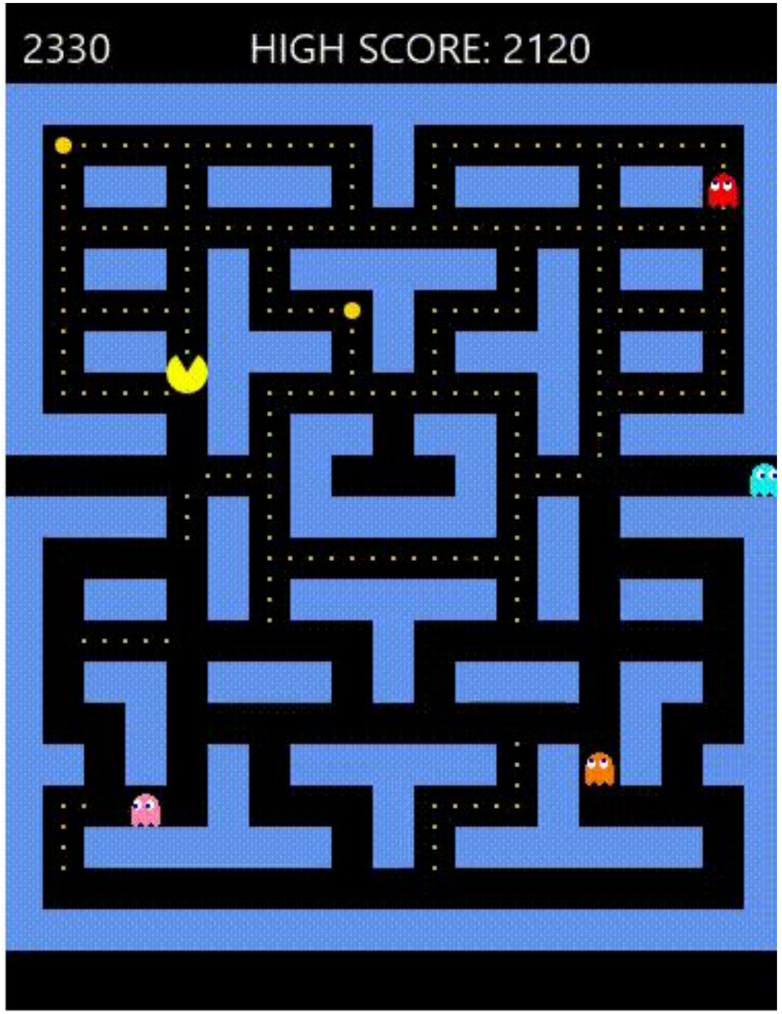
Working with neural networks at CERN

Roope Niemi

CERN | NexTGen Next Generation Triggers

Who I am & how I got here

- Sotungin lukio 2008-2011
- 2012-2017: non-IT work
- 2017-2022: Studies at University of Helsinki, Kumpula (Computer Science, Data Science)
- 2021-2024: Work at Nokia, applying to CERN a few times
- September 2024 → Software engineer / data scientist (QUEST) at CERN



What I do

- **Optimize neural networks for hardware, so they are fast and accurate → design methods to train compressed neural networks**

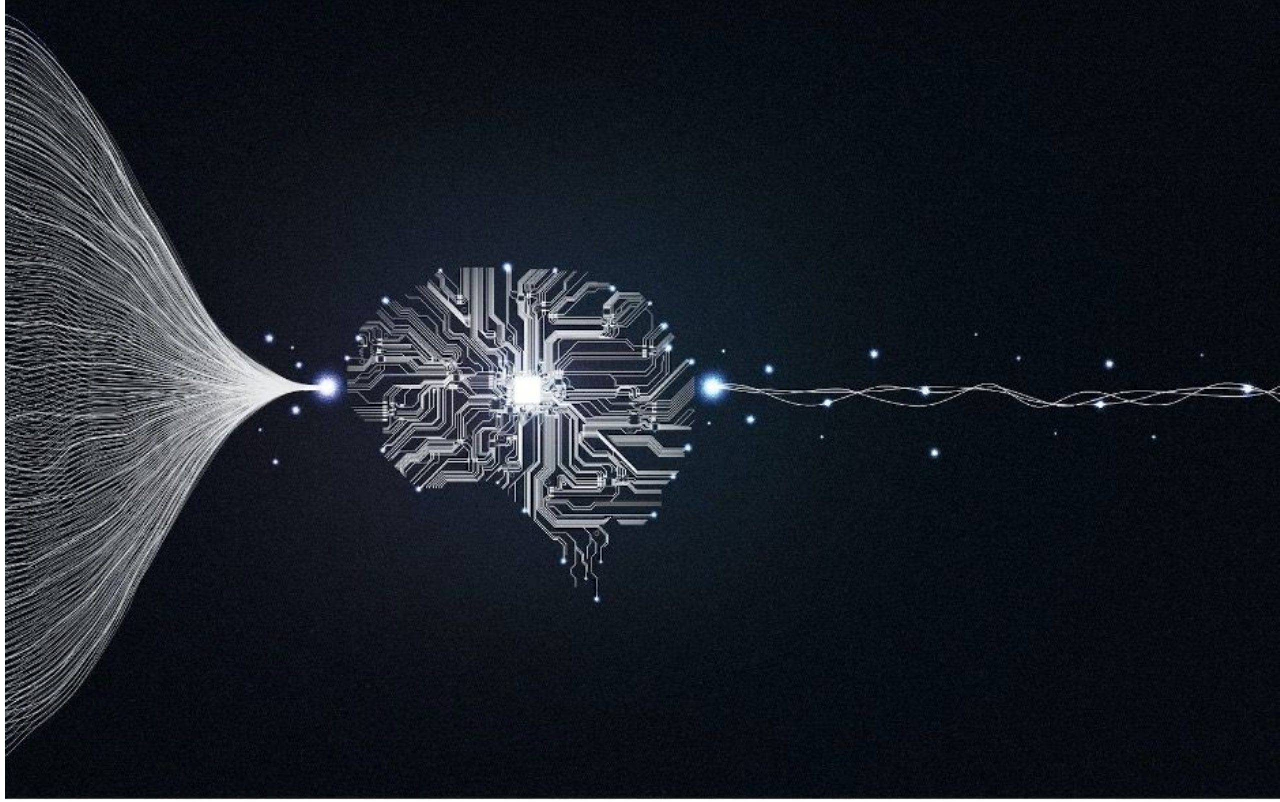
- **Programming with Python, read scientific papers, implement algorithms from them and compare them to other algorithms. Possibly improve them**

