Artificial Intelligence at Fermilab

Tia Miceli
Fermilab's 56th Annual Users Meeting
29 June 2023

Since last year…

The New York Times

SUBSCRIBER-ONLY NEWSLETTER

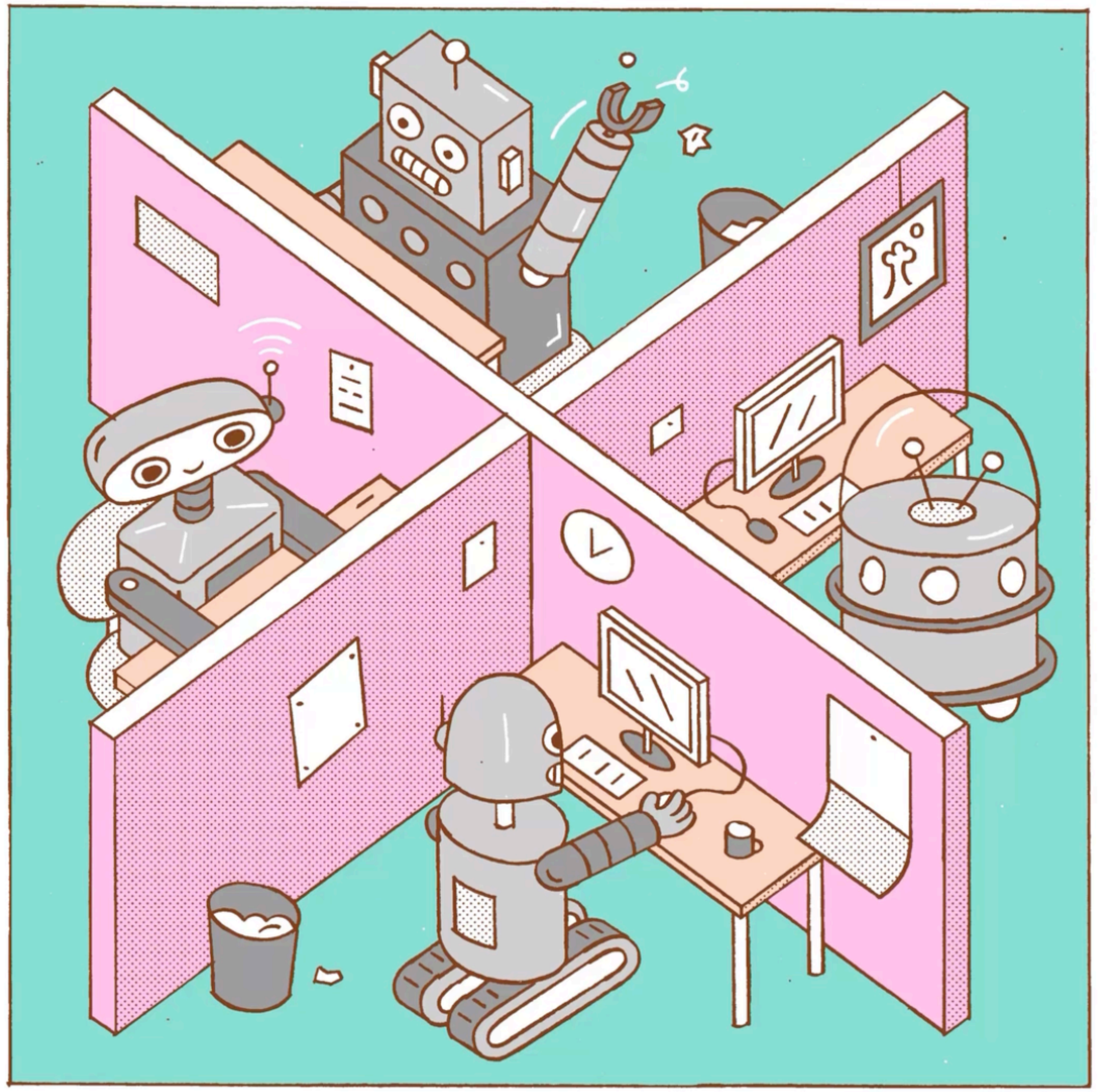
On Tech: A.I.

How to Use A.I. to Automate the Dreaded Office Meeting

Generating a slide deck, talking points and meetings minutes can all be done in a snap. All you need are the right prompts.

By Brian X. Chen

June 9, 2023



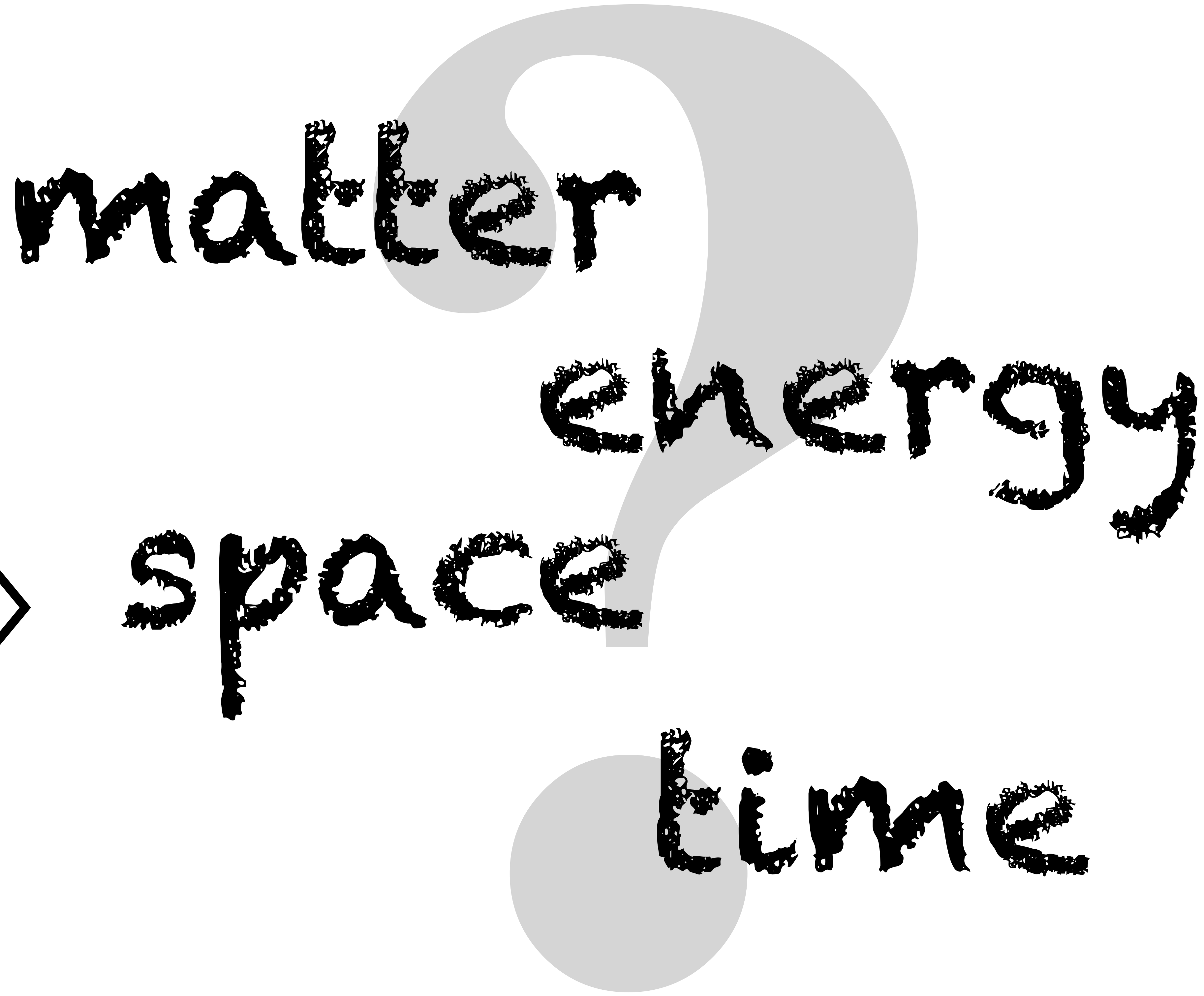
Oscar Nimmo

Fermilab's AI Vision

AI for physics, physics for AI

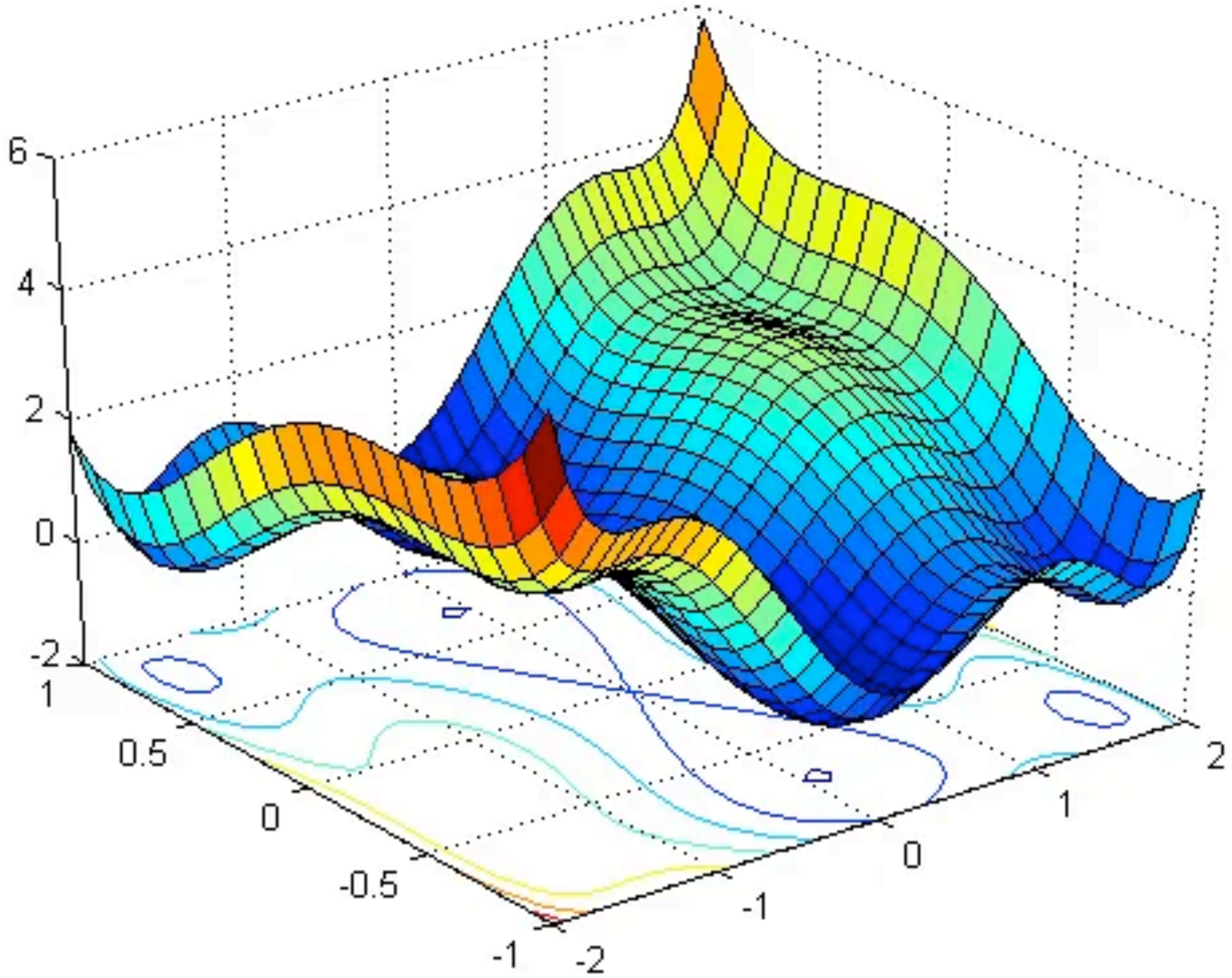
@Fermilab: AI Technology applied to HEP

Machine learning, neural networks, domain adaptation, surrogate models, active learning, continuous learning, inference at-the-edge

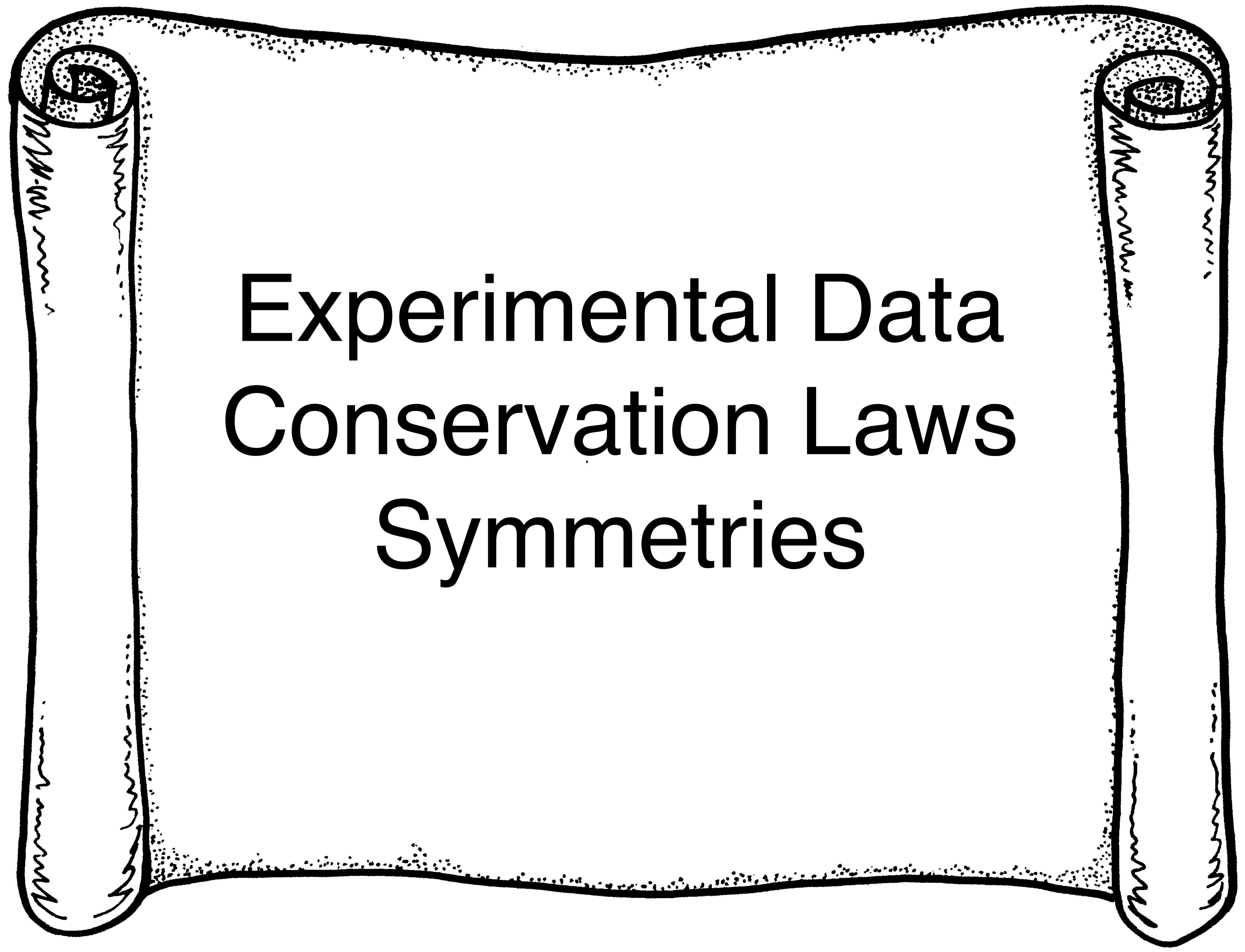


Fermilab's Research Missions
In High Energy Physics
(Neutrino, collider, dark energy, dark matter, theory, quantum, detector and instrumentation…)

@Fermilab: Physics applied to develop new AI algorithms



New constrained loss functions, physics inspired neural networks, enhanced uncertainty quantification



Laws of Physics

Setting the scene… circa “the before times” (pre-March 2020)

- Lots of AI/ML developments around the lab and groups may not be aware of each other

- Want to learn and excel together. Support our colleagues so that Fermilab can lead in AI for HEP.

- Grants! DOE starts to invest in AI, and Fermilab has lots of ideas!

