Fast ML in the NSF HDR Institute A3D3

Shih-Chieh Hsu

University of Washington

FastML Workshop ICCAD

Nov 2 2023

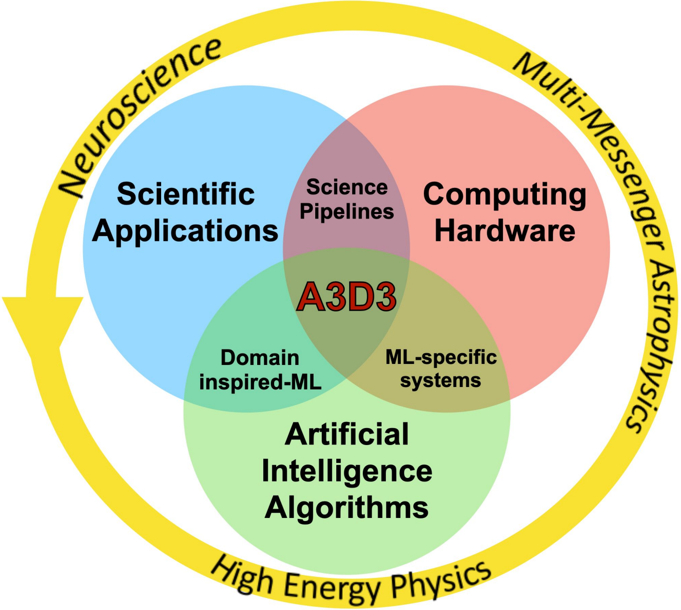
https://fastmachinelearning.org/iccad2023/program.html

OAC-2117997

https://a3d3.ai/

NSF HDR Institute: Accelerated Artificial Intelligence Algorithms for Data-Driven Discovery

(since 2021)

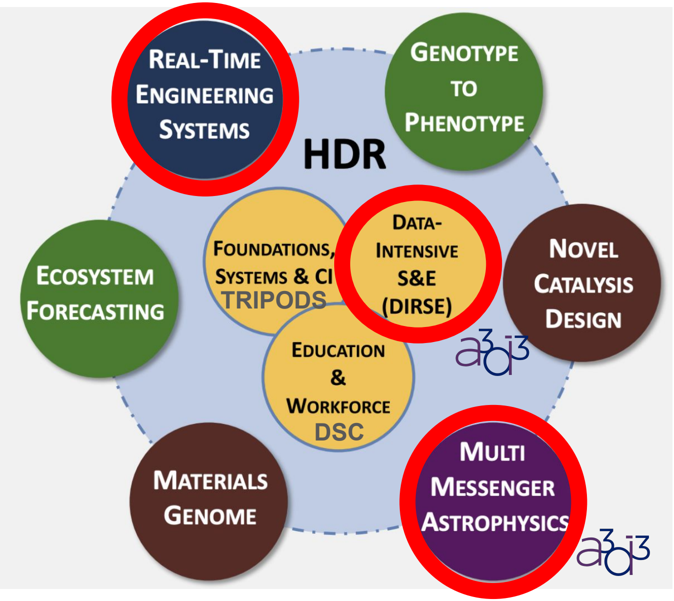


- **Our mission:**
  To enable real-time AI techniques for scientific and engineering discovery by uniting three core components: Scientific Applications, Artificial Intelligence Algorithms, and Computing Hardware

- **Our vision:**
  To make real-time AI accessible to the scientific and engineering community in order to accelerate discovery.

Harnessing the Data Revolution

- A national-scale initiative to enable new modes of **data-driven discovery** addressing fundamental questions in science & engineering
- Three parallel tracks:
  - Institutes (**5** awards, $75M)
    - **A3D3**
    - I-GUIDE
    - iHARP
    - Imageonics
    - ID4
  - Ideas Labs + Frameworks (28, $53M)
  - TRIPODS (28, $42M) & DSC (19, $25M)


REAL-TIME ENGINEERING SYSTEMS
GENOTYPE TO PHENOTYPE
HDR
ECOSYSTEM FORECASTING
FOUNDATIONS, SYSTEMS & CI TRIPODS
DATA-INTENSIVE S&E (DIRSE)
NOVEL CATALYSIS DESIGN
EDUCATION & WORKFORCE DSC
MATERIALS GENOME
MULTI MESSENGER ASTROPHYSICS
a3d3
a3d3

Multi-disciplinary multi-institution

Spread across **16** institutions globally and **106** members (**70%** students + postdocs).

ICCAD FastML organizers associated to A3D3

- Nhan Tran (EAB)
- Mia Liu
- Javier Duarte

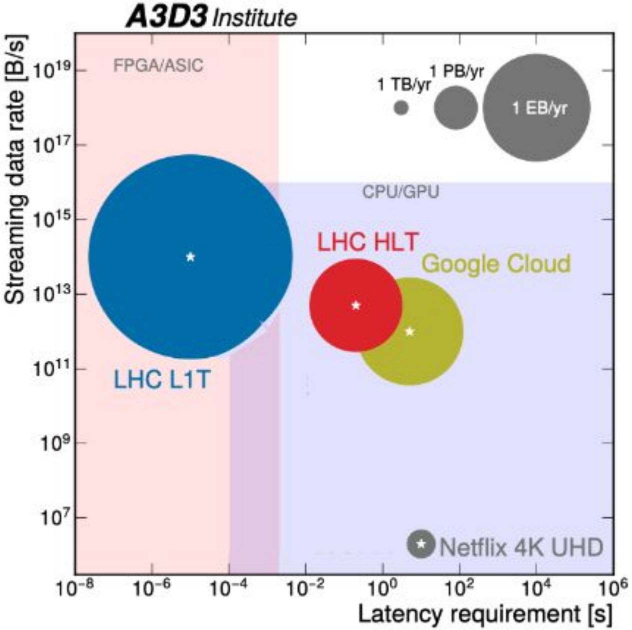
CANADA
Edmonton
Regina
Winnipeg
Toronto
Salem
Oregon
Washington
Montana
Helena
Idaho
Boise
North Dakota
Bismarck
South Dakota
Pierre
Wyoming
Cheyenne
Nebraska
Lincoln
Iowa
Des Moines
Michigan
Lansing
Detroit
Maine
New York
Albany
Mass.
Providence
Conn.
Pennsylvania
Philadelphia
New Jersey
Virginia
Charleston
Frankfort
Kentucky
Nashville
Tennessee
Carson City
Nevada
Salt Lake City
Utah
Denver
Colorado
Topeka
Kansas
Jefferson City
Missouri
Arizona
Phoenix
Santa Fe
New Mexico
Oklahoma
Oklahoma City
Dallas
Texas
Austin
Houston
Baton Rouge
Louisiana
Jackson
Alabama
Georgia
South Carolina
Tallahassee
Florida
Miami
THE BAHAMAS
Nassau
San Francisco
Vegas
North Pacific Ocean
Atlantic Ocean
Gulf of Mexico
MEXICO
Chihuahua
Monterrey
Hermosillo
BERKELEY LAB
WESTMONT COLLEGE
CHRISTUS PRIMATUM TENENS
CALIFORNIA INSTITUTE OF TECHNOLOGY
UNIVERSITY OF CALIFORNIA SAN DIEGO
DUKE
★ National capital
U.S. State
● State capital
Scale 1:13,500,000
Albers Equal-Area Projection
500 Kilometers
500 Miles

ETH Zürich

NYCU NATIONAL YANG MING CHIAO TUNG UNIVERSITY

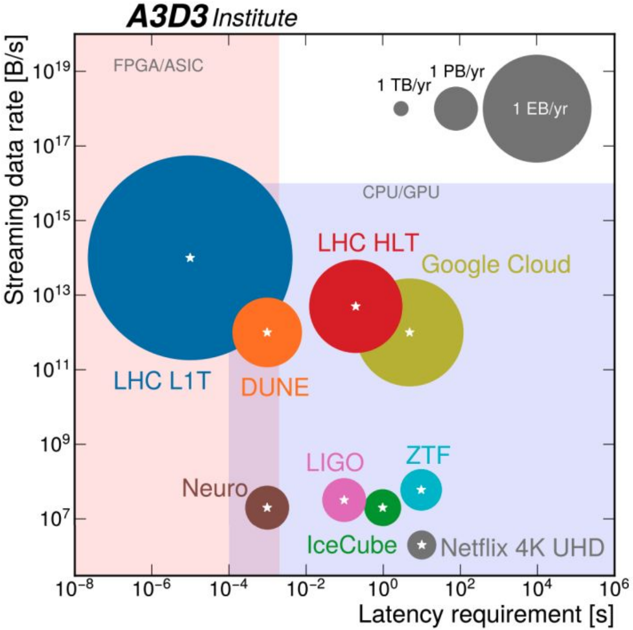
Next generation of big data challenge

- The broader use of **AI/ML** in industry and academia is fueling rapid innovation in hardware accelerators.
- **High Energy Physics** at the LHC driving technology frontier
  - Both data size and streaming rates exceed those handled by industry leaders.