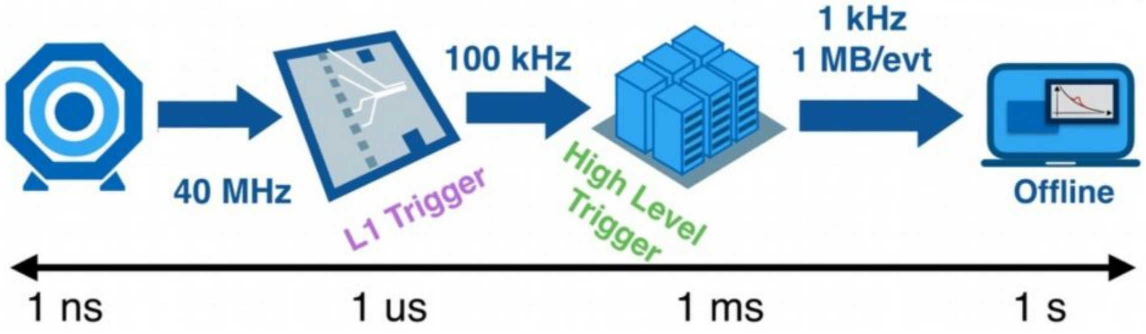
$$S_g(w, s) := \text{sign}(w) \cdot \text{ReLU}(|w| - g(s))$$

```python
class STR(PruningLayer):
    def __init__(self, config, layer, out_size):
        super(STR, self).__init__()
        self.config = config
        threshold_size = get_threshold_size(config, out_size, layer.weight.shape)
        self.s = nn.Parameter(torch.ones(threshold_size) * -self.config.threshold_init)
        self.g = torch.sigmoid

    def forward(self, weight):
        """
        sign(W) * ReLu(|W| - g(s))
        """
        mask = self.get_mask(weight)
        return torch.sign(weight) * mask.view(weight.shape)

    def get_mask(self, weight):
        return torch.relu(torch.abs(weight).view(weight.shape[0], -1) - self.g(self.s))
```

Triggers

- A lot of data from particle collisions. ATLAS has data volume of over 60TB/s
- Use triggers to save only relevant data. Has to be quick, but sensitive enough to signs of rare processes
- Selection based on heuristics such as energy, charge, direction, momentum
- Use neural networks to get better results than traditional rule-based methods?
- A good neural network means nothing if it cannot be run efficiently in hardware

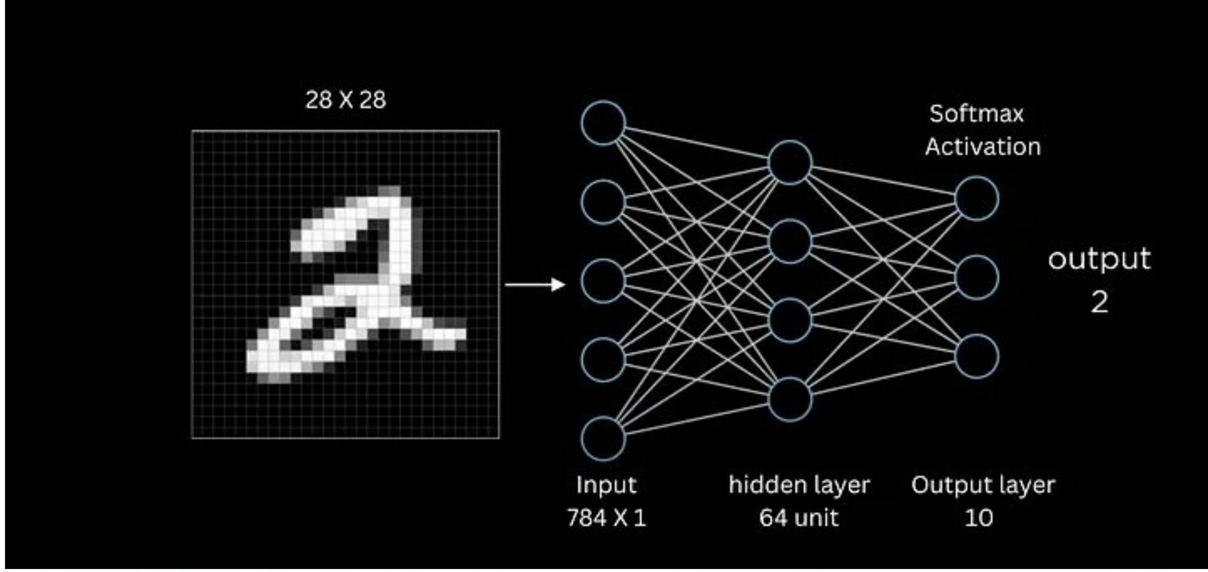


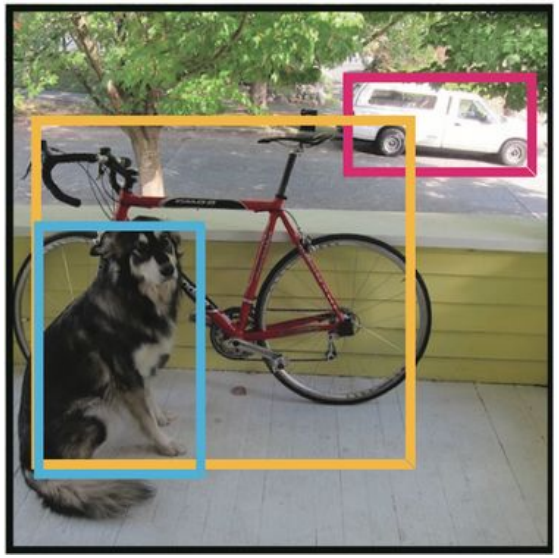


Neural networks

- **Neural networks learn from data**

- **Can be trained to do tasks such as classification, text, image or video generation, regression, pattern recognition**





