AI at Fermilab and beyond (in HEP)

Gabriel Perdue // Fermilab
perdue@fnal.gov -- @gnperdue

LAWSCHEP 2019

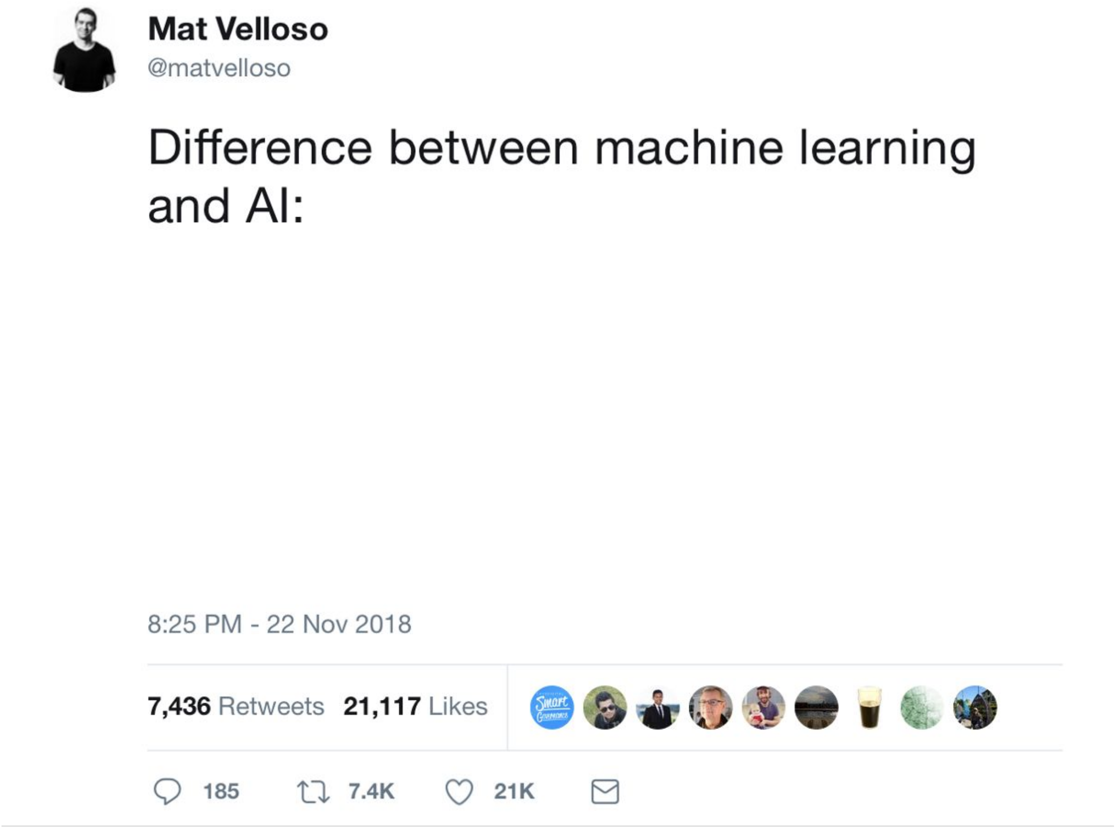
November, 2019

Special thanks to and N. Tran, B. Nord for slides and figures, and to B. Holzman and F. Fahim for additional supporting materials.

Overview

- The data landscape, HPC facilities
- HEP activities in AI
  - “Domain” applications
    - Colliders
    - Neutrinos
    - Cosmology
  - “General” applications
    - Control of complex devices
    - Simulations
    - Realtime AI
- Conclusions and future effort, thoughts on collaboration, etc.

Wait, what *is* AI?



Mat Velloso
@matvelloso

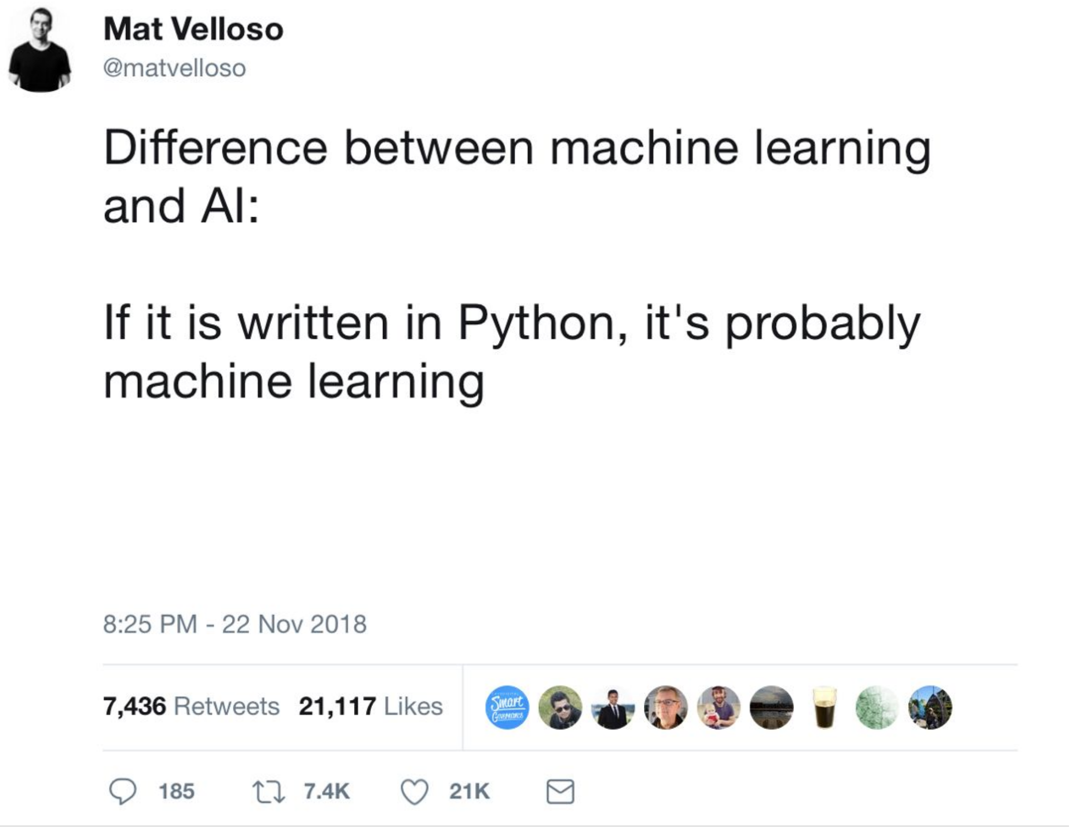
Difference between machine learning and AI:

8:25 PM - 22 Nov 2018

7,436 Retweets **21,117** Likes

185 7.4K 21K

Wait, what *is* AI?



Mat Velloso
@matvelloso

Difference between machine learning and AI:

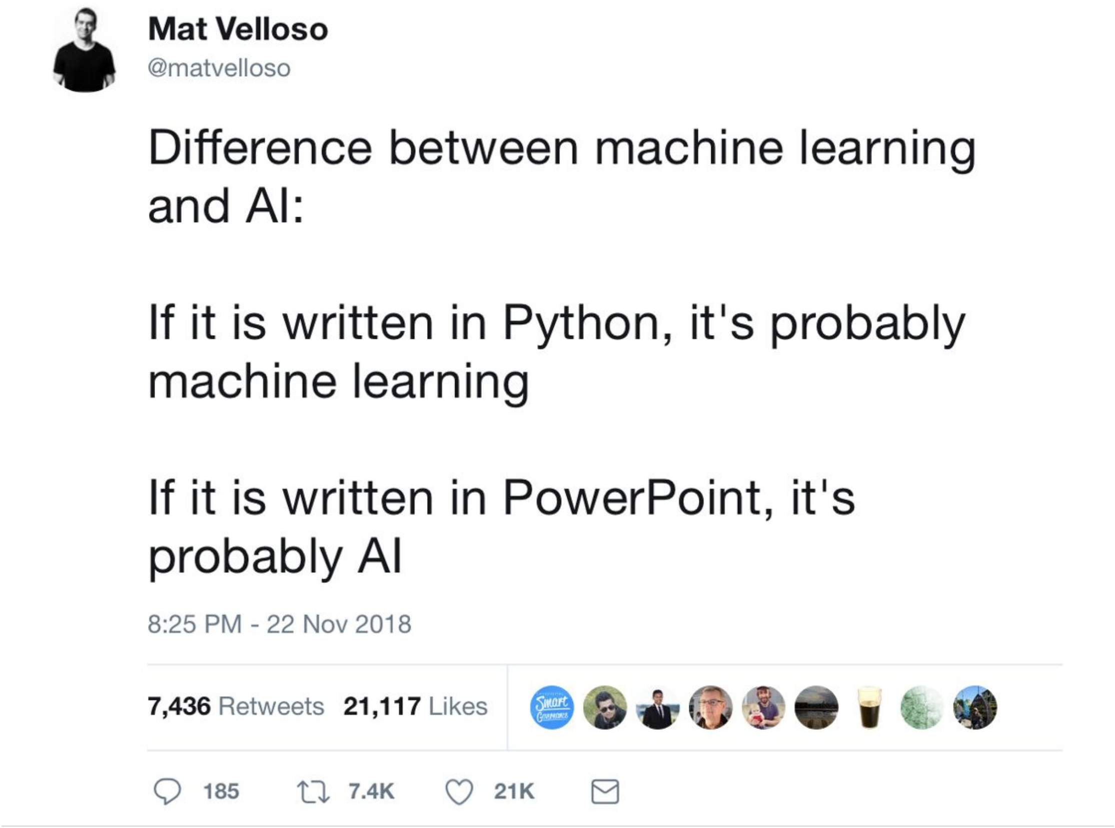
If it is written in Python, it's probably machine learning

8:25 PM - 22 Nov 2018

7,436 Retweets **21,117** Likes

185 7.4K 21K

Wait, what *is* AI?



