



Status of AI/ML at Fermilab

Nhan Tran

for the AI Project Office: Farah Fahim, Burt Holzman, Tia Miceli, Brian Nord, Gabriel Perdue, Tingjun Yang

Fermilab PAC meeting 9 January 2024

Executive summary

Charge: Review the status of the AI/ML activities at the laboratory and of the recommendations made at past meetings: Formulate a strategy to respond to future AI/ML *research* calls, not necessarily just for AI/ML centers.

Framing: Al research is advancing rapidly; one primary area of Fermilab strength is in intelligent sensing and real-time efficient Al

Vision: Accelerate scientific discovery at unprecedented data scales while creating enabling technology for society

Mission: Efficient, robust, autonomous ML codesign

- A. Catalyze inclusive, multidisciplinary *Fast ML* community around grand challenges and benchmark tasks
- B. Leverage relevant Fermilab core capabilities and strengths to build tools to support the community

Strategy:

- A. Identify and grow appropriate sustainable funding streams to support community tools
- B. Advance cutting-edge intelligent sensing, real-time AI research
- C. Develop industry/academic partnerships to support the core mission



Key performance indicators:

- 1. Sustainable funding sources for supporting community tools and users on 2 year timescale
- New and existing partnerships & collaborations resulting in: research output; new projects on Al technology and research; technology transfer; and community growth (users, downloads, etc.)



Outline

- Framing AI in the world, at the DOE
- Dual Visions & Fast ML
- Fast ML Mission
 - 2 core elements of the strategy
- Strategy & Key Performance Indicators

Charge -

We ask the PAC to review the status of the AI/ML activities at the laboratory and of the recommendations made at past meetings: Formulate a strategy to respond to future AI/ML calls, not necessarily just for AI/ML centers.



AI in the world and at the DOE

• Al is a big space: industry & academia

- The DOE AI strategy has gone through multiple evolutions
 - Recent excitement on foundation models
 - ChatGPT came out in late 2022
- The leadership class HPC facilities are a key resource (ORNL, ANL, LBNL)
 - most recently some effort to nucleate around the Trillion Parameter Consortium (TPC)
- Where does HEP mission and Fermilab fit into this picture?

ADVANCED RESEARCH **DIRECTIONS ON** AI FOR SCIENCE, ENERGY, AND SECURITY Report on Summer 2022 Workshops Jonathan Carter Lawrence Berkelev National Laboratory John Feddema Sandia National Laboratories Doug Kothe Oak Ridge National Laboratory Rob Neely Lawrence Livermore National Laboratory Jason Pruet Los Alamos National Laboratory **Rick Stevens AI APPROACHES** Argonne National Laboratory New AI-Empowered Computing Paradigms, known in this report as AI Approaches The scale of data and computation for training AI models is opening the potential today for new paradigms in computation, including the following AI Approaches: 01. Al and Surrogate Models for Scientific Computing 02. AI Foundation Models for Scientific Knowledge Discovery, Integration, and Synthesis 03. Al for Advanced Property Inference and Inverse CENERGY CENERGY Office of N Design 04. AI-Based Design, Prediction, and Control of Complex Engineered Systems 05. Al and Robotics for Autonomous Discovery

06. Al for Programming and Software Engineering



Outline

- Framing AI in the world, at the DOE
- Dual Visions & Fast ML
- Fast ML Mission
- Strategy & Key Performance Indicators





We have consistently focused on:

Al for physics ⇔ physics for Al

- Develop AI capabilities to accelerate HEP science and contribute greater science + industry AI ecosystem
- Build diverse, inclusive community; assemble multi-disciplinary collaborations around cross-cutting HEP AI challenges



We have consistently focused on:



Al for physics ⇔ physics for Al

This space is massive -

In <u>2023 PAC talk</u>, we summarized all the exciting activities at Fermilab.

It continues to be our goal to support **all directions of Al research** to advance HEP science, e.g. see recent FNAL Al Jamboree



We have consistently focused on:



Al for physics ⇔ physics for Al

This space is massive –

In <u>2023 PAC talk</u>, we summarized all the exciting activities at Fermilab.

It continues to be our goal to support **all directions of Al research** to advance HEP science, e.g. see recent FNAL Al Jamboree

screenshot from last PAC AI talk

Al program in ~15 minutes

- Algorithms for HEP science
 - · Physics-inspired models and data
 - Graph learning
 - Generative models
 - SBI/likelihood-free inference
 - Accelerating theory
 - Robust and generalizable learning
 - Domain adaptation
 - Anomaly detection
 - Semi-/self-supervision
 - Fast and efficient algorithms
 - Multi-objective optimization
 - Quantization/sparsity
 - Knowledge distillation

Operations and controls

- · Real-time accelerator controls
- · Telescope design and operations
- Quantum machine learning
- Computing hardware and infrastructure
 - · Resources for AI practitioners
 - Efficient AI-in-production
- Real-time systems at the edge
 - Hardware-algorithm codesign for HEP and beyond
 - Near-detector, low latency AI
 - On-sensor/detector Al



We have consistently focused on:

Al for physics ⇔ physics for Al

What is our unique value proposition in the AI space as it pertains to the PAC charge?

Accelerate scientific discovery at unprecedented data scales

This will be the focus of this talk.



We have consistently focused on:

Al for physics ⇔ physics for Al

What is our unique value proposition in the AI space as it pertains to the PAC charge?

