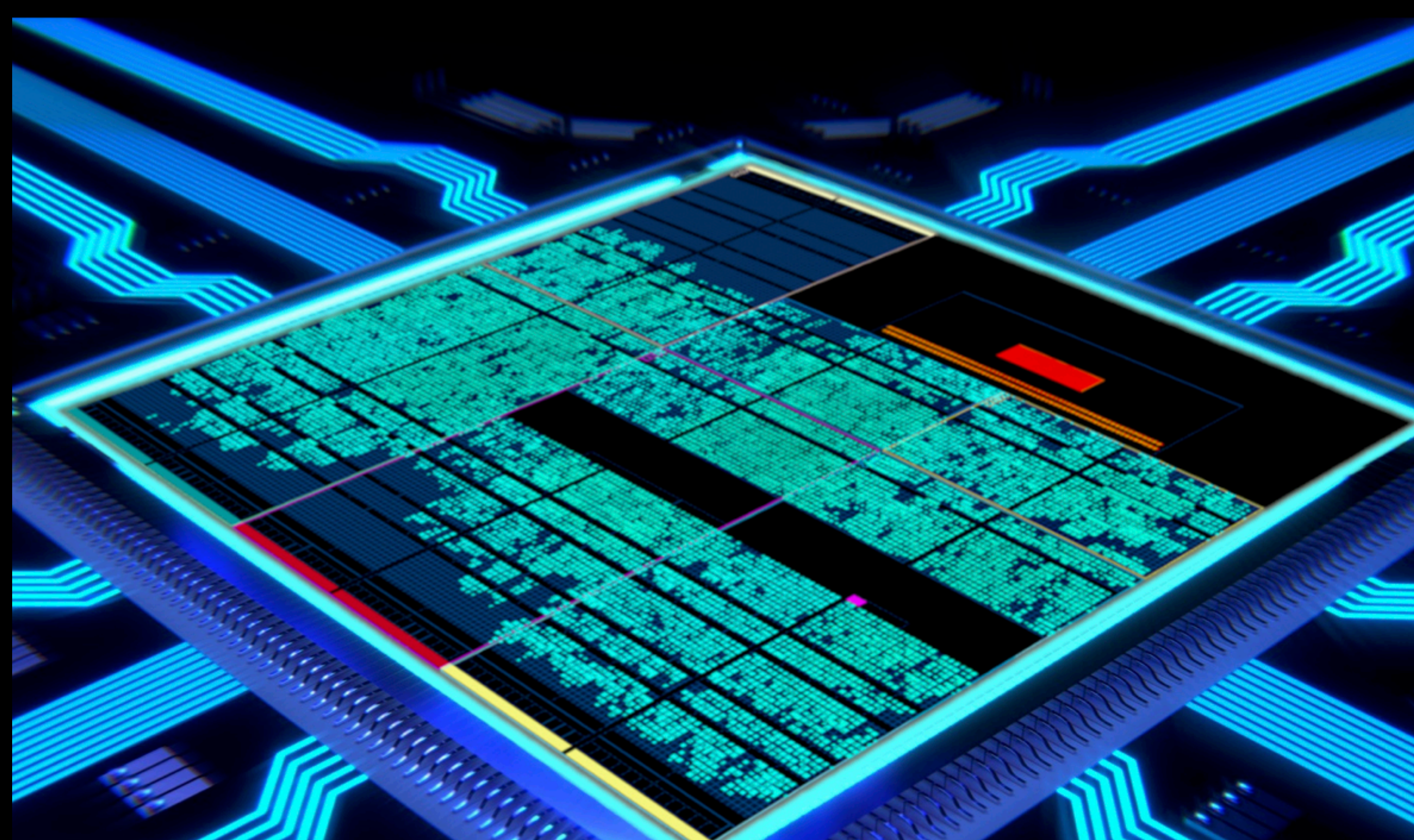
Exploring Data Challenges and Leveraging Codabench

A Practical Journey With Unsupervised New Physics Detection at 40 MHz



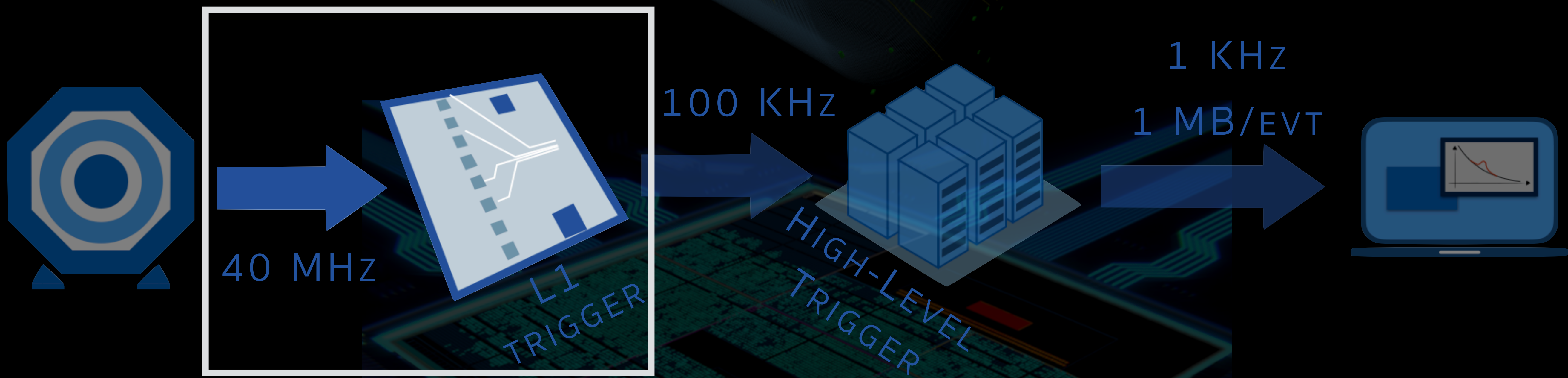
Katya Govorkova
MIT

29 November 2023
AISSAI

Unsupervised New Physics detection at 40 MHz Challenge

Idea is to **look for** something **very rare and unusual** directly in the **Level-1 Trigger** without any signal hypothesis in mind

The challenge is to find a-priori **unknown** and **rare New Physics** hidden in a data sample dominated by ordinary Standard Model processes



The deliverable is a developed **algorithm** that can be deployed and run **in L1** with strict **latency** requirement of **< 1 microsecond**

The task is therefore to design an architecture that maximises the **sensitivity for New Physics** but at the **lowest** possible **resource** and **latency** budget

Unsupervised New Physics detection at 40 MHz

Step-by-Step

- **Gather** in a team/by yourself
- **Get a cool name** for your team, for example team " **DeepAnomaly** " 👨‍👨‍👦
- Get yourself familiar with the **details on the challenge webpage**
- **Investigate** available **datasets** and **example codes**

SM Dataset: Training	BSM Dataset: A to 4 leptons	BSM Dataset: h^+ to τν
Blackbox Dataset	BSM Dataset: LQ to bτ	BSM Dataset: h^0 to ττ

- **Design** your AD **model**
- **Evaluate performance** and submit **results**
- **Best models** will be **published** in a White Paper (and perhaps deployed in L1 trigger of CMS!!)

