

Simulating Reality & Searching for the Unknown

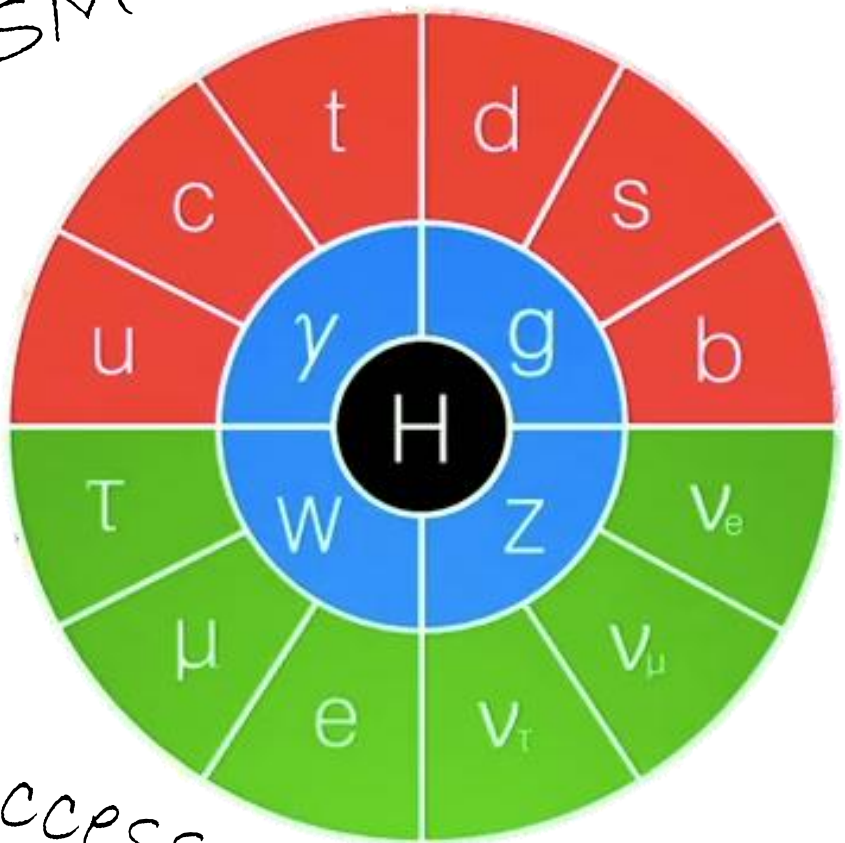
Tobias Golling,
University of Geneva

Outline

- Establish the goal: maximize LHC's sensitivity to new physics
- The need for accurate and fast background modeling
- Extend LHC's physics portfolio to model-agnostic searches
- How machine learning can help to overcome the challenges
 - Automate
 - Reduce complexity

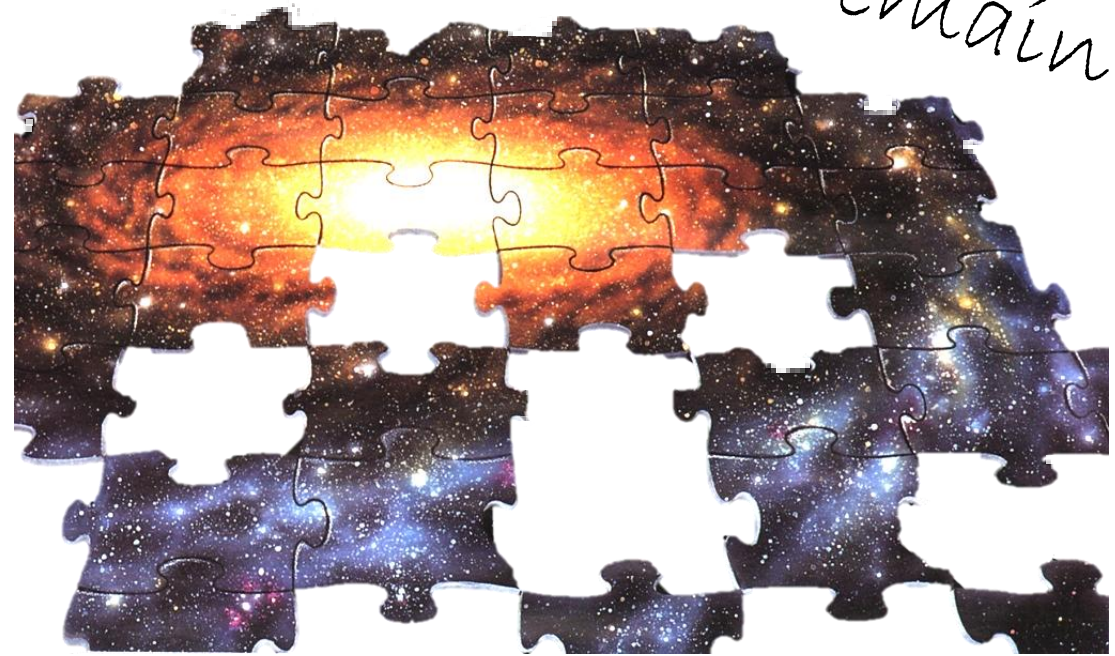
The current situation

The SM is complete



Success story!

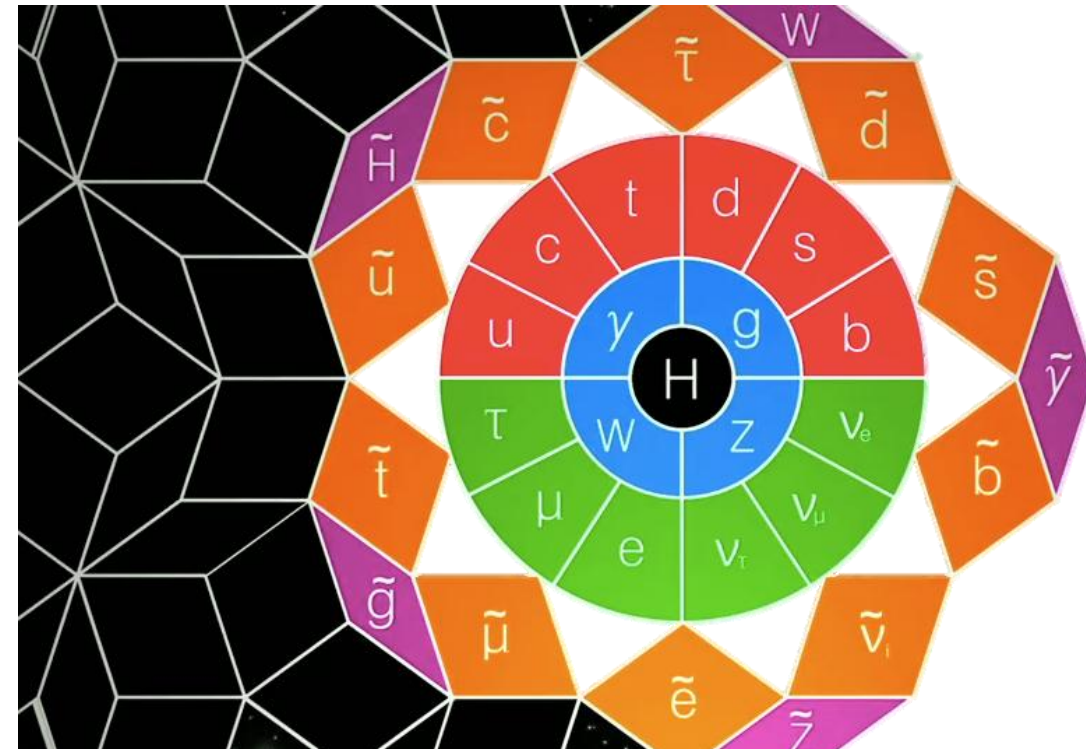
Open mysteries remain



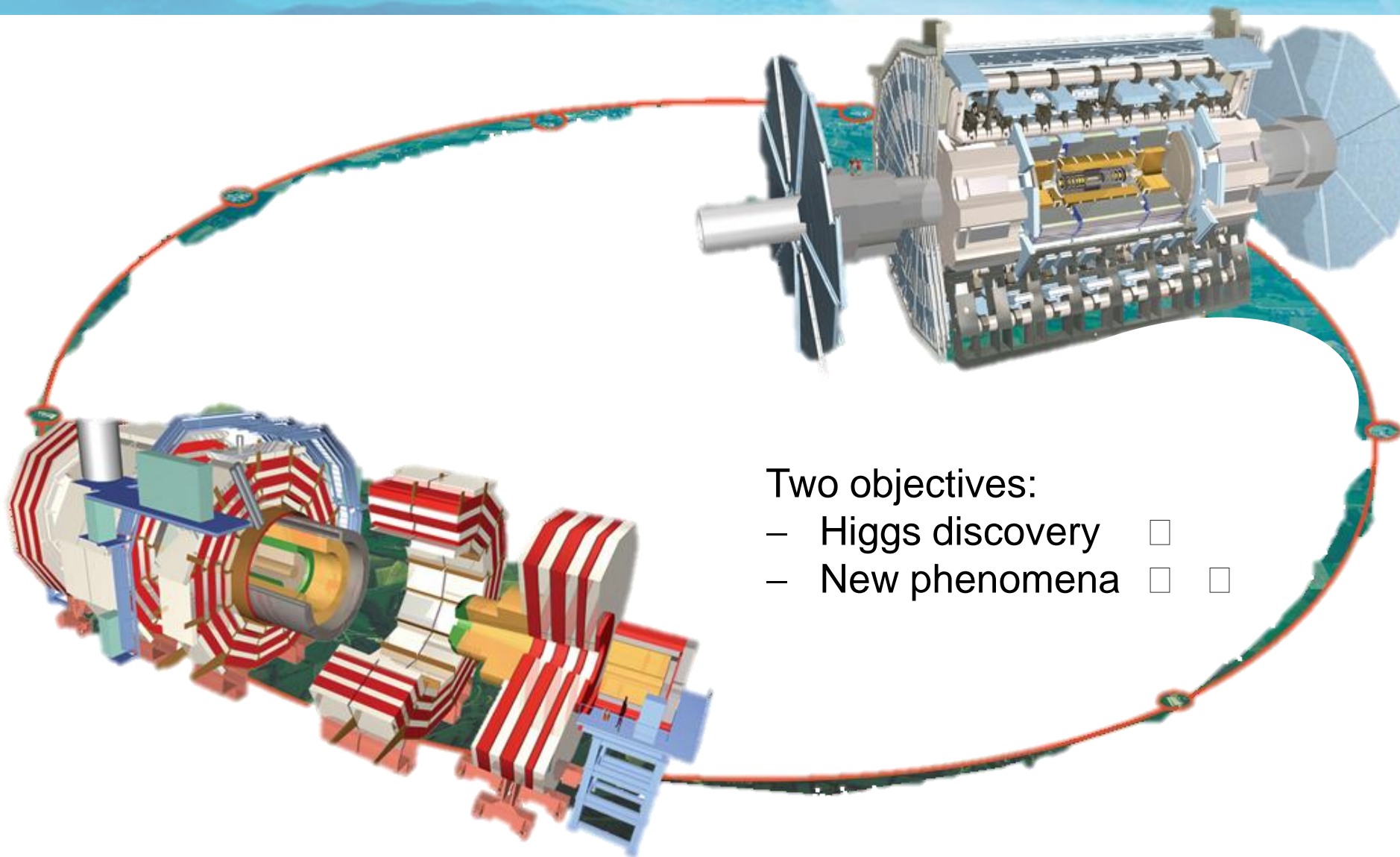
Dark matter, dark energy, quantum gravity,...

The theory guidance

- Hypothesize extensions of the SM
 - Addressing SM shortcomings
 - Leading to *testable* predictions
- Plethora of Beyond-the-SM extensions...



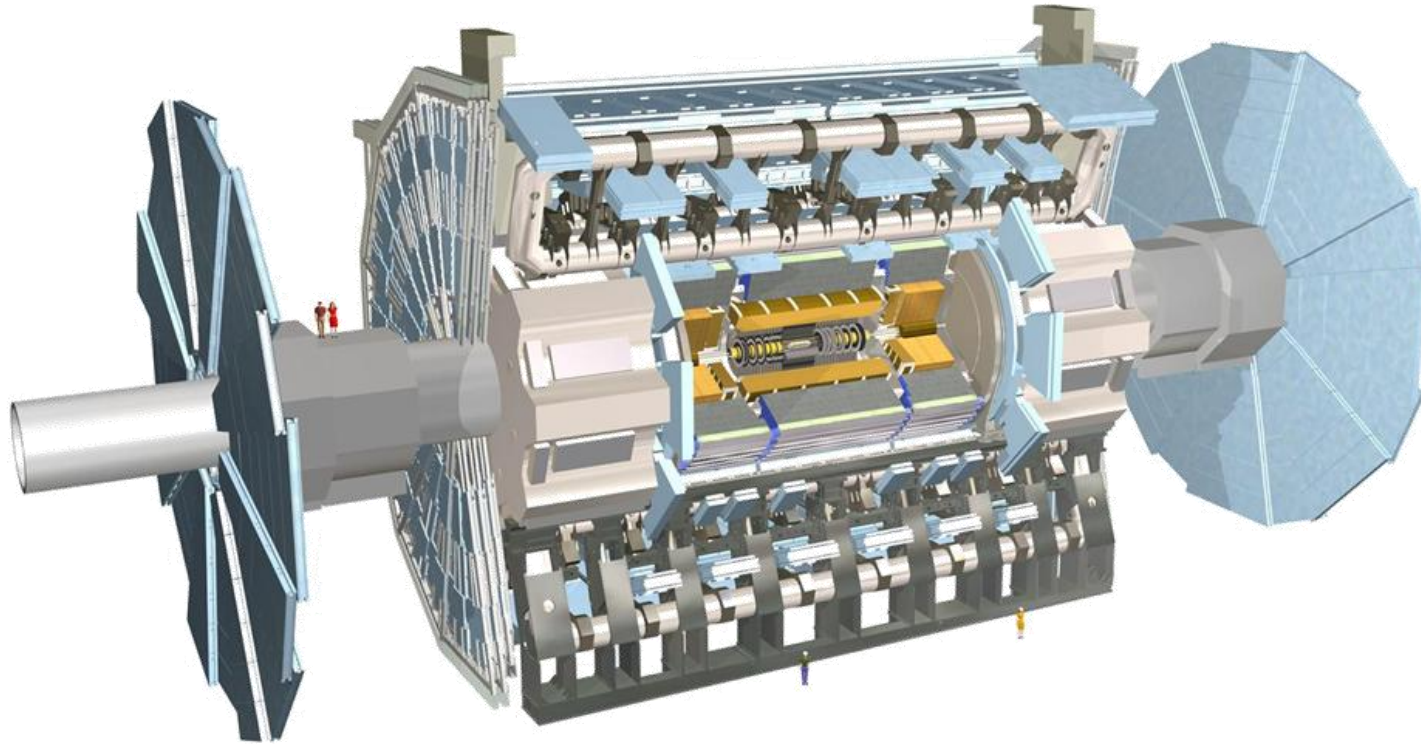
The Large Hadron Collider (LHC)



