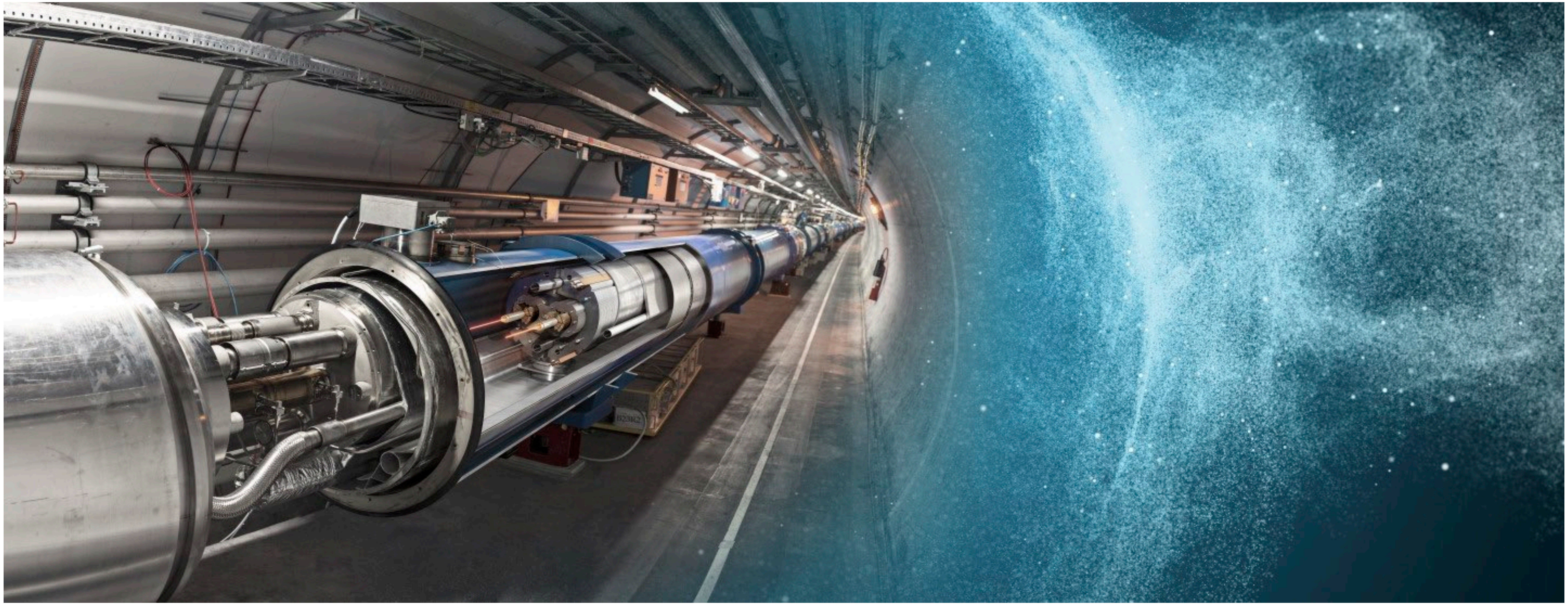
Data Analysis and Bayesian Methods
Lecture 5

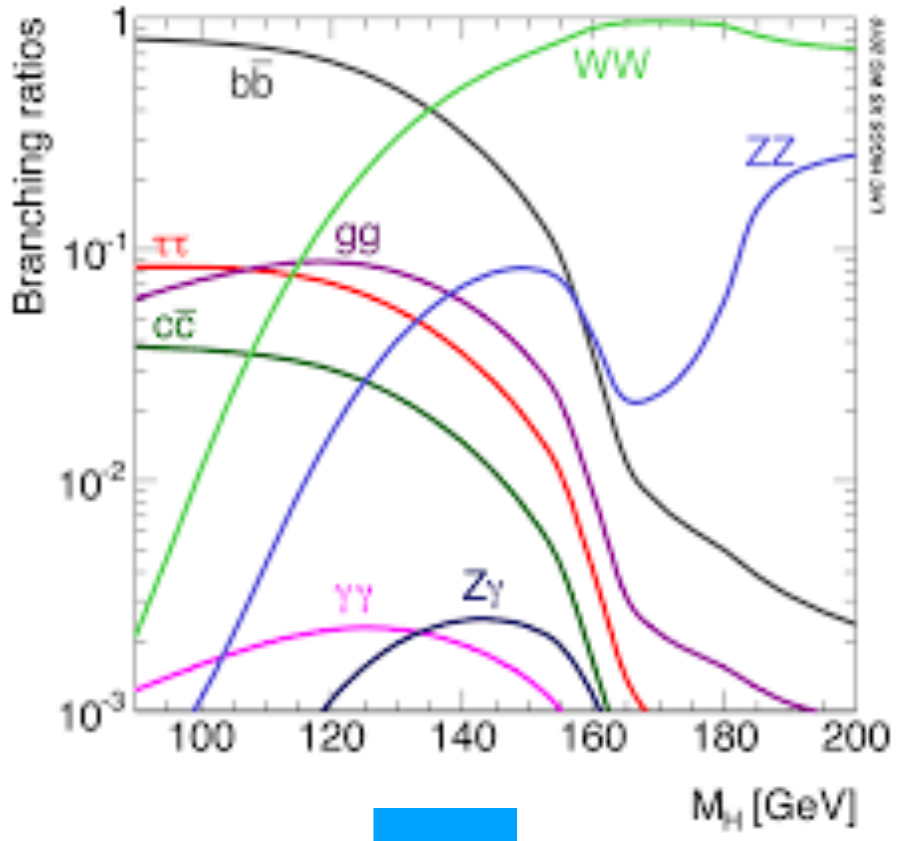
Maurizio Pierini

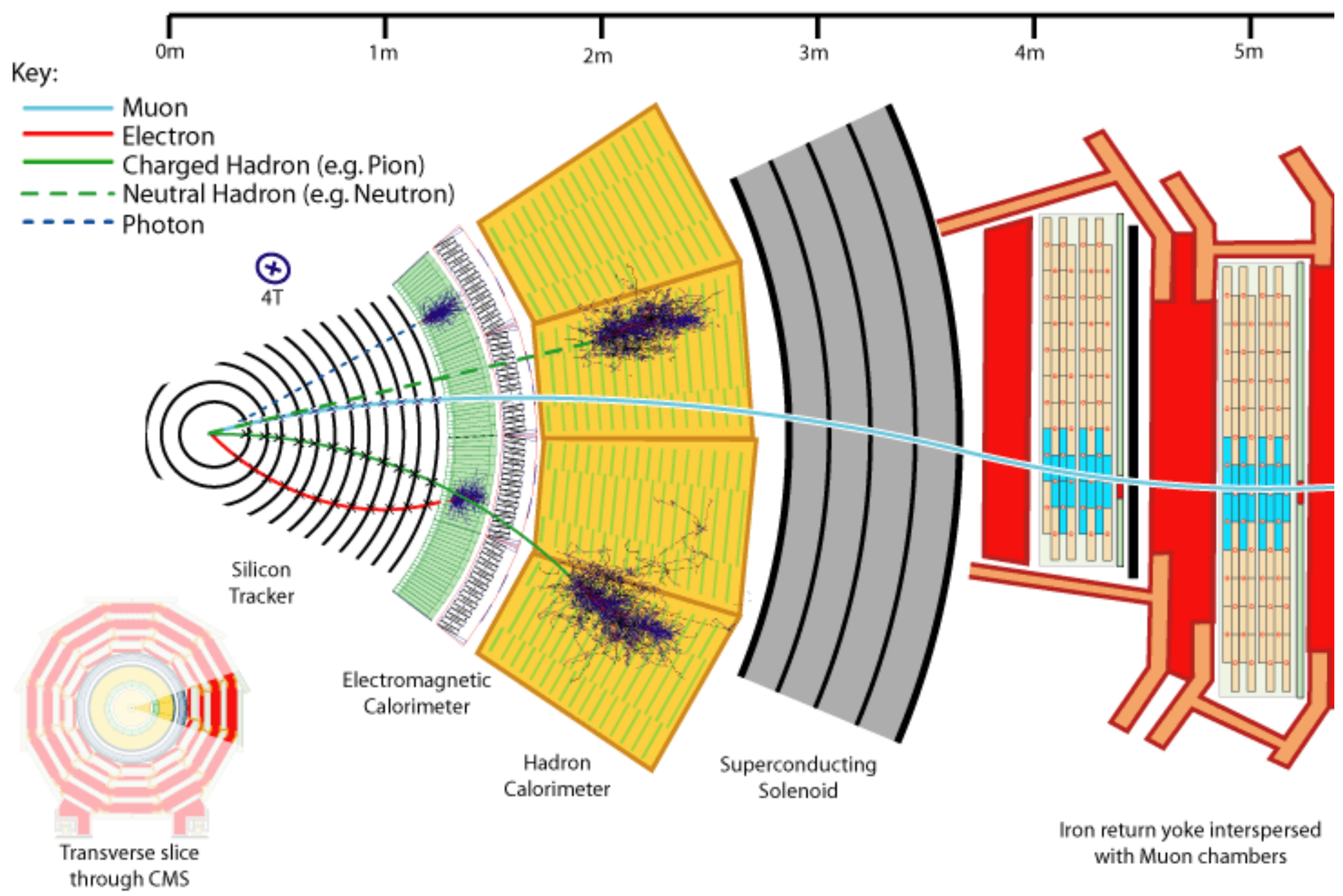


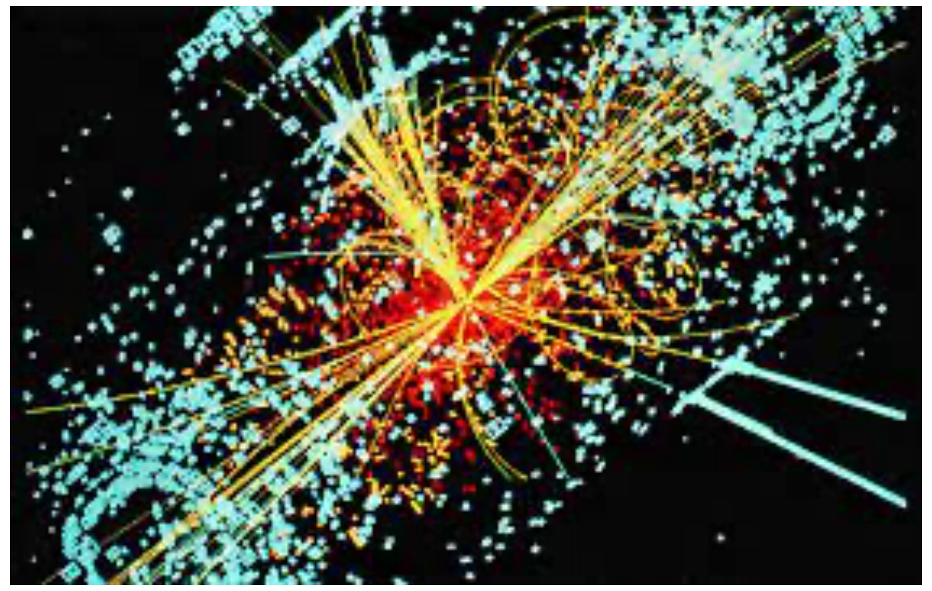
Anomaly detection with Deep Learning

LHC as a discovery machine

- The LHC was mainly built to discover the Higgs boson
  - ATLAS & CMS were designed to cover the meaningful mass range for a particle that was fully characterized

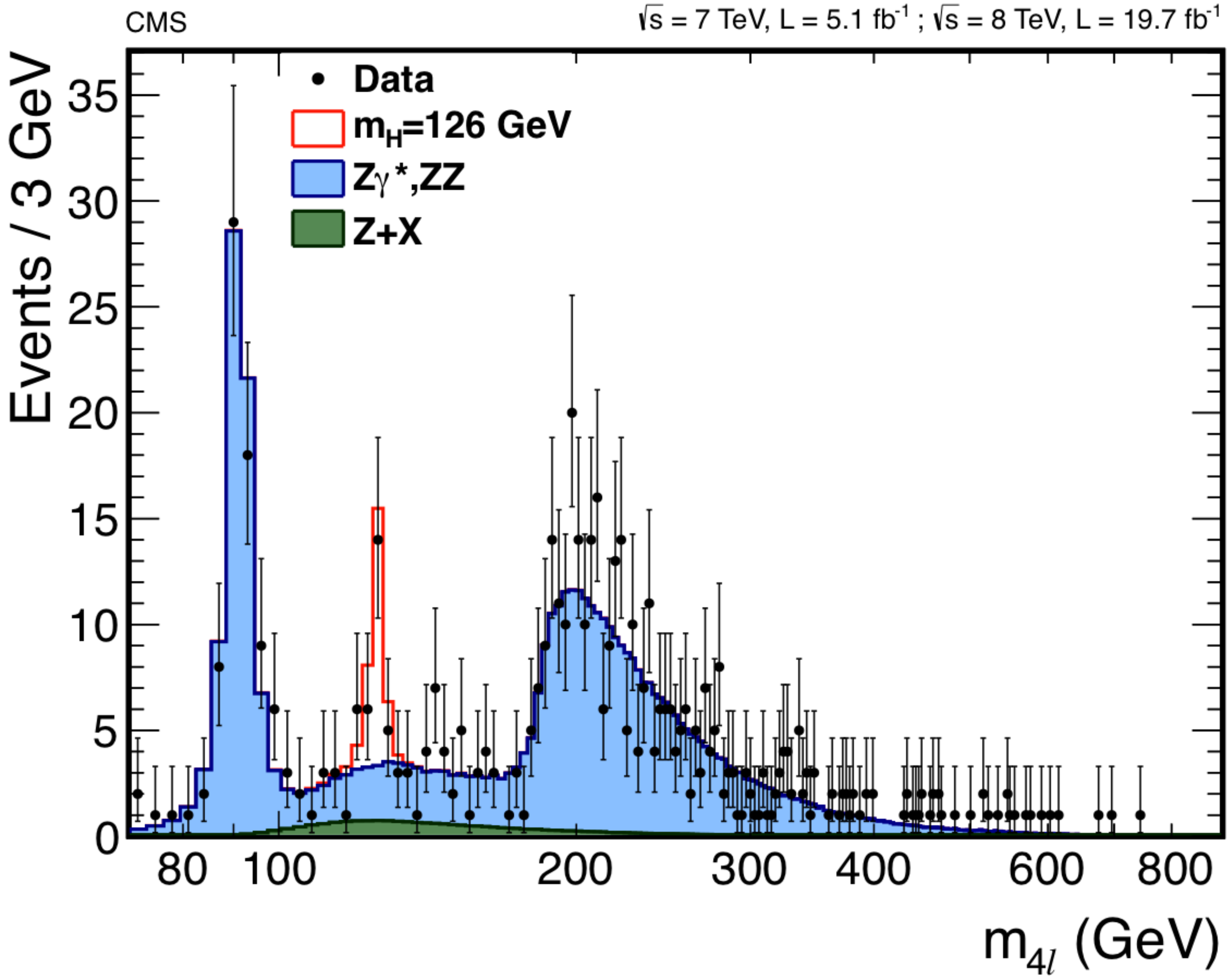


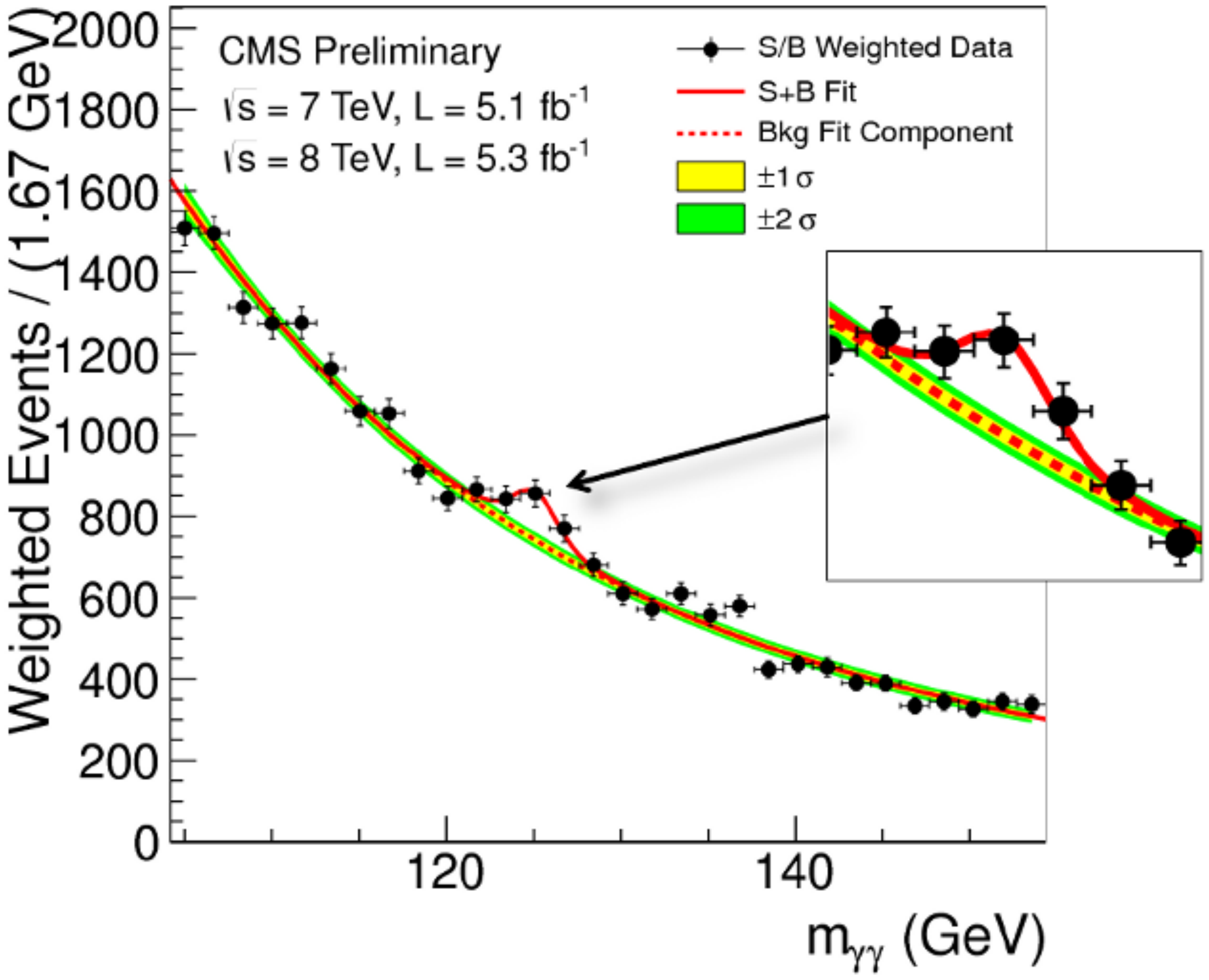




identify and measure muons, photons and electrons with high precision. The energy resolution for the above particles will be better than 1% at 100 GeV. At the core of the CMS detector sits a large superconducting solenoid generating a uniform magnetic field of 4 T. The choice of a strong magnetic field leads to a compact design for the muon spectrometer without compromising the momentum resolution up to rapidities of 2.5. The inner tracking system will measure all high p_t charged tracks with a momentum precision of $\Delta p/p \approx 0.1\ p_t$ (p_t in TeV) in the range $|\eta| < 2.5$. A high resolution crystal electromagnetic calorimeter, designed to detect the two photon decay of an intermediate mass Higgs, is located inside the coil. Hermetic hadronic calorimeters surround the intersection region up to $|\eta| = 4.7$ allowing tagging of forward jets and measurement of missing transverse energy.

And clearly it worked





Searches for something...

- At the LHC, you need a signal hypothesis
  - To design a trigger
  - To optimize your cuts
  - To compute the test statistics
  - To interpret the results
- so far so good...

CMS AN AN-11-065

CMS Draft Analysis Note

The content of this note is intended for CMS internal use and distribution only

2011/11/08
Head Id: 83705
Archive Id: 83789
Archive Date: 2011/11/07
Archive Tag: trunk

Trigger strategies for Higgs searches

The Higgs PAG

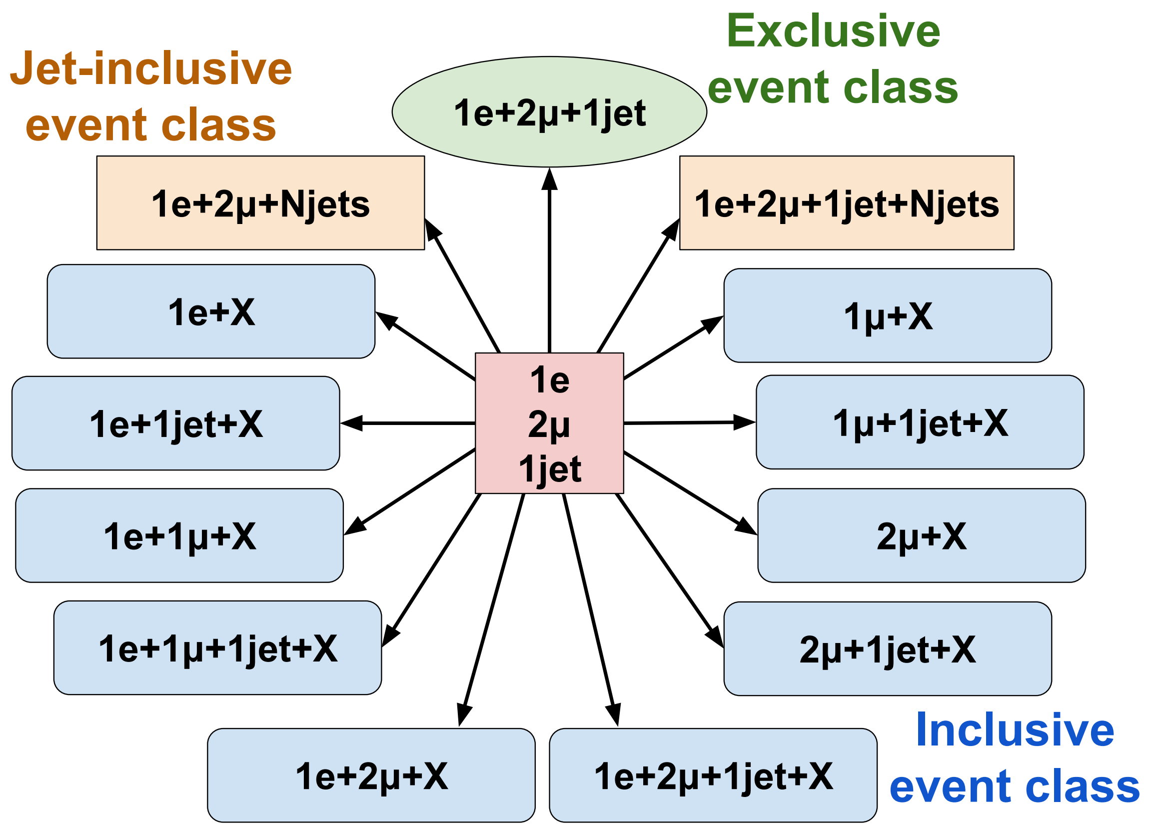
Abstract

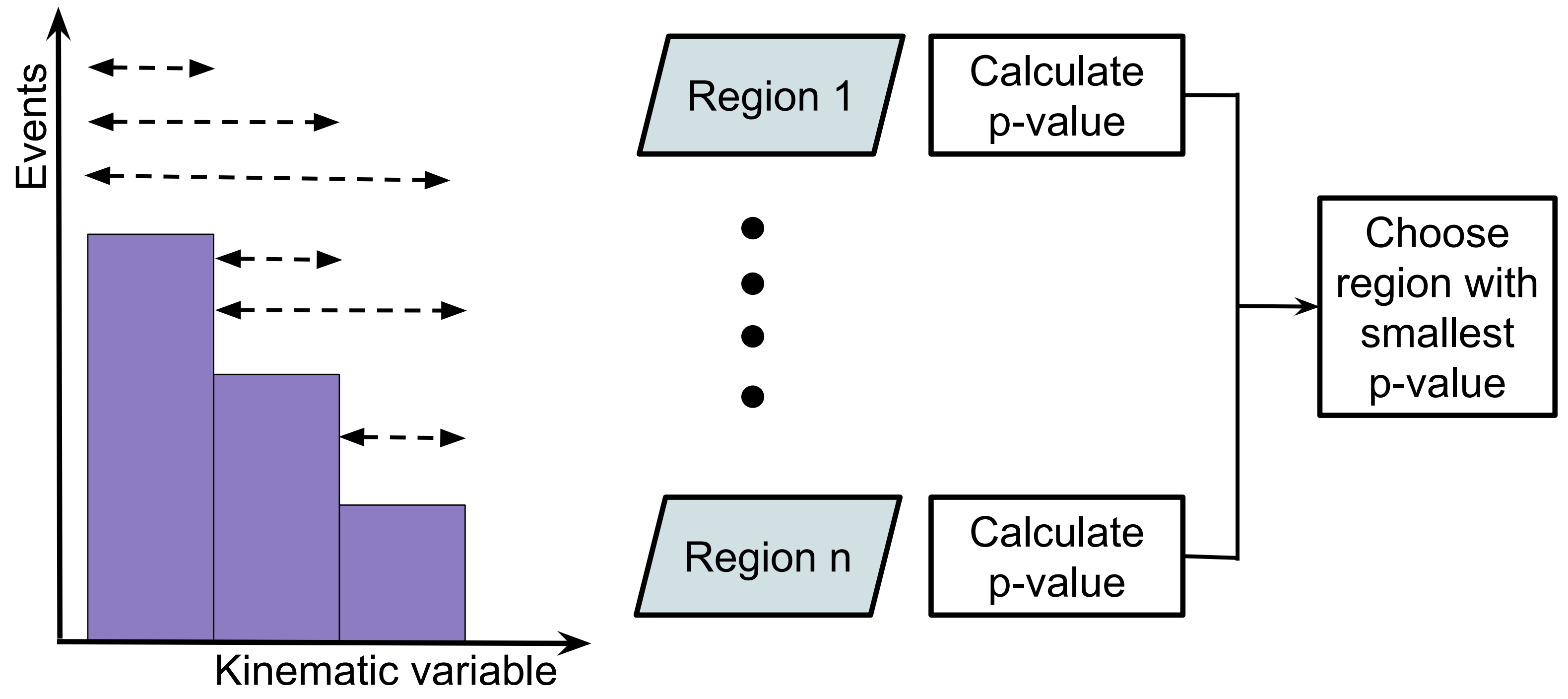
This document describes the triggers used in the Higgs analyses.

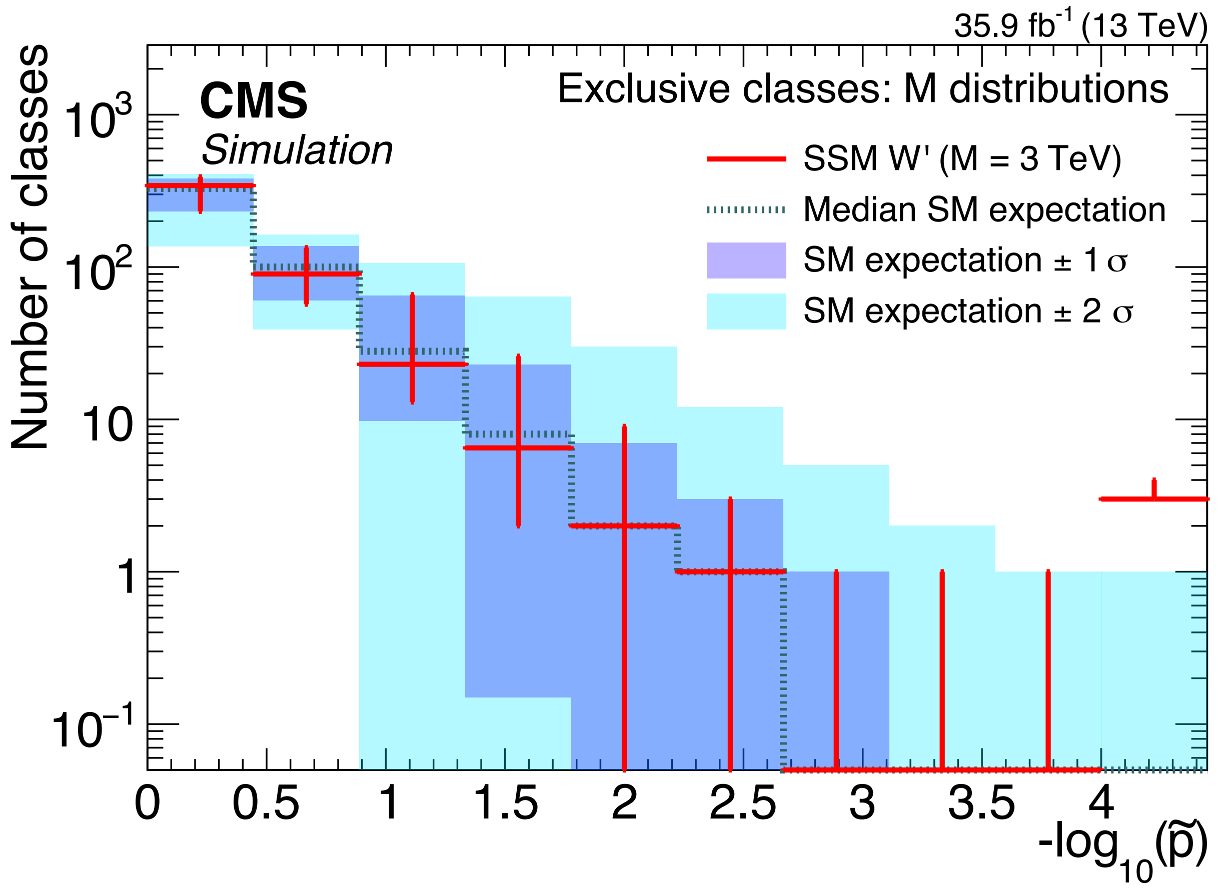
DRAFT

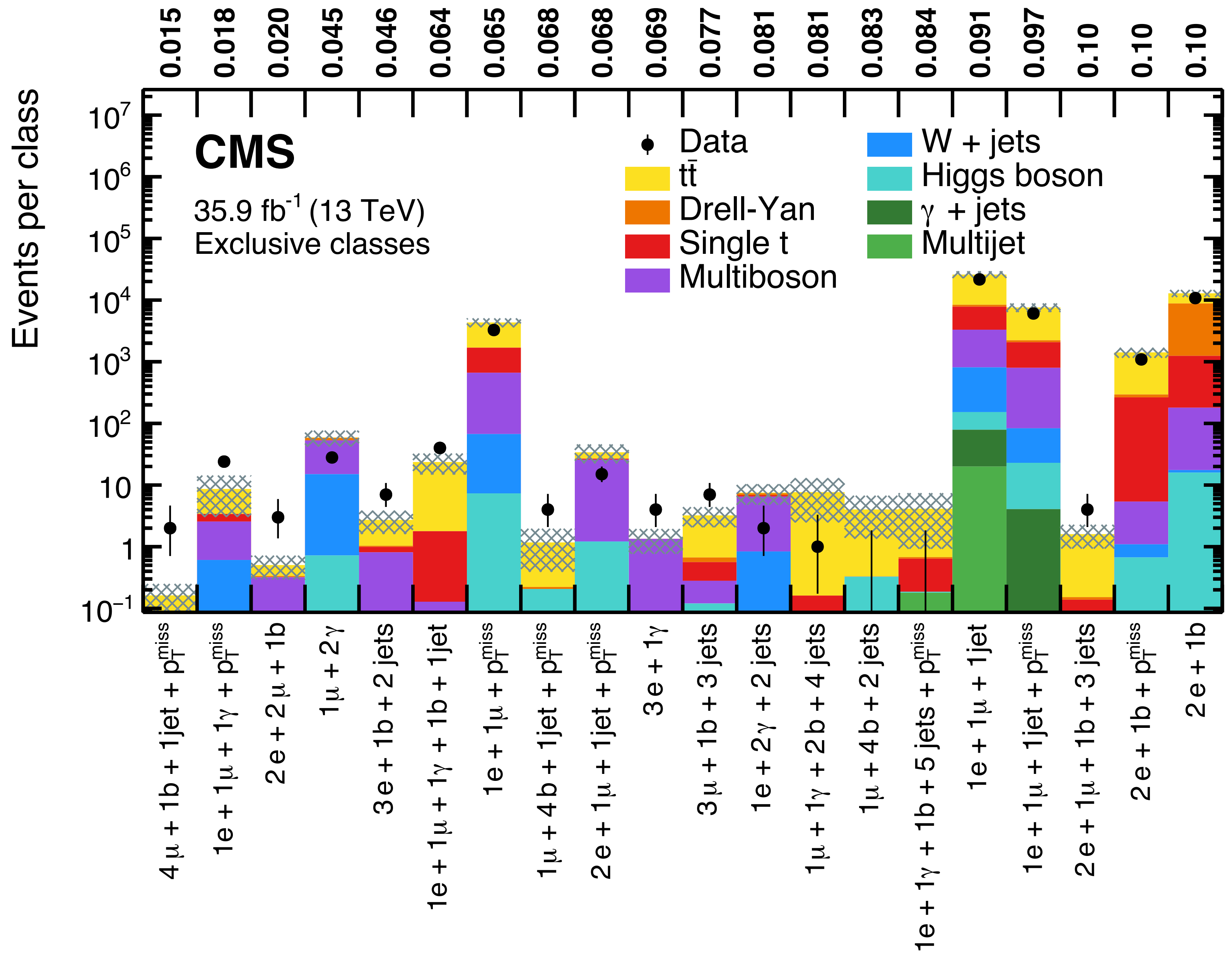
Searches for anything…

- What do you do when you don't know what to search for?
  - Any cut could be a signal killer
  - You need to look at as many signatures as possible
  - You can only look for some deviation from an expected distribution
- ***<u>How do you know that the "right events" are there to start with?</u>***