Unsupervised New Physics detection at 40 MHz

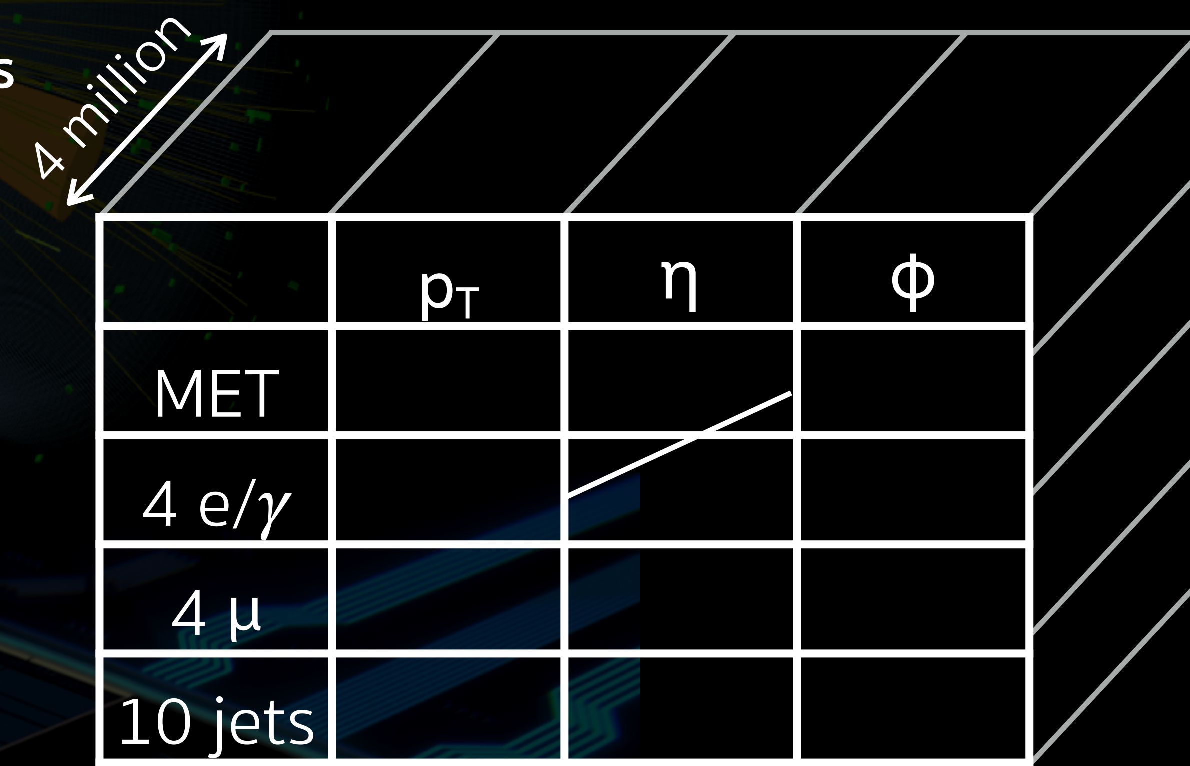
Data Samples

The data is represented as an array of MET, up to 4 e/γ, 4 μ and 10 jets each described by p_T, η and ϕ to **mimic L1 data format**

Train using provided **4 million background-like events** simulated with Delphes 💾

Events are **pre-filtered** to have **at least one lepton**

- Inclusive W production, with W → lν (59.2%)
- Inclusive Z production, with Z → ll (6.7%)
- tt production (0.3%)
- QCD multijet production (33.8%)



Evaluate performance on several different **New Physics simulated samples**

- Neutral scalar boson A, 50 GeV → 4 l 💾
- Leptoquark, 80 GeV → b τ 💾
- Scalar boson, 60 GeV → τ τ 💾
- Charged scalar boson, 60 GeV → τ ν 💾
- Black Box 🎁

We choose a **low-mass, rare signals** that would pass pre-feltiring

Overview and Timeline of the Challenge

2020 — we had a dataset that represents data in the L1 trigger of CMS and we started working on an algorithm for the L1 trigger

February 2021 — developed an unsupervised Anomaly Detection (AD) algorithm
The idea was to release the datasets and set up a public challenge based on it

July 2021 — released the datasets paper on arXiv
July 2021 — published the datasets in Zenodo
July 2021 — set up a challenge using GitHub io https://mpp-hep.github.io/ADC2021/
July 2021 — announced the challenge at the ML4Jets2021

