AI for Particle Physics: Better, Smarter, Faster

Kevin Pedro
Associate Scientist
Scientific Computing Division / Particle Physics Division
Fermilab

May 6, 2020

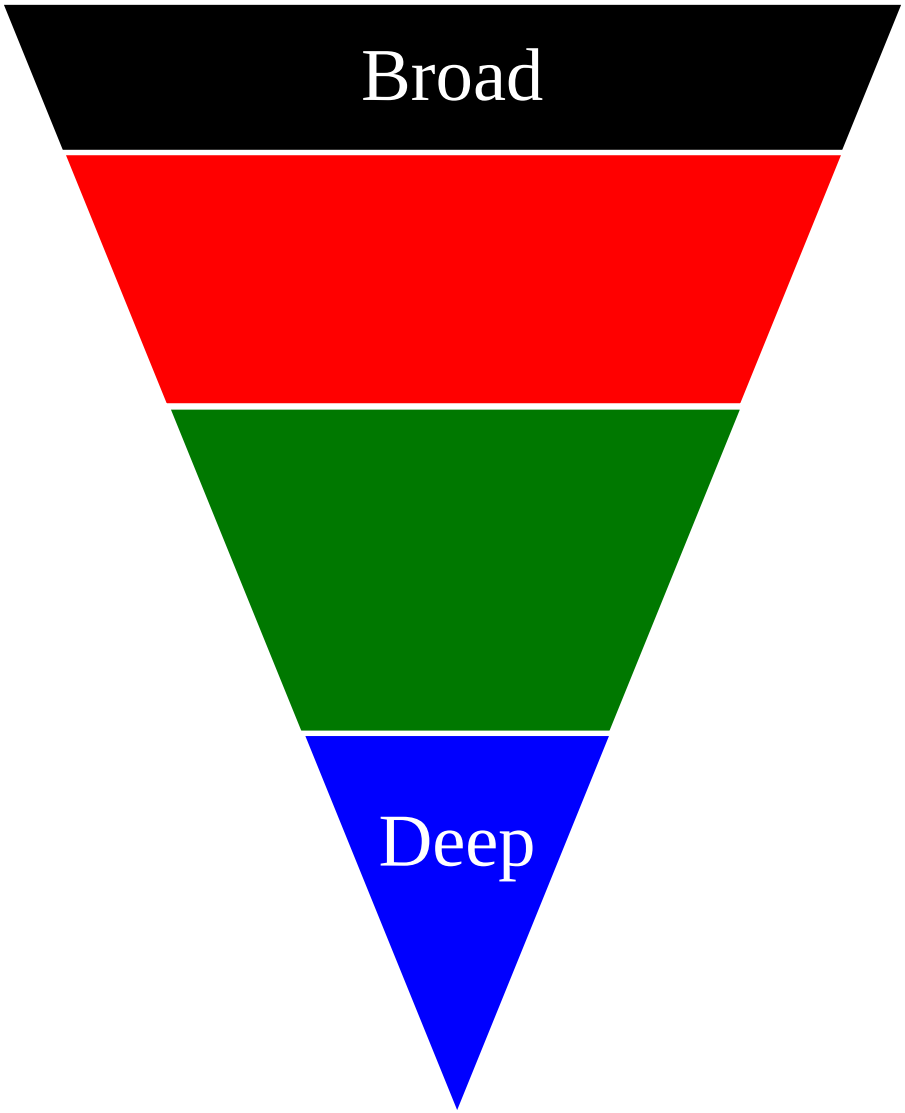
Outline

Particle Physics & AI:
- Particle detectors
- Artificial intelligence

Better:
- Recent improvements in AI
- Physics use case: tagging
- Open questions

Smarter:
- Cutting-edge R&D
- Graph-based algorithms
- Preliminary results

Faster:
- High Luminosity Upgrade
- Accelerating inference
- Coprocessors as a service



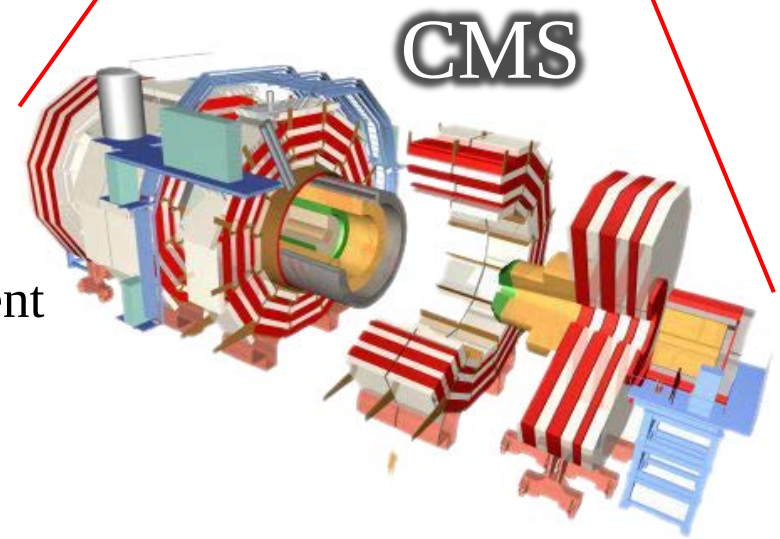
Particle Physics & AI

Collider Physics

- Largest, highest-energy particle collider
  - Circumference = 27 km (17 mi)
  - Center-of-mass energy = 13 TeV
- High data rate requires multiple levels of triggers

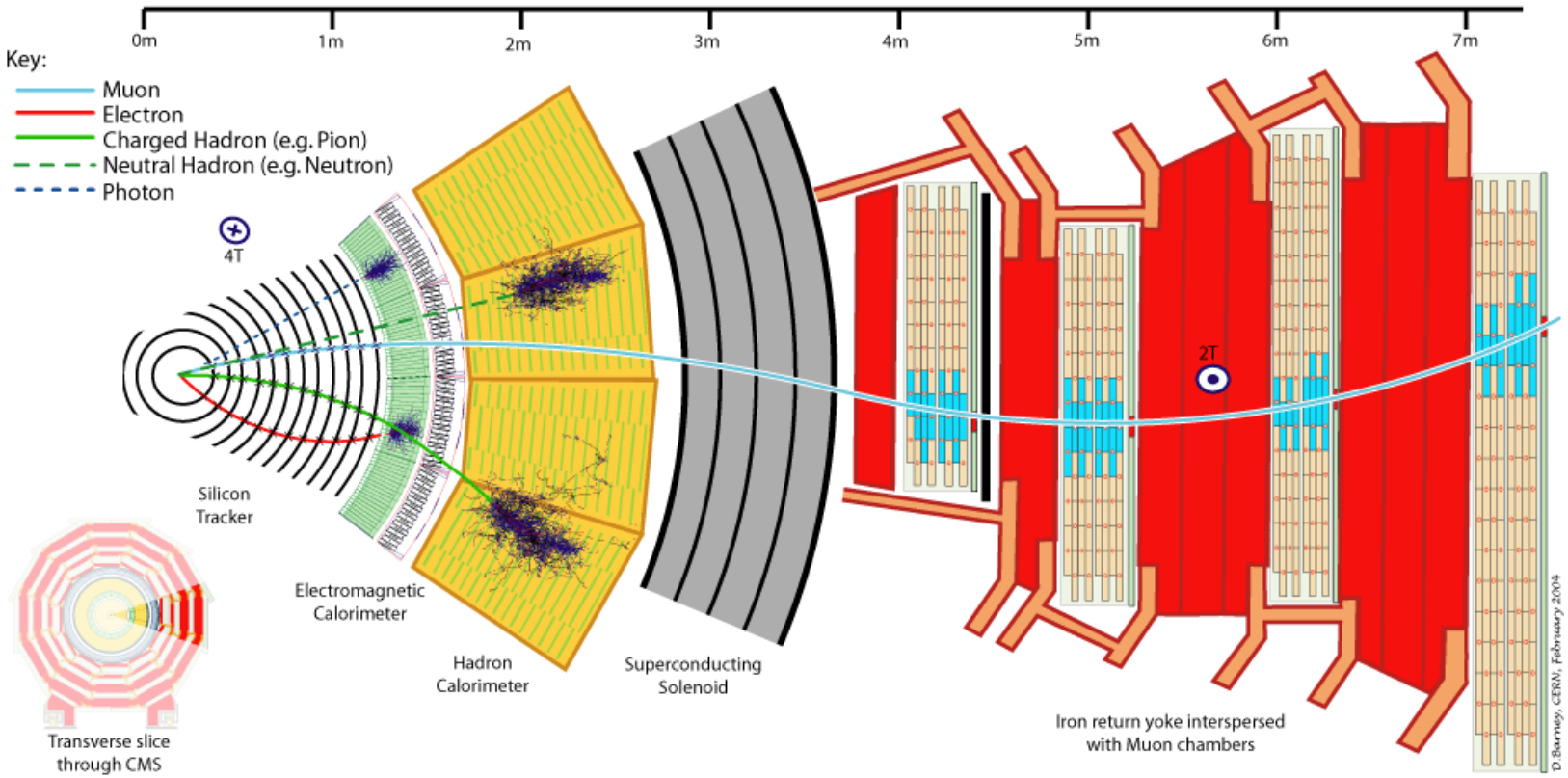
Collider Physics

- Largest, highest-energy particle collider
  - Circumference = 27 km (17 mi)
  - Center-of-mass energy = 13 TeV



➢ Focus on AI results from CMS experiment
  - Many items also applicable to ATLAS, neutrino physics, cosmology, etc.

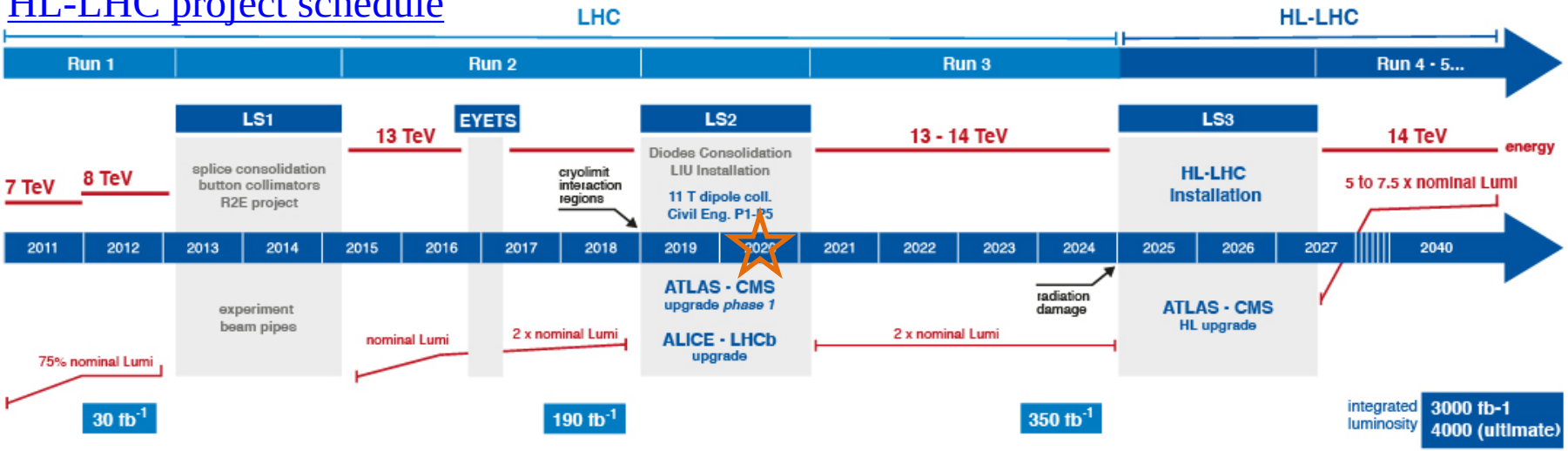
CMS Detector



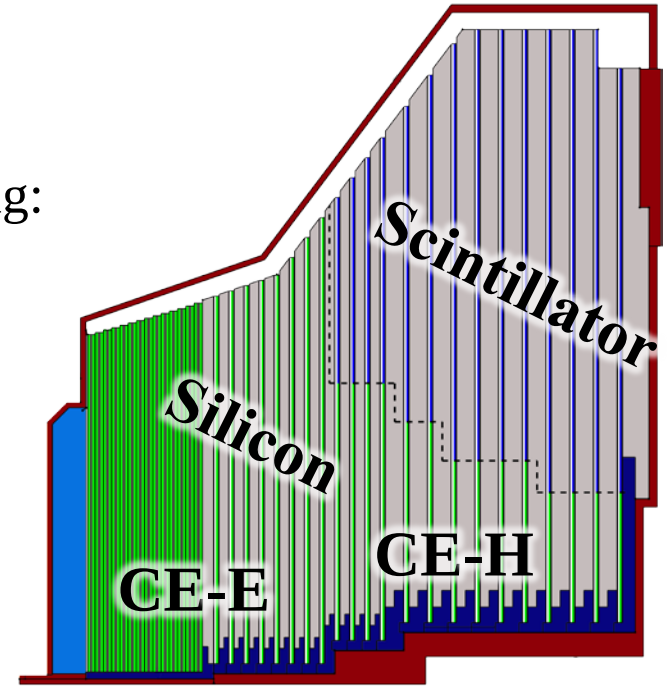
- "Hit": energy deposit in single channel
- Tracks: built from consistent hits in tracker, muon chambers
- Clusters: built from nearby hits in calorimeters
- Particles: built from linked tracks and clusters
- Jets: collimated sprays of particles

Upgrades

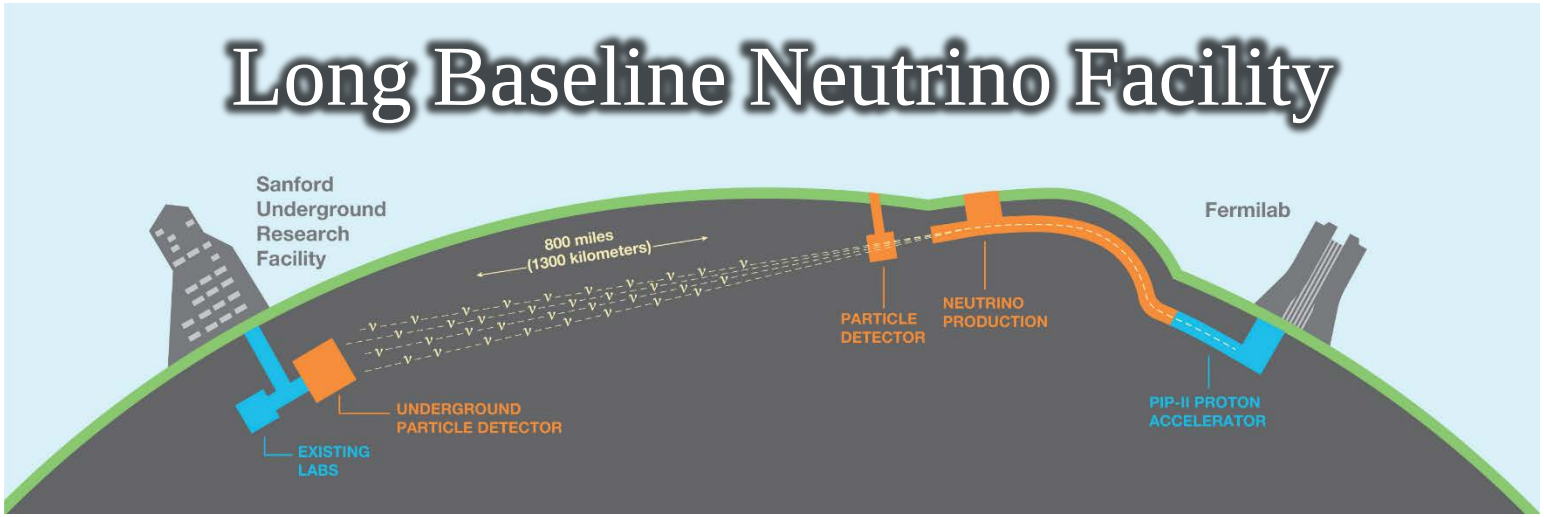
[HL-LHC project schedule](#)



- Increase in luminosity → more data!
  - Also more radiation…
- Corresponding CMS detector upgrades, including:
  - Pixel (innermost tracker):
    66M → 1947M channels
  - Outer tracker:
    9.6M → 215M channels
  - High Granularity Calorimeter (endcaps):
    85K → 6M channels



Neutrino Physics

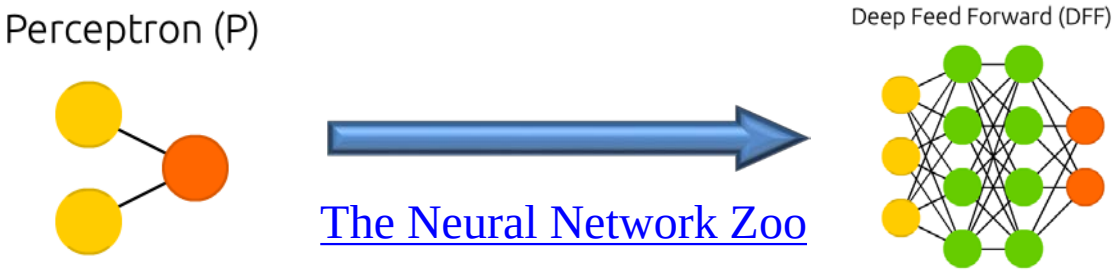


- Neutrinos interact very rarely and weakly
  - ➢ Detectors need large volume of material and long exposure
    - o Try to reduce *backgrounds* (unwanted hits) from cosmic rays, etc.
- Neutrinos have mass and therefore *oscillate* between different flavors
  - o Near and far detectors compare proportions
- Upcoming LBNF: most intense neutrino beam, 120 GeV

What is Artificial Intelligence?

