Intro to Anomaly Detection in Particle Physics

Oz Amram
Aug 15th, 2024
ML4FP School

Fermilab

Overview

- What is anomaly detection?
- Method 1 : Outlier Detection
- Method 2 : Overdensity methods
- Hands on tutorial

Warning : Decent amount of personal / LHC bias!

Intro

- What is anomaly detection?
  - “Finding something interesting without specifying exactly what you are looking for”
  - Classification without specifying your signal class
- Why would you want to do it?
  - Many possible signals in your data (or failure modes of your detector) → cannot search for them all one by one
  - Don’t want to miss a discovery because we didn’t think to look for it!
  - Science is full of many unexpected discoveries! Non-trivial to make this possible for modern complex data analysis

HEP Data Analysis

Basic Selection → **Signal Enhancing Selection** → **Statistical Analysis**

Basic Selection

- Good quality data
- Contains objects of interest
- Passes trigger*
- ...

Signal Enhancing Selection

- Simple cuts ↔ NN's
- Optimized with a chosen signal model
- Don't spoil background prediction

Statistical Analysis

- Predict remaining bkg
- Test B vs S+B model → significances, limits …

HEP Data Analysis

Basic Selection → **Signal Enhancing Selection** → **Statistical Analysis**

Basic Selection
- Good quality data
- Contains objects of interest
- Passes trigger*
- ...

Signal Enhancing Selection
- Simple cuts ↔ NN's
- Optimized with a chosen signal model
- Don't spoil background prediction

Statistical Analysis
- Predict remaining bkg
- Test B vs S+B model → significances, limits …

Can we do this part without specifying a signal model?

HEP Data Analysis

Basic Selection → **Signal Enhancing Selection** → **Statistical Analysis**

Basic Selection:
- Good quality data
- Contains objects of interest
- ...

Signal Enhancing Selection:
- Simple cuts ↔ NN's
- Optimized with a chosen signal model
- Don't spoil background prediction

Statistical Analysis:
- Predict remaining bkg
- Test B vs S+B model → significances, limits …

Can we do this part without specifying a signal model?

NB : There are methods which combine the statistical analysis w/ the classification part (eg 'New Physics Learning Machine')

Classification

The optimal classifier is the **Likelihood Ratio**

Read about the
Neyman-Pearson lemma
if you are unfamiliar

$$\boxed{L_{S/B}(X) = \frac{P_s(X)}{P_b(X)}}$$

Prob. distribution of
signal

Prob. distribution of
background

Classification

The optimal classifier is the **Likelihood Ratio**

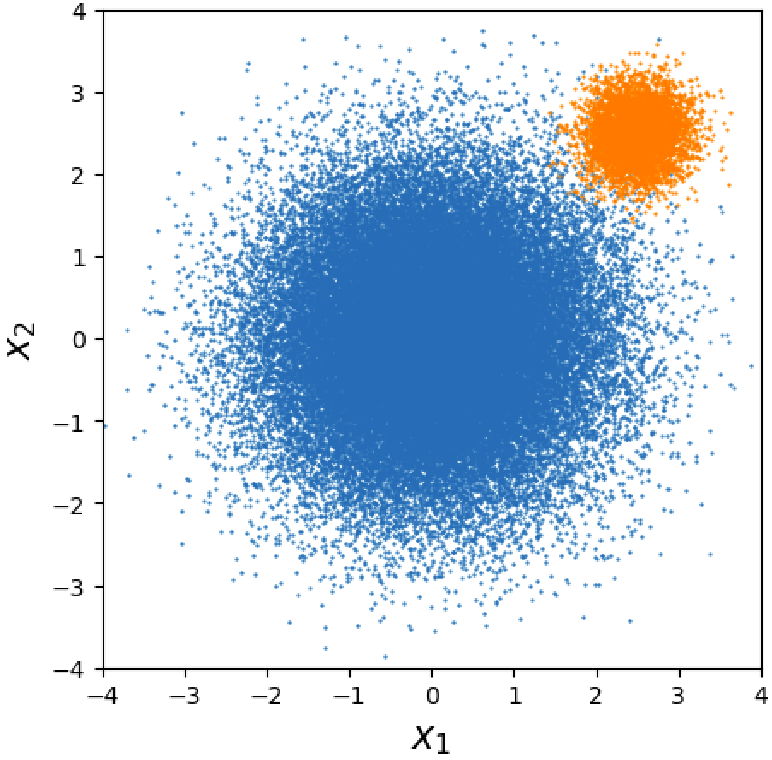
Prob. distribution of **signal**

$$L_{S/B}(X) = \boxed{\frac{P_s(X)}{P_b(X)}}$$

Prob. distribution of **background**

- In anomaly detection we do not know P_s
- How can we approximate the likelihood ratio then?
- Outlier Detection : **Learn P_b**, take anomaly score as $1/P_b$
- Data-driven likelihood ratio : Leverage localization of signal to $L_{S/B}$ **from data**

Outlier Detection

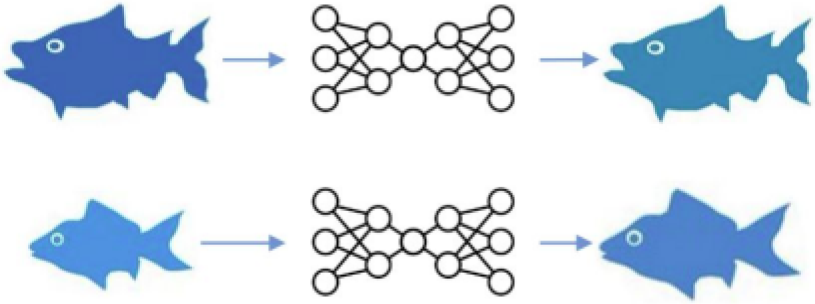


Outlier Detection

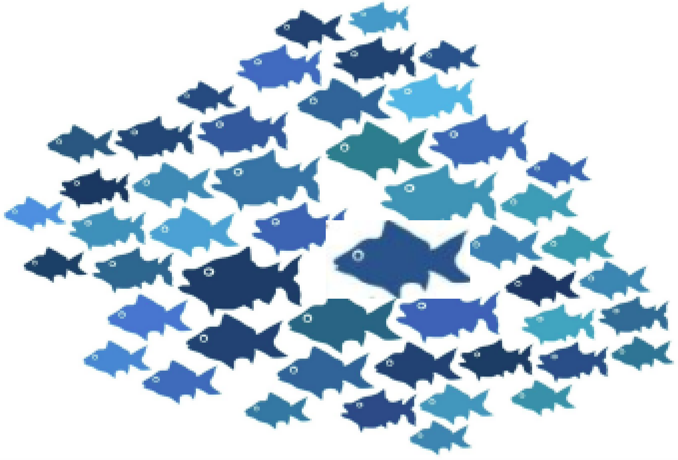
- We don’t know a signal → focus only on bkg (denom. of $L_{S/B}$)
  - Low $P_b(X)$ → anomalous
  - Ie, things that are rare / impossible to be background are anomalous
- Often have many examples of background, but don’t know explicit prob. dist.
  - First thing to try : simple tools to estimate bkg pdf (KDE, GP, …)
- For complex high dim. data can be hard to explicitly model P_b
  - Sometimes sophisticated generative models can be used to learn P_b (normalizing flows, diffusion) → covered already in other tutorial
  - Or train a model on bkg data to learn a proxy for P_b, like an **autoencoder**

Looking for Outliers

Train ‘Autoencoder’



Training Sample



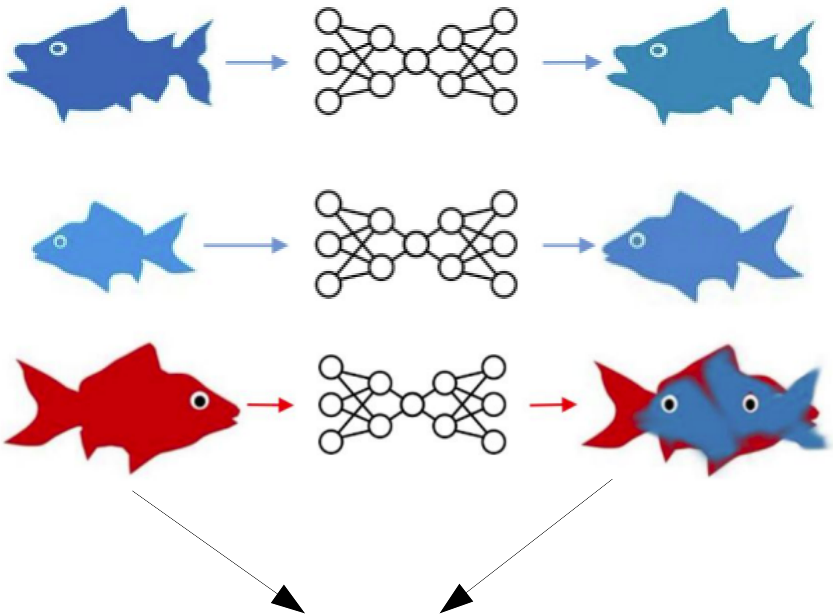
Autoencoder learns to compress data into a smaller representation & then decompress

