Introduction to anomaly detection

Barry Dillon

Universität Heidelberg
Zukunft. Seit 1386.

ML4Jets 2022 - Rutgers University
02/11/22

Model-agnostic new physics searches

Pre-ML anomaly detection

ATLAS → General search strategy [Eur. Phys. J. C 79 (2019) 120 (1807.07447)]

Compares data to simulation in (a few) 1D distributions of high-level observables

Looking for statistically significant deviations of the event counts in a data selection

Data selection → over 700 event classes considered, 10^5 signal regions

Outcome: data-derived signal regions for a dedicated analysis on new data

ATLAS EXPERIMENT

CERN

Eur. Phys. J. C 79 (2019) 120
DOI: 10.1140/epjc/s10052-019-6540-y

CERN-EP-2018-070
19th February 2019

A strategy for a general search for new phenomena using data-derived signal regions and its application within the ATLAS experiment

The ATLAS Collaboration

Model-agnostic new physics searches

Pre-ML anomaly detection

ATLAS → General search strategy [Eur. Phys. J. C 79 (2019) 120 (1807.07447)]

Compares data to simulation in (a few) 1D distributions of high-level observables

Looking for statistically significant deviations of the event counts in a data selection

Data selection → over 700 event classes considered, 10^5 signal regions

Outcome: data-derived signal regions for a dedicated analysis on new data

The Pros

- **wide coverage of possible signals**
- independent data selections can trigger in combination
- only a few regions chosen for dedicated study
  ⇒ reduced look-elsewhere effect
- truly global interpretation of the data

The Cons

- reliance on simulation
- **sensitive to choice of observables!**
- only uses a few high-level observables
- only uses 1D distributions

… & machine learning?

Neural networks ⇒ interpolating power

Deep-learning has demonstrated potential for analyses with…

- no restriction to a few, or even high-level, observables
- the ability to process high-dimensional datasets
- less reliance on simulation
- reduced look-elsewhere effect

Interesting questions!

- Physics bias in deep-learning approaches?
- How model-agnostic are the approaches?
- What would an analysis with these tools look like?

How do we define ‘anomalous’?

- Deviations from Standard Model simulations

  Regions of phase space, or individual jets/events, which are outliers wrt the SM

- Group anomalies

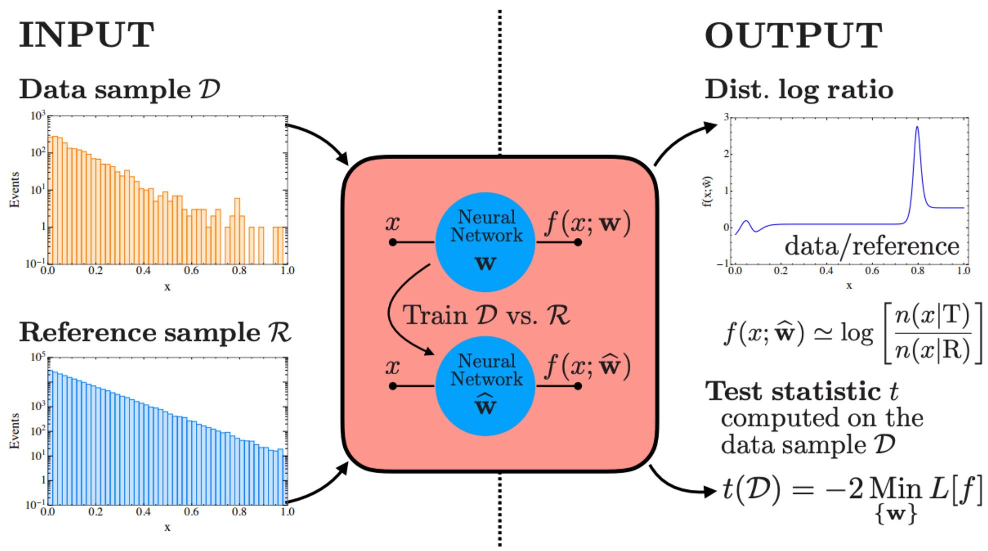
  Groups of events that form local over-densities in some observable

- Low density regions of the feature space

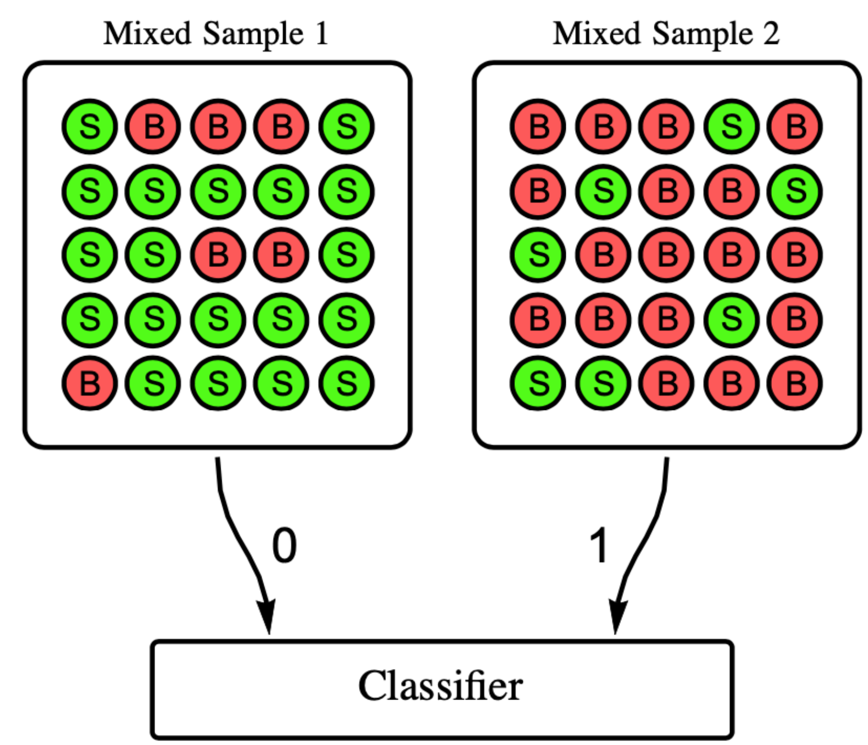
  Jet/events which are assigned a low-density according to some density model of either the data, or some reference model