> "AI is whatever hasn't been done yet."
> – Douglas Hofstadter

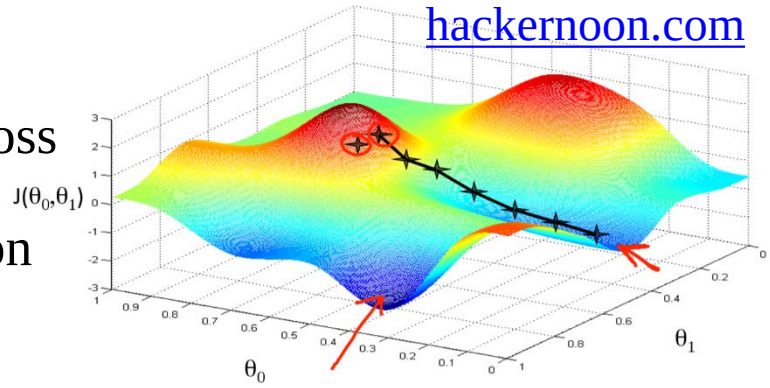
- In this colloquium: *machine learning* (ML)
- ML is *function approximation*:
  - ➢ map inputs to outputs, $\vec{x} \mapsto \vec{y}$
  - o $\vec{y} = F(\vec{x})$ unknown, probably not analytic
    → try to find approximation $\vec{y} \approx F'(\vec{x}; \vec{w})$ by optimizing *weights* $\vec{w}$
- Deep learning:
  - o Use thousands, even millions of weights
  - o Use many *layers* with intermediate *features* derived from inputs
    - ▪ More "neurons" → more multiplications

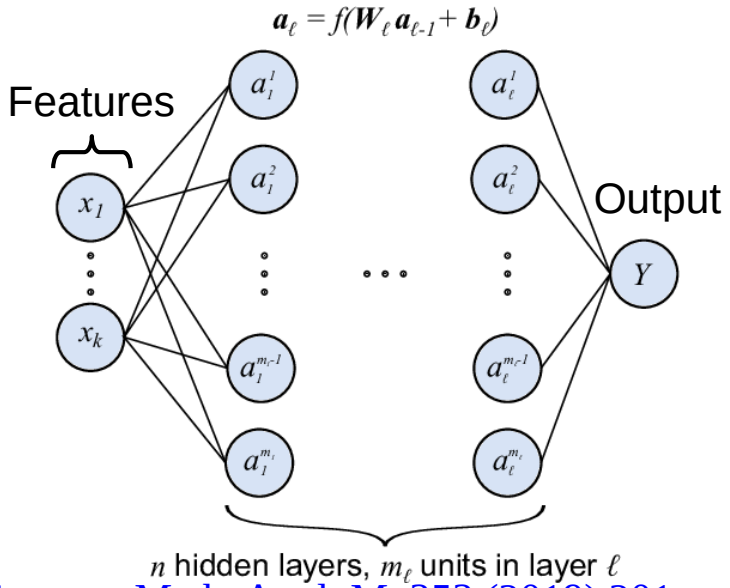


Training an AI

- Iteratively modify weights so F′ gets "closer" to $\vec{y}$ (training data)
  - "Closer" defined by a *loss function*
  - Use *gradient descent* to follow change in loss
- Keep separate datasets for testing & validation
  - Otherwise, AI could be *overtrained*

hackernoon.com





Comput. Meth. Appl. M. 353 (2019) 201

- Training is very intensive: large datasets, billions (!) of multiplications
  - GPUs are optimized for these operations
- *Inference*: applying trained AI to (new) input data to get output
  - Output: classification, regression, etc.

AI at FNAL: A Long History

FermiNews
November 18, 1988 Vol. XI, No. 21 Fermi National Accelerator Laboratory

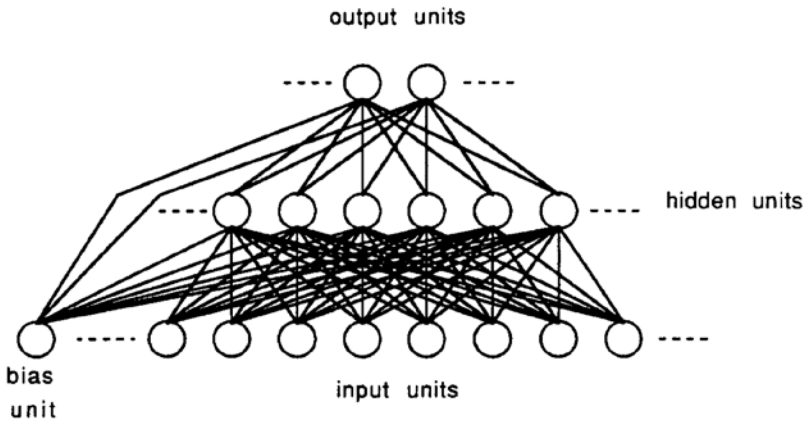
Neural Network Startup

In the past few years, there has been a tremendous resurgence in research on neural networks, the name given to arrays of single-bit, quasi-digital processors whose high level of interconnectivity resembles that of nerve cells in the brain. Neural nets seem to be good at problems that humans solve easily, but that conventional computers are notoriously bad at, such as pattern recognition and decision making based on incomplete or faulty data.

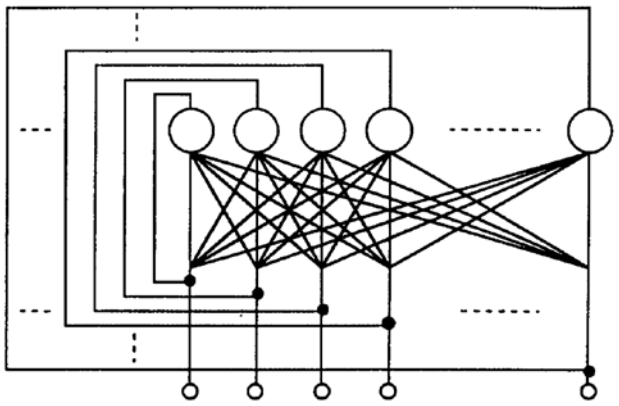
Bruce Denby, who has recently joined the Lab as a Wilson Fellow based in the Computing Department, is beginning a project to explore the possibility of using artificial neural networks and other fine-grained SIMD architecture devices in experimental triggers or offline pattern recognition engines.

Networks implemented in VLSI have demonstrated enormous speedups over conventional microprocessors for certain applications. Also, because of the high redundancy in the interconnection network, neural sets are relatively insensitive to localized faults caused by point defects in silicon substrate or by errors in the data input.

Persons wishing to find out more about neural networks should contact Bruce Denby at FNAL::DENBY or drop a note to him at MS 120. If there is sufficient interest, regular discussion sessions can be set up.



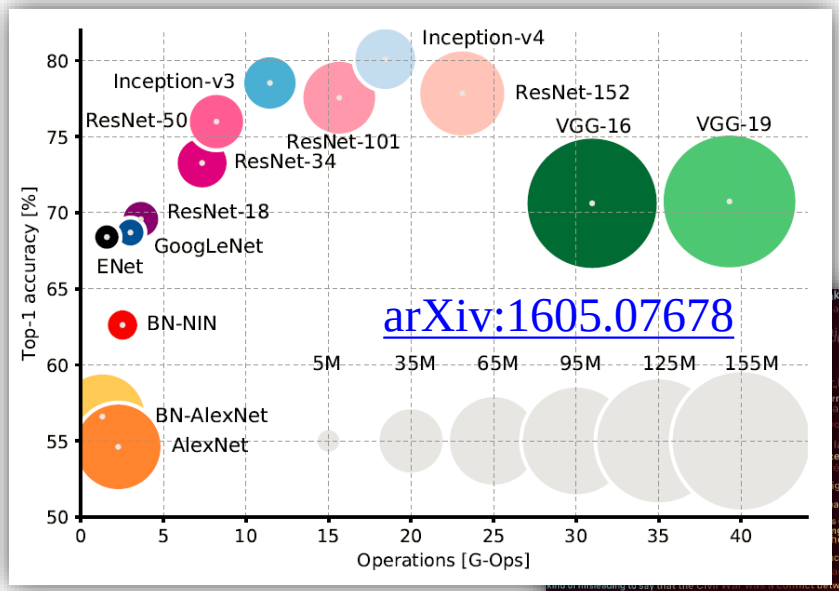
Feed forward neural network

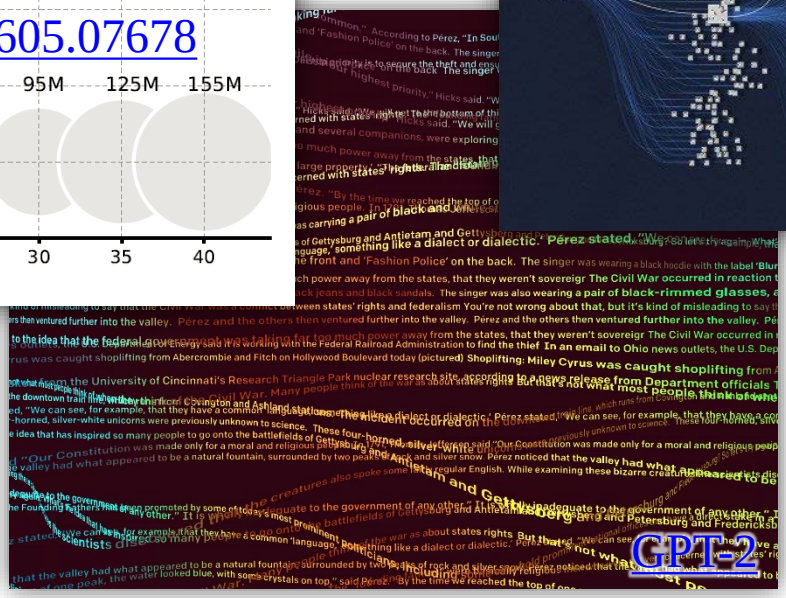


B. Denby, “Neural Network Tutorial for High Energy Physicists”, FERMILAB-Conf-90/94, May 1990

Better

AI Today

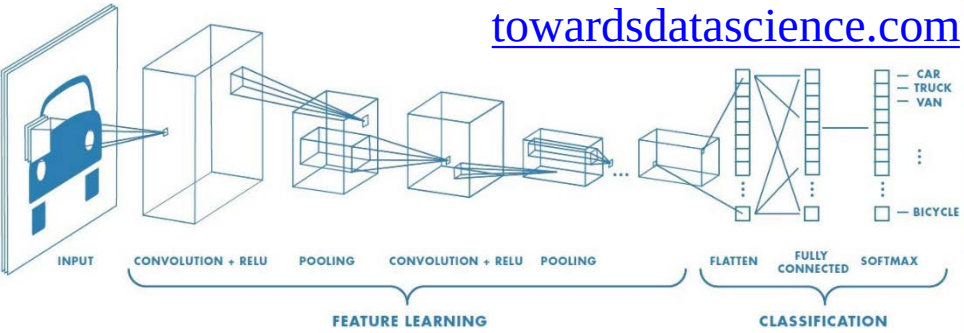




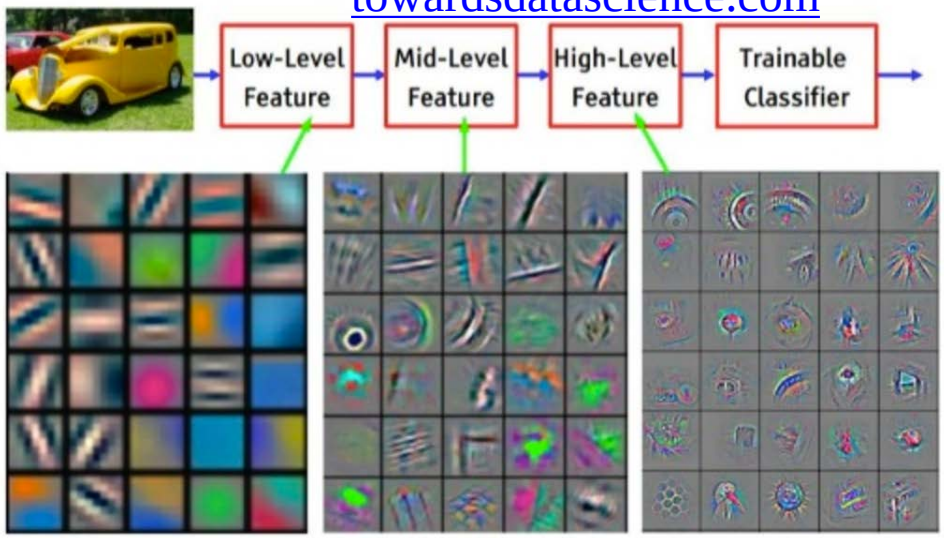
- Massive industry efforts in R&D for deep neural networks
  - Many frameworks: TensorFlow, PyTorch, MXNet, scikit-learn, etc.
- Giant leaps in image recognition, language processing, even game playing
  - Similar leaps in computational requirements…

