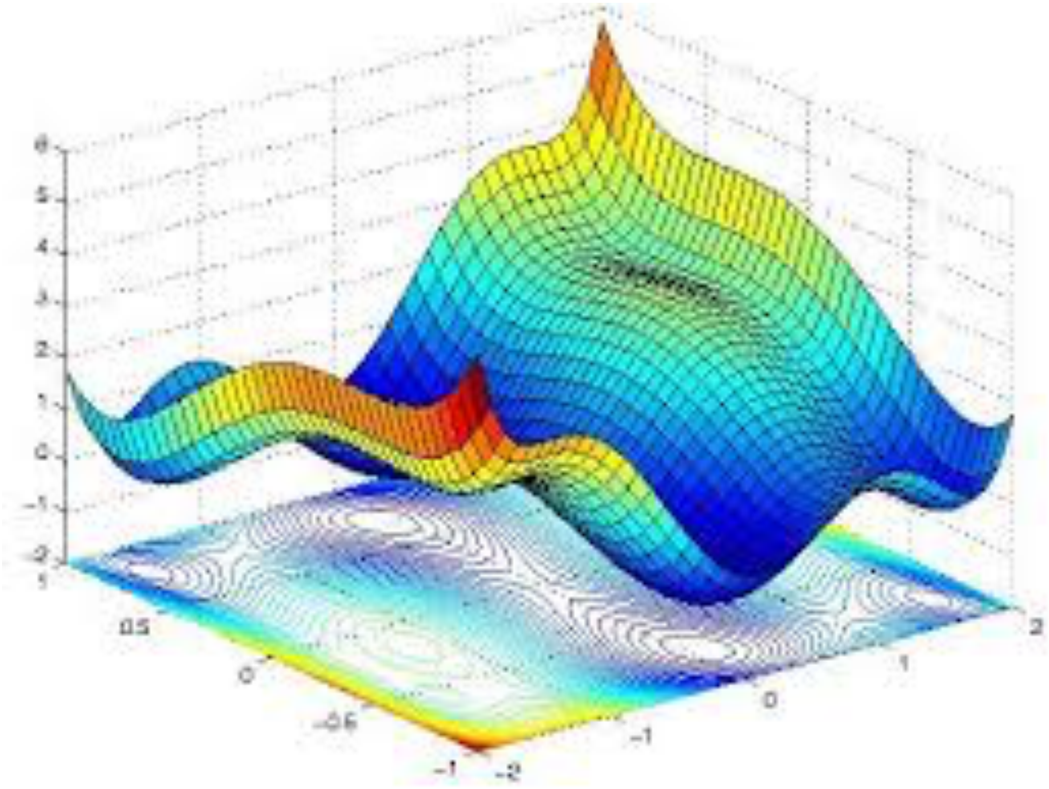
Machine Learning for future e^+e^- colliders

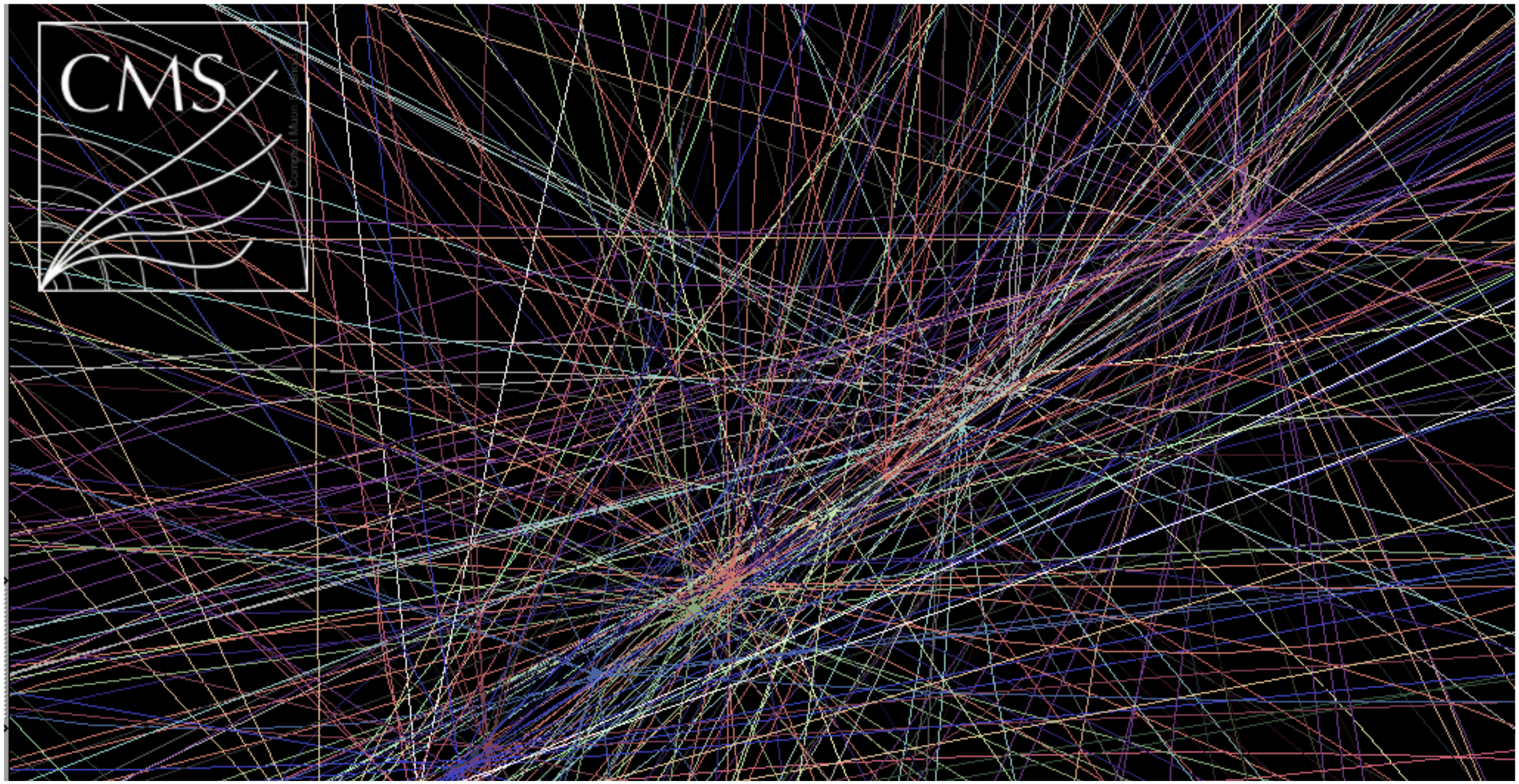
Maurizio Pierini

What is Machine Learning

- *Machine learning is a technique by which an algorithm in trained with example to accomplish a task*
  - *as opposed of being programmed to do so, by specified rules*
- *Deep Learning is the cutting-edge ML technology*
  - *based on "old-school" neural networks + augmented computational capabilities (e.g. GPUs) and "new" architectures (recurrent nets, convolutional nets, autoencoders, …)*
  - *the breakthrough is fast differentiability (back-propagation) allowing fast optimization*
  - *DL networks are good in learning non-linear functions: can be a fast shortcut to replace heavy processing tasks*



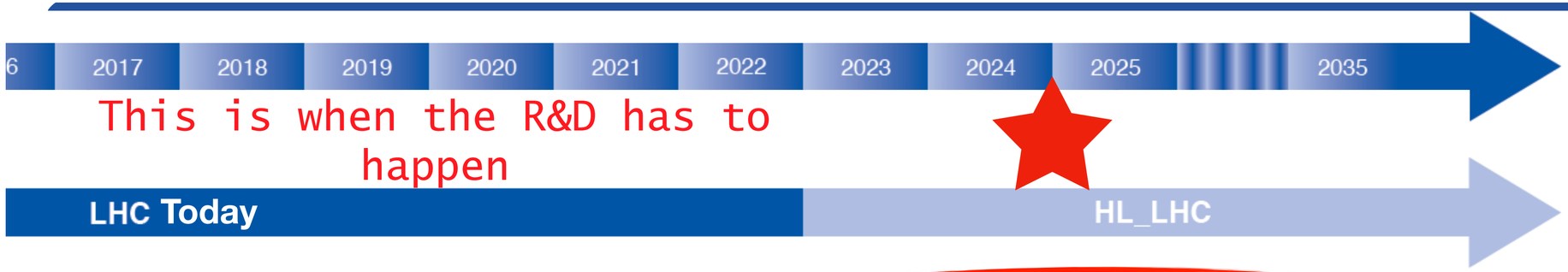
Training a Machine Learning algorithm consists in minimizing a complicated multi-dimensional function



Why ML for HEP?

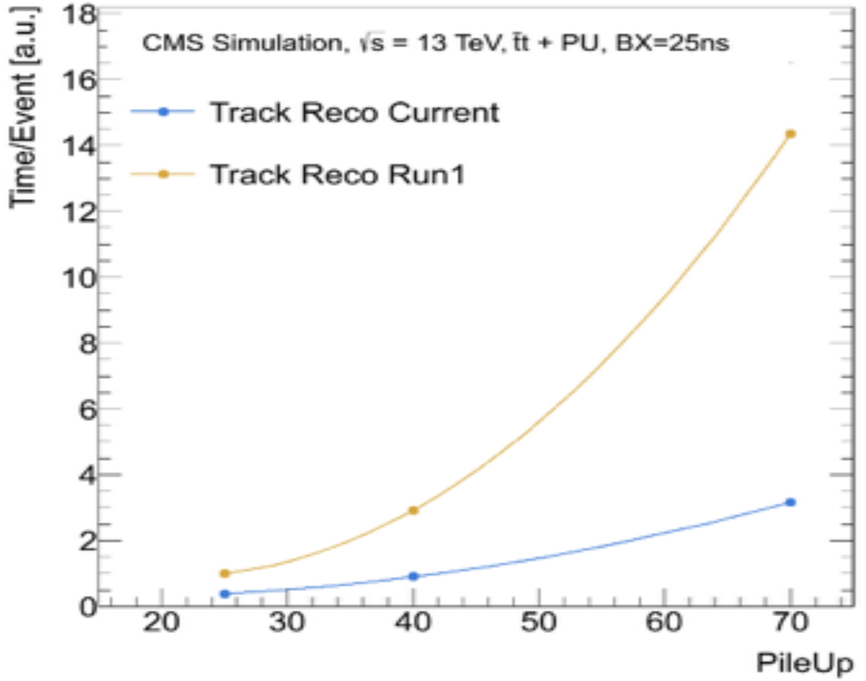
European Research Council

HL-LHC & ML as a must



- ~40 collisions/event
- ~10 sec/event processing time
- (at best)Same computing resources as today

- ~200 collisions/event
- ~minute/event processing time(*)
- (at best)Same computing resources as today

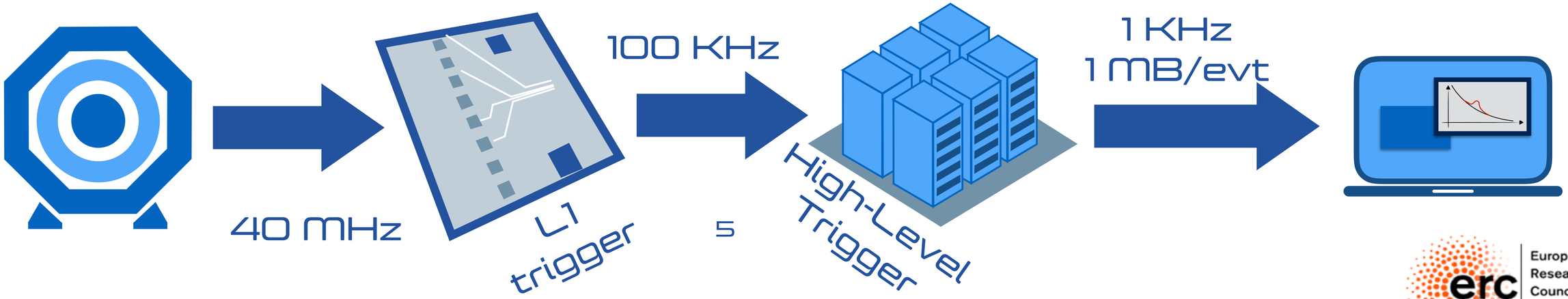


- *Flat budget vs. more needs = **current rule-based reconstruction algorithms will not be sustainable***
- ***Adopted solution:** more granular and complex detectors → more computing resources needed → more problems*
- ***Modern Machine Learning might be the way out***

(*)With nowadays software development

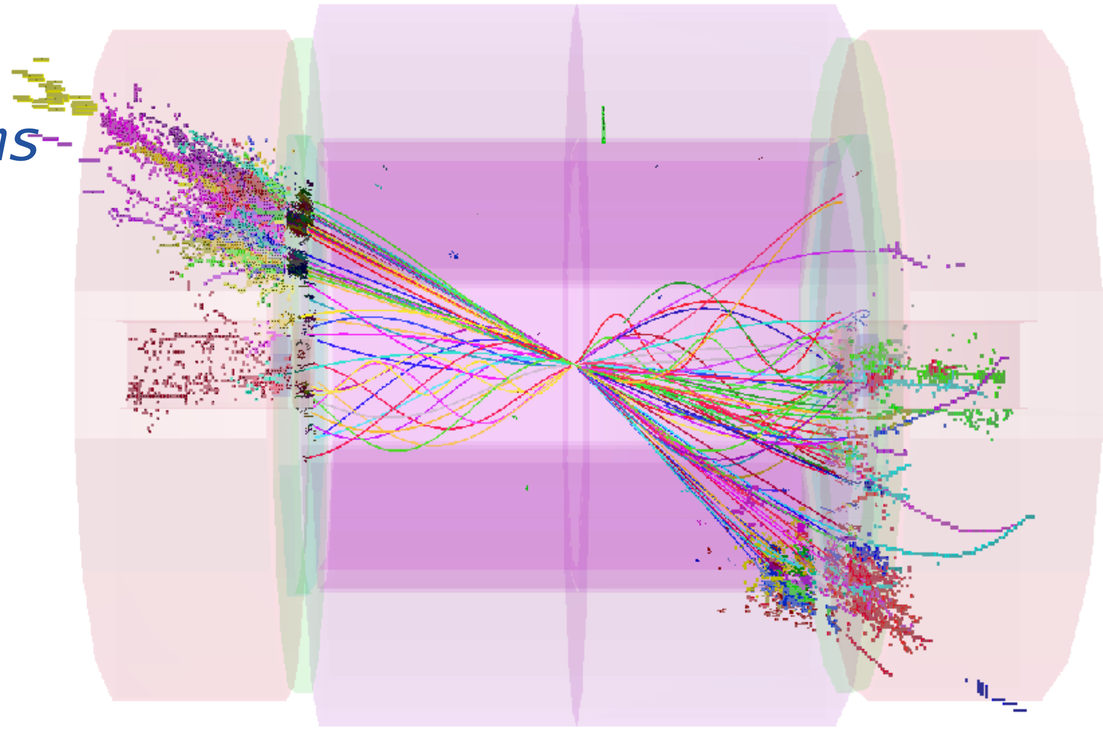
HL-LHC & ML as a must

- *Machine Learning can act as a short-cut to reduce CPU needs to accomplish computational tasks:*
  - *express the answer of a traditional algorithm as a function learned by examples*
- *ML deployment is happening at any level: L1, HLT, reconstruction, Monte Carlo generation & analysis*



Why ML for Future e^+e^-?

