Accelerated ***A***rtificial Intelligence ***A***lgorithms for ***D***ata-***D***riven ***D***iscovery (***A3D3***): Connection to Industry & Impact Beyond HEP

Shih-Chieh Hsu
University of Washington

OAC-2117997

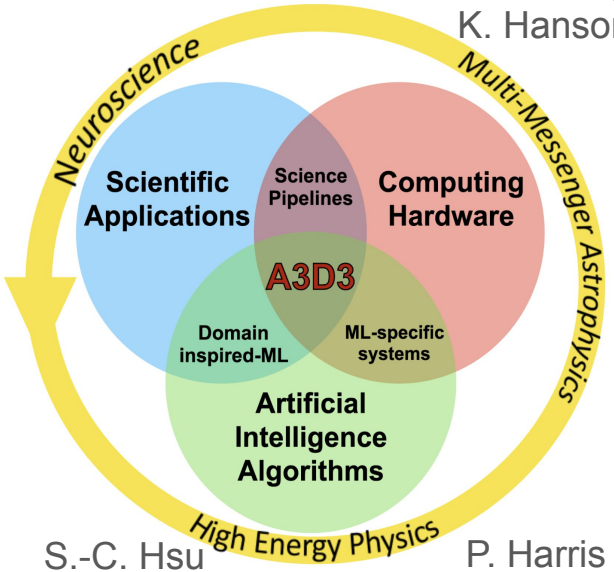
P5 Town Hall, April 12 2023
Brookhaven National Lab

https://a3d3.ai/

NSF HDR Institute **A3D3**

- **A cross-discipline and cross-institutional national institute**
  - Launched in 2021, 10 institutions, 81 members
  - Particle Physics senior personnel: 7/17
- **Our vision** is to establish a tightly coupled organization of **domain scientists**, **computer scientists**, and **engineers** that unite three core components which are essential to achieve **real-time AI** to transform science and engineering discoveries.

K. Scholberg
K. Hanson



S.-C. Hsu
M. Neubauer

P. Harris
J. Duarte, M. Liu

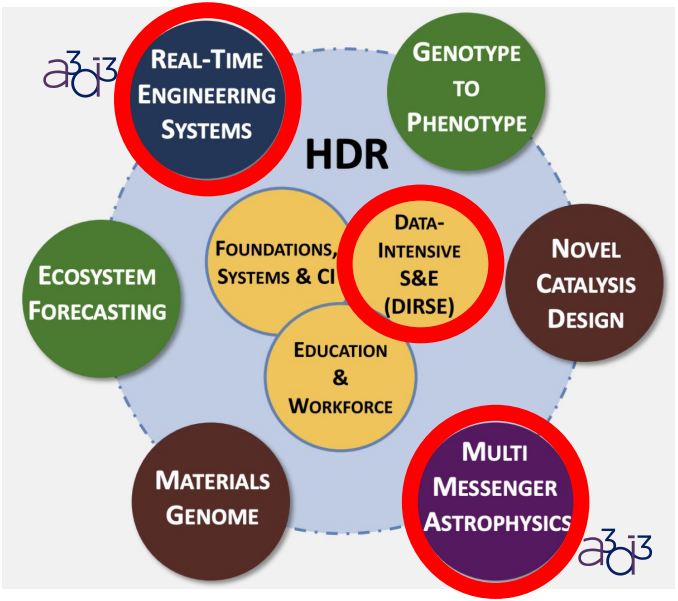
Caltech Duke UNIVERSITY GT MIT P UC San Diego I M W W

Amy Walton 2022

HDR

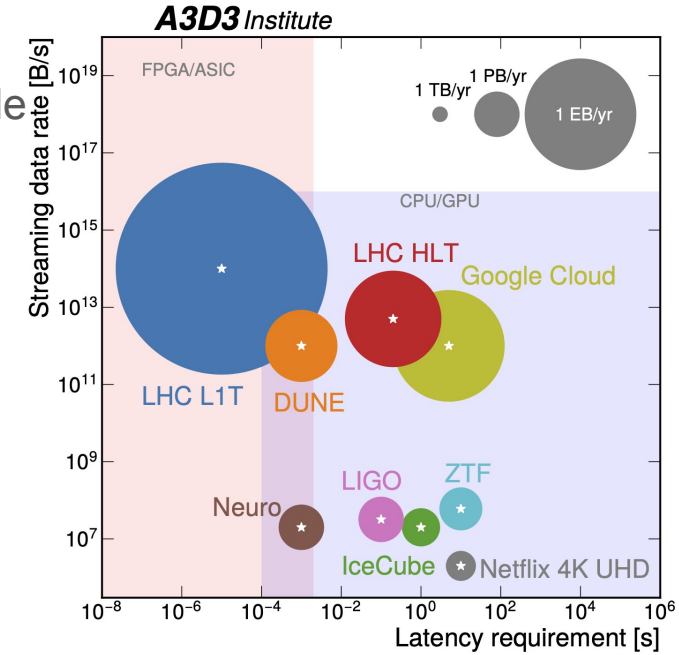
Harnessing the Data Revolution (HDR)

