Highlights and Opportunities in AI at Fermilab

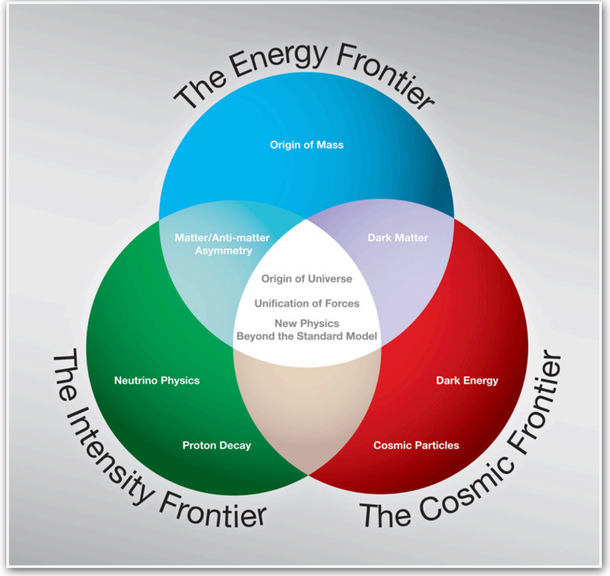
Nhan Tran, Fermilab

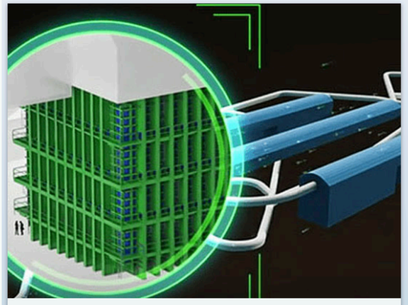
November 12, 2019

UC/ANL/FNAL Computational Science Seminar

Fermilab is America's particle physics and accelerator laboratory

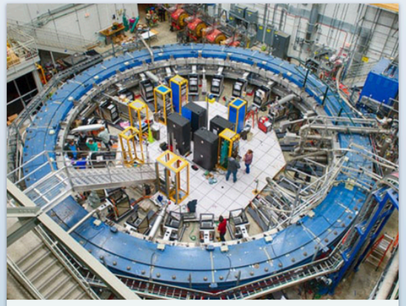
We bring the world together to solve the mysteries of matter, energy, space and time.





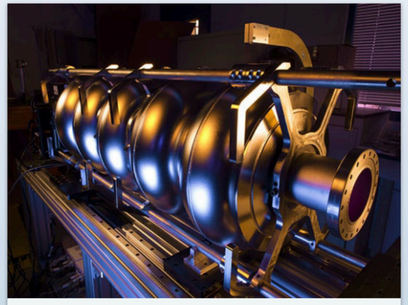
Deep Underground Neutrino Experiment

Fermilab hosts DUNE and the Long-Baseline Neutrino Facility, being built by scientists and engineers from more than 30 countries.



Particle physics

Fermilab explores the universe at the smallest and largest scales, studying the fundamental particles and forces that govern our universe.



Accelerator science and technology

Fermilab designs, builds and operates powerful accelerators to investigate nature's building blocks, advancing technology for science and society.

Detectors, computing and quantum science

Fermilab pioneers the research and development of particle detection technology and scientific computing applications and facilities.

Outline

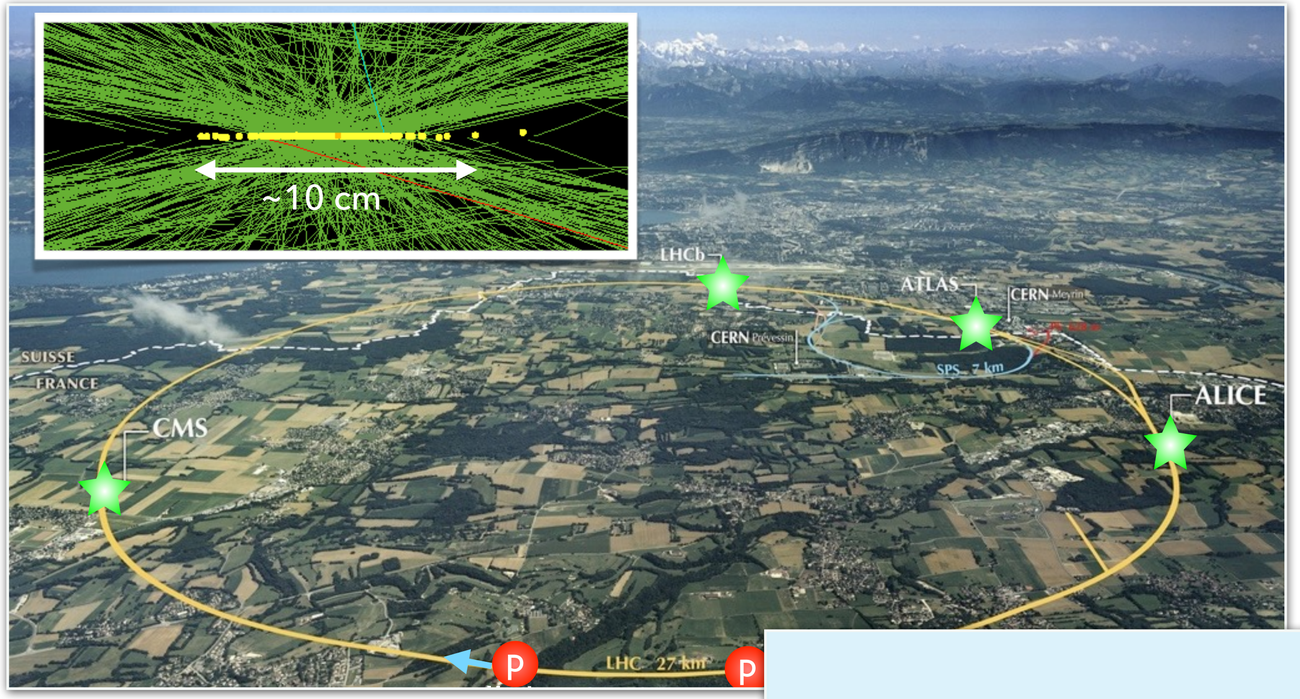
Fermilab through the lens of AI
scientific mission
enabling technologies at the bleeding edge

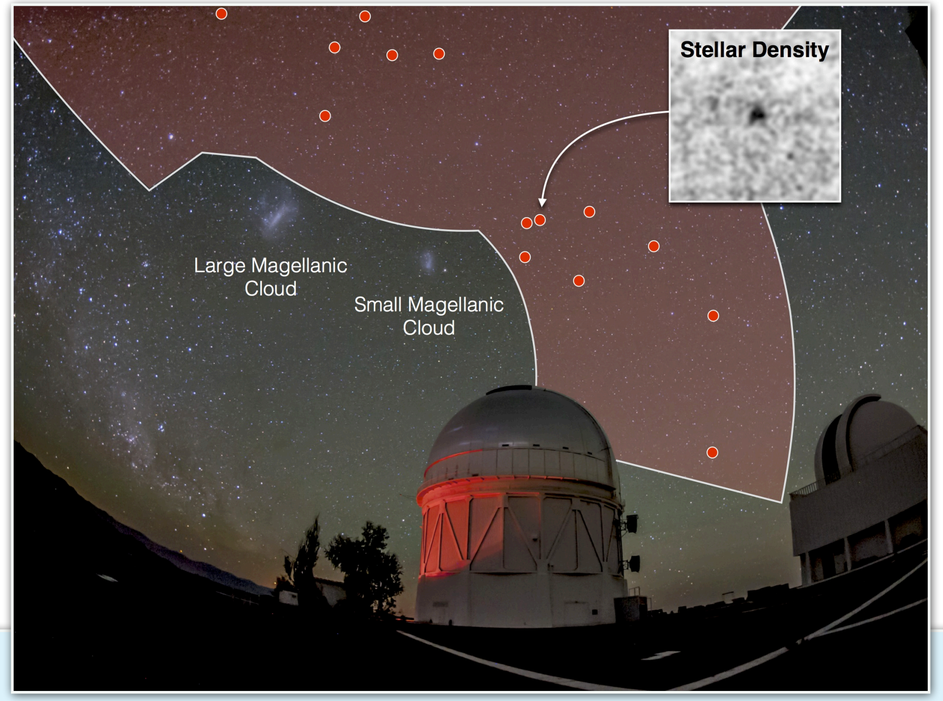
Capabilities and opportunities
AI development for HEP
opportunities to share and collaborate

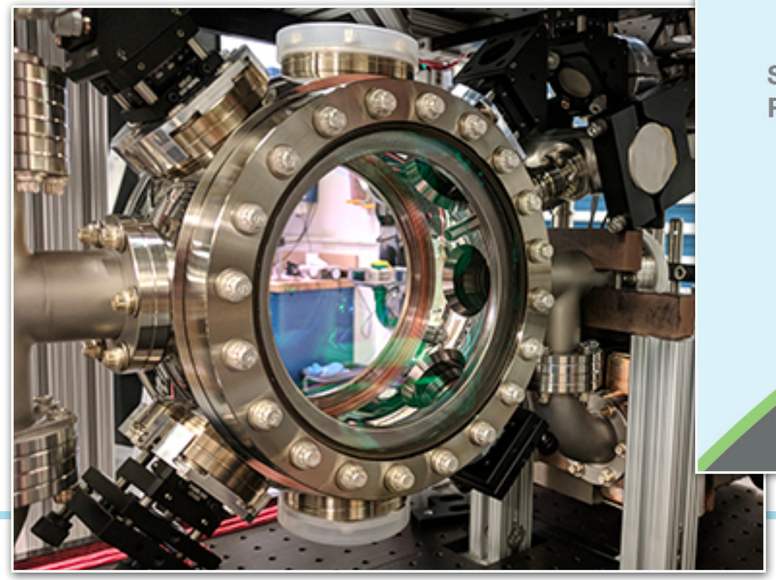
Outline

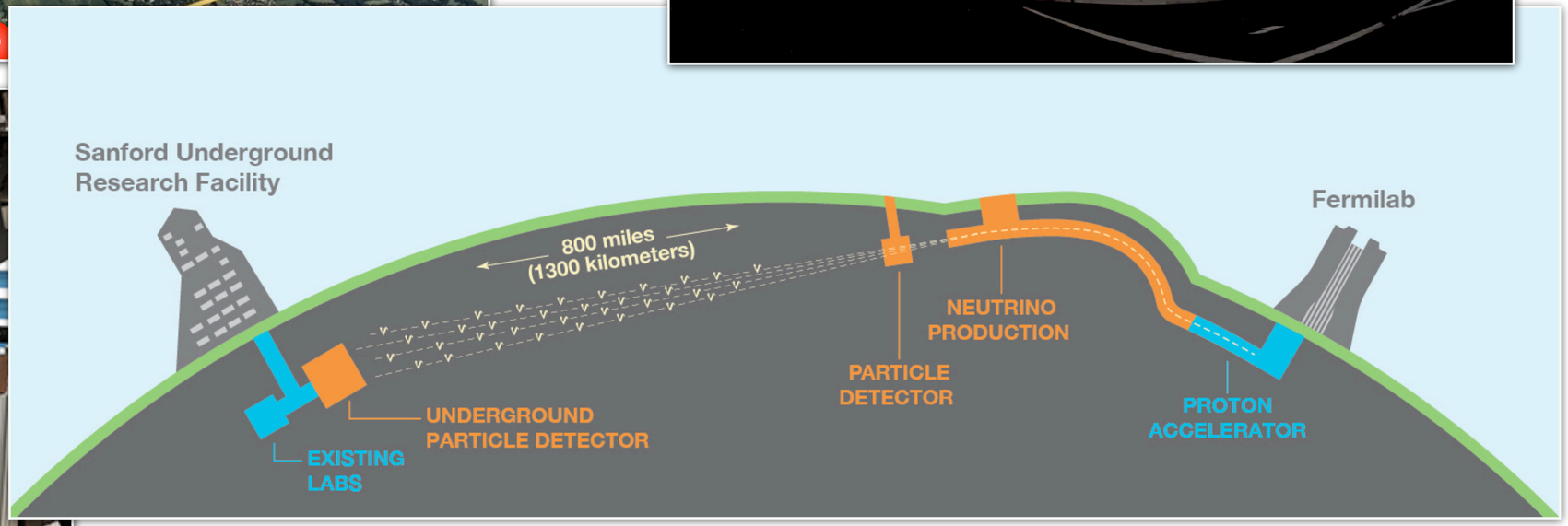
Fermilab through the lens of AI
scientific mission
enabling technologies at the bleeding edge

Capabilities and opportunities
AI development for HEP
opportunities to share and collaborate


~10 cm
LHCb
ATLAS
CERN Meyrin
CERN Prévessin
SPS 7 km
SUISSE
FRANCE
CMS
ALICE
P
LHC 27 km
P


Stellar Density
Large Magellanic Cloud
Small Magellanic Cloud




Sanford Underground Research Facility
800 miles (1300 kilometers)
Fermilab
NEUTRINO PRODUCTION
PARTICLE DETECTOR
UNDERGROUND PARTICLE DETECTOR
EXISTING LABS
PROTON ACCELERATOR

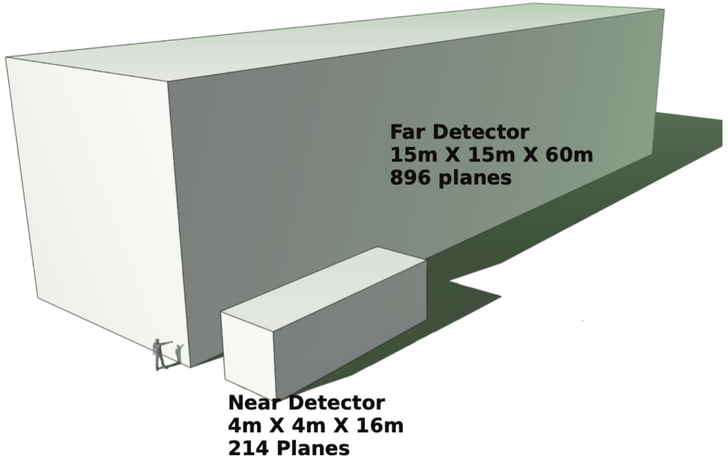
Neutrinos at NOvA

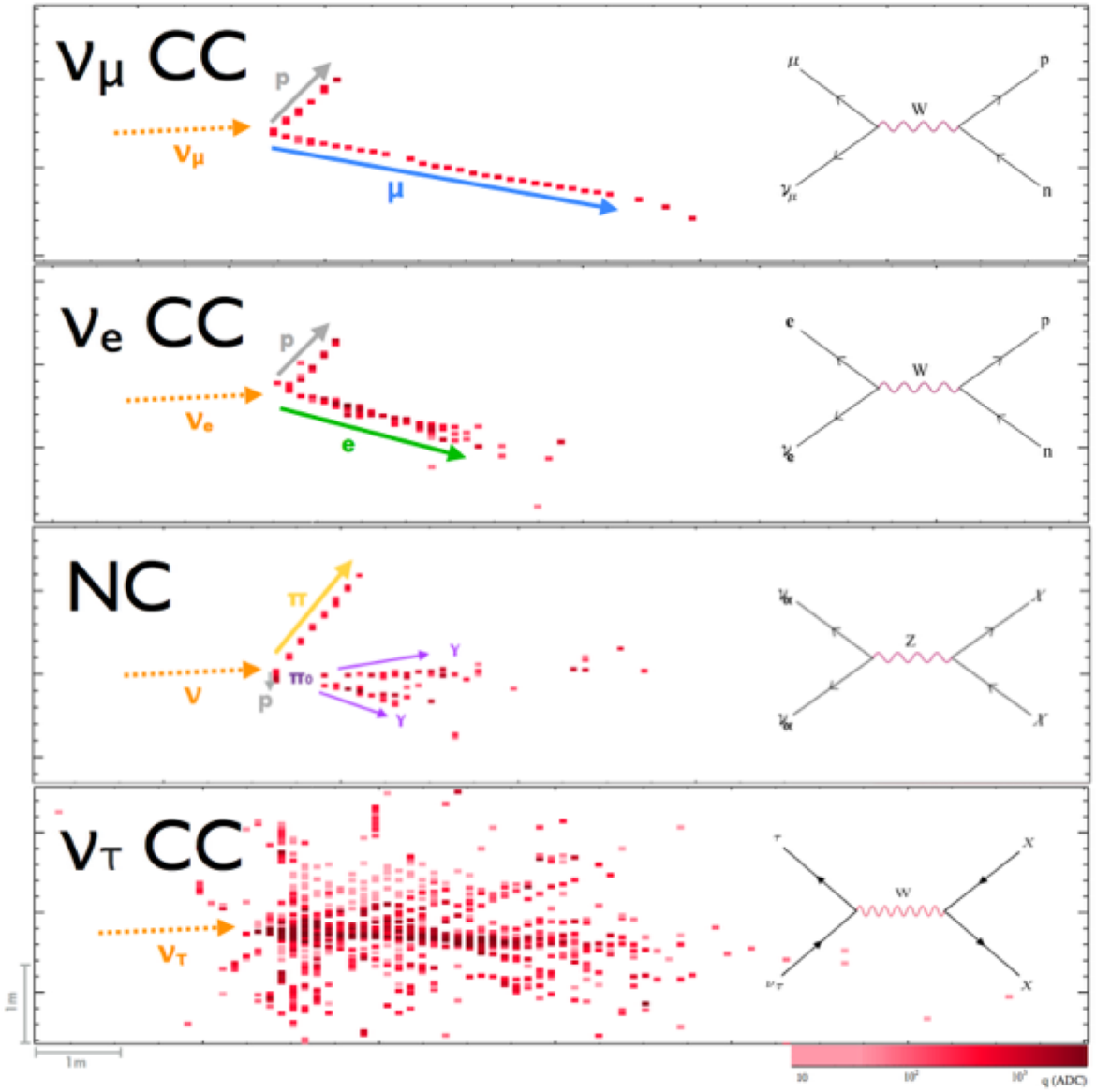
Near Detector

0.3 kt, > 20,0000 channels
100 m below service
1 km from NuMI

Far Detector

14 kt, > 344,000 channels
On surface
810 km from NuMI

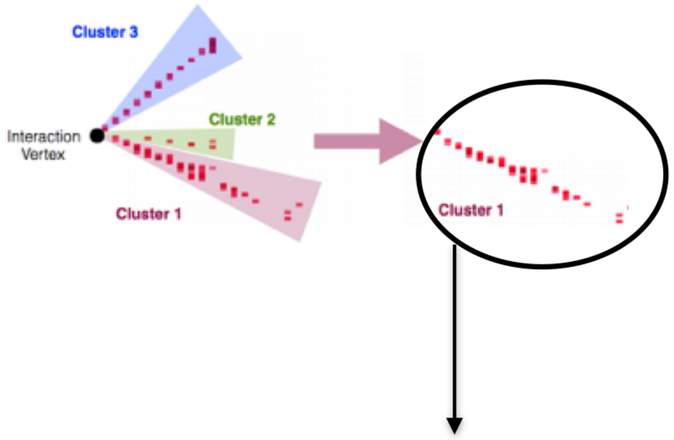


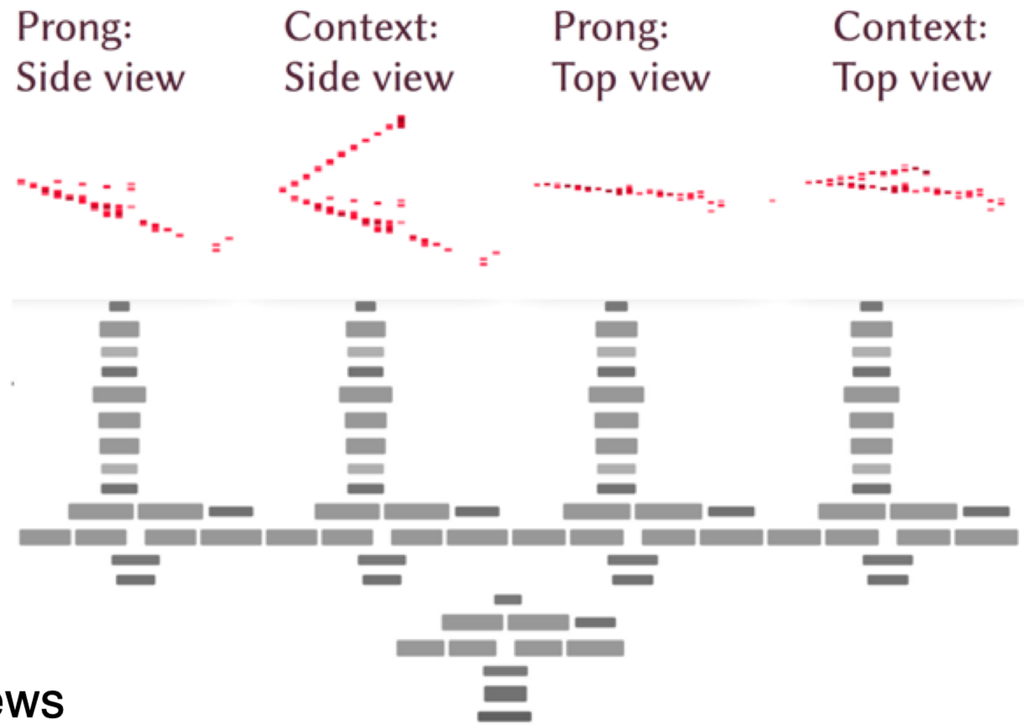


Example: NOvA Reconstruction

https://arxiv.org/abs/1604.01444
https://arxiv.org/abs/1703.03328

Single particles are separated using geometric reconstruction methods.