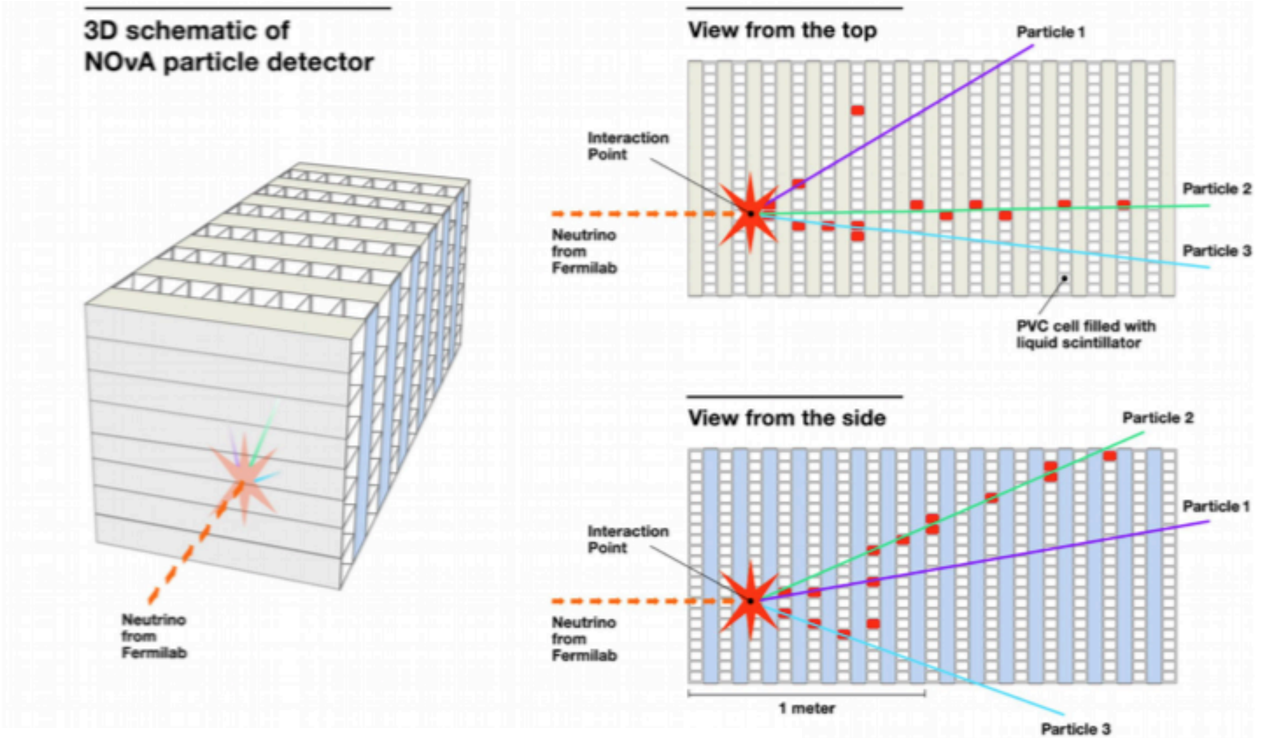
- *A-priori, there is real need to go in this direction for e^+e^- colliders*
  - *traditional rule-based algorithms should scale up easily*
- *On the other hand, this is where HEP is going (pushed by LHC needs, but not only)*
- *Moreover, there might be a performance gain ~ for free with these techniques*
  - *See the successful story with neutrinos*

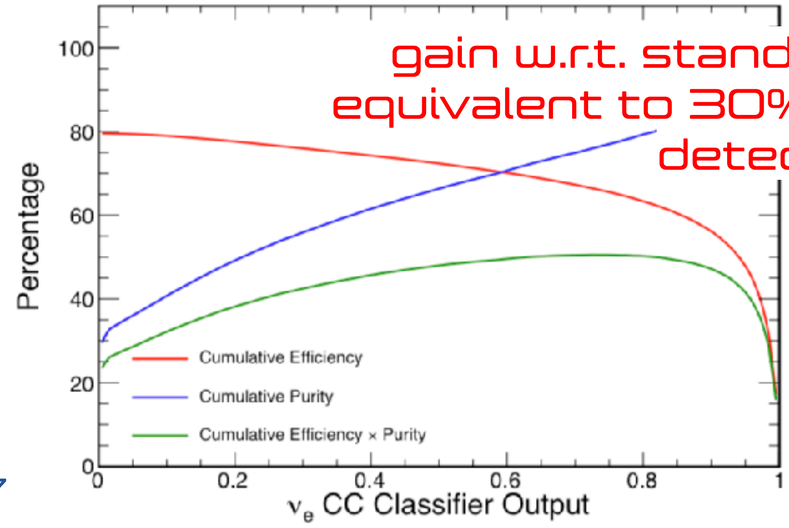


European Research Council

Example: PID for ν experiments

- *Many HEP detectors (particularly underground) more and more structured as regular arrays of sensors*
- *Modern computer-vision techniques work with images as arrays of pixel sensor (in 1D, 2D, and 3D)*
- *These techniques were applied by Nova on electron and muon ID*
- *Impressive gain over traditional techniques (comparable to +30% detector == $$$ saved)*



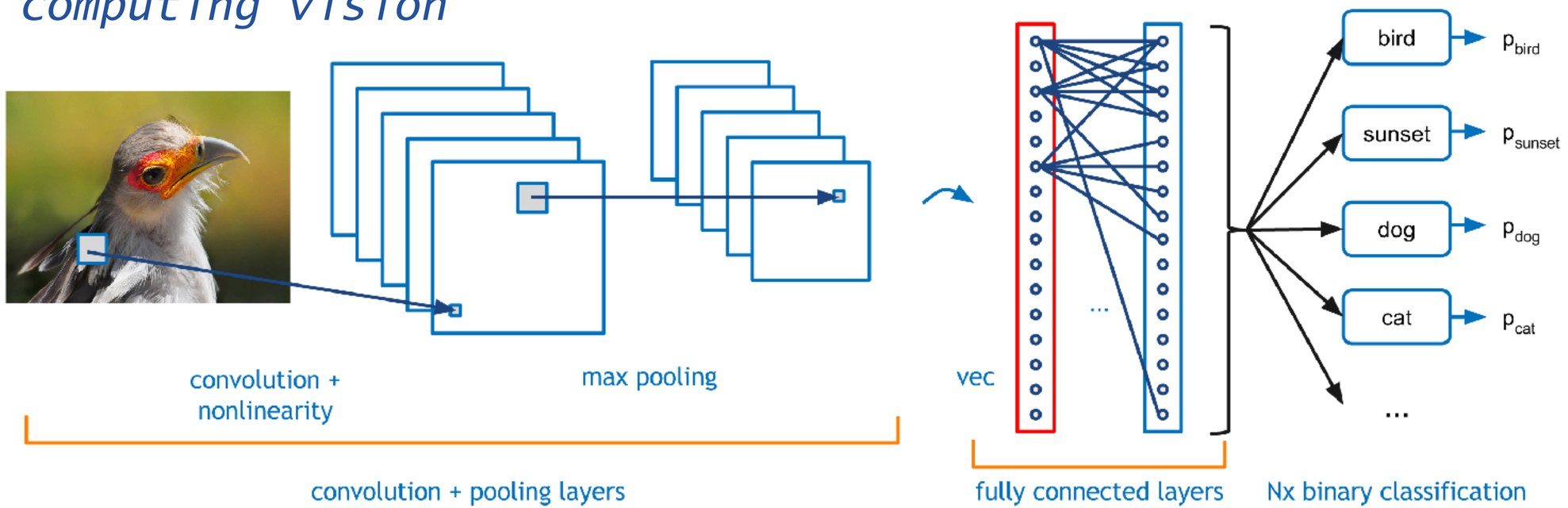


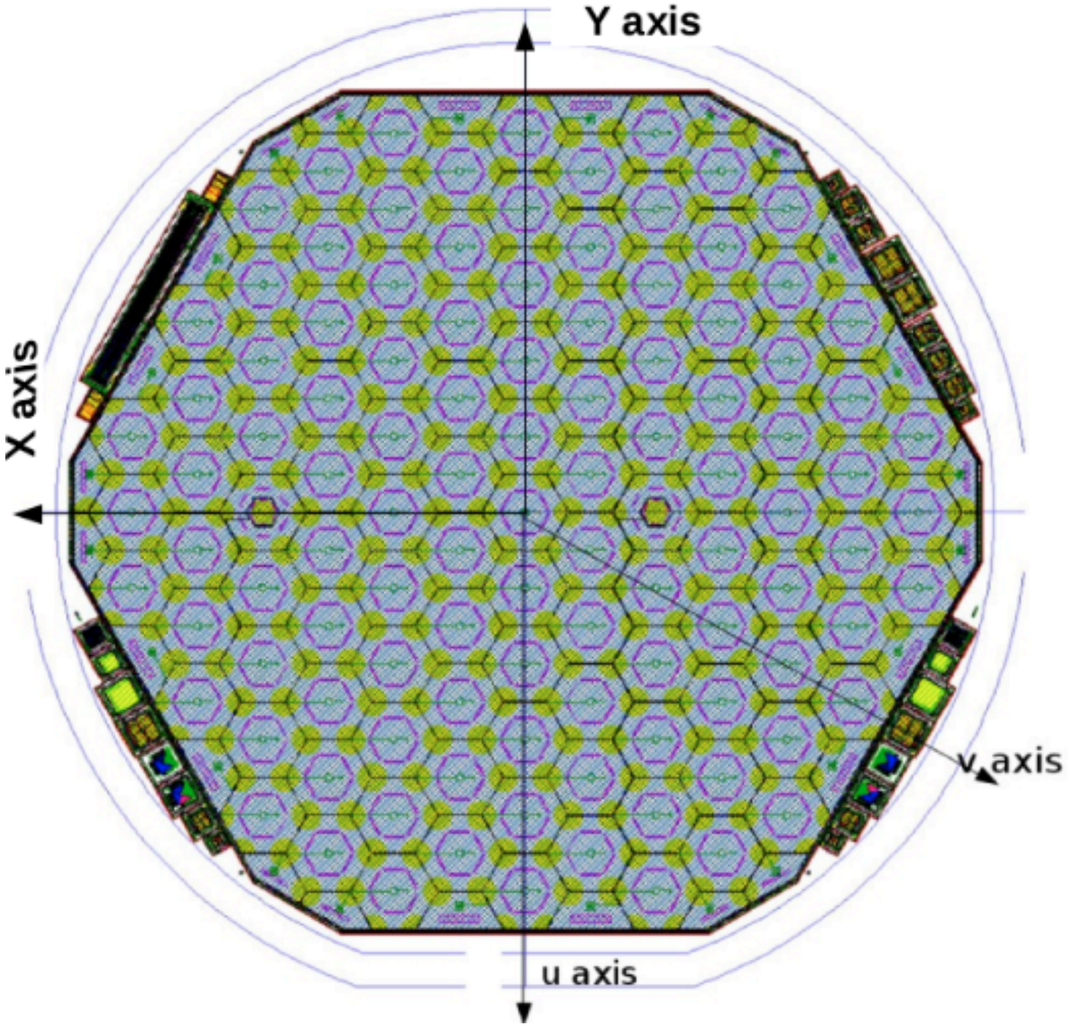
gain w.r.t. standard approach:
equivalent to 30% bigger mass of detector

European Research Council

Which computing-vision technique?

- *The main breakthrough of Deep Leaning is the capability of processing directly raw data*
- *Special architectures read the raw information (e.g., images) and convert them into "smart variables" (high-level features) to accomplish the task*
- *Typical example: convolutional neural networks for image processing & computing vision*





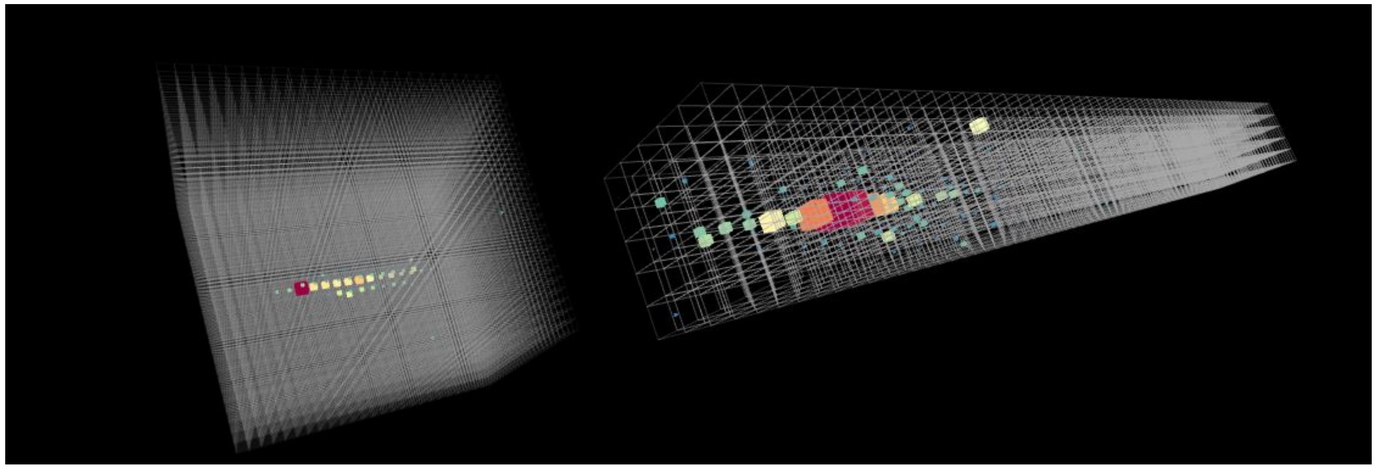
Faster Particle Reconstruction With Computing Vision

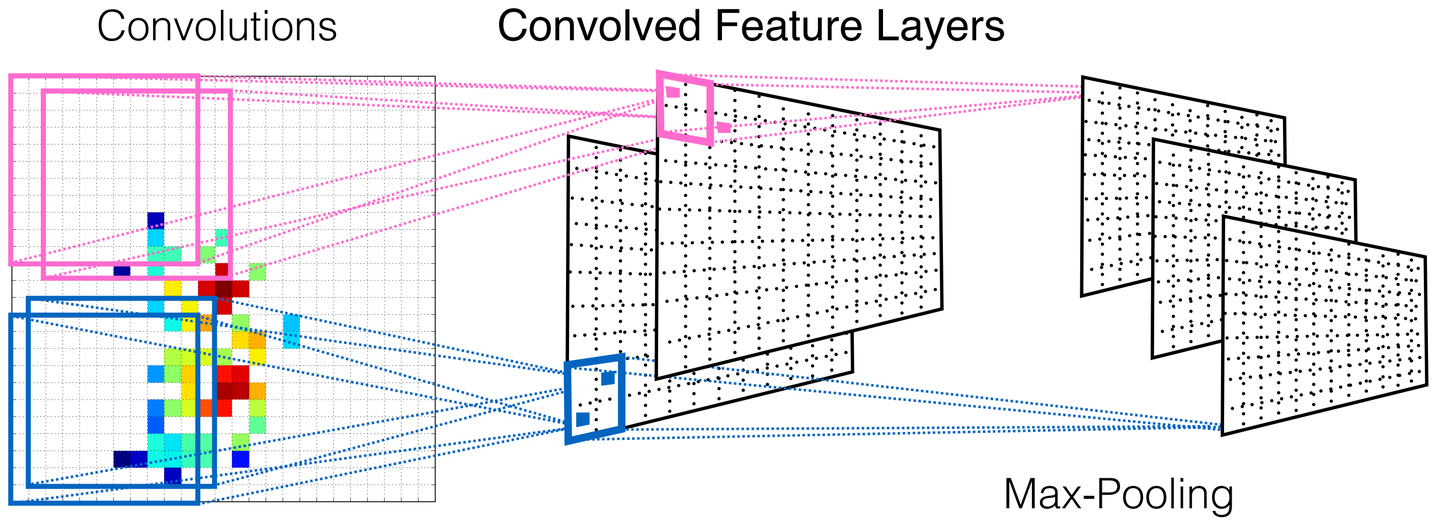
European Research Council

Particle reconstruction as image detection

- *Future detectors will be 3D arrays of sensors with regular geometry*
- *Ideal configuration to apply Convolutional Neural Network*
  - *speed up reconstruction at similar performances*
  - *and possibly improve performances*

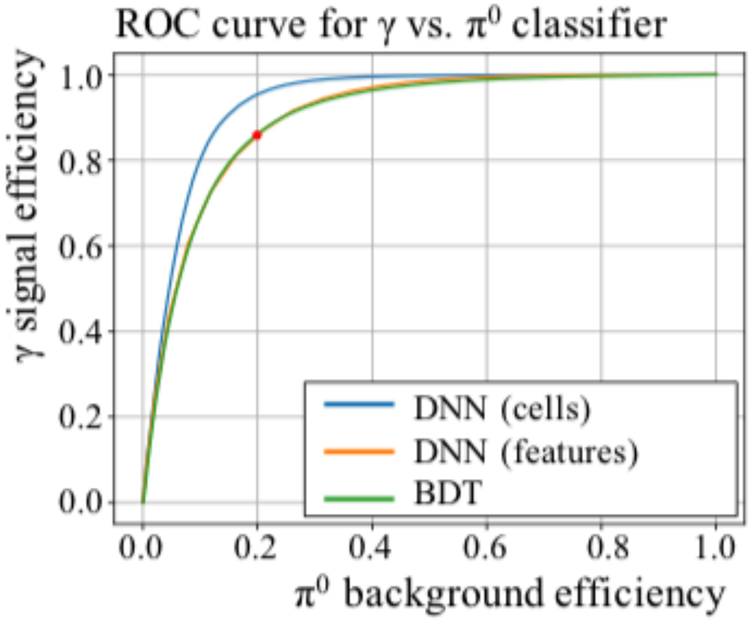
See contribution to NIPS workshop

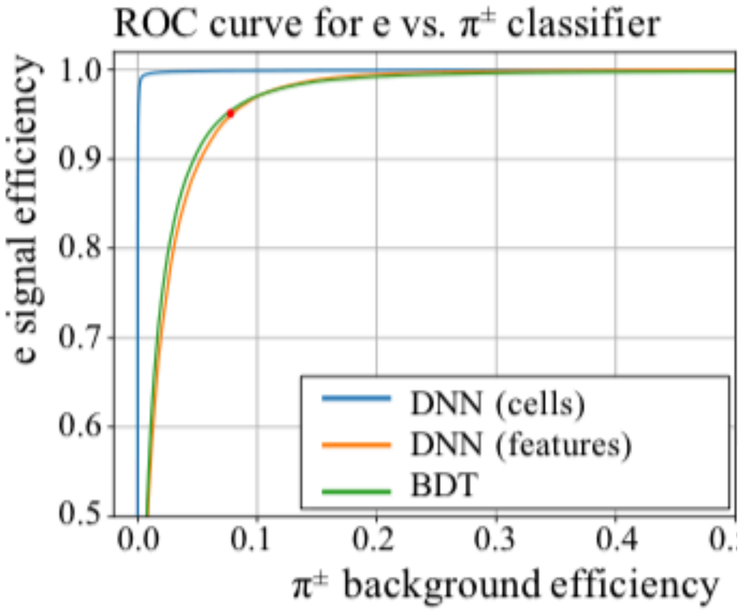




Proof of Principle: Particle ID

- *We tried particle ID on a sample of simulated events*
  - *one particle/event (e, γ, π^0, π)*
- *Different event representations*
  - *high-level features related to event shape (moments of X,Y, and Z projections, etc)*
  - *raw data (energy recorded in each cell)*
- *Pre-filtered pion events to select the nasty ones and make the problem harder*