Neural networks

- **Everything is numbers**

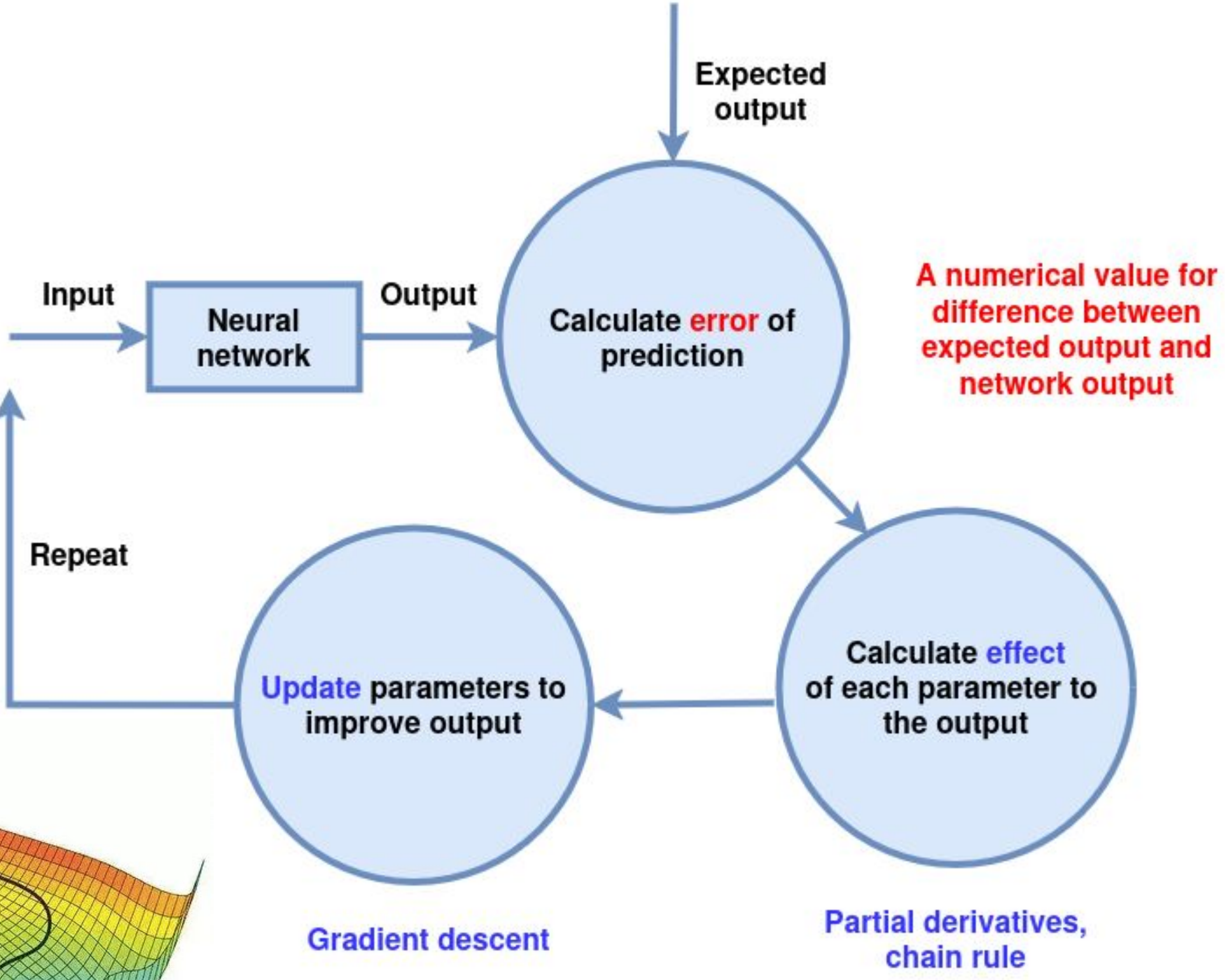
- **Neural networks produce an output by doing mathematical operations such as multiplication and addition**

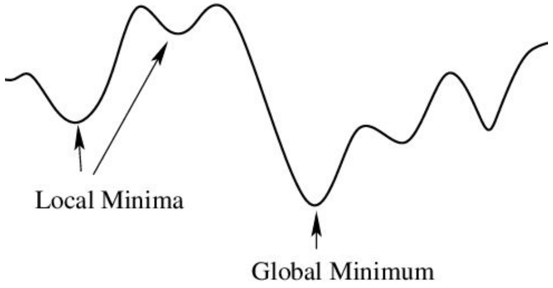
104	105	105	105	105	105	105	105	105	105	105	105	105	105	191	191	191	191	191	191	191	191	191
105	105	105	105	105	105	105	105	105	105	105	105	105	104	102	157	184	191	191	191	191	191	191
104	104	104	83	92	105	105	104	104	105	105	105	105	105	105	104	107	191	191	191	191	191	191
87	84	84	84	83	87	83	83	83	104	105	105	105	105	105	105	107	101	191	122	105	105	105
84	84	84	84	84	90	84	84	83	104	105	105	105	105	105	105	104	104	105	105	105	105	105
84	84	84	84	84	84	84	84	84	104	105	105	105	105	105	105	105	105	105	105	105	105	105
84	84	84	84	84	84	84	84	97	105	105	105	105	105	105	105	105	105	105	105	105	105	105
91	84	84	84	83	83	83	120	105	95	104	105	104	104	104	112	117	203	203	59	57	57	57
102	84	84	84	101	104	168	168	104	79	79	115	115	104	104	102	74	206	205	60	93	115	57
104	105	89	104	111	115	167	167	116	79	80	61	61	104	104	102	61	61	61	61	61	60	100
115	116	93	116	116	167	168	167	60	80	80	61	60	104	59	71	61	61	61	61	61	61	61
116	116	85	60	59	168	168	168	56	79	80	78	61	61	61	61	61	61	61	61	61	61	61
61	61	61	61	71	168	168	168	97	80	79	60	61	61	61	61	61	61	61	61	61	61	61
61	61	61	61	57	168	168	168	107	80	80	63	61	61	61	61	61	61	61	61	61	61	61
61	61	61	61	74	168	168	168	91	80	80	80	60	61	61	61	61	61	61	61	61	61	61
61	61	61	61	72	168	168	168	146	77	80	79	60	61	61	61	61	61	61	61	61	61	61
61	61	61	61	56	168	168	168	170	60	79	79	60	61	61	61	61	61	61	61	61	61	61
61	61	61	61	56	168	122	167	168	60	79	80	61	61	61	61	61	61	61	61	61	61	61
61	61	61	61	71	167	61	160	168	79	59	79	60	61	61	61	61	61	61	61	61	61	61
61	61	61	61	162	167	60	78	168	60	66	70	61	61	61	61	61	61	61	61	61	61	61
61	61	61	61	166	66	61	60	168	60	79	61	61	61	61	61	61	61	61	61	61	61	61
61	61	61	61	170	60	61	60	170	141	56	61	61	61	61	61	61	61	61	61	61	61	61
61	61	61	61	167	167	60	61	157	62	59	61	61	61	61	61	61	61	61	61	61	61	61
61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61
61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61
61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61
61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61