Grassroots coordination begins

AI Project Office supports AI activities across Fermilab Physics

Fermilab’s AI Project Portfolio

Algorithms for HEP science

Physics-inspired data & models; Robust & generalizable learning; Fast and efficient algorithms

Computing hardware and infrastructure

Operations and control systems

Real-time AI systems at edge

Fermilab’s AI Project Portfolio

Algorithms for HEP science

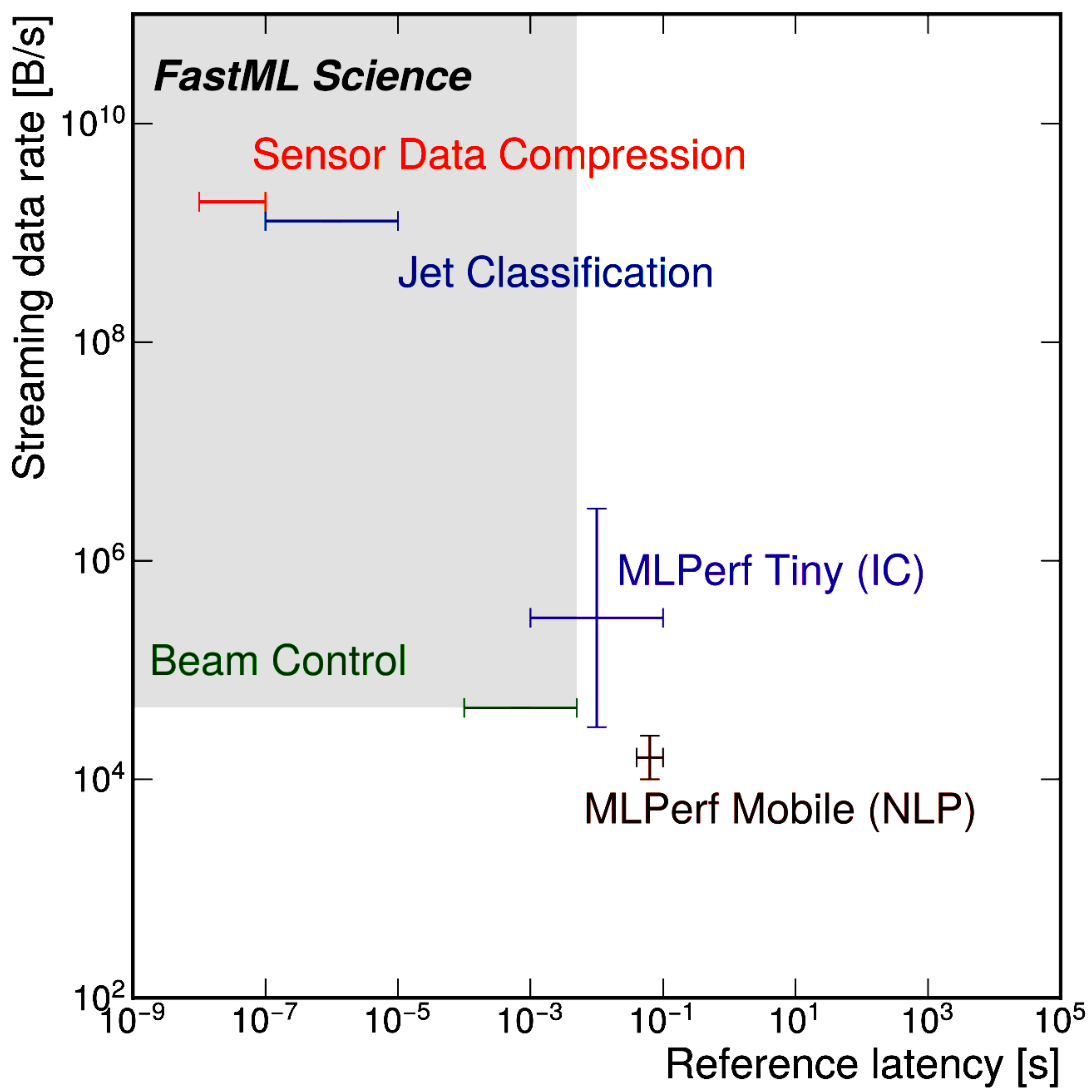
Physics-inspired data & models; Robust & generalizable learning; Fast and efficient algorithms

Computing hardware and infrastructure

Operations and control systems

Real-time AI systems at edge

“Fast” ML at the extreme edge



Cutting-edge scientific experiments explore nature at the **finest temporal and spatial scales**

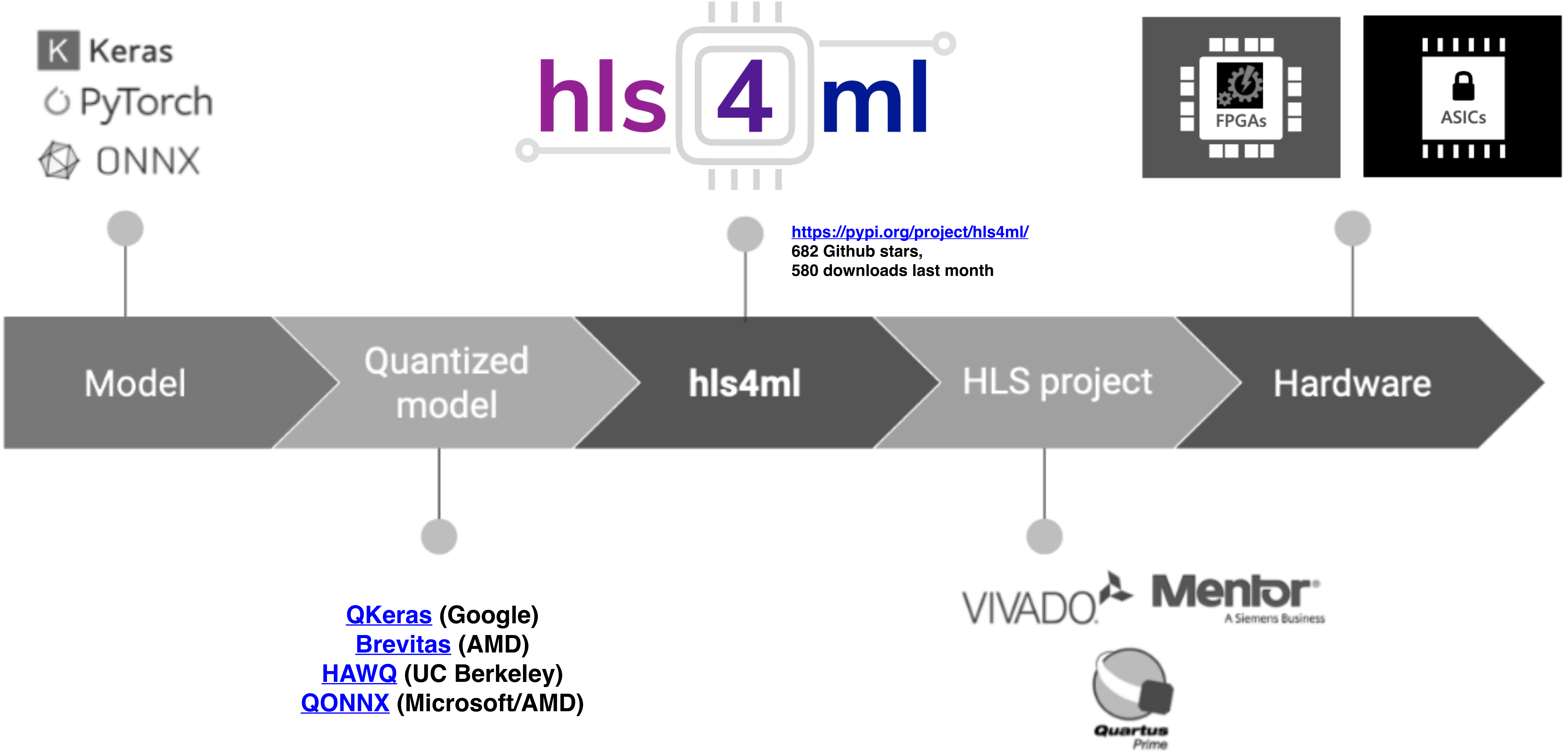
Leads to data rates far surpassing industry — requires developing **innovative techniques**

- ML in specialized embedded architectures require in ***real-time*** to reduce and filter data
- Optimal data selection enables **more efficient operation and control, saves lost data, and accelerates time-to-discovery**