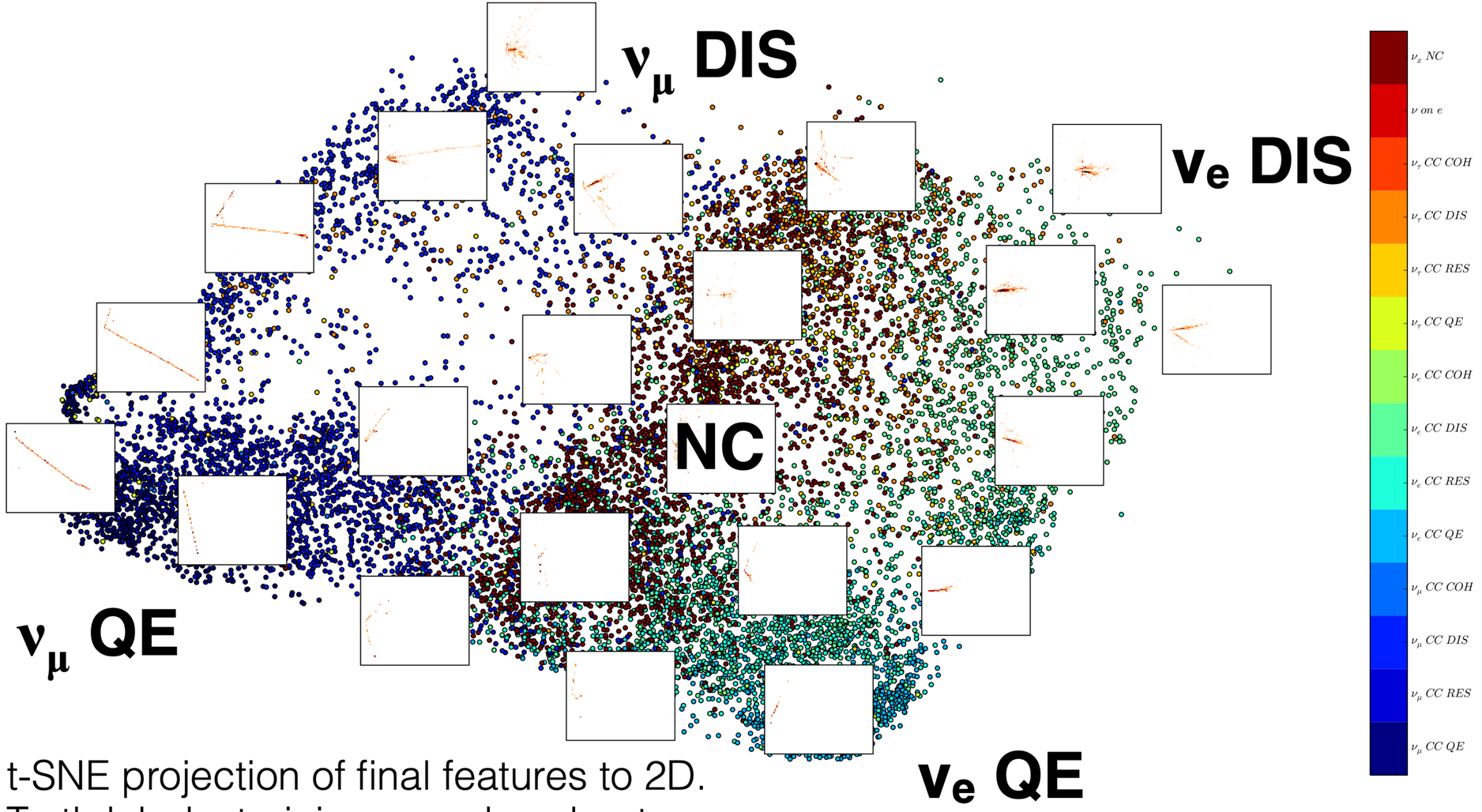




Classify particles using full event topology from both views as well as reconstructed cluster information (***4 views***)

Performance improvement equivalent to **4.2 kilotons of additional detector mass** with traditional particle identification algorithms.

NOvA: what did the network learn?

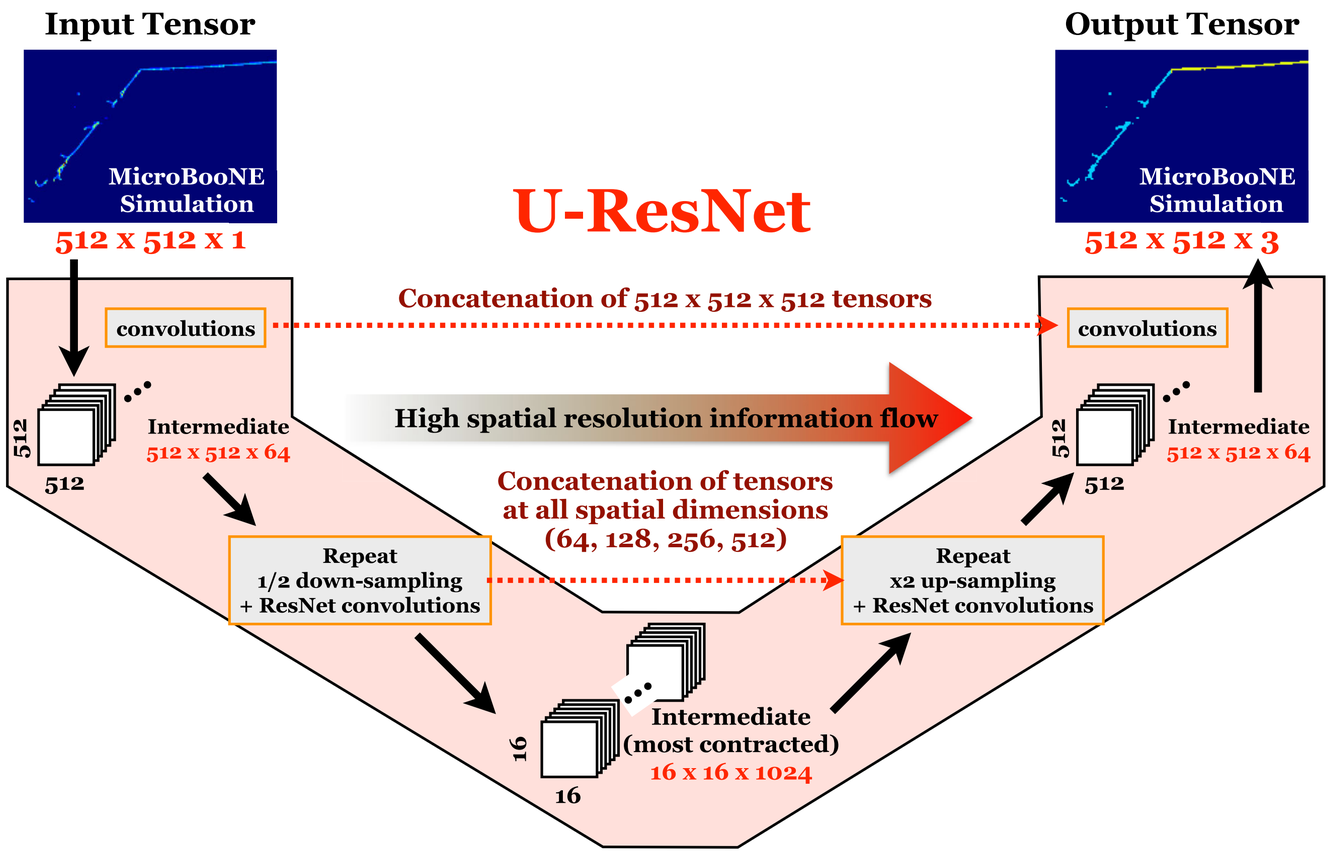


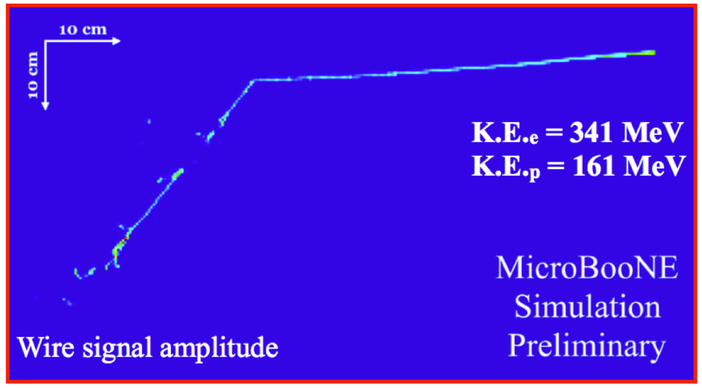
t-SNE projection of final features to 2D.
Truth labels, training sample subset.

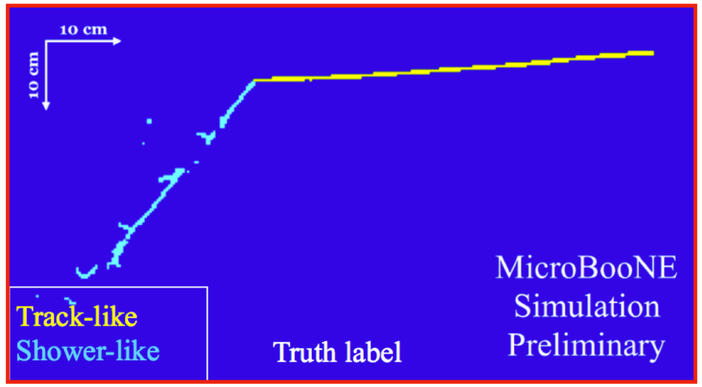
μBooNE: semantic segmentation

https://arxiv.org/abs/1808.07269

Liquid Argon Time-Projection Chamber detector

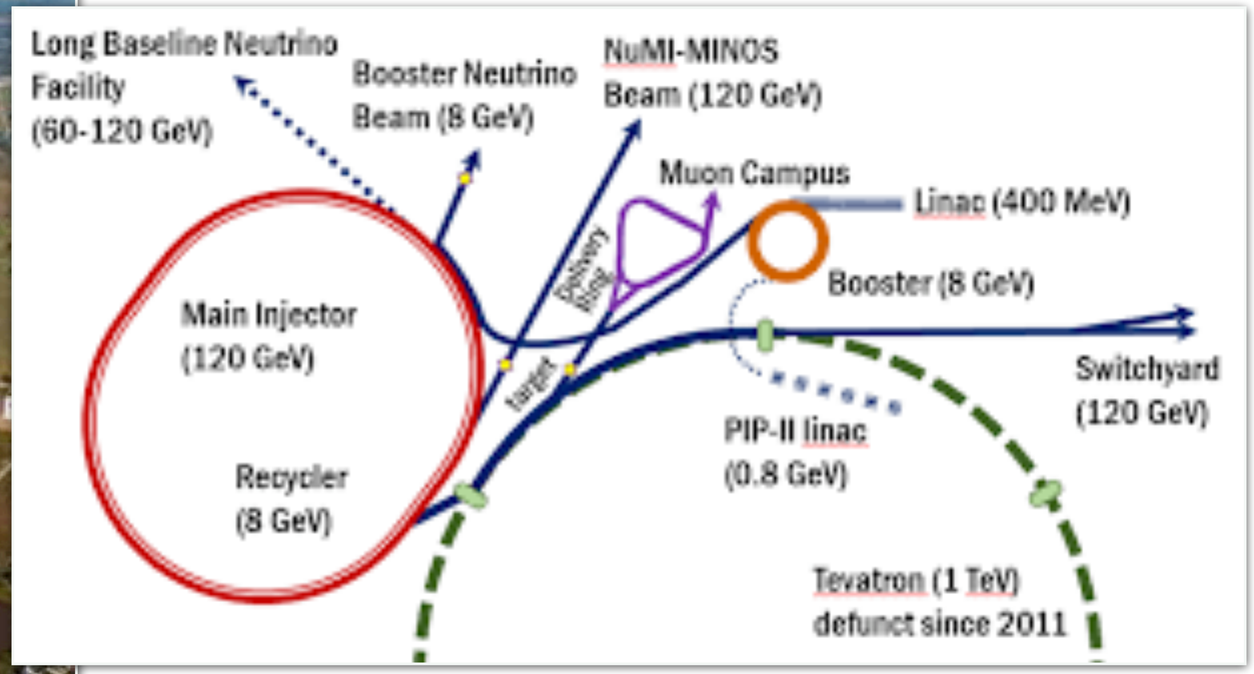




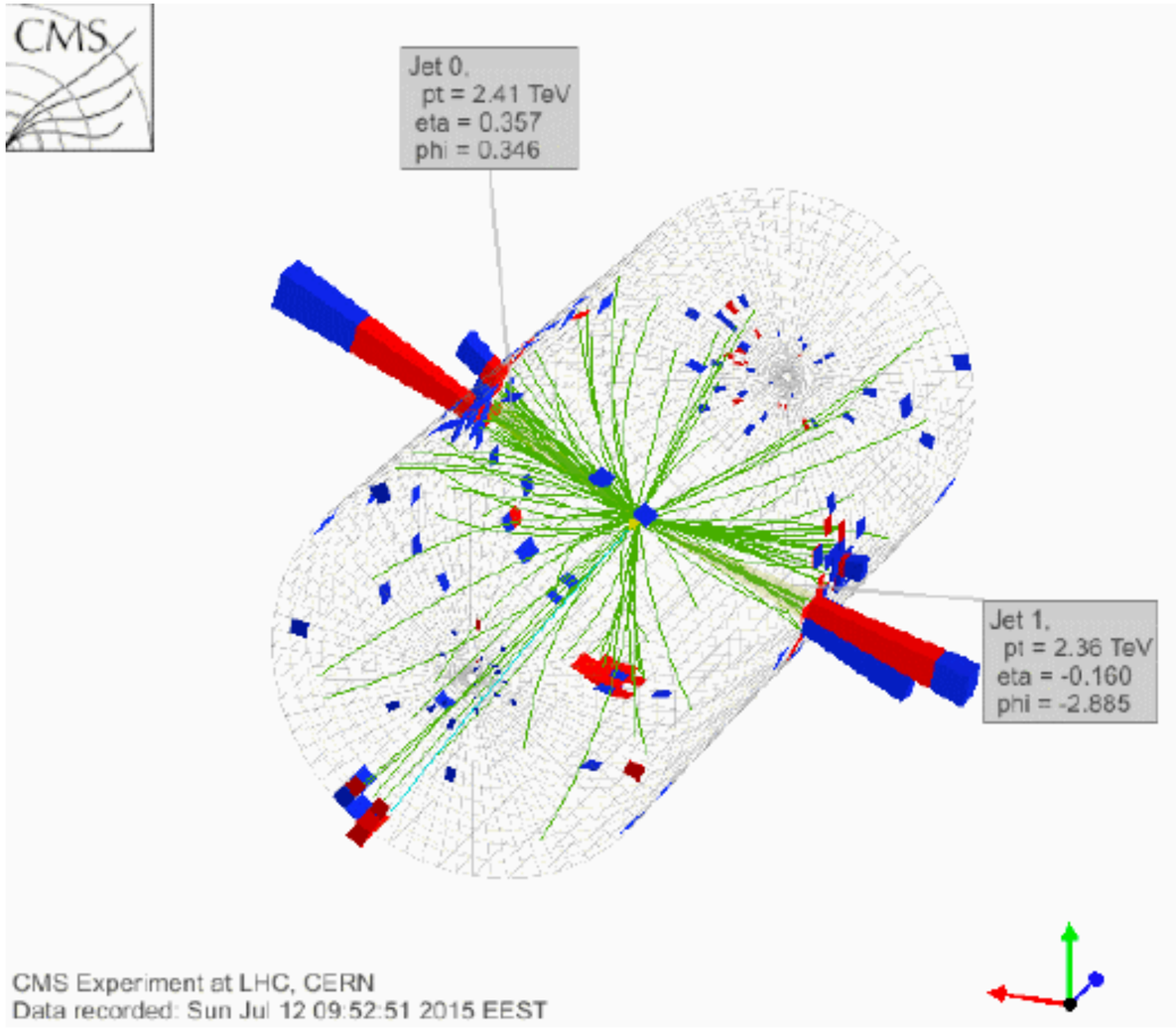


Encoder-decoder network pixel-wise classification of track/shower/background

Large, integrated systems and operations

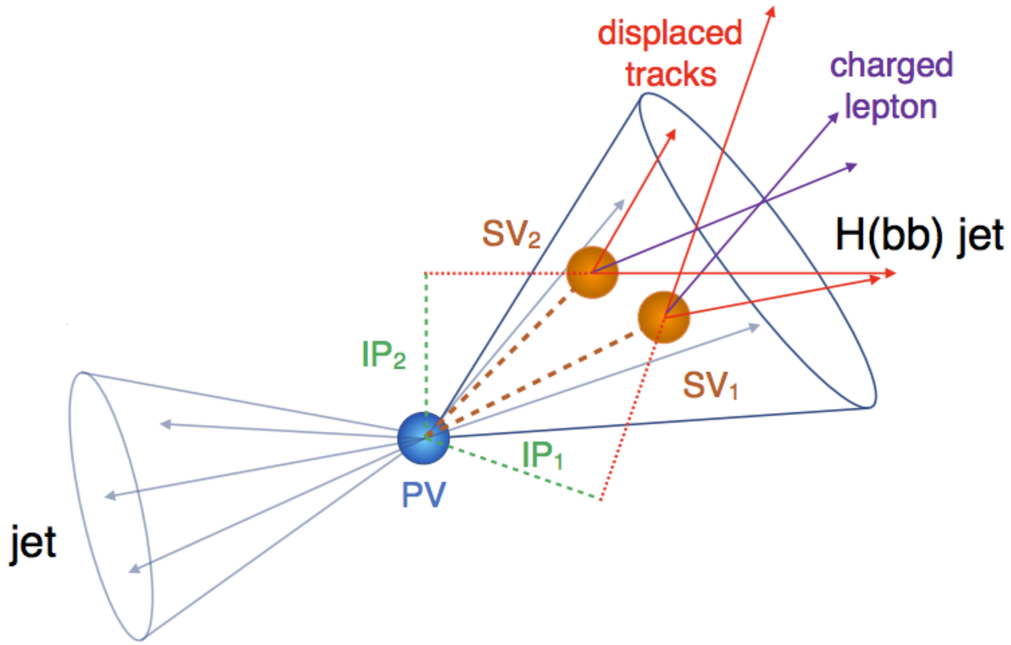


At the energy frontier



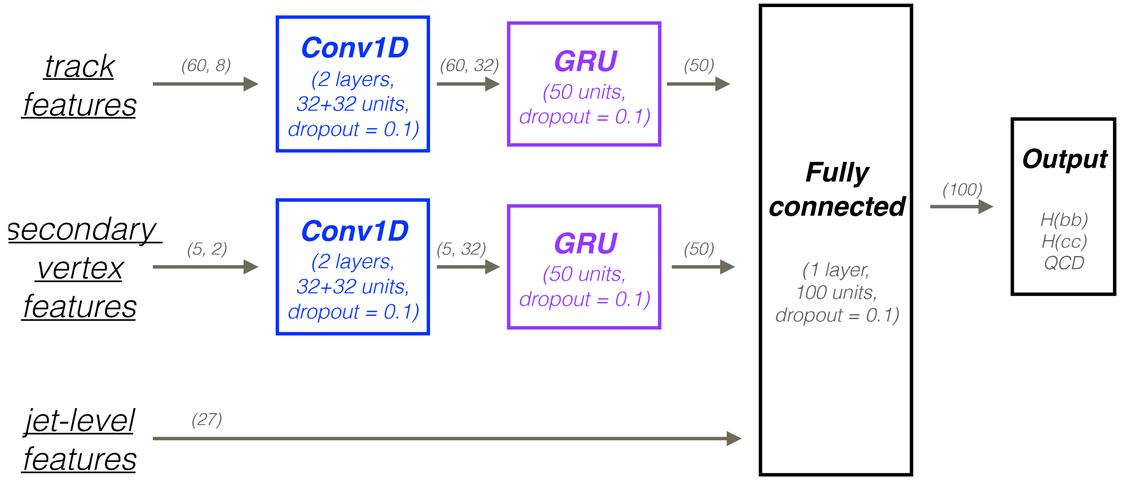
Fermilab on Compact Muon Solenoid experiment at the LHC

