Artificial Intelligence & Machine Learning at Fermilab

Abhijith Gandrakota

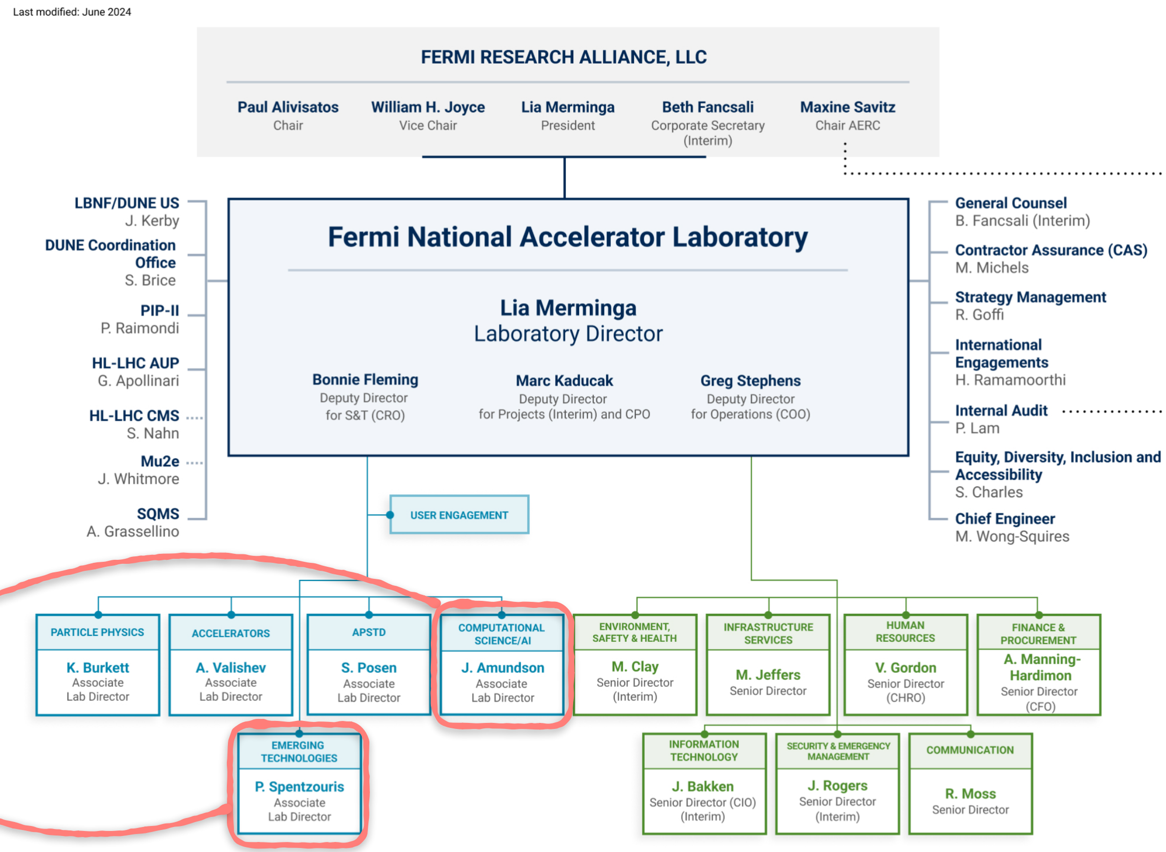
Fermilab's 57th annual users meeting
10 July 2024

AI Project Office

- Cross directorate: CSAID and Emerging Technologies Directorate

AI Project Office

(Nhan Tran), head, CSAID
(Burt Holzman), deputy head, CSAID
(Farah Fahim), ETD
(Tia Miceli), AD
(Brian Nord), CSAID
(Gabriel Perdue), ETD
(Tingjun Yang), PPD



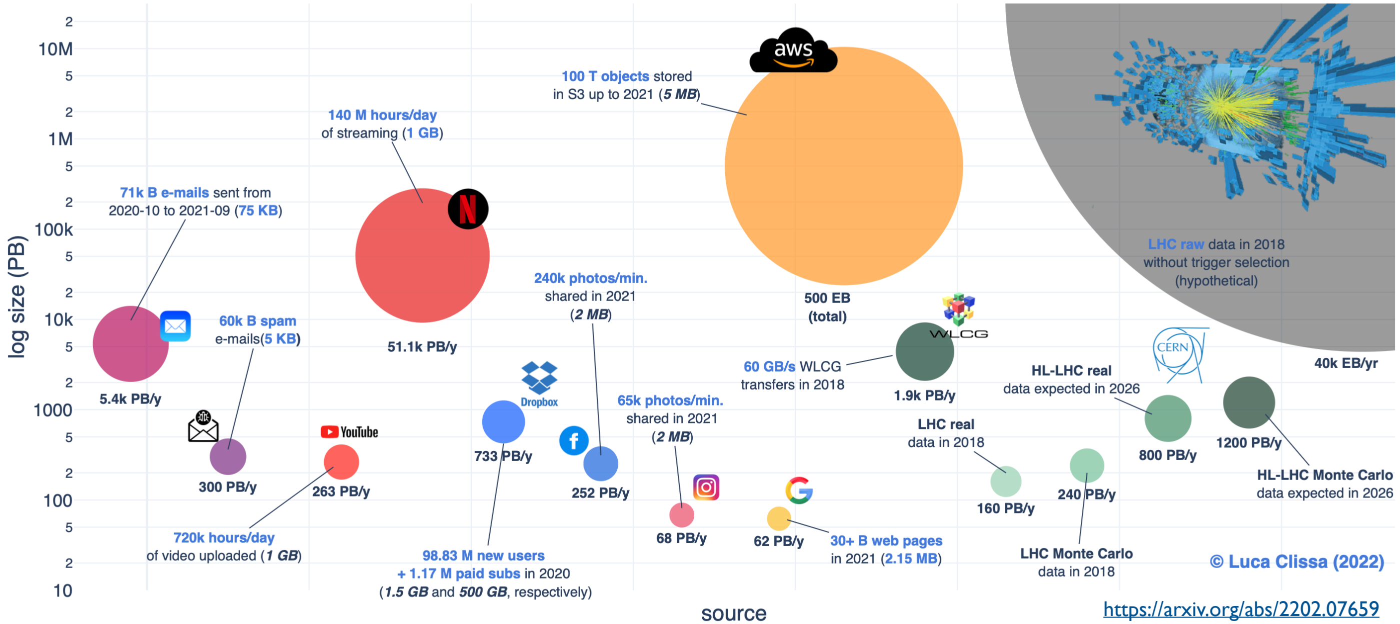
Preamble on AI Project Office

Self-defined goals:

- Developing **strategic capabilities** within the (inter)national AI ecosystem
  - AI to advance lab scientific mission, and where Fermilab can advance AI research
- Building **community** around cross-cutting problems, tools, and educational opportunities
  - Connecting teams across the lab and keeping a big-picture view of what is going on
  - Develop infrastructure for AI research — both people (e.g. AI associate program) and hardware (e.g. GPU access)
- Establish a strategy to support a **strong resource profile** through network of stakeholders and partners
- **Sharing** Fermilab and HEP's AI work with the world

Why AI & Ml for physics ?

