Machine Learning in High Energy Physics

Mark Neubauer

University of Illinois at Urbana-Champaign

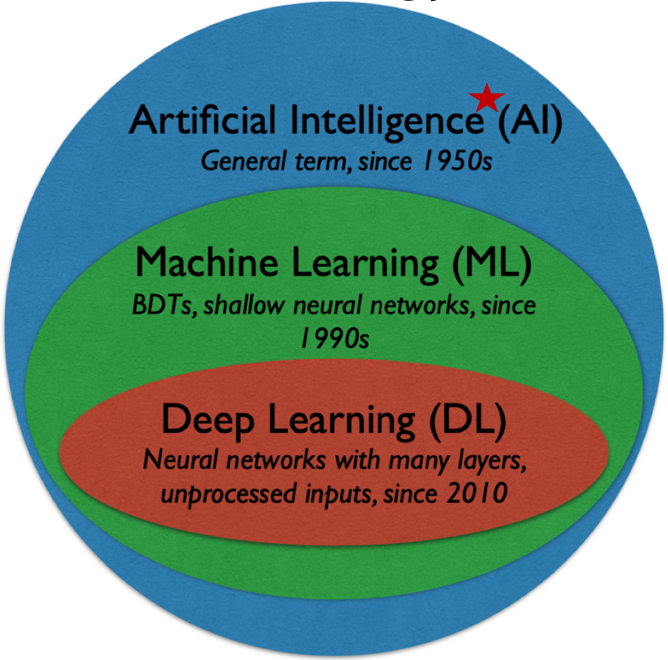
AI for Nuclear Physics Workshop

JLab / March 5, 2020

Preliminaries

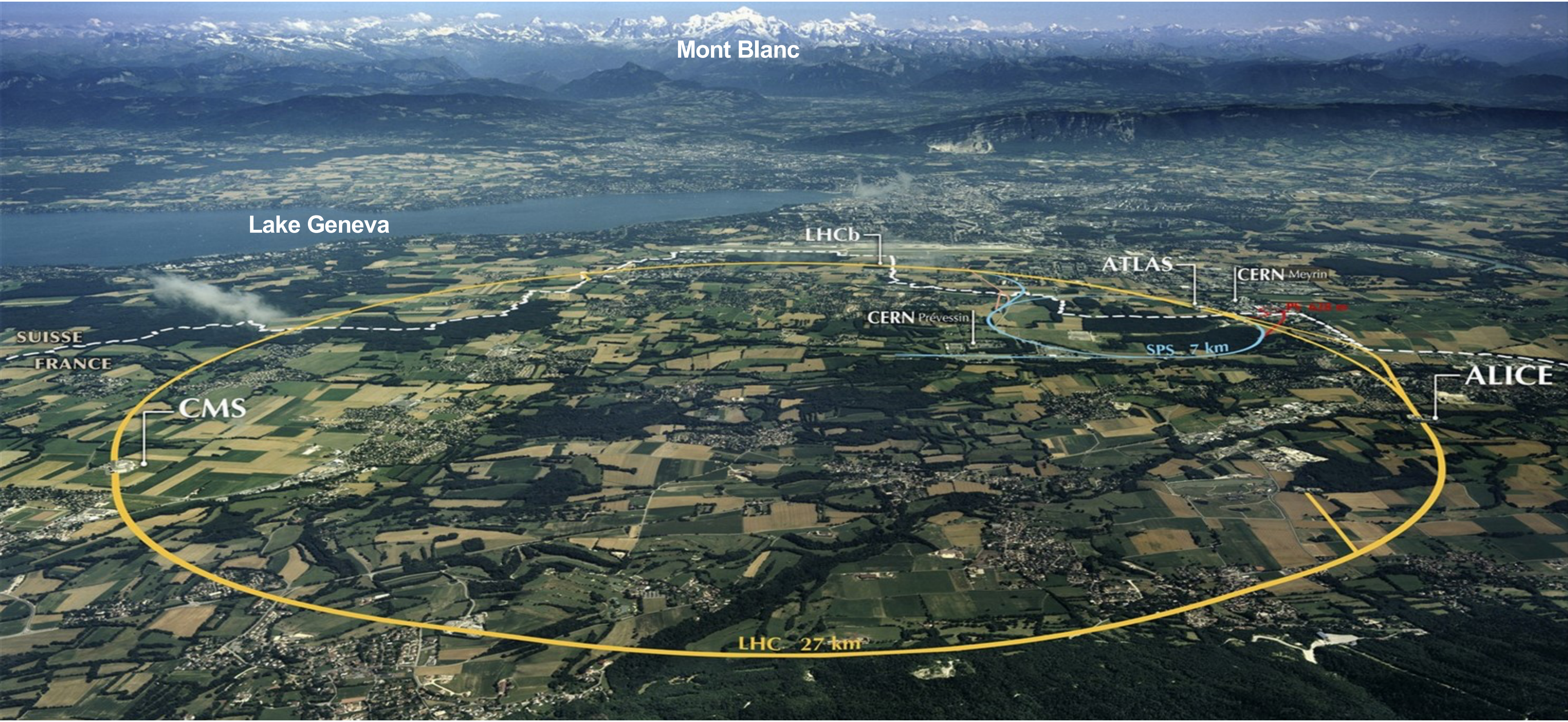
- I will not attempt to summarize all of the past and ongoing work on using ML for High Energy Physics
- I will focus primarily on the experiments at CERN's Large Hadron Collider (LHC)
    - They have big data challenges which ML applications can help to address
- I will try to emphasize ML aspects that I view as most forward-facing toward the next decade

Terminology

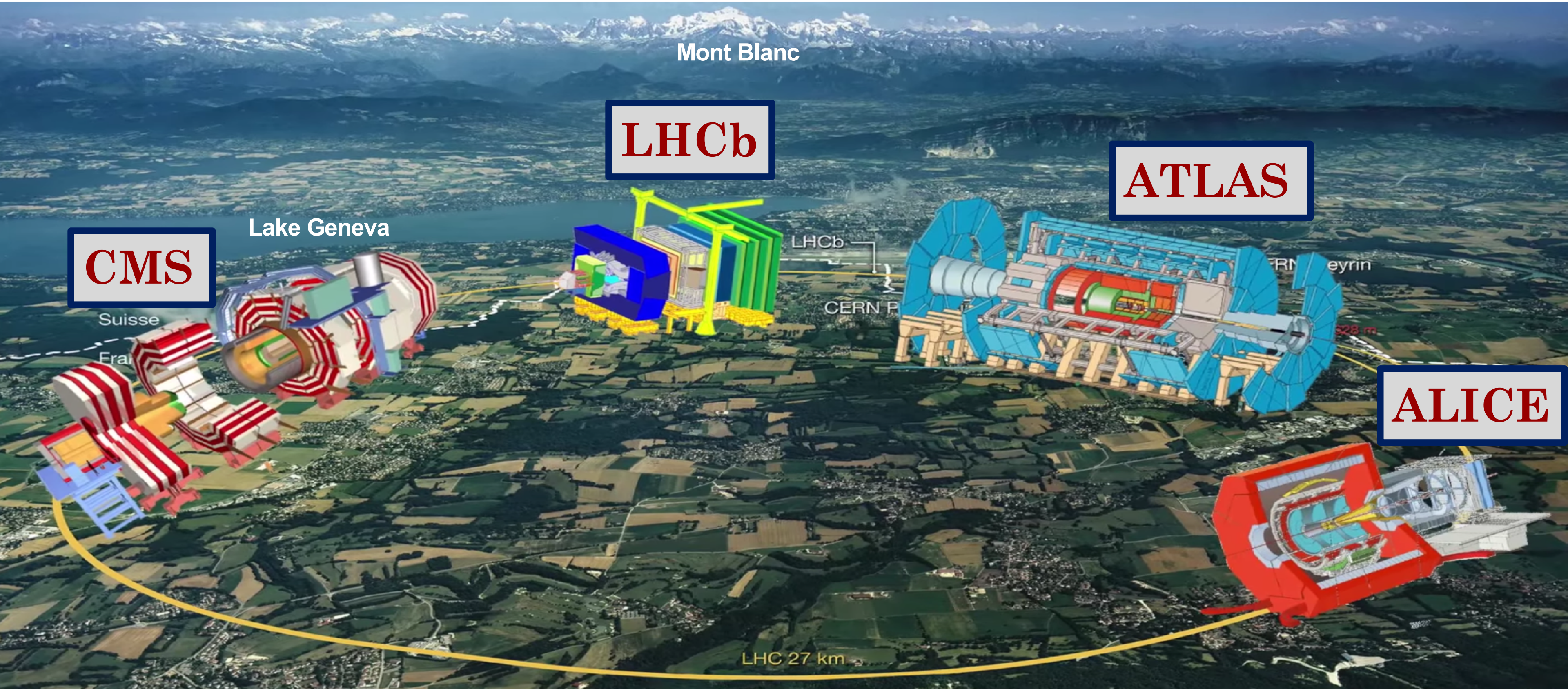


★ *Systems that make decisions usually requiring a human level of expertise, possessing the qualities of intentionality, intelligence and adaptability*

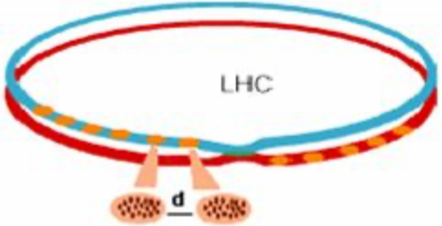
CERN's Large Hadron Collider (LHC)



LHC Experiments



CERN's Large Hadron Collider (LHC)