August 2021 — released the AD algorithm paper on arXiv

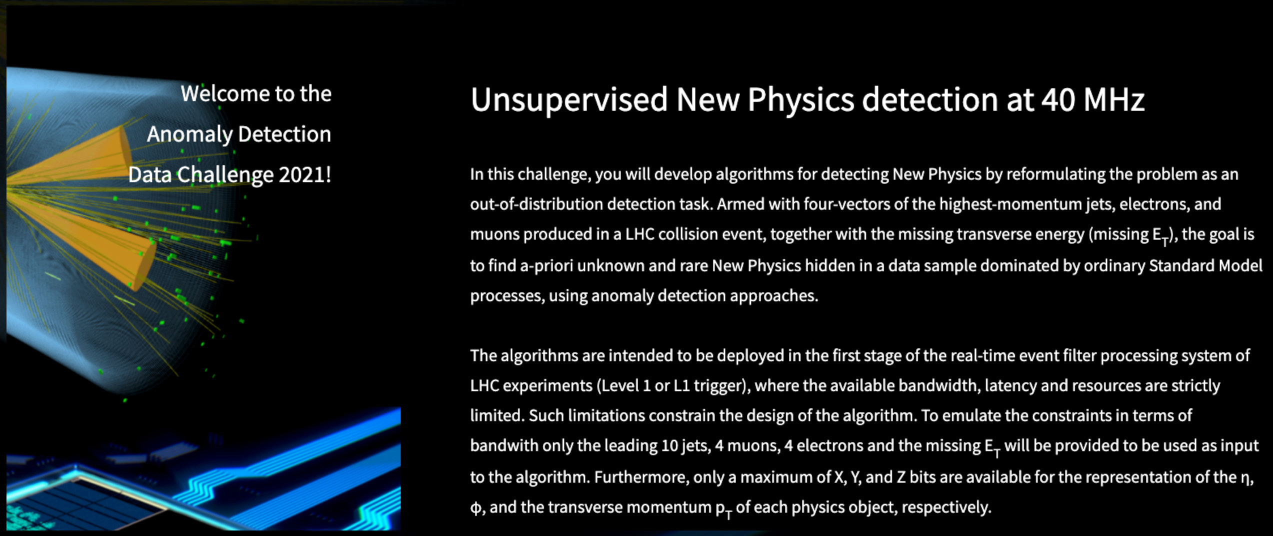
February 2022 — published the AD algorithm paper in nature Machine Intelligence
March 2022 — published the datasets paper in nature Scientific Data

2022/2023 — the data has been used to publish other research, but not to participate in the challenge

CHALLENGE SETUP

We provided

- A website with all the details https://mpp-hep.github.io/ADC2021/
- Datasets
  1. Signal h+ -> tau nu
  2. Signal h^0 -> tau tau
  3. Signal LQ -> b tau
  4. Signal A -> 4 leptons
  5. Training dataset
  6. BlackBox dataset

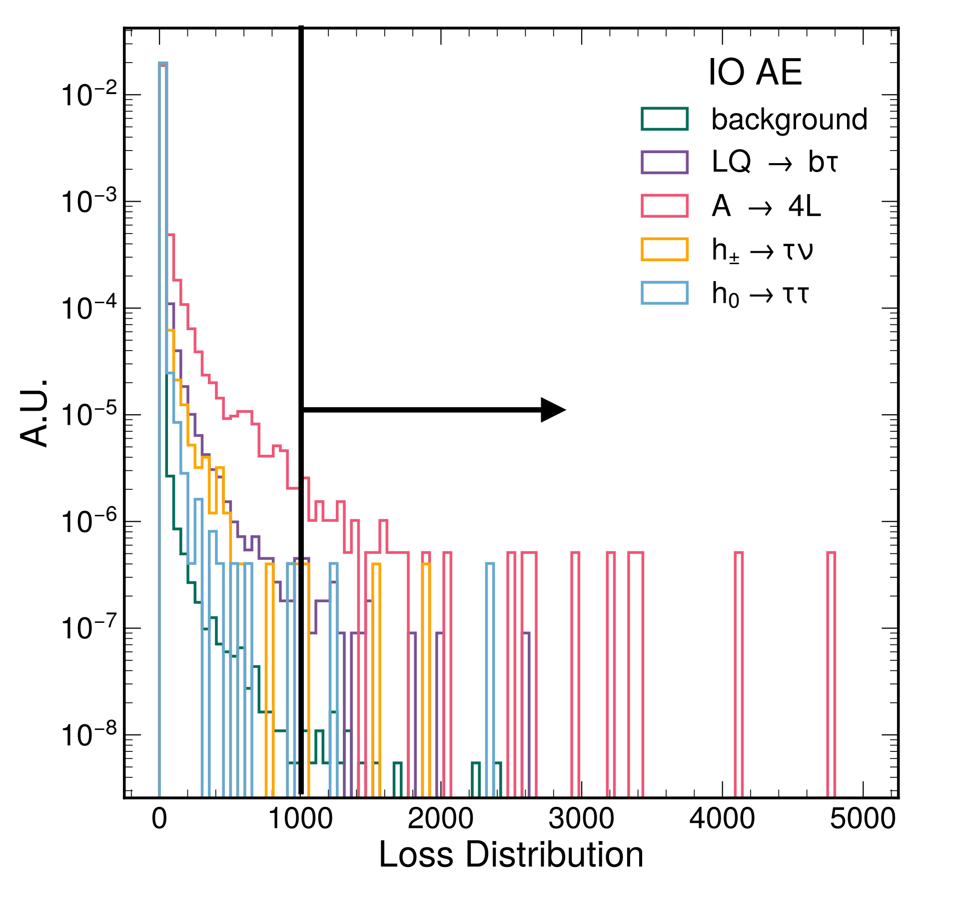


- A paper with details about the data
- Example algorithms
- Example analysis pipeline
- A script to estimate FLOPS
- A repo to upload the contributions to https://github.com/mpp-hep/ADC2021-results
- “Further reading” mostly about hls4ml and AD