Convolutional Neural Networks

towardsdatascience.com



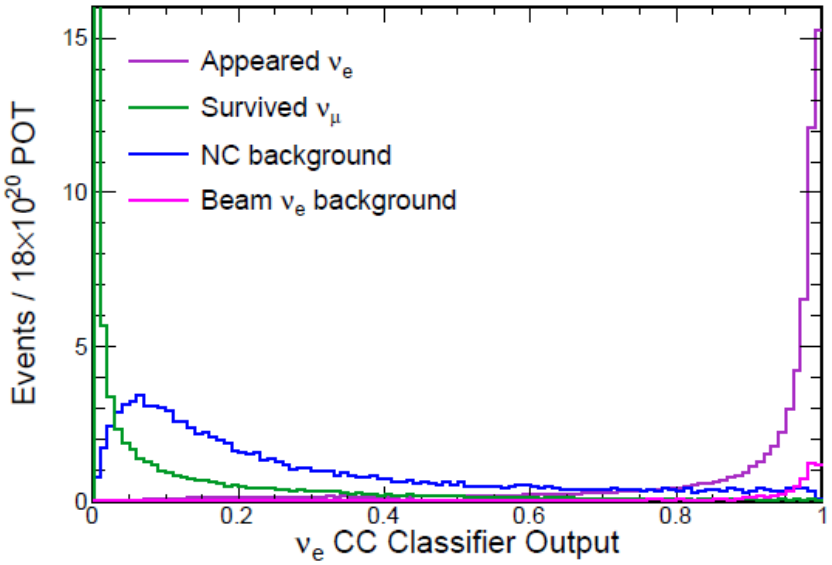
towardsdatascience.com

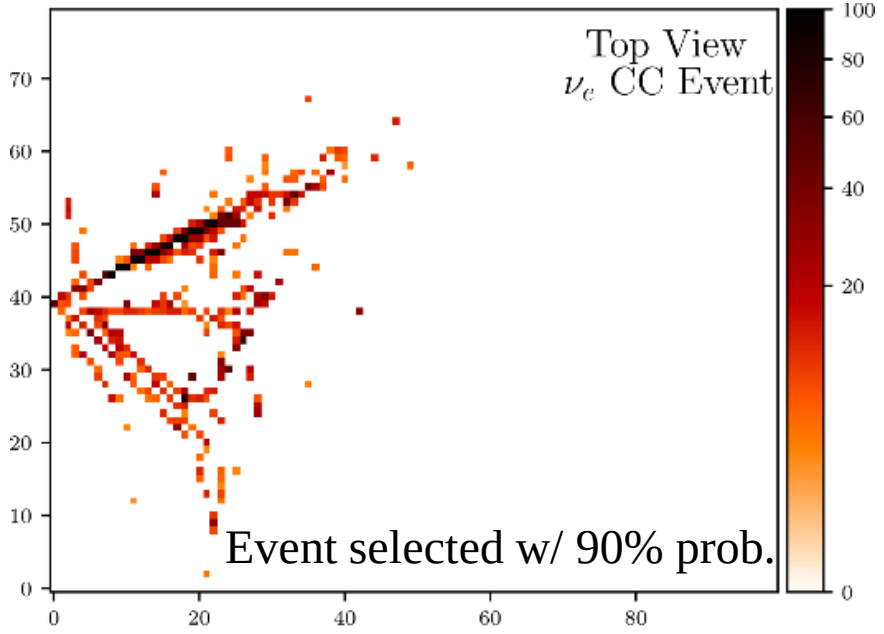


- Image recognition started modern AI revolution
- Innovation: *convolutional* neural networks (CNNs)
  - Combine neighboring pixels according to matrix of weights
    - Same convolution applied to whole image → reduce # weights
  - Derive *features* at different scales: edges, corners, etc.

CNNs for Neutrinos

- Neutrino detector data naturally image-like





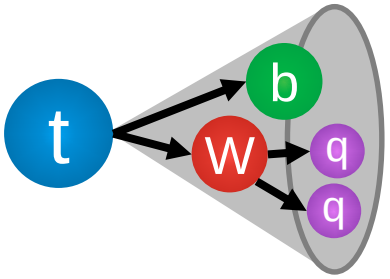
- NOvA was first particle physics experiment to publish† result from CNN
- ResNet50 can distinguish charged current events from cosmic background

† JINST 11 (2016) P09001, Phys. Rev. Lett. 118 (2017) 231801

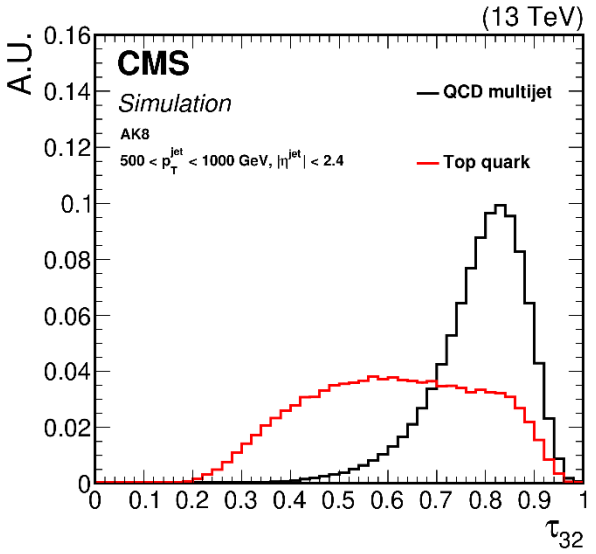
Collider Physics Example: Tagging

Prototypical case: tagging top quarks

- Many models of physics beyond the standard model (SM) include new particles that can decay to top quarks
  - Heavy new particles → boosted top quarks, decay products merge into a single wide jet

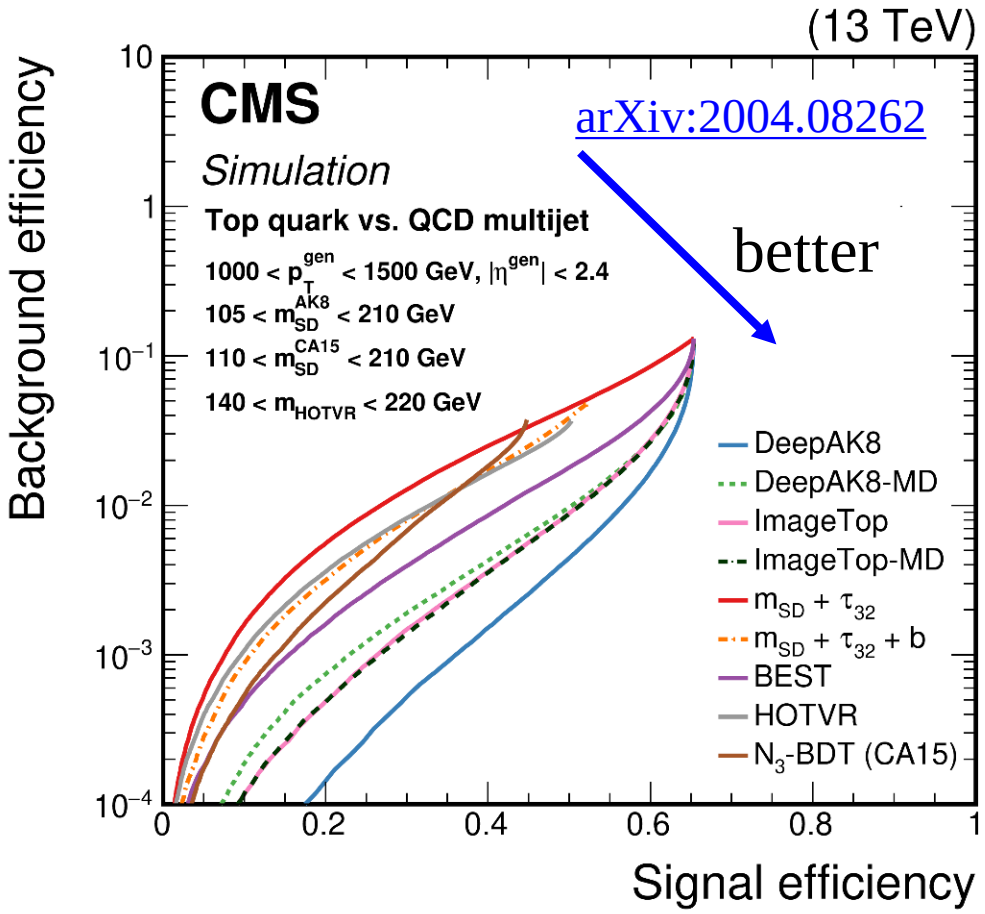


- Clear signature of new physics
  - But background events (e.g. SM QCD) have much higher rate
- Traditionally identified using jet substructure:
  - $\vec{x}$ = Nsubjettiness, groomed jet mass, etc. ("expert" variables)
  - $\vec{y}$ = top quark or QCD
  - $F(\vec{x})$ = selection criteria
    - Example: $\tau_{32} < 0.6$



AI for Tagging

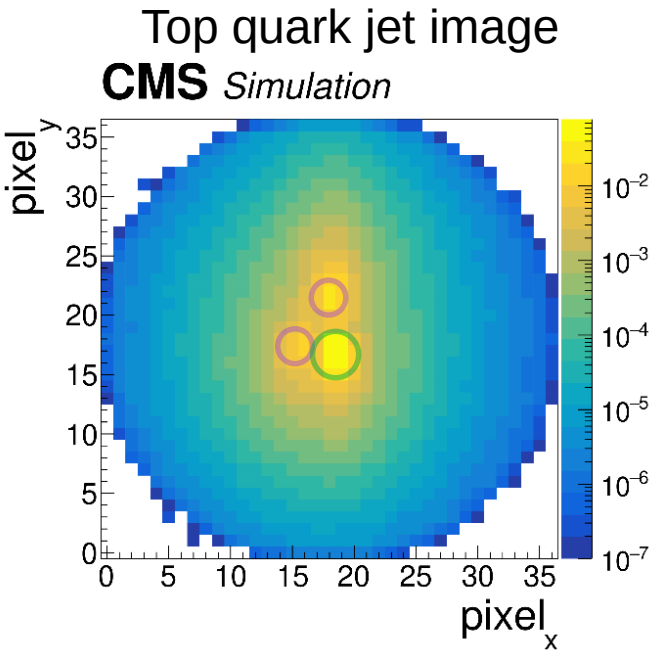
- Can machine learning algorithms do a better job than experts?
- Usual progression:
  - Combine expert variables in boosted decision tree (BDT)
  - Combine expert variables in deep neural network (DNN)
  - Use lower-level variables (reconstructed tracks, particles, etc.) in DNN
  - Use more advanced neural network architectures

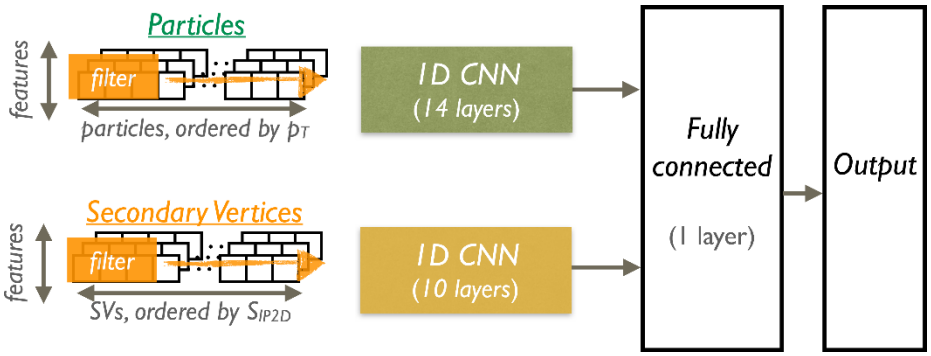


Different Approaches

ImageTop: build “image” out of jet constituents

- Pros: leverage ubiquitous industry tools for image recognition, convolutional neural networks (CNNs)
- Cons: some information is lost (jets aren’t “really” 2D images)



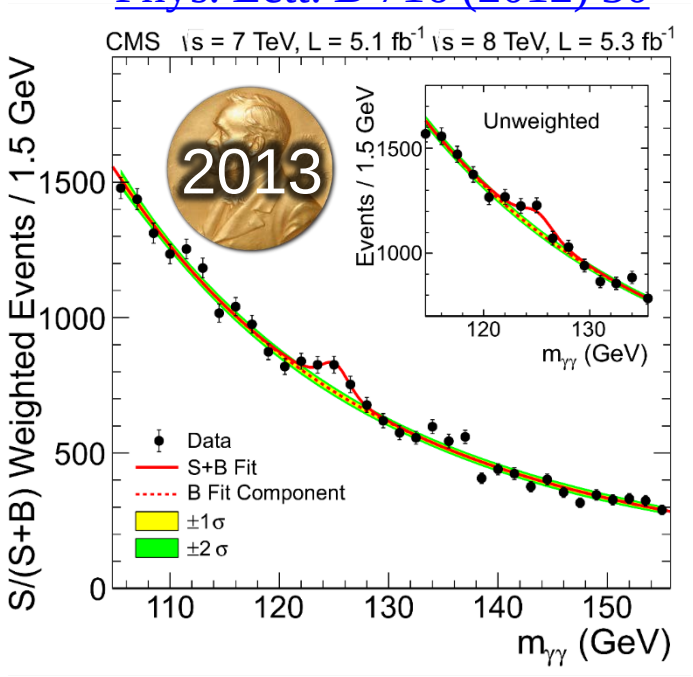


DeepAK8: learn from particle, vertex variables directly

- Pros: keep more information
- Cons: 1D convolutions may not fully capture all relationships between quantities

AI Enables Discovery

Phys. Lett. B 716 (2012) 30



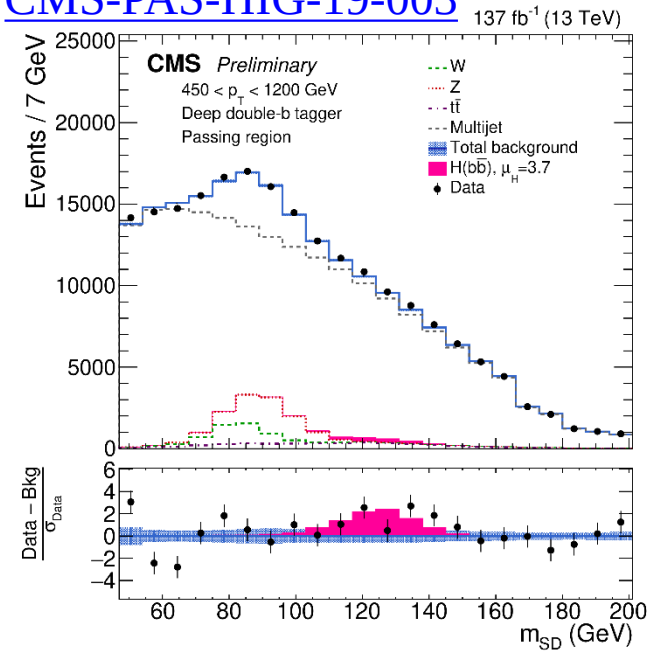
Higgs → γγ: rare process, but clear signature and clean background

- Boosted decision trees critical for Higgs discovery in 2012
- Event-level classification enhances resonance

