Fermilab U.S. DEPARTMENT OF ENERGY Office of Science

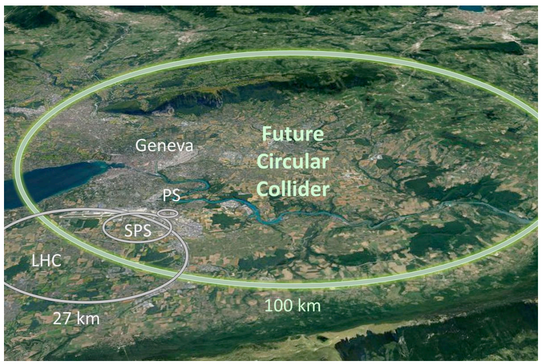
Artificial Intelligence & Machine Learning at Fermilab

Jennifer Ngadiuba

FermiFusion Workshop: Uniting Minds for Scientific Advancement
6 August 2024

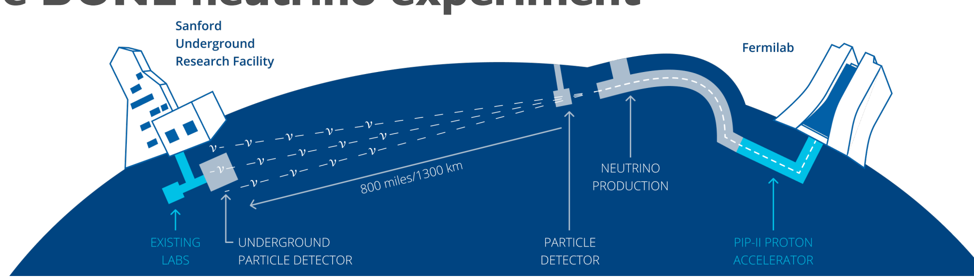
Big Science = Big Data

Probing the **fundamental structure of nature** requires complex experimental devices, large infrastructures and big collaborations.



International MUON Collider Collaboration

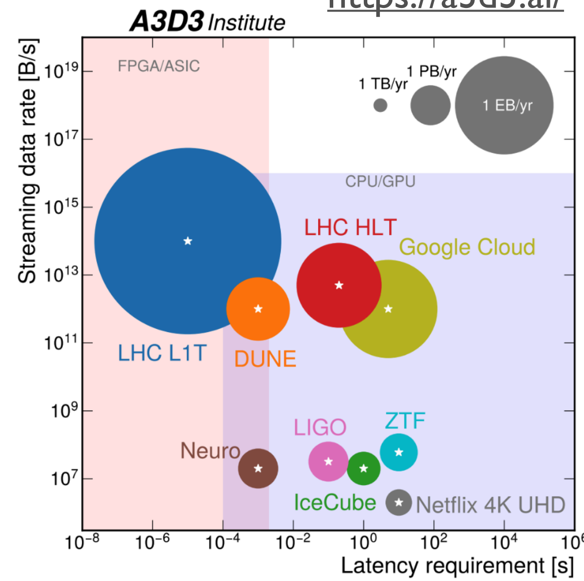
The DUNE neutrino experiment

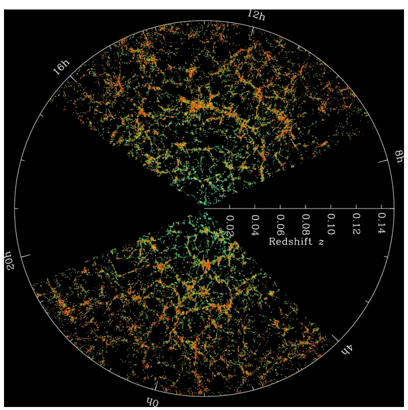


Big Science = Big Data

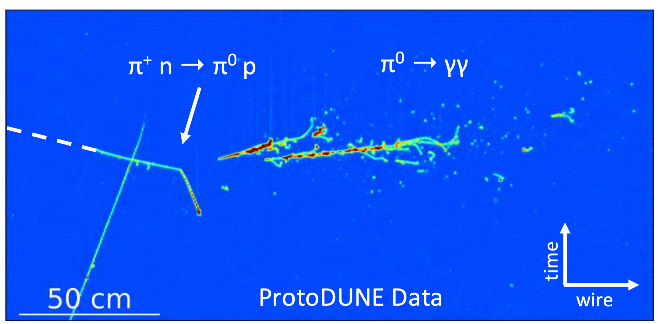
- Increasingly complex data both in **volume and dimensionality**
- Increasing need for **efficient and accurate data processing pipelines**
- Challenge in **simulating expectations** for what experiments may observe
- But also need for innovative **data & discovery driven** physics analyses approaches

https://a3d3.ai/

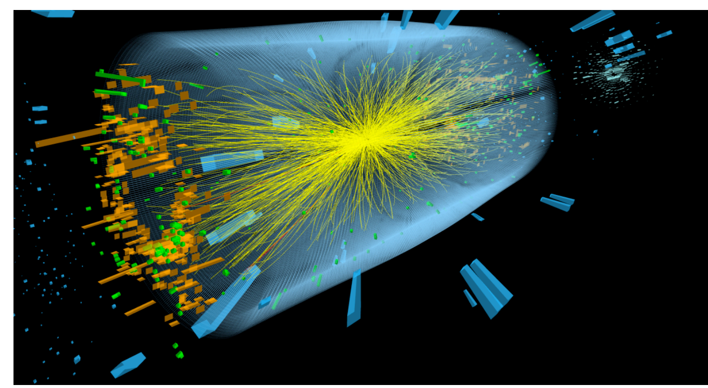




Sloan Digital Sky Survey



Interactions in LArTPC

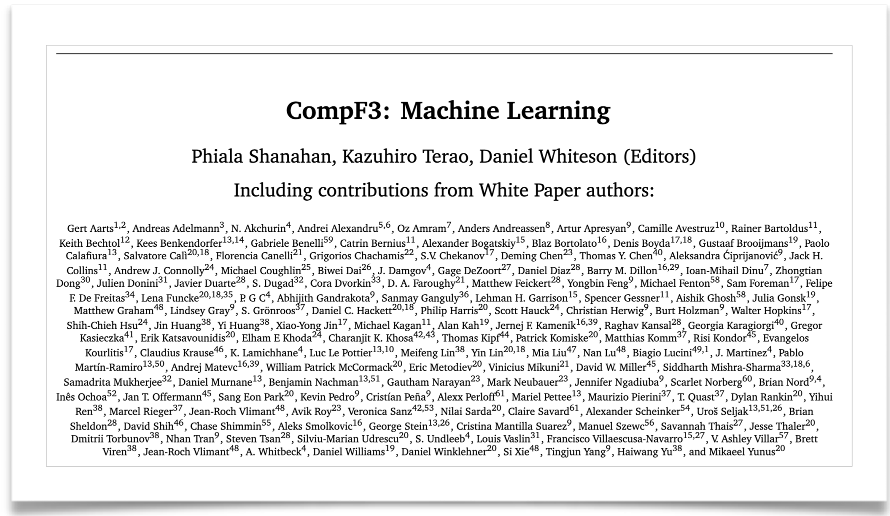


A LHC collision

The role of AI in HEP

- In this era of science **Artificial Intelligence can accelerate time to discovery**
  - efficient analysis of large amounts of highly-dimensional data to find subtle patterns
- With such capability it will allow us:
  - enhance control and operations of detectors and accelerators
  - automate online and offline experimental workflows
  - save and maximize potentially lost data
  - accelerate detector R&D
  - test hypotheses significantly faster

[Dedicated part of Snowmass computational frontier]



CompF3: Machine Learning

Phiala Shanahan, Kazuhiro Terao, Daniel Whiteson (Editors)

Including contributions from White Paper authors:

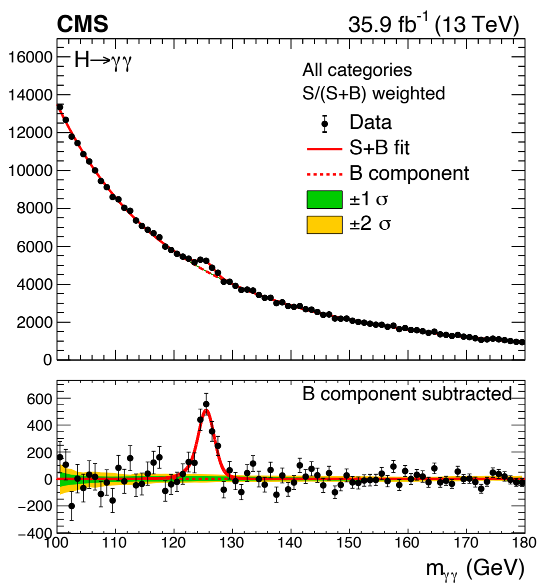
Gert Aarts[1,2], Andreas Adelmann[3], N. Akchurin[4], Andrei Alexandru[5,6], Oz Amram[7], Anders Andreassen[8], Artur Apresyan[9], Camille Avestruz[10], Rainer Bartoldus[11], Keith Bechtol[12], Kees Benkendorfer[13,14], Gabriele Benelli[59], Catrin Bernius[11], Alexander Bogatskiy[15], Blaz Bortolato[16], Denis Boyda[17,18], Gustaaf Brooijmans[19], Paolo Calafiura[13], Salvatore Calì[20,18], Florencia Canelli[21], Grigorios Chachamis[22], S.V. Chekanov[17], Deming Chen[23], Thomas Y. Chen[40], Aleksandra Ćiprijanović[9], Jack H. Collins[11], Andrew J. Connolly[24], Michael Coughlin[25], Biwei Dai[26], J. Damgov[4], Gage DeZoort[27], Daniel Diaz[28], Barry M. Dillon[16,29], Ioan-Mihail Dinu[7], Zhongtian Dong[30], Julien Donini[31], Javier Duarte[28], S. Dugad[32], Cora Dvorkin[33], D. A. Faroughy[21], Matthew Feickert[28], Yongbin Feng[9], Michael Fenton[58], Sam Foreman[17], Felipe F. De Freitas[34], Lena Funcke[20,18,35], P. G C[4], Abhijith Gandrakota[9], Sanmay Ganguly[36], Lehman H. Garrison[15], Spencer Gessner[11], Aishik Ghosh[58], Julia Gonsk[19], Matthew Graham[48], Lindsey Gray[9], S. Grönroos[37], Daniel C. Hackett[20,18], Philip Harris[20], Scott Hauck[24], Christian Herwig[9], Burt Holzman[9], Walter Hopkins[17], Shih-Chieh Hsu[24], Jin Huang[38], Yi Huang[38], Xiao-Yong Jin[17], Michael Kagan[11], Alan Kah[19], Jernej F. Kamenik[16,39], Raghav Kansal[28], Georgia Karagiorgi[40], Gregor Kasieczka[41], Erik Katsavounidis[20], Elham E Khoda[24], Charanjit K. Khosa[42,43], Thomas Kipf[44], Patrick Komiske[20], Matthias Komm[37], Risi Kondor[45], Evangelos Kourlitis[17], Claudius Krause[46], K. Lamichhane[4], Luc Le Pottier[13,10], Meifeng Lin[38], Yin Lin[20,18], Mia Liu[47], Nan Lu[48], Biagio Lucini[49,1], J. Martinez[4], Pablo Martín-Ramiro[13,50], Andrej Matevc[16,39], William Patrick McCormack[20], Eric Metodiev[20], Vinicius Mikuni[21], David W. Miller[45], Siddharth Mishra-Sharma[33,18,6], Samadrita Mukherjee[32], Daniel Murnane[13], Benjamin Nachman[13,51], Gautham Narayan[23], Mark Neubauer[23], Jennifer Ngadiuba[9], Scarlet Norberg[60], Brian Nord[9,4], Inês Ochoa[52], Jan T. Offermann[45], Sang Eon Park[20], Kevin Pedro[9], Cristían Peña[9], Alexx Perloff[61], Mariel Pettee[13], Maurizio Pierini[37], T. Quast[37], Dylan Rankin[20], Yihui Ren[38], Marcel Rieger[37], Jean-Roch Vlimant[48], Avik Roy[23], Veronica Sanz[42,53], Nilai Sarda[20], Claire Savard[61], Alexander Scheinker[54], Uroš Seljak[13,51,26], Brian Sheldon[28], David Shih[46], Chase Shimmin[55], Aleks Smolkovic[16], George Stein[13,26], Cristina Mantilla Suarez[9], Manuel Szewc[56], Savannah Thais[27], Jesse Thaler[20], Dmitrii Torbunov[38], Nhan Tran[9], Steven Tsan[28], Silviu-Marian Udrescu[20], S. Undleeb[4], Louis Vaslin[31], Francisco Villaescusa-Navarro[15,27], V. Ashley Villar[57], Brett Viren[38], Jean-Roch Vlimant[48], A. Whitbeck[4], Daniel Williams[19], Daniel Winklehner[20], Si Xie[48], Tingjun Yang[9], Haiwang Yu[38], and Mikaeel Yunus[20]

[From Snowmass summary:]

- The pervasive use of artificial intelligence and machine learning, AI/ML, in nearly every aspect of our software. Hardly mentioned in the 2013 report, these revolutionary machine-learning approaches are transforming the way we work.

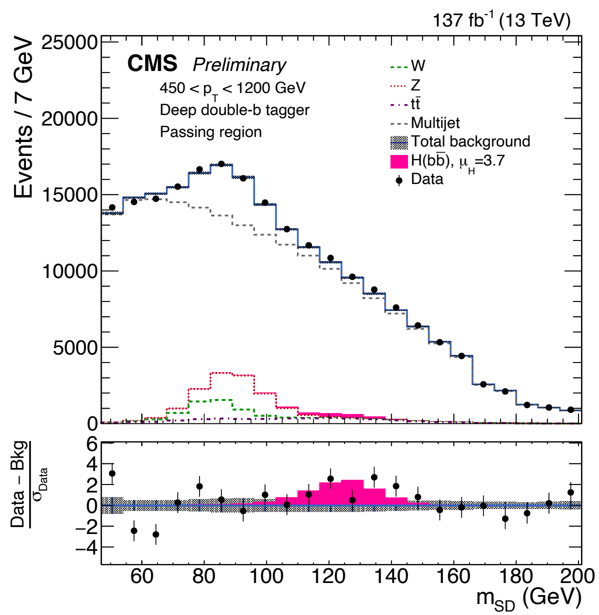
AI in HEP

- **Machine Learning is used in particle physics since the ‘80s**
  - Shallow networks back then, mostly BDTs since ~ 2004 (e.g., Higgs boson discovery)



Higgs → photons

[Phys. Lett. B 805 (2020) 135425]

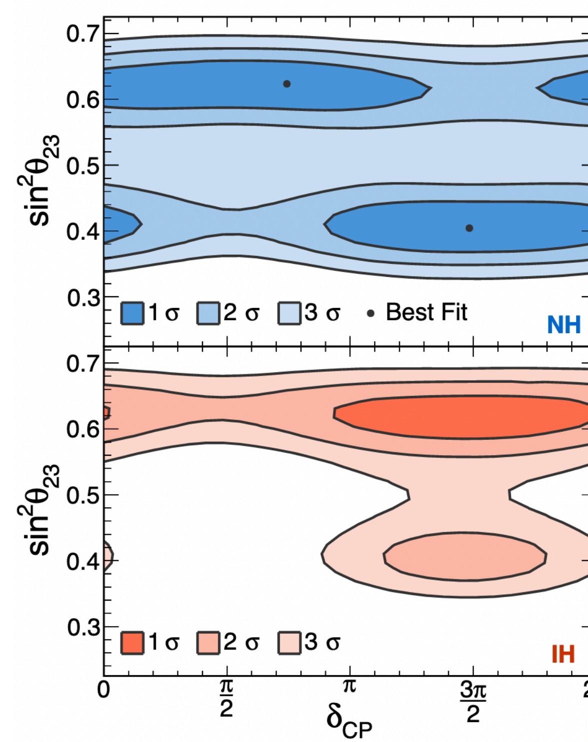


Higgs → bottom quarks

[JHEP 12 (2020) 085]

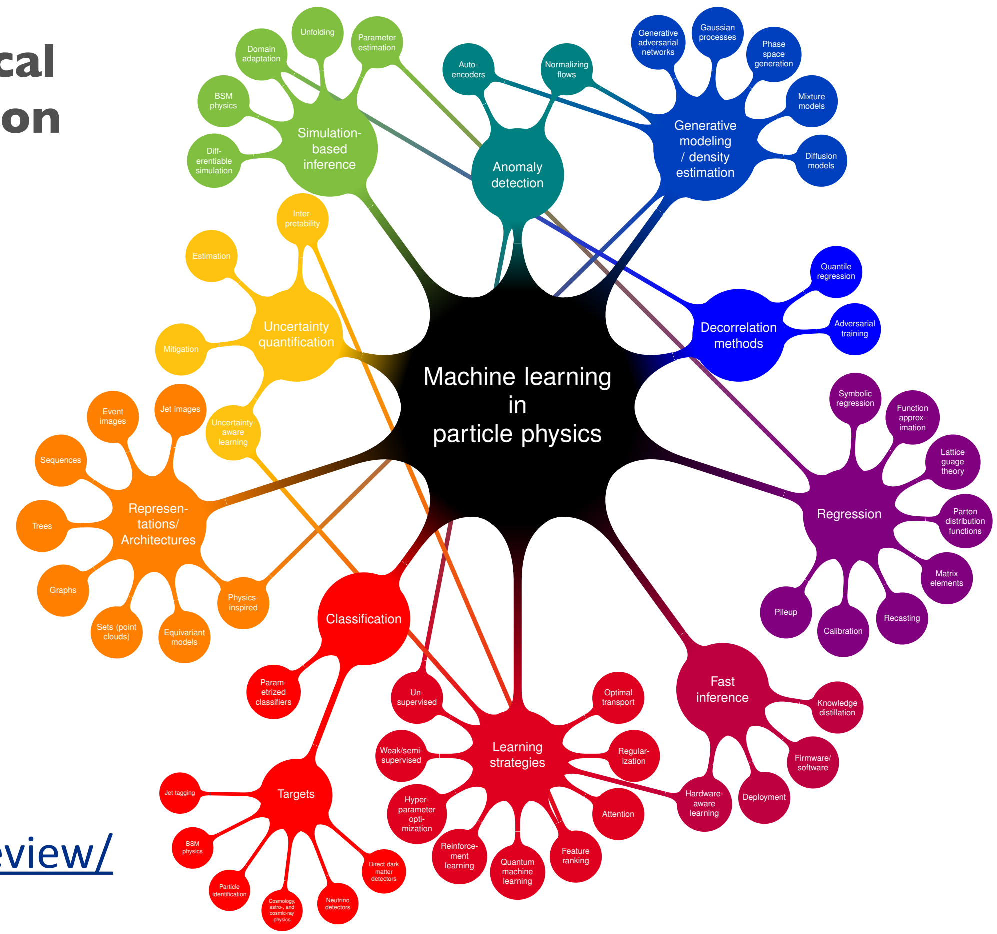
Measurement of neutrino oscillation parameters @ NovA