Two objectives:

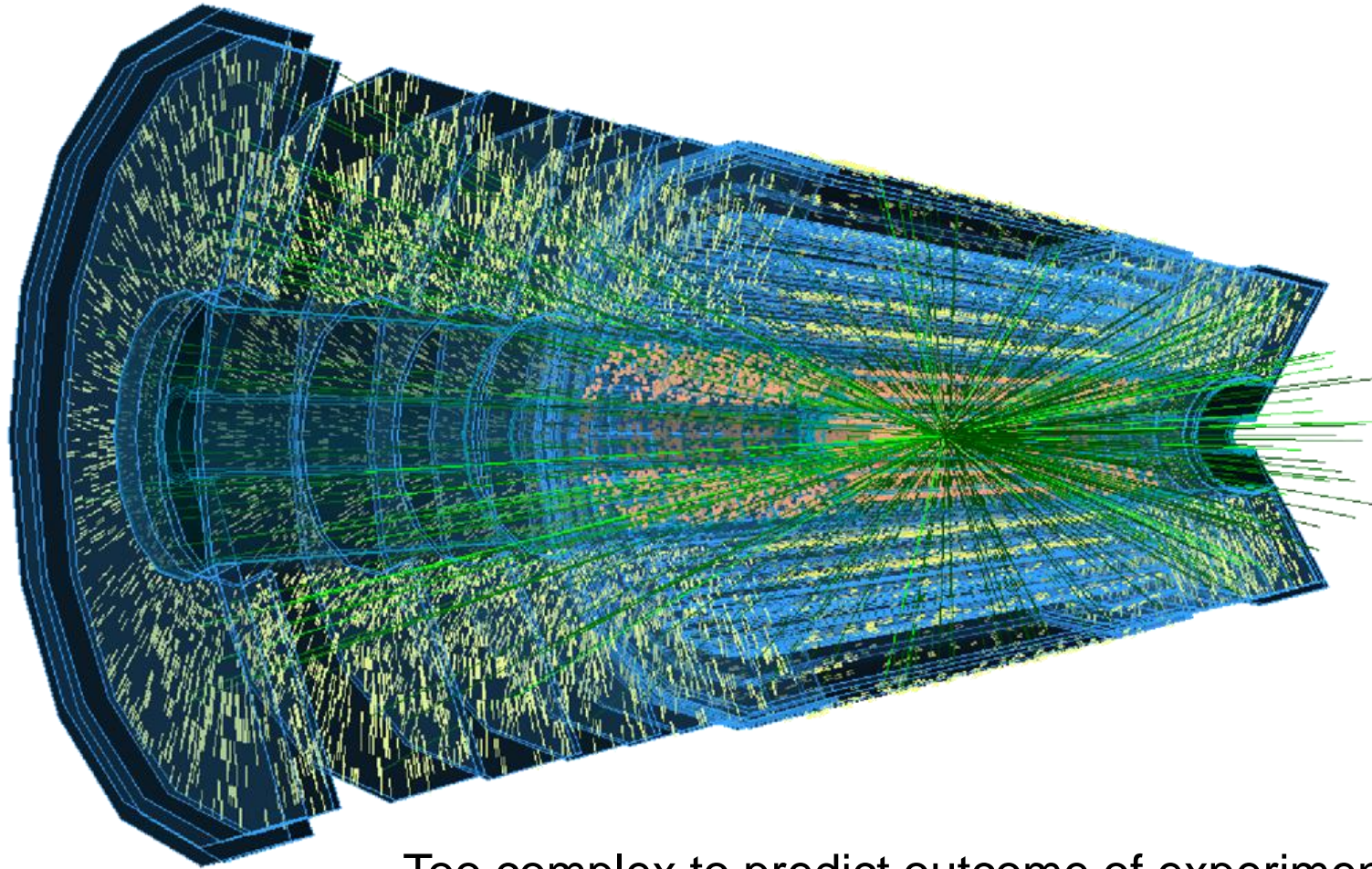
- Higgs discovery
- New phenomena

The ATLAS detector



- 40 MHz collision rate – online filter to record ~ 1 kHz
- Thousands of particles per collision
- 100M readout channels, $\sim 1\%$ occupancy
- Trillions of collisions in data & simulation – hundreds of petabytes

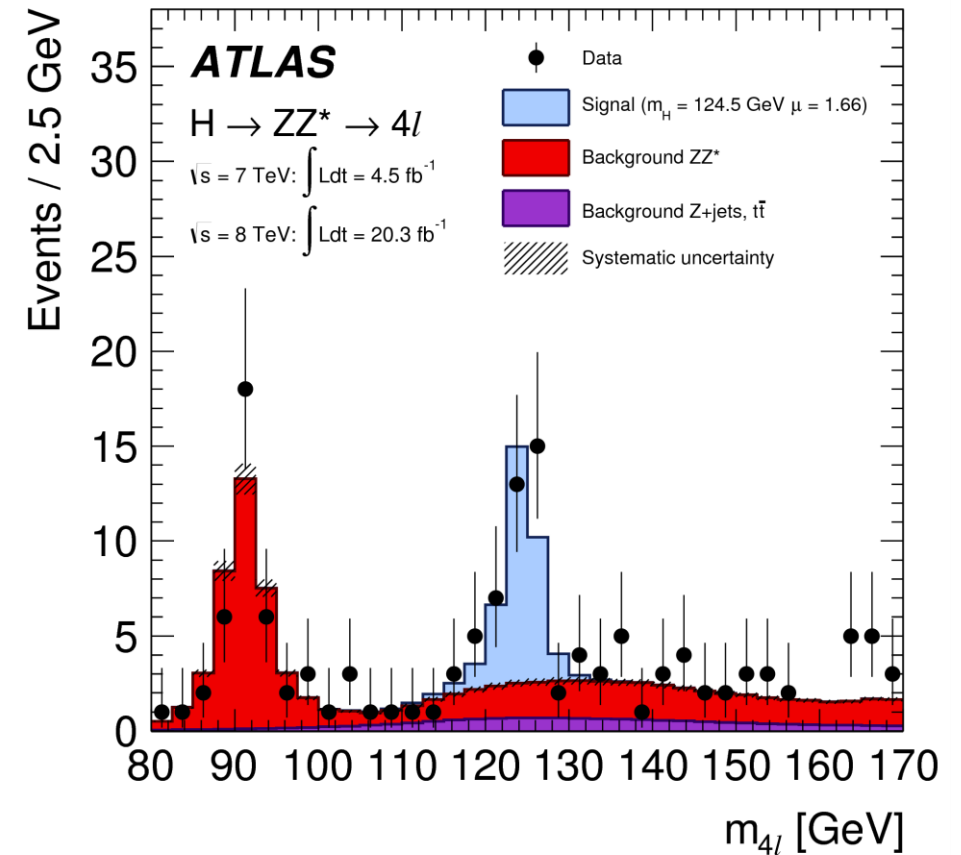
The need for synthetic data



Too complex to predict outcome of experiment from first principles
→ **Monte Carlo simulation**

The method of hypothesis testing

- Example: Higgs boson discovery:
 - H_0 : no Higgs
 - H_1 : null+Higgs
- Our standard inference approach:
 - Reduce input data $O(10^6)$ to $O(1)$ human-engineered feature
 - *Far from ideal*



Toolbox: what is ML good for?

Search for something *rare* in a *deluge* of data:

1. We *know* the signal (i.e. label) – **supervised ML**
2. We *do not know* the signal (no labels) – **unsupervised ML / anomaly detection**
 - i. Partial/noisy labels - **weakly-/semi-supervised ML**
3. High-fidelity and *high-speed* modeling – **generative ML**

- Use *Deep Neural Networks* to make the best out of the data we have

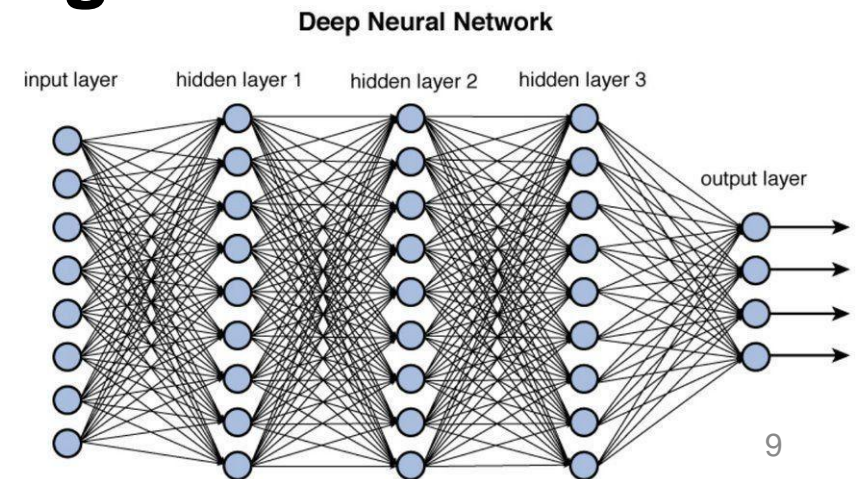


Figure 12.2 Deep network architecture with multiple layers.

Analogy: searching the needle in the hay



1. Searching for the **known**

- Take theory guidance at face value
 - We **know** how a needle & hay *look like*
- **Supervised** approach to fully exploit this knowledge



The *blemish*: No sign of physics Beyond the SM

ATLAS Exotics Searches* - 95% CL Upper Exclusion Limits

Status: July 2018

ATLAS Preliminary

$\int \mathcal{L} dt = (3.2 - 79.8) \text{ fb}^{-1}$ $\sqrt{s} = 8, 13 \text{ TeV}$

Model	ℓ, γ	Jets †	E_{T}^{miss}	$\int \mathcal{L} dt [\text{fb}^{-1}]$	Limit	Reference	
Extra dimensions	ADD $G_{KK} + g/g$	$0, e, \mu$	1-4 j	Yes	36.1	M_{Pl} 7.7 TeV	$n = 2$ 1711.03301
	ADD non-resonant $\gamma\gamma$	2 γ	-	-	36.7	M_{Pl} 8.6 TeV	$n = 3$ HLZ NLO 1707.04147
	ADD QBH	-	2 j	-	37.0	M_{Pl} 8.9 TeV	$n = 6$ 1703.09217
	ADD BH high Σp_T	$\geq 1, e, \mu$	≥ 2 j	-	3.2	M_{Pl} 8.2 TeV	$n = 6, M_0 = 3 \text{ TeV}$, rot BH 1606.02265
	ADD BH multijet	-	≥ 3 j	-	3.6	M_{Pl} 9.55 TeV	$n = 6, M_0 = 3 \text{ TeV}$, rot BH 1512.02586
	RS1 $G_{KK} \rightarrow \gamma\gamma$	2 γ	-	-	36.7	G_{KK} mass 4.1 TeV	$k/M_{\text{Pl}} = 0.1$ 1707.04147
	Bulk RS $G_{KK} \rightarrow WW/ZZ$	multi-channel	-	-	36.1	G_{KK} mass 2.3 TeV	$k/M_{\text{Pl}} = 1.0$ CERN-EP-2018-179
	Bulk RS $E_{KK} \rightarrow t\bar{t}$	$1, e, \mu$	≥ 1 b, ≥ 1 J †	Yes	36.1	E_{KK} mass 3.8 TeV	$f/m = 15\%$ 1604.10823
	2UED / RPP	$1, e, \mu$	≥ 2 b, ≥ 3 j	Yes	36.1	E_{KK} mass 1.8 TeV	Tier (1,1), $2\mathcal{A}(A^{(1)} \rightarrow t\bar{t}) = 1$ 1803.09678
	Gauge bosons	SSM $Z' \rightarrow \ell\ell$	2 e, μ	-	-	36.1	Z' mass 4.5 TeV
SSM $Z' \rightarrow \tau\tau$		2 τ	-	-	36.1	Z' mass 2.42 TeV	1709.07242
Leptophobic $Z' \rightarrow b\bar{b}$		-	2 b	-	36.1	Z' mass 2.1 TeV	1805.09299
Leptophobic $Z' \rightarrow t\bar{t}$		$1, e, \mu$	≥ 1 b, ≥ 1 J †	Yes	36.1	Z' mass 3.0 TeV	$f/m = 1\%$ 1804.10823
SSM $W' \rightarrow \ell\nu$		$1, e, \mu$	-	Yes	79.8	W' mass 5.6 TeV	ATLAS-CONF-2018-017
SSM $W' \rightarrow \tau\nu$		1 τ	-	Yes	36.1	W' mass 3.7 TeV	1801.06992
HVT $V' \rightarrow WW \rightarrow qq\bar{q}\bar{q}$ model B		$0, e, \mu$	2 J	-	79.8	V' mass 4.15 TeV	$g_V = 3$ ATLAS-CONF-2018-016
HVT $V' \rightarrow WH/ZH$ model B		multi-channel	-	-	36.1	V' mass 2.93 TeV	1712.06518
LRSM $W_R' \rightarrow t\bar{b}$	multi-channel	-	-	36.1	W_R' mass 3.25 TeV	CERN-EP-2018-142	
CI	CI $qq\bar{q}\bar{q}$	-	2 j	-	37.0	A 21.8 TeV	1703.09217
	CI $\ell\ell q\bar{q}$	2 e, μ	-	-	36.1	A 40.0 TeV η_{LL}	1707.02424
	CI $t\bar{t}t\bar{t}$	$\geq 1, e, \mu$	≥ 1 b, ≥ 1 j	Yes	36.1	A 2.57 TeV	CERN-EP-2018-174
DM	Axial-vector mediator (Dirac DM)	$0, e, \mu$	1-4 j	Yes	36.1	m_{DM} 1.55 TeV	$g_e = 0.25, g_b = 1.0, m(\chi) = 1 \text{ GeV}$ 1711.03301
	Colored scalar mediator (Dirac DM)	$0, e, \mu$	1-4 j	Yes	36.1	m_{DM} 1.67 TeV	$g = 1.0, m(\chi) = 1 \text{ GeV}$ 1711.03301
	$VV_{\chi\chi}$ EFT (Dirac DM)	$0, e, \mu$	1 j, ≤ 1 j	Yes	3.2	M_{Pl} 700 GeV	$m(\chi) < 150 \text{ GeV}$ 1608.02372
LO	Scalar LQ 1 st gen	2 e	≥ 2 j	-	3.2	LQ mass 1.1 TeV	$\beta = 1$ 1605.06035
	Scalar LQ 2 nd gen	2 μ	≥ 2 j	-	3.2	LQ mass 1.05 TeV	$\beta = 1$ 1605.06035
	Scalar LQ 3 rd gen	$1, e, \mu$	≥ 1 b, ≥ 3 j	Yes	20.3	LQ mass 640 GeV	$\beta = 0$ 1508.04735
Heavy quarks	VLQ $TT \rightarrow Ht/Zt/Wb + X$	multi-channel	-	-	36.1	T mass 1.37 TeV	SU(2) doublet ATLAS-CONF-2018-XXX
	VLQ $BB \rightarrow Wt/Zb + X$	multi-channel	-	-	36.1	B mass 1.34 TeV	SU(2) doublet ATLAS-CONF-2018-XXX
	VLQ $T_{5,3} T_{6,3} \rightarrow Wt + X$	2(SS) $\geq 3, e, \mu$	≥ 1 b, ≥ 1 j	Yes	36.1	$T_{5,3}$ mass 1.64 TeV	$2(T_{5,3} \rightarrow Wt) = 1$ CERN-EP-2018-171
	VLQ $V \rightarrow Wb + X$	$1, e, \mu$	≥ 1 b, ≥ 1 j	Yes	3.2	V mass 1.44 TeV	$2(V \rightarrow Wb) = 1, c(VWb) = 1/\sqrt{2}$ ATLAS-CONF-2018-072
	VLQ $B \rightarrow Hb + X$	$0, e, \mu, 2, \gamma$	≥ 1 b, ≥ 1 j	Yes	79.8	B mass 1.21 TeV	$g_B = 0.5$ ATLAS-CONF-2018-XXX
VLQ $QQ \rightarrow WqWq$	$1, e, \mu$	≥ 4 j	Yes	20.3	Q mass 680 GeV	1509.04261	
Excited fermions	Excited quark $q^* \rightarrow qg$	-	2 j	-	37.0	q^* mass 6.0 TeV	only u' and d' , $A = m(q')$ 1703.09127
	Excited quark $q^* \rightarrow q\gamma$	1 γ	1 j	-	36.7	q^* mass 5.3 TeV	only u' and d' , $A = m(q')$ 1709.10440
	Excited quark $b^* \rightarrow bg$	-	1 b, 1 j	-	36.1	b^* mass 2.6 TeV	1805.09299
	Excited lepton ℓ^*	3 e, μ	-	-	20.3	ℓ^* mass 3.0 TeV	$A = 3.0 \text{ TeV}$ 1411.2921
	Excited lepton ν^*	3 e, μ, τ	-	-	20.3	ν^* mass 1.6 TeV	$A = 1.6 \text{ TeV}$ 1411.2921
Other	Type III Seesaw	$1, e, \mu$	≥ 2 j	Yes	79.8	Δ^0 mass 560 GeV	ATLAS-CONF-2018-020
	LRSM Majorana ν	2 e, μ	2 j	-	20.3	Δ^0 mass 2.0 TeV	$m(W_2) = 2.4 \text{ TeV}$, no mixing 1506.06020
	Higgs triplet $H^{\pm\pm} \rightarrow \ell\ell$	2,3,4 e, μ (SS)	-	-	36.1	$H^{\pm\pm}$ mass 870 GeV	DY production 1710.09748
	Higgs triplet $H^{\pm\pm} \rightarrow \ell\tau$	3 e, μ, τ	-	-	20.3	$H^{\pm\pm}$ mass 400 GeV	DY production, $2(H^{\pm\pm} \rightarrow \tau\tau) = 1$ 1411.2921
	Monopole (non-res prod)	$1, e, \mu$	1 b	Yes	20.3	spin-1 invisible particle mass 657 GeV	$\mu_{\text{DM}} = 0.2$ 1415.5464
	Multi-charged particles	-	-	-	20.3	multi-charged particle mass 795 GeV	DY production, $ \ell = 5e$ 1504.04168
	Magnetic monopoles	-	-	-	7.0	monopole mass 1.34 TeV	DY production, $ \ell = 1g_0$, spin 1/2 1509.08059