Mat Velloso
@matvelloso

Difference between machine learning and AI:

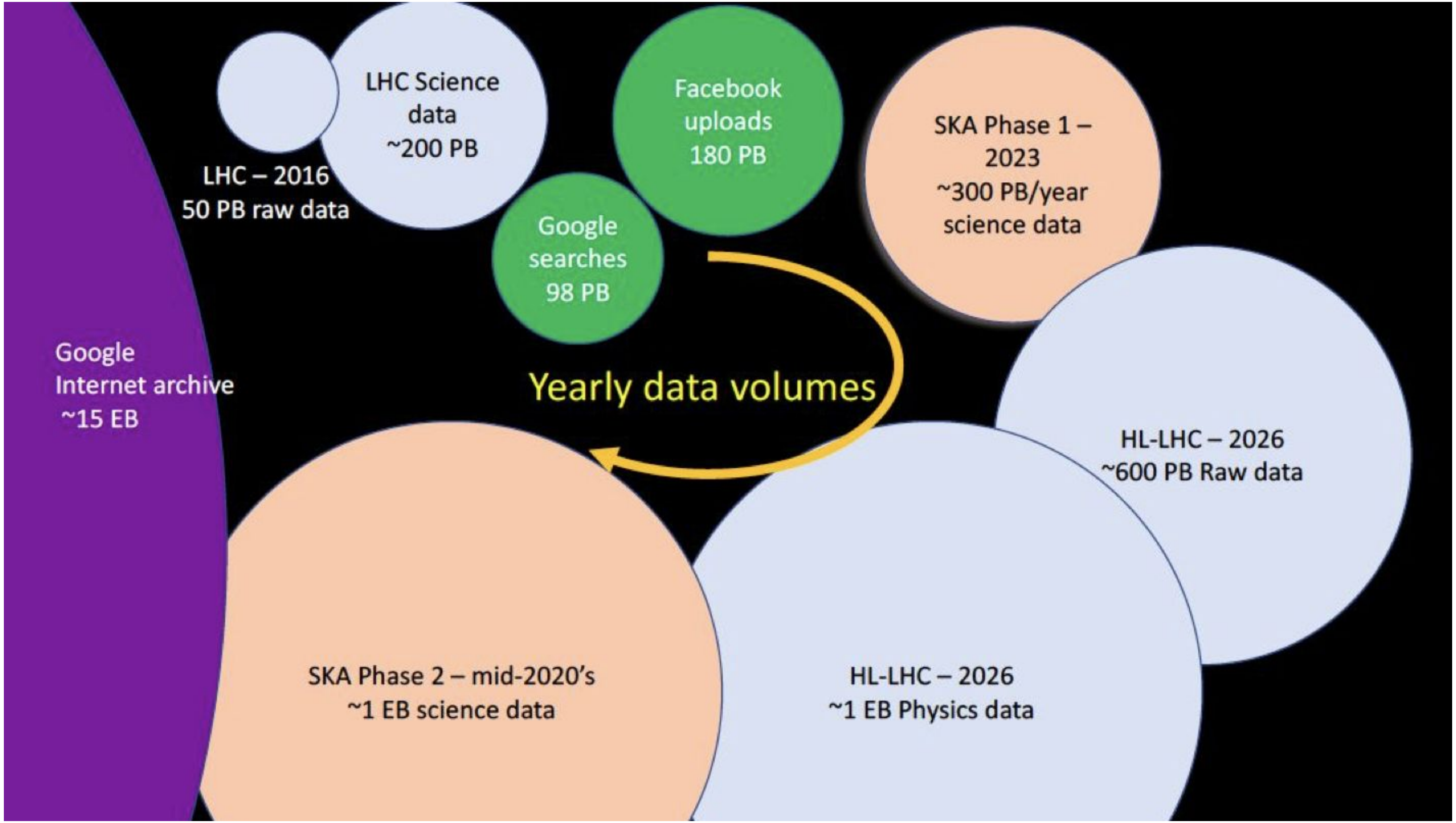
If it is written in Python, it's probably machine learning

If it is written in PowerPoint, it's probably AI

8:25 PM - 22 Nov 2018

7,436 Retweets **21,117** Likes

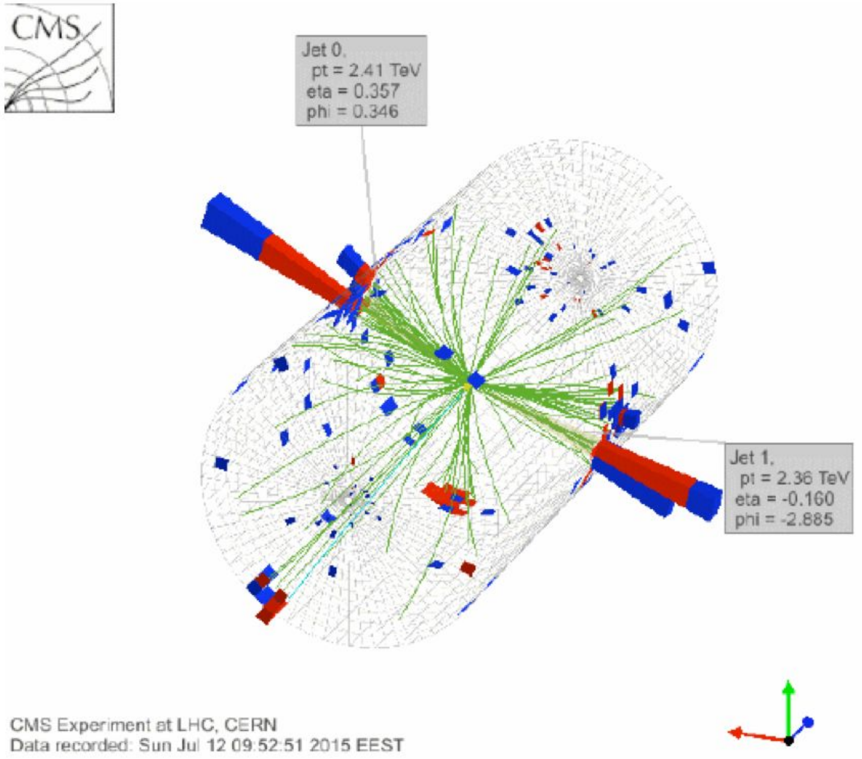
If data is the fuel for modern AI...



HPC Facilities and AI

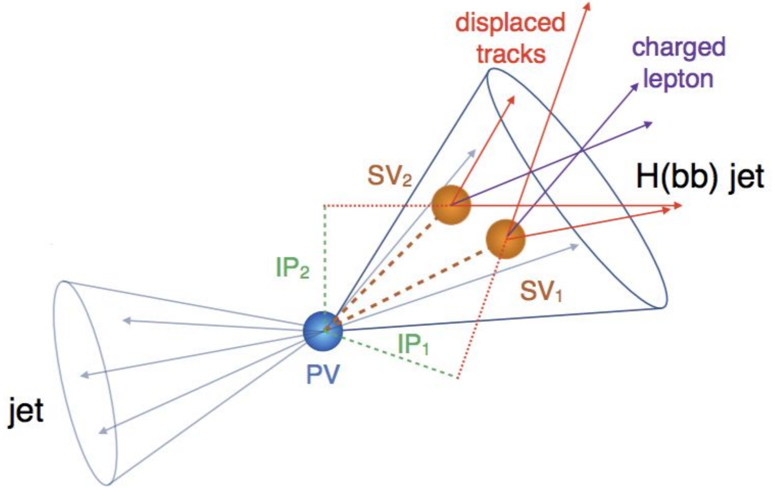
- HPC facilities are evolving towards AI and analysis workloads.
  - Long tradition in simulation applications.
  - Large implications for interconnect, storage systems, etc. -> Workloads change from compute intensive to IO intensive.
  - HPC + AI is a *dominant paradigm* in the US DOE -> major implications for HEP.
- AI provides a path to adapt HEP workloads to HPC facilities:
  - Map problems into machine learning problems.
  - Map imperative algorithms into AI models.
- Interesting, complex pros and cons:
  - Saves us the need to rewrite all of our code to support GPU/TPU/etc. platforms, but we need to understand biases and uncertainties in AI algorithms at a much deeper level.
    - Domain translation bias, efficient uncertainty propagation, model systematics.
    - Rich area for collaboration (within and without HEP).
    - HEP has an opportunity to provide leadership in these areas.

Higgs at the LHC



Fermilab is the lead lab on the Compact Muon Solenoid experiment at the LHC

A broad range of physics from studying the Higgs boson to searches for dark matter

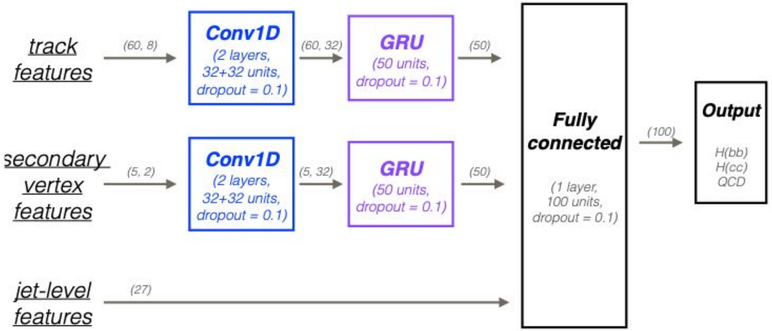


New deep learning techniques to identify the Higgs boson in dense, energetic decays to bottom quarks

Figures courtesy of N. Tran

COLLIDER PHYSICS

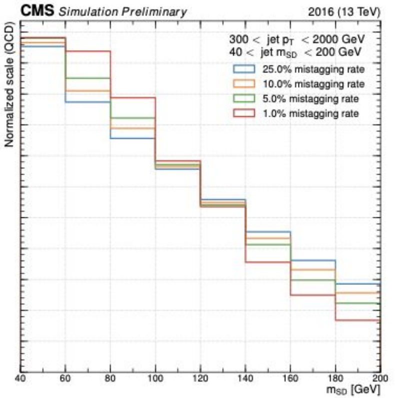
Example: Identifying the Higgs


track features
(60, 8)
Conv1D (2 layers, 32+32 units, dropout = 0.1)
(60, 32)
GRU (50 units, dropout = 0.1)
(50)

secondary vertex features
(5, 2)
Conv1D (2 layers, 32+32 units, dropout = 0.1)
(5, 32)
GRU (50 units, dropout = 0.1)
(50)

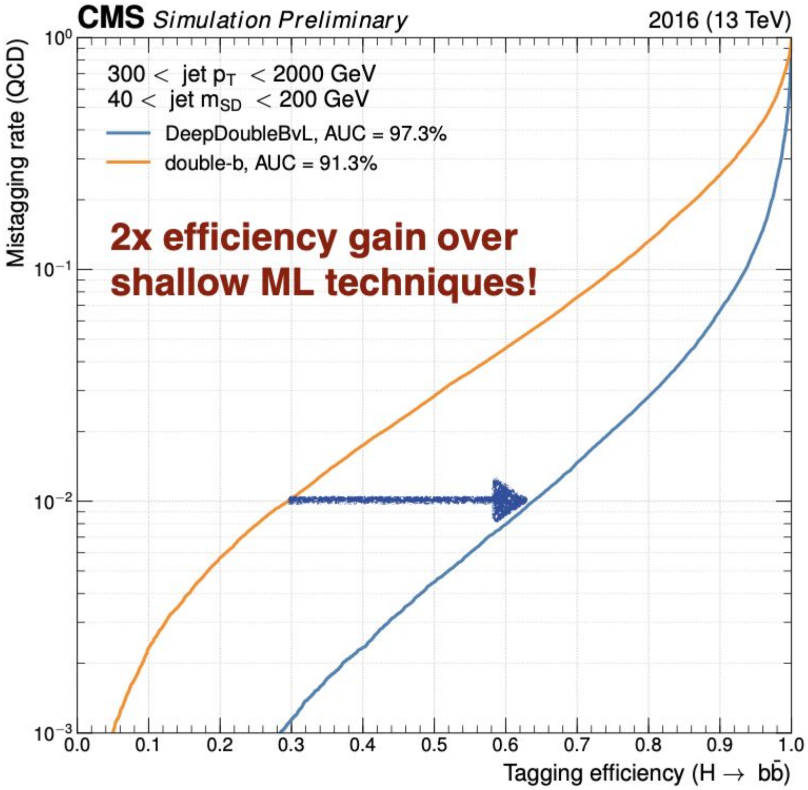
jet-level features
(27)