ML-based anomaly detection

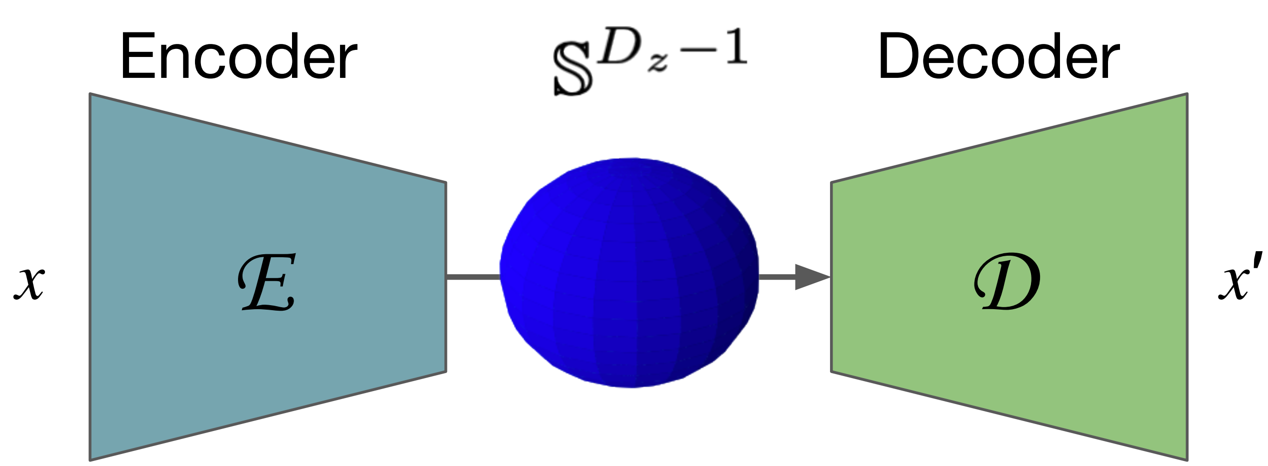
1 - Simulation vs experiment



2 - Classification Without Labels



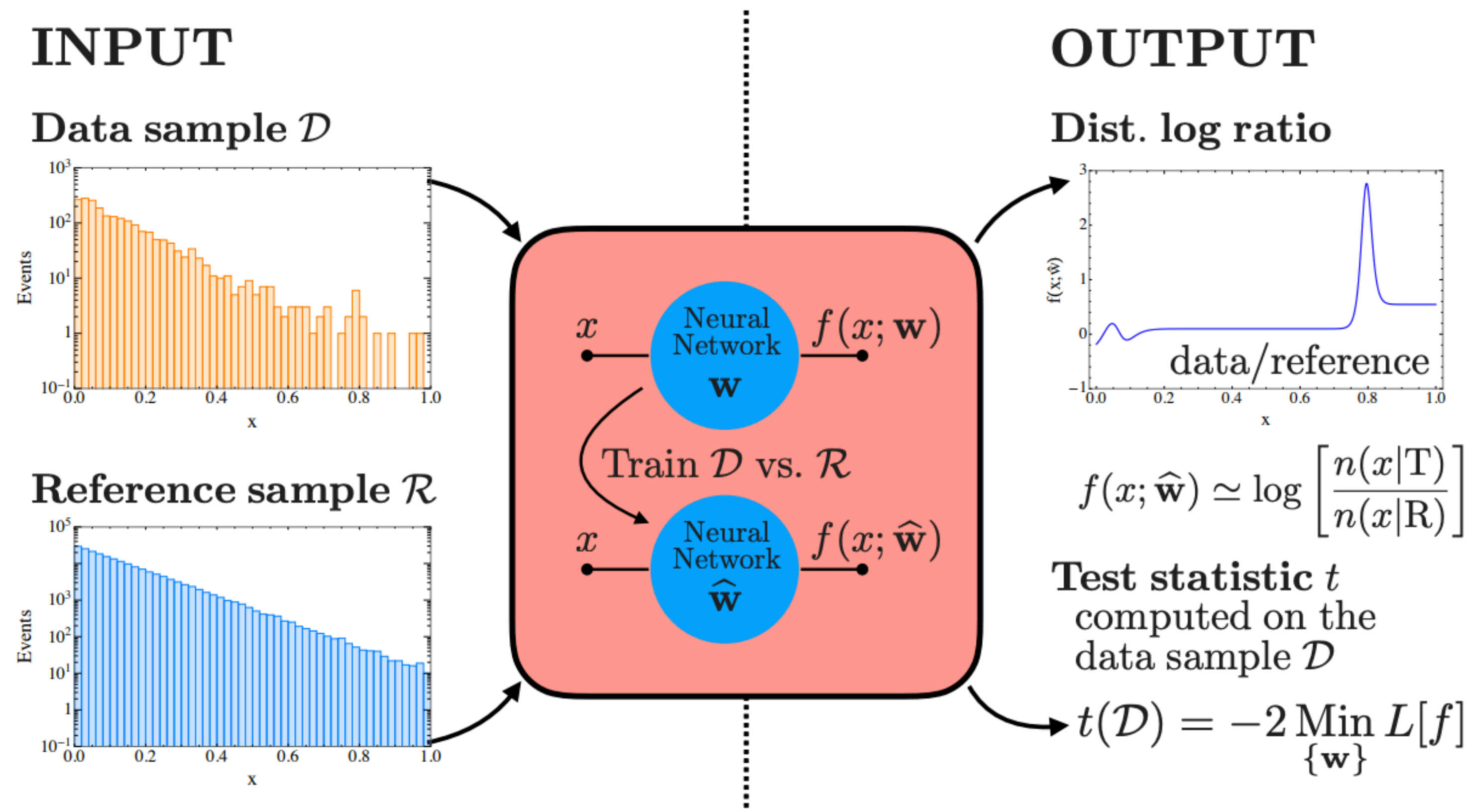
3 - AutoEncoders



Simulation vs data

['Learning new physics from a machine' - D'Agnolo & Wulzer]

- Classifier trained between data sample and a reference model
- Binned distributions are replaced by smooth functions approximated by neural networks.
- Procedure gives a p-value and the likelihood ratio



- Sensitive to both group and low-density anomalies
- Reliant on simulation
- But scales badly for many observables…

$\rightarrow$ perhaps ways to overcome this
['Learning new physics efficiently with nonparametric methods' - Letizia et al]

The CWoLa method [Metodiev, Nachman, Thaler (1708.02949)]