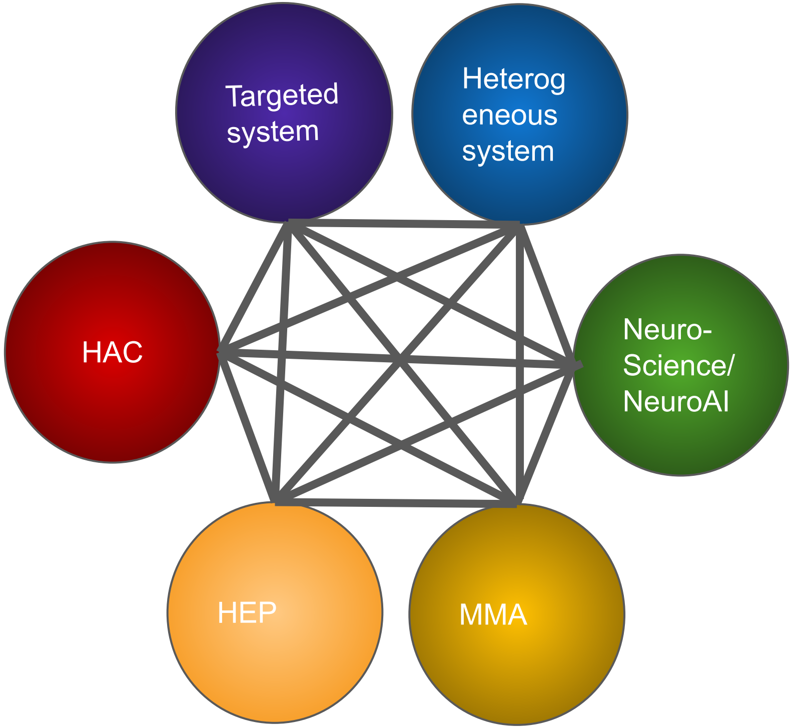


Common challenge cross disciplinary

- **Multi-messenger Astrophysics** facilities rapidly increasing detection rates due to transformative network growth
- **Neuroscience** entering massive data analysis and interpretation thanks to neural recordings at scale



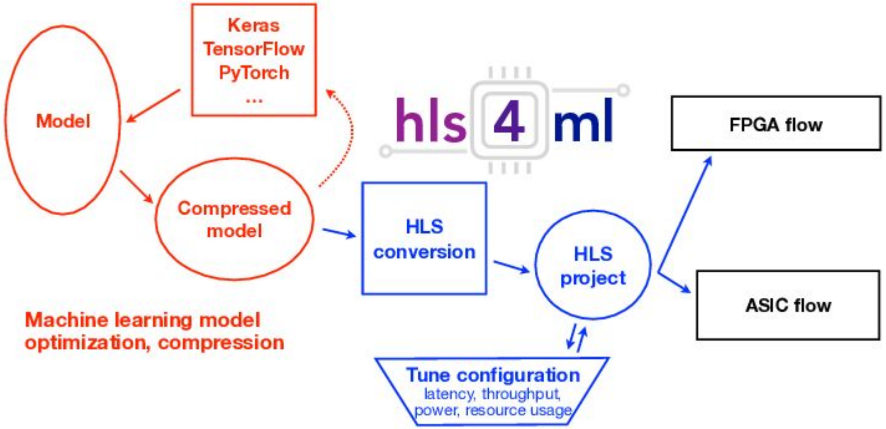
Four focus areas supported by core expertise for sustainability.



Two Integrated systems to facilitate integration and deployment.

Targeted system for low latency/power

- hls4ml: an open-source package enabling FPGAs & ASICs deployment of ML/AI algorithms
- A3D3 members are **core contributors and maintainers of package**, as well as **building a community of users**
  - AMD (FINN), TinyML, Imperial College London, University of Toronto, University of Zurich, CERN, FNAL, …, etc.



Model

Keras
TensorFlow
PyTorch
…

Compressed model

Machine learning model optimization, compression

hls 4 ml

HLS conversion

HLS project

Tune configuration
latency, throughput, power, resource usage

FPGA flow

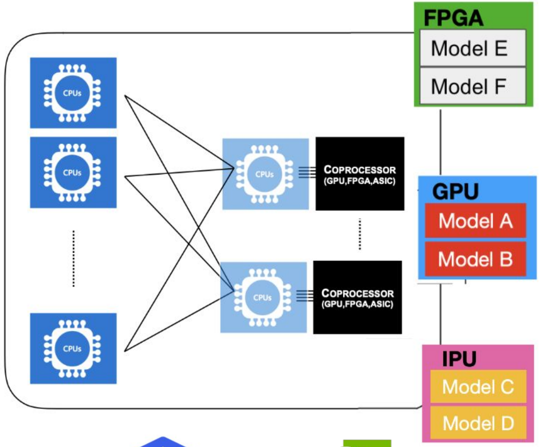
ASIC flow

AMD XILINX

intel

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



kubernetes

NVIDIA TRITON INFERENCE SERVER

Microsoft Azure

aws

Google Cloud Platform

GRAPHCORE

IT Cloud Providers

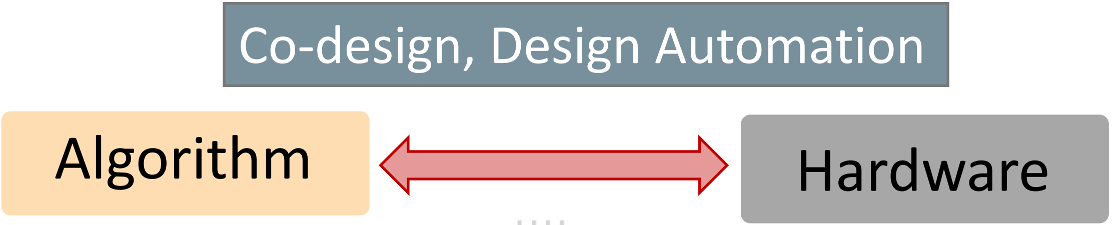
NRP SDSC

HOLL-I

I NCSA

High Performance Computing

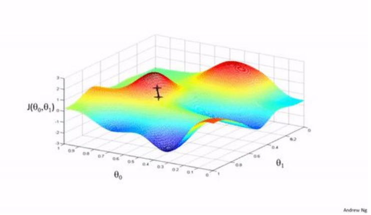
Hardware-Algorithm Co-design (HAC)



Co-design, Design Automation

Algorithm

Hardware



Challenges in Algorithm Design:

- Irregular data (graphs, point clouds)
- Label scarsity
- AI models are hard to be interpreted
- …

Challenges in Deployment in Hardware:

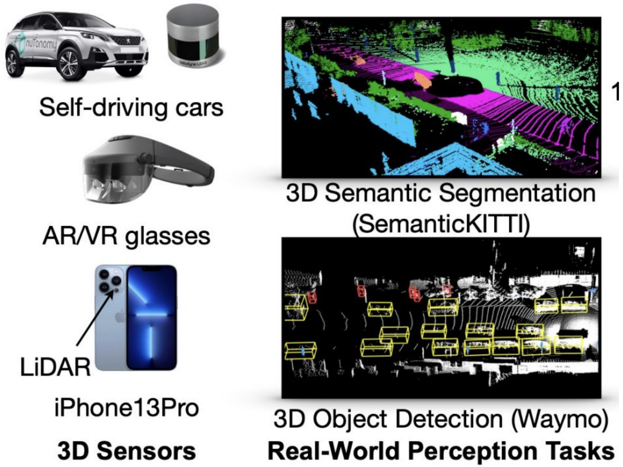
- Computation efficiency issues (e.g. see Caroline Johnson’s talk)
- Power/memory constraints
- Hard to be implemented on FPGA/ASIC

…

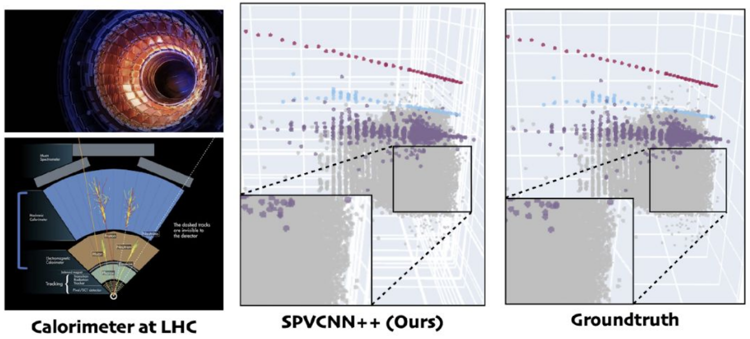
--> hardware design automation tools

HAC: Innovative application

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


Self-driving cars
AR/VR glasses
LiDAR
iPhone13Pro
3D Sensors
1
3D Semantic Segmentation (SemanticKITTI)
3D Object Detection (Waymo)
Real-World Perception Tasks

Torchsparse



Credit: Z. Liu

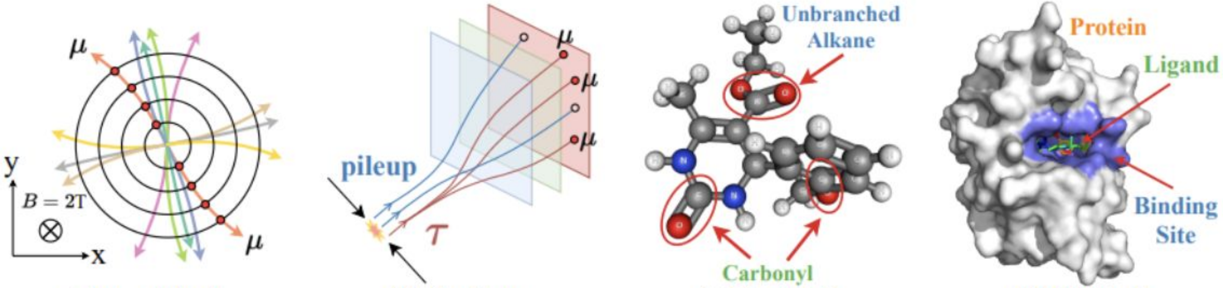
Torchsparse/ Torchsparse++ (Haotian Tang, et al. @ MLSys’22)

SPVCNN++ (Zhijian Liu et al . + HEP team)

HAC: ML Algorithms development

- GSAT & LRI (Siqi Miao, et al., @ ICML’22, ICLR’23)

How to build interpretable and generalizable graph/geometric learning models?