[Phys. Rev. Lett. 118, 231801 (2017)]



The AI revolution

- **Machine Learning is used in particle physics since the '80s**
  - Shallow networks back then, mostly BDTs since ~ 2004 (e.g., Higgs boson discovery)
- **Over the last decade a rapid progress guided by technological breakthrough led to a revolution in this area**
  - this is the era of Deep Learning



https://iml-wg.github.io/HEPML-LivingReview/

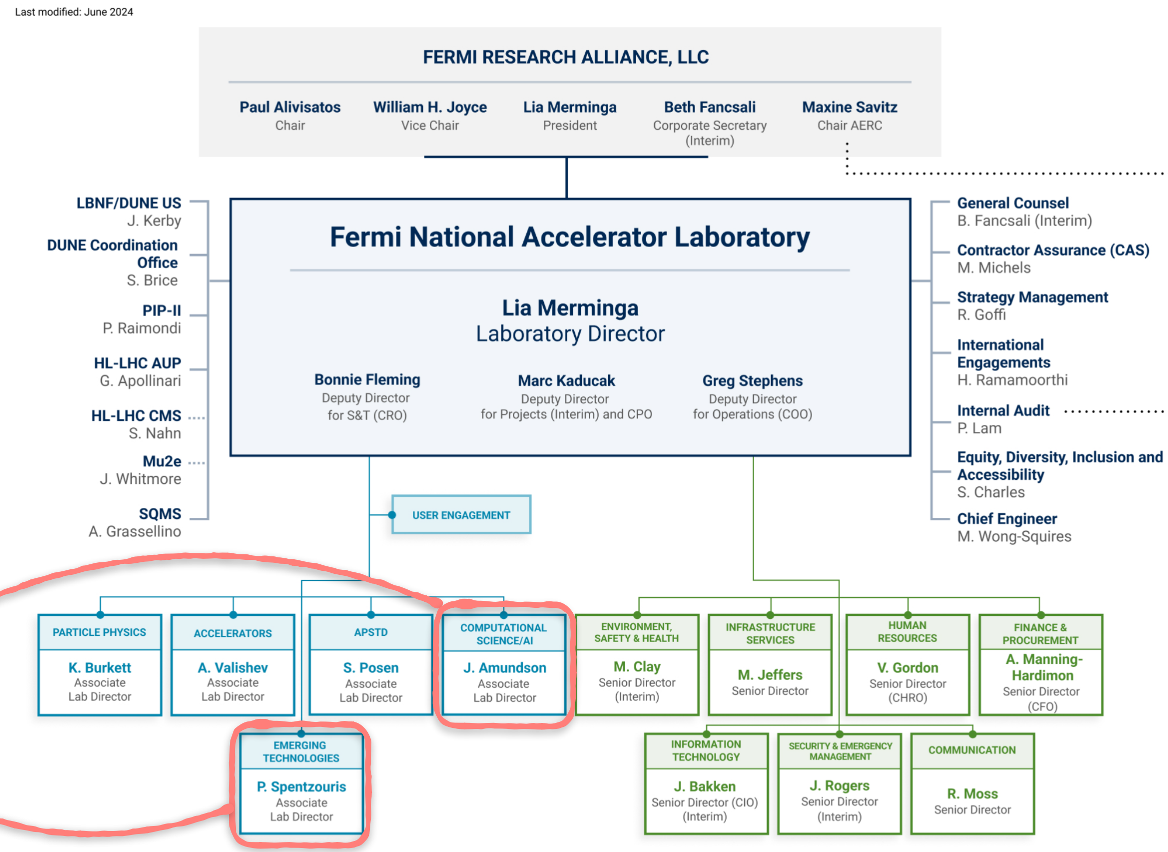
Fermilab

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
Jennifer Ngadiuba, PPD



https://computing.fnal.gov/artificial-intelligence/

Fermilab

AI Project Office goals

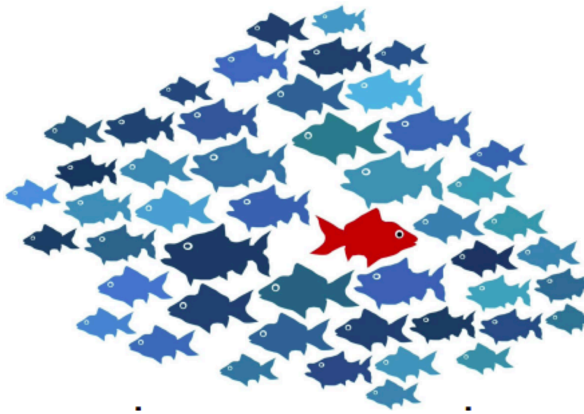
- **Accelerate HEP research** with the goal of solving the mysteries of matter, energy, space and time
- Developing **strategic capabilities** within the (inter)national AI ecosystem
  - AI to advance lab scientific mission, and where Fermilab can advance AI research
- Building **community** around cross-cutting problems, tools, and educational opportunities
  - By keeping a big-picture view of AI research and applications in and outside HEP, we connect teams across the lab and with teams at other labs/universities
  - Develop resources for AI research — both people (e.g. AI associate program) and hardware (e.g. GPU access)
- **Sharing** Fermilab AI related products with the world

AI for Physics ⇔ Physics for AI

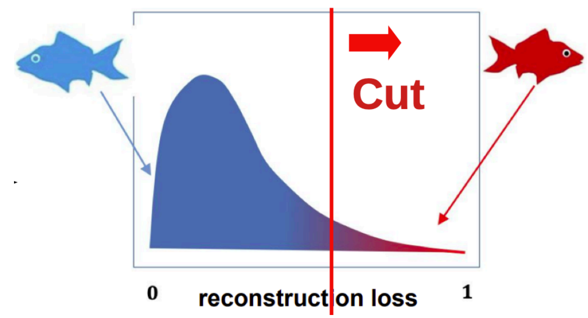
Outline

- AI for physics
    - Recent Highlights
- Physics for AI
    - Robust & Fast ML
    - AI @ Extreme Edge
- AI for user community
    - Computing Resources for AI training and inference
    - Engage with Fermilab AI community
        - Lab Wide AI meetings & Jamboree

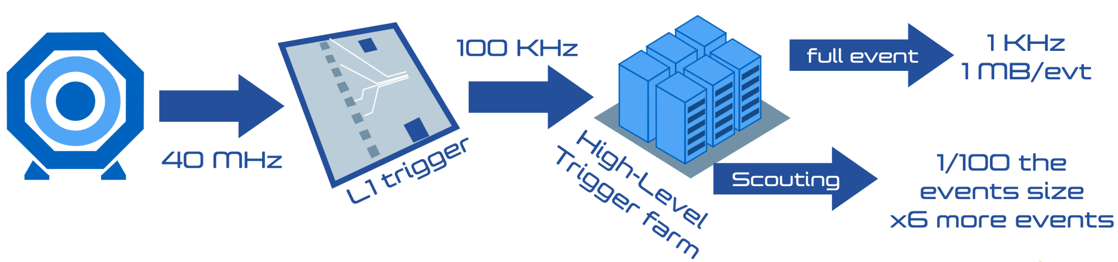
AI @ Energy Frontier: LHC triggers



- LHC detectors creates more data than we can handle !
  - Need to throw away 99.75% of data at first stage!
  - We are interested in rare physics processes
  - 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



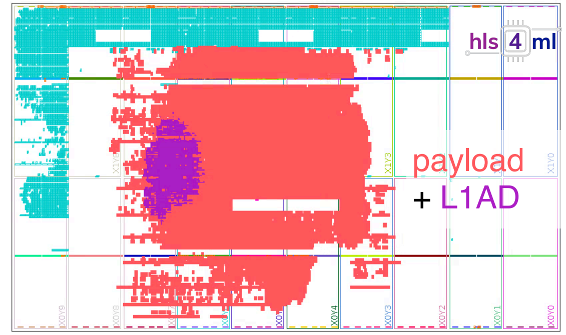
- Unsupervised ML technique such as Anomaly Detection can catch effectively the deviations from SM
  - Demonstrated for offline data analysis for new physics searches by 3-7x !
  - Triggering on "anomalousness" of collision event

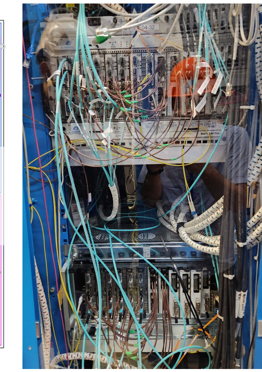


Records only ~ 0.01% of the data!