UNSUPERVISED NEW PHYSICS DETECTION AT 40 MHZ

EXAMPLE TEAM " DEEPANOMALY " 👨‍👨‍👦

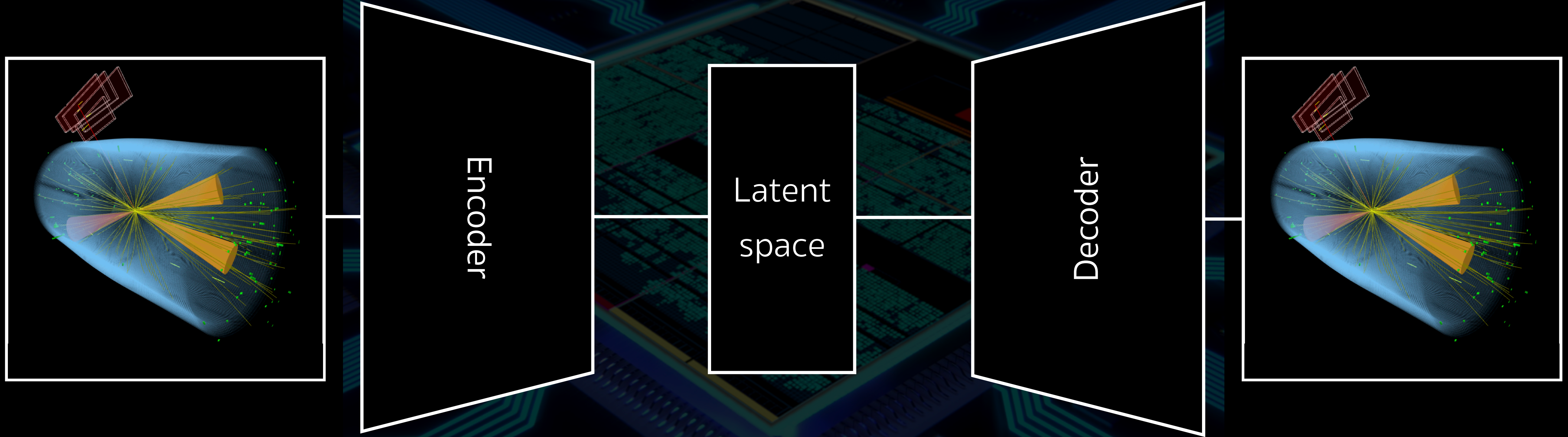
DESIGNS ALGORITHM



Decide on the **algorithm** that you want to explore

The Example Team " **DeepAnomaly** " has chosen **Autoencoders**

- **Encode** input in **smaller** dimensional **space**
- **Train** on typical LHC **background**
- **Anomalous** data will have **higher loss**
- **Calculating** the loss requires to **store the input** until the output is computed



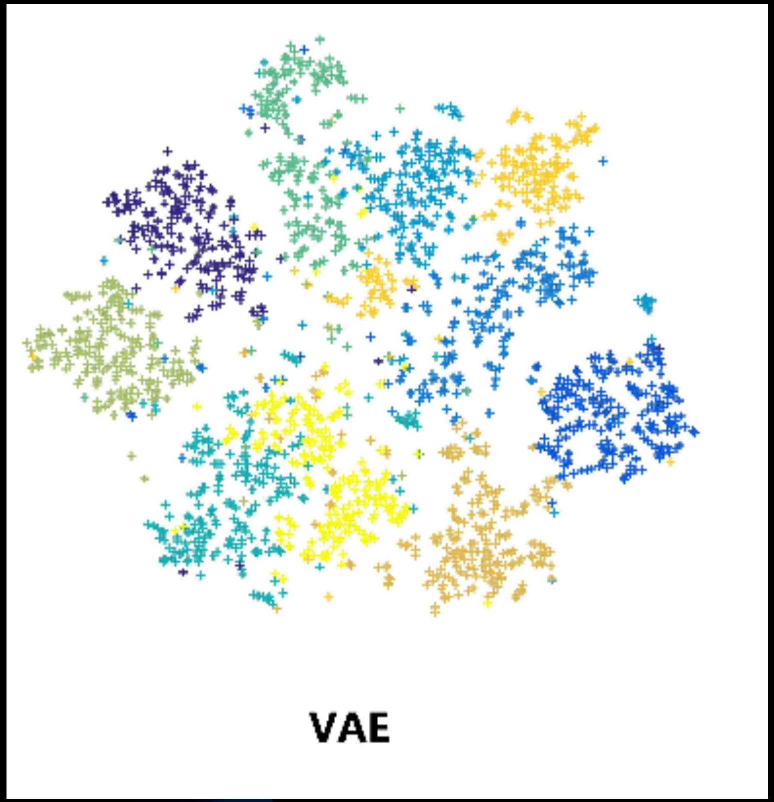
Unsupervised New Physics detection at 40 MHz

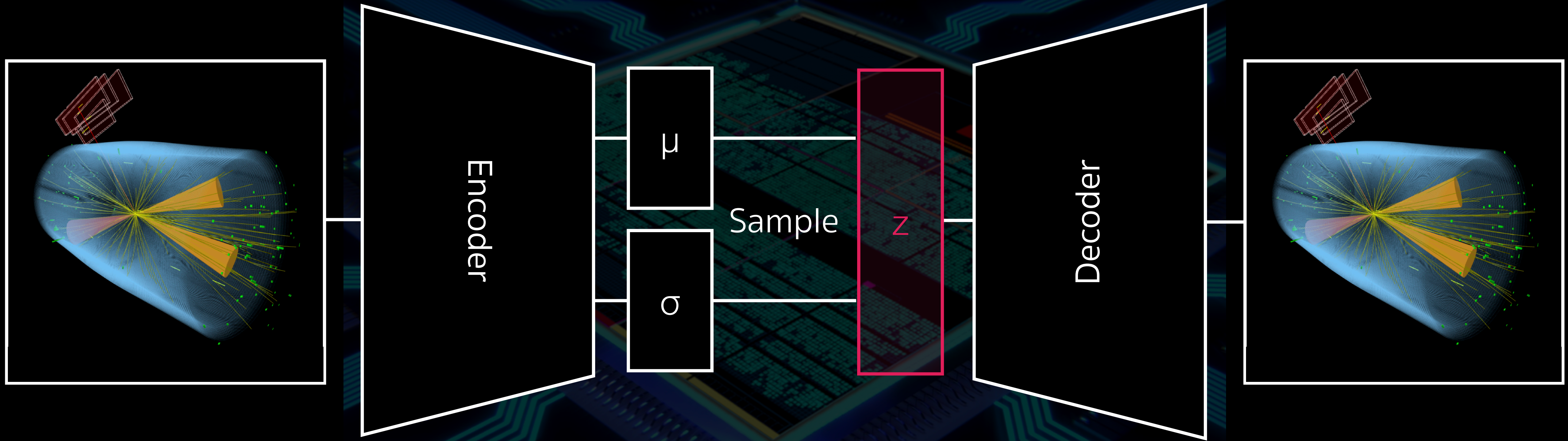
Example Team " DeepAnomaly " 👨‍👨‍👦

Designs algorithm

The " **DeepAnomaly** " team has also considers **Variational Autoencoders**

- **The latent space** is sampled from Encoder output
- Can be used **to generate new samples**
- **Inference** can be done only **on the latent space**
- **No** need to store **input** and deployment of **Encoder is enough**
  (e.g. saves resources and latency in comparison to AE)