(a) ActsTrack (b) Tau3mu (c) SynMol (d) PLBind

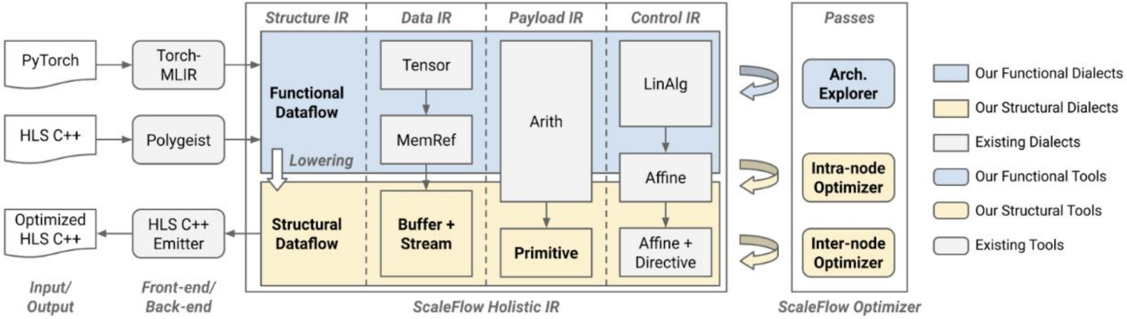
A good model should capture the truly effective data patterns

- Theoretically grounded by the principle of information bottleneck
- Outperform baselines with a 10% improvement in detection accuracy of effective patterns and a 3% improvement in out-of-distribution generalization prediction accuracy

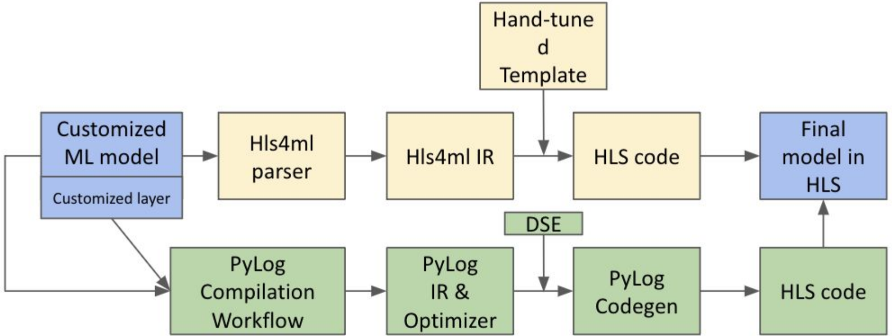
Credit: Pan Li

Design Automation

- ScaleHLS / ScaleHLS 2.0 (Hanchen Ye, et al.)
  - generate highly-efficient hardware accelerators for scientific algorithms without much design effort



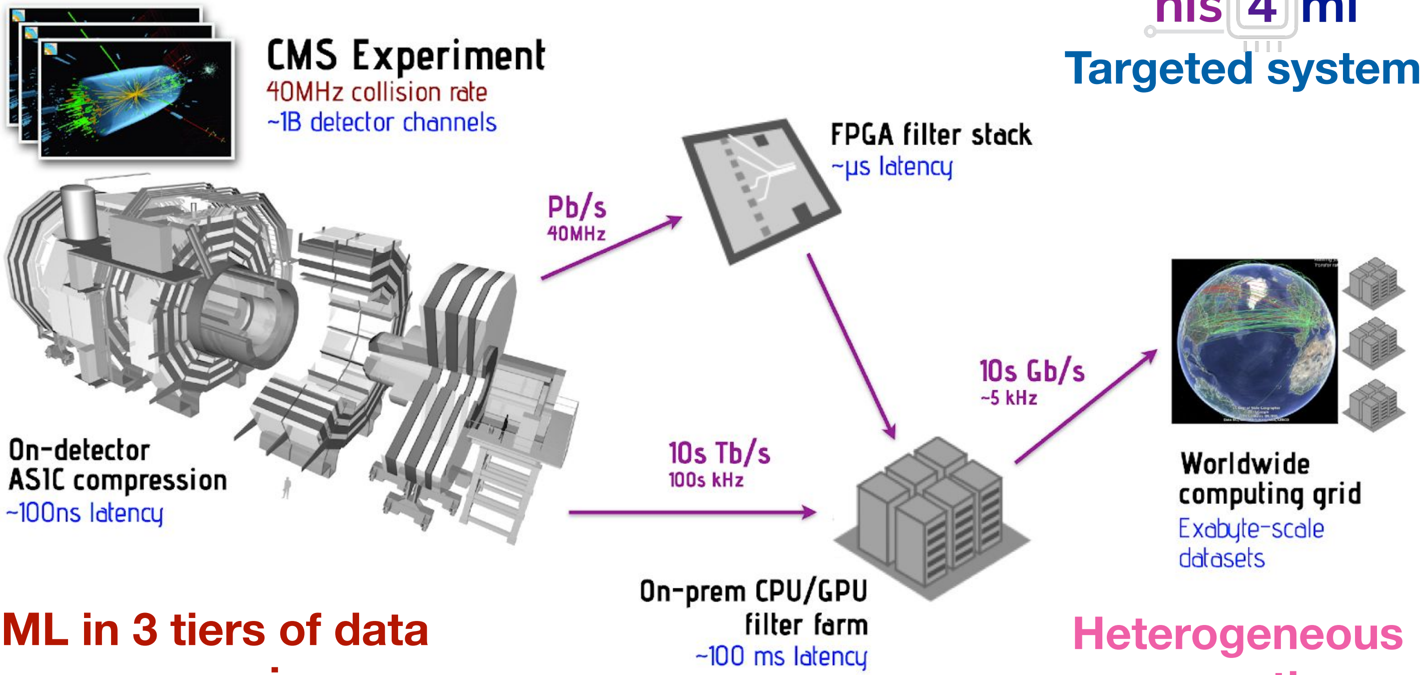
- PyLog + HLS4ML (Tim Zhang, et al.)
  - Integration of PyLog and HLS4ML enables significant code reduction in FPGA-oriented ML model development



High Energy Physics (HEP)

hls4ml

Targeted systems



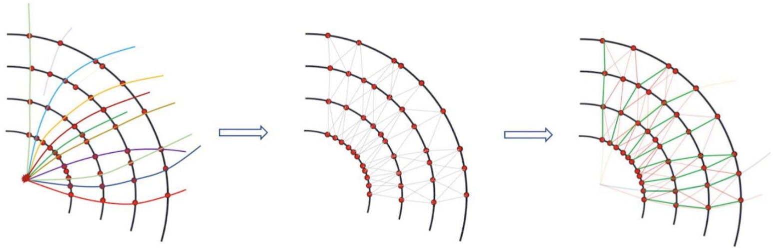
ML in 3 tiers of data processing

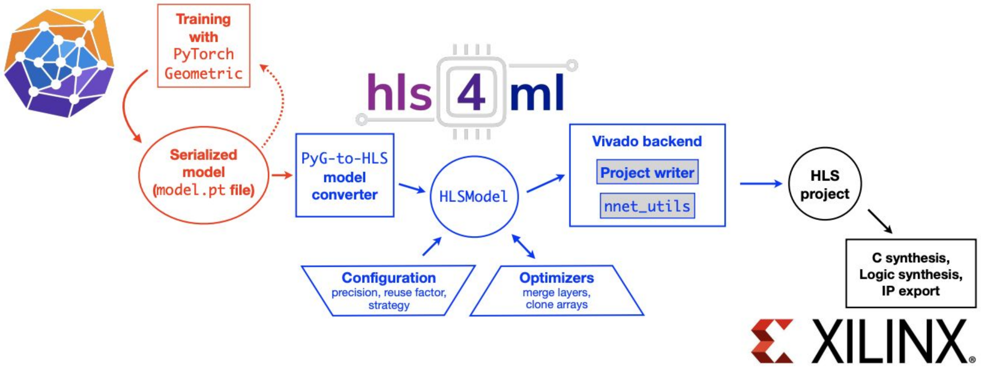
Heterogeneous computing

Credit M. Liu

Graph Neural Network for tracking

- Algorithms making tracking highly parallelizable both low latency FPGA version and GPU version
    - [Front. Big Data 5 (2022) 828666]
    - [2306.11330]
- Can be used at various tiers of track reconstruction
    - ExaTrkX as a service [CTD2023]