→ Will learn this well for ‘in distribution’ training set, will do poorly on ‘out of distribution’ (anomalies)

Illustrations: J Gonski, A Kahn

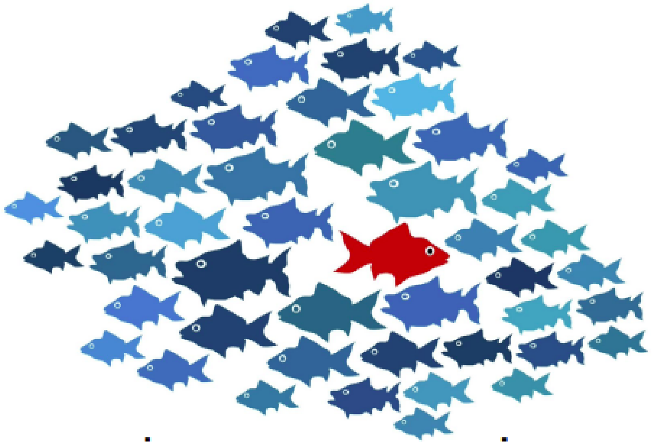
Looking for Outliers

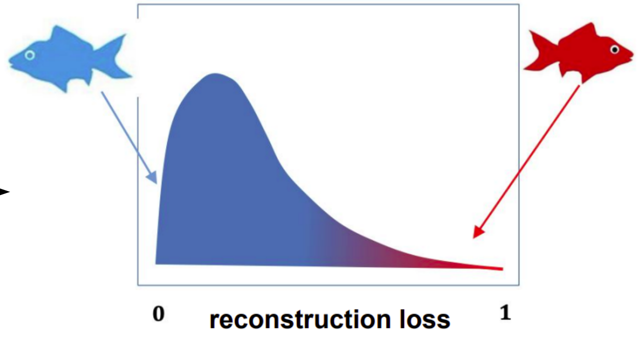
Apply Autoencoder



Take difference

Data from signal region



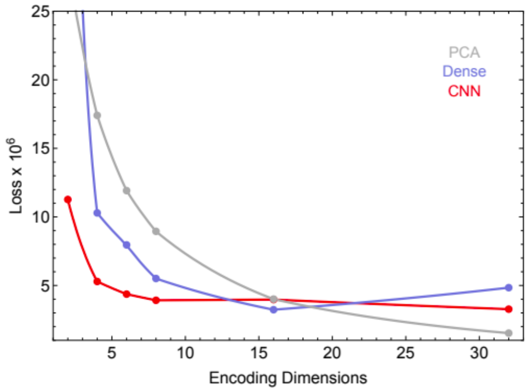


Illustrations: J Gonski, A Kahn

1808.08979
1808.08992

Autoencoder Practicalities

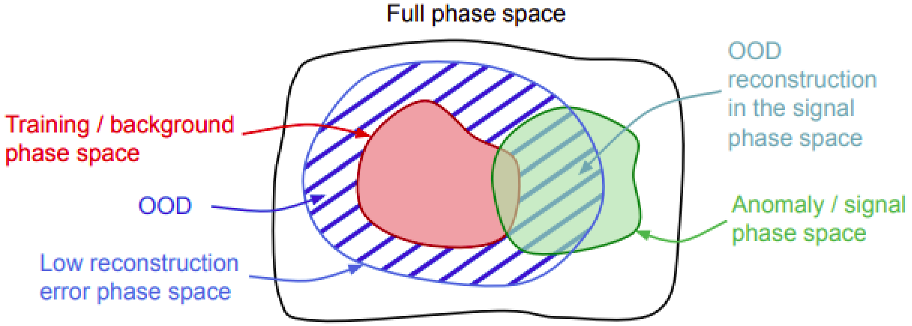
- Training loss is (typically) MSE between input & output
- Size of compressed (latent) dim is an important hyperparameter
  - No exactly method to pick it
  - Often look for ‘elbow’ in loss vs. dim distribution
- Can train directly from data!
  - Performance resilient to small amount of signal presence
- Can use **variational** autoencoder (VAE)
  - Same idea but force latent space to be Gaussian
  - Doesn’t seem to be a huge performance gain



1808.08979
1808.08992

Challenge 1 : Autoencoder Biases

- Autoencoders do not directly model P_b, suffer from biases
  - **Complexity bias** → more 'complex' data (higher intrinsic dim) harder to compress, seen as more anomalous
  - **Over generalization**: AE can reconstruct things well even outside training phase space because no penalty to do this



- **Normalized autoencoders** attempt to solve these issues 2206.14225
- Methods that directly model bkg pdf (NF's, diffusion) don't have these same issues

Challenge 2: Coordinate Invariance

Probability densities (eg $P_b(X)$) not invariant under coordinate transformations

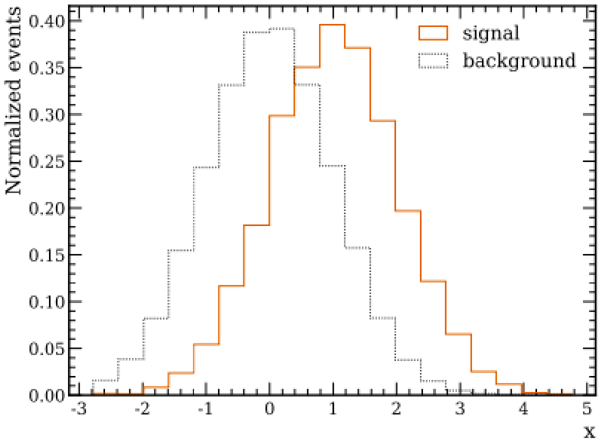
$y = f(x)$ $\longrightarrow$ $p_y(y) = p_x(f^{-1}(y))|\frac{d}{dy}|f^{-1}(y)|$

2209.06225

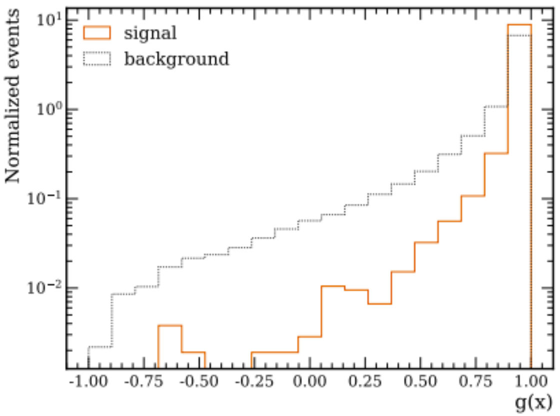
Challenge 2: Coordinate Invariance

Probability densities (eg $P_b(X)$) not invariant under coordinate transformations

$y = f(x)$ $\longrightarrow$ $$p_y(y) = p_x(f^{-1}(y))|\frac{d}{dy}|f^{-1}(y)|$$



y = tanh(x+2)

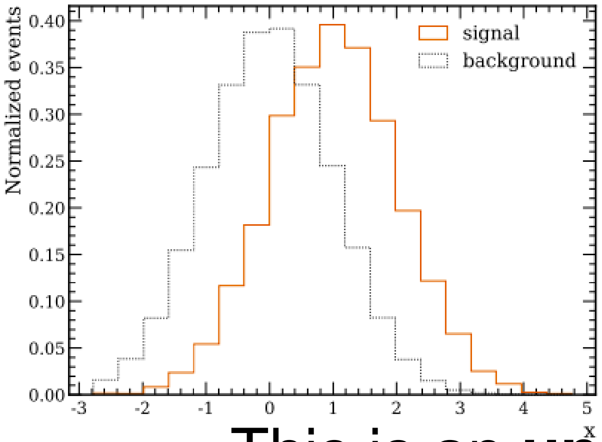


2209.06225

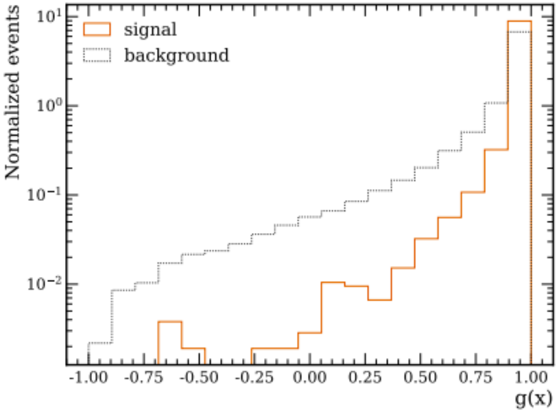
Challenge 2: Coordinate Invariance

Probability densities (eg $P_b(X)$) not invariant under coordinate transformations

