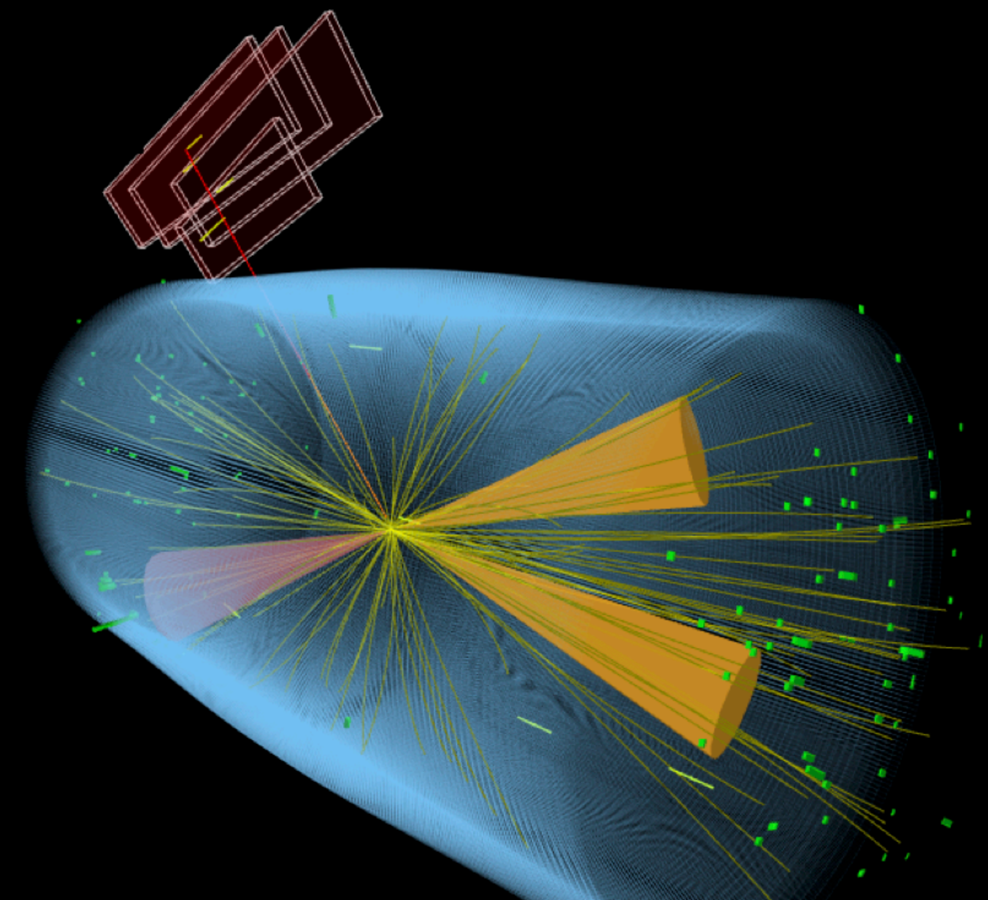
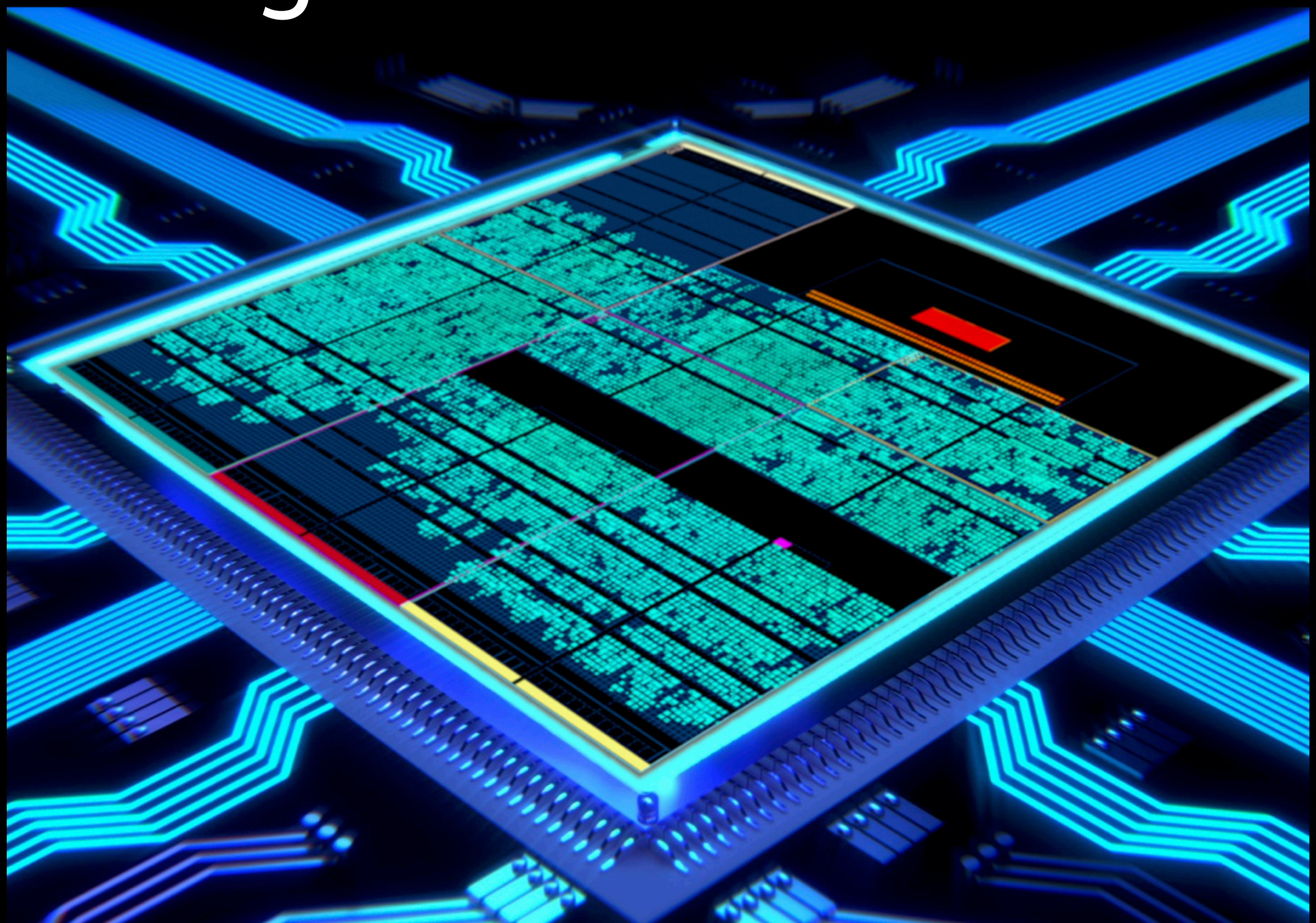