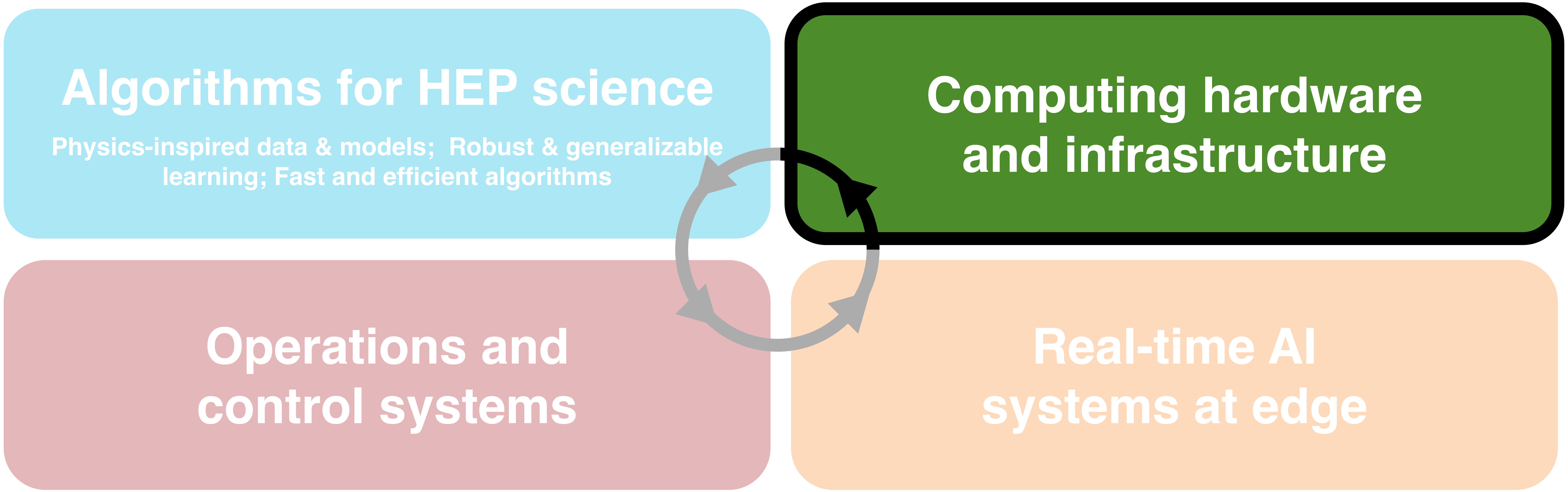
Efficient ML hardware software codesign

https://fastmachinelearning.org/hls4ml

Enabling efficient algorithms and workflows for non-experts into hardware



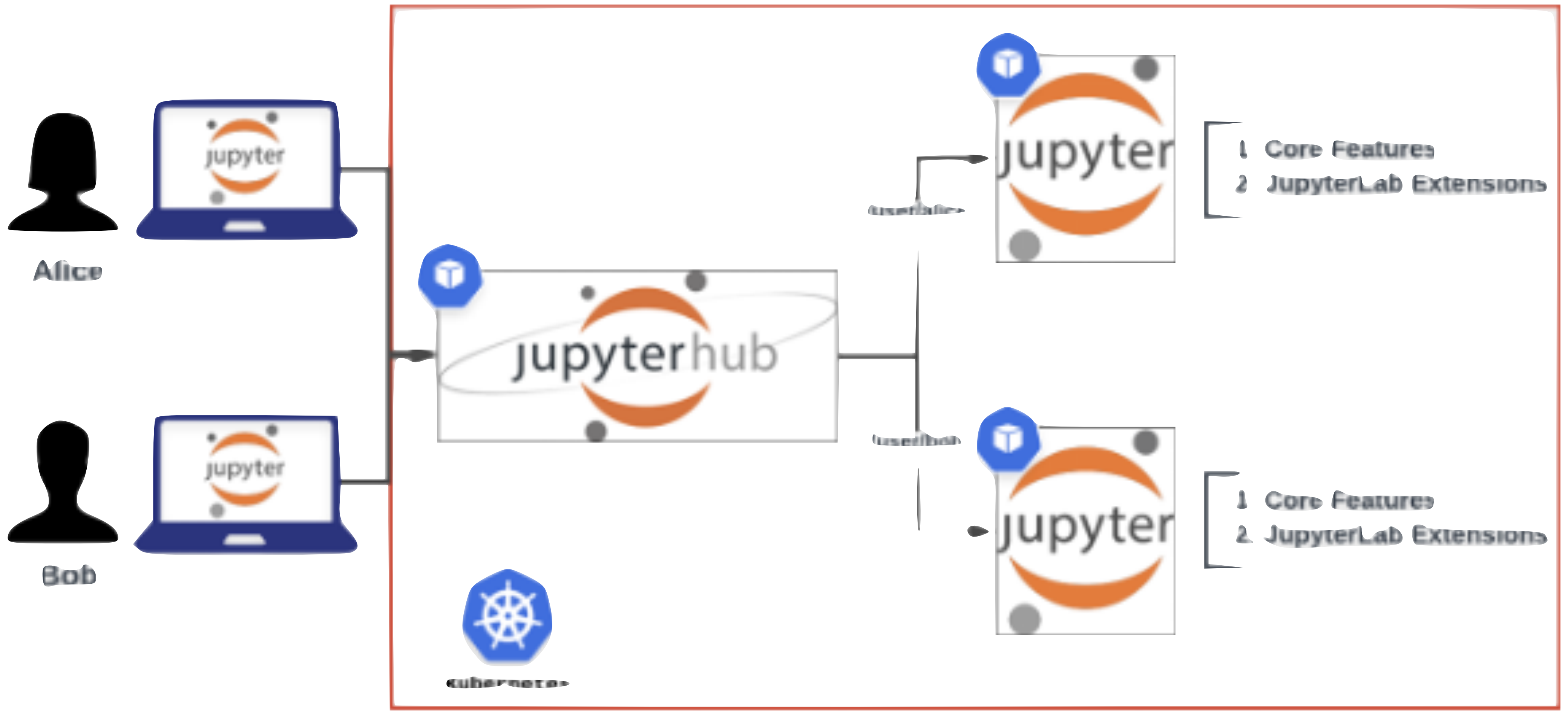
Fermilab's AI Project Portfolio

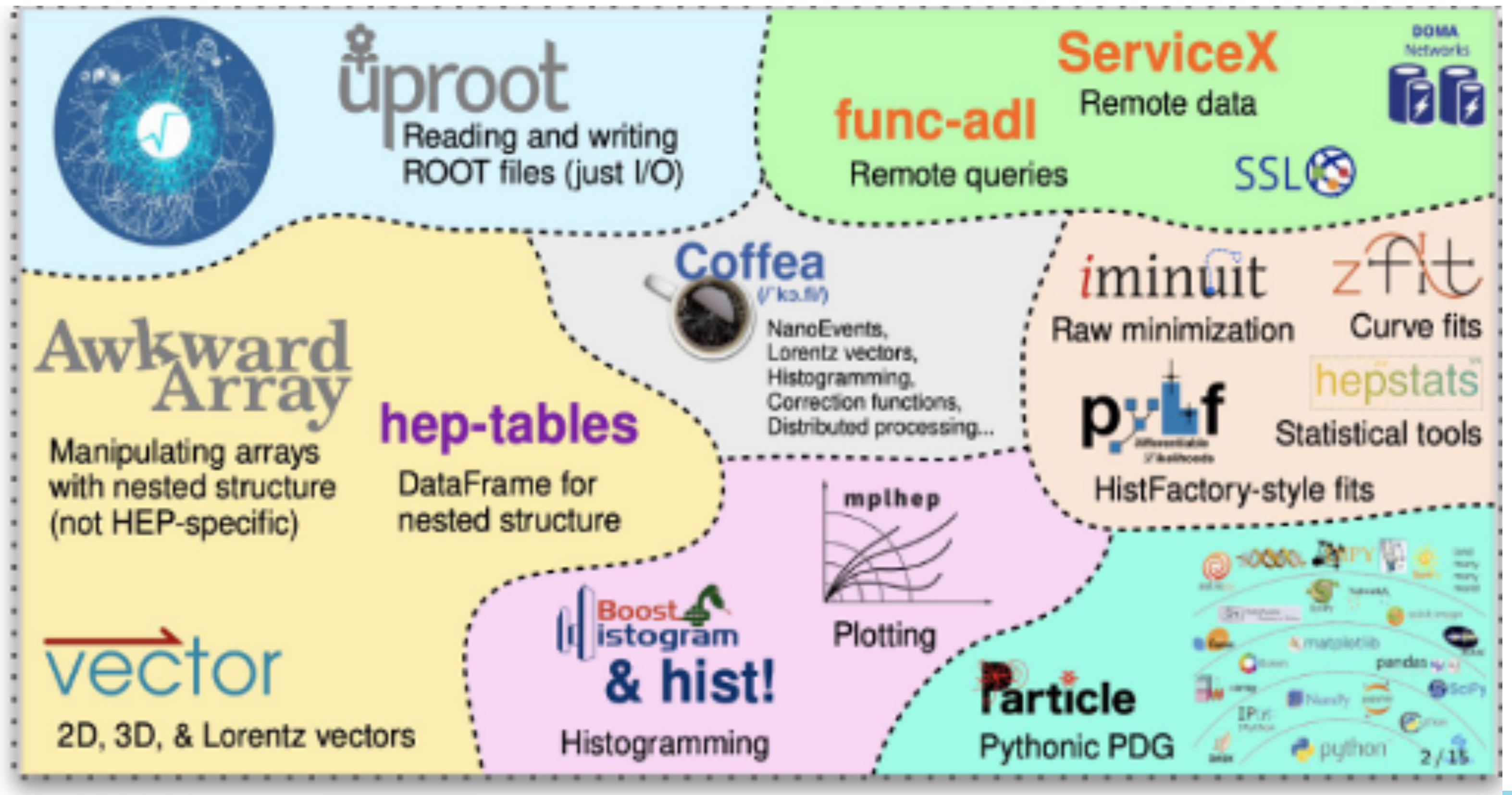


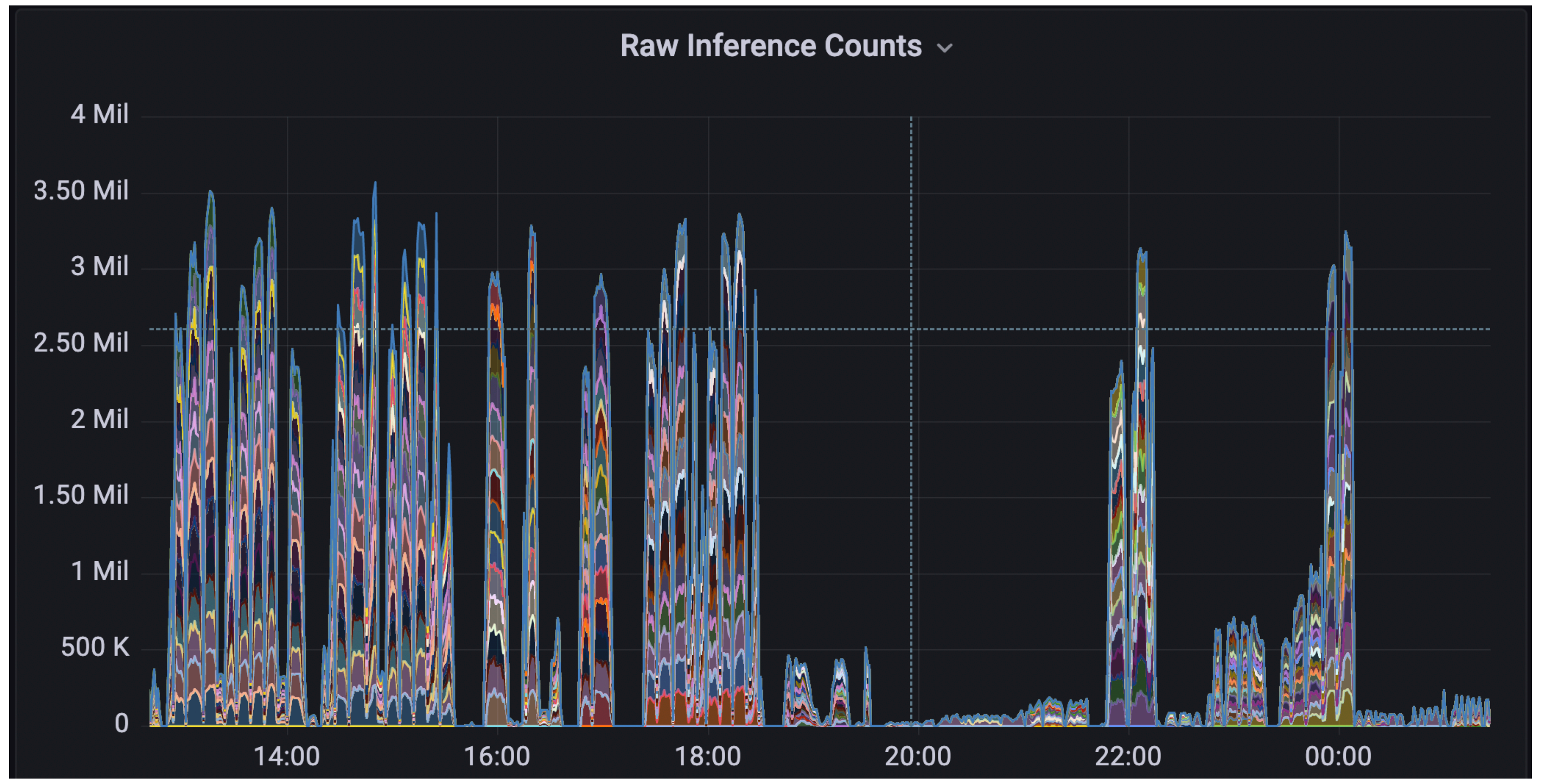
Elastic Analysis Facility & Fermilab Computing Facilities

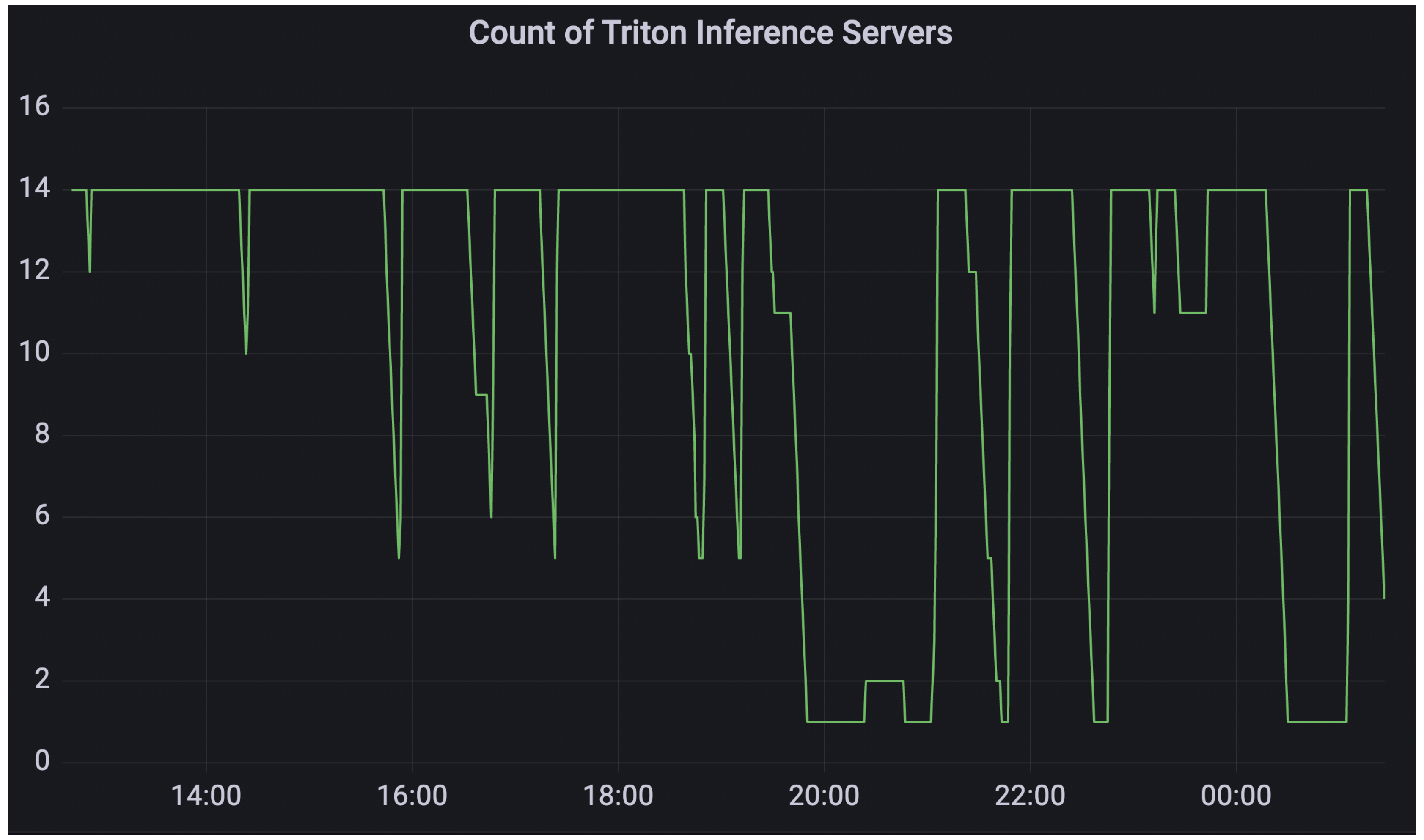
Flechas et al., arXiv:2203.10161
Benjamin et al., arXiv:2203.08010

- **Elastic Analysis Facility** @ Fermilab provides **resources** and **data-science standard industry tools** for AI training and inference
- Additional GPU resources available on CMS LPC, Wilson Cluster
- Capable of **bursting** to O(100k) batch computing CPU cores







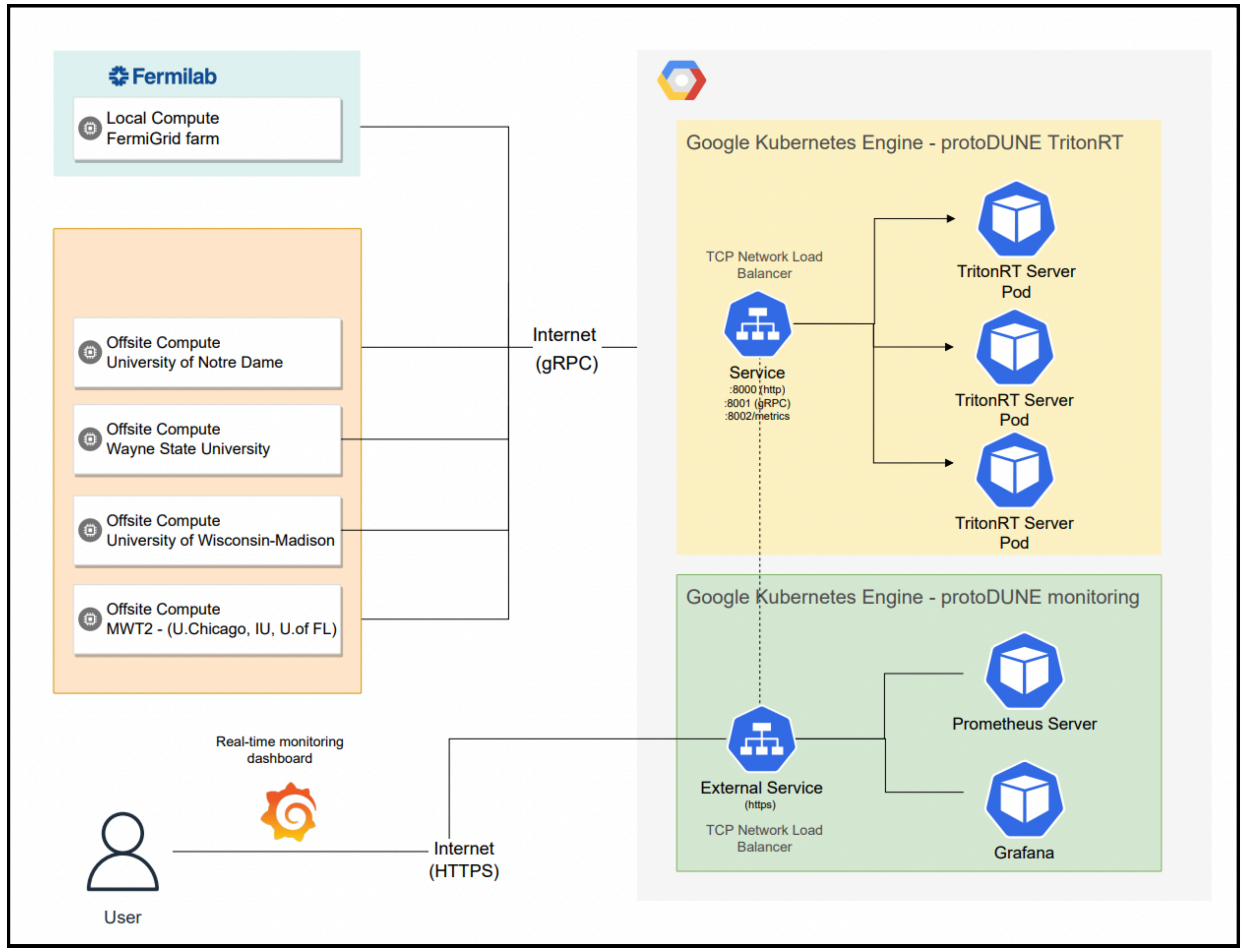


Fermilab

Accelerating ML processing

Jindariani, Ngadiuba, Pedro, Tran, Comput Softw Big Sci (2019) 3:13
Kljinsma, Pedro, Tran, Mach. Learn.: Sci. Technol. 2 (2021) 035005
Kljinsma, Pedro, Tran, IEEE/ACM H2RC 2020
Wang,Yang,Flechas,Hawks,Holzman,Knoepfel,Pedro,Tran, arXiv:2009.04509
Cai,Herner,Yang,Wang,Flechas,Holzman,Pedro,Tran, arXiv:2301.04633

- To alleviate future HEP **computing will be bottlenecks** - **enable more powerful algorithms** on optimal hardware
- **Coprocessors** (GPUs, FPGAs, ASICs, …) naturally accelerate ML workloads **by orders of magnitude**
- No way to guarantee access to HW at all production sites
- Leverage industry hardware and tools - provide **coprocessors as-a-service**

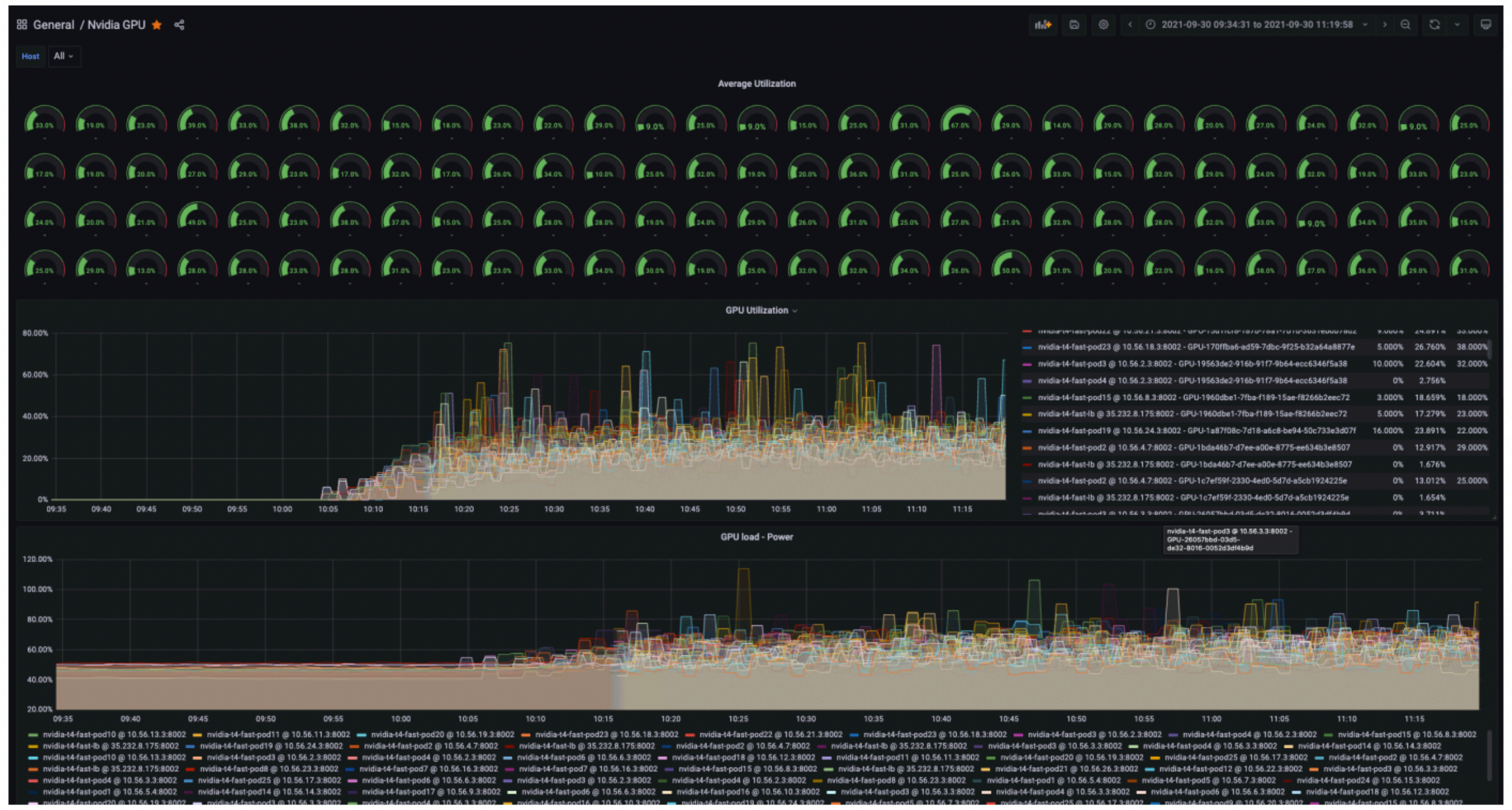


SONIC:

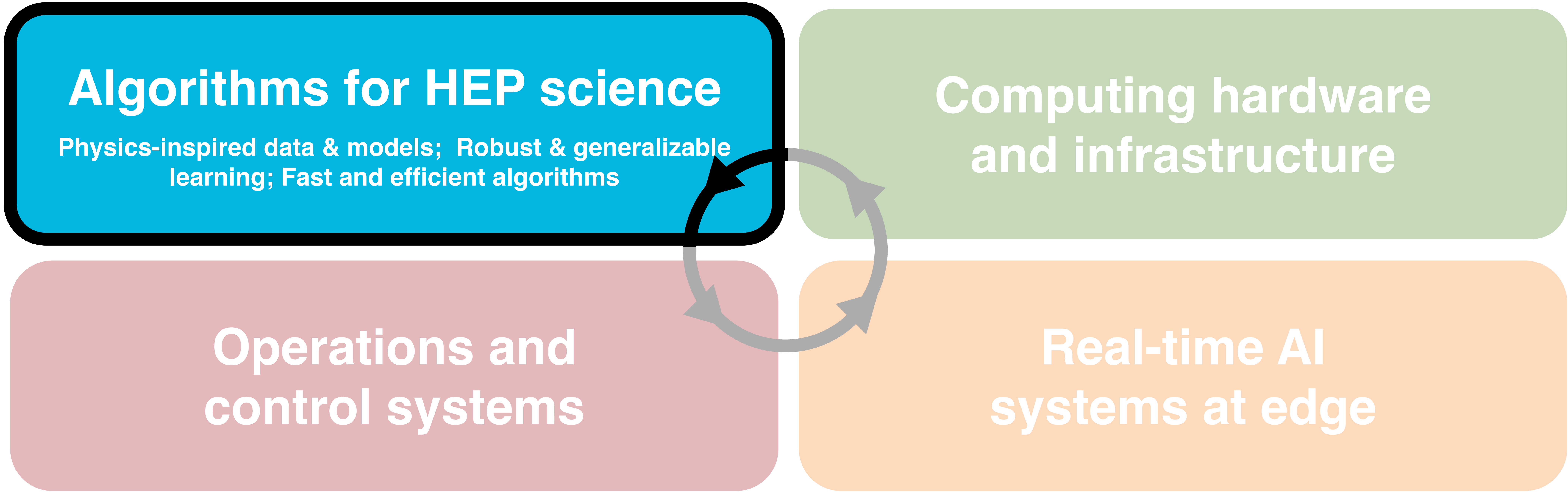
Services for Optimized Network Inference on Coprocessors

- Explore with on-prem, clouds, HPC and also for analysis facilities for all types of emerging hardware
- Testing now on CMS production workflows for Run 3
- ProtoDUNE production run (~7M) events demonstrates > 2x acceleration with GPU

Monitoring of 100 GPU run



Fermilab's AI Project Portfolio



Algorithms for HEP research projects

AI Techniques	HEP Projects	Impact
CNN	LArTPC Reconstruction	Uboldi et al, Nucl. Instrum. Meth. A 1028 (2022) 166371 ArgoNeuT JINST 17 (2022) P01018 DUNE Eur.Phys.J.C 82 (2022) 10, 903
GNN	CMS Reconstruction: HGCal, ECal, +	2x signal H->bb γγ improve 7%
SBI flexible likelihoods	Cosmic analyses	10^5 x faster
Generative models	Particle sim through matter	20-50x faster than GEANT4
Neural networks & importance sampling	Many-body schrodinger equation	Rocco et al., arXiv: 2206.10021 Issacson et al., arXIv:2212.06172
Deep Universal Domain Adaptation	Cosmic analyses, LHC Stealth SUSY background estimation	Mitigate bias, reduce hyper parameter tuning
Auto Encoders for anomaly detection	LHC QCD showers, Accelerator controls @ Linac (L-CAPE)	Pedro et al., JHEP 02 (2022) 074 Ngadiuba et al., arXiv: 2107.02157 Ngadiuba et al., Nature Machine Intelligence 4, 154 (2022) Ngadiuba et al., arXiv: 2110.08508
GNN	CMS pileup mitigation	Improve algo > 20%

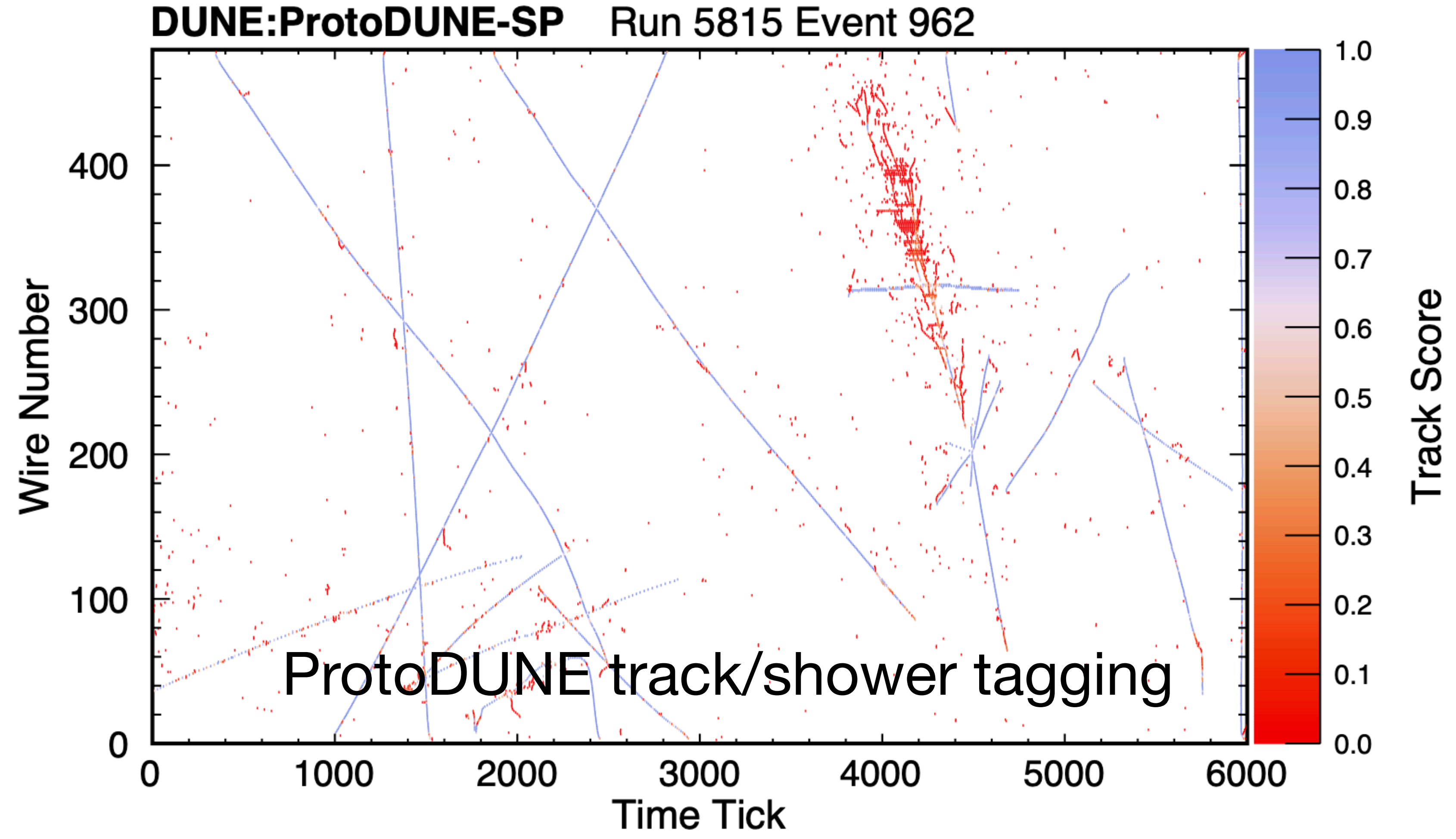
⋮ ⋮ ⋮

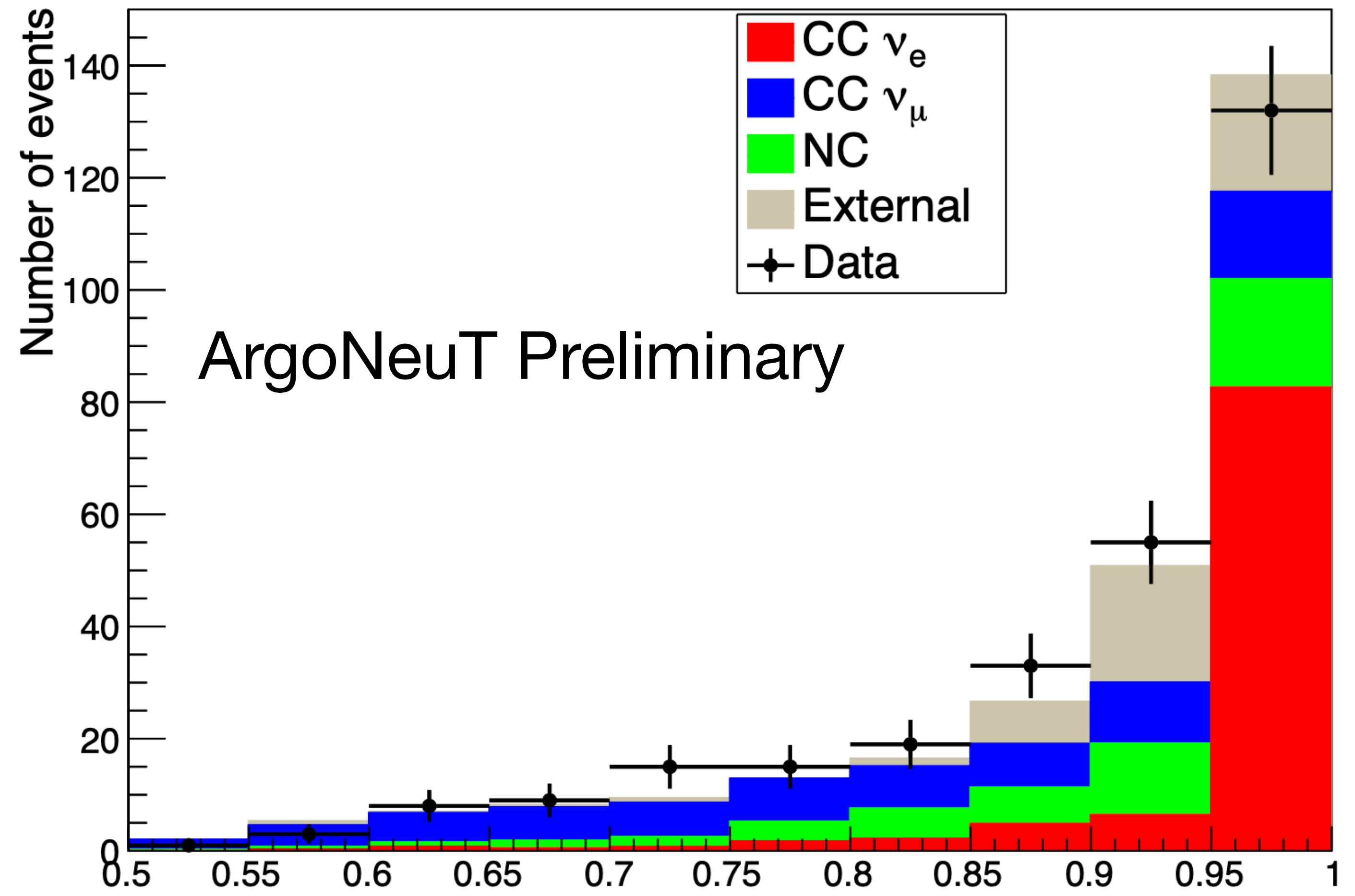
Reconstruction and pattern recognition

Uboldi et al, Nucl. Instrum. Meth. A 1028 (2022) 166371
ArgoNeuT JINST 17 (2022) P01018
DUNE Eur.Phys.J.C 82 (2022) 10, 903

Convolutional NNs to provide crucial information in neutrino interactions

- **Waveform ROI identification**
  - 1D CNN to identify signals in the raw waveforms.
  - Works for both TPC and photon detector waveforms.
- **Hit tagging**
  - 2D CNN to flag each hit as track, shower or Michel activity.
  - Validated using ProtoDUNE data.
- **Neutrino ID**
  - 2D CNN to flag each neutrino interaction as numu, nue or NC interaction.
  - Developed for DUNE and validated using ArgoNeuT data.
- **MicroBooNE open data!**
  - A tool for collaborative AI developments
  - https://microboone.fnal.gov/documents-publications/public-datasets/