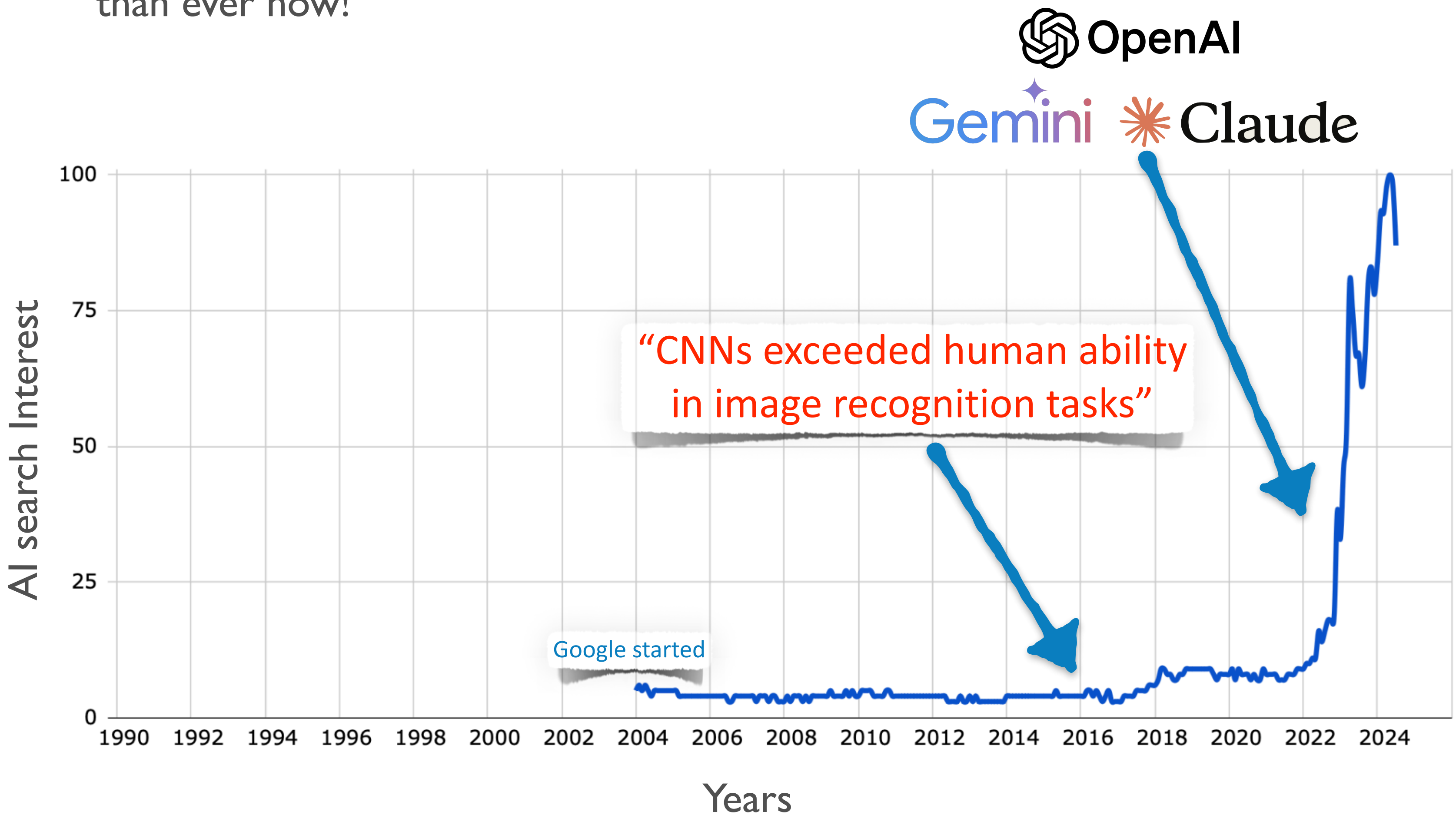
- We are collecting with huge amounts of data in HEP experiments
  - Challenging set up to find insights & new physics in these huge datasets
  - Need to maximize science by getting the most out of experiments
- AI & ML approaches are key to addressing these problems
  - Accelerate time-to-physics & discovery
  - Improve operational efficiency



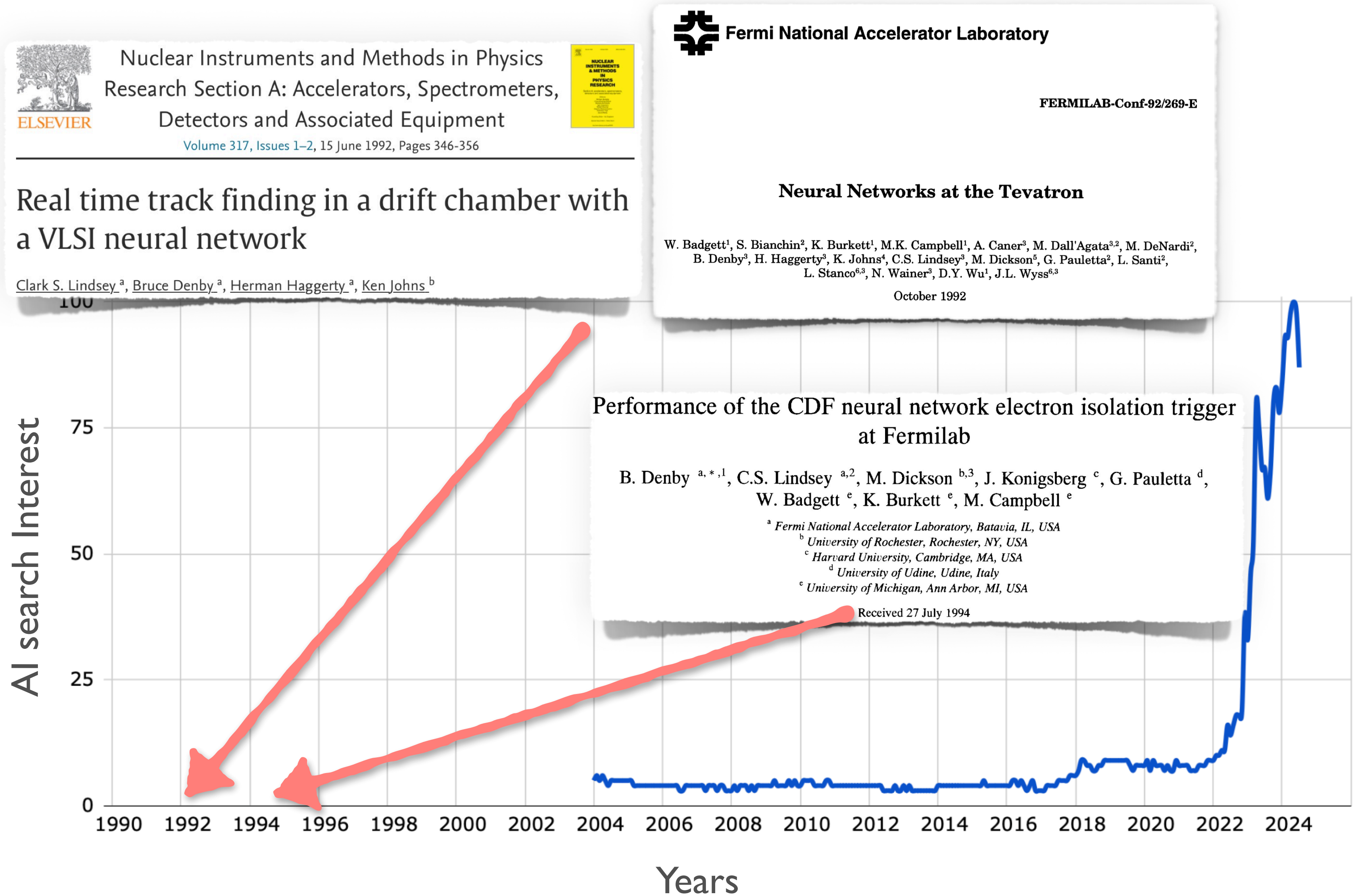
https://arxiv.org/abs/2202.07659

AI over the years

- Searches for "Artificial Intelligence" in google over the years
  - Artificial Intelligence and Machine learning is more main stream than ever now!



History of AI @ Fermilab



AI for Physics ⇔ Physics for AI

Outline

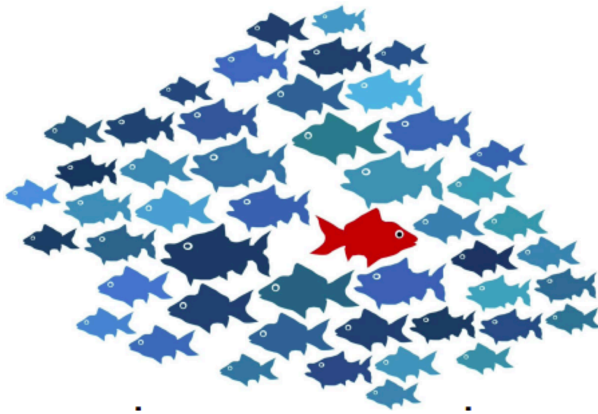
- AI for physics
  - Recent Highlights

- Physics for AI
  - FastML
  - AI @ Extreme Edge

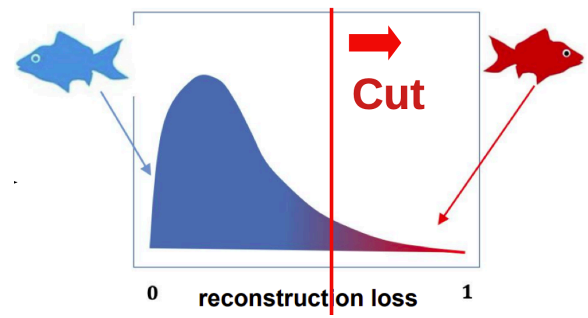
- AI for user community
  - Computing Resources for AI training and inference
  - Engage with Fermilab AI community
    - Lab Wide AI meetings & Jamboree

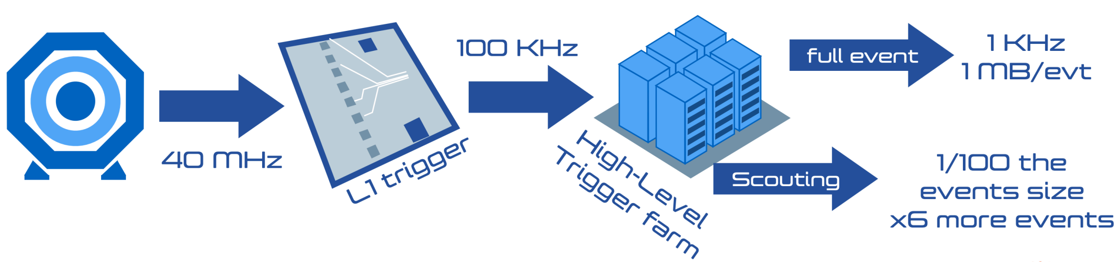
Learning from data: Unsupervised ML on chip !