- A national-scale activity to enable new modes of **data-driven discovery** addressing fundamental questions in Sci. & Eng.
- Three parallel tracks:
  - **Institutes (5 awards, $75M)**
    - Ideas Labs+Framework (28, $53M)
  - TRIPODS (28, $42M) & DSC (19, $25M)
- **A3D3** chosen to be the lead institute for **HDR Ecosystem** empowerment.
- A3D3 focusing on data science, while other AI institutes like IAIFI are in different initiatives focusing on intersections between physics and foundation AI



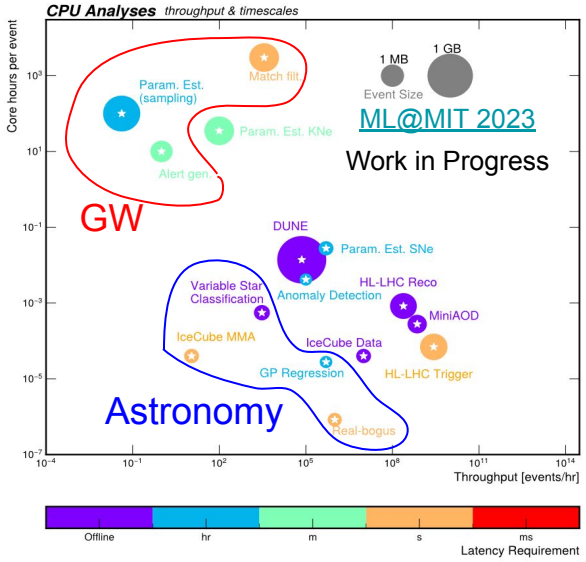
Common challenges across disciplines

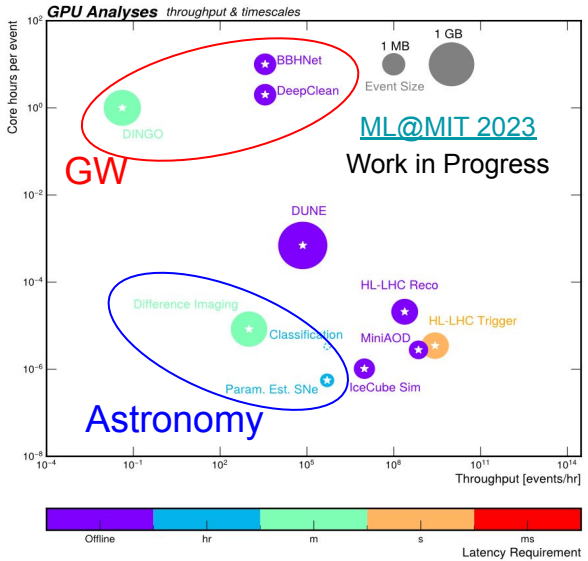
- **Next data revolution around the corner**
  - Both data size and streaming rates of large-scale experiments exceed those handled by industry leaders.
- **New opportunities for applications by accelerating ML/AI algorithms with co-processors:**
  - Classification, regression, parameter estimate, anomaly detection of High Energy Physics and Multi-Messenger Astrophysics
  - Sleeping spindle detection in Neuroscience
  - Opportunity to share next generation computing/hardware across scientific domains



Computational requirement

- GPU accelerating computing by one or more order of magnitudes than CPU
- Potential to explore common computing solution for all domains





How does A3D3 establish strong academic-industry connection?

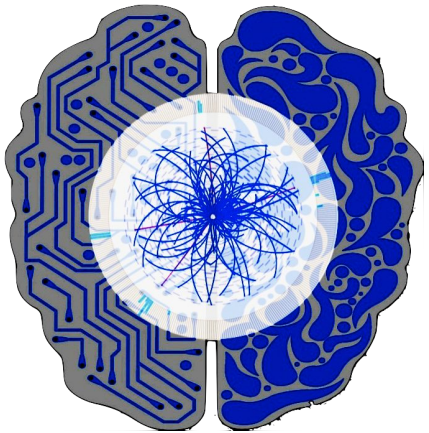
Growing and strengthening existing community

- **Strong connections with industry through Fast ML**
- **A3D3 grew out of the Fast ML Laboratory**
  - FastML: a community-driven research collective of physicists, engineers, and computer scientists interested in deploying machine learning algorithms for unique and challenging scientific applications Slack space
  - 750 members from wide scientific domains, e.g. HEP, Accelerator, Fusion, Material Science
  - A3D3 expanding the scope of the community (e.g. ++neuroscience) and bringing in vision and efforts

FastML Lab

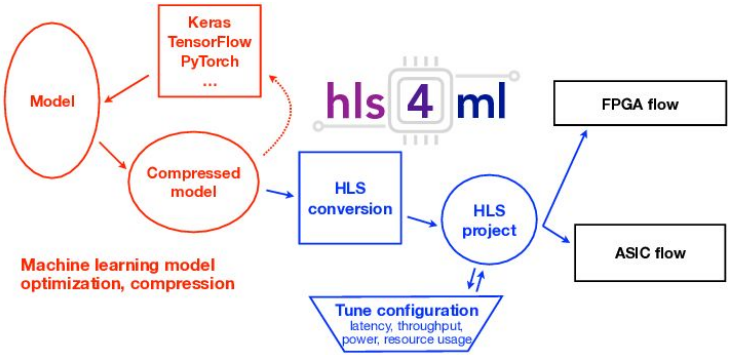
Real-time and accelerated ML for fundamental sciences

https://fastmachinelearning.org/



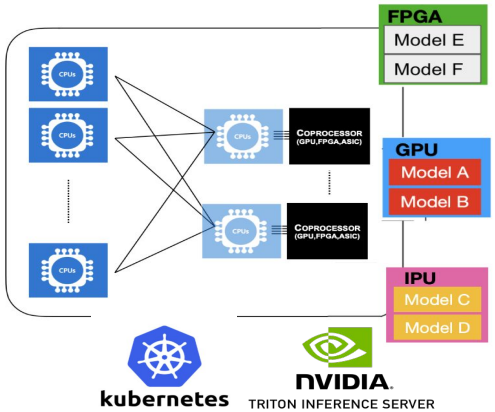
Hardware level: Targeted system for low latency/power