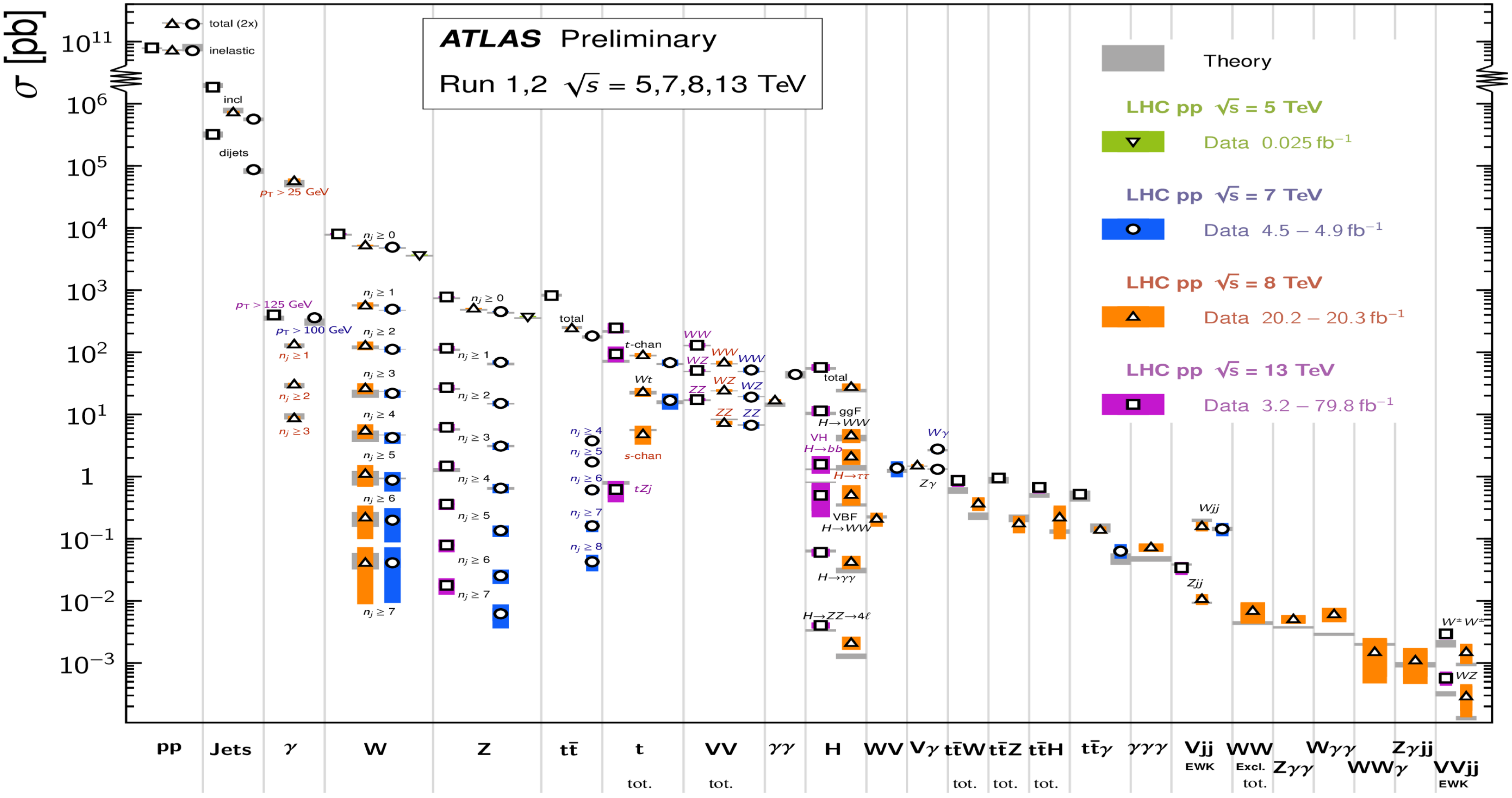
- At the LHC, counter-rotating proton beams cross with a frequency of 40 MHz
- The beams consist of ~2500 bunches of O(100 billion) protons per bunch that are steered into one another

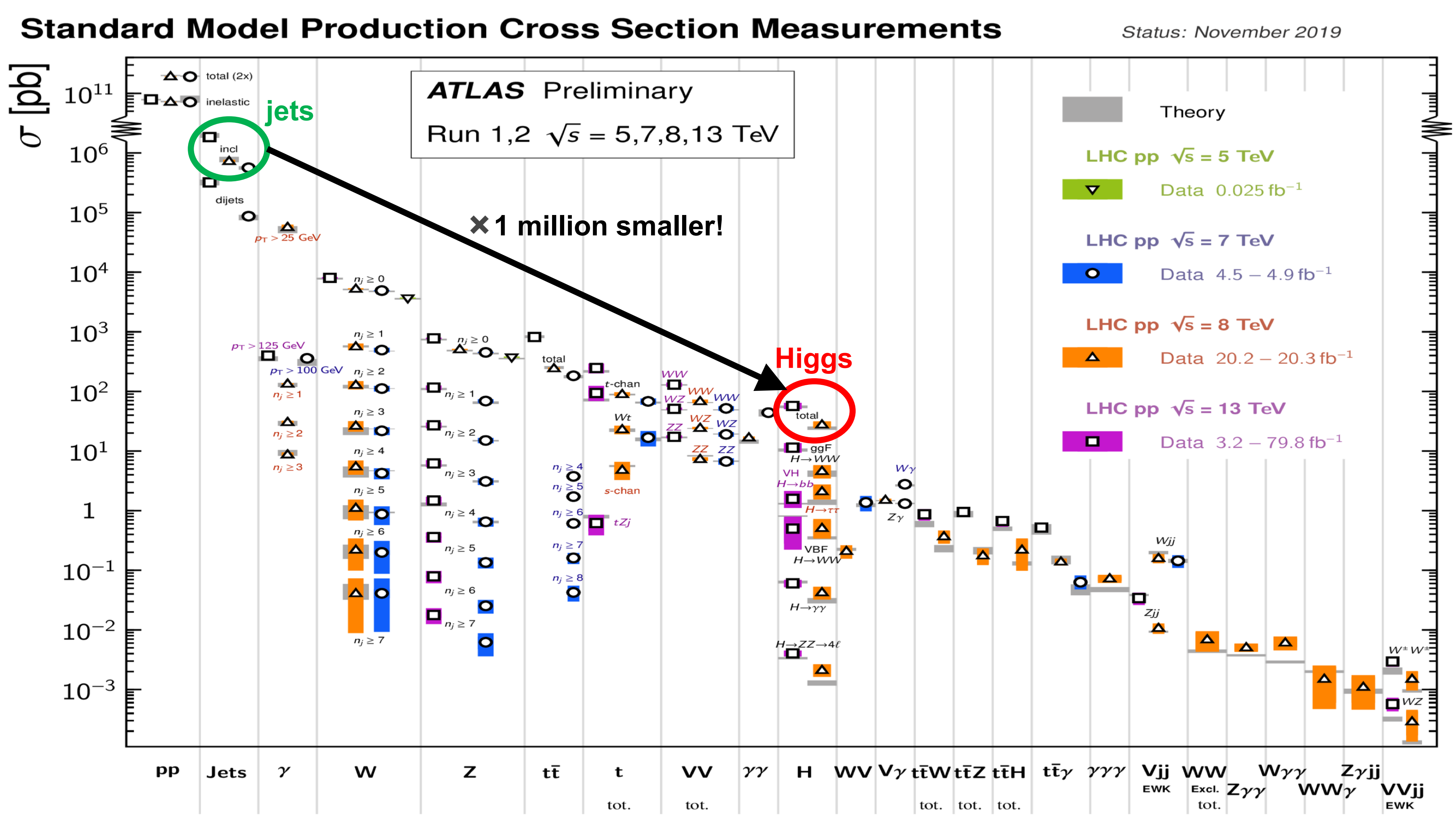
  → Each *pp* collision produces of O(10^3) particles!
- These beams are squeezed to increase the collision rate and therefore increase the chance of producing interesting but rare physics for discovery science

  → comes with a price: multiple *pp* interactions per bunch crossing ("pile-up") that can obscure the most interesting (hard scattering) physics

Standard Model Production Cross Section Measurements

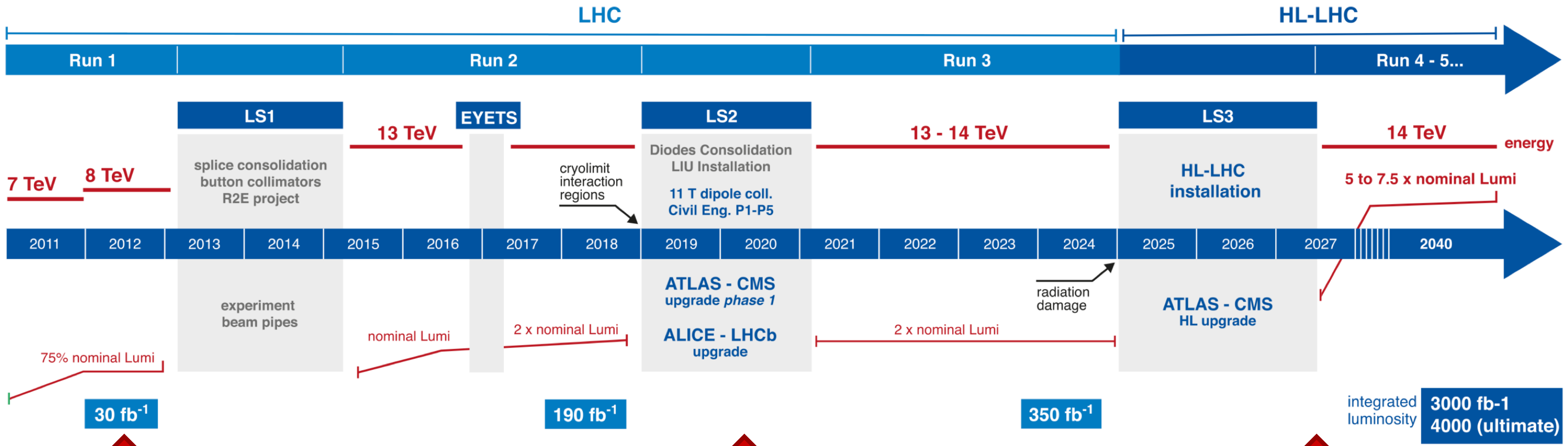
Status: November 2019



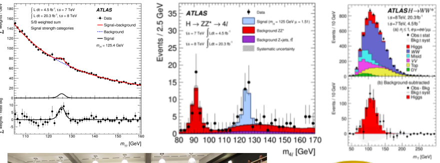


Standard Model Production Cross Section Measurements
Status: November 2019
σ [pb]
ATLAS Preliminary
Run 1,2 √s = 5,7,8,13 TeV
jets
×1 million smaller!
Higgs
Theory
LHC pp √s = 5 TeV
Data 0.025 fb⁻¹
LHC pp √s = 7 TeV
Data 4.5 – 4.9 fb⁻¹
LHC pp √s = 8 TeV
Data 20.2 – 20.3 fb⁻¹
LHC pp √s = 13 TeV
Data 3.2 – 79.8 fb⁻¹
total (2x)
inelastic
incl
dijets
total
pp
Jets
γ
W
Z
tt̄
t
VV
γγ
H
WV
Vγ
tt̄W
tt̄Z
tt̄H
tt̄γ
γγγ
Vjj EWK
WW Excl. tot.
Zγγ
Wγγ
WWγ
Zγjj
VVjj EWK