y = f(x) → $p_y(y) = p_x(f^{-1}(y))|\frac{d}{dy}|f^{-1}(y)|$



y = tanh(x+2) →

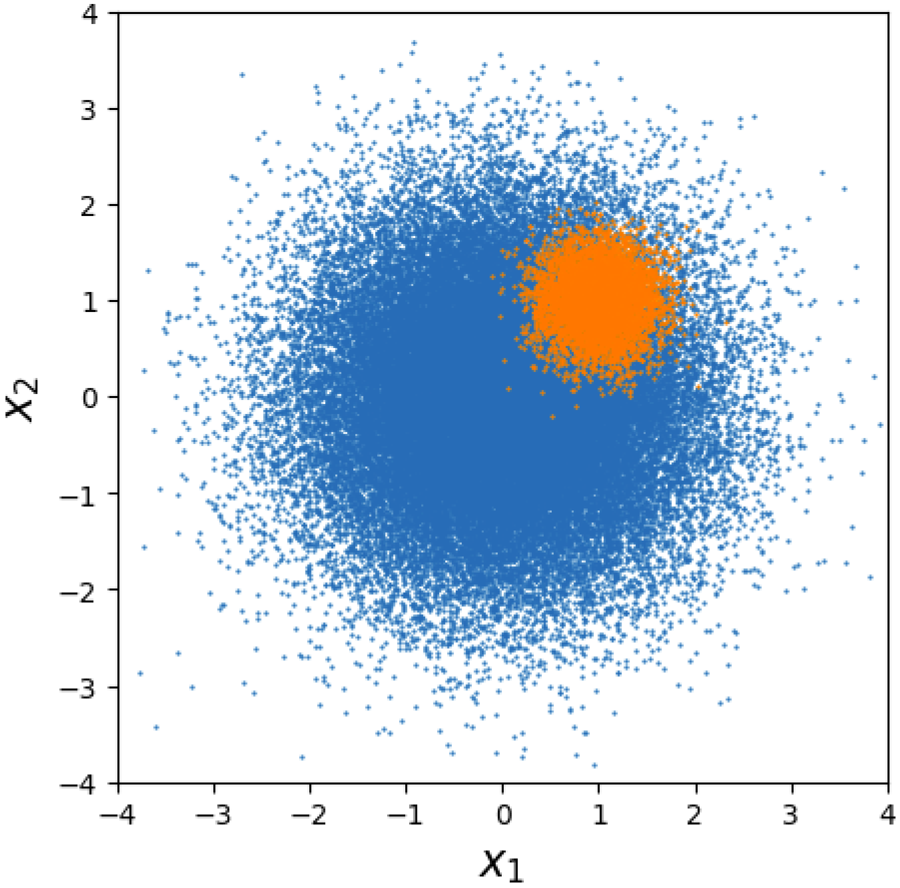


This is an **unavoidable** limitation of using only P_b

→ Data representation is an inductive bias for anomalies!

2209.06225

Data-Driven Likelihood Ratio



Advantages of the likelihood ratio?

- Often in HEP, signals are **within** the bkg distribution rather than full outliers
  - What makes them anomalous is **a cluster of similar events**
  - These **cannot** be found with outlier detection methods
- Likelihood ratio is **coordinate invariant**
- Outlier methods have upper bound on sensitivity because never learn about P_s

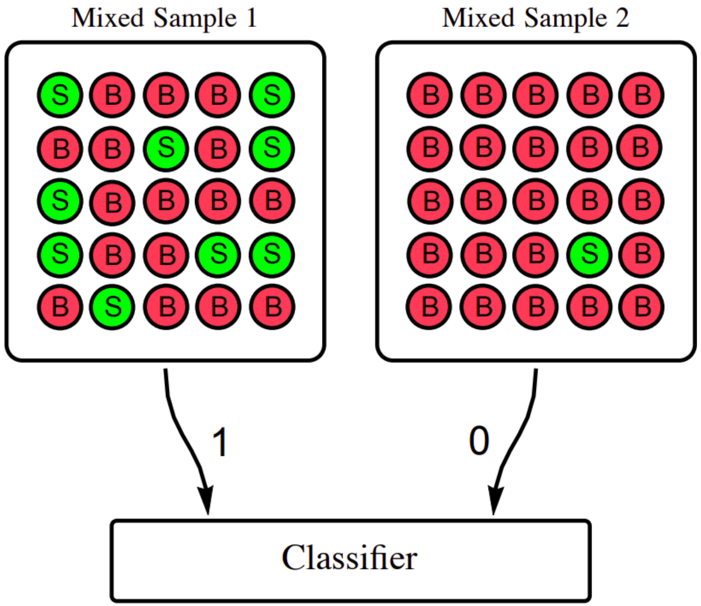
The Challenge

- A fully supervised NN trained with typical binary cross entropy will learn an approximation to the **likelihood ratio***
- But this requires labels for each data event, which we don't have!
- How can learn the likelihood ratio from **unlabeled data**?

* really a monotonic rescaling as the ratio, but this is identical for classification

Learning the Likehood Ratio

- Suppose someone gives you two samples of mixed signal and bkg
- Assuming the bkg in the two samples has the same underlying distribution
- The optimal classifier for distinguishing these mixed samples is also $L_{s/b}$!
  - Ie training a classifier with these mixed samples will mimic a supervised classifier!



1708.02949

Short Proof

$$L_{S/B}(X) = \frac{P_s(X)}{P_b(X)}$$

Two mixed samples (M_1, M_2) with signal fractions (f_1, f_2)

$$L_{M1/M2}(X) = \frac{P_{M1}(X)}{P_{M2}(X)} = \frac{f_1 P_s(X) + (1 - f_1) P_b(X)}{f_2 P_s(X) + (1 - f_2) P_b(X)}$$

Short Proof

$$L_{S/B}(X) = \frac{P_s(X)}{P_b(X)}$$

Two mixed samples (M_1, M_2) with signal fractions (f_1, f_2)

$$L_{M1/M2}(X) = \frac{P_{M1}(X)}{P_{M2}(X)} = \frac{f_1 P_s(X) + (1-f_1)P_b(X)}{f_2 P_s(X) + (1-f_2)P_b(X)} = \frac{f_1 L_{S/B} + (1-f_1)}{f_2 L_{S/B} + (1-f_2)}$$

Monotonically related to $L_{S/B}$

Short Proof

$$L_{S/B}(X) = \frac{P_s(X)}{P_b(X)}$$

Two mixed samples (M_1, M_2) with signal fractions (f_1, f_2)

$$L_{M1/M2}(X) = \frac{P_{M1}(X)}{P_{M2}(X)} = \frac{f_1 P_s(X) + (1 - f_1) P_b(X)}{f_2 P_s(X) + (1 - f_2) P_b(X)} = \frac{f_1 L_{S/B} + (1 - f_1)}{f_2 L_{S/B} + (1 - f_2)}$$

If $f_2 \to 0$ (ie one sample is 'background pure') then simplifies

$$L_{M1/M2}(X) = \frac{f_1 P_s(X) + (1 - f_1) P_b(X)}{P_b(X)} = f_1 + (1 - f_1) L_{S/B}$$

Weak Supervision

- This method of training between mixed samples is called **weak supervision** (or Classification Without Labels, CWoLa)

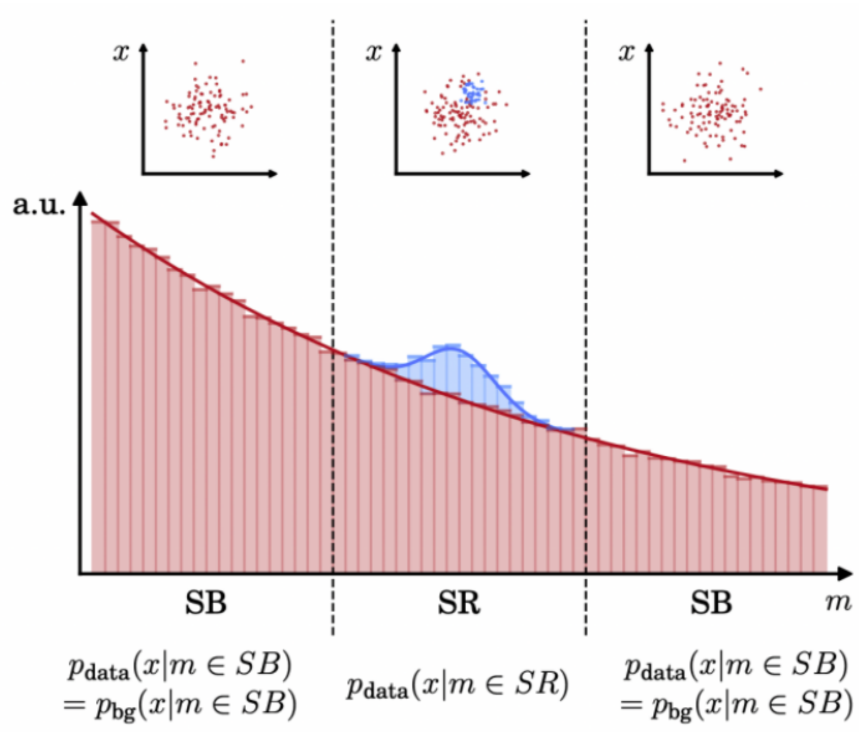
- In practice, convergence to full supervision depends
  - On how large the signal fraction is
  - On **how many** training samples you have
  - On how '**distinctive**' the signal is compared to the background
- Good performance can be achieved with realistic ~1% signal fractions!