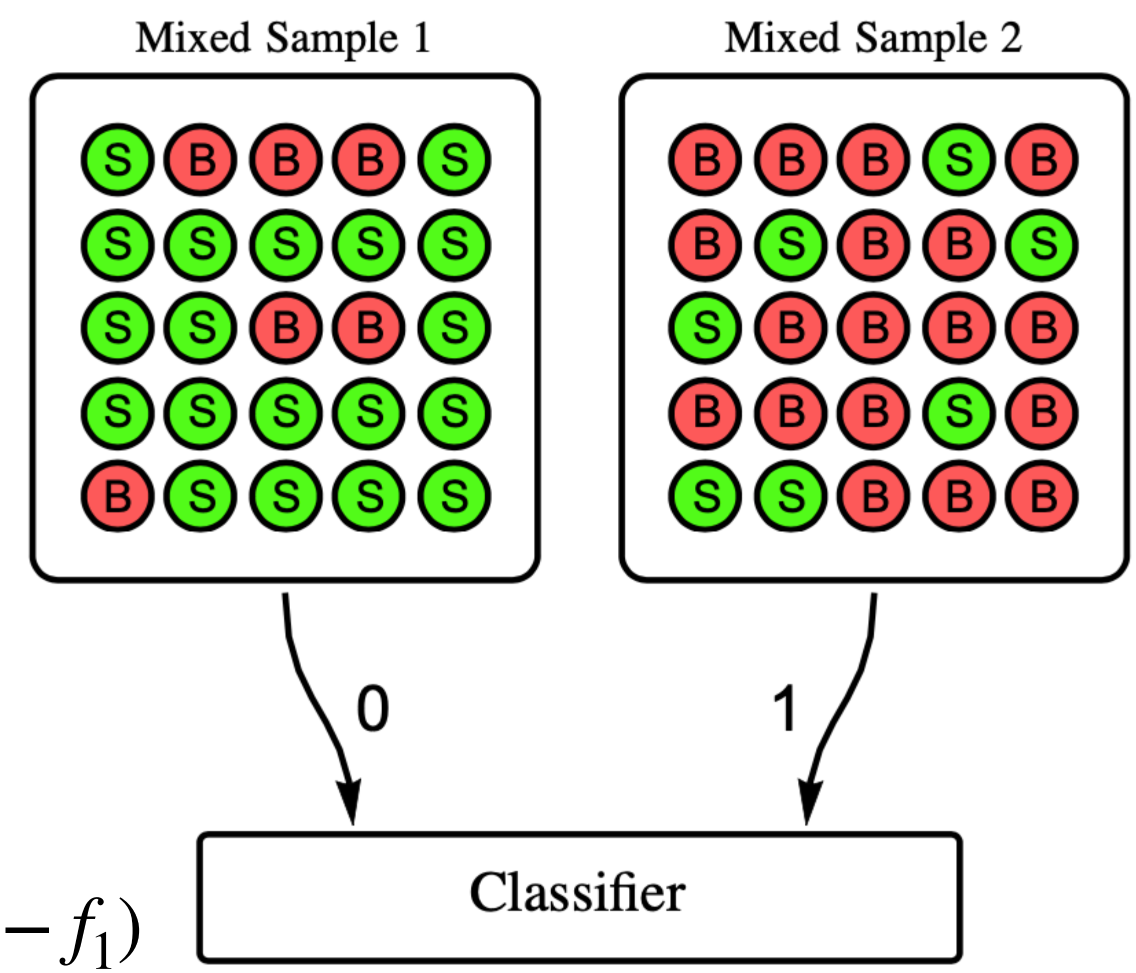
- Two samples, one signal-enriched

- Train a supervised classifier to distinguish between them

- If signals and backgrounds come from the same underlying distributions, then we learn

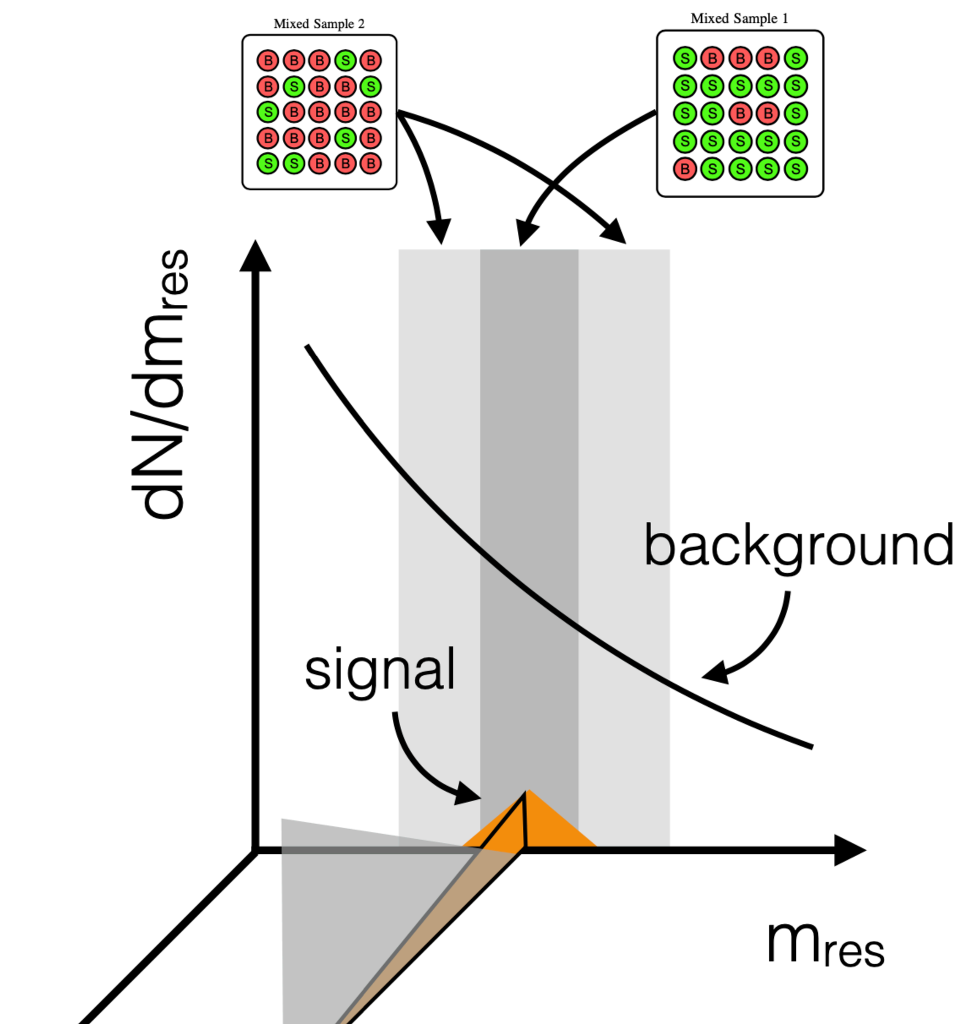
$$L_{M_1/M_2} = \frac{f_1 p_S + (1-f_1)p_B}{f_2 p_S + (1-f_2)p_B} = \frac{f_1 L_{S/B} + (1-f_1)}{f_2 L_{S/B} + (1-f_2)}$$

- Monotonically related to $L_{S/B}$ for $f_1 > f_2$



The CWoLa bump hunt [Collins, Howe, Nachman (1902.02634)]

- Looking for group anomalies in m_{res} spectrum
- Di-jet search: $m_{j1}, m_{j_2}, \tau_{2/1}, \tau_{2/1}$
- Background enriched-region obtained from sidebands
- Cut on classifier to keep Y% of events
- Fit smoothly falling backgrounds
- Estimate p-value



[Stolen from a Ben Nachman talk]

Talks:
CURTAINs for your Sliding Window - Johnny Raine
Resonant anomaly detection without background sculpting - Manuel Sommerhalder
HEP-Sim2Real: creating background templates with normalizing flows - Radha Mastandrea
Weakly supervised methods for LHC analyses - Thorben Finke