- CMS detector creates more data than we can handle !
  - Need to throw away 99.75% of data at first stage!
  - We are interested in rare and beyond SM physics
  - Trigger make real-time decision on which data to record
    - Runs on FPGAs within O(100) nano seconds!
    - Needs to be unbiased to maximize discovery



- Unsupervised ML technique such as Anomaly Detection can catch effectively the deviations from SM
  - Demonstrated for data analysis for new physics searches by 3-7x !
  - Triggering on "anomalousness" of collision event

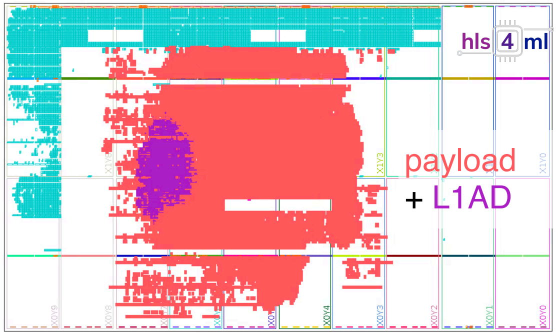


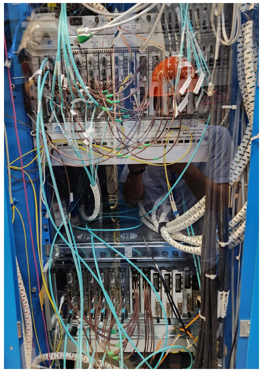


Records only ~ 0.01% of the data!

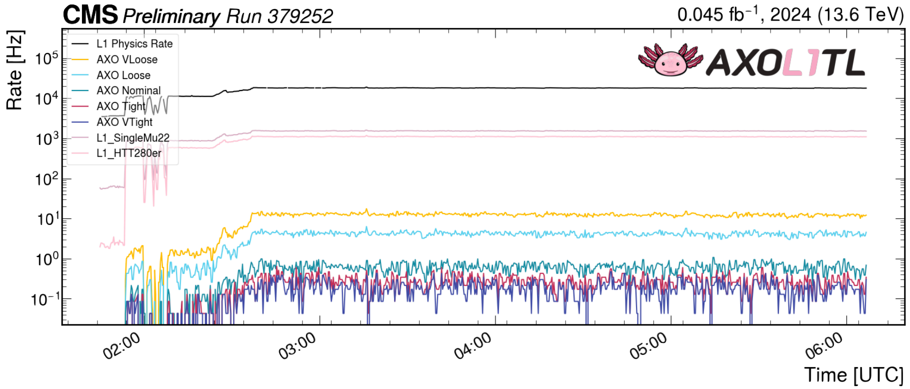
Learning from data: AXOL1TL

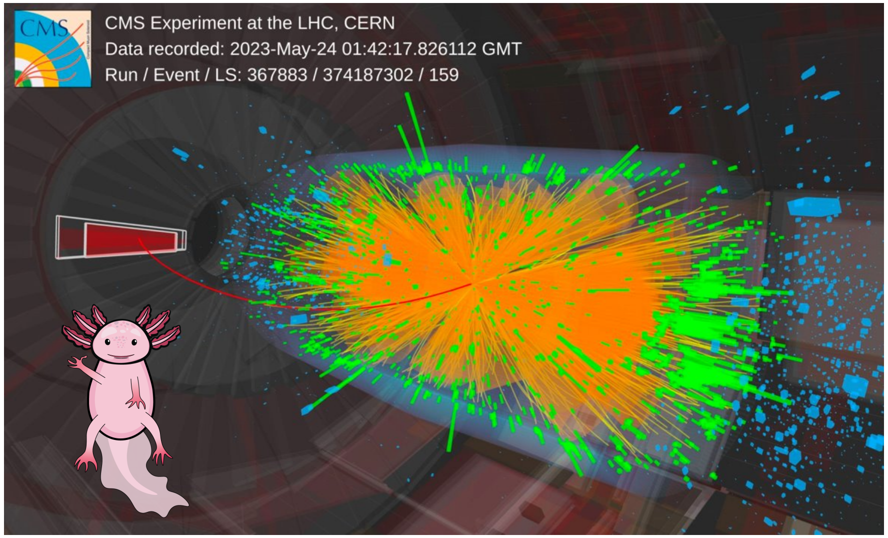
- AXOL1TL: triggering on "anomalousness"
  - Trained a ML model called Autoencoder directly on data to find "atypical" signatures
  - 10x more efficient that conventional trigger algorithms at CMS





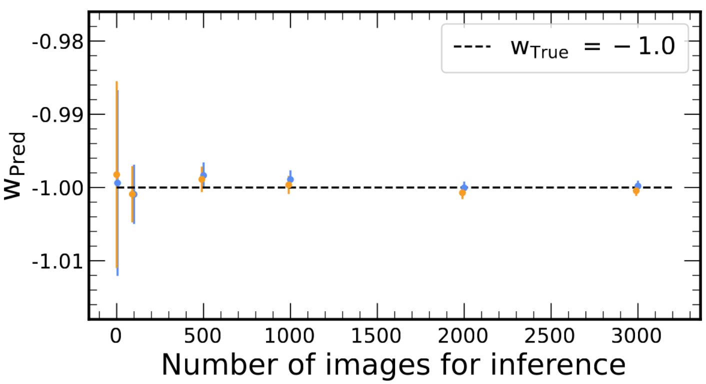
- AXOL1TL is running on CMS L1 Trigger FPGAs in at LHC, collecting the data
  - Performs inference in as little as 50 ns !
  - First ever full unsupervised ML trigger

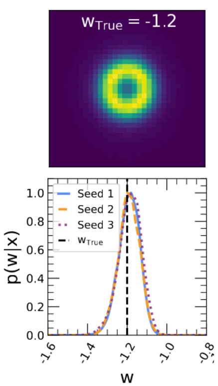




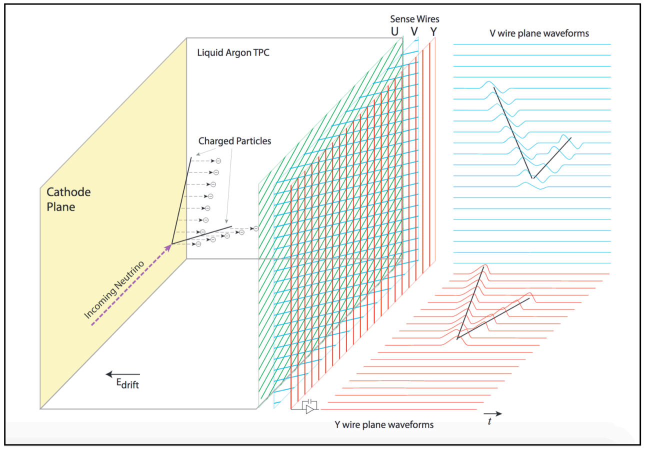
AI for Inference and reconstruction

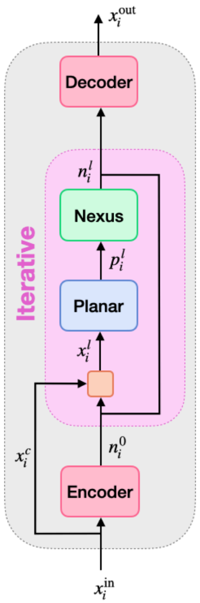
- [*Simulation Based inference for cosmology*]
- Goal: Rapidly find/measure objects, dark matter/energy from surveys
  - Inferring dark energy from population of strong lensed galaxies
  - Scalable for method for inference from O(1000) lenses from future surveys
  - Much faster than traditional MCMC