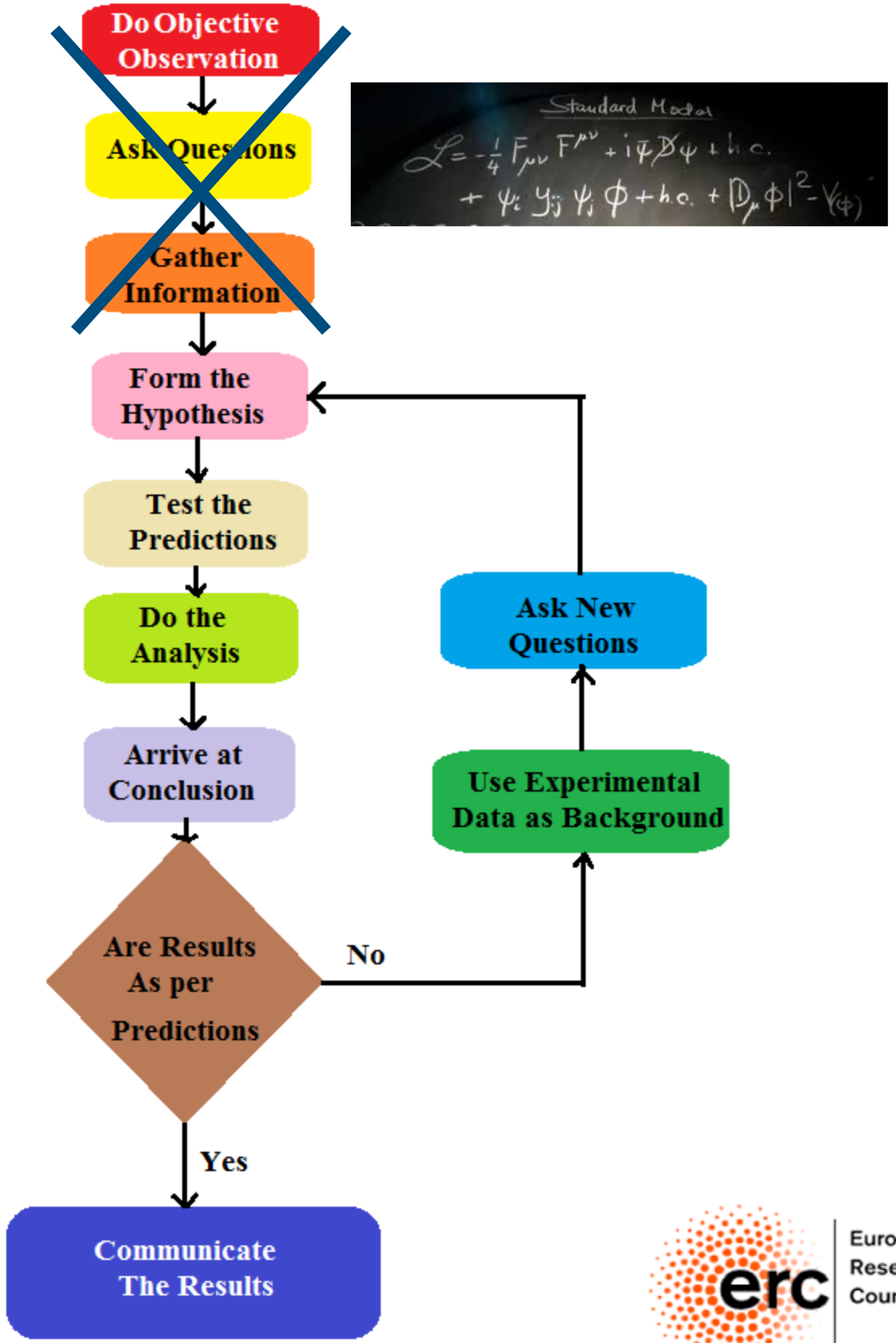


https://arxiv.org/pdf/2010.02984.pdf

New Physics searches & Scientific method

HEP searches in LHC era

- *Research under the scientific method starts gathering information about nature*
- *Instead, our baseline is the SM, which was formed once these informations were gathered*
- *We are victim of our success:*
  - *Since 1970s, we start always from the same point*
  - *We have lost the value of learning from data*
  - *Not by chance, we totally endorsed blind analysis as the ONLY way to search*

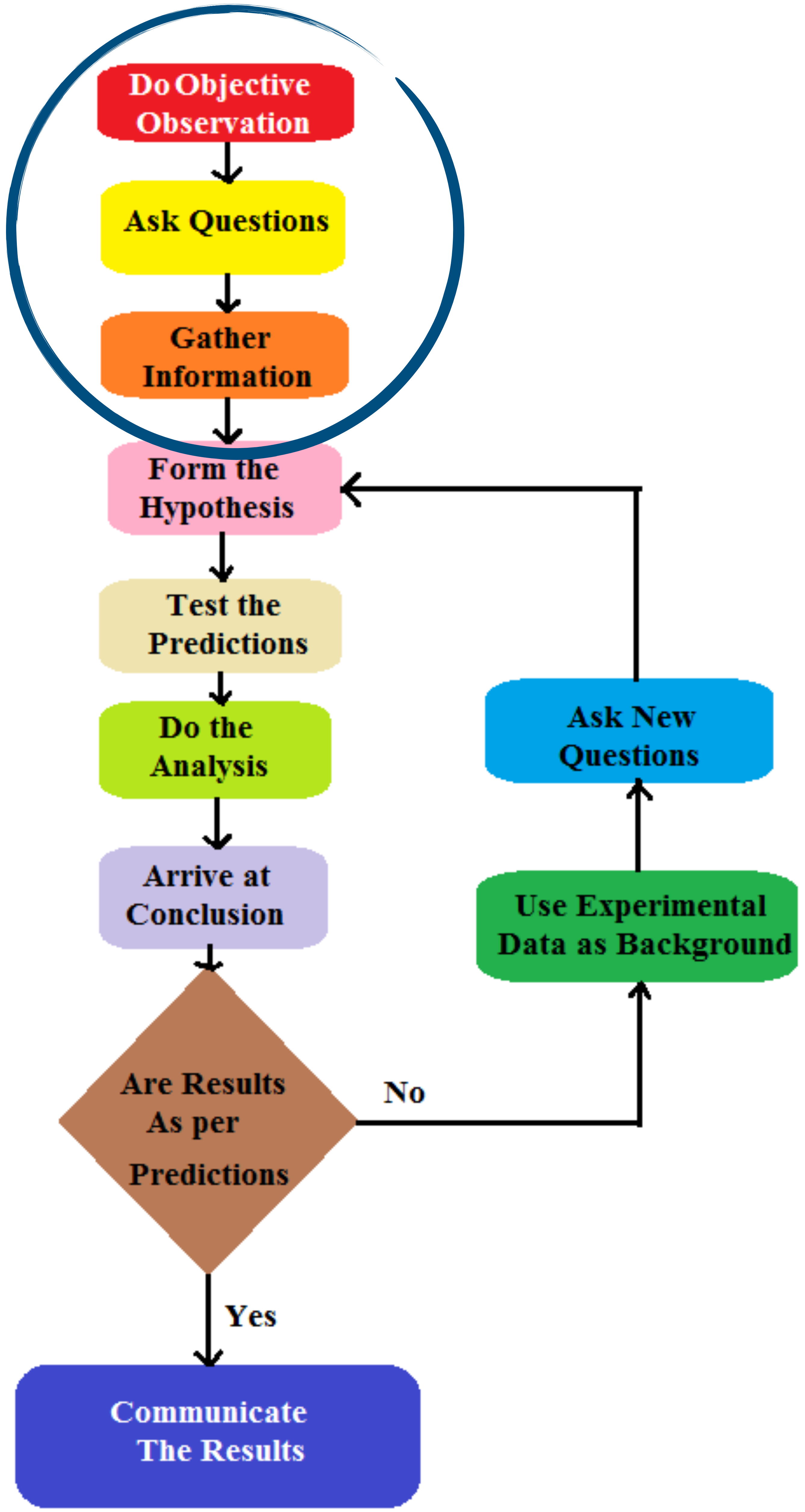


Learning from Data

- Rather than specifying a signal hypothesis upfront, we could start looking at our data
- Based on what we see (e.g., clustering alike objects) we could formulate a signal hypothesis
- EXAMPLE: star classification was based on observed characteristics

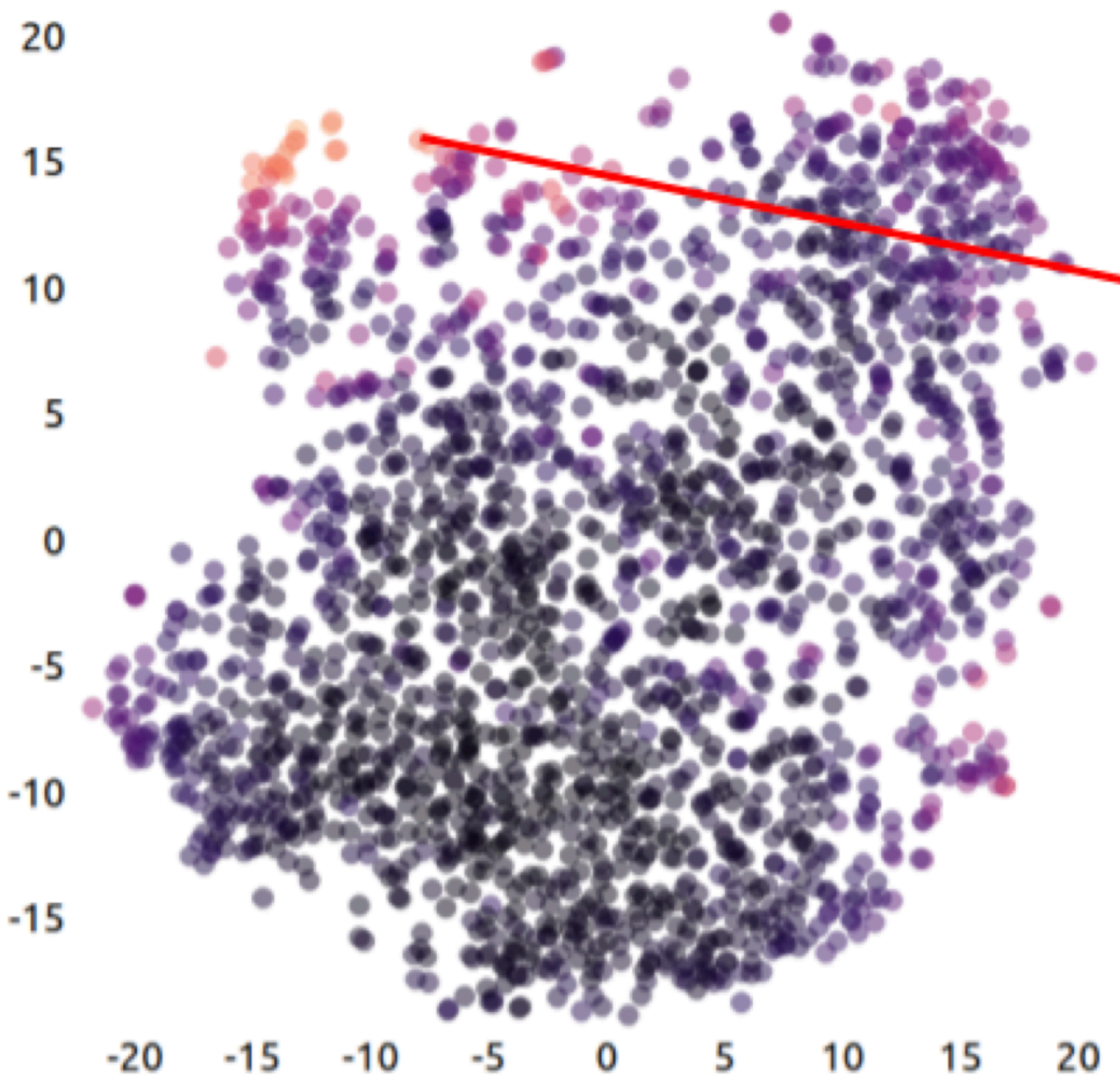
Class	Effective temperature[1][2]	Vega-relative chromaticity[3][4][a]	Chromaticity (D65)[5][6][3][b]	Main-sequence mass[1][7] (solar masses)	Main-sequence radius[1][7] (solar radii)	Main-sequence luminosity[1][7] (bolometric)	Hydrogen lines	Fraction of all main-sequence stars[8]
O	≥ 30,000 K	blue	blue	≥ 16 $M_\odot$	≥ 6.6 $R_\odot$	≥ 30,000 $L_\odot$	Weak	~0.00003%
B	10,000–30,000 K	blue white	deep blue white	2.1–16 $M_\odot$	1.8–6.6 $R_\odot$	25–30,000 $L_\odot$	Medium	0.13%
A	7,500–10,000 K	white	blue white	1.4–2.1 $M_\odot$	1.4–1.8 $R_\odot$	5–25 $L_\odot$	Strong	0.6%
F	6,000–7,500 K	yellow white	white	1.04–1.4 $M_\odot$	1.15–1.4 $R_\odot$	1.5–5 $L_\odot$	Medium	3%
G	5,200–6,000 K	yellow	yellowish white	0.8–1.04 $M_\odot$	0.96–1.15 $R_\odot$	0.6–1.5 $L_\odot$	Weak	7.6%
K	3,700–5,200 K	light orange	pale yellow orange	0.45–0.8 $M_\odot$	0.7–0.96 $R_\odot$	0.08–0.6 $L_\odot$	Very weak	12.1%
M	2,400–3,700 K	orange red	light orange red	0.08–0.45 $M_\odot$	≤ 0.7 $R_\odot$	≤ 0.08 $L_\odot$	Very weak	76.45%

- Afterwords, it was realised that different classes correspond to different temperatures

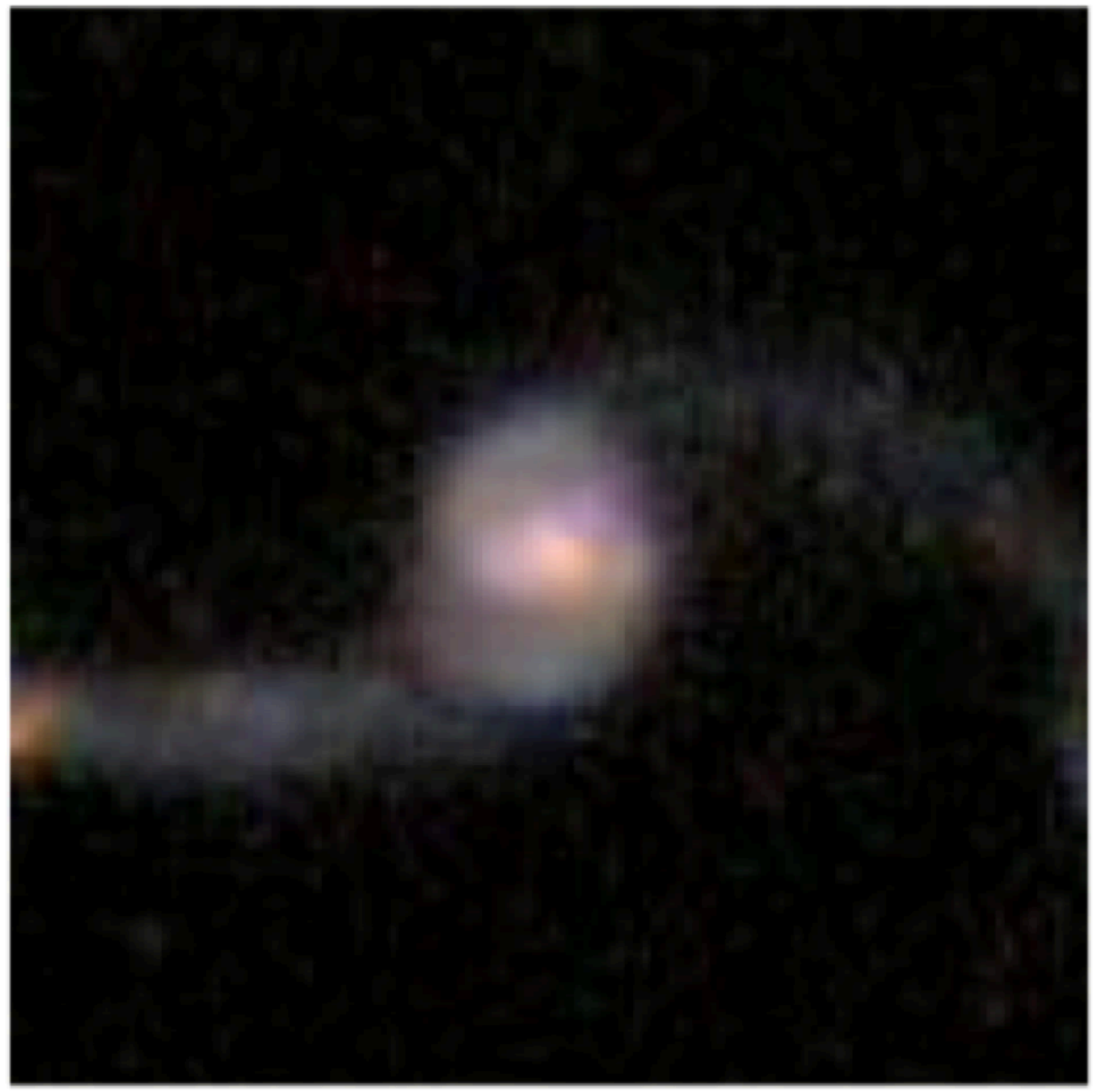


Learning from Anomalies

- Anomaly detection is one kind of data mining technique
  - One defines a metric of “typicality” to rank data samples
  - Based on this ranking, one can identify less typical events, tagging them as anomalies
  - By studying anomalies, one can make hypotheses on new physics mechanisms



Object ID: 960415



Anomaly Score: 4.470837

Back to 1984

- In the 1984 the UA1 experiment reported an excess of events with large missing transverse energy
- Before than, events with this signatures were extensively discussed with theorists (see “” for a first hand account of this)
- The community was looking for explanations (which eventually was provided by a combination of calorimeter cracks and tau decays)

EXPERIMENTAL OBSERVATION OF EVENTS WITH LARGE MISSING TRANSVERSE ENERGY ACCOMPANIED BY A JET OR A PHOTON (S) IN $p\bar{p}$ COLLISIONS AT $\sqrt{s}$ = 540 GeV

UA1 Collaboration, CERN, Geneva, Switzerland

G. ARNISON [m], O.C. ALLKOFER [g], A. ASTBURY [m,1], B. AUBERT [b], C. BACCI [ℓ], G. BAUER [p], A. BÉZAGUET [d], R.K. BOCK [d], T.J.V. BOWCOCK [h], M. CALVETTI [d], P. CATZ [b], P. CENNINI [d], S. CENTRO [2], F. CERADINI [ℓ], S. CITTOLIN [d], D. CLINE [p], C. COCHET [n], J. COLAS [b], M. CORDEN [c], D. DALLMAN [d,o], D. DAU [d,g], M. DeBEER [n], M. DELLA NEGRA [b,d], M. DEMOULIN [d], D. DENEGRI [n], D. DiBITONTO [d], A. DiCIACCIO [ℓ], L. DOBRZYNSKI [j], J. DOWELL [c], K. EGGERT [a], E. EISENHANDLER [h], N. ELLIS [d], P. ERHARD [a], H. FAISSNER [a], M. FINCKE [g,1], P. FLYNN [m], G. FONTAINE [j], R. FREY [k], R. FRÜHWIRTH [o], J. GARVEY [c], S. GEER [e], C. GHESQUIÈRE [j], P. GHEZ [b], W.R. GIBSON [h], Y. GIRAUD-HÉRAUD [j], A. GIVERNAUD [n], A. GONIDEC [b], G. GRAYER [m], T. HANSL-KOZANECKA [a], W.J. HAYNES [m], L.O. HERTZBERGER [i], D. HOFFMANN [a], H. HOFFMANN [d], D.J. HOLTHUIZEN [i], R.J. HOMER [c], A. HONMA [h], W. JANK [d], G. JORAT [d], P.I.P. KALMUS [h], V. KARIMÄKI [f], R. KEELER [h,1], I. KENYON [c], A. KERNAN [k], R. KINNUNEN [f], W. KOZANECKI [k], D. KRYN [d,j], P. KYBERD [h], F. LACAVA [ℓ], J.-P. LAUGIER [n], J.-P. LEES [b], H. LEHMANN [a], R. LEUCHS [g], A. LÉVÊQUE [d], D. LINGLIN [b], E. LOCCI [n], M. LORET [n], T. MARKIEWICZ [p], G. MAURIN [d], T. McMAHON [c], J.-P. MENDIBURU [j], M.-N. MINARD [b], M. MOHAMMADI [p], M. MORICCA [ℓ], K. MORGAN [k], F. MULLER [d], A.K. NANDI [m], L. NAUMANN [d], A. NORTON [d], A. ORKIN-LECOURTOIS [j], L. PAOLUZI [ℓ], F. PAUSS [d], G. PIANO MORTARI [ℓ], E. PIETARINEN [f], M. PIMIÄ [f], D. PITMAN [k], A. PLACCI [d], J.-P. PORTE [d], E. RADERMACHER [a], J. RANSDELL [k], H. REITHLER [a], J.-P. REVOL [d], J. RICH [n], M. RIJSSENBEEK [d], C. ROBERTS [m], J. ROHLF [e], P. ROSSI [d], C. RUBBIA [d], B. SADOULET [d], G. SAJOT [j], G. SALVINI [ℓ], J. SASS [n], A. SAVOY-NAVARRO [n], D. SCHINZEL [d], W. SCOTT [m], T.P. SHAH [m], I. SHEER [k], D. SMITH [k], J. STRAUSS [o], J. STREETS [c], K. SUMOROK [d], F. SZONCSO [o], C. TAO [j], G. THOMPSON [h], J. TIMMER [d], E. TSCHESLOG [a], J. TUOMINIEMI [f], B. Van EIJK [i], J.-P. VIALLE [b], J. VRANA [j], V. VUILLEMIN [d], H.D. WAHL [o], P. WATKINS [c], J. WILSON [c], C.-E. WULZ [o] and M. YVERT [b]

Aachen [a]–Annecy (LAPP) [b]–Birmingham [c]–CERN [d]–Harvard [e]–Helsinki [f]–Kiel [g]–Queen Mary College, London [h]–NIKHEF, Amsterdam [i]–Paris (Coll. de France) [j]–Riverside [k]–Roma [ℓ]–Rutherford Appleton Lab. [m]–Saclay (CEN) [n] Vienna [o]–Wisconsin [p] Collaboration

Received 30 March 1984

Back to 1984

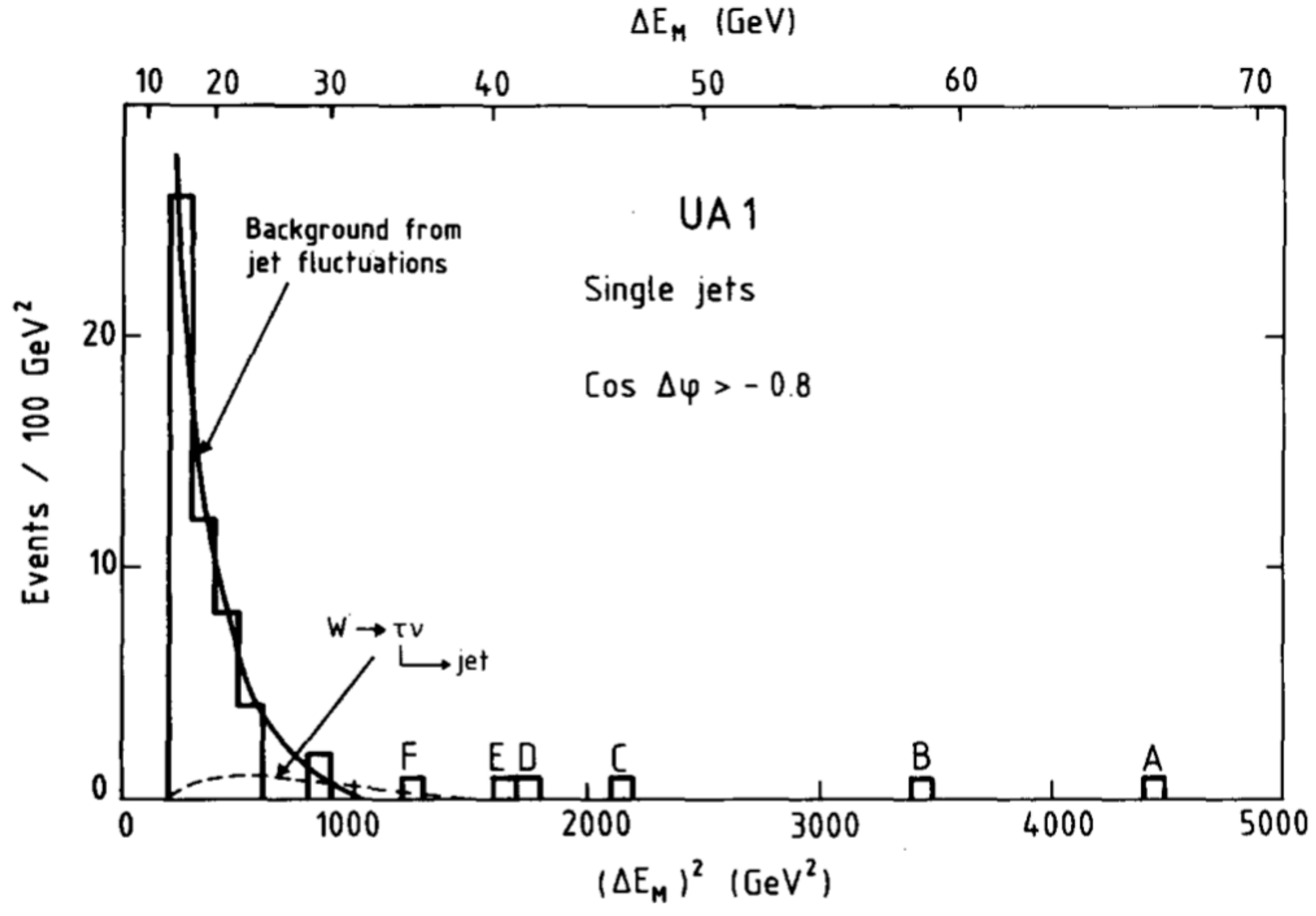
- *In the article, one sees the seeds of modern large-scale data analysis techniques*

- *But the paper is more about single events, event displays, etc. and not just significance, limits, p-value and interpretation*

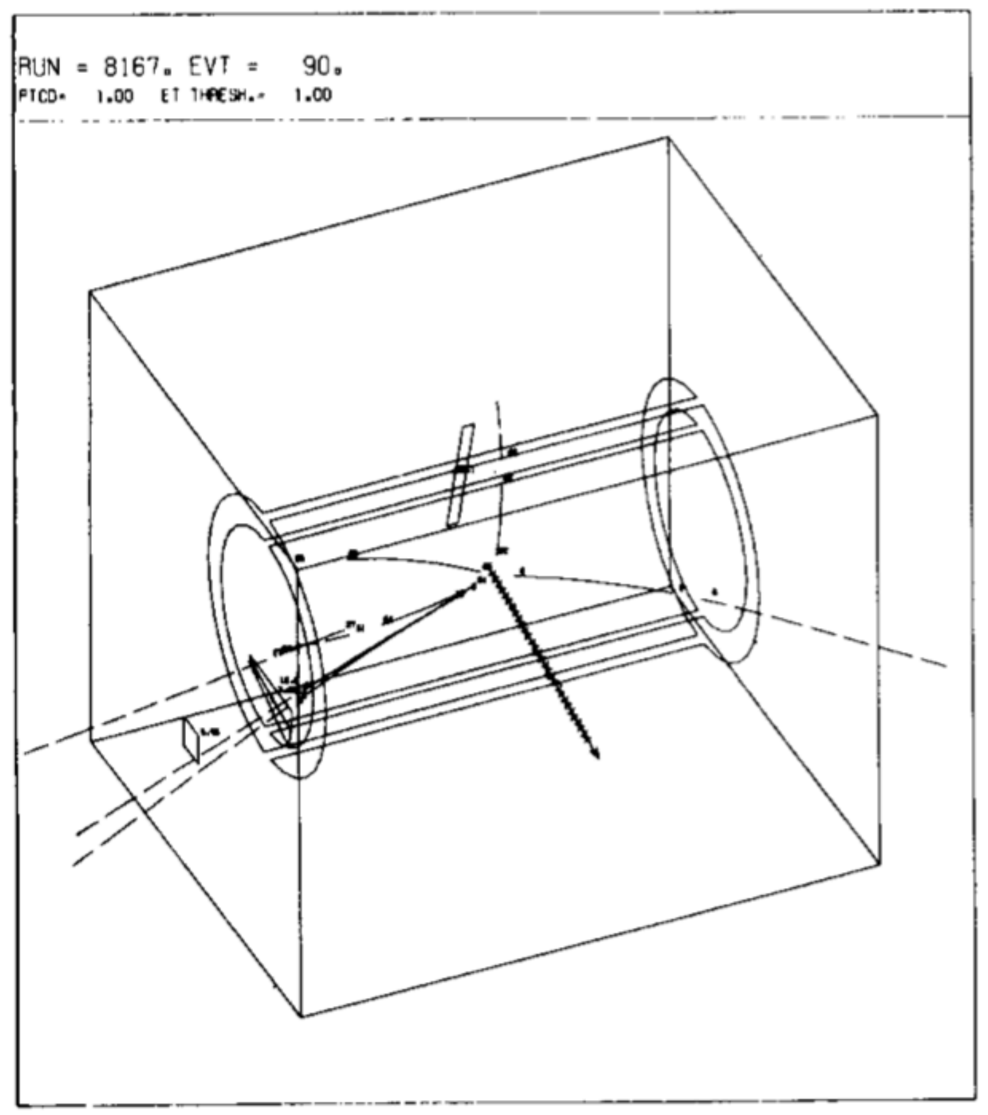
- *Data, and not their statistical interpretation, was central*

- *Certainly, we moved away from that for good reason (blind analysis, etc.)*

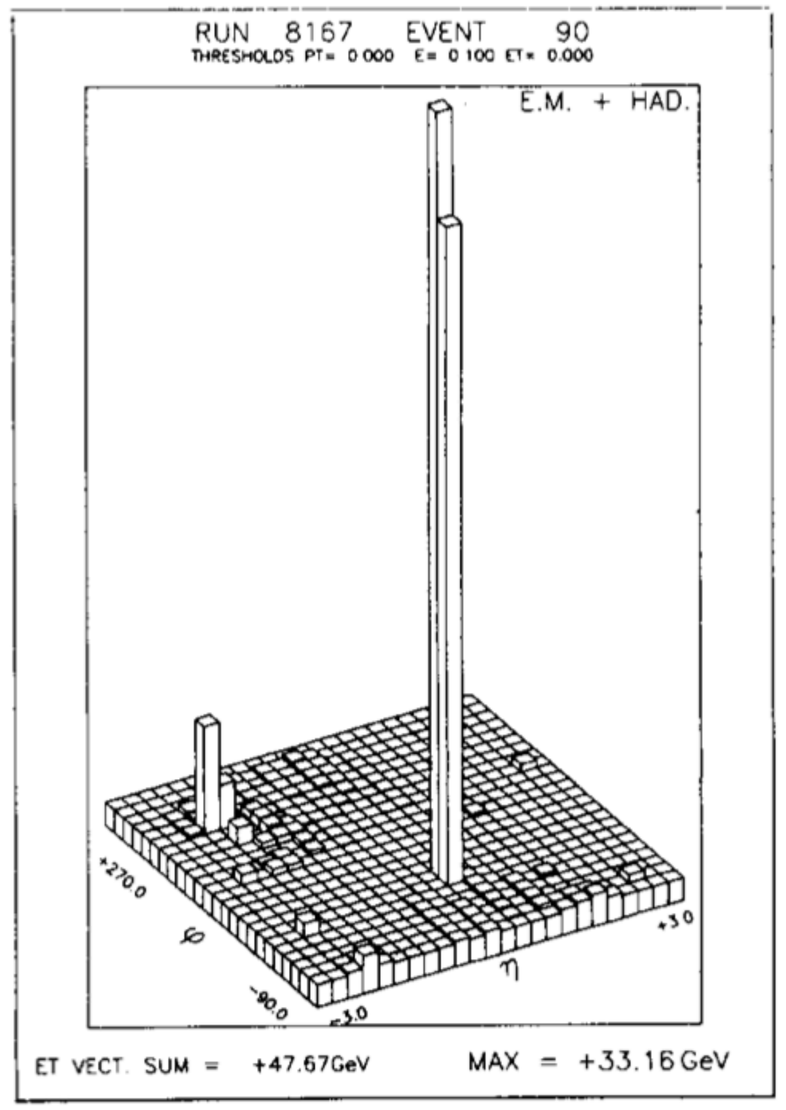
- *On the other hand, aren't we missing something?*



a



Event H



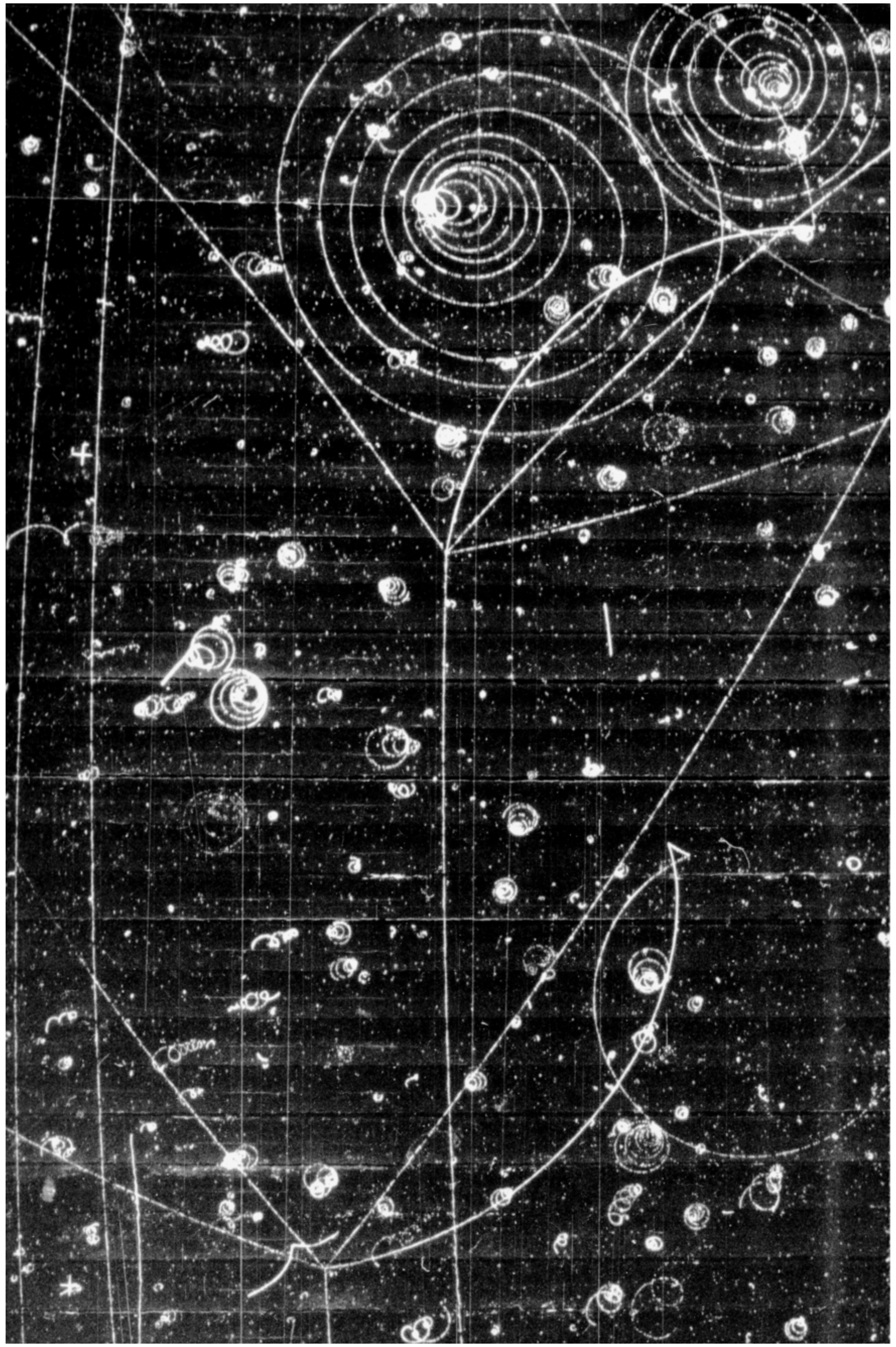
Looking at data used to be OK

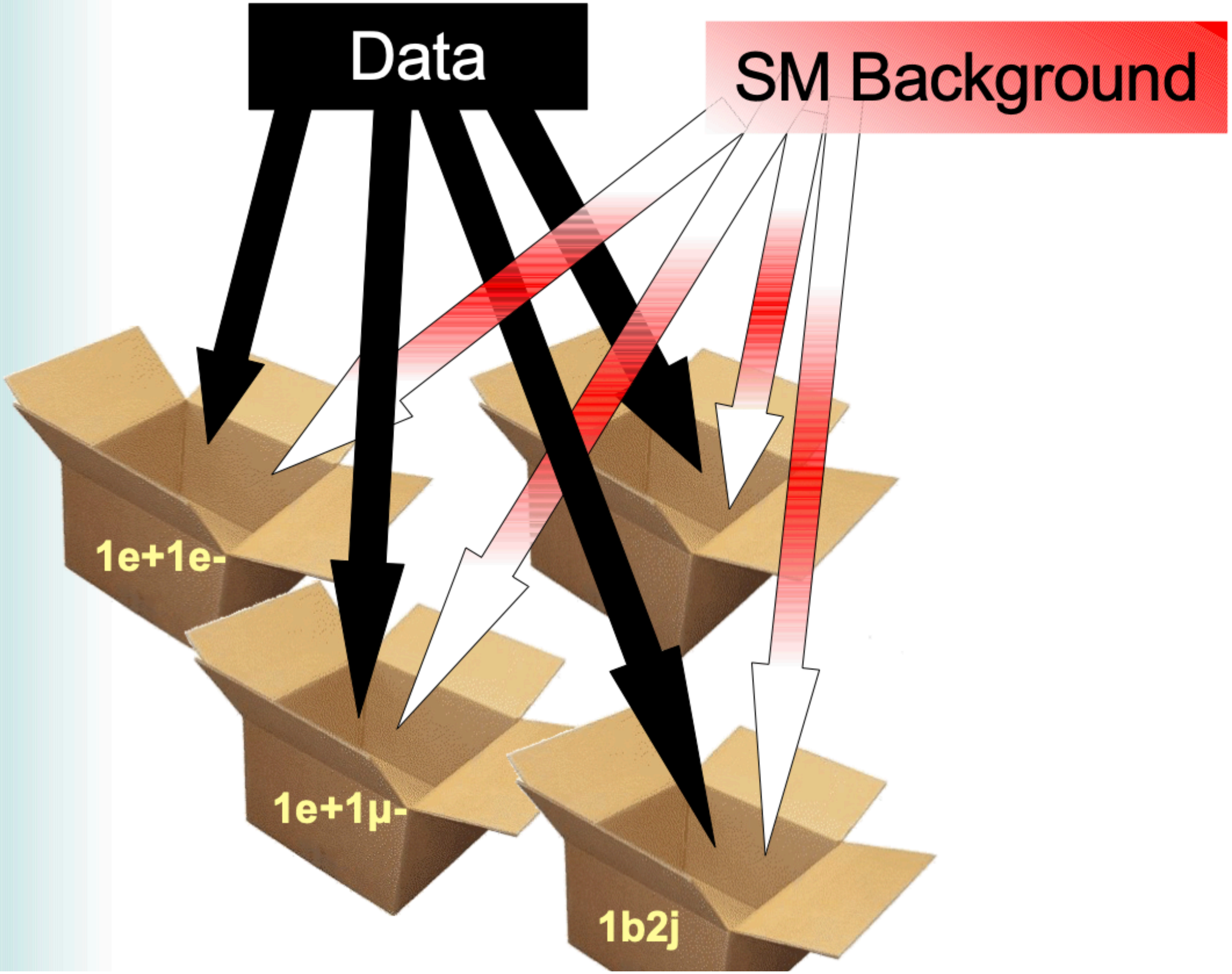
- *Our community looked at data for decades. It was the standard before the new standard (large-scale blind statistical analyses) became a thing*

