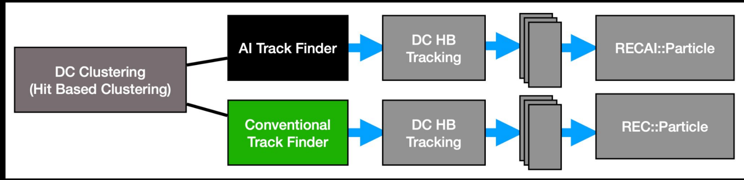
	Features	TP	FP	PA	TA	Time (ms)
ERT	6	100%	6.14%	100%	100%	0.36
MLP	6	99.96%	10.77%	98.88%	99.65%	0.12
CNN	36x112	96.11%	28.11%	94.26%	94.26%	1.2
RNN	36	88.40%	11.60%	-	-	-

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

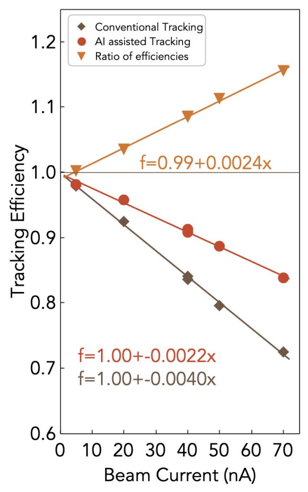


AI track classification and segment recovery network was implemented as a CLARA service. Tracking code was modified to separate clustering from track finding.



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%.

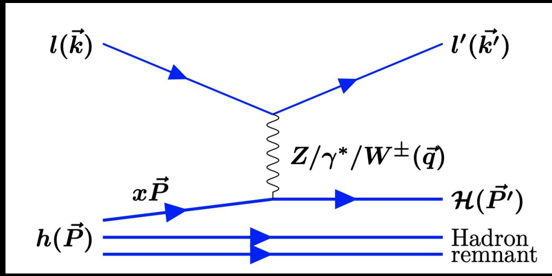


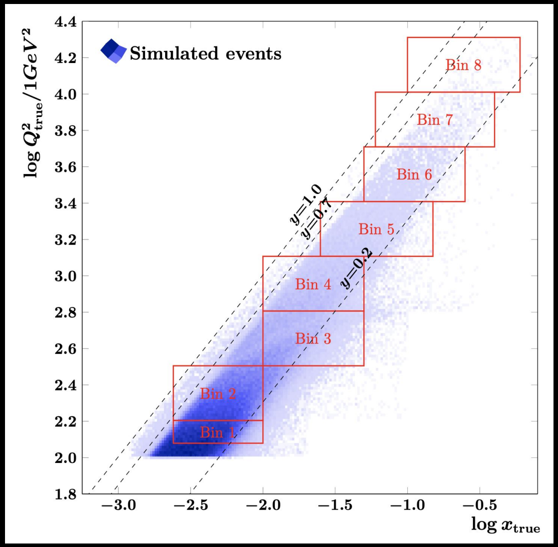
5

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

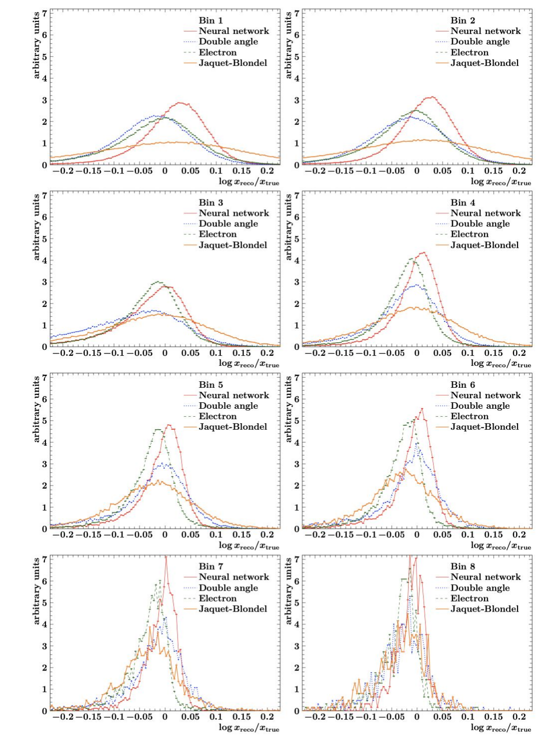


Figure 6: *Distributions of* $\log x - \log x_{\text{true}}$ *for different reconstruction methods in individual analysis bins. For a better visibility, the centers of bins in each distribution connected with straight lines.*