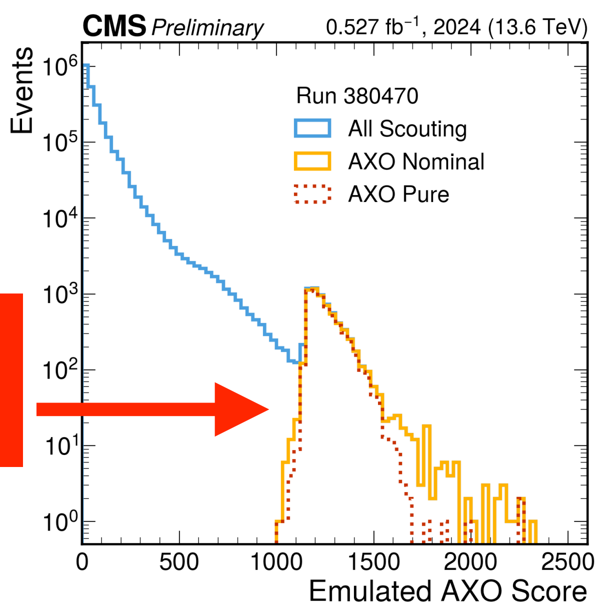
AI @ Energy Frontier: LHC triggers

- AXOL1TL: triggering on "anomalousness"
  - Trained a ML model called Autoencoder directly on data to find "atypical" signatures
- 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

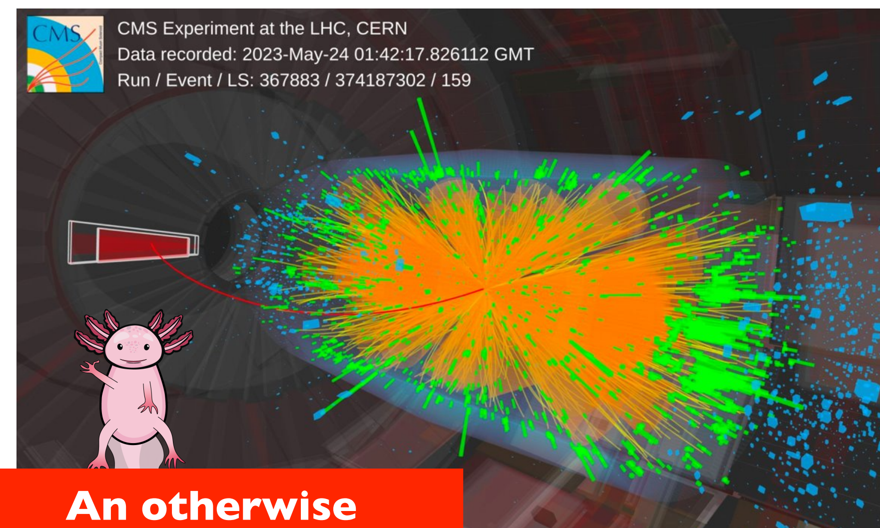




CMS-DP-2023-079
CMS-DP-2024-059



Otherwise untriggered events!



An otherwise untriggered high-multiplicity event!

AI @ Energy Frontier: fast simulation

- *Goal: address computational challenge of expensive simulation at (HL-)LHC experiments*
  - Diffusion based models to generate calorimeter shower simulations
  - SOTA model in CaloChallenge with a 10-1000x speed compared to **`Geant4`**

https://calochallenge.github.io/

Fast Calorimeter Simulation Challenge 2022

View on GitHub

Welcome to the home of the first-ever Fast Calorimeter Simulation Challenge!

The purpose of this challenge is to spur the development and benchmarking of fast and high-fidelity calorimeter shower generation using deep learning methods. Currently, generating calorimeter showers of interacting particles (electrons, photons, pions, ...) using GEANT4 is a major computational bottleneck at the LHC, and it is forecast to overwhelm the computing budget of the LHC experiments in the near future. Therefore there is an urgent need to develop GEANT4 emulators that are both fast (computationally lightweight) and accurate. The LHC collaborations have been developing fast simulation methods for some time, and the hope of this challenge is to directly compare new deep learning approaches on common benchmarks. It is expected that participants will make use of cutting-edge techniques in generative modeling with deep learning, e.g. GANs, VAEs and normalizing flows.

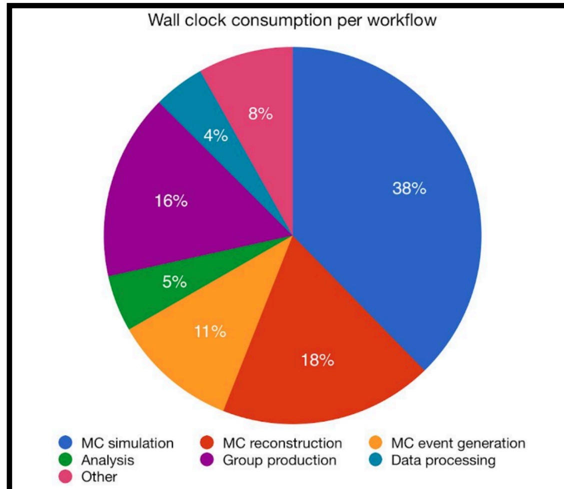
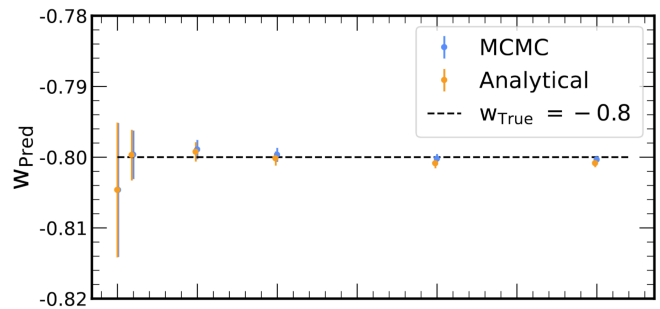


Figure 1: ATLAS CPU hours used by various activities in 2018

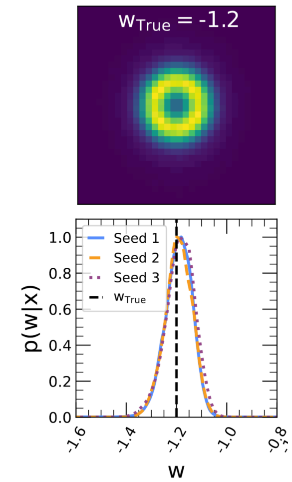
Many different generative models approaches being explored:
- Variational Autoencoders
- Generative Adversarial Networks
- Normalizing Flows
- Diffusion models

AI @ Cosmic Frontier: simulation-based inference

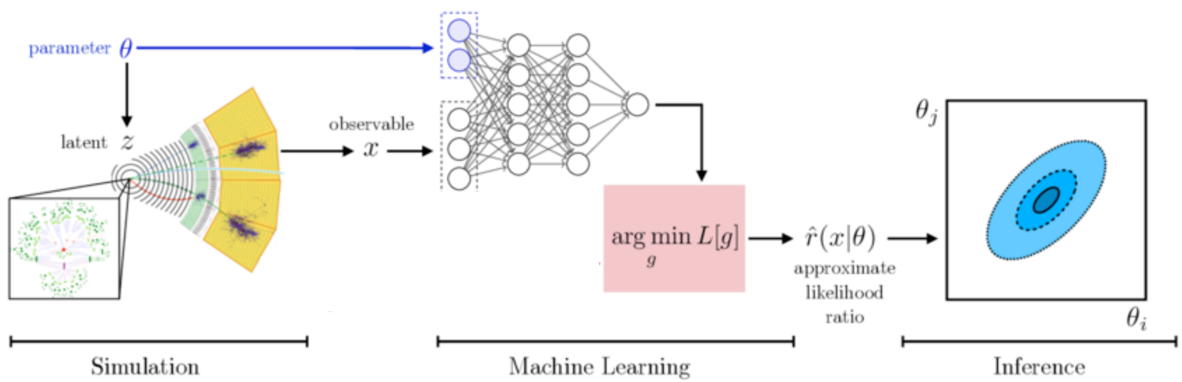
- Goal: infer the dark energy equation-of-state parameter w from a population of strong gravitational lens
  - Approximate an intractable likelihood with a Neural Network
  - Scalable for inference from O(1000) lenses from future surveys
  - Much faster than traditional MCMC



https://arxiv.org/abs/2407.17292



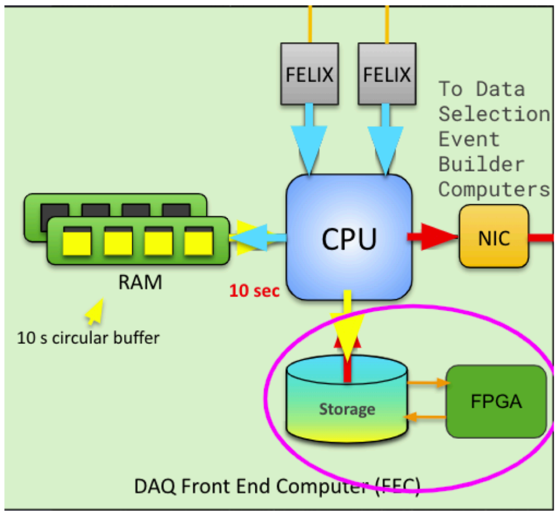
Neural Ratio Estimation [K. Cranmer et al. arxiv.1506.02169]

