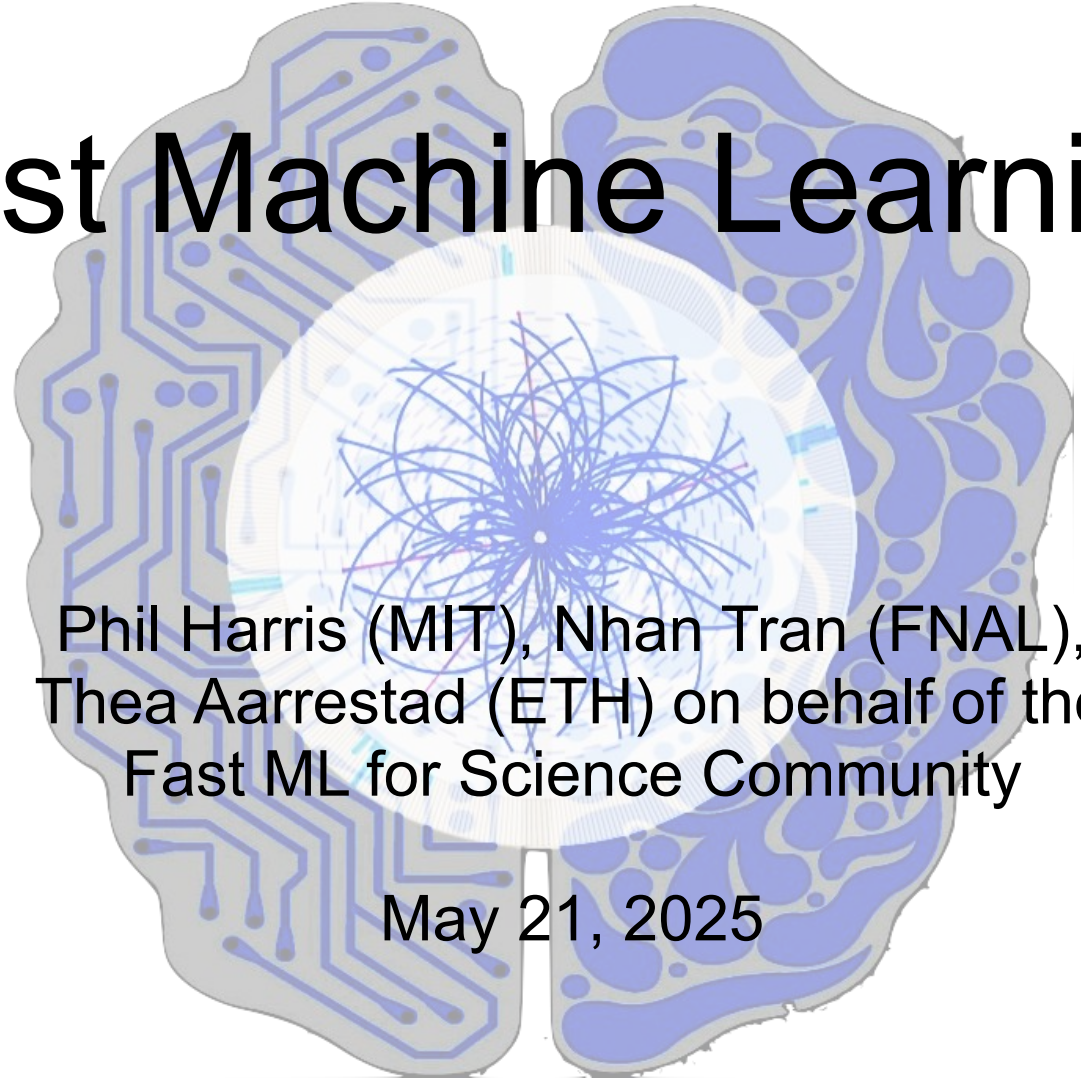
Fast Machine Learning

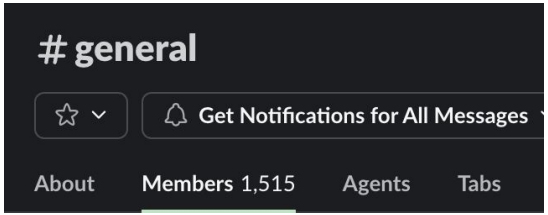


Phil Harris (MIT), Nhan Tran (FNAL),
Thea Aarrestad (ETH) on behalf of the
Fast ML for Science Community

May 21, 2025

What is Fast Machine Learning?