Accepted to ICML 2024, talk by S. Jarugula





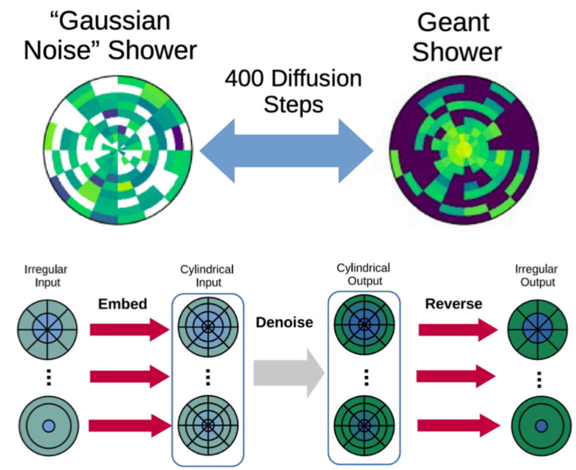
- *GNNs for Reconstruction in LArTPC*
  - Computationally efficient compared to previous CNN approaches
  - Adapted from HEPTrkX for tracking at LHC
  - Archived 98% efficiency in filtering background

https://arxiv.org/html/2403.11872v1#S1

AI for Theory and Accelerating Simulation

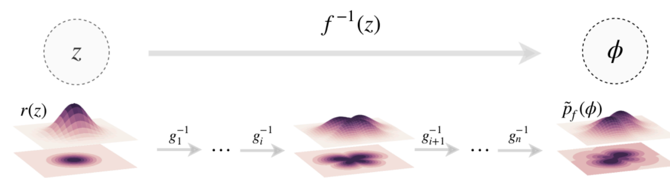
- *AI for fast detector simulation*
  - Addresses the computational challenge of simulation **Geant4** for HL-LHC
  - Diffusion based models to generate calorimeter shower simulations
    - Introduces novel geometry latent mapping
  - SOTA model in CaloChallenge with a 10-1000x speed compared to **Geant4**

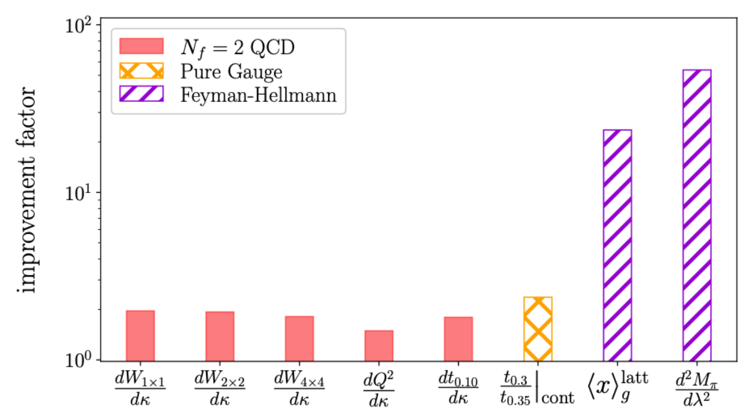
arXiv:2308.03876



- *Machine Learning for the lattice gauge theory*
  - Normalizing Flows to generate correlated lattice gauge field ensembles
  - Demonstrates variance reduction in the computation of observables
  - Significantly reduces statistical uncertainties

https://arxiv.org/pdf/2401.10874

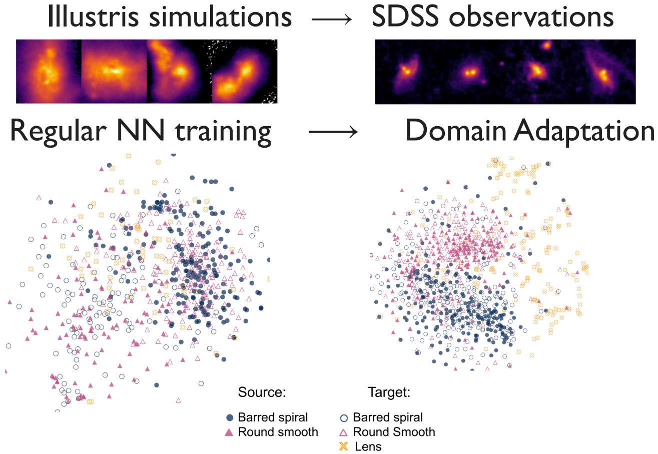




Robustness NNs

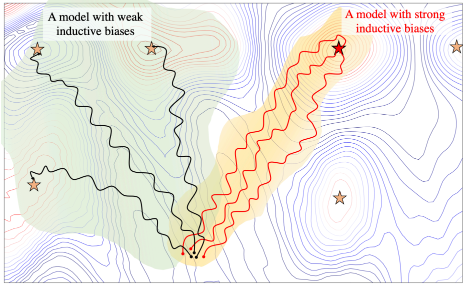
Domain Adaptation

Bridges difference between simulation & Obs. Data



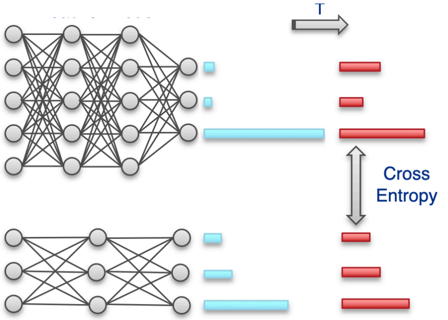
https://arxiv.org/abs/2302.02005

Robustness in Fast AI w/ Knowledge distillation of inductive bias



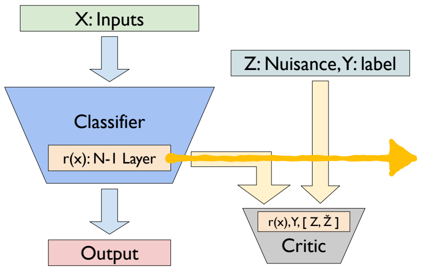
Imparts physics knowledge of the system into the fast and efficient ML models

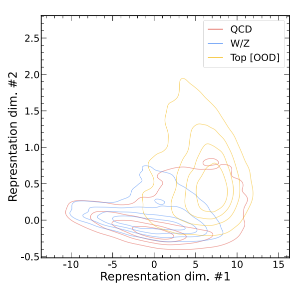
https://arxiv.org/abs/2311.14160



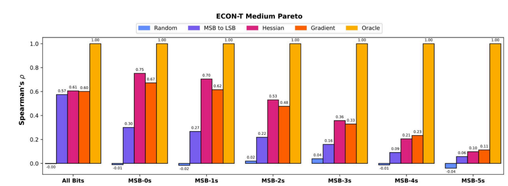
Nuisance invariant NNs w/ NuRD

Robust nuisance invariant Rep. learning





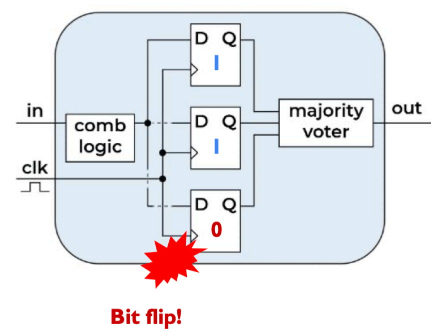
https://arxiv.org/abs/2401.08777



Robustness for NN on microelectronics

protects NNs on chip against bit flips in high radiation environments

https://arxiv.org/abs/2406.19522

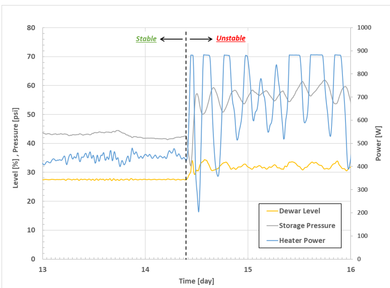


AI for operations and control system

Navigating Cryogenic System Transients at CMTF Using a Digital Twin

Predict the future state of a 3000-Liter Dewer level

Design a predictive LSTM reinforcement learning agent



Self Driving telescope

Train an algorithm to reactively move a ground telescope w/ Stone Edge Observatory (controllable with a Slack API)

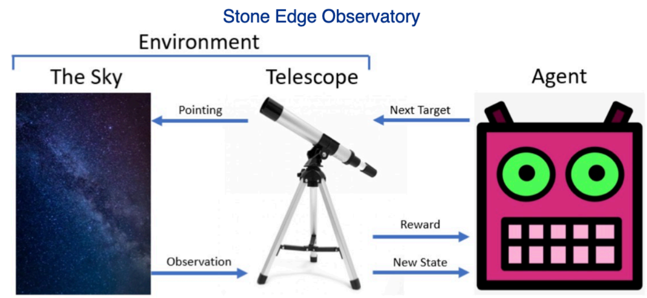


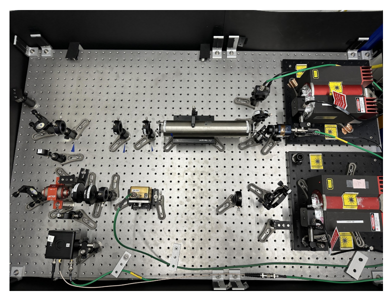
Diagram of the RL setup. A trained agent takes the current state of the sky, and moves the telescope to the next optimal target --