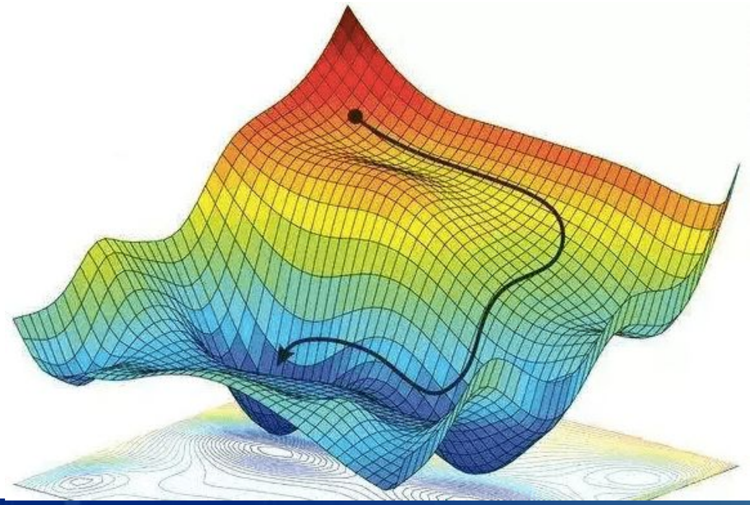
Neural networks

- In the beginning, the neural network produces outputs that make no sense

- During training the neural network learns to produce better outputs





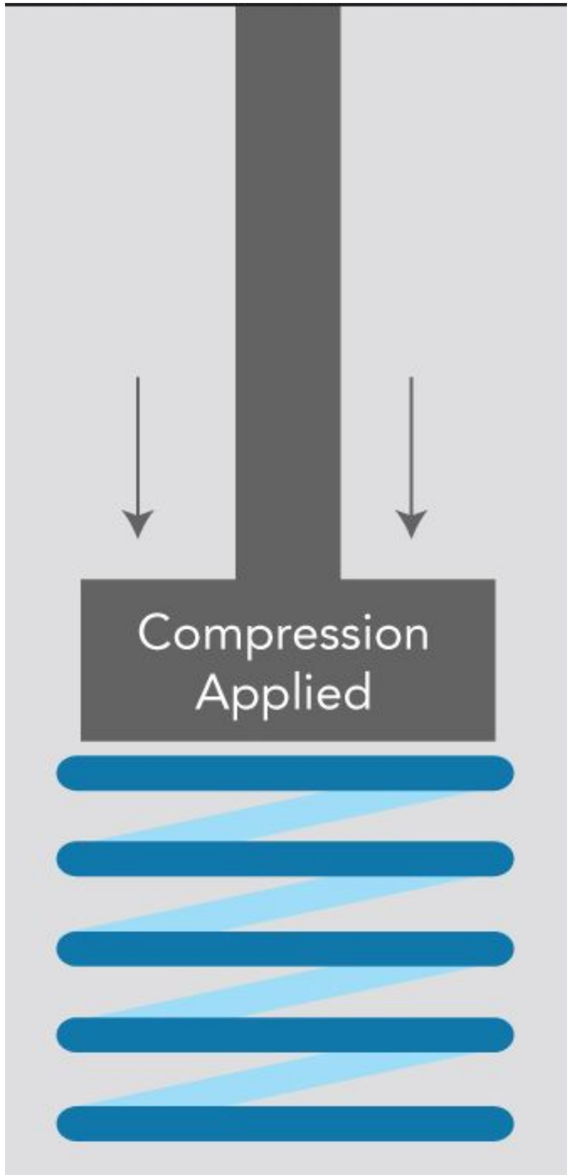


Neural networks

- Bigger neural networks can be more powerful, but slow

- In limited-resource or low-latency environments, this won't do.

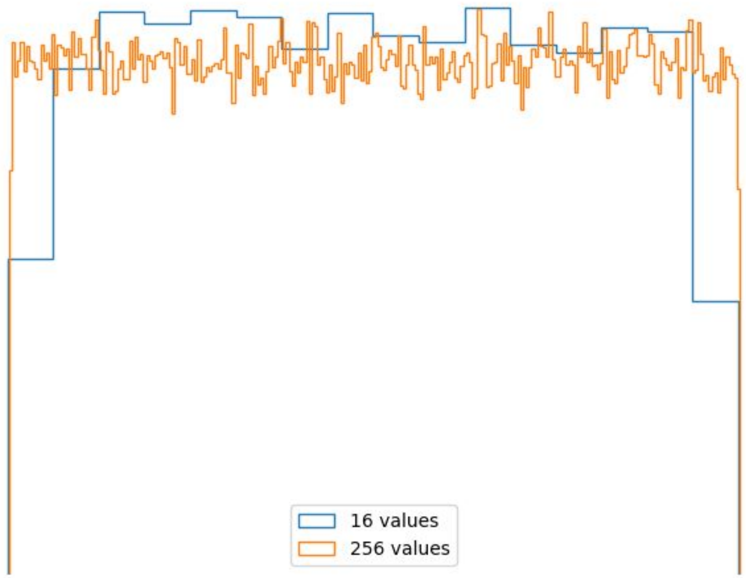
- Use compression to make them faster and use fewer resources



Compressing neural networks

- **Restrict parameters to be one of 2,4,16,256 etc. values**

- **Teach neural network to work with fewer parameters**



64 112 991 = 11110100100100100101011111 = 26 bits
25 812 = 110010011010100 = 15 bits
54 = 110110 = 6 bits

min	-1
max	0.9921875
mean	0.03131510416666667
std	0.2122100147954903
sparsity	91.7%

99	93	1	•	85	79	55	$= 99 \times 85 + 93 \times 79 + 1 \times 55 = 15817$
99	93	0	•	85	79	55	$= 99 \times 85 + 93 \times 79 + 0 \times 55 = 15762$

Optimizing (compressed) neural networks

- Hyperparameters: a set of parameters that configure the neural network architecture and how it is trained

