Machine Learning @ CMS

Abhijith Gandrakota
for CMS MLG group

MLFP school, Georgia Tech

*Many slides from Javier Duarte

Machine Learning

Machine learning is the use of data to make predictions or decisions without explicit programming

Why? When did we use it in HEP ?

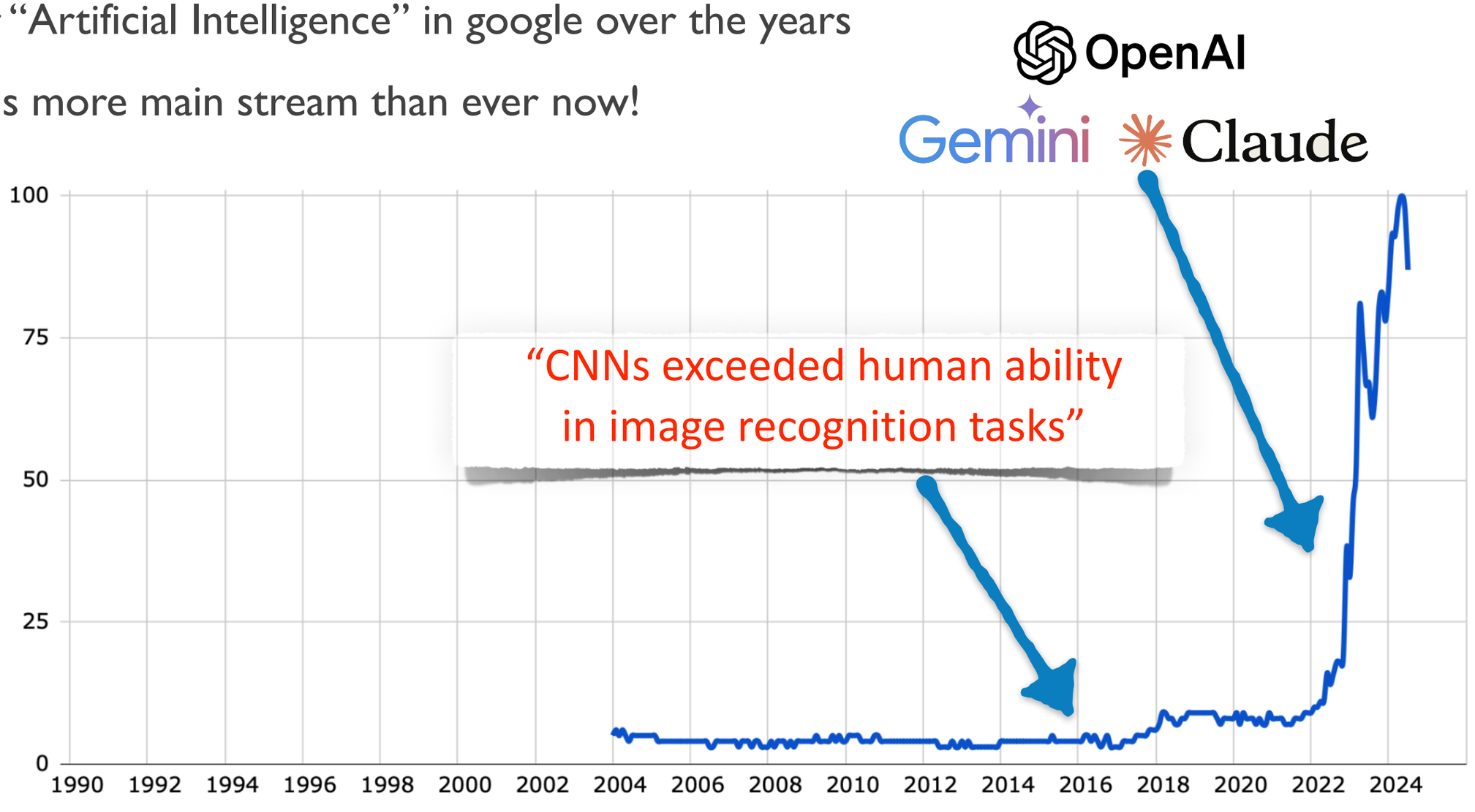
Machine Learning

Machine learning is the use of data to make predictions or decisions without explicit programming

Why? When did we use it in HEP ?

Searches for "Artificial Intelligence" in google over the years

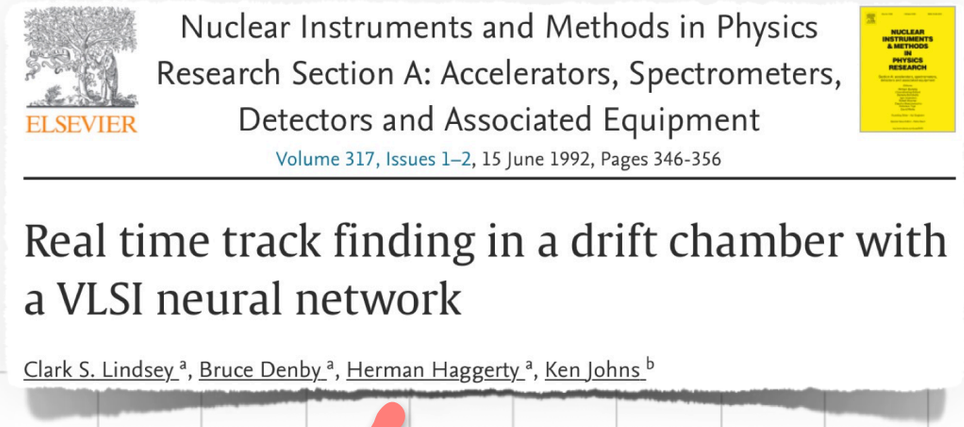
AI /ML is more main stream than ever now!



Machine Learning

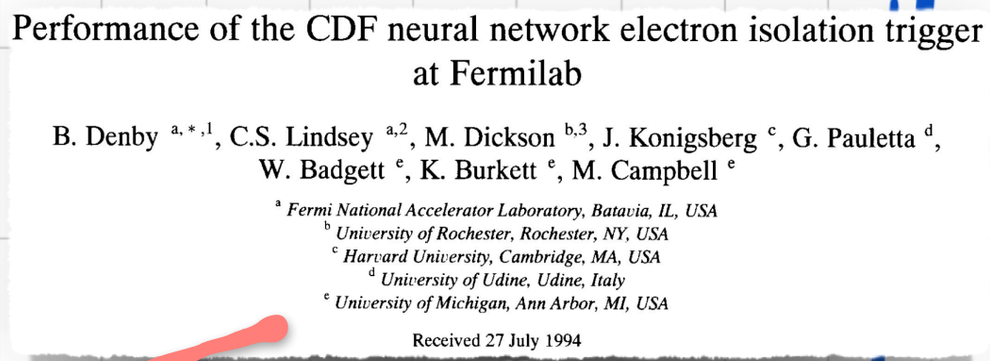
Machine learning is the use of data to make predictions or decisions without explicit programming

Why? When did we use it in HEP ?


Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment
Volume 317, Issues 1–2, 15 June 1992, Pages 346-356

Real time track finding in a drift chamber with a VLSI neural network

Clark S. Lindsey [a], Bruce Denby [a], Herman Haggerty [a], Ken Johns [b]


Performance of the CDF neural network electron isolation trigger at Fermilab

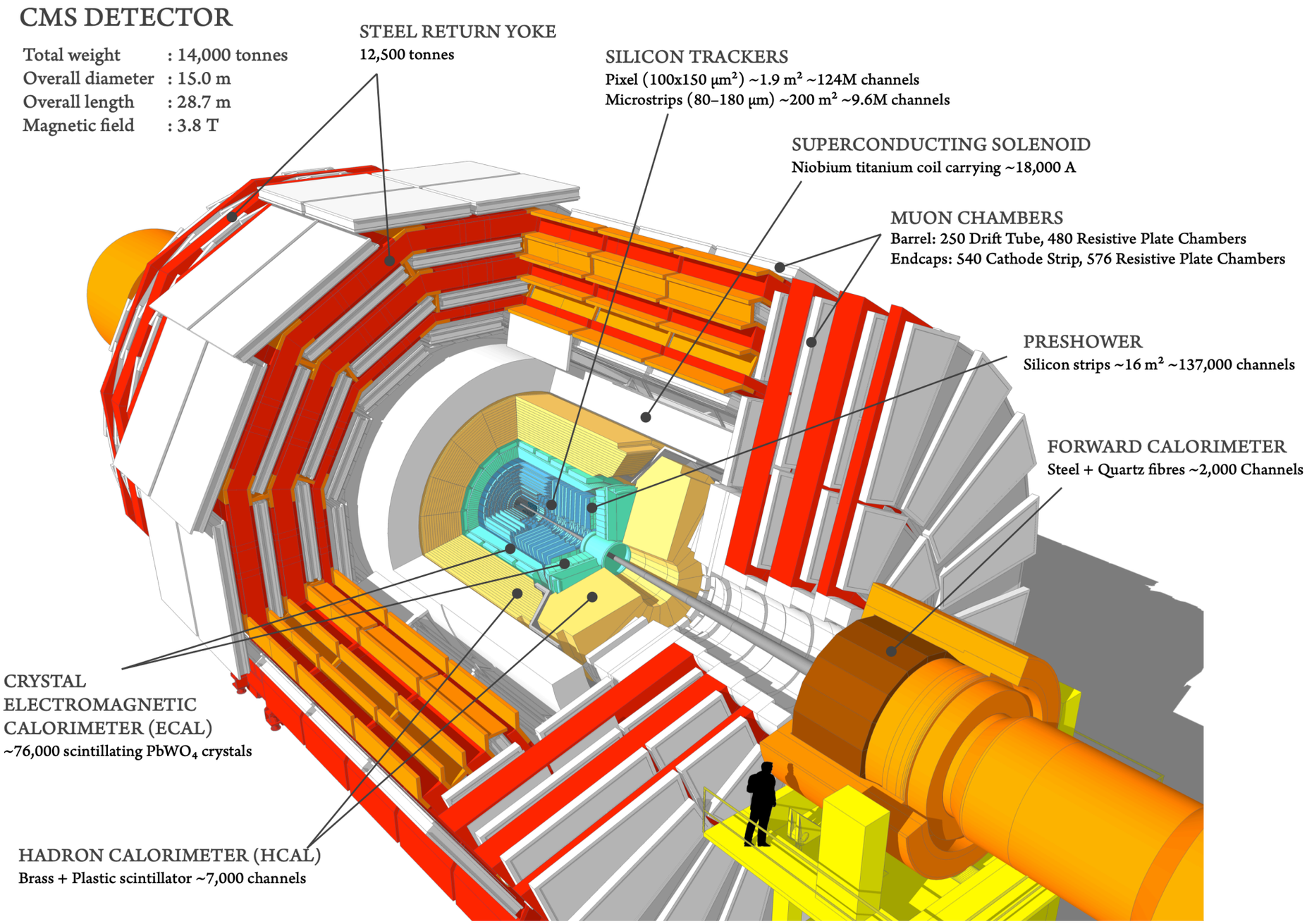
B. Denby [a,*,1], C.S. Lindsey [a,2], M. Dickson [b,3], J. Konigsberg [c], G. Pauletta [d], W. Badgett [e], K. Burkett [e], M. Campbell [e]

[a] Fermi National Accelerator Laboratory, Batavia, IL, USA
[b] University of Rochester, Rochester, NY, USA
[c] Harvard University, Cambridge, MA, USA
[d] University of Udine, Udine, Italy
[e] University of Michigan, Ann Arbor, MI, USA

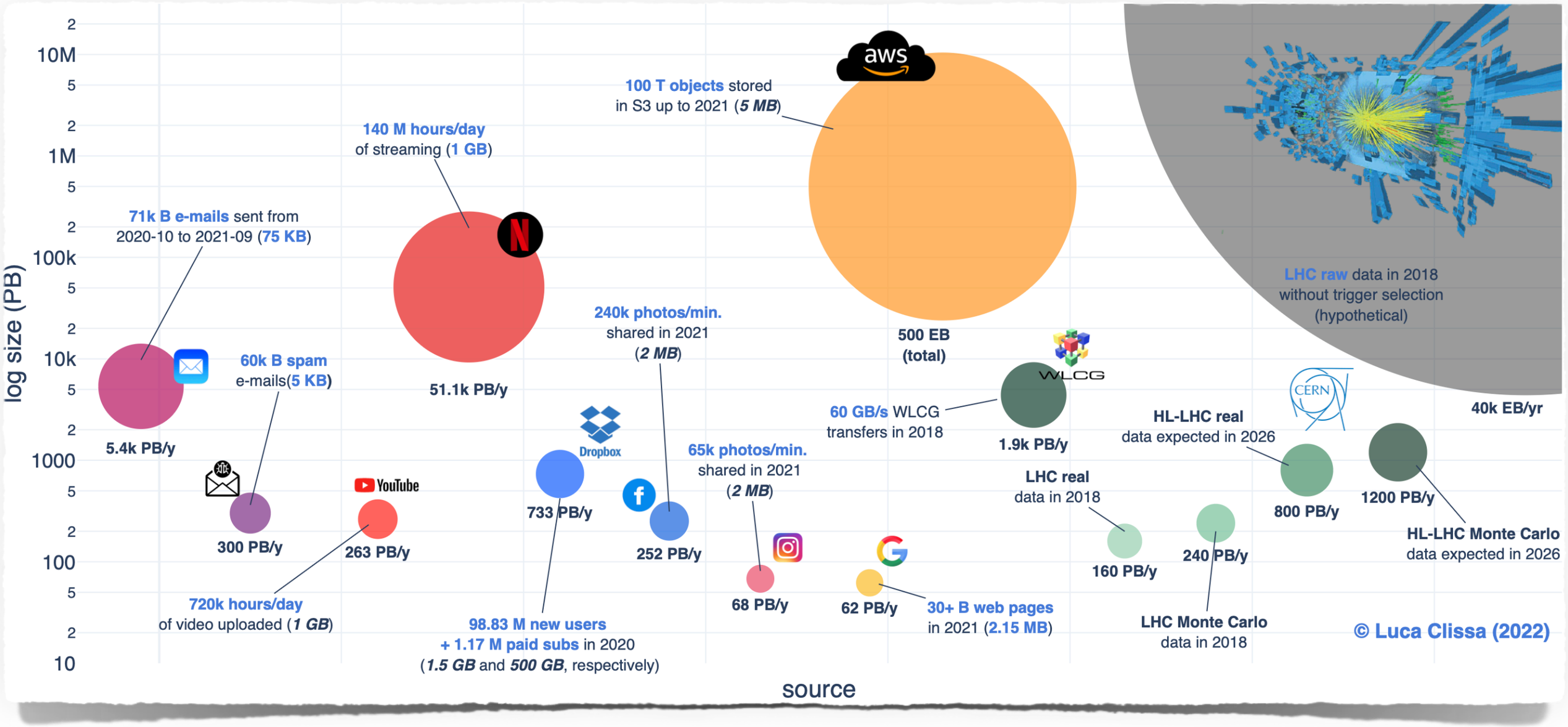
Received 27 July 1994

Compact Muon Solenoid

- Lots of readout channels from the detector
  - Data rates ~ 10x more than North American internet traffic!



Data from the LHC



https://arxiv.org/abs/2202.07659

Fermilab

Big data @ LHC

40 MHz proton-proton collision frequency
O(10³) particles per collision
O(10⁸) sensors
➜ **O(100) TB/s** data rates!

The Compact Muon Solenoid

Data reduction workflow @ LHC

99.75% events rejected!

Hardware based
Runs on FPGAs in real time
O(μs) latency

99% events rejected!

Software based
Runs on CPUs in real time
O(100 ms) latency

Software based
Run on data centers (LHC grid)
O(s) computing time

No data stored before this line

40 MHz

L1 trigger

100 kHz

High-Level trigger

1 kHz

1 MB/evt

Offline reconstruction

1 KHz
3-30 kB/evt

Run by individuals on the grid

Data volume

Computing time

Information

High-level analysis

Make physics discoveries with 0,0025% of the events!

(the rest is lost…)

No data
before th

Software based
Run on data centers (LHC grid)
O(s) computing time

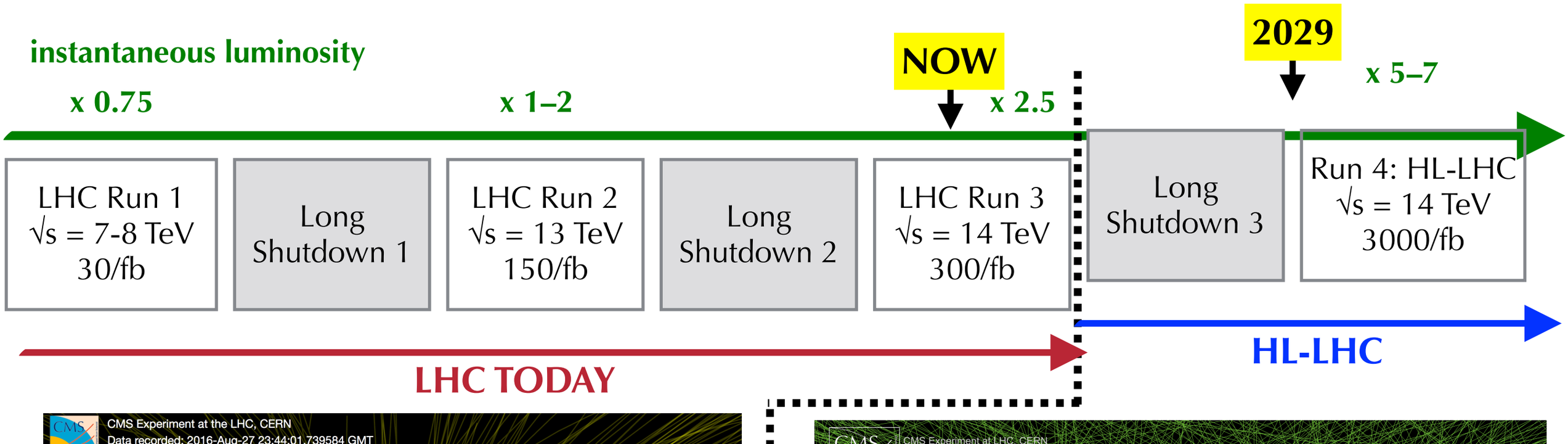
1 KHz
3-30 kB/evt

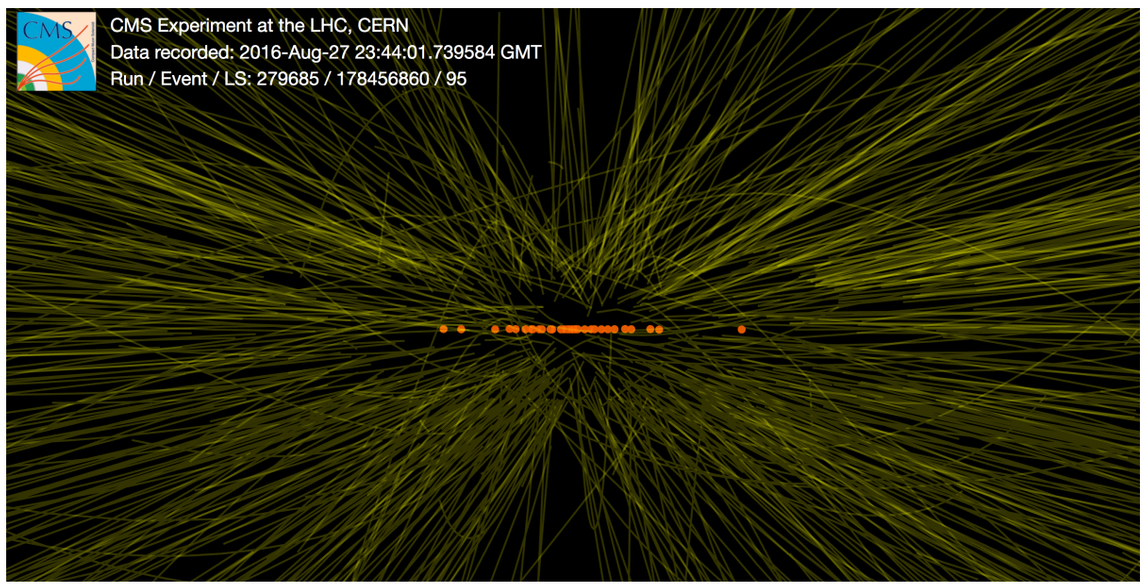
Run by
individuals
rid

40

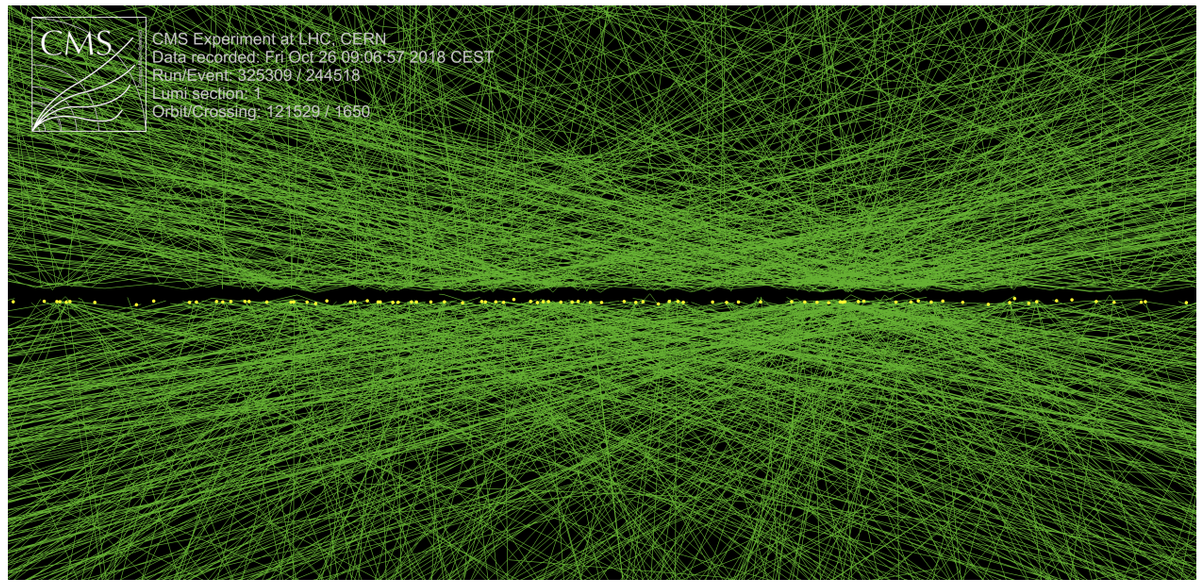
Data volume

The High-Luminosity LHC challenge





40 simultaneous collisions per bunch crossing

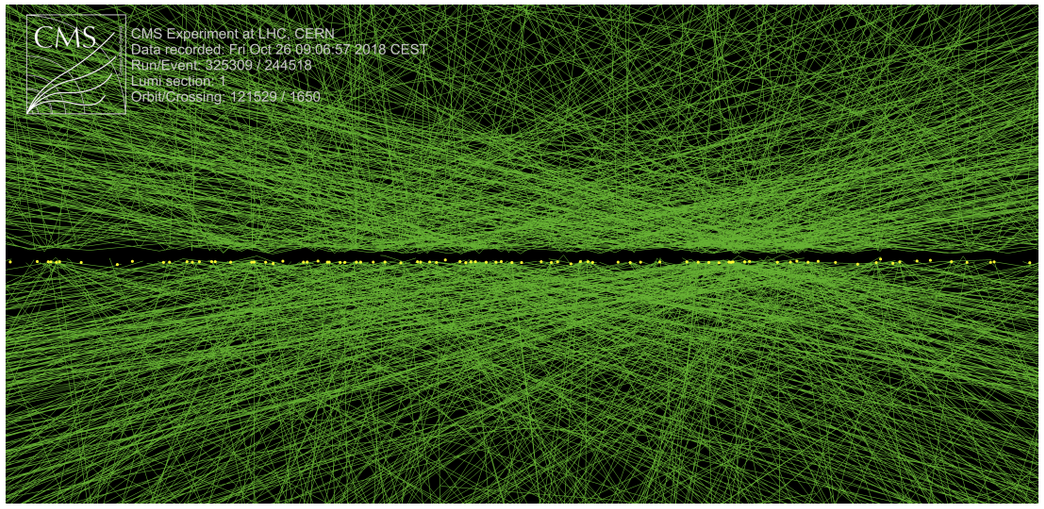


200 simultaneous collisions per bunch crossing + more granular detector!