- hls4ml: an open-source package enabling FPGAs & ASICs deployment of ML/AI algorithms (github ★798)
  - A3D3 members are **driving development and applications**, as well as **building a community of users**
  - A3D3 collaborates a lot with hardware research community:
    - e.g. AMD (FINN), TinyML, Imperial College London, University of Toronto, University of Zurich, CERN, FNAL, …, etc.



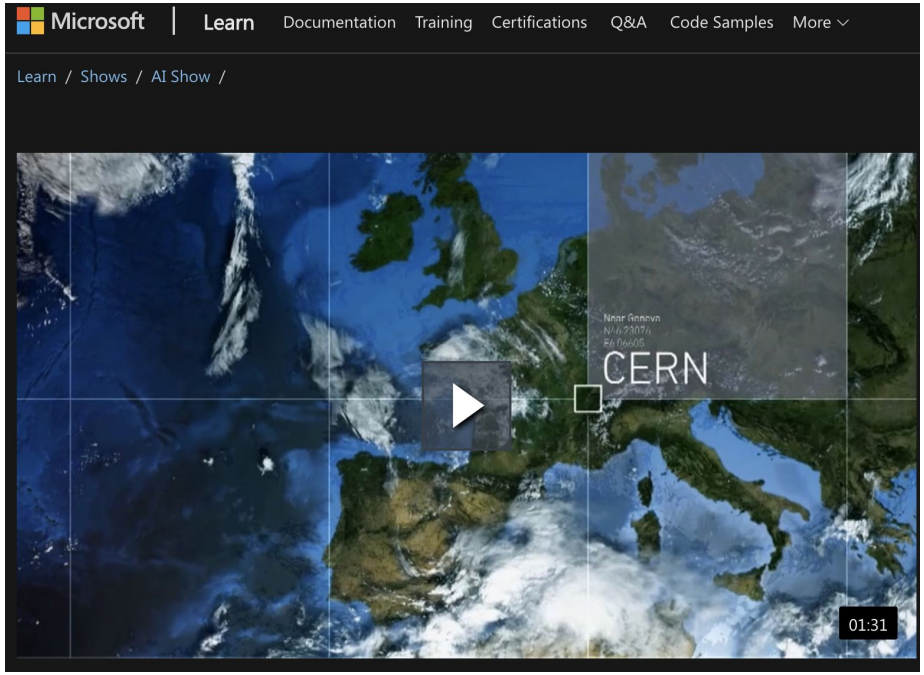
Computing level: Heterogeneous system for high throughput

- **ML as-a-Service** enabling users in sync with the most up-to-date AI model, and the inference server handling job execution in heterogeneous computing system.
  - A3D3 develops workflow platforms (SONIC, hermes) using standard industry tools and collaborates with IT Cloud providers & HPCs to evaluate performance



IT Cloud Providers

High Performance Computing



EPISODE

LHC scientists prototype data analysis solution on Azure Machine Learning

AI Show

May 3, 2019

The Large Hadron Collider at CERN is the largest physics machine ever built, and experiments using the collider generate close to an exabyte (one billion gigabytes) of data in the quest to understand the mysteries of the universe. Machine learning on the global-sized network and speed of FPGAs have the scope to improve data analysis for particle physics. LHC scientists from Fermilab, CERN, MIT, the University of Washington and other institutions worked with Microsoft to prototype a solution to their zettabyte LHC data challenge.

Have feedback? Submit an issue here.

[MicroSoft AI Show]

HPC Wire

HPC wire

Since 1987 - Covering the Fastest Computers in the World and the People Who Run Them

- Home
- Topics
- Sectors
- Exascale
- Specials
- Resource Library
- Podcast
- Events
- Solution Channels
- Job Bank
- About

Internet2 Taps Two Research teams for Final Phase of E-CAS Project

September 3, 2020

WASHINGTON, D.C., Sept. 3, 2

research teams using an extern

final phase of the Exploring Clo

project that was first announced

are:

1. **Investigating Heterogeneou**

Collider, *Philip Harris, MIT.* Onl

per second at the Large Hadron

to the huge volumes of data an

This project proposes a redesig

learning techniques that can be

systems, allowing more data to

and potentially foundational dis

2. **Deciphering the Brain's Ne**

Simulation of Cortical Circuit

SUNY Downstate. This project

Google Cloud

Google Cloud

Blog Solutions & technology Ecosystem Developers & Practitioners Transform with Google Cloud

Transform with Google Cloud

Scaling to infinity and beyond: Using cloud to explore the origins of the universe

October 5, 2022

Matt A.V. Chaban
Industries Editor

The IceCube Neutrino Observatory is one of a growing numbe

initiatives to use scalable cloud infrastructure to process mass

Google Cloud

AMD case study

AMD

CASE STUDY

Artificial Intelligence Accelerates Dark Matter Search

Integrating Inference Acceleration with Sensor Pre-processing in AMD FPGAs Delivers Performance Unachievable by GPUs and CPUs

AT A GLANCE:

Customer: High energy physics researchers from an association of leading international institutions (CMS Institute) conducting experiments at the European particle physics laboratory, CERN.