*Only a selection of the available mass limits on new states or phenomena is shown.

† Small-radius (large-radius) jets are denoted by the letter (J).

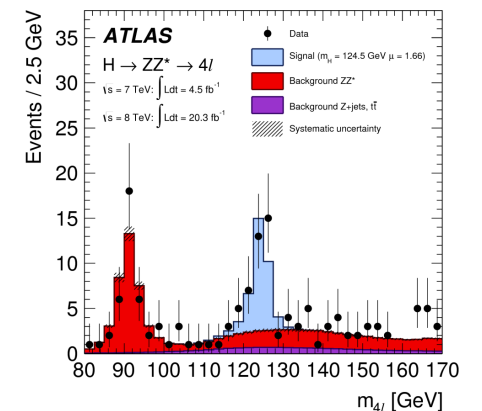
2. Searching for the **unknown**

- **Discard** theory guidance
 - Don't know **what** we're looking for in the hay
- **Unsupervised** approach to search for **structure** in the data

- Anomaly detection
 - **Outlier** *easy*: Not a needle but maybe a shiny object...
 - **Inlier/over-density** *much harder but closer to reality*: a tiny bit of *special* hay in a humongous haystack

Assumptions

- **Anomalies are rare** – otherwise we would have seen them already
 - No issues of *overlapping anomalies*
- **Anomalies are localized** – most prominent are resonances
 - Can define signal region (SR) with enhanced anomalous events
 - Control region (CR) depleted in anomalies
- **The data is smooth** – BG features vary slowly between SR & CR
 - Can use CR data to estimate BG in SR
- Only interested in statistical statement of group anomaly
 - Not trying to identify individual outliers



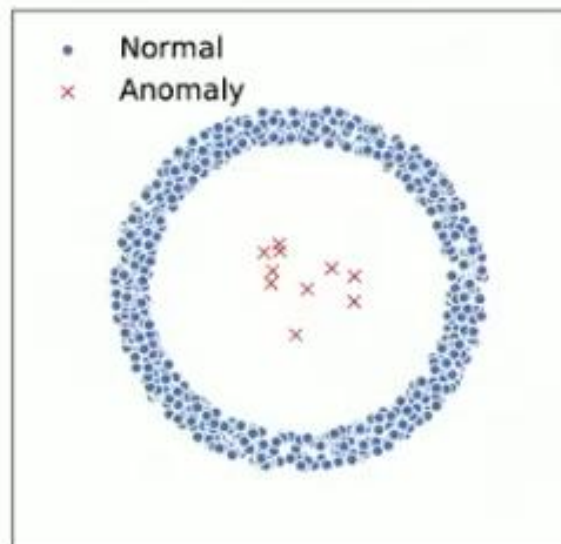
Analogy: searching for anomalies in the desert



- Grain of sand \triangleq LHC data collision
- What is an **outlier**
- What is an **inlier / over-density**

Example of an outlier

- *Anomalous monolith* in the desert
- Imagine each data point is a
 - *photo* of a grain of sand
 - equivalent *grain of monolith*
- *Grain of sand* easily separable from *grain of monolith*



[<https://www.vox.com/culture/22062796/monoliths-utah-california-romania>]