Higgs → $b\bar{b}$: most common decay, but huge background

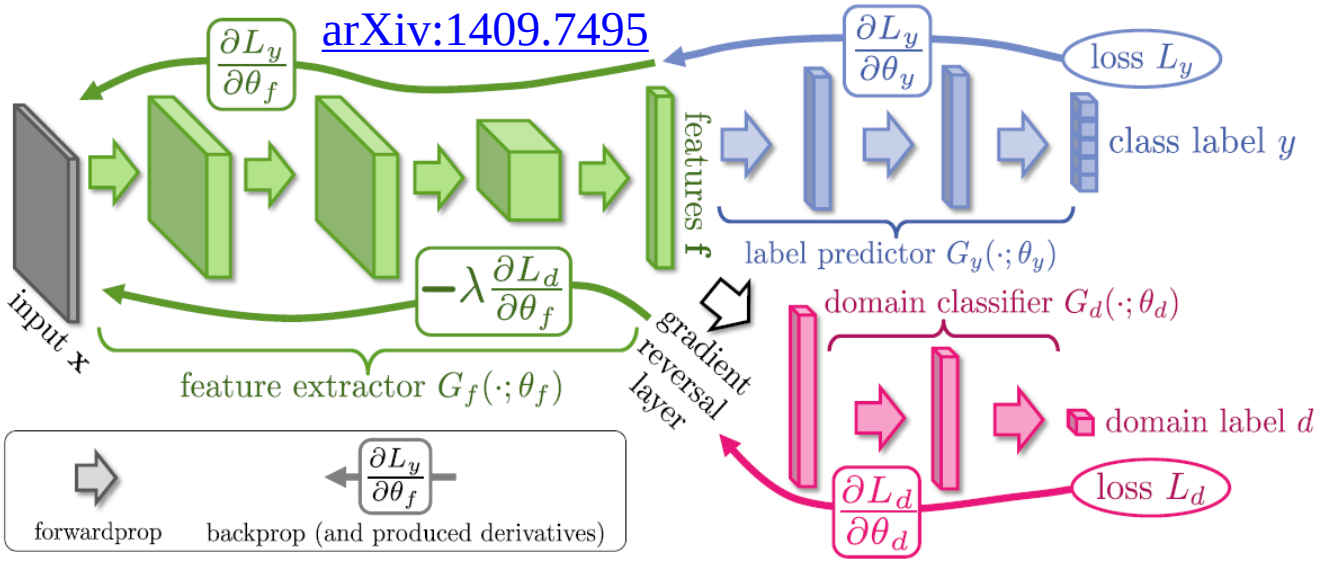
- Dedicated DNN tags boosted Higgs with two secondary vertices
- Once thought impossible
  - Now 2.5σ evidence in 2020

CMS-PAS-HIG-19-003



Open Questions

- How to handle differences between data and simulation?
  - Gradient reversal (domain adaptation) very promising
    - Can also help avoid other unwanted behavior, e.g. mass dependence
- How to explain what the network learns?
  - Should probably be an entire academic field in itself
  - See e.g. The Building Blocks of Interpretability for CNNs
- Far from an exhaustive list (of algorithms or questions)

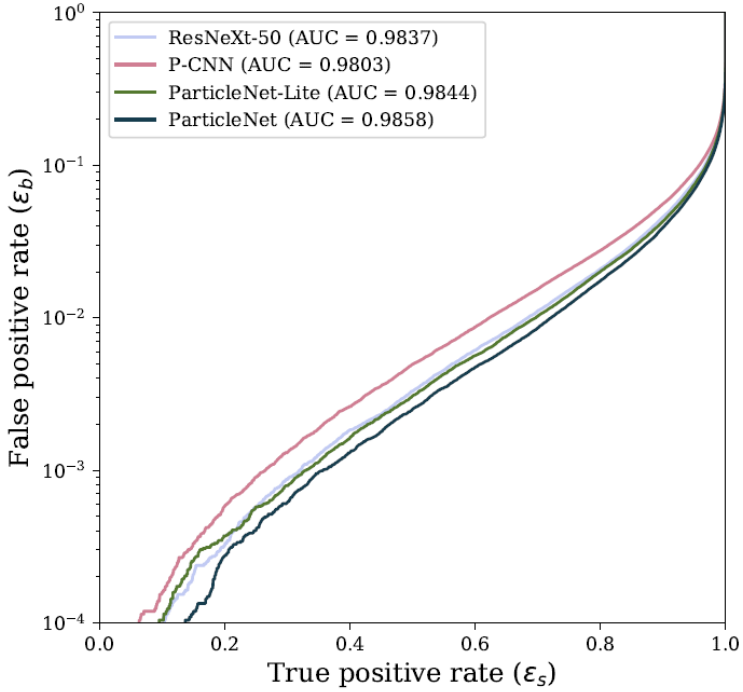


Smarter

Cutting Edge Tagging

Tagger	AUC	Acc	$1/\varepsilon_B$*	# Params
P-CNN	0.980	0.930	759	348K
ResNet50†	0.983	0.935	1000	25M
ResNeXt	0.984	0.936	1147	1.46M
ParticleNet	0.986	0.940	1615	366K

† CSBS 3 (2019) 13 * ($\varepsilon_S = 0.3$)

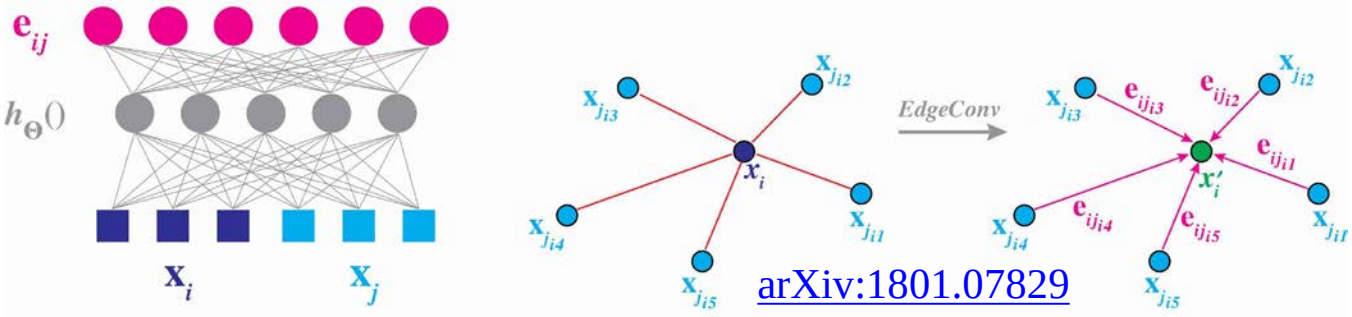


Phys. Rev. D 101 (2020) 056019
see also SciPost Phys. 7 (2019) 014

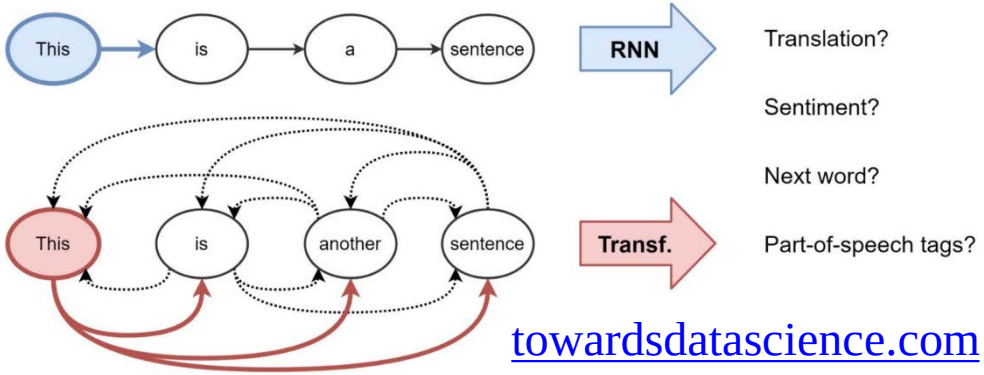
- P-CNN = simplified DeepAK8
- Apply massive image recognition networks (from industry) for significant gains
  - But regular grids unnatural for collider data
    - sparse occupancy, varying geometry, etc.
- ParticleNet does even better with far fewer parameters…
  - But more operations: 3–4× ResNeXt

All Roads Lead to Graphs

- Generalize convolutions → *message passing* w/ graphs (*nodes* & *edges*)
  - Derive new features for node x_i using neighbors x_j
  - Can even assign features to edges

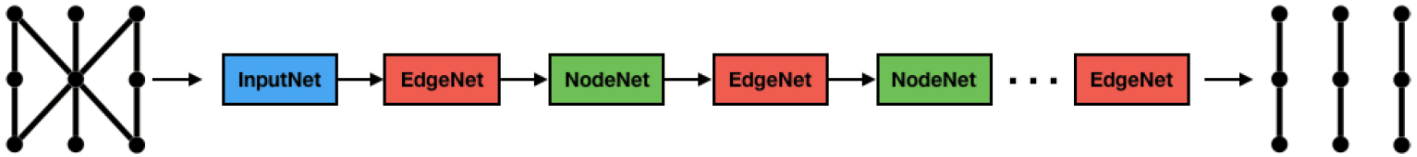


- Aside: recurrent networks (RNNs) for language processing now supplanted by “Transformers” that use “attention”
  - These are just graphs!



Graph Networks (GNNs) for Physics

- ParticleNet (leading top tagger) uses “point cloud” (also a GNN)
  - Also called “interaction networks”, “graph CNNs”, etc.
  - Same techniques applicable to many other tasks…



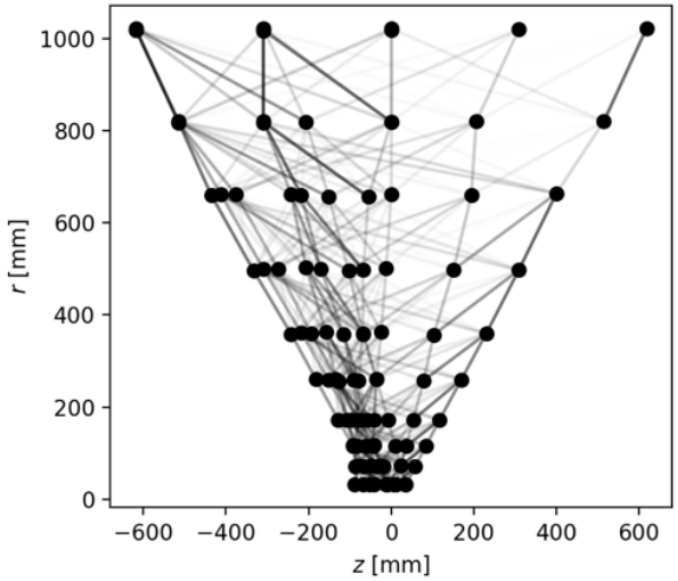
Input graph Output graph

- Most fundamental problems in event reconstruction:
  tracking, *vertexing*, *clustering*

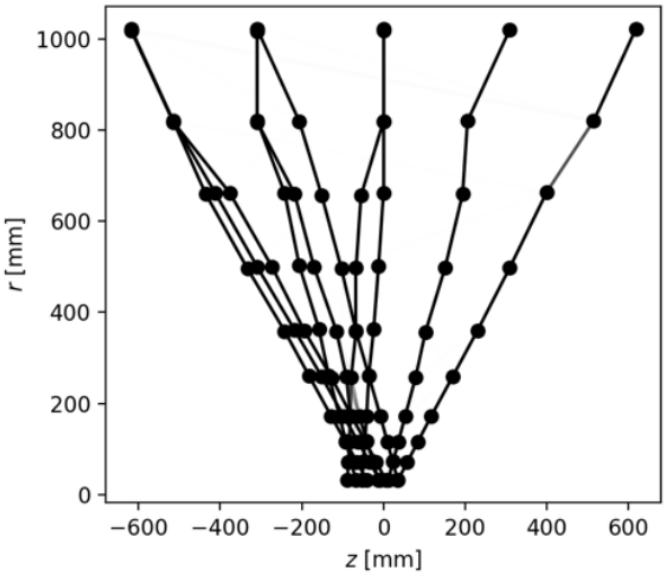
➢ How to associate detector hits with other detector hits

  - Detector geometry very important!

GNNs for Tracking

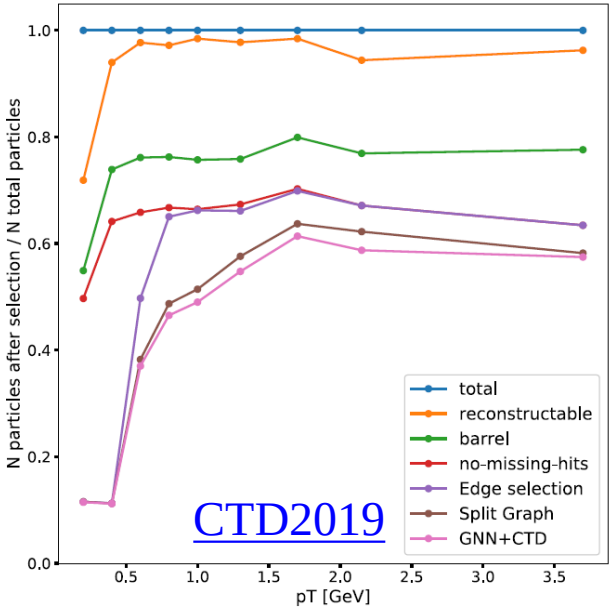


GNN



CTD2018

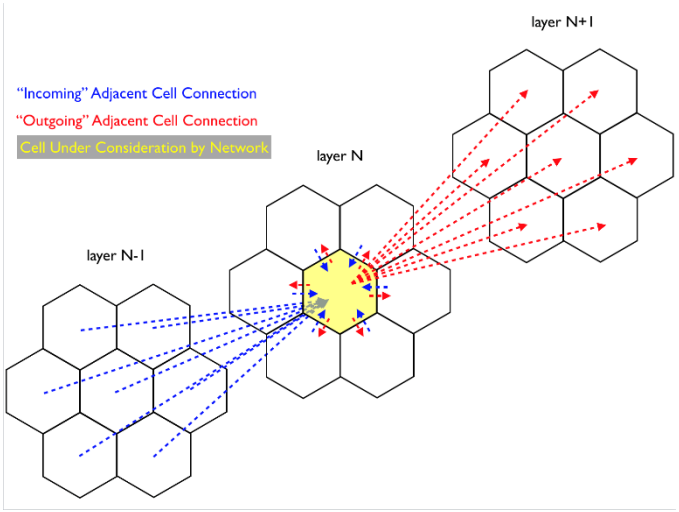
- First application to reconstruction: tracking
- Start w/ possible connections between hits
- *Edge classification*:
  GNN decides which edges are correct
- Work in progress: 97% efficient

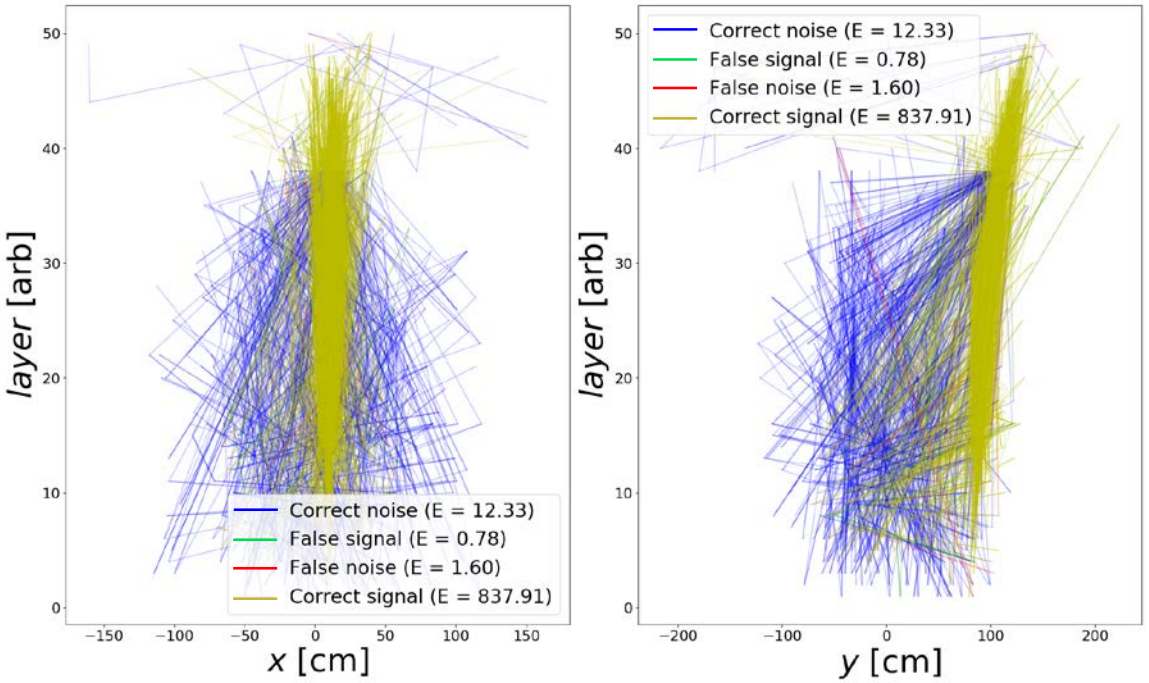


CTD2019

GNNs for Clustering

- CMS upgrades (~2026) include integrated endcap High Granularity Calorimeter
  - o Hexagonal wafers increase silicon yield
    - ➢ Especially non-grid-like geometry