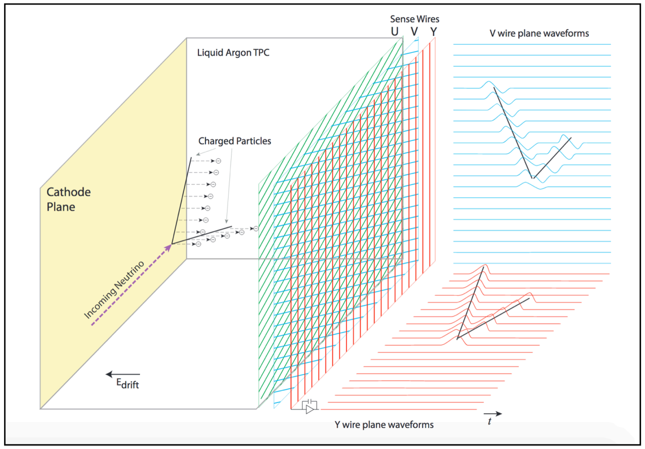


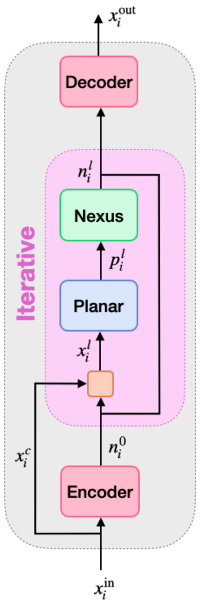
Fermilab

AI @ Intensity Frontier: LArTPC at DUNE

- ***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







- ***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

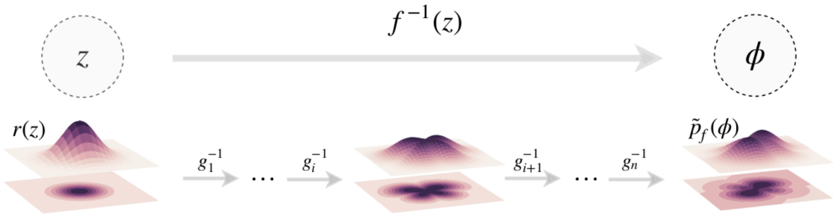
https://arxiv.org/html/2403.11872v1#S1

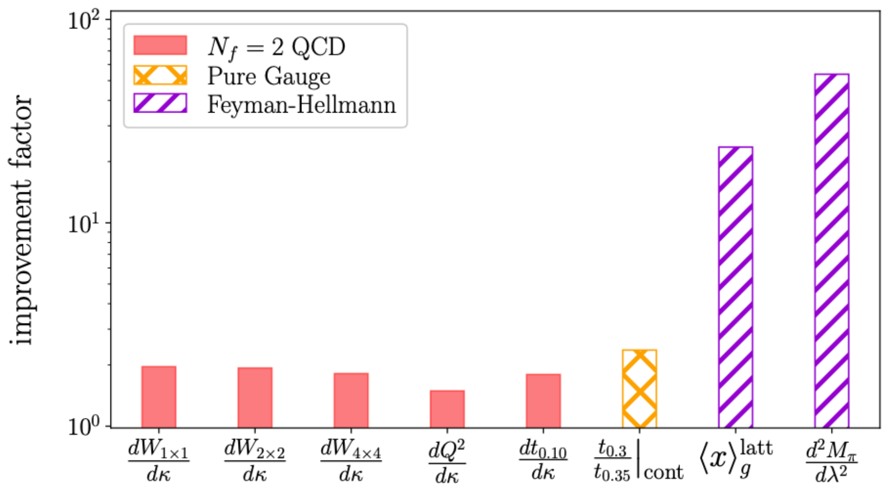
Fermilab

AI @ Theory Frontier

- *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 while accelerating the sampling of lattice field configuration

Normalizing flows can model complex distributions by transforming a simple distribution through a series of learned, invertible functions



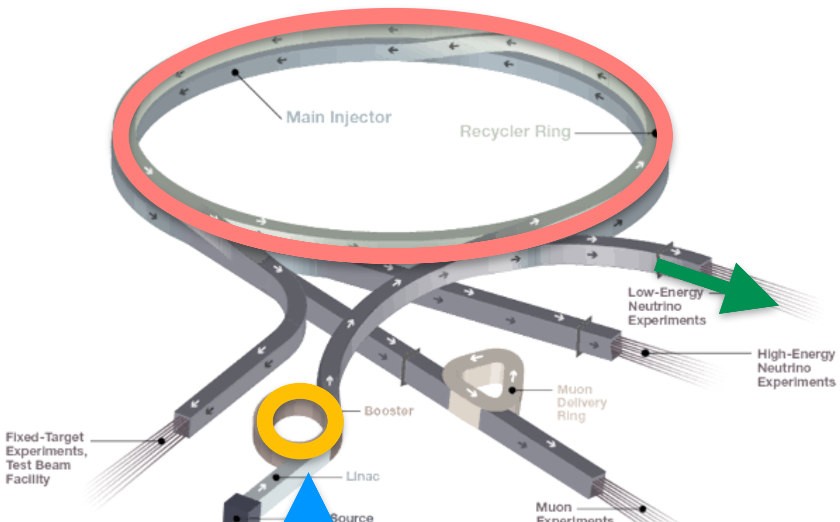


https://arxiv.org/pdf/2401.10874

AI @ Accelerator Frontier

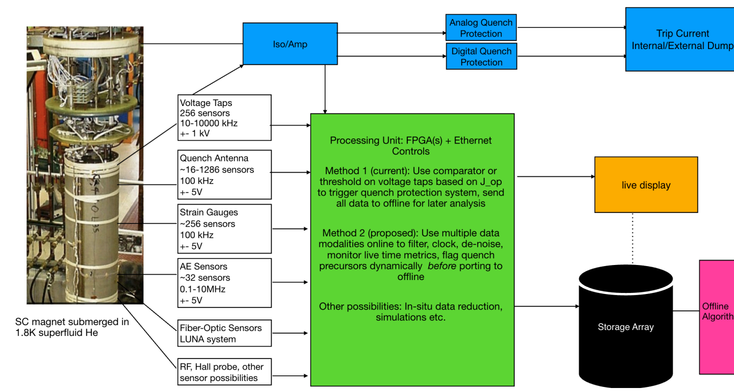
Real-time Edge AI Distributed System

- 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



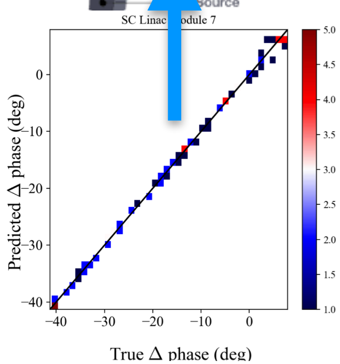
Magnet Quench Detection

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



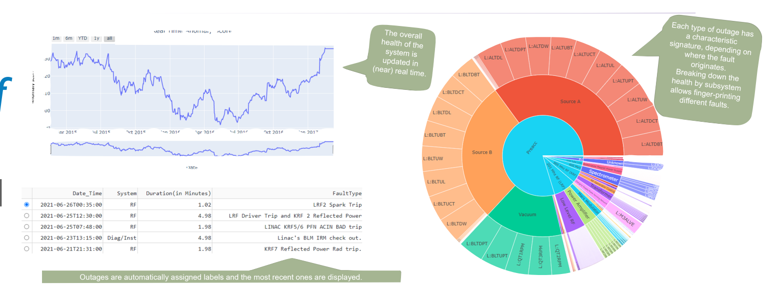
Linac RF Optimization

Predict RF parameters to keep beam energy constant and minimize emittance



Linac Condition Anomaly Prediction of Emergency

Predict anomalies and identify causing beam downtime



Physics for AI : Robust & Fast ML

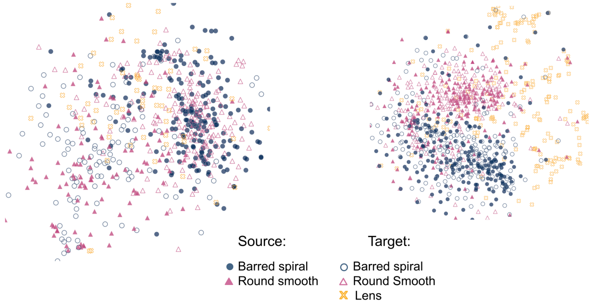
Robust Machine Learning

Domain Adaptation

Bridges difference between simulation & Obs. Data

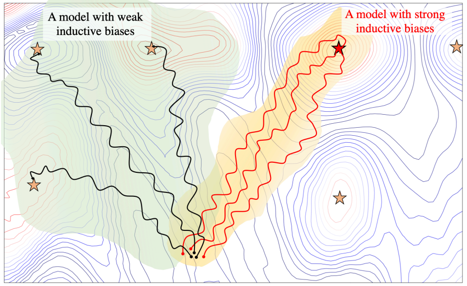
Illustris simulations ⟶ SDSS observations