- Individual examples **not** anomalous
- **Anomalous collective behaviour**

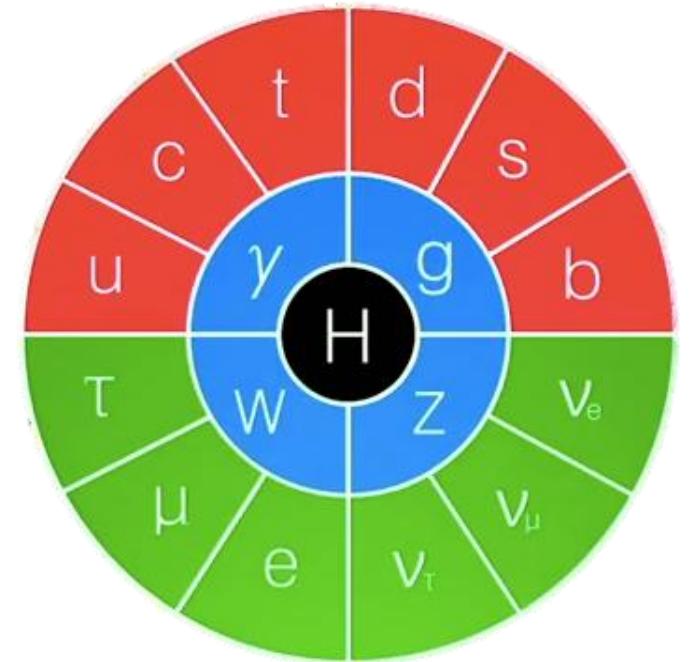
Example of an inlier / over-density

Anomalous tracks in the
desert

Need to know your **normal** events before you can look for **anomalous** events



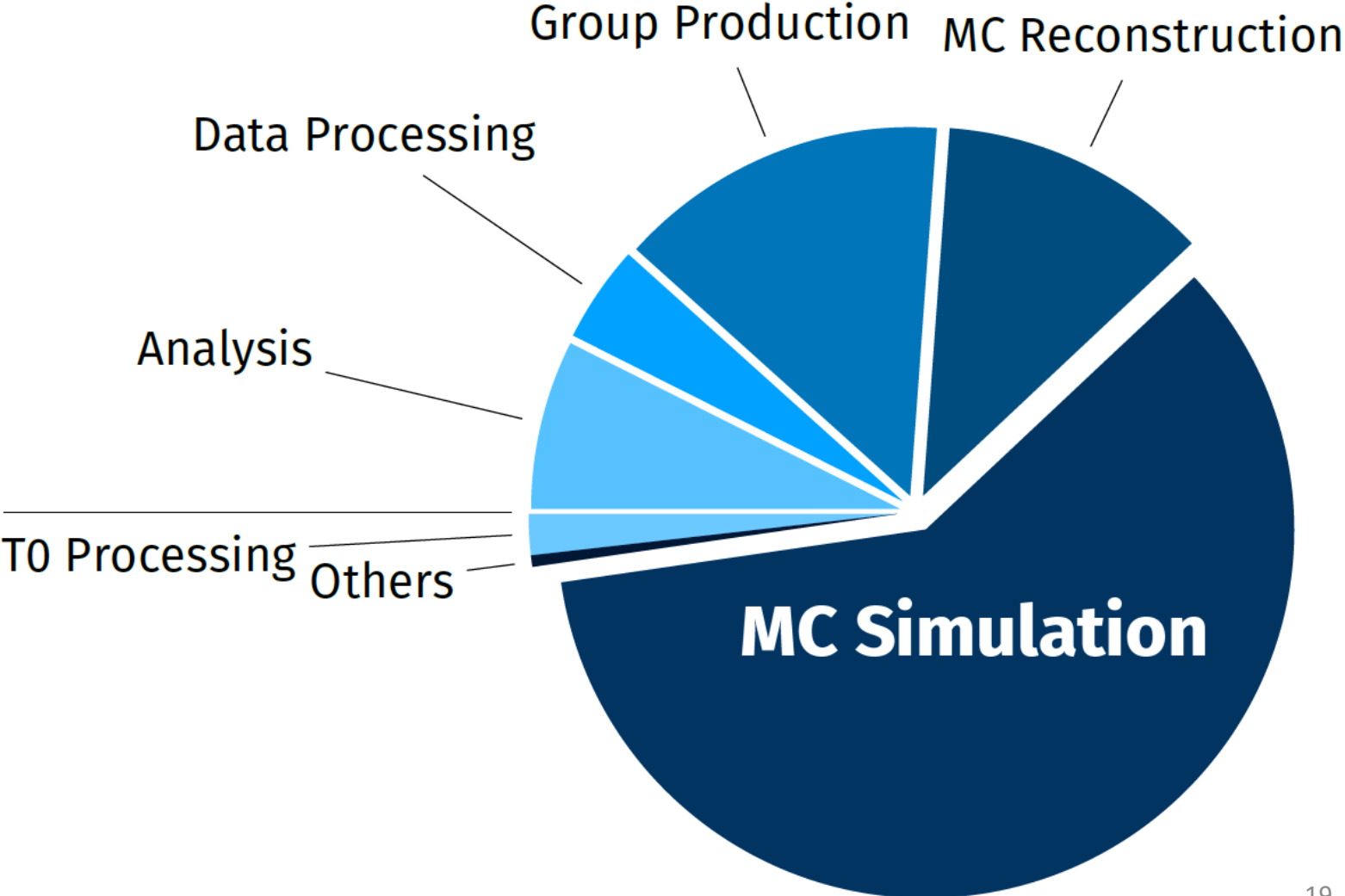
- Model of the desert



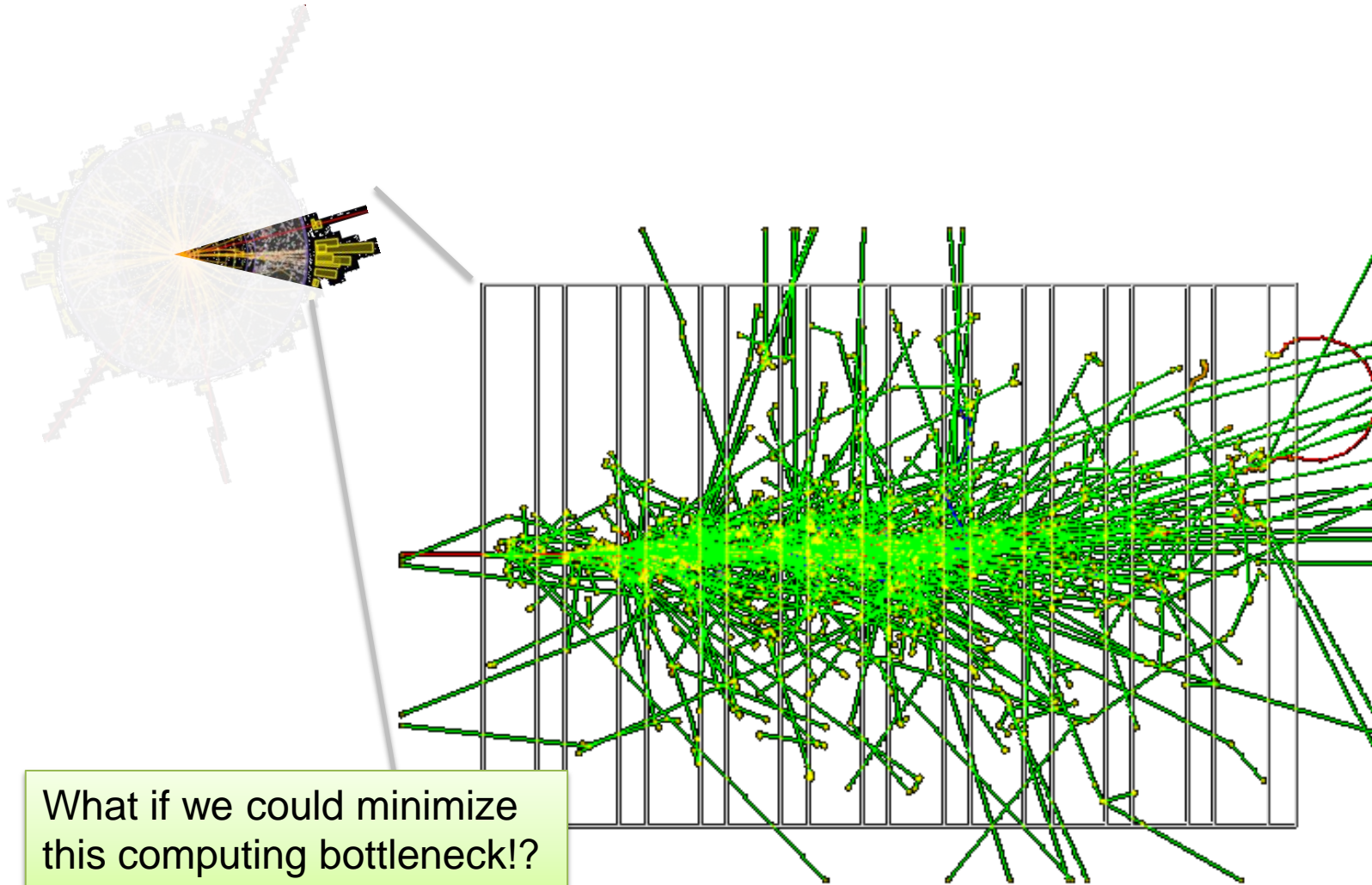
- Model of our SM events

Forward Monte Carlo modeling

**Computing bottleneck:
Monte Carlo simulation**



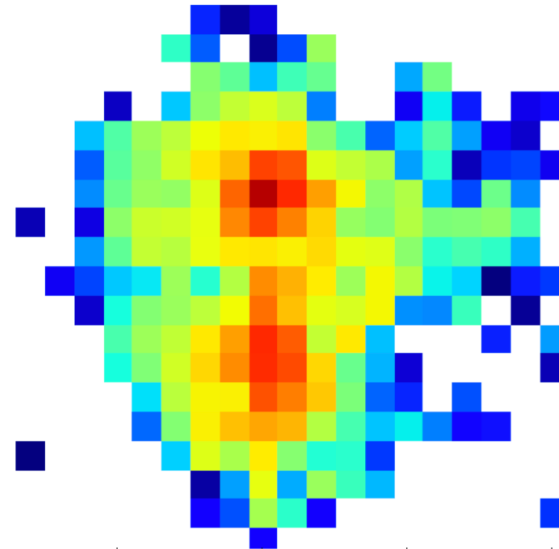
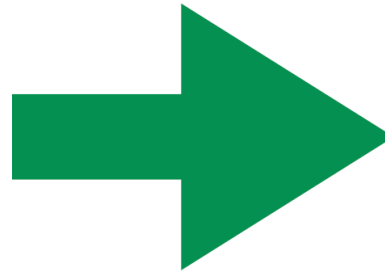
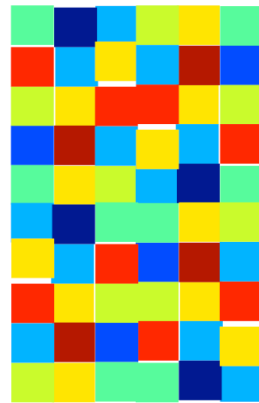
One particle entering the calorimeter...



- **Geant4**: simulate at **microscopic level** interaction of particles with matter
- Bottleneck: calorimeter simulation - **$O(10 \text{ min})$ per 1 event**
- \Rightarrow Need **trillions** of simulated events

Toolbox: generative modeling

Build a **generator*** which maps random numbers to structure



$$p_{\text{model}} \approx p_{\text{data}}$$

*Deep generative NN model:

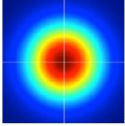
- Generative Adversarial Network (GANs)
- Normalizing Flows (NFs)
- **Variational Autoencoders (VAEs)**

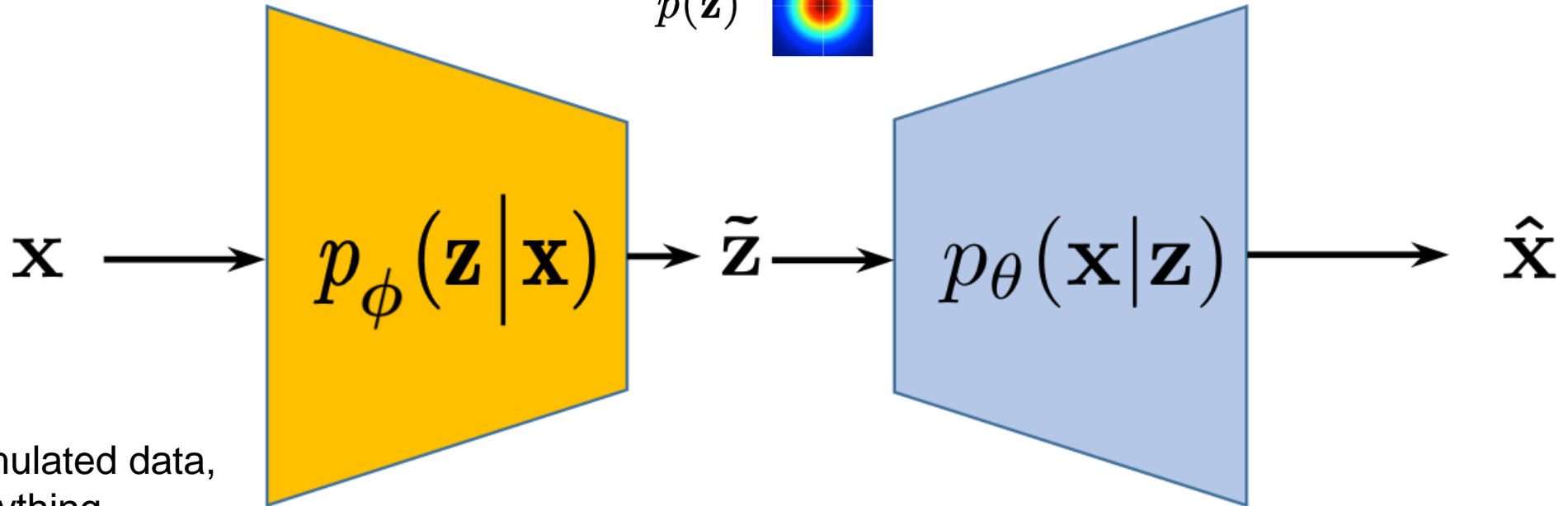
Toolbox: Variational Autoencoder (VAE)

Probabilistic encoder:
reduce dimensions

Latent space (with given prior):
easy to sample from

Probabilistic decoder:
Reconstruct input

$p(\mathbf{z})$ 

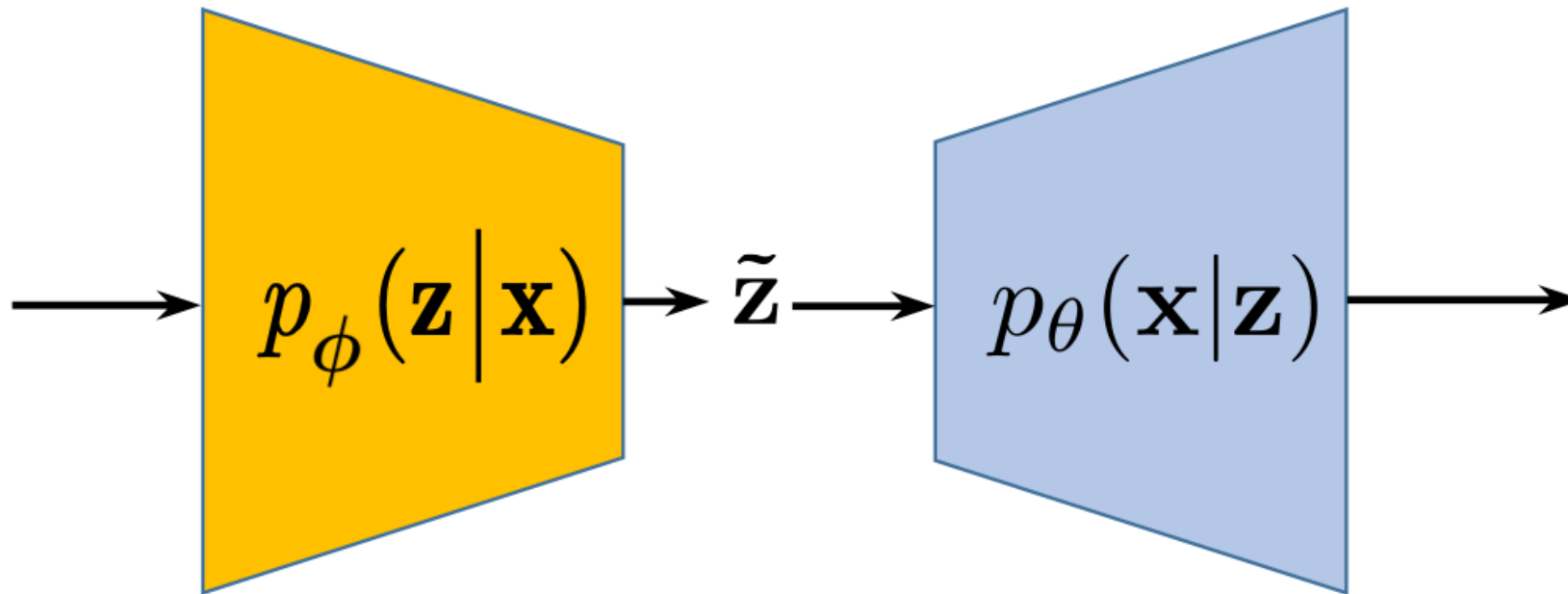


Input x:
Raw data, simulated data,
features,...anything

Information bottleneck:
maximize encoded information

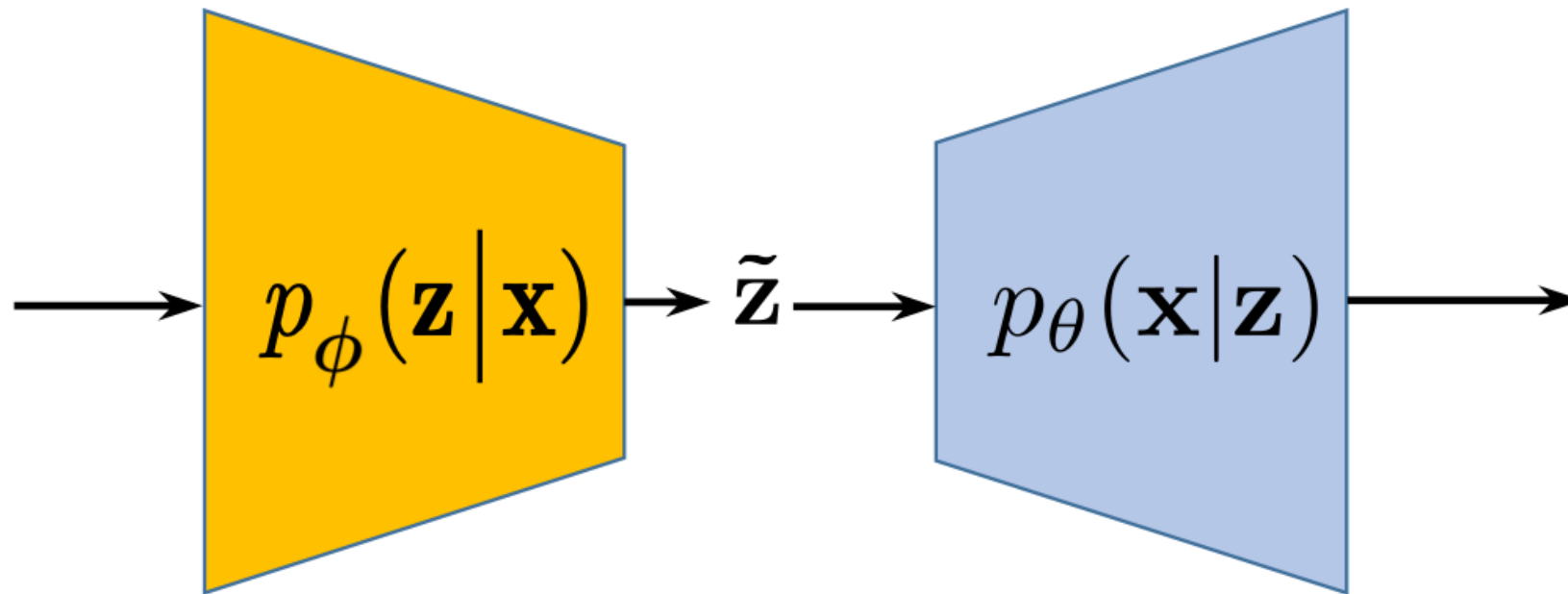
1. Reconstruction mode
2. Generation mode

[Data volume reduction]



- Lossy compression with auto encoders
- Only maintain key features in data
- Example application in PP: trigger
 - reduce bandwidth to increase event rate

Reconstruction mode

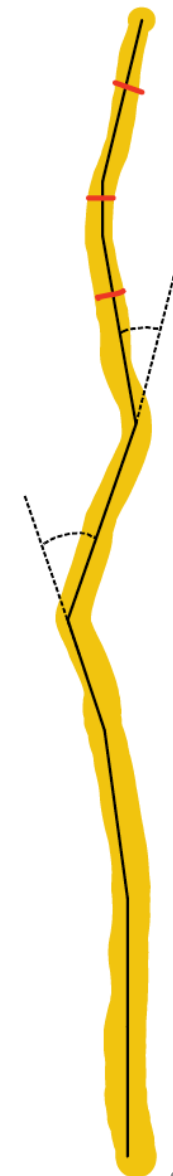
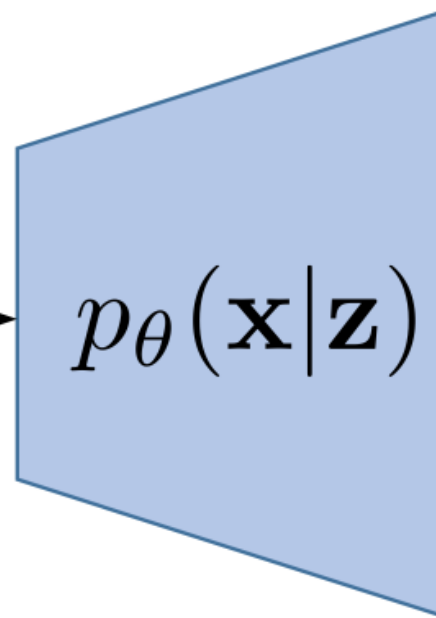
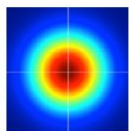


- Train on hay
- Apply to data: poor reconstruction for non-hay = anomaly

Generation mode

Sample from:

$p(\mathbf{z})$

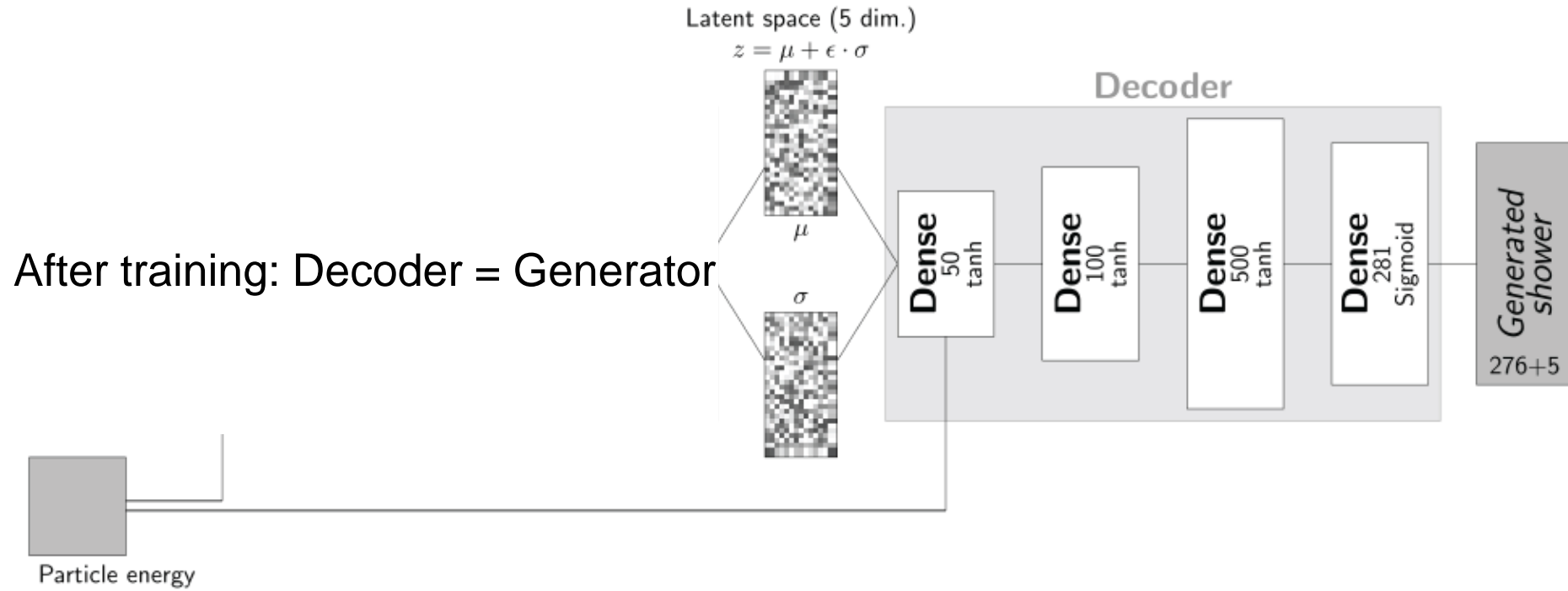


- Train on hay in reco mode
- **Rapidly** sample hay from a normal distribution

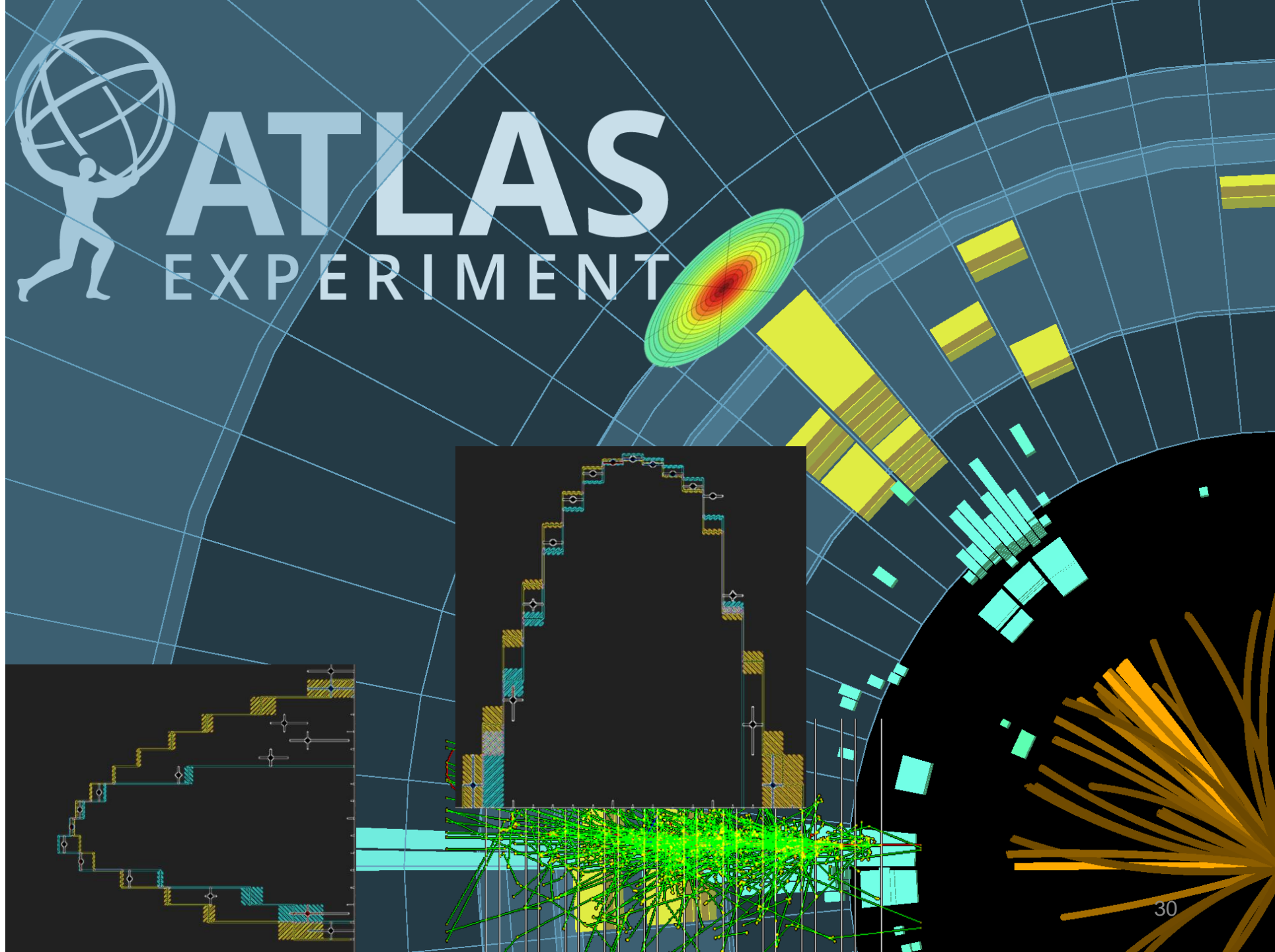
These faces do not exist !



VAE architecture



Validation:
marginals

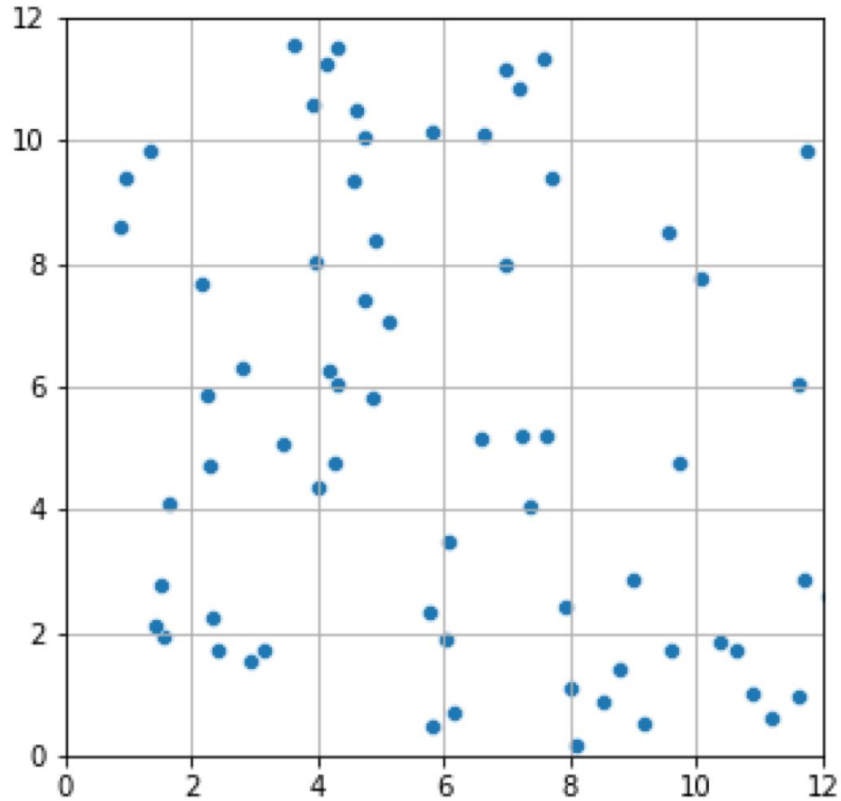


Generative modeling assessment

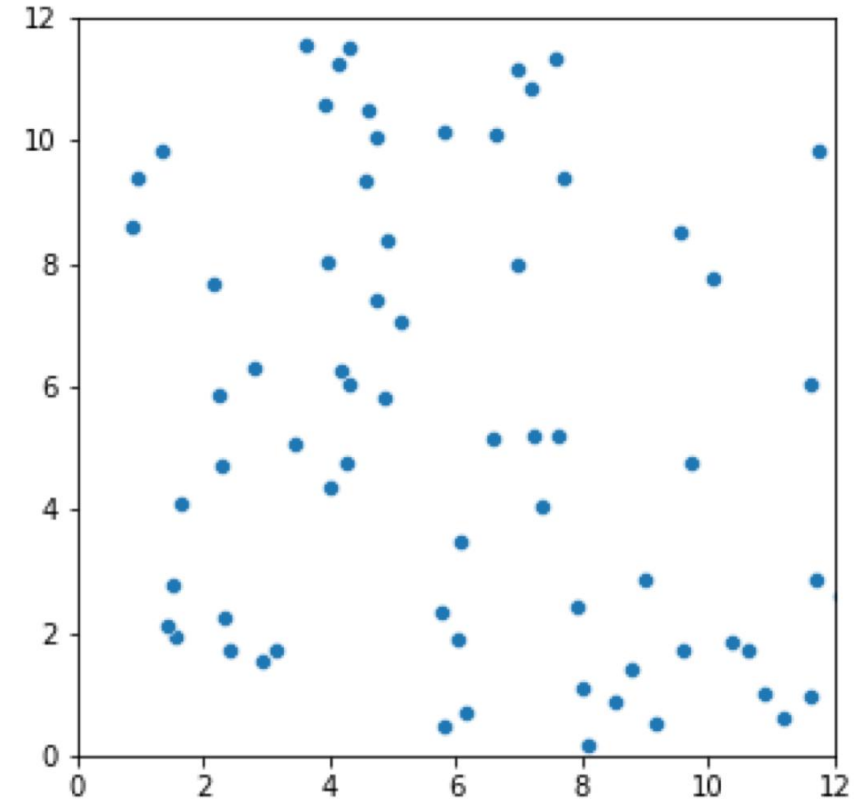
- Promising results but bottlenecks exist:
 - **Slow** development cycle
 - **Expensive & inflexible** training data (Geant4)
 - **Non-portable solution** highly dependent on detector geometry*
- Objectives:
 - Faster R&D
 - Decouple modeling from detector geometry → **point cloud format**

* A Common Tracking Software (ACTS) – portable tracking solution

Geant4 point cloud exists already



Current: mapping to fixed cells (**sparse**)
Intensity = sum of energy in each cell



Geant4 raw output: point cloud

The world of point-cloud data sets



[\[source\]](#)

Sweet
spot
?



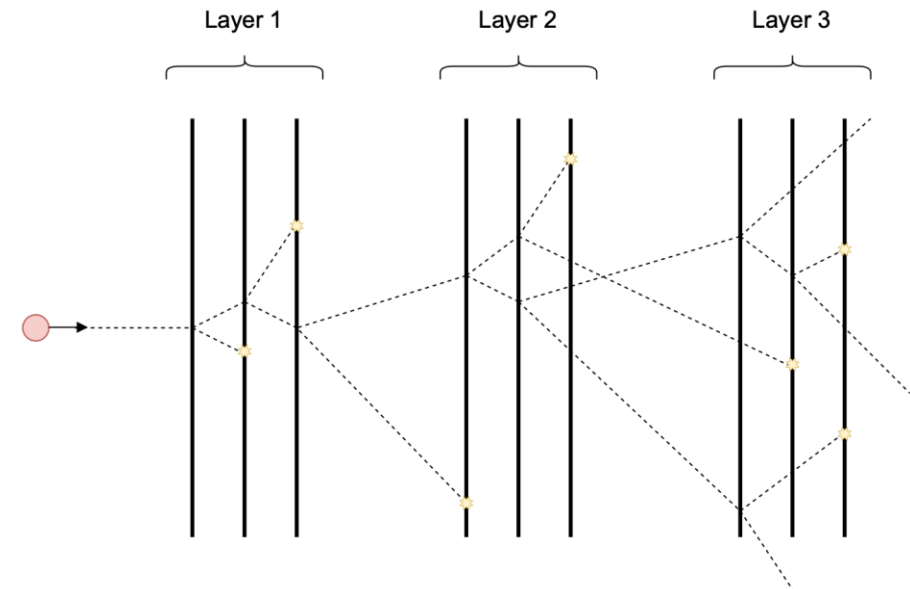
- Existing public point cloud data sets
 - Not a good proxy for physics data
 - Improvements don't *generalize*
- Costly and expertise-requiring Geant4 simulation
 - Hard to scale complexity, change geometry, detector,...