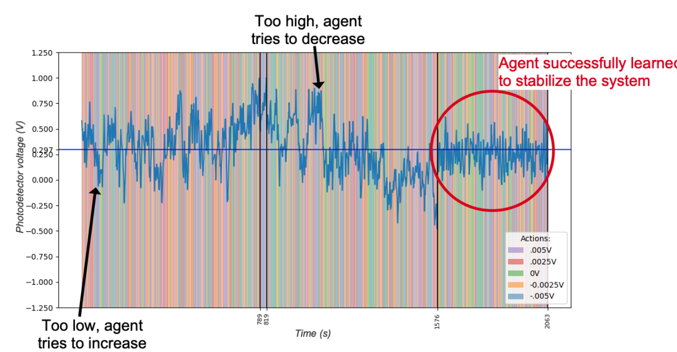
https://arxiv.org/abs/2311.18094

Optical Cavities for Quantum Gravity Searches

Control a laser interferometer set-up to photon counter

Reinforcement Learning to keep 2L/λ constant

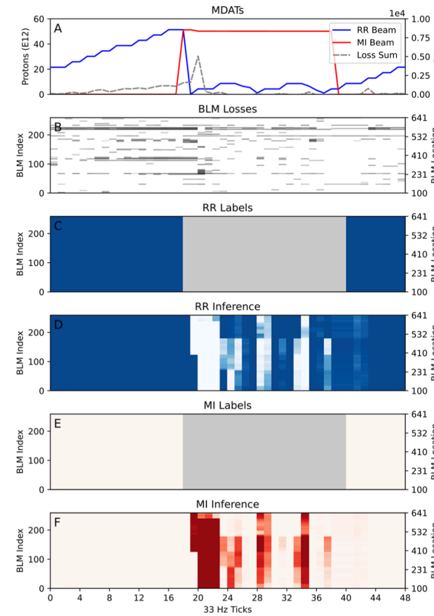


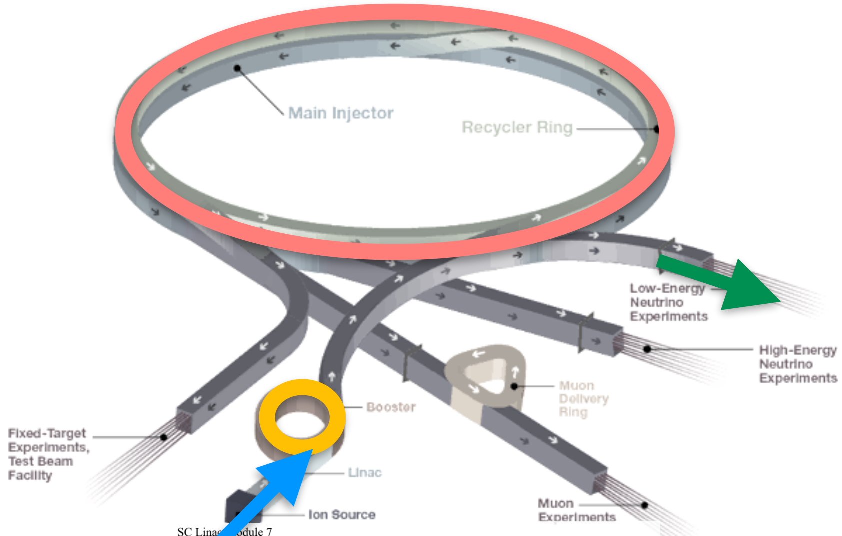


AI for Accelerator Controls

Real-time Edge AI Distributed System

- Differentiate beam loss monitor signals around the ring
  - Identify if main injector or recycler ring is the source
  - Deployed to FPGA on a custom card



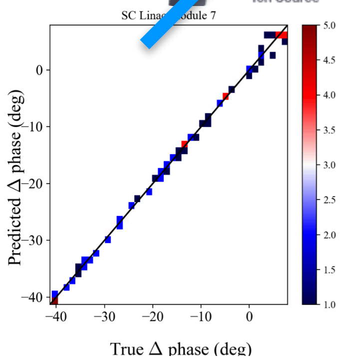


NuMI Beam Variability Prediction & LBNF monitoring

Predict NuMI proton beam position, intensity and horn current

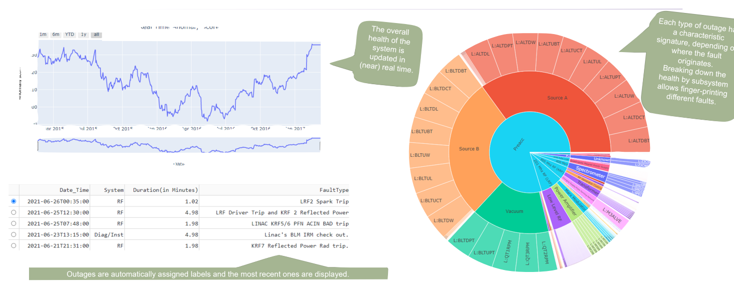
Linac RF Op

Predict PF parameters to keep beam energy constant and minimize eminence



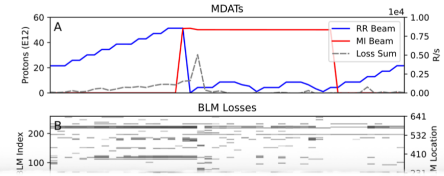
L-Cape

Predict anomalies and identify causing beam downtime

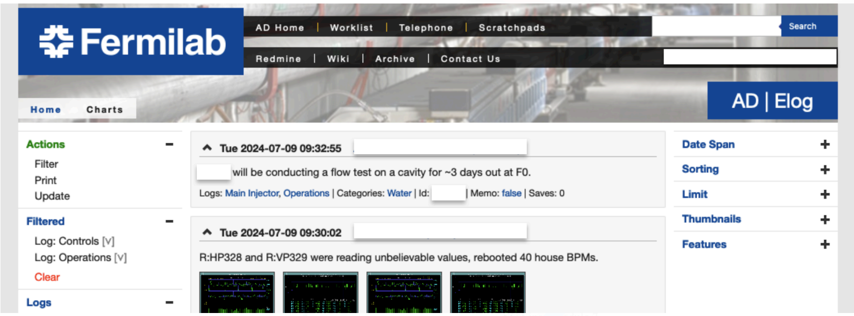


Fermilab

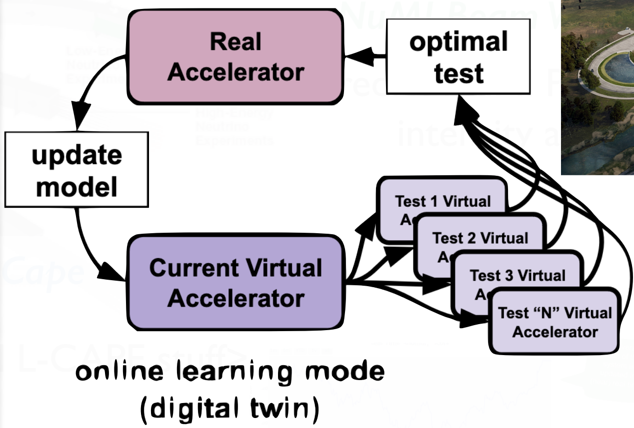
AI for Accelerator Controls



Large language model integration for main control room e-log



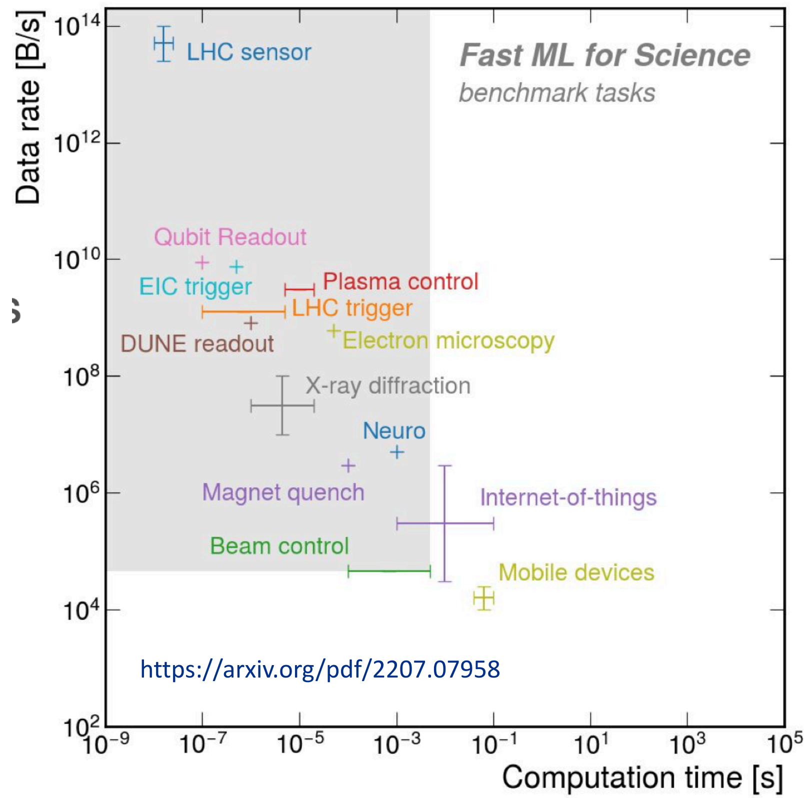
Digital Twin / Virtual Accelerator for PIP-II

