DEEP LEARNING

EPISODE II

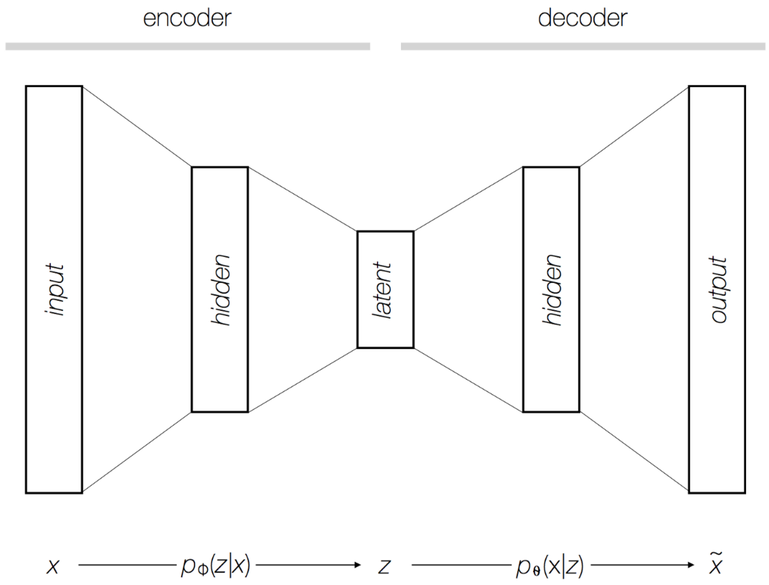
Maurizio Pierini

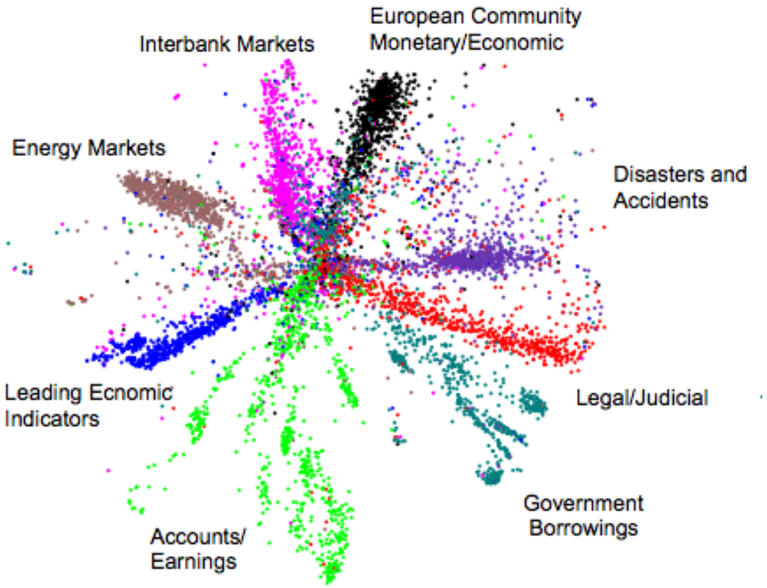
Anomaly Detection With Autoencoders

European Research Council

Autoencoders in a nutshell

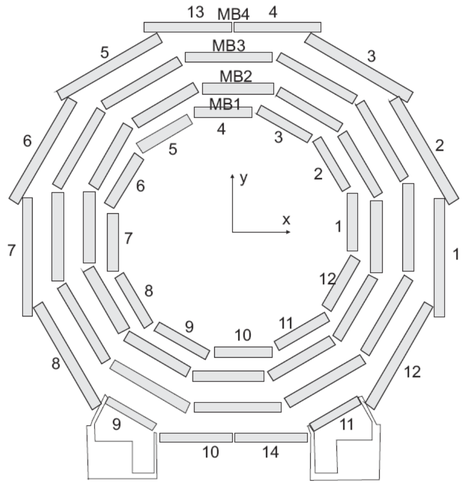
- *Autoencoders are compression-decompression algorithms that learn to describe a given dataset in terms of points in a lower-dimension latent space*
- *UNSUPERVISED algorithm, used for data compression, generation, clustering (replacing PCA), etc.*
- *Used in particular for anomaly detection: when applied on events of different kind, compression-decompression tuned on refer sample might fail*
- *One can define anomalous any event whose decompressed output is "far" from the input, in some metric (e.g., the metric of the auto-encoder loss)*



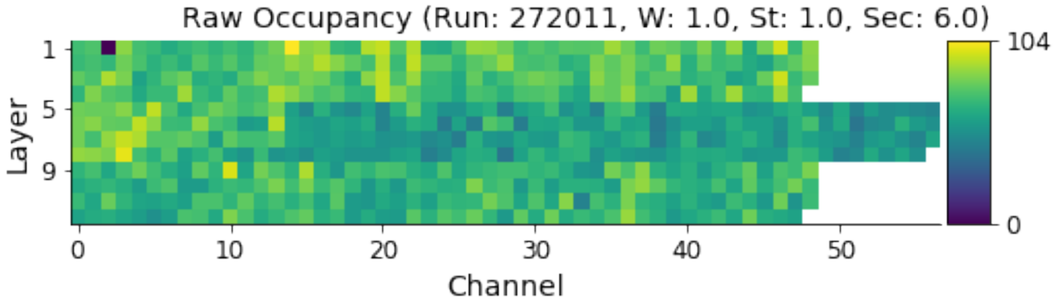


Data Quality Monitoring

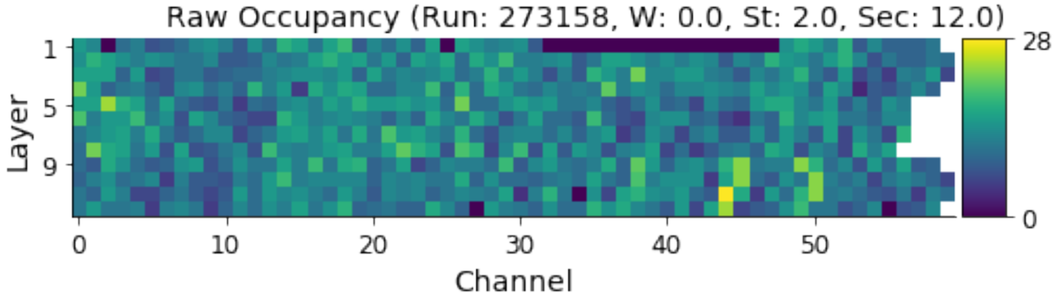
- *When taking data, >1 person watches for anomalies in the detector 24/7*
- *At this stage no global processing of the event*
- *Instead, local information from detector components available (e.g., detector occupancy in a certain time window)*



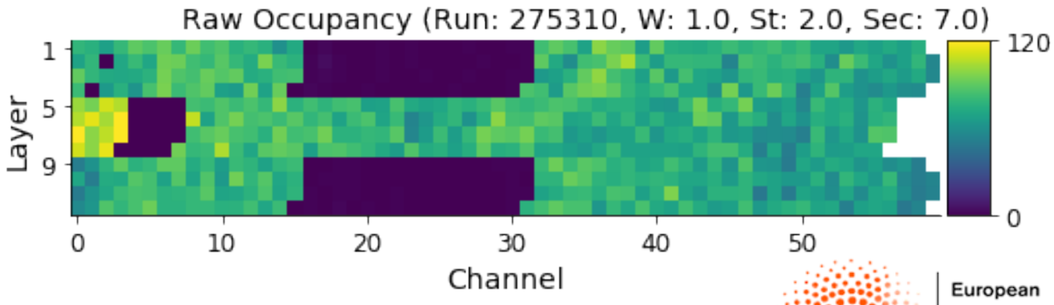
A



B



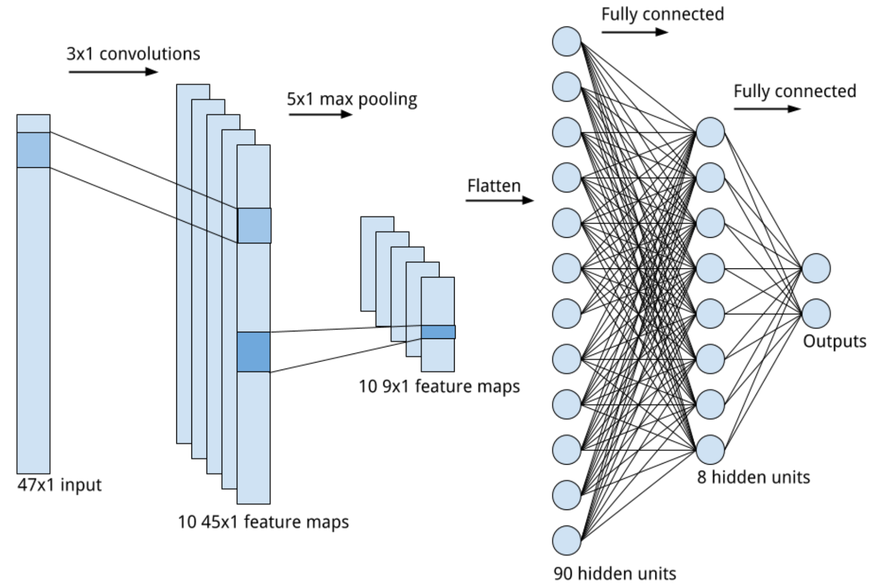
C

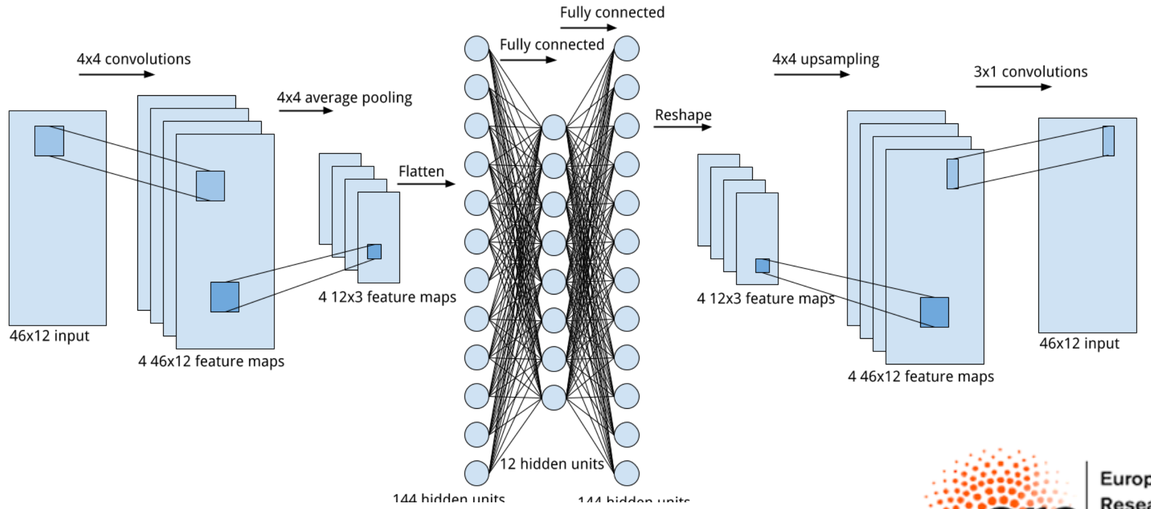


Two approaches

- *Given the nature of these data, ConvNN are a natural analysis tool. Two approaches pursued*
  - *Classify good vs bad data. Works if failure mode is known*
  - *Use autoencoders to assess data "typicality". Generalises to unknown failure modes*

A. Pol et al., to appear soon

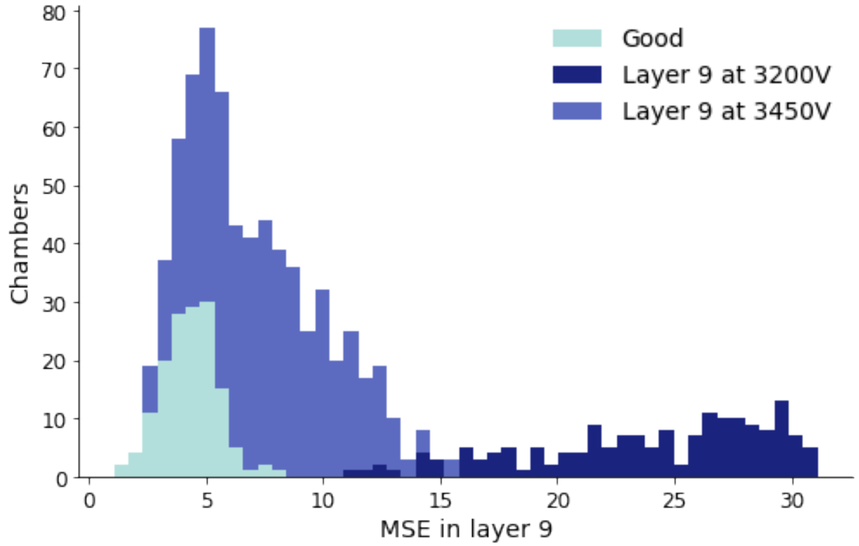


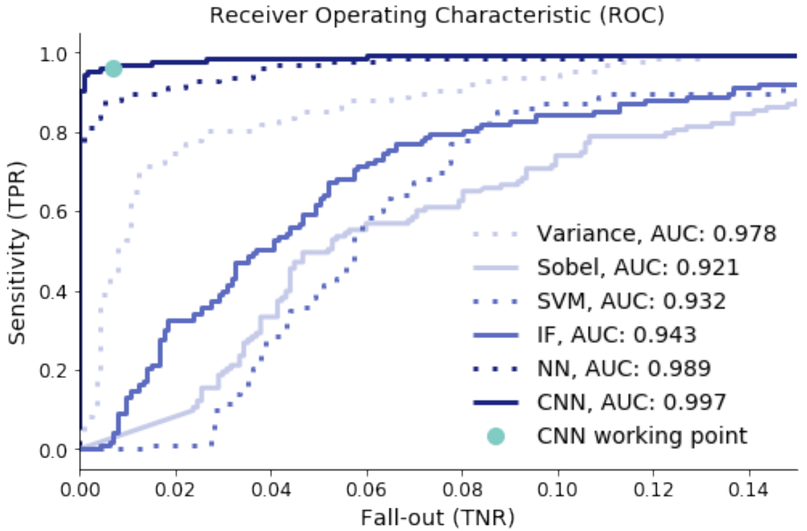


Two approaches

- *Given the nature of these data, ConvNN are a natural analysis tool. Two approaches pursued*
  - *Classify good vs bad data. Works if failure mode is known*
  - *Use autoencoders to assess data "typicality". Generalises to unknown failure modes*

A. Pol et al., to appear soon

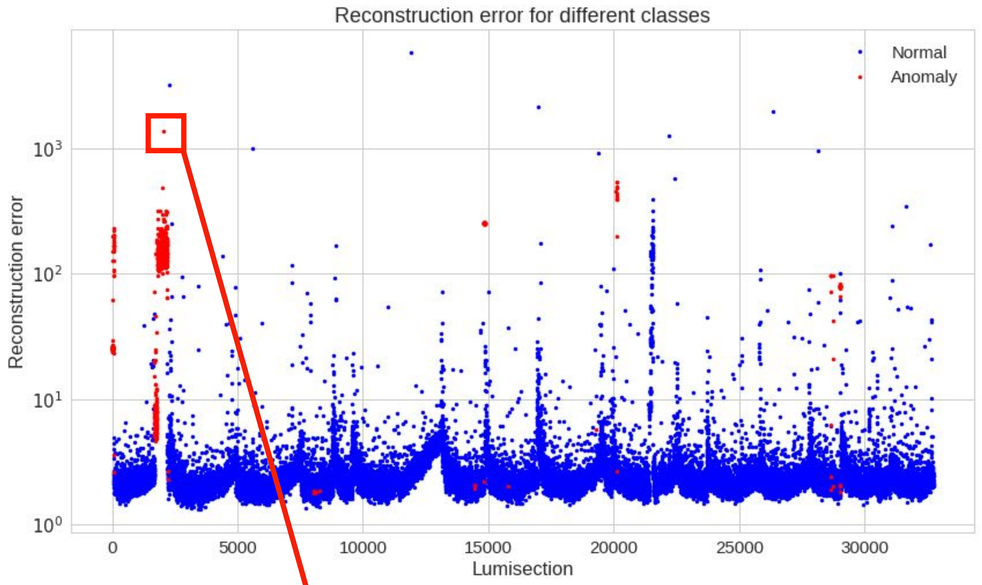


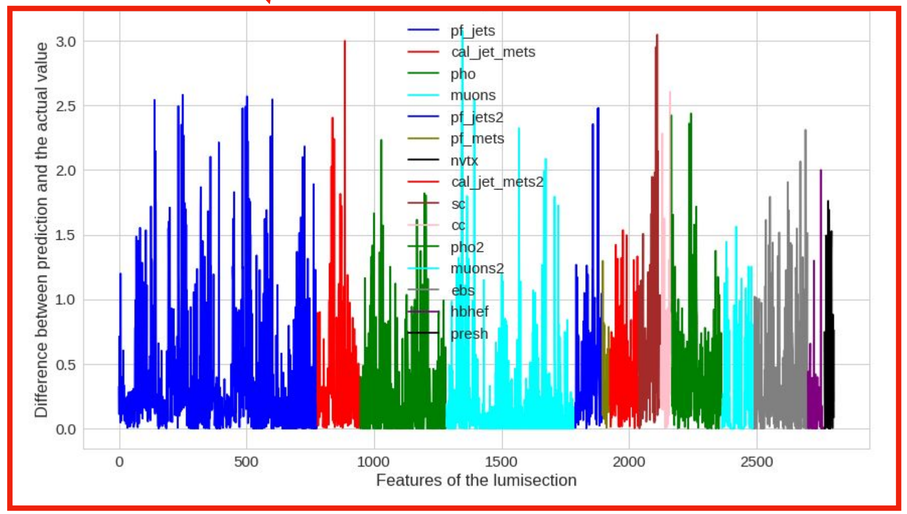


European Research Council

Data Quality Certification

- *Autoencoder-based 1-class approach generalises to later stages of quality assessment*
  - *after reconstruction of the events, event reconstruction allows a global assessment (w.g., looking at electrons, muons, etc rather than hits in the detector)*
  - *A global autoencoder can spot all these features*
  - *Monitoring individual contributions to loss function (e.g., MSE) one can track the problem back to a specific physics object/detector component*

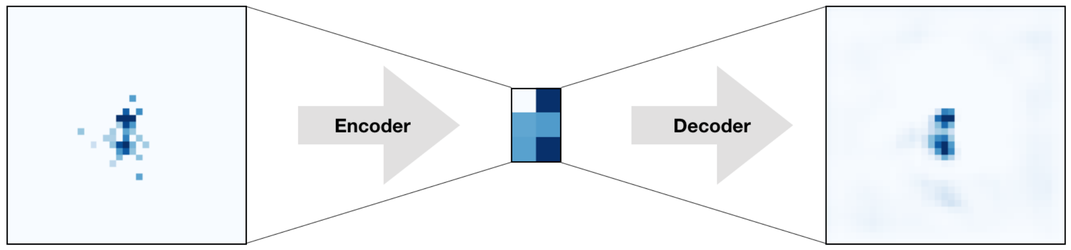


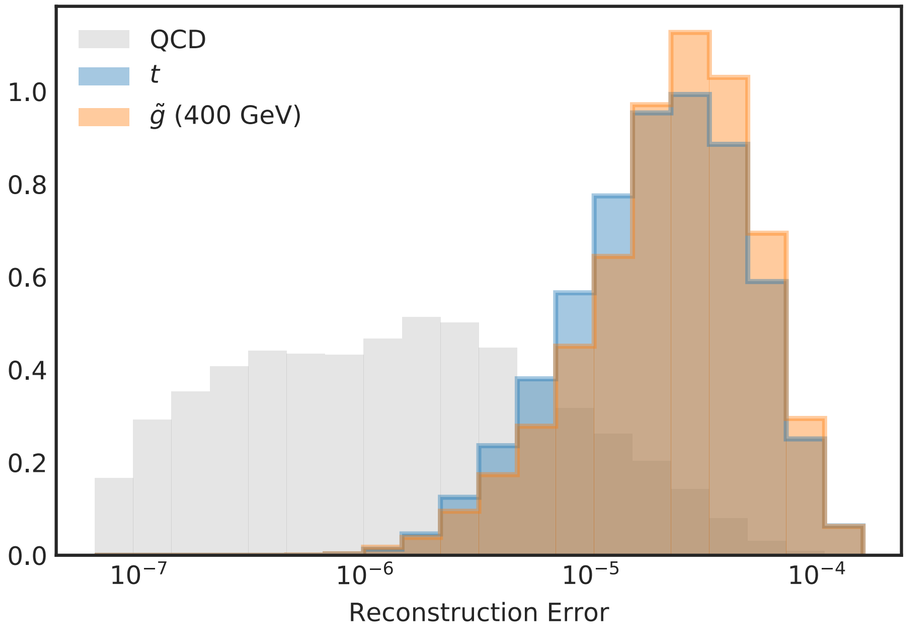


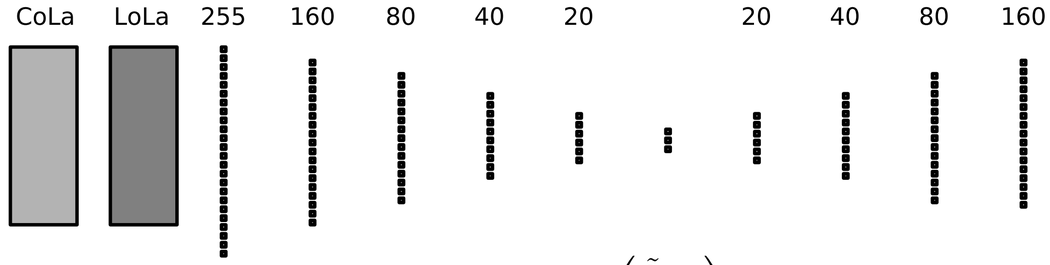
F. Široký et al., to appear sooner or later

Example: Jet autoencoders

- *Idea applied to tagging jets, in order to define a QCD-jet veto*
- *Applied in a BSM search (e.g., dijet resonance) could highlight new physics signal*
- *Based on image and physics-inspired representations of jets*