Fully connected (1 layer, 100 units, dropout = 0.1)
(100)
Output
H(bb)
H(cc)
QCD



decorrelation:

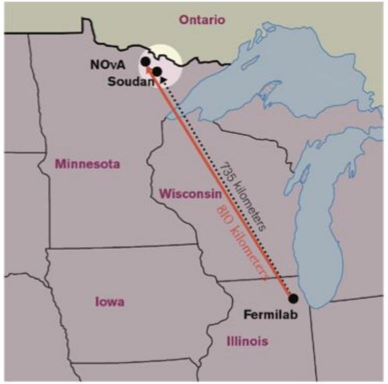
teach the network how to *not* learn certain physical features; important for controlling systematic uncertainties



2x efficiency gain over shallow ML techniques!

Figures courtesy of N. Tran

Neutrinos at NOvA



Near Detector

0.3 kt, > 20,0000 channels

100 m below service

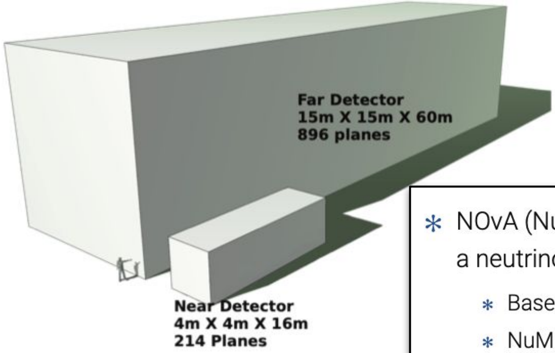
1 km from NuMI

Far Detector

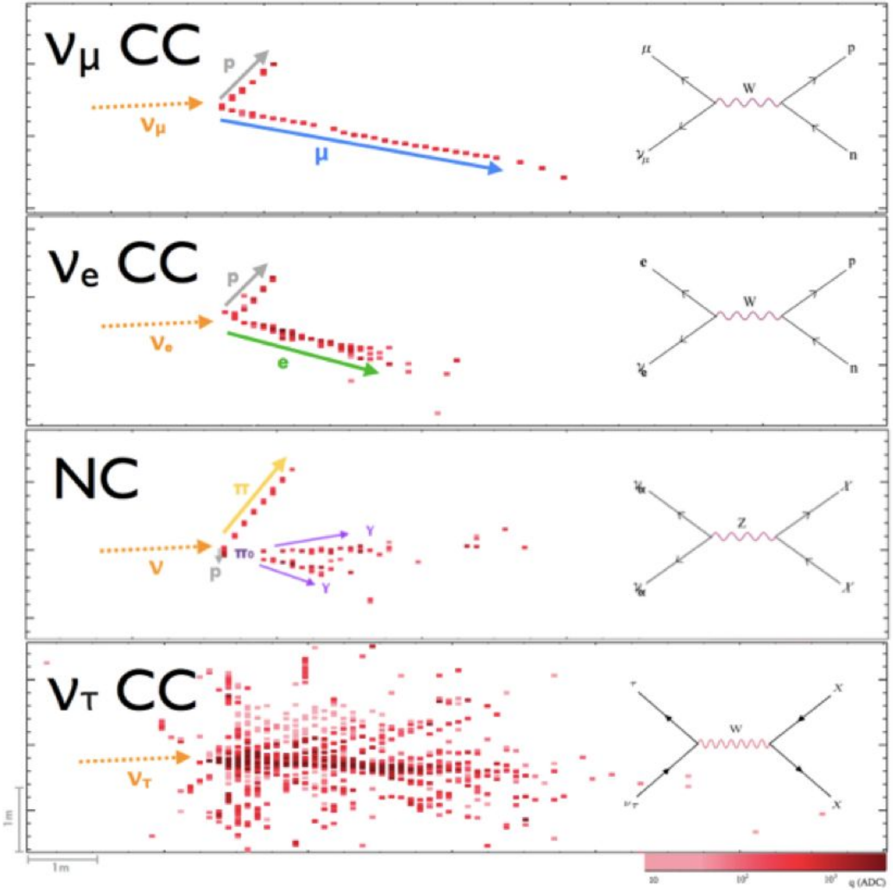
14 kt, > 344,000 channels

On surface

810 km from NuMI

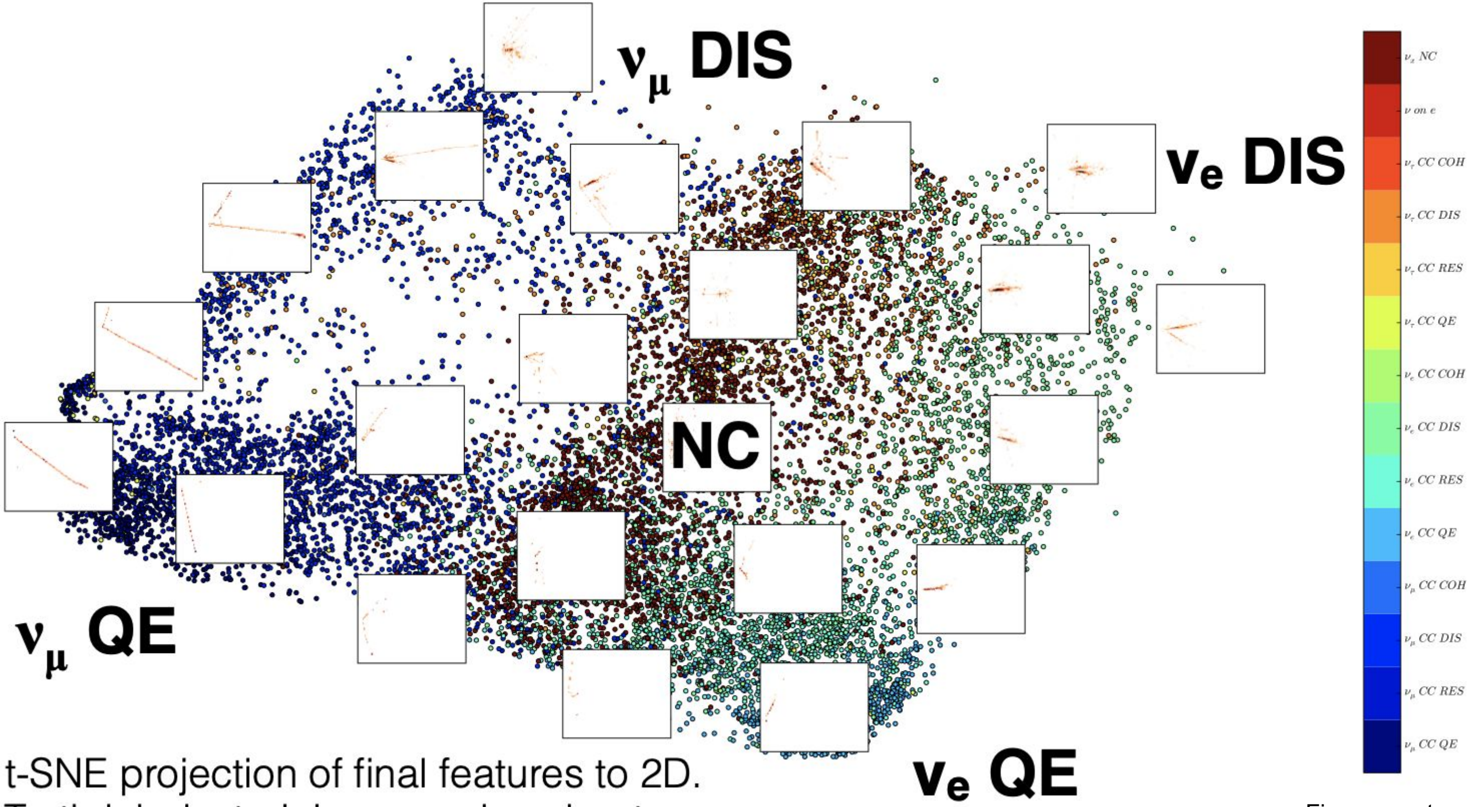


- NOvA (NuMI Off-axis ν_e Appearance) is a neutrino oscillation experiment
  - Baseline of 810 km
  - NuMI, beam of mostly ν_μ
  - 14 mrad off-axis from the beam
  - Two functionally identical detectors



Plots and figures courtesy of A. Tsaris, A. Himmel

NOvA: explainable AI



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

Figure courtesy of A. Radovic

Neutrino physics

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

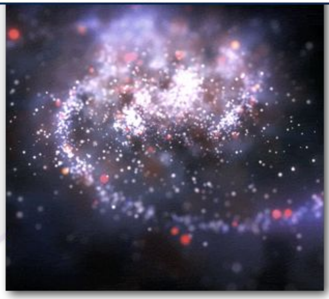
Simulating the Universe

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)