- Neural networks are a black box. Training can take a long time. Have to wait until training ends to see how well a set of hyperparameters work → automate

- With compressed neural networks that run directly on hardware, have to also consider requirements by hardware

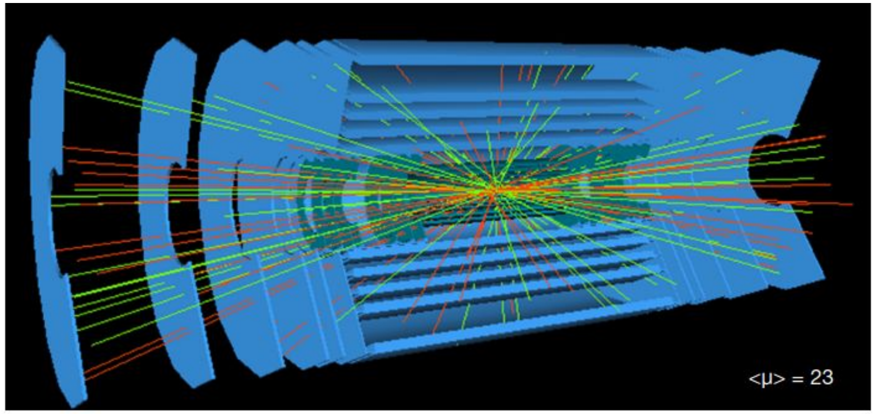
AI use cases

Track reconstruction

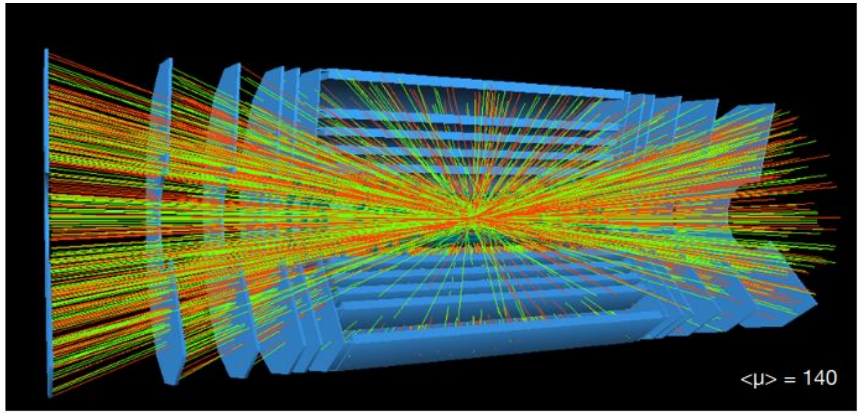
Offline:
- spacepoint formation
- track seeding
- track following
- track fitting