Industry: Scientific Research

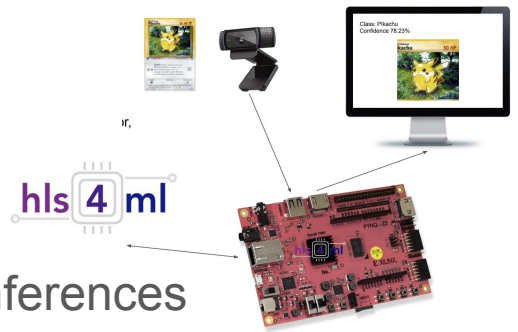
Employees: CMS Institute has more than 4,000 global scientific collaborators representing 200 institutes and universities from more than 40 countries.

https://cms.cern/

What is impact beyond HEP by A3D3?

Workshop, Training and Community planning

- Bringing together developers and stakeholders with an interest in fully integrating ML-based tools from experiments to data analyses and results
  - Fast Machine Learning workshop Oct 2022
  - Accelerating Physics with ML@MIT Jan 2023
- Active participation in community planning events
  - Snowmass Community Planning Exercise
  - IRIS-HEP Blueprint
  - 4NRP
- Tutorials and Demo at non-HEP events e.g. FPGA conferences
- Whitepapers to identify common challenges/solutions cross disciplinaries



Applications and Techniques for Fast Machine Learning in Science
Front. Big Data 5, 787421 (2022)

Physics Community Needs, Tools, and Resources for Machine Learning,
arxiv:2203.16255

Snowmass 2021

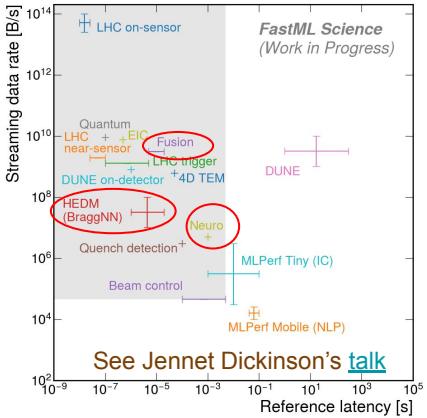
Data science and machine learning education
arxiv:2207.09060

Snowmass 2021

Engagement outside of HEP/Academia

- APS Data Science Education Community of Practice (DSECOP) group
- FAIR4HEP discussions on AI/datasets Planning around challenge datasets
- Empowerment of HDR Ecosystem activities
  - The 1st HDR PI meetings led by A3D3 (Oct 2022)
    - a scale-up version to be hosted by UIUC (2024)
  - HDR Postbaccalaureate workshop (June 2023)
  - HDR Machine Learning challenge within MLCommons in collaboration with FAIR Universe