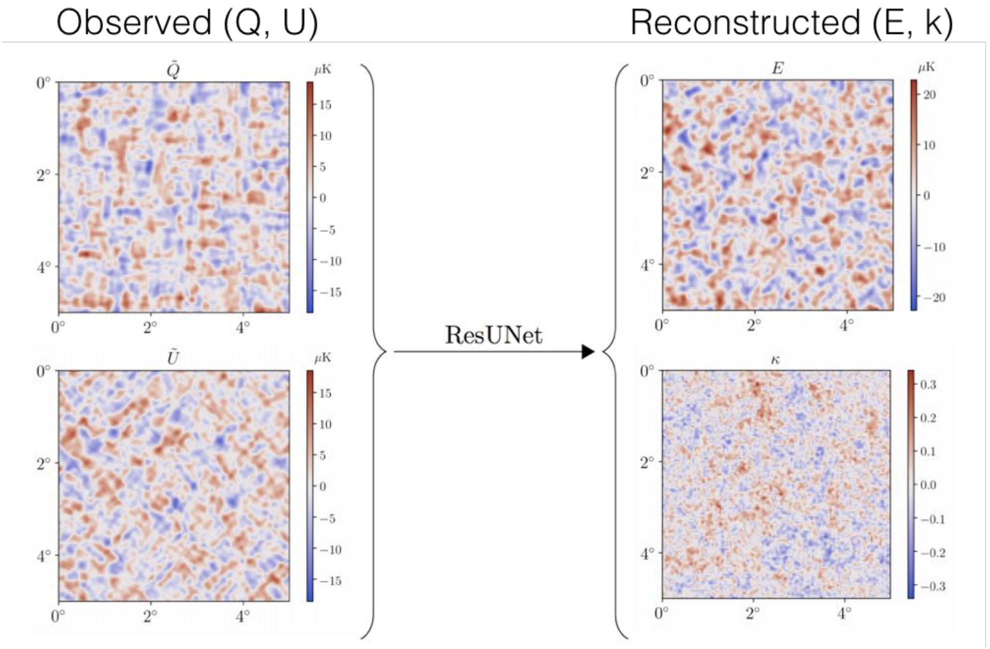
Task: Identify and measure objects in noisy data

Task: Separate data structures into components

COSMOLOGY

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

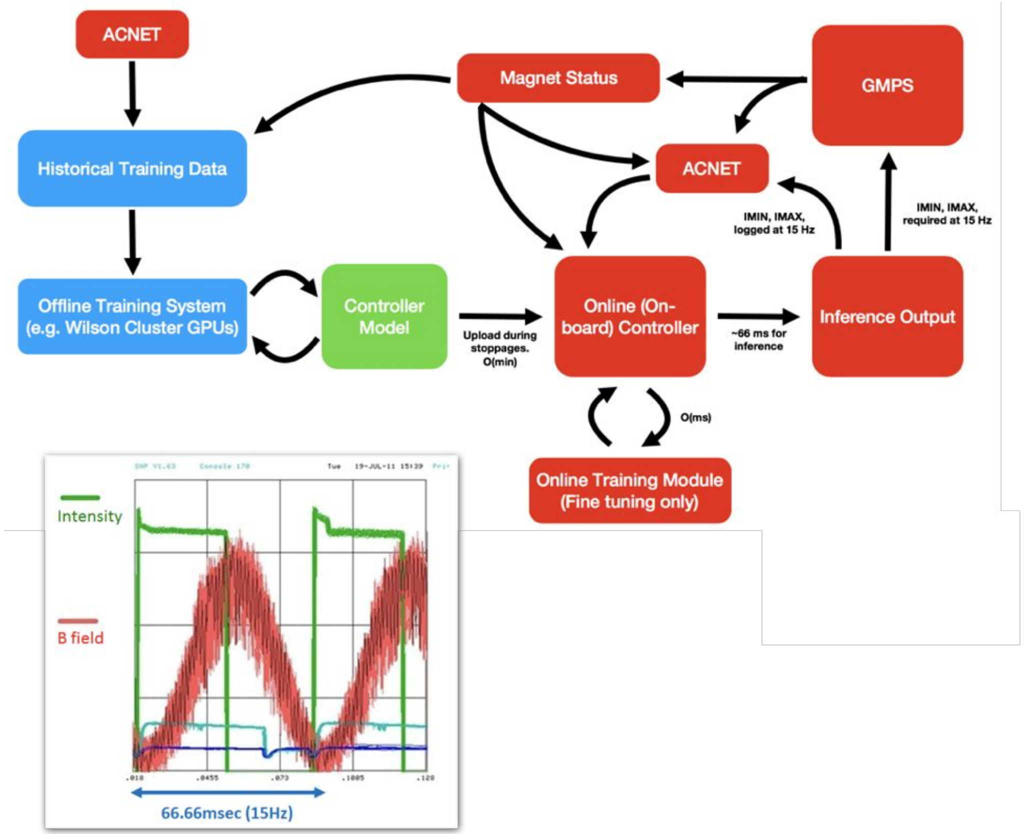


- Noise and other foregrounds obfuscate primordial
- Pioneered the use of Residual UNets to separate signals (k) from noise (E)

COSMOLOGY

Accelerator controls with reinforcement learning

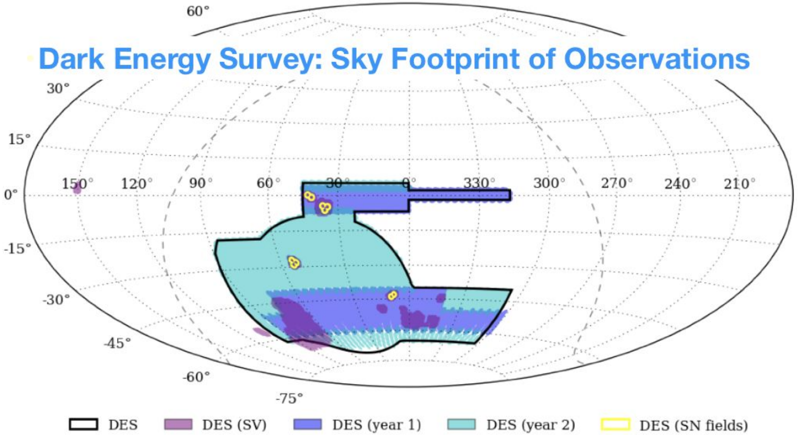
- Seek to reduce losses in Fermilab Booster, reduce operational burdens
- Build a self-tuning system that consumes low-latency environment and accelerator state behavior and runs an AI algorithm on a custom FPGA board to control magnet power supplies
  - Leverages hls4ml tool and CMS trigger technology
- Scope of the project is a single-crate control system, but the effort lays the foundation for more ambitious future programs
- Collaborative effort with Pacific Northwest National Laboratory


ACNET
Historical Training Data
Offline Training System (e.g. Wilson Cluster GPUs)
Controller Model
Upload during stoppages. O(min)
Online (On-board) Controller
~66 ms for inference
Inference Output
Magnet Status
GMPS
ACNET
IMIN, IMAX, logged at 15 Hz
IMIN, IMAX, required at 15 Hz
O(ms)
Online Training Module (Fine tuning only)
Intensity
B field
.018
.0455
.073
.1005
.128
66.66msec (15Hz)

Control of Complex Systems

Self-driving telescopes: Control machines, schedule observations

Precision Machine Control
DES: 2012-2018



Dark Energy Survey: Sky Footprint of Observations

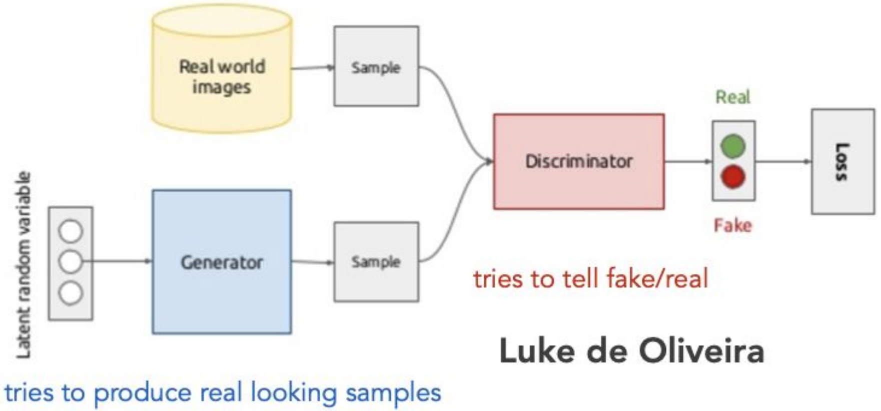
- Dark Energy Survey: Optical images, Chilean Andes
- Terabytes of data per night
- 500 million galaxis, thousands of exploding stars
- *Competing target requirements and environmental constraints*

Challenge of scheduling on multiple time scales:

- Long: competition between faint galaxies (green) and transient objects (yellow)
- Short: weather, annual modulation of sky positions
- Exploring reinforcement learning for optimal scheduling and control (balance short term rewards with long-term cumulative gain)

Generative models for simulation

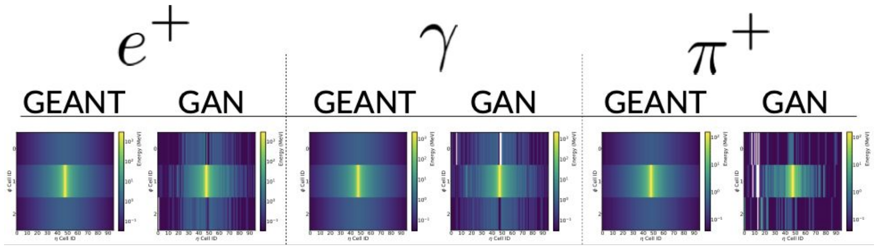
- Future simulation needs (e.g., HL-LHC) appear likely to outstrip even optimistic resource projections.
- Shower libraries face problems rooted in incompleteness and heavy data access.
- Generative models offer a potentially incredible speed-up along with better flexibility by modeling very complex distributions.