Accelerate scientific discovery at unprecedented data scales

This will be the focus of this talk.

The strategy:

- is complementary to HPCs (e.g. edge vs cloud)
- leverages unique HEP strengths in cutting-edge sensing technology
- cuts across many scientific domains and industry



Fast ML for Science vision:

Accelerate scientific discovery at unprecedented data scales

"

Scientific discoveries come from groundbreaking ideas and the capability to validate those ideas by testing nature at new scales—finer and more precise temporal and spatial resolution. This is leading to an explosion of data that must be interpreted, and ML is proving a powerful approach. The more efficiently we can test our hypotheses, the faster we can achieve discovery. To fully unleash the power of ML and accelerate discoveries, it is necessary to embed it into our scientific process, into our instruments and detectors.

"

Applications and Techniques for Fast Machine Learning in Science

Core ML Mission: Efficient, robust, autonomous ML codesign



Outline

- Framing AI in the world, at the DOE
- Dual Visions & Fast ML
- Fast ML Mission
- Strategy & Key Performance Indicators





Mission: Efficient, robust, autonomous ML codesign

- Catalyze inclusive, multidisciplinary Fast ML community around grand challenges and benchmark tasks
- Leverage strength and scale of national laboratories to develop critical technologies that support the community

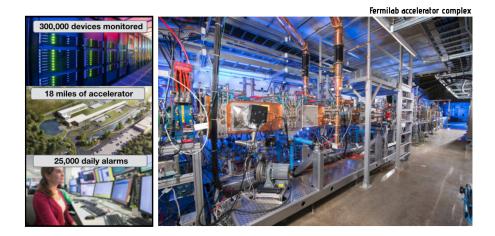


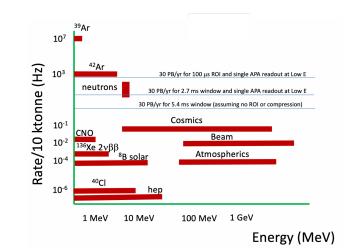
Grand challenges for HEP, examples

Optimal, continuous readout for DUNE for neutrino physics, multi-messenger astronomy, and other rare measurements

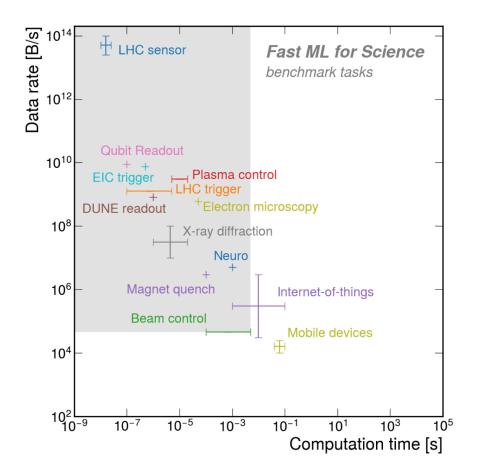
Analyze all 40 MHz of LHC data for the full detector for new physics searches, Higgs measurements, and more

Al-assisted, real-time operation of the Fermilab accelerator complex



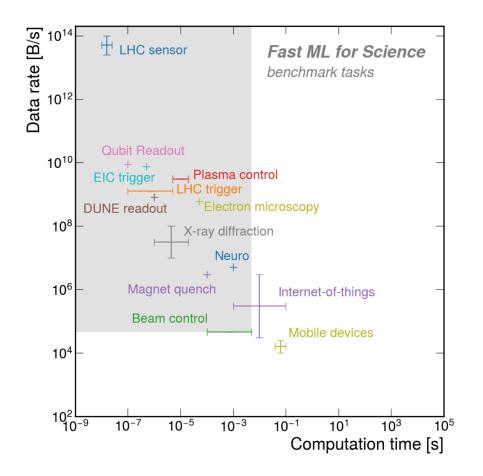


🚰 Fermilab



Grand challenges spark imaginations! Benchmarks bring innovation





Grand challenges spark imaginations! Benchmarks bring innovation

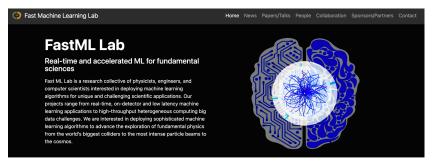
Benefits to HEP: bring new resources to bear on HEP grand challenges (industry partnerships, computer science & engineering researchers)

HEP-born technology brings **transformative technology** to new material research, fusion energy, neuroscience, or industry applications and so on...



The Fast ML community

- Weekly meetings
- Annual workshop
 - 1st edition at FNAL, now on 5th edition in 2024 (bidding for venue now)
 - 1st ICCAD (computer aided design workshop) this year
- **Projects** supported from DOE and NSF coming from the community
 - Example: NSF HDR A3D3 institute
 - DOE funding from HEP, ASCR, NP, SBIR
 - will discuss in more detail later
 - Includes strong local university connections: IIT, Northwestern, Purdue, UC, UIC, UIUC,...
 - Includes international collaborators
- Connected to wider communities
 - MLCommons
 - Microelectronics initiatives
 - Industry collaborations and engagements





The Fast ML community

Fast ML for Science @ ICCAD, 1st ed. (screenshot)