The HL-LHC challenge

With more particles per collision and more readout channels to combine, the reconstruction to become even more computing intensive

We cannot throw away more data
We must increase the throughput
with at most flat budget for computing resources
We must do more with less
to preserve the physics!

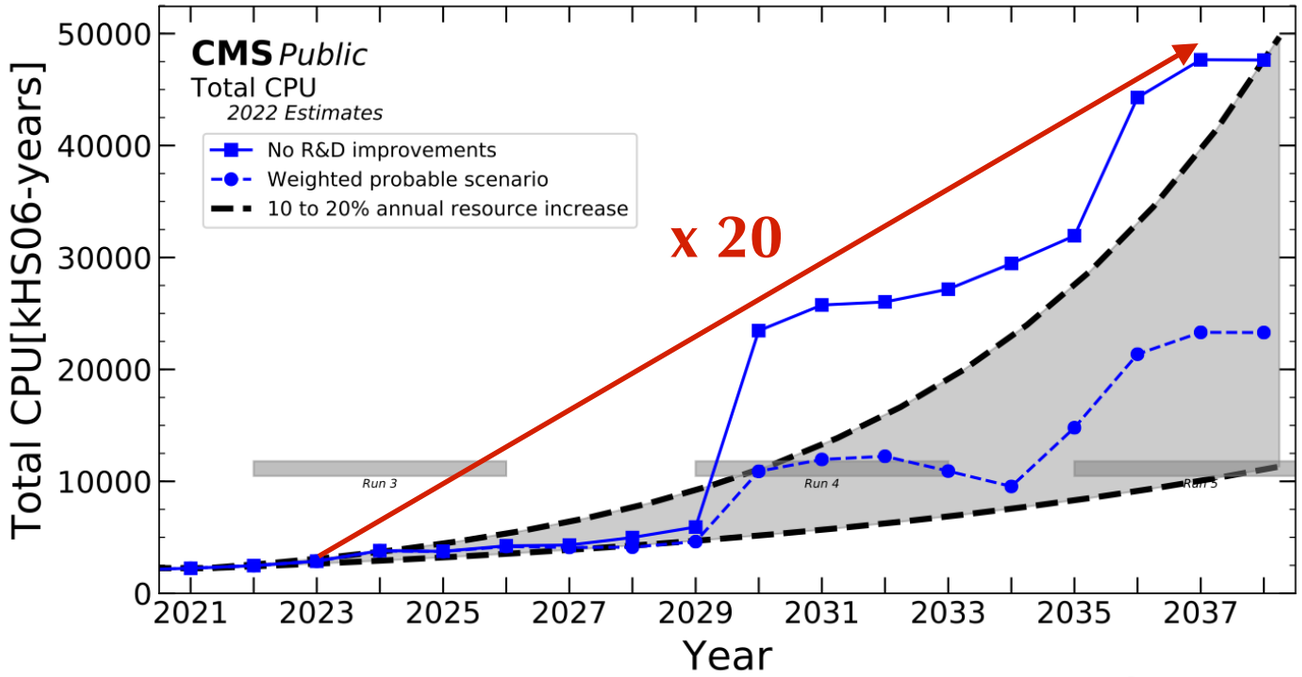


x5 → **Event complexity**

x50 → **Processing time**

x10 → **Data**

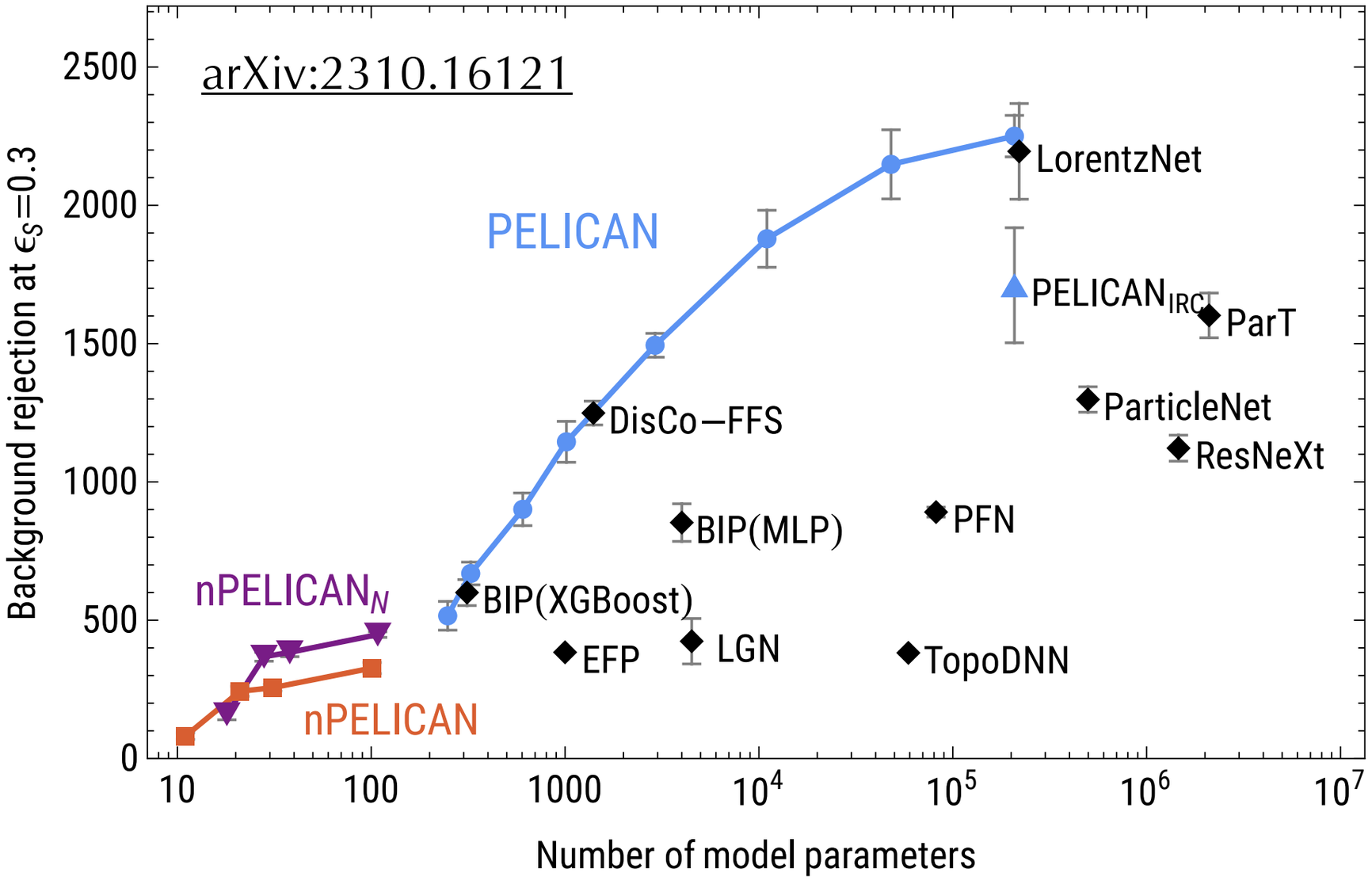
x20 → **Computing resources/ Disk storage**



Source

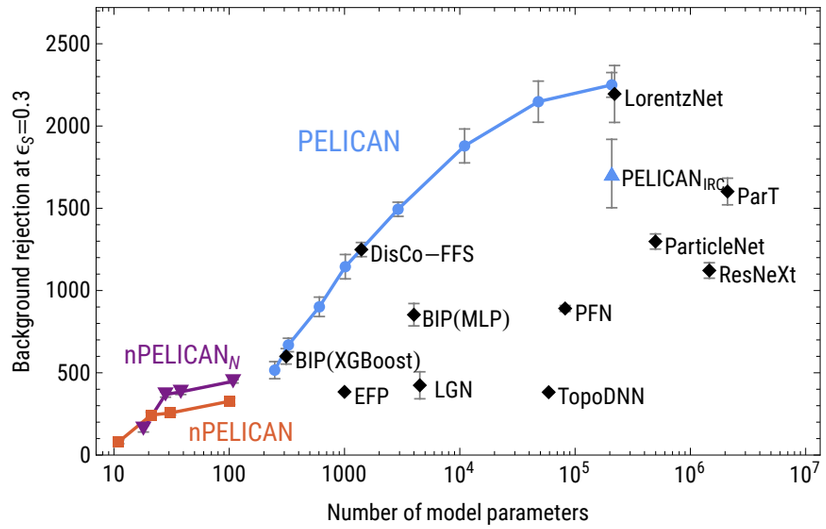
ML Advantages

- High accuracy
  - improved physics performance:
    expressive models can extract most useful info from complex datasets

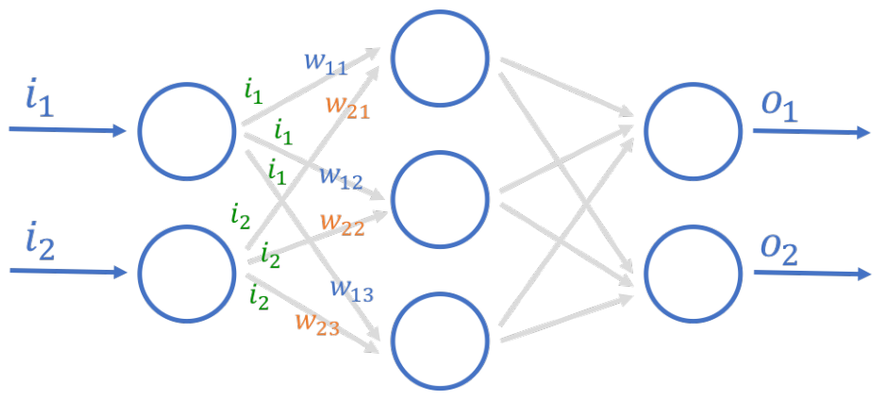


ML Advantages

- High accuracy
  - improved physics performance: capacity to extract most useful info from complex datasets



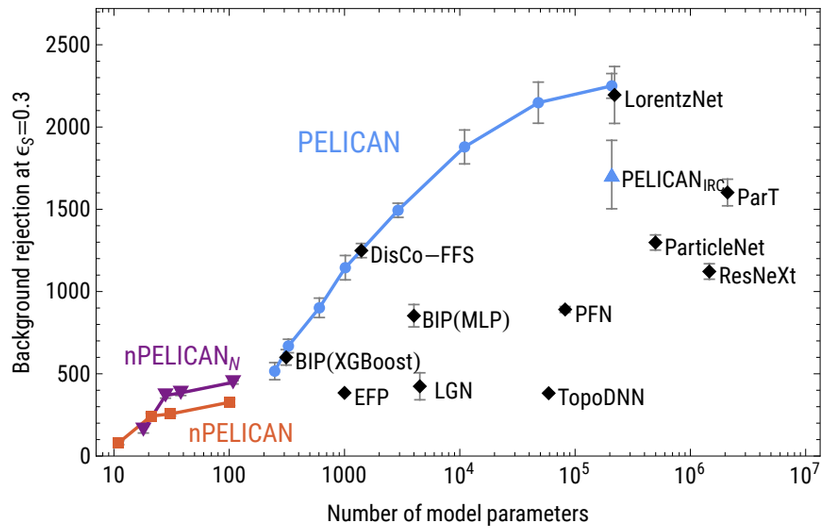
- Fast speed
  - matrix multiplication can be massively parallelized
    - Leveraging GPUs/FPGAs, and other modern AI chips
  - take advantage of reduced precision
  - reduced development time: adaptive, generalizable



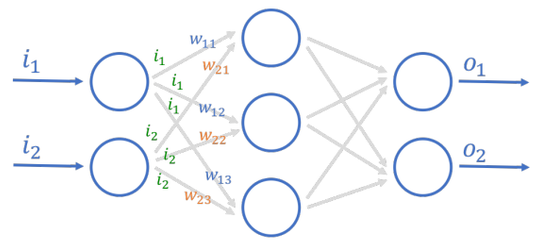
$$\begin{bmatrix} w_{11} & w_{21} \\ w_{12} & w_{22} \\ w_{13} & w_{23} \end{bmatrix} \cdot \begin{bmatrix} i_1 \\ i_2 \end{bmatrix} = \begin{bmatrix} (w_{11} \times i_1) + (w_{21} \times i_2) \\ (w_{12} \times i_1) + (w_{22} \times i_2) \\ (w_{13} \times i_1) + (w_{23} \times i_2) \end{bmatrix}$$

ML Advantages

- High accuracy
  - improved physics performance: capacity to extract most useful info from complex datasets
  - better scaling with data complexity (HL-LHC)

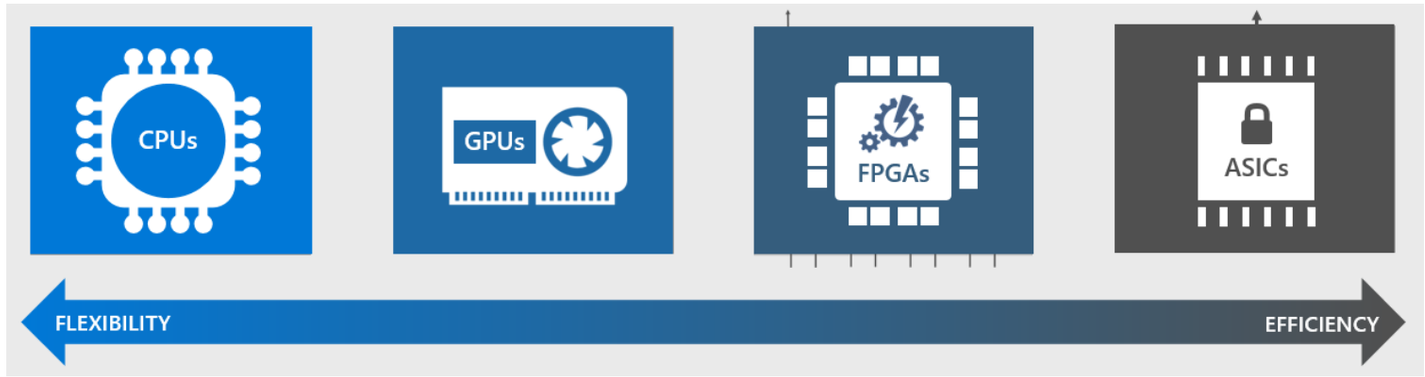


- Fast speed
  - reduced development time
  - matrix multiplication can be massively parallelized
  - take advantage of reduced precision



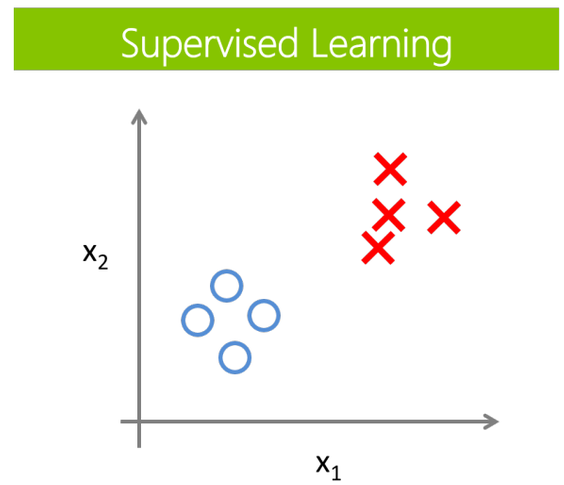
$$\begin{bmatrix} w_{11} & w_{21} \\ w_{12} & w_{22} \\ w_{13} & w_{23} \end{bmatrix} \cdot \begin{bmatrix} i_1 \\ i_2 \end{bmatrix} = \begin{bmatrix} (w_{11} \times i_1) + (w_{21} \times i_2) \\ (w_{12} \times i_1) + (w_{22} \times i_2) \\ (w_{13} \times i_1) + (w_{23} \times i_2) \end{bmatrix}$$

- Better portability
  - many dedicated processors: GPUs, FPGAs, TPU, IPU, …
  - large investment in tools to compile and optimize ML models for the hardware

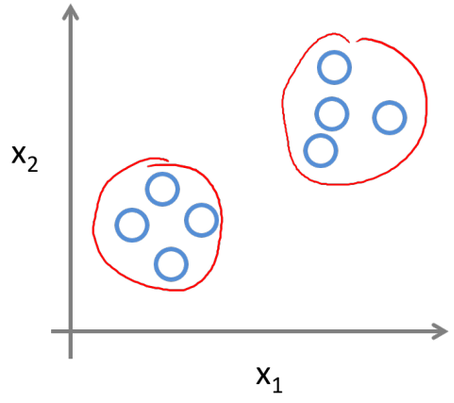


Types of learning

- **Supervised learning:**
  - attempt to predict/infer some target (truth label)
  - classification or regression target
  - train on data with known label (often MC simulation)



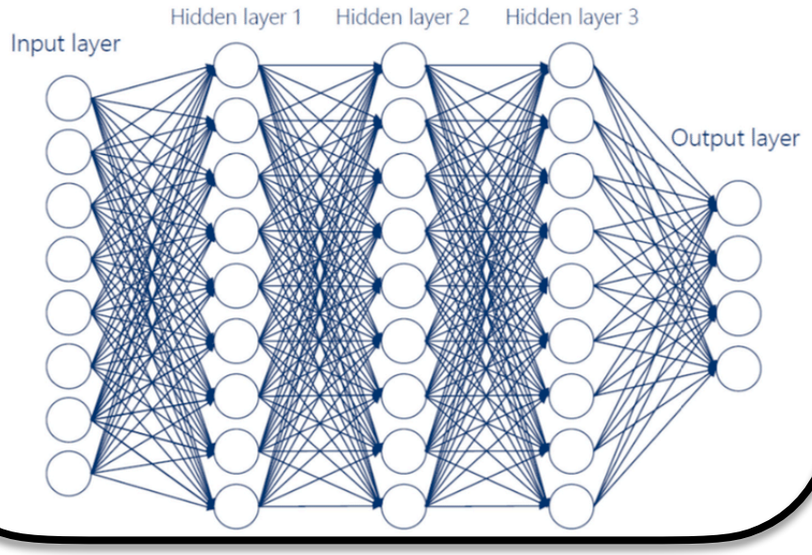
- **Unsupervised learning:**
  - target is not given
  - learn the data underlying distribution directly from data
  - can be used for sampling, clustering, anomaly detection, …



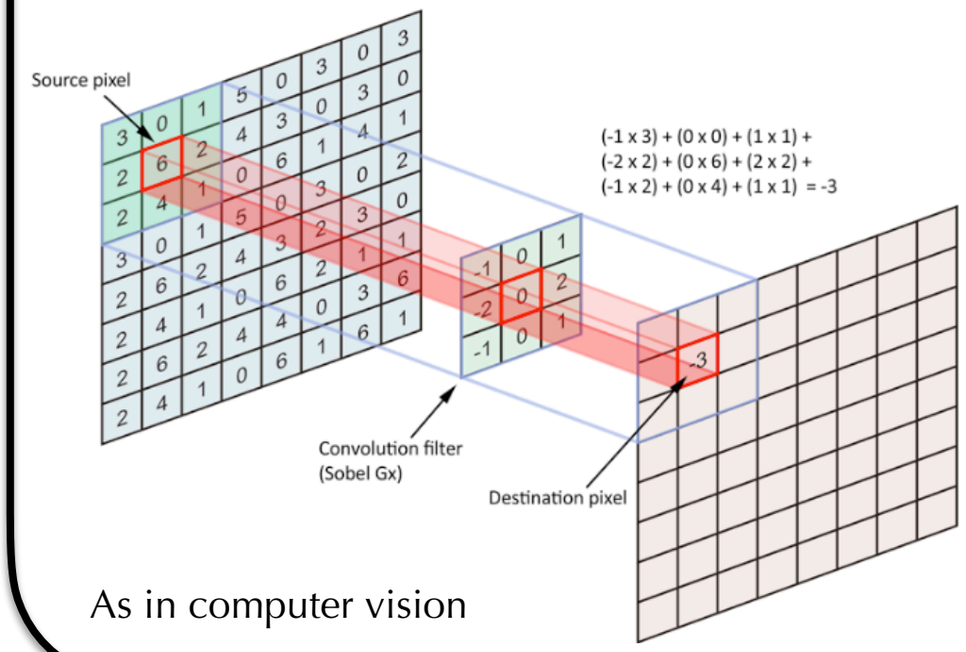
- **Weakly/semi-supervised learning**:
  - Only noisy or partial targets are known