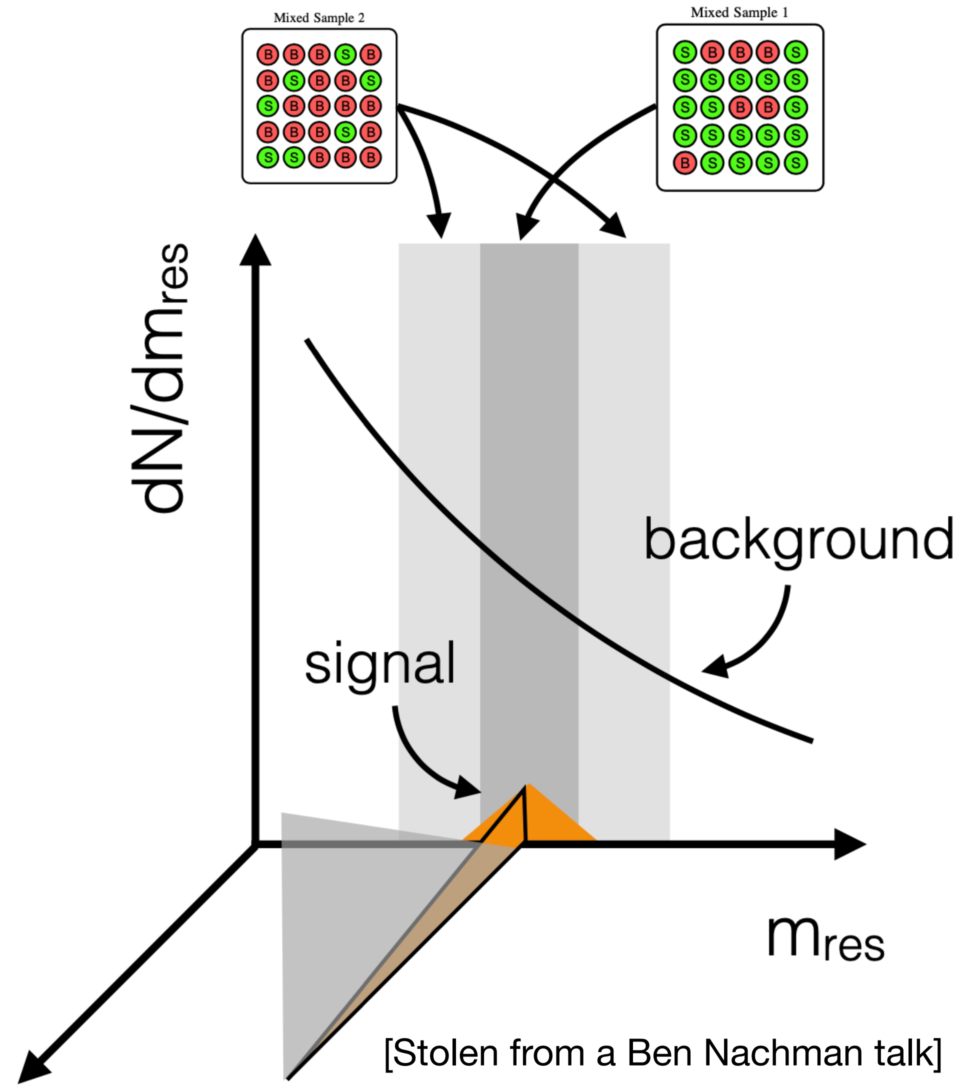
The CWoLa bump hunt [Collins, Howe, Nachman (1902.02634)]

The pros

- Background estimation
- Data-driven
- Completely agnostic wrt auxiliary observables

Model-dependence

- Assumptions on how anomalies are ‘grouped’
- Bad scaling with many observables
  ⇒ choice of observables is v. important



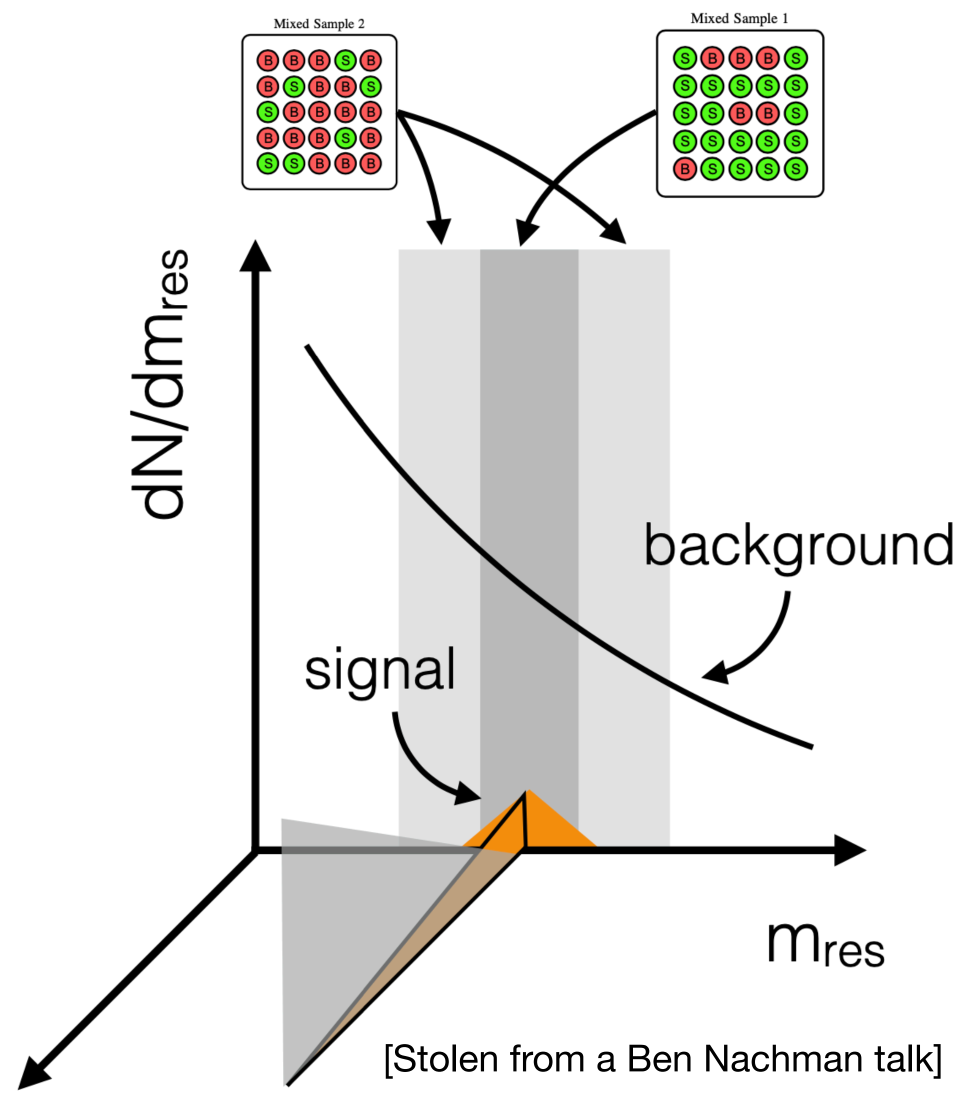
[Stolen from a Ben Nachman talk]

Talks:
CURTAINs for your Sliding Window - Johnny Raine
Resonant anomaly detection without background sculpting - Manuel Sommerhalder
HEP-Sim2Real: creating background templates with normalizing flows - Radha Mastandrea
Weakly supervised methods for LHC analyses - Thorben Finke

The CWoLa bump hunt [Collins, Howe, Nachman (1902.02634)]

ANODE & CATHODE [Hallin et al (2109.00546)] [Nachman et al (2001.04990)]

- Correlations with m_{res} are a big problem
- ANODE:
  interpolates probability densities from sidebands into the signal region & constructs the likelihood ratio
- CATHODE: (& LaCATHODE)
  also interpolates, but samples signal-region background events from the model and builds a classifier to estimate the likelihood ratio