Unsupervised New Physics detection at 40 MHz

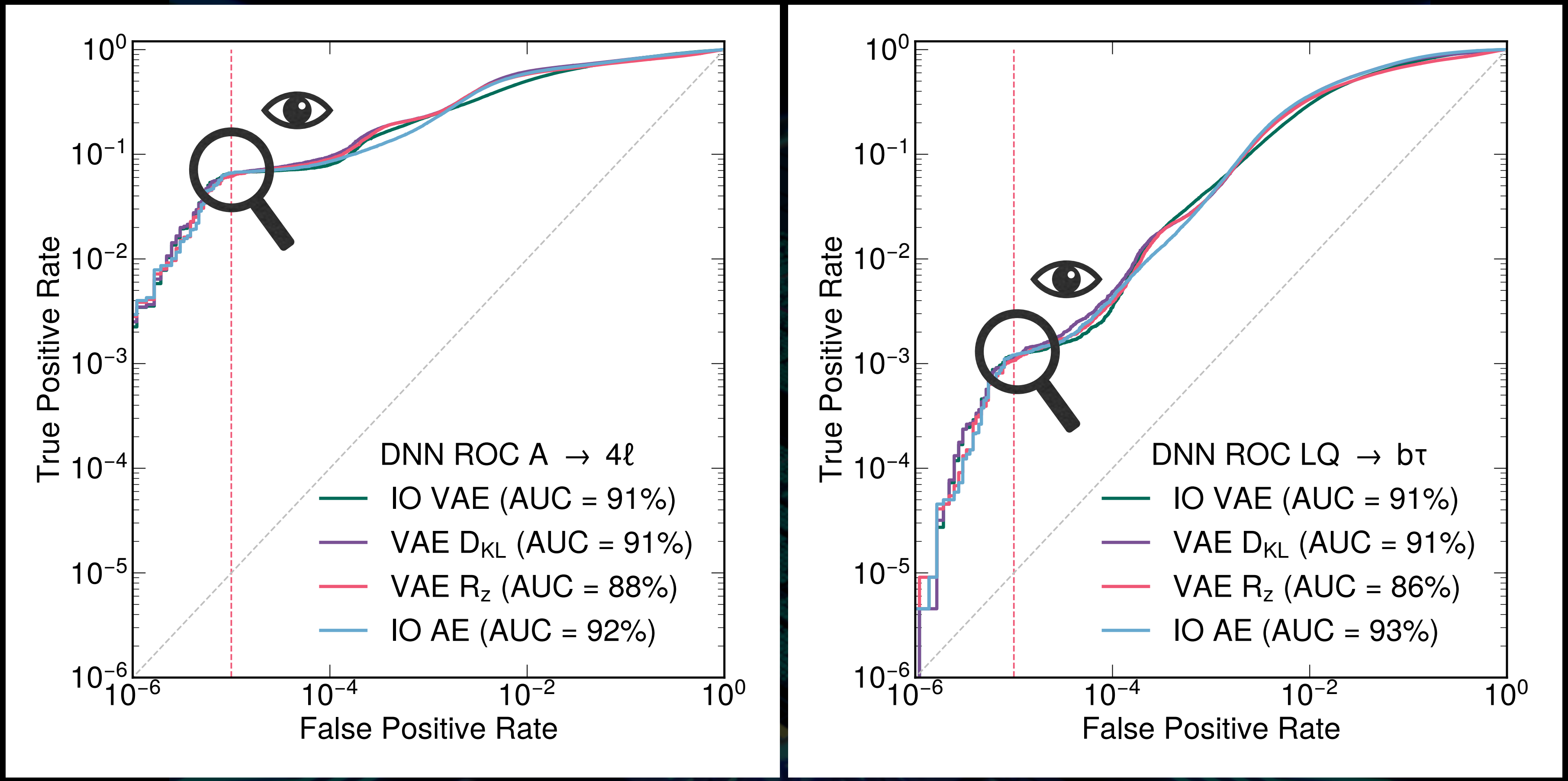
Example Team " DeepAnomaly " 👨‍👨‍👦

Evaluates Performance on Signal samples

Goal is to **maximise** TPR at FPR 10^{-5} (roughly corresponding to the available output data rate budget for a trigger algorithm) for each of the provided anomaly

The Team " **DeepAnomaly** " checks AE vs VAE

- The **Inference** can be done only **on the latent space,** either with D_{KL} or R_z
  Shoutout to DarkMachines for the R_z idea
- **No** need to store **input** and deployment of **Encoder is enough**



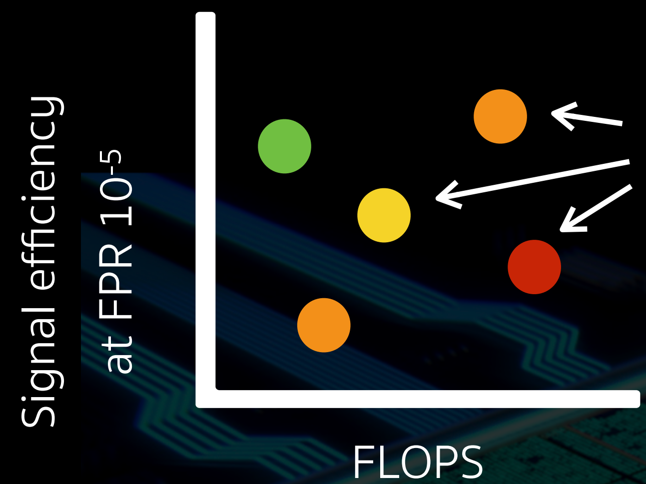
Unsupervised New Physics detection at 40 MHz

Example Team " DeepAnomaly " 👨‍👨‍👦

Submits results

- For the **Blackbox** dataset get the event number of the **1000 most anomalous** events based on the algorithm metrics

- An estimate of the **algorithm efficiency** can be obtained by calculating **the floating-point operations per second (FLOPS)** 🔗



Different algorithm performance

The best one is chosen by min(FLOPS) & max(sig eff)

- **The submission** should be in a form of a HDF5 file, DeepAnomaly.h5**,** containing a **numpy array** with the **identification numbers** of each selected event, plus a **dictionary** with the algorithm deployment **performance** 🔗

Upload contribution!