Time (PDT)	Duration	Presentations		
8:15	15'	Welcome and introduction Javier Duarte, UCSD		[Slides]
8:30	30'	Community Vision, Needs, and Progress Vladimir Loncar, MIT		[Slides]
9:00	30'	Design Tools Perspective: Catapult + hls4ml for Inference at the Edge David Burnette, Siemens		[Slides]
9:30	30'	Designing Hardware for Machine Learning John Wawrzynek, UC Berkeley		[Slides]
10:00	30'	Coffee		
10:30	30'	Design Tools Perspective: Mapping ML to the AMD RyzenAl Architecture Elliott Delaye, AMD		[Slides]
11:00	30'	Fast ML in the NSF HDR Institute: A3D3 Shih-Chieh Hsu, UW		[Slides]
11:30	30'	Real-time ML at the Linac Coherent Light Source Jana Thayer, SLAC		[Slides]
12:00	60'	Lunch		
1:20	30'	Robust and Efficient Machine Learning for Mission-Critical Applications Bhavya Kailkhura, LLNL		[Slides]
1:50	20'	Quantifying the Efficiency of High-Level Synthesis for Machine Learning Inference <u>Caroline Johnson (UW)</u> , Scott Hauck, Shih-Chieh Hsu, Waiz Khan, Stephany Ayala-Cerna, Geoff Jones, Anatoliy Martynyuk, Matthew Bavier, Oleh Kondratyuk, Trinh Nguyen, Jan Silva, Aidan Short (UW)	[Paper]	[Slides]

Fast ML fo	or Science, 4th ed. (screenshot)	for offline
Fast Mac	hine Learning at the Large Hadron Collider experiments	reading only
Blackett I	aboratory, Lecture Theatre 1, Imperial College London	09:45 - 10:30
Break		
Blackett I	aboratory, Foyer, Imperial College London	10:30 - 11:00
Fast Mac	hine Learning at European XFEL	Steve Aplin
Blackett I	aboratory, Lecture Theatre 1, Imperial College London	11:00 - 11:4
Bridging	Al and biomedicine: Towards Al-driven scientific discoveries	Maria Brbi
Blackett I	aboratory, Lecture Theatre 1, Imperial College London	11:45 - 12:30
Fast Ma	chine Learning for accelerator control	Karin Rathsman

Blackett Laboratory, Lecture Theatre 1, Imperial College London	09:00 - 09:45
Fast Machine Learning for laser wakefield acceleration	Matt Streeter 🥝
	-
Blackett Laboratory, Lecture Theatre 1, Imperial College London	09:45 - 10:30
Break	
Blackett Laboratory, Foyer, Imperial College London	10:30 - 11:00
Fast ML for fusion simulation, optimization, and control	Jonathan Citrin 🥝
Blackett Laboratory, Lecture Theatre 1, Imperial College London	11:00 - 11:45
Machine Learning in Exoplanet Characterisation	Ingo Waldmann et al. 🖉



09:45 - 10:30

10:30 - 11:00 Steve Aplin 0

> 11:00 - 11:45 Maria Brbic

11:45 - 12:30 Karin Rathsman 🥝

Mission: objectives

- Catalyze inclusive, multidisciplinary Fast ML community around grand challenges and benchmark tasks
- Leverage strength and scale of national laboratories to develop critical technologies that support the community



Core Fermilab AI capability, strong synergy with microelectronics initiative

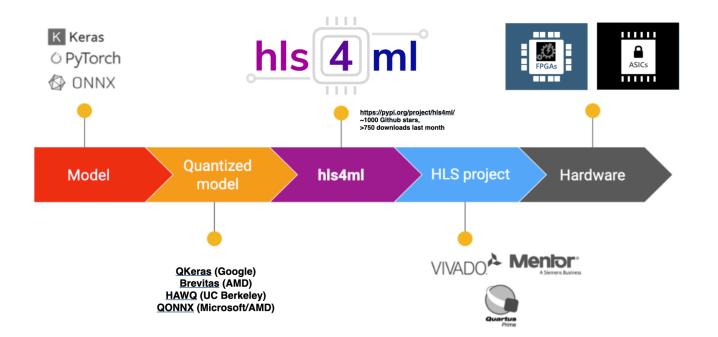
Synergies with different AI research areas across Fermilab, focus area for growth

Efficient, robust, autonomous ML codesign



Efficient, robust, autonomous ML codesign

Efficient {algorithms, tools, hardware, workflows, collaborations} for ML codesign





Efficient, robust, autonomous ML codesign

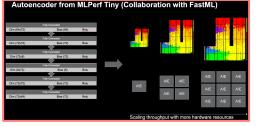
for offline reading only

Efficient {algorithms, tools, hardware, workflows, collaborations} for ML codesign

General techniques: sparsity, quantization Physics informed techniques: distillation, inductive bias

Knowledge Distillation		
Ryan Liu	Abhijith Gandrakota	
University of California, Berkeley	Fermi National Accelerator Laboratory	
Berkeley, CA 94720	Batavia, IL 60510	
Jennifer Ngadiuba	Maria Spiropulu	
Fermi National Accelerator Laboratory	California Institute of Technology	
Batavia, IL 60510	Pasadena, CA 91125	



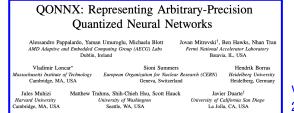


recent ICCAD workshop

SIEMENS

Design Tools Perspective:
Catapult + HLS4ML for
Inference at the Edge
David Burnette Director of Engineering – Catapult