Online:
- pattern recognition
- latency O(10)us

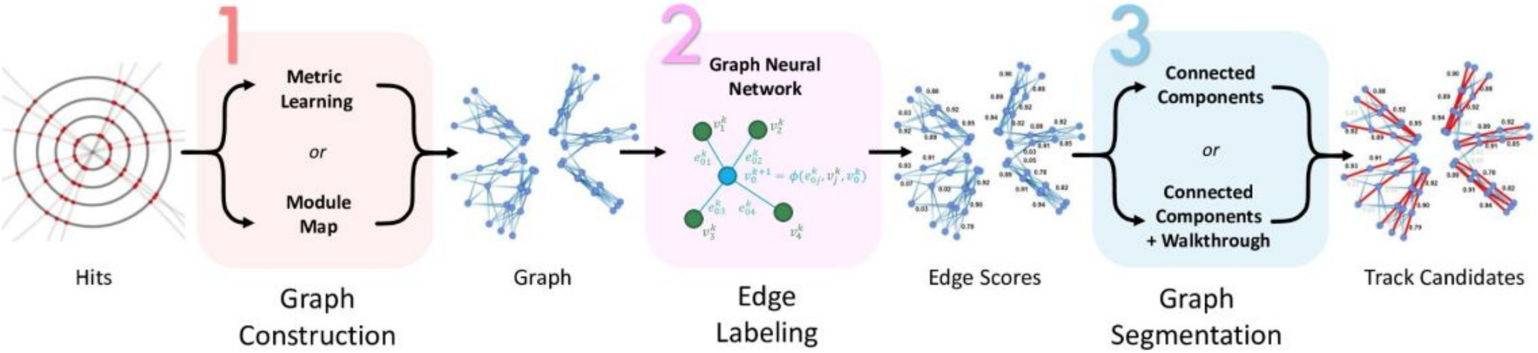
Neural networks



Current environment inside ATLAS at LHC



Expected environment inside ATLAS at HL-LHC



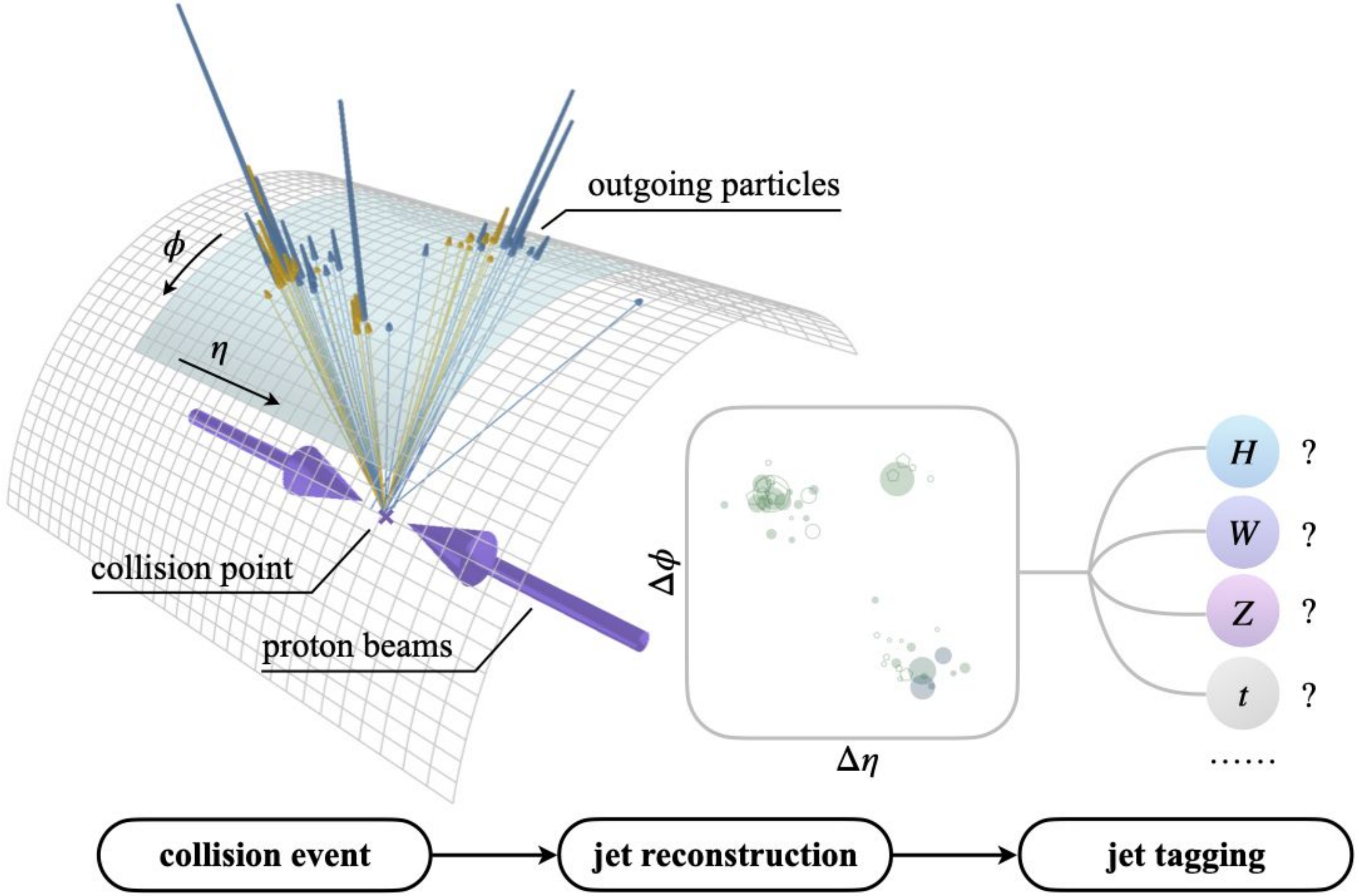
Jet tagging

Identify the type of the particle that initiates the jet.

Challenge: particles can radiate, radiated particles produce more particles

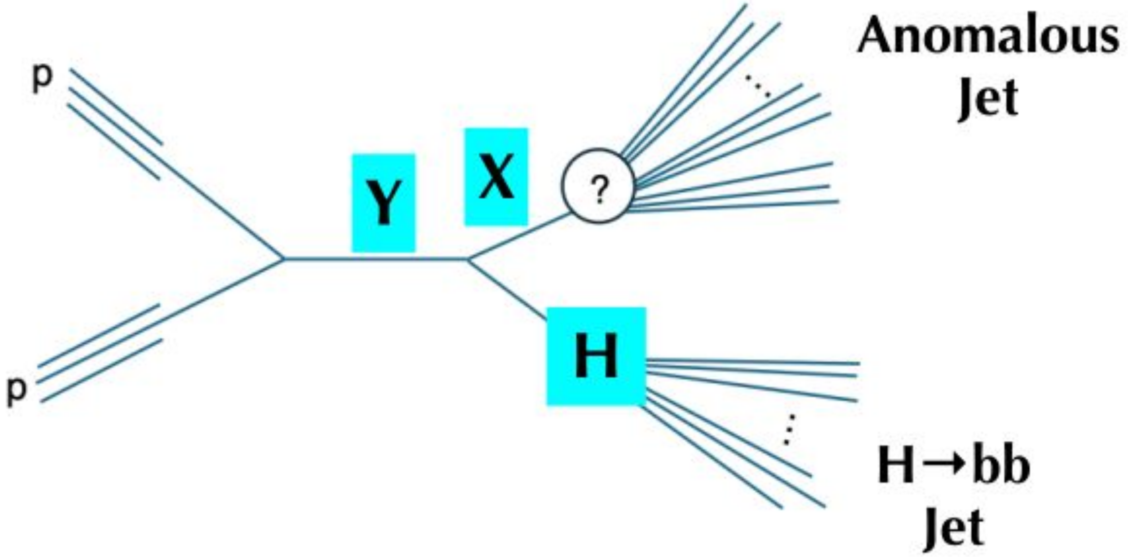
Neural network:

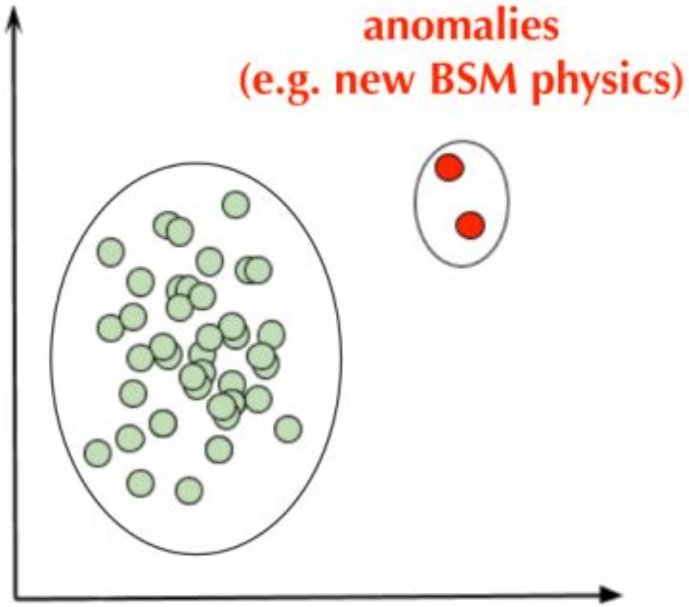
- **use measured particle properties and particle pair interactions to identify particle types**