HEP PIs: 3/72

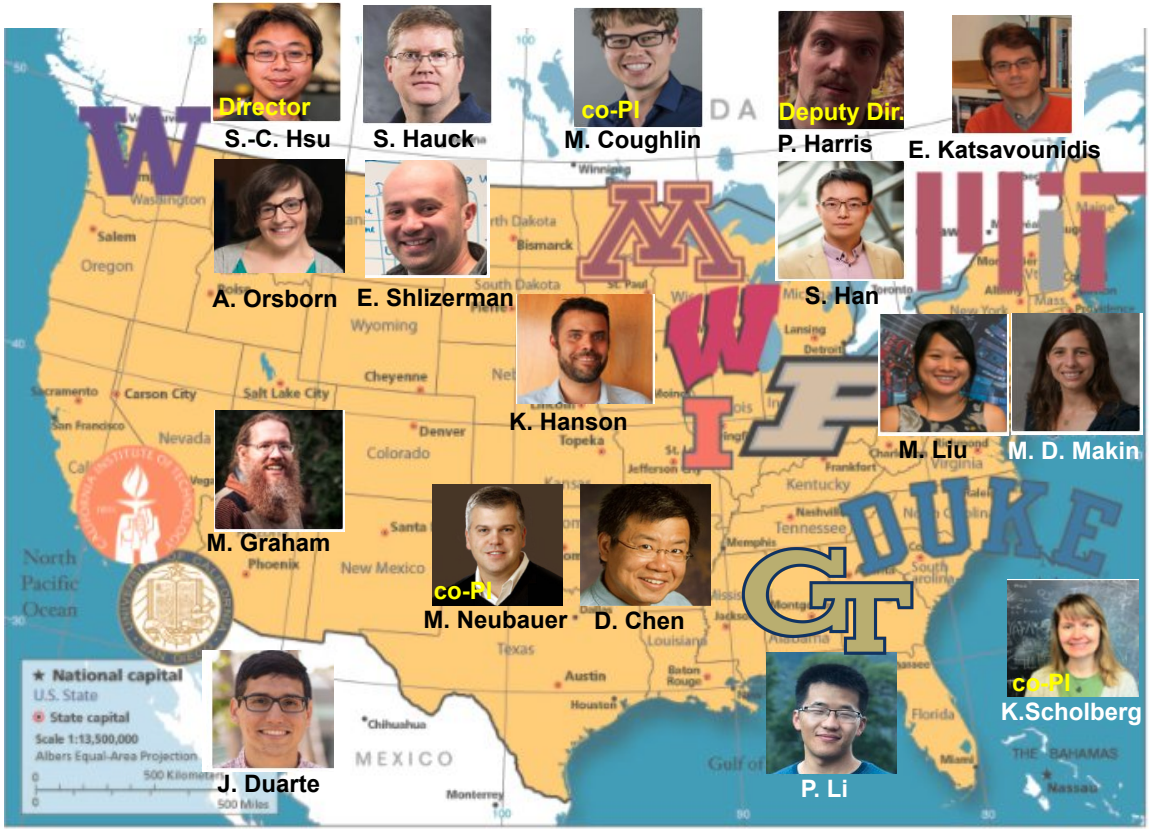


Summary

- **A3D3 utilizing novel AI algorithms and new hardware** to shift paradigm of real time AI processing in order to address upcoming big data revolution.
  - Physics and AI have important complementary roles to play. Foundation AI addressed by IAIFI, software/computing/algorithm investigated by IRIS-HEP
- **A3D3 growing and strengthening existing FastML community** to build strong academia-industry connection
  - At the hardware level, HEP requires the lowest latency AI inference which allows us to lead the industry in this direction
  - At the computing level, HEP requires different/competitive demands with industry which allows us to give feedback for future AI/ML use in physics and industry
- **A3D3 using HEP as the science driver** to empower HDR Ecosystem and bringing broader impact to fields beyond HEP
  - e.g. neuroscience, material science, fusion, etc.

Backup

A Nationwide Institute



trainees.html

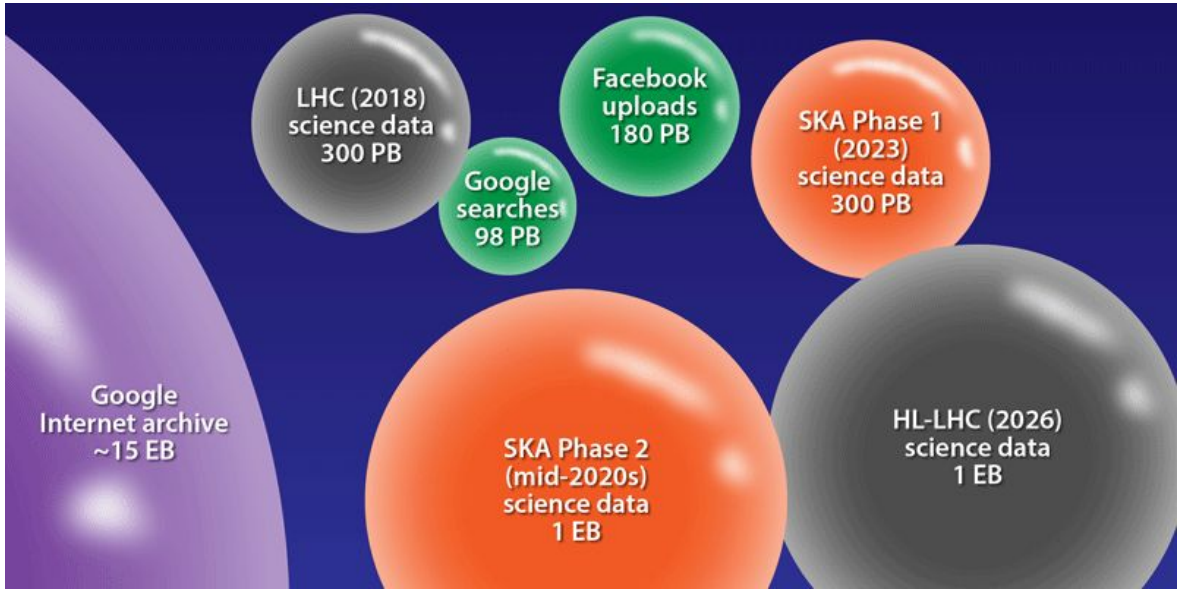
79 Members/10 institutions:

- 17 Senior Personnel
- 3 Research Scientists
- 11 Postdocs
- 27 PhD
- 3 Master
- 12 Undergrad
- 4 Postbacs (Sum ‘22)
- 1 High School