Short Courses	
- SC1 : Real-time m	achine learning on FPGAs (hls4ml)
Course title:	Real-time machine learning on FPGAs (hls4ml)
Course organizer:	Ben Hawks, Fermilab
Date/time/venue:	Saturday, November 4 – 8:00 am – 6:00 pm – 109 on Level 1



with AMD research, 20k downloads last month! End-to-end codesign of Hessian-aware quantized neural networks for FPGAs and ASICs

Siemens EDA

JAVIER CAMPOS, JOVAN MITREVSKI, and NHAN TRAN, Fermi National Accelerator Laboratory, USA ZHEN DONG, AMIR GHOLAMI⁺, and MICHAEL W. MAHONEY[†], University of California Berkeley, USA

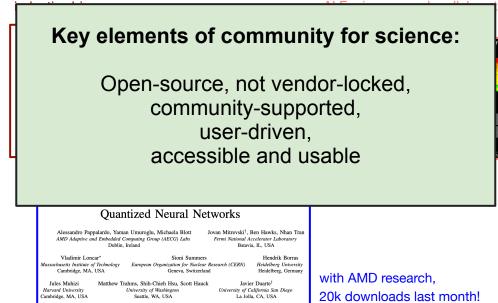
JAVIER DUARTE, University of California San Diego, USA



Efficient, robust, autonom

Efficient {algorithms, tools, hardware, workflov

General techniques: sparsity, quantization Physics informed techniques: distillation,



	offline
import hls4ml	ig only
# Configure the converter	
<pre>config = hls4ml.utils.config_from_keras_model(model, granularity='mov print(config)</pre>	del')
# Convert the model	ksho
<pre>hls_model = hls4ml.converters.convert_from_keras_model(model,</pre>	/e:
hls_config=co	nfig,
output_dir='h	
fpga_part='xc	vu9p-fl
# Print the model configuration to check	
<pre># Pint the model configuration to check hls4ml.utils.plot_model(hls_model, show_shapes=True, show_precision=""""""""""""""""""""""""""""""""""""</pre>	True. t
In the `hls4ml.converters.convert_from_keras_model` function:	EME
`hls_config`: is the configuration for the conversion. We generated it using	
`config_from_keras_model` for simplicity, but you can customize this as neede	d
• `output_dir`: is the directory where the HLS project will be created.	
• `fpga_part`: specifies the FPGA part. Adjust this based on your target FPGA.	
Build the HLS Project:	
Once the model is converted, you can compile it into HLS:	on Leve
python 🗋 Co	py code
hls_model.compile()	s
Run the HLS Simulation (Optional):	
Before synthesizing for FPGA, you can run a C-simulation to check if the model work	ks USA
correctly in HLS:	celey,
	py code
hls_model.build(csim=True)	
After this, you'll have an HLS project in the specified `output dir ` that you can use y	

FPGA development tools to generate bitstreams for FPGA deployment.

3. Convert with hls4ml:

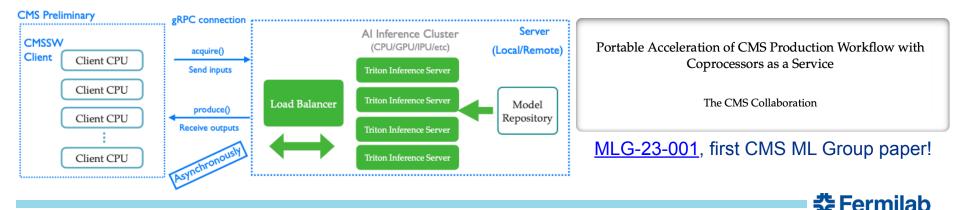
21

Efficient, robust, autonomous ML codesign

Efficient {algorithms, tools, hardware, workflows, collaborations} for ML codesign

Another example: system codesign for heterogeneous grid computing to accelerate ML workflows

- 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
- Leverage industry hardware and tools provide coprocessors as-a-service
- SONIC: Services for Optimized Network Inference on Coprocessors





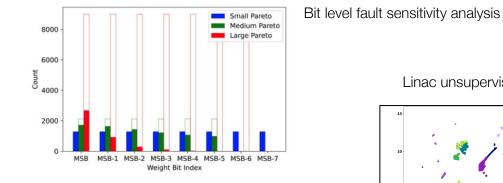
Efficient, robust, autonomous ML codesign

Example images of simulated galaxy morphologies with different levels of telescope noise.

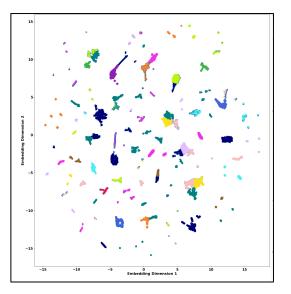
Anomaly detection algorithm

running online test bench on

CMS Run 3 data!