Data representations

High level/tabular: fully connected NN, BDTs

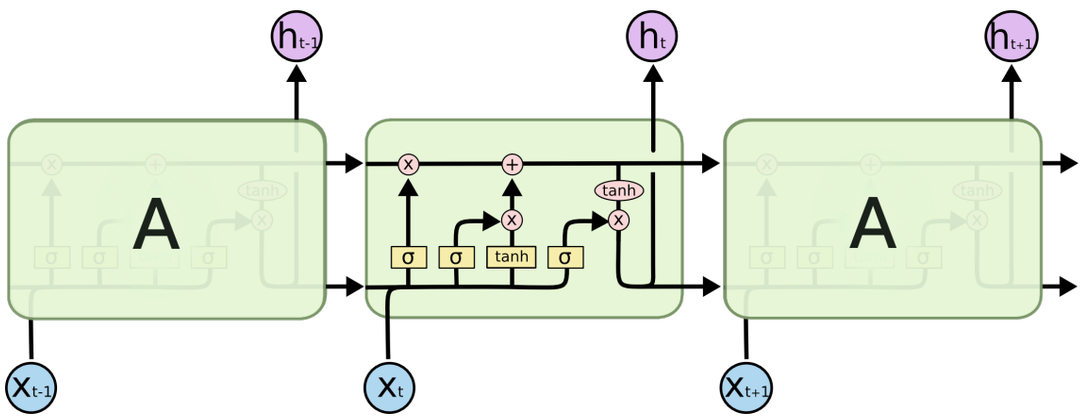


Regular grid: convolutional NN



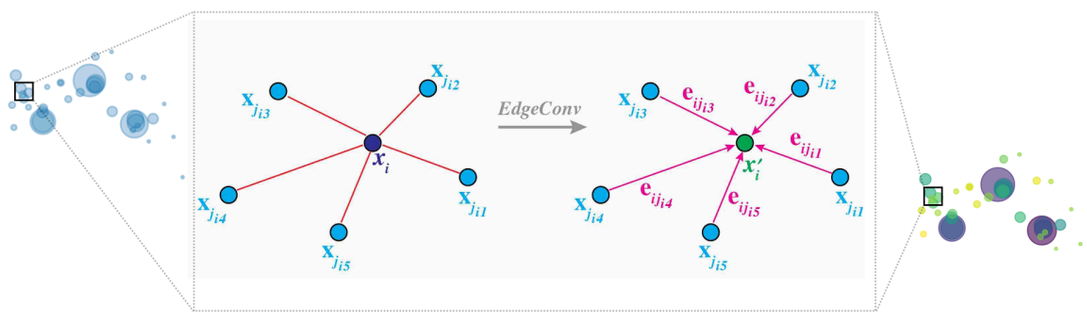
As in computer vision

Ordered sequence/time series: recurrent NN, transformers



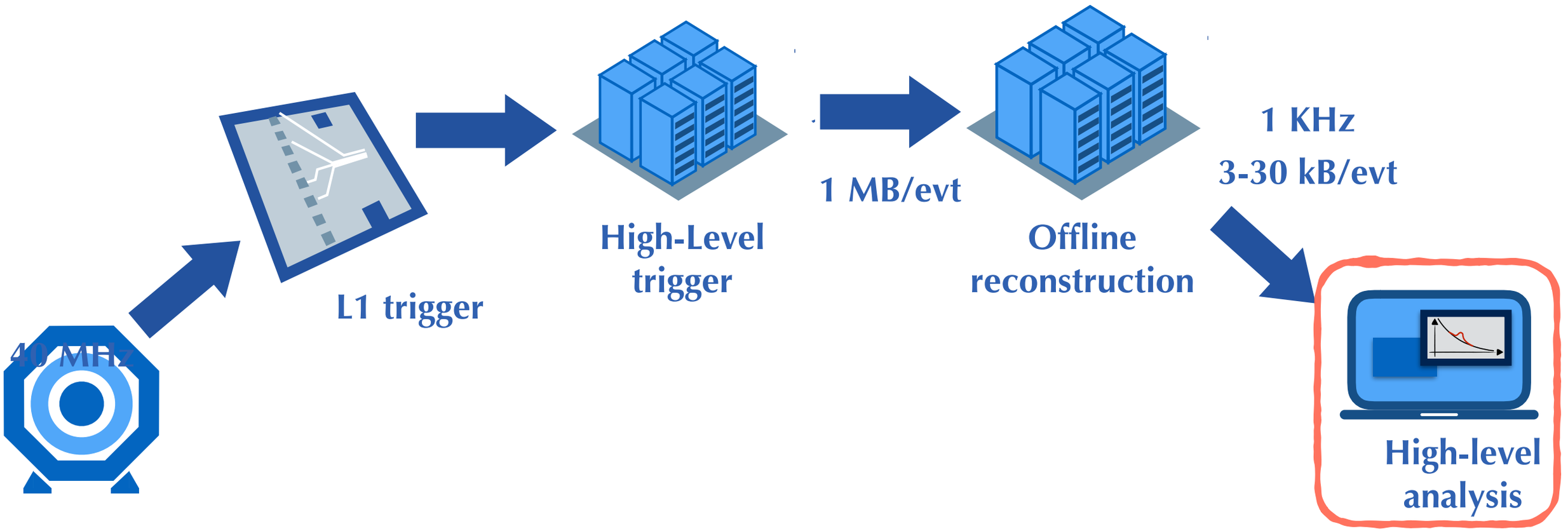
As in natural language processing

Point cloud: Deep sets, graph NN, transformers



As in social media analysis

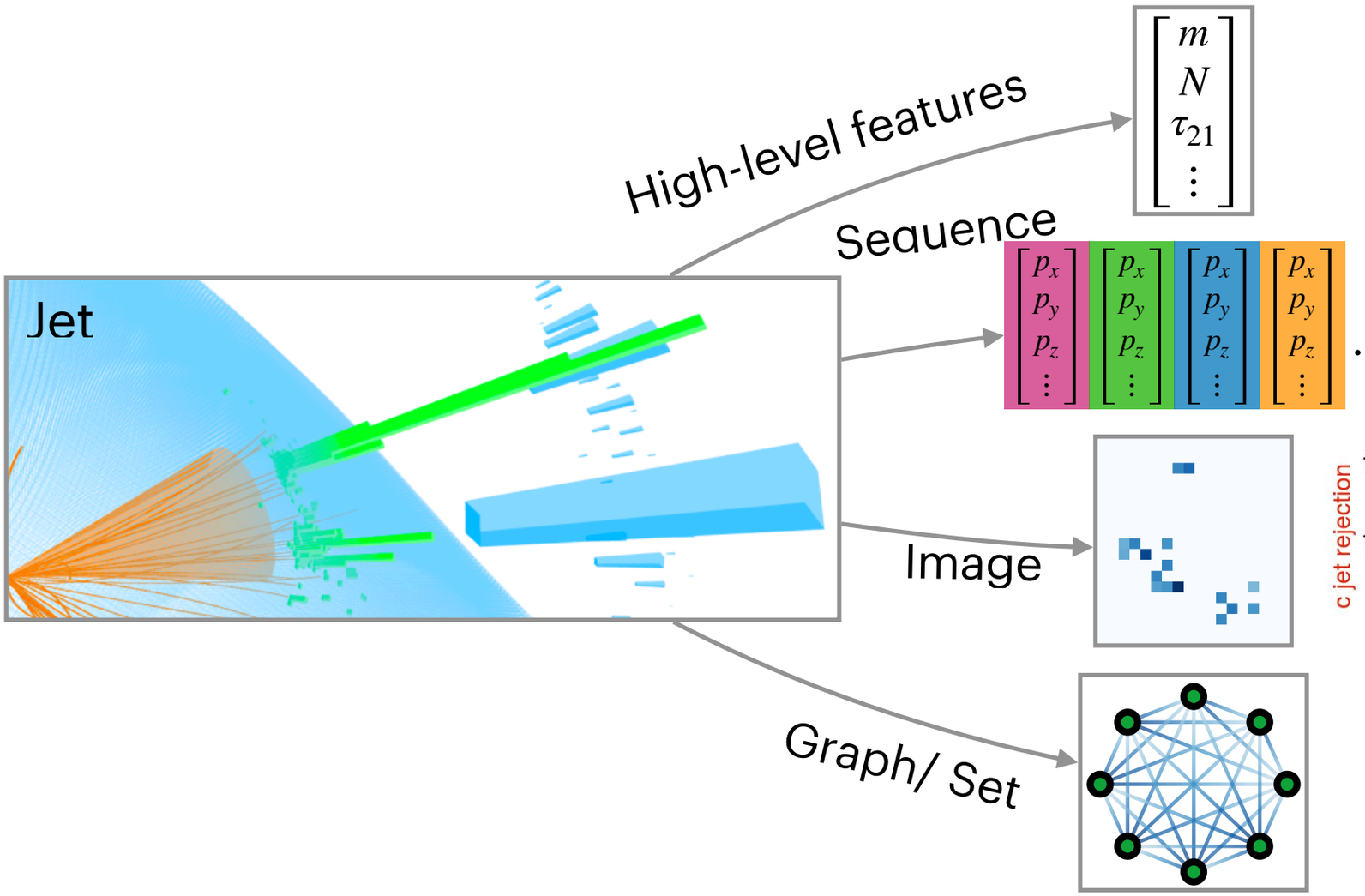
Data reduction workflow @ LHC



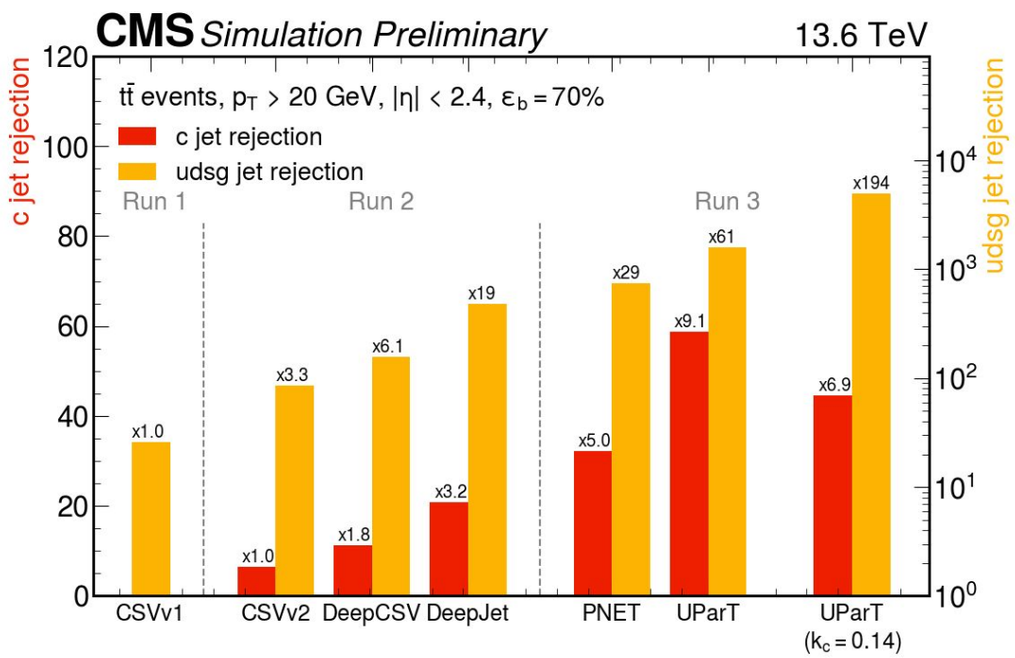
Taggers: From BDTS toTransformers

- Great strides made to leverage **rich low-level information** with **graph neural networks** and **transformers** for a variety of tasks including **jet classification**

CMS-DP-2024-066

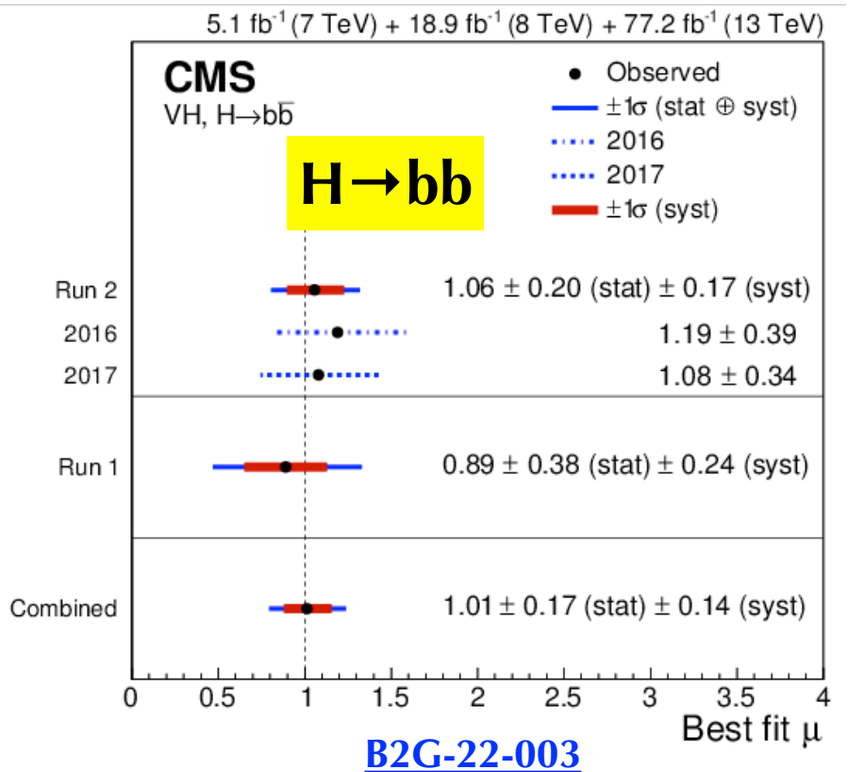


- Major improvements demonstrated in CMS (ParticleNet/Transformer)

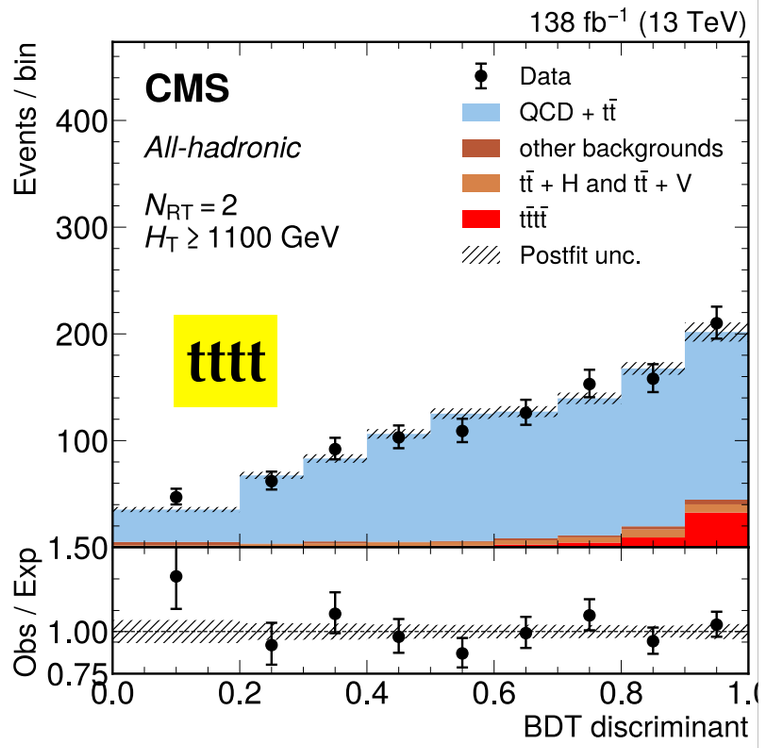


Impact: Better searches and measurements

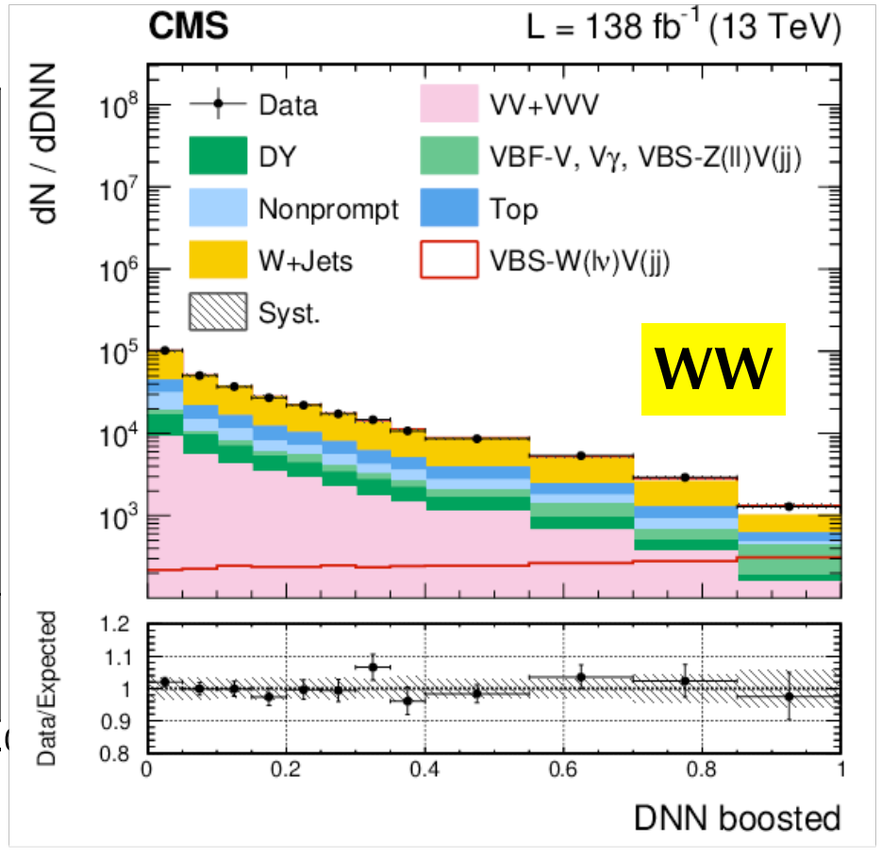
HIG-18-016



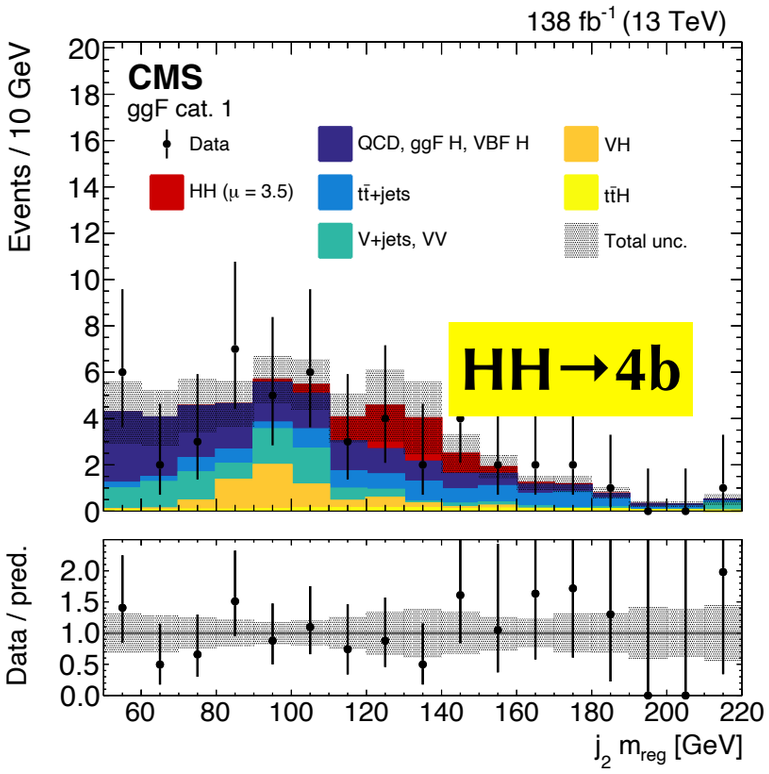
TOP-21-005



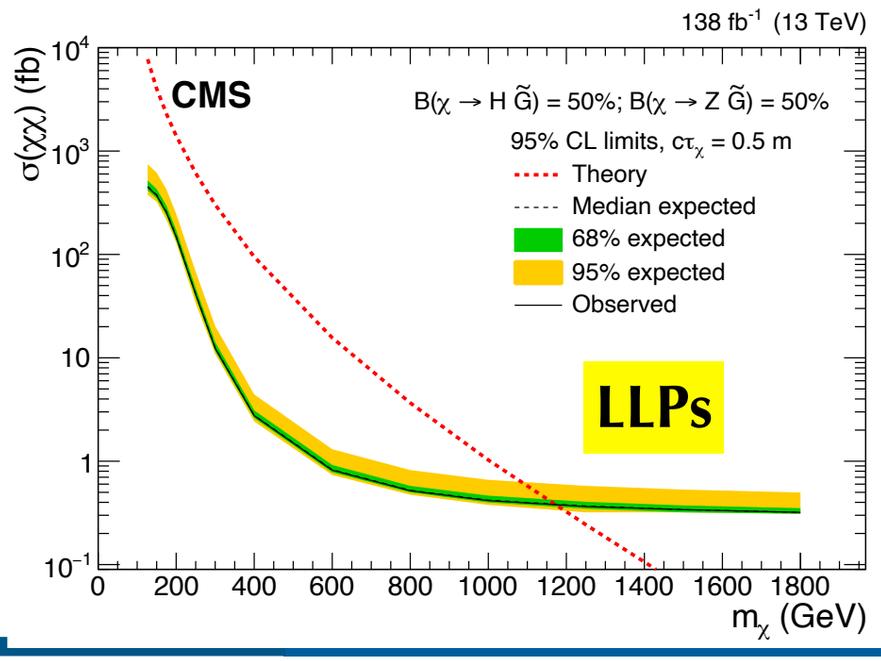
SMP-20-013



B2G-22-003



EXO-21-014



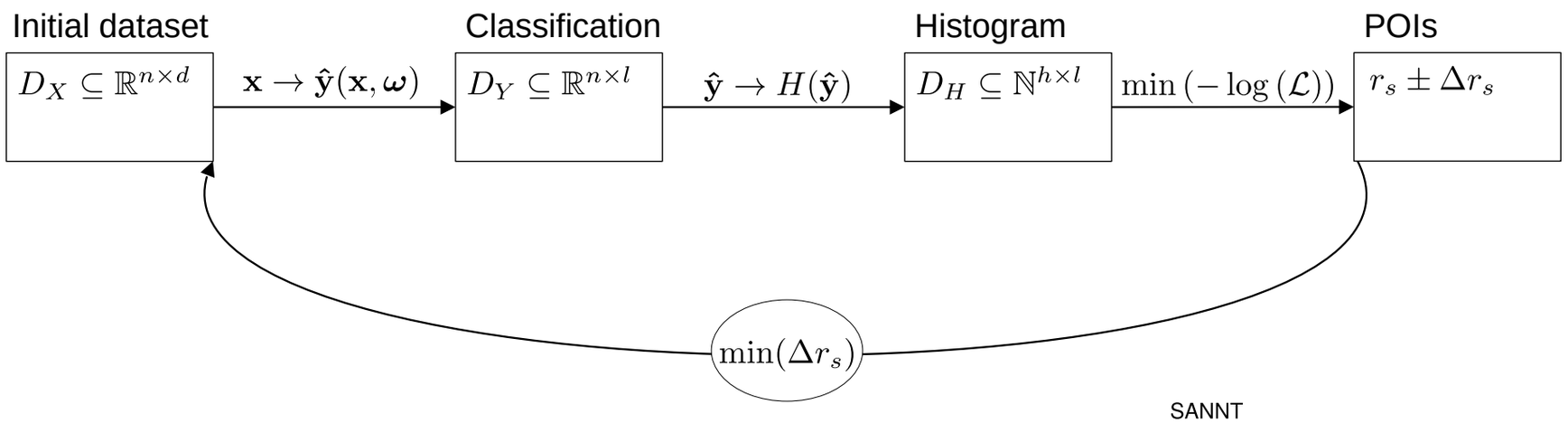
- Increased sensitivity with ML used by each PAG (mostly supervised):
  - improved event classification and object reconstruction