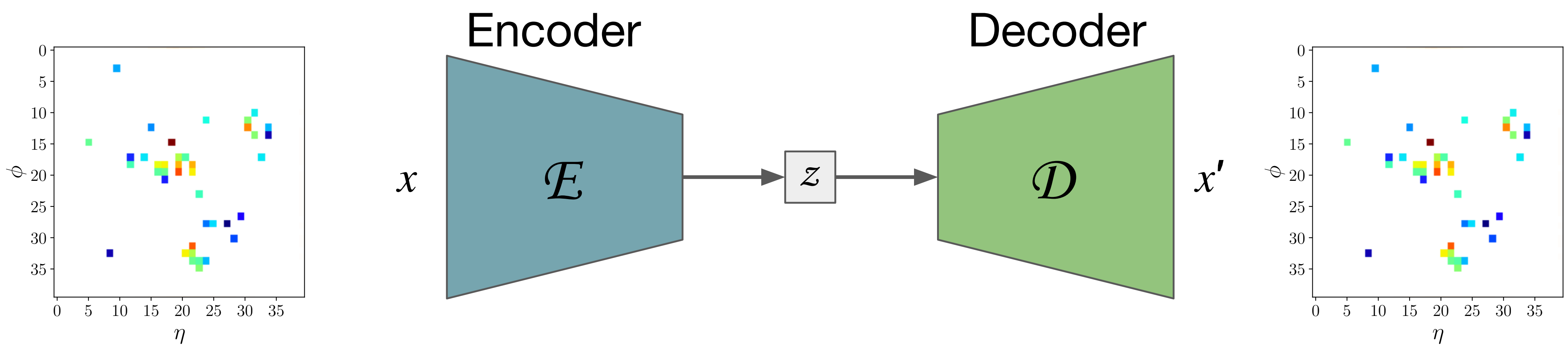
[Stolen from a Ben Nachman talk]

Talks:
CURTAINs for your Sliding Window - Johnny Raine
Resonant anomaly detection without background sculpting - Manuel Sommerhalder
HEP-Sim2Real: creating background templates with normalizing flows - Radha Mastandrea
Weakly supervised methods for LHC analyses - Thorben Finke

AutoEncoder networks

[‘QCD or What?' Heimel et al]

[‘Searching for new physics with deep auotencoders’ Farina et al]

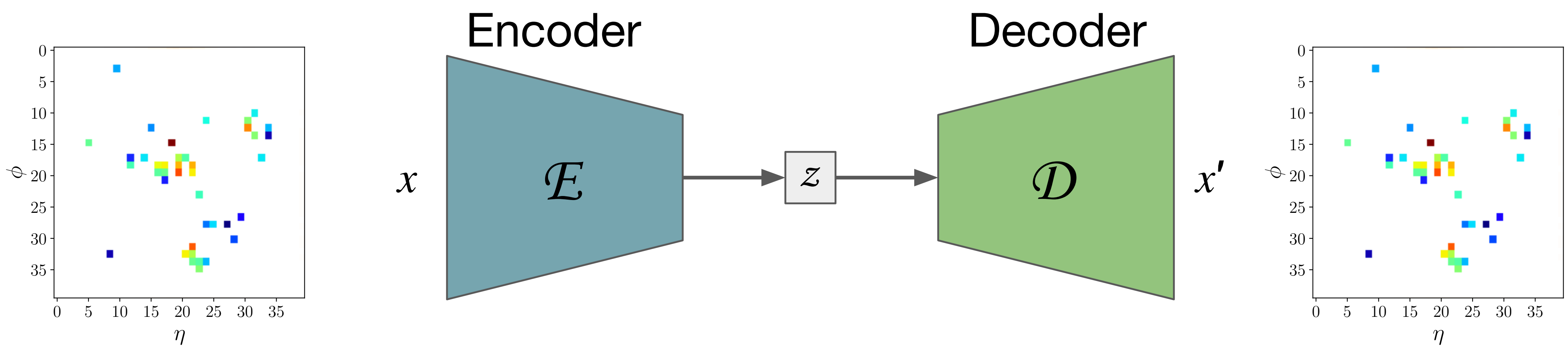


- Trained to reconstruct the data they are trained on
- Encode the most general features of the data in a latent space z
- Optimised on background-dominant data
- Unsupervised $\longrightarrow$ model-agnostic, no labels
- Reconstruction loss: $\mathcal{L} = ||x - x'||^2$
- Anomalous data $\Rightarrow$ data the network has seen least $\Rightarrow$ larger reconstruction loss

AutoEncoder networks

['QCD or What?' Heimel et al]

['Searching for new physics with deep auotencoders' Farina et al]



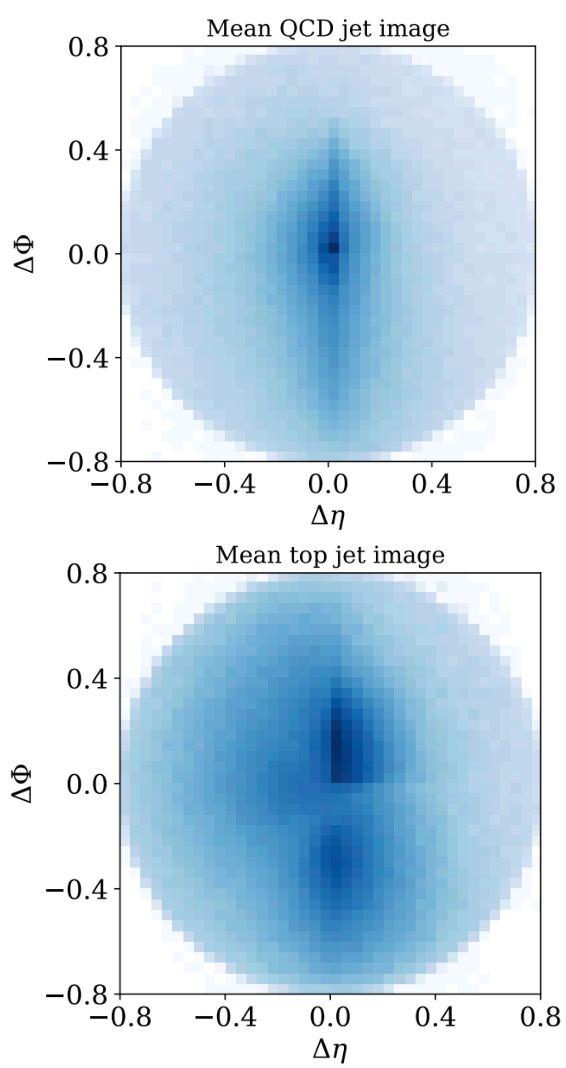
- Trained to reconstruct the data they are trained on
- Encode the most general features of the data in a latent space z
- Optimised on background-dominant data
- Unsupervised ⟶ model-agnostic, no labels
- Reconstruction loss: $\mathcal{L} = ||x - x'||^2$
- Anomalous data ⇒ data the network has seen least ⇒ larger reconstruction loss