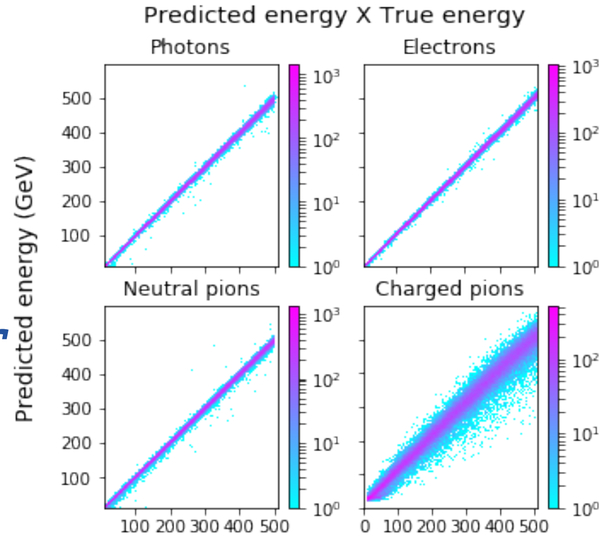


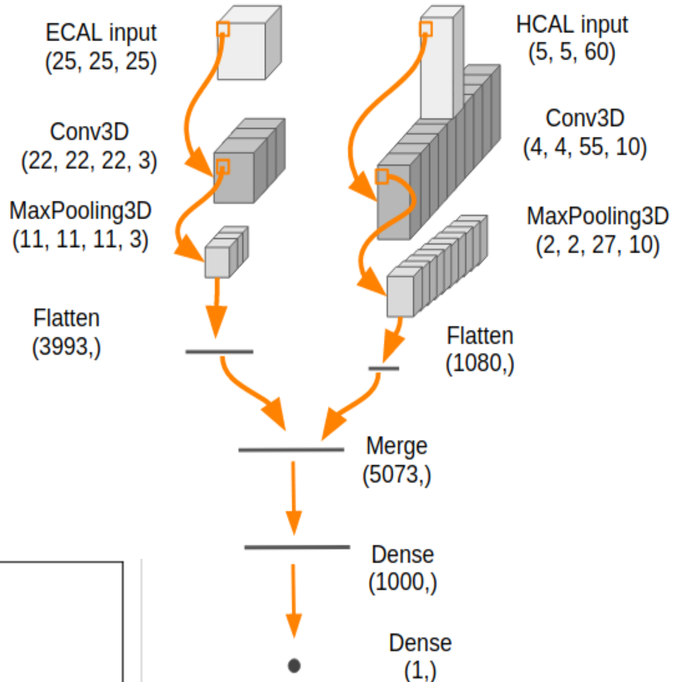
See contribution to NIPS workshop

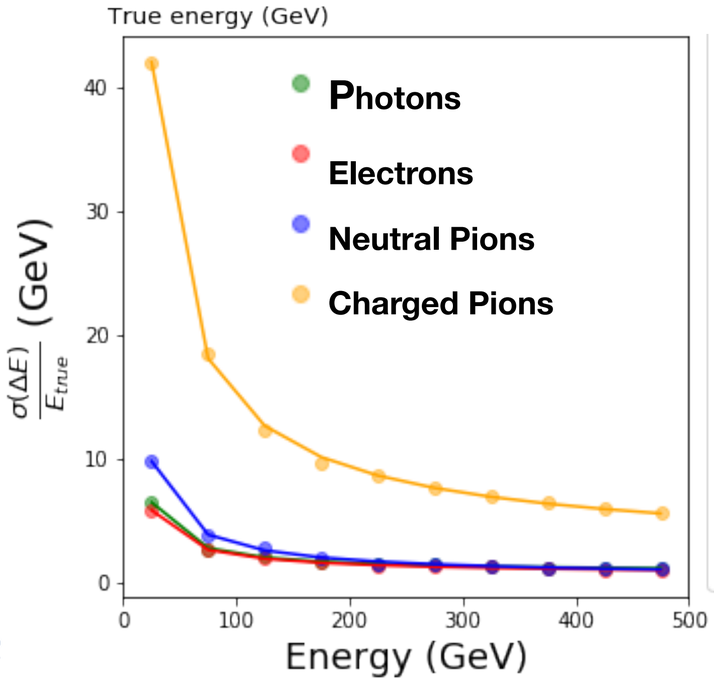
European Research Council

Proof of Principle: Energy Regression

- *Correctly reconstruct energy*
- *ECAL performances better than HCAL (as expected)*
- *π^0 resolution ~ $\sqrt{2}$ γ resolution (as expected)*
- *used only RAW data as inputs*
- *Processing time reduced by <msec*







$\frac{\sigma(\Delta E)}{E_{true}}$ (GeV)

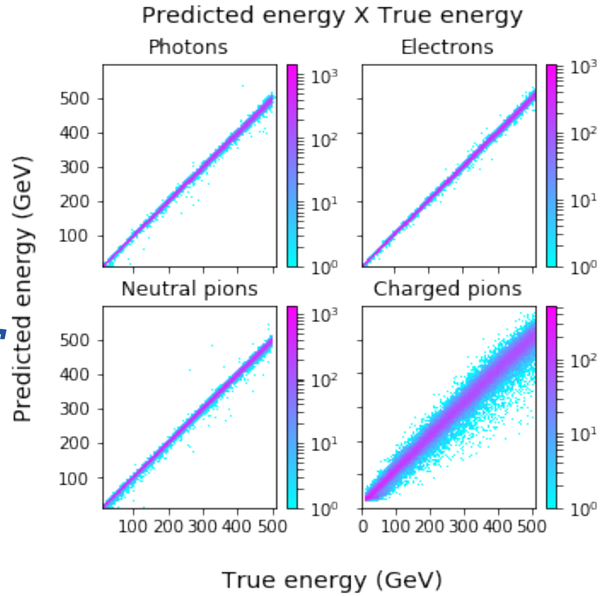
See contribution to NIPS workshop

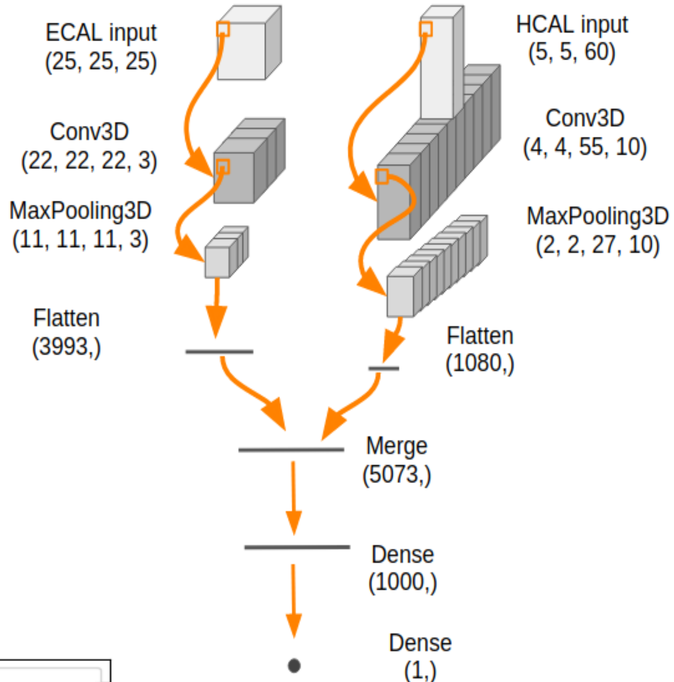
European Research Council

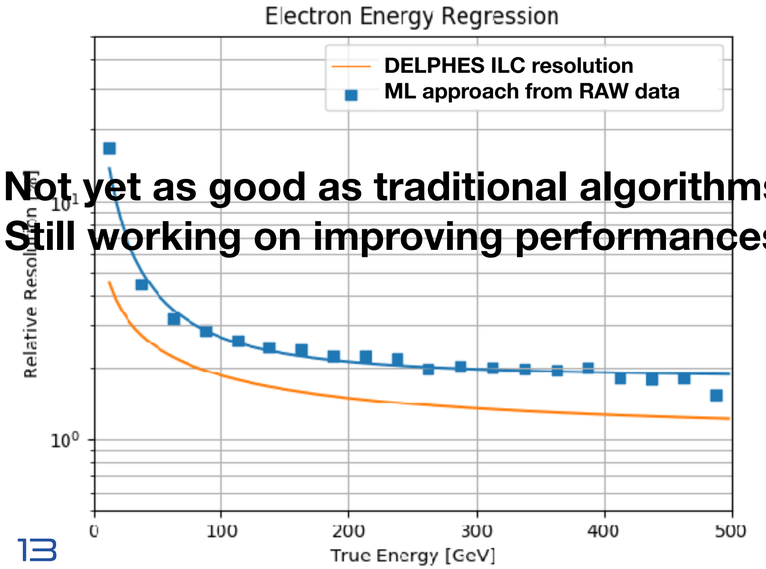
Proof of Principle: Energy Regression

- *Correctly reconstruct energy*
- *ECAL performances better than HCAL (as expected)*
- *π^0 resolution ~ $\sqrt{2}$ γ resolution (as expected)*
- *used only RAW data as inputs*
- *Processing time reduced by <msec*

See contribution to NIPS workshop







Not yet as good as traditional algorithms
Still working on improving performances

$\frac{\sigma(\Delta E)}{E_{true}}$ (GeV)

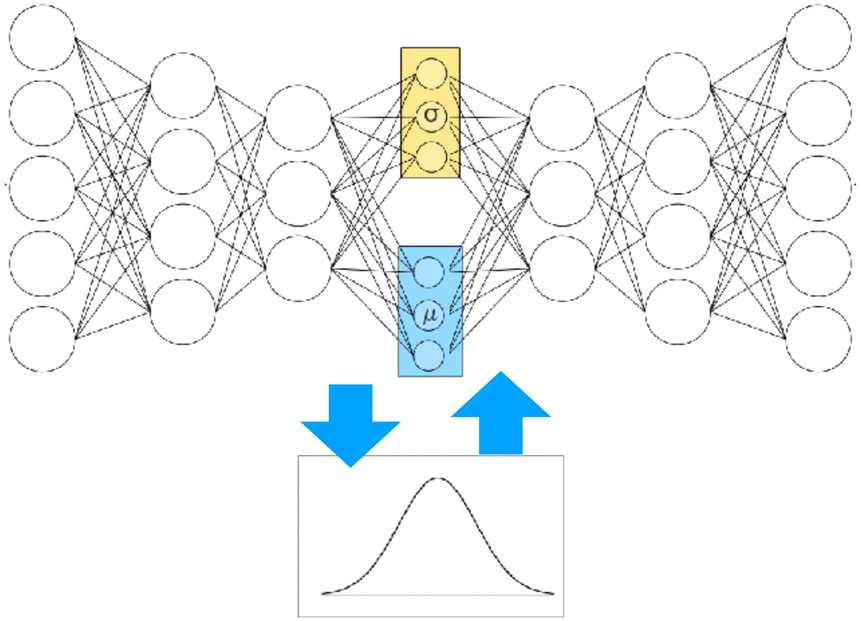
European Research Council

Generating large datasets with small resources

European Research Council

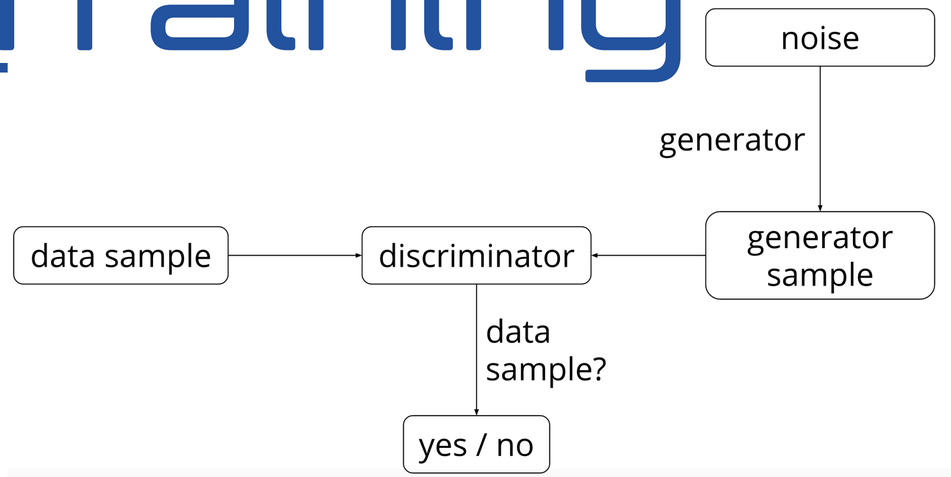
Generative Models

- Neural Networks can be trained to generate events similar to those they are trained on



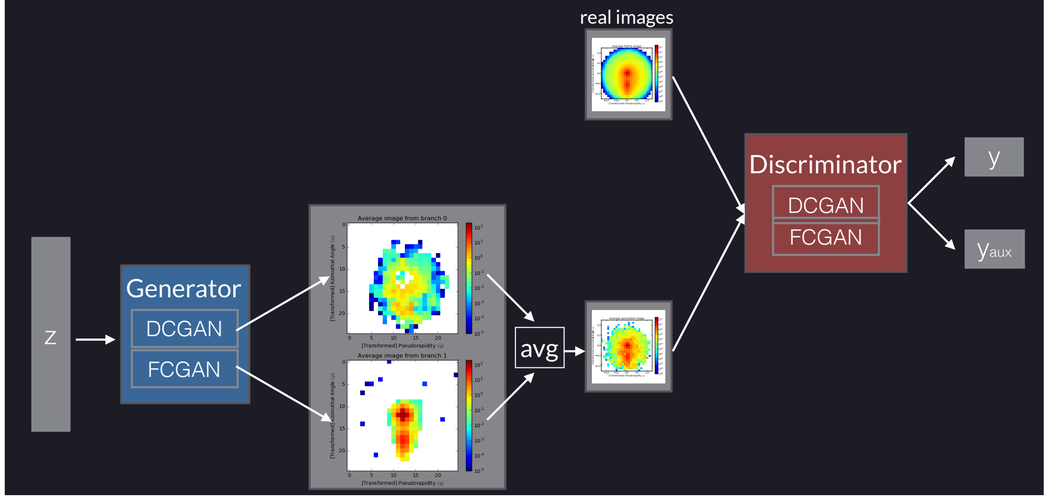
- Two main approaches these days
  - Variational Autoencoder: train the encoder, sample from the latent distribution & decode the sampled point
  - Generative Adversarial networks: train a generator by fooling a classifier

Adversarial Training



- *Two networks trained simultaneously*
  - *Generator: from noise to an event (e.g. image)*
  - *Discriminator: distinguish real events from those created by generator*
- *Loss function given by discriminator*
- *Parameter space = SUM of weights of the two networks*

- DCGAN — convolutional layers in both *G* and *D*
- FCGAN — fully-connected layers in both *G* and *D*
- HYBRIDGAN — a combination of the two:

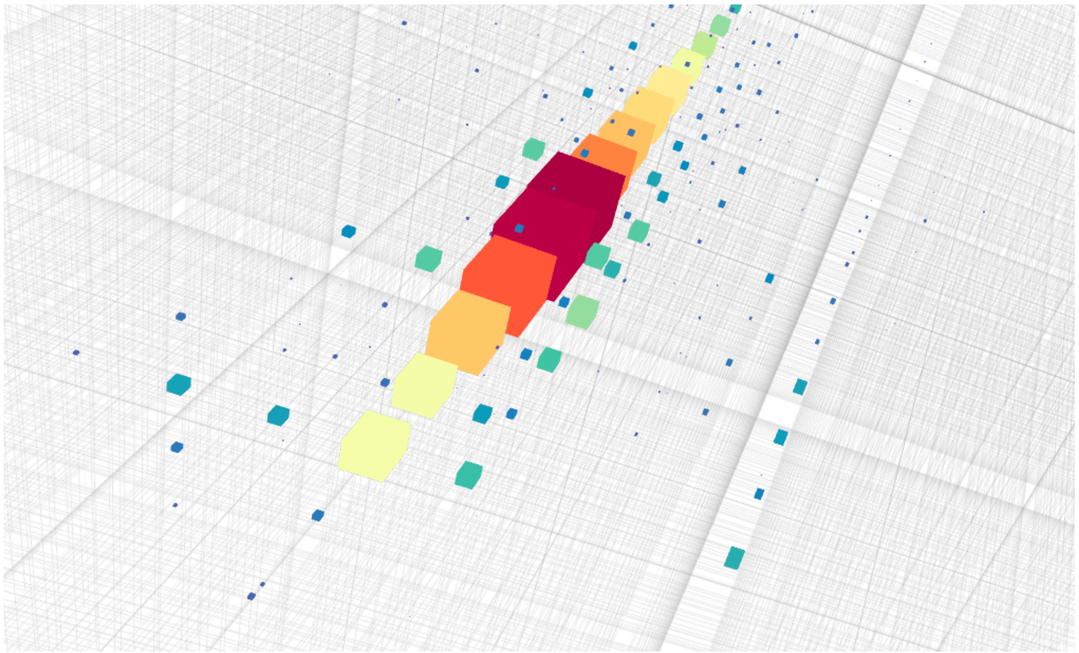


arXiv:1701.05927

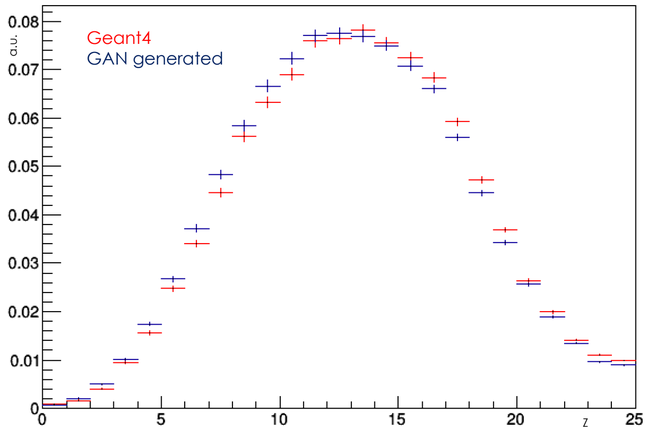
Particle shower generation

See contribution to NIPS workshop

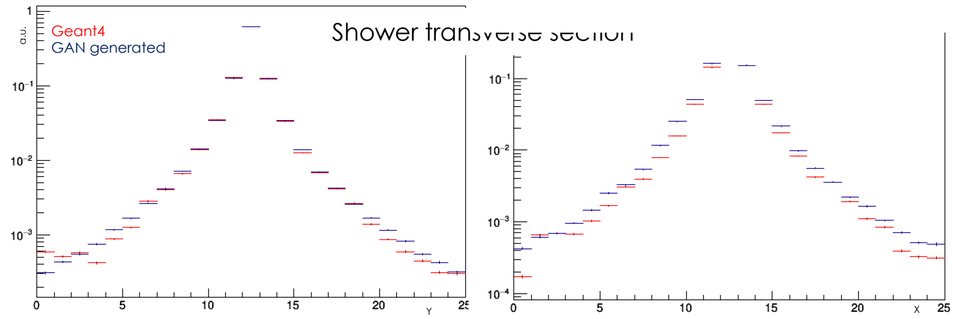
- Start from random noise
- Works very well with images
  - Applied to electron showers in digital calorimeters as a replacement of GEANT

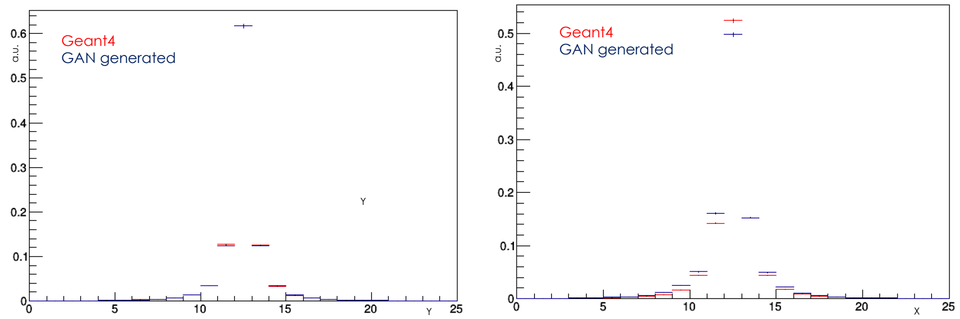


Shower longitudinal section



Shower transverse section



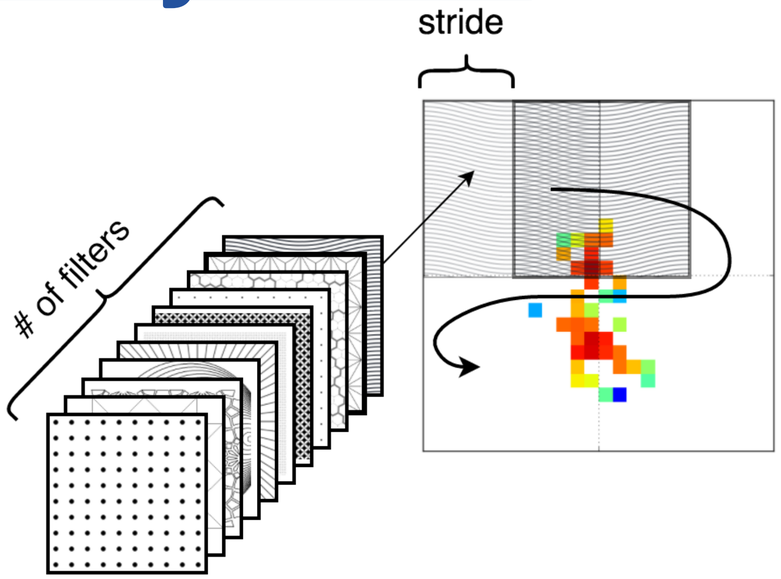


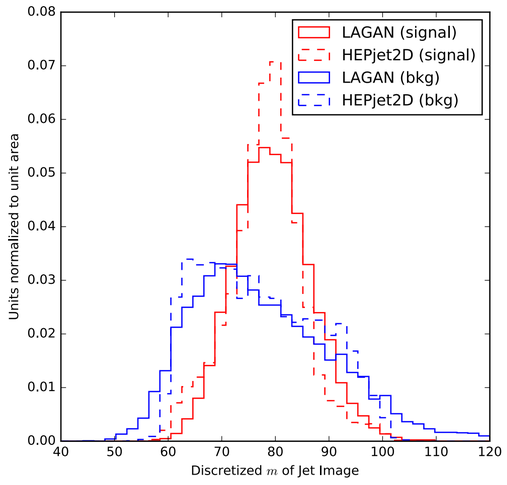
see also de Olivera, Paganini, and Nachman
https://arxiv.org/abs/1712.10321

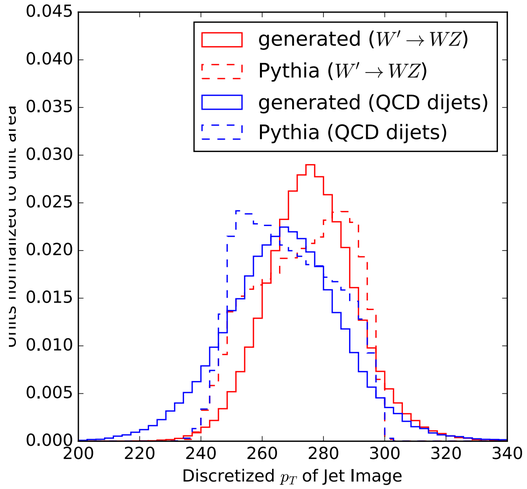
European Research Council

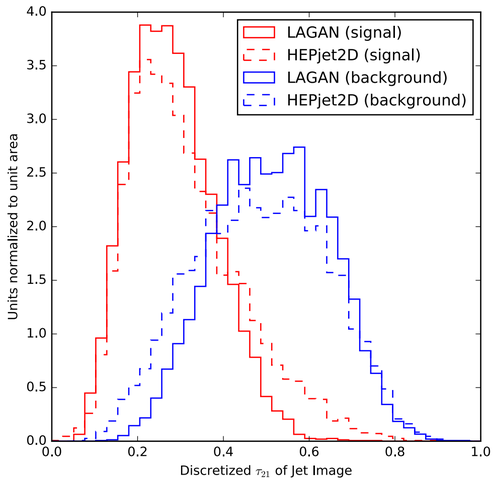
Generating full jets

- Start from random noise
- Works very well with images
  - Applied to electron showers in digital calorimeters as a replacement of GEANT







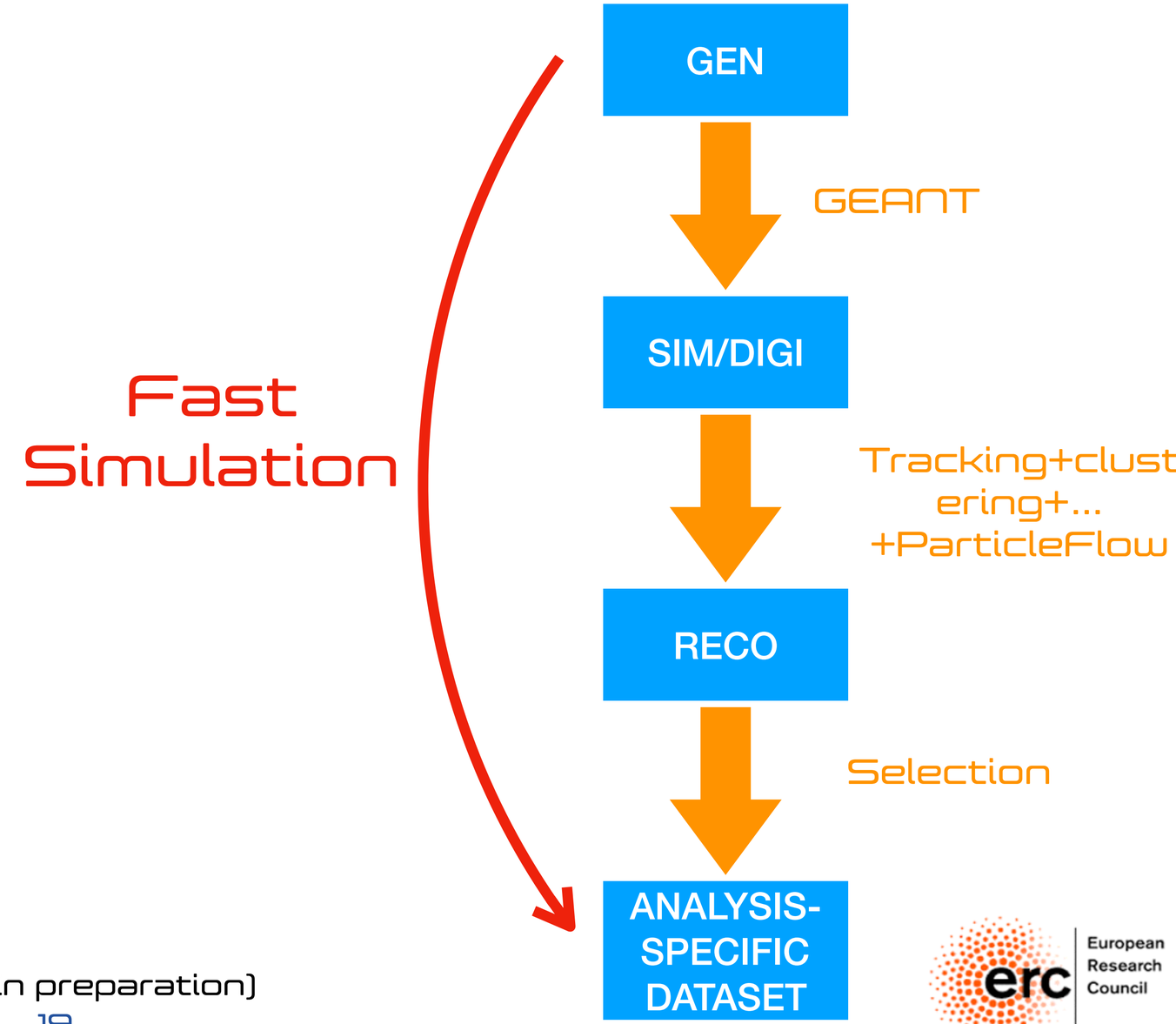


de Olivera, Paganini, and Nachman
https://arxiv.org/pdf/1701.05927.pdf

Figure 6: The distributions of image mass $m(I)$, transverse momentum $p_{\mathrm{T}}(I)$, and n-subjettiness $\tau_{21}(I)$. See the text for definitions.

Analysis-specific dataset generation

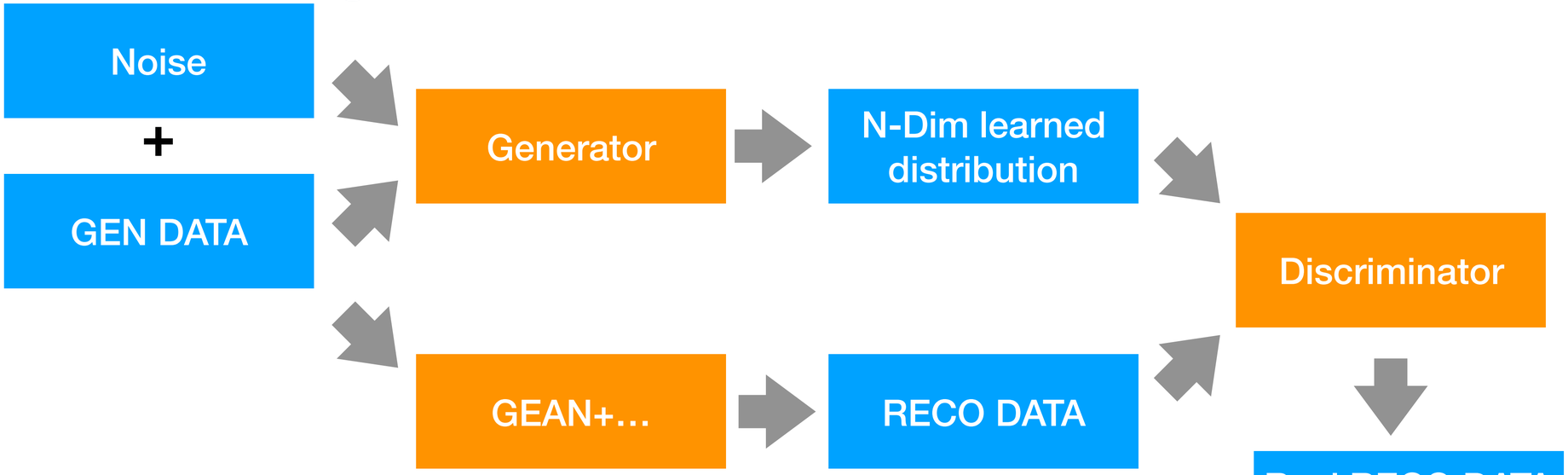
- In view of large statistics needs, one can use generative models as statistics augmentation tools
- For instance, could generate expert-feature quantities used in an analysis (muon four-momenta, jet momenta, etc.)
- Like sampling from histogram with two main advantages
  - no need to bin
  - generalizes to multi-dimensional problems



with K. Dutta, N. Amin, B. Hashemi and D. Olivito (in preparation)

An example

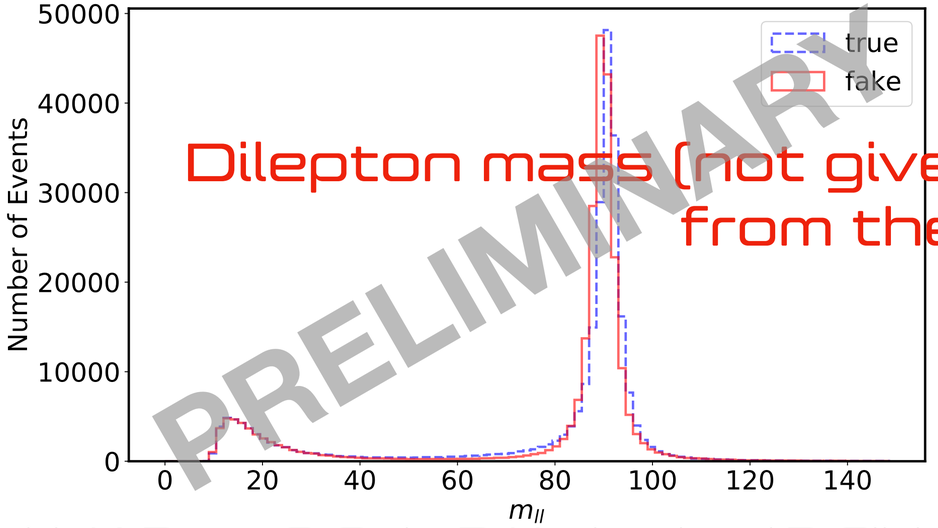
- *Dimuon events at LHC*
- *Typical analysis would use a few handful of quantities (muon momenta, isolation, jet pTs, etc)*
- *Can learn the N-dim distribution of these quantities with GAN setup*
- *Can use the generator network as a fastsim tool*