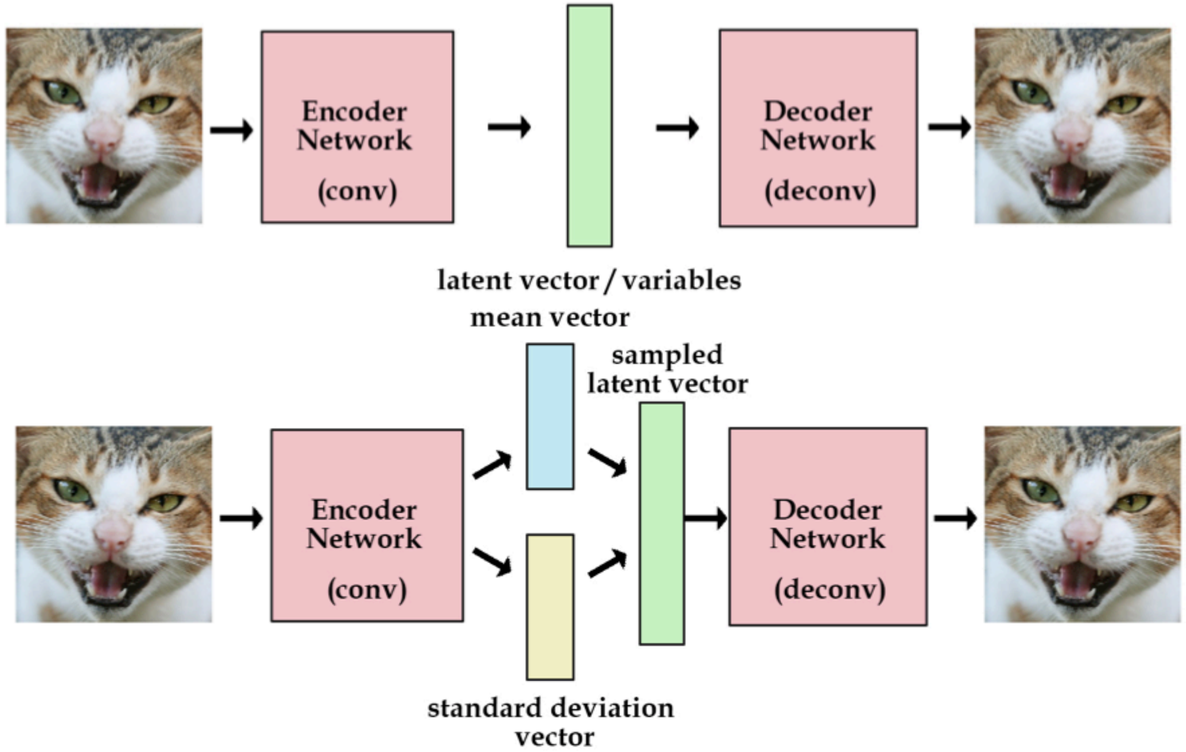
$$\tilde{k}_j = \begin{pmatrix} \tilde{k}_{0,j} \\ \tilde{k}_{1,j} \\ \tilde{k}_{2,j} \\ \tilde{k}_{3,j} \end{pmatrix} \xrightarrow{\text{LoLa}} \begin{pmatrix} \tilde{k}_{0,j} \\ \tilde{k}_{1,j} \\ \tilde{k}_{2,j} \\ \tilde{k}_{3,j} \\ \sqrt{\tilde{k}_j^2} \end{pmatrix} .$$

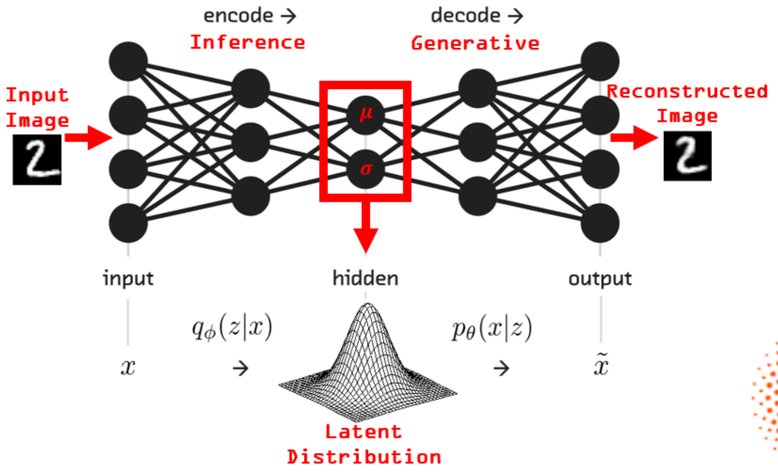
Farina et al., arXiv:1808.08992

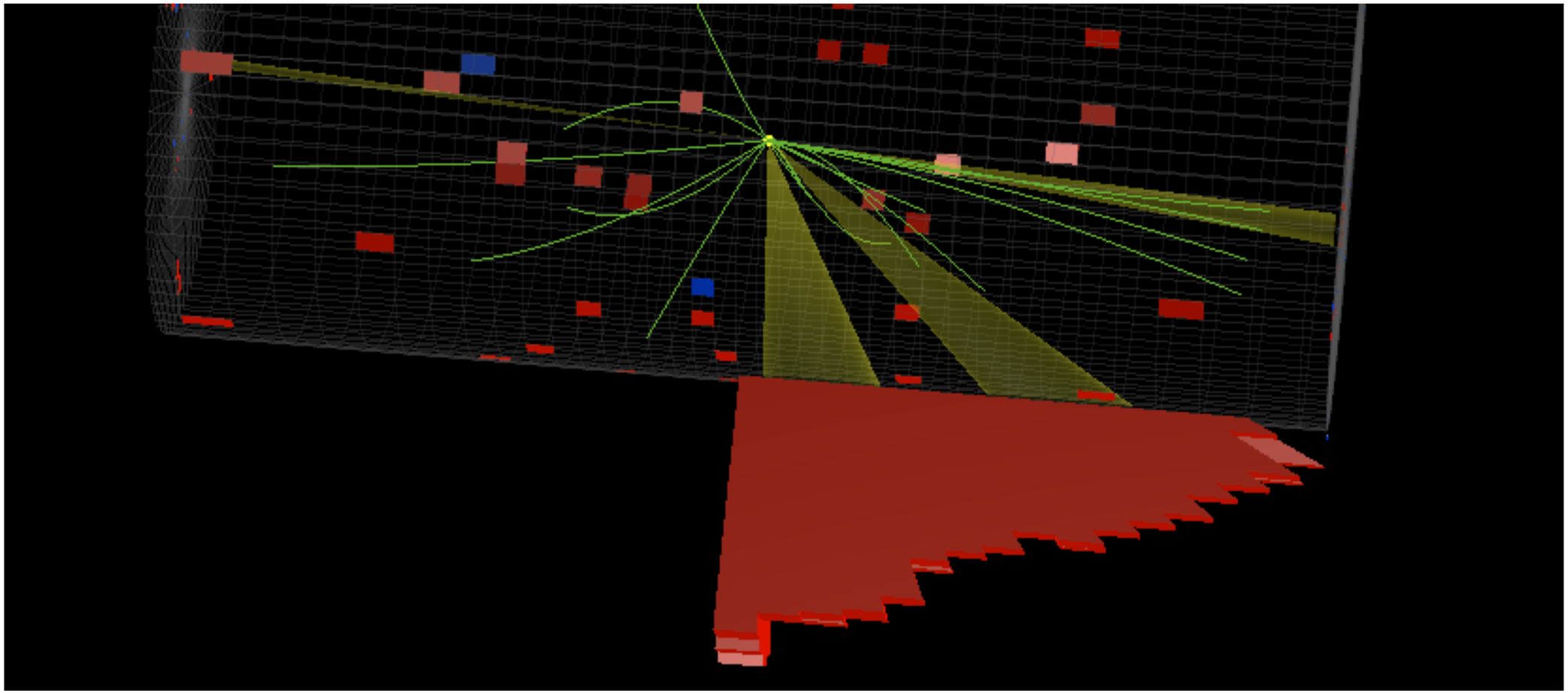
Heimel et al., arXiv:1808.08979

Variational Autoencoders

- We investigated variational autoencoders
- Unlike traditional AEs, VAEs try to associate a multi-Dim pdf to a given image
  - can be used to generate new examples
  - comes with a probabilistic description of the input
  - tends to work better than traditional AEs





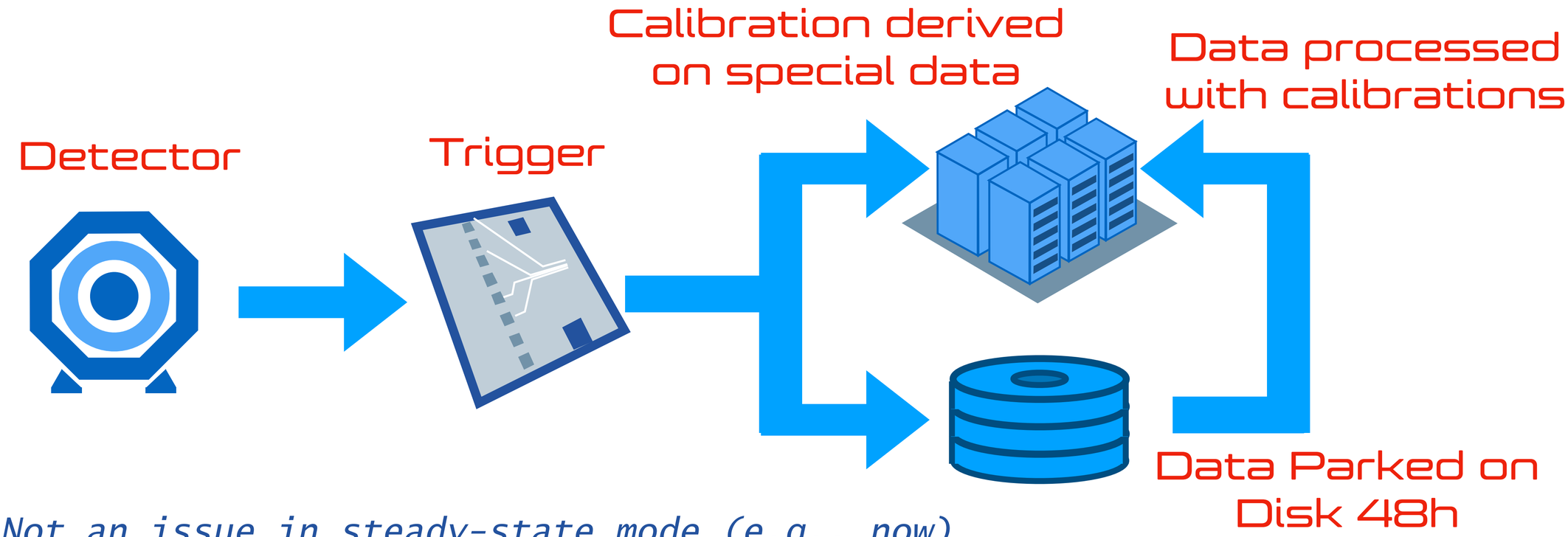


Anomaly detection at LHC Run 1

European Research Council

The CMS data-flow

- *Traditionally, data take 2-7 days before becoming available for analysis*



  - *Not an issue in steady-state mode (e.g., now)*
  - *A substantial delay at startup, particularly if you hope for an early discovery*
- *This is why we implemented in 2009 an alarm system for special physics events*

The Exotica Hotline

- *Back in 2009, we implemented a set of triggers to catch rare (and possibly interesting events)*
  - *high-Pt jets, muons, electrons, photons or taus*
  - *large lepton multiplicity*
  - *large di-object invariant mass*
- *Stored O(10) events/day, processed in real time*
- *Studied by experts (visual inspection of event displays)*

- Jets
  - at least 1 jet with p_T > 200 GeV [updated from 150
  - at least 5 jets with pT > 40 GeV
- MET/HT
  - MET > 250 GeV
  - HT > 300 GeV , calculated summing jets with pT > 30 GeV
- Electrons
  - at least 1 electron with pT > 100 GeV
  - at least 2 electrons with pT > 15 GeV
- Photons
  - at least 1 photon with pT > 100 GeV
  - at least 3 photons with pT > 20 GeV
- Muons
  - at least 1 muon with pT > 100 GeV
  - at least 2 muons with pT > 15 GeV
- Tracks
  - at least 600 tracks
- dE/dX
  - at least 1 track with pT > 50 GeV, dE/dx > 5.6 MeV/cm

European Research Council

The ringing bell...

- *The system was deployed in 2010, with a real-time alert system and a team of expert scanners*
- *It got some attention back then*
- *It was actually very effective in discovering something (which unfortunately was not new physics)*

CERN Bulletin

News Articles | Official News | Training | General Information | Staff Association

english | français

search

Issue No. 46-47/2010 - Monday 15 November 2010

Printable version - Subscribe:

"See you soon" – a retrospective on the first LHC proton run

A milestone in safety

The Latest from the LHC: Successful switch to ions

ATLAS: Exceeding all expectations

CMS: Beyond all possible expectations

ALICE: The best is yet to come

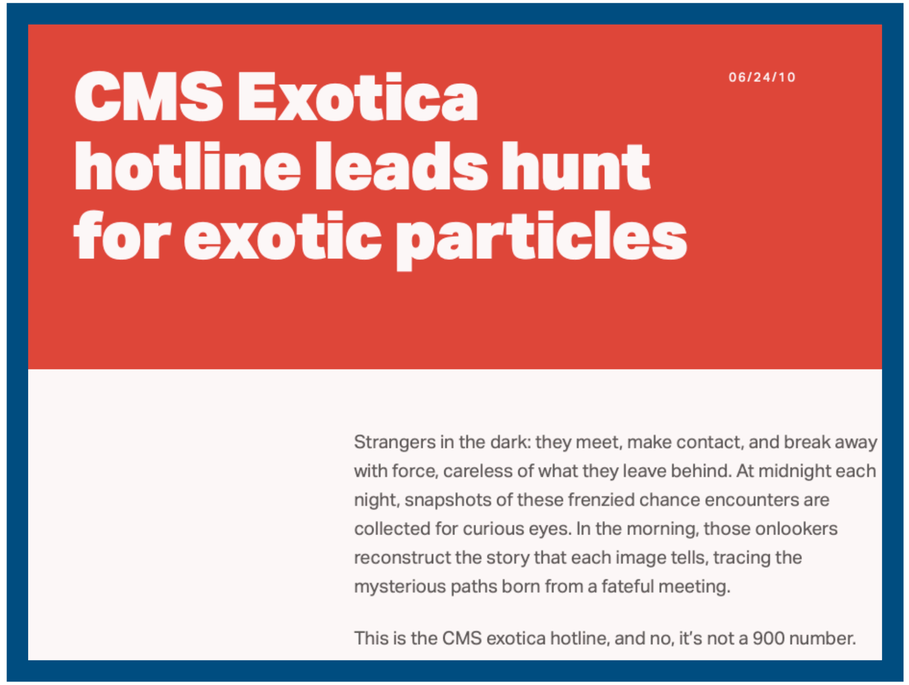
LHCb: Not just a precision experiment but also a detector ready for discoveries

TOTEM: Thousands of interesting events

IS THE BELL RINGING?

During the Nobel prize-winning UA1 experiment, scientists in the control room used to ring a bell if a particularly interesting event had occurred. Today, the "*CMS Exotica hotline*" routine produces a daily report that lists the exotic events that were recorded the day before.

Take just a very small fraction of the available data (max. 5%); define the events that you want to keep and set the parameters accordingly; run the *Exotica* routine and only look at the very few images



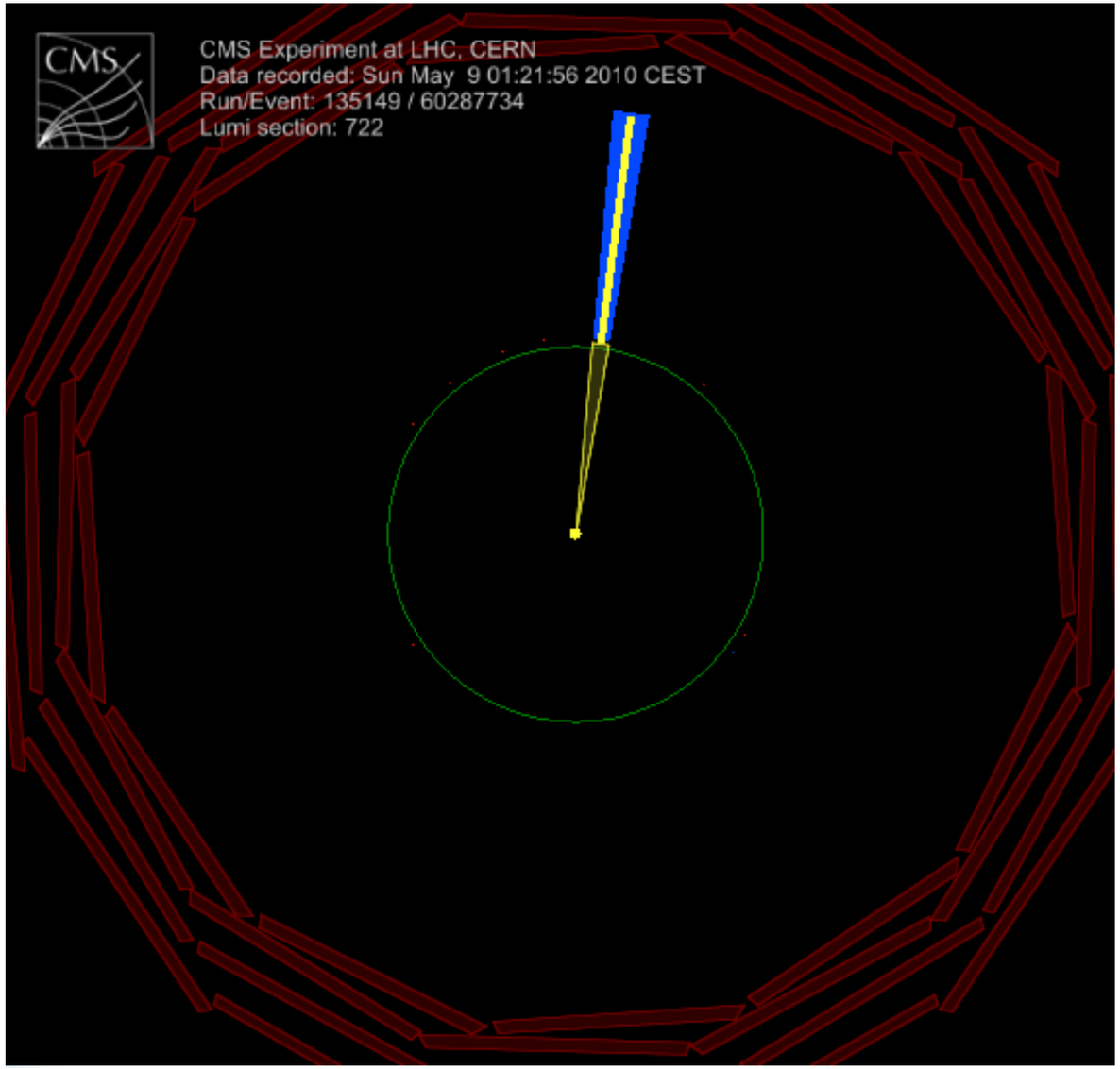
06/24/10

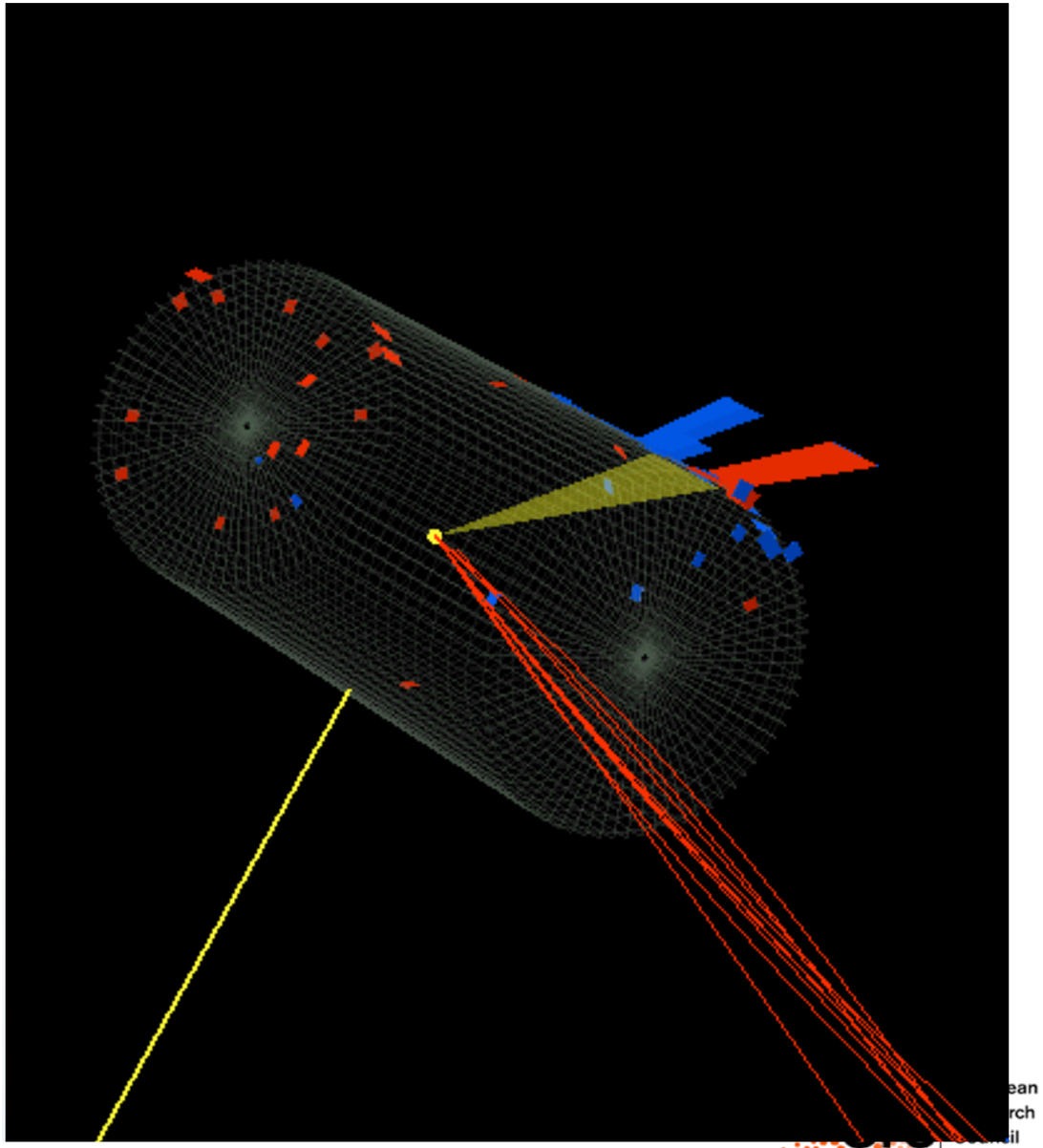
CMS Exotica hotline leads hunt for exotic particles

Strangers in the dark: they meet, make contact, and break away with force, careless of what they leave behind. At midnight each night, snapshots of these frenzied chance encounters are collected for curious eyes. In the morning, those onlookers reconstruct the story that each image tells, tracing the mysterious paths born from a fateful meeting.