1708.02949

Mixed Samples

- Where do I get these mixed samples from?
- This is where your physics knowledge comes in!
- Typically have a signal region where your signal might live
  - Can you find an orthogonal sample of very similar background events?
- Any difference between background events in signal region vs. background sample will be picked up by your classifier!

Weak Supervision + Bump hunt



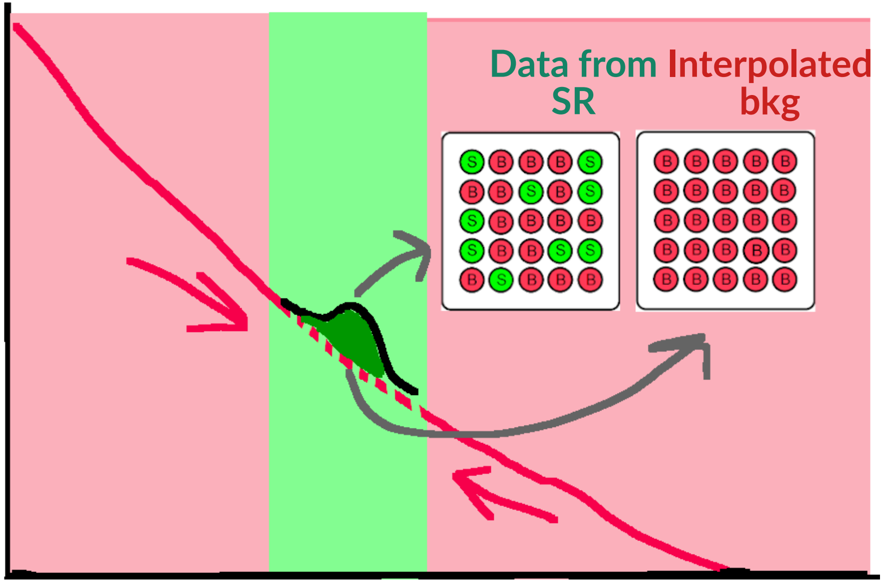
"CWoLa Hunting"

1902.02634

- Assume signal is a **narrow** resonance
  - → Will live in a localized region of mass
  - Sidebands will have very similar bkgs but minimal signal
- **Guess** a mass window where it lives
  - Train signal window vs. narrow sidebands using weak supervision
- **Repeat procedure**, scanning over different mass windows
- Need to be careful about correlations with Mjj

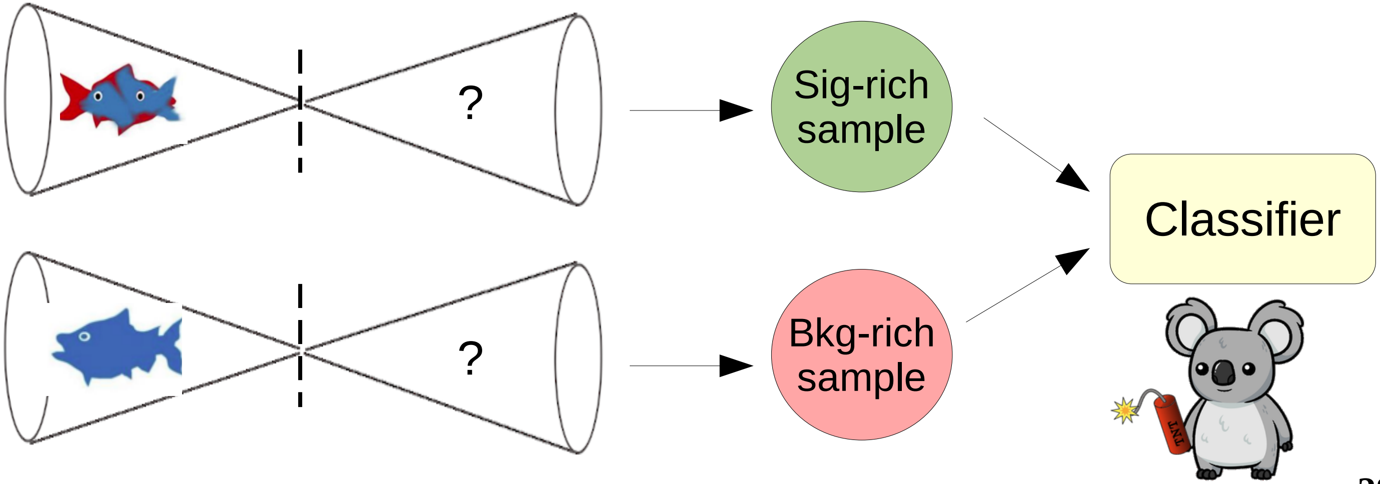
CATHODE
Interpolates bkg events into SR using **generative model**
Use gen. model. To construct bkg sample

2109.00546



Other variants with different interpolation methods (~similar performance)
CURTAINS, SALAD, FETA, ...

Tag N' Train
purifies samples by first tagging with AE



[**OA** & Suarez 2002.12376]

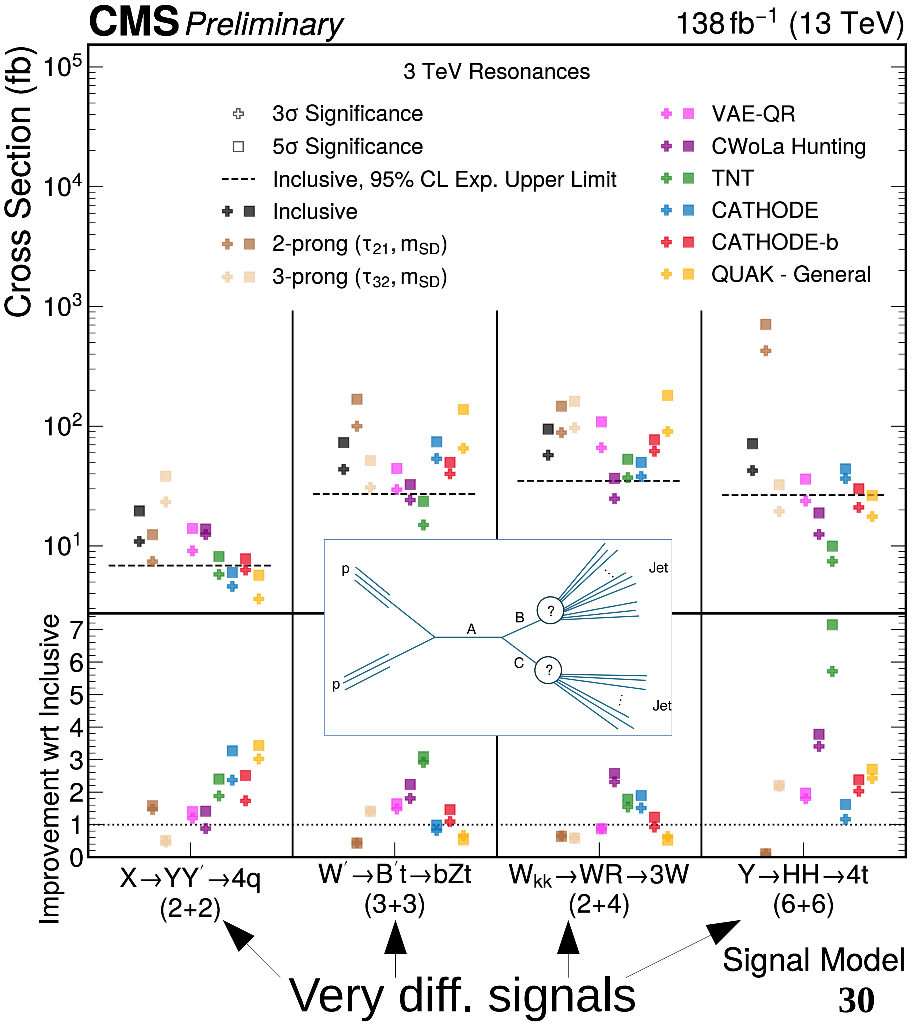
Challenges for Weak Supervision

- Weak supervision training is **noisy**
  - At low signal fraction, works better with high level features → less model independent
  - Ensembles of BDT's seem better than NN's!
- Not easy to create mixed samples
  - Biases in background samples will destroy method
  - **How can we apply this beyond bump hunts?**
- Performance varying with signal strength makes limit setting painful