CMS Preliminary 0.467 fb-1, 2023 (13.6 TeV) Rate [Hz] AXOL1TL Score > 250 L1 Physics Rate Score > 1250 Single Muon Trigge AYOI 1TL Score > 25 ารางการสุขมันสุของทระรุษัทให้สุขมาสกรณ์เป็นไปสารสินจารูไฟ 00:00 02:00 04:00 05:00 06:00 03:00 07-Jun Time [UTC] Linac unsupervised fault clustering





Elliptical

Impact: recent examples

Leverage core capabilities to deploy **ML** <u>at scale</u> - algorithms + facilities, tools, software, multidisciplinary teams

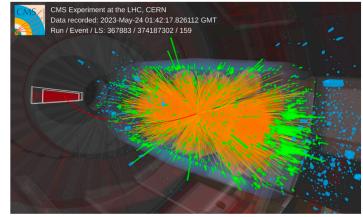
e.g. large scale user facilities & advanced instrumentation; advanced computer science, visualization, & data; microelectronics



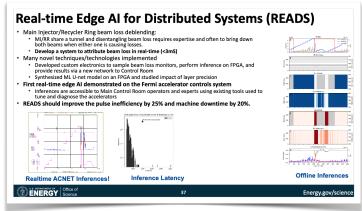
Impact: recent examples

Leverage core capabilities to deploy **ML** <u>at scale</u> - algorithms + facilities, tools, software, multidisciplinary teams *e.g. large scale user facilities & advanced instrumentation; advanced computer science, visualization, & data; microelectronics*

- First ever L1 trigger anomaly detection algorithm deployed for LHC CMS Run 3
 - Growth from community benchmarks and collaborations built from community efforts, investment in hls4ml (FastML, AMD, Siemens)
- CMS MLG-23-001 demonstration of accelerated ML workflows with SONIC; working with NVidia, Graphcore, computing operations experts
- First edge AI deployed in Fermilab accelerator complex; working with Intel/NU



Rameika, <u>HEPAP Aug23</u>

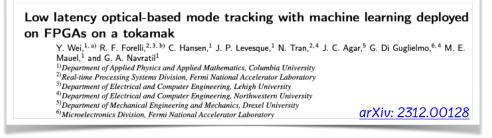


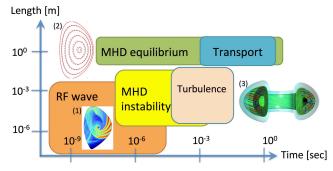


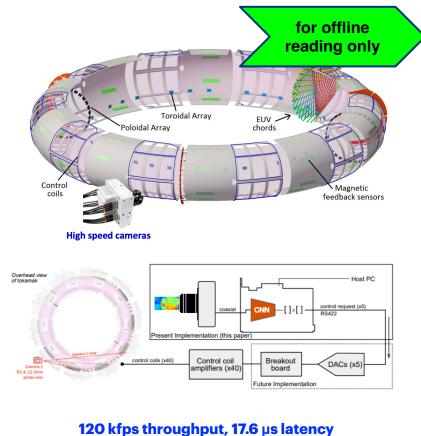
Impact: recent examples

Leverage core capabilities to deploy **ML** <u>at scale</u> - algorithms + facilities, tools, software, multidisciplinary teams

e.g. large scale user facilities & advanced instrumentation; advanced computer science, visualization, & data; microelectronics







Enabling new capabilities for fusion experiments!



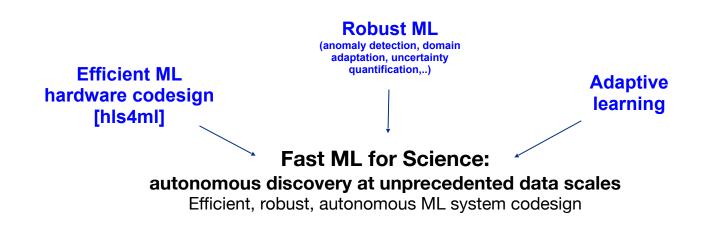
Outline

- Framing AI in the world, at the DOE
- Dual Missions & Fast ML
- Fast ML Vision
- Strategy & Key Performance Indicators