Regular NN training ⟶ Domain Adaptation

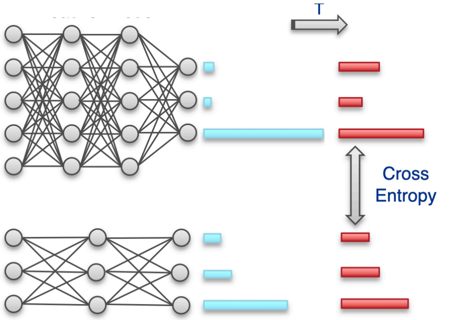


https://arxiv.org/abs/2302.02005

Robustness in Fast AI w/ Knowledge distillation of inductive bias



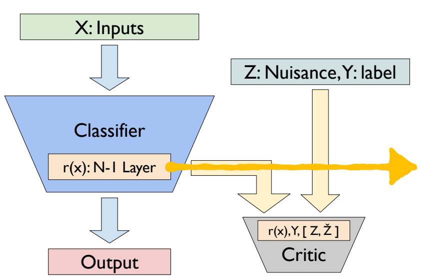
Include physics knowledge of the system into the fast and efficient ML models

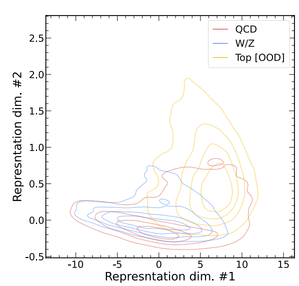


https://arxiv.org/abs/2311.14160

Nuisance invariant NNs w/ NuRD

Robust nuisance invariant Rep. learning

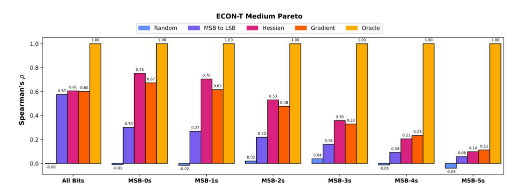


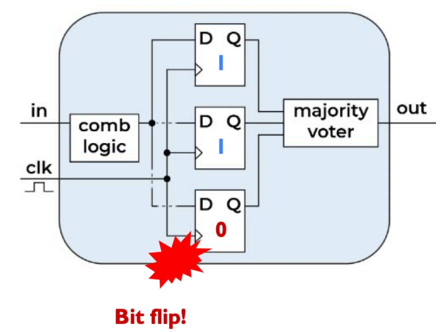


https://arxiv.org/abs/2401.08777

Robustness for NN on microelectronics

protects NNs on chip against bit flips in high radiation environments



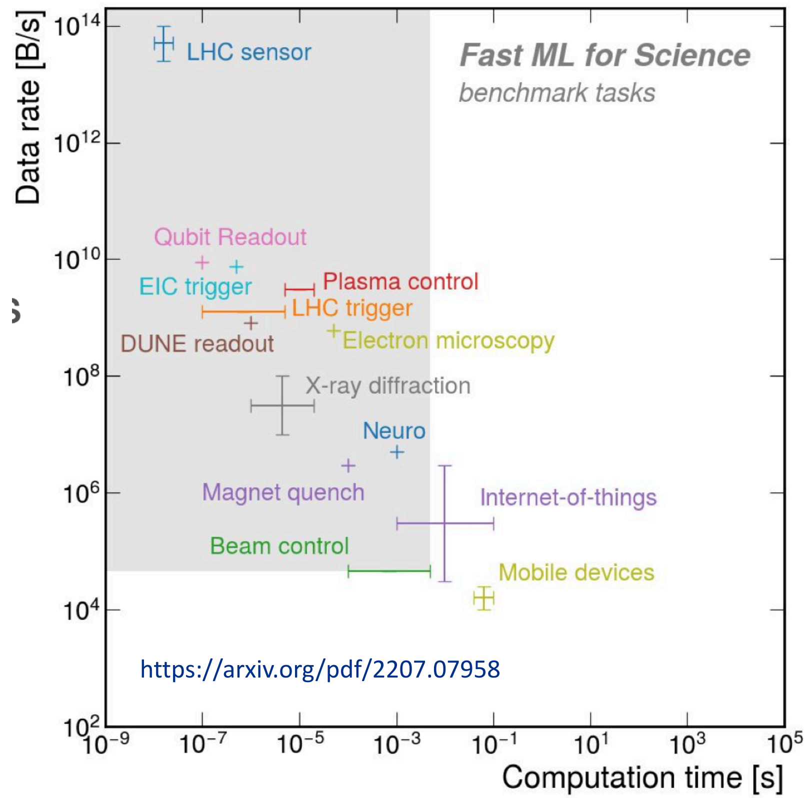


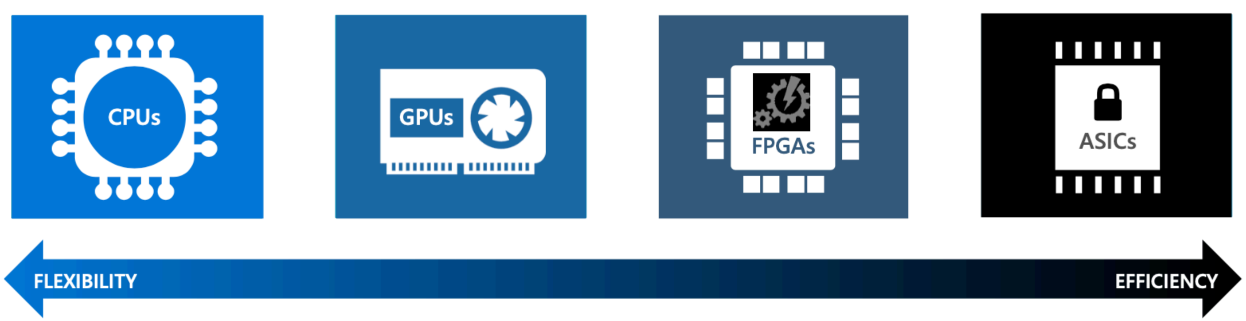
https://arxiv.org/abs/2406.19522

Fermilab

Fast Machine Learning

- 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

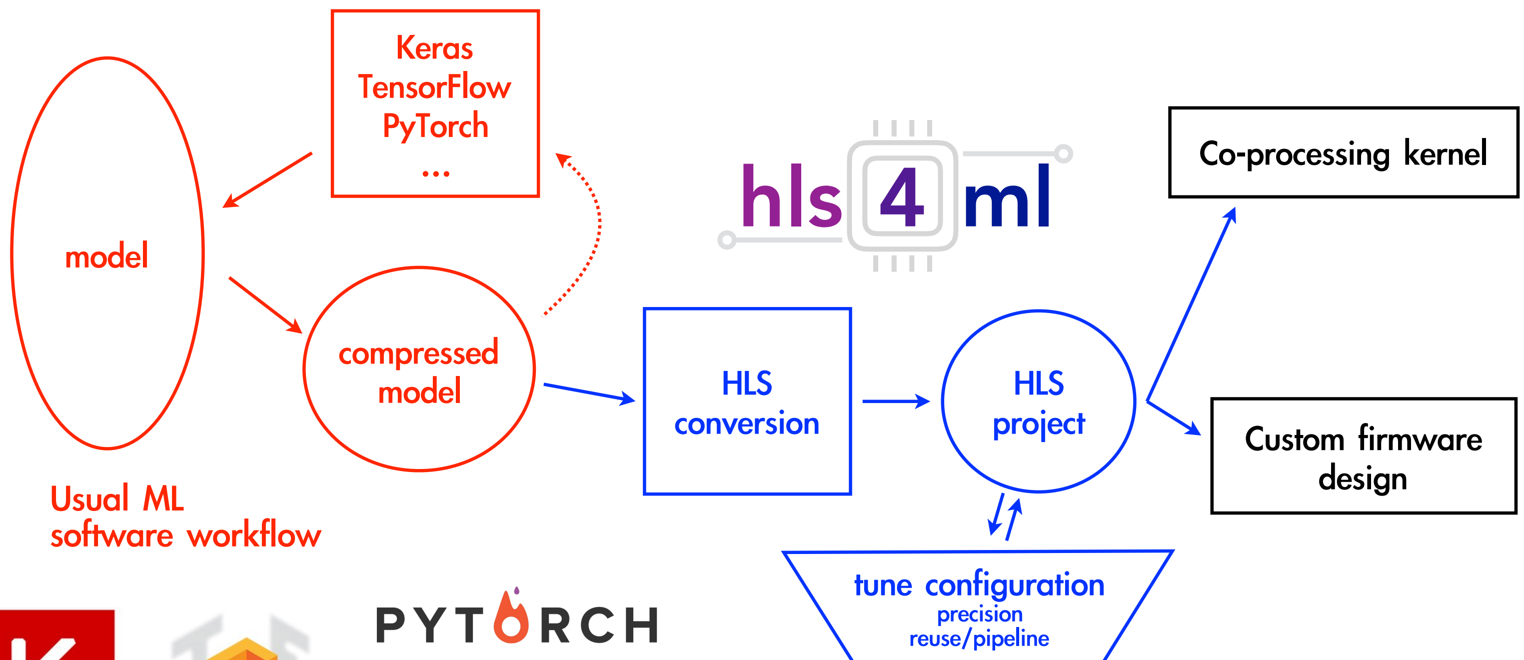




Bring ML models to hardware for real-time AI

high level synthesis for machine learning

A tool to efficiently program the FPGA hardware for Neural Networks with experimental constraints in mind!