In Action

CMS-PAS-EXO-22-026

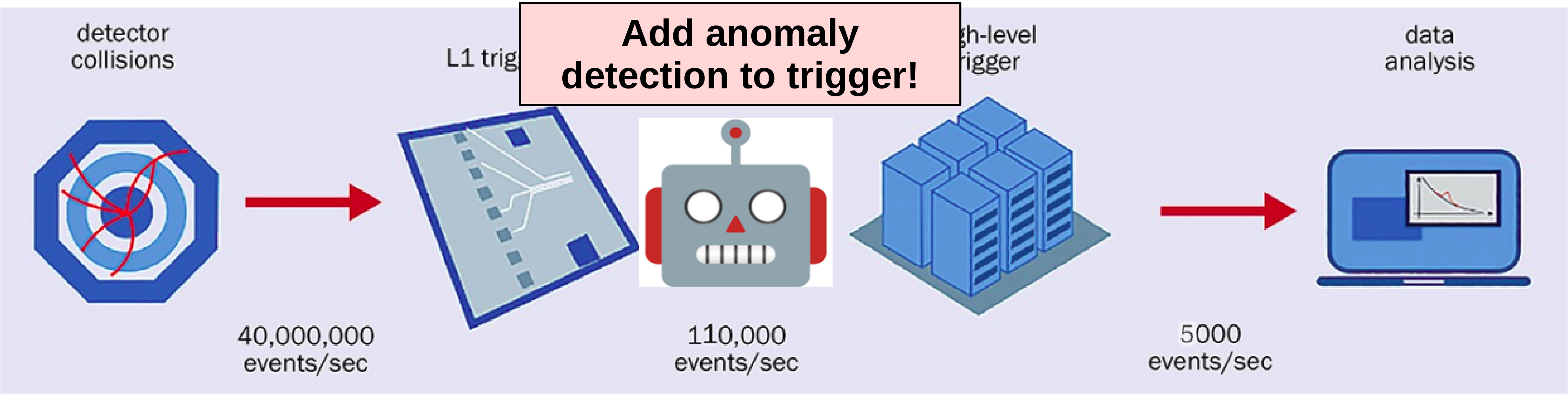
- CMS employed AD in recent search for dijet resonances
  - Anomaly tag substructure of the jets
- Compared multiple different anomaly methods
  - "What xsec do I need for 3/5σ of signal?"
  - Up to factor of 7 gain in discovery sensitivity!
- **Lesson : No one universal, 'best' method**



Trigger

Discarding 99.99% of events from trigger

→ could be missing signals!



Anomaly Detection in Trigger

- CMS has developed **two** an anomaly detection triggers
- Based on autoencoder's trained on zero bias data
- Many 'tricks' used to fit onto FPGA and operate at 40 MHz!!

Global Trigger

AXOL1TL

Calorimeter

CICADA

hls4ml QKeras

AXOL1TL led by FNAL postdoc Abhijith Gandrakota

AXOL1TL CMS-DP-2023-079
CICADA CMS-DP-2023-086

What should I use?

- Anomaly detection is underspecified problem → no single 'optimal' solution
- Method chosen should be tailored to use case
  - If model will only see one event at a time (eg trigger), **must** use outlier detection approaches
  - If you care about 'ultimate' sensitivity, consider weak supervision
  - Can't find suitable mixed samples in data → outlier detection is more universally applicable

Conclusions

- Anomaly detection tries to find signals without specifying them
- Two general philosophies
  - Outlier detection : Learns about background → anomalous = rare under bkg pdf
  - Weak supervision : Use mixed samples to learn S vs B classifier from data
- Both methods have pro's and con's
  - Which to use use depends on situation
- No single 'optimal' method

Tutorial

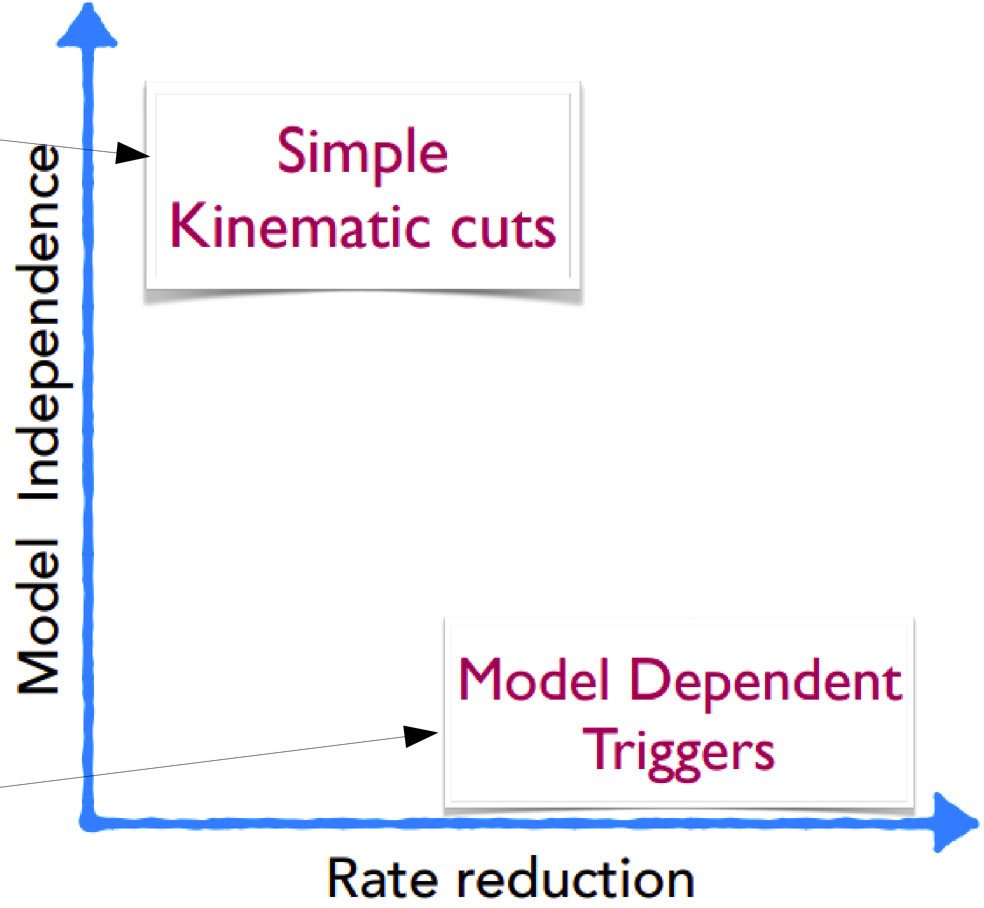
- ‘anomaly_tutorial’ directory includes much more material than we have time to cover
  - Full CATHODE demos and additional variants
  - Credits to Manuel Sommerhalder for building the repo
- We will focus on Gaussian data for simplicity to illustrate the main ideas
- Start with ‘autoencoder_gauss’ and then ‘weak_supervision_gauss’
  - After completing the main notebook, play around with different hyperparameters and see how results change!
  - Continue to other demos if you have time!