- *I am not saying we should go back (Discoveries have to be based on reasonable statistical procedures)*

- *I am saying that we should have a pre-analysis step in which we look at data to identify reasonable signatures.*

- *Model independent searches are a way to do this. But there are other ways, in which data are made more central*

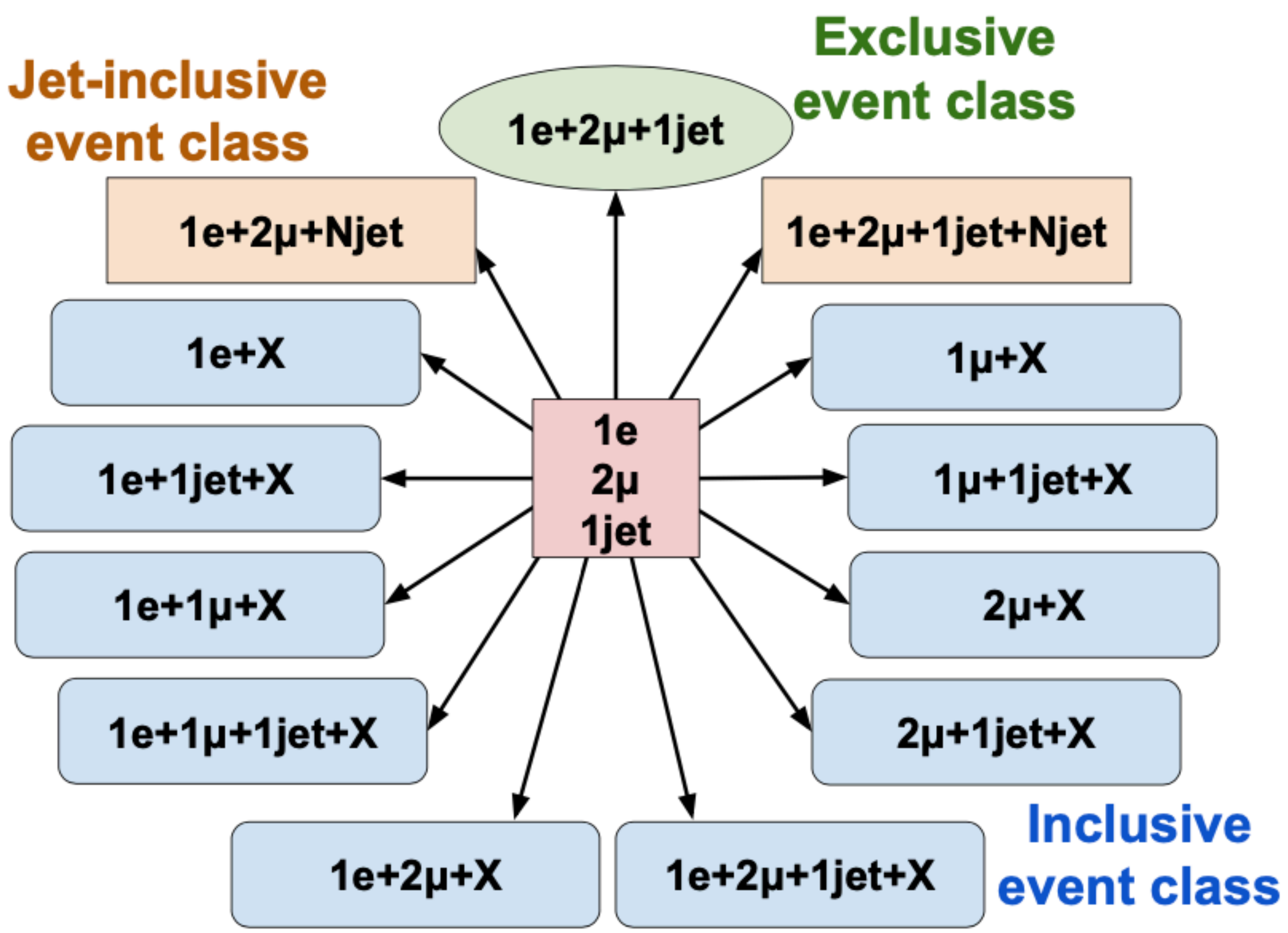




Going model agnostic

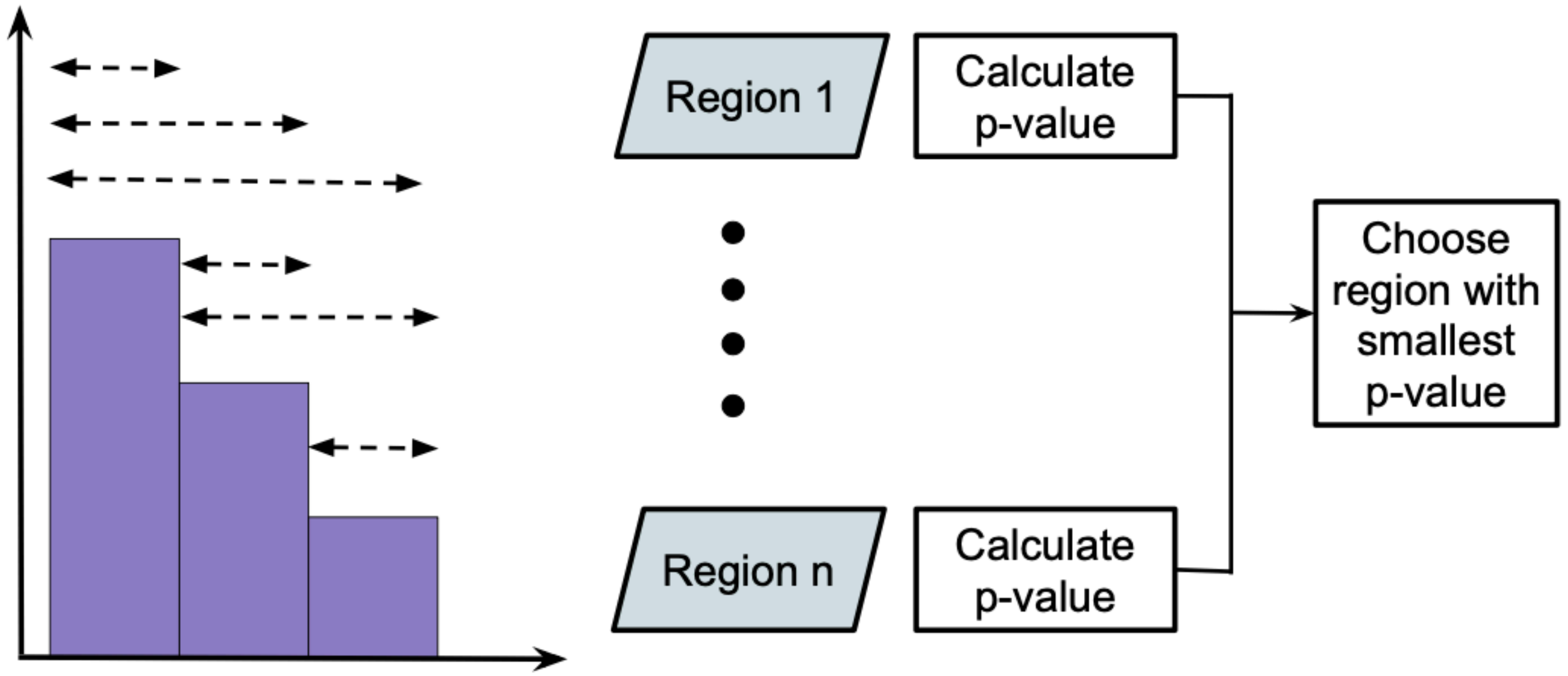
Model independence

- *Since Tevatron/Hera, people tried to go beyond a supervised setup*
  - *No signal specified upfront*
  - *multiple signatures considered at once*
  - *multiple quantities considered at once*

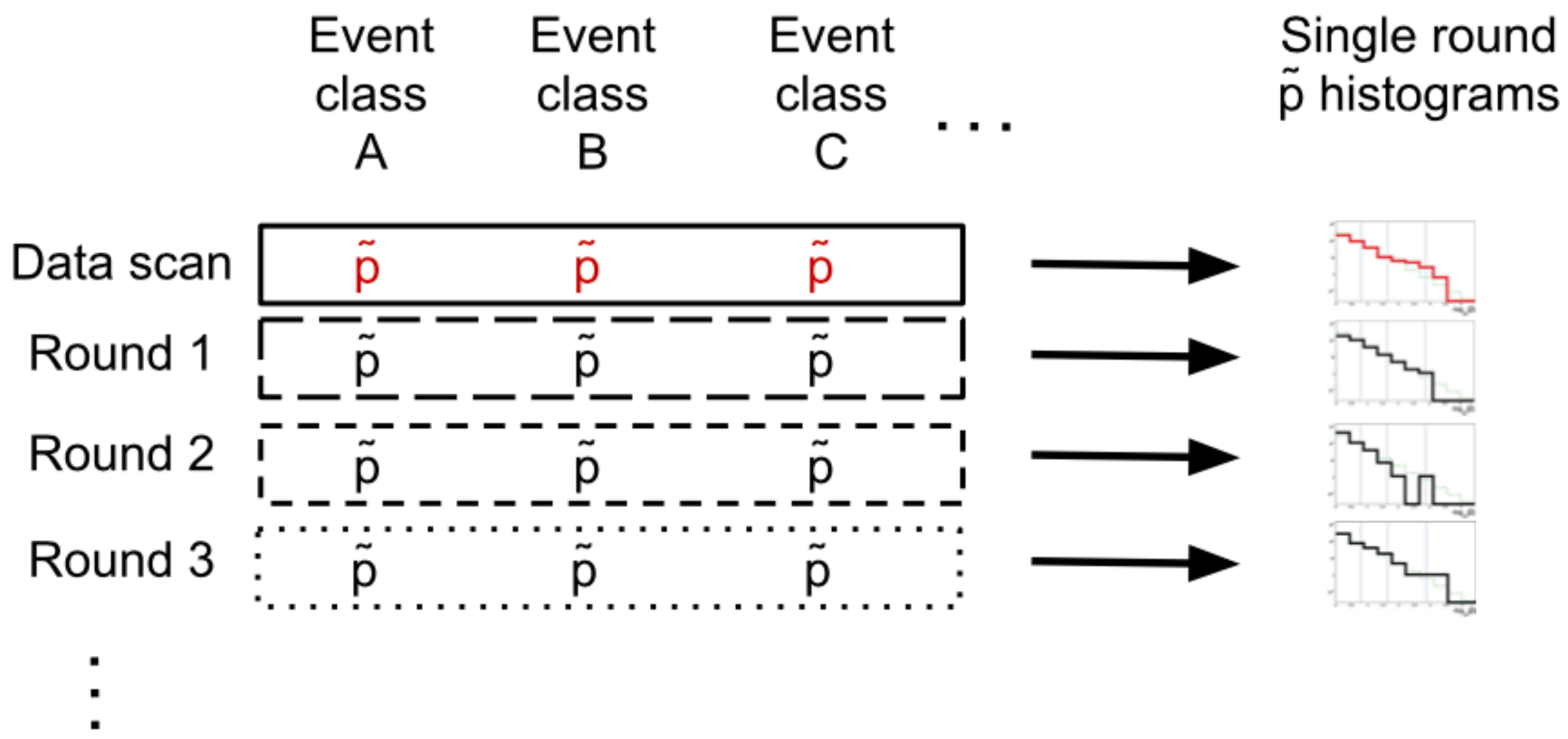


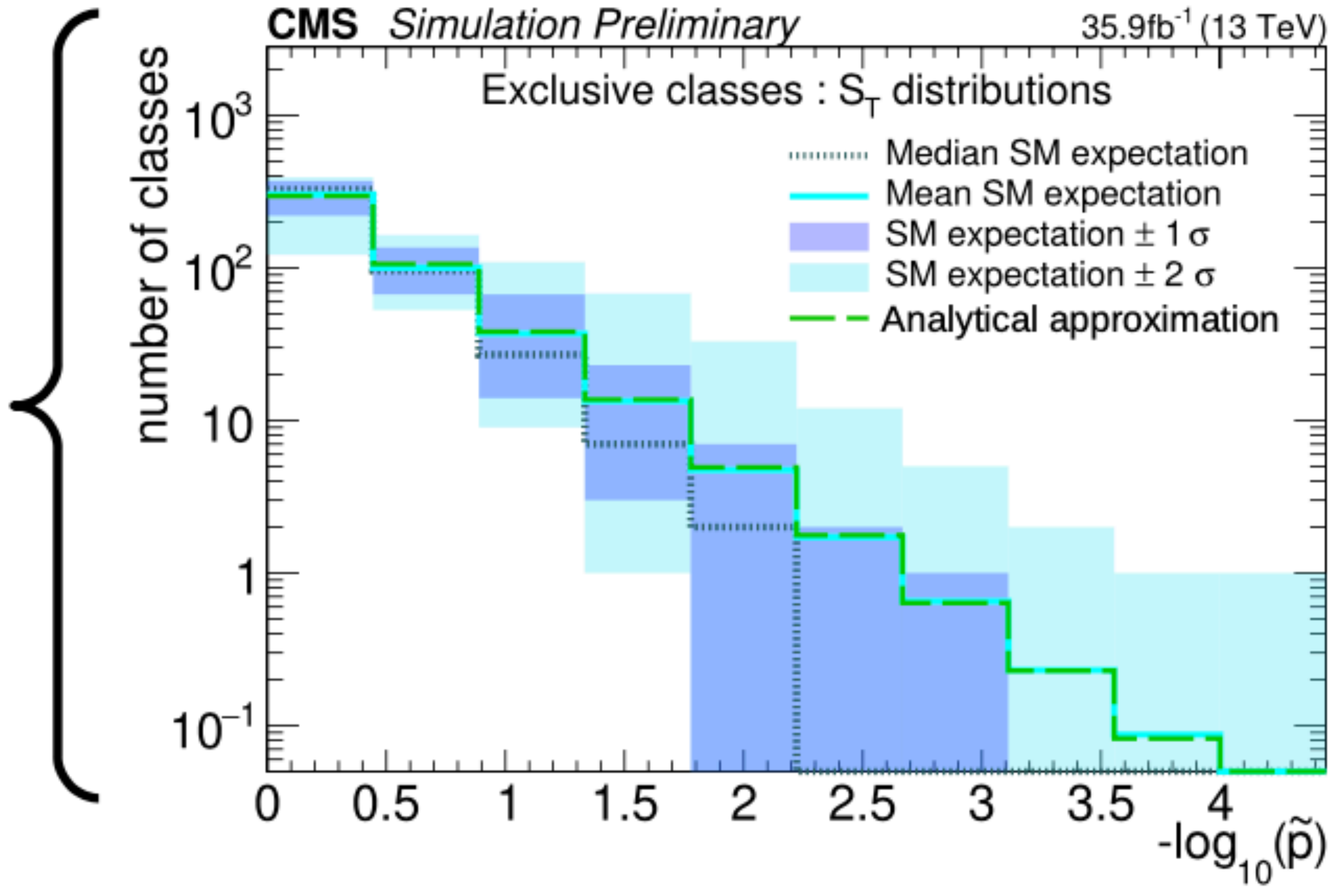
The pipeline

- *Run a goodness of fit test across these many histograms and focus on the smallest p-values to highlight possible anomalies*



- *Build a p-value distribution and look for an excess of low-value bins*



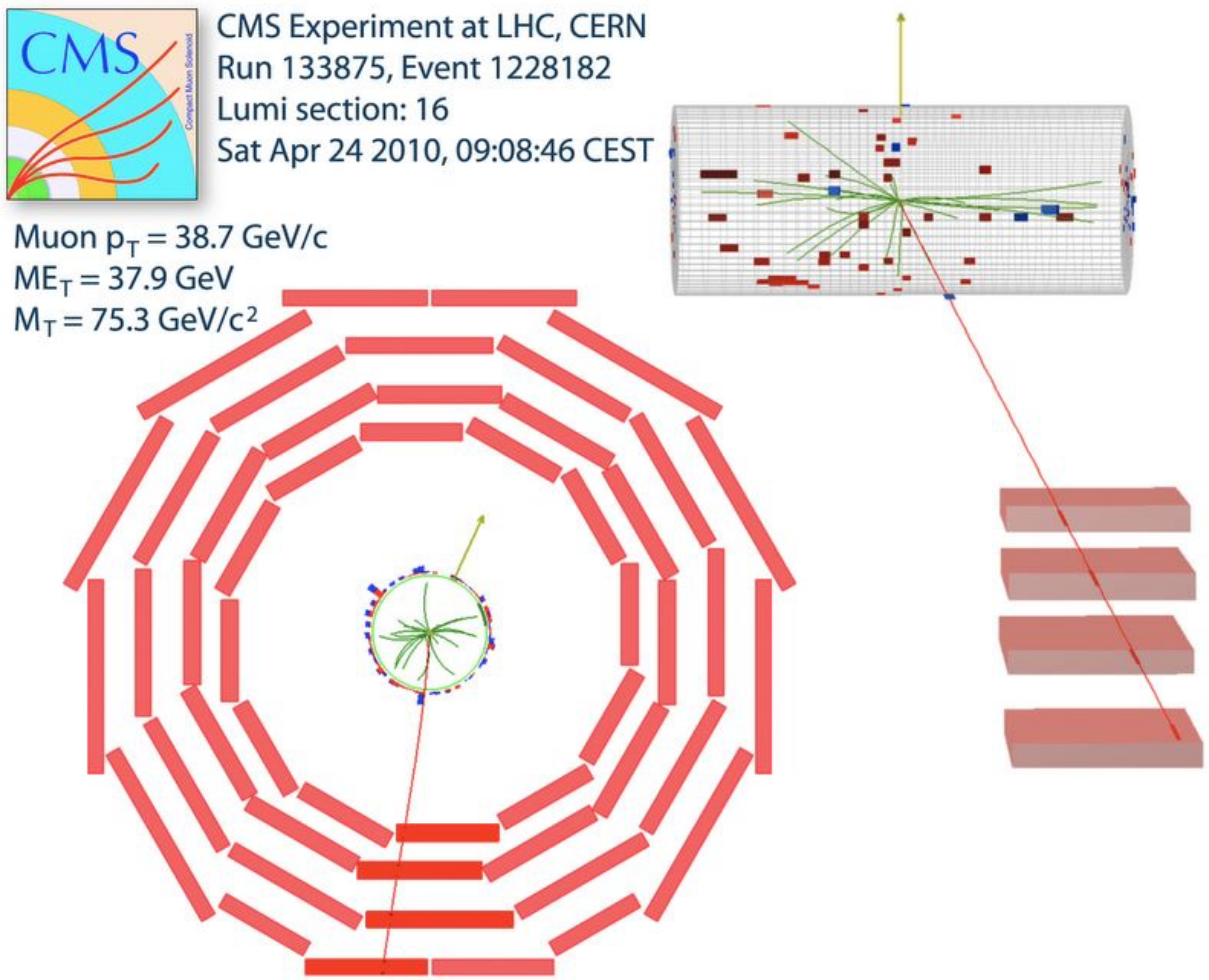


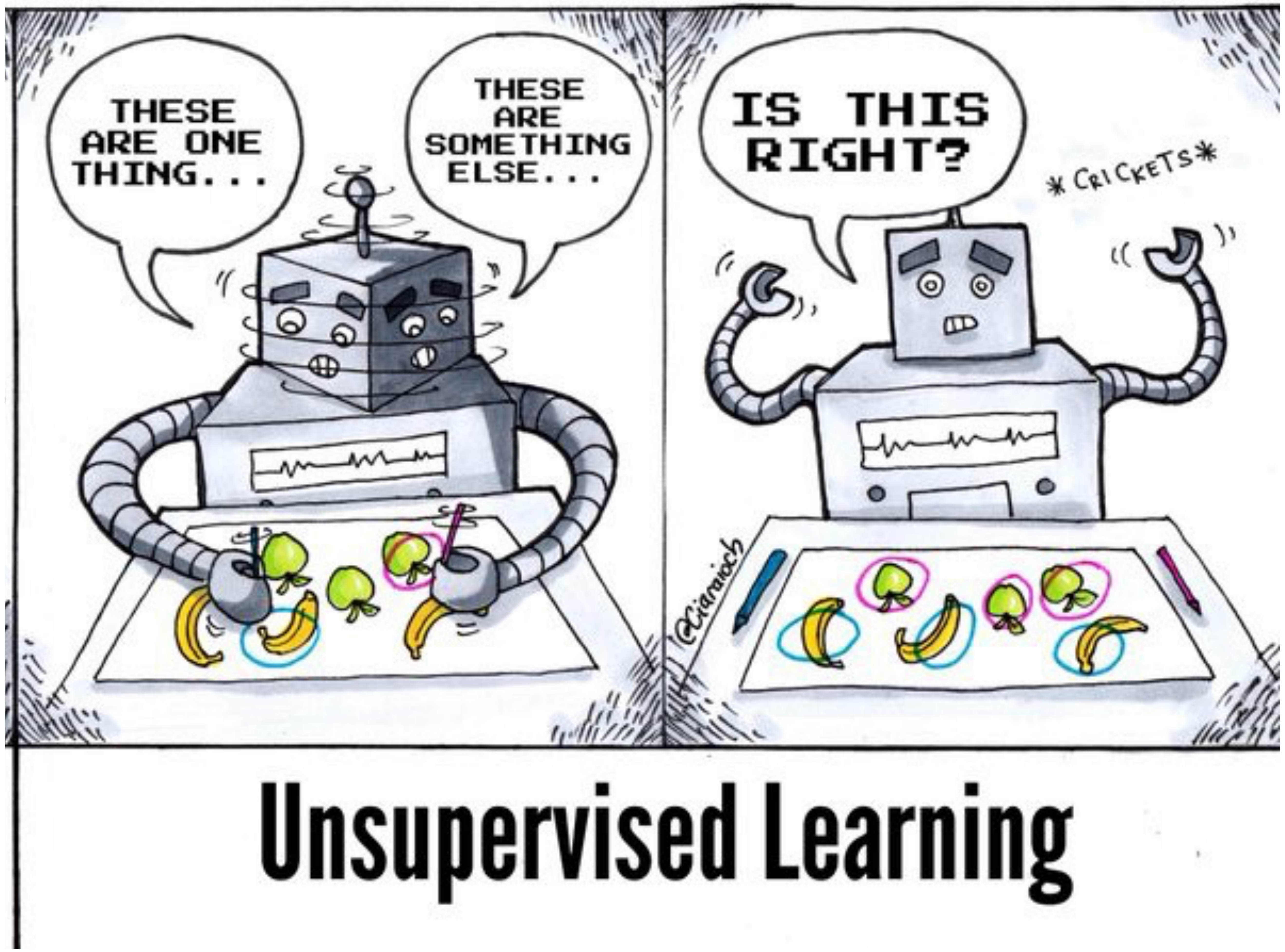
At work with real data

- *In practice, this has approach had limitations*
  - *Statistical fluctuations happen: low p-value bins will be found even in absence of a signal*
  - *Data/MC agreement: the whole strategy relies on MC simulation in low-statistics phase space. One might have issues with PDF, missing NLO contributions, etc.*
  - *Detector simulation: MC simulation might miss detector issues that would manifest as a large p-value. Certainly anomalies, but not of the kind one is targeting*
- *It certainly had its big value: helped finding issues with data, reconstruction software, etc. Particularly useful on first runs*

Physics-motivated anomaly detection

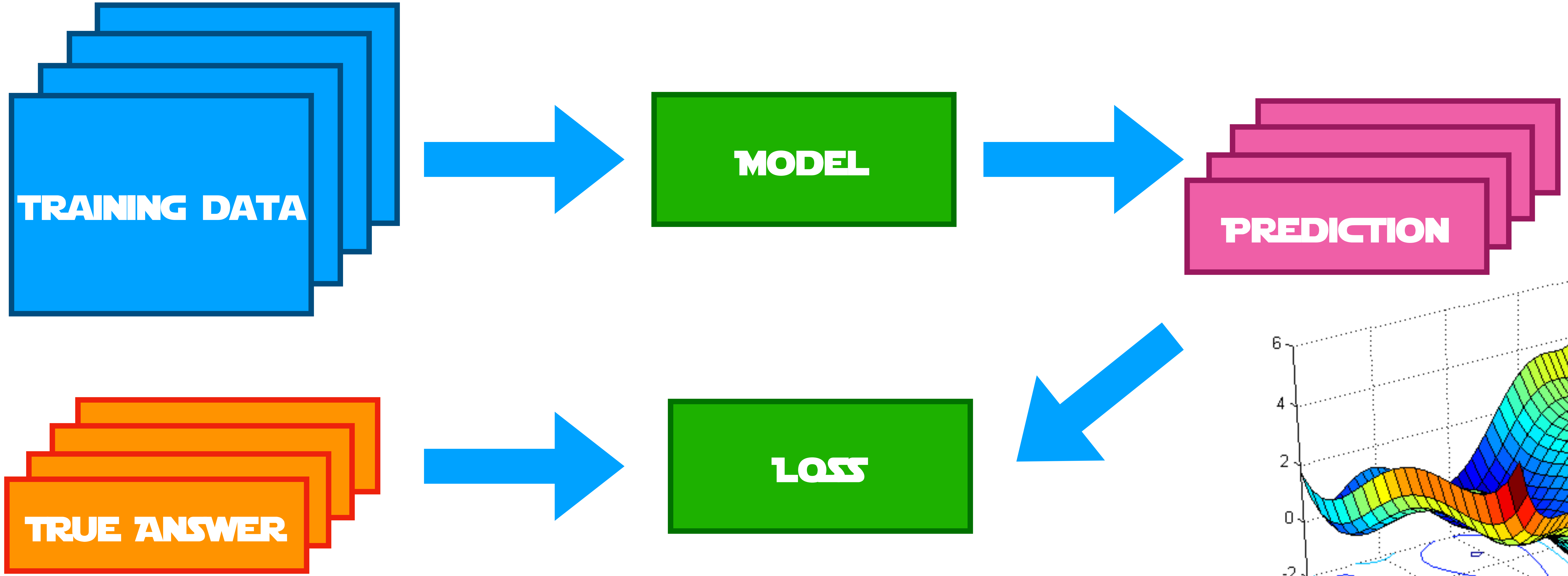
- At the beginning of the LHC, anomaly detection systems were put in place to identify possibly recurrence of low-probability events
  - Very high-pT objects
  - Large multiplicity of hard-to produce particles (leptons)
  - ..
- Even in this case
  - fluctuations happen
  - detector might malfunction
- It was great to find anomalies, but not of the kind one was looking for

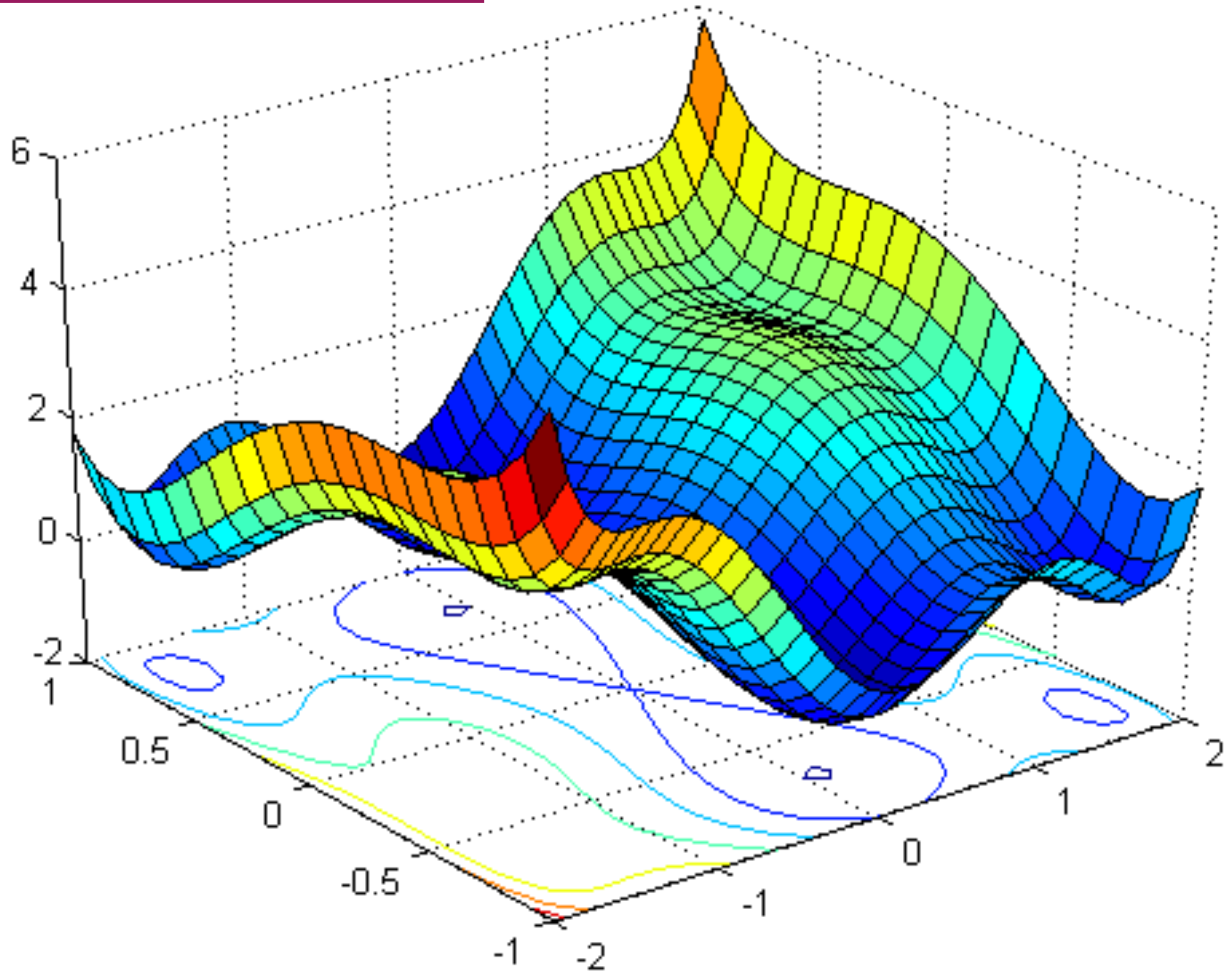




Unsupervised Learning

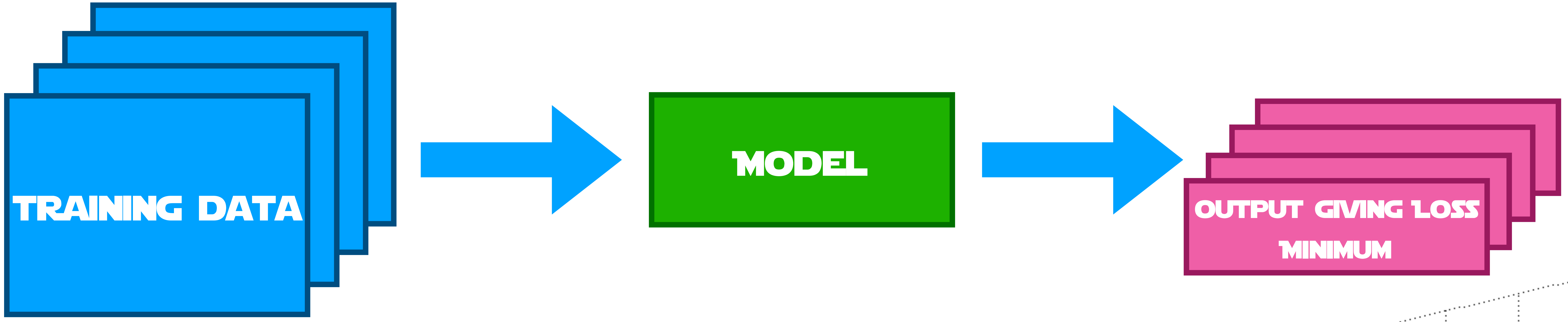
Supervised Learning

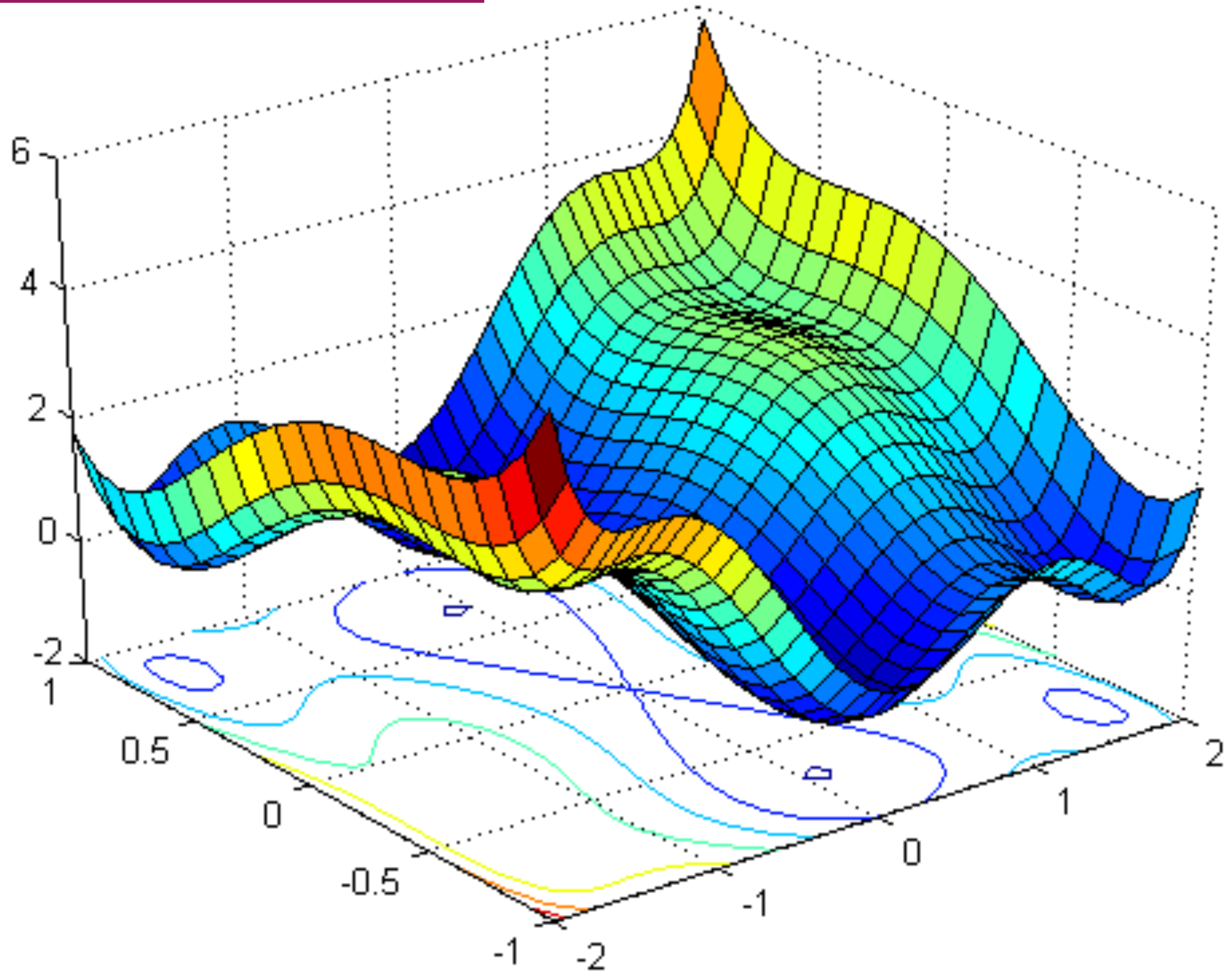




- *A training dataset* ***x***
- *A target y*
- *A model to go from* ***x*** *to y*
- *A loss function quantifying how wrong the model is*
- *A minimisation algorithm to find the model h that corresponds to the minimal loss*