Training with PyTorch Geometric

Serialized model (model.pt file)

hls 4 ml

PyG-to-HLS model converter

HLSModel

Vivado backend

Project writer

nnet_utils

HLS project

Configuration
precision, reuse factor, strategy

Optimizers
merge layers, clone arrays

C synthesis, Logic synthesis, IP export

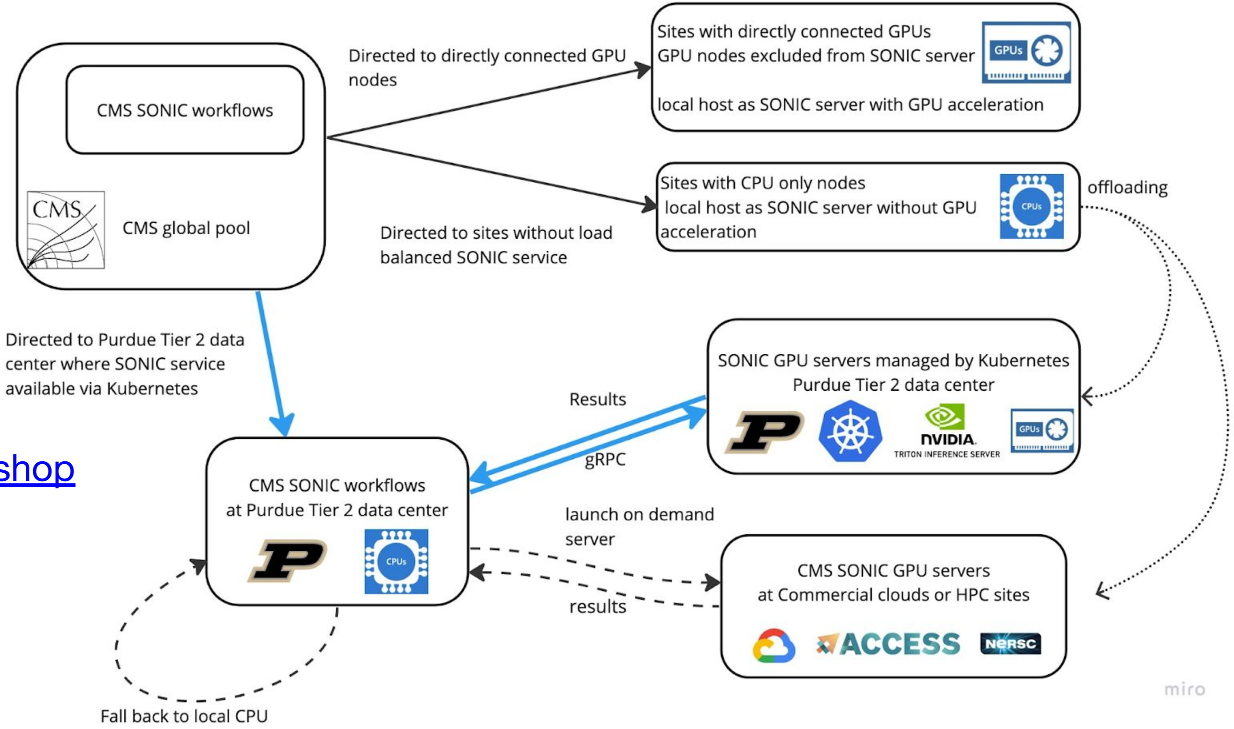
XILINX

Heterogeneous computing as-a-service (SONIC)

Significant progress in integration of SONIC in CMS for minAOD production

[Talk at fast ml workshop]()

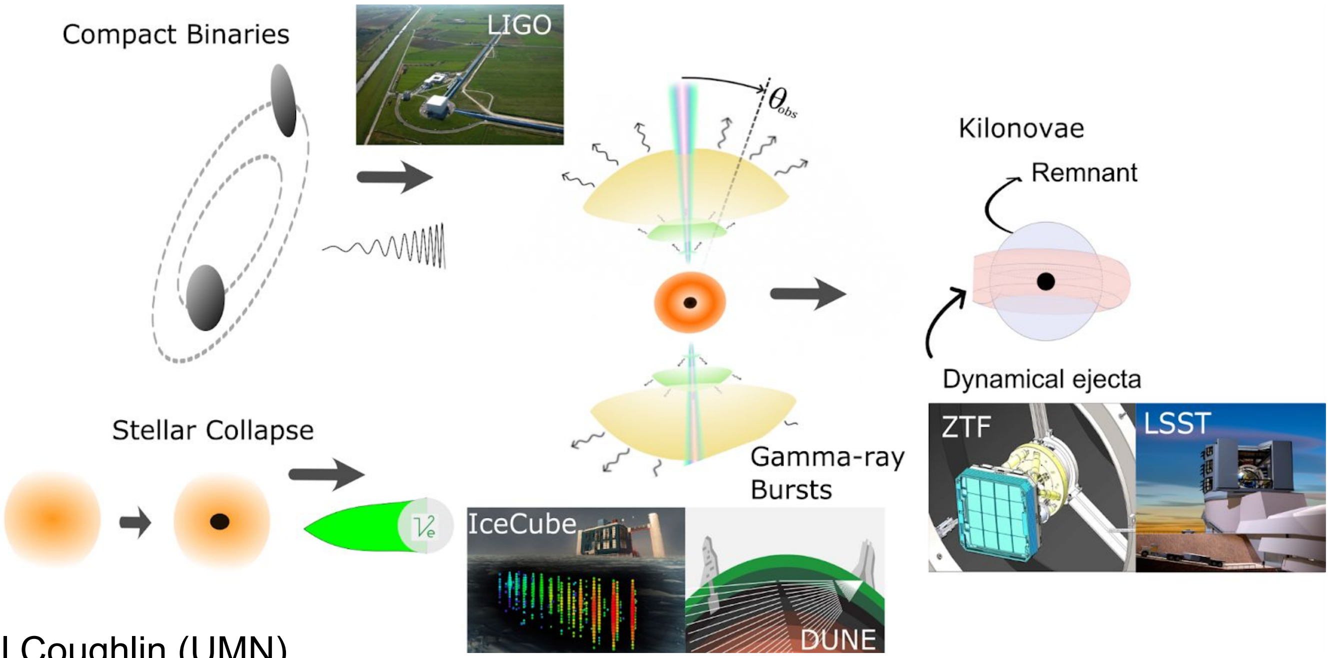
[CHEP 2023]()



CMS SONIC workflows
CMS
CMS global pool
Directed to directly connected GPU nodes
Sites with directly connected GPUs
GPU nodes excluded from SONIC server
GPUs
local host as SONIC server with GPU acceleration
Directed to sites without load balanced SONIC service
Sites with CPU only nodes
local host as SONIC server without GPU acceleration
CPUs
offloading
Directed to Purdue Tier 2 data center where SONIC service available via Kubernetes
SONIC GPU servers managed by Kubernetes
Purdue Tier 2 data center
NVIDIA
TRITON INFERENCE SERVER
GPUs
Results
gRPC
CMS SONIC workflows
at Purdue Tier 2 data center
CPUs
launch on demand server
results
CMS SONIC GPU servers
at Commercial clouds or HPC sites
ACCESS
NERSC
Fall back to local CPU
miro

Multi-messenger Astrophysics

- Develop and deploy software within astronomical facilities to enable discovery



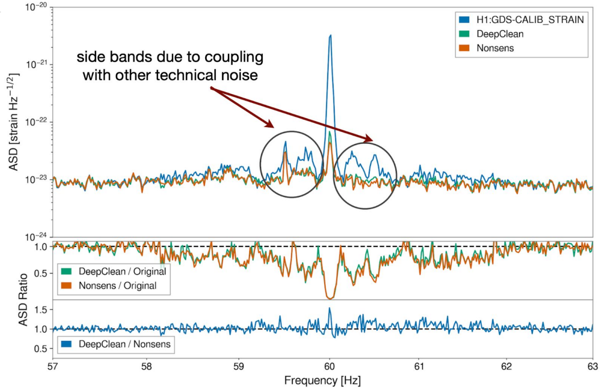
Credit: Michael Coughlin (UMN)

Gravitational Waves (LVK)

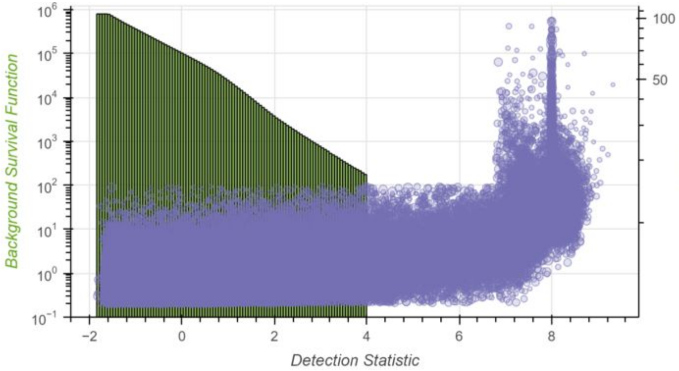
Github: ML4GW

All algorithms use our inference-as-a-service (IaaS) prototype to implement a real-time noise subtraction pipeline (DeepClean), detection (aframe/GWAK), and parameter estimation for use during the fourth observing run (O4) of LIGO-Virgo-KAGRA on dedicated hardware at the detector sites.