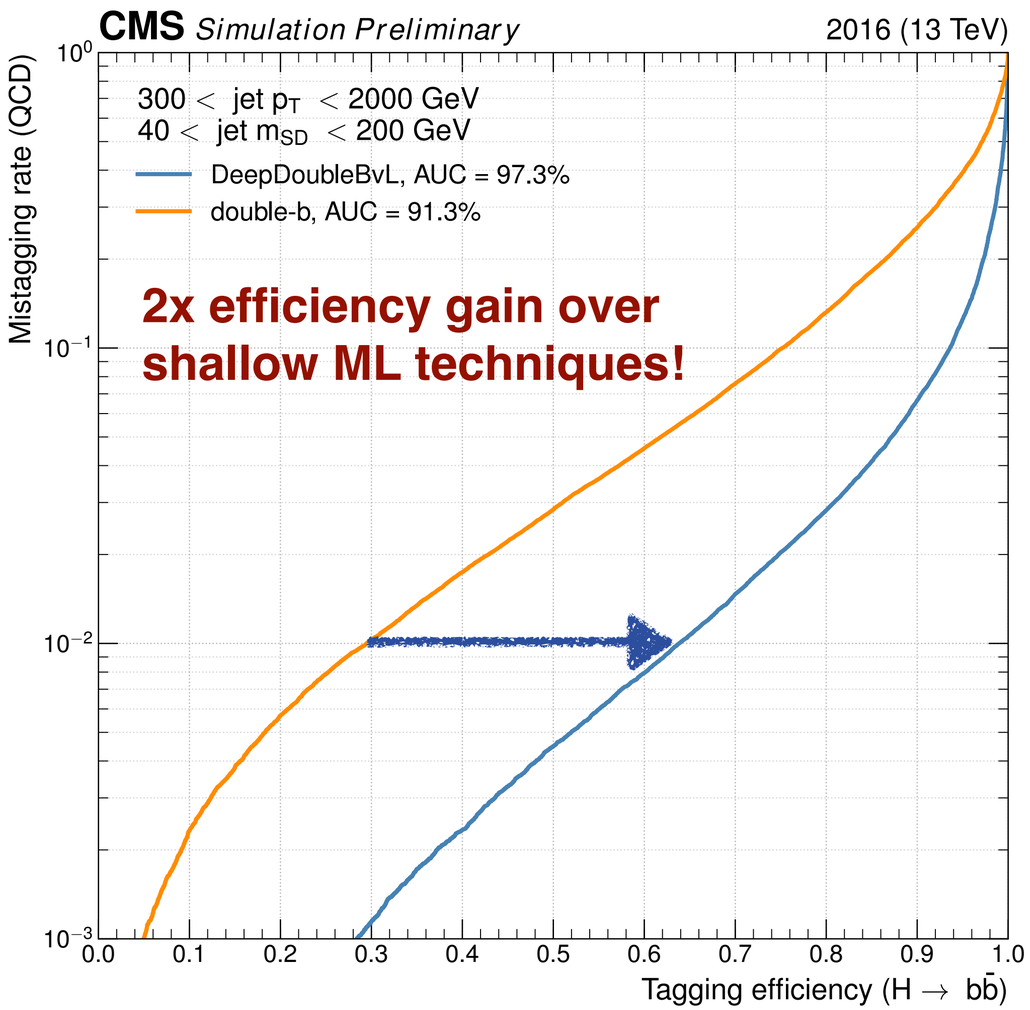
Broad range of physics from the Higgs boson to dark matter



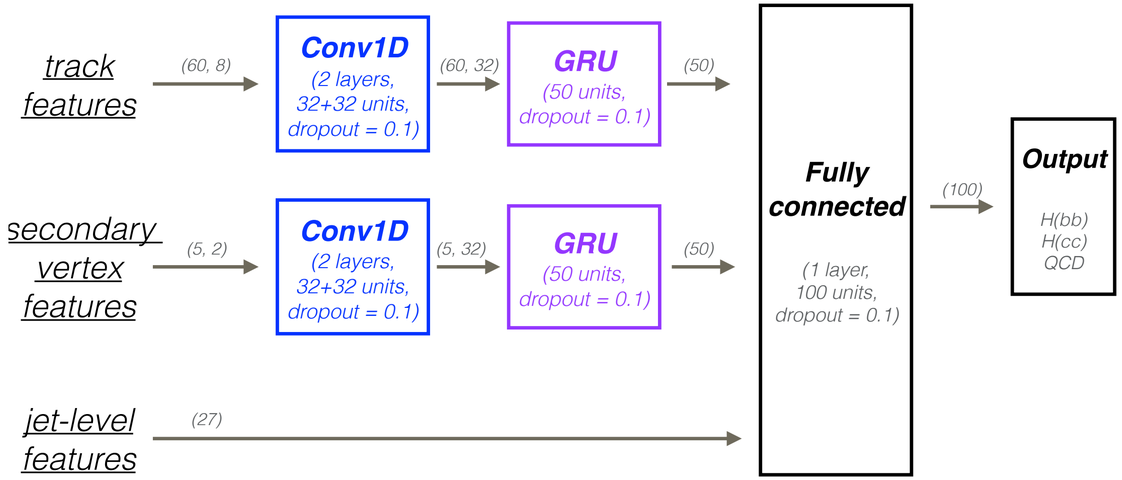
New deep learning techniques to identify the Higgs boson in dense, energetic decays to bottom quarks

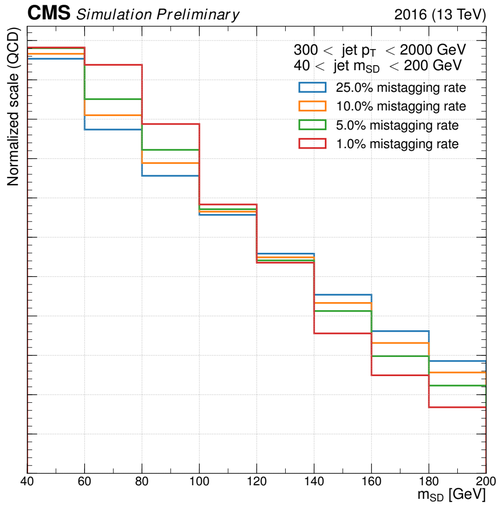
Example: Identifying the Higgs



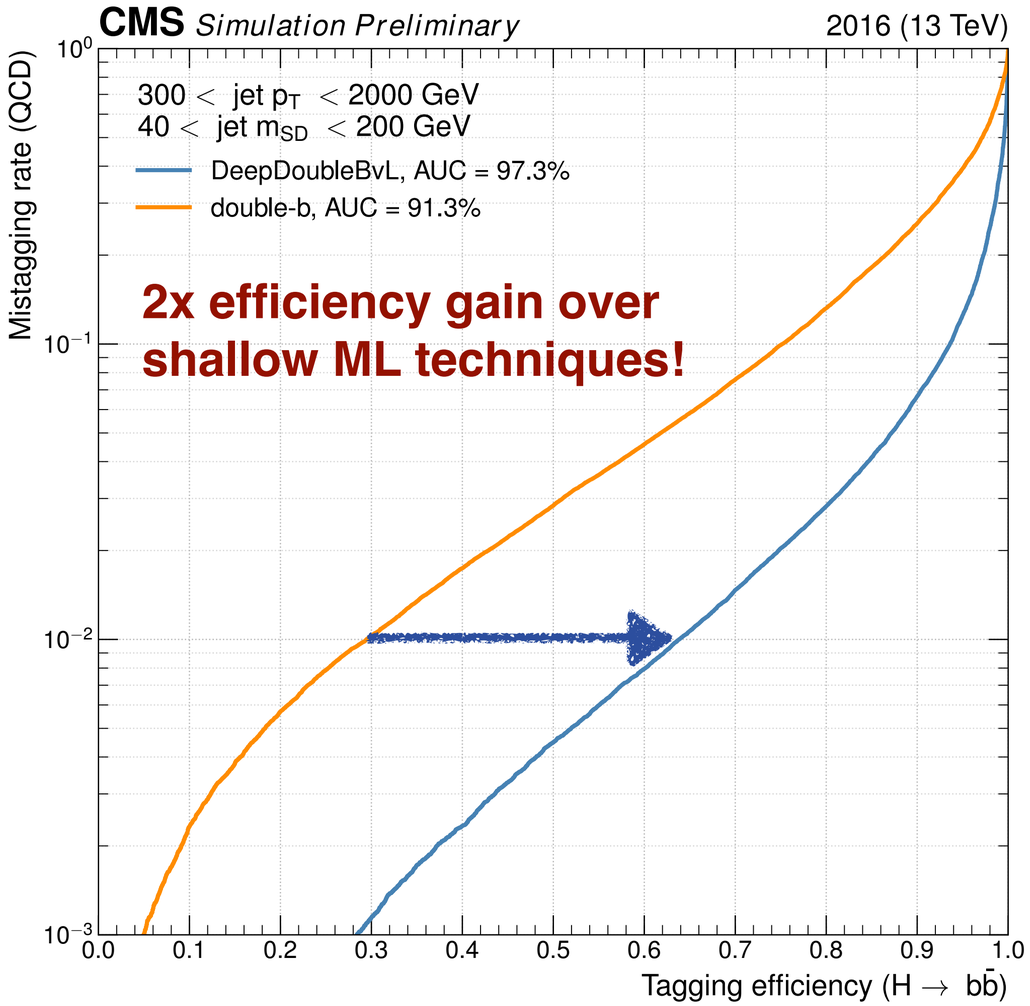


Example: Identifying the Higgs



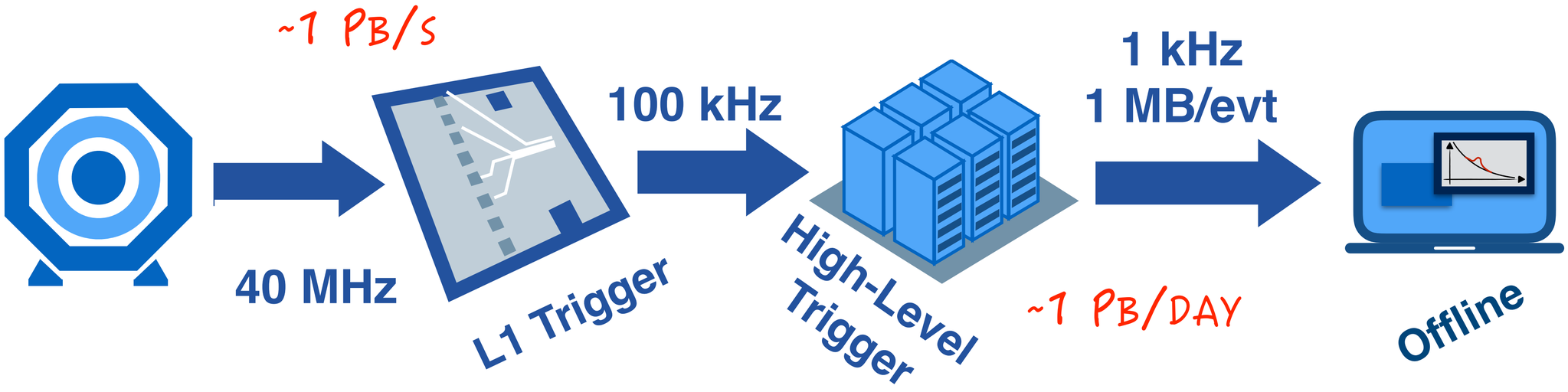


decorrelation:
teach the network how to *not* learn certain physical features; important for controlling systematic uncertainties



Low power, low latency, high radiation, real-time

Example: CMS processing chain



Resource constraints: massive data rates, on-detector low-latency processing

Extreme environments: low-power, cryogenic, high-radiation

AI In the Sky: Elements of Cosmic Experiments

Every main element of future cosmic experiments will be accelerated by AI.

Watching the sky

Data: Millions of sky pointings (≥2D)
Task: Optimize schedules over days and years
Task: Control the telescope and associated devices

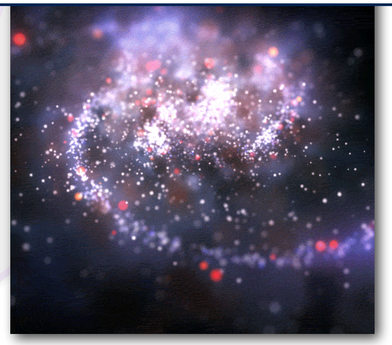
Simulating the Universe

Data: Trillions of particles in databases (≥6D)

Task: Recreate motions of particles under gravity, hydrodynamics

Task: Recreate observational noise elements in projected images of web

Analyzing Galaxies

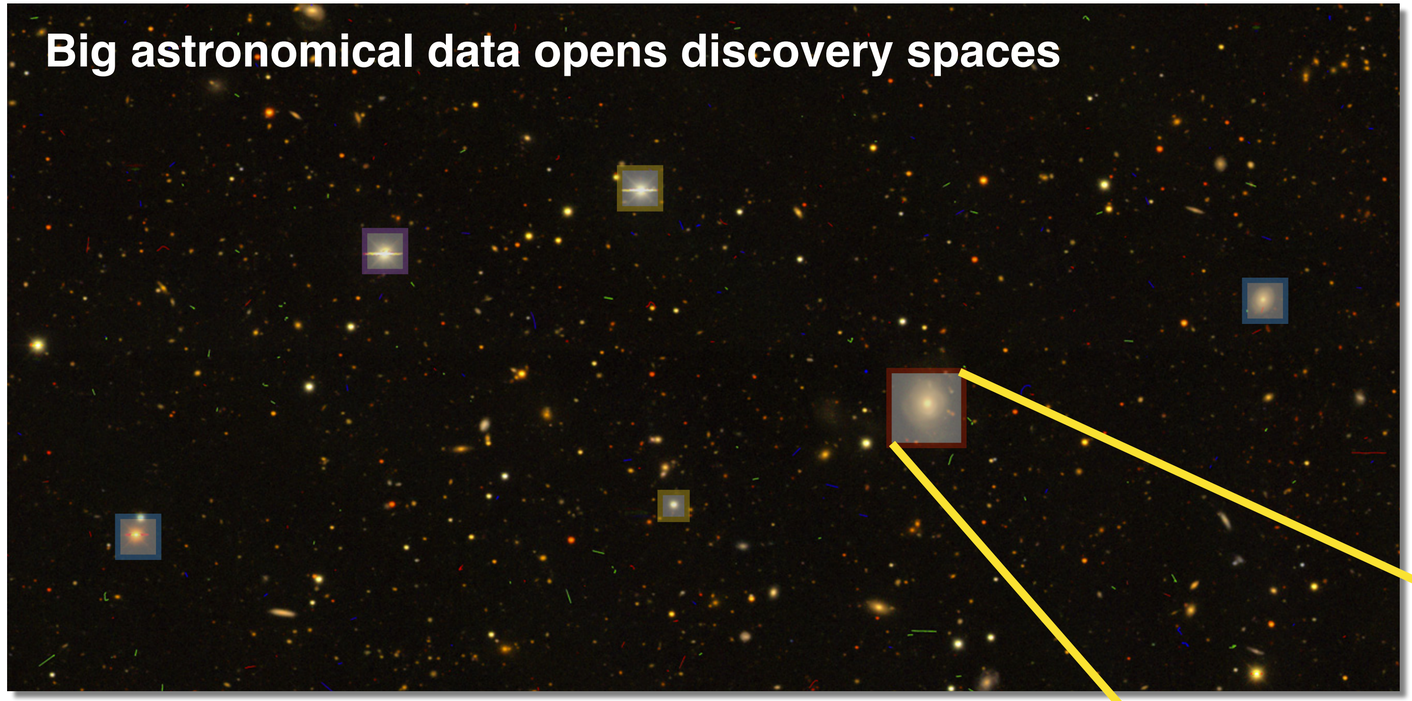


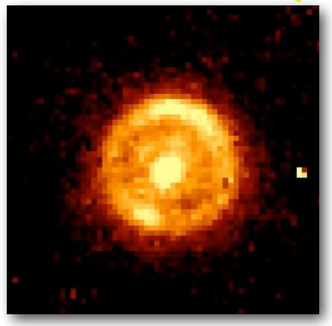
Data: Billions of objects in Images (2D) and Spectra (1D)

Task: Identify and measure objects in noisy data

Task: Separate data structures into components

“It’s full of stars”: find all the objects



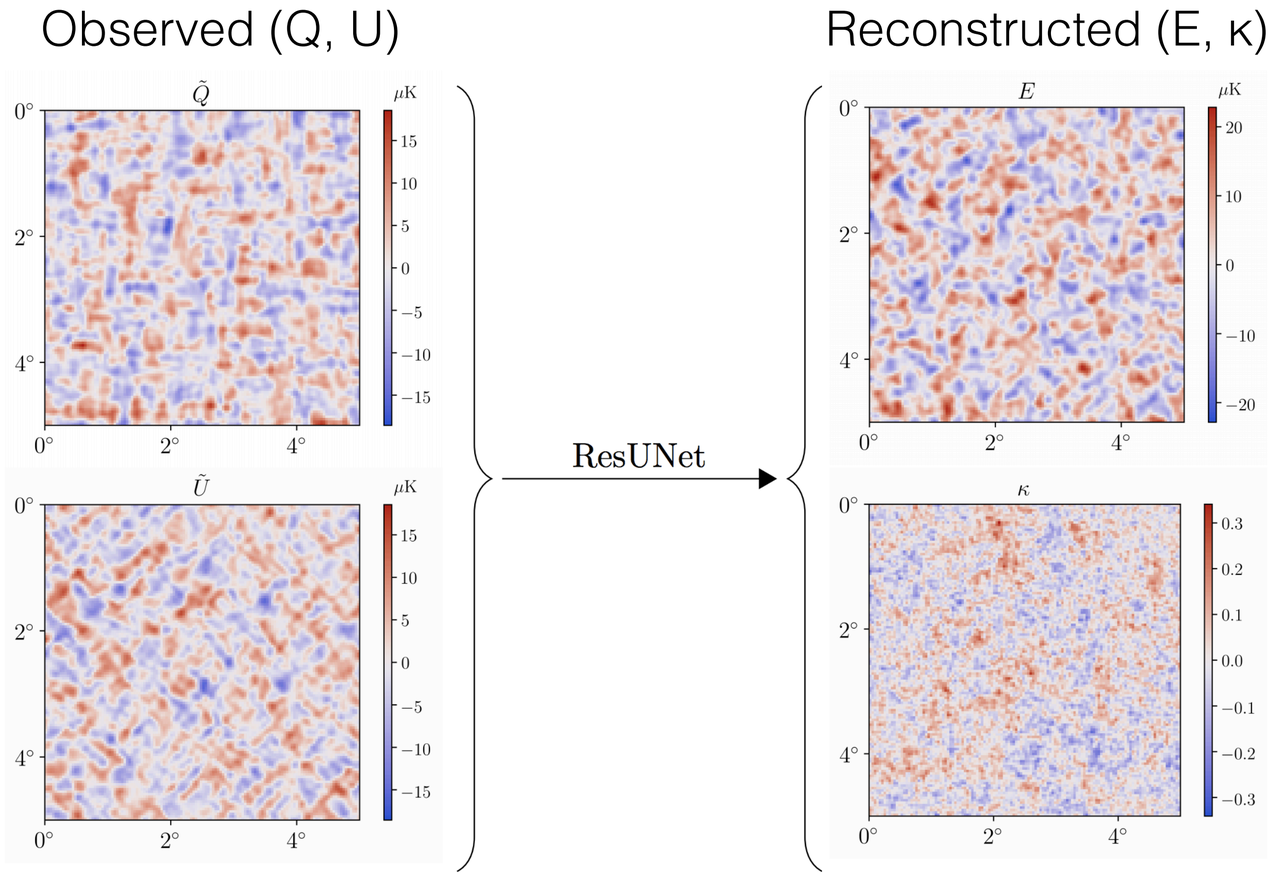


- Populations of objects show **dark matter, dark energy**
  Example: Strong Gravitational Lenses (right)
- Pursuing **Region-based CNNs** on heterogeneous compute devices for measurement of objects

- LSST:
  Optical Images, Chilean Andes
- 20 Tb / night
- 1 Billion transient alerts per night
- 10B galaxies, 10B stars over 10 years

Read between the layers: decompose microwave maps

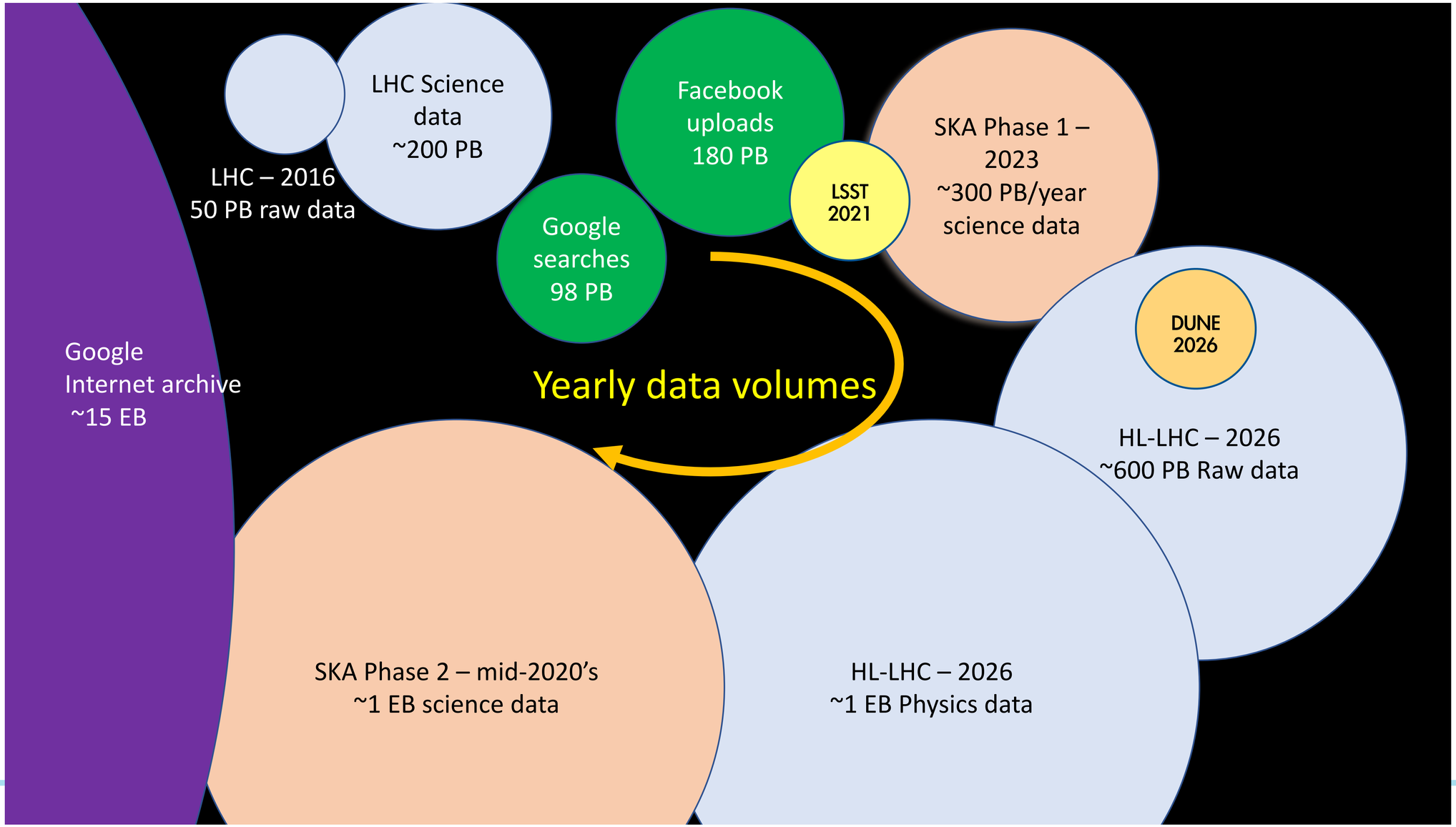
- South Pole Telescope (SPT): Polarized cosmic microwave background maps
- Earliest gravitational wave signatures that have very low signal
- Applicable for CMB-S4 next generation experiments