Anomaly Detection in HEP

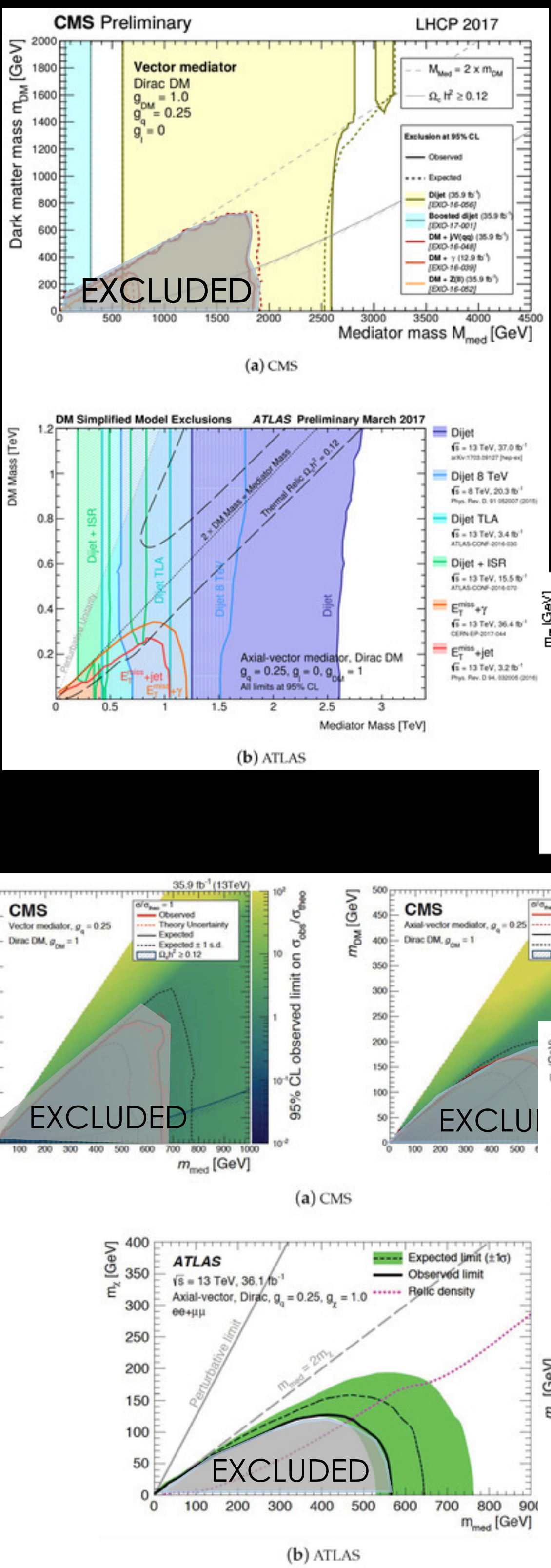


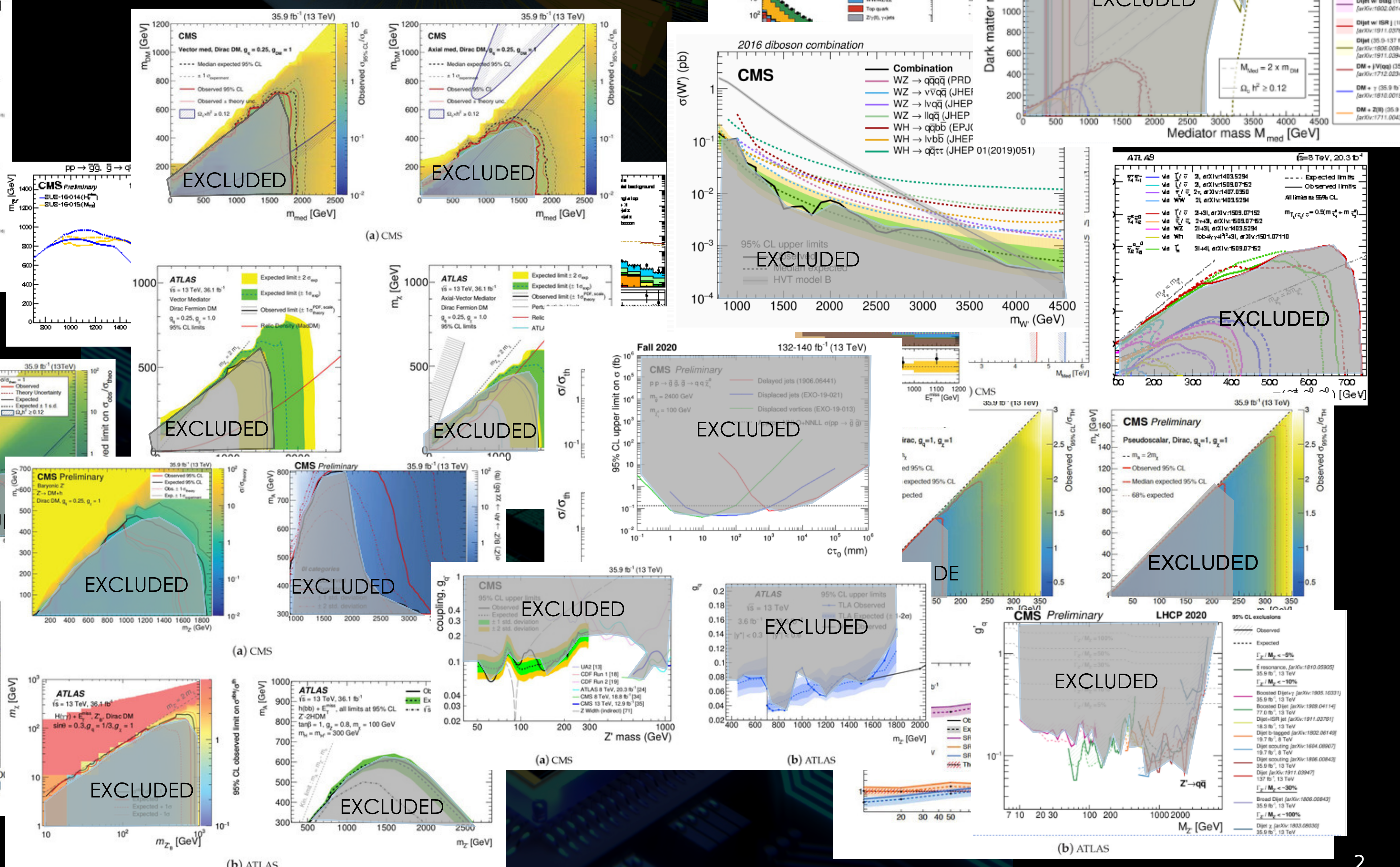
Katya Govorkova

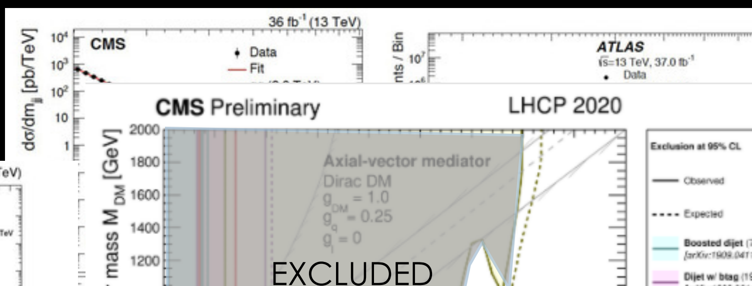
31 January 2023

Status of New Physics searches in HEP

We need **new physics (NP)** to explain many **phenomena**
Dark matter/energy, flavour puzzles, strong CP, hierarchy, baryogengesis. But **dedicated searches** are **not successful** so far