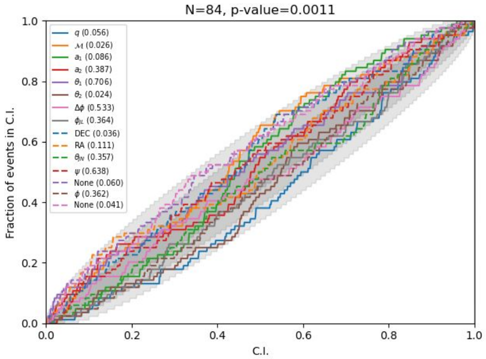
Clean the Data: DeepClean (CNN)



Detect the GWs: aframe (CNN)/GWAK (autoencoders)

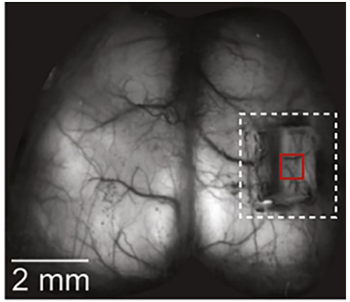


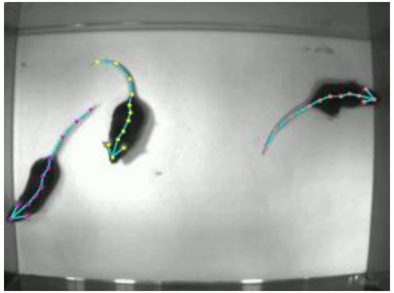
Characterize the GWs: (MAF*)



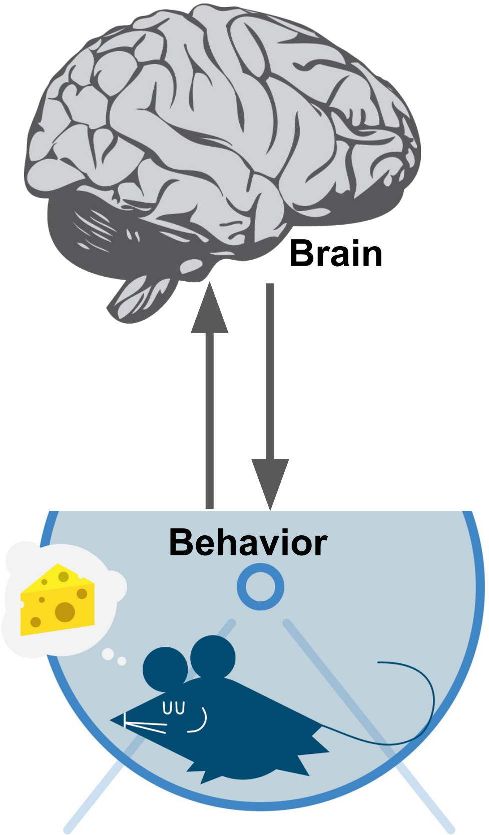
Neuroscience needs high-throughput & real-time AI

Rapid increase in number, type of measurements

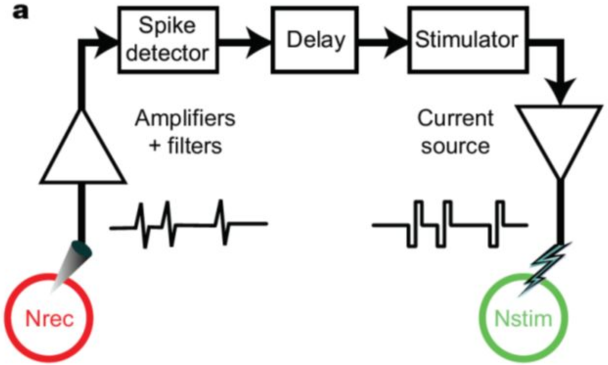




Need: data-driven discovery of relevant features, structure in data



Must *perturb* the system to disentangle causality, treat disorders.

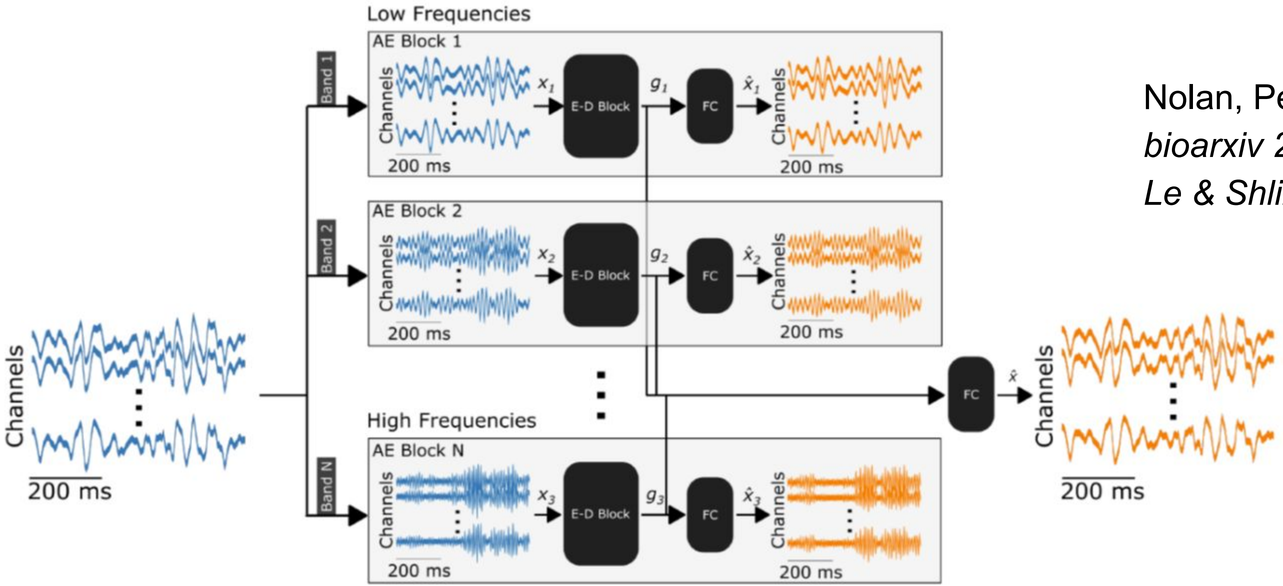


Need: low-latency algorithms (<1ms)

Improved time-series reconstruction methods

- Developed new Multi-block Recurrent Auto-Encoder (MRAE) to increase bandwidth more efficiently

- Developed Spatio-Temporal Transformer for Spiking Neural Data

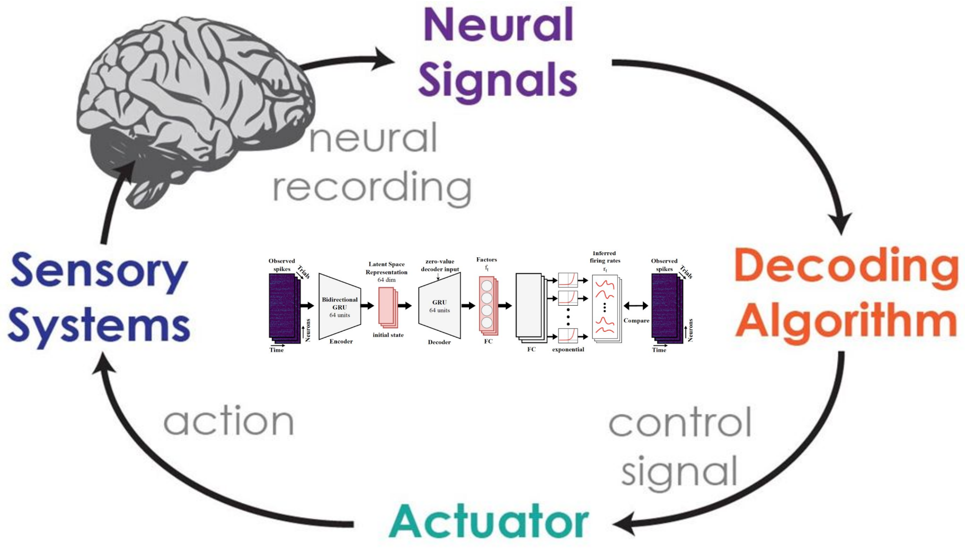


Nolan, Pesaran, Shlizerman & Orsborn, *bioarxiv 2022*

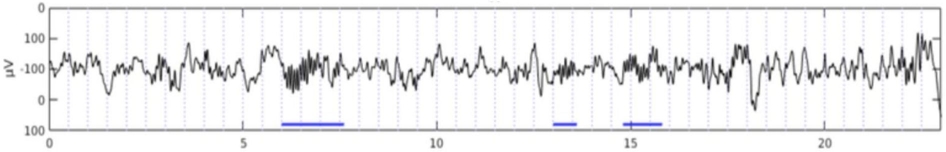
Le & Shlizerman, NeurIPS 2022

NeuroAI Integration

- A popular autoencoder model used on neural data (LFADS) in FPGA, Elham Khoda's talk