19

Software Developments

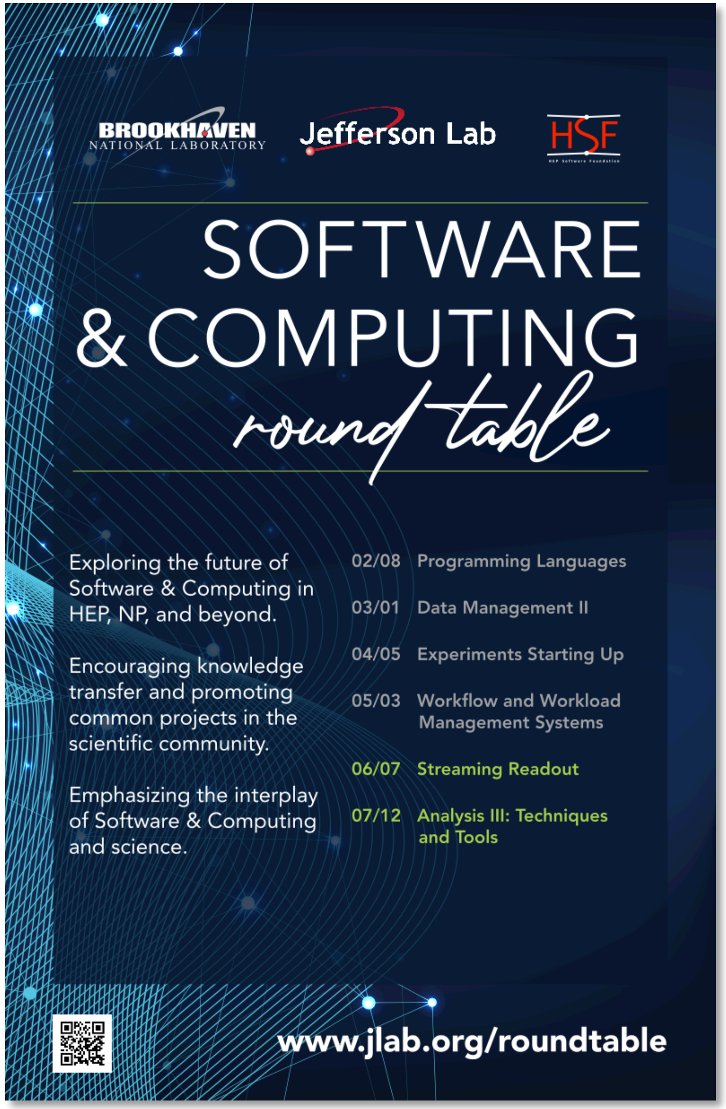
Software is in very early life stage

- Our focus on common software tools.
- We work with HEP standards, e.g., HepMC3 for MC event simulation,
- and engage with the wider community.

Community engagement

- **General-purpose MC event generators**: Herwig, Pythia, Sherpa
  - Community project on MC event simulation for ep and eA, current focus on validation with HERA data.
- **Geant4 collaboration**
  - Started collaboration on validation in 2017, since 2021 also involved in R&D.
- **HEP Software Foundation**
  - Started to collaborate on Software & Computing Round Table and software tutorials. Discussion of NP Software Forum as part of HSF.
- **Software Working Group** within the EIC User Group.
  - We co-chair the SWG and work with the community to address software needs and evolving R&D.

Software & Computing Round Table



- Seminar series on the **interplay of computing and science**
- With O(50) participants per month
- Initiated at Jefferson Lab after the first "Future Trends in NP Computing" workshop in 2016 with two main goals:
  - **Knowledge transfer**
  - **Encourage common projects**
- Since 2020 **jointly organized with BNL and the HSF** with software & computing topics from the wider NP and HEP community.
- Recordings available on YouTube:

Measurements and simulations in experimental Nuclear Physics

Measurement

Simulation

Role of simulations in experimental Nuclear Physics

MC event generator → Detector simulation → Digitization → Reconstruction → Physics Analysis