The for generative modeling

- SUPA*
 - Flexible & configurable proxy data sets
- Diagnostics tool to develop new generative surrogate simulators
- Point-cloud format promotes *GNN-based* generative models

Simplified

- particle propagation,
- scattering &
- shower development



*SURrogate PArticle propagation simulator

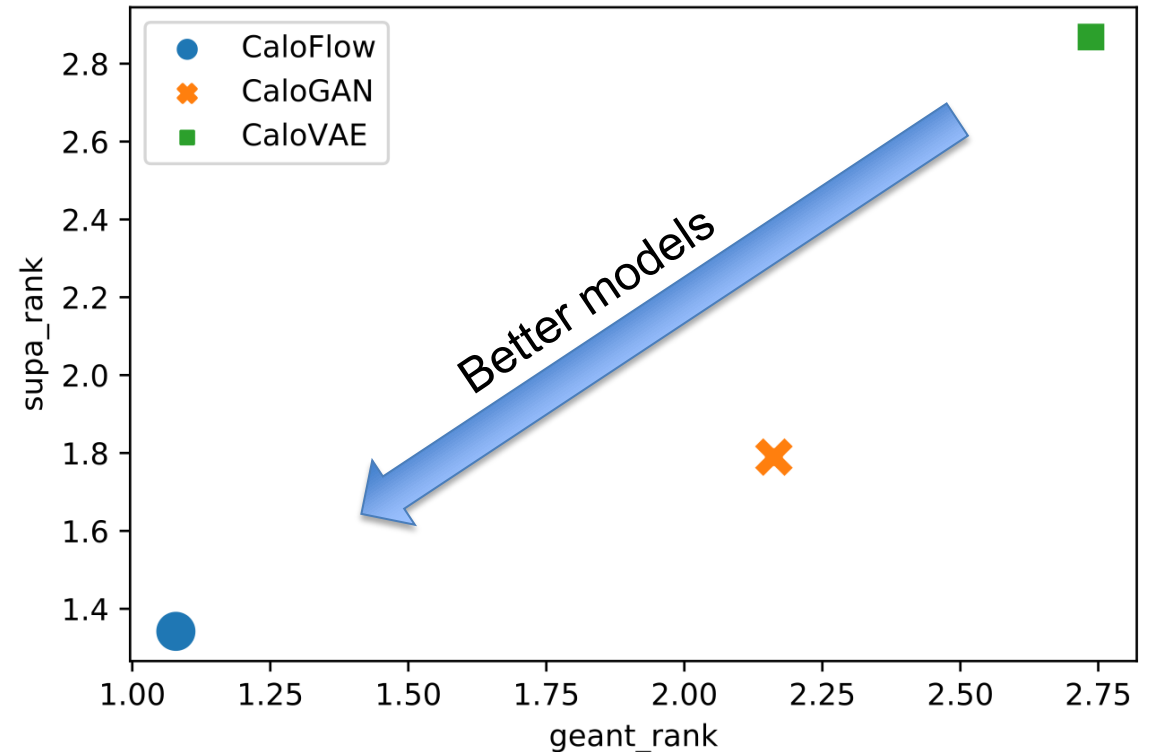
SUPA is *realistic enough*

- Improvements on SUPA translate to Geant4

Model design on SUPA:

- Vary data complexity
- Optimize model
- Validation metrics

SUPA tracks improvements of model on Geant4



Modeling vs. learning

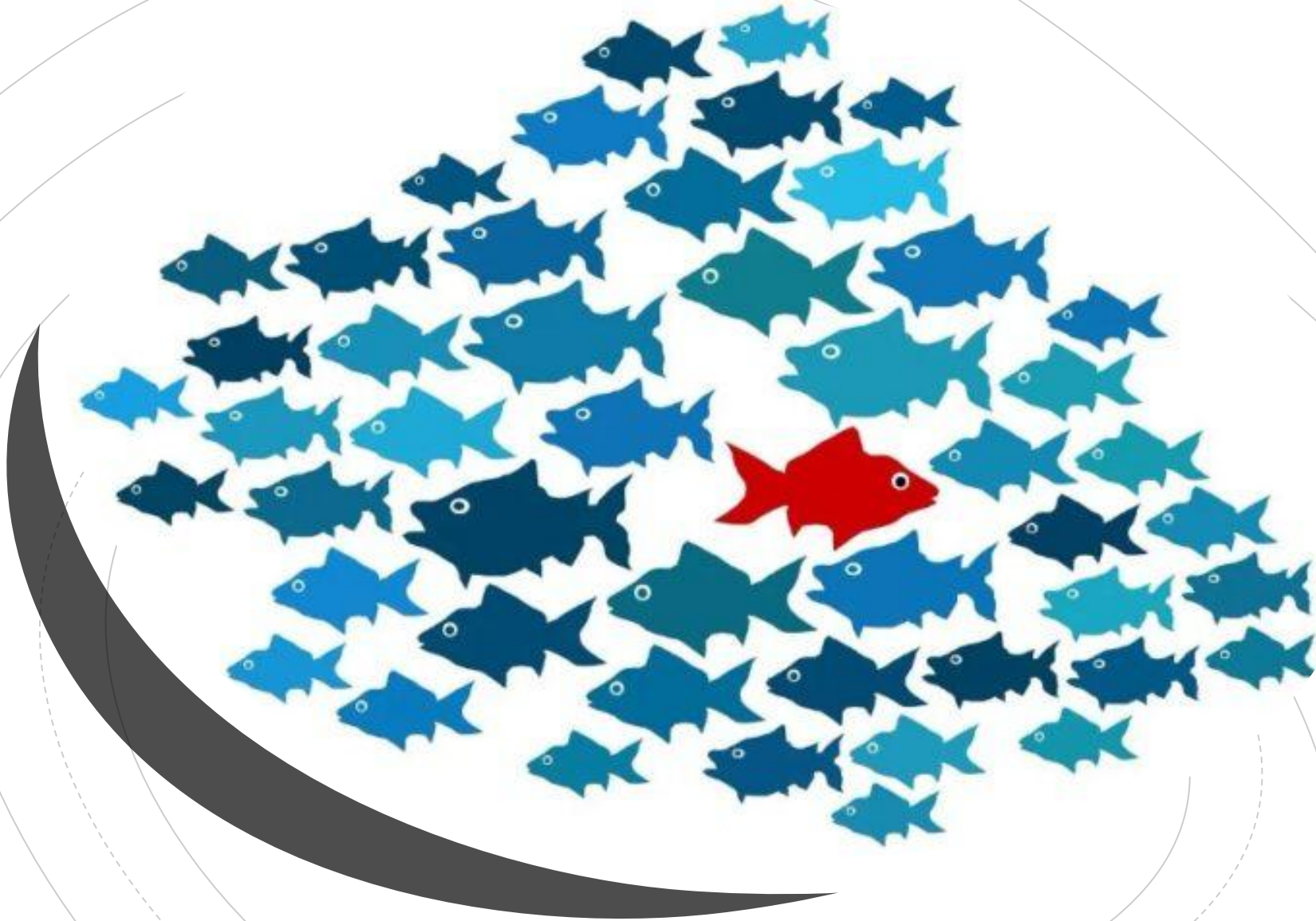
The world of modeling

- The Standard Model of particle physics
- High-fidelity Monte Carlo simulation
- Fast & accurate surrogate models

The world of learning

- Learning from **lots** of LHC data

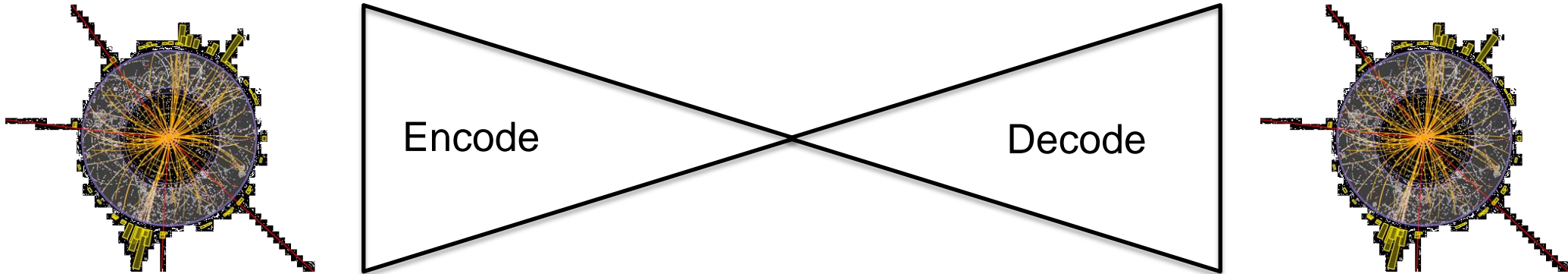
The best of both worlds?



**Outlier
detection**

VAE in reconstruction mode: search for anomalous boosted objects

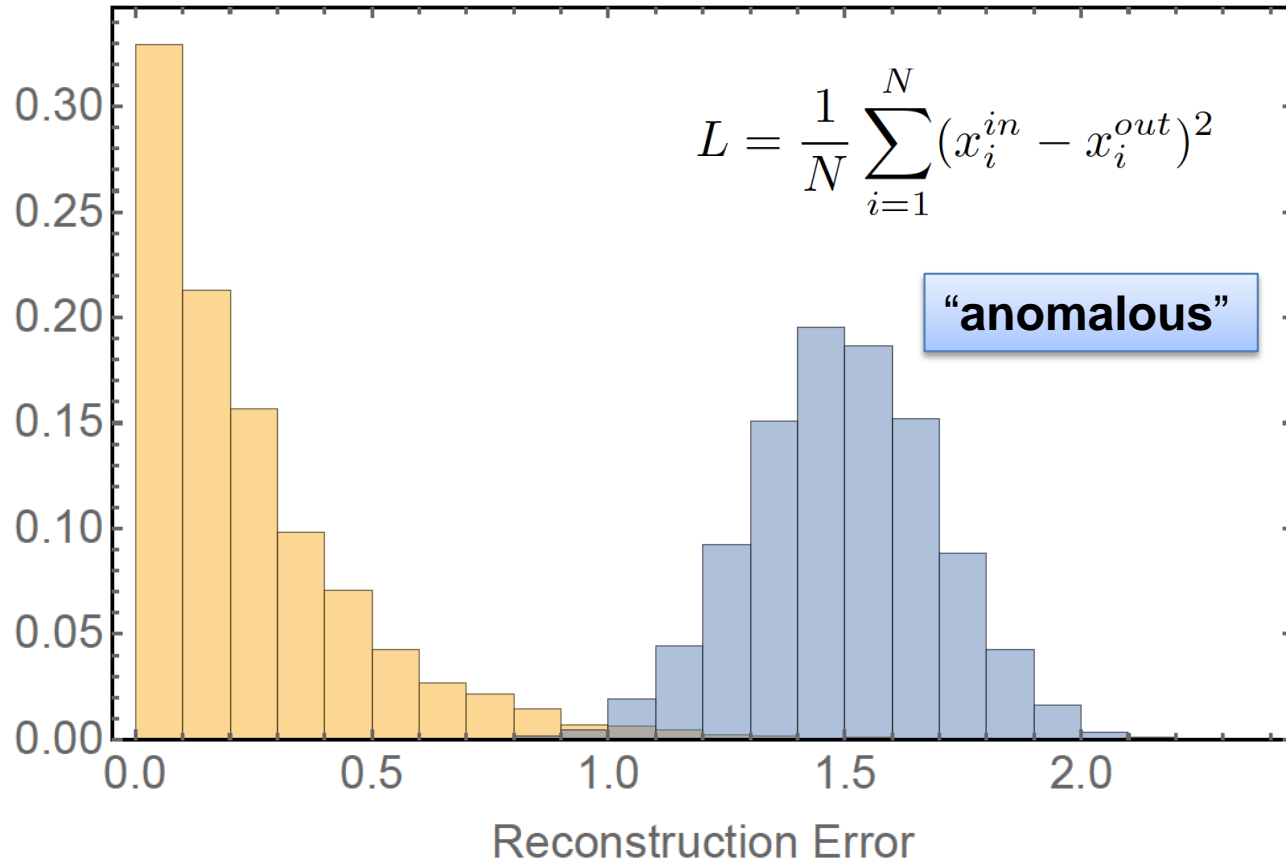
Encode and decode “normal” objects / events



Compare original and reconstructed image

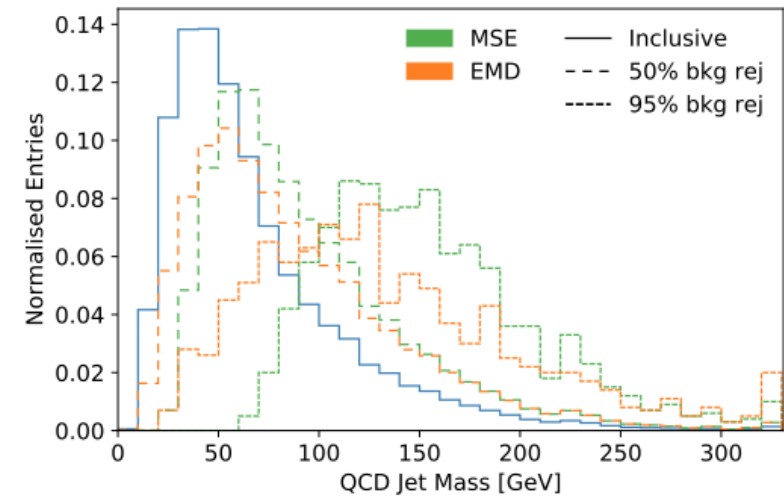
Anomalous jets

“normal”



Challenge:

- Tool picks up mainly on *dominant* difference, i.e. the mass of the anomalous jet



[https://ml4physicalsciences.github.io/2020/files/NeurIPS_ML4PS_2020_56.pdf]

[1709.01087, 1808.08979, 1808.08992, 1905.12651, 2007.01850]

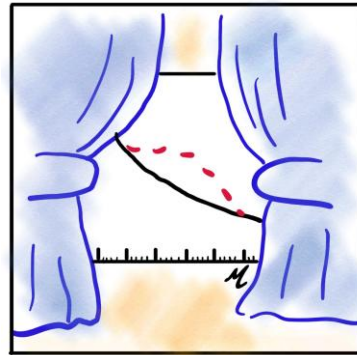
The problem with outlier detection

- Rarely *true* outliers in our data
- We look for an excess = over-density



Constructing Unobserved Regions by Transforming Adjacent Intervals

*All windows need **CURTAINS***



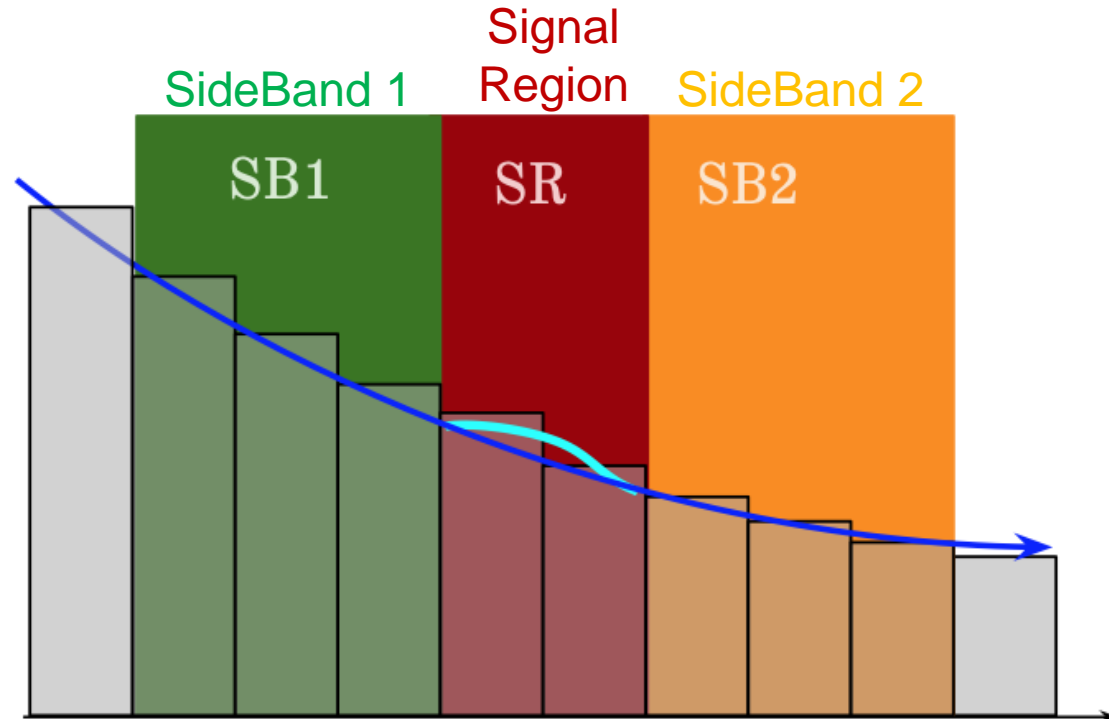
Data driven method for constructing
background templates with arbitrary variables