$15M for 5 years since 2021
$1.25M supplement to empower HDR Ecosystem

Trending of big data volume

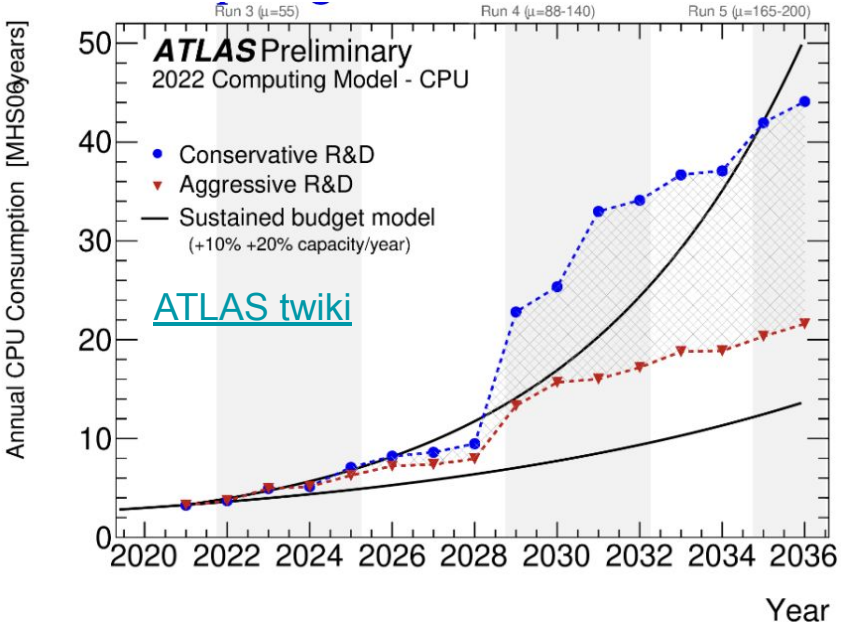
- Next-generation experiments will outpace industry data volumes

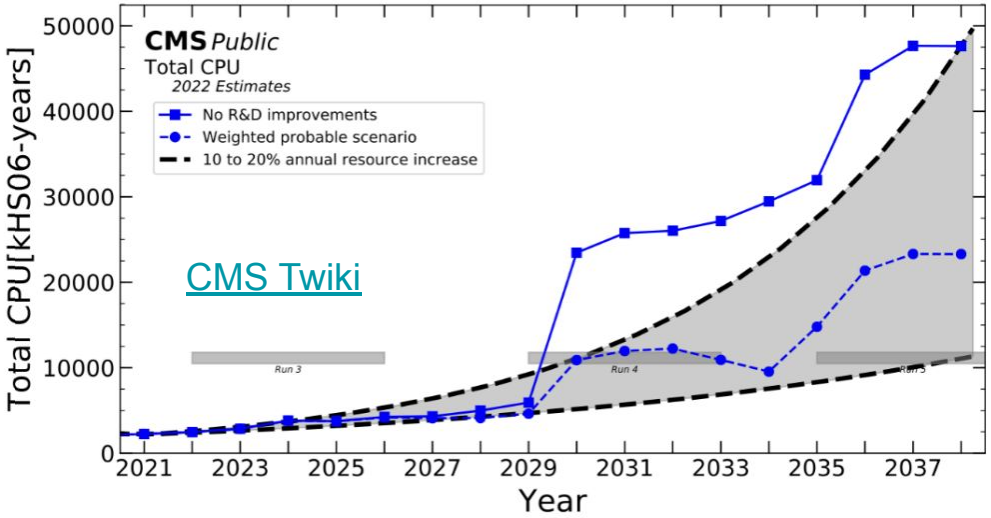


APS/Alan Stonebraker and V. Gülzow/DESY

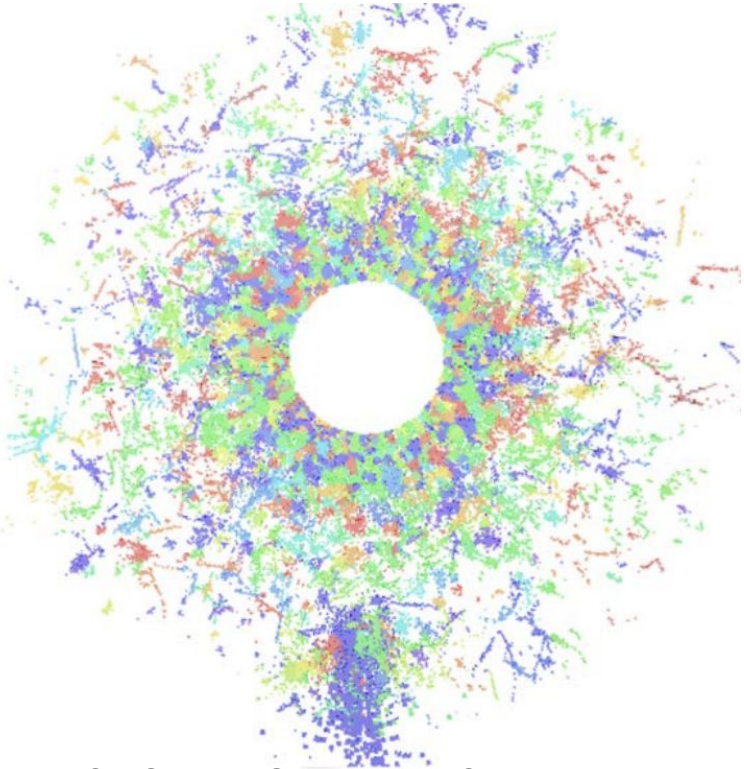
HL-LHC projection

- Substantial continued software R&D improvements starting to fit into optimistic resource regions
- Similar story for disk and tape
- Memory, network projects are uncertain but undeniably finite resources