Unsupervised New Physics detection at 40 MHz

Example Team " DeepAnomaly " 👨‍👨‍👦

Submits results

- For the **Blackbox** dataset get the event number of the **1000 most anomalous** events based on the algorithm metrics

- An estimate of the **algorithm eff**
point operations per second

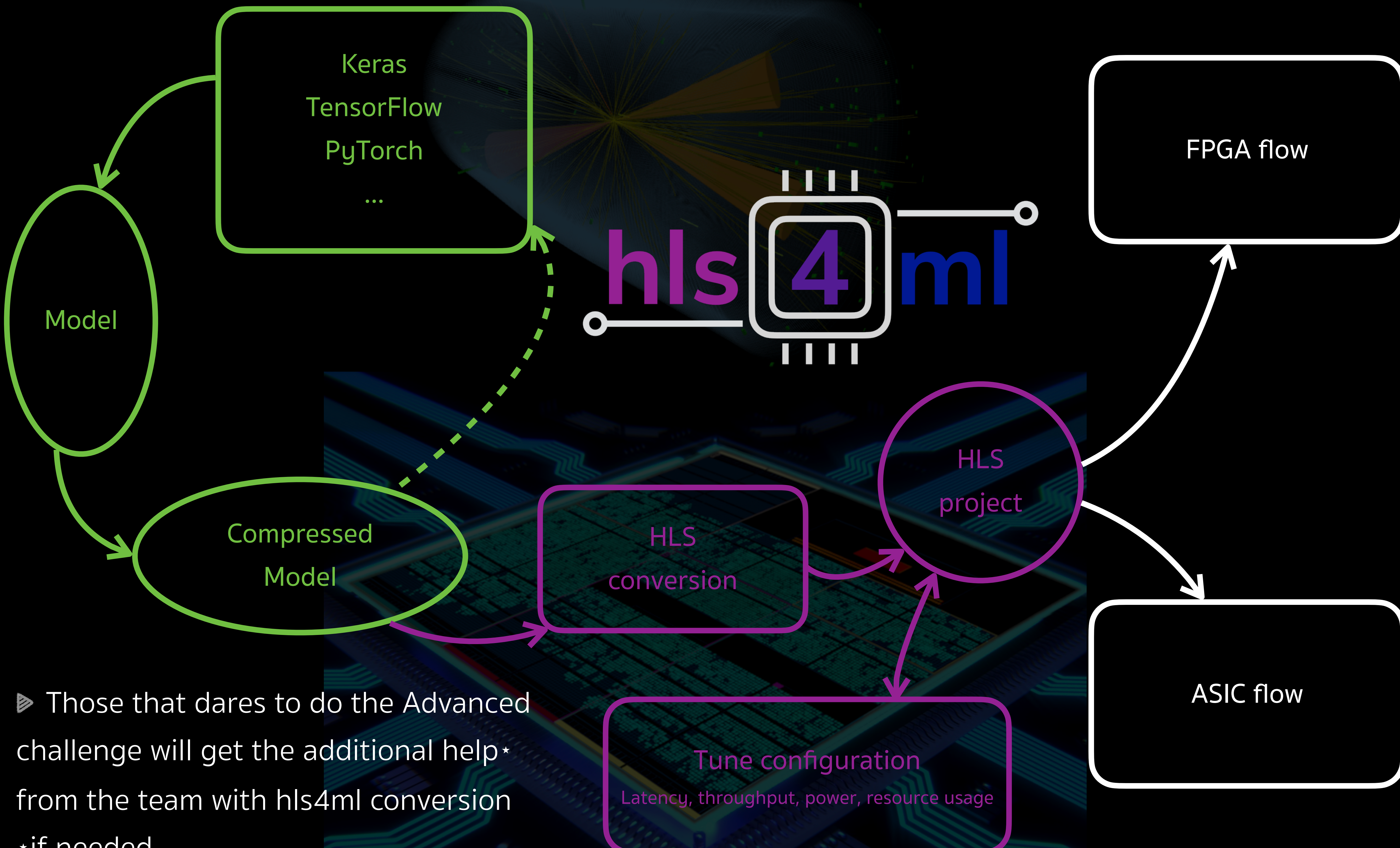
- **The submission** should be in a for
numpy array with the **identification numb** dictionary
with the algorithm deployment **performance** 🔗

We only accepted **the final submissions**, while many were asking us for an **intermediate** result to **iteratively improve** the algorithm

Upload contribution!