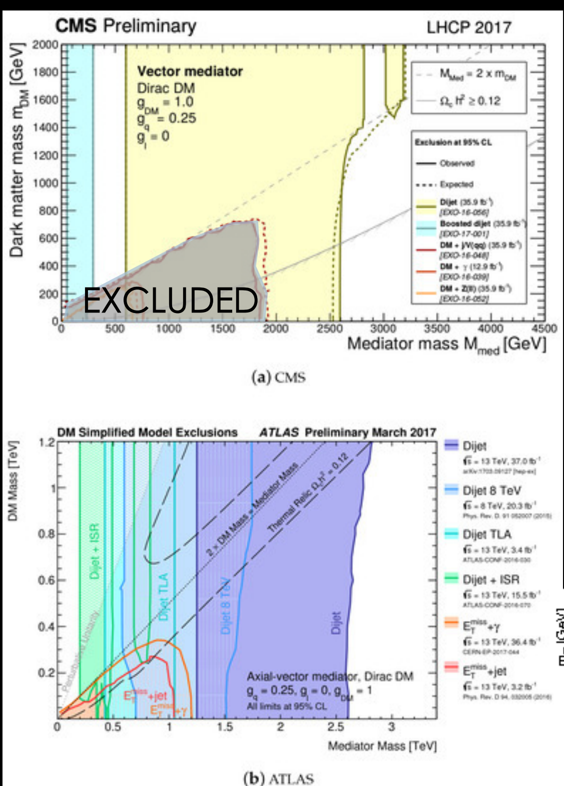


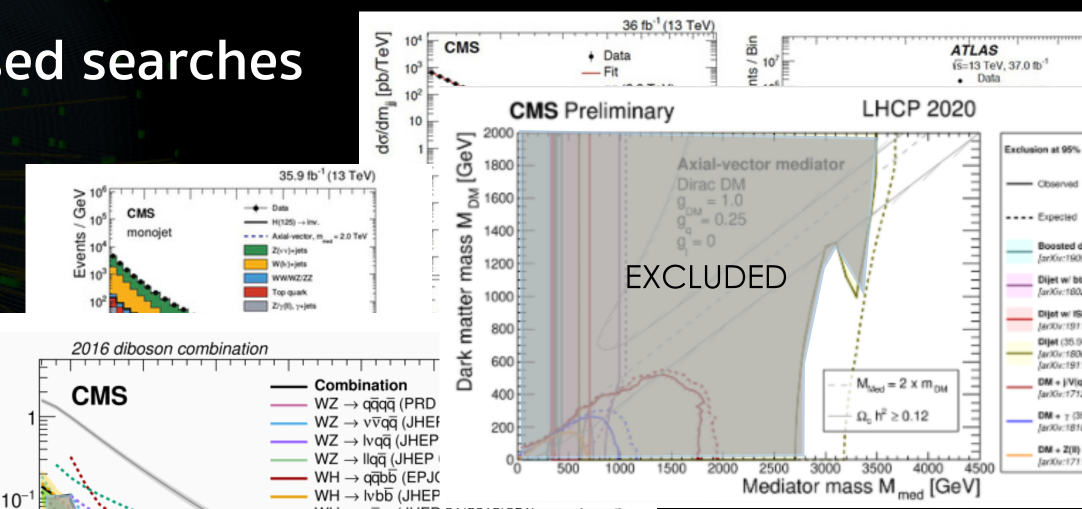


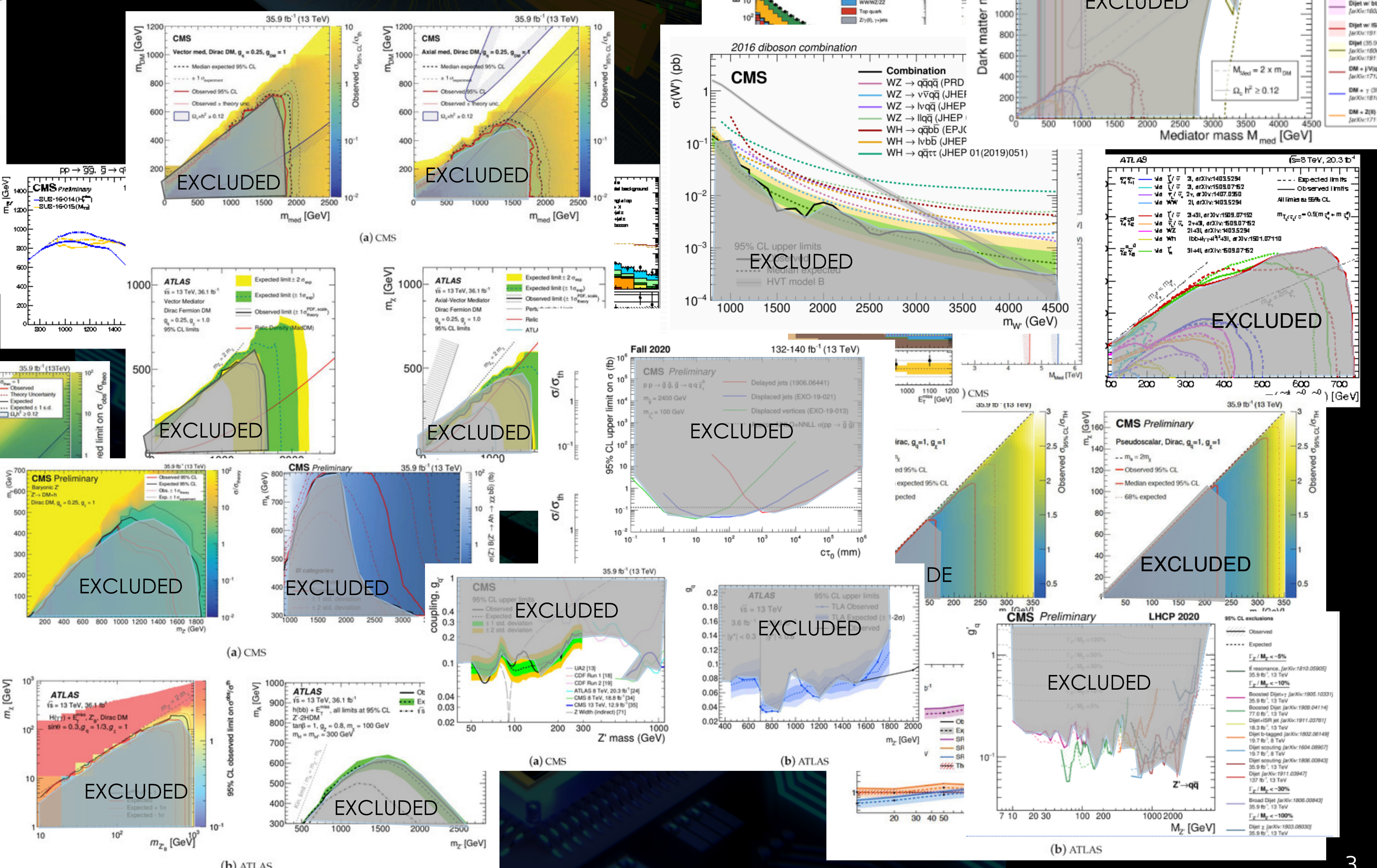


Status of New Physics searches in HEP

We should accept the possibility that **we do not know what to look for** and perform **unsupervised searches** for something that is off, **“anomalous”**



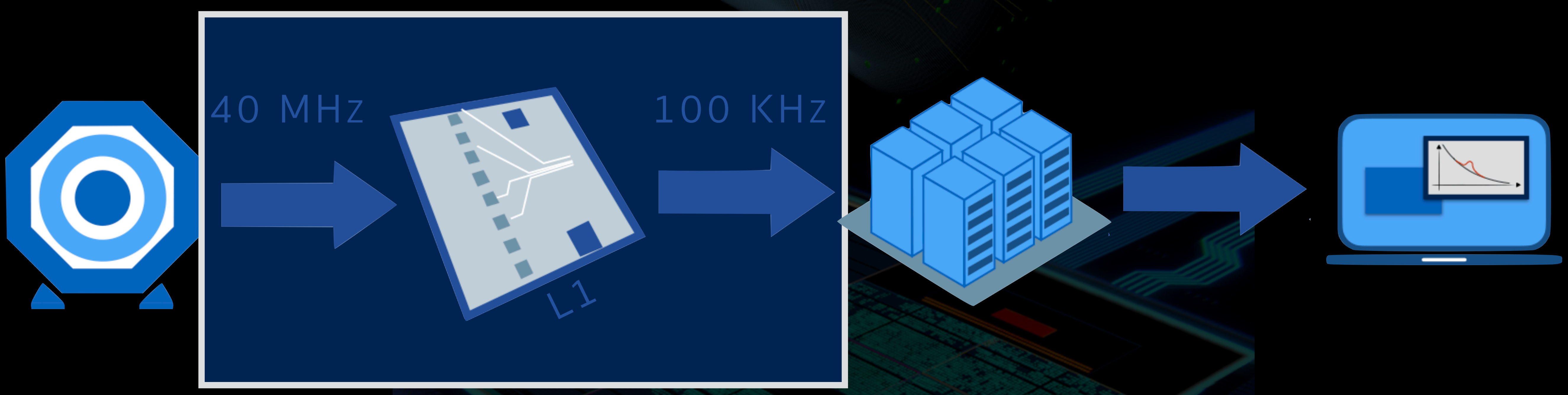




The LHC Big Data Problem

At the **first level (L1) of the trigger system**

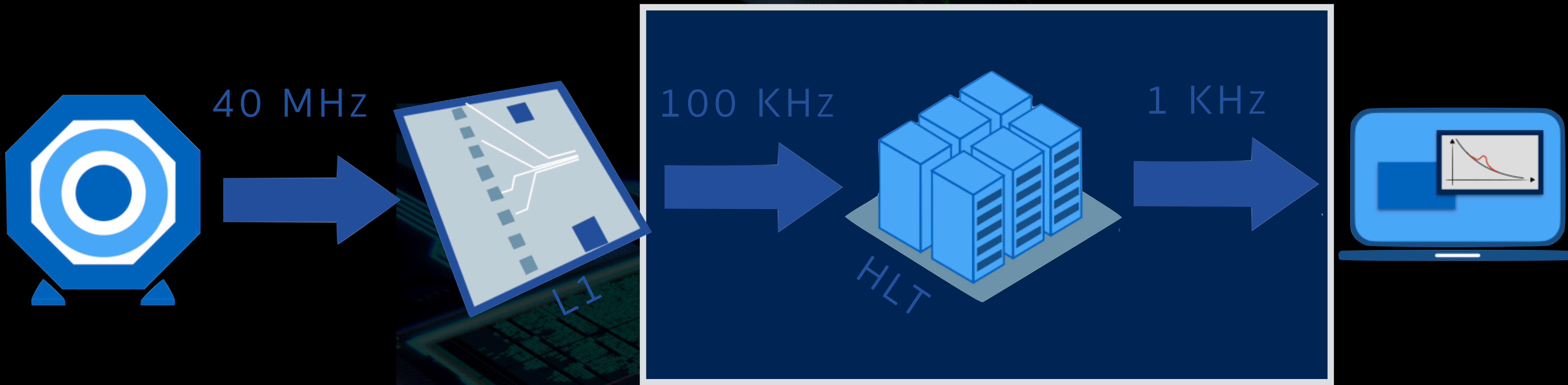
- **FPGAs/hardware** implemented
- **Processing time** 10 μsec
- Based on **coarse local reconstruction**
- Typical **event size** 500 KB/event