- Origin
  - Fast Machine Learning started not long after the HLS4ML paper
  - We were exploring a way to bring together the community around FastML
    - Clear interests were emerging from others at the LHC
- Fast ML was put together more concretely by end of 2019



https://indico.cern.ch/event/822126/

Mission

Cultivate resources in support of the multi-disciplinary Fast ML for Science community of domain science, machine learning (ML), and engineering researchers for the advancement of accelerated and autonomous scientific experimentation

- Organization and administration of regular meetings, events, and conferences to encourage the cross-pollination of ideas within the community and foster an inclusive environment for promoting new, multi-disciplinary collaborations
- Development and support of open-source software packages, firmware tools, hardware platforms, and benchmarks that increase productivity and accessibility to novel research techniques
- Community engagement and collection of feedback to ensure that activities are aligned with the needs and goals of community members

FastML Conferences and Workshops

J. Duarte

White paper

28,798

TOTAL VIEWS AND DOWNLOADS

97%
Views
rank

https://doi.org/10.3389/fdata.2022.787421

frontiers | Frontiers in Big Data

REVIEW
published: 12 April 2022
doi: 10.3389/fdata.2022.787421

Applications and Techniques for Fast Machine Learning in Science

Allison McCarn Deiana[1], Nhan Tran[2,3*], Joshua Agar[4], Michaela Blott[5], Giuseppe Di Guglielmo[6], Javier Duarte[7], Philip Harris[8], Scott Hauck[9], Mia Liu[10], Mark S. Neubauer[11], Jennifer Ngadiuba[7], Seda Ogrenci-Memik[3], Maurizio Pierini[12], Thea Aarrestad[13], Steffen Bähr[14], Jürgen Becker[15], Anne-Sophie Berthold[14], Richard J. Bonventre[16], Tomás E. Müller Bravo[16], Markus Diefenthaler[17], Zhen Dong[18], Nick Fritzsche[14], Amir Gholami[19], Ekaterina Govorkova[12], Dongning Guo[3], Kyle J. Hazelwood[2], Christian Herwig[2], Babar Khan[13], Sehoon Kim[19], Thomas Klijnsma[2], Yaling Liu[20], Kin Ho Lo[21], Tri Nguyen[3], Gianantonio Pezzullo[22], Seyedramin Rasoulinezhad[23], Ryan A. Rivera[2], Kate Scholberg[24], Justin Selig[25], Sougata Sen[26], Dmitri Strukov[27], William Tang[28], Savannah Thais[29], Kai Lukas Unger[13], Ricardo Vilalta[30], Belina von Krosigk[13,31], Shen Wang[32] and Thomas K. Warburton[33]

OPEN ACCESS

Edited by:
Elena Cuoco,
European Gravitational Observatory,
Italy

Reviewed by:
Andreea Arghiel,
IBM Research-Zurich, Switzerland
Rudy Raymond,
IBM, Japan

***Correspondence:**
Allison McCarn Deiana
adeiana@smu.edu
Nhan Tran
ntran@fnal.gov

Specialty section:
This article was submitted to Big Data and AI in High Energy Physics, a section of the journal Frontiers in Big Data

Received: 30 September 2021
Accepted: 31 January 2022
Published: 12 April 2022

Citation:
Deiana AM, Tran N, Agar J, Blott M, Di Guglielmo G, Duarte J, Harris P, Hauck S, Liu M, Neubauer MS, Ngadiuba J, Ogrenci-Memik S, Pierini M, Aarrestad T, Bähr S, Becker J, Berthold A-S, Bonventre RJ, Müller Bravo TE, Diefenthaler M, Dong Z, Fritzsche N, Gholami A, Govorkova E, Guo D, Hazelwood KJ, Herwig C, Khan B, Kim S, Klijnsma T, Liu Y, Lo KH, Nguyen T, Pezzullo G, Rasoulinezhad S, Rivera RA, Scholberg K, Selig J, Sen S, Strukov D, Tang W, Thais S, Unger KL, Vilalta R, von Krosigk B, Wang S and Warburton TK (2022) Applications and Techniques for Fast Machine Learning in Science. Front. Big Data 5:787421. doi: 10.3389/fdata.2022.787421

In this community review report, we discuss applications and techniques for fast machine learning (ML) in science—the concept of integrating powerful ML methods into the real-time experimental data processing loop to accelerate scientific discovery. The material for the report builds on two workshops held by the Fast ML for Science community and covers three main areas: applications for fast ML across a number of scientific domains; techniques for training and implementing performant and resource-efficient ML algorithms; and computing architectures, platforms, and technologies for deploying these algorithms. We also present overlapping challenges across the multiple scientific domains where common solutions can be found. This community report is intended to give plenty of examples and inspiration for scientific discovery through integrated and accelerated ML solutions. This is followed by a high-level overview and organization of technical advances, including an abundance of pointers to source material, which can enable these breakthroughs.