LHC Schedule

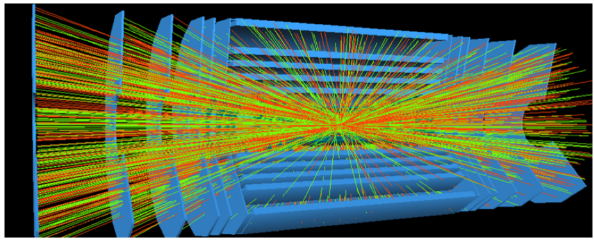


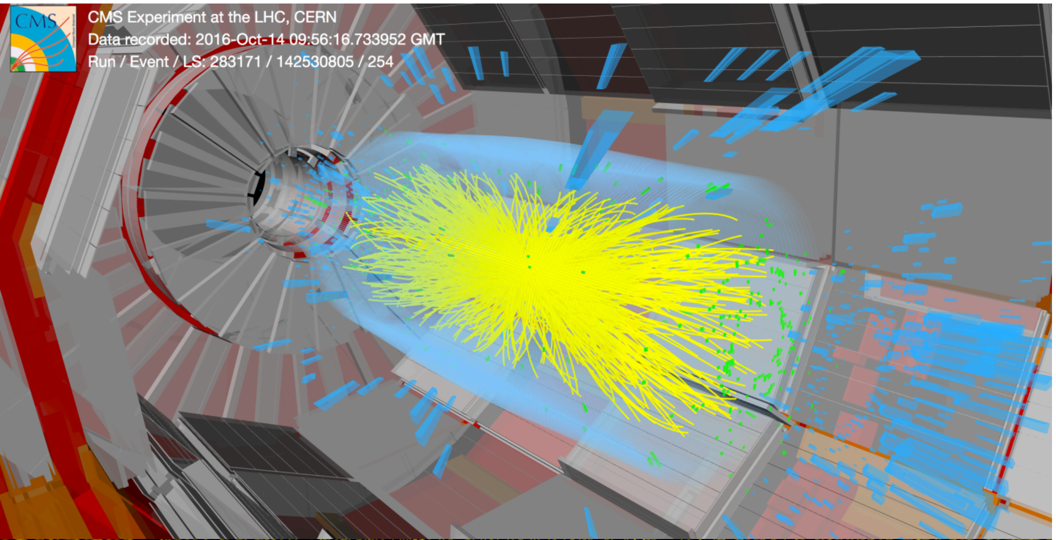
Higgs boson discovery

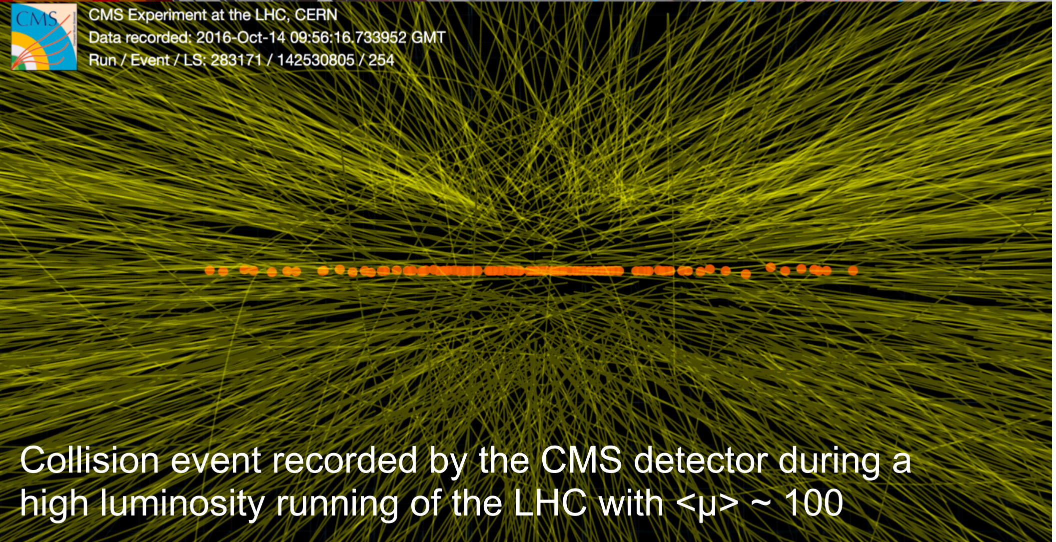


Today

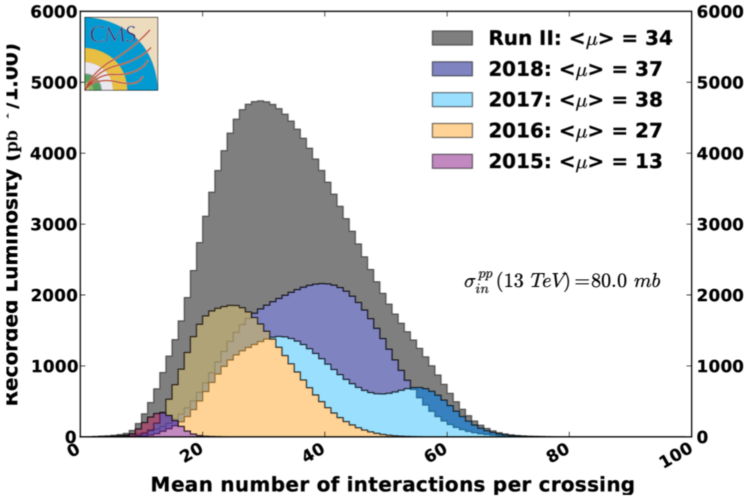
Start of HL-LHC physics







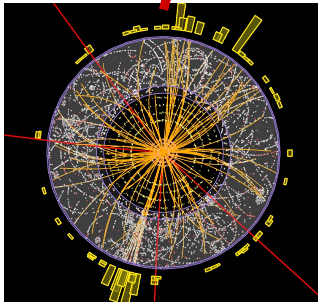
Collision event recorded by the CMS detector during a high luminosity running of the LHC with <μ> ~ 100



- Higher luminosity running leads to increase size and complexity collision data

→ a serious challenge for detector triggering and event reconstruction in the experiments during HL-LHC (<μ>~200) running!

Typical LHC data flow



Collisions

40 MHz
O(100 TB/s)

Level 1 Trigger
99.75% of events rejected!

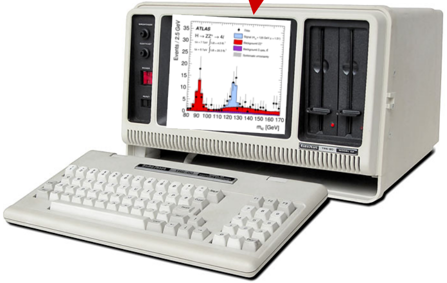
100 kHz

High Level Trigger
99% of events rejected!

1 kHz
1 MB / event

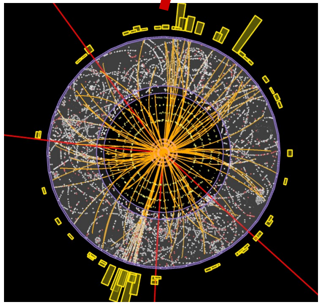
Offline Reconstruction @ Tier-0

Lots of human & artificial intelligence steps here!



Higgs Discovery!

Typical LHC data flow



Collisions

40 MHz
O(100 TB/s)

Level 1 Trigger

99.75% of events rejected!

100 kHz

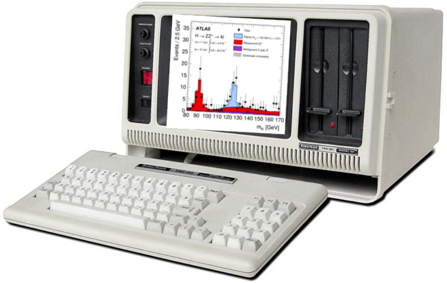
High Level Trigger

99% of events rejected!

1 kHz
1 MB / event

Offline Reconstruction @ Tier-0

Lots of human & artificial intelligence steps here!

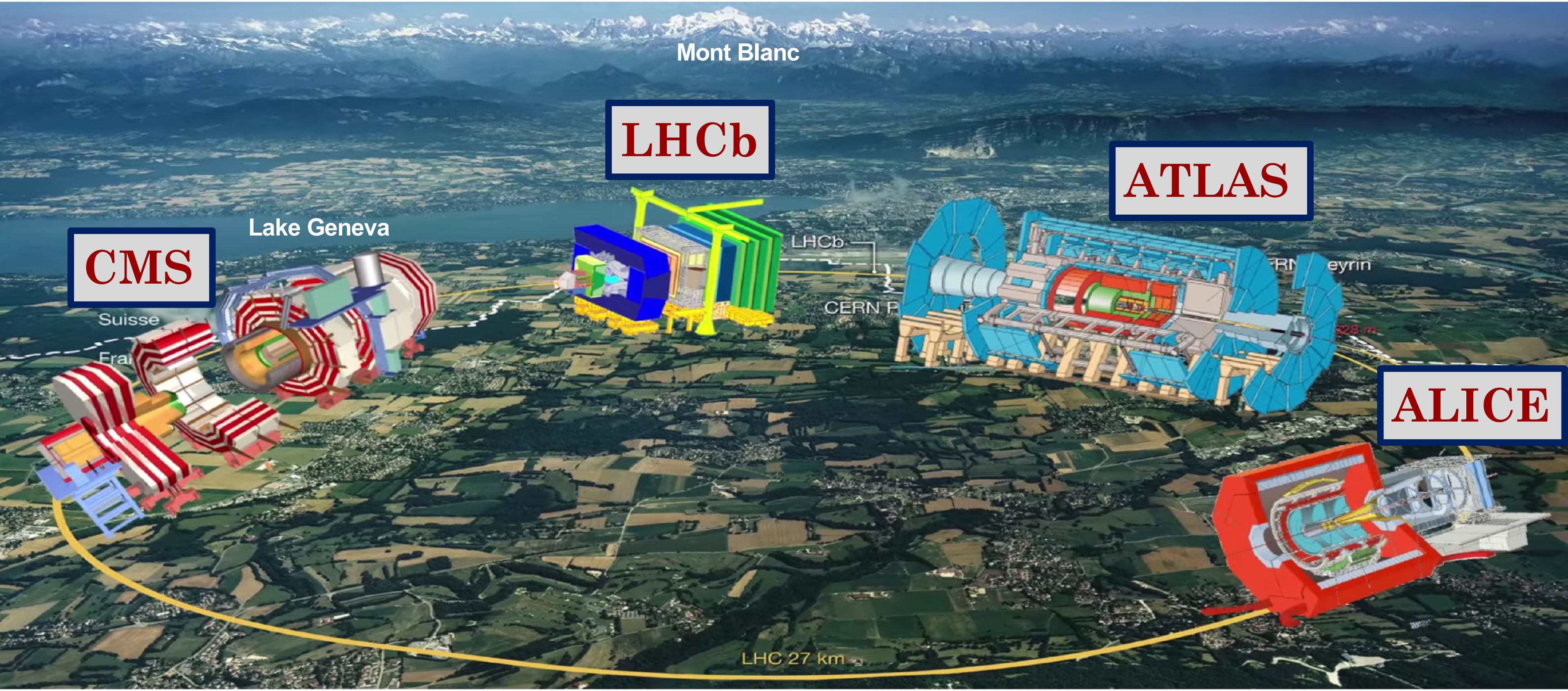


Higgs Discovery!