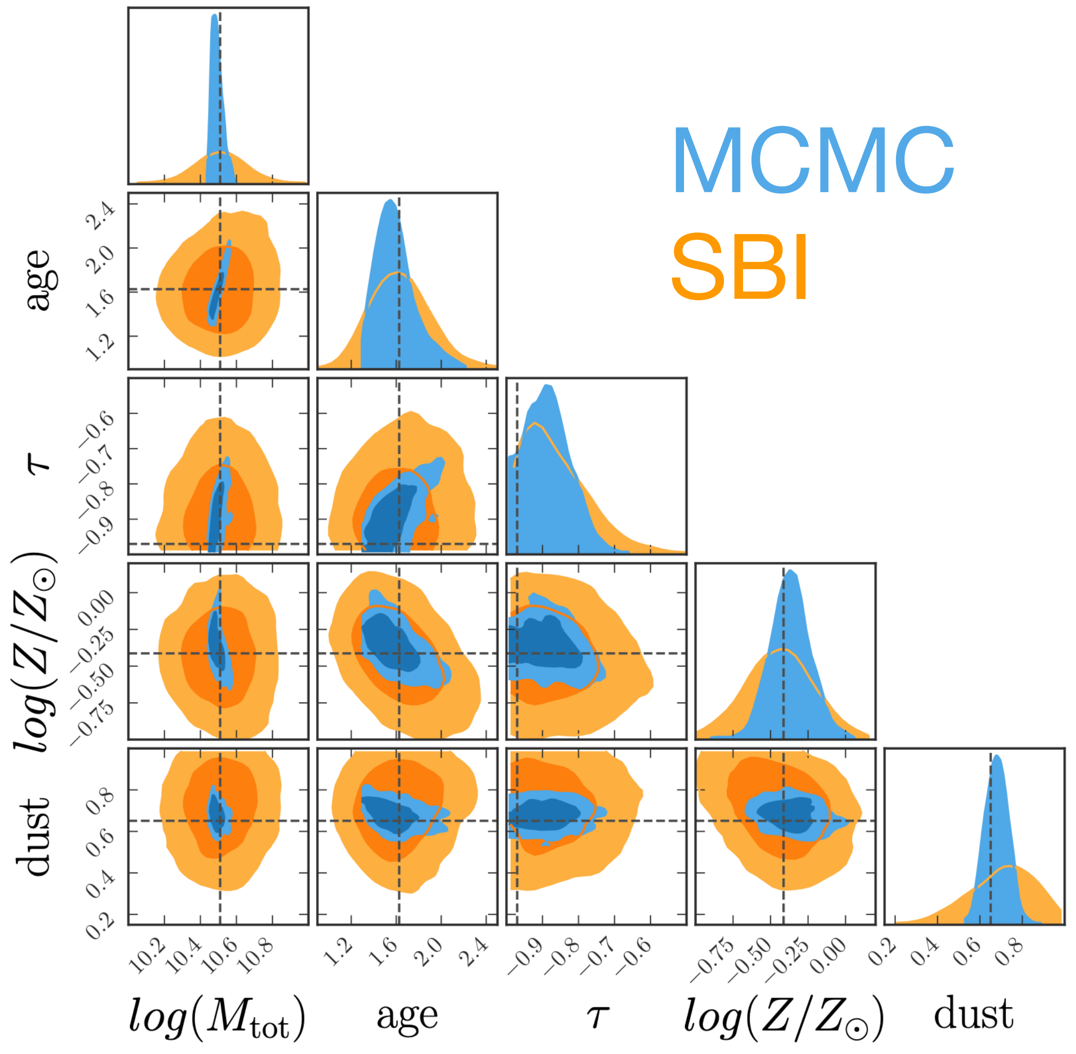


Fermilab

Simulation-based Inference (SBI) for Cosmic Analysis

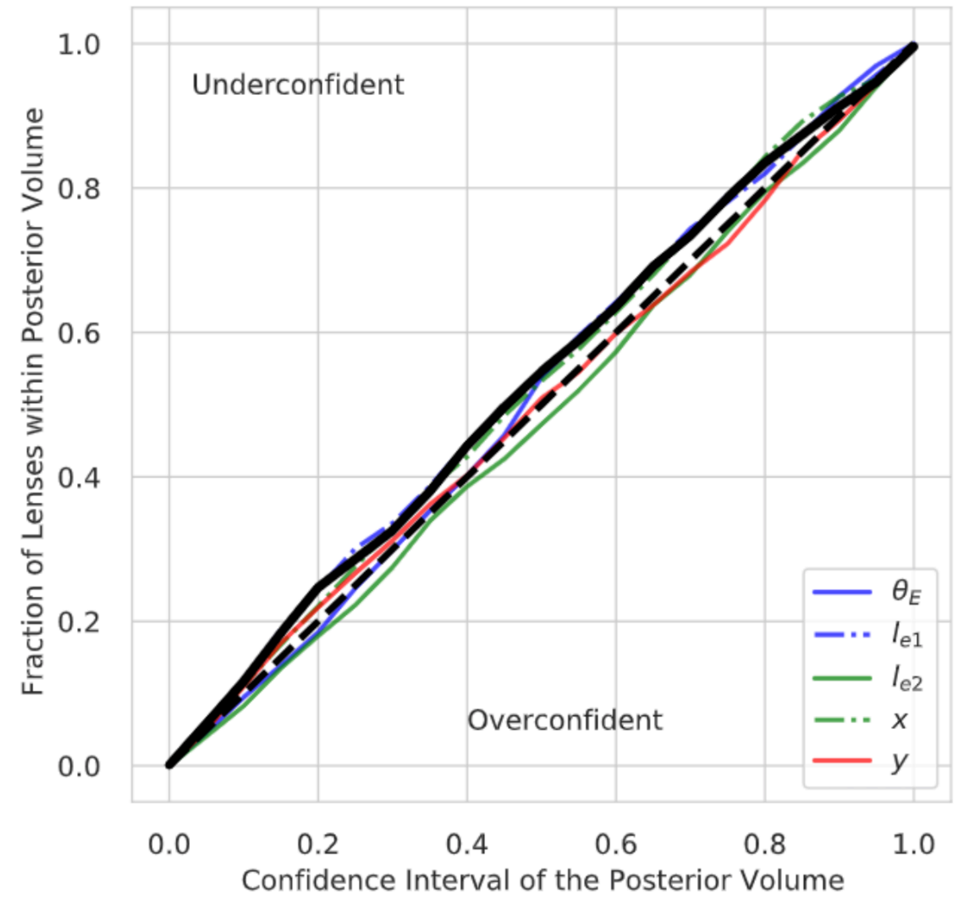
Nord ECA
Galaxy Spectra
Khullar, Nord, Ciprijanovic, Poh, Xu 2022 (MLST & Neurips)
Strong Lenses
Poh et a.l, 2022 in Neurips Workshop

- **Goal:** Maximal information extraction from high-dimensional data to **rapidly find/measure** objects, dark energy, dark matter
  - **Traditional methods use explicit analytic functions** with simplified assumptions; typically **slow** and **inaccurate**
- **Forward modeling and SBI** permits flexible likelihoods
  - Simulated datasets until matching observation
  - Can be 10^5 times faster than traditional methods
- Applications across many surveys (DES, LSST, CMB-S4) and objects (Strong Lenses, Spectra, Quasars, Galaxy Clusters)
  - Connections across all of HEP



Proof-of-concept:
Simple SBI method (not highly tuned) is just as accurate as MCMC, but much faster

SBI shows correct level of confidence in estimates.

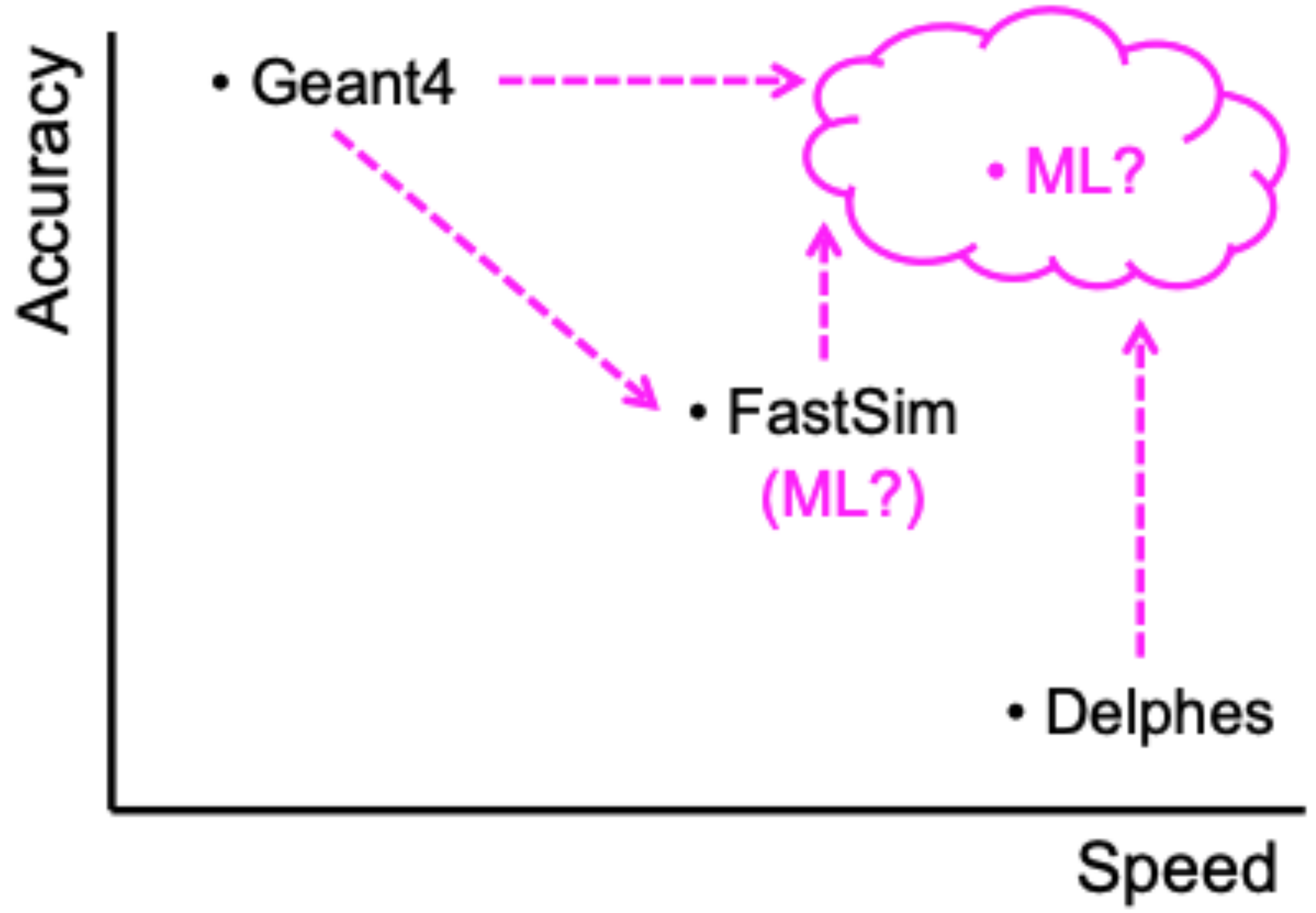


Fermilab

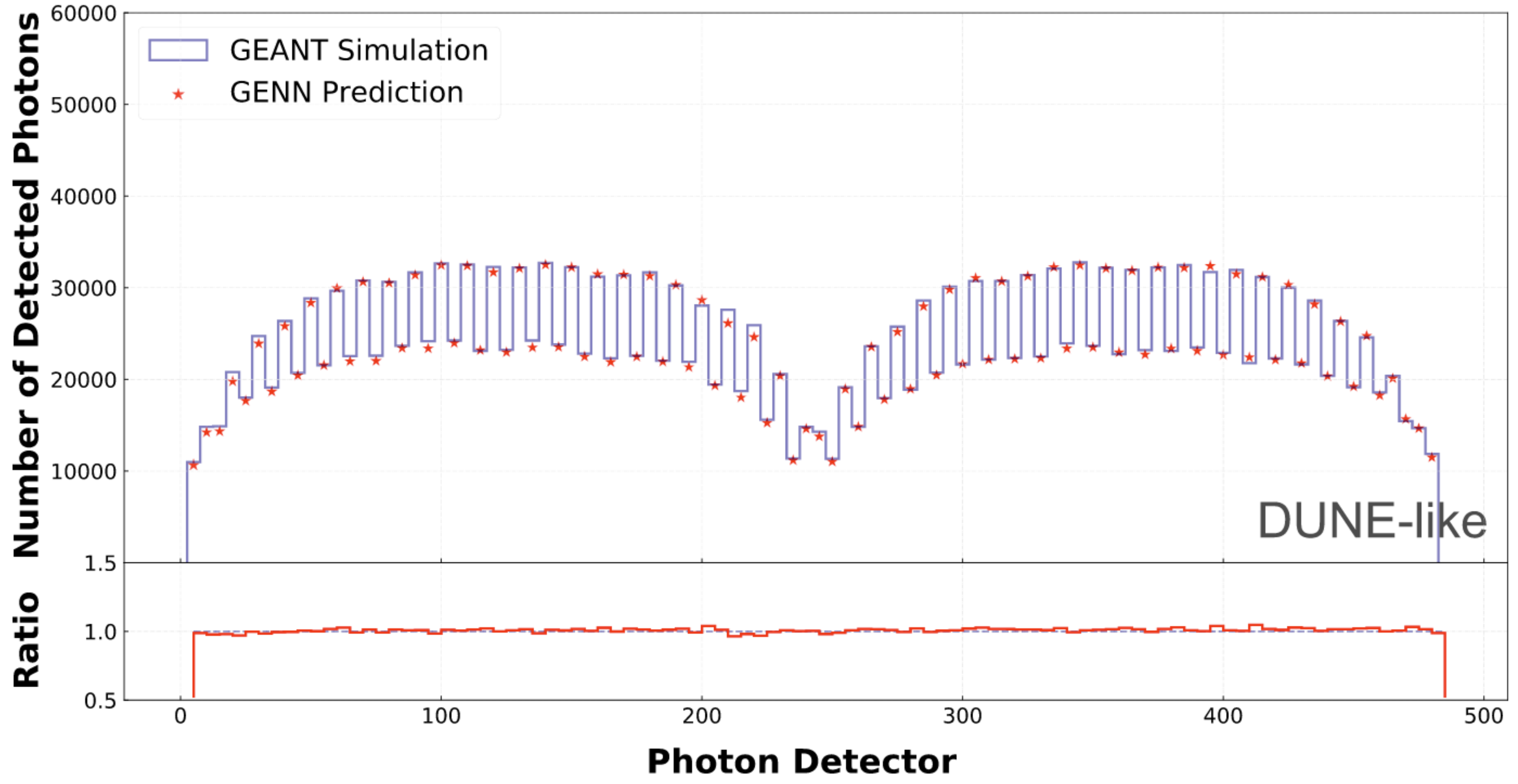
Generative models for simulation

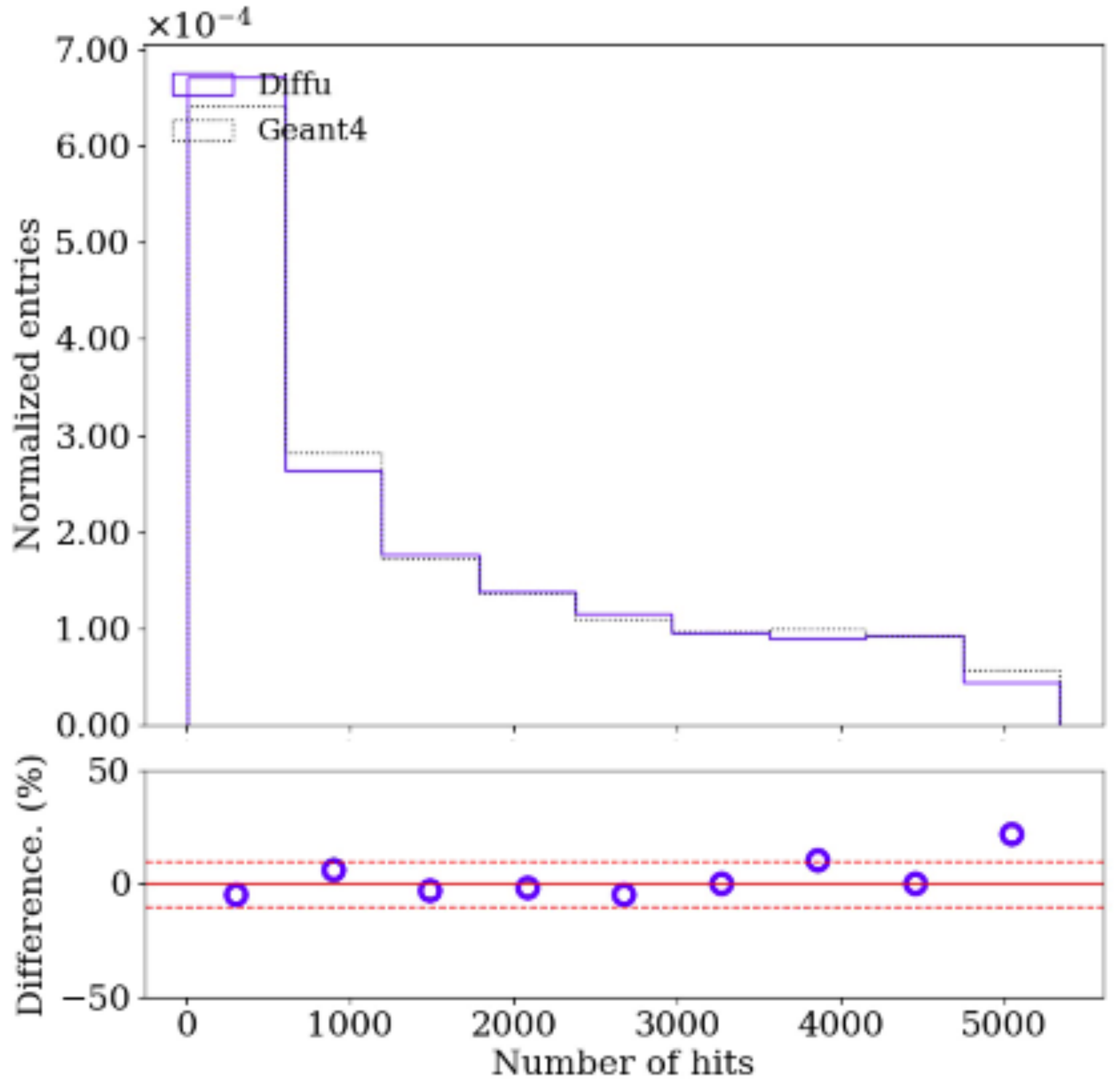
Pedro et al., arXiv:2202.05320, ACAT2021
Pedro et al., arXiv:2203.08806
Mu, Himmel, Ramson, Mach. Learn. Sci. Tech. 3 (2022) 1, 015033