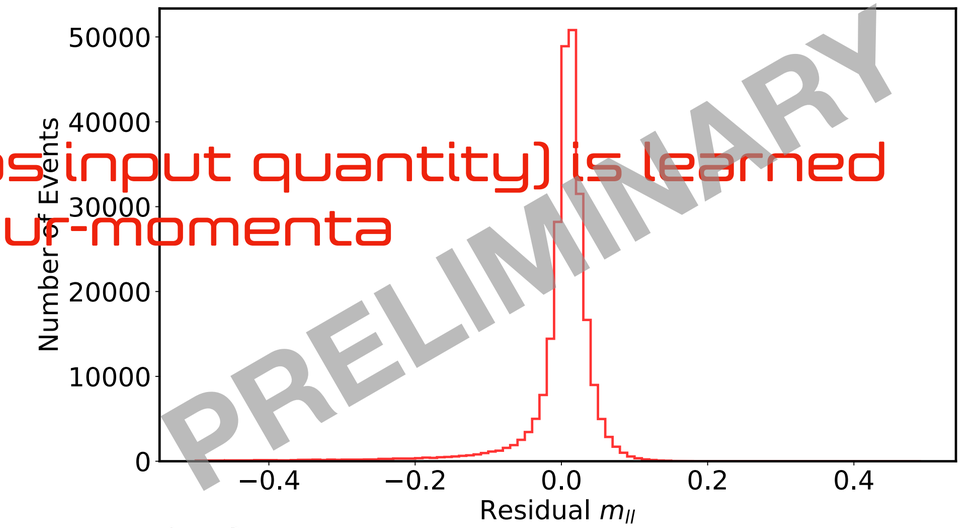


with K. Dutta, N. Amin, B. Hashemi and D. Olivito (in preparation)

An example

- *Dimuon events at LHC*
- *Typical analysis would use a few handful of quantities (muon momenta, isolation, jet pTs, etc)*
- *Can learn the N-dim distribution of these quantities with GAN setup*
- *Can use the generator network as a fastsim tool*



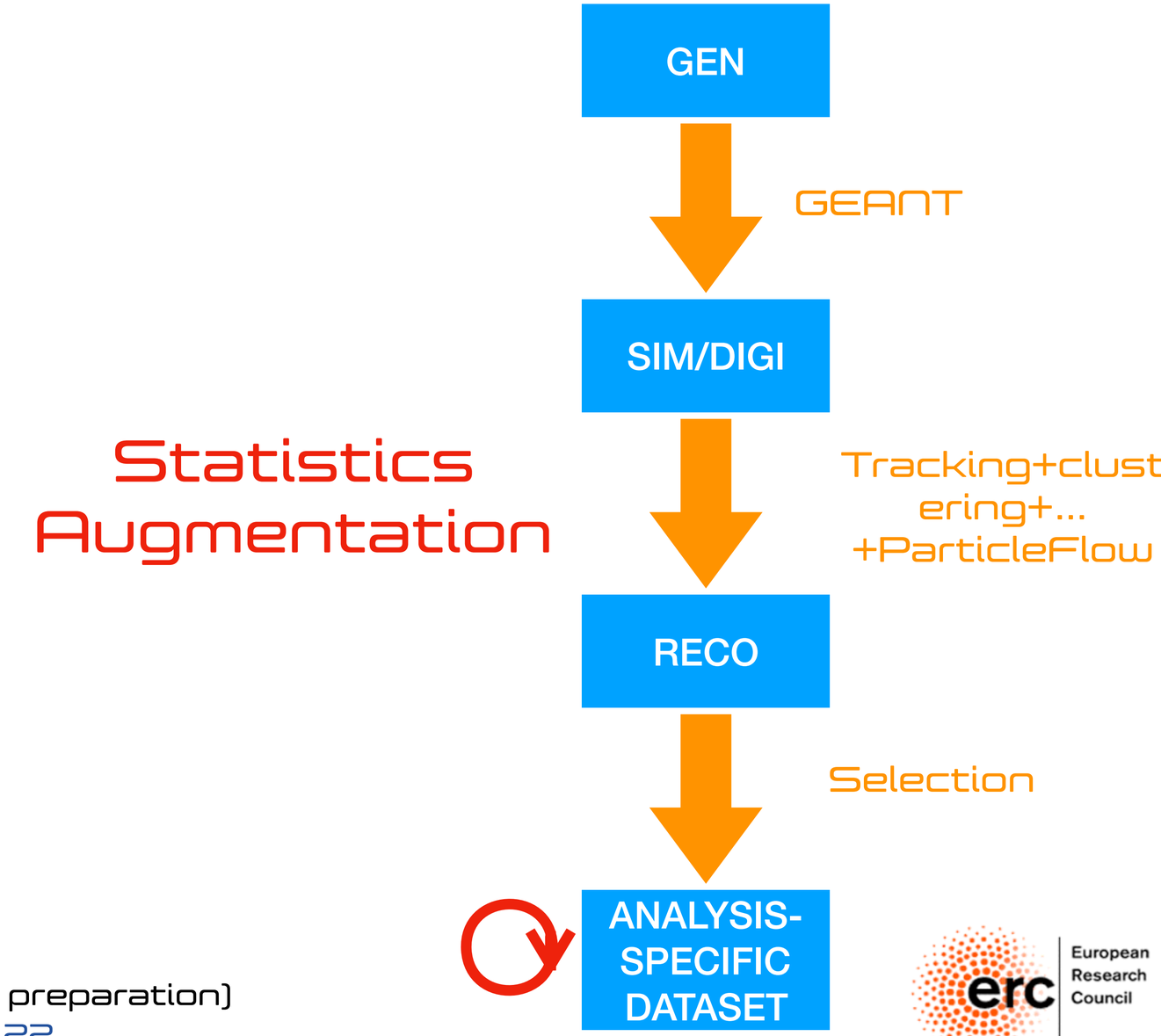


Dilepton mass (not given as input quantity) is learned from the four-momenta

with K. Dutta, N. Amin, B. Hashemi and D. Olivito (in preparation)

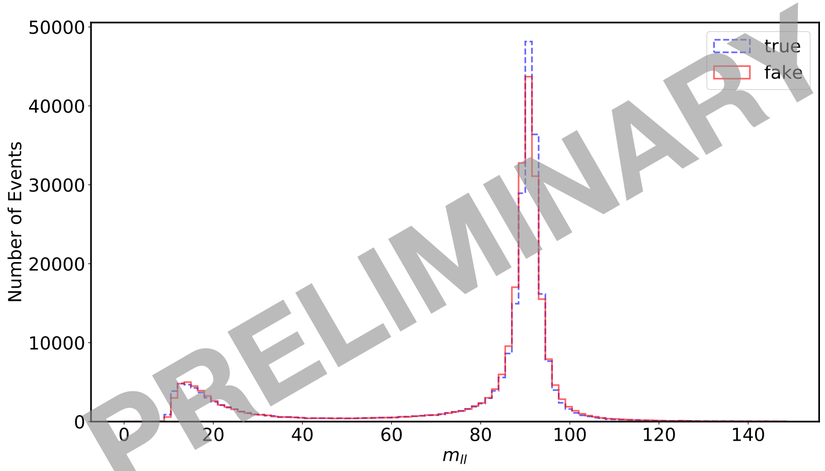
Analysis-specific dataset generation

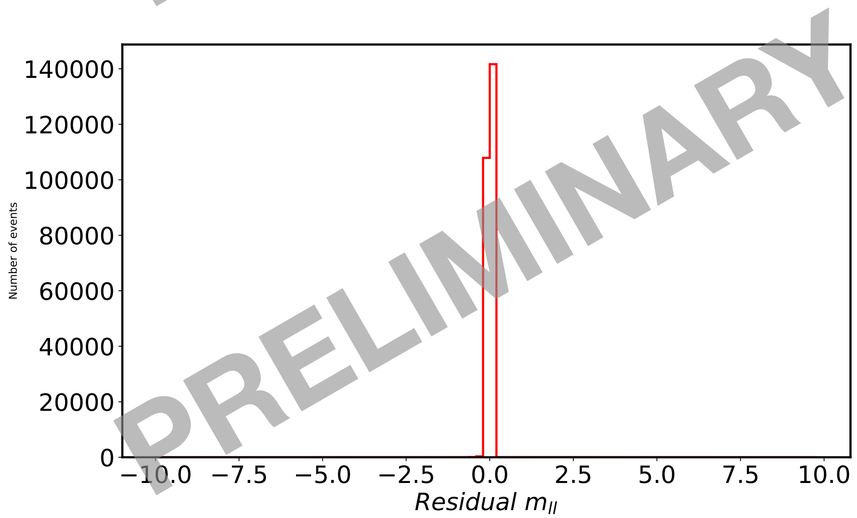
- In view of large statistics needs, one can use generative models as statistics augmentation tools
- For instance, could generate expert-feature quantities used in an analysis (muon four-momenta, jet momenta, etc.)
- Like sampling from histogram with two main advantages
  - no need to bin
  - generalizes to multi-dimensional problems



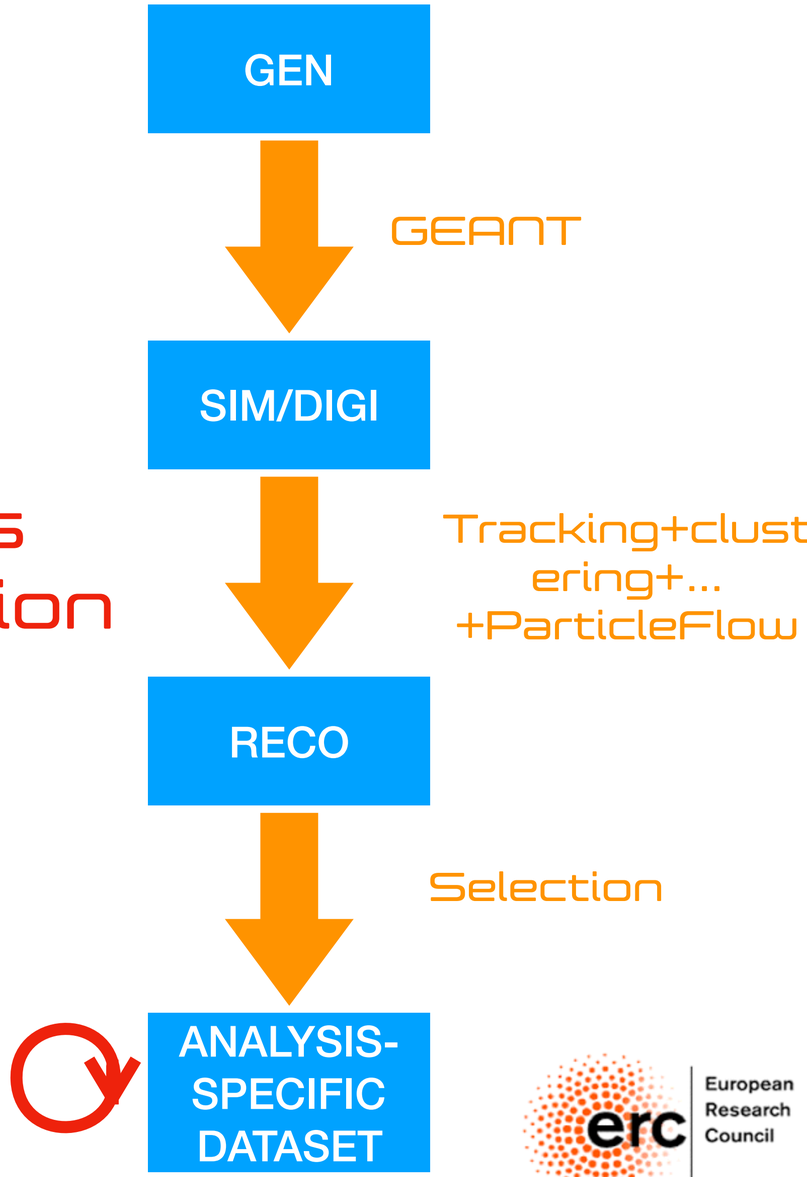
with K. Dutta, N. Amin, B. Hashemi and D. Olivito (in preparation)

Analysis-specific dataset generation





Statistics Augmentation

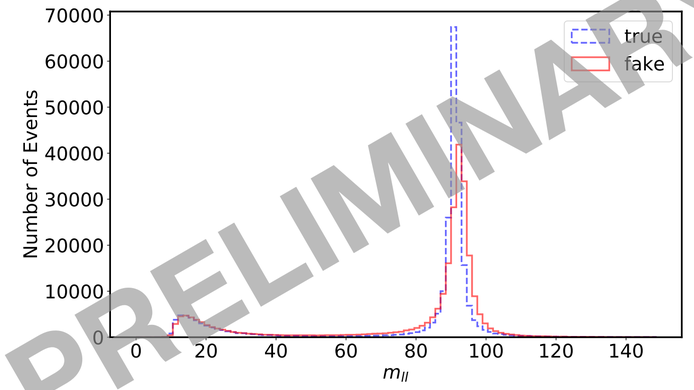


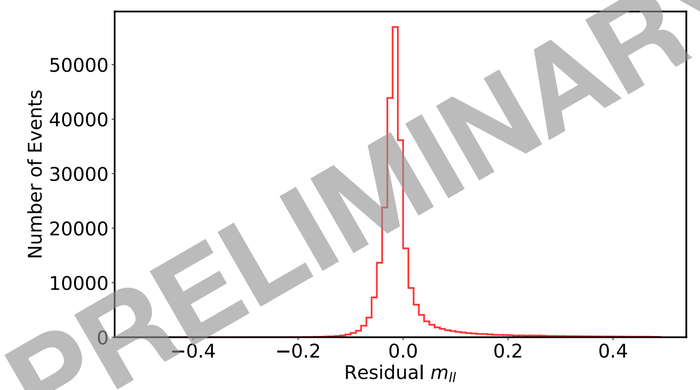
with K. Dutta, N. Amin, B. Hashemi and D. Olivito (in preparation)

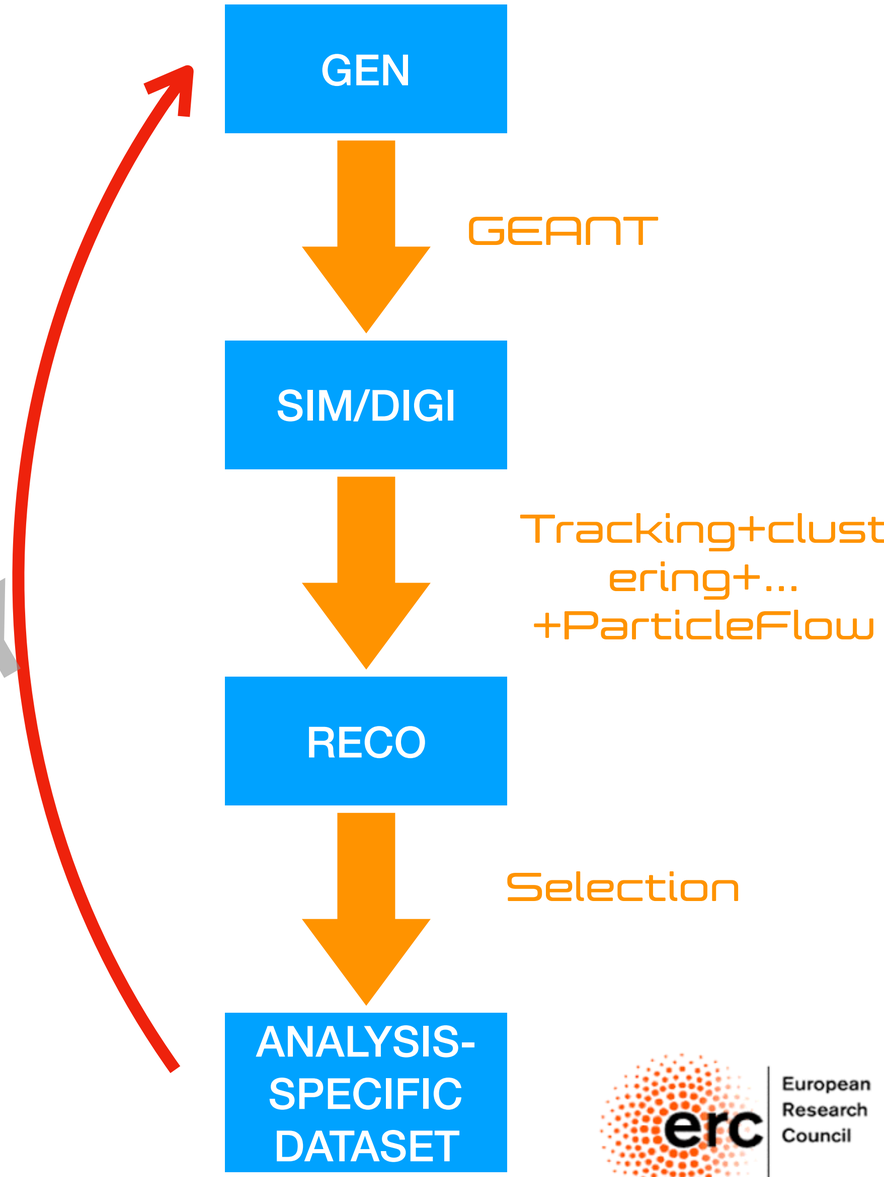
Analysis-specific unfolding

- One could invert the process and learn to predict the GEN features from the RECO ones
- Formally the same procedure
- Invert the role of the target and the input datasets