- Neuro A3D3 develops methods for reconstruction, forecasting and clustering of time-series
- Potential applications/uses:
  - Detect noise and artifacts
  - Detect rare neural events of interest (e.g., seizures, spindles, etc)



Not a spindle

spindle

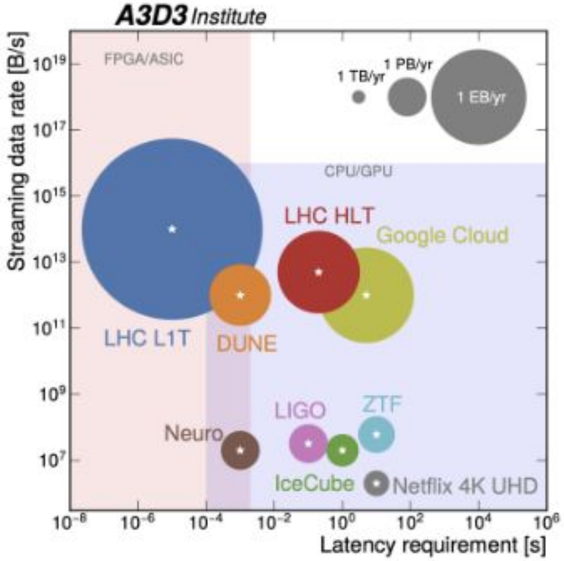
spindle

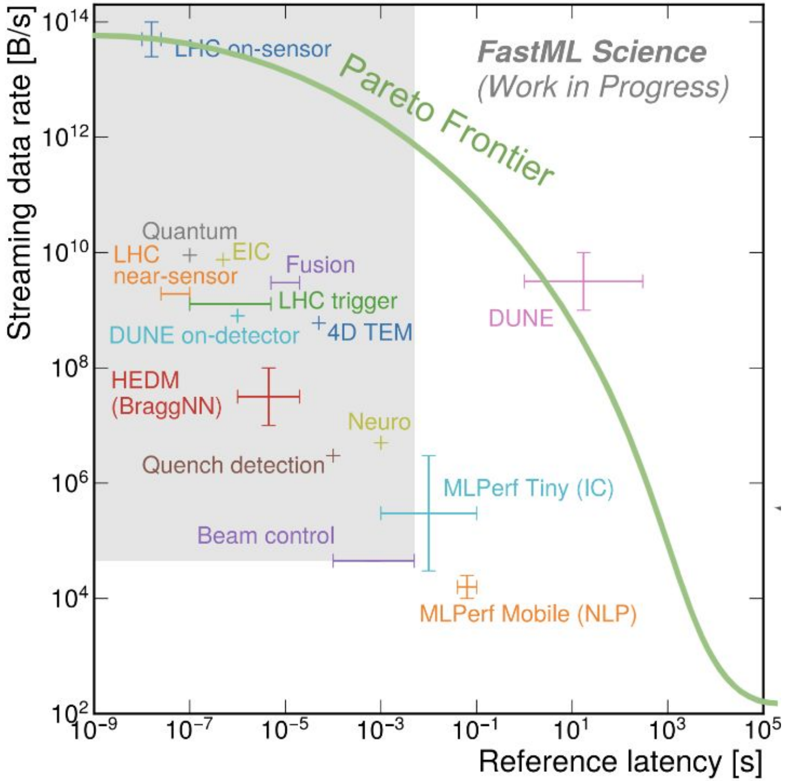
spindle

Not a spindle

Fast Machine Learning Community

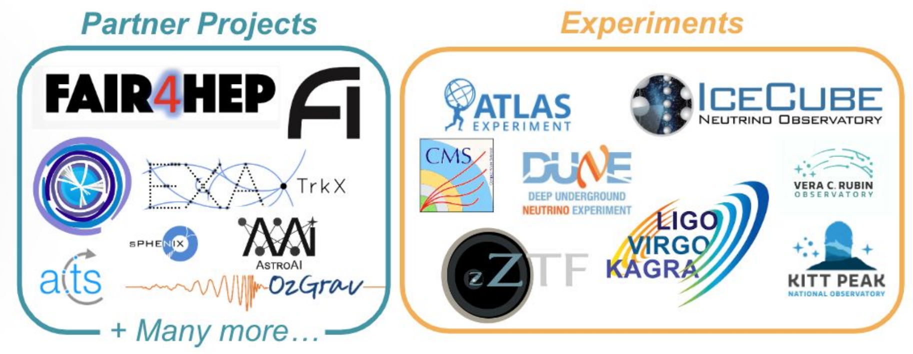
Our aim is to build a large-scale public scheme to advertise this work

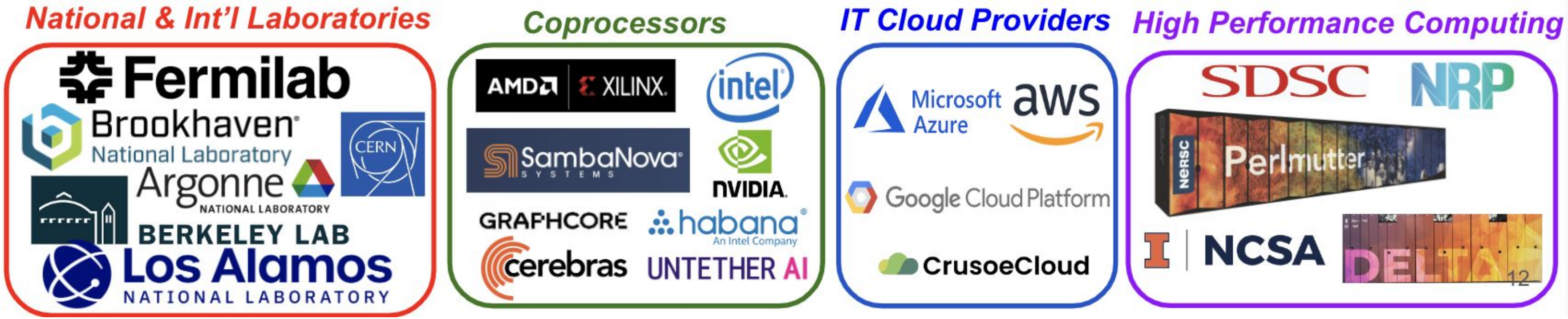




Partnership and FastML Ecosystem

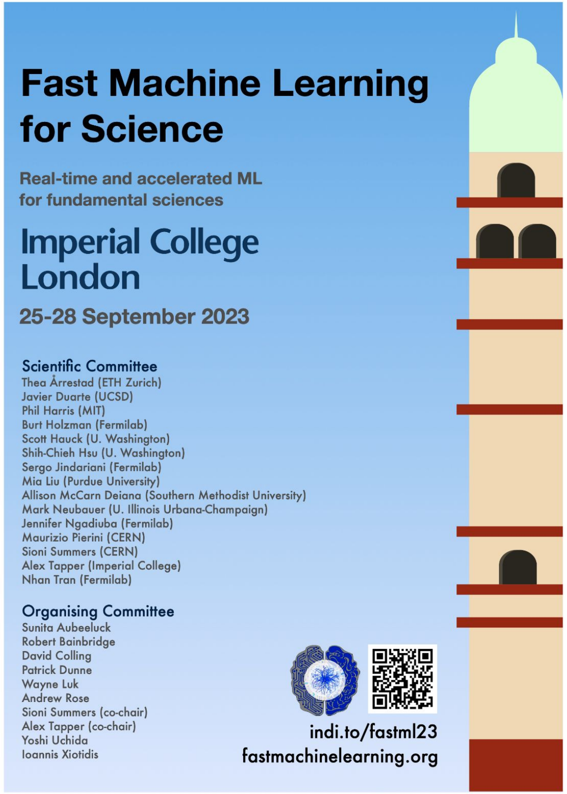
Growing strong industry connections with support through the Fast ML community


Partner Projects
FAIR4HEP
Exa.TrkX
sPHENIX
AstroAI
acts
OzGrav
+ Many more…
Experiments
ATLAS EXPERIMENT
ICECUBE NEUTRINO OBSERVATORY
CMS
DUNE DEEP UNDERGROUND NEUTRINO EXPERIMENT
VERA C. RUBIN OBSERVATORY
ZTF
LIGO VIRGO KAGRA
KITT PEAK NATIONAL OBSERVATORY