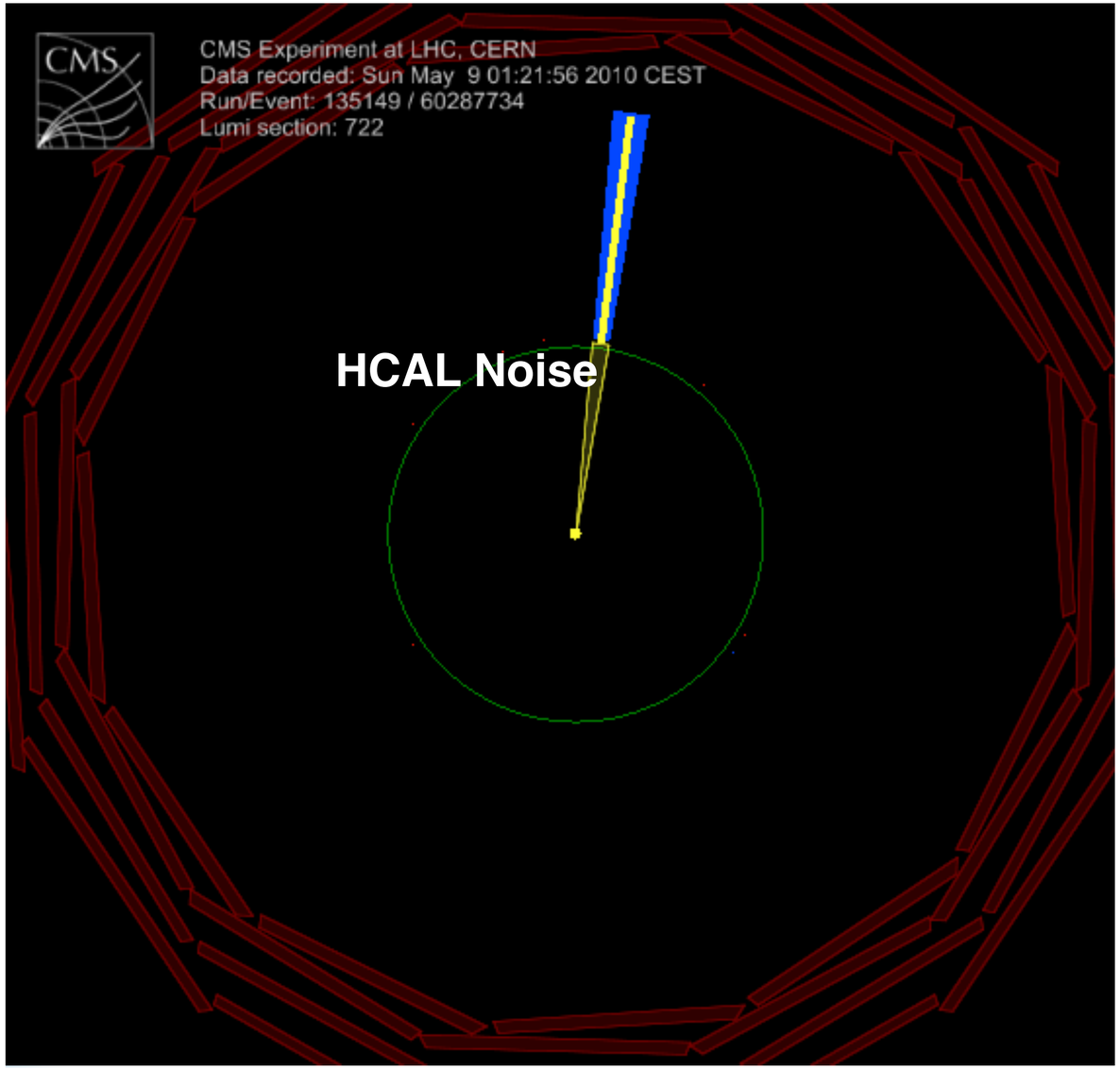
This is the CMS exotica hotline, and no, it's not a 900 number.

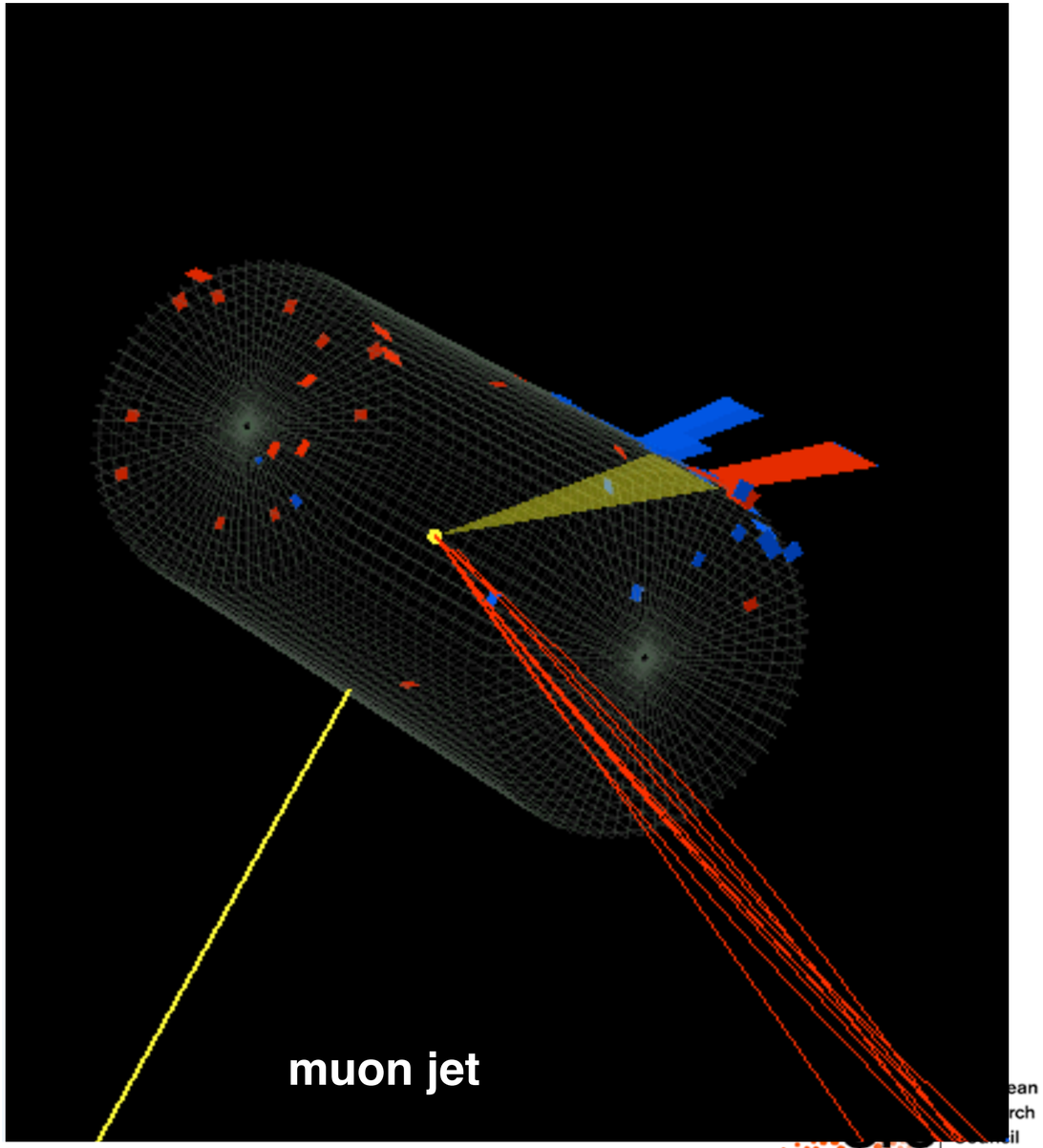
What was “found”



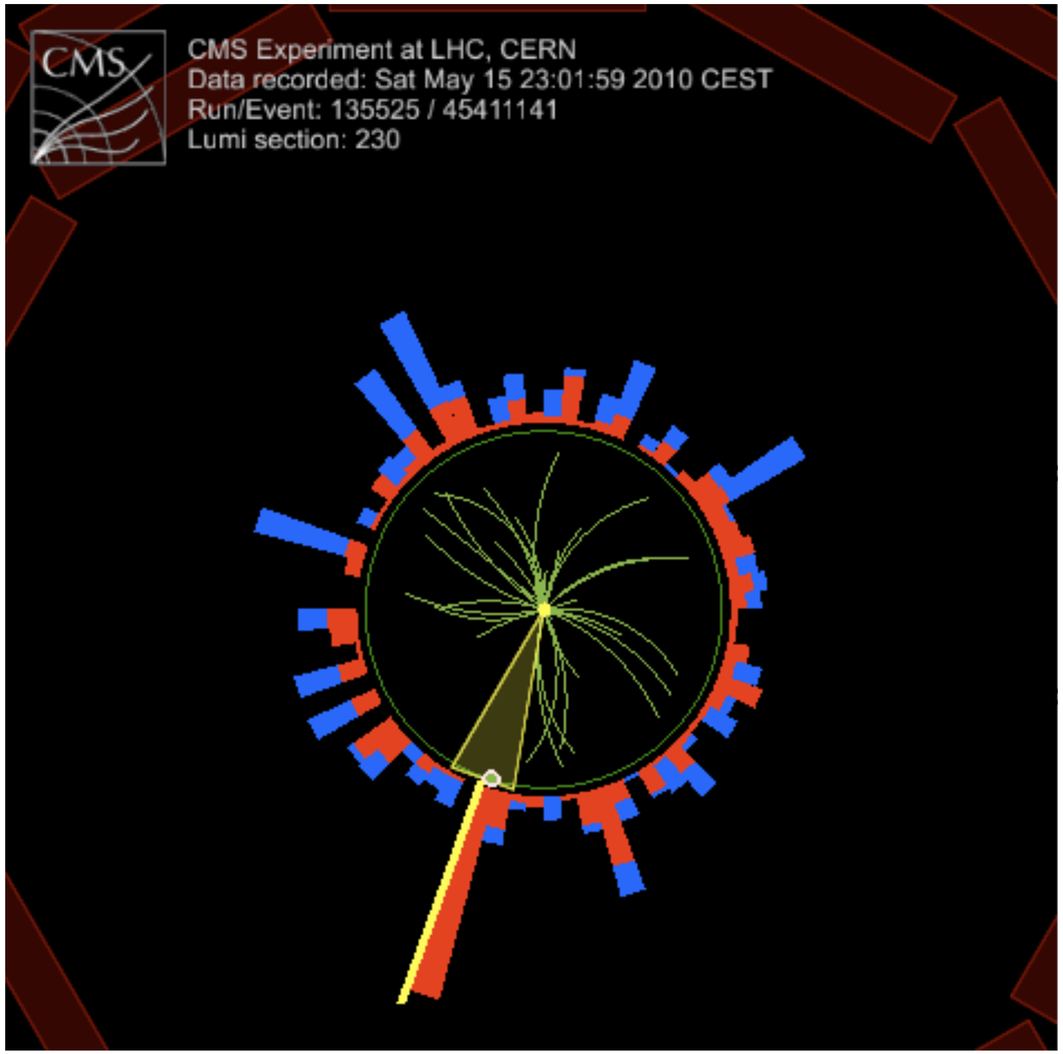


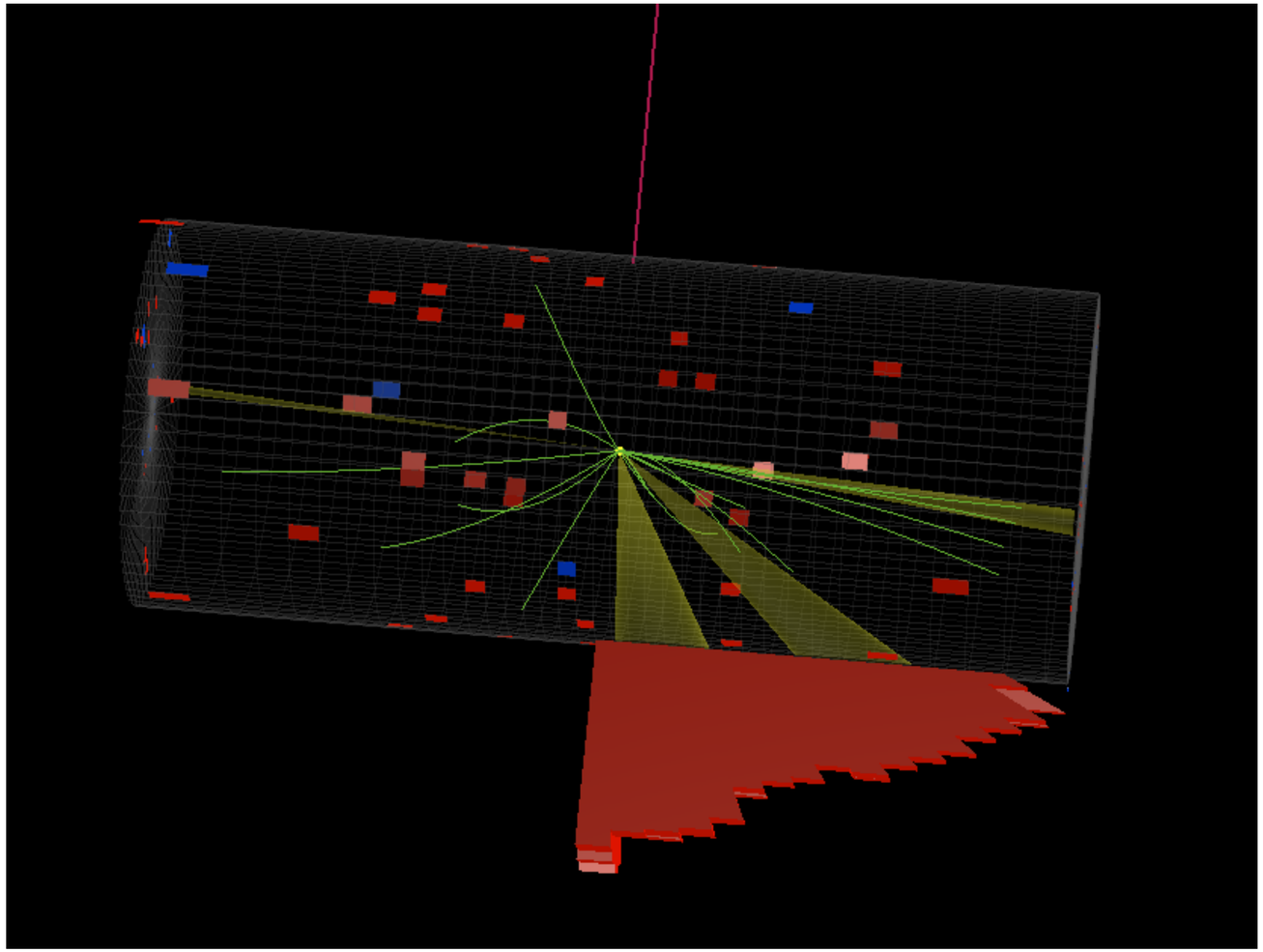
What was “found”



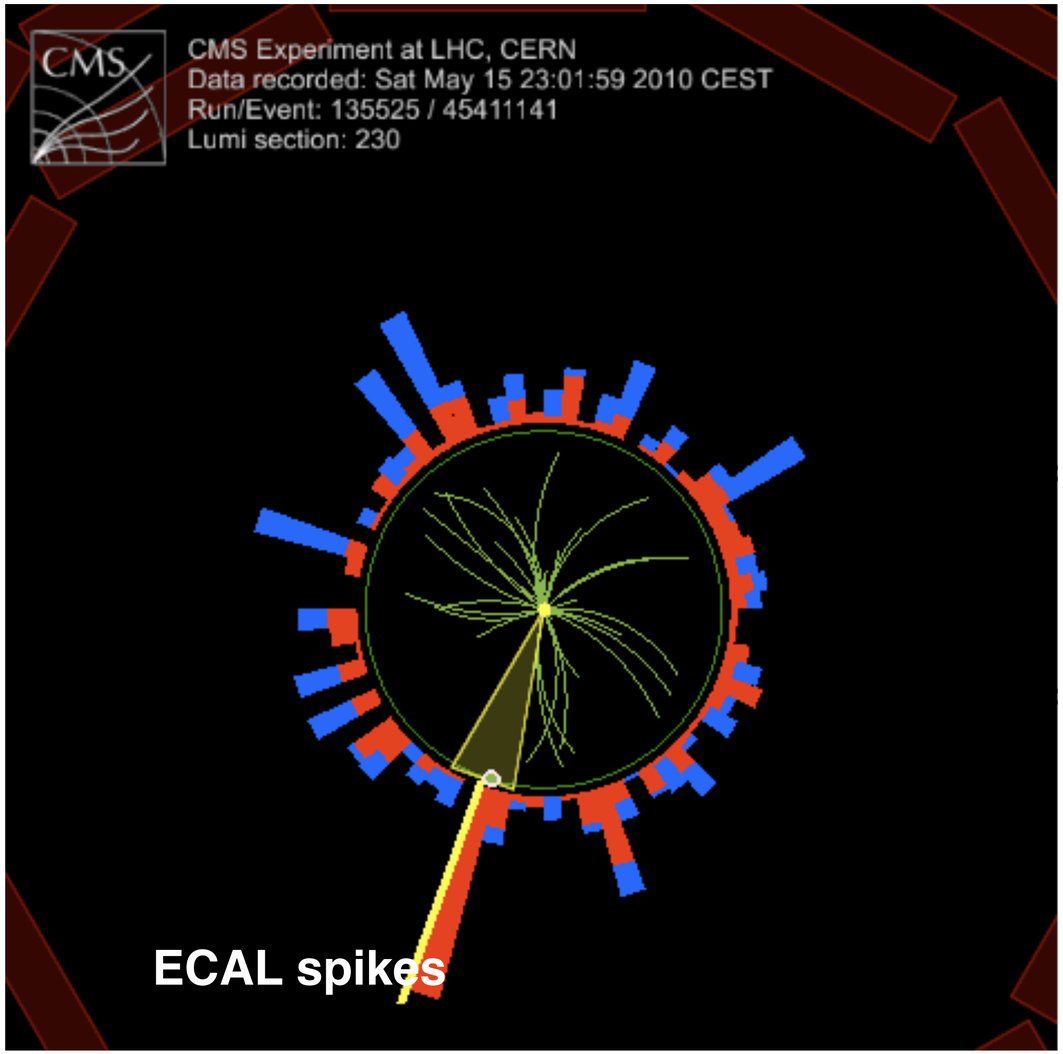


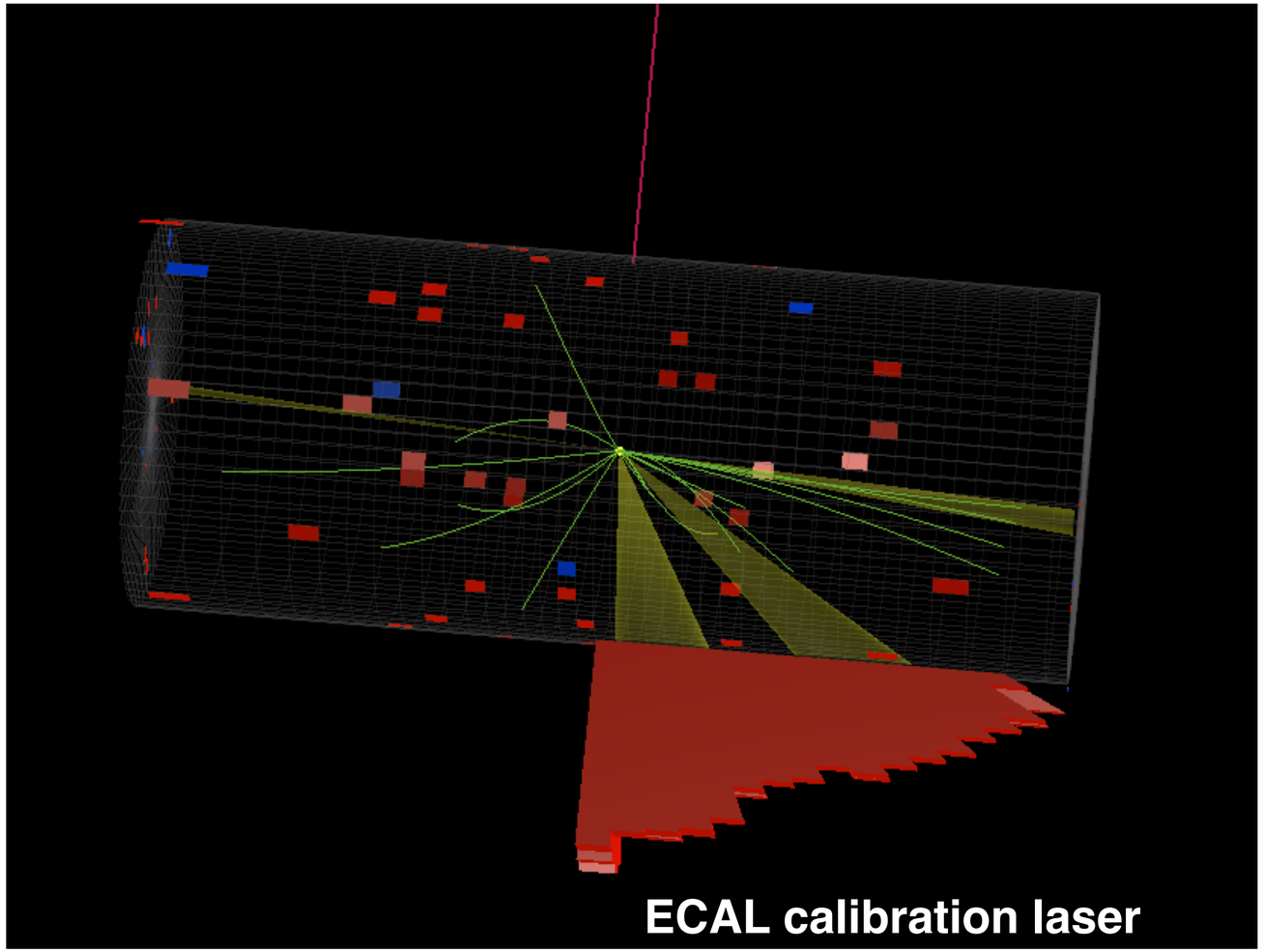
What was “found”



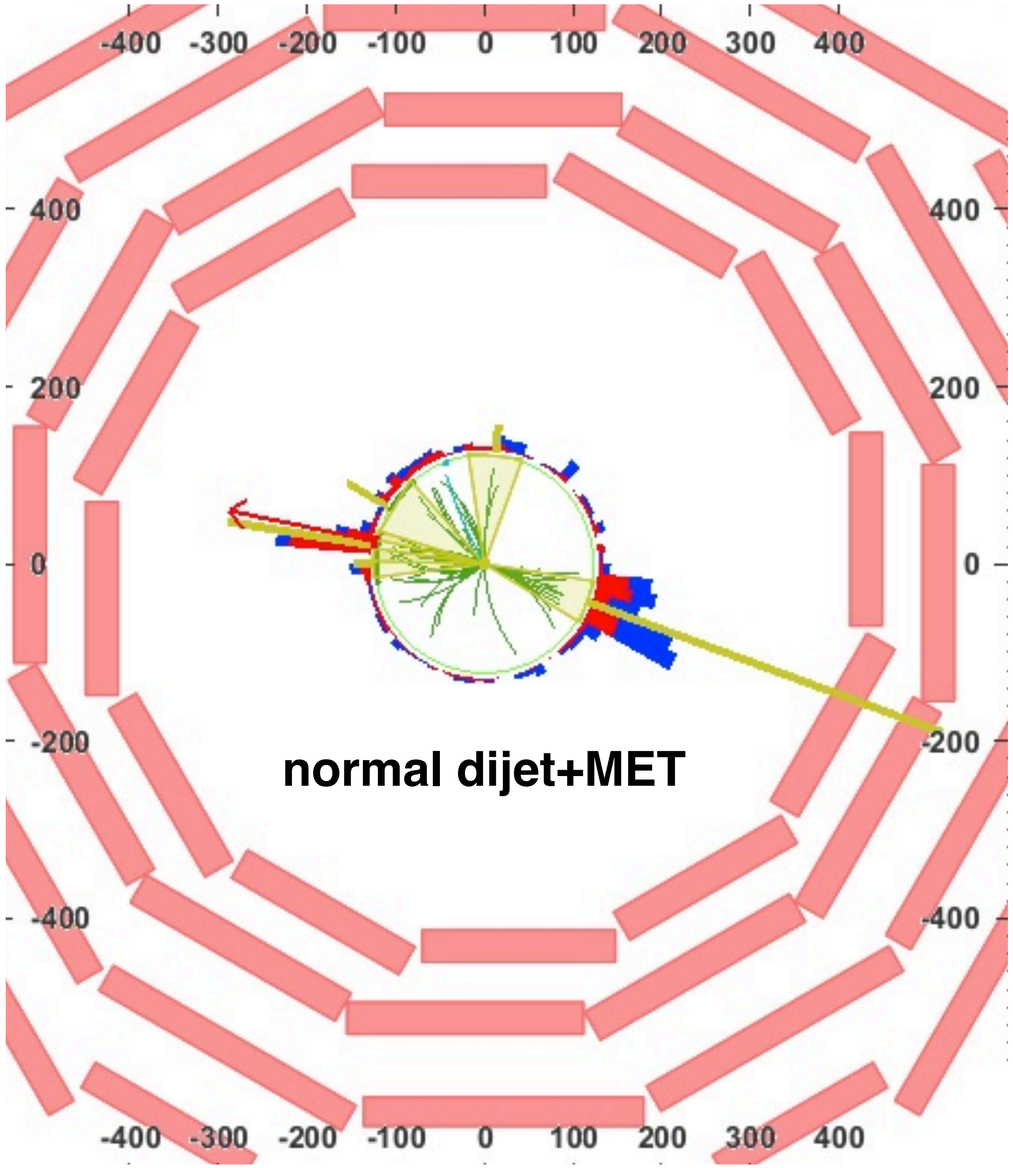


What was “found”