Unsupervised New Physics detection at 40 MHz

✪ Advanced Challenge

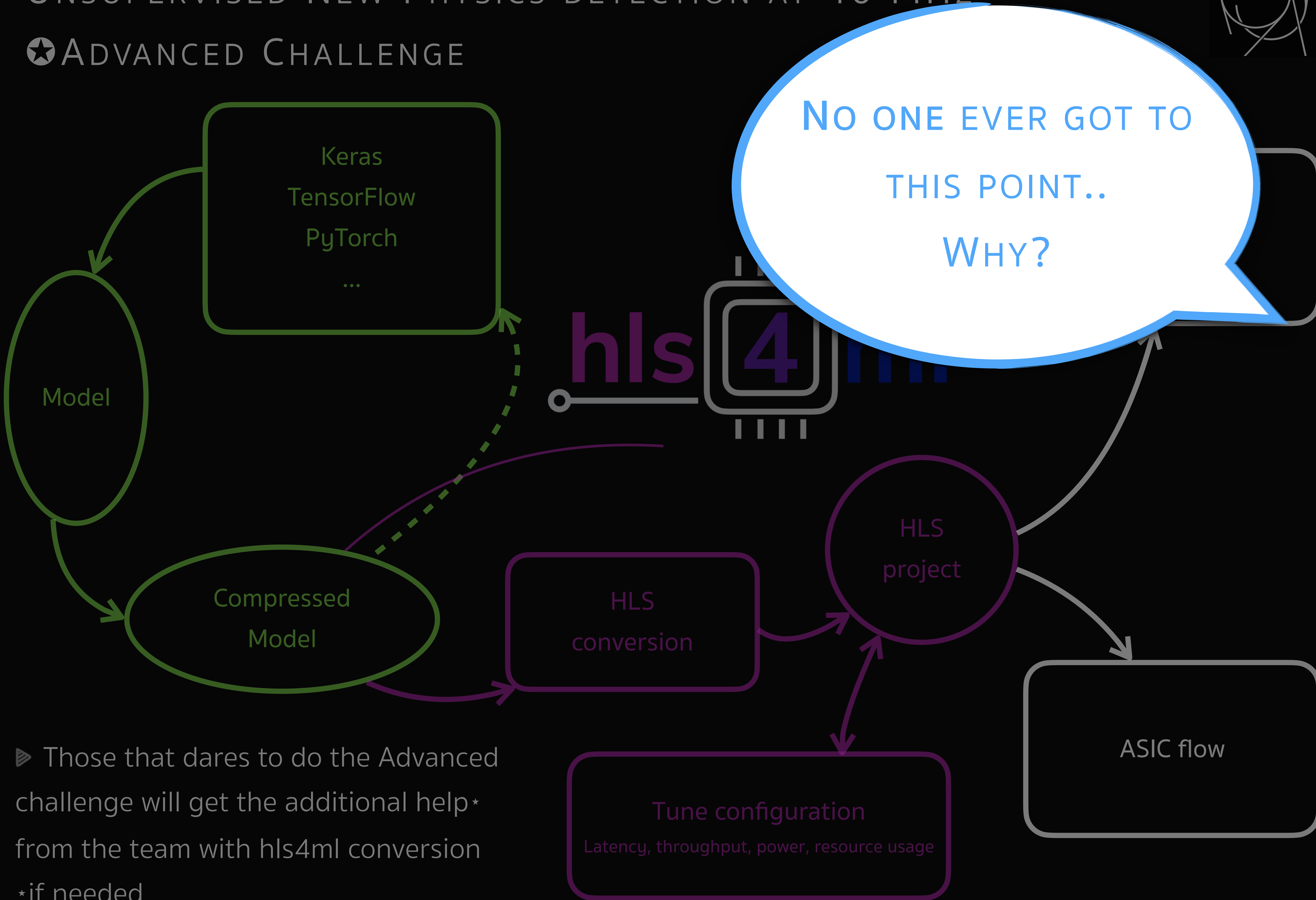


- Those that dares to do the Advanced challenge will get the additional help* from the team with hls4ml conversion

*if needed

Unsupervised New Physics detection at 40 MHz

✪ Advanced Challenge



- Those that dares to do the Advanced challenge will get the additional help* from the team with hls4ml conversion

*if needed

✪ Advanced Challenge

Keras
TensorFlow
PyTorch
...

Model

hls4ml

No one ever got to this point..
Why?

Most likely because it is a **very time consuming** task.
Perhaps, if we provided an **intermediate leaderboard** with results, participants would be **more interested** to invest their time

ASIC flow

- Those that c
challenge
from team with hls4

⋆if needed

Outcomes

There was only one official submissions in the GitHub repo

But several papers published using the dataset we provided

Anomalies, Representations, and Self-Supervision

Barry M. Dillon, Luigi Favaro, Friedrich Feiden, Tanmoy Modak, Tilman Plehn

Online-compatible Unsupervised Non-resonant Anomaly Detection

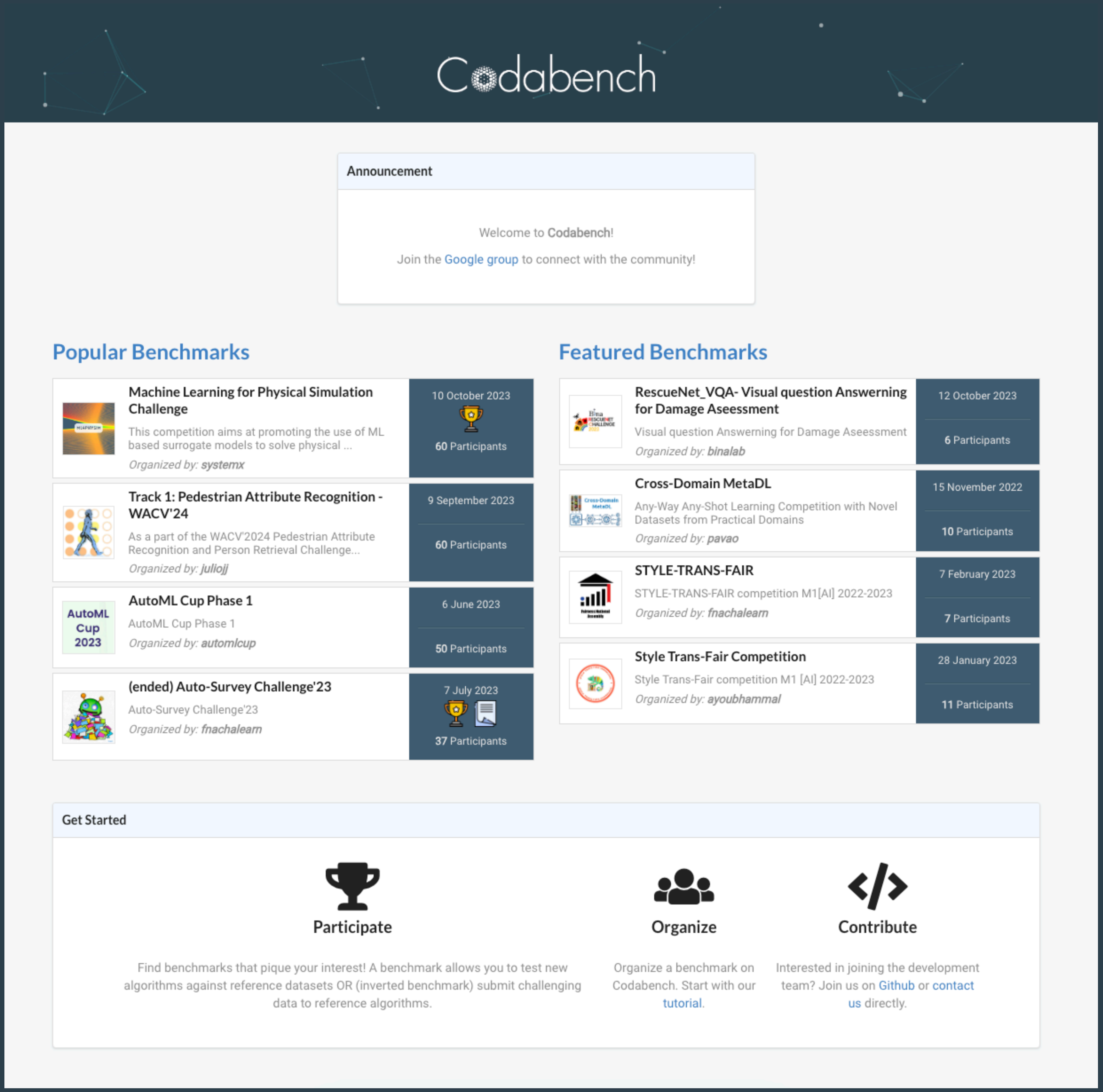
Vinicius Mikuni, Benjamin Nachman, David Shih

Nanosecond anomaly detection with decision trees for high energy physics and real-time application to exotic Higgs decays