Backup

L1 Trigger Strategies

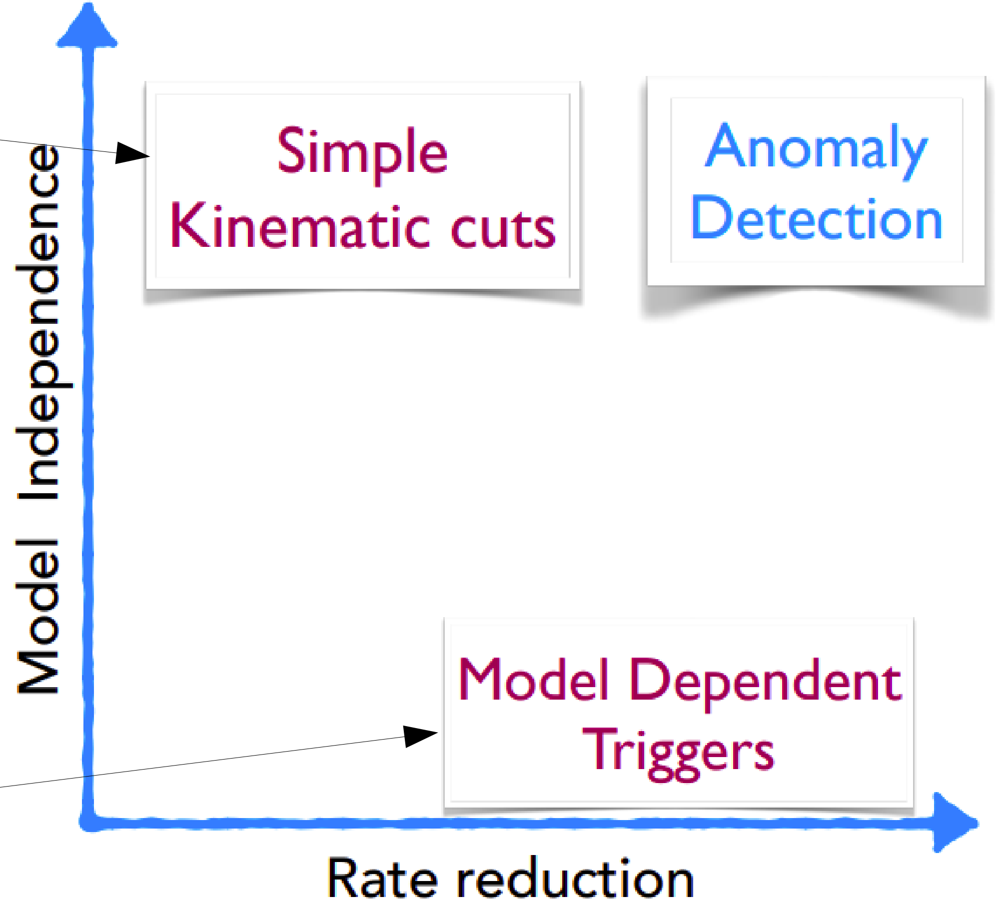
Eg.
Lepton’s with a given p_T
HT, MET, etc.

Eg.
LLP Triggers
VBF Triggers
Etc.



L1 Trigger Strategies

Eg.
Lepton’s with a given p_T
HT, MET, etc.



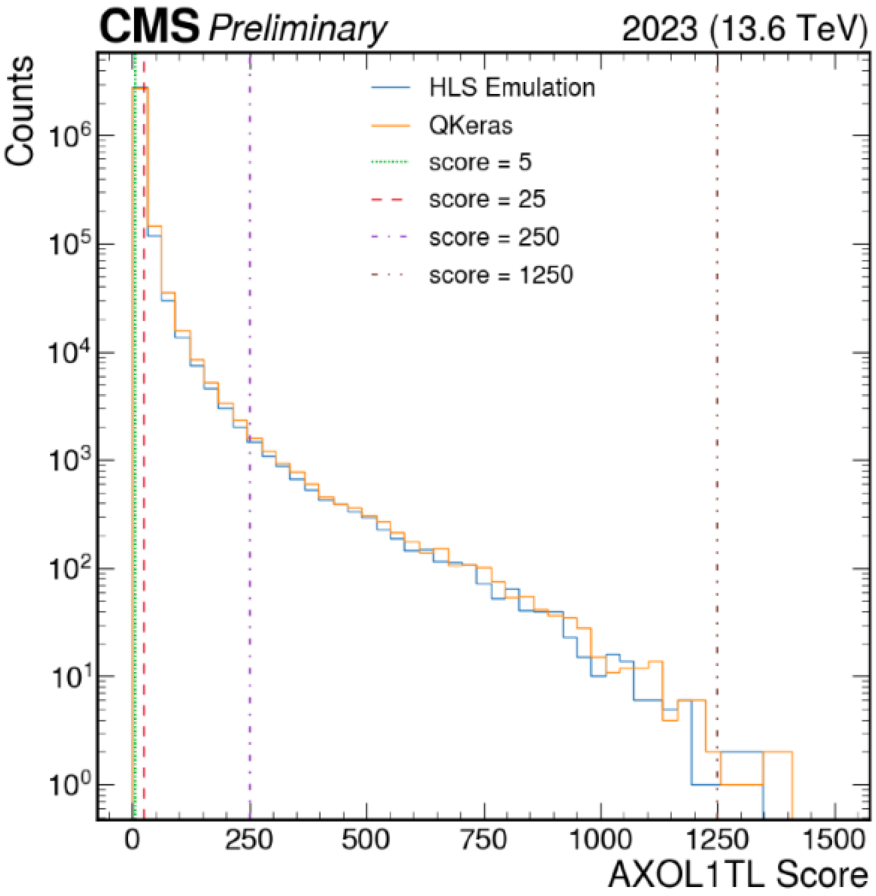
Best of both ?

Eg.
LLP Triggers
VBF Triggers
Etc.

Anomaly Detection at L1

AXOL1TL

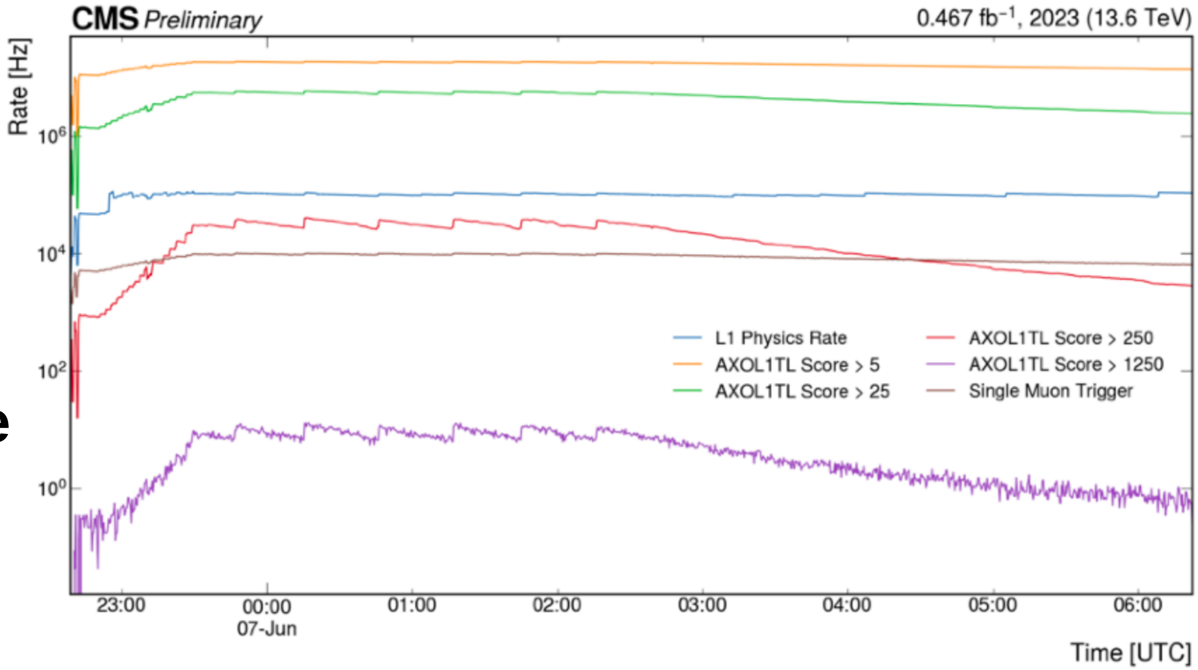
Thresholds on anomaly score chosen to achieve desired rate



CMS-DP-2023-079

In Action!