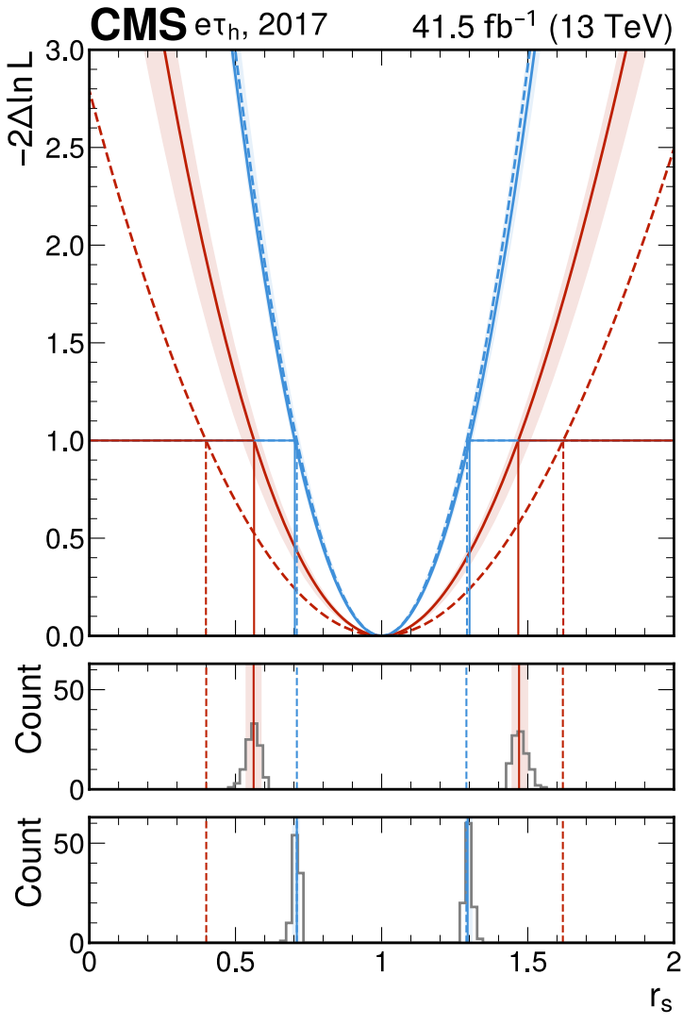
Systematic-aware learning

arXiv:2502.13047

- Standard classifier training (CENNT) optimizes for signal vs. background discrimination without considering systematics and other effects that affect the ultimate figure: uncertainty Δr_s on a physics parameter r_s

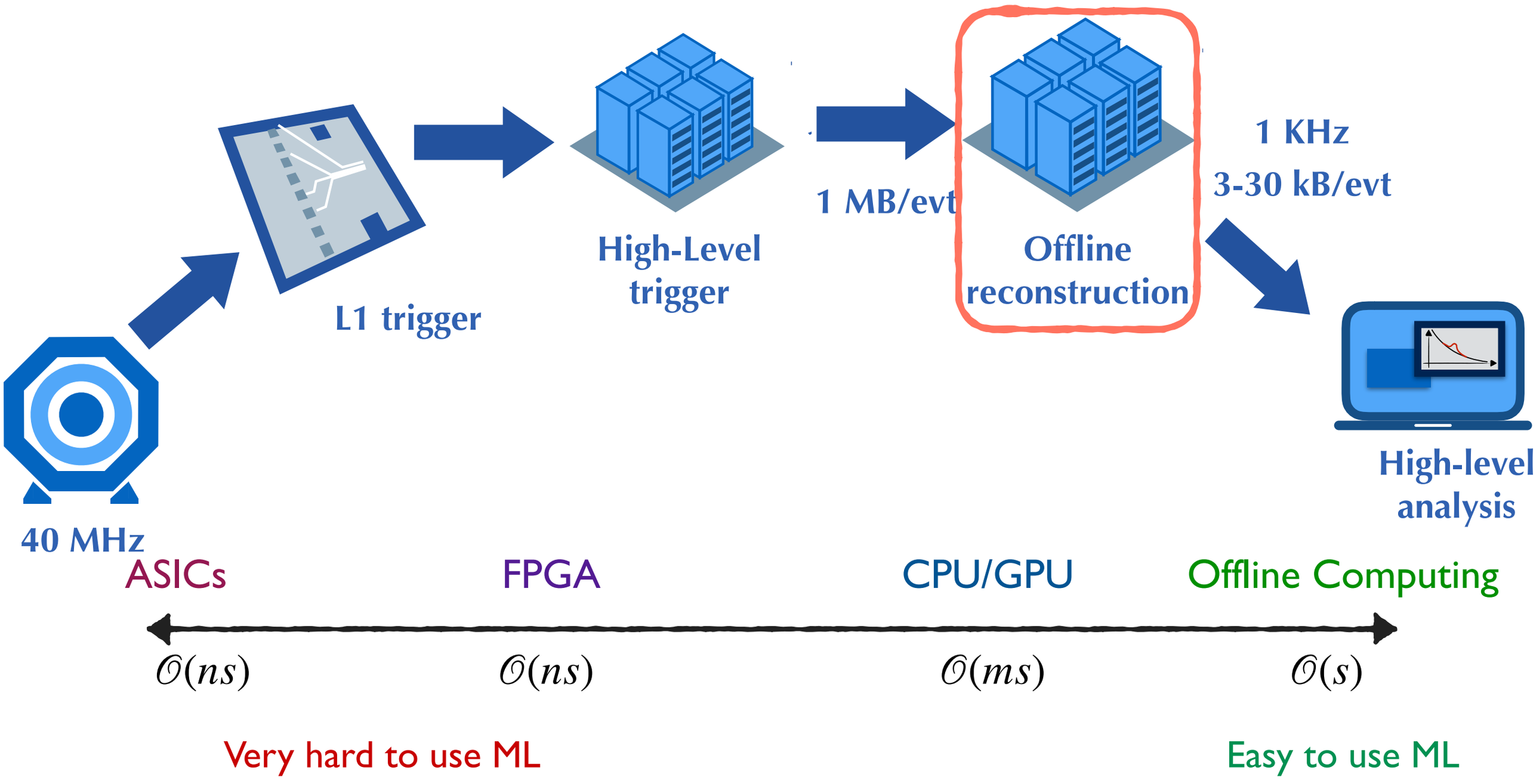
- By implementing the analysis chain (including systematics) in a **differentiable way**, we can directly optimize for min. Δr_s in the neural network training!





Fermilab

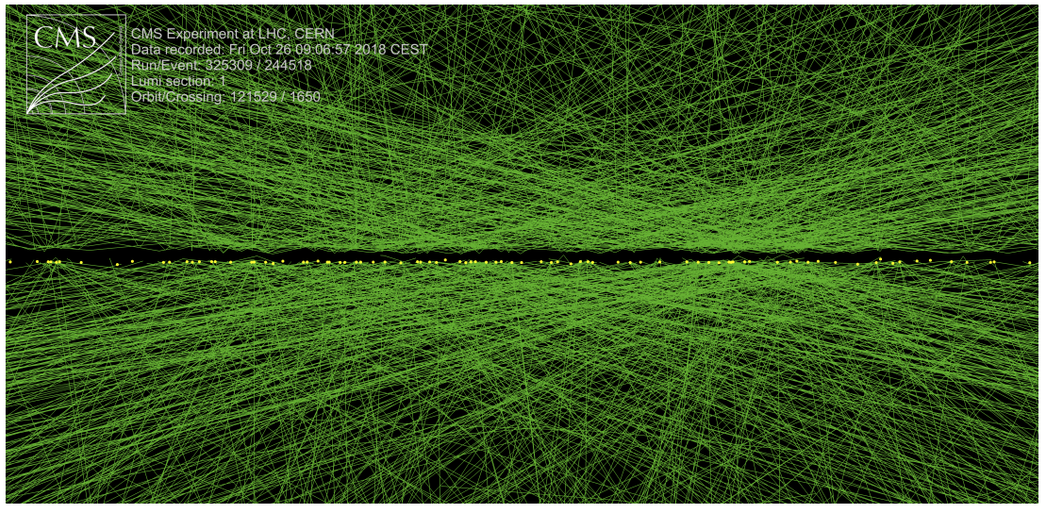
Data reduction workflow @ LHC



The HL-LHC challenge

With more particles per collision and more readout channels to combine, the reconstruction to become even more computing intensive

We cannot throw away more data
We must increase the throughput
with at most flat budget for computing resources
We must do more with less
to preserve the physics!

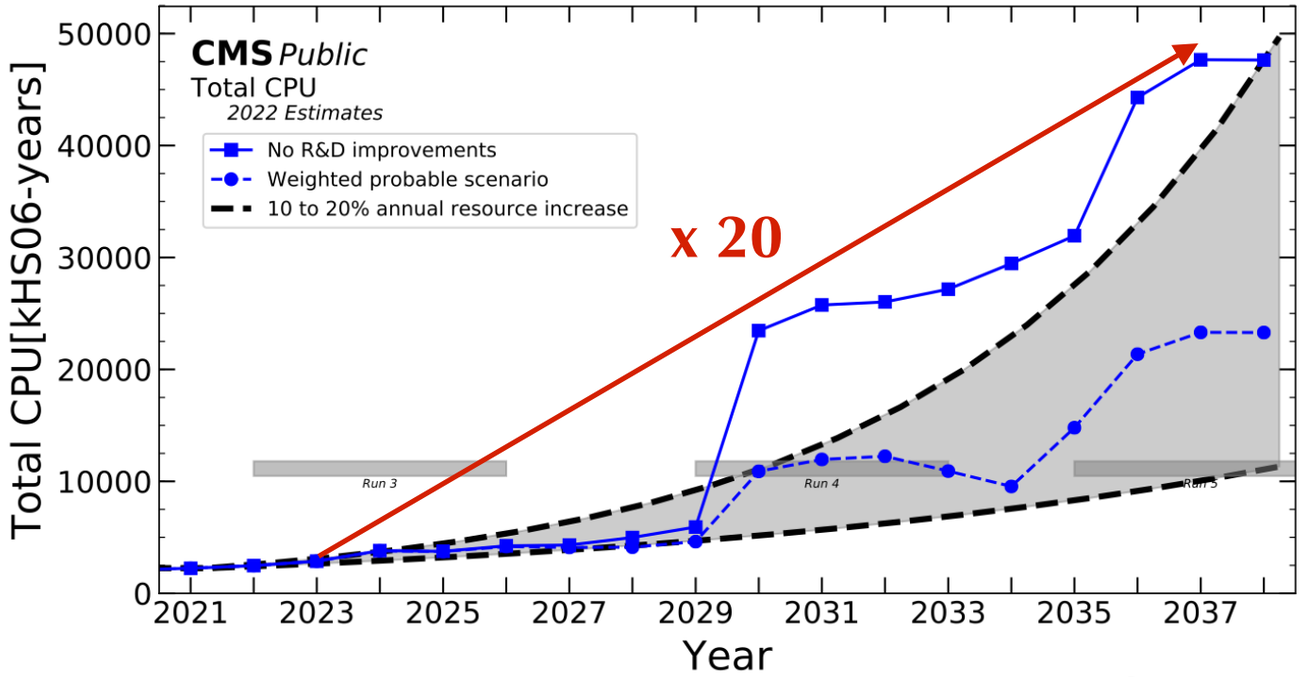


x5 → **Event complexity**

x50 → **Processing time**

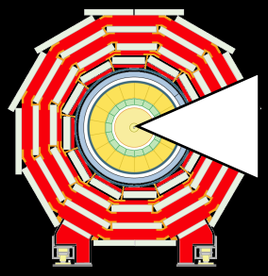
x10 → **Data**

x20 → **Computing resources/ Disk storage**



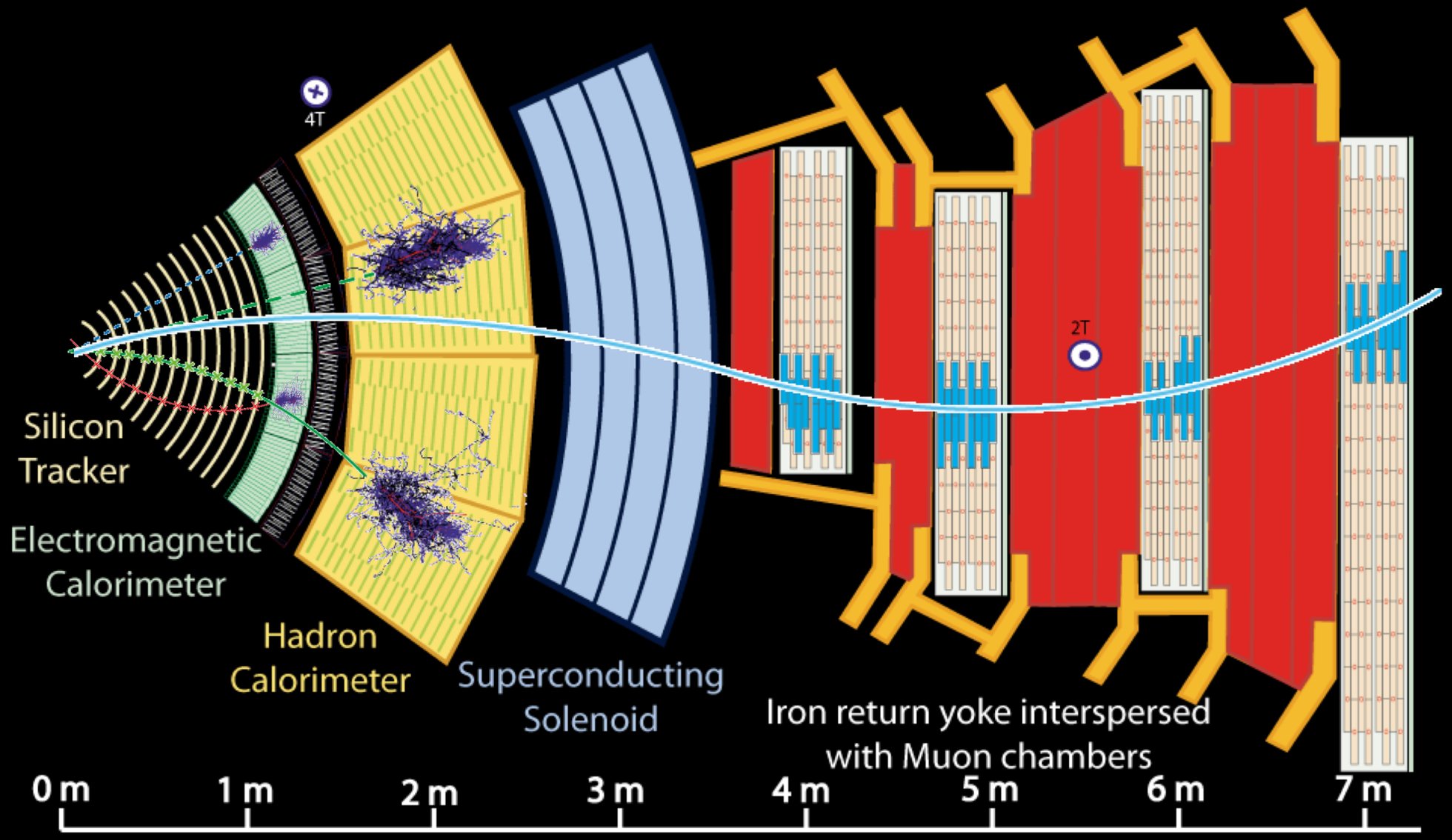
Source

Multilayered Detectors, e.g. CMS



Current and future multilayered detectors…

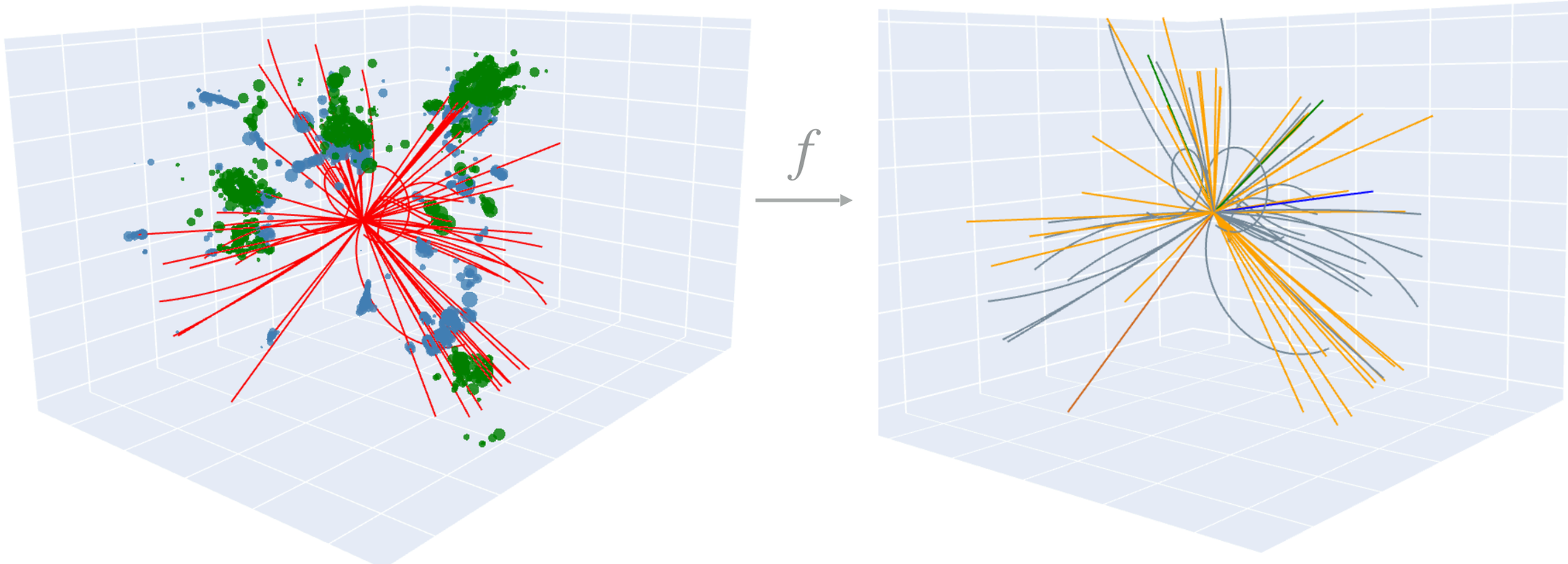
Require complex pattern recognition



Particle-flow as a Machine-Learning task

- Can we formulate PF as an ML task (naturally "tunable" through re-training and portable to new hardware)?

- Learn a "set-to-set" function $f: X \rightarrow Y$, where $\{\text{tracks, energy clusters}\} \in X$ and $\{\text{particles}\} \in Y$



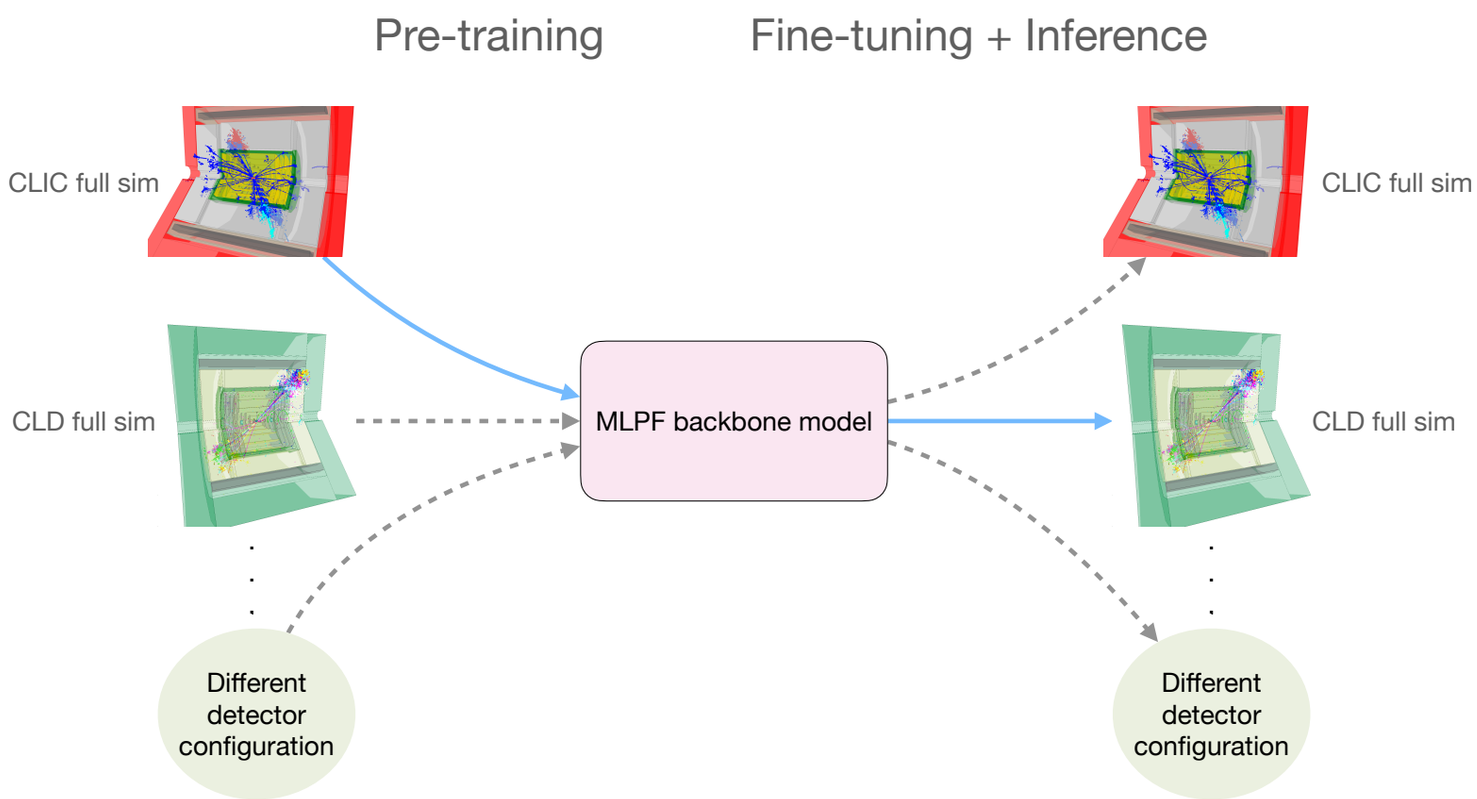
$\{\text{tracks, energy clusters}\} \in X$ $\{\text{particles}\} \in Y$

arXiv:2503.00131

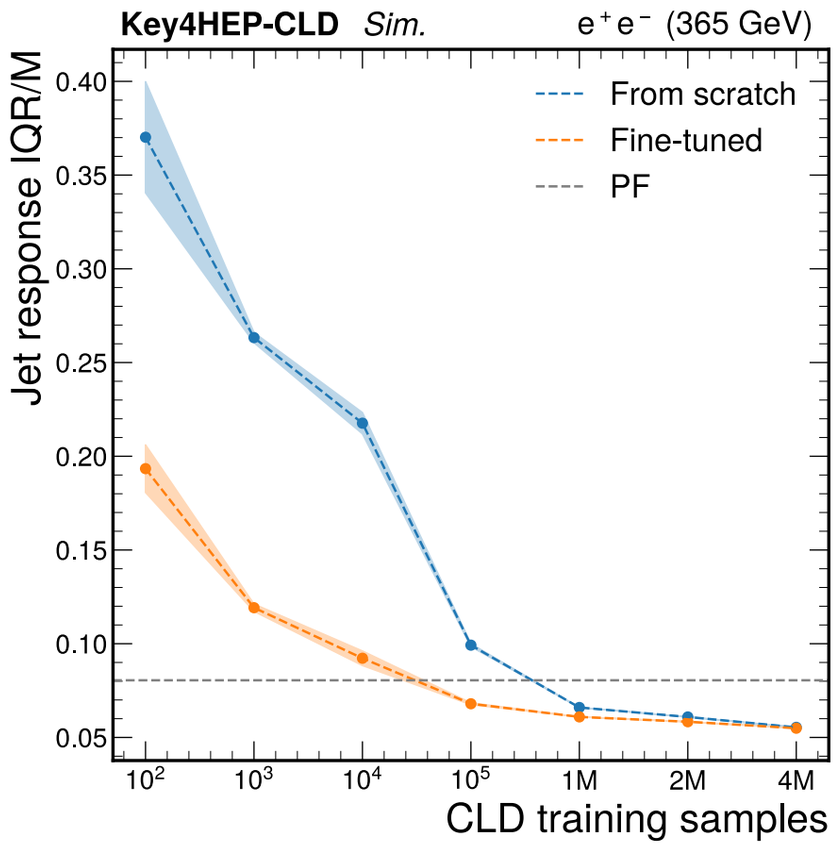
Cross-detector Transfer learning study

- Pre-train MLPF model on CLIC full sim. samples then fine-tune on full sim. samples from CLD at FCC-ee