Keywords: machine learning for science, big data, particle physics, codesign, coprocessors, heterogeneous computing, fast machine learning

> “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

Physics Community Needs, Tools, and Resources for Machine Learning

Snowmass https://arxiv.org/pdf/2203.16255

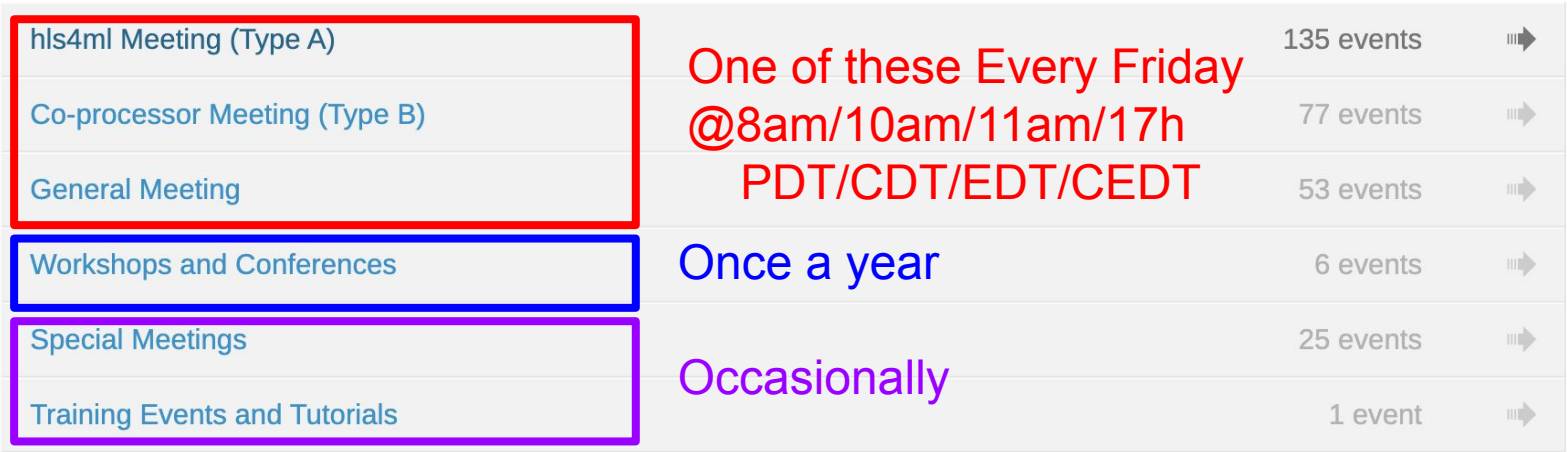
Snowmass https://arxiv.org/pdf/2204.13223

Transcendental Preprint
April 29, 2022

Smart sensors using artificial intelligence for on-detector electronics and ASICs

Communication within FastML at different tiers

- Organize regular meetings, events, and conferences to encourage the cross-pollination of ideas within the community
  - https://indico.cern.ch/category/11842/



What are the goals of fast machine learning

hls4ml

QONNX

Conifer

Promote interdisciplinary collaborations

physicists, computer scientists, electrical and computer engineers, software engineers

SuperSONIC

Custom embedded systems

Off-the-shelf coprocessors

Build open-source, multi-technology codesign workflows

Be nimble: abstraction, portability, containerization

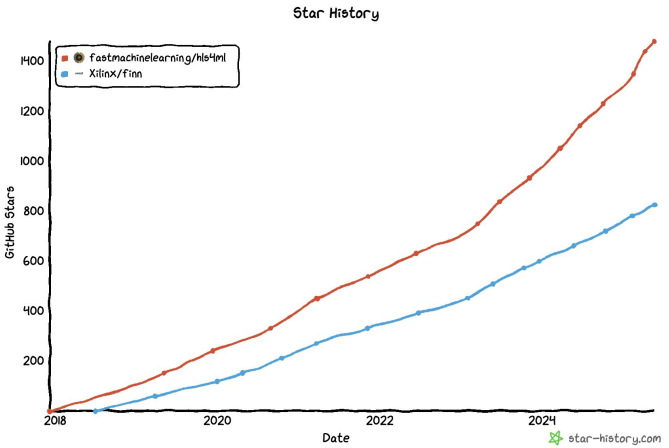
Novel ML research concepts: efficient, fault-tolerant, reliable

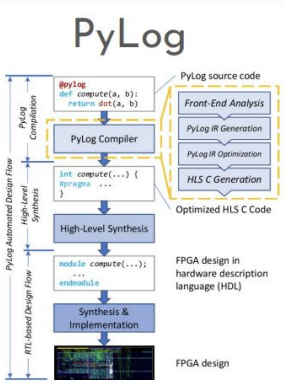
Open data, task-based, and data-based benchmarks

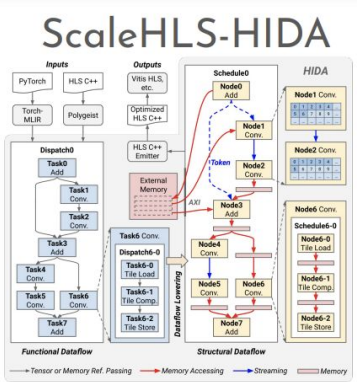
Support ecosystem integration and operation