- Noise and other foregrounds obfuscate primordial GW signatures
- Pioneered use of **Residual UNets** to separate lensing signals (κ) from CMB polarization map (E)

Caldeira, Nord, et al., https://arxiv.org/abs/1810.01483

Big datasets



HEP and AI

- **AI techniques accelerate discovery science**
  - Applications across the grandest challenges in physics

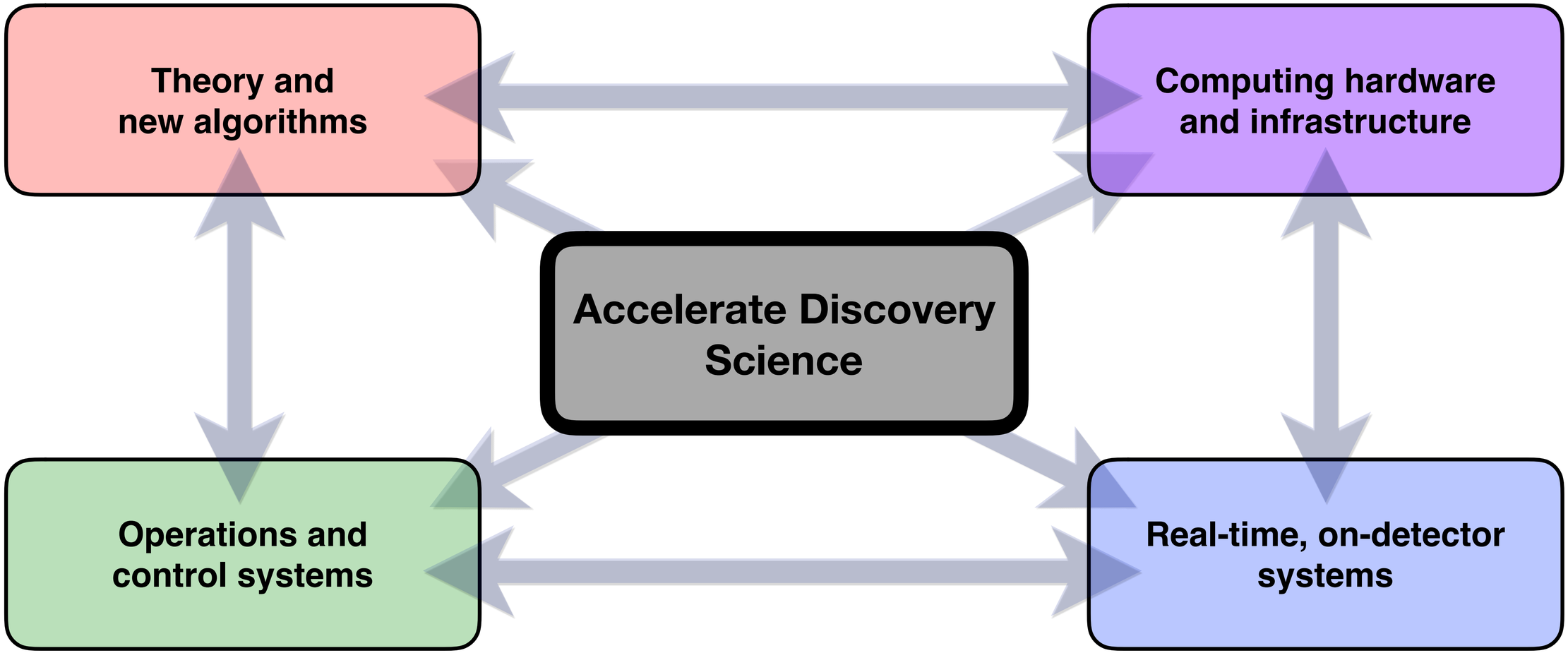
- **Key HEP technologies**
  - Requirements for real science including high-fidelity simulation, uncertainty quantification, new learned representations
  - Advanced sensors and real-time systems
  - Controls and operation of large, integrated detectors and accelerators
  - Large datasets for training and processing

Outline

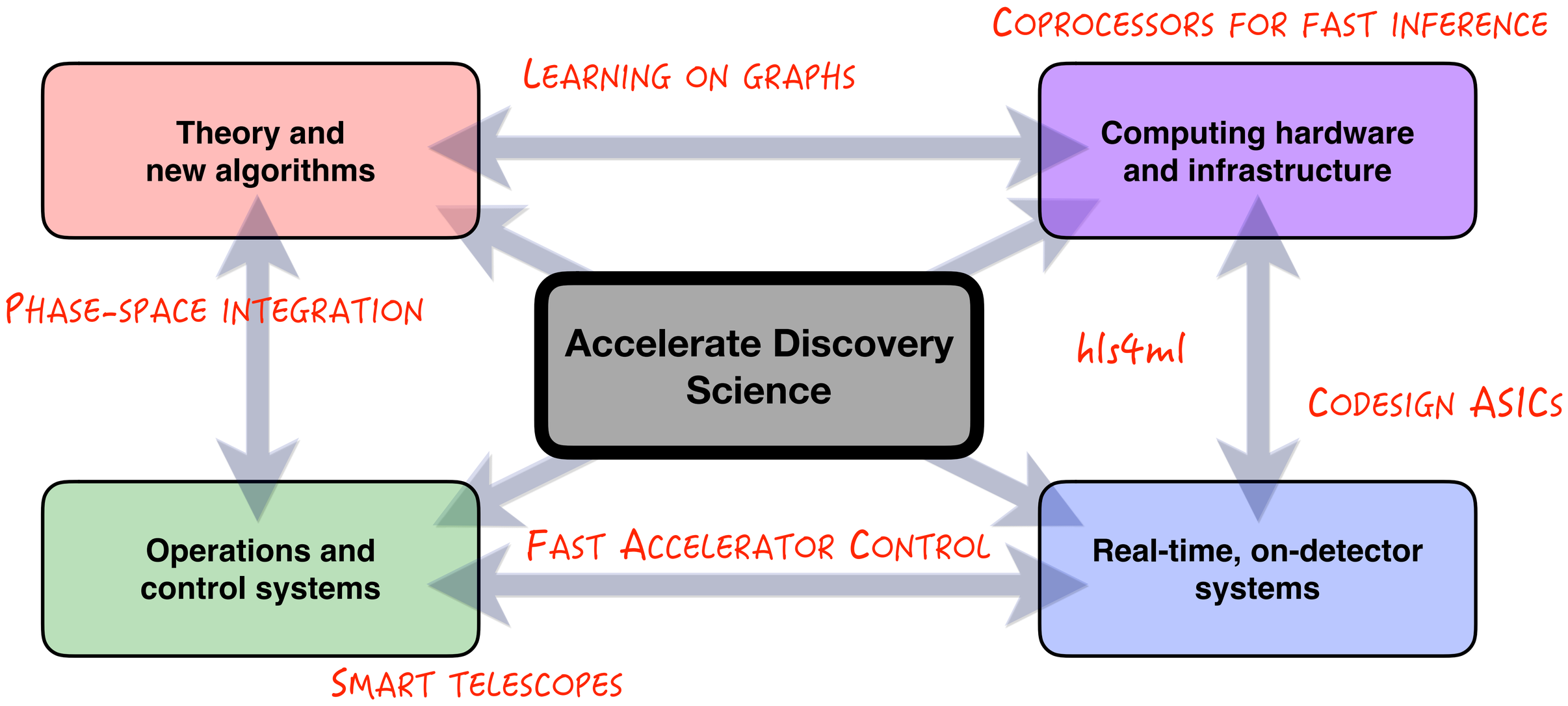
Fermilab through the lens of AI
scientific mission
enabling technologies at the bleeding edge

Capabilities and opportunities
AI development for HEP
opportunities to share and collaborate

Fermilab AI Capabilities



Fermilab AI Capabilities

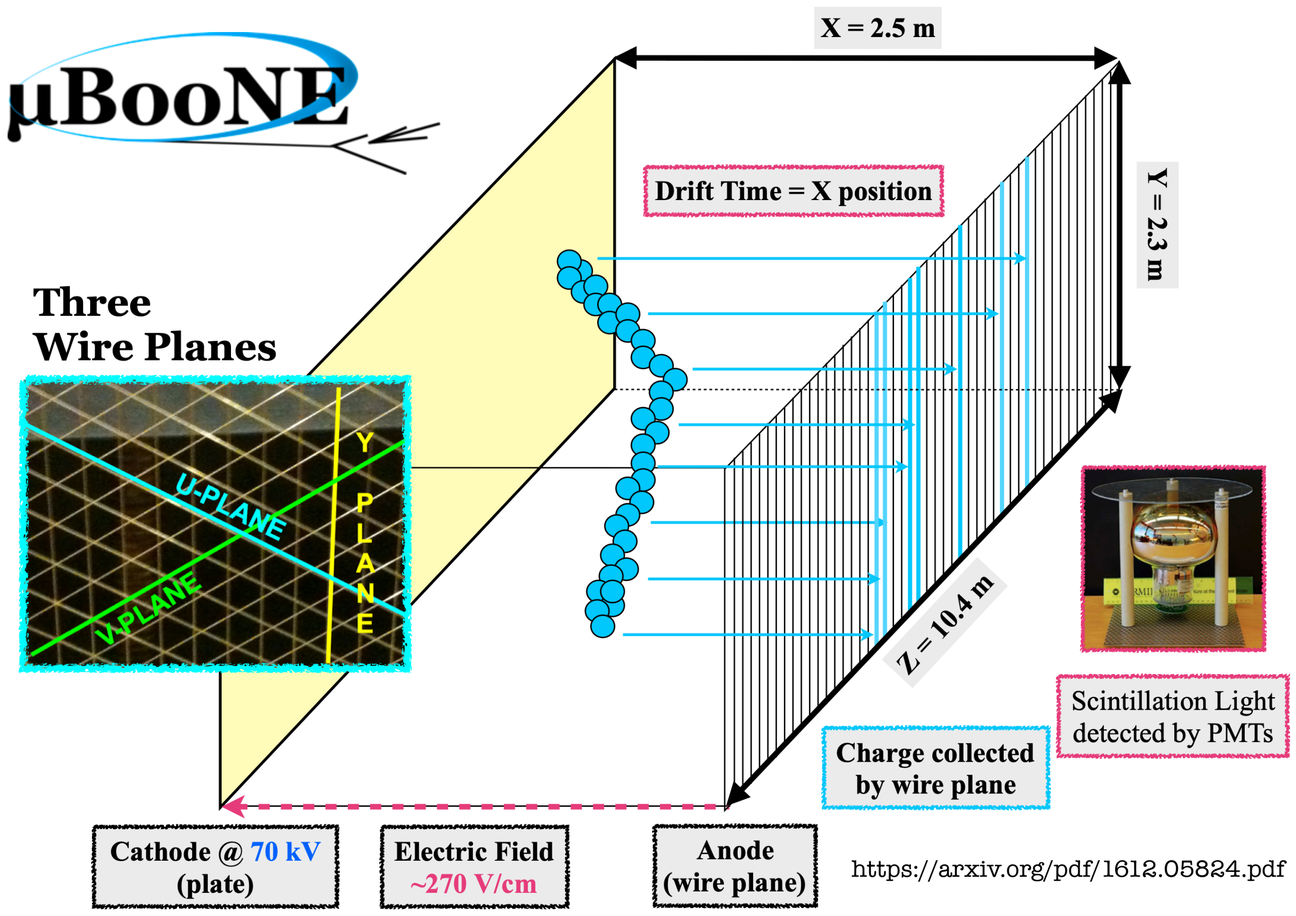


Beyond images

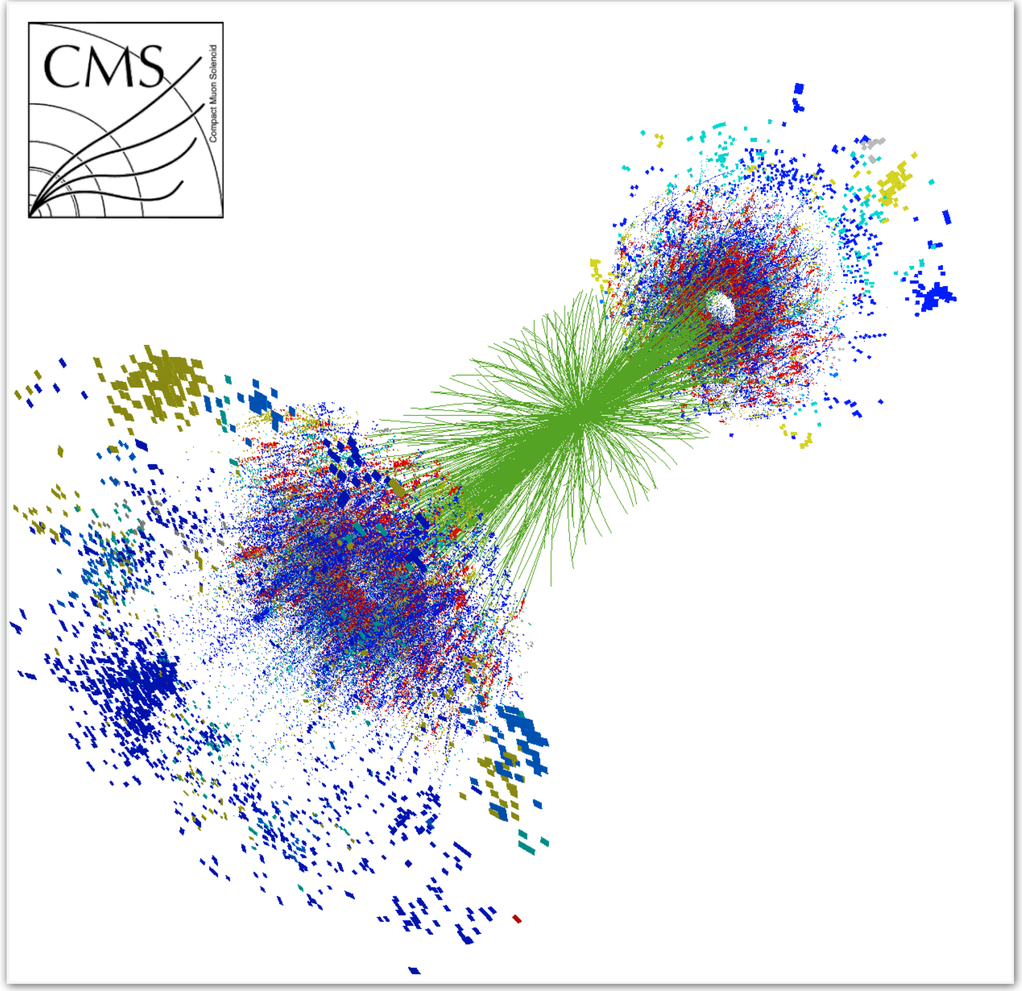
Theory and new algorithms

Sparse, multi-modal, high-dimensional

$\vec{v} = \{x,y,z,E,t,id,\delta,\ldots\}$



https://arxiv.org/pdf/1612.05824.pdf

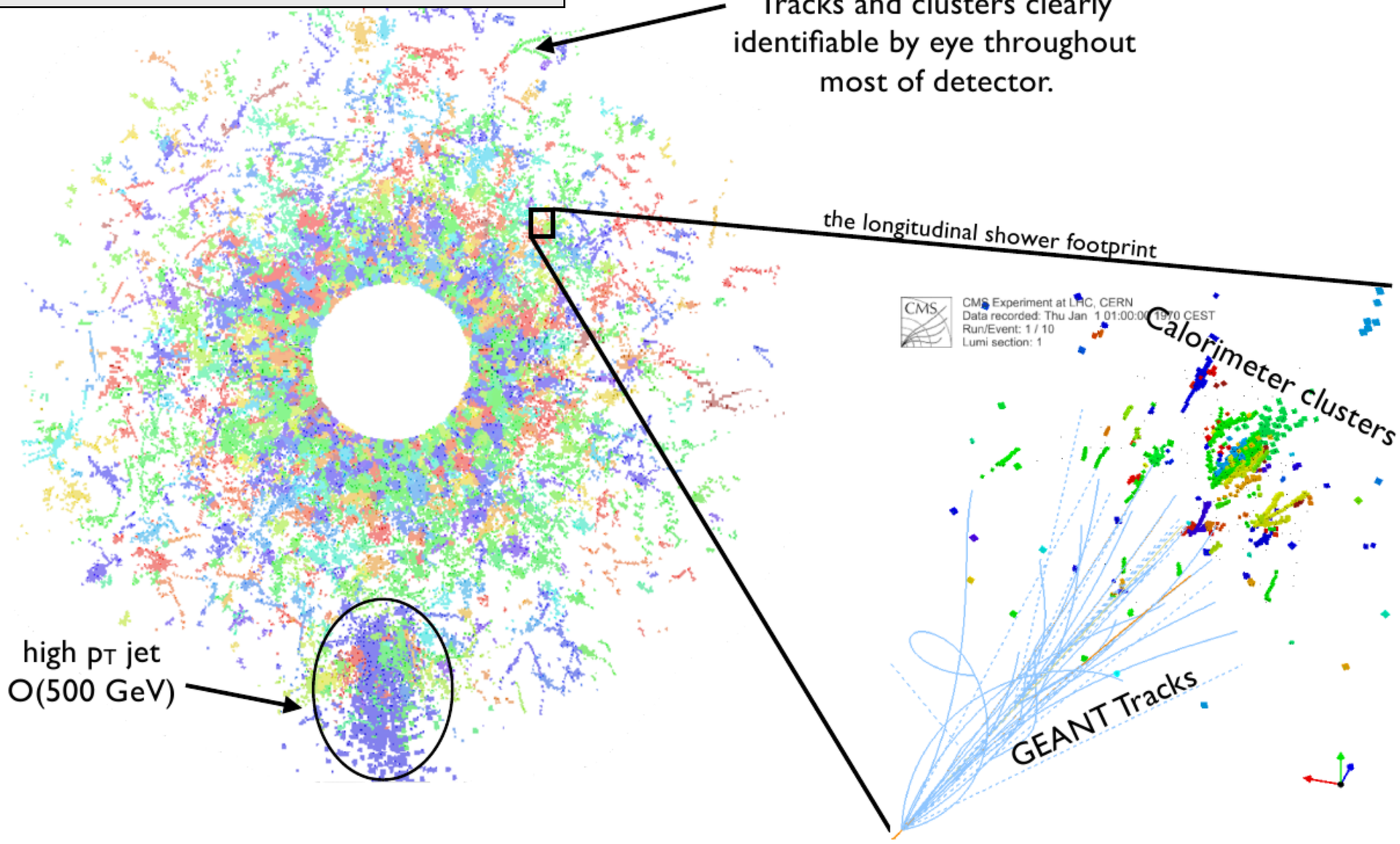


How to process 40 million of these per second?