- Fine-tuned model outperforms model trained from scratch for <2M events, with 30% better jet resolution at 100k events, and outperforms PF baseline

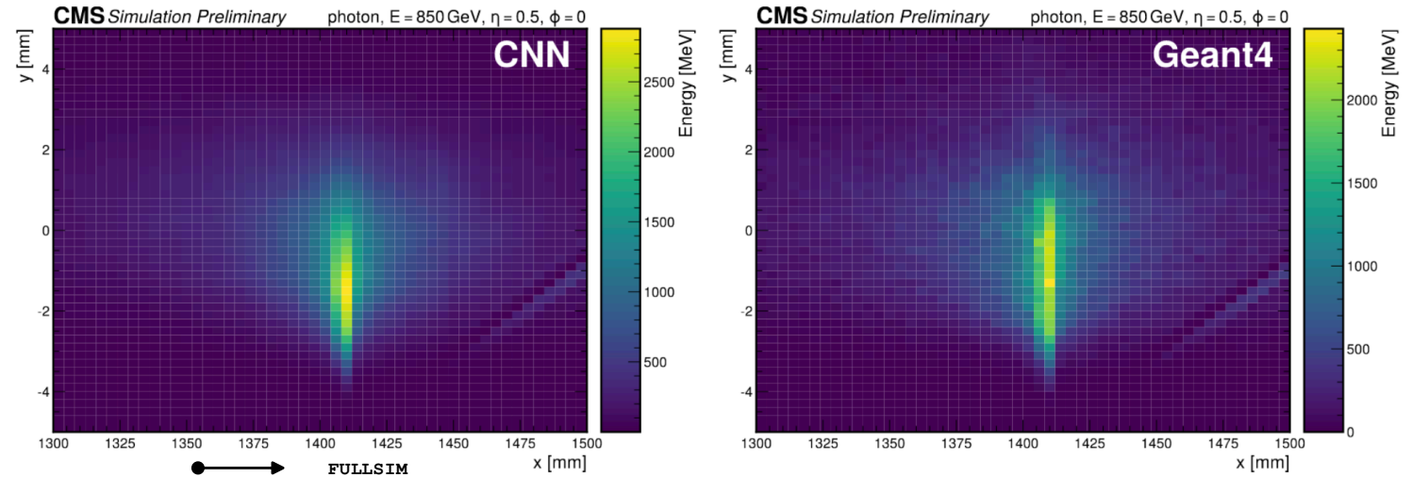


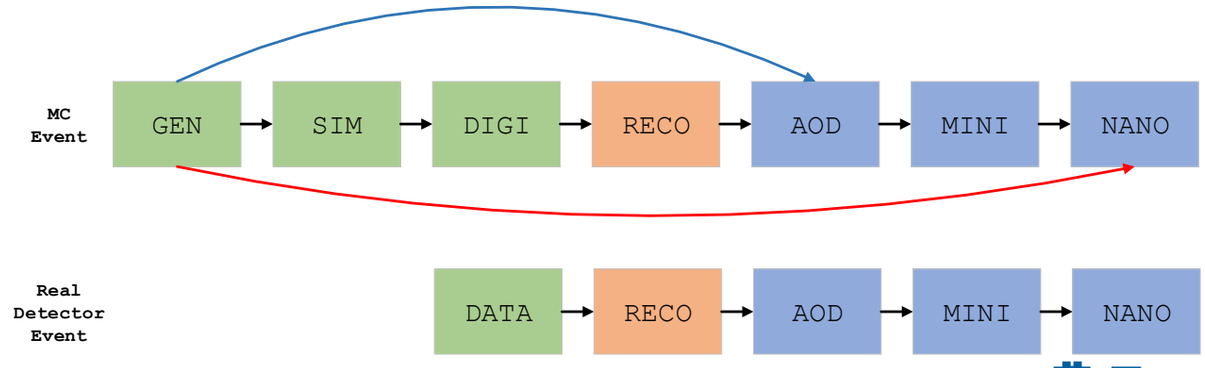
(b) A pre-training and fine-tuning MLPF approach



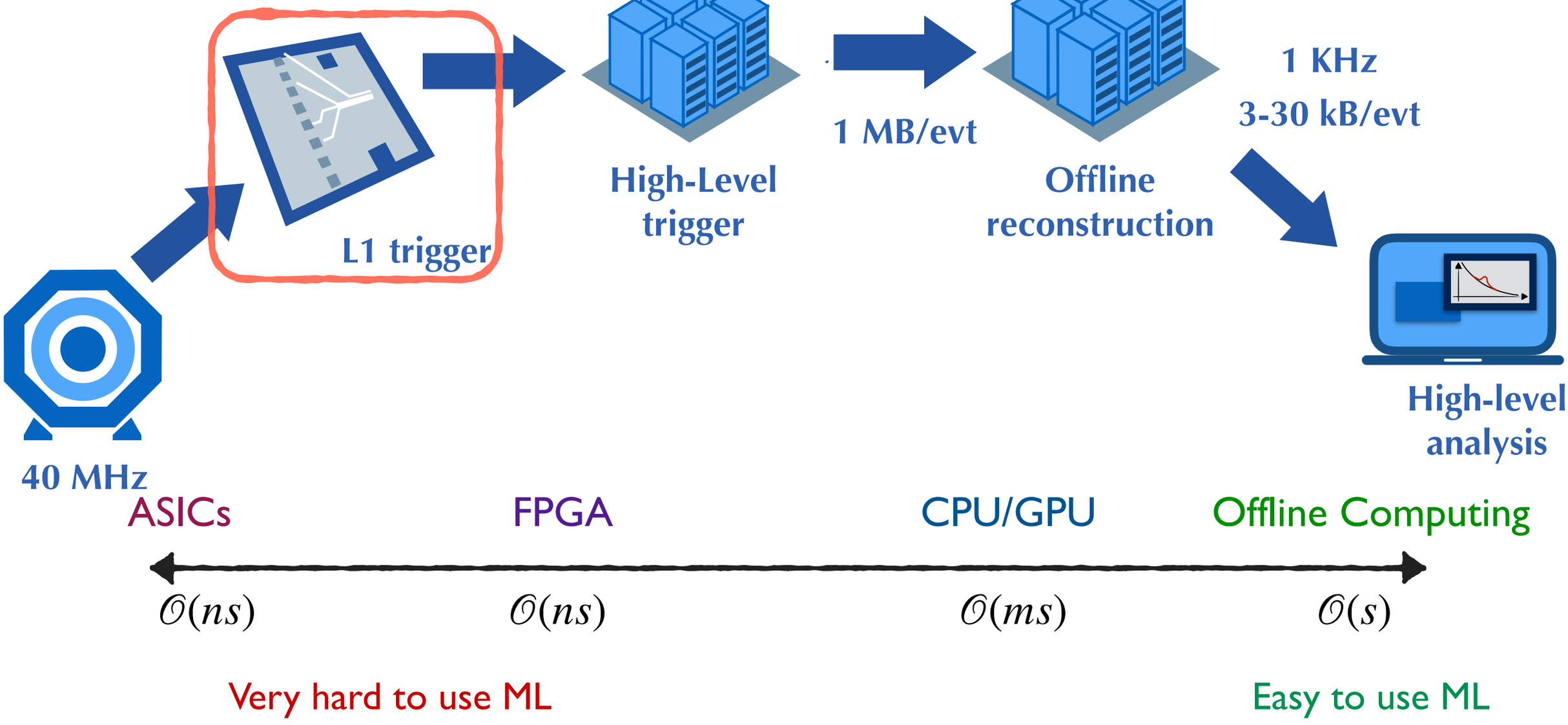
ML @ CMS: simulation

- Fully detailed simulation with generator+GEANT4 is computing intensive
- Generative models can though provide 1000x speed up
- Applicable at many levels: sampling, generation, detector model, analysis variables, etc.
- Careful study of statistical power of learned models over training samples
- A lot of R&D with examples in CMS: simulation of jets, photon and hadron showers in calorimeters
- A few ongoing use cases:
  - Denoising with convolution neural networks
  - FlashSim: end-to-end simulation using normalizing flows

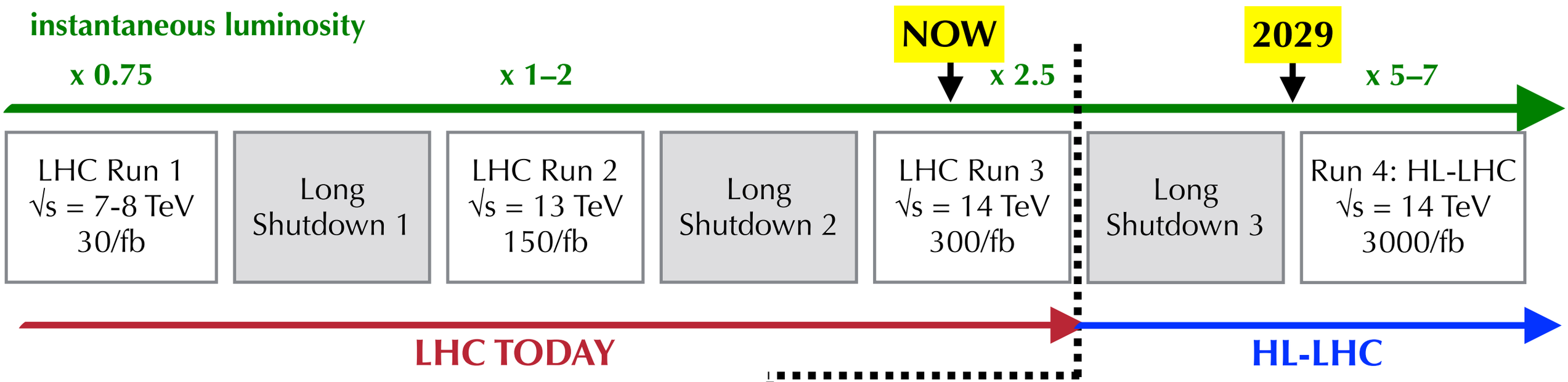


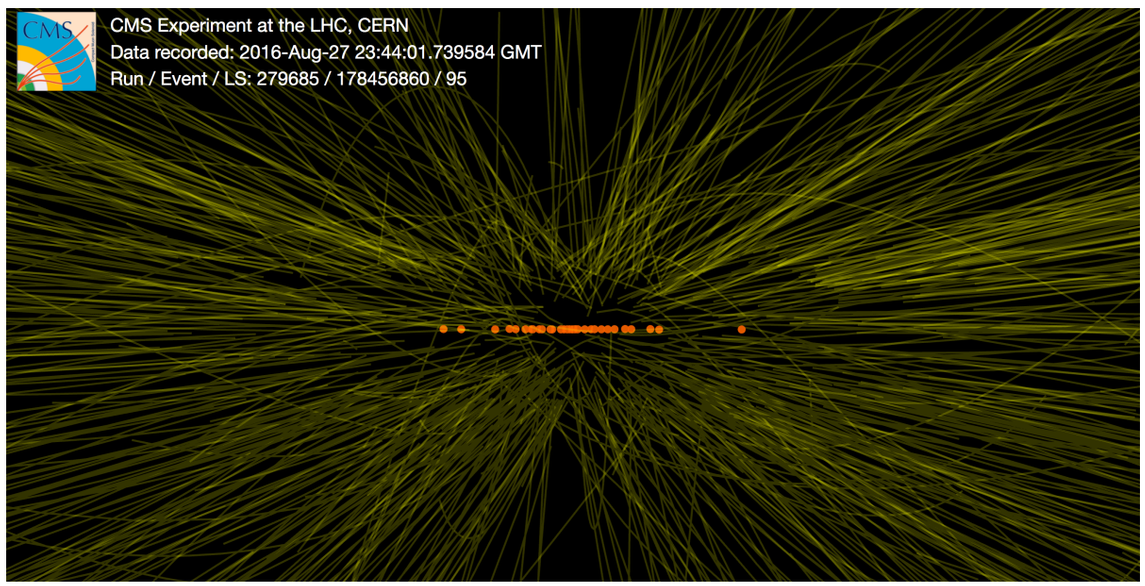


Data reduction workflow @ LHC

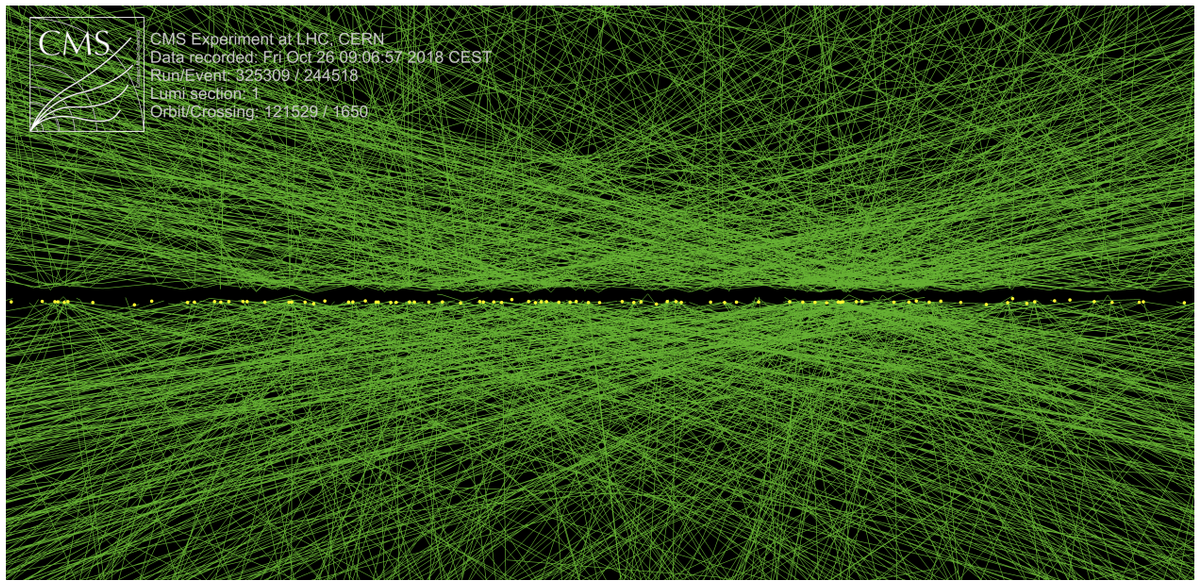


The High-Luminosity LHC challenge





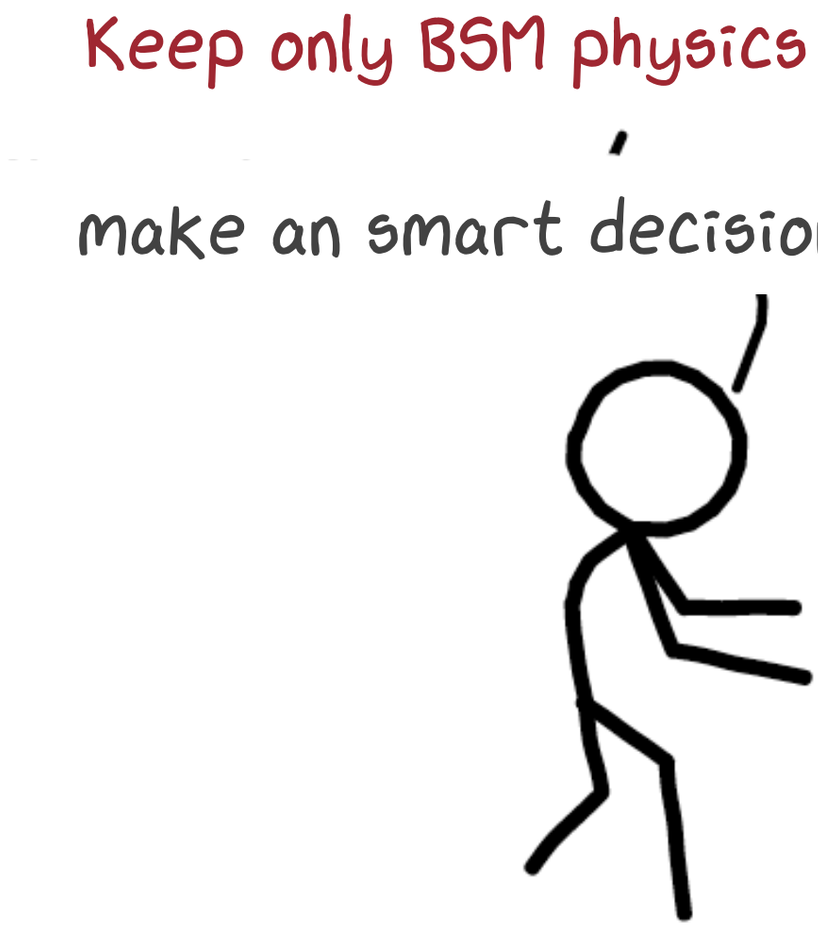
40 simultaneous collisions per bunch crossing

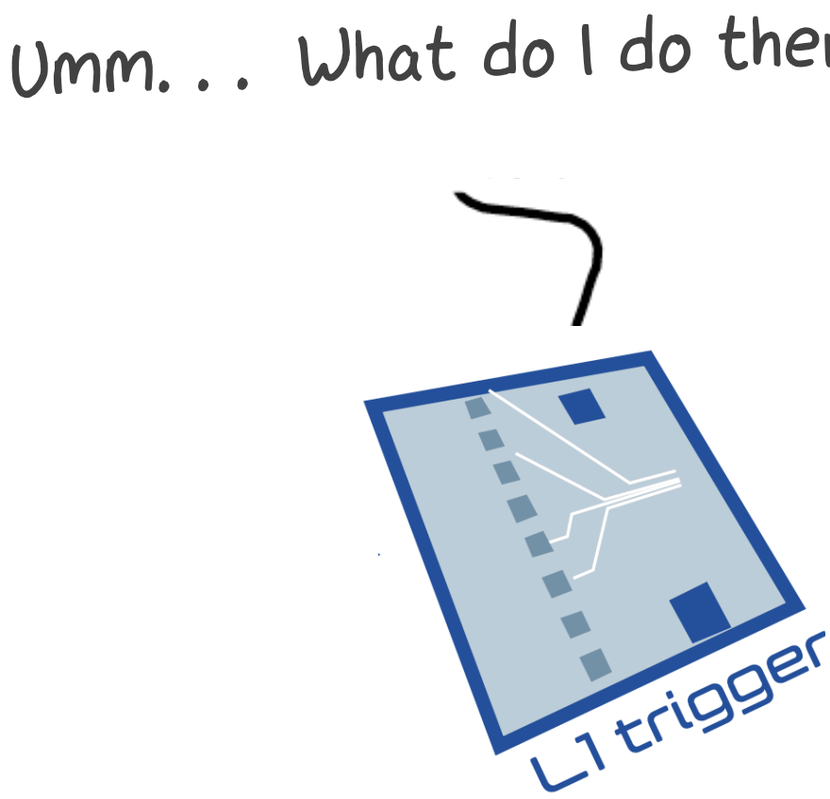


200 simultaneous collisions per bunch crossing + more granular detector!