HLS4ML(Embedded Systems) Subgroup

- Conveners: Benjamin Ramhorst (ETH-Z), Jan-Frederik Schulte (Purdue)
- Two types of meetings:
  - Type I : Survey the community for new ideas related to FPGA/ASIC programming
    - Here we often talk about community tools (not just HLS4ML)
    - Sometimes we hear about new scientific applications
  - Type II : Current status and planning of the HLS4ML software







Coprocessor Subgroup

SuperSONIC

- Conveners: Yongbin Feng (TTU), Yuantang Chou (UW), Ethan Marx (MIT)
- This meeting aims to cover our work with large scale compute clusters
  - How do we run Machine Learning Fast, but focusing on conventional tools and large scale

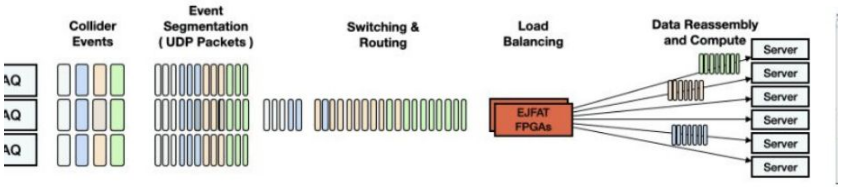
NVIDIA TRITON INFERENCE SERVER

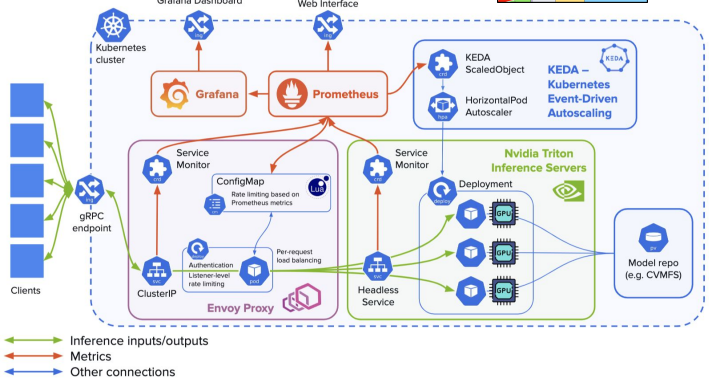
ICECUBE NEUTRINO OBSERVATORY

LIGO VIRGO KAGRA

CMS

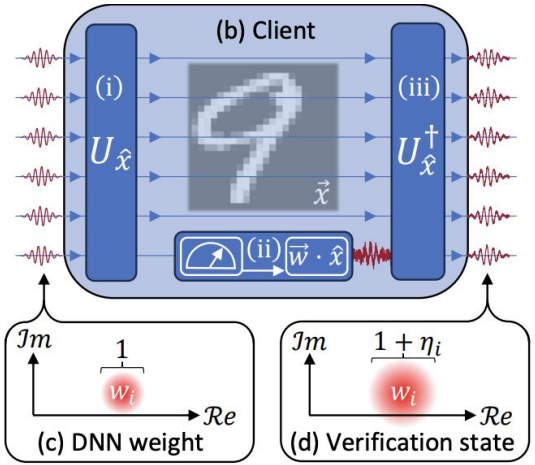
ATLAS EXPERIMENT

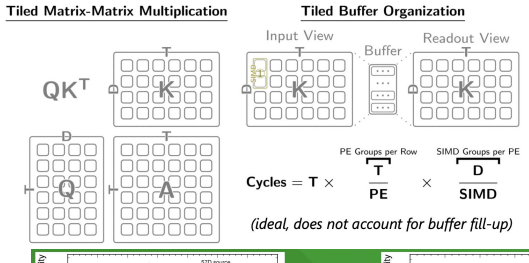


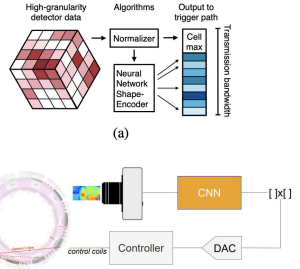


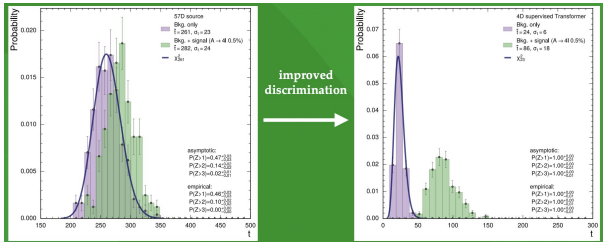
FastML General Meeting

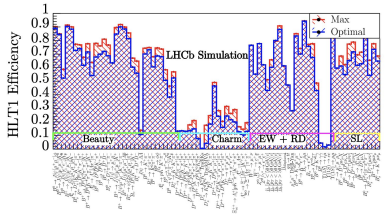
- Conveners: Phil, Nhan Thea
- Goal here is to present FastML(adjacent) studies that can have high impact
  - We want to highlight developments in the field that are relevant
  - Often this includes highlighting new technology and other scientific domains