Bump hunt

Focus on resonant signal = **bump**

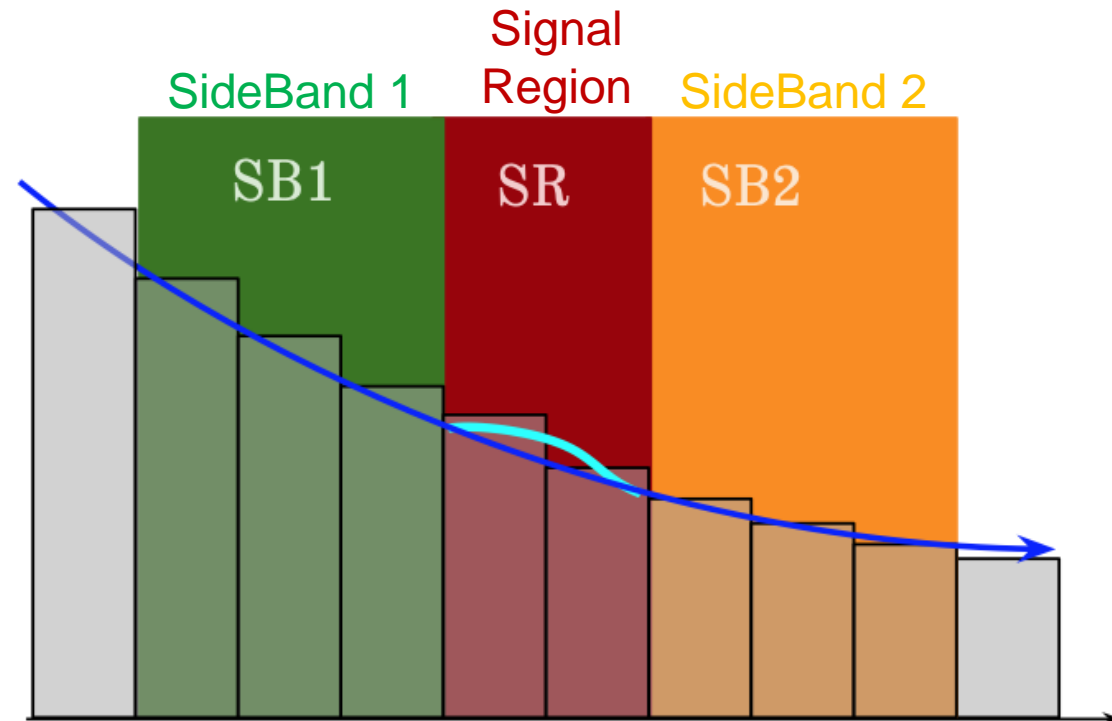
Method:

1. Split spectrum into sliding SBs
2. Fit the distribution in SBs
3. Interpolate into the SR
4. Look for an excess



Extended bump hunt

- Looking for tiny signal
- Increase sensitivity to new physics
 - \Rightarrow **use additional observables**
- Observables often **strongly correlated to the mass**
- **Interpolate** to find BG template in SR

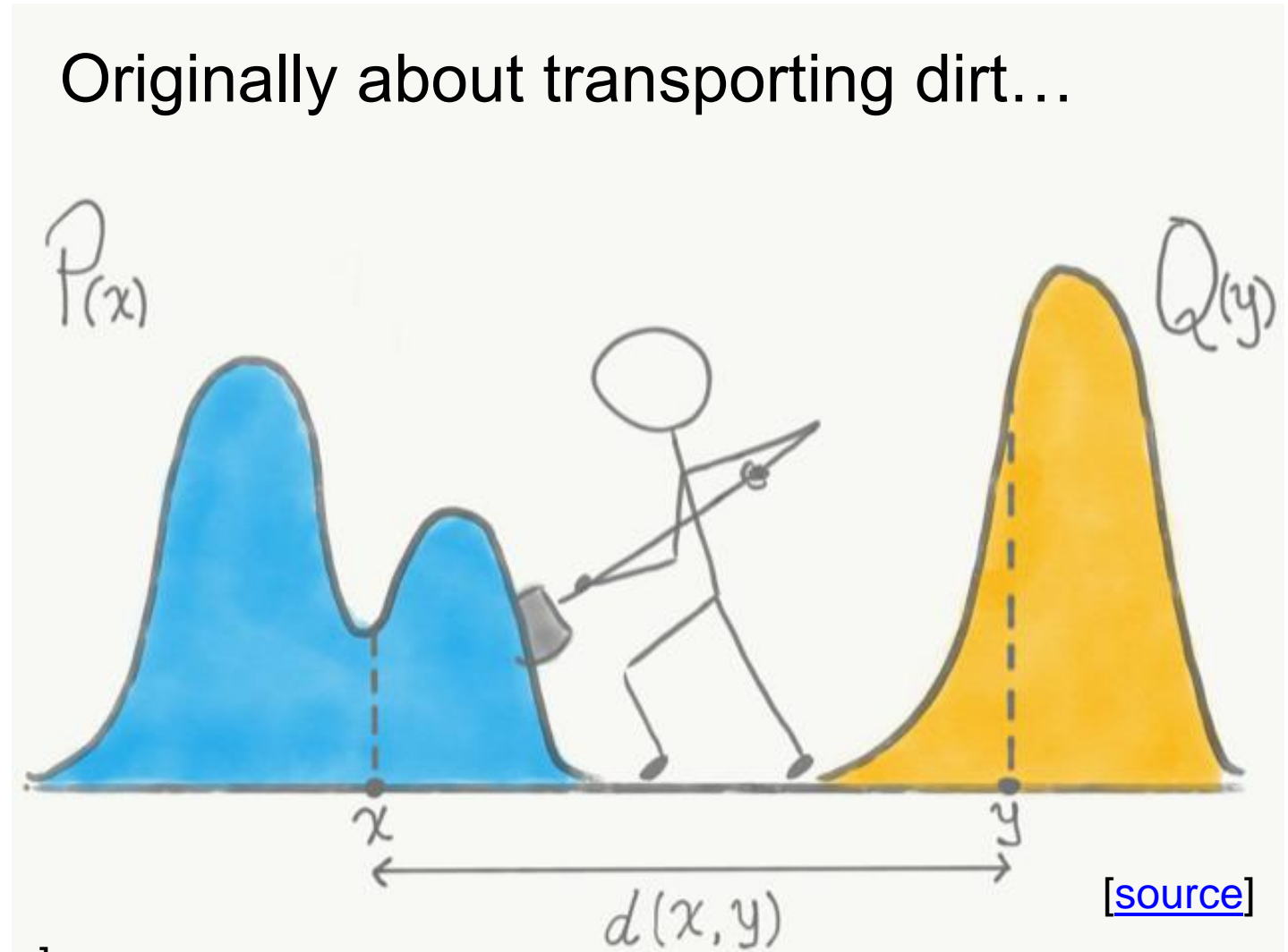


CURTAINS approach

1. **Transform** data from **the SBs** into the **SR**
2. Transformed side bands = background template
3. Train a classifier to separate background from *signal*

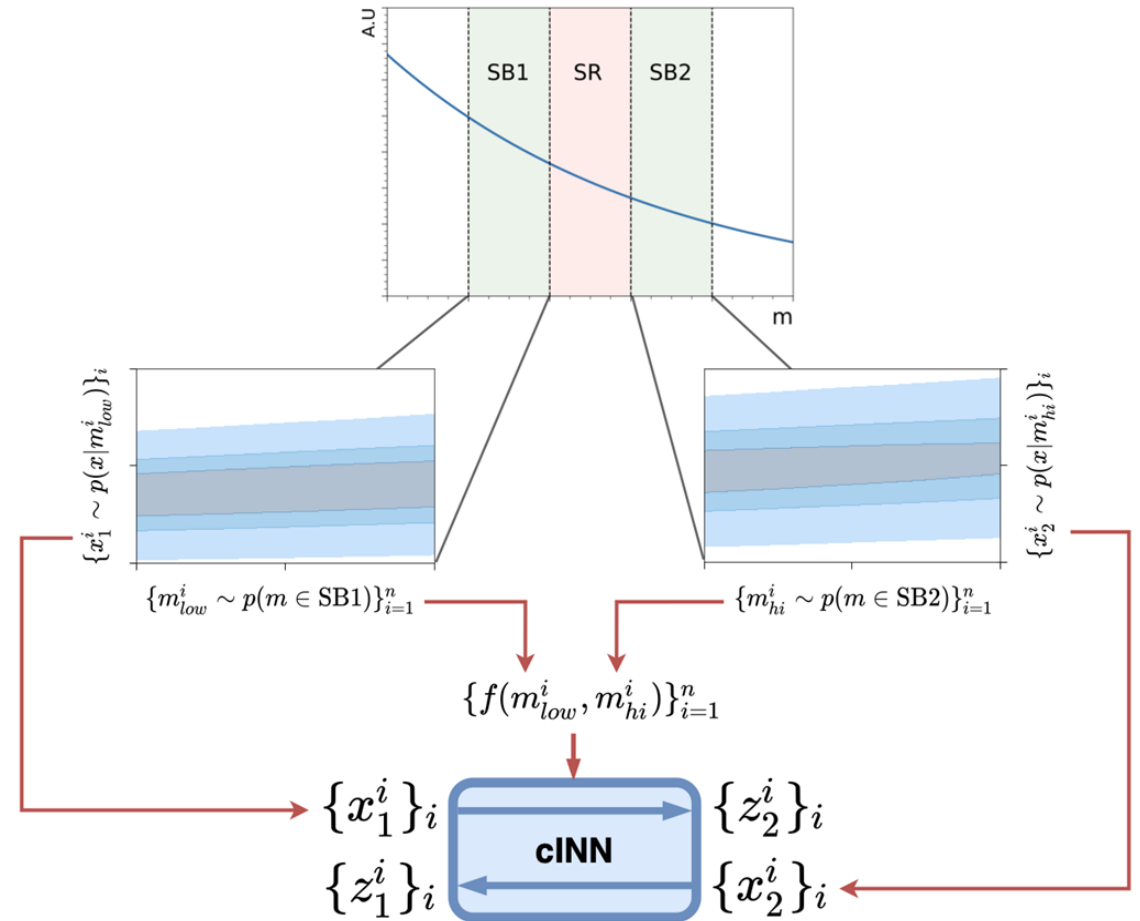
Toolbox: optimal transport

- Transforming \mathbf{P} into \mathbf{Q} while minimizing a cost
- Cost based on **distance d** between data points



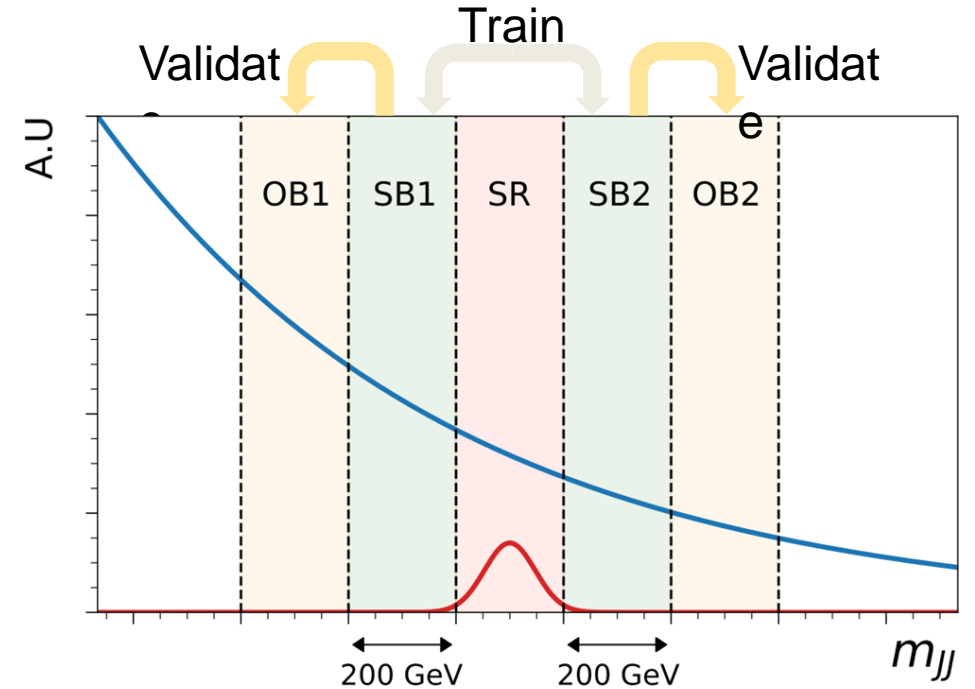
Training “SB-to-SR” transformation

- Use a **conditional invertible** neural network (cINN)
- Map from SB1 to SB2 and vice versa