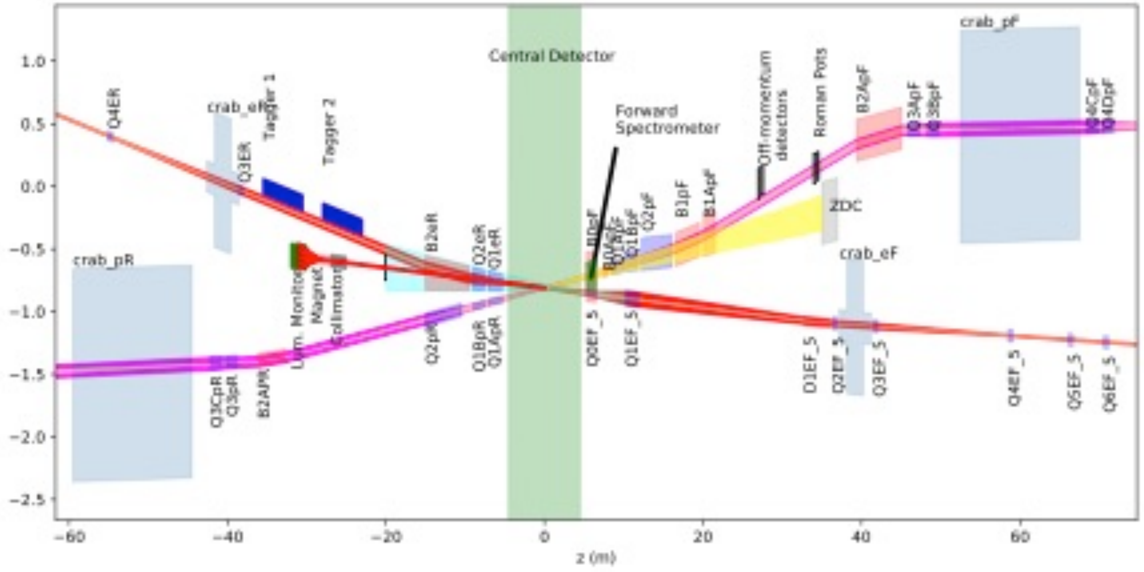
Design Experiments Design and develop detectors and large-scale detector systems based on key measurements / physics reach and background estimates. Optimize the design.

Analysis Develop and verify analysis methods and tools: Does the analysis method or tool give the correct result? Estimate systematic uncertainties.

Verify Measurements Detailed simulations essential for commissioning experiments and verify analyses.

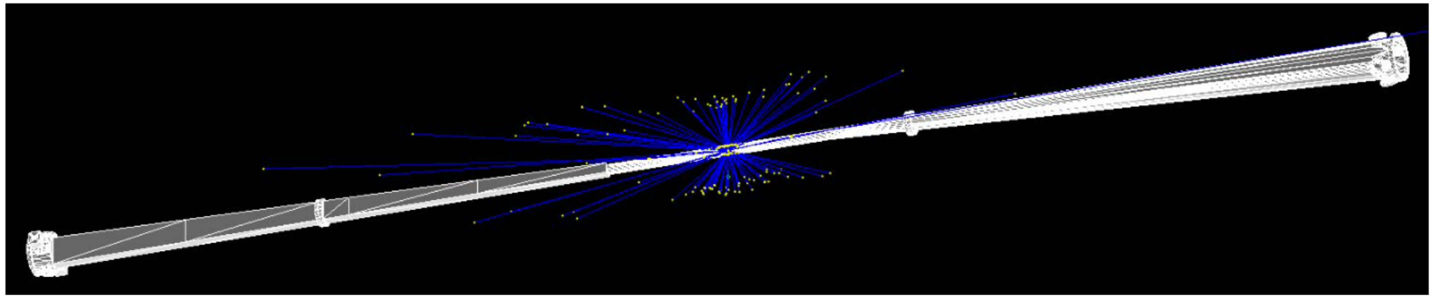
MDI in Simulations

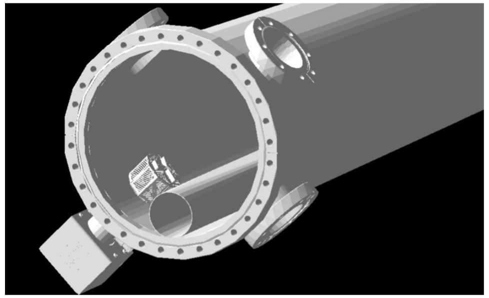
IR Layout



Accelerator and beam effects influence EIC measurements. Profound consequences on measurement capabilities of the EIC and layout of the detectors.