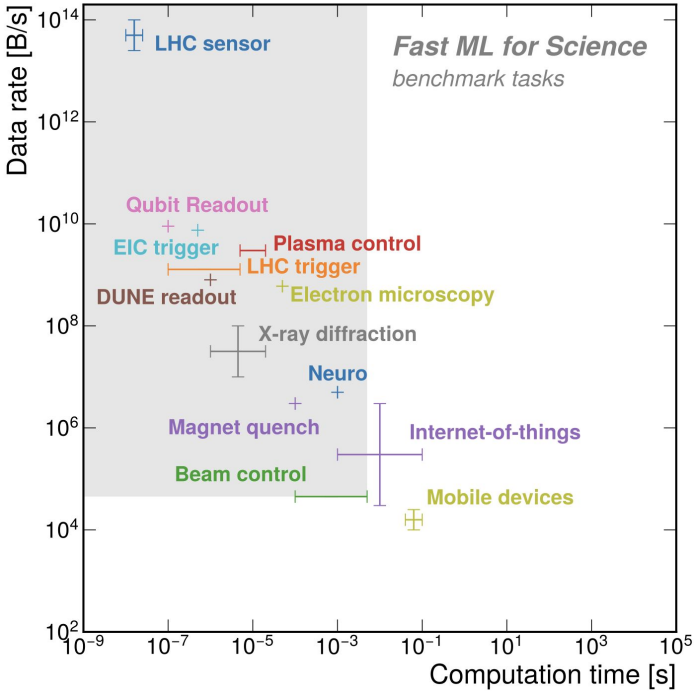




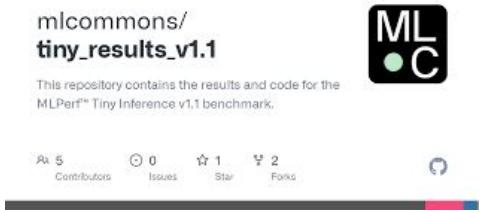


FastML benchmarking

- Large variety of different domains and benchmarks
  - Actively working to connect this with greater ML benchmarking efforts

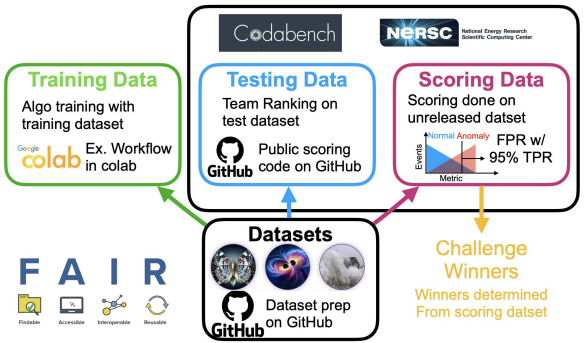


ML Commons



a3d3 imageomics iHARP

NSF HDR ML Challenge



Through A3D3 organized an challenge with 600 teams

FastML benchmarking

- Connecting FastML to the benchmarking community
  - Looking towards a website that can allow us to organize benchmarking
  - These elements are critical towards building robust Science drivers

MLPerf Tiny Benchmark

Colby Banbury* Vijay Janapa Reddi* Peter Torelli† Jeremy Holleman‡‖ Nat Jeffries§

Csaba Kiraly¶ Pietro Montino* David Kanter** Sebastian Ahmed†† Danilo Pau‡‡

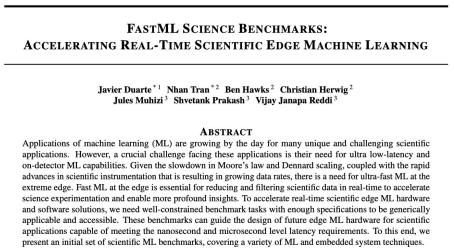
Urmish Thakker[I] Antonio Torrini[II] Peter Warden§ Jay Cordaro ‡ Giuseppe Di Guglielmo[III]

Javier Duarte[IV] Stephen Gibellini‡ Videet Parekh[V] Honson Tran[V] Nhan Tran[VI]

Niu Wenxu[VII] Xu Xuesong[VII]

Abstract

Advancements in ultra-low-power *tiny* machine learning (TinyML) systems promise to unlock an entirely new class of smart applications. However, continued progress is limited by the lack of a widely accepted and easily reproducible benchmark for these systems. To meet this need, we present MLPerf Tiny, the first industry-standard benchmark suite for ultra-low-power tiny machine learning systems. The benchmark suite is the collaborative effort of more than 50 organizations from industry and academia and reflects the needs of the community. MLPerf Tiny measures the accuracy, latency, and energy of machine learning inference to properly evaluate the tradeoffs between systems. Additionally, MLPerf Tiny implements a modular design that enables benchmark submitters to show the benefits of their product, regardless of where it falls on the ML deployment stack, in a fair and reproducible manner. The suite features four benchmarks: keyword spotting, visual wake words, image classification, and anomaly detection.