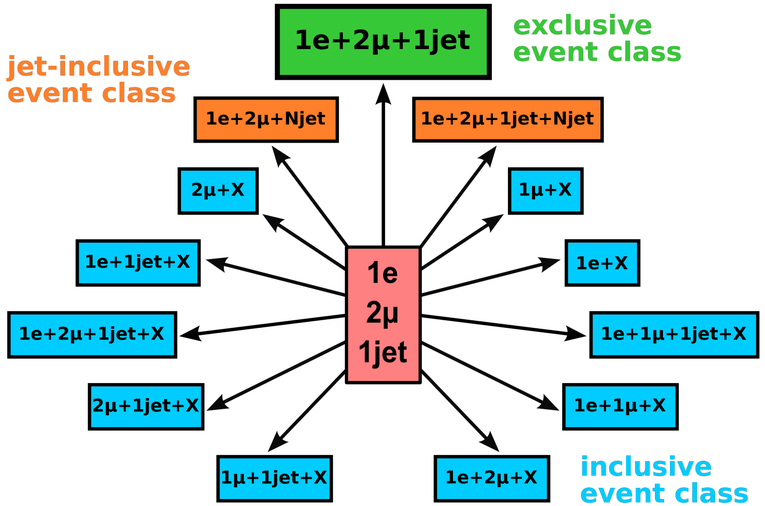


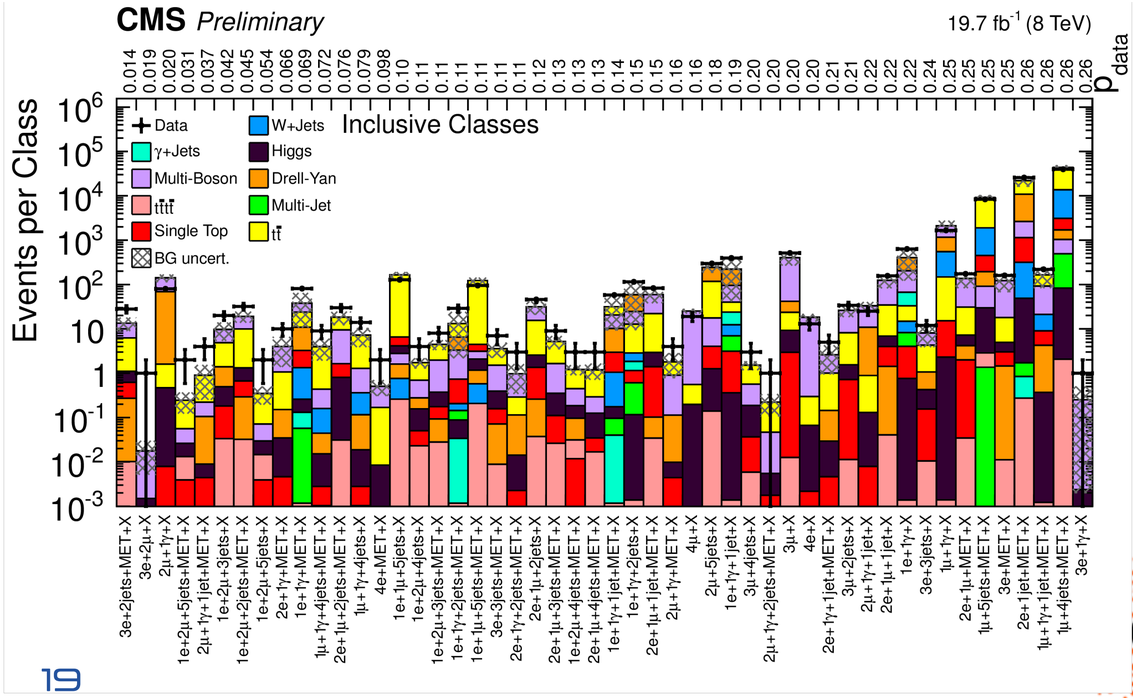
What was “found”



What we do today

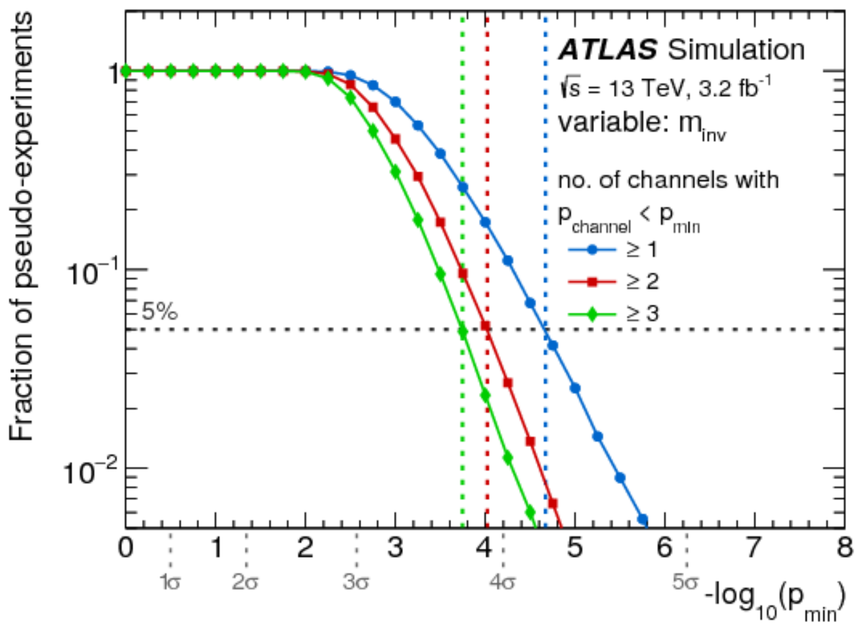
- *Model independent analyses are performed at colliders since Tevatron*
- *Plot a lot of histograms for data and compare them to what you expect on Monte Carlo*
- *Look for a discrepancy*
- *If you find it, try to exclude any instruments-driven explanation*

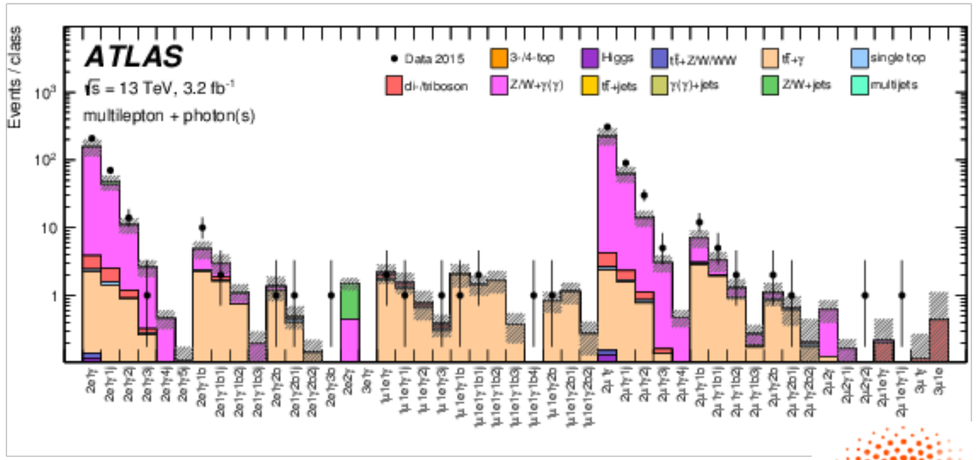




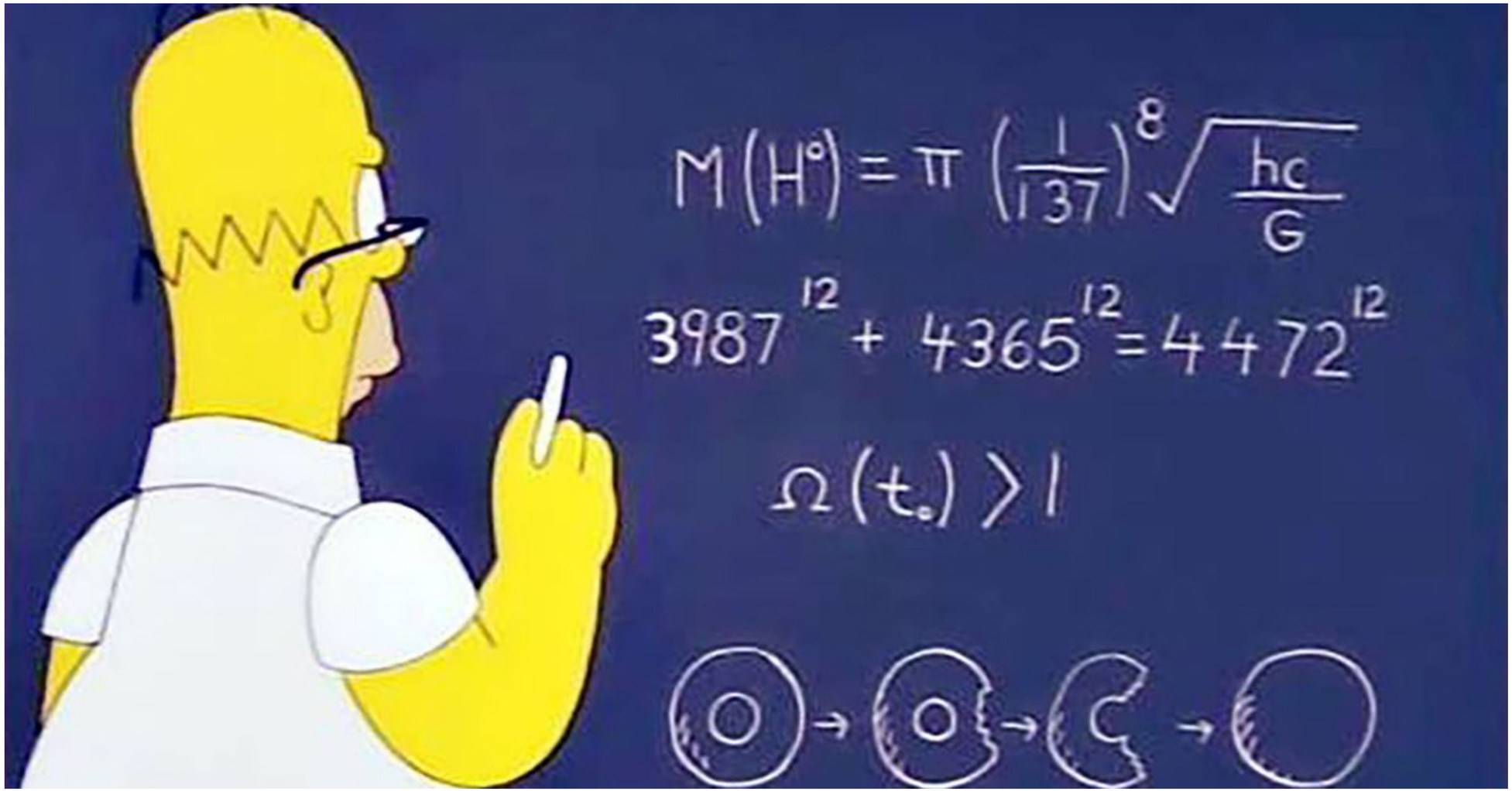
Trial factor

- *The issue with this approach is the trial factor (look-elsewhere effects)*
  - *one sets a p-value threshold α (e.g., 5%) to define the alarm*
  - *a fraction ~ α of the bins will be off even in absence of an anomaly*
  - *for large number of bins, this dilutes discovery power*
- *ATLAS came out with a proposal: use the analysis to identify an excess, but establish the significance with a traditional method on an independent dataset*
- *This is the same spirit we have in mind for what follows*





CERN-EP-2018-070



New Physics Anomalies

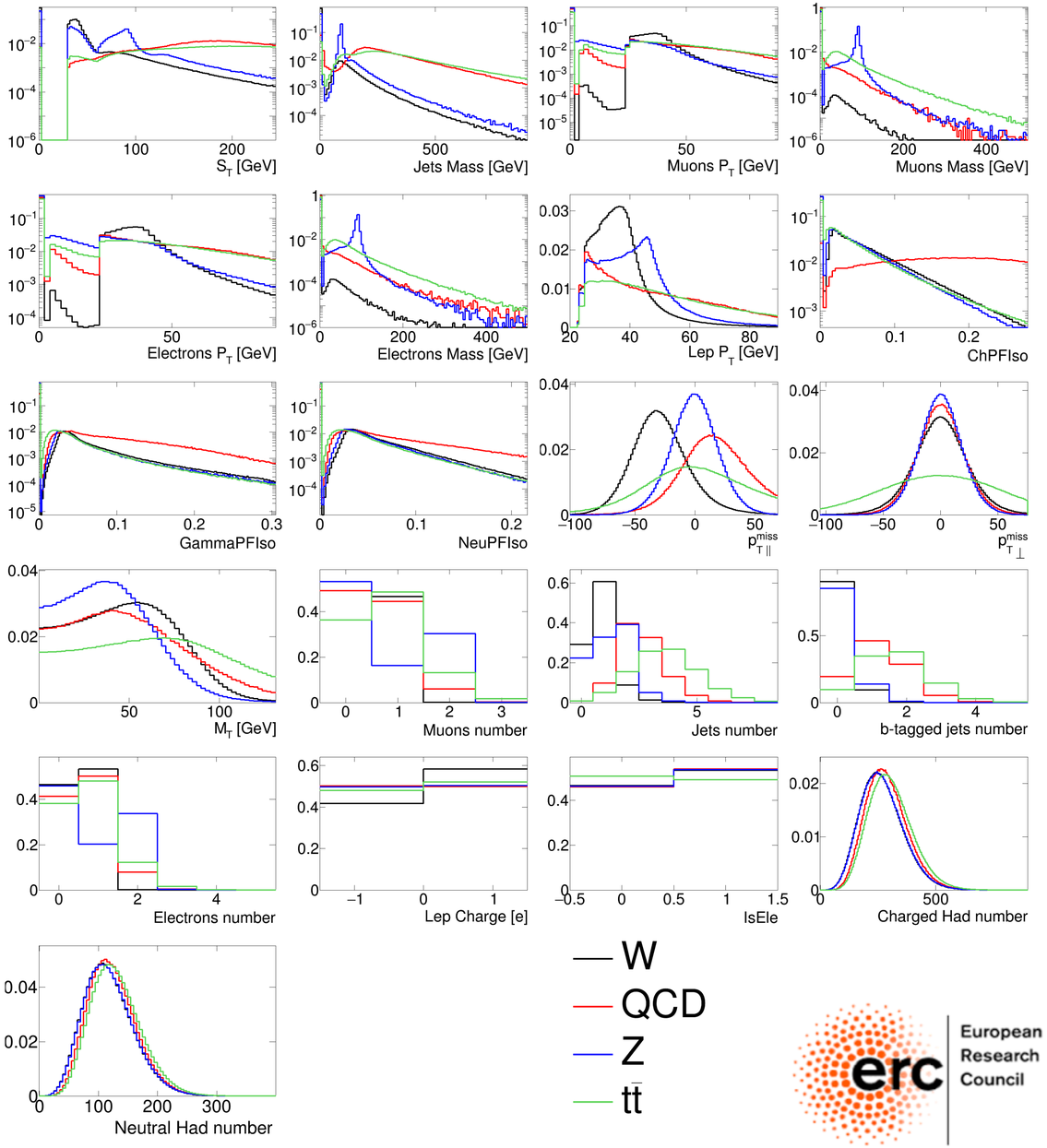
European Research Council

Our use case: ℓ+X @HLT

- Consider a stream of data coming from L1
  - Passed L1 because of 1 lepton (e,m) with pT>23 GeV
  - At HLT, very loose isolation applied
  - Sample mainly consists of W, Z, tt & QCD (for simplicity, we ignore the rest)

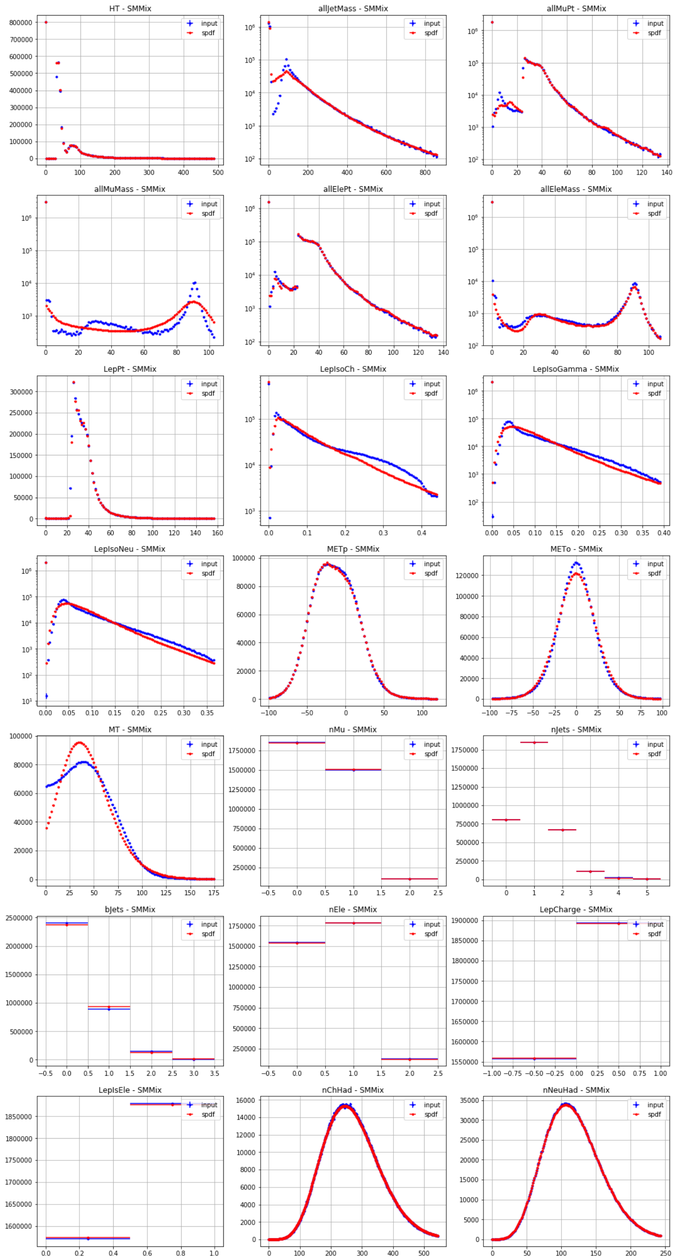
Standard Model processes					
Process	Acceptance	Trigger efficiency	Cross section [nb]	Events fraction	Event /month
W	55.6%	68%	58	59.2%	110M
QCD	0.08%	9.6%	$1.6 \cdot 10^5$	33.8%	63M
Z	16%	77%	20	6.7%	12M
$t\bar{t}$	37%	49%	0.7	0.3%	0.6M

  - We consider 21 features, typically highlighting the difference between these SM processes (no specific BSM signal in mind)



Standard Model encoding

- *First post-training check consists in verifying encoding-decoding capability, comparing input data to those generated sampling from decoder*
- *Reasonable agreement observed, with small discrepancy here and there*
- *NOTICE THAT: this would be a suboptimal event generator, but we want to use it for anomaly detection*
  - *no guarantee that the best autoencoder is the best anomaly detector (no anomaly detection rate in the loss function)*
  - *pros & cons of an unsupervised/ semisupervised approach*

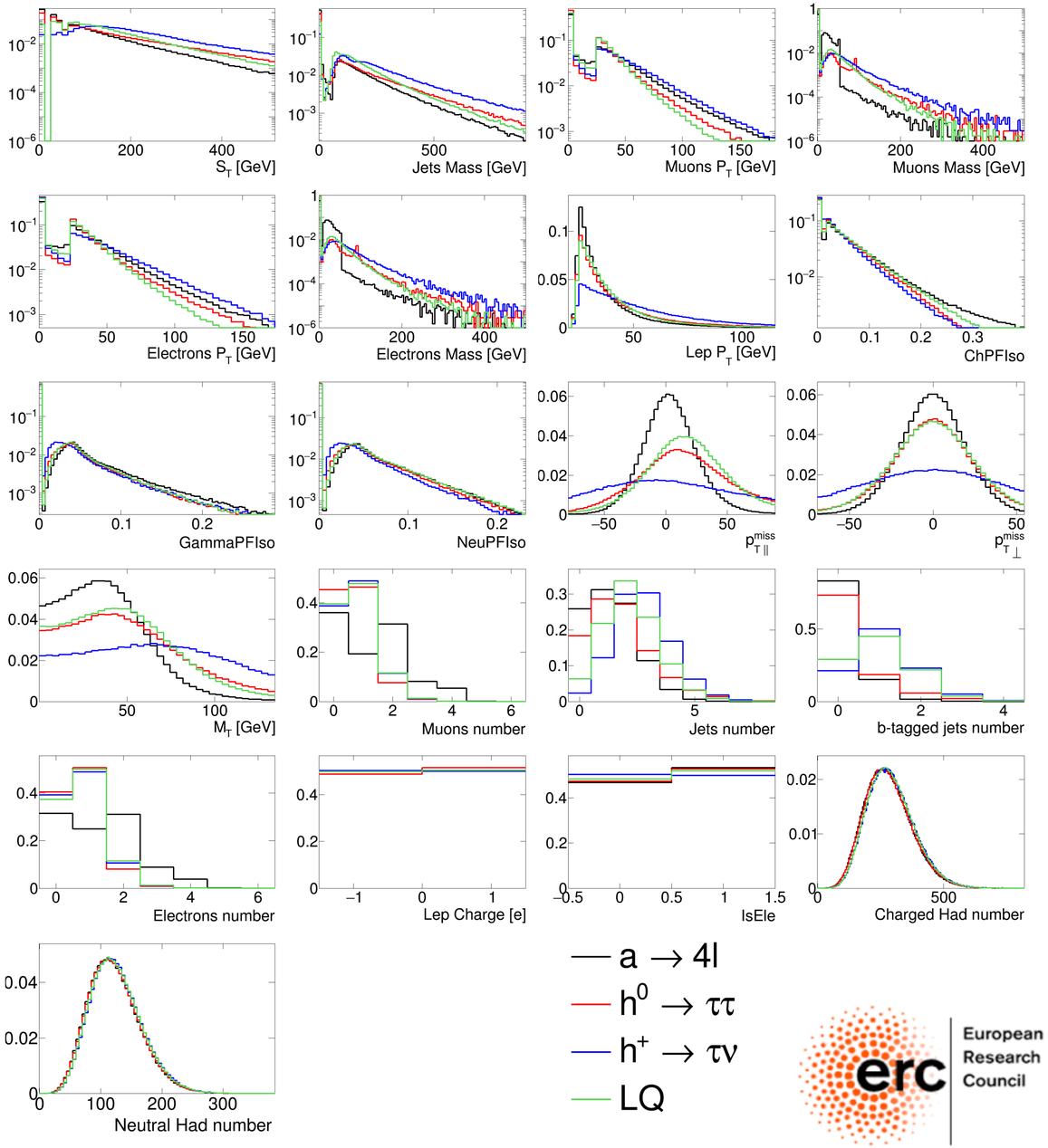


Some BSM benchmark

- We consider four BSM benchmark models, to give some sense of VAEs potential
  - leptoquark with mass 80 GeV, LQ→bτ
  - A scalar boson with mass 50 GeV, a→Z*Z*→4ℓ
  - A scalar scalar boson with mass 60 GeV, h→ττ
  - A charged scalar boson with mass 60 GeV, $h^{\pm}$→τν