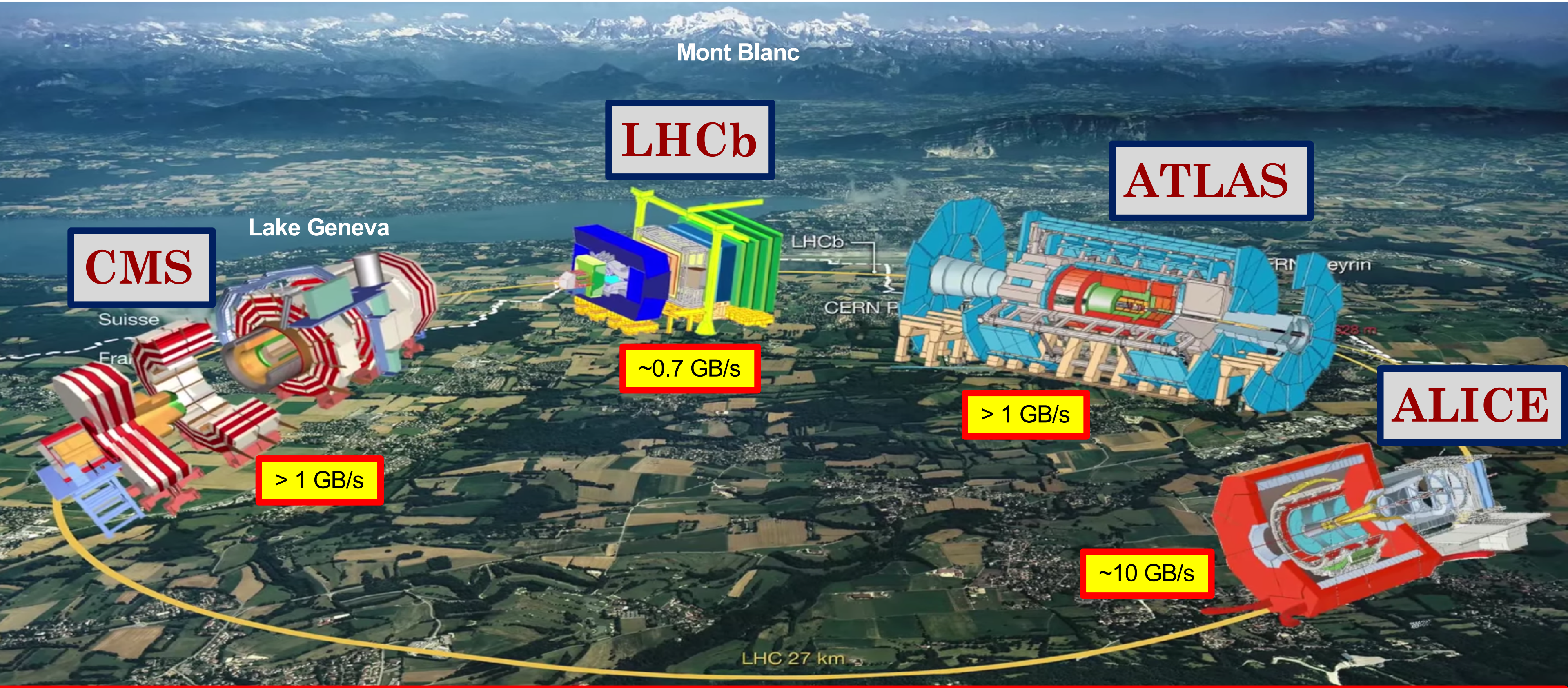
99.9975% of collision events are rejected while retaining those essential for our science!

→ Must continually choose a winner out of 40,000 (avg.) very wisely and must choose it very fast!

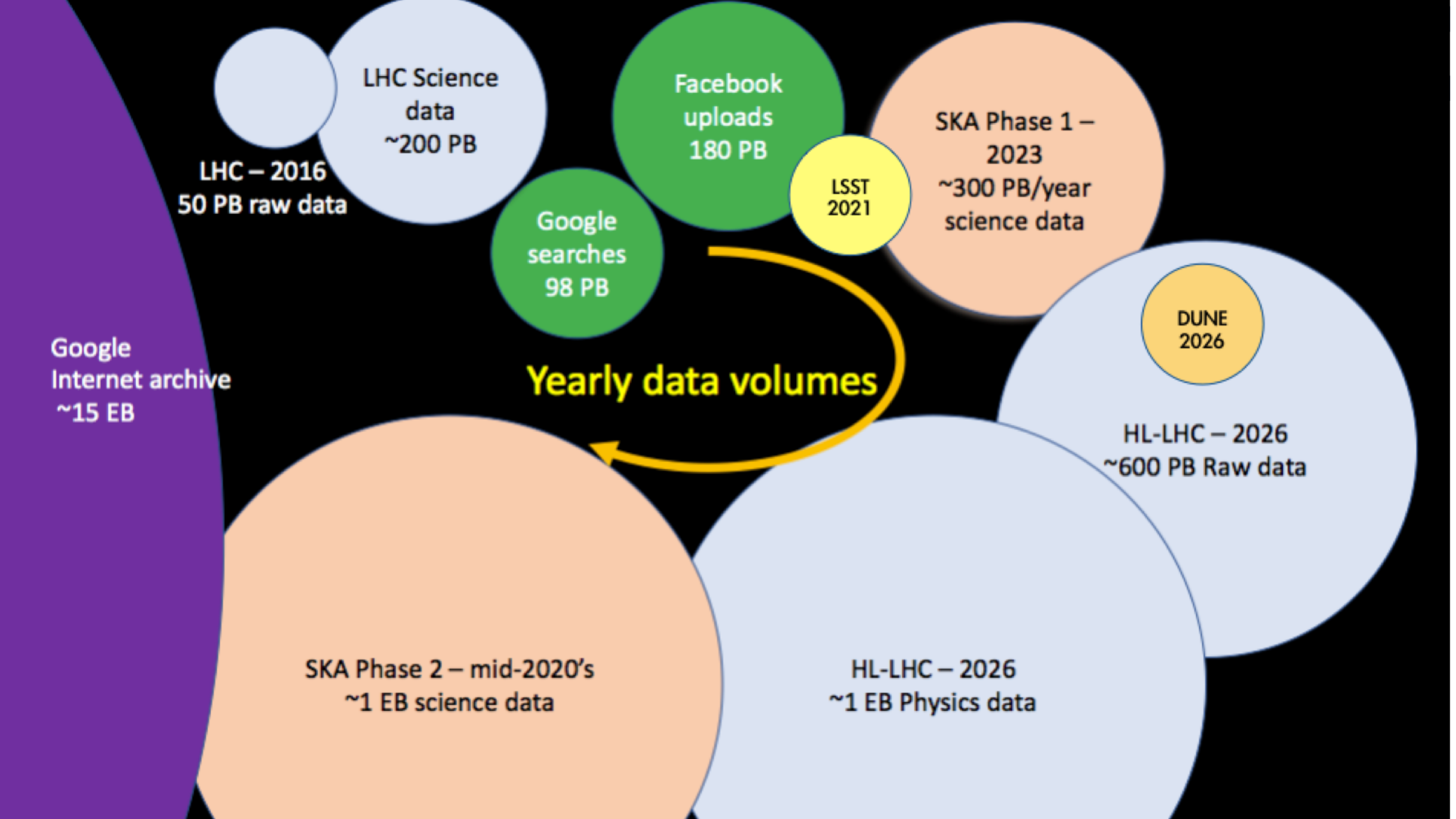
LHC Experiments



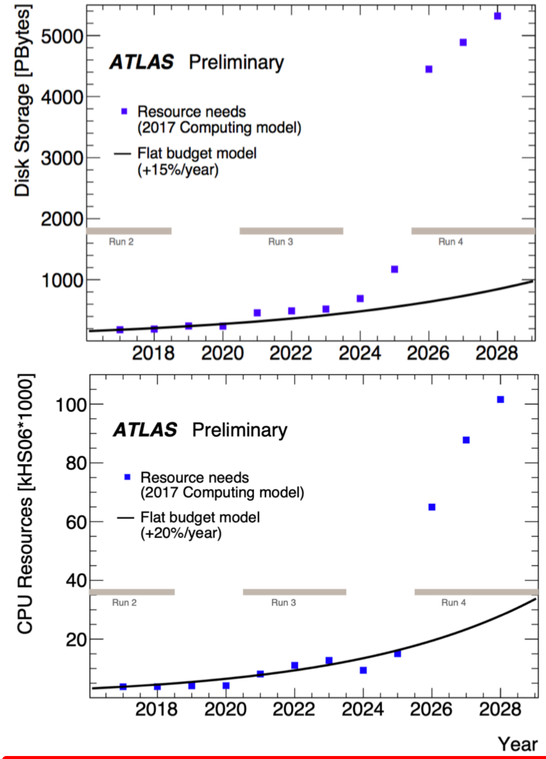
LHC Experiments



LHC Experiments generate ~50 PB/year of science data (during Run 2)


LHC Science data ~200 PB
LHC – 2016 50 PB raw data
Facebook uploads 180 PB
Google searches 98 PB
LSST 2021
SKA Phase 1 – 2023 ~300 PB/year science data
DUNE 2026
HL-LHC – 2026 ~600 PB Raw data
Google Internet archive ~15 EB
Yearly data volumes
SKA Phase 2 – mid-2020's ~1 EB science data
HL-LHC – 2026 ~1 EB Physics data

The HL-LHC Challenge



ATLAS & CMS will record ~10 times as much data from ~100 times as many collisions as were used to discover the Higgs boson (and at twice the energy).