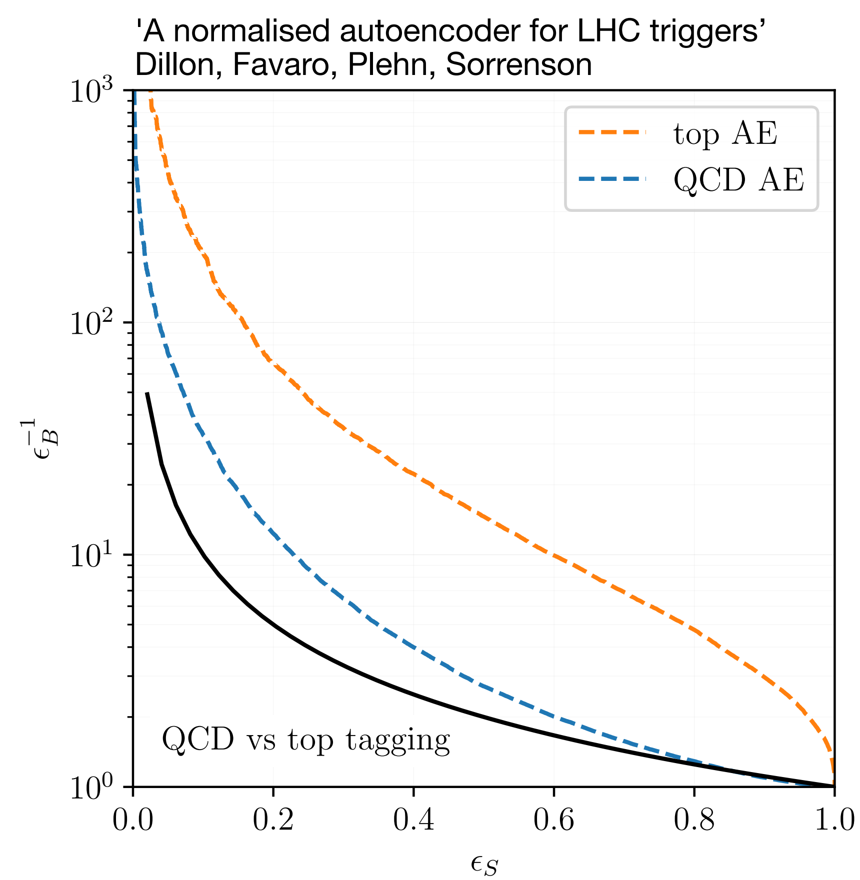
Has proved quite successful, but…

AutoEncoder networks - the problems

They don't robustly identify anomalous jets.

They do robustly identify complex jets,   e.g anomalous top/QCD jets





An AE trained on only top jets learns to reconstruct QCD jets…

AutoEncoder networks - the problems

They don't robustly identify anomalous jets.

They do robustly identify complex jets,   e.g anomalous top/QCD jets

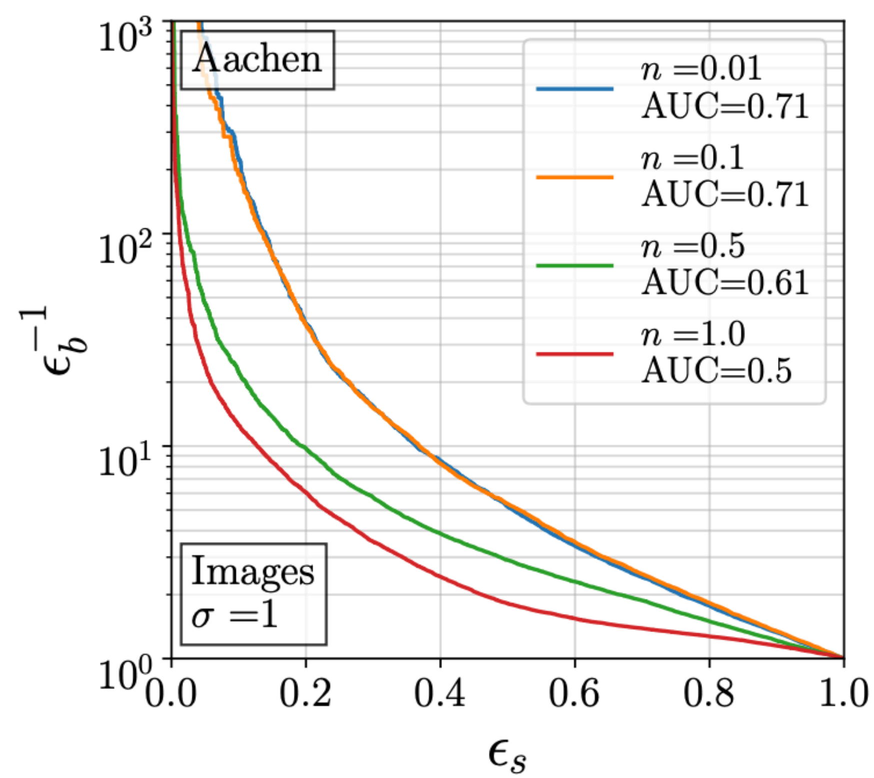
Very sensitive to the choice of representation / observables

e.g. under re-mapping of p_T's, $p_T \to p_T^n$

results vary a lot

['What's anomalous in LHC jets?' Buss et al]

['Anomaly detection under coordinate transformations' Kasieczka et al]



AutoEncoder networks - the problems

They don’t robustly identify anomalous jets.

They do robustly identify complex jets,   e.g anomalous top/QCD jets

Very sensitive to the choice of representation / observables

e.g. under re-mapping of p_T’s, $p_T \to p_T^n$

results vary a lot

[‘What’s anomalous in LHC jets?’ Buss et al]

[‘Anomaly detection under coordinate transformations’ Kasieczka et al]

On low level data: not invariant to physical symmetries in the problem.

AE can’t reconstruct something the latent space is invariant to…

Preprocessing is necessary, but approximate.

Density-based anomaly detection

Reconstruction is a very vague way to define anomalies

More accurately: anomalies are events/jets in low density regions of the feature space

$\Rightarrow$ not invariant to transformations in feature space

Machine-learned density estimation:

1 - some parameterisation of the density $p_{\text{data}}(\vec{x})$

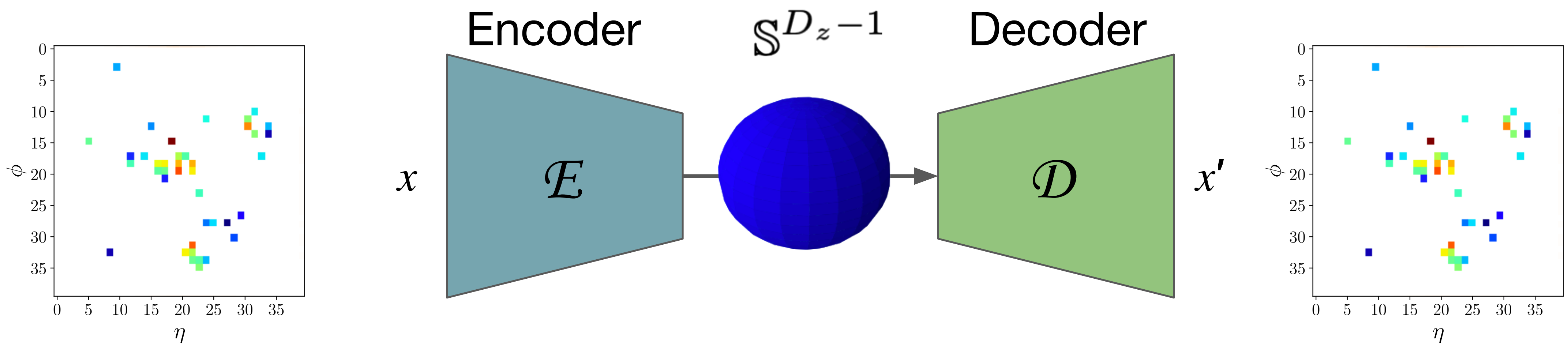
2 - a scheme to minimise $-\log p_{\text{data}}(\vec{x})$ wrt to the parameters

Can be difficult in high-dimensions!