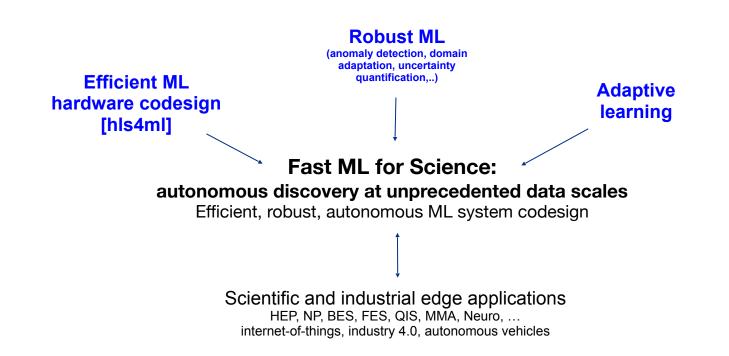


Fast ML ecosystem



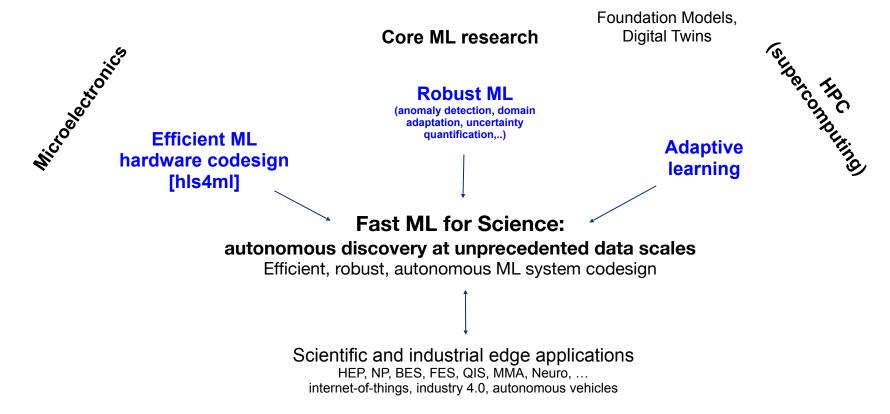


Fast ML ecosystem

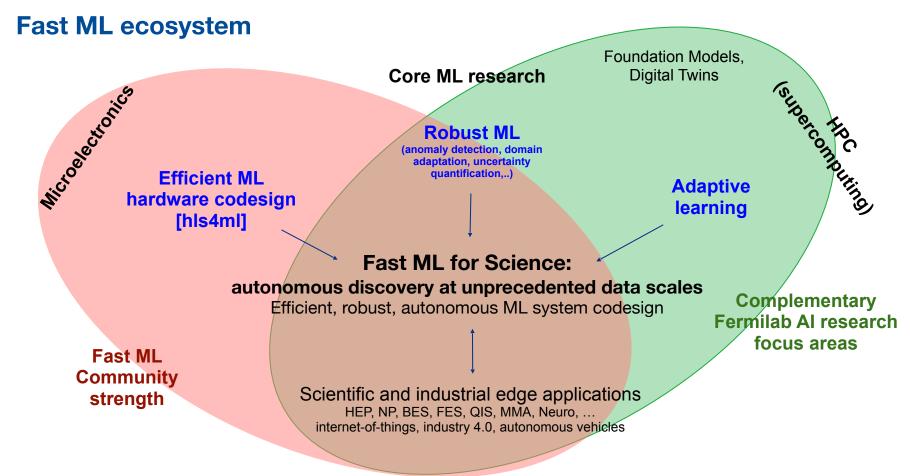




Fast ML ecosystem









Project support and funding strategy

Support from DOE and community focused around research topics

- DOE HEP (Lab and University awards), ASCR, NP
- Additional sources: JTFI (UChicago/ANL), DPI (UIIlinois system), LDRD
- NSF include HDR institute (A3D3)

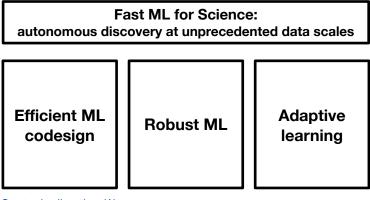
Strategic directions for support:

A. Identify and develop sustainable funding streams to support broadening community tools and techniques

B. Advance cutting-edge intelligent sensing, real-time AI research and hardware codesign C. Develop strategic industry/academic partnerships to support the core mission



Strategic growth

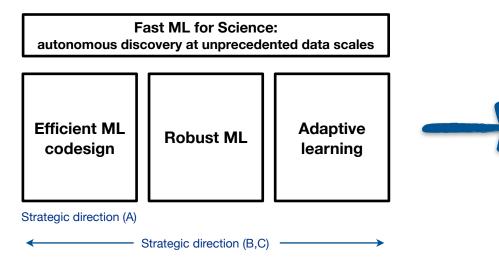


Strategic direction (A)

Strategic direction (B,C)



Strategic growth



Grand challenges in HEP:

- Optimal, continuous readout for DUNE for neutrino physics, multi-messenger astronomy, and other rare measurements
- Analyze all 40 MHz of LHC data for the full detector for new physics searches, Higgs measurements, and more
- Al-assisted, real-time operation of the Fermilab accelerator complex

And more in HEP and beyond!



Executive summary

Charge: Review the status of the AI/ML activities at the laboratory and of the recommendations made at past meetings: Formulate a strategy to respond to future AI/ML *research* calls, not necessarily just for AI/ML centers.

Framing: Al research is advancing rapidly; one primary area of Fermilab strength is in intelligent sensing and real-time efficient Al

Vision: Accelerate scientific discovery at unprecedented data scales while creating enabling technology for society

Mission: Efficient, robust, autonomous ML codesign

- A. Catalyze inclusive, multidisciplinary *Fast ML* community around grand challenges and benchmark tasks
- B. Leverage relevant Fermilab core capabilities and strengths to build tools to support the community

Strategy:

- A. Identify and grow appropriate sustainable funding streams to support community tools
- B. Advance cutting-edge intelligent sensing, real-time AI research
- C. Develop industry/academic partnerships to support the core mission



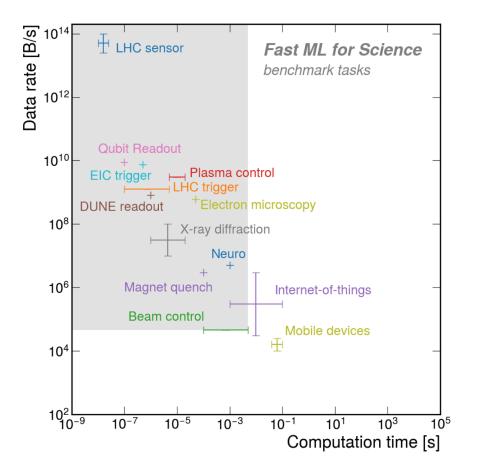
Key performance indicators:

- 1. Sustainable funding sources for supporting community tools and users on 2 year timescale
- New and existing partnerships & collaborations resulting in: research output; new projects on Al technology and research; technology transfer; and community growth (users, downloads, etc.)