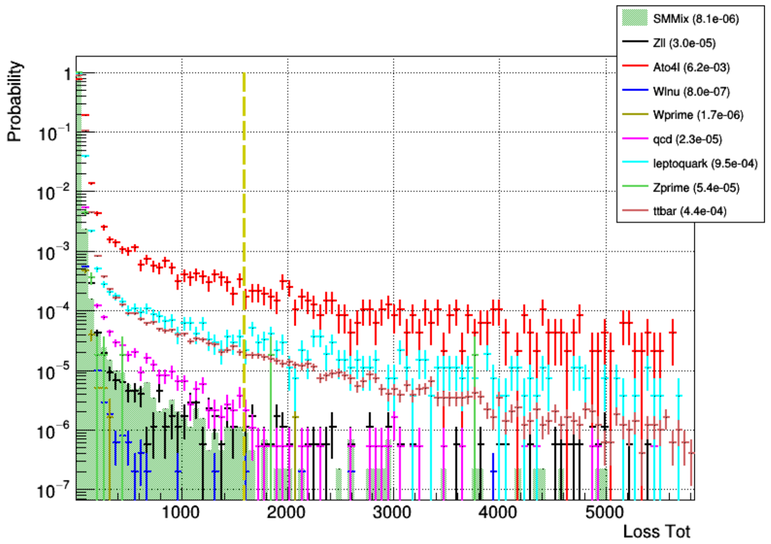
BSM benchmark processes

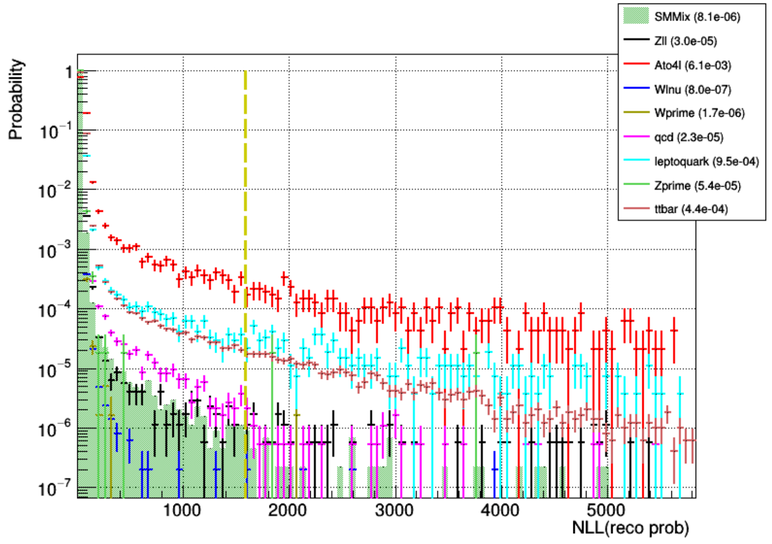
Process	Acceptance	Trigger efficiency	Total efficiency	Cross-section 100 events/month
$h^0 \to \tau\tau$	9%	70%	6%	335 fb
$h^0 \to \tau\nu$	18%	69%	12%	163 fb
$LQ \to b\tau$	19%	62%	12%	166 fb
$a \to 4\ell$	5%	98%	5%	436 fb

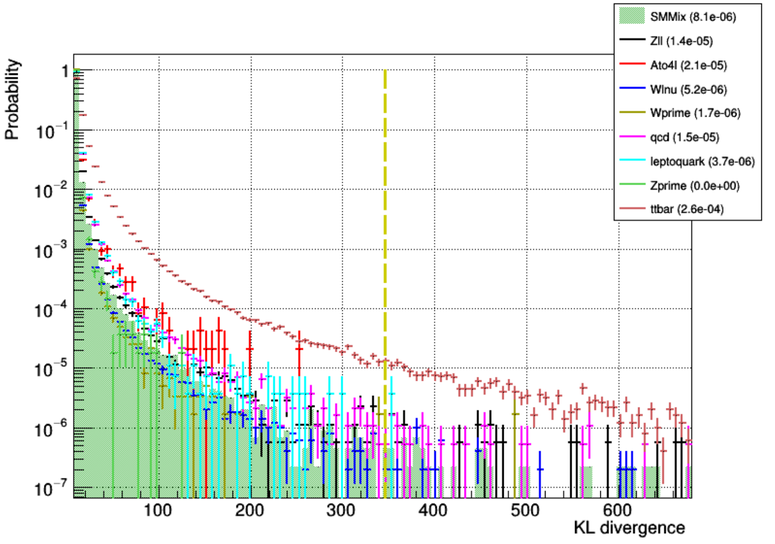


Defining anomaly

- *Anomaly defined as a p-value threshold on a given test statistics*
  - *Loss function an obvious choice*
  - *Some part of a loss could be more sensitive than others*
  - *We tested different options and found the total loss to behave better*

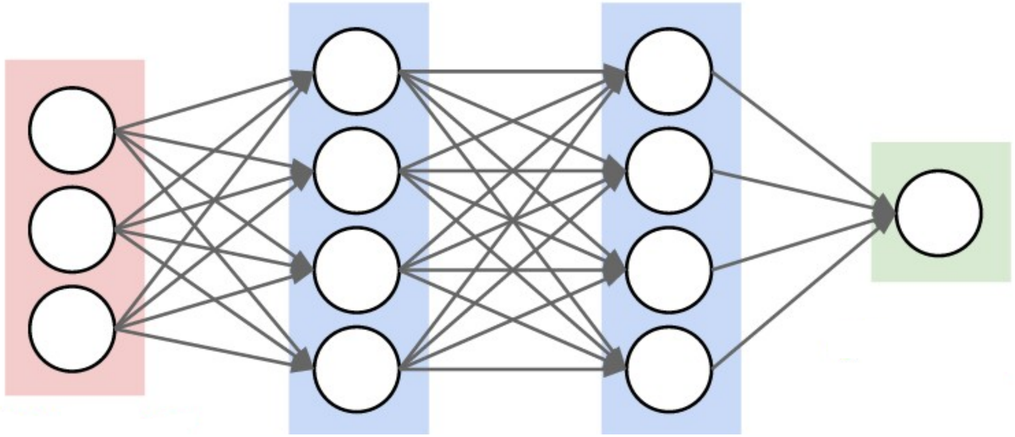


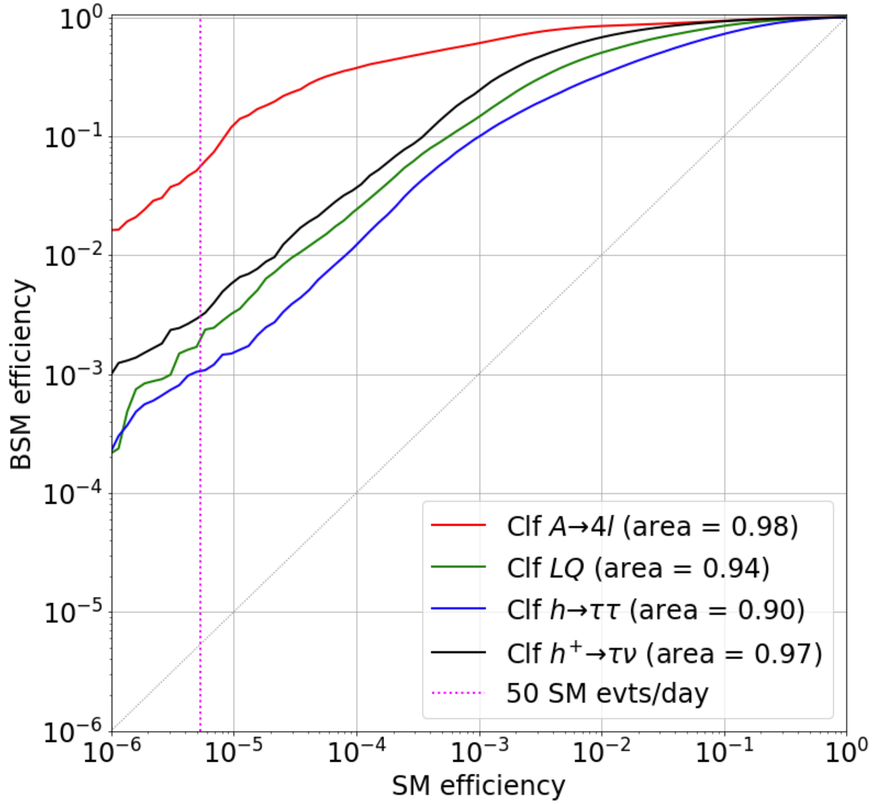




Benchmark comparison

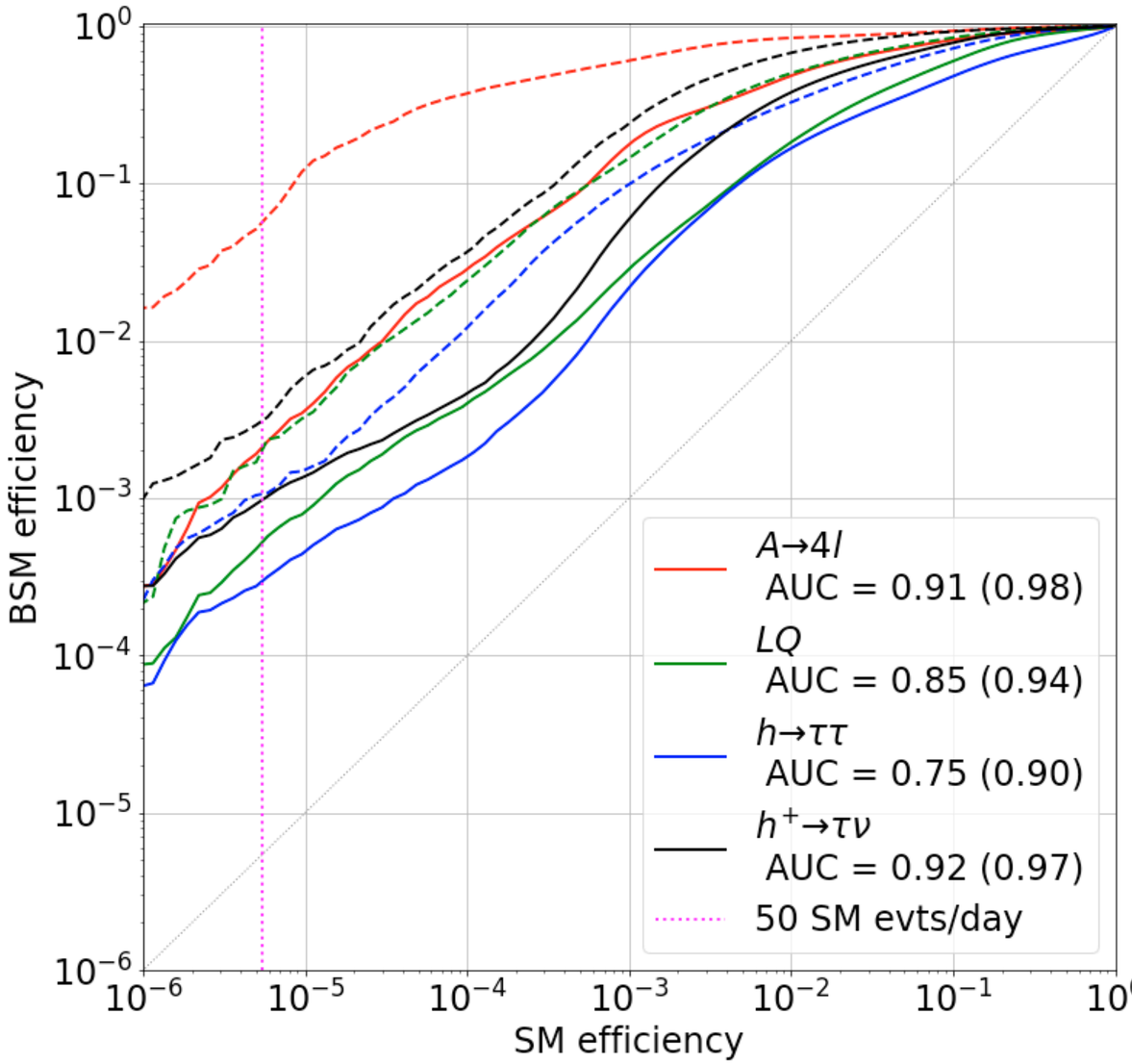
- *VAE's performances benchmarked against supervised classifiers*
- *For each BSM model*
  - *take same inputs as VAE*
  - *train a fully-supervised classifier to separate signal from background*
  - *use supervised performances as a reference to aim to with the unsupervised approach*
  - *Done for our 4 BSM models using dense neural networks*





Performances

- *Evaluate general discrimination power by ROC curve and area under curve (AUC)*
  - *clearly worse than supervised*
  - *but not so far*
- *Fixing SM acceptance rate at 50 events/day (assuming L=XXX*
  - *competitive results considering unsupervised nature of the algorithm*



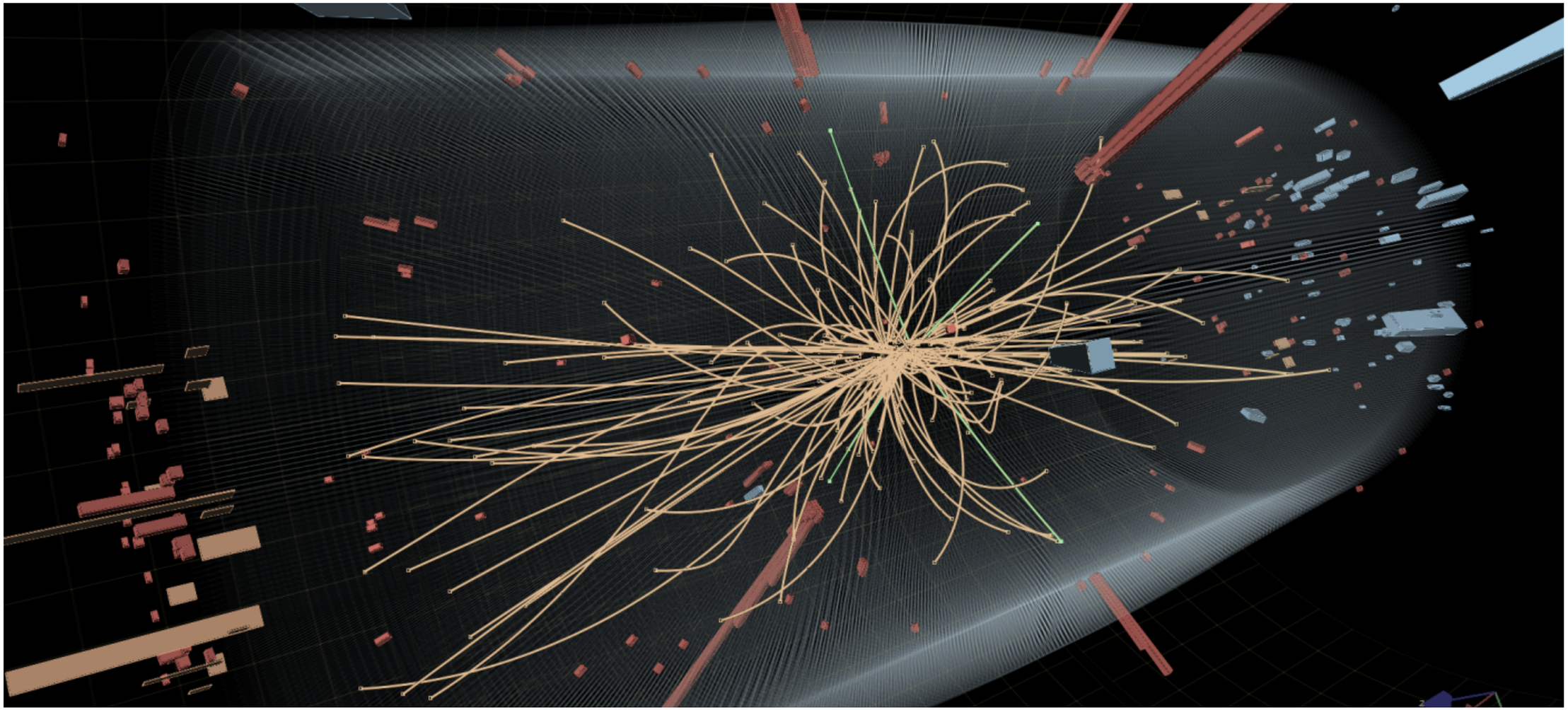
Performances

- *Small efficiency but still much larger than for SM processes*
- *Allows to probe 10-100 pb cross sections for reasonable amount of collected signal events*

Process	Efficiency for ~30 evt/day	xsec for 100 evt/ month [pb]	xsec for S/B~1/3 [pb]
$a \to 4\ell$	$2.8 \cdot 10^{-3}$	7.1	27
$LQ \to \tau b$	$6.5 \cdot 10^{-4}$	31	120
$h \to \tau\tau$	$3.6 \cdot 10^{-4}$	56	220
$h^{\pm} \to \tau\nu$	$1.2 \cdot 10^{-3}$	17	67

1/2 way to model independence

- *Procedure designed to be model independent*
  - *Training done only on SM*
  - *Algorithm that defines anomaly tuned only on number of selected SM events (false positive rate)*
- *Still, residual model dependence present*
  - *Based on physics-motivated observables*
  - *List not tailored on specific models and general enough to offer good performances in principle*
  - *But one cannot prove that performances on specific BSM models will generalise*
- *Can we go beyond this limitation and define something really BSM agnostic?*

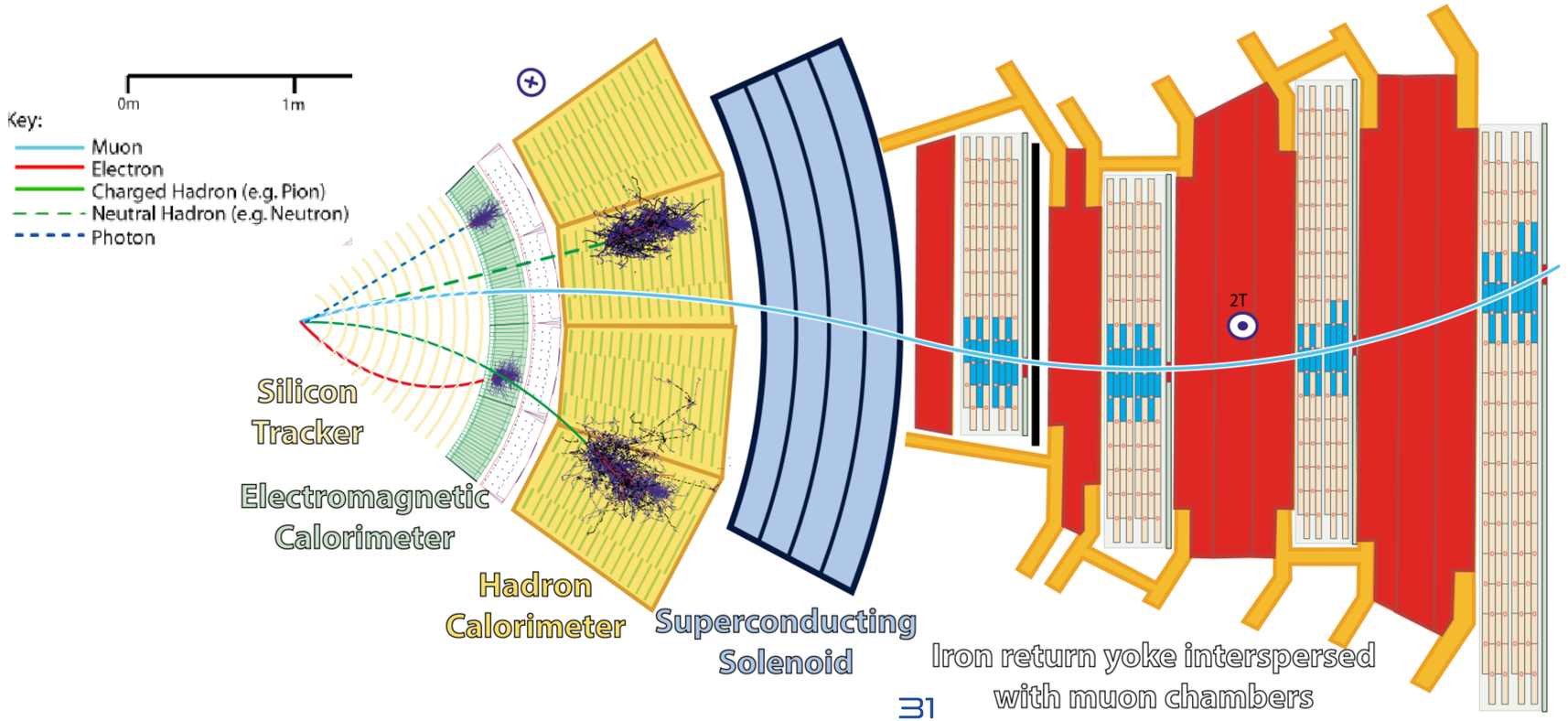


Language processing for particle physics

European Research Council

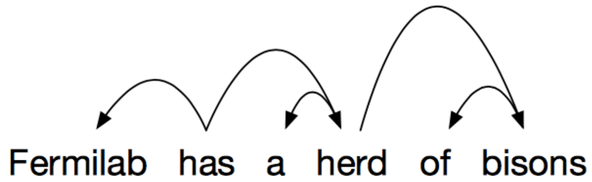
Particle Flow

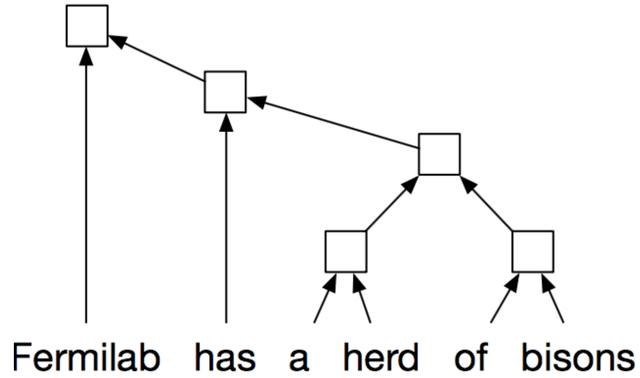
- *CMS uses PF to combine sub-detector information and produce a list of reconstructed particles*
- *Anything (jets, MET, resonances, etc) is reconstructed from these particles*
- *One could generalise the VAE new-physics-detection algorithm and make it PF compliant*
  - *integrated in the reconstruction flow @HLT*
  - *can abstract from model dependence inherited by any physics-motivated HLF choice*