- **High fidelity ML-based parameterized simulation** to mitigate computing bottleneck for DUNE and LHC
  - Find way to fuse GEANT full-sim with ML
  - More naturally run on coprocessors
- GENN for photon transport simulation
- Stable diffusion (CaloDiffusion) for LHC calorimeter



20-50 times faster than Geant4 simulation





Diffusion model: avoids pitfalls of GANs, high quality output

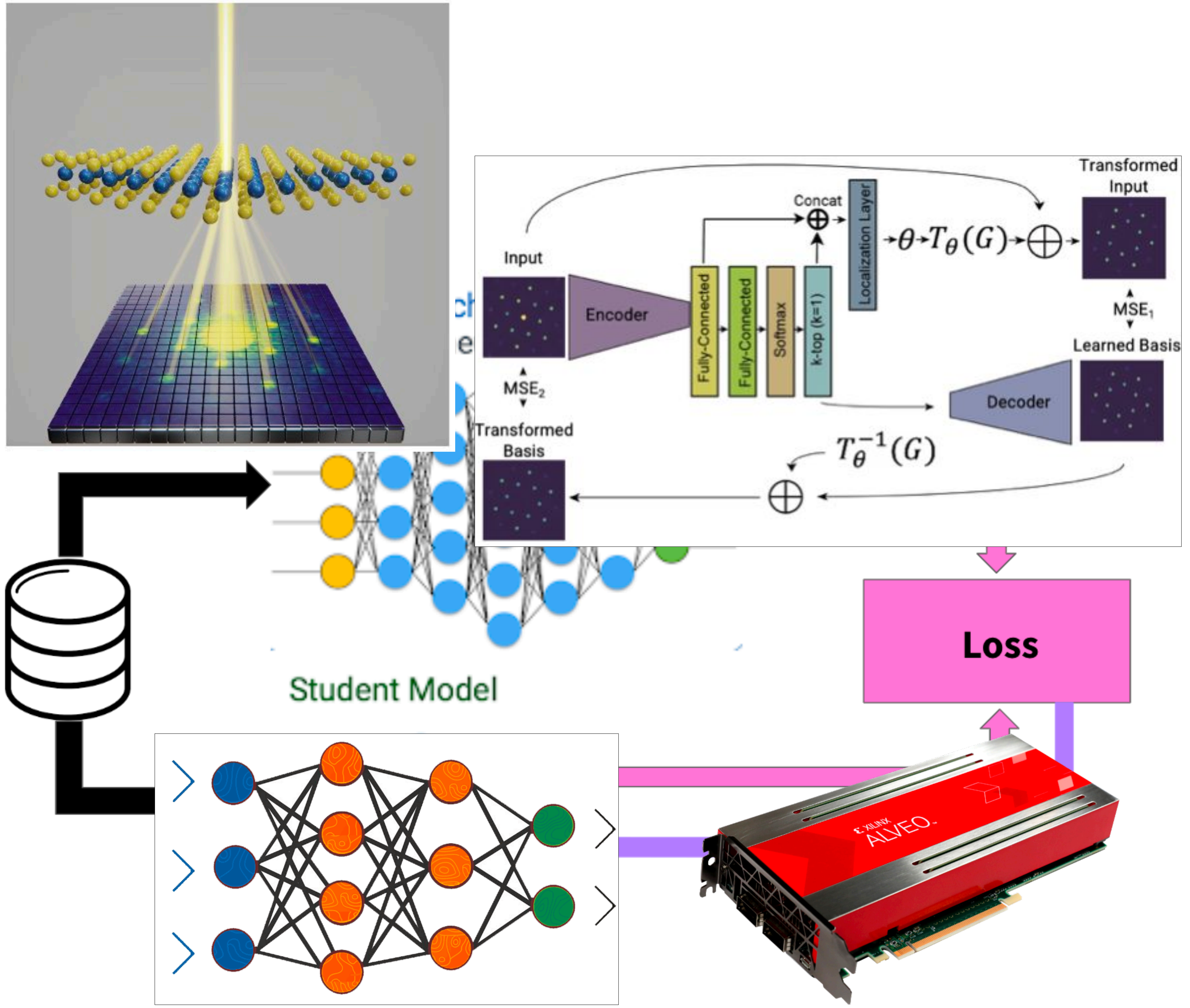
Competitive results on the CaloChallenge dataset

Fermilab

Fast and efficient algorithms

Hawks, Tran, Quantization-aware pruning, arXiv:2102.11289
Mitrevski, Hawks, Muhizi, Tran, QONNX, arXiv:2206.07527
An end-to-end codesign workflow of Hessian-aware quantized neural networks for FPGAs and ASICs
Campos, Hawks, MItrevski, Tran
Quantized Distilled Autoencoder Model for 4D Transmission Edge Microscopy
Forelli, Muhizi, Tran

- **Real-time and efficient AI**: driver for scientific sensing/compute
  - Core research into quantization and sparsity and optimization techniques
  - Important for hardware implementation (more on this later)
    - Developing training frameworks for quantization-aware AI and hardware translation
    - QONNX - build industry standards - interchange formats for quantized AI
- Building techniques for broader scientific community
  - Quantized model distillation for microscopy



Fermilab’s AI Project Portfolio

Algorithms for HEP science

Physics-inspired data & models; Robust & generalizable learning; Fast and efficient algorithms

Computing hardware and infrastructure

Operations and control systems

Real-time AI systems at edge

AI for Operations and Control Systems

Cosmology	Quantum	Accelerator Controls
- Experiment automation for self driving telescopes (GNN & RL) - instrument design (replace expensive optics simulations with SBI and decision trees)	- AI/ML for controlling & optimizing quantum computers with micro electronics and edge AI - Theoretical & experimental work on quantum detectors	- Linac RF optimization (prevent the need for constant tuning to reduce beam losses at injection to Booster) - Booster GMPS (reinforcement learning agent on FPGA to supplement traditional PID loop) - Real-time Edge AI Distributed Systems (READS) • Disentangle Main Injector and Recycler Ring beam losses with a U-Net • Increase muon resonant extraction spill uniformity for Mu2e with reinforcement learning

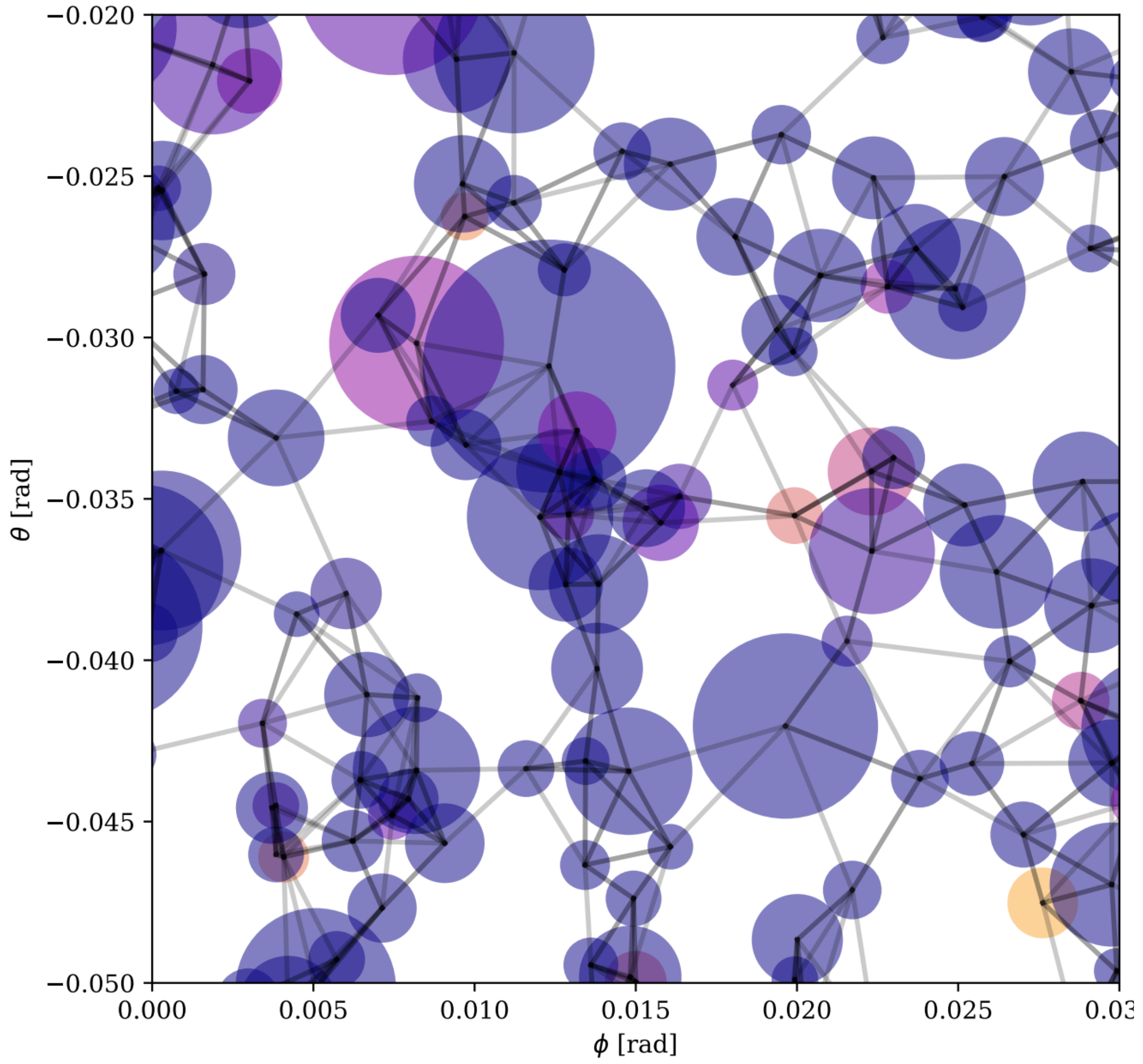
Fermilab

Automation for cosmology experiments

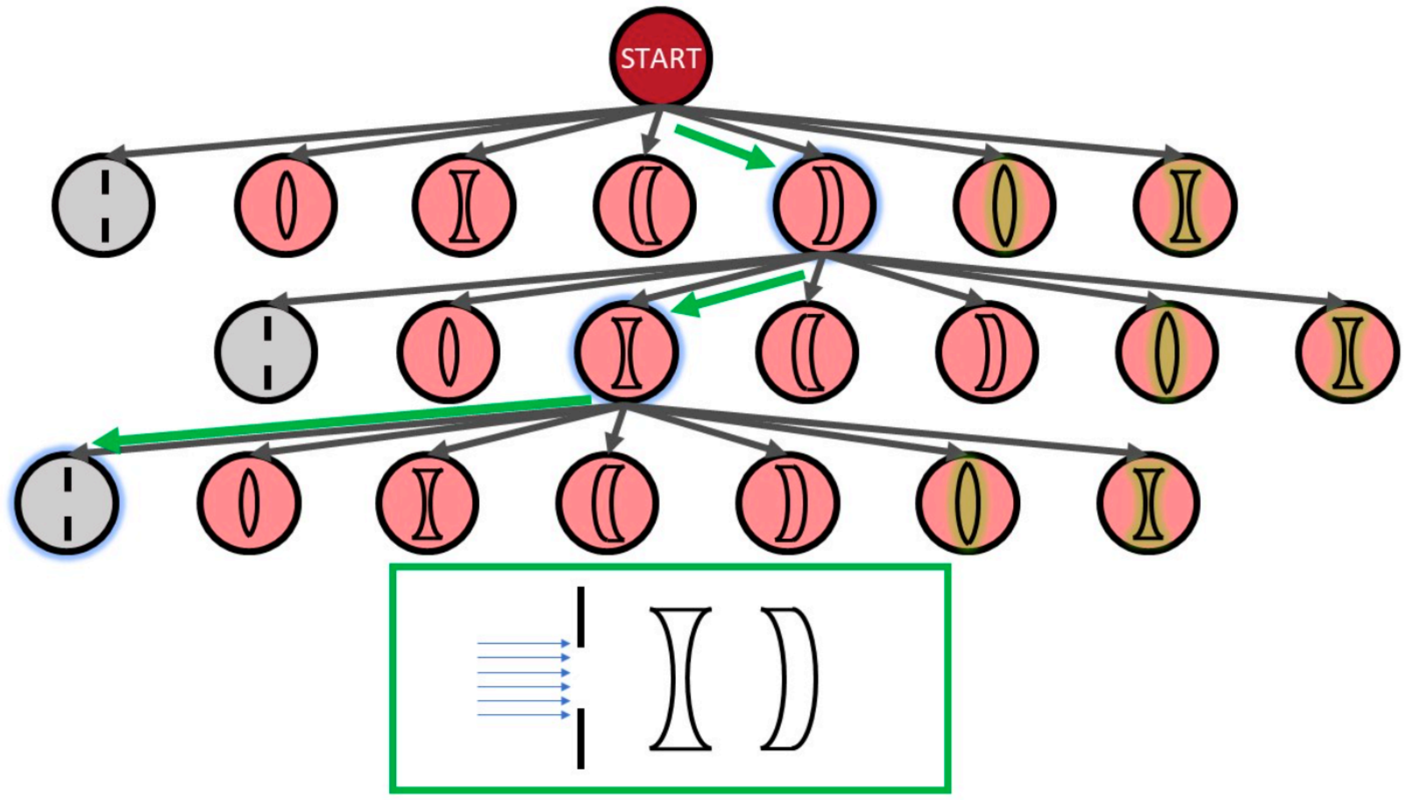
Spectroscopic Survey Optimization
Cranmer, Melchior, Nord, 2021 (Neurips workshop)
Optical System Design
Cohen (HS student) and Nord, 2023 (in prep.)

Self-driving telescopes:
Adaptive optimization for survey scheduling

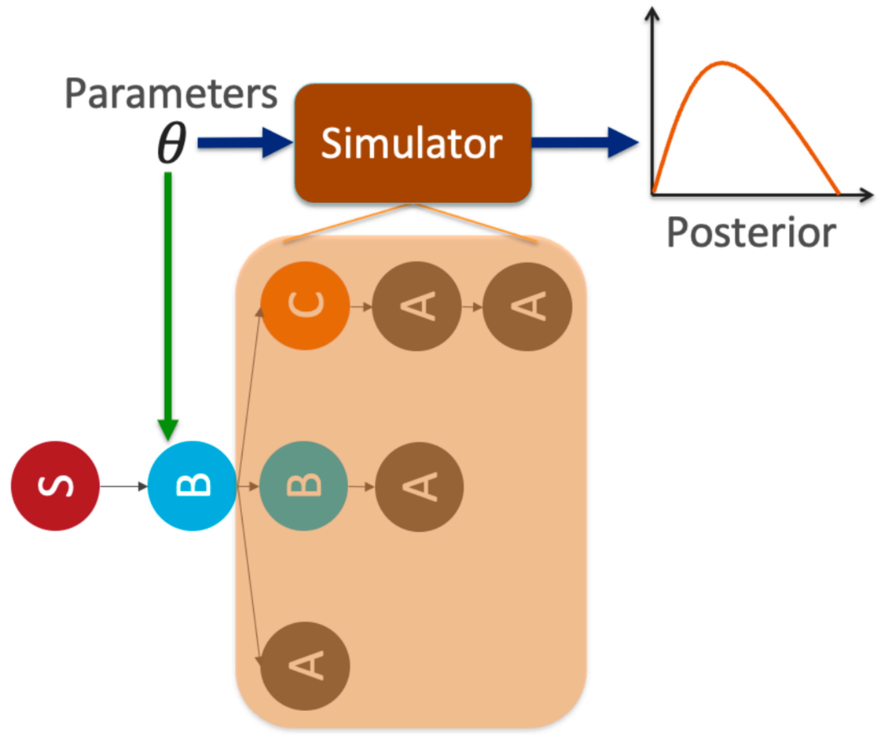
- **Unsupervised Graph Neural Networks**: optimize an observation strategy to constrain cosmological parameters
- **Supervised Reinforcement Learning**: build a decision-making algorithm to prepare or adapt observations



A network of galaxies optimally selected for cosmic matter estimation



Schematic example of generating an optical system - Green arrows show optimized tree traversal



Overview: tree produces optical system; posteriors are of element shape parameters