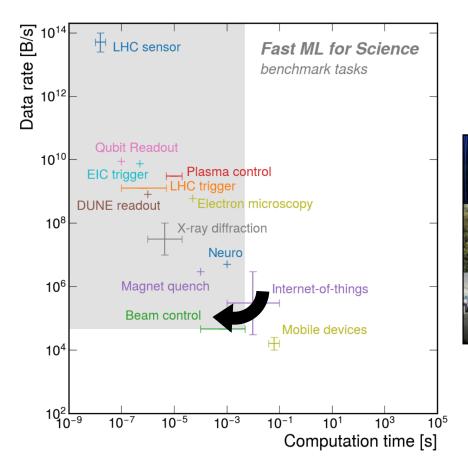


Additional material

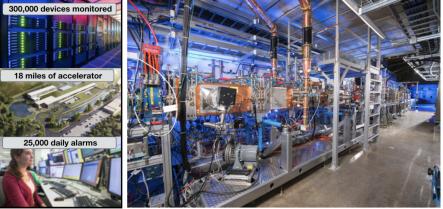






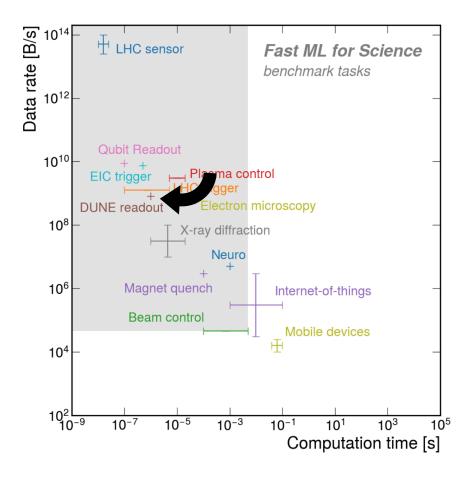


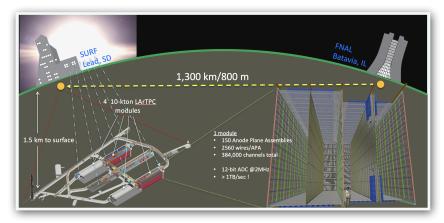
Fermilab accelerator complex



Particle accelerator controls Talk by J. Mitrevski

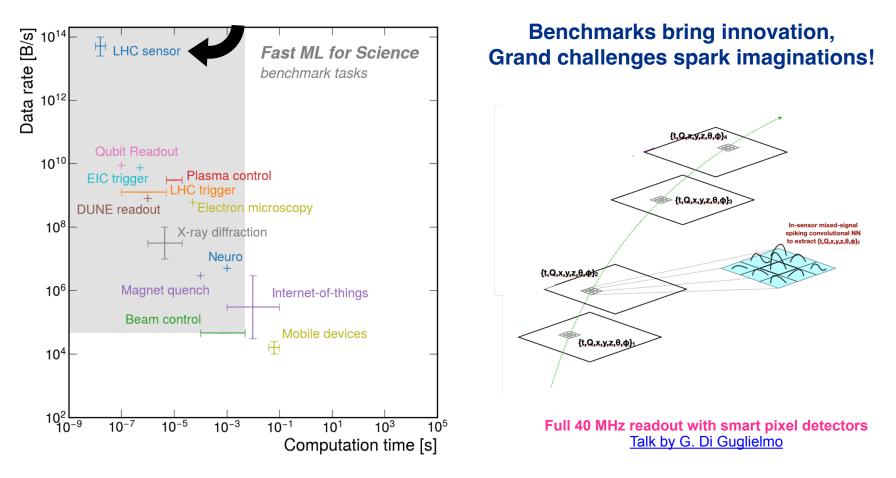




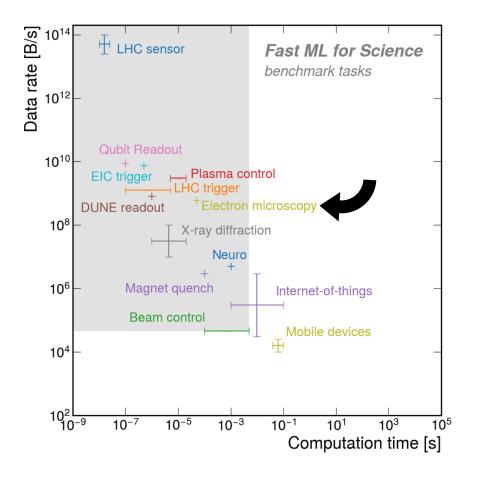


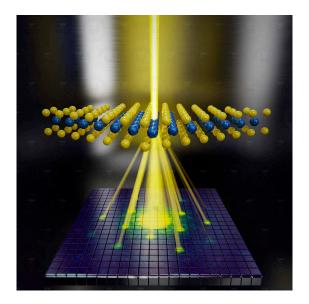
Supernova detection and multi-messenger astronomy Talk by M. Kahn