The HL-LHC will produce exabytes of science data per year, with increased complexity: an average of 200 overlapping *pp* collisions per event

Several large US software and computing projects have been initiated to help address these challenges, including:

Institute for Research & Innovation in Software for HEP (IRIS-HEP)

- IRIS-HEP begin in September 2018 and resulted from a 2-year community-wide effort involving 18 workshops and 8 position papers, most notably a HEP Software Foundation Community White Paper and a Strategic Plan.

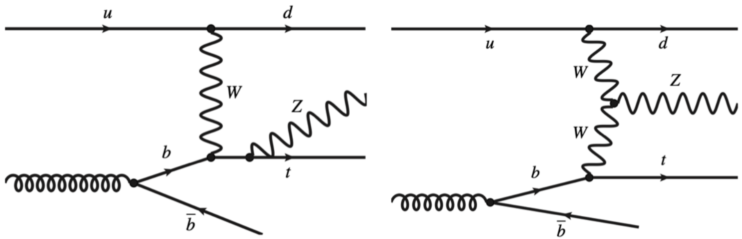
HEP Center for Computational Excellence (HEP-CCE)

- A cross-cutting initiative to promote excellence in high performance computing (HPC) including data-intensive applications, scientific simulations, and data movement and storage.

Many HEP problems can be recast as ML problems → ML is rapidly playing a key role in addressing HL-LHC challenges & enabling new science capabilities!

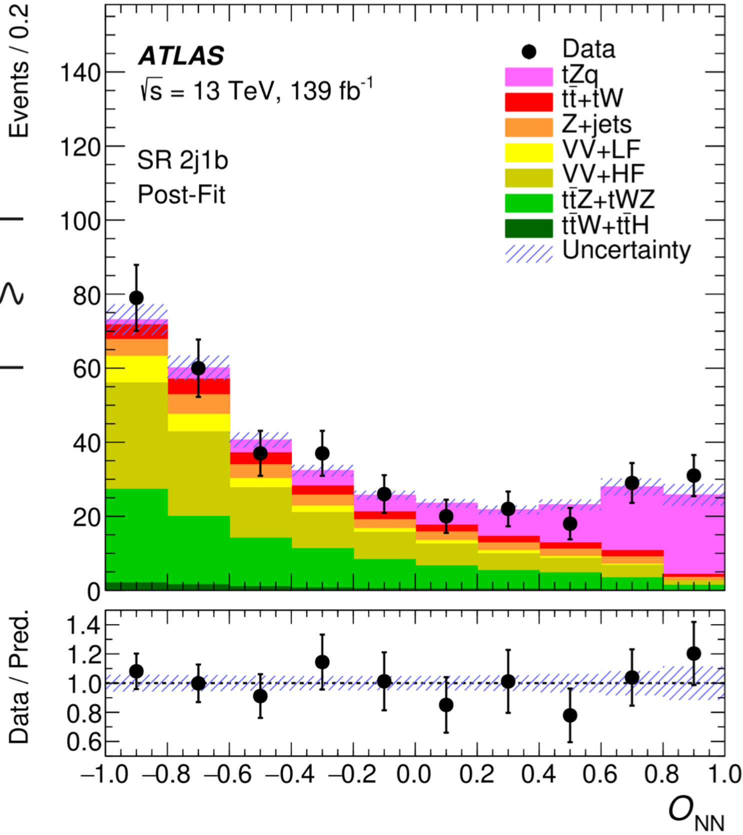
E.g. Event Classification

- Observation of electroweak single top quark production in association with a Z boson and a quark (*pp*→*tZq*) by the ATLAS Collaboration
- Sensitive probe of new physics (e.g. FCNCs)



- Trained an artificial neural network (ANN) to classify tZq events using simulated data for improved signal and background separation
  - ANN performance validated by studying events with similar signatures (control regions) that are dominated by background processes

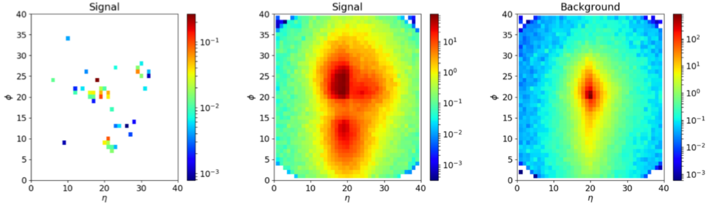
arXiv:2002.07546

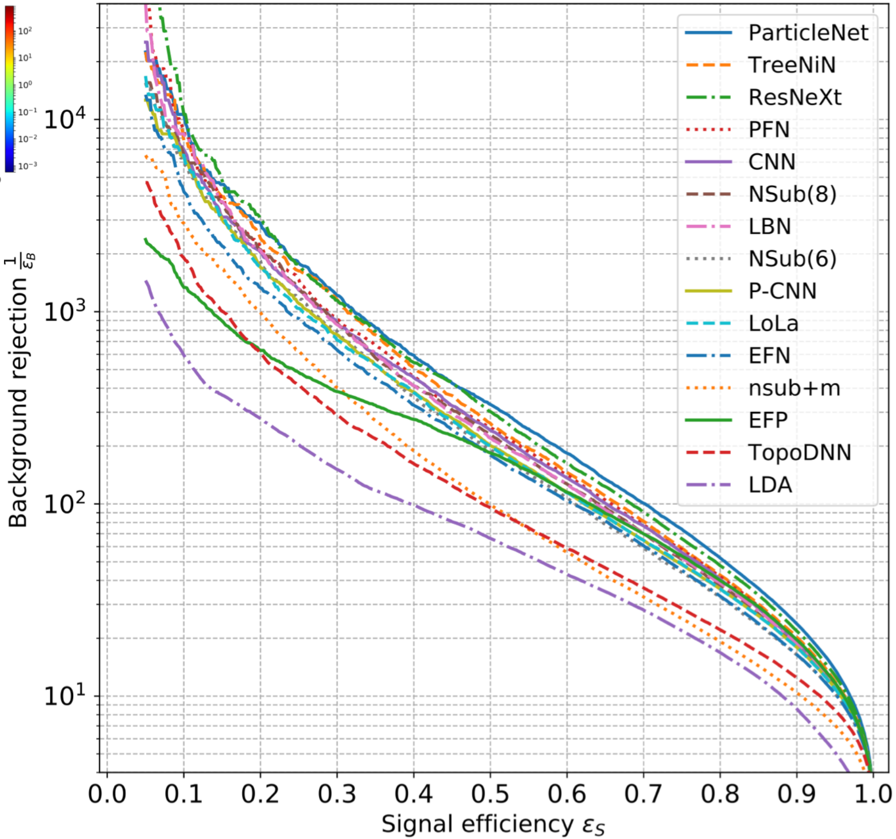


E.g. "Object" (t-quark) Classification

Comparison of many "state-of-the-art" ML-based top-quark taggers   arXiv:1902.09914

- **Image-based**
  - CNN
  - ResNeXt
- **4-vector-based**
  - TopoDNN
  - Multi-Body N-Subjettiness (Nsub)
  - TreeNiN
  - Particle-level CNN (P-CNN)
  - ParticleNet
- **Theory-inspired taggers**
  - Lorentz-boost network (LBN)
  - Lorentz-Layer (LoLa)
  - Latent Dirichlet Allocation (LDA)
  - Energy Flow Polynomials (EFP)
  - Energy Flow Network (EFN)
  - Particle Flow Network (PFN)





E.g. ML-based Fast Shower Simulation

ATLAS: Simulation is the largest use of distributed computing resources and ~80% of that is calorimeter simulation

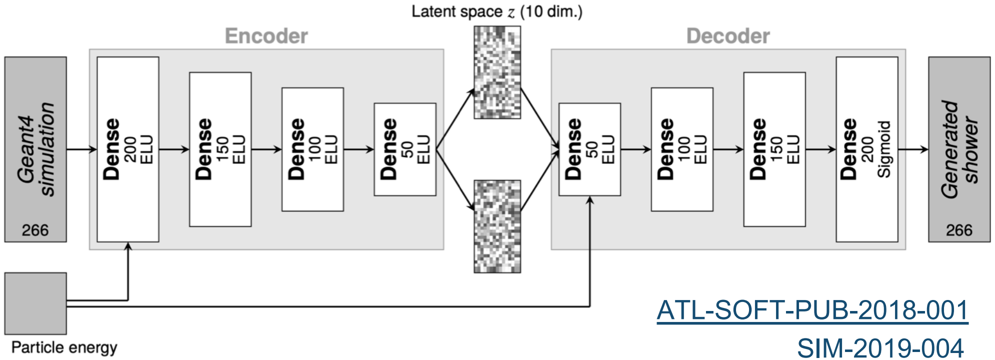
Two of the most promising ML-based approaches studied thus far:

Variational Autoencoder (VAE)

- Encode representation of Geant4 showers into latent space
- Use decoder to generate new showers

Generative Adversarial Networks (GAN)

- Train a generator for new showers
- Critic: Difference between generated shower and Geant 4
- Second Critic for total shower energy

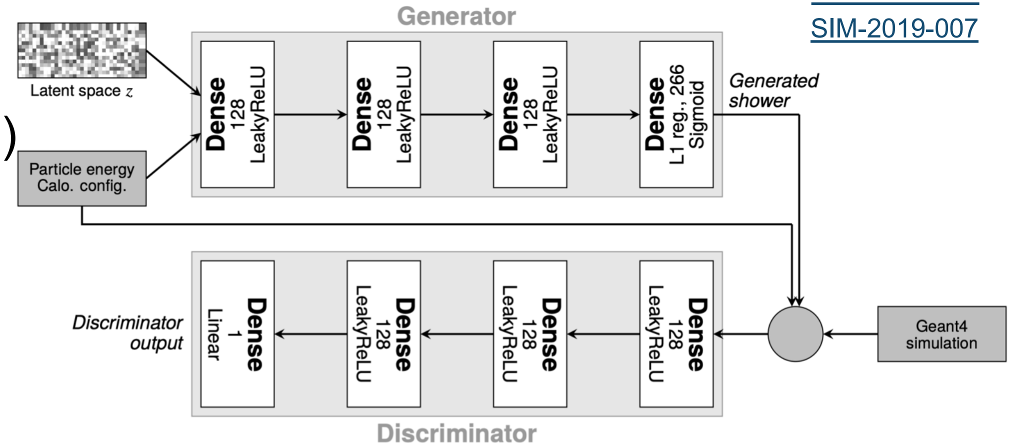


ATL-SOFT-PUB-2018-001

SIM-2019-004

SIM-2019-006

SIM-2019-007



E.g. ML-based Fast Shower Simulation

ATLAS: Simulation is the largest use of distributed computing resources and ~80% of that is calorimeter simulation