Unsupervised learning

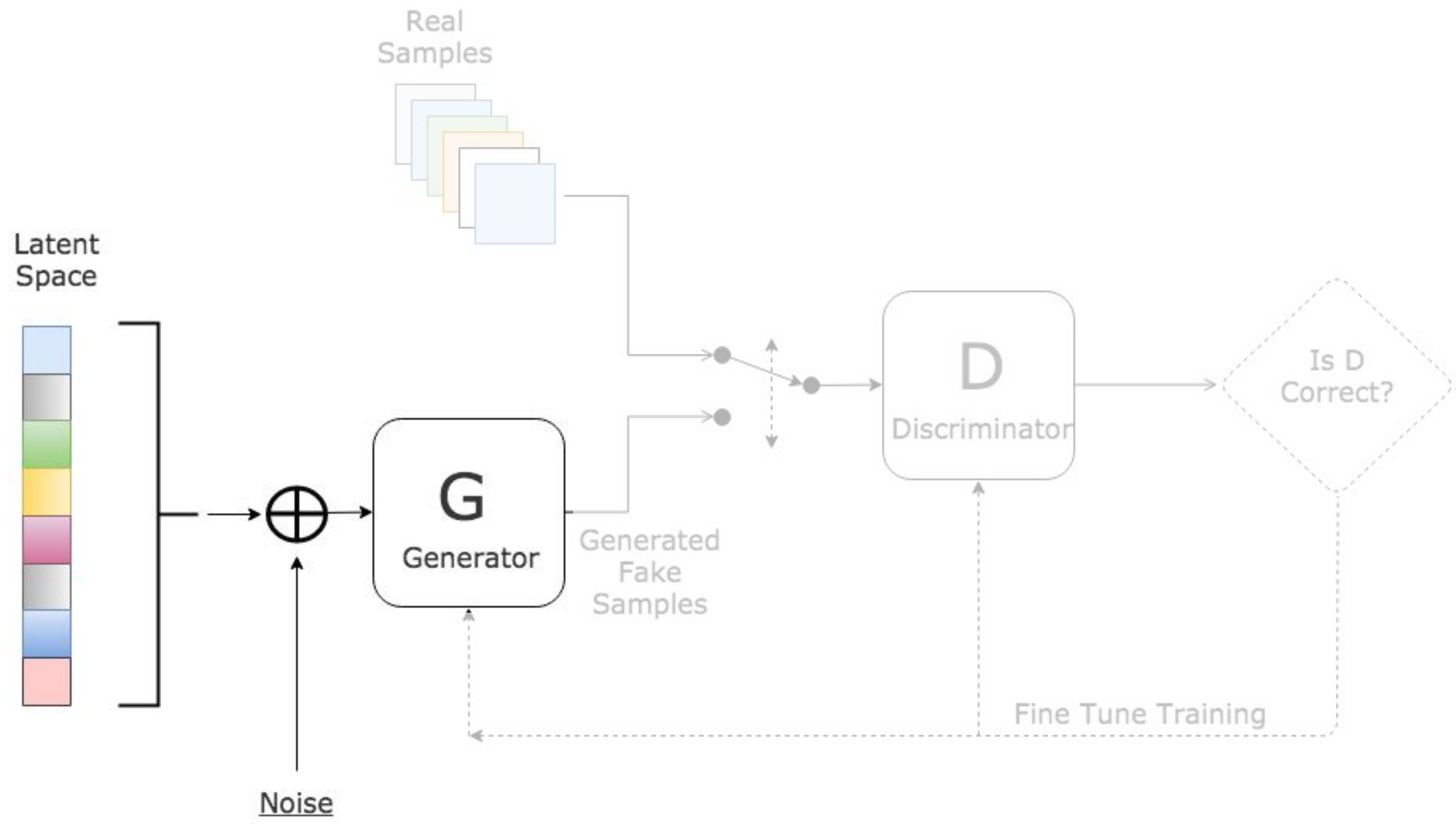




- A training dataset **x**
- No target y
- A model providing an output y at the minimum of the loss
- A loss function of x and y specifying the task
  - e.g., clustering: group similar objects together

Generative Adversarial Training

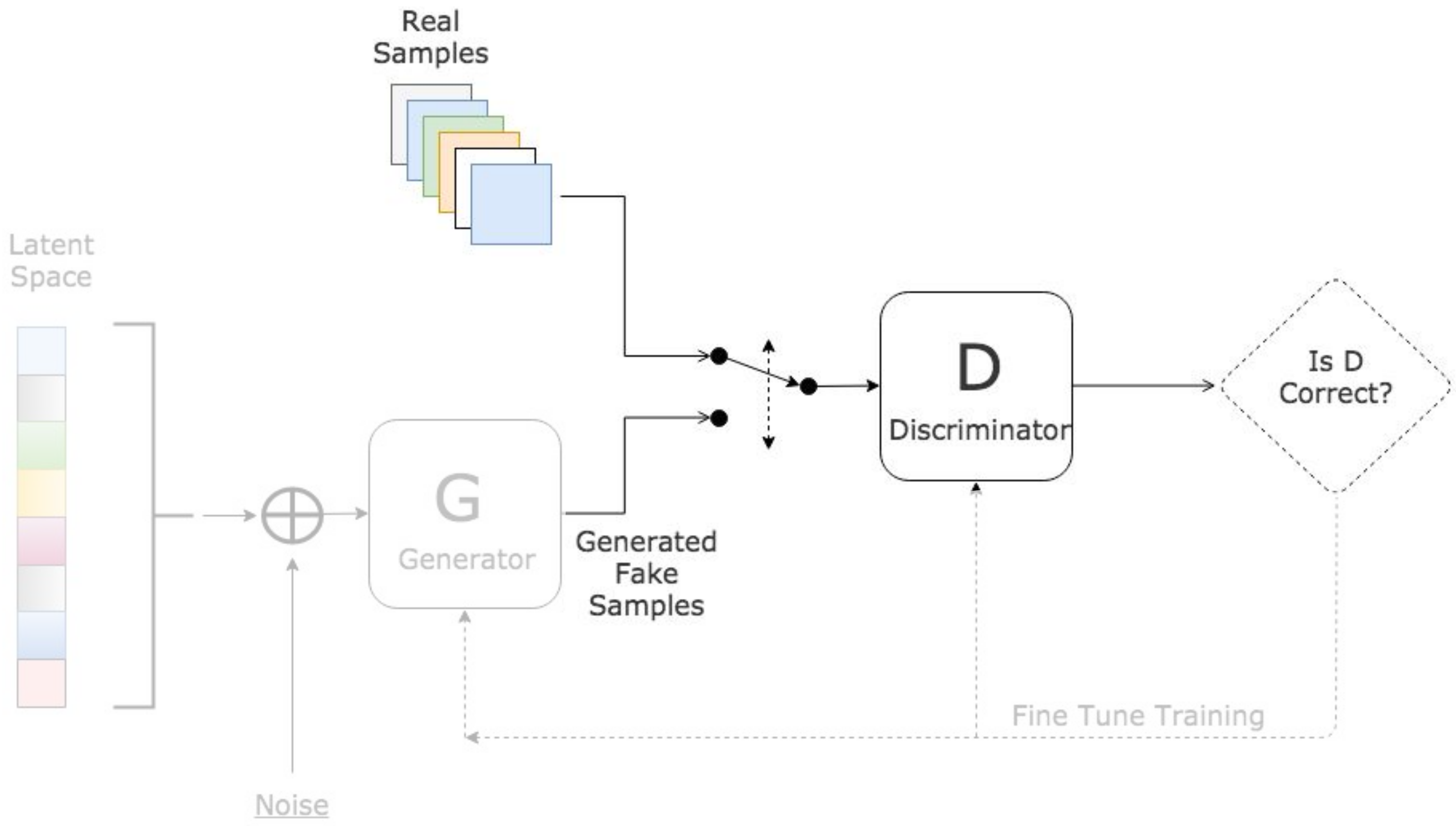
- *Two networks trained against each other*
  - ***Generator: create images (from noise, other images, etc)***
  - *Discriminator: tries to spot which image comes from the generator and which is genuine*



- *Loss function to minimise: Loss(Gen)-Loss(Disc)*
  - *Better discriminator -> bigger loss*
  - *Better generator -> smaller loss*
- *Trying to full the discriminatore, generatore learns how to create more realistic images*

Generative Adversarial Training

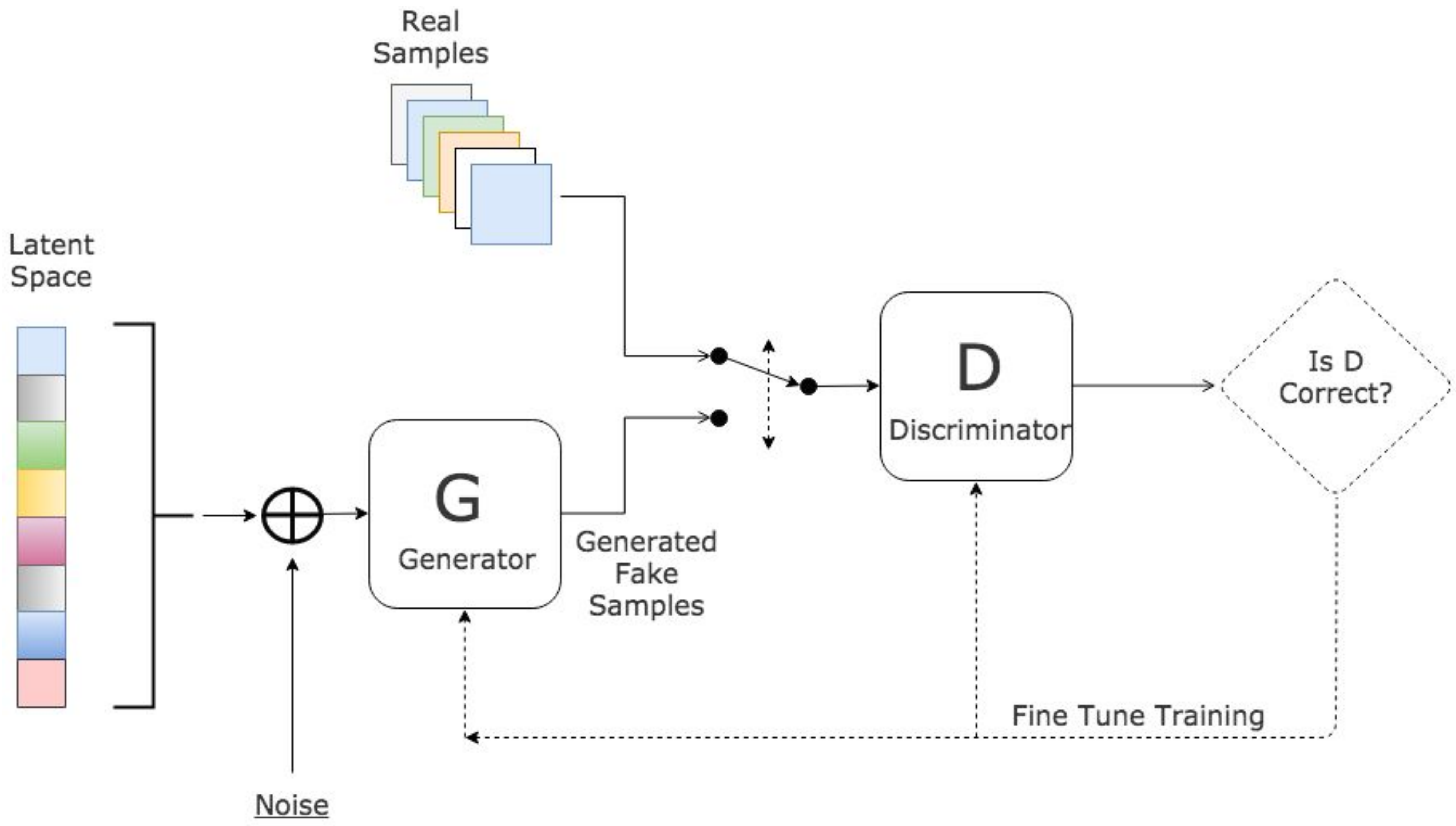
- *Two networks trained against each other*
  - *Generator: create images (from noise, other images, etc)*
  - ***Discriminator: tries to spot which image comes from the generator and which is genuine***



- *Loss function to minimise: Loss(Gen)-Loss(Disc)*
  - *Better discriminator -> bigger loss*
  - *Better generator -> smaller loss*
  - *Trying to full the discriminatore, generatore learns how to create more realistic images*

Generative Adversarial Training

- *Two networks trained against each other*
  - *Generator: create images (from noise, other images, etc)*
  - *Discriminator: tries to spot which image comes from the generator and which is genuine*



- ***Loss function to minimise: Loss(Gen)-Loss(Disc)***
  - ***Better discriminator -> bigger loss***
  - ***Better generator -> smaller loss***
- ***Trying to full the discriminatore, generatore learns how to create more realistic images***

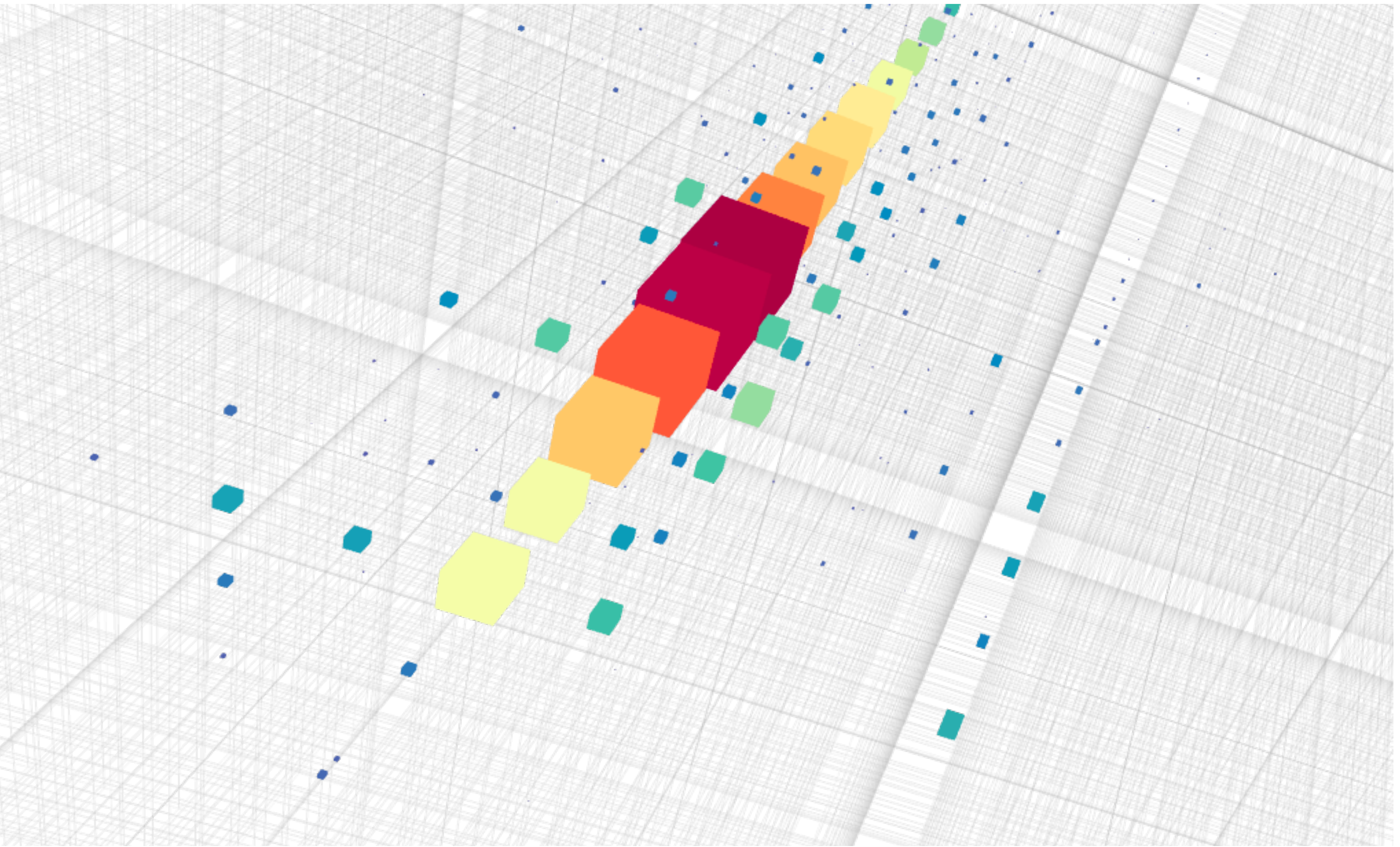
Generative Adversarial Training

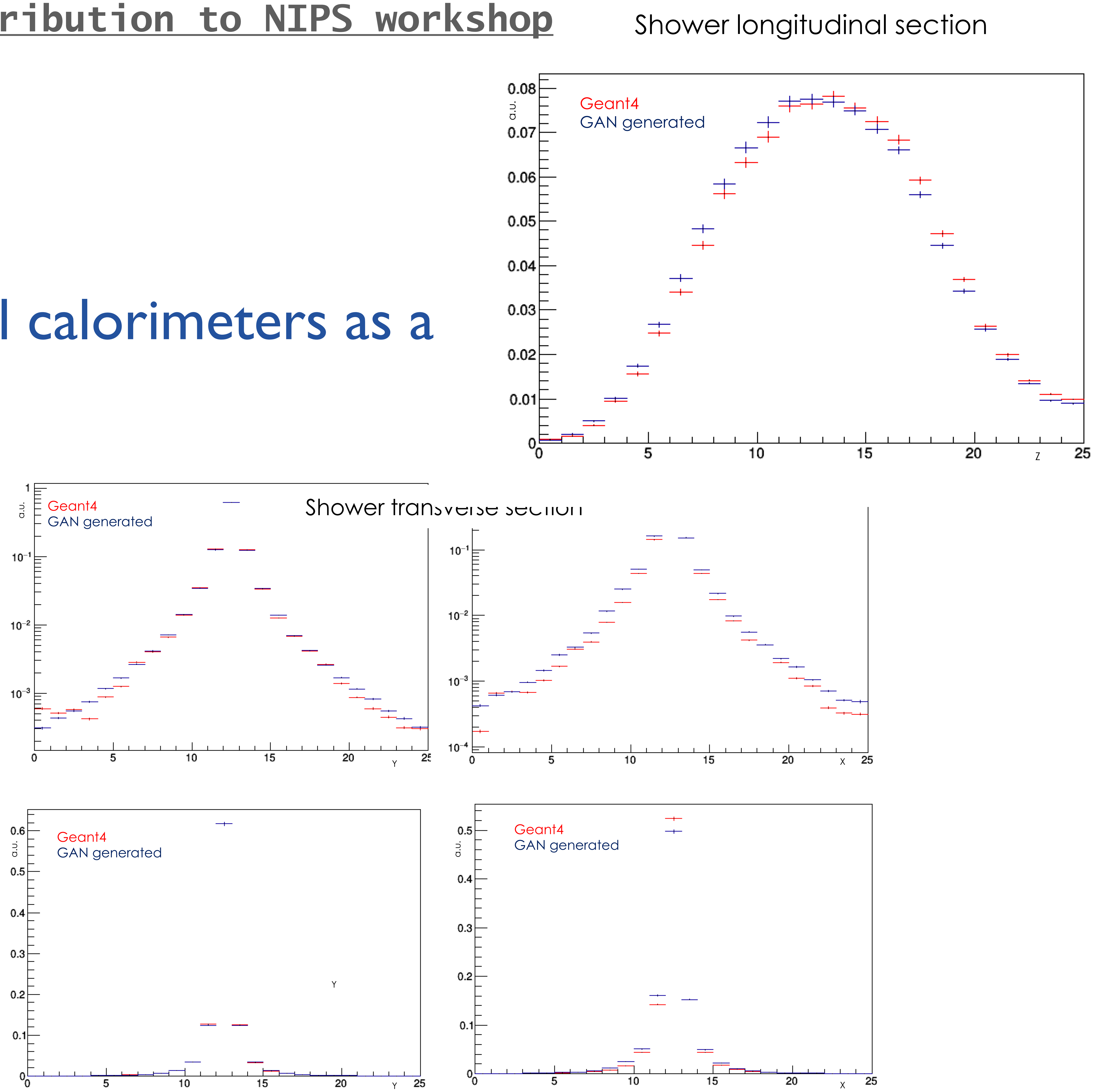
European
Research
Council

Particle shower generation

See contribution to NIPS workshop

- Start from random noise
- Works very well with images
  - Applied to electron showers in digital calorimeters as a replacement of GEANT

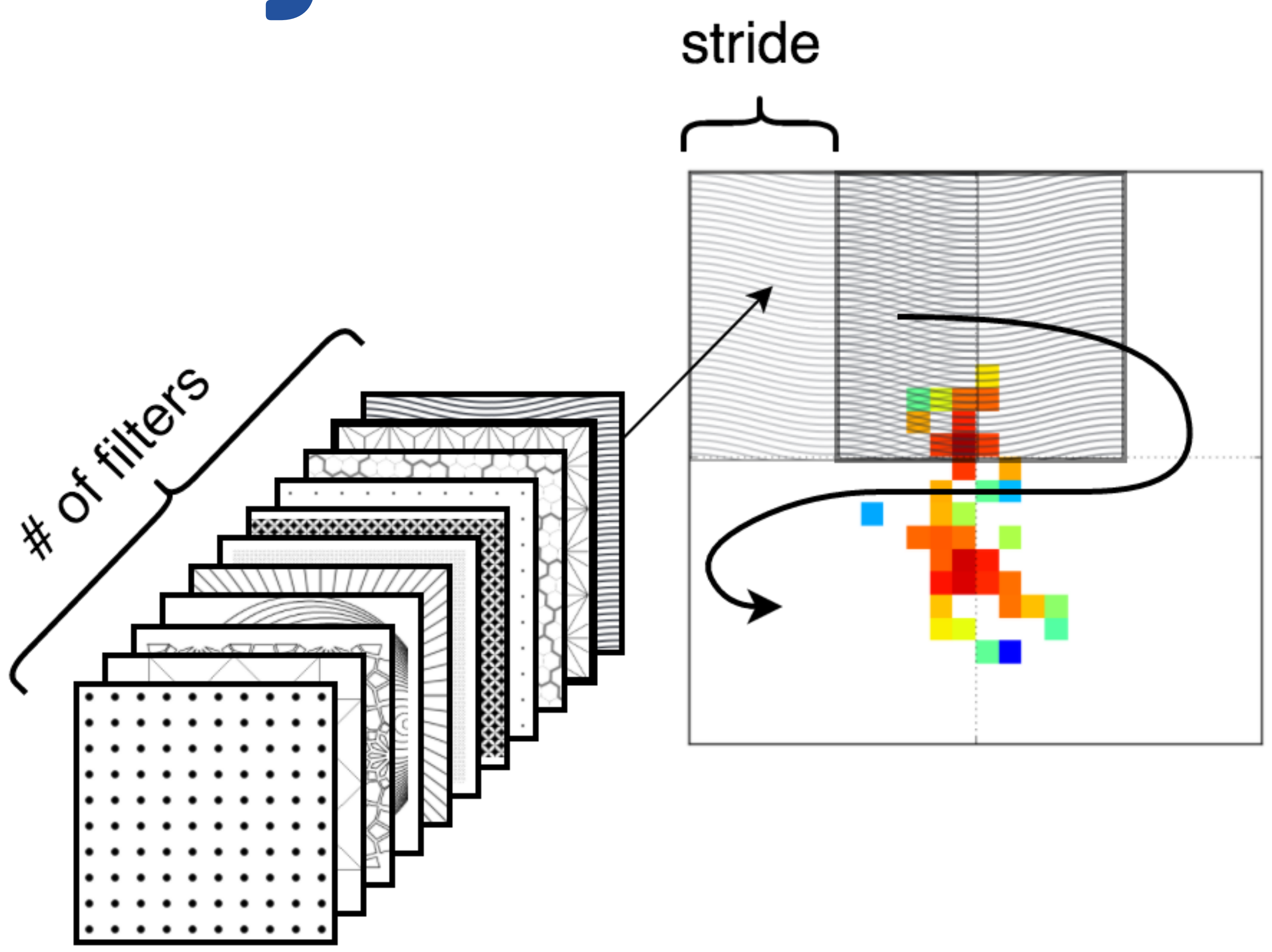


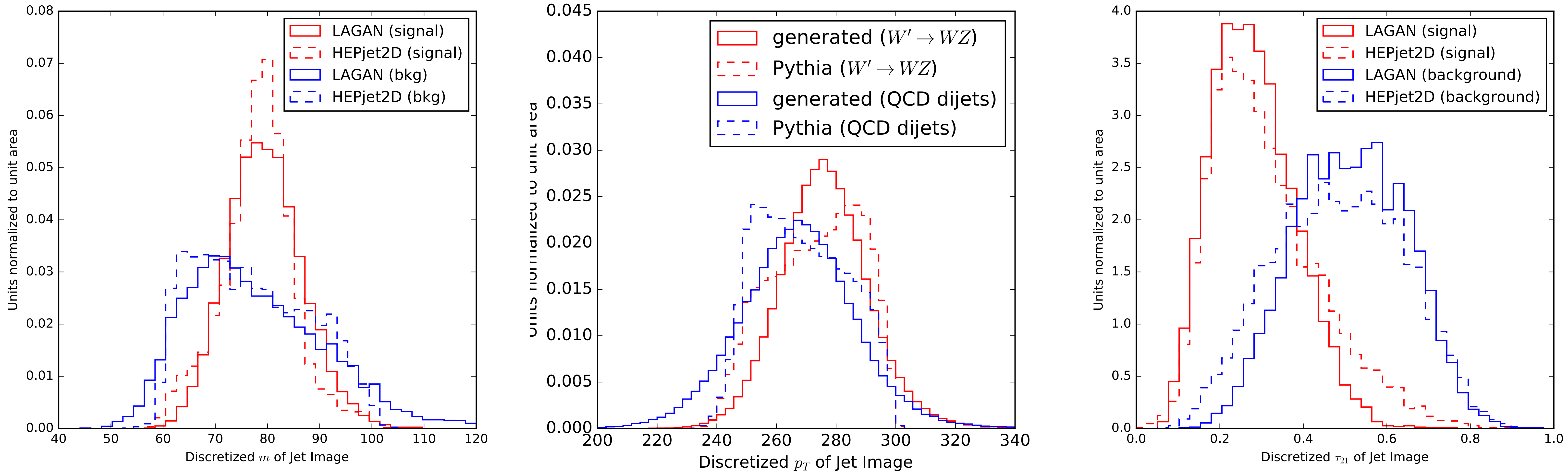


see also de Olivera, Paganini, and Nachman
https://arxiv.org/abs/1712.10321

Generating full jets

- Start from random noise
- Works very well with images
  - Applied to electron showers in digital calorimeters as a replacement of GEANT





de Olivera, Paganini, and Nachman
https://arxiv.org/pdf/1701.05927.pdf

Figure 6: The distributions of image mass $m(I)$, transverse momentum $p_{\mathrm{T}}(I)$, and n-subjettiness $\tau_{21}(I)$. See the text for definitions.

Autoencoders

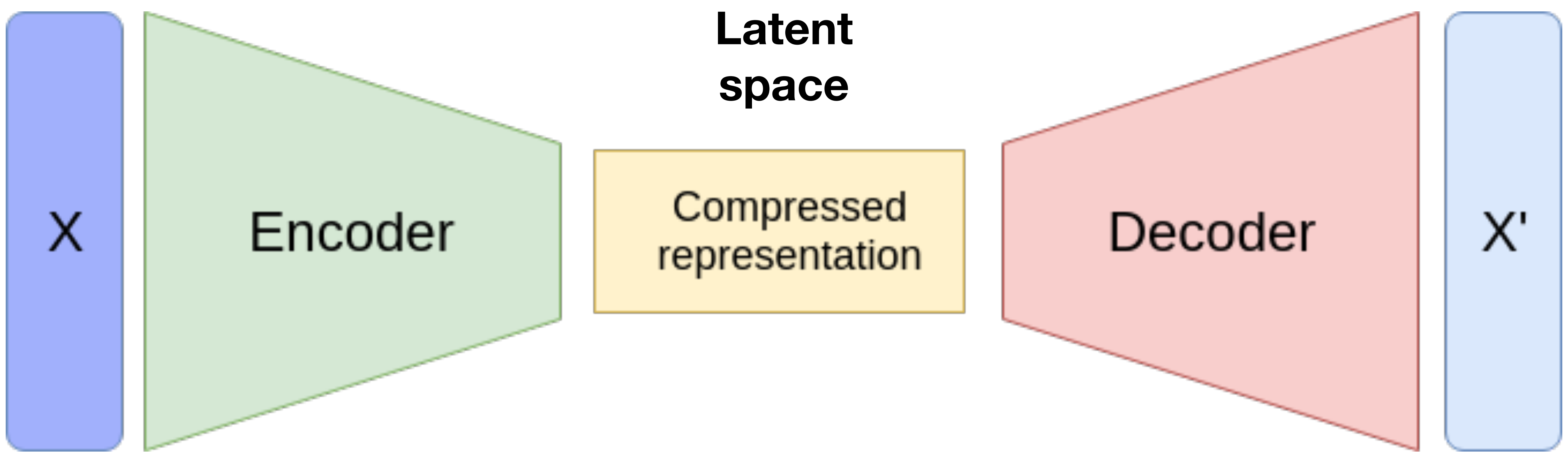
- *Autoencoders are networks with a typical "bottleneck" structure, with a symmetric structure around it*

  - *They go from* $\mathbb{R}^n \rightarrow \mathbb{R}^n$

  - *They are used to learn the identity function as* $f^{-1}(f(x))$

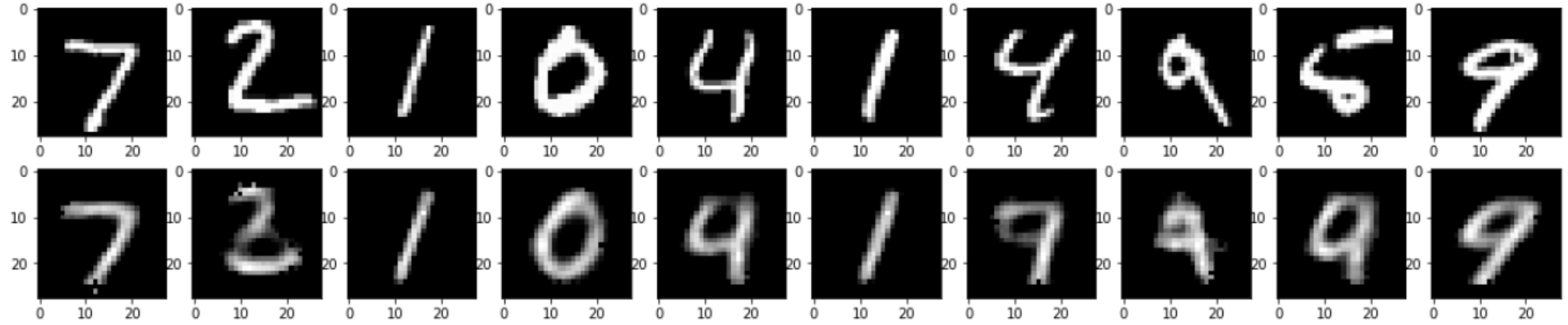
where $f: \mathbb{R}^n \rightarrow \mathbb{R}^k$ *and* $f^{-1}: \mathbb{R}^k \rightarrow \mathbb{R}^n$