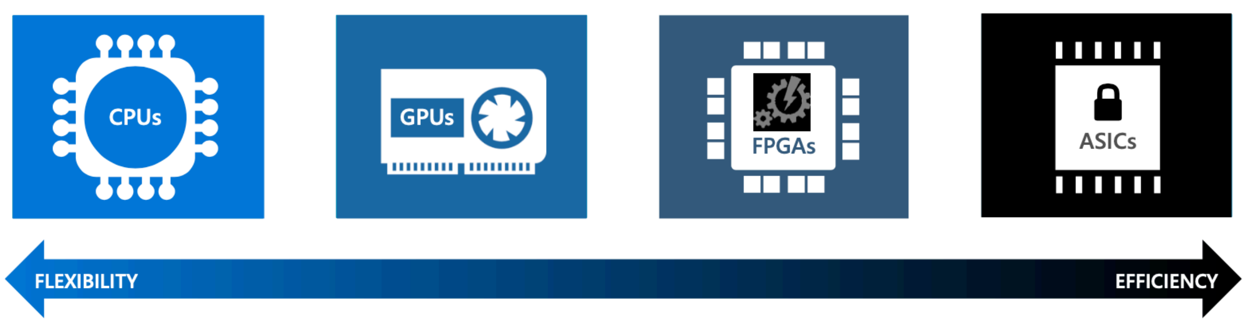


Physics for AI : Fast ML

Landscape of Fast ML

- Many experiments, particularly at Fermilab require custom made AI/ML methods
- Typically needs to process huge amounts of data in a very short time scale
  - Beyond the benchmarks in industry
  - Need: Real-time and efficient AI
- CPUs can not keep with these demands
  - Special hardware FPGAs/ASIC provide huge flexibility through parallel compute
  - Challenging to run ML models on these

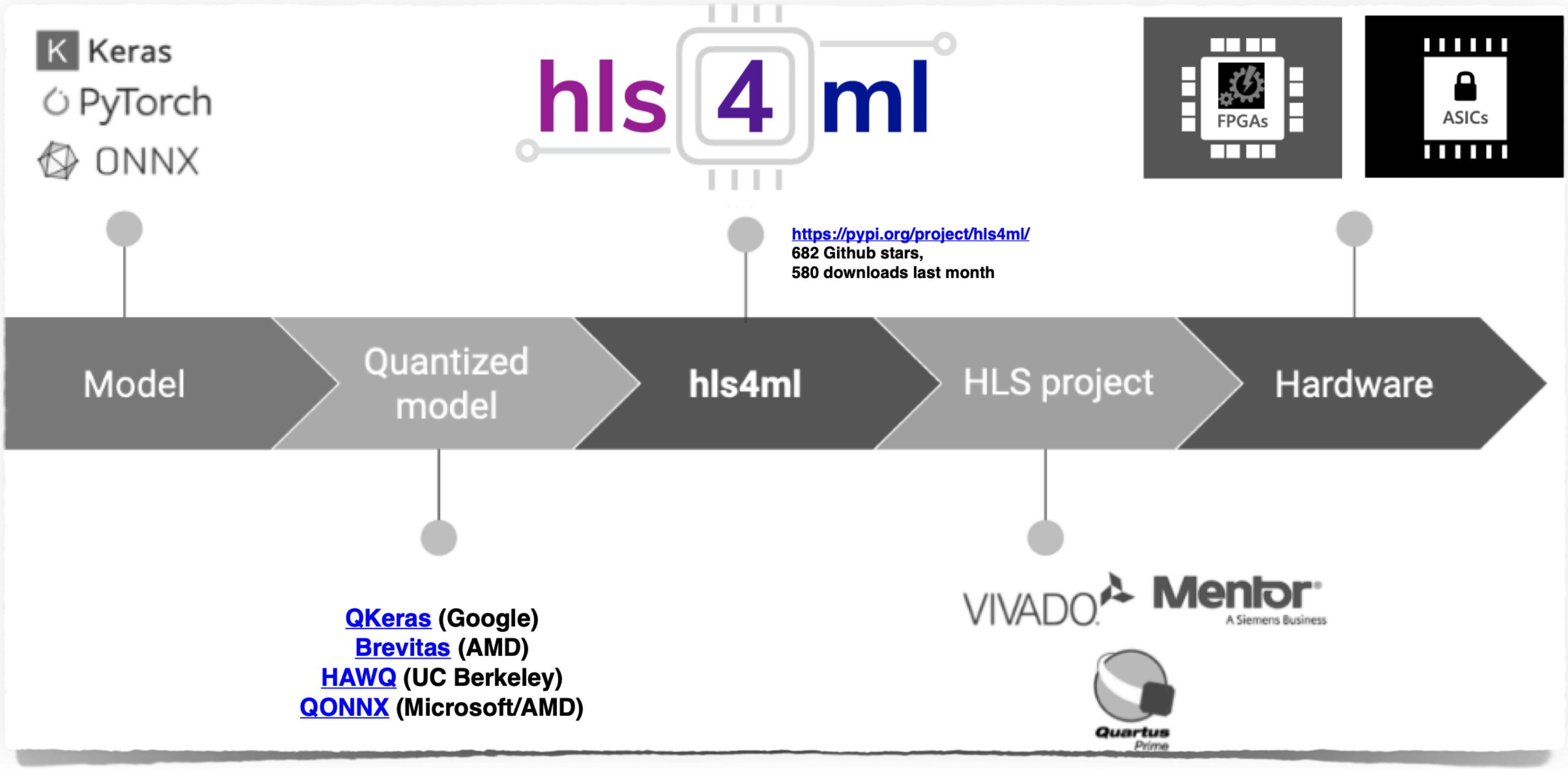




Fermilab

Efficient ML hardware software codesign

- Enabling efficient algorithms and workflows
  - Accessible point of entry and easy to use tools for non-experts into hardware

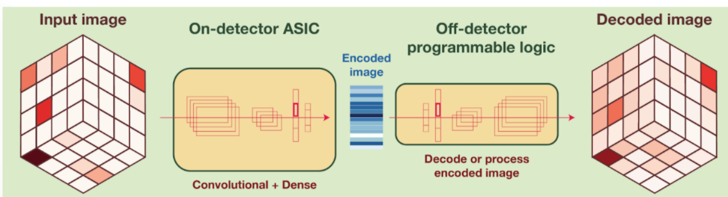


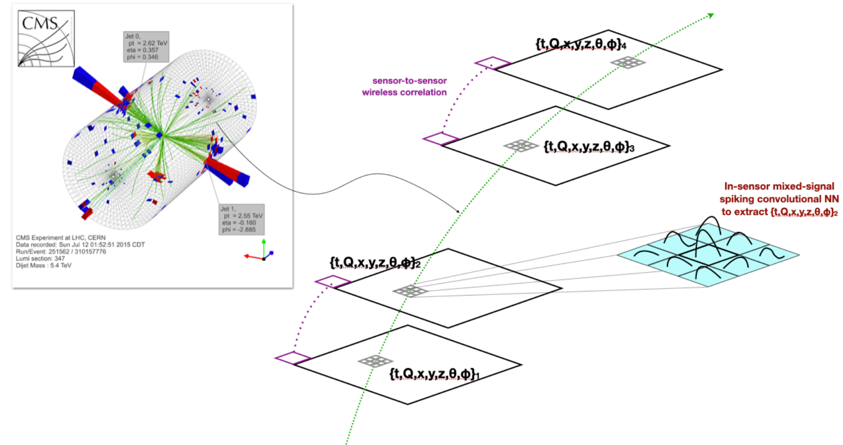
JINST 13 P07027 (2018)
arXiv:2102.11289
arXiv:2206.07527

https://fastmachinelearning.org/hls4ml

AI @ Extreme Edge

- Data compression w/ Rad. hard ASICs
  - First use of DL for HEP on ASICs
  - Developed for use in CMS High Granularity CALorimeter
  - Powerfull nonlinear data compression schemes