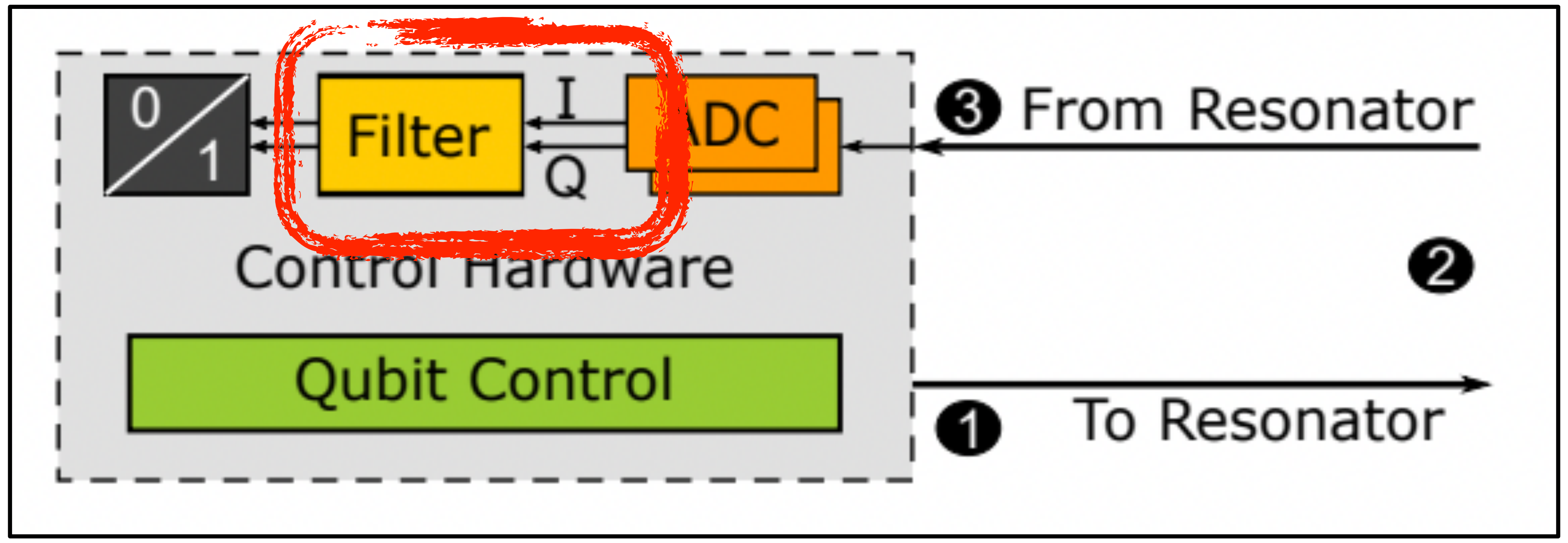
Automated instrument design: replace expensive optics simulation

Use decisions trees + simulation-based inference to *arrange optics* and *choose optical element*

Fermilab

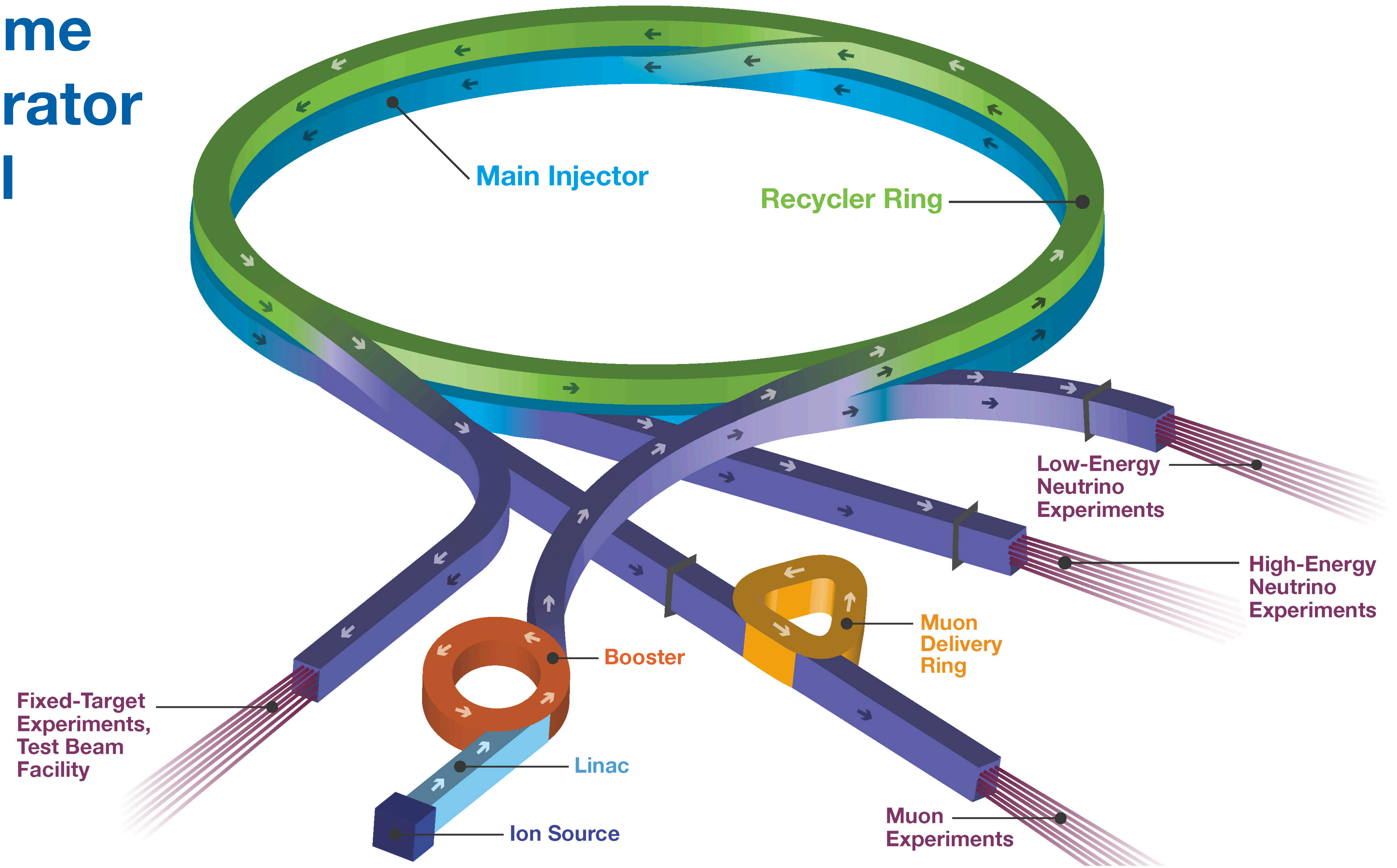
Practical QML (QC and QQ) at Fermilab

- AI/ML for controlling and optimizing quantum computers
  - Exciting effort couples to microelectronics and edge AI applications to improve quantum readout
  - Classical AI for de-noising quantum computations in theory calculations and event generators — QuantISED program studying quantum computing for neutrino scattering calculations
  - Classical AI for predicting quantum circuit fidelity on noisy hardware - important for HEP field theory problems involving extremely deep quantum circuits



- Quantum AI for quantum data
  - Exciting efforts involve theoretical work on enhancing the sensitivity of quantum sensors connected by a quantum network (SQMS and FQI).
    - Very early days although proof of principle theoretical and experimental work has been done on optical test benches.
  - Quantum ML techniques for enhancing signal extraction from quantum simulation (FQI, joint with U. Trento, CERN).
    - No clear advantages discovered yet - may be a hammer searching for nails, but potentially interesting.

Real-time accelerator control



Fermilab

Real-time accelerator control

Main Injector

Recycler Ring

High-Energy Neutrino Experiments

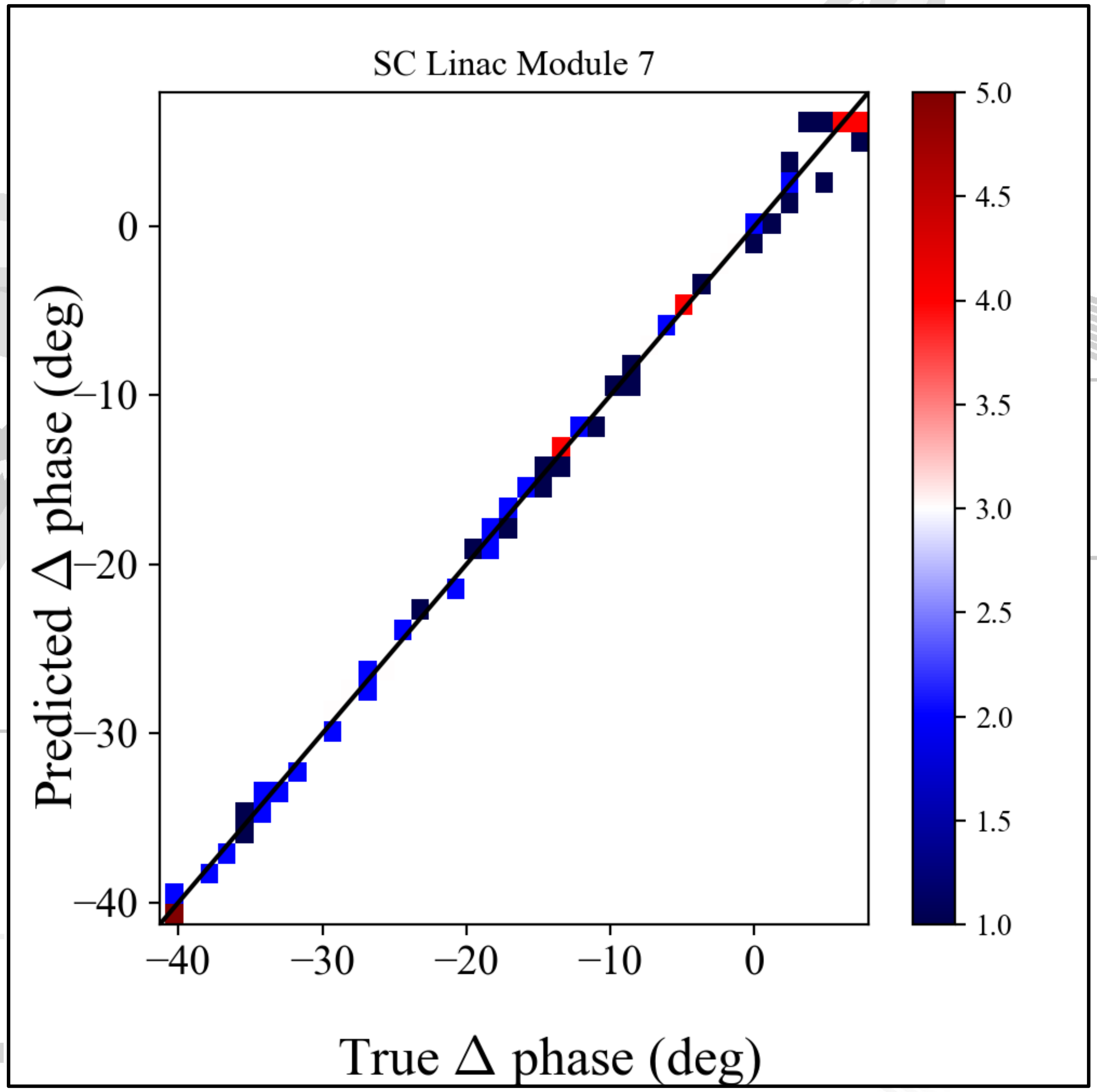
Fixed-Target Experiments, Test Beam Facility

Ion Sou

Linac RF optimization

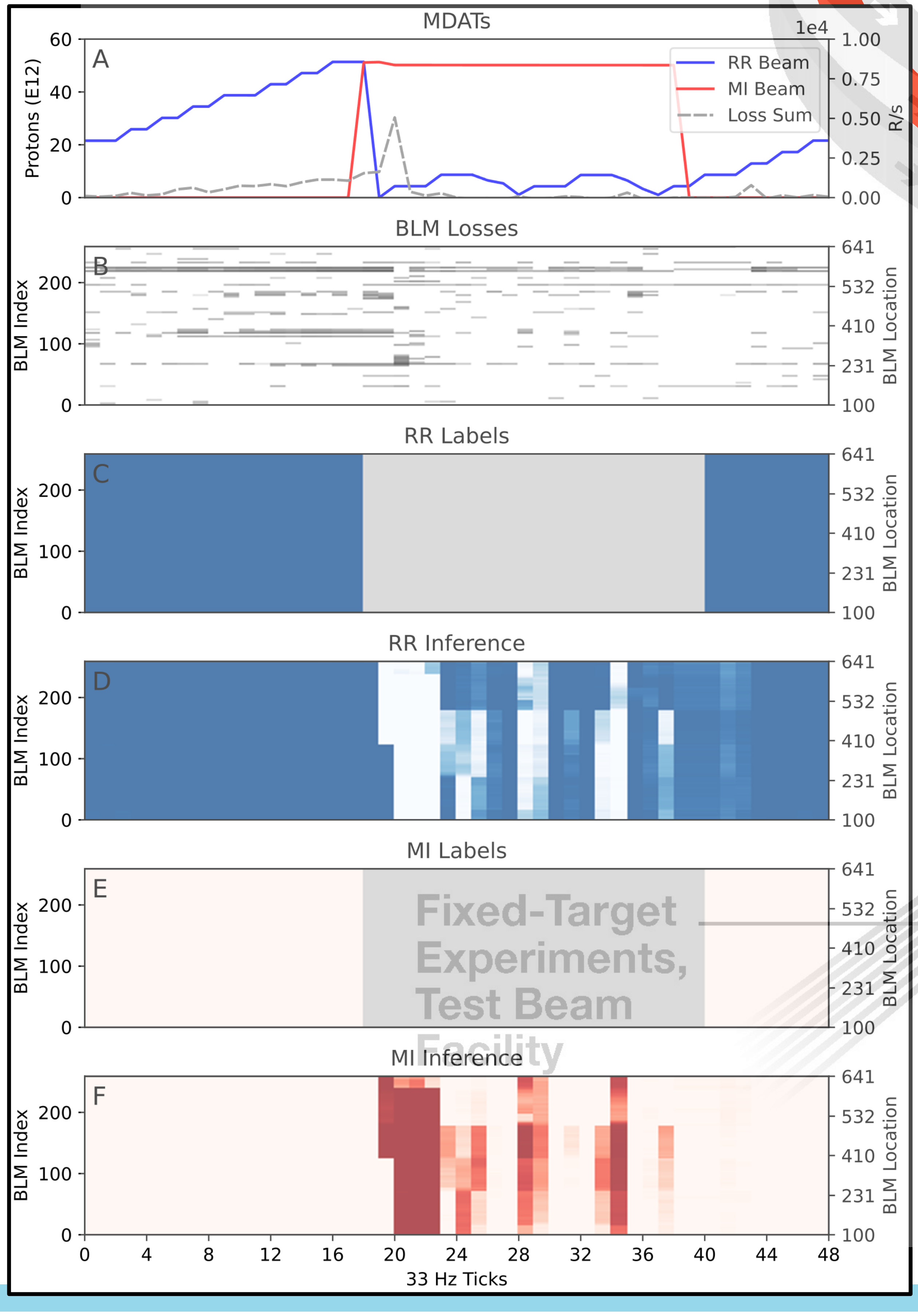
Predict RF parameters to keep beam energy constant and minimize emmitance