- *Autoencoders are essential tools for unsupervised studies*



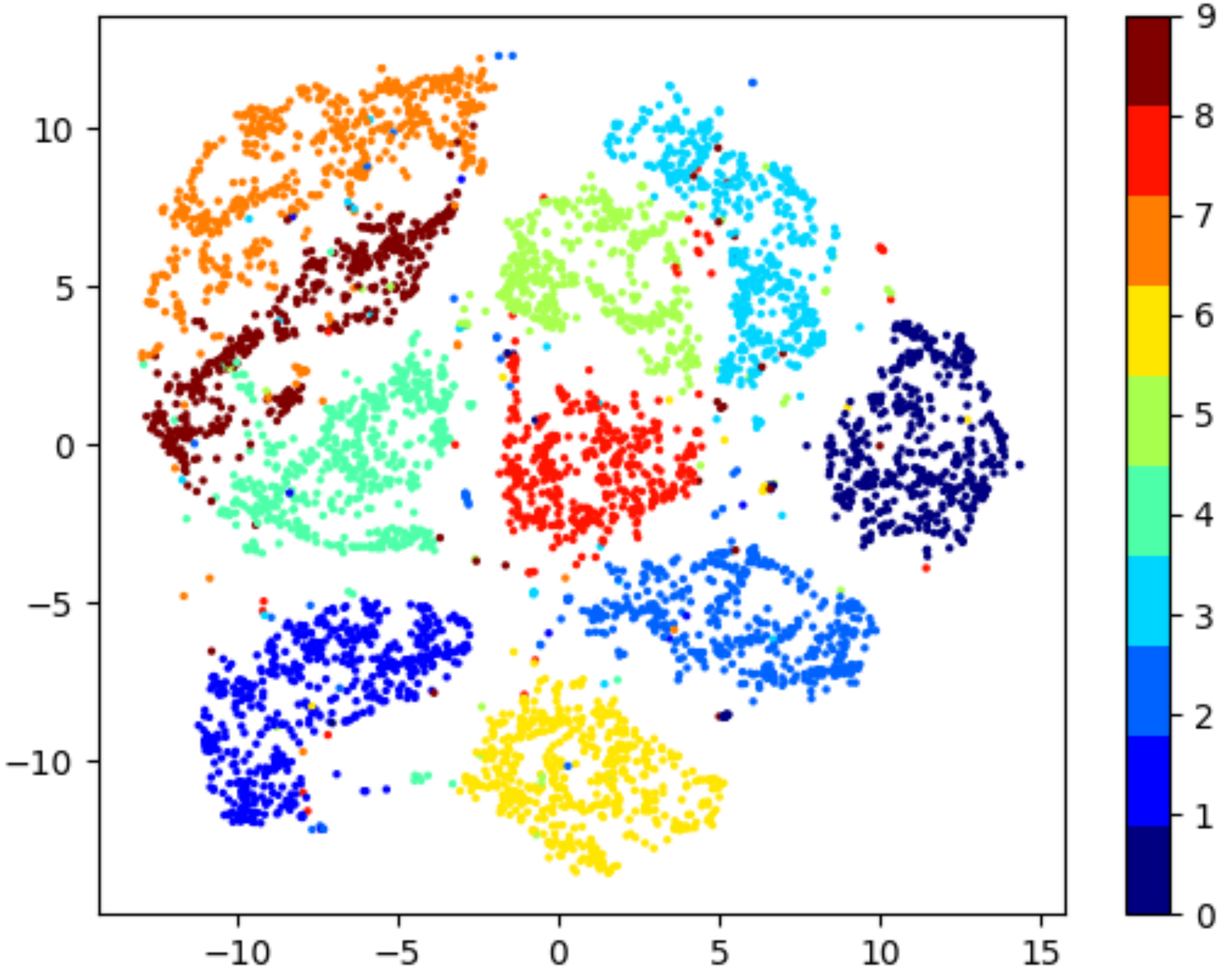
Dimensional Reduction

- *Autoencoders can be seen as compression algorithms*
  - *The n inputs are reduced to k quantities by the encoder*
  - *Through the decoder, the input can be reconstructed from the k quantities*
- *As a compression algorithm, an auto encoder allows to save (n-k)/n of the space normally occupied by the input dataset*



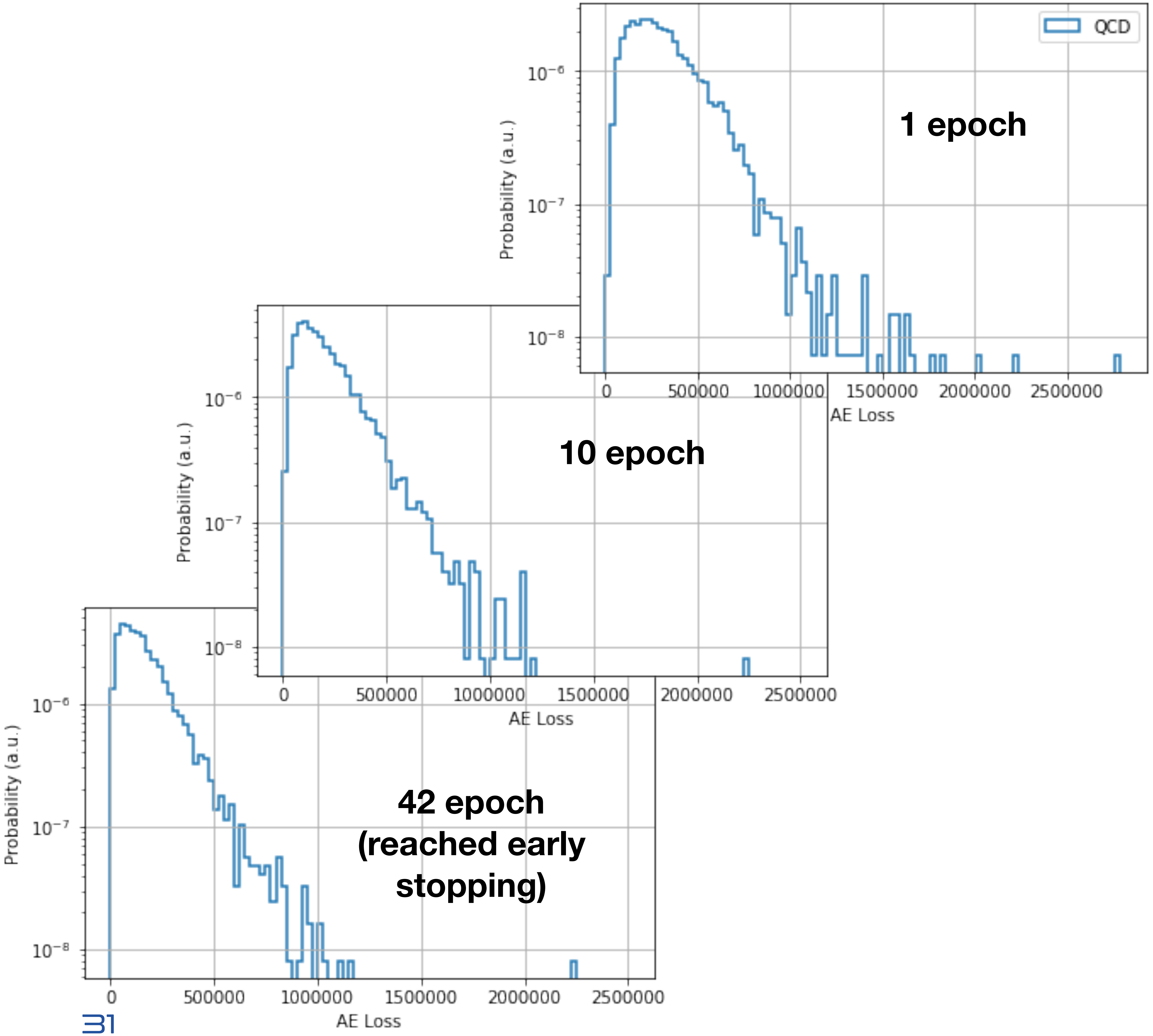
Clustering

- The auto encoder can be used as a clustering algorithm
  - Alike inputs tend to populate the same region of the latent space
  - Different inputs tend to be far away



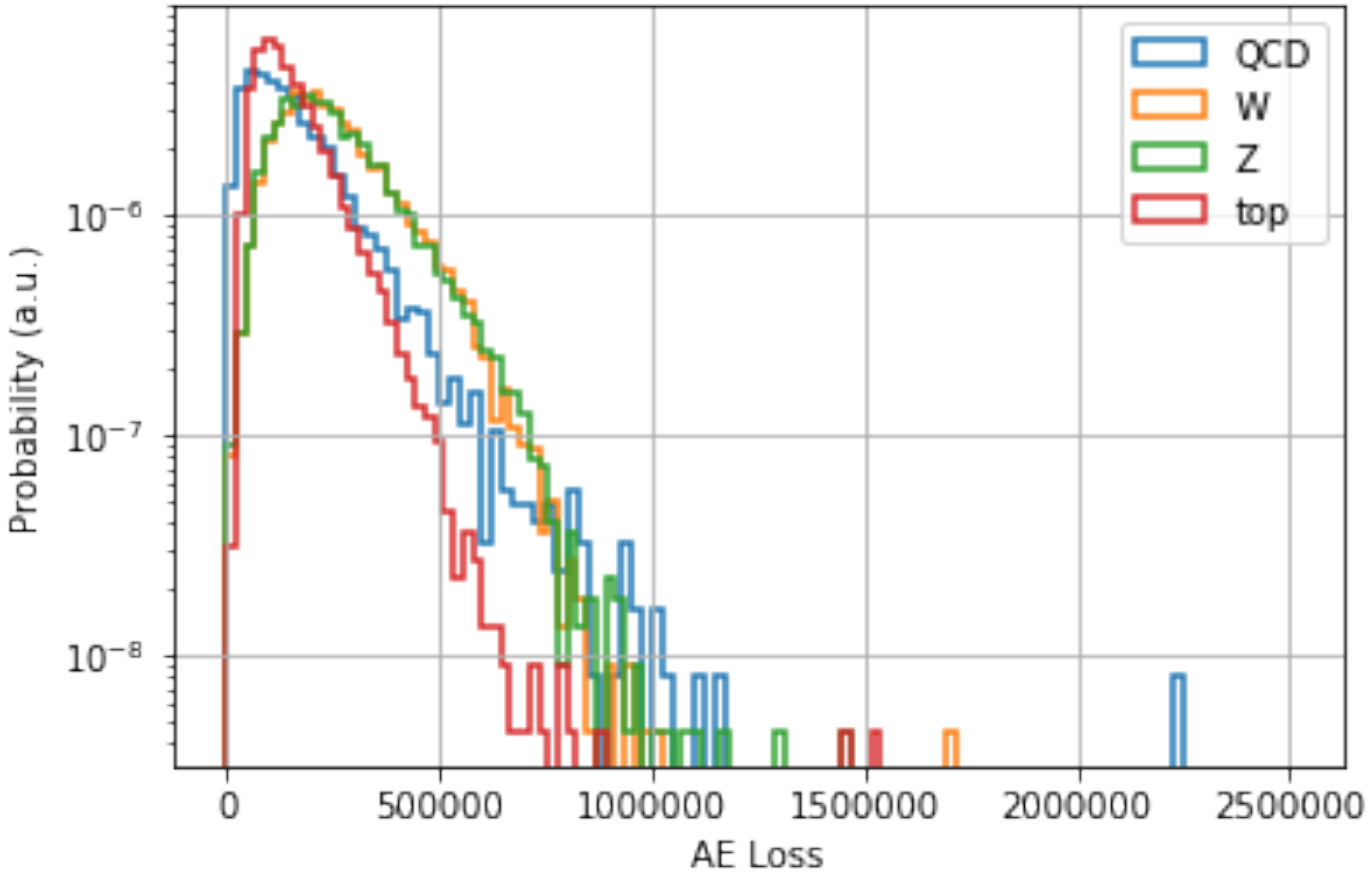
Training an Autoencoder

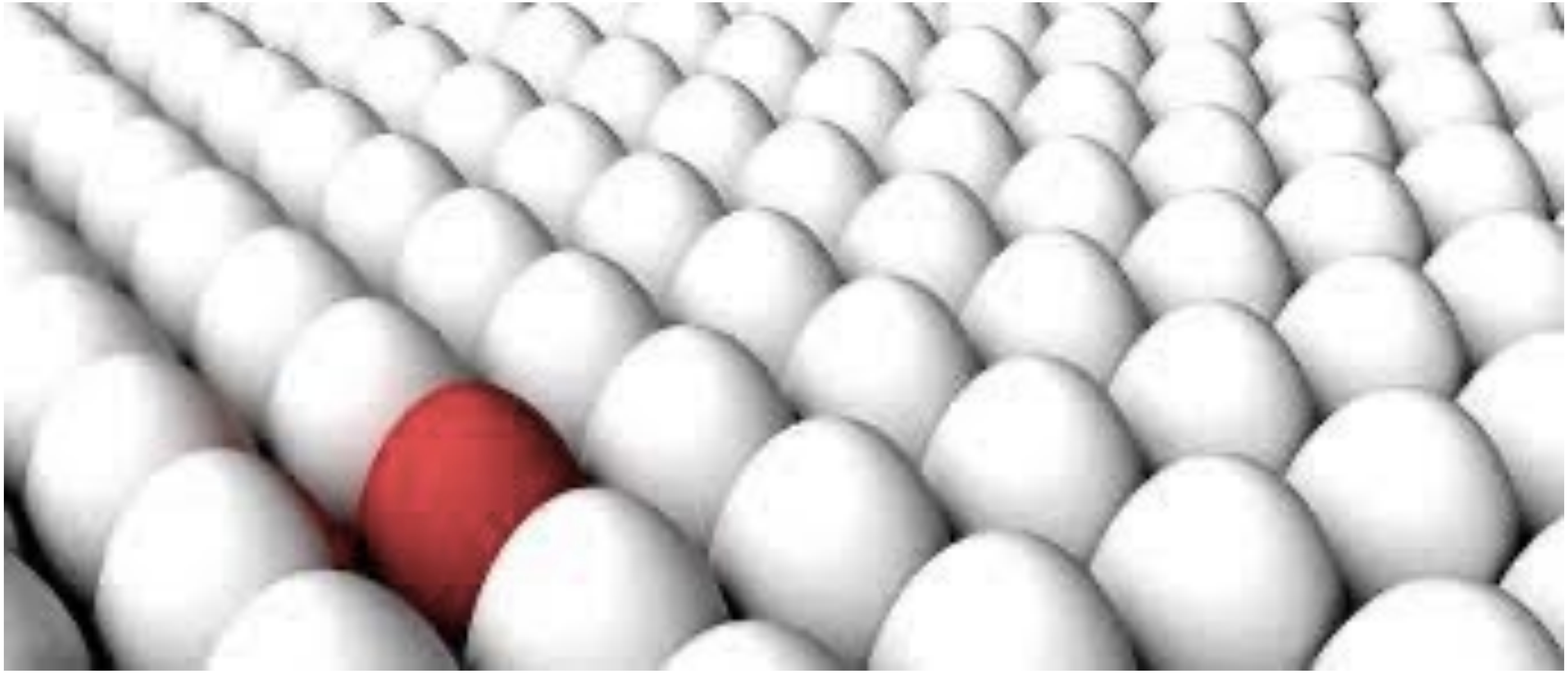
- AEs are training minimizing the distance between the inputs and the corresponding outputs
- The loss function represents some distance metric between the two
  - e.g., MSE loss
- A minimal distance guarantees that the latent representation + decoder is enough to reconstruct the input information



Anomaly detection

- *Once trained, an autoencoder can reproduce new inputs of the same kind of the training dataset*
  - *The distance between the input and the output will be small*
- *If presented an event of some new kind (anomaly), the encoding-decoding will tend to fail*
  - *In this circumstance, the loss (=distance between input and output) will be bigger*

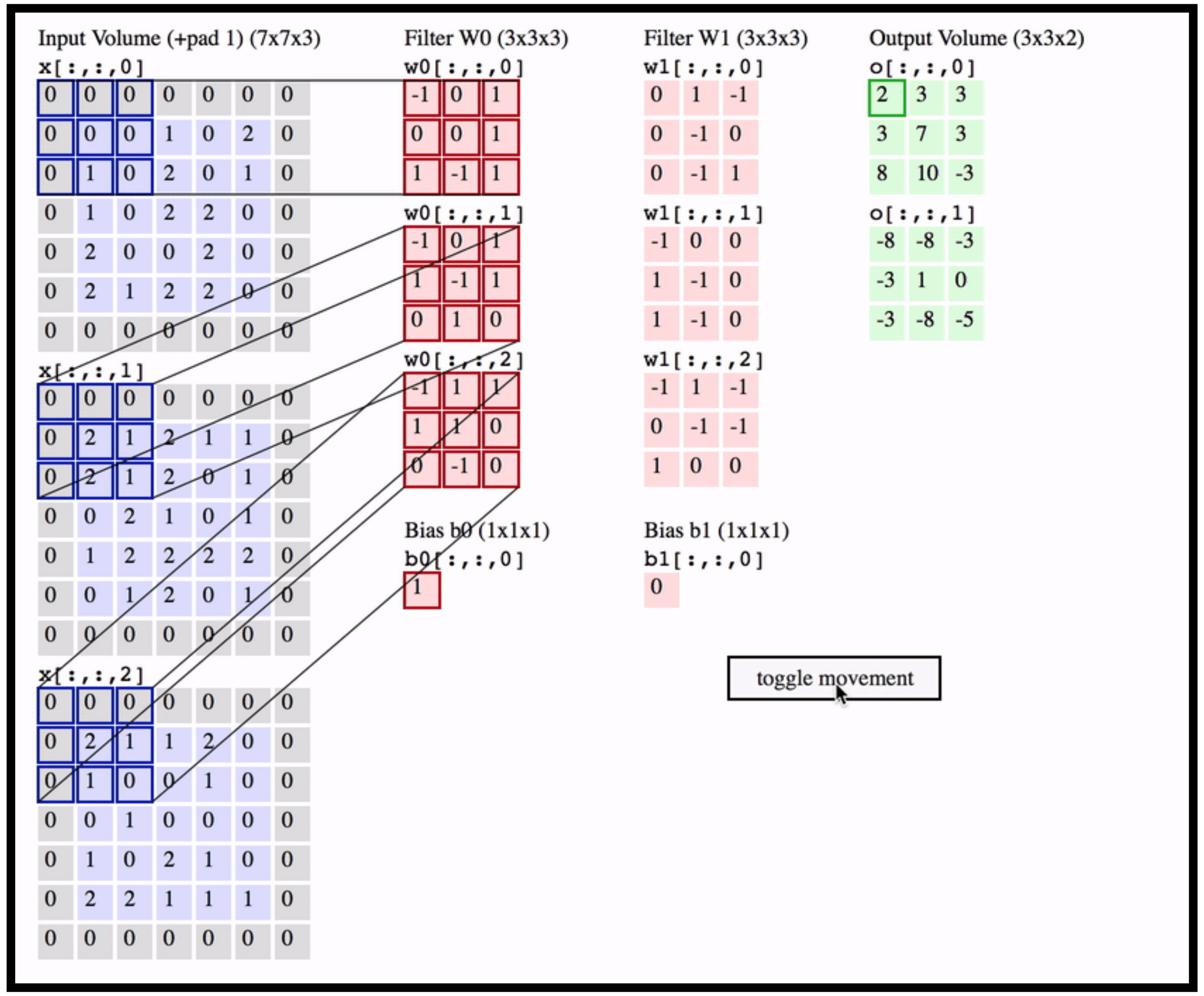


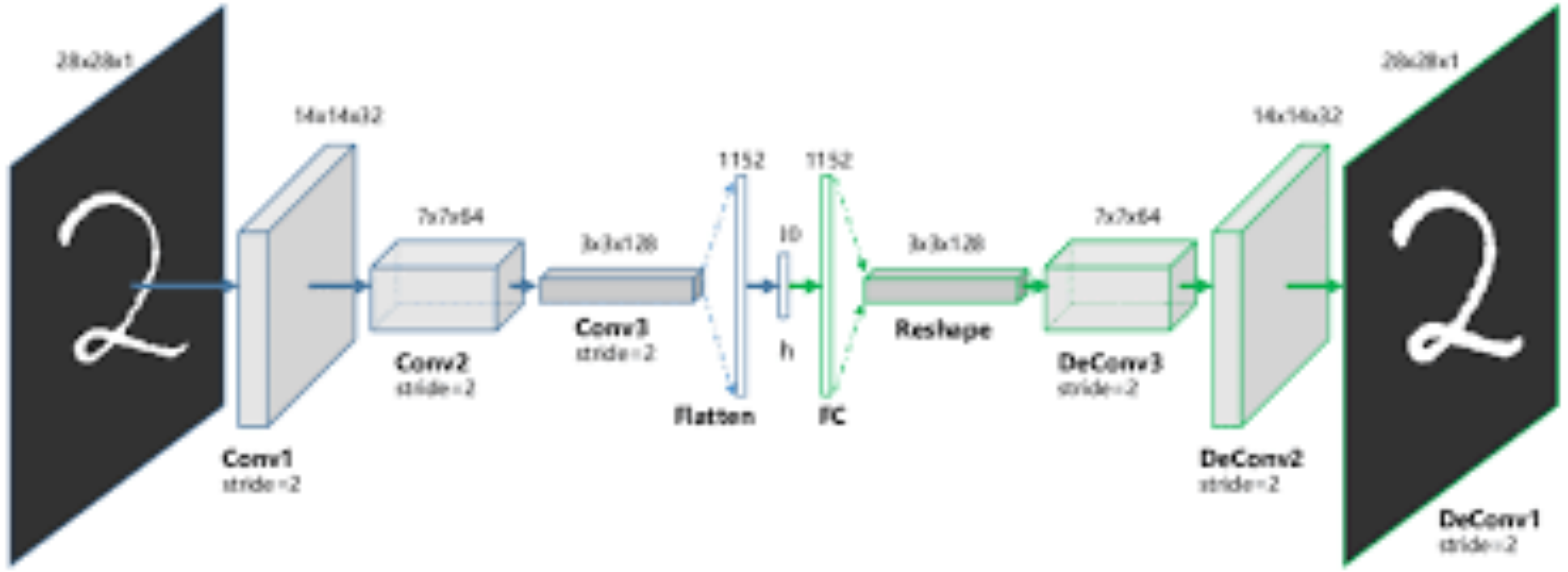


Looking at (a lot of) data with Anomaly Detection Algorithms

Convolutional Autoencoders

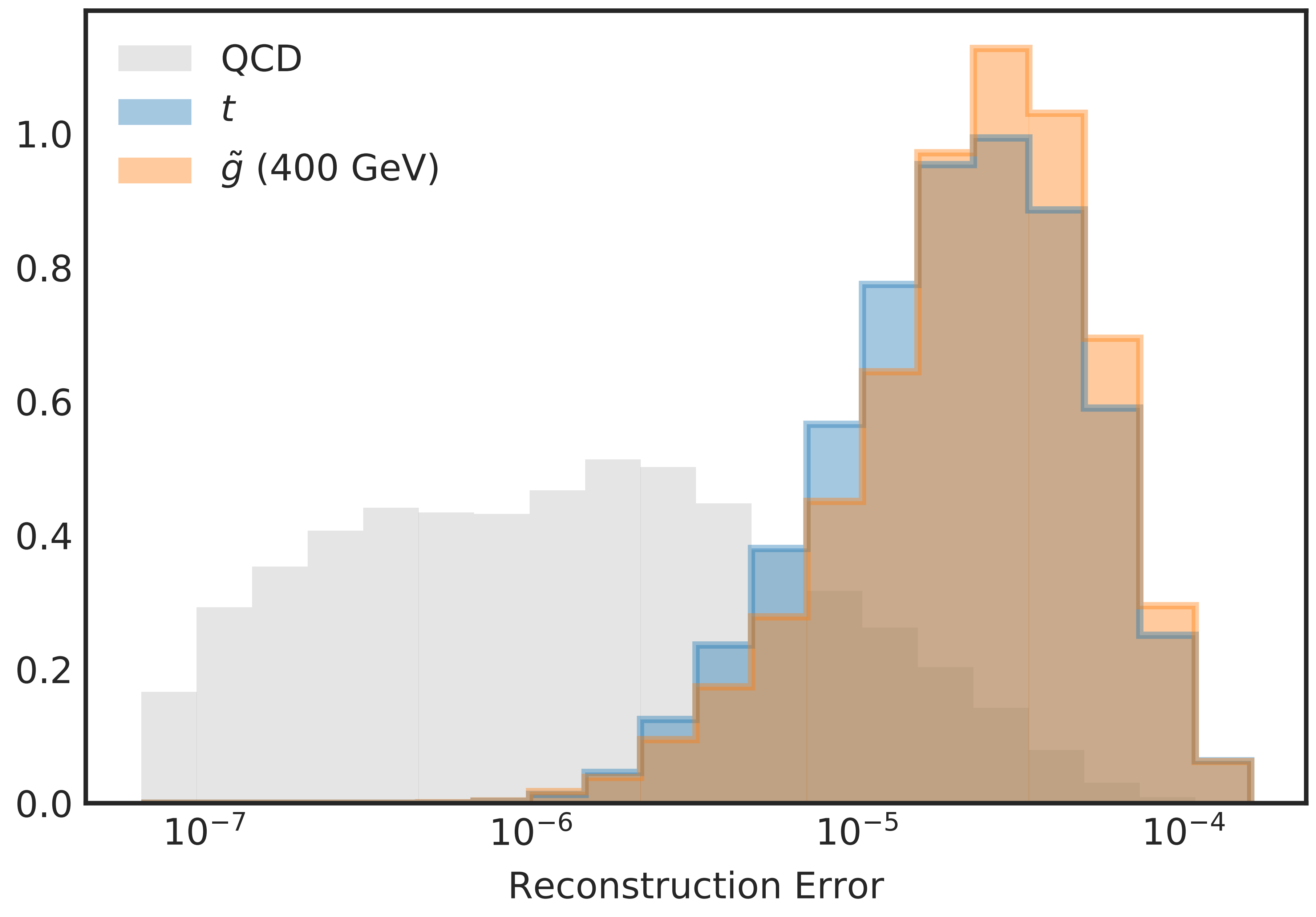
- *Conv Autoencoders take images as input*
- *They use convolutional layers to process these images and learn from them*
- *In the decoder ConvTranspose layers perform the inverse operation*

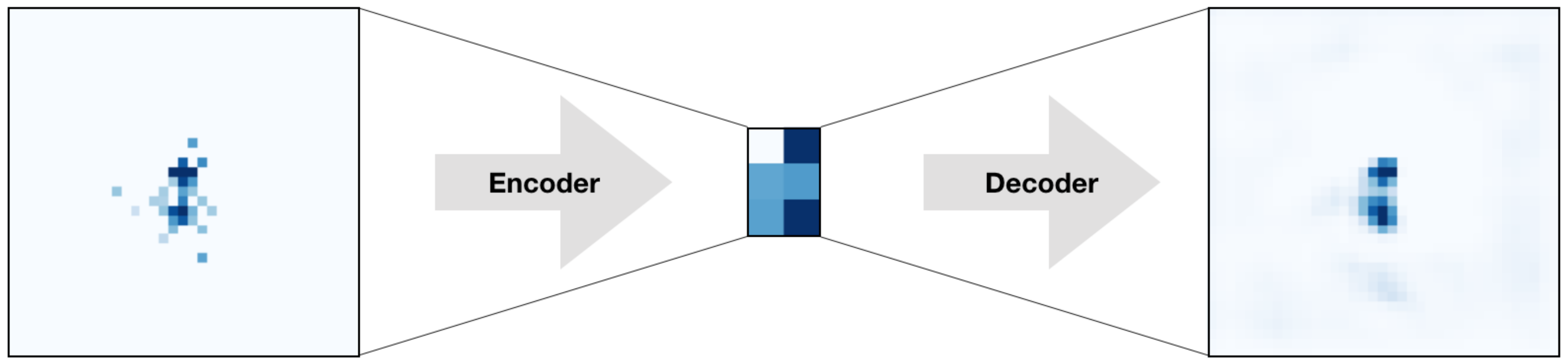


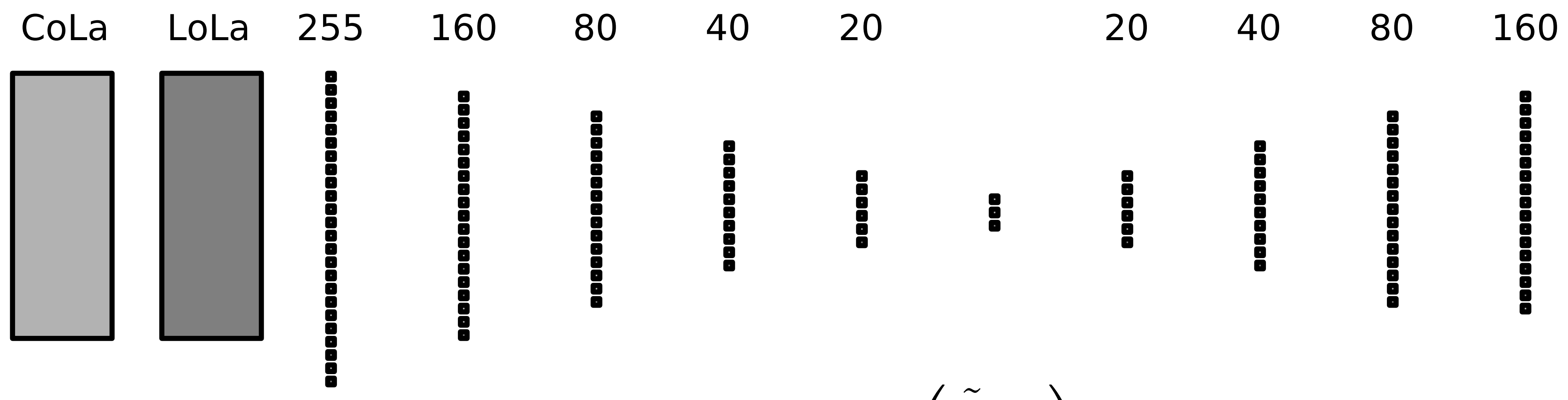


Example: Jet autoencoders

- *Idea applied to tagging jets, in order to define a QCD-jet veto*
- *Applied in a BSM search (e.g., dijet resonance) could highlight new physics signal*
- *Based on image and physics-inspired representations of jets*







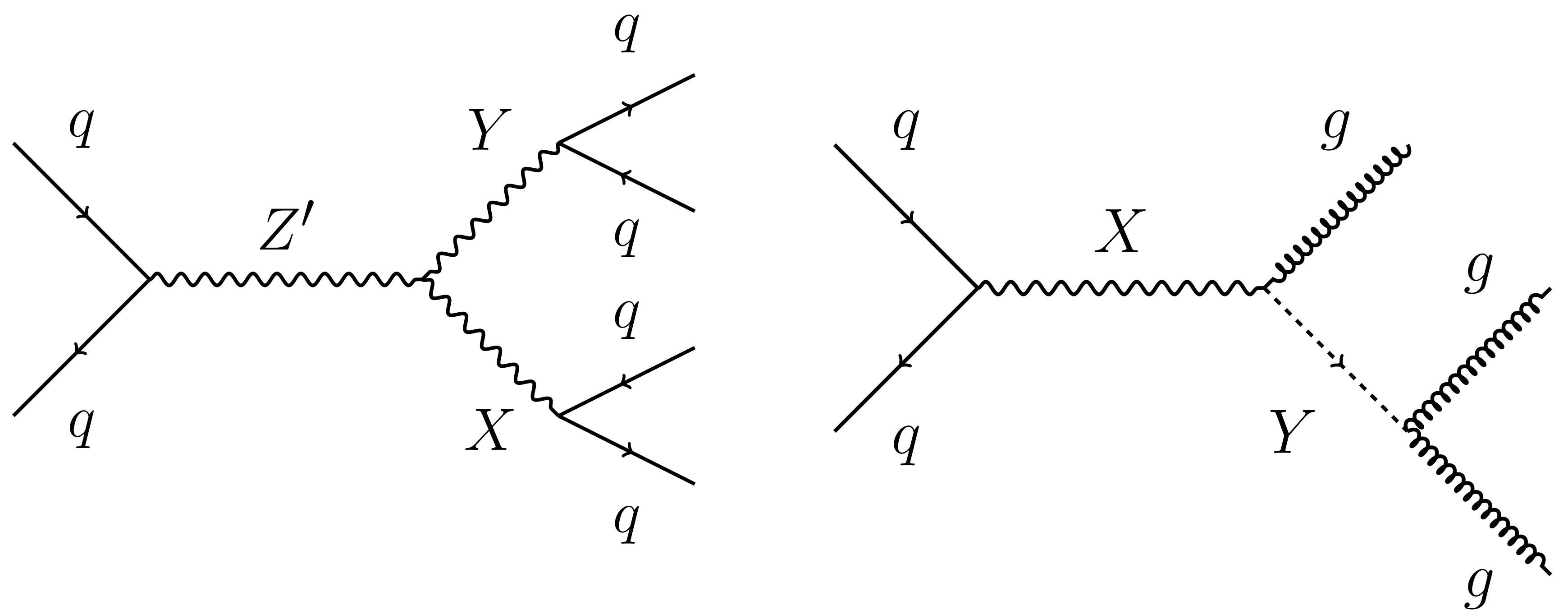
$$\tilde{k}_j = \begin{pmatrix} \tilde{k}_{0,j} \\ \tilde{k}_{1,j} \\ \tilde{k}_{2,j} \\ \tilde{k}_{3,j} \end{pmatrix} \xrightarrow{\text{LoLa}} \begin{pmatrix} \tilde{k}_{0,j} \\ \tilde{k}_{1,j} \\ \tilde{k}_{2,j} \\ \tilde{k}_{3,j} \\ \sqrt{\tilde{k}_j^2} \end{pmatrix}.$$

[Farina et al., arXiv:1808.08992]

[Heimel et al., arXiv:1808.08979]

LHC Olympics challenge

- Autoencoders are only one of the many possibilities to define an anomaly detection score
- A broad overview of possibilities in the 2020 LHC Olympics report
  - New particles produced and decaying to all-jets final states
  - Arranged in several Black boxes
  - Challengers asked to characterize the signal