Flows work well in low dimensional problems

['What's anomalous in LHC jets?' Buss et al]

Density-based anomaly detection



→ the normalised autoencoder

['Autoencoding under normalization constraints' - Yoon et al]

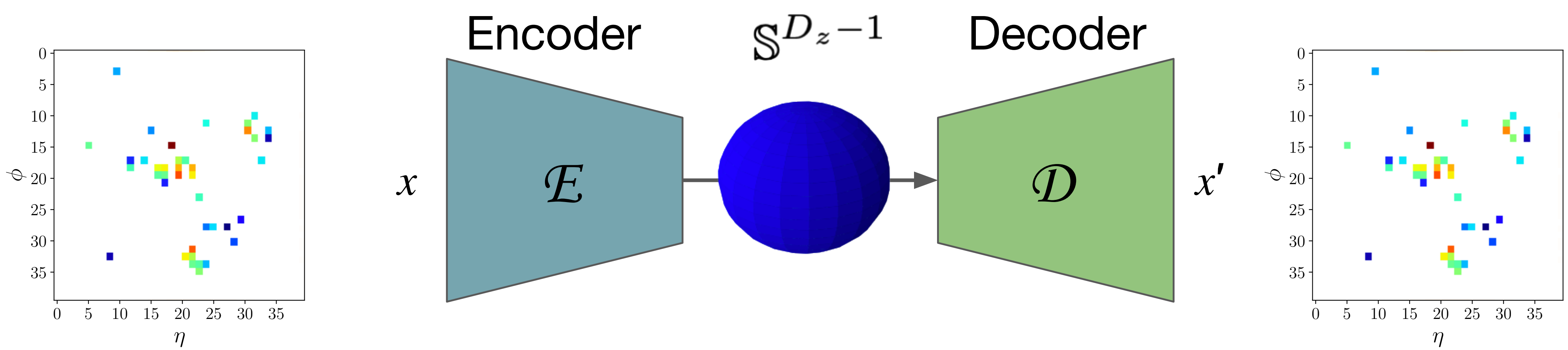
['A Normalized Autoencoder for LHC triggers' - Dillon et al]

[See Luigi Favaro's talk today: A Normalized Autoencoder for LHC triggers]

→ density estimation with an AutoEncoder architecture

- Energy-based model
- Overcomes the complexity bias in AEs
- Better understanding of preprocessing dependence
- More robust

Density-based anomaly detection

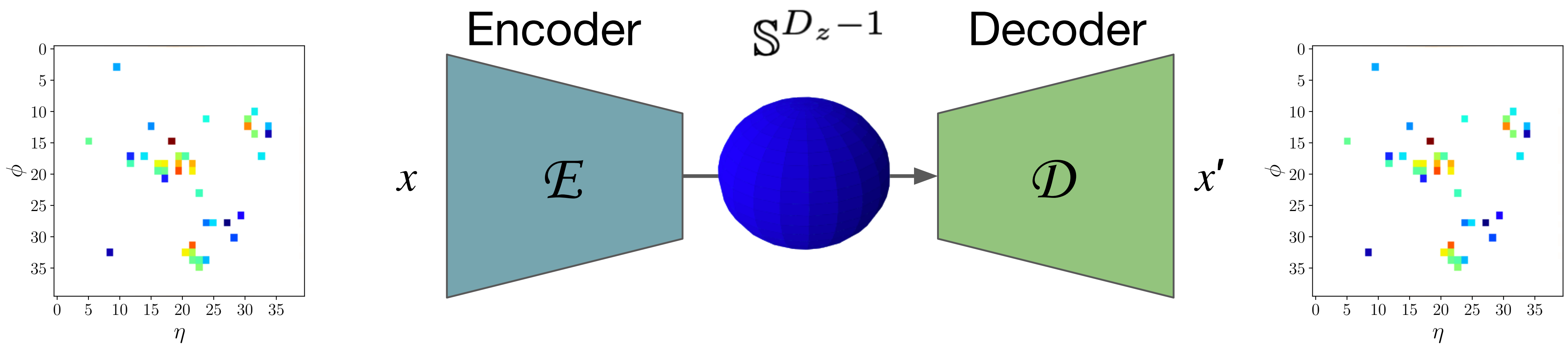


What about an analysis?

- Autoencoders are just tools, need to compute p-values
- Could compare to simulation, use data-driven methods, …
- Could use the ABCD technique:
  - train two (DisCo) decorrelated AEs simultaneously on the dataset
  - use cuts to determine A-B-C-D regions
  - background estimation in signal-region

[‘Online-compatible Unsupervised Non-resonant Anomaly Detection’ Mikuni et al]

Density-based anomaly detection



The Pros

- Can process high-dimensional data
- Complexity-bias issue is solved
- Highly performant
  … even for very small signal fractions
- Can be run online, i.e. on trigger

Model dependence

- Not invariant to transformations in the feature space
- Additional model dependence, or dependence on simulation, arises at the analysis stage

[see talk ‘Challenges for unsupervised anomaly detection in particle physics’ - Katherine Fraser]

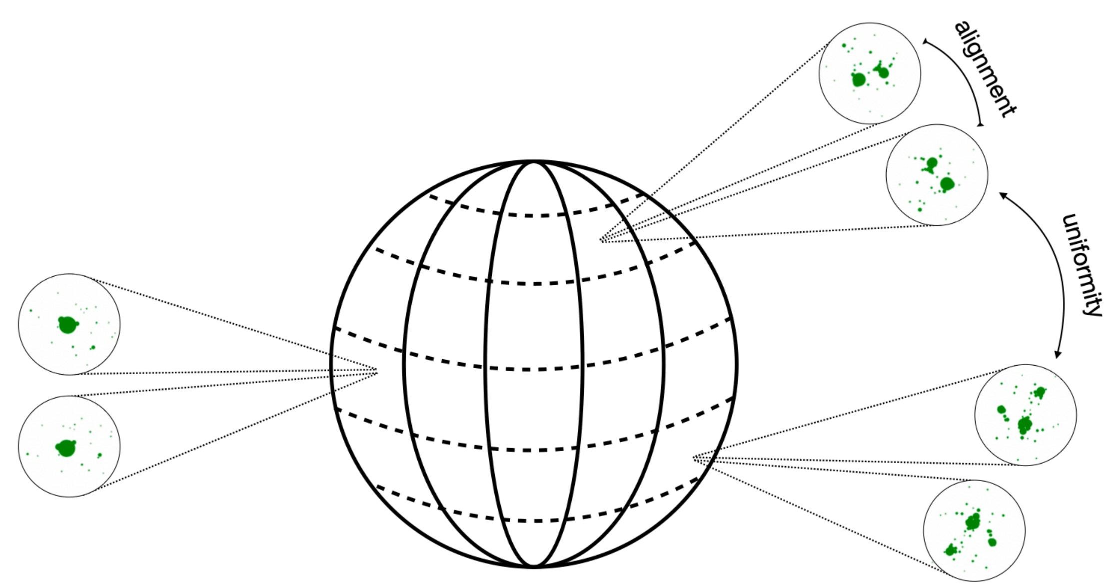
Representation learning

['Symmetries, Safety, and Self-Supervision' Dillon et al]
['Self-supervised anomaly-detection' Dillon, Mastandrea, Nachman]
['Invariant representation driven neural classifier for anti-QCD jet tagging' Cheng et al]

The CWoLa method and density estimation both have issues with choosing the best set of observables.