Vivado™ HLS

HLS COMPILER

Catapult

Mentor® A Siemens Business

TensorFlow

PYTORCH

ONNX

https://fastmachinelearning.org/hls4ml/

Bring ML models to hardware for real-time AI

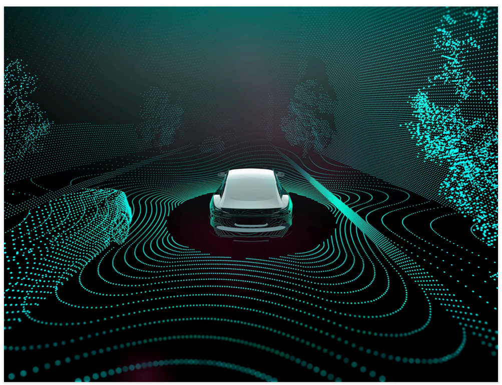
high level synthesis for machine learning

Sparking the interest of industry
(e.g., Google, Volvo, Siemens, AMD, …)

Colliding particles not cars: CERN's machine learning could help self-driving cars

CERN and software company Zenseact wrap up a joint research project that could allow autonomous-driving cars to make faster decisions, thus helping avoid accidents

25 JANUARY, 2023 | By Priyanka Dasgupta

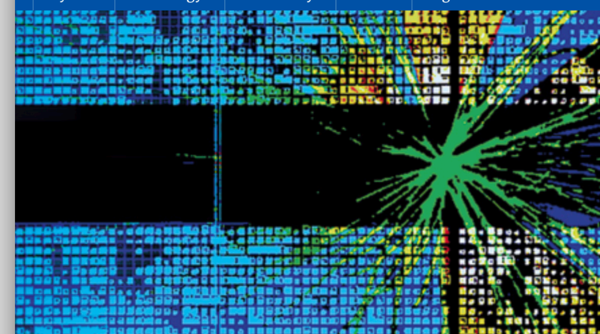


CERN's expertise in machine learning could help the field of autonomous driving (Image: Zenseact)

zenseact VOLVO CERN

CERNCOURIER Reporting on international high-energy physics

Physics ▾ Technology ▾ Community ▾ In focus Magazine



COMPUTING | FEATURE

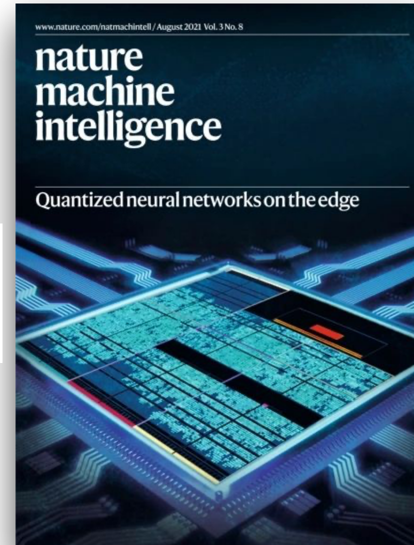
Hunting anomalies with an AI trigger

31 August 2021

Jennifer Ngadiuba and Maurizio Pierini describe how 'unsupervised' machine learning could keep watch for signs of new physics at the LHC that have not yet been dreamt up by physicists.

CALIFORNIA INSTITUTE OF TECHNOLOGY 1891

CERN



Siemens Digital Industries Software Newsroom Overview All news Blogs

PRESS RELEASE

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

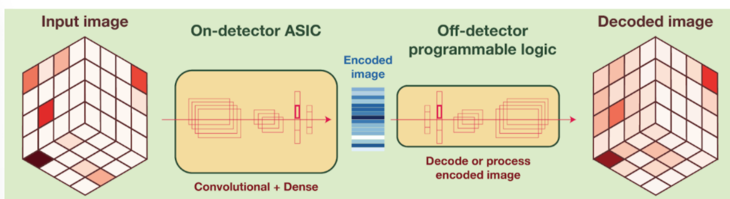
May 21, 2024
Plano, Texas

SIEMENS

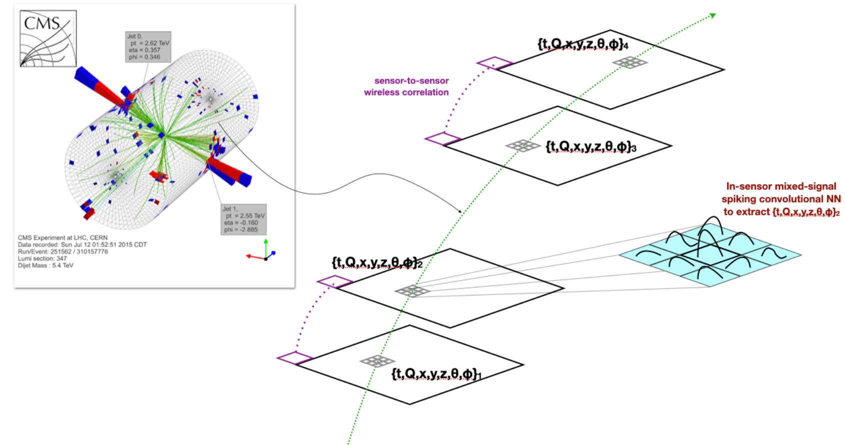
Fermilab

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



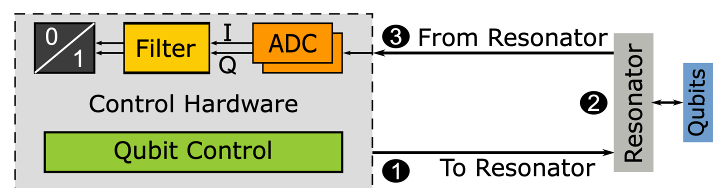
https://arxiv.org/abs/2105.01683



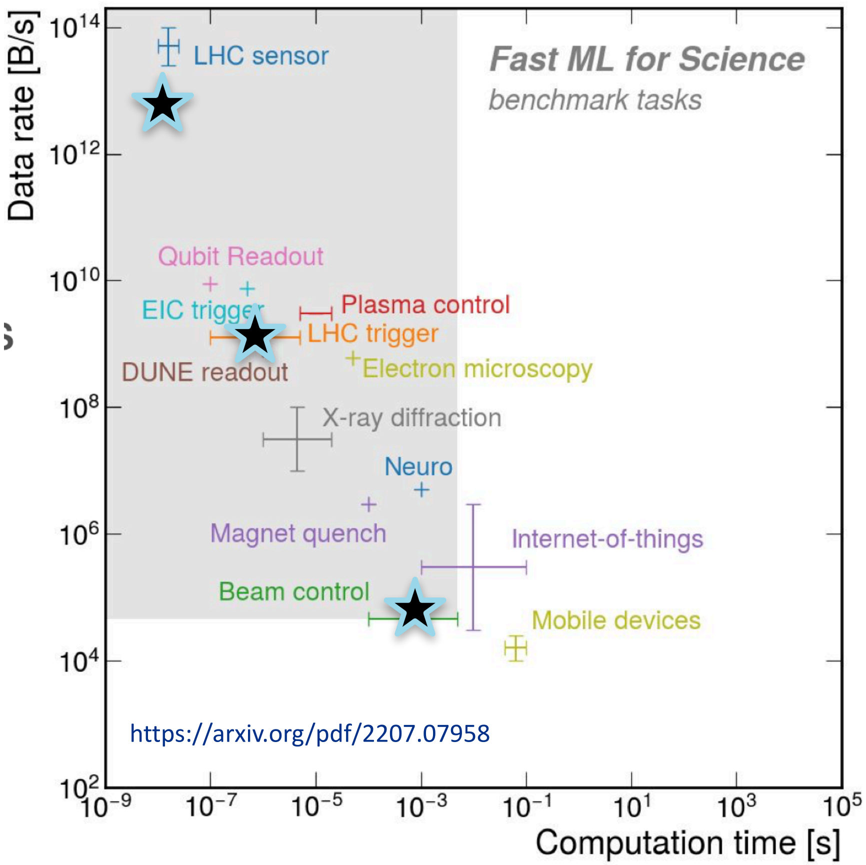
- Smart pixels: Pixel sensors w/ AI on chip
  - Efficiently filter low p_T tracks
  - Saving up to 75% of data bandwidth
  - Crucial for future colliders
    e.g: Reducing beam background in μC

https://arxiv.org/abs/2406.14860

- AI/ML for control and readout in quantum systems
  - Edge AI to improve readout of qubits
  - Denoising computations in theory calculations
  - Predicting quantum circuit fidelity on noisy hardware