Intelligent Triggering

- Intelligent selection is most impact full at L1 trigger





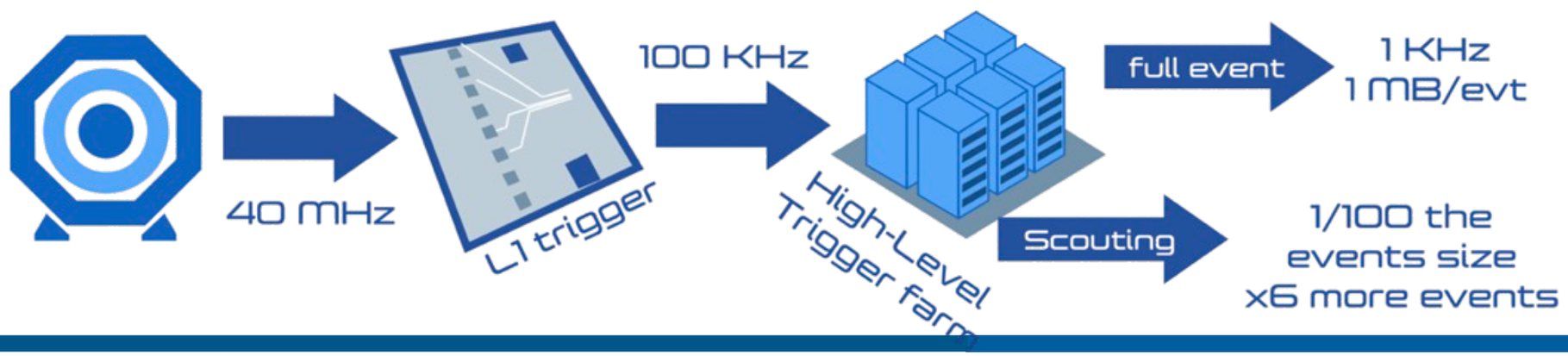
L1 Trigger in CMS

- CMS L1 trigger processes every collision
  - Need quick decisions w/ large data rate
    - L1 runs on FPGAs
      (w/ $3.8\mu s$ latency &
      selects 1 in 400 pp collisions)

Retain much of this

(Needs to be unbiased)

Discard a lot of this

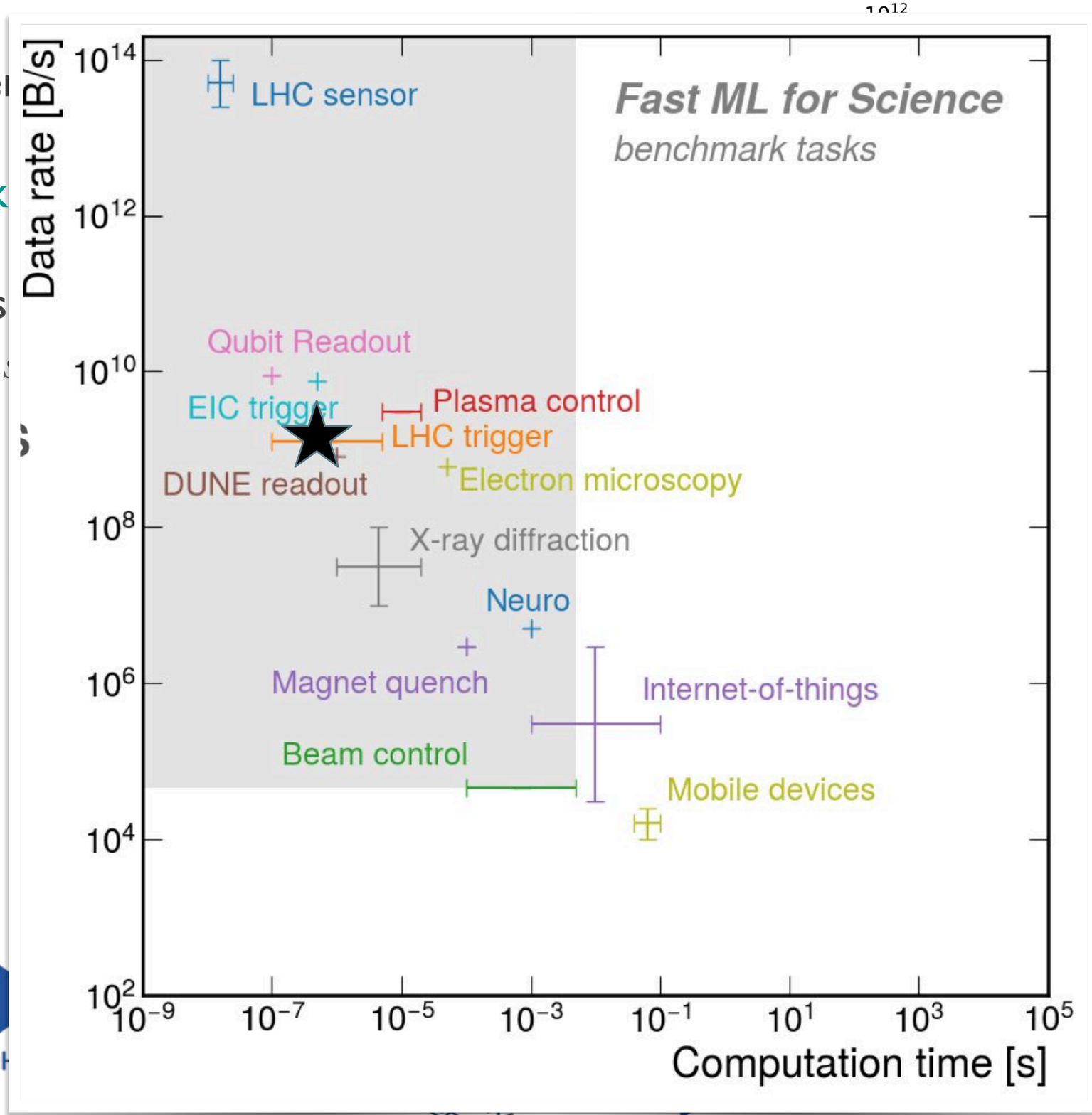


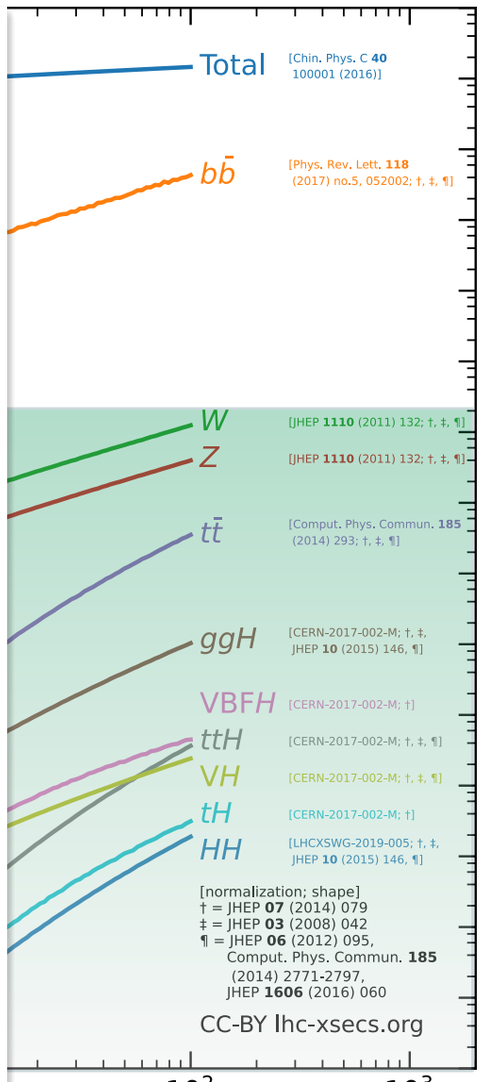
99.75% gets thrown away in first stage!

Credits: Noah P.

L1 Trigger in CMS

- CMS L1 trigge
  - Need quick
    - L1 runs
      (w/ $3.8\mu s$
      selects 1





er-of-mass $\sqrt{s}$ / TeV

9.75% gets thrown
away in first stage!

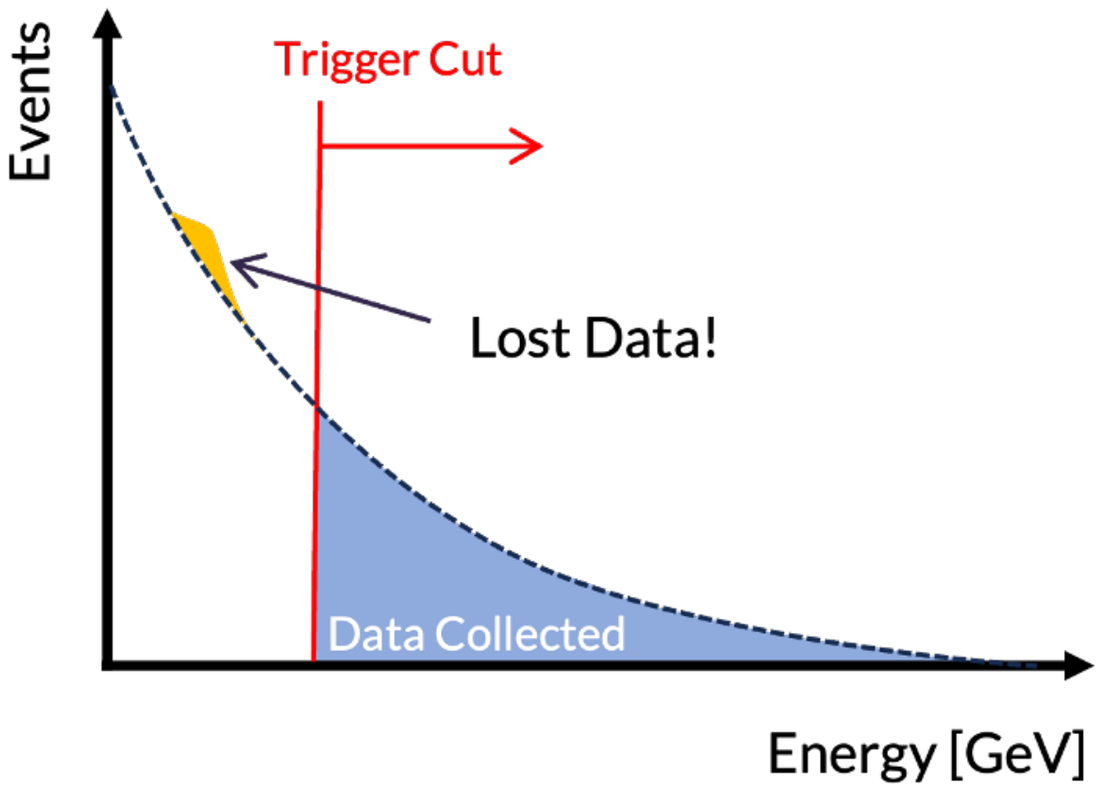
40 MH

x6 more events

Fermilab

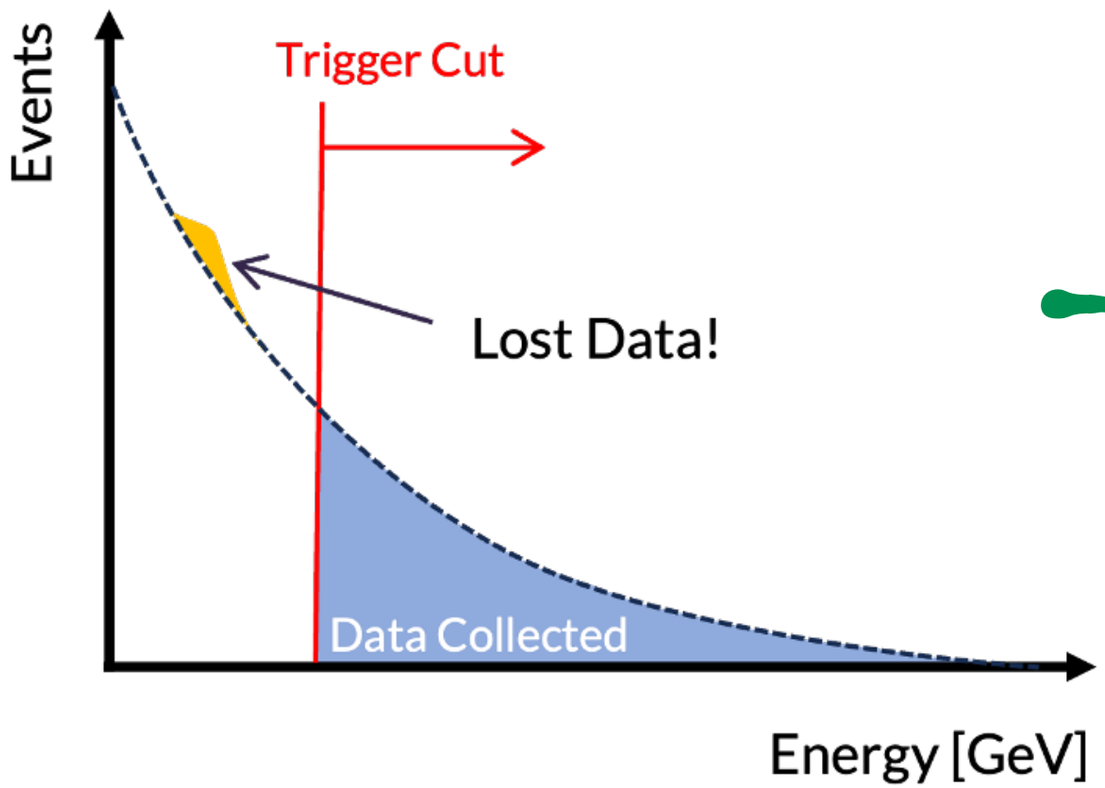
A new kind of data stream

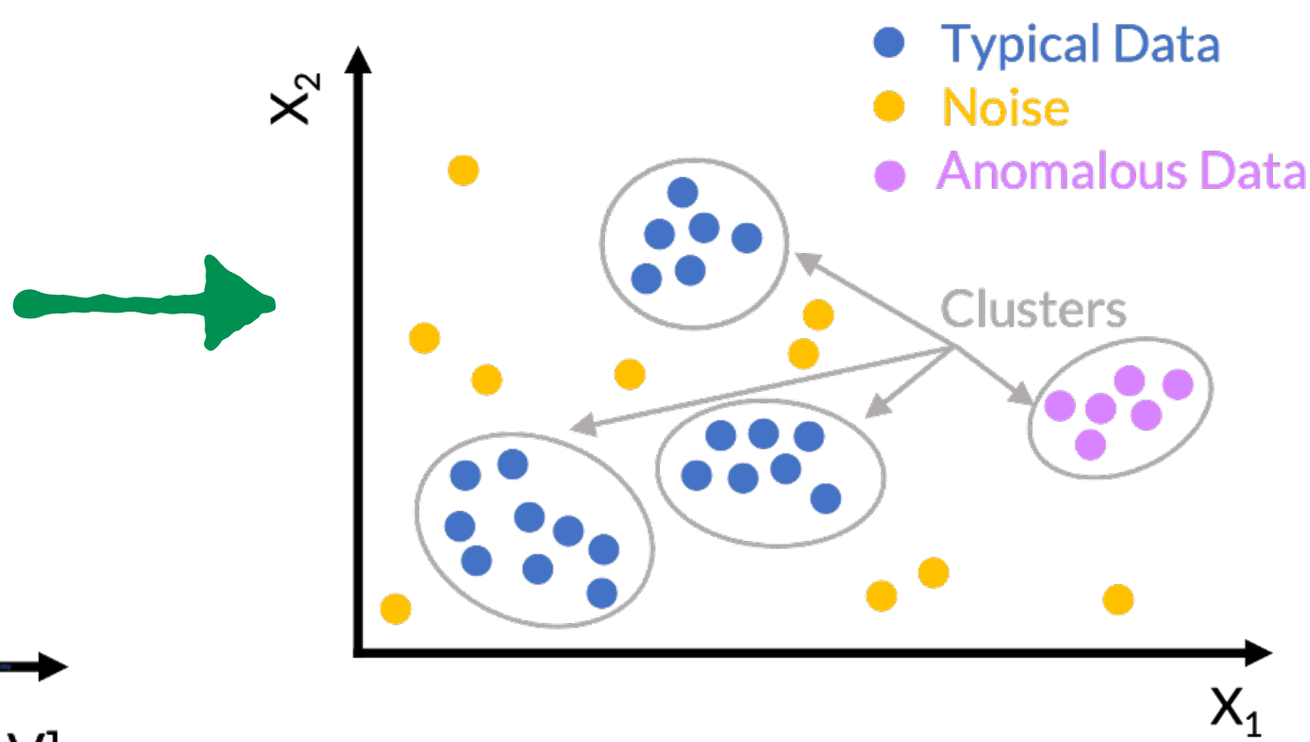
- Simple kinematic selections to define trigger algorithms
  - What of we miss it? We might not know what we are looking for ….



L1 Anomaly Detection

- New Approach: Anomaly Detection
  - What if we select "anomalous" events based on decay topology
  - With Machine Learning @ L1 trigger



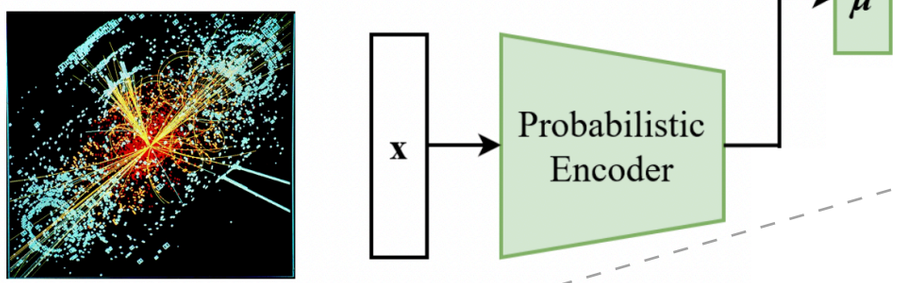


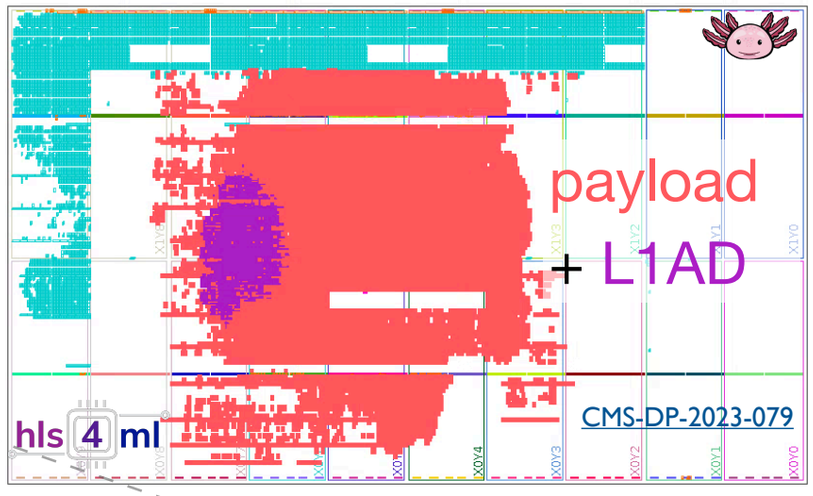
Enter AXOL1TL

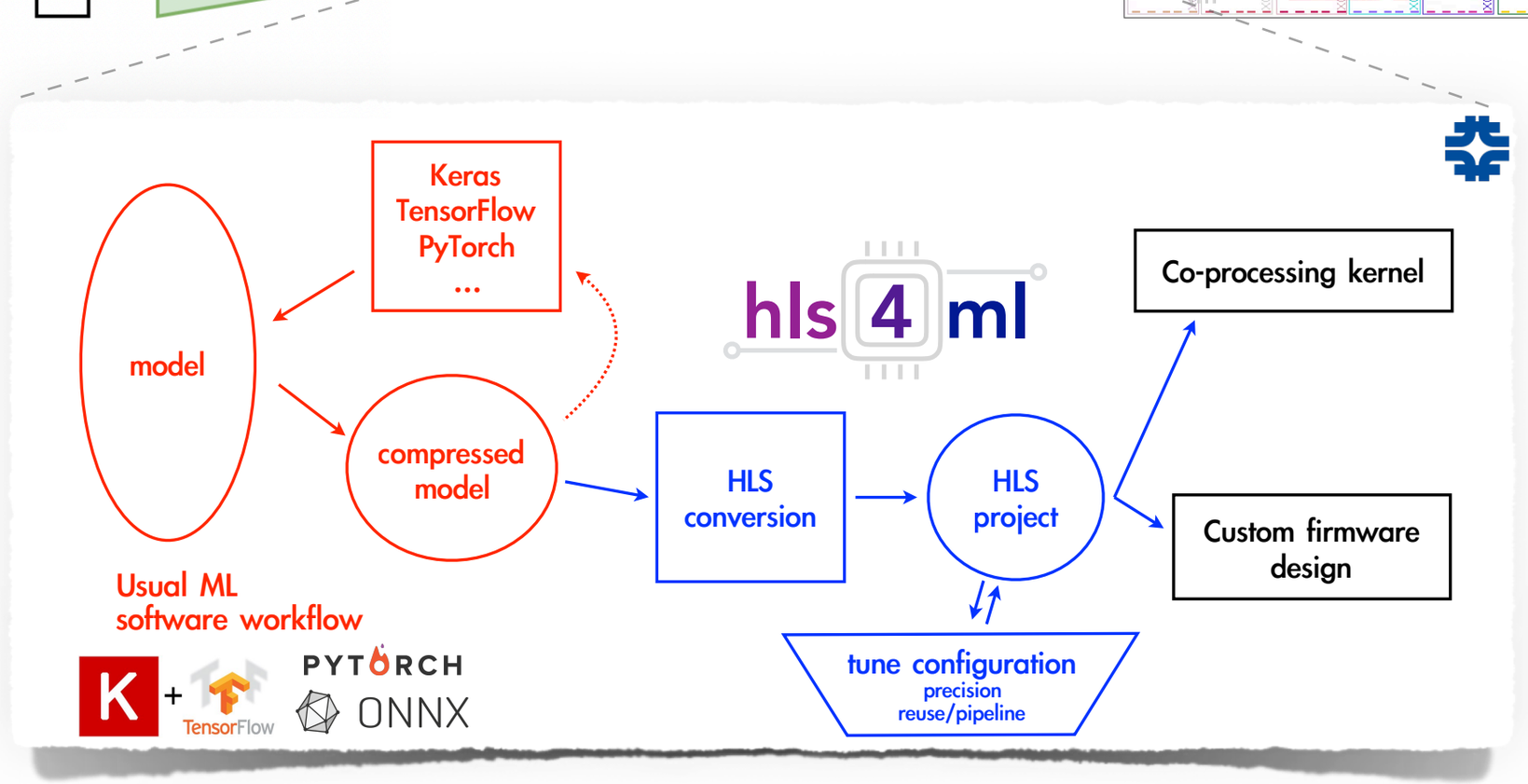
Anomaly eXtraction Online L1 Trigger aLgorithm

Fitting AXOL1TL on FPGA

- Instead of vanilla autoencoder, we use Variational Autoencoder
  - Use just the encoder part of the network for **AXOL1TL**
  - With hls4ml to deploy it on a FPGA



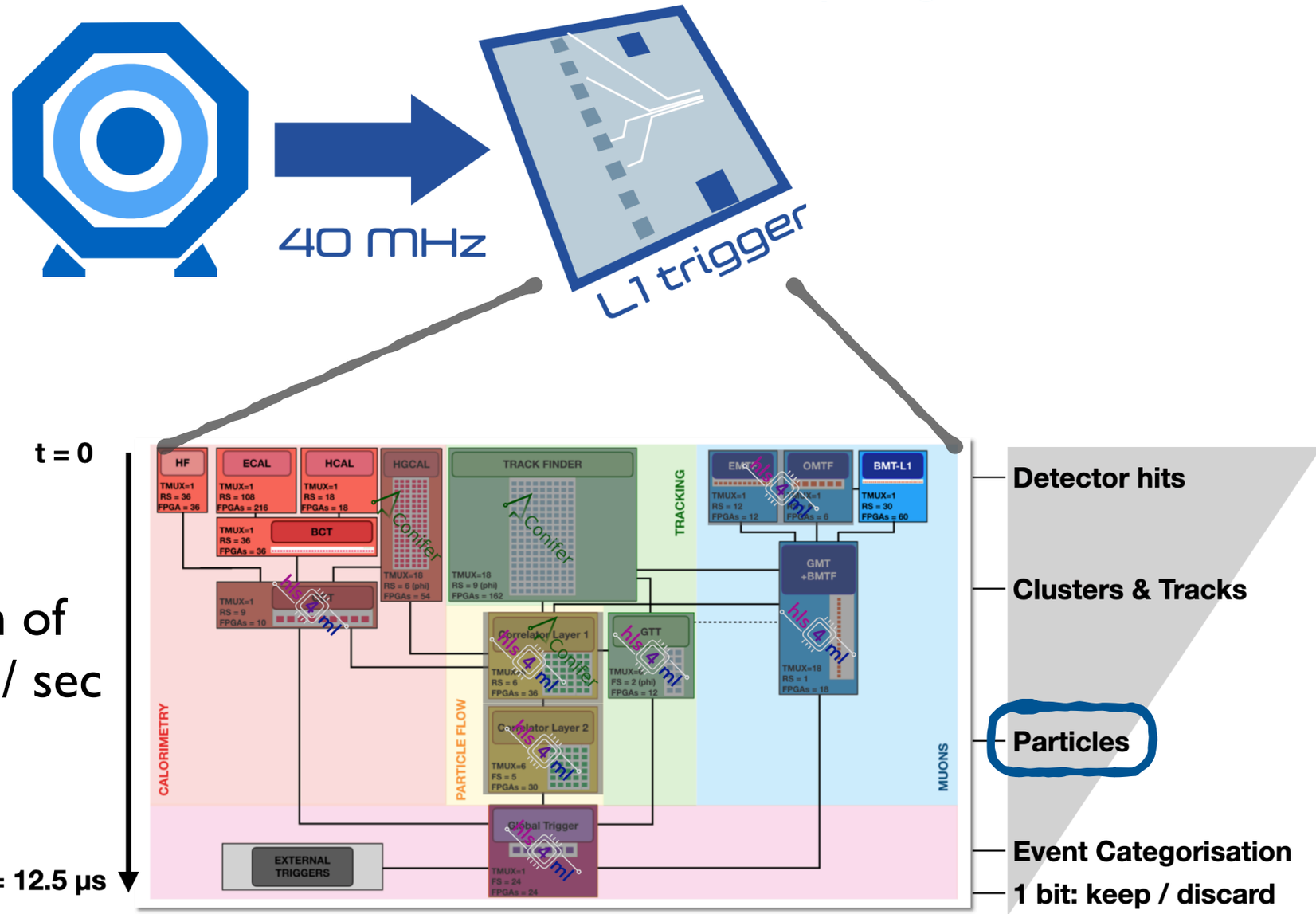




https://fastmachinelearning.org/hls4ml/

CMS @ HL-LHC

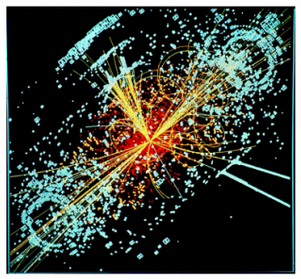
- CMS L1 Trigger @ HL-LHC is super charged with latest FPGAs + Algorithms
  - Can reconstruct as well as current High-Level Trigger @ 40 MHz
    - Even reconstruct individual particles for every single collision!