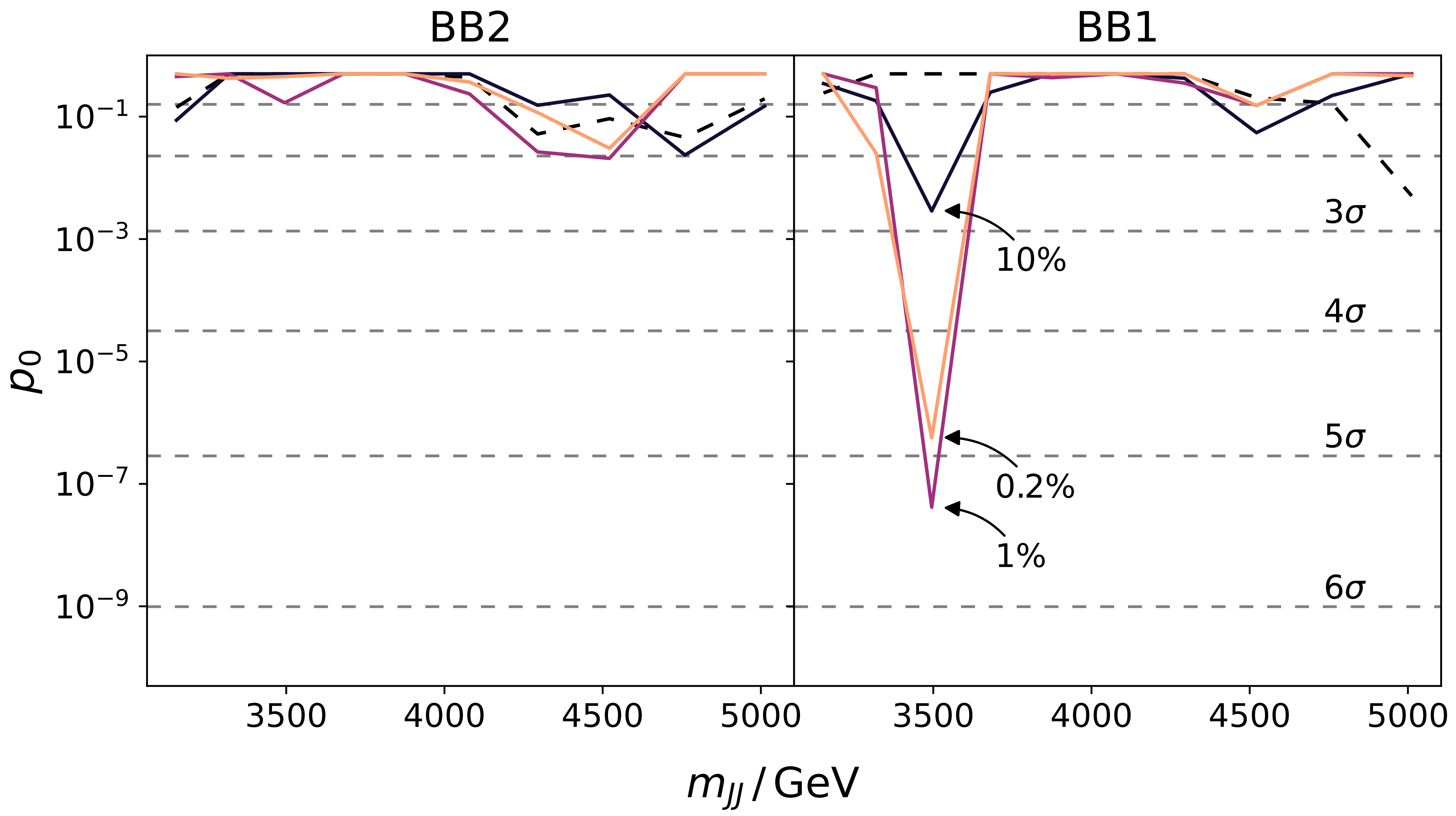


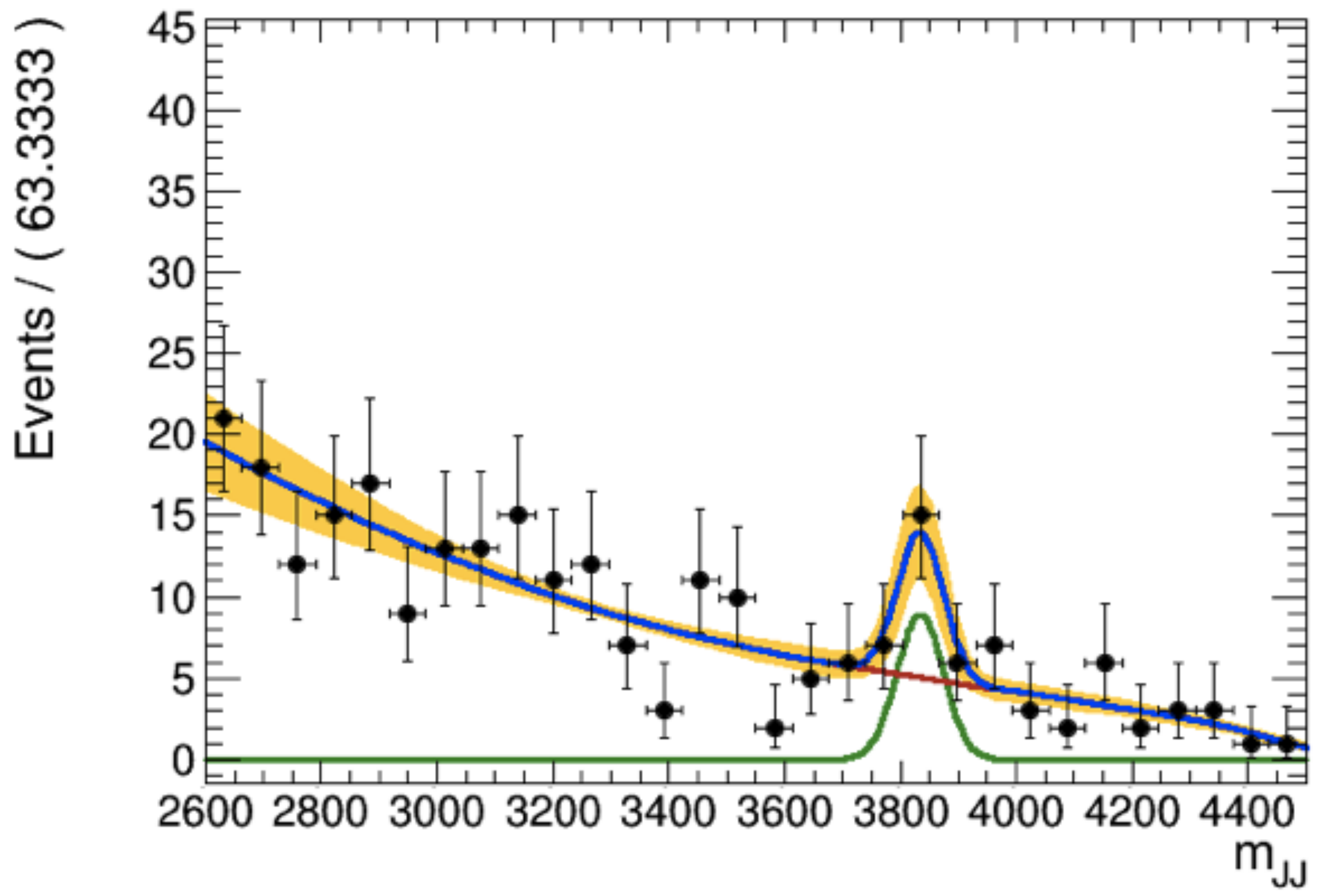
Section	Short Name	Method Type	Results Type
3.1	VRNN	Unsupervised	(i) (BB2,3) and (ii) (BB1)
3.2	ANODE	Unsupervised	(iii)
3.3	BuHuLaSpa	Unsupervised	(i) (BB2,3) and (ii) (BB1)
3.4	GAN-AE	Unsupervised	(i) (BB2-3) and (ii) (BB1)
3.5	GIS	Unsupervised	(i) (BB1)
3.6	LDA	Unsupervised	(i) (BB1-3)
3.7	PGA	Unsupervised	(ii) (BB1-2)
3.8	Reg. Likelihoods	Unsupervised	(iii)
3.9	UCluster	Unsupervised	(i) (BB2-3)
4.1	CWoLa	Weakly Supervised	(ii) (BB1-2)
4.2	CWoLa AE Compare	Weakly/Unsupervised	(iii)
4.3	Tag N' Train	Weakly Supervised	(i) (BB1-3)
4.4	SALAD	Weakly Supervised	(iii)
4.5	SA-CWoLa	Weakly Supervised	(iii)
5.1	Deep Ensemble	Semisupervised	(i) (BB1)
5.2	Factorized Topics	Semisupervised	(iii)
5.3	QUAK	Semisupervised	(i) (BB2,3) and (ii) (BB1)
5.4	LSTM	Semisupervised	(i) (BB1-3)

https://arxiv.org/pdf/2101.08320.pdf

LHC Olympics challenge

- *Autoencoders are only one of the many possibilities to define an anomaly detection score*
- *A broad overview of possibilities in the 2020 LHC Olympics report*
  - *New particles produced and decaying to all-jets final states*
  - *Arranged in several Black boxes*
  - *Challengers asked to characterize the signal*

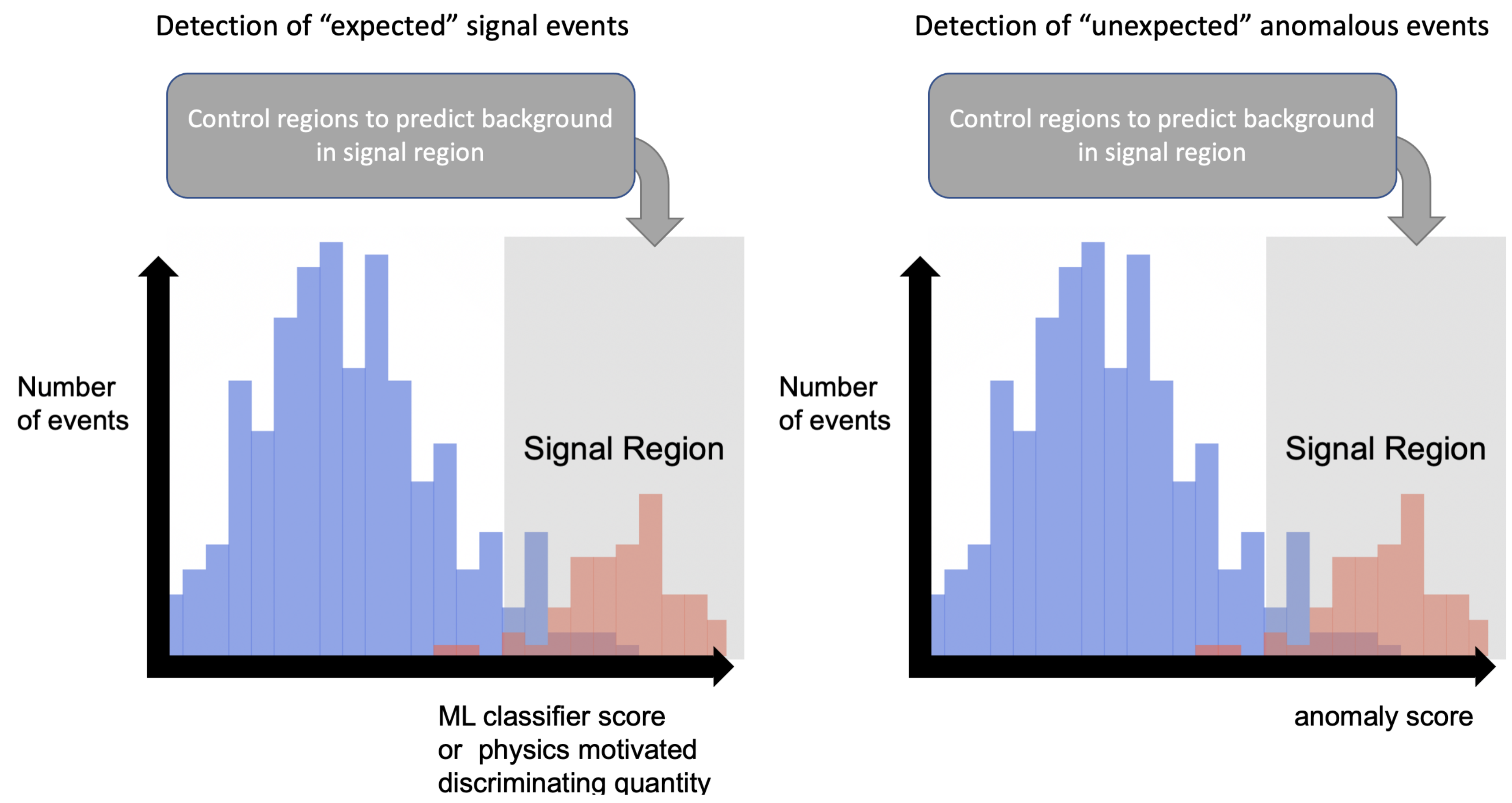




https://arxiv.org/pdf/2101.08320.pdf

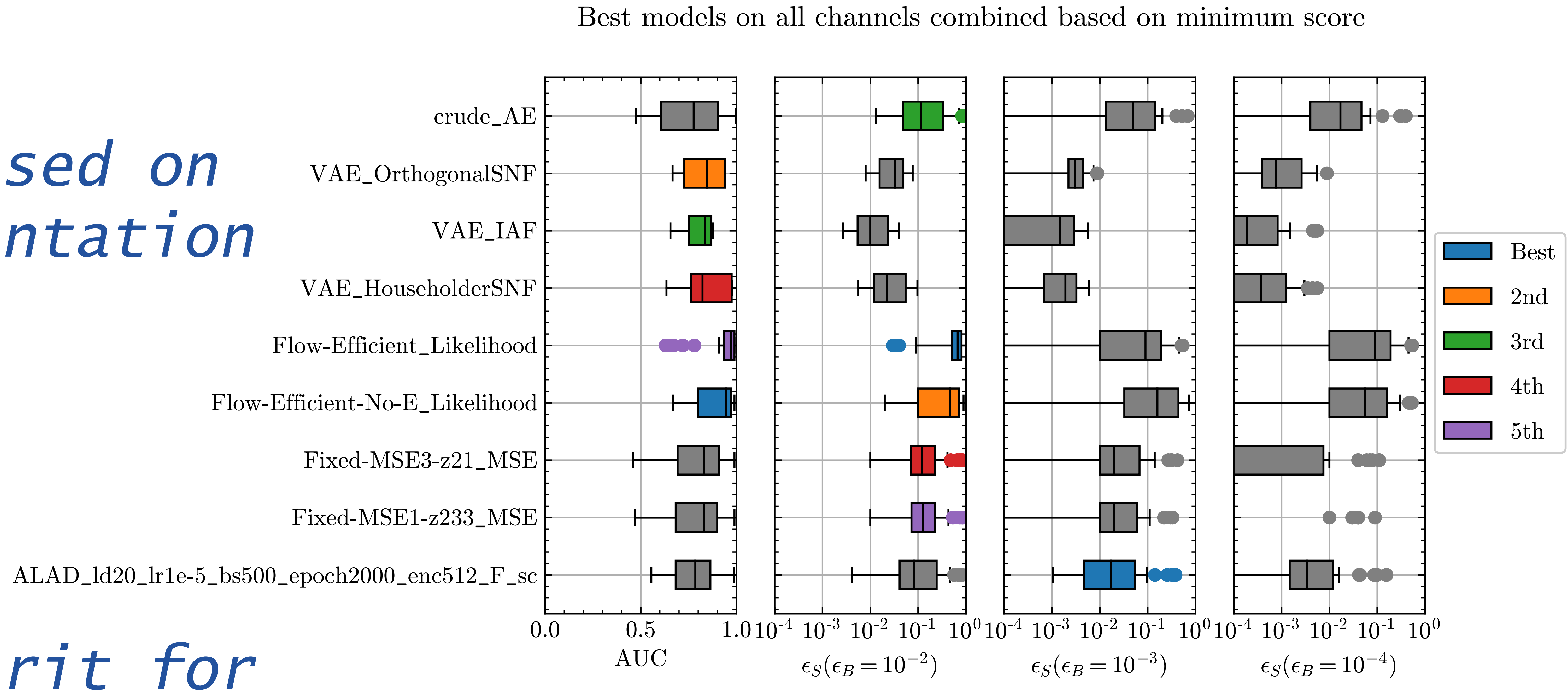
Dark Machine Challenge

- *Similar challenge, focusing on non-resonant signatures (e.g., SUSY)*
- *Similar methods, but based on the whole event representation*
- *Multiple final states considered (hadronic, leptonic, etc)*
- *Different figures of merit for different anomaly detection algorithms (signal efficiency @ different rejection values)*
- *Report coming soon on arXiv*



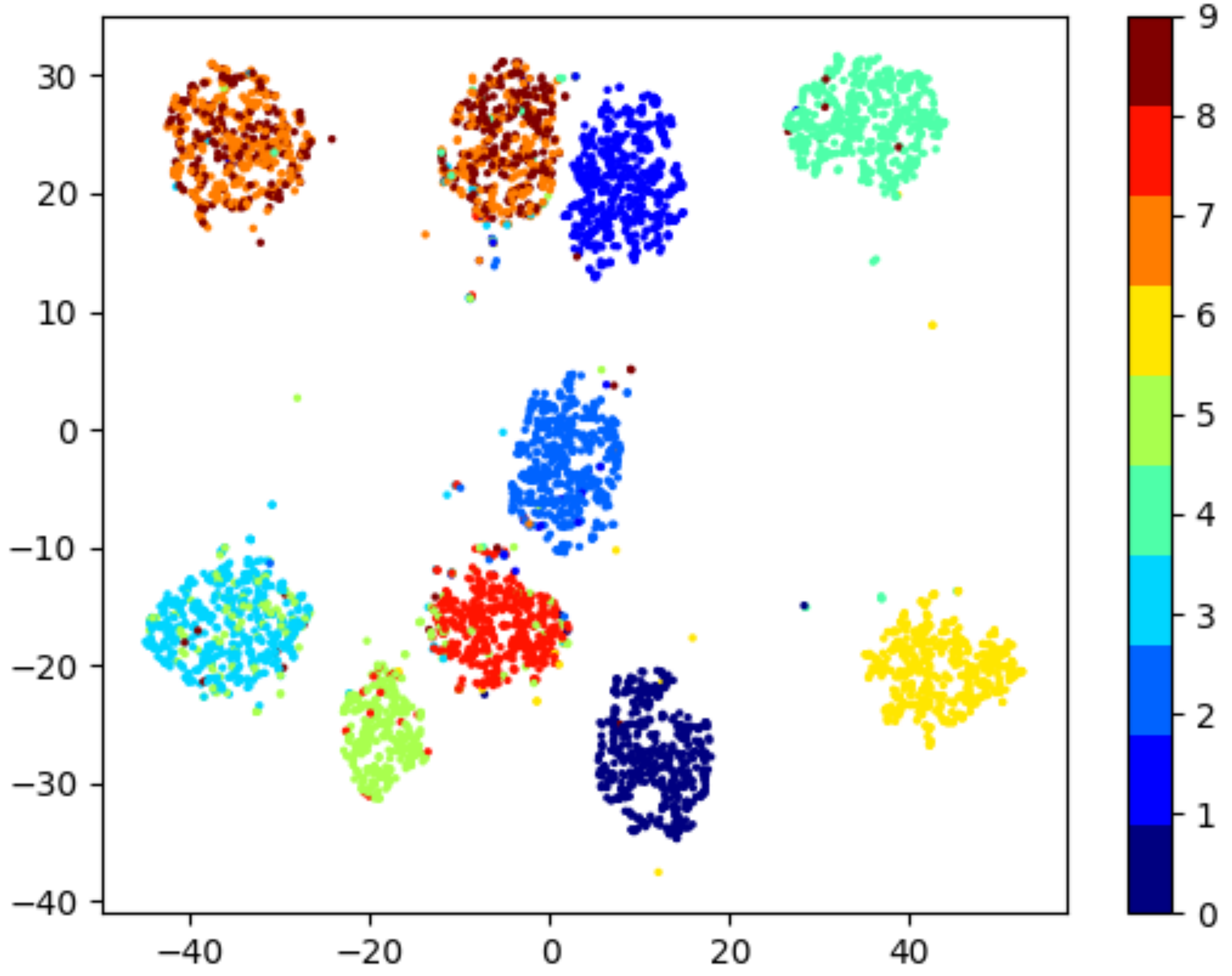
Dark Machine Challenge

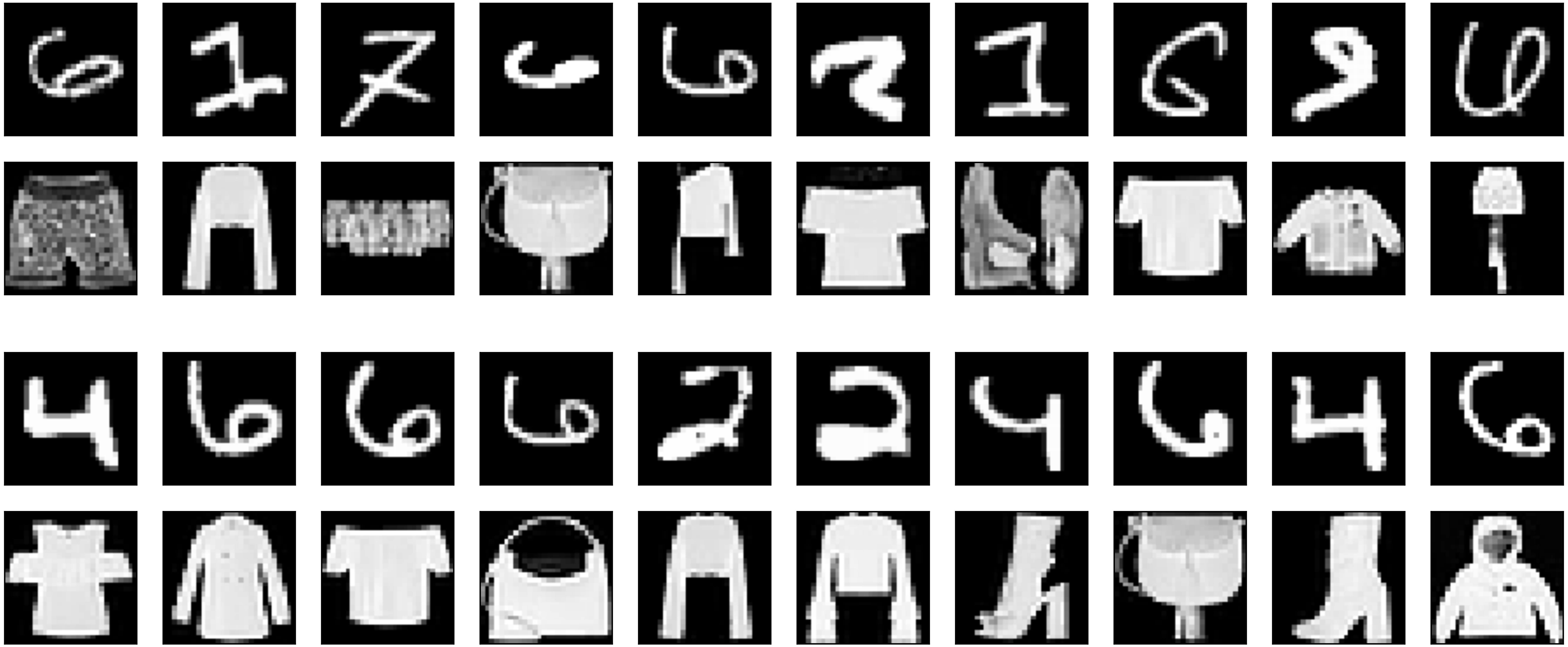
- *Similar challenge, focusing on non-resonant signatures (e.g., SUSY)*
- *Similar methods, but based on the whole event representation*
- *Multiple final states considered (hadronic, leptonic, etc)*
- *Different figures of merit for different anomaly detection algorithms (signal efficiency @ different rejection values)*
- *Report coming soon on arXiv*



What to do with these data?

- We could learn a lot running clustering algorithms (KNN, etc) on these data
  - In the latent space of the AE
  - In the natural space of the input
  - With any other similar technique
- In my mind, a descriptive paper on such an analysis would be a valuable publication, particularly before a long shutdown.
- Provided control on the background distribution (not for granted), we could run a statistical analysis on them and quote a significance (e.g., with https://arxiv.org/abs/1806.02350)
- Publishing the dataset as a catalog could incentive new ideas in view of HL-LHC
- While we sort out the technical details (e.g., with TSG and L1), we would like to request the EXO PAG to support the idea





Unsupervised Hypothesis Test

"Model-independent" hypothesis test

- *Deep Learning could help relaxing the underlying hypotheses of a new-physics search*
  - *stay within the hypothesis test framework*
  - *replace the fully specified (model dependent) signal hypothesis with a neural network trained on data*
  - *exploit neural networks to express different model shapes at once*
- *Training setup to learn the likelihood ratio of a traditional search*
- *Formally, still a fully-supervised learning process*

$$\boxed{L[f] = \sum_{(x,y)} \left[(1-y)\frac{N(\mathrm{R})}{\mathcal{N}_{\mathcal{R}}}\left(e^{f(x)}-1\right) - y\, f(x) \right]}$$

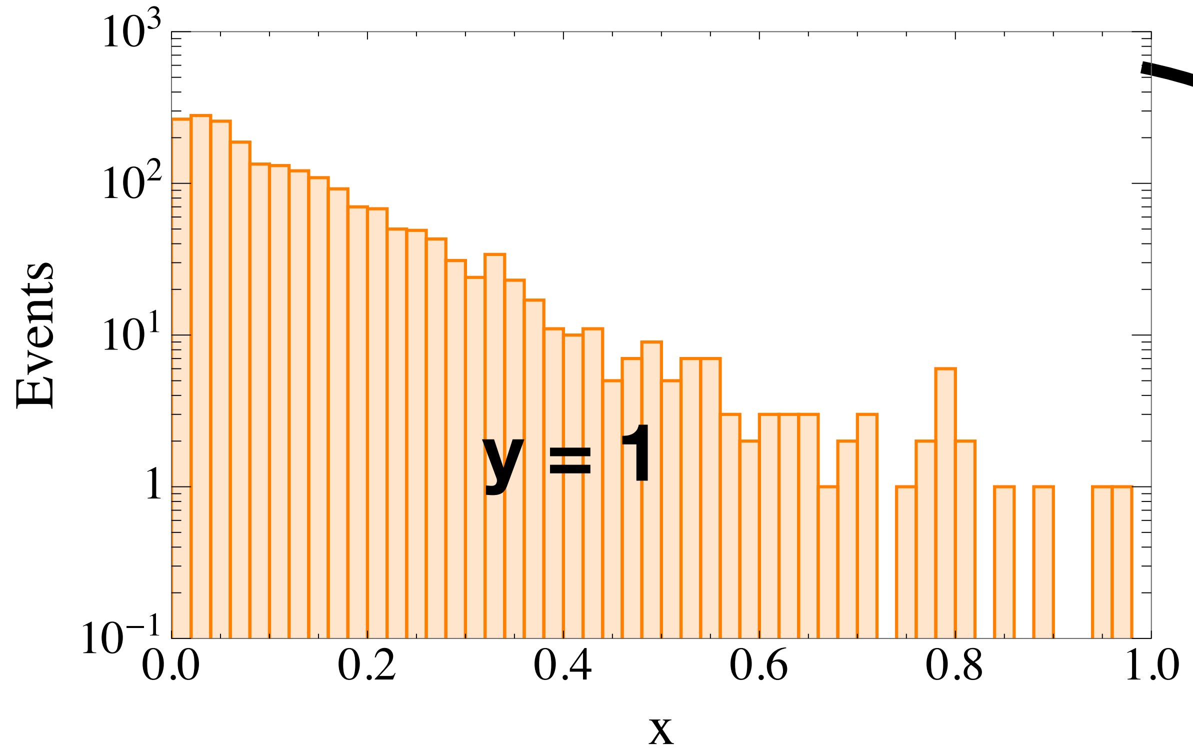
$$\underset{\{\mathbf{w}\}}{\mathrm{Min}}\, L = -\underset{\{\mathbf{w}\}}{\mathrm{Max}} \left\{ \log\left[\frac{e^{-N(\mathbf{w})}}{e^{-N(\mathrm{R})}} \prod_{x\in\mathcal{D}} \frac{n(x|\mathbf{w})}{n(x|\mathrm{R})} \right] \right\} = -\frac{t(\mathcal{D})}{2}\,.$$

[D'Agnolo et al., arXiv:1806.02350](https://arxiv.org/abs/1806.02350)
[D'Agnolo et al., arXiv:1912.12155](https://arxiv.org/abs/1912.12155)

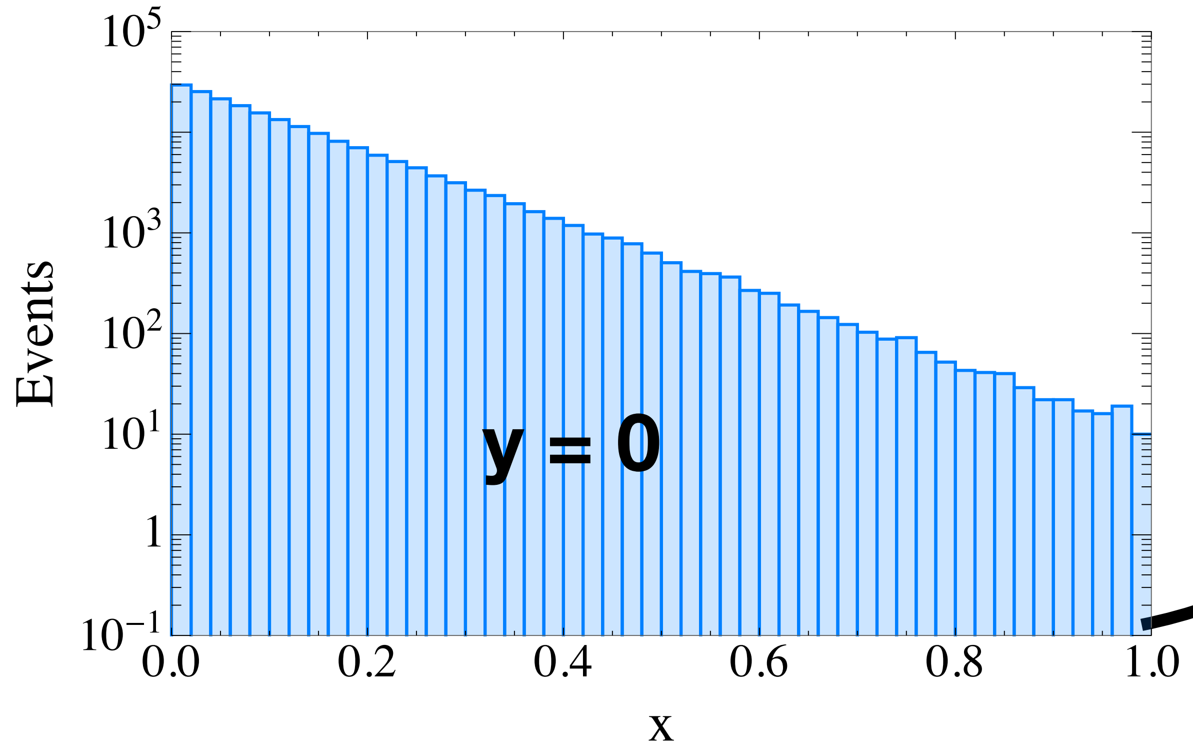
New Physics Learning Machine

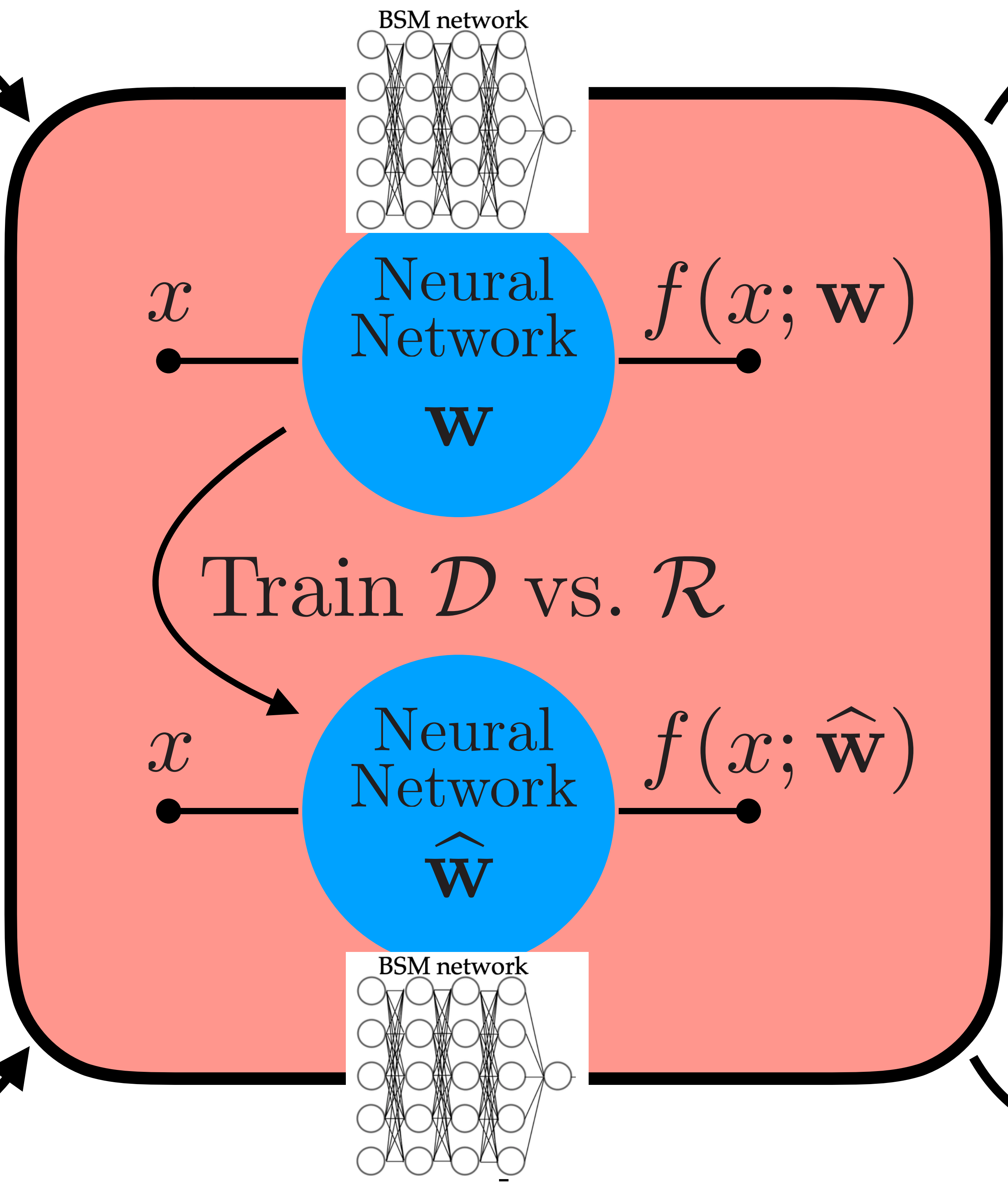
INPUT

Data sample $\mathcal{D}$



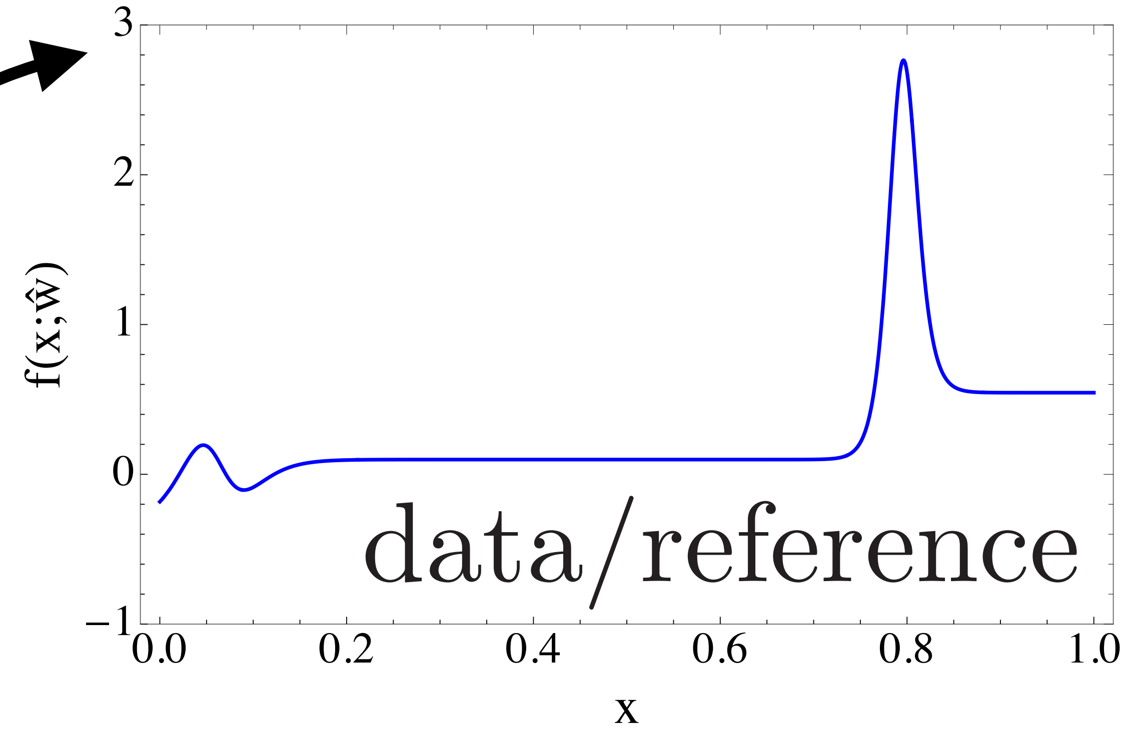
Reference sample $\mathcal{R}$





OUTPUT

Dist. log ratio



$$f(x;\widehat{\mathbf{w}}) \simeq \log\left[\frac{n(x|\mathrm{T})}{n(x|\mathrm{R})}\right]$$