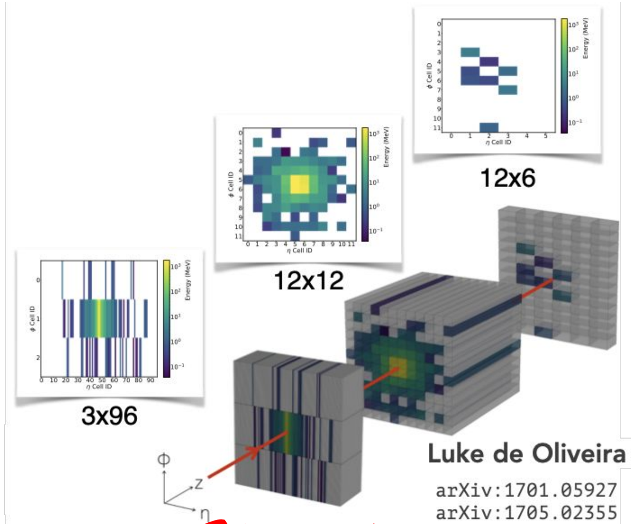
tries to tell fake/real

tries to produce real looking samples

Luke de Oliveira



Michela Paganini*, Luke de Oliveira, Ben Nachman DS@HEP 2017

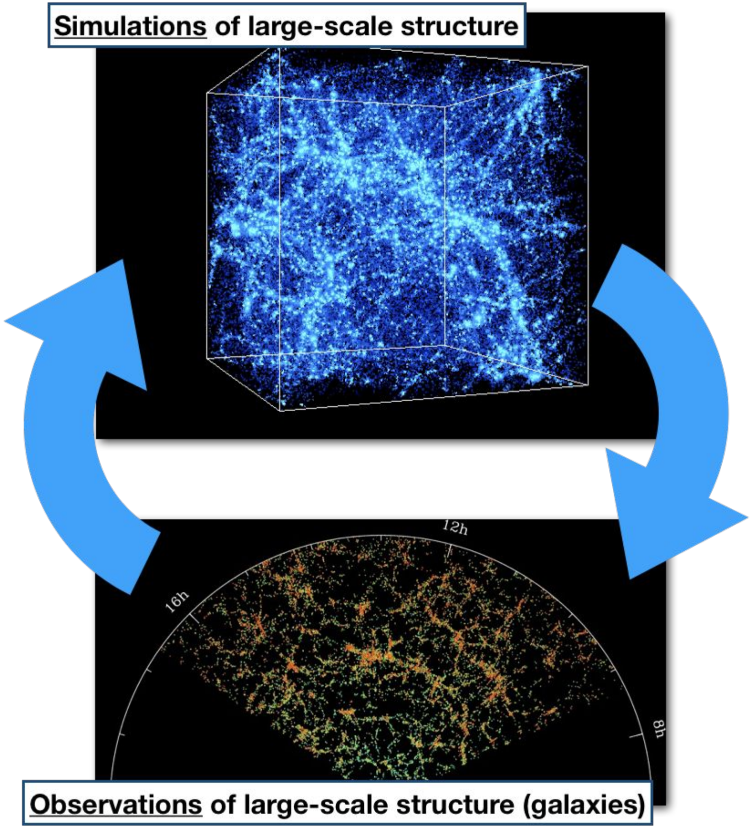


Luke de Oliveira

arXiv:1701.05927
arXiv:1705.02355

SIMULATION

Simulations represent theory


Simulations of large-scale structure

Observations of large-scale structure (galaxies)

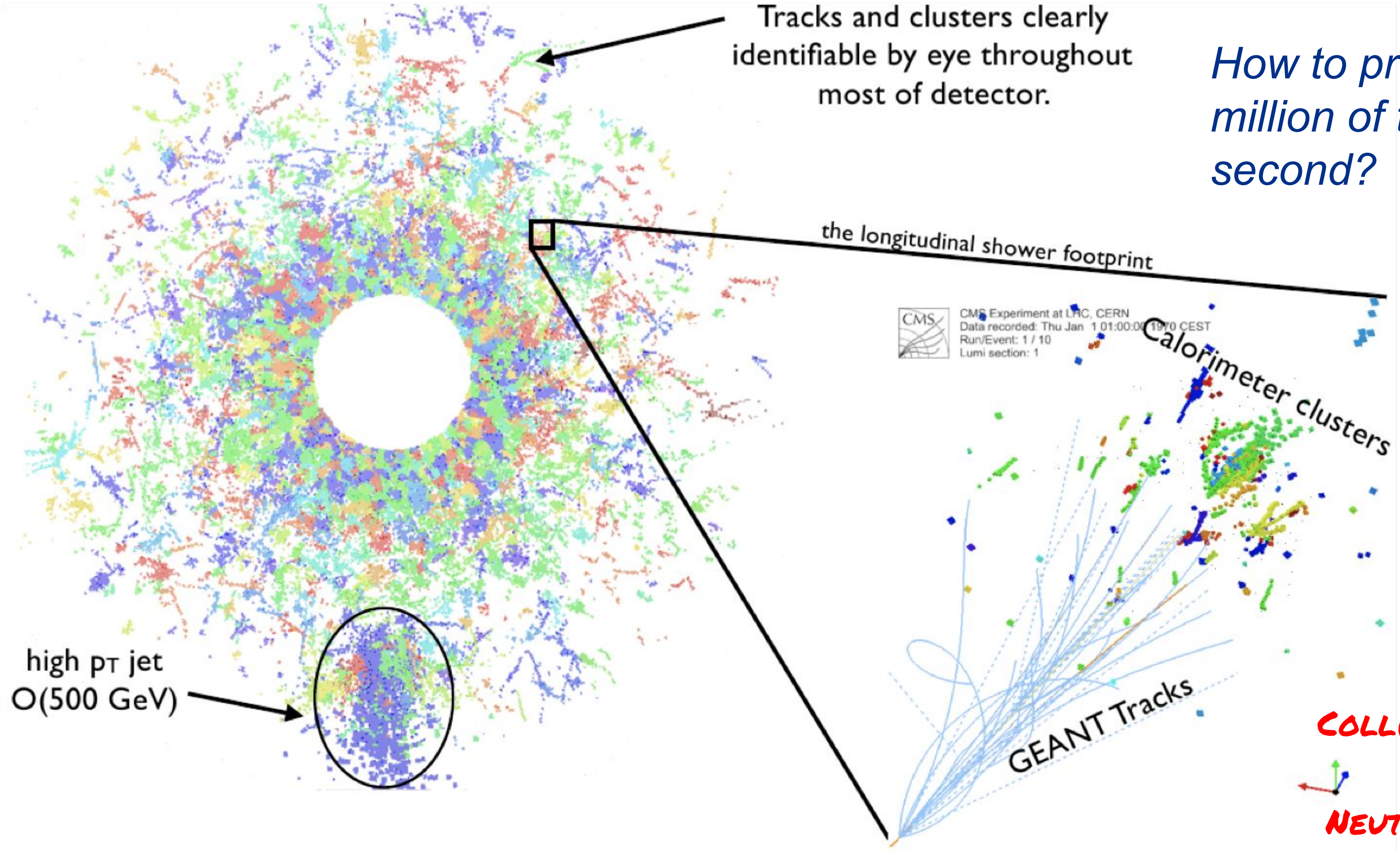
We compare simulated universes (blue, upper) to observed cosmic web of galaxies (red/yellow, lower)

- Across time scales: 3 orders of magnitude
- Across spatial scales: 6 orders of magnitude
- With forces: gravity, magnetohydrodynamics

- A single conventional simulation: millions of CPU hours and petabytes of data
- We need 1000’s of simulations (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 up calculations

SIMULATION


Tracks and clusters clearly identifiable by eye throughout most of detector.
the longitudinal shower footprint
CMS
CMS Experiment at LHC, CERN
Data recorded: Thu Jan 1 01:00:00 1970 CEST
Run/Event: 1 / 10
Lumi section: 1
Calorimeter clusters
GEANT Tracks
high pT jet
O(500 GeV)

How to process 40 million of these per second?