THE LHC BIG DATA PROBLEM

At the **High-Level Trigger stage**

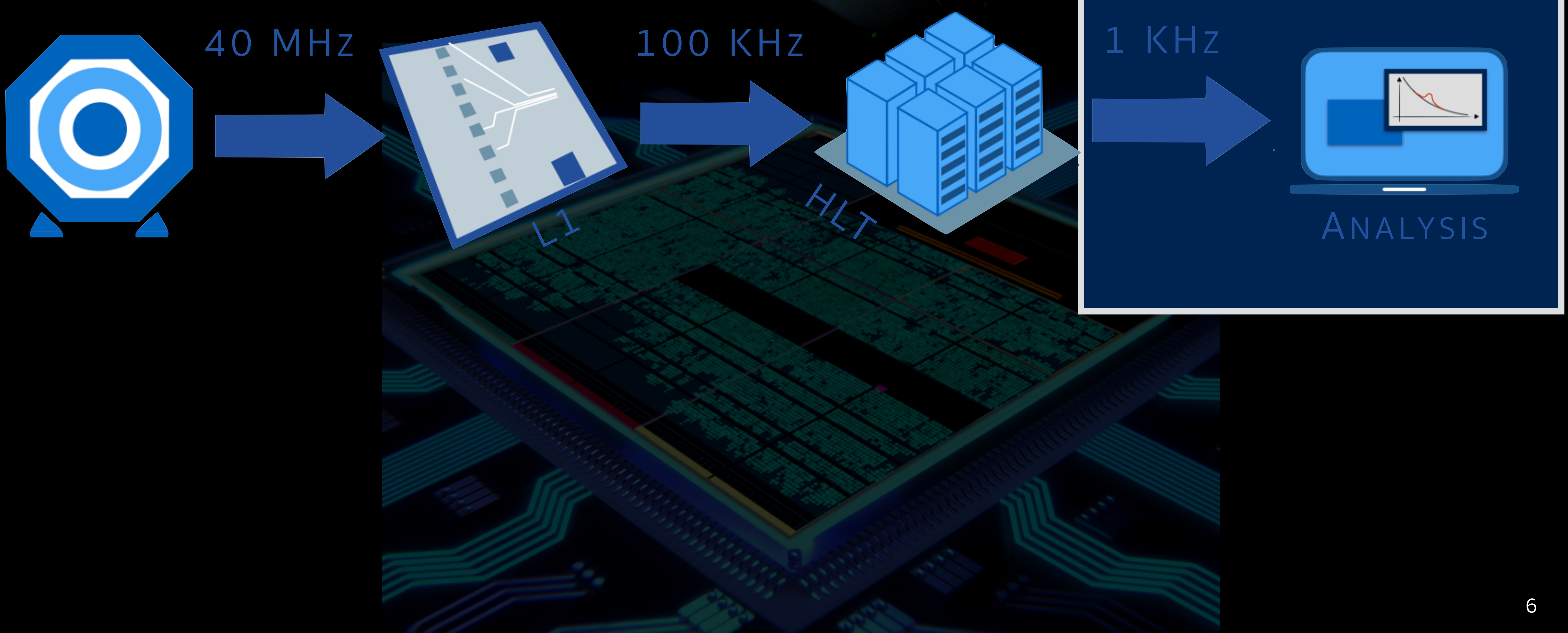
- **Software** implemented on **CPUs**
- **Processing time** 30 milliseconds
- Typical **event size** 500 KB/event



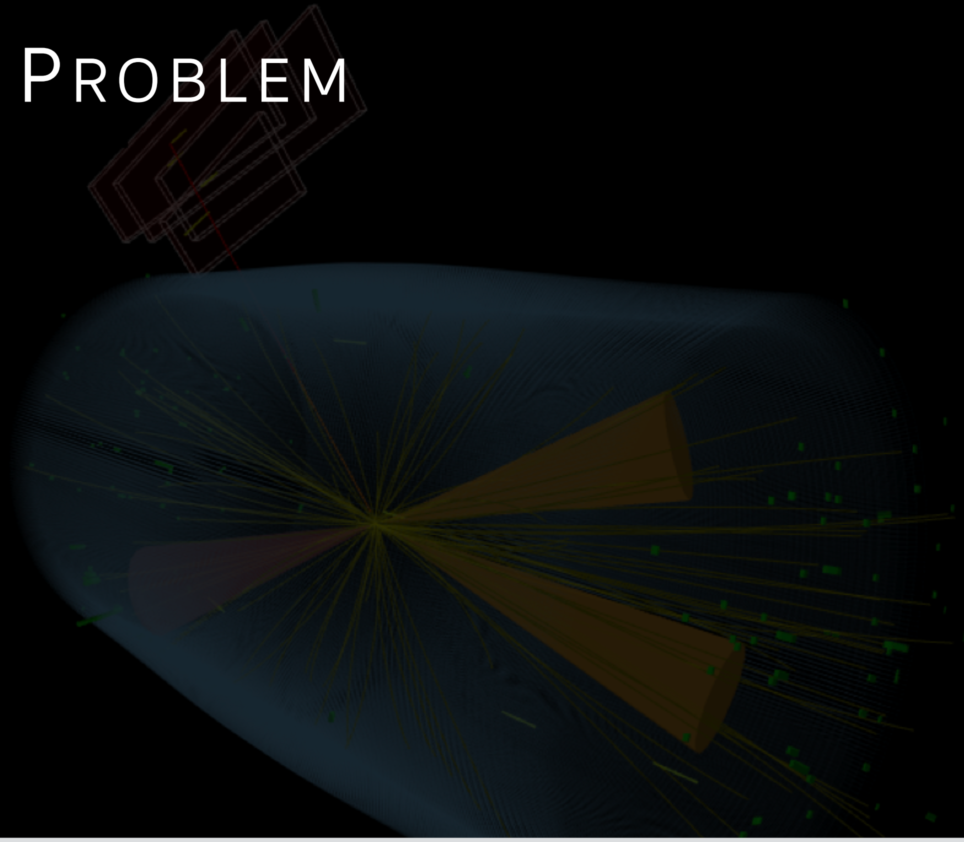
The LHC Big Data Problem

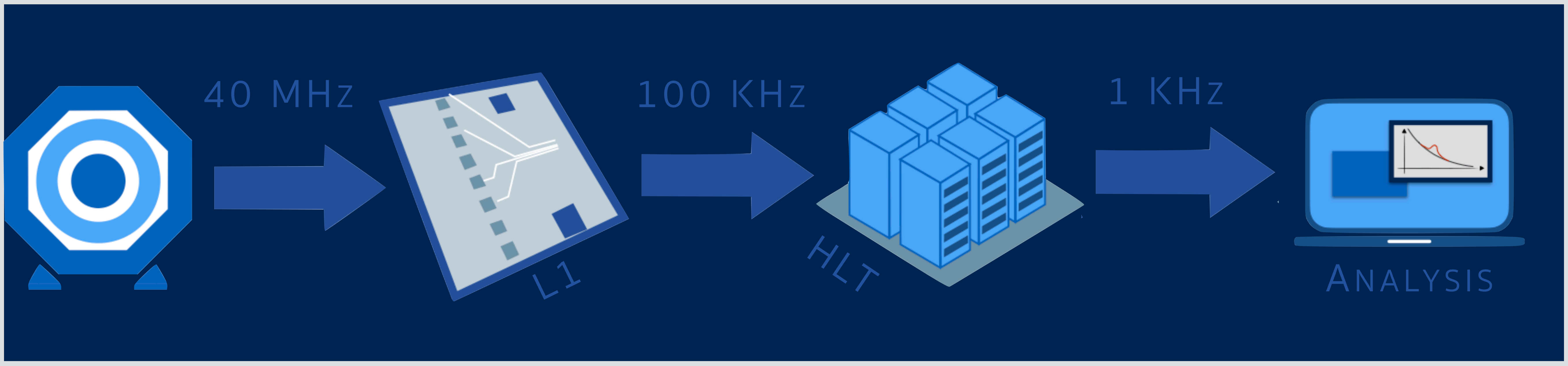
At the **offline analysis** stage

- Typically **100-1000 events out**
- Processing **time is irrelevant**
- **User-written code** and centrally produced selection algorithms
- **< 30 KB per event**