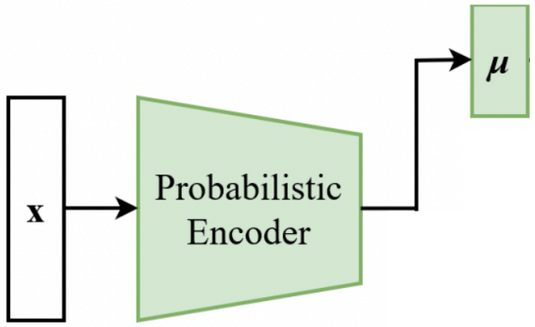


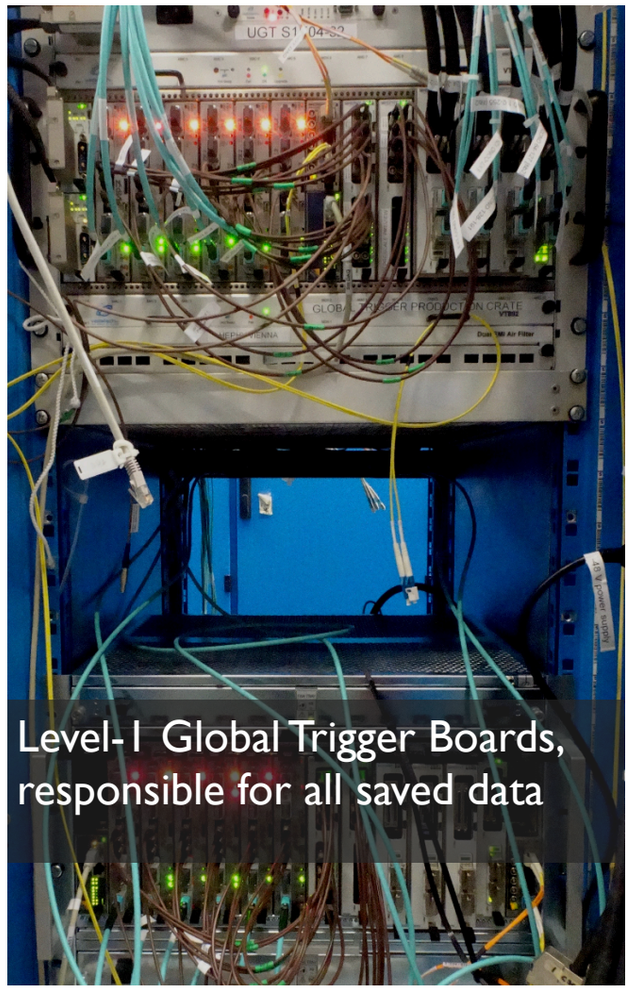
Credits: Sioni Summers

Fitting AXOL1TL on FPGA

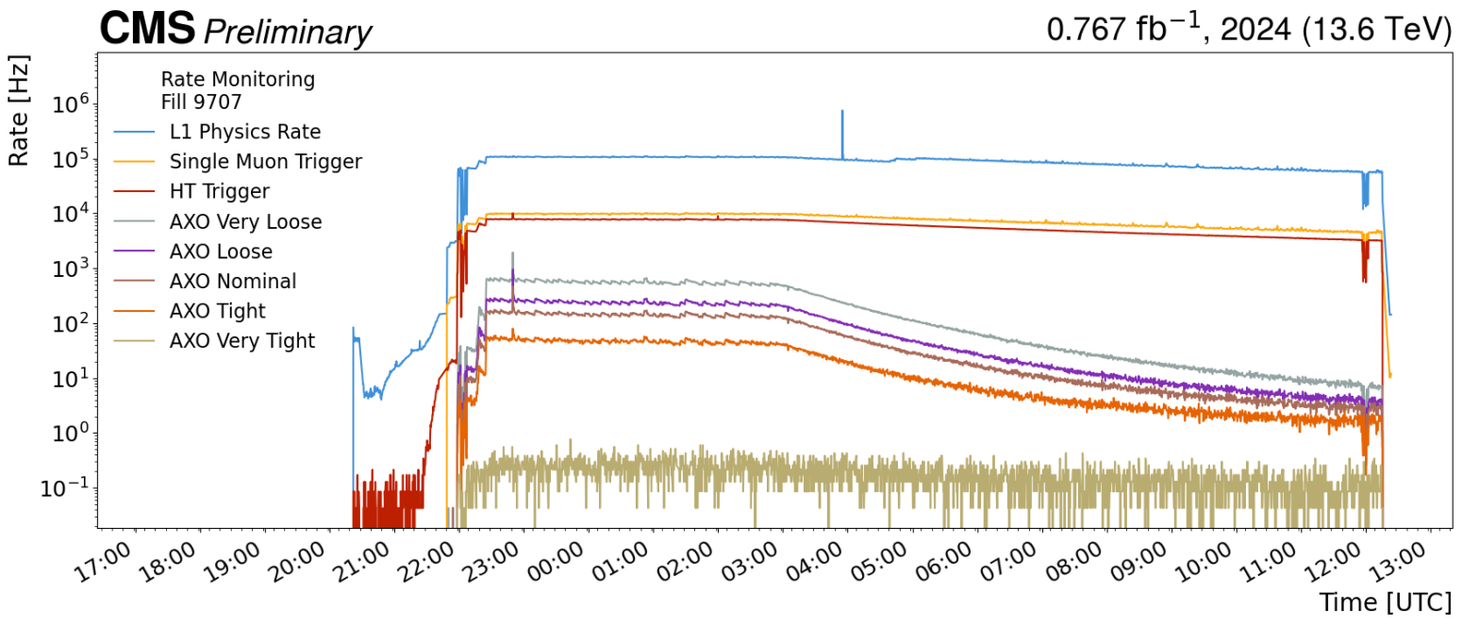
- Instead of vanilla autoencoder, we use Variational Autoencoder
  - With hls4ml to deploy it on a FPGA
  - *Inference in just 50 nano seconds!*







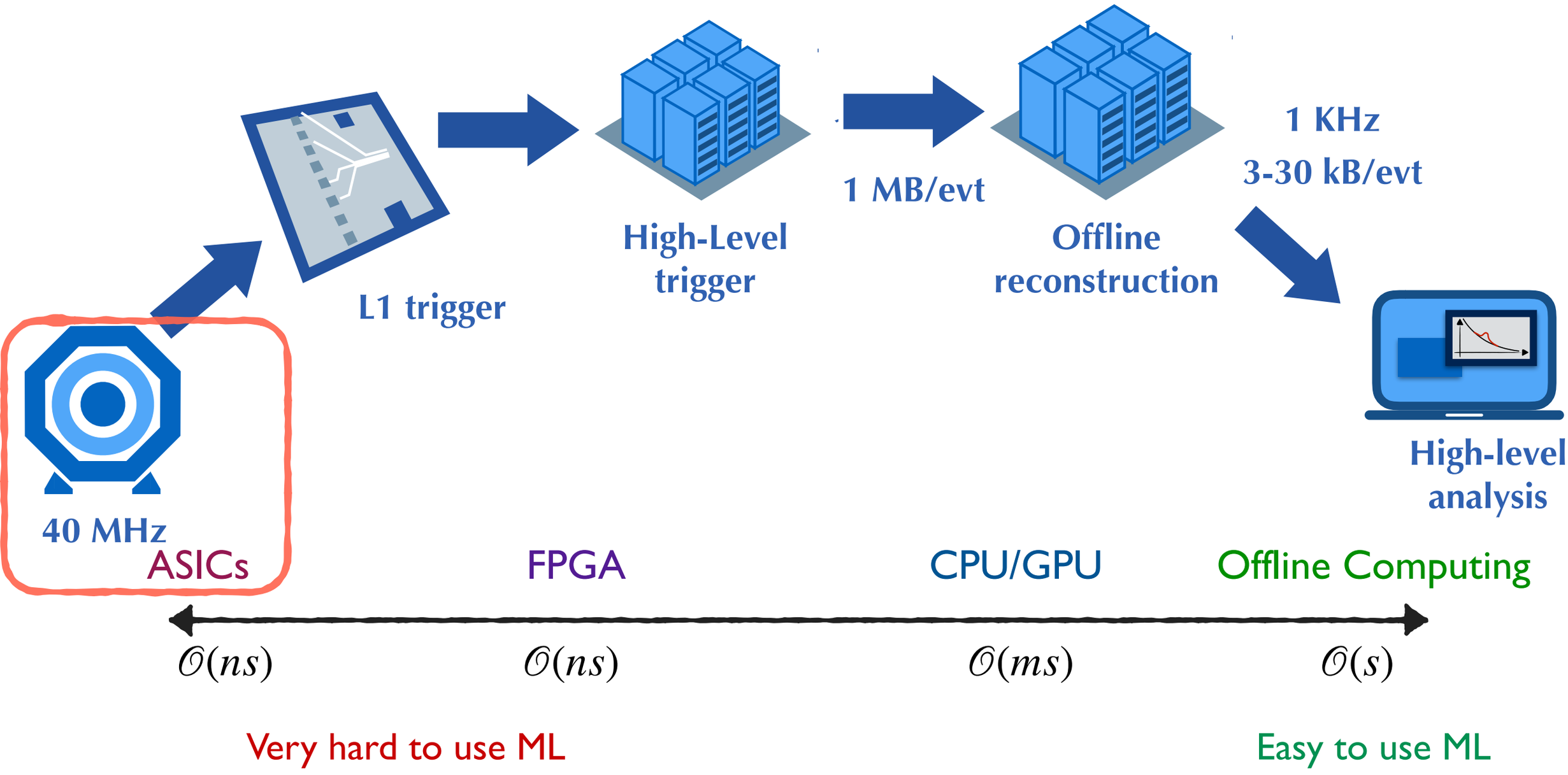
- **Collecting data since 2024 @ LHC !**



CMS-DP-2023-079

Fermilab

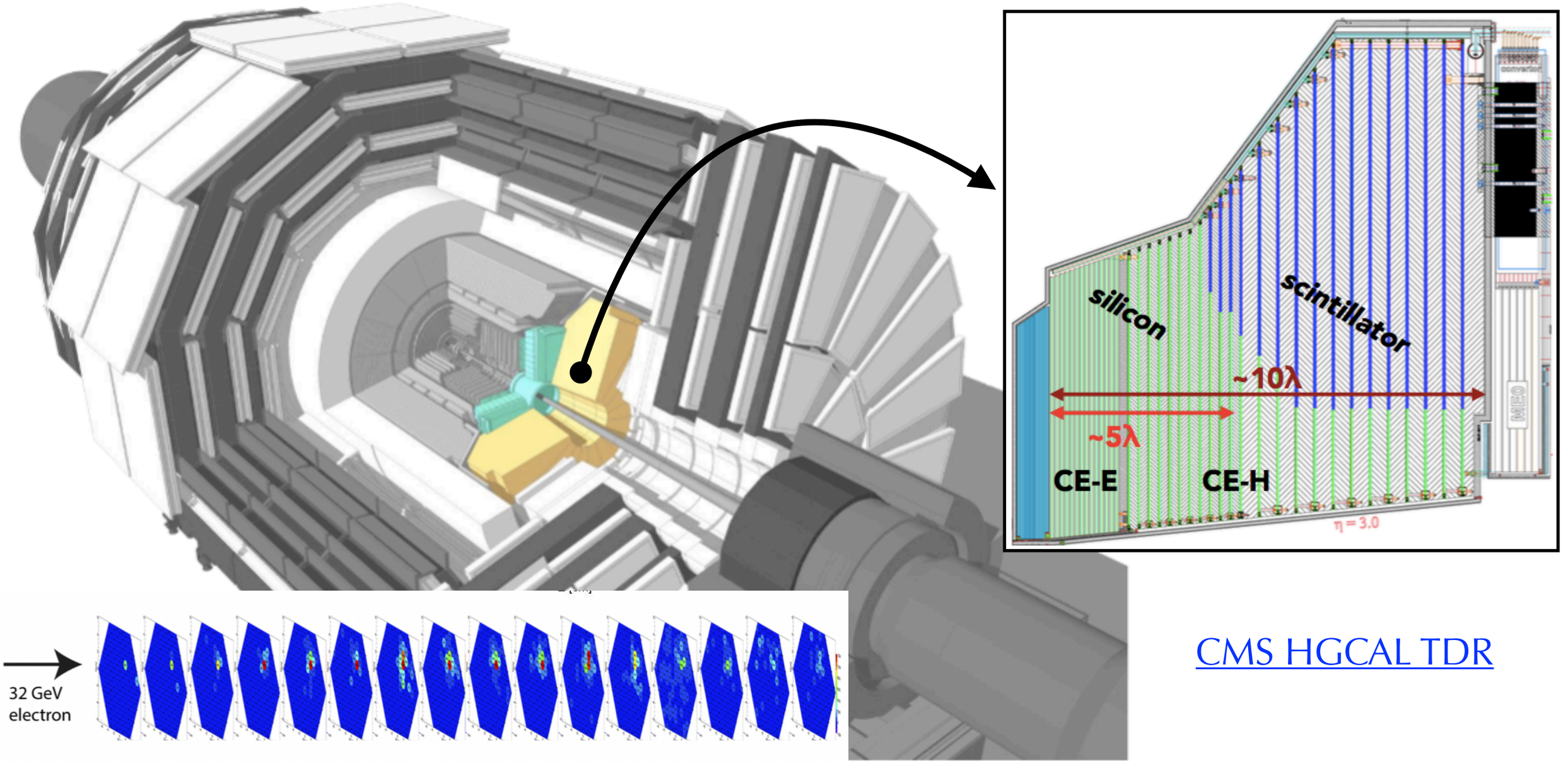
Data reduction workflow @ LHC



ML @ CMS: front-end electronics

The CMS High-Granularity Calorimeter

Novel technology for CMS endcap calorimeter:
50 layers with unprecedented number of readout channels (6M)!



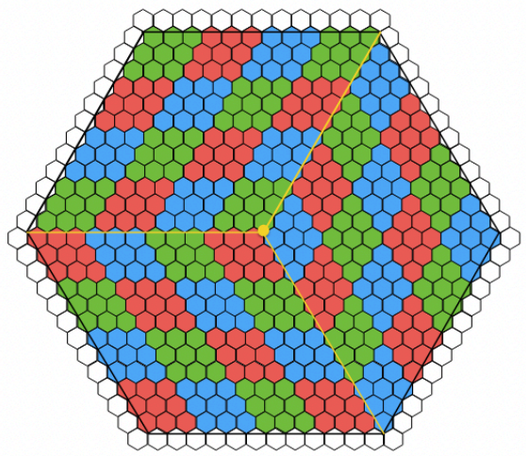
[CMS HGCAL TDR]

arXiv:2105.01683, arXiv:2306.04712

ML @ CMS: front-end electronics

Input

HGCAL 8" hex module



432 silicon sensors ➔ 48 trigger cells (TC) @ 7b per TC

(336b in total)

ASIC

Output:

"Super trigger cell" algo

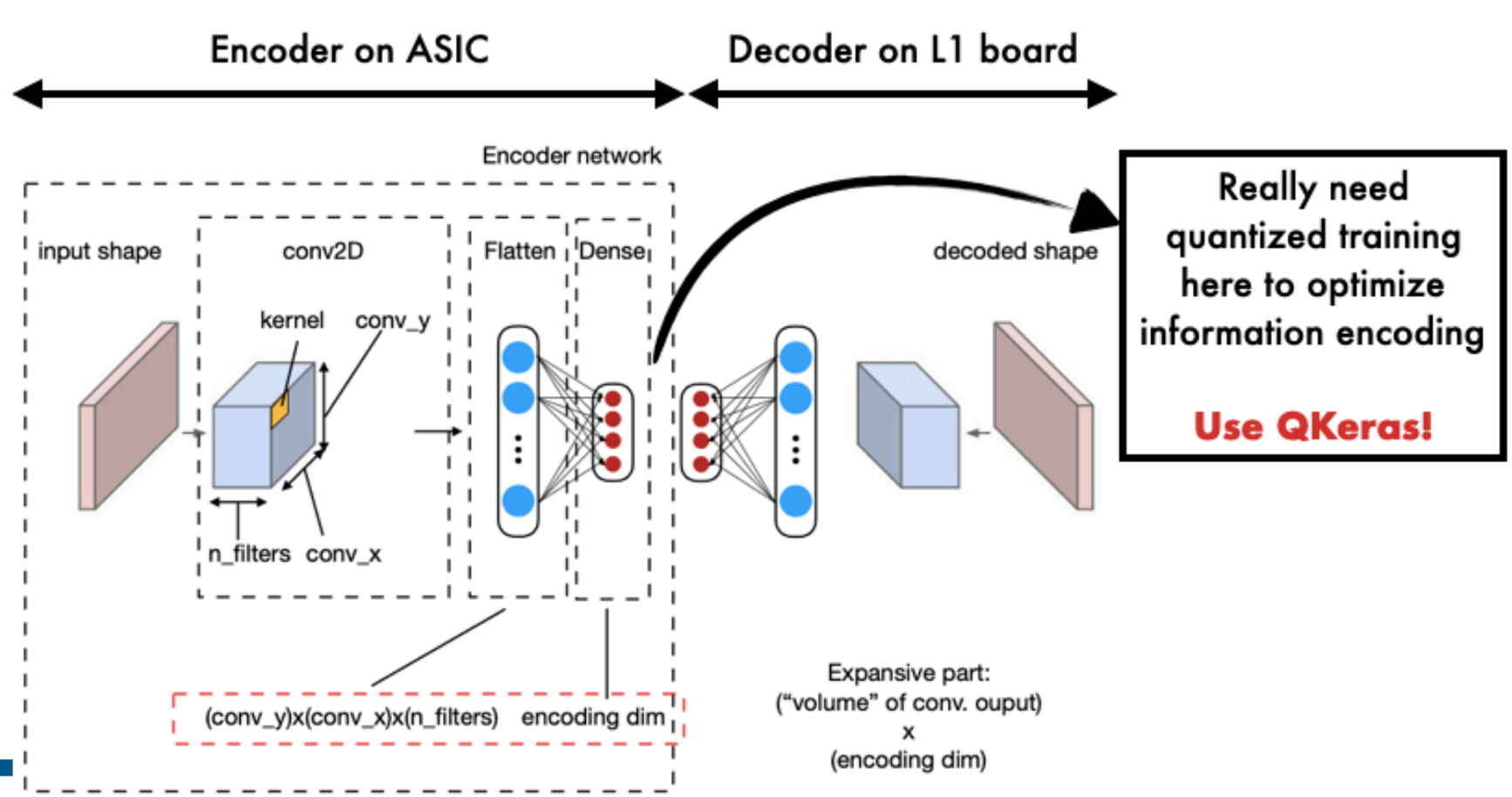
3[16 TC sum] x 16–48 bits

= 48–144 bits

(depending on the position)

Can we do a better job of encoding the info in those bits w/o so much loss in granularity?

hls4ml used to obtain ASIC ML design!