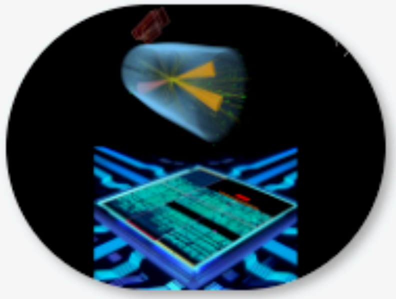
Stephen Roche, Quincy Bayer, Benjamin Carlson, William Ouligian, Pavel Serhiayenka, Joerg Stelzer, Tae Min Hong

CODABENCH.ORG

A platform to create your own challenges, see the talk by Wahid



Experience with codabench.org



CLASSIFY STANDARD MODEL EVENTS IN HEP DATASET

5 PARTICIPANTS

1 SUBMISSIONS

ORGANIZED BY: Katyag
CURRENT PHASE ENDS: Never
CURRENT SERVER TIME: 19 November 2023 At 15:53 CET
Docker image: codalab/codalab-legacy:gpu

I've made a test challenge to learn how to use Codabench

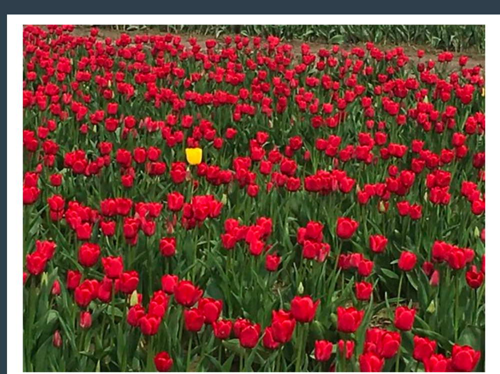
Things that are very useful:

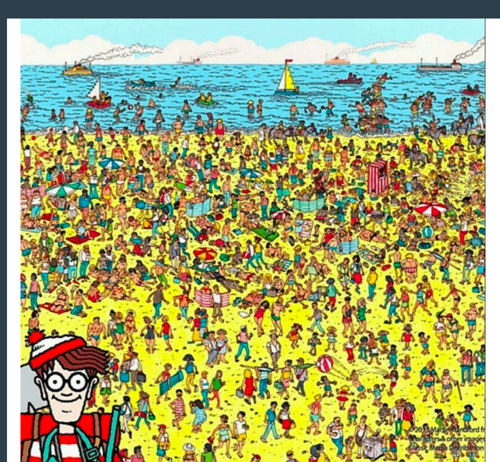
- ✓ Datasets+code examples in one place
- ✓ Clear deadline that is easily updated if need to
- ✓ Easy to reach out and to advertise the challenge to the communit
- ✓ Infinite submissions and automatic check of it
- ✓ Leaderboard
- ✓ Past challenges can be used as educational materials

Perhaps if our challenge was hosted on Codabench, we would get more submissions (and we could automate FLOPS estimation)

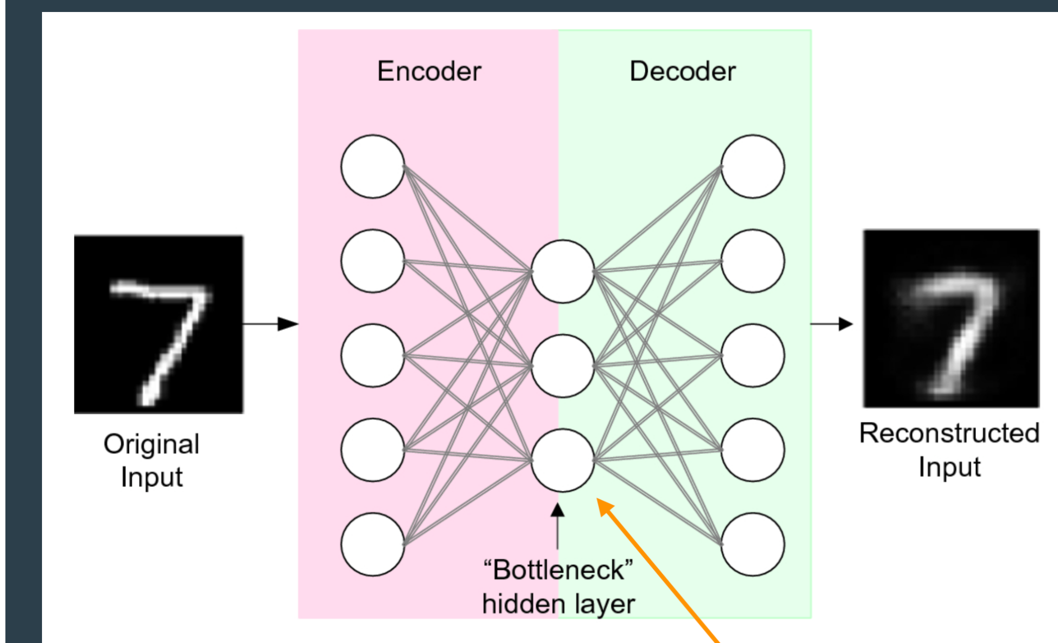
Future Anomaly data Challenges

"Can we make an anomaly detection algorithm capable of anomaly detection in many domains? "

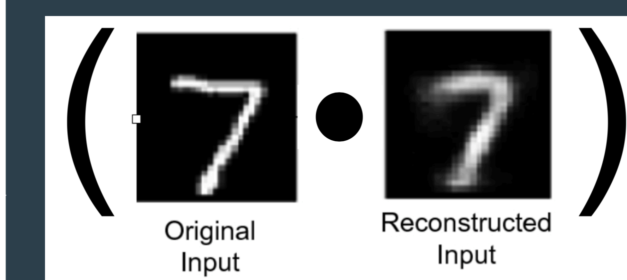




Many different datasets



Anomaly algorithm



Anomaly metric

From the slides of Phil Harris

Future Data Challenges

At **A3D3** we are looking to lead a challenge across **NSF HDR initiative** institutes on **anomaly detection** and would like to host this on Codabench.org

Possible Suggested datasets

- **Imageomics:** Images of animals for classification and new species
- **Iguide:** Rainfall data to predict anomalous local weather
- **iHARP:** Ice data from anomalous melts?
- **ID4:** new materials identification?
- **A3D3:** Astrophysical anomalies (LIGO/...)

And a PRIZE!! 10K$ 💸 for development

The plan is to write **two white papers**

1. A white paper on **making ML Workflows fair** not just the dataset, but we want the whole workflow reproducible
2. A second whitepaper **comparing and contrasting anomaly detection** on the different datasets