THE LHC BIG DATA PROBLEM





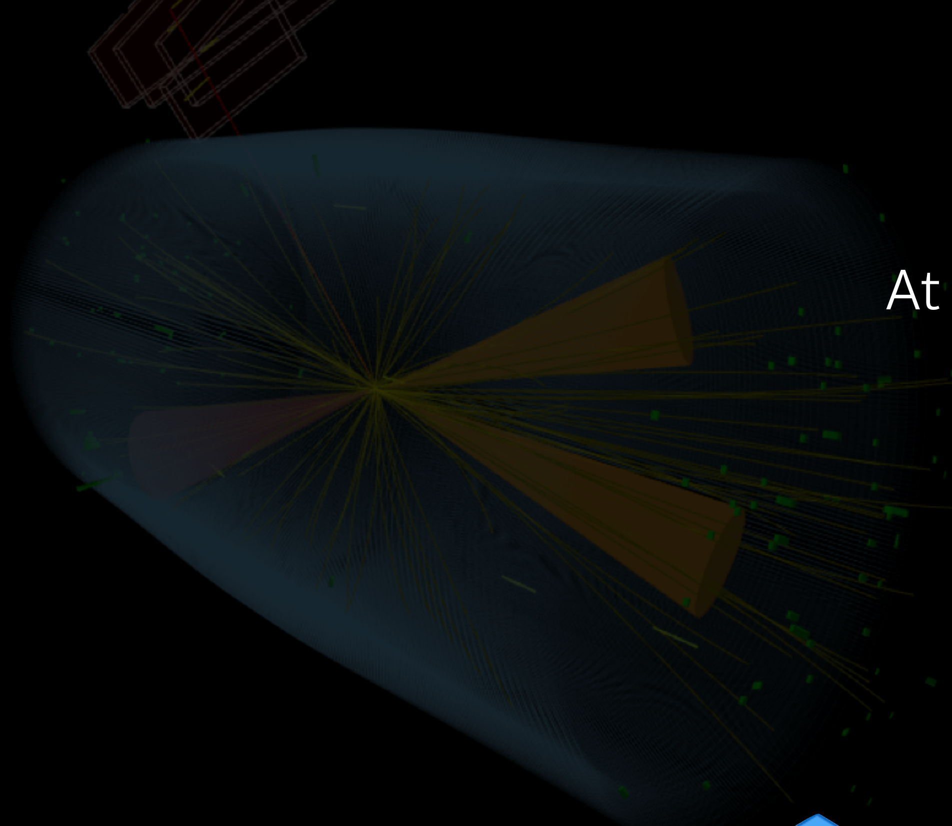
With such **a tight selection** to be made, **the risk of discarding events** is not negligible

The problem starts with **the need to assume a specific physics model**, to then make sure that we trigger on it

What if we never consider the right model?

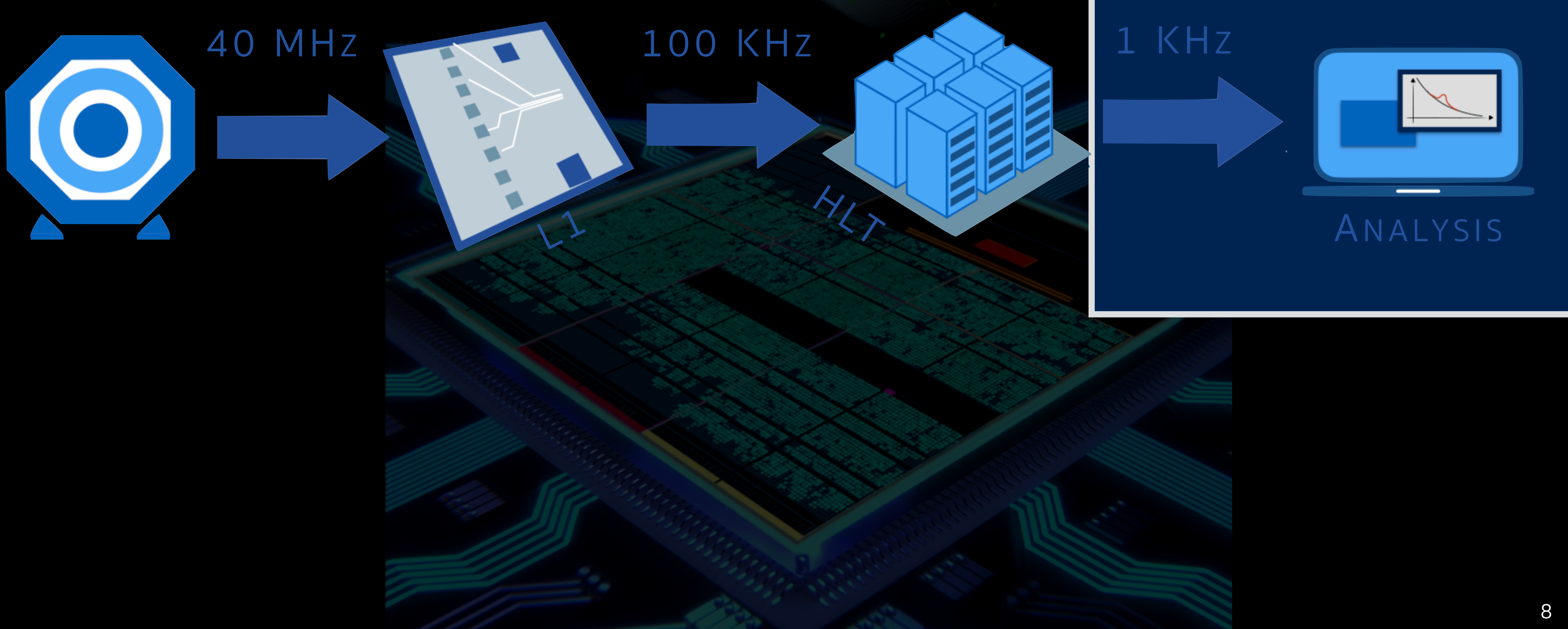
We will deploy in the trigger system **an algorithm** that selects **anomalous events with deep learning**

When to look for Anomalies?



At the **offline analysis** stage

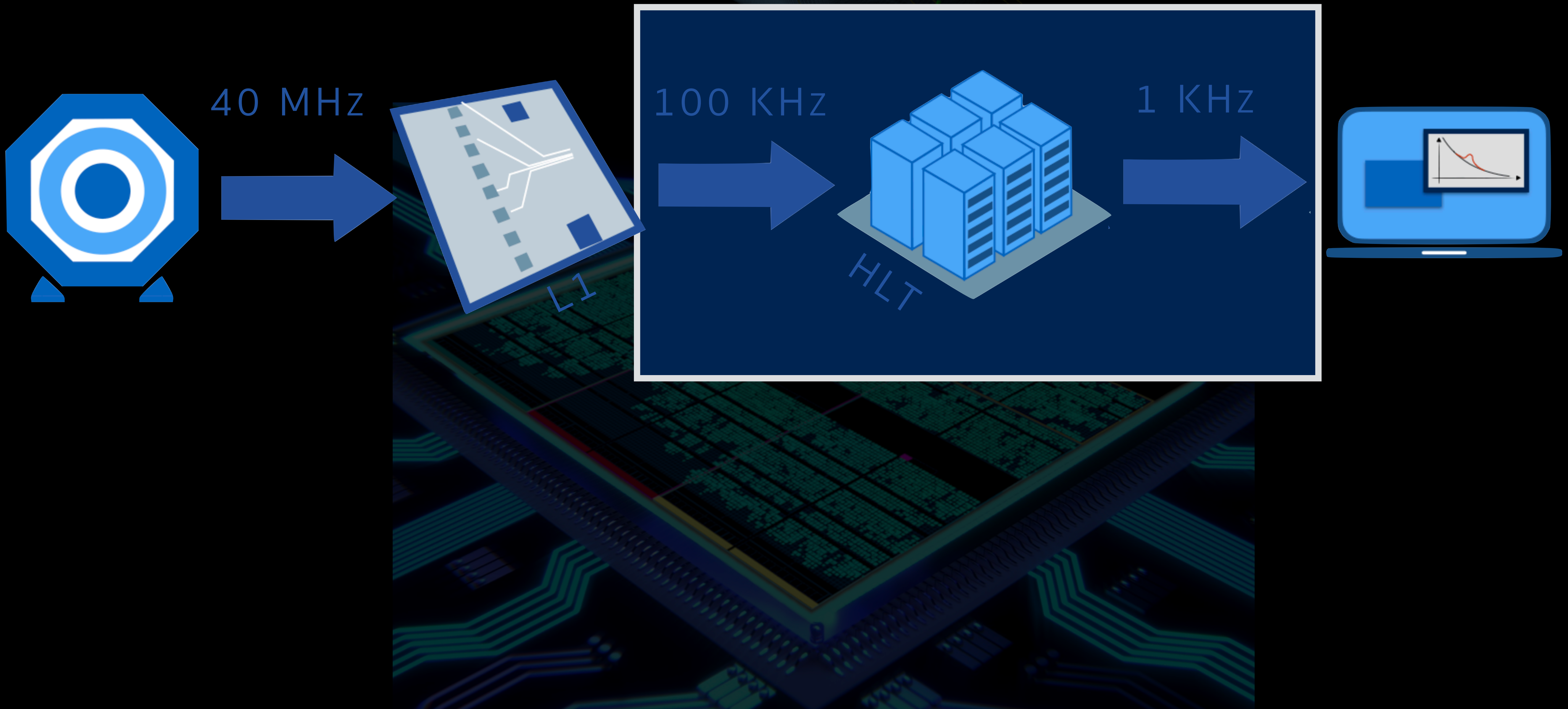
- Can use **very large ML models**
- But **NP can already be discarded**