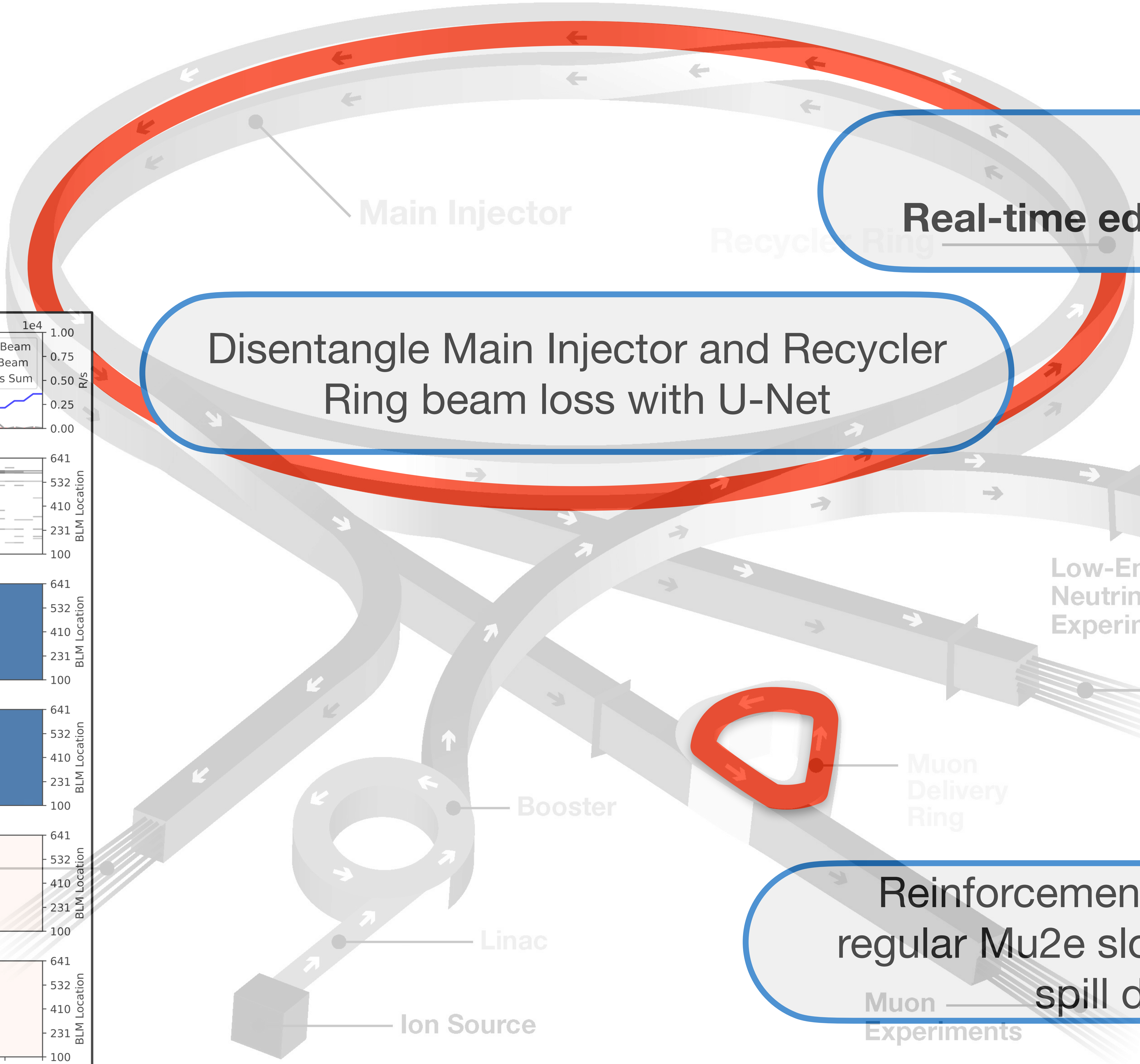
Proof-of-concept with single cavity phase regulation; multi-cavity promising



Fermilab

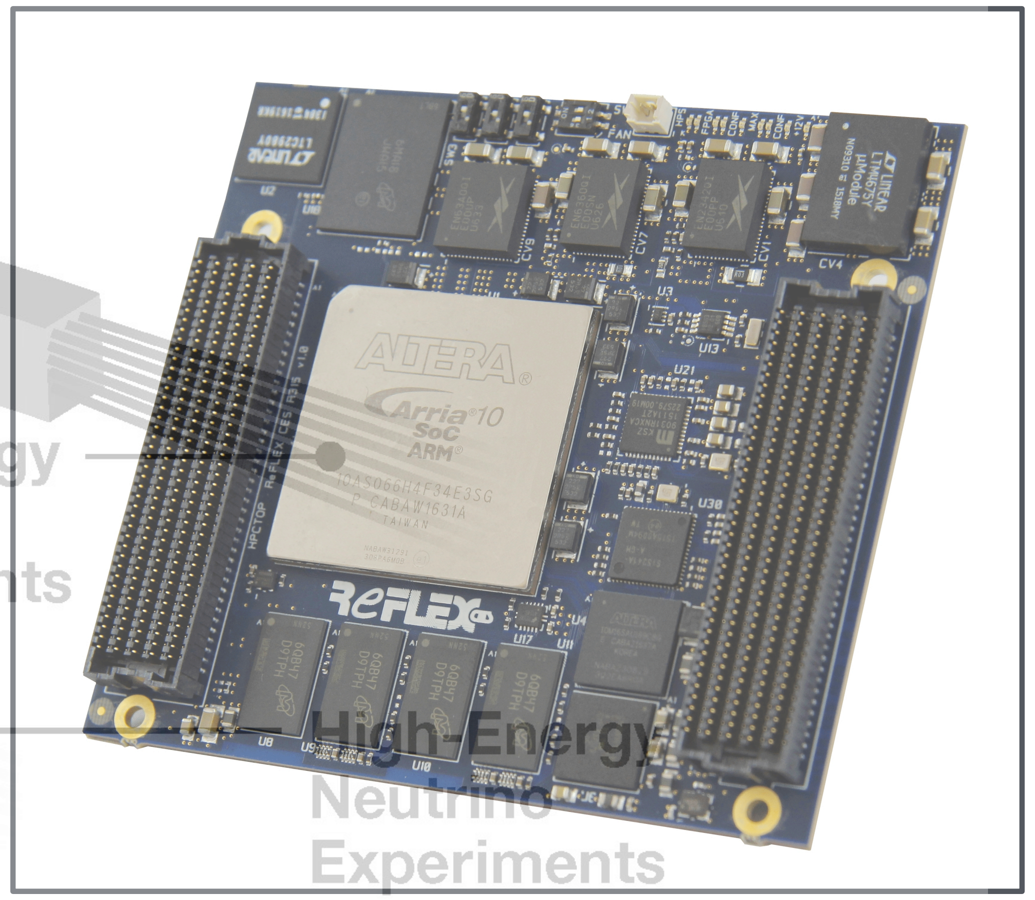
Real-time accelerator control





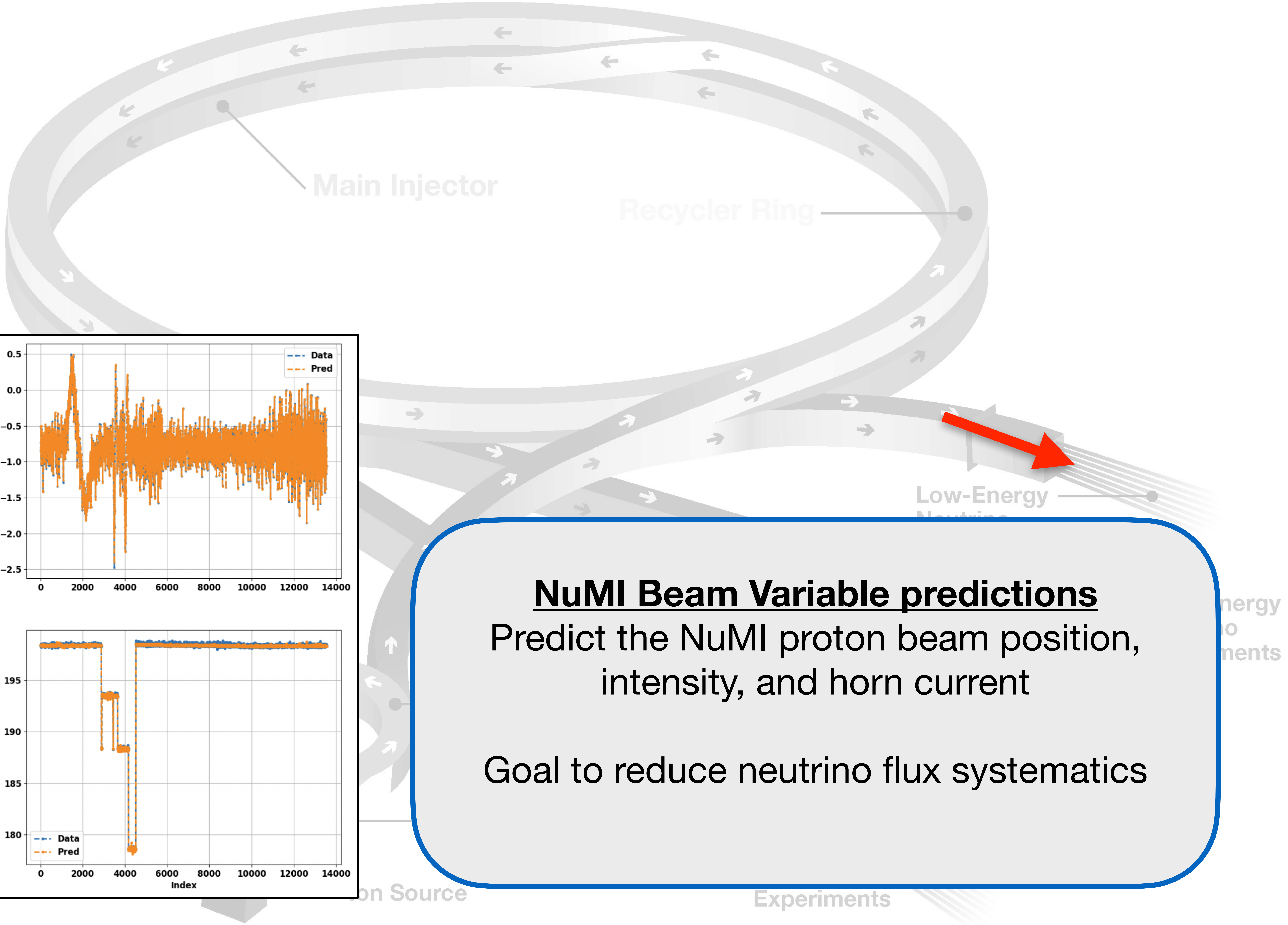
READS
Real-time edge AI distributed system

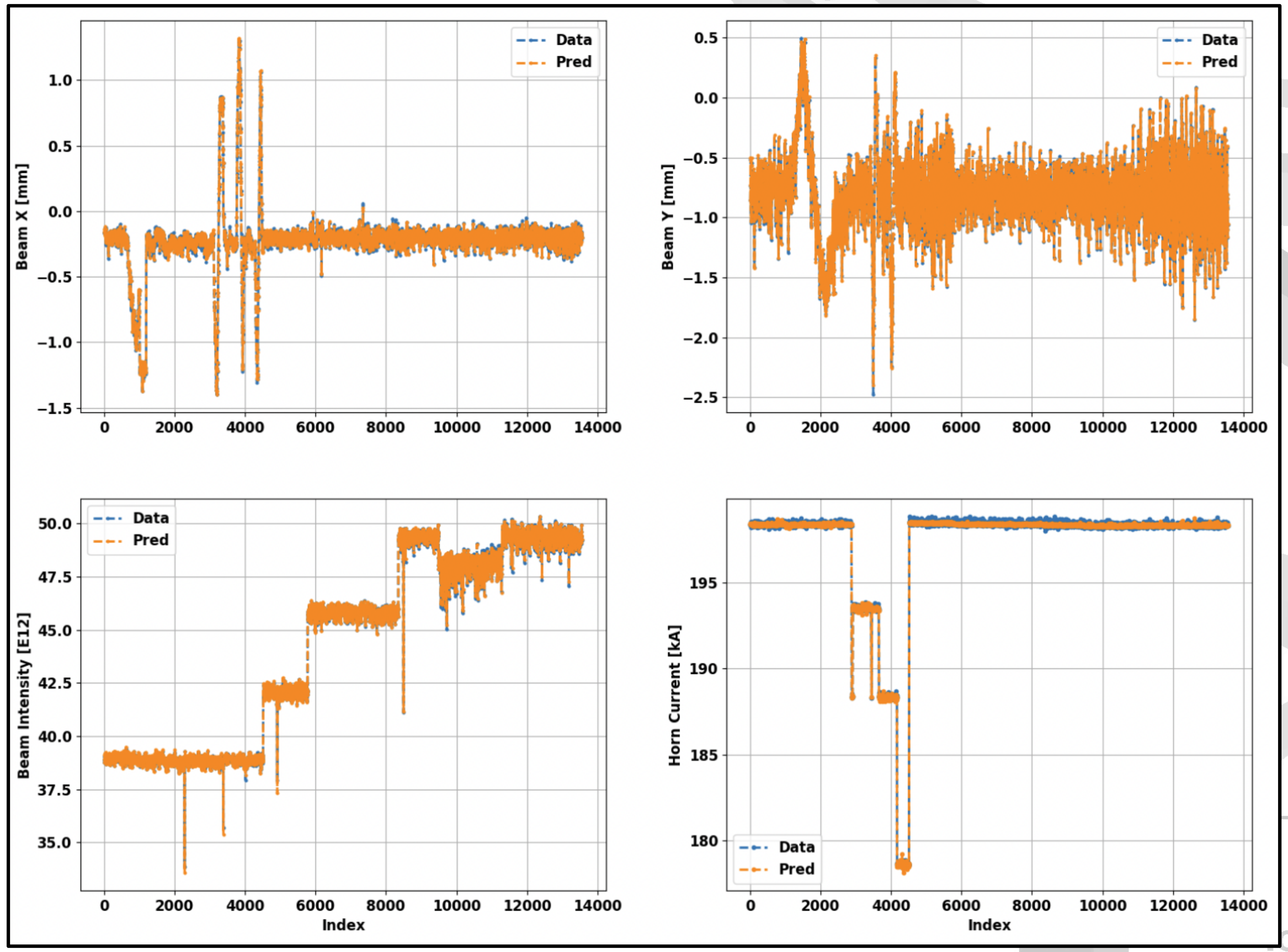
Disentangle Main Injector and Recycler Ring beam loss with U-Net



Reinforcement learning agent to regular Mu2e slow spill and increase spill duty factor

Real-time accelerator control





NuMI Beam Variable predictions

Predict the NuMI proton beam position, intensity, and horn current

Goal to reduce neutrino flux systematics

Fermilab

Extensive development in Fermilab's Accelerator Controls Department for MLOps

- Standardized accel. data logger & filter.
- Standardized format.
- Interface for ML Engineer: data filter.
- Dataset Management System
  - Versioning
  - Track derivative datasets
  - Metadata

- Use Common Tools
- Model Development System
  - ~MLFlow / hyper p. tune
  - VC: model with references to env., data, results

- Interfaces for ML Engineer:
  - Getting "live data"
  - Setting "actions"
  - Monitoring input, model predictions/performance
- Model Registry
  - All data, env., model, and performance assets

- Server with proper specifications
- Monitoring services
- Control services
- (Logging services)
- Automate deployment of model assets from Model Registry

Extensive development in Fermilab's Accelerator Controls Department for MLOps

- Collaborations across National Labs
  - SLAC
  - BNL
  - JLAB
  - Oak Ridge
  - PNNL

- Internationally
  - CERN
  - European Spallation Source, Sweden
  - JPARC

AI Project Office supports AI activities across Fermilab Physics

Fermilab's AI Project Office is the glue connecting our AI research so we can support each other

AI Project Office supports you!

Computing Resources

- Elastic Analysis Facility: https://analytics-hub.fnal.gov
- Wilson/Institutional Cluster: https://computing.fnal.gov/

Community Building

- Workforce development, AI Researcher job family
- Future AI Jamboree
- Engage broader AI & HEP community
- Foster existing and growing collaborations with laboratories, universities, industry

Advice/education

- Seminar series, Tutorials
- Lab-wide AI meeting

Coordinate responses to AI grants

- Tactical and strategic planning with lab
- Align proposals with lab strengths

Invitation to join Fermilab's AI Project!

Announcements:
ai-project@fnal.gov

Lab-Wide AI Seminars:
aimeetings@fnal.gov
https://indico.fnal.gov/category/1446/

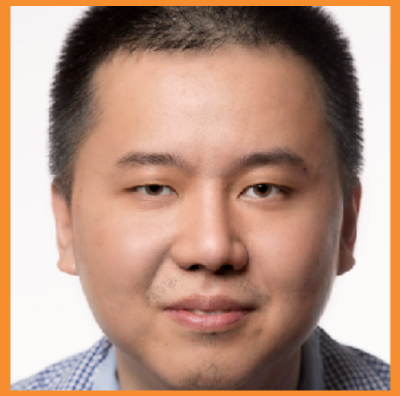
Learn more at:
ai.fnal.gov

Connect with AI Project officers for research collaborations!


Nhan Tran (Head)
(Collider +)


Burt Holzman
(Computing)


Farah Fahim
(Microelectronics)


Tinjun Yang
(Neutrino)


Gabe Perdue
(Quantum)


Brian Nord
(Cosmic)


Tia Miceli
(Accelerator)