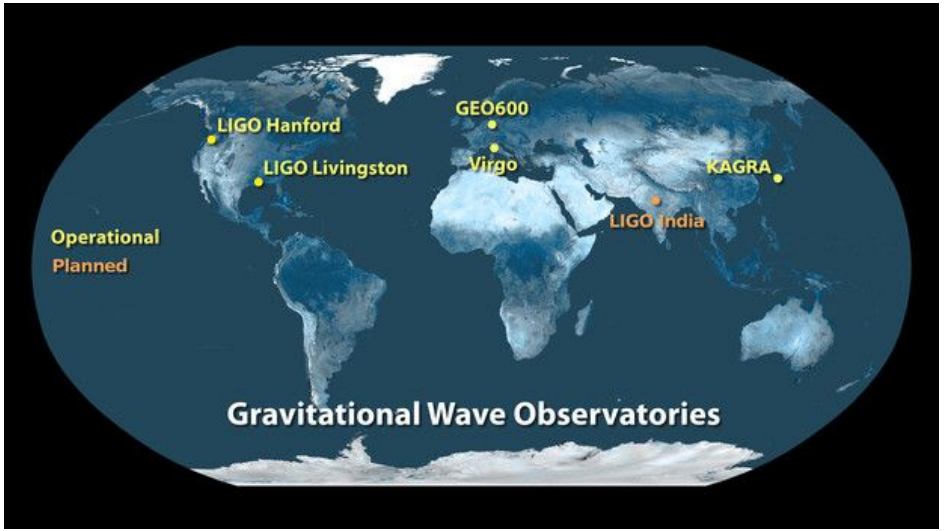




Increasing complexity of data



CMS High Granularity Calorimeter
w/ 200 simultaneous pp collisions

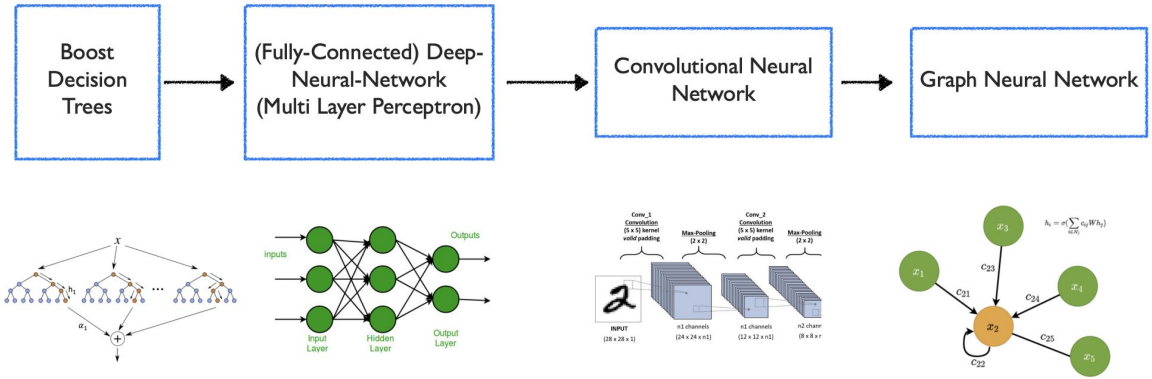


Global LIGO-VIRGO-KAGRA Gravitational Wave
detection and parameter inference analysis

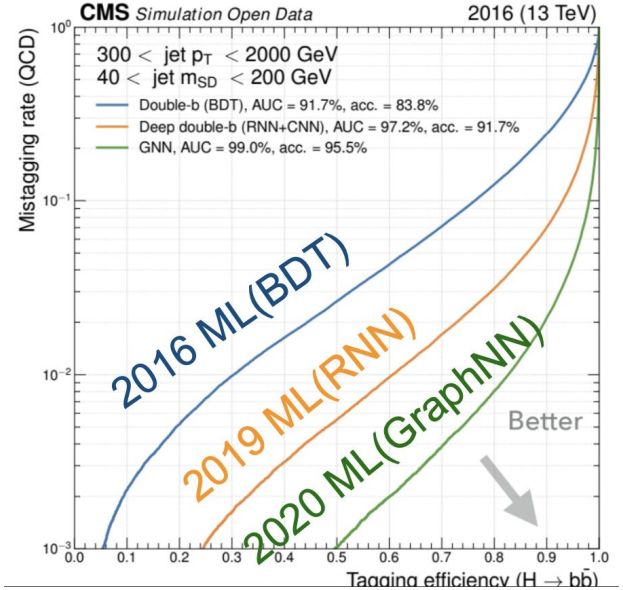
Revolution of AI

AI algorithms have the ability to go beyond algorithms

- Using low level features with deep neural networks and more advanced data structures lead to long latency



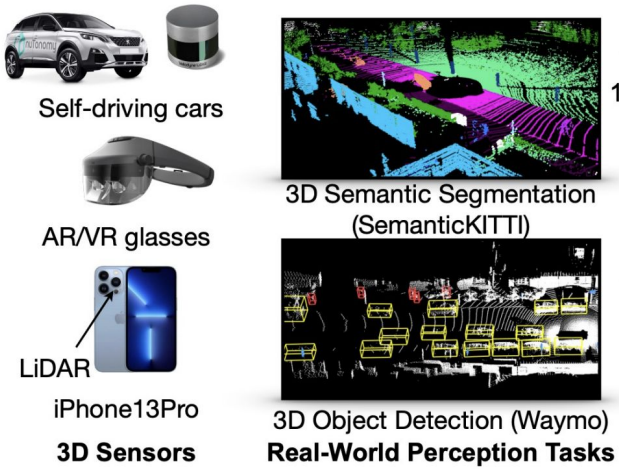
AI algorithms can naturally be accelerated by coprocessors.
The question is HOW!