When to look for Anomalies?

At the **High-Level Trigger stage**

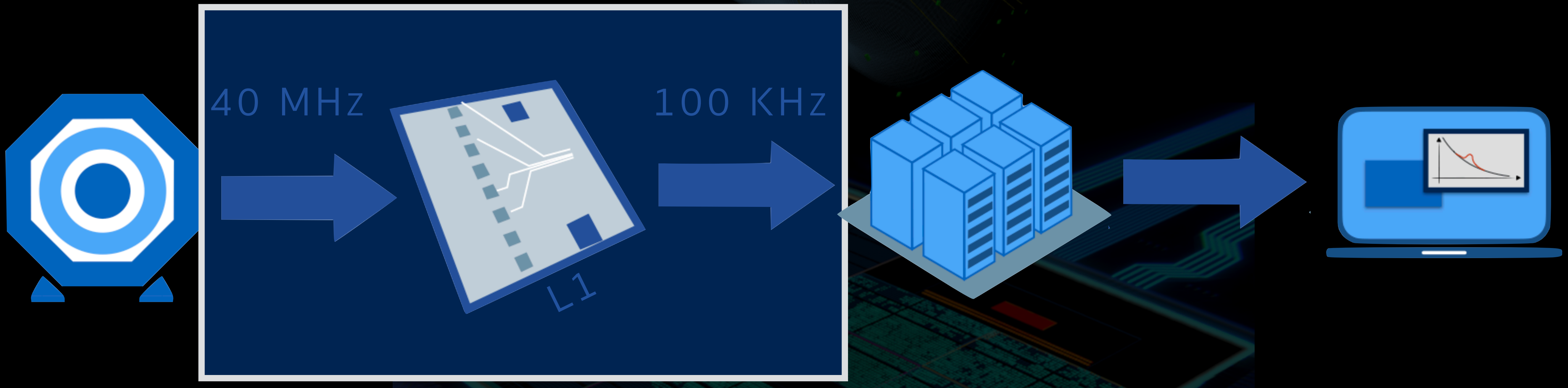
- 100 times more data so **more chance to find NP**
- However **stickier requirements on latency**



When to look for Anomalies?

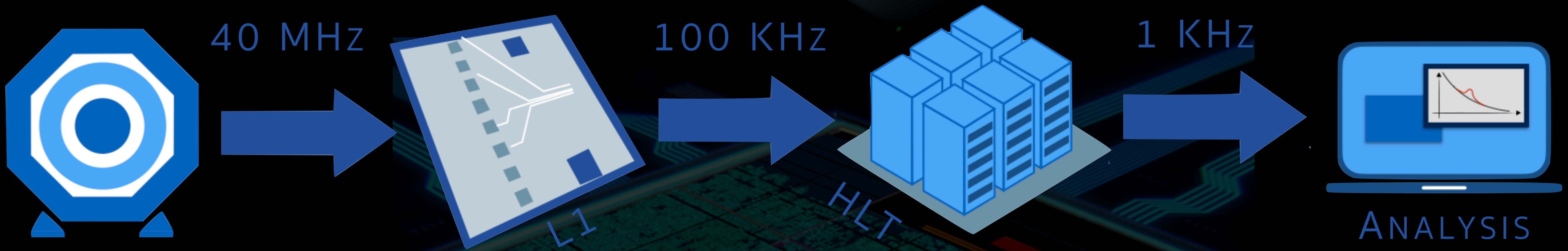
At **the first level of the trigger system**

- **Extreme latency (50ns)** and **resource usage requirements**
- Only **partial event information is available**
- But the algorithm will see **ALL the data** produced at the LHC



Public challenges

- LHC Olympics 2020 2101.08320
- The Dark Machines Anomaly Score Challenge 2105.14027



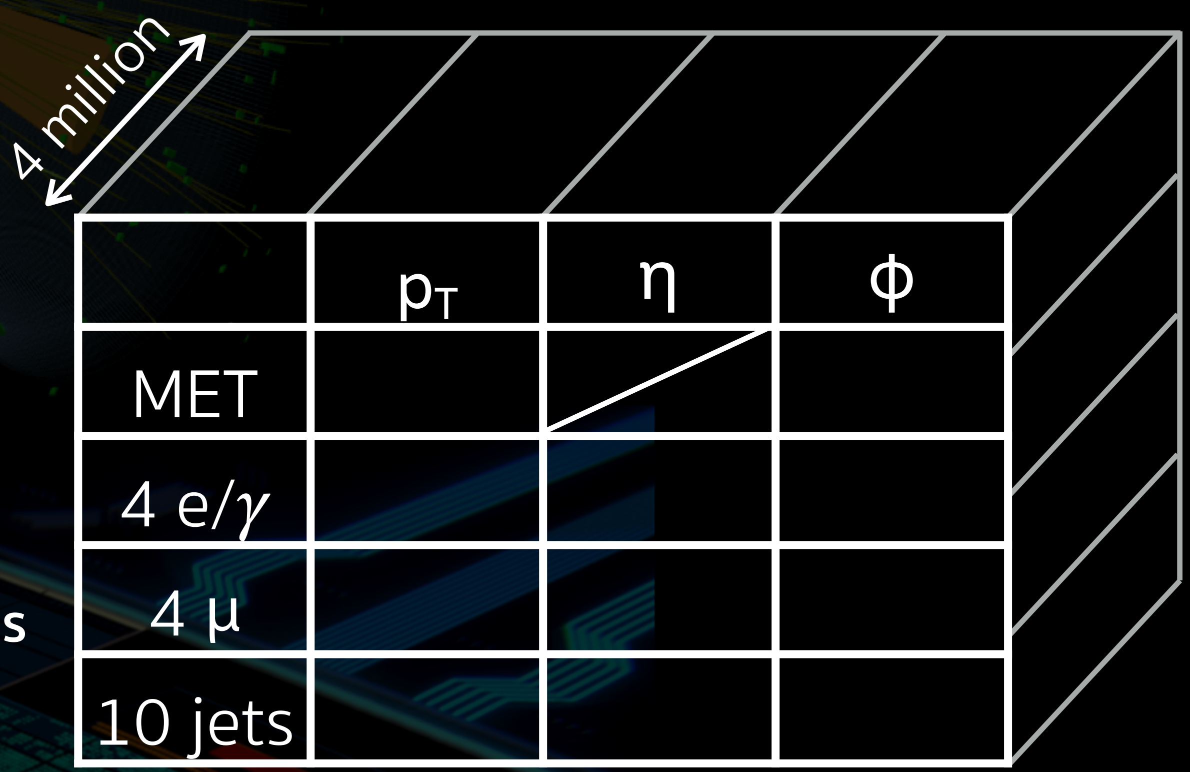
- Unsupervised New Physics detection at 40 MHz

Training and Datasets

The data mimic L1 trigger data format

Train using **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%)