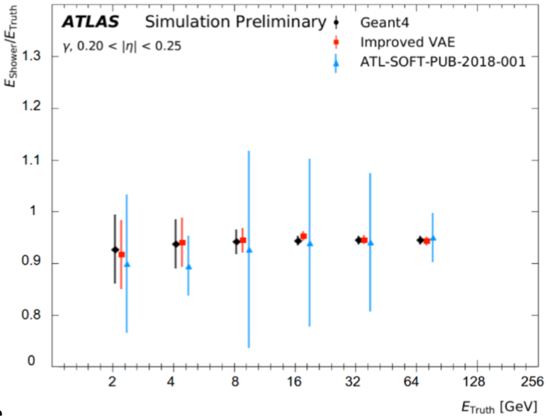
Two of the most promising ML-based approaches studied thus far:

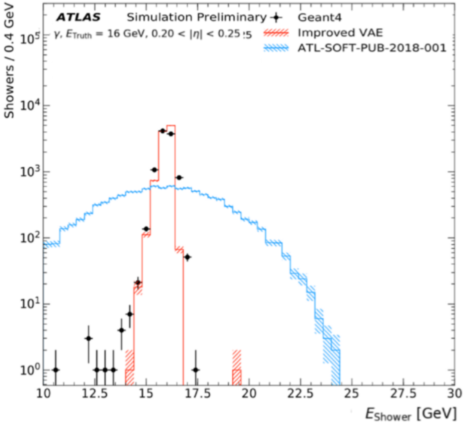
Variational Autoencoder (VAE)

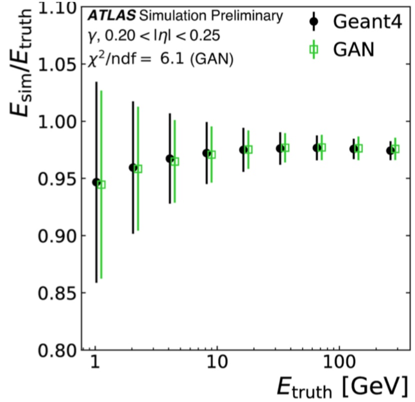
- Encode representation of Geant4 showers into latent space
- Use decoder to generate new showers

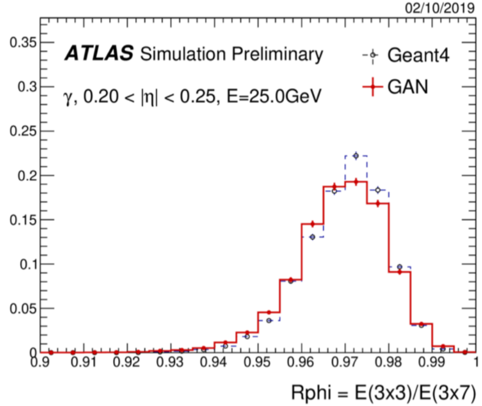
Generative Adversarial Networks (GAN)

- Train a generator for new showers
- Critic: Difference between generated shower and Geant 4
- Second Critic for total shower energy





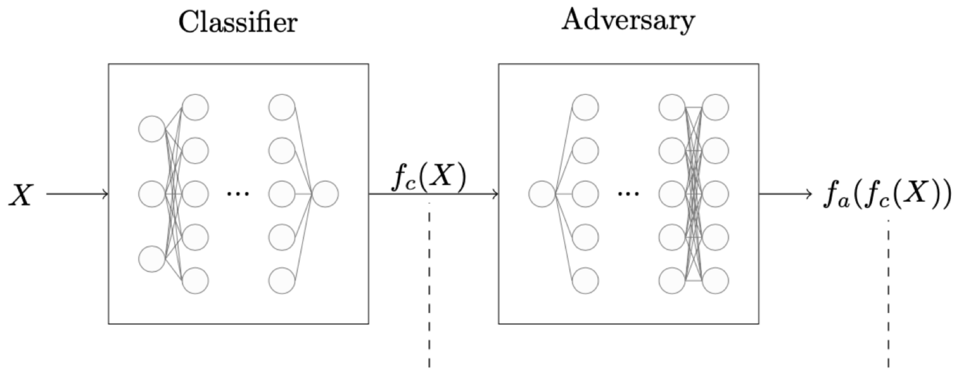




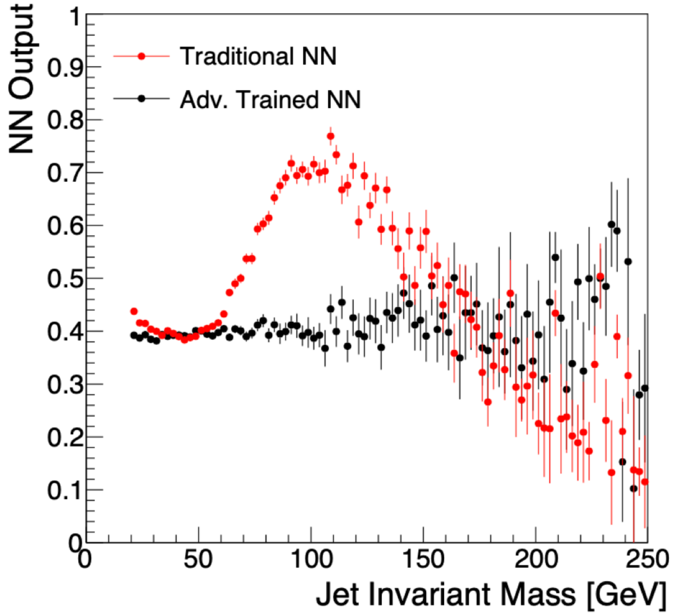
E.g. NN Stability / Decorrelation

- Goal: Stabilize discriminator against systematic uncertainty or other effect
- Adversarial training (e.g. decorrelate jet mass)

S Bollweg, BSc Thesis
1703.03507 1802.03325



$$L_{\text{tagger}} = L_{\text{classification}} - \lambda L_{\text{adversary}}$$

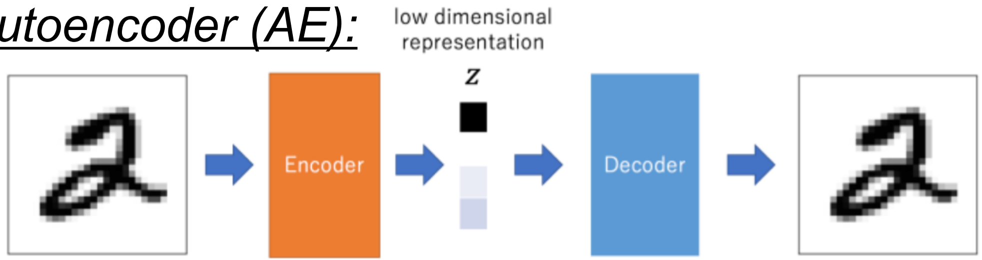


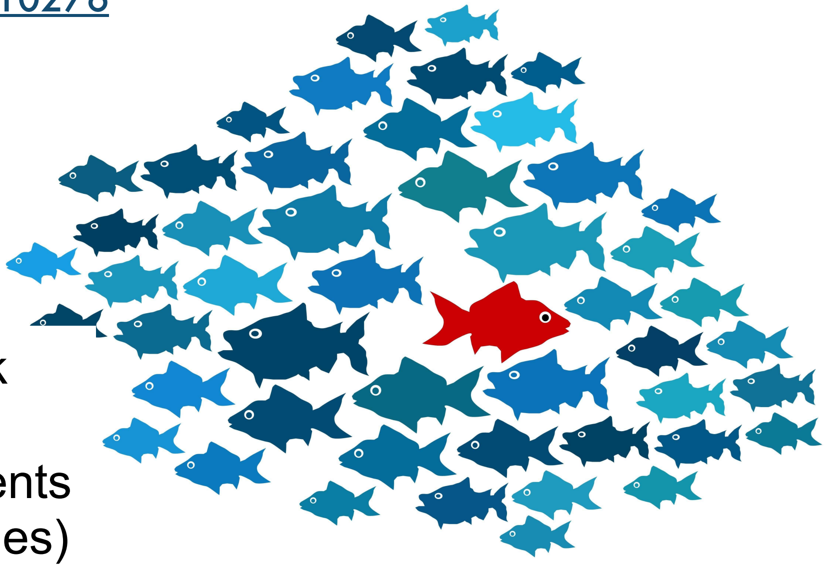
E.g. Triggering with Autoencoders

- Without algorithmic or other (e.g. hardware) improvements, trigger requirements become more restrictive at the HL-LHC to fit into computing constraints → decreasing sensitivity to some beyond-the-SM (BSM) signatures
- Unsupervised learning can be used to train the trigger to identify BSM physics as anomalies in the data stream

arXiv:1811.10276

Autoencoder (AE):

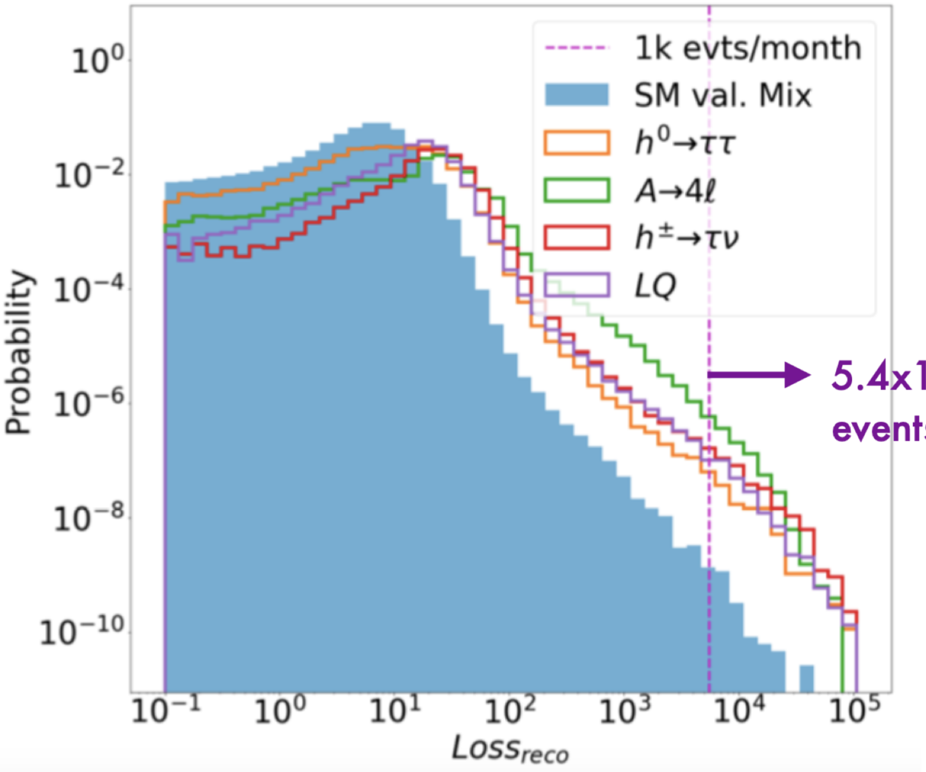




Since the compression capability of an AE network does not generalize well to other data, we can use the loss (encoder-decoder distance) to identify events not representative of the training data (i.e. anomalies)