A3D3 Ecosystem & Engagement

- [***High-Throughput AI Methods and Infrastructure Workshop***]
- [***Postbaccalaureate Workshop***]



Fast Machine Learning for Science

Real-time and accelerated ML
for fundamental sciences

Imperial College
London

25-28 September 2023

Scientific Committee
Thea Årrestad (ETH Zurich)
Javier Duarte (UCSD)
Phil Harris (MIT)
Burt Holzman (Fermilab)
Scott Hauck (U. Washington)
Shih-Chieh Hsu (U. Washington)
Sergo Jindariani (Fermilab)
Mia Liu (Purdue University)
Allison McCarn Deiana (Southern Methodist University)
Mark Neubauer (U. Illinois Urbana-Champaign)
Jennifer Ngadiuba (Fermilab)
Maurizio Pierini (CERN)
Sioni Summers (CERN)
Alex Tapper (Imperial College)
Nhan Tran (Fermilab)

Organising Committee
Sunita Aubeeluck
Robert Bainbridge
David Colling
Patrick Dunne
Wayne Luk
Andrew Rose
Sioni Summers (co-chair)
Alex Tapper (co-chair)
Yoshi Uchida
Ioannis Xiotidis

indi.to/fastml23
fastmachinelearning.org

https://indico.cern.ch/event/1283970

Summary

- A3D3 focusing on accelerating real-time AI to solve common challenges through interdisciplinary collaboration
  - 4 focus areas: HAC, HEP, MMA, Neuros
  - 2 integrated systems: Targeted system, Hetereogenous computing
- A3D3 is closely connected with the FastML Community
  - Leverage our leadership in FastML to connect to main different domains
  - Touches on many fields in industry/science not part of A3D3 scope
    - Plasma Physics/Materials Science/…/ASIC design
- Welcome to participate in A3D3 activities
  - HDR Ecosystem Workshops
  - Postbac Program Enhancements
  - Machine Learning Challenges
    - Nov 17 planning meeting https://indico.cern.ch/event/1342015/

Shih-Chieh Hsu
http://faculty.washington.edu/schsu/
schsu@uw.edu

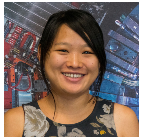
Cross-discipline

HEP

Hsu
PI

Harris
co-PI

Neubauer
co-PI


Liu

Duarte

MMA

Coughlin
co-PI

Scholberg
co-PI

Graham

Hanson

Katsavounidis

Neuros

Orsborn

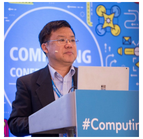
Shlizerman

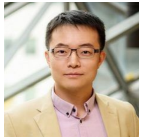
Dadarlat Makin

CS/EE

Hauck

Li

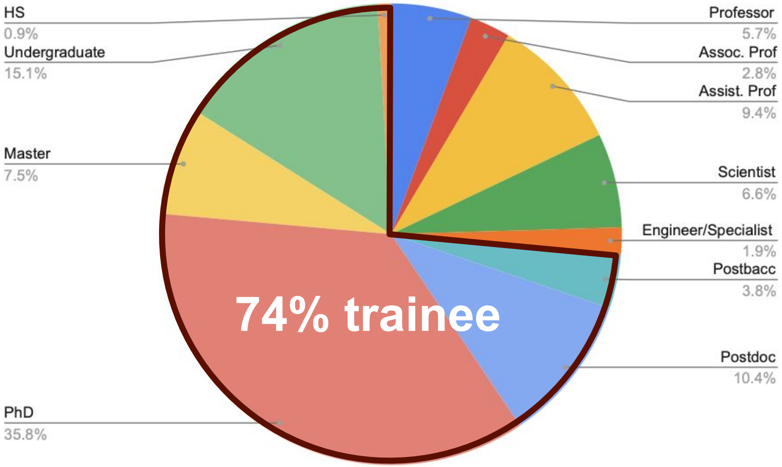

Chen


Han

17 Senior Personal

A3D3 fully staffed

106 Members (including 5 affiliate)



Prog. Ope. Spec.

Zhang (UW)

Affiliate faculty/staff

Rankin (Upenn)
A3D3 Alumni

Sravan (Drexel)
A3D3 Alumni

Ju (LBNL)

Lai (NYCU)

Carlson (Westmont)

PostBacc fellow

Gray (UMN)

Peterson (UW)

Lian (Duke)

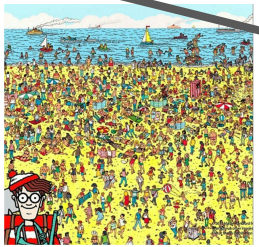
Skivington (UCSD)

ML Challenge: Unifying across domains

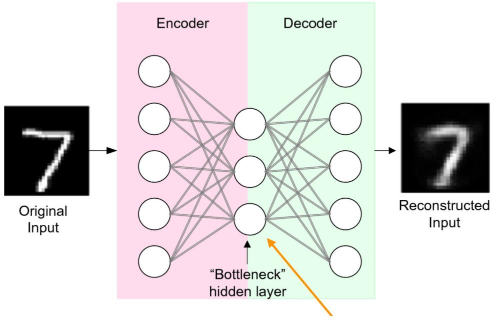
- challenge across HDR domains
    - Try to find anomalies over many different datasets with one metric

Would be a FAIR workflow challenge?

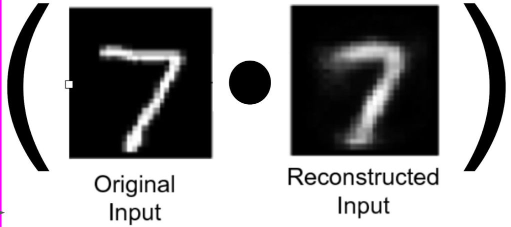
Could extend this to semi/self-supervised learning (foundation models)



Many Datasets covering whole HDR



Anomaly Algorithm



Dot product the input and output

Large Value : Good

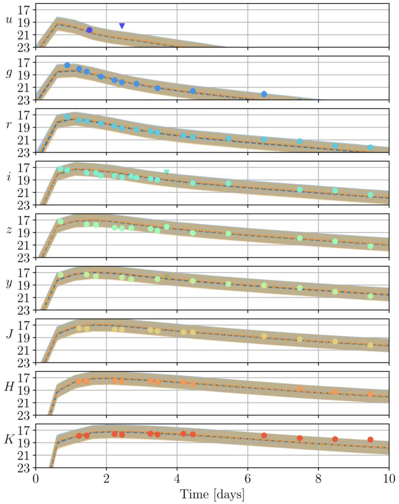
Small Value : Anomaly

Anomaly Metric

Optical Astronomy - Overview

Accelerated AI Algorithms for Data-Driven Discovery

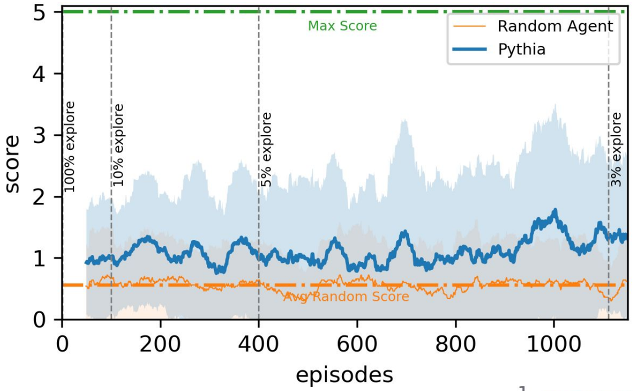
Simulate Observations: NMMA (emulator)



Github work areas:

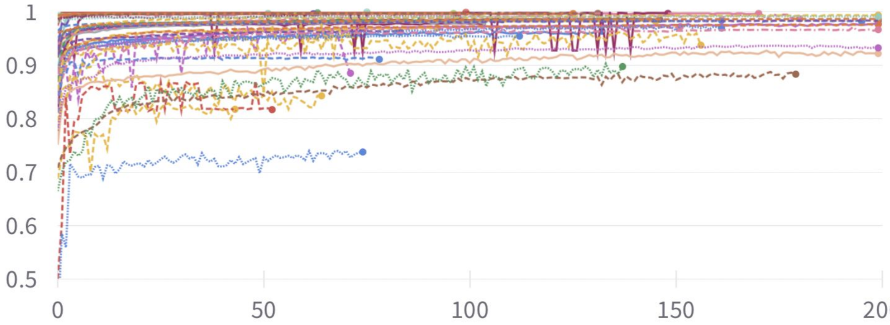
NMMA SCOPE Pythia

Optimize Observations: Pythia (RL)