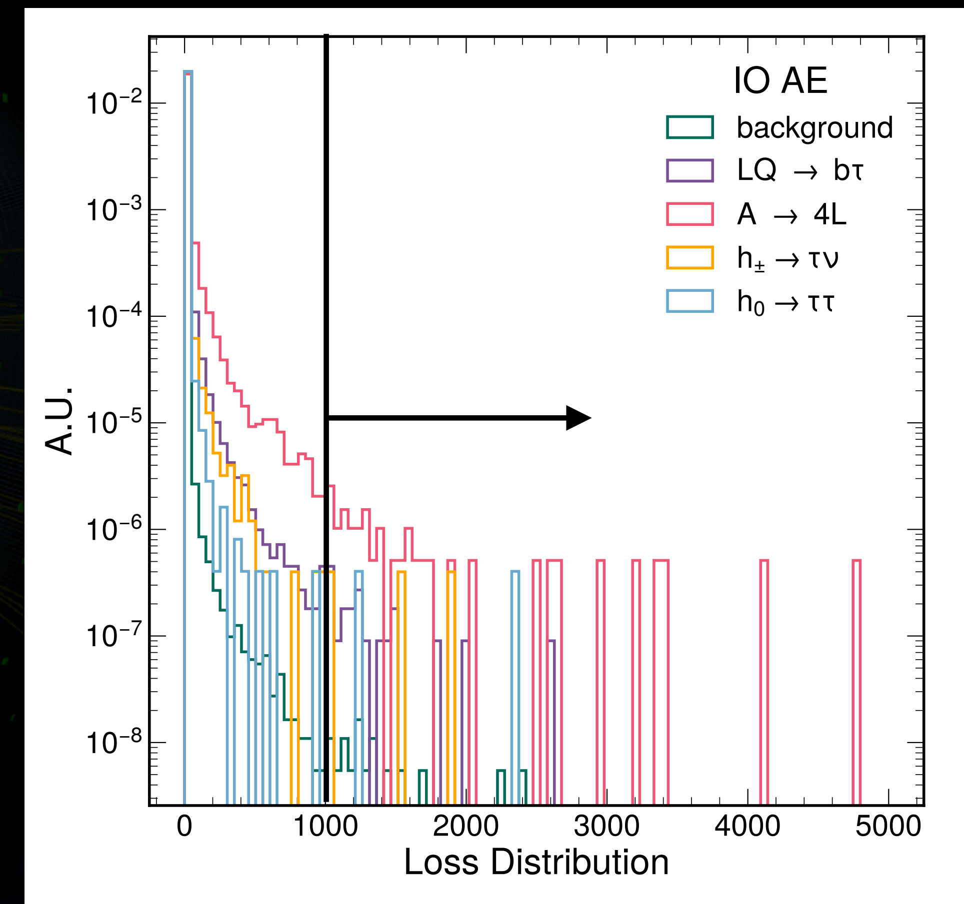


Evaluate performance on several different **New Physics simulated samples**

- 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 🎁

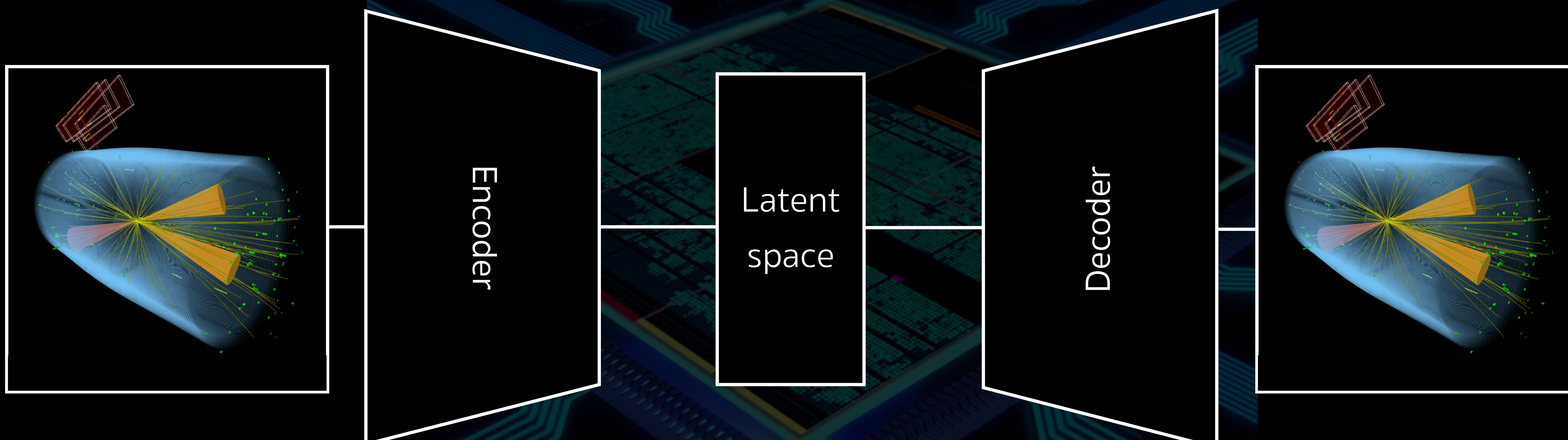
Autoencoders for Anomaly Detection

- **Autoencoders** are **compression-decompression algorithms** that learn to describe a given dataset in terms of **points in a lower-dimension latent space**



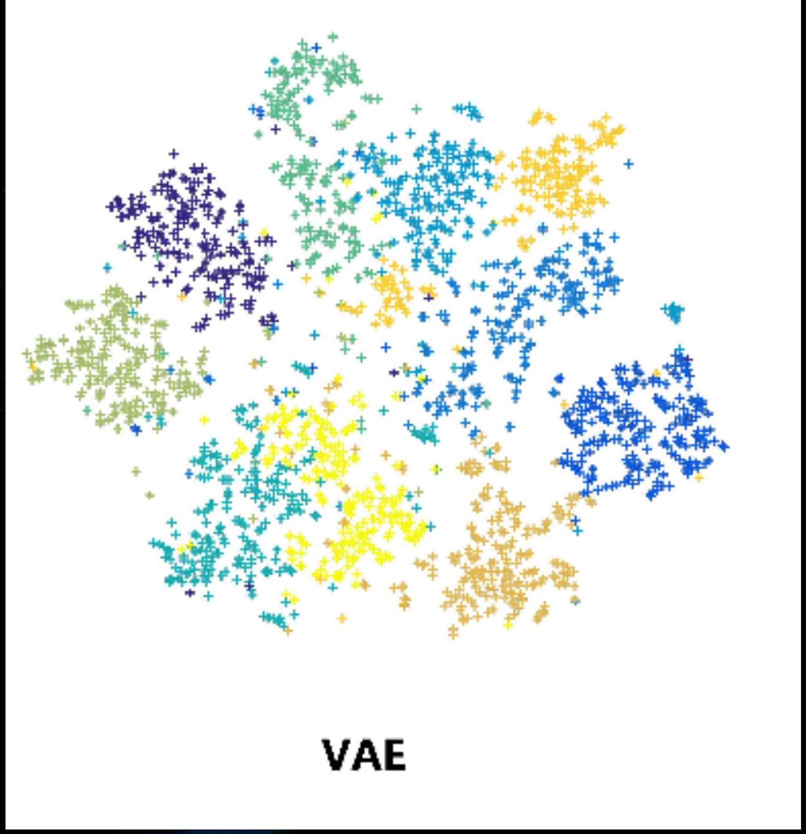
Input-output anomaly detection with loss

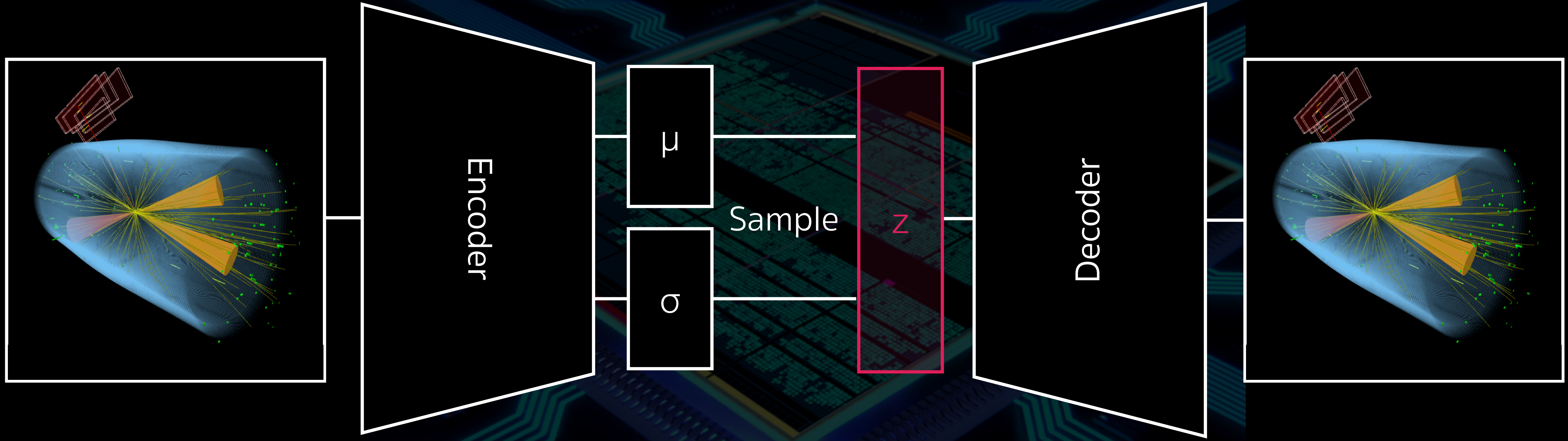
$$\text{MSE(input, output)} = \frac{1}{N}\sum_{i=0}^{N}(input - output)^2$$