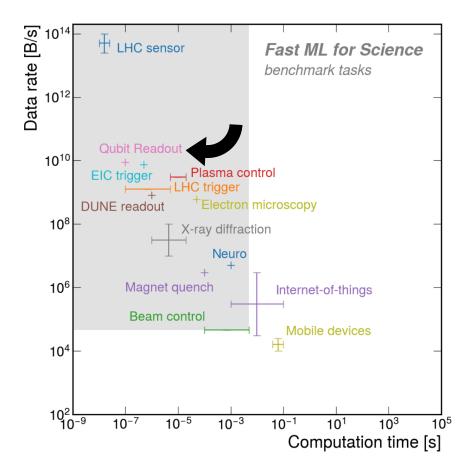


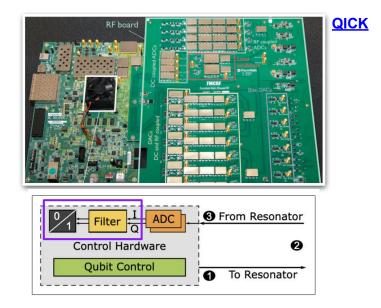




New materials for quantum and energy Talk by J. Agar

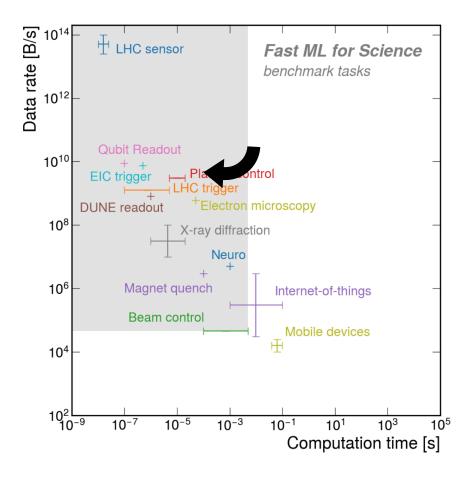


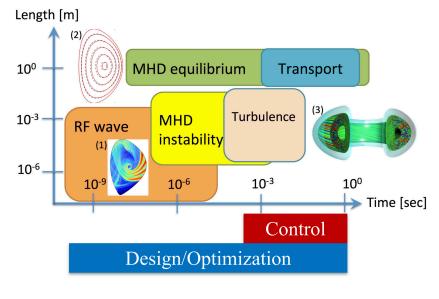




Qubit readout and control Talk by J. Campos

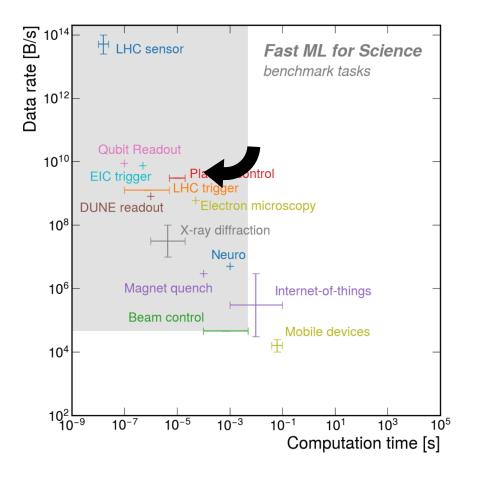


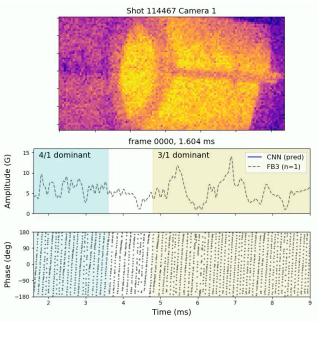




Magnetohydrodynamics Instabilities Talk by R. Forelli

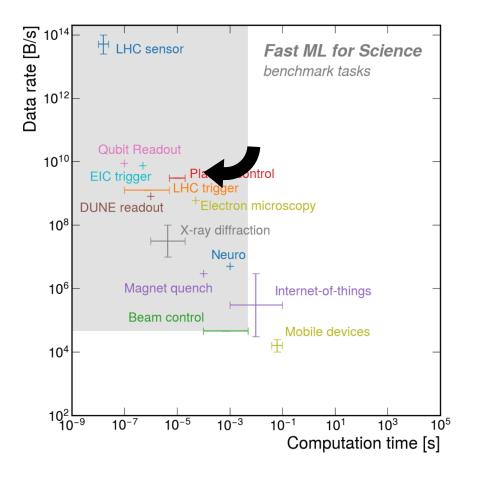


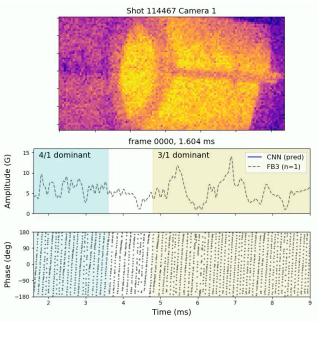




Magnetohydrodynamics Instabilities Talk by R. Forelli

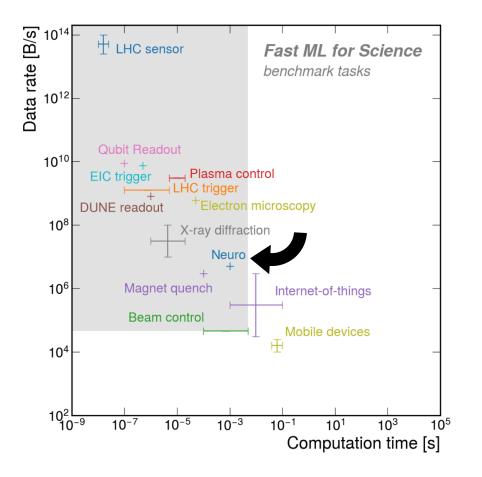


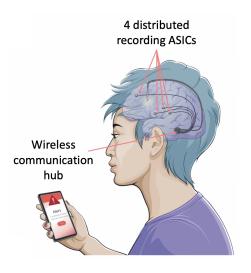




Magnetohydrodynamics Instabilities Talk by R. Forelli







Real-time seizure detection Talk by W. Lemaire