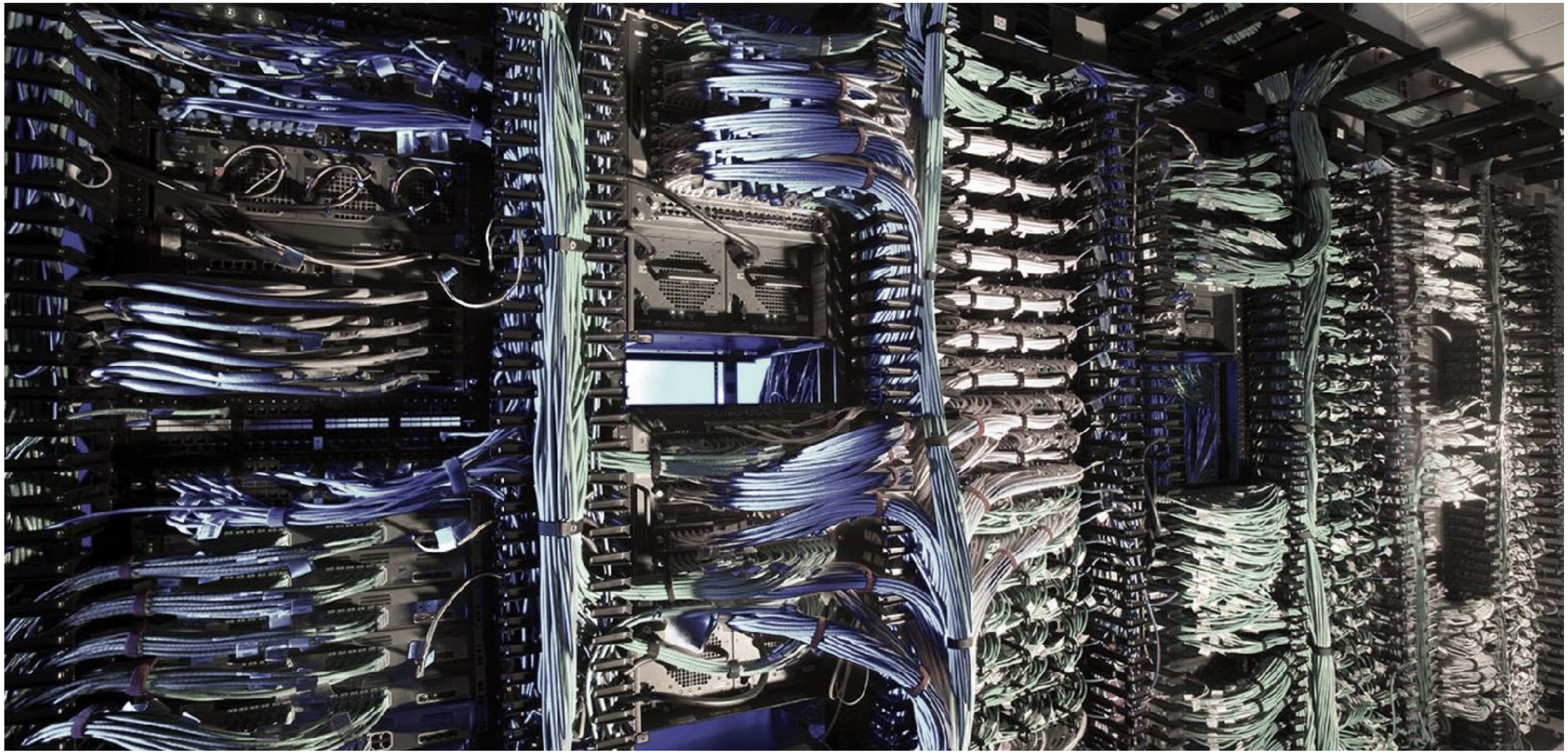
Unfolding







with K. Dutta, N. Amin, B. Hashemi and D. Olivito (in preparation)

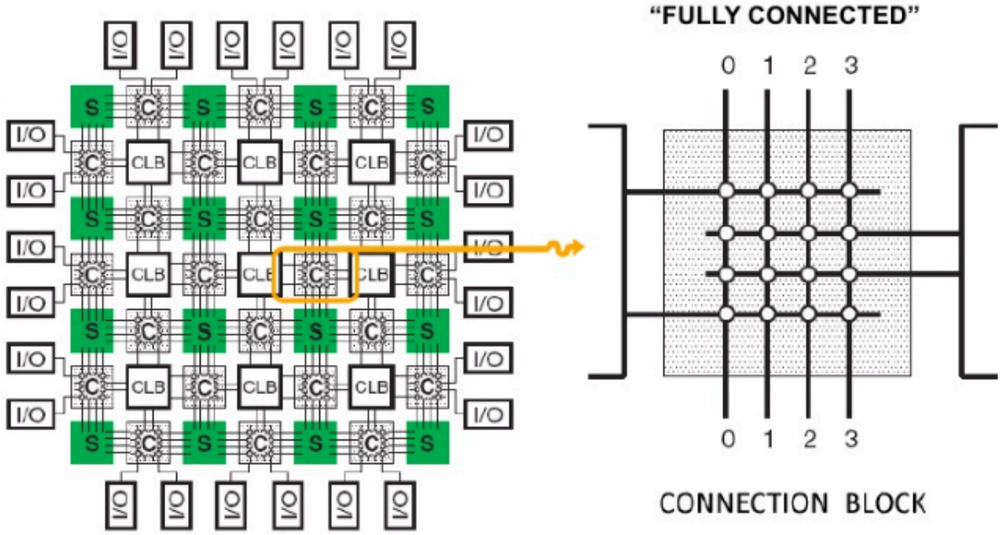


Machine Learning for trigger systems

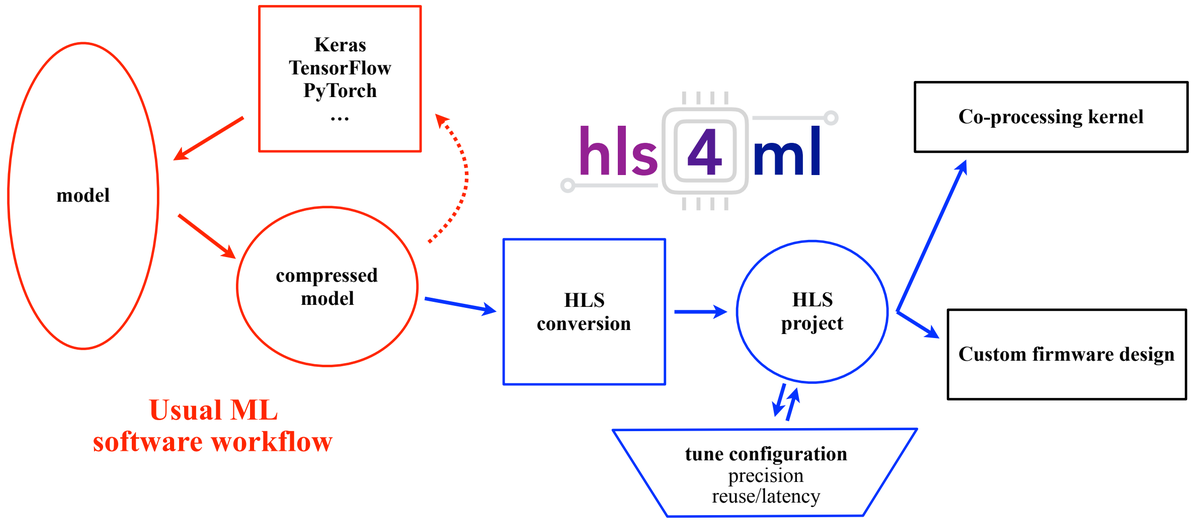
European Research Council

The frontier: bring DL to L1

- *The L1 trigger is a complicated environment*
  - *decision to be taken in ~10 µsec*
  - *only access to local portions of the detector*
  - *processing on Xilinx FPGA, with limited memory resources*
- *Some ML already running @L1*
  - *CMS has BDT-based regressions coded as look-up tables*
- *Working to facilitate DL solutions @L1 with dedicated library*

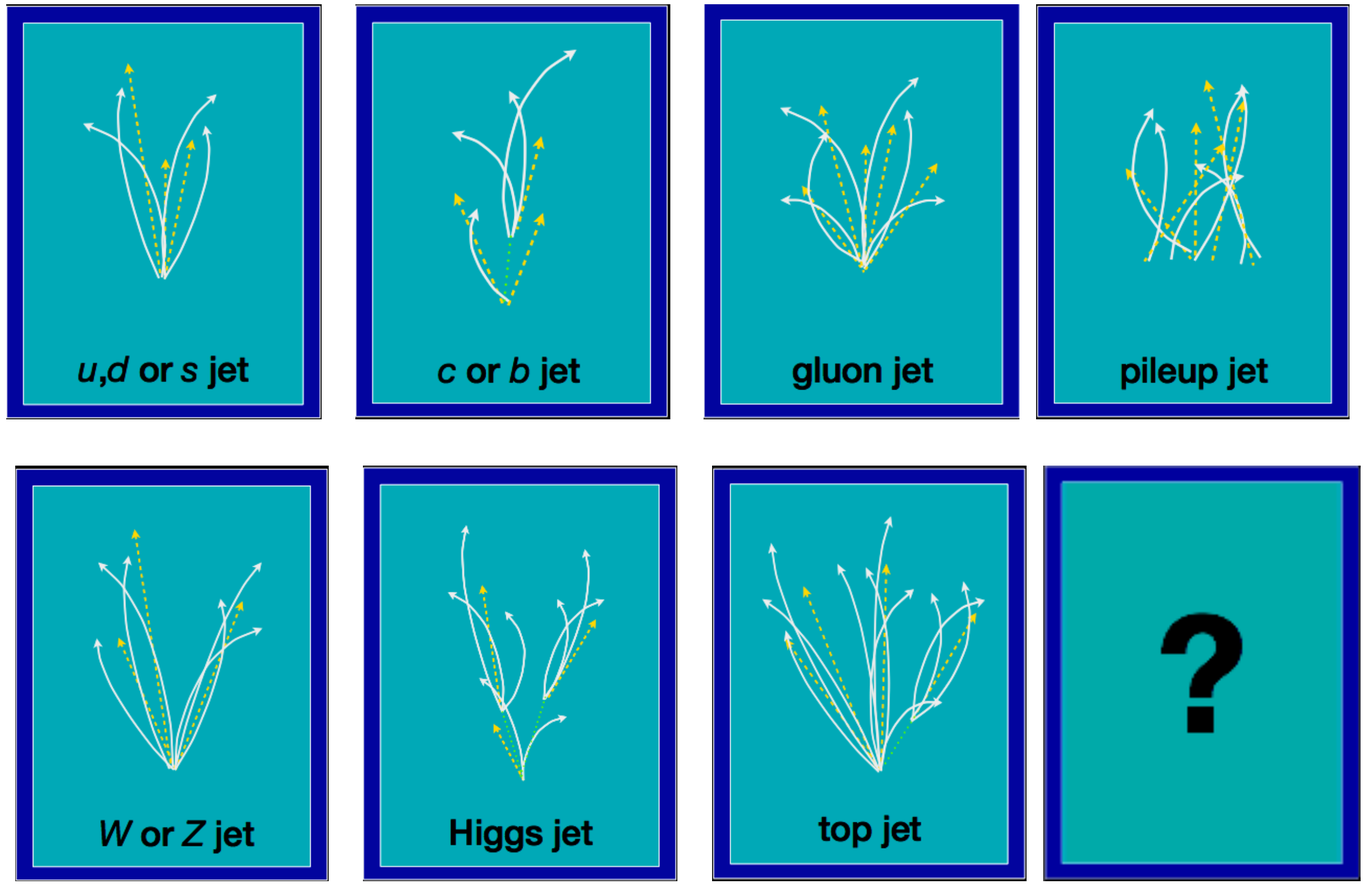


HLS4ML: CERN/FNAL/MIT collaboration



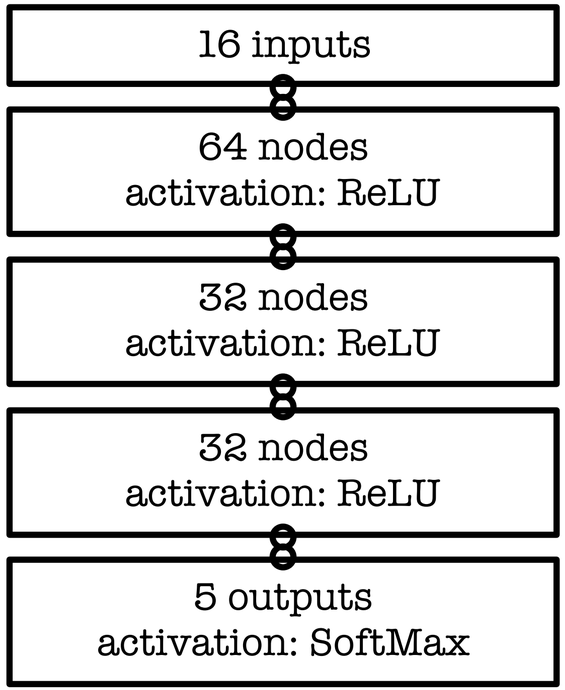
The use case: jet tagging

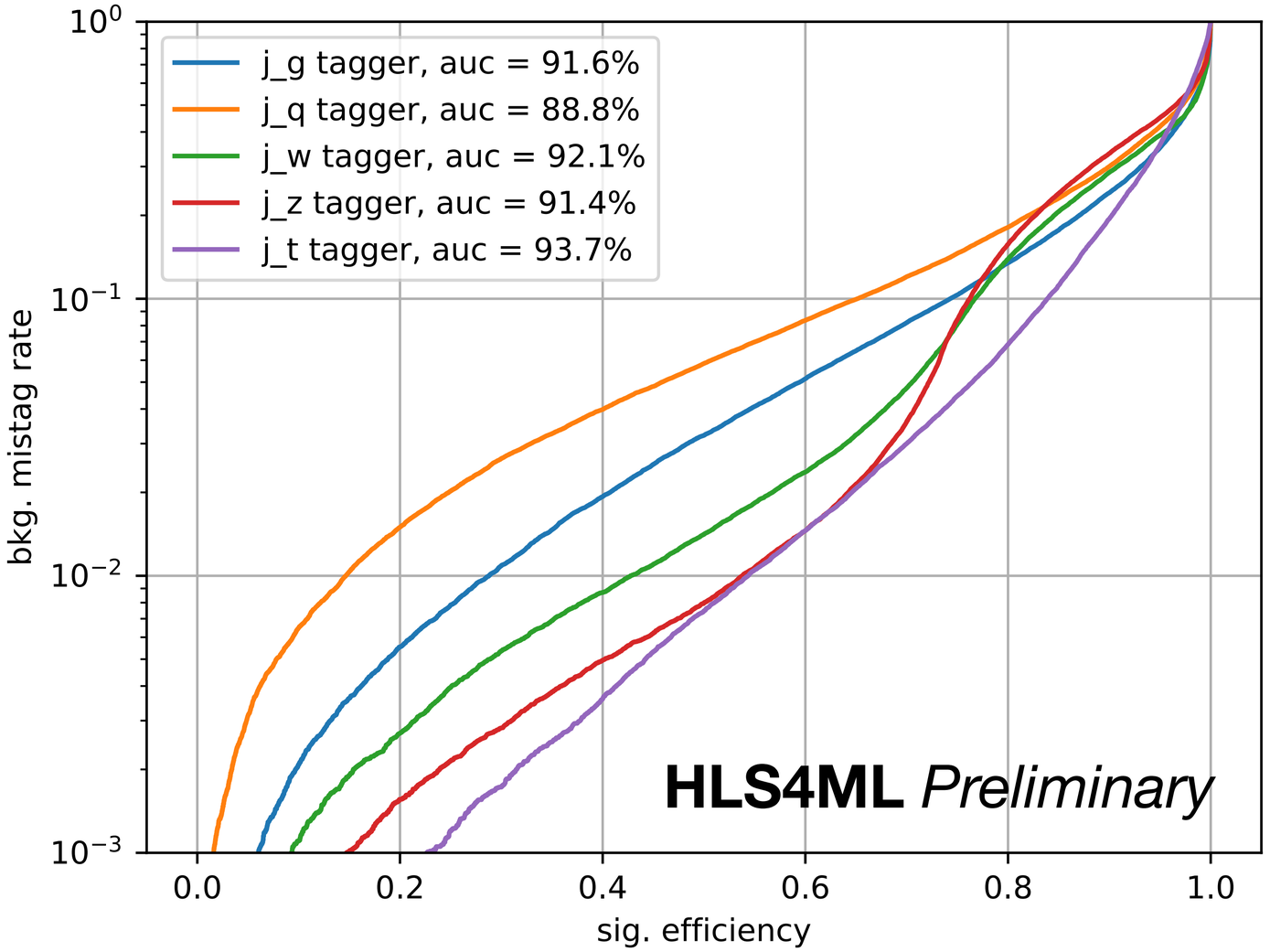
- *An LHC collision produces many kinds of jets*
- *Usually, jet tagging is resources demanding and happens late in the game*
- *We took this as a use case for a NN to be deployed @L1*