Theory and new algorithms

Computing hardware and infrastructure

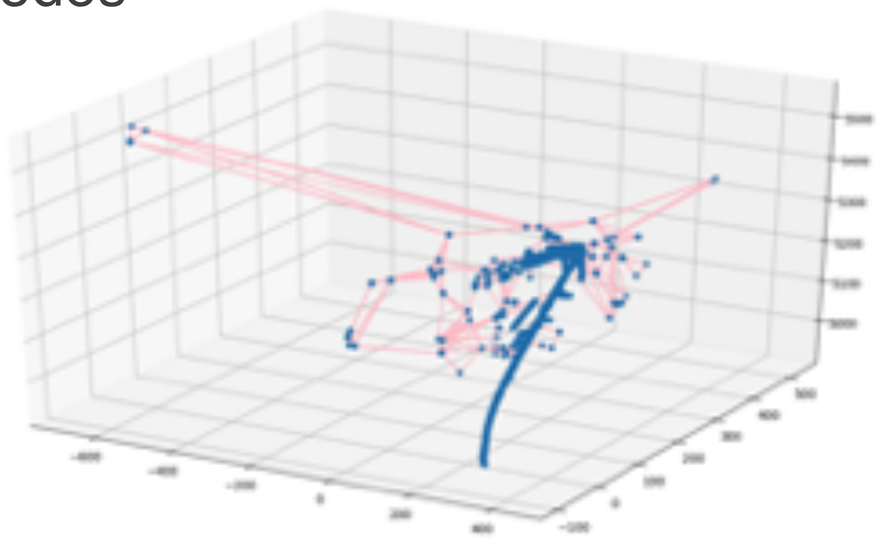
Real-time, on-detector systems



Beyond images

Theory and new algorithms

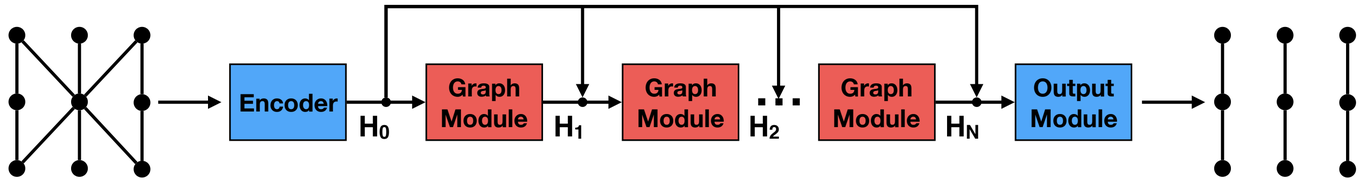
- Multiple activities into **learning new representations of detector data** for different physics applications
- Explore neural network architectures based on **point clouds and graphs**; n-dimensional inputs in non-Euclidean space
- Promising first results for multiple applications
  - Learn the strength of connections (edges) between nodes
  - Charged particle tracking [1]
  - Calorimetry for irregular geometries [2]



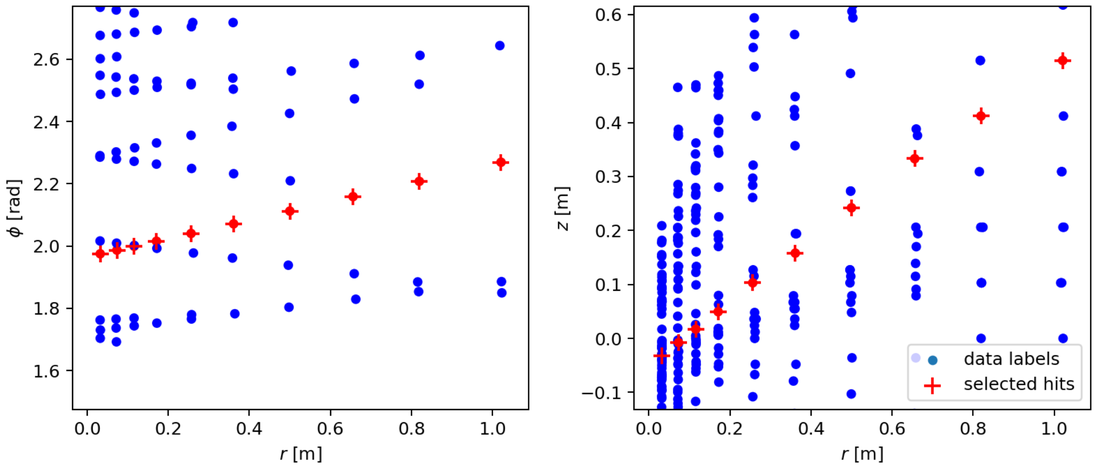
[1] https://arxiv.org/abs/1810.06111 , [2] https://arxiv.org/abs/1902.07987

Beyond images

Theory and new algorithms



Tracking

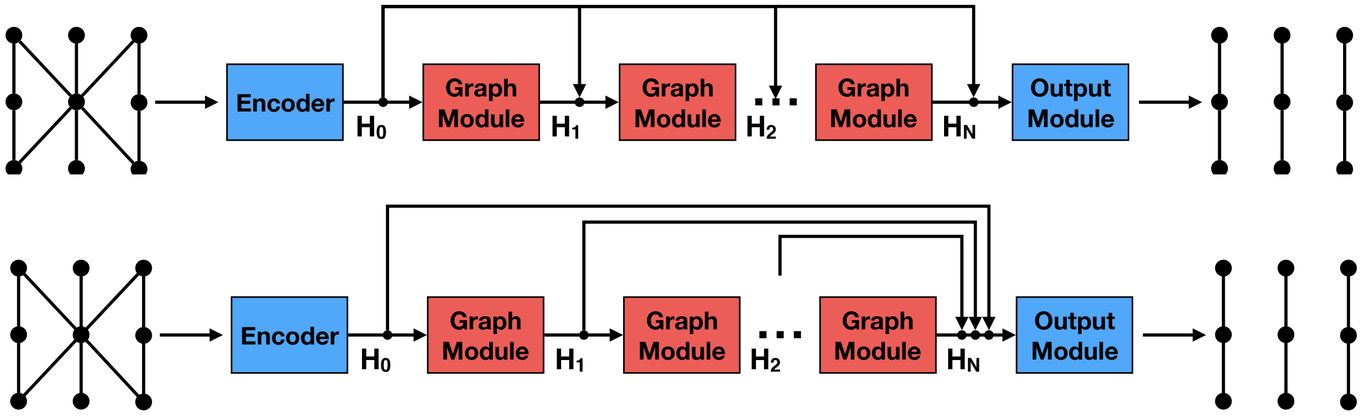


[https://heptrkx.github.io/], [https://github.com/exatrkx]

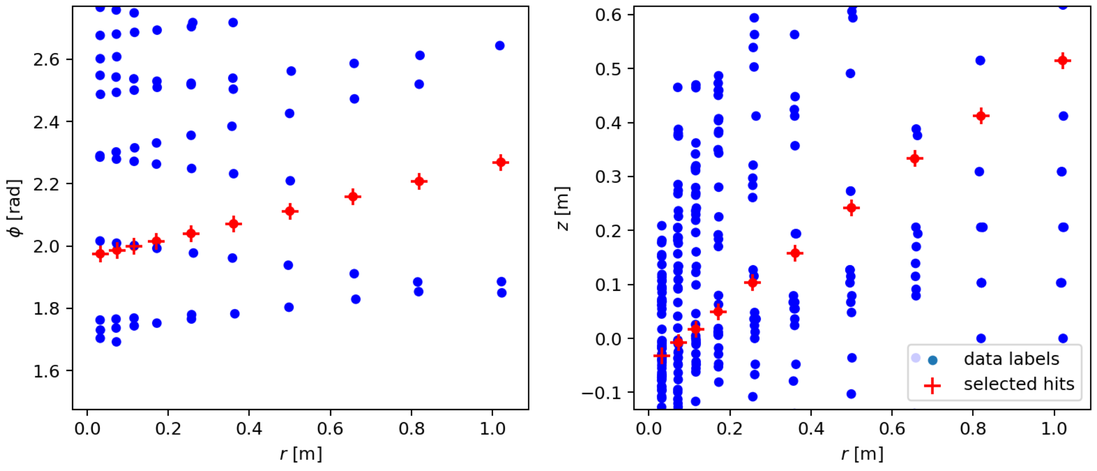
Beyond images

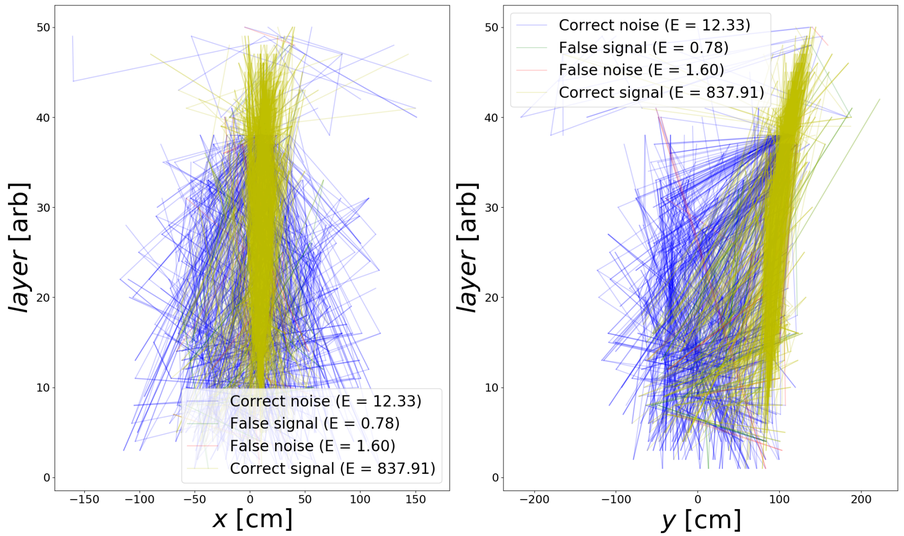
Theory and new algorithms

Tracking



Clustering



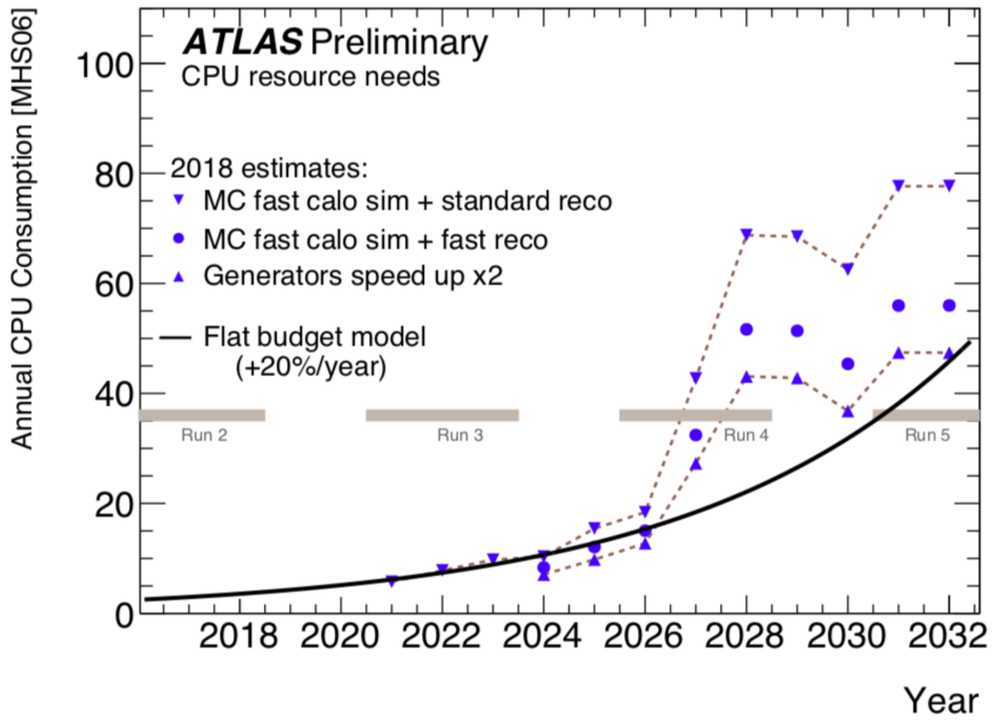


[https://heptrkx.github.io/], [https://github.com/exatrkx]

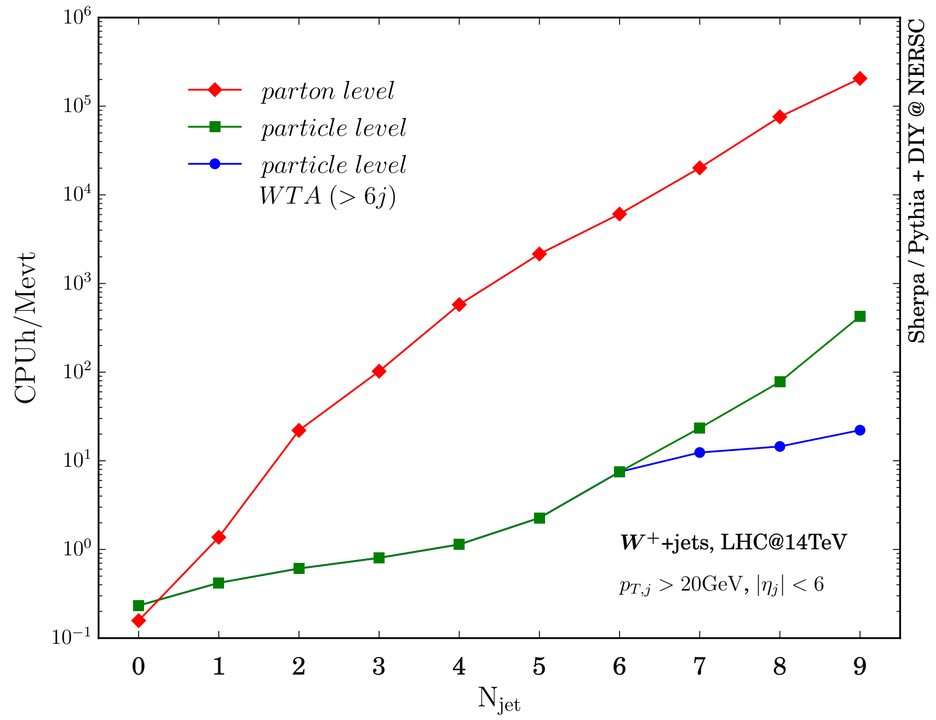
Teaching a computer to integrate

Theory and new algorithms

- Simulations are vital piece of physics experiments — encode our knowledge
- Needs for both more efficient and more accurate simulations



Growing computing needs



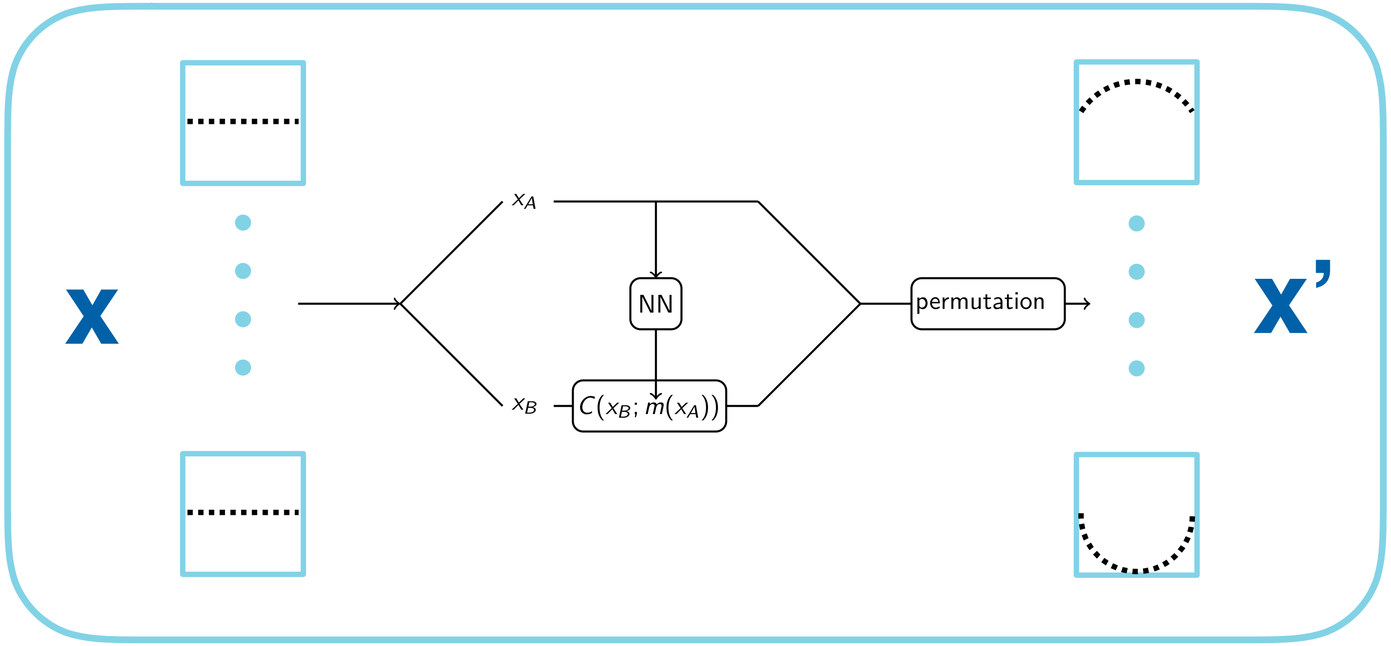
Compute scales with accuracy

[https://indico.cern.ch/event/849141/]

Teaching a computer to integrate

Theory and new algorithms

- **Importance sampling**: to produce simulated events efficiently — learn the *correlated, high-dimensional* space
  - Simulating N particles in a (3N - 4) dimensional space



Neural Importance Sampling

A normalizing flow models a complex probability density as an invertible transformation of a simple base density.

https://arxiv.org/pdf/1505.05770.pdf
https://arxiv.org/pdf/1808.03856.pdf
https://arxiv.org/abs/1906.04032

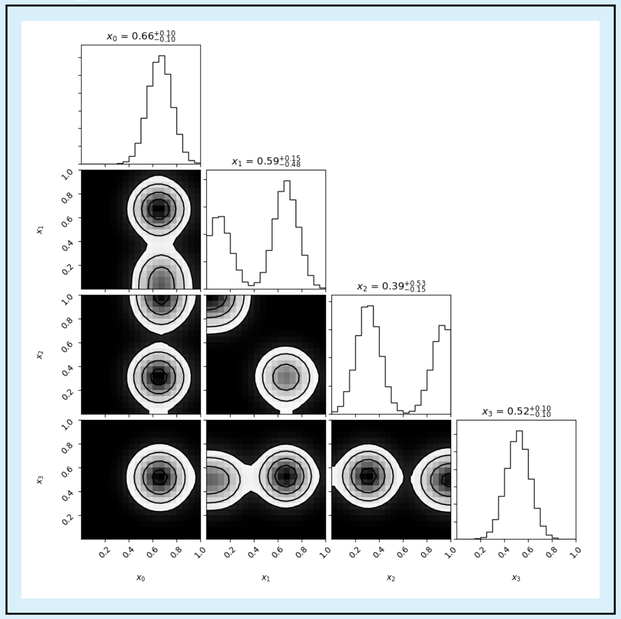
[https://indico.cern.ch/event/849141/]

Teaching a computer to integrate

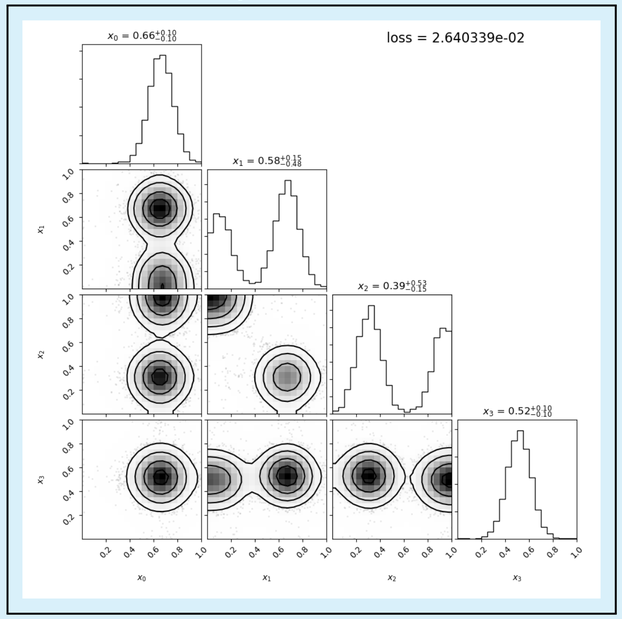
Theory and new algorithms

- **Importance sampling**: to produce simulated events efficiently — learn the *correlated, high-dimensional* space
  - Simulating N particles in a (3N - 4) dimensional space

Target Distribution:



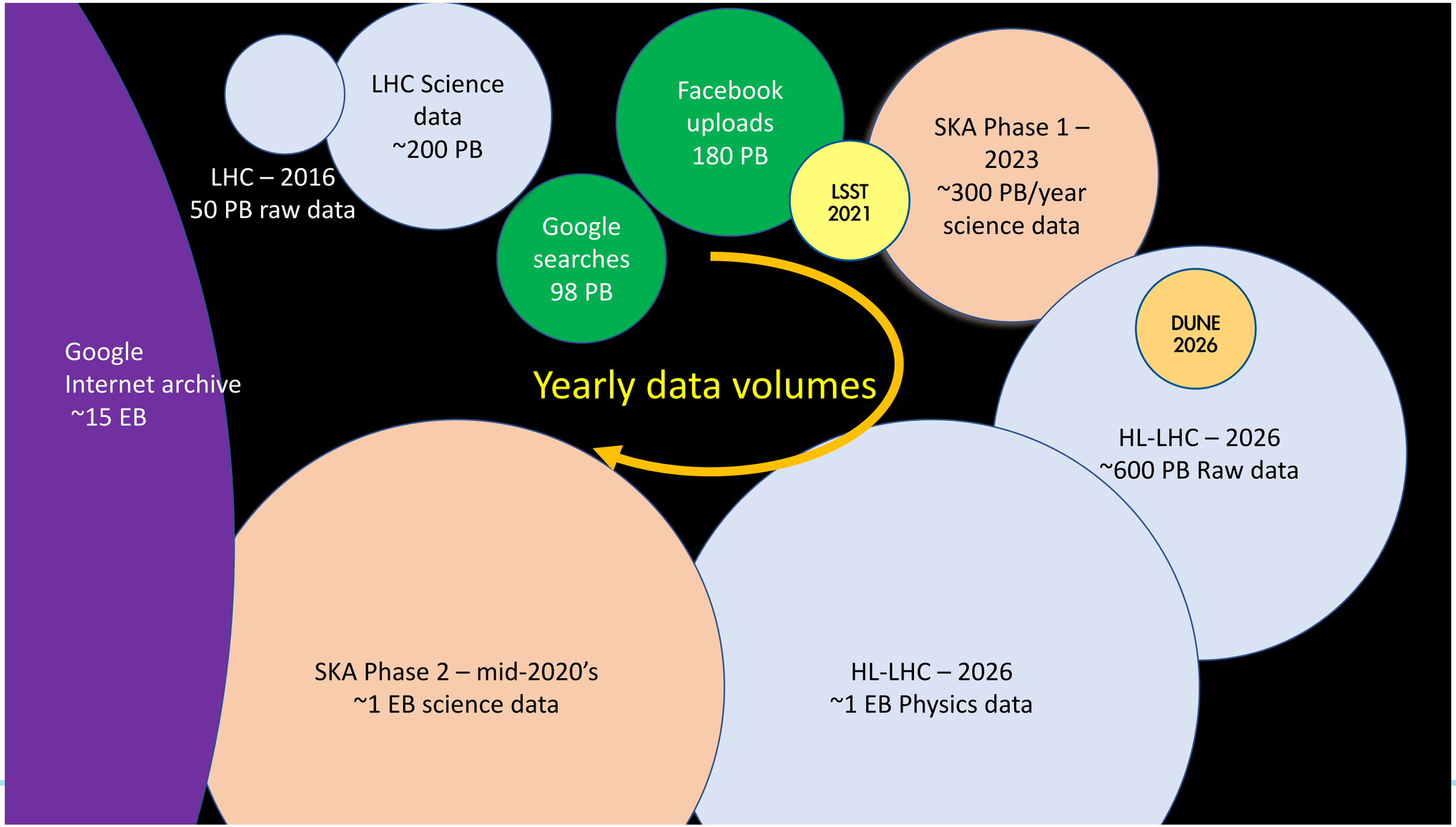
After 200 epochs:



First study shows **6x speed-up** for simple $e^+e^- \rightarrow$ 3-jet final state topology

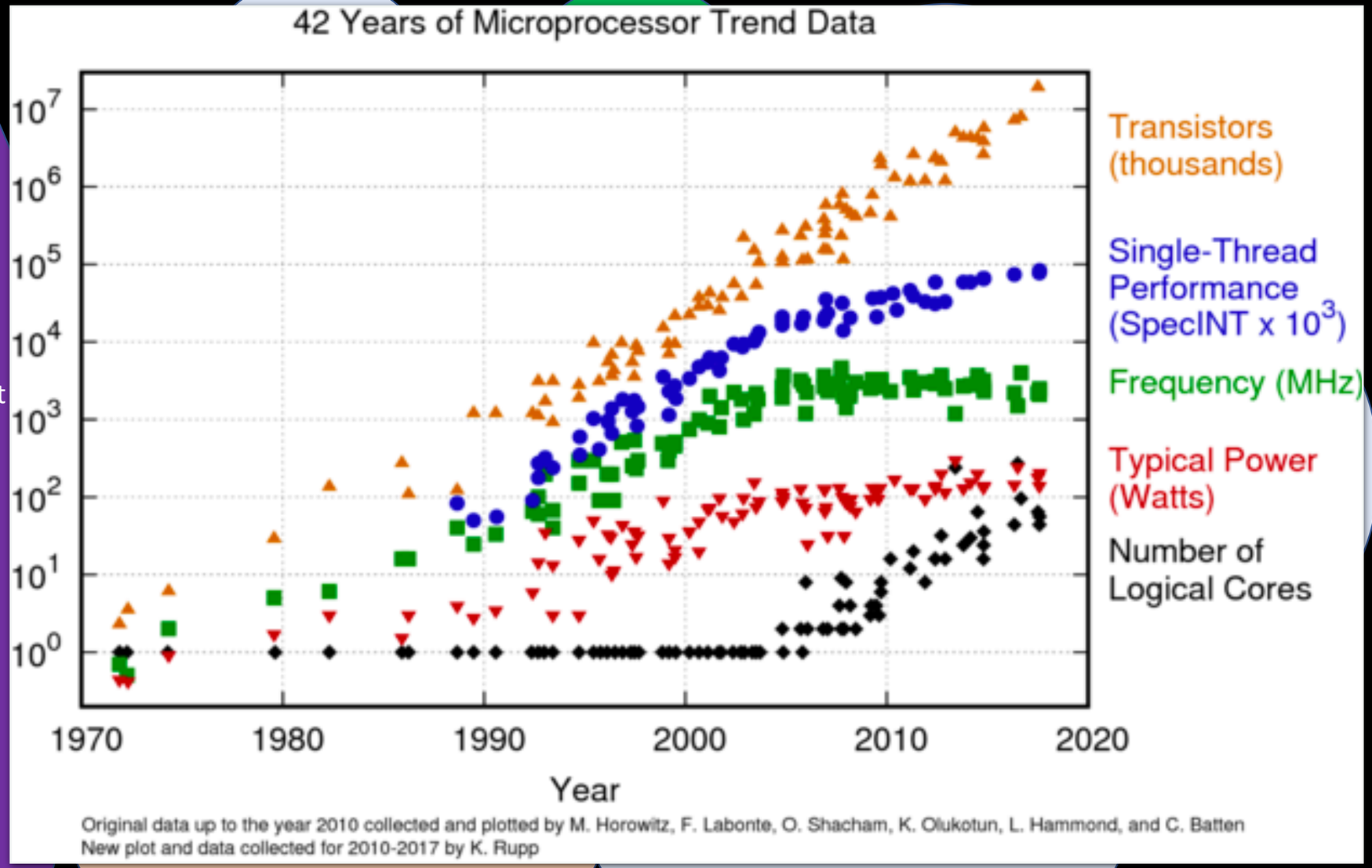
[https://indico.cern.ch/event/849141/]

Big datasets



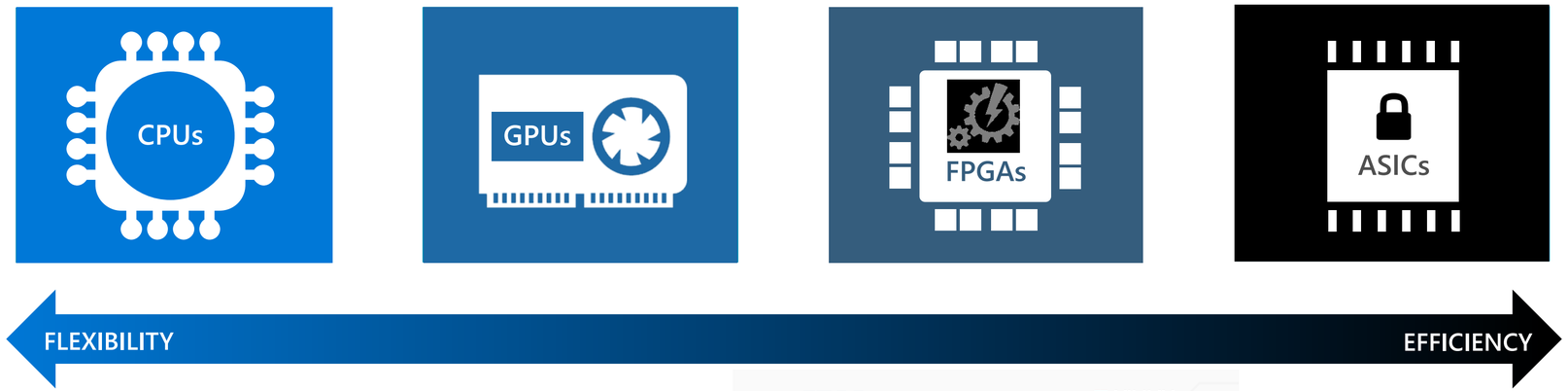
Big datasets

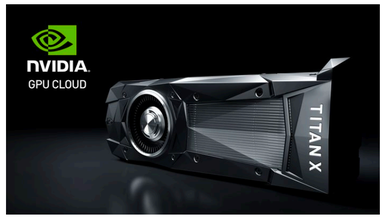
Google
Internet
~15 EB

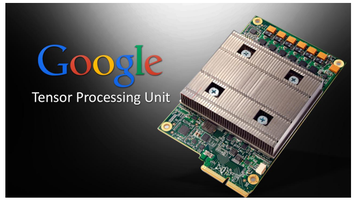


Computing infrastructure and hardware

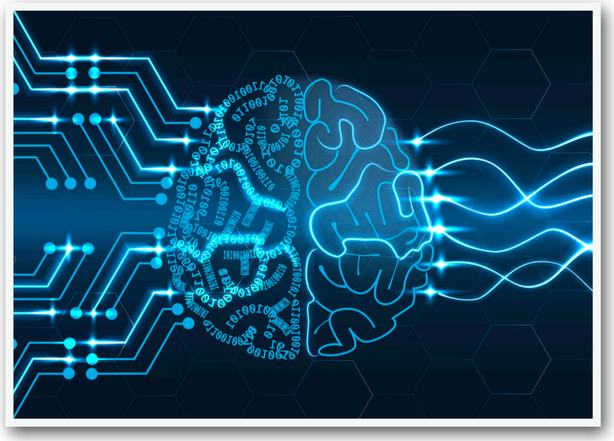
Computing hardware and infrastructure







Advances in heterogeneous computing driven by **machine learning and big data explosion**



Complex and massive datasets

Computing hardware and infrastructure

- Big science requires *both* high-performance and high-throughput compute
  - Translation: accelerated computing technologies for **training** and **inference**

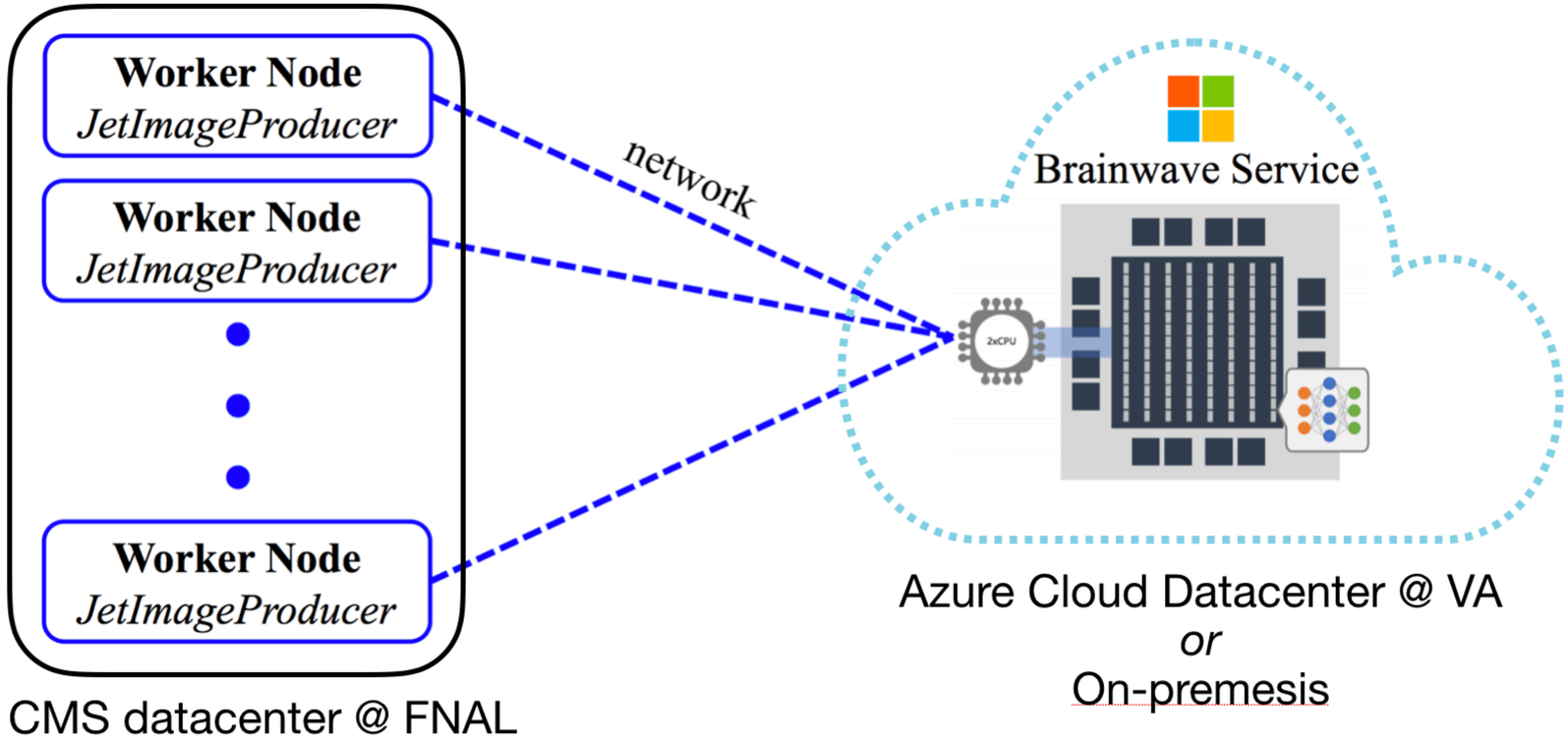
- Proof-of-concept study for high-throughput
  - In collaboration with Microsoft and many university partners, FPGA acceleration of machine learning inference in the cloud and the edge
  - https://arxiv.org/abs/1904.08986

story

Complex and massive datasets

Computing hardware and infrastructure

- Study found **30x (175x)** speed-up for cloud (edge) inference of ResNet50 over experimental software framework



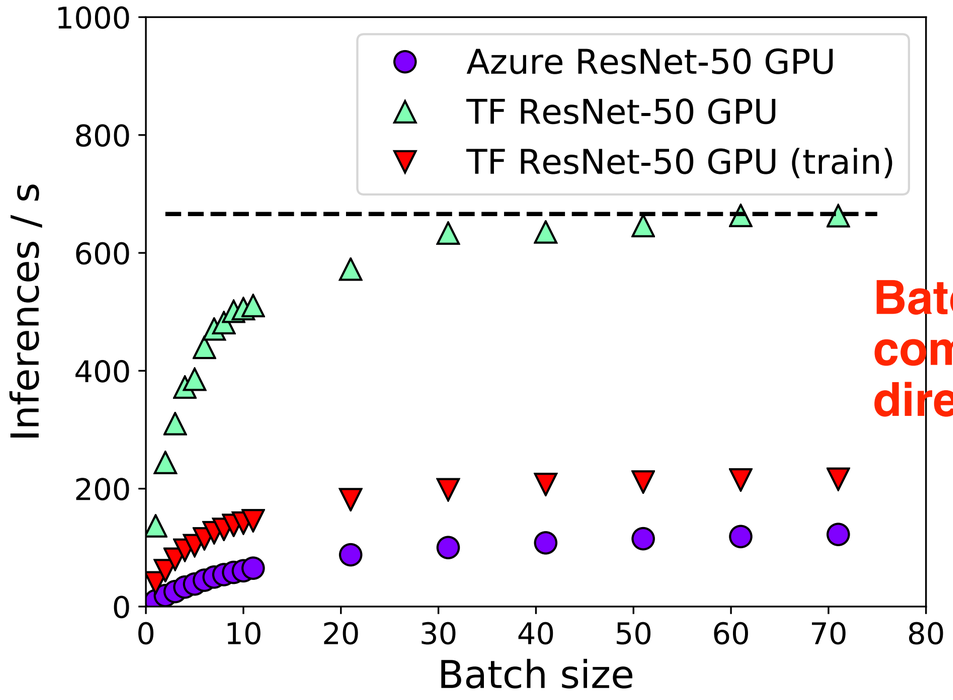
Non-disruptive integration of heterogenous computing resources into HEP computing model
Deploy as a service (many CPUs to few FPGAs) can be **more cost-effective**

Exploring applications across many experiments (LHC, neutrinos, cosmology, gravitational waves)

Complex and massive datasets

Computing hardware and infrastructure

- Study found **30x (175x)** speed-up for cloud (edge) inference of ResNet50 over experimental software framework



Batch-of-1 FPGAaaS comparable (or better) to directly connected GPU!

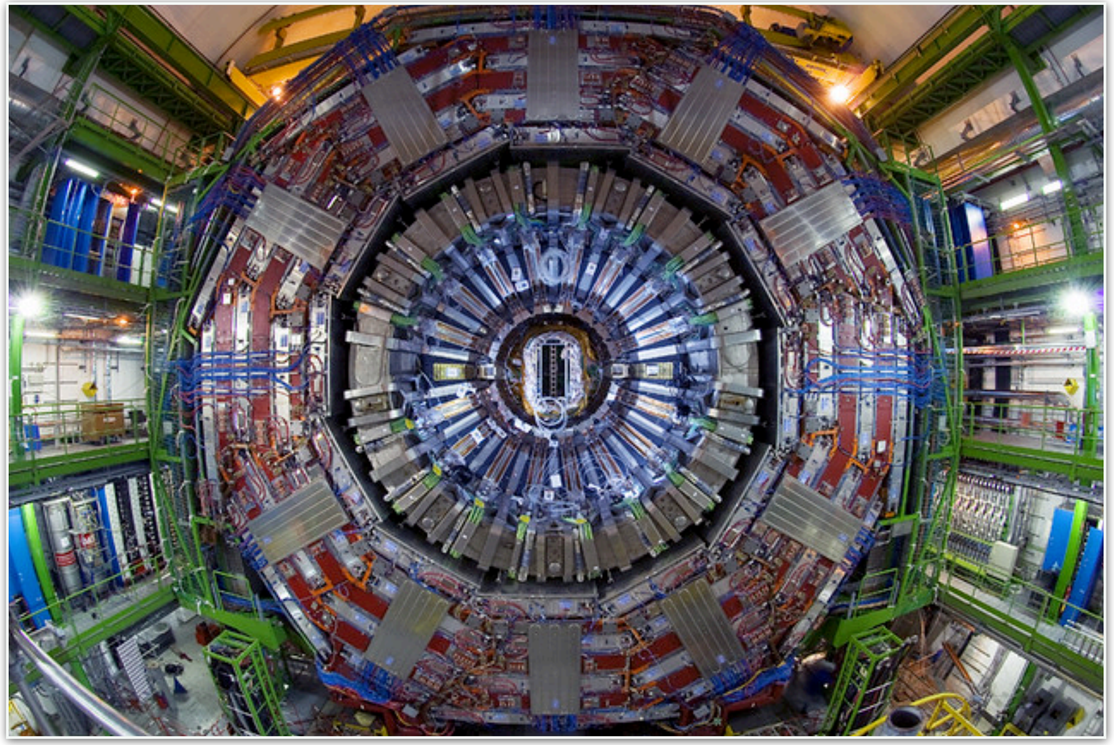
Non-disruptive integration of heterogenous computing resources into HEP computing model
Deploy as a service (many CPUs to few FPGAs) can be **more cost-effective**

Exploring applications to other experiments (neutrinos, cosmology, gravitational waves), and exploring accelerated training

Real-time, on-detector systems

Real-time, on-detector systems

- Resource-constrained AI
- Low-latency, low-power, high bandwidth
- Cryogenics, high-radiation



Real-time, on-detector systems

Real-time, on-detector systems

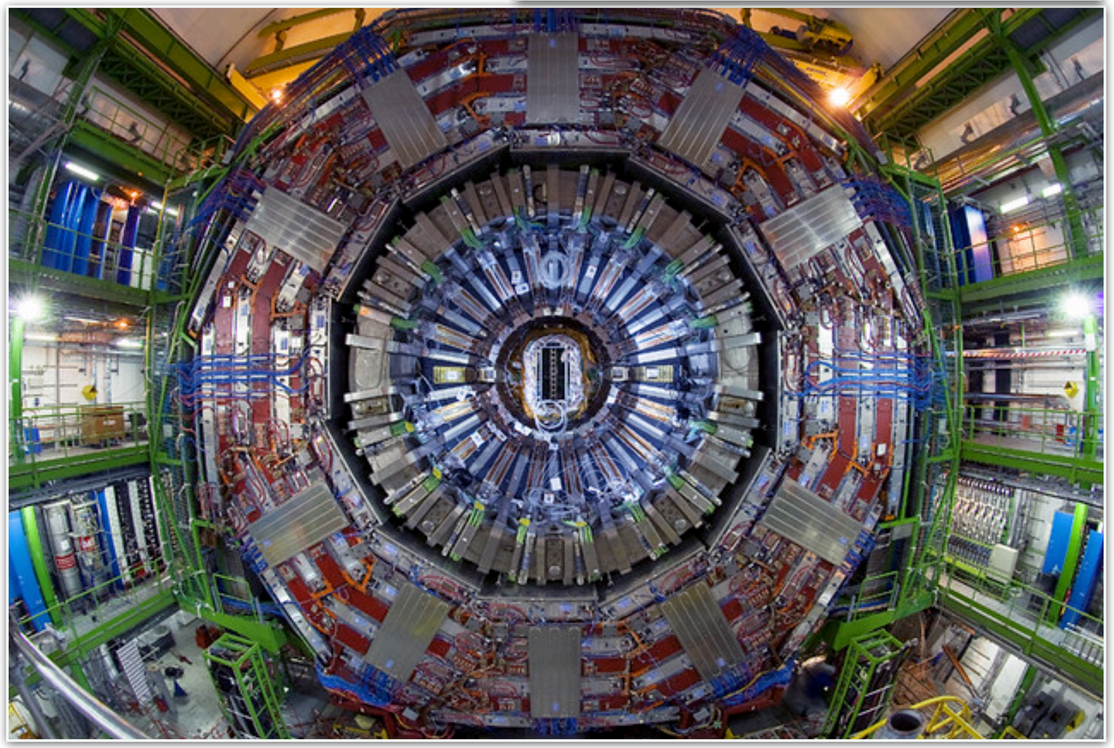
- Resource-constrained AI
- Low-latency, low-power, high bandwidth
- Cryogenics, high-radiation