CURTAINS validation

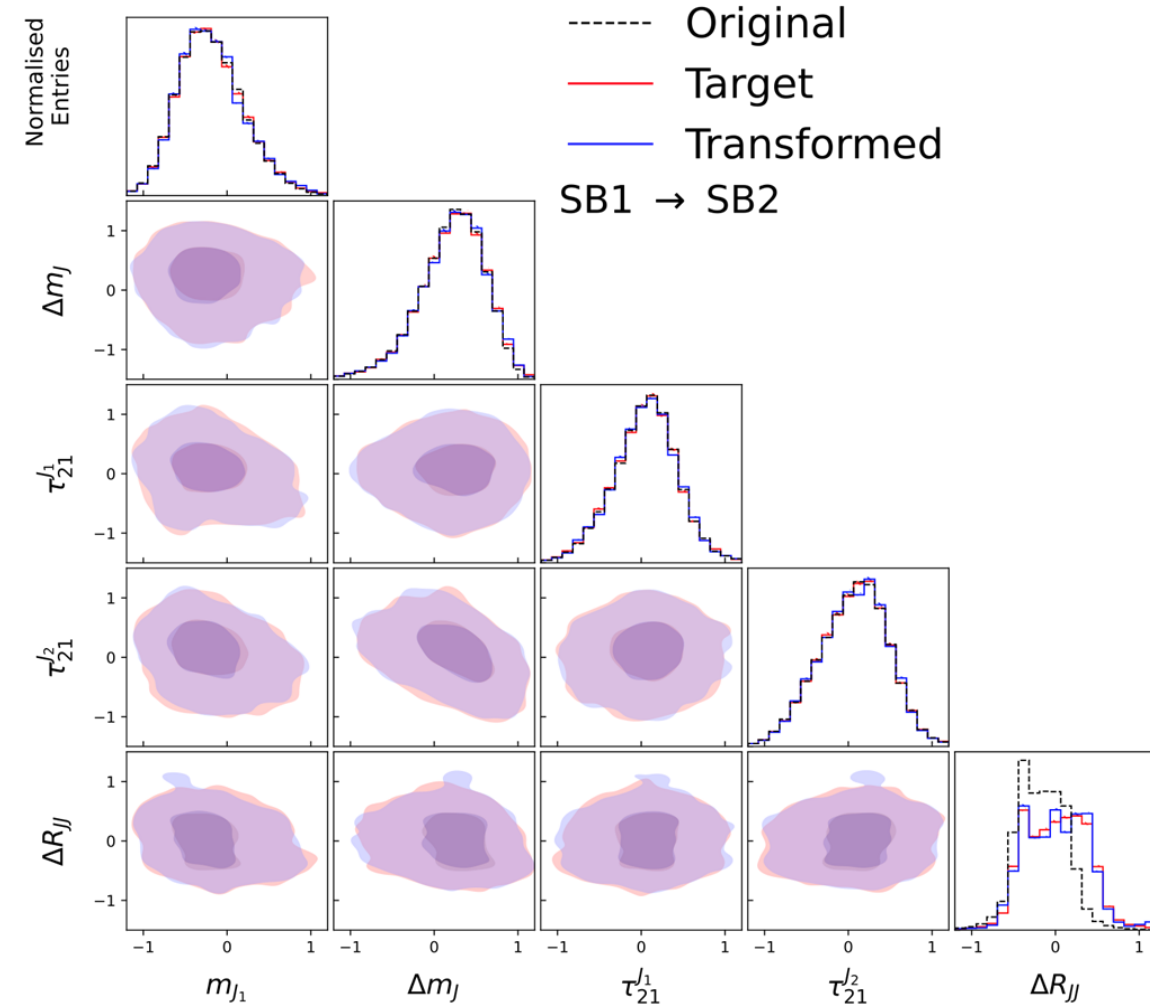
- Fix sidebands
- Define OuterBand (OB) validation regions
- Train CURTAINS transformer
- Validate on OBs



Training data

SB1: [3200, 3400] GeV
SB2: [3600, 3800] GeV

- Training on the LHC Olympics R&D dijet dataset*
 - Based on jet substructure & ΔR_{jj}
- SB1 \rightarrow SB2
 - as good for SB2 \rightarrow SB1, OBs, SR

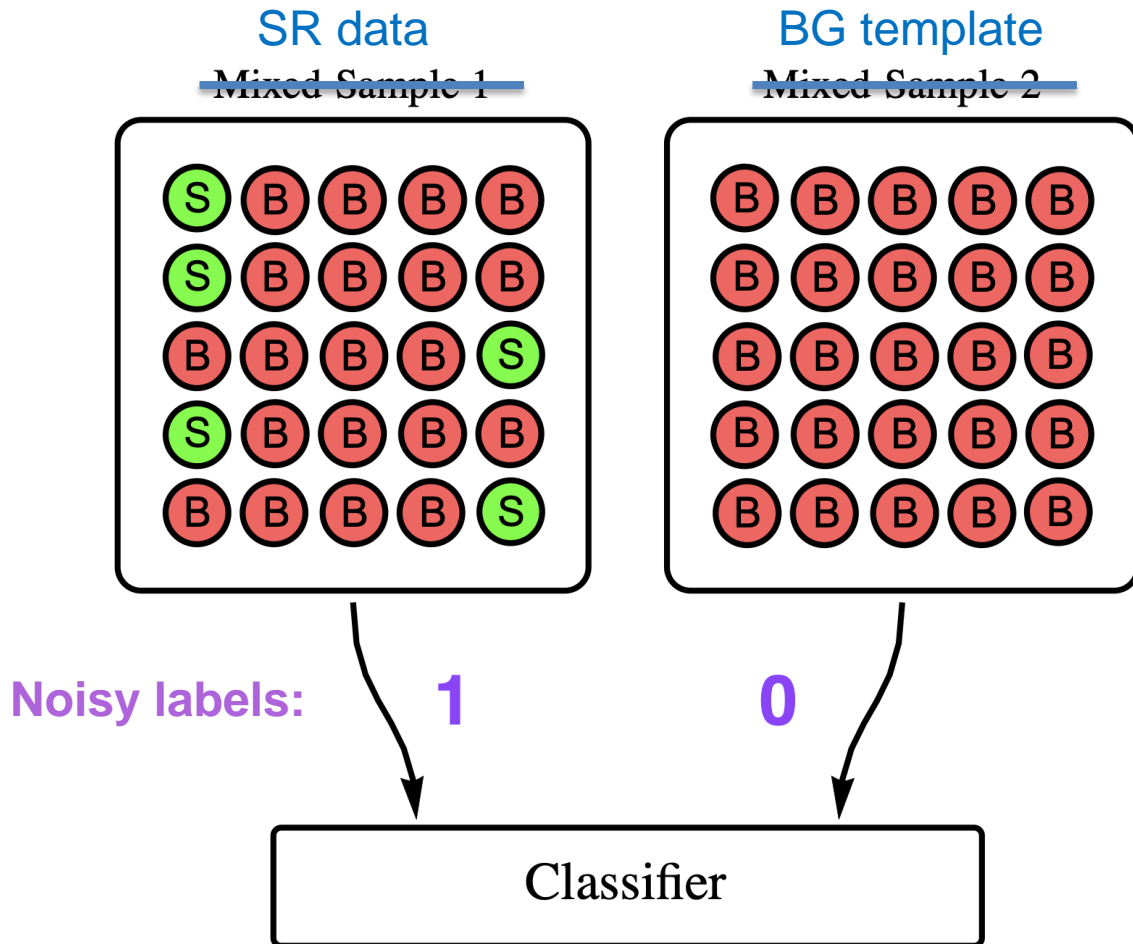


*[<https://doi.org/10.5281/zenodo.4536377>]

CURTAINS so far

- ✓ Transform data from the SBs into the SR
- ✓ Transformed side bands = background template
- Train a classifier to separate background from *signal*

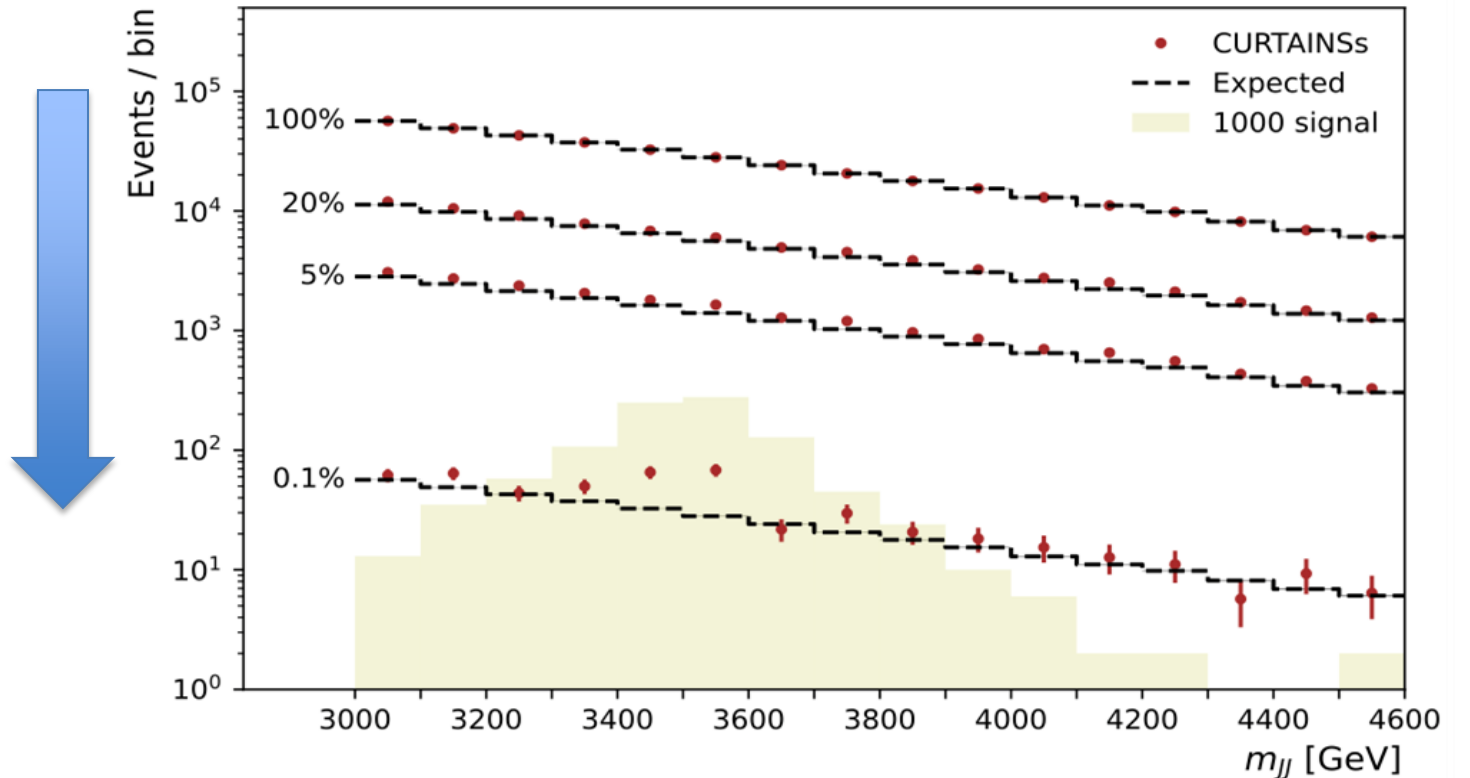
Classification without labeling (CWoLa)



- Use noisy labels
- Shown to be optimal classifier
- Apply to data-only
- CWoLa for CURTAINS

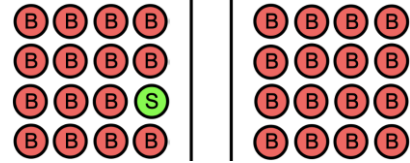
CURTAINS in action

- True BG (Expected)
- Predicted BG from CURTAINS
- Add signal

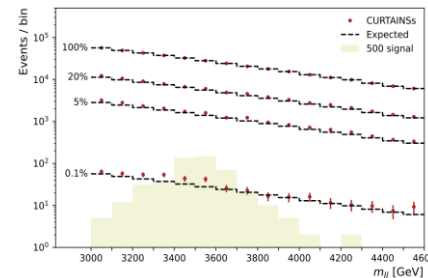
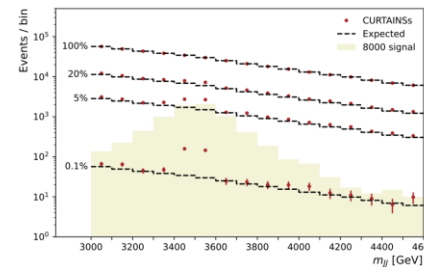
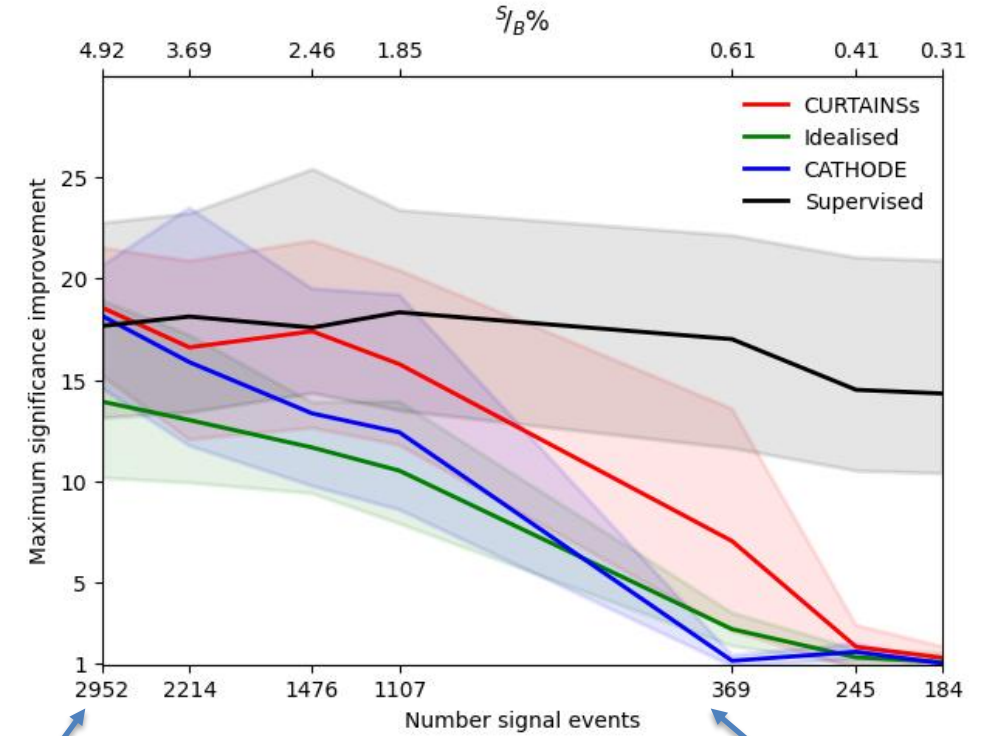


Apply cut on
CWoLa classifier

CURTAINS performance



- CURTAINSs
- Idealised: assume perfect BG template
- CATHODE
 - Competition generating BG template from density estimates
- Supervised



[CURTAINSs > Idealised due to oversampling]

Summary

- *Extend* LHC's physics portfolio to anomaly detection
- Key: robust background estimate
 - Data-derived: CURTAINS
 - MC modeling: speed & accuracy with generative models
 - Work in progress: **combine modeling & learning**
- Promote automation & reduce complexity