A jet multiclass classifier

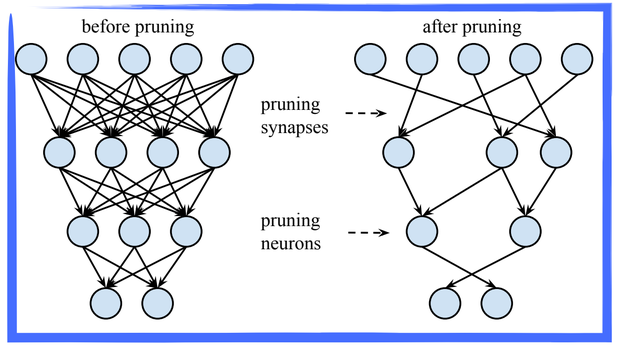
- *Simple DNN based on high-level features (jet masses, multiplicities, energy correlation functions)*

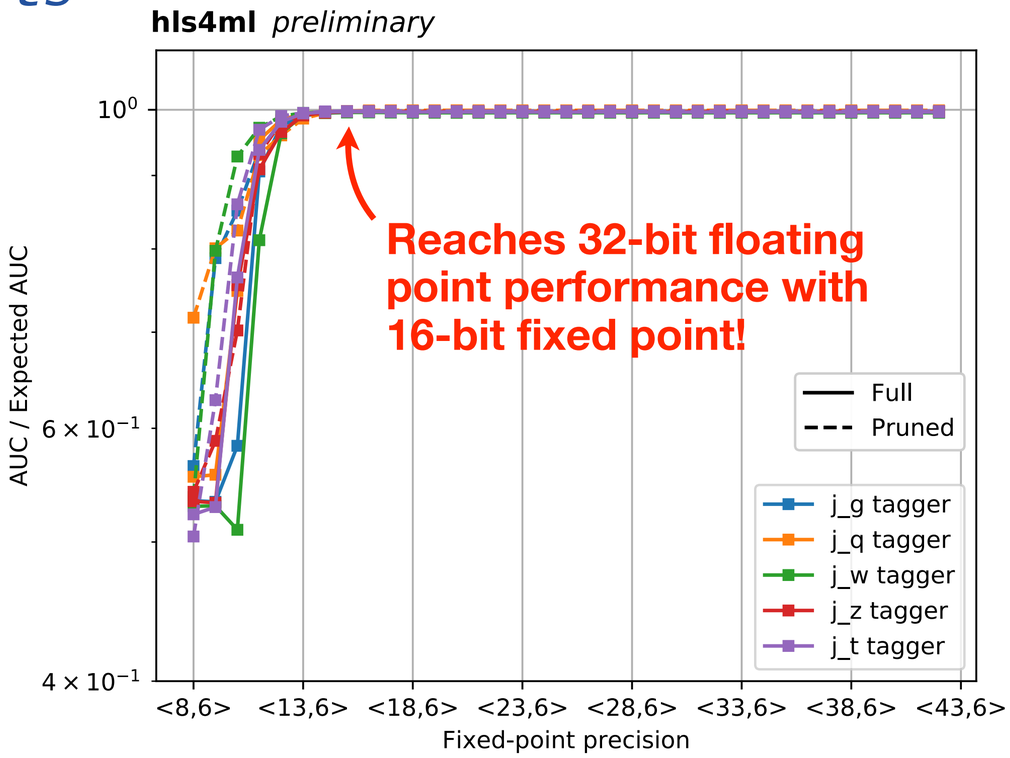


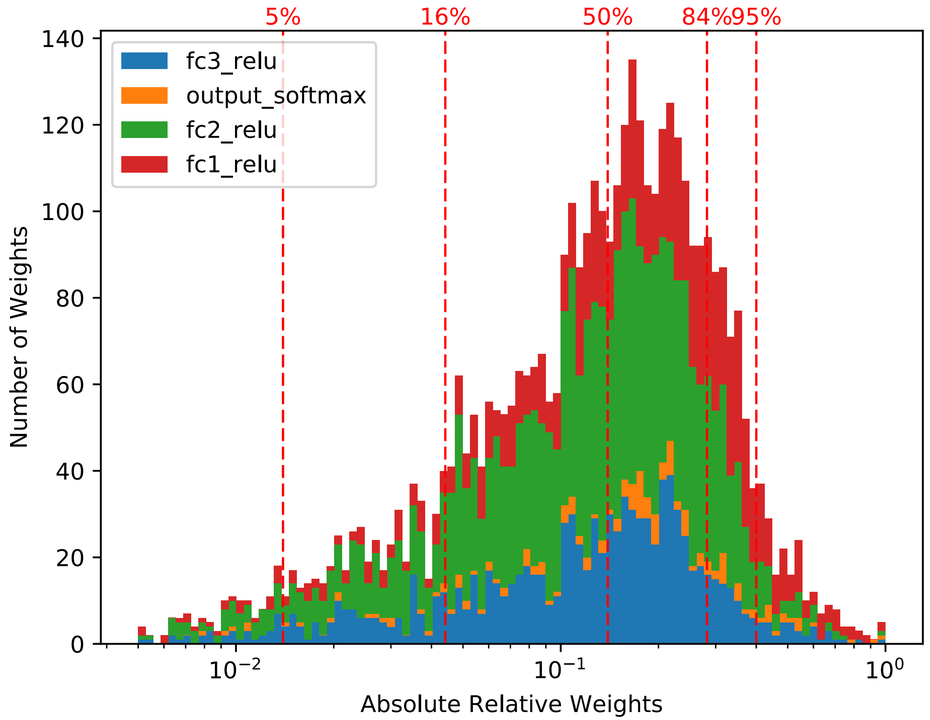


Online deployment: pruning

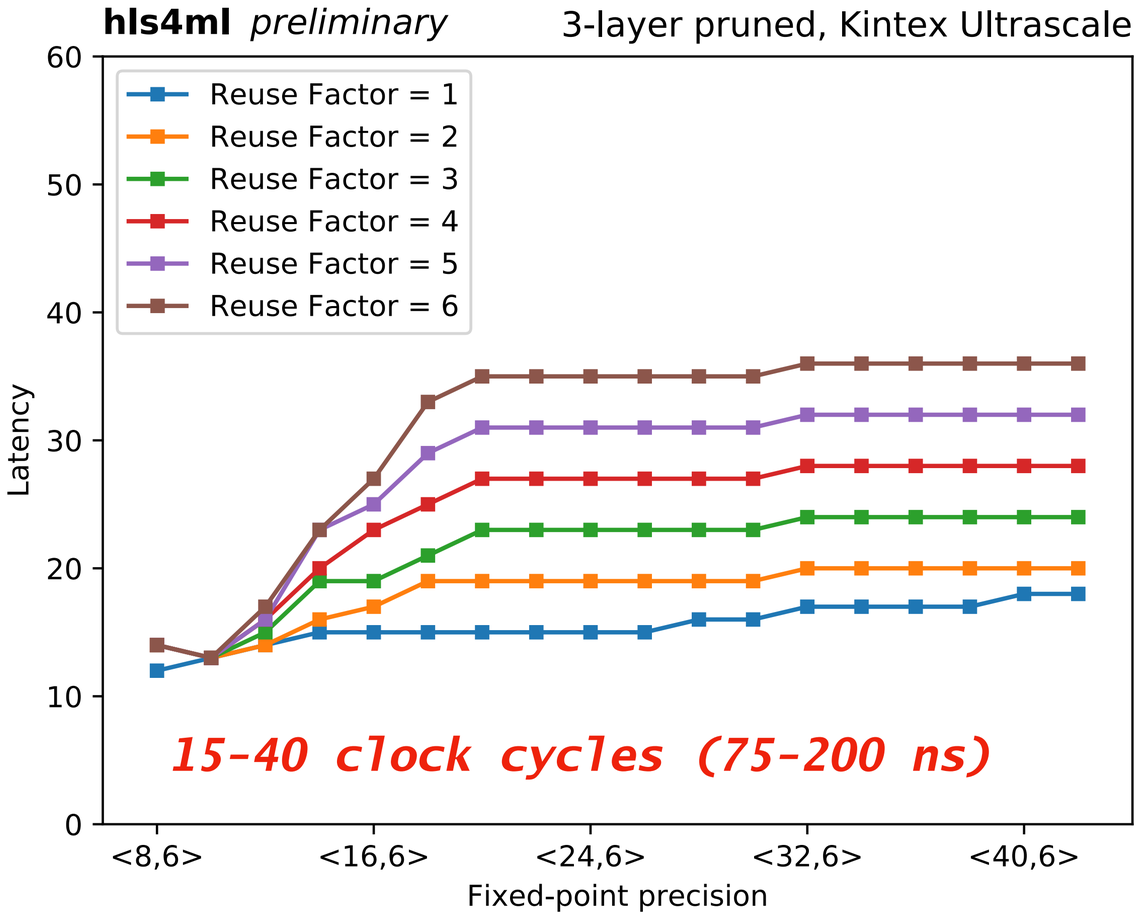
- *Reduced needed resources:*
  - *pruning: removes nodes and connections preserving performances*
  - *quantization: limit numerical precisions saving bits*

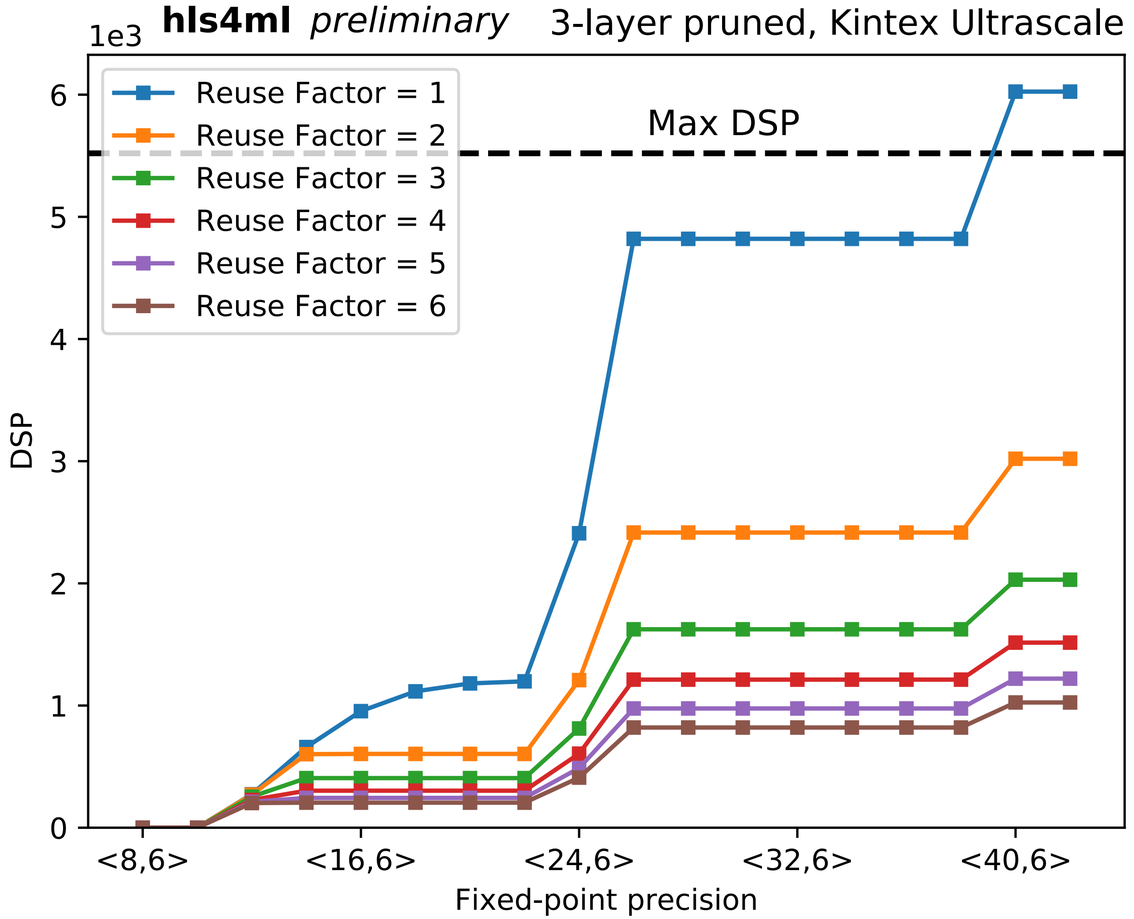






Performances





NB: FPGA emulator over-estimates resource needs by a factor 2-4 (tested our emulation vs actual deployment)

Conclusions

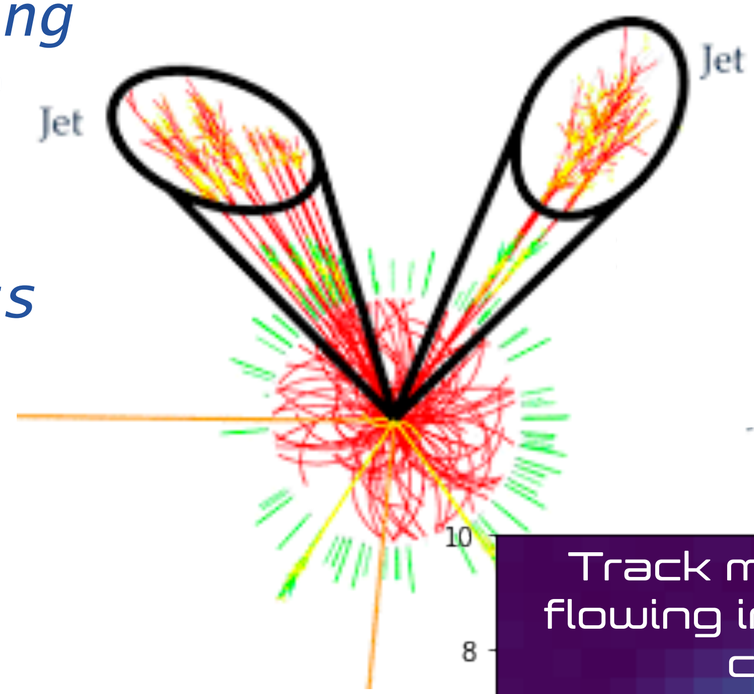
- *Modern Machine Learning is becoming a crucial ingredient to HEP, mainly due to the challenges of High-Luminosity LHC*
- *The clean e^+e^- environment might not necessarily call for advanced ML applications. Nevertheless*
  - *A clear advantage was prove in similar situations (neutrinos)*
  - *This is where the field is going. Why not profiting?*
- *Applications span across many typical HEP workflows*
  - *PID, energy measurement, simulation, DAQ/Trigger*
- *Work started, and a long way to go to meet very-high precision requirements of an intense e^+e^- collider*