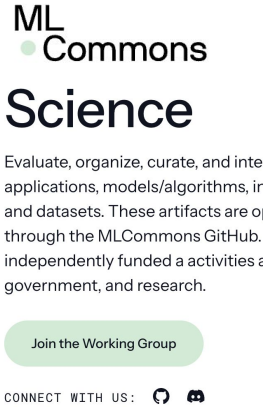


FastML Science Benchmarks: Accelerating Real-Time Scientific Edge Machine Learning

Javier Duarte *[1] Nhan Tran *[2] Ben Hawks [2] Christian Herwig [2]
Jules Muhizi [3] Shvetank Prakash [3] Vijay Janapa Reddi [3]

Abstract

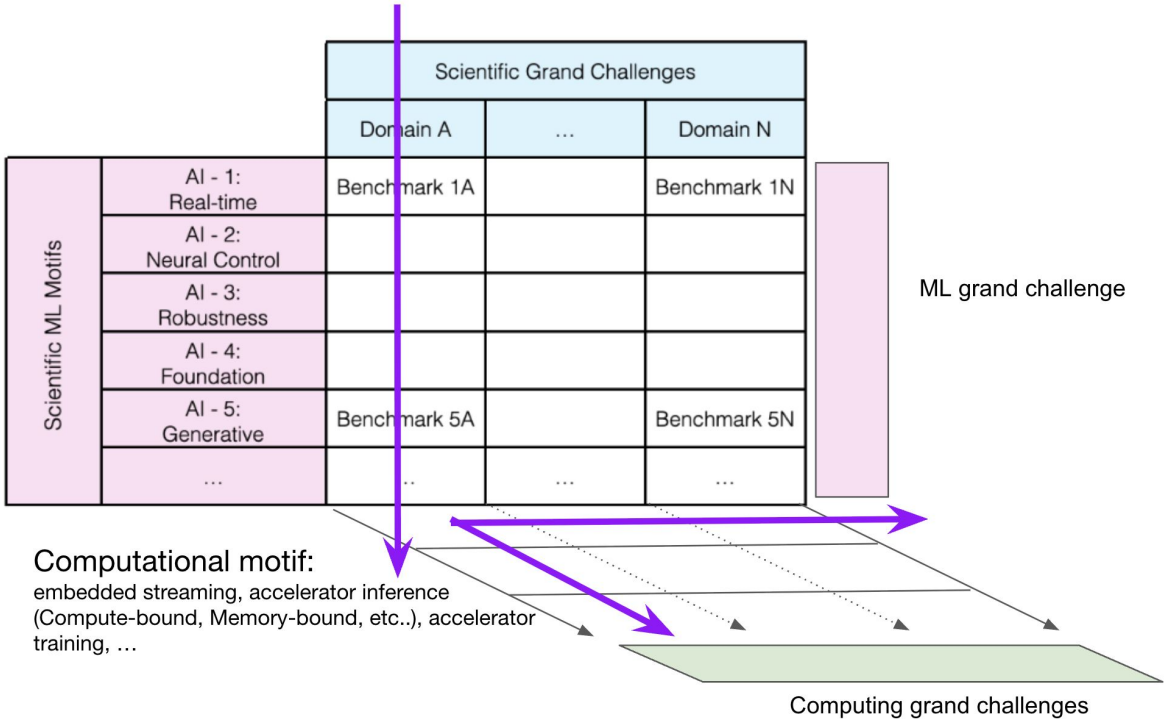
Applications of machine learning (ML) are growing by the day for many unique and challenging scientific applications. However, a crucial challenge facing these applications is their need for ultra low-latency and on-detector ML capabilities. Given the slowdown in Moore's law and Dennard scaling, coupled with the rapid advances in scientific instrumentation that is resulting in growing data rates, there is a need for ultra-fast ML at the extreme edge. Fast ML at the edge is essential for reducing and filtering scientific data in real-time to accelerate science experimentation and enable more profound insights. To accelerate real-time scientific edge ML hardware and software solutions, we need well-constrained benchmark tasks with enough specifications to be generically applicable and accessible. These benchmarks can guide the design of future edge ML hardware for scientific applications capable of meeting the nanosecond and microsecond level latency requirements. To this end, we present an initial set of scientific ML benchmarks, covering a variety of ML and embedded system techniques.