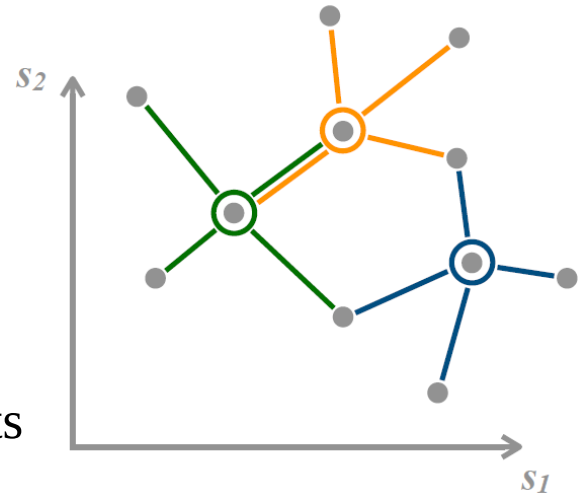


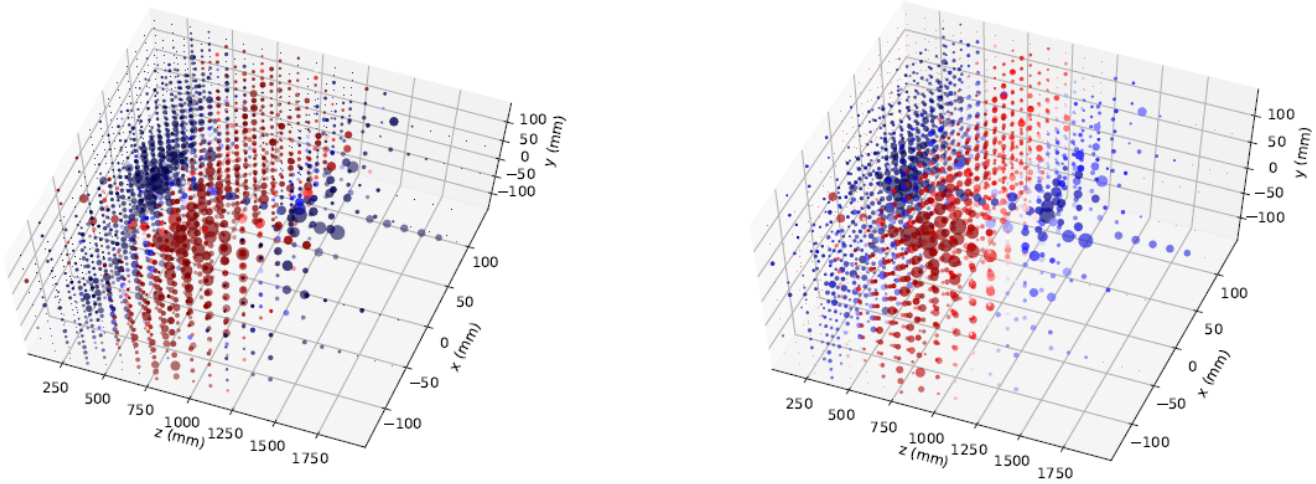
- Edge classification works well for clustering
- Charged pion: >90% efficiency to find correct edges (99% for photons and muons)
  - o 98–99% correct energy assignment

NeurIPS (ML4PS) 2019

Edge Determination

- Default edge assignment: use k-nearest neighbors or similar algorithm (based on detector geometry)
- GravNet: edges determined dynamically
  - k-nearest neighbors using *derived features*
    - ➢ GNN optimizes latent space to associate detector hits
- Open question: how to handle unknown number of clusters?



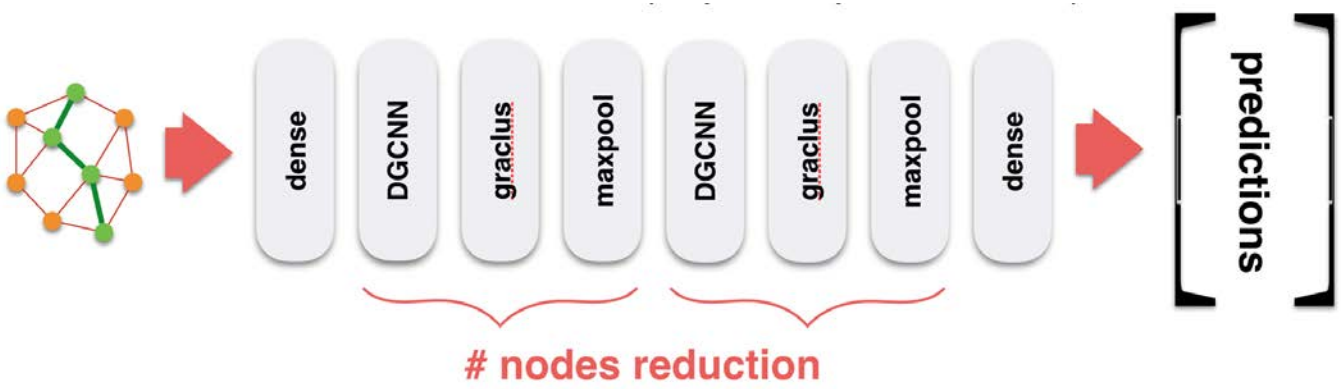


(a) Truth (b) Reconstructed

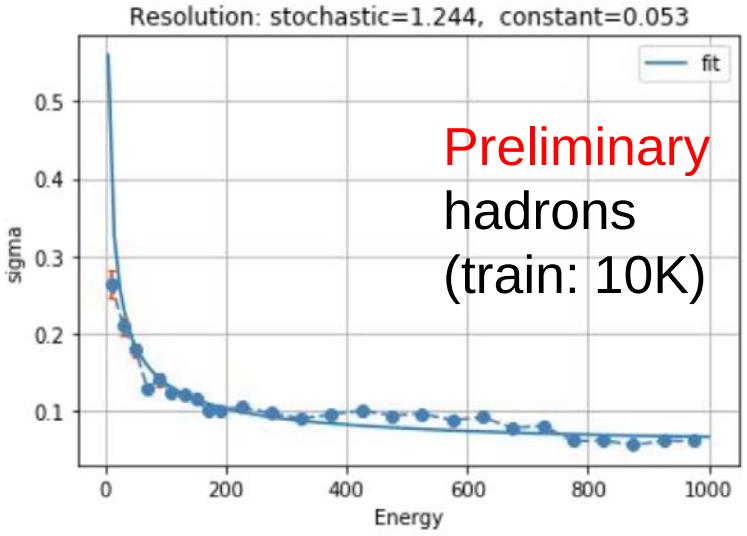
Eur. Phys. J. C 79 (2019) 608

Graphs for Calibration

- *Calibration*: another fundamental problem in physics
  - Raw measurements usually have some bias
- Dynamic reduction network: arXiv:2003.08013



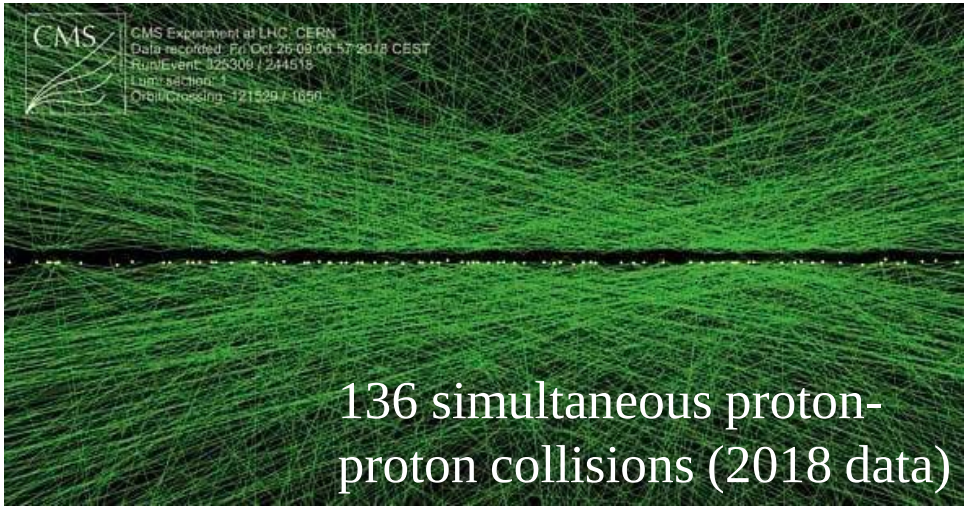
- Preliminary results (hadron resolution)
  - ➢ competitive w/ expert algorithms
  - Approach still being refined
  - Also studying industry benchmarks such as MNIST



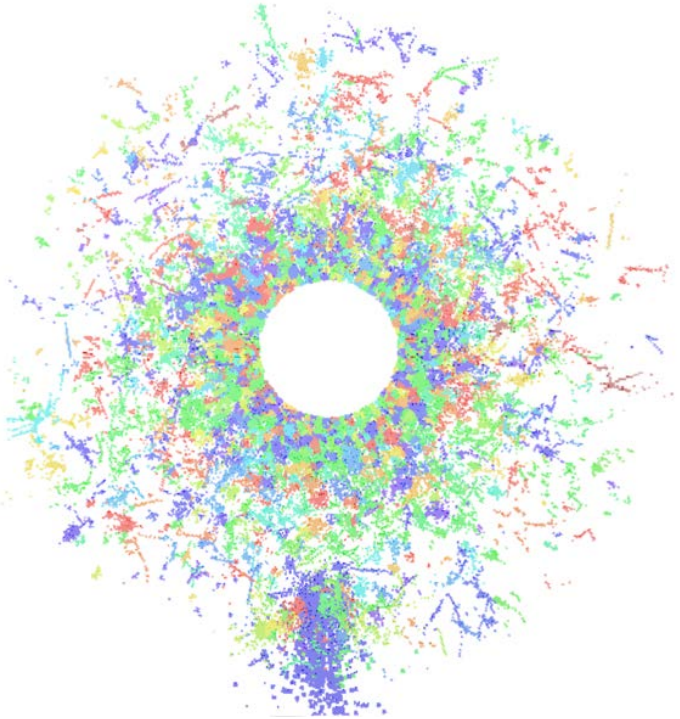
Faster

Computing for AI

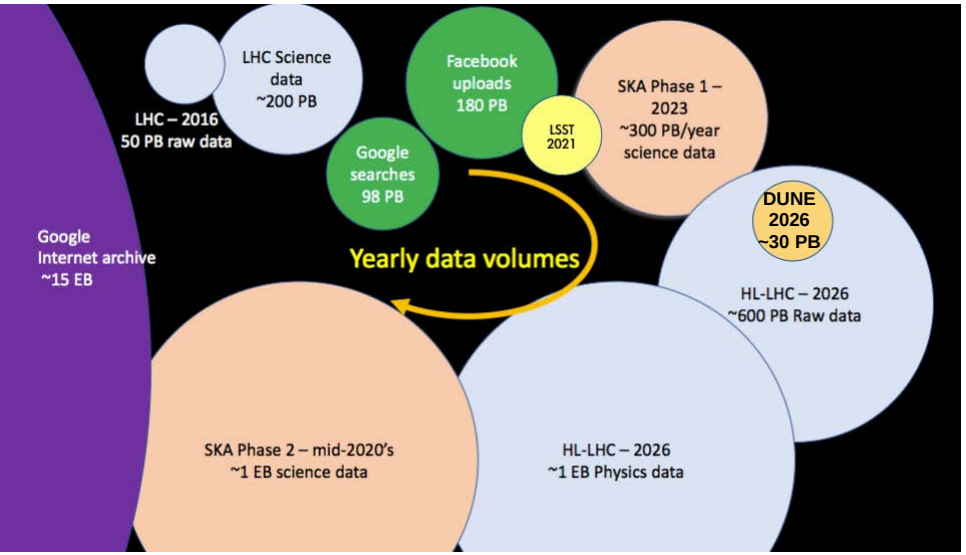
- AI has significant impacts on physics:
  - Helps us do things we couldn't do before
    - e.g. tag top and bottom quarks with unprecedented accuracy
  - Helps us do better at fundamental problems
    - Tracking, clustering, calibration, etc.
- But can we afford to keep doing all of this?
  - HL-LHC is just around the corner…

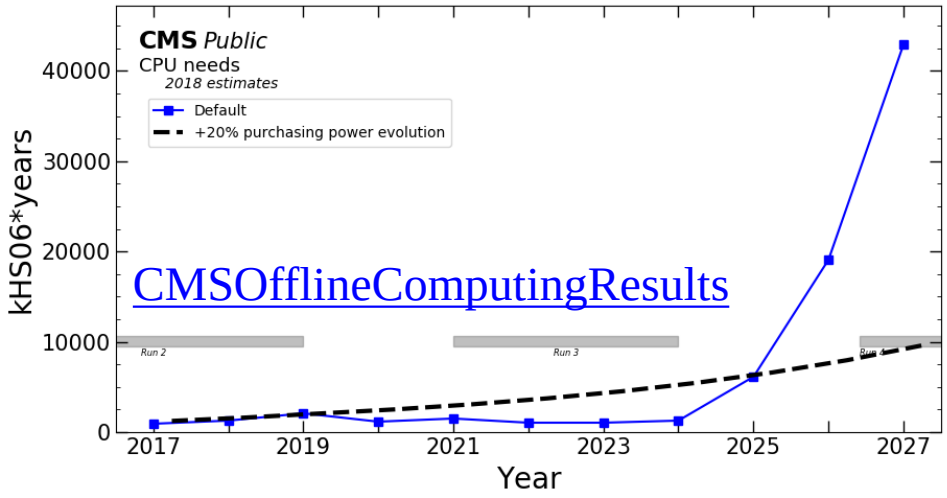


136 simultaneous proton-proton collisions (2018 data)


HGCal simulation, 200 simultaneous pp collisions

More Data, More Problems





- HL-LHC vital statistics:
  - 10× data vs. Run 2/3
  - 200 simultaneous collisions vs. ~30 in Run 2
  - Detector upgrades: 15–65× increase in channels