Test statistic t computed on the data sample $\mathcal{D}$

$$t(\mathcal{D}) = -2\,\underset{\{\mathbf{w}\}}{\mathrm{Min}}\, L[f]$$

[D'Agnolo et al., arXiv:1806.02350](https://arxiv.org/abs/1806.02350)

[D'Agnolo et al., arXiv:1912.12155](https://arxiv.org/abs/1912.12155)

New Physics Learning Machine

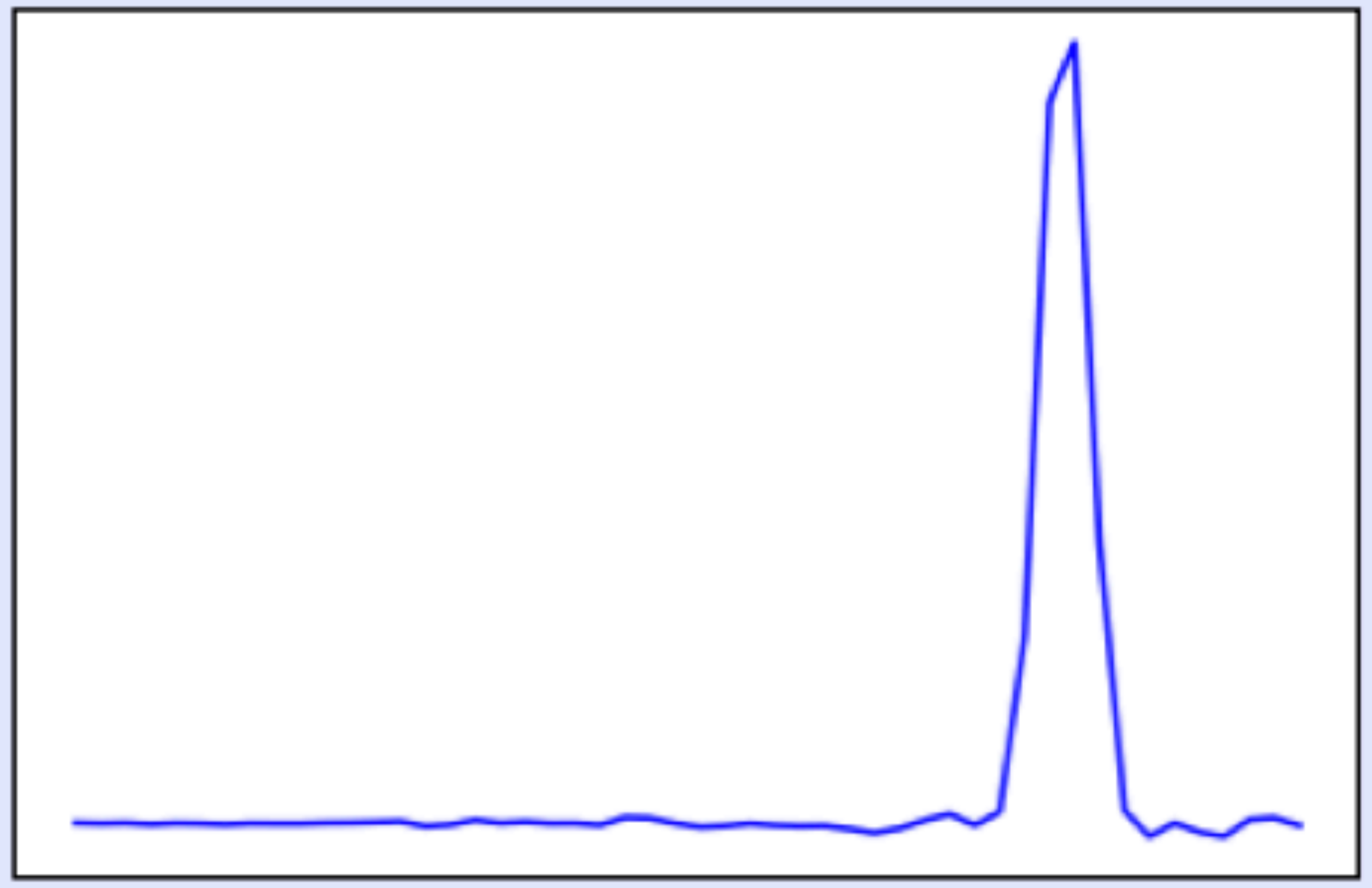
OUTPUT

Single training

$$t(D) = -2L\left[f(x;\hat{\mathbf{w}})\right]$$

$$f(x;\hat{\mathbf{w}}) = \log\left[\frac{n(x\,|\,\mathrm{H}_{\hat{\mathbf{w}}})}{n(x\,|\,\mathrm{R}_0)}\right]$$

$f(x;\hat{\mathbf{w}})$



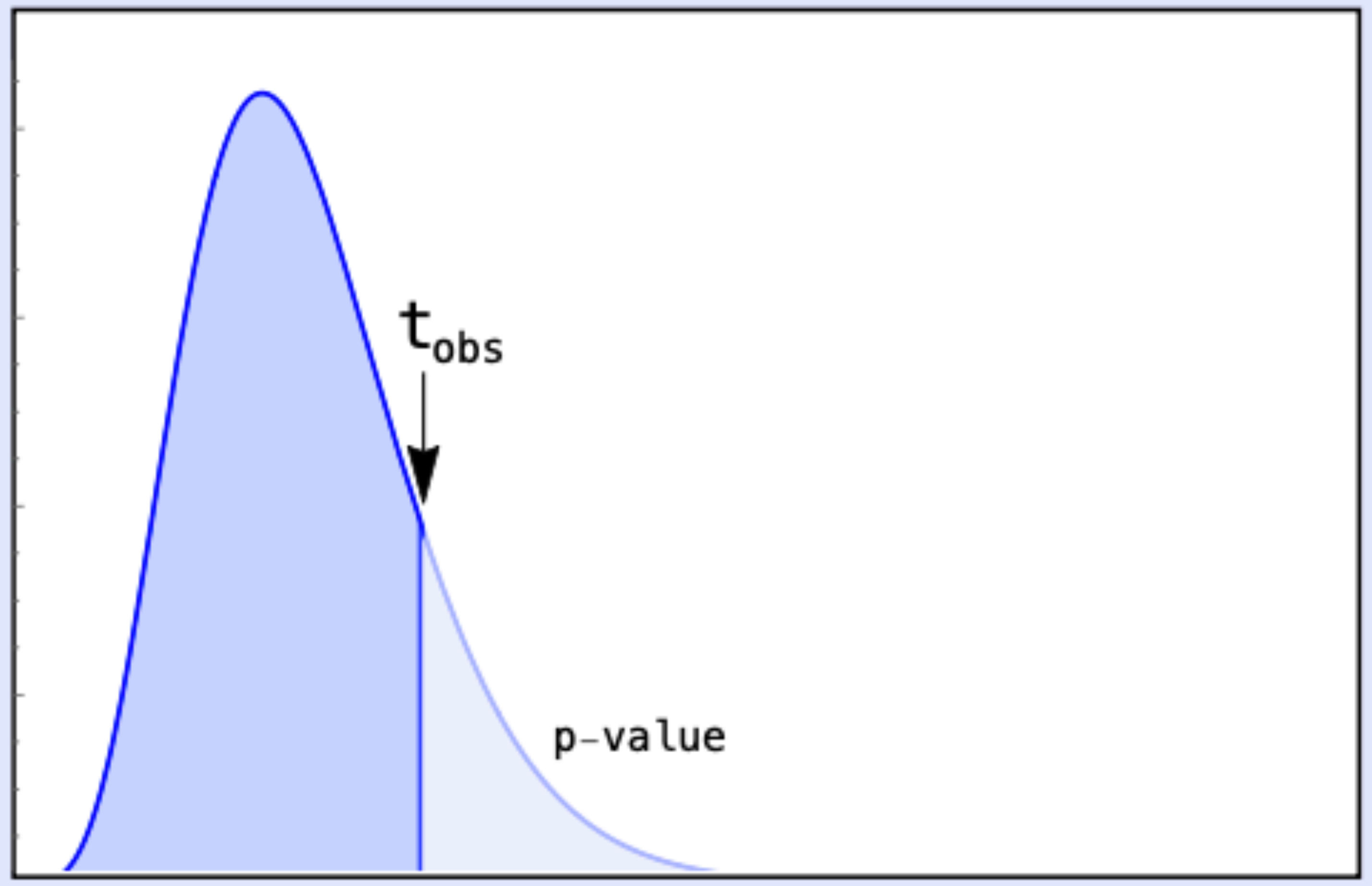
x

Many trainings (with pseudo-data)

Empirical distribution of t

→ p-value for new datasets

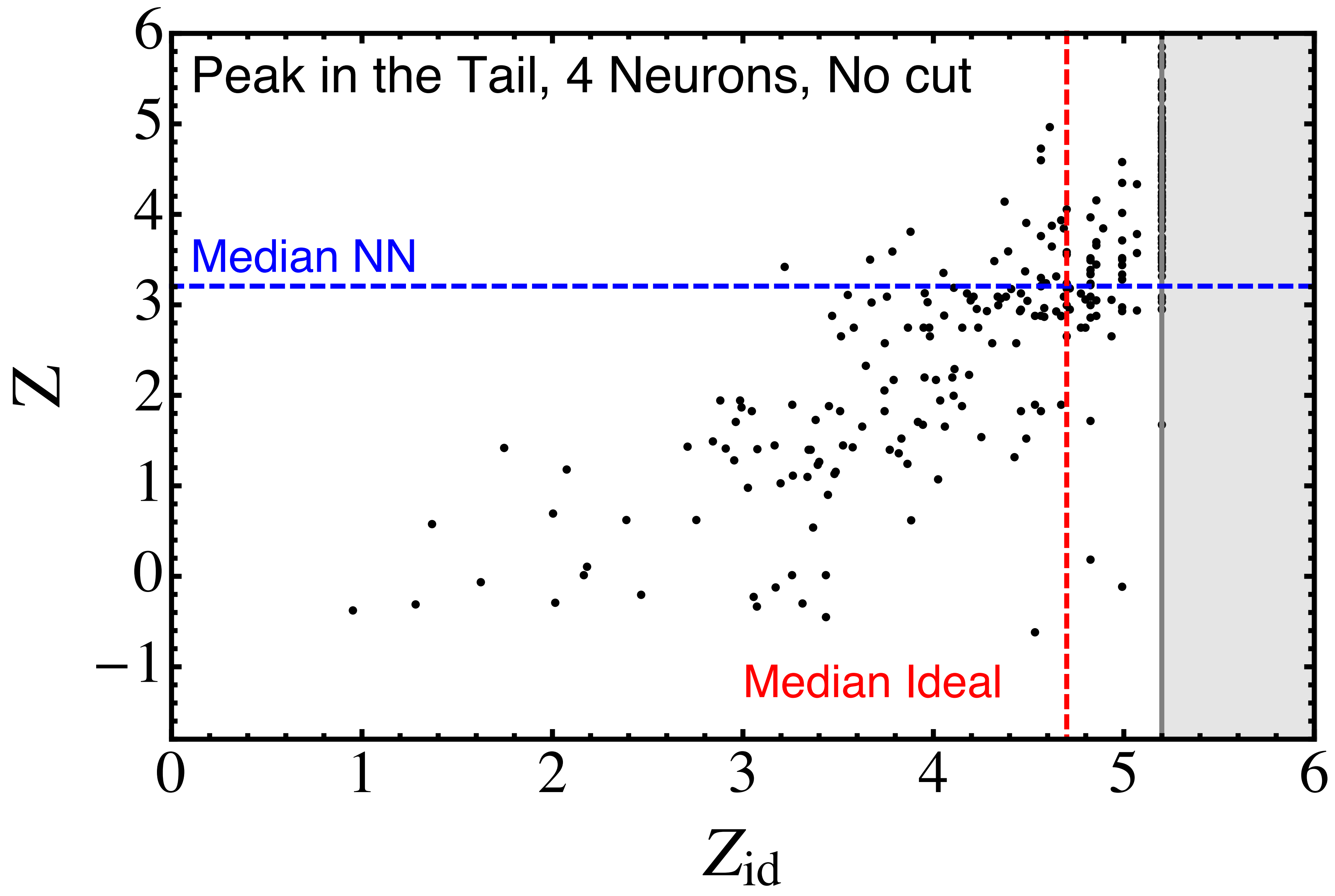
$P(\bar{t})$

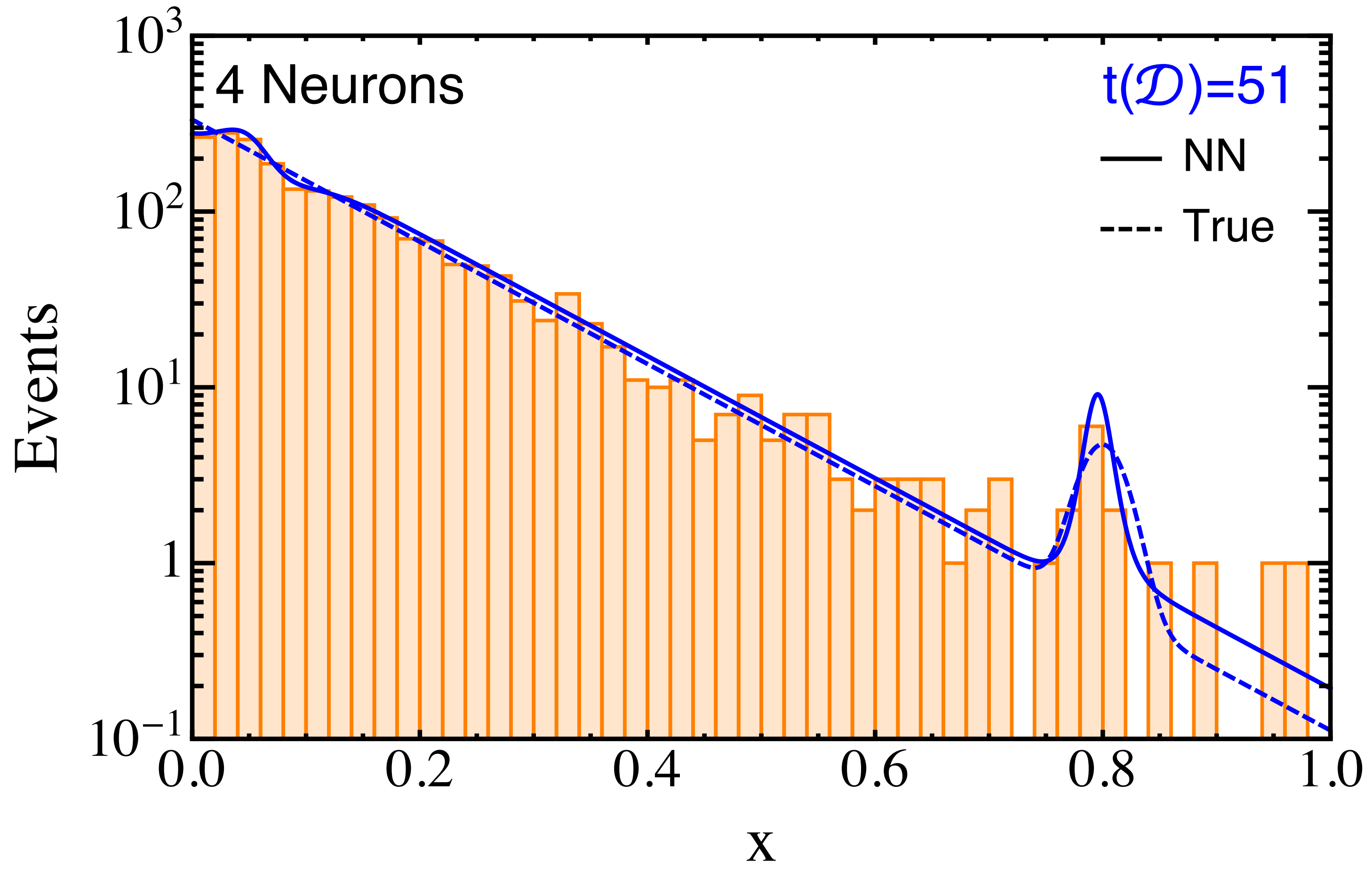


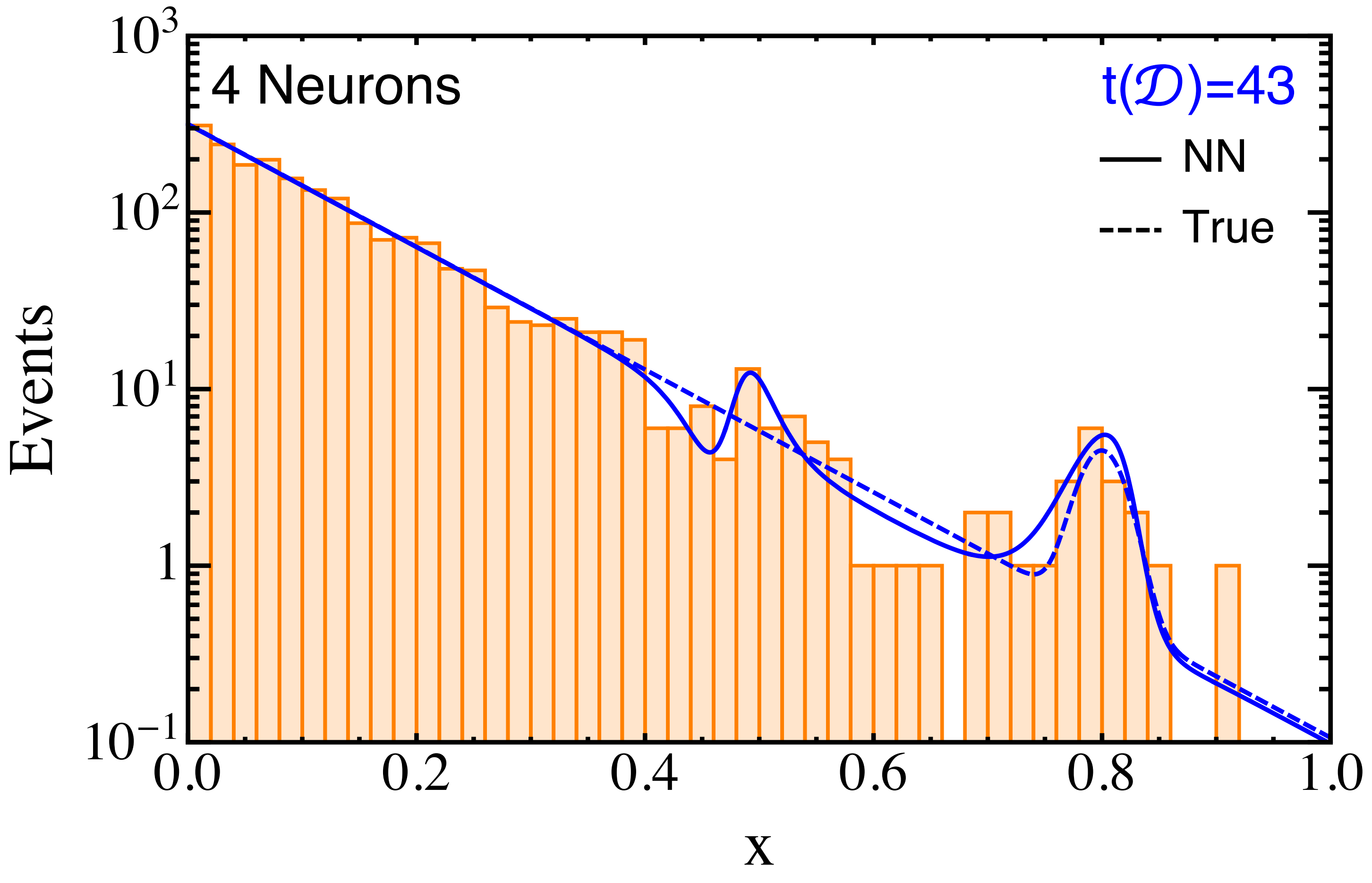
$\bar{t}$

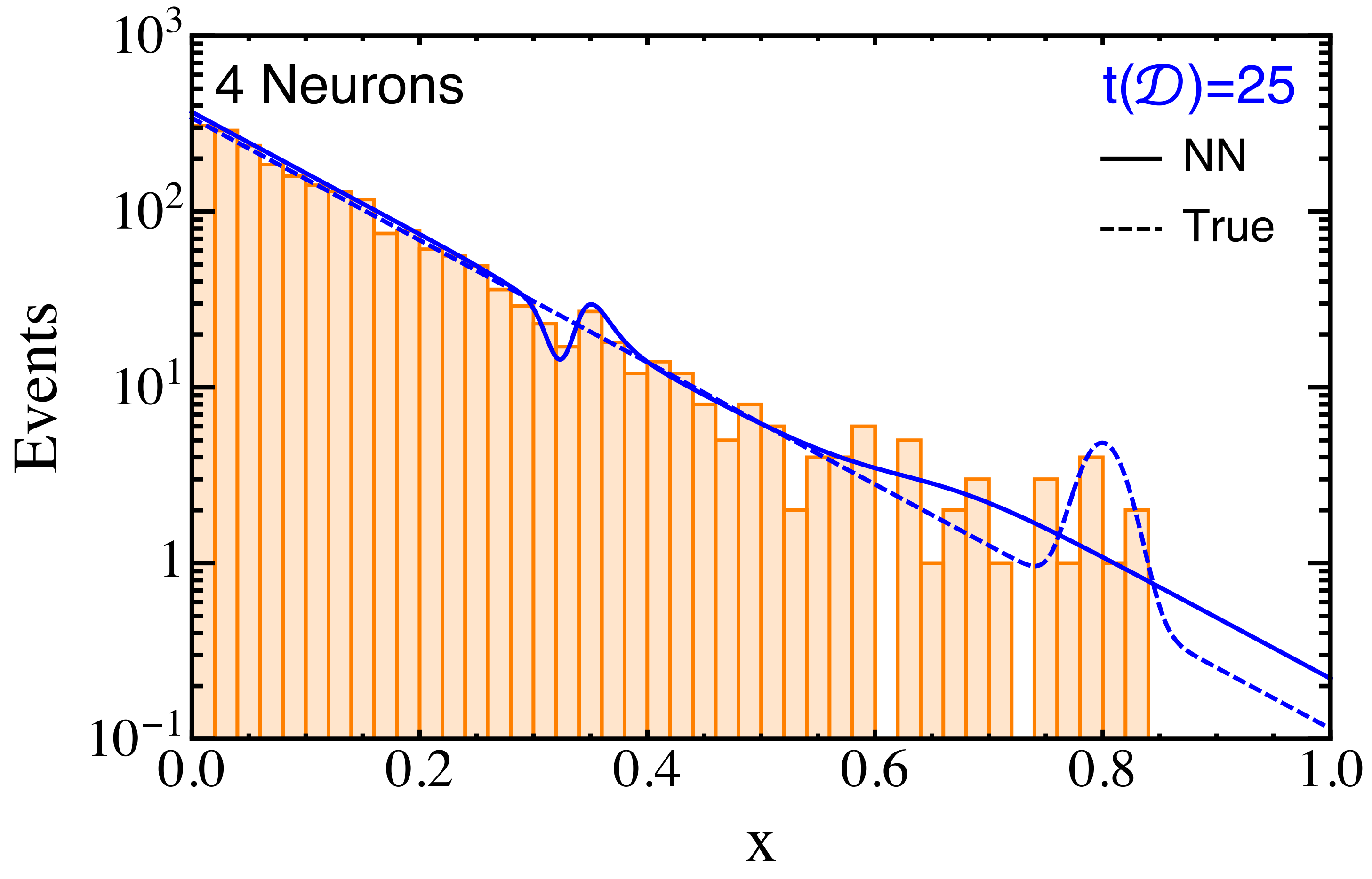
“Model-independent” hypothesis test

- In 1D, this method can detect new physics presence in D (but not in R)
  - performance reduced wrt fully-specified hypothesis test
  - still, sensitivity retained
  - no explicit assumption on signal shape



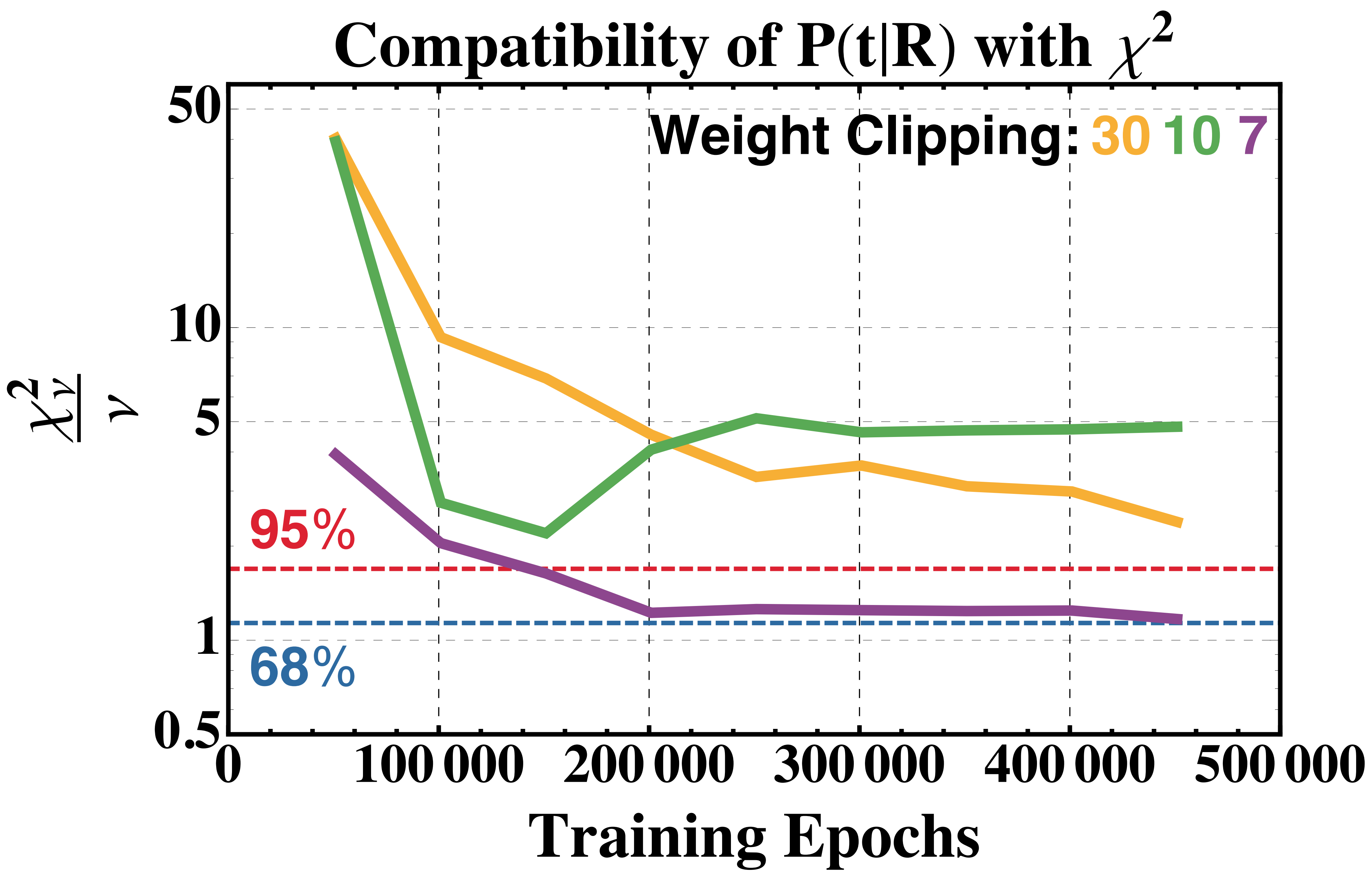


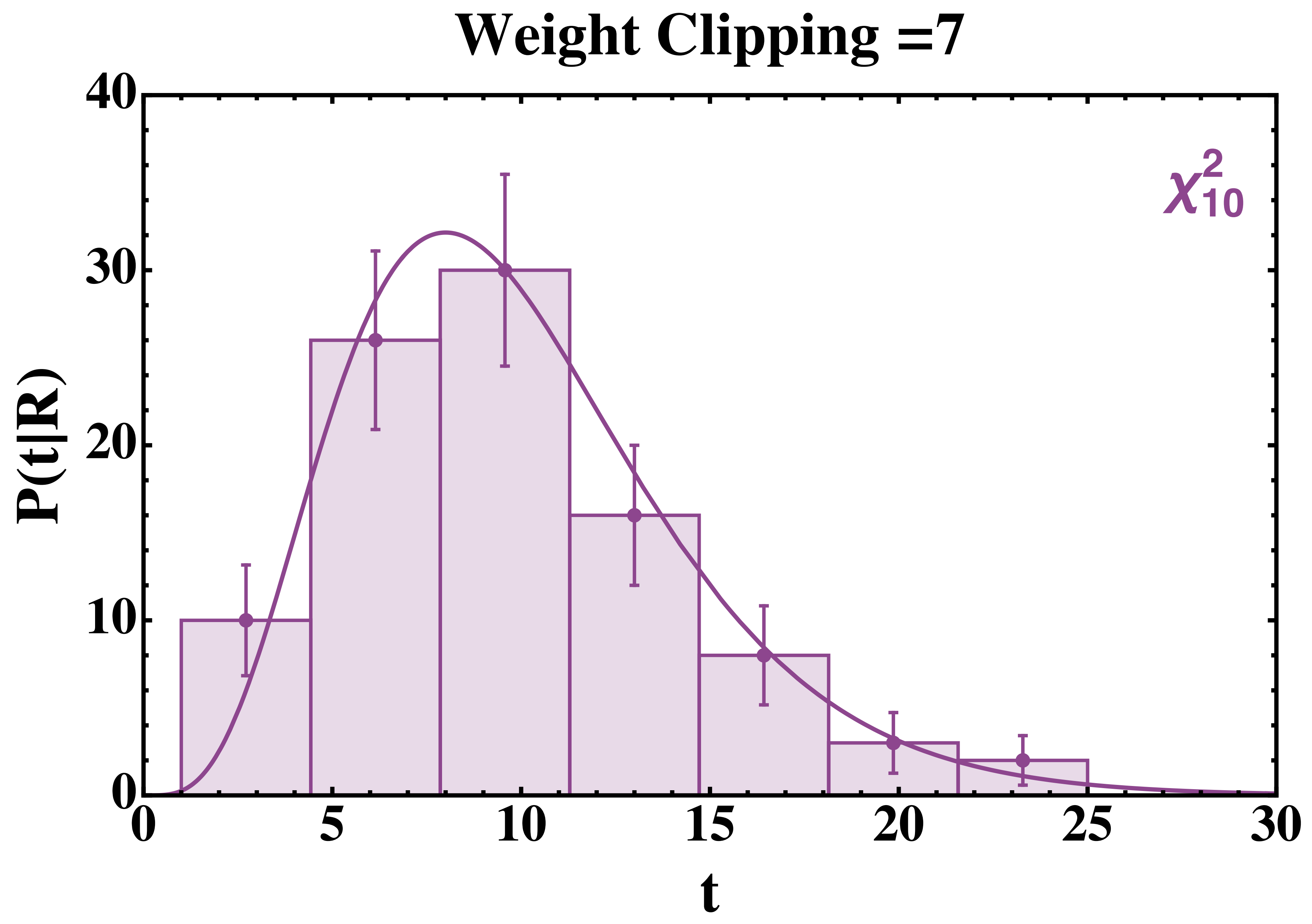




“Model-independent” hypothesis test

- The N-Dim generalization requires regularisation mechanism
  - weight clipping enforced to prevent over-fitting
  - with converge, test statistics recovers χ^2 distribution for standard events, with Ndof fixed by number of network parameters





D’Agnolo et al., arXiv:1806.02350

D’Agnolo et al., arXiv:1912.12155

“Model-independent” hypothesis test

- One would generate the expected distribution of the test statistics in absence of a signal, running the procedure on toy sets sampled from the reference-sample distribution (e.g., more MC samples)

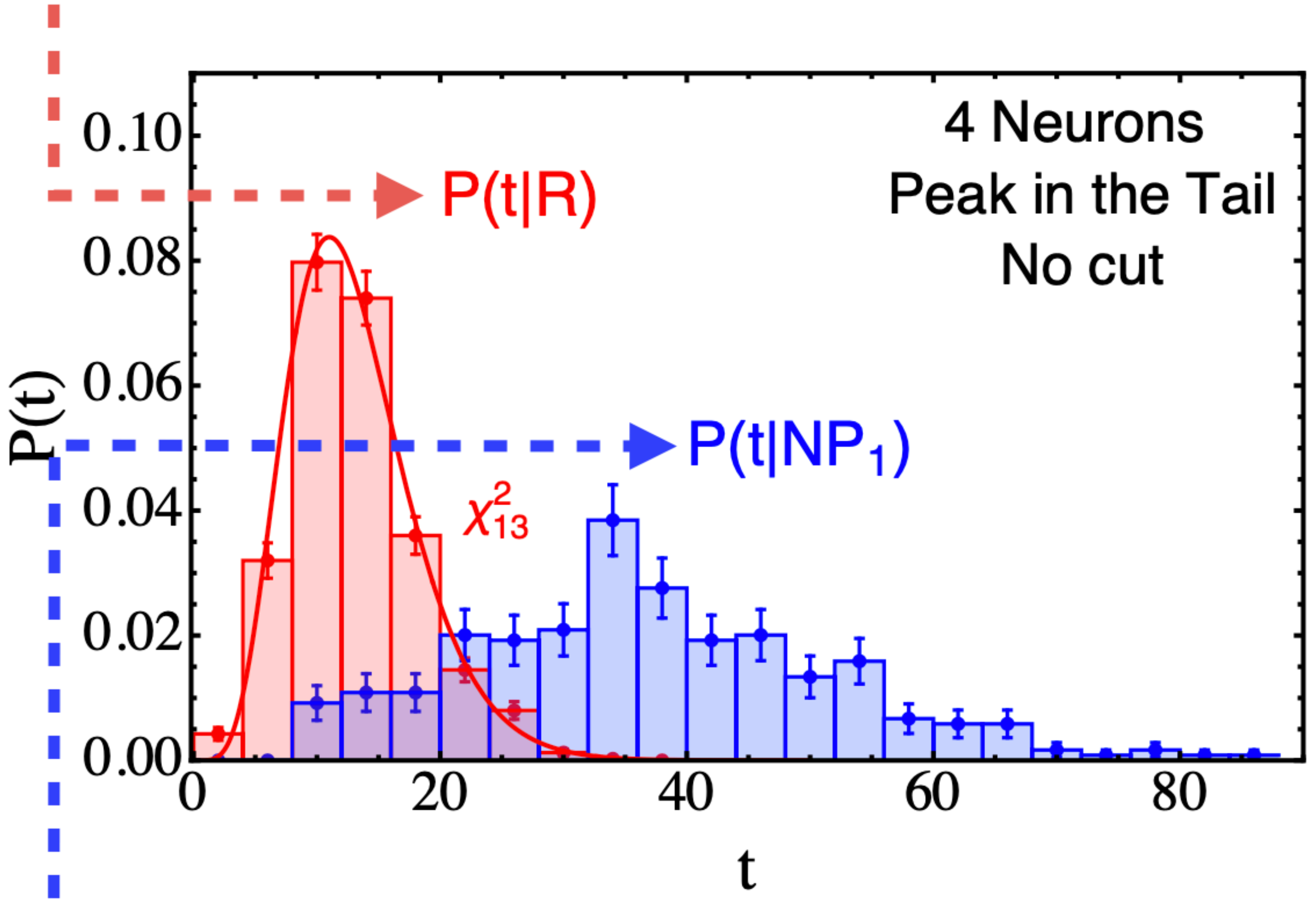
- [Wilks’ theorem] This distribution is ~ χ^2 (with dot given by the dot of the network)