Building Machine Learning Challenges for Anomaly Detection in Science

Editors: Elizabeth G. Campolongo[3], Yuan-Tang Chou[2], Ekaterina Govorkova[1], Wahid Bhimji[10], Wei-Lun Chao[3], Chris Harris[10], Shih-Chieh Hsu[2], Hilmar Lapp[4], Mark S. Neubauer[6], Josephine Namayanja[9], Aneesh Subramanian[5], Philip Harris[1],

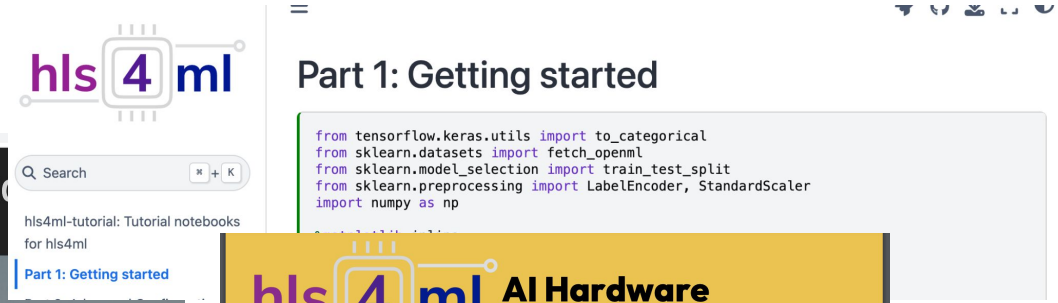
w/MLCommons Science



We are aiming to extend work on ML Challenges, benchmarking and performance to make a Computing axis for future scientific benchmarking

FastML Tutorials and Workshops

- We are regularly organizing tutorials and adjacent workshops
  - Recently hosted a workshop at ICCAD in 2023
  - Looking to organize a NeurIPS Wrokshop this year (fingers crossed)



FastML Engaging the public

symmetry
dimensions of particle physics

CERNCOURIER

MIT Technology Review

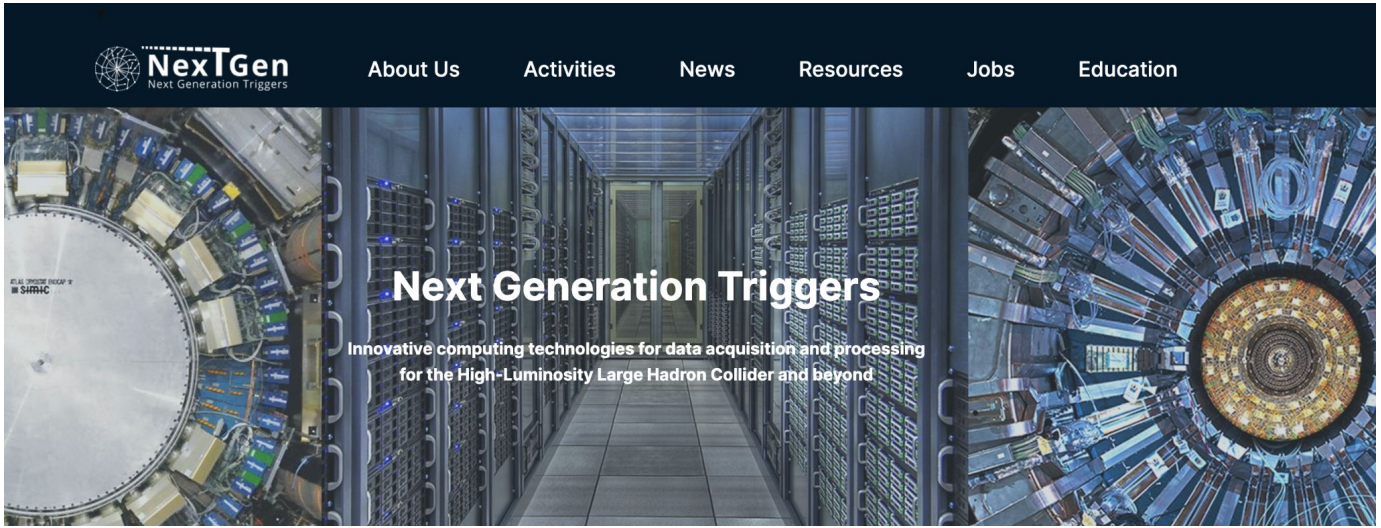
Quantamagazine

FastML Engaging Industry

- Active collaborations with many different members of industry
  - Working with members of AMD, Nvidia, Siemens, Microsoft (preivously), Intel, Graphcore,...
  - Much of our work has direct benefits with industry
- We see industry as a key player in helping scientific goals of FastML

FastML connections with the LHC

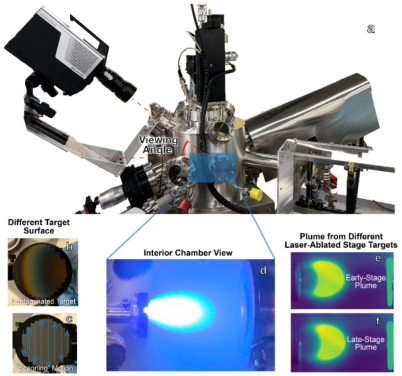
- Next Generation Trigger Project has elevated FastML at the LHC
  - https://nextgentriggers.web.cern.ch/
- Goal: establish working, cross-collaboration environment around triggering
  - Aligned workpackages and integration throughout all tiers of the LHC