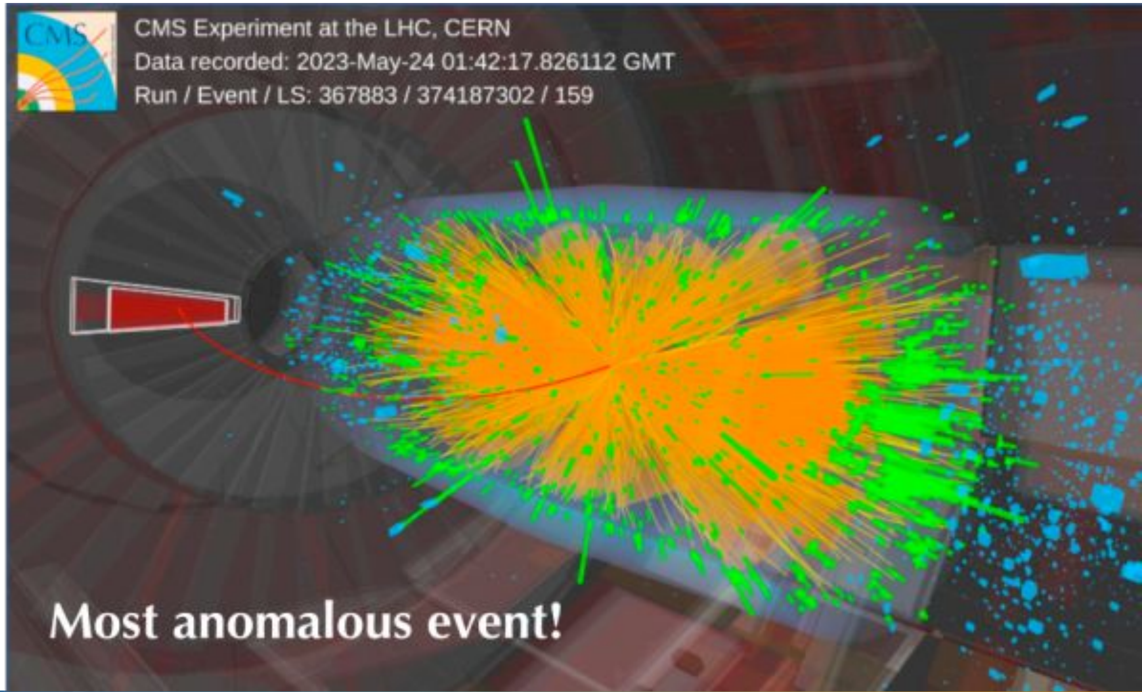


Anomaly detection

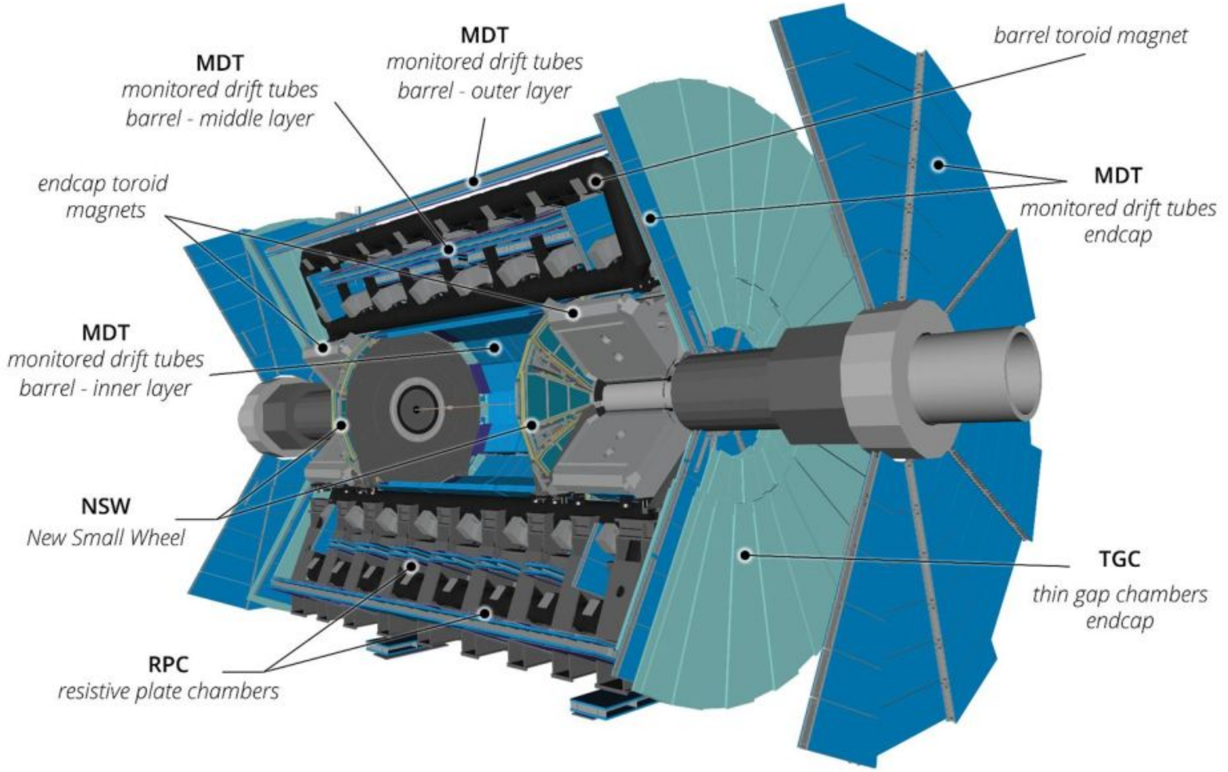
- What if the trigger discards events that show new physics?
- Same input as Global Trigger, has to run in 50ns
- Find events that are very unusual