E.g. Triggering with Autoencoders

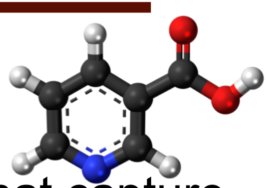
- SM cocktail dataset as collected with isolated single lepton trigger
- 21 input quantities: lepton momentum, isolation, charge, number of jets, missing transverse energy, etc... → very generic and intended to be BSM signal agnostic
- Triggered events could be written to a special "anomaly" data stream for additional analysis



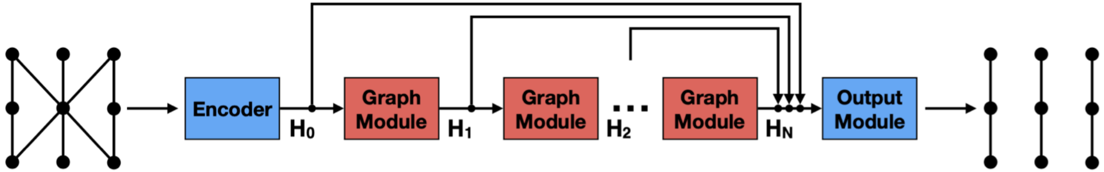
5.4x10^{-6} of the SM events are retained

arxiv.1811.10276

E.g. Tracking w/ Graph Neural Networks



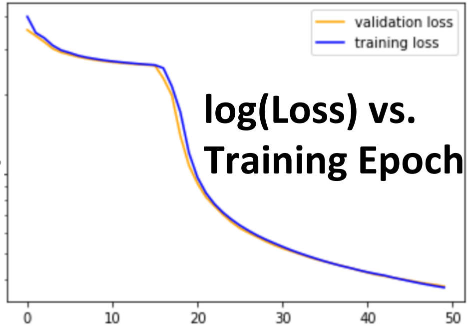
- Data relationships in many real-world applications can be naturally represented by graphs
- Graph Neural Networks (GNNs) are deep learning based methods that capture dependencies on graphs via message passing between the nodes of graphs
- GNNs are well-suited to pattern recognition—a key element of reconstructing charged particles in tracking detectors
- Work on GNN approaches to tracking and calorimetry initiated by HEP.TrkX and is being driven by the EXA.TrkX group (see their 2019 NeurIPS paper). Several members of IRIS-HEP are collaborating on this effort
- Detector measurements are represented as graph nodes which are associated with one another by learned graph edges that represent the particle tracks

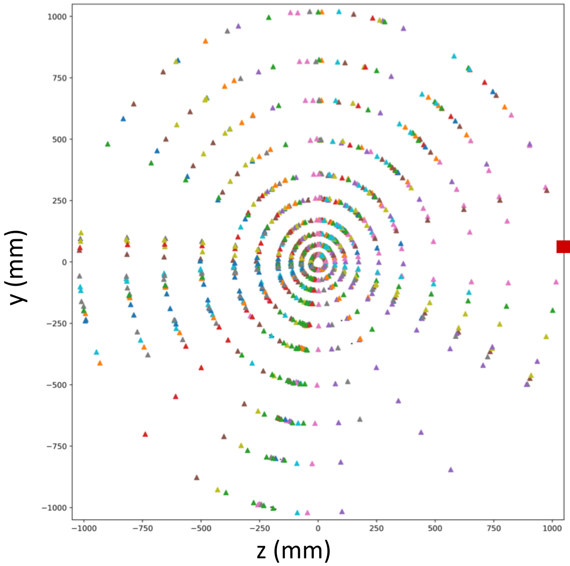


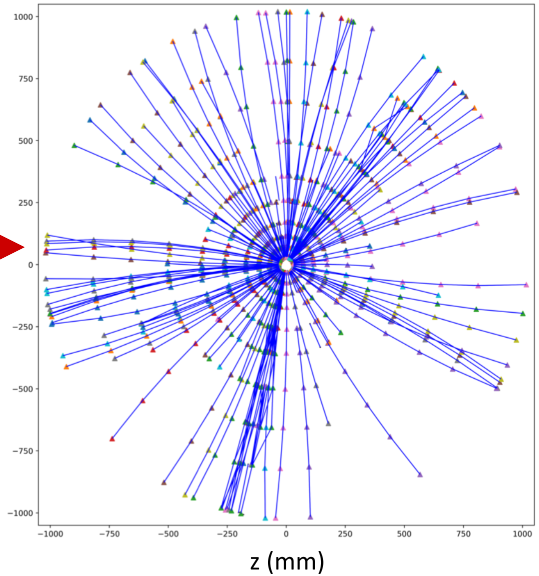
E.g. Tracking w/ Graph Neural Networks

- We use the TrackML Challenge Data for training & evaluation
- Preprocessing for GNN models use HEP.TrkX libraries with a truth particle p_T> 2 GeV cut

GeForce RTX™ 2080Ti Turbo 11G



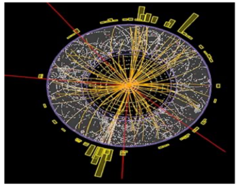




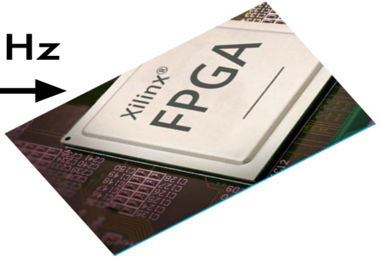
- GNN-based inference can be implemented on FPGAs to accelerate computationally expensive parts of the event reconstruction such as calorimetry and tracking in the ATLAS or CMS High-Level (software) trigger

E.g. Real-time ML-based Inference

1 ns 1 μs 100 ms 1 s



40 MHz



Level 1 trigger

100 KHz

High-Level Trigger

1 KHz

1 MB/event

Offline reconstruction @ Tier 0

Absorbs 100s TB/s

Trigger decision to be made in O(μs)

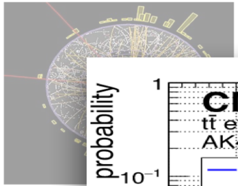
Latencies require all-FPGA design

Computing farm for detailed analysis of the full event

Latency O(100 ms)

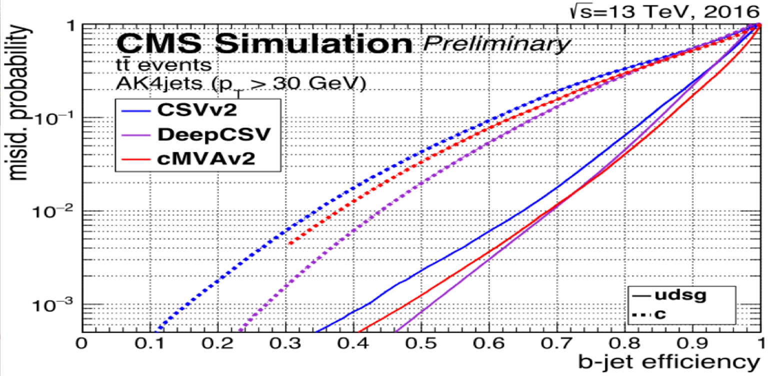
E.g. Real-time ML-based Inference

1 ns — 1 μs — 100 ms — 1 s



40 MHz

100 KHz



A

Tr

Latencies require all-FPGA design

Deep neural network
based on high-level features for b-quark jets identification (offline & HLT)

1 KHz

1 MB/event

Offline reconstruction @ Tier 0

High-Level Trigger

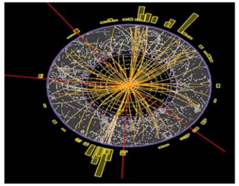
Computing farm for detailed analysis of the full event

Latency O(100 ms)

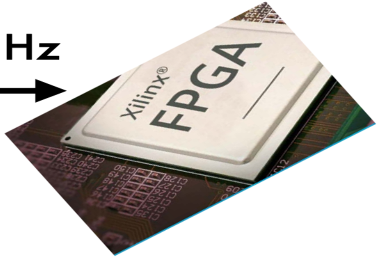
We are already applying Deep Learning here!