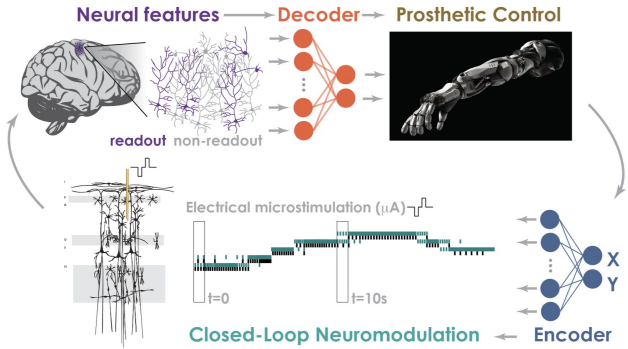


FastML connections with Other Domains

- Many domains are emerging that benefit from FastML
  - Astro, Plasma Physics, Neuroscience, Quantum Computing, Nuclear Physics , Health…
- We are constantly seeing interest to expand to many different domains

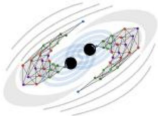
Materials Science





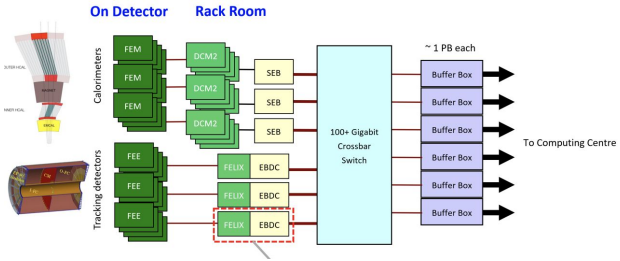
Neuroscience

Astrophysics



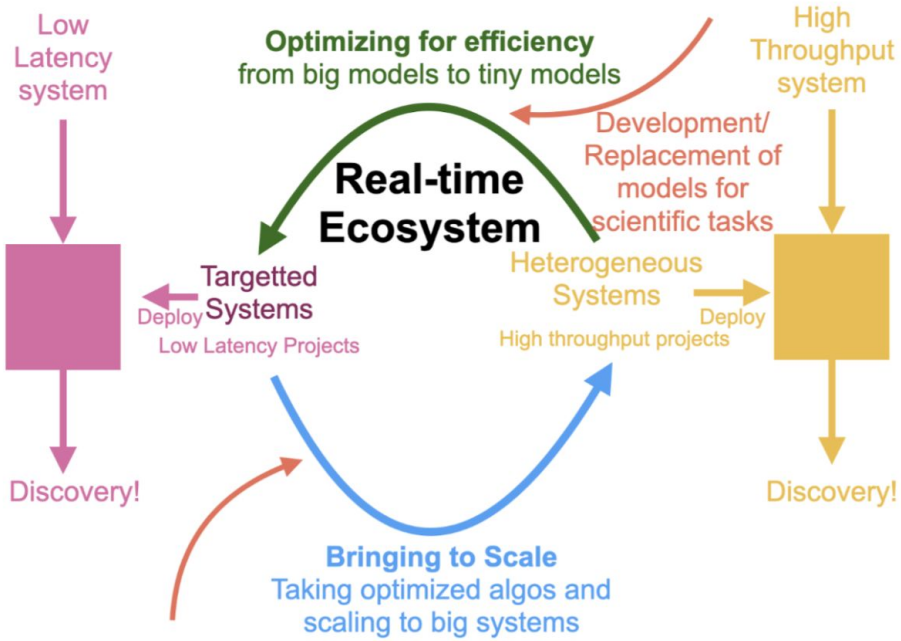
https://github.com/ML4GW

Nuclear Physics



FastML Sustainability Goals

- We are working to highlight the importance of FastML to the US



Cultivating an ecosystem is important

Workshops like this one are essential!

We are happy to see more interested people!

Conclusion and looking forward

- Towards a more sustainable and transparent community
  - We would like to continue growing the community
  - Your input is essential to keeping the community vibrant and dynamic
- Discussions in the last 1.5 years about how to evolve the community
  - At its heart, we are open-source and open science driven
    - We want to continue to support the open source environment
    - Looking for away to sustain FastML for a long period of time
      - Possibility of support of projects and resources
  - A complicated process but we have learned a lot about other open-source models through this exploration –
    - https://fastmachinelearning.org/pose/ – work in progress
- More Generally we are excited to hear from all of you
  - Keep in touch with the community throughout the year
  - Subscribe to the e-group hls-fml@cern.ch, join Slack, and attending meetings

Hope to see you there!