CAD Interface (accelerator elements and service structures)



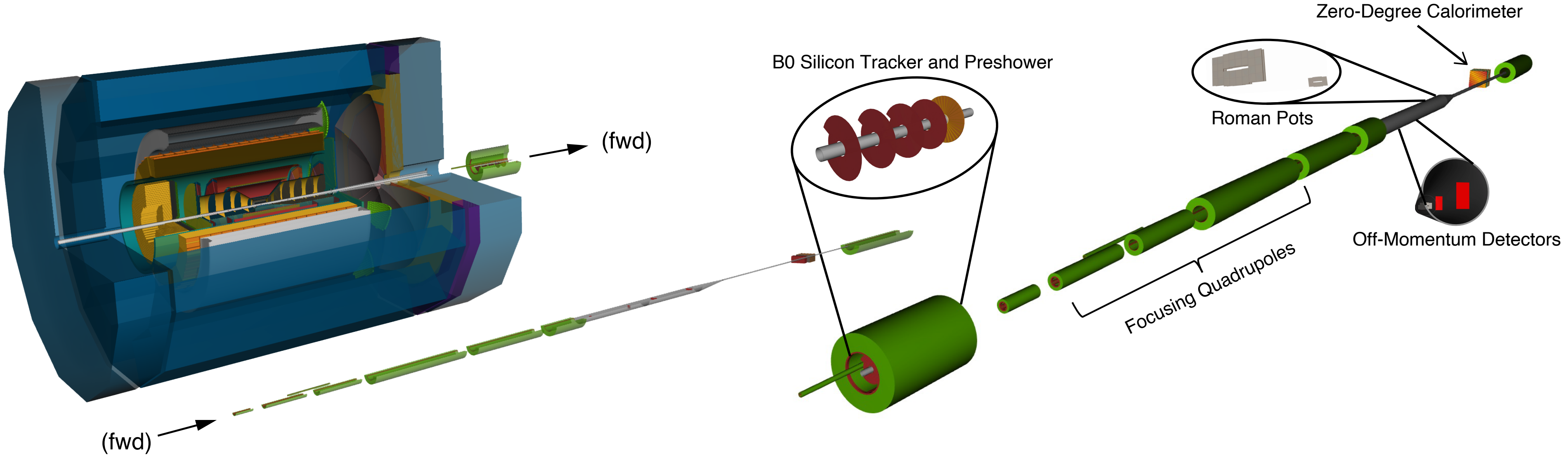


EIC Project

Simulation based (in part) on CAD files provided by EIC project engineering teams, rather than a bottoms-up reliance on constructive solid geometry (Screenshots from **eAST**)

Support Detector Design

- **Accelerate Simulations**
  - Detector design optimization using AI/ML as a service.
  - AI/ML for the simulation of calorimeters, Cherenkov detectors, etc.
    - Fast simulations fully integrated into Geant4.
- **Reconstruction**
  - AI/ML to accelerate reconstruction.
  - Reconstruction with far-forward detectors fully integrated (ATHENA example for IP6).



Our R&D Towards Next-Generation Detector Simulations: eA)ST

Detector Simulation

- **Turn-key application**
- **Built on top of Geant4 for full and fast simulations**
- **With library of potential detector options**

Requirements

- Ease of **leveraging new and rapidly evolving** technologies:
  - AI/ML to accelerate simulations
  - Heterogeneous architectures:
    - AI/ML is the best near term prospect for using LCF/Exascale effectively.
- Ability to reuse existing simulation work
- Ease of switching detector options
- Ease of switching between detailed and coarse detector descriptions

Project

- Support for high concurrency heterogeneous architectures and fast simulations integrated with full detector simulations allows to leverage AI/ML in Geant4.
- Makoto Asai, who led Geant4's multi-threaded reengineering to support high concurrency heterogeneous architectures, is now at Jefferson Lab and leading the next phase in concurrent Geant4, sub-event parallelism.
- We are building up team at Jefferson Lab on next-generation detector simulations with strong support from wider EIC community, in particular from BNL.