➢ More data and more complexity

  - DUNE, LSST, SKA will provide similarly huge datasets

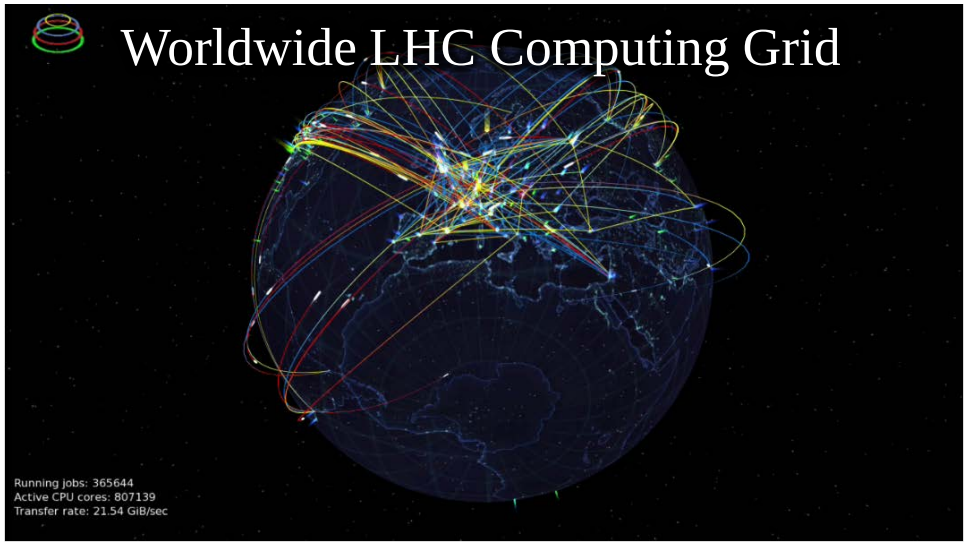
❖ Data volumes will approach scale of Google and Facebook

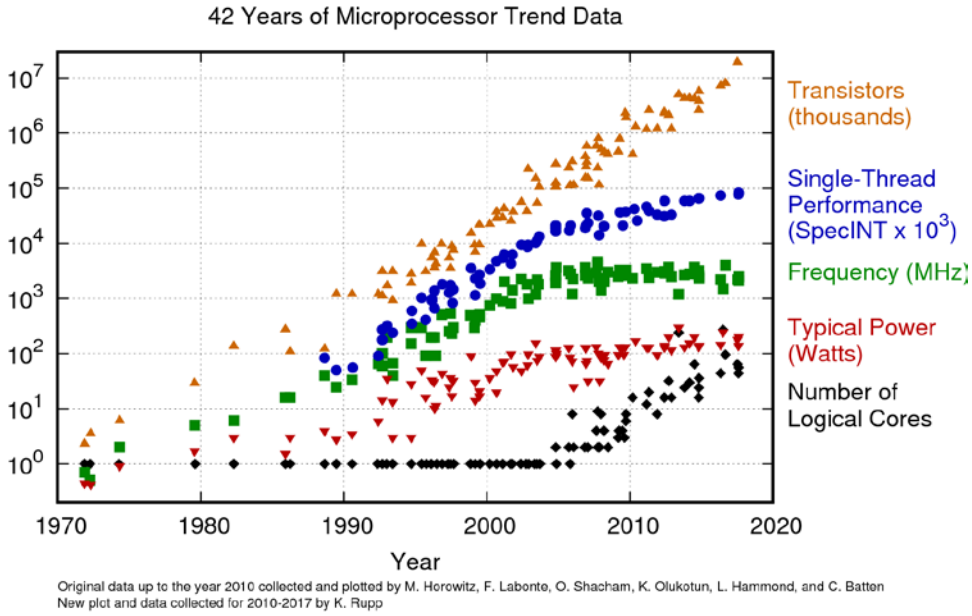
  - But computing resources won't…

CPU Stagnation

WLCG provides:

- 42 countries
- 170 computing centers
- > 2 million tasks/day
- 1 million CPU cores
- 1 exabyte of storage



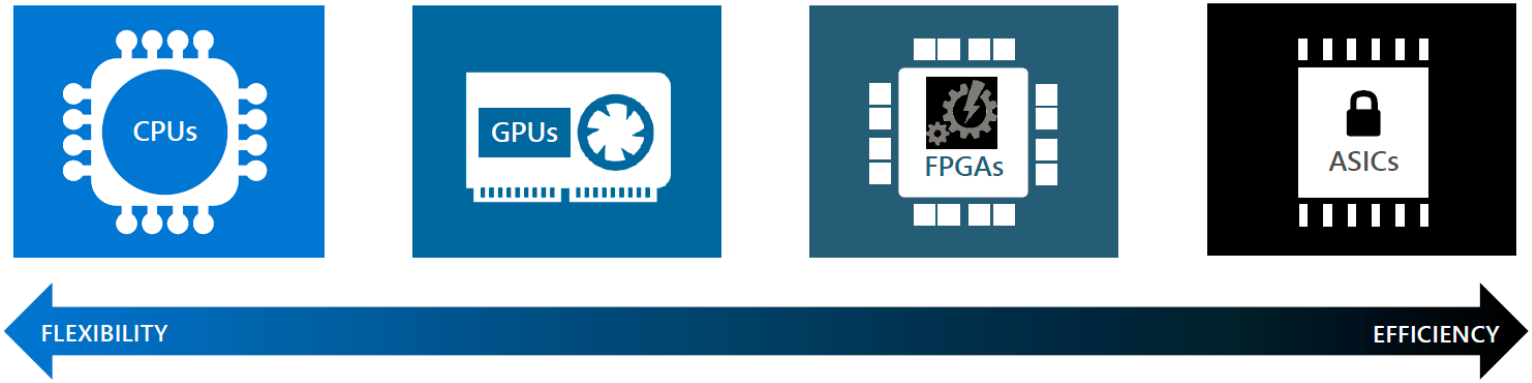


- Moore's Law continues
  - But without Dennard scaling

➢ Single-thread performance can't keep up with accelerator intensity

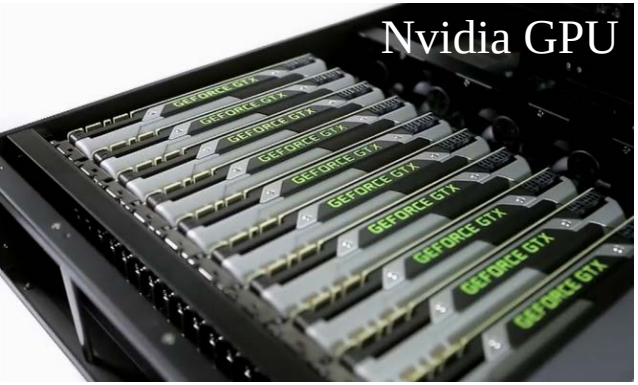
- Projected shortfalls 2–10×, depending on assumptions

Heterogeneous Revolution

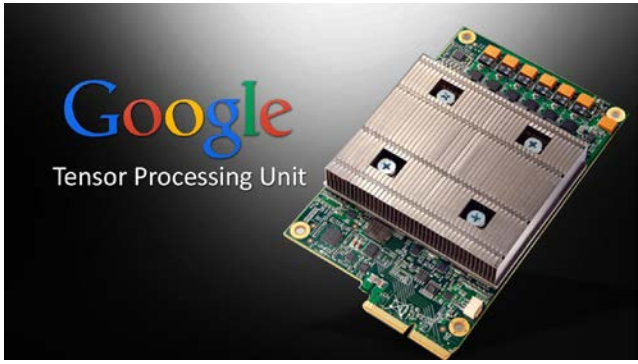


- New coprocessors provide efficiency at expense of flexibility
  - GPU: execute serial instructions on massive data
  - FPGA: spatial computing (execute many instructions simultaneously)

➢ Luckily, optimized for machine learning!

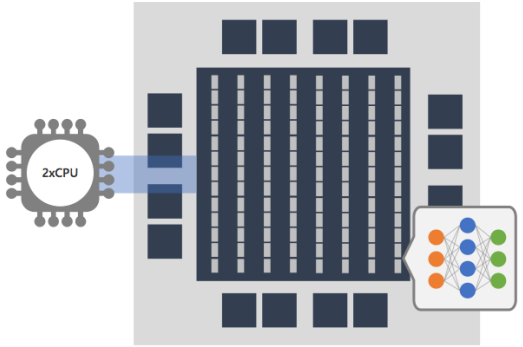


Microsoft FPGA



AI & Coprocessors: Two Great Tastes…

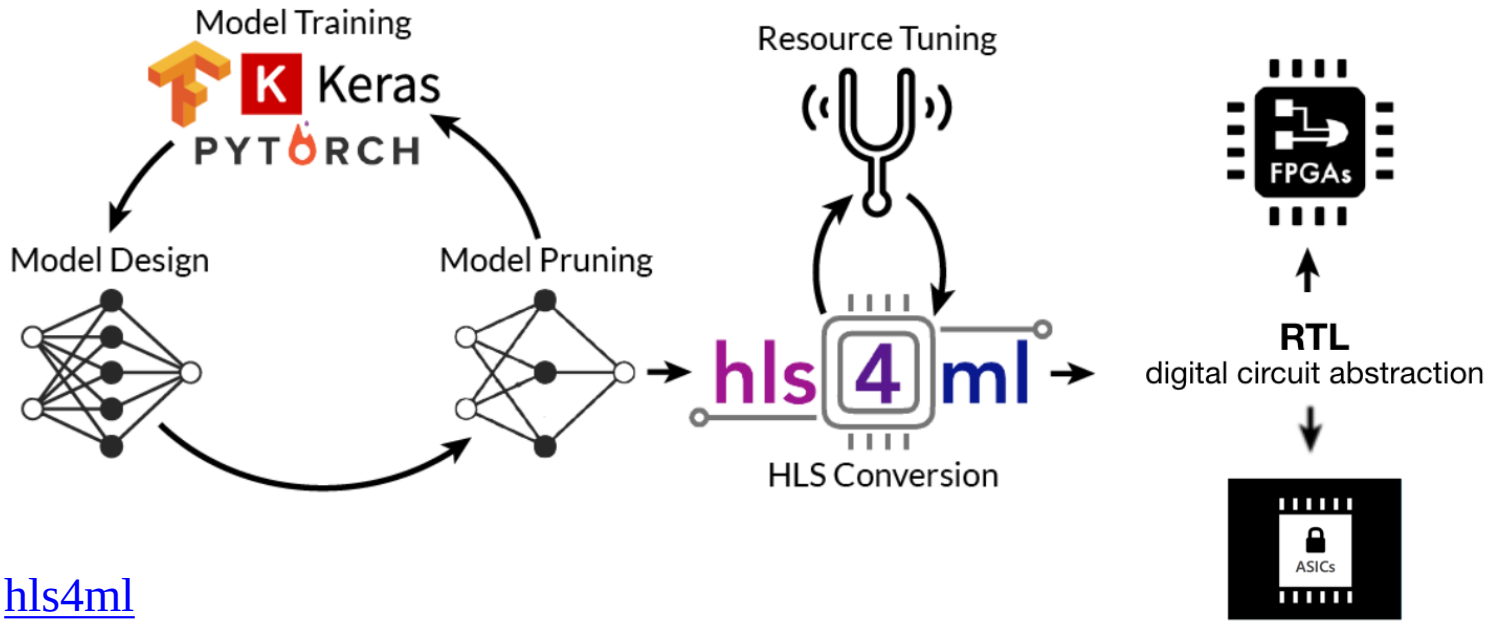
- Just adding new tagging algorithms doesn't speed up reconstruction
  - ResNet50 inference on CPU: ~1 sec/image
- Focus on *replacing* classical algorithms with AI
  - Fundamental problems (clustering, etc.) involve comparing all detector hits to all other detector hits
  - $O(N^2)$ operation, can be reduced to O(NlogN) w/ clever techniques
  - AI inference is O(N) → *much* better scaling w/ detector occupancy
- Use coprocessors to accelerate AI inference
  - GPUs also useful for training, but training only uses subset of data
  - Inference must be performed for every event (*billions*, at least)



AI for Triggers

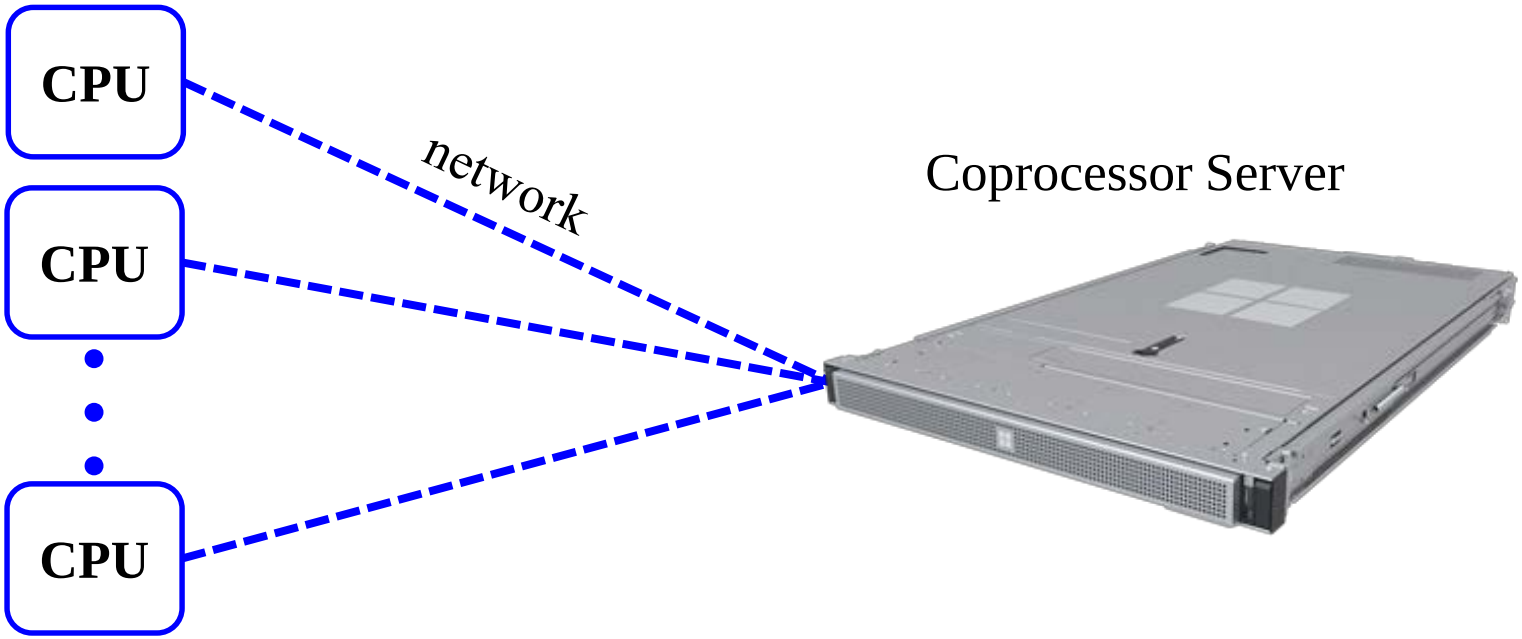
- CMS L1 trigger uses FPGAs to satisfy extreme latency requirements (~1 μs)
- **hls4ml**: open-source package
  - Optimize ML algorithms to run efficiently on FPGAs
  - Handles BDTs, various DNN architectures