Collider physics

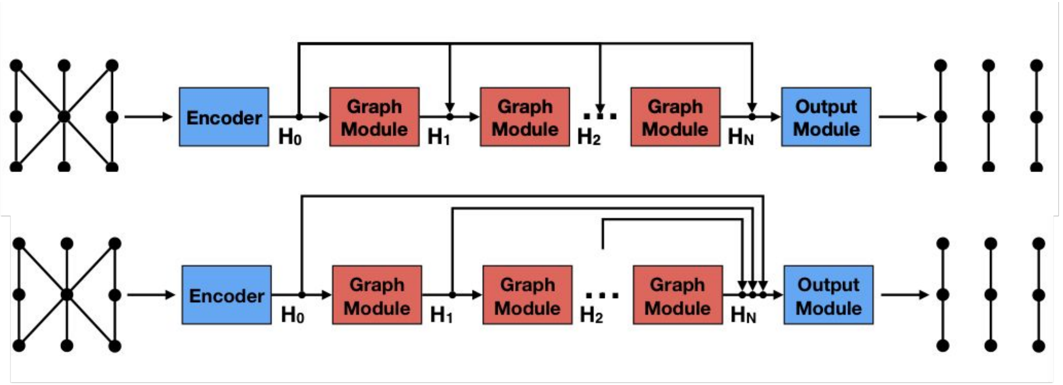
Neutrino physics

Graph Networks

Beyond images

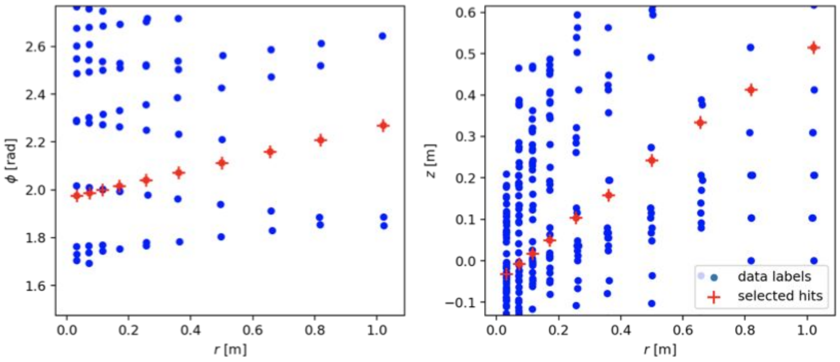
Neutrino physics

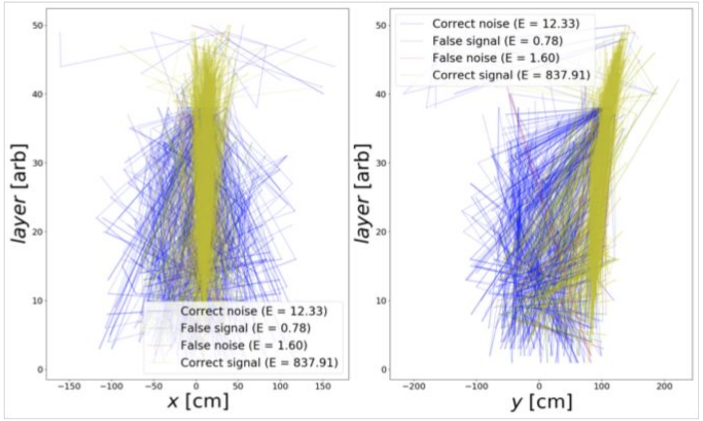
Collider physics



Tracking

Clustering

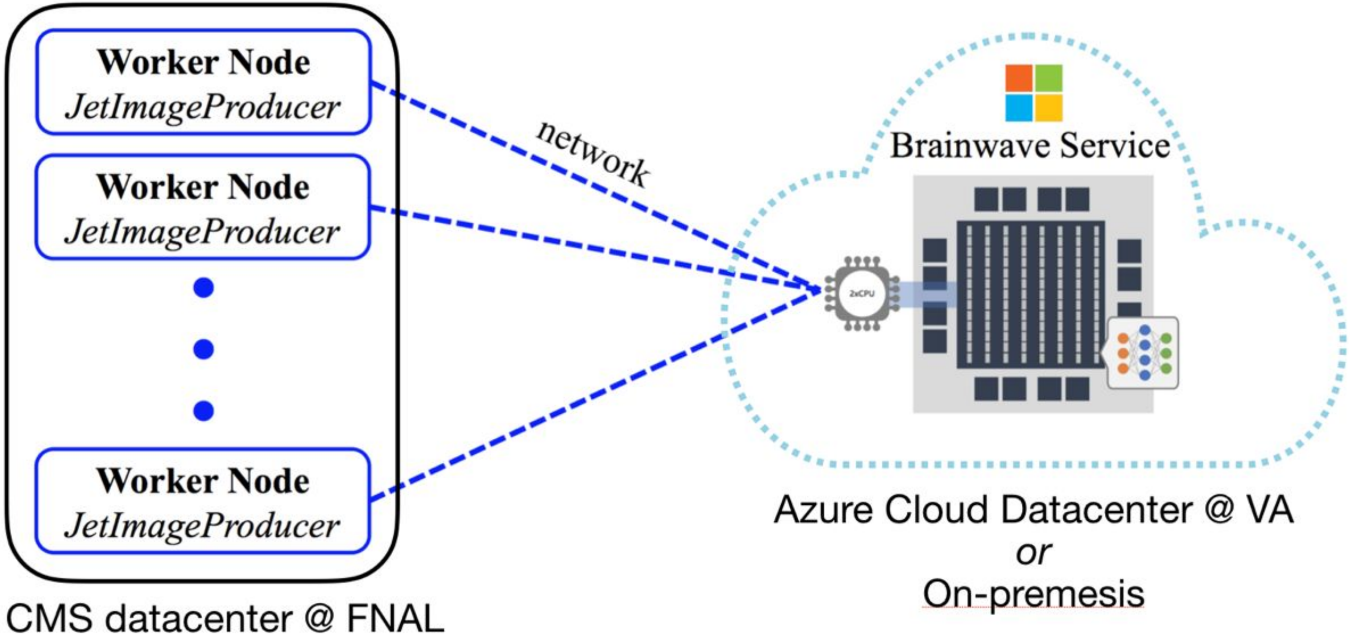




Graph Networks

Fast inference as a service

- Found 30x (175x) speed up for cloud (edge) inference using ResNet50 as compared to the experiment's software framework.



Non-disruptive integration of heterogenous computing resources into HEP computing model.

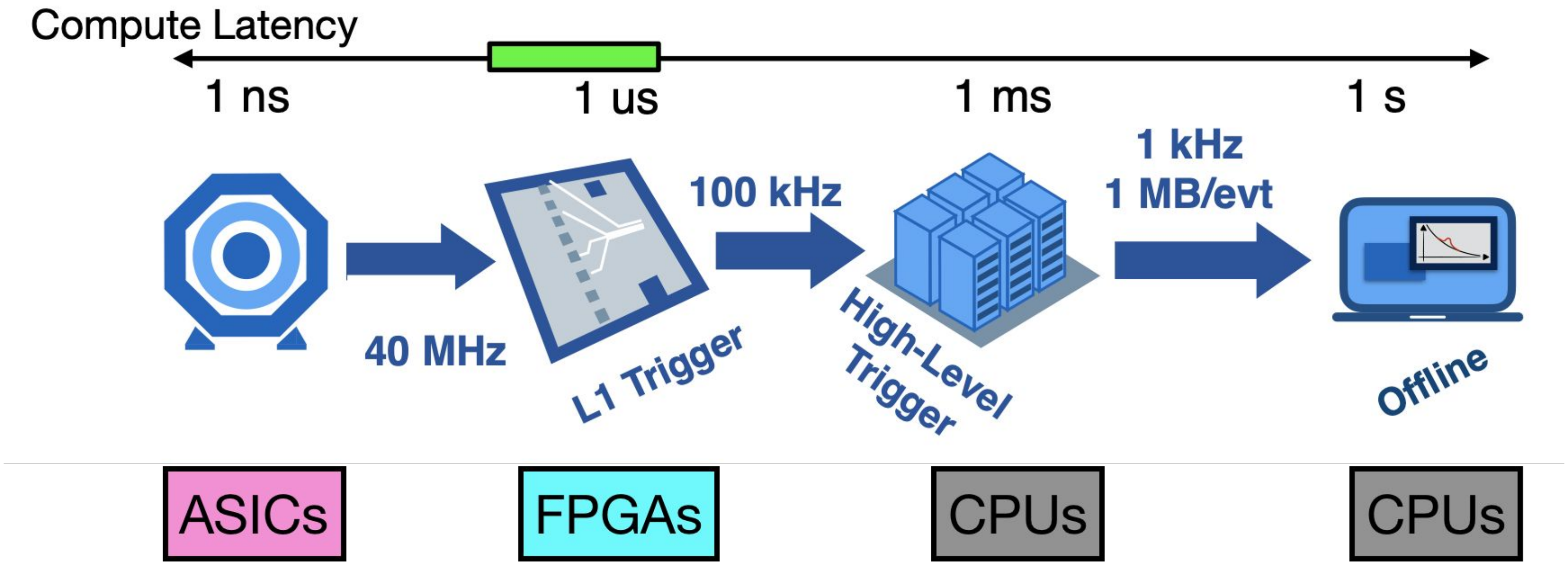
Deploy as a service (many CPUs to few FPGAs) can be more cost-effective.

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

CMS data processing chain

~1 PB/s

~1 PB/DAY



Compute Latency

1 ns

1 us

1 ms

1 s

40 MHz

L1 Trigger

100 kHz

High-Level Trigger

1 kHz

1 MB/evt

Offline

ASICs

FPGAs

CPUs

CPUs

REAL-TIME ON-DETECTOR AI

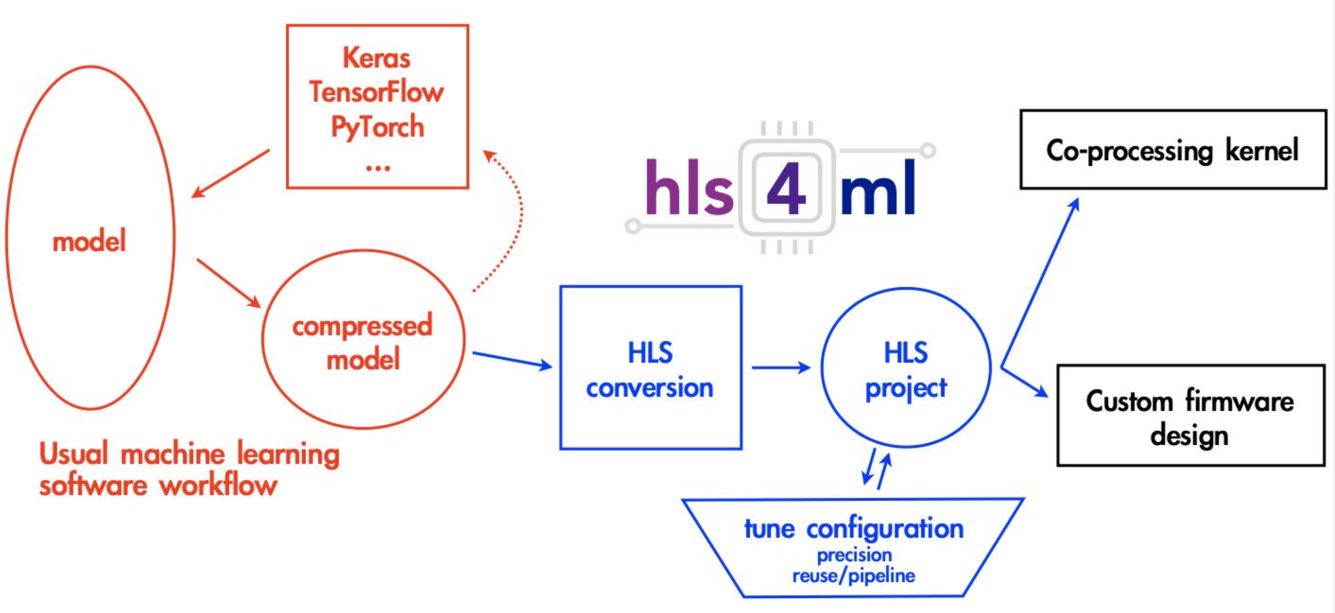
Figure by N. Tran

REALTIME AI

ML in the hardware trigger

https://arxiv.org/abs/1804.06913

- L1 Trigger: incorporate ML for on-detector real-time processing
- **hls4ml** - open-source automated translation tool: ML models to firmware
  - https://fastmachinelearning.org/hls4ml



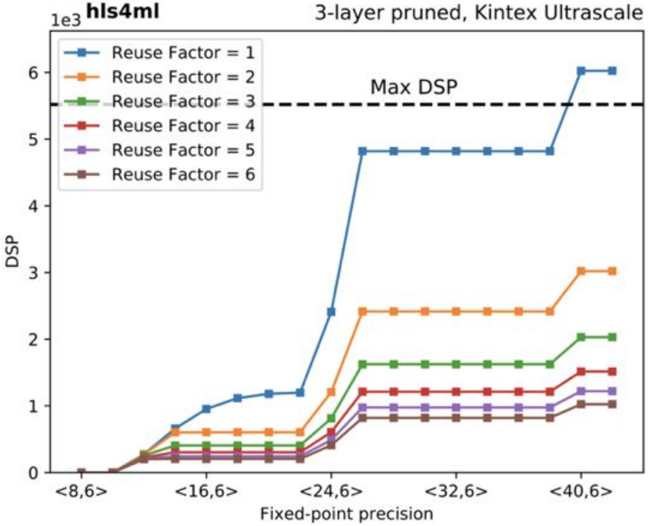
ML in the hardware trigger, cont.

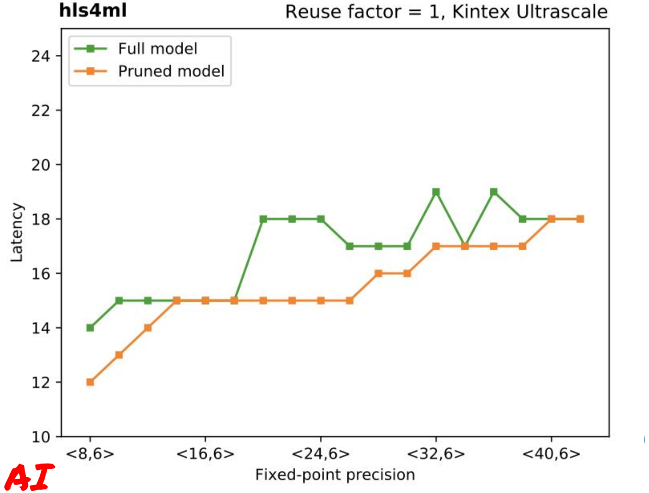
- All FPGA design
  - Flexible - many algorithm types for processing layers
- Application and adopting growing across the LHC
  - Firmware in hours instead of weeks/months
- Model compression and quantization showing no meaningful impact on classification performance.