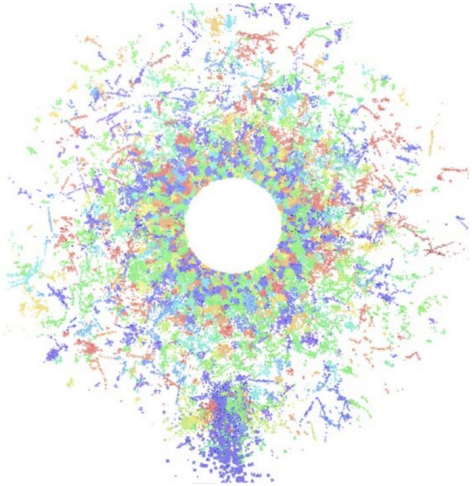
E. Moreno et al. Phys. Rev. D 102, 012010 (2020)

Hardware-Algorirthm co-design

- New algorithms and hardware being prototyped with computational benchmark dataset and applied to domain science.
  - A3D3 researchers proactively seeks synergy cross different data



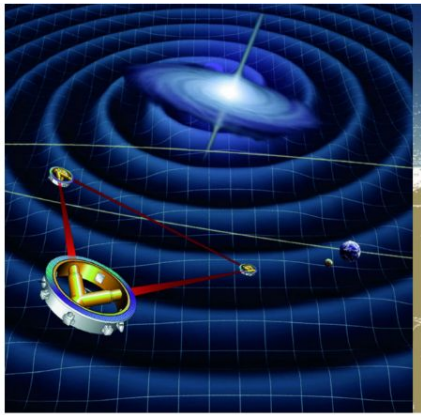
Torchsparse

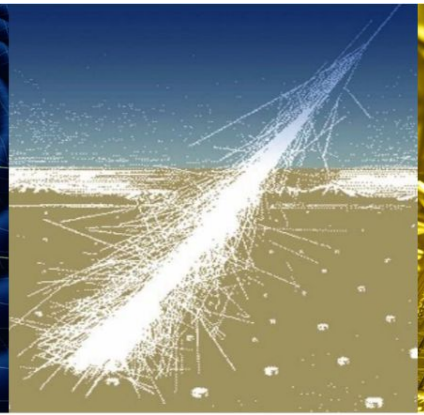


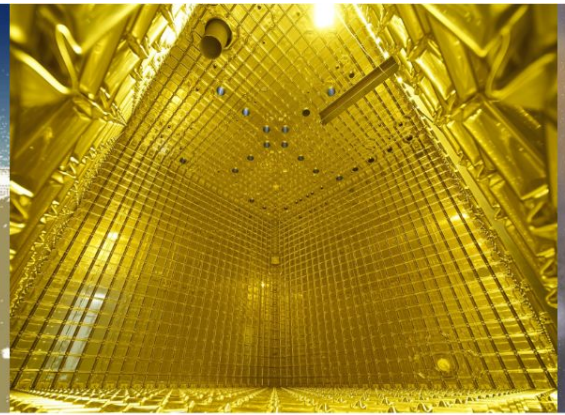
CMS HGCal

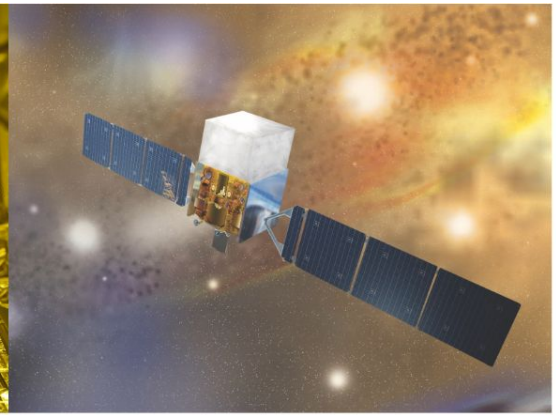
Multi-messenger astrophysics

- Representations of the 4 extrasolar messengers
- Gravitational wave detectors can act as triggers for other types of observatories


gravitational waves,


cosmic rays,


neutrinos (DUNE),


gamma rays (Fermi telescope)

A3D3 for Machine Learning Challenge

- A3D3 receives $1.25M supplement grant. One of the activities is to lead Machine Learning Challenge for the NSF HDR Ecosystem and looking for collaboration with HEP community
- Aim is to make a series of datasets released to public and explore common ML and data approaches
  a. Use these datasets to make a set of ML Challenges
  b. Use for education, training and outreach
  c. Engagement with industry partners to ensure challenges are aligned with real-world applications (training and professional development pipeline)
- We are lacking a clear framework for testing and validation
  a. There are potentially a few options:
    i. Hugging Face
    ii. https://www.modelshare.org/
- We are looking for building strong connections with MLCommons Science, FAIR4HEP and FAIR-Universe.