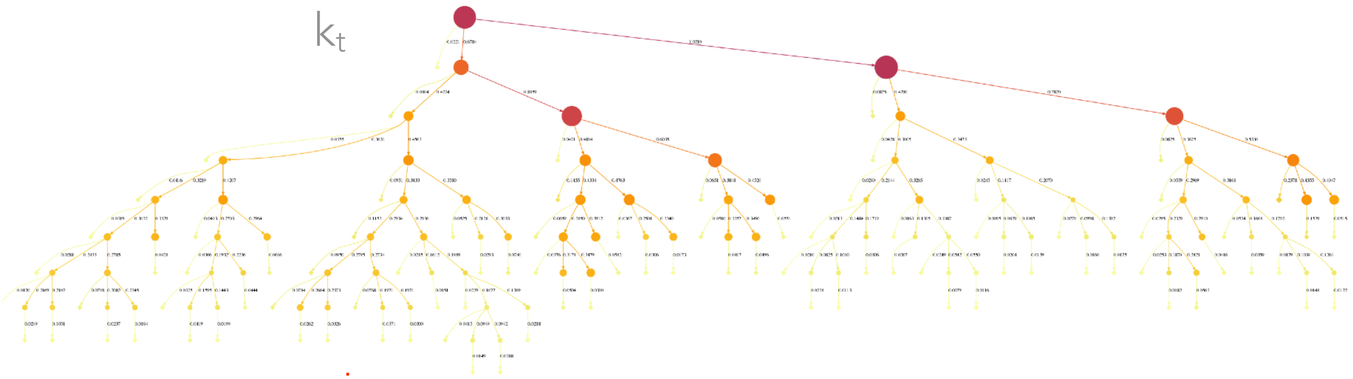


LHC events & language processing

- *PF reco is not the best match for computing vision techniques (e.g., convolutional neural networks) don't work*
  - *one would have to convert the particles to a pixelated images, loosing resolution*
- *Instead, list of particles can be processed by Deep Learning architectures designed for natural language processing (RNN, LSTMs, GRUs, …)*
  - *particles as words in a sentence*
  - *QCD is the grammar*

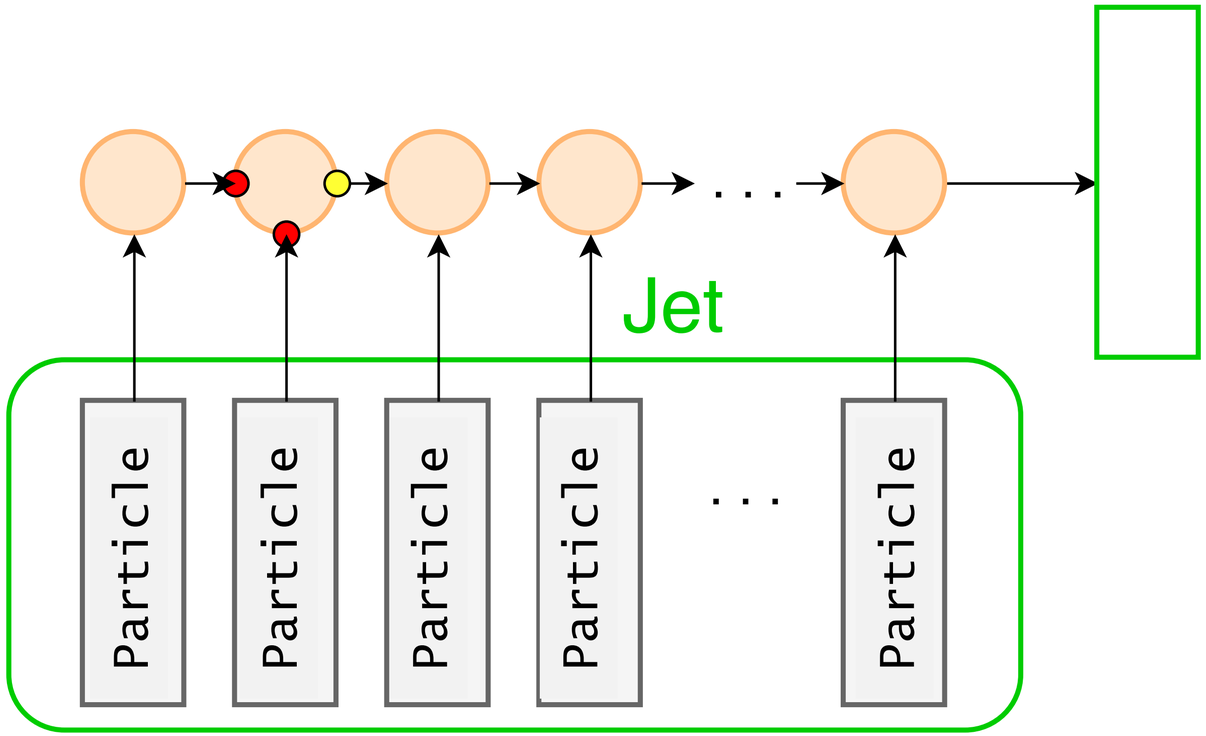






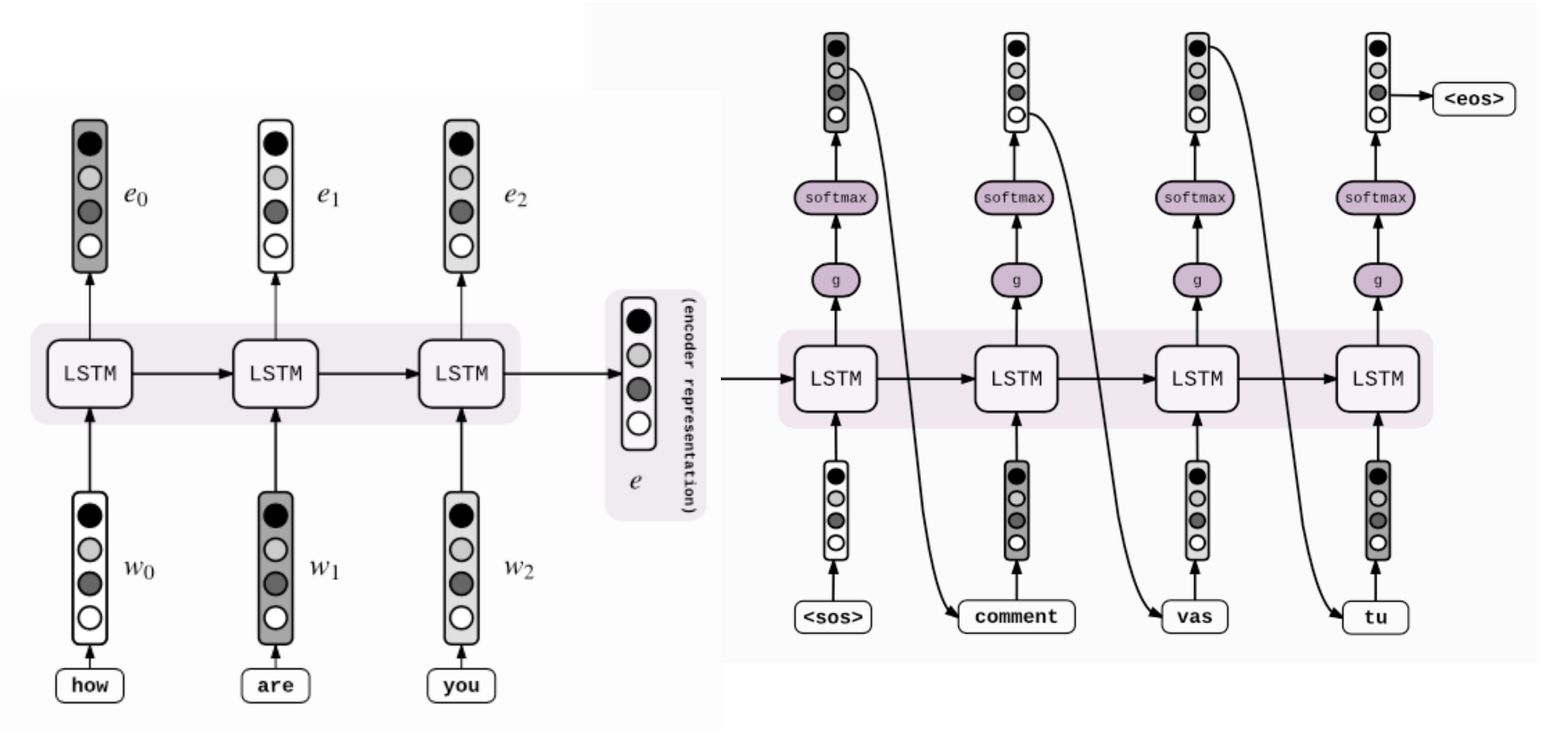
Recurrent Neural Networks

- A network architecture suitable to process an ordered sequence of inputs
  - words in text processing
  - a time series
  - particles in a list
- Could be used for a single jet or the full event
- Next step: graph networks (active research direction)



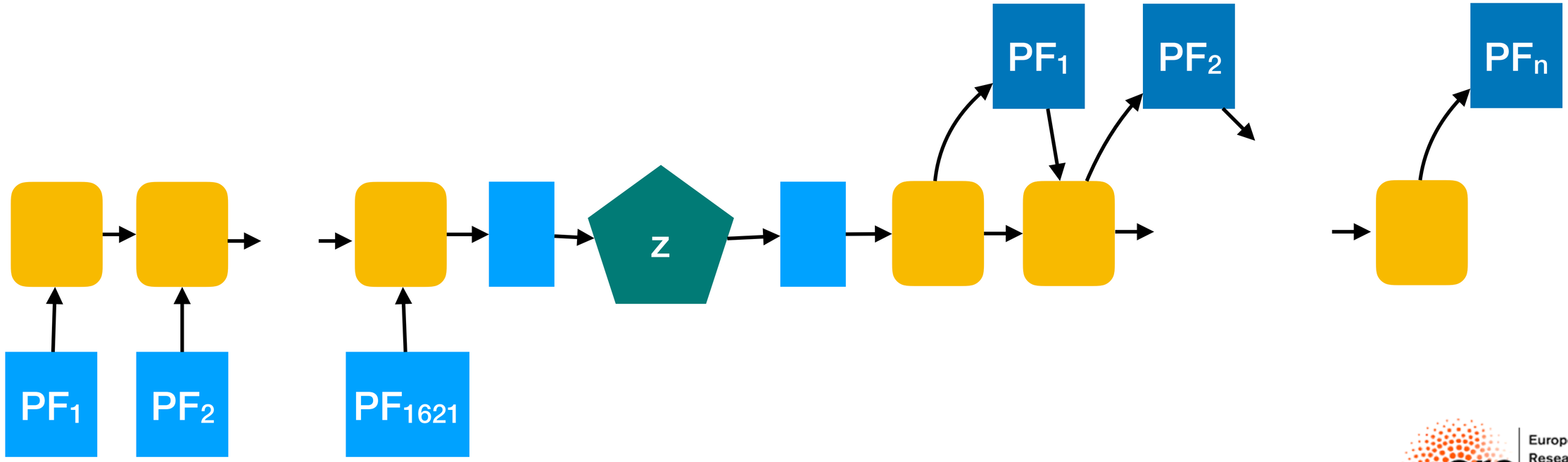
VAE with PF particles

- *Issues:*
  - *variable number of particles/event as input*
  - *need to return particles as output*
- *Networks used for translation*
  - *start from a sentence in language*
  - *code its meaning in some latent space z*
  - *translate to some other language, generating words from z*



VAE with PF particles

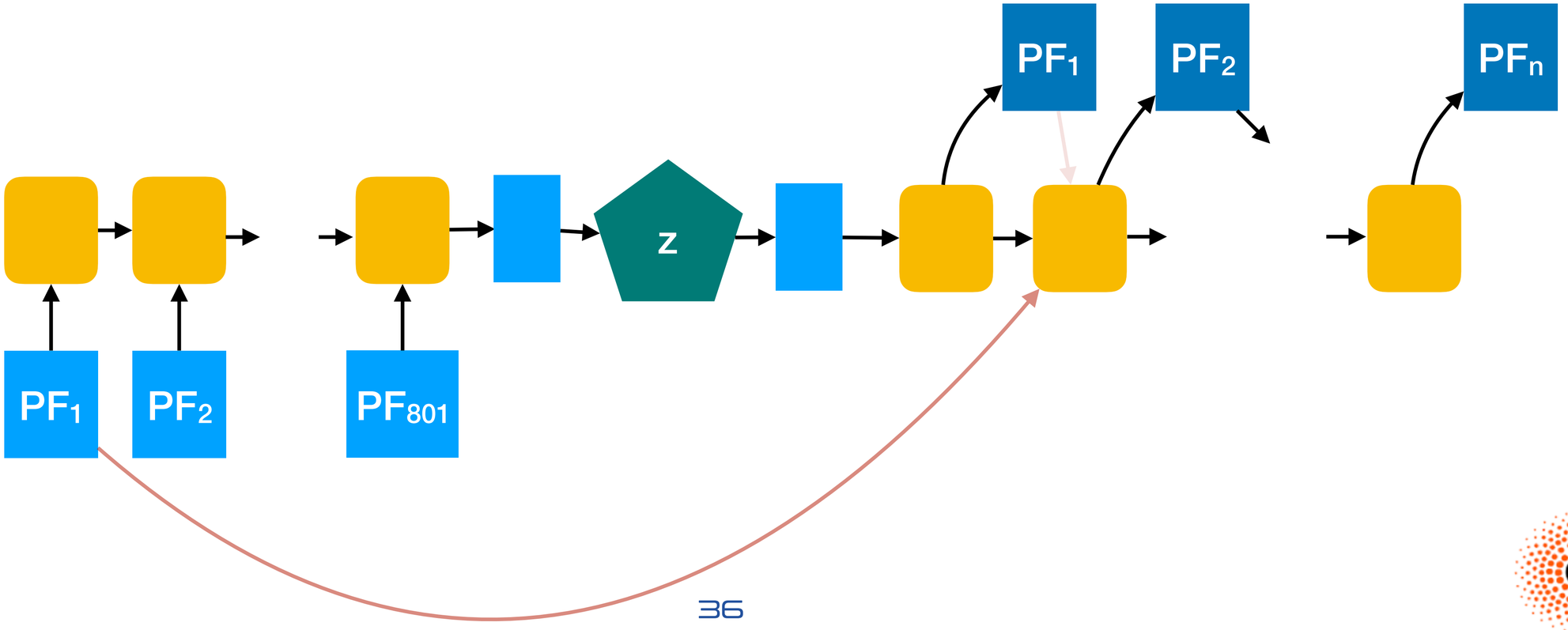
- *Issues:*
  - *variable number of particles/event as input*
  - *need to return particles as output*



Teacher forcing

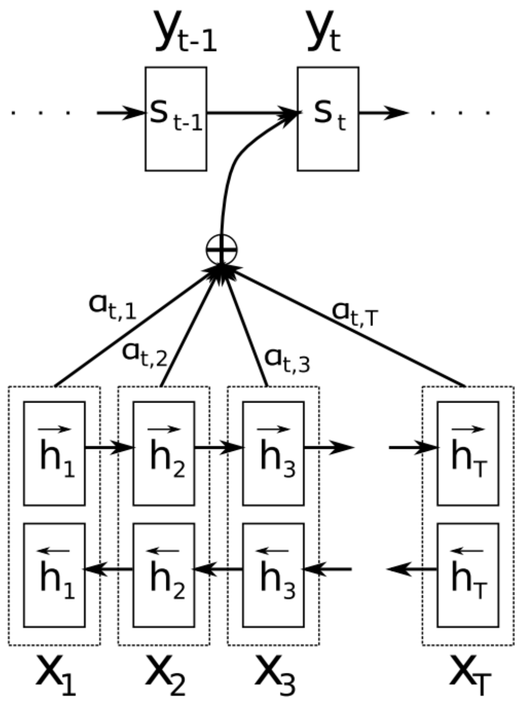
- *At early stage of training, the decoder can't reconstruct a reasonable first PF candidate; autoregressive mechanism propagates it into a wrong chain of particles.*

- ***Teacher-forcing:*** *under some probability k, feed the target as the next input instead of using the previous prediction. k decreases as the epoch number increases.*



Adding Attention

- *Attention allows the decoder to focus on which part of the inputs is relevant to the next prediction.*



the **Encoder** generates **h1,h2,h….hT** from the inputs **X1,X2,X3…XT**

a is the **Alignment model** which is a **feedforward neural network** that is trained with all the other components of the proposed system

$$e_{ij} = a(s_{i-1}, h_j)$$

The **Alignment model** scores (e) how well each encoded input (h) matches the current output of the decoder (s).

The alignment scores are normalized using a **softmax function.**

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$

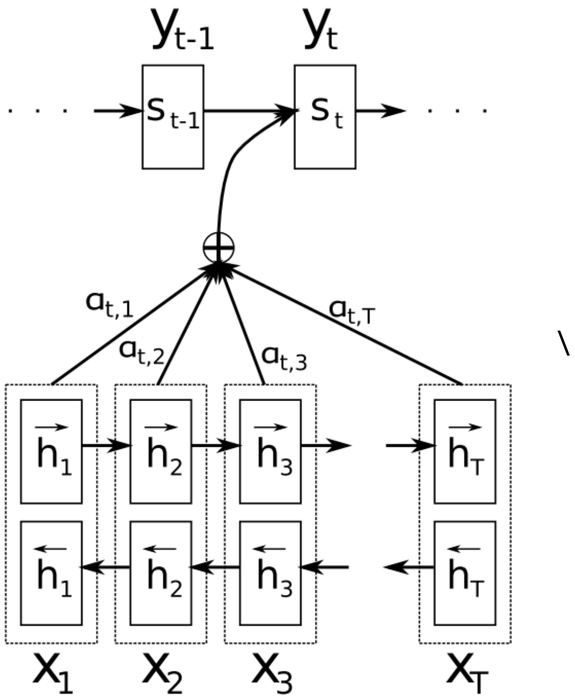
The context vector is a weighted sum of the **annotations** (hj) and **normalized alignment scores.**

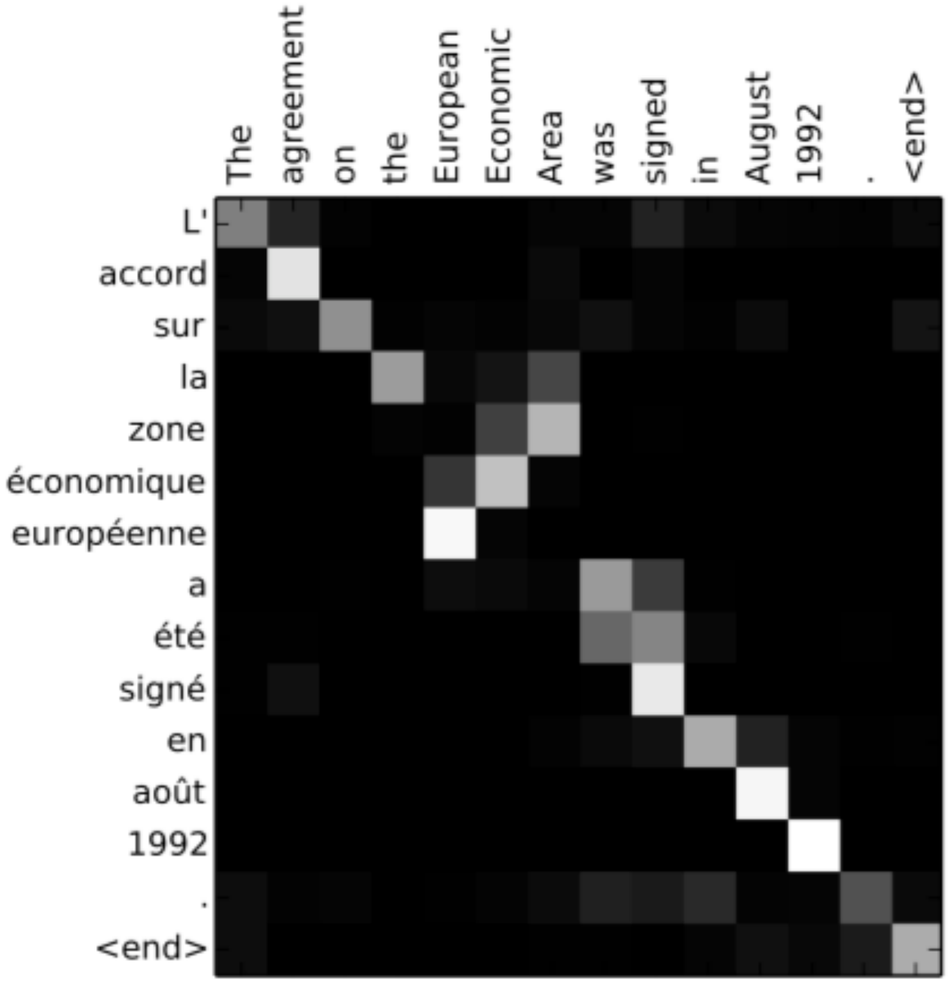
$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j.$$

Bahdanau et al., arXiv:1409.0473

Adding Attention

- *Attention allows the decoder to focus on which part of the inputs is relevant to the next prediction.*