- 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

https://arxiv.org/abs/1804.06913

>5000 parameter fully-connected network in 100 ns





Future directions

- AI is an important area to invest in.
- There is a real possibility of the world resolving into an AI duopoly dominated by the US (companies) and China (government). This has serious ethical and sociological consequences.
- Dominance in AI comes from dominance in data. HEP data, particularly large open datasets but also those that are “closed” within the experiments provide useful material for building expertise in this area and we should leverage this fact for funding support and actively seek collaboration with computer scientists, electrical engineers, and mathematicians to build interdisciplinary teams around HEP datasets that can grow new centers of excellence in this intellectual space.
- Opportunities for collaboration abound - it is important we all work together.

Thoughts on collaboration

- AI is a very challenging area to hire in. There is a very strong pull to companies that pay salaries we cannot match in science. We can compete on interesting problems! And we can continue to provide value as educators and for training.
  - “Workforce development” is a major “buzz phrase” in the US.
- The barriers to entry are (or at least can be) low. Ultimately large computing resources are required, but it is possible to bring real value to a collaboration around formal methods, algorithmic innovation, and creative applications.

Thoughts on collaboration, cont.

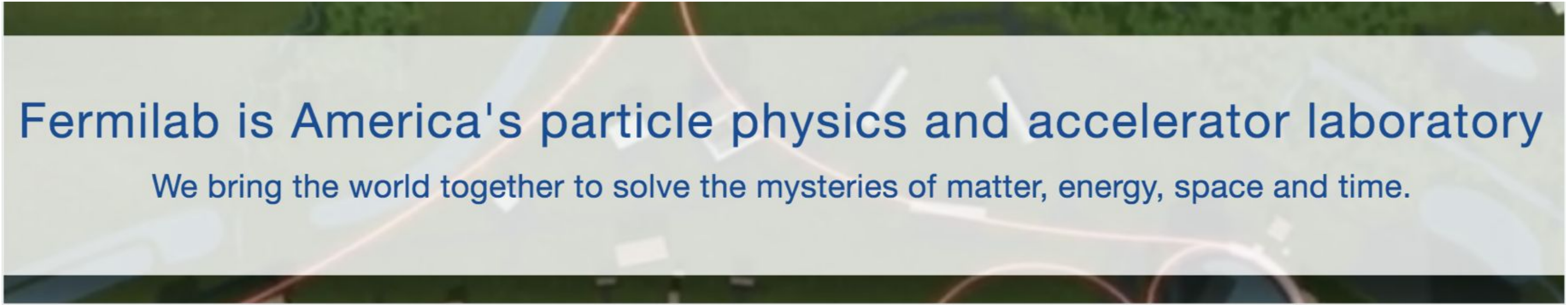
- We are very eager to collaborate in this space at Fermilab. Please talk to us!
- There are also excellent opportunities to build collaborative networks around specific, focused applications and algorithms in Latin America. Some infrastructure investments are required, but within an HEP context (and access to HEP data), there are opportunities to build a *world-leading program* in an area.

Conclusions

- The AI program in HEP is both deep and broad.
    - There are many topics I didn’t cover in theory, speeding up event generators, data quality monitoring, computing operations, and so on.
- We are moving past the era of novelty applications are really looking at hard at two questions:
    - What are the applications we could not pursue without this technology? What is the real extent of the power that AI can bring to HEP science?
        - Young people are the key to this! We need people who grew up with a technology to see how to apply it in truly novel ways. The current generation of leaders must enable this activity and avoid seeing AI as only a way to solve old problems better. We want to find changes in kind, not only in degree.
    - How can HEP help improve AI?
        - We bring novel data, problems, constraints, and interests to the endeavor.
        - We are also good at understanding problems from first principles - we may have much to offer here in particular in terms of explainable AI.

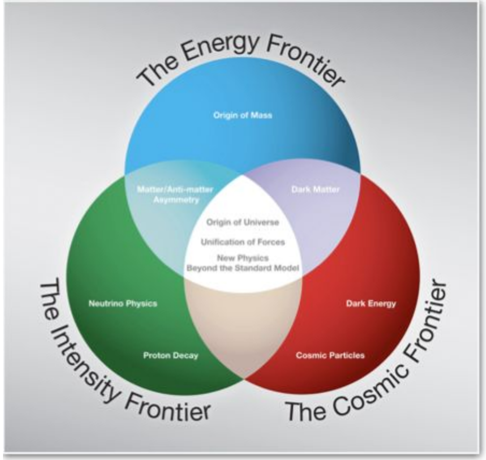
Thanks for listening!

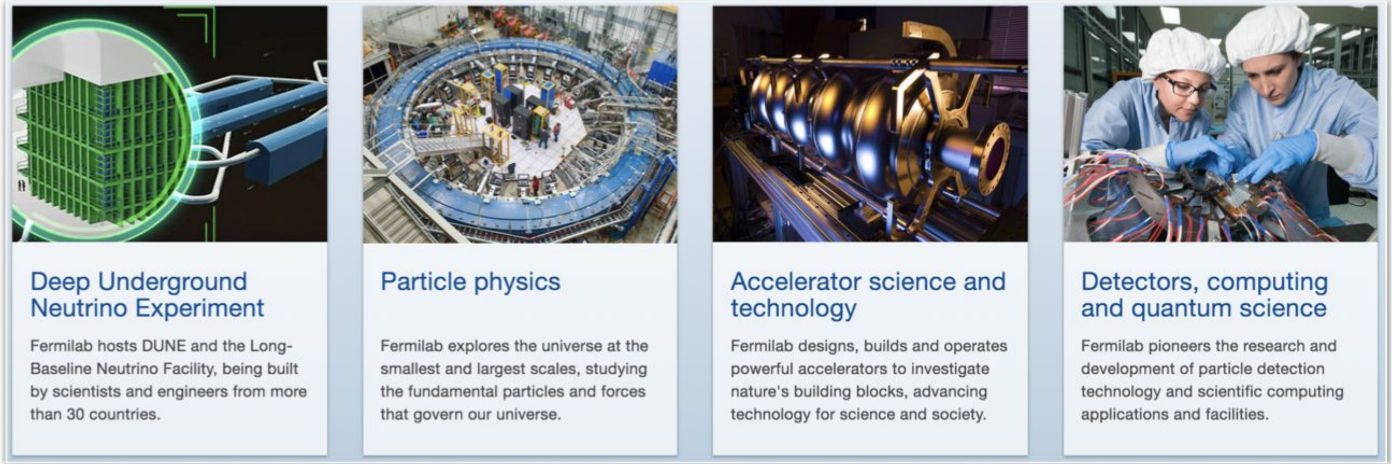
Introduction to Fermilab



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

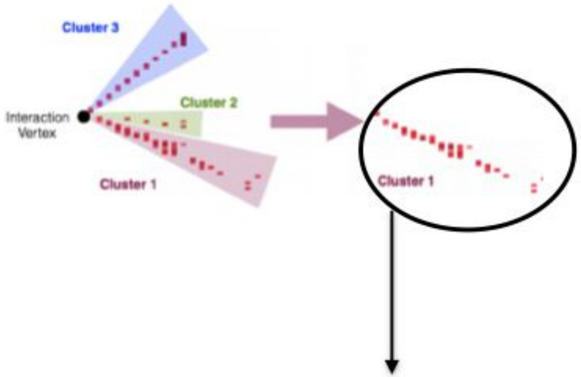
Fermilab designs, builds and operates powerful accelerators to investigate nature's building blocks, advancing technology for science and society.

Detectors, computing and quantum science

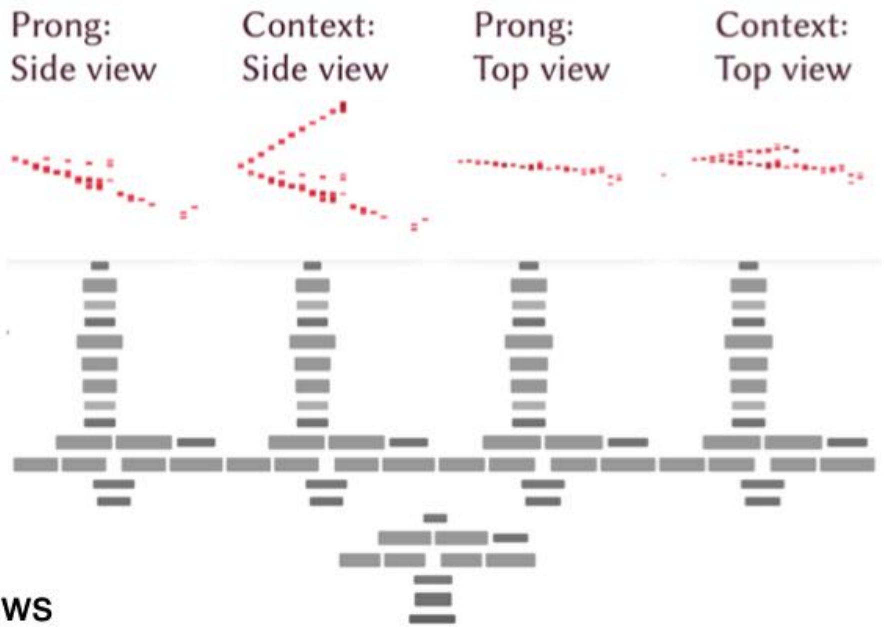
Fermilab pioneers the research and development of particle detection technology and scientific computing applications and facilities.

NOvA: Particle Reconstruction

Single particles are separated using geometric reconstruction methods.