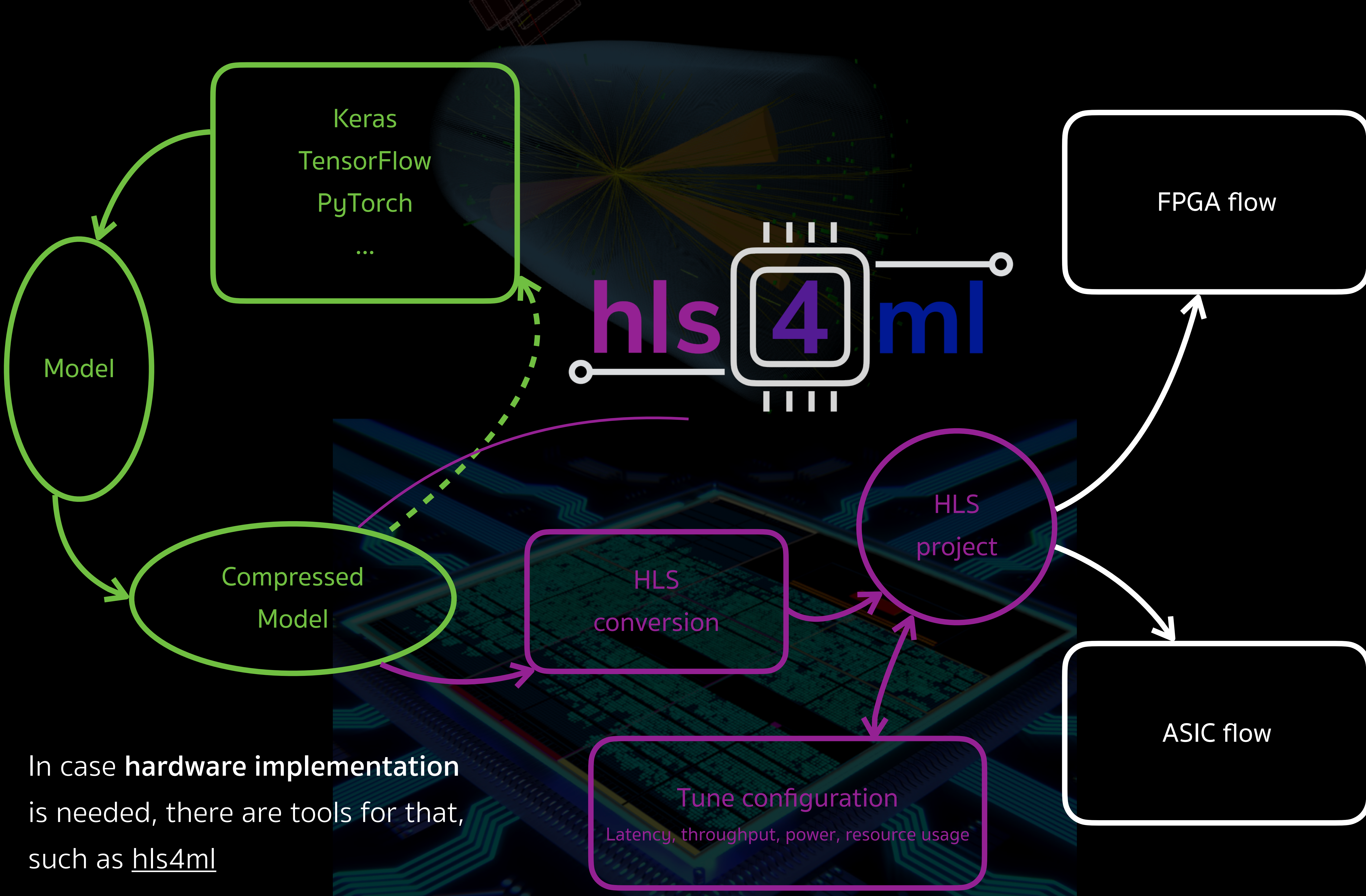
Enhancing Vanilla Autoencoders

- **Variational** autoencoders — deep generative models that learn a compact representation of the data 2108.03986
- **Normalising** flows — transform simple distributions into complex ones by applying a series of invertible transformations 2110.08508
- **Contrastive** autoencoders — learn compact representations by contrasting positive pairs of samples with negative pairs 2301.04660
- **Combining several AEs** — ABCD method, and also see the talk from Ryan



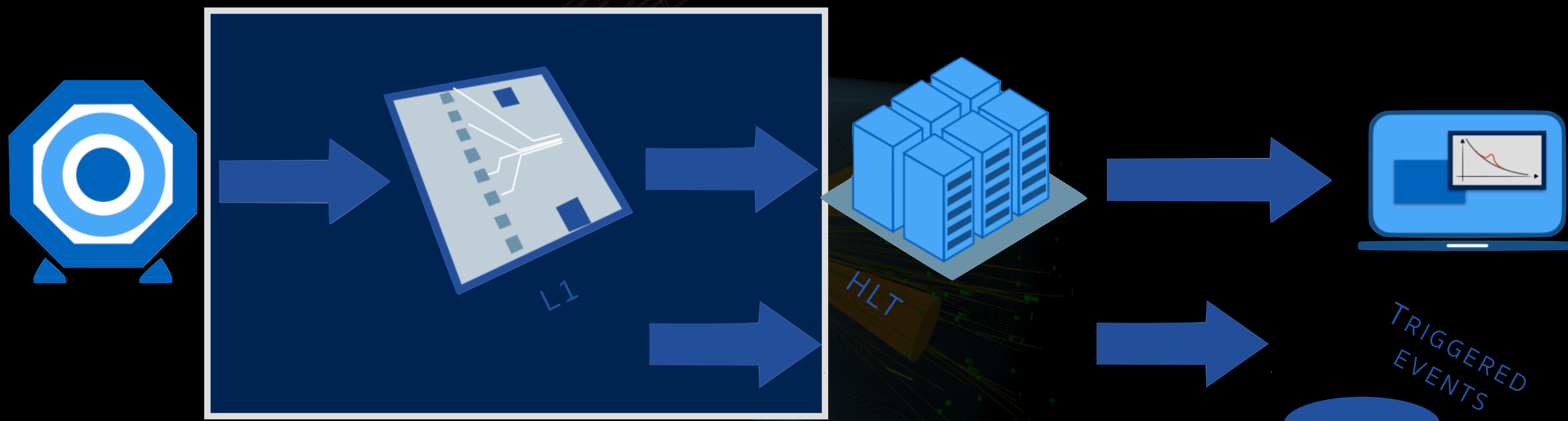


Hardware Implementation Workflow



In case **hardware implementation** is needed, there are tools for that, such as hls4ml

Anomaly Detection pipeline



- Once the event is passed L1 anomalous trigger by having a loss above the threshold, it is saved
- Offline we can cluster all the anomalous events and compare to expected clusters with simulation
- Once we see unexpected clusters in the data, we can have a closer look and potentially develop dedicated triggers for those phase space areas
- This can also be used to monitor the detector quality

SUMMARY

Anomaly detection for **new physics** searches is now **considered** at every stage of **data processing at the LHC**

With the constant development of **ML tools**, there are more **sophisticated unsupervised algorithms** that can help us finally shed a light on **new physics** (or on detector malfunctions)

There is still a lot of work to be done to create a **user-friendly established pipeline for AD algorithms in the trigger**, and **this work is being done now**