Interdisciplinary approach

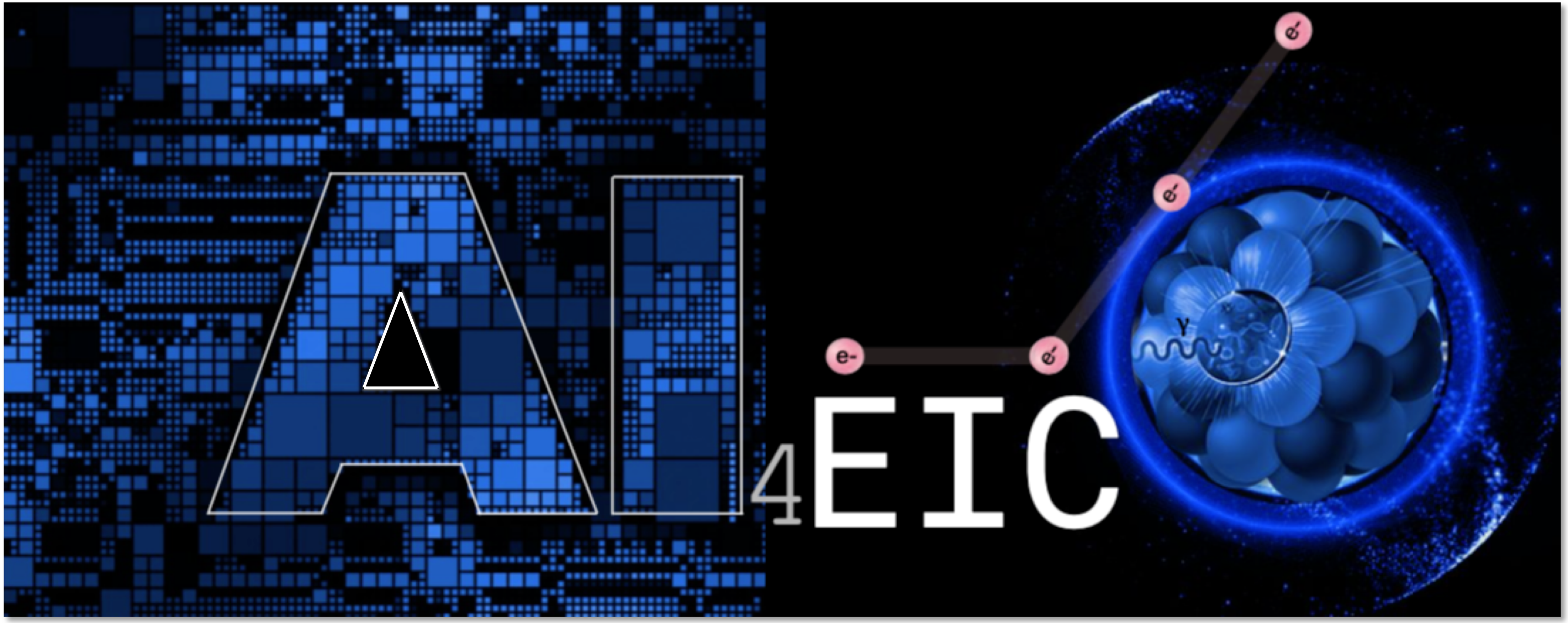
- interplay between **Data Science** and **Nuclear Physics**
  - data scientists need problems to solve
  - nuclear physicists want to extract information from **all** the data and find correlations / common features
  - Nuclear physics problems give insights into research in computer science and mathematics
  - great opportunity for education

- **AI / ML research**
  - scientific, systematic approach to applying AI / ML approaches to NP problems
  - activation functions, network design particular to Nuclear Physics applications
  - building efficient networks no more complex than necessary
  - **need to trust AI / ML**
    - drive for *explainable AI* and *uncertainty quantification*
    - human interaction could be applied with great benefit to better understand the requirements and dynamics of such criteria in the NP domain
    - debatable whether *explainable* is a useful criterion for a ML model. We don't have the words for theories we haven't discovered yet

- **reference datasets for AI / ML development in NP**

AI / ML for EIC

AI / ML already has an important presence in EIC, with one of the proto-collaborations (ECCE) applying it to detector design optimization, as well as applications such as streaming DAQ, and a AI Working Group as part of Software Working Group to explore and develop AI / ML's potential.



Jefferson Lab

Current challenges

- **Accelerating simulations**
  - calorimeter in particular
  - but also PID, e.g., Cerenkov detectors

- **AI driven detector design** Bayesian optimization for EIC detector R&D

- **HPC utilization**
  - Experimental NP, HEP have few or no payloads appropriate to the LCF/Exascale which are accelerator based.
  - ML is the best near term prospect for using them effectively.
  - Can we find the ML payloads? Do they use substantial processing resources?

- **ML for event generators**
  - replace models with ML as we do in detector simulations

Markus Diefenthaler
mdiefent@jlab.org

Scientific Computing for the EIC

- The EIC will enable us to embark on a **precision study of the nucleon and the nucleus at the scale of sea quarks and gluons**. Software & Computing will be an integral part of EIC science.
- We focus on a **federated approach** for distributed computing and **common software projects** with the NP and HEP community.
- Ware working to **accelerate science**:
  - AI/ML and heterogenous computing for next-generation simulations.
  - Seamless data processing from DAQ to analysis using streaming readout and AI/ML.