LHC at CERN

40 MHz collision rate, ~20 hrs/day

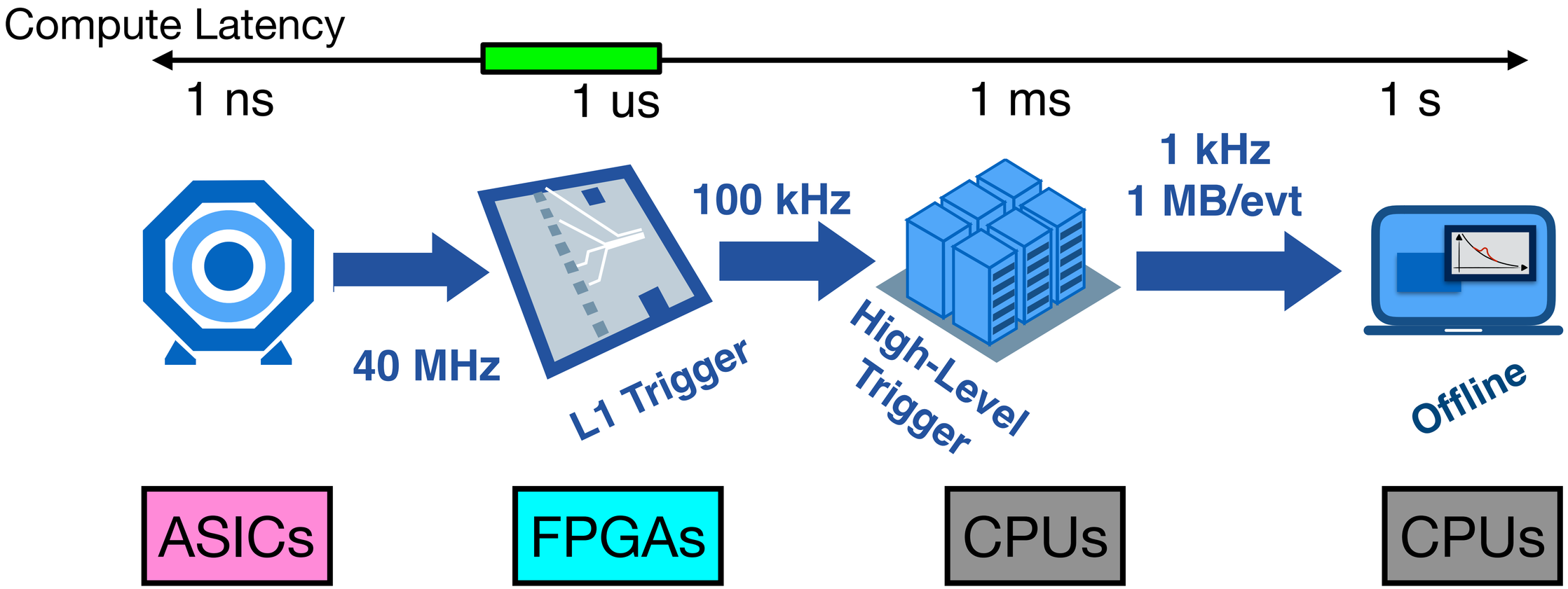
Compact Muon Solenoid (CMS)

→ > 1 billion channels

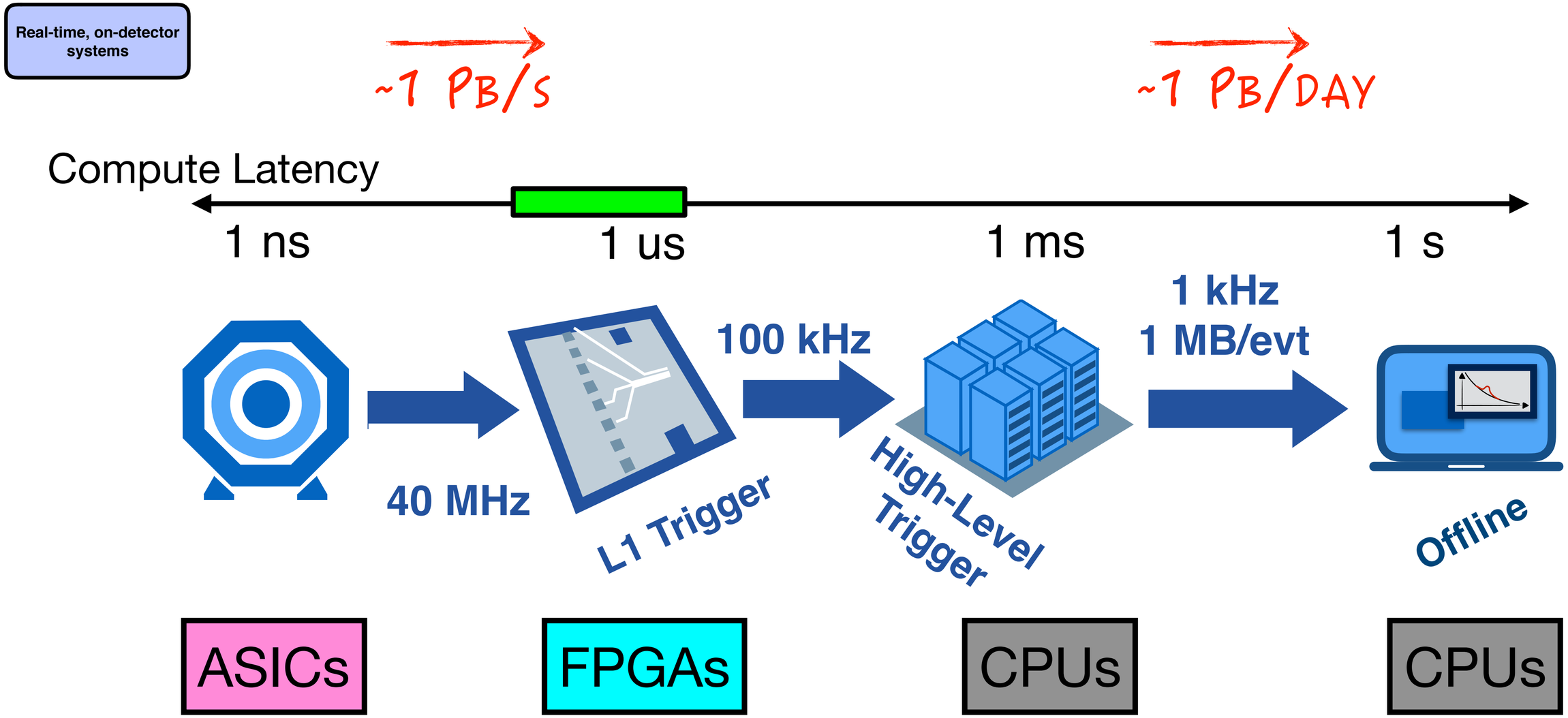


CMS data processing chain

Real-time, on-detector systems



CMS data processing chain



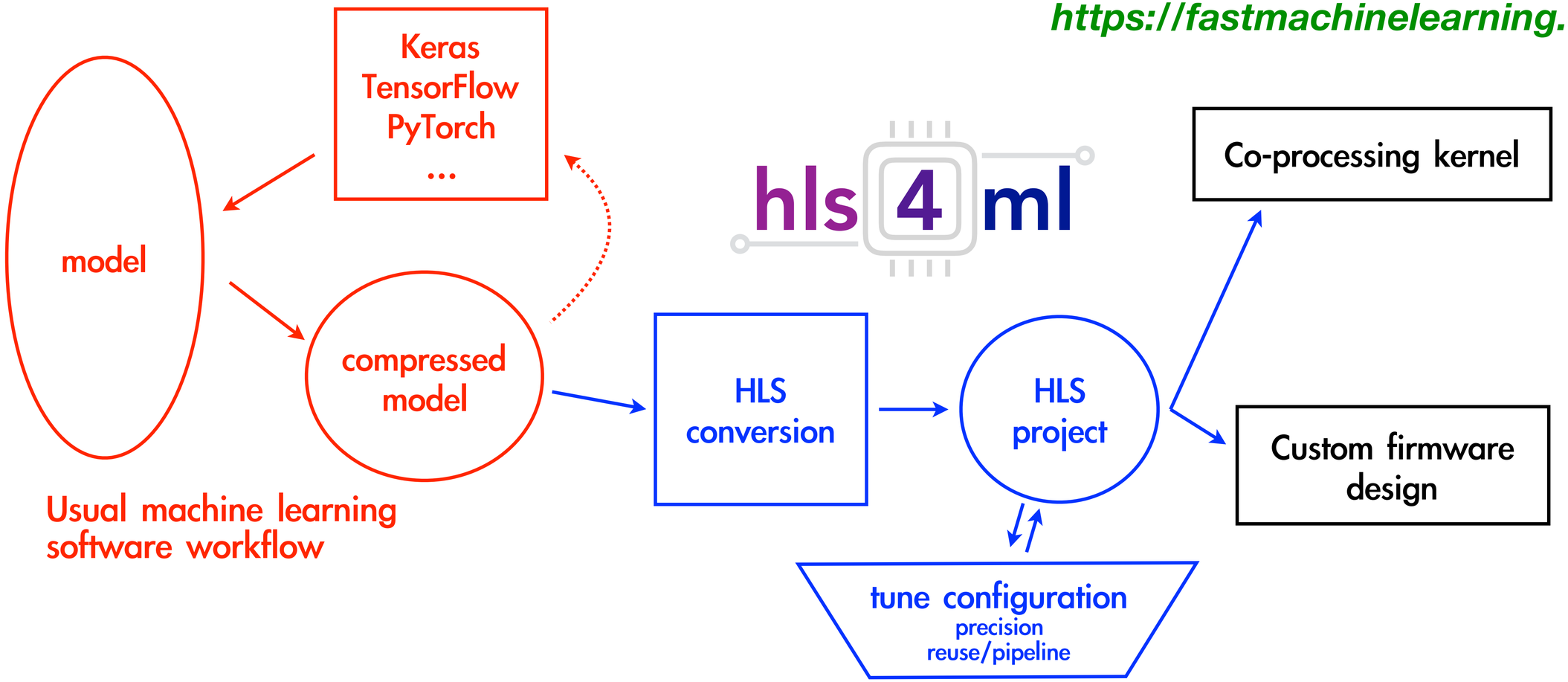
REAL-TIME ON-DETECTOR AI

ML in the hardware trigger

Real-time, on-detector systems

- **hls4ml** — open-source automated translation tool, ML models to firmware

https://fastmachinelearning.org/hls4ml



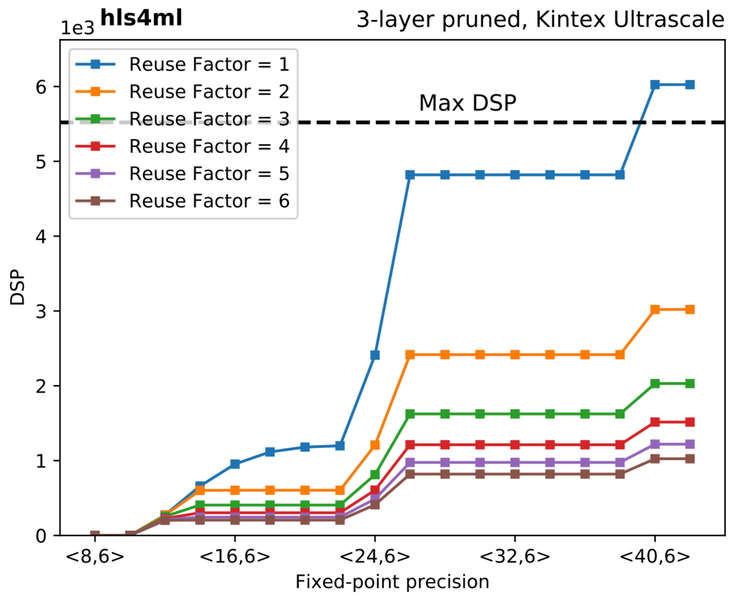
featured Xilinx case study!

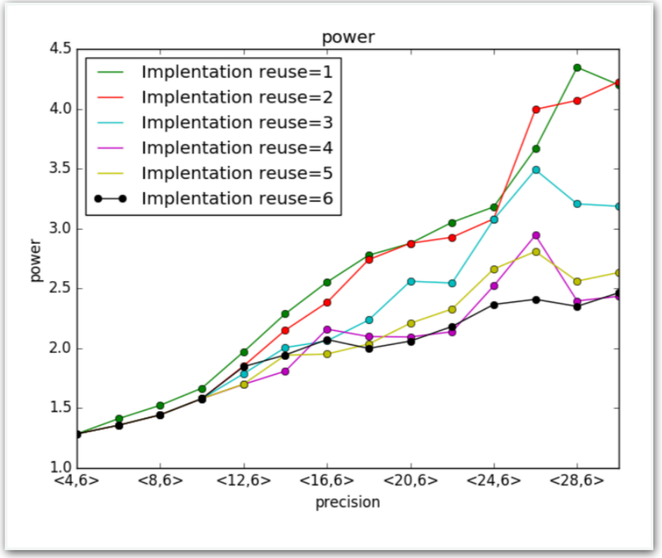
ML in the hardware trigger

Real-time, on-detector systems

[https://arxiv.org/abs/1804.06913]

- All FPGA design
  - Flexible: many algorithm types for layers of processing
- Application and adoption growing across the LHC
  - Firmware in hours instead of weeks/months
- **Growing interest with many on-going developments**
  - CNNs, Graphs, RNNs, auto-encoders, binary/ternary
  - Alternate HLS (Intel, Mentor, Cadence)
  - Co-processors, multi-FPGA
  - Intelligent ASICs





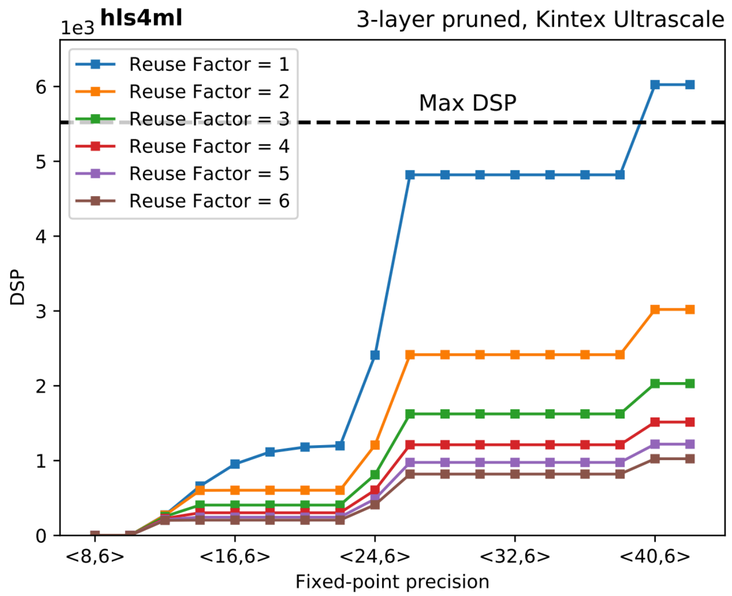
> 5000 parameter fully connected network in 100 ns

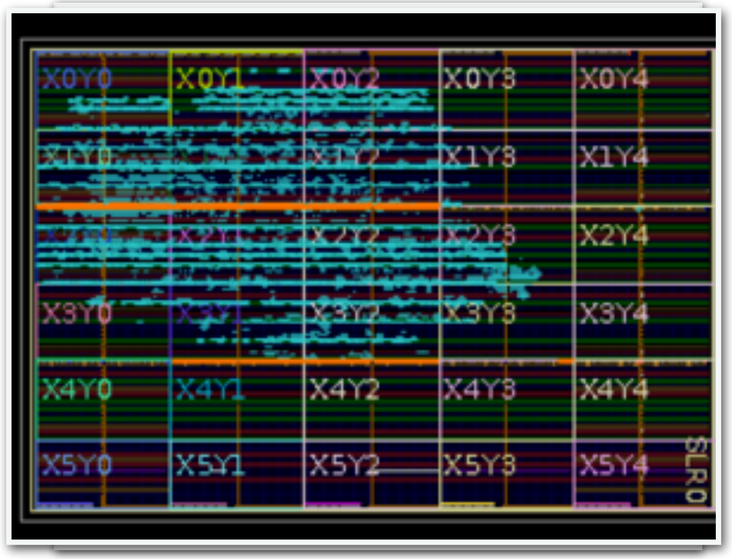
ML in the hardware trigger

Real-time, on-detector systems

[https://arxiv.org/abs/1804.06913]

- All FPGA design
  - Flexible: many algorithm types for layers of processing
- Application and adoption growing across the LHC
  - Firmware in hours instead of weeks/months
- **Growing interest with many on-going developments**
  - CNNs, Graphs, RNNs, auto-encoders, binary/ternary
  - Alternate HLS (Intel, Mentor, Cadence)
  - Co-processors, multi-FPGA
  - Intelligent ASICs





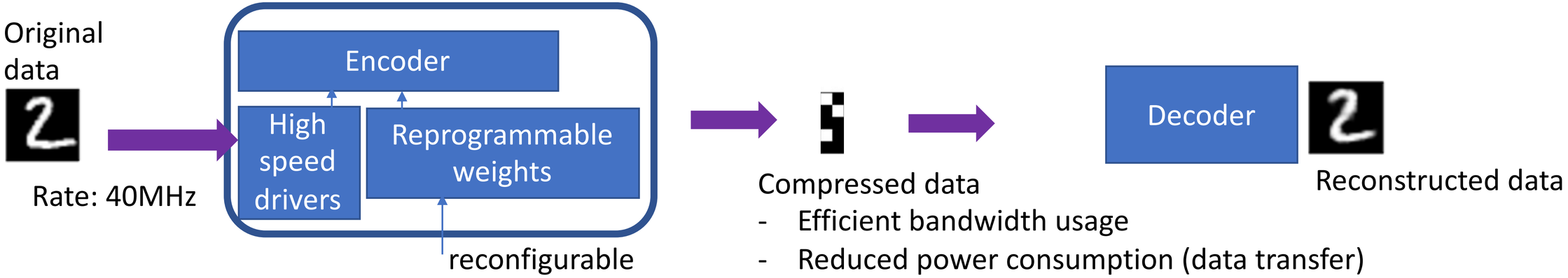
> 5000 parameter fully connected network in 100 ns

hls4ml…4asic?

Real-time, on-detector systems

Hardware acceleration with an emphasis on co-design and fast turnaround time

First project: Autoencoder with MNIST benchmark (28 x 28 x 8-bits @ 40 MHz)



Enable edge compute : e.g. data compression

Programmable and Reconfigurable: reprogrammable weights

Hardware – Software codesign: algorithm-driven architectural approach

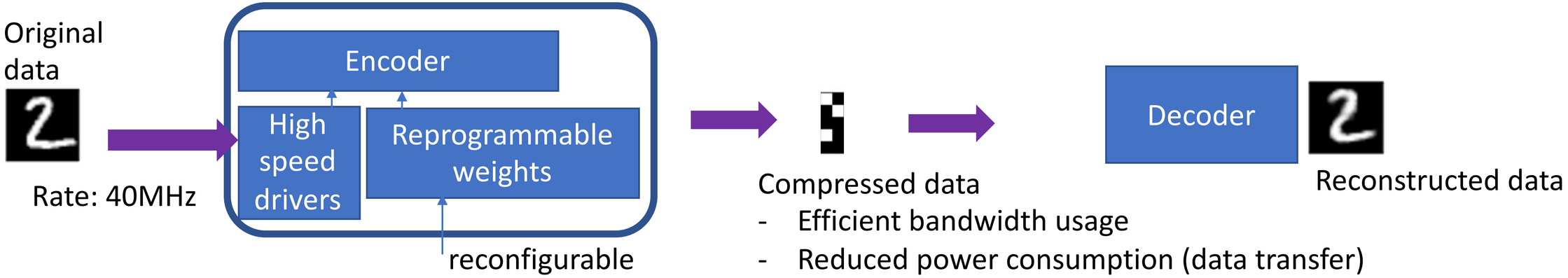
Optimized Mixed signal / Analog techniques: Low power and low latency for extreme environment (ionizing radiation, deep cryogenic)

hls4ml…4asic?