E.g. Real-time ML-based Inference

1 ns — 1 μs — 100 ms — 1 s



40 MHz



Level 1 trigger

100 KHz

High-Level Trigger

1 KHz

1 MB/event

Offline reconstruction @ Tier 0

Absorbs 100s TB/s

Trigger decision to be made in O(μs)

Latencies require all-FPGA design

Computing farm for detailed analysis of the full event

Latency O(100 ms)

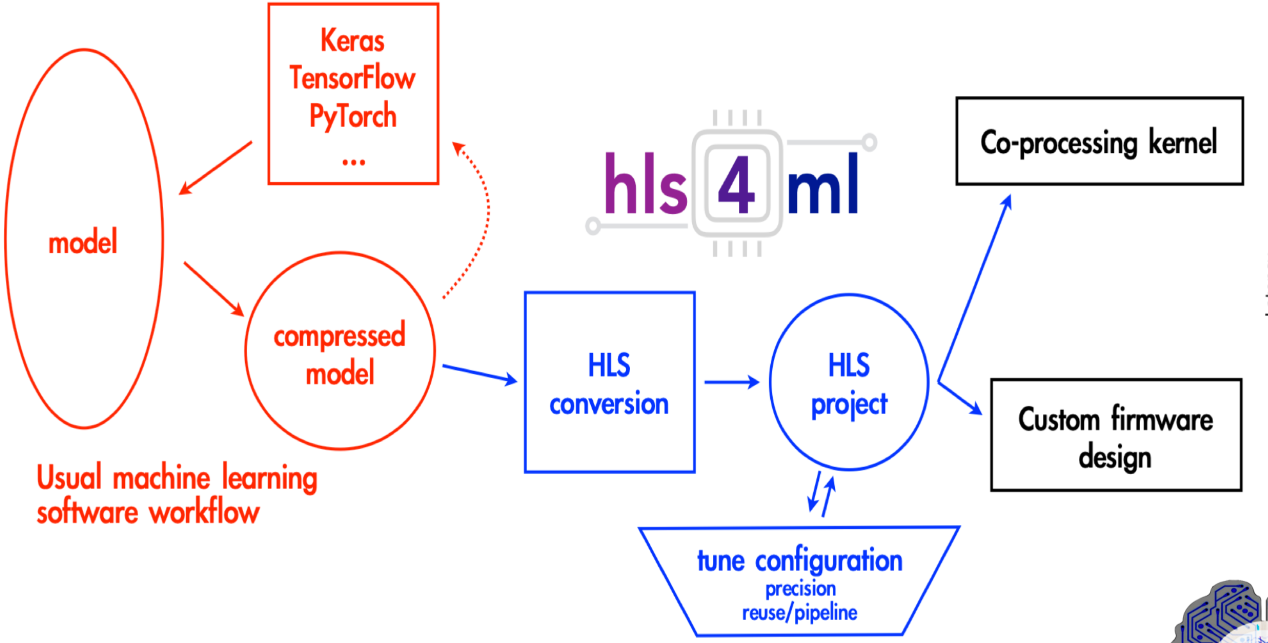
Can we do real-time AI in O(μs) on one FPGA?

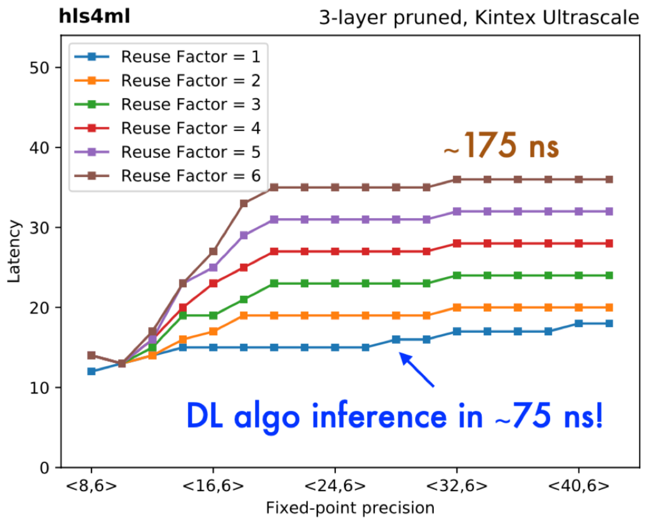
We are already applying Deep Learning here!

E.g. Real-time ML-based Inference

high level synthesis for machine learning

arxiv.1804.06913





Read our White Paper on how accelerated ML can be applied across many fields of fundamental physics!

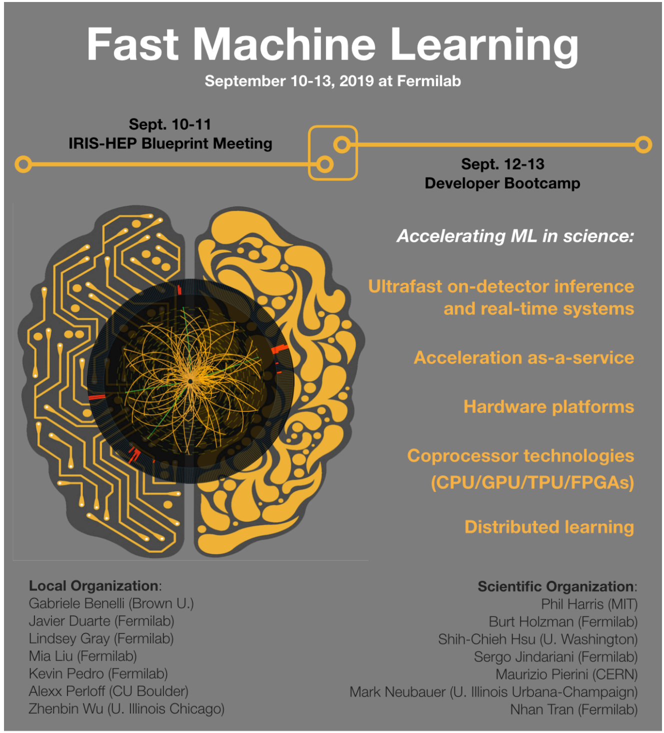
http://fastmachinelearning.org

E.g. Real-time ML-based Inference

Workshop on Sept 10-11, 2019* @ *Fermilab

Workshop Webpage & Report

Hosted at the Fermilab LPC and co-located with the FastML "developer bootcamp" which held tutorials and hackathon (195 registered participants!)



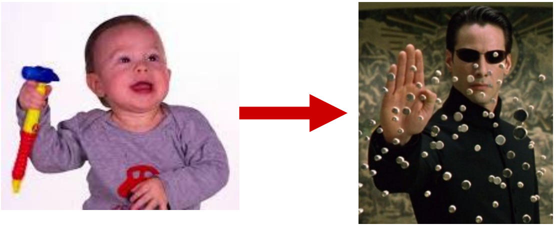
https://indico.cern.ch/e/FML

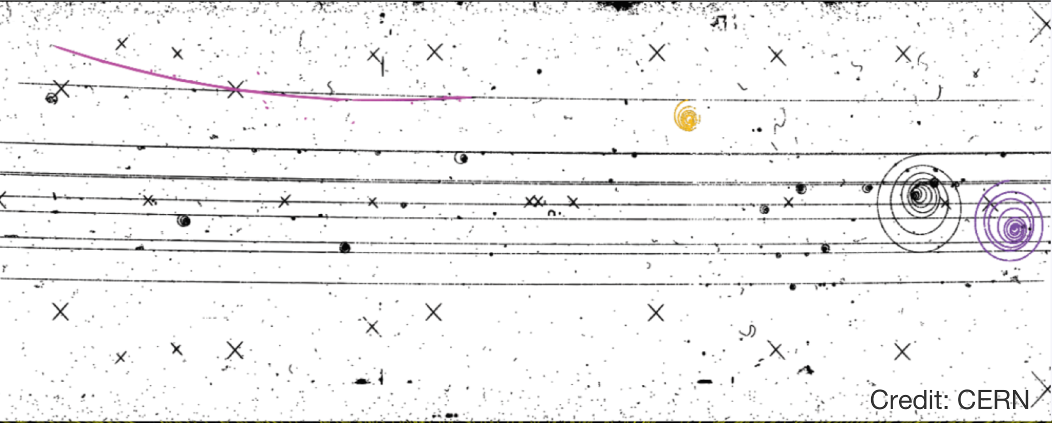
Read our White Paper on how accelerated ML can be applied across many fields of fundamental physics!

Fast Machine Learning and Inference

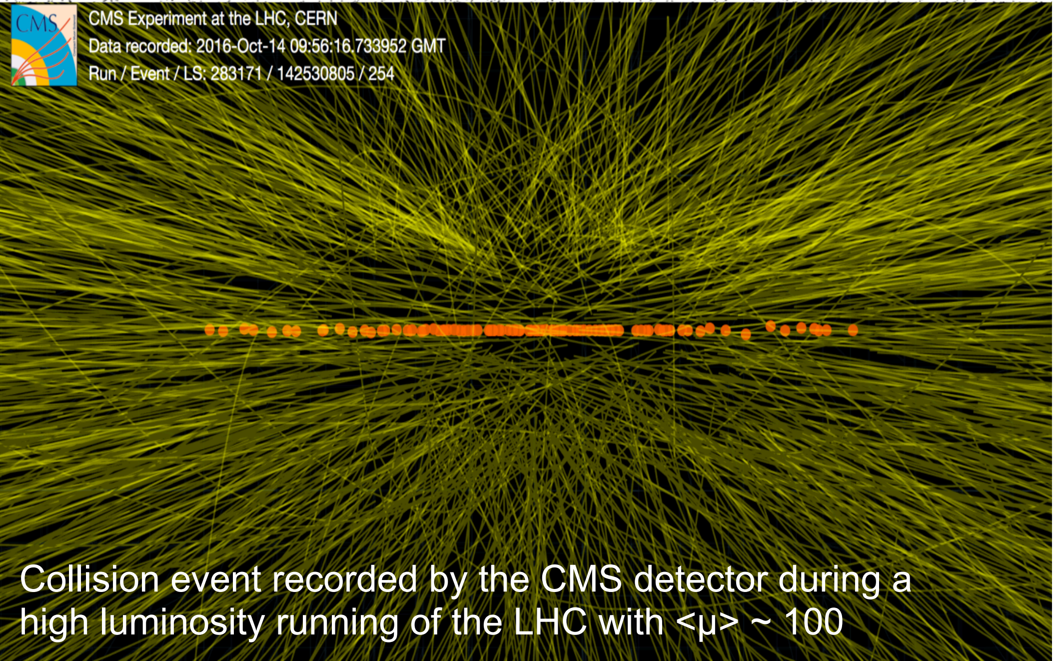
Looking forward...

- Many HEP problems can be recast as ML problems ML is rapidly playing a key role in addressing HL-LHC challenges & enabling new science capabilities
- Some key areas of ML for HEP going forward as I see it (some topics were covered in this talk, but many not):
    - Fast (accelerated, “real-time”) ML training and inference
    - Adversarial training, data augmentation
    - Supervised learning with raw detector information (“whole-event ML”)
    - Unsupervised learning, anomaly detection
    - Weakly supervised learning (e.g. CWoLa, Tag ‘N Track)
    - Reinforcement learning
    - Deep Generative Models for Fast Detector Simulation
    - Uncertainty quantification
    - Graph-based learning
    - Training and Workforce development
    - Physics-~~inspired~~ -driven network architectures





Credit: CERN



Collision event recorded by the CMS detector during a high luminosity running of the LHC with <μ> ~ 100

Bender: *Come on, Fry. I really wanna see it [the year 2000]. You know how I yearn for a simpler time... a time of barn dances and buggy rides before life was cheapened by heartless, high-tech machines.*

Leela: *But, Bender, you are a—*
Bender: *[dismissive] blah blah blah blah …*

Thanks!