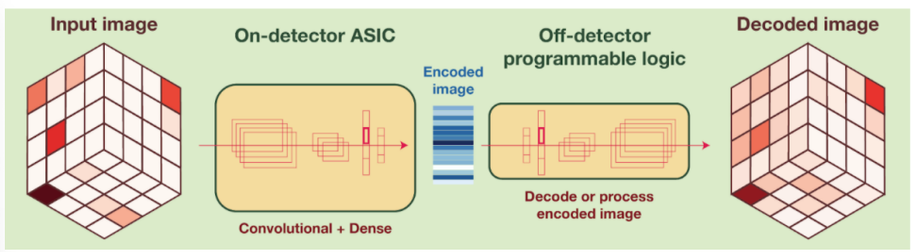


Really need quantized training here to optimize information encoding

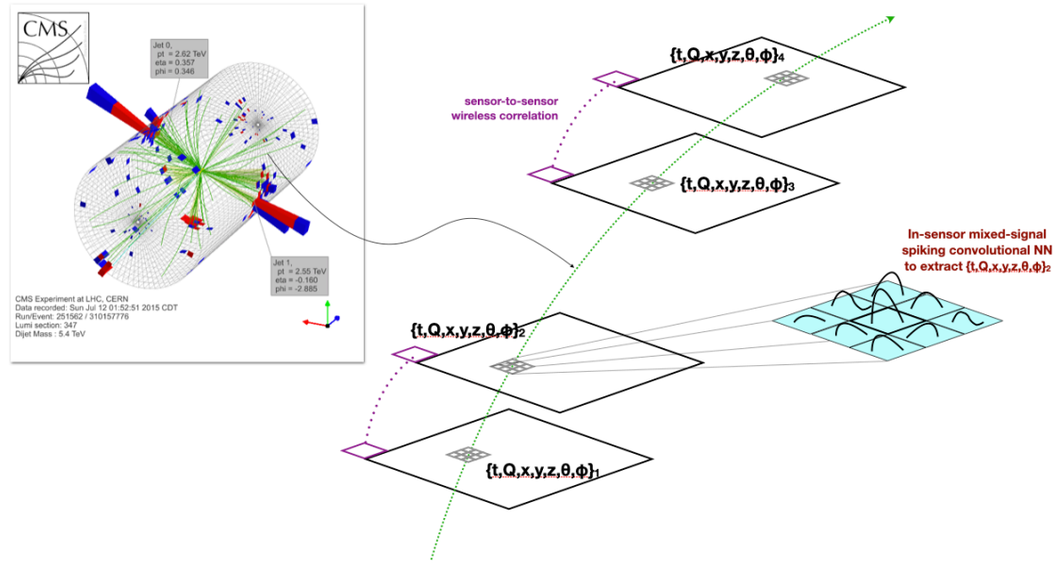
Use QKeras!

AI @ Extreme Edge

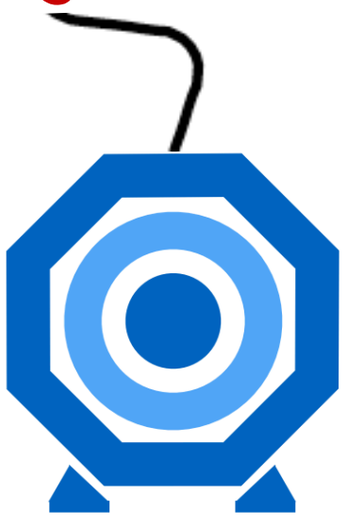
Data compression w/ Radiation hard ASICs



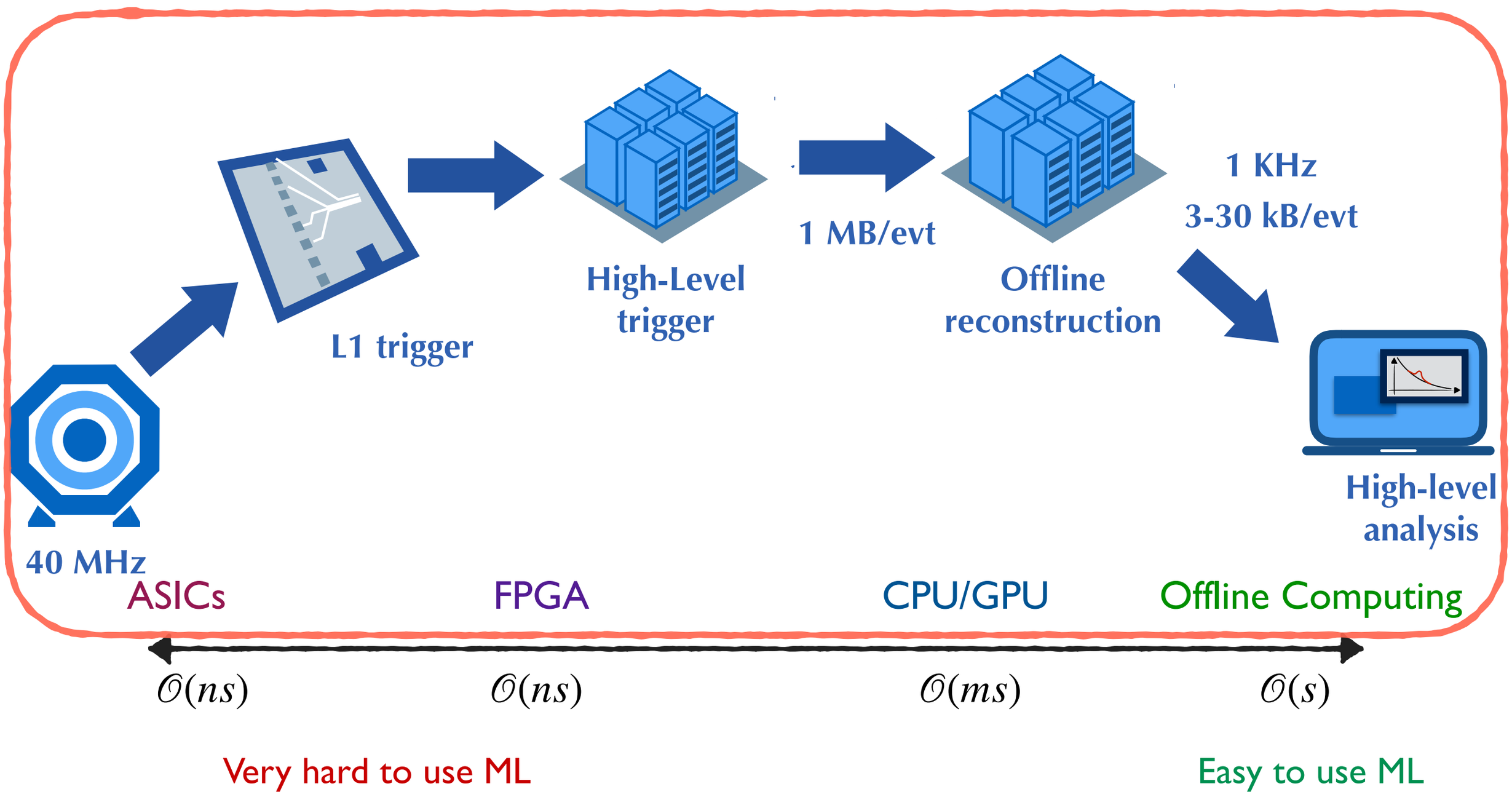
Smart pixels: Pixel sensors w/ AI on chip



Smart sensing w/ On chip AI

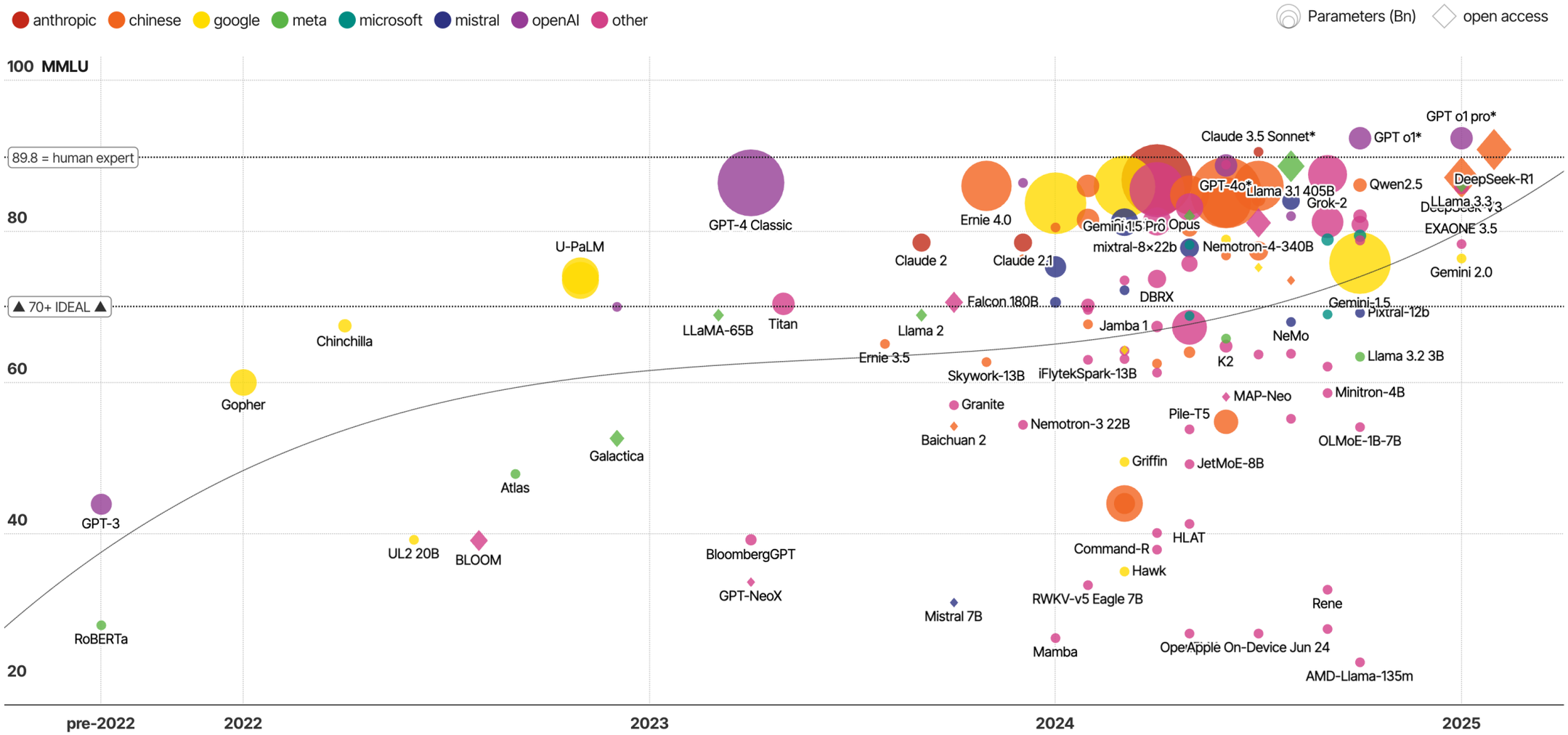


Data reduction workflow @ LHC



The Rise of Large Language Models

- Billion-parameter models pre-trained on massive unlabeled datasets can result in emergent capabilities surpassing human expertise in some areas

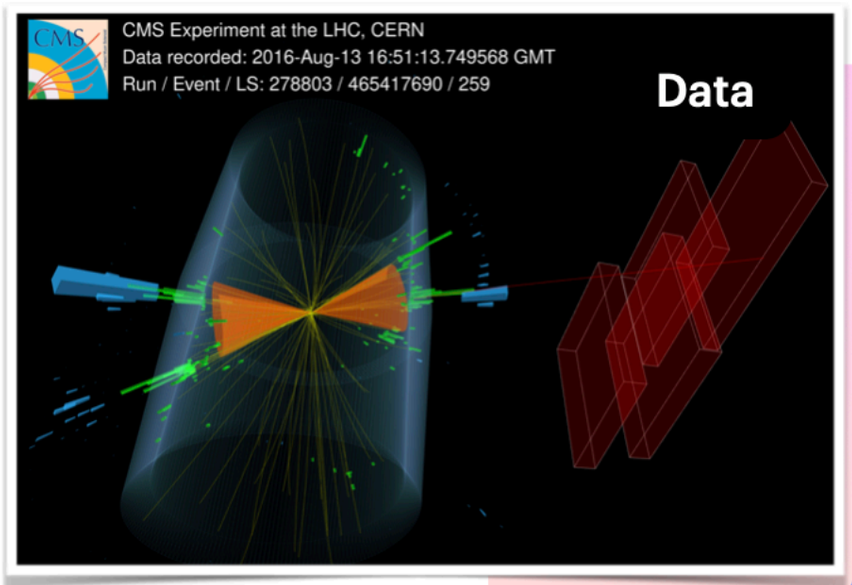


David McCandless, Tom Evans, Paul Barton
Informationisbeautiful // Jan 2024

https://informationisbeautiful.net/visualizations/the-rise-of-generative-ai-large-language-models-llms-like-chatgpt/

MMLU = benchmark for measuring LLM capabilities
* = parameters undisclosed // source: LifeArchitect // data

A Hep Foundation Model?



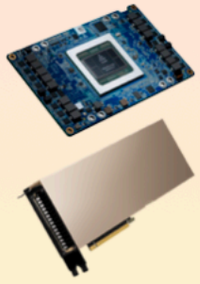
Tracks

EM Clusters

Had. Clusters

Muon Tracks

Training



Adaptation

HEP Foundation Model

Tasks

Jet Tagging

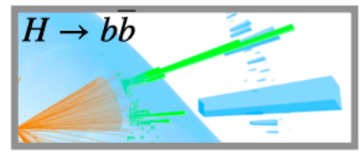
Jet Energy Correction

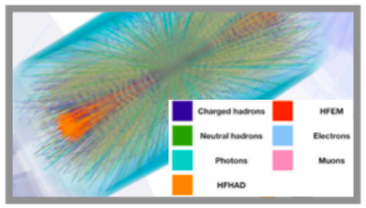
Particle-Flow Reconstruction

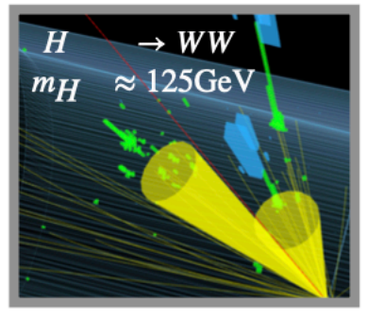
Pileup Mitigation

Event Analysis

Anomaly Detection







Fermilab

ML Group

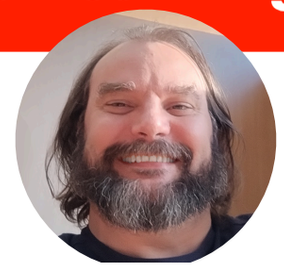
- State-of-the art machine learning methods are increasingly being deployed in CMS
- Many opportunities to integrate/develop proof of concepts methods into the experiment with many challenges ahead to bring the most promising into production
- CMS ML group was created to supervise this effort: cms-conveners-ml@cern.ch

Our goal: enable, support, guide and foster ML developments in computing, POGs, PAGs

Forum (bi-weekly):
https://indico.cern.ch/category/12412/
L3 subgroup meeting (weekly, rotating):
https://indico.cern.ch/category/12413/

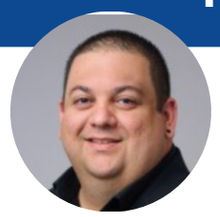
L2 Group: Machine Learning


Mia


Pietro

L3 topical groups

Knowledge


Daniel


Marius

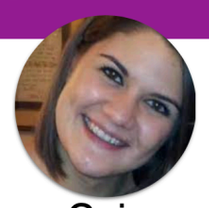
Production


Hyejin


Yao

Innovation

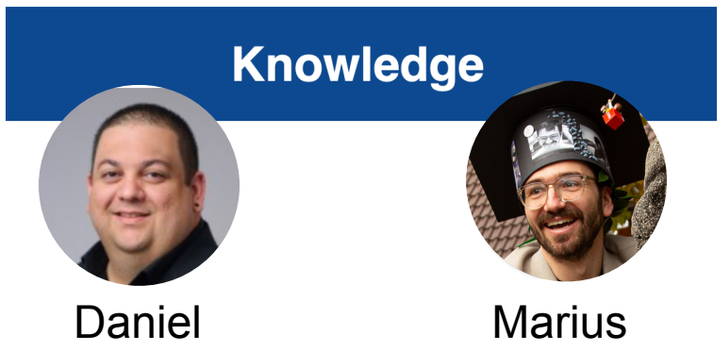

Abhijith


Gaia

+ contacts to POG/PAGs and external initiatives

ML Subgroup: Knowledge

- Knowledge subgroup **collects, maintains, and disseminates knowledge** of ML algorithms in the CMS collaboration
  - Development and maintenance of CMS ML **benchmarks**, comparing and tracking the performance of algorithms, platforms, and ML frameworks on a set of benchmark CMS ML applications in reconstruction, simulation, trigger and computing
  - Consolidate **tutorials** and **lectures** on ML for CMS with the School Committee
  - Maintain list of **in-house experts** in various ML topics.
  - Prepares the ML part of the analysis questionnaire in collaboration with the Statistics Committee and documents good ML practices
- **Documentation** webpage: https://cms-ml.github.io/documentation/
  - Contributions welcome: https://github.com/cms-ml/documentation/pulls

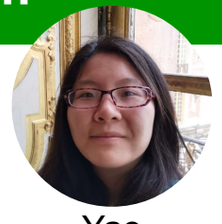


ML Subgroup: Production

- Production subgroup delivers production-level training and inference for CMS ML algorithms
  - Development and maintenance of **ML training and inference workflows** in the CMS software stack, including model inference and support for external frameworks such as TensorFlow, PyTorch, ONNX, and MXNet
  - Work closely with CMS framework experts and all relevant O&C groups, for example, overseeing framework-related aspects and relevant software/computing groups
  - Development of training infrastructure to satisfy the needs of as many collaborators as possible

Production

Hyejin



Yao

ML Frameworks

- **Keras/TensorFlow**
  - Conventionally one of the most beginner-friendly formats
  - Inference supported for production in CMSSW
- **PyTorch (Lightning)**
  - More "pythonic" way of building models (especially models within models)
  - Inference support in CMSSW forthcoming…
- **ONNX/ONNXRuntime**
  - ONNX intended to be a universal exchange format (can convert from all libraries)
  - For inference only; supported for CMSSW
- **Flax/JAX**
  - Newest library based on "autograd"
- **XGBoost**
  - For training boosted decision trees (fast)

ML Subgroup: Innovation

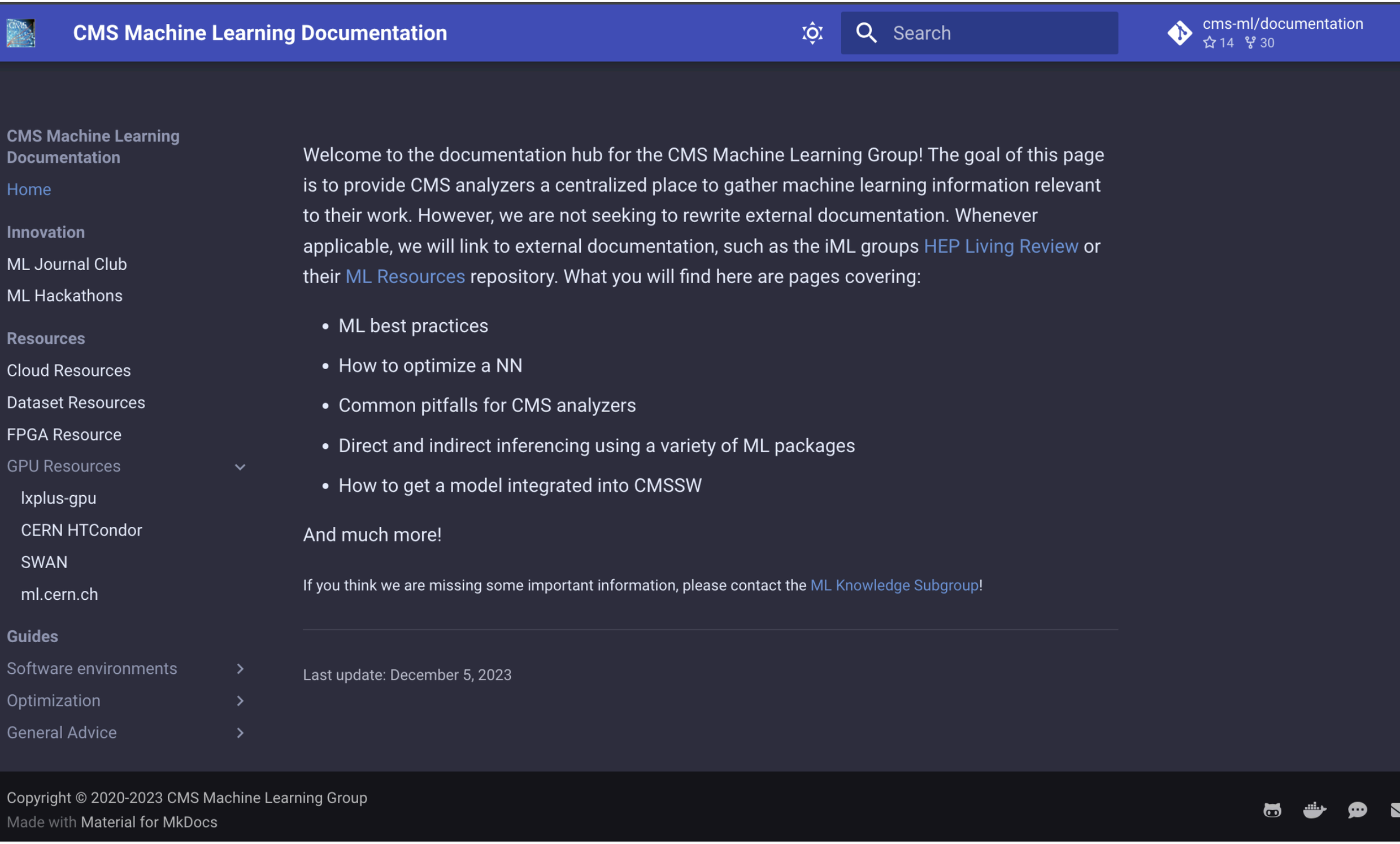
- Innovation subgroup identifies and applies new ML techniques to CMS challenges while working closely with CMS groups in areas where ML is expected to have a significant impact: reconstruction, trigger, simulation, and beyond
  - Discuss the relevance of new techniques and help with the adaptation and implementation of specific models
  - Develop specific methods for CMS that will lead to technical publications
  - Lead the organization of ML-oriented **hackathons** and **challenges** to help identify new applications and ML techniques in CMS
    - e.g. Upcoming hackathons in LLMs in CMS
  - Organize CMS internal **journal club** to discuss new and relevant ML results from inside and outside of particle physics

Innovation


Abhijith


Gaia

CMS ML Documentation



cms-ml.github.io/documentation

Summary

- Lots of R&D surrounding ML in CMS, which directly impact physics results
- Many activities coordinated by ML Group
  - We're here to help!
- Get involved!
  - Attend ML Forum (https://indico.cern.ch/category/12412/) and subgroup meetings (https://indico.cern.ch/category/12413/)
  - Contact conveners (cms-conveners-ml@cern.ch) or subgroup conveners for a slot to present your ML studies
  - Consider/discuss publishing your ML results in CMS
  - Release CMS datasets for ML studies through knowledge subgroup
  - Contribute to ML documentation (https://cms-ml.github.io/documentation/) by submitting pull requests or reporting issues
  - And more…