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

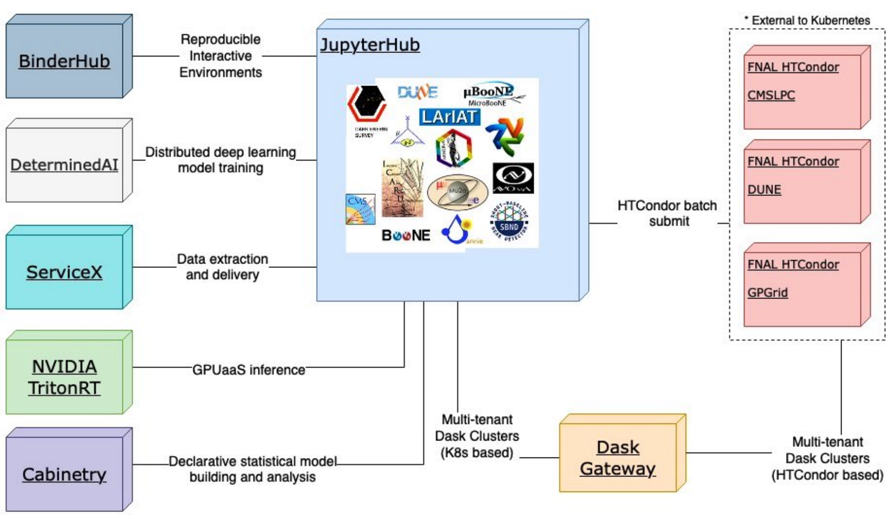
AI for Fermilab user community

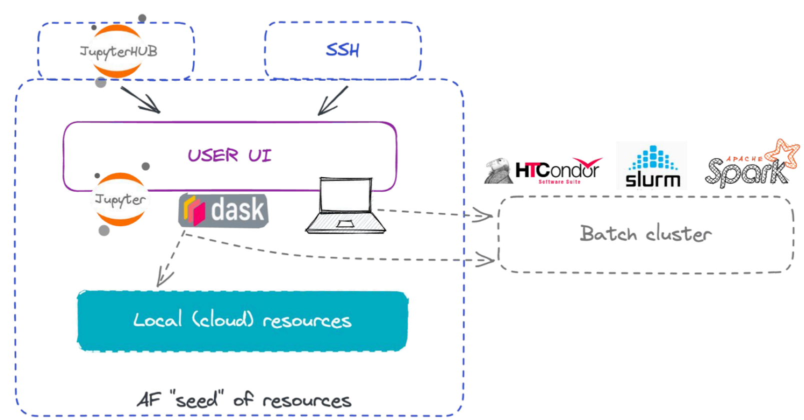
Elastic analysis facility ecosystem

- 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 up to O(100k) batch computing cores

Fermilab Elastic Analysis Facility Ecosystem



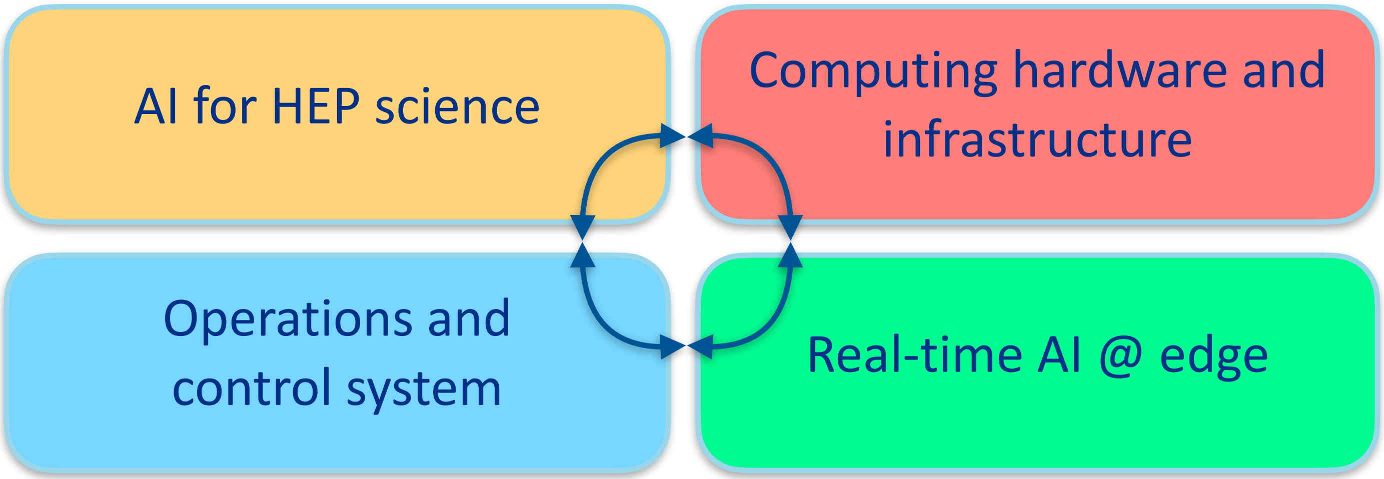


https://eafjupyter.readthedocs.io/en/latest/index.html

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

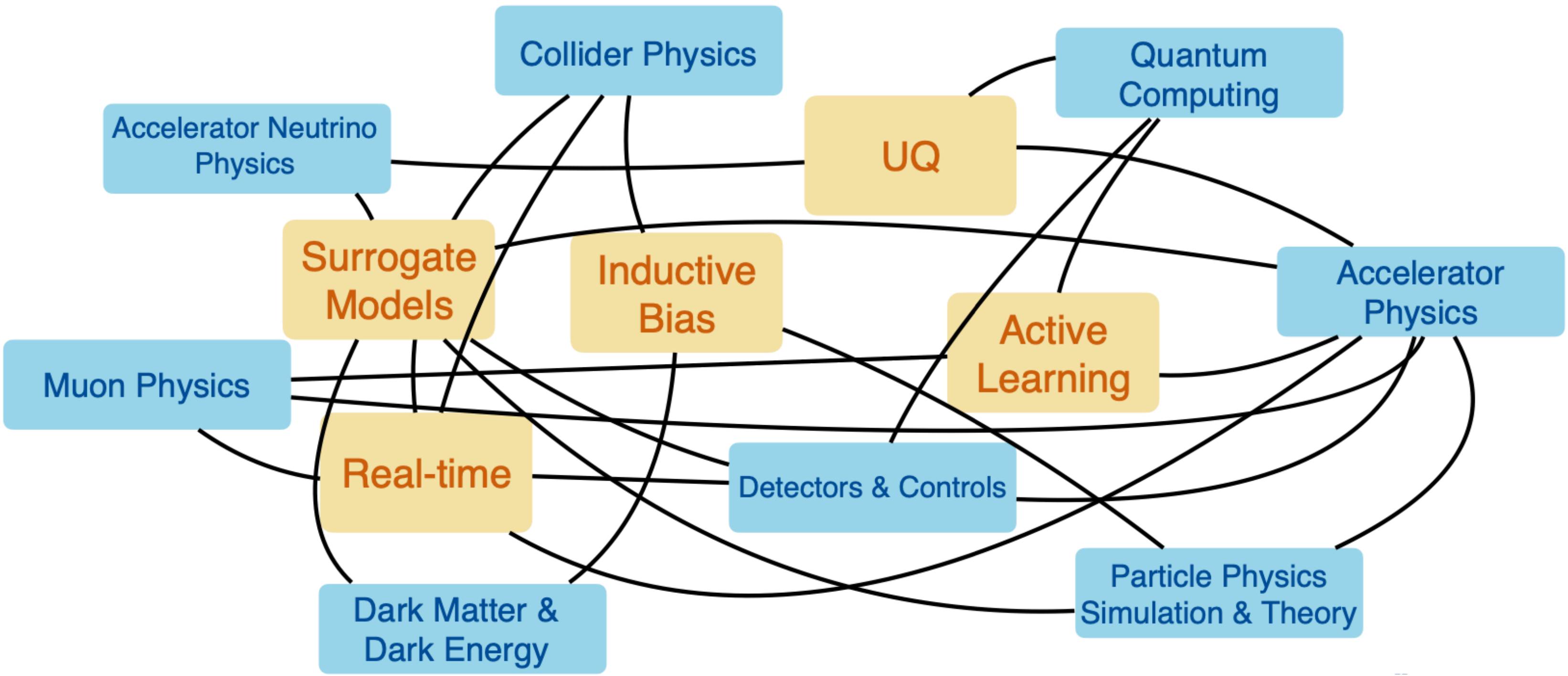
Landscape of AI @Fermilab



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.