Solutions?

- Manual choice of observables / representations
- Self-supervised (data-derived) representation
  → contrastive learning
  → invariant to symmetries
  → incorporate kinematics & particle IDs
- Still a new idea, more results to come!

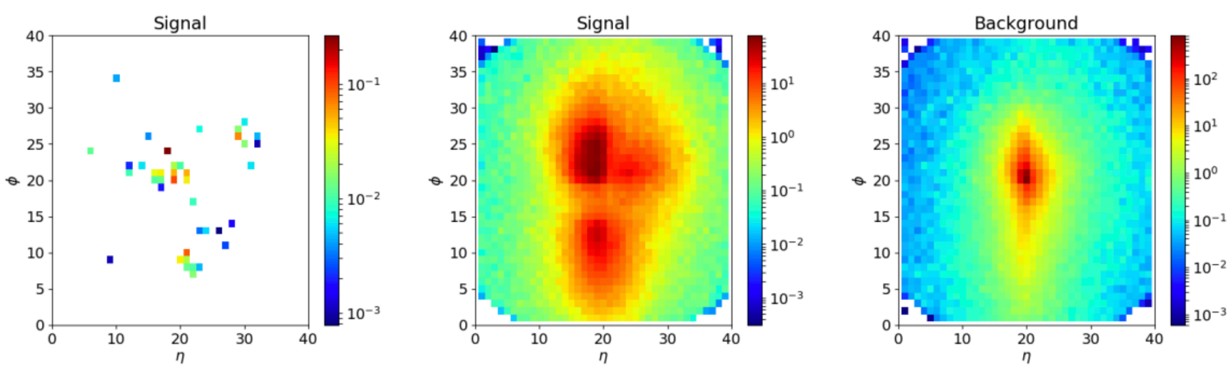


Talks:
Symmetries, Safety, and Self-Supervision - Peter Sorrenson
Topological Data Analysis for Collider Events - Tianji Cai
Robust anomaly detection using NuRD - Abhijith Gandrakota

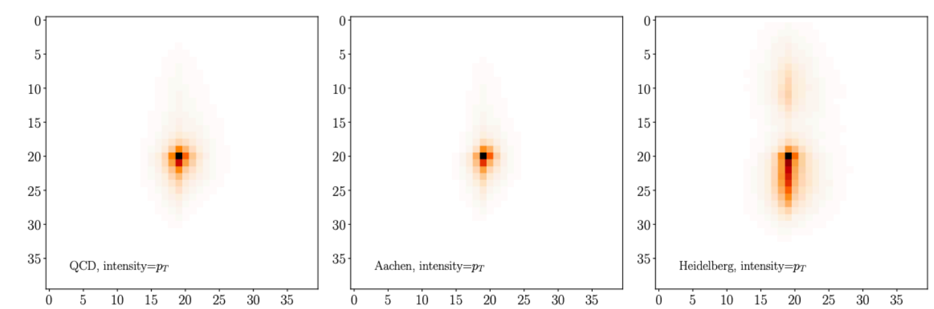
Public challenges & benchmark datasets

Jets

Top-tagging
[Kasieczka, Plehn et al (1902.09914)]

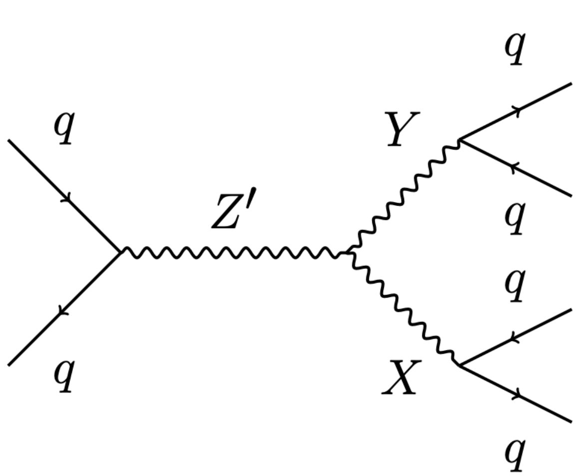


Semi-visible jets
[Buss et al (2202.00686)]

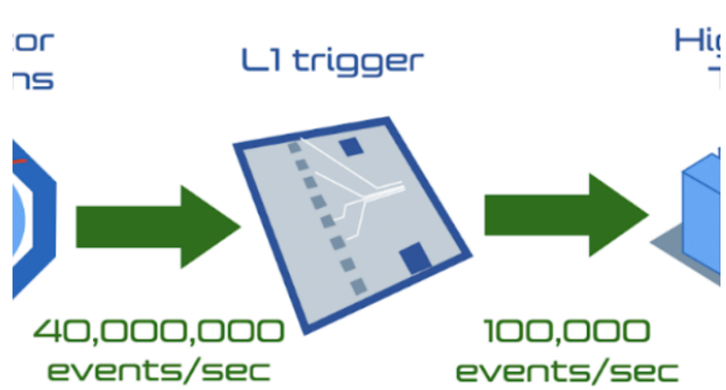


Events

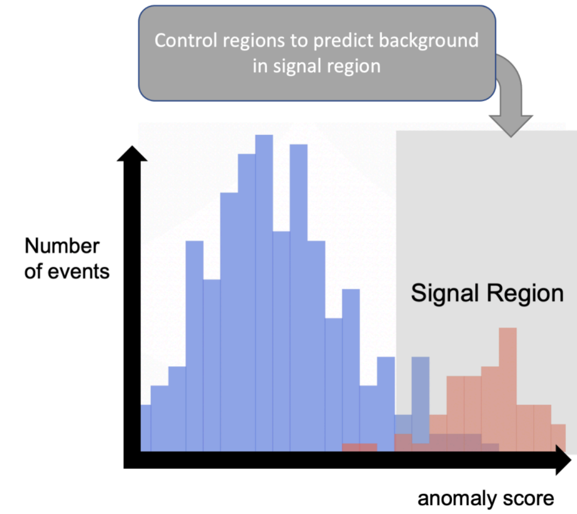
LHC Olympics
[Kasieczka, Nachman, Shih et al (2101.08320)]



CMS trigger challenge
[Govorkova et al (2107.02157)]



Dark Machines challenge
[Ostdiek et al (2105.14027)]



Outlook

- Anomaly detection is complicated
- Each technique and each physics scenario is very different
- But, some very interesting new ideas being developed:
  - ever more performant implementations of the CWoLa bump hunt
  - new applications of the CWoLa idea
  - better AutoEncoding tools
  - bette techniques for choosing representations for anomaly detection
- Experimental anomaly detection analyses
  - CMS trigger challenge [Govorkova et al (2107.02157)]
  - ATLAS weakly supervised di-jet search [ATLAS Collaboration (2005.02983)]
  - more I am not aware of…? See next talk by Julia Gonski!