Real-time, on-detector systems

Hardware acceleration with an emphasis on co-design and fast turnaround time

First project: Autoencoder with MNIST benchmark (28 x 28 x 8-bits @ 40 MHz)



Enable edge compute : e.g. data compression
Programmable and Reconfigurable: reprogrammable weights
Hardware – Software codesign: algorithm-driven architectural approach
Optimized Mixed signal / Analog techniques: Low power and low latency for extreme environment (ionizing radiation, deep cryogenic)

First tests of 1-layer design
Latency: 9ns
Power (FPGA, 28nm) ~ 2.5 W
Power (ASIC, 65nm) ~ 40 mW
Area = 0.5mm x 0.5mm

“Fast” ML for physics and beyond

Computing hardware and infrastructure

Real-time, on-detector systems

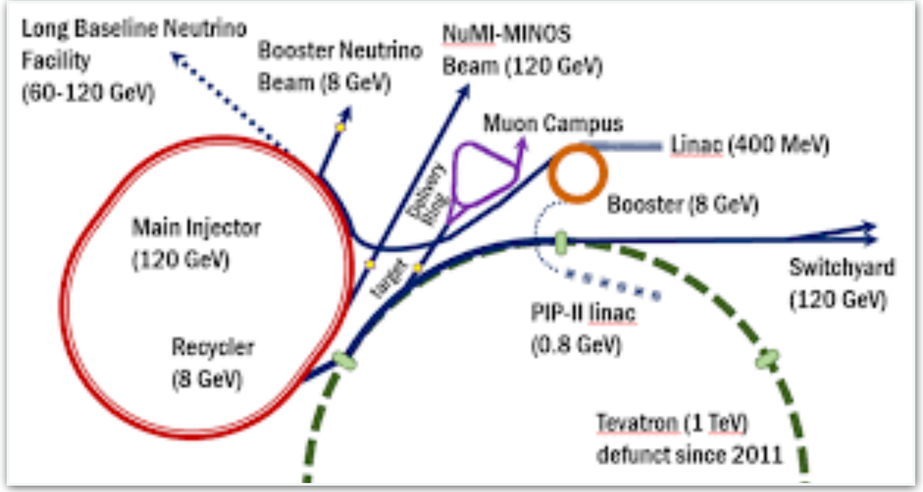
[https://indico.cern.ch/event/822126]

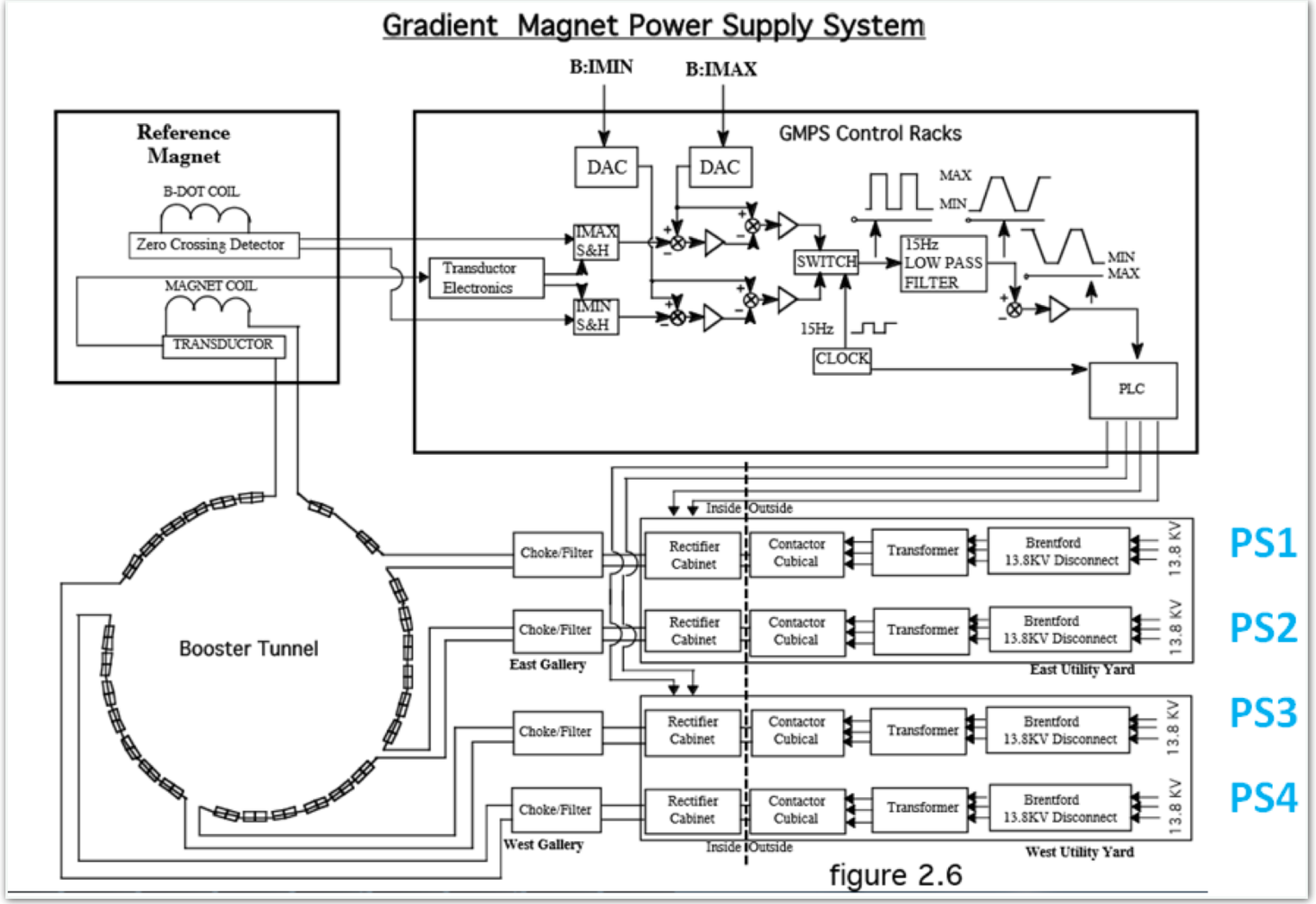
- Recent workshop at Fermilab with over 200 registered participants
- Fruitful discussions across physics for common challenges across domains
- Discussion also on where interesting intersections with industry and other fields

Get involved with the research community!
fastmachinelearning.org

Operations and control systems

Operations and control systems



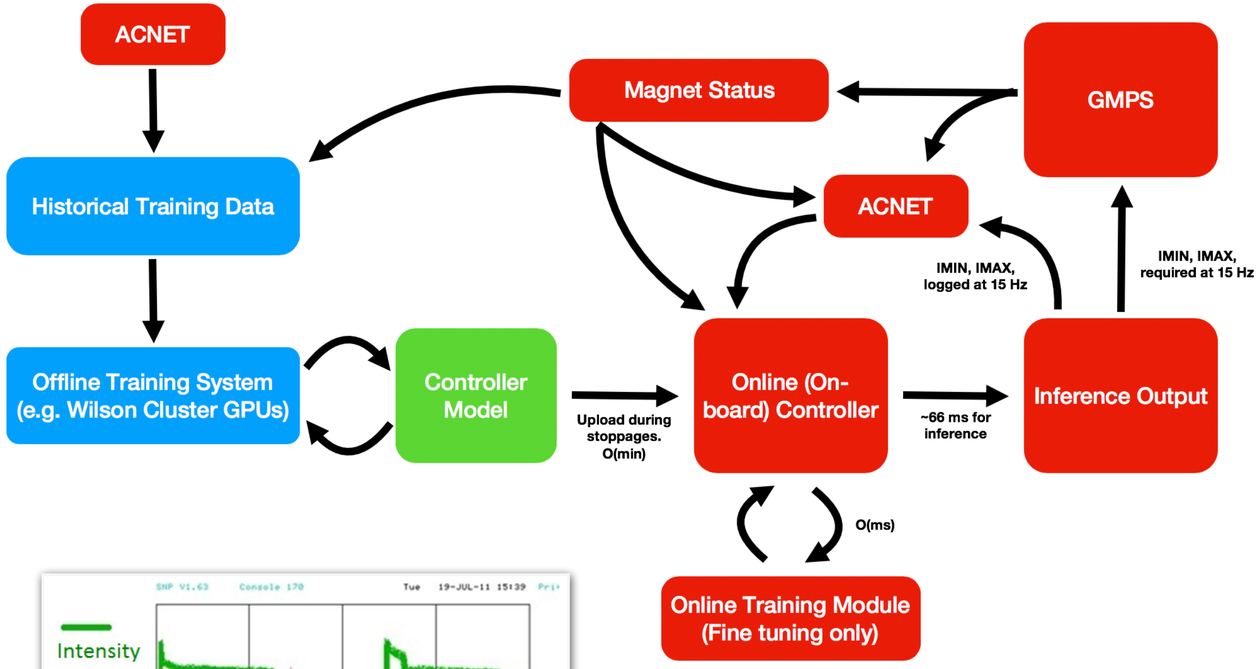


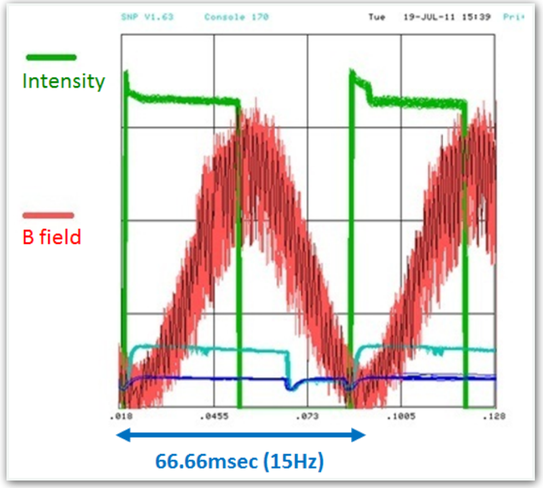
G. Perdue (FNAL) et al.

Accelerator controls with reinforcement learning

Operations and control systems

- Goal to reduce beam losses in Booster
- Solve problem using a ML algorithm on a custom FPGA board to control the magnet power supplies (GMPS) — deploy the hls4ml tool
- Scope is single crate control system (one board with back-up), but the project lays the foundation for a more ambitious future program.

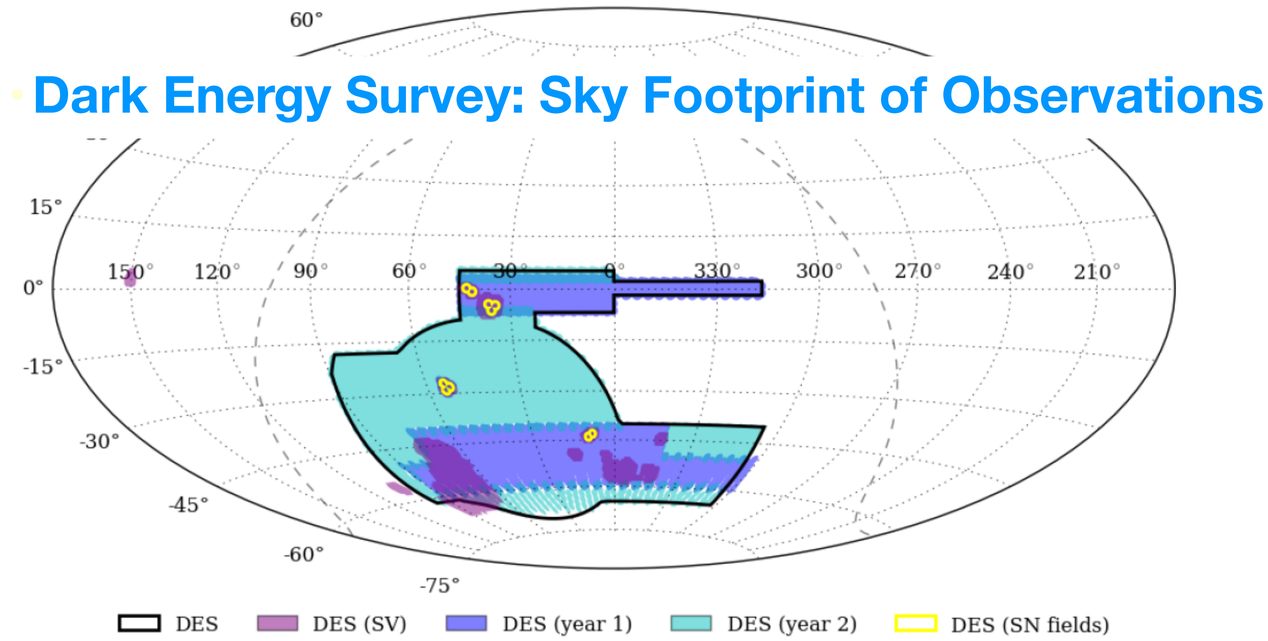




Self-Driving Telescopes: Control for observations

Operations and control systems

- Dark Energy Survey:
  Optical Images, Chilean Andes
- Terabytes per night
- 500 Million Galaxies, Thousands of exploding stars
- *Competing target requirements and environment constraints.*

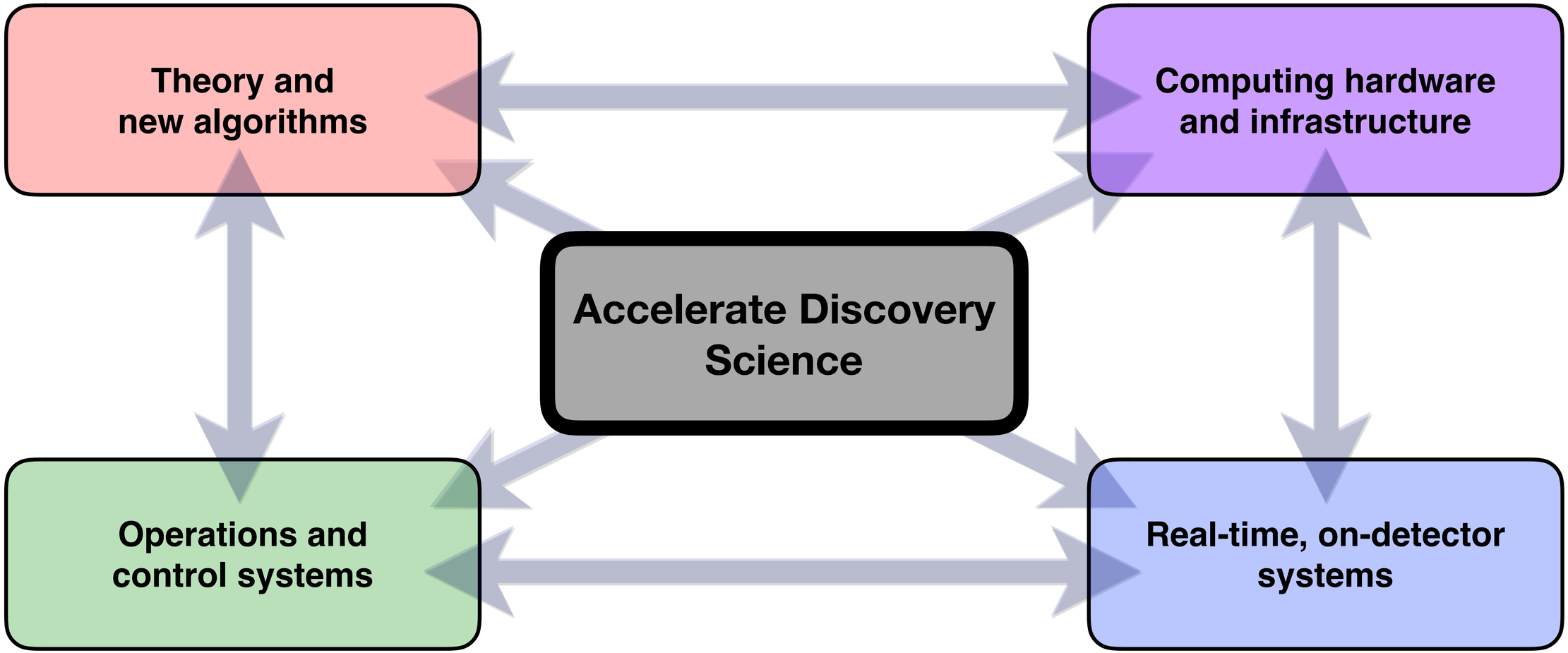


Challenge of scheduling on multiple time scales

- **Long**: competition between faint galaxies (green), transient objects (yellow)
- **Short**: Weather, annual modulation of sky positions
- Exploring reinforcement learning for optimal scheduling and control

Nord, Yuxin Chen (UC), JTFI award

AI Capabilities



Overflow and opportunities

- This talk was a *sampling* of activities going on in AI at Fermilab and more broadly in HEP
  - *Some other hot button topics — distributed training, uncertainty quantification, simulation for physics …*

- *Going forward?*
  - AI at Fermilab — living on the **cutting edge** to accelerate exciting science challenges
  - Built on general **AI capabilities applicable to multiple domains**

- **Fermilab AI program**
  contact us: ai@fnal.gov, team committed to building the AI community!
  - Interested in:
    - developing common technologies
    - collaborating on similar problems and working with experts in core AI and other domains

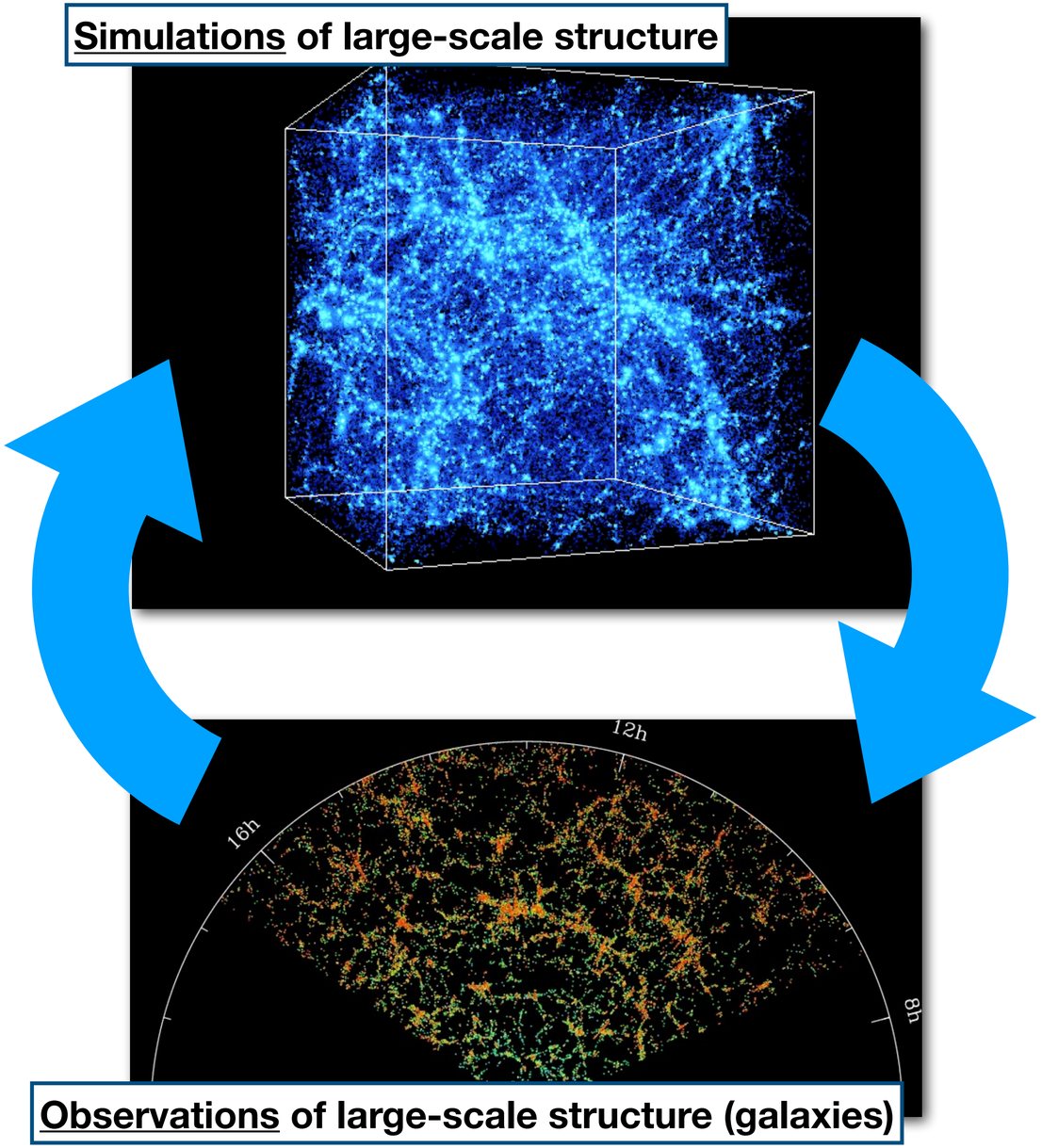
Summary and Outlook

- **AI to accelerate discovery science!**

- **Algorithms**
  - **Beyond Images**: higher-dimensional, sparse, multi-modal representation learning
  - Incorporating **physics into learning**: phase space integration, decorrelation, symmetries
- **Massive datasets**
  - Physics experiments are a excellent case for big data processing
  - Streaming real-time high-throughput systems to globally distributed offline compute
- **Real-time on-detector systems**
  - AI-integrated into the hardware filters, inferences at ~100ns
  - Applications for edge/IoT sensors — low-power, low-latency, cryogenic, and other extreme environments
- **Operations and controls**
  - **Improve efficiency** of site-wide operations from the accelerator complex to detector control

Extra

Simulations represent theory



We compare **simulated universes** (blue, upper) to **observed cosmic web of galaxies** (red/yellow, lower

- Across time scales: 3 orders of magnitude
- Across spatial scales: 6 orders of magnitude
- With several forces: Gravity, magnetism, hydrodynamics

- A single conventional simulation **>millions** of CPU hours, generates **>petabytes**
- We need 1000's of simulations (read: theoretical models) to test against real data.

→

- **Parameterized model-assisted GANs** for image and particle set generation
- Using **relationships between group symmetries and convolutions** to speed calculations