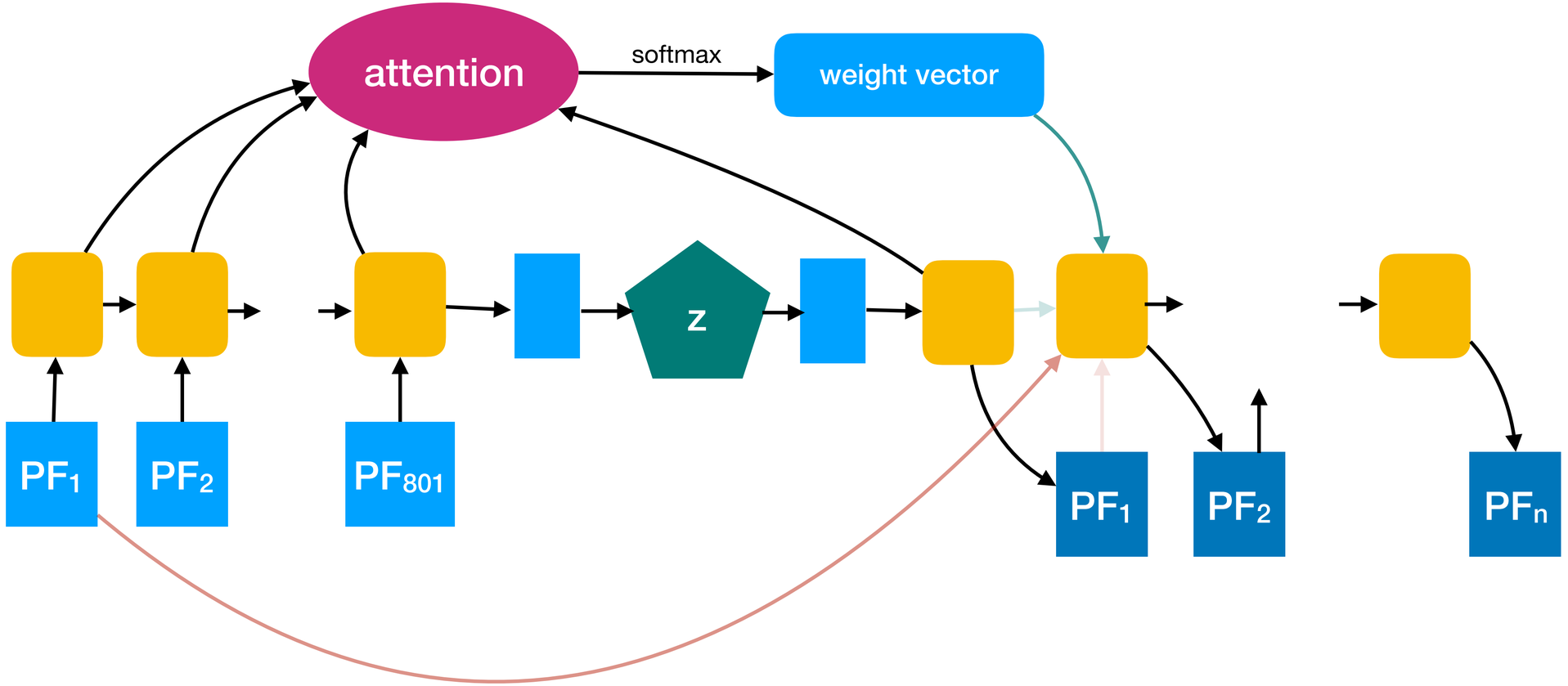




Bahdanau et al., arXiv:1409.0473

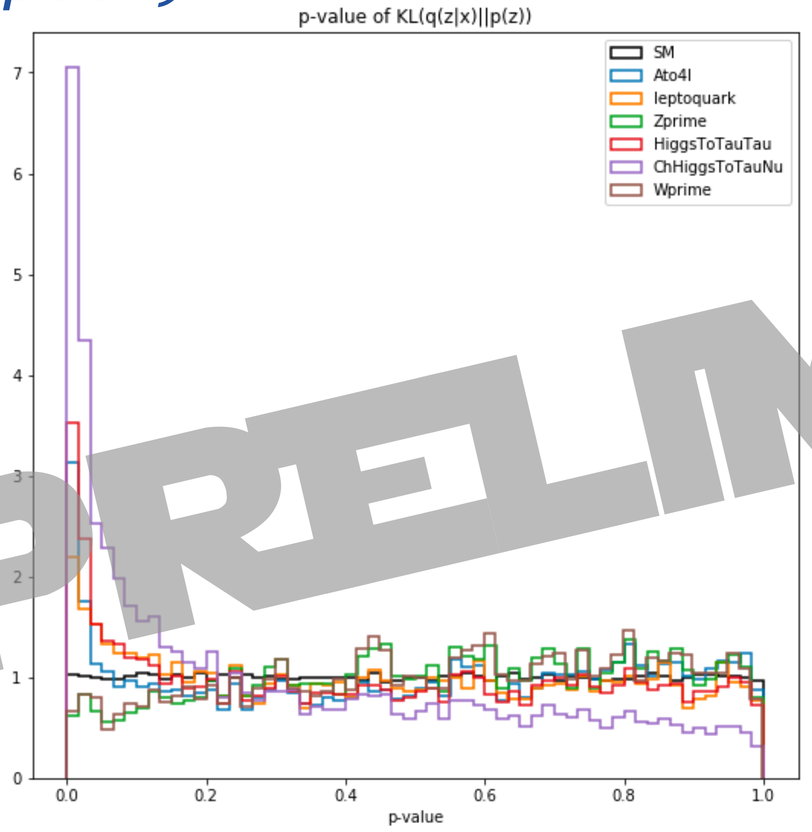
Adding Attention

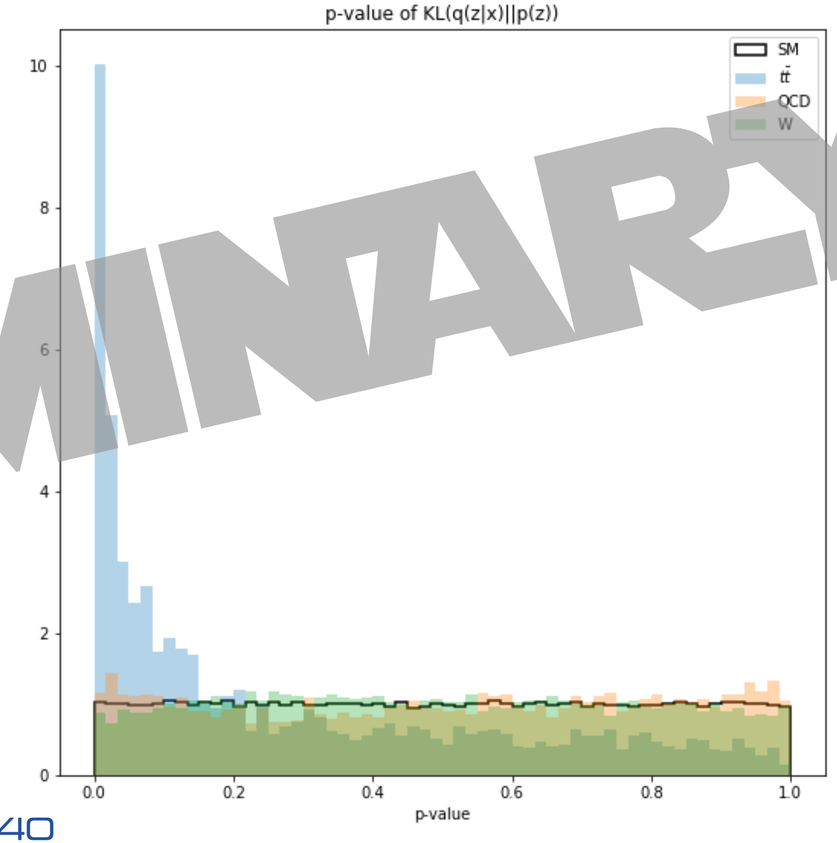
- *Attention allows the decoder to focus on which part of the inputs is relevant to the next prediction.*



Performances

- *(Preliminary) results trained on a small subset of the initial dataset (90K events)*
- *Due to architecture complexity, training is much slower (6h/epoch)*





Performances

- *(Preliminary) results trained on a small subset of the initial dataset (90K events)*
- *Due to architecture complexity, training is much slower (6h/epoch)*

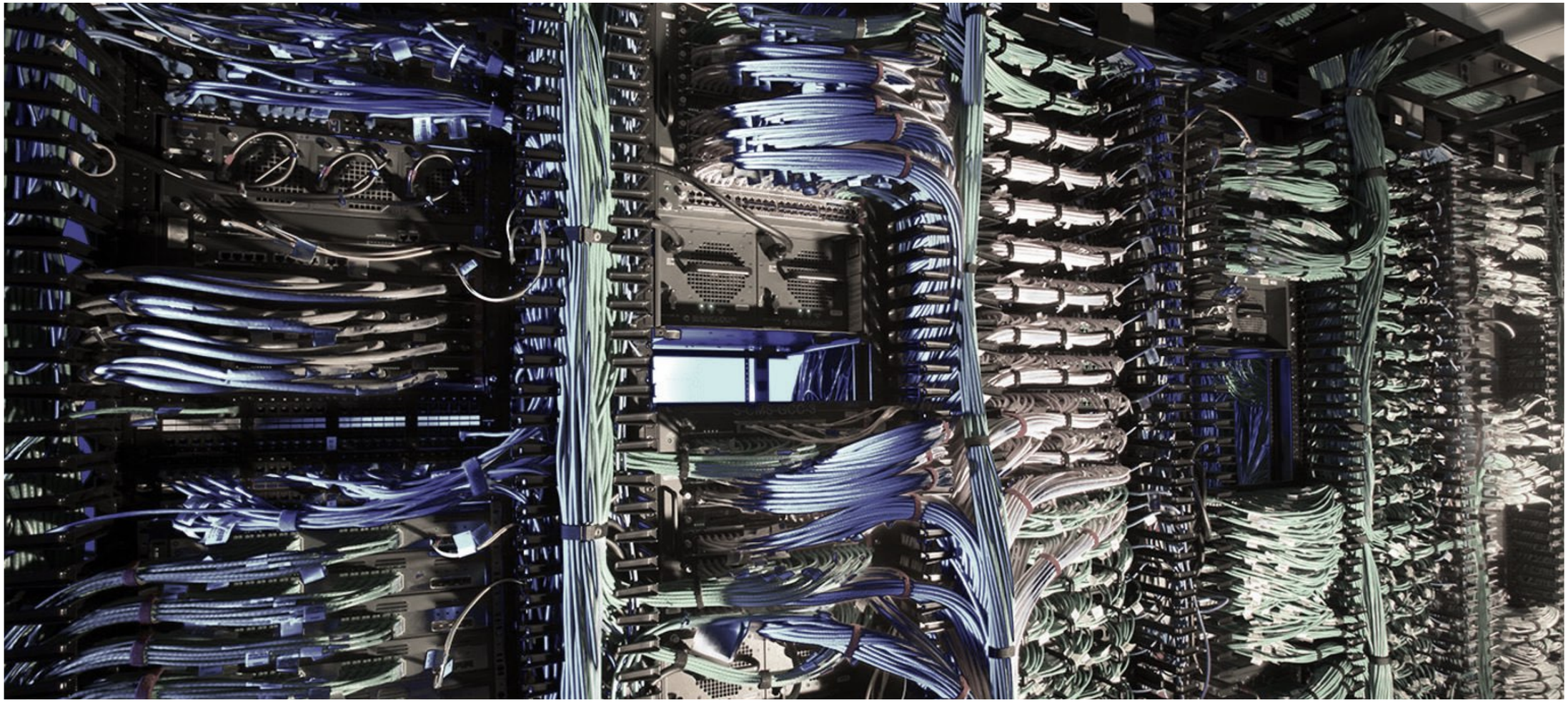
Process	p-value = 0.05	p-value = 0.01	p-value = 0.001	p-value = 0.0001
$a \to 4\ell$	0.100	0.036	0.007	0.002
$LQ \to \tau b$	0.090	0.021	0.003	0.001
$h \to \tau\tau$	0.124	0.040	0.010	0.004
$h^{\pm} \to \tau\nu$	0.232	0.079	0.018	0.006

European Research Council

Performances

- *(Preliminary) results trained on a small subset of the initial dataset (90K events)*
- *Due to architecture complexity, training is much slower (6h/epoch)*

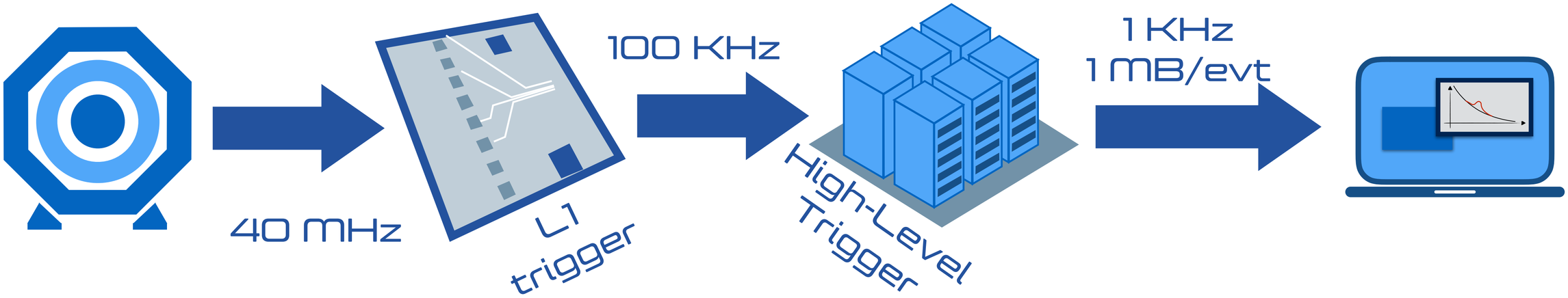
Process	Efficiency for ~300 evt/day	xsec for 10 evt/ month [pb]	xsec for S/B~1/3 [pb]
$a \to 4\ell$	$3.3 \cdot 10^{-4}$	7.2	$1.5 \cdot 10^{3}$
$LQ \to \tau b$	$5.8 \cdot 10^{-4}$	4.1	850
$h \to \tau\tau$	$1.1 \cdot 10^{-3}$	2.2	450
$h^{\pm} \to \tau\nu$	$1.4 \cdot 10^{-3}$	1.7	340



Fast Decision Taking

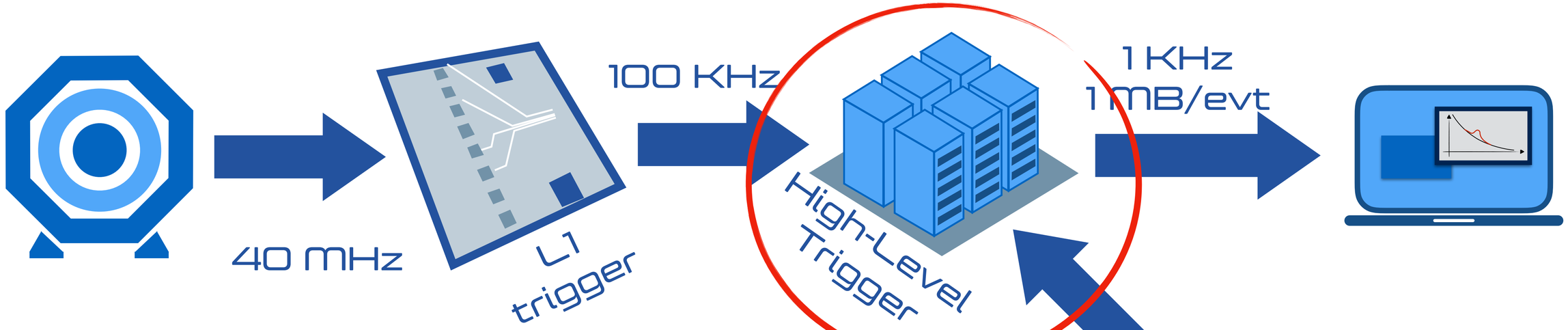
European Research Council

Future Architecture

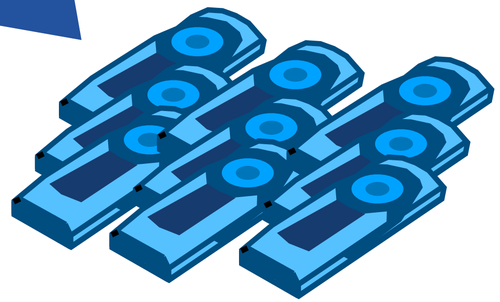


- *The more we use Machine Learning in our centralised reconstruction, the more we need to reconsider the design of our architecture*
  - *new performance/latency balance*
- *(to some extent) Not an issue after the event is written on disk (latency long enough that inference on CPU is OK*
- *This is not true for the trigger*

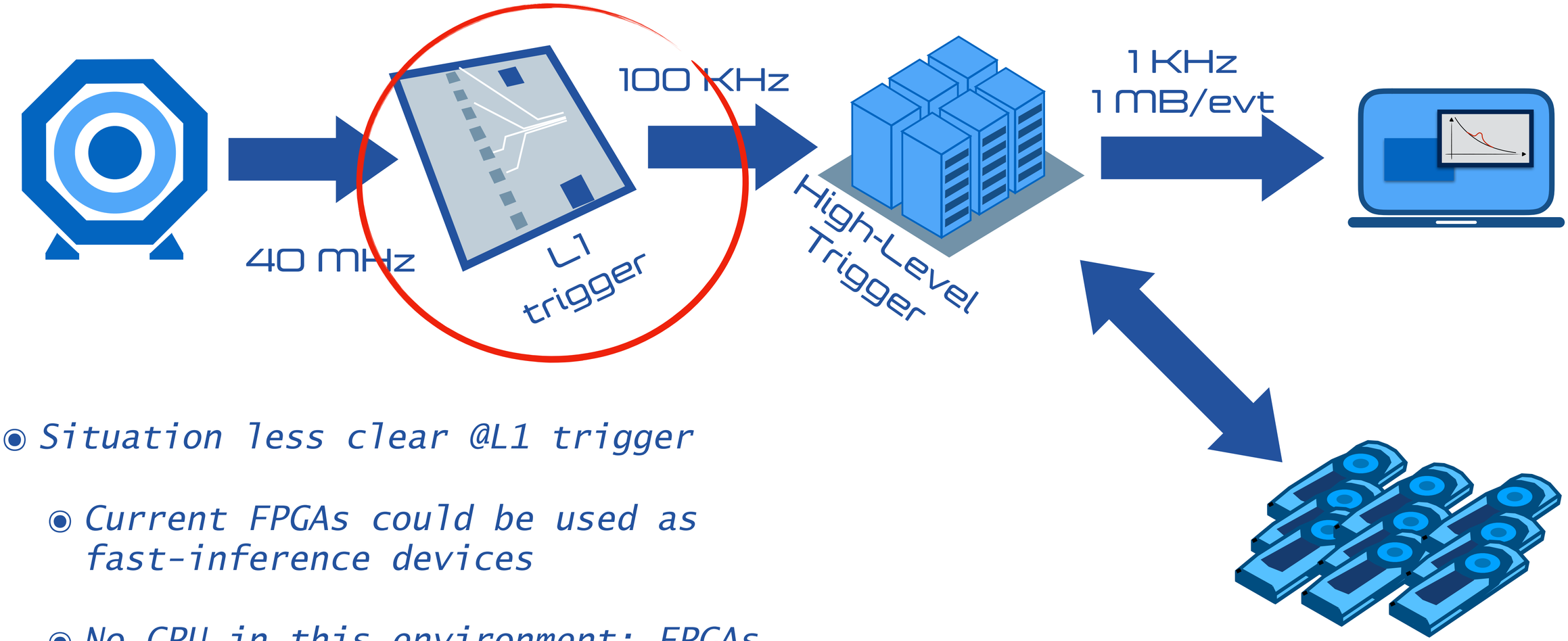
Not just a matter of training



- *Example: the seed-selection for tracking I showed you before*
  - *1 µsec to know if a seed is good or not*
  - *1M seeds/event -> 1sec to process an event serially*
- *This is why we are studying GPU clusters for our HLT farms*
  - *ML inference, tracking, etc delegated to GPUs as a service*



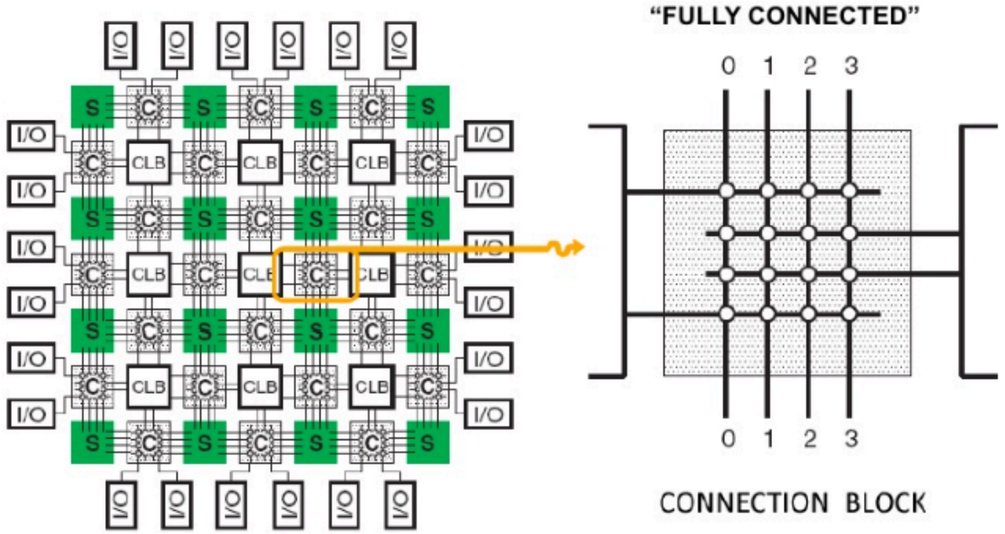
Not just a matter of training



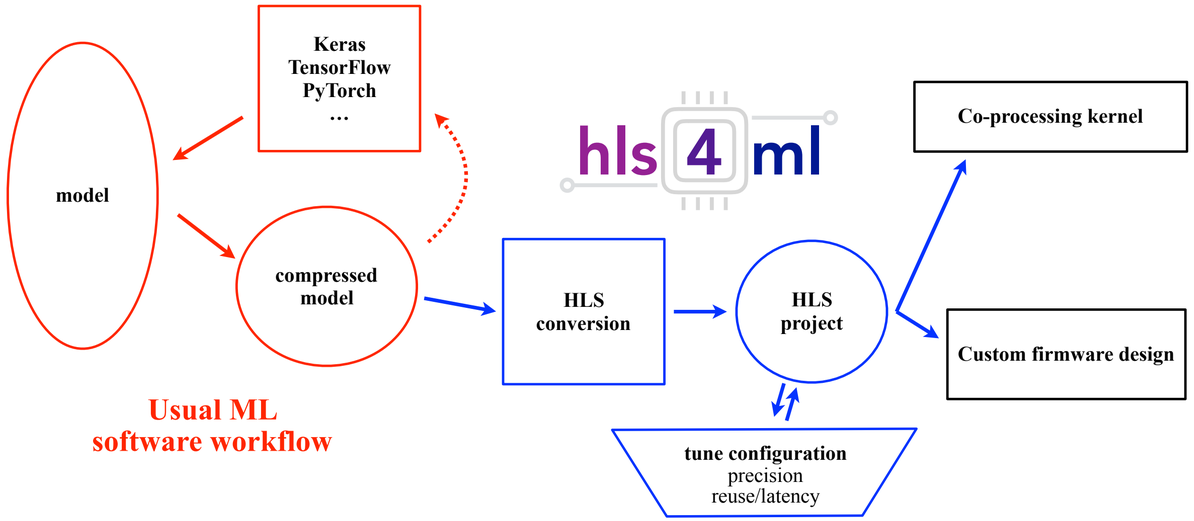
- *Situation less clear @L1 trigger*
  - *Current FPGAs could be used as fast-inference devices*
  - *No CPU in this environment: FPGAs attached to direct through optic fibres (10s of nsec latency)*

Bring DL to L1

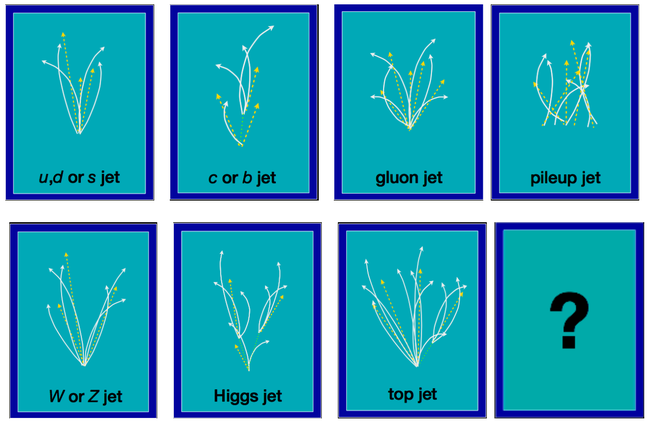
- *The L1 trigger is a complicated environment*
  - *decision to be taken in ~10 µsec*
  - *only access to local portions of the detector*
  - *processing on Xilinx FPGA, with limited memory resources*
- *Some ML already running @L1*
  - *CMS has BDT-based regressions coded as look-up tables*
- *Working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