AXOL1TL was deployed in CMS trigger test crate during 2023 →
rates found to be stable

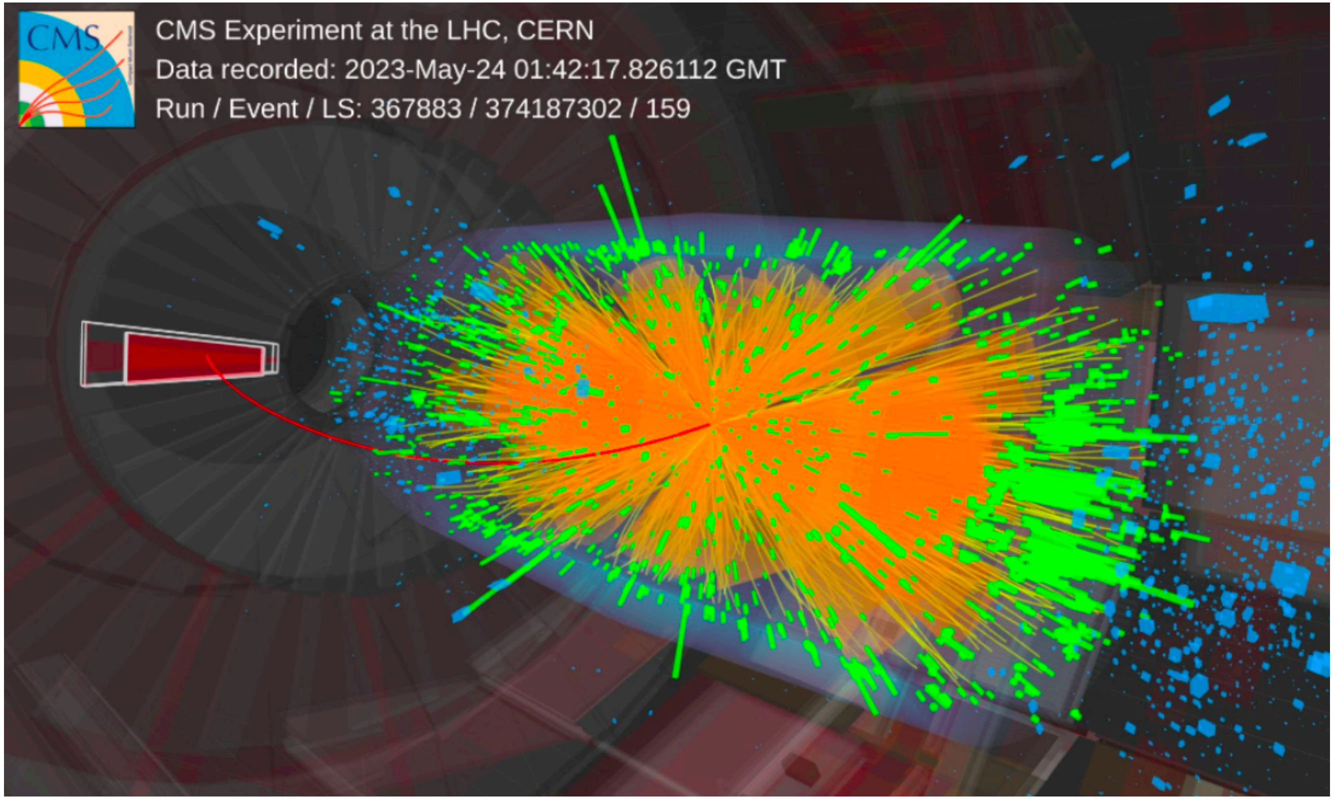


Deployed for real data taking in 2024 !

A L1 Anomalous Event

2023 event triggered only by **AXOL1TL**

Very busy, 11 jets + 1 muon



History

VOLUME 86, NUMBER 17 PHYSICAL REVIEW LETTERS 23 APRIL 2001

Quasi-Model-Independent Search for New High p_T Physics at D0

We apply a quasi-model-independent strategy ("Sleuth") to search for new high p_T physics in $\approx$100 pb^{-1} of $p\bar{p}$ collisions at $\sqrt{s} = 1.8$ TeV collected by the D0 experiment during 1992–1996 at the Fermilab Tevatron. We systematically analyze many exclusive final states and demonstrate sensitivity to a variety of models predicting new phenomena at the electroweak scale. No evidence of new high p_T physics is observed.

PHYSICAL REVIEW D **78,** 012002 (2008)

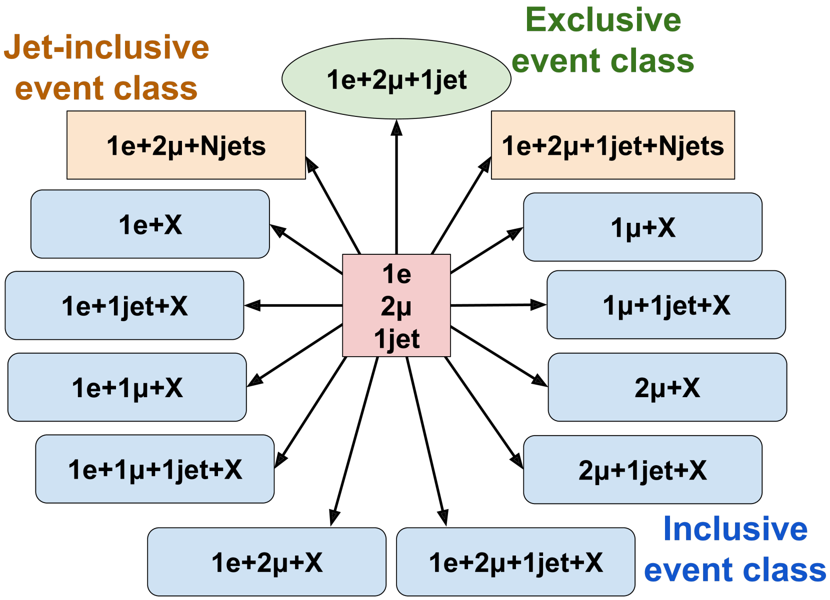
Model-independent and quasi-model-independent search for new physics at CDF

"Vista"