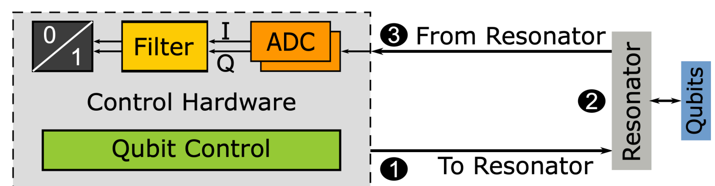




- Smart pixels: Pixel sensors w/ AI on chip
  - Efficiently filter low p_T tracks
  - Saving up to 75% of data bandwidth
  - Curial for future colliders e.g: Reducing beam background in μC

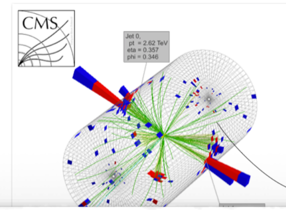
10.1109/ISCAS46773.2023.10182033

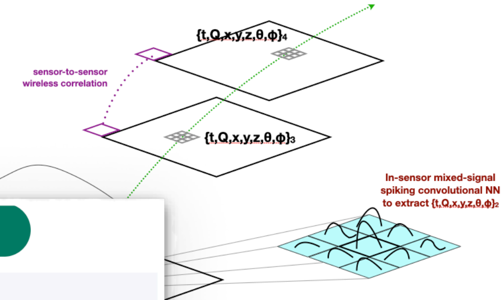
- AI/ML for controlling and optimizing quantum computers
  - Edge AI to improve qubit readout
  - Denoising computations in theory calculations
  - Predicting quantum circuit fidelity on noisy hardware



AI @ Edge

- Data compression w/ Rad. hard ASICs
  - Fi
  - D C
  - Po sc





yahoo!finance

Search for news, symbols or companies

My Portfolio | News | Markets | Sectors | Screeners | Personal Finance | Videos

Siemens simplifies development of AI accelerators for advanced system-on-chip designs with Catapult AI NN

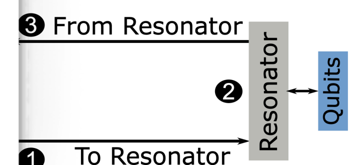
PR Newswire
Tue, May 21, 2024, 8:00 AM CDT • 5 min read

In This Article:

> Catapult AI NN brings together hls4ml, an open-source package for machine learning hardware acceleration, and Siemens' Catapult™ HLS software for High-Level Synthesis. Developed in close collaboration with Fermilab, a U.S. Department of Energy Laboratory, and other leading contributors to hls4ml, Catapult AI NN addresses the unique requirements of machine learning accelerator design for power, performance, and area on custom silicon.

- …s w/ AI on chip
  - p_T tracks
  - data bandwidth
  - …liders
  - …background in μC

- AI/ML f…
  - Ed…
  - De…
  - Predicting quantum circuit fidelity on noisy hardware



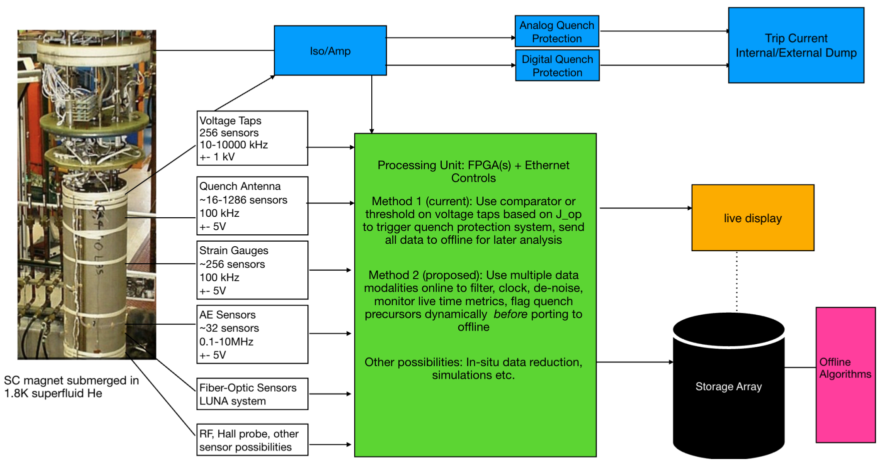
Fast ML for control systems

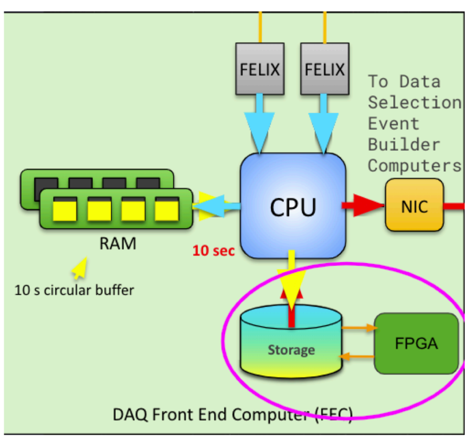
Magnet Quench Detection

- Efficiently detect quenches in SC magnets
  - Predicitve models to take preventive measures and decrease downtime
  - Critical for enabling future energy and intensity frontier experiments

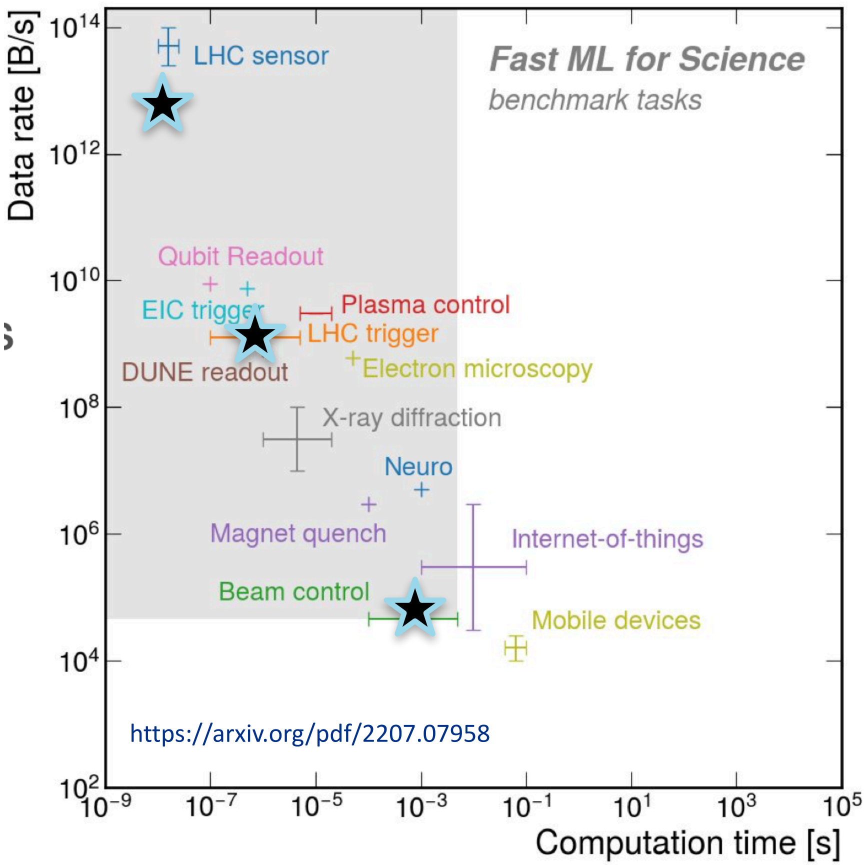
Supernova Detection with DUNE

- Quickly detect and point to the Supernova bursts
  - Uses FPGAs to bring power efficient processing to the data
  - Prompt detection enables multi-messenger astronomy for follow up w/ other detectors





Fast ML for Science Benchmarks



- Development of open source tools helps democratize the (edge) AI for all of HEP (hls4ml, DeepBench, SONIC, Open Data …)

- Benchmarks for HEP challenges will leads to more AI/ML solutions and broader engagement