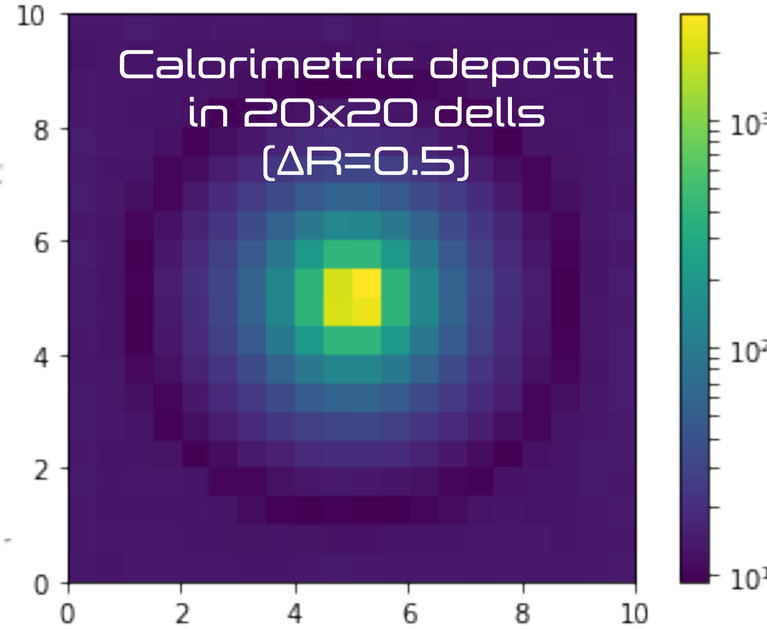
European Research Council

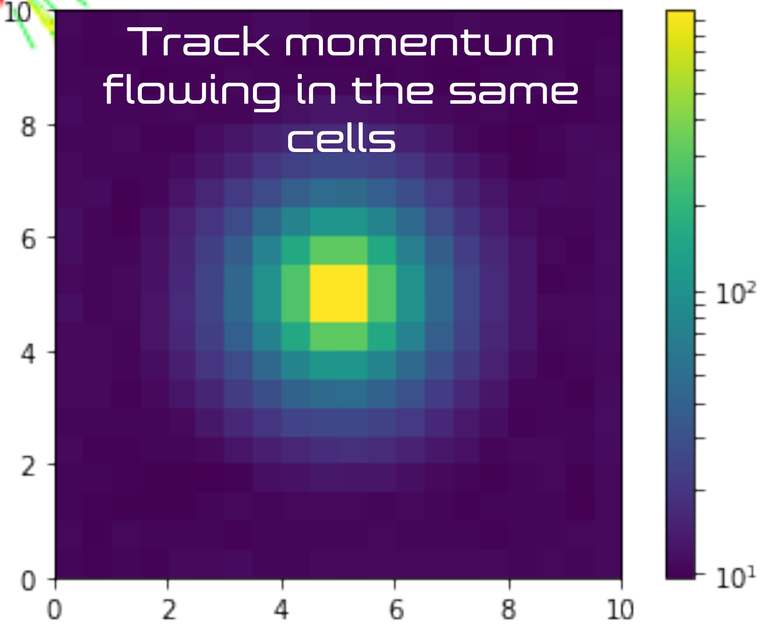
Backup

Faster/better Jet Reco

- *Offline, reconstructed from full particles using Particle Flow (accurate but slow)*
- *At L1, low-level objects are used*
  - *map of energy deposit on calorimeter*
  - *(in the future) local tracking (maybe pixel only)*
- ***ML could make jet reco at trigger faster & better***



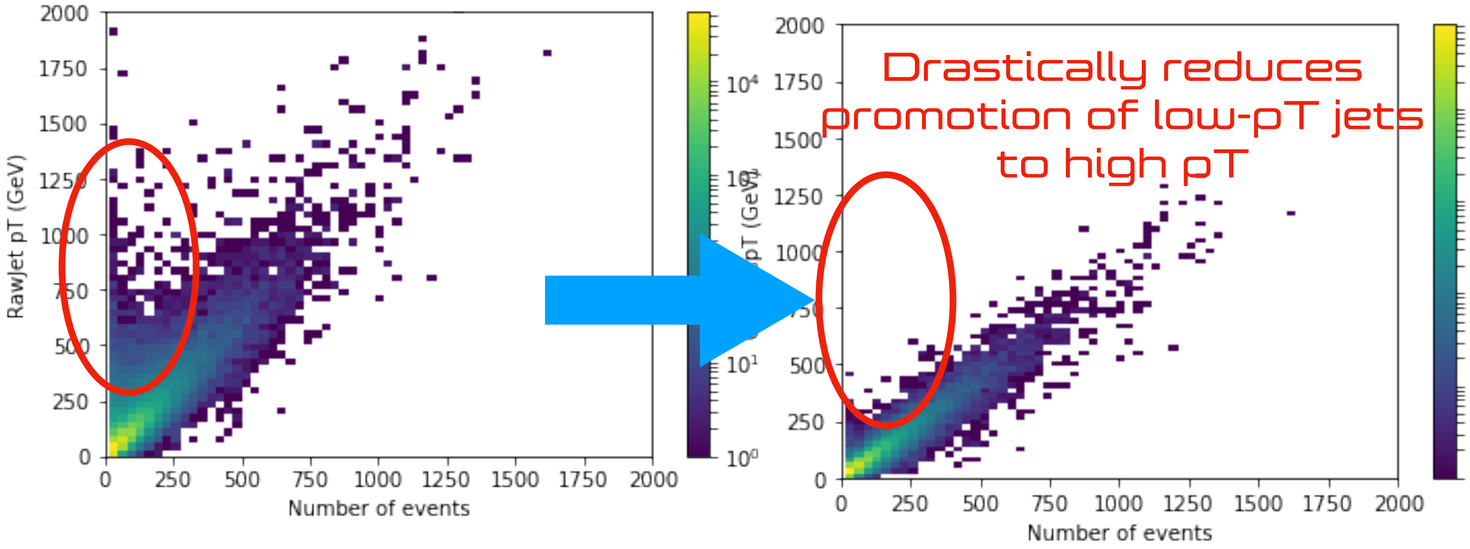


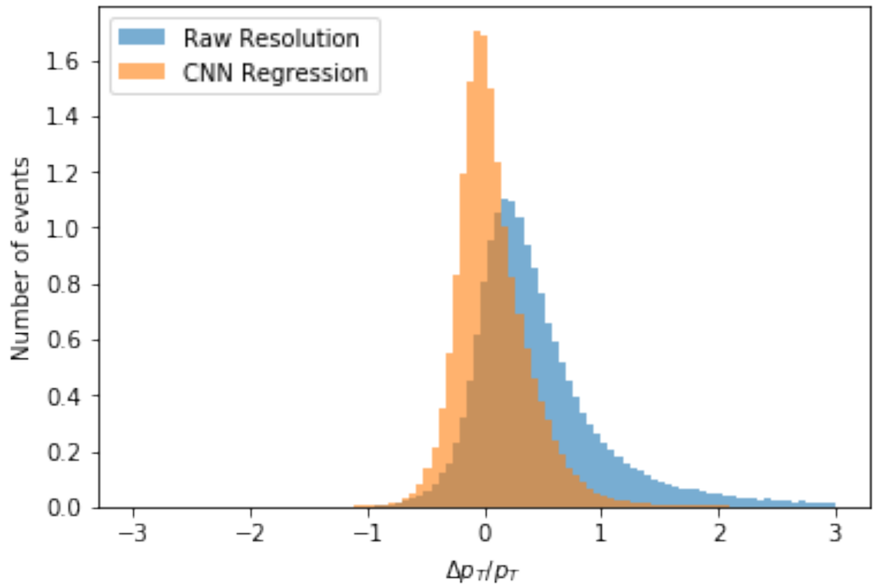


European Research Council

Faster/better Jet Reco

- *With a simple network, substantial improvement observed (could be pushed further, optimizing the network architecture)*
- *Consequences downstream:*
  - *better select the 100000 events to write /sec*
  - *With better input, HLT could work even better*
  - *We could recover lost territory/gain new territory and extend our search capability*

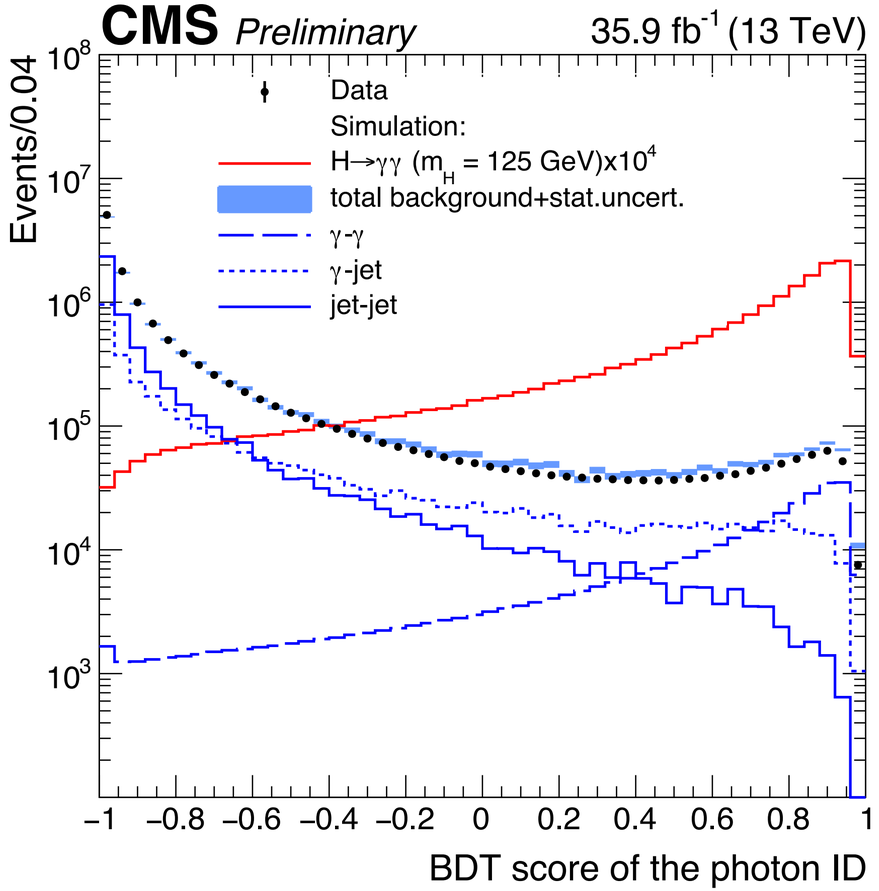




RAW resolution: 125%
CNN resolution: 36%

What we do with ML today

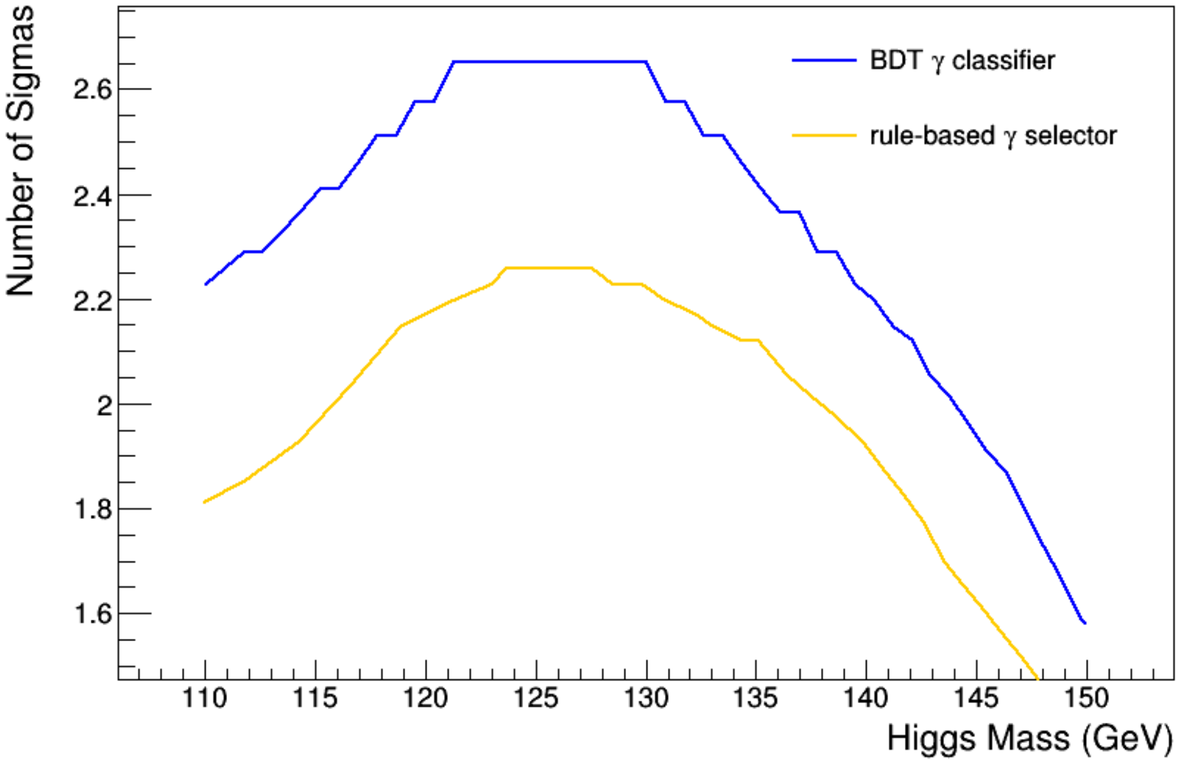
- *Classification:*
  - *identify a particle & reject fakes*
  - *identify signal events & reject background*
- *Regression:*
  - *Measure energy of a particle*
- *We typically use BDTs for these task*
  - *moved to Deep Learning for analysis-specific tasks*
  - *same will happen for centralised tasks (eventually)*

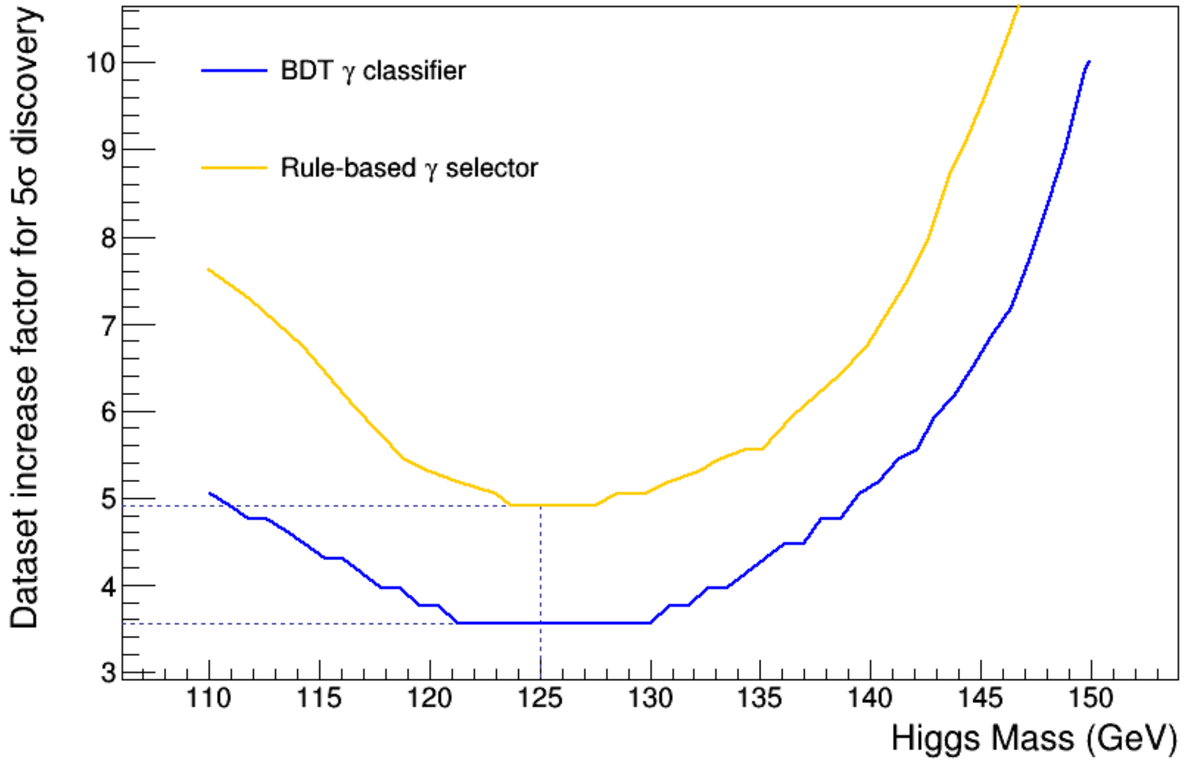


Centralised task (in online or offline reconstruction)
Analysis-specific task (by users on local computing infrastructures)

Example: ML for Higgs discovery

- *We were not supposed to discover the Higgs boson as early as 2012*
- *Given how the machine progressed, we expected discovery by end 2015 /mid 2016*
- *We made it earlier thanks (also) to Machine Learning*





HL4ML: FPGA details

Xilinx Vivado 2017.2

Results are slightly different in other versions of Vivado

e.g. 2016.4 optimization is less performant for Xilinx ultrascale FPGAs

Clock frequency: 200 MHz

Latency results can vary (~10%) with different clock choices

FPGA: Xilinx Kintex Ultrascale (XCKU115-FLVB2104)

Results are slightly different in other FPGAs

e.g. Virtex-7 FPGAs are slightly differently optimized

European Research Council