Classic Strategy

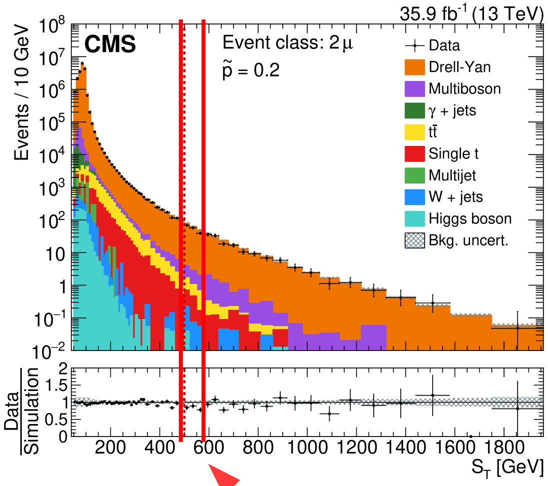
Using CMS MUSiC Search as an example

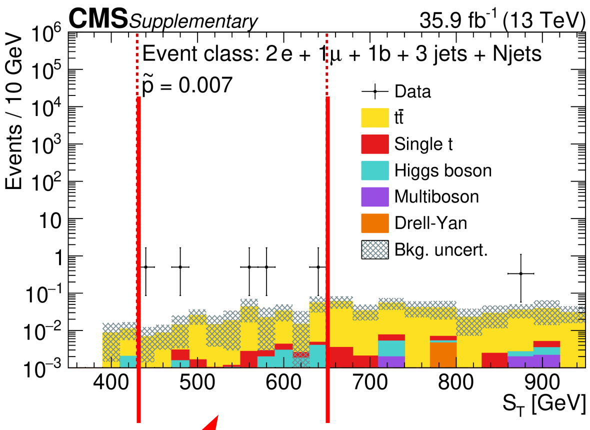
Categorize



~1.5k event classes

Data-MC Comparison



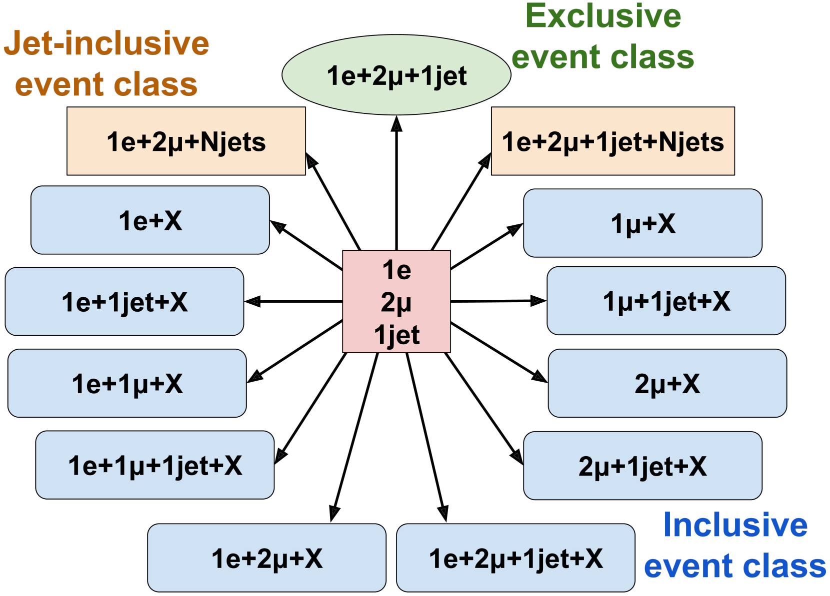


Find Largest Local Deviations

Classic Strategy

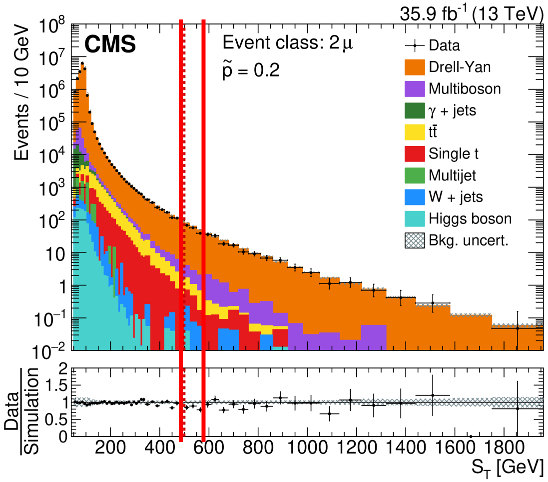
Using CMS MUSiC Search as an example

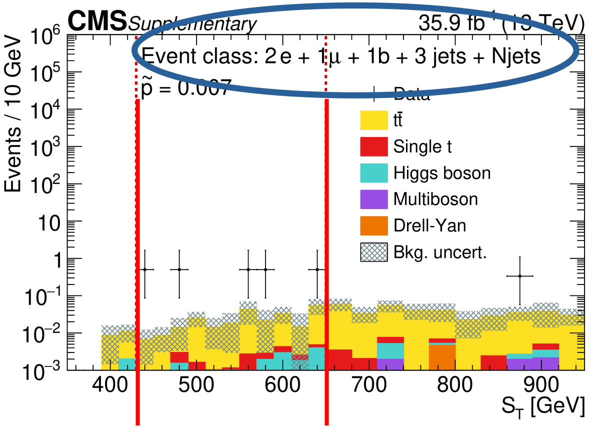
Categorize

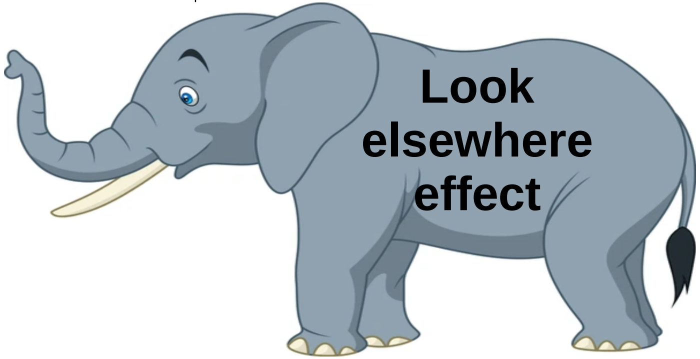


~1.5k event classes

Data-MC Comparison







Modern ‘Anomaly Detection’

The LHC Olympics 2020

A Community Challenge for Anomaly Detection in High Energy Physics

arXiv: 2101.08320

- Focus on a single topology at a time
- Entirely **data-driven**
- Novel ML methods to reduce bkg

Modern 'Anomaly Detection'

The LHC Olympics 2020

A Community Challenge for Anomaly Detection in High Energy Physics

- Focus on a single topology at a time

The Philosophy

"No free lunch" → Drop full model independence

But "**discounts for buying in bulk"!**

→ Cover a large model space in an efficient way

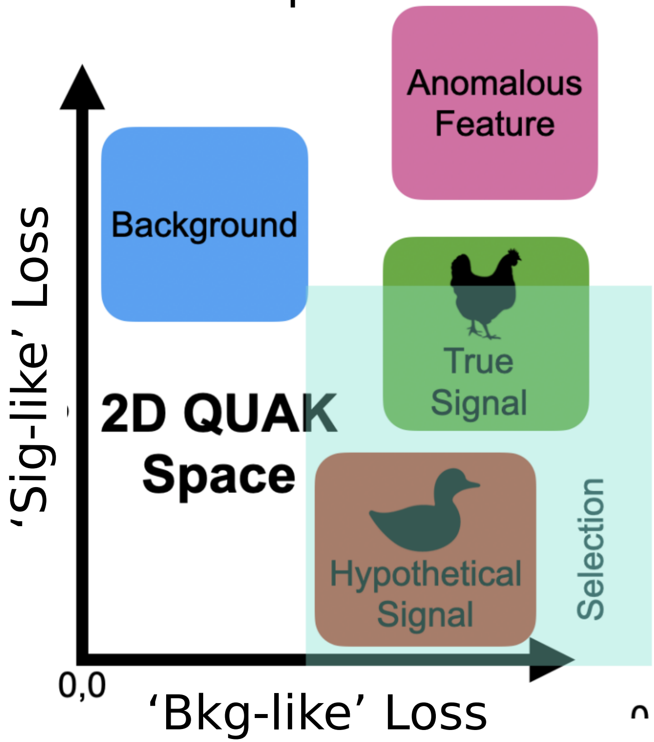
o

arXiv: 2101.08320

Quasi Anomalous Knowledge (QUAK)

- **Hybrid approach** between fully model-indep. and standard search
- **Encode a prior** on what a potential signal may look like
  - Use an AE trained on a variety of different signal MC's
- Construct 'QUAK space':
  - Loss of signal AE vs bkg AE
- Select events with low sig loss and high bkg loss

Hypothetical QUAK Space



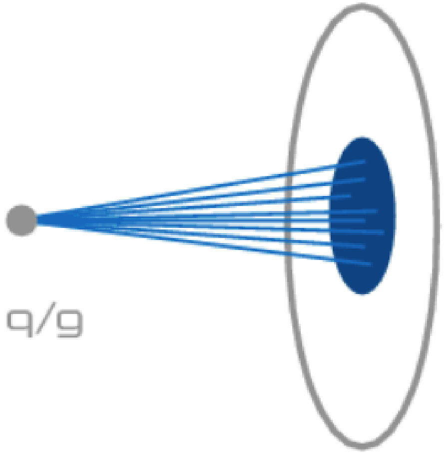
[Park et al 2011.03550]

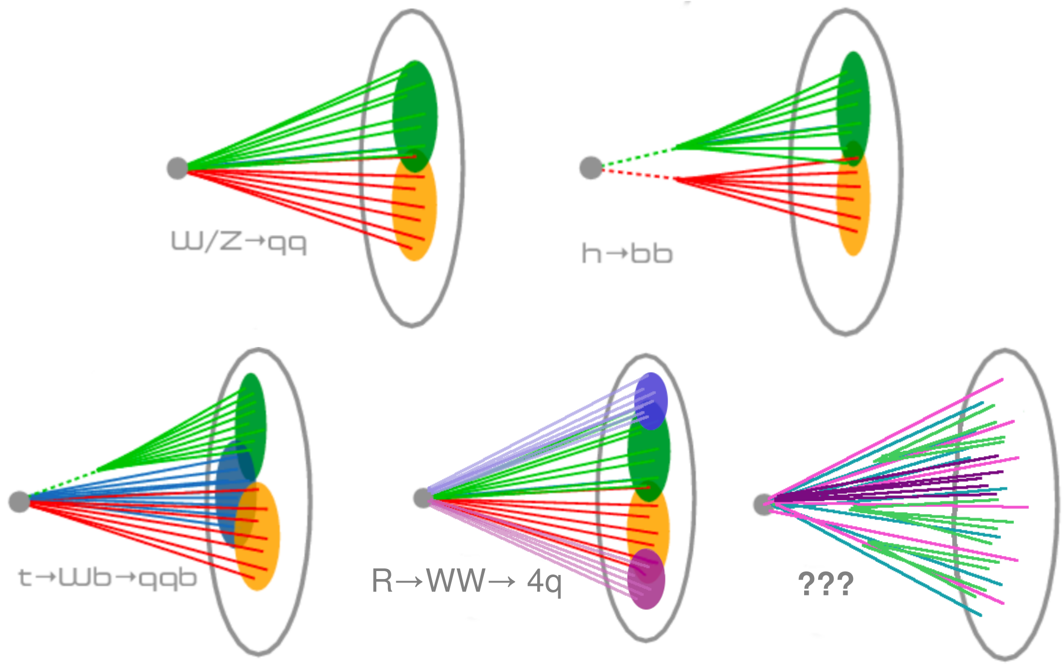
Input Features

Low-level features	Hand-picked high-level features			
VAE	**CWoLa Hunting**	**TNT**	**CATHODE**	**QUAK**
Jet Constituents p_x, p_y, p_z	Jet mass τ_{21} τ_{32} τ_{43} N_{const} Leptonic energy frac. Sub-jets b-tag score	Same as CWoLa Hunting	Jet masses τ_{41}'s ------------------- **CATHODE-b** **+** Subjet b-tag scores	ρ = jet mass / p_T τ_{21}'s τ_{32}'s τ_{43}'s N_{const}'s $\sqrt{\tau_{21}}/\tau_1$ Sub-jets b-tag scores

Graphics source

Jet Substructure





Typical jet

- One central axis (prong)
- From primary vertex
- ...

Anomalous jets

- Multiple prongs
- Displaced vertices
- ???