  – Fast ML Science benchmarks takes a step in this direction

  – Tasks with well defined real-time system and resource constraints

  – Challenges for broader AI community w/ datasets and baseline models

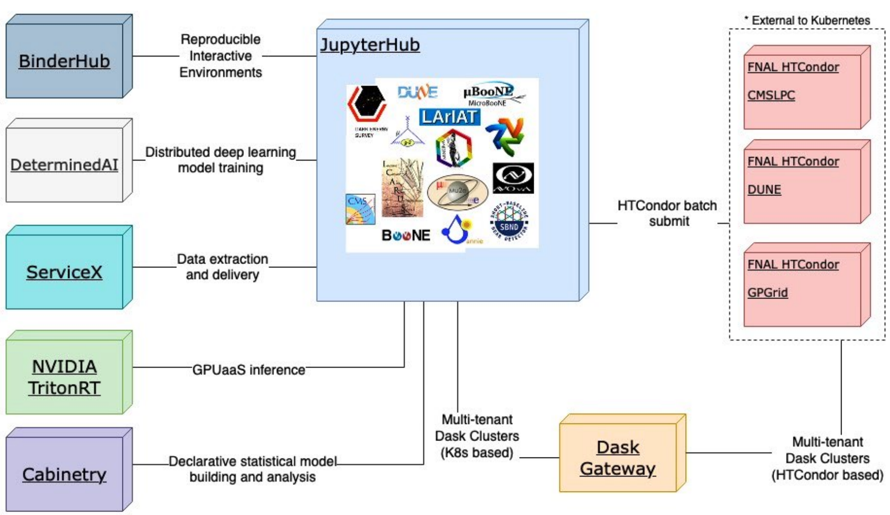
AI for Fermilab user community

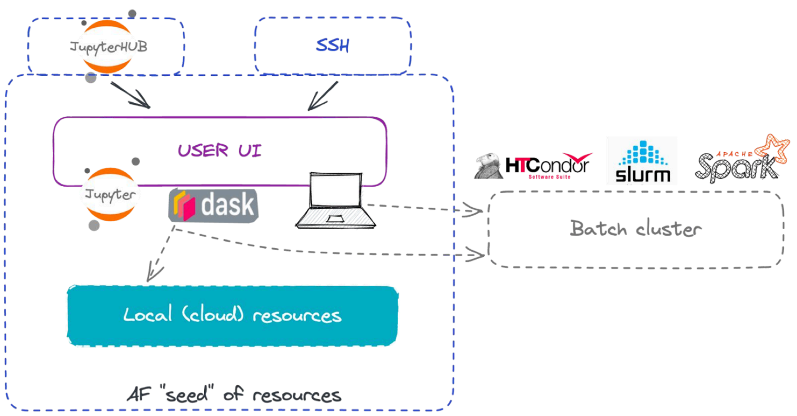
Elastic analysis facility ecosystem

- Platform for rapid scientific analysis with modern web and container technologies
  - Equipped with industry leading GPUs for AI training and inference

- Highly scalable, customizable computing infrastructure
  - Capable of bursting to O(100k)batch computing cores

Fermilab Elastic Analysis Facility Ecosystem



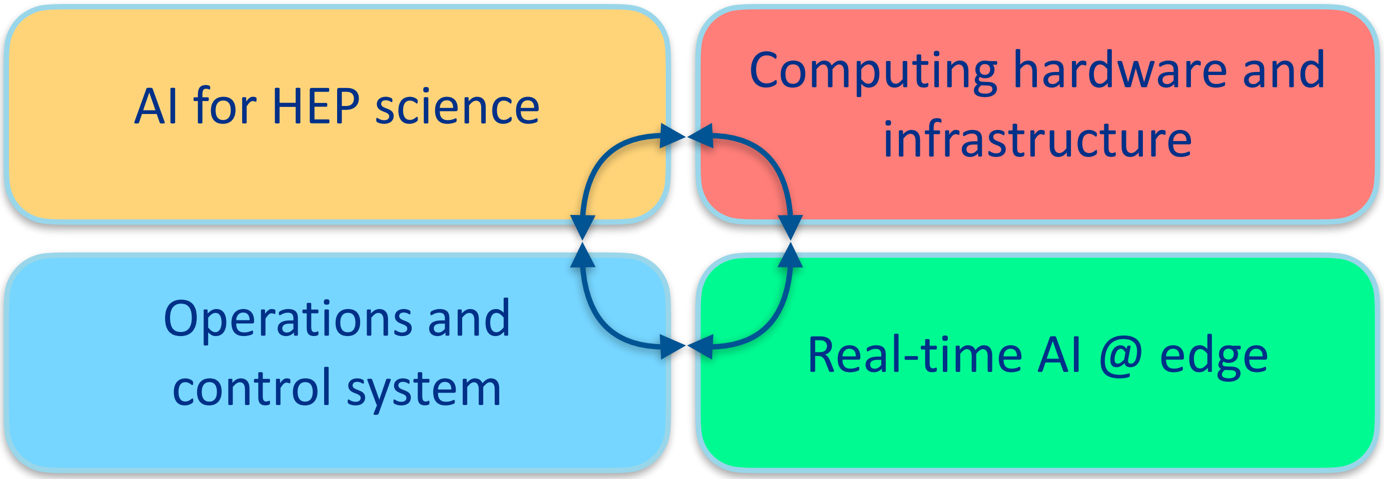


https://eafjupyter.readthedocs.io/en/latest/index.html

AI community @ Fermilab

- Bi weekly lab-wide AI meetings
  - Discuss the latest development in AI and cutting edge AI/ML projects across the lab
  - Great avenue to learn and collaborate
  - https://indico.fnal.gov/category/1446/
  - Announcements: aimeetings@listserv.fnal.gov
- AI Jamboree
  - Highlight current AI activities at the lab
  - Panel discussions and Idea incubator
- Engage with broader AI and HEP community

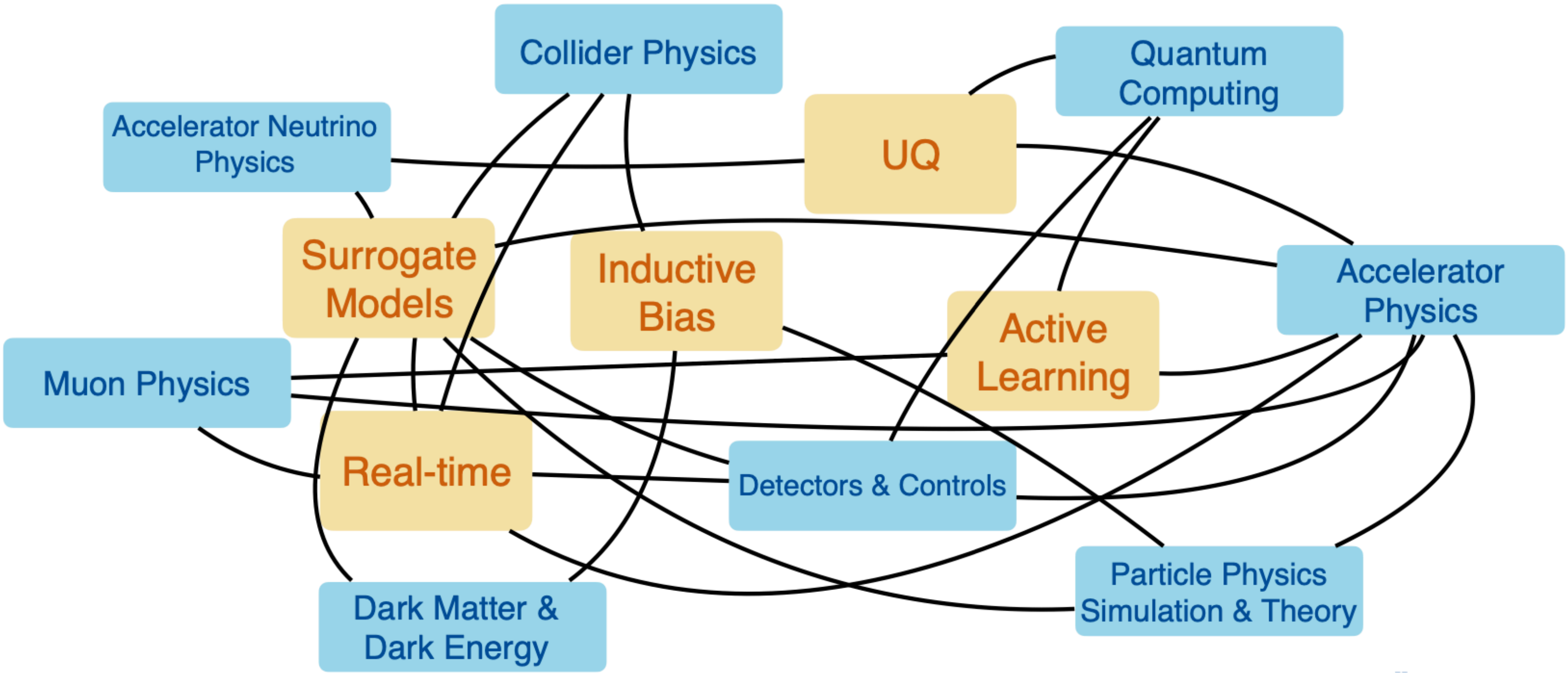
Landsacpe of AI @Fermilab



Using *Fast*, *Efficient* , *Robust* and *Generalizable* AI approaches

Broad view of Fermilab AI efforts

Connect with the AI project office!



Learn more at: ai.fnal.gov

Subscribe to meeting announcements: aimeetings@listserv.fnal.gov.

Thank you !