- When applied to data, this distribution would give a value

- If the value is on the tail, one get a low p-value (large number of sigmas)

- For a given scenario, one can estimate the expected sensitivity looking at the distribution of the test statistics in sig+bkg toys

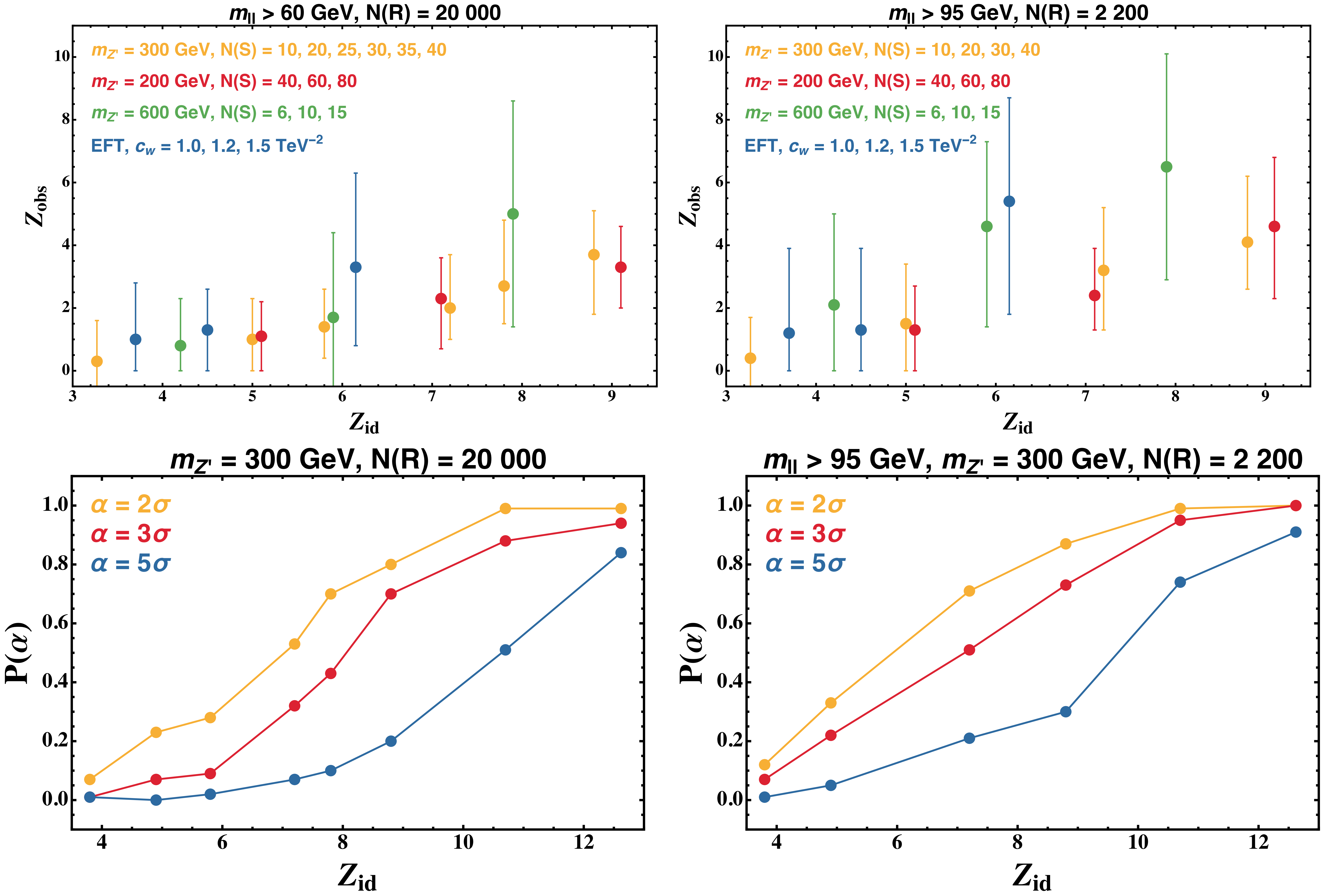
Distribution of the test statistic “t” in Reference Hypothesis



Distribution of “t” in one New Physics Model Hypothesis

t → p → Z-score (we use $Z = \Phi^{-1}(1-p)$)

“Model-independent” hypothesis test

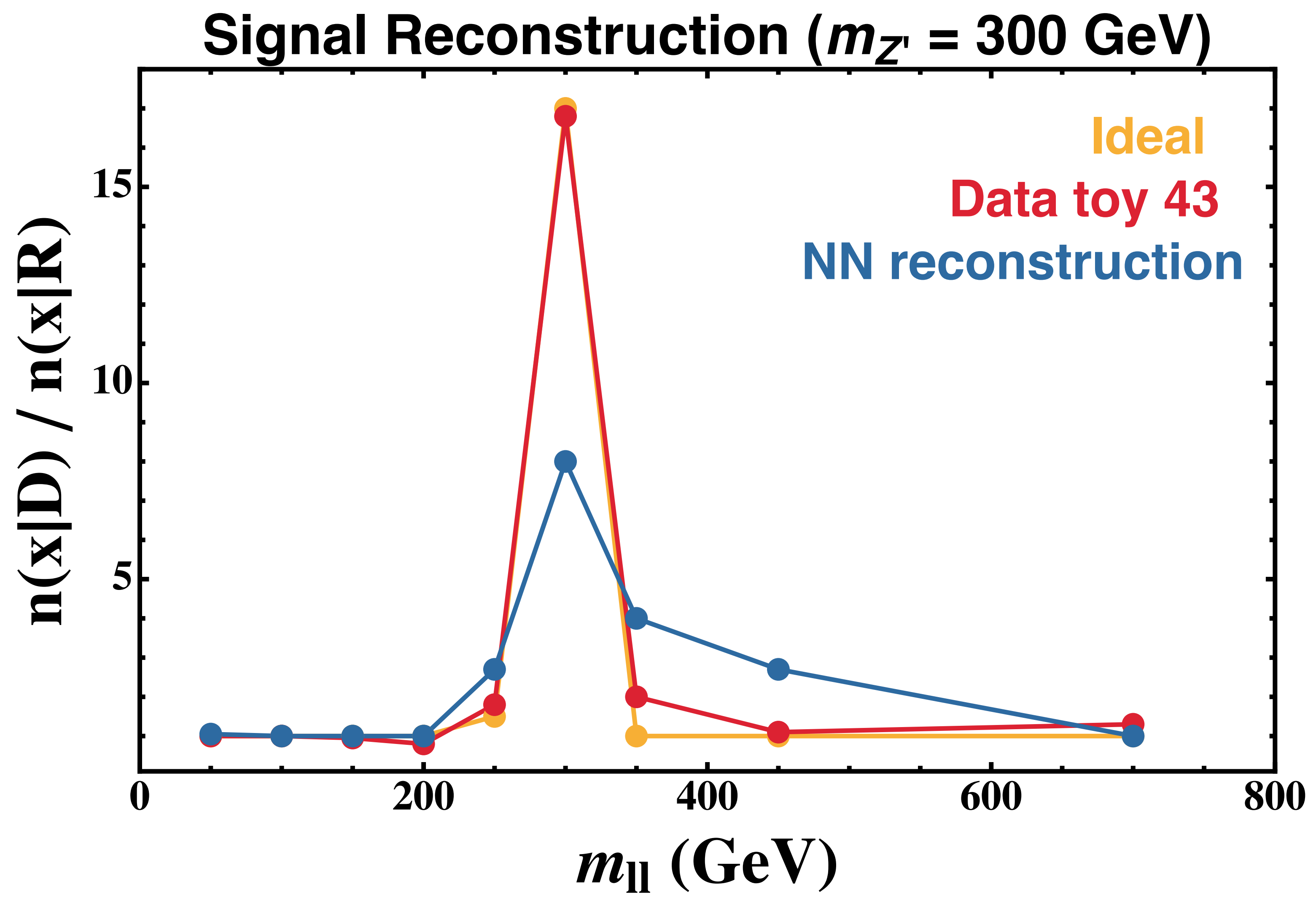


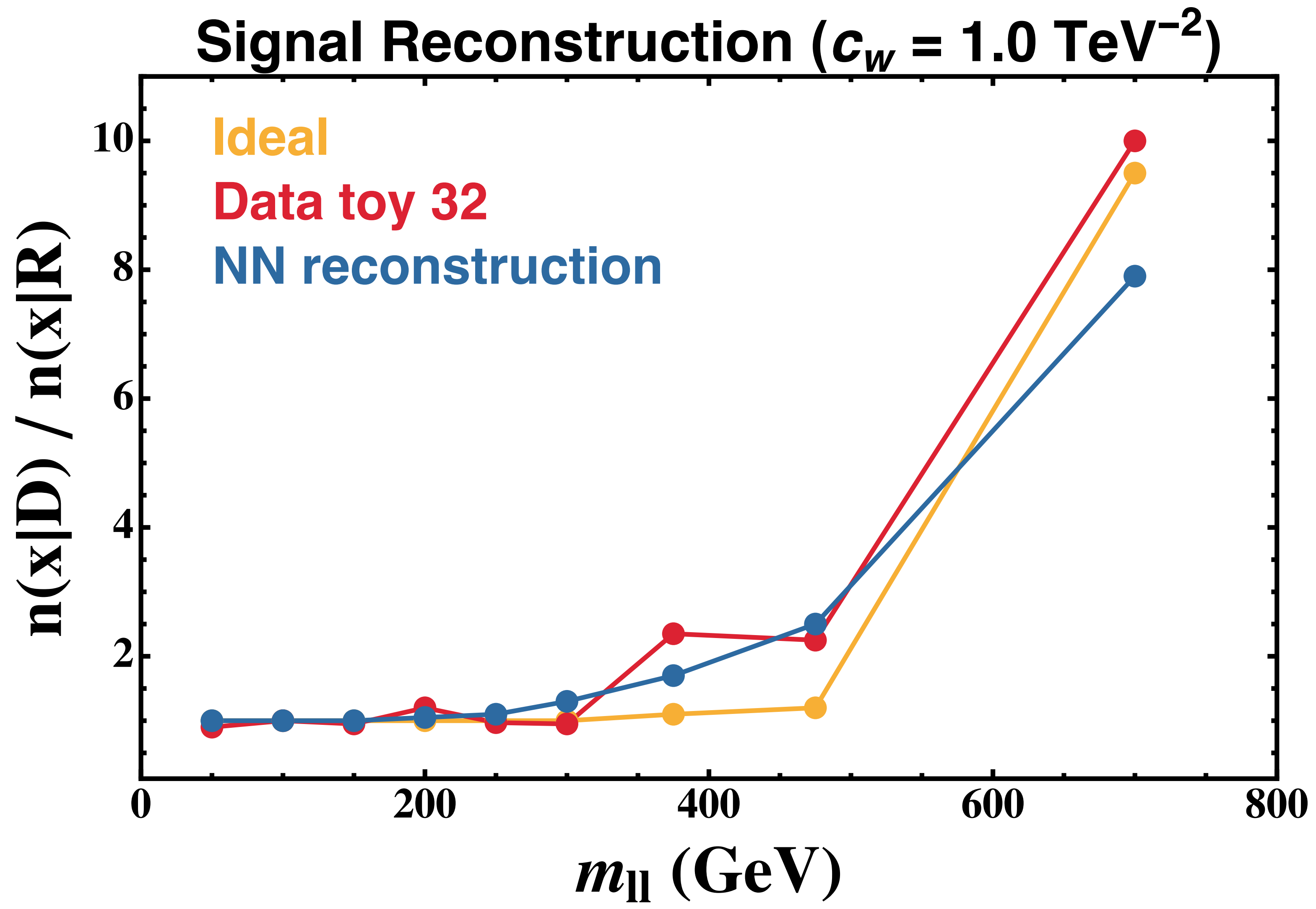
D’Agnolo et al., arXiv:1806.02350

D’Agnolo et al., arXiv:1912.12155

Characterizing the excess

- *A post-training analysis allows to characterize the nature of an excess that might have been found*
- *t(D) vs relevant quantities (not necessarily inputs to training) highlights clustering of signal events*
  - *Invariant mass peak for resonance signal*
  - *Tail excess for EFT signal*
- *The network is learning the nature of the underlying new physics and could guide its characterisation*



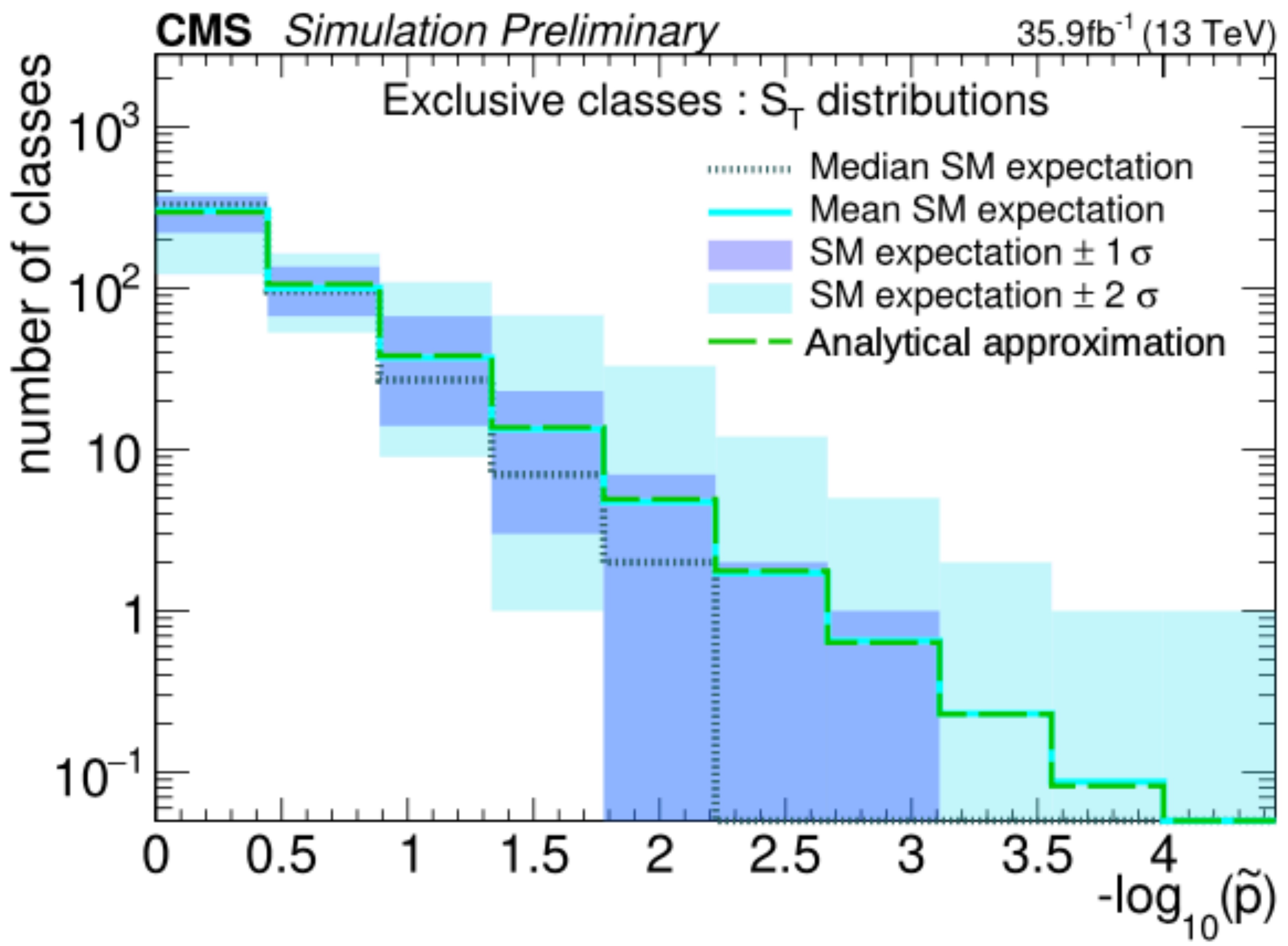


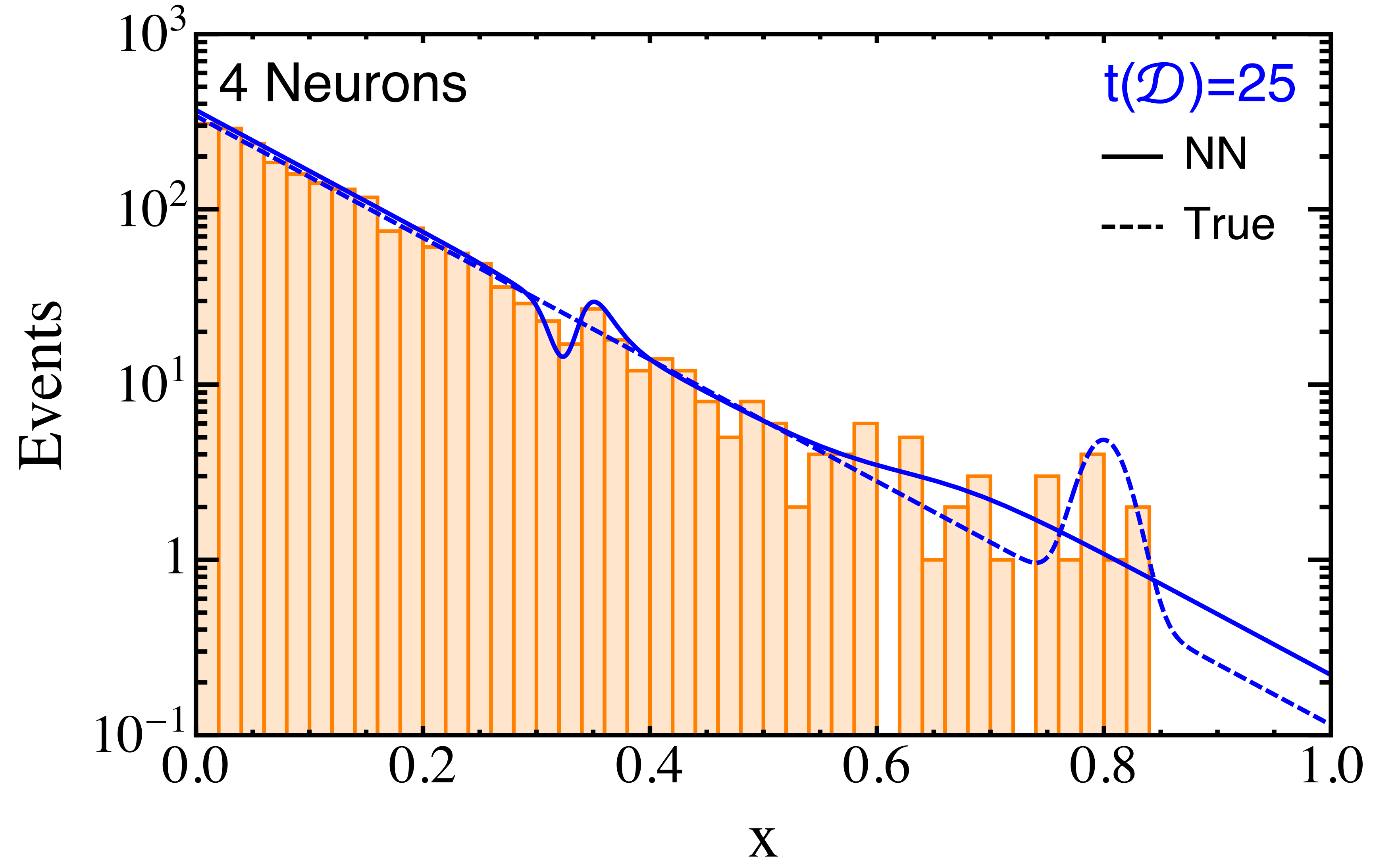
D'Agnolo et al., arXiv:1806.02350

D'Agnolo et al., arXiv:1912.12155

A goodness of Fit Test

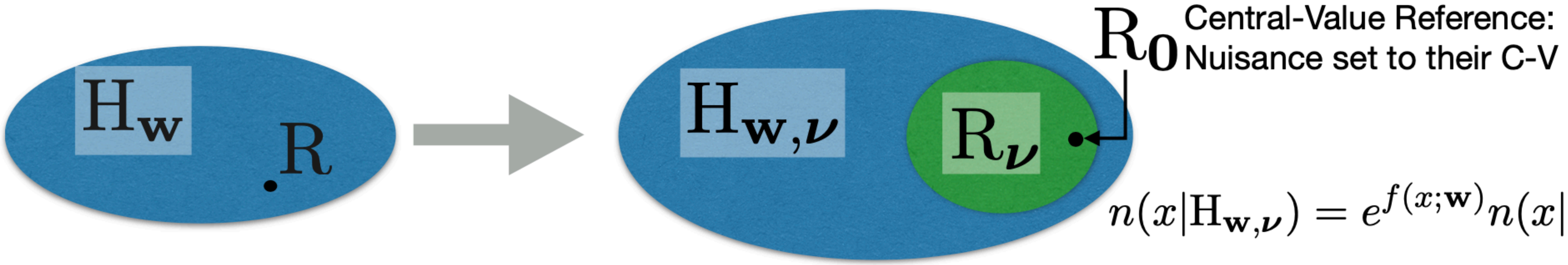
- *What we are doing is not really a hypothesis testing*
  - *NNs can be very expressive so H_1 is loosely defined*
- *In rigorous terms, NPLM is a goodness-of-fit test*
  - *We are given a dataset D and a model (the SM)*
  - *We want to test the compatibility between the two*
- *The setting is similar to that of the physics-inspired model-independent searches*
  - *But the approach has many differences (e.g., binned vs unbinned) at is in general more poewrwful*
  - *Also, it can account for systematic uncertainties (in a few slides)*





Imperfect Machine

- *The presence of systematic uncertainties would introduce anomalies in the data wrt reference sample*
  - *One could make false discovery claims (type-2 error)*
- *But the method can be generalized to include systematic uncertainties*
  - *Data is allowed to deviate from the reference in ways that are described by nuisance parameters*
  - *Deviations of different kind will not be accommodated: discovery potential retained*



$$n(x|\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}}) = e^{f(x;\mathbf{w})} n(x|\mathrm{R}_{\boldsymbol{\nu}})$$

Imperfect Machine

- *The new test statistics depends on the nuisance parameter*

$$t(\mathcal{D},\mathcal{A}) = 2\,\log \frac{\max\limits_{\mathbf{w},\boldsymbol{\nu}}\left[\mathcal{L}(\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}}|\mathcal{D})\cdot\mathcal{L}(\boldsymbol{\nu}|\mathcal{A})\right]}{\max\limits_{\boldsymbol{\nu}}\left[\mathcal{L}(\mathrm{R}_{\boldsymbol{\nu}}|\mathcal{D})\cdot\mathcal{L}(\boldsymbol{\nu}|\mathcal{A})\right]}$$

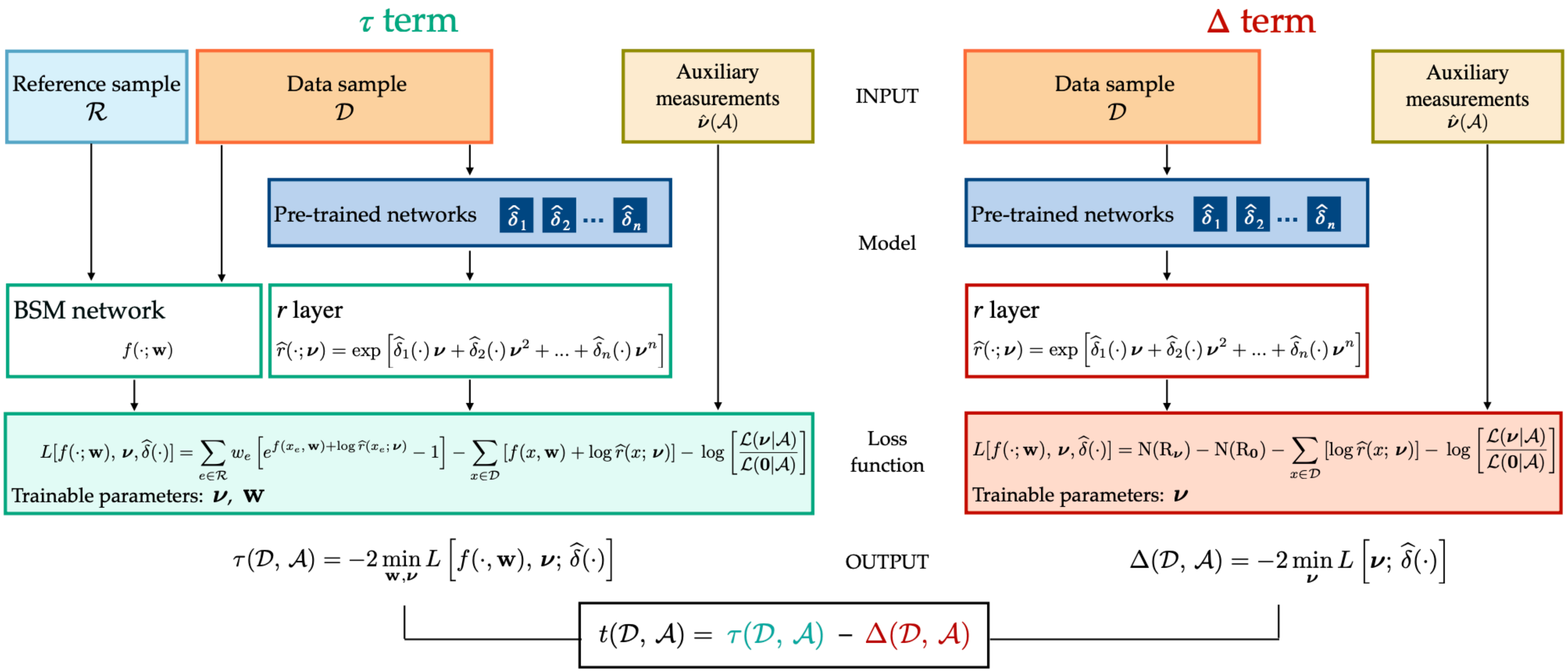
- *But it can be written as the sum of two terms*
  - *The previous one*
  - *A correction term, induced by the nuisance parameters*

$$= 2\,\max_{\mathbf{w},\boldsymbol{\nu}}\log\left[\frac{\mathcal{L}(\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}}|\mathcal{D})}{\mathcal{L}(\mathrm{R}_{\mathbf{0}}|\mathcal{D})}\cdot\frac{\mathcal{L}(\boldsymbol{\nu}|\mathcal{A})}{\mathcal{L}(\mathbf{0}|\mathcal{A})}\right]$$

$$-\,2\,\max_{\boldsymbol{\nu}}\log\left[\frac{\mathcal{L}(\mathrm{R}_{\boldsymbol{\nu}}|\mathcal{D})}{\mathcal{L}(\mathrm{R}_{\mathbf{0}}|\mathcal{D})}\cdot\frac{\mathcal{L}(\boldsymbol{\nu}|\mathcal{A})}{\mathcal{L}(\mathbf{0}|\mathcal{A})}\right]$$

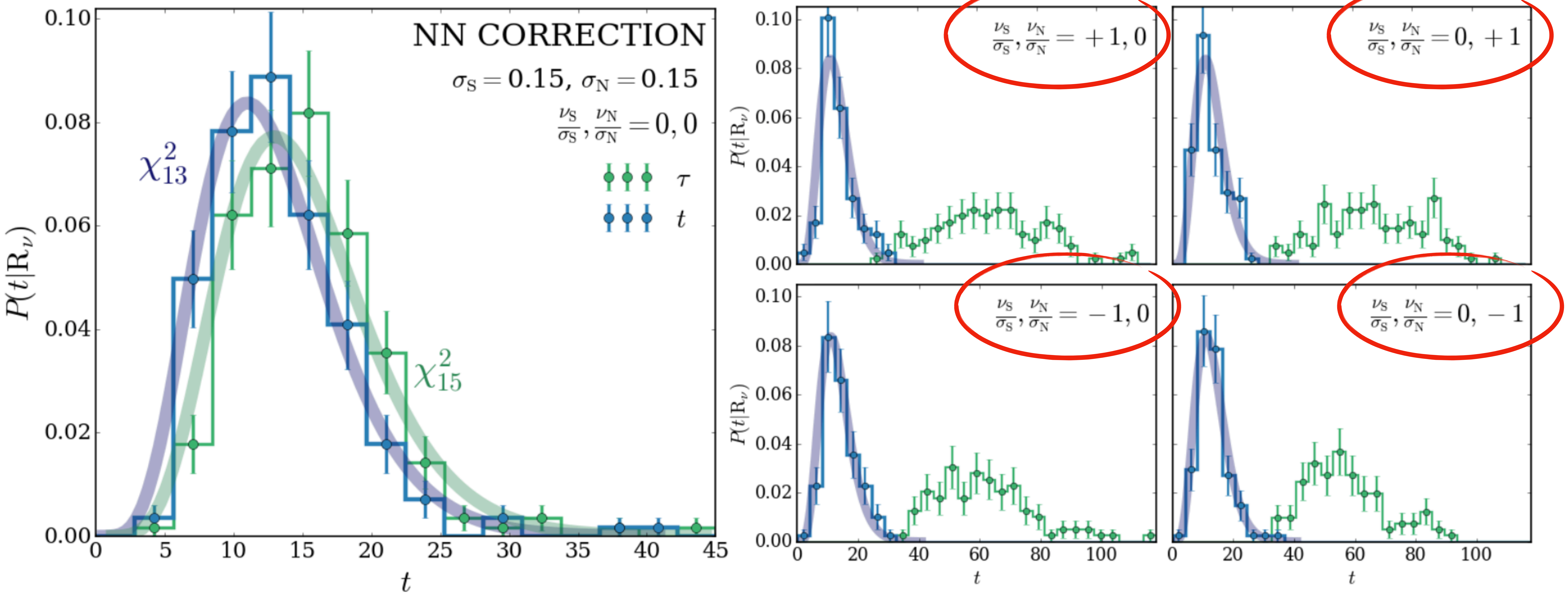
$$= \tau(\mathcal{D},\mathcal{A}) - \Delta(\mathcal{D},\mathcal{A})$$

Imperfect Machine



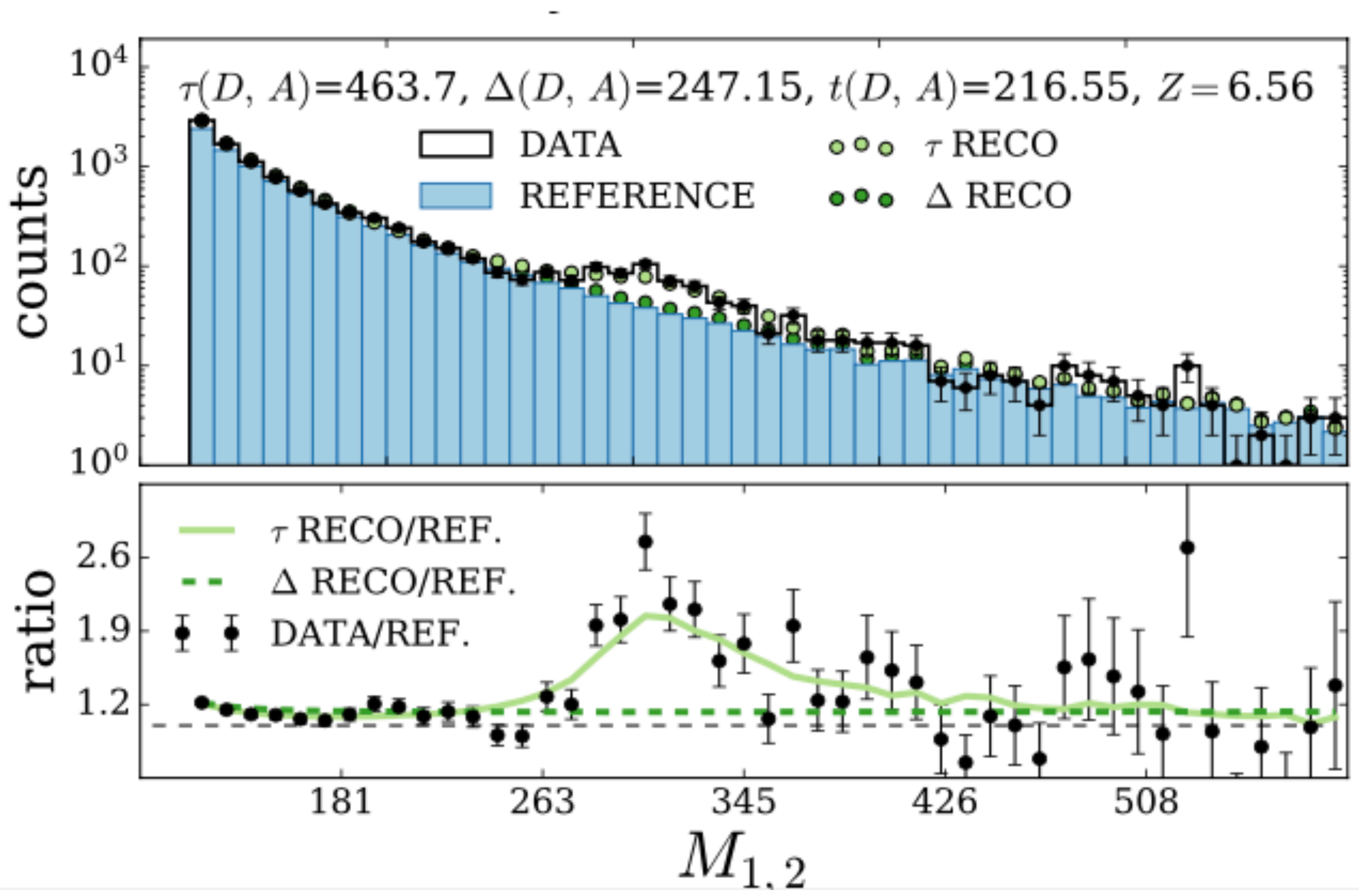
Imperfect Machine

- When nuisances are pulled away from 0 in toy generation
- The original τ distribution deviates from χ^2
- The Δ correction bring the distribution of $t = \tau-\Delta$ back to a χ^2



Anomaly Detection vs. NPLM

- *An anomaly detection technique*
  - *AD analysis (e.g., VAE) would be exploited as a selection to enrich a dataset of potential anomalies*
  - *But then one would typically run a normal fit to extract the dignal*
- *NPLM is instead an alternative fit strategy of a traditional analysis*
  - *Same signal selection as a supervised search*
  - *NPLM as a gof test, as an alternative to combine*
  - *Could be potentially performed by any traditional search, asa complementary/additional result*



Summary

- *Signal agnostic searches are a powerful tool to complement the typical LHC search strategy*
  - *Traditional techniques use binned histograms and bin-by-bin χ^2 test*
  - *Unsupervised/semisupervised techniques can be used to enrich the final fit sample of unspecified anomalies*
  - *NPLM can be used as a gof test to probe the presence of new physics in a data fit alternative to a traditional hypothesis testing*
- *In general, these approaches have less sensitivity on a specific scenario, but better performance in average across scenarios (generalization)-> complementarity to the traditional approaches*

References

- *Michael Kagan, CERN OpenLab classes on Machine Learning*
    - *Source of inspiration for this first lesson*
- *Pattern Recognition and Machine learning (Bishop)*
- *I. Goodfellow and Y. Bengio and A. Courville, "Deep Learning" MIT press*

- *Main reference for tutorial exercise: https://arxiv.org/abs/1908.05318*
- *All notebooks and classes are/will be on GitHub: https://github.com/pierinim/tutorials/tree/master/SMARTHEP*
- *Full dataset available at: https://zenodo.org/record/3602260*