~4 faculty, 3 postdocs, 5 grad students, 3 postbac/undergraduat

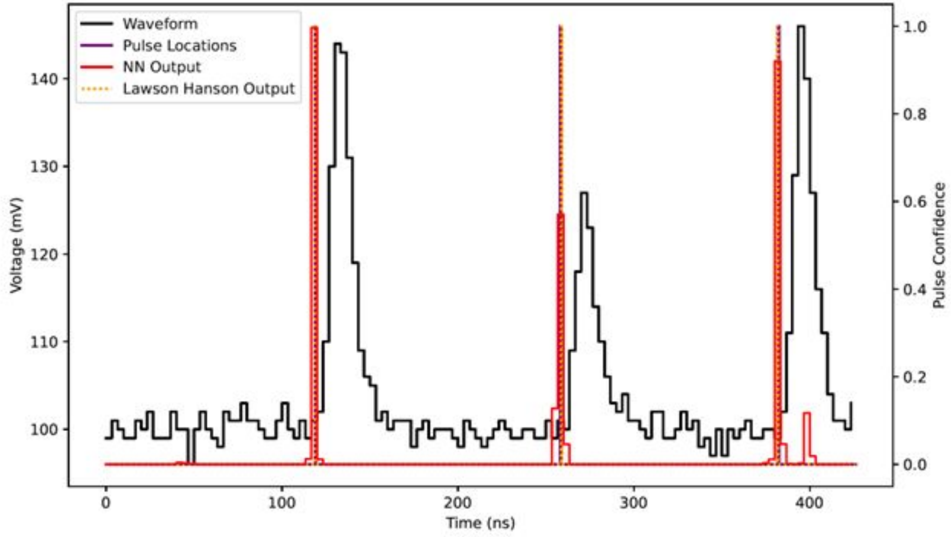
Classify the sources: Scope (CNN)



Main focus: Deploy ML algorithms throughout the observation preparation and follow-up for source identification and characterization

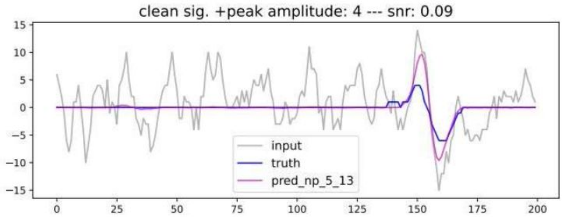
Neutrinos - Overview

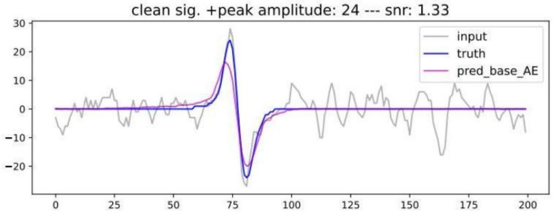
PMT Voltage Picking (CNN)



~2 faculty, 2 postdocs, 2 grad students, 2 postbac/undergraduates

Supernova Reconstruction (1DCNN autoencoder + pointing)





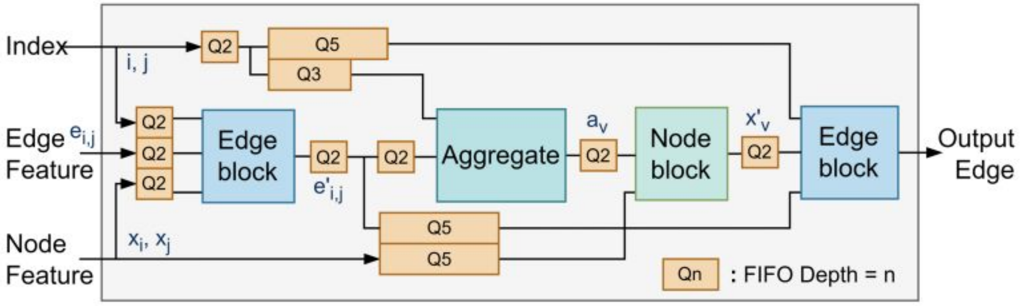
Main focus: Porting existing algorithms to GPUs and FPGAs for the purpose of detection and localization reconstruction.

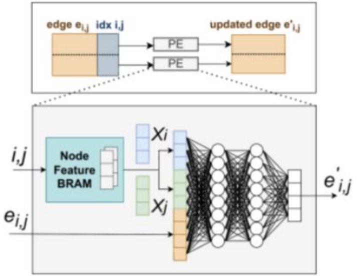
See: See Pan's Talk in Hardware-Algorithm Co-Development

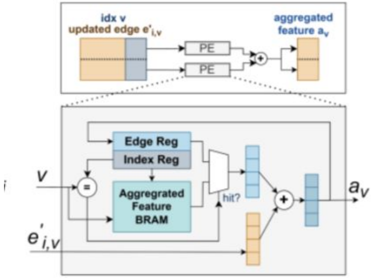
LOW LATENCY EDGE CLASSIFICATION GNN

Shi-Yu Huang, Yun-Chen Yang, Yu-Ru Si, et. al. FPL 2023

Modularized parallel architecture for each computational pipelines







Achieving 2.07 us Latency with 3.225 Throughput (MGPS)

- Xilinx Virtex UltraScale+ VU9P HLS 2019.2

National Lab: HLS4ML for Analog AI

- Project: “*Democratizing AI Hardware with an Open Source, Automated AI-Chip Design Toolkit* ”

- Joint initiative with Discovery Partners Institute and Fermilab

DISCOVERY PARTNERS INSTITUTE
PART OF THE UNIVERSITY OF ILLINOIS SYSTEM

Farah Fahim
Fermi Lab,
ASIC Research &
Development Head

Ben Parpillon
Fermi Lab,
Senior ASIC
Engineer

AI-Chip Prototyping and Analog Primitive Automation

Amit R. Trivedi
UIC,
Electrical and
Computer Engineering

Nhan Tran
Fermi Lab,
Accelerator-based
Experiments

High-Level Synthesis and Digital Automation Flow

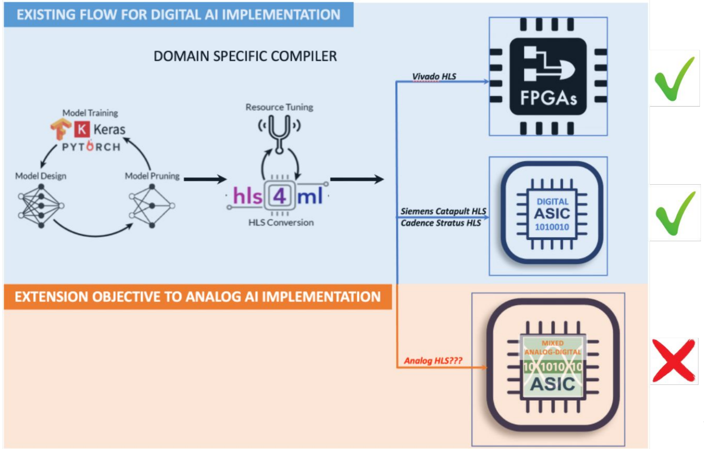
Ahmet Cetin
UIC,
Electrical and
Computer Engineering

Mark Neubauer
UIUC,
High Energy
Physics

Application Studies: Low Barrier Custom-AI for Small Businesses

Why Analog AI?
More efficient, Better Latency, Less Area

Why Automate Analog AI?
Cheaper, faster, less risky implementation



EXISTING FLOW FOR DIGITAL AI IMPLEMENTATION

EXTENSION OBJECTIVE TO ANALOG AI IMPLEMENTATION

Industry: Real-time Blood Cell Id

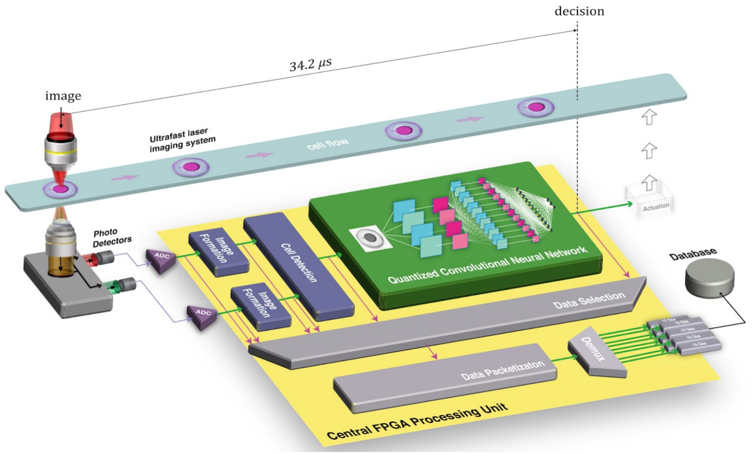


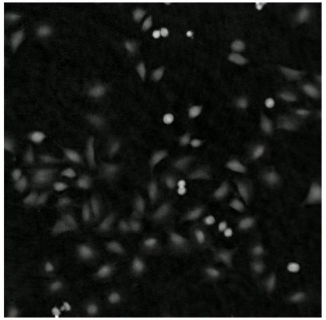
Diagram from: ieee paper

- Collaboration between MIT, CERN and Phiab
  - Led/initiated by Vladimir Loncar
- Working to bring HLS4ML to cell identification
  - Working directly with industry to deploy
  - Builds on A3D3 AI initiatives

- Collaboration with https://phiab.com/

- Key Ideas
  - Real time tagging of blood cells
  - Can be used for cell therapy
    - Cancers/….
  - Non-invasive
    - No chemicals
    - All electronics based

Original holography info



Segmented cell instances