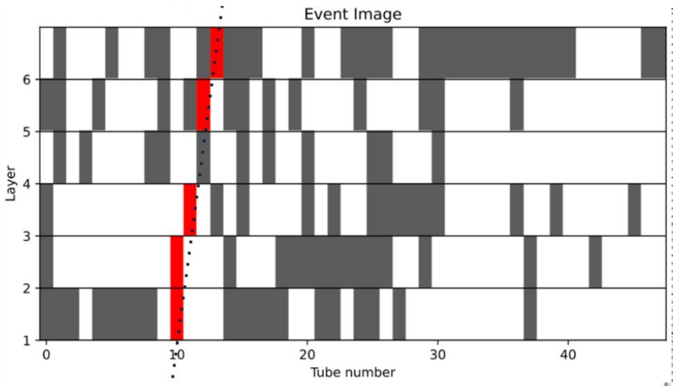


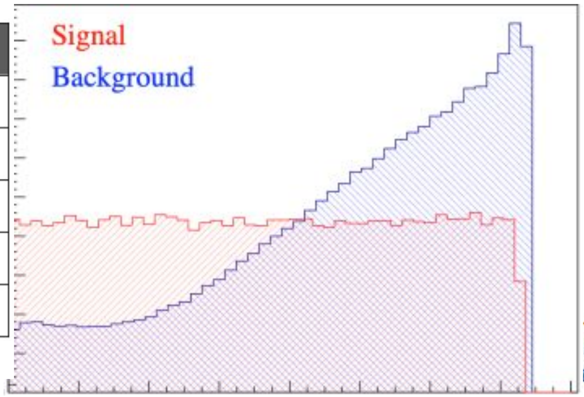
L0 Muon trigger

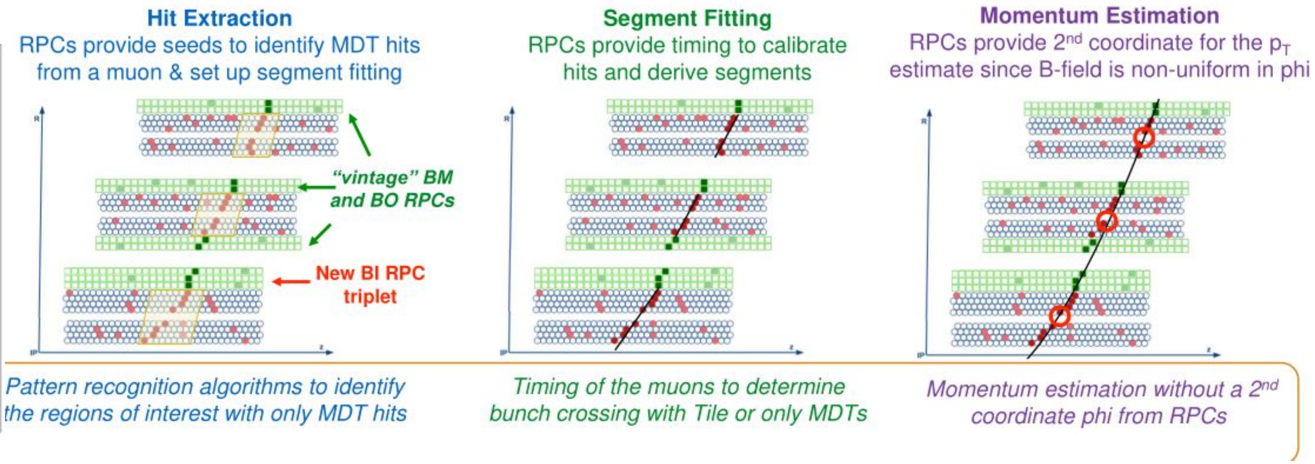


For cases with reduced RPC performance:

- Pattern recognition neural networks, distinguish muon hits from backgrounds
- Momentum estimation





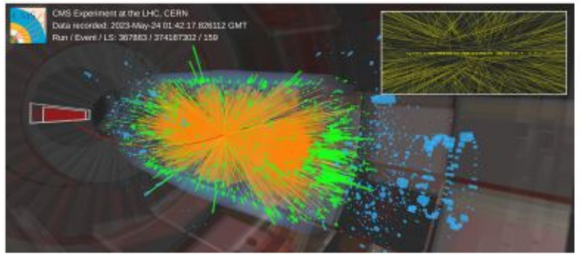


Additional

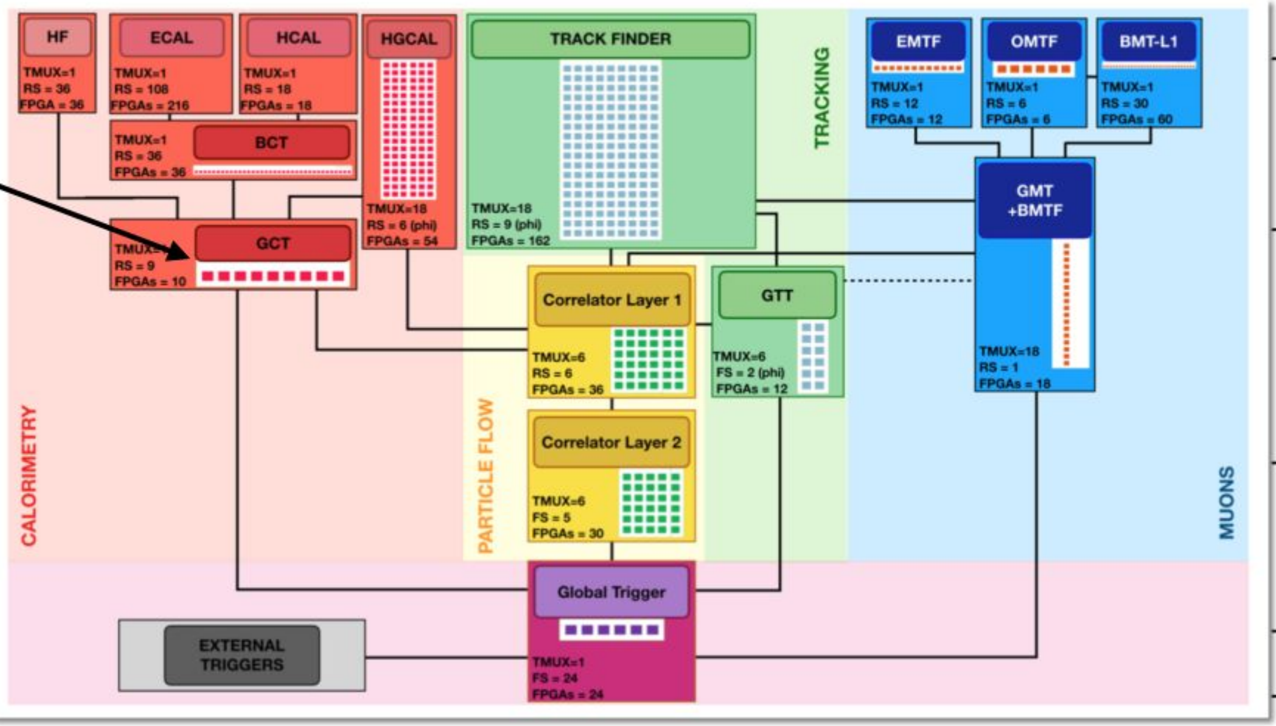
AI where?

t = 0

1 small box = 1 FPGA board with AMD VU13P FPGA



t < 12.5 μs



0 μs	**Detector hits**
5 μs	**Clusters & Tracks**
6 μs	**Particles**
7 μs	**Event Categorisation**
8 μs	**1 bit: keep / discard**