➢ Planned for use during LHC Run 3



hls4ml

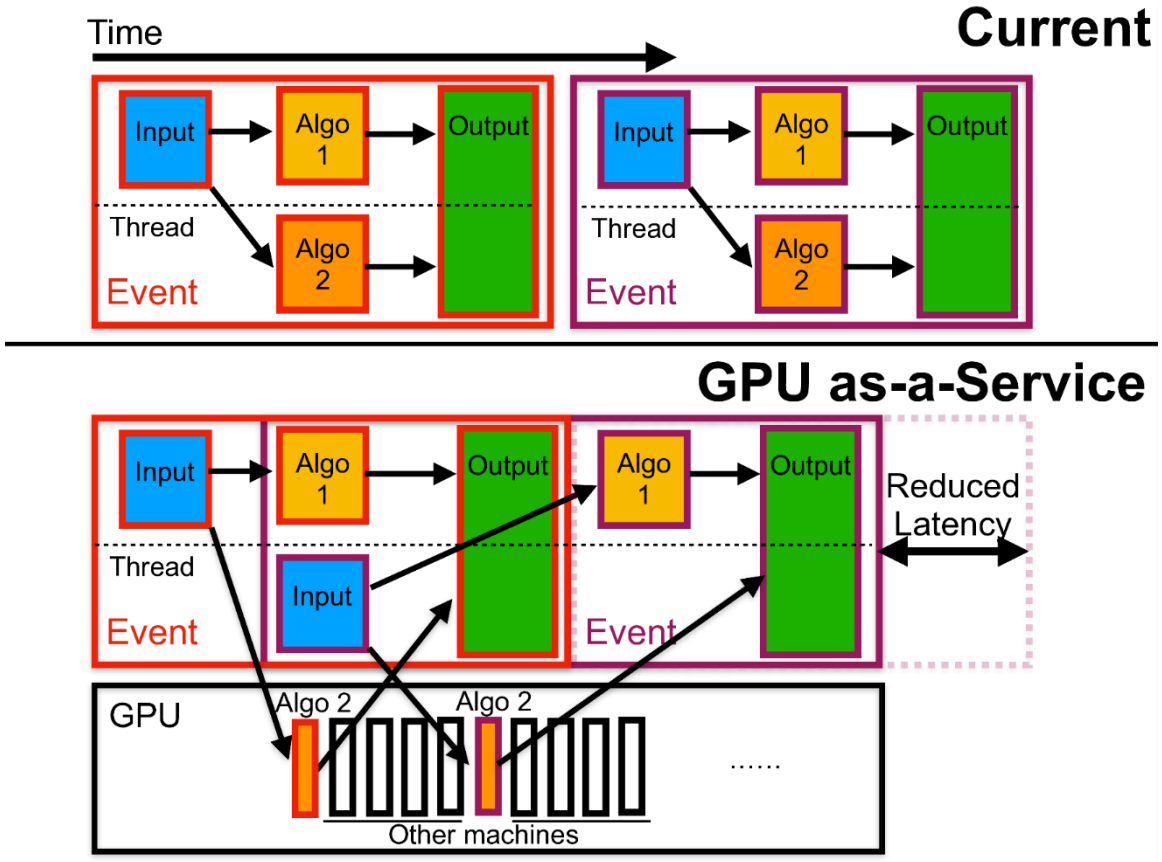
Inference As A Service



- Offline computing: looser latency requirements
- Multiple CPUs send inference requests to coprocessor server
- Ensures optimal utilization of GPUs/FPGAs, along with flexibility
- One coprocessor could serve ~100 CPUs
  - Depending on conditions and requirements:
    latency, bandwidth, memory, inference time, etc.
  - Much more cost effective than buying 1 GPU for every CPU in the grid…

LHC Computing Model

- Inference as a service naturally fits into existing computing model
- Reconstruction process involves 100s of algorithms
  - Only a few worth accelerating



- Most efficient method: asynchronous, non-blocking calls
  - Enabled by task-based multithreading
- CPU can do other work while inference request is ongoing
  - Significantly reduces impact of network latency

SONIC Approach

- SONIC (Services for Optimized Network Inference on Coprocessors): inference as a service in experiment software frameworks

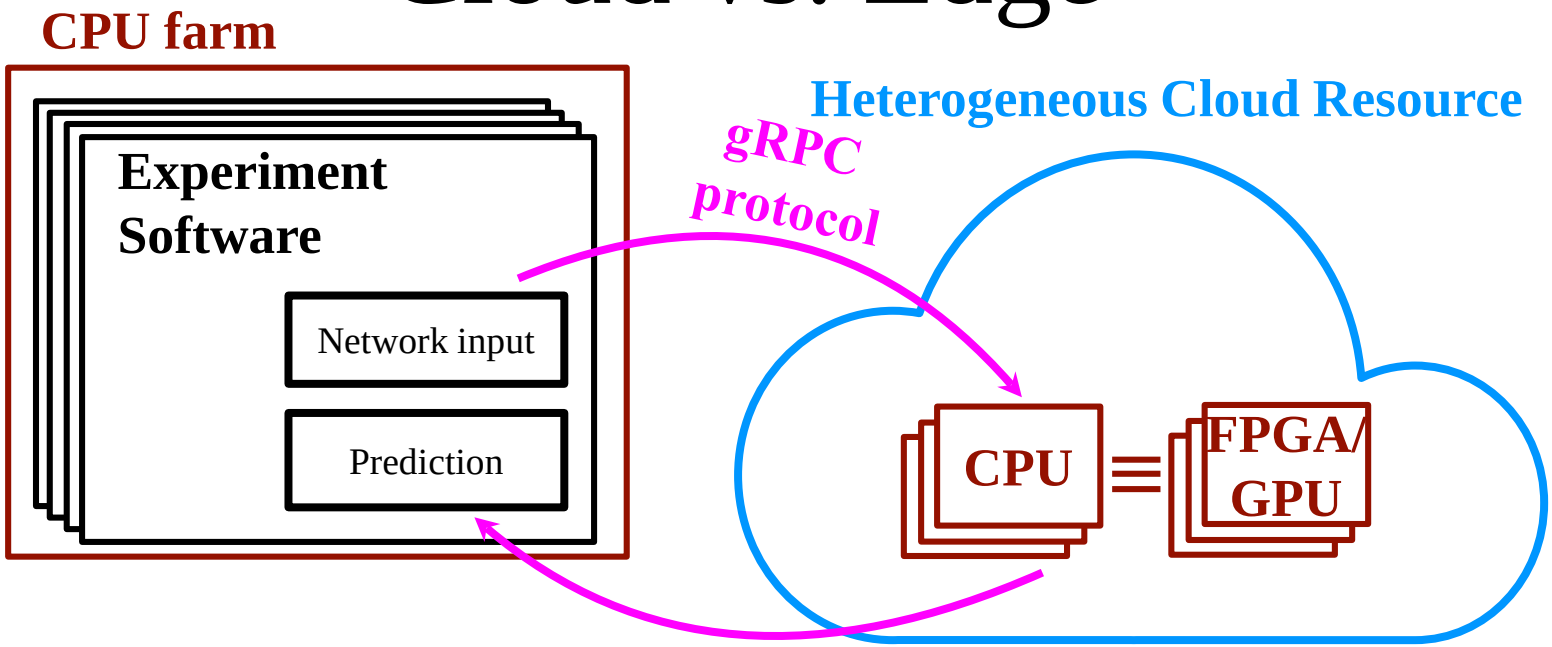
✓ Use industry tools:

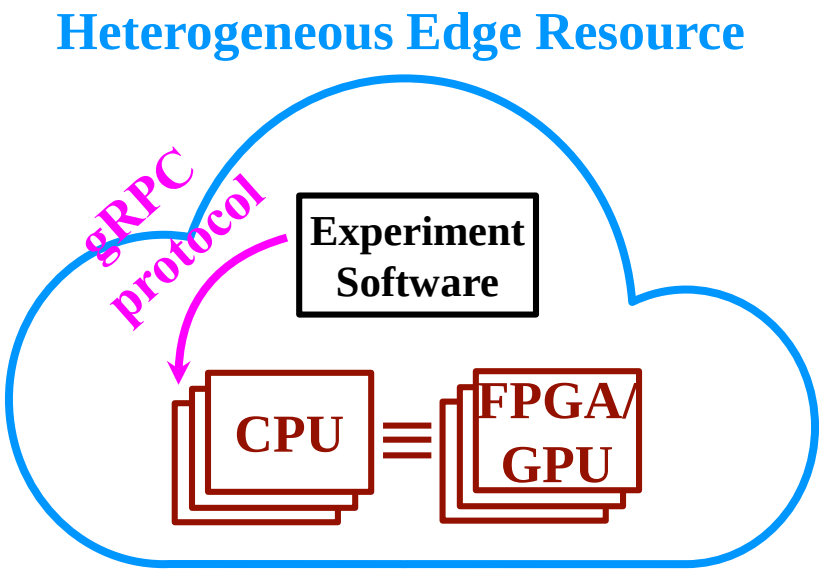
  - gRPC communication
  - TensorFlow or Nvidia Triton inference servers
  - Kubernetes for dynamic scaling of resources
    - Interact with cloud services: Azure, AWS, GCP

✓ Avoid rewriting millions of lines of C++ algorithm code in specialized coprocessor languages

  - User code just converts input and output data into desired formats

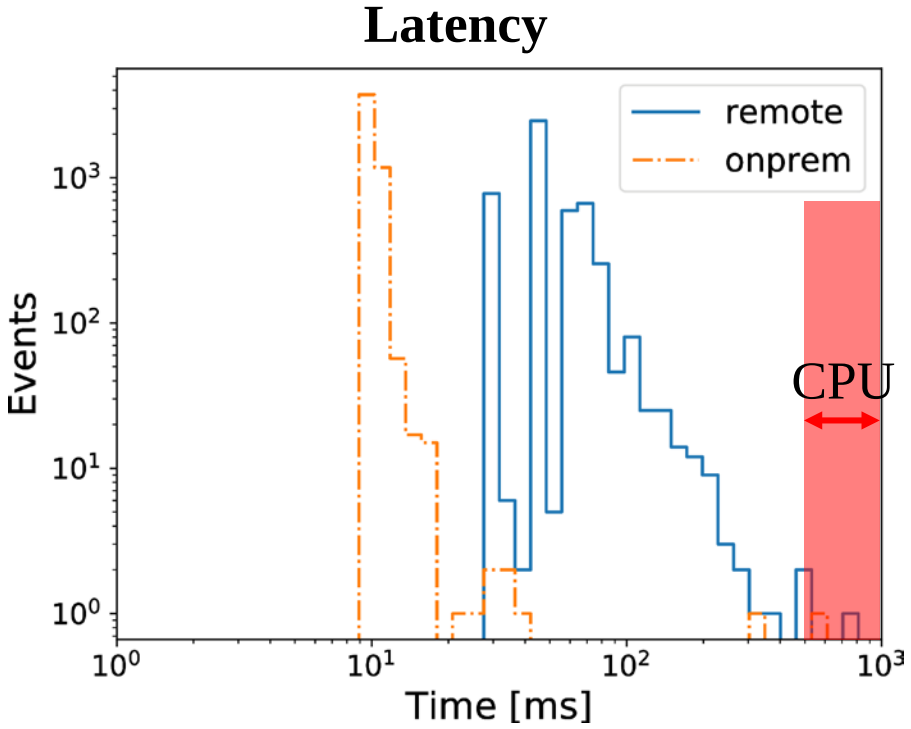
Cloud vs. Edge

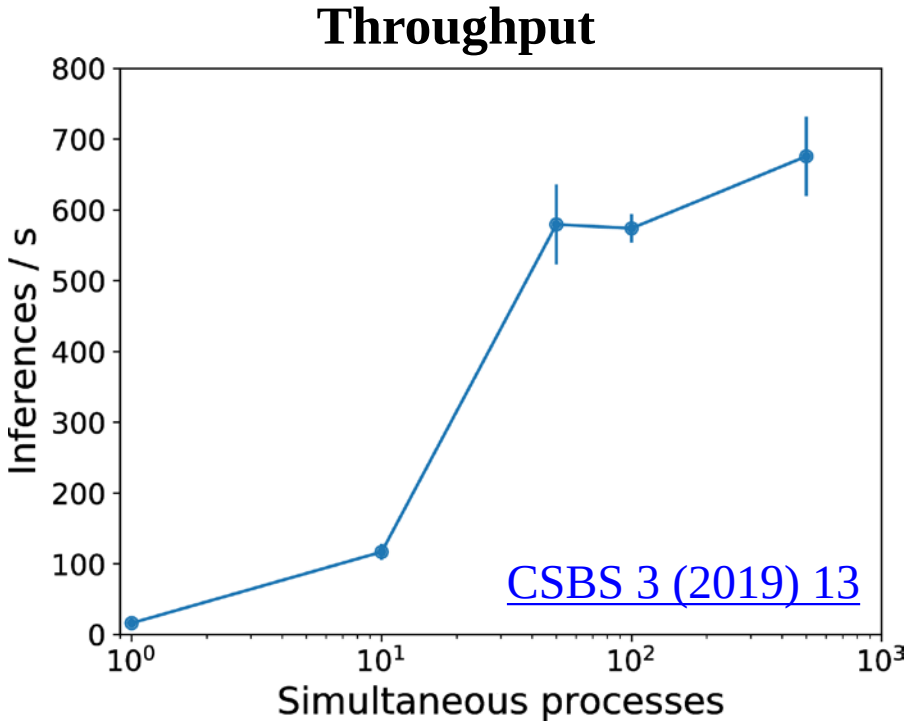




- Cloud service has higher latency
- Local installation of coprocessors: “on-prem” or “edge”
- Provides test of ultimate performance
- Use gRPC protocol either way

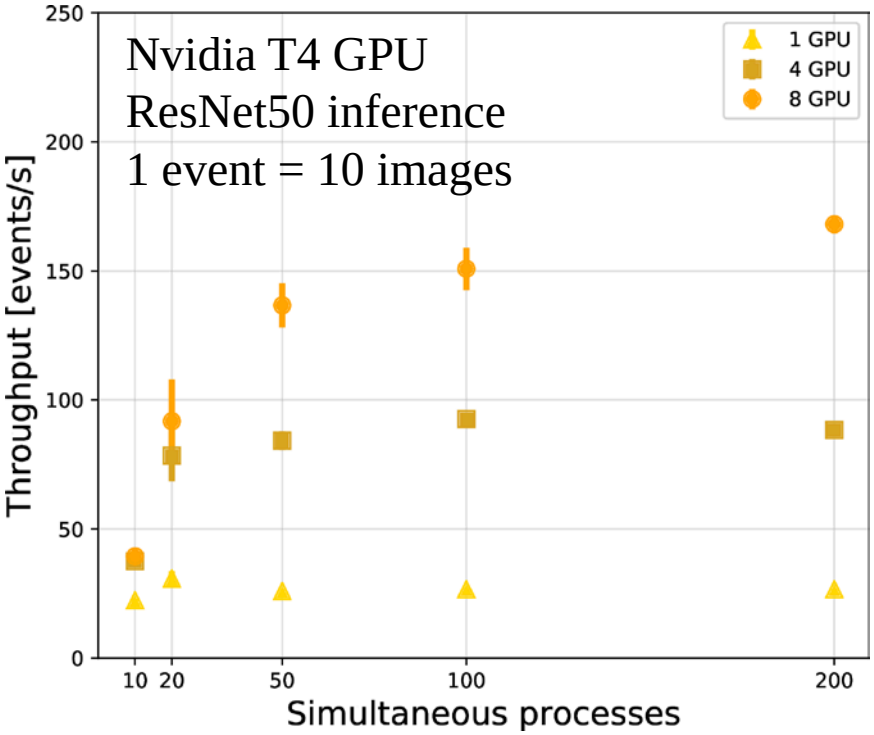
FPGA Results

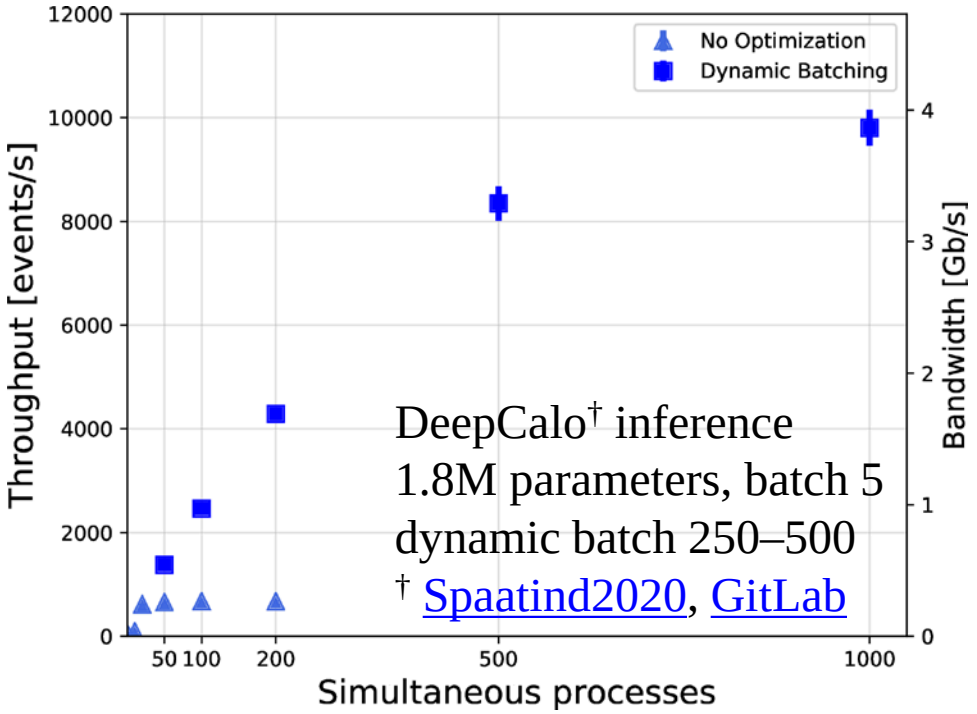




- Microsoft Brainwave FPGA, ResNet50 top tagger inference
- Latency: time for single request to complete
  - ‹CPU› = 500–1000 ms, ‹remote› = 60 ms, ‹on-prem› = 10 ms
- Throughput: requests per second
  - FPGA processes one image at a time, very quickly (1.8 ms)
  - GPU (GTX 1080) needs batch of ~50 images to attain similar throughput