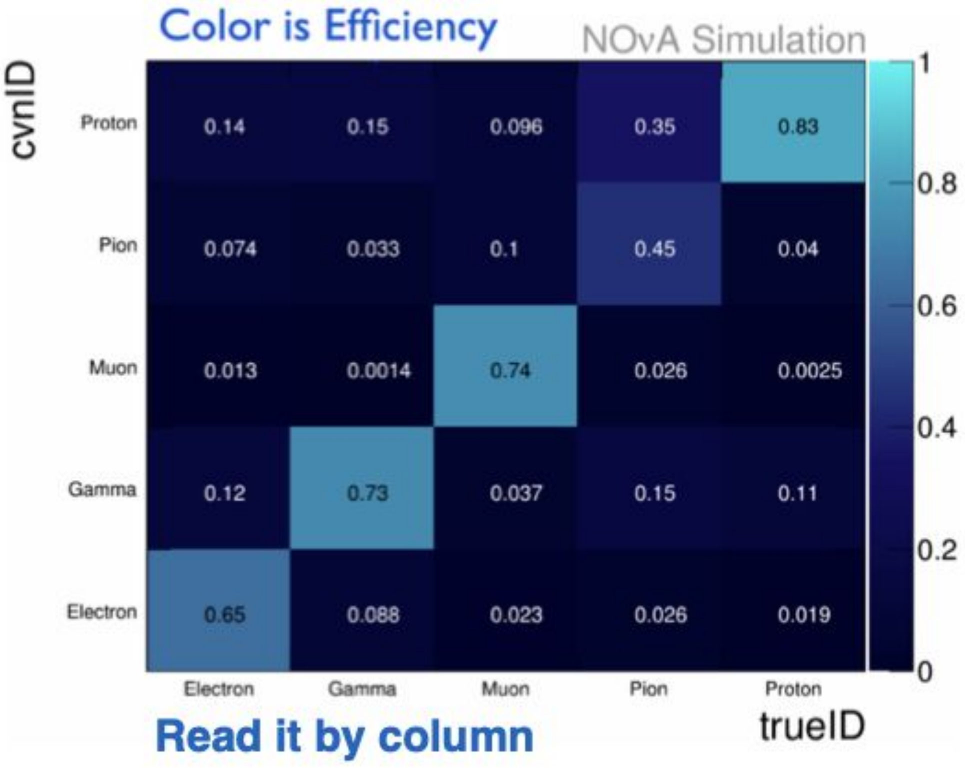


Classify particles using full event topology from both views as well as reconstructed cluster information (***4 views***)



Plots and figures courtesy of A. Tsaris, A. Himmel

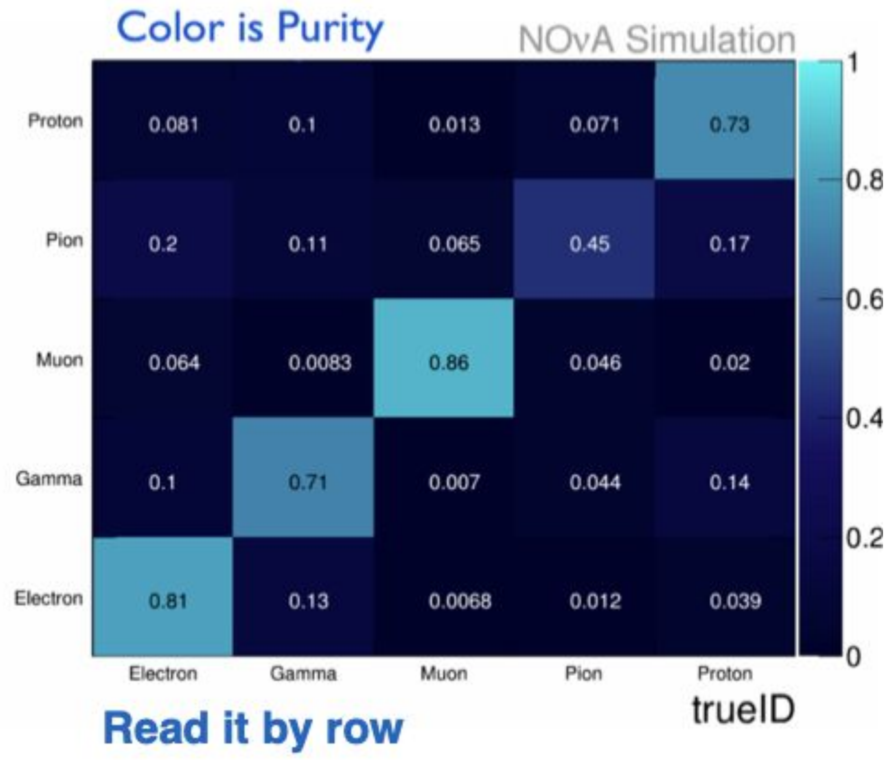
NOvA: Particle Reconstruction



Color is Efficiency

cvnID \ trueID	Electron	Gamma	Muon	Pion	Proton
Proton	0.14	0.15	0.096	0.35	0.83
Pion	0.074	0.033	0.1	0.45	0.04
Muon	0.013	0.0014	0.74	0.026	0.0025
Gamma	0.12	0.73	0.037	0.15	0.11
Electron	0.65	0.088	0.023	0.026	0.019

Read it by column



Color is Purity

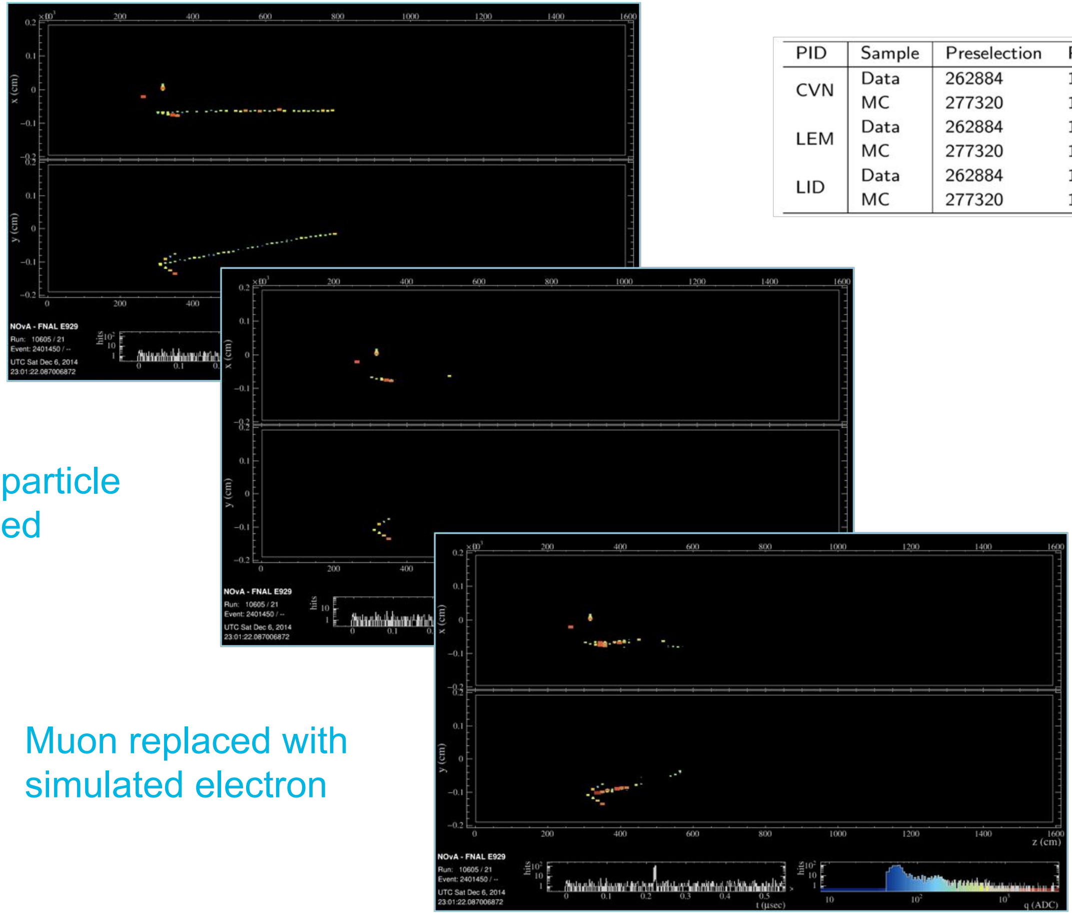
cvnID \ trueID	Electron	Gamma	Muon	Pion	Proton
Proton	0.081	0.1	0.013	0.071	0.73
Pion	0.2	0.11	0.065	0.45	0.17
Muon	0.064	0.0083	0.86	0.046	0.02
Gamma	0.1	0.71	0.007	0.044	0.14
Electron	0.81	0.13	0.0068	0.012	0.039

Read it by row

Plots and figures courtesy of A. Tsaris, A. Himmel

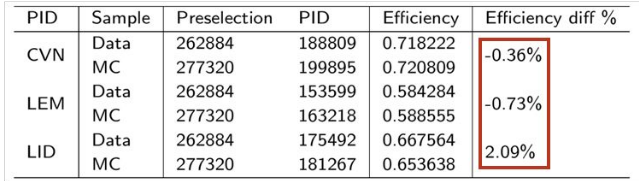
NOvA

Muon event from data

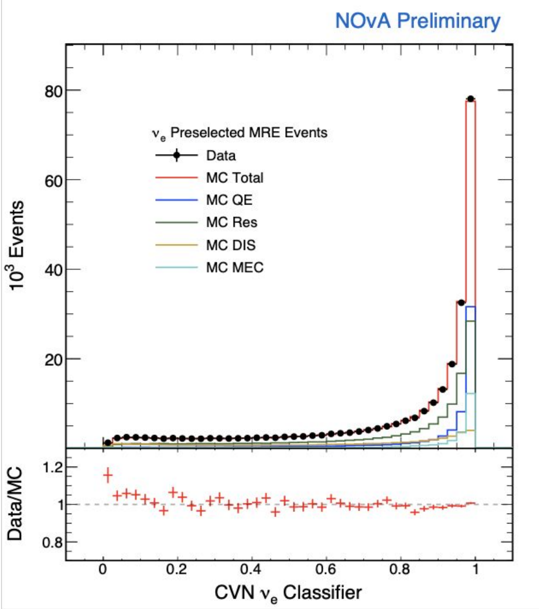


Muon particle removed

Muon replaced with simulated electron



PID	Sample	Preselection	PID	Efficiency	Efficiency diff %
CVN	Data	262884	188809	0.718222	-0.36%
	MC	277320	199895	0.720809	
LEM	Data	262884	153599	0.584284	-0.73%
	MC	277320	163218	0.588555	
LID	Data	262884	175492	0.667564	2.09%
	MC	277320	181267	0.653638	



Plots and figures courtesy of A. Tsaris, A. Himmel

Content Slide [22pt Bold]

Bullet points are optional. If preferred, only first level bullets can be used or bullets can be set to “NONE.” [18pt Regular]

- First level bullet [18pt Regular]
    - Second level bullet [16pt Regular]
        - Third level bullet [15pt Regular]
            - Fourth level bullet [14pt Regular]
                - Fifth level bullet [14pt Regular]

Comparison Slide [22pt Bold]

- First level bullet [18pt Reg]
  - Second level bullet [16pt]
    - Third level bullet [15pt]
      - Fourth level bullet [14pt]
        - Fifth level bullet [14pt]

Click to add caption text [13pt Bold]

- First level bullet [18pt Reg]
  - Second level bullet [16pt]
    - Third level bullet [15pt]
      - Fourth level bullet [14pt]
        - Fifth level bullet [14pt]

Click to add caption text [13pt Bold]

Content and Caption Slide [22pt Bold]

Click to add caption text [13pt Bold]

- First level bullet [18pt Reg]
  - – Second level bullet [16pt]
    - Third level bullet [15pt]
      - – Fourth level bullet [14pt]
        - Fifth level bullet [14pt]

Picture and Caption Slide [22pt Bold]

Click to add caption text [13pt Bold]