HLS4ML: CERN/FNAL/MIT collaboration



Back to our first example



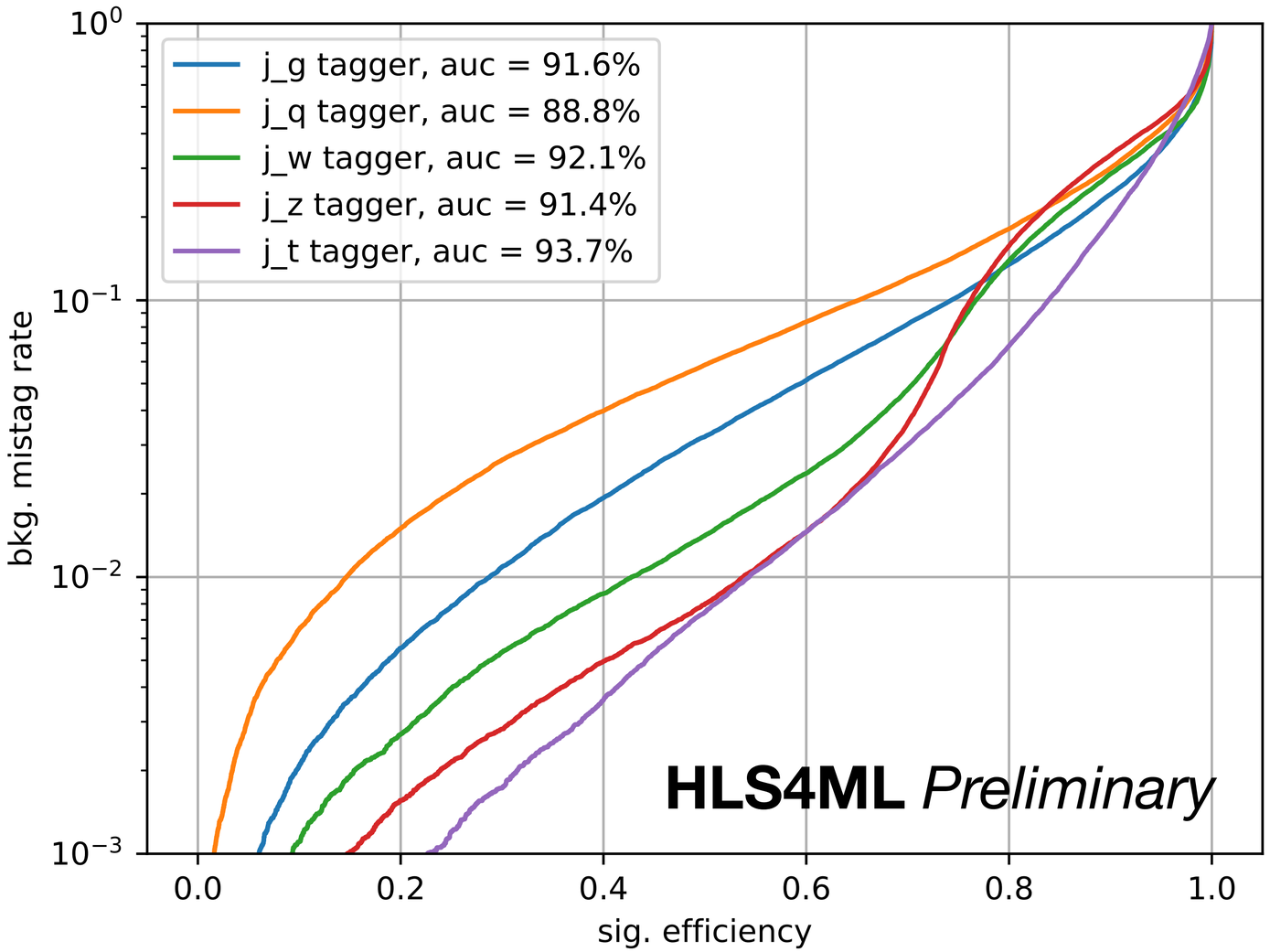
16 inputs

64 nodes
activation: ReLU

32 nodes
activation: ReLU

32 nodes
activation: ReLU

5 outputs
activation: SoftMax

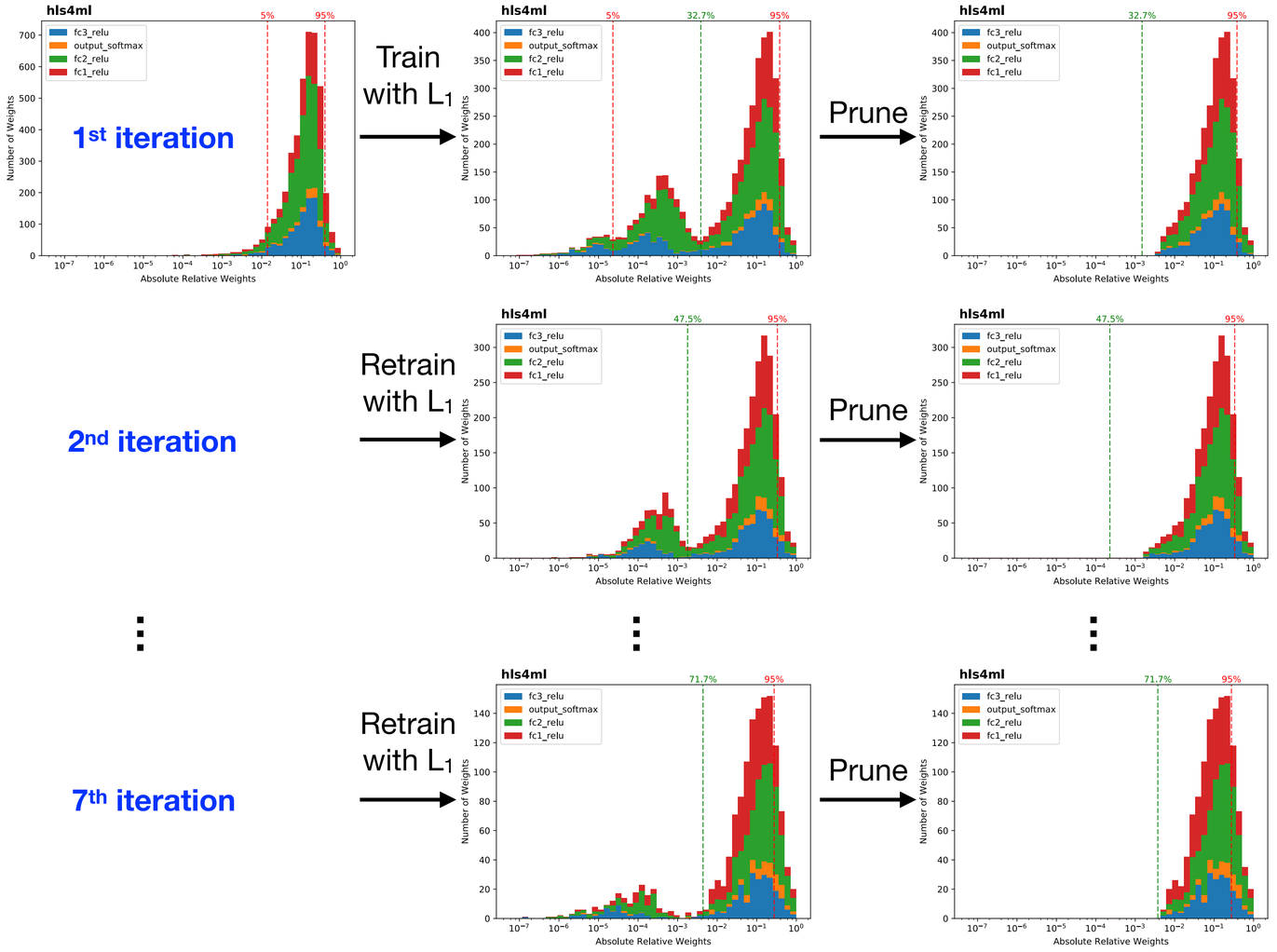


Make the model cheaper

- ***Pruning:*** *remove parameters that don't really contribute to performances*
  - *force parameters to be as small as possible (regularization)*

$$L_\lambda(\vec{w}) = L(\vec{w}) + \lambda||\vec{w}_1||$$

  - *Remove the small parameters*
  - *Retrain*

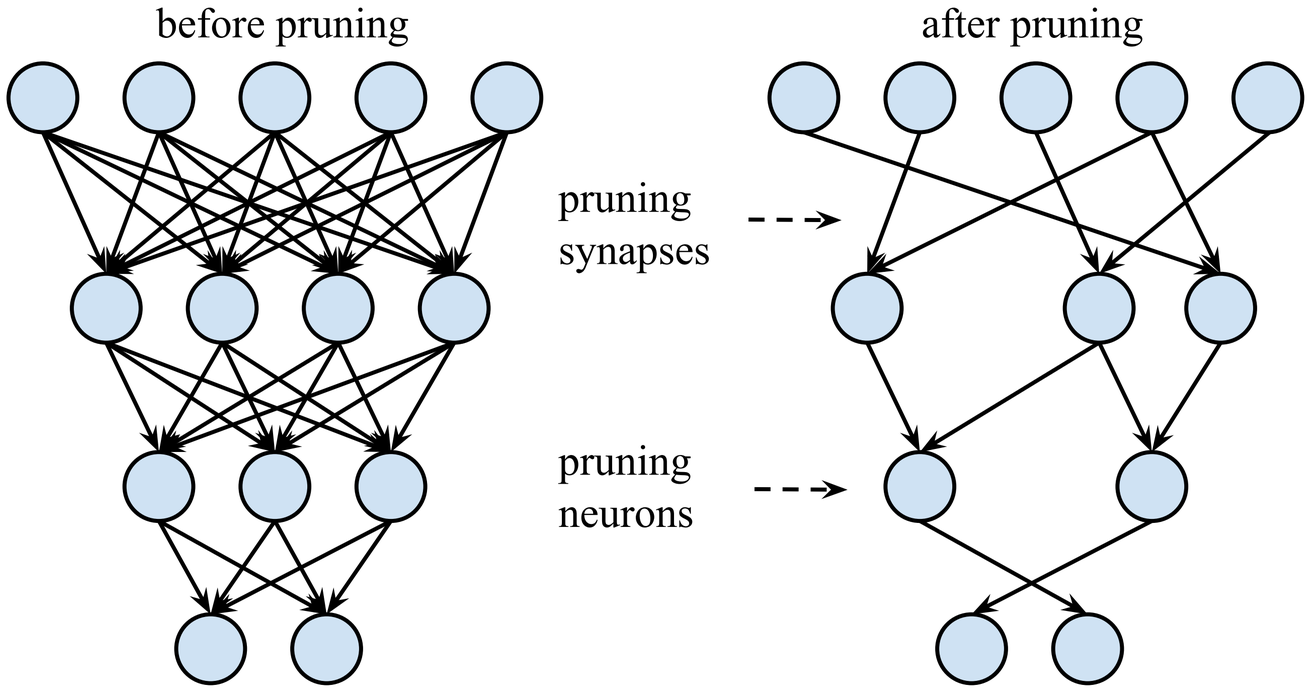


Make the model cheaper

- ***Pruning:*** *remove parameters that don't really contribute to performances*
  - *force parameters to be as small as possible (regularization)*

$$L_\lambda(\vec{w}) = L(\vec{w}) + \lambda||\vec{w}_1||$$

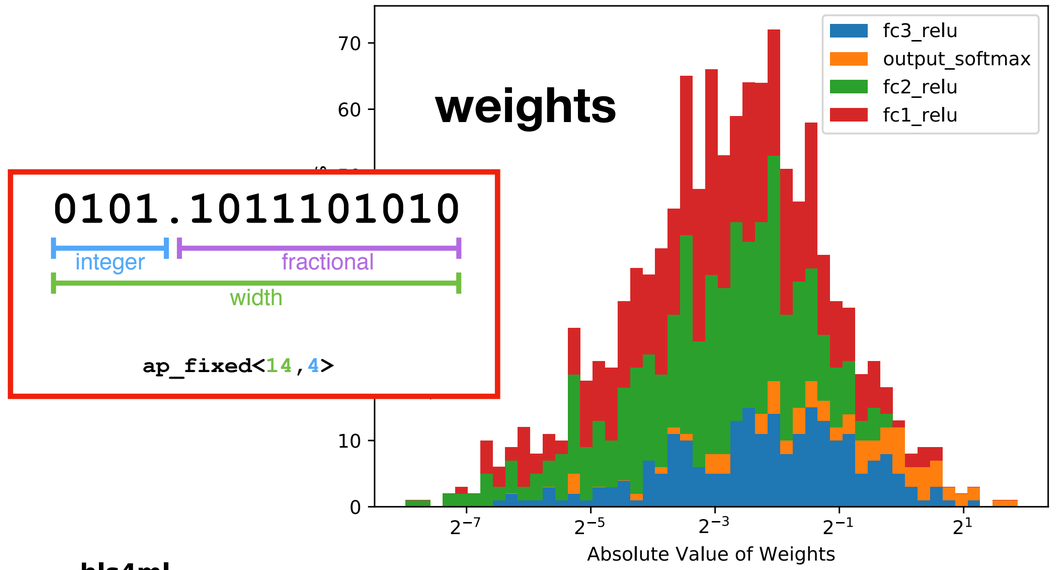
  - *Remove the small parameters*
  - *Retrain*

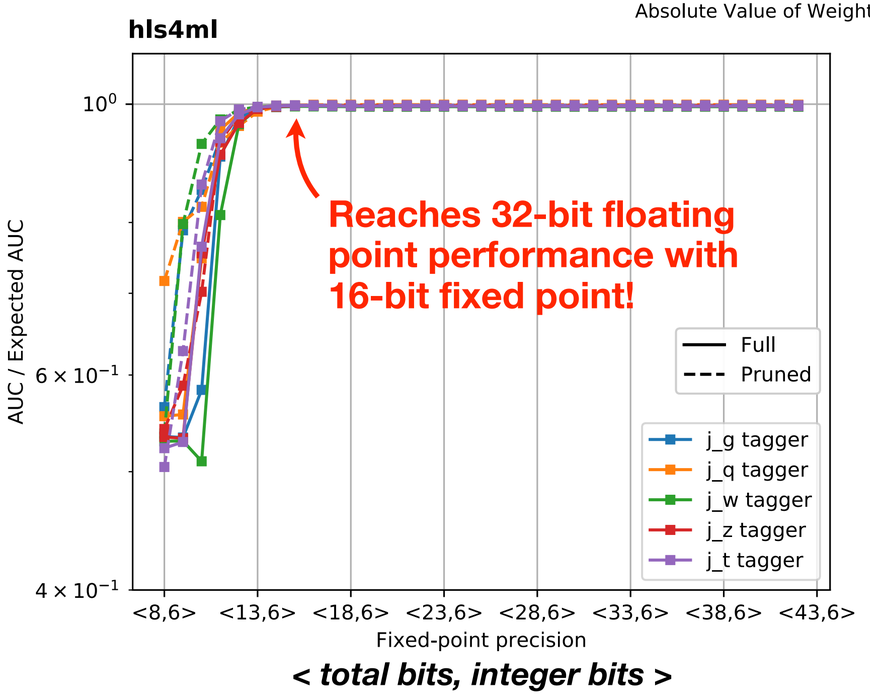


→ 70% reduction of weights and multiplications w/o performance loss

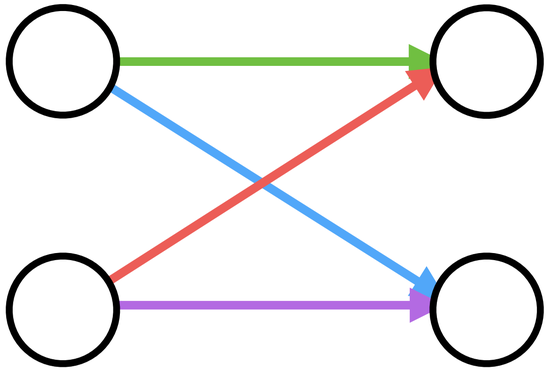
Make the model cheaper

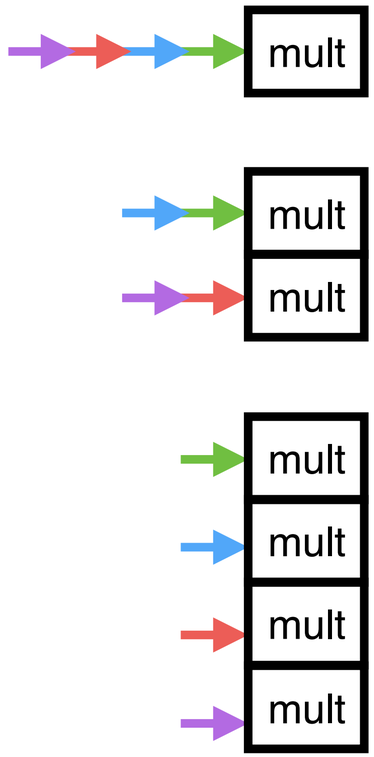
- ***Quantization***: *reduce the number of bits used to represent numbers (i.e., reduce used memory)*
  - *models are usually trained at 64 or 32 bits*
  - *this is not necessarily needed in real life*
- *In our case, we could reduce to 16 bits w/o loosing precision*
- *Beyond that, one would have to accept some performance loss*





Speed vs Memory





Fully serial

reuse = 4
use 1 multiplier 4 times

reuse = 2
use 2 multipliers 2 times each

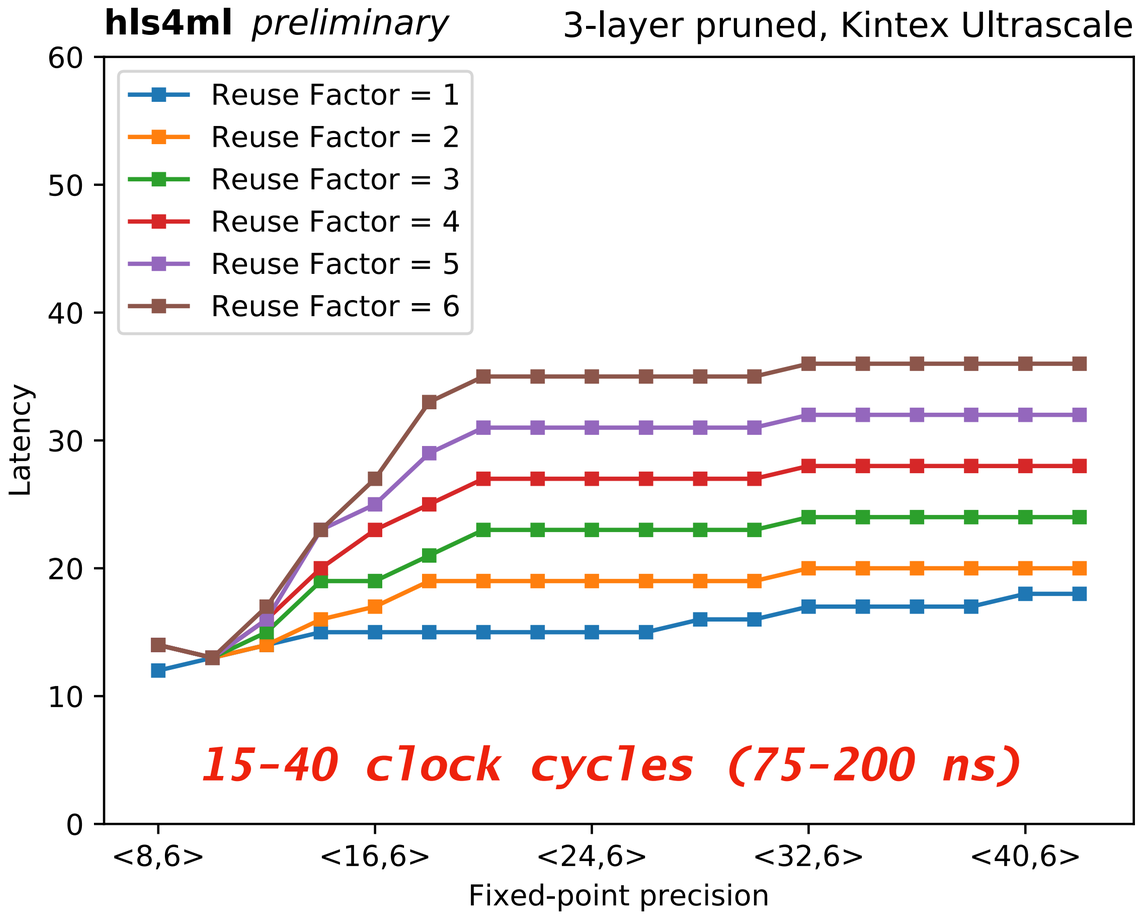
reuse = 1
use 4 multipliers 1 time each

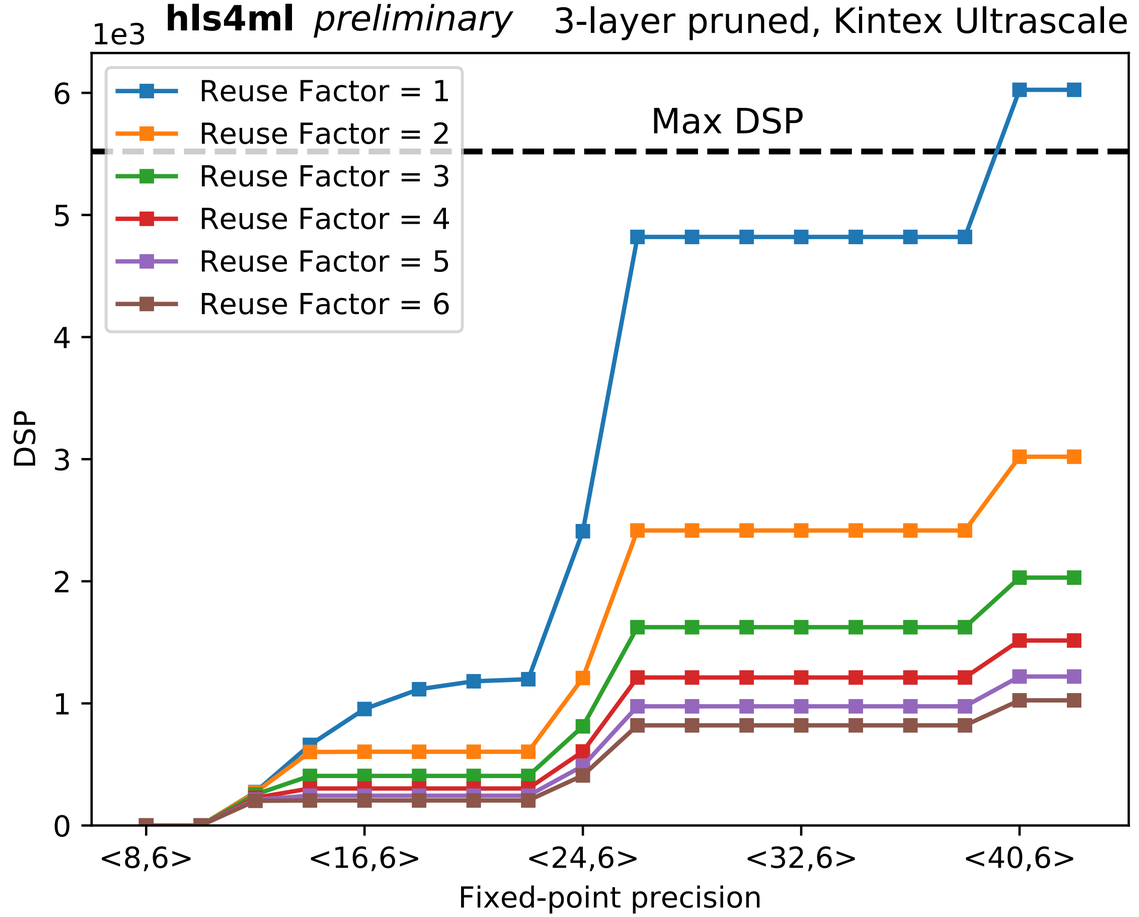
Fully parallel

Less resources/
Less throughput

Reuse factor: how much to parallelize operations in a hidden layer

Fast-inference





Foreseen architecture (FPGAs) will handle these networks
Inference-optimized GPUs could break the current paradigm
Looking forward to R&D projects with nVidia & E4 on this

Backup