Scaling Up GPUs

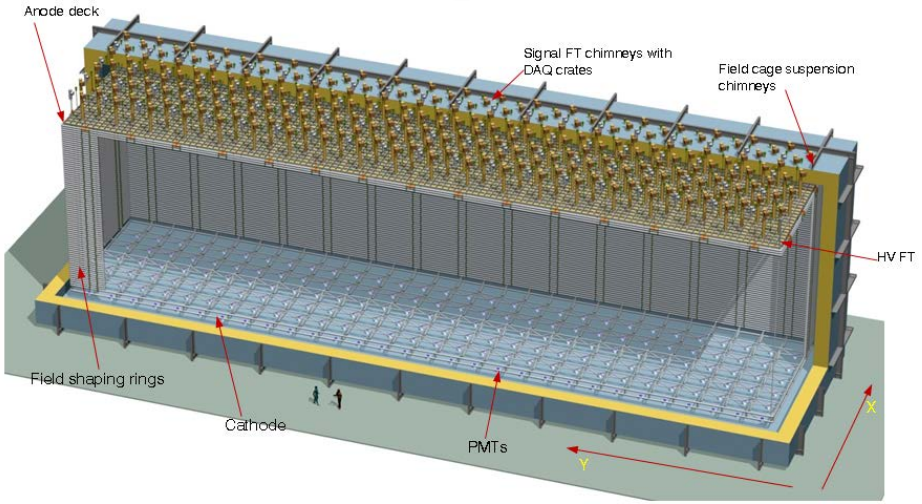




- Use Kubernetes + Triton to deploy multi-GPU server
  - More GPUs support more CPUs, higher throughput
- Triton supports dynamic batching: combine requests from multiple CPUs
  - Huge increase in throughput for large networks with small batch size

Neutrino Challenges

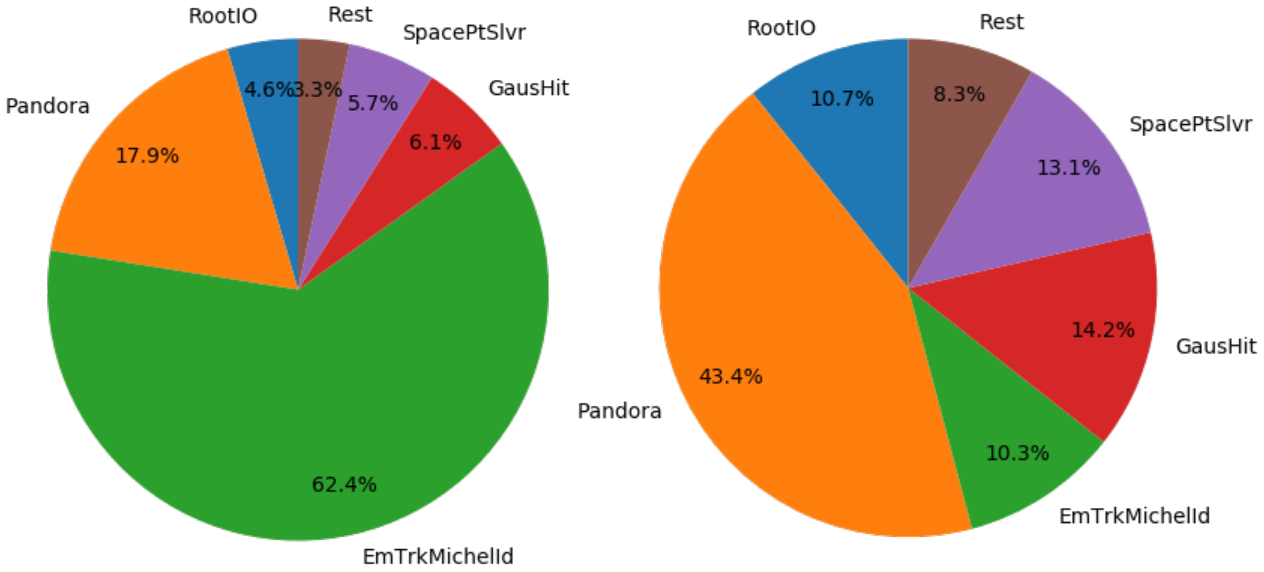
DUNE



DUNE: Deep Underground Neutrino Experiment

- Largest liquid argon detector ever designed
- ~1M channels, 1 ms integration time w/ MHz sampling → 30+ petabytes/year
  - Rate ultimately limited by available computing
- ProtoDUNE operating at CERN (5% size of DUNE)

ν-SONIC



all CPU

w/ GPU

- ProtoDUNE reconstruction dominated by single ML algorithm
- Offload to GPU as a service: >2× overall improvement!
  - Simple implementation w/ blocking, synchronous call
  - Latency preferable to CPU inference

Process	Time [s] (all CPU)	Time [s] (w/ GPU)
Full event	227	99
ML algorithm (EmTrackMichelId)	142	10

Conclusions

Better

- Major strides in deep learning have been incorporated in particle physics
- Significant improvements in top quark tagging (and other tasks)
- AI enables new avenues for discovery, such as boosted $H \rightarrow b\bar{b}$
- Many open questions remain

Smarter

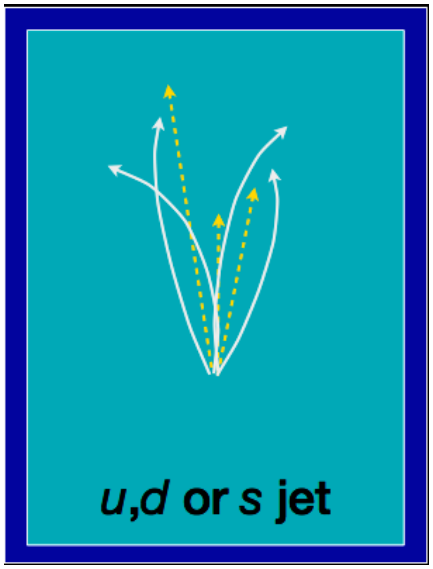
- Moving beyond fully-connected and convolutional neural networks → generalize by embedding data in graphs
- Cutting-edge techniques can handle fundamental tasks: tracking, clustering, calibration

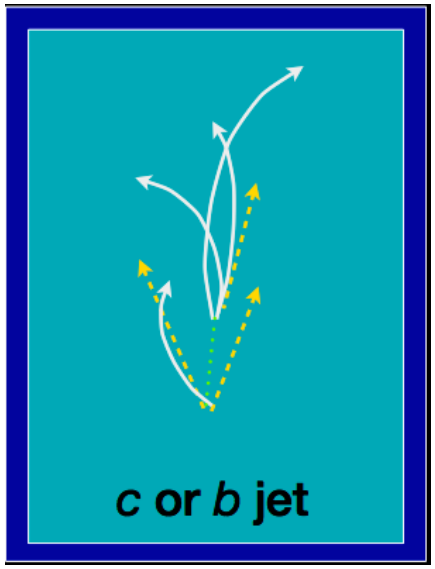
Faster

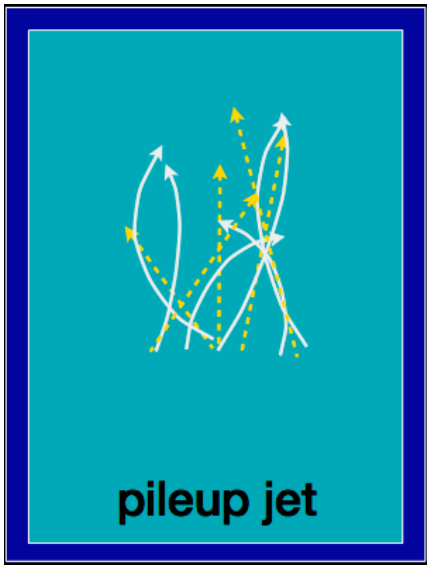
- Need to accomplish fundamentals *and* encourage new capabilities, while coping with unprecedented floods of data
- Solution: accelerate AI inference with coprocessors as a service
- Promising and achievable path for colliders, neutrinos, & beyond!

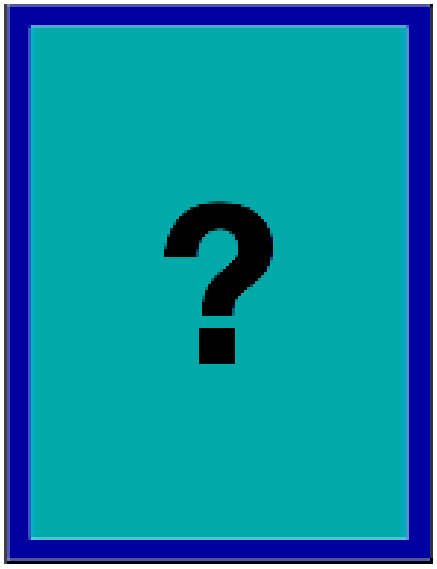
Backup

Jet Substructure



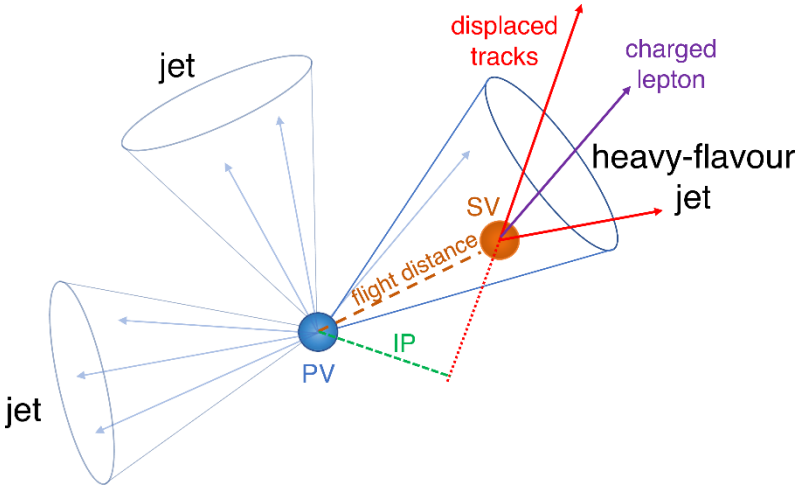




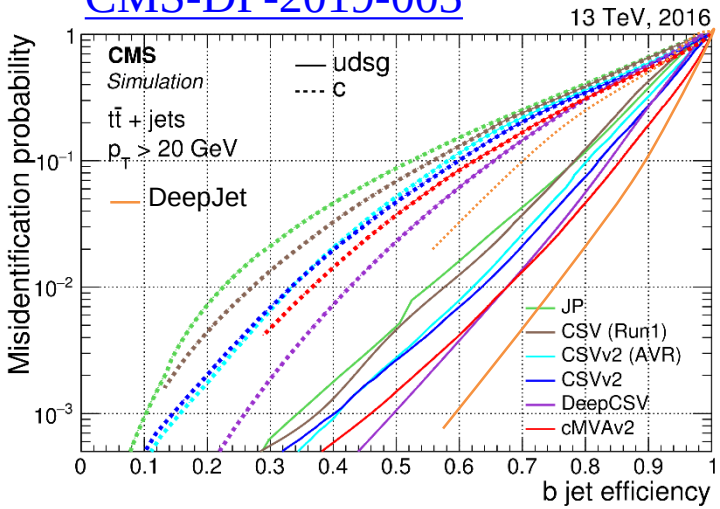


AI for b-tagging

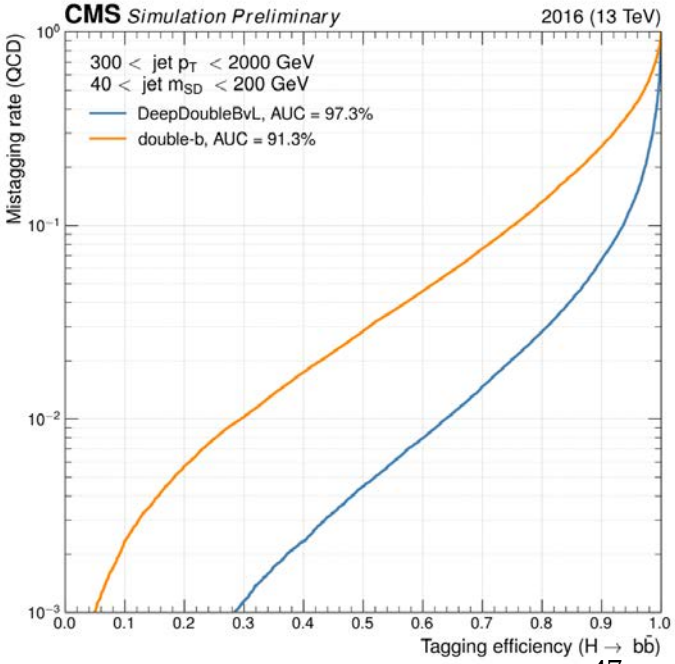
- Similar progression to top tagging:
  - Expert variables
  - Expert variables combined in BDT
  - Expert variables combined in DNN
  - Low-level variables improve DNN
- Double-b-tagging benefits similarly
  - "Expert" corresponds to b-tagging subjets



JINST 13 (2018) P05011
CMS-DP-2019-003



CMS-DP-2018-046



Tagging New Particles

- Many new AI approaches being developed all the time
  - Access to tools, frameworks, computing constantly increasing
- ➢ Can even tag new particles (beyond the SM)
  - e.g. displaced jets from long-lived particles
- Gradient reversal employed to avoid data/simulation discrepancies